WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at
http://www.mrw.interscience.wiley.com/eot

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor.
p. cm.
includes index.
ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G.
TK5102 .W55 2002
621.382′03 — dc21

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

I am pleased to welcome the readers to the *Wiley Encyclopedia of Telecommunications*. The Encyclopedia contains 275 tutorial articles focused on modern telecommunications topics. The contents include articles on communication networks, source coding and decoding, channel coding and decoding, modulation and demodulation, optical communications, satellite communications, underwater acoustic communications, radio propagation, antennas, multiuser communications, magnetic storage systems, and a variety of standards. Additional details on these topics are given below. The authors of these articles were selected for their expertise and leadership in their respective areas in the field of telecommunications. All of the authors possess advanced graduate degrees and have published widely in prestigious international journals and conferences.

COMMUNICATION NETWORKS

There are approximately 60 articles on the subject of communication networks, including several articles on protocols, such as the wireless application protocol (WAP) and MAC protocols; network flow; congestion control; admission control broadband integrated digital networks; local area networks and standards; satellite networks; network reliability and fault tolerance; DWDM ring networks, wireless ad hoc networks; and multi-protocol label switching (MPLS).

MODULATION AND DEMODULATION

There are over 40 articles covering various basic modulation and demodulation techniques, including analog amplitude modulation (AM), frequency modulation (FM) and phase modulation; digital modulation techniques, namely, pulse amplitude modulation (PAM); phase-shift keying (PSK), quadrature amplitude modulation (QAM), continuous-phase modulation (CPM); continuous-phase frequency shift-keying (CPFSK); partial response signals; spread spectrum modulation; adaptive equalization; turbo equalization; and orthogonal frequency-division multiplexing (OFDM).

OPTICAL COMMUNICATIONS

There are approximately 30 articles about optical communication, including articles on optical modulation; optical detectors; optical amplifiers; optical correlators; optical filters; photonic A/D conversion; optical transport; optical multiplexers and demultiplexers, optical switching; and characterization of optical fibers.

ANTENNAS

Various types of antennas and antenna arrays are described in 10 articles, including parabolic antennas; microstrip antennas; waveguide antennas; television antennas; loop antennas; horn antennas; leaky wave antennas; and helical and spiral antennas.

PROPAGATION

Six articles are devoted to electromagnetic radio signal propagation, including propagation at very low frequencies (VLF), low frequencies (LF), medium frequencies (MF), high frequencies (HF), very high frequencies (VHF), microwave frequencies, and millimeter wave frequencies.

CHANNEL CODING AND DECODING

Approximately 35 articles cover various channel codes and decoding algorithms, including BCH codes; convolutional codes; concatenated codes; trellis codes; space-time codes; turbo codes; Gold codes; Kasami codes; Golay codes; finite geometry codes; codes for magnetic recording channels; Viterbi decoding algorithm; and sequential decoding algorithm.

SOURCE CODING AND DECODING

Eight articles cover various data compression and source coding and decoding methods, including waveform coding techniques such as pulse code modulation (PCM) and differential PCM (DPCM); linear predictive coding (LPC); Huffman coding; and high definition television (HDTV).

MULTIUSER COMMUNICATION

There are 12 articles focused on multiuser communications, including multiple access techniques such as code-division multiple access (CDMA), frequency-division multiple access (FDMA), time-division multiple access (TDMA), and carrier-sense multiple access (CSMA); Ethernet technology; multiuser detection algorithms; and third-generation (3G) digital cellular communication systems.

ACOUSTIC COMMUNICATIONS

There are 5 articles on acoustic communications dealing with acoustic transducers; underwater acoustic communications and telemetry; underwater acoustic moderns, and acoustic echo cancellation.

SATELLITE COMMUNICATIONS

Two articles focus on geosynchronous satellite communications and on low-earth-orbit (LEO) and medium-earth-orbit (MEO) satellite communication systems.

John G. Proakis, Editor
Northeastern University

CONTRIBUTORS

Behnaam Aazhang, *Rice University, Houston, Texas,* Multiuser Wireless Communication Systems

Ali N. Akansu, *New Jersey Institute of Technology, Newark, New Jersey,* Orthogonal Transmultiplexers: A Time-Frequency Perspective

Nail Akar, *Bilkent University, Ankara, Turkey,* BISDN (Broadband Integrated Services Digital Network)

Arda Aksu, *North Carolina State University, Raleigh, North Carolina,* Unequal Error Protection Codes

Naofal Al-Dhahir, *AT&T Shannon Laboratory, Florham Park, New Jersey,* Space–Time Codes for Wireless Communications

Edward E. Altshuler, *Electromagnetics Technology Division, Hanscom AFB, Massachusetts,* Millimeter Wave Propagation

Abeer Alwan, *University of California at Los Angeles, Los Angeles, California,* Speech Coding: Fundamentals and Applications

Moeness G. Amin, *Villanova University, Villanova, Pennsylvania,* Interference Suppression in Spread-Spectrum Communication Systems

John B. Anderson, *Lund University, Lund, Sweden,* Continuous-Phase-Coded Modulation

Alessandro Andreadis, *University of Siena, Siena, Italy,* Wireless Application Protocol (WAP)

Peter Andrekson, *Chalmers University of Technology, Göthenburg, Sweden,* Optical Solitons

Oreste Andrisano, *University of Bologna, DEIS, Italy,* Communications for Intelligent Transportation Systems

A. Annamalai, *Virginia Tech, Blacksburg, Virginia,* Wireless Communications System Design

Cenk Argon, *Georgia Institute of Technology, Atlanta, Georgia,* Turbo Product Codes for Optical CDMA Systems

Hüseyin Arslan, *Ericsson Inc., Research Triangle Park, North Carolina,* Channel Tracking in Wireless Communication Systems

Tor Aulin, *Chalmers University of Technology, Göteborg, Sweden,* Serially Concatenated Continuous-Phase Modulation with Iterative Decoding

James Aweya, *Nortel Networks, Ottawa, Ontario, Canada,* Transmission Control Protocol

Ender Ayanoglu, *University of California, Irivne, California,* BISDN (Broadband Integrated Services Digital Network)

Krishna Balachandran, *Lucent Technologies Bells Labs, Holmdel, New Jersey,* Wireless Packet Data

Constantine A. Balanis, *Arizona State University, Tempe, Arizona,* Antennas

Stella N. Batalama, *State University of New York at Buffalo, Buffalo, New York,* Packet-Rate Adaptive Receivers for Mobile Communications

Rainer Bauer, *Munich University of Technology (TUM), Munich, Germany,* Digital Audiobroadcasting

S. Benedetto, *Politecnico di Torino, Torino (Turin), Italy,* Serially Concatenated Codes and Iterative Algorithms

Toby Berger, *Cornell University, Ithaca, New York,* Rate-Distortion Theory

Steven Bernstein, *MIT Lincoln Laboratory, Lexington, Massachusetts,* Communication Satellite Onboard Processing

Claude Berrou, *ENST Bretagne, Brest, France,* Turbo Codes

Randall Berry, *Northwestern University, Evanston, Illinois,* Information Theory

H. L. Bertoni, *Polytechnic University, Brooklyn, New York,* Path Loss Prediction Models in Cellular Communication Channels

Christian Bettstetter, *Technische Universität München, Institute of Communication Networks, Munich, Germany,* General Packet Radio Service (GPRS); GSM Digital Cellular Communication System

Ravi Bhagavathula, *Wichita State University, Wichita, Kansas,* Modems

Andrea Bianco, *Politecnico di Torino, Torino (Turin), Italy,* Multimedia Networking

Jayadev Billa, *BBN Technologies, Cambridge, Massachusetts,* Speech Recognition

Bjørn A. Bjerke, *Qualcomm, Inc., Concord, Massachusetts,* Pulse Amplitude Modulation

Fletcher A. Blackmon, *Naval Undersea Warfare Center Division Newport, Newport, Rhode Island,* Acoustic Telemetry

Ian F. Blake, *University of Toronto, Ontario, Toronto, Canada,* Cryptography

Martin Bossert, *University of Ulm, Ulm, Germany,* Hadamard Matrices and Codes

Gregory E. Bottomley, *Ericsson Inc., Research Triangle Park, North Carolina,* Channel Tracking in Wireless Communication Systems

Torsten Braun, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Madhukar Budagavi, *Texas Instruments, Incorporated, Dallas, Texas,* Wireless MPEG-4 Videocommunications

Kenneth Budka, *Lucent Technologies Bells Labs, Holmdel, New Jersey,* Wireless Packet Data

R. Michael Buehrer, *Virginia Tech, Blacksburg, Virginia,* Mobile Radio Communications

Julian J. Bussgang, *Signatron Technology Corporation, Concord, Massachusetts,* HF Communications

Jens Buus, *Gayton Photonics, Gayton, Northants, United Kingdom,* Optical Sources

Søren Buus, *Northeastern University, Boston, Massachusetts,* Speech Perception

Maja Bystrom, *Drexel University, Philadelphia, Pennsylvania,* Image Processing

Henning Bülow, *Optical Systems, Stuttgart, Germany,* Polarization Mode Dispersion Mitigation

A. R. Calderbank, *AT&T Shannon Laboratory, Florham Park, New Jersey,* Space–Time Codes for Wireless Communications

Gilberto M. Camilo, *OmniGuide Communications, Cambridge, Massachusetts,* Characterization of Optical Fibers

G. Cariolaro, *Università di Padova, Padova, Italy,* Pulse Position Modulation

Jeffrey B. Carruthers, *Boston University, Boston, Massachusetts,* Wireless Infrared Communications

John H. Carson, *George Washington University, Washington, District of Columbia,* Local Area Networks

Anne Cerboni, *France Télécom R&D, Issy Moulineaux, France,* IMT-2000 3G Mobile Systems

Kavitha Chandra, *Center for Advanced Computation and Telecommunications, University of Massachusetts Lowell, Lowell, Massachusetts,* Statistical Multiplexing

Sekchin Chang, *University of Texas at Austin, Austin, Texas,* Compensation of Nonlinear Distortion in RF Power Amplifiers

Matthew Chapman Caesar, *University of California at Berkeley, Berkeley, California,* IP Telephony

Jean-Pierre Charles, *France Télécom R&D, Issy Moulineaux, France,* IMT-2000 3G Mobile Systems

Chi-Chung Chen, *University of California, Los Angeles, California,* Chaos in Communications

Po-Ning Chen, *National Chi Tung University, Taiwan,* Sequential Decoding of Convolutional Codes

Thomas M. Chen, *Southern Methodist University, Dallas, Texas,* ATM Switching

Zhizhang (David) Chen, *Dalhousie University, Halifax, Nova Scotia, Canada,* Millimeter-Wave Antennas

Andrew R. Chraplyvy, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Nonlinear Effects in Optical Fibers

Christos G. Christodoulou, *University of New Mexico, Albuquerque, New Mexico,* Antennas for Mobile Communications

Michael T. Chryssomallis, *Democritus University of Thrace, Xanthi, Greece,* Antennas for Mobile Communications

Keith M. Chugg, *University of Southern California, Los Angeles, California,* Iterative Detection Algorithms in Communications

Habong Chung, *Hongik University, Seoul, Korea,* Gold Sequences

Leonard J. Cimini Jr., *AT&T Labs-Research, Middletown, New Jersey,* Orthogonal Frequency-Division Multiplexing

J. Cioffi, *Stanford University, Stanford, California,* Very High-Speed Digital Subscriber Lines (VDSLs)

Wim M. J. Coene, *Philips Research Laboratories, Eindhoven, The Netherlands,* Constrained Coding Techniques for Data Storage

Robert A. Cohen, *Troy, New York,* Streaming Video

Giovanni Emanuele Corazza, *University of Bologna, Bologna, Italy,* cdma2000

Steven Cummer, *Duke University, Durham, North Carolina,* Extremely Low Frequency (ELF) Electromagnetic Wave Propagation

Milorad Cvijetic, *NEC America, Herndon, Virginia,* Optical Transport System Engineering

Nelson L. S. da Fonseca, *Institute of Computing, State University of Campinas Brazil,* Bandwidth Reduction Techniques for Video Services; Network Traffic Modeling

Dirk Dahlhaus, *Communication Technology Laboratory, Zurich, Switzerland,* Chirp Modulation

Roger Dalke, *Institute for Telecommunication Sciences, Boulder, Colorado,* Local Multipoint Distribution Services (LMDS)

Marc Danzeisen, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Pankaj K. Das, *University of California, San Diego, La Jolla, California,* Surface Acoustic Wave Filters

Héctor J. De Los Santos, *Coventor, Inc., Irvine, California,* MEMS for RF/Wireless Applications

Filip De Turck, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Piet Demeester, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Jing Deng, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

Michael Devetsikiotis, *North Carolina State University, Raleigh, North Carolina,* Network Traffic Modeling

Olufemi Dosunmu, *Boston University, Boston, Massachusetts,* High-Speed Photodetectors for Optical Communications

Alexandra Duel-Hallen, *North Carolina State University, Raleigh, North Carolina,* Fading Channels

Tolga M. Duman, *Arizona State University, Tempe, Arizona,* Interleavers for Serial and Parallel Concatenated (Turbo) Codes

K. L. Eddie Law, *University of Toronto, Toronto, Canada,* Optical Switches

Thomas F. Eibert, *T-Systems Nova GmbH, Technologiezentrum, Darmstadt, Germany,* Antenna Modeling Techniques

Evangelos S. Eleftheriou, *IBM Zurich Research Laboratory, Rueschlikon, Switzerland,* Signal Processing for Magnetic Recording Channels

Amro El-Jaroudi, *University of Pittsburgh, Pittsburgh, Pennsylvania,* Linear Predictive Coding

Matthew Emsley, *Boston University, Boston, Massachusetts,* High-Speed Photodetectors for Optical Communications

T. Erseghe, *Università di Padova, Padova, Italy,* Pulse Position Modulation

Sonia Fahmy, *Purdue University, West Lafayette, Indiana,* Network Traffic Management

David R. Famolari, *Telecordia Technologies, Morristown, New Jersey,* Wireless IP Telephony

Li Fan, *OMM, Inc., San Diego, California,* Optical Crossconnects

András Faragó, *University of Texas at Dallas, Richardson, Texas,* Medium Access Control (MAC) Protocols

Aiguo Fei, *University of California at Los Angeles, Los Angeles, California,* Multicast Algorithms

Robert J. Filkins, *University of California, San Diego, La Jolla, California,* Surface Acoustic Wave Filters

John P. Fonseka, *University of Texas at Dallas, Richardson, Texas,* Quadrature Amplitude Modulation

M. Fossorier, *University of Hawaii at Manoa, Honolulu, Hawaii,* Finite-Geometry Codes

Roger Freeman, *Independent Consultant, Scottsdale, Arizona,* Community Antenna Television (CATV) (Cable Television); Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH)

Fabrizio Frezza, *"La Sapienza" University of Rome, Roma, Italy,* Leaky-Wave Antennas

Thomas E. Fuja, *University of Notre Dame, Notre Dame, Indiana,* Automatic Repeat Request

Alessandro Galli, *"La Sapienza" University of Rome, Roma, Italy,* Leaky-Wave Antennas

Costas N. Georghiades, *Texas A&M University, College Station, Texas,* EM Algorithm in Telecommunications

Leonidas Georgiadis, *Aristotle University of Thessaloniki, Thessaloniki, Greece,* Carrier-Sense Multiple Access (CSMA) Protocols

Mario Gerla, *University of California at Los Angeles, Los Angeles, California,* Multicast Algorithms

Pierre Ghandour, *France Télécom R&D, South San Franciso, California,* IMT-2000 3G Mobile Systems

K. Ghorbani, *RMIT University, Melbourne, Australia,* Microstrip Patch Arrays

Dipak Ghosal, *University of California at Davis, Davis, California,* IP Telephony

Giovanni Giambene, *University of Siena, Siena, Italy,* Wireless Application Protocol (WAP)

Arthur A. Giordano, *AG Consulting Inc., LLC, Burlington, Massachusetts,* Statistical Characterization of Impulsive Noise

Stefano Giordano, *University of Pisa, Pisa, Italy,* Multimedia Networking

Alain Glavieux, *ENST Bretagne, Brest, France,* Turbo Codes

Savo G. Glisic, *University of Oulu, Oulu, Finland,* Cochannel Interference in Digital Cellular TDMA Networks

Dennis L. Goeckel, *University of Massachusetts, Amherst, Massachusetts,* Bit-Interleaved Coded Modulation

Virgilio E. Gonzalez-Lozano, *Department of Electrical and Computer Engineering, University of Texas at El Paso, El Paso, Texas,* Optical Fiber Local Area Networks

Vivek Goyal, *Digital Fountain Inc., Fremont, California,* Transform Coding

Larry J. Greenstein, *AT&T Labs-Research, Middletown, New Jersey,* Orthogonal Frequency-Division Multiplexing

Marcus Greferath, *San Diego State University, San Diego, California,* Golay Codes

Manuel Günter, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Jaap C. Haartsen, *Ericsson Technology Licensing AB, Emmen, The Netherlands,* Bluetooth Radio System

Zygmunt J. Haas, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

David Haccoun, *Ecole Polytechnique de Montréal, Montréal, Quebec, Canada,* High-Rate Punctured Convolutional Codes

Abdelfatteh Haidine, *Dresden University of Technology, Dresden, Germany,* Powerline Communications

M. Hajian, *Delft University of Technology, Delft, The Netherlands,* Microwave Waveguides

Mounir Hamdi, *Hong Kong University of Science and Technology, Hong Kong,* Multimedia Medium Access Control Protocols for WDM Optical Networks

Yunghsiang S. Han, *National Chi Yan University, Taiwan,* Sequential Decoding of Convolutional Codes

Marc Handlery, *Lund University, Lund, Sweden,* Tailbiting Convolutional Codes

Eberhard Hänsler, *Darmstadt University of Technology, Darmstadt, Germany,* Acoustic Echo Cancellation

Fred Harris, *San Diego State University, San Diego, California,* Sigma-Delta Converters in Communication Systems

Christian Hartmann, *Technische Universität München, Institute of Communication Networks, Munich, Germany,* General Packet Radio Service (GPRS); GSM Digital Cellular Communication System

Mark Hasegawa-Johnson, *University of Illinois at Urbana-Champaign, Urbana, Illinois,* Speech Coding: Fundamentals and Applications

Homayoun Hashemi, *Sharif University of Technology, Teheran, Iran,* Wireless Local Loop Standards and Systems

Dimitrios Hatzinakos, *University of Toronto, Toronto, Ontario, Canada,* Spatiotemporal Signal Processing in Wireless Communications

Michelle C. Hauer, *University of Southern California, Optical Communications Laboratory, Los Angeles, California,* Digital Optical Correlation for Fiberoptic Communication Systems

Simon Haykin, *McMaster University, Hamilton, Ontario, Canada,* Maximum-Likelihood Estimation

Da-ke He, *University of Waterloo, Waterloo, Ontario, Canada,* Huffman Coding

Juergen Heiles, *Siemens Information & Communication Networks, Munich, Germany,* DWDM Ring Networks

Tor Helleseth, *University of Bergen, Bergen, Norway,* Ternary Sequences

Thomas R. Henderson, *Boeing Phantom Works, Seattle, Washington,* Leo Satellite Networks

Naftali Herscovici, *Anteg, Inc., Framingham, Massachusetts,* Microstrip Antennas

Pin-Han Ho, *Queen's University at Kingston, Ontario, Canada,* Survivable Optical Internet

Henk D. L. Hollmann, *Philips Research Laboratories, Eindhoven, The Netherlands,* Constrained Coding Techniques for Data Storage

R. Hoppe, *Institut Fuer Hochfrequenztechnik, University of Stuttgart, Stuttgart, Germany,* Propagation Models for Indoor Communications

Jiongkuan Hou, *New Jersey Institute of Technology, University Heights, Newark, New Jersey,* Admission Control in Wireless Networks

Halid Hrasnica, *Dresden University of Technology, Dresden, Germany,* Powerline Communications

Laura L. Huckabee, *Time Domain Corporation, Huntsville, Alabama,* Ultrawideband Radio

Abbas Jamalipour, *University of Sydney, Sydney, Australia,* Satellites in IP Networks

Pertti Järvensivu, *VTT Electronics, Oulu, Finland,* Cellular Communications Channels

Bahram Javidi, *University of Connecticut, Storrs, Connecticut,* Secure Ultrafast Data Communication and Processing Interfaced with Optical Storage

Rolf Johannesson, *Lund University, Lund, Sweden,* Tailbiting Convolutional Codes

Thomas Johansson, *Lund University, Lund, Sweden,* Authentication Codes

Sarah J. Johnson, *University of Newcastle, Callaghan, Australia,* Low-Density Parity-Check Codes: Design and Decoding

Douglas L. Jones, *University of Illinois at Urbana — Champaign, Berkeley, California,* Shell Mapping

Biing-Hwang Juang, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Hidden Markov Models

Edward V. Jull, *University of British Columbia, Vancouver, British Columbia,, Canada,* Horn Antennas

Peter Jung, *Gerhard-Mercator-Universität Duisburg,, Duisburg, Germany,* Time Division Multiple Access (TDMA)

Apostolos K. Kakaes, *Cosmos Communications Consulting Corporation, Centreville, Virginia,* Communication System Traffic Engineering

Dimitris N. Kalofonos, *Northeastern University, Boston, Massachusetts,* Multicarrier CDMA

Güneş Karabulut, *University of Ottawa, School of Information Technology and Engineering, Ottawa, Ontario, Canada,* Waveform Coding

Khalid Karimullah, *Hughes Network Systems, Germantown, Maryland,* Geosynchronous Satellite Communications

Magnus Karlsson, *Chalmers University of Technology, Göthenburg, Sweden,* Optical Solitons

Tadao Kasami, *Hiroshima City University, Hiroshima, Japan,* Kasami Sequences

Timo Kaukoranta, *Turku Centre for Computer Science (TUCS), University of Turku, Turku, Finland,* Scalar and Vector Quantization

Mohsen Kavehrad, *Pennsylvania State University, University Park, Pennsylvania,* Diversity in Communications

Haruo Kawakami, *Antenna Giken Corp., Laboratory, Saitama City, Japan,* Television and FM Broadcasting Antennas

Jürgen Kehrbeck, *Head, Division of e-Commerce and Mobile Communications, LStelcom Lichtenau, Germany,* Cell Planning in Wireless Networks

Gerd Keiser, *PhotonicsComm Solutions, Inc., Newton Center, Massachusetts,* Optical Couplers; Optical Fiber Communications

John Kieffer, *University of Minnesota, Minneapolis, Minnesota,* Data Compression

Dennis Killinger, *University of South Florida, Tampa, Florida,* Optical Wireless Laser Communications: Free-Space Optics

Kyungjung Kim, *Syracuse University, Syracuse, New York,* Adaptive Antenna Arrays

Ryuji Kohno, *Hiroshima City University, Hiroshima, Japan,* Kasami Sequences

Israel Korn, *University of New South Wales, Sydney, Australia,* Quadrature Amplitude Modulation

Sastri Kota, *Loral Skynet, Palo Alto, California,* Trends in Broadband Communication Networks

Hamid Krim, *ECE Department, North Carolina State University Centennial Campus, Raleigh, North Carolina,* Wavelets: A Multiscale Analysis Tool

Frank R. Kschischang, *University of Toronto, Toronto, Canada,* Product Codes

Erozan M. Kurtas, *Seagate Technology, Pittsburgh, Pennsylvania,* Design and Analysis of Low-Density Parity-Check Codes for Applications to Perpendicular Recording Channels

Alexander V. Kuznetsov, *Seagate Technology, Pittsburgh, Pennsylvania,* Design and Analysis of Low-Density Parity-Check Codes for Applications to Perpendicular Recording Channels

Henry K. Kwok, *University of Illinois at Urbana — Champaign, Berkeley, California,* Shell Mapping

Hyuck Kwon, *Wichita State University, Wichita, Kansas,* Modems

Cedric F. Lam, *Opvista Inc., Irvine, California,* Modern Ethernet Technologies

Paolo Lampariello, *"La Sapienza" University of Rome, Roma, Italy,* Leaky-Wave Antennas

F. Landstorfer, *Institut Fuer Hochfrequenztechnik, University of Stuttgart, Stuttgart, Germany,* Propagation Models for Indoor Communications

Greg D. LeCheminant, *Agilent Technologies, Santa Rosa, California,* Test and Measurement of Optically Based High-Speed Digital Communications Systems and Components

Frederick K. H. Lee, *Queen's University, Kingston, Ontario, Canada,* Nonuniformly Spaced Tapped-Delay-Line Equalizers for Sparse Multipath Channels

Jhong Sam Lee, *J.S. Lee Associates, Inc., Rockville, Maryland,* CDMA/IS95

Lin-Nan Lee, *Hughes Network Systems, Germantown, Maryland,* Geosynchronous Satellite Communications

Ralf Lehnert, *Dresden University of Technology, Dresden, Germany,* Powerline Communications

Hanoch Lev-Ari, *Northeastern University, Boston, Massachusetts,* Digital Filters

Allen H. Levesque, *Marlborough, Massachusetts,* BCH Codes — Binary; BCH Codes — Nonbinary and Reed-Solomon

Shipeng Li, *Microsoft Research Asia, Beijing, P.R. China,* Image and Video Coding

Weiping Li, *WebCast Technologies, Inc., Sunnyvale, California,* Image and Video Coding

Ben Liang, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

L. P. Ligthart, *Delft University of Technology, Delft, The Netherlands,* Microwave Waveguides

Jae S. Lim, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* High-Definition Television

Dave Lindbergh, *Polycom, Inc., Andover, Massachusetts,* H.324: Videotelephony and Multimedia for Circuit-Switched and Wireless Networks

K. J. Ray Liu, *University of Maryland, College Park, Maryland,* Multimedia Over Digital Subscriber Lines

Stephen S. Liu, *Verizon Laboratories, Waltham, Massachusetts,* ATM Switching

Alfio Lombardo, *University of Catania, Catania, Italy,* Multimedia Networking

Steven H. Low, *California Institute of Technology, Pasadena, California,* Network Flow Control

M. Luise, *University of Pisa, Dipartimento Ingegneria Informazione, Pisa, Italy,* Synchronization in Digital Communication Systems

Steven S. Lumetta, *University of Illinois Urbana — Champaign, Urbana, Illinois,* Network Reliability and Fault Tolerance

Wei Luo, *Lucent Technologies Bells Labs, Holmdel, New Jersey,* Wireless Packet Data

Maode Ma, *Nanyang Technological University, Singapore,* Multimedia Medium Access Control Protocols for WDM Optical Networks

Rangaraj Madabhushi, *Agere Systems, Optical Core Networks Dividion, Breinigsville, Pennsylvania,* Optical Modulators — Lithium Niobate

Aarne Mämmelä, *VTT Electronics, Oulu, Finland,* Cellular Communications Channels

Elias S. Manolakos, *Northeastern University, Boston, Massachusetts,* Neural Networks and Applications to Communications

Jon W. Mark, *University of Waterloo, Waterloo, Ontario, Canada,* Wideband CDMA in Third-Generation Cellular Communication Systems

Donald P. Massa, *Massa Products Corporation, Hingham, Massachusetts,* Acoustic Transducers

James L. Massey, *Consultare Technology Group, Bethesda, Denmark,* Threshld Decoding

Osamu Matoba, *University of Tokyo, Tokyo, Japan,* Secure Ultrafast Data Communication and Processing Interfaced with Optical Storage

John E. McGeehan, *University of Southern California, Optical Communications Laboratory, Los Angeles, California,* Digital Optical Correlation for Fiberoptic Communication Systems

Peter J. McLane, *Queen's University, Kingston, Ontario, Canada,* Nonuniformly Spaced Tapped-Delay-Line Equalizers for Sparse Multipath Channels

Steven W. McLaughlin, *Georgia Institute of Technology, Atlanta, Georgia,* Turbo Product Codes for Optical CDMA Systems

Donald G. McMullin, *Broadcom Corporation, Irvine, California,* Cable Modems

Muriel Médard, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* Network Reliability and Fault Tolerance

Seapahn Megerian, *University of California at Los Angeles, West Hills, California,* Wireless Sensor Networks

U. Mengali, *University of Pisa, Dipartimento Ingegneria Informazione, Pisa, Italy,* Synchronization in Digital Communication Systems

Lazaros Merakos, *University of Athens, Panepistimiopolis, Athens Greece,* Wireless ATM

Jan De Merlier, *Ghent University — IMEC, Ghent, Belgium,* Optical Signal Regeneration

Alfred Mertins, *University of Wollongong, Wollongong, Australia,* Image Compression

John J. Metzner, *Pennsylvania State University, University Park, Pennsylvania,* Aloha Protocols

Alan R. Mickelson, *University of Colorado, Boulder, Colorado,* Active Antennas

Arnold M. Michelson, *Marlborough, Massachusetts,* BCH Codes — Binary; BCH Codes — Nonbinary and Reed-Solomon

Leonard E. Miller, *Wireless Communications Technologies Group, NIST, Gaithersburg, Maryland,* CDMA/IS95

Mario Minami, *University of São Paulo, São Paulo, Brazil,* Low-Bit-Rate Speech Coding

Joseph Mitola III, *Consulting Scientist, Tampa, Florida,* Software Radio

Urbashi Mitra, *Communication Sciences Institute, Los Angeles, California,* Adaptive Receivers for Spread-Spectrum Systems

Eytan Modiano, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* Wavelength-Division Multiplexing Optical Networks

Peter Monsen, *P.M. Associates, Stowe, Vermont,* Tropospheric Scatter Communication

G. Montorsi, *Politecnico di Torino, Torino (Turin), Italy,* Serially Concatenated Codes and Iterative Algorithms

Tim Moors, *University of New South Wales, Sydney, Australia,* Transport Protocols for Optical Networks

Pär Moqvist, *Chalmers University of Technology, Göteborg, Sweden,* Serially Concatenated Continuous-Phase Modulation with Iterative Decoding

M. Morelli, *University of Pisa, Dipartimento Ingegneria Informazione, Pisa, Italy,* Synchronization in Digital Communication Systems

Geert Morthier, *Ghent University — IMEC, Ghent, Belgium,* Optical Signal Regeneration

Hussein T. Mouftah, *Queen's University at Kingston, Ontario, Canada,* Survivable Optical Internet

Biswanath Mukherjee, *University of California, Davis, Davis, California,* Design and Analysis of a WDM Client/Server Network Architecture

Hannes Müsch, *GN ReSound Corporation, Redwood City, California,* Speech Perception

Rohit U. Nabar, *Stanford University, Stanford, California,* Mimo Communication Systems

Ayman F. Naguib, *Morphics Technology Inc., Campbell, California,* Space–Time Codes for Wireless Communications

Masao Nakagawa, *Keio University, Japan,* Communications for Intelligent Transportation Systems

Hisamatsu Nakano, *Hosei University, Koganei, Tokyo, Japan,* Helical and Spiral Antennas

A. L. Narasimha Reddy, *Texas A&M University, College Station, Texas,* Differentiated Services

Krishna R. Narayanan, *Texas A&M University, College Station, Texas,* Turbo Equalization

Tomoaki Ohtsuki, *Tokyo University of Science, Noda, Chiba, Japan,* Optical Synchronous CDMA Systems

Yasushi Ojiro, *Antenna Giken Corp., Laboratory, Saitama City, Japan,* Television and FM Broadcasting Antennas

Rolf Oppliger, *eSECURITY Technologies Rolf Oppliger, Bern, Switzerland,* Network Security

Alessandro Orfei, *CNR, Istituto di Radioastronomia, Bologna, Italy,* Parabolic Antennas

Douglas O'Shaughnessy, *INRS-Telecommunications, Montreal, Quebec, Canada,* Speech Processing

Tony Ottosson, *Chalmers University of Technology, Goteborg, Sweden,* Signature Sequences for CDMA Communications

Sebnem Ozer, *MeshNetworks, Inc., Orlando, Florida,* Admission Control in Wireless Networks

Ryan A. Pacheco, *University of Toronto, Toronto, Ontario, Canada,* Spatiotemporal Signal Processing in Wireless Communications

K. Pahlavan, *Center for Wireless Information Network Studies Worcester Polytechnic Institute, Worcester, Massachusetts,* Trends in Wireless Indoor Networks

Algirdas Pakštas, *London Metropolitan University, London, England,* Intranets and Extranets

Constantinos B. Papadias, *Global Wireless Systems Research, Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Multiple Antenna Transceivers for Wireless Communications: A Capacity Perspective

Panagiotis Papadimitratos, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

Symeon Papavassiliou, *New Jersey Institute of Technology, University Heights, Newark, New Jersey,* Admission Control in Wired Networks; Admission Control in Wireless Networks

Peter Papazian, *Institute for Telecommunication Sciences, Boulder, Colorado,* Local Multipoint Distribution Services (LMDS)

Matthew G. Parker, *University of Bergen, Bergen, Norway,* Golay Complementary Sequences; Peak-to-Average Power Ratio of Orthogonal Frequency-Division Multiplexing

So Ryoung Park, *Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea,* Polyphase Sequences

Steen A. Parl, *Signatron Technology Corporation, Concord, Massachusetts,* HF Communications

Gianni Pasolini, *University of Bologna, DEIS, Italy,* Communications for Intelligent Transportation Systems

Nikos Passas, *University of Athens, Panepistimiopolis, Athens Greece,* Wireless ATM

Kenneth G. Paterson, *University of London, Egham, Surrey,* Golay Complementary Sequences

Arogyaswami J. Paulraj, *Stanford University, Stanford, California,* Mimo Communication Systems

Fotini-Niovi Pavlidou, *Aristotle University of Thessaloniki, Thessaloniki, Greece,* Frequency-Division Multiple Access (FDMA): Overview and Performance Evaluation

Menelaos K. Perdikeas, *National Technical University of Athens, Athens, Greece,* Distributed Intelligent Networks

Lance C. Perez, *University of Nebraska, Lincoln, Omaha,* Soft Output Decoding Algorithms

Athina P. Petropulu, *Drexel University, Philadelphia, Pennsylvania,* Interference Modeling in Wireless Communications

Stephan Pfletschinger, *Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany,* DMT Modulation

Raymond L. Pickholtz, *George Washington University, Washington, District of Columbia,* Code-Division Multiple Access

Pekka Pirinen, *University of Oulu, Oulu, Finland,* Cochannel Interference in Digital Cellular TDMA Networks

Leon Poladian, *University of Sydney, Eveleigh, Australia,* Optical Filters

Anastasis C. Polycarpou, *Arizona State University, Tempe, Arizona,* Antennas

Dimitrie C. Popescu, *Rutgers WINLAB, Piscataway, New Jersey,* Interference Avoidance for Wireless Systems

Miodrag Potkonjak, *University of California at Los Angeles, West Hills, California,* Wireless Sensor Networks

Edward J. Powers, *University of Texas at Austin, Austin, Texas,* Compensation of Nonlinear Distortion in RF Power Amplifiers

John G. Proakis, *Northeastern University, Boston, Massachusetts,* Amplitude Modulation; Companders; Intersymbol Interference in Digital Communication Systems; Matched Filters in Signal Demodulation; Power Spectra of Digitally Modulated Signals; Sampling of Analog Signals; Shallow-Water Acoustic Networks; Spread Spectrum Signals for Digital Communications

Chunming Qiao, *SUNY at Buffalo, Buffalo, New York,* Optical Switching Techniques in WDM Optical Networks

Hayder Radha, *Troy, New York,* Streaming Video

Harold Raemer, *Northeastern University, Boston, Massachusetts,* Atmospheric Radiowave Propagation

Daniel Ralph, *BTexact Technologies, Ipswich, Suffolk, United Kingdom,* Services Via Mobility Portals

Miguel Arjona Ramírez, *University of São Paulo, São Paulo, Brazil,* Low-Bit-Rate Speech Coding

Carey Rappaport, *Northeastern University, Boston, Massachusetts,* Reflector Antennas

Theodore S. Rappaport, *The University of Texas at Austin, Austin, Texas,* Mobile Radio Communications

Lars K. Rasmussen, *University of South Australia, Mawson Lakes, Australia,* Iterative Detection Methods for Multiuser Direct-Sequence CDMA Systems

Joseph A. Rice, *Northeastern University, Boston, Massachusetts,* Shallow-Water Acoustic Networks

Matti Rintamäki, *Helsinki University of Technology, Helsinki, Finland,* Power Control in CDMA Cellular Communication Systems

Apostolos Rizos, *AWARE, Inc., Bedford, Massachusetts,* Partial-Response Signals for Communications

Patrick Robertson, *Institute for Communications Technology, German Aerospace Center (DLR), Wessling, Germany,* Turbo Trellis-Coded Modulation (TTCM) Employing Parity Bit Puncturing and Parallel Concatenation

Ulrich L. Rohde, *Synergy Microwave Corporation, Paterson, New Jersey,* Frequency Synthesizers

Kai Rohrbacher, *Head, Department of Mobile Communication Software, LStelcom Lichtenau, Germany,* Cell Planning in Wireless Networks

Christopher Rose, *Rutgers WINLAB, Piscataway, New Jersey,* Interference Avoidance for Wireless Systems; Paging and Registration in Mobile Networks

George N. Rouskas, *North Carolina State University, Raleigh, North Carolina,* Routing and Wavelength Assignment in Optical WDM Networks

Michael Ruane, *Boston University, Boston, Massachusetts,* Optical Memories

William E. Ryan, *University of Arizona, Tucson, Arizona,* Concatenated Convolutional Codes and Iterative Decoding

Ashutosh Sabharwal, *Rice University, Houston, Texas,* Multiuser Wireless Communication Systems

John N. Sahalos, *Radiocommunications Laboratory, Aristotle University of Thessaloniki, Thessaloniki, Greece,* Antenna Arrays

S. Sajama, *Cornell University, Ithaca, New York,* Wireless Ad Hoc Networks

Magdalena Salazar Palma, *Universidad Politecnica de Madrid, Madrid, Spain,* Adaptive Antenna Arrays

Masoud Salehi, *Northeastern University, Boston, Massachusetts,* Frequency and Phase Modulation

Burton R. Saltzberg, *Middletown, New Jersey,* Carrierless Amplitude–Phase Modulation

Sheldon S. Sandler, *Lexington, Massachusetts,* Linear Antennas

Hikmet Sari, *Juniper Networks, Paris, France,* Broadband Wireless Access

Tapan K. Sarkar, *Syracuse University, Syracuse, New York,* Adaptive Antenna Arrays

Iwao Sasase, *Keio University, Yokohama, Japan,* Optical Synchronous CDMA Systems

Ali H. Sayed, *University of California, Los Angeles, California,* Wireless Location

Christian Schlegel, *University of Alberta, Edmonton, Alberta, Canada,* Trellis Coding

Jeffrey B. Schodorf, *MIT Lincoln Laboratory, Lexington, Massachusetts,* Land-Mobile Satellite Communications

Thomas A. Schonhoff, *Titan Systems Corporation, Shrewsbury, Massachusetts,* Continuous Phase Frequency Shift Keying (CPFSK); Statistical Characterization of Impulsive Noise

Henning Schulzrinne, *Columbia University, New York, New York,* Session Initiation Protocol (SIP)

Romed Schur, *Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany,* DMT Modulation

Kenneth Scussel, *Benthos, Inc., North Falmouth, Massachusetts,* Acoustic Modems for Underwater Communication

Randall G. Seed, *MIT Lincoln Laboratory, Lexington, Massachusetts,* Multibeam Phased Arrays

M. Selim Ünlü, *Boston University, Boston, Massachusetts,* High-Speed Photodetectors for Optical Communications

Husrev Sencar, *New Jersey Institute of Technology, Newark, New Jersey,* Orthogonal Transmultiplexers: A Time-Frequency Perspective

Mehdi Shadaram, *Department of Electrical and Computer Engineering, University of Texas at El Paso, El Paso, Texas,* Optical Fiber Local Area Networks

Ippei Shake, *NTT Network Innovation Laboratories, Kanagawa, Japan,* Signal Quality Monitoring in Optical Networks

K. Sam Shanmugan, *University of Kansas, Lawrence, Kansas,* Simulation of Communication Systems

John M. Shea, *University of Florida, Gainesville, Florida,* Multidimensional Codes

Chris Shephard, *BTexact Technologies, Ipswich, Suffolk, United Kingdom,* Services Via Mobility Portals

Barry L. Shoop, *United States Military Academy, West Point, New York,* Photonic Analog-to-Digital Converters

Mark Shtaif, *Tel-Aviv University, Tel-Aviv, Israel,* Modeling and Analysis of Digital Optical Communications Systems

Marvin K. Simon, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California,* Minimum-Shift-Keying

Kazimierz (Kai) Siwiak, *Time Domain Corporation, Huntsville, Alabama,* Loop Antennas; Ultrawideband Radio

David R. Smith, *George Washington University, Ashburn, Virgina,* Terrestrial Microwave Communications

Josep Sole i Tresserres, *France Télécom R&D, South San Franciso, California,* IMT-2000 3G Mobile Systems

Hong-Yeop Song, *Yonsei University, Seoul, South Korea,* Feedback Shift Register Sequences

Iickho Song, *Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea,* Polyphase Sequences

Ethem M. Sozer, *Northeastern University, Boston, Massachusetts,* Shallow-Water Acoustic Networks

Andreas Spanias, *Arizona State University, Tempe, Arizona,* Vocoders

Predrag Spasojević, *Rutgers, The State University of New Jersey, Piscataway, New Jersey,* EM Algorithm in Telecommunications

Joachim Speidel, *Institute of Telecommunications, University of Stuttgart, Stuttgart, Germany,* DMT Modulation

Per Ståhl, *Lund University, Lund, Sweden,* Tailbiting Convolutional Codes

Alexandros Stavdas, *National Technical University of Athens, Athens, Greece,* Optical Multiplexing and Demultiplexing

Marc-Alain Steinemann, *University of Bern, Bern, Switzerland,* Virtual Private Networks

Dimitrios Stiliadis, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Packet-Switched Networks

Milica Stojanovic, *Massachusetts Institute of Technology, Cambridge, Massachusetts,* Acoustic (Underwater) Communications; Shallow-Water Acoustic Networks

Detlef Stoll, *Siemens ICN, Optisphere Networks, Boca Raton, Florida,* DWDM Ring Networks

Erik Strom, *Chalmers University of Technology, Goteborg, Sweden,* Signature Sequences for CDMA Communications

Carl-Erik W. Sundberg, *iBiquity Digital Corp., Warren, New Jersey,* Continuous-Phase-Coded Modulation

Arne Svensson, *Chalmers University of Technology, Goteborg, Sweden,* Signature Sequences for CDMA Communications

Violet R. Syrotiuk, *University of Texas at Dallas, Richardson, Texas,* Medium Access Control (MAC) Protocols

Chintha Tellambura, *University of Alberta, Edmonton, Alberta,* Golay Complementary Sequences; Peak-to-Average Power Ratio of Orthogonal Frequency-Division Multiplexing; Wireless Communications System Design

Hemant K. Thapar, *LSI Logic Corporation, San Jose, California,* Magnetic Storage Systems

Ioannis Tomkos, *Athens Information Technology, Peania, Greece,* WDM Metropolitan-Area Optical Networks

Lang Tong, *Cornell University, Ithaca, New York,* Channel Modeling and Estimation

Brent Townshend, *Townshend Computer Tools, Menlo Park, California,* V.90 MODEM

William H. Tranter, *Virginia Tech, Blacksburg, Virginia,* Mobile Radio Communications

H. J. Trussell, *North Carolina State University, Raleigh, North Carolina,* Image Sampling and Reconstruction

Jitendra K. Tugnait, *Auburn University, Auburn, Alabama,* Blind Equalization Techniques

B. E. Usevitch, *University of Texas at El Paso, El Paso, Texas,* JPEG2000 Image Coding Standard

Steven Van Den Berghe, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Pim Van Heuven, *Ghent University, Ghent, Belgium,* Multiprotocol Label Switching (MPLS)

Richard van Nee, *Woodside Networks, Breukelen, The Netherlands,* Wireless LAN Standards

Alessandro Vanelli-Coralli, *University of Bologna, Bologna, Italy,* cdma2000

Emmanouel Varvarigos, *University of Patras, Patras, Greece,* Computer Communications Protocols

Theodora Varvarigou, *National Technical University, Patras, Greece,* Computer Communications Protocols

Bane Vasic, *University of Arizona, Tucson, Arizona,* Design and Analysis of Low-Density Parity-Check Codes for Applications to Perpendicular Recording Channels

Iakovos S. Venieris, *National Technical University of Athens, Athens, Greece,* Distributed Intelligent Networks

Roberto Verdone, *University of Bologna, DEIS, Italy,* Communications for Intelligent Transportation Systems

A. J. Viterbi, *Viterbi Group, San Diego, California,* Viterbi Algorithm

Emanuele Viterbo, *Politechnico di Torino, Torino* (*Turin*)*, Italy,* Permutation Codes

Branimir R. Vojčić, *George Washington University, Washington, District of Columbia,* Code-Division Multiple Access

John L. Volakis, *University of Michigan, Ann Arbor, Michigan,* Antenna Modeling Techniques

John C. H. Wang, *Federal Communications Commission, Washington, District of Columbia,* Radio Propagation AT LF, MF, and HF

Xiaodong Wang, *Columbia University, New York, New York,* Blind Multiuser Detection

R. B. Waterhouse, *RMIT University, Melbourne, Australia,* Microstrip Patch Arrays

Steven R. Weller, *University of Newcastle, Callaghan, Australia,* Low-Density Parity-Check Codes: Design and Decoding

Wushao Wen, *University of California, Davis, Davis, California,* Design and Analysis of a WDM Client/Server Network Architecture

Lih-Jyh Weng, *Maxtor Corporation, Shrewsbury, Massachusetts,* Coding for Magnetic Recording Channels

Richard D. Wesel, *University of California at Los Angeles, Los Angeles, California,* Convolutional Codes

Krzysztof Wesołowski, *Poznań University of Technology, Poznań, Poland,* Adaptive Equalizers

Stephen B. Wicker, *Cornell University, Ithaca, New York,* Cyclic Codes

Werner Wiesbeck, *Director, Institute for High Frequency Technology and Electronics Karlsruhe University, Germany,* Cell Planning in Wireless Networks

Alan E. Willner, *University of Southern California, Optical Communications Laboratory, Los Angeles, California,* Digital Optical Correlation for Fiberoptic Communication Systems

Stephen G. Wilson, *University of Virginia, Charlottesvaille, Virginia,* Trellis-Coded Modulation

Bernhard Wimmer, *Siemens AG, Munich, Germany,* H.324: Videotelephony and Multimedia for Circuit-Switched and Wireless Networks

Peter J. Winzer, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Optical Transmitters, Receivers, and Noise

G. Woelfle, *Institut Fuer Hochfrequenztechnik, University of Stuttgart, Stuttgart, Germany,* Propagation Models for Indoor Communications

Tan F. Wong, *University of Florida, Gainesville, Florida,* Multidimensional Codes

Thomas Wörz, *Audens ACT Consulting GmbH, Wessling, Germany,* Turbo Trellis-Coded Modulation (TTCM) Employing Parity Bit Puncturing and Parallel Concatenation

William W. Wu, *Consultare Technology Group, Bethesda, Denmark,* Threshld Decoding

Yiyan Wu, *Communications Research Centre Canada, Ottawa, Ontario, Canada,* Terrestrial Digital Television

Jimin Xie, *Siemens ICN, Optisphere Networks, Boca Raton, Florida,* DWDM Ring Networks

Fuqin Xiong, *Cleveland State University, Cleveland, Ohio,* Digital Phase Modulation and Demodulation

En-hui Yang, *University of Waterloo, Waterloo, Ontario, Canada,* Huffman Coding

Jie Yang, *New Jersey Institute of Technology, University Heights, Newark, New Jersey,* Admission Control in Wired Networks

Xueshi Yang, *Drexel University, Philadelphia, Pennsylvania,* Interference Modeling in Wireless Communications

Kung Yao, *University of California, Los Angeles, California,* Chaos in Communications

Bülent Yener, *Rensselaer Polytechnic University, Troy, New York,* Internet Security

Ikjun Yeom, *Korea Advanced Institute of Science and Technology, Seoul, South Korea,* Differentiated Services

Abbas Yongaçoğlu, *University of Ottawa, School of Information Technology and Engineering, Ottawa, Ontario, Canada,* Waveform Coding

Myungsik Yoo, *Soongsil University, Seoul, Korea,* Optical Switching Techniques in WDM Optical Networks

Nabil R. Yousef, *Adaptive Systems Laboratory, Department of Electrical Engineering, University of California, Los Angeles, California,* Wireless Location

Jens Zander, *Royal Institute of Technology, Stockholm, Sweden,* Radio Resouce Management in Future Wireless Networks

Yimin Zhang, *Villanova University, Villanova, Pennsylvania,* Interference Suppression in Spread-Spectrum Communication Systems

Haitao Zheng, *Bell Laboratories, Lucent Technologies, Holmdel, New Jersey,* Multimedia Over Digital Subscriber Lines

Shihua Zhu, *Xian Jiaotong University, Xian, Shaanxi, People's Republic of China,* Wideband CDMA in Third-Generation Cellular Communication Systems

Rodger E. Ziemer, *University of Colorado, Colorado Springs, Colorado,* Mobile Radio Communications

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 1

A

ACOUSTIC ECHO CANCELLATION

EBERHARD HÄNSLER
Darmstadt University of Technology
Darmstadt, Germany

1. INTRODUCTION

In 1877 the front page of *Scientific American* showed a picture of a man using "the new Bell telephone" [1]. He held a microphone in front of his mouth and an identical-looking device — the loudspeaker — close to one of his ears. So, at the beginning of telecommunications both hands were busy while making a telephone call. This troublesome way of operation was due to the lack of efficient electroacoustic converters and amplifiers. The inconvenience, however, guaranteed optimal conditions: a high signal-to-(environmental) noise ratio at the microphone input, a perfect coupling between loudspeaker and the ear of the listener, and — last but not least — a high attenuation between the loudspeaker and microphone. The designers of modern speech communication systems still dream of getting back those conditions.

It did not take long until the microphone and the loudspeaker of a telephone were mounted in a handset. Thus, one hand had been freed. To provide a fully natural communication between two users of a speech communication system — to allow both to speak at the same time and to interrupt each other, with both hands free — is still a problem that keeps hundreds of researchers and industrial developers busy.

This article is meant to explain the problem of acoustical echoes and their cancellation. It will focus on the hands-free telephone as one of the applications mostly asked for. The statements, however, hold for other applications such as hearing aids, voice input systems, and public-address systems as well.

The problem of acoustic echo cancellation arises wherever a loudspeaker and a microphone are placed such that the microphone picks up the signal radiated by the loudspeaker and its reflections at the borders of the enclosure. As a result, the electroacoustic circuit may become unstable and produce howling. In addition, the users of telecommunication systems are annoyed by listening to their own speech delayed by the round-trip time of the system. To avoid these problems, the attenuation of the acoustic path between loudspeaker and microphone has to be sufficiently high.

In general, acoustic echo cancellation units as used in hands-free communication systems consist of three subunits: (1) a loss control circuit (LC), (2) a filter parallel to the loudspeaker–enclosure–microphone system (LEMS) — the echo cancellation filter (ECF), and (3) a second filter — the residual echo-suppressing filter (RESF) — within the path of the output signal (see Fig. 1).

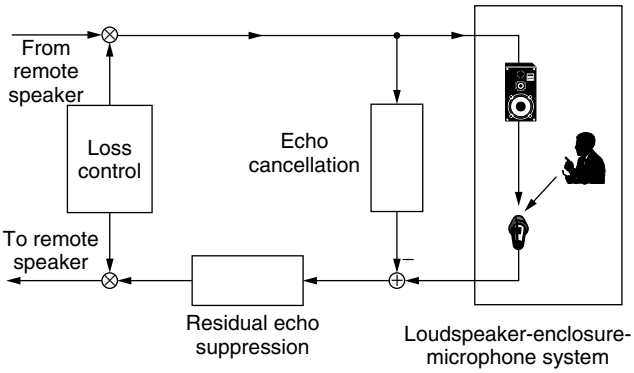


Figure 1. General structure of an acoustic echo cancellation system.

Their functions are obvious — the loss control circuit can attenuate the input and/or output signal such that the communication loop always remains stable. The ECF in parallel to the LEMS is able to cancel echoes picked up by the microphone according to the degree to which this filter is matched to the LEMS. The RESF within the path of the output signal can be used to suppress residual echoes and background noise. In the early days of acoustic echo control a so-called center clipper (see Fig. 2) took the place of this filter.

Of these subunits, the loss control circuit has the longest history in hands-free communication systems. In its simplest form it reduces the usually full-duplex communication system to a half-duplex one by alternatively switching the input and output lines on and off. Apart from preventing howling and suppressing echoes, any natural conversation was prevented, too. Only the ECF in parallel to the LEMS can help to provide full-duplex (i.e., fully natural) communication.

A device for hands-free telephone conversation using voice switching was presented in 1957 [2]. The introduction of a center clipper in 1974 [3] resulted in a noticeable improvement. Laboratory experiments applying an adaptive filter for acoustic echo cancellation were reported in 1975 [4]. At that time, however, an economically feasible implementation of such a filter for acoustic echo cancellation was far out of sight.

Because of the high interest in providing hands–free speech communication, an enormous number of papers has been published since the early 1980s. Among those are a number of bibliographies [5–8], overview papers [9–11], and books [12,13].

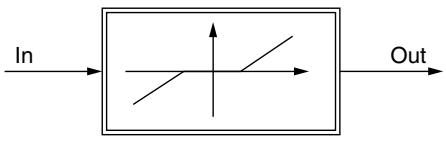


Figure 2. Center clipper.

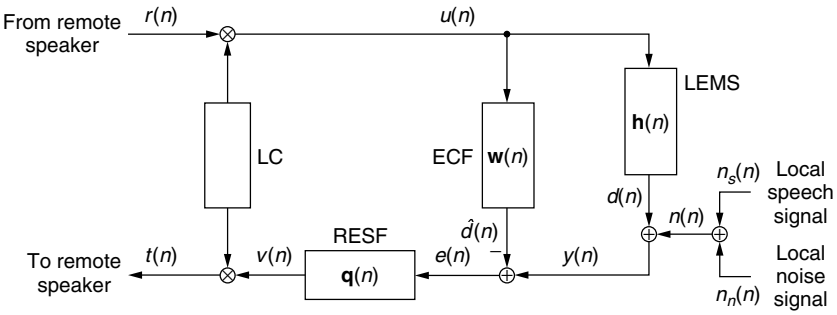


Figure 3. Notation used in this contribution: LC = loss control circuit, LEMS = loudspeaker–enclosure–microphone system, ECF = echo-canceling filter, RESF = residual echo suppressing filter.

For the following considerations, the notation as given in Fig. 3 will be used. Lowercase boldface letters will indicate column vectors; uppercase boldface letters will denote matrices.

2. SCENARIO

2.1. Systems

2.1.1. Loudspeaker–Enclosure–Microphone System (LEMS). In a LEMS the loudspeaker and the microphone are connected by an acoustical path formed by a direct connection (if both can "see" each other) and in general a large number of reflections at the boundaries of the enclosure. For low sound pressure and no overload of the converters, this system may be modeled with sufficient accuracy as a linear system. The impulse response can be described by a sequence of delta impulses delayed proportionally to the geometric length of the related path and the inverse of the sound velocity. The amplitudes of the impulses depend on the reflection coefficients of the boundaries and on the inverse of the pathlengths. As a first-order approximation one can assume that the impulse response decays exponentionally. A measure for the degree of this decay is the *reverberation time* T_{60}, which specifies the time necessary for the sound energy to drop by 60 dB after the sound source has been switched off [14]. Depending on the application, it may be possible to design the boundaries of the enclosure such that the reverberation time is small, resulting in a short impulse response. Examples are telecommunication studios. For ordinary offices, the reverberation time T_{60} is typically in the order of a few hundred milliseconds. For the interior of a passenger car, this quantity is a few tens of milliseconds long. Figure 4 shows the impulse responses of LEMSs measured in an office (left) and in a passenger car (right). The microphone signals have been sampled at 8 kHz according to the standards for telephone signals. It becomes obvious that the impulse response of an office exhibits amplitudes noticeably different from zero even after 1000 samples, that is to say, after 125 ms. In comparison, the impulse response of the interior of a car decays faster because of the smaller volume of this enclosure.

The impulse response of a LEMS is highly sensitive to any changes such as the movement of a person within it. This is explained by the fact that, assuming a sound velocity of 343 m/s and 8 kHz sampling frequency, the distance traveled between two sampling instants is 4.3 cm. Therefore, a 4.3-cm change in the length of an echo path, the move of a person by only a few centimeters, shifts the related impulse by one sampling interval. Thus, the echo cancellation filter (ECF) that has to mimic the LEMS must be an adaptive filter.

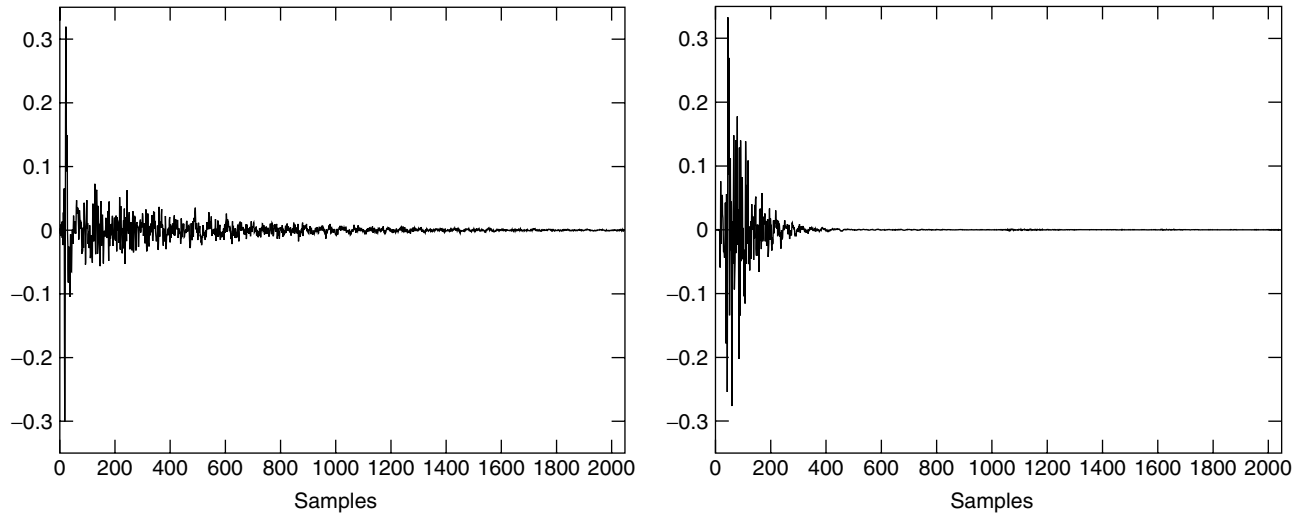


Figure 4. Impulse responses measured in an office (left) and in a car (right) (sampling frequency = 8 kHz).

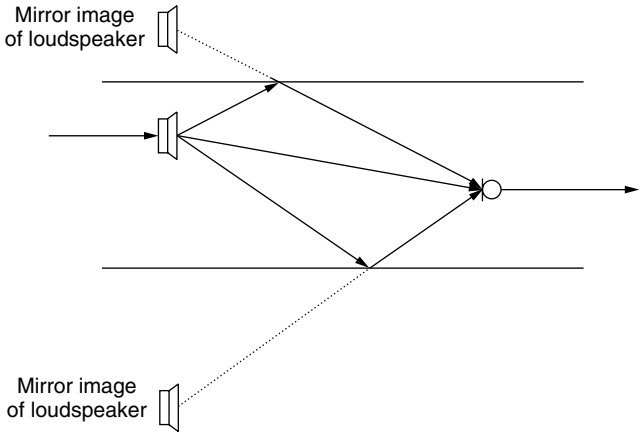


Figure 5. Simulation of an impulse response: direct path and two reflections.

2.1.2. Simulation of a LEMS. The impulse response of a LEMS can be simulated by using the principle of an *image source* [15] just as used in geometric optics. A reflected sound ray can be viewed as originating from a mirror image of the original source (see Fig. 5). At the reflection the sound pressure is attenuated according to a *reflection coefficient* that depends on the material at the reflection point. Typical reflection coefficients are in the order of 0.4 for "soft" material and close to but smaller than one for rigid walls [16]. Reflections occur at all boundaries of the enclosure, and sound rays may be reflected several times before reaching the microphone. In addition to (multiple) reflections, the sound pressure is also attenuated proportionally to the inverse of the length of its path.

2.1.3. Electronic Replica of LEMSs. From a control engineering point of view, acoustic echo cancellation is a system identification problem. However, the system to be identified — the LEMS — is highly complex; its impulse response exhibits up to several thousand sample values noticeably different from zero and it is time-varying at a speed mainly according to human movements. The question of the optimal structure of the ECF has been discussed intensively. Since a long impulse response has to be modeled by the ECF, a recursive (IIR) filter seems best suited at first glance. At second glance, however, the impulse response exhibits a highly detailed and irregular shape. To achieve a sufficiently good match, the replica must offer a large number of adjustable parameters. Therefore, an IIR filter does not have an advantage over a nonrecursive (FIR) filter [17,18]. The even more important argument in favor of an FIR filter is its guaranteed stability during adaptation.

Figure 6 shows an FIR filter of length N. The N values of the input signal $u(n)$ can be combined in a column vector $\mathbf{u}(n)$:

$$\mathbf{u}(n) = [u(n), u(n-1), u(n-2), \ldots, u(n-N+2), \times u(n-N+1)]^T \quad (1)$$

If we also combine the filter coefficients, the tap weights, in a column vector $\mathbf{w}(n)$

$$\mathbf{w}(n) = [w_0(n), w_1(n), w_2(n), \ldots, w_{N-2}(n), w_{N-1}(n)]^T \quad (2)$$

the output signal $\hat{d}(n)$ can be written as an inner product:

$$\hat{d}(n) = \sum_{k=0}^{N-1} w_k(n)\, u(n-k) = \mathbf{w}^T(n)\, \mathbf{u}(n) = \mathbf{u}^T(n)\, \mathbf{w}(n) \quad (3)$$

A measure to express the effect of an ECF is the *echo return loss enhancement* (*ERLE*):

$$ERLE = 10 \log \frac{E[d^2(n)]}{E[(d(n) - \hat{d}(n))^2]} \text{ dB} \quad (4)$$

where the echo $d(n)$ is equal to the microphone output signal $y(n)$ in case the loudspeaker is the only signal source within the LEMS [i.e., the local speech signal $n_s(n)$ and the local noise $n_n(n)$ are zero], and $\hat{d}(n)$ describes the ECF output. Denoting the (assumed to be time invariant for the moment) impulse responses of the LEMS by $h_i, i = 0, \ldots, \infty$, and the ECF by $\mathbf{w}$ respectively, it follows that

$$d(n) = \sum_{i=0}^{\infty} h_i u(n-i) \quad (5)$$

and

$$\hat{d}(n) = \sum_{i=0}^{N-1} w_i u(n-i) \quad (6)$$

where $N-1$ is the degree of the nonrecursive ECF. Assuming, also for simplicity, a stationary white input signal $u(n)$, the *ERLE* can be expressed as

$$ERLE = 10 \log \frac{E[u^2(n)] \sum_{i=0}^{\infty} h_i^2}{E[u^2(n)] \left(\sum_{i=0}^{\infty} h_i^2 - 2 \sum_{i=0}^{N-1} h_i w_i + \sum_{i=0}^{N-1} w_i^2 \right)} \text{ dB} \quad (7)$$

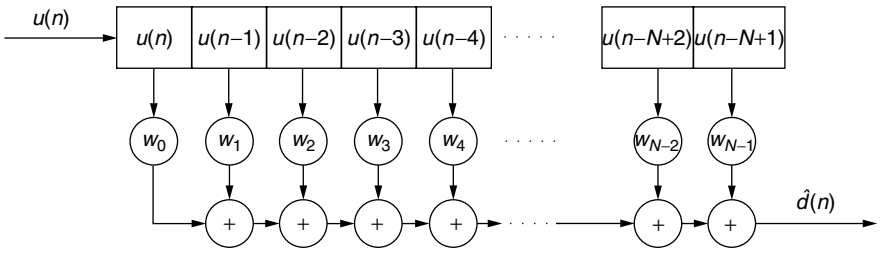


Figure 6. Transversal (FIR) filter as used to model the LEMS.

An upper bound for the effect of an ECF of degree $N-1$ can be calculated by assuming a perfect match of the first N coefficients of the ECF with the LEMS:

$$w_i = h_i \quad \text{for} \quad 0 \le i < N \tag{8}$$

In this case Eq. (7) reduces to

$$ERLE_{\max}(N) = 10 \log \frac{\sum_{i=0}^{\infty} h_i^2}{\sum_{n=N}^{\infty} h_i^2} \text{ dB} \tag{9}$$

Figure 7 shows the upper bounds of the *ERLE* achievable with transversal ECFs of length N for an office and a car with impulse responses as given in Fig. 4. An attenuation of only 25 dB needs filter lengths of about 1100 for the office and about 250 for the car.

2.2. Signals

2.2.1. Speech Signals. Acoustic echo cancellation requires the adaptation of FIR filters. In general, the performance of adaptation algorithms crucially depends on the properties of the signals involved. In the application considered here, one has to deal primarily with speech signals additively disturbed by other speech signals (in the case of doubletalk, i.e., if both communication partners talk simultaneously) and by noise. Performing signal processing with this type of signals turns out to be very difficult.

Speech is characterized by nearly periodic segments, by noiselike segments, and by pauses. The signal envelope fluctuates enormously. In speech processing it is widely accepted that parameters derived from a speech signal have to be updated after intervals of about 20 ms. Short-time variances may differ by more than 40 dB [19]. Sampling frequencies range from 8 kHz in telephone systems up to about 40 kHz in high-fidelity systems. Even in the case of 8 kHz sampling frequency, consecutive samples are highly correlated. The normalized autocorrelation coefficient $s_{uu}(1)/s_{uu}(0)$ of neighboring samples assumes values in the range of 0.8–0.95. Short-time autocorrelation matrices very often become singular. Thus, special precautions are necessary to prevent instability of algorithms that use — directly or indirectly — the inverse of the autocorrelation matrix. To summarize, speech signals are *nonpersistent*. Figure 8 shows a segment of a speech signal (left) and an estimate of a power spectral density of speech signals (right). The spectrum clearly indicates that the statistical properties of speech are quite different from those of a white process.

2.2.2. Noise. The noise signals involved in echo-canceling applications are typically those existing in offices or in moving cars. Especially the noise in a passenger car moving at constant speed exhibits a power density spectrum that is decaying slightly faster than the one of speech (Fig. 9). In both cases the major part of the energy is concentrated at low frequencies.

2.3. Side Constraints

Acoustic echo cancelers used in telephones have to comply with a number of standards issued by the international standardization organizations like the International Telecommunication Union (ITU) or the European Telecommunications Standards Institute (ETSI). Especially important are requirements concerning delay and echo attenuation [20–22]. For ordinary telephones the ITU allows only 2 ms additional delay for front-end processing. In case of mobile telephones 39 ms are permitted. The maximum of 2 ms prohibits the application of efficient frequency domain or block processing methods. Concerning the overall echo attenuation a minimum of 45 dB is necessary in singletalk situations. In case of doubletalk, the attenuation can be reduced to 30 dB taking into consideration that the echo of the far-end signal is masked by the local speech signal. This high echo attenuation has

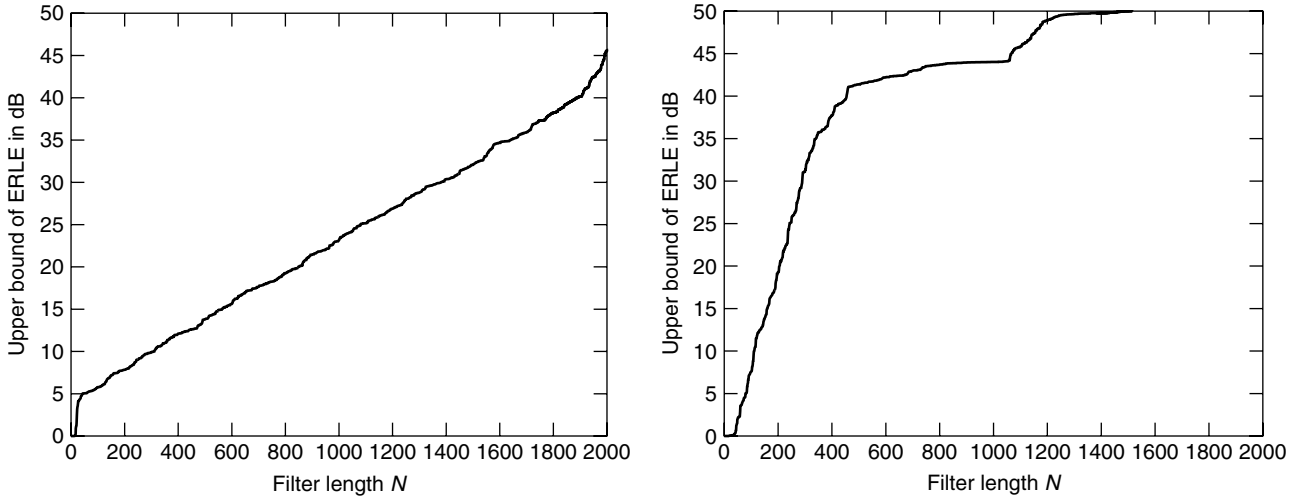


Figure 7. Maximal attenuations achievable with a transversal filter of length N in an office (left) and in a car (right) (sampling frequency = 8 kHz).

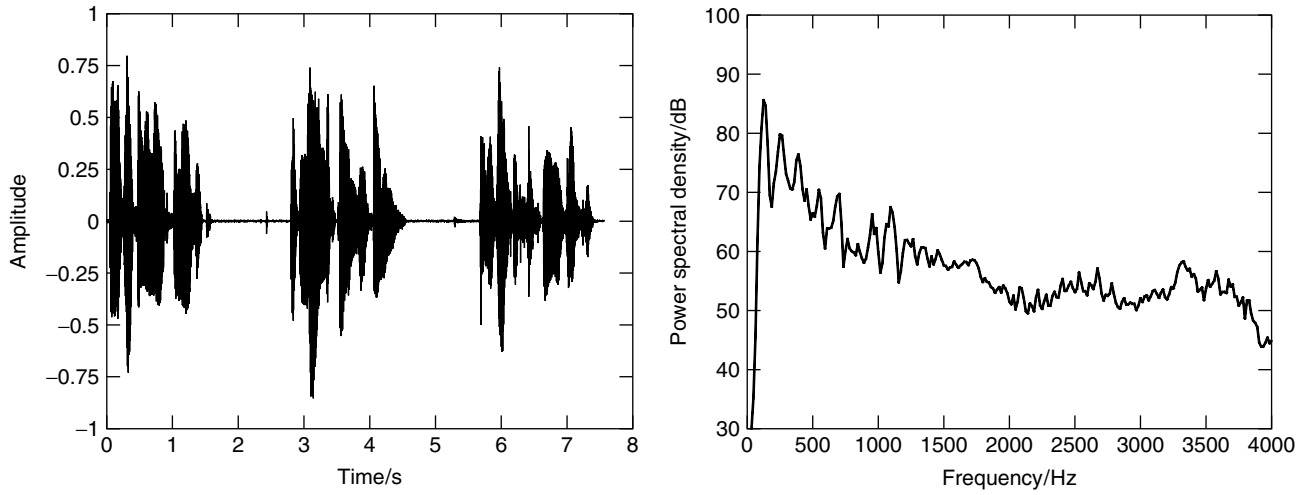


Figure 8. Section of a speech signal and estimate of the power spectral density of speech.

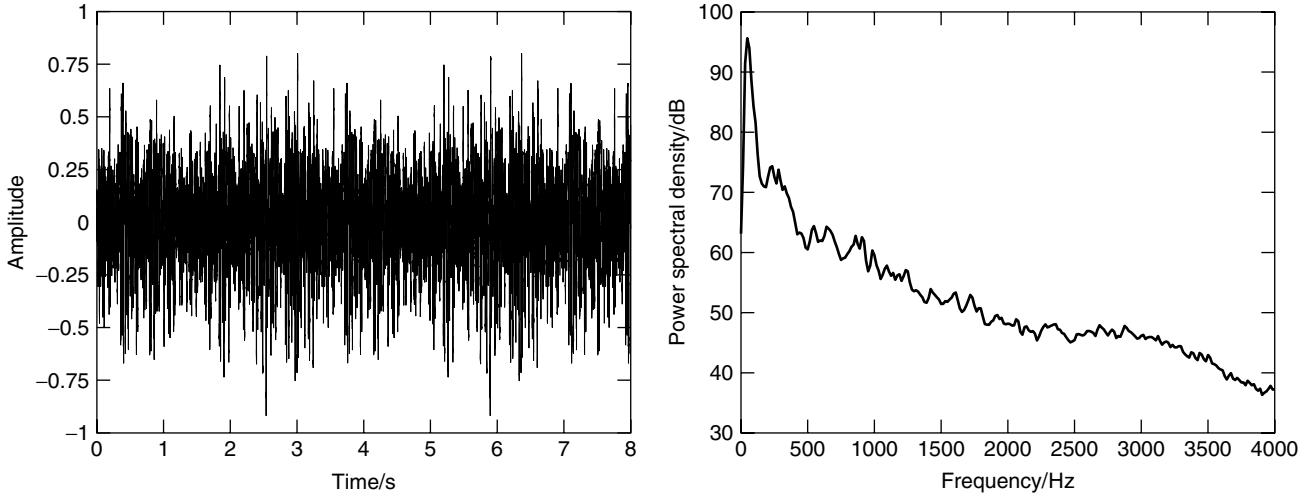


Figure 9. Section of a car noise signal and estimate of the power spectral density.

to be provided at all times of the connection and at all situations. Since adaptive filters (the ECF and the RESF; see Fig. 3) may not (yet) have converged to their optimal settings the attenuation required can be guaranteed only with the help of an additional loss control device.

3. METHODS TO STABILIZE THE ELECTROACOUSTIC LOOP

3.1. Traditional Methods

Most voice communication systems operate as a closed loop. In the case of the telephone system with customers using handsets, the attenuations between loudspeakers and microphones crucially contribute to the stabilization of this loop. Therefore, with the introduction of hands-free devices the provision of a stable electroacoustic loop became a major problem.

The simplest way to achieve stability is reducing the full-duplex connection to half-duplex. This can be done manually — as it is still done by astronauts — or by voice-controlled switches. The problem related to those switches is that they do not distinguish between speech and noise signals. Therefore, a strong noise can mislead the switching circuit. Consequently an active speaker may be switched off in favor of a noise at the remote location.

A moderate form of switching off one direction of a connection consists of artificially increasing its attenuation. This results in shifting an attenuation to the incoming or the outgoing circuit depending on where the loss control device detects the lower speech activity. In case no activity is detected on both circuits, the additional attenuation can be distributed equally between both directions. As mentioned before, even modern echo-canceling devices cannot fulfill the ITU requirements without loss insertion. However, the attenuation already provided by echo-canceling (and other) circuits can be estimated and only the lacking attenuation has to be inserted. In the case of a well-adapted ECF, this may only be a few decibels that do not disturb the speakers.

A method to stabilize an electroacoustic loop has been proposed [23]. It is especially designed for systems like public-address systems where the loudspeaker output signal feeds back into the talker microphone directly. It consists of a frequency shift of a few hertz, implemented by a single-sideband modulation — within the microphone–loudspeaker circuit. Thus, stationary howling cannot build up. Echoes are not canceled. They are, however, shifted to higher or lower frequencies — depending on whether the modulation frequency is positive or negative — until they "fall" into a minimum of the transfer function of the LEMS and become inaudible. In speech communication systems frequency shifts of ~3–5 Hz are scarcely audible. The stability gain achievable with this method depends on the signal and the acoustical properties of the enclosure. For speech signals and rooms with short reverberation times the gain is in the order of 3–5 dB; for rooms with long reverberation times it can go up to ~10 dB.

3.2. Adaptive Filters

With the availability of powerful digital signal processors, the application of an adaptive filter to cancel acoustic echoes, the ECF, and a second adaptive filter, the RESF, to suppress residual echoes not canceled by the ECF became feasible (see Fig. 1). As explained in the previous section, a transversal filter of high order is used for the ECF. The RESF, typically, is implemented in the frequency domain by an adaptive filter as well.

3.2.1. The Echo-Canceling Filter (ECF). For the following considerations we assume that the impulse responses of the ECF and of the LEMS both have the same length N. In reality, the impulse response of the LEMS may be much longer than that of the ECF. Nevertheless, this assumption means no restriction because the shorter impulse response can always be extended by zeros. Equivalent to Eqs. (2) and (3), one can write the impulse response of the LEMS at time n as a vector $\mathbf{h}(n)$

$$\mathbf{h}(n) = [h_0(n), h_1(n), h_2(n), \ldots, h_{N-2}(n), h_{N-1}(n)]^T \quad (10)$$

and the output signal $d(n)$ as an inner product:

$$d(n) = \sum_{k=0}^{N-1} h_k(n)\, u(n-k) = \mathbf{h}^T(n)\, \mathbf{u}(n) = \mathbf{u}^T(n)\, \mathbf{h}(n) \quad (11)$$

The mismatch between LEMS and ECF can be expressed by a *mismatch vector* $\boldsymbol{\varepsilon}(n)$:

$$\boldsymbol{\varepsilon}(n) = \mathbf{h}(n) - \mathbf{w}(n) \quad (12)$$

Later, the squared L_2-norm $\boldsymbol{\varepsilon}^T(n)\, \boldsymbol{\varepsilon}(n)$ of the system mismatch vector will be called the *system distance*:

$$\Delta(n) = \boldsymbol{\varepsilon}^T(n)\, \boldsymbol{\varepsilon}(n) = \|\boldsymbol{\varepsilon}(n)\|^2 \quad (13)$$

The quantity $e_u(n)$

$$e_u(n) = d(n) - \hat{d}(n) = \boldsymbol{\varepsilon}^T(n)\, \mathbf{u}(n) \quad (14)$$

represents the *undisturbed error*, that is, the error signal when the locally generated signals $n_s(n)$ and $n_n(n)$ are zero. Finally, the error signal $e(n)$ is given by

$$e(n) = y(n) - \hat{d}(n) = e_u(n) + n(n) = e_u(n) + n_s(n) + n_n(n) \quad (15)$$

This error will enter the equation used to update the coefficients of the ECF. Obviously, only the fraction expressed by the undisturbed error $e_u(n)$ contains "useful" information. The locally generated signal $n(n)$, however, causes the filter to diverge and thus to increase the system distance. Therefore, independent of the used adaptive algorithm a control procedure is necessary to switch off or slow down the adaptation when $n(n)$ is large compared to the echo $d(n)$.

3.2.2. The Residual Echo-Suppressing Filter (RESF). The impact of the ECF on the acoustical echo is limited by — at least — two facts: (1) only echoes due to the linear part of the transfer function of the LEMS can be canceled and (2) the order of the ECF typically is much smaller than the order of the LEMS (see Section 2.1.3). Therefore, a second filter — the RESF — is used to reduce the echo further. The transfer function of this filter is given by the well-known Wiener equation [24,25]:

$$Q(\Omega, n) = \frac{S_{en}(\Omega, n)}{S_{ee}(\Omega, n)} \quad (16)$$

In this equation Ω is a normalized frequency, $S_{ee}(\Omega, n)$ is the short-term auto-power spectral density of the error signal $e(n)$, and $S_{en}(\Omega, n)$ is the short-term cross-power spectral density of $e(n)$ and the locally generated signal $n(n)$. In good agreement with reality one can assume that the undisturbed error $e_u(n)$ and $n(n)$ [see Eq. (15)] are orthogonal. Then the power spectral density $S_{ee}(\Omega, n)$ reduces to

$$S_{ee}(\Omega, n) = S_{e_u e_u}(\Omega, n) + S_{nn}(\Omega, n) \quad (17)$$

Furthermore, the cross-power spectral density $S_{en}(\Omega, n)$ is given by

$$S_{en}(\Omega, n) = S_{nn}(\Omega, n) \quad (18)$$

Then, it follows from Eq. (16) and after some manipulations for the transfer function of the RESF that

$$Q(\Omega, n) = 1 - \frac{S_{e_u e_u}(\Omega, n)}{S_{ee}(\Omega, n)} \quad (19)$$

The impulse response of the RESF is found by an inverse Fourier transformation.

Since the signals involved are highly nonstationary, the short-term power spectral densities have to be estimated for time intervals no longer than 20 ms. The overwriting problem, however, is that the locally generated signal $n(n)$ is observable only during the absence of the remote excitation signal $u(n)$. Since $n(n)$ is composed of local speech and local noise the RESF suppresses local noise, as well. It should be noted, however, that any impact of the RESF on residual echoes also impacts the local speech

signal $n_s(n)$ and, thus, reduces the quality of the speech output of the echo-canceling unit.

When applying Eq. (19), the power spectral densities $S_{ee}(\Omega, n)$ and $S_{e_u e_u}(\Omega, n)$ have to be replaced by their estimates $\hat{S}_{ee}(\Omega, n)$ and $\hat{S}_{e_u e_u}(\Omega, n)$. Therefore, it is possible that the quotient becomes larger than one. Consequently, the filter exhibits a phase shift of π. To prevent that, Eq. (19) of the filter transfer function is (heuristically) modified to

$$Q(\Omega, n) = 1 - \min\left[\frac{\hat{S}_{e_u e_u}(\Omega, n)}{\hat{S}_{ee}(\Omega, n)}, Q_{\min}\right] \tag{20}$$

where $Q_{\min}$ determines the maximal attenuation of the filter. Details can be found in, for example, Quatieri's book [26]. The problem of residual echo suppression is very similar to the problem of noise suppression and both are treated simultaneously [27,28].

4. ADAPTIVE ALGORITHMS

4.1. Normalized Least-Mean-Square (NLMS) Algorithm

The majority of implementations of acoustic echo-canceling systems use the NLMS algorithm to update the ECF. This gradient type algorithm minimizes the mean-square error [24]. The update equation is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\mu(n)}{\mathbf{u}^T(n)\,\mathbf{u}(n)}\,\mathbf{u}(n)\,e(n) \tag{21}$$

The term $\mathbf{u}^T(n)\,\mathbf{u}(n)$ in the denominator represents a normalization according to the energy of the input vector $\mathbf{u}(n)$. This is necessary because of the high variance of this quantity for speech signals. The step size of the update is controlled by the *step-size factor* $\mu(n)$. In general, the algorithm is stable (in the mean square) for $0 < \mu < 2$. Reducing the step size is necessary to prevent divergence of the filter coefficients in case of strong local signals $n_s(n)$ and/or $n_n(n)$.

The NLMS algorithm has no memory; that is, it uses only signals that are available at the time of update. This is advantageous for tracking changes of the LEMS. The update is performed in the direction of the input signal vector $\mathbf{u}(n)$ (see Fig. 10). For speech signals, consecutive vectors may be highly correlated, meaning that their directions differ only slightly. This is the reason for the low speed of convergence of the NLMS algorithm in case of speech excitation. Additional measures such as decorrelation of the input signal $u(n)$ and/or controlling the step-size parameter $\mu(n)$ (see Section 5) are necessary to speed up convergence.

The motivation for using the NLMS algorithm in the application discussed here is its robustness and its low computational complexity that is only in the order of $2N$ operation per coefficient update.

Decorrelating the input signal offers a computationally inexpensive method to improve the convergence of the filter coefficients. To achieve this two (identical) decorrelation filters and an inverse filter have to be added to the echo-canceling system (see Fig. 11). The decorrelation filter has to be duplicated since the loudspeaker needs the original signal $u(n)$. Simulations show that even filters of first order approximately double the speed of convergence in acoustic echo-canceling applications [11]. Further improvements require adaptive decorrelation filters because of the nonstationarity of speech signals. Also, in case of higher-order filters the necessary interchange of the decorrelation filter and the time-varying LEMS causes additional problems. Therefore, only the use of a first-order decorrelation filter can be recommended. The curves in Fig. 11(a) are averages over several speech signals with pauses removed.

4.2. Affine Projection (AP) Algorithm

The AP algorithm [29] overcomes the weakness of the NLMS algorithm concerning correlated input signals by updating the filter coefficients not just in the direction of the current input vector but also within a hyperplane spanned by the current input vector and its $M-1$ immediate predecessors (see Fig. 10). To accomplish this an *input signal matrix* $\mathbf{U}(n)$ is formed

$$\mathbf{U}(n) = [\mathbf{u}(n), \mathbf{u}(n-1), \mathbf{u}(n-2), \ldots, \mathbf{u}(n-M+2) \times \mathbf{u}(n-M+1)] \tag{22}$$

and an error vector is calculated

$$\mathbf{e}(n) = [y(n), y(n-1), \ldots, y(n-M+1)]^T - \mathbf{U}^T(n)\,\mathbf{w}(n) \tag{23}$$

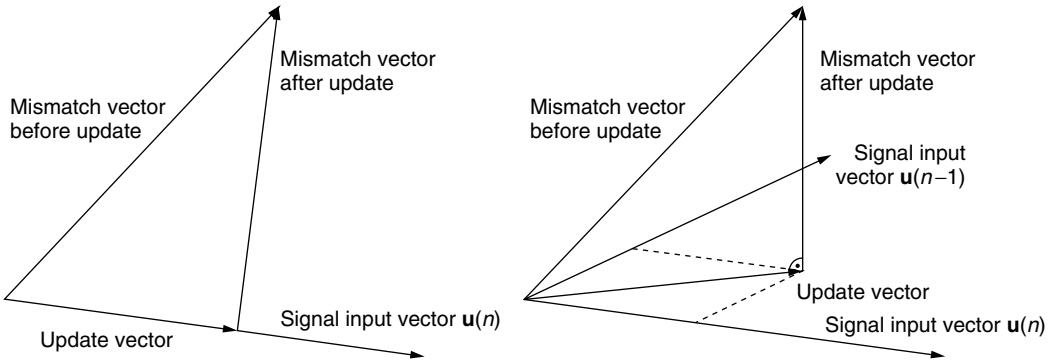


Figure 10. Updates of the system mismatch vector according to the NLMS algorithm (left) and to the AP algorithm (right).

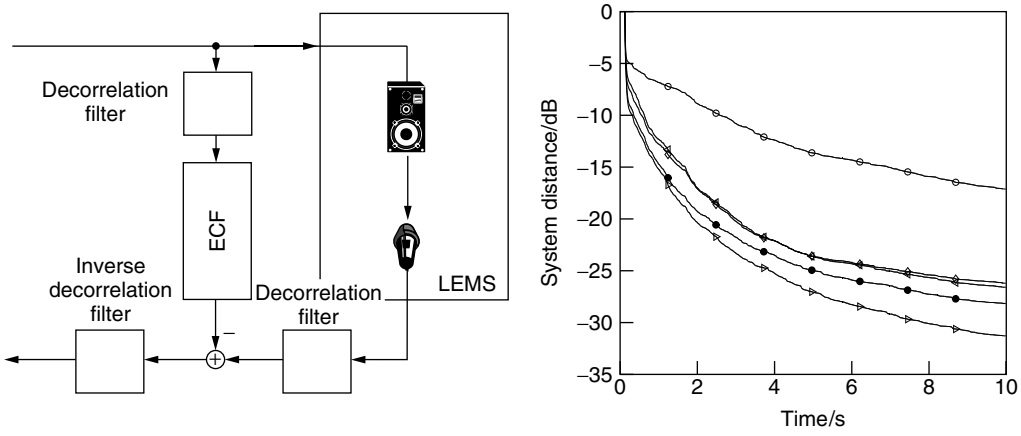


Figure 11. Insertion of decorrelation filters (**a**) and improved concergence of NLMS algorithm utilizing decorrelation filters (**b**): ○: none, ◇: fixed first order, ◁: fixed second order, ●: adaptive tenth order, ▷: adaptive 18th order (sampling frequency = 8 kHz).

collecting the errors for the current and the $M-1$ past input signal vectors applied to the ECF with the *current* coefficient setting. The price to be paid for the improved convergence is the increased computational complexity caused by the inversion of an $M \times M$ matrix required at each coefficient update. Fast versions of this algorithm are available [30,31].

Finally, the update equation is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(n)\,\mathbf{U}(n)\,(\mathbf{U}^T(n)\,\mathbf{U}(n))^{-1}\,\mathbf{e}(n) \qquad (24)$$

Numerical problems arising during the inversion of the matrix $\mathbf{U}^T(n)\,\mathbf{U}(n)$ can be overcome by using $\mathbf{U}^T(n)\,\mathbf{U}(n) + \delta\,\mathbf{1}$ instead of $\mathbf{U}^T(n)\,\mathbf{U}(n)$, where $\mathbf{1}$ is the unit matrix and δ is a small positive constant. For $M = 1$ the AP algorithm is equal to the NLMS procedure. For speech input signals even $M = 2$ leads to a considerably faster convergence of the filter coefficients (see Fig. 12). Suggested values for M are between 2 and 5 for the ECF update.

It should be noted, however, that faster convergence of the ECF coefficients also means faster divergence in case of strong local signals. Therefore, faster control of the step size is required as well. Its optimal value is based on estimated quantities (see Section 5). Since their reliabilities depend on the lengths of the data records usable for the estimation, a very high speed of convergence may not be desirable. Nevertheless, the AP algorithm seems to be a good candidate to replace the NLMS algorithm in acoustic echo-canceling applications.

4.3. Recursive Least-Squares (RLS) Algorithm

The RLS algorithm minimizes the sum of the squared error

$$\overline{e^2(n)} = \sum_{k=0}^{n} \lambda^{n-k}\, e^2(k) \qquad (25)$$

where $e(n)$ is given by Eqs. (15) and (3). It calculates an estimate $\hat{\mathbf{S}}_{uu}(n)$ of the autocorrelation matrix of the input

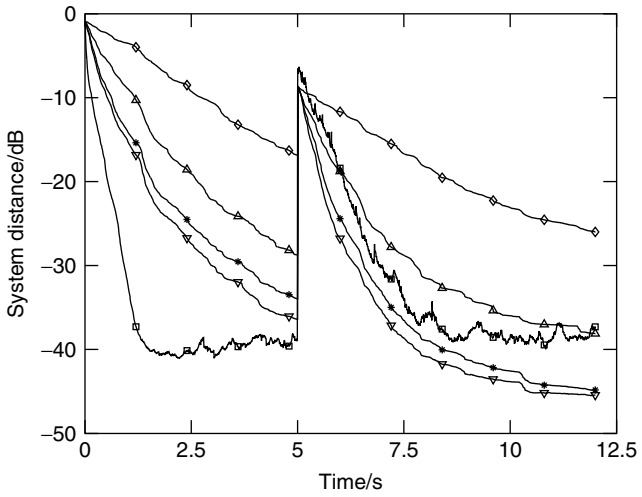


Figure 12. Convergence of the filter coefficients for different adaptation algorithms (filter length = 1024, sampling frequency = 8 kHz): ◇: NLMS ($\mu = 1$), △: AP of order two, ∗: AP of order 5, ▽: AP of order 10 (all AP algorithms with $\mu = 1$), □: RLS ($\lambda = 0.9999$). The impulse response of the LEMS is changed at $t = 5$ s.

signal vector

$$\hat{\mathbf{S}}_{uu}(n) = \sum_{k=0}^{n} \lambda^{n-k}\, \mathbf{u}(k)\, \mathbf{u}^T(k) \qquad (26)$$

and uses the inverse of this $N \times N$ matrix to decorrelate the input signal in the update equation. The factor λ is called the *forgetting factor*. It is chosen close to but smaller than one and assigns decreasing weights to the input signal vectors the further they are in the past. In addition, the *a priori error* $\tilde{e}(n)$, defined by

$$\tilde{e}(n+1) = d(n+1) - \mathbf{u}^T(n+1)\,\mathbf{w}(n) \qquad (27)$$

is calculated. This is the error calculated with the new input vector but with the not yet updated filter coefficients.

Finally, the update equation of the RLS algorithm is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(n)\,\hat{\mathbf{S}}_{uu}^{-1}(n+1)\,\mathbf{u}(n+1)\,\tilde{e}(n+1) \tag{28}$$

In contrast to the AP algorithm, now an $N \times N$ matrix has to be inverted at each coefficient update. This can be done recursively, and fast RLS algorithms are available [32] with numerical complexity in the order of $7N$.

The RLS algorithm exhibits considerable stability problems. One problem is caused by finite wordlength errors, especially when the procedure is executed in 16 bit fixed-point arithmetic. The second problem arises from the properties of speech signals (see Section 2.2.1), temporarily causing the estimate of the (short-term) autocorrelation matrix to become singular. A forgetting factor λ very close to one helps to overcome this problem. On the other hand, however, the long memory caused by a λ close to one slows down the convergence of the coefficients of the ECF after a change of the LEMS.

In spite of the speed of convergence achievable with the RLS algorithm, the numerical complexity and the stability problems so far prevented the use of this algorithm in acoustical echo-cancelation applications.

A unified analysis of least squares adaptive algorithms can be found in the paper by Glentis et al. [33].

5. STEP-SIZE CONTROL OF THE NLMS ALGORITHM

Independent of the specific algorithm used, the update of the coefficients of the ECF strongly depends on the error signal $e(n)$. This signal is composed of the undisturbed error $e_u(n)$ and the locally generated signal $n(n)$ [see Eq. (15)]. Only $e_u(n)$ steers the coefficients toward their optimal values. The step-size factor $\mu(n)$ is used to control the update according to the ratio of both contributions. Assuming the filter has converged, the error signal $e(n)$ has assumed a certain value. Suddenly the amplitude of $e(n)$ is increased. This may have two reasons that require different actions: (1) a local speaker became active or a local noise started — in this case the step size has to be reduced to prevent losing the degree of convergence achieved before, and (2) the impulse response of the LEMS has changed, for example, by the movement of the local talker. Now, the step size has to be increased to its maximal possible value in order to adapt the ECF to its new impulse response as fast as possible.

A major problem becomes visible with this consideration — the not-directly-observable undisturbed error signal $e_u(n)$ needs to be known in order to control the adaptation process. Another leading point should be mentioned here. The first situation requires immediate action. In the second case, a delayed action causes an audible echo but no divergence of the ECF.

5.1. Optimal Step Size for the NLMS Algorithm

In this section an optimal step size for the NLMS algorithm will be derived assuming that all required quantities are available. The following sections explain how to estimate them from measurable signals.

Using Eq. (21) and assuming that the impulse response of the LEMS does not change

$$\mathbf{h}(n+1) = \mathbf{h}(n) \tag{29}$$

the mismatch [see Eq. (12)] is given by

$$\begin{aligned}\boldsymbol{\varepsilon}(n+1) &= \mathbf{h}(n+1) - \mathbf{w}(n+1) \\ &= \mathbf{h}(n) - \mathbf{w}(n) - \frac{\mu(n)}{\|\mathbf{u}(n)\|^2}\,\mathbf{u}(n)\,e(n) \qquad (30)\\ &= \boldsymbol{\varepsilon}(n) - \frac{\mu(n)}{\|\mathbf{u}(n)\|^2}\,\mathbf{u}(n)\,e(n) \qquad (31)\end{aligned}$$

Using Eqs. (13) and (14), the expectation of the system distance can be expressed as

$$\begin{aligned}E\{\Delta(n+1)\} = E\{\Delta(n)\} &- 2\,\mu(n)\,E\left\{\frac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\} \\ &+ \mu^2(n)\,E\left\{\frac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\}\end{aligned} \tag{32}$$

For an optimal step size, it is required that

$$E\{\Delta(n+1)\} - E\{\Delta(n)\} \le 0 \tag{33}$$

Inserting Eq. (32) leads to

$$\mu^2(n)\,E\left\{\frac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\} - 2\,\mu(n)\,E\left\{\frac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\} \le 0 \tag{34}$$

Thus, the condition for the optimal step size is given by

$$0 \le \mu(n) \le 2\,\frac{E\left\{\dfrac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\}}{E\left\{\dfrac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\}} \tag{35}$$

A step size in the middle of this range achieves the fastest decrease of the system distance. The optimal step size μ_{opt} therefore is

$$\mu_{\text{opt}}(n) = \frac{E\left\{\dfrac{e(n)\,e_u(n)}{\|\mathbf{u}(n)\|^2}\right\}}{E\left\{\dfrac{e^2(n)}{\|\mathbf{u}(n)\|^2}\right\}} \tag{36}$$

To simplify this result, one can assume that the L_2 norm of the input signal vector $\mathbf{u}(n)$ is approximately constant. This can be justified by the fact that in echo-canceling applications the length of this vector typically is in the order of 512–2048. Then, the optimal step size is given by

$$\mu_{\text{opt}}(n) \approx \frac{E\{e(n)\,e_u(n)\}}{E\{e^2(n)\}} \tag{37}$$

Since the undisturbed error $e_u(n)$ and the locally generated signal $n(n)$ are uncorrelated, this expression further simplifies to

$$\mu_{\text{opt}}(n) \approx \frac{E\{e_u^2(n)\}}{E\{e^2(n)\}} \tag{38}$$

Finally, the denominator may be extended using Eq. (15), and again the property that $e_u(n)$ and $n(n)$ are orthogonal:

$$\mu_{\text{opt}}(n) \approx \frac{E\{e_u^2(n)\}}{E\{e_u^2(n)\} + E\{n^2(n)\}} \tag{39}$$

Equation (39) emphasizes the importance of the undisturbed error $e_u(n)$, specifically, if there is a good match between the LEMS and the ECF, this term is small. If at the same time the local signal $n(n)$ is large, the optimal step size approaches zero; the adaptation freezes.

We have discussed here only a scalar step-size factor $\mu(n)$. This means that the same factor is applied to all filter coefficients. Numerous suggestions have been made to replace the scalar step size factor by a diagonal matrix in order to apply distinct factors to different elements of the coefficient vector. An example is an exponentially weighted step size taking into account the exponential decay of the impulse response of LEMS [34,35].

The implementation of the optimal step size needs the solution of a number of problems that will be discussed in the following section.

5.2. Implementation of the Optimal Step Size

The implementation of the optimal step size derived in the previous section requires the estimation of several quantities, including

- The expectations of signal powers have to be approximated by short-term estimates.
- An estimation method for the not-directly-observable undisturbed error $e_u(n)$ has to be derived.

5.2.1. Estimation of Short-Term Signal Power. Short-term signal power can be easily estimated by squaring the signal amplitude and smoothing this value by an IIR filter. A filter of first order proved to be sufficient [36]. If a rising signal amplitude should be detected faster than a falling one, a shorter time constant for a rising edge can be used than for a falling one. Typically both constants are chosen out of [0.9, 0.999]. Applying different time constants gives rise to a (small) bias that can be neglected in this application. Where squaring the signal amplitude causes a problem because of fixed-point arithmetic, the square can be replaced by the magnitude of the amplitude. Both square and magnitude are related by a factor depending on the probability density function of the signal amplitude. If two short-term estimates of signal powers are compared with each other, as is done in most cases in controlling the ECF, this factor cancels out.

5.2.2. Estimation of the Undisturbed Error. Two methods will be described to estimate the undisturbed error. The first one will use so-called delay coefficients. The second procedure compares signal powers at the input and the output of the LEMS. Both need supporting measures in order to distinguish between local activities and alterations of the impulse response of the LEMS.

5.2.2.1. Estimation via Delay Coefficients. Estimating the undisturbed error needs an estimate of the mismatch vector $\varepsilon(n)$ [see Eq. (12)]. Obviously, the impulse response vector $\mathbf{h}(n)$ of the LEMS is not known. However, if an artificial delay of N_D samples is inserted before the loudspeaker [37], the ECF also models this part of the unknown impulse response. The impulse response coefficients related to this delay are zero:

$$h_i(n) = 0 \quad \text{for } i = 0, \ldots, N_D - 1 \tag{40}$$

The NLMS algorithm has the property to distribute coefficient errors equally over all coefficients. Therefore, from the mismatch of the first N_D coefficients, one can estimate the system distance [see Eq. (13)]:

$$\hat{\Delta}(n) = \frac{N}{N_D} \sum_{i=0}^{N_D-1} w_i^2(n) \tag{41}$$

Assuming statistical independence of the input signal $u(n)$ and the filter coefficients, the optimal step size according to Eq. (38) is approximately given by

$$\mu_{\text{opt}}(n) \approx \frac{E\{u^2(n)\}\,\hat{\Delta}(n)}{E\{e^2(n)\}} \tag{42}$$

The performance of this method proves to be quite reliable. It has one deficiency, however. The update of the ECF freezes in case of a change of the impulse response of the LEMS. The reason for this behavior is that the coefficients related to the artificial delay remain equal to zero in that case. Therefore, applying this method requires an additional detector for changes of the LEMS.

Several methods are known [36]. A reliable indicator is based on a so-called *shadow filter*. This is a short adaptive filter in parallel to the ECF. Its step size is controlled only by the excitation signal $u(n)$. Consequently, it does not stop in case of a change of the LEMS. Since the shadow filter has far fewer coefficients than the ECF, it converges (but also diverges) much faster. During normal operation periods, the output signal of the shadow filter is inferior to the output signal of the ECF due to the nonoptimal step size control and the insufficient degree compared to the degree of the LEMS. Only immediately after a system change, the error calculated from the output of the shadow filter and the microphone output is smaller than the error based on the output signal of the ECF. If this situation is detected, the step size of the ECF adaptation is increased artificially in order to restart adaptation.

5.2.2.2. Estimation via Power Comparison. A second estimation method of the short-term power of the undisturbed error $E\{e_u^2(n)\}$ is based on the estimation of a so-called power transfer factor $\beta(n)$. If there is sufficient remote excitation and if there are no local activities $[n(n) = 0]$ this factor is given by

$$\beta(n) = \frac{E\{e^2(n)\}}{E\{u^2(n)\}} \tag{43}$$

For the expectations, short-term estimates (see Section 5.2.1) can be used. The estimation of the power transfer factor requires intervals of remote singletalk. These have

to be determined by a *doubletalk detector*. One method is based on a correlation measure ρ between the microphone output signal $y(n)$ and the output signal $\hat{d}(n)$ of the ECF. In case of a sufficiently converged ECF $\hat{d}(n)$ is a good estimate of the echo signal $d(n)$. Also, it is synchronized with $d(n)$. Therefore, the correlation must be evaluated only for a delay of zero. In contrast, correlation of $u(n)$ and $y(n)$ requires to search for the maximum of the measure. Further, it is advisable to normalize the correlation measure $\rho(n)$ defined as

$$\rho(n) = \frac{\left| \sum_{k=0}^{N_c-1} \hat{d}(n-k)\, y(n-k) \right|}{\sum_{k=0}^{N_c-1} |\hat{d}(n-k)\, y(n-k)|} \tag{44}$$

The value for N_c has to be determined by a compromise between a reliable correlation measure and the delay required for its calculation. Remote singletalk is assumed if $\rho(n)$ is larger than a given threshold. Since even in the case of singletalk, low local noise $n_n(n)$ may be present in the microphone output $y(n)$. Consequently, the estimate for the power coupling factor β may be too large. Therefore, the optimal step size $\mu_{\text{opt}}(n)$ should not exceed a given upper bound. For a reliable operation it is necessary to smooth the result of Eq. (44).

During intervals with no remote singletalk condition the power coupling factor $\beta(n)$ has to be frozen. Let the most recent doubletalk interval start at time n_1. Then the power transfer factor $\beta(n)$ is set to $\beta(n_1 - 1)$ during this interval.

With this modifications the factor $\beta(n)$ can replace $\hat{\Delta}(n)$ in Eq. (42):

$$\mu_{\text{opt}}(n) \approx \begin{cases} \dfrac{E\{u^2(n)\}\,\overline{\beta(n)}}{E\{e^2(n)\}} & \text{during remote singletalk} \\ \dfrac{E\{u^2(n)\}\,\overline{\beta(n_1-1)}}{E\{e^2(n)\}} & \text{during doubletalk} \end{cases} \tag{45}$$

where $\overline{\beta(n)}$ is a smoothed value of $\beta(n)$.

In general, doubletalk detectors can be based on other measures such as the cepstral distance, coherence, or the likelihood ratio [38–41].

6. ECHO CANCELLATION FOR STEREOPHONIC SYSTEMS

In stereophonic telecommunication systems there are four ECFs (per location) necessary to model the four echo paths between left and right loudspeakers and left and right microphones (see Fig. 13). In addition to providing the increased processing power a problem specific to stereophonic systems has to be solved — the signals on the left and the right channel originate from the same signal source and are separated only by the convolution with the impulse responses $\mathbf{g}_R(n)$ and $\mathbf{g}_L(n)$ of the transmission paths between signal source and left and right microphone at the remote location. Typically, both impulse responses exhibit minimal phase components that are invertible. Therefore, the impulse responses of the ECFs do not necessarily converge to the impulse responses of the related echo paths. Preprocessing of the loudspeaker input signals is required in order to decorrelate both signals. A large number of methods to achieve this goal have been suggested [42–48].

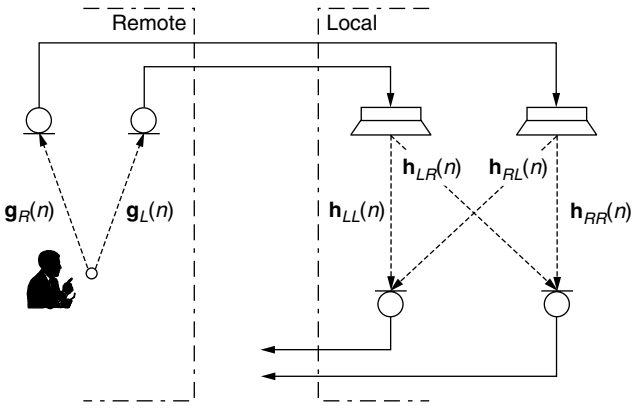


Figure 13. Signal paths in a stereophonic telecommunication system.

One method for decorrelation of left and right channels consists of introducing a nonlinearity into one or both of the channels. Up to a certain degree that depends on the quality of the audio equipment, this distortion proved to be not audible. A special suggestion is adding signals to the inputs of the loudspeakers that are nonlinear functions, such as half-wave rectifications of these signals [49]. Another way to achieve decorrelation is obtained by the insertion of a periodically varying delay [50,51]. The tolerable amount of delay has to be limited such that the spatial information is maintained. Other proposals consist of adding noise such that it is masked by the speech signal [52] and using audio coding methods [53].

7. SUBBAND ECHO CANCELLATION

Cancelling acoustic echoes in subbands (see Fig. 14) leads to a reduced processing load and increases the degrees of freedom for optimizing the echo cancelling system [54–57].

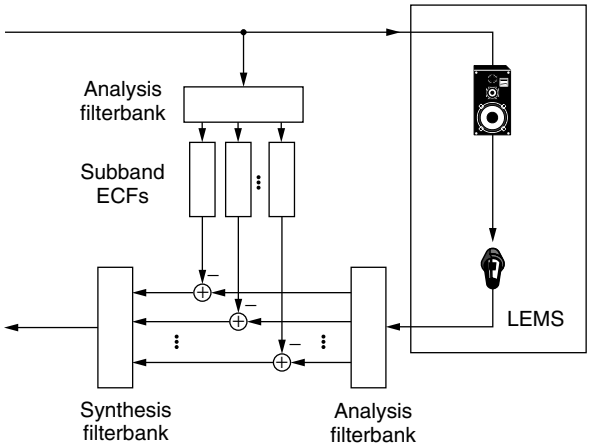


Figure 14. General structure of subband echo cancellation filters.

Prohibitive in many applications, however, is the enlarged signal delay (see Section 2.3) that is inevitably connected with the insertion of filterbanks.

Assume that the signal $u(n)$ will be split into K subbands. Then K filters have to be adapted. The subband signals can be decimated by a factor of K (critical decimation according to Nyquist's law) and the lengths of the subband ECFs can be reduced to N/K, assuming for simplicity that N is a multiple of K. Adaptation takes place after every K sampling intervals. Compared to full-band processing, the number of operations is given by

$$K(\text{filters}) \times \frac{1}{K}(\text{reduced filter lengths}) \times \frac{1}{K}(\text{extended updating interval}) \quad (46)$$

resulting in a reduction by the factor K. In addition, however, two analysis and one synthesis filterbanks are necessary. If the signals are split into equally wide subbands polyphase filters may be applied that can be efficiently implemented [58–62]. Critical decimation of subband signals may give rise to crosstalk problems between adjacent frequency bands. A decimation factor slightly smaller than K eases these problems [63–65]. Using the NLMS algorithm, the speed of convergence of the filter coefficients is inversely proportional to the filter length. For subband ECFs this length is N/K and the extended adaptation interval is compensated by the faster convergence. Splitting into subbands partially whitens the signals also causing a faster convergence (using the NLMS algorithm).

Additional savings in processing power are possible by using the degrees of freedom offered by subband processing for a better tuning of the echo-canceling system. The subband filters need not have the same lengths. Since the major part of the energy of speech signals is concentrated within the lower subbands (see Fig. 8) — and so is the echo — the ECFs of the upper bands can be shorter than those of the lower-frequency bands. In addition, the sound absorption of materials used in typical enclosures such as offices increases with frequency. Thus, echoes in higher-frequency bands are attenuated faster. A design criterion can be the contribution of the energy of the noncanceled echo tails to the energy of the total error $e(n)$ [66]. Finally, routines such as the doubletalk detector need to be implemented only in one subband. Choosing the band in which the speech signal has its highest energy enhances the reliability of the detection.

8. BLOCK PROCESSING SOLUTIONS

The adaptation of the coefficients of the ECF requires a huge amount of computational power. It can be reduced by applying block processing algorithms [67]. In this case an update takes place only after intervals of length $BT, B > 1$, where T is the sampling time. Between updates the data are assembled in blocks of length B. At each update, a block of B samples of the filter output signal is generated. The efficiency of block processing algorithms increases with the blocklength. On the other hand, however, collecting signal samples in blocks introduces a delay corresponding to the blocklength. Therefore, in time-critical applications, such as hands-free telephones, only short blocks can be tolerated.

Since adaptation takes place only once in B sampling intervals, the speed of convergence of the filter coefficients is reduced accordingly. To speed up convergence, it is possible to correct the error signal (i.e., a vector of length B in block processing procedures) such that it is identical to the error signal generated by adaptation at each sampling instant [68–70]. In this case the NLMS algorithm based on block processing behaves exactly such as the ordinary NLMS algorithm. The filter adaptation may be performed in the time or in the frequency domain [71].

An desirable choice of the blocklength B would be the length N of the ECF [72]. In typical applications, however, this filter has to cover echoes that are in the order of 32–125 ms long. A blocklength equal to N, therefore, would introduce a nontolerable signal delay. To overcome this problem, the ECF has to be partitioned into subfilters of smaller length [73]. Thus the blocklength and the delay related to it can be tailored to the specific needs of the application. An efficient algorithm is available that combines the error signal correction and the partitioning of the filter impulse response with an overlap-save implementation calculating the subfilter updates and the filter output in the frequency domain [11,12].

9. CONCLUSIONS AND OUTLOOK

Powerful and affordable acoustical echo-canceling systems are available. Their performance is satisfactory, especially if compared to solutions in other voice processing areas such as speech recognition or speech-to-text translation. The fact that echo control systems have not yet entered the market on a large scale seems not to be a technical but a marketing problem — a customer who buys a high-quality echo suppressing system pays for the comfort of his her communication partner. Using a poor system only affects the partner at the far end, who usually is too polite to complain.

Future research and development in the area of acoustic echo cancellation certainly will not have to take into account processing power restrictions. This has a number of consequences; the implementation of even sophisticated procedures on ordinary (office) PCs will be possible. This will make it easier to test modifications of existing procedures or completely new ideas in real time and in real environments. The performance of future systems will approach limits given only by the environment they have to work in. It will no longer be limited by the restricted capabilities of affordable hardware. It will depend only on the quality of the algorithms implemented.

This does not necessarily mean that future systems will be perfectly reliable in all situations. The reliability of estimation procedures used to detect system states such as a change of the impulse response of the LEMS or the beginning of doubletalk depends on the length of the usable data record. Since, however, the working environment is highly time-varying and nonstationary the usage of too long records can cause the loss of the real-time capability.

Up to now the NLMS algorithm plays the role of the "workhorse" for acoustic echo cancelling. The AP algorithm offers improved performance at modest additional implementation and processing cost. It does not cause stability problems that are difficult to solve. Rules for step-size control used for the NLMS algorithm, however, have to be reconsidered.

Customer demands are increasing with time. Using available systems, customers will certainly ask for better performance. Therefore, the need for new and better ideas will remain. Acoustic echo canceling will continue to be one of the most interesting problems in digital signal processing.

BIOGRAPHY

Eberhard Hänsler received his degrees (Dipl.-Ing., 1961, Dr.-Ing., 1968) in Electrical Engineering from Darmstadt University of Technology, Darmstadt, Germany. He worked with the Research Institute of the German PTT (1961–1963), the Electrical Engineering Department of Darmstadt University of Technology (1963–1968), and with the IBM Research Division at Zurich and Yorktown Heights (1968–1974). Since 1974 he has been Full Professor for Signal Theory at Darmstadt University of Technology.

His research interests are signal and system theory, adaptive systems, digital signal processing, echo cancellation, and noise reduction. He has been working on the hands-free telephone problem for several years. He is cofounder of the biennial International Workshop on Acoustic Echo and Noise Control (IWAENC), and has organized sessions on this topic at several international conferences and acted as guest editor of special issues on acoustic echo and noise control of *signal processing* (January 1998) and of the *European Transactions on Telecommunications* (March/April 2002).

Together with his group, he received the Annual European Group Technical Achievement Award in 2000 for "major contributions in the design and implementation of acoustic echo and noise control system."

BIBLIOGRAPHY

1. The new Bell telephone, *Sci. Am.* **37**: 1 (1877).
2. W. F. Clemency, F. F. Romanow, and A. F. Rose, The Bell system speakerphone, *AIEE. Trans.* **76**(I): 148–153 (1957).
3. D. A. Berkley and O. M. M. Mitchell, Seeking the ideal in "hands-free" telephony, *Bell Lab. Rec.* **52**: 318–325 (1974).
4. G. Pays and J. M. Person, Modèle de laboratoire d'un poste téléphonique à haut-parleur, *FASE* **75**: 88–102 (Paris) (1975).
5. E. Hänsler, The hands-free telephone problem — an annotated bibliography, *Signal Process.* **27**: 259–271 (1992).
6. E. Hänsler, The hands-free telephone problem — an annotated bibliography update, *Annales des Télécommunications* **49**: 360–367 (1994).
7. E. Hänsler, The hands-free telephone problem — a second annotated bibliography update, *Proc. 4th Int. Workshop on Acoustic Echo and Noise Control*, 1995, pp. 107–114.
8. A. Gilloire et al., Innovative speech processing for mobile terminals: An annotated bibliography, *Signal Process.* **80**(7): 1149–1166 (2000).
9. A. Gilloire, State of the art in acoustic echo cancellation, in A. R. Figueiras and D. Docampo, eds., *Adaptive Algorithms: Applications and Non Classical Schemes*, Univ. Vigo, 1991, pp. 20–31.
10. A. Gilloire, E. Moulines, D. Slock, and P. Duhamel, State of the art in acoustic echo cancellation, in A. R.Figueiras-Vidal, ed., *Digital Signal Processing in Telecommunications*, Springer, London, 1996, pp. 45–91.
11. C. Breining et al., Acoustic echo control, *IEEE Signal Process. Mag.* **16**(4): 42–69 (1999).
12. S. L.Gay and J.Benesty, eds., *Acoustic Signal Processing for Telecommunication*, Kluwer, Boston, 2000.
13. J. Benesty et al., *Advances in Network and Acoustic Echo Cancellation*, Springer, Berlin, 2001.
14. H. Kuttruff, Sound in enclosures, in M. J. Crocker, ed., *Encyclopedia of Acoustics*, Wiley, New York, 1997, pp. 1101–1114.
15. J. B. Allen and D. A. Berkley, Image method for efficiently simulating small-room acoustics, *J. Acoust. Soc. Am.* **65**: 943–950 (1975).
16. M. Zollner and E. Zwicker, *Elektroakustik*, Springer, Berlin, 1993.
17. M. Mboup and M. Bonnet, On the adequateness of IIR adaptive filtering for acoustic echo cancellation, *Proc. EUSIPCO-92*, Brussels, Belgium, 1992, pp. 111–114.
18. A. P. Liavas and P. A. Regalia, Acoustic echo cancellation: Do IIR filters offer better modelling capabilities than their FIR counterparts? *IEEE Trans. Signal Process.* **46**(9): 2499–2504 (1998).
19. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
20. International Telecommunication Union, *Acoustic Echo Controllers*, ITU-T Recommendation G.167, 1993.
21. International Telecommunication Union, *Control of Talker Echo*, ITU-T Recommendation G.131, 1996.
22. International Telecommunication Union, *Relation Between Echo Disturbances under Single Talk and Double Talk Conditions* (*Evaluated for One-Way Transmission Time of 100 ms*), ITU-T Recommendation G.131(App. II), 1999.
23. M. R. Schroeder, Improvement of acoustic-feedback stability by frequency shifting, *J. Acoust. Soc. Am.* **36**: 1718–1724 (1964).
24. S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice-Hall, Englewood Cliffs, NJ, 2002.
25. E. Hänsler and G. U. Schmidt, Hands-free telephones — joint control of echo cancellation and post filtering, *Signal Process.* **80**: 2295–2305 (2000).
26. T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice-Hall, Upper Saddle River, NJ, 2002.
27. R. Martin and P. Vary, Combined acoustic echo control and noise reduction for hands-free telephony — state of the art and perspectives, *Proc. EUSIPCO-96*, Triest, Italy, 1996, pp. 1107–1110.
28. S. Gustafsson, R. Martin, and P. Vary, Combined acoustic echo control and noise reduction for hands-free telephony, *Signal Process.* **64**: 21–32 (1998).

29. K. Ozeki and T. Umeda, An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, *Electron. Commun. Jpn.* **67-A**(5): 19–27 (1984).
30. S. Gay and S. Travathia, The fast affine projection algorithm, *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 3023–3027.
31. V. Myllylä, Robust fast affine projection algorithm for acoustic echo cancellation, *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, 2001, pp. 143–146.
32. D. Slock and T. Kailath, Fast transversal RLS algorithms, in N. Kalouptsidis and S. Theodoridis, eds., *Adaptive System Identification and Signal Processing Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
33. G.-O. Glentis, K. Berberidis, and S. Theodoridis, Efficient least squares adaptive algorithms for FIR transversal filtering: A unified view, *IEEE Signal Process. Mag.* **16**(4): 13–41 (1999).
34. S. Makino and Y. Kaneda, Exponentially weighted step-size projection algorithm for acoustic echo cancellers, *IECE Trans. Fund.* **E75-A**: 1500–1507 (1992).
35. S. Makino, Y. Kaneda, and N. Koizumi, Exponentially weighted step-size NLMS adaptive filter based on the statistics of a room impulse response, *IEEE Trans. Speech Audio Process.* **1**: 101–108 (1993).
36. A. Mader, H. Puder, and G. Schmidt, Step-size control for acoustic echo cancellation filters — an overview, *Signal Process.* **80**: 1697–1719 (2000).
37. S. Yamamoto and S. Kitayama, An adaptive echo canceller with variable step gain method, *Trans. IECE Jpn.* **E65**: 1–8 (1982).
38. T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, Double-talk detector based on coherence, *IEEE Trans. Commun.* **COM-44**: 1421–1427 (1996).
39. K. Ghose and V. U. Redd, A double-talk detector for acoustic echo cancellation applications, *Signal Process.* **80**: 1459–1467 (2000).
40. H. Ye and B. Wu, A new double-talk detection algorithm based on the orthogonality theorem, *IEEE Trans. Commun.* **COM-39**: 1542–1545 (1991).
41. A. H. Gray and J. D. Markel, Distance measures for speech processing, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-24**: 380–391 (1976).
42. J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, Adaptive filtering algorithms for stereophonic acoustic echo cancellation, *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 3099–3102.
43. S. Shimauchi and S. Makino, Stereo projection echo canceller with true echo path estimation, *Proc. ICASSP-95*, Detroit, MI, 1995, pp. 3059–3062.
44. F. Amand, J. Benesty, A. Gilloire, and Y. Grenier, A fast two-channel projection algorithm for stereophonic acoustic echo cancellation, *Proc. ICASSP-95*, Atlanta, GA, 1996, pp. 949–952.
45. S. Shimauchi, Y. Haneda, S. Makino, and Y. Kaneda, New configuration for a stereo echo canceller with nonlinear pre-processing, *Proc. ICASSP-98*, Seattle, OR, 1998, pp. 3685–3688.
46. S. Shimauchi et al., A stereo echo canceller implemented using a stereo shaker and a duo-filter control system, *Proc. ICASSP-99*, Phoenix, AZ, 1999, pp. 857–860.
47. T. Gänsler and J. Benesty, New insights to the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution, *IEEE Trans. Speech Audio Process.* **9**: 686–696 (1998).
48. A. Sugiyama, Y. Joncour, and A. Hirano, A stereo echo canceler with correct echo-path identification based on an input-sliding technique, *IEEE Trans. Signal Process.* **49**: 2577–2587 (2001).
49. J. Benesty, D. R. Morgan, and M. M. Sondhi, A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation, *IEEE Trans. Speech Audio Process.* **6**: 156–165 (1998).
50. Y. Joncour and A. Sugiyama, A stereo echo canceler with preprocessing for correct echo-path identification, *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 3677–3680.
51. M. Ali, Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation, *Proc. ICASSP-98*, Seattle, OR, 1998, pp. 3689–3692.
52. A. Gilloire and V. Turbin, Using auditory properties to improve the behavior of stereophonic acoustic echo cancellers, *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 3681–3684.
53. T. Gänsler and P. Eneroth, Influence of audio coding on stereophonic acoustic echo cancellation, *Proc. ICASSP-98*, Seattle, WA, 1998, pp. 3649–3652.
54. I. Furukawa, A design of canceller of broad band acoustic echo, *Int. Teleconf. Symp.*, Tokyo, Japan, Jan. 8–Aug. 8, 1984.
55. A. Gilloire, Adaptive filtering in sub-bands, *Proc. ICASSP-88*, New York, 1988, pp. 1572–1576.
56. W. Kellermann, Analysis and design of multirate systems for cancellation of acoustical echoes, *Proc. ICASSP-88*, New York, 1988, pp. 2570–2573.
57. W. Kellermann, Zur Nachbildung physikalischer Systeme durch parallelisierte digitale Ersatzsysteme im Hinblick auf die Kompensation akustischer Echos, *Fortschr.-Ber. VDI Reihe* 10(102), VDI Verlag, Düsseldorf, Germany, 1989.
58. R. E. Chrochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
59. P. Vary and G. Wackersreuther, A unified approach to digital polyphase filter banks, *AEÜ Int. J. Electron. Commun.* **37**: 29–34 (1983).
60. G. Wackersreuther, On the design of filters for ideal QMF and polyphase filter banks, *AEÜ Int. J. Electron. Commun.* **39**: 123–130 (1985).
61. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
62. P. P. Vaidyanathan, Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial, *Proc. IEEE* **78**: 56–93 (1990).
63. S. Weiss, R. W. Stewart, A. Stenger, and R. Rabenstein, Performance limitations of subband adaptive filters, *Proc. EUSIPCO-98*, Rhodos, Greece, 1998, pp. 1245–1248.
64. S. Weiss, *On Adaptive Filtering in Oversampled Subbands*, Ph.D. dissertation, Dept. Electronic and Electrical Engineering, Univ. Strathclyde, May 1998.
65. G. U. Schmidt, Entwurf und Realisierung eines Multiratensystems zum Freisprechen, *Fortschr.-Ber. VDI Reihe* 10(674), VDI Verlag, Düsseldorf, Germany, 2001.
66. G. U. Schmidt, Acoustic echo control in subbands — an application of multirate systems, *Proc. EUSIPCO-98*, Rhodos, Greece, 1998, pp. 1961–1964.

67. J. Shynk, Frequency-domain and multirate adaptive filtering, *IEEE Signal Process. Mag.* **9**(1): 14–37 (1992).

68. J. Benesty and P. Duhamel, A fast exact least mean square adaptive algorithm, *IEEE Trans. Signal Process.* **40**: 2904–2920 (1992).

69. B. Nitsch, The partitioned exact frequency domain block NLMS algorithm, a mathematically exact version of the NLMS algorithm working in the frequency domain, *AEÜ Int. J. Electron. Commun.* **52**: 293–301 (1998).

70. B. Nitsch, Real-time implementation of the exact block NLMS algorithm for acoustic echo control in hands-free telephone systems, in S. L. Gay and J. Benesty, eds., *Acoustic Signal Processing for Telecommunication*, Kluwer, Boston, 2000.

71. B. Nitsch, A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain, *Signal Process.* **80**: 1733–1745 (2000).

72. A. O. Ogunfumni and A. M. Peterson, Fast direct implementation of block adaptive filters, *Proc. ICASSP-89, Glasgow*, UK, 1989, pp. 920–923.

73. P. Estermann and A. Kaelin, A hands-free phone system based on partitioned frequency domain adaptive echo canceller, *Proc. EUSIPCO-96*, Trieste, Italy, 1996, pp. 1131–1134.

ACOUSTIC MODEMS FOR UNDERWATER COMMUNICATION

KENNETH SCUSSEL
Benthos, Inc.
North Falmouth, Massachusetts

1. INTRODUCTION

Conventional in-air wireless communications typical rely on RF or electromagnetic means to convey information. Because such signals do not propagate well under water, sound waves, or acoustic signals, are the obvious choice for wireless underwater communication. This article discusses the development of modems for acoustic communications (acomms) and provides an overview of illustrative applications of such modems. The acomms channel is considerably more difficult than those encountered in typical terrestrial applications. The acoustic modem has to overcome several major obstacles in using the underwater channel: slow propagation, limited bandwidth, nonuniform propagation conditions, multipath, and low signal-to-noise ratio (SNR). The speed of sound in water is approximately 1500 m/s, while electromagnetic signals travel at nearly the speed of light. Thus, at practical distances (>2 km), the propagation delay for acoustic signals is measured in seconds compared to the nearly instantaneous propagation of RF signals. Another problem is that high-frequency sound waves are severely attenuated, so operation at any practical distance requires that the acoustic signal be less than approximately 30 kHz. At these frequencies, the current state of the art in transducer development limits the available bandwidth, which is a major obstacle to achieving high data rates. A third major obstacle is that the speed of sound in water varies with temperature and salinity. Thus, if an acoustic signal encounters a large temperature or salinity gradient, its path is bent or refracted, leaving "holes" in the channel where acoustic energy may be greatly reduced. This can result in a shadow zone where no communication is possible. The next major obstacle is multipath reflections from the seabed, surface, or other boundary. This can cause destructive interference resulting in frequency-dependent fading or intersymbol interference (ISI). Finally, acoustic modems may be required to operate at very low SNR caused by a combination of ambient noise and interference, limited transmitter source level, and transmission loss. At short ranges, the transmitted signal spreads spherically from its source. This transmission loss (TL) can be expressed as a function of range, $\mathrm{TL} = 20 * \log_{10}$ (range). In addition, there may be high levels of ambient noise present because of weather conditions at the sea surface and shipping noise. Despite these challenges, several companies have developed commercially available acoustic modems. There is a small but growing demand for wireless underwater telemetry in applications where cables are impractical or simply too expensive. These include command and control for autonomous undersea vehicles (AUVs), instrumentation around oil fields, any application in areas that are frequently fished by bottom trawlers, and deep-water instrumentation, to name a few. This article discusses the architecture of acoustic communication system, the hardware and software implementation of commercially available acoustic modems, and some examples of real-world applications.

2. ARCHITECTURE OF ACOUSTIC COMMUNICATION SYSTEMS

Acoustic modems generally segment an incoming stream of data or information bits into smaller blocks called *packets*, each of which is transmitted as an individual waveform over the physical channel. The acoustic communication system can be divided into multiple layers similar to the Open System Interconnection (OSI) reference model. The lowest layer is the physical layer, which, on the transmitter side, applies error correction coding, modulates the message into a passband waveform and passes it into the channel. On the receive side, the physical layer consists of those functions that acquire and align the waveform, demodulate the message, and finally decode the message and present it to the next layer. The higher levels are dependent on the application. For point-to-point communications, where only two modems are involved, the link makes use only of the protocol layer, which is similar to the media access control (MAC) layer in the OSI model. This layer handles any packet retransmissions and is responsible for presenting the data to the users. In the case of networks of acoustic modems, there is a MAC layer and a network layer. The implementation of the acoustic modem networks uses a communication protocol that contains many elements similar to the IEEE 802.11 protocol. This section focuses on the physical layer.

The first portion of a packet processed by the physical layer is the acquisition. The acquisition signal is prepended to the beginning of every packet and can

be any waveform that is used to detect the presence of the packet and to synchronize the receiver. This is the most important part of the receive processing, as without properly detecting and aligning the receiver to the start of a packet, it is impossible to receive any of the data that follow. Therefore it is important that this waveform be as robust as possible, to assure detection of the packet. The most robust signal is one that uses all the available bandwidth and is as long as practical. The temporal duration of the waveform is limited by additional receiver complexity associated with processing longer signals. In addition, acquisition-related portions of the waveform do not convey any data and thus are overhead, which reduces the actual data throughput. Therefore, the selection of the waveform is a tradeoff between reliability, overhead, and receiver complexity. A couple of examples of possible acquisition waveforms are a linear frequency-modulated (LFM) chirp, or a pseudorandom broadband signal. Both types of signals are processed with a replica correlator (an envelope-detected matched filter). The replica correlator transforms a T second waveform with an input SNR of S_{in} to a compressed peak with an output SNR of $S_{\text{out}} \sim 2TWS_{\text{in}}$, where W is the effective signal bandwidth. The effective duration of the compressed peak is approximately $1/W$ (second). Given a priori knowledge of the temporal offset from the edge of the chirp to the edge of the message portion, one judges the start of the received message by the same distance relative to the correlator peak location.

Following the acquisition are the modulated data. Modulation is a technique used to transmit information or data using an "elemental" signal occupying a specific portion of the time–frequency band. Historically, the elemental waveform is usually called a "chip." The (information) data are usually obtained as a binary string of 1s and 0s. Generally, modulation of digital data is accomplished by varying the amplitude, frequency, or phase (or a combination thereof) of a sinusoidal chip. Amplitude modulation is problematic in the ocean environment and requires a high SNR, so it is seldom used in underwater acoustic modems. Frequency-shift-keyed (FSK) modulation is the predominant method used in low-data-rate, noncoherent modems.

The simplest form of FSK techniques is binary FSK, or BFSK. BFSK uses two discrete frequency "slots" within an allocated time–frequency block, where a logic 1 is represented by the first frequency and a logic 0 is represented by the second frequency. By switching between the two frequencies, a stream of digital data can be sent. There are several variations of frequency modulation: (1) *multiple frequency shift keying* (MFSK), using multiple frequency slots within a block; and (2) *frequency hopping*, in which blocks are hopped about the available signal band, so that only a few (generally one) tones are transmitted at one baud interval. To achieve the densest mapping of frequency slots, the width of the slots should be equal to 1/(chip duration). This orthogonal signaling will prevent adjacent frequencies from interfering with each other and will make the maximum use of the available bandwidth. All the FSK methods can be received simply by measuring the signal energy and ignoring the phase of the signal, and thus are usually referred to as "noncoherent." Typically, the signal energy in each of the M slots is compared, and the slot with the most energy is selected. If there is broadband noise, it will affect each frequency slot equally and the correct decision will still be the frequency bin with the most energy. However, in a multipath environment the transmitted tone could be lost as a result of frequency-dependent fading, resulting in selection of the wrong slot, thereby causing multiple bit errors.

Another way to map the data to the available spectrum is to divide the entire band into N slots without regard to blocks. We map sequential clusters of 5 data bits into one group of 20 slots, which is drawn from $2^5 = 32$ possible combinations of 20 slots. These 20 slot codewords are derived from a family of Hadamard codes. Each codeword consists of 10 ones and 10 zeros, and has the property that each has a minimum distance of 10. The advantage is that if a tone is lost, the receiver employs a soft decision algorithm to pick the codeword that is the closest match to one of the 32 possible codewords. This method is referred to as *Hadamard MFSK* and is effective in both the presence of noise and multipath, at the expense of bandwidth efficiency. This method provides for coding gain, which means that the modem can operate at a lower SNR, with fewer errors, all at the expense of a lower data rate.

Among the noncoherent techniques, those that provide the most transmitted energy to a single tone, and those that can provide some immunity to frequency-dependent fading are generally more reliable in a given channel. Thus, for a given chip duration, frequency hopping will be more reliable at low SNRs than will the other techniques, with Hadamard MFSK performing midway between the other two. In all cases, this conclusion assumes that the electronic and channel spectral response is flat across the signal band. With substantial frequency-dependent attenuation, Hadamard signaling will be degraded more than the other techniques.

Yet another technique for message modulation is to vary the phase of the carrier. The simplest form is binary phase shift keying (BPSK), where a 1 is represented by one phase and a 0 is represented by a 180° phase shift. Information bits can be clustered and transformed into an "alphabet," permitting modulation that is more compact. For example, there are precisely four ways to combine 2 bits. We can therefore modulate the phase in 90° increments to reflect any of the four combinations. This is referred to as *quadrature PSK*, *QPSK*, or *4PSK*. The phase can be broken up into even more divisions. The process of receiving phase-shifted data is referred to as "coherent" processing. Coherent techniques require much more sophisticated processing as the acoustic channel can severely affect the phase of the received signal. The state-of-the-art receiver uses a decision feedback equalizer (DFE) to remove the effects of multipath (both amplitude and phase distortion), in an attempt to convert the received signal into one similar to what was transmitted. This method is usually successful, but requires relatively high SNR, and cannot tolerate very rapid variation in the channel multipath.

3. HARDWARE IMPLEMENTATION

Designing acoustic modems requires overcoming two challenges: (1) an intensive amount of processing is required to overcome the obstacles presented by the underwater acoustic channel and (2) most acoustic modems are battery-powered and thus must operate with a minimum of power. Advancements in low-power digital signal processors (DSPs) have made commercially available acoustic modems possible. Although there are several vendors producing commercially available acoustic modems Benthos (formally Datasonics) was one of the first and is the leading supplier of acoustic modems. This article discusses the hardware and signal processing software of a typical Benthos acoustic modem.

Figure 1 is a block diagram of a typical acoustic modem. The DSP is the main component of the modem and implements the signal processing software required to generate the transmit signals and the processing required to make sense of the received signals. A fixed-point DSP is used, since it provides the required processing power with less energy and memory demand than a floating-point DSP.

The DSP generates the transmit signal and sends it to the D/A converter. The analog output of the D/A is then put into a power amplifier, which boosts the level of the transmit signal to generate sufficient acoustic energy or source level at the transducer. The transmit signal level is adjustable, allowing power control to deliver sufficient but not excessive SNR at the receiver. Power control provides transmission security, power conservation, and improved multiple-access networking. The power amplifier drives the transmitter/receiver (T/R) network and matching network. The T/R network prevents the sensitive receiver from being damaged by the large transmitted waveform, and the matching network is used to match the output impedance of the power amplifier to the transducer's impedance. The transducer is a piezoelectric ceramic, which converts the electrical transmit signal to an acoustic signal.

Received acoustic signals are generally very small and pass through the T/R network to the preamplifier. The output of the preamplifier goes into either the receiver or a wideband amp, depending on the mode of the DSP. The DSP operates in either a low-power mode or an active mode. In the low-power mode, the DSP runs off the slow clock and all functions are shut down except the wakeup logic. In this mode, the DSP is capable of processing only coded wakeup signals. This allows the modem to operate in a standby mode with a very low current drain. In active mode, the DSP runs off the fast clock, allowing all incoming acoustic signals to be processed. The received signal can have a large dynamic range. Thus, automatic gain control (AGC) is used to maintain a nearly constant signal level at the input of the A/D. The AGC is controlled by the DSP, which measures the signal level at the A/D. Large signals are attenuated and gain is applied

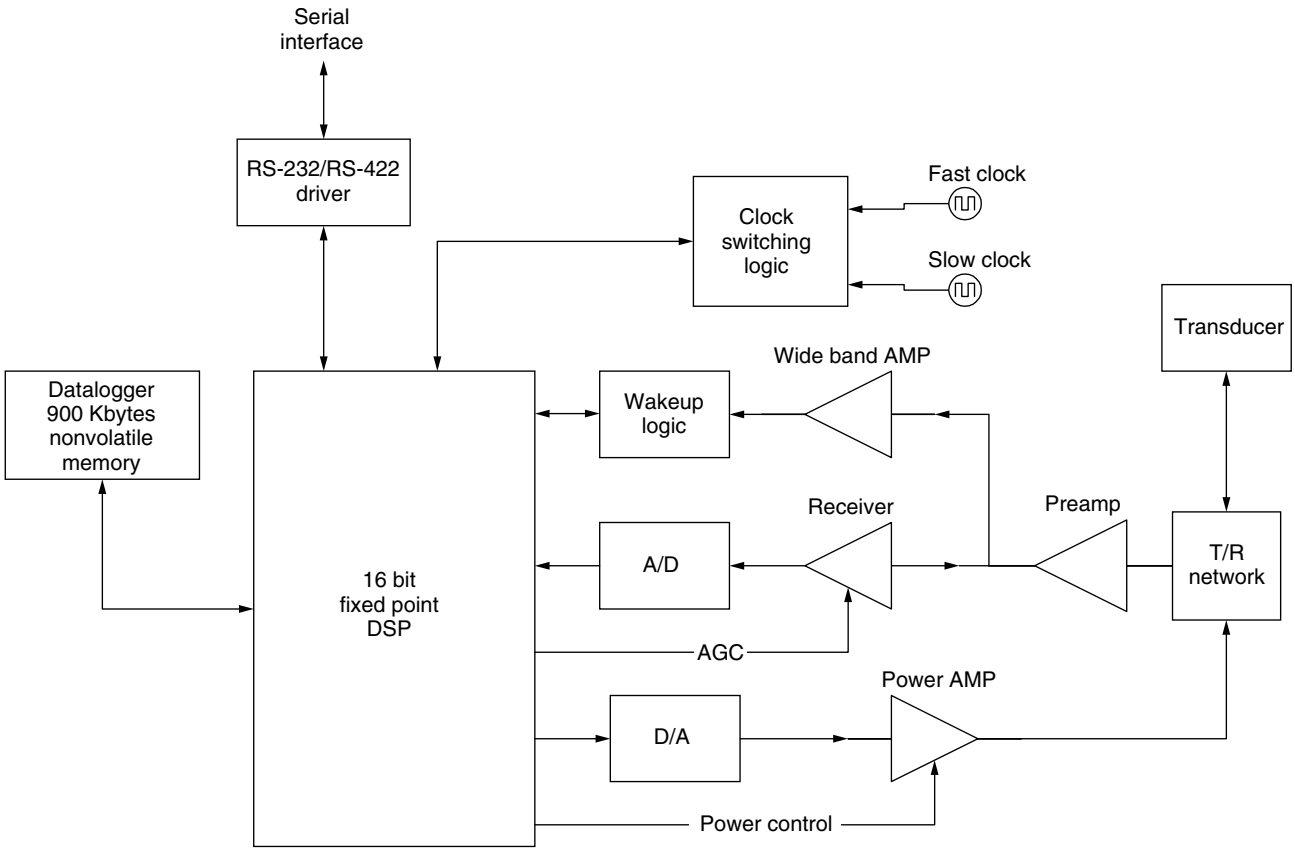


Figure 1. Top-level block diagram of a typical Benthos acoustic modem.

to small signals. The sampling rate of the A/D is set to greatly oversample incoming signals to reduce the need for expensive antialiasing filters.

Other peripherals include a serial interface and a datalogging capability. The serial interface allows control of the modem by a host PC or other processor through a standard RS-232 or RS-422 interface. RS-232 allows connection to most PCs and instruments at rates of $\leq$9600 baud. RS-422 allows the modem electronics to be placed closer to the transducer, allowing the host PC or instrument to be several kilometers from the electronics. This provides an alternative to noisy, lossy analog connections to the transducer via long cables. A datalogging capability using 900 kB (kilobytes) of nonvolatile memory is available for buffering and storage of incoming data. In many applications, it is desirable to store data for some time before acoustic transmission.

The modem electronics can be packaged in a variety of configurations. The simplest configuration is the modem board set shown in Fig. 2. In this configuration, the printed-circuit boards are mounted to a chassis, and are externally powered. Usually this configuration is used by original equipment manufacture (OEM) applications, where the modem is integrated with the instrumentation of another manufacturer. Another configuration is to package the modem with batteries in a self-contained pressure housing, for deployment underwater. The required water depth or pressure determines the type of housing. Figure 2 shows an ATM-885 acoustic modem packaged in a hardcoat anodized aluminum housing with a maximum depth rating of 2000 m. Note that the housing contains a connector for external power and the serial interface for connection to host equipment. For shipboard operation the electronics can also be installed in an AC-powered shipboard deckbox or 19-in. rack. The shipboard equipment makes use of an over the side or remote transducer. Figure 2 shows photographs of the AC-powered shipboard deckbox and a remote transducer. The final option is for the modem electronics to be packaged in a buoy containing a RF modem and a GPS receiver. The modem's transducer hangs below the buoy, and any data received by the modem are relayed via the RF link to a shore station, and vice versa.

4. SIGNAL PROCESSING IMPLEMENTATION

The transmit signal processing is as is shown in Fig. 3. The data or information bits are first convolutionally encoded for error correction. The output of the data encoder is mapped to MFSK frequency tones. Tones for Doppler tracking are also added. The phase of the frequency-domain signal is randomized to avoid large peak : average power ratio signals at the output. The spectrum is inverse Fast Fourier–transformed to obtain a time-domain baseband signal sampled at 10,240 Hz. The baseband signal is then interpolated to 163 kHz rate and quadrature-mixed to a passband carrier prior to digital-to-analog conversion.

The receiver signal processing is shown in Figure 4. Acoustic data sampled at 163 kHz are obtained from the A/D converter and resampled at a slightly different sample rate depending on the Doppler shift present in the communication channel. Automatic gain control detects the signal level and adjusts it as necessary via external hardware. A quadrature mixer converts the signal to complex baseband. The baseband signal is decimated to a 10,240-Hz sample rate. During the acquisition period, matched-filter processing is performed to look for the acquisition signal. Once the acquisition signal is detected, it is used to synchronize to the incoming data. Adjustments

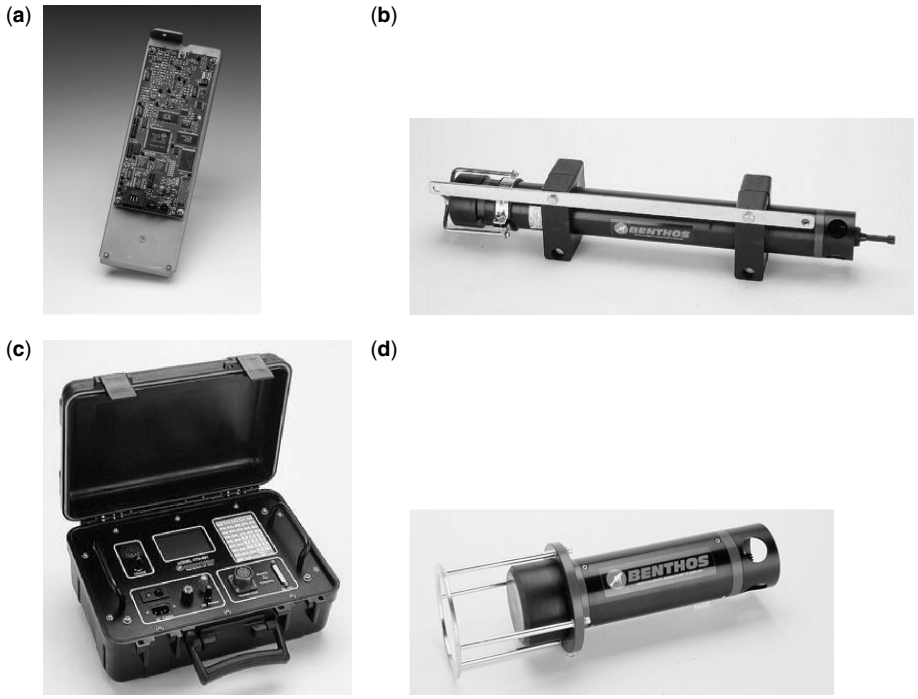


Figure 2. ATM-88x modem components: (**a**) OEM board set; (**b**) ATM-885; (**c**) ATM-881 deckbox; (**d**) remote transducer.

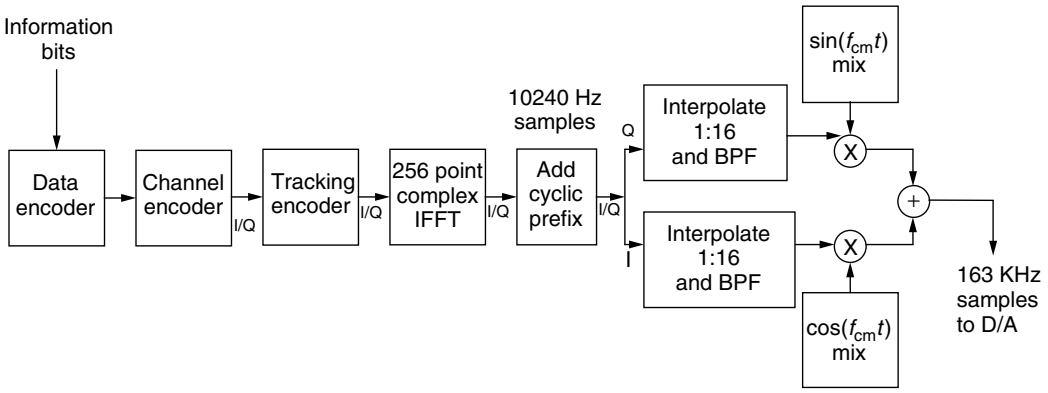


Figure 3. Transmit signal processing.

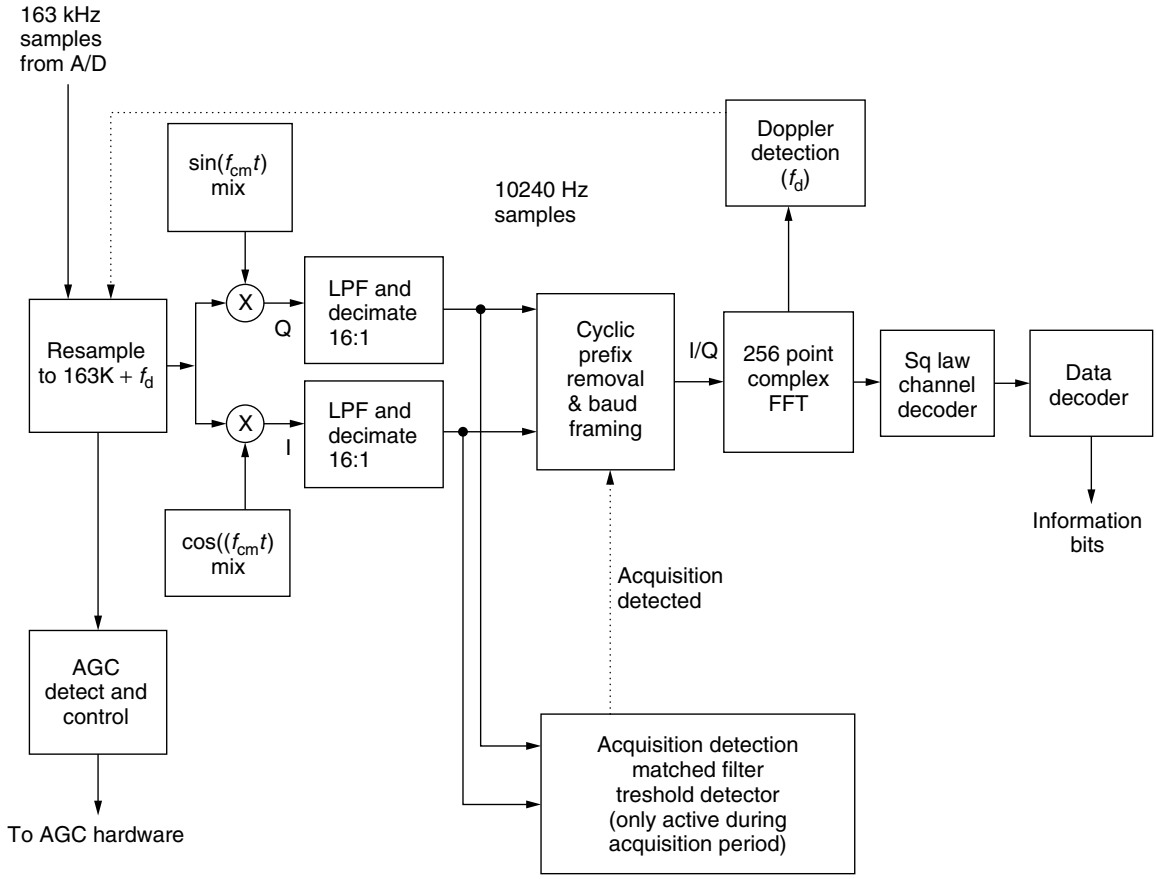


Figure 4. Receive signal processing.

are made for timing errors and to eliminate the prepended signal used to guard against multipath. A complex 256-point FFT converts the signal to the frequency domain, which is used for MFSK decoding and Doppler tracking. If a Doppler shift is detected, the next incoming samples are resampled to adjust for the shift. The frequency-domain data are then run through a square-law channel decoder to obtain the magnitude of the signal. This decodes the MFSK data. Finally, a Viterbi decoder interprets the convolutionally encoded data. The final information bits are sent to the user.

5. MODEM APPLICATIONS

The are numerous diverse applications for underwater acoustic modems. A few sample applications are described below.

5.1. Cone Penetrometer

In one modem application, wireless control and real-time data recovery are effected using a deep-water seabed cone penetrometer (CPT). Guardline Surveys, located in Great Yarmouth, England, uses the CPT for penetrating the

seabed in water depths of ≤2000 meters. In operation, the instrument is lowered over the side from a ship and to the seabed. The instrument is fitted with pressure and temperature sensors as well as inclinometers to assure proper attitude and stability. In the past, the CPT used an expensive electromechanical cable both to lower the instrument and for communication purposes. Guardline removed the constraints imposed by the electrical umbilical cable, replacing it with Benthos acoustic modems. With the modems the operator can communicate with the CPT all the way down during deployment to the seabed, sending commands, and receiving status information. During penetration, data from the sensors are sent to the operator in real time. With real-time remote recovery of data, the CPT can be lifted just off the bottom and maneuvered to another nearby site. Figure 5 is an illustration of the CPT in operation.

5.2. Pipeline Bending

Another acoustic modem application is the remote acquisition of pipeline bending stresses and vortex-induced vibration (VIV) from an offshore oil–gas platform. A monitoring project was established by the Petrobras R&D center to collect data vibrations and tensions on the mooring lines and risers. Petrobras contracted with Scientific Marine Services (SMS) to provide the instrumentation. It was determined that cables connecting the subsea instrumentation to the surface would have a minimum survival probability during the difficult pipe laying operations; therefore acoustic communication was selected. Benthos acoustic modems provide both the downlink command and control signaling and uplink data acquisition. The acoustic modems provide the means for controlling the VIV bottle synchronization, data uplink repetition rate, as well as other operating parameters. The data rate used is 600 bps (bits per second) at a range of 1300 m. Figure 6 is an illustration of the deployment.

5.3. Imaging and Telemetry

Command and control of an autonomous underwater vehicle is a growing application for acoustic modems. During one experiment,[1] a robotic crawler[2] carrying an acoustic modem, camera, and a digital signal processing unit was used to search autonomously for an underwater object. When the object was found, the robot informed the user using acomms that "something" had been found. The robot was then acoustically commanded to take a still-frame picture, compress it, and transmit using the acoustic modem. The grayscale images shown in Fig. 7 each consist

[1] The U.S. Navy's "AUVfest 2001" held off of Gulfport, Mississippi in October 2001.
[2] Developed by Foster-Miller Inc. for the U.S. Navy's Coastal Systems Station (CSS).

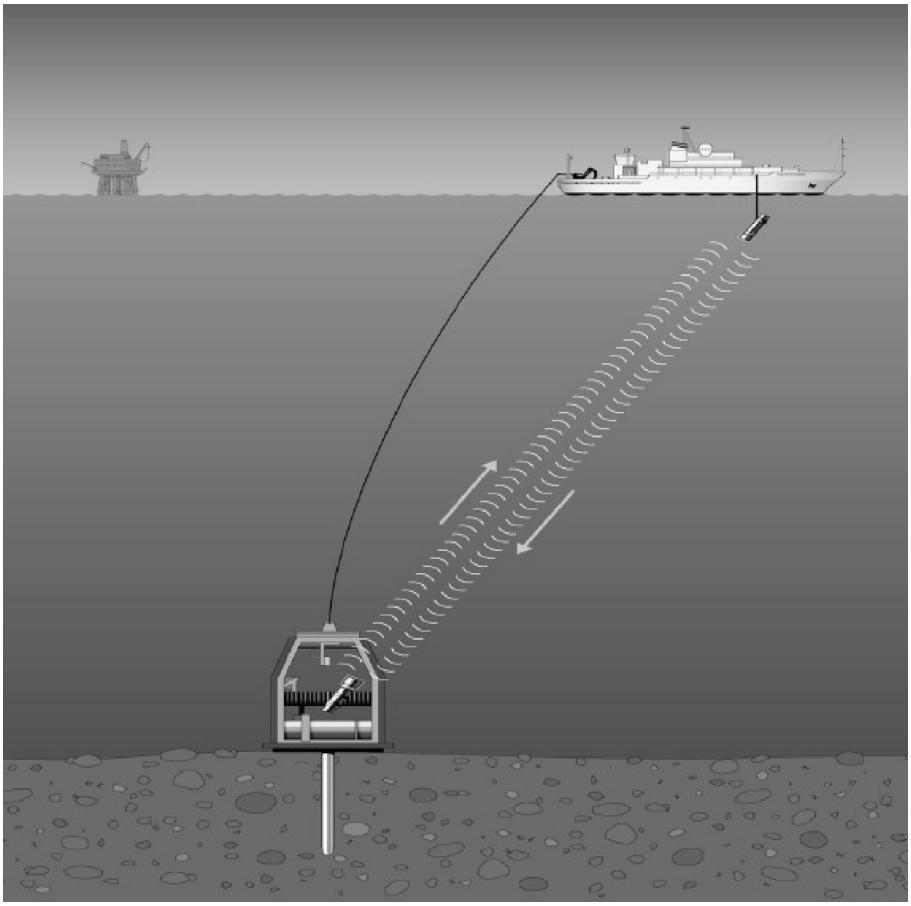


Figure 5. Illustration of penetrometer operation incorporating the telesonar acoustic modems for transmission of real-time command, control, and penetrometer data.

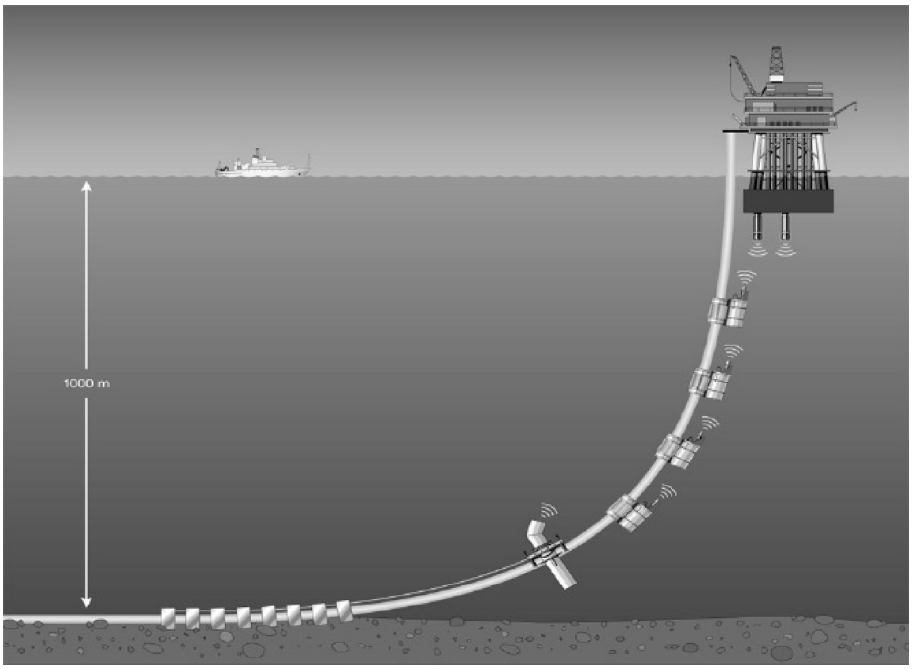


Figure 6. Illustration of instrumentation deployment scenario incorporating the Telesonar Acoustic Modems.

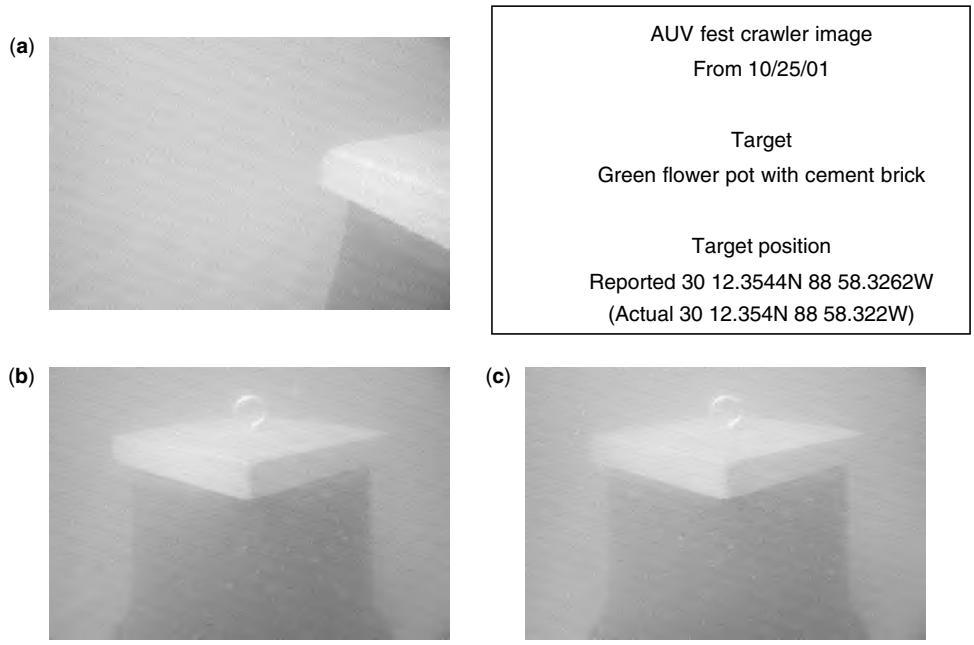


Figure 7. Automated detection, commanded acomms telemetry of compressed images: (**a**) acquisition image from raster pattern, acomm 600 bps, 50/1 compression; (**b**,**c**) ID image after vehicle reposition, acomm 600 bps, 50/1 compression (b) and acomm 1200 bps, 50/1 compression (c).

of 100 kB. With the 50 : 1 compression ratio used for this experiment, the transmitted image was only 2 kB, which, at 1200 bps transmission rate, required only 13 s to transmit. We note that the revolutionary wavelet-based compression technology used here[3] easily supports 200 : 1 and higher compression for such images.

[3] Developed by Professor Truong Nguyen of the University of California at San Diego.

We observe that, using coherent communications at a nominal bit rate of 4000 bps, and an image compression ratio of 200 : 1, we can maintain a real-time rate of one frame every 1.5 s. Thus, "slow scan" video is distinctly possible in those channels that support high-rate acomms.

BIOGRAPHY

Kenneth Scussel received a B.S. degree in Electrical Engineering in 1988 from the University of Connecticut,

and an M.S. degree in Electrical Engineering from Rensselaer Polytechnic Institute in 1992. In 1988 he joined General Dynamics, Electric Boat Division, where he developed embedded software for submarine launched cruise missile and torpedo tube control console simulators. Following General Dynamics, Mr. Scussel joined Datamarine International in 1993. At Datamarine he was a Project Engineer in charge of the development of consumer marine electronics. His accomplishments include putting a system that integrated depth, wind, boatspeed, and navigation sensors into production. This work included the development of depth sounder algorithms, which dramatically improved the performance of the depth sensor. Since 1995, he has been with Benthos (formerly Datasonics) as a Staff Software Engineer responsible for all aspects of software development in the acoustic modem product line. Mr. Scussel was part of the team that developed the original Datasonics ATM-87X series of acoustic modems. He has made significant enhancements to the signal processing, and enhanced the networking protocol algorithms that lead to the release of the ATM-88X series, a new generation of acoustic modems. His areas of interest are digital signal processing and developing software for real-time embedded processors.

FURTHER READING

R. Edwards, SMS deploys first successful acoustically coupled VIV measurement system in Campos Basin, published in the *Marine Analyst*, OTC-2000. SMS newsletter article available on their website www.SCIMAR.com.

M. Green and J. Rice, Channel-tolerant FH-MFSK acoustic signaling for undersea communications and networks, *IEEE J. Ocean. Eng.* **25**(1): 28–39 (Jan. 2000).

J. Preisig, Underwater acoustic communications, *IEEE Signal Process. Mag.* (July 1998).

J. Proakis, *Digital Communications*, McGraw-Hill, 1989.

K. Scussel, J. Rice, and S. Merriam, (1997). A new MFSK acoustic modem for operation in adverse underwater channels, *Oceans'97 MTS/IEEE Conf. Proc.*, 1997, Vol. 1, pp. 247–254.

R. Urick, *Principles of Underwater Sound*, McGraw-Hill, 1983.

ACOUSTIC TELEMETRY

FLETCHER A. BLACKMON
Naval Undersea Warfare Center
Division Newport
Newport, Rhode Island

1. INTRODUCTION

Humankind has always felt the need to communicate. It is a basic human desire that began most primitively as oral tradition and cave writings. The forms of communication were made more elaborate in the writings and hieroglyphics of ancient cultures such as the Egyptians and the Greeks. More modern cultures have refined the art of communication through language and pictures. The focus then became one of how to communicate over exceedingly larger distances. This human desire for the transfer of the printed word and audio and visual effects led to the creation of the telegraph, the U.S. Postal Service, and the telephone, radio, and television. More recently, the desire for information — the reason one communicates — has extended to wireless forms of communication using mobile and cellular phone technology, satellite communications, and communications from deep-space probes. It is no wonder that electronic mail (email) and the Internet, which currently provide a worldwide communications/information network, has literally taken over the world by involving everyone in the dialog of humankind. It is therefore natural and a vital step in this dialog to communicate and transfer information into, within, and out of the underwater environment, which covers more than 70% of the earth's surface.

2. BACKGROUND

The underwater environment has provided and is still providing one of the most interesting and challenging mediums for communication. These challenges include limited bandwidth, multipath induced time and spectral dispersion, and channel time variability. The available bandwidth is limited because of the frequency-dependent absorption characteristics of the underwater environment since higher frequencies are attenuated more strongly than lower frequencies and also as a function of range from the transmitter to the receiver. Another consideration relating to bandwidth is the receive signal strength, which decreases as a function of range as well as the noise present at the receiver due to ambient and human-made (synthetic) noise components. Bandwidth, signal transmission loss due to spreading and absorption, and noise are parameters that are used to determine the signal-to-noise ratio (SNR) at the receiver. Complicating this underwater picture is the presence of unwanted multipath in addition to the desired direct path signal. *Multipath* can be defined as one or more delayed signal replicas arriving at the receiver with time-varying amplitude and phase characteristics. These multipaths are responsible for temporal spreading and spectral spreading of the transmitted signal. These distorted signal replicas are produced by boundary reflection from the surface, bottom, and other objects as well as from acoustic ray bending in response to sound speed variation in the underwater environment. Add to these complications the time variability of these phenomena and a very interesting medium for acoustic telemetry results.

One of the first acoustic communication systems that was employed was the underwater telephone or UQC. This is a low-bandwidth (8–11 kHz) voice link that was developed by the United States government in 1945. It was used to communicate to and from submarines at speed and depth over a range of several kilometers. This method of acoustic communication is still today the standard for submarines and surface vessels. However,

modern technology with the advent of miniaturized, low power, digital signal processor (DSP) electronics and portable personal computers that can implement and support complex signal processing/communications algorithms has provided the capability to improve the quality, increase the data throughput, and increase the number of military and commercial telemetry applications that are possible.

Next, a brief telemetry system overview will be presented as a framework for the discussion to follow, which will focus on commercially available acoustic telemetry modems, past and present acoustic telemetry applications, the navy's range-based telemetry modems, several specific range-based telemetry applications, and finally current and future research in underwater acoustic telemetry.

3. TELEMETRY SYSTEM OVERVIEW

Telemetry data can take many different forms. Current speed data, sound velocity speed data, salinity data, pressure data, accelerometer data, temperature data, system health data, instrument command data, videos and images, data bearing files, digital voice, and interactive text are all clear examples of the various types of data that a telemetry system can be used to convey from one point to another point. In the underwater acoustic telemetry case, the data are telemetered from one point underwater to another point underwater that may be more accessible to those interested in the data. A telemetry system is comprised of a transmitter and a receiver. A typical transmission format is shown in Fig. 1, and a typical transmitter block diagram is shown in Fig. 2. The first portion of the transmission is a synchronization signal that is used to detect the presence of a telemetry signal and the location of data within the telemetry stream. The synchronization signal may be a large time bandwidth "chirp," or a differential or binary phase-shift keyed (DPSK or BPSK) signal with good auto- and cross-correlation properties. Guard time intervals are used after the synchronization signal and following the data prior to the next synchronization signal to mitigate the acoustic channel's multipath effects. A short training sequence follows the first guard time interval and is used to train the receiver's adaptive equalizer, specifically, the channel compensation capability for the modulated data that are to follow. Alternatively, the data may be transmitted by frequency shift-keyed (FSK) or multifrequency shift-keyed (MFSK) or spread spectrum modulation waveforms. A Doppler tracking tone may be superimposed over the telemetry packet for Doppler compensation at the receiver side of the link or may be derived from the synchronization signal itself. An optional channel probe in some telemetry systems can be used periodically prior to sending the telemetry packet or in an integral fashion to measure the channel characteristics so that the receiving system may take advantage of it. The transmitter block diagram in Fig. 2 shows the generation of the synchronization signal, coding and interleaving of the binary data, and the modulation of the binary data into transmission symbols as well as the power amplifier and electrical to acoustic transducer.

The acoustic telemetry receiver block diagram is shown in Fig. 3. The task of the receiver is to mitigate, compensate, and/or undo the effects of the underwater channel on the transmitted telemetry signal. The receiver incorporates acoustic to electric conversion via hydrophone, filtering, amplification, detection, synchronization, and Doppler compensation. The filtered, amplified, synchronized, and Doppler-corrected signal is then presented to a demodulator to provide baseband processing. The baseband signal is then passed to a symbol detector that in the case of coherent signals takes the form of an adaptive equalizer or a bank of filters for frequency-based incoherent systems or a set of correlators for direct-sequence spread-spectrum (DSSS)-based signals. Following the bit or symbol detection process, a decoding step can be performed if the data were coded at the transmitter. This decoding step seeks to correct errors that may still be present after the symbol/bit detection stage.

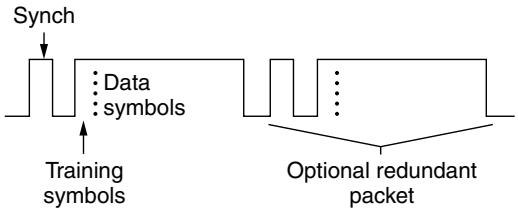


Figure 1. Typical telemetry transmit packet.

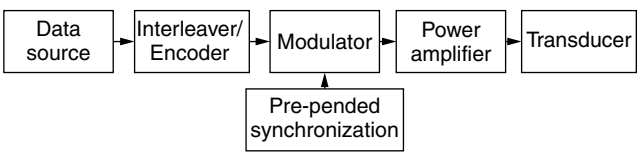


Figure 2. Acoustic telemetry transmitter block diagram.

4. ACOUSTIC TELEMETRY MODEMS

There are a number of commercially available acoustic telemetry modems at present. These modems span the gamut of communications techniques such as FSK, MFSK, spread-spectrum, and coherent signaling schemes such as BPSK and QPSK. In addition, these modems provide varying capabilities, chief among these are the bandwidth and data rate. Table 1 shows a brief comparison of these telemetry modems and their salient features.

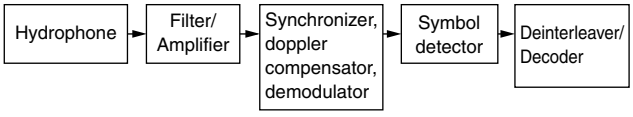


Figure 3. Acoustic telemetry receiver block diagram.

Table 1. Commercially Available Telemetry Modem Systems

Modem Manufacturer	Application	Modulation Format	Data Rate (bps)	Link Type
Benthos	Oil, environmental, military	FSK, HFSK, MFSK, BPSK, QPSK, TCM8PSK	10–10,000	Half -duplex
Linkquest	Oil, military	Spread-spectrum	200–2000	Half-duplex

5. TELEMETRY APPLICATIONS

A number of telemetry applications can be found in the fairly recent history of acoustic telemetry that chiefly began late in the 1980s after feasibility research was conducted in the area of incoherent FSK systems in the early 1980s [1]. The DATS system [1] was capable of operating at data rates up to 1200 bits per second (bps). As early as 1984, the concept of using acoustic telemetry to monitor and control oil wellheads was being tested and used in 100-m water depths and at horizontal ranges of 2 nautical miles [2]. Oil monitoring and control since this time and at present has been employing acoustic telemetry to reduce maintenance and operations costs. In 1989, it was demonstrated that a vertical acoustic telemetry link could be established to an undersea robot that was used to replace divers in the maintenance of submerged platforms [3]. In 1991, the Woods Hole Oceanographic Institute (WHOI) conducted a long-term acoustic telemetry experiment using their utility acoustic modem (UAM) to collect sensor data during a 6-month moored deployment [4]. This telemetry system, still used today, is based on TMS30C44 DSP technology and represents a versatile, configurable device for autonomous and moored telemetry applications. Again in 1991, acoustic data links employing telemetry were being explored for unmanned undersea vehicle (UUV) applications [5] using a 1250-bps data link with a range of 2 nautical miles using a bandwidth of 10 kHz. In 1992, a submersible was used to telemeter high-quality video imagery from a deep ocean trench to a surface vessel [6]. Throughout the 1990s, acoustic telemetry technology employed noncoherent modems. One of the more advanced of these systems, called the Telesonar system developed by Datasonics, Inc., and the U.S. Naval Command, Control and Ocean Surveillance Center, employed MFSK techniques using Hadamard coding as well as convolutional coding and frequency-hopping patterns [7]. This system is currently being used for a variety of applications and is commercially available from Benthos Inc.

First attempts were made in 1989 using phase-coherent signaling and minimal equalization techniques to telemeter images and commands between a surface vessel and a subsea robot in a very benign channel. Quadrature amplitude modulation (QAM) signaling was used with a transmission rate of 500 kbps with a bandwidth of 125 kHz centered at 1 MHz [8]. New and innovative ground breaking research in coherent acoustic communication techniques that allowed for tracking channel time variability in the early 1990s [9,10] made it possible to robustly communicate successfully in horizontal as well as vertical channels with higher data rates than were previously possible through the use of bandwidth-efficient MPSK and MQAM modulation schemes. In 1993, long-range, low-error-rate ($<10^{-4}$) telemetry over horizontal distances in excess of 200 km was shown to be possible [9]. In 1994, a prototype digital, acoustic underwater phone was described and demonstrated that compressed the data prior to transmission [11]. This type of system has not received much attention but still represents an important and much needed upgrade to the much older UQC voice system used aboard surface and subsurface vessels. In 1998, a practical coherent telemetry system for use onboard an autonomous underwater vehicle (AUV) was demonstrated using a bandwidth of 3 kHz and a bandwidth of 25 kHz with data rates of 2500 and 10,000 bps, respectively, in shallow water depths of 10–30 m [12]. Also in 1998, the WHOI UAM telemetry system was deployed on the MIT Odyssey and the Florida Atlantic University (FAU) Ocean Explorer UUVs [13]. The Odyssey UUV was integrated as part of an autonomous ocean sampling network (AOSN). Sensor data were transmitted from the vehicle to another telemetry modem. The data were then sent up a mooring cable to a surface buoy and then relayed via RF link to a support vessel for real time monitoring. In 2000, a surf-zone acoustic telemetry experiment (SZATE) was conducted alongside the pier at the Scripps Institution of Oceanography to demonstrate the ability to coherently telemeter data in the challenging very-shallow-water surf-zone environment [14]. It is envisioned that commercial and military applications of small size and/or miniature lemmings or crawlers with video and other sensors will be used in the surf zone in the near future and will have a need to telemeter this data to one or more remote undersea sites.

One of the most advanced acoustic communication/telemetry systems has been developed as part of the Acoustic Communications Advanced Technology Demonstration (ACOMMS ATD) funded by the U.S. Navy Advanced System Technology Office (ASTO). This tactical system can employ noncoherent as well as coherent modulation/demodulation techniques with multiple array sensor inputs and has been used for a multitude of naval demonstrations and applications involving transmission of voice, text, and video between UUVs, UUV and surface vessels, UUV and submarine, surface vessel and submarine, two submarines, and surface buoys. These telemetry links have been established at various data rates of $\leq$20 kbps in some cases and has operated within a

number of low-, medium-, and high-frequency bands over ranges of 2 km in shallow water, 3.7–5.6 km at high frequency, and 37–124 km at medium frequency [15] in deep water.

More recently, there has been growing interest in the application of acoustic telemetry to remote undersea networks given the success of point-to-point telemetry links. One such example of an undersea network is the Autonomous Oceanographic Surveillance Network (AOSN), which has been developed through funds from the Office of Naval Research (ONR) to network surface buoys and autonomous AUVs in order to sample the underwater environment [16]. The U.S. Navy is also developing sensor networks that employ acoustic telemetry with power control [17], protocol layers [18], and the ability for the sensor nodes to adaptively learn the network parameters such as node numbers, link quality, and connectivity tables [19]. This particular network has been demonstrated showing its surveillance capability as well as RF and satellite gateway capability to offload collected sensor data. A pictoral representation of an undersea acoustic telemetry network is shown in Fig. 4. Typically, these systems employ one or more of the following schemes for multiuser access: time-division multiple access (TDMA), frequency-division multiple access (FDMA), and code-division multiple access (CDMA) with a higher-level protocol layer that frequently uses ARQ and handshaking methodologies.

An excellent tutorial on acoustic communications presenting the history as well as an eye to the future has been presented in a review article [20]. Another excellent and also very readable acoustic underwater communications tutorial has been presented by Milica Stojanovic in this very same edition of the Wiley Encyclopedia of Telecommunications for the interested reader.

Next, we will take a more detailed look into a specific NUWC range-based modem telemetry system. Following this telemetry modem discussion, a number of range-based telemetry applications will be presented in detail as specific examples.

6. NUWC RANGE-BASED MODEM

The NUWC range-based telemetry system is a set of underwater acoustic telemetry modems developed by the Engineering, Test and Evaluation Department at the Naval Undersea Warfare Center Division, Newport (NUWCDIVNPT). The modems were developed as part of the Submarine Underwater Telemetry project that was tasked to provide robust bidirectional, full-duplex underwater acoustic communication subsystems to undersea test and evaluation and training ranges now under development. The goal was to produce acoustic modems that can reliably communicate with subsurface range participants at a throughput data rate of approximately 1 kbps at ranges out to 2 nautical miles in shallow and deep water while maintaining vehicle track.

The fact that navy ranges are designed for tracking exercise participants and not specifically for telemetry placed numerous constraints on the modem's design. The system's bandwidth and center frequencies were constrained, as was the choice of transducers to be used. The receiving hydrophones available on a range are typically widely spaced and omnidirectional, and available transmitters are also often omnidirectional. The benefits of beamforming, spatial diversity combining, and other directive techniques developed by researchers [23–25] in both the United Kingdom and the United States are not practical options under these conditions. Instead, the range-based modems must rely on adaptive equalization, paired with time redundancy and error correction schemes to achieve robust underwater communication. The cost of these techniques is reduced data rate. Although the system transmits at 4000 bps, the maximum sustained throughput data rates are approximately 900 bps with half-rate coding and 1800 bps without coding. The performance of this telemetry system has been documented in the literature [21].

Physically, each modem is a VME chassis with a notebook computer which serves as a user interface. The VME chassis is controlled by a FORCE-5CE CPU running SUN OS 4.1.3-u1. The chassis also contains a hard-disk drive, a VME clock and a timing board (VCAT) that serves as a master external clock, an optional GPS timing board for synchronizing transmissions with a GPS time reference, and two octal TMS320C40 digital signal processor boards as shown in Fig. 5.

Each modem has a transmitter and a receiver. The transmitter resides on two TMS320C40 processors. It reads binary data from a buffer on the UNIX host, then

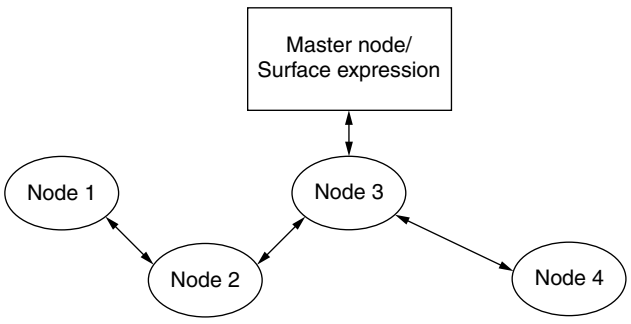


Figure 4. Underwater acoustic telemetry network diagram.

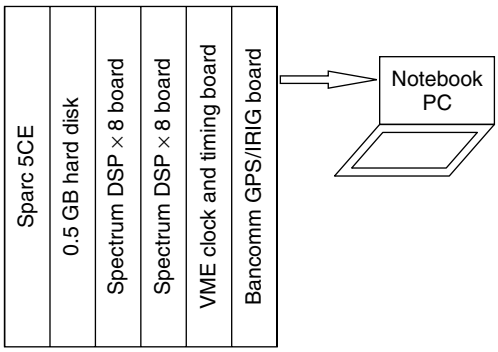


Figure 5. Block diagram of a NUWC range-based VME telemetry modem chassis.

packetizes and (optionally) convolutionally encodes the data. The transmission waveform is digitally synthesized, digital to analog converted and broadcast through the water. Figure 1 shows the format for the transmission. Each one second data packet is sent redundantly in a time diversity scheme to improve the robustness of the receiver.

The receiver for the system is a jointly adaptive decision feedback equalizer with a digital phase-locked loop (DFE-DPLL). The DFE-DPLL is patterned after a receiver first proposed by Stojanovic, Catipovic, and Proakis [24] with modifications to allow for efficient real-time implementation [25]. The receiver is implemented using eight TMS320C40 digital signal processors and consists of four functional blocks: packet detection and synchronization, Doppler estimation and compensation, complex demodulation, and equalization. Viterbi decoding (if required) is performed on the UNIX host.

The pairs of redundant data packets are jointly equalized in a manner similar to the spatial diversity combining technique presented in Ref. 24. Spatial diversity combining assumes that multiple spatially separated sensors are available to receive a given telemetry packet, and that the transmission travels through independent paths (or channels) to arrive at each sensor. The output of the sensors can then be jointly equalized to recover more of the signal than a single sensor. Since these telemetry modems cannot rely on the availability of multiple sensors, it employs multiple, time-separated transmissions of the same signal packets (i.e., time diversity). The ocean is a highly nonstationary, time-varying environment. Therefore, two transmissions of the same data packet, spaced sufficiently in time, travel through independent channels to arrive at the single sensor. Once the redundant data packets have arrived at the sensor, the net effect is essentially equivalent to spatial diversity. The adaptive equalizer algorithm has multiple inputs with each input containing the same signal information as received across an independent path. The cost of time diversity is a reduction in the throughput data rate, but diversity is often essential for low-error-rate telemetry.

7. RANGE-BASED TELEMETRY APPLICATIONS

A number of range-based acoustic telemetry applications will now be discussed. These include the mobile deep-range (MDR) application, the synthetic environment tactical integration virtual torpedo program (SETI VTP) application, and the underwater range data communication (URDC) application.

7.1. MDR

In November 1997, the NUWC range-based modems were used by personnel from the Naval Surface Warfare Center (NSWC), Detachment Annapolis to acoustically transfer data files collected during their mobile deep-range (MDR) exercises from a submerged submarine to a moored surface vessel as shown in Fig. 6. The MDR trial represents one of the initial uses of this modem system. One modem chassis with its notebook PC interface was set up in a data analysis station aboard the moored ship.

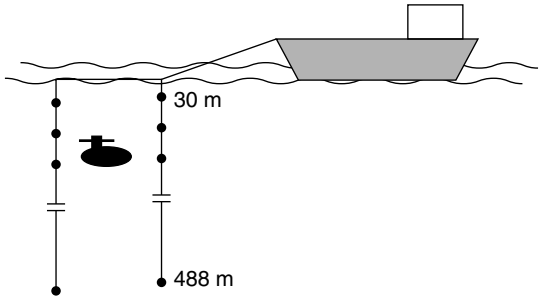


Figure 6. Configuration of the mobile deep-range telemetry.

At the analysis station, there was access to the outputs of 18 omnidirectional hydrophones contained in the two vertical legs of the MDR array shown in Fig. 6. A number of ASCII data files containing position and test configuration data were transmitted from the submarine under test and were received on the vertical hydrophone arrays. The acoustic telemetry modem receiver was successful in decoding the data files with few or no errors. These test configuration data files were then used by analysts aboard the measurement vessel to reconstruct and better analyze the data recorded with the arrays' acoustic and electromagnetic sensors in near real time.

7.2. SETI VTP

The submarine virtual torpedo program is the initial implementation of the synthetic environment tactical integration (SETI) project. The SETI project promotes the use of advanced distributed simulation (ADS) capabilities by creating high-fidelity antisubmarine warfare (ASW) training opportunities using live targets, synthetic torpedoes, and onboard submarine tactical and training systems in realistic external environments. The goal of the VTP project is to enable the real-time interaction of live submarines with high-fidelity simulated torpedoes. The SETI VTP conceptual drawings are shown in Figs. 7 and 8. In June 1998, the SETI VTP test conducted at the Atlantic Undersea Test and Evaluation Center (AUTEC) demonstrated the full-duplex capabilities of the NUWC range-based modems. The demonstrated capabilities include (1) encrypted, bidirectional, full-duplex data exchange between a submerged submarine operating on an instrumented navy range and remote modeling and simulation facilities via a wide-area network (WAN), which includes the range-based modems and a satellite link; (2) submarine launch control of simulated torpedoes from the remote modeling and simulation facilities and reception of tactical weapon data from the launched weapon; and (3) use of two alternate communication methods, distributed interactive simulation (DIS) protocols and high level architecture (HLA) runtime infrastructure (RTI) to transfer both weapon and positional data between the submarine and the remote modeling and simulation facilities.

7.3. URDC

The underwater range data communications (URDC) project is a currently ongoing U.S. Navy project that

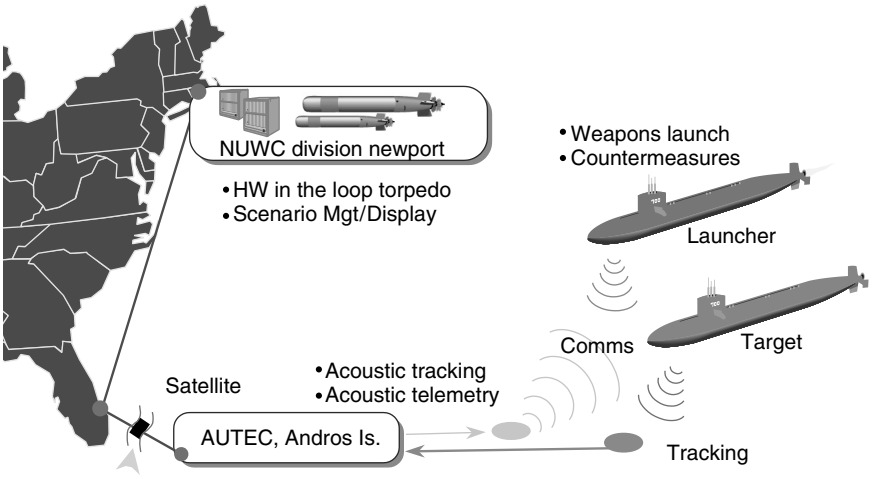


Figure 7. Illustration of the SETI VTP telemetry link concept.

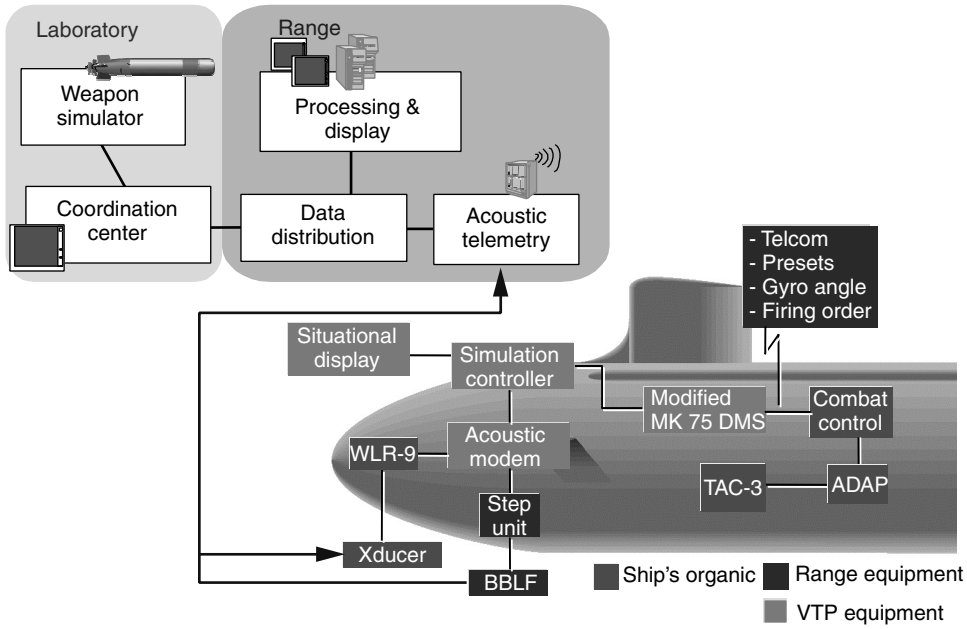


Figure 8. Ship/shore SETI VTP telemetry configuration.

is funded to provide acoustic telemetry capability on shore at the Command and Control (CC) building at AUTEC as well as a roll on/roll off utility on submarines engaged in Test and Evaluation exercises on range. This system will expand and modernize the older NUWC range-based telemetry modems discussed earlier. The URDC system will interface to range receive and transmit node infrastructure on shore as well as organic submarine sensors. In addition, this system will integrate submarine tracking capability with telemetry capability. This new telemetry system will integrate state-of-the-art SHARC-based DSP technology with other telemetry components in a small user-friendly package for navy sailor and range use. The system will incorporate multiple telemetry modes of operation including high-, medium-, and low-data-rate transfer in half-duplex and full-duplex FDMA modes. These modes will employ coherent signaling techniques coupled with strong error correction techniques and selective time diversity automatic repeat request (ARQ) capability. Initial prototype performance results have been reported previously in the literature [22]. In addition, for low probability of intercept (LPI) and potentially multiuser applications, variants of direct-sequence spread-spectrum (DSSS) techniques will be used. The URDC applications that are envisioned include rapid mission/exercise debrief, interactive text (chat) capability, digital voice, and file transfer as well as test and evaluation specific applications such as transferring ground truth track of ship's position from shore and exercise coordination information.

8. CURRENT AND FUTURE RESEARCH

Current and future acoustic telemetry research is focused on a number of aspects of the telemetry problem. One area of active research is the development of iterative and

integral equalization and decoding methodologies to obtain performance at or better than that possible by conventional equalization techniques [26,27]. These procedures require more complex processing by virtue of the feedback nature and number of iterations inherent in these schemes. Future work in this area will involve novel methods of Turbo equalization employing Turbo codes and decoders while reducing overall complexity while maintaining improved performance. Another active area of current and future research is that of multiuser and networked node telemetry, associated protocols, routing, and multiuser techniques. The goal of this daunting task is to include as many underwater nodes, such as environmental devices, UUVs, remotely operated vehicles (ROVs), and submarines in increasingly larger communication nets while demanding performance approaching that of single-point links as well as conserving precious bandwidth. Finally, another active area of telemetry research that this author with other researchers is currently involved in is the remote optoacoustic and acoustooptic telemetry technology that connects submerged platforms at speed and depth with in-air platforms.

9. FINAL REMARKS

Once again, the dialog of humankind demands the search, collection, and transfer of knowledge and information to and from the underwater world. Our need to communicate will ultimately drive the technology and applications that solve the problems of the future that in the past were thought impossible to solve.

BIOGRAPHY

Fletcher A. Blackmon received his B.S. degree in electrical engineering in 1988 from Southeastern Massachusetts University (now known as University of Massachusetts at Dartmouth), his M.S. degree in electrical engineering in 1991 from the University of Massachusetts at Dartmouth, and is currently enrolled as a Ph.D. student in electrical engineering at the University of Massachusetts at Dartmouth. He joined the Naval Undersea Warfare Center in 1989 as an electronics engineer. At NUWC he worked on the research, design, and development of underwater acoustic communications and telemetry modem systems for Navy ranges. Since 1995, he has been involved in research at NUWC, where he has been working on opto-acoustic systems and applications. Mr. Blackmon holds a number of patents in the area of signal generation as well as patents pending in the areas of iterative and integral equalization and coding/decoding for underwater acoustic communication systems as well as the areas of opto-acoustic methods for communication and sonar applications. His areas of interest are the design and performance of equalization and decoding algorithms for underwater acoustic communications and opto-acoustic/acousto-optic methods for sonar and communications applications.

BIBLIOGRAPHY

1. A. Baggeroer, D. E. Koelsch, K. V. Der Heydt, and J. Catipovic, DATS — a digital acoustic telemetry system for underwater communications, *Proc. Oceans '81*, Boston, MA, 1981.
2. F. C. Jarvis, Description of a secure reliable acoustic system for use in offshore oil blowout (BOP) or wellhead control, *IEEE. J. Ocean. Eng.* **OE-9**: 253–258 (1984).
3. A. Kaya and S. Yauci, An acoustic communication system for subsea robot, *Proc. Oceans '89*, 1989, pp. 765–770.
4. L. Freitag, S. Meriam, D. Frye, and J. Catipovic, A long term deep water acoustic experiment, *Proc. Oceans '91*, Honolulu, Hawaii, 1991.
5. G. Mackelburg, Acoustic data links for UUVs, *Proc. Oceans '91*, Honolulu, Hawaii, 1991.
6. M. Suzuli and T. Sasaki, Digital acoustic image transmission system for deep sea research submersible, *Proc. Oceans '92*, 1992, pp. 567–570.
7. K. Scussel, J. Rice, and S. Meriam, A new MFSK acoustic modem for operation in adverse underwater channels, *Proc. Oceans '97*, Halifax, Nova Scotia, Canada, 1997.
8. A. Kaya and S. Yauchi, An acoustic communication system for subsea robot, *Proc. Oceans '89*, Seattle, WA, 1989.
9. M. Stojanovic, J. Catipovic, and J. G. Proakis, Adaptive multichannel combining and equalization for underwater acoustic communications, Part 1, *J. Acoust. Soc. Am.* **94**(3): 1621–1631 (Sept. 1993).
10. M. Stojanovic, J. Catipovic, and J. G. Proakis, Phase-coherent digital communications for underwater acoustic channels, *IEEE. J. Ocean. Eng.* **19**: 100–111 (1994).
11. A. Goalic et al., Toward a digital acoustic underwater phone, *Proc. Oceans '94*, Brest, France, 1994.
12. L. Freitag et al., A bidirectional coherent acoustic communication system for underwater vehicles, *Proc. Oceans '98*, Nice, France, 1998.
13. L. Freitag, M. Johnson, and J. Preisig, Acoustic communications for UUVs, *Sea Technol.* **40**(5): (May 1999).
14. D. Green and F. Blackmon, Performance of channel-equalized acoustic communications in the surf zone, *Proc. Oceans '01*, Honolulu, Hawaii, 2001.
15. T. Curtin and R. Benson, ONR program in underwater acoustic communications, *Sea Technol.* **4**(5): (May 1999).
16. T. Curtin, J. Bellingham, J. Catipovic, and D. Webb, Autonomous oceanographic sampling networks, *Oceanography* **6**: 86–94 (1993).
17. J. Proakis, M. Stojanovic, and J. Rice, Design of a communication network for shallow water acoustic modems, *Proc. Ocean Community Conf. '98*, Baltimore, MD, 1998.
18. M. Green and J. Rice, Handshake protocols and adaptive modulation for underwater communication networks, *Proc. Oceans '98*, Nice, France, 1998.
19. E. Soizer, M. Stojanovic, and J. Proakis, Underwater acoustic networks, *IEEE. J. Ocean. Eng.* **25**: 72–83 (2000).
20. D. Kilfoyle and A. Baggeroer, The state of the art in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **25**: 4–27 (2000).
21. S. M. Jarvis, F. A. Blackmon, K. Fitzpatrick, and R. Morrissey, Results from recent sea trials of the underwater digital acoustic telemetry system, *Proc. Oceans '97*, Oct. 1997.
22. F. Blackmon and W. Canto, Performance comparison of several contemporary equalizer structures applied to selected field test data, *Proc. Oceans '00*, Sept. 2000.

23. D. Thompson et al., Performance of coherent PSK receivers using adaptive combining, beamforming, and equalisation in 50 km underwater acoustic channels, *Proc. Oceans '96*, Sept. 1996.

24. J. A. Catipovic and L. E. Freitag, Spatial diversity processing for underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **16**(1): 86–97 (Jan. 1991).

25. S. M. Jarvis and N. A. Pendergrass, Implementation of a multichannel decision feedback equalizer for shallow water acoustic telemetry using a stabilized fast transversal filters algorithm, *Proc. Oceans '95*, Oct. 1995.

26. F. Blackmon et al., Performance comparison of iterative/integral equalizer/decoder structures for underwater acoustic channels, *Proc. Oceans '01*, Honolulu, Hawaii, Nov. 2001.

27. E. Sozer, J. Proakis, and F. Blackmon, Iterative equalization and decoding techniques for shallow water acoustic channels, *Proc. Oceans '01*, Honolulu, Hawaii, Nov. 2001.

ACOUSTIC TRANSDUCERS

DONALD P. MASSA
Massa Products Corporation
Hingham, Massachusetts

1. INTRODUCTION AND HISTORICAL OVERVIEW

The purpose of this article is to provide a brief overview of the very extensive topic of acoustic transducers. Transducers are devices that transform one form of energy into another. A few acoustic transducers, such as whistles or musical instruments, transform mechanical energy into sound, but the following discussion is concerned primarily with electroacoustic transducers. They are classified as either transmitters that convert electricity to sound, or receivers that change acoustic energy into electrical signals.

The invention of the telephone in the late 1800s resulted in the first widespread use of electroacoustic transducers. The microphone in the telephone converted the acoustical energy of the human voice into electrical signals. The earpiece in the telephone converted the electrical signals back into acoustic energy so the voice of the person at the other end of the line can be heard.

New requirements for different types of electroacoustic transducers were created by the development of the phonograph at the turn of the last century, followed by increased consumer use of radio in the 1920s and the advent of sound motion pictures in the 1930s. Improved loudspeakers and microphones were required to meet the demands of these new industries, and the science of sound was transformed into the applied science of electroacoustics.

During the 1920s, electrical engineers began applying the concepts of "equivalent circuits" to characterize acoustic transducers. The mechanical and acoustical portions of the transducer were modeled by converting them to equivalent electric circuit components of inductors, capacitors, and resistors. These equivalent-circuit elements of the acoustic portions were coupled to the pure electrical portions of the transducer by means of an electromechanical transformer. This modeling allowed the pioneering generation of electroacoustic engineers to not only better understand how transducers operated but also to optimize transducer designs by using the well-known methods of electric circuit analysis. In 1929 the Acoustical Society of America was formed, and in 1934 the first engineering-based textbook on transducers, entitled *Applied Acoustics*, was published by Olson and Massa.

While significant improvement in the design of electroacoustic transducers for use in the audible frequency band in air were achieved during 1900–1940, a new requirement for electroacoustic transducers to operate underwater for sonar applications was only in its infancy. However, the military threat of submarines during World War II caused sonar transducer development to rapidly advance during the 1940s.

Following World War II, new types of electroacoustic transducers designed to operate in the ultrasonic frequency range were developed for a wide variety of new industrial applications, such as noncontact distance or level measurement, collision avoidance, communication, remote control, intrusion alarms, ultrasonic cleaning, ultrasonic welding, ultrasonic flow detection, and ultrasonic imaging. Different transducers were designed to operate at frequencies as low as 20 kHz, the upper frequency limit of human hearing, to 10 MHz and higher [1–3].

2. FUNDAMENTALS OF ELECTROACOUSTIC TRANSDUCERS

Many factors affect the design of an electroacoustic transducer. For example, a transducer designed to operate in a gaseous medium, such as air, is very different from one designed to operate in a liquid medium, such as water. Likewise, differences in acoustical requirements, such as frequency of operation or radiation pattern, will influence the design. It is, therefore, necessary to first understand some basic acoustical principles in order to properly understand how electroacoustic transducers operate.

2.1. Generation of Sound

Sound is a transfer of energy caused by the vibration of particles in the transmission medium. The particles vibrate back and forth a small distance, which causes a longitudinal wave to travel in the same direction as the vibrating particles.

An electroacoustic transmitting transducer produces sound by vibrating a portion of its surface, which therefore affects the molecules in the transmission medium. When the radiating surface moves forward the molecules are pushed closer together, thus increasing the instantaneous pressure (condensation). When the radiating surface moves back, the molecules expand, thus decreasing the instantaneous pressure (rarefaction). The vibrating molecules near the transducer push against their neighbors, causing them to also vibrate. This process continues, creating a propagating wave in which the instantaneous

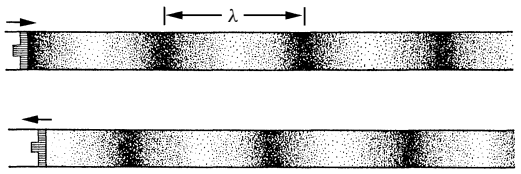


Figure 1. Illustration showing how a piston moving forward (top) compresses the molecules in the transmission medium in a pipe creating condensation, and moving backward (bottom) expands the molecules creating rarefaction.

pressures oscillate between condensation and rarefaction as it progresses outward from the transducer. This is illustrated in Fig. 1, which shows two moments of time of a piston moving back and forth, pushing against the molecules in a transmission medium contained in a pipe. The top picture shows the molecules when the piston is moving forward creating condensation, while the bottom diagram shows the piston moving backward, causing rarefaction.

2.2. Differences in the Characteristics of Sound Propagation in Gaseous and Liquid Media

There are fundamental differences in many of the properties of sound radiating in a liquid as compared to sound propagating in a gas, but there are typically only minor variations in the acoustic properties of sound radiating among various gases or among various liquids. Since the applications of most sound transmissions in gases occur in air, and in liquids occur in water, the characteristics of these two media will be used to illustrate the difference between acoustic radiation in liquids and in gases.

2.2.1. Speed of Sound. The velocity with which the sound waves travel through a transmission medium is called the *speed of sound, c*. The nominal value of c is primarily a function of the composition of the particular medium, but slight changes occur for each medium because of variations in parameters such as temperature or pressure. However, the velocity of sound is much greater in liquids than in gasses. As an example, in air at 20 °C the speed of sound is 343 m/s, and in freshwater at 20 °C it is 1483 m/s [4].

2.2.2. Wavelengths of Sound. The wavelength of sound traveling in a medium is the distance between condensation peaks, as shown in Fig. 1, and is a function of both the frequency and the speed of the sound wave. The wavelength is

$$\lambda = \frac{c}{f} \tag{1}$$

where λ is the wavelength, c is the speed of sound, and f is the frequency.

Figure 2 shows a plot of the wavelength of sound in air and water at room temperature as a function of frequency. As can be seen from the curve, since the speed of sound in water is approximately 4.3 times greater than in air, the wavelength for a given frequency in water is approximately 4.3 times longer than in air.

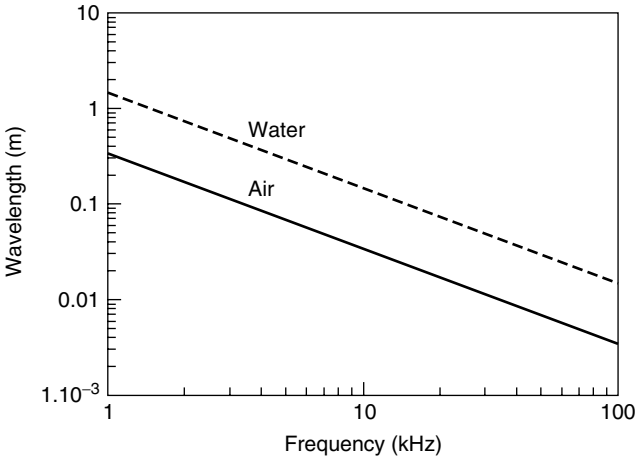


Figure 2. Plot of the wavelength, λ, as a function of frequency for sound in air and water.

2.2.3. Attenuation of Sound. The attenuation of sound traveling through a medium increases as the frequency increases, and the attenuation in a gas at a given frequency is much greater than in a liquid. Figure 3 shows plots of typical attenuations for sound in both air and water as a function of frequency [5,6]. Because the attenuation is much less in water than in air, objects can be detected at much greater ranges using echo location in water than in air. Table 1 compares propagation distances and wavelengths for sound at different frequencies in air and water.

2.2.4. Density. Gasses are much less dense than liquids. The density, ρ_0, of air is only 1.2 kg/m^3. The

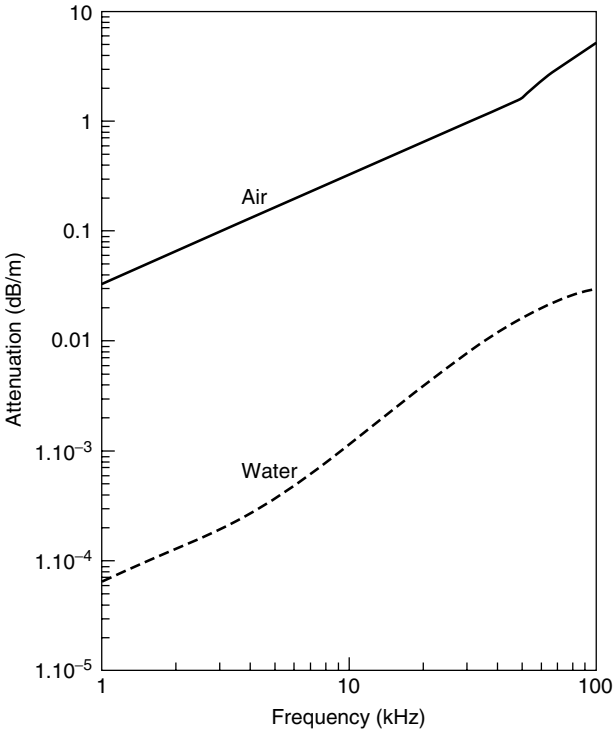


Figure 3. Plot of the attenuation, λ, as a function of frequency for sound in air and water.

Table 1. Sound Propagation Distance and Wavelength Comprison for Air and Water

	Propagation Distance for Planar Sound Wave to Attenuate to Half Its Initial Pressure (No Spreading Loss) (m)		Wavelength (m)	
Frequency (kHz)	Air	Water	Air	Water
0.5	375	330,000	0.68	3
1.0	180	130,000	0.34	1.5
5.0	40	20,000	0.068	0.3
10	18	7,000	0.034	0.15
50	3.8	330	0.0068	0.03
100	1.2	160	0.0034	0.015

density of water is 1000 kg/m^3, which is over 800 times greater [4].

2.2.5. Analogies Between Acoustical Properties of a Transmission Medium and the Electrical Properties of a Circuit. Most people are familiar with Ohm's law in electricity. This law states that for a given voltage E applied across an electrical component that has an impedance Z, the electric current I flowing through the component will be directly proportioned to the value of the electrical impedance Z. Ohm's law can be written as

$$E = IZ \tag{2}$$

where E is the voltage in volts, I is the current in amperes, and Z is the electrical impedance in ohms. Acoustic transmission media have properties analogous to electrical properties in circuits. In acoustics, the sound pressure p of the acoustic wave is equivalent to voltage in an electric circuit, and the particle velocity u is analogous to current. The acoustical impedance of a transmission medium is the product of the density times the speed of sound. It is written as $\rho_0 c$, and the units are acoustic ohms or rayls (after Lord Rayleigh). Ohm's law in acoustics is

$$p = u\rho_0 c \tag{3}$$

where p is the pressure of the sound wave in pascals, u is the particle velocity in meters per second, and $\rho_0 c$ is the acoustic impedance in rayls.

Because gases have much slower sound velocities and lower densities than do liquids, their acoustic impedances are much less. For example, the acoustic impedance of air is nominally 415 rayls, while $\rho_0 c$ for water is approximately 1.48×10^6 rayls. This large disparity causes the fundamental design concepts to be much different from those for use in water. Table 2 summarizes the analogies between acoustical and electrical properties.

2.2.6. Relationship Between Sound Pressure, Particle Displacement, and Partial Velocity. As was discussed previously, sound is transmitted in a wave caused by the particles in the medium vibrating back and forth. In a plane acoustic wave traveling in the x direction, the particle velocity is the real part of

$$\xi = |\xi| e^{j(\omega t - (\omega/c)x)} \tag{4a}$$

which is

$$\xi = |\xi| \cos\left(\omega t - \frac{\omega}{c}x\right) \tag{4b}$$

where ξ is the particle displacement in meters, ω is the frequency of the sound wave in radians per second, and c is the speed of sound in the medium in meters per second.

The particle velocity is the derivative of the displacement, therefore

$$u = -\omega|\xi| \sin\left(\omega t - \frac{\omega}{c}x\right) \tag{5}$$

where u is the particle velocity in meters per second.

Substituting Eq. (5) into Eq. (3), the sound pressure becomes

$$p = -\rho_0 c\omega|\xi| \sin\left(\omega t - \frac{\omega}{c}x\right) \tag{6}$$

where p is the sound pressure of the acoustic wave in pascals and $\rho_0 c$ is the acoustic impedance in rayls. From Equation 6, it can be seen that the magnitude of the sound pressure of an acoustic wave is directly proportional to the acoustic impedance, the frequency, and the magnitude of the particle displacement. To transmit sound at a specific pressure and frequency, an acoustical transducer must vibrate with an amplitude equal to the particle displacement required for the particular medium. Because the acoustic impedance of water is 3600 times greater than that of air, the particle displacement in air must be 3600 times greater than the particle displacement in water to produce the same sound pressure at the same frequency. For either medium, the particle displacement required to produce a constant sound pressure will decrease as the frequency increases.

Because of these relationships, transducer designs are different for operation in fluids than in gases, or for

Table 2. Ohm's Law Analogies Between Electrical and Acoustical Properties

Electrical			Acoustical		
Quantity	Symbol	Units	Quantity	Symbol	Units
Voltage	E	Volts	Pressure	p	μPa
Charge	q	Coulomb	Particle displacement	ξ	m
Current	I	Amperes	Particle velocity	u	m/s
Impedance	Z	Ohms	Acoustic impedance	$\rho \cdot c$	rayls

operation at high frequencies than low frequencies. The radiating surfaces in underwater sonar transducers have to move only small displacements to generate large sound pressures. However, they must generate a relatively large amount of force to compress the dense water. Transducers that operate in air have to vibrate over much larger displacements to generate high sound pressures, but very little force is necessary to compress the gas.

Transducers that radiate at low frequencies in either medium must vibrate for greater distances than those that operate at high frequencies, which can be verified by observing loudspeakers in a typical stereo system. The low-frequency "woofer" can be seen moving large amplitudes when base notes are played, while the motion of the high-frequency "tweeters" appears to be negligible.

2.3. Sound Pressure Levels

Because sound pressures vary more than 10 orders of magnitude, they are expressed by acoustical engineers as logarithmic ratios, which are called sound pressure levels (SPLs). The SPL in decibels for a sound pressure p is calculated as $20\log(p/p_{\text{ref}})$, where p_{ref} is a standard reference sound pressure. Some confusion can occur because several different reference pressures are in use, which results in a given sound pressure being expressed with several different possible sound pressure levels.

Most of the early work in acoustical engineering was associated with the development of audio equipment, so it was natural to use the threshold of human hearing for the reference pressure. In the cgs system, that sound pressure is 0.0002 dyn/cm^2 (0.0002 μbar), so sound pressure levels were expressed in terms of dB/0.0002 μbar [SPL $= 20\log(p/0.0002)$ dB/0.0002 μbar, where p is the pressure in microbars].

During World War II, there were major advances in the development of sonar for detecting submarines. Since the sounds produced by sonar systems are not heard directly by people, sonar engineers began using 1 μbar as a more logical standard reference pressure. Sound pressure levels therefore began being expressed in terms of dB/1 μbar [SPL $= 20\log(p/1)$ dB/1 μbar, where p is pressure in microbars].

In the early 1970s, the SI system of units was adopted in acoustical engineering, so the micropascal (μPa), which is equal to 10^{-6}N/m^2, became the reference pressure. Sound pressure levels therefore began to be expressed in terms of dB/1 μPa [SPL $= 20\log(p/1)$ dB/1 μPa, where p is pressure in μPa]. This is now the most often used reference pressure for acoustic measurements, but it is not unusual to encounter data using any of these three standard reference pressures. To add to the confusion, sometimes the SPL will be improperly stated in terms of decibels only, without indicating the reference pressure used to compute the ratio. It is obviously important to know which reference pressure was used whenever an SPL is expressed, and when comparing transducer responses, all sound pressure levels should be converted to the same reference pressure. It is quite simple to convert sound pressure levels among the three reference pressures by using Table 3.

Table 3. Sound-Pressure-Level Conversion Table

To Convert SPL in	To SPL in	
dB/0.0002 μbar	dB/1 μbar	Subtract 74 dB
dB/0.0002 μbar	dB/1 μPa	Add 26 dB
dB/1 μPa	dB/0.0002 μbar	Subtract 26 dB
dB/1 μPa	dB/1 μbar	Subtract 100 dB
dB/1 μbar	dB/0.0002 μbar	Add 74 dB
dB/1 μbar	dB/1 μPa	Add 100 dB

2.4. Radiation Patterns of Transducers

The acoustic radiation pattern, or beam pattern, is the relative sensitivity of a transducer as a function of spatial angle. This pattern is determined by factors such as the frequency of operation and the size, shape, and acoustic phase characteristics of the vibrating surface. The beam patterns of transducers are reciprocal, which means that the beam will be the same whether the transducer is used as a transmitter or as a receiver.

Transducers can be designed to radiate sound in many different types of patterns, from omnidirectional to very narrow beams. The beam pattern of a transducer is usually calculated and graphed showing the relative reduction in sensitivity as a function of angle, with the maximum sensitivity of the transducer along the main acoustic axis set to equal 0 dB. The beam angle of the transducer is equal to the total arch encompassed by the beam between the angles when the pressure has fallen to a level of −3 dB on either side of the main acoustic axis.

For a transducer with a circular radiating surface vibrating in phase, the narrowness of the beam pattern is a function of the ratio D/λ, the diameter of the radiating surface over the wavelength of sound at the operating frequency. The larger the diameter of the transducer as compared to a wavelength of sound, the narrower the sound beam. For example, if the diameter is twice the dimension of the wavelength, the total beam angle will be approximately 30°, but if either the diameter or frequency is changed so that the ratio becomes 10, the total beam angle will be reduced to approximately 6°. Since the wavelength of sound at a given frequency in water is approximately 4.3 times larger than in air, the diameter of an underwater transducer must be approximately 4.3 times larger than an air transducer to produce the same beam angle at the same frequency.

A transducer large in size compared to a wavelength produces not only a narrow main beam, but also secondary lobes separated by nulls. Figure 4 is a three-dimensional (3D) representation of the beam pattern produced by a transducer with a radiating diameter that is large compared to a wavelength. As can be seen, each secondary lobe is sequentially lower in amplitude than the previous one. The equation for the radiation pattern of a circular rigid piston in an infinite baffle as a function of spatial angle is [7].

$$P(\theta) = \left[\frac{2J_1\left(\pi\dfrac{D}{\lambda}\sin\theta\right)}{\pi\dfrac{D}{\lambda}\sin\theta}\right]^2 \tag{7a}$$

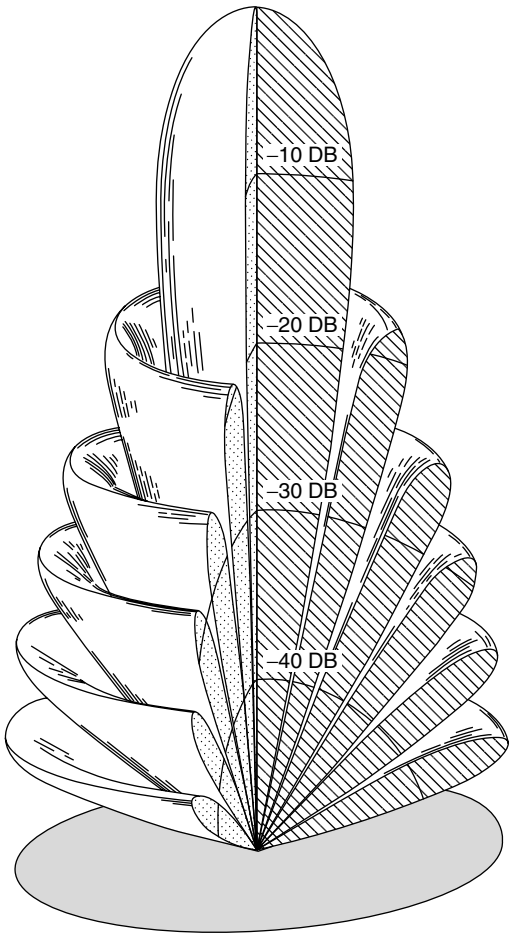


Figure 4. 3D beam pattern produced by a transducer with a circular radiating surface where the diameter is large compared to a wavelength.

where $P(\theta)$ is the relative sound pressure as a function of the angle, θ is the angle of the sound pressure from an axis perpendicular to the center of the piston, D is the diameter of the piston, λ is the wavelength of the sound, and J_1 is the first-order Bessel function.

Beam patterns are usually plotted on a decibel scale where the sound pressure as a function of spatial angle is

$$P_{\mathrm{dB}}(\theta) = 20 \log \left[\frac{2J_1\left(\pi \dfrac{D}{\lambda}\right) \sin\theta}{\pi \dfrac{D}{\lambda} \sin\theta} \right] \tag{7b}$$

where $P_{\mathrm{dB}}(\theta)$ is the relative sound pressure as a function of spatial angle in decibels. The beam angle is usually defined as the measurement of the total angle where the sound pressure level of the main beam has been reduced by 3 dB on both sides from the peak that occurs along the axis perpendicular to the piston. When describing transducer beam patterns, two-dimensional (2D) plots are most commonly used. These show the relative sensitivity of the transducer versus angle in a single plane cut through the axis perpendicular to the center of the piston in the 3D beam pattern. Figure 5 shows 2D plots on rectilinear coordinates of the beam patterns of circular piston radiators for several different values of D/λ.

2.5. Resonance

Many electroacoustic transducers are designed to operate at resonance, which is the natural frequency of vibration of the transducer structure. Transducers will produce a much greater displacement for a given drive voltage when operated at frequencies in the vicinity of resonance. Likewise, when used as receivers they will produce a larger electrical signal for a given sound pressure.

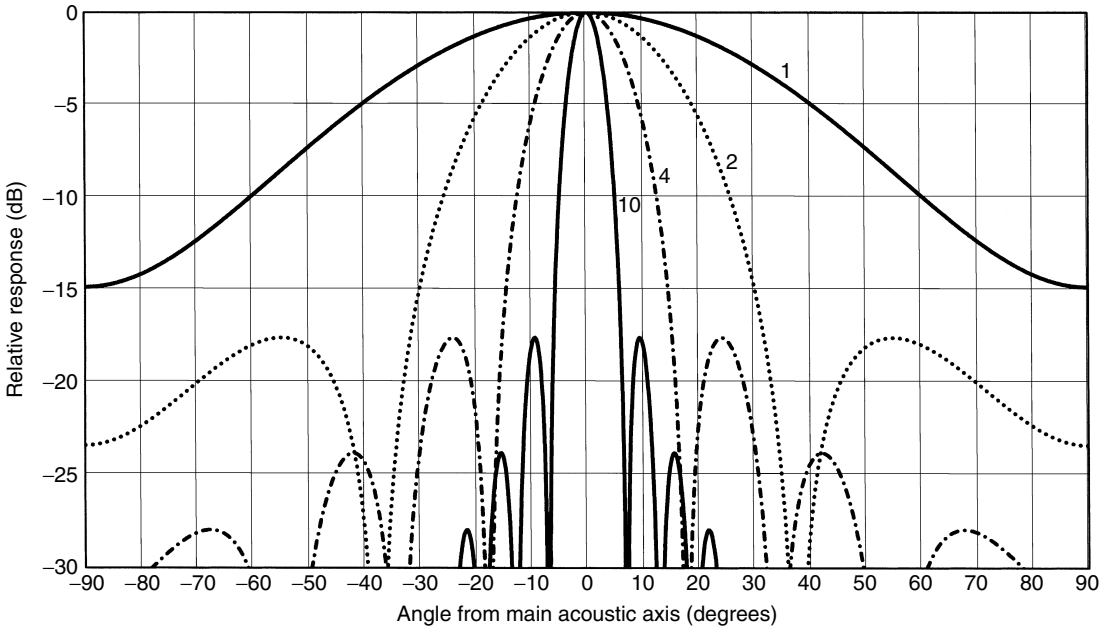


Figure 5. 2D graph showing the beam patterns of four different transducers radiating with circular pistons in an infinite baffle having diameter to wavelength ratios (D/λ) of 1, 2, 4, and 10 [From Eq. (7b)].

The quality factor, Q, of a transducer is a value that indicates the width of the frequency band in the vicinity of resonance over which it can operate with high output. The Q is calculated by dividing the resonant frequency by the bandwidth, which is defined as the frequency band over which the response of the transducer is within 3 dB of the peak response.

The receiving response of transducers is usually constant at frequencies well below resonance, so the output voltage will be constant for a given sound pressure, and proportional to changing sound pressures at all frequencies in this region. Transducers that are used only for receiving (microphones in air, and hydrophones in water) are often operated well below resonance to take advantage of this broadband flat response.

3. TRANSDUCTION

Electroacoustic transducers operate by using a variety of different transduction materials or mechanisms to transform electrical energy into sound and vice versa. For example, in transducers that employ magnetics, an alternating electric current flowing through a coil of wire produces a varying magnetic force that causes the transducer structure to vibrate. In like manner, a sound wave will vibrate the transducer, which moves the coil in a magnetic field, thus generating an electrical signal.

Transducers can also be designed using magnetostrictive materials for transduction. When these materials are placed in a magnetic field, their mechanical dimensions change as a function of the strength of the magnetic field, which in turn can be used to generate sound. Other transducers employ piezoelectric crystals, such as quartz, Rochelle salt, or ammonium dihydrogen phosphate (ADP) for transduction. They develop an electric charge between two surfaces when the crystal is mechanically compressed, and they expand and contract in size in the presence of an applied electric field.

The most commonly used transduction materials for transducers are electrostrictive ceramics. These ceramic materials, such as barium titinate and lead zirconate titanate, are often referred to as *piezoelectric ceramics* and also produce an electric charge when a mechanical stress is applied, and vice versa. However, they must have an internal polarizing electric field established in order for transduction to occur. Their popularity is due to relatively low cost, coupled with the ability to be fabricated into a wide variety of shapes and sizes.

4. A FEW EXAMPLES OF SOME ELECTROACOUSTIC TRANSDUCERS

The following sections contain short descriptions of the construction of a few electroacoustic transducers. Since there are such a wide variety of different types of electroacoustic transducers, it is not possible to provide a description of most of them in this brief overview. Some of the publications in the reading list at the end of this article contain detailed information on many specific types of transducers.

4.1. Moving-Coil Electrodynamic Loud Speaker

The most common loudspeakers used in stereo or public address systems are electrodynamic transducers, which contain a coil of wire suspended in a magnetic field. When an alternating electrical current is passed through the coil, mechanical forces are developed between the coil's electromagnetic field and the field in which it is mounted.

Figure 6 is a cross-sectional sketch illustrating the schematic construction of an electrodynamic speaker [8]. As can be seen, the voice coil (4) is a coil of wire fashioned into a cylindrical tube. It is rigidly connected to a radiating diaphragm (1), which is resiliently mounted to an enclosure (3). This holds the coil within the magnetic field produced by the permanent magnet (4), but allows it to freely vibrate within this field. The magnet is shaped like a disk with a circular groove cut into the surface facing the diaphragm. The tubular voice coil is mounted so that it is held within this groove. A varying electrical current in the coil produces proportional changes in its electromagnetic field, which in turn modulates the magnetic forces between the coil and the permanent magnet. This causes the coil to move back and forth, thus vibrating the diaphragm and generating sound.

4.2. Condenser Microphone

The condenser microphone produces a variation in its electrical capacitance in the presence of an acoustic wave. Figure 7 illustrates the construction of such a transducer. The stretched thin metallic membrane is separated from the rigid backplate by a small airgap. When a sound wave vibrates the membrane, it causes the airgap to change in thickness, producing a variation in the electrical capacitance between it and the backplate. This varying capacitance is converted into an electrical signal that is proportional to the sound pressure wave.

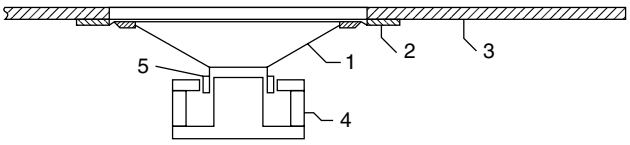


Figure 6. Illustration showing the construction of an electrodynamic speaker (Fig. 1 of U.S. Patent 2,445,276 [8]).

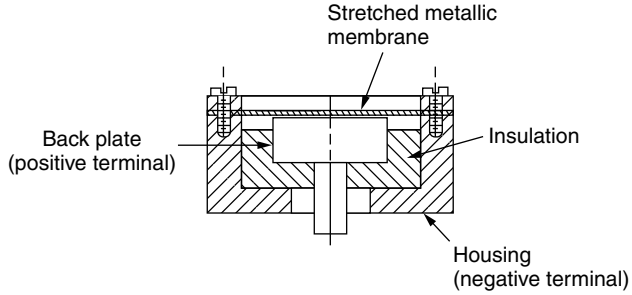


Figure 7. Illustration showing the construction of a condenser microphone.

4.3. Flexural Air Ultrasonic Transducer Using Electrostrictive Ceramics

Flexural ultrasonic transducers use the resonance of a mechanical diaphragm to produce the motion required to generate the required sound pressure. Figure 8 shows a cross-sectional illustration of a typical flexural ultrasonic transducer [9]. The housing is an aluminum cup consisting of an outer cylindrical shell (1) with relatively thick sides and a thin circular diaphragm. This produces a rigid clamped circular disk in which the resonant frequency is controlled primarily by its diameter and stiffness. A thin ceramic disk (5) is cemented to the radiating diaphragm.

As a receiver, the diaphragm is mechanically vibrated by sound pressure, causing it to buckle up and down. Since the ceramic is rigidly attached to the diaphragm, it stretches as the diaphragm buckles, which produces an electrical voltage across it. In like manner, when the ceramic is excited by an electrical voltage it will stretch, causing the diaphragm to vibrate and transmit sound. Because the diaphragm can move large displacements while creating only minor strains in the ceramic, this design allows for generation of large sound pressures without over stressing and cracking the ceramic.

This particular transducer design operates at an overtone of the fundamental resonant frequency of the clamped diaphragm. The frequency of operation can be adjusted by varying the diaphragm thickness.

4.4. Tonpilz Sonar Transducer

A mass-loaded vibratile transducer (Tonpilz transducer) is a common design used in sonar. A typical example is shown in Figure 9 [10]. A ceramic cylinder (12) is cemented between a light aluminum head mass (11), and a heavy steel tail mass (15). The ceramic has electrodes on its two ends. This transducer resonates in much the same way as a large mass attached to a spring. If the mass is reduced, the resonant frequency will lower. If the stiffness of the spring increases, the resonant frequency will be higher.

In the transducer of Fig. 9, the ceramic cylinder is the spring connected to the head mass and the tail mass. If it is made more compliant, for example, by reducing the wall thickness or increasing the length, the resonant frequency will lower. If the head and tail masses are made smaller, the resonant frequency will increase. At resonance, the length of the ceramic will increase and decrease relatively large amounts, causing the head and tail masses to vibrate. Because the head mass is much lighter than the tail mass, it vibrates at much larger amplitude.

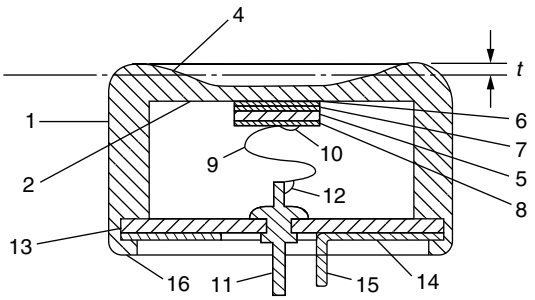


Figure 8. Illustration showing the construction of a flexural ultrasonic transducer designed for operation in air using an electrostrictive ceramic for transduction (Fig. 1 of U.S. Patent 3,943,388 [9]).

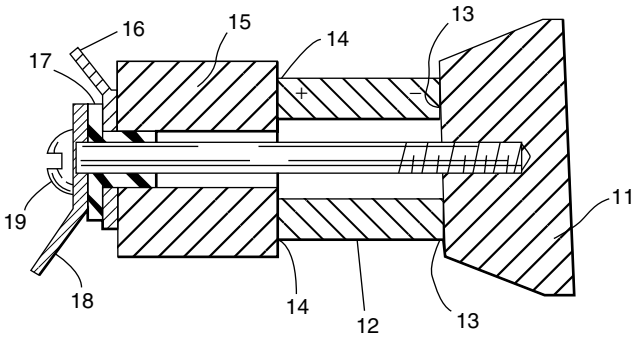


Figure 9. Illustration showing the construction of a tonpilz sonar transducer (Fig. 1 of U.S. Patent 3,739,327 [10]).

In operation the structure is encapsulated in waterproof material, such as rubber, and the radiating head is acoustically coupled to the water. When used as a transmitter, an oscillating electrical voltage is connected across the electrodes of the ceramic, causing it to alternately lengthen and contract. This in turn causes the head mass, which is coupled to the water, to vibrate large amplitudes and produce a sound pressure wave. As a receiver, a sound pressure wave pushes the head mass, causing the transducer structure to vibrate. This in turn causes the length of the ceramic tube to alternately contract and expand, which generates a voltage across the ceramic stack.

BIOGRAPHY

Donald P. Massa received a B.S. and an M.S. degree in electrical engineering from Northeastern University in 1969 and 1972, respectively. His professional career started at Woods Hole Oceanographic Institution as a co-op student, and in the late 1960s he assumed full-time responsibilities as a development engineer for advanced electroacoustic products at Massa Products Corporation. He has been the president and chief technical officer of the company since 1976. He has been responsibilities for the development and production design of more than 100 electroacoustic transducers and systems for both air ultrasonic and underwater sonar applications. Mr. Massa holds 14 patents on electroacoustic devices and systems, and he has published numerous technical articles and presented many invited and contributing papers on electroacoustics to professional societies, the U.S. Navy, and Allied Navies. He was awarded the Outstanding Alumni Award in the Field of Science and Technology by Northeastern University in 1997, and he also serves as a member of Northeastern's Board of Trustees.

BIBLIOGRAPHY

1. F. Massa, Some personal recollections of early experiences on the new frontier of electroacoustics during the late 1920's

and early 1930's, *J. Acoust. Soc. Am.* **77**(4): 1296–1302 (April 1985).

2. F. Massa, Sonar transducers: A history, *Sea Technol.* (Nov. 1989).
3. H. Olson and F. Massa, *Applied Acoustics*, Blakston, Philadelphia, 1934.
4. L. Kinsler and A. Frey, *Fundamentals of Acoustics*, Wiley, New York, 1962.
5. D. Massa, Choosing an ultrasonic sensor for proximity or distance measurement, Parts 1, 2, *Sensors* (Feb., March 1999).
6. R. Urick, *Principles of Underwater Sound for Engineers*, McGraw-Hill, New York, 1967.
7. F. Massa, Radiation of sound, in *American Institute of Physics Handbook*, McGraw-Hill, New York, 1963, Sect. 3, pp. 118–122.
8. U.S. Patent 2,445,276 (July 13, 1948), F. Massa, Electrodynamic loudspeakers.
9. U.S. Patent 3,943,388 (to D. Massa, Trustee Stoneleigh Trust) (March 9, 1976), F. Massa, Electroacoustic transducer of the flexural vibrating diaphragm type.
10. U.S. Patent 3,739,327 (to Massa Div. DCA) (June 12, 1973), D. Massa, Electroacoustic transducers of the mass loaded vibratile piston type.

ACOUSTIC (UNDERWATER) COMMUNICATIONS

MILICA STOJANOVIC
Massachusetts Institute of Technology
Cambridge, Massachusetts

The need for underwater wireless communications exists in applications such as remote control in offshore oil industry, pollution monitoring in environmental systems, collection of scientific data recorded at ocean-bottom stations and unmanned underwater vehicles, speech transmission between divers, and mapping of the ocean floor for detection of objects and discovery of new resources. Wireless underwater communications can be established by transmission of acoustic waves. The underwater acoustic communication channels, however, have limited bandwidth, and often cause signal dispersion in time and frequency [2–7]. Despite these limitations, underwater acoustic communications are a rapidly growing field of research and engineering.

Acoustic waves are not the only means for wireless communication underwater, but they are the best known so far. Radiowaves that will propagate any distance through conductive seawater are the extra-low-frequency ones (30–300 Hz), which require large antennaes and high transmitter powers [1]. Optical waves do not suffer as much from attenuation, but they are affected by scattering. Transmission of optical signals requires high precision in pointing the narrow laser beams, which are still being perfected for practical use. Thus, acoustic waves remain the single best solution for communicating underwater, in applications where tethering is not acceptable.

The idea of sending and receiving information under water is traced back all the way to the time of Leonardo Da Vinci, who is quoted for discovering the possibility of detecting a distant ship by listening on a long tube submerged under the sea. In the modern sense of the word, underwater communications began to develop during World War II, for military purposes. One of the first underwater communication systems was an underwater telephone, developed in 1945 in the United States for communicating with submarines [4]. This device used a single-sideband (SSB) suppressed carrier amplitude modulation in the frequency range of 8–11 kHz, and it was capable of sending acoustic signals over distances of several kilometers. However, it was not until the development of VLSI (very large-scale integration) technology that a new generation of underwater acoustic communication systems began to emerge. With the availability of compact digital signal processors (DSPs) with their moderate power requirements, it became possible for the first time to implement complex signal processing and data compression algorithms at the submerged ends of an underwater communication link.

Since the late 1990s, significant advancements have been made in the development of underwater acoustic communication systems [7], in terms of their operational range and data throughput. Acoustically controlled robots have been used to replace divers in performing maintenance of submerged platforms [16], high-quality video transmission from the bottom of deepest ocean trenches (6500 km) to a surface ship was established [17], and data telemetry over horizontal distances in excess of 200 km was demonstrated [25].

As efficient communication systems are developing, the scope of their applications continues to grow, and so do the requirements on the system performance. Many of the developing applications, both commercial and military, are calling for real-time communication with submarines and autonomous, or unmanned underwater vehicles (AUVs, UUVs). Setting the underwater vehicles free from cables will enable them to move freely and refine their range of operation. The emerging communication scenario in which the modern underwater acoustic systems will operate is that of an underwater data network consisting of both stationary and mobile nodes. This network is envisaged to provide exchange of data, such as control, telemetry, and eventually video signals, between many network nodes. The network nodes, located on underwater moorings, robots, and vehicles, will be equipped with various sensors, sonars, and videocameras. A remote user will be able to access the network via a radio link to a central node based on a surface station.

In attempts to achieve these goals, current research is focusing on the development of efficient communications and signal processing algorithms, design of efficient modulation and coding schemes, and techniques for mobile underwater communications. In addition, multiple-access communication methods are being considered for underwater acoustic networks, as well as the design of network protocols, suited for long propagation delays and strict power requirements encountered in the underwater environment. Finally, data compression algorithms suitable

for low-contrast underwater images, and related image processing methods [18], are expected to enable image transmission through band-limited underwater acoustic channels.

1. SYSTEM REQUIREMENTS

The achievable data throughput and the reliability of an underwater acoustic communication system, as measured by the bit error rate, vary from system to system, but are always subject to bandwidth limitations of the ocean channel. Unlike the situation in the majority of other communication media, the use of underwater acoustic resources has not been regulated yet by standards.

In the existing systems, four kinds of signals usually are transmitted: control, telemetry, speech, and video signals.

1. Control signals include navigation, status information, and various on/off commands for underwater robots, vehicles, and submerged instrumentation such as pipeline valves or deep-ocean moorings. The data rates up to about 1 kilobit per second (kbps) are sufficient for these operations, but very low bit error rates (BERs) may be required.
2. Telemetry data are collected by submerged acoustic instruments such as hydrophones, seismometers, sonars, current meters, and chemical sensors, and it also may include low-rate image data. Data rates on the order of one to several tens of kbps are required for these applications. The reliability requirements are not so stringent as for the command signals, and a probability of bit error of 10^{-3}–10^{-4} is acceptable for many applications.
3. Speech signals are transmitted between divers and a surface station or among divers. While the existing, commercially available diver communication systems use mostly analog communications, based on single-sideband modulation of the 3-kHz audio signal, research is advancing in the area of synthetic speech transmission for divers, as digital transmission is expected to provide better reliability. Transmission of digitized speech by linear predictive coding (LPC) methods requires rates on the order of several kbps to achieve close-to-toll quality. The BER tolerance of $\sim 10^{-2}$ makes it a viable technology for poor-quality band-limited underwater channels [19,20].
4. Video transmission over underwater acoustic channels requires extremely high compression ratios if an acceptable frame transmission rate is to be achieved. Fortunately, underwater images exhibit low contrast and detail, and preserve satisfactory quality if compressed even to 2 bits per pixel. Compression methods, such as the JPEG (Joint Photographic Experts Group) standard discrete cosine transform, have been used to transmit 256×256-pixel still images with 2 bits per pixel, at transmission rates of about one frame per 10 seconds (s^{-1}) [17]. Further reduction of the required transmission rate seems to be possible by using dedicated compression algorithms, such as the discrete wavelet transform [18]. Current achievements report on the development of algorithms capable of attaining compression ratios in excess of 100 : 1. On the other hand, underwater acoustic transmission of television-quality monochrome video would require compression ratios in excess of 1000 : 1. Hence, the required bit rates for video transmission are greater than 10 kbps, and possibly up to several hundreds of kbps. Performance requirements are moderate, as images will have satisfactory quality at bit error rates on the order of 10^{-3}–10^{-4}.

2. CHANNEL CHARACTERISTICS

Sound propagation under water is determined primarily by transmission loss, noise, reverberation, and temporal and spatial variability of the channel. Transmission loss and noise are the principal factors determining the available bandwidth, range, and signal-to-noise ratio. Time-varying multipath influences signal design and processing, which determine the information throughput and communication system performance.

2.1. Range and Bandwidth

Transmission loss is caused by energy spreading and sound absorption. While the energy spreading loss depends only on the propagation distance, the absorption loss increases not only with range but also with frequency, thus limiting the available bandwidth.

In addition to the nominal transmission loss, link condition is influenced largely by the spatial variability of the underwater acoustic channel. Spatial variability is a consequence of the waveguide nature of the channel, which results in such phenomena as formation of shadow zones. Transmission loss at a particular location can be predicted by many of the propagation modeling techniques [2] with various degrees of accuracy. Spatial dependence of transmission loss imposes particularly severe problems for communication with moving sources or receivers.

Noise observed in the ocean consists of human-made noise and ambient noise. In deep ocean, ambient noise dominates, while near shores, and in the presence of shipping activity, human-made noise significantly increases the noise level. Unlike the human-made noise, most of the ambient noise sources can be described as having a continuous spectrum and Gaussian statistics [2]. As a first approximation, the ambient noise power spectral density is commonly assumed to decay at 20 dB/decade in both shallow and deep water, over frequencies that are of interest to communication systems design. The exception are biological sources of noise, such as snapping shrimp, which lives only in certain geographic areas and produces impulsive noise within the range of frequencies used by a typical communication system.

Frequency-dependent transmission loss and noise determine the relationship between the available range, bandwidth, and SNR (signal-to-noise ratio) at the receiver input. This dependence is illustrated in Fig. 1, which shows the frequency dependent portion of SNR for several

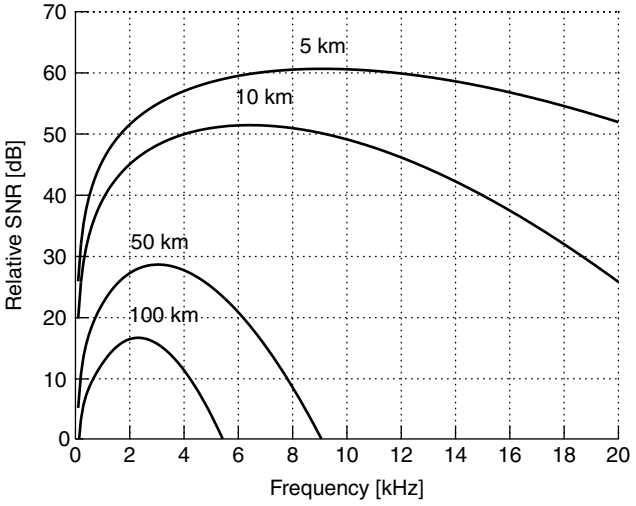


Figure 1. Frequency-dependent portion of SNR.

transmission ranges. (The SNR is evaluated assuming spherical spreading, absorption according to Thorp [2], and a 20-dB/decade decay of the noise power spectral density.) Evidently, this dependence influences the choice of a carrier frequency for the desired transmission range. In addition, it determines the relationship between the available range and frequency band. Underwater acoustic communication links can be classified according to range as very long, long, medium, short, and very short links. For a long-range system, operating over 10–100 km, the bandwidth is limited to few kilohertz (for a very long distance on the order of 1000 km, the available bandwidth falls below 1 kHz). A medium-range system operating over 1–10 km has a bandwidth on the order of 10 kHz, while only at very short ranges below about 100 m, more than 100 kHz of bandwidth may be available.

Within this limited bandwidth, the signal is subject to multipath propagation through a channel whose characteristics vary with time and are highly dependent on transmitter and receiver location. The multipath structure depends on the link configuration, which is designated primarily as vertical or horizontal. While vertical channels exhibit little time dispersion, horizontal channels may have extremely long multipath spreads. Most notable in the long- and medium-range channels, multipath propagation causes severe degradation of the acoustic communication signals. Combating the underwater multipath to achieve a high data throughput is without exception considered to be the most challenging task of an underwater acoustic communication system.

2.2. Multipath

In a digital communication system that uses a single carrier, multipath propagation causes intersymbol interference (ISI), and an important figure of merit is multipath spread in terms of symbol intervals. While typical multipath spreads in the commonly used radio channels are on the order of several symbol intervals, in the horizontal underwater acoustic channels they increase to several tens, or a hundred of symbol intervals for moderate to high data rates. For example, a commonly encountered multipath spread of 10 ms in a medium-range shallow-water channel, causes the ISI to extend over 100 symbols if the system is operating at a rate of 10 kilosymbols per second (ksps).

The mechanisms of multipath formation in the ocean are different in deep and shallow water, and also depend on the frequency and range of transmission. Understanding of these mechanisms is based on the theory and models of sound propagation. Depending on the system location, there are several typical ways of multipath propagation. It is mostly the water depth that determines the type of propagation. The delineation between shallow and deep water is not a strict one, but usually implies the region of continental shelves, with depth less than about 100 m, and the region past the continental shelves, where the water gets deeper. Two fundamental mechanisms of multipath formation are reflection at boundaries (bottom, surface, and any objects in the water), and ray bending (rays of sound always bend towards regions of lower propagation speed). If the water is shallow, propagation will occur in surface–bottom bounces in addition to a possible direct path. If the water is deep, as in the regions past the continental shelves, the sound channel may form by bending of the rays toward the location where the sound speed reaches its minimum, called the *axis of the deep sound channel*. Because there is no loss due to reflections, sound can travel in this way over several thousands of kilometers. Alternatively, the rays bending upward may reach the surface focusing in one point where they are reflected, and the process is repeated periodically. The region between two focusing points on the surface is called a *convergence zone*, and its typical length is 60–100 km.

The geometry of multipath propagation and its spatial dependence are important for communication systems that use array processing to suppress multipath [e.g., 22,23]. The design of such systems is often accompanied by the use of a propagation model for predicting the multipath configuration. Ray theory and the theory of normal modes provide basis for such propagation modeling.

2.3. Time Variation

Associated with each of the deterministic propagation paths (macromultipaths), which can be modeled accurately, are random signal fluctuations (micromultipath), which account for the time variability of the channel response. Some of the random fluctuations can be modeled statistically [2,3]. These fluctuations include surface scattering due to waves, which is the most important contributor to the overall time variability of the shallow-water channel. In deep water, in addition to surface scattering, internal waves contribute to the time variation of the signal propagating along each deterministic path.

Surface scattering is caused by the roughness of the ocean surface. If the ocean were calm, a signal incident on the surface would be reflected almost perfectly, with the only distortion in the form of a phase shift of π. However, wind-driven waves act as the displacement of the reflection point, resulting in signal dispersion. Vertical displacement of the surface can be accurately modeled as a zero-mean Gaussian random variable, whose power spectrum

is completely characterized by the windspeed [2]. Motion of the reflection point results in frequency spreading of the surface-reflected signal, significantly larger than that caused by many other phenomena. Doppler spread of a signal component of frequency f caused by a single surface reflection occurring at an incidence angle θ is $0.0175(f/c)w^{3/2}\cos\theta$, where c is the speed of sound, nominally taken to be 1500 m/s, and w is the windspeed in meters per second [2]. A moderate windspeed is on the order of 10 m/s. Highest Doppler spreads are most likely to be found in short-range links, which use relatively high frequencies. For longer ranges, at which lower frequencies are used, the Doppler spread will be lower; however, multipath spread will increase as there will be more significant propagation paths. The exact values of multipath and Doppler spreads depend on the geometry of multipath on a particular link. Nevertheless, it can be said that the channel spread factor, that is, the product of the Doppler spread and the multipath spread, can in general be expected to decrease with range.

As an example, Figs. 2–4 each show an ensemble of channel impulse responses, observed as functions of delay over an interval of time. These figures describe channel responses obtained at three fundamentally different locations with different mechanisms of multipath formation. Figure 2 shows the impulse responses recorded in deep water of the Pacific Ocean, off the coast of California. In this channel, propagation occurs over three convergence zones, which span 110 nautical miles (nmi). At each fixed time instant, the figure shows a realization of the channel impulse response magnitude as a function of delay. Looking at one channel response reveals that two or more signals arrive at the receiver at any given time. The multipath delay spread in this channel is on the order of 20 ms. The multiple arrivals have comparable energy, thus causing strong ISI. The amplitudes and phases of distinct arrivals may vary independently in time. Along the time axis, variation of the channel response is observed

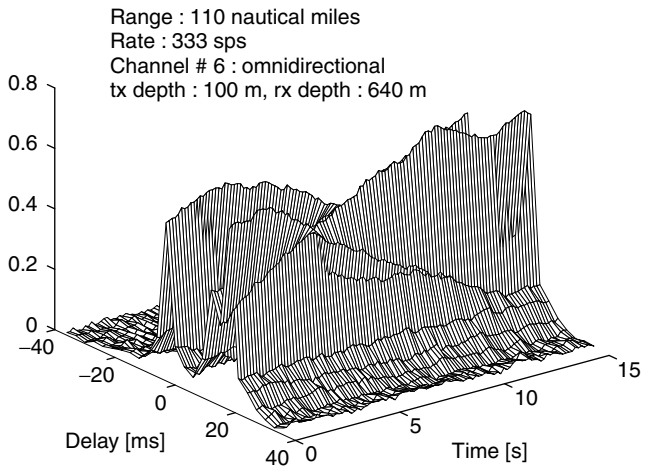


Figure 2. Ensemble of long range channel responses in deep water (~2000 m) off the coast of California, during the month of January. Carrier frequency is 1 kHz. Rate at which quaternary date symbols used for channel estimation were transmitted is given in symbols per second (sps).

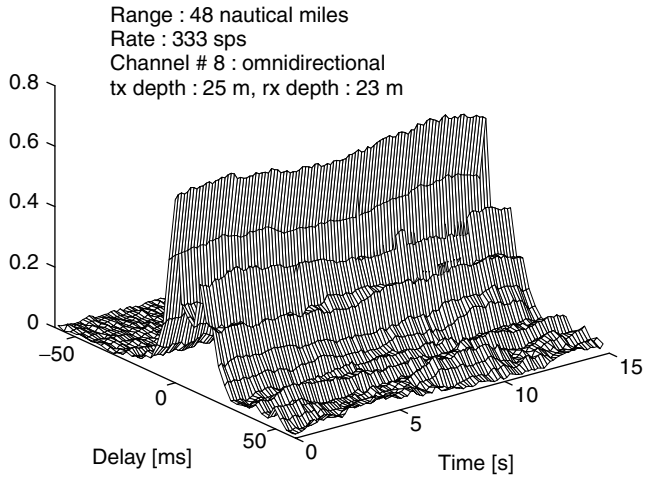


Figure 3. Ensemble of long-range channel responses in shallow water (~50 m) off the coast of New England, during the month of May. Carrier frequency is 1 kHz.

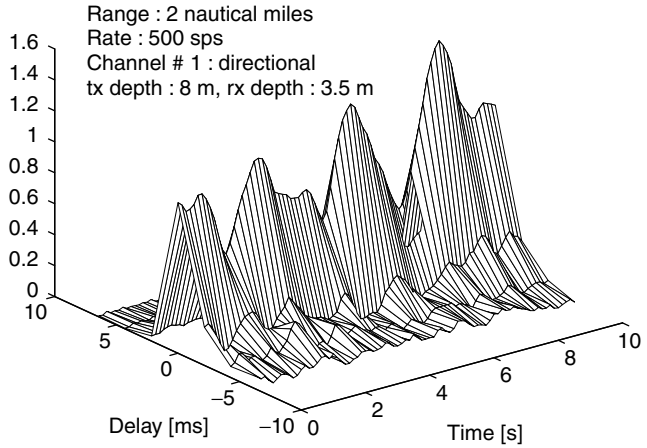


Figure 4. Ensemble of medium-range channel responses in shallow water (~20 m) near the coast of New England, during the month of February. Carrier frequency is 15 kHz.

for each given delay. In this example, significant variation occurs over the shown 15-s interval. This channel does not have a well-defined principal, or strongest, arrival, as evidenced by the fact that the maximum amplitude does not always occur at the same delay. The channel responses shown in Figs. 2–4 are obtained by adaptive channel estimation techniques. In particular, a recursive least-squares algorithm is applied to 4-PSK (phase shift keying) signals transmitted over the channels at rates indicated in the figures. Figure 3 shows the impulse responses obtained in shallow water of the Atlantic Ocean continental shelf, off the coast of New England, over a long distance (48 nmi). This example shows a channel with a well-defined principal arrival, followed by a multipath of lower energy. The extent of multipath is up to 50 ms. It is worth noting that even though the extended multipath may appear to have negligible energy, its contribution to the overall ISI cannot be neglected. This channel shows a slower time-variation than the one observed in Fig. 2. In contrast,

Fig. 4 provides an example of a rapidly time-varying channel. These response were recorded in the shallow water of Buzzards Bay near the coast of New England, over a distance of 2 nmi. Of the three examples shown, this channel demonstrates the fastest time variation, which is typical of a medium-range shallow water environment.

The factor that determines the performance of a digital communication system on a frequency-spread channel is the Doppler spread normalized by the symbol rate. In underwater acoustic channels, the normalized Doppler spread can approach values as high as 10^{-2}. The implications that the time-varying multipath bears on the communication system design are twofold. On one hand, signaling at a high rate causes many adjacent symbols to interfere at the receiver, and requires sophisticated processing to compensate for the ISI. On the other hand, as pulse duration becomes shorter, channel variation over a single symbol interval becomes slower. This allows an adaptive receiver to efficiently track the channel on a symbol-to-symbol basis, providing, of course, a method for dealing with the resulting time dispersion. Hence, time-varying multipath causes a tradeoff in the choice of signaling rate for a given channel. Experimental results obtained on a rapidly varying shallow-water channel [27] demonstrate these observations.

While there exists a vast knowledge of both deterministic and statistical modeling of sound propagation underwater, the use of this knowledge in modeling of communication channels has only recently received more attention [e.g., 8–12]. A time-varying multipath communication channel is commonly modeled as a tapped delay line, with tap spacing equal to the reciprocal of twice the channel bandwidth, and the tap gains modeled as stochastic processes with certain distributions and power spectral densities. While it is known that many radio channels fit well within the model of Rayleigh fading, where the tap gains are derived from complex Gaussian processes, there is no single model accepted to date for any of the underwater acoustic channels. Modeling of the shallow-water medium-range channel has received most attention, as this channel is known to be among the most rapidly varying ones. Most authors consider that this channel is fully saturated, meaning that it exhibits Rayleigh fading [3,5,9]. The deep-water channel has also been modeled as a Rayleigh fading channel; however, the available measurements are scarce, often making channel modeling a controversial issue [10].

The statistical channel measurements available today focus mostly on stationary communication scenarios. In a mobile underwater acoustic channel, vehicle speed will be the primary factor determining the time-coherence properties of the channel, and consequently the system design. Knowledge of a statistical channel model has proved useful in the design and analysis of land-mobile radio systems, and it remains for the future to develop such models for underwater mobile acoustic channels.

3. SYSTEM DESIGN

To overcome the difficulties of time-varying multipath dispersion, the design of commercially available underwater acoustic communication systems has so far relied mostly on the use of noncoherent modulation techniques and signaling methods that provide relatively low data throughput. More recently, phase-coherent modulation techniques, together with array processing for exploitation of spatial multipath diversity, have been shown to provide a feasible means for a more efficient use of the underwater acoustic channel bandwidth. These advancements are expected to result in a new generation of underwater communication systems, with at least an order of magnitude increase in data throughput.

Approaches to system design vary according to the technique used for overcoming the effects of intersymbol interference and signal phase variations. Specifically, these techniques may be classified according to (1) the signal design (i.e., the choice of modulation/detection method) and (2) the transmitter/receiver structure (i.e., the choice of array processing method and the equalization method, if any). In this section, the design of several systems that have been implemented is described. While most of the existing systems operate on the vertical, or the very short-range channels, the systems under development often focus on the severely spread horizontal shallow-water channels. Signal processing methods used in these systems are addressed in the following section.

3.1. Systems Based on Noncoherent Modulation

Noncoherent detection of FSK (frequency shift keying) signals has been used for channels exhibiting rapid phase variation such as the shallow-water long-range and medium-range channels. To overcome the ISI, the existing noncoherent systems employ signal design with guard times, that are inserted between successive pulses to ensure that all the reverberation will vanish before each subsequent pulse is to be received. The insertion of idle periods of time obviously results in a reduction of the available data throughput. In addition, because fading is correlated among frequencies separated by less than the coherence bandwidth (the inverse of the multipath spread), it is desired that only those frequency channels that are separated by more than the coherence bandwidth be used at the same time. This requirement further reduces the system efficiency unless some form of coding is employed so that the adjacent, simultaneously transmitted frequencies belong to different codewords. A representative system [13] for telemetry at a maximum of 5 kbps uses a multiple FSK modulation technique in the 20–30-kHz band. This band is divided into 16 subbands, in each of which a 4-FSK signal is transmitted. Hence, out of a total of 64 channels, 16 are used simultaneously for parallel transmission of 32 information bits (2 information bits per one 4-channel subband). This system has successfully been used for telemetry over a 4-km shallow-water horizontal path, and a 3-km deep-ocean vertical path. It was also used on a <1 km long shallow-water path, where probabilities of bit error on the order of 10^{-2}–10^{-3} were achieved without coding. The system performance may be improved by using error-correction coding (ECC); however, its data throughput will be reduced. This multiple FSK system is commercially available with a maximum data rate of 1200 bps (bits per second). Although bandwidth efficiency of this system does

not exceed 0.5 bps/Hz, noncoherent FSK is a good solution for applications where moderate data rates and robust performance are required. An improved FSK system [14] uses 128 subbands and employs coding. The essence of its coding method is a Hadamard H(20,5) code, in which each 5 input bits are encoded into 20 output bits (the minimum distance of this code is 10). The encoded bits dictate the choice of active subbands for transmission of the given codeword. The 20 subbands that are simultaneously used are chosen (among the 128 available) to be maximally separated, which ensures the least correlated fading, and thus provides diversity on time-varying underwater channels. Because of their robustness and simplicity of implementation, the noncoherent signaling methods are being further developed, and a system has been implemented [15] that uses orthogonal frequency-division multiplexing (OFDM) realized with DFT (discrete-time Fourier transform)-based filter banks. This system was used on a medium-range channel; however, because of the high-frequency separation among the channels (only every fourth channel is used) and relatively long guard times (10-ms guard following a 30-ms pulse), needed to compensate for the multipath fading distortion, the effective data rate is only 250 bps.

3.2. Systems Based on Differentially Coherent and Coherent Modulation

With the goal of increasing the bandwidth efficiency of an underwater acoustic communication system, research focus has shifted toward phase-coherent modulation techniques, such as PSK (phase shift keying) and QAM (quadrature amplitude modulation). Phase-coherent communication methods, previously considered infeasible, were demonstrated to be a viable way of achieving high-speed data transmission over many of the underwater channels, including the severely time-spread horizontal shallow-water channels [24–27]. These methods have the capability to provide raw data throughputs that are an order of magnitude higher than those of the existing noncoherent systems.

Depending on the method for carrier synchronization, phase-coherent systems fall into two categories: differentially coherent and purely phase-coherent. The advantage of using differentially encoded PSK (DPSK) with differentially coherent detection is the simple carrier recovery it allows; however, it has a performance loss as compared to coherent detection. Most of the existing systems employ DPSK methods to overcome the problem of carrier phase extraction and tracking. Real-time systems have been implemented mostly for application in vertical and very short-range channels, where little multipath is observed and the phase stability is good.

In the very short-range channel, where bandwidth in excess of 100 kHz is available, and signal stability is good, a representative system [16] operates over 60 m at a carrier frequency of 1 MHz and a data rate of 500 kbps. This system is used for communication with an undersea robot that performs maintenance of a submerged platform. A 16-QAM modulation is used, and the performance is aided by an adaptive equalizer. A linear equalizer, operating under a least-mean-squares (LMS) algorithm suffices to reduce the bit error rate from 10^{-4} to 10^{-7} on this channel.

A deep-ocean, vertical-path channel is used by an image transmission system [17]. This is 4-DPSK system with carrier frequency 20 kHz, capable of achieving 16 kbps bottom–surface transmission over 6500 m. The field tests of this system indicate the achievable bit error rates on the order of 10^{-4} with linear equalizer operating under an LMS algorithm.

Another example of a successfully implemented system for vertical-path transmission is that of an underwater image and data transmission system [29]. This system uses a binary DPSK modulation at a rate of 19.2 kbps. The carrier frequency of 53 kHz was used for transmission over 2000 m.

More recent advances in digital underwater speech transmission are represented by a prototype system described in Ref. 19. This system uses a code-excited linear prediction (CELP) method to transmit the speech signal at 6 kbps. The modulation method used is 4-DPSK. A decision-feedback equalizer, operating under LMS algorithm is being used in the pool tests. Field tests have not been reported yet. A similar approach has been considered [20].

For applications in shallow-water medium-range channel, a binary DPSK system [21] uses a direct-sequence spread-spectrum (DSSS) method to resolve a strong surface reflection observed in the 1-km-long, 10-m-deep channel. The interfering reflection is only rejected, and not used for multipath recombining. Data throughput of 600 bps within a bandwidth of 10 kHz is achieved. Such high spreading ratios are justified in interference-suppression applications.

Current state of the art in phase-coherent underwater communications is represented by the system described by Johnson et al. [30]. This system is based on purely phase-coherent modulation and detection principles [24] of 4-PSK signals. The signals are transmitted at 5 kbps, using a carrier frequency of 15 kHz. The system's real-time operation in configuration as a six-node network was demonstrated in the under-ice shallow-water environment. To overcome the ISI caused by shallow-water multipath propagation, the system uses a decision feedback equalizer operating under an RLS (recursive least squares) algorithm.

4. SIGNAL PROCESSING METHODS FOR MULTIPATH COMPENSATION

To achieve higher data rates, bandwidth-efficient systems based on phase-coherent signaling methods must allow for considerable ISI in the received signal. These systems employ either some form of array processing, equalization methods, or a combination thereof, to compensate for the distortions. Three main approaches have been taken toward this end. The first two approaches use differentially coherent detection and rely on array processing to eliminate, or reduce, multipath. The third approach is based on purely phase-coherent detection and the use of equalization together with array processing for exploitation of the multipath and spatial diversity.

Array processing for multipath suppression has been used at both the transmitter and receiver ends. Transmitter arrays can be used to excite only a single path of propagation, but very large arrays are required. To overcome the need for a large array, the use of parametric sources has been studied extensively [22]. These highly directive sources rely on the nonlinearity of the medium in the vicinity of a transducer where two or more very high frequencies from the primary projector are mixed. The resulting difference frequency is transmitted by a virtual array formed in the water column in front of the projector. A major limitation of such a source is in its high power requirements. High directivity implies the problem of pointing errors, and careful positioning is required to ensure complete absence of multipath. These systems have been employed in shallow-water channels where equalization is not deemed feasible because of rapid time variation of the signal. Instead, a receiving array is employed to compensate for the possible pointing errors. Binary and quaternary DPSK signals were used achieving data rates of 10 and 20 kbps, respectively, with a carrier frequency of 50 kHz. The estimated bit error rate was on the order 10^{-2}–10^{-3}, depending on the actual channel length. In general, the technique was found to be more effective at shorter ranges.

Multipath rejection using adaptive beamforming at the receiver end only is another possibility. The beamformer [23] uses an LMS algorithm to adaptively steer nulls in the direction of a surface-reflected wave. Similarly as in the case of the transmitter array, it was found that the beamformer encounters difficulties as the range increases relative to depth. To compensate for this effect, the use of an equalizer was considered to complement the performance of the beamformer. The equalizer operates under an LMS algorithm whose low computational complexity permits real-time adaptation at the symbol rate. A separate waveform is transmitted at twice the data rate for purposes of time synchronization. The system was tested in shallow-water at 10 kbps, using a carrier frequency of 50 kHz, and showed the estimated bit error rate of 10^{-2} without, and 10^{-3} with, the equalizer.

A different method, based on purely phase-coherent detection, uses joint synchronization and equalization for combating the effect of phase variations and ISI [24,25]. The equalization method is that of fractionally spaced decision feedback equalization, used with an RLS algorithm. The system incorporates spatial signal processing in the form of multichannel equalization based on diversity combining. The phase-coherent methods have been tested in a variety of underwater channels with severe multipath, showing satisfactory performance regardless of the link geometry. The achieved data rates of up to 2 kbps over long-range channels, and up to 40 kbps over shallow-water medium-range channels, are among the highest reported to date. These methods are discussed in more detail below.

4.1. Design Example: Multichannel Signal Processing for Coherent Detection

In many of the underwater acoustic channels multipath structure may exhibit one or more components that carry the energy similar to that of the principal arrival. As the time progresses, it is not unusual for these components to exceed in energy the principal arrival (e.g., see Fig. 2). The fact that the strongest multipath component may not be well defined makes the extraction of carrier reference a difficult task in such a channel. To establish coherent detection in the presence of strong multipath, a technique based on simultaneous synchronization and multipath compensation may be used [24]. This technique is based on joint estimation of the carrier phase and the parameters of a decision feedback equalizer, where the optimization criterion is minimization of the mean-squared error (MSE) in the data estimation process. In addition, the equalizer/synchronizer structure can be extended to include a number of input array channels [25,26]. Spatial diversity combining has shown superior performance in a number of channels, as well as potentials for dealing with several types of interference. In Fig. 5,

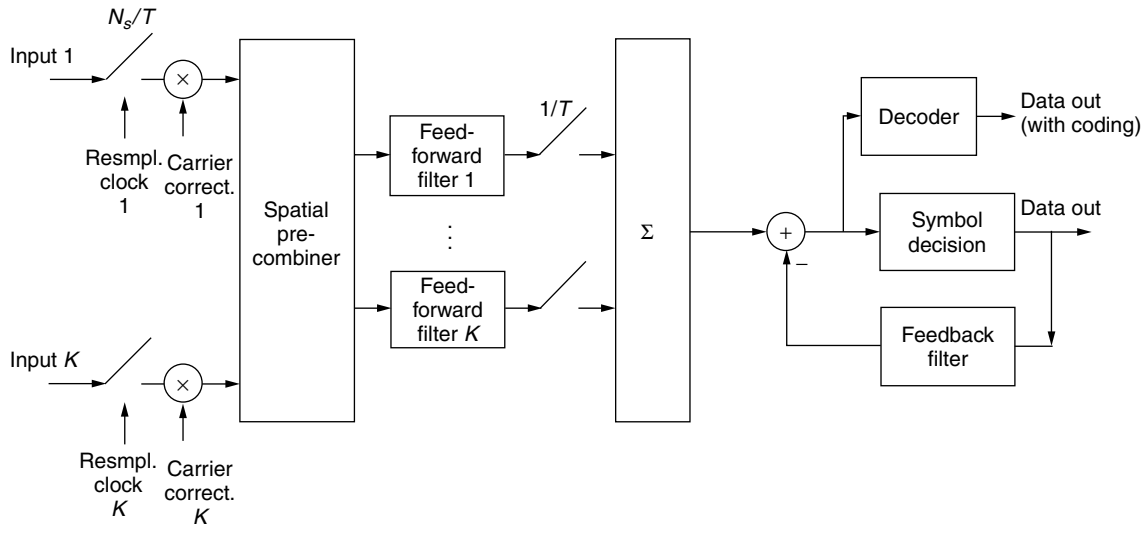


Figure 5. A multichannel receiver for phase-coherent detection.

the multichannel equalizer is shown, preceded by an additional precombiner, which may or may not be used depending on the application and the number of available received channels.

The input signals to the baseband processor are the A/D (analog/digital)-converted array signals, brought to baseband using nominal carrier and lowpass filtering. The signals are frame-synchronized using a known channel probe (usually a short Barker sequence transmitted in phase and quadrature at the data rate). Baseband processing begins with downsampling, which may be carried out to an interval of as few as 2 samples per symbol ($N_s = 2$), since the signals are shaped at the transmitter to have a raised-cosine spectrum that limits their maximal frequency to less than $1/T$. Since there is no feedback to the analog part of the receiver, the method is suitable for an all-digital implementation.

For applications where transmitter and receiver are not moving, but only drifting with water, no explicit adjustment of the sampling clock is needed. This will implicitly be accomplished during the process of adaptive fractionally spaced equalization. The front section of the equalizer will also perform adaptive matched filtering and linear equalization. To correct for the carrier offset, the signals in all channels are phase-shifted by the amount estimated in the process of joint equalization and synchronization. After coherent combining, the ISI resulting from the previously transmitted symbols (postcursors) is canceled in the feedback section of the equalizer. This receiver structure is applicable to any linear modulation format, such as M-PSK, or M-QAM; the only difference is in the way in which symbol decision is performed.

In addition to combining and equalization, signal processing at the receiver includes the operation of decoding if the signal at the transmitter was encoded. For example, in a DSP implementation of the receiver [28] two coding methods are used: concatenated coding of an outer Reed–Solomon code and an inner cyclic block code (Hamming, BCH), and punctured convolutional coding with interleaving. Alternatively, trellis coded modulation, compatible with PSK and QAM signals, provides an effective means of improving performance on a band-limited channel.

The receiver parameters that are adaptively adjusted are the weights of the precombiner, the tap weights of the feedforward filters, the carrier phase estimates, and the tap weights of the feedback filter. A single estimation error is used for the adaptation of all parameters. This error is the difference between the estimated data symbol at the input to the decision device and its true value. During the initial training mode, the true data symbols are known. After the training period, when the receiver parameters have converged, the online symbol decisions are fed back to the equalizer and used to compute the error. The adaptive algorithm used to update the receiver parameters is a combination of the second-order digital phase-locked loop (PLL) for the carrier phase estimates, and the RLS algorithm for the multichannel equalizer tap weights. The complexity of the multichannel equalizer grows with the number of receiver array sensors. For this reason, the spatial precombiner may be used to limit the number of equalizer channels, but still make use of the diversity gain. The precombiner weights can be estimated jointly with the rest of adjustable parameters. The details of the joint adaptation are given in a 1995 paper [26].

The receiver is adaptively adjusted to coherently combine the multiple signal arrivals and thus exploit both spatial and temporal, or multipath, diversity gain. In this manner, it differs from a receiver based on adaptive beamforming that is adjusted to null out the signal replicas arriving from angles different from those of the desired path. The signal isolated by a beamformer usually has to be processed by a separately optimized equalizer to compensate for the residual ISI that arises because the beamformer cannot completely eliminate the multipath interference. Since it is not constrained by angular resolution, the method of multichannel equalization may be used with as few as two input channels, and is applicable to a variety of underwater acoustic channels, regardless of the range : depth ratio. In applications where large arrays are available, the precombiner reduces receiver complexity, while preserving the multichannel diversity gain.

The method of adaptive multichannel combining and equalization was demonstrated to be effective in underwater channels with fundamentally different mechanisms of multipath formation. Experimental results include data rates of 2 kbps over three convergence zones (200 km or 110 nmi) in deep water; 2 kbps over 90 km (50 nmi) in shallow water, and up to 40 kbps over 1–2 km in rapidly varying shallow-water channels [7].

5. ACTIVE RESEARCH TOPICS

At this stage in the development of underwater acoustic communication techniques, with the feasibility of high-rate communications established, a number of research topics are foreseen that will influence the development of future systems. These topics include reduced-complexity receiver structures and algorithms suitable for real-time implementation, techniques for interference suppression, multiuser underwater communications, system self-optimization, development of modulation/coding methods for improved bandwidth efficiency, and mobile underwater acoustic communication systems.

5.1. Reducing the Receiver Complexity

Although the underwater acoustic channels are generally confined to low data rates as compared to many other communication channels, the encountered channel distortions require complex signal processing methods, resulting in high computational load that may exceed the capabilities of the available programmable DSP platforms. Consequently, reducing the receiver complexity to enable efficient real-time implementation has been a focus of many studies.

The problem of reducing the receiver complexity may be addressed on two levels: the design of an efficient receiver structure and the design of an efficient adaptive algorithm. For application to time-varying channels,

the receiver — whether it is based on array processing, equalization, or both methods — must use an adaptive algorithm for adjusting its parameters. Two commonly used types of algorithm are based on the LMS and the RLS estimation principles.

In a majority of the more recent studies, the LMS-based algorithms are considered as the only alternative because of their low computational complexity, which is linear in the number of coefficients N [20,23,33]. However, the LMS algorithm has a convergence time that may become unacceptably long when large adaptive filters are used ($20N$ as opposed to $2N$ of the RLS algorithm). The total number of coefficients N may be very large (more than 100 taps is often needed for spatial and temporal processing in medium and long-range shallow-water channels). In addition, the LMS algorithm is very sensitive to the choice of step size. To overcome this problem, self-optimized LMS algorithms may be used [33], but this results in increased complexity, and increased convergence time.

RLS algorithms, on the other hand, have better convergence properties but higher computational complexity. The quadratic complexity of the standard RLS algorithm is too high when large adaptive filters need to be implemented. In general, it is desirable that the algorithm be of linear complexity, a property shared by the fast RLS algorithms. A numerically stable fast RLS algorithm [31] has been used for the multichannel equalizer [25]. Despite its quadratic complexity, a square-root RLS algorithm [32] has been used for real-time implementation [30]. The advantage of this algorithm is that it allows the receiver parameters to be updated only periodically, rather than every symbol interval, thus reducing the computational load per each detected symbol. In addition, the updating intervals can be determined adaptively, based on monitoring the mean-squared error. Such adaptation methods are especially suitable for use with high transmission rates, where long ISI requires large adaptive filters, but eliminates the need to update the receiver parameters every symbol interval. The square-root RLS algorithm has excellent numerical stability, which makes it a preferable choice for a practical implementation. A different class of adaptive filters, which also have the desired convergence properties and numerical stability, are the lattice filters that use RLS algorithms. These algorithms have been proposed [34], but have not yet been applied to underwater acoustic channel equalization. Choosing an appropriate receiver adaptation method is expected to receive more attention in the future acoustic modem design.

Regardless of the adaptive algorithm used, its computational complexity is proportional to the number of receiver parameters (tap weights). Rather than focusing on low-complexity algorithms only, one may search for a way to reduce the receiver size. Although the use of spatial combining reduces residual ISI and allows shorter-length equalizers to be used, a broadband combiner may still require a large number of taps to be updated, limiting the practical number of receiving channels to only a few. The use of a precombiner [26] is a method for reducing a large number of input channels to a smaller number for subsequent multichannel equalization. By careful design, full diversity gain can be preserved by this technique. More than one channel at the output of the combiner is usually required, but this number is often small (e.g., three). The fact that diversity gain may be preserved is explained by multipath correlation across the receiver array. In addition to the reduced computational complexity, smaller adaptive filters result in less noise enhancement, contributing to improved performance.

A different approach in the design of reduced-complexity receiver structures has been investigated [35], where the focus is on reducing the number of equalizer taps. A conventional equalizer is designed to span all the channel responses. However, if the channel is characterized by several distinct multipath arrivals separated in time by intervals of negligible reverberation, an equalizer may be designed to have fewer taps. By reducing the number of adaptively adjusted parameters, this approach also makes it possible to use simple updating algorithms, such as standard RLS algorithms, which have good numerical stability. Finally, in channels that are naturally sparse, discarding the low-magnitude equalizer taps in fact results in improved performance since no unnecessary noise is processed.

5.2. Interference Cancelation

The sources of interference in underwater acoustic channels include external interference and internal interference, generated within the system. The external sources of interference include noise coming from onboard machinery or other nearby acoustic sources, as well as the propulsion and flow noise associated with the underwater vehicle launch process. The internal noise, which has signal-like characteristics, arises in the form of echo in full-duplex systems, and in the form of multiple-access interference generated by other users operating within the same network.

Methods for cancelation of interference in the form band-limited white noise and multiple sinusoidal interference have been investigated [36]. It was found that the multichannel receiver of Fig. 5 was most effective in canceling the interference while simultaneously detecting the desired signal. Noise cancellation is performed simply by providing a reference of the noise signal to one of the multichannel combiner inputs, while cancellation of the sinusoidal interferer may be performed even without the reference signal. By virtue of having the training sequence, the multichannel combiner is able to adaptively filter the interfering signal out, and extract the desired signal.

5.3. Multiuser Communications and Underwater Networks

A multiple-access communication system represents a special case of structured interference environment. Because of the bandwidth limitation of the underwater acoustic channel, frequency-division multiple access (FDMA) may not be an efficient technique. Time-division multiple access (TDMA) is associated with the problem of efficient time-slot allocation, which arises because of the long propagation delays. A possible solution in such a situation is to allow a number of users to transmit simultaneously in both

time and frequency. The receiver then has to be designed to deal with the resulting multiple-access interference, which may be very strong in an underwater acoustic network. The fact that transmission loss varies significantly with range, and that only very low code-division processing gains are available as a result of bandwidth constraints, both contribute to the enhanced near–far effect in the underwater acoustic channel. The multiuser detection methods suitable for underwater acoustic channels rely on the principles of joint synchronization, channel equalization, and multiple-access interference cancellation [37]. Two categories of multiuser receivers that have been considered are the (1) *centralized receiver*, in which the signals of all the users are detected simultaneously (e.g., uplink reception at a surface buoy, which serves as a central network node), and (2) the *decentralized receiver*, in which only the desired user's signal needs to be detected (e.g., downlink reception by an ocean-bottom node). Similarly as in the case of interference cancellation, the adaptive multichannel receiver of Fig. 5 was experimentally shown to have excellent capabilities in the role of a decentralized multiuser detector, operating without any knowledge of the interfering signal. Array processing plays a crucial role in the detection of multiuser signals, but is associated with the problem of computational complexity.

The advancements in point-to-point communication links have sparked an interest in the development of underwater acoustic communication networks. In addition to the choice of a multiple-access strategy, network design has been addressed on the levels of the data-link layer and the network layer [layers 2 and 3, respectively, of the seven-layer OSI (Open Systems Interconnection) reference model] [8,38]. Typically, packet transmission in a store-and-forward network is considered, and the design of automatic repeat request (ARQ) protocols and routing protocols is influenced by the long propagation times in the underwater channels. Underwater acoustic networks are a young area of research that is only awaiting new developments.

5.4. System Self-Optimization

A receiver algorithm must use a number of parameters that need to be adjusted according to the instantaneous channel conditions before the actual signal detection can begin. These parameters include the number and location of array sensors that provide good signal quality, the sizes of the equalizer filters, and their tracking parameters. The optimal values of receiver parameters depend not only on the general link configuration and location but also on the time of operation. In addition, an increase the background noise level, caused, for example, by a passing ship, may temporarily disable the communication. If the adaptive receiver algorithms are to be used in autonomous systems, external assistance in algorithm initialization, or reinitialization should be minimized. For this reason, the development of self-optimized receiver algorithms is of interest to future research.

The first steps in this direction are evident in the implementation of self-optimized LMS algorithms [23,33], in which the step size is adaptively adjusted, and the periodically updated RLS algorithm [30], self-adjusted to keep a predetermined level of performance by increasing the tracking rate if the channel condition worsens. These strategies provide the receiver with the capability to adjust to the fine channel changes. However, they depend on the availability of a reliable reference of the desired signal. Since a training sequence is inserted only so often in the transmitted signal, a loss of synchronization or convergence during detection of a data packet will cause the entire packet to be lost. An alternative to periodic reinsertion of known data, which increases the overhead, methods for self-optimized, or blind, recovery may be considered.

A blind equalization method based on using the cyclostationary properties of oversampled received signals [39], which requires only the estimation of second-order signal statistics, provides a practical solution for recovering the data sequence in the absence of clock synchronization. Originally developed for linear equalizers, this method has been extended to the case of the decision feedback equalizer, necessary for application in underwater acoustic channels with extreme multipath. These methods have proven successful in preliminary tests with real data [7]. Blind decision feedback equalization for application to underwater acoustic channels has also been investigated [40]. Further work on blind system recovery for underwater acoustic channels will focus on methods for array processing and carrier phase tracking.

5.5. Modulation and Coding

Coding techniques are known to be one of the most powerful tools for improving the performance of digital communication systems on both the additive white Gaussian noise channels and the fading channels. Several well-known techniques have been used for underwater communications with both noncoherent and coherent detection. Turbo codes are also being considered for use in underwater communications. While the performance of various codes is known on Gaussian noise channels and fading channels that can be described by Rayleigh or Rice statistics, it is not known as well on underwater acoustic channels. Future work should provide experimental results necessary for a better understanding of the performance of coded systems on these channels.

Achieving high throughputs over band-limited underwater acoustic channels is conditioned on the use of bandwidth-efficient modulation and coding techniques [41]. Related results documented in contemporary literature are confined to signaling schemes whose bandwidth efficiency is at most 3–4 bps/Hz. Higher-level signal constellations, together with trellis coding, are being considered for use in underwater acoustic communications. While trellis-coded modulation is well suited for vertical channels that have minimal dispersion, their use on the horizontal channels requires further investigation. In the first place, conventional signal mapping into a high-level PSK or QAM constellation may be associated with increased sensitivity of detection on a time-varying channel. Certain results in radio communications show that certain types of high-level constellations are more robust to the channel fading and phase variations than are the conventional rectangular QAM constellations [42]. Another

issue associated with the use of coded modulation on the channels with long ISI is the receiver design that takes full advantage of the available coding gain. Namely, the delay in decoding poses problems for an adaptive equalizer that relies on the feedback of instantaneous decisions. Receiver structures that deal with this problem as it applies to underwater channels are a subject of current studies.

In addition to bandwidth-efficient modulation and coding techniques, the future underwater communication systems will rely on data compression algorithms to achieve high data rates over severely band-limited underwater acoustic channels. This is another active area of research, which, together with sophisticated modulation and coding techniques, is expected to provide solutions for high-rate underwater image transmission.

5.6. Mobile Underwater Communications

The problem of channel variability, already present in applications with a stationary transmitter and receiver, becomes a major limitation for the mobile underwater acoustic communication system. The ratio of the vehicle speed to the speed of sound ($1/10^3$ for a vehicle speed of 30 knots or 54 km/h) often exceeds its counterpart in the mobile radio channels ($1/10^8$ for a mobile moving at 60 mi/h or 100 km/h), making the problem of time synchronization very difficult in the underwater acoustic channel. Apart from the carrier phase and frequency offset, the mobile underwater acoustic systems will have to deal with the motion-induced pulse compression and dilation (time scaling). Successful missions of experimental AUVs that use commercial FSK acoustic modems for vehicle-to-vehicle communication have been reported [43]. In a coherent acoustic modem, a method based on estimating the time-scaling factor from a signal preamble has been implemented and successfully demonstrated in operation with a remotely controlled underwater vehicle [44]. Rather than estimating the motion-induced distortion on a packet-to-packet basis, algorithms for continuous tracking of the time-varying symbol delay in the presence of underwater multipath are under development. One approach is based on a model that relates the instantaneous vehicle speed to the signal phase distortion. Using this relationship and the phase estimate from the PLL, the vehicle speed is calculated, and the corresponding time-scaling factor is used to resample the received signal before equalization. The resampling operation is efficiently implemented using polyphase filters. Other approaches are possible that do not rely on explicit estimation of the vehicle speed to perform adaptive resampling for highly mobile communication scenarios.

While many problems remain to be solved in the design of high-speed acoustic communication systems, more recent advances in this area serve as an encouragement for future work, which should facilitate remote exploration of the underwater world.

BIOGRAPHY

Milica Stojanovic graduated from the University of Belgrade, Belgrade, Yugoslavia, in 1988 and received her M.S. and Ph.D. degrees in electrical engineering from Northeastern University, Boston, Massachusetts, in 1991 and 1993. She is a principal research scientist at the Massachusetts Institute of Technology and a guest investigator at the Woods Hole Oceanographic Institution. Her research interests include digital communications theory and statistical signal processing, and their applications to wireless communication systems.

BIBLIOGRAPHY

1. R. Coates, *Underwater Acoustic Systems*, Wiley, New York, 1989.
2. L. Brekhovskikh and Y. Lysanov, *Fundamentals of Ocean Acoustics*, Springer, New York, 1982.
3. S. Flatte, ed., *Sound Transmission through a Fluctuating Ocean*, Cambridge Univ. Press, Cambridge, UK, 1979.
4. A. Quazi and W. Konrad, Underwater acoustic communications, *IEEE Commun. Mag.* 24–29 (1982).
5. J. Catipovic, Performance limitations in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **15**: 205–216 (1990).
6. A. Baggeroer, Acoustic telemetry — an overview, *IEEE J. Ocean. Eng.* **9**: 229–235 (1984).
7. M. Stojanovic Recent advances in high rate underwater acoustic communications, *IEEE J. Ocean. Eng.* **21**: 125–136 (1996).
8. D. Kilfoyle and A. Baggeroer, The state of the art in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **25**: 4–27 (2000).
9. R. Owen, B. Smith, and R. Coates, An experimental study of rough surface scattering and its effects on communication coherence, *Proc. Oceans'94*, 1994, Vol. III, pp. 483–488.
10. A. Essebbar, G. Loubet, and F. Vial, Underwater acoustic channel simulations for communication, *Proc. Oceans'94*, 1994, Vol. III, pp. 495–500.
11. A. Falahati, B. Woodward, and S. Bateman, Underwater acoustic channel models for 4800 b/s QPSK signals, *IEEE J. Ocean. Eng.* **16**: 12–20 (1991).
12. C. Bjerrum-Niese, L. Bjorno, M. Pinto, and B. Quellec, A simulation tool for high data-rate acoustic communication in a shallow-water, time-varying channel, *IEEE J. Ocean. Eng.* **21**: 143–149 (1996).
13. J. Catipovic, M. Deffenbaugh, L. Freitag, and D. Frye, An acoustic telemetry system for deep ocean mooring data acquisition and control, *Proc. Oceans'89*, 1989, pp. 887–892.
14. K. Scussel, J. Rice, and S. Merriam, A new MFSK acoustic modem for operation in adverse underwater channels, *Proc. Oceans'97*, 1997, Vol. I, pp. 247–254.
15. S. Coatelan and A. Glavieux, Design and test of a multicarrier transmission system on the shallow water acoustic channel, *Proc. Oceans'94*, 1994, Vol. III, pp. 472–477.
16. A. Kaya and S. Yauchi, An acoustic communication system for subsea robot, *Proc. Oceans'89*, 1989, pp. 765–770.
17. M. Suzuki and T. Sasaki, Digital acoustic image transmission system for deep sea research submersible, *Proc. Oceans'92*, 1992, pp. 567–570.
18. D. Hoag, V. Ingle, and R. Gaudette, Low-bit-rate coding of underwater video using wavelet-based compression algorithms, *IEEE J. Ocean. Eng.* **22**: 393–400 (1997).

19. A. Goalic et al., Toward a digital acoustic underwater phone, *Proc. Oceans'94*, 1994, Vol. III, pp. 489–494.

20. B. Woodward and H. Sari, Digital underwater voice communications, *IEEE J. Ocean. Eng.* **21**: 181–192 (April 1996).

21. J. Fischer et al., A high rate, underwater acoustic data communications transceiver, *Proc. Oceans'92*, 1992, pp. 571–576.

22. R. F. W. Coates, M. Zheng, and L. Wang, BASS 300 PARACOM: A model underwater parametric communication system, *IEEE J. Ocean. Eng.* **21**: 225–232 1996.

23. G. S. Howe et al., Sub-sea remote communications utilising an adaptive receiving beamformer for multipath suppression, *Proc. Oceans'94*, 1994, Vol. I, pp. 313–316.

24. M. Stojanovic, J. A. Catipovic, and J. G. Proakis, Phase coherent digital communications for underwater acoustic channels, *IEEE J. Ocean. Eng.* **19**: 100–111 (1994).

25. M. Stojanovic, J. A. Catipovic, and J. G. Proakis, Adaptive multichannel combining and equalization for underwater acoustic communications, *J. Acoust. Soc. Am.* **94**(3)(Pt. 1): 1621–1631 (1993).

26. M. Stojanovic, J. A. Catipovic, and J. G. Proakis, Reduced-complexity multichannel processing of underwater acoustic communication signals, *J. Acoust. Soc. Am.* **98**(2)(Pt. 1): 961–972 (1995).

27. M. Stojanovic, J. G. Proakis, and J. A. Catipovic, Performance of a high rate adaptive equalizer on a shallow water acoustic channel, *J. Acoust. Soc. Am.* **100**(4)(Pt. 1): 2213–2219 (1996).

28. L. Freitag, M. Grund, S. Singh, and M. Johnson, Acoustic communication in very shallow water: Results from the 1999 AUV Fest, *Proc. Oceans'00*, 2000.

29. G. Ayela, M. Nicot, and X. Lurton, New innovative multimodulation acoustic communication system, *Proc. Oceans'94*, 1994, Vol. I, pp. 292–295.

30. M. Johnson, D. Herold, and J. Catipovic, The design and performance of a compact underwater acoustic network node, *Proc. Oceans'94*, 1994, Vol. III, pp. 467–471.

31. D. Slock and T. Kailath, Numerically stable fast transversal filters for recursive least squares adaptive filtering, *IEEE Trans. Signal Process.* **39**: 92–114 (1991).

32. F. Hsu, Square root Kalman filtering for high-speed data received over fading dispersive HF channels, *IEEE Trans. Inform. Theory* **28**: 753–763 (1982).

33. B. Geller et al., Equalizer for video rate transmission in multipath underwater communications, *IEEE J. Ocean. Eng.* **21**: 150–155 (1996).

34. F. Ling and J. G. Proakis, Adaptive lattice decision-feedback equalizers–their performance and application to time-variant multipath channels, *IEEE Trans. Commun.* **33**: 348–356 (1985).

35. M. Stojanovic, L. Freitag, and M. Johnson, Channel-estimation-based adaptive equalization of underwater acoustic signals, *Proc. Oceans'99*, 1999, pp. 590–595.

36. J. Catipovic, M. Johnson, and D. Adams, Noise canceling performance of an adaptive receiver for underwater communications, *Proc. 1994 Symp. AUV Technology*, 1994, pp. 171–178.

37. M. Stojanovic and Z. Zvonar, Multichannel processing of broadband multiuser communication signals in shallow water acoustic channels, *IEEE J. Ocean. Eng.* **21**: 156–166 (1996).

38. E. Sozer, M. Stojanovic, and J. Proakis, Underwater acoustic networks, *IEEE J. Ocean. Eng.* **25**: 72–83 (2000).

39. L. Tong, G. Xu, and T. Kailath, Blind identification and equalization based on second-order statistics, *IEEE Trans. Inform. Theory* **40**: 340–349 (1994).

40. J. Gomes and V. Barroso, Blind decision-feedback equalization of underwater acoustic channels, *Proc. Oceans'98*, 1998, pp. 810–814.

41. J. Proakis, Coded modulation for digital communications over Rayleigh fading channels, *IEEE J. Ocean. Eng.* **16**: 66–74 (1991).

42. W. T. Webb and R. Steele, Variable rate QAM for mobile radio, *IEEE Trans. Commun.* **43**: 2223–2230 (1995).

43. S. Chappell et al., Acoustic communication between two autonomous underwater vehicles, *Proc. 1994 Symp. AUV Technology*, 1994, pp. 462–469.

44. L. Freitag et al., A bidirectional coherent acoustic communication system for underwater vehicles, *Proc. Oceans'98*, 1998, pp. 482–486.

ACTIVE ANTENNAS

ALAN R. MICKELSON
University of Colorado
Boulder, Colorado

The present article is an introduction to the topic of active antennas. The first section is a description of the field suitable for reading by almost any undergraduate science major. The next section is an in-depth reexamination of the subject, including equations and some derivations. Its basic idea is to provide the readers with enough tools to enable them to evaluate whether it is an active antenna that they might need for a specific application. The final section is a discussion of where active antennas are finding and will find application.

We should mention here that, if one really needs to design active antennas, one will need to go further than this article. The set of references to the primary research literature given in this article is by no means complete, nor is it meant to be. A good way to get started on the current literature on this topic would be a reading of the overview monograph of Navarro and Chang [1]. We will not cover active amplifiers in this article. However, this topic is treated in the book edited by York and Popović [2].

1. AN INTRODUCTION TO ACTIVE ANTENNAS

An antenna is a structure that converts electromagnetic energy propagating in free space into voltage and current in an electric circuit and/or vice versa. In a transceiver system, the antenna is used both to receive and to transmit free-space waves. At minimum, a transceiver then must consist of a signal source that serves to drive the antenna as well as a receiver circuit that reads out the signal from the antenna. Previously, practically all antenna systems operating in the microwave frequency regime (operation frequencies greater than 1 billion cycles per second, or 1 GHz) were designed mostly to isolate the antenna from the circuits — that is, to find ways to make system operation independent of the antenna's

electrical characteristics. In contradistinction, an active antenna is one in which the antenna actually serves as a circuit element of either the driver or the readout circuit. To understand why this is different from conventional antenna driving or readout will require us to take a brief historical trip through the last century or so.

Actually, the first antenna was an active one. Heinrich Hertz, back in 1884 [2a], was the first to demonstrate that one could generate radiowaves and that they would propagate from a transmitter to a receiver at the speed of light. The apparatus used is schematically depicted in Fig. 1. The idea of the transmitter is that, by discharging an induction coil (a wire looped about a magnetic core such that the composite device can store significant amounts of magnetic energy) into a spark gap, one can generate a current in the 5-mm-diameter wire. The voltage in the spark gap induces a current in the wires, which in turn induces a voltage in the wires, and this voltage in turn induces current, so that the voltage and current propagate along the two pieces of the wire to either side of the gap as waves, appearing much like a one-dimensional slice through a water wave propagating away from the point where a pebble has struck the water's surface (the spark gap). A wave will propagate rectilinearly until it encounters an obstruction, at which point it can suffer reflection from or transmission into the barrier that the obstruction presents. There will then be reflections off the metal spheres on the ends of the wire. The spark will generate a broad spectrum of frequencies or wavelengths. The reflections off the two ends, though, will tend to cancel each other except at certain special frequencies. The effect at these wrong frequencies is much like the effect of throwing a handful of pebbles into the pond and noting that, in between the points where the pebbles struck, the waves are much more indistinct than they are far from where the handful struck the surface. The special frequencies are ones which just fit into the region between the spheres. The current needs to be zero at the two ends in order to fit, whereas the voltage needs to be maximum

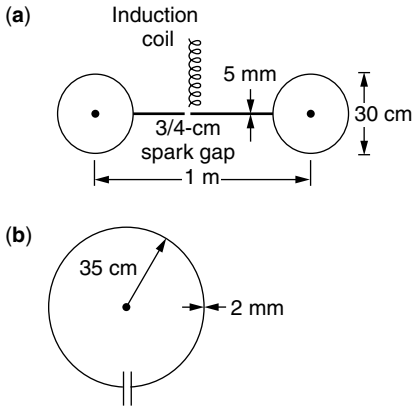


Figure 1. Hertz apparatus for (**a**) transmitting and (**b**) receiving radiowaves, where the transmitting antenna serves to choose a specific frequency of the spark-gap voltage to transmit to the receiving antenna, which also serves to pick out this special frequency from the free-space waveform and turn this electromagnetic disturbance into a voltage across the receiver antenna gap.

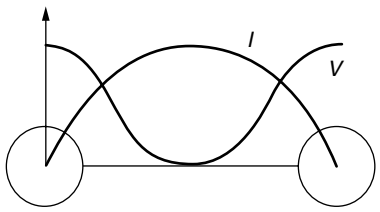


Figure 2. Current and voltage waveforms for the lowest-order (least number of zeros) waveform for the Hertz transmitter of Fig. 1a. The current must go to zero at the points where the wire ends, whereas the potential will be highest there.

at the ends. The current and voltage waves at the right frequencies may appear as depicted in Fig. 2.

The Hertz transmitter is the archetypical active antenna. The source is the spark gap, which is actually placed in the antenna. The antenna then acts as a filter to pick the right frequency out of a large number of frequencies that could be launched from the gap. The receiver is picked to be of a length to also select this primary frequency.

Hertz-style spark-gap transmitters, after further development and popularization by Marconi, were in use for 50 years after Hertz. However, such transmitters exhibit some rather severe drawbacks. The main problem is that the simple resonant dipole antenna (i.e., a straight-wire antenna with a gap or a feeder cable used to feed in current) is a filter with a poor frequency selection. Namely, if one increases the frequency by 50%, there is 75% as much power transmitted at this frequency as at the first resonance, which is called the *fundamental*. There is a second resonance at twice the frequency of the first resonance, and another at each integer multiple of the fundamental. With increasing frequency, the transmitted power decreases a little and then flattens out around the second resonance, decreases a little, flattens out at the third resonance, and so on, as illustrated in Fig. 3. If the spark discharge is really broadband (i.e., if it generates a large number of frequencies where the highest frequency may be many times the lowest), then what is transmitted by the antenna will also be broadband, although with somewhat higher transmission at the fundamental frequency and its harmonics than in between. In the very early days of radio, this was somewhat acceptable, although any information impressed on such a broadband carrier would be rather

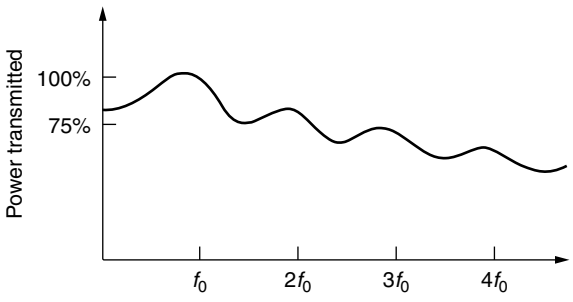


Figure 3. A sketch of what the transmission as a function of frequency might look like for the Hertzian dipole antenna of Figs. 1 and 2.

severely degraded on reception. However, the demise of the spark-gap transmitter was really instigated by the early success of radio, which caused the available frequency bands to begin to fill up rapidly. This band filling led to the formation of the Federal Communications Commission (FCC) in 1934, which was charged with allocation of frequency bands. The allocation by nature led to a ban on spark-gap transmitters, which were needlessly wasting bandwidth.

In a later experiment, Hertz noticed that the waves he was generating would tend to have a component that hugged the ground and could therefore travel over the horizon and, in fact, across the Atlantic Ocean, skimming along the surface of the water. Other researchers noticed that the effect became more pronounced at wavelengths longer than the roughly 2-m wavelength that Hertz originally used. (For the frequencies and wavelengths of some important frequency bands, see Table 1.) In order for wave transmission to be useful, however, the transmitted signal needs to carry information. Impressing information on the wave is called *modulating* the carrier. One can modulate the height (amplitude), the frequency, and so on. The discovery of a technique to *amplitude-modulate* the waves coming off an antenna (in 1906) then led to the inception of AM radio in bands with wavelengths greater than 300 m, which corresponds to roughly 1 MHz. AM radio became commercial in 1920. By the 1930s, other researchers noted that waves with frequencies around 10 MHz, corresponding to a wavelength around 30 m, could be quite efficiently propagated over the horizon by bouncing the wave off the ionosphere. This led to the radio bands known as *shortwave*. In 1939, a researcher realized a technique to modulate the frequency of the wave. This realization led in the 1950s to FM radio, which was allocated the band around 100 MHz with a corresponding wavelength around 3 m. However, the FM technique was used first during World War II as a radar modulation technique. Radars today are at frequencies above roughly 1 GHz or wavelengths below 30 cm.

Table 1. A Listing of the Allocated Microwave and Millimeter-Wave Bands as Defined by the Frequency and Wavelength Range within Each Band

Band Designation	Frequency (GHz)	Wavelength
L	1–2	15–30 cm
S	2–4	7.5–15 cm
C	4–8	3.75–7.5 cm
X	8–12	2.5–3.75 cm
Ku	12–18	1.67–2.5 cm
K	18–26	1.15–1.67 cm
Ka	26–40	0.75–1.15 cm
Q	33–50	6–9 mm
U	40–60	5–7.5 mm
V	50–75	4–6 mm
E	60–80	3.75–5 mm
W	75–110	2.7–4 mm
D	110–170	1.8–2.7 mm
G	140–220	1.4–2.1 mm
Y	220–325	0.9–1.4 mm

There is a fundamental difference between circuits that operate at frequencies whose corresponding wavelengths are less than the maximum circuit dimension and those that are large compared to the carrier wavelength. The effect is closely related to the concept of impedance. As was mentioned above, in the wire antenna, the voltage and current reinforce each other and thereby travel on the antenna as waves. The same effect takes place in a circuit. At any point along the path (line) in a circuit, one defines the ratio of voltage at one frequency to the current at the same frequency as the impedance at that frequency. For a sinusoidal waveform, if the impedance tends to preserve the phase relationship (where the wave peaks lie, relatively), then we say that the impedance is *resistive*. If the impedance tends to drive the voltage peaks forward with respect to the current peaks, we say that the impedance is *capacitive*; in the opposite case we say that the impedance is *inductive*. In a small circuit (small compared to a wavelength), one generally tries to carefully design passive components — resistors, capacitors, and inductors — so that they exhibit large local impedance, that is, large impedance within their physical dimensions. When the circuit is small, one would like to control the phase and amplitude of the wave at discrete points by using lumped elements and thereby minimizing line effects. The lines (wires) between the components have little or no effect on the electromagnetic disturbances passing through the circuit, then, as the impedances in the wires are small and reasonably independent of their lengths. When the circuit is large, the lines themselves effectively become circuit elements, and they themselves must be carefully designed in order to exhibit the proper impedances. To illustrate, consider the parallel-plate capacitor of Fig. 4. The capacitance is maximized by maximizing the permittivity ε (a material parameter equal to the ratio of electrical displacement to applied electric field) and area A while minimizing the plate spacing d. However, the fact that the capacitance depends on the plate spacing d is the important point here. Consider the circuit of Fig. 5 as an example. The only ground in the figure is the one on the battery, but the wires connecting the circuit elements together in essence form at each point a capacitor, with a point on the wire that is carrying charge as the upper plate and the ground as the

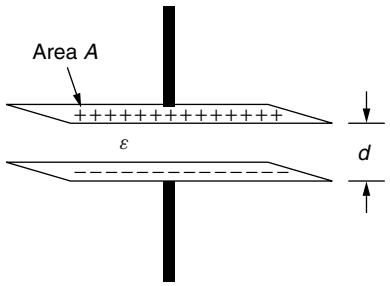


Figure 4. Schematic depiction of a parallel-plate capacitor in which the flow of a current will tend to change the upper plate, causing a voltage difference between upper and lower plates. The capacitance is defined as the ratio of the amount of change of the upper plate to the magnitude of the voltage this change induces between the plates.

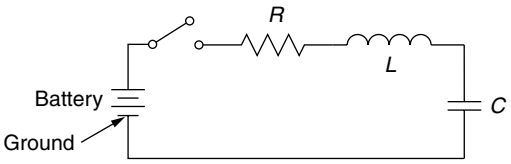


Figure 5. A circuit with lumped elements connected by wire segments.

lower. This capacitance changes as a function of position along the wire. For a small enough circuit (relative to the wavelength of the highest frequency carried by the circuit), the effect is not too important, as the wire–ground pair has small capacitance and the position-varying effect is small. For a large circuit, the effect is disastrous, as we shall consider below. The effect is identical to the effect of Fresnel coefficients in optics.

Consider the circuit of Fig. 6. We will now discuss what happens when impedances are not carefully controlled. This leads to the concept of *impedance matching*. Let us first say that the circuit is short (compared to a wavelength). If the load resistor, R_L, is not matched to (i.e., is not equal to, or, one could say, *not impedance-matched to*) the resistance of the source, R_S, some amount of reflection will occur at R_L, propagate back to R_S, be reflected with a reversal of sign at R_L, propagate back to R_L, and so on. The reflections add up perfectly out of phase (i.e., simply subtract from one another) at the source and load, and the amount of power supplied to the load is less than optimal. In this limit of a small circuit, it is as if the load will not allow the source to supply as much power as it is capable of. Let us now say that the line is "well designed" but long compared to the wavelength used. Then the same argument applies to the reflections, but in this case the source does not know that the load is there until several wave periods have passed (several maxima and minima of the waveform have left the source), so the source supplies all the power it can. The power, though, is not allowed to be fully absorbed by the load, and some of it will rattle around the line until it is radiated or absorbed. As we mentioned above, in a long enough circuit the wire itself becomes a distributed element — that is, one with an impedance of its own. If the distance to the nearest ground is not kept fixed along the line, the inductance and capacitance become dependent on the position. In this case, we have distributed reflections all along the line

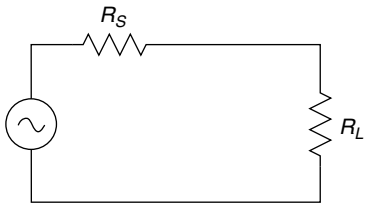


Figure 6. A circuit in which one is trying to supply power from a source with internal resistance R_S to a load with resistance R_L. The power transfer is maximized only when R_S and R_L are equal, in which case half the power supplied by the source is supplied to the load, the other half being dissipated in the source and causing it to heat.

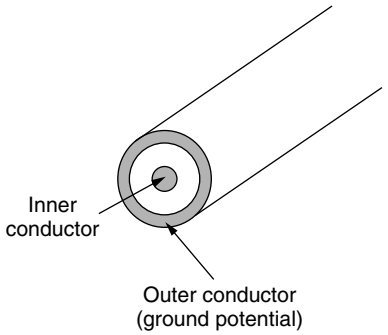


Figure 7. A coaxial cable in which signals are carried on an inner conductor and in which the grounded outer conductor serves to carry the ground plane along with the signal in order to give a constant impedance along the line.

and the circuit will probably not work at all. This spatial variability of the line impedance is remediable, though, as illustrated by the drawing of a coaxial cable in Fig. 7. The idea is that, if the line brings along its own ground plane in the form of a grounded outer conductor, the characteristic impedance of the line can be kept constant with distance. Such a line, which carries its own ground plane, is called a *transmission line*. The problem becomes the connection of the line to the source and load (i.e., impedance matching).

Before going on to discuss the conventional solution versus the new active-antenna solution, perhaps we should summarize a bit. In AM, shortwave, and FM applications, the wavelengths are of order greater than meters. If one considers typical receivers, the whole circuit will generally be small compared to the carrier wavelength. This is also to say that in all of these cases, the antennas will be active in the sense that the antenna presents an impedance to the circuit. (Recall that an active antenna is any antenna in which an active element lies within a wavelength of the antenna and is used as an element to match the antenna impedance to the decoder impedance.) To passively match an antenna to the receiver circuit, one needs pieces of line comparable to a wavelength. However, from here on we shall not be interested in the low-frequency case but rather in the well-above-1-GHz case, as AM, FM, and TV technologies are mature technologies. During World War II, radar was the application that drove the frequencies above 1 GHz (wavelength less than 30 cm). In a radar, one sends out a pulse and, from the returned, scattered wave, tries to infer as much as possible about the target. Target resolution is inversely proportional to wavelength. There has been a constant drive to shorten wavelength. Therefore as is indicated by Table 1, bands have been allocated out to hundreds of gigahertz. Presently, however, there are a plethora of nonmilitary drivers for pushing to higher-frequency communication systems that are compact and have lower power dissipation. However, the conventional solution, which was developed originally for radars, is really not conducive to compactness or to the pressures of cost minimization of the commercial market.

A typical conventional transmitter is schematically depicted in Fig. 8. A main concept here is that the transmission lines and matching networks are being used to isolate the oscillator from the amplifier and the

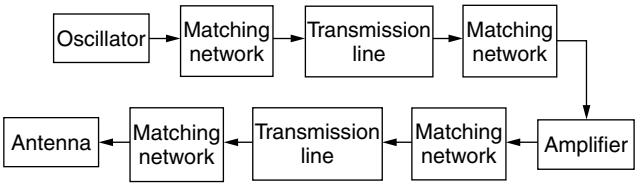


Figure 8. Schematic of a conventional RF microwave transmitter in which each individual element of the transmitter is matched to each other element.

amplifier from the antenna, in contrast to the situation in an active antenna. There were a number of reasons why the conventional solution took on the form it did. Among them was the urgency of World War II. Radar was developed rapidly in both Great Britain and the United States in the 1930s and 1940s. Rapid development required numerous researchers working in parallel. When operating frequencies exceeded 1 GHz (corresponding to 30 cm wavelengths), passive matching networks, whose main requirement is that they must consist of lines of lengths comparable to a wavelength, became convenient to construct (in terms of size) for ground-based radar. In this case, then, the oscillators could be optimized independently of the amplifiers, which in turn could be optimized independently of the antennas and the receiver elements. The impedances of the individual pieces didn't matter, as the matching networks could be used to effectively transform the effective impedances looking into an element into something completely different for purposes of matching pieces of the network to each other. There are costs associated with such a solution, though, such as total system size as well as the tolerances that components must satisfy. However, once the technique was in place, the industry standardized on the conventional solution and perfected it to the point where it was hard to challenge. The reemergence of the active solution owes itself to two independent technologies, the emergence of high-frequency solid-state devices and the development of planar circuit and planar antenna technology.

A single frequency of electromagnetic energy must be generated in a so-called oscillator — that is, a circuit that converts DC electrical power to AC electromagnetic power at the proper frequency. The basic operation of an oscillator can be described with respect to Fig. 9. What is shown here schematically is an amplifier in which a portion $b(<1)$ of the output is fed back to the input with either a plus or a minus sign. When the feedback is off ($b = 0$), then the signal out will be just G times the input. When the feedback is negative, the output will be less than G times the input. However, in the negative-feedback mode, the stability to noise increases, since fluctuations will be damped. That is, if the output fluctuates up, this lowers the effective input, whereas if the output fluctuates down, the output is driven up. The opposite is true in the positive-feedback case. In the positive-feedback case, if there were no fluctuations, any input would cause the output to increase until all of the DC power in as well as all the input signal showed up at the output. (This is all the power that can show up at the output. Such behavior is typical of unstable

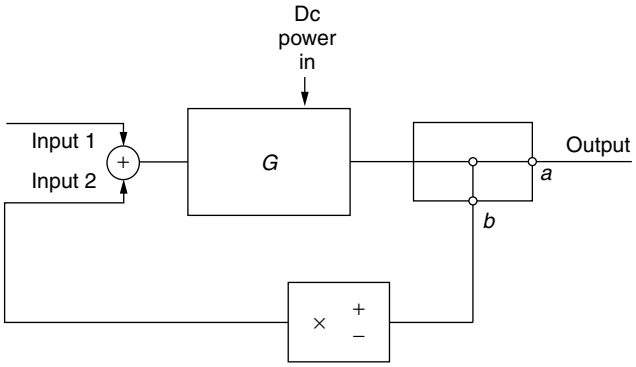


Figure 9. Schematic depiction of a feedback system that can operate as an oscillator when G is greater than 1, the feedback is positive, and there is a delay in feeding back the output to the input.

operation.) This would not be such an interesting case; however, there are always fluctuations of the input, and the positive feedback will cause these to grow. If there is a delay from output to input, then fluctuations with a period corresponding to this delay will be favored, as a rise in the input will show up as a rise in the output one period later, and all the DC power in will be rapidly converted to power out at this magic frequency.

A real circuit operates a bit more interestingly than our ideal one. In a real circuit, as the fluctuations build up, the gain is affected and some elements absorb power, but the oscillations still take place, although perhaps with a different frequency and amplitude from what one would have predicted from nondynamic measurements.

The transistor was first demonstrated in 1947, with publication in 1948 [3], and the diode followed shortly [4]. Although the field-effect transistor (FET) was proposed in 1952 [5], it was not until the mid-1960s that the technology had come far enough that it could be demonstrated [6]. The FET (and variations thereof) is presently the workhorse microwave three-terminal device. Two-terminal transfer electron devices (TEDs) were used before the FET for microwave applications and are still in use, but tend to have a much lower wall plug efficiency (DC/AC conversion), especially as the amplifying device of an oscillator. Radar systems, however, were already in use in the late 1930s. Essentially all of the microwave sources in radars up until the 1970s operated on principles that required that the source have physical dimensions larger than a wavelength, and perhaps many wavelengths. This fact almost required the conventional solution to be used. Transistors, though, can have active areas with dimensions of micrometers; even packaged hybrid devices can have complete packages of dimensions smaller than a millimeter. The transistor can therefore act as an amplifier with dimensions much smaller than a wavelength and does not, therefore, need to be placed in a conventional (passive) solution design.

The last piece of our story of the new active-antenna era involves the development of printed-circuit technology, along with slot and patch antennas. The two most common planar "open waveguide" designs are microstrip line and coplanar waveguide (CPW). Depictions of these waveguide lines are given in Fig. 10. The idea behind the microstrip

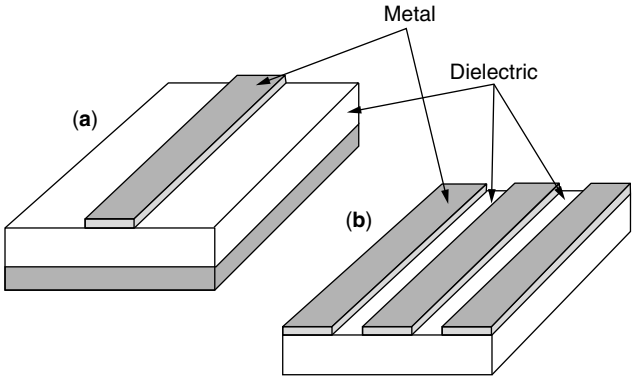


Figure 10. Views of (**a**) a microstrip and (**b**) a coplanar waveguide line. In the microstrip, the ground plane is the lower electrode, whereas in the coplanar waveguide the ground plane is placed on the surface of the dielectric substrate.

line is to propagate electromagnetic energy along the lines by confining the electric field between the upper signal line and a lower ground plane. As the upper line carries current, a magnetic field encircles the upper line. As power flow takes place in a direction perpendicular to the electric and magnetic fields, the power flow is mostly between the signal line and the ground line in the dielectric. On a low-frequency wire (a line whose transverse dimensions are small compared to a wavelength), the voltage and current waveforms reinforce each other. The coupling of the electric and magnetic fields in the microstrip is analogous to the coupling of voltage and current on the Hertz antenna wire, except that the microstrip line can be electrically long in the sense that the distance from the signal line to the ground plane is kept constant so that the impedance can be kept constant, as with the earlier-discussed coaxial cable. Lines that carry along their ground planes are generally referred to as *transmission lines*. Components (i.e., capacitors and inductors) can be built into the line by changing the width, cutting gaps into the upper line, or putting slits in the ground plane. In this sense, we can still describe transmission-line circuits by conventional circuit theory if we use a special circuit model for the line itself. The CPW line is quite similar to the microstrip line except that there the ground planes are on top of the dielectric slab. Either of these line types is reasonably easy to fabricate, as one needs only to buy a metal-coated dielectric plate and then pattern the needed shapes by photographically defining the patterns using a technique known as *photolithography*, a process common to all present-day circuit fabrication. These planar structures are quite compatible with transistor technology, as is indicated by the simple transistor oscillator circuit depicted in Fig. 11. The gap in the line on the drain side is there in order to provide the proper feedback for oscillation. In this case, the total oscillator linear dimension can be less than a wavelength.

In order to have an active antenna, one needs to have a radiating element — that is, a passive antenna element in the active antenna. Certain antenna technologies are compatible with microstrip and CPW technologies, and the resulting antenna types are illustrated in Fig. 12. The idea behind either of these antenna types is that the patch

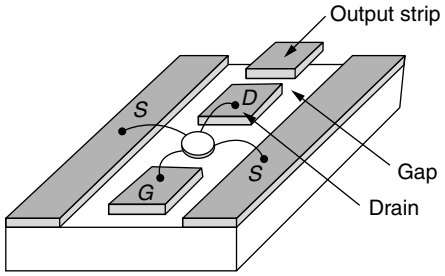


Figure 11. A simple transistor oscillator implemented in CPW technology.

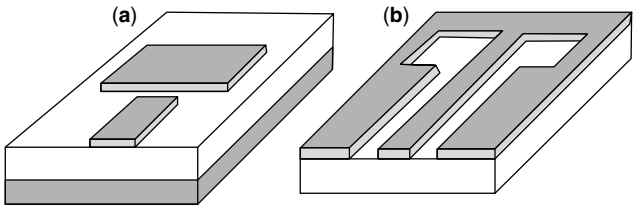


Figure 12. A depiction of (**a**) a patch antenna in a microstrip line and (**b**) a slot antenna in a CPW line.

(slit) is designed to have a transverse length that matches the operating wavelength (as we discussed in conjunction with Hertz dipole antennas). In the case of the patch, the electric field points primarily from the patch to the ground plane, as is illustrated in Fig. 13. The edges of the transverse (to the input line) dimension will then have a field pattern as sketched in Fig. 13a, and the longitudinal edges will have a field pattern as sketched in Fig. 13b, with a composite sketch given in Fig. 13c. The important part of the sketches, however, is really the so-called fringing fields in Fig. 13a — that is, the fields that point neither up nor down but rather across. Beyond the longitudinal edges of the patch are fields, in phase for the two edges, that are

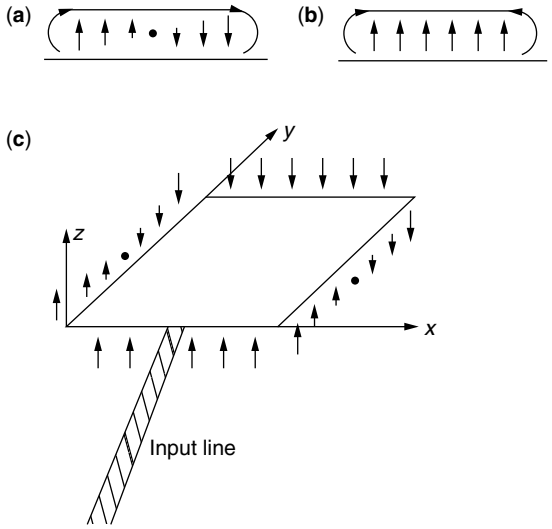


Figure 13. Illustration of the electric field directions along (**a**) the nonradiating edge and (**b**) the radiating edge, and (**c**) a schematic depiction of the edge fields around the patch.

normal to the surface. It is these fields (when combined with transverse magnetic fringe fields in the same strips) that give rise to the upward radiation. Similar arguments describe the operation of the slit antenna if one exchanges the electric and magnetic fields in the argument.

We have now introduced all of the pieces necessary to describe the new resurgence in active antenna research. A possible active-antenna design could appear as in Fig. 14 [7], where the transistor is actually mounted right into the patch antenna element, and therefore the design can be quite compact; that is, the source plus oscillator plus antenna can all be fitted into less than a wavelength. The design of Fig. 14, which comes from R. Compton's group at Cornell [31,32], will be discussed further in the next section.

There are a number of advantages to the use of active antennas. One is that an active antenna can be made compact. Compactness in itself is advantageous, as throughout the history of microelectronics, miniaturization has led to lowered costs. There are two more advantages, though, that relate to compactness. One is that the power-handling capabilities of a device go down with increasing frequency. We would therefore like to find ways to combine the power from several devices. One can try to add together outputs from various oscillators in the circuit before feeding them to the elements, but this goes back to the conventional solution. A more advantageous design is to make an array of antennas, with proper spacing relative to the wavelength and antenna sizes, and add the power of the locked oscillators in the array quasioptically in free space. (In other words, optical radiation tends to radiate into free space, whereas radiofrequency in microwave radiation needs to be kept in guiding waveguides until encroachment on radiating elements. *Quasioptics* uses the principle of the optical interferometer to combine multiple coherent microwave fields in free space.) The locking requires that the oscillators talk to each other so that the phases of all the array elements stay in a given relation. As will be discussed in more detail in the next section, however, an important problem at present in the active-antenna field relates to keeping elements locked yet still being able to modulate the output as well as steer the beam in order to be able to electronically determine on output direction. These issues will be discussed in Section 2 and taken up in more detail in Section 3.

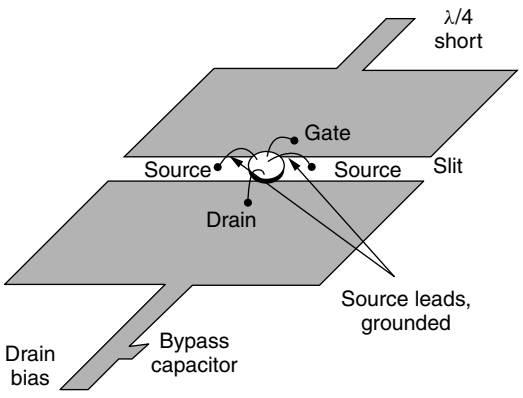


Figure 14. Depiction of the upper surface metallization of a microstrip active patch antenna discussed in Ref. 7. The short circuit on the gate together with the slit between gate and drain provides the proper feedback delay to cause oscillation.

2. SOME QUANTITATIVE DISCUSSION OF ASPECTS OF ACTIVE ANTENNAS

In order to be able to make calculations on active antennas, it is important to know what level of approximation is necessary in order to obtain results. An interesting point is that, although the operating frequency of active antennas is high, the circuit tends to be small in total extent relative to the operating wavelength, and therefore the primary design tool is circuit theory mixed with transmission-line theory. These techniques are approximate, and a most important point in working with high frequencies is to know where a given technique is applicable. Exact treatments of all effects, however, prove to be impossible to carry out analytically. Numerical approaches tend to be hard to interpret unless one has a framework to use. The combined circuit transmission-line framework is the one generally applied. When it begins to break down, one tends to use numerical techniques to bootstrap it back to reality. We will presently try to uncover the basic approximations of transmission-line and circuit theory.

Maxwell's equations are the basic defining equations for all electromagnetic phenomena, and they are expressible in mksA (meter-kilogram-second-ampere) units as [8]

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}$$

$$\nabla \cdot \boldsymbol{D} = \rho$$

$$\nabla \cdot \boldsymbol{B} = 0$$

where $\boldsymbol{E}$ is the electric field vector, $\boldsymbol{B}$ is the magnetic induction vector, $\boldsymbol{H}$ is the magnetic field vector, $\boldsymbol{D}$ is the electric displacement vector, $\boldsymbol{J}$ is the current density vector, and ρ is the volume density of charge. An additional important quantity is $\boldsymbol{S}$, the Poynting vector, defined by

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$$

If one takes the divergence of $\boldsymbol{S}$, one finds

$$\nabla \cdot \boldsymbol{S} = \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H})$$

If one assumes a free-space region

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E}$$

$$\boldsymbol{B} = \mu_0 \boldsymbol{H}$$

which is therefore lossless

$$\boldsymbol{J} = 0$$

and charge-free

$$\rho = 0$$

(where ε_0 is the permittivity of free space and μ_0 is the permeability of free space), one can use vector identities and Maxwell's equations to obtain

$$\nabla \cdot \boldsymbol{S} = -\frac{\varepsilon_0}{2}\frac{\partial}{\partial t}(\boldsymbol{E}\cdot\boldsymbol{E}) - \frac{\mu_0}{2}\frac{\partial}{\partial t}(\boldsymbol{H}\cdot\boldsymbol{H})$$

Integrating this equation throughout a volume V and using Gauss' theorem

$$\int \nabla \cdot \boldsymbol{S}\, dV = \int \boldsymbol{S}\cdot \boldsymbol{dA}$$

where $\boldsymbol{dA}$ is the differential area times the unit normal pointing out of the surface of the volume V, one finds that

$$\int \boldsymbol{S}\cdot\boldsymbol{dA} = -\frac{\partial}{\partial t}W_e - \frac{\partial}{\partial t}W_m$$

where W_e is the electric energy density

$$W_e = \frac{\varepsilon_0}{2}\int \boldsymbol{E}\cdot\boldsymbol{E}\, dV$$

and W_m is the magnetic energy density

$$W_m = \frac{\mu_0}{2}\int \boldsymbol{H}\cdot\boldsymbol{H}\, dV$$

The interpretation of the above is that the amount of $\boldsymbol{S}$ flowing out of V is the amount of change of the energy within. One therefore associates energy flow with $\boldsymbol{S} = \boldsymbol{E}\times\boldsymbol{H}$. This is important in describing energy flow in wires as well as transmission lines and waveguides of all types. As was first described by Heaviside [9], the energy flow in a wire occurs not inside the wire but around it. That is, as the wire is highly conductive, there is essentially no field inside it except at the surface, where the outer layer of oscillating charges have no outer shell to cancel their effect. There is therefore a radial electric field emanating from the surface of the wire, which combines with an azimuthal magnetic field that rings the current flow to yield an $\boldsymbol{E}\times\boldsymbol{H}$ surrounding the wire and pointing down its axis.

It was Pocklington in 1897 [10], who made the formal structure of the fields around a wire a bit more explicit and, in the effort, also formed the basis for the approximation on which most of circuit and transmission-line theory rests, the *quasistatic approximation*. A simplified version of his argument is as follows. Assume an $x-y-z$ Cartesian coordinate system where the axis of the wire is the z axis. One then assumes that all of the field quantities $\boldsymbol{f}(x, y, z, t)$ vary as

$$\boldsymbol{f}(x,y,z,t) = \boldsymbol{f}(x,y)\cos(\beta z - \omega t + \phi)$$

If one assumes that the velocity of propagation of the above-defined wave is $c = (\mu_0\varepsilon_0)^{-1/2}$, the speed of light, then one can write that

$$\beta = \frac{\omega}{c}$$

The assumption here that $\boldsymbol{f}(x,y)$ is independent of z, by substitution of the equation above into Maxwell's equations, can be shown to be equivalent to the assumption that the transverse field components E_x, E_y, B_x, and B_y all satisfy relations of the form

$$\left|\frac{\partial E_x}{\partial z}\right| \ll \beta |E_x|$$

which is the crux of the quasistatic approximation. With the preceding approximation, one finds that

$$\nabla_t \times \boldsymbol{E}_t = \rho$$
$$\nabla_t \times \boldsymbol{H}_t = \boldsymbol{J}$$

where

$$\nabla_t = \hat{\boldsymbol{e}}_x\frac{\partial}{\partial x} + \hat{\boldsymbol{e}}_y\frac{\partial}{\partial y}$$

which is just the transverse, and therefore two-dimensional, gradient operator. These equations are just the electro- and magnetostatic equations for the transverse fields, whereas the propagation equation above shows that these static transverse field configurations are propagated forward as if they corresponded to a plane wave field configuration. If the magnetic field is caused by the current in the wire, it rings the wire, whereas if the electric field is static, it must appear to emanate from charges in the wire and point outward at right angles to the magnetic field. If this is true, then the Poynting vector $\boldsymbol{S}$ will point along the direction of propagation and the theory is self-consistent, if approximate.

If we wish to guide power, then the quasistatic picture must come close to holding, as the Poynting vector is in the right direction for guidance. The more general approximate theory that comes from Pocklington's quasistatic approximation is generally called *transmission-line theory*. To derive this theory, first consider the two-wire transmission line of Fig. 15. If we are to have something that we can actually call a transmission line, then we would hope that we can find equiphase fronts of the electromagnetic disturbance propagating in the gap crossing the gap conductor and that we can find lines along which the current flows on the current-carrying conductor. Otherwise (if the equiphases closed on themselves and/or we had eddies in the current), it would be hard to think of the structure as any form of guiding structure. Let us say

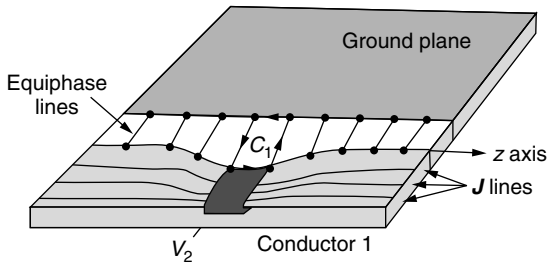


Figure 15. A sketch of a two-conductor transmission line where some equipotentials and some current lines are drawn in, as well as a volume V_1 with outward-pointing normal $d\boldsymbol{A}_1$. There is also an outward-pointing normal $d\boldsymbol{A}_2$ associated with the area bounded by contour C_2.

we form an area in the gap with two walls of the four-sided contour C_1 surrounding this area following equiphases an infinitesimal distance dz from each other. We can then write

$$\int \nabla \times \boldsymbol{E} \cdot \boldsymbol{dA}_1 = -\int \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{dA}_1$$

where $\boldsymbol{dA}_1$ corresponds to an upward-pointing normal from the enclosed area. One generally defines the integral as

$$\int \boldsymbol{B} \cdot \boldsymbol{dA}_1 = \phi$$

where ϕ is the magnetic flux. We often further define the flux as the inductance of the structure times the current:

$$\phi = Li$$

The integral with the curl in it can be rewritten by Stokes' theorem as

$$\int \nabla \times \boldsymbol{E} \cdot \boldsymbol{dA}_1 = \oint_{c_1} \boldsymbol{E} \cdot \boldsymbol{dl}$$

where C_1 is the contour enclosing the area. If we define

$$v = \int \boldsymbol{E} \cdot \boldsymbol{dl}$$

on the two equiphase lines of the contour C_1, where v is an AC voltage (this is the main approximation in the above, as it is only strictly true for truly static fields), then, noting that v does not change along two of the boundaries of the contour (because they are the infinitesimal walls on constant-voltage plates) and making the other two connecting lines infinitesimal, we note that the relation between the curl of $\boldsymbol{E}$ and the magnetic field reduces to

$$v(z + dz) - v(z) = \frac{\partial}{\partial t}(Li)$$

where it has been tacitly assumed that geometric deviations from rectilinearity are small enough that one can approximately use Cartesian coordinates, which can be rewritten in the form

$$\frac{\partial v}{\partial z} = l \frac{\partial i}{\partial t} \tag{1}$$

where l is an inductance per unit length, which may vary with longitudinal coordinate z if the line has longitudinal variation of geometry. A similar manipulation can be done with the second and third of Maxwell's equations. Taking

$$\nabla \cdot (\nabla \times \boldsymbol{H}) = \nabla \cdot \boldsymbol{J} + \frac{\partial}{\partial t} \nabla \cdot \boldsymbol{D}$$

and noting that the divergence of a curl is zero, substituting for $\nabla \cdot \boldsymbol{D}$, we find

$$\nabla \cdot \boldsymbol{J} + \frac{\partial \rho}{\partial t} = 0$$

which is the equation of charge conservation. Integrating this equation over a volume V_2 that encloses the current-carrying conductor whose walls lie perpendicular to the current lines gives

$$\int \nabla \cdot \boldsymbol{J} dV_2 = -\frac{\partial}{\partial t} \int \rho dV_2$$

where the total change Q, given by

$$Q = \int \rho dV_2$$

is also sometimes defined in terms of capacitance C and voltage v by

$$Q = Cv$$

Nothing that

$$\int \nabla \cdot \boldsymbol{J} dV_2 = \int \boldsymbol{J} \cdot \boldsymbol{dA}_2$$

where $\boldsymbol{dA}_2$ is the outward-pointing normal to the boundary of the volume V_2 and where one usually defines

$$i = \int \boldsymbol{J} \cdot \boldsymbol{dA}_2$$

and letting the volume V have infinitesimal thickness, one finds that

$$\int \boldsymbol{J} \cdot \boldsymbol{dA}_2 = i(z + dz) - i(z)$$

Putting this together with the preceding, we find

$$\frac{\partial i}{\partial z} = c \frac{\partial v}{\partial t} \tag{2}$$

where c is the capacitance per length of the structure, and where longitudinal variations in line geometry will lead to a longitudinal variation of c. The system of partial differential equations for the voltage and current have a circuit representation, as is schematically depicted in Fig. 16a. One can verify this by writing Kirchhoff's laws for the nodes with $v(z + dz)$ and $v(z)$ using the relations

$$v = l \frac{\partial i}{\partial t}$$

and

$$i = c \frac{\partial v}{\partial t}$$

Figure 16b illustrates the circuit equivalent for a lossy (and therefore dispersive) transmission line, where r represents the resistance encountered by the current in the metallization and where g represents any conductance of the substrate material that might allow leakage to ground. A major point of the diagram is that the structure need not be uniform in order to have a transmission-line representation, although one may find that irregularities in the structure will lead to longitudinally varying inductances and capacitances.

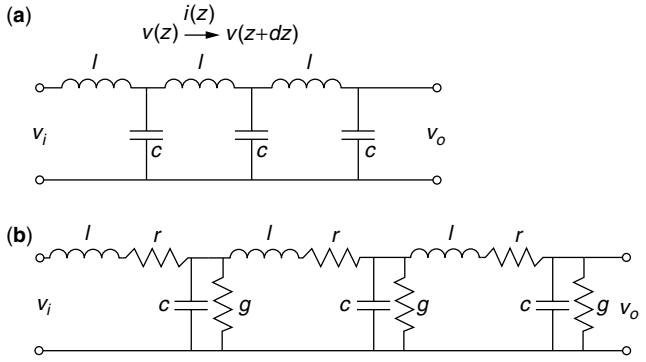


Figure 16. A circuit equivalent for (**a**) a lossless and (**b**) a lossy transmission line. The actual stages should be infinitesimally long, and the l and c values can vary with distance down the line. In reality, one can find closed-form solutions for the waves in nominally constant l and c segments and put them together with boundary conditions.

The solution to the circuit equations will have a wave nature and will exhibit propagation characteristics, which we discussed previously. In a region with constant l and c, one can take a z derivative of Eq. (1) and a t derivative of Eq. (2) and substitute to obtain

$$\frac{\partial^2 v}{\partial z^2} - lc\frac{\partial^2 v}{\partial t^2} = 0$$

which is a wave equation with solutions

$$v(z,t) = v_f \cos(\omega t - \beta z + \phi_f) + v_b \cos(\omega t + \beta z + \phi_b) \qquad (3)$$

where v_f is the amplitude of a forward-going voltage wave, v_b is the amplitude of a backward-going voltage wave, and

$$\frac{\omega}{\beta} = \sqrt{lc}$$

Similarly, taking a t derivative of Eq. (1) and a z derivative of Eq. (2) and substituting gives

$$\frac{\partial^2 i}{\partial z^2} - lc\frac{\partial^2 i}{\partial t^2} = 0$$

which will have a solution analogous to the one in Eq. (3) above, but with

$$v_f = \sqrt{\frac{l}{c}} i_f$$

$$v_b = \sqrt{\frac{l}{c}} i_b$$

which indicates that we can make the identification that the line phase velocity v_p is given by

$$v_p \triangleq \frac{\omega}{\beta} = \sqrt{lc}$$

and the line impedance Z_0 is given by

$$Z_0 = \sqrt{\frac{l}{c}}$$

Oftentimes, we assume that we can write (the sinusoidal steady-state representation)

$$v(z,t) = \mathrm{Re}[v(z)e^{j\omega t}]$$

$$i(z,t) = \mathrm{Re}[i(z)e^{j\omega t}]$$

so that we can write

$$\frac{\partial v}{\partial z} = -j\omega l i$$

$$\frac{\partial i}{\partial z} = -j\omega c v$$

with solutions

$$v(z) = v_f e^{-j\beta z} + v_b e^{j\beta z}$$

$$i(z) = i_f e^{-j\beta z} - i_b e^{j\beta z}$$

Let us say now that we terminate the line with a lumped impedance Z_l at location l. At the coordinate l, then, the relations

$$Z_l i(l) = v_f e^{-j\beta l} + v_b e^{j\beta l}$$

$$Z_0 i(l) = v_f e^{-j\beta l} - v_b e^{j\beta l}$$

hold, and from them we can find

$$v_f = \tfrac{1}{2}(Z_l + Z_0) i(l) e^{j\beta l}$$

$$v_b = \tfrac{1}{2}(Z_l - Z_0) i(l) e^{-j\beta l}$$

which gives

$$v(z) = \frac{i(l)}{2}[(Z_l + Z_0)e^{j\beta(l-z)} + (Z_l - Z_0)e^{-j\beta(l-z)}]$$

$$i(z) = \frac{i(l)}{2Z_0}[(Z_l + Z_0)e^{j\beta(l-z)} - (Z_l - Z_0)e^{-j\beta(l-z)}]$$

allowing us to write that

$$Z(z-\ell) = \frac{v(z-l)}{i(z-l)} = Z_0 \frac{Z_l + jZ_0 \tan\beta(z-l)}{Z_0 + jZ_l \tan\beta(z-l)} \qquad (4)$$

This equation allows us to, in essence, move the load from the plane l to any other plane. This transformation can be used to eliminate line segments and thereby use circuits on them directly. However, note that line lengths at least comparable to a wavelength are necessary in order to significantly alter the impedance. At the plane $z = l$, then, we can further note that the ratio of the reflected voltage coefficient v_b and the forward-going v_f, which is the voltage reflection coefficient, is given by

$$R = \frac{Z_l - Z_0}{Z_l + Z_0}$$

and has the meaning of a Fresnel coefficient [8]. This is the reflection we discussed in the last section, which causes the difference between large and small circuit dimensions.

One could ask what the use was of going at some length into Poynting vectors and transmission lines when

the discussion is about active antennas. The answer is that any antenna system, at whatever frequency or of whatever design, is a system for directing power from one place to another. To direct power from one place to another requires constantly keeping the Poynting vector pointed in the right direction. As we can surmise from the transmission-line derivation, line irregularities may cause the Poynting vector to wobble (with attendant reflections down the line due to attendant variations in the l and c), but the picture must stay close to correct for power to get from one end of the system to another. For this reason, active antennas, even at very high frequencies (hundreds of gigahertz), can still be discussed in terms of transmission lines, impedances, and circuit equivalents, although ever greater care must be used in applying these concepts at increasingly higher frequencies.

The next piece of an active antenna that needs to be discussed is the active element. Without too much loss of generality, we will take our device to be a field-effect transistor (FET). The FET as such was first described by Shockley in 1952 [5], but the MESFET (metal semiconductor FET), which is today's workhorse active device for microwave circuitry, was not realized until 1965 [6], when gallium arsenide (GaAs) fabrication techniques became workable albeit only as a laboratory demonstration. [Although we will discuss the MESFET in this section, it should be pointed out that the silicon MOSFET (metal oxide semiconductor FET) is the workhorse device of digital electronics and therefore the most common of all electronic devices presently in existence by a very large margin.] A top view of an FET might appear as in Fig. 17. As is shown clearly in the figure, an FET is a three-terminal device with gate, drain, and source regions. A cross section of the active region (i.e., where the gate is very narrow) might appear as in Fig. 18. The basic idea is that the saturation-doped n region causes current to flow through the ohmic contacts from drain to source (i.e., electrons flow from source to drain), but the current is controlled in magnitude by the electric field generated by the reverse bias voltage applied to the gate electrode. The situation is described in a bit more detail in Fig. 19, where bias voltages are defined and a typical I–V curve for DC operation is given. Typically the bias is supplied by a circuit such as that of Fig. 20. In what follows, we will simply assume that the biases are properly applied and isolated, and we will consider the AC operation. An AC circuit model is given in Fig. 21. If one uses the proper number of circuit values, these models can be quite accurate, but the values do vary from

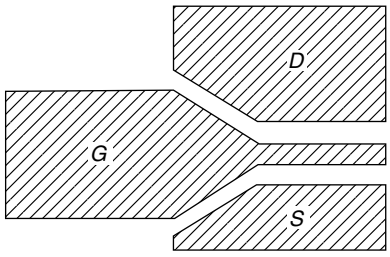


Figure 17. Schematic depiction of a top view of the metallized surface of an FET, where G denotes gate; D, drain; and S, source.

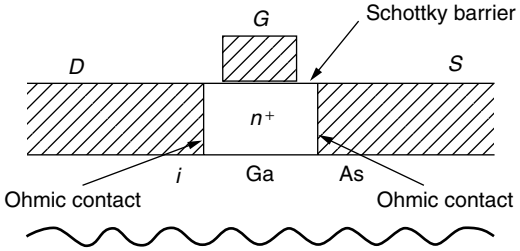


Figure 18. Schematic depiction of the cross section of the active region of a GaAs FET. Specific designs can vary significantly in the field-effect family.

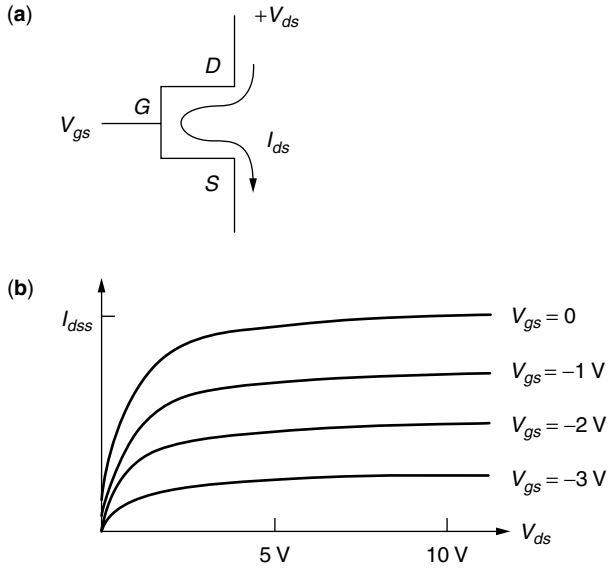


Figure 19. (**a**) Circuit element diagram with voltages and currents labeled for (**b**), where a typical I–V curve is depicted.

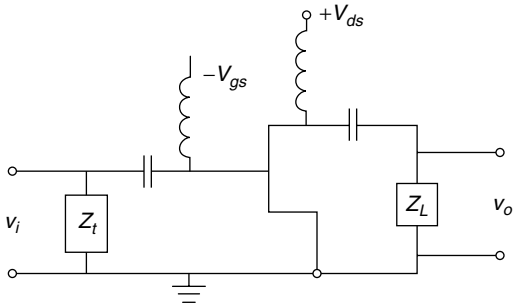


Figure 20. Typical FET circuit including the bias voltages v_{gs} and v_{ds} as well as the AC voltages v_i and v_o, where the conductors represent AC blocks and the capacitors, DC blocks.

device to device, even when the devices were fabricated at the same time and on the same substrate. Usually, the data sheet with a device, instead of specifying the circuit parameters, will specify the parameters of the device S, which are defined as in Fig. 22 and that can be measured in a straightforward manner by a network analyzer. The S parameters are defined by the equation

$$\begin{pmatrix} V_1^- \\ V_2^- \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} V_1^+ \\ V_2^+ \end{pmatrix} \tag{5}$$

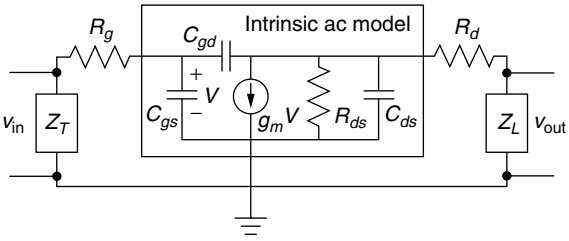


Figure 21. Intrinsic model for a common-source FET with external load and termination impedances and including gate and drain resistive parasitics, where Z_T is the gate termination impedance, R_g is the gate (metallization) resistance, C_{gs} is the gate-to-source capacitance, C_{gd} is the gate-to-drain capacitance, g_m is the channel transconductance, R_{ds} is the channel (drain-to-source) resistance, C_{ds} is the channel capacitance, R_d is the drain (metallization) resistance, and Z_L is the load impedance.

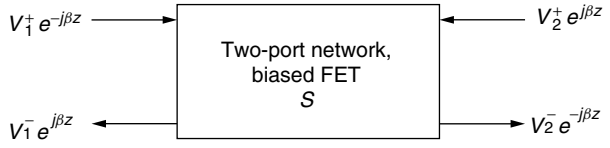


Figure 22. Schematic depiction of an FET as a two-port device that defines the quantities used in the S matrix of Eq. (5).

An important parameter of the circuit design is the transfer function of the transistor circuit, which can be defined as the ratio of v_o to v_i as defined in Fig. 21. To simplify further analysis, we will ignore the package parasitics R_g and R_d in comparison with other circuit parameters, and thereby we will carry out further analysis on the circuit depicted in Fig. 23. The circuit can be solved by writing a simultaneous system of equations for the two nodal voltages v_i and v_o. These sinusoidal steady-state equations become

$$v_i = v$$

$$j\omega C_{gd}(v_o - v_i) + g_m v_i + j\omega C_{ds} v_o + \frac{v_o}{R_{ds}} + \frac{v_o}{Z_L} = 0$$

The system can be rewritten in the form

$$v_o\left(j\omega(C_{gd} + C_{ds}) + \frac{1}{R_{ds}} + \frac{1}{Z_L}\right) = v_i(-g_m + j\omega C_{gd})$$

which gives us our transfer function T in the form

$$T = \frac{v_o}{v_i} = \frac{-g_m + j\omega C_{gd}}{j\omega(C_{gd} + C_{gs}) + \dfrac{1}{R_{ds}} + \dfrac{1}{Z_L}}$$

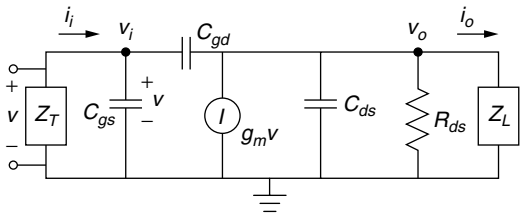


Figure 23. Simplified transistor circuit used for analyzing rather general amplifier and oscillator circuits, where the circuit parameter definitions are as in Fig. 22.

Oftentimes we are interested in open-circuit parameters — for example, the circuit transfer function when Z_L is large compared to other parameters. We often call this parameter G the open-circuit gain. We can write this open-circuit gain in the form

$$G = \left.\frac{v_o}{v_i}\right|_{oc} = \frac{-g_m R_{ds} + j\omega C_{gd} R_{ds}}{j\omega(C_{gd} + C_{gs})R_{ds} + 1}$$

It is useful to look at approximate forms. It is generally true that

$$C_{gd} \ll C_{ds}, C_{gs}$$

and for usual operating frequencies it is also generally true that

$$\frac{1}{\omega C_{ds}} \ll R_{ds}$$

Using both of these in our equations for T and G, we find

$$T = \frac{-g_m R_{ds}}{1 + \dfrac{R}{Z_L}}$$

$$G = -g_m R_{ds}$$

Clearly, one sees that the loaded gain will be lower than the unloaded gain, as we would expect. Making only the first of our two approximations above, we can write the above equations as

$$T = \frac{-g_m R_{ds}}{1 + j\omega\tau_{ds} + \dfrac{R_{ds}}{Z_L}}$$

$$G = \frac{-g_m R_{ds}}{1 + j\omega\tau_{ds}}$$

where τ_{ds} is a time constant given by

$$\tau_{ds} = \frac{1}{C_{ds} R_{ds}}$$

We see that, in this limit, the high-frequency gain is damped. Also, an interesting observation is that, at some frequency ω, an inductive load could be used to cancel the damping and obtain a purely real transfer function at that frequency. This effect is the one that allows us to use the transistor in an oscillator.

Let us now consider an oscillator circuit. The basic idea is illustrated in the one-port diagram of Fig. 24. The transistor's gain, together with feedback to the input loop through the capacitor C_{gd}, can give the transistor an effective negative input impedance, which can lead to oscillation if the real and imaginary parts of the total impedance (i.e., Z_T in parallel with the Z_i of the transistor plus load) cancel. The idea is much like that illustrated in Fig. 25 for a feedback network. One sees that the output of the feedback network can be expressed as

$$v_o = G(j\omega)[v_i - H(j\omega)v_o]$$

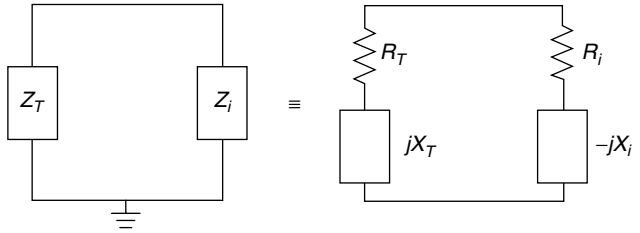


Figure 24. Diagram depicting the transistor and its load as a one-port device that, when matched to its termination so that there is no real or imaginary part to the total circuit impedance, will allow for oscillations.

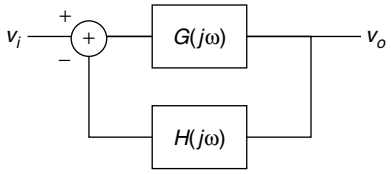


Figure 25. Depiction of a simple feedback network.

or, on rearranging terms

$$\frac{v_o}{v_i} = \frac{G(j\omega)}{1 + G(j\omega)H(j\omega)}$$

which clearly will exhibit oscillation — that is, have an output voltage without an applied input voltage — when

$$H(j\omega) = -\frac{1}{G(j\omega)}$$

What we need to do to see if we can achieve oscillation is to investigate the input impedance of our transistor and load seen as a one-port network. Clearly we can write the input current of Fig. 23 as

$$i_i = j\omega C_{gs} v_i + j\omega C_{gd}(v_i - v_o)$$

and then, using the full expression for T to express v_o as a function of v_i, one finds

$$Z_i = \frac{i_i}{v_i} = j\omega C_{gs} + j\omega C_{gd}\left(1 + \frac{g_m - j\omega C_{gd}}{j\omega(C_{gd} + C_{ds}) + \dfrac{1}{R_{ds}} + \dfrac{1}{Z_L}}\right)$$

which can be somewhat simplified to yield

$$Z_i = j\omega C_{gs} + j\omega C_{gd}\frac{g_m R_{ds} + 1 + j\omega\tau_{ds} + \dfrac{R_{ds}}{Z_L}}{1 + j\omega\tau_{ds} + \dfrac{R_d}{Z_L}}$$

We can again invoke a limit in which $\omega\tau_{ds} \ll 1$ and then write

$$Z_i = j\omega C_{gs} + j\omega C_{gd}\frac{Z_L(1 + g_m R_{ds} + R_{ds})}{R_{ds} + Z_L}$$

Perhaps the most interesting thing about this expression is that if

$$Z_L = j\omega L$$

and

$$g_m R_{ds} \gg 1$$

then clearly

$$R_i < 0$$

Whether X_i can be made to match any termination is another question, which we will take up in the next paragraph.

As was mentioned earlier, generally the data sheet one obtains with an FET has plots of the frequency dependence of the S parameters rather than values for the equivalent-circuit parameters. Oscillator analysis is therefore usually carried out using a model of the circuit such as that depicted in Fig. 26, where the transistor is represented by its measured S matrix. The S matrix is defined as the matrix of reflection and transmission coefficients. That is to say, with reference to the figure, S_{11} would be the complex ratio of the field reflected from the device divided by the field incident on the device. S_{21} would be the field transmitted from the device divided by the field incident on the device. S_{12} would be the field incident from the load side of the device divided by the power incident on the device, and S_{22} would be the power reflected from the load side of the device divided by the power incident on the device. For example, if there is only an input from Z_T, then

$$\Gamma_i = S_{11}$$

If there is only an input from Z_L, then

$$\Gamma_o = S_{22}$$

The condition for oscillation in such a system can be expressed in either of the forms

$$\Gamma_i \Gamma_T = 1$$

or

$$\Gamma_o \Gamma_L = 1$$

where the Γs are defined in the caption of Fig. 26. If both Z_T and Z_L were passive loads — that is, loads consisting of

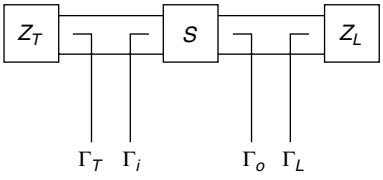


Figure 26. Schematic depiction of an oscillator circuit in which the transistor is represented by its S matrix and calculation is done in terms of reflection coefficients Γ_T looking into the gate termination, Γ_i looking into the gate source port of the transistor, Γ_o looking into its drain source port, and Γ_L looking into the load impedance.

resistance, inductance, and capacitance — then we would have that

$$|\Gamma_T| < 1$$
$$|\Gamma_L| < 1$$

and the conditions for unconditional stability (nonoscillation at any frequency) would be that

$$|\Gamma_i| < 1$$
$$|\Gamma_o| < 1$$

Clearly, we can express Γ_i and Γ_o as series of reflections such that

$$\Gamma_i = S_{11} + S_{12}\Gamma_L S_{21} + S_{12}\Gamma_L S_{22}\Gamma_L S_{21} + S_{12}\Gamma_L S_{22}\Gamma_L S_{22}\Gamma_L S_{21} + \cdots$$
$$\Gamma_o = S_{22} + S_{21}\Gamma_T S_{12} + S_{21}\Gamma_T S_{11}\Gamma_T S_{12} + S_{21}\Gamma_T S_{11}\Gamma_T S_{11}\Gamma_T S_{12} + \cdots$$

Using the fact that

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

we can reexpress the Γs as

$$\Gamma_i = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}$$
$$\Gamma_o = S_{22} + \frac{S_{12}S_{21}\Gamma_T}{1 - S_{22}\Gamma_T}$$

If we denote the determinant of the S matrix by

$$\Delta = S_{11}S_{22} - S_{12}S_{21}$$

and define a transistor parameter κ by

$$\kappa = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|}$$

then some tedious algebra leads to the result that stability requires

$$\kappa > 1$$
$$\Delta < 1$$

At frequencies where the above are not satisfied, oscillation can occur if the load and termination impedances, Z_L and Z_T respectively, are chosen properly. Oscillator design is discussed in various texts [11–14]. Generally, though, oscillator design involves finding instability points and not predicting the dynamics once oscillation is achieved. Here we are discussing only oscillators that are self-damping. External circuits can be used to damp the behavior of an oscillator, but here we are discussing only those that damp themselves independent of an external circuit. The next paragraph will discuss these dynamics.

If a transistor circuit is designed to be unstable, then, as soon as the DC bias is raised to a level where the circuit achieves the set of unstable values, the circuit's output within the range of unstable frequencies rises rapidly and dramatically. The values that we took in the equivalent AC circuit, though, were small-signal parameters. As the circuit output increases, the signal will eventually no longer be small. The major thing that changes in this limit is that the input resistance to the transistor saturates, so that [14]

$$R_i = -R_{i\phi} + mv^2$$

where the plus sign on the nonlinearity is necessary, for if it were negative, the transistor would burn up or else burn up the power supply. Generally, m has to be determined empirically, as nonlinear circuit models have parameters that vary significantly from device to device. For definiteness, let us assume that the Z_T is resistive and the Z_L is purely inductive. At the oscillation frequency, the internal capacitance of the transistor then should cancel the load inductance, but to consider dynamics we need to put in both C and L, as dynamics take place in the time domain. The dynamic circuit to consider is then as depicted in Fig. 27. The loop equation for this circuit in the time domain is

$$L\frac{\partial i}{\partial t} + (R_i + R_T)i + \frac{1}{C}\int i dt = 0$$

Recalling the equivalent circuit of Fig. 23 and recalling that

$$C_{gs} \gg C_{gd}$$

we see that, approximately at any rate, we should have a relation between v_i and i_i of the form

$$i_i = C_{gs}\frac{\partial v_i}{\partial t}$$

Using this $i - v$ relation above, we find that

$$\frac{\partial^2 v}{\partial t^2} - \frac{R_i - R_T}{L}\left(1 - \frac{mv^2}{R_i - R_T}\right)\frac{\partial v}{\partial t} + \frac{v}{LC} = 0$$

which we can rewrite in terms of other parameters as

$$\frac{\partial^2 v}{\partial t^2} - \varepsilon(1 - \gamma^2 v^2)\frac{\partial v}{\partial t} + \omega_0^2 v = 0$$

which is the form of Van der Pol's equation [15,16], which describes the behavior of essentially any oscillator.

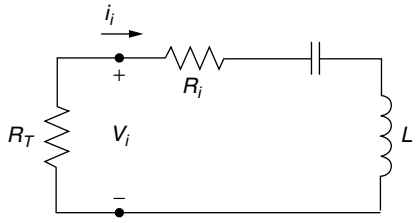


Figure 27. Circuit used to determine the dynamical behavior of a transistor oscillator.

Now that we have discussed planar circuits and dynamical elements that we can put into theory, the time has arrived to discuss planar antenna structures. Perhaps the best way to gain understanding of the operation of a patch antenna is by considering a cavity resonator model of one. A good review of microstrip antennas is given in Carver and Mink [17] and is reprinted in Pozar and Schaubert [18]. Let us consider a patch antenna and coordinate system as is illustrated in Fig. 28. The basic idea behind the cavity model is to consider the region between the patch and ground plane as a resonator. To do this, we need to apply some moderately crude approximate boundary conditions. We will assume that there is only a z-directed electric field underneath the patch and that this field achieves maxima on the edges (open-circuit boundary condition). The magnetic field $\boldsymbol{H}$ will be assumed to have both x and y components, and its tangential components on the edges will be zero. (This boundary condition is the one consistent with the open-circuit condition on the electric field and becomes exact as the thickness of the layer approaches zero, as there can be no component of current normal to the edge at the edge, and it is the normal component of the current that generates the transverse $\boldsymbol{H}$ field.) The electric field satisfying the open-circuit condition can be seen to be given by the modes

$$\boldsymbol{e}_{mn} = \hat{\boldsymbol{e}}_z \frac{\chi mn}{\sqrt{\varepsilon abt}} \cos k_n x \cos k_m y$$

where

$$k_n = \frac{n\pi}{a}$$

$$k_m = \frac{m\pi}{b}$$

$$\chi_{mn} = \begin{cases} 1, & m = 0 \quad \text{and} \quad n = 0 \\ \sqrt{2}, & m = 0 \quad \text{or} \quad n = 0 \\ 2, & m \neq 0 \quad \text{and} \quad n \neq 0 \end{cases}$$

The $\boldsymbol{H}$ field corresponding to the $\boldsymbol{E}$ field then will consist of modes

$$\boldsymbol{h}_{mn} = \frac{1}{j\omega\mu} \frac{\xi_{mn}}{\varepsilon abt} (\hat{e}_x k_m \cos k_n x \sin k_m y - \hat{e}_y k_n \sin k_n x \cos k_m y)$$

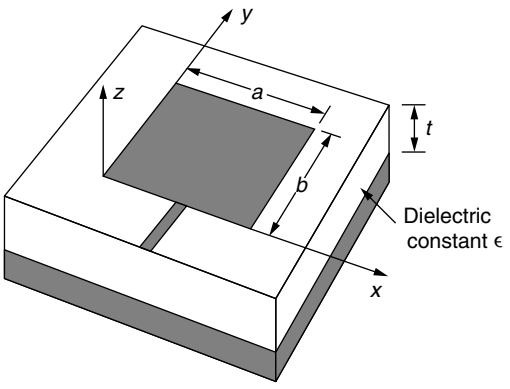


Figure 28. A patch antenna and Cartesian coordinate system.

As can be gathered from Fig. 13, the primary radiation mode is the mode with $m = 1$ and $n = 0$.

The basic operation is described by the fact that the boundary conditions are not quite exact. Recall from the earlier argument that accompanied Fig. 13 that the z-directed field gives rise to a fringe field at the edges $y = 0$ and $y = b$ such that there are strips of y-directed electric field around $y \leq 0$ and $y \geq b$. Because the boundary conditions are not quite correct on $\boldsymbol{H}$, there will also be strips of x-directed magnetic fields in these regions. As the Poynting vector is given by $\boldsymbol{E} \times \boldsymbol{H}$, we note that these strips will give rise to a z-directed Poynting vector. Similar arguments can be applied to the edges at $x = 0$ and $x = a$. However, the x-directed field at $x \leq 0$ has a change of sign at the center of the edge and is pointwise oppositely directed to the x-directed electric field at $x = 0$. These fields, therefore, only give rise to very weak radiation, as there is significant cancellation. Analysis of the slot antenna requires only that we interchange the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields.

The picture of the patch antenna as two radiating strips allows us to represent it with a transmission line as well as a circuit model. The original idea is due to Munson [19]. The transmission-line model is depicted in Fig. 29. The idea is that one feeds onto an edge with an admittance (inverse impedance) $G_1 + jB_1$ and then propagates to a second edge with admittance $G_2 + jB_2$. When the circuit is resonant, then the length of transmission line will simply complex-conjugate the given load [see Eq. (4)], leading to the circuit representation of Fig. 29b. The slot admittance used by Munson [19] was just that derived for radiation from a slit in a waveguide [20] as

$$G_1 + jB_1 = \frac{\pi a}{\lambda_0 Z_0} (1 - j0.636 \ln k_0 t)$$

where Z_0 is the impedance of free space ($\sqrt{\mu_0/\varepsilon_0} = 377\ \Omega$), λ_0 is the free-space wavelength, and k_0 is the free-space propagation vector, and where a and t are defined as in Fig. 28. When the edges are identical (as for a rectangular patch), one can write

$$G_2 + jB_2 = G_1 + jB_1$$

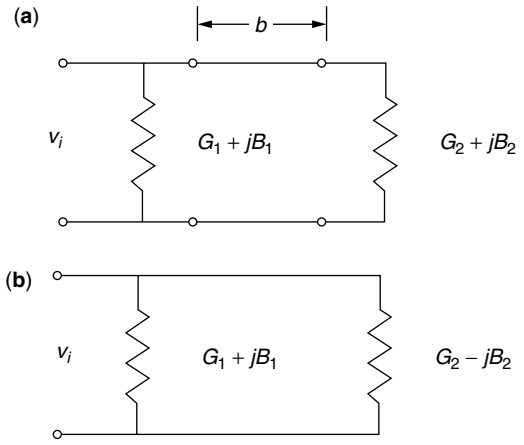


Figure 29. (**a**) A transmission-line model for a patch antenna, and (**b**) its circuital equivalent as resonance.

to obtain the input impedance in the form

$$Z_i = \frac{1}{Y_i} = \frac{1}{2G_1}$$

We have now considered all of the pieces, and therefore it is time to consider a couple of actual active antenna designs. Figure 30 depicts one of the early designs from Kai Chang's group at Texas A&M [21]. Essentially, the patch here is being used precisely as the feedback element of an amplifier circuit (as was described in connection with Fig. 9). A more compact design is that of Fig. 14 [7]. There, the transistor is actually mounted directly into the patch antenna. The slit between the gate and the drain yields a capacitive feedback element such that the effective AC circuit equivalent of this antenna may appear as depicted in Fig. 31. The capacitor–inductor pair attached to the gate lead forms what is often referred to as a *tank circuit*, which (if the load were purely real) defines a natural frequency through the relation

$$\omega = \sqrt{\frac{1}{LC}}$$

As was discussed at some length in Section 1 of this article, a major argument for the use of active antennas is that they are sufficiently compact that they can be arrayed together. Arraying is an important method for free-space power combining, which is necessary because as the frequency increases, the power-handling capability of active devices decreases. However, element size also decreases with increasing frequency so that use of multiple coherently combined elements can allow one to fix the total array size and power more or less independently of frequency, even though the number of active elements to combine increases. In the next paragraph, we shall consider some of the basics of arrays.

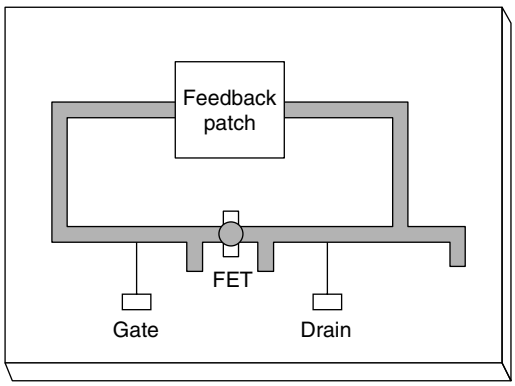


Figure 30. A design of a microstrip active radiating element.

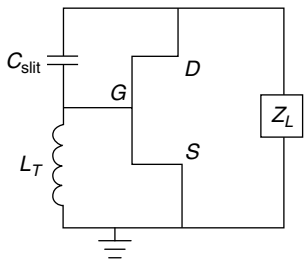


Figure 31. AC circuit equivalent of the active antenna of Fig. 14.

Consider a linear array such as is depicted in Fig. 32. Now let us say that the elements are nominally identical apart from phases that are set by the array operator at each of the elements. The complex electric field far from the nth element due to only the nth element is then given by

$$\boldsymbol{E}_n = \boldsymbol{E}_e e^{i\phi n}$$

where $\boldsymbol{E}_e$ is the electric field of a single element. To find out what is radiated in the direction θ due to the whole array, we need to sum the fields from all of the radiators, giving each radiator the proper phase delay. Each element will get a progressive phase shift $kd \sin\theta$ due to its position (see Fig. 32), where k is the free-space propagation factor, given by

$$k = \frac{2\pi}{\lambda}$$

where λ is the free-space wavelength. With this, we can write for the total field radiated into the direction θ due to all n elements

$$\boldsymbol{E}_t(\theta) = \boldsymbol{E}_e \sum_{n=0}^{N-1} e^{-inkd\sin\theta} e^{i\phi n}$$

The sum is generally referred to as the *array factor*. The intensity, then, in the θ direction is

$$\boldsymbol{I}_t(\theta) = \boldsymbol{I}_e \left| \sum_{n=0}^{N-1} e^{-iknd\sin\theta} e^{i\phi n} \right|^2$$

One notes immediately that, if one sets the phases ϕ_n to

$$\phi_n = nkd\sin\theta$$

then the intensity in the θ direction is N^2 times the intensity due to a single element. This is the effect of coherent addition. One gets a power increase of N plus a directivity increase of N. To illustrate, let us consider the

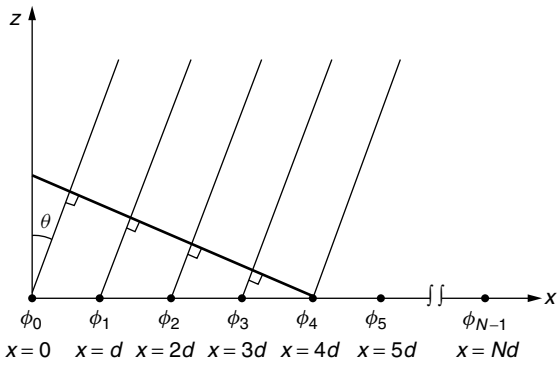


Figure 32. Depiction of a linear array of N identical radiating elements.

broadside case where we take all the ϕ_n to be zero. In this case, we can write the array factor in the form

$$\left|\sum_{n=0}^{N-1} e^{-ind\sin\theta}\right|^2 = \left|\frac{1-e^{-iNkd\sin\theta}}{1-e^{-ikd\sin\theta}}\right|^2$$

which in turn can be written as

$$\mathrm{AF} = \frac{\sin^2\left(N\dfrac{kd}{2}\sin\theta\right)}{\sin^2\left(\dfrac{kd}{2}\sin\theta\right)} \tag{6}$$

which is plotted in Fig. 33. Several interesting things can be noted from the expression and plots. For kd less than π, there is only one central lobe in the pattern. Also, the pattern becomes ever more directed with increasing N. This is called the *directivity effect*. If the array has a power-combining efficiency of 100% (which we have built into our equations by ignoring actual couplings, etc.), then the total power radiated can be only N times that of a single element. However, it is radiated into a lobe that is only $1/N$ times as wide as that of a single element.

If we are to realize array gain, however, we need to be certain that the array elements are identical in frequency and have fixed phase relations in time. This can take place only if the elements are locked together. The idea of locking is probably best understood in relation to the Van der Pol equation [16], with an injected term, such that

$$\frac{\partial^2 v}{\partial t^2} - \frac{R_{i\phi}-R_T}{L}\left(1-\frac{mv^2}{R_{i\phi}-R_T}\right)\frac{\partial v}{\partial t} + \omega_0^2 v = A\cos\omega_i t$$

where $R_{i\phi}$ is the input resistance of the transistor circuit as seen looking into the gate source port and R_T is the external termination resistor placed between the gate and common source. In the absence of the locking term, one can see that oscillation will take place with a primary frequency (and some harmonics) at angular frequency ω_0 with amplitude $\sqrt{R_{i0}-R_T/m}$ such that

$$v(t) \approx \sqrt{\frac{R_{i0}-R_T}{m}}\cos\omega_0 t$$

Without being too quantitative, one can say that, if ω_i is close enough to ω_0 and A is large enough, the oscillation will lock to ω_i in frequency and phase. If ω_i is not quite close enough and A not quite big enough (how big A needs to be is a function of how close ω_i is), then the oscillation frequency ω_0 will be shifted so that

$$v(t) = A_0\cos[(\omega_0+\Delta\omega)t+\phi]$$

where $\Delta\omega$ and ϕ are functions of ω_i and A. These ideas are discussed in a number of places including Refs. 1, 15, 16, 22, 23, and 24. In order for our array to operate in a coherent mode, the elements must be truly locked. This locking can occur through mutual coupling or through the injection of an external signal to each of the elements.

Ideally, we would like to be able to steer the locked beam. A number of techniques for doing this are presently under investigation. Much of the thinking stems from the work Stephan [25–28] and Vaughan and Compton [28a]. One of the ideas brought out in these works was that, if the array were mutually locked and one were to try to inject one of the elements with a given phase, all the elements would lock to that phase. However, if one were to inject two elements at the locked frequency but with different phases, then the other elements would have to adjust themselves to these phases. In particular, if one had a locked linear array and one were to inject the two end elements with phases differing by ϕ, then the other elements would share the phase shift equally so that there would be a linear phase taper of magnitude ϕ uniformly distributed along the array.

A different technique was developed by York [29,30], based on work he began when working with Compton [31,32]. In this technique, instead of injecting the

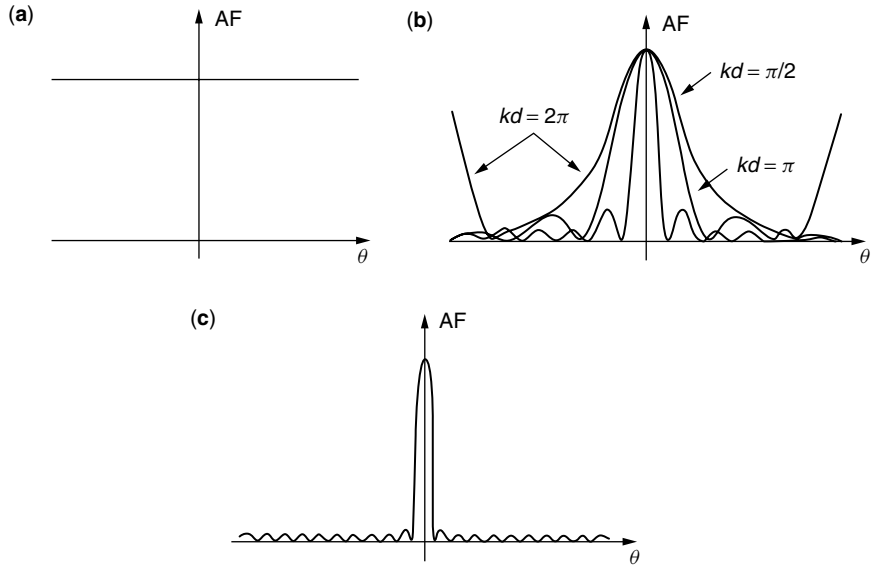


Figure 33. Plots of the array factor of Eq. (6), where (**a**) $N = 1$, (**b**) $N = 5$ and $kd = \pi/2, \pi$, and 2π, and (**c**) $N = 10$ and $kd = \pi$.

end elements with the locked frequency and different phase, one injects with wrong frequencies. If the amplitudes of these injected frequencies are set to values that are not strong enough to lock the elements to this wrong frequency, then the elements will retain their locked frequencies but will undergo phase shifts from the injected signal. If the elements of the array are locked because of mutual feedback, trying to inject either end of the array with wrong frequencies will then tend to give the elements a linear taper — that is, one in which the phase varies linearly with distance down the array — with much the same result as in the technique of Stephan. This will just linearly steer the main lobe of the array off broadside and to a new direction. Such linear scanning is what is needed for many commercial applications such as tracking or transmitting with minimum power to a given location.

Another technique, which again uses locking-type ideas, is that of changing the biases on each of the array's active devices [33–35]. Changing the bias of a transistor will alter the ω_0 at which the active antenna wants to oscillate. For an element locked to another frequency, then, changing the bias will just change the phase. In this way one can individually set the phase on each element. There are still a couple of problems with this approach (as with all the others so far, which is why this area is still one of active research). One is that addressing each bias line represents a great increase in the complexity that we were trying to minimize by using an active antenna. The other is that the maximum phase shift obtainable with this technique is $\pm\pi$ from one end of the array to the other (a limitation that is shared by the phase-shifts-at-the-ends technique). In many phased-array applications, of which electronic warfare is a typical one, one wants to have true time delay, which means that one would like to have as much as a π phase shift between adjacent elements. I do not think that the frequency shifting technique can achieve this either. Work, however, continues in this exciting area.

3. APPLICATIONS OF AND PROSPECTS FOR ACTIVE ANTENNAS

Perhaps the earliest application of the active antenna concept (following that of Hertz) was aimed at solving the small-antenna problem. As we recall, an antenna can be modeled (roughly) by a series RLC network, where the R represents the radiation resistance. The input impedance of such a combination is given by

$$Z_i = \frac{1 - \omega^2/\omega_0^2 + j\omega RC}{j\omega C}$$

and so we see that, when the operation frequency ω is well below the resonant frequency

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

and the reciprocal of the RC time constant

$$\tau = RC$$

then the antenna appears as a capacitor and radiates quite inefficiently. The problem of reception is similar. Apparently already in 1928 Westinghouse had a mobile antenna receiver that used a pentode as an inductive loading element in order to boost the amount of low-frequency radiation that could be converted to circuit current. In 1974, two works discussed transistor-based solutions to the short aerial problem [36,37]. In Ref. 37, the load circuit appeared as in Fig. 34. The idea was to generate an inductive load whose impedance varied with frequency, unlike a regular inductor, but so as to in crease the antenna bandwidth. The circuit's operation is not intuitively obvious. I think that it is possible that most AM, shortwave, and FM receivers employ some short-antenna solution regardless of whether the actual circuit designers were aware that they were employing active antenna techniques.

Another set of applications where active devices are essentially used as loading elements is in the >100-GHz regime. Reviews of progress in this regime are given in Refs. 1 and 38. To date, most work at frequencies greater than 100 GHz has involved radioastronomical receivers. A problem at such frequencies is a lack of components, including circuit elements so basic as waveguides. Microstrip guides already start having extramode problems at Ku band. Coplanar waveguides can go higher, although to date, rectangular metallic waveguides are the preferred guiding structures past about 60 GHz. In W band (normally narrowband, about 94 GHz — see Table 1), there are components, as around 94 GHz there is an atmospheric window of low propagation loss. However, waveguide tolerances, which must be a small percentage of the wavelength, are already severe in W band, where the wavelength is roughly 3 mm. Higher frequencies have to be handled in free space or, as one says, quasioptically. Receivers must therefore by nature be downconverting in this >100-GHz regime. Indeed, these types of solutions are the ones being demonstrated by the group at Michigan [38], where receivers will contain multipliers and downconverting mixers right in the antenna elements in order that CPW can be used to carry the downconverted signals to the processing electronics. Millimeter-wave-terahertz radioastronomy seems to be a prime niche for quasioptical active-antenna solutions.

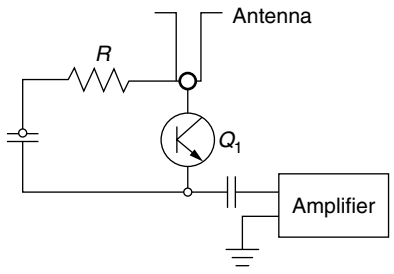


Figure 34. A circuit taken from Ref. 37 in which a transistor circuit is used to load a short antenna. Analysis shows that, in the frequency regime of interest, the loading circuit appears, when looking toward the antenna from the amplifier terminals, to cancel the strongly capacitive load of the short antenna.

The first applications of active antennas where solid-state components were used as gain elements were primarily for power boosting [39–44]. Power combining (see reviews in Refs. 45 and 46) can be hard to achieve. There is a theorem that grew out of the early days of radiometry and radiative transfer (in the 1800s), known variously as the brightness theorem, the Lagrange invariant, or (later) the second law of thermodynamics. (See, e.g., Ref. 8, Chap. 5.) The theorem essentially states that one cannot increase the brightness of a source by passive means. This theorem practically means that, if one tries to combine two nominally identical sources by taking their outputs, launching them into waveguides, and then bringing the two waveguides together in a Y junction into a single waveguide, the power in the output guide, if the output guide is no larger than either of the input guides, can be no greater than that of either of the nominally identical sources. This seems to preclude any form of power combining. There is a bit of a trick here, though. At the time the brightness theorem was first formulated, there were no coherent radiation sources. If one takes the output of a coherent radiation source, splits it in two, and adds it back together in phase, then the brightness, which was halved, can be restored. If two sources are locked, they are essentially one source. (As P. A. M. Dirac said, a photon only interferes with itself. Indeed, the quantum-mechanical meaning of locking is that the locked sources are sharing a wavefunction.) Therefore, locked sources can be coherently added if they are properly phased. We will take this up again in a following paragraph.

An alternative to power combining that obviates the need for locking and precise phase control is amplification of the signal from a single source at each element. By 1960, solid-state technology had come far enough that antennas integrated with diodes and transistors could be demonstrated. The technology was to remain a laboratory curiosity until the 1980s, when further improvements in microwave devices were to render it more practical. More recent research, however, has been more concentrated on the coherent power combining of self-oscillator elements. This is not to say that the element-mounted amplifier may not still be of practical use. The main research issue at present, though, is the limited power available from a single active element at millimeter-wave frequencies.

Another application area is that of proximity detection [47]. The idea is that an oscillator in an antenna element can be very sensitive to its nearby (several wavelengths) environment. As was discussed previously, variation in distances to ground planes changes impedances. The proximity of any metal object will, to some extent, cause the oscillator to be aware of another ground plane in parallel with the one in the circuit. This will change the impedance that the oscillator sees and thereby steer the oscillator frequency. The active antenna of Ref. 47 operated as a self-oscillating mixer. That is, the active element used the antenna as a load, whereas the antenna also used a diode mixer between itself and a low-frequency external circuit. The antenna acted as both a transmitting and a receiving antenna. If there were something moving near the antenna, the signal reflected off the object and rereceived might well be at a different frequency than the shifting oscillator frequency. These two frequencies would then beat in the mixer, be downconverted, and show up as a low-frequency beat note in the external circuit. If such a composite device were to be used in a controlled environment, one could calibrate the output to determine what is occurring. Navarro and Chang [1, p. 130] mention such applications as automatic door openers and burglar alarms. The original paper [47] seemed to have a different application in mind, as the term *Doppler sensor* was in the title. If one were to carefully control the immediate environment of the self-oscillating mixer, then reflections off more distant objects that were received by the antenna would beat with the stable frequency of the oscillator. The resulting beat note of the signals would then be the Doppler shift of the outgoing signal on reflection off the surface of the moving object, and from it one could determine the normal component of the object's velocity. It is my understanding that some low-cost radars operate on such a principle. As with other applications, though, the active-antenna principle, if only due to size constraints, becomes even more appealing at millimeter-wave frequencies, and at such frequencies power constraints favor use of arrays.

An older antenna field that seems to be going through an active renaissance is that of retroreflection. A retroreflector is a device that, when illuminated from any arbitrary direction, will return a signal directly back to the source. Clearly, retroreflectors are useful for return calibration as well as for various tracking purposes. An archetypical passive retroreflector is a corner cube. Another form of passive reflector is a Van Atta array [48]. Such an array uses wires to interconnect the array elements so that the phase progression of the incident signal is conjugated and thereby returned in the direction of the source. As was pointed out by Friis already in the 1930s, though, phase conjugation is carried out in any mixer in which the local oscillator frequency exceeds the signal frequency [49]. (A *phase conjugate* signal is one that takes on negative values at each phase point on the incoming wave.) This principle was already being exploited in 1963 for implementing retroreflection [50]. This work did not catch on, perhaps for technical reasons. A review in 1994 [51] and designs for such arrays were demonstrated and presented at the 1995 International Microwave Symposium [52,53]. Although both demonstrations used transistors and patch-type elements, both also employed circulators for isolation and therefore were not actually active array demonstrations. It would seem that retroreflection should motivate an active self-oscillating mixer solution, which will perhaps appear in the future.

As was mentioned earlier in this article, a quite important application area for active antennas is free-space power combining. As was pointed out then, a number of groups are working on developing compact elements such as those of Fig. 14 [7] and Fig. 30 [21]. As was also mentioned previously, in order to do coherent power combining, the elements must be locked. In designs where the elements are spatially packed tightly enough, proximity can lead to strong enough nearest-neighbor coupling so that the array will lock to a common frequency

and phase. Closeness of elements is also desirable in that arrays with less than $\lambda/2$ spacing will have no sidelobes sapping power from the central array beam. In designs that do not self-lock, one can inject a locking signal either on bias lines or spatially from a horn to try to lock to all elements simultaneously. Of course, the ultimate application would be for a high-bandwidth, steerable, low-cost transceiver.

Another method of carrying out power combining is to use the so-called *grid oscillator* [54,55]. The actual structure of a grid appears in Fig. 35. The operating principle of the grid is quite a bit different from that of the arrays of weakly coupled individual elements. Note that there is no ground plane at all on the back, and there is no ground plane either, per se, on the front side. Direct optical measurements of the potentials on the various lines of the grid [56], however, show that the source bias lines act somewhat like AC grounds. In this sense, either a drain bias line together with the two closest source biases, or a gate bias line together with the two horizontally adjacent bias lines, appears somewhat like CPW. The CPW lines, however, are periodically loaded ones with periodic active elements alternated with structures that appear like slot antennas. The radiating edges of the slots are, for the drain bias lines, the vertical AC connection lines between drain and drain or, for the gate bias CPW, the horizontal AC gate-to-gate connection lines. Indeed, the grid is known to lock strongly between the rows and more weakly between columns. As adjacent row elements are sharing a patch radiator, this behavior should be expected.

In a sense, this strong locking behavior of the grid is both an advantage and a disadvantage. It is advantageous that the grid is compact (element spacing can be $\leq \lambda/6$) and further that it is easy to get the rows to lock to each other. However, the compactness is also a disadvantage in that it is quite hard to get any more functionality on the grid. Much effort has been made in this area to generate functionality by stacking various grid-based active surfaces such as amplifying surfaces, varactor surfaces for frequency shifting and modulation; and doubling surfaces. A problem with stacking is, of course, diffraction as well as alignment. Alignment tolerance adds to complexity. Diffraction tends to ease alignment tolerance, but in an inelegant manner. A 100-transistor array with $\lambda/6$ spacing will have an extent of roughly 1.5λ per side. As the diffraction angle is something like the wavelength divided by the array diameter, the diffraction angle for such an array is a good fraction of a radian. One can say that grids are quasioptical, but in optics one generally doesn't use apertures much smaller than a millimeter (center optical wavelength of micrometers), for which the diffraction angle would be roughly a thousandth of a radian. As far as pure combining efficiency goes, grids are probably the optimal solution. However, more functionality may well be hard to obtain with this solution.

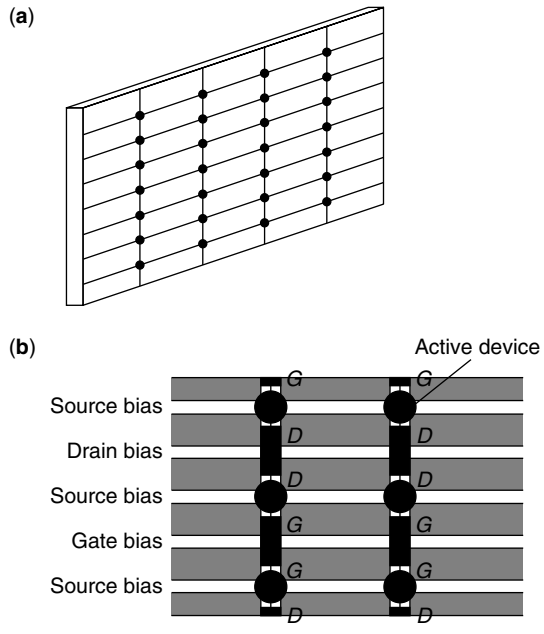


Figure 35. Schematic depiction of (**a**) the active surface of a grid oscillator and (**b**) a breakout of an internal region of the grid showing the active device placement relative to the bias lines.

As we have mentioned, there are a number of techniques for steering being investigated. There seems to be less work on modulation, and I do not know of any simultaneous steering of modulated beams to date. Although the field of active antennas began with the field of radiofrequency, it still seems to be in its infancy. However, as I hope this article has brought across, there is a significant amount of work ongoing, and the field of active antennas will grow in the future.

BIOGRAPHY

Alan Mickelson received his B.S.E.E. degree from the University of Texas, El Paso, in 1973 and his M.S. and Ph.D. degrees in electrical engineering from California Institute of Technology in 1974 and 1978, respectively. He was a National Academy of Sciences, Washington, D.C. visiting scientist at the Byurakan Astrophysical Observatory, Byurakan, Armenian S.S.R. in 1979–1980. Dr. Mickelson was a postdoctoral fellow at the Norwegian Institute of Technology, Norway, in 1980 and 1981 under a grant from the Norwegian National Science and Engineering Foundation and a staff scientist at the Electronics Laboratory of the same institute in 1982 and 1983. In 1984, he joined the faculty of the Electrical and Computer Engineering Department of the University of Colorado, Boulder. Since settling in Colorado, Professor Mickelson has continued to apply his background in electromagnetic theory and technique to a number of technological problems, including novel techniques to control microwave signal transmission and reception.

BIBLIOGRAPHY

1. J. A. Navarro and K. Chang, *Integrated Active Antennas and Spatial Power Combining*, Wiley, New York, 1995.
2. R. A. York and Z. B. Popović, eds., *Active and Quasi-Optical Arrays for Solid-State Power Combining*, Wiley, New York, 1997.

2a. H. Hertz, *Electric Waves*, Macmillan, New York, 1983. (This is a book of reprints of Hertz' work in the 1890s.)

3. J. Bardeen and W. Bratain, The transistor: A semiconductor triode, *Phys. Rev.* **74**: 435 (1948).
4. W. Shockley, The theory of $p-n$ junctions in semiconductors and $p-n$ junction transistors, *Bell Syst. Tech. J.* **28**: 435 (1949).
5. W. Shockley, A unipolar field-effect transistor, *Proc. IEEE* **40**: 1365–1376 (1952).
6. C. A. Mead, Schottky-barrier gate field-effect transistor, *Proc. IEEE* **54**: 307–308 (1966).
7. R. A. York, R. D. Martinez, and R. C. Compton, Active patch antenna element for array applications, *Electron. Lett.* **26**: 494–495 (March 1990).
8. A. R. Mickelson, *Physical Optics*, Van Nostrand Reinholt, New York, 1992, Chap. 2.
9. B. J. Hunt, *The Maxwellians*, Cornell Univ. Press, Ithaca, NY, 1991, Chap. 3.
10. H. C. Pocklington, Electrical oscillations in wires, *Proc. Cambridge Phil. Soc.* 324–333 (1897).
11. D. M. Pozar, *Microwave Engineering*, Addison-Wesley, Reading, MA, 1990.
12. P. E. Gray and C. L. Searle, *Electronic Principles, Physics, Models and Circuits*, Wiley, New York, 1967.
13. R. E. Collin, *Foundations for Microwave Engineering*, 2nd ed., McGraw-Hill, New York, 1992.
14. K. Chang, *Microwave Solid-State Circuits and Applications*, Wiley, New York, 1994.
15. K. Y. Chen et al., Analysis of an experimental technique for determining Van der Pol parameters of a transistor oscillator, *IEEE Trans. Microwave Theory Tech.* **46**: 914–922 (1998).
16. B. Van der Pol, Forced oscillations in a circuit with a nonlinear resistance, *Phil. Mag.* **3**: 65–80 (1927).
17. K. R. Carver and J. W. Mink, Microstrip antenna technology, *IEEE Trans. Antennas Propag.* **AP-29**: 2–24 (1981).
18. D. M. Pozar and D. H. Schauber, eds., *Microstrip Antennas*, IEEE Press, Piscataway, NJ, 1995.
19. R. E. Munson, Conformal microstrip antennas and microstrip phased arrays, *IEEE Trans. Antennas Propag.* **AP-22**: 74–78 (1974).
20. R. F. Harrington, *Time Harmonic Electromagnetic Fields*, McGraw-Hill, New York, 1961, p. 276.
21. K. Chang, K. A. Hammer, and G. K. Gopalakrishnan, Active radiating element using FET source integrated with microstrip patch antenna, *Electron. Lett.* **24**: 1347–1348 (1988).
22. R. Adler, A study of locking phenomena in oscillators, *Proc. IRE* **34**: 351–357 (1946).
23. R. Adler, A study of locking phenomena in oscillators, *Proc. IEEE* **61**: 1380–1385 (1973). (This is a reprint of Ref. 22.)
24. K. Kurokawa, Injection locking of microwave solid-state oscillators, *Proc. IEEE* **61**: 1386–1410 (1973).
25. K. D. Stephan, Inter-injection-locked oscillators for power combining and phased arrays, *IEEE Trans. Microwave Theory Tech.* **34**: 1017–1025 (1986).
26. K. D. Stephan and W. A. Morgan, Analysis of inter-injection-locked oscillators for integrated phased arrays, *IEEE Trans. Antennas Propag.* **35**: 771–781 (1987).
27. K. D. Stephan and S. L. Young, Mode stability of radiation-coupled inter-injection-locked oscillators for integrated phased arrays, *IEEE Trans. Microwave Theory Tech.* **36**: 921–924 (1988).
28. W. A. Morgan and K. D. Stephan, An x-band experimental model of a millimeter-wave inter-injection-locked phase array system, *IEEE Trans. Antennas Propag.* **36**: 1641–1645 (1988).

28a. M. J. Vaughan and R. C. Compton, 28 GHz omnidirectional quasioptical transmitter array, *IEEE Trans. Microwave Theory Tech.* **MTT-43**: 2507–2509 (1995).

29. R. A. York, Nonlinear analysis of phase relationships in quasioptical oscillator arrays, *IEEE Trans. Microwave Theory Tech.* **41**: 1799–1809 (1993).
30. P. Liao and R. A. York, A new phase-shifterless beam scanning technique using arrays of coupled oscillators, *IEEE Trans. Microwave Theory Tech.* **41**: 1810–1815 (1993).
31. R. A. York and R. C. Compton, Quasi-optical power combining using mutual synchronized oscillator arrays, *IEEE Trans. Microwave Theory Tech.* **39**: 1000–1009 (1991).
32. R. A. York and R. C. Compton, Coupled-oscillator arrays for millimeter-wave power-combining and mode-locking, *IEEE MTT-S Int. Microw. Symp. Digest*, 1992, pp. 429–432.
33. P. S. Hall and P. M. Haskins, Microstrip active patch array with beam scanning, *Electron. Lett.* **28**: 2056–2057 (1992).
34. P. S. Hall et al., Phase control in injection locked microstrip active antennas, *IEEE MTT-S Int. Microw. Symp. Digest*, 1994, pp. 1227–1230.
35. A. Zarrang, P. S. Hall, and M. Cryan, Active antenna phase control using subharmonic locking, *Electron. Lett.* **31**: 842–843 (1995).
36. T. S. M. Maclean and P. A. Ransdale, Short active aerials for transmission, *Int. J. Electron.* **36**: 261–269 (1974).
37. P. K. Rangole and S. S. Midha, Short antenna with active inductance, *Electron. Lett.* **10**: 462–463 (1974).
38. G. M. Rebeiz, Millimeter-wave and terahertz integrated circuit antennas, *Proc. IEEE* **80**: 1748–1770 (1996).
39. A. D. Frost, Parametric amplifier antennas, *Proc. IRE* **48**: 1163–1164 (1960).
40. J. R. Copeland and W. J. Robertson, Antenna-verters and antennafiers, *Electronics* 68–71 (1961).
41. M. E. Pedinoff, The negative conductance slot amplifier, *IRE Trans. Microwave Theory Tech.* **9**: 557–566 (1961).
42. W. J. Robertson, J. R. Copeland, and R. G. Verstraete, Antennafier arrays, *IEEE Trans. Antennas Propag.* **2**: 227–233 (1964).
43. K. Fujimoto, Active antennas: Tunnel-diode-loaded dipole, *Proc. IEEE* **53**: 174 (1964).
44. H. H. Meinke, Tunnel diodes integrated with microwave antenna systems, *Radio Electron. Eng.* **31**: 76–80 (1966).
45. K. J. Russell, Microwave power combining techniques, *IEEE Trans. Microwave Theory Tech.* **27**: 472–478 (1979).
46. K. Chang and C. Sun, Millimeter-wave power-combining techniques, *IEEE Trans. Microwave Theory Tech.* **31**: 91–107 (1983).
47. B. M. Armstrong et al., Use of microstrip impedance-measurement technique in the design of BARITT piplex Doppler sensor, *IEEE Trans. Microwave Theory Tech.* **28**: 1437–1442 (1980).

48. E. D. Sharp and M. A. Diab, Van Atta reflector array, *IRE Trans. Antennas Propag.* **8**: 436–438 (1960).

49. H. T. Friis and C. Feldman, A multiple-unit steerable antenna for short-wave reception, *Bell Syst. Tech. J.* **16**: 337–419 (1937).

50. C. Y. Pon, Retrodirective array using the heterodyne technique, *IEEE Trans. Antennas Propag.* **12**: 176–180 (1964).

51. B. S. Hewitt, The evolution of radar technology into commercial systems, *IEEE MTT-S Int. Microw. Symp. Digest*, 1994, pp. 1271–1274.

52. C. W. Poblanz and T. Itoh, A conformal retrodirective array for radar applications using a heterodyne phase scattering element, *IEEE MTT-S Int. Microw. Symp. Digest*, 1995, pp. 905–908.

53. Y. Chang, D. C. Scott, and H. R. Fetterman, Microwave phase conjugation using antenna coupled nonlinear optically pumped surface, *IEEE MTT-S Int. Microw. Symp. Digest*, 1995, pp. 1303–1306.

54. Z. B. Popovic, M. Kim, and D. B. Rutledge, Grid oscillators, *Int. J. Infrared Millimeter Waves* **9**: 647–654 (1988).

55. Z. B. Popovic et al., A 100-MESFET planar grid oscillator, *IEEE Trans. Microwave Theory Tech.* **39**: 193–200 (1991).

56. K. Y. Chen et al., Noninvasive experimental determination of charge and current distributions on an active surface, *IEEE Trans. Microwave Theory Tech.* **44**: 1000–1009 (1996).

ADAPTIVE ANTENNA ARRAYS

KYUNGJUNG KIM
TAPAN K. SARKAR
Syracuse University
Syracuse, New York

MAGDALENA SALAZAR PALMA
Universidad Politecnica de Madrid
Madrid, Spain

1. INTRODUCTION

Adaptive array signal processing has been used in many applications in such fields as radar, sonar, and wireless mobile communication. One principal advantage of an adaptive array is the ability to recover the desired signal while also automatically placing deep pattern nulls along the direction of the interference.

In conventional adaptive algorithms, the statistical approach based on forming an estimate of the covariance matrix of the received antenna voltages (measured voltages at the antenna terminals) without the signal is frequently used. However, these statistical algorithms suffer from two major drawbacks. First, they require independent identically distributed secondary data to estimate the covariance matrix of the interference. The formation of the covariance matrix is quite time-consuming, and so is the evaluation of its inverse. Unfortunately, the statistics of the interference may fluctuate rapidly over a short distance, limiting the availability of homogeneous secondary data. The resulting errors in the covariance matrix reduce the ability to suppress interference. The second drawback is that the estimation of the covariance matrix requires the storage and processing of the secondary data. This simply cannot be accomplished in real time for most applications.

Recently, a direct data domain algorithm has been proposed to overcome these drawbacks of a statistical technique [1–7]. In that approach one adaptively minimizes the interference power while maintaining the gain of the antenna array along the direction of the signal. Not having to estimate a covariance matrix leads to an enormous savings in memory and computer processor time and makes it possible to carry out an adaptive process in real time. The novelty of the proposed approach is that we analyze the antenna systems as spatial filters instead of treating them as temporal channels.

The use of real antenna elements and not omnidirectional point sources in an actual antenna array will also require an investigation into the capabilities of the direct data domain algorithms to perform adaptivity in nonideal situations such as in the presence of mutual coupling between the elements of the array, near-field scatterers, and obstacles located close to the array. This could also involve the various platform effects on which the antenna array is mounted.

Most adaptive algorithms assume that the elements of the receiving array are independent isotropic omnidirectional point sensors that do not reradiate the incident electromagnetic energy. It is further assumed that the array is isolated from its surroundings. However, in a practical case, array elements have a finite physical size and reradiate the incident fields. The reradiated fields interact with the other elements, causing the antennas to be mutually coupled. Adve and Sarkar [7] observed the degradation in the capabilities of direct data domain algorithms and suggested ways to improve it under some circumstances.

Gupta and Ksienski [8] and Pasala and Friel [9] compensate for the effects of mutual coupling by relating the open-circuit voltages (voltages at the ports of the array as if all were open-circuited) with the voltages measured at the ports in an adaptive antenna array used for direction of arrival (DoA) estimation. Adve and Sarkar [7] used the method of moments (MOM) to analyze the antenna array in which the entries of the MOM impedance matrix measure the interaction between the basis functions; that is, they quantize the mutual coupling. In these works the compensation matrix is in general considered to be independent of the angle of arrival of the signals. However, in a more practical environment the presence of near field scatterers (i.e., buildings the structure on which the array is mounted) will have effects on the array elements. The effects of these near field elements are similar to the effects of mutual coupling between the elements of the array. These environmental scatterers necessitate the development of a compensation matrix, which depends on the direction of arrival of signals including the undesired ones. In this article, we shall use the measured steering vector in an interpolation technique, which is contaminated by the presence of near field scatters as well as by the mutual coupling between

the elements of the real array, to obtain the compensation matrix for a more accurate numerical analysis.

This presentation is divided into two distinct parts. In the first part we use the electromagnetic analysis along with an interpolation algorithm to transform the voltages that are measured or computed in a real array containing realistic antenna elements receiving signals in the presence of near field scatterers to a uniform linear virtual array (ULVA) consisting of isotropic omnidirectional point radiators. In this way we take into account not only the effects of mutual coupling and the near-field scattering effects of the antenna array produced by the signal of interest but also those due to the strong coherent interferers whose directions of arrival are unknown. The first stage preprocesses the voltages induced in the real elements and transforms them to a set of voltages that would be induced in an ULVA of isotropic omnidirectional point radiators radiating in free space.

During the second phase of the processing, these transformed voltages induced in the ULVA are processed by a direct data domain least-squares method. In this phase, the goal is to estimate the complex signal amplitude given the direction of arrival in a least-squares fashion when the signal of interest (SoI) is contaminated by strong interferers that may come through the mainlobe of the array, clutter and thermal noise. The advantage of this methodology is that no statistical description of the environment is necessary, and since we are processing the data on a snapshot-by-snapshot basis, this new technique can be applied to a highly dynamic environment.

The article is organized as follows. In Section 2 we formulate the problem. In Section 3 we present the transformation technique incorporating mutual coupling effects between the array elements and near field scatterers. Section 4 describes the direct data domain least-squares approach. In Section 5 we present simulation results illustrating the performance of the proposed method in a real environment. Finally, in Section 6 we present the conclusion.

2. PROBLEM FORMULATION

Consider an array composed of N sensors separated by a distance d as shown in Fig. 1. We assume that narrowband signals consisting of the desired signal plus possibly coherent multipaths and jammers with center frequency f_0 are impinging on the array from various angles θ, with the constraint $0 \leq \theta \leq 180^\circ$. For sake of simplicity we assume that the incident fields are coplanar and that they are located in the far field of the array. However, this methodology can easily be extended to the noncoplanar case without any problem including the added polarization diversity.

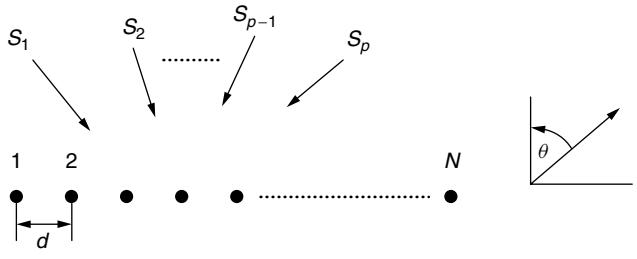


Figure 1. A linear uniform array.

Using the complex envelope representation, the $N \times 1$ complex vectors of phasor voltages $[\mathbf{x}(t)]$ received by the antenna elements at a single time instance t can be expressed by

$$[\mathbf{x}(t)] = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_N(t) \end{bmatrix} = \sum_{k=1}^{P} [\mathbf{a}(\theta_k)] s_k(t) + [n(t)] \tag{1}$$

where $s_k(t)$ denotes the incident signal from the kth source directed towards the array at the instance t, and P stands for the number of sources, $[\mathbf{a}(\theta)]$ denotes the steering vector of the array toward direction θ, and $[n(t)]$ denotes the noise vector at each of the antenna elements. It is important to note that the array elements can be dissimilar and they can be noncoplanar and may even be nonuniformly spaced. Here, the angle θ is measured from the broadside direction as shown in Fig. 1. We now analyze the data using a single snapshot of voltages measured at the antenna terminals.

Using a matrix notation, (1) becomes

$$[\mathbf{x}(t)] = [\mathbf{A}(\theta)][s(t)] + [n(t)] \tag{2}$$

where $[\mathbf{A}(\theta)]$ is the $N \times p$ matrix of the steering vectors, referred to as the array manifold

$$[\mathbf{A}(\theta)] = [a(\theta_1), a(\theta_2), \ldots, a(\theta_p)] \tag{3}$$

In a typical calibration methodology, a far-field source $s_k(t)$ is placed along the directions θ_k and then $[\mathbf{x}(t)]$ is the voltage measured at the feed point of the antenna elements in the array. Here $[s(t)]$ is a $p \times 1$ vector representing the various signals incident on the array at time instance t. In practice, this array manifold for a real array is contaminated by both effects of nonuniformity in the individual elements, and the interelement spacing may also be nonuniform to achieve greater aperture efficiency. Furthermore, there are mutual couplings between the antenna elements in the array, which undermine the performance of any conventional adaptive signal processing algorithm.

Hence, our problem can be stated as follows. Given the sampled data vector snapshot $[\mathbf{x}(t)]$ at a specific instance of time, how do we recover the desired signal arriving from a given look direction while simultaneously rejecting all other interferences that may be coherent? Most signal processing techniques are based on the fact that a far-field source presents a linear phase front at the elements of the antenna array. However, we shall demonstrate that the nonuniformity of a real array and the presence of mutual coupling between the elements of the real array and scatterers located close to the array undermines the ability of any adaptive algorithm to maintain the gain of the array along the direction of the signal while

simultaneously rejecting the interferences. To compensate for the lack of nonuniformity of the real array and mutual coupling effects, we propose an interpolation technique based on the method of least squares that incorporates all the electromagnetic coupling effects as outlined in Section 3. With appropriate preprocessing using Maxwell's equations, any adaptive technique can be applied to real antenna arrays located in any arbitrary environment. However, the use of a direct data domain least-squares procedure makes it possible to implement the algorithm in hardware, and the solution can be obtained in almost real time.

3. AN ARRAY TRANSFORMATION TECHNIQUE USING LEAST SQUARES THAT ACCOUNTS FOR ALL THE ELECTROMAGNETIC EFFECTS SUCH AS MUTUAL COUPLING AND PRESENCE OF NEAR FIELD SCATTERERS

For the first step of this adaptive method we transform the voltages that are induced in the actual antenna elements operating in any environment to a set of equivalent voltages that would be induced in a ULVA consisting of omnidirectional point radiators located in free space. The presence of mutual coupling between the antenna elements and existence of near field scatterers also disturb the capability of any algorithm to maintain the gain of the array along the direction of the signal while simultaneously rejecting the strong time varying coherent interferences. Hence, we need to preprocess the data to account for these undesired electromagnetic effects.

The preprocessing is to compensate for the lack of nonuniformity in a real array contaminated by the mutual coupling effects between the various elements. The methodology is similar to the one described by Friedlander [10–12]. The procedure is based on transforming the nonuniformly spaced array into a uniform linear virtual array (ULVA) consisting of isotropic omnidirectional point radiators operating in vacuum through the use of a transformation matrix. Our basic assumption is that electrical characteristics of the array corresponding to the ULVA can be obtained through an interpolation of the real array, which is disturbed by various undesired electromagnetic couplings. The goal is to select the best-fit transformation $[T]$ between the real array manifold $[A(\theta)]$ and the array manifold corresponding to a uniform linear virtual array (ULVA) consisting of isotropic omnidirectional point radiators $[\overline{A}(\theta)]$ such that $[T][A(\theta)] = [\overline{A}(\theta)]$ for all possible angles θ within a predefined sector. In this way we not only compensate for the various electromagnetic effects associated with the SoI but also correct for the interactions associated with coherent strong interferers whose direction of arrival we do not know. Since such a transformation matrix is defined within a predefined sector, the various undesired electromagnetic effects such as nonuniformity in spacing and mutual coupling between the elements and presence of near field obstacles for an array is made independent of the angular dependence.

The following is a step-by-step description of what needs to be done to obtain the transformation matrix $[T]$ that will transform the real array manifold that is disturbed by various undesired electromagnetic effects such as mutual coupling and various near-field effects to that of a ULVA:

1. The first step in designing the ULVA is to divide the field of view of the array into Q sectors. If the field of view is 180°, it can be divided into 6 sectors of 30° each. Then, each of the Q sectors is defined by the interval $[\theta_q, \theta_{q+1}]$, for $q = 1, 2, \ldots, Q$. Or equivalently, only one sector of 180° extent can also be used. In that case $Q = 1$.

2. Next we define a set of uniformly defined angles to cover each sector:

$$\mathbf{\Theta}_q = [\theta_q, \theta_q + \Delta, \theta_q + 2\Delta, \ldots, \theta_{q+1}] \tag{4}$$

where Δ is the angular step size.

3. We measure/compute the steering vectors associated with the set $\mathbf{\Theta}_q$ for the real array. This is done by placing a signal in the far field for each angle of arrival θ_q, $\theta_q + \Delta, \theta_q + 2\Delta, \ldots, \theta_{q+1}$. The measured/computed vector is different from the ideal steering vector, which is devoid of any undesired electromagnetic effects such as the presence of the mutual coupling between the nonuniformly spaced elements and other near-field coupling effects. Then, we obtain either through measurement or by using an electromagnetic analysis tool such as WIPL-D [13], to obtain the measured voltages at the antenna elements from

$$[\mathbf{A}_q(\mathbf{\Theta}_q)] = [a(\theta_q), a(\theta_q + \Delta) \cdots a(\theta_{q+1})] \tag{5}$$

This can be either actually measured or simulated and includes all the undesired electromagnetic coupling effects. Hence, each column of $[\mathbf{A}_q(\mathbf{\Theta}_q)]$ represents the relative signal strength received at each of the antenna elements for an incident signal arriving the angular direction θ_q. The elements of the matrix are a function of only the incident angle of an incoming plane wave within that predefined sector.

4. Next we fix the virtual elements of the interpolated array. We always assume that the ULVA consists of omnidirectional isotropic sources radiating in free space. We denote by the section of the array manifold of the virtual array obtained for the set of angles $\mathbf{\Theta}_q$:

$$[\overline{\mathbf{A}}_q(\mathbf{\Theta}_q)] = [\overline{a}(\theta_q), \overline{a}(\theta_q + \Delta), \ldots, \overline{a}(\theta_{q+1})] \tag{6}$$

where $[\overline{\mathbf{a}}(\theta)]$ is a set of theoretical steering vectors corresponding to the uniformly spaced linear array.

5. Now we evaluate the transformation matrix $[\mathbf{T}_q]$ for the sector q such that $[\mathbf{T}_q][\mathbf{A}(\Theta_q)] = [\overline{\mathbf{A}}(\Theta_q)]$ using the least-squares method. This is achieved by minimizing the functional

$$\min_{T_q} \|[\overline{\mathbf{A}}_q] - [\mathbf{T}_q][\mathbf{A}_q]\| \tag{7}$$

In order to have a unique solution for (7), the number of direction vectors in a given sector must be greater than or equal to the number of the elements of array. The least square solution to (7) is given by [14]

$$[\mathbf{T}] = [\overline{\mathbf{A}}(\Theta_q)][\mathbf{A}(\Theta_q)]^H\{[(\mathbf{A}(\Theta_q)][\mathbf{A}(\Theta_q)]^H\}^{-1} \tag{8}$$

where the superscript H represents the conjugate transpose of a complex matrix. Computationally it is more efficient and accurate to carry out the solution of (7)

through the use of the total least squares implemented through the singular value decomposition [15]. The transformation matrix needs to be computed only once *a priori* for each sector and the computation can be done offline. Hence, once [**T**] is known, we can compensate for the various undesired electromagnetic effects such as mutual coupling between the antenna elements, including the effects of near-field scatterers, as well as nonuniformity in the spacing of the elements in the real array simultaneously. The transformation matrix [**T**] is thus characterized within the predefined angle. However, if there is only one sector, specifically, $Q = 1$, then there will be only one transformation matrix [**T**].

6. Finally, using (8), one can obtain the corrected input voltages in which all the undesired electromagnetic effects are accounted for and the measured snapshot of the voltages are transformed to that which will be obtained for a ULVA. Let that set be denoted by $[\mathbf{x}_c(t)]$. Its value can be obtained through

$$[\mathbf{x}_c(t)] = [\mathbf{T}][\mathbf{x}(t)] \tag{9}$$

Once (9) is obtained, we can apply the direct data domain algorithms to the preprocessed corrected voltages $[\mathbf{x}_c(t)]$ without any significant loss of accuracy.

Next a direct data domain least-squares algorithm is applied to the processed voltage sets $[\mathbf{x}_c(t)]$ to obtain the complex amplitude corresponding to the signal of interest in a least-squares fashion.

4. THE DIRECT DATA DOMAIN LEAST-SQUARES PROCEDURE

Let us assume that the signal of interest (SoI) is coming from the angular direction θ_d and that our objective is to estimate its complex amplitude while simultaneously rejecting all other interferences. The signal arrives at each antenna at different times dependent on the direction of arrival of the SoI and the geometry of the array. We make the narrowband assumption for all the signals including the interferers. At each of the N antenna elements, the received signal given by (9) is a sum of the SoI, interference, and thermal noise. The interference may consist of coherent multipaths of SoI along with clutter and thermal noise. Here we model clutter as a bunch of electromagnetic waves coming through an angular sector. Hence, this model of clutter does not require any statistical characterization of the environment [1–7]. Therefore, we can reformulate (9) as

$$[\mathbf{x}(t)] = [\mathbf{s}_d] + \sum_{p=1}^{P-1} s_p\, a(\theta_p) + [\mathbf{n}(t)] \tag{10}$$

where s_p and θ_p are the amplitude and direction of arrival of the pth interference, respectively, and s_d is the SoI. If θ_d is the assumed DoA of the SoI, then we can represent the received voltage solely due to the desired signal at the kth sensor element as

$$s_d = s_d(t)e^{j\psi(\theta_d)} \tag{11}$$

The strength of the SoI, $s_d(t)$, is the desired unknown parameter that will be estimated for the given snapshot at the time instance t. $\psi(\theta_d)$ does not provide a linear phase regression along the elements of the real array, when the elements deviate from isotropic omnidirectional point sensors. This deviation from phase linearity undermines the capabilities of the various signal processing algorithms. For a conventional adaptive array system, we can now estimate the SoI by a weighted sum given by

$$y(t) = \sum_{k=1}^{K} w_k x_k(t) \tag{12}$$

or in a compact matrix form as

$$[\mathbf{y}(t)] = [\mathbf{W}]^T[\mathbf{X}] = [\mathbf{X}]^T[\mathbf{W}] \tag{13}$$

where the superscript T denotes the transpose of a matrix. The two vectors [**W**] and [**X**] are given by

$$[\mathbf{W}]^T = [w_1, w_2, \ldots, w_K] \tag{14}$$

$$[\mathbf{X}]^T = [x_1, x_2, \ldots, x_K] \tag{15}$$

Let [**V**] be a matrix whose elements consist of the complex voltages measured at a single time instance t at all the N elements of the array simultaneously. The received signals may also be contaminated by thermal noise. Let us define another matrix [**S**] whose elements comprise of the complex voltages received at the antenna elements of the ULVA due to a signal of unity amplitude coming from the desired direction θ_d. However, the actual complex amplitude of the signal is α, which is to be determined. Then if we form the matrix pencil using these two matrices, we have

$$[\mathbf{V}] - \alpha[\mathbf{S}] \tag{16}$$

where

$$[\mathbf{V}] = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & & & \\ x_K & x_{K+1} & \cdots & x_N \end{bmatrix}_{K\times K} \tag{17}$$

$$[\mathbf{S}] = \begin{bmatrix} s_{d1} & s_{d2} & \cdots & s_{dK} \\ s_{d2} & s_{d3} & \cdots & s_{dK+1} \\ \vdots & & & \\ s_{dK} & s_{dK+1} & \cdots & s_{dN} \end{bmatrix}_{K\times K} \tag{18}$$

represent only the undesired signal components. The difference between each element of $\{[\mathbf{V}]-\alpha[\mathbf{S}]\}$ represents the contribution of all the undesired signals due to coherent mutipaths, interferences, clutter, and thermal noise (i.e., all undesired components except the signal). It is assumed that there are K equivalent interferers, and so the number of degrees of freedom is $K = (N+1)/2$. One could form the undesired noise power from (16) and estimate a value of α by using a set of weights [**W**], which minimizes the noise power. This results in [1–7]

$$([\mathbf{V}] - \alpha[\mathbf{S}])[\mathbf{W}] = [\mathbf{0}] \tag{19}$$

Alternately, one can view the left-hand side of (19) as the total noise signal at the output of the adaptive processor due to interferences and thermal noise:

$$[\mathbf{N}_{\text{out}}] = [\mathbf{R}][\mathbf{W}] = \{[\mathbf{V}] - \alpha[\mathbf{S}]\}[\mathbf{W}] \tag{20}$$

Hence, the total undesired power would be given by

$$[\mathbf{P}_{\text{undesired}}] = [\mathbf{W}]^H\{[\mathbf{V}] - \alpha[\mathbf{S}]\}^H\{[\mathbf{V}] - \alpha[\mathbf{S}]\}[\mathbf{W}] \tag{21}$$

where the superscript H denotes the conjugate transpose of a matrix. Our objective is to set the undesired power to a minimum by selecting [**W**] for a fixed signal strength α. This yields the generalized eigenvalue equation given by (19). Therefore

$$[\mathbf{V}][\mathbf{W}] = \alpha[\mathbf{S}][\mathbf{W}] \tag{22}$$

where α, the strength of the signal, is given by the generalized eigenvalue and the weights [**W**] are given by the generalized eigenvector. Even though (22) represents a $K \times K$ matrix, the matrix [**S**] is only of rank 1. Hence, (22) has only one eigenvalue, and that generalized eigenvalue is the solution for the SoI.

For real-time applications, it may be computationally difficult to solve the reduced-rank generalized eigenvalue problem in an efficient way, particularly if the dimension K, i.e., the number of weights is large. For this reason we convert the solution of a nonlinear eigenvalue problem in (22) to the solution of a linear matrix equation.

We observe that the first and the second elements of the matrix [**R**] in (20) is given by

$$R(1) = x_1 - \alpha s_{d1} \tag{23}$$

$$R(2) = x_2 - \alpha s_{d2} \tag{24}$$

where x_1 and x_2 are the voltages received at the antenna elements 1 and 2 due to the signal, interferences and thermal noise, whereas s_{d1} and s_{d2} are the values of the signals only at the same elements due to an assumed incident signal of unit strength.

Define

$$Z = \exp\left[j2\pi\frac{d}{\lambda}\sin\theta_d\right] \tag{25}$$

where θ_d is the angle of arrival corresponding to the desired signals. Then $R(1) - Z^{-1}R(2)$ contains no components of the signal as

$$s_{d1} = \exp\left[j2\pi\frac{id}{\lambda}\sin\theta_d\right] \quad \text{with } i = 1 \tag{26}$$

$$s_{d2} = \exp\left[j2\pi\frac{id}{\lambda}\sin\theta_d\right] \quad \text{with } i = 2 \tag{27}$$

Therefore one can form a reduced-rank matrix $[\mathbf{U}]_{(K-1)\times K}$ generated from [**R**] such that

$$[\mathbf{U}] = \begin{bmatrix} X_1 - Z^{-1}X_2 & X_2 - Z^{-1}X_3 & \cdots & X_K - Z^{-1}X_{K+1} \\ \vdots & \vdots & & \\ X_{K-1} - Z^{-1}X_K & X_K - Z^{-1}X_{K+1} & \cdots & X_{N-1} - Z^{-1}X_N \end{bmatrix}_{(K-1)\times K} = [0] \tag{28}$$

In order to make the matrix full rank, we fix the gain of the subarray by forming a weighted sum of the voltages $\sum_{i=1}^{K} W_iX_i$ along the direction of arrival of the SoI. Let us say that the gain of the subarray is C in the direction of θ_d. This provides an additional equation resulting in

$$\begin{bmatrix} 1 & \cdots & Z^{K-1} \\ X_1 - Z^{-1}X_2 & \cdots & X_K - Z^{-1}X_{K+1} \\ \vdots & \vdots & \vdots \\ X_{K-1} - Z^{-1}X_K & \cdots & X_{N-1} - Z^{-1}X_N \end{bmatrix}_{K\times K} \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_K \end{bmatrix}_{K\times 1} = \begin{bmatrix} C \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{K\times 1} \tag{29}$$

or, equivalently,

$$[\mathbf{F}][\mathbf{W}] = [\mathbf{Y}] \tag{30}$$

Once the weights are solved by using (30), the signal component α may be estimated from

$$\alpha = \frac{1}{C}\sum_{i=1}^{K} W_iX_i \tag{31}$$

The proof of (29–31) is available in Ref. 1. As noted in that article [1], (29) can be solved very efficiently by applying the FFT and the conjugate gradient method, which may be implemented to operate in real time utilizing, for example, a DSP32C signal processing chip [16].

So for the solution of $[\mathbf{F}][\mathbf{W}] = [\mathbf{Y}]$ in (30), the conjugate gradient method starts with an initial guess $[W]_0$ for the solution and lets [16]

$$[\mathbf{P}]_0 = -b_{-1}[\mathbf{F}]^H[\mathbf{R}]_0 = -b_{-1}[\mathbf{F}]^H\{[\mathbf{F}][\mathbf{W}]_0 - [\mathbf{Y}]\}, \tag{32}$$

At the nth iteration the conjugate gradient method develops

$$t_n = \frac{1}{\|[\mathbf{F}][\mathbf{P}]_n\|^2} \tag{33}$$

$$[\mathbf{W}]_{n+1} = [\mathbf{W}]_n + t_n[\mathbf{P}]_n \tag{34}$$

$$[\mathbf{R}]_{n+1} = [\mathbf{R}]_n + t_n[\mathbf{Z}][\mathbf{P}]_n \tag{35}$$

$$b_n = \frac{1}{\|[\mathbf{F}]^H[\mathbf{R}]_{n+1}\|^2} \tag{36}$$

$$[P]_{n+1} = [\mathbf{P}]_n - b_n[\mathbf{F}]^H[\mathbf{R}]_{n+1} \tag{37}$$

The norm is defined by

$$\|[\mathbf{F}][\mathbf{P}]_n\|^2 = [\mathbf{P}]_n^H[\mathbf{F}]^H[\mathbf{F}][\mathbf{P}]_n \tag{38}$$

These iterative procedures continue until the error criterion is satisfied. In our case, the error criterion is defined by

$$\frac{\|[\mathbf{F}][\mathbf{W}]_n - [\mathbf{Y}]\|}{\|\mathbf{Y}\|} \le \sigma \tag{39}$$

where σ denotes the number of effective bits of data associated with the measured voltages. Hence, the iteration is stopped when the normalized residuals are of the same order as the error in the data. The computational bottleneck in the application of the conjugate gradient method is to carry out the various matrix vector products. That is where the FFT comes in as the equations involved have a Hankel structure and therefore use of the FFT reduces the computational complexity by an order of magnitude without sacrificing accuracy [17].

The advantage of using the conjugate gradient method is that the iterative solution procedure will converge even if the matrix [**F**] is exactly singular. Hence, it can be used for real time implementations. Also, the number of iterations taken by the conjugate gradient method to converge to the solution is dictated by the number of independent eigenvalues of the matrix [**F**]. This often translates into the number of dominant signal components in the data. So, the conjugate gradient method has the advantage of a direct method as it is guaranteed to converge after a finite number of steps and that of an iterative method as the roundoff and the truncation errors in the computations are limited only to the last stage of iteration.

Next we illustrate how to increase the degrees of freedom associated with (30). It is well known in the parametric spectral estimation literature that a sampled sequence can be estimated by observing it in either the forward or reverse direction [1–7]. This we term as the backward model as opposed to the forward model just outlined. If we now conjugate the data and form the reverse sequence, then we get an equation similar to (29) for the solution of weights W_m:

$$\begin{bmatrix} 1 & Z & \cdots & Z^{K-1} \\ X_N^* - Z^{-1}X_{N-1}^* & X_{N-1}^* - Z^{-1}X_{N-2}^* & \cdots & X_K^* - Z^{-1}X_{K+1}^* \\ \vdots & \vdots & & \\ X_{K+1}^* - X_K^* & X_K^* - X_{K-1}^* & \cdots & X_2^* - Z^{-1}X_1^* \end{bmatrix}_{K\times K} \times \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_K \end{bmatrix}_{K\times 1} = \begin{bmatrix} C \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{K\times 1} \tag{40}$$

where $*$ denotes the complex conjugate, or equivalently

$$[\mathbf{B}][\mathbf{W}] = [\mathbf{Y}] \tag{41}$$

The signal strength α can again be determined by (31), once (40) is solved for the weights. C is the gain of the antenna array along the direction of the arrival of the signal. Note that in both cases of equations (30) and (41) [**F**] and [**B**], $K = (N+1)/2$, where N is the total number of antenna elements. So we now increase the number of weights significantly by combining the forward–backward model. In this way we double the amount of data by not only considering the data in the forward direction but also conjugating it and reversing the increment direction of the independent variable. This type of processing can be done as long as the series to be approximated can be fit by exponential functions of purely imaginary argument. This is always true for the adaptive array case. There is an additional benefit. For both the forward and the backward methods, the maximum number of weights we can consider is given by $(N-1)/2$, where N is the number of antenna elements. Hence, even though all the antenna elements are being utilized in the processing, the number of degrees of freedom available is essentially half that of the number of antenna elements. For the forward–backward method, the number of degrees of freedom can be significantly increased without increasing the number of antenna elements. This is accomplished by considering the forward–backward version of the array data. For this case, the number of degrees of freedom M can reach $(N+0.5)/1.5$. This is approximately equal to 50% more weights or numbers of degrees of freedom than the two previous cases. The equation that needs to be solved for the weights in the combined forward–backward model is obtained by combining (29) and (40) into

$$\begin{bmatrix} 1 & Z & \cdots & Z^{M-1} \\ X_1 - Z^{-1}X_2 & X_2 - Z^{-1}X_3 & \cdots & X_M - Z^{-1}X_{M+1} \\ \vdots & \vdots & \vdots & \\ X_{M-1} - Z^{-1}X_M & X_M - Z^{-1}X_{M+1} & \cdots & X_{M-1} - Z^{-1}X_N \\ X_N^* - Z^{-1}X_{N-1}^* & X_{N-1}^* - Z^{-1}X_{N-2}^* & \cdots & X_M^* - Z^{-1}X_{M+1}^* \\ \vdots & \vdots & & \\ X_{M+1}^* - X_M^* & X_M^* - X_{M-1}^* & \cdots & X_2^* - Z^{-1}X_1^* \end{bmatrix}_{M\times M} \times \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix}_{M\times 1} = \begin{bmatrix} C \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{M\times 1} \tag{42}$$

or, equivalently,

$$[\mathbf{U}][\mathbf{W}] = [\mathbf{Y}] \tag{43}$$

Once the increased degrees of freedom are used to compute the weights the complex amplitude for the signal of interest is determined from Eq. (31), where in the summation N is replaced by the new degrees of freedom M. Also as before the matrix [**U**] now has a block Hankel structure.

This illustrates how the direct data domain least-squares approach can be implemented in real time by using single snapshots of data.

5. NUMERICAL EXAMPLES

In this section we illustrate the principles described above through some numerical simulations.

5.1. Application in Absence of Mutual Coupling

As a first example consider a signal of unit amplitude arriving from $\theta = 0^\circ$. We consider a 17-element array of element spacing of $\lambda/2$ as shown in Fig. 1. The magnitude of the incident signal is varied from 1 to 10.0 V/m in steps of 0.1 V/m while maintaining the jammer intensities constant, which are arriving from -50°, -30°, 20°. The signal-to-thermal noise ratio at each antenna element is set at 20 dB. All signal intensities and directions of arrival are summarized in Table 1.

Table 1. Parameters of the Incident Signals

	Magnitude(V/m)	Phase	DoA
Signal	1.0–10.0	0.0	0°
Jammer	1.0	0.0	−50°
Jammer	1.0	0.0	−30°
Jammer	1.0	0.0	20°

Here, we assume that we know the direction of arrival of the signal but need to estimate its complex amplitude.

If the jammers have been nulled correctly and the signal recovered properly, it is expected that the recovered signal will have a linear relationship with respect to the intensity of the incident signal. Figure 2 plots the results of using the direct data domain approach presented in Section 3. The magnitude and the phase are shown. As can be seen, the magnitude displays the expected linear relationship, and the phase varies within very small values. The beam pattern associated with this example is shown in Fig. 3. Here, we set the magnitude of the desired signal to be 1 V/m and the other parameters are as given in Table 1. The nulls are deep and occur along the correct directions.

For the second example, the intensity of the jammer signal is varied from 1 to 1000 V/m in 10-V/m increments while the intensity of the desired signal is fixed at 1 V/m. All signal intensities and directions of arrival are summarized in Table 2.

The signal-to-thermal noise ratio at each antenna element is set at 20 dB.

Figure 4 shows the results of using the direct data domain approach. The fluctuations of the magnitude and phase of the estimated signal are very small even when there is a very strong interference.

The beam pattern associated with this adaptive system when the field strength of the strong jammer is 500 V/m is shown in Fig. 5. This demonstrates that the strong jammer

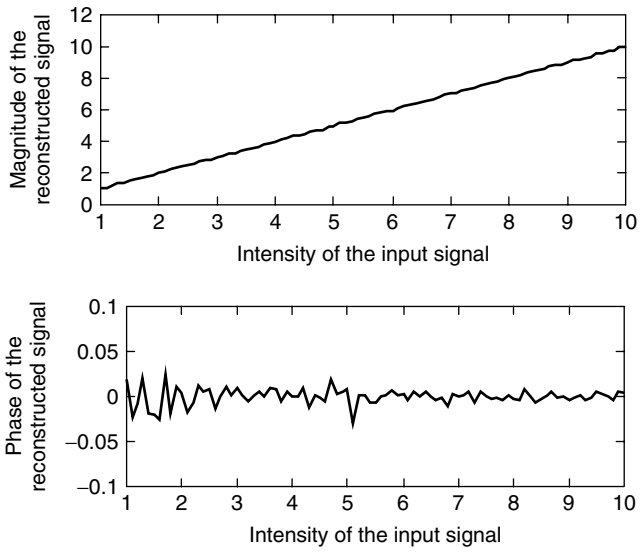


Figure 2. Estimation of the signal of interest (SoI) in the presence of jammers and thermal noise.

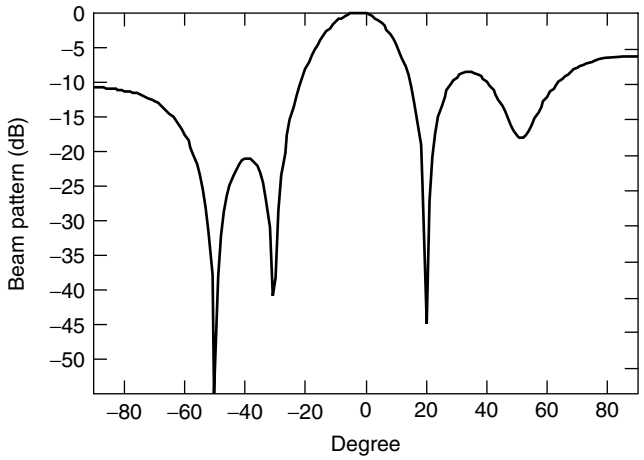


Figure 3. Adapted beam pattern in the presence of the jammers and thermal noise.

Table 2. Parameters for the SoI and Interference

	Magnitude(V/m)	Phase	DoA
Signal	1.0	0.0	0°
Jammer	1	0.0	−50°
Jammer	1	0.0	−30°
Jammer	1–1000	0.0	20°

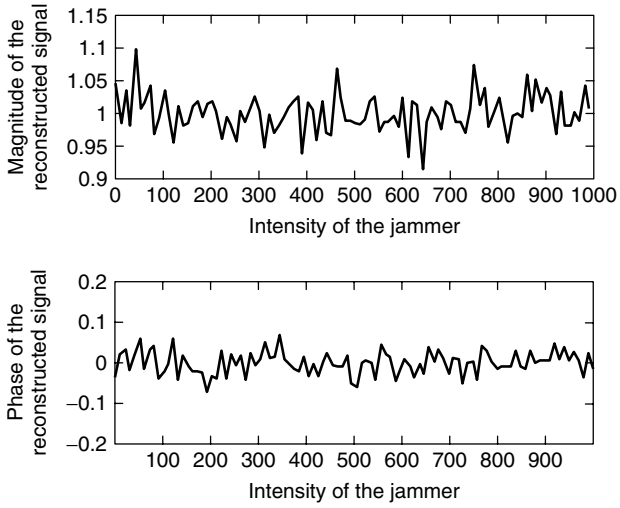


Figure 4. Estimation of the reconstructed signal in the presence of the strong jammer.

has been suppressed enough so as to recover the proper amplitude of the signal.

5.2. Application in Presence of Mutual Coupling

Next we consider a semi circular array consisting of half-wave dipoles as shown in Fig. 6. It consists of 24 half-wave thin-wire dipole antenna elements. Each element is identically loaded at the center by 50 Ω. The radius of the semicircular array is 3.82 wavelength. The dipoles are z-directed, of length $L = \lambda/2$ and radius $r = \lambda/200$, where λ is the wavelength of operation. The details of the semicircular are presented in Table 3.

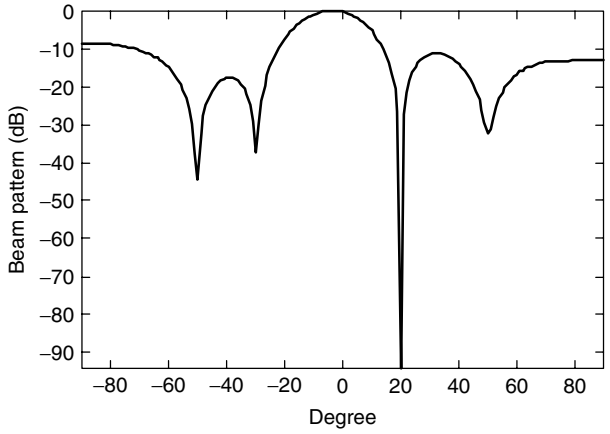


Figure 5. Adapted Beam pattern in the presence of the jammers.

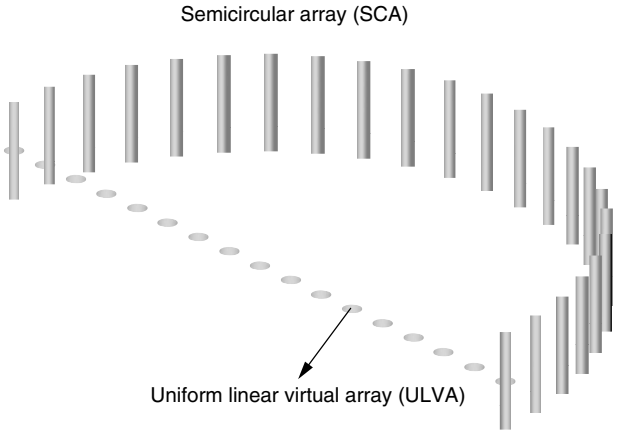


Figure 6. A semicircular array.

Table 3. Physical Sizes of the Elements for the Semicircular Array

Number of elements in semicircular array	24
Length of z-directed wires	$\lambda/2$
Radius of wires	$\lambda/200$
Loading at the center	50Ω

Then, as described in Section 4, the real array is interpolated into a ULVA consisting of $M = 17$ isotropic omnidirectional point sources separated by a distance d/λ. Typically d is chosen to be close to $\lambda/2$. By choosing the reference point of the ULVA at the center of the real array, the steering vectors associated with the virtual array are given by

$$\overline{a}(\theta) = \begin{bmatrix} \exp\left(\frac{j2\pi kd}{\lambda}\cos\theta\right), \ldots, \exp\left(\frac{j2\pi 2d}{\lambda}\cos\theta\right), \\ \times \exp\left(\frac{j2\pi d}{\lambda}\cos\theta\right), \quad 1, \\ \exp\left(\frac{j2\pi d}{\lambda}\cos\theta\right), \exp\left(\frac{j2\pi 2d}{\lambda}\cos\theta\right) \ldots, \\ \times \exp\frac{j2\pi kd}{\lambda}\cos\theta \end{bmatrix}^T_{N\times 1}, \tag{44}$$

The distance d between the elements of the ULVA was chosen to be 0.4775λ. The incremental size Δ in the interpolation region, $\Theta = [-\theta_q, \theta_{q+1}] = [-60, 60^\circ]$, is chosen to be 1°. In this case $Q = 1$. The sector chosen here, for example, is of 120° symmetrically located. Then, a set of real steering vectors are computed for the sources located at each of the angles $\theta_q, \theta_q + \Delta, \theta_q + 2\Delta, \ldots, \theta_{q+1}$. The computed vector $[\overline{a}(\theta)]$ is then distorted from the ideal steering vector due to the presence of mutual coupling between the elements of the real array. The actual steering vectors having all the undesired electromagnetic effects are computed using WIPL-D [13]. Then, using (8) we obtain the transformation matrix to compensate for the effects of nonuniformity in spacing and the presence of mutual coupling between the elements of the real array. Finally, using (69), we can obtain the corrected input voltage in which the nonuniformity in spacing and mutual coupling effects are eliminated from the actual voltage.

Next, the magnitude of the incident SoI is varied from 1 to 10.0 V/m in increments of 0.01 V/m while maintaining the jammer intensities constant, which are arriving from -20°, 40°, 50°. The direction of arrival of the SoI is 10°. The signal-to-thermal noise ratio at each antenna element is set at 20 dB. All signals intensities and directions of arrival are given by the data in Table 4.

If the jammers have been nulled correctly and the SoI recovered properly, it is expected that the recovered signal will have a linear relationship with respect to the intensity of the incident signal, implying that the various electromagnetic effects have been properly accounted for. The estimate for the SoI in Fig. 7a shows that the mutual coupling between the elements of the real array undermines the performance of the direct data domain approach if the various electromagnetic effects are not accounted for. Fig. 7b illustrates the superiority of the results when the direct data domain least-squares approach is used after preprocessing the data to take into account the various mutual coupling effects. The estimated amplitude and the phase for the SoI are shown in Fig. 7b. Here, the amplitude displays the expected linear relationship and the phase changes very little from zero degrees.

The beam patterns associated with this example are shown in Fig. 8a,b. Here, we set the amplitude of the SoI to be 1 V/m, and the other parameters are as given in Table 4. Figure 8b illustrates that the nulls are deep and are located along the correct directions. This indicates that the two-step procedure illustrated in this article have properly modeled the real environment nulling the relevant interferers. However, we see that in Fig. 8a the nulls in the beam pattern are shallow and are displaced from their desired locations, as the undesired

Table 4. Parameters of the Signals

	Magnitude(V/m)	Phase	DoA
Signal	1.0 −10.0	0.0	10°
Jammer	1.0	0.0	-20°
Jammer	1.0	0.0	40°
Jammer	1.0	0.0	50°

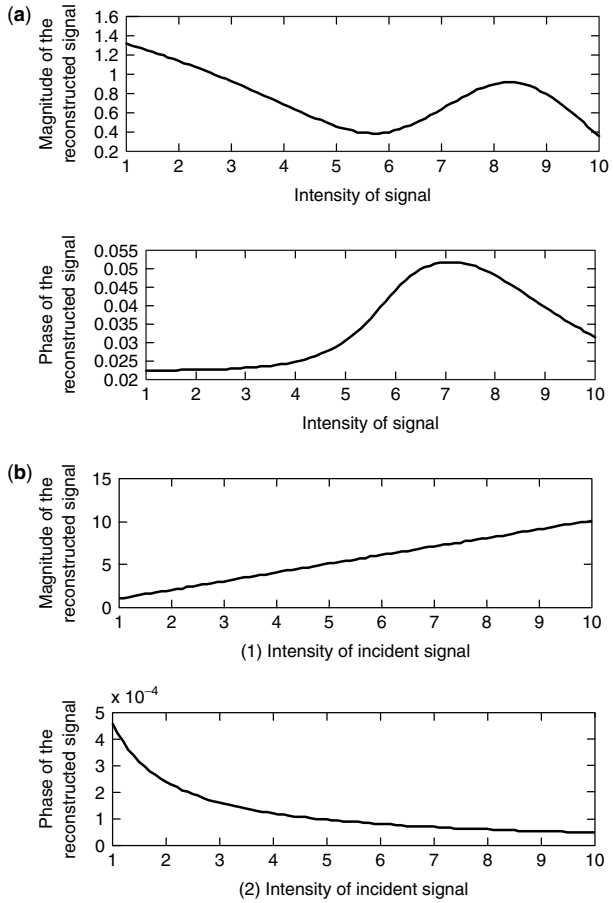


Figure 7. Estimation of the SoI (**a**) without and (**b**) after compensating for mutual coupling.

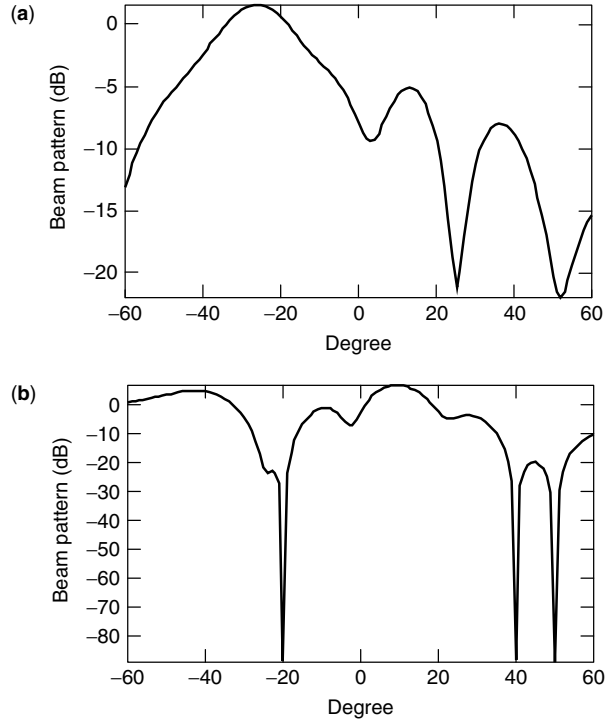


Figure 8. Adapted beam pattern (**a**) without and (**b**) after compensating for the mutual coupling.

electromagnetic effects have not been properly taken into account. Hence, in that case the SoI cannot be recovered.

For the final example we consider the effects of large near-field scatterers located close to the semicircular array. As shown in Fig. 9, there is a large structure located within a distance, that is 5 times the radius of the semicircular array and is oriented along the direction of -20°. The width of the structure is 7.64 wavelengths, and the height is 15.28 wavelengths. Hence, the semicircular array and the scatterer have strong electromagnetic coupling in addition to the presence of mutual coupling between the elements. We again consider the case of four incoming signals from -20°, 10°, 40°, 50°. The parameters for the desired signal and the jammers are summarized in Table 4.

After we compensate for the various electromagnetic couplings and project the data to that due to a ULVA, we solve for the weights [**W**] using (43). Then, we estimate the amplitude of the desired signal through (31). Fig. 10a,b plots the amplitude and the phase for the SoI in both presence and absence of mutual coupling between the

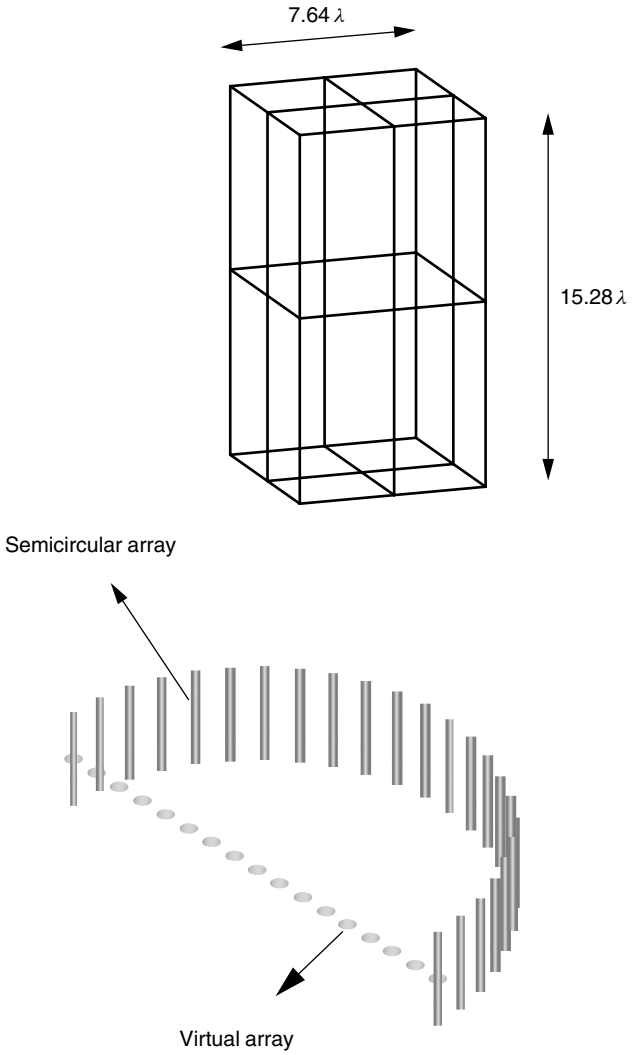


Figure 9. A semicircular array operating in the presence of a large obstacle.

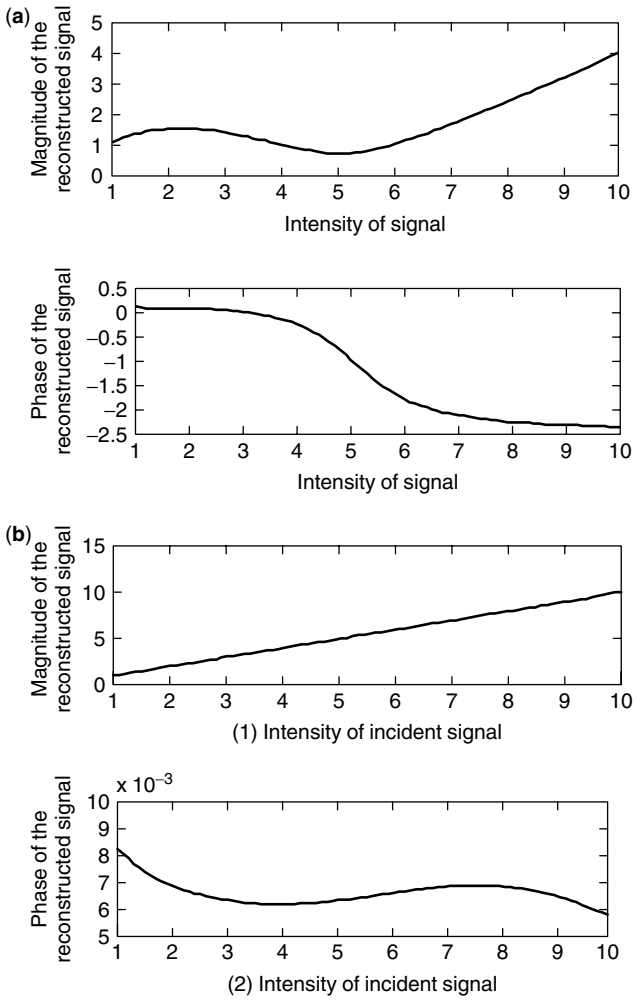


Figure 10. Estimation of SoI (**a**) without and (**b**) after compensating for the mutual coupling and the near-field scatterer.

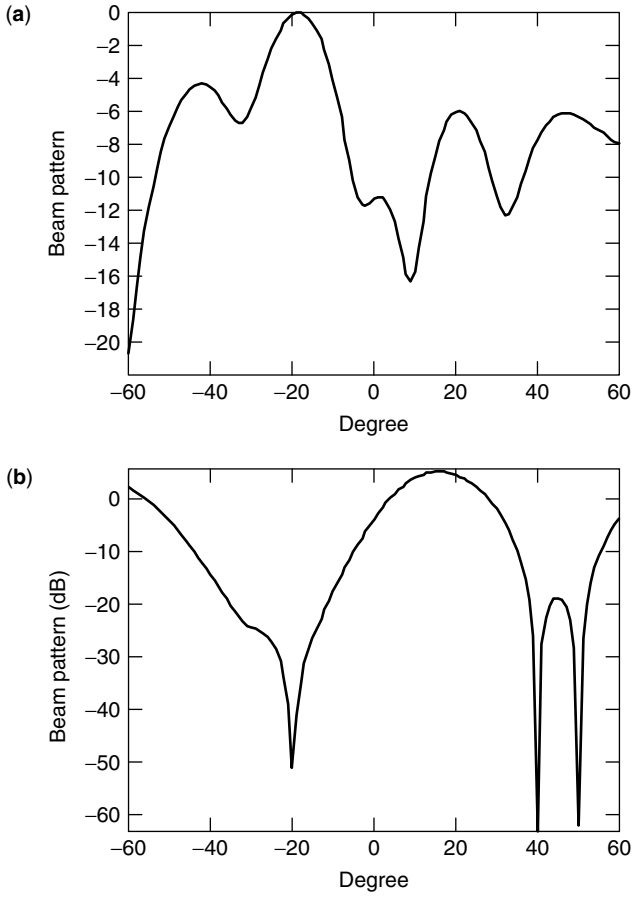


Figure 11. Adapted beam pattern (**a**) without and (**b**) after compensating for the mutual coupling and the near-field scatterer.

elements of the array and the scatterer located close to the array. As can be seen, after compensation of the undesired electromagnetic effects, the expected linear relationship is clearly seen, implying that the jammers have been nulled and the SoI estimated with a good accuracy.

The adapted beam patterns associated with this example are shown in Fig. 11a,b for the two cases considered above. When the mutual coupling is neglected, the beam pattern in Fig. 11a clearly indicates that the interferers have not been nulled in a correct fashion. However, when the electromagnetic effects have been appropriately accounted for, the beam points correctly along the direction of SoI while simultaneously placing deep nulls along the direction of the interferers. By comparing the adapted beam patterns in Figs. 8b and 11b it is seen that the presence of a large near-field scatterer has indeed produced a wide null along that direction due to the diffraction effects of the interferer.

6. CONCLUSION

The objective of this article has been to present a two-step process for using adaptive antenna arrays operating in a real environment. In the first step a transformation matrix is determined that transforms the voltages induced at the feed points of the antenna elements operating in the presence of near-field scatterers and the presence of mutual coupling between the antenna elements to that of the voltages induced in a uniform linear virtual array consisting of isotropic omnidirectional point radiators operating in free space. Such a transformation takes into account all the electromagnetic interactions between the antenna elements and its environment. The next step in the solution procedure involves applying a direct data domain least-squares approach that estimates the complex amplitude of the signal of interest given its direction of arrival. The signal of interest may be accompanied by coherent multipaths and interferers, which may be located in an angle quite close to the direction of arrival of the signal. In addition, there may be clutter and thermal noise at each of the antenna elements. In this approach, since no statistical methodology is employed, there is no need to compute a covariance matrix. Therefore, this procedure can be implemented on a general-purpose digital signal processor for real-time implementations.

BIOGRAPHIES

Kyungjung Kim was born in Seoul, Korea. He received the B.S. degree from Inha University, Korea, and the

M.S. degree from Syracuse University, Syracuse, New York. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering at Syracuse University.

He was a research assistant at Syracuse University from 1998 to 2001 and a graduate fellow from 2001 to 2002. His current research interests include adaptive array signal processing and wavelet transform.

Tapan Kumar Sarkar received the B. Tech. degree from the Indian Institute of Technology, Kharagpur, India, in 1969; the M.Sc.E. degree from the University of New Brunswick, Fredericton, Canada, in 1971; and the M.s. and Ph.D. degrees from Syracuse University; Syracuse, New York, in 1975. He is now a professor in the Department of Electrical and Computer Engineering at Syracuse University. His current research interests deal with numerical solutions of operator equations arising in electromagnetics and signal processing with application to system design. He obtained one of the "best solution" awards in May 1977 at the Rome Air Development Center (RADC) Spectral Estimation Workshop. He has authored or coauthored more than 250 journal articles and numerous conference papers and has written chapters in 28 books and 10 books including the latest ones on *Iterative and Self Adaptive Finite-Elements in Electromagnetic Modeling* and *Wavelet Applications in Engineering Electromagnetics* by Artech House.

Dr. Sarkar is a registerd professional engineer in the State of New York. He received the Best Paper Award of the IEEE Transactions on Electromagnetic Compatibility in 1979 and in the 1997 National Radar Conference. He received the College of Engineering Research Award in 1996 and the chancellor's citation for excellence in research in 1998 at Syracuse University. He received the title Docteur Honoris Causa from Universite Blaise Pascal, Clermont Ferrand, France, in 1998, and he was awarded the medal of Friends of Clermont Ferrand by the mayor of the city in 2000.

Magdalena Salazar-Palma receive dhte degree of *Ingeniero de Telecomunicación* and the Ph.D. degree from the *Universidad Politècnica de Madrid* (Madrid, Spain), where she is a *Profesor Titular* of the *Departamento de Señales, Sistemas y Radiocomunicaciones* (Signals, Systems and Radiocommunications Department) at the *Escuela Técnica Superior de Ingenieros de Telecomunicación*. Her research is in the areas of electromagnetic field theory, computational and numerical methods for microwave passive components and filter design, antenna analyis and design. A number of times she has been a visiting professor at the Electrical Engineering and Computer Science, Syracuse University (Syracuse, New York).

She has authored one book and a total of 15 contributions for chapters and articles in books, 25 papers in international journals, and 113 papers in international conferences, symposiums, and workshops. She is a member of the editorial board of three scientific journals. She is a registered engineer in Spain. She is a senior member of the Institute of Electrical and Electronics Engineers (IEEE). She has served as vice chairman and chairman of IEEE MTT-S/AP-S (Microwave Theory and Techniques Society/Antennas and Propagation Society) Spain joint chapter and chairman of IEEE Spain Section. She is a member of IEEE Region 8 Nominations and Appointments Committee, IEEE Ethics and Member Conduct Committee, and IEEE Women in Engineering Committee (WIEC). She is acting as liaison between IEEE Regional Activities Board and IEEE WIEC.

BIBLIOGRAPHY

1. T. K. Sarkar et al., A pragmatic approach to adaptive antennas, *IEEE Antennas Propag. Mag.* **42**(2): 39–55 (April 2000).
2. T. K. Sarkar, S. Park, J. Koh, and R. A. Schneible, A deterministic least squares approach to adaptive antennas, *Digital Signal Process. Rev. J.* **6**: 185–194 (1996).
3. S. Park and T. K. Sarkar, Prevention of signal cancellation in an adaptive nulling problem, *Digital Signal Process. Rev. J.* **8**: 95–102 (1998).
4. T. K. Sarkar, S. Nagaraja, and M. C. Wicks, A deterministic direct data domain approach to signal estimation utilizing nonuniform and uniform 2-D arrays, *Digital Signal Process. Rev. J.* **8**: 114–125 (1998).
5. T. K. Sarkar et al., A deterministic least squares approach to space time adaptive processing (STAP), *IEEE Trans. Antennas Propag.* **49**: 91–103 (Jan. 2001).
6. T. K. Sarkar and R. Adve, Space time adaptive processing using circular arrays, *IEEE Antennas Propag. Mag.* **43**: 138–143 (Feb. 2001).
7. R. S. Adve and T. K. Sarkar, Compensation for the effects of mutual coupling on direct data domain adaptive algorithms, *IEEE Trans. Antennas Propag.* **48**(1): (Jan. 2000).
8. I. J. Gupta and A. A. Ksienski, Effects of mutual coupling on the performance of adaptive arrays, *IEEE Trans. Antennas Propag.* **31**: 785–791 (Sept. 1983).
9. K. M. Pasala and E. M. Friel, Mutual coupling effects and their reduction in wideband direction of arrival estimation, *IEEE Trans. Aerospace Electron. Syst.* **30**: 1116–1122 (April 1994).
10. B. Friedlander, The root-MUSIC algorithm for direction finding with interpolated arrays, *Signal Process.* **30**: 15–29 (1993).
11. T.-S. Lee and T.-T. Lin, Adaptive beamforming with interpolation arrays for multiple coherent interferes, *Signal Process.* **57**: 177–194 (1997).
12. M. Wax and J. Sheinvald, Direction finding of coherent signals via spatial smoothing for uniform circular arrays, *IEEE Trans. Antennas Propag.* **42**(5): (May 1994).
13. B. M. Kolundzija, J. S. Ognjanovic, and T. K. Sarkar, *WIPL-D: Electromagnetic Modeling of Composite Metallic and Dielectric Structures*, Artech House, Norwood, MA, 2000.
14. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Press, Baltimore, 1989.
15. S. Van Huffel, *Analysis of the Total Least Squares Problem and Its Use in Parameter Estimations*, PhD thesis, Dept. Electrical Engg, Katholicke Univ. Leuven, 1990.
16. R. Brown and T. K. Sarkar, Real time deconvolution utilizing the fast Fourier transform and the Conjugate Gradient

method, 5th *ASSP Workshop on Spectral Estimation and Modeling*, Rochester, NY, 1980.

17. T. K. Sarkar, Application of the conjugate gradient method to electromagnetics and signal analysis, *Progress in Electromagnetics Research*, Vol. 5, Elsevier, 1991.

ADAPTIVE EQUALIZERS

KRZYSZTOF WESOŁOWSKI
Poznań University of Technology
Poznań, Poland

1. INTRODUCTION

Physical channels used in transmission of digital signals can be rarely represented by a nondistorting channel model with additive noise as the only impairment. However, the vast majority of channels are characterized by a limited bandwidth in which particular frequency components of transmitted signals are nonequally attenuated (causing *amplitude distortion*) and nonequally delayed (creating *delay distortion*). These effects are the results of the physical properties of the transmission medium and of the imperfect design of transmit and receive filters applied in the transmission system. A good example of the first is the radio channel, in which the transmitted signal reaches the receiver along many different paths through reflections, diffractions, and dispersion on the terrain obstacles. As a result, particular signal path components arriving with various attenuations and delays are combined at the receiver. The delayed components can be considered as echoes that cause time dispersion of the transmitted signal. If time dispersion is greater than a substantial fraction of the signaling period, the channel responses to the subsequent data signals overlap. This effect is known as *intersymbol interference*. Thus, the signal observed at the receiver input contains information on a certain number of data signals simultaneously. In many cases the channel impulse response spans even tens of signaling periods and intersymbol interference appears to be a major impairment introduced by the channel.

The destructive influence of intersymbol interference on a digital communication system performance has to be counteracted by special receiver and/or transmitter design. Thus, a fundamental part of the receiver is the channel *equalizer*. Very often transmission channel characteristics are either not known at the beginning of a data transmission session or they are time-variant. Therefore, it is advantageous to make the equalizer *adaptive*. The adaptive equalizer is able to adaptively compensate the distorting channel characteristics and simultaneously track the changes of channel characteristics in time. The latter property is a key feature of equalizers used in digital transmission over nonstationary radio channels.

Since the invention of an equalizer in the early 1960s, hundreds of papers have been devoted to this subject. Adaptive equalization is usually a topic of a separate chapter in leading books on digital communication systems [1–3], and separate books tackle this subject as well [5,6]. Adaptive equalization is also a well-documented application example in books devoted to adaptive filters [4,7]. In this tutorial we are able to present a general survey of adaptive equalizers only and give reference to the key papers and books. Interested readers are advised to study the wide literature, quoted, for example, in such papers as those by Qureshi [8] and Taylor et al. [9].

In this tutorial we will concentrate on basic structures and algorithms of adaptive equalizers. We will start with the model of a transmission system operating in the intersymbol interference channel. Subsequently, we will divide the equalizers into several classes. We will continue our considerations with the basic analysis of adaptation criteria and algorithms for linear and decision feedback equalizers. Then we will concentrate on adaptive algorithms and equalizer structures applying the MAP (*maximum a posteriori*) symbol-by-symbol detector and the MLSE (*maximum-likelihood sequence estimation*) detector. Finally, we will describe basic structures and algorithms of adaptive equalization without a training sequence (*blind equalization*).

2. SYSTEM MODEL

Generally, we can consider two types of data transmission. In the first one the channel transmits signal spectral components close to zero frequency and no modulation of the sinusoidal carrier is necessary. A good example is data transmission over a PSTN (public switched telephone network) subscriber loop, realizing the basic or primary ISDN access. Figure 1 shows a model of the data transmission system operating in the baseband channel. The binary sequence is transformed into the data symbol sequence in a symbol mapper. Data symbols are fed to the transmit filter $p(t)$ with the signaling period of T seconds. This filter shapes the transmitted signal spectrum. The data pulses are subsequently transmitted through the channel with the impulse response $g(t)$. The receive filter is usually a filter matched to the transmit filter, so its impulse response is $p(-t)$ (assuming that the additive white Gaussian noise $n(t)$ is added at the output of the channel). Replacing the cascade of the transmit filter, the transmission channel and the receive filter by a single equivalent

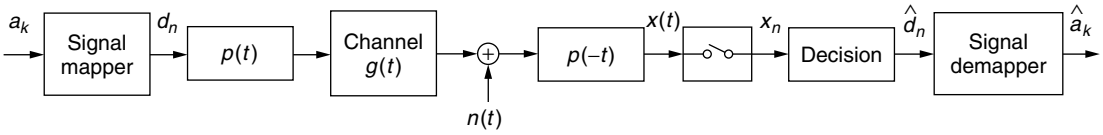


Figure 1. Model of the baseband transmission system.

filter, we receive the following equation describing the operation of the transmission system

$$x(t) = \sum_{i=-\infty}^{+\infty} d_i h(t - iT) + \nu(t) \tag{1}$$

where $h(t)$ is a convolution of the filter and channel impulse responses $[h(t) = p(t) * g(t) * p(-t)]$ and $\nu(t)$ is the noise $n(t)$ filtered by the receive filter $(\nu(t) = n(t) * p(-t))$.

In the second type of data transmission system a bandpass channel is used, so data signals modulate a sinusoidal carrier. Figure 2 presents a system model in which sinusoidal and cosinusoidal carriers of the frequency f_c are modulated by a pair of in-phase and quadrature data symbols d_n^I and d_n^Q, respectively. These symbols are received from the signal mapper, which associates the binary data blocks with a pair of data symbols resulting from the applied modulation format. As in the baseband transmission system, the spectrum of the transmitted signal is shaped in the baseband by the transmit filters $p(t)$. One can easily prove that the system model contained between lines A and B can be represented by the same equation as (1); however, in this case the variables and functions in Eq. (1) are complex: $d_i = d_i^I + jd_i^Q$, and $h(t) = h_{\text{re}}(t) + jh_{\text{im}}(t)$. In the passband transmission system model the channel impulse response $h(t)$ is a convolution of the transmit and receive filter responses $p(t)$ and $p(-t)$, respectively, and the so called *baseband equivalent channel* impulse response $g_B(t)$. The baseband equivalent channel impulse response is associated with the impulse response $g_P(t)$ of the bandpass channel (see Fig. 2) by the equation

$$g_P(t) = g_B(t) \exp[j2\pi f_c t] + g_B^*(t) \exp[-j2\pi f_c t] \tag{2}$$

Concluding, with respect to intersymbol interference, both baseband and passband transmission systems can be modeled by Eq. (1), in which variables and functions of time are real- or complex-valued depending on whether the system is implemented in the baseband or whether it uses the bandpass channel.

3. INTERSYMBOL INTERFERENCE

The task of the digital system receiver is to find the most probable data symbols d_n^I and d_n^Q on the basis of the observed signal $x(t)$ at its input. If the channel were nondistorting, then, assuming appropriate shaping of the transmit filter $p(t)$ and the filter $p(-t)$ matched to it, it would be possible to find such periodic sampling moments at the outputs of the receive filters that the samples of signals $x^I(t)$ and $x^Q(t)$ would contain information on single data symbols only. However, the distortion introduced by the channel renders the finding of such sampling moments impossible. Thus, it is necessary to apply a special system block denoted in Fig. 3 as equalizer, which is able to detect data symbols on the basis of $x(t)$ [or equivalently $x^I(t)$ and $x^Q(t)$] or its samples. An equalizer is in fact a kind of receiver that either minimizes the influence of intersymbol interference or uses it constructively in the decisions concerning the transmitted data.

Although the signals $x^I(t)$ and $x^Q(t)$ are represented in Fig. 3 as continuous time functions, typically, due to digital implementation, the equalizer accepts their samples only. Let us temporarily assume that the equalizer input samples are given with the symbol period of T seconds with the time offset τ with respect to the zero moment. Then the equalizer input signal is expressed by the equation

$$\begin{aligned} x_n = x(nT + \tau) &= \sum_{i=-\infty}^{\infty} d_i h(nT + \tau - iT) + \nu(nT + \tau) \\ &= \sum_{i=-\infty}^{\infty} d_i h_{n-i} + \nu_n = \sum_{i=-\infty}^{\infty} h_i d_{n-i} + \nu_n \end{aligned} \tag{3}$$

$$x_n = h_0 d_n + \sum_{i=-\infty, i\neq 0}^{\infty} h_i d_{n-i} + \nu_n \tag{4}$$

where $h_{n-i} = h(nT + \tau - iT)$. The first term in Eq. (4) is proportional to the data symbol to be decided on. The second term is a linear combination of previous and future data symbols and expresses intersymbol interference. It should be eliminated or constructively used by the

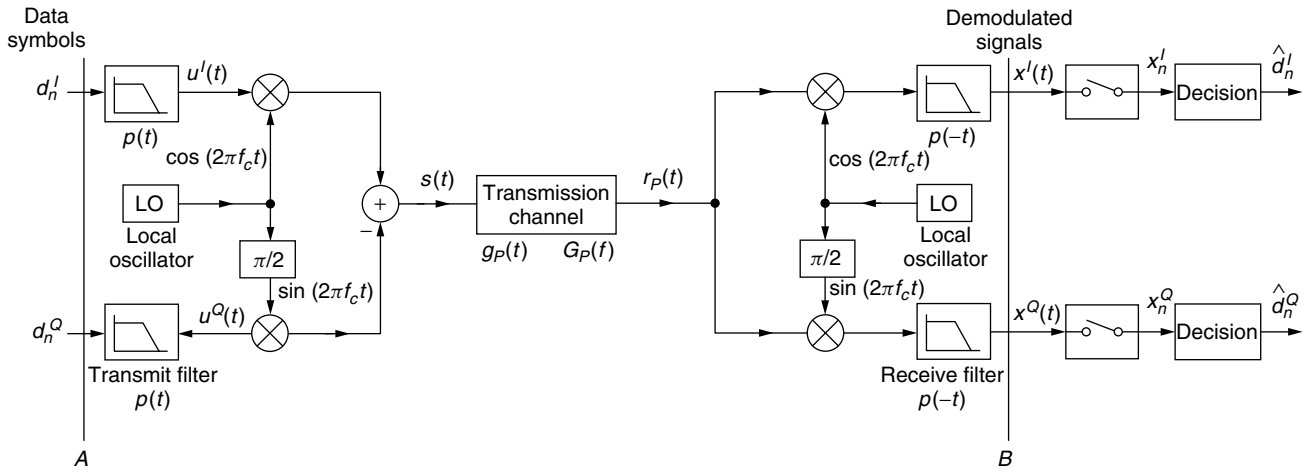


Figure 2. Model of the passband transmission system.

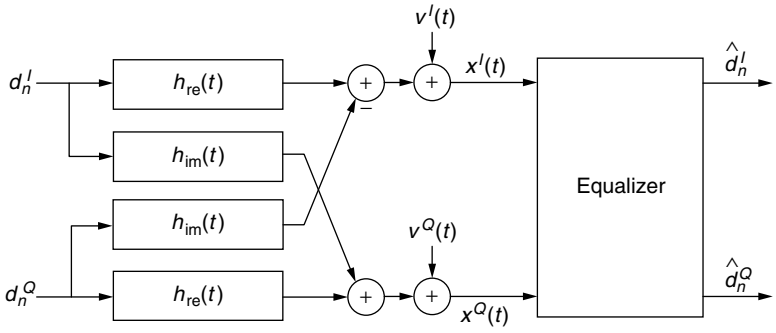


Figure 3. Equivalent transmission system model with the equalizer.

equalizer. The third term is the additive noise and cannot be eliminated.

4. CLASSIFICATION OF EQUALIZER STRUCTURES AND ALGORITHMS

Channel equalization can be performed by linear or nonlinear methods. Decisions regarding the data symbols can be made by the equalizer on the symbol-by-symbol basis or can be performed on the whole sequence. Figure 4 presents the classification of equalization structures.

Within the class of linear receivers the equalizer based on an FIR (Finite impulse response) *transversal filter* is of great importance. It is implemented using symbol-spaced or fractionally spaced taps. A lot of attention has also been paid in the literature to the linear equalizer applying a *lattice filter* [10]. The latter, although more complicated than the transversal filter, assures faster convergence of the adaptation algorithm.

In case of channels characterized by the occurrence of deep notches, nonlinear receivers are used. The simplest version of a nonlinear receiver is the *decision-feedback equalizer* (DFE) [11]. The MLSE equalizer, which is more computationally intensive is applied for example in GSM receivers and high-speed telephone modems. It detects a whole sequence of data symbols, usually using the *Viterbi algorithm* [12]. If the intersymbol interference is caused by a long-channel impulse response or if the data symbol alphabet is large, the MLSE equalizer becomes unfeasible due to excessive computational complexity. Several suboptimal structures and procedures can be applied instead, such as *reduced state sequence estimation* (RSSE) [13], *delayed decision feedback sequence estimation* (DDFSE) [14], or the *M* algorithm [15]. Another approach is a nonlinear symbol-by-symbol detection using the *maximum a posteriori* (MAP) criterion. The algorithm of Abend and Fritchman [22] is one example of such an approach. The MAP algorithms are usually computationally complex.

The key feature of all equalization structures is their ability of initial adaptation to the channel characteristics (*startup equalization*) and tracking it in time. In order to acquire the initial adaptation, an optimization criterion has to be defined. Historically, the first criterion was minimization of the maximum value of intersymbol interference (*minimax criterion*) resulting in the *zero-forcing* (ZF) equalizer. The most popular adaptation criterion is minimization of the *mean-squared error*, resulting in the *MSE* equalizer. In this criterion the expectation of the squared error signal at the equalizer output is minimized. Finally, the criterion used in the fastest adaptation algorithms relies on minimization of the *least squares* (LS)

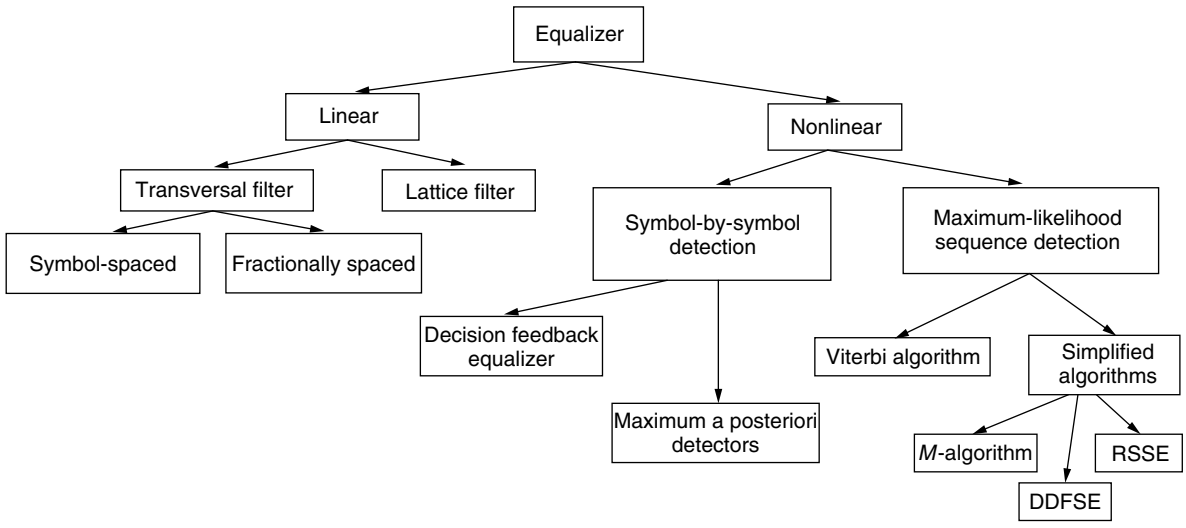


Figure 4. Classification of the equalization structures.

of errors. The equalizer using the algorithm based on this criterion is called an *LS* equalizer. The equalizer parameters are selected to minimize the squared sum of the equalizer output signal errors that would be achieved if these parameters were used starting from the initial moment of adaptation. Some other cost functions can be selected if the equalizer coefficients are derived without the knowledge of the transmitted data symbols.

The equalizer parameters are derived in accordance with a chosen adaptation criterion by an *adaptation algorithm*. Most of the algorithms are *recursive* — the adaptation at a given moment is performed iteratively, taking advantage of the results achieved in the previous adaptation step. In special cases *fast startup equalization* algorithms are applied, resulting in extremely fast calculation of coarse equalizer parameters that are good enough to start regular data transmission and that are later refined. Some of these algorithms are known as *noniterative*; some others are *recursive least-squares* (RLS) algorithms.

The adaptation of a typical equalizer can be divided into two phases. In the first phase, the training data sequence known to the receiver is transmitted. The adaptation algorithm uses this sequence as a reference for the adjustment of the equalizer coefficients; thus the equalizer is in the *training mode*. After achieving the equalizer parameters that result in a sufficiently low probability of errors made by the equalizer decision device, the second phase of adaptation begins in which the equalizer starts to use the derived decisions in its adaptation algorithm. We say that the equalizer is then in the *decision — directed* mode.

In some cases, in particular in point-to-multipoint transmission, sending the training sequence to initiate a particular receiver is not feasible. Thus, the equalizer must be able to adapt without a training sequence. Its algorithm is based exclusively on the general knowledge about the data signal statistics and on the signal reaching the receiver. Such an equalizer is called *blind*. Blind adaptation algorithms are generally either much slower or much more complex than data-trained algorithms.

5. LINEAR ADAPTIVE EQUALIZERS

The linear equalizer is the simplest structure, most frequently used in practice. Let us consider the receiver applying a transversal filter. The scheme of such an equalizer is shown in Fig. 5. The output signal y_n at the nth moment depends on the input signal samples x_{n-i} and the equalizer coefficients $c_{i,n}(i = -N, \ldots, N)$ according to the equation

$$y_n = \sum_{i=-N}^{N} c_{i,n} x_{n-i} \tag{5}$$

The equalizer output signal is a linear combination of $2N + 1$ subsequent samples of the input signal. Indexing the equalizer coefficients from $-N$ to N reflects the fact that the reference tap is located in the middle of the tapped delay line of the equalizer and that, typically, not only previous data symbols with respect to the reference one but also some future symbols influence the current input signal sample.

5.1. ZF Equalizers

Historically, the earliest equalizers used the *minimax* adaptation criterion. It resulted in the simplest algorithm that is still used in the equalizers applied in line-of-sight microwave radio receivers. Let us neglect the additive noise for a while. Taking into account Eqs. (4) and (5), we obtain

$$y_n = \sum_{i=-N}^{N} c_{i,n} \sum_{k=-\infty}^{\infty} h_k d_{n-i-k} \tag{6}$$

Substituting $j = i + k$, we get

$$y_n = \sum_{i=-N}^{N} c_{i,n} \sum_{j=-\infty}^{\infty} h_{j-i} d_{n-j} \tag{7}$$

or equivalently

$$y_n = \sum_{j=-\infty}^{\infty} g_{j,n} d_{n-j} \qquad \text{where} \qquad g_{j,n} = \sum_{i=-N}^{N} c_{i,n} h_{j-i} \tag{8}$$

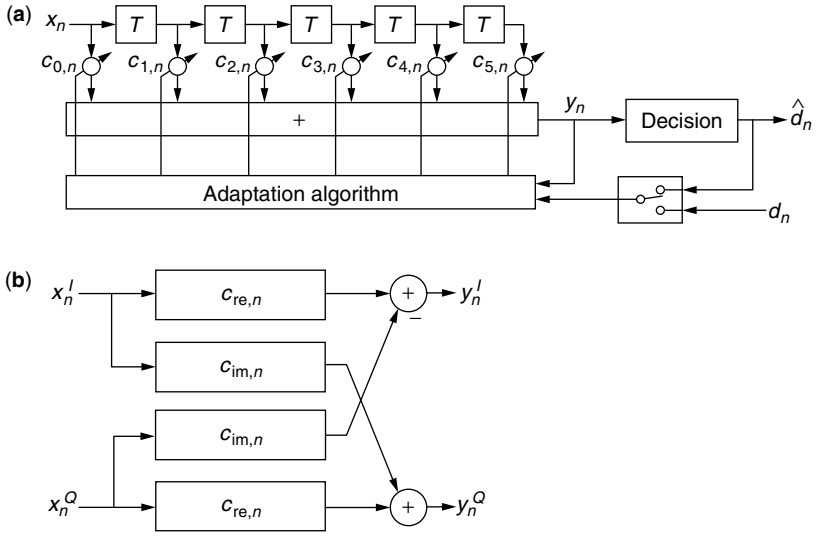


Figure 5. Linear adaptive equalizer: (**a**) basic structure; (**b**) structure equivalent to the complex filter applying real filters.

and $g_{j,n}$ are the samples of cascade connection of the discrete channel and the equalizer. In the *minimax* criterion the equalizer coefficients $c_{i,n} (i = -N, \ldots, N)$ are adjusted to minimize the expression

$$I = \frac{1}{g_{0,n}} \sum_{j=-\infty,\ j\neq 0}^{\infty} |g_{j,n}| \tag{9}$$

Let us note that, because of the finite number of the adjustable equalizer coefficients, it is possible to set to zero only part of the intersymbol interference samples observed at the output of the equalizer filter. One can show that in order to set the intersymbol interference samples to zero, at the assumption that the data symbols are uncorrelated and equiprobable, it suffices to set the equalizer coefficients to force to fulfil the following equality

$$E[e_n d_{n-i}] = 0 \qquad \text{for} \qquad i = -N, \ldots, N \tag{10}$$

where the error e_n in the training mode is given by the expression

$$e_n = y_n - d_n \tag{11}$$

or $e_n = y_n - \mathrm{dec}(y_n)$ in the decision-directed mode. In fact, substituting in (10) the expression for e_n and y_n, we obtain from (8)

$$E[e_n d_{n-i}] = E\left[\left(\sum_{j=-\infty}^{\infty} g_{j,n} d_{n-j} - d_n\right) d_{n-i}\right]$$
$$= \begin{cases} 0 & \text{for} \quad i = 0, \quad \text{if} \quad g_{0,n} = 1 \\ 0 & \text{for} \quad i \neq 0, i \in \langle -N, N\rangle, \quad \text{if} \quad g_{i,n} = 0 \end{cases} \tag{12}$$

Forcing condition (12), $2N$ intersymbol interference samples can be set to zero. Therefore, such an equalizer is called *zero-forcing equalizer*. If the equalizer was infinitely long it would be able to completely eliminate the ISI at its output. The cascade connection of the channel and equalizer would have the discrete impulse response in form of a unit pulse. Therefore, the equalizer would ideally inverse the channel frequency characteristics. Such an equalizer could be adjusted iteratively according to the equation

$$c_{i,n+1} = c_{i,n} - \alpha E[e_n d_{n-i}] \qquad \text{for} \qquad i = -N, \ldots, N \tag{13}$$

However, replacing the ensemble average with its stochastic estimate, we receive the following equation for the coefficients' adjustment, which is easily implementable even at a very high symbol rate

$$c_{i,n+1} = c_{i,n} - \alpha e_n d_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{14}$$

for real equalizers, and

$$c_{i,n+1} = c_{i,n} - \alpha e_n d^*_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{15}$$

for complex ones. More details on the ZF equalizer can be found in Ref. 16. The ZF equalizer attempting to inverse the channel characteristics amplifies the noise in these frequency regions in which the channel particularly attenuates the signal.

5.2. MSE Equalizers

As we have already mentioned, the most frequent adaptation criterion is minimization of the mean square error:

$$\min_{\{c_{i,n}, i=-N,\ldots,N\}} E[|e_n|^2] \tag{16}$$

where the error is given by equation (11). Direct calculations of the mean square error (MSE) $\mathcal{E}_n^{\mathrm{MSE}} = E[|e_n|^2]$ with respect to the equalizer coefficients $\mathbf{c}_n = [c_{-N,n}, \ldots, c_{0,n}, \ldots, c_{N,n}]^T$ lead to the following dependence of the MSE on the coefficients for the real equalizer

$$\mathcal{E}_n^{\mathrm{MSE}} = \mathbf{c}_n^T A \mathbf{c}_n - 2\mathbf{b}^T \mathbf{c}_n + E[|d_n|^2] \tag{17}$$

where $A = E[\mathbf{x}_n \mathbf{x}_n^T]$ ($\mathbf{x}_n = [x_{n+N}, \ldots, x_n, \ldots, x_{n-N}]^T$) is the input signal autocorrelation matrix and $\mathbf{b} = E[d_n \mathbf{x}_n]$ is the vector of cross-correlation between the current data symbol and the equalizer input samples. The autocorrelation matrix A is positive definite (its all eigenvalues are positive). It is well known from algebra that for such a matrix expression (17) has a single and global minimum. The minimum can be found if we set the condition

$$\frac{\partial \mathcal{E}_n^{\mathrm{MSE}}}{\partial c_{i,n}} = 0 \qquad \text{for} \qquad i = -N, \ldots, N \tag{18}$$

The result is the well-known Wiener–Hopf equation for the optimum equalizer coefficients

$$A\mathbf{c}_{\mathrm{opt}} = \mathbf{b} \tag{19}$$

An efficient method of achieving the optimum coefficients and the minimum MSE is to update the equalizer coefficients iteratively with adjustments proportional to the negative value of the gradient of $\mathcal{E}_n^{\mathrm{MSE}}$ calculated for the current values of the coefficients:

$$c_{i,n+1} = c_{i,n} - \alpha_n \frac{\partial \mathcal{E}_n^{\mathrm{MSE}}}{\partial c_{i,n}} \qquad \text{for} \qquad i = -N, \ldots, N \tag{20}$$

where α_n is a small positive value called the adjustment step size. Generally, it can be time variant, which is expressed by the time index n. The calculation of the gradient $\partial \mathcal{E}_n^{\mathrm{MSE}} / \partial c_{i,n}$ leads to the result

$$\frac{\partial \mathcal{E}_n^{\mathrm{MSE}}}{\partial c_{i,n}} = \frac{\partial E\left[|e_n|^2\right]}{\partial c_{i,n}} = 2E\left[e_n \frac{\partial e_n}{\partial c_{i,n}}\right] = 2E[e_n x_{n-i}]$$
$$\text{for} \qquad i = -N, \ldots, N \tag{21}$$

Replacing the gradient calculated in Eq. (21) by its stochastic estimate $e_n x_{n-i}$ $(i = -N, \ldots, N)$ for the real equalizer we receive the stochastic gradient [*least mean-square*(LMS)] algorithm

$$c_{i,n+1} = c_{i,n} - \gamma_n e_n x_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{22}$$

where $\gamma_n = 2\alpha_n$. One can show that the analogous equation for the complex equalizer is

$$c_{i,n+1} = c_{i,n} - \gamma_n e_n x^*_{n-i} \qquad \text{for} \qquad i = -N, \ldots, N \tag{23}$$

Figure 6 presents the scheme of the linear transversal equalizer with the tap coefficients adjusted according to algorithm (23). The switch changes its position from 1 to 2 after a sufficiently long training mode.

The convergence rate of the LMS algorithm depends on the value of the step size γ_n. This problem has been thoroughly researched. Generally, the value of the step size depends on the eigenvalue distribution of the input signal autocorrelation matrix A [1]. G. Ungerboeck [17] derived a simple "engineering" formula for the step size, which results in fast and stable convergence of the LMS adaptive equalizer. The initial step size is described by the formula

$$\gamma_0 = \frac{1}{(2N+1)E[|x_n|^2]} \tag{24}$$

where $E[|x_n|^2]$ is the mean input signal power and is equal to the elements of the main diagonal of the autocorrelation matrix A. When the equalizer taps are close to their optimum values, the step size should be decreased in order to prevent an excessively high level of the residual mean square error (e.g., $\gamma_\infty = 0.2\gamma_0$).

5.3. LS Equalizers

Particularly fast initial equalizer convergence is achieved if the *least-squares* adaptation criterion is applied. The coefficients of a linear equalizer are set in order to minimize the following cost function with respect to the filter coefficient vector $\mathbf{c}_n$:

$$\mathcal{E}_n^{LS} = \sum_{i=0}^{n} \lambda^{n-i} |\mathbf{c}_n^T \mathbf{x}_i - d_i|^2 \tag{25}$$

For each moment n, that weighted summed squared error starting from the initial moment up to the current moment n is minimized, which would be achieved if the current coefficient vector calculated on the basis of the whole signal knowledge up to the nth moment were applied in the equalizer from the initial moment. The window coefficient $\lambda^{n-i} (\lambda \leq 1)$ causes gradual forgetting of the past errors and is applied for nonstationary channels to follow the changes in the channel characteristics. The calculation of (25) leads to equations similar to (17) and (19):

$$\varepsilon_n^{LS} = \mathbf{c}_n^T R_n \mathbf{c}_n - 2\mathbf{c}_n^T \mathbf{q}_n + \sum_{i=0}^{n} \lambda^{n-i} |d_i|^2 \tag{26}$$

$$R_n \mathbf{c}_{n,\text{opt}} = \mathbf{q}_n \tag{27}$$

where

$$R_n = \sum_{i=0}^{n} \lambda^{n-i} \mathbf{x}_i^T \mathbf{x}_i = \lambda R_{n-1} + \mathbf{x}_n^T \mathbf{x}_n \qquad \text{and}$$

$$\mathbf{q}_n = \sum_{i=0}^{n} \lambda^{n-i} d_i \mathbf{x}_i \tag{28}$$

Instead of solving the set of linear equations (27) at each subsequent moment, the optimum coefficients can be found iteratively using the results derived at the previous time instant. Below we list only the equations of the standard RLS (Kalman) algorithm proposed by Godard [18] for fast adaptive equalization. The algorithm is quoted after Proakis [1].

For convenience, let us define $P_n = R_n^{-1}$. Let us also assume that before adaptation at the nth moment we have the filter coefficients $\mathbf{c}_{n-1}$ and the inverse matrix P_{n-1} at our disposal. The algorithm steps are as follows:

- Initialization: $\mathbf{c}_0 = [0, \ldots, 0]^T$, $\mathbf{x}_0 = [0, \ldots, 0]^T$, $R_0 = \delta I$.

Do the following for $n \geq 1$

- Shift the contents of the filter tapped delay line by one position and accept the new input signal x_n,
- Compute the filter output signal:

$$y_n = \mathbf{c}_{n-1}^T \mathbf{x}_n \tag{29}$$

- Compute the error at the filter output:

$$e_n = d_n - y_n \tag{30}$$

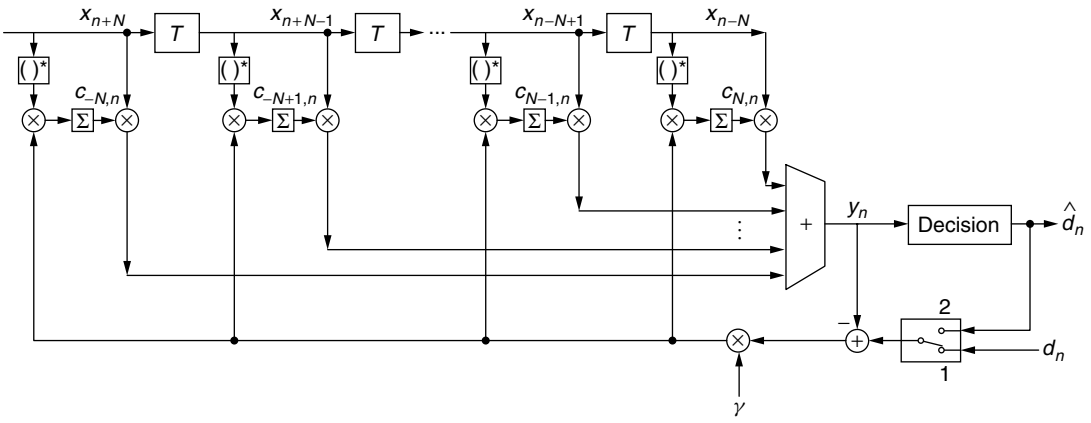


Figure 6. Adaptive MSE gradient equalizer.

- Compute the so called Kalman gain vector $\mathbf{k}_n = P_n\mathbf{x}_n$:

$$\mathbf{k}_n = \frac{P_{n-1}\mathbf{x}_n}{\lambda + \mathbf{x}_n^T P_n \mathbf{x}_n} \tag{31}$$

- Update the inverse of the autocorrelation matrix:

$$P_n = \frac{1}{\lambda}\left[P_{n-1} - \mathbf{k}_n \mathbf{x}_n^T P_{n-1}\right] \tag{32}$$

- Update the filter coefficients:

$$\mathbf{c}_n = \mathbf{c}_{n-1} + \mathbf{k}_n e_n \tag{33}$$

Formulas (29)–(33) summarize the RLS Kalman algorithm for the real equalizer. The complex version of this algorithm can be found in Proakis' handbook [1]. Knowing that $\mathbf{k}_n = P_n\mathbf{x}_n$, we find that the coefficients update is equivalent to the formula

$$\mathbf{c}_n = \mathbf{c}_{n-1} + R_n^{-1}\mathbf{x}_n e_n \tag{34}$$

Comparing the equalizer update using the LMS algorithm (23) and the RLS algorithm (34) we see that the Kalman algorithm speeds up its convergence because of the inverse matrix $P_n = R_n^{-1}$ used in each iteration. In the LMS algorithm this matrix is replaced by a single scalar γ_n. Figure 7 presents the convergence rate for both the LMS and RLS algorithms used in the linear transversal equalizer. The step size of the LMS algorithm was constant and selected to ensure the same residual mean-square error as that achieved by the RLS algorithm. The difference in the convergence rate is evident. However, we have to admit that for the channel model used in the simulations shown in Fig. 7, the application of the step size according to formula (24) and switching it to a small fraction of the initial value after the appropriate number of iterations improves the convergence of the LMS equalizer considerably. On the other hand, tracking abilities of the Kalman algorithm are much better than these of the LMS algorithm. However, the RLS Kalman algorithm is much more demanding computationally. Moreover, because of the roundoff noise, it becomes numerically unstable in the long run, particularly if the forgetting factor λ is lower than one. Solving the problem of excessive computational complexity and ensuring the numerical stability have been the subject of intensive research. Cioffi and Kailath's paper [19] is only one representative example of numerous publications in this area.

Besides the transversal filter, a lattice filter can also be applied in the adaptive equalizer using both the LMS [10] and RLS [20] adaptation algorithms.

5.4. Choice of the Reference Signal

The selection of the reference signal plays an important role in the equalizer adaptation process. In fact, the reference signal tests the unknown channel. Its spectral properties should be selected in such a way that the channel characteristics is fully reflected in the spectrum of the signal at the input of the equalizer. Thus far in our analysis we have assumed that the data symbols are uncorrelated and equiprobable:

$$E[d_n d_{n-k}^*] = \begin{cases} \sigma_d^2 & \text{for} \quad k = 0 \\ 0 & \text{for} \quad k \neq 0 \end{cases} \tag{35}$$

This means that the power spectrum of the test signal is flat and the channel characteristic is "sampled" by a constant power spectrum of the input signal. In practice this theoretical assumption is only approximately fulfilled. Typically, the data sequence is produced on the basis of the *maximum-length* sequence generator. The test generator is usually implemented by a scrambler contained in the transmitter that is based on a *linear feedback shift register* (LFSR). As a result, a *pseudonoise* (PN) binary sequence is generated. Typically, subsets of very long PN sequences are used as a training sequence.

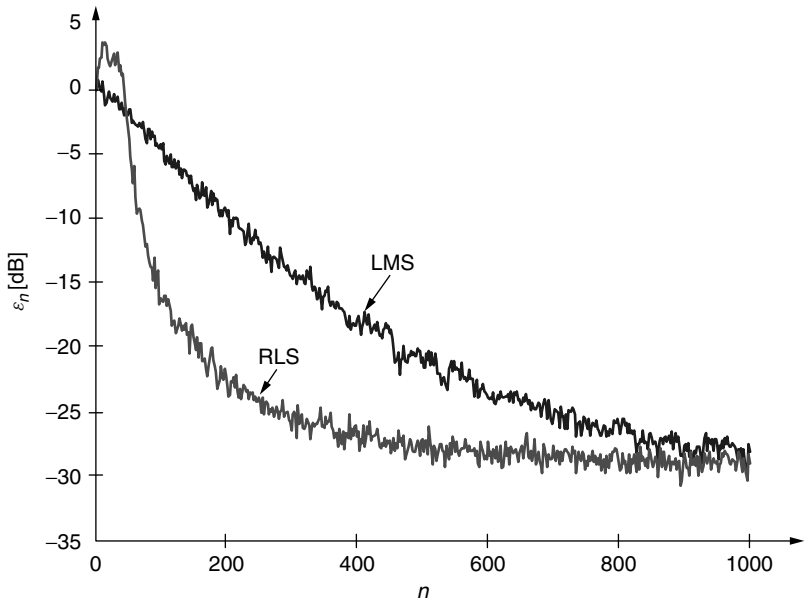


Figure 7. Convergence of the constant-step-size LMS and Kalman RLS equalizer of the length $2N = 30$.

Special attention has focused on very short test sequences that allow for fast, coarse setting of the equalizer coefficients. These sequences are periodic and are constructed in such a way that their deterministic autocorrelation function is zero apart from its origin.

5.5. Fast Linear Equalization Using Periodic Test Signals

In certain applications extremely fast initial equalization is of major importance. A good example is the master modem in a computer network, which receives data blocks from many tributary modems communicating through different channels. Such communication can be effective if the block header is a small part of the whole transmitted data block. A part of the header is a training sequence necessary to acquire the equalizer settings. Let us neglect the influence of the noise for a while. In many cases the SNR in the channel is around 30 dB and the ISI plays a dominant role in the signal distortion. Let the reference signal be periodic. In fact, no more than two periods of the test signal should be transmitted in order to acquire coarse equalizer settings. The period M of the test signal is at least as long as the highest expected length of the channel impulse response L. With periodic excitation the channel output (neglecting the influence of the additive noise) is also a periodic signal. This fact is reflected by the formula

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \end{bmatrix} = \begin{bmatrix} d_0 & d_1 & d_2 & \cdots & d_{M-1} \\ d_{M-1} & d_0 & d_1 & \cdots & d_{M-2} \\ \vdots & & \ddots & & \vdots \\ d_1 & d_2 & \cdots & d_{M-1} & d_0 \end{bmatrix} \cdot \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{M-1} \end{bmatrix} \tag{36}$$

If the length of the channel impulse response is shorter than the length of the test signal, we can assume that some of the last elements in the vector $\mathbf{h}^T = [h_0, h_1, \ldots, h_{M-1}]$ are equal to zero. Due to the periodic nature of the signal transmitted through the channel, a cyclic convolution of the sequence $\mathbf{h}$ and the data sequence $\mathbf{d} = [d_0, d_1, \ldots, d_{M-1}]$ is realized. In the frequency domain this operation is equivalent to the multiplication of two *discrete Fourier transform* (DFT) spectra:

$$X(k\Delta f) = D(k\Delta f) \cdot H(k\Delta f), k = 0, 1, \ldots, M-1 \tag{37}$$

where $\Delta f\ T = 1/M$ and

$$X(k\Delta f) = \frac{1}{M} \sum_{i=0}^{M-1} x(iT) \exp\left[-j2\pi k \Delta f i T\right] \tag{38}$$

Dependencies similar to (38) are held for the data and channel impulse response sequences. Knowing the spectrum of the data sequence, one can easily calculate the spectrum of the channel and, after reversing it, the characteristics of the ZF equalizer can be achieved, i.e.

$$C(k\Delta f) = \frac{1}{H(k\Delta f)} = \frac{D(k\Delta f)}{X(k\Delta f)} \quad k = 0, 1, \ldots, M-1 \tag{39}$$

On the basis of the equalizer characteristics $\mathbf{C}^T = [C(0), C(\Delta f), \ldots, C((M-1)\Delta f)]$, the equalizer coefficients $\mathbf{c}^T = [c_0, c_1, \ldots, c_{M-1}]$ can be calculated using the inverse DFT. If the length of the training sequence and of the equalizer is a power of 2 then all the DFT and IDFT calculations can be effectively performed by the FFT/IFFT algorithms. More detailed considerations on fast startup equalization using the periodic training sequence can be found [21].

5.6. Symbol-Spaced Versus Fractionally Spaced Equalizers

Thus far we have considered equalizers which accepted one sample per symbol period at their input. In fact the spectrum of the transmitted signal, although usually carefully shaped, exceeds half of the signaling frequency by 10–50%. Thus, the Nyquist theorem is not fulfilled, and as a result of sampling at the symbol rate, the input signal spectra overlap. In consequence, the symbol spaced equalizer is able to correct the overlapped spectrum only. In some disadvantageous cases the overlapping spectra can result in deep nulls in the sampled channel characteristic, which is the subject of equalization. In these spectral intervals the noise will be substantially amplified by the equalizer, which results in deterioration of the system performance.

Derivation of the optimum MSE receiver in the class of linear receivers results in the receiver structure consisting of a filter matched to the impulse observed at the receiver input and an infinite T-spaced transversal filter (see Ref. 3 for details). This derivation also shows that the characteristics $W_0(f)$ of the optimum MSE linear receiver are given by the formula

$$W_0(f) = \frac{\sigma_d^2}{\sigma_v^2} H^*(f) \left(\sum_{i=-\infty}^{\infty} c_i \exp\left[-j2\pi f i T\right] \right) \exp[-j2\pi f t_0] \tag{40}$$

where σ_d^2 is the data symbol mean power and σ_v^2 is the noise power. Lack of a matched filter preceding the transversal filter results in the suboptimality of the receiver and in performance deterioration. In practice, a sufficiently long but finite transversal filter is applied.

The question of whether an optimum receiver can be implemented more efficiently was answered by Macchi and Guidoux [23] as well as by Qureshi and Forney [24].

As we have mentioned, typically the spectrum of the received input signal is limited to the frequency $f_{\max} = (1/2T)(1+\alpha)$, where $\alpha \leq 1$. Let us assume that the noise is also limited to the same bandwidth because of the band-limiting filter applied in the receiver front end. Thus, the bandwidth of the optimal receiver is also limited to the same frequency $f_{\max}$. Because the input signal is spectrally limited to $f_{\max}$, the optimum linear receiver can be implemented by the transversal filter working at the input sampling frequency equal at least to $2f_{\max}$. Let the sampling period $T' = (KT/M)$ be selected to fulfill this condition: $(1/2T') \geq f_{\max}$. K and M are integers of possibly small values. As a result, the following equation holds:

$$H(f) \cdot W_0(f) = H(f) \cdot C_{\text{opt}}(f) \tag{41}$$

where

$$C_{\text{opt}}(f) = \sum_i W_0\left(f - i\frac{1}{T'}\right) \tag{42}$$

We must stress that although the input sampling frequency is $1/T'$, the data symbols are detected each T seconds, so the output of the equalizer is processed at the rate of $1/T$. It is important to note that the channel characteristics is first equalized by the T'-spaced filter and then its output spectrum is overlapped due to sampling the output at the symbol rate. Figure 8 illustrates these processes for $K = 1$ and $M = 2$, specifically, the equalizer is $T/2$-spaced. One can also show that the performance of the fractionally spaced equalizer is independent of the sampling phase [25].

Because the input signal spectrum is practically limited to $|f_{\max}|$, the equalizer can synthesize any characteristics in the frequency range

$$\left(-\frac{1}{2T'}, -f_{\max}\right) \cup \left(f_{\max}, \frac{1}{2T'}\right)$$

without any consequences for the system performance. Thus the optimum fractionally spaced equalizer can have many sets of the optimum coefficients. This phenomenon is disadvantageous from the implementation point of view because the values of the coefficients can slowly drift to unacceptable values. To stabilize the operation of the gradient algorithm, a *tap leakage algorithm* was introduced [26].

6. DECISION-FEEDBACK EQUALIZER

The decision-feedback equalizer (DFE) is the simplest nonlinear equalizer with a symbol-by-symbol detector. It was first described by Austin in 1967 [27]. The in-depth treatment of decision feedback equalization can be found in Ref. 28. It was found that intersymbol interference arising from past symbols can be canceled by synthesizing it using already detected data symbols and subtracting the received value from the sample entering the decision device. Figure 9 presents the basic scheme of the decision-feedback equalizer.

The equalizer input samples are fed to the linear (usually fractionally spaced) adaptive filter which performs matched filtering and shapes the ISI on its output in such a way that the symbol-spaced samples given to the decision device contain the ISI arising from the past symbols only. The ISI resulting from the joint channel and linear filter impulse response is synthesized in the transversal decision-feedback filter. The structure of the DFE is very

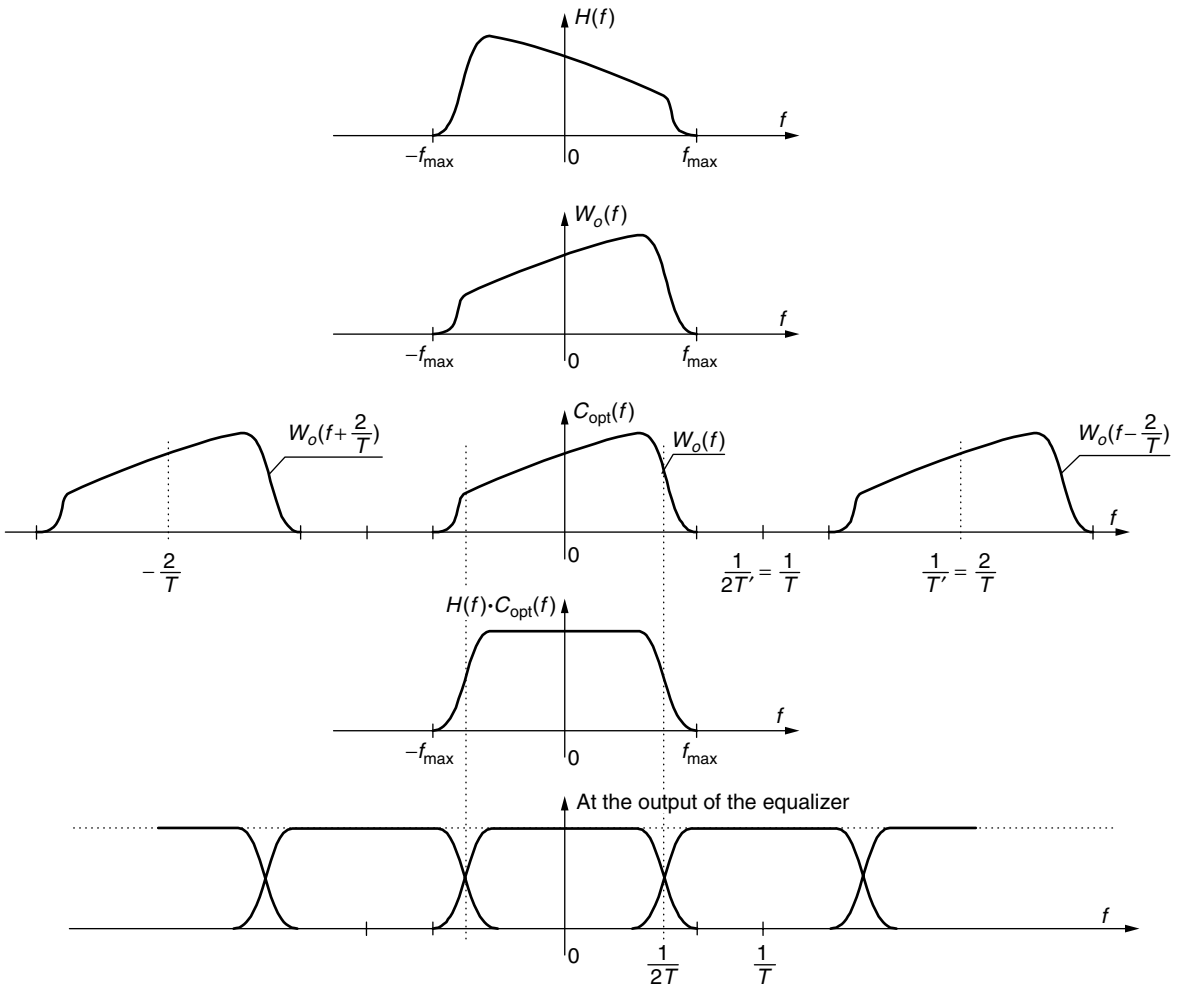


Figure 8. Equalization of the channel spectrum using $T/2$-spaced equalizer.

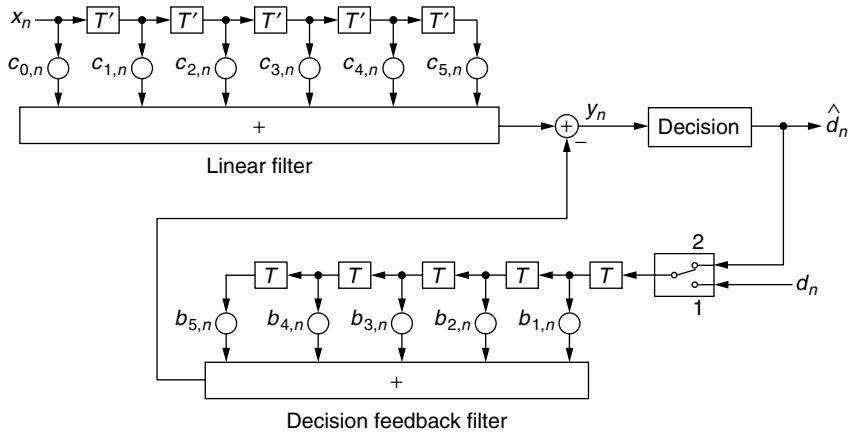


Figure 9. Structure of the decision-feedback equalizer.

similar to the infinite impulse response filter; however, the decision device is placed inside the filter loop, causing the whole structure to be nonlinear. Generally, the operation of the decision feedback equalizer is described by the equation

$$y_n = \sum_{k=-N_1}^{N_2} c_{k,n} x(nT - kT') - \sum_{j=1}^{N_3} b_{j,n} \hat{d}_{n-j} \tag{43}$$

where $c_{k,n}$ are the tap coefficients of the linear filter, $b_{j,n}$ are the tap coefficients of the decision-feedback filter and $\hat{d}_n$ is a data symbol estimate produced by the decision device. In the training mode the data estimates are replaced by the training data symbols.

The decision-feedback equalizer is applied in digital systems operating on channels with deep nulls [29]. Such channels cannot be effectively equalized by the linear equalizers attempting to synthesize the reverse channel characteristics. Instead, the DFE cancels a part of the ISI without inverting the channel and, as a result, the noise in the frequency regions in which nulls in channel characteristics occur is not amplified. Although the DFE structure is very simple and improves the system performance in comparison to that achieved for the linear equalizer, it has some drawbacks as well: (1) part of the signal energy is not used in the decision process because of its cancellation by the decision feedback filter; and (2) because of the decision feedback, errors made in the decision device take part in the synthesis of the ISI as they propagate along the decision feedback filter delay line. Thus, the errors contained in the tapped delay line increase the probability of occurrence of next errors. The phenomenon of error propagation effect can be observed if the signal to noise ratio is not sufficiently high. This effect is discussed, for example, by Lee and Messerschmitt [2].

The DFE tap coefficients can be adjusted according to the ZF or MSE criterion. As for the linear equalizer, the LMS and RLS adaptation algorithms can be used in the DFE. The DFE can be based on transversal or lattice filter structures [30]. Let us concentrate on the LMS algorithm only. We can combine the contents of the tapped delay lines of the linear and decision feedback filters as well as the filter coefficients into single vectors:

$$\mathbf{z}_n = \begin{bmatrix} \mathbf{x}_n \\ \cdots \\ \mathbf{d}_n \end{bmatrix} \quad \mathbf{w}_n = \begin{bmatrix} \mathbf{c}_n \\ \cdots \\ -\mathbf{b}_n \end{bmatrix} \tag{44}$$

where $\mathbf{x}_n = [x_{n+N_1}, \ldots, x_{n-N_2}]^T$, $\mathbf{d}_n = [d_{n-1}, \ldots, d_{n-N_3}]^T$, $\mathbf{c}_n = [c_{-N_1,n}, \ldots, c_{N_2,n}]^T$ and $\mathbf{b}_n = [b_{1,n}, \ldots, b_{N_3,n}]^T$. Then equation (43) can be rewritten in the form

$$y_n = \mathbf{z}_n^T \mathbf{w}_n \tag{45}$$

and the LMS gradient algorithm can be described by the recursive expression

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \beta_n e_n \mathbf{z}_n^* \tag{46}$$

where $e_n = y_n - d_n$. Knowing (44), we can break Eq. (46) into two separate LMS adjustment formulas for the feedforward and feedback filters.

Besides the regular DFE structure shown in Figure 9 there exists the so called predictive DFE [1,28], which, although featuring slightly lower performance, has some advantages in certain applications. Figure 10 presents the block diagram of this structure. The feedforward filter works as a regular linear equalizer according to the ZF or MSE criterion. Its adaptation algorithm is driven by the error signal between the filter output and the data decision (or training data symbol). As we remember, the linear equalizer more or less inverts the channel characteristics, which results in noise amplification. The noise contained in the feedforward filter output samples is correlated due to the filter characteristics. Therefore, its influence can be further minimized applying the linear predictor. Assuming that the decision device makes correct decisions, the noise samples contained in the feedforward filter output are the error samples used in the adaptation of this filter. The linear combination of the previous noise samples allows to predict the new sample, which is subsequently subtracted from the feedforward filter output. This way the effective SNR is increased. The result of subtraction constitutes the basis for decisionmaking.

Let us note (see Fig. 10) that the feedforward filter and the predictor are adjusted separately, so the performance

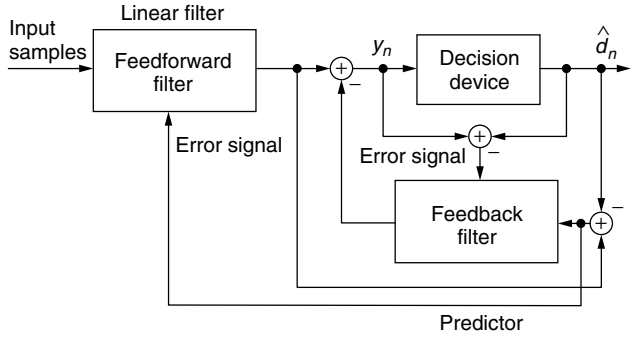


Figure 10. Predictive DFE.

of the predictive DFE is worse than the performance of the conventional DFE for which the taps adjustments are realized on the basis of the final output error. It has been shown that the predictive DFE is useful in realization of the joint trellis code decoder and channel equalizer [31].

7. EQUALIZERS USING MAP SYMBOL-BY-SYMBOL DETECTION

The decision-feedback equalizer is a particularly simple version of a nonlinear receiver in which the decision device is some kind of an M-level quantizer, where M is the number of data symbols. Much more sophisticated detectors have been developed which minimize the symbol error probability. This goal is achieved if the *maximum a posteriori probability* (MAP) criterion is applied. Let us consider the receiver structure shown in Fig. 11. The linear filter preceding the detection algorithm is a *whitened matched filter* (WMF). Its function is very similar to that of the linear filter applied in the decision feedback equalizer. It shapes the joint channel and linear filter impulse response to receive ISI arising from the past data symbols only. At the same time the noise samples at the output of the WMF are white. We say that the signal at the output of the whitened matched filter constitutes a *sufficient statistic* for detection. This roughly means that that part of the received signal that has been removed by the WMF is irrelevant for detection. Assuming that the number of interfering symbols is finite, we can write the following equation describing the sample y_n at the detector input:

$$y_n = \sum_{i=0}^{N} b_i d_{n-i} + \nu_n \tag{47}$$

where ν_n is a white Gaussian noise sample. Let us note that the information on the data symbol d_n is "hidden" in the samples $y_n, y_{n+1}, \ldots, y_{n+N}$. Generally, according to the MAP criterion the detector finds that $\hat{d}_n$ among

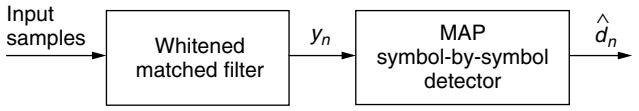


Figure 11. Basic scheme of the MAP symbol-by-symbol equalizer.

all possible M data symbols for which the following a posteriori probability is maximum:

$$\Pr\{d_n|\mathbf{y}_{n+N}\} \tag{48}$$

where $\mathbf{y}_{n+N} = (y_{n+N}, y_{n+N-1}, \ldots, y_1)$ is the vector of the observed input samples. From the Bayes theorem we know that for expression (48) the following equality holds:

$$\Pr\{d_n = m|\mathbf{y}_{n+N}\} = \frac{p(\mathbf{y}_{n+N}|d_n = m)\Pr\{d_n = m\}}{p(\mathbf{y}_{n+N})} \tag{49}$$

Because $p(\mathbf{y}_{n+N})$ is common for all possible probabilities (49), it has no meaning in the search for the data symbol featuring the MAP probability. Thus the task of the MAP detector can be formulated in the following manner

$$\hat{d}_n = \arg\left\{\max_{d_n} p(\mathbf{y}_{n+N}|d_n)\Pr\{d_n\}\right\} \tag{50}$$

Finding the data estimate (50) is usually computationally complex. Several algorithms have been proposed to realize (50). Abend and Fritchman [22] as well as Chang and Hancock [32] algorithms (the last being analogous to the well known BCJR algorithm [33] applied in convolutional code decoding) are good examples of these methods. We have to stress that all of them require the knowledge of the impulse response $\{b_i\}, (i = 1, \ldots, N)$ to calculate values of the appropriate conditional probability density functions. This problem will also appear in the MLSE receiver discussed in the next section.

8. MAXIMUM-LIKELIHOOD EQUALIZERS

Instead of minimizing the data symbol error, we could select minimization of error of the whole sequence as the optimization goal of the receiver. Thus, the MAP criterion yields the form

$$\max_{\mathbf{d}_n} P(\mathbf{d}_n|\mathbf{y}_n) \tag{51}$$

If the data sequences are equiprobable, our criterion is equivalent to the selection of such a data sequence that maximizes the conditional probability density function $p(\mathbf{x}_n|\mathbf{d}_n)$. Namely, we have

$$\begin{aligned}\hat{\mathbf{d}}_n &= \arg\left\{\max_{\mathbf{d}_n} P(\mathbf{d}_n|\mathbf{y}_n)\right\} = \arg\left\{\max_{\mathbf{d}_n} \frac{p(\mathbf{y}_n|\mathbf{d}_n)P(\mathbf{d}_n)}{p(\mathbf{y}_n)}\right\} \\ &= \arg\left\{\max_{\mathbf{d}_n} p(\mathbf{y}_n|\mathbf{d}_n)\right\}\end{aligned} \tag{52}$$

where, as before, $\mathbf{y}_n = (y_1, \ldots, y_n)^T$, $\mathbf{d}_n = (d_1, \ldots, d_n)^T$. Because noise at the WMF output is white and Gaussian, the conditional probability density function can be expressed by the formula

$$\begin{aligned}p(\mathbf{y}_n|\mathbf{d}_n) &= \prod_{i=1}^{n} p(y_i|\mathbf{d}_i) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \prod_{i=1}^{n} \exp\left[-\frac{\left|y_i - \sum_{k=0}^{N} b_k d_{i-k}\right|^2}{2\sigma^2}\right]\end{aligned} \tag{53}$$

Calculating the natural logarithm of both sides of (53), we obtain

$$\hat{\mathbf{d}}_n = \arg\left\{\max_{\mathbf{d}_n} \ln p(\mathbf{y}_n|\mathbf{d}_n)\right\} = \arg\left\{\min_{\mathbf{d}_n} \sum_{i=1}^{n} \left| y_i - \sum_{k=0}^{N} b_k d_{i-k} \right|^2 \right\} \tag{54}$$

Concluding, from all possible equiprobable data sequences $\mathbf{d}_n$ this sequence $\hat{\mathbf{d}}_n$ is selected for which the sum

$$S_n = \sum_{i=1}^{n} \left| y_i - \sum_{k=0}^{N} b_k d_{i-k} \right|^2 \tag{55}$$

is minimum. It was found by Forney [12] that the effective method of searching for such a sequence is the *Viterbi algorithm* See VITERBI ALGORITHM. Let us note that in order to select the data sequence the samples of the impulse response $\{b_k\}$, $k = 0, \ldots, N$ have to be estimated. They are usually derived on the basis of the channel impulse response $\{h_k\}$, $k = -N_1, \ldots, N_2$. The scheme of such a receiver is shown in Figure 12. The heart of the receiver is the Viterbi detector fed with the impulse response samples $\{b_k\}$ calculated on the basis of the channel impulse response samples $\{h_k\}$ estimated in the *channel estimator*. The channel estimator is usually an adaptive filter using the LMS or RLS algorithm for deriving the impulse response samples. From the system theory point of view it performs system identification. The channel estimator input signal is the data reference signal or the final or preliminary decision produced by the Viterbi detector. The channel output signal acts as a reference signal for the channel estimator. Usually, the reference signal has to be appropriately delayed in order to accommodate the decision delay introduced by the Viterbi detector.

For example, let us consider the channel estimator using the LMS algorithm and driven by ideal data symbols. Let us neglect the delay with which the data symbols are fed to the estimator. Assume that the data symbols are uncorrelated. Then, applying the mean-square error as the criterion for the estimator, we have

$$\mathcal{E}_n = E\left[|e_n|^2\right] = E\left[\left| x_n - \sum_{j=-N}^{N} \hat{h}_{j,n} d_{n-j} \right|^2\right] \tag{56}$$

where x_n is the channel output sample [see Eq. (4)] and $\hat{h}_{j,n}, j = -N, \ldots, N$ are the estimates of the channel impulse response at the nth moment. The calculation of the gradient of error $\mathcal{E}_n$ with respect to the channel impulse response estimate $\hat{h}_j$ gives

$$\frac{\partial \mathcal{E}_n}{\partial \hat{h}_{j,n}} = -2E[e_n d^*_{n-j}] \tag{57}$$

Therefore the stochastic gradient algorithm for the adjustment of channel impulse response estimates is

$$\hat{h}_{j,n+1} = \hat{h}_{j,n} - \alpha_n e_n d^*_{n-j} \quad j = -N, \ldots, N \tag{58}$$

where α_n is an appropriately selected step size. It can be shown that the initial step size should be $\alpha_0 = 1/(2N+1)E[|d_n|^2]$.

Another solution for deriving the channel impulse response is to use a zero-autocorrelation periodic training sequence. A fast channel estimator using such sequence is applied, for example, in the GSM receiver. Part of the known sequence placed in the middle of the data burst, called *midamble*, is a zero-autocorrelation periodic training sequence. In this case the channel impulse response samples are estimated on the basis of the following formula:

$$\hat{h}_i = \sum_{j=-N}^{N} x_j d^*_{i-j} \tag{59}$$

Thus, the received signal, which is the response of the channel to the periodic training signal, is cross-correlated

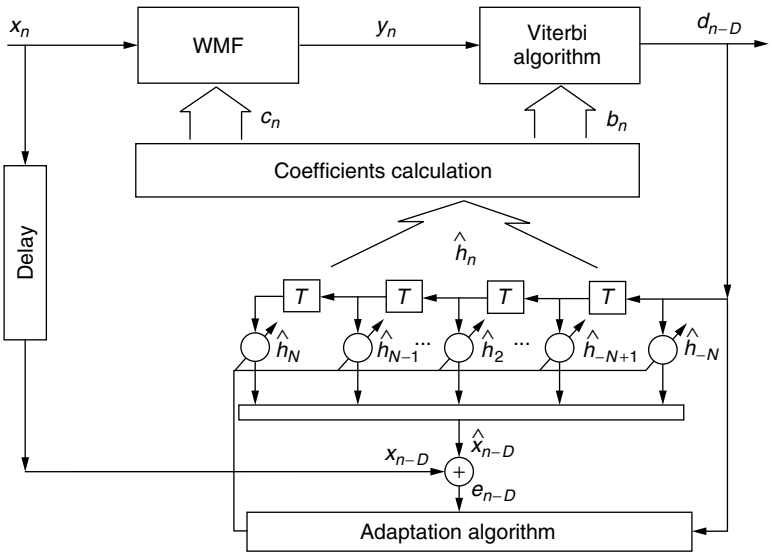


Figure 12. Basic scheme of the MLSE receiver with the whitened matched filter and the Viterbi algorithm.

with the complex conjugate of the training sequence. On the basis of the estimated impulse response samples $\hat{h}_i$ the receiver calculates the WMF coefficients and the weights $\{b_k\}$ used by the Viterbi detector.

An alternative equivalent structure of the MLSE equalizer was proposed by Ungerboeck [34]. Its derivation following Ungerboeck's considerations can be also found in Proakis' handbook [1]. Instead of the whitened matched filter, the filter matched to the channel impulse response is applied and the Viterbi detector maximizes the following cost function C_n with respect to the data sequence

$$C_n = \sum_{i=1}^{n} \mathrm{Re}\left[d_i^* \left(2y_i - g_0 d_i - 2\sum_{k=1}^{N} g_k d_{i-k} \right) \right] \qquad (60)$$

where y_i is the sample at the matched filter output at the ith moment and g_k $(k = 0, \ldots, N)$ are the samples of the channel impulse response autocorrelation function. See Ref. 34, 1, or 3 for details and for derivation of the algorithm.

Closer investigation of formula (54) allows us to conclude that in order to minimize the cost function and find the optimum data sequence, M^N operations (multiply and add, compare etc.) have to be performed for each timing instant. M is the size of the data alphabet. If modulation is binary $(M = 2)$ and the length of ISI is moderate, the detection algorithm is manageable. This is the case of the GSM receiver. However, if M is larger and/or the ISI corrupts a larger number of modulation periods, the number of calculations becomes excessive and suboptimal solutions have to be applied. Three papers [13,14,35] present examples of suboptimum MLSE receivers.

9. EQUALIZERS FOR TRELLIS-CODED MODULATIONS

In 1982 Ungerboeck published a paper [36] in which he proposed a joint approach to modulation and coding applied on band-limited channels. At the cost of expansion of the data signal constellation, application of a convolutional code and an appropriate binary data block-to-data symbol mapping, interdependence of subsequent data signals is obtained. Therefore, in order to select the maximum likelihood sequence among all possible sequences, whole data sequences have to be compared in the decision process. The distance between two closest data sequences is larger than between two uncoded data symbols and, in consequence, the system using *trellis-coded modulation* (TCM) is more robust against errors than the uncoded system transmitting data at the same data rate. The detection of the trellis-coded data stream requires sequential algorithm such as the Viterbi algorithm.

Using TCM signals on the ISI channels requires adaptive equalization and TCM decoding. The TCM detection process of the whole symbol sequences creates some problems in the selection of the equalizer structure and in the adjustment of the equalizer coefficients. The standard solution is to apply a linear equalizer minimizing the ISI followed by the TCM Viterbi decoder. The equalizer coefficient updates can be done using unreliable tentative decisions or the reliable but delayed decisions from the

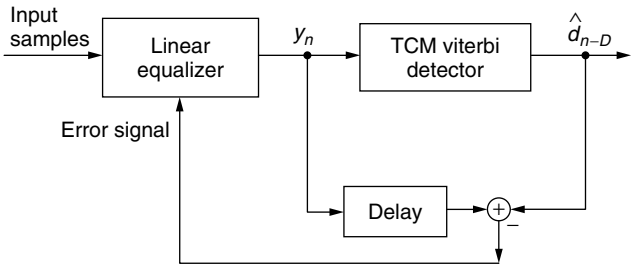


Figure 13. Linear equalizer with trellis-coded modulation decoder.

TCM Viterbi decoder [37]. In case of the LMS algorithm applied in the equalizer, the consequence of the delayed error signal (see Fig. 13) is the necessity of decreasing the step size [37].

On some channels, in particular those featuring a long tail in the channel impulse response or possessing deep nulls in their characteristics, applying a decision-feedback equalizer is more advantageous. Using joint DFE and trellis coding requires some special solutions because the DFE uses symbol-by-symbol decisions in its feedback filter with a single delay. One solution to this problem is to apply an interleaver between the TCM encoder and the modulator at the transmitter and the predictive DFE with the deinterleaver between the linear part of the equalizer and the decision-feedback part incorporating the TCM Viterbi decoder and predictor [1]. Another solution which is applicable in systems with a feedback channel and operating on transmission channels that are stationary or slowly change in time, is to share the DFE equalization between transmitter and receiver. In this case the concept of Tomlinson precoding applied jointly with the TCM coding is very useful [38].

The optimum receiver for TCM signals received in the presence of the ISI has been shown [31]. Its structure is basically the same as that shown in Fig. 12; however, now the Viterbi detector operates on the supertrellis resulting from concatenation of the ISI and TCM code trellises. See VITERBI ALGORITHM. Because the number of supertrellis states and the computational complexity associated with them are very high, suboptimum solutions have to be applied. The most efficient one is to incorporate intersymbol interference into the decision feedback for each supertrellis state, using the data sequences that constitute the "oldest" part of the maximum-likelihood data sequence associated with each state (so called *survivor*). In fact, this idea is already known from the delayed decision-feedback sequence estimation [14] used for uncoded data.

10. BLIND ADAPTIVE EQUALIZATION

As we have already mentioned, in some cases sending a known data sequence to train the equalizer can result in wasting of a considerable part of transmission time. One of the cases where adaptive equalization without a training sequence is applied is the transmission of *digital video broadcasting* (DVB) datastream in a DVB cable distribution system. A DVB cable receiver, after being

switched on, has to compensate intersymbol interference on the basis of the received signal and the general knowledge of its properties.

Blind equalization algorithms can be divided into three groups:

- The Bussgang [39] algorithms, which apply the gradient-type procedure with nonlinear processing of the filter output signal in order to obtain a reference signal conforming to the selected criterion,
- Second- and higher-order spectra algorithms, which apply higher-order statistics of the input signals in order to recover the channel impulse response and subsequently calculate the equalizer coefficients,
- Probabilistic algorithms, which realize the ML or MAP sequence estimation or suboptimum methods.

The algorithms belonging to the first category are easiest to implement. They will be described below.

The theory of blind equalization presented by Benveniste et al. [40] shows that in order to adjust the linear equalizer properly, one should drive its coefficients in such a way that the instantaneous probability distribution of the equalizer output y_n converges to the data input signal probability distribution $p_D(y)$. However, one important condition has to be fulfilled — the probability density function of the input signal d_n must be different from the Gaussian one. It has been found that the ISI introduced by the channel distorts the shape of the input probability density function unless it is Gaussian.

The main difficulty in designing the equalizer's adaptation algorithm is finding a criterion which, when minimized with respect to equalizer's coefficients, results in (almost) perfect channel equalization. One approach is to calculate the error

$$e_n = y_n - g(y_n) \tag{61}$$

which is to be minimized in the MSE sense, where $g(y_n)$ is an "artificially" generated "reference signal" and $g(.)$ is the memoryless nonlinearity. Thus, the general criterion that is the subject of minimization with respect to the coefficient vector $\mathbf{c}_n$ is

$$\mathcal{E}_n = E[|e_n|^2] = E[|y_n - g(y_n)|^2] \tag{62}$$

A typical approach to finding the minimum of $\mathcal{E}_n$ is to change the equalizer's coefficients in the direction opposite that indicated by the current gradient of ε_n, calculated with respect to $\mathbf{c}_n$. If we assume that all the signals and filters are complex, we get the following "reference" and error signals:

$$\tilde{y}_n = g(\mathrm{Re}(y_n)) + jg(\mathrm{Im}(y_n)) \quad \tilde{e}_n = y_n - \tilde{y}_n \tag{63}$$

Calculation of the gradient of $\mathcal{E}_n$ leads to the result

$$\operatorname*{grad}_{\mathbf{c}_n} \mathcal{E}_n = 2E\left\{\left[\mathrm{Re}(\tilde{e}_n)(1 - g'(\mathrm{Re}(y_n))) + j\,\mathrm{Im}(\tilde{e}_n)(1 - g'(\mathrm{Im}(y_n)))\right]\mathbf{x}_n^*\right\} \tag{64}$$

In practice the derivative $g'(.)$ is equal to zero except for a few discrete values of its argument. Thus, the stochastic version of the gradient algorithm achieves the well-known form

$$\mathbf{c}_{n+1} = \mathbf{c}_n - \alpha\tilde{e}_n\mathbf{x}_n^* \tag{65}$$

where this time the error signal $\tilde{e}_n$ is described by Eq. (63). Unfortunately, the optimum nonlinear function $g(.)$ is difficult to calculate. Bellini [39] investigated this function with several simplifying assumptions. Generally, function $g(.)$ should vary during the equalization process. Most of the gradient-based adaptation algorithms are in fact examples of the Bussgang technique, although they were found independently of it. Below we list the most important versions of the gradient algorithms, quoting the error signals that are characteristic for them.

- *Sato algorithm*: $\tilde{e}_n = e_n^S = y_n - A_S\mathrm{csgn}(y_n)$, where $\mathrm{csgn}(y_n) = \mathrm{sgn}(\mathrm{Re}(y_n)) + j\mathrm{sgn}(\mathrm{Im}(y_n))$, A_S is the weighting center of the in-phase and quadrature data signal components,
- *Benveniste–Goursat algorithm*: $\tilde{e}_n = e_n^B = k_1 e_n + k_2 |e_n| e_n^S$, $e_n = y_n - \mathrm{dec}(y_n)$, k_1, k_2 are properly selected weighting coefficients,
- *Stop-and-go algorithm*: $\tilde{e}_n = e_n^{SG} = f_n^R\mathrm{Re}(e_n) + jf_n^I\mathrm{Im}(e_n)$, $e_n = y_n - \mathrm{dec}(y_n)$, the weighting factors f_n^R, f_n^I turn on and off the in-phase and quadrature components of the decision error depending on the probability of the event that these components indicate the appropriate direction of the coefficients' adjustment,
- *Constant-modulus (CM) algorithm*: $\tilde{e}_n = e_n^G = (|y_n|^2 - R_2)y_n$, where R_2 is a properly selected data constellation radius.

The CM algorithm, although the most popular among the four described above, loses information about the phase of the received signal. Therefore, it has to be supported by the phase-locked loop in order to compensate for the phase ambiguity.

The second group of blind algorithms applied in channel estimation or equalization are the algorithms using the methods of the higher-order statistics of the analyzed signal.

A survey of the higher-order statistics applied in adaptive filtering can be found in Haykin's book [4]. Let us concentrate on the second-order statistics methods, showing an example of such an algorithm [41]. If the signal on the channel output

$$x(t) = \sum_{k=0}^{L-1} d_k h(t - kT) + n(t) \tag{66}$$

is sampled once per data symbol period, it is not possible to identify the samples of the channel impulse response based on the autocorrelation function of these samples. However, it is possible to do this if the signal is oversampled or received from the antenna arrays, that is, if more samples per data symbol are processed by the algorithm. If the

signal is sampled m times in each data period T, then it can be expressed in the vector form as

$$\mathbf{x}_n = \begin{bmatrix} x_{1,n} \\ \cdots \\ x_{m,n} \end{bmatrix} = \sum_{k=0}^{L-1} \begin{bmatrix} h_{1,k} \\ \cdots \\ h_{m,k} \end{bmatrix} d_{n-k} + \begin{bmatrix} \nu_{1,n} \\ \cdots \\ \nu_{m,k} \end{bmatrix} = \sum_{k=0}^{L-1} \mathbf{h}_i d_{n-i} + \boldsymbol{\nu}_n \tag{67}$$

On the basis of the signal vectors $\mathbf{x}_n$, we can estimate the matrices

$$C_x(i) = \frac{1}{N-i} \sum_{k=i+1}^{N} \mathbf{x}_k \mathbf{x}_{k-i}^* \; i = -L+1, \ldots, L-1 \tag{68}$$

and calculate the power density spectrum estimate in the matrix form

$$Q(e^{j\omega}) = \sum_{i=-L+1}^{L-1} C_x(i) e^{j\omega i} \tag{69}$$

By eigendecomposition of Q we can obtain the principal eigenvector, which is also described by the equation

$$\mathbf{c}(\omega) = e^{j\alpha(\omega)} \sum_{k=0}^{L-1} \mathbf{h}_k e^{j\omega k} \tag{70}$$

In order to calculate the channel impulse response samples, we take $N > 2L+1$ uniform samples of $\mathbf{c}(\omega_i)$ and form an appropriate system of linear equations. The solution of the system is the set of samples of the channel impulse response and weights $a_k = \exp(j\alpha(\omega_i))$. The estimated channel impulse response can be subsequently applied in the sequential algorithm or serve as the basis for the calculation of the equalizer coefficients. The simulation results reported by Xu et al. [41] show that satisfactory results can be achieved already with $N = 50$ sample blocks. A low number of input signal samples necessary for reliable channel estimation is the main advantage of the methods using second-order statistics as compared with the Bussgang techniques using the gradient algorithm. The price paid for fast channel estimation and equalization is high computational complexity of the algorithms.

The algorithms applying higher than second-order statistics generally use the cumulants of the input signal samples and their Fourier transforms called polyspectra. Polyspectra provide the basis for the nonminimum phase channel identification, thanks to their ability to preserve phase information of the channel output signal. The main drawback of the higher-order statistical methods is an extensive number of signal samples necessary to estimate the cumulants with sufficient accuracy and high computational complexity of the algorithms.

The third group of blind equalization algorithms relies on the joint channel estimation and data detection. Usually the maximum-likelihood criterion is applied, which in the blind case is expressed in the form

$$\arg p(\mathbf{x}_n | \mathbf{h}, \mathbf{d}_n) = \arg_{\mathbf{h},\mathbf{d}_n} \frac{1}{(2\pi\sigma^2)^n} \times \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left| x_i - \sum_{k=0}^{L-1} h_k d_{i-k} \right|^2 \right] \tag{71}$$

The channel identification can be performed after reception of the signal sequence $\mathbf{x}_n = (x_1, \ldots, x_n)$ by averaging the probability density function over all possible data sequences of length n. Subsequently, the Viterbi algorithm can be performed that finds the best data estimate in the ML sense. There are several other versions of joint channel data sequence estimation (see Ref. 42 as a representative example).

11. CONCLUSIONS

In this tutorial we have concentrated on the problem of equalization for point-to-point transmission. Limited space did not allow us to describe many other important issues such as *multiple-input/multiple-output* (MIMO) equalizers [43], the principle of per survivor processing [44], or equalization and MLSE detection performed jointly with diversity reception in mobile radio channels [e.g., 45]. Other interesting subjects are adaptive equalization in the frequency domain and adaptive channel equalization in the OFDM (orthogonal frequency-division multiplexing) systems.

BIOGRAPHY

Krzysztof Wesołowski was born in 1952. He received a M.Sc. degree in electrical engineering from Poznań University of Technology, Poznań, Poland, in 1976, a M.A. in Mathematics (*cum laude*) from Adam Mickiewicz University, Poznań, Poland, in 1978, a Ph.D. in 1982, and a Dr *Habilitus* degrees in 1989 in telecommunications. Currently, he holds a position of the professor of electrical engineering at the same university and leads the research group of wireless communications. He has published over 90 papers in Polish, English, and German. He is the author of the book *Mobile Communication Systems* published in Polish (1998, 1999). Its updated translation has been published in 2002 by John Wiley & Sons, Ltd. He spent his sabbatical leaves at Northeastern University, Boston, (Postdoctoral Fulbright Scholarship) and at the University of Kaiserslautern, Germany, (Alexander von Humboldt Scholarship). At the latter university he also served as a visiting professor teaching courses on adaptive equalization, information theory, and coding.

His main interests concentrate on the physical layer of digital communication systems, in particular on adaptive equalization, signal detection, error control coding, and other transmit and receive techniques applied to wireless communications.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1996.
2. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, 2nd ed., Kluwer, Boston, 1995.
3. R. D. Gitlin, J. F. Hayes, and S. B. Weinstein, *Principles of Data Communications*, Plenum Press, New York, 1992.
4. S. Haykin, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, Englewood-Cliffs, NJ, 1991.

5. A. P. Clark, *Equalizers for Digital Modems*, Pentech Press, London, 1985.
6. Zh. Ding and Ye Li, *Blind Equalization and Identification*, Marcel Dekker, New York, 2001.
7. O. Macchi, *Adaptive Processing*, Wiley, Chichester, UK, 1995.
8. S. U. H. Qureshi, Adaptive equalization, *Proc. IEEE* **53**: 1349–1387 (1985).
9. D. P. Taylor, G. M. Vitetta, B. D. Hart, and A. Mämmelä, Wireless channel equalisation, *Eur. Trans. Telecommun.* **9**: 117–143 (1998).
10. E. H. Satorius and S. T. Alexander, Channel equalization using adaptive lattice algorithms, *IEEE Trans. Commun.* **COM-27**: 899–905 (1979).
11. P. Monsen, Feedback equalization for fading dispersive channels, *IEEE Trans. Inform. Theory* **IR-17**: 56–64 (1971).
12. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363–378 (1972).
13. M. V. Eyuboglu and S. U. H. Qureshi, Reduced-state sequence estimation with set partitioning and decision feedback, *IEEE Trans. Commun.* **36**: 13–20 (1988).
14. A. Duel-Hallen and C. Heegard, Delayed decision-feedback sequence estimation, *IEEE Trans. Commun.* **37**: 428–436 (1989).
15. J. B. Anderson and S. Mohan, Sequential decoding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**: 169–176 (1984).
16. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*, McGraw-Hill, New York, 1968.
17. G. Ungerboeck, Theory on the speed of convergence in adaptive equalizers for digital communication, *IBM J. Res. Devel.* **16**: 546–555 (1972).
18. D. N. Godard, Channel equalization using a Kalman filter for fast data transmission, *IBM J. Res. Devel.* **18**: 267–273 (1974).
19. J. M. Cioffi and T. Kailath, Fast recursive least-squares transversal filter for adaptive filtering, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-32**: 304–337 (1984).
20. E. H. Satorius and J. D. Pack, Application of least squares lattice algorithms to adaptive equalization, *IEEE Trans. Commun.* **COM-29**: 136–142 (1981).
21. P. R. Chevillat, D. Maiwald, and G. Ungerboeck, Rapid training of a voiceband data-modem receiver employing an equalizer with fractional-T spaced coefficients, *IEEE Trans. Commun.* **COM-35**: 869–876 (1987).
22. K. Abend and B. D. Fritchman, Statistical detection for communication channels with intersymbol interference, *Proc. IEEE* 779–785 (1970).
23. O. Macchi and L. Guidoux, A new equalizer and double sampling equalizer, *Ann. Telecommun.* **30**: 331–338 (1975).
24. S. U. H. Qureshi and G. D. Forney, Jr., Performance and properties of a T/2 equalizer, *Conf. Record, National Telecommunication Conf.*, 1977.
25. G. Ungerboeck, Fractional tap-spacing equalizer and consequence for clock recovery in data modems, *IEEE Trans. Commun.* **COM-24**: 856–864 (1976).
26. R. D. Gitlin, H. C. Meadors, and S. B. Weinstein, The tap leakage algorithm: An algorithm for the stable operation of a digitally implemented, fractionally spaced adaptive equalizer, *Bell Syst. Tech. J.* **61**: 1817–1839 (1982).
27. M. E. Austin, *Equalization of Dispersive Channels Using Decision Feedback*, Research Laboratory of Electronics, MIT, Cambridge, MA., QPR 84, 1967, pp. 227–243.
28. C. A. Belfiore and J. H. Park, Jr., Decision feedback equalization, *Proc. IEEE* **67**: 1143–1156 (1979).
29. P. Monsen, Feedback equalization for fading dispersive channels, *IEEE Trans. Inform. Theory* **IT-17**: 56–64 (1971).
30. F. Ling and J. G. Proakis, Adaptive lattice decision-feedback equalizers — their performance and application to time-variant multipath channels, *IEEE Trans. Commun.* **COM-33**: 348–356 (1985).
31. P. R. Chevillat and E. Eleftheriou, Decoding of trellis-encoded signals in the presence of intersymbol interference and noise, *IEEE Trans. Commun.* **37**: 669–676 (1989).
32. R. W. Chang and J. C. Hancock, On receiver structures for channel having memory, *IEEE Trans. Inform. Theory* **IT-12**: 463–468 (1966).
33. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (1974).
34. G. Ungerboeck, Adaptive maximum-likelihood receiver for carrier-modulated data transmission systems, *IEEE Trans. Commun.* **COM-22**: 624–636 (1974).
35. K. Wesolowski, An efficient DFE & ML suboptimum receiver for data transmission over dispersive channels using two-dimensional signal constellations, *IEEE Trans. Commun.* **COM-35**: 336–339 (1987).
36. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**: 55–67 (1982).
37. G. Long, F. Ling, and J. G. Proakis, The LMS algorithm with delayed coefficient adaptation, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-37**: (1989).
38. A. K. Aman, R. L. Cupo, and N. A. Zervos, Combined trellis coding and DFE through Tomlinson precoding, *IEEE J. Select. Areas Commun.* **9**: 876–883 (1991).
39. S. Bellini, Bussgang techniques for blind equalization, *Proc. GLOBECOM'88*, 1988, pp. 1634–1640.
40. A. Benveniste, M. Goursat, and G. Ruget, Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications, *IEEE Trans. Autom. Control* **AC-25**: 385–398 (1980).
41. G. Xu, L. Tong, and H. Liu, A new algorithm for fast blind equalization of wireless communication channels, *Proc. GLOBECOM'94*, 1994, pp. 544–548.
42. N. Seshadri, Joint data and channel estimation using blind trellis search techniques, *IEEE Trans. Commun.* **42**: 1000–1011 (1994).
43. A. Duel-Hallen, Equalizers for multiple input/multiple output channels and PAM Systems with cyclostationary input sequences, *IEEE J. Select. Areas Commun.* **10**: 630–639 (1992).
44. R. Raheli, A. Polydoros, and Ch.-K. Tzou, The principle of per-survivor processing: A general approach to approximate and adaptive MLSE, *Proc. GLOBECOM'91*, 1991, pp. 1170–1175.
45. R. Krenz and K. Wesołowski, Comparative study of space-diversity techniques for MLSE receivers in mobile radio, *IEEE Trans. Vehic. Technol.* **46**: 653–663 (1997).

ADAPTIVE RECEIVERS FOR SPREAD-SPECTRUM SYSTEMS

URBASHI MITRA
Communication Sciences Institute
Los Angeles, California

1. INTRODUCTION

The explosive growth of wireless communications has motivated the "re"consideration of spread-spectrum techniques for multiuser communications. As the phrase suggests, "multiuser" communication systems offer communication services to multiple users simultaneously. Our focus is on a system as depicted by Fig. 1. In such a system, multiple users share a communications channel. The term "channel" is both abstract and physical; it describes the link between the transmitter(s) and the receiver(s). Thus, it could refer to free space or even a body of water. For free space, the channel is typically defined by a band of frequencies and characterized by the physical topology between the transmitter(s) and the receiver(s). This multiuser system can also be termed a *multipoint-to-point* communications system in contrast with a *point-to-multipoint* or *broadcast* system as employed for broadcast radio transmission and broadcast television. In broadcast channels, a single information stream is transmitted from a centralized transmitter to be received by multiple receivers. The objective of the receiver in our multipoint-to-point or *multiple-access* system is to ultimately demodulate the information stream of one, some, or all of the active users in the system. The receiver is thus the recipient of the different information signals of multiple users. Examples of multiuser communication systems include cellular communications, local-area networks, computer communications networks (such as the Internet), telephone networks, and packet-radio networks. Note that while Fig. 1 distinguishes the interference, or additive noise, that can be contributed by the channel, from the contributions from the individual users, in a sense, each active user in the system can represent a noise source for every other user in the system. The challenge of designing a receiver that operates well in a multiuser environment is to mitigate the effects of both the interfering users as well as the effects inherent to the wireless channel due to propagation and ambient channel noise.

Spread-spectrum signaling while somewhat bandwidth inefficient relative to more traditional narrowband signaling schemes offers certain advantages for radio communication systems. The wideband nature of the signal facilitates channel estimation and enables the resolution of multipath (described in more detail in Section 5). Multipath occurs to the presence of obstructions such as buildings, trees, and vehicles in the path between transmitter and receiver. Because of these obstructions, the transmitted signal is reflected, absorbed, and diffused; the received signal is in fact a sum of delayed and attenuated replicas of the originally transmitted signal. The access schemes associated with spread-spectrum technology tend to be more flexible. Statistical multiplexing can be exploited since all active users have bursty communication. Thus, there is potential for a capacity increase relative to narrowband signaling systems.

We shall focus on adaptive schemes for data detection; however, adaptive algorithms can also be developed for adaptive estimation of key communication parameters such as the channel, the number of active users, timing information, etc.

1.1. Access Methods

The emphasis of this entry is on signaling and detection methods appropriate for multiuser radio communications; however, it is observed that multiuser demodulation methods have found application in a variety of other fields including radar signal processing and medical imaging. There are many ways in which the radio channel resource can be shared. Two more classical methods are frequency-division multiple access (FDMA) and time-division multiple access (TDMA) (see Fig. 2). For illustrative purposes, one can view the communications resource as having two dimensions: frequency and time. In reality, other dimensions are available such as space [48,60].

The first wireless mobile communications system, the advanced mobile phone service (AMPS) employed FDMA as the multiuser access technology in 1977. In FDMA, each user is assigned a frequency band; the user can communicate continuously employing this frequency "slot." Commercial radio and broadcast television employ FDMA as a method of transmitting many different station signals to an individual receiver. In TDMA, each active user is assigned a nonoverlapping time slot. During its

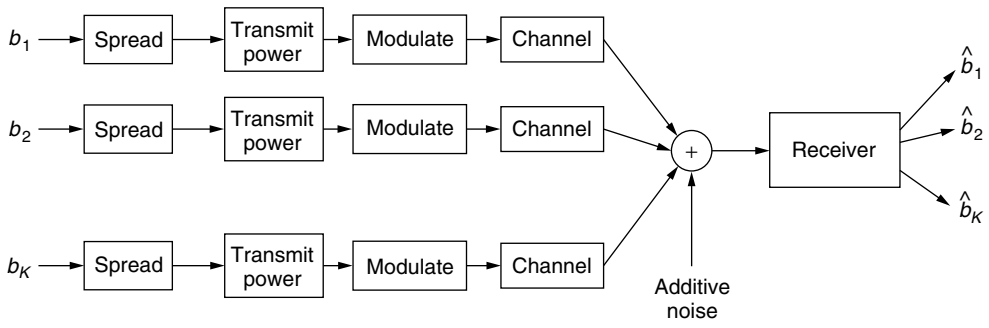


Figure 1. Multiuser system.

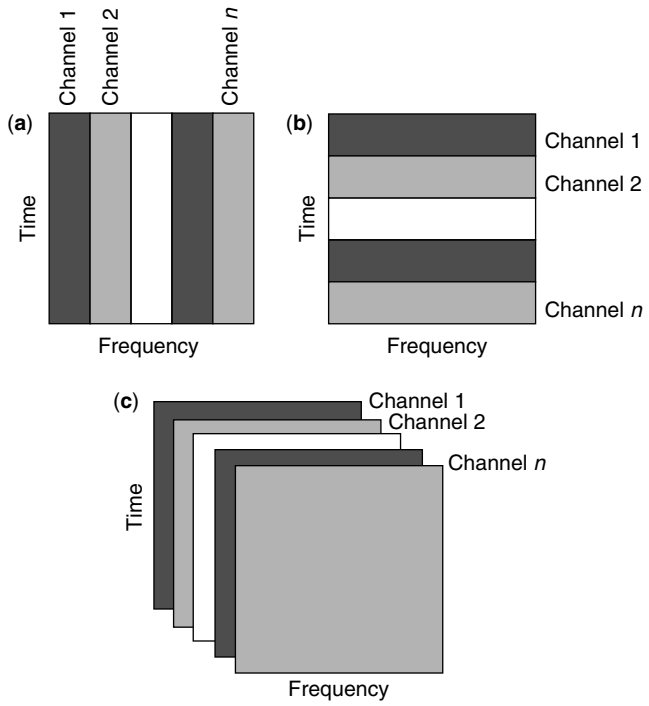


Figure 2. Multiple-access methods: (**a**) frequency division; (**b**) time division; (**c**) code division.

assigned time slot, the active user transmits over the entire frequency band allocated to the TDMA service. The TDMA access scheme can be considered to be the dual of FDMA: users are assigned non-overlapping time slots and utilize the entire frequency band during this time slot. Users share the resource by communicating one at a time, in round-robin fashion. Currently, there are three variations of TDMA in use for commercial wireless communications. The Global System for Mobile Communications (GSM) is widely deployed in Europe, North America, and parts of Asia in the 900-, 1800-, and 1900-MHz frequency bands. In place in Japan, is the Pacific Digital Cellular system, which also employs TDMA. And finally, North American Digital Cellular exists in North America at the 800- and 1900-MHz bands.

Both of these methods, FDMA and TDMA, assign orthogonal channels to each active user and have a predetermined maximum number of users that can be serviced simultaneously. We thus consider TDMA and FDMA to be fixed resource allocation schemes. We note that if there are fewer users than the maximum number of "slots," resources can be wasted (not used) for these fixed resource allocation schemes.

The multiple access strategy of interest for this work, is code-division multiple access (CDMA), also illustrated in Fig. 1. In CDMA, each active user is assigned a waveform that exploits the total time and frequency bands allocated for the service. Both TDMA and FDMA can be considered as special cases of CDMA. By allowing for user waveforms with more general characteristics, CDMA signals can be designed to be more immune to the effects of the wireless channel and to allow more users to share the radio channel. This additional user capacity, however, will come at the expense of degradation in performance or at the expense of a more sophisticated and thus generally more complex receiver structure.

The particular type of CDMA considered here is direct-sequence CDMA (DS-CDMA). In the present implementations of standardized DS-CDMA, (long code) the spreading sequence is time-varying and has a period that is equal to that of many symbol intervals. In short code DS-CDMA, the waveform assigned to each user is generally a sequence, $\mathbf{s}$, drawn from a finite alphabet (e.g., $\{\pm 1/\sqrt{N}\}$, where N is the length of the sequence) and modulated onto a pulseshape $p(t)$ (e.g., a rectangular pulse shape or a raised cosine pulseshape). To provide a consistent definition of signal-to-noise ratio per bit, the spreading waveforms are normalized $\|\mathbf{s}\|^2 = 1$. The parameter N is the length of the spreading sequence and is also known as the *processing gain*. It is a measure of the bandwidth expansion offered by the spreading operation. In distinguishing "short code" DS-CDMA, we focus on systems where the same spreading sequence is used for each bit transmitted. An example is

$$\mathbf{s} = \frac{1}{\sqrt{N}}[-1, -1, +1, -1, +1, -1, +1, +1]$$

$$p(t) = \begin{cases} 1 & t \in [0, T_c) \\ 0 & \text{else} \end{cases}$$

The parameter T_c is called the *chip duration* and the symbol duration is thus $T = NT_c$. The spreading sequences are chosen to have desirable autocorrelation and cross-correlation properties [50].

While versions of FDMA and TDMA have been standardized for some years, standards for DS-CDMA are relatively recent. In 1993, IS95 was the first interim standard for the CDMA protocol. Since then several revisions have occurred. The current, second generation, CDMA personal communications system (PCS) is in the 1.8- and 2.0-GHz bands.

The focus on DS-CDMA is motivated by the imminent adoption of the DS-CDMA-type signaling in a variety of third generation wireless standards [1,11,38]. It is observed that for both the frequency-division duplex (FDD) and the time-division duplex (TDD) modes of UMTS, DS-CDMA multiple access is laid over the FDD and TDD duplexing schemes [16].

As a concluding note to the discussion of multiple-access schemes, we observe that TDMA and FDMA are special cases of CDMA, where typically the "spreading waveforms" are mutually orthogonal. Thus, with proper signal description, the methods described herein have utility for TDMA and FDMA systems where there is adjacent or co-channel interference (CCI) caused by dispersion or insufficient frequency reuse.

1.2. The Need for Adaptive Systems in Wireless Communications

The objective of this chapter is to introduce methods for the adaptive demodulation of data in DS-CDMA systems. Adaptive algorithms are instrumental for providing consistent performance in unknown or time-varying environments. Adaptive methods can implicitly reveal unknown parameters of a system or can be

used to track these parameters as they change over time. These characteristics of adaptive methods make them particularly suitable for wireless communications systems. In contrast to communication environments with relatively fixed characteristics, as is found in the wired environment of classical telephony, the wireless communication channel is time-varying [60]. This time-variation stems from a variety of sources: mobility of the user, changes in the active user population, and the potential mobility of interference sources. As a mobile user moves, the orientation of the user and the structures that absorb, reflect, and diffuse the user's transmitted radio signal change, thus changing the number of replica signals impinging on the receiver from the mobile user and further changing the amount of attenuation experienced by each signal. Further variability is induced by the fact that users rarely maintain a truly constant speed, especially in an urban environment. In a wireless communications system, users enter and exit communication in a bursty fashion. As noted earlier, each active user can be considered a form of interference for other users in a DS-CDMA system. Thus, with a time-varying user population, the interference experienced by a single user also varies with the population. Furthermore, as vehicles and other sources of scattering move themselves, this movement will also affect the channel experienced by the user. Another possible scenario is one in which the receiver has only partial knowledge of the communication or interference environment. Thus, in a cellular system, a base station may have accurately estimated parameters for the active users within that cell, but may not have information about out-of-cell interferers. Thus, it is clear that the wireless channel is diverse and changing. Because of these inherent characteristics of the wireless communications environment, we shall see that adaptive algorithms can be used to mitigate the effects of this channel.

2. SIGNAL MODEL

For illustrative purposes, we initially focus on a simplified communication scenario: bit-synchronized multiple users communicating over an additive white Gaussian noise (AWGN) channel. We shall assume that the front-end filter of the receiver is synchronized to the users and is coherent. The chip-matched filtered signal can be represented as a sequence of vectors:

$$\begin{aligned}\mathbf{r}(i) &= \sum_{k=1}^{K} A_k b_k(i)\mathbf{s}_k + \mathbf{n}(i)\\ &= \mathbf{SAb}(i) + \mathbf{n}(i)\end{aligned}$$

$$\text{where } \mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K]$$

$$\mathbf{A} = \text{diag}\,[A_1, A_2, \ldots, A_K]$$

$$\mathbf{b}(i) = [b_1(i), b_2(i), \ldots, b_K(i)]^T$$

where K is the number of active users and $\mathbf{n}(i)$ is the additive noise process, modeled as a white, Gaussian random vector process with zero mean and covariance $\sigma^2\mathbf{I}_N$. The matrix $\mathbf{I}_L$ is an $L \times L$ identity matrix. User k's received amplitude, data bit at time i and spreading sequence are denoted by A_k, $b_k(i)$, and $\mathbf{s}_k$, respectively. Herein, we shall assume binary phase shift keying (BPSK) data; that is $b_k(i) = \pm 1$. If required, the statistical model for the data is that $b_k(i)$ takes on its binary values with equal probability ($\frac{1}{2}$).

To facilitate the description of some classical nonadaptive multiuser receivers, we introduce the following notions. It can be shown that a set of sufficient statistics for detecting the data of a single user or all users in a multiuser DS-CDMA system is the output of a bank of filters matched to the spreading code of each active user. These matched filter outputs are defined as

$$\mathbf{y}(i) = \mathbf{S}^H\mathbf{r}(i) = \mathbf{RAb}(i) + \tilde{\mathbf{n}}(i)$$

where $\mathbf{R} = \mathbf{S}^H\mathbf{S}$. The noise present in the system is now colored, $\tilde{\mathbf{n}}(i) \sim \mathcal{N}(\mathbf{0}_K, \sigma^2\mathbf{R})$, where $\mathbf{0}_L$ is a $L \times 1$ vector of zeros. The $K \times K$ cross-correlation matrix $\mathbf{R}$ can be interpreted as a catalog of how far apart pairs of spreading codes are in the multiuser system. The Euclidean distance between two spreading codes is

$$\begin{aligned}\|\mathbf{s}_j - \mathbf{s}_k\|^2 &= \|\mathbf{s}_j\|^2 + \|\mathbf{s}_k\|^2 - 2\mathbf{s}_j^H\mathbf{s}_k\\ &= 2 - 2\mathbf{R}[j,k]\end{aligned}$$

We shall see that the performance of any multiuser or single-user receiver is very dependent on the values of the components of $\mathbf{R}$. Note that a set of K mutually orthogonal spreading codes leads to $\mathbf{R} = \mathbf{I}_K$.

Another key matrix for multiuser detection is the data correlation matrix:

$$\mathbf{C} = \mathbf{E}\{\mathbf{rr}^H\} = \mathbf{SA}^2\mathbf{S}^H + \sigma^2\mathbf{I}_N. \tag{1}$$

The operator $\mathbf{E}\{\cdot\}$ denotes expectation over all random quantities in the argument, unless otherwise specified. We next briefly review five important static multiuser receivers: the conventional receiver, the decorrelating detector, the minimum-mean-squared-error receiver, the jointly optimal detector, and the individually optimal detector. These receivers vary in their offered performance and attendant complexity. In the subsequent sections, we study various adaptive versions of a subset of these receivers.

2.1. Conventional Receiver

This is the conventional receiver that was considered optimal prior to a deeper understanding of multiple-access interference (MAI). If the central-limit theorem [e.g., 42,59] holds, then, it was originally argued, the MAI could be modeled as AWGN Gaussian noise [e.g., 43,56]. The optimal single-user receiver for transmission in AWGN is the matched-filter receiver. The *conventional* or *matched filter* receiver output is given by

$$\hat{\mathbf{b}}(i) = \text{sgn}\,(\mathbf{y}(i)) \tag{2}$$

where the operator $\text{sgn}(\cdot)$ outputs the sign of each component of its argument. The conventional receiver

is MAI-limited and incurs a high probability of error if the power of the received signal of the desired user is significantly less than that of the interfering users. This undesirable property of the conventional receiver is termed the *near-far problem*. We note that if a receiver is insensitive to the near far problem it is deemed *near-far-resistant*. The conventional receiver also suffers if the system is highly loaded. We note that there exists one scenario in which the conventional receiver is actually optimal in terms of probability of bit error — this occurs when the spreading codes of the active users are mutually orthogonal. In a realistic wireless system, this property is difficult to maintain. It is noted, however, that the conventional receiver is very straightforward to implement and requires knowledge only of the desired user's spreading waveform and timing.

2.2. Decorrelating Detector

The decorrelating detector [29,30,55] is the receiver that *zero-forces* the MAI — that is, it completely nulls out the MAI at the possible expense of removing some of the signal energy of the user of interest. It can also be viewed as the maximum-likelihood estimator of the vector $\mathbf{Ab}(i)$. This receiver is formed by

$$\hat{\mathbf{b}}(i) = \text{sgn}\,(\mathbf{R}^{-1}\mathbf{y}(i)) \tag{3}$$

The direct construction of the decorrelator requires the knowledge of the spreading waveforms of all the active users and the associated timing information. Despite the simplicity of this receiver, the decorrelator shares many properties with that of the optimal detector to be discussed in the sequel.

2.3. Minimum Mean-Squared Error Receiver

To introduce the linear minimum-mean-squared error (MMSE) receiver [32,67], we first discuss a generic linear receiver. Let $\mathbf{z}(i)$ be the soft output of a general linear receiver $\mathbf{M}$; then

$$\mathbf{z}(i) = \mathbf{M}\mathbf{r}(i) \tag{4}$$

$$\hat{\mathbf{b}}(i) = \text{sgn}\,(\mathbf{z}(i)) \tag{5}$$

Then, the MMSE receiver is determined by

$$\mathbf{M} = \arg\min_{\mathbf{M}} \mathbf{E}\left\{\|\mathbf{b}(i) - \mathbf{z}(i)\|^2\right\} \tag{6}$$

Two equivalent solutions, ignoring positive scalings, are given by

$$\mathbf{M} = \mathbf{S}^T(\mathbf{S}\mathbf{A}^2\mathbf{S}^T + \sigma^2\mathbf{I}_N)^{-1} \tag{7}$$

$$= (\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}\mathbf{S}^T \tag{8}$$

The two forms of the MMSE receiver can be shown to be equivalent through the use of the matrix inversion lemma [15,25]. Thus, the MMSE estimate of the data is given by

$$\hat{\mathbf{b}}(i) = \text{sgn}\,\left((\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}\mathbf{y}(i)\right) \tag{9}$$

In addition to the information required by the decorrelating detector, the MMSE receiver is also a function of the received amplitudes of the active users and the noise variance. In general, the MMSE receiver outperforms the decorrelating detector. As will be observed below, decentralized implementations of the MMSE receiver (and decorrelator) exist. In considering these decentralized receivers, a few observations can be made in regard to asymptotic behavior. As the noise variance grows ($\sigma^2 \to \infty$) or as the interfering amplitudes diminish ($A_2, \ldots, A_K \to 0$), the MMSE receiver approaches the conventional receiver. Alternatively, as the noise variance decreases ($\sigma^2 \to 0$), the MMSE receiver converges to the decorrelating detector. The MMSE receiver performs the optimal tradeoff in combating multiple access interference versus suppressing ambient channel noise in a linear receiver.

2.4. A Few Points on Linear Receivers

We note that the prior receiver algorithms are all linear in nature. To summarize, the bit decision for a particular user — say, user 1 — can be written as

$$\hat{b}_1(i) = \text{sgn}\,\left(\mathbf{c}_1^H\mathbf{r}(i)\right) \tag{10}$$

Note that for data demodulation, a positive scaling of the receiver does not affect the decision. Thus for $\alpha > 0$, both $\mathbf{c}_1$ and $\alpha\mathbf{c}_1$ yield the same decision. For the sequel, we shall note the soft-decision statistic for user k as

$$z_k(i) = \mathbf{c}_k^H\mathbf{r}(i) \tag{11}$$

Thus, for the joint receivers discussed above, we can define the following single user (decentralized) linear receiver vectors:

1. Conventional receiver $\mathbf{c}_k = \mathbf{s}_k$
2. Decorrelating detector $\mathbf{c}_k = \mathbf{S}\rho_k$ where ρ_k the kth column of $\mathbf{R}^{-1}$
3. MMSE receiver $\mathbf{c}_k = \mathbf{S}\mathbf{m}_k$, where $\mathbf{m}_k$ the kth column of $(\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}$

In addition, we also define the error sequence for a time-varying linear receiver $\mathbf{c}_k(i)$:

$$e_k(i) = b_k(i) - \mathbf{c}_k(i)^H\mathbf{r}(i) \tag{12}$$

Many of the adaptive algorithms to be discussed herein update the ith instantiation of the parameter vector by a function of the value of the error.

We next consider nonlinear receivers. These receivers are sometimes more complex to implement than the linear receivers discussed above; the benefit of this added complexity is improved performance.

2.5. Jointly Optimal Detector

The jointly optimal multiuser receiver is formed by determining the maximum likelihood estimate of

$\mathbf{b}$ [63,64]. Thus

$$\hat{\mathbf{b}}(i) = \arg \max_{\mathbf{b}} p(\mathbf{y}(i)|\mathbf{b}) \tag{13}$$

$$= \arg \max_{\mathbf{b}(i)} 2\mathbf{b}(i)^H \mathbf{A}\mathbf{y}(i) - \mathbf{b}(i)^H \mathbf{ARAb}(i) \tag{14}$$

$$= \arg \max_{\mathbf{b}(i)} \Omega(\mathbf{b}(i)) \tag{15}$$

where

$$\Omega(\mathbf{b}) = 2\mathbf{b}^T \mathbf{A}\mathbf{y} - \mathbf{b}^T \mathbf{ARAb} \tag{16}$$

The spreading waveforms and amplitudes of all active users are necessary to construct the jointly optimal receiver. Note that the decorrelating detector requires only the spreading waveforms of the active users and not the amplitudes.

2.6. Individually Optimal Detector

The individually optimum receiver achieves the minimum probability of error for the user of interest [63,64]. Without loss of generality, we shall assume that user 1 is the intended user. Then, the individually optimum receiver is obtained by

$$\hat{\mathbf{b}}_1(i) = \text{sgn}\left[\sum_{\mathbf{b},\mathbf{b}_1=1} \exp\left(\frac{\Omega(\mathbf{b})}{2\sigma^2}\right) - \sum_{\mathbf{b},\mathbf{b}_1=-1} \exp\left(\frac{\Omega(\mathbf{b})}{2\sigma^2}\right)\right] \tag{17}$$

The individually optimal detector requires the same set of parameter knowledge as the MMSE detector: user spreading waveforms, amplitudes, the noise variance, and timing. We observe that the individually optimal detector converges to the jointly optimum receiver as $\sigma \to 0$.

Other nonlinear receiver structures exist such as serial or parallel interference cancellation schemes. In such algorithms, estimates of certain bits are used to reconstruct the contribution due to a subset of users, which is in turn subtracted from the received signal, thus diminishing the multiple access interference. Adaptive versions of such receivers have been considered [e.g., 26,40,68], although we do not focus on them here.

3. ADAPTIVE SYSTEMS

We begin by discussing a generic adaptive receiver algorithm and desirable properties of such an adaptive system. As in any equalizer design, adaptive multiuser receivers can be *direct* or *indirect*. In the direct adaptive detectors, the adaptive algorithm demodulates the data directly. In the indirect implementations, the adaptive subsystems adaptively estimate parameters that are then utilized in a receiver of fixed form. An illustration of these two methods is provided in Fig. 3.

Let the *objective function*[1] of interest be defined as $J(\mathbf{c})$. Our goal is to determine the parameter vector $\mathbf{c}$ such that the objective function is minimized. Typical cost

[1] The objective function can also be termed the *cost function*.

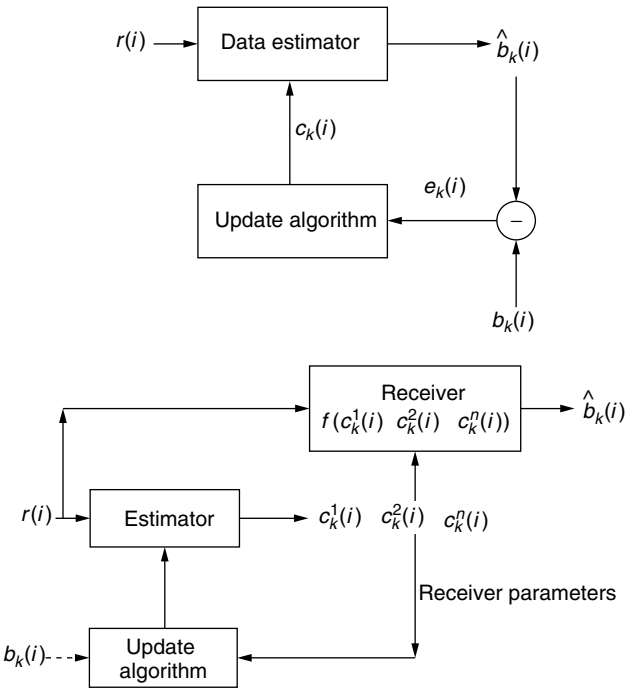


Figure 3. Direct and indirect implementations of adaptive receiver structures.

functions include the mean-squared error and the mean output energy. These cost functions will be discussed in more detail in the sequel.

Because of the stochastic nature of communication signals, the cost functions of interest will typically have the following form:

$$J(\mathbf{c}) = \mathbf{E}\{j(\mathbf{c}, \mathbf{r})\} \tag{18}$$

The expectation is taken with respect to all random quantities present in the received signal $\mathbf{r}$. If the cost function is convex in the unknown parameter vector, we can consider the method of steepest descent to determine the desired stationary point [17]. The method of steepest descent updates the parameter vector estimate in the direction opposite to the gradient of the cost function.

Ideally, update of the multidimensional parameter vector $\mathbf{c}(i)$ is conducted by

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu \nabla J(\mathbf{c}(i-1)) \tag{19}$$

The scalar μ, called the *step size*, is selected to ensure convergence of the adaptive algorithm while offering a reasonable convergence rate. These two objectives are conflicting.

The true gradient is often impossible to determine as it may require the statistics of the received signal (which are often presumed unknown as they could be employed to derive a nonadaptive receiver, if such quantities were known). Thus an approximation to the gradient is used:

$$\nabla J(\mathbf{c}(i-1) \approx \nabla j(\mathbf{c}(i-1), \mathbf{r}(i)) \tag{20}$$

The justification for such an approximation is provided by the fact that under sufficient regularity of the cost

functions and the probability distribution functions of the embedded random processes

$$\nabla J(\mathbf{c}(i-1)) = \mathbf{E}\{\nabla j(\mathbf{c}(i-1), \mathbf{r}(i))\} \tag{21}$$

4. ADAPTIVE DETECTION ALGORITHMS

Two types of adaptive algorithm are possible. In the first case, a *training signal* is used to update the adaptive algorithm. This data signal is known at both the transmitter and the receiver. This training signal is essentially a known data sequence for one or some of the active users [e.g., $b_k(i)$]. The adaptive algorithm typically employs this training signal to form an error signal that is used to drive the update. In contrast, in *blind* adaptive algorithms, there is no training signal and cost functions based on the statistics of the signal are formed and updated. In training-based systems, no meaningful data are transmitted during training; thus it is sometimes argued that training-based systems are not bandwidth-efficient. However, many blind algorithms require a considerable amount of observations (with embedded unknown data) before reliable estimates are achieved. Our focus is on training-based algorithms.

4.1. Adaptive Least Mean Squares

We shall assume a linear receiver for user 1. Thus the data are estimated via

$$\hat{b}_1(i) = \text{sgn}\,(\mathbf{c}_{\text{LMS}}(i)^H \mathbf{r}(i)) \tag{22}$$

For the adaptive least-mean-squares (LMS) algorithm, the cost function is the mean-squared error:

$$J(\mathbf{c}) = \mathbf{E}\left\{|b_1(i) - \mathbf{c}^H \mathbf{r}(i)|^2\right\} \tag{23}$$

The parameter vector is the desired linear receiver. The estimate of the parameter vector at time i ($\mathbf{c}(i)$) is obtained by employing the method of steepest descent. Thus

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu \nabla J(\mathbf{c}(i-1)) \tag{24}$$

$$= \mathbf{c}(i-1) - \mu \mathbf{E}\{\mathbf{r}(i)(b_1(i-1) - \mathbf{c}(i-1)^H \mathbf{r}(i))\} \tag{25}$$

$$\approx \mathbf{c}(i-1) - \mu \mathbf{r}(i)(b_1(i-1) - \mathbf{c}(i-1)^H \mathbf{r}(i)) \tag{26}$$

The cost function of interest here, the mean-squared error, is a special case of the least-mean p norm:

$$J(\mathbf{c}) = \mathbf{E}\left\{|b_1(i) - \mathbf{c}^H \mathbf{r}(i)|^p\right\} \tag{27}$$

It is noted that the cost function above is convex for $1 \le p < \infty$. The nonstochastic form of the cost function above was investigated for determining adaptive multiuser receivers for systems with non-Gaussian additive noise modeled as a symmetric alpha-stable process [27]. The resultant adaptive algorithm is given by

$$\mathbf{c}(i) = \mathbf{c}(i-1) + \mu p |b_1(i) - \mathbf{c}(i-1)^H \mathbf{r}(i)|^{p-1} \times \text{sgn}\,(b_1(i) - \mathbf{c}(i-1)^H \mathbf{r}(i)) \mathbf{r}(i)$$

Clearly when $p = 2$, the algorithm above reduces to the adaptive LMS algorithm noted above.

4.1.1. Convergence Analysis of LMS. The sequence $\{\mathbf{c}(i)\}_{i=1}^{\infty}$ is a sequence of random vectors. It is of interest to investigate the limiting behavior of this random sequence to determine the efficacy of the update algorithm. The typical convergence analysis provided for adaptive LMS algorithms is the study of the asymptotic bias of the update algorithm. Thus, we seek to determine whether the following is in fact true:

$$\lim_{i \to \infty} \mathbf{E}\{\mathbf{c}(i)\} \stackrel{?}{=} \mathbf{c}_{\text{MMSE}} \tag{28}$$

Recall that the optimal MMSE receiver is given by

$$\mathbf{c}_{\text{MMSE}} = (\mathbf{S}\mathbf{A}^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{s}_1 \tag{29}$$

To study the convergence behavior of LMS to this desired parameter vector, we define the parameter error vector, $\mathbf{v}(i) = \hat{\mathbf{c}}(i) - \mathbf{c}_{\text{MMSE}}$. The evolution of the mean error vector can thus be described as

$$\mathbf{E}\{\mathbf{v}(i)\} = \mathbf{E}\{\mathbf{v}(i-1)\} - \mu \mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H \mathbf{v}(i-1)\} + \mu \mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H\} \mathbf{c}_{\text{MMSE}} \tag{30}$$

$$= [\mathbf{I}_N - \mu(\mathbf{S}A^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N)] \mathbf{E}\{\mathbf{v}(i-1)\} \tag{31}$$

where

$$\mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H\} = \mathbf{S}\mathbf{A}^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N \tag{32}$$

The vector $\hat{\mathbf{c}}(i-1)$ is a function of all prior received signal vectors, $\mathbf{r}(0), \mathbf{r}(1), \ldots, \mathbf{r}(i-1)$, but is independent of the current observation $\mathbf{r}(i)$. We define the following matrix and its associated eigenvalue decomposition:

$$\mathbf{C}(\mu) = \mathbf{I}_N - \mu(\mathbf{S}\mathbf{A}^2\mathbf{S}^H + \sigma^2 \mathbf{I}_N) \tag{33}$$

$$= \mathbf{V} \Lambda(\mu) \mathbf{V}^H \tag{34}$$

where $\Lambda(\mu)$ is a diagonal matrix of the eigenvalues ($\lambda_i(\mu)$) of $\mathbf{C}(\mu)$ and $\mathbf{V}$ is a marix whose columns correspond to the eigenvectors of $\mathbf{C}(\mu)$. Thus

$$\mathbf{E}\{\mathbf{v}(i)\} = \mathbf{C}(\mu) \mathbf{E}\{\mathbf{v}(i-1)\} \tag{35}$$

Because of the orthonormal property of the eigenvectors, we obtain

$$\mathbf{V}^H \mathbf{E}\{\mathbf{v}(i)\} = \Lambda(\mu) \mathbf{V}^H \mathbf{E}\{\mathbf{v}(i-1)\} \tag{36}$$

$$\mathbf{E}\{\tilde{\mathbf{v}}(i)\} = \Lambda(\mu) \mathbf{E}\{\tilde{\mathbf{v}}(i-1)\} \tag{37}$$

where

$$\mathbf{V}^H \mathbf{E}\{\mathbf{v}(i)\} = \mathbf{E}\{\tilde{\mathbf{v}}(i)\} \tag{38}$$

We can now rewrite the linear transformation of the error vector at time i as a function of the initial error:

$$\mathbf{E}\{\tilde{\mathbf{v}}(i)\} = \Lambda(\mu)^i \mathbf{E}\{\tilde{\mathbf{v}}(0)\} \tag{39}$$

$$= \text{diag}\,[\lambda_1^i(\mu), \lambda_2^i(\mu), \ldots, \lambda_N^i(\mu)] \mathbf{E}\{\tilde{\mathbf{v}}(0)\} \tag{40}$$

The system is stable, that is, the expected error vector converges to zero if $0 < |\lambda_i(\mu)| < 1$, this implies that

$$0 < \mu < \frac{2}{\lambda_i(\mu)} \forall i \tag{41}$$

If we denote $\lambda_{\max}$ as the maximum eigenvalue of the matrix $\mathbf{SA}^2\mathbf{S}^H$, then a fixed step size that achieves the desired convergence must fall within the following range

$$0 < \mu < \frac{2}{\sigma^2 + \lambda_{\max}} \tag{42}$$

A key concern about the implementation of an adaptive algorithm is that the rate of adaptation of the algorithm be matched to the underlying rate of change of the time-varying system. For certain fast-fading channels (see Sections 4.5 and 5.1), certain families of adaptive algorithms cannot be used because they cannot track the channel variations.

4.2. Recursive Least-Squares Methods

The recursive least-squares algorithm differs from the adaptive LMS algorithm just derived in that the cost function is a deterministic one. For an observation record corresponding to M symbols, it is desired to minimize the metric

$$J(\mathbf{c}) = \sum_{i=1}^{M} \lambda^{M-i} |e(i)|^2, \tag{43}$$

where λ is deemed the *forgetting factor* and is assumed to be $0 < \lambda < 1$. With this form of weighting, more recent observations are given more weight than are previous observations. As before, the desired parameter vector is a linear receiver, thus

$$\hat{b}_1(i) = \text{sgn}\,(\mathbf{c}_{\text{RLS}}(i)^H \mathbf{r}(i)) \tag{44}$$

The resultant algorithm, which exploits the matrix inversion lemma (to avoid a computationally expensive direct matrix inverse) is given as follows:

$$\mathbf{k}(i) = \frac{\lambda^{-1}\mathbf{P}(i-1)\mathbf{r}(i)}{1 + \lambda^{-1}\mathbf{r}(i)^H\mathbf{P}(i-1)\mathbf{r}(i)} \tag{45}$$

$$\zeta(i) = b_1(i) - \mathbf{c}(i-1)^H\mathbf{r}(i) \tag{46}$$

$$\mathbf{c}(i) = \mathbf{c}(i-1) + \mathbf{k}(i)\zeta(i) \tag{47}$$

$$\mathbf{P}(i) = \lambda^{-1}\mathbf{P}(i-1) - \lambda^{-1}\mathbf{k}(i)\mathbf{r}(i)^H\mathbf{P}(i-1) \tag{48}$$

The typical initialization of the algorithm is with $\mathbf{P}(0) = \delta^{-1}\mathbf{I}_N$ and $\mathbf{c}(0) = \mathbf{0}_N$, where δ is a small positive constant. In the multiuser receiver context, we can also initialize the weight vector to be $\mathbf{c}(0) = \mathbf{s}_1$; thus, the receiver is initialized as the conventional matched-filter receiver. The vector $\mathbf{k}(i)$ is called the *gain vector*. We also note that $\mathbf{P}(i)$ is the current estimate of the inverse of the weighted (each observation is weighted by λ) data correlation matrix.

There is a subtle distinction between the *tentative* error sequence $\zeta(i)$ and the *current* error sequence $e(i)$, in that $\zeta(i)$ employs the past value of receiver vector $\mathbf{c}(i-1)$ while $e(i)$ is formed with $\mathbf{c}(i)$. This error is often termed the *prediction error* as it uses the past estimate of the receiver to predict the current data.

For the general case (linear estimation of a scalar process), one can show that the RLS algorithm has a rate of convergence that is far superior to that of the LMS algorithm [17]; however, this comes as the expense of greater computational complexity. Theoretically, the RLS algorithm produces zero excess mean-squared error. This result is dependent on the presence of a statistically stationary environment, which is not typically experienced in a wireless communications channel. Finally, the convergence of the RLS algorithm is not dependent on the eigenvalue spread of the data correlation matrix. In contrast, the LMS algorithm's convergence speed is dependent on this eigenvalue spread. The implication for DS-CDMA systems is the fact that the eigenvalue spread can be quite large in a system where there is a great disparity in received powers amongst the users. Thus, if tight power control is not in place, the LMS algorithm will experience slow convergence. However, for many scenarios, this improved convergence speed for RLS is in fact not observed [65].

We conclude this subsection by noting a work [5] that considers the steady-state behavior of LMS and RLS operating in a multiuser environment by employing the results on steady-state excess mean-squared error provided elsewhere [12,17].

4.3. Adaptive Linear Minimum Probability of Error Receivers

Through performance comparisons and analysis, it can be shown that the MMSE receiver (and thus the convergent adaptive LMS receiver) offers strong performance and can combat multiple-access interference (MAI). However, in a digital communications system, the true performance metric of interest is the probability of bit detection error rather than the mean-squared error. Thus, we next consider a set of adaptive linear receivers that endeavor to minimize the probability of bit detection error.

Let the transmitted, noiseless, multiuser signal be represented as

$$\mathbf{m}(i) = \sum_{k=1}^{K} A_k b_k(i)\mathbf{s}_k = \mathbf{SAb}(i) \tag{49}$$

Then the received signal is simply $\mathbf{r}(i) = \mathbf{m}(i) + \mathbf{n}(i)$. Conditioned on knowing all of the parameters of the interfering users, that is, if we know the matrices $\mathbf{S}$, $\mathbf{A}$, $\mathbf{b}$ completely, then the probability of error for the bit interval i for a linear receiver $\mathbf{c}(i)$ is given by,

$$P_e(\mathbf{b}, \mathbf{c}) = Q\left(\frac{b_1(i)\mathbf{c}(i)^H\mathbf{m}(i)}{\sigma\|\mathbf{c}(i)\|}\right) \tag{50}$$

where $Q(x)$ is the complementary cumulative distribution function[2] of a zero-mean, unit-variance Gaussian random variable.

[2] $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} dv.$

The goal of the adaptive minimum probability of error receiver is to determine the optimal receiver vector $\mathbf{c}^*$ that satisfies [34]

$$\mathbf{c}^* = \arg \min_{\mathbf{c}} J(\mathbf{c}) \tag{51}$$

where

$$J(\mathbf{c}) = \mathbf{E}\{P_e(\mathbf{b}, \mathbf{c})\} \tag{52}$$

The method for updating is also based on steepest descent; thus we invoke Eq. (19). We also approximate $\nabla \mathbf{E}\{P_e(\mathbf{b}, \mathbf{c})\} \approx \nabla P_e(\mathbf{c}(i), \mathbf{b}(i))$. It has been shown [34] that this approximation is an unbiased estimator of the true gradient. The desired update procedure is

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu \nabla P_e(\mathbf{c}(i), \mathbf{b}(i)) \tag{53}$$

$$= \mathbf{c}(i-1) - \mu - \frac{1}{2\sqrt{2\pi}} \times \left\{ \exp\left(-\frac{1}{2} \left(\frac{b_1(i)\mathbf{c}(i-1)^H \mathbf{m}(i)}{\sigma \|\mathbf{c}(i-1)\|} \right)^2 \right) \right. \tag{54}$$

$$\left. \times \left[\frac{b_1(i)\mathbf{m}(i)}{\sigma \|\mathbf{c}(i-1)\|} + \frac{b_1(i)\mathbf{c}(i-1)^H \mathbf{m}(i)}{\sigma^2 \|\mathbf{c}(i-1)\|^3} \mathbf{c}(i-1) \right] \right\} \tag{55}$$

This detector is deemed the *clairvoyant* adaptive receiver [34] as it requires the knowledge of the transmitted signal $\mathbf{m}(i)$ and not just the training sequence $b_1(i)$. This unrealistic assumption is removed by performing a maximum-likelihood estimation (MLE) operation to determine an estimate for $\mathbf{m}(i)$. Thus $\mathbf{m}(i)$ above is replaced with $\hat{\mathbf{m}}(i)$, where

$$\hat{\mathbf{m}}(i) = \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^H \mathbf{r}(i) \tag{56}$$

Therefore, the estimated transmitted signal vector is simply the projection of the received signal onto the subspace spanned by the spreading codes of the active users.

The function $J(\mathbf{c})$ given above is not strictly a convex function. However, in [34], a series of conditions are established that ensure convexity. In brief, at each iteration of the update algorithm, the receiver must be near-far resistant.

An alternative approach to minimizing the probability of error through an adaptive linear receiver has been investigated in [49], where the cost function of interest is the expected value of the *single-letter distortion measure*, $\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c})$. For simplicity, we assume that the prior probabilities of the BPSK transmitted data for the desired user are $\pi_0 = \pi_1 = \frac{1}{2}$. Then

$$\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c}) = \tfrac{1}{4}\{[1 + \operatorname{sgn}(\mathbf{c}^H \mathbf{r}_0)] + [1 - \operatorname{sgn}(\mathbf{c}^H \mathbf{r}_1)]\} \tag{57}$$

The vectors $\mathbf{r}_0, \mathbf{r}_1$ represent training signals given that $b_1 = -1$ and $b_1 = 1$ respectively:

$$\mathbf{r}_0 = \mathbf{r}(i) | b_1 = -1 \tag{58}$$

$$\mathbf{r}_1 = \mathbf{r}(i) | b_1 = +1 \tag{59}$$

Thus, if the receiver vector, $\mathbf{c}$, achieves correct decisions, $\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c}) = 0$. The cost function is strictly positive if errors occur. Note that

$$P_e = \mathbf{E}\{\zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c})\} = J(\mathbf{c}) \tag{60}$$

This cost function is independent of the underlying noise distribution. Thus the methods described below can be applied in scenarios where non-Gaussian (impulsive) noise is present. Steepest descent methods can be considered where the noisy estimate of the gradient of $J(\mathbf{c})$ is used. In this case

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu_n \nabla \zeta(\mathbf{r}_0, \mathbf{r}_1; \mathbf{c}) \tag{61}$$

However, in contrast to the adaptive minimum probability of error receiver discussed earlier, determining the true gradient poses difficulty. Thus the adaptive receiver update algorithm is constructed by determining one or two-sided difference approximations to the gradient of $J(\mathbf{c})$. For example, construct the following vector function at time i

$$\mathbf{x}(\mathbf{c}(i)) = [\mathbf{x}(\mathbf{c}(i))_1, \mathbf{x}(\mathbf{c}(i))_2, \ldots, \mathbf{x}(\mathbf{c}(i))_N,] \tag{62}$$

$$\mathbf{x}(\mathbf{c}(i))_j = \frac{1}{2\alpha(i)}\{[\zeta(\mathbf{r}_0(i), \mathbf{r}_1(i); \mathbf{c}(i) + \alpha(i)\mathbf{e}_j)] - [\zeta(\mathbf{r}_0(i), \mathbf{r}_1(i); \mathbf{c}(i) - \alpha(i)\mathbf{e}_j)]\} \tag{63}$$

where the vector $\mathbf{e}_j$ is the jth coordinate unit vector; $\alpha(i)$ is selected such that $\alpha(i) = \beta i^{-(1/4)}$ for some $\beta > 0$. The receiver update algorithm with the approximate gradient in place is given by

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu_i \mathbf{x}(\mathbf{c}(i-1)) \tag{64}$$

With key conditions on the step size sequence (μ_i) and the perturbation rate $\alpha(i)$, it can be shown that the recursion in Eq. (64) converges with probability 1 to the minimizing value of $\mathbf{c}^*$. Under the assumption of AWGN, necessary conditions can be established to ensure the convexity of the cost function of interest.

Numerical results show that for the scenario of a strong desired user, the adaptive algorithm in Eq. (64) achieves a significant performance improvement over the adaptive LMS algorithm, the true MMSE solution and the decorrelating detector. However, as the near-far ratio increases, these three algorithms achieve comparable probability of error.

Finally, it is noted that the work of Yeh and Barry [69] can be viewed as the generalization of these two approaches for multiuser detection to the equalization of single-user intersymbol interference channels. Thus a steepest-descent approach for the probability of error in Gaussian channels as well as a steepest-descent method based on a single-letter distortionlike measure are considered. Further, low-complexity approximations and convergence rate increasing alternatives are also investigated.

4.4. Adaptive Optimal Receivers

Because of the assumption of additive Gaussian noise, the form of the individually optimal receiver described by Eq. (17) is the same as a radial basis function (RBF) network. Thus, methods previously considered to adaptively update the parameters of such a network can

be employed to form an adaptive multiuser detector [36]. The general form of a RBF network output is given by

$$z(i) = \sum_{j=1}^{N} w_j \Phi\left(\frac{\|\mathbf{r}(i) - \mathbf{m}_j\|}{\sigma_j}\right). \tag{65}$$

Here $\Phi(\cdot)$ is a continuous, nonlinear function from $\Re^+ \rightarrow \Re^+$ (other conditions on Φ, which stem from regularization and approximation theory can be found in [45]). The input vector is $\mathbf{r}(i)$, $\mathbf{m}_j$ is called the *center* of the RBF neuron, σ_j is the *spread* of the neuron, and the w_j are the *weights* that optimize some performance criterion. The methods by which the $\Phi(\cdot)$, $\mathbf{m}_j$, σ_j, and w_j are selected, constitute much of the research on RBF networks. Traditionally, the centers were randomly chosen from the given data set; the spreads were then calculated by determining the minimum distance between centers using the appropriate distance metric and the weights could be solved for given a simple error criterion (e.g., MMSE) [28]. Methods for determining these network parameters are often unique to the application for which the RBF network is used.

The application of RBF networks as multiuser receivers is inspired by the work of Chen et al. [7] and Chen and Mulgrew [8], who used these networks and modifications thereof to perform equalization of intersymbol interference (ISI) channels. While intersymbol interference is analogous to MAI, the distinctions between these two noise sources imply that modifications of the previous RBF techniques are necessary to ensure good performance from the RBF network as an adaptive multiuser detector.

By inspecting Eq. (17), the decision rule for the individually optimum receiver, we see that we can rewrite the decision rule as a function of the received signal (versus the matched filter outputs) as follows:

$$\hat{b}_1(i) = \operatorname{sgn}\left[\sum_{j=1}^{2^{K-1}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{r}(i) - \underline{\Theta}_1 - \underline{\mu}_j\|^2\right\} - \sum_{j=1}^{2^{K-1}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{r}(i) - \underline{\Theta}_0 - \underline{\mu}_j\|^2\right\}\right]$$

This decision rule can then be mapped to the RBF network with the following definitions of key functions and parameters:

$$\Phi(x) = \exp\{-x^2\} \tag{66}$$

$$\sigma_j = \sqrt{2}\sigma \tag{67}$$

(where σ^2 is the Gaussian noise variance)

$$\underline{\Theta}_d = (-1)^{d+1} A_1 \mathbf{s}_1 \tag{68}$$

$$\underline{\mu}_j = \sum_{k=2}^{K} A_k b_k \mathbf{s}_k \tag{69}$$

(for some permutation of the b_k values)

$$\mathbf{m}_j \in \{\underline{\mu}_j + \Theta_0, \underline{\mu}_j + \Theta_1 | j = 1, \ldots, 2^{K-1}\} \tag{70}$$

$$w_j = \begin{cases} 1 & \text{if } \mathbf{m}_j = \underline{\mu}_l + \Theta_1 \\ -1 & \text{if } \mathbf{m}_j = \underline{\mu}_l + \Theta_0 \end{cases} \tag{71}$$

With this mapping in hand, we turn to methods used to train RBF networks to determine the centers and the weights. We first present a *supervised* learning algorithm whereby the data of *all active users* must be known; this is akin to the clairvoyant receiver of Ref. [34]. With known bits, it is known to which center a received signal corresponds. Given i observations of the received signal, we update the centers as follows:

$$\mathbf{b}(i) \leftrightarrow \text{index}, j \tag{72}$$

$$\hat{\mathbf{m}}_j(i) = \frac{i-1}{i}\hat{\mathbf{m}}_j(i-1) + \frac{1}{i}\mathbf{r}(i) \tag{73}$$

This supervised algorithm is somewhat impractical as it requires coordination of all active users to send training data simultaneously. Next the *k-means clustering algorithm* [31] is described:

$$\text{index}, j^* = \arg\min_l \|\hat{\mathbf{m}}_l(i-1) - \mathbf{r}(i)\|^2 \tag{74}$$

$$\hat{\mathbf{m}}_{j^*}(i) = \frac{i-1}{i}\hat{\mathbf{m}}_{j^*} l(i-1) + \frac{1}{i}\mathbf{r}(i) \tag{75}$$

The convergence of the supervised algorithm to the true centers trivially follows from the law of large numbers [42,59]. The convergence of the k-means algorithm has been investigated [31]. To speed up convergence, the k-means algorithm can be initialized with estimates of the centers using matched-filter outputs to perform coarse amplitude estimation. Then all possible permutations of the noiseless received signal are constructed.

To adaptively determine the weights, a LMS update is considered:

$$\mathbf{w}(i+1) = \mathbf{w}(i) + \mu(z_1(i) - b_1(i))\underline{\Phi}(\mathbf{r}(i))$$

where μ is the adaptation gain, $z_1 = \mathbf{w}(i)^H \underline{\Phi}(\mathbf{r}(i))$ is the output of the RBF network at time i, and $\underline{\Phi}(\cdot)$ is the vector RBF nonlinear functions applied to the input. It has been shown [36] that even with estimated centers, the mean weight vector is a positively scaled version of the desired weights.

4.5. Reduced-Rank Adaptive MMSE Filtering

We return to linear adaptive receivers to consider reduced-rank adaptive MMSE algorithms. To introduce the reduced-rank methods, we recall the full rank-fixed MMSE receiver for the desired user 1. This linear MMSE receiver is an $N \times 1$ vector $\mathbf{c}$ that minimizes the mean-squared error (MSE):

$$\mathbf{c}_{\text{MMSE}} = \arg\min_{\mathbf{c}} \text{MSE} = \arg\min_{\mathbf{c}} \mathbf{E}\{|b_1(i) - \mathbf{c}^H \mathbf{r}(i)|^2\} = \mathbf{C}^{-1}\mathbf{p} \tag{76}$$

Recall that $\mathbf{C} = \mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H\}$ is the data cross-correlation matrix and $\mathbf{p} = \mathbf{E}\{b_1(i)\mathbf{r}(i)\}$ is termed the *steering vector* [35].

The adaptive algorithms [35,70] can be used to estimate $\mathbf{c}_{\text{MMSE}}$, even in a time-varying channel. When N is large, convergence is slow. Reduced rank techniques reduce the

number of taps to be adaptively tracked by projecting the received signal vector onto a lower-dimensional subspace. Let D be the resultant lower dimension, where $D < N$, the projection is

$$\tilde{\mathbf{r}}(i) = \mathbf{P}_D^H \mathbf{r}(i) \tag{77}$$

where $\mathbf{P}_D$ is the $N \times D$ projection matrix and the D dimensional signal is denoted by a tilde. The vector $\tilde{\mathbf{r}}(i)$ is then the input to a length D tap delay linear filter. When the MMSE criterion is applied, the optimum coefficients for the D dimensional space are given by

$$\tilde{\mathbf{c}}_{\text{MMSE}} = \tilde{\mathbf{C}}^{-1}\tilde{\mathbf{p}}, \quad \text{where} \quad \tilde{\mathbf{C}} = \mathbf{P}_D^H \mathbf{C} \mathbf{P}_D, \tilde{\mathbf{p}} = \mathbf{P}_D^H \mathbf{p}. \tag{78}$$

4.5.1. Projection Matrix Selection. A number of methods for selecting the projection matrix $\mathbf{P}_D$ are considered here.

The multistage Wiener filtering (MWF) algorithm for DS-CDMA has been presented [20]. The resultant algorithm is a specialization of the work by Goldstein et al. [14]. The MWF projection matrix is given by

$$\begin{aligned} \mathbf{P}_D^{MW} &= [\mathbf{g}_{MW,1}\, \mathbf{g}_{MW,2} \cdots \mathbf{g}_{MW,D}] \\ &= \left[\mathbf{h}_1 \; \mathbf{B}_1^H \mathbf{h}_2 \; \cdots \prod_{j=1}^{D-1} \mathbf{B}_j^H \mathbf{h}_D \right] \end{aligned} \tag{79}$$

where $\mathbf{g}_{MW,j}$, $j = 1, \ldots, D$ are implicitly defined. The matrix $\mathbf{B}_j$ is an $(N-j) \times (N-j+1)$ blocking matrix, namely, $\mathbf{B}_j \mathbf{h}_j = 0$. The vector $\mathbf{h}_j$ is the normalized correlation vector $E\{\tilde{d}_{j-1}(i)\mathbf{r}_{j-1}(i)\}$, where $\mathbf{r}_j(i) = \mathbf{B}_j \mathbf{r}_{j-1}(i)$ with $\mathbf{r}_0(i) = \mathbf{r}(i)$, and $\tilde{d}_j(i) = \mathbf{h}_i^H \mathbf{r}_{j-1}(i)$ with $\tilde{d}_0(i) = b_1(i)$ [20].

The projection matrix for the auxiliary vector filtering (AVF) algorithm is given by [41]

$$\mathbf{P}_D^{AV} = [\mathbf{g}_{AV,1}\, \mathbf{g}_{AV,2} \cdots \mathbf{g}_{AV,D}] \tag{80}$$

where $\mathbf{g}_{AV,1}$ is also the normalized correlation vector $E\{b_1(i)\mathbf{r}(i)\} = \mathbf{h}_1$, and $\mathbf{g}_{AV,j}$, $j = 2, \ldots, D$ are the auxiliary vectors, given by [41]

$$\begin{aligned} &\mathbf{g}_{AV,j+1} \\ &= \frac{\mathbf{C}\mathbf{g}_{AV,j}^{Eq} - (\mathbf{g}_{AV,1}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,1} - \sum_{l=2}^{j} (\mathbf{g}_{AV,l}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,l}}{\| \mathbf{C}\mathbf{g}_{AV,j}^{Eq} - (\mathbf{g}_{AV,1}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,l} - \sum_{l=2}^{j} (\mathbf{g}_{AV,l}^H \mathbf{C} \mathbf{g}_{AV,j}^{Eq})\mathbf{g}_{AV,l} \|} \end{aligned} \tag{81}$$

where $\mathbf{g}_{AV,j}^{Eq} = \mathbf{g}_{AV,1} - \sum_{l=2}^{j} w_l \mathbf{g}_{AV,j}, w_l, l = 2, \ldots, j$ are the optimized constants [41] and $\| \cdot \|$ is the vector norm. By construction, the $\mathbf{g}_{AV,j}$, $j = 1, \ldots, D$ are normalized orthogonal vectors.

Using the Cayley–Hamilton (CH) theorem, Moshavi et al. proposed the following projection matrix [37]:

$$\mathbf{P}_D^{CH} = [\mathbf{g}_{CH,1}\; \mathbf{g}_{CH,2} \; \cdots \; \mathbf{g}_{CH,D}] = [\mathbf{h}_1 \; \mathbf{C}\mathbf{h}_1 \cdots \mathbf{C}^{D-1}\mathbf{h}_1] \tag{82}$$

The vector $\mathbf{h}_1$ is the one defined previously.

The projection matrix can also be selected by performing an eigendecomposition on $\mathbf{C}$:

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^H \tag{83}$$

The matrix $\Lambda = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_N]$ contains the eigenvalues associated with the eigenvectors which are the columns of $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$. If the eigenvalues are ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, the desired projection matrix is then $\mathbf{P}_D = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D]$. This method is titled the method of *principal components*. A related method, termed the *cross-spectral reduced-rank method*, selects the D eigenvectors that result in a minimum mean-squared error estimate of the bit. This method requires knowledge of the desired user's spreading waveform, but offers improved performance over the principal-components methods [13]. Eigendecomposition methods, in general, are not attractive due to their high attendant computational complexity.

Finally, a very simple method of rank reduction is to employ partial despreading as proposed in [57]. The desired user's spreading code is decomposed into multiple subvectors: $\mathbf{s}_1 = [\mathbf{s}_1^{(1)}, \mathbf{s}_1^{(2)}, \ldots, \mathbf{s}_1^{(D)}]$. The projection matrix is then constructed as

$$\mathbf{P}_D = \begin{bmatrix} \mathbf{s}_1^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{s}_1^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{s}_1^{(D)} \end{bmatrix} \tag{84}$$

Through a key simplification of the AVF algorithm, it can be shown that the AVF method is equivalent to the MWF algorithm [10]. The work by Chen et al. [10] provides a simplified proof of the equivalence of the static MWF receiver and that based on the Cayley Hamilton expansion of the correlation matrix inverse [37]. This observation was made previously [20]. Because of its simplicity of presentation, we focus on adaptive implementations of the MWF method as presented [20].

4.5.2. Adaptive Reduced-Rank Detection. In general, the following two cost functions can be set up for determining adaptive reduced rank receivers:

$$J(\tilde{\mathbf{c}})_{\text{LS}} = \sum_{i=1}^{M} \| b_1(i) - \tilde{\mathbf{c}}^H \tilde{\mathbf{r}}(i) \|^2 \tag{85}$$

$$J(\tilde{\mathbf{c}})_{\text{MMSE}} = \mathbf{E}\{ \| b_1(i) - \tilde{\mathbf{c}}^H \tilde{\mathbf{r}}(i) \|^2 \} \tag{86}$$

where

$$\tilde{\mathbf{r}}(i) = \hat{\mathbf{P}}^H(i)\mathbf{r}(i) \tag{87}$$

Note that $\hat{\mathbf{P}}(i)$ is an estimate at time i of the projection matrix $\mathbf{P}$. The methods discussed in the previous sections can be employed to derive adaptive algorithms.

A host of adaptive algorithms for the reduced rank multistage Wiener filter for DS-CDMA signaling has been provided [20]. Batch and recursive algorithms based on least-squares as well stochastic gradient descent algorithms are considered. In addition, both blind and training-based methods are provided. The training-based stochastic gradient method, which offers strong performance, is described here. The stochastic gradient algorithm of Ref. 20 is given by two sets of "recursions". We note that the subscript n indicates the stage number

of the multistage system, while the index i corresponds to the symbol interval timing.

Initialization:

$$d_0(i) = b_1(i)\,\mathbf{r}_0(i) = \mathbf{r}(i) \tag{88}$$

Forward recursion: at each i, for $n = 1, \ldots D$

$$\hat{\mathbf{p}}_n(i) = (1-\mu)\hat{\mathbf{p}}_n(i-1) + \mu d^*_{n-1}(i)\mathbf{r}_{n-1}(i) \tag{89}$$

$$\hat{\mathbf{c}}_n(i) = \frac{\hat{\mathbf{p}}_n(i)}{\|\hat{\mathbf{p}}_n(i)\|} \tag{90}$$

$$\mathbf{B}_n(i) = \text{null}\,[\hat{\mathbf{c}}^H_n(i)] \tag{91}$$

$$d_n(i) = \hat{\mathbf{c}}_n(i)^H \mathbf{r}_{n-1}(i) \tag{92}$$

$$\mathbf{r}_n(i) = \mathbf{B}^H_n(i)\mathbf{r}_{n-1}(i) \tag{93}$$

Backward recursion: decrementing $n = D, \ldots, 1$

$$\zeta_n(i) = (1-\mu)\zeta_n(i-1) + \mu|\varepsilon_n|^2 \tag{94}$$

$$w_n(i) = \frac{\|\hat{\mathbf{p}}_n(i)\|}{\zeta_n(i)} \tag{95}$$

$$\varepsilon_{n-1}(i) = d_{n-1}(i) - w^*_n(i)\varepsilon_n(i) \tag{96}$$

$$\text{where } \varepsilon_D(i) = d_D(i) \tag{97}$$

The estimated data is given by

$$\hat{b}_1(i) = \text{sgn}(w^*_1(i)\varepsilon_1(i)) \tag{98}$$

For the MWF, the matrices $\mathbf{B}_n$ are *blocking matrices*, which are selected to be orthogonal to the nested filers $\hat{\mathbf{c}}_n$. The scalar sequence $d_n(i)$ is the output of the filter $\hat{\mathbf{c}}_n$ and the filter in the next stage would be selected as

$$\hat{\mathbf{c}}_{n+1} = \frac{\mathbf{E}\,\{d^*_n\mathbf{r}_n\}}{\|\mathbf{E}\,\{d^*_n\mathbf{r}_n\}\|} \tag{99}$$

where $\mathbf{r}_n$ is the output of the blocking matrix $\mathbf{B}_n$. Thus, the stochastic gradient algorithm above provides adaptive estimators for these key quantities. The choice of blocking matrices is not unique. For example, one can select $\mathbf{B}_n = \mathbf{I} - \mathbf{c}_n\mathbf{c}^H_n$. However, it is shown in [10], that the projection matrix for the D-stage MWF algorithm is in fact independent of the choice of the blocking matrices within the class of row orthonormal blocking matrices.

4.6. Decision-Directed and Decision Feedback Methods

In the most simplistic of views, one could convert the training based adaptive algorithms previously described into blind adaptive algorithms by considering the "hard" decision of the receiver to be the true data and comparing this to the associated "soft" decision. This notion is depicted in Fig. 4. However, it should be noted that if this adaptive receiver is not initialized to a receiver vector that is close in some sense to the desired receiver vector, the algorithm could converge to the receiver of a more powerful user in a near-far environment. Algorithms that employ this idea are termed *decision-directed*.

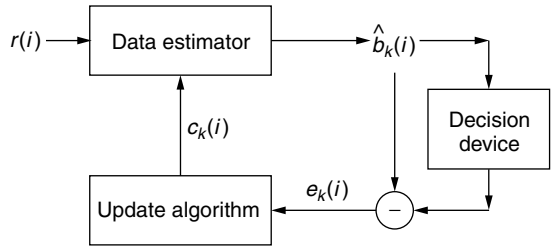


Figure 4. Decision-directed adaptive receiver.

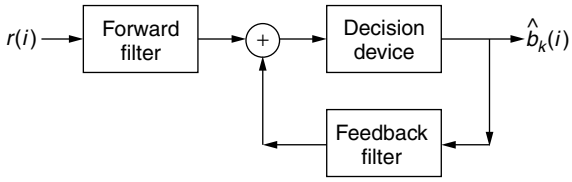


Figure 5. Decision feedback adaptive receiver.

Another set of algorithms that also feed back decisions, but are far more robust than the simple decision-directed scheme exhibited in Fig. 4, are decision feedback algorithms, depicted in Fig. 5. Adaptive schemes are used to determine the feedforward and feedback weight vectors. Examples of the design of such structures specifically for multiuser spread-spectrum systems can be found in the literature [44,51,52,58].

5. MULTIPATH FADING ENVIRONMENTS

In practice, the AWGN channel is too idealized of a model. In this section, the multipath channel model is introduced and the resultant signal model is provided. In addition to the assumption of a more realistic channel, we also consider possible asynchronism amongst the active users.

The received baseband signal corresponding to a single-symbol interval, which is coherently demodulated, chip-matched filtered and sampled at the chip rate is now described as

$$\mathbf{r}(i) = \sum_{k=1}^{K}\sum_{l=1}^{L_k} A_k[h_{kl}(i)b_k(i)\mathbf{s}^+_{kl} + h_{kl}(i-1)b_k(i-1)\mathbf{s}^-_{kl}] + \mathbf{n}(i) \tag{100}$$

We note that this received vector is of dimension $N \times 1$. In general for asynchronous, multipath systems, such short observations lead to degraded performance. We note, however, that the signal descriptions provided are easily generalized to the consideration of received signal vectors that correspond to multiple symbol intervals. The complex coefficients $h_{kl}(i)$ correspond to the multipath coefficients. The number of multipaths for user k is denoted by L_k. The partial spreading vectors $\mathbf{s}^+_{kl}$ and $\mathbf{s}^-_{kl}$ correspond to truncated and shifted versions of the spreading codes. For example, if we consider a two user system, with $L_1 = 2$

and $L_2 = 1$, employing spreading codes of length $N = 7$, a possible realization of the partial spreading codes would be

$$\begin{array}{lccccccc}
\mathbf{s}_{11}^{+} = [& 0 & 0 & \mathbf{s}_1(1) & \mathbf{s}_1(2) & \mathbf{s}_1(3) & \mathbf{s}_1(4) & \mathbf{s}_1(5) &] \\
\mathbf{s}_{11}^{-} = [& \mathbf{s}_1(6) & \mathbf{s}_1(7) & 0 & 0 & 0 & 0 & 0 &] \\
\mathbf{s}_{21}^{+} = [& 0 & 0 & 0 & 0 & \mathbf{s}_2(1) & \mathbf{s}_2(2) & \mathbf{s}_2(3) &] \\
\mathbf{s}_{22}^{+} = [& 0 & 0 & 0 & 0 & 0 & \mathbf{s}_2(1) & \mathbf{s}_2(2) &] \\
\mathbf{s}_{21}^{-} = [& \mathbf{s}_2(4) & \mathbf{s}_2(5) & \mathbf{s}_2(6) & \mathbf{s}_2(7) & 0 & 0 & 0 &] \\
\mathbf{s}_{22}^{-} = [& \mathbf{s}_2(3) & \mathbf{s}_2(4) & \mathbf{s}_2(5) & \mathbf{s}_2(6) & \mathbf{s}_2(7) & 0 & 0 &]
\end{array} \tag{101}$$

Here we note that user 1 has a delay of $\tau_1 = 2$ chips with respect to the receiver clock, while user 2 has a delay of $\tau_2 = 4$ chips. We can denote the discrete baseband equivalent representation of the multipath channel for user k as the complex vector $\mathbf{h}_k = [h_{k1}, h_{k2}, \ldots h_{kL_k}]^T$. This model assumes that the multipath channel can be represented as a finite-impulse response filter whose taps are spaced one chip (T_c) apart [48,60]. The *effective spreading code* is the convolution of the transmitted spreading sequence with this channel vector and is denoted by $\overline{\mathbf{s}}_k = \mathbf{h}_k \star \mathbf{s}_k$, where $\star$ denotes convolution. Note that $\overline{\mathbf{s}}_k$ is of length $N + L_k - 1$. If we let

$$\overline{\mathbf{s}}_k^{+} = \sum_{l=1}^{L_k} h_{kl}\mathbf{s}_{kl}^{+} \tag{102}$$

$$\overline{\mathbf{s}}_k^{-} = \sum_{l=1}^{L_k} h_{kl}\mathbf{s}_{kl}^{-} \tag{103}$$

$$\overline{\mathbf{S}}^{+} = \left[\overline{\mathbf{s}}_1^{+}, \overline{\mathbf{s}}_2^{+}, \cdots \overline{\mathbf{s}}_K^{+}\right] \tag{104}$$

$$\overline{\mathbf{S}}^{-} = \left[\overline{\mathbf{s}}_1^{-}, \overline{\mathbf{s}}_2^{-}, \cdots \overline{\mathbf{s}}_K^{-}\right] \tag{105}$$

then we can rewrite the received signal vector as

$$\mathbf{r}(i) = \overline{\mathbf{S}}^{+}\mathbf{A}\mathbf{b}(i) + \overline{\mathbf{S}}^{-}\mathbf{A}\mathbf{b}(i-1) + \mathbf{n}(i) \tag{106}$$

From this description it is clear to see that the K user asynchronous multiuser system can be considered as a $2K$ user synchronous system if observations corresponding to N chips (one symbol interval) are employed. We also observe the following relationship between $\overline{\mathbf{s}}_k$ and the vectors $\overline{\mathbf{s}}_k^{+}, \overline{\mathbf{s}}_k^{-}$. Let τ_k be the delay of user k in integer multiples of a chip; then

$$[\overline{\mathbf{s}}_k^{+}, \overline{\mathbf{s}}_k^{-}] = \left[\underbrace{0, 0, \ldots 0}_{1\times\tau_k}, \overline{\mathbf{s}}_k, \underbrace{0, 0, \ldots, 0}_{(N-\tau_k-L_k+1)\times 1}\right] \tag{107}$$

We note that if the channel of the desired user is completely known, that is, if we have knowledge of $\mathbf{h}_k$ and τ_k, then the previous algorithms can be applied directly where observations of length $N + L_k - 1$ are collected and the desired user's spreading code, $\mathbf{s}_k$, is replaced by the effective spreading code, $\overline{\mathbf{s}}_k$. If information about the interfering users is necessary, the interfering users are modeled as $2(K-1)$ synchronous users with spreading waveforms $\overline{\mathbf{s}}_k^{+}$ and $\overline{\mathbf{s}}_k^{-}$. The information required to implement the various adaptive receivers considered herein in the asynchronous multipath environment is noted in Table 1. This table is constructed with the view that each user is viewed as a *single* user with spreading code $\overline{\mathbf{s}}_k$. We note that an alternative view is possible, which can obviate the need for the channel vector $\mathbf{h}_k$, but necessitates knowledge of the channel length L_k and the relative delay τ_k. In this approach, each user is considered to be L_k users with a spreading code that is a shifted version of the other spreading codes. Then, receivers can be designed for each path. The final decision is made by combining the soft decisions from each L_k adaptive receiver. The choice of combining coefficients remains an open question. A typical approach is to consider *equal-gain combining* as in Barbosa and Miller [2]. This approach can be applied to the bulk of the receivers considered.

5.1. MMSE Receivers for Multipath Fading Channels

Because of its simplicity of implementation, strong performance, and amenability to adaptive implementation, there has been significant interest in applying the linear MMSE receiver to multipath fading channels. The construction of the true linear MMSE receiver requires knowledge of all the active users' spreading codes, timing, and channel state information [33,66]. However, an adaptive implementation of the this receiver can be achieved with prior information comparable to that of the conventional matched filter (MF) receiver, namely, information of the user of interest only and not that of the interfering users. Table 3 summarizes the applicability of the algorithms in section 4 to the multipath channel case.

Barbosa and Miller [2] proposed a modified MMSE receiver for flat fading channels in which the channel phase of the desired user is estimated and then compensated for in the MMSE receiver input. However, in frequency-selective fading channels, determining accurate estimates of the channel phases for all resolvable paths is at best challenging, and often impossible. As a result, noncoherent MMSE receivers become a more favorable choice for rapidly fading multipath environments. Several works have considered training-signal-independent, or blind approaches to developing MMSE-based receivers for multipath channels. Such receivers are robust to deep

Table 1. Information Required to Construct Nonadaptive Multiuser Receivers

Receiver	Signature of User 1, $\mathbf{s}_1$	Signature of AllUsers, $\mathbf{S}$	Relative Amplitudes, $\mathbf{A}$	Noise Variance, σ^2	Timing Information, τ_i
Matched filter	Yes	No	No	No	Yes
MMSE receiver	Yes	Yes	Yes	Yes	Yes
Decorrelator	Yes	Yes	No	No	Yes
Jointly optimal	Yes	Yes	Yes	No	Yes
Individually optimal	Yes	Yes	Yes	Yes	Yes

Table 2. Taxonomy of Adaptive Receivers in Terms of Direct or Indirect Implementation

Direct	Indirect
LMS	RBF [36]
RLS	—
Linear/letter distortion MPER[a]	—
MWF/AVF	—

[a]Minimum probility of error rate.

Table 3. Knowledge Required for Implementation of Adaptive Multiuser Receivers in an Asynchronous Multipath Environment

Need τ_k	Need $\bar{\mathbf{s}}_k, \tau_k$	Need $\bar{\mathbf{s}}_k, \tau_k \forall k$
LMS/RLS	RBF[a]	RBF
Letter distortion MPER	MWF/AVF	Linear MPER
MWF[a]	—	—

[a]An implementation is possible, but degraded performance will be experienced.

fades as they do not attempt to track the amplitude and phase fluctuations of the user of interest [18,21,46]. However, this robustness comes at the cost of higher excess MSE for adaptive implementations of such blind receivers. This feature is in contrast to training sequence based adaptive MMSE algorithms [46]. The training sequence based differential least-squares (DLS) algorithm proposed in [21] suffers from robustness to deep fades. Thus, to achieve acceptable performance and robustness simultaneously, the DLS algorithm is switched to a blind adaptive implementation when the receiver is in deep fade [21,46].

We next discuss two advances in the development of adaptive DS-CDMA receivers with a view to moderately fast fading environments. With the challenge of accurately estimating the channel phase in a fast fading environment, both methods consider differentially encoded phase shift keying (DPSK) to avoid estimating the channel phase. The first method [70] is initially developed for flat fading channels and then extended to multipath channels using the multipath combining technique noted above. The second method makes a key observation based on the technique of Zhu et al. [70] to develop an improved method for estimating the data correlation matrix, $\mathbf{C}$ [9].

5.1.1. Differential MMSE. As noted previously, the development of the differential MMSE criterion presumes a flat fading channel, thus the contribution in the received signal due to user k can be modeled as, $\alpha_k(i)b_k(i)\mathbf{s}_k$. The random process $\alpha_k(i)$ is complex-valued and represents the channel fading. The proposed modified MMSE-based cost function is

$$J(\mathbf{c}) = \mathbf{E}\left[|b_1(i-1)\mathbf{c}^H\mathbf{r}(i) - b_1(i)\mathbf{c}^H\mathbf{r}(i-1)|^2\right] \tag{108}$$

subject to

$$\mathbf{c}^H\mathbf{C}\mathbf{c} = 1 \tag{109}$$

where $\mathbf{C}$ remains the data correlation matrix; however, expectation is now taken over the channel coefficients as well as the data and the noise. The objective of this cost function is to suppress the multiple access interference while endeavoring to recover a scaled version of the datastream of the desired user. Thus the resultant soft decision will include an unknown complex scalar. By assuming that $\alpha_1(i) \approx \alpha_1(i-1)$, we can employ heuristic arguments to show the suppression of the multiple-access interference. By invoking some simple assumptions, it can be shown [70] that the solution to the cost function above is a scaled version of the true MMSE receiver in Eq. (76). The required assumptions are

1. $\mathbf{E}\left[\text{Re}\ (\alpha_1(i)\alpha_1^*(i-1))\right] > 0$
2. $\mathbf{E}[\text{Re}\ (b_1(i)b_1^*(i-1)\alpha_k(i)d_k(i)\alpha_j^*(i-1)d_j^*(i-1))]$
   $= \gamma\delta(k-1)\delta(j-1)$

where γ is a positive constant and $\delta(\cdot)$ is the Krönecker delta function.

The general solution to the minimization of Eq. (109) is the determination of the following generalized eigenvalue problem:

$$\mathbf{Q}\mathbf{c} = \lambda\mathbf{C}\mathbf{c} \tag{110}$$

where

$$\mathbf{Q} = \text{Re}\ \{\mathbf{E}\left[b_1(i)b_1^*(i-1)\mathbf{r}(i)\mathbf{r}(i-1)^H + b_1(i-1)b_1^*(i)\mathbf{r}(i-1)\mathbf{r}(i)^H\right]\} \tag{111}$$

An efficient algorithm, denoted the *power algorithm* [15], can be utilized to determine the desired generalized eigenvector $\mathbf{c}$. The power algorithm requires a key matrix $\mathbf{M} = \mathbf{C}^{-1}\mathbf{Q}$. Either block adaptive or recursive methods can be employed to estimate $\hat{\mathbf{C}}$ and $\hat{\mathbf{Q}}$ from the data. In addition, a gradient-based, recursive least-squares-type algorithm can be employed:

$$\mathbf{k}(i) = \frac{\mathbf{P}(i-1)\mathbf{r}(i)b_1^*(i-1)}{\lambda + |b_1(i-1)|^2\mathbf{r}(i)^H\mathbf{P}(i-1)\mathbf{r}(i)} \tag{112}$$

$$\zeta(i) = b_1(i)\mathbf{c}(i-1)^H\mathbf{r}(i-1) - b_1(i-1)\mathbf{c}(i-1)^H\mathbf{r}(i) \tag{113}$$

$$\mathbf{P}(i) = \lambda^{-1}\mathbf{P}(i-1) - \lambda^{-1}b_1(i-1)\mathbf{k}(i)\mathbf{r}(i)^H\mathbf{P}(i-1) \tag{114}$$

$$\mathbf{c}(i) = \frac{\mathbf{c}(i-1) + \beta_a\mathbf{k}(i)\zeta(i)^*}{|\mathbf{c}^H(i-1)\mathbf{r}(i-1)|} \tag{115}$$

The differential MMSE adaptive methods for flat fading channels can be extended to the multipath environment by constructing a correlator for each path and then combining the soft outputs.

5.1.2. Improved Correlation Matrix Estimation. While the modified differential MMSE criterion proposed in Ref. 70 offers solid performance in flat fading channels, and also enables various adaptive implementations, the receiver experiences significant degradation over the true MMSE receiver in frequency-selective fading channels. A further challenge to consider is that in the presence of unknown multipath (or imperfectly estimated multipath), there is performance degradation for MMSE based

receivers [21,35]. This is because in the fast multipath fading environment, an interfering user appears as multiple virtual users for the adaptive MMSE receiver, a phenomenon known as *interferer multiplication* [70]; it has been observed that the performance of the MMSE receiver degrades with the number of effective users in the system [2]. We note that this problem is not an issue in the flat fading environment where both training-sequence-based and blind adaptive MMSE detectors can achieve performance close to that of the true MMSE receiver, although the detector may not track the channel parameter of each interfering user perfectly [21,35].

The approach to be discussed herein makes the following key observation about the cost function of [70]. The new approach is based on observations for flat fading channels, however, the method offers strong performance improvements even in multipath fading channels [9]. In flat fading channels, where $L = 1$ [see Eq. (100)], the true $\mathbf{R}$ is given by

$$\mathbf{R} = \mathbf{R}_u + \mathbf{R}_I = P_1\mathbf{s}_1\mathbf{s}_1^T + \left\{\sum_{k=2}^{K} P_k[\mathbf{s}_k^+(\mathbf{s}_k^+)^T + \mathbf{s}_k^-(\mathbf{s}_k^-)^T] + \sigma^2\mathbf{I}_N\right\} \quad (116)$$

where $\mathbf{R}_u = P_1\mathbf{s}_1\mathbf{s}_1^T$ is the correlation matrix for user 1, $P_k = A_k^2$ and $\mathbf{R}_I$ is the interference correlation matrix which is defined implicitly in this equation. For this scenario, it can be shown through use of the matrix inversion lemma

$$\frac{\mathbf{R}^{-1}\mathbf{s}_1}{\mathbf{s}_1^T\mathbf{R}^{-1}\mathbf{s}_1} = \frac{\mathbf{R_I}^{-1}\mathbf{s}_1}{\mathbf{s}_1^T\mathbf{R_I}^{-1}\mathbf{s}_1} \quad (117)$$

In other words, $\mathbf{c}_{\text{MMSE}}$ can be expressed as a function of $\mathbf{R}_I$ only and not as a function of $\mathbf{R}_u$ [22]. This important property will form the basis of the proposed correlation matrix estimation scheme.

We next recall the objective function of Zhu et al. [70] in Eq. (109). One can show that

$$E\{\hat{b}_1(m)\hat{b}_1(m-1)\mathbf{y}(m)\mathbf{y}(m-1)^H\} \approx (\check{\mathbf{s}}_{1,1:L}^+)^H\check{s}_{1,1:L}^+ = \mathbf{R}_u \quad (118)$$

where the assumptions $\hat{b}_1(m) \approx b_1(m)$, $\gamma_{1l}(m) \approx \gamma_{1l}(m-1)$, $l = 1, \ldots, L$ and $E\{|\gamma_{1l}(m)|^2\} = 1$ have been used, and $\check{\mathbf{s}}_{1,1:L}^+ = \mathbf{s}_{1,1:L}^+\mathbf{A}_1$, $\mathbf{A}_1 = \text{diag}([A_{11}, A_{12}, \ldots, A_{1L_1}])$ and diag(.) denotes to diagonalize.

Now, the new correlation matrix estimation scheme is given by

$$\hat{\mathbf{R}}_I(m) = \hat{\mathbf{R}}(m) - \hat{\mathbf{R}}_u(m) = \lambda\hat{\mathbf{R}}_I(m-1) + \mathbf{r}(m)\mathbf{z}(m)^H \quad (119)$$

where

$$\mathbf{z}(m) = \mathbf{r}(m) - \hat{b}_1(m)\hat{b}_1(m-1)\mathbf{r}(m-1) \quad (120)$$

for the blind adaptive MMSE detector. The new correlation matrix estimate given in Eq. (120) results in significantly improved performance for the blind adaptive MMSE receiver. One might think that for asynchronous systems, the performance improvement is due to an equivalent observation window enlargement since $\hat{\mathbf{R}}_u(m)$ uses $\mathbf{r}(m-1)$ as well as $\mathbf{r}(m)$ for estimation. However, it turns out that even for synchronous flat fading environments, the performance gain is still evident.

We note that data decisions of $b_1(m)$ are needed for the adaptation of the receivers, as shown in Eq. (120). Similar to the decision-directed receivers, the proposed MMSE receivers also start with the conventional RAKE receiver [47] if the initialization of the estimate of $\mathbf{R}$ is given by a small identity matrix [17]. However, in contrast to the decision-directed adaptive MMSE receivers, which might lose track of the user of interest and lock on the user with the strongest signal instead, the blind adaptive MMSE receivers will always lock on the intended user since the steering vector is assumed to be known and we are always in the right direction.

6. DIFFERENT ENVIRONMENTS AND FURTHER EXTENSIONS

In this section, we discuss modifications of the previously discussed linear adaptive receiver algorithms based on minimizing the mean-squared error. These modifications are inspired by characteristics of the particular communications environment, signaling and/or reception scheme. In particular, we shall consider receivers tailored to: multiple data rates, binary phase shift keying (BPSK), and multiple sensors.

6.1. Multiple Data Rates

With the discussion of future standards [1,11,39], there has been heightened interest in wireless systems that offer multiple data rates in an integrated fashion. In such a system, users can consider transmitting at one of a class or possible data rates. The methods by which one can modify a data rate are varied. Herein, we shall focus on systems where the symbol rates among users are different. Digital communication signals are, in general, cyclostationary; that is, the received signal $r(t)$, is wide-sense cyclostationary with period T if

$$\mathbf{E}\{r(t)\} = \mathbf{E}\{r(t+T)\}$$

$$\mathbf{E}\{r(t_1)r^*(t_2)\} = \mathbf{E}\{r(t_1+T)r^*(t_2+T)\}$$

For a digital communications signal, this period T is the symbol duration. In a multiuser environment where all users transmit at the same data rate, the period of cyclostationarity is also T; if multiple users have different symbol rates, then the received sum signal retains its cyclostationary nature. However, the period of cyclostationarity is now the least common multiple of the individual symbol rates, that is, $T^* = \text{LCM}(T_1, T_2, \ldots, T_K)$. It can be shown that the optimal MMSE detector is time-varying, but periodic with period T^* [4,53,54]. For the

design of MMSE receivers, the desired time-varying filter will be the solution to

$$\mathbf{c}_k(i)_{\text{MMSE}} = \arg\ \min \mathbf{E}\{|b_k(i) - \mathbf{c}(i)^H \mathbf{r}(i)|^2\}$$

The optimal receiver $\mathbf{c}_k(i)_{\text{MMSE}}$, due to the fact that it is periodic, will have the following Fourier series representation:

$$\mathbf{c}_k(i)_{\text{MMSE}} = \sum_{q=1}^{R} \mathbf{c}_k^{(q)} \exp\left(\frac{j2\pi qi}{R}\right)$$

Let R be such that $T = R \min_i T_i$. The subfilters $\mathbf{c}_k^{(q)}$ are not time-varying. Thus, we can rewrite our desired criterion in (121) as

$$\mathbf{c}_k(i)_{\text{MMSE}} = \arg\ \min \mathbf{E}\{|b_k(i) - \tilde{\mathbf{c}}^H \tilde{\mathbf{r}}(i)|^2\}$$

where

$$\tilde{\mathbf{c}} = \left[\mathbf{c}_k^{(0)^H}, \mathbf{c}_k^{(1)^H}, \ldots, \mathbf{c}_k^{(R-1)^H}\right]^H$$

$$\tilde{\mathbf{r}}(i) = \mathbf{r}(i) \odot \theta(i)$$

$$\theta(i) = \left[1, e^{j2\pi i}, \ldots, e^{j2\pi(R-1)i}\right]^T$$

The Schur product operator is denoted by $\odot$; in the sequel, the Krönecker product operator will also be required, $\otimes$. The desired static receiver is given by

$$\tilde{\mathbf{c}} = \mathbf{E}\left[\tilde{\mathbf{r}}(i)\tilde{\mathbf{r}}(i)^H\right]^{-1} \mathbf{E}\left[b_k(i)\tilde{\mathbf{r}}(i)\right]$$

where

$$\mathbf{E}\left[b_K(i)\tilde{\mathbf{r}}(i)\right] = \mathbf{s}_1 \otimes \theta(i)$$

The receiver can be implemented as depicted in Fig. 6. These receiver structures have been investigated for both multimedia applications and narrowband interference suppression [4,53,54]. The extension to the case of of multipath channels is straightforward if the channel is known; the strategies for unknown channels as discussed above can be applied. The adaptive implementation of this multirate receiver is discussed by Buzzi et al. [4]. A host

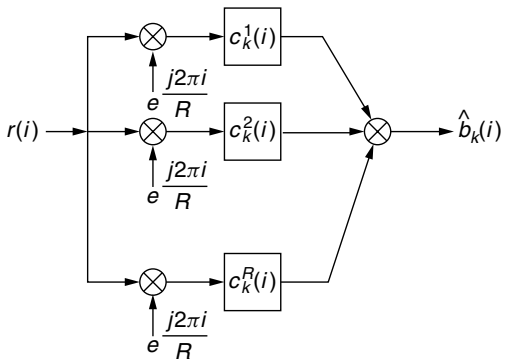


Figure 6. Implementation of MMSE receiver for multirate signaling.

of adaptive implementations are possible, as described in the prequel, with the modification that the observation is no longer $\mathbf{r}(i)$, but rather $\tilde{\mathbf{r}}(i)$.

6.2. Binary Signaling

In the presence of flat fading or multipath, the received signal is complex-valued; thus the use of the decision statistic for a linear receiver, under the assumption of BPSK signaling, is

$$\hat{b}_1(i) = \text{sgn}\{\text{Re}(\mathbf{c}^H \mathbf{r})\} \tag{121}$$

Thus, in contrast to the MMSE cost function in Eq. (76), it is of interest to consider optimizing the following cost function:

$$J(\mathbf{c}) = \mathbf{E}\left\{[b_1(i) - \text{Re}(\mathbf{c}^H \mathbf{r}(i))]^2\right\} \tag{122}$$

The desired receiver is found by forming the alternative, but equivalent optimization problem:

$$\mathbf{c} = \arg \min_{\mathbf{c}_a} \mathbf{E}\left\{|b_1(i) - \mathbf{c}_a^H \mathbf{r}_a(i)|^2\right\} \tag{123}$$

where

$$\mathbf{c}_a = [\mathbf{c}^H, \mathbf{c}^T]^H \tag{124}$$

$$\mathbf{r}_a(i) = [\mathbf{r}(i)^H, \mathbf{r}(i)^T]^H \tag{125}$$

The resulting solution is

$$\mathbf{c}_a = \mathbf{C}_a^{-1} \mathbf{p}_a \tag{126}$$

where

$$\mathbf{p}_a = \mathbf{E}\{b_1(i)\mathbf{r}(i)\} \tag{127}$$

$$\mathbf{C}_a = \begin{bmatrix} \mathbf{C} & \mathbf{C}' \\ \mathbf{C}'^* & \mathbf{C}^* \end{bmatrix} \tag{128}$$

$$\mathbf{C}' = \mathbf{E}[\mathbf{r}(i)\mathbf{r}(i)^T] \tag{129}$$

As in the case of multirate data signals as described in Section 6.1, the previous adaptive implementations of an MMSE-based algorithm can be constructed by replacing the observation $\mathbf{r}(i)$ with $\mathbf{r}_a(i)$. Given that the algorithm exploits further structure in the received signal, the binary signaling-based MMSE receiver outperforms the conventional MMSE receiver [3].

7. CONCLUSIONS

In this article, methods for adaptive reception of DS-CDMA multiuser signals has been provided. Key static receivers were reviewed and their adaptive counterparts provided. As there is much structure embedded in multiuser DS-CDMA receivers, a variety of specialized adaptive receivers are possible and have been pursued. For further reading, several tutorials on

adaptive multiuser detection have been written. Perhaps the most extensive is the paper by Woodward and Vucetic [65]. The Honig–Tsatsanis article [19] focuses on blind adaptive algorithms based on second order statistics as well as the reduced-rank methods described in Section 4.5. An extensive bibliography, coupled with the derivation and description of the MOE algorithm [23], has been provided [62]. In addition, two texts on adaptive receivers are suggested: Refs. 24 and 17. These texts consider the derivation of a host of adaptive algorithms based on a variety of cost functions and also describe the properties of the derived adaptive receivers.

BIBLIOGRAPHY

1. F. Adachi, M. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communication systems, *IEEE Commun. Mag.* 56–69 (Sept. 1998).
2. A. N. Barbosa and S. L. Miller, Adaptive detection of DS-CDMA signals in fading channels, *IEEE Trans. Commun.* **46**(5): 115–124 (Jan. 1998).
3. S. Buzzi, M. Lops, and A. Tulino, A new family of MMSE multiuser receivers for interference suppression in DS-CDMA systems employing BPSK modulation, *IEEE Trans. Commun.* **49**(1): 154–167 (Jan. 2001).
4. S. Buzzi, M. Lops, and A. Tulino, Partially blind adaptive MMSE interference rejection in asynchronous DS-CDMA networks over frequency selective fading channels, *IEEE Trans. Commun.* **49**(1): 94–108 (Jan. 2001).
5. G. Caire, Adaptive linear receivers for DS-CDMA, Part 1: Steady-state performance analysis, *IEEE Trans. Commun.* **48**(10): 1712–1724 (Oct. 2000).
6. D. S. Chen and S. Roy, An adaptive multi-user receiver for CDMA systems, *IEEE J. Select. Areas Commun.* **12**(5): 808–816 (June 1994) (issue on code-division multiple-access networks II).
7. S. Chen, G. J. Gibson, C. F. N. Cowan, and P. M. Grant, Reconstruction of binary signals using an adaptive radial-basis-function equalizer, *Signal Process.* **22**: 77–93 (1991).
8. S. Chen and B. Mulgrew, Overcoming co-channel interference using an adaptive radial basis function equalizer, *Signal Process.* **28**: 91–107 (1992).
9. W. Chen and U. Mitra, An improved blind adaptive MMSE receiver for fast fading DS-CDMA channels, *IEEE J. Select. Areas Commun.* **2**: 758–762 (2001).
10. W. Chen, U. Mitra, and P. Schniter, Reduced rank detection schemes for DS-CDMA communication systems, *IEEE Trans. Inform. Theory* (in press).
11. E. Dahlman, B. Gudmundson, M. Nilsson, and J. Sköld, UMTS/IMT2000 based on wideband CDMA, *IEEE Commun. Mag.* 70–80 (Sept. 1998).
12. E. Eleftherious and D. D. Falconer, Tracking properties and steady-state performance of RLS adaptive filter algorithms, *IEEE Trans. Acoust. Speech Signal Process.* **34**(5): 1097–1110 (Oct. 1986).
13. J. S. Goldstein and I. S. Reed, Reduced rank adaptive filtering, *IEEE Trans. Signal Process.* **45**(2): 492–496 (Feb. 1997).
14. J. S. Goldstein, I. S. Reed, and L. L. Scharf, A multistage representation of the Wiener filter based on orthogonal projections, *IEEE Trans. Inform. Theory* **44**(7): 2943–2959 (Nov. 1998).
15. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins Univ. Press, Baltimore, 1989.
16. M. Haardt et al., The TD-CDMA based UTRA TDD mode, *IEEE J. Select. Areas Commun.* **18**(8): 1375–1386 (Aug. 2000).
17. S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
18. M. Honig, U. Madhow, and S. Verdu, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **41**(4): 944–960 (July 1995).
19. M. Honig and M. Tsatsanis, Adaptive techniques for multiuser CDMA receivers, *IEEE Signal Process. Mag.* **17**(3): 49–61 (May 2000).
20. M. L. Honig and J. S. Goldstein, Adaptive reduced-rank interference suppression based on the multi-stage Weiner filter, 2000.
21. M. L. Honig, S. L. Miller, M. J. Shensa, and L. B. Milstein, Performance of adaptive linear interference suppression in the presence of dynamic fading, **49**(4): 635–645 (April 2001).
22. M. L. Honig and W. Xiao, Performance of reduced-rank linear interference suppression for DS-CDMA, **47**(5): 1928–1946 (July 2001).
23. Michael Honig, Upamanyu Madhow, and Sergio Verdú, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **41**(4): 944–960 (July 1995).
24. M. L. Honig and D. G. Messerschmitt, *Adaptive Filters*, Kluwer, Boston, 1984.
25. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK, 1985.
26. W.-S. Hou and B.-S. Chen, Adaptive detection in asynchronous code-division multiple-access system in multipath fading channels, *IEEE Trans. Commun.* **48**(5): 863–874 (May 2000).
27. S. Lee and J. Dickerson, Adaptive minimum dispersion interference suppression for DS-CDMA systems in non-gaussian impulsive channels, *Proc. of Milcom'97*, IEEE, Nov. 1997, Vol. 2, pp. 857–861.
28. D. Lowe, Adaptive radial basis function nonlinearities, and the problem of generalization, *Proc. 1st IEE Int. Conf. Artificial Neural Networks*, IEE, Oct. 1989, pp. 171–175.
29. R. Lupas and S. Verdú, Linear multiuser detectors for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **35**(1): 123–136 (Jan. 1989).
30. R. Lupas and S. Verdú, Near-far resistance of multi-user detectors in asynchronous channels, *IEEE Trans. Commun.* **38**: 496–508 (April 1990).
31. J. B. MacQueen, Some methods of classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.
32. U. Madhow and M. L. Honig, MMSE interference suppression for direct-sequence spread spectrum CDMA, *IEEE Trans. Commun.* **42**(12): 3178–3188 (Dec. 1994).

33. U. Madhow and M. L. Honig, MMSE interference suppression for DS/SS CDMA, *IEEE Trans. Commun.* **42**: 3178–3188 (Dec. 1994).

34. N. B. Mandayam and B. Aazhang, Gradient estimation for sensitivity analysis and adaptive multiuser interference rejection in code-division multiple-access systems, *IEEE Trans. Commun.* **45**(7): 848–858 (July 1997).

35. M. L. Miller, S. L. Honig, and L. B. Milstein, Performance analysis of MMSE receivers for DS-CDMA in frequency-selective fading channels, **48**(11): 1919–1929 (Nov. 2000).

36. U. Mitra and H. V. Poor, Neural network techniques for adaptive multi-user demodulation, *IEEE J. Select. Areas Commun.* **12**(9): 1460–1470 (Dec. 1994) (issue on intelligent communications systems).

37. S. Moshavi, E. G. Kanterakis, and D. L. Schilling, Multistage linear receivers for DS-CDMA systems, *Int. J. Wireless Inform. Networks* **3**(1): 1–17 (1996).

38. T. Ojanperä and R. Prasad, An overview of air interface multiple access for IMT-2000/UMTS, *IEEE Commun. Mag.* 82–95 (Sept. 1998).

39. Y. Okamura and F. Adachi, Variable-rate data transmission with blind rate detection for coherent DS-CDMA mobile radio, *IEICE Trans. Commun.* **E81-B**: 71365–71373 (July 1998).

40. T.-B. Oon, R. Steele, and Y. Li, Performance of an adaptive successive serial-parallel CDMA cancellation scheme in flat rayleigh fading channels, *Vehic. Technol. Conf.* **49**(1): 130–147 (Jan. 2000).

41. D. A. Pados, F. J. Lombardo, and S. N. Batalama, Auxiliary-vector filters and adaptive steering for DS-CDMA single-user detection, *IEEE Trans. Vehic. Technol.* **48**(6): 1831–1839 (Nov. 1999).

42. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed., McGraw-Hill, New York, 1984.

43. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread spectrum communications — a tutorial. *IEEE Trans. Commun.* **30**(5): 855–884 (May 1982).

44. H. Ping, T. Tjhung, and L. Rasmussen, Decision-feedback blind adaptive multiuser detector for synchronous CDMA system, *Vehic. Technol.* **49**(1): 159–166 (Jan. 2000).

45. T. Poggio and F. Girosi, Networks for approximation and learning, *Proc. IEEE* **78**(9): 1481–1497 (1990).

46. H. V. Poor and X. Wang, Code-aided interference suppression for DS-CDMA communications — Part II: Parallel blind adaptive implementations, *IEEE Trans. Commun.* **45**(9): 1112–1122 (Sept. 1997).

47. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

48. J. G. Proakis, *Digital Communications*, 2nd ed., McGraw Hill Series in Communications and Signal Processing, McGraw-Hill, New York, 1989.

49. I. N. Psaromiligkos, S. N. Batalama, and D. A. Pados, On adaptive minimum probability of error linear filter receivers for DS-CDMA channels, *IEEE Trans. Commun.* **47**(7): 1092–1102 (1999).

50. M. B. Pursley and D. V. Sarwate, Performance evaluation of phase-coded spread-spectrum multiple-access communication — Part II: Code sequence analysis, *IEEE Trans. Commun.* **25**(8): 800–803 (Aug. 1977).

51. P. Rapajic, M. Honig, and G. Woodward, Multiuser decision-feedback detection: performance bounds and adaptive algorithms, *Proc. ISIT*, IEEE, Nov. 1998, p. 34.

52. P. B. Rapajic and B. S. Vucetic, Adaptive receiver structures for asynchronous CDMA systems, *IEEE J. Select. Areas Commun.* **12**(4): 685–697 (May 1994).

53. A. Sabharwal, U. Mitra, and R. Moses, Low complexity MMSE receivers for multirate DS-CDMA systems, *Proc. 2000 Conf. Information Sciences and Systems*, Princeton, NJ, March 2000, Vol. 1, pp. TA3–TA18.

54. A. Sabharwal, U. Mitra, and R. Moses, MMSE receivers for multirate DS-CDMA systems, *IEEE Trans. Commun.* **49**(12): 2184–2197 (Dec. 2001).

55. K. S. Schneider, Optimum detection of code division multiplexed signals, *IEEE Trans. Aerospace Electron. Syst.* **16**: 181–185 (Jan. 1979).

56. K. Simon, J. Omura, R. Scholtz, and B. Levitt, *Spread Spectrum Communications*, 2nd ed., Vol. III, Computer Science Press, New York, 1985.

57. R. Singh and L. B. Milstein, Interference suppression for DS-CDMA, *IEEE Trans. Commun.* **47**(3): 446–453 (March 1999).

58. J. E. Smee and S. C. Schwartz, Adaptive feedforward/feedback architectures for multiuser detection in high data rate wireless CDMA networks, *IEEE Trans. Commun.* **48**(6): 996–1011 (June 2000).

59. H. Stark and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1994.

60. G. L. Stüber, *Principles of Mobile Communication*, Kluwer, Boston, 1996.

61. S. Ulukus and R. D. Yates, A blind adaptive decorrelating detector for CDMA systems, *IEEE J. Select. Areas Commun.* **16**(8): 1530–1541 (Oct. 1998).

62. S. Verdú, Adaptive multiuser detection, *Proc. IEEE ISSSTA*, IEEE, July 1994, Vol. 1, pp. 43–50.

63. S. Verdú, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **32**(1): 85–96 (Jan. 1986).

64. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, Cambridge, UK, 1998.

65. G. Woodward and B. S. Vucetic, Adaptive detection for DS-CDMA, *Proc. IEEE* **86**(7): 1413–1434 (July 1998).

66. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multi-user communications, *IEEE J. Select. Areas Commun.* **8**: 683–690 (May 1990).

67. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multiuser communications, *IEEE J. Select. Areas Commun.* **8**(4): 683–690 (May 1990).

68. G. Xue, J. Weng, T. Le-Ngoc, and S. Tahar, Adaptive multistage parallel interference cancellation for CDMA, *IEEE J. Select. Areas Commun.* **17**(10): 1815–1827 (Oct. 1999).

69. C.-C. Yeh and J. R. Barry, Adaptive minimum bit-error rate equalization for binary signaling, *IEEE Trans. Commun.* **48**(7): 1226–1235 (July 2000).

70. L. J. Zhu, U. Madhow, and L. Galup, Differential MMSE: Two applications to interference suppression for direct-sequence CDMA, manuscript in preparation, 2000.

ADMISSION CONTROL IN WIRED NETWORKS

SYMEON PAPAVASSILIOU
JIE YANG
New Jersey Institute of Technology
University Heights
Newark, New Jersey

1. INTRODUCTION

Within the wired communications infrastructure, where switches and/or routers are deployed to support voice and data transmission, there are basically two key switching techniques: circuit switching and packet switching.[1] Traditionally, circuit switching is mostly used in telephone networks and packet switching is used in data networks including today's Internet. With the development of Broadband Integrated Services Digital Network (B-ISDN), where voice and various data services are provided in a common network infrastructure, packet switching has been widely deployed. Independent of the applied switching technique, to provide end-to-end connection and communication through the network, some network resources, such as link capacity, switching bandwidth, and buffers, are utilized. Since the amount of such resources are limited compared to the fast-increasing demand on voice and data communication, if at some point the requests on the resources exceed the available network resources, the network is considered as "congested." When the network is congested, either the connection can not go through or the quality of service (QoS) degrades. The technique to avoid congestion in a network is referred to as "congestion control." In general, there are two approaches to perform congestion control: (1) the *preventive approach*, where each connection reserves resources in advance, from which the idea of admission control stems; and (2) the *reactive approach*, where when congestion occurs flow control is performed via end-to-end closed-loop control mechanism or open-loop control mechanism at the intermediate network nodes.

Admission control is a more effective way to perform congestion control in high-speed networks because when congestion occurs, even though the network could react promptly, a large amount of data may be affected. Moreover, combined with resource allocation, admission control can reserve sufficient resources in advance for a connection so that its QoS can be guaranteed. Admission control can also be utilized as a mechanism to check and enforce policies before providing services to the users.

In traditional telephone networks, where circuit-switching is applied, congestion control is usually achieved via the first approach: admission control. Before users can talk to each other, a path from the sender to the receiver has to be set up and the required resources have to be reserved first. If many users want to use the telephone network at the same time and there are insufficient resources to be allocated, the admission control mechanism will turn down some new requests so that the admitted calls can be supported with satisfactory QoS.

In data networks including B-ISDN networks and the Internet, two scenarios may occur. If the network provides connection-oriented service, such as the ATM used in B-ISDN networks, usually both admission control and flow control are deployed. If the network provides connectionless service, such as today's best-effort Internet service, only flow control will be applied. However, newly proposed Internet service models that provide more services and better QoS also require admission control mechanisms.

2. OVERVIEW OF ADMISSION CONTROL

An admission control scheme basically consists of signaling messages and admission control units that perform the admission control algorithm or policy. The signaling is usually a part of the call setup/release signaling procedure. The connection to be set up could be either duplex or simplex. Three types of signaling may be involved in this procedure: (1) the signaling between the end user and the network access point, (2) the signaling inside the network, and (3) the signaling between networks.[2] The signaling between the user and network access point is used to send and respond to the call setup request. The signaling inside the network can be used to identify the current resources available for the new calls, inform the switches/routers to prepare for the new call, or inquire as to whether they are able to admit it. If the call needs to trespass several network domains, signaling between networks is required to check whether this new call can be supported by all the networks.

An admission control algorithm or policy is performed by the admission control units in the switch/router or some specified components/devices in the network. Admission control algorithms can be categorized into resource-based and policy-based. *Resource-based algorithms* base their decision on the current resource usage in the network, the resources needed to be allocated to the new call and whether the required QoS can be guaranteed by the network. Moreover, the already admitted calls shall not be affected if a new call is admitted.

Policy-based admission control is needed when different policies are enforced in the network. Its basic purpose is to determine whether a user is qualified to access the network service at a specific time.

3. ADMISSION CONTROL SCHEMES

In the following paragraphs we discuss admission control schemes in two types of networks: B-ISDN and Internet. The admission control for B-ISDN reflects the typical

[1] In today's wired telecommunication networks, especially telephone and data networks, another switching technique, namely the message switching is not widely used.

[2] The signaling protocols for the three cases do not have to be the same.

schemes of current connection-oriented networks that utilize common channel signaling. The schemes designed for the Internet reflect the tendency of future Internet services and models.

3.1. Admission Control in B-ISDN Networks

In B-ISDN networks where ATM is utilized as the transport technology, admission control procedure is implemented through signaling messages and admission control units of the switches. Two types of signaling are involved in the admission control procedure: the signaling at the user–network interface (UNI)[3] and the signaling at the network–network interface (NNI).[4] The ITU-T recommendation for UNI access signaling protocol is Q.2931 [1], which is a modified version of Q.931 — the access signaling protocol for ISDN. In ITU-T recommendations, the signaling protocol for NNI is B-ISUP [2], the B-ISDN user part of Signaling System 7 (SS7) — a widely used signaling protocol in today's telephone networks. ATM forum also defines Private Network-Network Interface (PNNI) protocol to address the issues of interswitch, internetwork, and routing operations, which are not specified in the ITU-T recommendations. In the following paragraphs we describe the signaling procedure that involves admission control based on ITU-T recommendations.

Generally, admission control is coupled with resource management that is involved in two procedures: call setup and call release. An example of a successful point-to-point call setup and release procedure is shown in Fig. 1. In the call setup procedure, the signaling messages SETUP, CALL PROCEEDING, ALERTING, CONNECT, and CONNECT ACK comply with the specification of Q.2931 and the signaling messages IAM, IAA, ACM, and ANM comply with B-ISUP.

When the switch at the UNI (switch A in the example) receives the SETUP message, which includes the calling and called party identities, such as the ongoing traffic descriptor, switch A will reply with CALL PROCEEDING

[3] UNI refers to the interface between the user and the network.
[4] NNI refers to the interfaces between the switches.

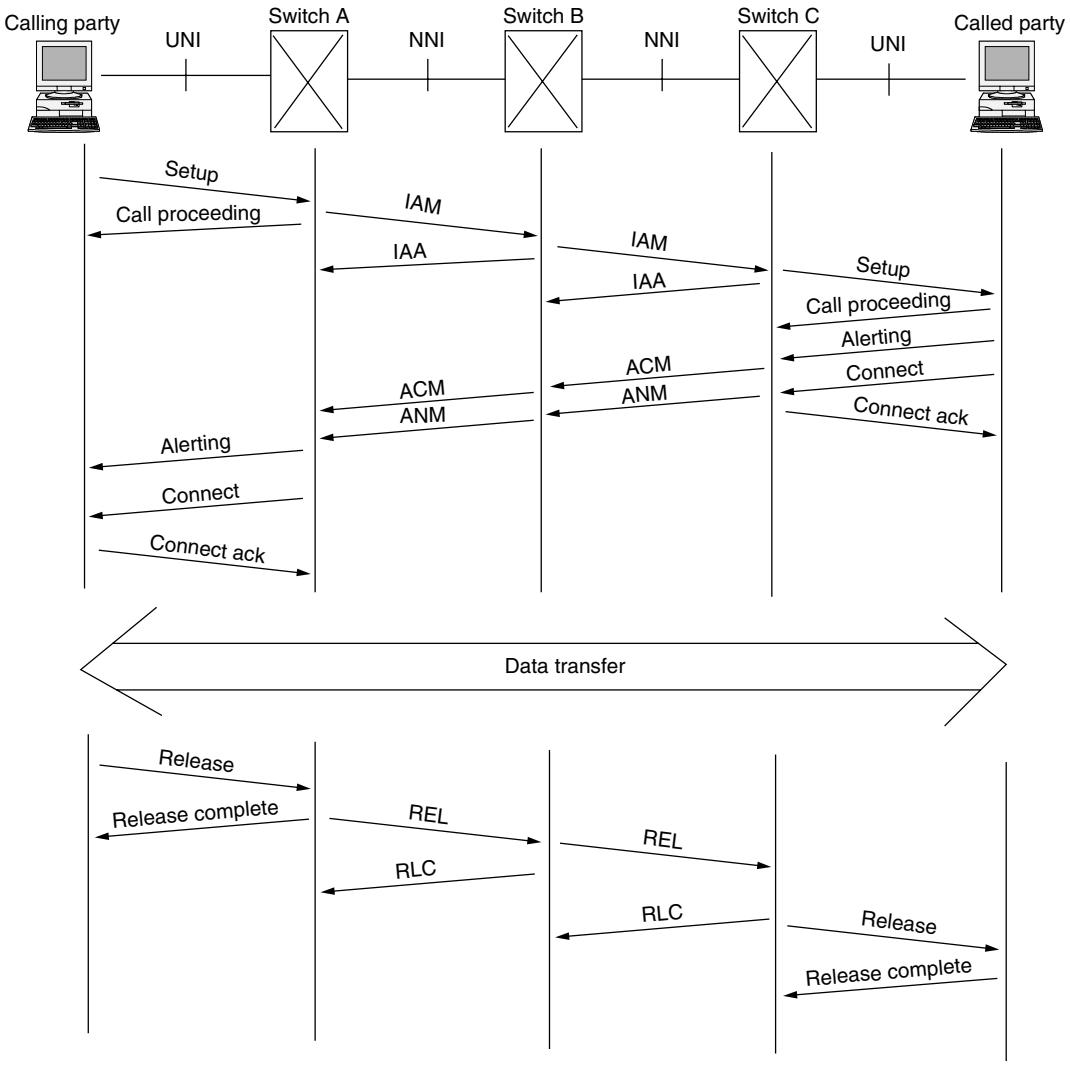


Figure 1. An example of successful point-to-point call setup and release procedures.

message if its admission control algorithm decides that this call can be accepted by the switch. This decision is made based on the identities of calling and called party, the route and resource availability, the traffic descriptor and the corresponding QoS requirements. If the route to the called party trespasses several switches, B-ISUP will be invoked. The switch will send an IAM (initial address message) to its next-hop switch (switch B in the example), which contains all the information in the call SETUP message. An IAA (IAM acknowledgment) will be sent back to switch A. If the next-hop switch can admit this call, and it is not the switch through which the called party accesses the network, it will forward the IAM to its next-hop switch (switch C). The same procedure will repeat until IAM reaches the switch connected to the called party, which is switch C in our example. If switch C is able to admit the call, it will issue a Q.2931 SETUP message through the UNI to the called party. The called party will reply in any of the following three ways: CALL PROCEEDING, ALERTING, and CONNECT if the call can be accepted. If the called party replies with CALL PROCEEDING, the network will wait for the following ALERTING or CONNECT messages from the called party. If the called party replies with ALERTING message, switch C will issue an address complete message (ACM) backward to switch A, where an ALERTING message will be issued to the calling party. Then the network will wait for the CONNECT message from the called party. If the called party replies with CONNECT message, switch C will issue an ANswer Message (ANM) backward to switch A, which indicates the connection is activated. Correspondingly, switch A will issue a CONNECT message to notify the calling party on the activation of the call. The calling party and switch C will also reply with a CONNECT ACK message to switch A and the called party respectively. Thereafter the data transfer phase can start.

The call can be terminated by either the users (calling party or called party) or the network. At the UNI, a RELEASE message will be initiated and a RELEASE COMPLETE message will be responded. Correspondingly, at NNI, a RELease (REL) message and a ReLease Complete (RLC) message will be exchanged between the switches. On receiving the Release message from UNI or REL message from NNI, a switch will release the resource reserved for the connection and these resources become available to the network.

During the setup procedure, if any switch detects that the call cannot be admitted, the call will be rejected. The switch will issue an IAM reject (IAR) message to its previous-hop switch from which it receives the IAM request. If no alternate route can be found, this IAR message will be sent backward until it reaches the edge of the network, where the switch will send a RELEASE message to the calling party.

To support point-to-multipoint connections, ATM Forum defines some supplementary messages for UNI [3]. It also defines PNNI protocol for Private NNI, which includes routing and signaling protocols across private ATM networks. For further details on signaling and its implementation, readers may refer to the corresponding standards and recommendations [1,2,4].

Admission control decisions are made and resource management is performed at each switch by its admission control unit, which can be a centralized module or a decentralized component deployed at each input or output module of the switch. The algorithm performed at the admission control unit is decided by the specific network administrators or equipment vendors and is not standardized. In Section 4 we briefly introduce and discuss some of the proposed approaches.

3.2. Admission Control on the Internet

Traditionally, the Internet only provides connectionless best-effort service; therefore, no admission control at the IP level is required. However, with the wide deployment of the Internet, many applications including many real-time applications may choose the TCP/IP protocol suite as their transport technology. These applications will require better QoS guarantees than what the best-effort service can provide. To address such a demand, new Internet service models have been proposed and are being developed that include the Integrated Service model (Intserv) [5] and the Differentiated service model (Diffserv) [6]. In the Intserv model, traffic is identified by a flow, a concept similar to the virtual connection in ATM; and a signaling protocol, namely, the resource ReSerVation Protocol (RSVP) [7] was proposed so that admission control and resource allocation can be performed in a dynamic manner. In the Diffserv model, traffic is aggregated and classified by service classes, which is significantly different from that in the Intserv model. In the Diffserv model services can be provided based on static or dynamic admission control and resource allocation.

3.2.1. Admission Control via RSVP in the Intserv Model. RSVP is used to setup a simplex path between two hosts. If duplex communication is required, two separate paths in each direction have to be setup via RSVP signaling procedure. A successful point-to-point path setup procedure is shown in Fig. 2. At the beginning of the procedure, the sender (host A) sends a PATH message to the receiver, which sets up a path from the source to the destination (downstream). The PATH message contains the "previous hop" information that the message has trespassed, the information about the sender, the traffic descriptor and the information about the status of the network. It installs a "path state" at each intermediate node. When the receiver (host D) receives the path message, it will send a RESV message back to the sender along the path that was set up by the PATH message (upstream). The RESV message makes resource reservation request at each node, which will decide whether to accept this request, based on both the availability of the resources and the policy enforced in the network. If the request can be admitted, a "reservation state" will be created at the node and the RESV message will continue to be forwarded along the path until it reaches the sender. Otherwise the request will be rejected and an error message will be sent to the receiver. Once the sender has received the RESV message, it can start its data transmission to the receiver. The path state and reservation state at the intermediate routers form a "soft state," which must be refreshed by PATH and

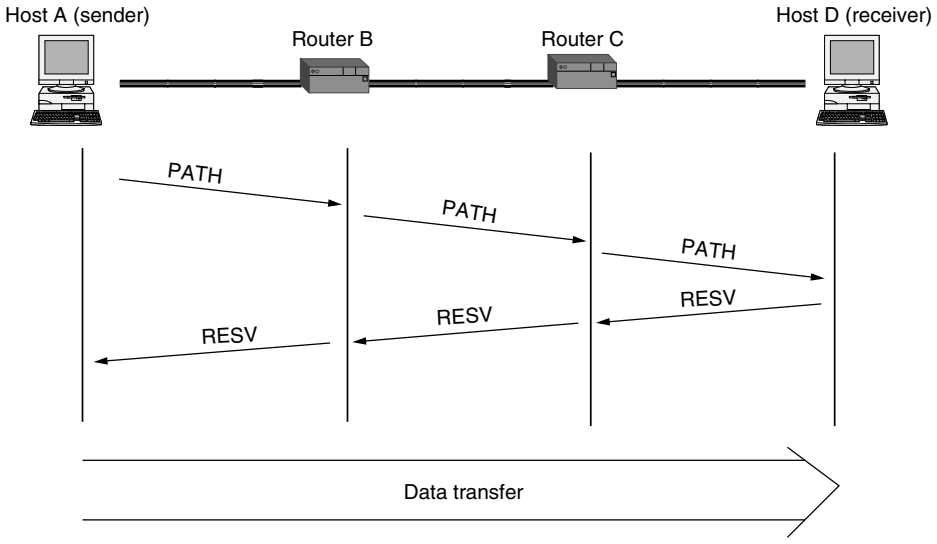


Figure 2. An example of successful point-to-point call setup via RSVP.

RESV messages periodically during data transmission. The state is timed out and deleted if no matching PATH or RESV messages arrive within the timeout interval. In this case the reservation is released. The state is also able to be torn down explicitly by the hosts (sender or receiver) when data transfer finishes. RSVP also supports multicast and flow aggregation.

The resource reservation request should be mapped to the link-layer technology where the resource-based admission control and resource allocation algorithms are actually deployed. The IETF workgroup ISSLL have developed some specifications on how to map an RSVP reservation request onto specific link layer technologies such as ATM [8] and Ethernet [9].

3.2.2. Admission Control in Diffserv Networks. The underlying principle of the Diffserv model is that service is provided to users based on the service-level agreement (SLA) between the user and the service provider. Traffic flows will be marked by the host or leaf router and classified, metered, shaped, and possibly re-marked at the ingress router of the Diffserv network where the flows will be aggregated according to the service class set by SLA and forwarded to the core routers. At core routers, the QoS is provided by the "Per hop behavior" (PHB) associated with service classes. SLA can be static, half-static or dynamic.

In static SLA, the service provided to the user is negotiated in advance and may be manually configured by the network administrator periodically. At the ingress node of the network, the flows of a user will be monitored to ensure its conformance to the SLA. Therefore, admission control can be performed in a static and implicit manner by which there is no signaling procedure to request resource reservation.

In half-static and dynamic SLA, there exists an agent in the Diffserv network, called "bandwidth broker" (BB), to manage the resources in its domain. In the half-static mode, resources have been preallocated to the users according to their SLA; however, it needs a signaling procedure to activate and install states at the border (ingress and egress) routers before transmission, and to deactivate and clear states after transmission. Usually the duration of such a state is in the time-scale of hours. In the dynamic mode, there are no preallocated resources to the users. The user has to use an explicit signaling protocol to request admission to the network. In both modes, there are two options to perform admission control and resource management. First, only boundary nodes are signaling-aware. The resource is managed through resource management agents, for example, BB. The interior routers are not signaling-aware. In this case, an admission request will be forwarded to the BB from the edge node by using signaling messages. The second option is that the interior routers are also signaling-aware and the signaling procedure is hop by hop similar to the procedure described in Intserv. However, the interior nodes will schedule and forward traffic only on the basis of the traffic class, while in the Intserv model traffic is scheduled and forwarded according to the flow specifications. The signaling protocol in Diffserv can be RSVP, extensions of RSVP or any other customized protocol.

3.2.3. Policy-Based Admission Control on the Internet. Policy-based admission control is still in its early stage. Generally, it is complementary to the resource-based admission control to resolve those issues not addressed by resource-based admission control, which may include priority of users and applications, security, or time-of-day traffic. RFC 2753 [10] provides a framework for policy-based admission control in which two basic network elements were proposed: policy enforcement point (PEP) and policy decision point (PDP). PEP is used to enforce policy for admission and PDP is the component that actually makes policy decisions. Usually PEP is deployed as a function unit in the edge routers and PDP is deployed as a centralized server in the network. An optional local decision point (LDP) can be deployed together with PEP to provide local policy decision. PDP may also need to access other servers in the network to reach a policy decision. A typical configuration of a QoS-enabled network with policy-based admission control capability is shown in

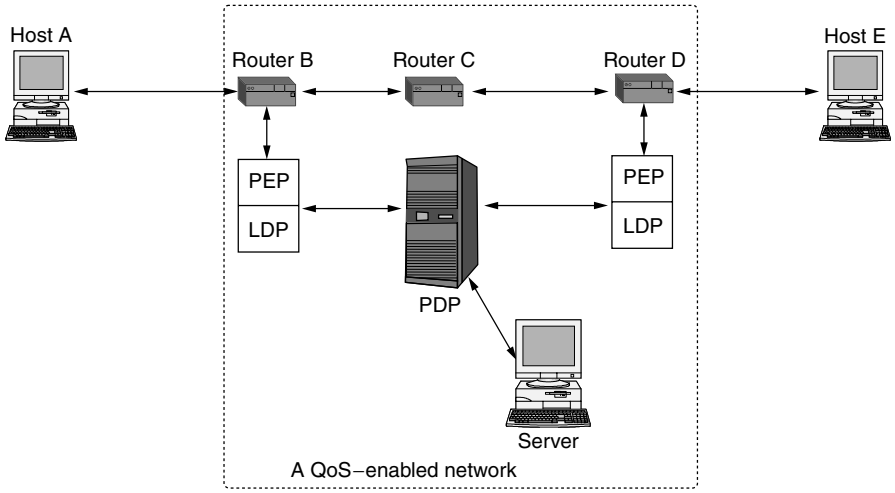


Figure 3. A typical configuration of QoS-enabled network with policy-based admission control.

Fig. 3, which can be applied to both Intserv networks and Diffserv networks.

In a network that enables both resource-based and policy-based admission control, when a signaling message reaches the edge router requesting admission to the network and reserving resources, the router will first check whether there are sufficient resources for the request through the local resource management unit or through the remote resource management components. Then it will check through PEP whether there are policies to be enforced to the request. The PEP will inquire LDP and PDP as to possible policy decisions. This can be referred to as a dynamic mode. PDP and LDP are also able to issue policies to be enforced at PEP simply according to the conditions of the network or the policies of the service provider, which can be referred to as a static mode. Since PDP is usually deployed at a remote server, a protocol between PEP and PDP is needed for them to communicate. A candidate for such a protocol is COPS [11]. COPS employs a client/server model and uses TCP as its transport protocol for reliable communication between PEP and PDP. A PEP can send its request through COPS REQ message and receive decision from PDP through COPS DEC message. An RSVP signaling message is also able to carry policy requests so that end-to-end policies can be enforced. The extension of RSVP to support policy-based admission control is proposed in RFC 2750 [12], in which a new data object, policy data object, is defined for this purpose.

3.2.4. MPLS and Its Admission Control. In connectionless networks such as today's Internet, each router analyzes the network layer header of a packet and makes its routing and forwarding decision independently, which is referred to as "hop by hop." However, a router will handle packets with the same forwarding equivalence class (FEC) in the same manner and these packets are actually indistinguishable to the router [13]. Moreover, it is believed that the network layer header provides much more information than necessary to perform routing and forwarding [13], which also makes it difficult to achieve fast analysis of the header and makes routing of packets a bottleneck in high-performance routers. Therefore, multiprotocol label switching (MPLS) was proposed, in which a small fixed label was bound to the packet to locally identify FEC. Similar to an ATM virtual circuit, a label-switched path (LSP) has to be set up so that the adjacent label−switching routers (LSRs) along the LSP can forward packets by simply looking into the locally bound labels without having to analyze the entire network layer header. A label distribution protocol (LDP) is used by an LSR to inform the other LSR about the label binding to a specific FEC. During the label assignment and distribution procedure, attributes of the FEC can be set up so that QoS and policies of the FEC can be specified. Since an LSP is similar to an ATM virtual circuit, admission control is performed during the LSP setup and LDP is used as the signaling. Although Rosen [13] does not specify the LDP scheme, it requires that label binding is determined by the downstream LSR. MPLS is a technology between layers 2 and 3 (referred to as layer 2.5 technology), whose resource management and admission control algorithm depend on the underlying layer 2 technology.

4. ADMISSION CONTROL ALGORITHMS

As can be seen from the previous discussion, resource-based admission control algorithms are closely coupled with the problem of resource management. Generally, a resource-based admission control algorithm consists of two parts: estimation of resource usage and QoS performance, and optimized decisionmaking on whether to accept the connection request. Resource-based admission control algorithms have been under intensive research and development efforts since the emergence of ATM. This is due to several reasons:

1. Data networks, including ATM and the Internet, support and encourage statistical multiplexing so that the resource utilization can be enhanced, which, however, increases the algorithm's complexity. The estimation of resource usage must be accurate so that the QoS performance will not be harmed as a result of the statistical multiplexing. This, in turn, requires the development and use of a good model for

estimation of resource usage and QoS performance, and an optimized solution to it.

2. There are many different services provided by the data networks that have quite different statistical nature and QoS requirements. This complicates the modeling of the problem and makes the optimization of the solution quite difficult.
3. Admission control must be performed in real time. Therefore, the algorithm needs to be accurate on one hand, and simple enough on the other hand. However, these two requirements are sometimes contradictory to each other.

Compared to resource-based admission control, policy-based admission control is an approach that manages the network resources in a relatively static manner. A connection will be accepted if it matches the policies of the network. When a connection falls into several policies, the network node needs to find which policy fits the connection best and basis its admission decision on that policy.

4.1. Model-Based Admission Control

Resource-based admission control algorithms can be further divided into model-based algorithms and measurement-based algorithms. Some traffic descriptors are used to model the traffic and measure the performance in both types of algorithms. In ATM, these descriptors include peak cell rate, sustainable cell rate (SCR), and maximum burst size (MBS) [3]. Two approaches have been adopted in model-based algorithms: deterministic approach and statistical approach.

The *deterministic* approach allocates resources simply according to peak data rate while assuming no data loss. This approach was adopted by telephone networks and CBR traffic in ATM. The advantage of the deterministic approach is that it is easy to be implemented and can be performed in real time. However, although QoS of each connection will be guaranteed, the resources, such as bandwidth and buffer allocated to each connection, cannot be shared, which in turn may lead to underutilization of resources because usually arriving traffic in data networks is bursty in nature rather than of constant rate; therefore a lot of resources will be wasted. For ATM networks, various statistical approaches have been proposed that take the statistical nature of the traffic into account, so that resources can be shared among different connections efficiently.

4.1.1. Traffic Models. The basic objective of the statistical approach is to accept as many connections as possible so that resources can be efficiently utilized while the QoS of each connection is still guaranteed. To achieve this objective, the arriving traffic needs to be modeled accurately so that resources can be allocated properly. It has been found that traffic from many applications such as voice and video present bursty characteristics that can be described by various "ON/OFF" models. These ON/OFF models consist of two states: a busy state in which data are transmitted from the traffic source, and an idle state in which no data are transmitted from the source. Various on-off models differ from each other in the distribution of the duration at each state, and the arrival process in the busy state. Commonly used ON/OFF models for a virtual connection in ATM and B-ISDN include Interrupted Poisson Process (IPP), Interrupted Bernoulli Process (IBP), and Interrupted Fluid Process (IFP). Further studies on ATM traffic introduce more complex distributions into the modeling such as Markov modulated poisson process (MMPP), Markov modulated bernoulli process (MMBP), and Markov modulated fluid process, each of which consists of several different states with state-dependent arrival rates.

Through measurement on Ethernet traffic, it has been recognized that IP traffic presents the characteristic of self-similarity which can be modeled by fractional Gaussian noise and fractional autoregressive integrated moving-average processes [14]. The self-similar traffic can be obtained by superposition of many on-off sources whose ON and OFF states strictly alternate and have high variability [15].

According to the various traffic model assumptions, such as the ON/OFF traffic model or its Gaussian approximation on traffic aggregation, many model-based admission control algorithms have been proposed that can be categorized into single-link approach and multiple-link approach.

4.1.2. Single-Link Approach. The single-link approach studies the admission control problem on a single link, specifically, the output of a multiplexor at the output port of a switch. The objective is to optimize or maximize the utilization on the link. By using traffic descriptors and traffic models, an admission region can be calculated assuming a target QoS performance [e.g., cell loss probability (CLP)].

A direct approach is to try to find the relationship between the arrival pattern of incoming traffic and CLP. As an example, an upper bound of CLP that is based on the average cell rate and peak cell rate or the rate variance in a fixed interval can be obtained [16]. A new connection is admitted if the resulting upper bound is below a pre-defined threshold.

Since traffic rate fluctuates between the minimum rate and peak rate, equivalent capacity has been proposed to describe the bandwidth needed to accept N connections for a given CLP threshold. In other words, this problem can be rephrased as, given a link capacity C, a predefined CLP threshold ε, and a buffer length K, how many connections can be accepted. For instance, if the connection requests are homogeneous with same SCR and PCR, by using equivalent capacity, the number of connections that can be accepted is between C/PCR and C/SCR depending on the value of ε. Admission control via equivalent capacity of a single service class and multiple service classes has been extensively studied. For a detailed review and comparison of single-link admission control algorithms interested readers may refer to Refs. 17 and 18.

4.1.3. Multiple-Link Approach. From the perspective of the service provider, the concern is how to efficiently utilize the resources not only on a single link but also in the entire network. For instance, when a specific link to a destination node can no longer accept connections, the connection may still be accepted by carefully rerouting the

connection through another path to the same destination node. Therefore, the optimization objective function should also take into account the routing algorithms. Further consideration of this problem leads to the observation that in the multiple service networks, the optimization objective could be the maximization of network revenue or gain, with multiple constraints such as routing, QoS, and policies. Solutions to the problem include decomposition of the multiple-link problem into single-link problems and the use of various techniques such as neurodynamic programming and reinforcement learning.

4.2. Measurement-Based Admission Control

Model-based admission control algorithms rely heavily on the corresponding traffic model assumptions to achieve the desired objective. However, this type of approach presents several problems:

1. The traffic model may not be sufficiently accurate, which may lead to the problem that the admission region is not properly designed. Although more accurate traffic models can be found by in-depth study of the traffic pattern, it may be too complex to be incorporated into the admission control model, or make the admission control algorithms too complicated to be performed in real time.
2. A user may overestimate or underestimate the traffic it will generate, which leads to an inefficient admission decision made by the network node.
3. Model-based admission control is usually a conservative scheme, which allocates resources based on worst-case scenario and may result in waste of the resources. Therefore, more recently measurement-based admission control algorithms have been extensively studied, most of which differ in three aspects: (a) the objective of measurement, (b) the approach to estimate and evaluate the measurement, and (c) the approach to make the admission decision.

Two types of objectives are used to perform measurement: to evaluate the CLP and to evaluate the equivalent bandwidth of aggregate traffic flows. Compared to the evaluation of CLP, measurement to estimate the equivalent bandwidth requires less computational power and is more straight-forward. The basic idea of this approach is as follows. The current equivalent bandwidth of aggregate traffic flows is estimated through measurements. Combined with the parameters declared in the new connection request, the equivalent bandwidth required is calculated assuming that the new connection is accepted. If the result is less than the bandwidth of the link, the new connection can be accepted; otherwise it is rejected. In this approach, CLP or CLR (cell loss ratio) is implicitly taken into account when estimating and evaluating the equivalent bandwidth.

An advantage of measurement of equivalent bandwidth or flow aggregation is that this approach does not need to track per-flow information of existing connections. Therefore, it can be applied not only in ATM but also in Internet where scalability of the algorithm is a big concern. For example, Cetinkaya et al. [19] proposed an admission control architecture in which an admission control algorithm is performed only at egress routers where the aggregate traffic envelopes are measured and estimated by using an adaptive algorithm. In this scheme, there is no involvement of backbone routers or per-flow management in the admission control procedure. As a result, the algorithm can achieve a good scalability.

To further address the scalability concern, a new approach based on measurement was proposed for the Internet. This approach is significantly different from conventional schemes in that it does not use signaling messages to make connection requests and the network node is not responsible for admission decisions any more. It is the host or end system that actually makes the decision as to whether to access the network. This is achieved by sending probe packets through the network to check the congestion level. If the probe packet indicates the current congestion level is low and will not harm the QoS of the existing and new connections, the host will admit the new flow; otherwise it will hold and give up the connection request. This approach is referred to as *endpoint admission control* or *distributed admission control* [20].

5. CONCLUDING REMARKS

In this article we discussed the problem of admission control in wired networks, with reference to the most current and future key networking infrastructures. The main objective of admission control in wired networks is to control the access of the users to the network resources in order to maximize the network utilization while providing the required quality of service and avoiding the occurrences of network congestion. An admission control scheme consists mainly of signaling messages and admission control units that perform the admission control algorithm or policy. In this article we introduced several typical admission control schemes and signaling procedures that can be applied in connection-oriented networks and on the next-generation Internet. We also presented some admission control algorithms that can be used in the admission control schemes, and discussed their characteristics. In general the admission control algorithms and schemes can be categorized into resource-based and policy-based admission control. *Resource-based admission control* algorithms base their decisions on the current resource usage, the resources needed to be allocated to the new calls, and whether the required quality of service can be guaranteed by the network. The basic purpose of *policy-based control* is to determine whether a user is qualified to access the network service at a specific time, and are required when different policies are enforced in the network. The implementation of admission control for Internet is still under development and deployment. As communication infrastructures are evolving into multiplexed and multiple service-class networks, network resource sharing by multiple service-classes correlates the performances of all classes that are supported in logically partitioned networks, and therefore additional scalable, less complicated and of high-efficiency admission control approaches and architectures are required.

BIOGRAPHIES

Symeon Papavassiliou received a diploma in electrical engineering from the National Technical University of Athens, Greece, in 1990 and his M.Sc. and Ph.D. degrees in electrical engineering from Polytechnic University, Brooklyn, New York in 1992 and 1995, respectively. From 1995 to 1996 Dr. Papavassiliou was a technical staff member at AT&T Bell Laboratories in Holmdel, New Jersey, and from 1996 to August 1999 he was a senior technical staff member at AT&T Laboratories in Middletown, New Jersey. From June 1996 till August 1999 he was also an adjunct professor at the Electrical Engineering Department of Polytechnic University, Brooklyn, New York. Since August 1999, he has been an assistant professor at the Electrical and Computer Engineering Department of New Jersey Institute of Technology, Newark, New Jersey. Dr. Papavassiliou was awarded the Best Paper Award in INFOCOM'94 and the AT&T Division Recognition and Achievement Award in 1997. Dr. Papavassiliou has an established record of publications in his filed of expertise, he is the director of the Broadband, Mobile, and Wireless Networking Laboratory at NJIT, and one of the founding members of the New Jersey Center for Wireless Networking and Security (NJWINS). His main research interests lie in the areas of computer and communication networks with emphasis on wireless communications and high-speed networks, network design and management, TCP/IP and internetworking, computer network modeling and performance evaluation and optimization of stochastic systems.

Jie Yang received a B.S. degree in information engineering, and an M.S. degree in communication and information system from Xidian University, P.R.China, in 1996 and 1999, respectively. He is currently a Ph.D. candidate in electrical engineering at the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, and a research assistant in the Broadband, Mobile and Wireless Networking Laboratory at NJIT, as well as a member of the New Jersey Center for Wireless Networking and Internet Security. From September 1999 till December 2001 he was a member of the New Jersey Center for Multimedia Research where he had been working on the design of high-speed networks. His current research interests are high-speed switch/router architectures, admission control, resource allocation and traffic engineering, and Internet security.

BIBLIOGRAPHY

1. ITU-T Recommendation Q.2931, *Digital Subscriber Signalling System No. 2 (DSS 2)—User-Network Interface (UNI) Layer 3 Specification for Basic Call/Connection Control*, 1995.
2. ITU-T Recommendation Q.2761, *Functional Description of the B-ISDN User Part (B-ISUP) of Signaling System No. 7*, 1999.
3. ATM Forum, *ATM User-Network Interface Specification V3.1*, 1994.
4. ATM Forum, *Private Network-Network Interface Specification Version 1.0 (PNNI 1.0)*, March 1996.
5. R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, June 1994.
6. S. Blake et al., *An Architecture for Differentiated Services*, RFC 2475, Dec. 1998.
7. R. Braden et al., *Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification*, RFC 2205, Sept. 1997.
8. E. Crawley et al., *A Framework for Integrated Services and RSVP over ATM*, RFC 2382, Aug. 1998.
9. R. Yavatkar et al., *SBM (Subnet Bandwidth Manager): A Protocol for RSVP-Based Admission Control over IEEE 802-Style Networks*, RFC 2814, May 2000.
10. R. Yavatkar, D. Pendarakis, and R. Guerin, *A Framework for Policy-Based Admission Control*, RFC 2753, Jan. 2000.
11. D. Durham et al., *The COPS (Common Open Policy Service) Protocol*, RFC 2748, Jan. 2000.
12. S. Herzog, *RSVP Extensions for Policy Control*, RFC 2750, Jan. 2000.
13. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, Jan. 2001.
14. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (Extended version), *IEEE/ACM Trans. Network.* **2**(1): 1–15 (Feb. 1994).
15. W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level, *IEEE/ACM Trans. Network.* **5**(1): 71–86 (Feb. 1997).
16. H. Saito, Call admission control in an ATM network Using upper bound of cell loss probability, *IEEE Trans. Commun.* **9**(40): 1512–1521 (Sept. 1992).
17. H. G. Perros and K. M. Elsayed, Call admission control schemes: A review, *IEEE Commun. Mag.* 82–91 (Nov. 1996).
18. E. W. Knightly and N. B. Shroff, Admission control for statistical QoS: Theory and practice, *IEEE Network* 20–29 (March/April 1999).
19. C. Cetinkaya, V. Kanodia, and E. W. Knightly, Scalable services via egress admission control, *IEEE Trans. Multimedia* **3**(1): 69–81 (March 2001).
20. F. P. Kelly, P. B. Key, and S. Zachary, Distributed admission control, *IEEE J. Select. Areas Commun.* **18**(12): 2617–2628 (Dec. 2000).

ADMISSION CONTROL IN WIRELESS NETWORKS

SYMEON PAPAVASSILIOU
JIONGKUAN HOU
New Jersey Institute of Technology
University Heights
Newark, New Jersey

SEBNEM OZER
MeshNetworks, Inc.
Orlando, Florida

1. INTRODUCTION

The goal of wireless communications is to provide users with ubiquitous information access, that is, to allow the

Table 1. Service Classification

	Real-Time		Nonreal-Time	
	Conventional	Streaming	Interactive	Background
Examples	Voice	Videostreaming	Web browsing	Email
Delay	Bounded		Sensitive	Tolerable
Rate	Guaranteed		Not guaranteed	
BER	$\leq 10^{-3}$	$\leq 10^{-6}$	≈ 0	

users to access the capabilities and resources of the global network at any time without regard to their location and mobility. Technological advances such as time and space diversity systems, low noise filters, efficient equalizers, advanced modulation and coding schemes, and the rapid development of handheld wireless terminals have facilitated the rapid growth of wireless communications and mobile computing.

Compared with fixed networks the most salient features of the wireless networks include the users' mobility, the limited bandwidth and power resources, the highly dynamic network (re)configuration, and the higher link bit error rates. In cellular wireless networks, due to the mobility of wireless subscribers, the network configuration is rearranged every time a subscriber moves into the coverage region (cell) of a base station or a new network. Furthermore, for current and future wireless networks designed to support high-data-rate applications, among the major limitations are the propagation conditions, such as fading and multipath, and power consumption that determine the communication range. In general, the tradeoff is between creating a dense infrastructure with high handoff rate and a more sparse deployment at the expense of high power consumption. Therefore in order to maximize the utilization efficiency of the limited radio resources, while meeting the quality-of-service (QoS) requirements of mobile users, efficient admission control and resource management schemes are required in the emerging wireless network architectures.

Wireless network management services can be categorized as call management, radio resource management, and mobility management [1]. Call management, for setting up and terminating communication sessions is necessary for both conventional information networks and the wireless communication systems. The other two categories of network management tasks are new and specific to wireless communications. The distinctive features in terms of mobility, traffic patterns and QoS requirements, and availability of the limited resource are the factors that govern admission control in wireless networks.

QoS is the ability of a network element to provide some level of assurance that its traffic and service requirements can be satisfied. For admission control depending on the QoS requirements, two main service classes can be considered: real-time and non-real-time services. Each class can be further divided into subclasses corresponding to different applications with given traffic characteristics and QoS parameters. For instance, in UMTS (Universal Mobile Telecommunications Systems), the supported traffic types are divided into four different service classes such as conventional class, streaming class, interactive and background class. Table 1 summarizes this classification and highlights the respective traffic characteristics of each class.

Features specific to the mobile environment such as the particular problems of highly variable connection quality, management of data location, the restrictions of battery life and cost, all impact the delivery of the required QoS. Therefore, admission control mechanisms combined with effective resource allocation schemes are crucial for the efficient design and use of wireless systems.

2. OVERVIEW OF ADMISSION CONTROL

The wireless network management techniques perform various processes such as power control, channel allocation, and handoff. The call and radio resource management problem is to assign, at different timescales, to each terminal, a base station, a physical channel and transmitter power levels, for both uplink (from mobile terminal to base station) and downlink (from base station to mobile terminal) communication. Depending on the access technology, the channels may take the form of time slots in time-division multiple access (TDMA) systems, or frequency bands in frequency-division multiple access (FDMA) systems, or different codes in code-division multiple access (CDMA) systems. According to these management techniques, admission control can be implemented at call (circuit switching), packet (packet switching) or burst (burst switching) level, or a common access scheme (random/common packet access) may be deployed. As mentioned before due to the user mobility, a user might move across the cell boundary while a call is in progress. In this case the system automatically transfers the call to a new channel belonging to the new base station. This process is called handoff or handover. In order to provide uninterrupted service to the mobile subscribers, handoffs must be performed successfully and should be imperceptible to the users. Users may have to change their radio cells a number of times during the lifetime of their connections, and as a result, availability of wireless network resources at the connection setup time does not necessarily guarantee that wireless network resources will be available throughout the whole lifetime of the connection. Thus the handoff events make the call admission control process for wireless networks more complicated than those for wired networks.

More specifically the admission control steps in wireless networks can be summarized as follows [2]: (1) assign one or more (e.g., soft handoff) base stations for a new or handoff call if the call has been accepted; (2) assign one

or more channels (e.g., frequency, time slots and codes or a combination of them) according to rate requirements; (3) assign transmitting powers for the base station and mobile nodes (power levels are adjusted according to the channel conditions, user location and required QoS); and (4) allocate resources according to the traffic classes (a time scheduler decides when to use the allocated resources).

The resources are estimated from the measured interference conditions, radio channel characteristics, current load in the cell site, sessions' traffic characteristics, and quality of service requirements. These inputs along with historical values and capacity models are used for the admission control tasks. An important consideration in specifying these functions is the interplay of the granularity of their response, and the load they create on the system. Too little monitoring may cause out of specification performance for a period of time while these measurement and management functions themselves will place a load on the systems they are monitoring. Another measurement required for wireless systems is done in order to estimate the link quality for highly variable air interface and user mobility. Bursty data transmission poses a new problem as the link quality cannot be measured efficiently at long idle times where the distance between transmissions can be considerably changed. A tradeoff between using estimated average link qualities versus keeping the link alive at a minimum idle power level must be taken into account for different ratios of idle rate to mobility rate.

In the following we examine in more detail each one of the main elements involved in the admission control process in wireless networks.

3. POWER CONTROL

The objective of power control is to deliver to each radio receiver a signal that is strong enough to overcome noise and interference from other signals but not so strong as to cause excessive interference to other communications. In general power control guards against changes in the system load, jamming, slow and fast variations in the channel conditions, and sudden improvements or degradations in the links. The gain from power control can be seen in conserving energy for prolonged power supply, in satisfying stable QoS for multimedia services, in efficient handling of mobility (handoffs), in increasing overall capacity, and in other applications.

The carrier-to-interference ratio (CIR) [or signal-to-interference ratio (SIR)] balancing technique for power control purposes has been presented in several early power control schemes [3,4]. The power control algorithms in the literature can be classified as distributed and centralized algorithms. For mainly practical considerations, most efforts have concentrated more on distributed power control schemes than on centralized schemes, because the centralized power control suffers from problems such as large-scale data management, complexity, network vulnerability, and latency, etc.

In any of these algorithms either the path gains are assumed to be known a priori, or measured SIRs on the active links are utilized. In general two main distributed power control approaches have been proposed. In the first approach, the receiver's signal-to-interference ratio (SIR) is measured and the transmission power is adjusted according to whether the SIR is below or above some target value. The drawback of this algorithm is that the local adjustments, without a global consistency, increase the interference to the neighboring areas, which, in turn, results in an increase of power in this area, and finally in an increase of interference. In the second approach, the transmitted power is adjusted in order to balance the SIRs of all links and to maximize the worst SIR in the channel. The drawback of this approach is that during the iterations of power adjustment or after reaching the steady state, the SIRs of the links may fall below the required value.

Besides the ability to compute capacity margin, the desired properties of a power control scheme are stated [5] as to be distributed (at the node or link level) in order to require minimal usage of network resources for control signaling, simple to be suitable for real-time implementation, agile for fast tracking and adaption to the channel changes and mobility, robust to be able to adapt to stressful contingencies, and scalable to perform at various network scales of interest. In predictive call admission control, a predictor is used to predict the future traffic from its present and past values. If the call setup time is longer relative to the traffic variations, the advantage of the predictive algorithm is higher.

Power control in code division multiple access (CDMA) cellular systems is a crucial issue since the capacity of CDMA networks is mainly interference limited [6,7]. Present CDMA cellular systems have been optimized for voice transmission. For voice CDMA systems based on the Interim Standard (IS95) standard, power control is used to combat the near–far problem by maintaining nearly constant received power at the base station. Power control is used as a means of minimizing multiuser interference and improving capacity by adjusting the powers to obtain the same carrier to interference power ratio on all links.

In the IS95 reverse link (uplink), the signal from each mobile unit should arrive at the base station with the minimum signal-to-noise ratio (SNR) needed to maintain the desired quality. In reverse-link–open-loop control, the mobile unit estimates the path loss from the cell site by comparing the received power to the transmitted power. Then the mobile adjusts its power such that the transmitted power is lowered if the signal is determined to be too strong or is increased slightly otherwise. In reverse-link–closed-loop control, the demodulator at each cell site compares the received SNR to the desired value and commands the appropriate adjustments. In the forward (downlink) link, at certain locations, the signal received by a mobile unit may be too weak to accurately decode data due to the excessive shadowing and interference from a neighboring cell. The cell periodically reduces the transmitted power in order not to transmit high power if not necessary. When a mobile detects an increase in its frame error rate, it requests higher power and the cell site increases the power by a predetermined amount.

Several power control algorithms have been proposed to address the problem of admission control in a DS-CDMA (direct-sequence code-division multiple access) network

with integrated services [8,9]. The main objective is to achieve optimality in the sense of maintaining active link quality (QoS of active users) while maximizing free capacity of new admissions. Bursty packet applications can introduce high interference during active periods. Multiaccess interference is regulated by controlling the transmit powers of the users for active link quality protection. This is done by computing the "interference margin," that is, the amount of excess interference that can be tolerated by active users without violating their SNR thresholds.

In many systems, transmission power control and channel allocations according to the traffic classes are managed jointly in order to maintain the SIR's of all links above the required quality factor at all times. For instance, the conventional power control scheme can be used with one power level for each slot if packets with equal or similar bit error rate (BER) requirements are transmitted in the same slots [10]. Otherwise, an optimal power distribution can be computed to provide the required BER of media with high priority and achieve the maximum throughput and minimum BER of media with low priority [11].

4. CHANNEL ALLOCATION AND ADMISSION CONTROL

In general the channel allocation problem can be viewed as a combinatorial optimization problem. Frequency-division multiplexing and time-division multiplexing provide a "channelization" of the spectrum. In code-division multiplexing schemes waveform allocations are permuted in a random fashion [12]. Depending on the dedication of bandwidth, we can categorize the channel allocation schemes into four sets as follows: dedication of channels for call duration (circuit switching), dedication of channels for packet duration (packet switching), dedication of channels for the duration of burst data (burst switching), and random access transmission (common channel packet switching) that do not require a reservation of channels. While circuit switching is a fixed dedicated assignment, packet and burst switching are demand-based assignments. For each one of these sets and switching schemes, different channel allocation and admission control processes are required.

4.1. Circuit Switching

In circuit switching, the users are allocated a dedicated channel and a continuous connection is guaranteed during a session. Hence, circuit switching is a static admission control where the negotiation is done for the call duration in the specific cell. The steps are specification of QoS requirements, negotiation for an agreed specification between all parties, admission control for prediction of the capability to meet the users' requirements, and resource reservation for allocation of resources to connections. These functions are supported by a database of multimedia documents that has information such as historical characteristics of a link, a profile manager that maintains QoS-related information for different classes of users, and a network monitor that monitors the system's state at the new and handoff call arrivals.

The user first sends a request message containing the information for the specification of the QoS requirements. The cell site decides to accept or reject the user according to the active users QoS requirements. If feasible power and rate vectors are found, the new user is accepted to the system. A power vector is feasible if for every active link, a positive transmission power level smaller than the peak transmit power can be found. Similarly, a rate vector is feasible if for every active link a rate level greater than the minimum required rate can be assigned. If the user is accepted, an acknowledgment message with the assigned channel and the required power level is send to the mobile. Each session arrival is either allocated to a dedicated channel or blocked. If a negative acknowledgment is received or no response is received within a predetermined time interval, the user resends the request message after a random delay. A blocked user is lost if the waiting time exceeds the tolerance time or a maximum number of access attempts is reached.

4.2. Packet Switching

Dynamic admission control is more effective for bursty multimedia data and highly variable wireless channel conditions. Its corresponding functions include: monitoring of the QoS parameters, policing for ensuring that all parties satisfy the QoS contracts, maintenance of QoS by modification of some network parameters, renegotiation of a contract, and adaptation to the changes in the system. Depending on whether the renegotiation is done on a packet or burst basis, the packet and burst switching techniques are performed. In packet switching, data terminals must contend for a channel for each packet that must be sent. Therefore the network utilization is maximized while access delay per packet is increased.

4.3. Burst Switching

Circuit and packet switching techniques are insufficient in meeting the quality of service (QoS) requirements of bursty long multimedia messages due to the poor channel utilization and high per-packet delay, respectively. For instance, the proposed burst switching technique in cdma2000 MAC layer [13] attempts to overcome these problems by allocating the dedicated channels to the burst of data and releasing them at the end of the bursts. This ideal burst switching system would immediately release the circuit at the beginning of the idle period following the packet burst, so that the allocation delay constraints would be satisfied, while the channel utilization is maximized [14].

Since the traffic channels are allocated for the duration of a burst, admission control at burst level is considered. Admission control must be dependent not only on the active users but also on the registered users in the inactive state (since they can reaccess the system), in order to foresee the potential to admit a new user. Depending on the traffic and control channels allocation and registration process, a terminal can be in different states where the state transitions are controlled by the base station via "timer" values. The optimal timer that determines the burst length depends on the user traffic characteristics, the timescale of interest, the system load, and the corresponding QoS requirements.

4.4. Common Channel Packet Switching

For short bursty messages, the exchange of resource allocation control information can be avoided by using ALOHA-type random-access methods where terminals compete for radio resources. This approach requires the resolution of collisions and the use of retransmission policies.

The common packet channel (CPCH) mechanism has been shown to be an efficient transfer mechanism of packet data in wireless environments for non-real time applications such as email, HTTP, and FTP [15]. CPCH message transmission typically operates in power controlled CDMA systems. Each message can have variable length where the maximum length is a higher-layer parameter. Since error control via acknowledgments and retransmission in non-real time applications is crucial, especially in the environments where message losses are usually higher, a retransmission scheme is used. The additional delays caused by retransmissions are likely to be tolerable in applications with less stringent delay bounds, while the loss of some of the messages is often intolerable since completeness of information delivery is essential.

There is no guaranteed QoS during the packet transmission. A positive or negative acknowledgment is sent to the user after the reception of each packet according to the packet error at the receiver side. If negative acknowledgment is received or no response is received within a predetermined timeout value, the user retransmits the packet after a random delay. Acknowledgments can be send for each packet or for a burst of packets.

In general, the advantage of burst reservation schemes for data services is minimization of the interference for voice and data packets at the expense of higher overhead to control and measure the channel load. On the other hand, ALOHA-type common packet transmission requires a higher rate of retransmission for data users while a simpler control mechanism is needed.

4.5. Hybrid Schemes

Hybrid channel assignment schemes are used for integrated voice/data services where voice traffic is transmitted in a circuit mode while data traffic is transmitted in packet/burst mode or on a common channel. In this case, data users can use the voice channels during OFF time of voice users without degrading the QoS of the voice users. Some proposed channel assignment techniques are described below.

For Dynamic TDMA/TDD mode, users send transmission requests to the base station that processes them with a schedule table based on the QoS parameters of user traffic. For constant-bit-rate (CBR) and variable-bit-rate (VBR) traffic, slot allocation is performed once during call establishment, as in circuit mode. For available-bit-rate (ABR) and unspecified-bit-rate (UBR) traffic, slot allocation is performed on a burst-by-burst basis via dynamic reservation of ABR/UBR slots and unused CBR/VBR slots as in burst switching.

In packet reservation multiple access (PRMA), each of the slots are classified as being either reserved or available. Reservations are limited to terminals transmitting real-time data such as voice or video. Data terminals must contend for a time slot for each packet that must be sent as in packet switching. Speech activity detectors are used to hold reservations only for the duration of the talk spurt and to release them during quiet spurts so that the channel bandwidth can be used by other terminals with packets to send. One drawback of this algorithm is that while voice terminals are able to minimize collisions by reserving slots, terminals must still contend for initial access and data packets must contend for each slot. Furthermore, the permission probability of the different terminals must be controlled so that terminals with time-sensitive data are able to access the channel without excessive delays. Various improvements to the basic PRMA protocol have been proposed for multimedia cellular systems. For TDMA systems, dynamic and centralized PRMA are proposed where time slots are assigned to users according to the amount of bandwidth required and their priority levels. After a contention period, the base station allocates as much of the user's requested rate as possible. For time-division CDMA (TDCDMA) systems, PRMA-based techniques further divide each time slot into subslots using up to eight spreading codes. These subslots are used for contention and data transmission. The reservation will last until the end of the voice burst or for a certain numbers of data frames. By controlling the contention access and allocation for data services, the protocol is able to track delay requirements and dropping probabilities for different services.

For CDMA systems, the packets are classified according to their traffic rate and queued according to their priority levels. The terminal can be at three states: idle, active, and blocked. Different techniques are studied to meet different rate requirements by using variable processing gain or multiple codes. In hybrid systems [16], short packets are transmitted on an ALOHA basis using random access with no access delay and minimum overhead. In the case of larger packets the terminal will request a dedicated channel (code) on the access channel. The base station will evaluate if the request resources are available to assign a transmission format with the time that the user can start transmitting. Once the transmission is finished the terminal will maintain the link for a certain time. If a packet is generated within that time, the user transmits immediately, but if the packet is very large, the user has to request the channel again.

5. HANDOFF AND ADMISSION CONTROL

According to the call initiation position, in a wireless network two types of calls submit admission requests to a base station: new calls, which are initiated by mobile subscribers in the current cell; and handoff calls, which are initiated in other cells and handed off into the current cell. The function of admission control as mentioned before is to determine whether to grant radio resources to an incoming new/handoff call on the basis of information such as the current channel occupation, the bandwidth and QoS requirements of calls in service, and the characteristics of the call that requests admission.

When a call hands off to a neighboring cell whose admission control process decides to reject its admission

request, the call is forced to be terminated prematurely (dropped). One of the important tasks of call admission control is to limit the probability of such forced termination of ongoing calls, because from the viewpoint of mobile subscribers, having a call abruptly terminated in the middle of the conversation is less desirable than new call attempts being blocked occasionally. Hence most of the wireless admission control schemes are handoff-prioritized schemes, which offer handoff calls higher priorities over new calls to access the limited radio resources.

It should be noted here that a lower handoff call blocking probability is obtained at the cost of an increase in the new-call blocking probability. Therefore the admission control schemes must be carefully designed to balance these two types of blocking in order to achieve a better performance. Many handoff priority-based admission control schemes, that range from static to dynamic, have been proposed in the literature and they can be roughly classified into three categories [17]: guard channel schemes, queuing priority schemes, and channel borrowing schemes.

5.1. Guard Channel Schemes

In guard channel schemes (also called cutoff priority schemes), some of the radio channels are reserved for the exclusive use of handoff calls while the rest of the channels are shared equally by both new calls and handoff calls. One critical factor that influences the performance of guard channel schemes is the number of channels that need to be reserved. If the reservation is low (underreservation), the QoS requirements on handoff call blocking probability cannot be met as shown in Fig. 1a. On the other hand, a higher level of reservation (over-reservation) may result in a large number of new-call attempts being rejected.

Depending on how the number of guard channels is determined, guard channel schemes can be further classified into static schemes and dynamic schemes. For static guard channel schemes the number of channels reserved for handoff purposes is fixed for each cell, and it is calculated based on the knowledge of the traffic pattern of the area and the estimation of channel occupancy time distribution at the system design stage. The major advantage of this static approach is its simplicity since no communication and computation overheads are involved. However, the problems of overreservation and underreservation are unavoidable if the cell traffic does not conform to the prior knowledge. Therefore dynamic reservation schemes are designed to overcome these problems. Through the use of current system information, such as user mobility information and channel occupation information, dynamic schemes can determine the number of guard channels in a real-time fashion, and therefore they can easily adjust to the changing conditions of the system.

It should be noted here that no matter how the reservation is made, the guard channel schemes may result in a reduction of the total carried traffic (as shown in Fig. 1b). In general it is the originating calls and not the ongoing calls that really add to the total traffic [18]. Because fewer channels are available to new calls, the larger the number of the guard channels, the higher the probability the originating calls being blocked and, hence the less the overall traffic carried by the system.

5.2. Queuing Priority Schemes

The basic idea of the queuing priority scheme is that when a new call or a handoff call cannot be granted the required channels at its arrival time, the call is put into a queue waiting for its admission conditions to be met. Queuing of handoff calls is possible due to the fact that there is a finite time interval between the time that the received signal level drops below the handoff threshold and the time the call is terminated due to insufficient signal level. As shown in Fig. 2, handoff can occur at any time during the time interval Δt.

Queuing of new calls is possible because of the use of common channel signaling in digital communication systems. In the standard Public Switched Telephone Network (PSTN) the queuing of new calls is impractical since the signaling needed for the dialing is done on the communication channel itself. Queuing of a new call would therefore result in multiple redials that would unnecessarily occupy some communication channels. In cellular systems, however, the setup of a call is done on a separate control

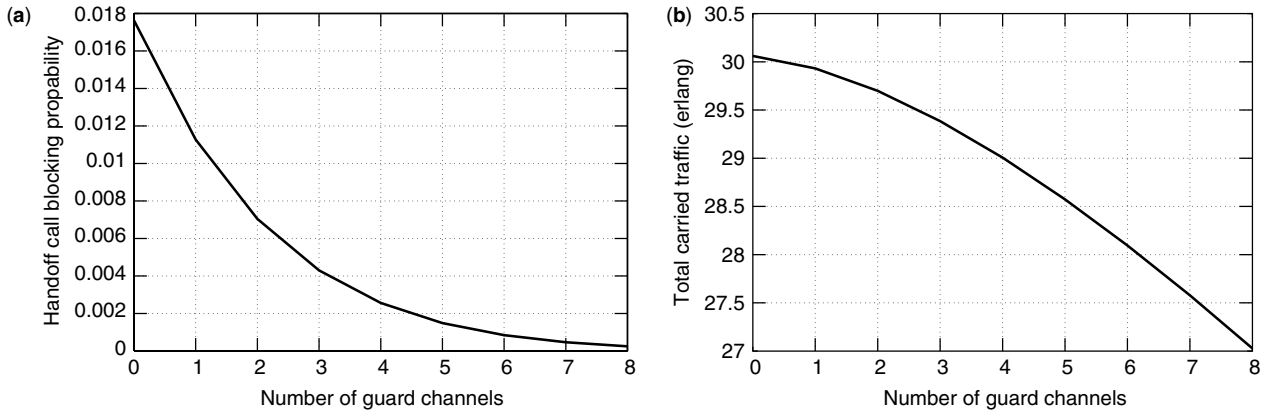


Figure 1. The system performance versus number of guard channels: (**a**) handoff call blocking probability; (**b**) total carried traffic.

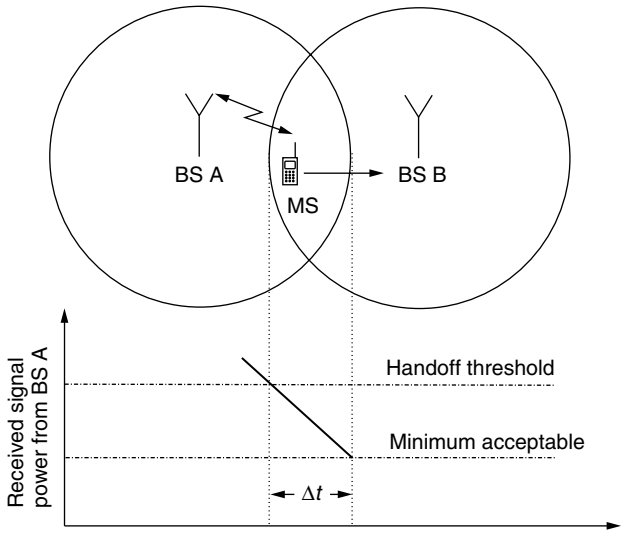


Figure 2. Handoff time interval.

channel, which can provide the system with a way of queuing new calls without affecting the transmission channels.

According to the types of calls that are queued, the queuing priority schemes can be further classified into handoff call queuing, new-call queuing, and new/handoff call queuing. Handoff call queuing schemes put the incoming handoff call in the queue and block new call attempts if there are no available channels. New call queuing is often used in combination with guard channel schemes to increase the carried traffic; if the new-call admission conditions are not met, then the arriving new calls are put into a queue to wait for the channels to be released. In the new/handoff call queuing schemes, both new calls and handoff calls are queued in the same queue and handoff calls are given non-preemptive priorities over new calls.

5.3. Channel Borrowing Schemes

The channel borrowing scheme is a combination of fixed and dynamic channel assignment schemes. The channel borrowing schemes work as follows: when all the channels in a cell are occupied, the cell borrows channels from other cells to accommodate the incoming handoff calls, as long as the borrowed channels do not interfere with the ones used by existing calls. The channel borrowing schemes are more flexible in the sense that by "moving" (borrowing) channels from less busy cells to more busy cells, a balanced performance throughout in the system can be achieved.

One problem associated with the channel borrowing scheme is channel locking. This occurs when cells within the required minimum channel reuse distance from a cell that is using a borrowed channel cannot use the same channel. Reuse distance in a cellular system is defined as the minimum distance between two cells that may use the same channels. Reuse factor is a cell plan parameter equivalent to reuse distance. It is the minimum number of channels needed to establish one call connection at each cell without reusing a channel in cells closer than the reuse distance [19]. Figure 3 shows a part of a wireless network that has frequency reuse factor 7. When cell B

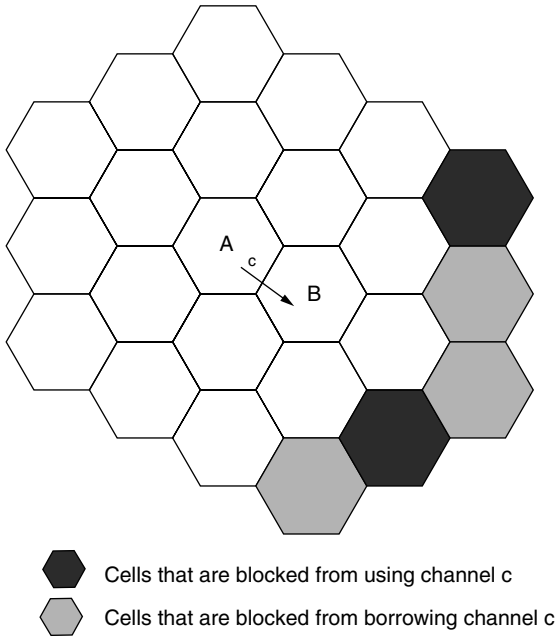


Figure 3. Channel locking caused by borrowing action.

borrows channel c from cell A to serve an incoming handoff call, the predefined frequency reuse pattern is temporarily violated: those darkest shadowed cells are prohibited from using channel c, although channel c is originally assigned to these cells and the darker shadowed cells cannot borrow channel c due to the cochannel interference requirements.

Channel locking problems make the channel borrowing schemes more complicated than the guard channel schemes and queuing priority schemes because the decreased handoff call blocking probability is obtained at the cost of decreasing the capacity of other cells, which, in turn, will cause QoS degradation in these cells. In order to achieve a better performance, a channel borrowing scheme must try to minimize this cost through cell coordination and centralized control. Two commonly used borrowing protocols are the minimum influence borrowing and channel reallocation. The minimum influence borrowing algorithm aims to borrow the channel that has minimum impact on the overall performance of the system. When channel borrowing in necessary, all the borrowable channels are compared in terms of the traffic conditions in the blocked cells of each borrowable channel, and predictions are made accordingly. Then the channel that will cause the least QoS degradation in the expected future is chosen. The channel reallocation process aims at minimizing the time that a borrowed channel is used. Instead of returning the channel when the call that uses the borrowed channel completes or hands off, if there is a channel that is released by another call in the borrower cell, the borrowed channel is returned and the released channel is allocated to this call.

6. ADMISSION CONTROL IN 3G STANDARDS

Second–Generation (2G) wireless systems have focused on the development of mobile networks in order to support

conventional telephony services. The aim of next generation systems, such as 3G technologies wideband CDMA (W-CDMA), cdma2000, and the Universal Mobile Telecommunications System (UMTS), [International Mobile Telecommunications in 2000 (IMT-2000)], is to support a variety of data services while increasing the system capacity. As an interim solution, the global system for mobile communications (GSM) operators are moving toward the general packet radio system (GPRS) technology. At the same time the TDMA (and some GSM) operators are planning for enhanced data rate for global evolution (EDGE). The IS95 CDMA operators are considering 1XRTT, which is the interim step toward CDMA-2000. Two other interesting approaches being developed are higher data rate (HDR) and "1 EXTREME." One industry group, the 3rd Generation Wireless Partnership Project (3GPP), is developing the 3G standards for GSM-based and wideband CDMA (WCDMA) air interface, while the 3rd Generation Partnership Project 2 (3GPP2), is developing 3G standards for cdma2000-based systems and the Universal Wireless Communications Consortium (UWCC) for the evolution of North American-TDMA (NA-TDMA) technology.

For the establishment of a packet data session, a GPRS UE (user equipment) must activate a packet data protocol (PDP) context where QoS parameter values are negotiated according to the availability of resources [20]. The QoS profile consists of five attributes: delay, service precedence, reliability, mean throughput, and peak throughput. However, since there is no per-flow prioritization and only best effort traffic is supported, end-to-end QoS, such as delay attribute, is not implemented. In UMTS, the PDP context mechanism has been improved to support QoS for multiple application flows with enhanced QoS negotiation and setup, and as a result, network QoS for end-to-end services can be realized.

For a CDMA air interface, QoS requirements are satisfied for different traffic classes. For instance, hard QoS guarantees are provided to realtime applications such as voice and video, and best-effort service to non-real-time applications such as packet data. Therefore, the resources are offered in accordance with the specific group characteristics. Group behavior of a class is implemented by power control and spreading control. An extensive research is done for both uplink and downlink cases with class-based bandwidth scheduling schemes to attain differentiated QoS on the CDMA air interface. The admission control in these schemes requires a radio resource allocation framework that characterizes the capacity model of a CDMA air interface and QoS models of various traffic classes.

In WCDMA, the power and rate adaptation algorithms can be implemented by the power/rate scheduler and the transmission time control by the time scheduler. The necessary radio channel characteristics can be provided by a resource estimator and the built-in capacity models in a call admission control module, and/or the resource estimator that can translate the gain from optimal power and rate allocation into better capacity estimation and utilization. In TD-CDMA, the admission control also has a time-slot assignment for uplink or downlink. This flexibility provides better adaptation to different scenarios, such as traffic asymmetry between uplink and downlink. Furthermore, allocation of CDMA codes and TDMA time slots provides higher granularity.

In IS95B, higher-data-rate service is provided through code aggregation. Specifically, up to eight codes can be assigned for the duration of a burst (one fundamental code and seven supplemental codes) requiring a burst-level admission control. A call origination with the packet data service option is used to establish a packet data service level registration. When the user remains idle for a predetermined inactivity time, the air interface resource is deallocated but the packet data registration remains established. The fundamental channel is assigned for the duration of the packet data call, whereas the supplemental channels are assigned for the duration of the packet data burst. Some of the major enhancements of CDMA2000 include the addition of a pilot channel on the reverse link and closed loop power control on the forward link. Furthermore, two additional states are added. In the active state, both a traffic and a control channel are assigned to a mobile. In the control hold state, the traffic channel is released but the control channel is maintained. In the suspended state, no interface channels are assigned, but the radiolink protocol state is remembered to avoid delays during reinitialization. The CDMA2000 physical layer provides a single supplemental channel with variable spreading gain. CDMA2000 defines a multimode enhanced random, and reservation-access scheme [13]. It is a combination of well-known packet reservation multiple access and common channel multiple access concepts. For an efficient high-speed packet data traffic with variable duration and data rates, admission control is enhanced with fast congestion control, fast capture feedback, interference control, and closed-loop power control. Depending on the service type and packet size, the mobile may need to request a dedicated channel or a reserved common channel, or simply include its packet in its random-access probe.

Standardization activities also include defining 3G wireless networks based on a TDMA air interface, namely EGPRS. It uses a TDMA-based packet-switched radio technology and a new air interface, EDGE, and GPRS core network designed for best-effort packet data services. Most of the research and development efforts in supporting QoS for multiple service classes in TDMA networks include packet scheduling schemes and multiple time-slot assignments.

7. CONCLUDING REMARKS

The explosive growth of the Internet and the continued dramatic increase for all wireless services are fueling the demand for increased capacity, data rates, supported multimedia services, and support for different QoS requirements for different classes of services. The scarcity of available radio spectrum limits the obtainable user data rates, and therefore issues associated with the QoS, network management and control, and system adaptability are rapidly gaining critical research and commercial importance. In this article we discussed the problem of admission control in circuit-switched and packet-switched wireless networks, and presented

efficient admission control schemes in order to maximize the utilization of the limited radio resources, while maintaining the QoS requirements of mobile users. The various elements and processes associated with the admission control in wireless networking technologies include the assignment and allocation of limited network resources, such as bandwidth, channels, and transmission powers. Depending on the dedication of bandwidth and channels, we classified the channel allocation mechanisms and the corresponding system operation in the following modes: circuit switching, packet switching, burst switching, and common channel packet switching. For each one of these various modes we described in detail the overall admission control process. We have also provided an overview of the available handoff priority-based admission control schemes, in order to address tradeoffs between the handoff call blocking probability and the new call blocking probability. Finally we discussed the role and operation of admission control in the various standards for the current and next-generation wireless networks.

BIOGRAPHIES

Symeon Papavassiliou received a diploma in electrical engineering from the National Technical University of Athens, Greece, in 1990 and the M.Sc. and Ph.D. degrees in electrical engineering from Polytechnic University, Brooklyn, New York in 1992 and 1995, respectively. From 1995 to 1996 Dr. Papavassiliou was a technical staff member at AT&T Bell Laboratories in Holmdel, New Jersey, and from 1996 to August 1999 he was a senior technical staff member at AT&T Laboratories in Middletown, New Jersey. From June 1996 till August 1999 he was also an adjunct professor at the Electrical Engineering Department of Polytechnic University, Brooklyn, New York. Since August 1999 he has been an assistant professor at the Electrical and Computer Engineering Department of New Jersey Institute of Technology, Newark, New Jersey. Dr. Papavassiliou was awarded the Best Paper Award in INFOCOM'94 and the AT&T Division Recognition and Achievement Award in 1997. Dr. Papavassiliou has an established record of publications in his filed of expertise, he is the Director of the Broadband, Mobile, and Wireless Networking Laboratory at NJIT, and one of the founding members of the New Jersey Center for Wireless Networking and Security (NJWINS). His main research interests lie in the areas of computer and communication networks with emphasis on wireless communications and high-speed networks, network design and management, TCP/IP and internetworking, computer network modeling and performance evaluation and optimization of stochastic systems.

Sebnem Zorlu Ozer received a B.S. degree in electronics and telecommunications engineering with highest honors in 1992 from Istanbul Technical University, Turkey, an M.S. degree in electrical engineering in 1995 from Bogazici University, Turkey, and a Ph.D. degree in electrical engineering in 2001 from New Jersey Institute of Technology. Since 2001, she has been a systems engineer at Meshnetworks Inc, where she has been working on the design and development of ad-hoc networks. Her primary responsibilities are focused on integrating quality of service within a dynamic mobile network. Her areas of interest are design and management of wireless networking systems, performance analysis of computer networks, optimization of stochastic systems and quality of service management in mobile networks.

Jiongkuan Hou received his B.S. and his M.S. degrees in electrical engineering from Northern Jiaotong University, Beijing, China, in 1993 and 1999, respectively. He is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at New Jersey Institute of Technology, and a research assistant in the Broadband, Mobile, and Wireless Networking Laboratory at NJIT, as well as a member of the New Jersey Center for Wireless Telecommunications. His research interests lie in the areas of design and management of mobile cellular networks, with emphasis on resource allocation, call admission control, pricing, and multimedia service support.

BIBLIOGRAPHY

1. D. J. Goodman, *Wireless Personal Communications*, Addison Wesley, Longman, 1997.
2. L. Jorguseski, E. Fledderus, J. Farserotu, and R. Prasad, Radio resource allocation in third-generation mobile communication systems, *IEEE Commun. Mag.* 117–123 (Feb. 2001).
3. R. W. Nettleton and H. Alavi, Power control for a spread spectrum cellular mobile radio system, *Proc. IEEE Vehicular Technology Conf.*, 1983, pp. 242–246.
4. J. Zander, Performance of optimum transmitter power control in cellular radio systems, *IEEE Trans. Vehic. Technol.* **41**(1): 57–62 (Feb. 1992).
5. N. Bambos, Toward power-sensitive network architectures in wireless communications: Concepts, issues, and design aspects, *IEEE Pers. Commun.* 50–59 (June 1998).
6. W. C. Y. Lee, Overview of cellular cdma, *IEEE Trans. Vehic. Technol.* **40**(2): 291–302 (May 1991).
7. K. S. Gilhousen and I. M. Jacobs, On the capacity of a cellular cdma system, *IEEE Trans. Vehic. Technol.* **40**(2): 303–312 (May 1991).
8. S. Ramakrishna and J. M. Holtzman, A scheme for throughput maximization in a dual-class cdma system, *IEEE J. Select. Areas Commun.* **16**(6): 830–844 (Aug. 1998).
9. J. M. Capone and L. F. Merakos, Integrating data traffic into a cdma cellular voice system, *Wireless Networks* **1**(4): 389–401 (1995).
10. I. F. Akyildiz, D. A. Levine, and I. Joe, A slotted cdma protocol with ber scheduling for wireless multimedia networks, *IEEE/ACM Trans. Network.* **7**(2): 146–158 (April 1999).
11. J. Wu and R. Kohno, A wireless multimedia cdma system based on transmission power control, *IEEE J. Select. Areas Commun.* **14**(4): 683–691 (May 1996).
12. J. Zander, Radio resource management in future wireless networks — requirements and limitations, *IEEE Commun. Mag.* **35**: 30–36 (Aug. 1997).
13. Telecommunications Industry Association, *The cdma2000 ITU-R RTT Candidate Submission-TR45-5.5*, 1998.

14. S. Z. Ozer, S. Papavassiliou, and A. Akansu, On performance of switching techniques for integrated services in cdma wireless systems, *Proc. IEEE Vehicular Technology Conf.*, 1973, 2000, Vol. 4, pp. 1967–1973.

15. ETSI TS 125 322 V3.1.2 (2000–01), *Technical Specification Universal Mobile Telecommunications System (UMTS); RLC Protocol Specification*.

16. C. Roobol, P. Beming, J. Lundsjo, and M. Johansson, A proposal for an rlc/mac protocol for wideband cdma capable of handling real time and non real services, *Proc. IEEE Vehicular Technology Conf.*, May 1998, pp. 107–111.

17. J. Hou, J. Yang, and S. Papavassiliou, Integration of pricing with call admission control for wireless networks, *Proc. IEEE Vehicular Technology Conf.*, 2001, Vol. 3, pp. 1344–1348.

18. I. Katzela and M. Naghshineh, Channel assignment schemes for cellular mobile telecommunication system: A comprehensive survey, *IEEE Pers. Commun.* **3**: 10–31 (June 1996).

19. S. Papavassiliou, L. Tassiulas, and P. Tandon, Meeting QoS requirements in a cellular network with reuse partitioning, *IEEE J. Select. Areas Commun.* **12**(8): 1389–1400 (Oct. 1994).

20. R. Koodli and M. Puuskari, Supporting packet-data QoS in next-generation cellular networks, *IEEE Commun. Mag.* **39**(2): 180–188 (Feb. 2001).

ALOHA PROTOCOLS

JOHN J. METZNER
Pennsylvania State University
University Park, Pennsylvania

1. INTRODUCTION

The ALOHA method originated at the University of Hawaii as a means for multiple users to send short data packets via radio transmission over a common channel to a central station, in a largely uncoordinated manner [1]. The central station (see Fig. 1) would rebroadcast all its receptions on a different channel, so that the sending users would be fed back their own transmissions along with others. If the user observes that its data packet has collided with another, it would retransmit some time later. If the sender sees its own packet unaltered, this serves as an acknowledgment.

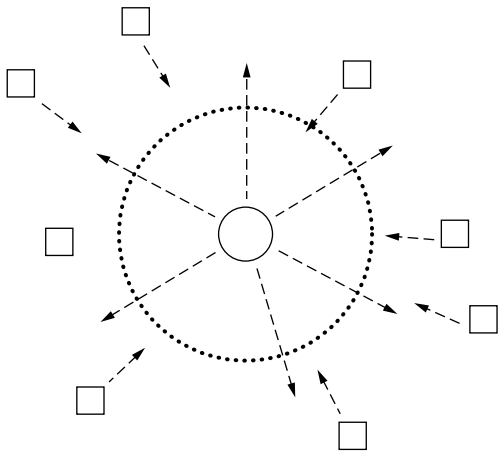


Figure 1. Rebroadcast by the central station (circle) of signals sent by the ALOHA participants (squares).

Two major categories of ALOHA are unslotted ALOHA and slotted ALOHA. In unslotted ALOHA, a transmitted packet can be sent at any starting time and can be of varied duration. In slotted ALOHA, there are fixed-sized synchronized time slots. A sender can send in any time slot, and the data packets should all be slightly smaller than a time slot duration. Collisions can be complete or partial in unslotted ALOHA, but are complete in slotted ALOHA. Even a partial collision usually calls for retransmission of the packet.

If the broadcast method does not provide enough information for the sending user to be certain of successful transmission, another option is for the destination or the central station to send back an individual acknowledgment signal.

ALOHA is very effective in light traffic situations; the sender can use the whole channel, and send very rapidly, usually without collision. With heavier traffic, ALOHA suffers from problems of efficiency and stability. Various protocols and signal processing methods have been invented to alleviate these problems. These include collision resolution algorithms, control of arrival rate and retransmission time, reservation ALOHA, multichannel ALOHA, capture ALOHA, and diversity reception. Also, a modification of unslotted ALOHA using carrier-sense multiple access with collision detection (CSMA/CD) is the basis for the popular Ethernet local-area network protocol [2].

2. QUANTITATIVE ANALYSIS OF IDEALIZED SLOTTED ALOHA COMMUNICATION

Suppose slotted ALOHA is used and all users observe the results of all transmissions. Models have assumed either (1) an infinite number of users having a finite total message arrival rate or (2) a finite number of users. The time delay for retransmission is assumed to be independently randomized for each retransmission; if it were the same for two colliders, they would collide again for certain.

2.1. The Infinite Number of Users Model

Packets (including packet retransmissions) are presumed to be generated by the infinite set of all users at a total finite rate of G packets per time slot. The number of transmissions in a slot is assumed to obey a Poisson distribution:

$$P[k] = \frac{G^k}{k!}e^{-G} \tag{1}$$

There is a successful transmission in a slot if and only if exactly one transmission occurs. Let S be the fraction of successful slots, also called *the throughput per slot*.

$$S = Ge^{-G} \tag{2}$$

S is maximum at $G = 1$, where it equals e^{-1} or 0.368.

G consists of two components: an average new packet arrival rate denoted as λ, and an average retransmission rate denoted as r: ($G = \lambda + r$). This model does not properly

reflect the effect of statistical variations. For a given average arrival rate λ, there is a finite probability that a short-term packet transmission rate G will occur that is large enough to cause the success (departure) rate to fall below λ. This will cause increased r (more retransmissions needed), which reduces S, leading to still higher r, until almost all traffic is retransmission traffic and hardly any are successful. This potential instability problem can be alleviated by blocking new packet transmissions and/or increasing average retransmission delay on observing excessive collision.

2.2. The Finite Number of Users Model

Assume that there is a fixed number M of active users. Suppose that k of these users are backlogged, which means that they have a need to retransmit a previously collided packet. Backlogged users do not send a new packet until after the backlogged packet is successful, at which time it becomes a nonbacklogged user. Assume that each $M - k$ nonbacklogged user has, independently of other users or its own past events, a probability P_A of sending a packet in the next slot, and that each of the k backlogged users has, independently, a probability P_R of sending its retransmission in the next slot.

This model is artificial because there is no fixed number of "active" users in practice, and active users are bursty in their needs for transmission. A large number of users tend to average out the behavior, however, so the model may approximate the real situation.

Suppose at a given time there are k backlogged users. Call this state k. The offered load in state k is

$$G(k) = (M - k)P_A + kP_R \tag{3}$$

For moderately large M, the success rate $S(k)$ can be approximated by [3]

$$S(k) \approx G(k)e^{-G(k)} \tag{4}$$

where $G(k)$ is as given by (3).

In state k, there is an average arrival rate of $(M - k)P_A$. The success rate $S(k)$ can be thought of as a departure rate. When arrivals exceed departures, the backlog increases and G increases, since, normally, $P_R > P_A$. The reverse happens when departures exceed arrivals. Figure 2 illustrates how system behavior tends to drift.

The arrows along the departure curve denote the drift direction. Point A is a desirable stable operating point. However, occurrence of a short-term jump in arrival rate can easily move the system past the unstable equilibrium point B, after which k increases to send the system to the undesired stable operating point C.

For very small P_R there can be a single desirable stable point, but very small P_R means long delay. Ideal operation would be to control P_R as a function of k so as to keep $G(k)$ as close to 1 as possible, but the state is difficult to know exactly. A simple control scheme could increase P_R when an idle slot is observed and decrease it when a collision is observed.

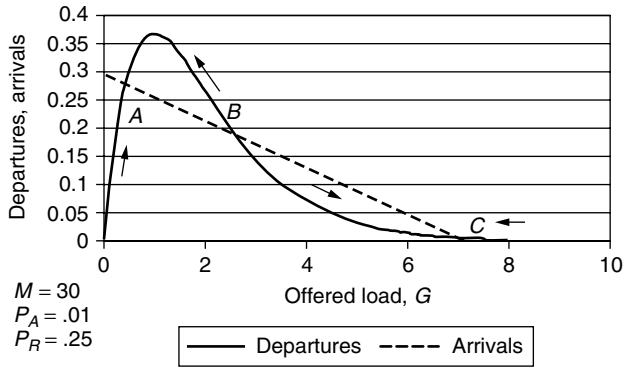


Figure 2. Drift analysis of ALOHA stability.

3. UNSLOTTED ALOHA

Unslotted ALOHA packets can start at any time and can be of variable size. If one assumes that any overlap of two frames causes both to be lost, unslotted ALOHA is considerably less efficient than slotted ALOHA. For equal-size packets the maximum efficiency is $\frac{1}{2}e$, or half that of slotted ALOHA. With unequal-size packets the maximum efficiency is slightly higher than $\frac{1}{2}e$. However, variable size packets often are desired, in which case slotted ALOHA would suffer inconvenience and/or inefficiency if packets had to be broken up or only partially filled slots.

4. IMPROVING THE EFFICIENCY OF ALOHA

There are three avenues for modifying ALOHA communication to improve its efficiency:

1. Use protocols to eliminate or reduce collision frequency.
2. Tolerate collision through signal design, signal processing, and/or diversity reception.
3. Increase the number of bits per slot by using nonbinary transmission.

Techniques in the first avenue include (a) unslotted — send only when no activity is sensed on the channel, called carrier-sense multiple access (CSMA); (b) unslotted — stop sending if a collision is detected, called collision detection (CD); (c) slotted — collision resolution algorithms.

Avenues 1a and 1b are employed as CSMA/CD in the Ethernet protocol for local-area networks. The sensing of the channel does not result in perfect collision avoidance because of the nonzero propagation time between potential senders. Wireless networks can use CSMA but seldom CD, because the locally strong signal of a transmitting station prevents detection of a locally weak signal transmitted remotely in the same timeframe and frequency band. With the central station model the central station broadcasts on a different band than the multiaccess users. The collisions are detected, but the round-trip propagation time may make it too late to benefit from a halt in transmission.

5. COLLISION RESOLUTION ALGORITHMS

These algorithms [4] improve the efficiency of slotted ALOHA by forcing a resolution of a collision. After a collision, only the colliders (the backlogged set) are permitted to send until the collision is resolved by their success. The algorithm ensures successes by the colliders within a minimum average number of slot times, while also allowing all users to know when the collision has been resolved. All potential senders are assumed to be informed of whether the prior slot experienced a success or a collision, or was idle; they are not presumed to know how many have collided.

One may question how the potential senders all get to know of the collision just before the next slot. The slotted ALOHA channel could occupy a slot in a TDM frame whose duration is longer than the round-trip time needed to learn of the collision. Figure 3 illustrates such a slot.

The gist of the algorithms, without going into the details, is as follows. Each collider picks a number in some agreed-on range; the range is split into two subsets, and those in the first subset send. If there is no collision, the first subset is resolved and attention turns to the second subset. If there is a collision, the first subset is split in two, eventually leading to resolution of the first subset.

With these algorithms, the maximum efficiency is increased from 0.368 to close to 0.5, with the exact gain dependent on the algorithm details and assumptions. Perhaps more importantly, the algorithms ensure stability at all arrival rates less than the maximum efficiency, which is the limit as $n \to \infty$ of n divided by the expected number of slots to resolve n colliders. This is because, with n colliders, the average number of arrivers during the average number of slots needed to resolve the colliders remains less than n. Thus n always tends to drift lower. Delay also is reduced, because it is not necessary to use a small P_r for a backlogged packet; the backlogged packet delay is no more than the time to resolve its collision.

6. RESERVATION ALOHA AND HYBRID SYSTEMS

In reservation systems, the information transmitted to make a reservation of data to be sent afterward is normally a minute fraction of the total information flow. Thus, a small portion (subchannel) of the channel can be devoted to reservation traffic, and even that subchannel can be lightly used. ALOHA methods work well in a light traffic channel, because transmissions are fast and collisions are rare. A collision merely becomes a delayed reservation, which is not too troublesome if rare.

ALOHA can also be used to implicitly reserve a slot in a TDM system for an indefinite period of successive frames by successfully transmitting data in that slot. Others could defer attempting to send in that slot until it was observed that the slot had become idle. This method is called *reservation ALOHA* [5].

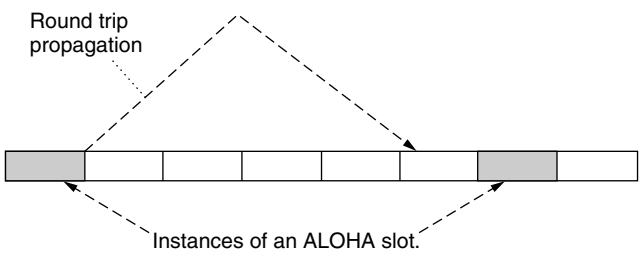


Figure 3. Occurrences of successive instances of an ALOHA slot.

7. MULTICHANNEL ALOHA AND MULTICOPY ALOHA

The ALOHA idea can be extended to multiple channels. These channels might be different frequency bands (FDM), different time slots of a synchronized frame (TDM), or different orthogonal code sequences in CDMA. If there are n frequency channels instead of 1, the time to send a packet increases by a factor of n. This would be a slight disadvantage under light traffic, since it would take much longer to send a burst of packets unless more than one channel was used simultaneously. There is no gain in total capacity by splitting into multiple channels.

Consider TDM with multiple ALOHA channels. One possibility is that if a collision occurs, the sender should resend in a randomly chosen channel. But this is just like picking at random to send in one of n slots, which is not much different than the one-channel strategy of sending a random time later after a collision.

Another thought is to send multiple copies of a packet in the hope that at least one will get through. This may be okay to do for a small subset of important, delay-sensitive messages. However, it is highly inefficient as a general policy. This is because (1) the limit of a fraction 0.368 of total successful packets can't be exceeded and (2) multiple successes of the same packet are included in this fraction, so the proportion of different successful packets will be much lower than 0.368.

8. CAPTURE ALOHA AND ABILITY TO TOLERATE COLLISION

The discussion so far has assumed that all colliding packets are lost. This may be overly pessimistic. There is a capture effect such that if one of two received signals is a certain amount stronger than an interferer, the strongest signal can be decoded. More sophisticated decoding and signal processing techniques may even allow both of two colliding packets to be decoded. If the stronger of two interfering signals is decoded, the effect of this now known signal could be subtracted out to a large extent, possibly allowing successful decoding of the weaker signal.

With slotted ALOHA, the number of simultaneously transmitted signals is rarely greater than 2, so simultaneous decoding may not be that complex. If two could be decoded simultaneously, but not more than two, the throughput per slot would be

$$S = Ge^{-G} + G^2e^{-G} \tag{5}$$

where S is maximized at $G = 1.618$ and is

$$S_{\max} = 0.840 \text{ packets/slot} \tag{6}$$

One simple way of using the capture effect to increase the throughput of slotted ALOHA is to use two different packet transmission power levels. Assume that if exactly one strong signal packet is sent in a slot, along with any number of weak signal packets, the strong signal packet is successfully decoded, but the weak one(s) will be lost. In any other collision event, all packets are assumed to be lost. This allows a higher efficiency than ALOHA without capture, and for an optimum proportion of high-power senders the maximum throughput is increased [6] from $1/e$ or 0.368 to $(e^{-(1-1/e)})$, or about 0.53.

Even without intentionally designing for two different power levels, received signals will come in with different powers. With fading, individual senders will sometimes come in at high power, sometimes with low power. This is an advantage [7]. A receiver could have multiple antennas that experience independent fading on different antennas. With this diversity reception feature, a signal A could be received more strongly than signal B on one antenna, while signal B could be received more strongly than signal A on another antenna, such that both could be captured. In ALOHA systems, fading may actually create greater throughput than without fading.

9. MULTIBASE ALOHA

Extending the idea of multiple antennas for diversity reception in ALOHA systems, it is possible to have a network of cooperating base stations [8]. A mobile sender transmits a packet, and if at least one base station can decode it, the packet could be successful. Duplicates could be recognized by the network, and the base station with the strongest reception could supply the feedback acknowledgment. This way several mobile senders can be successful simultaneously. Also, handoff would not be necessary unless the mobile left the entire base station network.

10. ALOHA WITH A TIME CONSTRAINT

ALOHA systems rely on being able to repeat collided packets. This is a drawback to time-constrained traffic that must meet some deadline. Still, the time constraint can allow several retransmissions if round-trip acknowledgment is much shorter than the delay tolerance. On mobile-base communication the distances are relatively short, allowing short round-trip times. If the deadline for a packet is not met, no further retransmissions are attempted. As long as these losses are tolerable, this helps system stability, since backlogged packets are removed from the offered load by the deadline.

The capture effect with two power levels can be used to give priority to one class of signals. For example, time-constrained traffic such as real-time voice can be transmitted with higher power than non-real-time data transfer. Another option, if there is time for multiple retransmissions, is to use the high power only if the deadline is imminent [9]. Also, in the multichannel case, it has been suggested [10] to send copies of the same packet on multiple channels when the deadline is imminent. Both the higher-power and the multiple-copy techniques use the principle of devoting more signal energy to the transmission of the packet when the deadline nears.

11. NONBINARY TRANSMISSION WITH ALOHA

In systems where collision is a more serious problem than noise, signal-to-noise ratio is relatively high. Channel capacity would then indicate that more bits per hertz could be sent by using nonbinary transmission. Consider the channel capacity formula for white Gaussian noise channels:

$$C = F \times \log\left(1 + \frac{P}{N_0 F}\right) \tag{7}$$

where C is capacity in bits per second, F is the bandwidth in hertz, $N = N_0 F$ is the noise power, and P is the signal power. When P/N is large, C is proportional to P/N in dB (decibels). Thus there would be 5 times the capacity at 40 dB as at 8 dB. If ALOHA sent at 5 times the rate, its same size packets would only be one-fifth as long in duration, and thus would be much less likely to collide if data were generated at the same long-term average rate. If P/N varied widely, however, adaptation to the current channel capacity could be complex and difficult to accomplish.

12. CDMA AND SPREAD ALOHA

Code-division multiple access [11] is a technique whereby many users send simultaneously over a wide frequency band. In the direct sequence version, a sender sends a bit by sending its unique pseudorandom L-bit binary code sequence or its inverse. The individual data rate is lower by a factor of L (spreading factor) than the data rate at which a single user could send. The receiver decodes a particular sender by correlating with the sender's known code. Abramson [12] has suggested an idea of everyone using one code, in a scheme he calls "spread ALOHA." Suppose that two different users start sending a sequence of data at starting times d seconds apart. A correlator will output two interleaved streams of spikes spaced d seconds apart corresponding to the two senders. The approximate output is illustrated in Fig. 4, where the solid pulses represent the bits of one stream and the dashed line pulses represent the bits of the other stream. If the spacing d is not very close to zero or to a multiple of the period, the two senders' data can be readily decoded. A similar effect can be achieved by sending a narrow, low-duty-cycle pulsetrain for a packet [13]; no code or correlator would then be needed. This latter alternative could be called pulse time-spread ALOHA. Spread ALOHA is simpler than everyone having their own code, but it suffers from the usual ALOHA problem of collisions when the pulsetrains are too closely spaced. With unequal powers, the capture effect could result in one of two or more colliders being successfully decoded. For example, if the two pulsetrains in Fig. 4 were in step, the net sum of the two responses would always have the correct sign of the larger pulse, in the absence of noise.

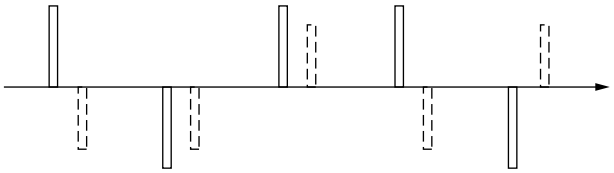


Figure 4. Two decorrelated streams or two user pulsetrains.

BIOGRAPHY

John J. Metzner is a professor of computer engineering, with appointments in both the Department of Electrical Engineering and the Department of Computer Science and engineering. He received his B.E.E., M.E.E., and Eng.Sc.D. degrees from New York University in 1953, 1954, and 1958, respectively. He has held faculty and research appointments at New York University, Polytechnic University, Brooklyn, New York, Wayne State University, Detroit, Michigan, Oakland University, Rochester, Michigan, and, since 1986, The Pennsylvania State University, University Park, Pennsylvania. He served a year as acting dean of the School of Engineering and Computer Science at Oakland University, and two years as acting director of the computer engineering program at Penn State. In research, Dr. Metzner has devised various ARQ protocols for reliable and efficient data communication, techniques for efficient comparison of remote replicated data files, efficient acknowledgement protocols for slotted ring networks, improved broadcast retransmission protocols, methods for improved utilization of ALOHA and spread spectrum multiaccess, and techniques for simpler and more effective error correction.

BIBLIOGRAPHY

1. N. Abramson, The ALOHA system, in N. Abramson and F. Kuo, eds., *Computer Communication Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
2. R. Metcalf and D. Boggs, Ethernet: Distributed packet switching for local computer networks, *Commun. ACM* **19**(7): 395–403 (July 1976).
3. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
4. J. I. Capetanakis, Tree algorithms for packet broadcast channels, *IEEE Trans. Inform. Theory* **IT-25**(5): 505–515 (Sept. 1979).
5. D. J. Goodman, Cellular packet communications, *IEEE Trans. Commun.* **38**(8): 1272–1280 (Aug. 1990).
6. J. Metzner, On improving utilization in ALOHA networks, *IEEE Trans. Commun.* **COM-24**(4): 447–448 (April 1976).
7. J. Arnbak and W. van Blitterswijk, Capacity of slotted ALOHA in Rayleigh-fading channels, *IEEE J. Select. Areas Commun.* **SAC-5**: 261–269 (Feb. 1987).
8. M. Sidi and I. Cidon, A multi-station packet radio network, *Performance Evaluation*, Vol. 8, no. 1, North-Holland, Feb. 1988, pp. 65–72.
9. J. Metzner and J.-M. Chung, Efficient energy utilization with a time constraint and time-varying channels, *IEEE Trans. Vehic. Technol.* **48**(12): 2005–2013 (Dec. 2000).
10. Y. Birk and Y. Keren, Judicious use of redundant transmissions in multichannel ALOHA networks with deadlines, *IEEE J. Select. Areas Commun.* **17**(2): 257–269 (Feb. 1999).
11. S. Tantaratana and K. Ahmed, eds., *Wireless Applications of Spread Spectrum Systems: Selected Readings*, IEEE Press, Piscataway, NJ, 1998.
12. N. Abramson, Multiple access in wireless digital networks, *Proc. IEEE* **82**(9): 1360–1369 (Sept. 1994).
13. J. Metzner, *Reliable Data Communications*, Academic Press, San Diego, CA, 1998.

AMPLITUDE MODULATION

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Speech, images, and video are examples of analog signals that are transmitted routinely over wireline and wireless (radio-) communication channels. In spite of the general trend toward digital transmission of these types of analog signals, there is still today a significant amount of analog signal transmission, specifically, audio and video broadcast. In this article, we describe the transmission of analog signals by amplitude modulation of a sinusoidal carrier. Methods for demodulation of the amplitude modulated sinusoidal carrier to recover the analog signal are also described.

The analog signal to be transmitted is generally characterized as an information-bearing message signal, which is denoted as $m(t)$. The message signal $m(t)$ is an electrical signal that may represent either an audio signal, or a still image, or a video signal. Such a signal is assumed to be a lowpass signal with frequency content that extends from $f = 0$ to some upper frequency limit, say, B Hz. Hence, if the voltage spectrum (Fourier transform) of $m(t)$ is denoted as $M(f)$, then $M(f) = 0$ for $|f| > B$. The bandwidth B of the message signal depends on the type of analog signal. For example, the bandwidth of an audio signal is typically approximately 4 kHz and that of an analog video signal is approximately 6 MHz.

The sinusoidal carrier which is modulated by $m(t)$ is expressed as

$$c(t) = A_c \cos 2\pi f_c t$$

where A_c is the (unmodulated) carrier amplitude and f_c is the carrier frequency. Basically, the modulation of the carrier $c(t)$ by the message signal $m(t)$ converts the message signal from lowpass to bandpass, in the neighborhood of the carrier f_c. This frequency translation resulting from the modulation process is performed in order to achieve one or both of the following objectives: (1) to translate lowpass signal in frequency to the passband of the channel so that the spectrum of the frequency-translated message signal matches the passband characteristics of the channel and (2) to accommodate for the simultaneous transmission of

signals from several message sources, where each message signal modulates a different carrier and, thus, occupies a different frequency band, as in frequency division multiplexing.

2. AMPLITUDE MODULATION

In amplitude modulation, the message signal $m(t)$ is impressed on the amplitude of the carrier signal $c(t)$. There are several different ways to modulate the amplitude of the carrier by the message signal $m(t)$, each of which results in different spectral characteristics for the transmitted signal. Specifically, these methods are called (1) *double-sideband*, *suppressed-carrier AM*, (2) *conventional double-sideband AM*, (3) *single-sideband AM*, and (4) *vestigial-sideband AM*.

2.1. Double-Sideband Suppressed-Carrier AM

A double-sideband, suppressed-carrier (DSB-SC) AM signal is obtained by multiplying the message signal $m(t)$ with the carrier signal $c(t)$. Thus, we have the amplitude modulated signal

$$\begin{aligned} u(t) &= m(t)c(t) \\ &= A_c m(t) \cos 2\pi f_c t \end{aligned} \tag{1}$$

The voltage spectrum of the modulated signal can be obtained by computing the Fourier transform of $u(t)$. The result of this computation is

$$U(f) = \frac{A_c}{2}[M(f - f_c) + M(f + f_c)] \tag{2}$$

Figure 1 illustrates the magnitude and phase spectra for $M(f)$ and $U(f)$.

We observe that the magnitude of the spectrum of the message signal $m(t)$ has been translated or shifted in frequency by an amount f_c. The phase of the message signal has been translated in frequency the same amount. Furthermore, the bandwidth occupancy of the amplitude-modulated signal is $2B$, whereas the bandwidth of the message signal $m(t)$ is B. Therefore, the channel bandwidth required to transmit the modulated signal $u(t)$ is $B_c = 2B$.

The frequency content of the modulated signal $u(t)$ in the frequency band $|f| > f_c$ is called the *upper sideband* of $U(f)$, and the frequency content in the frequency band $|f| < f_c$ is called the *lower sideband* of $U(f)$. It is important to note that either one of the sidebands of $U(f)$ contains all the frequencies that are in $M(f)$. Thus, the frequency content of $U(f)$ for $f > f_c$ corresponds to the frequency content of $M(f)$ for $f > 0$, and the frequency content of $U(f)$ for $f < -f_c$ corresponds to the frequency content of $M(f)$ for $f < 0$. Hence, the upper sideband of $U(f)$ contains all the frequencies in $M(f)$. A similar statement applies to the lower sideband of $U(f)$. Therefore, the lower sideband of $U(f)$ contains all the frequency content of the message signal $M(f)$. Since $U(f)$ contains both the upper and the lower sidebands, it is called a *double-sideband* (DSB *AM signal*.

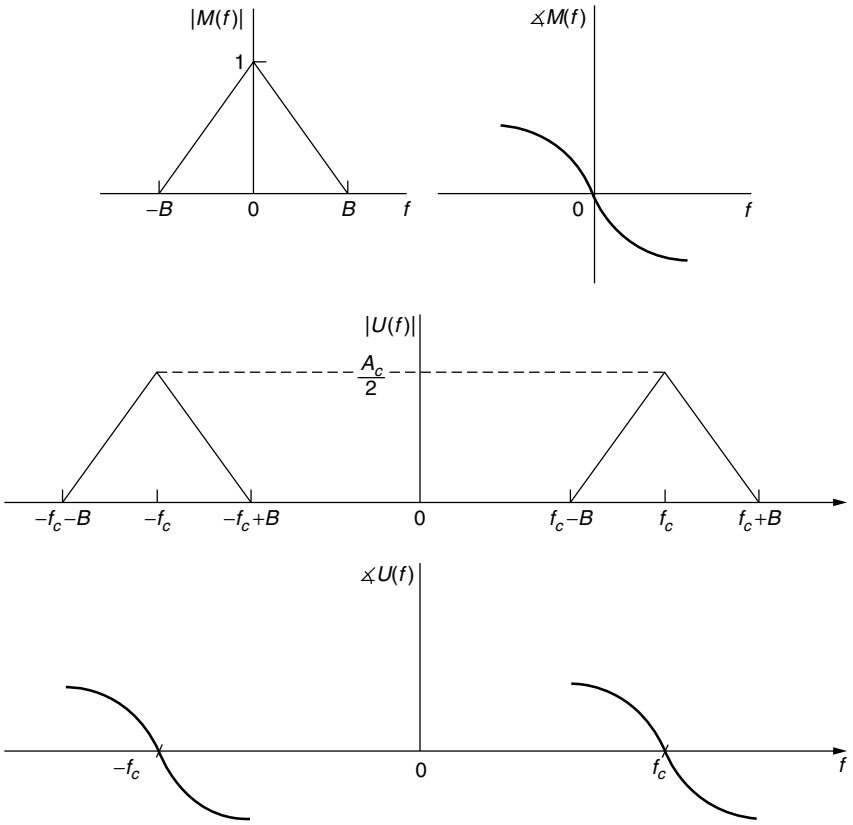


Figure 1. Magnitude and phase spectra of the message signal $m(t)$ and the DSB AM modulated signal $u(t)$.

The other characteristic of the modulated signal $u(t)$ is that it does not contain a carrier component; that is, all the transmitted power is contained in the modulating (message) signal $m(t)$. This is evident from observing the spectrum of $U(f)$. We note that, as long as $m(t)$ does not have any DC component, there is no impulse in $U(f)$ at $f = f_c$, which would be the case if a carrier component was contained in the modulated signal $u(t)$. For this reason, $u(t)$ is called a *suppressed-carrier signal*. Therefore, $u(t)$ is a DSB-SC AM signal.

In the propagation of the modulated signal through the communication channel, the signal encounters a propagation time delay, which depends on the characteristics of the propagation medium (channel). Generally, this time delay is not precisely known to the signal receiver. Such a propagation delay results in a received signal, in the absence of any channel distortion or additive noise, of the form

$$r(t) = A_c m(t) \cos(2\pi f_c t + \phi_c)$$

where ϕ_c is a carrier phase manifested by the propagation delay.

Suppose that we demodulate the received signal by first multiplying $r(t)$ by a locally generated sinusoid $\cos 2\pi f_c t + \phi)$, where ϕ is the phase of the sinusoid, and then passing the product signal through an ideal lowpass filter having a bandwidth B. The multiplication of $r(t)$ with $\cos(2\pi f_c t + \phi)$ yields

$$\begin{aligned} r(t)\cos(2\pi f_c t + \phi) &= A_c m(t)\cos(2\pi f_c t + \phi_c)\cos(2\pi f_c t + \phi) \\ &= \tfrac{1}{2}A_c m(t)\cos(\phi_c - \phi) \\ &\quad + \tfrac{1}{2}A_c m(t)\cos(4\pi f_c t + \phi + \phi_c) \end{aligned} \tag{3}$$

A lowpass filter rejects the double frequency components and passes only the lowpass components. Hence, its output is

$$y_o(t) = \tfrac{1}{2}A_c m(t)\cos(\phi_c - \phi) \tag{4}$$

Note that $m(t)$ is multiplied by $\cos(\phi_c - \phi)$. Thus, the desired signal is scaled in amplitude by a factor that depends on the phase difference between the phase ϕ_c of the carrier in the received signal and the phase ϕ of the locally generated sinusoid. When $\phi_c \neq \phi$, the amplitude of the desired signal is reduced by the factor $\cos(\phi_c - \phi)$. If $\phi_c - \phi = 45^\circ$, the amplitude of the desired signal is reduced by $\sqrt{2}$ and the signal power is reduced by a factor of 2. If $\phi_c - \phi = 90^\circ$, the desired signal component vanishes.

The discussion above demonstrates the need for a *phase-coherent or synchronous demodulator* for recovering the message signal $m(t)$ from the received signal. Thus, the phase for the locally generated sinusoid should ideally be equal to the phase ϕ_c of the received carrier signal.

A sinusoid that is phase-locked to the phase of the received carrier can be generated by use of a phase-locked loop (PLL), which is described in Refs. 1–3.

2.2. Conventional Double-Sideband AM

A conventional AM signal consists of a large carrier component in addition to the double-sideband AM modulated signal. The transmitted signal is expressed mathematically as

$$u(t) = A_c[1 + m(t)]\cos 2\pi f_c t \tag{5}$$

where the message waveform is constrained to satisfy the condition that $|m(t)| \leq 1$. We observe that $A_c m(t)\cos 2\pi f_c t$ is a double-sideband AM signal and $A_c \cos 2\pi f_c t$ is the carrier component. Figure 2 illustrates an AM signal in the time domain.

As long as $|m(t)| \leq 1$, the amplitude $A_c[1 + m(t)]$ is always positive. This is the desired condition for conventional DSB AM that makes it easy to demodulate, as described next. On the other hand, if $m(t) < -1$ for some t, the AM signal is said to be *overmodulated* and its demodulation is rendered more complex. In practice, $m(t)$ is scaled so that its magnitude is always less than unity.

The voltage spectrum of $u(t)$ given by (5) is obtained by computing the Fourier transform of $u(t)$. The result of this computation is

$$U(f) = \frac{A_c}{2}[M(f - f_c) + M(f + f_c) + \delta(f - f_c) + \delta(f + f_c)] \tag{6}$$

This spectrum is sketched in Fig. 3. We observe that spectrum of the conventional AM signal occupies a bandwidth twice the bandwidth of the message signal. As in the case of DSB-SC carrier, conventional AM consists of both an upper sideband and a lower sideband. In addition, the spectrum of a conventional AM signal contains impulses at $f = f_c$ and $f = -f_c$, which correspond to the presence of the carrier component in the modulated signal.

The major advantage of conventional AM signal transmission is the ease with which the signal can be demodulated. There is no need for a synchronous demodulator. Since the message signal $m(t)$ satisfies the condition $|m(t)| < 1$, the envelope (amplitude) $1 + m(t) > 0$. If we rectify the received signal, we eliminate the negative values without affecting the message signal as shown in Fig. 4. The rectified signal is equal to $u(t)$ when $u(t) > 0$ and zero when $u(t) < 0$. The message signal is recovered by passing the rectified signal through a lowpass filter whose bandwidth matches that of the message signal. The combination of the rectifier and the lowpass filter is called an *envelope detector*.

Ideally, the output of the envelope detector is of the form

$$d(t) = g_1 + g_2 m(t)$$

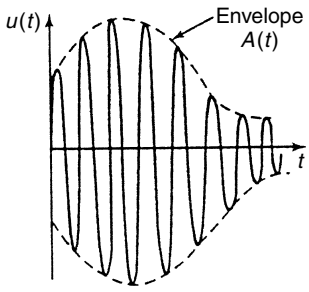


Figure 2. A conventional AM signal in the time domain.

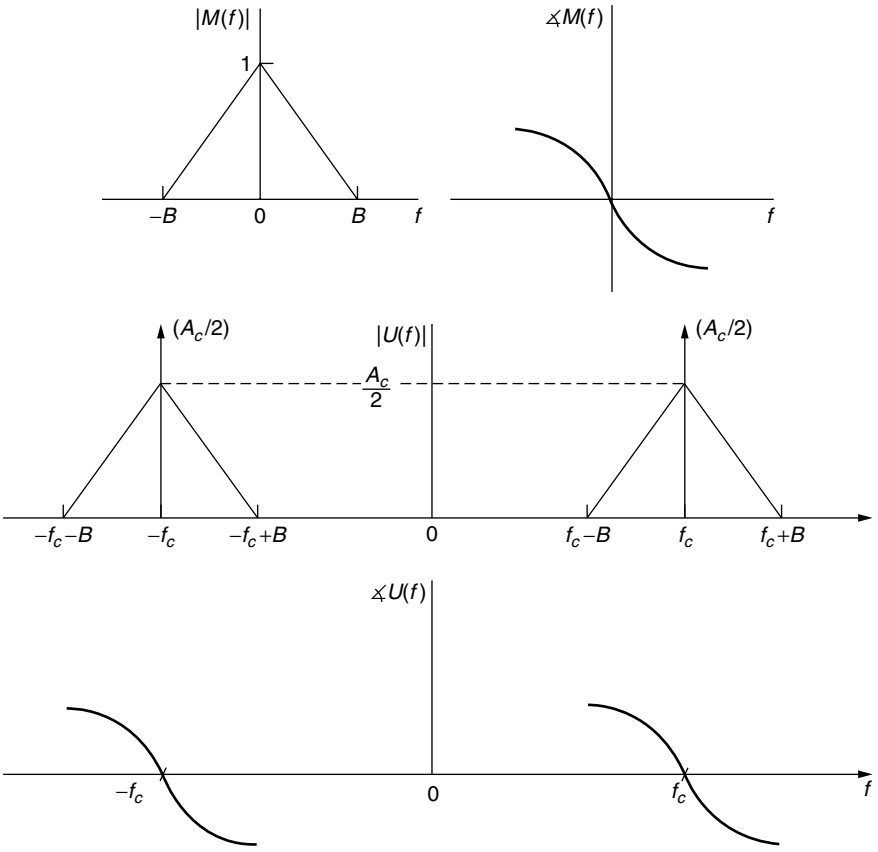


Figure 3. Magnitude and phase spectra of the message signal $m(t)$ and the conventional AM signal.

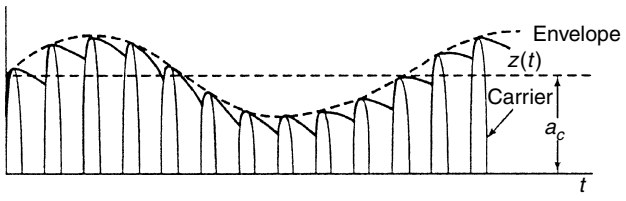


Figure 4. Envelope detection of conventional AM signal.

which g_1 represents a dc component and g_2 is a gain factor due to the signal demodulator. The dc component can be eliminated by passing $d(t)$ through a transformer, whose output is $g_2 m(t)$.

The simplicity of the demodulator has made conventional DSB AM a practical choice for AM radiobroadcasting. Since there are literally billions of radio receivers, an inexpensive implementation of the demodulator is extremely important. The power inefficiency of conventional AM is justified by the fact that there are few broadcast transmitters relative to the number of receivers. Consequently, it is cost-effective to construct powerful transmitters and sacrifice power efficiency in order to simplify the signal demodulation at the receivers.

2.3. Single-Sideband AM

In Section 2.1 it was observed that a DSB-SC AM signal required a channel bandwidth of $B_c = 2B$ for transmission, where B is the bandwidth of the message signal $m(t)$. However, the two sidebands are redundant. In this section, it is demonstrated that the transmission of either sideband is sufficient to reconstruct the message signal $m(t)$ at the receiver. Thus, the bandwidth of the transmitted signal is reduced to that of the message signal $m(t)$.

It can be demonstrated by the use of the Fourier transform that a single-sideband (SB) AM signal can be represented mathematically as

$$u(t) = A_c m(t) \cos 2\pi f_c t \mp A_c \hat{m}(t) \sin 2\pi f_c t \tag{7}$$

where $\hat{m}(t)$ is the Hilbert transform of $m(t)$ and the plus-or-minus sign determines which sideband we obtain (+ for the lower sideband and − for the upper sideband). The Hilbert transform may be viewed as a linear filter with impulse response $h(t) = 1/\pi t$ and frequency response

$$H(f) = \begin{cases} -j, & f > 0 \\ j, & f < 0 \\ 0, & f = 0 \end{cases} \tag{8}$$

Therefore, the SSB AM signal $u(t)$ may be generated by using the system configuration shown in Fig. 5.

The method shown in Fig. 5 for generating a SSB AM signal is one that employs a Hilbert transform filter. Another method, illustrated in Fig. 6, generates a DSB-SC AM signal and then employs a filter that selects either the upper sideband or the lower sideband of the double-sideband AM signal.

To recover the message signal $m(t)$ in the received SSB AM signal, a phase coherent or synchronous demodulator

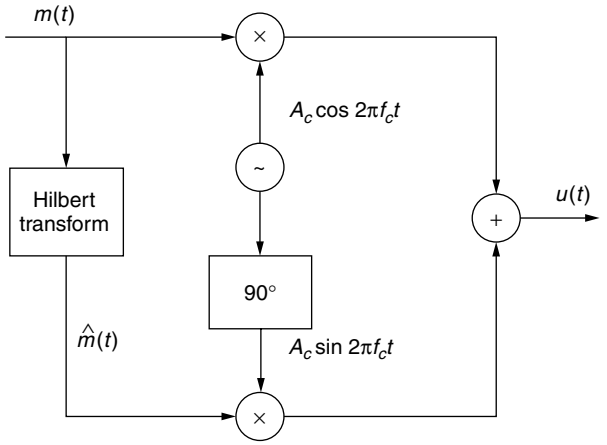


Figure 5. Generation of a single-sideband signal.

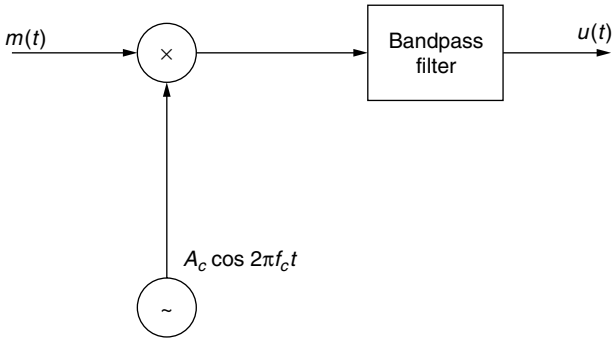


Figure 6. Generation of a single-sideband AM signal by filtering one of the sidebands of a DSB-SC AM signal.

is required, as was the case of DSB-SC AM signals. Thus, for the received SSB signal as given in (7), when demodulated by multiplying $r(t)$ with a sinusoid that has a phase offset ϕ, we obtain

$$\begin{aligned} r(t)\cos(2\pi f_c t + \phi) &= u(t)\cos(2\pi f_c t + \phi) \\ &= \tfrac{1}{2}A_c m(t)\cos\phi \pm \tfrac{1}{2}A_c \hat{m}(t)\sin\phi \\ &\quad + \text{double-frequency terms} \end{aligned} \tag{9}$$

By passing the product signal in (9) through an ideal lowpass filter, the double-frequency components are eliminated, leaving us with

$$y(t) = \tfrac{1}{2}A_c m(t)\cos\phi \pm \tfrac{1}{2}A_c \hat{m}(t)\sin\phi \tag{10}$$

Note that the effect of the phase offset not only is to reduce the amplitude of the desired signal $m(t)$ by $\cos\phi$ but also results in an undesirable sideband signal due to the presence of $\hat{m}(t)$ in $y(t)$. The latter component was not present in a DSB-SC signal and, hence, it was not a factor. However, it is an important element that contributes to the distortion of the demodulated SSB signal.

The transmission of a pilot tone at the carrier frequency is a very effective method for providing a phase-coherent reference signal for performing synchronous demodulation at the receiver. Thus, the undesirable sideband signal component is eliminated. However, this means that a portion of the transmitted power must be allocated to the transmission of the carrier.

The spectral efficiency of SSB AM makes this modulation method very attractive for use in voice communications over telephone channels (wirelines and cables). In this application, a pilot tone is transmitted for synchronous demodulation and shared among several channels.

The filter method shown in Fig. 6 for selecting one of the two signal sidebands for transmission is particularly difficult to implement when the message signal $m(t)$ has a large power concentrated in the vicinity of $f = 0$. In such a case, the sideband filter must have an extremely sharp cutoff in the vicinity of the carrier in order to reject the second sideband. Such filter characteristics are very difficult to implement in practice.

2.4. Vestigial-Sideband AM

The stringent frequency-response requirements on the sideband filter in a SSB AM system can be relaxed by allowing a part, called a vestige, of the unwanted sideband to appear at the output of the modulator. Thus, the design of the sideband filter is simplified at the cost of a modest increase in the channel bandwidth required to transmit the signal. The resulting signal is called *vestigial-sideband* (VSB) *AM*.

To generate a VSB AM signal we begin by generating a DSB-SC AM signal and passing it through a sideband filter with frequency response $H(f)$ as shown in Fig. 7. In the time domain the VSB signal may be expressed as

$$u(t) = [A_c m(t)\cos 2\pi f_c t] * h(t) \tag{11}$$

where $h(t)$ is the impulse response of the VSB filter and the asterisk denotes convolution. In the frequency domain, the corresponding expression is

$$U(f) = \frac{A_c}{2}[M(f - f_c) + M(f + f_c)]H(f) \tag{12}$$

To determine the frequency-response characteristics of the filter, consider the demodulation of the VSB signal $u(t)$. Multiply $u(t)$ by the carrier component $\cos 2\pi f_c t$ and pass the result through an ideal lowpass filter, as shown in Fig. 8. Thus, the product signal is

$$v(t) = u(t)\cos 2\pi f_c t$$

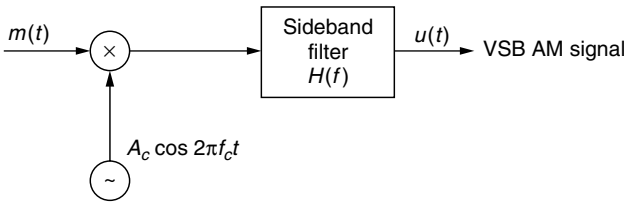


Figure 7. Generation of a VSB AM signal.

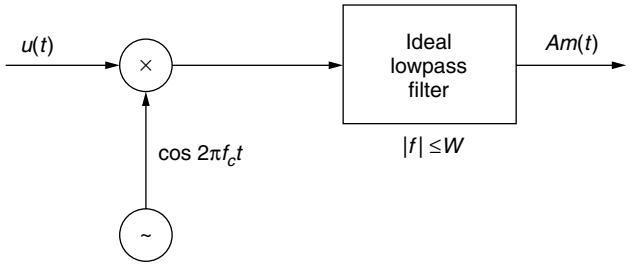


Figure 8. Demodulation of VSB signal.

or, equivalently

$$V(f) = \tfrac{1}{2}[U(f - f_c) + U(f + f_c)] \tag{13}$$

When $U(f)$ is substituted from (12) into (13), the result is

$$\begin{aligned} V(f) = {} & \frac{A_c}{4}[M(f - 2f_c) + M(f)]H(f - f_c) \\ & + \frac{A_c}{4}[M(f) + M(f + 2f_c)]H(f + f_c) \end{aligned} \tag{14}$$

The lowpass filter rejects the double-frequency terms and passes only the components in the frequency range $|f| \leq B$. Hence, the signal spectrum at the output of the ideal lowpass filter is

$$V_o(f) = \frac{A_c}{4}M(f)[H(f - f_c) + H(f + f_c)] \tag{15}$$

We require that the message signal at the output of the lowpass filter be undistorted. Hence, the VSB filter characteristic must satisfy the condition

$$H(f - f_c) + H(f + f_c) = \text{constant}, \quad |f| \leq B \tag{16}$$

This condition is satisfied by a filter that has the frequency-response characteristic shown in Fig. 9. We note that $H(f)$ selects the upper sideband and a vestige of the lower sideband. It has odd symmetry about the carrier frequency f_c, in the frequency range $f_c - f_a < f < f_c + f_a$, where f_a is a conveniently selected frequency that is some small fraction of B; that is, $f_a \ll B$. Thus, we obtain an undistorted version of the transmitted signal. Figure 10 illustrates the frequency response of a VSB filter that selects the lower sideband and a vestige of the upper sideband.

In practice, the VSB filter is designed to have some specified phase characteristic. To avoid distortion of the message signal, the VSB filter should be designed to have linear phase over its passband $f_c - f_a \leq |f| \leq f_c + B$.

3. IMPLEMENTATION OF AM MODULATORS AND DEMODULATORS

There are several different methods for generating AM modulated signals. We shall describe the methods most commonly used in practice. Since the process of modulation involves the generation of new frequency components, modulators are generally characterized as nonlinear and, or, time-variant systems.

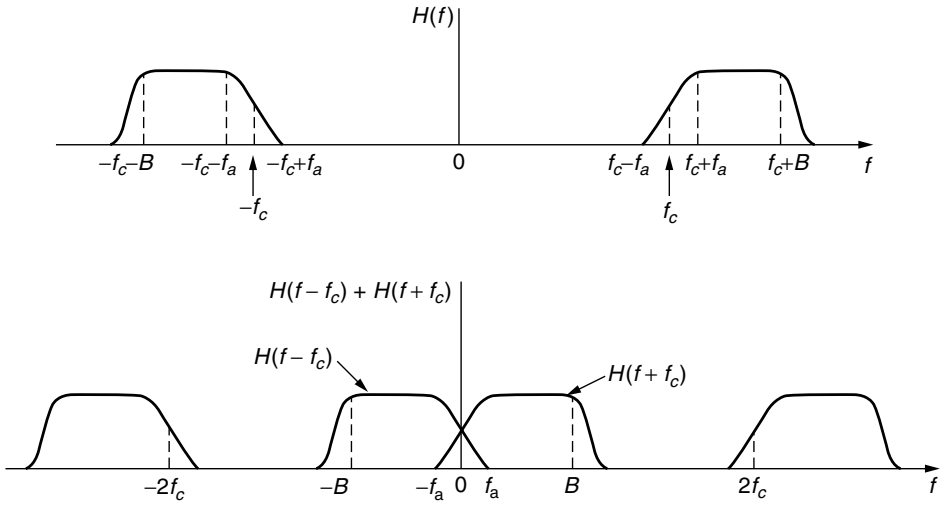


Figure 9. VSB filter characteristics.

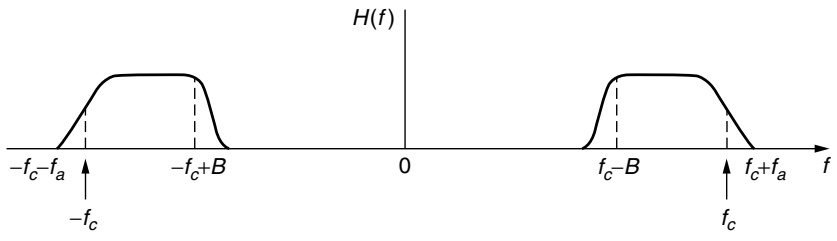


Figure 10. Frequency response of VSB filter for selecting the lower sideband of the message signal.

3.1. Power-Law Modulation

Consider the use of a nonlinear device such as a p-n diode that has a voltage–current characteristic as shown in Fig. 11. Suppose that the voltage input to such a device is the sum of the message signal $m(t)$ and the carrier $A_c \cos 2\pi f_c t$, as illustrated in Fig. 12. The nonlinearity will generate a product of the message $m(t)$ with the carrier, plus additional terms. The desired modulated signal can be filtered out by passing the output of the nonlinear device through a bandpass filter.

To elaborate on this method, suppose that the nonlinear device has an input–output (square-law) characteristic of the form

$$v_0(t) = a_1 v_i(t) + a_2 v_i^2(t) \tag{17}$$

where $v_i(t)$ is the input signal $v_0(t)$ is the output signal, and the parameters (a_1, a_2) are constants. Then, if the input to the nonlinear device is

$$v_i(t) = m(t) + A_c \cos 2\pi f_c t \tag{18}$$

its output is

$$\begin{aligned} v_0(t) &= a_1[m(t) + A_c \cos 2\pi f_c t] \\ &\quad + a_2[m(t) + A_c \cos\cos 2\pi f_c t]^2 \\ &= a_1 m(t) + a_2 m^2(t) + a_2 A_c^2 \cos 2\pi f_c t \\ &\quad + A_c a_1 \left[1 + \frac{2a_2}{a_1} m(t)\right] \cos 2\pi f_c t \end{aligned} \tag{19}$$

The output of the bandpass filter with bandwidth $2B$ centered at $f = f_c$ yields

$$u(t) = A_c a_1 \left[1 + \frac{2a_2}{a_1} m(t)\right] \cos 2\pi f_c t \tag{20}$$

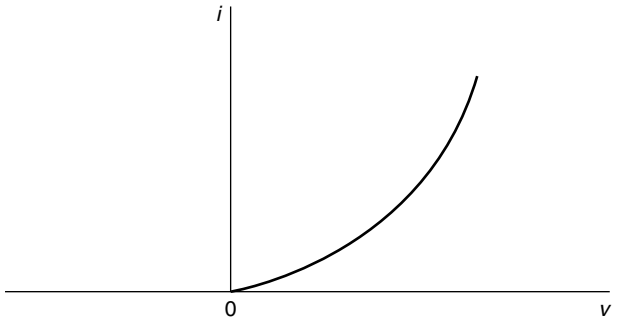


Figure 11. Voltage–current characteristic of p-n diode.

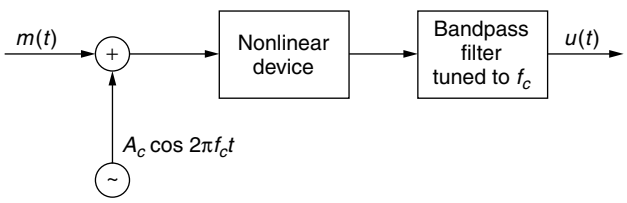


Figure 12. Block diagram of power-law AM modulator.

where $2a_2|m(t)|/a_1 < 1$ by design. Thus, the signal generated by this method is a conventional DSB AM signal.

3.2. Switching Modulator

Another method for generating an AM modulated signal is by means of a switching modulator. Such a modulator can be implemented by the system illustrated in Fig. 13a. The sum of the message signal and the carrier; i.e., $v_i(t)$ given by (18), are applied to a diode that has the input–output voltage characteristic shown in Fig. 13b, where $A_c \gg m(t)$. The output across the load resistor is simply

$$v_0(t) = \begin{cases} v_i(t), & c(t) > 0 \\ 0, & c(t) < 0 \end{cases} \tag{21}$$

This switching operation may be viewed mathematically as a multiplication of the input $v_i(t)$ with the switching function $s(t)$

$$v_0(t) = [m(t) + A_c \cos 2\pi f_c t] s(t) \tag{22}$$

where $s(t)$ is as shown in Fig. 13c.

Since $s(t)$ is a periodic function, it is represented in the Fourier series as

$$s(t) = \frac{1}{2} + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{2n-1} \cos[2\pi f_c t(2n-1)] \tag{23}$$

Hence

$$\begin{aligned} v_0(t) &= [m(t) + A_c \cos 2\pi f_c t] s(t) \\ &= \frac{A_c}{2} \left[1 + \frac{4}{\pi A_c} m(t)\right] \cos 2\pi f_c t + \text{other terms} \end{aligned} \tag{24}$$

The desired AM modulated signal is obtained by passing $v_0(t)$ through a bandpass filter with center frequency $f = f_c$

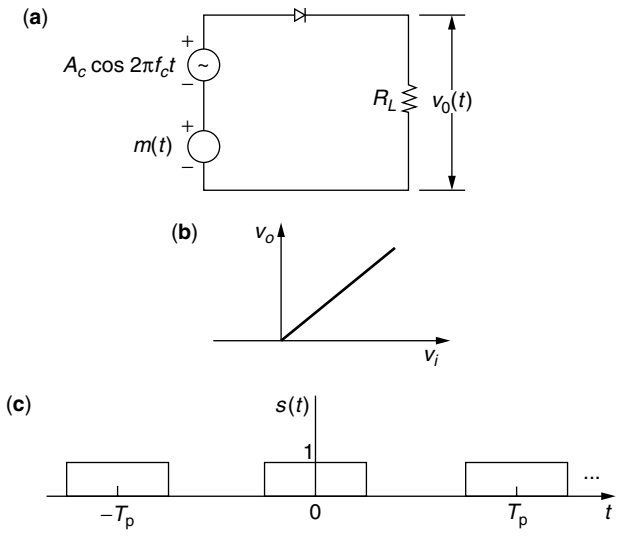


Figure 13. Switching modulator and periodic switching signal.

and bandwidth $2B$. At its output, we have the desired conventional DSB AM signal:

$$u(t) = \frac{A_c}{2}\left[1 + \frac{4}{\pi A_c} m(t)\right] \cos 2\pi f_c t \tag{25}$$

3.3. Balanced Modulator

A relatively simple method for generating a DSB-SC AM signal is to use two conventional AM modulators arranged in the configuration illustrated in Fig. 14. For example, we may use two square-law AM modulators as described above. Care must be taken to select modulators with approximately identical characteristics so that the carrier component cancels out at the summing junction.

3.4. Ring Modulator

Another type of modulator for generating a DSB-SC AM signal is the ring modulator illustrated in Fig. 15. The switching of the diodes is controlled by a square wave of frequency f_c, denoted as $c(t)$, which is applied to the center taps of the two transformers. When $c(t) > 0$, the top and bottom diodes conduct, while the two diodes in the crossarms are off. In this case, the message signal $m(t)$ is multiplied by $+1$. When $c(t) < 0$, the diodes in the crossarms of the ring conduct, while the other two are switched off. In this case, the message signal $m(t)$ is multiplied by -1. Consequently, the operation of the ring modulator may be described mathematically as a multiplier of $m(t)$ by the square-wave carrier $c(t)$:

$$v_0(t) = m(t)c(t) \tag{26}$$

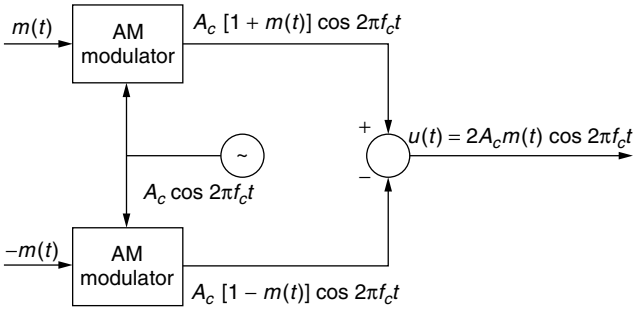


Figure 14. Block diagram of a balanced modulator.

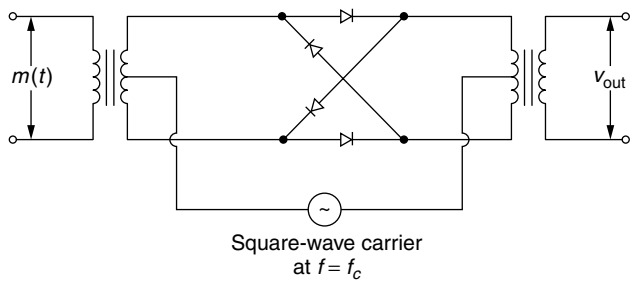


Figure 15. Ring modulator for generating DSB-SC AM signals.

Since $c(t)$ is a periodic function, it is represented by the Fourier series

$$c(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{(-1)n^{-1}}{2n-1} \cos[2\pi f_c(2n-1)t] \tag{27}$$

Hence, the desired DSB-SC AM signal $u(t)$ is obtained by passing $v_0(t)$ through a bandpass filter with center frequency f_c and bandwidth $2B$.

From the discussion above, we observe that the balanced modulator and the ring modulator systems, in effect, multiply the message signal $m(t)$ with the carrier to produce a DSB-SC AM signal. The multiplication of $m(t)$ with $A_c \cos w_c t$ is called a *mixing operation*. Hence, a mixer is basically a balanced modulator.

The method shown in Fig. 5 for generating a SSB signal requires two mixers, specifically, two balanced modulators, in addition to the Hilbert transformer. On the other hand, the filter method illustrated in Fig. 6 for generating a SSB signal requires a single balanced modulator and a sideband filter.

Let us now consider the demodulation of AM signals. We begin with a description of the envelope detector.

3.5. Envelope Detector

As indicated previously, conventional DSB AM signals are easily demodulated by means of an envelope detector. A circuit diagram for an envelope detector is shown in Fig. 16. It consists of a diode and an RC circuit, which is basically a simple lowpass filter.

During the positive half-cycle of the input signal, the diode is conducting and the capacitor charges up to the peak value of the input signal. When the input falls below the voltage on the capacitor, the diode becomes reverse-biased and the input becomes disconnected from the output. During this period, the capacitor discharges slowly through the load resistor R. On the next cycle of the carrier, the diode conducts again when the input signal exceeds the voltage across the capacitor. The capacitor charges up again to the peak value of the input signal and the process is repeated again.

The time-constant RC must be selected so as to follow the variations in the envelope of the carrier-modulated signal. In effect

$$\frac{1}{f_c} \ll RC \ll \frac{1}{B}$$

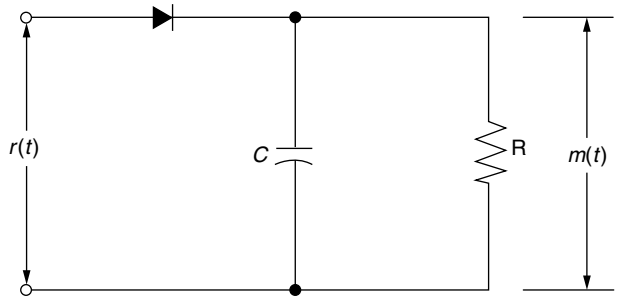


Figure 16. An envelope detector.

In such a case, the capacitor discharges slowly through the resistor, and thus, the output of the envelope detector closely follows the message signal.

3.6. Demodulation of DSB-SC AM Signals

As indicated earlier, the demodulation of a DSB-SC AM signal requires a synchronous demodulator. Thus, the demodulator must use a coherent phase reference, which is usually generated by means of a phase-locked loop (PLL) to demodulate the received signal.

The general configuration is shown in Fig. 17. A PLL is used to generate a phase-coherent carrier signal that is mixed with the received signal in a balanced modulator. The output of the balanced modulator is passed through a lowpass filter of bandwidth B that passes the desired signal and rejects all signal and noise components above B Hz. The characteristics and operation of the PLL are described in Refs. 1–3.

3.7. Demodulation of SSB Signals

The demodulation of SSB AM signals also requires the use of a phase-coherent reference. In the case of signals such as speech, that have relatively little or no power content at d_c, it is straightforward to generate the SSB signal, as shown in Fig. 6, and then to insert a small carrier component that is transmitted along with the message. In such a case we may use the configuration shown in Fig. 18 to demonstrate the SSB signal. We observe that a balanced modulator is used for the purpose of frequency conversion of the bandpass signal to lowpass or baseband.

3.8. Demodulation of VSB Signals

In VSB a carrier component is generally transmitted along with the message sidebands. The existence of the carrier component makes it possible to extract a phase-coherent reference for demodulation in a balanced modulator, as shown in Fig. 18.

In some applications such as TV broadcasting, a large carrier component is transmitted along with the message in the VSB signal. In such a case, it is possible to recover the message by passing the received VSB signal through an envelope detector.

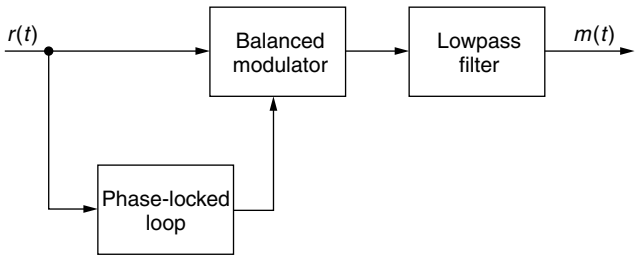


Figure 17. Block diagram of demodulator for DSB-SC AM signals.

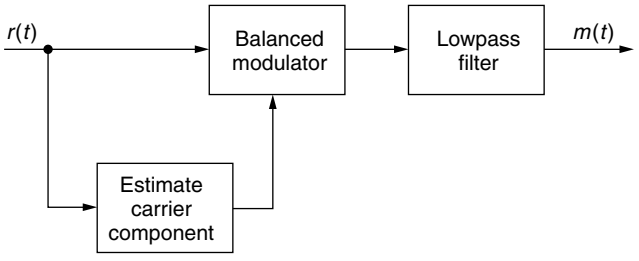


Figure 18. Block diagram of demodulator for SSB AM signal with a carrier component.

4. CONCLUDING REMARKS

This article has covered the different types of amplitude modulation techniques that are used in the transmission of analog signals. Conventional AM is most widely used in radiobroadcasting. It is anticipated that conventional AM in radiobroadcasting will be phased out eventually and replaced by a digital modulation method that provides better signal fidelity. Current analog television broadcasting will also undergo a similar transformation. Single-sideband AM was widely used in telephone systems for audio signal transmission over many decades. However, today, audio signal transmission in the telephone systems is performed by digital modulation after the voice signals are converted to digital form by use of pulse code modulation (PCM) or differential PCM (DPCM).

Treatments of AM modulation can be found in many undergraduate-level textbooks in communication systems. References 4–6 are cited as typical.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the

author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. W. C. Lindsey, *Synchronization Systems in Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1972.
2. F. M. Gardner, *Phaselock Techniques*, Wiley, New York, 1979.
3. W. C. Lindsey and C. M. Chie, A survey of digital phase-locked loops, *Proc. IEEE* **69**: 410–432 (1981).
4. H. Taub and D. L. Schilling, *Principles of Communication Systems*, McGraw-Hill, New York, 1971.
5. J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
6. H. Stark, F. B. Tuteur and J. B. Anderson, *Modern Electrical Communications*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1988.

ANTENNA ARRAYS

JOHN N. SAHALOS
Radiocommunications Laboratory
Aristotle University of Thessaloniki
Thessaloniki, Greece

1. INTRODUCTION

Antennas occupy a palmary position in radiocommunication systems. It is not an overemphasis to say that antennas have become ubiquitous devices. This has occurred because radio and TV as well satellite and mobile communications have experienced the largest growth among the industry systems. The strongest economic and social impact nowadays is coming from cellular telephony, personal communications, and satellite navigation systems. All these applications have served on the motivation for engineers to achieve elegant antennas to be incorporated into handy and portable systems.

Many textbooks provide in-depth resources on antennas. Especially on antenna arrays there are digests and books containing extensive data and techniques. The references cited here [1–20] include some of the most well known and recommended books.

A device able to receive or transmit the electromagnetic energy is called an antenna, which consists of one or more elements. A single-element antenna is usually not enough to cover the technical needs. That happens because its performance is limited. A set of discrete elements made up of an antenna array offers the solution to the transceiving of electromagnetic energy. An antenna array is characterized by the geometry and the type of the elements. A major role in the antenna array is played by the mutual coupling between the elements and their input impedance. For simplicity reasons in both the fabrication and the synthesis, the elements are chosen, if it is possible, to be identical and parallel. For the same reasons uniformly spaced linear arrays are mostly encountered in practice.

In the following paragraphs the properties of various antenna arrays will be presented and the synthesis method will be discussed.

2. ANTENNA ARRAY FACTOR

The radiation characteristics of antennas have to do with the far-field region. In this region the field is separated in two parts: one containing the distance r of the observation point (receiver location) and the other, its spherical coordinate angles θ and φ. The far electric field of a typical antenna element (see Fig. 1) can be expressed as

$$\boldsymbol{E}_n(r) \cong -j\omega\mu \frac{e^{-j\beta r}}{4\pi r} \boldsymbol{f}_n(\theta, \varphi) \tag{1}$$

The angular — dependent vector $\boldsymbol{f}_n(\theta, \varphi)$ gives the directional characteristics of the nth-element electric field [11]:

$$\boldsymbol{f}_n(\theta, \varphi) = (\hat{\theta}\hat{\theta} + \hat{\varphi}\hat{\varphi}) \cdot \int_{\text{element}} \boldsymbol{J}_n(\mathbf{r'_n}) e^{j\beta \hat{r} \cdot (\mathbf{r_n} - \mathbf{r'_n})} dv' \tag{2}$$

where $\boldsymbol{J}_n(r'_n)$ = electric current density of the nth element
r'_n = distance of a source point from the origin

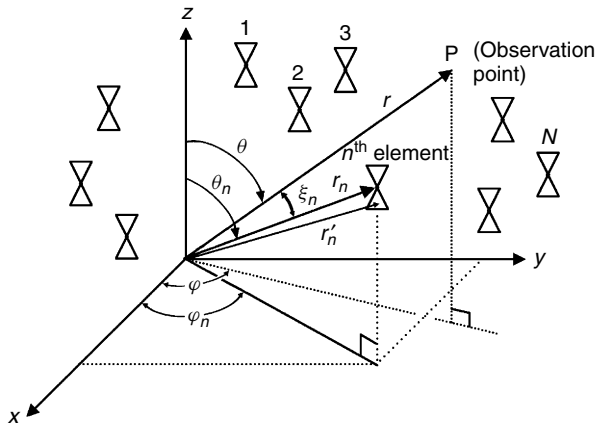


Figure 1. Geometry of a general antenna array.

r = distance of the observation point from the origin
$\beta = \dfrac{2\pi}{\lambda}$, the free-space wavenumber
ω = angular frequency
μ = magnetic permeability of the space

The total electric field of the N-element antenna array is

$$\boldsymbol{E}(r) = \sum_{n=1}^{N} \boldsymbol{E}_n(r) \tag{3}$$

Also the total magnetic field is [6]

$$\boldsymbol{H}(r) = \frac{1}{\eta}\hat{\boldsymbol{r}} \times \boldsymbol{E}(r) \tag{4}$$

where $\eta = \sqrt{\mu/\varepsilon}$ (ε is the electric permeability).

For identical — and, if possible, identically oriented — elements, the current distribution of these elements is approximately the same except for a constant complex multiplier. In Eq. (1), $\boldsymbol{f}_n(\theta, \varphi)$ can be expressed as

$$\boldsymbol{f}_n(\theta, \varphi) = I_n \boldsymbol{f}(\theta, \varphi) \tag{5}$$

$\boldsymbol{f}(\theta, \varphi)$ is called the pattern function of the element and I_n is the complex excitation of the nth element of the array.

Combining Eqs. (1), (2), and (5), we have

$$\boldsymbol{E}(r) = -j\omega\mu \frac{e^{-j\beta r}}{4\pi r} \boldsymbol{f}(\theta, \varphi) \sum_{n=1}^{N} I_n e^{j\beta r_n \cos \xi_n} \tag{6}$$

where $(r_n,\ \theta_n,\ \varphi_n)$ are the spherical coordinates of a convenient reference point of the nth element and $\cos \xi_n = \sin\theta \sin\theta_n \cos(\varphi - \varphi_n) + \cos\theta \cos\theta_n$. The last term of (6) is expressed separately as

$$\mathrm{AF}(\theta, \varphi) = \sum_{n=1}^{N} I_n e^{j\beta r_n \cos \xi_n} \tag{7}$$

$\mathrm{AF}(\theta, \varphi)$ is the array factor. This factor is actually the array pattern of N isotropic point sources positioned at the reference points of the elements of the original array.

From Eqs. (6) and (7) we know that

$$\boldsymbol{E}(r) = -j\omega\mu \frac{e^{-j\beta r}}{4\pi r} \boldsymbol{f}(\theta, \varphi)\mathrm{AF}(\theta, \varphi) \tag{8}$$

This expression states the following pattern multiplication principle: "An array consisting of identically oriented similar elements has a pattern which can be expressed as the product of the element pattern and the array factor."

An antenna engineer must select the element according to the technical requirements. Since the element pattern is known, the design effort is mainly directed to the array factor.

3. ELEMENTS AND ARRAY TYPES

Element types of antenna arrays can be found in the literature [1–14]. Monopoles, dipoles, loops, slots, microstrip patches, and horns are the most common types. More recent studies and innovations have resulted in new types of elements, such as the monolithic, the superconducting, the active, and the electronically and functionally small elements.

In parallel with the development of elements, antenna arrays have experienced a tremendous growth. Their list starts from the linear broadside and endfire arrays, the planar, the circular, and the conformal and goes up to the adaptive arrays. Some of the more recent types are flat-plate slot arrays, digital beamforming, dichroic, slotted, and fractal arrays.

As mentioned above, antenna analysis and synthesis focus mostly on the array factor. Consequently, in the following paragraphs we devote our analysis mainly to this factor.

4. ANTENNA CHARACTERISTICS AND INDICES

One of the main characteristics of an antenna is its radiation pattern. This characteristic graphically presents the radiation properties and can be measured by moving a probe antenna around the antenna under test at a constant distance in the far field (see Fig. 2a). The response as a function of the angular coordinates constitutes the radiation pattern. Depending on the probe type and orientation, the appropriate component of the electric or the magnetic field can be measured. If the probe is moved over the spherical surface, its terminal voltage will present the 3D radiation pattern. A pattern taken in one plane is known as the *plane pattern*. The pattern that contains the electric field vector is the E-plane pattern, while the pattern that contains the magnetic field vector is the H plane. The above two are referred as the *principal plane patterns*. As an example, Fig. 2b presents the 3D radiation pattern of an electric dipole while Figs. 2c and 2d show the E- and H-plane pattern.

Polarization of an antenna is the polarization of the wave transmitted by the antenna. Polarization has to do with a certain direction. If the direction is not stated, then it is assumed that it corresponds in the direction of maximum. The polarization is characterized by the curve traced by the endpoint of the arrow representing the instantaneous electric field. The field is observed in the direction of propagation. Polarization is classified as linear, circular, or elliptical.

For the sake of convenience, some of the antenna indices will be defined as follows:

1. The directive gain $D(\theta_0,\ \varphi_0)$ is a dimensionless quantity that is defined by

$$D(\theta_0, \varphi_0) = \frac{\text{radiation intensity for the direction } (\theta_0, \varphi_0)}{\dfrac{1}{4\pi}(\text{radiation power of the antenna})} \tag{9}$$

Radiation intensity is the power radiated in a given direction per unit solid angle. The maximum $D(\theta_0, \varphi_0)$ is the directivity D of the antenna.

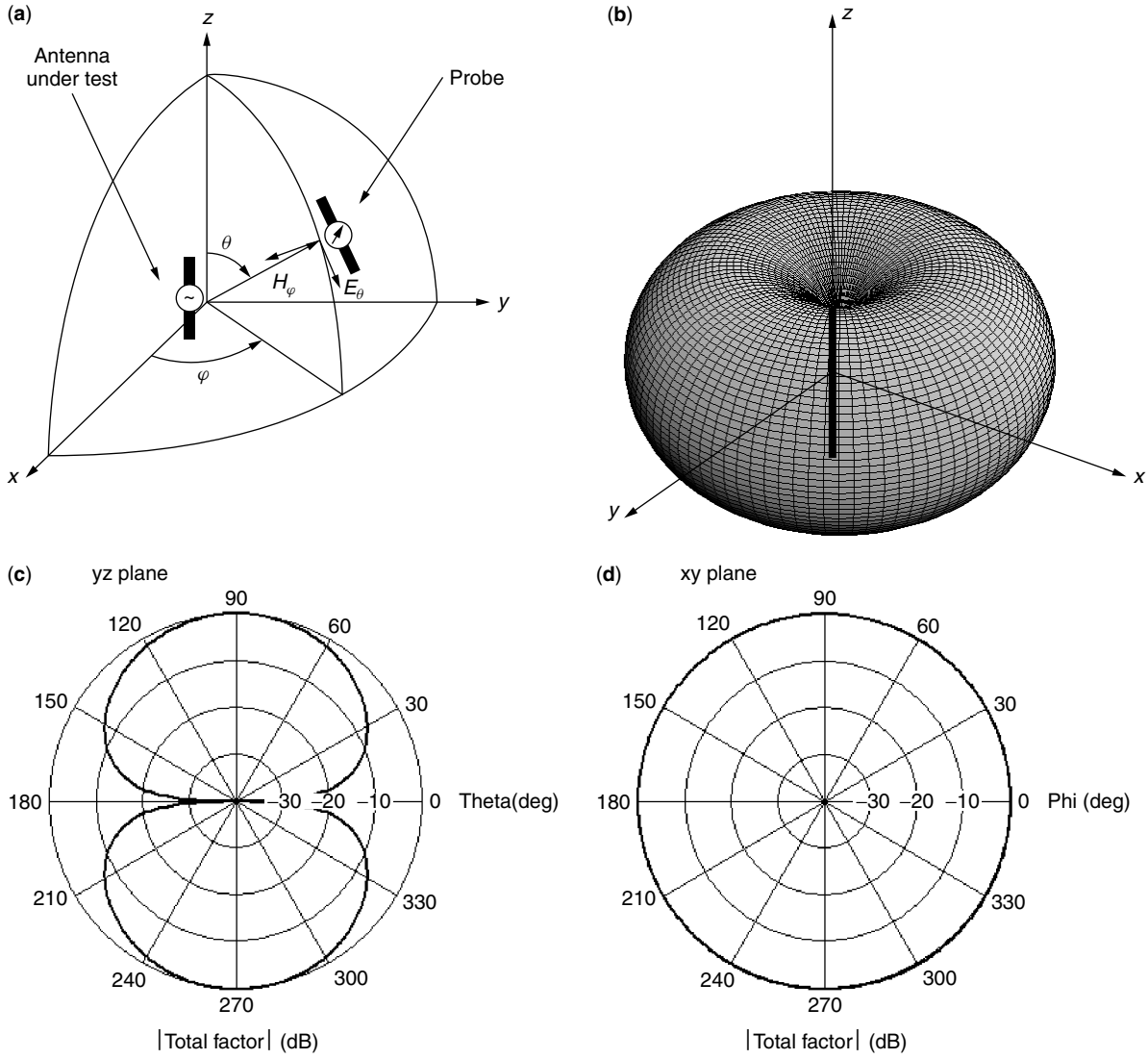


Figure 2. Radiation pattern of an electric $\lambda/2$ dipole: (**a**) pattern measurement scheme; (**b**) three-dimensional plot; (**c**) E-plane radiation pattern; (**d**) H-plane radiation pattern.

2. The power gain $G(\theta_0, \varphi_0)$ is defined by

$$G(\theta_0, \varphi_0) = \frac{\text{radiation intensity for the direction } (\theta_0, \varphi_0)}{\frac{1}{4\pi}(\text{power input to the antenna})} \quad (10)$$

3. The signal-to-noise ratio SNR applies for the receiving antennas and is defined by

$$\text{SNR} = \frac{\text{received power of the desired signal}}{\frac{1}{4\pi}(\text{noise plus interference power})} \quad (11)$$

4. The radiation efficiency η of an N-element array antenna is defined by

$$\eta = \frac{\text{radiation intensity to the direction of maximum}}{N\,(\text{sum of the excitation current magnitude squared})} \quad (12)$$

5. The quality factor Q of an N-element array antenna is defined by

$$Q = \frac{\text{sum of the excitation current magnitude squared}}{\frac{1}{4\pi}\,(\text{power input to the array})} \quad (13)$$

Combining (13) and (12), we find that

$$G_{\max} = N\eta Q \quad (14)$$

6. The half-power (or 3-dB power) beamwidth (HPBW) is the angular width between the angular points half-power (3 dB) below that of the main beam maximum of the antenna.
7. The first null beamwidth (BW_{null}) is defined as the angular width between the first zero crossing of either side of the main-beam maximum of the antenna.

8. The bandwidth of an antenna is defined by the frequency limits at which the maximum gain is reduced to half-power (3 dB). The fractional bandwidth is given by $\Delta f/f$.
9. The sidelobe level (SLL) is the ratio of the pattern value of a sidelobe peak to the corresponding value of the mainlobe. Usually SLL in an antenna is defined as the largest sidelobe level for the whole pattern.

5. LINEAR ARRAYS

One method to obtain directive antennas is to use several individual antennas that add their contributions in preferred directions and cancel in others. This arrangement is known as an *array*, and the individual antennas are called *elements*. The most practical array that consists of a number of elements set up along a straight line is the linear array.

Consider a typical linear array placed along the z axis as shown in Fig. 3. The array factor $\mathrm{AF}(\theta, \varphi)$ depends only on the angle θ and is written as

$$\mathrm{AF}(\theta) = \sum_{n=0}^{N} I_n e^{j\beta d_n \cos\theta} \tag{15}$$

If the elements are placed in the same interelement distance d, then Eq. (15) yields

$$\mathrm{AF}(\theta) = \sum_{n=0}^{N} I_n e^{j\beta n d \cos\theta} = \sum_{n=0}^{N} I_n z^n \tag{16}$$

where

$$z = e^{j\beta d \cos\theta} \tag{17}$$

For $0 \le \theta \le \pi$, AF (θ) is a polynomial of z that moves on a unit circle with a phase bounded between $-\beta d$ and $+\beta d$. The bounded region is called the *visible region*. The unit circle approach, proposed by Schelkunoff [1] visually indicates how the element contributions combine.

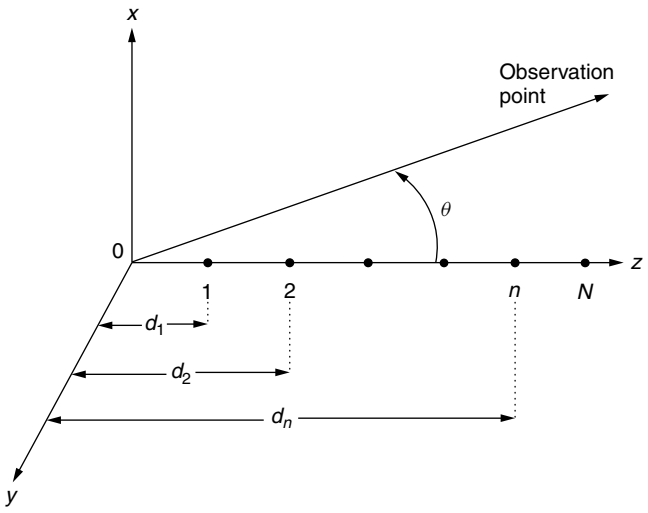


Figure 3. A linear N-element array.

A linear array that cancels noise from N different directions has a pattern of the following form:

$$\mathrm{AF}(\theta) = C\prod_{n=1}^{N}(z - z_n) = C\sum_{k=0}^{N} I_k z^k \tag{18}$$

C is the normalization factor such that $|\mathrm{AF}(\theta)|_{\max} = 1; z_n = e^{j\beta d \cos\theta_n}$, where θ_n is the direction of the nth interference, and I_k is the required excitation of the kth element.

The type and orientation of the elements of the array are selected for receiving the maximum of the desired signal. By varying z_n, it is possible to steer the nulls. This will alter the element excitations. Also some roots can be outside the visible region or the unit circle.

A linear array with all roots equal is known as a *binomial array*. $\mathrm{AF}(\theta)$ is of the form

$$\mathrm{AF}(\theta) = C(z - z_1)^N = C\sum_{k=0}^{N}\binom{N}{k}(-z_1)^k z^{N-k} \tag{19}$$

where

$$\binom{N}{k} = \frac{N!}{(N-k)!k!} \tag{20}$$

The binomial array has low minor lobes. However a wide and difficult to be realized variation between the amplitudes of the elements is present. This variation increases as the number of elements increases.

5.1. Uniform Arrays

Linear arrays with equally spaced elements, identical magnitude and progressive phase, are referred to as *uniform arrays*. For N elements we have $I_k = (e^{i\alpha})^k$ and Eq. (18) becomes

$$\mathrm{AF}(\theta) = C\sum_{n=0}^{N-1}(ze^{j\alpha})^n = C\frac{(ze^{j\alpha})^N - 1}{(ze^{j\alpha}) - 1} \tag{21}$$

According to [6], Eq. (21) may be transformed to

$$\mathrm{AF}(\theta) = e^{j(N-1)\psi/2}\frac{\sin(N\psi/2)}{N\sin(\psi/2)} \tag{22}$$

where

$$\psi = \beta d\cos\theta + \alpha \tag{23}$$

$|\mathrm{AF}(\theta)|$ as a function of ψ, known also as $|F(\psi)|$, is presented for a few values of N in Fig. 4.

The main characteristics of $F(\psi)$ are the following:

1. Maximum values occur at $\psi = \pm 2k\pi$, where $k = 0, 1, 2, \ldots$.
2. Nulls of the array are at $\psi = \pm 2k\pi/N$ where $k = 1, 2, 3, \ldots$ and $k \ne N, 2N, 3N, \ldots$.
3. The 3-dB points of the array factor are at ψ, which gives

$$F(\psi) = \pm\frac{\sqrt{2}}{2} \Rightarrow |F(\psi)|^2 = \frac{1}{2}$$
$$\Rightarrow 10\log|F(\psi)|^2 = 3 \text{ dB}$$

Therefore, $N\psi/2 = \pm 1.391$ rad.

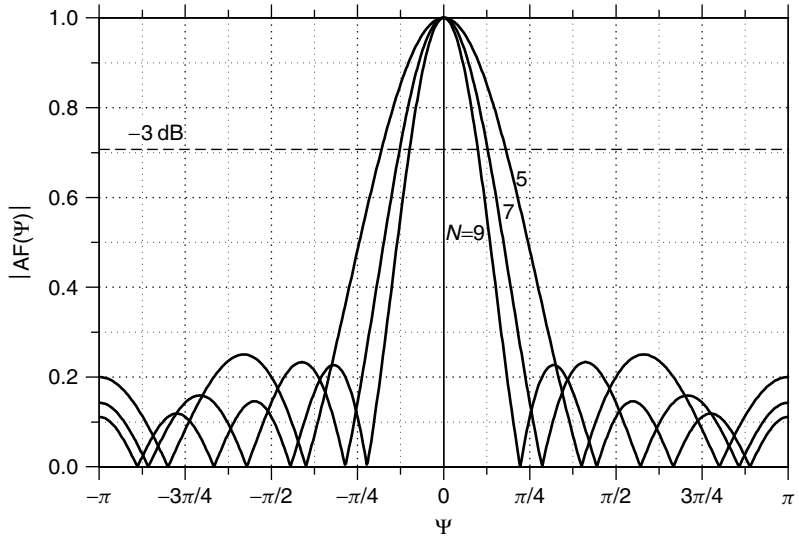


Figure 4. The magnitude of the array factor as a function of ψ.

4. Secondary maxima (minor lobes) occur when $\psi = (2k+1)\pi/N$, where $k = 1, 2, 3, \ldots$.

It is noticed that the first minor lobe is at $\psi = \pm 3\pi/N$, which gives

$$|F(\psi)| = \frac{1}{N \sin \dfrac{3\pi}{2N}} = \text{SLL}$$

When the beam maximum appears at $\theta = \pi/2$, the array is called *broadside*. *Endfire* is an array with the beam maximum at $\theta = 0$ or π. The array beam can be steered at any direction θ_0 if the phase shift is $\alpha = -\beta d \cos\theta_0$. For the endfire array it is $\alpha = \mp\beta d$. In addition to the ordinary endfire, there is the Hansen–Woodyard (HW) endfire array, which is more directive [Eq. (6)]. In HW the progressive phase shift is

$$\alpha = \mp\left(\beta d + \frac{2.94}{N}\right) \cong \mp\left(\beta d + \frac{\pi}{N}\right) \tag{24}$$

Also

$$|\psi| = \pi \begin{cases} \text{for } \theta = \pi \text{ if maximum occurs at } \theta = 0 \\ \text{for } \theta = 0 \text{ if maximum occurs at } \theta = \pi \end{cases} \tag{25}$$

Condition (25) gives

$$d = \frac{\lambda}{4}\left(1 - \frac{2.94}{\pi N}\right) \cong \frac{\lambda}{4}\left(1 - \frac{1}{N}\right) \tag{26}$$

HW is useful for very long arrays with small interelement distance. Useful formulas for the prescribed linear arrays are given in Table 1.

5.2. Chebyshev Arrays

5.2.1. Chebyshev Polynomials. Chebyshev arrays are uniformly spaced linear arrays with nonuniform excitation. They make use of the Chebyshev polynomials [21].

A Chebyshev polynomial $T_m(x)$ of an independent variable x is an orthogonal one of mth order. It contains equal ripples in the region $-1 \le x \le 1$ and the amplitude varies between $+1$ and -1. The polynomial outside this region rises exponentially:

$$T_m(x) = \begin{cases} \cos(m \cos^{-1} x) & |x| \le 1 \\ \left(\dfrac{x}{|x|}\right)^m \cosh(m \cosh^{-1}|x|) & |x| > 1 \end{cases} \tag{27}$$

and

$$\left.\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \end{aligned}\right\} \tag{28}$$

Equation (28) and the recursion relation

$$T_m(x) = 2xT_{m-1}(x) - T_{m-2}(x) \tag{29}$$

create the Chebyshev polynomials of any order.

5.2.2. Dolph–Chebyshev Arrays. Dolph [22] recognized that Chebyshev polynomials could be used to have arrays with maximum directivity for a given sidelobe level. The equal ripples of the polynomials present the sidelobes, while the main beam comes from the exponential increase beyond $|x| = 1$.

The linear array is fed symmetrically about the centerline (see Fig. 5). The array factor can be expressed in terms of $\cos(\psi/2)$, where $\psi = \beta d \cos\theta + \alpha$. The independent variable of the Chebyshev polynomial is

$$x = x_0 \cos\left(\frac{\psi}{2}\right) \tag{30}$$

At $x = x_0$ the polynomial has its maximum value R:

$$T_m(x_0) = R \text{ or } x_0 = \cosh\left(\frac{1}{m}\cosh^{-1} R\right) \tag{31}$$

The zeros of $T_m(x)$ are at

$$x_k = \pm \cos\frac{(2k-1)\pi}{2m} \qquad k = 1, 2, \ldots, m \tag{32}$$

Table 1. Useful Formulas for Uniform Linear Arrays

	Broadside	Endfire	Hansen–Woodyard Endfire	Intermediate with Maximum at $\theta = \theta_0$
Directivity	$\sim 2N\mathrm{d}/\lambda$	$\sim 4N\mathrm{d}/\lambda$	$\sim 1.789\,(4N\mathrm{d}/\lambda)$	Depending on θ_0
HPBW	$\pi - 2\cos^{-1}\left(\dfrac{2.782}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{2.782}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{0.2796\pi}{N\beta d}\right)$	$\left\lvert\cos^{-1}\left(\cos\theta_0 + \dfrac{2.782}{N\beta d}\right) - \cos^{-1}\left(\cos\theta_0 - \dfrac{2.782}{N\beta d}\right)\right\rvert$
Beamwidth between nulls	$\pi - 2\cos^{-1}\left(\dfrac{2\pi}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{2\pi}{N\beta d}\right)$	$2\cos^{-1}\left(1 - \dfrac{\pi}{N\beta d}\right)$	$\left\lvert\cos^{-1}\left(\cos\theta_0 + \dfrac{2\pi}{N\beta d}\right) - \cos^{-1}\left(\cos\theta_0 - \dfrac{2\pi}{N\beta d}\right)\right\rvert$
Null angular positions	$\cos^{-1}\left(\pm\dfrac{2k\pi}{N\beta d}\right)$	$\cos^{-1}\left(1 - \dfrac{2k\pi}{N\beta d}\right)$	$\cos^{-1}\left[1 + (1-2k)\dfrac{\pi}{N\beta d}\right]$ $k = 1, 2, 3, \ldots, \quad k \neq N, 2N, 3N, \ldots$	$\cos^{-1}\left(\cos\theta_0 \pm \dfrac{2k\pi}{N\beta d}\right)$
Sidelobe maximum positions	$\cos^{-1}\left(\pm\dfrac{(2k+1)\pi}{N\beta d}\right)$	$\cos^{-1}\left(1 - \dfrac{(2k+1)\pi}{N\beta d}\right)$	$\cos^{-1}\left(1 - \dfrac{2k\pi}{N\beta d}\right)$ $k = 1, 2, 3, \ldots, \quad k \neq N, 2N, 3N, \ldots$	$\cos^{-1}\left(\cos\theta_0 \pm \dfrac{(2\ k+1)\pi}{N\beta d}\right)$

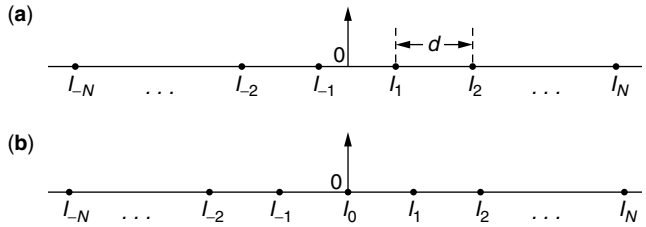


Figure 5. Linear array with (**a**) even and (**b**) odd number of elements, uniformly spaced and symmetrically excited ($I_k = I_{-k}$).

which correspond to

$$\psi_k = \pm 2 \cos\left(\frac{x_k}{x_0}\right) \tag{33}$$

The excitations I_k are found from (18) by using $z_k = e^{j\psi_k}$. The order of the Chebyshev polynomial is one less than the total number of elements of the array.

An example of a nine-element array with SLL $= -25$ dB is given. The polynomial is the $T_8(x)$. For $R = 25$ dB it is $R = 10^{25/20} = 17.7828$ and from (31), $x_0 = 1.1013$. A broadside array has the excitation coefficients presented in Table 2.

An intermediate array with maximum at $\theta = \theta_0$ has the same amplitude as before with a phase shift $\alpha = -\beta d \cos\theta_0$. Figure 6 shows the patterns for $d/\lambda = 0.5$ of a broadside and an intermediate array with $\theta_0 = \pi/3$.

5.2.3. Riblet Arrays. Dolph–Chebyshev arrays are suitable for $d \geq \lambda/2$ and fail to give the optimum design for $d < \lambda/2$. Riblet [23] devised a method to overcome the problem. He used $(2m + 1)$ elements and an independent variable of the form

$$x = a \cos\psi + b \tag{34}$$

The polynomial is $T_m(x)$ with a visible region $-1 \leq x \leq x_0$:

$$\left.\begin{aligned} x_0 &= a + b \\ -1 &= a \cos\beta d + b \end{aligned}\right\} \tag{35}$$

Solving (35), we obtain

$$a = \frac{1 + x_0}{1 - \cos\beta d} \quad \text{and} \quad b = -\frac{1 + x_0 \cos\beta d}{1 - \cos\beta d} \tag{36}$$

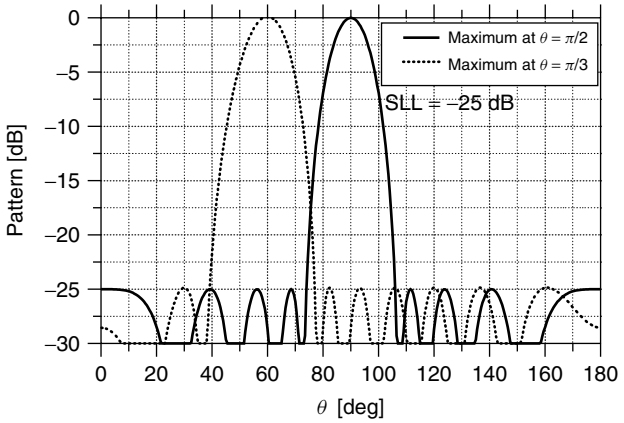


Figure 6. Radiation patterns of a Dolph–Chebyshev array with nine elements in $d/\lambda = 0.5$ with max. at $\theta = \pi/2$ and $\theta = \pi/3$.

Zeros of $T_m(x)$ are given by (32) and the corresponding ψ_k are

$$\psi_k = \pm \cos^{-1}\left(\frac{x_k - b}{a}\right) \tag{37}$$

A nine-element array with $d/\lambda = 0.4$ and SLL $= -20$ dB is presented. The excitation is given in Table 3 and the pattern, in Fig. 7.

5.2.4. Other Chebyshev Arrays. Equal-sidelobe designs with the help of Chebyshev polynomials can be made by

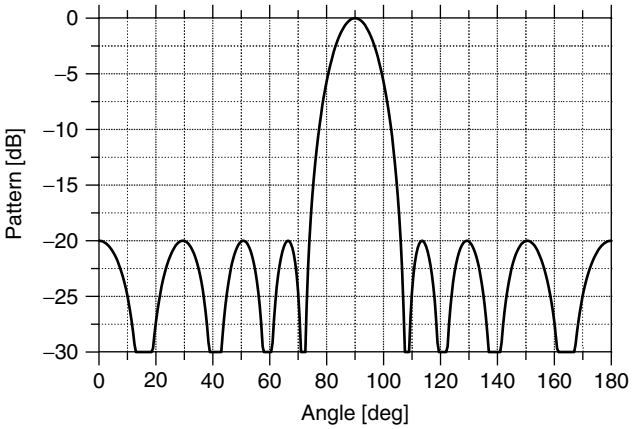


Figure 7. Pattern of a nine-element Riblet array with $d/\lambda = 0.4$ and SLL $= -20$ dB.

Table 2. Normalized Excitation Coefficients for a Nine-Element Dolph–Chebyshev Array

Element number	1, 9	2, 8	3, 7	4, 6	5
Excitation coefficient	3.783×10^{-1}	5.310×10^{-1}	7.639×10^{-1}	9.363×10^{-1}	1

Table 3. Excitation Coefficient of Nine-Element Riblet Array with $d/\lambda = 0.4$

Element number	1, 9	2, 8	3, 7	4, 6	5
Excitation coefficient	5.519×10^{-1}	8.913×10^{-2}	8.393×10^{-1}	1.566×10^{-1}	1

extending the Riblet technique, where x is written as

$$x = a\cos(\beta d\cos\theta + \alpha) + b \tag{38}$$

To find a, b, and α, three characteristic angles θ_1, θ_2 and θ_3, which correspond to x_1, x_2, and x_3, respectively, must be defined.

$$\left.\begin{aligned} x_1 &= a\cos(\beta d\cos\theta_1 + \alpha) + b \\ x_2 &= a\cos(\beta d\cos\theta_2 + \alpha) + b \\ x_3 &= a\cos(\beta d\cos\theta_3 + \alpha) + b \end{aligned}\right\} \tag{39}$$

Solving this equation, one can find

$$\tan\alpha = \frac{\sin y_{21} - \lambda\sin y_{31}}{\lambda\cos y_{31} - \cos y_{21}} \tag{40}$$

where

$$\left.\begin{aligned} y_{21} &= \beta\frac{d}{2}(\cos\theta_2 + \cos\theta_1) \\ y_{31} &= \beta\frac{d}{2}(\cos\theta_3 + \cos\theta_1) \\ \lambda &= \frac{(x_2 - x_1)\sin\left[\dfrac{\beta d}{2}(\cos\theta_3 - \cos\theta_1)\right]}{(x_3 - x_1)\sin\left[\dfrac{\beta d}{2}(\cos\theta_2 - \cos\theta_1)\right]} \end{aligned}\right\} \tag{41}$$

$$a = \frac{x_2 - x_1}{\cos(\beta d\cos\theta_2 + \alpha) - \cos(\beta d\cos\theta_1 + \alpha)} \tag{42}$$

$$b = x_1 - a\cos(\beta d\cos\theta_1 + \alpha) \tag{43}$$

A general approach not only for Chebyshev but also for Legendre arrays can be found in Ref. 24. Table 4 and Figs. 8–11 present four common endfire cases expressed in Eqs. (39)–(43).

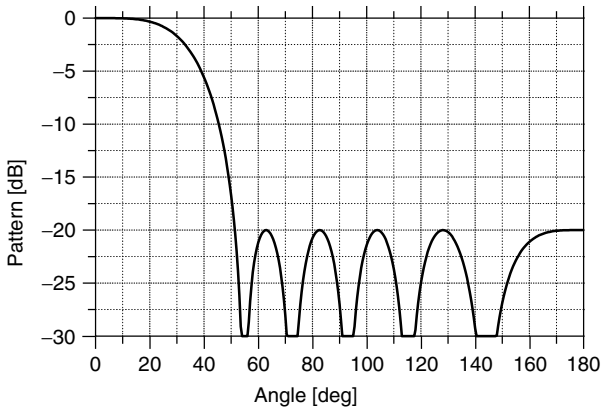


Figure 8. Pattern of the endfire case 1 array for $N = 11$, $d/\lambda = 0.25$, SLL $= -20$ dB.

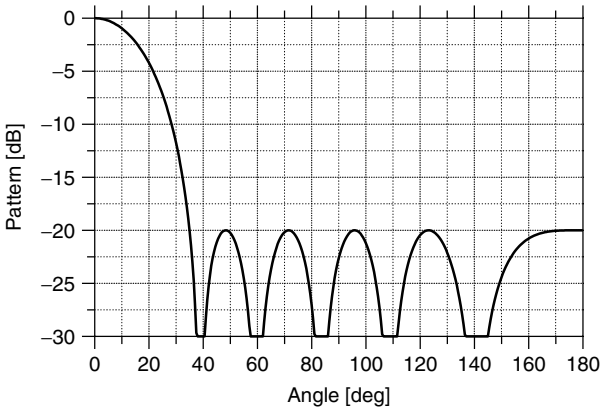


Figure 9. Pattern of the endfire case 2 array for $N = 11$, $d/\lambda = 0.2$, SLL $= -20$ dB.

6. PLANAR ARRAYS

Individual radiators positioned on a plane constitute a planar array. The usual planar array is rectangular. The elements are positioned along a rectangular grid. The grid can be constructed from the combination of two perpendicular linear arrays. Because of the increase in variables, the pattern can be controlled and scanned at any point in space. The more symmetrical patterns and the lower

Table 4. Coefficients of Chebyshev Endfire Arrays

Endfire Case 1	Endfire Case 2	Endfire Case 3 (Optimum)	Endfire Case 4
$x_1 \to \theta_1 = 0$	$x_1 \to \theta_1 = 0$	$x_1 \to \theta_1 = 0$	$x_1 \to \theta_1 = 0$
$x_2 \to \theta_2 = \pi$	$x_2 \to \theta_2 = \pi$	$x_2 \to \theta_2 = \pi$	$-x_2 \to \theta_2 = \pi$
x_3 #	x_3 #	$x_3 = -(a + b)$	$x_3 \to \theta_3 = \theta_{HP}$
$\mathrm{T}_m(x_1) = R$	$\mathrm{T}_m(x_1) = R$	$\mathrm{T}_m(x_1) = R$	$\mathrm{T}_m(x_1) = R, \mathrm{T}_m(x_3) = R\sqrt{2}/2$
$\alpha = -\beta d$	$\alpha = \beta d$	$\alpha = 2\tan^{-1}\left[\sin(\beta d)\dfrac{x_1 + x_2 + 2x_3 - 2\sqrt{(x_1 + x_3)(x_2 + x_3)}}{(x_1 - x_2)[1 + \cos(\beta d)]}\right]$	$\alpha = \cot^{-1}\left[\dfrac{(2\lambda + 1)\sin\beta d - \sin(\beta d\cos\theta_3)}{\cos\beta d - \cos(\beta d\cos\theta_3)}\right]$
			$\lambda = \dfrac{x_3 - x_1}{x_1 + x_2}$
$a = \dfrac{x_1 + x_2}{1 - \cos 2\beta d}$	$a = \dfrac{x_1 - x_2}{\cos 2\beta d - 1}$	$a = \dfrac{x_2 - x_1}{2\sin\beta d\cdot\sin\alpha}$	$a = \dfrac{x_1 + x_2}{2\sin\beta d\cdot\sin\alpha}$
$b = x_1 - a$	$b = x_2 - a$	$b = -x_3 - a$	$b = x_1 - \cos(\beta d + \alpha)$
$d_{\max} = \lambda/4$	$d_{\max} = \lambda/4$	$d_{\max} = \dfrac{\lambda}{2} - \dfrac{\lambda}{2\pi}\sin^{-1}\left(\sqrt{\dfrac{x_1 + x_3}{x_1 + 1}}\right)$	$d_{\max} = \dfrac{\lambda}{2\pi} \times \left[\alpha - \sin^{-1}\left(\sqrt{\dfrac{x_1 + x_2}{x_1 + 1}}\right)\right]$

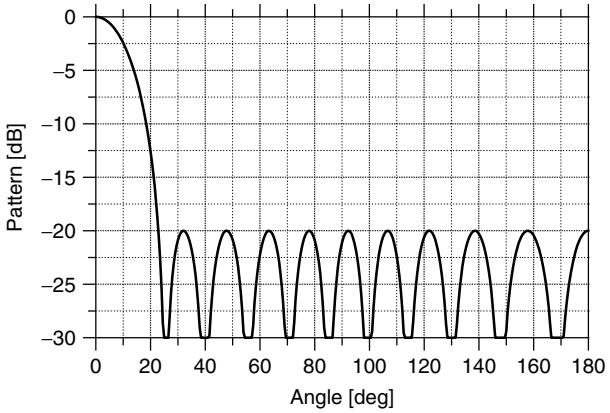


Figure 10. Pattern of the endfire case 3 array for $N = 11$, $d/\lambda = 0.35$, SLL $= -20$ dB.

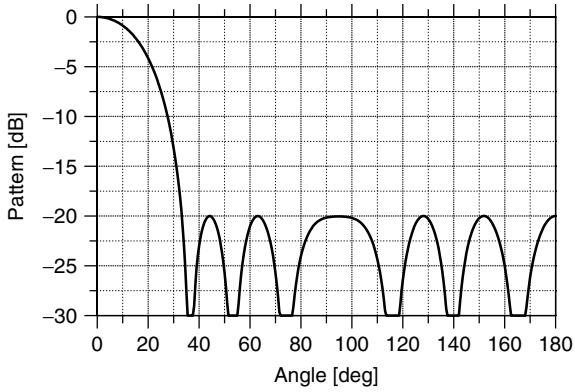


Figure 11. Pattern of the endfire case 4 array for $N = 11$, $d/\lambda = 0.4$, HPBW $= 35°$, SLL $= -20$ dB.

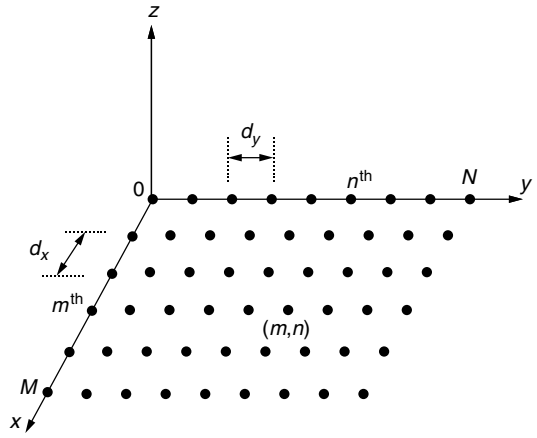


Figure 12. Planar array geometry.

sidelobes are the two main advantages of the planar over the linear arrays. Let us refer to Fig. 12.

The element corresponding to the mth row and the nth column is the mnth element with excitation I_{mn}. The array factor is

$$\mathrm{AF}(\theta, \varphi) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{mn} z_x^m z_y^n \tag{44}$$

where

$$z_x = e^{j\beta d_x \sin\theta \cos\varphi} \text{ and } z_y = e^{j\beta d_y \sin\theta \sin\varphi}$$

For uniform excitation and progressive phase α_x along x axis and α_y along y axis, the array factor becomes

$$\mathrm{AF}(\theta, \varphi) = I_0 \left(\sum_{m=1}^{M} e^{j(m-1)(\beta d_x \sin\theta \cos\varphi + \alpha_x)} \right) \times \left(\sum_{n=1}^{N} e^{j(n-1)(\beta d_y \sin\theta \sin\varphi + \alpha_y)} \right) \tag{45}$$

According to Eqs. (22) and (23), expression (45) gives

$$|\mathrm{AF}(\theta, \varphi)| = \left| \frac{\sin(M\Psi_x/2)}{M \sin(\Psi_x/2)} \right| \left| \frac{\sin(N\Psi_y/2)}{N \sin(\Psi_y/2)} \right| \tag{46}$$

where

$$\left.\begin{aligned} \Psi_x &= \beta d_x \sin\theta \cos\varphi + \alpha_x \\ \Psi_y &= \beta d_y \sin\theta \sin\varphi + \alpha_y \end{aligned}\right\} \tag{47}$$

For d_x and/or $d_y \geq \lambda$, the in-phase addition of the radiated field is made in more than one direction and grating lobes are produced. The grating lobes are located at

$$\left.\begin{aligned} \Psi_x &= \pm 2m\pi \\ \Psi_y &= \pm 2n\pi \quad m \text{ and } n = 1, 2, \ldots \end{aligned}\right\} \tag{48}$$

If the mainlobe direction is at (θ_0, φ_0), then

$$\left.\begin{aligned} \alpha_x &= -\beta d_x \sin\theta_0 \cos\varphi_0 \\ \alpha_y &= -\beta d_y \sin\theta_0 \sin\varphi_0 \end{aligned}\right\} \tag{49}$$

Solving (48) for the direction $(\theta_{mn}, \varphi_{mn})$ where grating lobes occur, we obtain

$$\varphi_{mn} = \tan^{-1}\left[\frac{\sin\theta_0 \sin\varphi_0 \pm n\lambda/d_y}{\sin\theta_0 \cos\varphi_0 \pm m\lambda/d_x} \right] \tag{50}$$

$$\theta_{mn} = \sin^{-1}\left[\frac{\sin\theta_0 \sin\varphi_0 \pm n\lambda/d_y}{\sin\varphi_{mn}} \right] = \sin^{-1}\left[\frac{\sin\theta_0 \cos\varphi_0 \pm m\lambda/d_x}{\cos\varphi_{mn}} \right] \tag{51}$$

A 6×6-element array with $\alpha_x = \alpha_y = 0$ will be given. Figures 13 and 14 show the corresponding patterns for $d_x = d_y = \lambda/2$ and $d_x = d_y = \lambda$. It is obvious that for the large spacing there are grating lobes at $\theta = \pi/2$ and $\varphi = 0$, $\pi/2$, π, $3\pi/2$.

Finally, the pattern of a 6×6 rectangular array that combines two different Dolph–Chebyshev arrays with SLL $= -20$ dB for $d_x = 0.5\lambda$ and $d_y = 0.82\lambda$ is shown in Fig. 15.

7. CIRCULAR ARRAYS

A *circular array* is a planar array with the elements positioned on a circular ring. A circular array with N isotropic elements (see Fig. 16) produces the array factor:

$$\mathrm{AF}(\theta, \varphi) = \sum_{n=1}^{N} I_n e^{j[\beta R \sin\theta \cos(\varphi - \varphi_n) + \alpha_n]} \tag{52}$$

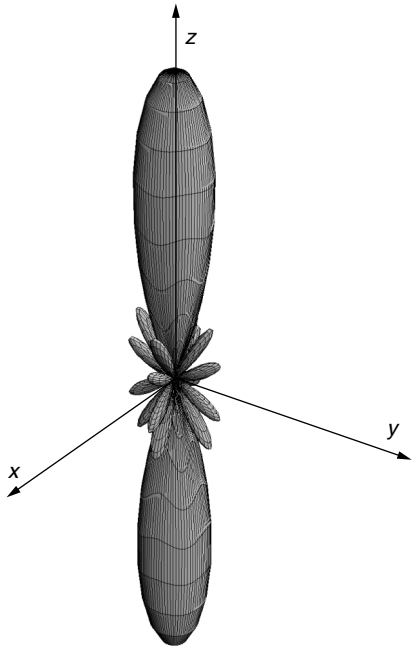


Figure 13. Three-dimensional pattern of a 6×6 uniform planar array with $a_x = a_y = 0$ and $d_x = d_y = 0.5\lambda$.

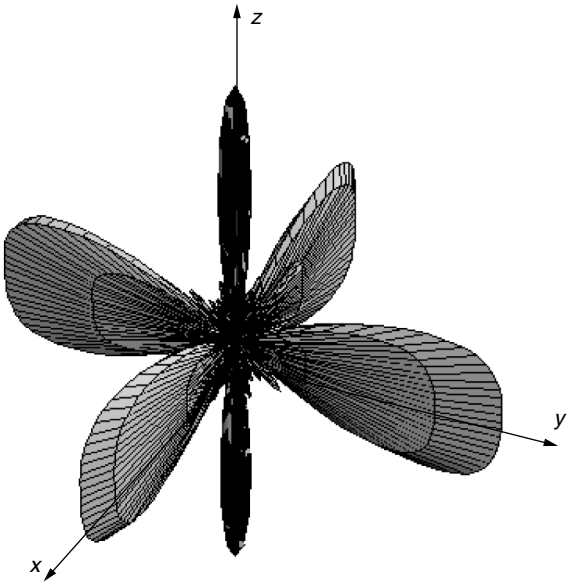


Figure 14. Three-dimensional pattern of a 6×6 uniform planar array with $a_x = a_y = 0$ and $d_x = d_y = 1\lambda$.

where I_n is the amplitude of the excitation of the nth element and α_n is the corresponding phase. To have a peak at (θ_0, φ_0), it must be

$$\alpha_n = -\beta R \sin\theta_0 \cos(\varphi_0 - \varphi_n) \tag{53}$$

and

$$\mathrm{AF}(\theta, \varphi) = \sum_{n=1}^{N} I_n e^{j\beta R[\sin\theta\cos(\varphi-\varphi_n) - \sin\theta_0\cos(\varphi_0-\varphi_n)]} \tag{54}$$

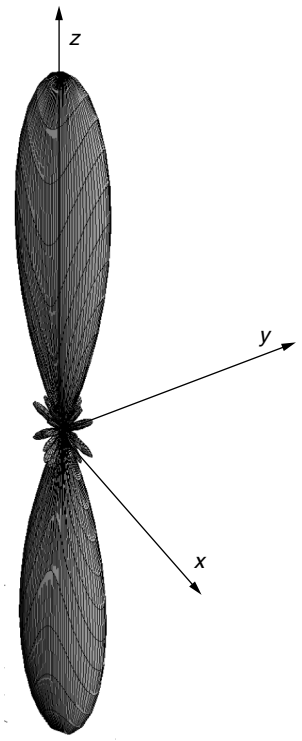


Figure 15. Three-dimensional pattern of a 6×6 Chebyshev planar array with SLL $= -20$ dB, $a_x = a_y = 0$, $d_x = 0.5\lambda$, and $d_y = 0.82\lambda$.

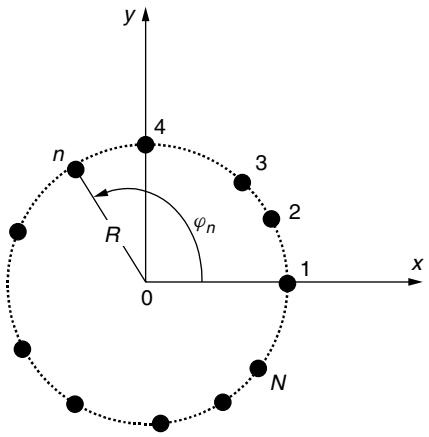


Figure 16. Geometry of a circular array with N elements.

The pattern of a 12-element uniform circular array with $\beta R = 10$, $\theta_0 = 0$, $\varphi_0 = 0$ is shown in Fig. 17.

An interesting array comes from a dipole positioned at the bisector of a corner reflector with angle $\omega_N = 2\pi/\mathrm{N}$. The reflector creates a circular array with the $N - 1$ images (see Fig. 18). Figure 19 presents the corresponding H pattern. It is noticed that the pattern outside the reflectors angle does not exist.

M circular arrays in concentric rings produce an array factor

$$\mathrm{AF}(\theta, \varphi) = \sum_{m=1}^{M} \sum_{n=1}^{N} I_{mn} e^{j[\beta R_m \sin\theta\cos(\varphi-\varphi_{mn}) + \alpha_{mn}]} \tag{55}$$

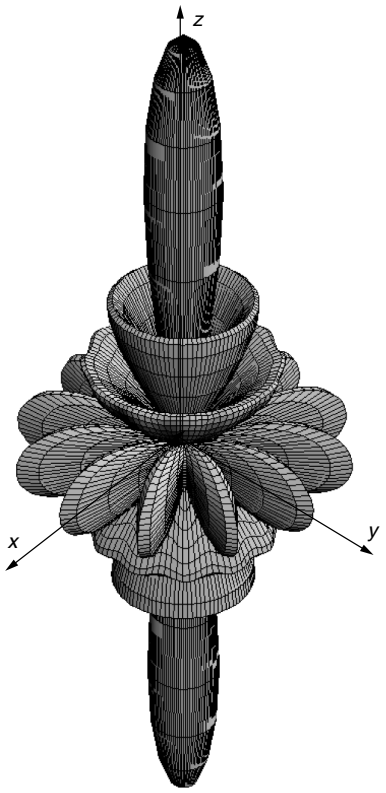


Figure 17. Three-dimensional pattern of a circular array with $N = 12$, $\beta R = 10$, $\theta_0 = 0^\circ$, and $\varphi_0 = 0^\circ$.

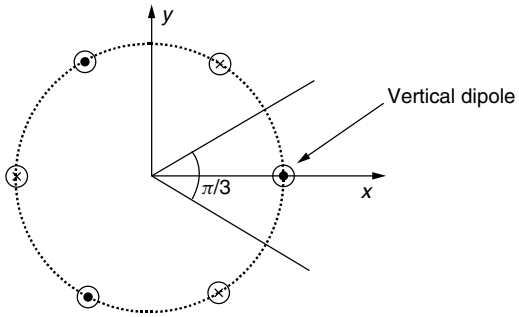


Figure 18. Vertical dipole in front of a corner reflector with its images.

where $I_{mn}e^{j\alpha_{mn}}$ is the excitation of the nth element of the mth ring.

A corner reflector with a linear array positioned in front of it (see Fig. 20) creates concentric rings. For a 2-dipole uniform linear array the pattern is as presented in Fig. 21.

8. 3D ARRAYS

N elements positioned in 3D space constitute a three-dimensional array. The array factor is

$$\mathrm{AF}(\theta,\varphi) = \sum_{n=1}^{N} I_n e^{j[\alpha_n + \beta r_n[\sin\theta\sin\theta_n\cos(\varphi-\varphi_n)+\cos\theta\cos\theta_n]]} \qquad (56)$$

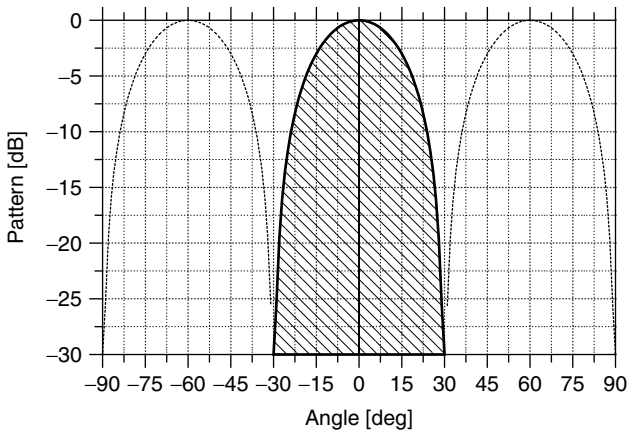


Figure 19. H-Pattern of a dipole in front of a $\pi/3$ corner reflector.

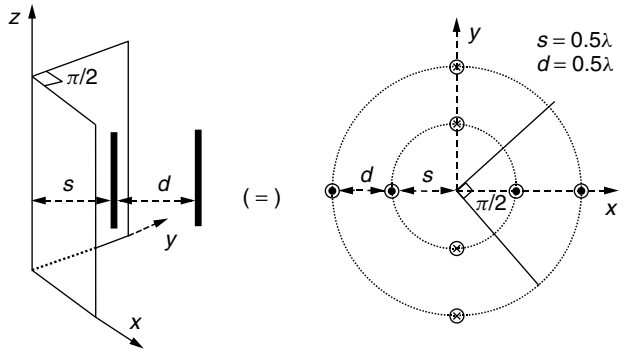


Figure 20. A parallel dipole linear array in front of a $\pi/2$ corner reflector.

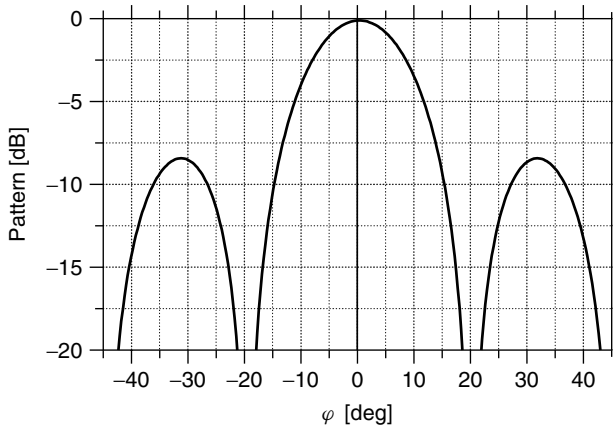


Figure 21. Pattern of a 2-dipole array (see Fig. 20) in front of a $\pi/2$ reflector.

where $(r_n, \theta_n, \varphi_n)$ are the spherical coordinates of the nth element and $I_n e^{j\alpha_n}$ is the corresponding excitation. Parallel circular arrays with their centers on the same axis constitute a cylindrical array. The array factor is simplified to

$$\mathrm{AF}(\theta,\varphi) = \sum_{m=1}^{M}\sum_{I=1}^{L} I_{mI} e^{j[\alpha_{mI} + \beta R_0 \sin\theta\cos(\varphi-\varphi_{mI})+\beta z_m\cos\theta]} \qquad (57)$$

where $I_{mI}e^{j\alpha_{mI}}$ is the excitation of the Ith element of the mth circular array, R_0 is the radius of the circular arrays, and z_m is the position of the mth array on the z axis. A cylindrical array can be made from an array of collinear dipoles in front of a corner reflector (Fig. 22).

To have a maximum at $\theta = \pi/2$, $\varphi = \pi/4$, the array can be uniform. If there are additional constraints on the SLL, then the excitations must be nonuniform. Figure 23 shows the E pattern of a Chebyshev array with $N = 9$ collinear dipoles in front of a $\pi/2$ corner reflector. $d = 0.7\lambda$, SLL ≤ -20 dB and maximum occurs at $\theta = \pi/2$, $\varphi = \pi/4$.

9. CONFORMAL ARRAYS

Arrays with requirements in conformality to a shaped surface are known as *conformal*. Conformal arrays are used in mobile platforms for aerodynamic reasons. Also for specified angles of coverage, arrays can be conformal to stationary shaped surfaces.

Analysis of conformal arrays differs from that of planar ones. The pattern of the array can not be given by multiplying the element pattern and the array factor. These are not separable, and, of course, the pattern is not a simple polynomial. Also, the illumination around the radiating surface as well as the polarization and the pattern of each radiating element must be taken into account separately.

Practical communication and surveillance systems with scanning requirements use conformal arrays of cylindrical shape. In this case one part of the array is illuminated by means of a switching network. For the commutation of a given illumination region around the cylinder, one can use either mechanical rotation, or switch networks, or lens scanning, or matrix beam formers, or digital beam formers [25]. If the array compared to the radius R takes up a small sector and the radius is large ($R \gg \lambda$), then the element pattern is approximated by that of a planar array. If the sector is large compared to the radius R or if the element is in an illuminated region of an array fully wrapped around the cylinder, then the pattern must be carefully calculated and is much different than that of a planar array.

An example of a cylindrical array of 16 microstrip patches is shown in Fig. 24. For a uniform excitation and element phase given by Eq. (53) for a maximum at ($\theta_0 = \pi/2$, $\varphi_0 = 0$) and ($\theta_0 = \pi/2$, $\varphi_0 = \pi/3$), the patterns of the array are presented in Fig. 25.

Conical arrays are used mainly for missiles and aircrafts. The design follows the corresponding one of the cylindrical arrays. Spherical, hemispherical, and conical

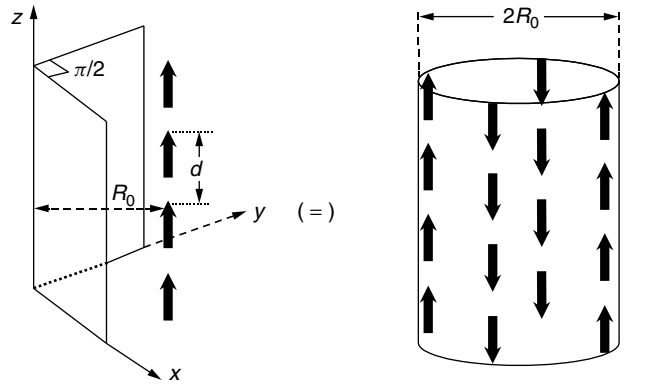


Figure 22. Array of collinear dipoles in front of a $\pi/2$ reflector.

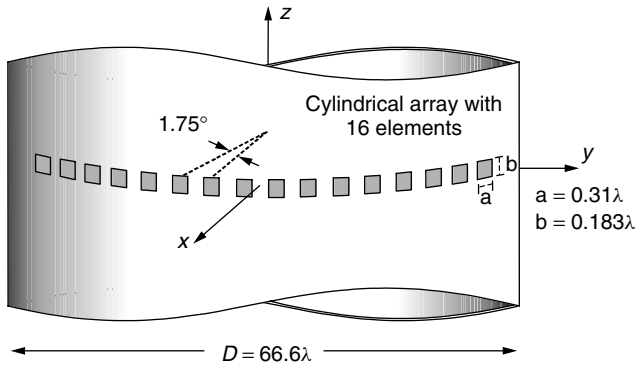


Figure 24. A cylindrical array of 16 rectangular microstrip patches.

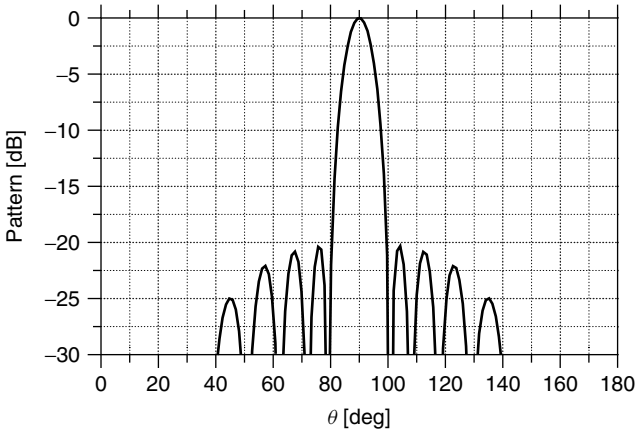


Figure 23. E pattern of a Chebyshev linear array of $N = 9$ collinear dipoles with $d = 0.7\lambda$ and SLL ≤ -20 dB.

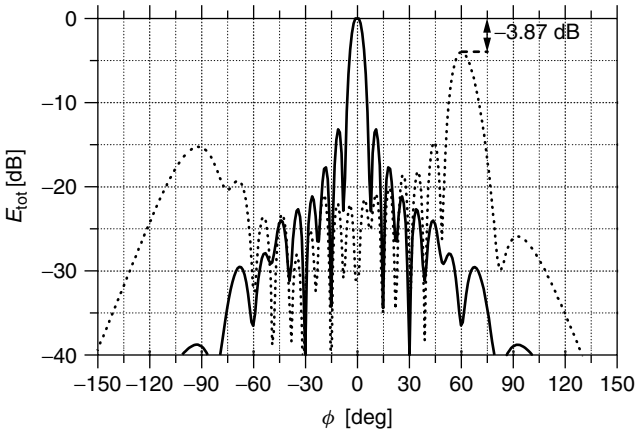


Figure 25. Radiation patterns for a uniform array of 16 axially polarized rectangular patches given in Fig. 24.

are conformal arrays, which are excited in groups of subarrays. Usually they are fed by switch matrices in the same way as the cylindrical arrays.

10. PATTERN SYNTHESIS FOR ARRAYS

The main advantage of arrays is that they can produce accurate approximations of desired radiation patterns. Several techniques have been given in the past for synthesizing array factors. Most of them relate to the synthesis of narrowbeam and low-sidelobe patterns.

Synthesis is based primarily on the antenna engineer experience. The meaningful method, which will result in a realizable solution, must approach the desired property and not the exact requirement. Several synthesis procedures will be described in the following paragraphs.

10.1. Uniform Linear Array Synthesis

Uniform arrays can be used for SLL ≥ -13.3 dB. A linear scanning array with maximum at $\theta = \theta_0$ and half power beamwidth θ_H must have (see Table 1)

$$\theta_0 > \cos^{-1}\left(\cos^2\frac{\theta_H}{2}\right) \tag{58}$$

and

$$N\frac{d}{\lambda} = 0.4428\left[\frac{2(1+\cos\theta_H)}{\sin^4\theta_0 - (\cos\theta_H - \cos^2\theta_0)^2}\right]^{1/2} \tag{59}$$

For example

$$\left.\begin{aligned} &\text{For } \theta_H = 10^\circ \text{ and } \theta_0 = \frac{\pi}{2} \Rightarrow N\frac{d}{\lambda} \\ &\quad = 5.08 \text{ (broadside array)} \\ &\text{For } \theta_H = 10^\circ \text{ and } \theta_0 = \frac{\pi}{6} \Rightarrow N\frac{d}{\lambda} \\ &\quad = 10.28 \text{ (intermediate array)} \end{aligned}\right\} \tag{60}$$

An endfire array where $\theta_0 = 0^\circ$ (Table 1) needs

$$N\frac{d}{\lambda} = \frac{0.4428}{1-\cos\dfrac{\theta_H}{2}} \text{ (ordinary)} \tag{61}$$

$$N\frac{d}{\lambda} = \frac{0.1398}{1-\cos\dfrac{\theta_H}{2}} \text{ (Hansen–Woodyard)} \tag{62}$$

Uniform arrays are useful in practice. For example, an array for mobile communications or with $\theta_0 = 96^\circ$ and $\theta_H = 8^\circ$ must have

$$N\frac{d}{\lambda} = 6.383 \tag{63}$$

Eight $\lambda/2$ collinear dipoles with phase difference of 30° in $d/\lambda = 0.798$ can be used. The antenna E pattern (see Fig. 26) is found by multiplying the array factor by the element pattern.

10.2. Chebyshev Arrays Synthesis

Arrays with constraints on the SLL are useful in communications and radar systems. The Dolph method can be used for $d/\lambda \geq 0.5$. To avoid grating lobes we must have $d \leq d_{\max}$, where

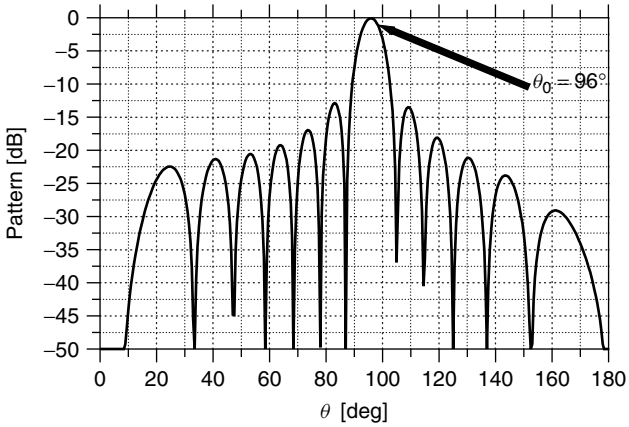


Figure 26. E pattern of a uniform array of eight collinear dipoles in $d/\lambda = 0.798$, $\theta_0 = 96^\circ$, and HPBW $= 8^\circ$.

$$\frac{d_{\max}}{\lambda} = 1 - \frac{\cos^{-1}(1/x_0)}{\pi} \tag{64}$$

x_0 is the distance where the maximum of the array occurs.

For $d/\lambda < 0.5$ the Riblet method must be used. For the Dolph array the HPBW has been related with that of the uniform one (HPBW$_u$) of the same length. The so-called broadening factor f was found to be [7]

$$f = \frac{(\text{HPBW})}{(\text{HPBW}_u)} = 1 + 0.632\left[\frac{2}{R}\cosh\sqrt{(\cosh^{-1}R)^2 - \pi^2}\right]^2 \tag{65}$$

f is valid in the range of -60 dB $\leq$ SLL ≤ -20 dB and for scanning near broadside.

An estimate to the directivity D with the help of f is possible:

$$D = \frac{2R^2}{1 + (R^2 - 1)f\dfrac{\lambda}{Nd}} \tag{66}$$

For the Riblet case the HPBW is given approximately by

$$\text{HPBW} \cong 10.3^\circ\frac{\lambda}{Nd}\sqrt{s + 4.52}\csc\theta_0 \tag{67}$$

where s is the SLL in dB and θ_0 is the scan angle. Also the directivity D [7] is

$$D = \frac{2R^2}{1 + R^2\dfrac{\lambda}{Nd}\sqrt{\dfrac{\ln(2R)}{\pi}}\cdot\sin\left(\beta\dfrac{d}{2}\right)} \tag{68}$$

Array factors of the form [24]

$$\text{AF}(\theta) = T_m(x)T_1^n(x) \tag{69}$$

are able to give either equal or unequal sidelobes. The number of nulls depends on m and n. If we compare (69) with an array factor

$$\text{AF}_1(\theta) = T_{m+n}(x) \tag{70}$$

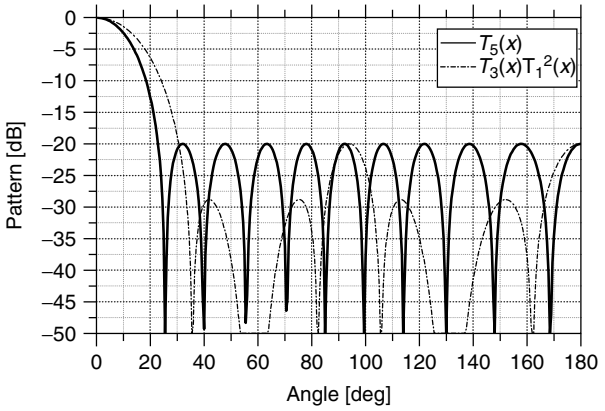


Figure 27. Patterns of case 3 endfire Chebyshev array with array factor $T_5(x)$ and $T_3(x)T_1^2(x)$ for SLL $= -20$ dB.

we see that $\mathrm{AF}_1(\theta)$ has $(m+n)$ roots while $\mathrm{AF}(\theta)$ has either $(m+1)$ roots for $m = 2k$ or m roots for $m = 2k+1$.

Figure 27 presents for comparison the factors $T_5(x)$ and $T_3(x)T_1^2(x)$ of the case 3 endfire array for $N = 11$, $d/\lambda = 0.35$, and SLL $= -20$ dB.

10.3. Synthesis by Sampling or by Root Matching

Continuous distributions create excellent patterns with low sidelobes. Discrete arrays coming from sampling them can give similar patterns. For large-element spacing, the patterns of the array and the line source do not match well. To avoid this problem the method of root matching is used. In other words, the nulls of the continuous distribution pattern appear in the pattern of the discrete array. If the pattern does not yield the desired accuracy, a perturbation technique [26] can be applied. In this case the distribution of the discrete-element array varies to improve the accuracy.

10.3.1. Simple Distributions. A discrete-element array of a fixed length is transformed to a continuous distribution as the number of elements approaches to infinity (Fig. 28).

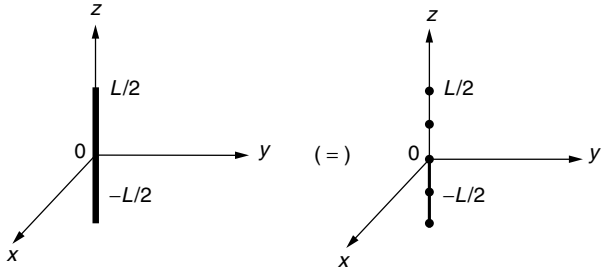


Figure 28. Line source and its equivalent discrete-element array.

The array factor reduces to an integral and is called the space factor (SF):

$$\mathrm{SF}(\theta) = \int_{-L/2}^{L/2} I(z')e^{j(\beta\cos\theta-\alpha)z'}dz' = \int_{-L/2}^{L/2} I(z')e^{j\xi z'}dz' \qquad (71)$$

where $I(z')$ and α are the amplitude distribution and phase progress along the source. This equation is the finite one-dimensional Fourier transform relating the far field to the excitation.

Changing the bounds of integration, Eq. (71) becomes

$$\mathrm{SF}(\theta) = \int_{-\infty}^{\infty} I(z')e^{j\xi z'}dz' \qquad (72)$$

$I(z')$ is zero outside of $-L/2 \le z' \le L/2$. Using the Fourier transform we have

$$I(z') = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{SF}(\theta)e^{-jz'\xi}d\xi \qquad (73)$$

If L is large enough, Eq. (71) gives the desired pattern within a certain error. In the discrete-element array $I(z')$ is sampled in appropriate intervals. It is observed that a small difference between the two patterns appears. Useful distributions with their characteristics are presented in Table 5.

10.3.2. Taylor Distribution (Chebyshev Error). Chebyshev arrays provide an optimum relation between the SLL

Table 5. Radiation Characteristics for Line Sources and Linear Arrays with Uniform, Triangular, Cosine, and Cosine-Squared Distributions

Distribution	Uniform	Triangular	Cosine	Cosine-Squared
Distribution I_n	I_0	$I_1\left(1-\frac{2}{L}\mid z'\mid\right)$	$I_2\cos\left(\frac{\pi}{L}z'\right)$	$I_3\cos^2\left(\frac{\pi}{L}z'\right)$
Space factor (SF) $u=\left(\frac{\pi L}{\lambda}\right)\sin\theta$	$I_0L\frac{\sin u}{u}$	$I_1\frac{L}{2}\left[\frac{\sin\left(\frac{u}{2}\right)}{\frac{u}{2}}\right]^2$	$I_2L\frac{\pi}{2}\frac{\cos(u)}{(\pi/2)^2-u^2}$	$I_3\frac{L}{2}\frac{\sin(u)}{u}\left[\frac{\pi^2}{\pi^2-u^2}\right]$
Half-power beamwidth (degrees) $L\lambda$	$\frac{50.6}{(L\lambda)}$	$\frac{73.4}{(L\lambda)}$	$\frac{68.8}{(L\lambda)}$	$\frac{83.2}{(L\lambda)}$
First null beamwidth (degrees) $L\lambda$	$\frac{114.6}{(L\lambda)}$	$\frac{229.2}{(L\lambda)}$	$\frac{171.9}{(L\lambda)}$	$\frac{229.2}{(L\lambda)}$
First side lobe max. (to main max.) (dB)	-13.2	-26.4	-23.2	-31.5
Directivity factor (Llarge)	$2\left(\frac{L}{\lambda}\right)$	$0.75\left[2\left(\frac{L}{\lambda}\right)\right]$	$0.810\left[2\left(\frac{L}{\lambda}\right)\right]$	$0.667\left[2\left(\frac{L}{\lambda}\right)\right]$

and the HPBW. Another distribution characterized by low SLL of the first N sidelobes next to the main beam is the Taylor distribution. The other sidelobes gradually fall off in value. The space factor of the Taylor distribution comes from Dolph–Chebyshev if the elements of the array become infinite [27]:

$$\mathrm{SF}(\theta) = \frac{\cosh\left[\sqrt{(\pi A)^2 - u^2}\right]}{\cosh(\pi A)} \tag{74}$$

where

$$\left.\begin{aligned} u &= \pi \frac{L}{\lambda}(\cos\theta - \cos\theta_0) \\ \cosh(\pi A) &= R \end{aligned}\right\} \tag{75}$$

Since Eq. (74) cannot be realized physically, Taylor [27] presented a space factor whose roots are the zeros of SF (θ). Because the factor is the approximation of the ideal Chebyshev, it is also known as the *Chebyshev error*. The space factor is

$$\mathrm{SF}(u, A, \overline{n}) = \frac{\sin u}{u} \frac{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1 - \left(\frac{u}{u_n}\right)^2\right]}{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1 - \left(\frac{u}{n}\right)^2\right]} \tag{76}$$

where $(\overline{n} - 1)$ is a parameter that defines the number of pairs of inner nulls.

$$u_n = \overline{n} \frac{\sqrt{A^2 + \left(n - \frac{1}{2}\right)^2}}{\sqrt{A^2 + \left(\overline{n} - \frac{1}{2}\right)^2}} \quad n = 1, 2, \ldots, \overline{n} - 1 \tag{77}$$

The Taylor distribution can be found by the Fourier transform:

$$I(z') = \mathrm{SF}(0, A, \overline{n}) + 2\sum_{m=1}^{\overline{n}-1} \mathrm{SF}(m, A, \overline{n}) \cos\left(m\pi \frac{z'}{L}\right) \tag{78}$$

Sampling $I(z')$, we create the discrete array. Problems that arise and cause inaccuracies, even for large arrays, were addressed earlier [28].

Figure 29 presents the Taylor pattern for SLL $= -25$ dB, $L = 7\lambda$, and $\overline{n} = 5$. The amplitudes of the elements at $d/\lambda = 0.5$ are given in Fig. 30.

10.3.3. Taylor One-Parameter Distribution. In low-noise systems it is desirable to have the first sidelobes at a certain level while the others decay as the angle increases. Taylor [6] developed a procedure for synthesizing such a pattern. The distribution is referred as the *Taylor one-parameter* and is of the following form:

$$I_n(z') = I_0\left[\pi B \sqrt{1 - \left(\frac{2z'}{L}\right)^2}\right] \tag{79}$$

where z' = distance from the center of the line source
L = length of the line source

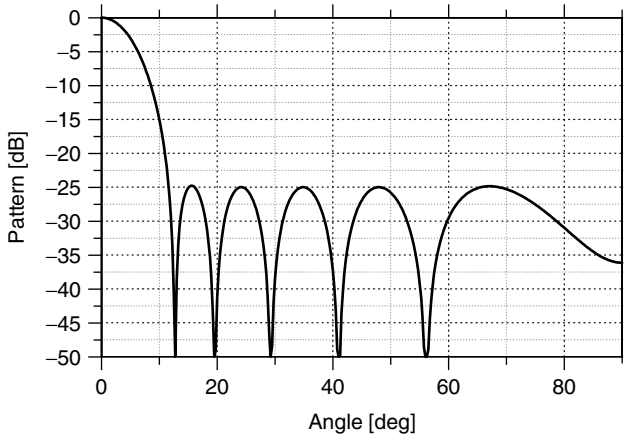


Figure 29. Pattern of a Taylor distribution with SLL $= -25$ dB, $\overline{n} = 5$, $d/\lambda = 0.5$, and $N = 14$.

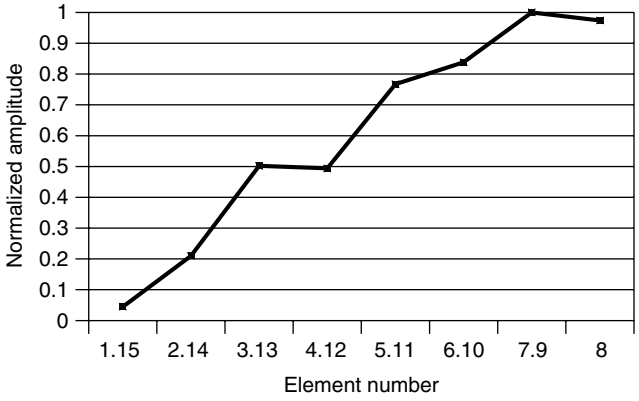


Figure 30. Amplitudes of the elements at $d/\lambda = 0.5$ of the Taylor distribution with the pattern given in Fig. 29.

B = parameter that determines the sidelobe
I_0 = modified Bessel function of the first kind and zero order.

The space factor associated with (79) can be obtained by the Fourier transform:

$$\mathrm{SF}(\theta) = \begin{cases} \dfrac{\sin[\sqrt{(\pi B)^2 - u^2}]}{\sqrt{(\pi B)^2 - u^2}}, & u^2 < (\pi B)^2 \\ \dfrac{\sinh[\sqrt{u^2 - (\pi B)^2}]}{\sqrt{u^2 - (\pi B)^2}}, & u^2 > (\pi B)^2 \end{cases} \tag{80}$$

where

$$u = \beta \frac{L}{2}(\cos\theta - \cos\theta_0) \tag{81}$$

Parameter B is found from [29]

$$R = 4.60333 \frac{\sinh \pi B}{\pi B} \tag{82}$$

10.3.4. Bayliss Line Source. A pattern null on boresight with the appropriate sidelobe level was developed by Bayliss [30]. Monopulse tracking systems use an auxiliary pattern of the form of Bayliss in coincidence with a beam

peak of the main pattern. The Bayliss pattern is described in terms of two parameters, A and $\overline{n}$. The pattern is

$$\mathrm{SF}(\theta) = u\cos(\pi u)\frac{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{u}{u_n}\right)^2\right]}{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{u}{n+\frac{1}{2}}\right)^2\right]} \tag{83}$$

$(\overline{n}-1)$ is the parameter that defines the number of inner nulls:

$$u_n = \begin{cases} \left(\overline{n}+\frac{1}{2}\right)\left(\frac{\xi_n^2}{A^2+\overline{n}^2}\right)^{1/2} & n = 1,2,3,4 \\ \left(\overline{n}+\frac{1}{2}\right)\left(\frac{A^2+n^2}{A^2+\overline{n}^2}\right)^{1/2} & n = 5,6,\ldots,\overline{n}-1 \end{cases} \tag{84}$$

A, ξ_n, and the location $u_{\max}$ where SF is maximized are found as a function of $S = |\text{sidelobe level (dB)}|$ (see Table 6):

$$x = \alpha_1 + S\{\alpha_2 + S[\alpha_3 + S(\alpha_4 + S\cdot\alpha_5)]\} \tag{85}$$

The aperture distribution by the sine Fourier series with $\overline{n}$ terms is

$$I(z') = \sum_{m=0}^{\overline{n}-1} B_m \sin\left[(2m+1)\frac{\pi z'}{L}\right] \tag{86}$$

where

$$B_m = \frac{(-1)^m\left(m+\frac{1}{2}\right)^2}{2j}\frac{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{m+\frac{1}{2}}{u_n}\right)^2\right]}{\displaystyle\prod_{n=1}^{\overline{n}-1}\left[1-\left(\frac{m+\frac{1}{2}}{n+\frac{1}{2}}\right)^2\right]} \tag{87}$$

A Bayliss and a Taylor pattern ($\overline{n}=5$) for SLL $= -30$ dB, $N = 14$ and $d/\lambda = 0.5$ are shown in Fig. 31.

10.3.5. Modified Patterns by Iteration. Taylor and Bayliss patterns with individual different sidelobes can be made by using a perturbation procedure. According to Elliot [5], we express the patterns in more general forms:

$$\text{Taylor:}\quad \mathrm{SF}(u) = C_0\frac{\sin\pi u}{u}\frac{\displaystyle\prod_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1}\left(1-\frac{u}{u_n}\right)}{\displaystyle\prod_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1}\left(1-\frac{u}{n}\right)} \tag{88}$$

Table 6. Coefficients of the Parameters A, ξ_n, and $u_{\max}$

x	α_1	α_2	α_3	$\alpha_4\cdot 10^5$	$\alpha_5\cdot 10^7$
A	0.3038753	0.05042922	−0.00027989	0.343	−0.2
ξ_1	0.9858302	0.0333885	0.00014064	−0.19	0.1
ξ_2	2.00337487	0.01141548	0.0004159	−0.373	0.1
ξ_3	3.00636321	0.00683394	0.00029281	−0.161	0
ξ_4	4.00518423	0.00501795	0.00021735	−0.088	0
$u_{\max}$	0.4797212	0.01456692	−0.00018739	0.218	−0.1

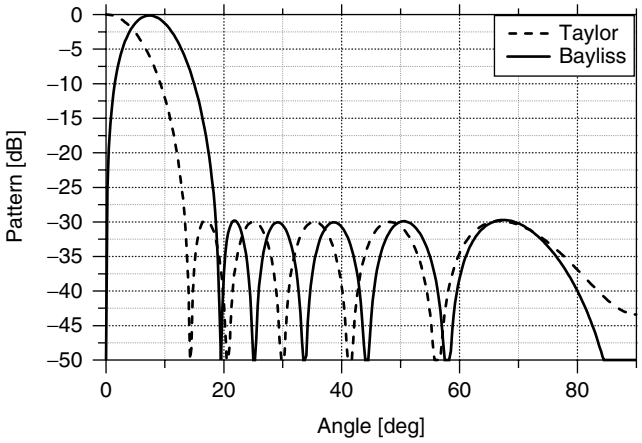


Figure 31. Pattern of a Taylor and a Bayliss array to give SLL $= -30$ dB ($\overline{n}=5$) for $N = 14$ and $d/\lambda = 0.5$.

where

$$\left.\begin{aligned} u_n &= \overline{n}_R\frac{\sqrt{A_R^2+\left(n-\frac{1}{2}\right)^2}}{\sqrt{A_R^2+\left(\overline{n}_R-\frac{1}{2}\right)^2}} \\ u_n &= -\overline{n}_L\frac{\sqrt{A_L^2+\left(n+\frac{1}{2}\right)^2}}{\sqrt{A_L^2+\left(\overline{n}_L-\frac{1}{2}\right)^2}} \end{aligned}\right\} \tag{89}$$

where R and L identify the right and left side of the pattern. $\overline{n}_R$ and $\overline{n}_L$ denote the transition roots of the two sides, and A_R and A_L are the corresponding SLL parameters.

$$\text{Bayliss:}\quad \mathrm{SF}(u) = C_0 u\cos\pi u\frac{\displaystyle\prod_{n=-(\overline{n}_{L-1})}^{\overline{n}_R-1}\left(1-\frac{u}{u_n}\right)}{\displaystyle\prod_{n=-\overline{n}_L}^{\overline{n}_R-1}\left(1-\frac{u}{n+\frac{1}{2}}\right)} \tag{90}$$

We start from a pattern $\mathrm{SF}_0(u)$ with the SLL on both sides to be the average of the desired ones. All the roots of Eqs. (88) and (90) u_n^0 are known.

We assume that the roots of the desired pattern are

$$u_n = u_n^0 + \delta u_n \tag{91}$$

with a small perturbation δu_n. Then if

$$C = C_0 + \delta\, C$$

$\mathrm{SF}(u)$ becomes

$$\frac{\mathrm{SF}(u)}{\mathrm{SF}_0(u)} - 1 = \frac{\delta C}{C_0} + \sum_{n=-(\overline{n}_{L-1})}^{\overline{n}_{R-1}}\frac{\dfrac{u}{(u_n^0)^2}}{1-\dfrac{u}{u_n^0}}\delta u_n \tag{92}$$

The peak positions u_m^p give

$$\frac{\mathrm{SF}(u_m^p)}{\mathrm{SF}_0(u_m^p)} - 1 = \frac{\delta C}{C_0} + \sum_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1} \frac{\dfrac{u_m^p}{(u_n^0)^2}}{1 - \dfrac{u_m^p}{u_n^0}} \delta u_n \tag{93}$$

For the $\overline{n}_R + \overline{n}_L - 1$ lobes we have an equal number of linear equations of the form (93). The system is solved for $\delta C/C_0$ and the $\overline{n}_R + \overline{n}_L - 2$ values δu_n since $\delta u_0 = 0$. The new values of u_n are substituted in (88) and the new pattern is checked. The process is repeated until the new pattern differs from the desired by a minimum predefined amount.

The same procedure is applied for the Bayliss distribution. Equation (92) is modified to

$$\frac{\mathrm{SF}(u_m^p)}{\mathrm{SF}_0(u)} - 1 = \frac{\delta C}{C_0} - \frac{\delta u_0}{u_m^p} + \sum_{n=-(\overline{n}_L-1)}^{\overline{n}_R-1} \frac{\dfrac{u_m^p}{(u_n^0)^2}}{1 - \dfrac{u_m^p}{u_n^0}} \delta u_n \tag{94}$$

which gives a system of $\overline{n}_R + \overline{n}_L$ unknowns, which is solved. The perturbation process is repeated in the same way as before.

Figures 32 and 33 show the pattern of two modified distributions for $\overline{n} = 6$, SLL $= -20$ dB and three intermost pairs of lobes -30 dB.

In the preceding cases, an iterative procedure can be applied for power pattern synthesis where we have additional degrees of freedom. Orchard et al. [13] proposed a technique by dividing the pattern in the shaped beam region and the sidelobe region.

The array factor in general is

$$\mathrm{AF}(\theta) = \prod_{n=1}^{N} (z - z_n) = \sum_{n=0}^{N} I_n z^n \tag{95}$$

It is assumed that the zero locations are complex of the form

$$z_n = \exp(a_n + jb_n) \tag{96}$$

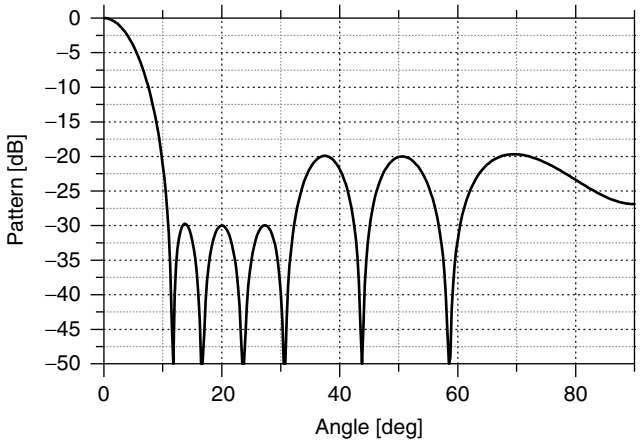


Figure 32. Modified Taylor pattern for $\overline{n} = 6$, three intermost pairs of lobes with -30 dB level and the other lobes with -20 dB level ($N = 14$, $d/\lambda = 0.5$).

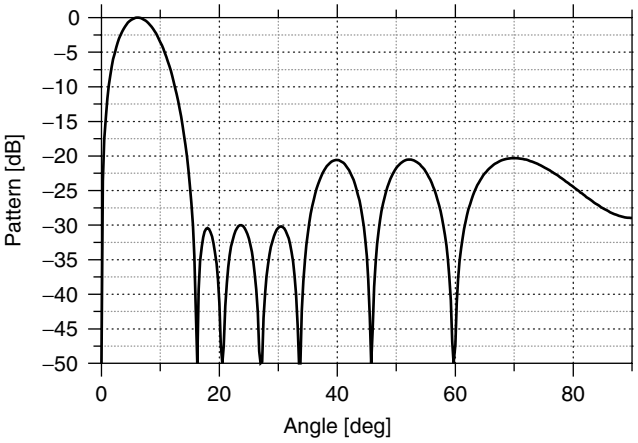


Figure 33. Modified Bayliss pattern for $\overline{n} = 6$, three intermost pairs of lobes with -30 dB level and the other lobes with -20 dB level ($N = 14$, $d/\lambda = 0.5$).

and z is written as

$$z = \exp(j\varphi) \tag{97}$$

Orchard sets the Nth root $z_N = 1$ and expresses the power pattern in decibels:

$$G = \sum_{n=1}^{N-1} 10 \log[1 - 2e^{a_n} \cos(\varphi - b_n) + e^{2a_n}] + 10 \log[2(1 + \cos\varphi)] + C_1 \tag{98}$$

C_1 is a constant that allows G to have at the main beam a given value.

The unknown coefficients a_n, b_n, and φ are found by using an iterative scheme. This scheme uses the derivatives of G and the difference between the existing and the desired power pattern. The procedure does not produce an optimum result. However it offers flexibility and control to the ripple and the sidelobe level as well as to the entire radiation pattern.

11. FOURIER TRANSFORM AND THE ORTHOGONAL METHOD

A linear uniformly spaced array with nonuniform excitation has an array factor of the form

$$\mathrm{AF}(\psi) = \sum_{n=1}^{N} I_n e^{jn\psi} \tag{99}$$

We expand a desired $\mathrm{AF}_d(\psi)$ in a Fourier series with infinite terms of the form (99). The first N coefficients of the two series are equated to approximate the desired pattern. The coefficients are found by using the orthogonality of the expansion functions:

$$I_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{AF}_d(\psi) e^{-jn\psi} d\psi \tag{100}$$

The Fourier method is adequate for spacing $d = 0.5\lambda$. For $d > 0.5\lambda$ it fails, while for $d < 0.5\lambda$ and a sufficient

number of elements the pattern more closely matches the desired one.

The general expression of (99) comes from nonuniformly spaced arrays, [33–37]:

$$\mathrm{AF}(u) = \sum_{n=1}^{N} I_n e^{jx_n u} \tag{101}$$

where

$$\left.\begin{aligned} u &= \pi \cos\theta \\ x_n &= \frac{d_n}{\lambda/2} \end{aligned}\right\} \tag{102}$$

The basis functions of (101) are not orthogonal. Their inner product is

$$k_{in} = \int_{-\pi}^{\pi} e^{j(x_i - x_n)u} du = \frac{\sin(x_i - x_n)\pi}{(x_i - x_n)\pi} \tag{103}$$

$\mathrm{AF}(u)$ can be expressed by the Gram–Schmidt procedure [33] in an orthogonal basis $\{\Psi_n(u)\}$:

$$\Psi_n(u) = \sum_{i=1}^{n} c_i^{(n)} e^{jx_i u} \tag{104}$$

and

$$\mathrm{AF}(u) = \sum_{n=1}^{N} B_n \Psi_n(u) \tag{105}$$

With the aid of the orthogonality, we have

$$B_n = \int_{-\pi}^{\pi} \mathrm{AF}_d(u) \cdot \Psi_n^*(u)\, du \tag{106}$$

$\Psi_n^*(u)$ is the conjugate of $\Psi_n(u)$.

Combining (101), (104), and (105), we have

$$I_n = \sum_{i=n}^{N} c_n^{(i)} B_i \tag{107}$$

For comparison purposes, an 7-element array with 0.85λ spacing and a constant beam between 85° and 95° is designed. Figure 34 presents the pattern obtained by the orthogonal method and the Fourier transform.

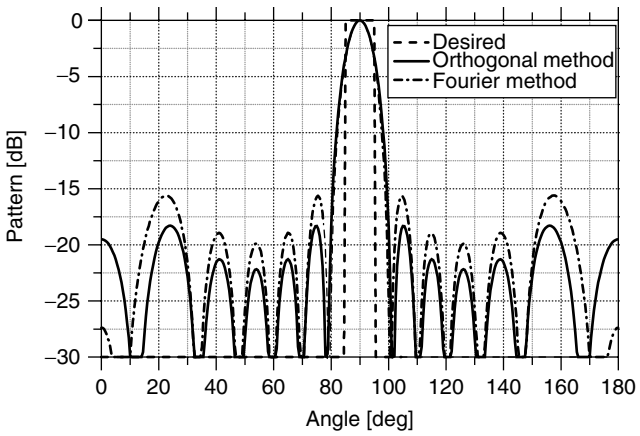


Figure 34. Pattern of an 7-element array with $d/\lambda = 0.85$ by the orthogonal and Fourier methods.

12. WOODWARD–LAWSON (WL) METHOD AND ORTHOSYNTHESIS

A uniform linear array with an array factor of the form $\sin(N\psi/2)/N\sin(\psi/2)$ has the narrowest pattern that can be achieved with an array. The uniform pattern is a useful tool for synthesis because it can be a member of an orthogonal set of beams. By devising lossless networks one can superimpose groups of beams in order to synthesize a desired pattern. A uniform array with N elements in equal distance d/λ produces a normalized beam pattern:

$$f_m(\theta) = b_m \frac{\sin(N\psi_m/2)}{N\sin(\psi_m/2)} \tag{108}$$

where

$$\psi_m = \beta d(\cos\theta - \cos\theta_m) \tag{109}$$

If we assume that a desired factor is the superposition of terms of the form (108), then

$$\mathrm{AF}(\theta) = \sum_{m=-M}^{M} b_m \frac{\sin(N\psi_m/2)}{N\sin(\psi_m/2)} \tag{110}$$

where

$$b_m = \mathrm{AF}(\theta_m) \tag{111}$$

and

$$\theta_m = \cos^{-1}\left(m\frac{\lambda}{Nd}\right) \tag{112}$$

The excitation of each element becomes

$$I_n = \frac{1}{N}\sum_{m=-M}^{M} \mathrm{AF}(\theta_m) e^{-j\beta d_n \cos\theta_m} \tag{113}$$

For a line source, we again can superimpose groups of beams of the form [6]

$$f_m(\theta) = b_m \frac{\sin[\beta L/2(\cos\theta - \cos\theta_m)]}{\beta L/2(\cos\theta - \cos\theta_m)} \tag{114}$$

The space factor is

$$\mathrm{SF}(\theta) = \sum_{m=-M}^{M} b_m \frac{\sin[\beta L/2(\cos\theta - \cos\theta_m)]}{\beta L/2(\cos\theta - \cos\theta_m)} \tag{115}$$

where

$$b_m = \mathrm{SF}(\theta_m) \tag{116}$$

and

$$\theta_m = \cos^{-1}\left(m\frac{\lambda}{L}\right) \tag{117}$$

The excitation distribution is

$$I(z') = \sum_{m=-M}^{M} b_m e^{-j\beta z' \cos\theta_m} \tag{118}$$

Instead of sampling $\mathrm{AF}(\theta)$ and $\mathrm{SF}(\theta)$ in θ_m we could apply the orthogonal method termed *orthosynthesis*. In this case

θ_m can be different from that in Eq. (112) or (117) and can have values that optimize the solution.

Figure 35 presents a cosecant-squared power pattern of a line source with $L = 10\lambda$. The same pattern with discrete elements ($N = 20$ and $N = 30$) is presented in Fig. 36. Figure 37 presents the desired of a modified cosecant-squared pattern and the pattern by orthosynthesis and WL for $N = 16$ and $d/\lambda = 0.5$. From the value of the mean-square error it appears that orthosynthesis is better than the WL.

13. ORTHOGONAL PERTURBATION METHOD

In the shape design of an array, an adjustment of the spacing between the elements can be made. A procedure that combines the adjustment with the orthogonal method is known as the *orthogonal perturbation method* [38,39]. A linear array has an array factor of the form

$$\mathrm{AF}_0(\theta) = \sum_{i=1}^{N} I_i e^{j\beta d_i \cos\theta} = \sum_{i=1}^{N} I_i \Phi_i(\theta) \tag{119}$$

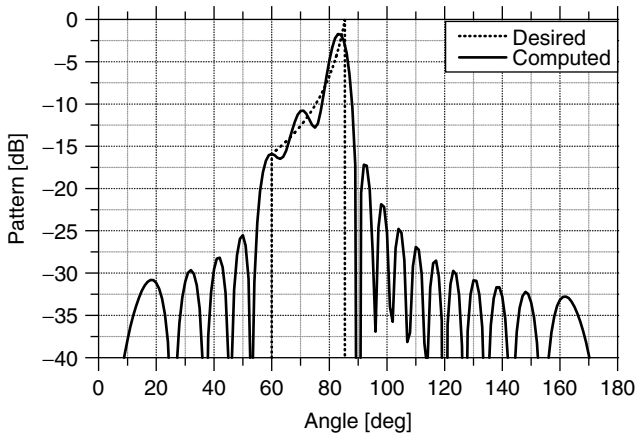


Figure 35. A cosecant-squared power pattern of a line source with $L = 10\lambda$ taken by WL method.

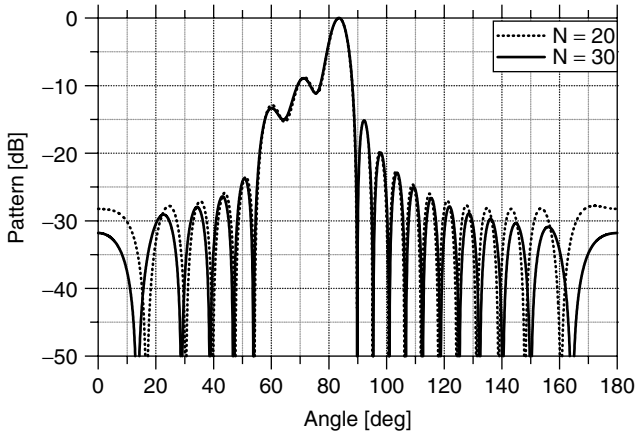


Figure 36. A cosecant-squared power pattern by sampling the line source with pattern of Fig. 35 to have $N = 20$ and 30 elements.

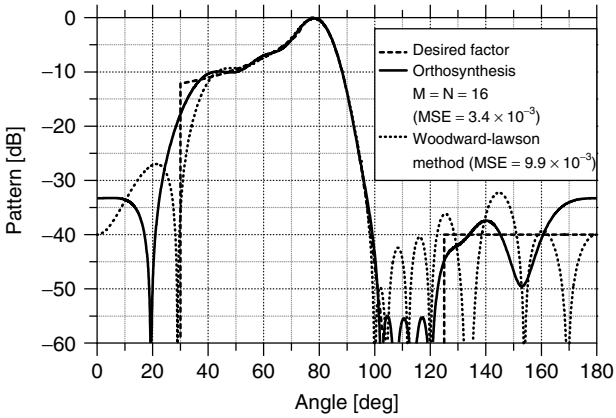


Figure 37. Cosecant-squared power pattern for $N = 16$ elements and $d/\lambda = 0.5$ by orthosynthesis and WL.

If we perturb the position d_i of each element such that $\beta(\delta d_i) \gg 1$, then the array factor becomes

$$\mathrm{AF}_1(\theta) \cong \sum_{i=1}^{N} [1 + j\beta(\delta d_i)\cos\theta] I_i e^{j\beta d_i \cos\theta} \tag{120}$$

Substituting (119) into (120) and dividing by $\cos\theta$ we have

$$F(\theta) = \frac{\mathrm{AF}_1(\theta) - \mathrm{AF}_0(\theta)}{\cos\theta} = \sum_{i=1}^{N} A_i \Phi_i(\theta) \tag{121}$$

where

$$A_i = j\beta(\delta d_i) I_i \tag{122}$$

It is clarified that for $\theta = \pi/2$, $F(\theta)$ is already kept equal to zero. By the orthogonal method we have

$$\left.\begin{aligned} \mathrm{AF}_0(\theta) &= \sum_{i=1}^{N} B_i^0 \Psi_i(\theta) \\ F(\theta) &= \sum_{i=1}^{N} B_i \Psi_i(\theta) \end{aligned}\right\} \tag{123}$$

I_i and A_i are

$$I_i = \sum_{j=i}^{N} B_j^0 c_i^{(j)} \tag{124}$$

$$A_i = \sum_{j=i}^{N} B_j c_i^{(j)} \tag{125}$$

Instead of I_i, we can use quantized approximate values for the initial array. After the quantization the array is perturbed and from $F(\theta)$ we take A_i, which gives δd_i. The perturbation continues by an iterative procedure until the desired approximation is achieved. If the result is not the expected, the procedure is repeated for a larger number of amplitudes.

An example with a Chebyshev pattern $T_5(x)$ with $\mathrm{SLL} = -30$ dB and $\mathrm{HPBW} = 15^\circ$ is presented. Three quantized amplitudes and 11 elements are used. The results are shown in Fig. 38 and Table 7.

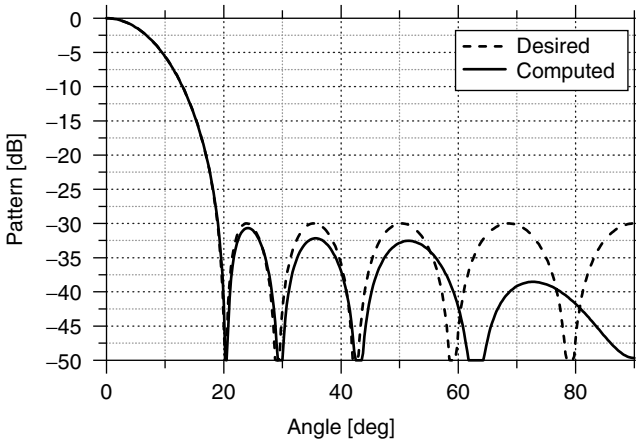


Figure 38. Chebyshev pattern with SLL = −30 dB and HPBW = 15° taken by the orthogonal perturbation method.

Table 7. Amplitudes and Positions of an 11-Element Array that Produces a Chebyshev Pattern

Element Number (i)	Distance (λ)	Quantized Current $I(i)$
1,11	±1.975	2
2,10	±1.603	2
3,9	±1.204	5
4,8	±0.800	5
5,7	±0.378	8
6	0	5

14. SYNTHESIS AS AN OPTIMIZATION PROBLEM

The antenna synthesis is mainly a nonlinear optimization procedure. In this procedure a convenient real function, which takes an optimum value at the reached properties of the desired antenna, is constructed. More than one function can be used to fit several antenna properties [40]:

1. Radiation pattern at a single frequency or at a number of frequencies.
2. Antenna impedance at a single frequency or at a number of frequencies.
3. Antenna index without or under constraint on another index.
4. Antenna impedance and/or radiation pattern in a given frequency range.
5. Coupling of antennas.

The optimization parameters may characterize the excitation, the shape, the size, the loadings, and the current distribution of the antenna elements. Any variation of some parameters requires completely new solutions.

Most of the optimization methods are divided into two categories. The first makes use of the values of the optimization function itself. The second looks at the gradient of the above function. The optimization function in some cases is not an explicit function but it is simply computed numerically.

Except for the abovementioned methods, procedures based on random search are available. These are based on the use of a random-number generator by which the successive points are determined. Finally, the simulated annealing and the genetic algorithms are two global optimizers.

15. OPTIMIZATION OF AN INDEX

An antenna index, I, such as directivity, gain, or quality factor, can be written as

$$I = \frac{[\tilde{a}]^*[A][a]}{[\tilde{a}]^*[M][a]} \tag{126}$$

where $[\tilde{a}]^* = [a_1, a_2, \ldots, a_N]^*$ is the conjugate transpose of $[a]$. By $[a]$ one can represent the current or the voltage excitation vector of the array. $[A]$ and $[M]$ with

$$\left.\begin{aligned}[A] &= [\alpha_{ij}] \\ [M] &= [m_{ij}]\end{aligned}\right\} \tag{127}$$

are both Hermitian $N \times N$ square matrices. Also $[M]$ is positive-definite.

An index I of the form (126) will be optimized under the constraint that another index I_1 is

$$I_1 = \frac{[\tilde{a}]^*[M_2][a]}{[\tilde{a}]^*[M_3][a]} = \gamma \tag{128}$$

According to several authors [41,42], a solution can be found by using the Lagrange multiplier and setting the quantity L

$$L = \frac{[\tilde{a}]^*[A][a]}{[\tilde{a}]^*[M][a]} + \lambda \left\{ \frac{[\tilde{a}]^*[M_2][a]}{[\tilde{a}]^*[M_3][a]} - \gamma \right\} \tag{129}$$

stationary with respect to $[a]$ and λ.

Zeroing the first variation of L, we have

$$[a] = q[K]^{-1}[\tilde{B}]^* \tag{130}$$

where q is a constant and

$$[K] = [M] + p\{\gamma[M_3] - [M_2]\} \tag{131}$$

where p is found from (128) by solving an eigenvalue equation [43,44]. If there is no constraint on I_1, $p = 0$. In the optimization procedure, pattern values and index constraints can also be combined.

An example of a wire dipole array shown in Fig. 39 with maximum gain G_1 in $\theta_1 = 0^\circ$ at the frequency f_1 under the constraint that at $f_2 = f_1/2$ the gain is G_2 in the direction θ_2 is presented in Fig. 40.

An interesting case is the optimization of the directivity of a 21-element array of 1λ length, which gives $D = 48.77$, $Q = 2.258875 \times 10^5$, and $\eta = 1.028 \times 10^{-13}\%$. The minimum to maximum excitation is 4.2044×10^{-4}. A five-element uniform array with the same length has maximum $D = 5$, $Q = 1$, and $\eta = 100\%$. The first array is superdirective, usually known as *supergain*. In supergain arrays the ohmic losses are extremely large. This is the

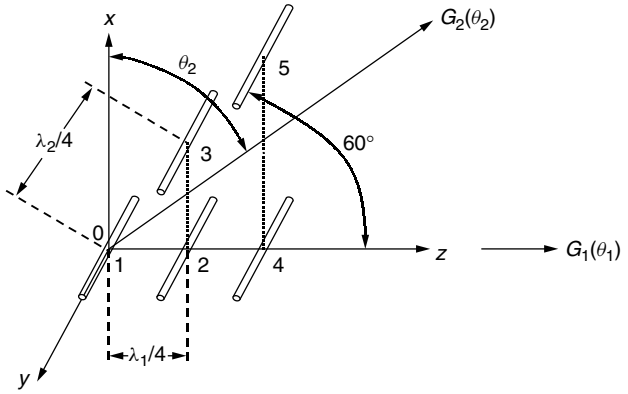


Figure 39. A 5-parallel-wire dipole triangular array.

penalty than one must pay in loss of efficiency if reduction of length is important.

16. OPTIMIZATION BY SIMPLEX AND GRADIENT METHODS

Simplex and gradient [45–47], are both local optimizer methods. A *simplex* is a body in multidimensional space. The optimization function at the vertices of a simplex is computed. On this basis, a new smaller simplex is chosen within which an optimum should be situated. The optimum depends on the initial simplex [45].

Gradient methods are known as steepest-descent methods. A starting point is chosen and the direction where the optimization function decreases most rapidly is found. Adopting a new point in that direction at a desired distance and repeating the process, a minimum of the optimization function is achieved.

It is noticed that it is difficult to judge if the minimum is the global one or a local one. From an antenna engineering point of view, we are usually interested in a suitable solution and not necessarily the global optimum.

The following example illustrates the optimization of the radiation pattern and the antenna impedance of a four-element Yagi–Uda array. For the pattern, the antenna gain obtained was larger than a prescribed value. For the impedance, the mutual coupling was taken into account. Use of initial values for the optimization parameters was based on experience. After 12 simplex iterations, an antenna was obtained with (see Fig. 41) $L_R = 0.50\lambda$, $L_F = 0.468\lambda$, $L_D = 0.45\lambda$, $d_r = 0.25\lambda$, $d = 0.30\lambda$, $a = 0.002\lambda$.

It was found that $G = 9.15$ dB and $Z_{in} = (36 + j0)\Omega$. The front-to-back-ratio was 14.1 dB and the HPBW = 65.2°. Figure 42 shows the pattern of the antenna. A similar optimization can be applied for log-periodic dipole arrays [9].

17. OPTIMIZATION BY SIMULATED ANNEALING METHODS

The basic idea in simulated annealing (SA) is to combine local search with Monte Carlo techniques in analogy

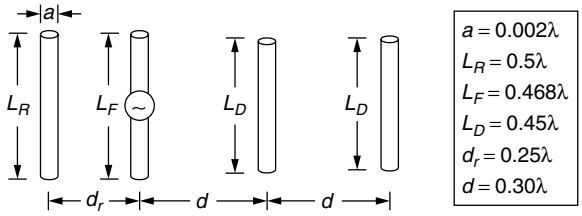


Figure 41. Yagi–Uda antenna with four elements.

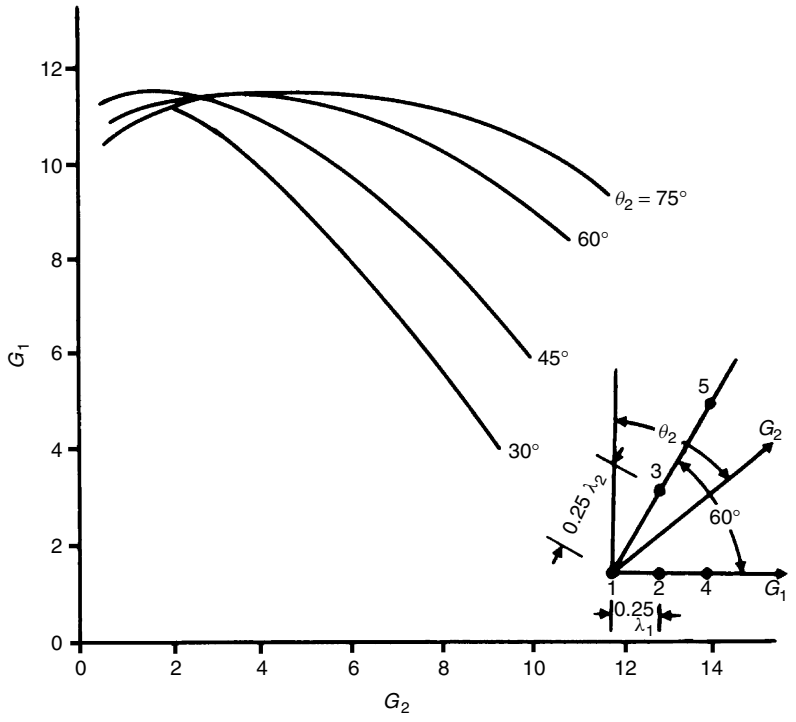


Figure 40. Maximum gain G_1 in frequency f_1 versus G_2 in frequency $f_1/2$. $\theta_1 = 0°$ and θ_2 is 30°, 45°, 60°, and 75°.

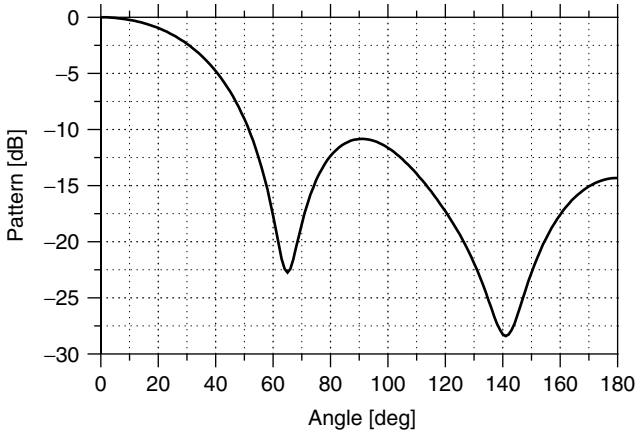


Figure 42. H pattern of the optimized four-element Yagi antenna with $G = 9.15$ dB, $F/B = 14.1$ dB and $Z_{\text{in}} = 36\Omega$.

to cooling processes in thermodynamics. *Simulated annealing* [48] refers to a process used to reveal the low-temperature state of some material. At high temperatures the molecules of a liquid move freely with respect to one another. If the liquid is cooled slowly, the thermal mobility is lost. The atoms are able to line up in a crystal, which represents the minimum-energy state for the system. The time spent at each temperature must be sufficiently long to allow a thermal equilibrium to be realized. If the system is cooled quickly, it does not reach the minimum energy state but one having higher energy.

In optimization by SA we simulate the annealing process by a Monte Carlo method where the global minimum of the objective function represents the low-energy configuration [49,50].

Variations of the simulated annealing process can include parallelization techniques with the use of multiple CPUs [51]. SA has been used in various combinatorial optimization problems.

An example of an array of eight $\lambda/2$ collinear dipoles is presented. The initial array is a uniform array with $d/\lambda = 0.93$ and phase shift $\alpha = 52^\circ$. The main-beam maximum is at $\theta = 99^\circ$. It is observed that in the area $\theta \le \pi/2$, SLL ≥ -10 dB occurs. In practice, this is not desirable. Mobile and radio stations aim at lower upward of horizon. By using the appropriate cost function with the geometry constraints for SLL ≤ -18 dB in $\theta \le \pi/2$, a new array is found. Table 8 and Fig. 43 present the array and the pattern. Applications for wire antenna arrays as well as slot arrays can be found in the literature [52,53].

Table 8. Antenna Array with Collinear Dipoles by Simulated Annealing

Dipole Number	Dipole Position	Dipole Phase (degree)
1	0 λ	0
2	0.70 λ	73
3	1.53 λ	116
4	2.37 λ	155
5	3.23 λ	−172
6	4.05 λ	−130
7	4.92 λ	−93
8	5.62 λ	−19

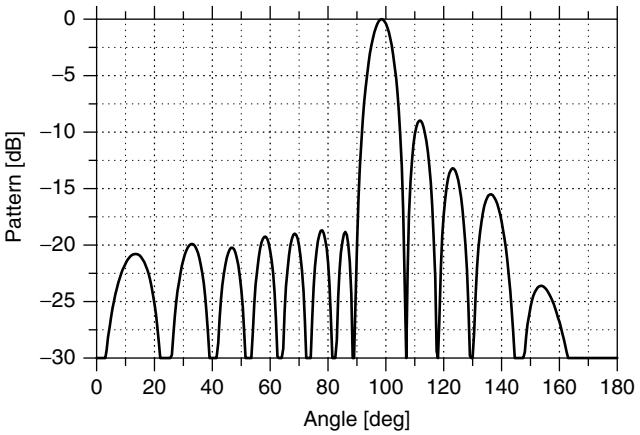


Figure 43. E pattern of $N = 8$ collinear dipole for SLL ≤ -18 dB at $\theta \le \pi/2$ and maximum at $\theta = 99^\circ$.

18. OPTIMIZATION BY GENETIC ALGORITHMS (GAs)

Genetic algorithms (GAs) are global optimizers. GAs [54] follow two main principles: the ability to encode complex structures and the use of simple transformations to improve such structures. GAs are well suited for a wide range of problems in electromagnetics [54]. They have the advantage of quick and easy programming and implementation. They are also suitable for constrained optimization. GAs are based on Darwin's principle [55]: survival of the fittest. The basic idea is an analogy between an individual and a solution on one hand and between an environment and a given problem on the other. The function to be minimized or maximized represents the fitness. This is computed for a given individual and determines how that person "fits" or, in other words, how good this solution for the given problem is.

Many categories of GAs have been designed. A simple GA [56] has nonoverlapping populations. Very popular for electromagnetics are the steady-state GAs with overlapping populations. The best individuals survive to the next generation. Another approach is the deme GA [57], which involves parallel evolving populations with migration.

GAs have been successfully applied in many engineering problems [58–61]. GAs can be applied to thinned arrays. A *thinned array* is a subset of aperiodic arrays. Thinning an array means turning off some elements in a uniformly spaced or periodic array. The OFF elements remain in the array, so the mutual coupling for the interior elements remains the same.

A thinned array offers essentially the same beamwidth with less directivity and fewer elements than does a uniform array of the same size [13,16,20]. The most realistic applications of GAs to array thinning have to do with optimizing the SLL of a large number of elements.

Concluding GAs, an example of an 11-element endfire (case 1) array with SLL $= -20$ dB and HPBW $\cong 72^\circ$ is presented. The elements have the same amplitude. Figure 44

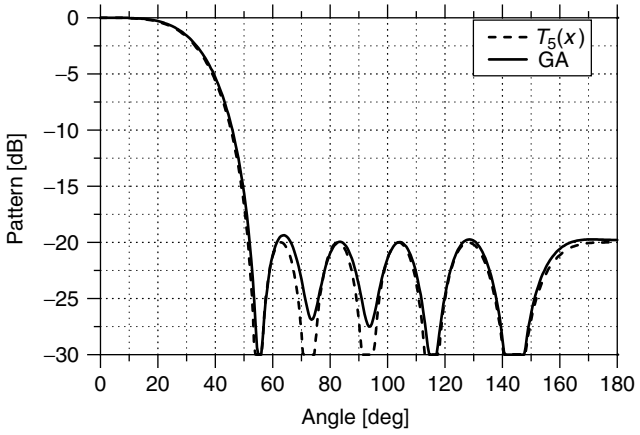


Figure 44. Pattern of 11-element Chebyshev endfire array with SLL = −20 dB, HPBW = 72° (case 1).

Table 9. Position and Phase of an 11-Element Array Able to Create a $T_5(x)$ End-Fire Pattern

Element Position	Distance (λ)	Phase (degree)
1	0.000	0
2	0.400	−140.1
3	0.719	111.8
4	0.855	47.4
5	1.080	−36.4
6	1.300	−108.4
7	1.611	156.8
8	1.860	88.5
9	2.025	10.8
10	2.131	−63.7
11	2.410	−156.5

presents the pattern, and Table 9 gives the position and the phase of the elements.

19. SPACE AND TIME OPTIMIZATION AND SMART ANTENNAS

Antenna arrays combined with signal processing in space and real time are known as "smart" antennas. The low-cost and fast digital processors now available have made possible the implementation of smart antennas. Smart antennas can be used with great success in cellular and satellite mobile communications. They improve the system performance by increasing the spectrum efficiency and the channel capacity. They also extend the range of coverage by multiple-beam steering and electronic compensation of the distortion. Smart antennas can reduce propagation problems such as multipath fading, cochannel interference, and delay spread as well as communication indices such as bit error rate (BER) and outage probability. Their main advantage is the capability to provide a certain channel at a certain direction. This results in *spatial-division multiple access* (SDMA), which performs differently from the frequency (FDMA), the time (TDMA), and the code (CDMA) division multiple accesses.

Smart antennas are known as *adaptive arrays*, *intelligent antennas*, *spatial processing*, *digital beamforming antennas*, and by other terms, [62]. They direct their main-beam maximum to the user while the pattern nulls are in the direction of possible interference [17]. Two main types of beam patterns are available: (1) the switched-beam and (2) the adaptive system. The switched beam divides the communication sector in microsectors. Each microsector contains a predetermined fixed beam pattern. The adaptive systems dynamically alter the patterns to optimize the communications performance. They utilize sophisticated signal processing algorithms [17,63], which update the beam patterns on the basis of changes in both the desired and the interfering signal directions.

Adaptive array theory is based on the optimization methods given before and on the real-time response in a transient environment.

In the literature one can find a lot of special journal issues, books, and specialized research papers in the area of smart antennas [64–69].

Consider a uniform linear array immersed in a homogeneous medium in the far field of M uncorrelated sinusoidal point sources of frequency f_0 (see Fig. 45).

The time difference taken of a plane wave, coming from the ith source in the direction (θ_i, φ_i), to arrive from the kth element to the origin, is

$$r_k = \frac{d}{c}(k-1)\cos\theta_i \tag{132}$$

The signal induced on the first element due to the same source is $m_i(t)e^{j2\pi f_0 t}$. The function $m_i(t)$ depends on the type of the access used:

$$m_i(t) = A_i e^{j\xi_i(t)} \quad \text{(frequency modulation for FDMA)} \tag{133}$$

$$m_i(t) = \sum_n d_i(n)p(t-n\Delta) \quad \text{(TDMA)} \tag{134}$$

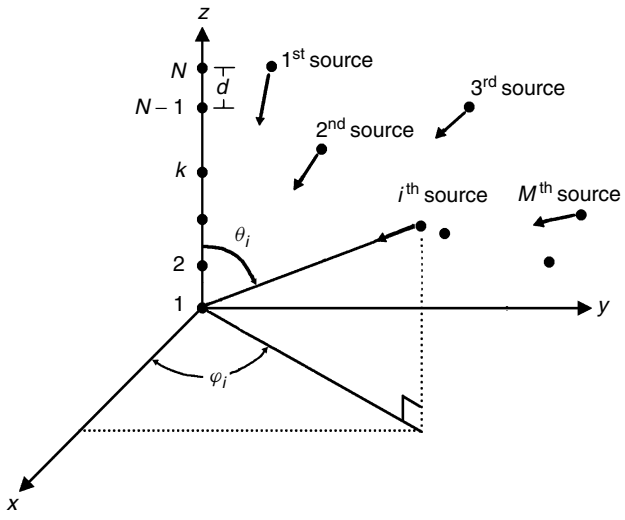


Figure 45. Signal model of a linear array.

where $p(t)$ is the sampling pulse, $d_i(n)$ is the message symbol, and Δ is the sampling interval:

$$m_i(t) = d_i(t)g(t) \quad \text{(CDMA)} \tag{135}$$

where $d_i(t)$ is the message sequence and $g(t)$ is a pseudorandom binary sequence.

The signal induced at the kth element is $m_i(t)e^{j2\pi f_0(t+r_k)}$. It is assumed that the bandwidth of the signal is narrow enough and the array dimensions are small enough for the modulating function $m_i(t)$ to stay almost constant during r_k.

The total signal induced at the kth element due to all sources plus the noise $n_k(t)$ is

$$x_k = \sum_{i=1}^{M} m_i(t)e^{j2pf_0(t+r_k)} + n_k(t) \tag{136}$$

Let us now consider a narrowband beamformer where signals from each element are multiplied by a complex weight and summed to form the array output $y(t)$ (see Fig. 46):

$$y(t) = \sum_{k=1}^{N} w_k^* x_k(t) = [\tilde{w}^*][x(t)] \tag{137}$$

where $[w]$ and $[x(t)]$ are column vectors containing the weights and the inputs of the elements of the array. The values of $[w]$ are determined by using one of the optimization methods.

Smart antennas are analyzed for different network topologies and mobility scenarios. The array geometries

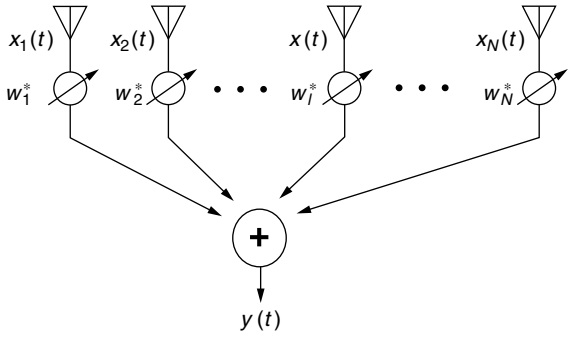


Figure 46. Narrowband beamformer structure.

will be realized with the feed networks and the algorithms for fast beamforming and direction of arrival. A smart antenna system is presented in Fig. 47.

Smart antennas have undergone significant progress since the early 1990s, and their future looks bright. Cost will continue to be the most critical point. It is believed that an explosive development of array processing algorithms within communication systems will appear.

20. ELEMENT PATTERN AND MUTUAL COUPLING

Analysis and synthesis of antenna arrays are given for array elements with known current or aperture field characteristics. It was assumed that these characteristics are proportional to the excitations, the same for similar elements, and unchanged as the array is scanned. In general, all the currents and fields differ in magnitude,

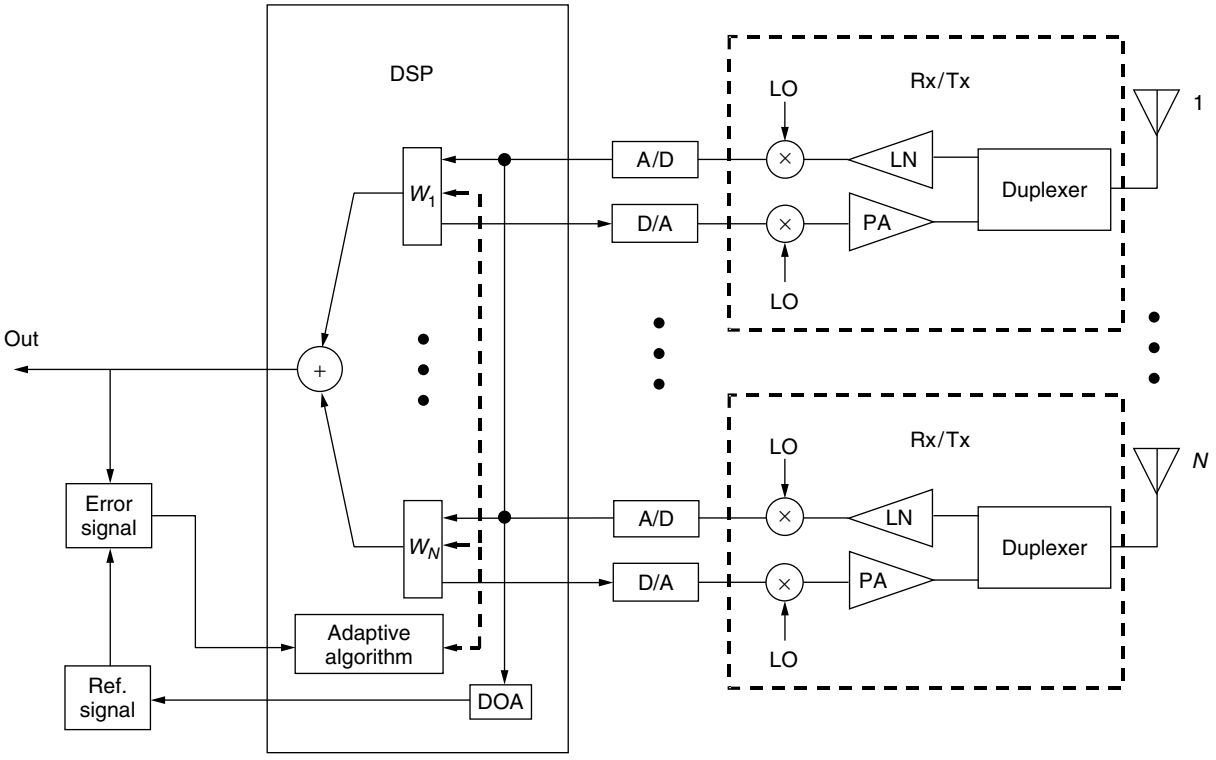


Figure 47. Diagram of a smart antenna system.

phase, and distribution from element to element. The differences depend on the frequency and the scan angle as well as on the geometry of the array. That happens because mutual coupling plays an important role in the behavior of the elements. Actually the radiated field can be expressed as generalized integrals that include the appropriate distributions over the radiating antenna elements and nearby diffracting bodies. The array characteristics are dominated by the mutual coupling between the elements.

Mutual coupling alters mainly the amplitudes and phases between various elements while the currents or aperture distributions remain very similar. In the antenna array synthesis the required distributions can be found by using different methods depending on the problem. Among these, the boundary-value, the transmission line, and the Poynting vector methods are the main ones [1,3,6]. In the late 1960s the integral equations with suitable numerical solutions were successively applied. The numerical techniques are collectively referred to as the *method of moments* (MoM), [6,41,70]. This method is simple and versatile and requires fast and large amounts of computation. The speed and storage capacity of the computer characterize the limitation of the method.

There are several forms of integral equations. For the time-harmonic EM fields, the electric (EFIE) and the magnetic field (MFIE) integral equations are popular [71]. The EFIE enforces the electric field, while the MFIE enforces the magnetic field boundary condition. MoM reduces the integral equations to a system of simultaneous linear algebraic ones in terms of the unknown current or aperture distribution. For radiation problems, especially for wire antennas, there are popular integral equations as the Pocklington, the Hallen, and the reaction integral equations. There are computer codes for the evaluation of the radiation characteristics of antennas. They make use of the abovementioned equations and compute the appropriate quantities in the near and far fields.

20.1. Finite and Infinite Arrays

Let a wire structure (Fig. 48) be composed of straight segments of circular cross section. For electrically thin wires it was found that the total current (conduction plus displacement) on the structure can be found by using one of the electric field integral equations [70,71]. The current on the wires is expanded in a finite series as follows:

$$I(\ell) = \sum_{n=1}^{N} I_n F_n(\ell) \tag{138}$$

where $F_n(\ell)$ are the current expansion functions.

Substituting (138) into the integral equation used, the following system of simultaneous linear algebraic equations yields

$$\sum_{n=1}^{N} I_n Z_{mn} = V_m \quad m = 1, 2, 3, \ldots, N \tag{139}$$

where V_m are the applied voltages and I_n are the complex amplitudes of the current distribution. The elements Z_{mn} are the mutual impedance elements. Equation (139) can be expressed in matrix form:

$$[Z][I] = [V] \tag{140}$$

The only nonzero elements V_m of $[V]$ are these where a generator is at the terminals of the mth segment.

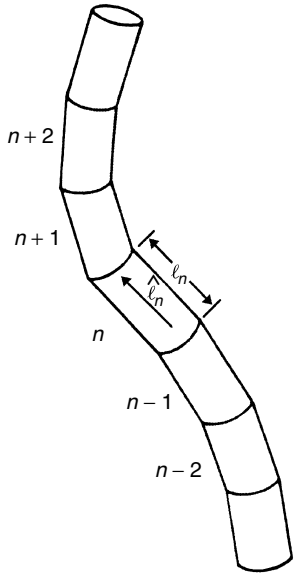


Figure 48. Wire structure of straight segments.

Let us assume that a −20-dB Chebyshev broadside array with five parallel $\lambda/2$ dipoles is desired. Table 10 shows the required currents and the corresponding voltages at the main ports for equal spacing $d = 0.25\lambda$ and $d = 0.5\lambda$. The resulting pattern for $d = 0.25\lambda$ is shown in Fig. 49. If we suppose that there is no coupling between the elements, then the currents have the same relative values as the voltages. In this case the resulting pattern is also given in Fig. 49 and is very different from the desired.

In addition to the numerical methods, one could use measured data to evaluate the mutual coupling effects. In this case the currents and voltages can be found by using one of the classical methods of synthesis [72].

The prediction of element impedance as a function of scan and element patterns in an infinite array is very different in a finite one. Elements away from the edge of large finite arrays have approximately the same characteristics to these of infinite arrays. So, the study of infinite arrays has a practical aim.

Table 10. Relative Input Currents and Voltages for a Five-Element Chebyshev Broadside Array with SLL = −20 dB

	I_i		V_i	
Element	$d/\lambda = 0.25$	$d/\lambda = 0.5$	$d/\lambda = 0.25$	$d/\lambda = 0.5$
1, 5	1	1	$1\angle 0°$	$1\angle 0°$
2, 4	−1.194	1.608	$1.902\angle 251°$	$1.383\angle 341°$
3	2.178	1.932	$3.027\angle 32°$	$1.804\angle -6°$

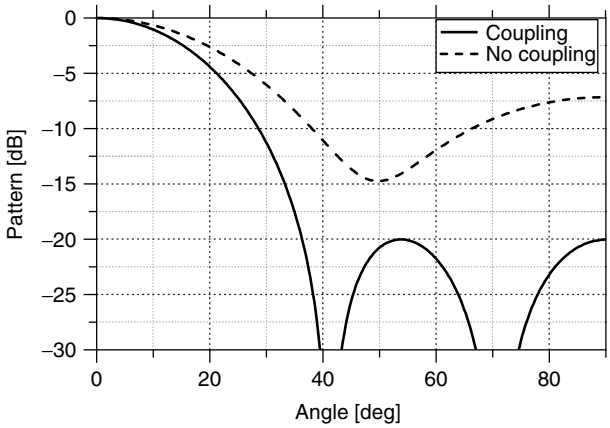


Figure 49. Pattern of 5 parallel $\lambda/2$ dipole linear array with $d = 0.25\lambda$.

In infinite arrays a wave-type formulation or a mode-matching approach with a direct solution of the differential equations can be used. Any of the abovementioned approaches is based on the periodic nature of the fields.

Infinite array theory is a good approximation of the impedance behavior of central elements in large arrays [6,13,16].

21. ARRAY FEEDS

Linear arrays or assemblies thereof making planar arrays are the most usual fixed-beam arrays. Using linear arrays as the building blocks, appropriate feed networks are developed. Two kinds of feeds are more usual: the "series" and the "shunt" ones.

In the series feed the elements of the array are in series along the transmission line. Similarly, in the shunt feed the elements are in parallel with the line or the network. Feeds must offer an acceptable in-band performance in relatively modest cost. The feed choice depends on the application as well as on its physical, processing, and electrical properties. The weight with the conformity and the material used characterize the physical properties. The fabrication and the availability of the materials define the processing properties. Finally the losses, the shielding, the design ability, and the performance over a specified bandwidth characterize the electrical properties. The ability of the array feed to control the power distribution allows the antenna engineer to meet the appropriate requirements.

A critical function of a feed network is that of impedance matching. By matching the impedances as closely as possible at each portion of the network, the reflection coefficients and therefore the VSWR of the feed is kept to with certain levels.

The feed network must keep the isolation between outputs. This means that any energy entering in the ith output port should not reappear at any other via the network.

An array feed should have the ability to steer its main beam. This is accomplished by using discrete phase shifters and attenuators located between the outputs of the feed network and the elements of the array.

The most common shunt feed is the corporate one (Fig. 50). A series feed can be constructed by using the transmission line junctions (Fig. 51). Multiple-beam feeds are made by series-fed beamforming networks (Fig. 52) or by parallel feed as the Butler matrix (Fig. 53). Planar arrays use arrangement of series-series or series–parallel topologies [73–79].

Finally, optical hardware for the care and feeding of an array can be used. An extended analysis of photonic feed systems can be found in the literature [80,81].

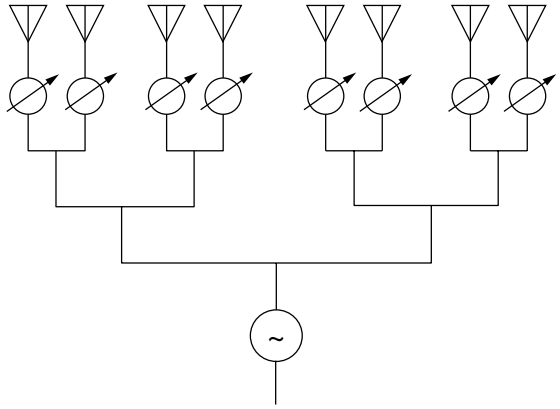


Figure 50. Parallel corporate feed.

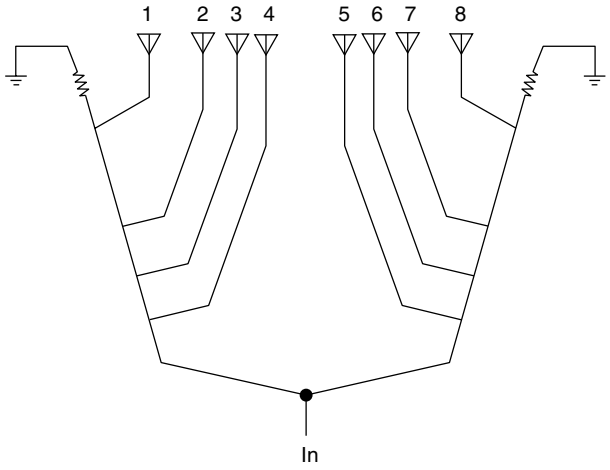


Figure 51. Series feed with transmission line junctions.

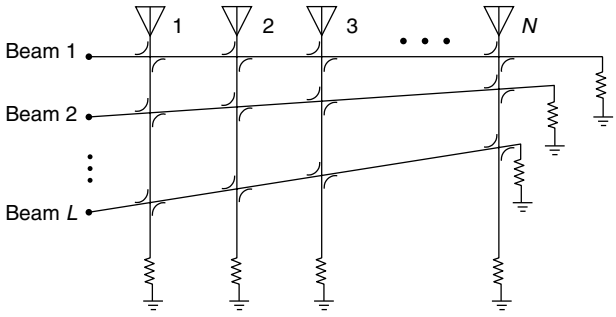


Figure 52. Series-fed beamforming network.

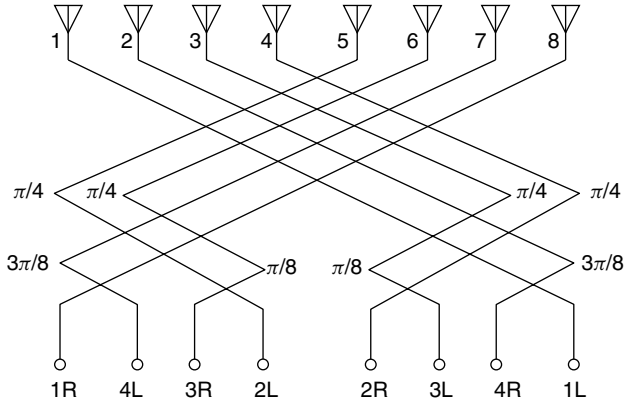


Figure 53. Eight beams and elements Butler matrix.

BIOGRAPHY

John N. Sahalos received his B.Sc. degree in physics and the Diploma in civil engineering from the University of Thessaloniki, Greece, in 1967 and 1975, respectively. He also received the Diploma of postgraduate studies in electronics in 1975 and a Ph.D. in electromagnetics in 1974. During 1976 he was with Electroscience Laboratory, the Ohio State University, Columbus, as a postdoctoral university fellow. From 1977–1985 he was a professor in the Electrical Engineering Department, University of Thrace, Greece. Since 1985 he has been a professor at the School of Science, University of Thessaloniki, Greece. During 1982 and 1989, he was a visiting professor at the University of Colorado, Boulder, and at the Universidad Politecnica de Madrid, Spain, correspondingly. He is the author of three books and more than 200 articles published in the scientific literature. His research interests are in the area of applied electromagnetics, antennas, high-frequency methods, communications, microwaves, and biomedical engineering.

Dr. Sahalos is a professional engineer and a consultant to industry. He is on the editorial board of two scientific journals. Since 1999 he has been the president of the Greek URSI committees, is a member of five IEEE Societies, and a member of both the New York Academy of Science and the Technical Chamber of Greece.

BIBLIOGRAPHY

1. S. A. Schelkunoff and H. T. Friis, *Antenna Theory and Practice*, Wiley, New York, 1952.
2. R. W. P. King, *The Theory of Linear Antennas*, Harvard Univ. Press, Cambridge, MA, 1956.
3. J. D. Kraus, *Antennas*, McGraw-Hill, New York, 1988.
4. R. C. Hansen, ed., *Microwave Scanning Antennas*, Academic Press, New York, Vol. I, 1964; Vols. II, III, 1966. (Peninsula Publishing, 1985).
5. R. S. Elliot, *Antenna Theory and Design*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
6. C. A. Balanis, *Antenna Theory*, Analysis and Design, Wiley, New York, 1997.
7. T. Milligan, *Modern Antenna Design*, McGraw-Hill, New York, 1985.
8. N. Amitay, V. Galindo, and C. P. Wu, *Theory and Analysis of Phased Arrays*, Wiley-Interscience, New York, 1972.
9. M. T. Ma, *Theory and Applications of Antenna Arrays*, Wiley-Interscience, New York, 1974.
10. A. W. Rudge, K. Milne, A. D. Olver, and P. Knight, eds., *The Handbook of Antenna Design*, IEE/Peregrinus, London, 1983.
11. Y. T. Lo and S. W. Lee, *Antenna Handbook*, Van Nostrand Reinhold, New York, 1988.
12. R. C. Johnson and H. Jasik, *Antenna Engineering Handbook*, McGraw-Hill, New York, 1993.
13. R. C. Mailloux, *Phased Array Antenna Handbook*, Artech House, Norwood, MA, 1994.
14. J. R. James and P. S. Hall, eds., *Handbook of Microstrip Antennas*, Vols. I, II, IEE/Peter Peregrinus, London, 1989.
15. N. Fourikis, *Phased Array-Based Systems and Applications*, Wiley-Interscience, New York, 1997.
16. R. C. Hansen, *Phased Array Antennas*, Wiley-Interscience, New York, 1998.
17. R. T. Compton, Jr., *Adaptive Antennas*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
18. T. S. Rappaport, ed., *Smart Antennas*, IEEE Press, 1998.
19. G. V. Tsoulos, ed., *Adaptive Antennas for Wireless Communications*, IEEE Press, 2001.
20. Y. Rahmat-Samii and E. Michielssen, *Electromagnetic Optimization by Genetic Algorithms*, Wiley-Interscience, New York, 1999.
21. M. Abramowitz and L. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.
22. C. L. Dolph, A current distribution for broadside arrays which optimizes the relationship between beam width and side-lobe level, *Proc. IRE* **34**: 335–338 (1946).
23. H. J. Riblet, Discussion on a current distribution for broadside arrays which optimizes the relationship between beam width and side-lobe level, *Proc. IRE* **35**: 489–492 (1947).
24. G. Miaris, M. Chryssomalis, E. Vafiadis, and J. N. Sahalos, A unified formulation for Chebyshev and Legendre superdirective end-fire array design, *Archiv Elektrotechnik* **78**(4): 271–280 (1995).
25. W. H. Kummer, general ed., *IEEE Trans. Antennas Propag.* (Special Issue on Conformal Arrays), **AP-22**(1) (1974).
26. R. S. Elliot, On discretizing continuous aperture distributions, *IEEE Trans. Antennas Propag.* **AP-25**(5): 617–621 (1977).
27. T. T. Taylor, Design of line source antennas for narrow beamwidth and low sidelobes, *IEEE Trans. Antennas Propag.* **AP-3**: 16–28 (1955).
28. A. T. Villeneuve, Taylor patterns for discrete arrays, *IEEE Trans. Antennas Propag.* **AP-32**(10): 1089–1093 (1984).
29. R. C. Hansen, Linear arrays, in A. Rudge, ed., *Handbook of Antenna Design*, Vol. 2, Peter Peregrinus, London, 1983, Chap. 9.
30. E. T. Bayliss, Design of monopulse antenna difference patterns with low sidelobes, *Bell Syst. Tech. J.* **47**: 623–650 (1968).
31. D. K. Cheng, *Analysis of Linear Systems*, Addison-Wesley, Reading, MA, 1959.

32. S. R. Laxpatti, Planar array synthesis with prescribed pattern nulls, *IEEE Trans. Antennas Propag.* **AP-30**(6): 1176–1183 (1982).

33. J. N. Sahalos, The orthogonal method of nonuniformly spaced arrays, *Proc. IEEE* **62**: 281 (1974).

34. H. Unz, Nonuniformly spaced arrays: The orthogonal method, *Proc. IEEE* **54**: 53–54 (1966).

35. J. N. Sahalos, A solution of nonuniformly linear array with the help of the Chebyshev polynomials, *IEEE Trans. Antennas Propag.* **AP-24**: 109–112 (1976).

36. J. N. Sahalos, K. Melidis, and S. Lampou, On the optimum directivity of general nonuniformly spaced broadside arrays of dipoles, *Proc. IEEE* **64**: 1706–1709 (1974).

37. J. N. Sahalos, A solution of the general nonuniformly spaced antenna array, *Proc. IEEE* **64**: 1292–1294 (1976).

38. G. Miaris and J. N. Sahalos, The orthogonal method for the geometry synthesis of a linear antenna array, *IEEE-AP Mag.* **41**(1): 96–99 (1999).

39. S. Goudos, G. Miaris, and J. N. Sahalos, On the quantized excitation and the geometry synthesis of a linear array by the orthogonal method, *IEEE Trans. Antennas Propag.* **AP-49**(2): 298–305 (2001).

40. B. D. Popovic, M. B. Dragovic, and A. R. Djordjevic, *Analysis and Synthesis of Wire Antennas*, Research Studies Press, Wiley, New York, 1982.

41. R. F. Harrington, *Field Computations by Moment Method*, IEEE Press, New York, 1993.

42. Y. T. Lo, S. W. Lee, and Q. H. Lee, Optimization of directivity and SNR of an arbitrary antenna array, *Proc. IEEE* **54**(8): 1033–1045 (1966).

43. L. P. Winkler and M. Schwartz, A fast numerical method for determining the optimum SNR of an array subject to a Q factor constraint, *IEEE Trans. Antennas Propag.* **AP-20**(4): 503–505 (1972).

44. P. Zimourtopoulos and J. N. Sahalos, On the gain maximization of the dual frequency and direction array consisting of wire antennas, *IEEE Trans. Antennas Propag.* **AP-33**: 874–880 (1985).

45. J. A. Nelder and R. Mead, A simplex method for function minimization, *Comput. J.* **7**: 308–313 (1965).

46. S. L. S. Jacoby, J. S. Kowalik, and J. T. Pizzo, *Iterative Methods for Nonlinear Optimization Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

47. P. R. Abdy and M. A. H. Dempster, *Introduction to Optimization Methods*, Chapman & Hall, London, 1974.

48. N. Metropolis, A. W. Rosenbluth, A. H. Teller, and E. Teller, Equations of state calculations by fast computing machines, *J. Chem. Phy.* **21**: 1087–1091 (1953).

49. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge Univ. Press, 1992.

50. A. Torn and A. Zilinskas, *Global Optimization, Lecture Notes in Computer Science*, Springer-Verlag, 1987.

51. R. Azencott, *Simulated Annealing Parallelization Techniques*, Willey, New York, 1992.

52. Z. Zaharis, E. Vafiadis, and J. N. Sahalos, On the design of a dual-band base station wire antenna, *IEEE Antennas Propag. Mag.* **42**(6): 144–151 (2000).

53. K. Kechagias, E. Vafiadis, and J. N. Sahalos, On the RLSA antenna optimum design for DBS reception, *IEEE Trans. Broadcast.* **44**(4): 460–469 (1998).

54. Y. Rahmat-Samii and E. Michielssen, *Electromagnetic Optimization by Genetic Algorithms*, Willey, New York, 1999.

55. C. Darwin, *On the Origin of Species*, John Murray, London, 1859.

56. D. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989.

57. M. Wall, *GAlib: A C++ Library of Genetic Algorithm Components*, Version 2.4, Document Revision B, MIT, 1996.

58. J. H. Holland, *Adaptation in Natural and Artificial Systems*, Univ. Michigan Press, 1975.

59. M. Mitchell, *An Introduction to Genetic Algorithms*, 2nd Pr., MIT Press, 1996.

60. L. Chambers, *Practical Handbook of Genetic Algorithms*, Vol. I, *Applications*, CRC Press, Boca Raton, FL, 1995.

61. K. F. Man, K. S. Tang, and S. Kwong, *Genetic Algorithms*, Springer-Verlag, London, 1999.

62. M. Chryssomallis, Smart antennas, *IEEE Antennas Propag. Mag.* **42**(3): 129–136 (2000).

63. J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, 2nd edn., Macmillan, New York, 1992.

64. R. Schreiber, Implementation of adaptive array algorithms, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**: 1038–1045 (1986).

65. S. Choi and T. K. Sarkar, Adaptive antenna array utilizing the conjugate gradient method for multipath mobile communications, *Signal Process.* **29**: 319–333 (1992).

66. A. El Zooghby, C. G. Christodoulou, and M. Georgiopoulos, Neural network-based adaptive beamforming for one and two dimensional antenna arrays, *IEEE Trans. Antennas Propag.* **AP-46** (1998).

67. Th. S. Rappaport, ed., *Smart Antennas: Adaptive Arrays, Algorithms and Wireless Position Location*, IEEE Press, 1998.

68. L. C. Godara, Application of antenna arrays to mobile communications, Part I: Performance improvement, feasibility and system considerations, *Proc. IEEE* **85**(7): 1031–1060 (1997).

69. L. C. Godara, Application of antenna arrays to mobile communications, Part II: Beam-forming and direction-of-arrival considerations, *Proc. IEEE* **85**(8): 1195–1245 (1998).

70. W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, Wiley, New York, 1998.

71. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.

72. H. Steyskal and J. S. Herd, Mutual coupling compensation in small array antennas, *IEEE Trans. Antennas Propag.* **AP-38**(12): 1971–1975 (1990).

73. L. Young, *Parallel Coupled Lines and Directional Couplers*, Artech House, Norwood, MA, 1992.

74. R. S. Elliot, An improved design procedure for small arrays of shunt slots, *IEEE Trans. Antennas Propag.* **AP-32**: 48–53 (1983).

75. R. C. Hansen and G. Brunner, Dipole mutual impedance for design of slot arrays, *Microwave J.* **22**: 54–56 (1979).

76. N. A. Begovich, Frequency scanning, in R. C. Hansen, ed., *Microwave Scanning Antennas*, Vol. III, Peninsula Publishing, 1983, Chap. 2.

77. J. S. Ajioka, Frequency-scan antennas, in R. C. Johnson, ed., *Antenna Engineering Handbook*, McGraw-Hill, New York, 1993, Chap. 19.

78. J. L. Butler, Digital, Matrix and intermediate frequency scanning, in R. C. Hansen, ed., *Microwave Scanning Antennas*, Peninsula Publishing, 1985, Chap. 3.

79. J. R. James and P. S. Hall, *Handbook of Microstrip Antennas*, Vols. 1, 2, IEE, Peter Peregrinus, London, 1989.

80. H. Zmuda and E. N. Toughlian, eds., *Photonic Aspects of Modern Radar*, Artech House, Norwood, MA, 1994.

81. A. Kumar, *Antenna Design with Fiber Optics*, Artech House, Norwood, MA, 1996.

ANTENNA MODELING TECHNIQUES

JOHN L. VOLAKIS
University of Michigan
Ann Arbor, Michigan

THOMAS F. EIBERT
T-Systems Nova GmbH
Technologiezentrum
Darmstadt, Germany

1. INTRODUCTION

Antennas are key components in any wireless communication system [1,2]. They are the devices that allow for the transfer of a signal (in a wired system) to waves that in turn propagate through space and can be received by another antenna. The receiving antenna is responsible for the reciprocal process: that of turning an electromagnetic wave into a signal or voltage at its terminals that can subsequently be processed by the receiver. The receiving and transmitting functionalities of the antenna structure itself are fully characterized by Maxwell's equations [3] and are fairly well understood. The dipole antenna (a straight wire fed at the center by a 2-wire transmission line) was the first antenna ever used and is also one of the best understood [1,2]. For effective reception and transmission, it must be approximately $\lambda/2$ long (λ = wavelength) at the frequency of operation (or multiples of this length). Thus, it must be fairly long when used at low frequencies ($\lambda = 1$ m at 300 MHz), and even at higher frequencies (UHF and above), its protruding nature makes it quite undesirable. Further, its low gain (2.15 dB), lack of directionality, and extremely narrow bandwidth make it even less attractive. Not surprisingly, the Yagi–Uda antenna (typically seen on the roof of most houses for television reception) was considered a breakthrough in antenna technology when introduced in the early 1920s because of its much higher gain of 8–14 dB. Log periodic wire antennas introduced in the late 1950s and 1960s and wire spirals allowed for both gain and bandwidth increases. On the other hand, high-gain antennas even today rely on large reflectors (dish antennas) and waveguide arrays [used for airborne/warning and control system (AWACS) radar] that are expensive and cumbersome to deploy.

Until the late 1970s, antenna design was based primarily on practical approaches using off-the-shelf antennas such as various wire geometries (dipoles, Yagi–Uda, log periodics, spirals), horns, reflectors and slots/apertures as well as arrays of some of these. The antenna engineer could choose or modify one of them based on design requirements that characterize antennas such as gain, input impedance, bandwidth, pattern beamwidth, and sidelobe levels (see, e.g., Refs. 1 and 2 or any of the several antenna textbooks for a description of these quantities). Antenna development required extensive testing and experimentation and was therefore funded primarily by the governments. However, more recently, dramatic growth in computing speeds and development of effective computational techniques [4–6] for realistic antenna geometries has allowed for low-cost virtual antenna design. Undoubtedly the explosive growth of wireless communications and microwave sensors, microwave imaging needs and radars has been the catalyst for introducing a multitude of new antenna designs since 1990 and an insatiable desire for using modern computational techniques for low cost designs. Requirements for conformal (nonprotruding) antennas for airborne systems, increased bandwidth requirements, and multifunctionality have led to heavy exploitation of printed (patch) or other slot-type antennas [7] and the use of powerful computational tools (commercial and noncommercial) for designing such antennas (see Fig. 1) [8]. The accuracy of these techniques is also remarkable, as seen by the results shown in Fig. 1 [9] for a cavity-backed slot spiral antenna. Needless to mention, the commercial mobile communications industry has been the catalyst for the recent explosive growth in antenna design needs. Certainly, the 1990s have seen an extensive use of antennas by the public for cellular, GPS, satellite, wireless LAN for computers, upcoming Bluetooth technology, and so on. However, future needs will be even greater when a multitude of antennas will be integrated into automobiles for all sorts of communication needs. Such antennas must be designed with the platform in mind (see Fig. 2) and must therefore satisfy gain and pattern requirements in the presence of the platform. Concurrent modeling of the large structure is therefore needed, resulting in a substantial increase of computational requirements for analysis and design purposes. For military applications, there is an increasing need for multifunctional antennas that can satisfy a plethora of communications needs using a single aperture as small as possible. Such apertures are also intended for unmanned airborne vehicles (UAVs) and small general aviation vehicles where real estate is even more limited. The multitude of design requirements for such antennas implies use of fast computational tools as well as optimization methods to arrive at designs that satisfy the specific mission or product needs.

In this article we summarize the most popular antenna analysis methods. These can be subdivided into time-domain and frequency-domain methods. *Time-domain methods* are appropriate for broadband analysis (many frequencies), and among them the finite-difference–time domain method [5,10] is the most popular technique. Time-domain–integral equation methods [11–13] are gaining attention by combining them with fast algorithms. However, in general, time-domain approaches are still slow and seldom attractive for narrowband analysis and design.

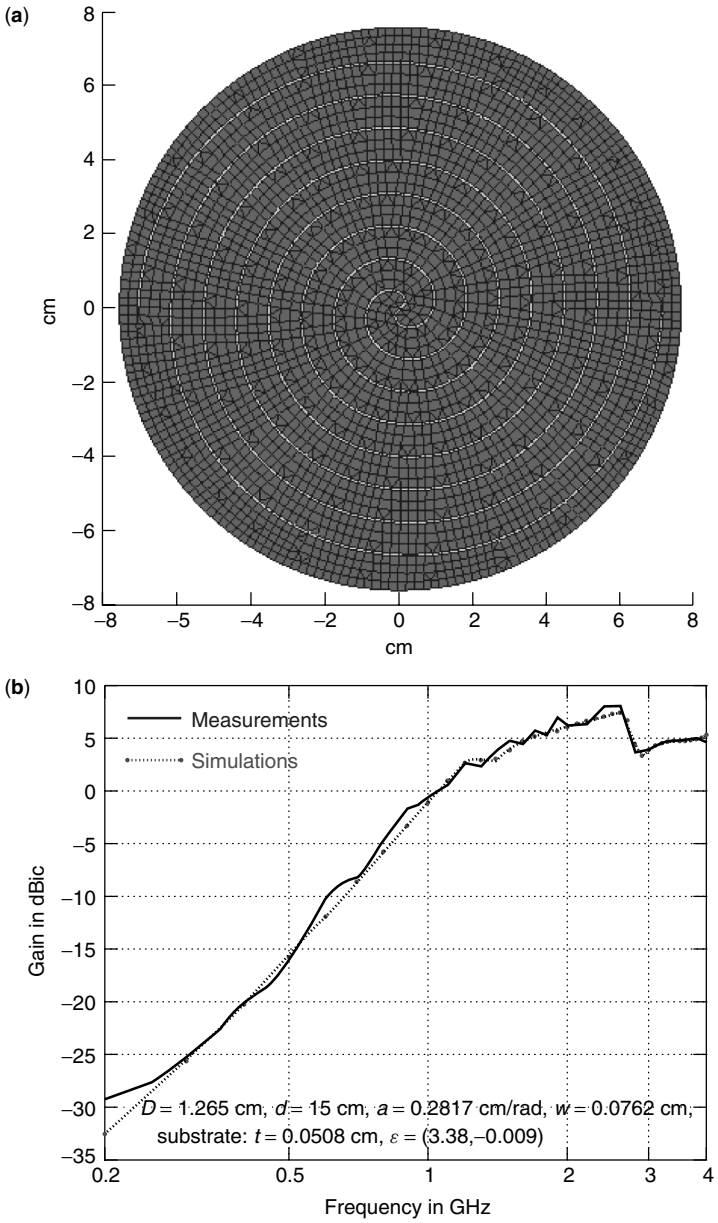


Figure 1. (**a**) Computational grid for a slot spiral antenna. (**b**) Comparison of gain calculations with measurements for slot spiral antennas [9]. The slot spiral is situated on the aperture of a circular cavity of diameter $d = 15$ cm and depth $D = 1.265$ cm. The slot spiral is a metal surface residing on a dielectric substrate ($\varepsilon_r = 3.38 - j0.009$ and of thickness $t = 0.0508$ cm) with the shown slot imprint of width $w = 0.0762$ cm.

In this article, we will discuss *frequency-domain methods* for antenna analysis. Like the time-domain methods, frequency domain approaches [14] can be categorized under (1) integral, (2) differential, and (3) hybrid techniques. The popular finite-element (FE) method [6,15] used in most branches of engineering belongs to the second category, and the ensuing procedure entails a direct solution of Maxwell's equations. Differential or FE methods are the choice modeling techniques for finite or inhomogeneous dielectric regions. In contrast, integral methods [4] are the choice techniques for modeling metallic structures situated in free space or on thin substrates (layers of dielectric). As can be understood, hybrid techniques involve a suitable combination of finite-element and integral or other modeling methods, including high-frequency techniques. The latter were actually the first to be used for accurate analysis of reflector and horn antennas [16] and for predicting antenna interactions on complex platforms such as aircraft [17], and work in this area continues to be explored [18]. More recently the combination of integral and FE methods [referred to as *finite-element–boundary integral* (FEBI) *methods*] has been successful for modeling complex antenna geometries constructed of metallic and nonmetallic materials [19]. The recent introduction of fast methods [20–22] has played an important role in the use of hybrid FEBI methods for design [23].

Below we proceed to discuss details associated with the implementation of integral and FE methods after we first present some basic electromagnetic concepts.

3. SOME BASIC EQUATIONS

Electromagnetic phenomena are governed by Maxwell's equations, a system of coupled time space partial differential equations established in the nineteenth century [3].

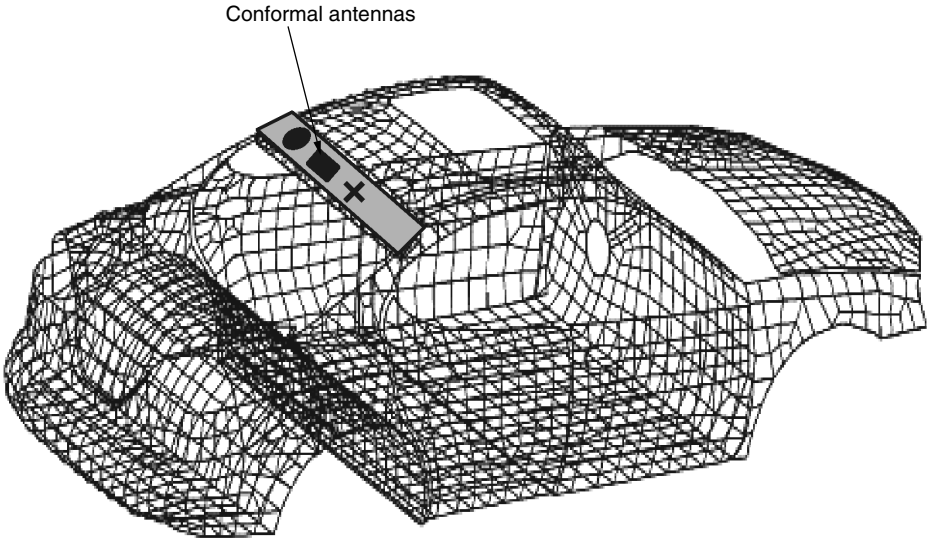


Figure 2. Quadrilateral grid used for modeling the automobile surface for onboard antenna evaluations. Antenna radiation can be introduced using the antenna aperture fields or currents computed for the isolated antenna.

Although Maxwell's equations allow for very general field variations in both space and time, for simplicity we shall consider only time-harmonic fields where the $e^{+j\omega t}$ time dependence is assumed and suppressed. In addition, we shall consider only fields in a linear and isotropic medium. With these assumptions, Maxwell's equations in differential (point) form can be written as

$$\nabla \times \mathbf{E} = -\boldsymbol{M}_i - jkZ\boldsymbol{H} \tag{1}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}_i + jkY\boldsymbol{E} \tag{2}$$

where $\boldsymbol{H}$ is the magnetic field in amperes per meter (A/m), $\boldsymbol{E}$ is the electric field in volts per meter (V/m), $\boldsymbol{J}_i$ is the impressed electric current (source antenna radiating current), and $\boldsymbol{M}_i$ is a fictitious magnetic current (source) often used for mathematical convenience. The radiating medium is completely described by its intrinsic impedance $Z = 1/Y = (\mu/\varepsilon)^{1/2}$, where ε and μ denote the medium's permittivity and permeability, respectively. The permittivity characterizes the medium's response in the presence of an electric field whereas the permeability is associated with the magnetic field. The wavenumber is denoted by $k = \dfrac{2\pi}{\lambda} = \omega(\mu\varepsilon)^{1/2}$, where λ is the wavelength and ω is the corresponding angular frequency. The faraday (1) and Ampère–Maxwell laws (2) are independent first-order vector equations. To solve for $\boldsymbol{E}$ and $\boldsymbol{H}$, we typically combine (1) and (2) to obtain the vector wave equation

$$\nabla \times \nabla \times \begin{Bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{Bmatrix} - k^2 \begin{Bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{Bmatrix} = -j\omega \begin{Bmatrix} \mu \boldsymbol{J}_i \\ \varepsilon \boldsymbol{M}_i \end{Bmatrix} \mp \nabla \times \begin{Bmatrix} \boldsymbol{M}_i \\ \boldsymbol{J}_i \end{Bmatrix} \tag{3}$$

The wave nature of $\boldsymbol{E}$ and $\boldsymbol{H}$ is easily surmised when we introduce the identity $\nabla \times \nabla \times \boldsymbol{E} = -\nabla^2 \boldsymbol{E} + \nabla(\nabla \cdot \boldsymbol{E}) = -\nabla^2 \boldsymbol{E}$, where we have assumed that $\nabla \cdot \boldsymbol{E} = 0$, true away from the source region (away from the antenna). With this replacement, the vector wave equation reduces to scalar wave or rather Helmholtz equations of the form $\nabla^2 E_\xi + k^2 E_\xi = f_i$, where E_ξ implies the ξth component of the vector field and f_i is the appropriate right-hand side reduced from (3). A solution of the wave equation can be accomplished provided the boundary conditions are enforced on canonical surfaces (spheres, cubes, infinite cylinders, and planes). For practical problems, however, boundary conditions must be enforced on noncanonical surfaces, and consequently a closed-form solution is not possible. This is the reason for resorting to a numerical solution of Maxwell's equations.

When dealing with arbitrary antenna structures, it is customary to invoke the surface equivalence principle [24], as illustrated in Fig. 3. The antenna itself is enclosed in a mathematical surface S, and equivalent or mathematical surface currents $\boldsymbol{J}_s$ and $\boldsymbol{M}_s$ are introduced on that surface. When $(\boldsymbol{J}_i, \boldsymbol{M}_i)$ in (1) and (2) are replaced by $(\boldsymbol{J}_s, \boldsymbol{M}_s)$, it follows that an integral representation of the fields everywhere is

$$\boldsymbol{E}(\boldsymbol{r}) = -jkZ \iint_S \left[\boldsymbol{J}_s(\boldsymbol{r}') G(\boldsymbol{r}, \boldsymbol{r}') + \frac{1}{k^2} (\nabla_s' \cdot \boldsymbol{J}_s(\boldsymbol{r}')) \nabla G(\boldsymbol{r}, \boldsymbol{r}') \right] ds' + \iint_S \left[\boldsymbol{M}_s(\boldsymbol{r}') \times \nabla G(\boldsymbol{r}, \boldsymbol{r}') \right] ds' + \boldsymbol{E}^{\text{inc}} = \boldsymbol{E}^{\text{rad}} + \boldsymbol{E}^{\text{inc}} \tag{4}$$

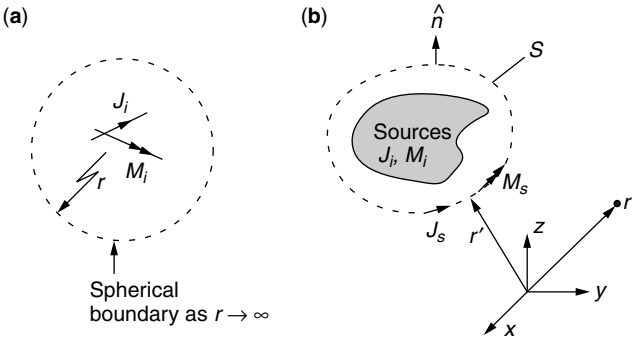


Figure 3. (**a**) Source currents enclosed by a spherical boundary. (**b**) Illustration of surface equivalence where the source current or field is replaced by equivalent surface currents located on a surface enclosing the antenna/source; $(\boldsymbol{J}_s, \boldsymbol{M}_s)$ can be found by enforcing the boundary conditions on the antenna structure.

where $\boldsymbol{r}$ and $\boldsymbol{r}'$ refer to the observation and source (i.e., $\boldsymbol{J}_s$ and $\boldsymbol{M}_s$) position vectors, respectively. Here, $\nabla_s \cdot$ denotes the surface divergence operator [25] and the equivalent surface current densities ($\boldsymbol{J}_s, \boldsymbol{M}_s$) are related to the surface fields via the relations

$$\begin{aligned} \boldsymbol{J}_s &= \hat{n} \times \boldsymbol{H} \\ \boldsymbol{M}_s &= \boldsymbol{E} \times \hat{n} \end{aligned} \tag{5}$$

where $\hat{n}$ denotes the unit normal to the integration surface S. With this identification, (4) becomes the Stratton–Chu equation [26]. Also, $\boldsymbol{E}^{\text{inc}}$ is the incident or excitation field intensity and is nonzero for radar scattering problems, but typically set to zero for antenna analysis. Further, G is the scalar Green function[1] given by

$$G(\boldsymbol{r}, \boldsymbol{r}') = \frac{1}{4\pi} \frac{e^{-jk|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|} \tag{6}$$

This Green function also incorporates the radiation condition stating that outgoing waves are of the form e^{-jkr}/r as $r \to \infty$. In mathematical form, the corresponding boundary condition is

$$\lim_{r\to\infty} r[\nabla \times \boldsymbol{E} + jk\hat{r} \times \boldsymbol{E}] = 0$$

a necessary condition for the unique solution of (3).

The analysis of antennas using integral equation methods amounts to finding ($\boldsymbol{J}_s, \boldsymbol{M}_s$) or some other predefined currents subject to boundary conditions satisfied by the antenna structure. To illustrate the implementation of integral equation methods, we refer to Fig. 4, showing a reflector and a patch antenna [24]. For the reflector antenna, on application of the equivalence principle (here the closed surface S is collapsed onto the reflector surface and $\boldsymbol{J}_s = \boldsymbol{J}_s^+ - \boldsymbol{J}_s^-$ is the net current), the reflector is removed and replaced by the equivalent current $\boldsymbol{J}_s$. There is no need for magnetic currents since $\hat{n} \times \boldsymbol{E} = 0$ on metallic surfaces in accordance with (5). The unknown $\boldsymbol{J}_s$ can subsequently be found by solving the integral equation

$$\hat{n} \times [\boldsymbol{E}^{\text{rad}}(\boldsymbol{r}) + \boldsymbol{E}^{\text{inc}}(\boldsymbol{r})] = 0, \qquad \boldsymbol{r} \in S \tag{7}$$

valid for $\boldsymbol{r}$ on the reflector's surface, where $\boldsymbol{E}^{\text{inc}}$ is the excitation field from the reflector feed. In the case of the patch antenna structure, the $\boldsymbol{M}_s$ current is also introduced across the antenna aperture where $\boldsymbol{E}$ is nonzero. Referring to Fig. 4, the corresponding integral equations for $\boldsymbol{J}_s$ and $\boldsymbol{M}_s$ are

$$\begin{aligned} &\hat{n} \times \boldsymbol{E}_1(\boldsymbol{r}) = 0, && \boldsymbol{r} \in \text{ exterior metallic surfaces} \\ &\hat{n} \times \boldsymbol{E}_1(\boldsymbol{r}) = \hat{n} \times \boldsymbol{E}_2(\boldsymbol{r}) && \boldsymbol{r} \in \text{ cavity aperture} \\ &\hat{n} \times \boldsymbol{H}_1(\boldsymbol{r}) = \hat{n} \times \boldsymbol{H}_2(\boldsymbol{r}) && \boldsymbol{r} \in \text{ cavity aperture} \\ &\hat{n} \times \boldsymbol{E}_2(\boldsymbol{r}) + \boldsymbol{E}^{\text{probe}}(\boldsymbol{r}) = 0, && \boldsymbol{r} \in \text{ interior metallic surfaces} \end{aligned} \tag{8}$$

[1] The Green function of a given solution domain characterizes the fields caused by a unit point source.

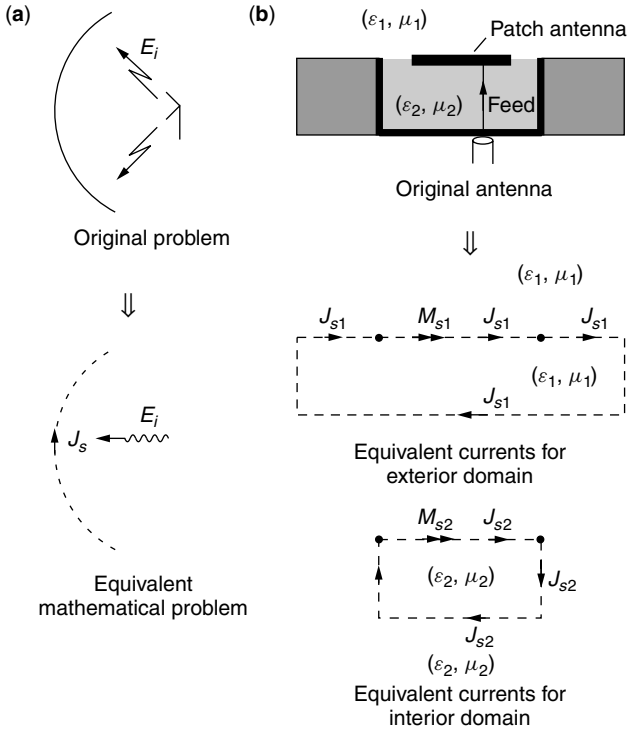


Figure 4. Surface equivalence principle applied to a reflector (left) and a patch (right) antenna.

where $\boldsymbol{E}_1$ is due to $\boldsymbol{J}_{s1}$ and $\boldsymbol{M}_{s1}$ radiating in a homogeneous medium (ε_1, μ_1) and $\boldsymbol{E}_2$ is the corresponding field due to $\boldsymbol{J}_{s2}$ and $\boldsymbol{M}_{s2}$ radiating in the homogeneous medium (ε_2, μ_2); that is, $(\boldsymbol{J}_{s1}, \boldsymbol{M}_{s1})$ and $(\boldsymbol{J}_{s2}, \boldsymbol{M}_{s2})$ are conveniently introduced to satisfy the boundary conditions on the boundary of the piecewise homogeneous region. We can readily conclude that the continuity conditions across the aperture (excluding the patch) can be satisfied a priori by setting $\boldsymbol{J}_{s1} = -\boldsymbol{J}_{s2}$ and $\boldsymbol{M}_{s1} = -\boldsymbol{M}_{s2}$ across that section of the surface. The equivalent currents for the other surfaces must be found via a numerical solution of the integral equation enforcing the boundary conditions.

4. INTEGRAL EQUATION TECHNIQUES

Integral equation (IE) techniques are among the oldest and most successful computational antenna modeling approaches. Their basic idea is to replace an antenna or scattering object by equivalent sources (currents) such that those sources radiate in a domain whose Green function is known. An IE is then derived by enforcing the boundary or continuity conditions for the fields such as those in (8) or (7). This is the first step (*step* 1) in any simulation test and typically involves use of the equivalence principle illustrated in Fig. 3 and 4. The equivalence principle relies on the uniqueness theorem, stating that the resulting solution is unique provided it satisfies Maxwell's equations and the boundary conditions. Use of the representation (4) guarantees the first, whereas the boundary conditions are enforced with the appropriate choice of the equivalent current. It is important to note the introduction of the equivalent currents implies removal of the

actual structure whose presence is only implied through the boundary conditions to be enforced.

The second step (*step* 2) in a numerical simulation is the discretization of the geometry (as done in Fig. 5 and 2) and the unknown equivalent sources using an appropriate set of basis or expansion functions. The coefficients of the expansion then become the unknowns or degrees of freedom (DoFs) used for setting up the linear system of equations. Such a linear system is formed by enforcing the boundary conditions associated with the original geometry using point matching, the method of moments (MoM) [27] procedure or Nyström's method [28]. For example, at metallic surfaces the boundary condition to be enforced is that the tangential electric fields vanish on that surface (viz., $\hat{n} \times \boldsymbol{E} = 0$ on perfect conductors). On a dielectric surface, the pertinent boundary conditions must enforce continuity of tangential $\boldsymbol{E}$ and $\boldsymbol{H}$ across the interface. To do so, it is necessary that both electric and magnetic equivalent currents be introduced at dielectric interfaces as done, for example, in Fig. 4(b) for the cavity aperture. On the other hand, for metallic surfaces only one equivalent current is required to satisfy the boundary conditions.

The third and final step (*Step* 3) in a numerical simulation is the solution of the linear system generated in step 2. For small linear systems involving less than

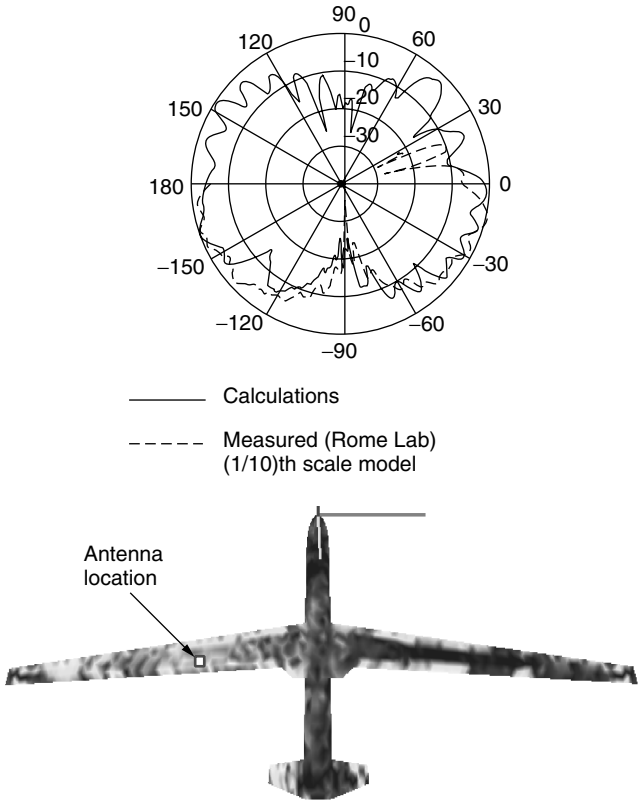


Figure 5. Comparison of measured and calculated patterns for a UHF antenna radiating on a Global Hawk unmanned aerial vehicle (UAV). The surface field plot shows the currents or surface magnetic field on the UAV (red implies highest strength). Measurements are courtesy of the Air Force Research Laboratory (Rome, New York, USA).

1000–5000 unknowns, direct inversion (Gauss elimination) and Lower-Upper (LU) decomposition [29] methods are typically the choice approaches. However, these methods require $O(N^3)$ central processing unit (CPU) solution time and $O(N^2)$ for storing the matrix (N: number of unknowns), and typically over a million discrete elements (and unknowns) are needed to discretize a simple full-scale aircraft at radar frequencies. Furthermore, design for antennas and microwave circuits can be realized only by using extremely fast algorithms that provide results in seconds or minutes rather than hours or days. In the mid-1990s, fast algorithms such as the fast multipole method (FMM) and its multilevel version [20] and adaptive integral method (AIM) [21,22] were introduced to alleviate the CPU bottleneck associated with realistic EM simulations. Basically, FMM and AIM overcome the $O(N^3)$ "curse of dimensionality," as it is often called, in solving dense matrix systems using direct solution methods. Their CPU and memory requirements reduce down to $O(N \log N)$ or so. The basic idea of AIM and FMM is the same as that used in the highly efficient fast Fourier transform (FFT) method. AIM does even make direct utilization of the FFT in its implementation, and the multilevel FMM can be considered as a fast FFT for data with unequal separation intervals.

Let us now proceed to implement the three steps mentioned above. For a metallic antenna structure as in Fig. 4, only electric equivalent currents are needed to express the radiated fields, and in this case the condition to be enforced on the metallic surface of the reflector or some other structure is $\hat{n} \times [\boldsymbol{E}^{\text{rad}} + \boldsymbol{E}^{\text{inc}}] = 0$, where $\boldsymbol{E}^{\text{inc}}$ denotes the field from the horn feed (see Fig. 5) and $\boldsymbol{E}^{\text{rad}}$ is the integral expression in (4). The enforcement of this condition gives the integral equation

$$jkZ\hat{n} \times \iint_S \left[\boldsymbol{J}_s(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}') + \frac{1}{k^2}(\nabla_s' \cdot \boldsymbol{J}_s(\boldsymbol{r}'))\nabla G(\boldsymbol{r},\boldsymbol{r}') \right] ds' \Bigg|_{\boldsymbol{r}\in S} = \hat{n} \times \boldsymbol{E}^{\text{inc}}(\boldsymbol{r})|_{\boldsymbol{r}\in S} \tag{9}$$

referred to as the *electric field integral equation* (EFIE) since it enforces the boundary condition on $\boldsymbol{E}$. The unknown here is the current density $\boldsymbol{J}_s$ and to solve for it, we proceed to step 2 of the analysis.

Step 2 of the analysis amounts to discretizing the current density and the associated geometry as displayed in Figs. 1 and 2 using triangles, quads, or other surface patches. Specifically, we introduce the expansion

$$\boldsymbol{J}_s(\boldsymbol{r}) = \sum_{n=1}^{N} I_n \boldsymbol{f}_n(\boldsymbol{r}) \tag{10}$$

where $\boldsymbol{f}_n$ are the expansion or basis functions defined on S. This current expansion is substituted into (9) and application of the MoM means testing the resulting equation with N different testing or weighting functions $\boldsymbol{w}_m$. Since testing is equivalent to multiplication with the individual testing functions and subsequent integration over S, N linear algebraic equations for the unique determination of

the N expansion coefficients I_n are obtained:[2]

$$jkZ\sum_{n=1}^{N} I_n \iint_S \iint_S \boldsymbol{w}_m(\boldsymbol{r})\cdot\left[\boldsymbol{f}_n(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}') + \frac{1}{k^2}(\nabla_s'\cdot\boldsymbol{f}_n(\boldsymbol{r}'))\nabla G(\boldsymbol{r},\boldsymbol{r}')\right] ds'\,ds = \iint_S \boldsymbol{w}_m(\boldsymbol{r})\cdot\boldsymbol{E}^{\text{inc}}(\boldsymbol{r})\,ds, \qquad m=1,\ldots,N. \quad (11)$$

For each mth weighting function we obtain a single equation for a total of N equations, leading to the matrix system

$$[Z_{mn}]\{I_n\} = \{V_m\} \quad (12)$$

where $[Z_{mn}]$ is an $N\times N$ fully populated matrix containing the coupling integrals on the left-hand side of (11), $\{I_n\}$ is a column vector containing the unknown coefficients I_n, and $\{V_n\}$ is a column vector containing the weights (moments) of the incident field $\boldsymbol{E}^{\text{inc}}$ on the right-hand side of (11). Depending on the choices for expansion and testing functions, an enormous number of different IE techniques have been developed. If the testing functions are the delta functions, the resulting IE method is called a *point-matching* or collocation method since the IE is enforced only at distinct observation points. The so-called *Galerkin method* performs testing using a weighting function that is identical to the expansion function in (10). If the expansion functions have a domain that spans the entire structure, they are referred to as *entire-domain basis functions*. Entire domain basis functions have only been used for specific cases where the structure shape can be of advantage. *Subdomain basis functions* are typically used and have been the most successful. Subdomain bases have their domain restricted to a single or a pair of surface discretization elements (triangles or quads as shown in Figs. 1 and 2). Typically, they approximate the surface current density as a constant, linear, or quadratic function over the element.

A very crucial point related to all integral equation techniques in electromagnetics is the evaluation of the coupling integrals Z_{mn}. An analytic evaluation of these integrals is usually not possible and numerical integration is plagued by the singular behavior of the Green function $G(\boldsymbol{r},\boldsymbol{r}')$ near $\boldsymbol{r}=\boldsymbol{r}'$. To overcome this difficulty, a series of specialized integrations has been developed over the years that combine analytic integration techniques for the extracted singularity [30,31] with numerical quadrature rules for the remainder integrands. Another possibility is the application of integration variable transformations such as the Duffy's transform [32] to allow for robust numerical integration.

MoM techniques have been very successful in modeling antenna structures and antennas on platforms (see Fig. 5). To better understand the moment method procedure, next we consider the solution of a rather simple integral equation, that associated with radiation by a wire dipole antenna.

[2] $\hat{n}\times$ can be omitted after testing.

4.1. Wire Modeling

A wire antenna is simply a cylindrical conductor whose diameter (a) is much smaller than the length (L) of the wire and also $a \ll \lambda$, where λ is the wavelength of operation. Since the wire is very thin (see Fig. 6), we can practically model the radiation from such a cylindrical antenna using a filamentary (equivalent) current flowing through the center of the wire instead of considering a surface current $\boldsymbol{J}_s$. The use of this filamentary current also implies that no currents exist that are transverse to the wire axis (z), a reasonable approximation for very thin wires. Thus, the integral equation can be rewritten using line rather than surface integrals. We have

$$jkZ\sum_{n=1}^{N} I_n \int_{-L/2}^{L/2}\int_{-L/2}^{L/2} w_m(z)f_n(z')\left[1+\frac{1}{k^2}\frac{\partial^2}{\partial z^2}\right]G_w(z-z')\,dz'\,dz = \int_{-L/2}^{L/2} w_m(z)E_z^{\text{inc}}(z)\,dz, \qquad m=1,\ldots,N \quad (13)$$

and by enforcing the boundary condition on the surface of the wire, $\rho = a$ (see Fig. 6b), the Green function takes the form

$$G_w(z-z') = \frac{1}{4\pi}\frac{e^{-jk[a^2+(z-z')^2]^{1/2}}}{[a^2+(z-z')^2]^{1/2}} \quad (14)$$

We remark that since the current $I(z)$ is not at the same location where the boundary condition is enforced, the Green function is nonsingular at $z=z'$. Another important advantage of thin-wire models is the very small number of expansion functions needed for modeling real-world configurations achieved by the one-dimensional (line current) representation of the originally two-dimensional equivalent surface current densities [33,34] (when away from the feed, usually a single wire of radius $a=w/4$, where w is the strip width, can be used to model a surface as a wire grid).

The wire antenna excitation is performed by an appropriate specification of E_z^{inc} in (13). For a voltage source excitation at the wire center, a good approximation is $E_z^{\text{inc}} = V_0\delta(z)$ as illustrated in Fig. 6d, where V_0 is the voltage source applied at $z=0$. An alternative magnetic frill–current excitation has been considered [35] and is appropriate for the wire monopole antenna seen on most

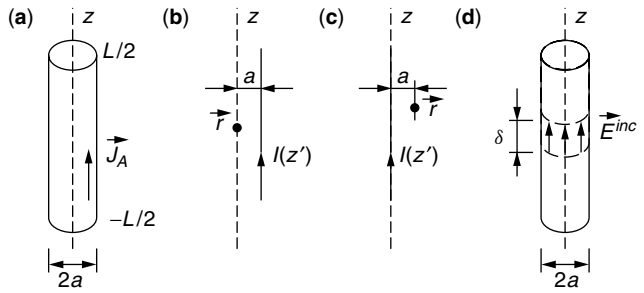


Figure 6. (**a**) Straight wire along z axis; (**b**, **c**) interpretations of thin-wire model; (**d**) δ-gap excitation of wire antenna.

automobiles. The input impedance of the antenna can then be calculated from $Z_{\text{in}} = V_0/I(z_{\text{feed}})$ once the equation system is solved, where the feed location z_{feed} is adjusted for impedance control.

To solve for $\{I_n\}$ using (13), we must choose the basis and weighting functions. The simplest functions useful for the line current approximation are the piecewise constant or pulse basis functions displayed in Fig. 7, but the triangular bases displayed in Fig. 8 provide a much improved approximation without discontinuities at the junctions of the wire segments. Testing is often done using point matching or Galerkin's methods. Figure 9 shows the complex input impedance Z_{in} of a center-fed wire antenna computed using pulse bases with point matching for testing. About 100 basis functions were equally distributed along the wire, and the excitation frequency was varied to investigate the L/λ dependence of the wire antenna. Wire resonances were found at $L/\lambda = 0.48$ and $L/\lambda = 0.84$. For this reason, the optimal operation of the dipole antenna occurs when the dipole is 0.48λ long (see Fig. 10).

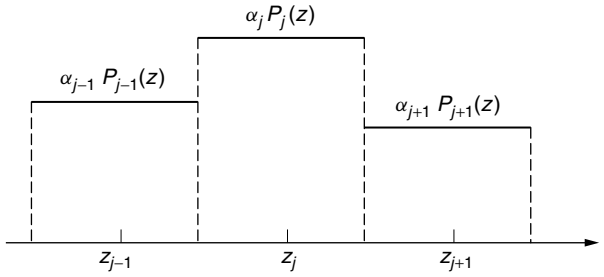


Figure 7. Pulse basis functions for line current modeling.

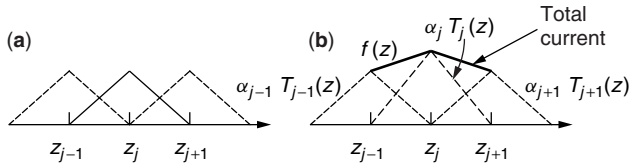


Figure 8. Triangular basis functions (**a**) and (**b**) the resulting piecewise linear current approximation.

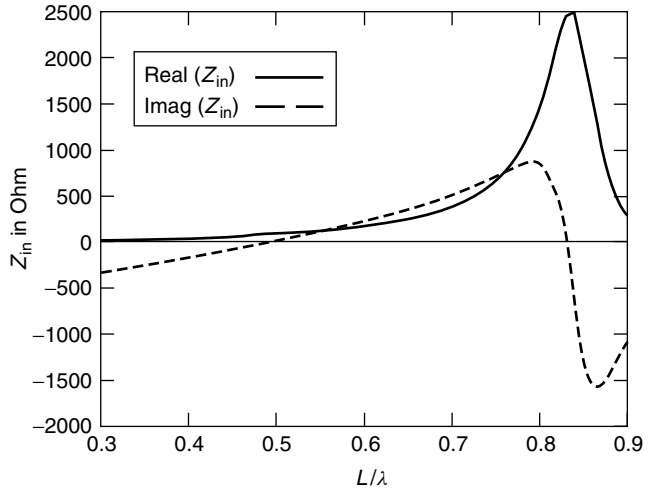


Figure 9. Complex input impedance as a function of length for the wire antenna in Fig. 6 ($a = 0.001\lambda$).

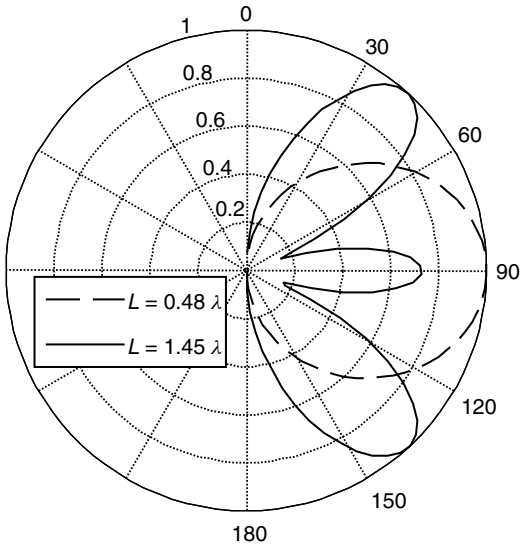


Figure 10. Normalized E-plane radiation patterns for two different wire lengths as shown in Fig. 6 ($a = 0.001\lambda$).

The wire integral equation (13) can be readily generalized to curved wires. In any case, once the line currents are computed along the wire contour, the far field [35] is found from

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{jkZ}{4\pi r}\int_C \hat{r} \times [\hat{r} \times \hat{l}]I(l')e^{-jk|\boldsymbol{r}-\boldsymbol{r}'|}\,dl' \qquad (15)$$

where $\hat{r}$ is the unit vector in the radial direction and $\hat{l}$ is the tangential unit vector along the wire contour.

Thin-wire IE models are useful not only for wire antennas and scatterers but also for computing the radiation of metallic antenna structures (reflectors, horns, etc.) when sufficiently dense wire-grid models are employed to represent the equivalent surface current densities on the structure's surface [33]. For this purpose, wire junctions must be included into the thin-wire theory. This is not an issue for pulse basis approximations, since no current continuity is enforced between wire segments. However, for triangular or higher-order basis functions, special junction conditions must be introduced to fulfill Kirchhoff's current continuity law at wire junctions.

4.2. Surface Modeling

For a variety of antenna problems, wire-grid models are accurate enough to produce useful simulation with high computational efficiency. For other applications such as accurate shielding or planar antenna analyses, the thin-wire approximations cannot produce acceptable results and a direct evaluation of the surface currents as given in (11) is necessary. The formal MoM IE solution is already given by (11), and a numerical implementation can be realized with appropriately assigned surface basis functions for the current representation. Again, popular surface current IE models are based on subdomain bases, where triangular and quadrilateral surface subdomains

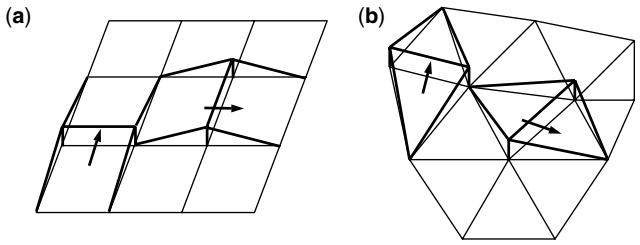


Figure 11. Illustration of mixed linear/constant (rooftop) surface current expansions on (**a**) rectangular and (**b**) triangular subdomains.

as displayed in Fig. 11 are often used. Such techniques are also known as *boundary element* or *boundary integral methods*. The surface current distribution on the individual subdomains is typically chosen to fulfill the necessary physical constraints (e.g., zero normal current components at the edge of a plate) for surface currents. Edge-based surface current basis functions with mixed interpolations are illustrated in Fig. 11 [36]. The degrees of freedom are assigned to the edges of the surface mesh and each basis function has constant normal surface current density components over exactly one edge. Perpendicular to the edge, the current density variation is linear (mixed linear/constant). The basis functions shown in Fig. 11(a) are the *rooftop* basis functions on rectangular subdomains, and those in Fig. 11(b) are the well-known *Rao–Wilton–Glisson* functions introduced in 1982 [36]. On applying coordinate transformations, these basis functions can be transformed to fit to curved surface patches, allowing for more accurate surface current representations [37–39]. Even higher accuracy can be achieved by higher-order basis functions that typically include high-order polynomials subject to constraints at the element junctions for current continuity [34].

The linear system in (11) was derived under the assumption of a metallic surface S. However, surface IE techniques can also be applied for the analysis of antennas involving dielectric and even lossy material objects. As stated earlier, for this case, two current densities must be introduced on S, such as the electric $\boldsymbol{J}_s$ and magnetic $\boldsymbol{M}_s$ surface currents, so that field continuity can be enforced as illustrated in Fig. 4 [24,36]. When dealing with antennas that include dielectrics, volume IEs can be combined with surface IEs for the analysis of the antenna volumetric sections [40,41].

5. FINITE-ELEMENT METHODS

As noted above, IE or boundary element methods result in fully populated matrix systems. For volumetric integral equations where materials must be handled, their storage and computational requirements become prohibitive as the size of the structure increases. This is because the Green function couples all boundary elements regardless of their separation distance. To avoid the introduction of a Green function, one can instead pursue a direct solution of Maxwell's equations in their differential form. Specifically, one could replace the continuous derivatives by finite differences (FDs) to construct a set of equations that can then be solved iteratively in conjunction with specified or natural boundary conditions. This simple, yet very powerful approach became especially popular for time-domain electromagnetic field analysis [e.g., 5,10].

The standard FE method can be derived by applying Galerkin's testing to the differential form of Maxwell's equation, that is, by weighting Maxwell's equations with a suitable (testing) function whose domain is usually restricted over a small region or element (test element) of the computational volume [6,38]. Similar elements can also be used for discretizing the volume region where simple geometric shapes such as quads, tetrahedrons, or triangular prisms (see Fig. 12) are often used. The tetrahedron provides the most flexible element for discretizing volumetric regions. Distorted hexahedra have also been successful for volumetric modeling and may lead to fewer unknowns. On the other hand, shapes with higher symmetry are associated with simple discretization algorithms and may result in better conditioned linear systems.

Although, as discussed above, the finite-element method is based on a direct discretization and solution of the wave equation (3), the weighted or weak form of (3) is used for discretization and solution [6]. The latter enforces Maxwell's equations on an average sense over the discrete element (tetrahedron, quadrilateral, etc.) and is obtained by first dotting (3) with the weighting function/basis $\boldsymbol{W}$ and then making use of the divergence theorem [3] to obtain

$$\iiint_V \left[\frac{1}{\mu_r}(\nabla \times \boldsymbol{E}) \cdot (\nabla \times \boldsymbol{W}) - k^2 \varepsilon_r \boldsymbol{E} \cdot \boldsymbol{W}\right] dv + jkZ \oiint_{S_V} \boldsymbol{W} \cdot (\boldsymbol{H} \times \hat{n})\, ds = 0 \tag{16}$$

Here V denotes the volume domain of interest, S_V is the surface enclosing V, and $\hat{n}$ is the outward normal to S_V. As in (11), $\boldsymbol{W}$ is some weighting function spanning the domain of, say, the eth discrete element. By choosing an expansion for the electric field such as

$$\boldsymbol{E} = \sum_{e=1}^{N} \sum_{i=1}^{N_e} E_i^e N_i^e(\boldsymbol{r}) \tag{17}$$

we can then follow the same steps done for integral equation methods to construct the linear system to find the unknown E_i^e. Although (17) is in principle the same as (10), its form is different but rather convenient in addition to

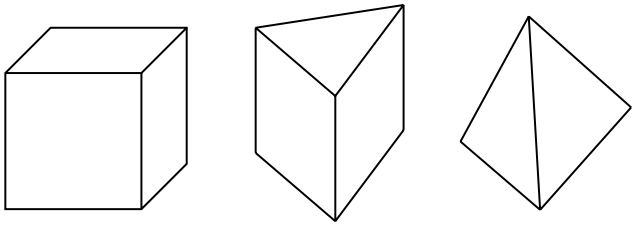


Figure 12. Popular finite-element subdomains: quad, triangular prism, tetrahedron.

having a physical meaning (see Ref. 6 for details). To explain it, let's assume that tetrahedra are used to model the volume of interest. Since the tetrahedron has six edges, we will then set $N_e = 6$, and if N represents the number of tetrahedra used to discretize the entire volume, then the sums in (17) run through all the edges of the tetrahedra constituting the volume. However, if an edge is common to, say, three tetrahedra, then it will appear as many times in the sum. Of particular interest for the expansion (17) is the choice of the basis functions $N_i^e(\boldsymbol{r})$, whose domain based on our notation is only within the eth element and is associated with the ith edge of that element. If we choose $N_i^e(\boldsymbol{r})$ so that

$$N_i^e(\boldsymbol{r}) = \begin{cases} 1 & \boldsymbol{r} = \boldsymbol{r}_i \\ 0 & \boldsymbol{r} = \boldsymbol{r}_j, \quad j \neq i \end{cases}$$

that is, if it is unity at the ith edge located along $\boldsymbol{r}_i$ (similar to Fig. 11) and goes to zero at all other edges, then the coefficient E_i^e is simply the field along the ith edge of the eth element.

The linear system for solving E_i^e is constructed via Galerkin's method by setting $\boldsymbol{W} = \boldsymbol{N}_i^e(\boldsymbol{r})$; then, for $i = 1, \ldots, N_e$ and $e = 1, \ldots, N$, we will obtain NN_e equations. Clearly, these are more than required ($N_{\text{total edges}} < NN_e$) because most of the edges are shared by multiple tetrahedra. For example, an edge shared by three tetrahedra will generate three equations, but actually only one is needed for the unknown field at that edge. Thus, the NN_e equations must be reduced down to $N_{\text{total edges}}$, and this is done via the so-called assembly process. The latter is nothing more than taking the average of the equations generated by testing with $N_j^e(\boldsymbol{r})$, which is unity at the same edge.

As we observe, (16) also includes the unknown magnetic field $\boldsymbol{H}$ at the boundary surface S_V of the domain. Because it is on the boundary, the unique solution of the wave equation requires that it be specified with an external condition unrelated to Maxwell's equations. Of course, if the surface S_V is far away from the radiating source, then the radiation condition [26] can be employed to relate $\boldsymbol{E}$ and $\boldsymbol{H}$ on S_V and thus obtain a deterministic system for the solution of E_i^e. Since the radiation condition can be used only for S_V far away from the source, the enclosed volume is enlarged significantly, requiring many elements for its discretization. Therefore, the number of unknowns becomes unmanageable when the classic radiation condition must be used [25,26], and this plagued the practical application of finite element methods to electromagnetics until the late 1980s (nearly two decades since the method was used by Silvester [42] for waveguide propagation).

The introduction of absorbing boundary conditions (ABCs) such as [6,43]

$$-jkZ\hat{n} \times \boldsymbol{H} = jk\boldsymbol{E}_t + \frac{1}{2jk}\{[\nabla \times [\hat{n}(\hat{n} \cdot \nabla \times \boldsymbol{E})] + \nabla_t(\nabla \cdot \boldsymbol{E}_t)]\} \tag{18}$$

where the subscript "t" denotes the tangential components of $\boldsymbol{E}$ or the nabla operator when evaluated on S_V, provided the means for practical implementation of the finite-element method. This ABC can be enforced with S_V placed on only a fraction of a wavelength from the source. As an alternative to the ABC, one could also use an integral equation to relate $\boldsymbol{E}$ and $\boldsymbol{H}$ on S_V. Such integral equations are of the same form as (4) with $(\boldsymbol{J}_s, \boldsymbol{M}_s)$ replaced by $(\hat{n} \times \boldsymbol{H}, \boldsymbol{E} \times \hat{n})$. Clearly, such an integral equation will lead to a dense submatrix for relating the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields on the surface S_V. Since the finite-element method used for discretizing the interior volume fields is a sparse matrix (usually having a bandwidth of 40 elements or so), the resulting overall system will be partly sparse and partly dense [44]. This is referred to as a *hybrid system*, and the associated methodology is the successful hybrid FE-BI method [44]. It is attractive because S_V can be placed as close to the radiator as desired without restrictions, thus leading to the least number of unknowns. The drawback of this advantage is the increased complexity due to the partly dense system, but more recent use of fast integral methods has alleviated the related issues of CPU and memory complexity [20,45].

The hybrid FE-BI method has been extensively used for the analysis of cavity-backed antennas (see Fig. 4) [44]. Here, the FE volume V includes the possibly inhomogeneous dielectric below the patch only and a BI is applied only at the aperture. Utilizing a half-space Green function, it is sufficient to place magnetic surface current densities $\boldsymbol{M}_s$ on the dielectric portion of the aperture only, and thus a very efficient implementation of the method is possible [6,44]. Fig. 13 shows the input impedance of a cavity-backed and coaxially fed rectangular microstrip patch antenna that has been computed with the hybrid FE-BI technique.

6. CONCLUDING REMARKS

During the 1990s several integral, differential, and hybrid methods matured, and a number of commercial level antenna simulation packages were developed for small

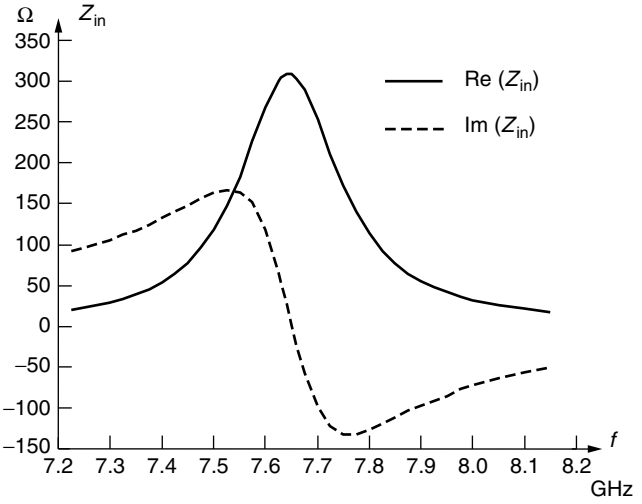


Figure 13. Input impedance of rectangular coaxially fed cavity-backed microstrip patch antenna computed using the hybrid finite-element–boundary-integral technique. Patch: length $L = 1.2$ cm, width $W = 0.8$ cm, feed distance to edge $d = 0.2$ cm, dielectric: thickness $t = 0.1$ cm, $\varepsilon_r = 2.2$.

size but practical antenna models. One can say that the introduction of fast methods during the second half of the 1990s has truly provided the community with computational tools that can allow for both geometric adaptability and material generality, as well as practical size simulations. As an example, using the fast multipole method, a problem involving as many as 170,000 unknowns and resulting in a dense matrix can now be solved in 3 h on a desktop PC using 700 MB (megabytes) of memory, whereas in the mid-1990s the same problem was unsolvable. However, further research on fast methods and related iterative solvers is required prior to their commercialization. Topics such as solver convergence, material modeling, robust and higher-order basis functions, and methods and solver hybridizations are all issues of current research for a variety of specific applications.

BIOGRAPHIES

John L. Volakis obtained his B.E. degree, summa cum laude, in 1978 from Youngstown State University, Ohio, his M.Sc. in 1979 from the Ohio State University, Columbus, Ohio, and a Ph.D. degree in 1982, also from the Ohio State University. Before joining the University of Michigan, Ann Arbor, faculty he worked at the Ohio State University. ElectroScience Laboratory and at Rockwell International. He is currently a professor in the Department of Electrical Engineering and Computer Science (EECS). His primary research deals with the development and application of computational and design techniques to scattering, antennas, and bioelectromagnetics. Dr. Volakis has published over 180 articles in major refereed journal articles, more than 220 conference papers, and 9 book chapters. In addition, he coauthored two books: *Approximate Boundary Conditions in Electromagnetics* (Institution of Electrical Engineers, 1995) and *Finite Element Method for Electromagnetics* (IEEE Press, 1998). In 1998, he received the University of Michigan College of Engineering Research Excellence award, and in 2001 he received the Department of Electrical Engineering and Computer Science Service Excellence Award. Dr. Volakis is a fellow of the IEEE and has served on the editorial boards of several journals. He is also listed in several *Who's Who* directories.

Thomas F. Eibert received the Dipl.-Ing.(FH) degree in 1989 from Fachhochschule Nuernberg, Germany, the Dipl.-Ing. degree in 1992 from the University of Bochum, Germany, and the Dr.-Ing. Degree in 1997 from the University of Wuppertal, Germany, all in electrical engineering. From 1997 to 1998 he was with the Radiation Laboratory at the Electrical Engineering and Computer Science (EECS) Department of the University of Michigan, Ann Arbor, and in 1998 he joined Deutsche Telekom, Darmstadt, Germany, as a research engineer where he has been working on wave propagation and coverage predictions for terrestrial mobile communications and radio broadcasting networks. His major areas of interest are numerical and analytical techniques for electromagnetic and terrestrial wave propagation problems from low frequencies up to millimeter waves.

BIBLIOGRAPHY

1. J. D. Kraus, *Antennas*, McGraw-Hill, 1988.
2. C. A. Balanis, *Antenna Theory*, 2nd ed., Wiley, 1997.
3. J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Dover Publications, New York, 1981 (republication of the 3rd Clarendon Press ed. of 1891).
4. E. K. Miller, L. Medgyesi-Mitchang, and E. H. Newman, eds., *Computational Electromagnetics, Frequency-Domain Method of Moments*, IEEE Press, 1992.
5. A. Taflove, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, Artech House, Boston, 1995.
6. J. L. Volakis, A. Chatterjee, and L. C. Kempel, *Finite Element Method for Electromagnetics*, IEEE Press, USA, 1998.
7. D. M. Pozar and D. H. Schaubert, eds., *Microstrip Antennas*, IEEE Press, 1995.
8. Special issue on advanced numerical techniques in electromagnetics, *IEEE Trans. Antennas and Propagation*, March 1997 issue.
9. D. Filipović and J. L. Volakis, Design and demonstration of a novel conformal slot spiral antenna for VHF to L-band operation, *2001 IEEE Antennas and Propagation Symp. Digest*, 2001.
10. K. S. Kunz and R. J. Luebbers, *The Finite Difference Time Domain Method for Electromagnetics*, CRC Press, Boca Raton, FL, 1993.
11. A. Arif Ergin, B. Shanker, and E. Michielssen, The plane-wave time-domain algorithm for the fast analysis of transient wave phenomena, *IEEE Antennas Propag. Mag.* **41**(4): 39–52 (Aug. 1999).
12. T. K. Sarkar, W. Lee, and S. M. Rao, Analysis of transient scattering from composite arbitrarily shaped complex structures, *IEEE Trans. Antennas Propag.* **48**(10): 1625–1634 (Oct. 2000).
13. T. Abboud, J.-C. Nedeléc, and J. L. Volakis, Stable solution of the retarded potential equations, *Proc. 17th Annual Progress Review in Applied Electromagnetism* (*ACES*), *Digest*, Monterey CA, March 2001, pp. 146–151.
14. A. Peterson, S. Ray, and R. Mittra, *Computational Methods for Electromagnetics*, IEEE Press, 1998.
15. P. Silvester and R. Ferrari, *Finite Elements for Electrical Engineers*, 2nd ed., Cambridge Univ. Press, 1990.
16. R. C. Hansen, ed., *Geometrical Theory of Diffraction*, IEEE Press, 1981.
17. R. J. Marhefka and W. D. Burnside, Antennas on complex platforms, *Proc. IEEE* **80**: 204–208 (Jan. 1992).
18. U. Jakobus and F. M. Landstorfer, Improvement of the PO-MoM hybrid method by accounting for effects of perfectly conducting wedges, *IEEE Trans. Antennas Propag.* **43** (1995).
19. J. L. Volakis, T. Özdemir, and J. Gong, Hybrid finite element methodologies for antennas and scattering, *IEEE Antennas Propag.* 493–507 (March 1997).
20. W. C. Chew, J.-M. Jin, E. Michielssen, and J. M. Song, *Fast and Efficient Algorithms in Computational Electromagnetics*, Artech House, Boston, 2001.
21. E. Bleszynski, M. Bleszynski, and T. Jaroszewicz, AIM: Adaptive integral method compression algorithm for solving large-scale electromagnetic scattering and radiation problems, *Radio Sci.* **31**(5): 1225–1251 (Sept./Oct. 1996).

22. T. F. Eibert and J. L. Volakis, Adaptive integral method for hybrid FE/BI modelling of 3D doubly periodic structures, *IEE Proc. Microwaves, Antennas Propag.* **146**(1): 17–22 (Feb. 1999).

23. Z. Li, Y. E. Erdemli, J. L. Volakis, and P. Y. Papalambros, Design optimization of conformal antennas with the hybrid finite element method, *IEEE Antennas Propag.*

24. E. Arvas, A. Rahhal-Arabi, A. Sadigh, and S. M. Rao, Scattering from multiple conducting and dielectric bodies of arbitrary shape, *IEEE AP Mag.* **33**(2): 29–36 (April 1991).

25. J. van Bladel, *Electromagnetic Fields*, Hemisphere Publishing, New York, 1985.

26. J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, New York, 1941.

27. R. F. Harrington, *Field Computation by Moment Methods*, Macmillan, New York, 1968.

28. L. S. Canino et al., Numerical solution of the Helmholtz equation in 2D and 3D using a high-order Nyström discretization, *J. Comput. Phys.* **146**: 627–633 (1998).

29. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Fannery, *Numerical Recipes in C*, Cambridge Univ. Press, Cambridge, UK, 1992.

30. D. R. Wilton et al., Potential integrals for uniform and linear source distributions on polygonal and polyhedral domains, *IEEE Trans. AP* **32**(3): 276–281 (March 1984).

31. T. F. Eibert and V. Hansen, On the calculation of potential integrals for linear source distributions on triangular domains, *IEEE Trans. Antennas Propag.* **43**(12): 1499–1502 (Dec. 1995).

32. M. G.Duffy, Quadrature over a pyramid or cube of integrands with a singularity at a vertex, *SIAM J. Num. Anal.* **19**(6): 1260–1262 (Dec. 1982).

33. E. K. Miller and E. J. Deadrick, Some computational aspects of thin wire modeling, in R. Mittra, ed., *Numerical and Asymptotic Techniques in Electromagnetics*, Springer-Verlag, 1975.

34. R. Mittra, *Computer Techniques for Electromagnetics*, Pergamon Press, New York, 1973.

35. W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, 2nd ed., Wiley, New York, 1998.

36. S. M. Rao, D. R. Wilton, and A. W. Glisson, Electromagnetic scattering by surfaces of arbitrary shape, *IEEE Trans. AP* **30**(3): 409–418 (May 1982).

37. S. Wandzura, Electric current basis functions for curved surfaces, *Electromagnetics* **12**: 77–91 (1992).

38. G. E. Antilla and N. G. Alexopoulos, Scattering from complex three-dimensional geometries by a curvilinear hybrid fine element–integral equation approach, *J. Opt. Soc. Am. A* **11**: 1445–1457 (April 1994).

39. R. D. Graglia, D. R. Wilton, and A. F. Peterson, Higher-order interpolatory vector bases for computational electromagnetics, *IEEE Trans. Antennas Propag.* **45**: 329–342 (March 1997).

40. D. H. Schaubert, D. R. Wilton, and A. W. Glisson, A tetrahedral modeling method for electromagnetic scattering by arbitrarily shaped inhomogeneous dielectric bodies, *IEEE Trans. Antennas Propag.* **32**(1): 77–85 (Jan. 1984).

41. T. J. Peters and J. L. Volakis, Application of a conjugate gradient FFT method to scattering by thin material plates, *IEEE Trans. Antennas Propag.* **AP-36**: 518–526 (April 1988).

42. P. Silvester, Fine element solution of homogeneous waveguide problems, *Alta Freq.* **38**: 313–317 (1969).

43. T. B. A. Senior and J. L. Volakis, *Approximate Boundary Conditions in Electromagnetics*, IEE Press, 1995.

44. J. M. Jin and J. L. Volakis, A hybrid finite element method for scattering and radiation from microstrip patch antennas and arrays residing in a cavity, *IEEE Trans. Antennas Propag.* **A-39**: 1598–1604 (1991).

45. J. L. Volakis, T. F. Eibert, and K. Sertel, Fast integral methods for conformal antenna and array modeling in conjunction with hybrid finite element formulations, *Radio Sci.* **35**(2): 537–546 (March–April 2000).

ANTENNAS

CONSTANTINE A. BALANIS
ANASTASIS C. POLYCARPOU
Arizona State University
Tempe, Arizona

1. INTRODUCTION

An antenna is the system component that is designed to radiate or receive electromagnetic waves. In other words, the antenna is the electromagnetic transducer that is used to convert, in the receiving mode, free-space waves to guided waves. In a modern wireless system, the antenna must also act as a directional device to optimize or accentuate the transmitted or received energy in some directions while suppressing it in the others. The antenna serves to the communication system the same purpose that eyes and eyeglasses serve to a human.

The history of antennas dates back to James Clerk Maxwell, who unified the theories of electricity and magnetism and eloquently represented their relations through a set of profound equations best known as *Maxwell's equation*. His work was first published in 1873. He also showed that light was electromagnetic, and that both light and electromagnetic waves travel by wave disturbances of the same speed. In 1886, Professor Heinrich Rudolph Hertz demonstrated the first wireless electromagnetic system. He was able to produce in his laboratory at a wavelength of 4 m a spark in the gap of a transmitting $\lambda/2$ dipole that was then detected as a spark in the gap of a nearby loop. It was not until 1901 that Guglielmo Marconi was able to send signals over large distances. He performed, in 1901, the first transatlantic transmission from Poldhu in Cornwall, England, to St. John's, Newfoundland [1].

From Marconi's inception through the 1940s, antenna technology was centered primarily on wire-related radiating elements and frequencies up to about UHF. It was not until World War II that modern antenna technology was launched and new elements (waveguide apertures, horns, reflectors, etc.) were primarily introduced. A contributing factor to this new era was the invention of microwave sources (such as the klystron and magnetron) with frequencies of 1 GHz and above.

While World War II launched a new era in antennas, advances made in computer architecture and wireless

communications technology during the 1960s–1990s have had a major impact on the advance of modern antenna technology, and they are expected to have an even greater influence on antenna engineering in the new millennium. Beginning primarily in the early 1960s, advanced numerical and computational methods were introduced that allowed previously intractable complex antenna system configurations to be analyzed and designed very accurately. Antenna design plays a critical role in overall system design since the success of a system strongly relies on the performance of the antenna. Detailed analysis, design, and measurements of antennas can be found in Ref. 2. A tutorial on antennas is described in Ref. 3.

2. ANTENNA ELEMENTS

Prior to World War II, most antenna elements were of the wire type (long wires, dipoles, helices, rhombuses, fans, etc.), and they were used either as single elements or in arrays. During and after World War II, many other radiators, some of which may have been known for some time and others of which were relatively new, were put into service. This created a need for better understanding and optimization of their radiation characteristics. Many of these antennas were of the aperture type (e.g., open-ended waveguides, slots, horns, reflectors, lenses), and they have been used for communication, radar, remote sensing, and deep-space applications on both airborne and earth-based platforms. Many of these operate in the microwave region.

Prior to the 1950s, antennas with broadband pattern and impedance characteristics had bandwidths not much greater than about 2 : 1. In the 1950s, a breakthrough in antenna evolution was created that extended the maximum bandwidth to as great as 40 : 1 or more. Because the geometries of these antennas are specified by angles instead of linear dimensions, they have ideally an infinite bandwidth. Therefore, they are referred to as *frequency-independent* [2]. These antennas are primarily used in the 10–10,000 MHz region in a variety of applications, including TV, point-to-point communications, and feeds for reflectors and lenses.

It was not until almost 20 years later that a fundamental new radiating element, which has received a lot of attention and many applications since its inception, was introduced. This occurred in the early 1970s when the *microstrip* or *patch* antenna was reported [2]. This element is simple, lightweight, inexpensive, low-profile, and conformal to the surface. Microstrip antennas and arrays can be flush-mounted to metallic and other existing surfaces. Operational disadvantages of microstrip antennas include low efficiency, narrow bandwidth, and low power handling capabilities. Major advances in millimeter wave antennas have been made, including integrated antennas where active and passive circuits are combined with radiating elements in one compact unit (monolithic form).

The unparalleled advances in telecommunications have brought a dramatic increased interest and activity in antenna design. This has resulted in many new elements and design concepts [4], including increased interest in adaptive arrays and "smart" antennas [5,6].

3. THEORY

To analyze an antenna system, the sources of excitation are specified, and the objective is to find the electric and magnetic fields radiated by the elements. Once this is accomplished, a number of parameters and figures of merit (directivity, input impedance, effective area, polarization, etc.) that characterize the performance of the antenna system can be found. To design an antenna system, the characteristics of performance are specified, and the sources to satisfy the requirements are sought.

3.1. Maxwell's Equations

An antenna system is an electromagnetic boundary problem. Therefore, the fields radiated must satisfy Maxwell's equations, which, for lossless medium ($\sigma = 0$) and time-harmonic fields (assuming an $e^{j\omega t}$ time convention), can be written as

$$\nabla \times \vec{E} = -\vec{M}_i - j\omega\mu\vec{H} \tag{1a}$$

$$\nabla \times \vec{H} = \vec{J}_i + j\omega\mu\vec{E} \tag{1b}$$

$$\nabla \cdot (\varepsilon\vec{E}) = q_{ve} \tag{1c}$$

$$\nabla \cdot (\mu\vec{H}) = q_{vm} \tag{1d}$$

In Eqs. (1a)–(1d) both electric $\vec{J}_i$ and magnetic $\vec{M}_i$ current densities, and electric q_{ve} and magnetic q_{vm} charge densities, are allowed to represent the sources of excitation. The respective current and charge densities are related by the continuity equations

$$\nabla \cdot \vec{J}_i = -j\omega q_{ve} \tag{2a}$$

$$\nabla \cdot \vec{M}_i = -j\omega q_{vm} \tag{2b}$$

Although magnetic sources are not physical, they are often introduced as *electrical equivalents* to facilitate solutions of physical boundary-value problems [2,7]. In fact, for some configurations, both electric and magnetic equivalent current densities are used to represent actual antenna systems. For a metallic wire antenna, such as a dipole, an electric current density is used to represent the antenna. However, an aperture antenna, such as a waveguide or horn, can be represented by either an equivalent magnetic current density or by an equivalent electric current density or both. For a radiation problem, the first step is to represent the antenna excitation by its source, represented by the current density $\vec{J}_i$ or $\vec{M}_i$ or both. The next step is to solve the Maxwell equations, subject to a given set of boundary conditions, for $\vec{E}$ and $\vec{H}$. This is a difficult step, and it usually involves an integral with a complicated integrand. This procedure is represented in Fig. 1 as path 1.

To reduce the complexity of the problem, it is common practice to break the procedure into two steps. This is represented in Fig. 1 by path 2. The first step involves an integration while the second involves a differentiation. To accomplish this, auxiliary vector potentials are introduced. The most commonly used potentials are $\vec{A}$ (magnetic vector potential) and $\vec{F}$ (electric vector potential). Although the electric and magnetic field intensities ($\vec{E}$ and $\vec{H}$) represent

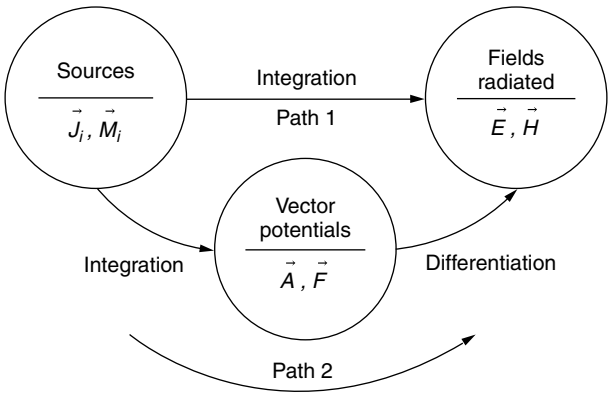


Figure 1. Procedure to solve antenna radiation.

physically measurable quantities, for most engineers the vector potentials are strictly mathematical tools [7]. The procedure along with the appropriate analytical formulations are detailed in Refs. 2 and 7.

3.2. Field Regions

The space surrounding an antenna is usually subdivided into three regions: the *reactive near-field* region, the *radiating near-field* (*Fresnel*) region, and the *far-field* (*Fraunhofer*) region. These regions are so designated to identify the field structure in each. Although no abrupt changes in the field configurations are noted as the boundaries are crossed, there are distinct differences among them. The boundaries separating these regions are not unique, although various criteria have been established and are commonly used to identify the regions [2]. The following definitions in quotations are from an IEEE standard [8].

1. The *reactive near-field region* is defined as "that region of the field immediately surrounding the antenna wherein the reactive field predominates." For most antennas, the outer boundary of this region is commonly taken to exist at a distance $R < 0.62\sqrt{D^3/\lambda}$ from the antenna, where λ is the wavelength and D is the largest dimension of the antenna.
2. The *radiating near-field* (*Fresnel*) *region* is defined as "that region of the field of an antenna between the reactive near-field region and the far-field region wherein radiation fields predominate and wherein the angular field distribution is dependent on the distance from the antenna." The radial distance R over which this region exists is $0.62\sqrt{D^3/\lambda} \leq R < 2D^2/\lambda$ (provided D is large compared to the wavelength). This criterion is based on the maximum phase error of $\pi/8$ radians (22.5°). In this region the field pattern is, in general, a function of the radial distance and the radial field component may be appreciable.
3. The *far-field* (*Fraunhofer*) *region* is defined as "that region of the field of an antenna where the angular field distribution is essentially independent of the distance from the antenna." In this region, the real part of the power density is dominant. The radial distance of this region is $R \geq 2D^2/\lambda$ (provided D is large compared to the wavelength). The outer boundary is ideally at infinity. The criterion is also based on the maximum phase error of $\pi/8$ radians (22.5°). In this region, the field components are essentially transverse to the radial direction, and the angular distribution is independent of the radial distance.

3.2.1. Far Field. The analytical formulation followed to find the field radiated by an antenna at any point, near field or far field, is in general complex and is outlined in detail in Ref. 2. Since antennas are primarily used to communicate over long distances, only the procedure used to find the fields in the far zone will be summarized; it is also less complex.

The following procedure can be followed to determine the electric and magnetic fields radiated by an antenna at an observation point in the far-field region. Once the current densities $\vec{J}_s$ and/or $\vec{M}_s$, either physical or equivalent, are selected to represent the physical antenna, then the vector potentials $\vec{A}$ and $\vec{F}$ are found according to

$$\vec{A} = \frac{\mu}{4\pi} \iint_S \vec{J}_s \frac{e^{-j\beta R}}{R}\, ds' \tag{3a}$$

$$\vec{F} = \frac{\varepsilon}{4\pi} \iint_S \vec{M}_s \frac{e^{-j\beta R}}{R}\, ds' \tag{3b}$$

where R is the distance from any point on the source to the observation point and β is the phase constant ($\beta = 2\pi/\lambda$). The surface current densities $\vec{J}_s$ and $\vec{M}_s$ have the units of A/m and V/m (amperes and volts per meter), respectively. If the current densities are distributed over a volume, the surface integrals of Eqs. (3a) and (3b) are replaced by volume integrals; line integrals are used for thin wire elements.

In the far-field (Fraunhofer) region, the radial distance R of Fig. 2a can be approximated by

$$R \cong \begin{cases} r - r' \cos\psi & \text{for phase terms} \quad (4a) \\ r & \text{for amplitude terms} \quad (4b) \end{cases}$$

Graphically, the approximation of (4a) is illustrated in Fig. 2b, where the radial vectors R and r are parallel to each other. Although such relation is strictly valid only at infinity, it becomes more accurate as the observation point is moved outward at radial distances exceeding $2D^2/\lambda$. Since the far-field region extends at radial distances of at least $2D^2/\lambda$, the approximation of (4a) for the radial distances R leads to phase errors that do not exceed $\pi/8$ radians (22.5°). It has been shown that such phase errors do not have a pronounced effect on the variations of the far-field amplitude patterns.

Using the approximations of (4a) and (4b) for observations in the far-field region, the integrals in (3a) and (3b)

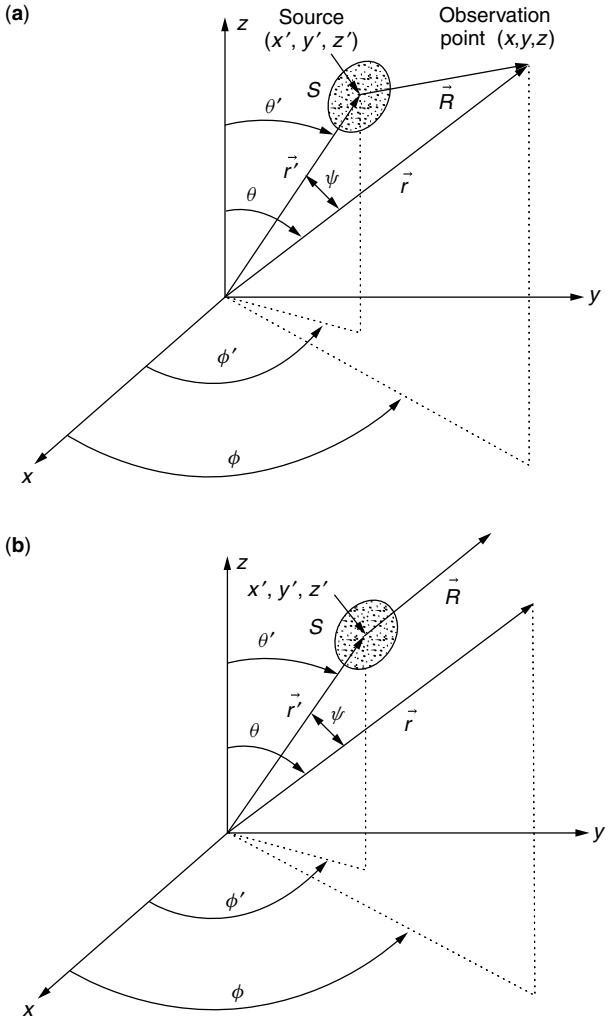


Figure 2. Coordinate system arrangements for (**a**) near-field and (**b**) far-field radiation.

can be reduced to

$$\vec{A} \cong \frac{\mu}{4\pi}\frac{e^{-j\beta r}}{r}\iint_S \vec{J}_s e^{j\beta r\cos\psi}\,ds' = \frac{\mu}{4\pi}\frac{e^{-j\beta r}}{r}\vec{N} \tag{5a}$$

$$\vec{N} = \iint_S \vec{J}_s e^{j\beta r'\cos\psi}\,ds' \tag{5b}$$

$$\vec{F} \cong \frac{\varepsilon}{4\pi}\frac{e^{-j\beta r}}{r}\iint_S \vec{M}_s e^{j\beta r\cos\psi}\,ds' = \frac{\varepsilon}{4\pi}\frac{e^{-j\beta r}}{r}\vec{L} \tag{6a}$$

$$\vec{L} = \iint_S \vec{M}_s e^{j\beta r'\cos\psi}\,ds' \tag{6b}$$

Once the vector potentials are determined, the corresponding spherical components of the electric and magnetic fields in the far-field region can be found in scalar form using [2]

$$E_r \cong 0 \tag{7a}$$

$$E_\theta \cong -j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}[L_\phi + \eta N_\theta] \tag{7b}$$

$$E_\phi \cong j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}[L_\theta - \eta N_\phi] \tag{7c}$$

$$H_r \cong 0 \tag{8a}$$

$$H_\theta \cong j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}\left[N_\phi - \frac{L_\theta}{\eta}\right] \tag{8b}$$

$$H_\phi \cong -j\frac{\beta}{4\pi}\frac{e^{-j\beta r}}{r}\left[N_\theta + \frac{L_\phi}{\eta}\right] \tag{8c}$$

where η is the intrinsic impedance of the medium ($\eta = \sqrt{\mu/\varepsilon}$) while N_θ, N_ϕ and L_θ, L_ϕ are the spherical θ and ϕ components of $\vec{N}$ and $\vec{L}$ from (5b) and (6b), respectively. In antenna theory, the spherical coordinate system is the one most widely used.

By examining (7a)–(8c), it is apparent that

$$E_\theta \cong \eta H_\phi \tag{9a}$$

$$E_\phi \cong -\eta H_\theta \tag{9b}$$

The relations of (9a) and (9b) indicate that in the far-field region the fields radiated by an antenna, and observed in a small neighborhood on the surface of a large-radius sphere, have all the attributes of a transverse electromagnetic (TEM) wave whereby the corresponding electric and magnetic fields are orthogonal to each other and to the radial direction.

To use the procedure described above, the sources representing the physical antenna structure must radiate into an infinite homogeneous medium. If that is not the case, then the problem must be reduced further (e.g., through the use of a theorem, such as the image theorem) until the sources radiate into an infinite homogeneous medium.

4. ANTENNA SOURCE MODELING

The first step in the analysis of the fields radiated by an antenna is the specification of the sources to represent the antenna. Here we will present two examples of source modeling: one for thin-wire modeling (such as a dipole) and the other for an aperture antenna (such as a waveguide). These are two distinct examples each with a different source modeling; the wire requires an electric current density while the aperture is represented by an equivalent magnetic current density.

4.1. Wire Source Modeling

Let us assume that the wire antenna is a dipole, as shown in Fig. 3. If the wire has a circular cross section with radius a, the electric current density induced on the surface of the wire will be symmetric about the circumference (no ϕ variations). If the wire is also very thin ($a \ll \lambda$), it is common to assume that the excitation source representing the antenna is a current along the axis of the wire. This current must vanish at the ends of the wire. For a center-fed resonant dipole, the excitation is often represented by [2]

$$\vec{I}_e = \begin{cases} \hat{a}_z I_0 \sin\left[\beta\left(\dfrac{l}{2} - z'\right)\right] & 0 \le z' \le \dfrac{l}{2} \\ \hat{a}_z I_0 \sin\left[\beta\left(\dfrac{l}{2} + z'\right)\right] & -\dfrac{l}{2} \le z' \le 0 \end{cases} \tag{10}$$

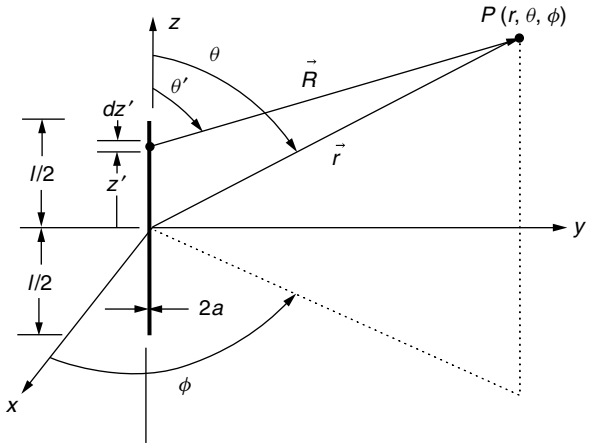


Figure 3. Dipole geometry for electromagnetic wave radiation.

No magnetic source representation is necessary for this type of antenna. For observations in the far-field region, the fields radiated by the antenna can be found using (5a)–(8c) where the surface integrals are replaced by line integrals. On the basis of these far-field expressions, the radiated electric and magnetic fields can be written as

$$E_\theta \cong j\eta \frac{I_0 e^{-j\beta r}}{2\pi r}\left[\frac{[\cos(\beta l/2)\cos\theta] - \cos(\beta l/2)}{\sin\theta}\right] \quad (11a)$$

$$H_\phi \cong \frac{E_\theta}{\eta} \quad (11b)$$

To illustrate the field variation of (11a), a three-dimensional graph of the normalized field amplitude pattern for a half-wavelength ($l = \lambda/2$) dipole is plotted in Fig. 4. A 90° angular section of the pattern has been omitted to illustrate the figure-eight elevation plane pattern variation. As the length of the wire increases, the pattern becomes narrower. When the length exceeds one wavelength ($l = \lambda$), sidelobes are introduced into the elevation plane pattern.

Wire type radiating elements include dipoles, monopoles, loops, and helices. Arrays of dipoles are very popular for wireless communication base stations. Monopoles and very thin helices are widely used in cellular phones and mobiles, and in many wireless communication systems. Loop antennas are used in pagers and also being suggested for cellular phones [9].

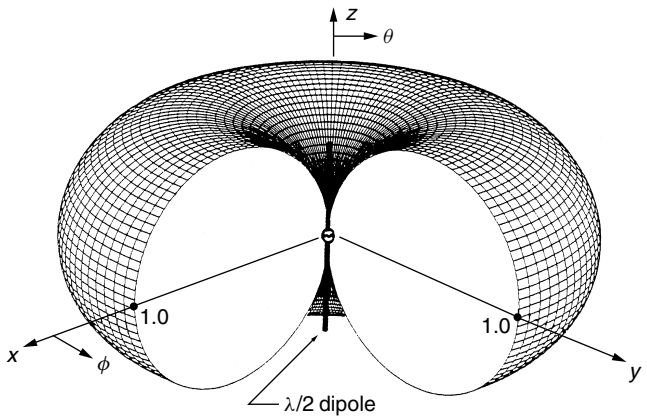


Figure 4. Three-dimensional amplitude pattern of a $\lambda/2$ dipole.

4.2. Aperture Source Modeling

To analyze aperture antennas, the most often used procedure is to model the source representing the actual antenna by the *field equivalence principle* (FEP), also referred to as *Huygen's principle* [2,7]. With this procedure, the actual antenna is replaced by equivalent sources (electric or magnetic or both) that, externally to a closed surface enclosing the actual antenna, produce the same fields as those radiated by the actual antenna. This procedure is analogous to the Thévenin equivalent of circuit analysis, which produces the same response, to an external load, as the actual circuit.

To demonstrate the use of FEP to calculate the fields radiated by an antenna, consider an open-ended rectangular waveguide aperture mounted on an infinite planar PEC (Perfect Electric Conductor) radiating in a semiinfinite homogeneous medium, as shown in Fig. 5a. Let us assume that the fields in the waveguide aperture are those of the dominant TE_{10} mode. Hence, the tangential electric field over the x–y plane is

$$\vec{E}_s = \begin{cases} \hat{a}_y E_0 \cos\left(\dfrac{\pi}{a}x'\right) & -\dfrac{a}{2} \le x' \le \dfrac{a}{2}, -\dfrac{b}{2} \le y' \le \dfrac{b}{2} \\ 0 & \text{elsewhere over the PEC} \end{cases} \quad (12)$$

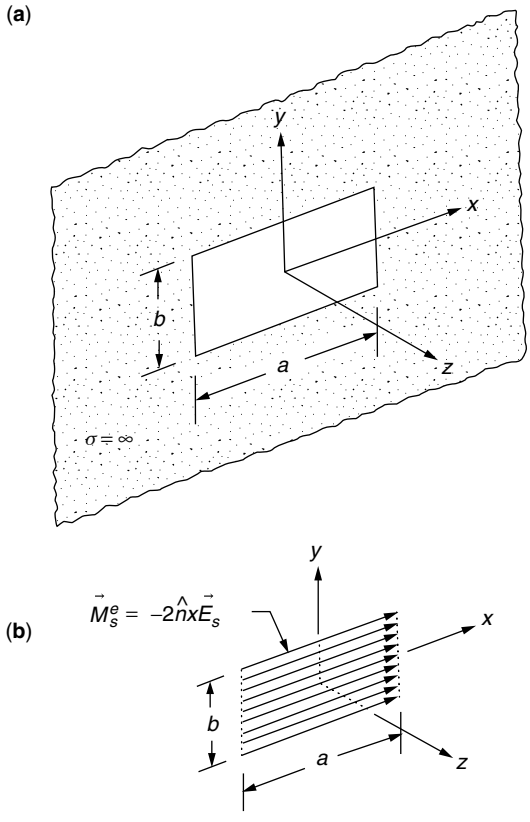


Figure 5. (**a**) Waveguide aperture on an infinite ground plane and (**b**) its equivalent.

By adopting the FEP, an imaginary closed surface is chosen to coincide with an infinite PEC (x–y plane) and covers also the waveguide aperture. The imaginary closed surface is chosen to coincide with the x–y plane because the tangential component of the electric field, and thus the equivalent magnetic current density, are nonzero only in the aperture. Using (12), we can write that

$$\vec{M}_s = -\hat{n} \times \vec{E}_s = \begin{cases} \hat{a}_x E_0 \cos\left(\frac{\pi}{a}x'\right) & -\frac{a}{2} \le x' \le \frac{a}{2}, -\frac{b}{2} \le y' \le \frac{b}{2} \\ 0 & \text{elsewhere over the PEC} \end{cases} \tag{13}$$

Using image theory, the infinite ground plane can be removed by replacing $\vec{M}_s$ with a magnetic current density of twice the strength of (13). The new $\vec{M}_s$ is now radiating into an infinite homogeneous medium, as shown in Fig. 5b. Using (5a)–(8c), the far-field electric and magnetic fields radiated by the waveguide can be written by

$$E_r \cong H_r \cong 0 \tag{14a}$$

$$E_\theta \cong -\frac{\pi}{2} C \sin\phi \frac{\cos(X)}{X^2 - \left(\frac{\pi}{2}\right)^2} \frac{\sin(Y)}{Y} \tag{14b}$$

$$E_\phi \cong -\frac{\pi}{2} C \sin\theta \cos\phi \frac{\cos(X)}{X^2 - \left(\frac{\pi}{2}\right)^2} \frac{\sin(Y)}{Y} \tag{14c}$$

$$H_\theta \cong -\frac{E_\phi}{\eta} \tag{14d}$$

$$H_\phi \cong +\frac{E_\theta}{\eta} \tag{14e}$$

$$X = \frac{\beta a}{2} \sin\theta \cos\phi \tag{14f}$$

$$Y = \frac{\beta b}{2} \sin\theta \sin\phi \tag{14g}$$

$$C = j\frac{ab\beta E_0 e^{-j\beta r}}{2\pi r} \tag{14h}$$

The three-dimensional normalized field pattern of a rectangular aperture with dimensions of $a = 3\lambda$ and $b = 3\lambda$ is shown in Fig. 6. The minor lobes formed throughout the space are clearly shown.

Aperture antennas include waveguides, horns, reflectors, and microstrips. Horns mounted on tall towers are used by telephone companies as transmitting and receiving antennas. Reflectors, because of their high gain, are utilized as ground-based antennas for spaceborne missions and at gateway stations of wireless communication systems. Microstrip antennas, because of their light weight, conformal shapes, low profile, versatility, and other attractive radiation characteristics, are excellent candidates for adaptive and smart antennas.

5. ANTENNA PARAMETERS AND FIGURES OF MERIT

Many different parameters and figures of merit characterize the performance of an antenna system. Some of the most important are included here.

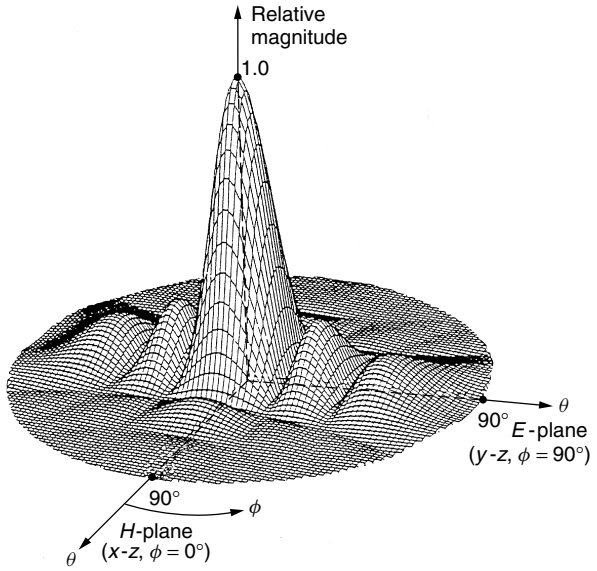


Figure 6. Three-dimensional amplitude radiation pattern for an aperture ($a = 3\lambda$, $b = 3\lambda$) on an infinite ground plane.

An *antenna pattern* is defined as a "graphical representation, usually in the far-field region, of one of the antenna's parameters. For a complete description, the parameters of interest are usually plotted as a function of the spherical angles θ, ϕ." Parameters of interest include amplitude, phase, polarization, and directivity. An amplitude pattern is usually comprised of a number of lobes.

A *main* (*major*) *lobe* is defined as "the radiation lobe containing the direction of maximum radiation. In certain antennas, such as multilobed or split-beam antennas, there may exist more than one major lobe." A *sidelobe* is defined as "a radiation lobe in any direction other than that of the major lobe." The amplitude level of a sidelobe relative to the main lobe (usually expressed in decibels) is referred to as *sidelobe level*.

An antenna, in the transmitting and receiving modes, is often represented by a Thévenin equivalent circuit with an antenna impedance Z_A, as shown in Fig. 7. The antenna impedance Z_A consists of the *radiation resistance* R_r, the *loss resistance* R_L, and an imaginary part $X_A [Z_A = R_A + jX_A = (R_r + R_L) + jX_A]$. The radiation resistance is the resistance that represents antenna radiation or scattering. The loss resistance is the resistance that accounts for the conductive and dielectric losses of the antenna. Expression for R_r and R_L for dipoles and small circular loops can be found, respectively, in Chaps. 4 and 5 of Ref. 2.

Input impedance is defined as "the impedance presented by an antenna at its terminals." It is expressed at the terminals as the ratio of the voltage to current or the ratio of the appropriate components of the electric to magnetic fields, and it is usually complex. When the antenna impedance Z_A is referred to the input terminals of the antenna, it reduces to the input impedance.

Radiation efficiency is defined as "the ratio of the total power radiated by an antenna to the net power accepted by an antenna from the connected transmitter." Using the

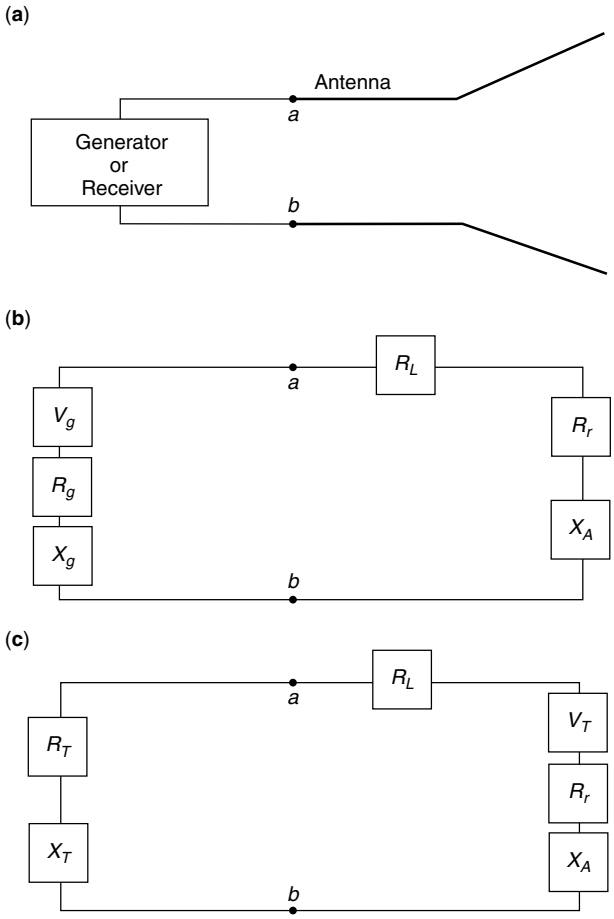


Figure 7. Thévenin equivalents for transmitting and receiving antennas: (**a**) antenna system; (**b**) Thévenin equivalent — transmitting; (**c**) Thévenin equivalent — receiving.

equivalent-circuit representation of an antenna of Fig. 7, the radiation efficiency can be written as

$$e_r = \frac{R_r}{R_r + R_L} \tag{15}$$

Power density S is defined as the power density [in watts per square meter (W/m^2)] of the fields radiated by the antenna. In general, the power density is complex. In the reactive near field, the imaginary component is dominant. In the far field, the real part is dominant. In equation form, the power density $\vec{S}$ is expressed as

$$\vec{S} = \tfrac{1}{2}\vec{E} \times \vec{H}^* = \vec{S}_r + j\vec{S}_i \tag{16}$$

where $\vec{E}$ and $\vec{H}$ are the fields radiated by the antenna (*indicates complex conjugate). The real part of (16) is usually referred to as *radiation density*.

Radiation intensity U is defined as "the power radiated from an antenna per unit solid angle (steredian)." The radiation intensity is usually defined in the far field and is related to the real part of the power density by

$$U = r^2 S_r \tag{17}$$

where r is the spherical radial distance.

Beamwidth is defined as the angular separation between two directions in which the radiation intensity is identical, with no other intermediate points of the same value. When the intensity is one-half of the maximum, it is referred to as *half-power beamwidth*.

An *isotropic radiator* is defined as "a hypothetical, lossless antenna having equal radiation intensity in all directions." Although such an antenna is an idealization, it is often used as a convenient reference to express the directive properties of actual antennas. The radiation intensity S_{r0} of an isotropic radiator and intensity U_0 are defined, respectively, as

$$S_{r0} = \frac{P_r}{4\pi r^2} \tag{18a}$$

$$U_0 = \frac{P_r}{4\pi} \tag{18b}$$

where P_r represents the power radiated by the antenna.

Directivity is one of the most important figures of merit that describes the performance of an antenna. It is defined as "the ratio of the radiation intensity in a given direction from the antenna to the radiation intensity averaged over all direction." Using (18b), it can be written as

$$D = \frac{U(\theta, \phi)}{U_0} = \frac{4\pi U(\theta, \phi)}{P_r} \tag{19}$$

where $U(\theta, \phi)$ is the radiation intensity in the direction θ, ϕ and P_r is the radiated power. For antennas radiating both electric field components (E_θ and E_ϕ), partial directivities D_θ and D_ϕ can be defined as associated, respectively, with E_θ and E_ϕ [2]. The total directivity is then the sum of the two ($D = D_\theta + D_\phi$). If the direction of observation is not specified, it implies the direction of maximum radiation intensity (maximum directivity) expressed as

$$D_0 = \frac{U_m(\theta, \phi)}{U_0} = \frac{4\pi U_m(\theta, \phi)}{P_r} \tag{20}$$

The directivity is an indicator of the relative directional properties of the antenna. As defined by Eqs. (19) and (20), the directional properties of the antenna in question are compared to those of an isotropic radiator. Figure 8 displays the directivity pattern of a $\lambda/2$ dipole and an isotropic source. In each angular direction, only the greater directivity between the two radiators is shown. This allows us to relate the directivity of the element in question to that of an isotropic radiator by simply adding (if expressed in decibels) the relative directivities of one element to another. This procedure is analogous to that used to determine the overall gain of cascaded amplifiers.

Gain is probably the most important figure of merit of an antenna. It is defined as "the ratio of the radiation intensity in a given direction, to the radiation intensity that would be obtained if the power accepted by the antenna were radiated isotropically." Antenna gain is expressed as

$$G = \frac{4\pi U(\theta, \phi)}{P_a} \tag{21}$$

where P_a is the accepted (input) power of the antenna. If the direction is not specified, it implies the direction of

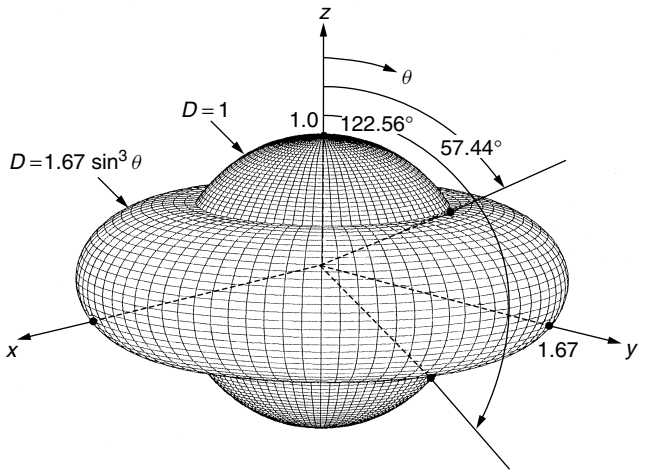


Figure 8. Three-dimensional directivity patterns of a $\lambda/2$ dipole and isotropic radiator.

maximum radiation (maximum gain). In simplest terms, the main difference between the definitions of directivity and gain is that the directivity is based on the radiated power while the gain is based on the accepted (input) power. Since all of the accepted (input) power is not radiated (because of losses), the two are related by

$$P_r = e_r P_a \tag{22}$$

where e_r is the radiation efficiency of the antenna as defined by Eq. (15). By using (19), (21), and (22) the gain can be expressed as

$$G = e_r \frac{4\pi U(\theta, \phi)}{P_r} = e_r D \tag{23}$$

For a lossless antenna, its gain is equal to its directivity.

Antenna polarization in a given direction is determined by the polarization of the fields radiated by the antenna. In general, the polarization of an antenna is classified as linear, circular, or elliptical. Although linear and circular polarizations are special cases of elliptical polarization, in practice they are usually treated separately. Circular and elliptical polarizations also are classified according to the rotation of the transmitted field vectors; the rotation can be either clockwise (right-hand) or counterclockwise (left-hand) as viewed in the direction of propagation.

Polarization efficiency (polarization mismatch or loss factor) is defined as "the ratio of the power received by an antenna from a given plane wave of arbitrary polarization to the power that would be received by the same antenna from a plane wave of the same power flux density and direction of propagation, whose state of polarization has been adjusted for a maximum received power." This is an important factor that must be included in the power budget of communications systems and is one that sometimes is neglected.

Effective area in a given direction is defined as "the ratio of the available power at the terminals of a receiving antenna to the power flux density of a plane wave incident on the antenna from that direction, *the wave being polarization-matched to the antenna*. If the direction is not specified, the direction of maximum radiation intensity is implied." The maximum effective area is related to the antenna gain by

$$A_{em} = pq \left(\frac{\lambda^2}{4\pi} \right) G_0 \tag{24}$$

where p is the polarization efficiency or polarization loss factor, G_0 is the maximum gain of the antenna, and q is the impedance matching efficiency between the transmission line and the antenna defined as

$$q = (1 - |\Gamma_{\text{in}}|^2) \tag{25}$$

where Γ_{in} is the reflection coefficient at the input terminals of the antenna. When multiplied by the power density of the incident wave that impinges on the antenna, the maximum effective area determines the maximum power that is delivered to a matched load connected to the antenna.

Aperture efficiency, usually expressed in percent, is defined as the ratio of antenna's maximum effective aperture to its physical aperture, which can also be expressed on the ratio of the maximum directivity of the aperture to the standard directivity, or

$$\varepsilon_{ap} = \frac{A_{em}}{A_p} = \frac{D_0}{D_s} \tag{26}$$

where A_p is the physical area of the antenna and D_s is the standard directivity of antenna ($4\pi A_p/\lambda^2$ when $A_p \gg \lambda^2$ and with radiation confined to a half-space).

For a rectangular aperture mounted on an infinite ground plane and with a triangular aperture distribution, its aperture efficiency is 75%. However, for an aperture with a sinusoidal aperture distribution, its aperture efficiency is 81%. Again, we see that the aperture distribution, which satisfies the wave equation and the boundary conditions of the structure, determines its aperture efficiency. If an aperture could support a uniform field distribution, its aperture efficiency would be 100%.

6. ARRAYS

Specific radiation pattern requirements usually cannot be achieved by a single antenna element, because single elements usually have relatively wide radiation patterns and low directivities. To design antennas with very large directivities, it is usually necessary to increase the electrical size of the antenna. This can be accomplished by enlarging the electrical dimensions of the chosen single element. However, mechanical problems are usually associated with very large elements.

An alternative way to achieve large directivities, without increasing the size of the individual elements, is to use multiple single elements to form an array. An array is really a sampled version of a very large single element. In an array, the mechanical problems of large single elements are traded for the electrical problems associated with the feed networks of arrays. However, with today's solid-state

technology, very efficient and low-cost feed networks can be designed.

Arrays are the most versatile antenna systems. They find wide applications not only in many spaceborne systems but also in many earth-bound missions. In most cases, the elements of an array are identical; this is not necessary, but it is often more convenient, simpler, and more practical. In general, the radiation characteristics of an array depend on many factors, some of which are:

1. The geometric configuration of the overall array (linear, circular, rectangular, spherical, etc.)
2. The relative displacement between the elements
3. The excitation amplitude of the individual elements
4. The excitation phase of the individual elements
5. The relative pattern of the individual elements

Therefore the designer has many controls or degrees of freedom that can be exercised in order to make the antenna very versatile and meet the specifications of the design.

With arrays, it is practical to not only *synthesize* almost any desired amplitude radiation pattern, but the main lobe can be scanned, resulting in a *scanning* array, by controlling the relative phase excitation between the elements. This is most convenient for applications where the antenna system is not readily accessible, especially for spaceborne missions. The beamwidth of the main lobe along with the sidelobe level can be controlled by the relative amplitude excitation (distribution) between the elements of the array. In fact, there is a tradeoff between the beamwidth and the sidelobe level, based on the amplitude distribution of the elements [2]. The spacing between the elements can be used to control many characteristics of an array, including the pattern, beamwidth, bandwidth, input impedance, and sidelobe level.

There are a plethora of array designs. Two classic array configurations include the Yagi–Uda and log-periodic arrays [2]. The Yagi–Uda is a popular antenna used by amateur radio enthusiasts and for TV. The log-periodic array, because of its large and attractive bandwidth, is probably the most widely used home TV antenna. Arrays of waveguides, horns, reflectors, and microstrips are also very popular [10]. Microstrip arrays will play a key role in the realization of unique designs of adaptive and smart antennas for wireless communications.

Designs of uniform distribution arrays include the broadside, end-fire, and scanning arrays. Classic nonuniform distribution arrays include the binomial, Dolph–Tschebyscheff, Woodward–Lawson, Fourier transform, and Taylor (Chebyshev error and line source) [2]. There are many other array designs, too numerous to name here.

7. ADAPTIVE ARRAYS AND SMART ANTENNAS

In *adaptive antenna arrays* [11,12], the amplitude and phase distribution between the elements are adaptively chosen to improve signal reception or transmission in certain directions and reduce noise and interference in all other directions. Adaptive signal processing algorithms are often used in conjunction with the array architecture to obtain an optimum set of weights that maximizes *signal-to-noise ratio* (SNR) and minimizes *mean-square error* (MSE). In this context, such adaptive arrays are commonly referred to as *smart antennas* [5]. In *code-division multiple-access* (CDMA) applications, such as cellular and mobile communications, smart antennas at the base station can form a main beam toward the subscriber and low-level sidelobes or nulls toward interfering signals. Through adaptive beamforming, smart antennas can penetrate through buildings and cover areas that are otherwise unattainable by a single-element antenna at the base station. In addition, smart antennas can more effectively alleviate problems due to multipath, fading, and time dispersion. This results in an improved system performance compared to an isotropic antenna. Also, dynamic beamforming in smart antennas enhances antenna gain in the direction of the subscriber, thus extending signal coverage. Coverage enhancement can reduce manufacturing cost in cellular and mobile communications by requiring a smaller number of base stations within a given area.

8. CONCLUSIONS

Antenna engineering has enjoyed a very successful period of development since 1950. Responsible for its success have been the introduction and technological advances of some new elements of radiation, such as aperture antennas, horns, reflectors, frequency-independent antennas, and microstrip/patch antennas. Excitement has been created by the advancement of numerical methods that have been instrumental in analyzing many previously intractable problems. Another major factor in the success of antenna technology has been the advances in the computer architecture and wireless communications. Today antenna engineering is considered a truly fine engineering art.

Although a certain level of maturity has been attained, many challenging opportunities and problems remain to be solved. Unique and innovative adaptive and smart antenna designs for wireless communication are creating new enthusiasm and interest in the exploding wireless communication technology. Phased array architecture integrating monolithic MIC (Monolithic Integrated Circuits) technology is still a challenging problem. Integration of new materials into antenna technology offers many advantages, and numerical methods will play key roles in their incorporation and system performance. Computational efficiency in numerical methods will allow modeling, design, and optimization of antennas on complex platforms without the need of supercomputing capabilities. Innovating antenna designs to perform complex and demanding system functions always remain a challenge. New basic elements are always welcomed and offer refreshing opportunities.

BIOGRAPHIES

Constantine A. Balanis received the B.S.E.E. degree from Virginia Tech, Blacksburg, Virginia, in 1964, the

MEE degree from the University of Virginia, Charlottesville, Virginia, in 1966, and the Ph.D. degree in Electrical Engineering from Ohio State University, Columbus, in 1969.

From 1964 to 1970 he was with NASA Langley Research Center, Hampton VA, and from 1970 to 1983 he was with the Department of Electrical Engineering, West Virginia University, Morgantown, West Virginia. Since 1983, he has been with the Department of Electrical Engineering, Arizona State University, Tempe, where he is now regents' professor. His research interests are in low- and high-frequency computational methods for antennas and scattering, smart antennas for wireless communication, and high-intensity radiated fields (HIRF). He received the 2000 IEEE Third Millennium Medal, 1996–97 Arizona State University Outstanding Graduate Mentor Award, 1992 Special Professionalism Award from the IEEE Phoenix Section, the 1989 IEEE Region 6 Individual Achievement Award, and the 1987–1988 Graduate Teaching Excellence Award, School of Engineering, Arizona State University.

Dr. Balanis is a fellow of the IEEE, and he has served as associate editor of the *IEEE Transactions on Antennas and Propagation* and the *IEEE Transactions on Geoscience and Remote Sensing*, and as editor of the Newsletter for the IEEE Geoscience and Remote Sensing Society. He is the author of *Antenna Theory: Analysis and Design* (Wiley, 1997, 1982) and *Advanced Engineering Electromagnetics* (Wiley, 1989).

Anastasis C. Polycarpou received his B.S., M.S., and Ph.D. degrees in electrical engineering from Arizona State University in 1992, 1994, and 1998, respectively. He then joined the Department of Electrical Engineering as an associate research faculty where he worked on various funded research projects. At Arizona State University, he worked on the development and enhancement of numerical methods, in particular the finite element method and the method of moments, for the analysis of complex electromagnetic problems such as microwave circuits, interconnects and electronic packaging, cavity-backed slot antennas in the presence of magnetized ferrites, and helicopter electromagnetics. He has published ten journal papers and 25 conference proceedings. He is currently an associate professor at a small private college in Cyprus. His areas of interest are antennas, electromagnetic theory, and numerical methods.

BIBLIOGRAPHY

1. J. D. Kraus, Antennas since Hertz and Marconi, *IEEE Trans. Antennas Propag.* **AP-33**: 131–137 (Feb. 1985).
2. C. A. Balanis, *Antenna Theory: Analysis and Design*, Wiley, New York, 1997, 1982.
3. C. A. Balanis, Antenna theory: A review, *Proc. IEEE* **80**: 7–23 (Jan. 1992).
4. *Special Issue on Wireless Communications*, *IEEE Trans. Antennas Propag.* **AP-46**: (June 1998).
5. J. C. Liberti, Jr. and T. S. Rappaport, *Smart Antennas for Wireless Communications: IS-95 and Third Generation CDMA Applications*, Prentice-Hall PTR, Englewood Cliffs, NJ, 1999.
6. T. S. Rappaport, ed., *Smart Antennas: Adaptive Arrays, Algorithms, & Wireless Position Location*, IEEE, 1998.
7. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.
8. IEEE, *IEEE Standard Definitions of Terms for Antennas*, IEEE Standard 145-1983, reprinted in *IEEE Trans. Antennas Propag.* **AP-31**(Pt. II of two parts): 5–29 (Nov. 1983).
9. K. D. Katsibas, C. A. Balanis, P. A. Tirkas, and C. R. Birtcher, Folded loop antenna for mobile hand-held units, *IEEE Trans. Antennas Propag.* **AP-46**: 260–266 (Feb. 1998).
10. Special Issue on Phased Arrays, *IEEE Trans. Antennas Propag.* **47**(3): (March 1999).
11. Special Issue on Adaptive Antennas, *IEEE Trans. Antennas Propag.* **AP-24**: (Sept. 1976).
12. Special Issue on Adaptive Processing Antenna Systems, *IEEE Trans. Antennas Propag.* **AP-34**: (Sept. 1986).

ANTENNAS FOR MOBILE COMMUNICATIONS

MICHAEL T. CHRYSSOMALLIS
Democritus University of Thrace
Xanthi, Greece

CHRISTOS G. CHRISTODOULOU
University of New Mexico
Albuquerque, New Mexico

1. ANTENNAS AND MOBILE COMMUNICATION SYSTEM REQUIREMENTS

Although some mobile communication systems are alleged to have originated in 1885, broadly recognized first real mobile services started around 1900 with wireless telegraph on ships, introduced by G. Marconi. He used long vertical wire antennas in various forms; wire antennas were the main type used in mobile systems up to 1970, when integrated antennas appeared as a consequence of the rapid progress in semiconductor integrated circuits. Printed antenna technology introduced the possibility of producing lightweight, less bulky, low-cost, easy-to-manufacture radiating structures, fully compatible with the newly integrated electronic packages [1].

Antennas are used to transmit or receive electromagnetic waves, and also serve as transducers converting guided waves into free-space waves and vice versa. Although all antennas, regardless of their type, operate on the same basic principles of electromagnetic theory, different antenna systems require careful design and a good understanding of the radiation mechanisms involved. Electrical, mechanical, and operating costs usually determine the proper type of antenna, which serves best specific application at hand.

Antennas can be classified into different categories. For our purposes we are interested in antennas in mobile communication systems. Antennas are one of the most important parts in a wireless communication system since they are responsible for the proper transmission and reception of signals. A successful design can relax some of the

complex system requirements involved in a communication link and increase the overall system performance. The choice of an antenna for a specific wireless communication application, either fixed or mobile, depends on (1) the platform to be used, which can be a tower or a building, car, ship, spacecraft, satellite, or other vehicle; (2) the environment in which the communication operates, such as urban, rural area of land, sea, or space; (3) the frequency of operation of the link; and (4) the nature of the application, for example, voice, data transmission, or video.

In the catalog of terrestrial or land mobile systems [2], two of the systems hold a prominent position if we take into account their rapid growth and commercialization: the cellular system and the cordless phones system. The main difference between cellular and cordless is that in a cellular system, the providing service is very similar to those of a normal telephony service with the advantage that the user can be anywhere and in motion, whereas the cordless service is simply a wireless telephone link to your home or your office. There is also a small personal base station that makes it possible to walk around with a direct link to this base station, but there is no provision for taking the cordless phone far away or to another city. Land or terrestrial mobile communications technology has come a long way since the pioneering work of AT&T Bell Laboratories researchers, around 1970, which lead to the first operational cellular system, in 1983. This early analog system, now described as *first-generation*, was relatively simple and developed to suit local requirements, became popular and demonstrated the benefits of communication on the move. The *second-generation* system (1990), designed to use digital transmission, has shown advantages of not only digital over analog processing but also the future of the global standardization. Nowadays third-generation mobile communication systems provide users with advanced communications services having wideband capabilities and using a single worldwide standard.

Mobile satellite communications began in 1976 with satellites in geostationary orbit to provide communications services to ships at sea, and later to aircraft and land-based terminals. The rapid growth of land-mobile communications systems resulted in more efforts for providing global mobile communications services through the use of mobile satellite communication systems in low, medium, and geostationary earth orbits. Second-generation terrestrial and satellite mobile communications systems have existed as two independent environments. However, these two environments are now being combined toward a third-generation global mobile communications system in which the two systems play complementary rather than independent roles, forming a single universal integrated system.

Today, the complexity facing the mobile antenna designer is compounded by the awareness that with clever design the antenna can improve system performance by embodying additional functions, such as diversity reception capability, multipath fading immunity, polarization characteristics selectivity, and matching to different environmental and operational conditions [3]. Modern antenna design for mobile services is no longer confined to small, lightweight, low-profile, or flush-mounted omnidirectional radiators on a well-defined perfect flat ground plane. It is rather the creation of a complicated electromagnetic device that may incorporate active elements and signal processing schemes while operating in a constantly varying time-varying environment.

Antennas now have also to be seen as an integral part of the overall system design, since the equipment onto which an antenna element is mounted can act as a radiator, so that the antenna element and the body of the equipment must be treated together as an antenna system. Another factor is the proximity effects caused by obstacles near to the antenna element, which affect antenna performance and must be considered in the design. The operator of a portable mobile unit can seriously perturb antenna performance, while the human hazard problem is another factor, which always must be of concern.

Mobile communication systems can be divided into two main categories: (1) terrestrial or land-mobile systems and (2) satellite systems. Accordingly, antennas for mobile communication systems can be considered to be part of one of these two main systems, although in many cases there are handsets that were designed to work in both systems.

2. ANTENNAS FOR LAND-MOBILE COMMUNICATION SYSTEMS

2.1. Design Considerations

The two main parts of a land-mobile communication system are the base stations and the mobile stations or units (Fig. 1). Since these stations employ different types of antennas, the key design items are also different. Although *antenna design* usually refers to electrical design, some other aspects must also be considered. Construction costs, easy manufacturing, low installation fees, as well as, key mechanical items such as wind load and seismic load design are important criteria that must be of prime concern. In practice, as a first step, an evaluation that compares electrical and mechanical characteristics and the tradeoff between performance and cost is done, followed in the second step by the determination of the electrical and mechanical design. Also, because these antennas will

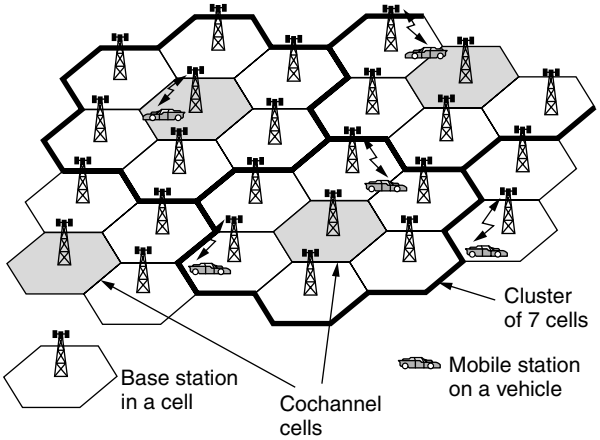


Figure 1. Illustration of a cellular system. Every cell, shown as a hexagon, contains a base station. The cells are organized in clusters of seven cells.

be used in a multipath environment instead of free space, their radiation patterns and gains must be designed for their operating environment. This means that it is useless for an antenna to have superb performance in free space or in an anechoic chamber if it cannot operate well in a real multipath environment. The mobile station antennas, which are classified into two categories, antennas for mobile mounting (on vehicles) and antennas for mounting on portable handsets, are designed for easy handling and subscriber convenience. Besides their electrical characteristics, they must have small volume and weight.

The frequencies used for land-mobile communication systems range from <200 MHz to >60 GHz. The most significant bands are the frequency ranges for analog and digital cellular radio systems from $\sim$800–1000 MHz and 1700–2200 MHz and for the wireless local-area network (WLAN) systems at around 2.4–2.5, 5.1–5.8, and 17 GHz. The bandwidths of cellular systems, which use frequency-domain duplexing (FDD) with two distinct frequency bands for each system, range from $\sim$8–17%, while those of WLAN systems have values of $\sim$5% [4].

2.2. Base-Station Antennas

2.2.1. Basic Requirements. A cellular system services a terrestrial area by dividing it into a number of cells, each containing a base station. The cells are organized in clusters, and every cluster uses all the available channels (Fig. 1). The use of many clusters in an area means an increased number of channels, and this depends on the number of cells per cluster and the area of every cell [5]. Since the same frequencies are used in every cluster, distributed in an ingenious way in its cells in order to reduce cochannel interference, every base station must communicate only with the mobile stations located in its service area, and its radiowave energy must be uniformly radiated inside this area. For this purpose, the radiation pattern and the output power of the base station antenna must be carefully adjusted. For efficient frequency reuse, the uniform illumination of a cell and the suppression of radiation outside of it are accomplished by the use of main beam tilting downward in the vertical plane (see Fig. 2). There are two methods for achieving beam tilting — one is a mechanical beam tilting by leaning the antenna, and the other is an electrical beam tilting by adjusting the relative phases of the antenna elements. The use of a shaped-beam array antenna whose sidelobes, facing the interference cell directions, are suppressed to very low levels assures that interference is kept to minimum possible level. The narrowing of the antenna beam in the vertical plane offers an additional advantage of increased gain [3]. Since the coverage area of each base-station antenna is given, antenna gain cannot be increased by narrowing the beam in the horizontal plane, but only in the vertical plane. So the design parameters for the base-station antenna are the vertical plane pattern shape, which is achieved by an array configuration, and the horizontal plane pattern, which is set by the correct choice of antenna element.

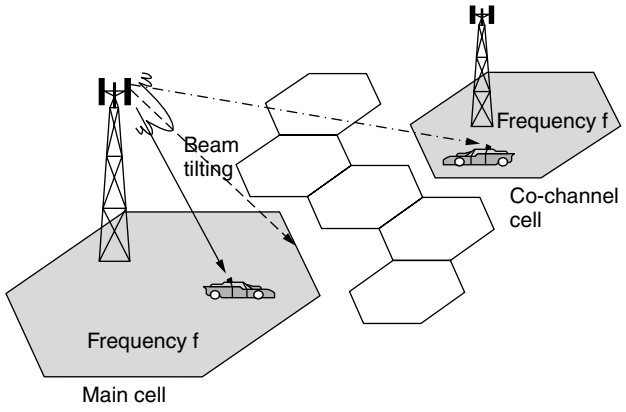


Figure 2. Beam tilt effect for reducing the frequency reuse distance in a cellular system.

Another requirement for the base-station antenna, in order to communicate in all the available channels with the mobile stations, is to be wideband and have a function for branching and combining the channels. Sometimes, because the antenna is shared by several systems, a wider frequency bandwidth is required, and in such cases dual-band or triple-band antennas are used. The multichannel function of a base-station antenna is assured by the use of a broadband antenna element.

As it is well known, communication between base-stations and mobiles antennas rarely occurs within line of sight of each other. A radio transmission link is established in a mobile channel by *multipath propagation*, in which surrounding objects reflect and scatter the transmitted energy causing several waves to arrive at the receiver via different routes. Since these waves have different phases as a result of small pathlength differences between rays coming from scatterers in the near vicinity, or significant time differences if they come from strong scatterers, narrow- or wideband fast fading is produced, respectively. Fast fading is added to slow fading or shadowing, which results from the varying nature of the particular terrain and with the path loss, which is caused by the spreading of waves in space, together constitute the mobile channel fading behavior. As a result, fading occurs constantly at the base and mobile stations and the receiving signal level may fluctuate by $\leq$40–50 dB. In order to keep constant the receiving level and reduce the delay spread, *diversity reception* is used, a technique whose effectiveness has been proved both experimentally and theoretically. It was shown that by placing two antennas $\sim$10 wavelengths apart, a reduction in fading could be achieved, and the value of this reduction depends basically on the correlation between the two antennas [3]. From the three configurations of diversity antennas — space, pattern, and polarization diversity — space diversity is the most widely used [2,6]. In a multipath environment, with Rayleigh distribution, the relationship between the correlation coefficient of the diversity terminals and the carrier-to-noise-level ratio (CNR), with a cumulative probability of 1%, shows that the improvement of CNR does not fall below 8 dB even if the value of correlation coefficient is as high as 0.6 [3]. That means that there is no need to design a diversity antenna with a lower correlation coefficient value. In order to achieve a correlation coefficient of $\leq$0.6 dB, greater distances between antennas for suburban or rural areas are usually required than in urban areas.

Also an increase in antenna height results an increase of the correlation coefficient value.

Finally, lightweight antennas that occupy a small space and have a low wind load are constructed by proper mechanical design. Another subject that must be of concern is the correct choice of materials in order to eliminate *passive intermodulation*, which can arise when the antenna is used for both transmission and reception [7]. According to this phenomenon, as a result of the nonlinear effect of the metal heterojunctions that exist between the antenna elements and the feedline, intermodulation of the transmitting channels can occur, resulting the presence of interference waves having the same frequency of the receiving waves. The use of printed instead of wire elements, the increased contact area between flanges and printed elements, good welding, and the use of measures to prevent oxide generation are some of the common methods used to suppress passive intermodulation.

2.2.2. Types of Base-Station Antennas. The correct choice of a base-station antenna depends on the size of the service area. For cellular systems, a common technique for better frequency reuse is the division of the cell area in zone sectors and the use of sector-beam antennas. It can be shown that in a cellular system, the frequency reuse distance, with a sector zone arrangement, is shorter than that of a circular zone arrangement illuminated by omnidirectional antennas [3]. For a service area limited within a restricted angle in the horizontal plane, a corner reflector antenna is often used either as a single radiator or as an element of a linear array antenna. A corner reflector antenna has the advantage of adjusting its beamwidth by controlling the aperture angle of the reflector. A reflector antenna with its parameters is shown in Fig. 3. Sector beams with beamwidths of 60–180° can be realized by adjusting the aperture angle from 60° to 270°. It should be noted that for the base-station antenna, it is more important to operate with large ratio of desired to undesired signal ratio than to have a high gain value. Fortunately, beam tilting, which is essential for frequency reuse, works in this direction, and if it is combined with the suppression of the sidelobes adjacent to the main beam, by appropriate array antenna pattern synthesis, a very effective antenna is produced. Another point to be noted is that in order to reduce the interference, only several sidelobes near the

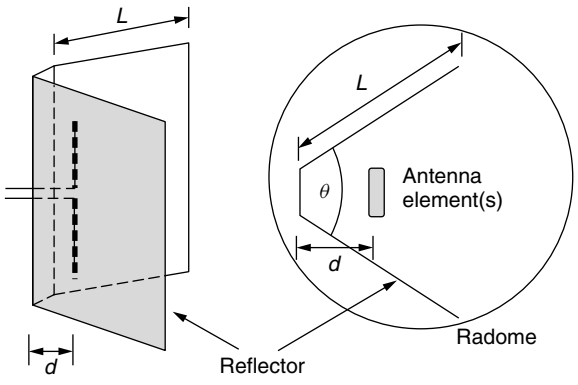


Figure 3. The geometry of a corner reflector antenna.

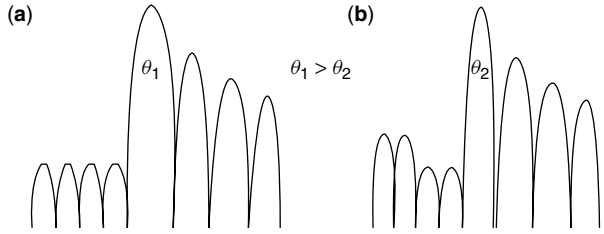


Figure 4. From a uniformly suppressed sidelobe pattern (**a**) a new one is received with the main beam 30% narrower (**b**) by suppressing only several sidelobes near the main beam and setting the others at a higher level.

main beam must be suppressed. In Fig. 4, from a uniformly excited array a new array with a main beam 30% narrower is produced by suppressing only several sidelobes near the main beam and setting the other sidelobes at a higher level [3].

Some basic types of base-station antennas that are used in many systems are the dual-frequency antenna, the dual-beam antenna, and the polarization diversity antenna. In third-generation cellular systems, digital beamforming or smart antennas are used.

Today, several terms are used to refer to the various aspects of smart-antenna system technology, including intelligent antennas, phased array, space-division multiple access (SDMA), spatial processing, digital beamforming, and adaptive antenna systems. A "smart antenna" consists of an antenna array combined with signal processing in both space and time. Spatial processing leads to more degrees of freedom in the system design, which can help improve the overall performance of the system. Smart-antenna systems are usually categorized as either switched-beam or adaptive array systems [8]. Although both systems attempt to increase gain in the direction of the user, only the adaptive array system offers optimal gain while simultaneously identifying, tracking, and minimizing interfering signals. It is the adaptive system's active interference capability that offers substantial performance advantages and flexibility over the more passive switched-beam approach. Smart antennas communicate directionally by forming specific antenna beam patterns. They direct their mainlobe, with increased gain, in the direction of the user, and nulls in directions away from the mainlobe. Different switched beam and adaptive smart antennas control the lobes and the nulls with varying degrees of accuracy and flexibility.

The traditional switched-beam method is considered as an extension of the current cellular sectorization scheme in which a typical sectorized cell site is composed of three 120° macrosectors. The switched-beam approach further subdivides the macrosectors into several microsectors. Each microsector contains a predetermined fixed beam pattern, with the greatest gain placed in the center of the beam. Typically, the switched-beam system establishes certain choices of beam patterns before deployment and selects from one of several choices during operation (Fig. 5). When a mobile user is in the vicinity of a macrosector, the switched-beam system selects the microsector containing the strongest signal. During the call, the system monitors the signal strength and switches to other fixed

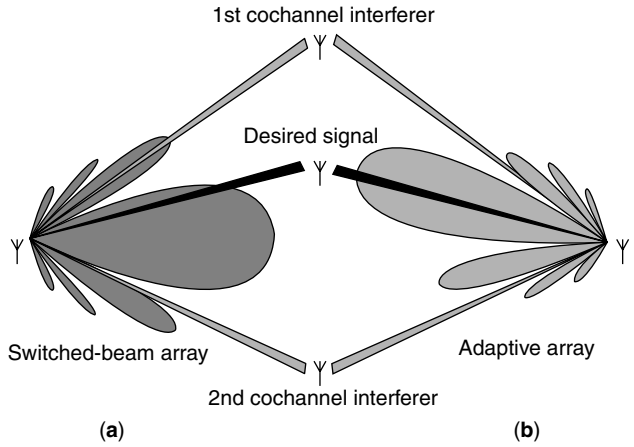


Figure 5. Beamforming lobes and nulls that switched beam (**a**) and adaptive array (**b**) systems might choose for identical user signals and cochannel interferers.

microsectors if required. All switched-beam systems offer similar benefits even though the different systems utilize different hardware and software designs. Compared to conventional sectored cells, switched-beam systems can increase the range of a base station from 20 to 200% depending on the circumstances of operation [8,9]. The additional coverage means that an operator can achieve substantial reduction in infrastructure costs.

There are, however, limitations to switched-beam systems. Since the beams are predetermined, the signal strength varies as the user moves through the sector. As a mobile unit approaches the far azimuth edges of a beam, the signal strength degrades rapidly before the user is switched to another microsector. Moreover, a switched-beam system does not distinguish between a desired signal and interfering ones. Thus, if an interfering signal is around the center of the selected beam and the user is away from the center, degradation in the quality of the signal for the mobile user occurs.

Adaptive antennas take a very different approach. By adjusting to an RF environment as it changes, adaptive antenna technology can dynamically alter the signal patterns to optimize the performance of the wireless system (Fig. 5). The adaptive antenna utilizes sophisticated signal processing algorithms to continuously distinguish between desired signals and multipath and interfering signals, and also, calculate their directions of arrival [8,9]. A block diagram of an adaptive antenna system is shown in Fig. 6. The adaptive approach continuously updates its beam pattern according to changes in both the desired and interfering signal locations. The ability to smoothly track users with mainlobes and interferers with null ensures that the link budget is constantly maximized.

2.3. Mobile-Station Antennas

2.3.1. Design Considerations. Mobile-station antennas or handheld antennas had to follow the dramatic decrease in size and weight of portable phones while maintaining the same antenna performance in terms of radiation pattern, gain, and bandwidth. These changes have necessitated a rapid evolution of antenna structures and techniques for

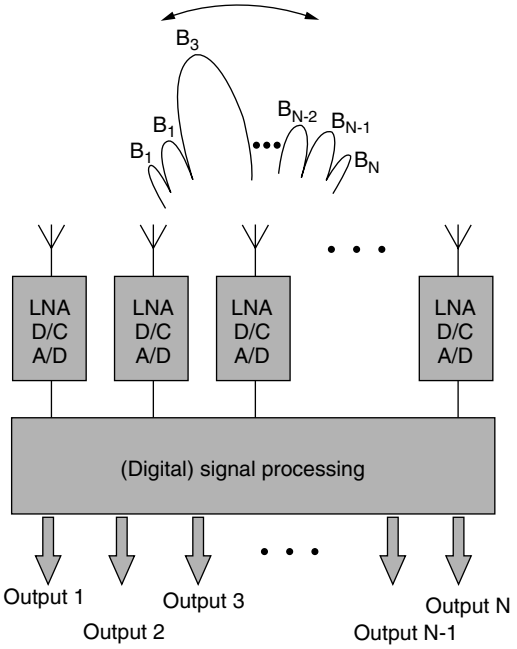


Figure 6. A block diagram of an adaptive antenna system (LNA — low-noise amplifier; D/C — downconverter; A/D — analog–digital converter).

the mobile stations or portable phones. With a volume of <100 cm^3 for a typical mobile phone today, it is difficult to achieve the necessary bandwidth of 10% without inducing currents on the handheld unit, which leads the antenna element to act more as a coupling structure than a radiating element. We can separate the antennas here according to their application: for vehicular and for portable phones. For the first category (vehicular), since the typical mobile user moves randomly in a radio cell, an omnidirectional azimuth pattern is required. Particularly in suburban or rural areas it is common for the mobile station and the base station to be in line of sight, so the omnidirectional pattern assures that the received signal level will not vary considerably. For the second category, although the requirements for operating frequency zones and bandwidth are the same, the transmitting power of portable phones must be less, because of the limited battery capacity and the size limitations. The achievable gain is also generally less than that possible in vehicular antennas, since only small antennas can be used for which adequate bandwidth and high efficiency are very difficult to exist simultaneously [11,12]. Also, their proximity to the human body is another factor of gain degradation. For these reasons, the main antenna requirement for mobile phones is to develop the highest possible gain over the required bandwidth.

It should be added here that the gain performance of a mobile antenna operating in its practical environment is different from that of the same antenna in isolation. For this reason the concept of the mean effective gain (MEG) in a multipath mobile radio environment is introduced. The *mean effective gain* of an antenna is defined as the ratio between the power that the mobile antenna actually receives to the total power available. All power values are

considered averages taken after the mobile station has moved along a path of several wavelengths, and the total average power incident on the mobile is composed of both horizontally and vertically polarized components [3,6].

Among the distinctive characteristics of antennas for portable phones is the fact that their radiation pattern and the polarization direction cannot be considered as fixed, because of the random direction of the phones when used and their proximity to the user's body, which absorbs and scatters the electromagnetic energy. This makes it very impractical for antennas of portable phones to be characterized by omnidirectional radiation patterns and by vertical polarization.

In conclusion, the design considerations to be considered are the relatively large bandwidth ($\sim$10%), the need for a uniform coverage over the azimuthal plane with a high mean effective gain value, and the operating frequency bands if the antenna is mounted on a multiple-band phone, while the small size is more than essential. Under these conditions a wide variety of antenna structures, such as the monopole antenna in different forms, the helical antenna, and the planar inverted-F antenna (PIFA), as well as the microstrip patch, meander line and chip-type built-in antennas, are used in the phone technology. The majority of portable phones use a monopole antenna as the main element and a PIFA as the subelement forming a two-element diversity antenna, while other combinations are a monopole antenna and a helical or a meander line. These antennas are described here.

2.3.2. Types of Practical Antennas for Mobile Stations. Cellular phone antennas were developed initially from communication radio antennas used at lower frequencies. A *monopole antenna* is a quarter-wave whip antenna, which was the original type from which by a distributed inductive loading a significantly shortened antenna was produced (Fig. 7). An antenna of this type, known as a "rubber duck" on communication radios, can be accomplished by using a spiral enclosed in plastic or rubber having a total length of 5–15% of a wavelength [1,3]. Electrically, these antennas are still quarter-wave whips, which are tuned to a shorter length by distributed inductive loading, which is embodied by a helically wound wire. Generally the loading, both inductive and capacitive, can be added to a basic monopole type antenna to obtain better operating characteristics with a reduced physical height by maintaining a more constant current distribution for larger field strength or effective height and improved impedance matching. For all these antennas, which are basically quarter-wave elements, the cavity of the phone and, partially, also the user, are very important parts of overall antenna function. These antennas can be considered electrically as asymmetrically fed half-wave elements, with the feeding point located at the point where the antenna element enters the phone. At the commonly used frequencies, around 900 MHz, the body of the phone is of the order of half wavelength, and a short stubby antenna element is widely used. This is an essential way to extend the electrical size of the antenna when a low frequency is used. The evident disadvantage is that the main antenna current is flowing through the phone case and consequently also through the user, causing losses and probable undesirable medical effects at high powers levels. In order to eliminate these effects, in another approach half-wave whips having high feeding impedance can be used, resulting in low currents along the phone case. Half-wave whip antennas can also be made shorter than half wavelength if an inductive load is used, taking care that sufficient bandwidth remains assured. For easy transportation, the still rather long whip can be retracted into the phone case. It has to be noted that with the widespread use of cellular phones and the continuing improvement of cellular networks, more emphasis is presently being placed on a handy design rather than on maximum performance.

Helical antennas are usually constructed by a wire helically wound around a dielectric core as shown in Fig. 8. This type of antenna has an axial mode of radiation and a normal mode of radiation that is perpendicular to the axis of the helix. Although the axial-mode helix has been widely used as endfire directional elements, resonant normal-mode helical Antennas are useful as short, vertically

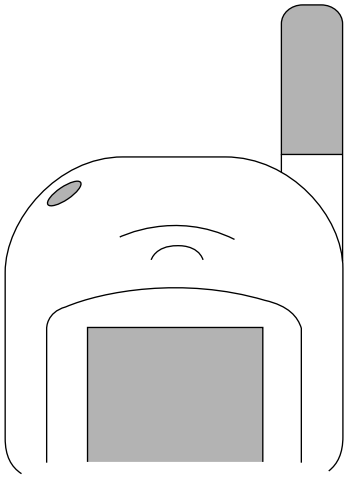


Figure 7. A reduced-length quarter-wave whip antenna by inductive loading.

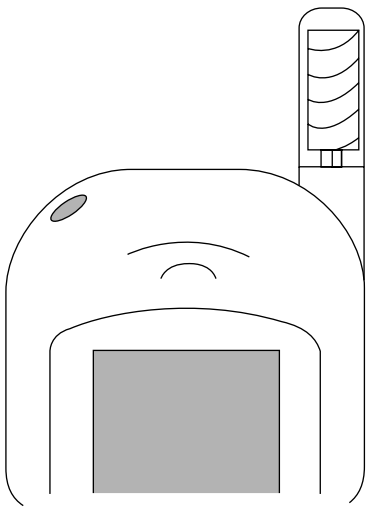


Figure 8. A normal-mode helical antenna.

polarized radiators, similar to the monopole. Using helical antennas, combined with inductive and dielectric loading, a considerable decrease in wave velocity can be achieved, which manifests a behavior similar to that of a quarter-wavelength antenna but with a physical length reduced to only 6–13% of a wavelength [1,3]. Another advantage is that the radiation resistance depends mainly on the exterior physical length and only slightly on the kind of helical winding. Also, the normal-mode helical antenna is more wideband in comparison to a straight monopole of the same length tuned by an internal inductance, because the axial current distribution gives up to 2.5 times higher radiation resistance. For cellular systems operating at 900 MHz, a normal-mode helical antenna has a length equal to 20–40 mm, with the minimum value determined by the necessary bandwidth (8–10%) [3].

A classical helical antenna with good performance, regardless of the grip and orientation of the phone, is shown in Fig. 9, where the helix configuration is on the bottom when the extendable whip is fully taken out. This antenna category is known as *helical Antennas with and without a whip* and can be developed for dual-band applications [3]. Since the whip has a nonmetallic top, when the whip is fully retracted the operation is similar to that of a fixed normal-mode helical antenna. When the whip is fully extended, it is connected in parallel to the helix, via a metallic connection that exists in its bottom. In this case, a part of the whip passes through the helix and detunes it so only the whip itself is fed. During this operation, the higher-frequency bands are slightly less sensitive to the influence of the user, and this can be explained by the fact that the distance from the skin of the user, expressed in wavelengths, is different. Moreover, a slight asymmetry in the azimuthal plane away from the head has been observed, which is desirable since it tends to decrease the losses in the head.

Instead of the classical helical conductor inside a normal-mode helical antenna, a printed meander pattern can be used. This kind of antennas is known as *meander patch antennas* [3,9]. Because of the easy fabrication process, virtually any meander shape can be obtained at low cost, and this advantage can be utilized for the creation of multiband operation and matching networks. The meander pattern can be printed on a flexible board, which then is rolled on a core. Because of the many possibilities for controlling the pattern, better optimization can be obtained than with a helical wire antenna, particularly for multiband functions. The different shapes that can be used as meander patterns include fractal patterns and patterns generated by genetic algorithms. For all these cases the added inductance is the important factor that guarantees the use of the small antenna at the lowest-frequency band, while the special shape of the pattern assures good multiband performance. A simple way to accomplish multiband operation is to connect two or more quarter-wavelength meanders in parallel, each tuned to its own desired frequency. The use of two or three frequency bands has become very common, typically with GSM 900, 1800, and 1900 MHz, allowing dual-band operation in GSM countries as well as in PCS (1900 MHz) areas in the United States. Because radiation resistance increases with the frequency quadratically, the result is that not only the absolute but also the relative bandwidth at higher frequencies will generally be higher. Thus for GSM, both systems at 1800 and 1900 MHz may very well be fitted within the higher band, while at 900 MHz a problem is eliminated if the stubby antenna is short. A dual-band meander antenna corresponding to a fixed normal-mode helical antenna is shown in Fig. 10.

In many new GSM phones there are *built-in antennas*; this means that the antennas are not visible from the exterior of the phone. Since at low-frequency bands (GSM 900 MHz) the antenna element plays the role of a kind of feed structure inducing currents on the phone body, which is actually the main radiator source, there

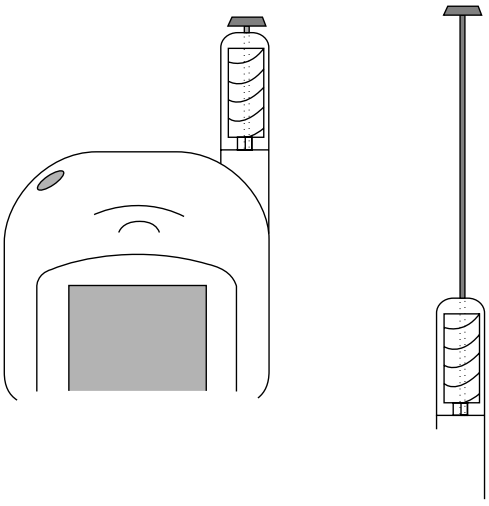


Figure 9. A helical antenna with a whip for dual-band applications.

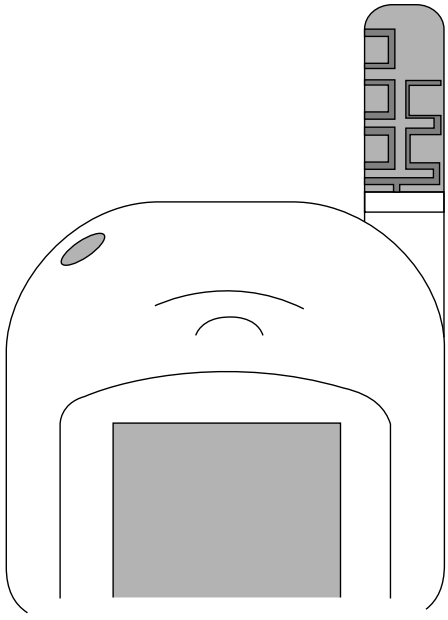


Figure 10. A meander patch antenna. The meander pattern has been printed on a flexible board and rolled on a core.

is really very little difference in whether the antenna elements are actually visible. At higher-frequency bands (GSM 1800/1900 MHz) the antenna element mainly operates independently from the phone body, but generally the same antenna element is used for both lower- and higher-frequency bands.

The radiation properties and the near field around the antenna elements are dependent mainly on the surface of the antenna element, while for the bandwidth and losses the most important characteristic is the volume occupied by the internal field. Usually, for installation, a surface on the upper back of the phone is preferred because the field very near to the antenna element should be kept away from the user to avoid unnecessary losses there. Also the antenna element should not be located too low on the back of the phone, as such a location could also result in increased losses in the hand of the user. The typical *planar inverted-F antenna* (PIFA) has been studied extensively because of its advantages, its low profile, and the easy incorporation into phone units [10]. It can be considered as a quarter-wave stub, which, when viewed from the open end, has a real admittance at quarter-wave resonance (Fig. 11). This L or inverted-L antenna can be considered as a variation of the monopole antenna where the monopole is bent over in an L shape with respect to a ground plane. The basic monopole structure is modified to reduce the height of the antenna while obtaining a lower resonant frequency than that of a comparable electrically short monopole. The short arm of the L antenna radiates in an omnidirectional pattern in the plane perpendicular to the vertical element, and some radiation is also obtained from the long arm of the L antenna.

The disadvantage of PIFA, which is its narrow bandwidth (1–2% in the relative bandwidth in free space) can be corrected when it is mounted on a finite ground plane. In practice, the bandwidth of an antenna system, which uses a built-in PIFA element in a phone, can be designed to have wider bandwidth due to the attribute from the phone body, which acts as a part of the radiator due to the current flow excited by the PIFA element. The physical length, which corresponds to a quarter-wavelength, can become shorter either by a capacitive loading (high ε) or by adding inductance (meandering, etc.). Thus, the characteristic impedance can be increased or decreased by inductive or capacitive loading, respectively, but the radiating conductance will remain the same. As a consequence, the bandwidth will be much wider for the inductively loaded case (meander, etc.) compared to the case with high ε. As already mentioned, volume will again be the critical parameter, and in both cases bandwidth will be smaller in comparison to that of the full-size PIFA. Regarding the implementation with meander patches, there is a limitation in that losses will increase if lines that are too narrow are used. The built-in antenna can also implement an optional whip, which can help in decreasing the influence of the hand of the user, combining the advantages of the built-in antenna with the performance of the extended whip.

It is worth mentioning that in the personal digital cellular system (PDC) used in Japan, mobile phones incorporate a system of a whip and a built-in PIFA antennas, with improved reception quality due to the diversity reception scheme. As can be seen in Fig. 12, the monopole element (whip), which has a length of either $\frac{3}{8}$ or $\frac{5}{8}$ wavelength in order to minimize the current flowing on the handset, has a normal-mode helical antenna on its top, which is encapsulated in a plastic cover. Because the two antenna elements are placed very closely, with a separation distance of only $\sim$0.1–0.2 wavelengths on the ground plane of the phone, the mutual coupling between them may degrade the radiation efficiency and also diversity function. Optimum values have been shown for the length of the whip element and the length of the phone unit in order to achieve low correlation coefficient without causing much degradation in radiation efficiency. The analyses have shown also that the diversity antenna performance depends not only on the correlation coefficient but also on the mean effective gain (MEG) of the antenna system.

Another interesting element is the *chip antenna*, which is a small normal-mode helical element molded in a ceramic chip, where ceramic materials, having a relative permeability of $\geq$20, are used. In order to assure wider bandwidth operation, a matching circuit or a PIN diode

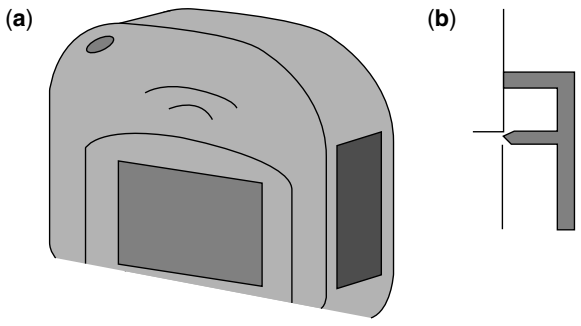


Figure 11. A PIFA antenna mounted on a side of a handset (**a**); side view of a PIFA (**b**).

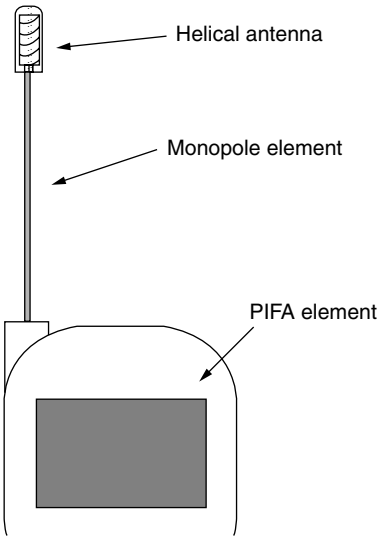


Figure 12. A handset of PDC with a PIFA element on its back, and a whip antenna with a normal-mode helical antenna on top.

circuit, which switches frequency to cover the necessary bandwidth, can be used.

3. ANTENNAS FOR MOBILE SATELLITE COMMUNICATION SYSTEMS

3.1. Introduction

In many cases terrestrial cellular communication systems cannot provide complete coverage over large global rural regions, due to the inability of installing base stations. A satellite-based system can fulfill this need by using either a few fixed geostationary satellites or a large number of low or medium earth-orbiting satellites.

The concept of artificial satellite was introduced in 1945 by A. C. Clarke, and since then more than 1000 satellites are in the geostationary or Clarke orbit, with more added at frequent intervals. Nearly all communication satellites are in geostationary earth orbit (GEO), for which a satellite appears stationary with respect to an observer on the ground because of its velocity match with that of the earth surface. Among the many existing GEO satellite systems, probably the best known is the *International Maritime Satellite System* (INMARSAT), which has provided international maritime satellite communication services since 1982, and is expanding its services to aircraft and land mobiles [9]. Many other systems such as AMSC in the United States, MSAT in many countries, AUSSAT in Australia, and ACES in Asia are using dedicated satellites to provide domestic satellite communication services mainly for land mobiles, such as voice and low-speed data.

Medium- and low-earth-orbiting satellite systems (MEOs and LEOs) have been introduced and use groups of low-altitude orbiting satellites (up to ~10,000 and ~1000 km for MEO and LEO, respectively). Because of the reduced distance between transmitting and receiving sites, as compared to those of GEO satellites systems (Fig. 13), less power and low-gain omnidirectional antennas can be used, and signal delay is also reduced. Typical examples of MEO/LEO satellite systems are Iridium (1.65 GHz, 66 satellites at 780 km), Globalstar (1.6/2.5 GHz, 48 satellites at 1414 km), Teledesic (19/29 GHz, 288 satellites at 1400 km), Ellipso, and ECCO [9].

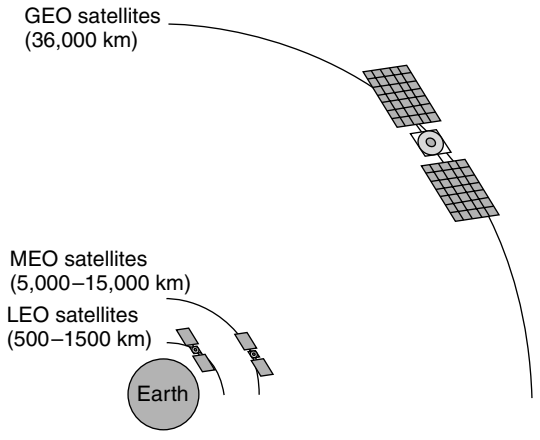


Figure 13. GEO, MEO, and LEO satellite orbits.

3.2. Antenna Design Considerations

Antennas play a significant role in the development and operation of satellite communications. They form the input and output ports to the satellite communication system, which can be divided generally in two segments: the ground and space segments. In every system, the signal is beamed into space by an uplink antenna and after the appropriate processing onboard the satellite is sent back to earth using a downlink antenna to be received by the earth-station antenna.

The types of antennas used depend on a number of factors related mainly to the distance between the satellite and the earth. Thus, for GEO satellite systems, because of the large distance (36,000 km), both satellite and ground-station antennas are characterized by high directive gain in order to overcome space loss, which translates into the demand for large aperture, pencil-beam antennas. Since earth-station antennas look upward at the sky, ground reflections are eliminated but the presence of the ground remains through its effect on the antenna noise temperature via sidelobe and backlobes. In all systems using GEO satellites, the L band is used (1.6/1.5 GHz) and the required frequency bandwidth to cover transmitting and receiving channels is ~8%. As a result, in using a narrowband antenna element such as a patch antenna, efforts have to be made to produce a wider bandwidth. The required gain is calculated by a link budget analysis taking into account the required channel quality and the satellite capability. Although gain is an essential parameter in antennas, the figure of merit G/T, which is the ratio of gain to system noise temperature, is a more commonly used factor in satellite communications. Since the antenna beam must cover $0-90^\circ$ in elevation and $0-360^\circ$ in azimuth directions, a tracking capability generally is needed, while in order to eliminate the need for polarization tracking, circular polarized waves are used.

Radiowave propagation over earth–space links for maritime mobile satellite systems lead to problems substantially different from those arising in the fixed satellite service. Thus, the effects of reflections and scattering by the sea surface become quite severe, especially in case of antennas with wide beamwidths, signal level attenuation due to blocking by the ship superstructure is not negligible, and the effect of ionospheric scintillation for L-band frequencies must also be taken into account. Multipath fading due to sea reflection must also be considered. The reflected waves are composed of a coherent (specular reflection) component and an incoherent (diffuse) component that fluctuates due to the motion of the sea waves [3].

The requirement for high-gain antennas for satellite systems can be relaxed using MEO or LEO satellites. The reduced distance also decreases signal delay, but because the satellites become nonstationary with respect to the ground surface, more complicated communications links are required. Shadowing by buildings is more severe for these systems than the GEO ones and generates polarization-sensitive reflection and diffraction that leads to multipath effects. Also, Doppler shifts (36 kHz–1.6 GHz and 55 kHz–2.5 GHz) have to be corrected with due regard to the relative direction of flight [3].

In implementing mobile satellite communications, of significant importance is the vehicle antenna. For an antenna system to be mounted on a mobile station, for satellite communications, it must be compact and lightweight. Additionally it must be easy to install and ensure mechanical strength. The installation requirement is not so severe for shipborne antennas, because even in small ships there is the space necessary to install an antenna system. However, in the case of automobiles, especially for small, private cars, low-profile and lightweight equipment is an essential requirement. The same demands generally hold for aircraft, plus the more severe required conditions in order to satisfy the standards for avionics, where one of the most important requirements for an aircraft antenna is low drag.

Antennas for mobile satellite communications can be classified as omnidirectional and directional antennas [3]. In the following sections typical examples of these antennas are described, and the most common types of antennas systems for aeronautical mobile communications are also presented.

3.2.1. Omnidirectional Antennas for Mobile Satellite Systems. In the category of omnidirectional antennas for mobile satellite systems, the most common ones are the quadrifiliar helical, the crossed–drooping dipole and printed antennas, and the microstrip patch and cavity-backed cross-slot antennas. These antennas, which are also used as elements in directional array antennas, are attractive because they have small size, are lightweight, and operate with circular polarization. Also, since their gain in the L band is 0–4 dBi, they do not require satellite tracking.

The *quadrifiliar helical antenna* (QHA) (Fig. 14) is composed of four identical helixes wound, equally spaced, on a cylindrical surface. The helixes are fed with signals equal in amplitude and 0°, 90°, 180°, and 270° in relative phase. The QHA has a height equal to 40 cm and presents a gain with a minimum value equal to −4 dBi, and axial ratio of ≤3 dB [3]. The *crossed-drooping dipole antenna* is the best choice for land-mobile satellite systems, when the required angular coverage must be narrow in elevation and almost constant in azimuth. The variation of the separation distance between the dipole elements and the ground plane adjusts the elevation pattern to ensure an optimum pattern for the coverage region of interest

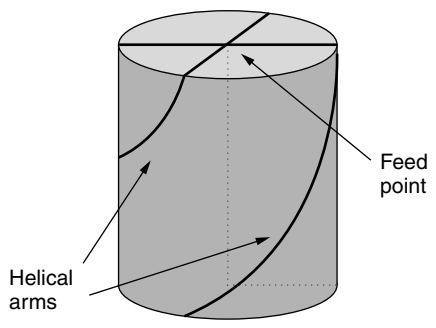


Figure 14. The quadrifiliar helical antenna.

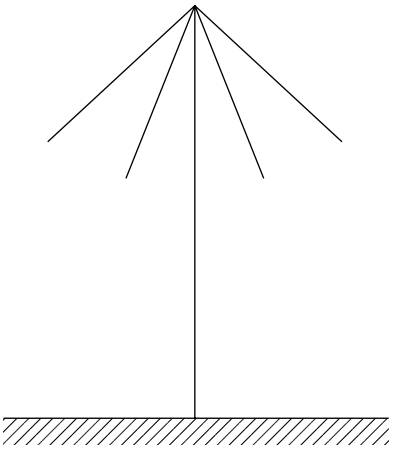


Figure 15. The crossed–drooping dipole antenna.

(Fig. 15). The crossed–drooping dipole antenna is characterized by a minimum gain equal to 4 dBi and a maximum axial ratio value equal to 6 dB for a height of 15 cm [3].

The *microstrip patch antenna* (MSA) [10,13,14] incorporates a circular metallic disk on a dielectric grounded substrate, and in order to produce circularly polarized waves, it is excited at two points orthogonal to each other with signals equal in amplitude and 90° phase difference (Fig. 16). A higher-mode patch antenna can also be designed with a similar radiation pattern to the drooping dipole. To produce conical radiation patterns (null on axis) suitable for land-mobile satellite applications, the antenna is excited at higher-order modes. The circular microstrip patch antenna is characterized by 3.5 dBi minimum gain value and 4 dB maximum axial ratios for a height of 1 cm when RT/Duroid is used as dielectric substrate. Because the available frequency bandwidth of this patch antenna is very narrow, the two-layer patch antenna is used in which the upper and lower parts play a role for transmission and reception, respectively. In a two-layer patch antenna each layer is individually fed at two points with a phase difference of 90° for circular polarization. Another useful form of printed antenna is the *cavity-backed cross-slot antenna* (XSA) [3]. In this case each slot antenna is fed with an equal amplitude and in-phase condition at two points equidistant from the center. One advantage of this antenna is that the input impedance can be matched for a wider frequency band than in the case of the slot antenna. The general

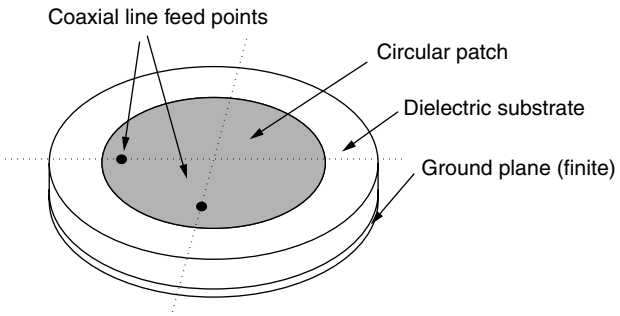


Figure 16. A circular microstrip antenna excited to produce circularly polarized waves.

characteristics of the MSA are antenna gain at boresight $\sim$15.2 dBi and at a scanning angle $\theta = 45^\circ$ the gain is equal to about 13.5 dBi, while for the XSA, the boresight gain is $\sim$15.7 dBi and the gain at $\theta = 45^\circ$ is $\sim$14 dBi [3].

One of the main and most common uses of these antennas is in the Navigation System with Time and Ranging/Global Positioning System (NAVSTAE/GPS), which is the most widely used navigation system at present. Since a GPS satellite transmits two radiofrequencies (1575.42 and 1227.6 MHz), with right-hand circular polarizations and for accurate positioning, both frequencies must be used in order to compensate for the excess delay of radiowaves in the ionosphere, and an antenna operating equally well at both frequencies is needed [3]. From the omnidirectional antennas, QHA and MSAs have been widely used because of their simplicity, small size, and low cost, with slightly modified design parameters to ensure that their radiation patterns are as uniform as possible over the upper hemisphere. Thus, for dual-frequency operation two QHAs are used, one for every frequency, into one structure by coaxially mounting them one into the other or by one on the top of the other. In the case of MSAs, which are extensively used as GPS antennas because of their printed antennas advantages, one effective method of obtaining a broadbeam for GPS reception is the reduction of the size of their ground planes.

3.2.2. Directional Antennas for Mobile Satellite Systems. Directional antennas for mobile satellite communication systems are used in all INMARSAT systems, as well as in PROSAT, ETS-V, MSAT, and MSAT-X research programs for mobile satellite communications. A short description of the basic antennas for the various INMARSAT systems will be given here.

The typical antennas for INMARSAT-A, -B, and -F systems (a description of available INMARSAT systems can be found in Ref. 9) are aperture antennas such as a *parabolic antenna*, which is characterized as a simple structure with high-aperture efficiency. The parabolic antenna can have a gain of 20–23 dBi, a bandwidth of $\sim 10^\circ$, while satellite tracking is required because of ship motions and the small half-power beamwidth. In the case of the INMARSAT-C antenna system, the simplest and most compact configuration is required, that is, without mount systems and tracking/pointing systems. Therefore the antennas used are usually omnidirectional antennas, and the most suitable ones are the antennas described above, namely, the QHA, the cross-drooping dipole, and the MSA. The QHA is the most appropriate for ships because of its good performance of widebeam coverage under the condition of ship motion, while in handheld or briefcase terminals, where a very low profile characteristic is necessary, the obvious choice is the MSA.

Since the INMARSAT-M is used mainly for small ships, such as fishing boats and land vehicles, the kinds of antennas used have to be of small size and low cost. Concerning the efficient utilization of satellite power and the required G/T, antenna gain ranges from 13 to 16 dBi. A *short-backfire antenna* (SBF) is one of the most commonly used M-terminal antennas for maritime applications, especially in the improved form in which the flat-disk

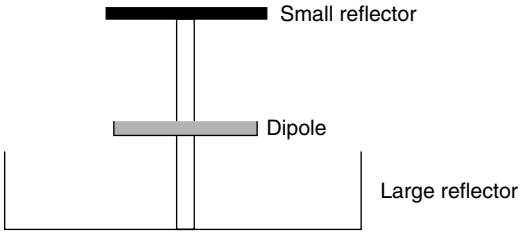


Figure 17. The short-backfire antenna.

main reflector is replaced by a conical or step plate plus a second small reflector in order to achieve better aperture efficiency (80%) and frequency bandwidth (20% instead of 8% of the normal SBF) (Fig. 17). Another antenna suitable for the INMARSAT-M is the *quad helical antenna*, used as a single antenna or as array elements. This antenna, which can be considered as a compromise between the dipole and loop antenna, operates in the so called axial mode, when the helix has a pitch angle of $12-15^\circ$ and the circumference is about one wavelength. The two-turn helical antenna is characterized by very good polarization characteristics for its size and high gain, affected mainly by the size of the reflector. The quad-helical array antenna is composed of four two-turn helical antennas in a square arrangement whose elements are oriented in the manner shown in Fig. 14. This antenna has gain of $\sim$13 dBi (at HPBW = 38°), an axial ratio of $\sim$1.0 dBi, and an aperture efficiency of $\sim$100%. Microstrip and slot antennas can also be used in the INMARSAT-M system in the form of planar arrays. Regarding the INMARSAT-Aero system, phased-array antennas are the best candidates for airborne antennas because of their low profile and mechanical strength. At the present time, two types of phased-array antennas have been used, the conformal type, with the important advantage of the low air drag because of its low profile, and the top-mount type. The first one consists of two sets of phased arrays on both sides of a fuselage, while the second one has a set of phased arrays on the top of the fuselage.

3.2.3. Antenna Systems for Aeronautical Mobile Communications. Today, many different types of antennas are mounted on modern aircraft and serve many individual functions associated with specific avionics systems such as navigation, identification, or radar. These antennas, which are characterized as airborne antennas, must be designed and manufactured in such a way as to satisfy specific electrical requirements under the presence of the stringent environmental and aerodynamic conditions. Thus, airborne antennas must have structures that do not increase the aerodynamic drag and must operate without degrading their basic electrical characteristics under the influence of great scale pressure, temperature, and humidity variations. Moreover, they have to withstand great acceleration differences, static electricity, and lightning. Finally, they must be lightweight, and of small size and low profile. Their individual electrical characteristics are dependent mainly on the function the antenna serves. Generally, these are strongly influenced by the shape and the size of the airframe in relation to the wavelength used. When the wavelengths are significantly

larger than the maximum dimensions of the aircraft, which occurs at low and medium frequencies, the antennas are characterized by low radiation efficiency, which results in a high Q value. In this case, careful matching of the antenna over the necessary frequency band is needed. When the airframe size can be considered larger than the wavelength, for very high and ultrahigh frequencies, the antenna or a part of the airframe can become resonant and in this case more degrees of freedom, regarding the design and the position of installation of the antenna, exist. In any case the influence of the airframe on the radiation pattern of antennas must be considered, since shadowing and reflection on the airframe can result in significant distortions and shape changes of the desired radiation pattern.

Between the various types of antennas installed on an aircraft, those used for satellite systems are the most interesting and complicated. A common antenna element is the microstrip patch antenna, which can be used as a high-gain circularly polarized radiator or as an element to form phased-array and shape-beam antennas [10]. In practical applications, circular and rectangular shapes of the patch radiator are used to achieve circularly polarized patterns. Its main advantages are that it can be made conformal to metallic surfaces; is low-profile, lightweight, and small-size; and can be produced at low cost. Its basic disadvantage of narrow frequency bandwidth, of the order of 2%, can be overcome by either (1) increasing the thickness of the substrate and decreasing its dielectric constant or (2) using stacked patches, electromagnetically coupled together [10,13]. In the first technique, countermeasures must be taken to improve the axial ratio, which degrades with the generation of higher-order modes. The second technique also allows the generation of a dual-frequency antenna system, when transmit and receive frequency bands are well separated, by assigning each frequency band to every patch. In both cases, the degenerate modes for circular polarization can be produced, by giving an appropriate perturbation to the patch dimensions and selecting the suitable feedpoint. This eliminates the necessity of an external circuit.

Other elements that can be used instead of the microstrip patch are the electromagnetically complementary radiators, and crossed-dipole and crossed-slot antennas. Using a pair of orthogonally positioned dipoles or slots fed with equal amplitudes and quadrature phase, circular polarization can be obtained. The two sets differ mainly in terms of the feeding method, which in both cases is quite complicated, while special techniques have been developed to improve the axial ratio off boresight.

Another element is the quadrifiliar helical antenna in the resonant form, which is characterized by small size, no ground-plane requirement, and immunity to effects due to the presence of nearby metal structures.

BIOGRAPHIES

Christos Christodoulou received the B.Sc. degree in physics and math from the American University of Cairo in 1979, and the M.S. and Ph.D. degrees in Electrical Engineering from North Carolina State University, Raleigh, in 1981 and 1985, respectively. He served as a faculty member in the University of Central Florida, Orlando, from 1985 to 1998, where he received numerous teaching and research awards. In 1999, he joined the faculty of the Electrical and Computer Engineering Department of the University of New Mexico, Albuquerque as a Chair. In 1991 he was selected as the AP/MTT Engineer of the year (Orlando Section). He is an IEEE Fellow and a member of URSI (Commission B). He has published over 150 papers in journals and conferences. He also has a book on "neural network applications in electromagnetics." He is, currently, the co-editor for a column on "Wireless Communications" for the IEEE AP Magazine and the associate editor for the IEEE Transactions on Antennas and Propagation. His research interests are in the areas of wireless Communications, modeling of electromagnetic systems, smart antennas, neural network applications in electromagnetics, and reconfigurable/MEMS antennas.

Michael T. Chryssomallis received the Diploma in Electrical Engineering from Democritus University of Thrace, Greece, in 1981. He also received the Ph.D. degree in Electrical Engineering from Democritus University in 1988. In 1982, he joined Democritus University as a Scientific Collaborator up to 1989, then up to 1994 as a Lecturer and up to now as an Assistant Professor.

He worked with the Communications Group (director Prof. P. S. Hall) of the University of Birmingham for the period of Oct. 1997–Jan. 1998, in the areas of Active Antennas, and with the Wireless Group (director Prof. C. G. Christodoulou) of the University of New Mexico for the periods of April to June 2000, and April to July 2002, in the areas of Microstrip Antennas and Arrays, and Smart Antennas.

He is serving as a reviewer for the IEEE Transactions of Antennas and Propagation and has published several journal and conference papers. His current research interests are in the areas of microstrip antennas, RF-Mems, smart antennas and propagation channel characterization. He is member of the IEEE since 1988 and senior member since 2000.

BIBLIOGRAPHY

1. C. A. Balanis, *Antenna Theory, Analysis and Design*, Wiley, New York, 1997, Chap. 14.
2. J. D. Gibson, ed., *The Mobile Communications Handbook*, CRC Press and IEEE Press, 1996, Sec. II.
3. K. Fujimoto and J. R. James, eds., *Mobile Antenna Systems Handbook*, Artech House, Boston, 2001.
4. J. D. Kraus and R. Marhefka, *Antennas*, McGraw-Hill, New York, 2001, Chap. 21.
5. T. S. Rappaport, *Wireless Communications, Principles and Practice*, IEEE Press and Prentice-Hall, New York, 1996, Chap. 2.
6. S. R. Saunders, *Antennas and Propagation for Wireless Communication Systems*, Wiley, Chichester, UK, 1999.
7. R. J. Holbeche, ed., *Aerials and Base Station Design*, IEE Telecommunications Series 14, Peter Peregrinus, 1985, Chap. 4.

8. M. Chryssomallis, Smart Antennas, *IEEE Antennas Propag. Mag.* **42**(3): 129–136 (June 2000).

9. L. C. Godara, ed., *Handbook of Antennas in Wireless Communications*, CRC Press, Boca Raton, FL, 2002.

10. R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, *Microstrip Antenna Design Handbook*, Artech House, Boston, 2001.

11. K. Siwiak, *Radiowave Propagation and Antennas for Personal Communications*, Artech House, Boston, 1998, Chap. 11.

12. K. Fujimoto, A. Henderson, K. Hirasaura, and J. R. James, *Small Antennas*, Research Studies Press, UK, 1987.

13. J. R. James and P. S. Hall, eds., *Handbook of Microstrip Antennas*, Peter Petegrinus, London, 1989, Vols. 1 and 2.

14. D. M. Pozar and D. H. Schaubert, *Microstrip Antennas, The Analysis and Design of Microstrip Antennas and Arrays*, IEEE Press, Piscataway, NJ, 1995.

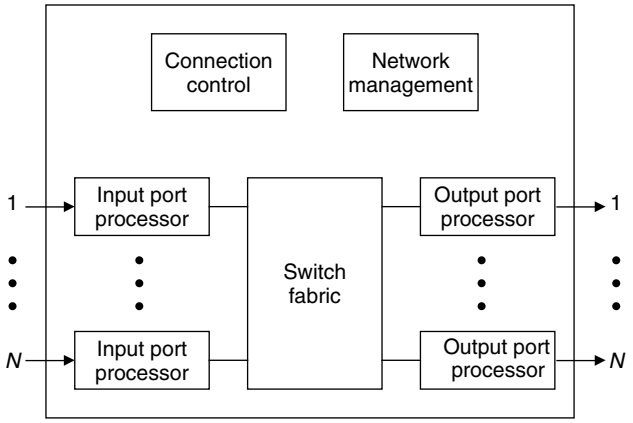


Figure 1. A generic ATM switch architecture.

ATM SWITCHING

THOMAS M. CHEN
Southern Methodist University
Dallas, Texas

STEPHEN S. LIU
Verizon Laboratories
Waltham, Massachusetts

1. INTRODUCTION

ATM (asynchronous transfer mode) is an internationally standardized connection-oriented packet switching protocol designed to support a wide variety of data, voice, and video services in public and private broadband networks [1,2]. ATM networks generally consist of ATM switches interconnected by high-speed transmission links. ATM switches are high-speed packet switches specialized to process and forward ATM cells (packets). Since ATM is a connection-oriented protocol, ATM switches must establish a virtual connection from one of its input ports to an output port before forwarding incoming ATM cells along that virtual connection.

An ATM cell consists of a 5-byte header followed by a 48-byte information field or payload. The main purpose of the ATM cell header is to identify the virtual connection of the cell that occupies most of the header bits. An ATM virtual connection is specified by the combination of a 12-bit virtual path identifier (except the first 4 bits are used for generic flow control at the user-network interface) and a 16-bit virtual channel identifier. Virtual paths are bundles of virtual channels. VP cross-connects are designed to route ATM traffic on the basis of virtual paths only, which is convenient when large amounts of traffic must be routed or rerouted at the same time. The VPI/VCI fields are followed by a 3-bit payload type (PT), 1-bit cell loss priority (CLP), and 8-bit header error control (HEC) field. The PT field is used to distinguish control cells from data cells, and explicit forward congestion indication (EFCI). The CLP flag is used to indicate that lower priority (CLP = 1) cells should be discarded before CLP = 0 cells in the event of congestion. The HEC field allows single bit error correction and multiple bit error detection over the cell header.

A generic ATM switch architecture with N input ports and N output ports is shown in Fig. 1 (note that switches can have any dimensions). The functions of an ATM switching system may be divided broadly into user cell forwarding, connection control, and network management [3]. ATM cells containing user data are received at the input ports, and the input port processors prepare the cells for routing through the switch fabric. The fabric in the center of the switching system provides the interconnections between input port processors and output port processors. The output port processors prepare the outgoing user cells for transmission from the switch. User cell forwarding is characterized by parallelism and high-speed hardware processing. The ATM protocol was intentionally streamlined to allow incoming cells to be processed simultaneously in hardware and routed through the switch fabric in parallel. Thus, ATM switches have been able to realize high-end performance in terms of throughput and cell forwarding delay.

Connection control, sometimes called the *control plane*, refers to the functions related to the establishment and termination of ATM virtual connections. Connection control functions generally encompass: exchange and processing of signaling information, participation in routing protocols, and decisions on admission or rejection of new connection requests.

Network management is currently carried out by SNMP (Simple Network Management Protocol), the standard protocol for managing data networks. ATM switches typically support an SNMP agent and an ATM MIB (management information base). The SNMP agent responds to requests from a network manager to report status and performance data maintained in the MIB. The agent might also send alarms to the network manager when prespecified conditions are detected. Since ATM switches can be viewed as a specific type of network element covered within the SNMP framework, the details of SNMP functions in ATM switches are not discussed in detail here.

Network management should also include standardized ATM-layer OAM (operations and maintenance) functions. ATM switches carry out OAM procedures by generating,

exchanging, processing, and terminating OAM cells. OAM cells are used for fault management, performance management, and possibly other ATM-layer management functions.

2. INPUT AND OUTPUT PORT PROCESSING

The input port processors carry out several important functions. First, the physical layer signal is terminated. For the common case of SONET/SDH (synchronous optical network/synchronous digital hierarchy), the SONET/SDH framing overhead fields are processed, and the payload is extracted from the frame. Individual 53-byte ATM cells are delineated in the payload.

Next, each cell header undergoes a number of processing steps. The cell header is checked for bit errors using the HEC field, and cells with uncorrectable header errors are discarded. The traffic rate of each virtual path or virtual channel is monitored according to an algorithm called the *generic cell rate algorithm* (GCRA), which is essentially a leaky-bucket algorithm. A switch may be configured to allow, discard, or "tag" cells (by setting CLP = 1) exceeding the allowed traffic rate. The VPI/VCI value in each cell header is used to index a routing table to determine the proper output port and outgoing VPI/VCI values. Incoming VPI/VCI values must be translated to outgoing VPI/VCI values by every ATM switch. Cells requiring special handling, such as signaling cells and OAM cells, must be recognized and routed to the appropriate processors in the switch. User cells are prepared for routing through the switch fabric, often by prefixing a routing tag to the cell. The routing tag may consist of the output port, service priority, type of cell, timestamp, or other information for routing and housekeeping purposes. Since the routing tag exists only within the switch, its contents may be chosen entirely by the switch designer. Before entering the switch fabric, cells may be queued in a buffer in the input port processor.

The output port processors have the opposite role of the input port processors, namely, preparing ATM cells for physical transmission from the switch. Cells from the switch fabric may be queued in a buffer in the output port processor, in which case the switch is called an *output-buffered switch*. If routing tags are used, the output port processors remove the routing tag from each user cell. If special cells, such as signaling cells and OAM cells, need to be transmitted, they are inserted into the outgoing cell stream. A new HEC field is calculated and inserted into each cell header. Finally, the ATM cells are transmitted as a physical layer signal. In the case of SONET/SDH, cells are mapped into the payloads of SONET/SDH frames.

3. SWITCH FABRICS

It is often convenient to visualize the basic operation of an $N \times N$ switch synchronized to periodic time intervals equal to the transmission time of one cell, referred to as a "cell time," assuming that the transmission rate on all links are equal. For example, the cell time for a 155-Mbps (megabit-per-second) transmission link would be approximately 53 bytes/155 Mbps = 2.7 μs. In each cell time, a new set of N incoming cells may appear at the input ports and up to N outgoing cells may depart from the output ports. The N incoming cells are processed in parallel by the input port processors and presented simultaneously to the switch fabric. The switch fabric attempts to route the cells in parallel to their appropriate output ports.

There is a chance that more than one cell may attempt to reach the same output port at the same time, called *output contention*, which has four significant consequences:

1. One cell may reach the output port but the other cells would be lost without buffers existing somewhere in the switch to temporarily store these cells. Switch fabric designs differ in their choice of buffer placement.
2. Queues may accumulate in the buffers resulting in random cell delay and cell delay variation, which are usually two performance metrics of interest.
3. Buffers are necessarily finite implying the possibility of buffer overflow and cell loss. Some fabric designs may not be able to handle a full traffic load without a probability of cell loss. A performance metric for switch fabrics is the normalized throughput or utilization defined as the overall fraction of a full traffic load that can be forwarded successfully through the fabric. Ideally, switch fabrics should be capable of 100 percent utilization.
4. Some switch fabrics must operate at a rate faster than the transmission link rate. The ratio of the switch fabric rate to the transmission link rate is sometimes referred to as a "speedup factor." A speedup factor is often related to the *scalability* of a switch fabric design, that is, the difficulty of constructing an arbitrarily large fabric [3].

3.1. Shared Memory and Shared Medium

A speedup factor of N is evident in switch fabric designs based on a shared memory or shared medium, which are shown in Fig. 2. In a shared memory design, incoming cells are first converted from serial to parallel form. They are written sequentially into a dual-port random-access memory [4]. Their cell headers with routing tags are directed to a memory controller that keeps track of the memory location of all cells associated with each output port. The memory controller links the memory location of outgoing cells to maintain virtual output queues. The outgoing cells are read out of the memory, demultiplexed, and converted from parallel to serial form for delivery to the output port processors. Since the cells must be written into and read out from the memory one at a time, the shared memory must operate at the total throughput rate. Hence, it must be capable of writing N cells and reading N cells in one cell time, implying a speedup factor of N. As a consequence, the size of the fabric, N, will be limited by the memory access time. On the other hand, a shared memory design has been popular due to its simplicity and efficient sharing of memory space.

Similarly, cells may be passed from input port processors to output port processors through a high-speed time-division multiplexed (TDM) bus. Incoming

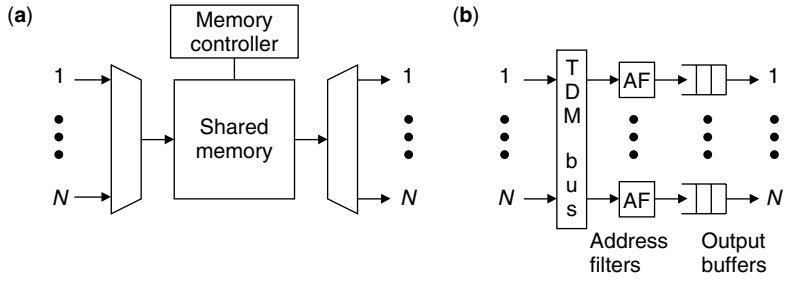


Figure 2. Prototypical switch fabric designs based on (**a**) shared memory and (**b**) shared medium.

cells are sequentially broadcast on the bus. At each output, address filters examine the routing tag on each cell to determine whether the cell is addressed for that output. The address filter passes the appropriate cells through to an output buffer. Shared bus designs have been used in traditional router architectures due to their simplicity and modularity. On the other hand, the buffer space is not shared as efficiently as a shared memory. Also, the bus speed must be fast enough to carry up to N cells in each cell time, corresponding to a speedup factor of N. The address filters and output queues must operate at the bus speed as well. The size of a shared bus fabric will be limited by the expense and complexity of high-speed hardware for the bus, address filters, and output queues.

3.2. Space Division

A simple example of a space division fabric is a crossbar switch shown in Fig. 3, which was originally developed for telephone switching. An $N \times N$ matrix of crosspoints can connect any of the N inputs to any of the N outputs. While a crossbar switch has the advantages of simplicity and no speedup factor, it has two major disadvantages:

1. A crossbar switch will have output blocking, meaning that only one cell may be delivered to an output port at a time. Other cells contending for the same output port may be queued at the input ports, but the normalized throughput for an input buffered fabric is well known to be only $2 - 2^{1/2} = 0.586$ for large N, assuming uniform random traffic; that is, an incoming cell attempts to go to any output port with equal probability independent of all other conditions [5].
2. The N^2 number of crosspoints does not scale well to large fabrics. As an alternative, multistage interconnection networks (MINs) have been studied extensively over many years of development of telephone switches [6]. MINs are constructed by connecting a number of small switching elements, often 2×2 switching elements, in a regular pattern. Banyan networks are a popular class of MINs used for ATM switch fabrics. Figure 4 shows an example of an 8×8 banyan network. The dashed outlines emphasize that the 8×8 banyan network is constructed by adding a third stage to interconnect 4×4 banyan networks, which are in turn constructed by an interconnection of two stages of 2×2 switching elements. An n-level banyan may be constructed by connecting several $(n-1)$-level banyans with an additional stage of switching elements. This recursive and modular construction of larger fabrics is a significant advantage for implementation.

Another advantage is the simplicity of the 2×2 switching elements. Each 2×2 switching element routes a cell according to a control bit. If the control bit is 0, the cell is routed to the upper output (address 0); otherwise, the cell is routed to the lower output (address 1). Delta networks are a subclass of banyan networks with the "self-routing" property: the output address of a cell also controls the route of that cell. For example, the cell shown in Fig. 4 is addressed to output port 010. The nth bit of the address "010" is used as the control bit in the nth stage to route the cell to the proper output port, and this self-routing works regardless of which input port the cell starts from.

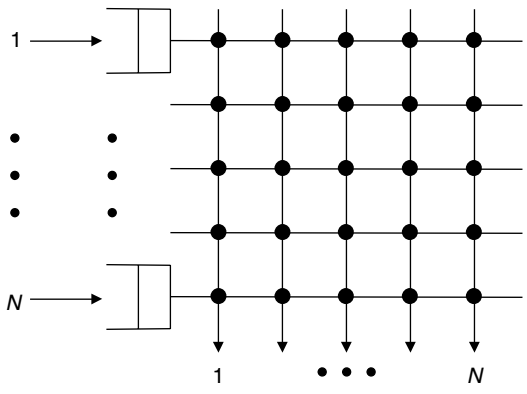


Figure 3. $N \times N$ crossbar switch fabric.

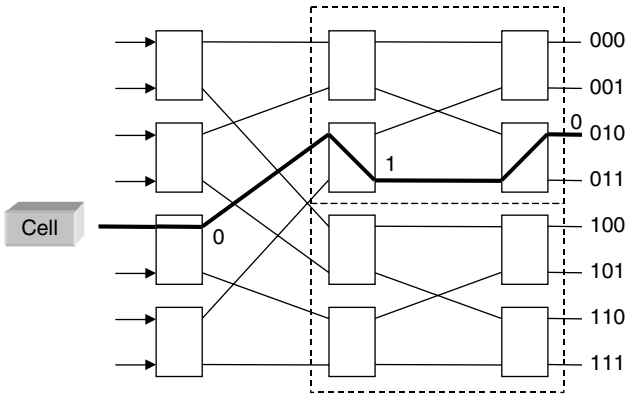


Figure 4. Example of an 8×8 banyan network.

The self-routing property simplifies the control of the delta network switch fabric.

Delta networks can take different forms, depending on their method of construction, including omega, flip, cube, shuffle exchange, and baseline networks [7]. A delta network of size $N \times N$ constructed of $M \times M$ switching elements will have $\log_M N$ stages, each stage consisting of N/M switching elements. Hence, if $M = 2$, the total number of crosspoints will be on the order of $N \log_2 N$, which compares favorably to N^2 crosspoints in a crossbar switch.

Unfortunately, the savings in number of crosspoints comes at the cost of possible internal blocking, meaning that the routes of two cells addressed to different outputs might conflict for the same internal link in the fabric before the last stage. In this situation, only one of the two cells for a link can be passed to the next stage, while the other cell stays behind, queued either in a buffer within each switching element or in an input buffer. Thus, internal blocking will cause a loss of throughput. A well-known solution is to add a Batcher sort network to rearrange the cells according to an increasing or decreasing order of addresses before the banyan network [8]. A combined Batcher–banyan network will be internally nonblocking in that a set of N cells addressed to N different outputs will not cause an internal conflict. However, output blocking can still occur if two cells are addressed to the same output, and it must be resolved by buffering.

An obvious possibility is input buffering before the Batcher–banyan network. If more than one cell is addressed to the same output, one cell is allowed to pass through the Batcher–banyan network while the other cells remain in the input buffers. Naturally, throughput will be lost due to the so-called head-of-line blocking, where a delayed cell prevents the other cells waiting behind it from going through the fabric. Many approaches are possible to overcome the head-of-line blocking problem and increase the throughput of the fabric, such as increasing the speedup factor, distributing the traffic load to multiple banyan networks in parallel, cascading multiple banyan networks in tandem, or virtual output queueing where N separate virtual queues corresponding to the output ports are maintained at each input port. Although these solutions add complexity to the fabric implementation, space-division fabrics are still attractive for their ability to scale to large sizes. Large fabrics may be constructed as MINs composed of small switching modules, where the small switching modules can be any type of fabric design.

3.3. Input and Output Buffering

The placement of buffers in the switch can have a significant effect on the switch performance. Fortunately, this issue has been studied extensively. Figure 5 shows three basic examples: input buffering, output buffering, and internal buffering. Input buffering is known to suffer from head-of-line blocking without special provisions to overcome it. Output buffering is generally agreed to be optimal in terms of throughput and delay [5]. However, output buffering often involves a speedup factor which limits the scalability to large fabrics.

The addition of buffers within the switching elements of a banyan network to resolve internal blocking has not been shown to improve the throughput substantially. An interesting fabric design is a crossbar switch with buffers at each crosspoint [9]. Incoming cells are dropped into the appropriate buffer corresponding to the output. Each output multiplexes the cells queued in N buffers. The buffered crossbar switch (also called *bus matrix switch*) is actually an output buffered fabric as illustrated in Fig. 6. It offers the desirable performance of output buffering with no speedup factor. However, it does not share buffer space efficiently, and the number of output buffers scales exponentially as N^2. The knockout switch shown in Fig. 6 reduces the number of output buffers to NL, where L is a constant by the addition of N : L concentrators at each output [10]. It has been noted that under uniform random traffic conditions, the probability of more than L cells

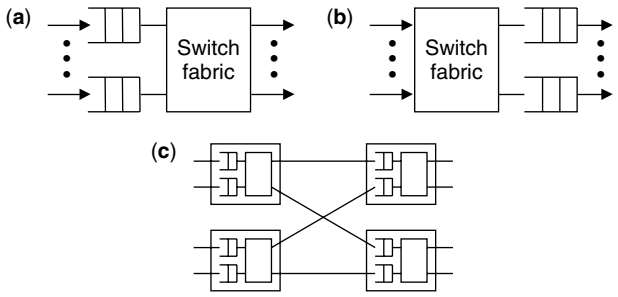


Figure 5. Examples of (**a**) input buffering, (**b**) output buffering, and (**c**) internal buffering.

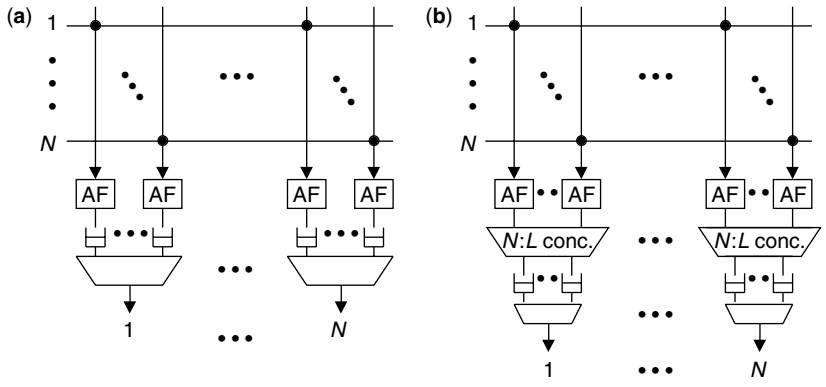


Figure 6. (**a**) Crossbar switch with buffers at each of N^2 crosspoints; (**b**) knockout switch with NL buffers.

addressed to the same output port in the same cell time will be very small if L is chosen appropriately large. The N : L concentrators allow up to only L cells to pass in one cell time to L output buffers at each output; additional cells are lost. However, if L is chosen to be 8 or greater, the cell loss ratio will be 10^{-6} or less. At the cost of a small cell loss, the scalability of the fabric becomes linear with N instead of exponential.

4. CONNECTION CONTROL

Since ATM is a connection-oriented protocol, virtual connections must be established before any user cells can be forwarded. Virtual connections may be permanent, semipermanently controlled through network management, or dynamically established by means of ATM signaling in response to user requests. ATM switches exchange signaling messages along a selected route and make decisions about allocation of switch resources to new user requests. Usually route selection is carried out by a separate process. Routes may be static or dynamically chosen through a routing protocol. The PNNI (private network node interface) routing protocol is a dynamic link-state routing protocol similar to the OSPF (open shortest path first) protocol used in the Internet.

4.1. Signaling

ATM switches must participate in signaling protocols, either access signaling between the user and edge switch or interoffice signaling between two switches. The ATM access signaling protocol is the ITU-T standard Q.2931, which was derived from the ISDN access signaling protocol Q.931. Q.2931 signaling messages are encapsulated in ATM cells using a signaling ATM adaptation layer (SAAL) protocol. Signaling cells are exchanged on a preestablished signaling virtual channel (VCI = 5) or another signaling virtual channel dynamically established through metasignaling (a preestablished metasignaling virtual channel identified by VCI = 1).

The high-layer interoffice signaling protocol is the ITU-T standard BISUP (broadband ISDN user part) derived from the ISDN user part of Signaling System 7 (SS7). BISUP messages may be exchanged directly between ATM switches, where BISUP messages would be encapsulated into ATM cells using SAAL, or sent through the existing SS7 packet-switched network.

Figure 7 shows a typical exchange of signaling messages between ATM switches to successfully establish and release a virtual connection. Basically, a Q.2931 "setup" message is first sent by the user to request a new virtual connection. If each switch decides to accept the request, a BISUP "initial address" message (IAM) is forwarded along a selected route. The IAM message includes all information required to route the connection request to the destination user, such as destination user address, service class, ATM traffic descriptor, connection identifier, quality-of-service (QoS) parameters, and additional optional parameters. A Q.2931 "setup" message notifies the destination user. If the connection is accepted, a series of signaling messages are returned in the reverse direction to alert the calling party that the connection is established. The reverse signaling messages also serve to finalize the resource reservations that were made earlier tentatively in each switch. When the virtual connection is no longer needed, a "release" message will free the reserved resources at each switch to be used for another connection.

The complete ATM signaling protocol, including additional signaling messages, options, and timing requirements, is elaborate to implement. Obviously, signaling cells require special processing within the switch. Incoming signaling cells are recognized and diverted to a signaling protocol engine for processing. Outgoing signaling cells from the signaling protocol engine are multiplexed into the outgoing cell streams.

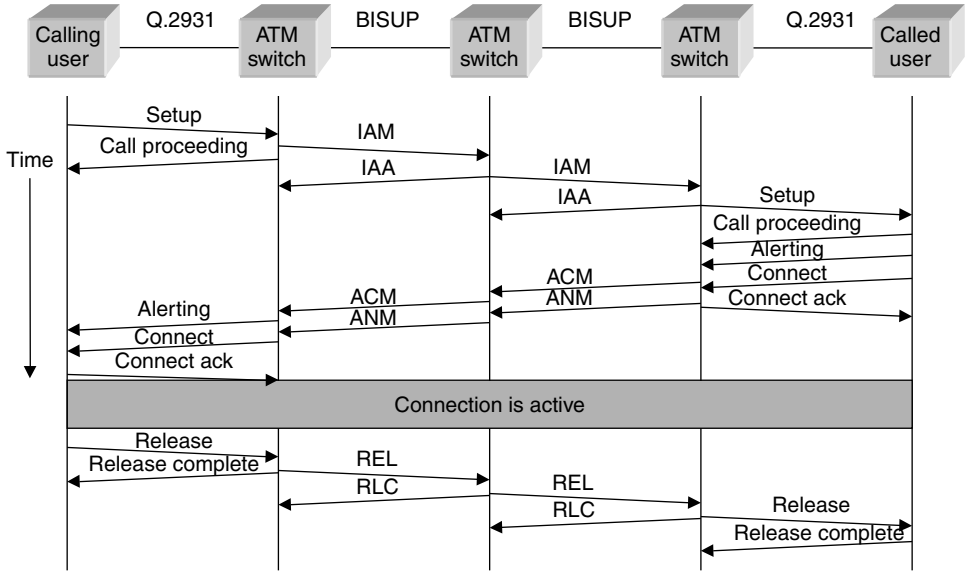


Figure 7. Exchange of signaling messages involved in a successful connection.

4.2. Connection Admission Control

ATM supports the notion that accepted virtual connections will be guaranteed their requested level of QoS — mainly in terms of maximum cell delay, cell delay variation, and cell loss ratio — or otherwise, a new connection request should be rejected. Hence, the acceptance of a virtual connection is an implicit agreement between the user and the network on a mutual understanding of their respective obligations, often called a "traffic contract." The user side of the traffic contract involves conformance to the ATM traffic descriptor or traffic rate parameters. The network side of the traffic contract is a guarantee of the requested QoS for the conforming traffic.

Naturally, not every connection request may be accepted because network resources are shared for higher efficiency. If too much traffic is admitted, the QoS for existing connections may deteriorate below their guaranteed levels. On the other hand, the network should attempt to accept as much traffic as possible to maximize efficiency and revenue. *Connection admission control* (CAC) refers to the general process for deciding acceptance or rejection of new connection requests. The main issue for an ATM switch is whether sufficient resources are available to satisfy the QoS requirements of the new connection and all existing connections. Because ATM traffic is random, the effect of a new connection cannot be known precisely during CAC. The switch follows a CAC algorithm chosen by the network provider to estimate the impact of a new connection.

The numerous CAC algorithms studied over the years can be broadly classified as deterministic or statistical. Deterministic approaches calculate the effect of a new connection on the basis of a deterministic traffic envelope characterizing a bound on the shape of the expected traffic, such as peak cell rate or a leaky-bucket-limited envelope. Statistical methods usually estimate the effect of a new connection by carrying out a stochastic analysis of a queueing model. Statistical methods can be classified as model-based or measurement-based (or a combination of both). Model-based approaches make an assumption about traffic models as inputs to a queueing model. Measurement-based approaches depend on measurements of actual traffic as inputs to an analytical model. In any case, the CAC algorithm is not a matter for standardization and should be chosen by the network provider.

4.3. Routing

The ATM protocol is not tied to a specific routing protocol. Indeed, a dynamic routing protocol is not needed if routes are static. Also, the concept of semipermanent virtual paths was intentionally included in ATM to simplify the routing process. Virtual paths can serve as large "pipes" with allocated bandwidth between pairs of nodes. If a new connection finds a convenient virtual path to its destination, it can make use of an available virtual channel within that virtual path with minimal setup overhead at intermediate switches.

For dynamic routes, PNNI routing is a link-state routing protocol. ATM switches will periodically advertise information about its links and maintain a topological view of the network constructed from link-state advertisements from other switches. These functions are carried out by a routing protocol engine within the connection control function.

5. TRAFFIC CONTROL CONSIDERATIONS

ATM switches are responsible for a comprehensive set of traffic control mechanisms to support QoS guarantees in addition to connection control [11,12]. For the most part, these other mechanisms operate in various parts of the switch independently of connection control.

5.1. Usage Parameter Control

Although the source traffic is expected to conform to the traffic descriptor negotiated during connection establishment, the actual source traffic may be excessive for various reasons. To protect the QoS of other connections, the source traffic rate needs to be monitored and regulated at the user-network interface by ATM edge switches. Usage parameter control (UPC) is the process for traffic regulation or "policing" carried out by a leaky-bucket algorithm called the *generic cell rate algorithm*. The generic cell rate algorithm involves two parameters, an increment I and a limit L, and is therefore denoted as GCRA (I,L). The parameter I is inversely proportional to the average rate allowed by the GCRA, while the parameter L determines its strictness. The GCRA is activated for a virtual connection after it has been accepted.

The operation of the GCRA is illustrated in Fig. 8. A bucket of capacity $I + L$ drains continuously at a rate of 1 per unit time. A cell is deemed to be conforming if the bucket contents can be incremented by I without overflowing; otherwise, the cell is deemed to be nonconforming or excessive. Conforming cells should be allowed to pass the GCRA without any effect. The network administrator can choose non-conforming cells to be allowed, discarded, or tagged by setting CLP = 1 in the cell header.

A virtual scheduling algorithm offers an alternative but equivalent view of the GCRA. The actual arrival time of the nth cell, $t(n)$, is compared with its theoretical arrival time $T(n)$, which is the expected arrival time assuming that all cells are spaced equally in time with separation I. Cells should not arrive much earlier than their theoretical arrival times, with some tolerance dependent on L. A cell is deemed to be conforming if $t(n) > T(n) - L$; otherwise, it is nonconforming (too early). The theoretical arrival time for the next cell, $T(n + 1)$, is calculated as a function of $t(n)$. If the nth cell is conforming and $t(n) < T(n)$, then the next theoretical arrival time is set to $T(n + 1) = T(n) + I$.

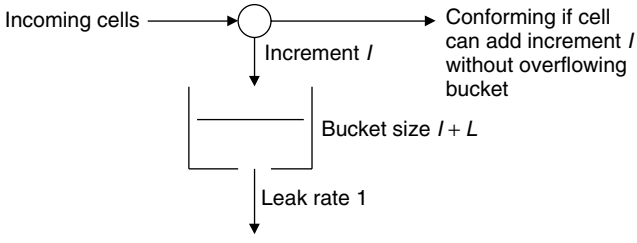


Figure 8. GCRA operation viewed as a leaky-bucket algorithm.

If the n-th cell is conforming and $t(n) \geq T(n)$, then the next theoretical arrival time is $T(n+1) = t(n) + I$. Nonconforming cells are not counted in the update of the theoretical arrival times.

Multiple GCRAs may be used in combination to regulate different sets of parameters. For example, a dual leaky-bucket may consist of a GCRA to regulate the peak cell rate followed by a second GCRA to regulate the sustainable cell rate (an upper bound on the average rate). A conforming cell must be deemed conforming by both GCRAs.

5.2. Packet Scheduling

ATM does not allow indication of service priority on the basis of individual cells. Although service priorities can be associated with virtual connections, it is common to group virtual connections according to their class of service, such as real-time constant bit rate (CBR), real-time variable bit rate (RTVBR), non-real-time VBR (NRTVBR), available bit rate (ABR), and unspecified bit rate (UBR). Real-time services would typically receive the highest service priority; NRTVBR, the second priority; and ABR and UBR, the lowest priority. Packet scheduling is not a matter for standardization and depends on the switch designer.

5.3. Selective Cell Discarding

Selective cell discarding is based on the cell loss priority indicated by the CLP bit in each cell header. CLP =1 cells should be discarded before CLP = 0 in the event of buffer overflows. CLP = 1 cells may be generated by a user who deliberately wants to take a risk with excess traffic or might be tagged cells from the UPC mechanism. In a pushout scheme, a CLP = 0 cell arriving to a full buffer may be allowed to enter the buffer if a queued CLP = 1 cell can be discarded to free space. If more than one CLP = 1 cell is queued, the discarding policy can push out CLP = 1 cells starting from the head or tail of the queue. Pushing out from the tail of the buffer tends to favor more CLP = 1 cells, because the CLP = 1 cells left near the head of the buffer are likely to depart successfully from the buffer. If CLP = 1 cells are pushed from the head of the buffer, the CLP = 1 cells left near the tail of the buffer will take longer to depart and will have a higher risk of being pushed out by the next arriving CLP = 0 cell.

More complicated buffer management strategies are possible. For example, a partial buffer sharing strategy can use a threshold; when the queue exceeds the threshold, only CLP = 0 cells will be admitted into the buffer and arriving CLP = 1 cells will be discarded. This strategy ensures a certain amount of space will always be available for CLP = 0 traffic. Similarly, it is possible to impose an upper limit on the number of CLP = 1 cells queued at any one time, which would ensure some space to be available for only CLP = 0 cells.

5.4. Explicit Forward Congestion Indication

ATM included EFCI as a means for ATM switches to communicate simple congestion information to the user to enable end-to-end control actions. User cells are generated with the second bit in the 3-bit PT field set to 0, signified as EFCI = 0. Any congested ATM switch can set EFCI = 1, which must be forwarded unchanged to the destination user. The algorithm for deciding when to activate EFCI is chosen by the network provider.

EFCI is used for the binary mode of the ABR service [12]. The ABR service is intended to allow rate-adaptable data applications to make use of the unused or "available bandwidth" in the network. An application using an ABR connection is obligated to monitor the receipt of EFCI = 1 cells and change its transmission rate according to a predefined rate adaptation algorithm. The objective is to match the transmission rate to the instantaneous available bandwidth. The ATM switch buffers should be designed to absorb the temporarily excessive traffic caused by mismatch between the actual transmission rate and the available bandwidth. In return for compliance to the rate adaptation algorithm, the ATM network should guarantee a low cell loss ratio on the ABR connection (but no guarantees on cell delay).

5.5. Closed-Loop Rate Control

The binary mode of the ABR rate adaptation algorithm involves gradual decrementing or incrementing of an application's transmission rate. The rate adaptation algorithm for the ABR service also allows an optional explicit mode of operation where the ATM switches along an ABR connection may communicate an exact transmission rate to the application. A resource management cell indicated by a PT = 6 field is periodically inserted into an ABR connection and makes a complete round trip back to the sender. It carries an "explicit rate" field that can be decremented (but not incremented) by any ATM switch along the ABR connection. The sender is obligated to immediately change its transmission rate to the value of the explicit rate field, or the rate dictated according to the binary mode of rate adaptation, whichever is lower.

6. ATM-LAYER OAM

The ATM protocol defines OAM cells to carry out various network management functions in the ATM layer such as fault management and performance management [13]. ATM switches are responsible for the generation, processing, forwarding, and termination of OAM cells according to standardized OAM procedures. OAM cells have the same cell header but their payloads contain predefined fields depending on the function of the OAM cell. F4 OAM cells share a virtual path with user cells. F4 OAM cells have the same VPI value as the user cells in the virtual path but are recognized by the preassigned virtual channels: VCI = 3 for segment OAM cells (relayed along part of a route) or VCI = 4 for end-to-end OAM cells (relayed along an entire route). F5 OAM cells share a virtual channel with user cells. F5 OAM cells have the same VPI/VCI values as the user cells in the virtual channel but have these preassigned PT values: PT = 4 for segment OAM cells and PT = 5 for end-to-end OAM cells.

6.1. Fault Management

OAM cells are used for these fault management functions: alarm surveillance, continuity checks, and loopback testing. If a physical layer failure is detected, a virtual connection failure will be reported in the ATM layer with two types of OAM cells: alarm indication signal (AIS) and

remote defect indicator (RDI). AIS cells are sent periodically "downstream" or in the same direction as user cells effected by the failure to notify downstream switches of the failure and its location. The last downstream ATM switch will generate RDI cells in the upstream direction to notify the sender of the downstream failure.

The purpose of continuity checking is to confirm that an inactive connection is still alive. If a failure has not been detected and no user cells have appeared on a virtual connection for a certain length of time, the switch on the sender's end of a virtual connection should send a continuity check cell downstream. If the switch on the receiver's end of the virtual connection has not received any cell within a certain time in which a continuity check cell was expected, it will assume that connectivity was lost and will send a RDI cell to the sender.

An OAM loopback cell is for testing the connectivity of a virtual connection on demand. Any switch can generate an OAM loopback cell to another switch designated as the loopback point. The switch at the loopback point is obligated to reverse the direction of the loopback cell to the originator. The failure of a loopback cell to return to its originator will be interpreted as a sign that a fault has occurred on the tested virtual connection.

6.2. Performance Management

OAM performance management cells are used to monitor the performance of virtual connections to detect intermittent or gradual error conditions caused by malfunctions. At the sender's end of a virtual connection, OAM performance monitoring cells are inserted between blocks of user cells. Nominal block sizes may range between 2^7, 2^8, 2^9, or 2^{10} cells but do not have to be exact. The OAM performance monitoring cell includes fields for the monitoring cell sequence number, size of the preceding cell block, number of transmitted user cells, error detection code computed over the cell block, and timestamp to measure cell delay. The switch at the receiver's end of the virtual connection will return the OAM cell in the reverse direction with additional fields to report any detected bit errors and any lost or misinserted cells. The timestamp will reveal the roundtrip cell delay. The measurements of cell loss and cell delay reflect the actual level of QoS for the monitored virtual connection.

BIOGRAPHY

Thomas M. Chen received the B.S. and M.S. degrees in electrical engineering in 1984 from Massachusetts Institute of Technology, Cambridge, Massachusetts, and the Ph.D. degree from the University of California, Berkeley, in 1990. He worked at GTE Laboratories on ATM traffic control and network management research until 1997. Since 1997, he has been an Associate Professor in Electrical Engineering at Southern Methodist University, Dallas, Texas, where he has been working on traffic modeling, network management, programmable networks, and network security. He was the recipient of the 1996 IEEE Communications Society's Fred W. Ellersick best paper award. He currently is an associate editor for *ACM Transactions on Internet Technology*, and senior technical editor for *IEEE Communications Magazine* and *IEEE Network*.

Stephen S. Liu received the B.S. degree in electrical engineering from National Cheng-Kung University in Taiwan, and the M. S. and Ph.D. degrees from Georgia Institute of Technology in Atlanta, Georgia. He co-developed the ISO and ANSI standard 32-degree error detection code polynomial used on Ethernet and various high-speed data networks, and co-authored the book entitled *ATM Switching Systems* published by Artech House in 1995. He joined the Verizon Labs (formerly GTE Laboratories) in 1981 and has since been working on packet-switching technology. Dr. Liu's current interest is in unified control plane technology for optical transport networks. He is a senior member of the IEEE.

BIBLIOGRAPHY

1. ITU-T Rec. I.361, *B-ISDN ATM-Layer Specification*, Geneva, July 1995.
2. ATM Forum, *ATM User-Network Interface (UNI) Specification Version 4.0*, April 1996.
3. T. Chen and S. Liu, *ATM Switching Systems*, Artech House, Boston, 1995.
4. N. Endo et al., Shared buffer memory switch for an ATM exchange, *IEEE Trans. Commun.* **41**: 237–245 (Jan. 1993).
5. M. Karol, M. Hluchyj, and S. Morgan, Input versus output queueing on a space-division switch, *IEEE Trans. Commun.* **35**: 1347–1356 (Dec. 1987).
6. T.-Y. Feng, A survey of interconnection networks, *IEEE Commun. Mag.* **14**: 12–27 (Dec. 1981).
7. X. Chen, A survey of multistage interconnection networks in fast packet switches, *Int. J. Digital Analog Cabled Syst.* **4**: 33–59 (1991).
8. J. Hui, Switching integrated broadband services by sort-banyan networks, *Proc. IEEE* **79**: 145–154 (Feb. 1991).
9. S. Nojima, E. Tsutsui, H. Fukuda, and M. Hashimoto, Integrated services packet network using bus matrix switch, *IEEE J. Select. Areas Commun.* **SAC-5**: 1284–1292 (Oct. 1987).
10. Y. Yeh, M. Hluchyj, and A. Acampora, The knockout switch: A simple, modular architecture for high-performance packet switching, *IEEE J. Select. Areas Commun.* **SAC-5**: 1274–1283 (Oct. 1987).
11. ITU-T Rec. I.371, *Traffic Control and Congestion Control in B-ISDN*, Geneva, July 1995.
12. ATM Forum, *Traffic Management Specification Version 4.0*, April 1996.
13. ITU-T Rec. I.610, *B-ISDN Operation and Maintenance Principles and Functions*, Geneva, July 1995.

FURTHER READING

Several good surveys of ATM switch fabric architectures can be found in the literature:

H. Ahmadi and W. Denzel, A survey of modern high-performance switching techniques, *IEEE J. Select. Areas Commun.* **7**: 1091–1103 (Sept. 1989).

A. Pattavina, Nonblocking architectures for ATM switching, *IEEE Commun. Mag.* **31**: 38–48 (Feb. 1993).

E. Rathgeb, T. Theimer, and M. Huber, ATM switches — basic architectures and their performance, *Int. J. Digital Analog Cabled Syst.* **2**: 227–236 (1989).

F. Tobagi, Fast packet switch architectures for broadband integrated services digital networks, *Proc. IEEE* **78**: 133–178 (Jan. 1990).

R. Awdeh and H. Mouftah, Survey of ATM switch architectures, *Comput. Networks ISDN Syst.* **27**: 1567–1613 (1995).

A wealth of papers can be found on performance analysis of switch architectures; examples are

A. Pattavina and G. Bruzzi, Analysis of input and output queueing for nonblocking ATM switches, *IEEE/ACM Trans. Network.* **1**: 314–327 (June 1993).

D. Del Re and R. Fantacci, Performance evaluation of input and output queueing techniques in ATM switching systems, *IEEE Trans. Commun.* **41**: 1565–1575 (Oct. 1993).

The difficulties of constructing large ATM switches are explored in

T. Banwell et al., Physical design issues for very large scale ATM switching systems, *IEEE J. Select. Areas Commun.* **9**: 1227–1238 (Oct. 1991).

T. Lee, A modular architecture for very large packet switches, *IEEE Trans. Commun.* **38**: 1097–1106 (July 1990).

ATMOSPHERIC RADIOWAVE PROPAGATION

HAROLD RAEMER
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Radiowave propagation in the terrestrial environment is a very mature science. Its basic principles were understood in the nineteenth century, after the work of Faraday, Maxwell, Hertz, Sommerfeld, and others had resulted in a formulation of the basic equations of electromagnetic theory and their application to radio wave propagation. However, it was in the early 1950s that these principles were intensively applied to the rapidly developing technology of microwave communications. The impetus for that technology was the radar research that occurred during and immediately after World War II. Much of what was learned about the propagation of radar waves [1] was directly applicable to line-of-sight communication links operated at frequencies from 300 MHz through the microwave bands up to about 30 GHz. With the rapid growth of wireless communication systems in the 1990s, it has become increasingly important to engineers involved in the development and improvement of those systems to understand the propagation environment within that band. Information obtained from propagation studies is directly applicable to decisions on such issues as siting of transmitters and receivers to optimize reception and prediction of fading rates and intersymbol interference due to multipath propagation in digital transmission. Since most wireless systems are mobile, it is particularly important, for a specific swath of terrain over which a system is operating, to know how the received signal varies with time as transmitters and/or receivers travel. That knowledge can be acquired through propagation analysis aided by topological information about the environment and can help reduce the amount of expensive and time-consuming field experimentation required for system design decisions.

Terrestrial radio wave propagation is a vast subject when it covers the entire radio spectrum from the kilohertz region through millimeter waves. The subject could not be adequately covered in an article of this length. This article is confined to the frequency range from 30 MHz to 30 GHz and to "space wave" propagation in the troposphere; the region below 16–18 km above the earth's surface [2, pp. 100–141], that is, ground wave [2, pp. 33–61] ionospheric ("sky wave;" see Ref. 2, pp. 62–99 or Ref. 3, pp. 218–255) and satellite transmission [2, pp. 263–295], are not included. The emphasis is on the propagation phenomena important in mobile wireless communication systems [4,5]. Today's wireless links operate at certain key frequencies [450, 900 MHz (mobile cellular), 2.4, 5.8 GHz (indoor wireless)], but much higher frequencies (e.g., 22 GHz) are being investigated for some applications and will probably be in use within the near future. In summary, the range of frequencies of greatest interest for terrestrial wireless communication is between 30 MHz and 30 GHz, namely, the VHF band (30–300 MHz), the UHF band (300 MHz–3 GHz), and the SHF band (3–30 GHz).

The modes of propagation not covered in this article operate primarily in frequency bands outside this range. Ground-wave propagation predominates at frequencies below 3 MHz, for instance, in the very low-frequency (VLF; 3–30 kHz), low-frequency (LF; 30–300 kHz), and medium-frequency (MF; 300 kHz–3 MHz) bands, while sky-wave propagation predominates in the high-frequency band (HF; 3–30 MHz) and also occurs in VLF, LF, and MF bands. Satellite links operate primarily in the SHF bands, overlapping our spectral region of interest, but satellite communication is a major topic in its own right, and its propagation aspects are not included in this article.

2. THEORY

2.1. Free-Space Propagation Equations

The idealized propagation geometry for a wireless communication link is illustrated in Fig. 1. The transmitter at T and the receiver at R are separated by a distance D_{TR}. If the transmitting and receiving antennas were in infinite free space, the vertically polarized (V) or horizontally polarized (H) components of the vector phasor of the electric field of the wave transmitted from T at the receiving point R would be

$$E_R^{(V,H)} = \frac{e^{-jk_0 D_{TR}}}{D_{TR}} \sqrt{\frac{P_T G_{T0} A_{eR0}}{4\pi}} f_T^{(V,H)}(\Delta\boldsymbol{\Omega}_{TR}) f_R^{(V,H)}(\Delta\boldsymbol{\Omega}_{RT}) \tag{1}$$

where P_T is the total power radiated by the transmitting antenna; G_{T0} and A_{eR0} are the peak values of antenna gain and effective aperture area of transmitting and receiving

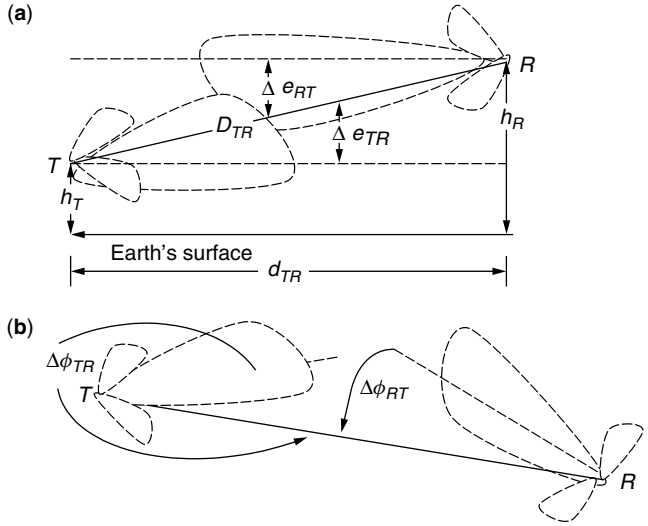


Figure 1. Idealized propagation geometry for wireless communication: (**a**) side view — antennas pointing horizontally in elevation; (**b**) downward view — antennas pointed arbitrarily in azimuth.

antennas, respectively; k_0 is the free-space wave number, equal to $2\pi/\lambda_0$, where λ_0 is the free-space wavelength; and $f_T^{(V,H)}(\Delta\boldsymbol{\Omega}_{TR})$ and $f_R^{(V,H)}(\Delta\boldsymbol{\Omega}_{RT})$ are the angular radiation pattern and receptivity pattern of transmitting and receiving antennas, respectively. The argument of f_T is a two-dimensional vector whose components are $\Delta e_{TR} = e_{TR} - e_{T0}$ and $\Delta\phi_{TR} = \phi_{TR} - \phi_{T0}$, where e_{TP} and ϕ_{TP} denotes the elevation and azimuth angle, respectively, of an arbitrary point P with respect to T and where the subscript 0 denotes the point where the pattern amplitude reaches its maximum. The vector $\Delta\boldsymbol{\Omega}_{RT}$, the argument of f_R, has the same meaning as $\Delta\boldsymbol{\Omega}_{TR}$ except that the reference point is R rather than T. These angles are illustrated in Fig 1.

2.2. Effect of Earth Reflection

The actual field component E_R in the presence of the earth is not given by Eq. (1) alone but includes an additional term due to the reflection of the transmitted wave from a specular point S on the earth's surface, as illustrated in Fig. 2. If the earth's surface is perfectly smooth within the region of interest, then the total field component at R is given by

$$E_R^{(H,V)} = E_{R0}^{(H,V)} F^{(H,V)} \tag{2}$$

where E_{R0} is the free-space wave field given by (1), (the "direct wave") and $F^{H,V}$, known as the *propagation factor* or *path-gain factor*, is given by

$$F^{H,V} = 1 + (R')^{H,V} e^{-jk_0(D_{TS}+D_{SR}-D_{TR})} \tag{3}$$

where, as shown in Fig. 2, D_{TS} and D_{SR} are separation distances between the specular point and transmitter and receiver, respectively, and $(R')^{H,V}$ is a factor of the general form

$$(R')^{H,V} = \frac{f_T^{(H,V)}(\Delta\Omega_{TS}) f_R^{(H,V)}(\Delta\Omega_{RS}) R^{H,V}}{f_T^{(H,V)}(\Delta\Omega_{TR}) f_R^{(H,V)}(\Delta\Omega_{RT})}$$

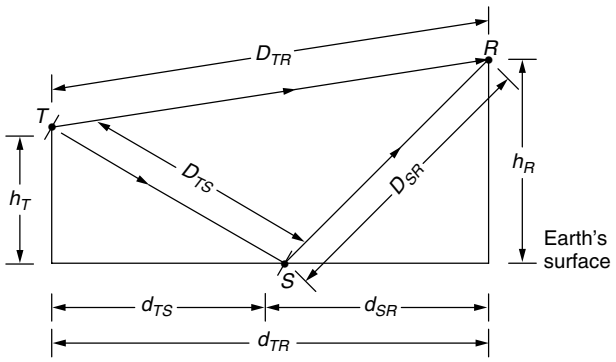


Figure 2. Effect of earth reflection — flat-earth approximation.

where R^H and R^V are the Fresnel reflection coefficients for $H(TE)$ and $V(TM)$ polarizations, respectively, given by

$$R^H = \frac{\mu_R \cos\Theta_i - \sqrt{\nu^2 - \sin^2\Theta_i}}{\mu_R \cos\Theta_i + \sqrt{\nu^2 - \sin^2\Theta_i}} \tag{4a}$$

$$R^V = \frac{\varepsilon_{CR} \cos\Theta_i - \sqrt{\nu^2 - \sin^2\Theta_i}}{\varepsilon_{CR} \cos\Theta_i + \sqrt{\nu^2 - \sin^2\Theta_i}} \tag{4b}$$

where $\varepsilon_{CR} = (\varepsilon_1/\varepsilon_0) - (j\sigma/\omega\varepsilon_0)$, $\mu_R = (\mu_1/\mu_0)$, where ε_0 and μ_0 are respectively the permittivity and magnetic permeability of free-space $\varepsilon_0 = 10^{-9}/36\pi$ farads per meter and $\mu_0 = 4\pi(10^{-7})$ henrys per meter; ε_1, μ_1, and σ, are respectively permittivity, permeability, and conductivity (in siemens per meter) of the earth medium; and ν is the medium's complex refractive index, equal to $\sqrt{\varepsilon_{CR}\mu_R}$.

Invoking the flat-earth approximation, valid for short-range communication links, we can express the pathlength difference appearing in the phase in (3) in the form (see Fig. 2)

$$\Delta L = D_{TS} + D_{SR} - D_{TR} = \sqrt{d_{TR}^2 + (h_R + h_T)^2} - \sqrt{d_{TR}^2 + (h_R - h_T)^2} \tag{5}$$

where h_T and h_R are antenna heights and d_{TR} the horizontal distance between T and R. Given the approximation

$$|h_T \pm h_R| \ll d_{TR} \tag{6}$$

nearly always valid for ground-based transmitting and receiving antennas, we have

$$k_0\Delta L \simeq k_0 d_{TR}\left[1 + \frac{1}{2}\left(\frac{h_R + h_T}{d_{TR}}\right)^2 - 1 - \frac{1}{2}\left(\frac{h_R - h_T}{d_{TR}}\right)^2\right] = \frac{4\pi h_T h_R}{\lambda_0 d_{TR}} \tag{7}$$

Approximation (6) implies that the antenna pattern function for transmitter to ground reflection point (GRP) is nearly the same as that for transmitter to receiver and the same applies to the patterns from receiver to GRP and receiver to transmitter. The assumption that

these patterns are the same and Eq. (7) gives us a simple approximation for (3):

$$F^{H,V} \simeq 1 + R^{H,V} \exp\left(-j\frac{4\pi h_T h_R}{\lambda_0 d_{TR}}\right) \tag{8}$$

The simplified form of the path-gain factor (8) can be used as the basis for "coverage diagrams," examples of which are shown in Ref. 6 (pp. 357–361), which show, for various values of h_T, h_R, d_{TR}, and λ_0, the locations and amplitudes of the peaks and the locations and depths of the troughs of the fields in the propagation plane (defined as the vertical plane containing transmitter, receiver, and GRP) due to interference between the direct and ground-reflected waves. These examples [6] include the effect of earth curvature, to be discussed in Section 2.3 (below). Flat-earth examples are given in Ref. 6 (p. 344).

A further simplification of (8) is achieved by assuming near-grazing incidence, [a further consequence of approximation (6)], namely, $\Theta_i \simeq \pi/2$, in (4), implying that $R^V \simeq R^H \simeq -1$. Under this assumption, the amplitude of (8) reduces to

$$\left|F^{H,V}\right| \simeq 2\left|\sin\left(\frac{2\pi h_T h_R}{\lambda_0 d_{TR}}\right)\right| \tag{9}$$

implying that nulls occur when $h_T h_R/d_{TR}$ is an integral multiple of a half-wavelength and peaks occur when $(h_T h_R/d_{TR}) = (n\lambda_0/2) + (\lambda_0/4)$, where n is an arbitrary integer. The simplified form (9) can be used to roughly estimate the locations of peaks and troughs of the path-gain factor.

2.3. Effect of Earth Curvature: Atmospheric Refraction

As illustrated in Fig. 3, the refractive index of the atmosphere immediately above the earth's surface, although it is very close to unity at all altitudes, decreases approximately exponentially with altitude. The propagation vector, indicating the direction of the radiowave's energy flow, is continuously changing direction if it has a locally vertical component. Governed by Snell's law, in a standard atmosphere, the wave travels in a curved path as shown in Fig. 3. The altitude dependence of the refractivity $N(z)$, defined as $N(z) = [\nu(z) - 1] \times 10^6$, where $\nu(z)$ is the refractive index, is

$$N(z) = N(z_0)e^{-\alpha(z-z_0)} \tag{10}$$

where z_0 is sea level and α is a positive number dependent on atmospheric conditions.

The curvature (reciprocal of the radius of curvature) of the raypath is equal to $-(d\nu(z)/dz)$. The curvature of the earth is $1/r_t$, where r_t is the true earth radius, approximately 6340 km. Using values of α and $N(z_0)$ from the literature, the difference between raypath curvature and earth curvature is about $1.1(10^{-7})\text{m}^{-1}$. The resulting modified earth radius r_e, is between about 1.3 and 1.4 times the true radius (due to variability of atmospheric conditions), but is usually assumed to be $\frac{4}{3}r_t$, giving rise to the concept of the "four-thirds earth radius," commonly used in radio propagation modeling.

Using r_e, the modified earth radius explained above, we then model the propagation as if the rays were straight lines. The maximum distance over which the receiver has a line-of-sight view of the transmitter unobstructed by the earth's curvature (illustrated in Fig. 4a) is given by (where $h_T \ll r_e$, $h_R \ll r_e$)

$$\begin{aligned} D_{\max} = D_{TG} + D_{GR} &= \sqrt{(r_e + h_T)^2 - r_e^2} + \sqrt{(r_e + h_R)^2 - r_e^2} \\ &\simeq \sqrt{2r_e h_T} + \sqrt{2r_e h_R} \end{aligned} \tag{11}$$

If $D_{TR} \le D_{\max}$, then the propagation model is line-of-sight plus a specular earth-reflected ray and the path-gain factor can still be modeled approximately by (8), with the aid of (3) through (7), but with significant modifications. If $D_{TR} > D_{\max}$, then R and T are below the "radio horizon" of T and R, respectively (the radio horizons for T and R are defined as $D_{TG} \simeq \sqrt{2r_e h_T}$ and $D_{GR} \simeq \sqrt{2r_e h_R}$ respectively). This implies that the raypaths between T and R are obscured from each other by the earth's curvature and line-of-sight transmission becomes infeasible. The only feasible modes of propagation in this case are diffraction around the earth [1, pp. 109–112; 6, pp. 369–372], tropospheric scatter [2, pp. 216–237], or satellite transmission [2, pp. 263–295; 3, pp. 100–103].

An important modification is the generalization of (7) to include earth curvature [3, pp. 56–65]. From Fig. 4b, we note that application of the law of sines to the triangles TOS and ROS, the law of cosines to the triangle TSR, and the law of reflection at the point S, together with the observation that $\alpha_1 = d_{TS}/r_e$ and $\alpha_2 = d_{SR}/r_e$, result in the expressions

$$\frac{D_{TS}}{\sin(d_{TS}/r_e)} = \frac{r_e + h_T}{\sin\Theta_i} = \frac{r_e}{\sin(\Theta_i - d_{TS}/r_e)} \tag{12a}$$

$$\frac{D_{SR}}{\sin(d_{SR}/r_e)} = \frac{r_e + h_R}{\sin\Theta_i} = \frac{r_e}{\sin(\Theta_i - d_{SR}/r_e)} \tag{12b}$$

$$D_{TR}^2 = D_{TS}^2 + D_{SR}^2 - 2D_{TS}D_{SR}\cos 2\Theta_i \tag{12c}$$

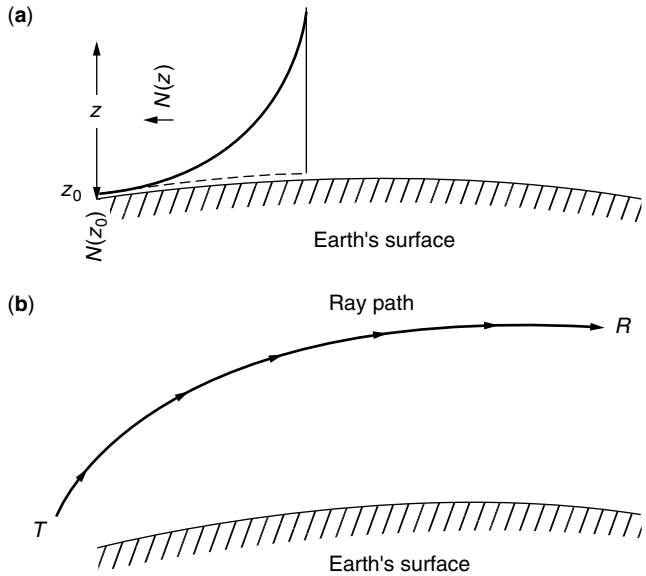


Figure 3. Atmospheric refraction and earth curvature effects: (**a**) variation of refractivity with altitude; (**b**) curvature of ray-path due to variable refractivity.

By adding the assumptions that $d_{TS}/r_e \ll 1$, $d_{SR}/r_e \ll 1$ and the grazing angle $[(\pi/2) - \Theta_i]$ is much smaller than $\pi/2$, applicable to low-altitude propagation paths, manipulations of Eqs. (12a–c) result in the approximate generalization of (5) and (7) to account for earth curvature [6, pp. 349–352]:

$$k_0 \Delta L \simeq \frac{4\pi h_T h_R}{\lambda_0 d_{TR}} \left[1 - \frac{1}{2r_e}\left(\frac{d_{TS}^2}{h_T} + \frac{d_{SR}^2}{h_R}\right)\right] \tag{13}$$

Another effect of earth curvature is the divergence of reflected waves due to the convexity of the reflecting surface near the specular point (Fig. 4c). This can be modeled by multiplication of the reflection coefficient in (4) by a "divergence factor." The general form of this factor is attributed to Vanderpol and Bremmer and its derivation can be found in Ref. 1, (pp. 404–406). Using the usual approximations $h_T \ll r_e$, $h_R \ll r_e$, $(\pi/2) - \Theta_i \ll (\pi/2)$ a simplified form is

$$F_d \simeq \left(1 + \frac{2 d_{TS} d_{SR}}{r_e (d_{TS} + d_{SR}) \cos \Theta_i}\right)^{-1/2} \tag{14}$$

The divergence factor reduces the reflected wave energy relative to the direct wave energy and hence decreases the importance of earth-reflection in the propagation factor.

The location of the ground reflection point and the incidence angle Θ_i at that point are also affected by earth curvature. For the flat-earth case, it is evident from Fig. 2 that

$$d_{TS} = \frac{d_{TR} h_T}{h_T + h_R}; \Theta_i = \cot^{-1}\left(\frac{h_T + h_R}{d_{TR}}\right) \tag{15a}$$

while for the curved earth case [1, p. 113], solution of a cubic equation is required to determine d_{TS}, with the result $d_{TS} = (d_{TR}/2) + p \cos[(\Phi + \pi)/3]$, where

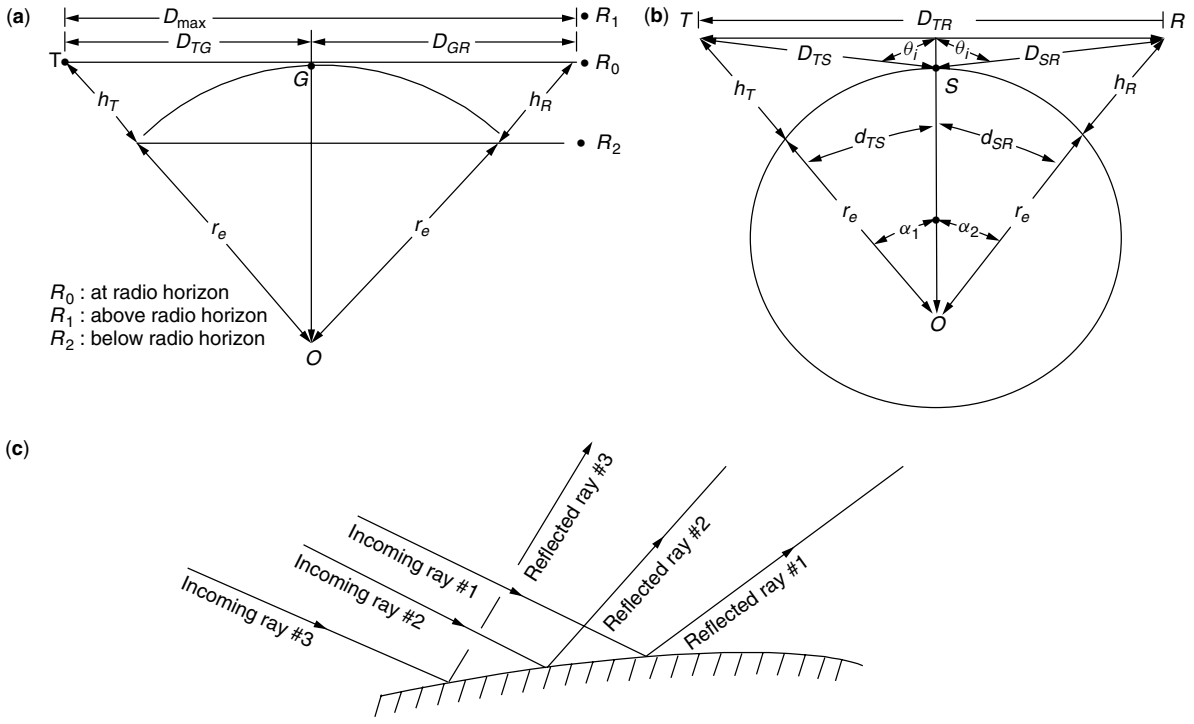
$$p = \frac{2}{\sqrt{3}} \sqrt{r_e (h_T + h_R) + \left(\frac{d_{TR}}{2}\right)^2};$$
$$\Phi = \cos^{-1}\left[\frac{2 r_e |h_R - h_T| d_{TR}}{p^3}\right] \tag{15b}$$

and the angle of incidence Θ_i is approximated by [6, p. 351]

$$\Theta_i \simeq \cot^{-1}\left(\left[\frac{h_T + h_R}{d_{TR}}\right] - \frac{1}{2r_e}\left[\frac{(h_T + h_R)(d_{SR}^2 h_R + d_{TS}^2 h_T)}{d_{TR}(h_T^2 + h_R^2)}\right]\right) \tag{15c}$$

2.4. Effect of Surface Roughness

The effect of surface roughness on the path-gain factor depends on the scale of the roughness. Figure 5 illustrates this in two dimensions by showing three different roughness scales and the corresponding changes in the numbers of specular reflections. It is intuitively evident that there will be a multiplicity of surface points that will result in a specular reflection of a wave from the transmitter into the direction of the receiver. Figure 5c shows that some of the specular reflections from a very rough surface may be mitigated or nullified by shadowing, thus establishing an upper limit on the effect. The theory that leads to (3) and (4a,b) is no longer strictly valid for rough surfaces. However, for a low level of roughness, the effect can



A bundle of nearly parallel incoming rays from transmitter diverges upon reflection due to local earth curvature.

Figure 4. Some effects of earth curvature: (**a**) radio horizon effect; (**b**) effect of pathlength difference; (**c**) divergence of rays reflected from earth's surface.

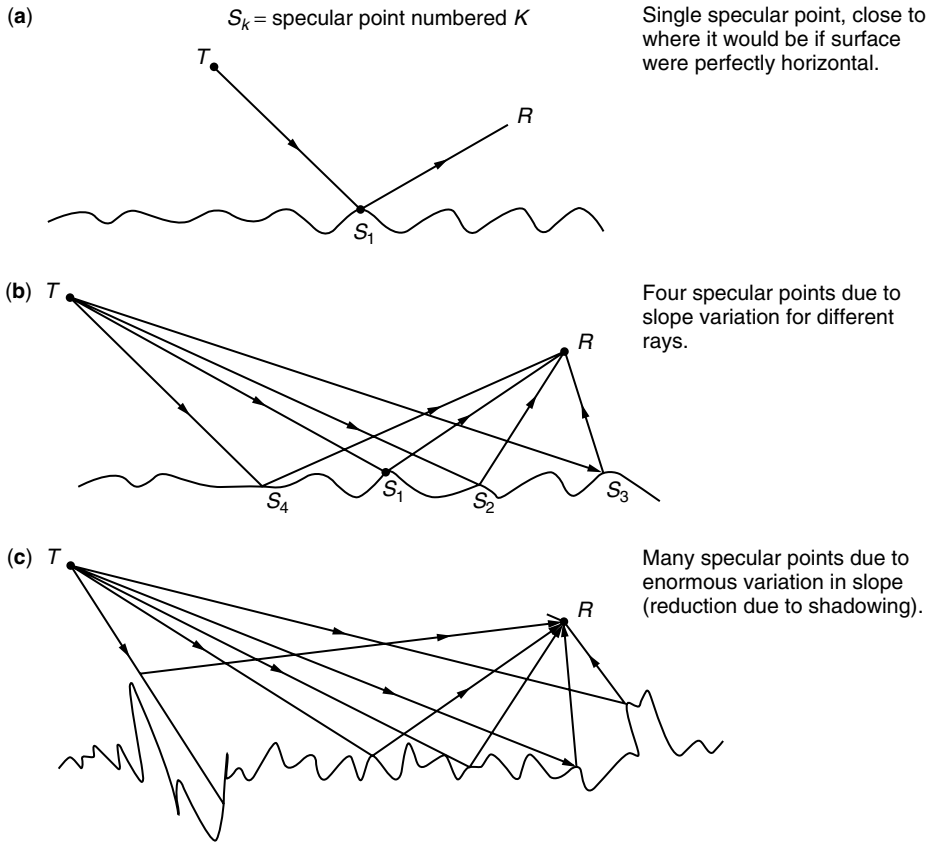


Figure 5. Specular reflections from roughness: (**a**) slightly, (**b**) moderately, and (**c**) very rough surfaces.

be approximated by a "roughness factor" multiplying the reflection coefficient.

Before accounting for roughness, it is desirable to invoke a criterion to determine whether it has a significant effect on the reflected wave, known as the "Rayleigh criterion" where σ is rms surface height, illustrated in Fig. 6a and given by the following rule:

If

$$\sigma \leq \frac{(0.1)\lambda_0}{4\pi \cos \Theta_i}$$

the surface can be approximated as smooth, where σ is the rms height of the rough surface. If σ exceeds

$$\frac{(0.1)\lambda_0}{4\pi \cos \Theta_i}$$

then roughness must be accounted for. The maximum rms height that justifies the smooth surface approximation, $\sigma_{\max}$, is inversely proportional to frequency and increases monotonically with angle of incidence. A surface considered "rough" near normal incidence can be approximated as "smooth" near grazing incidence at a given frequency.

From Fig. 6, it is evident that the difference in pathlength between two ground-reflected parallel rays between T and R, one from the mean surface and the other from a local peak, is $2\Delta h \cos \Theta_i$. The local deviation of the height from its mean value, Δh, is a random variable in a typical ground or sea surface and is often assumed to have a zero-mean Gaussian distribution in the absence of known terrain-specific statistics. The phase difference between two rays is

$$\frac{2\pi}{\lambda_0}(2\Delta h \cos \Theta_i) = \frac{4\pi \Delta h}{\lambda_0} \cos \Theta_i$$

which is also a zero-mean Gaussian random variable. It is shown by more sophisticated methods (e.g., Ref. 7, pp. 80–81 or Ref. 8, pp. 399–401) that this results in 'a "roughness factor" multiplying the reflection coefficient in (8) and given by

$$F_r = \exp\left(-\frac{1}{2}\left(\frac{4\pi}{\lambda_0} \cos \Theta_i\right)^2 \sigma^2\right) \tag{16}$$

where $\sigma^2 = \langle (\Delta h)^2 \rangle$. Like the divergence factor given by (14), this reduces the effect of the reflected wave.

Returning to the basis of the Rayleigh criterion, if $(4\pi \cos \Theta_i/\lambda_0)\sigma$ is less than one-tenth of a radian, or equivalently $\sigma < 0.1\lambda_0/4\pi \cos \Theta_i$, then roughness is considered negligible. Through (16), this implies that $|F_r - 1| < 0.01$, which in turn implies that F_r can be approximated as unity and hence roughness is negligible in the height-gain function. However it should be noted that (16) applies only for small-scale roughness. If σ is far in excess of the Rayleigh limit, then the effect can no longer be modeled in such a simple way and much more complicated theory is required to account for it.

If, as is often the case, large-scale roughness well beyond the Rayleigh limit is distributed throughout the antenna beam coverage region, then specular points with

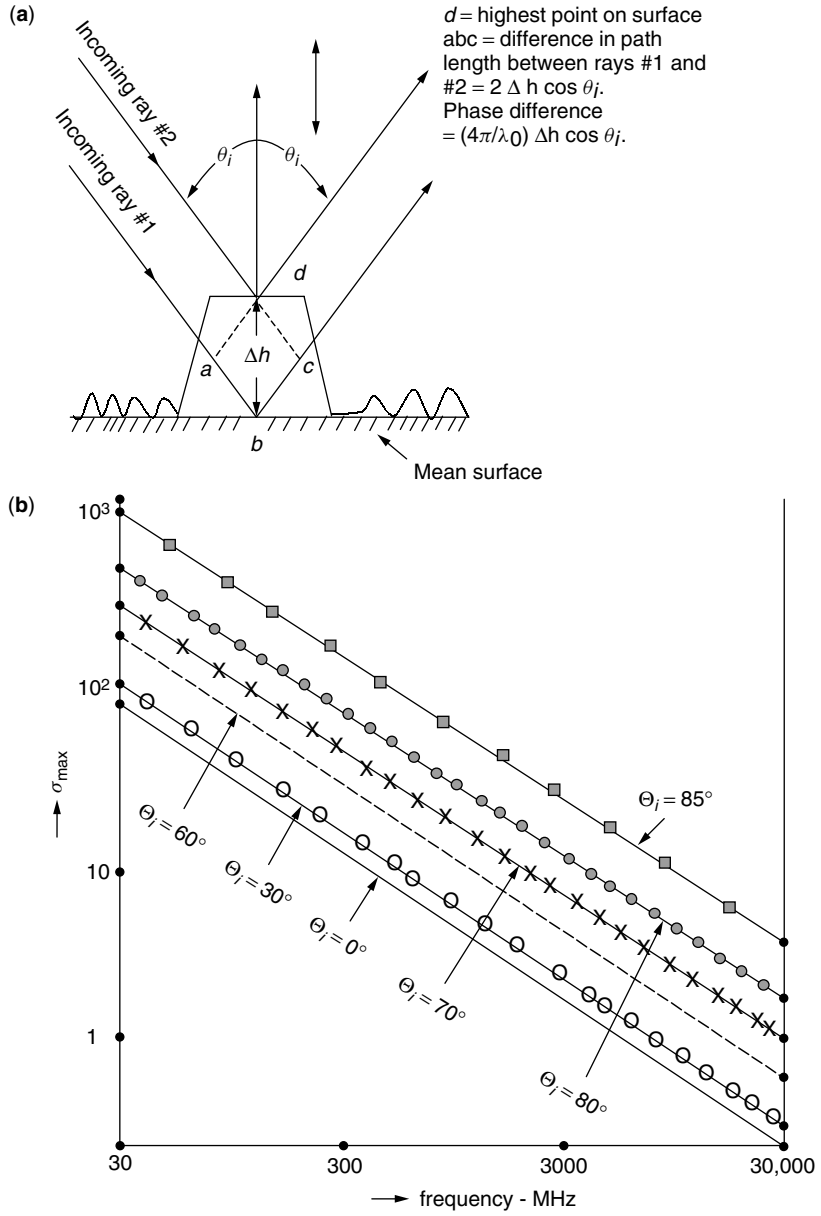


Figure 6. Rayleigh criterion for surface roughness: (**a**) basis of Rayleigh criterion; maximum allowable (**b**) frequency variation of RMS height for smooth surface approximation.

respect to transmitted waves reflected into the receiver's direction exist over a wide swath of terrain. In this case (Fig. 5c), more accurate modeling would require that the height-gain function of (3) be replaced by

$$F^{H,V} \simeq 1 + \sum_{1}^{N} (R')_n^{H,V} e^{-jk_0 \Delta L_n} \tag{17}$$

where $(R')_n^{H,V}$ and the path delay ΔL_n are of the form given in (3) but applicable to the nth specular point. In order to apply (17), one needs either a terrain-specific or a statistically generated surface-height distribution within the coverage area and algorithms to (1) locate the specular points, (2) determine for each such point the angles of incidence of the wave from T and reflection of those waves toward R, and (3) compute the superposition of terms of (17) using stationary phase.

This requires extensive computational power for a swath of very irregular land terrain or a very rough water surface. There are simulation programs available that perform these tasks within reasonable computation times and therefore provide the ability to obtain reality-based coverage diagrams quickly and easily (e.g., SEKE [9]).

2.5. Obstacles Along Propagation Path: Diffraction

Line-of-sight propagation is limited by obstacles along the propagation path, e.g., hills or foliage in irregular open terrain, high waves in a rough sea surface, or buildings in an urban area. The theory of diffraction around obstacles that are opaque to direct waves is required to analyze this situation (Figure 7). Standard line-of-sight theory would predict zero fields behind the obstacle, but diffraction theory shows nonzero fields in the region and can provide

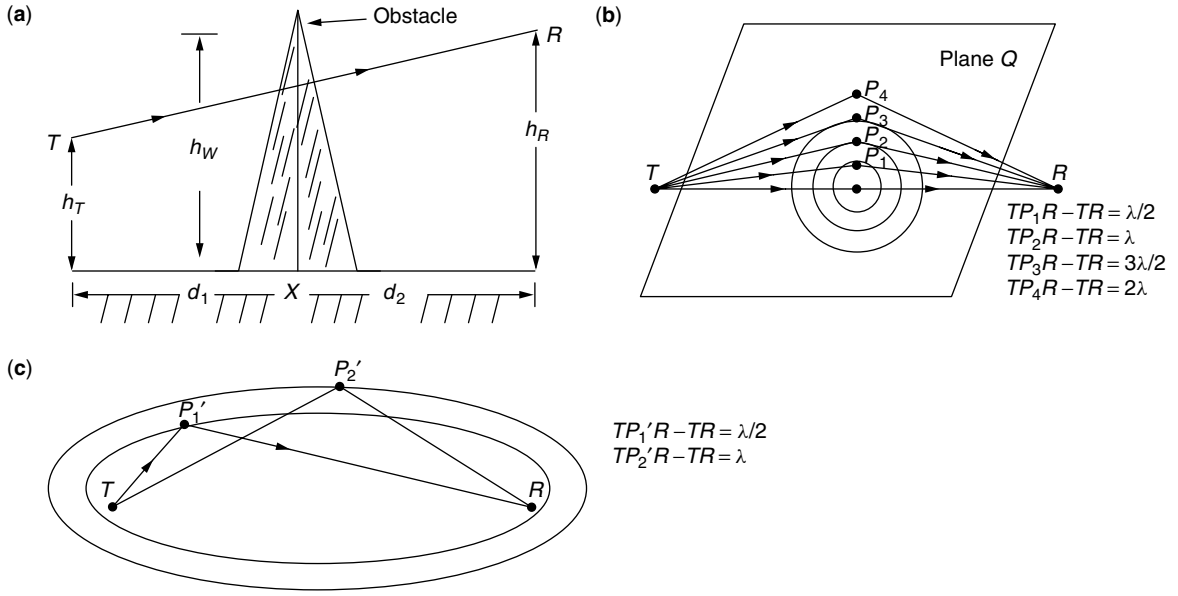


Figure 7. Diffraction around an obstacle: (**a**) obstacle intercepting raypath TR; (**b**) Fresnel zones on plane Q; (**c**) Fresnel ellipsoids.

good predictions of their magnitude [2, pp. 111–112, 122–130; 3, pp. 33–36, 46–50; 5, pp. 90–100].

In Fig. 7, we show a vertical plane Q, normal to the vertical propagation plane P. An opaque obstacle (e.g., hill or building) is placed along the raypath and centered on the plane Q. According to Huyghens' principle, the field at R is a superposition of waves from all points on the plane Q.

To determine whether consideration of diffraction around obstacles is required, we must determine whether an obstacle is within the first Fresnel zone. The Fresnel zone (Fig. 7b) for the transmitted wave is defined as the locus of the points for which the difference in the direct line-of-sight path TR and the indirect path TPR differ by an odd number of half-wavelengths. These 3D (three-dimensional) loci are ellipsoids (Fig. 7c). The circular rings on plane Q illustrated in Fig. 7b correspond to these half-wavelength pathlength differences.

From Fig. 7, with the assumptions that $h \ll d_{TP}$, d_{PR}, the Fresnel zones correspond to the condition

$$\Delta L = \sqrt{d_{TP}^2 + h^2} + \sqrt{d_{PR}^2 + h^2} - d_{TR} \simeq \frac{1}{2}h^2\left(\frac{1}{d_{TP}} + \frac{1}{d_{PR}}\right) = n\frac{\lambda_0}{2} \tag{18}$$

where n is any integer other than zero. The first Fresnel zone is that for which $n = 1$. It follows from (18) that

$$h_{cl} = \sqrt{\frac{\lambda_0 d_{TP} d_{PR}}{d_{TP} + d_{PR}}} \tag{19}$$

where h_{cl} is the "Fresnel zone clearance" height, that is, the minimum height of the transmitter and receiver (assumed to be the same in this simplified analysis) above the top of an obstacle such that line-of-sight conditions are approximated.

To determine the field strength in the shadow zone of an obstacle, diffraction theory must be invoked. Given the wedge shown in Fig. 7a as a simple example of such an obstacle, a field component at R behind the wedge is given approximately in Ref. [8], (pp. 402–406):

$$E_R = \int_{-\infty}^{\infty} dy' \int_{h_w}^{\infty} dz' A(y', z') e^{j\Delta\phi(y', z')} \tag{20}$$

where $A(y', z')$ and $\Delta\phi(y', z')$ are the amplitude and phase (relative to the phase of the straight-line raypath TR) of the field at a point (y', z') on the plane P. If we model the effect as two-dimensional (i.e., uniform in the y direction) and note that $d_1 + d_2 \gg |h_R - h_T|$, after some analysis, we arrive at the expression

$$E_R \simeq \frac{A_0}{\eta} \int_X^{\infty} du\, e^{j\left(\frac{\pi}{2}\right)\mu^2} \tag{21}$$

where $X = \eta[(h_w - h_{av}) + (\Delta h/2)\xi]h_{av} = (h_T + h_R)/2$, $\Delta h = h_R - h_T$

$$\eta = \sqrt{\frac{2}{\lambda_0}\left(\frac{d_1 + d_2}{d_1 d_2}\right)}, \quad \xi = \left(\frac{d_2 - d_1}{d_2 + d_1}\right)$$

Evaluation of the Fresnel integral in (21) leads to the diffraction loss in decibels as a function of the parameter X, shown in Fig 8. That loss is the ratio of field strength at R with the obstacle present to that if the obstacle were not present (i.e., the line-of-sight case). By assigning values to the parameters d_1, d_2, λ_0, h_w, h_{av}, and Δh, calculating X with these values, we can determine the loss in decibels corresponding to that value of X on the curve of Fig 8.

Knife-edge diffraction, as described above, provides a very rough estimate of field amplitude in the shadow zone of an obstacle such as a hill or a building along the propagation path. For realistic terrain, there are usually many vertical protrusions along the path and they are *not* usually simple wedges. There are methods of analysis that

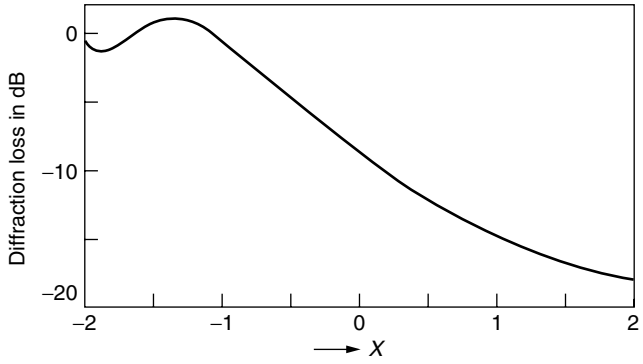


Figure 8. Diffraction loss versus parameter X in Eq. (21).

deal with more complicated diffraction models, summarized by Parsons [4] in 1992 and by Rappaport [5] in 1996.

2.6. Scattering

Scattering from objects, both natural (e.g., trees) and human-made (e.g., distant buildings, particularly in urban environments), occurs in most propagation paths. To differentiate scattering from earth reflection as discussed in Section 2.2, the latter is primarily specular and coherent; that is, the relative phase between direct and reflected waves is mostly deterministic, and hence their superposition contains peaks and troughs due to constructive and destructive interference, respectively. Scattering is from an object sufficiently small and far enough away from both T and R so that it subtends a negligibly small angle with respect to either T or R, appearing as a point when viewed from either site. The electric field at the receiver is determined from the bistatic radar equation [see derivation in, e.g., Ref. 8, (pp. 17–20)] and has the general form

$$E_S^{(V,H)} = \frac{e^{-jk_0(D_{TP}+D_{PR})}}{4\pi D_{TP}D_{PR}}\sqrt{P_T G_{T0} A_{eR0}\sigma_0}\, f_T^{(V,H)}(\Delta\boldsymbol{\Omega}_{TP}) \times f_R^{(V,H)}(\Delta\boldsymbol{\Omega}_{RP}) g^{(V,H)} \frac{\Delta\boldsymbol{\Omega}_{TP}}{\Delta\boldsymbol{\Omega}_{PR}} \tag{22}$$

where the subscript P denotes the location of the scatterer and where the quantities in (22) were all defined following Eq. (1), except σ_0, the peak value of the scattering cross section of the object and $g^{(V,H)}(\Omega_{TP}/\Omega_{PR})$, the angular complex field pattern of scattering from P into the direction of R given a wave originating at T and incident on P. An important feature of the scattered field is that the received power, proportional to $|E_S|^2$, is inversely proportional to the square of $D_{TP}D_{PR}$. This implies that it is usually small compared with the direct field as given by Eq. (1) or the specularly reflected field as obtained from Eq. (3). Hence, in order for this effect to be important in an outdoor communication link, there must be a large number of strong scatterers whose superposed fields generate a signal comparable in magnitude to the received signal. However, in urban or indoor environments, there may be large numbers of scatterers, such as buildings in urban scenes or particularly metallic objects within a room that compete in magnitude with direct signals and whose relative phases generate spatial peaks and troughs, the latter appearing as "dead zones" if the receiver is placed at their locations.

2.7. Tropospheric Scatter

A form of communication mode that was popular in the 1950s and 1960s was "tropospheric scatter" or "troposcatter" [7, pp. 418–453], based on scattering of the transmitted wave from random inhomogeneities in permittivity within a small tropospheric region where transmitter and receiver antenna beams intersect, such that significant power is scattered into the direction of the receiver. The advantage of this propagation mode is that it provides a clear transmission path well above the terrain, circumvents the problem of obstacles along the path, and greatly extends the effective radio horizon. However, satellite links provide those same features with greater reliability and for much longer propagation paths and have largely replaced troposcatter links within recent years.

2.8. Diffraction Around the Earth

In very long-range communication links, where the receiver is beyond the radio horizon of the transmitter, nonzero signal levels can be attained through the mechanism of diffraction around the earth. The theory behind "diffraction zone" propagation is developed in great detail in Ref. 1 (pp. 109–112), and the results are summarized in Ref. 6 (pp. 369–372).

2.9. Attenuation

An attenuation factor of the form $e^{-\alpha D}$ may be present in the electric or magnetic field of a propagating wave, where α is a positive real number measured in nepers per meter and D is the propagation pathlength. It is usually expressed in decibels: $20\log_{10} e^{-\alpha D} = -20(\log_{10} e)\alpha D = 8.686\alpha D$. The important number is $\gamma = 8686\alpha$, the attenuation in decibels per kilometer.

The real part of the refractive index of perfectly dry air is nearly unity, but if there is some degree of moisture content of the air, an imaginary part exists and gives rise to attenuation ranging from the order of 10^{-4} dB/km at 100 MHz to about 10^{-1} dB/km at 30 GHz. Hence within the frequency range of interest attenuation due to atmospheric gases is seldom important over the short pathlengths characteristic of mobile wireless links, even at the resonance peak that occurs at 22.24 GHz. Attenuation due to scattering from and absorption by the tiny water droplets that constitute a very dense fog may be significant at frequencies in the 20–30-GHz region if the density of water droplets is sufficiently high.

Raindrops are another source of attenuation, significant at frequencies above 10 GHz. With sufficiently high rainfall rates, the drops are large enough compared to wavelength to absorb energy in a propagating wave and to scatter it in directions other than that of propagation. Both of these mechanisms result in an exponential decay in wave amplitude, of the order of 0.01 dB/km in a light drizzle to as much as 3–4 dB/km in a very heavy rainfall at 20 GHz and possibly 6–10 dB/km at 30 GHz. Snow particles exhibit similar characteristics. The effects of precipitation on a

wireless link can sometimes be severe at the high end of the microwave region. The details of the theoretical background and key results on atmospheric attenuation in general can be found in Ref. 1 (Chap. 8, pp. 641–692) and on attenuation due to precipitation on in the same work [1, pp. 671–692]. This is still an authoritative source on the subject for engineers requiring data for design purposes.

Another source of attenuation is foliage along the propagation path in a forested environment. The attenuation is due to scattering of wave energy by trees and blockage of the raypath by large aggregates of trees. Since this is a highly specialized situation for wireless links, it will not be discussed further here.

3. PATH LOSS PREDICTION MODELS

A number of models have been used to determine path loss in wireless communication links. Some of these are physics-based and account for the propagation effects discussed above, namely, earth reflection, atmospheric refraction, effects of earth curvature and surface roughness, and diffraction losses due to obstacles in the propagation path. The last of these effects is so important in determination of path loss in real-world environments that it has received enormous emphasis in research on VHF and UHF propagation.

Since propagation environments are often too complex to model physically, some models have been developed from empirical data. Some of these empirical models are widely applied to loss computations for paths along irregular terrain, urban areas, and indoor environments. In what follows, some of the analytical and empirical models will be briefly summarized or at least mentioned with references to the literature for details.

3.1. Analytical Models

The analytical models differ primarily in the degree of complexity in accounting for diffraction. One of the earliest (1947) models, that of Bullington [10], accounted for two knife-edge-type obstacles along the path and set up a diffraction problem for an equivalent single knife edge. This was an attempt to model more than one obstacle. Diffraction from multiple obstacles was later (1953) treated by Epstein and Peterson [11] and much later by Deygout [12] and Longley and Rice [13], Edwards and Durkin [14], and others [15–17]. The Longley–Rice model has been widely used in the United States for irregular terrain modeling and was improved between 1967 and 1985, including a partially empirical extension to urban areas [18]. Covering frequencies from 40 to 100 GHz, the model included most of the effects discussed above — ground reflection, earth curvature effects, diffraction from isolated obstacles, and both tropospheric scatter and earth diffraction for very long propagation paths [4, pp. 57–61; 19].

Another model and associated computer program similar to Longley–Rice, is due to Edwards and Durkin [14] and Dadson et al. [20]. This method was adopted by the Joint Radio Committee (JRC) in the United Kingdom and has been widely used there.

Models like those indicated above are two-dimensional, confined to the vertical propagation plane, and can roughly predict path losses due to ground reflection and diffraction from widely separated obstacles within that plane. But they cannot account adequately for scattering and diffraction due to natural and synthetic terrain features very close to the receiver, and diffraction from obstacles that are close together (and hence interact with each other in a complicated manner). Finally, they do not account for three dimensional effects, such as the multipath arising from constructive and destructive interference between various scatterers distributed horizontally along the terrain both on and off the vertical propagation plane. The rapid fading prevalent in mobile communication links is due largely to these effects.

Another avenue of research to improve the realism of diffraction models is to consider rounded edges, more typical of real hills than the knife edge [e.g., 2, pp. 129,130; 4, pp. 45–47]. Still another is to use geometric theory of diffraction (GTD) or unified theory of diffraction (UTD) in lieu of knife-edge diffraction. There is a body of literature on these latter topics in the context of radio propagation modeling [e.g., 21–23].

3.2. Empirical Models

A set of empirical models primarily designed for urban areas was developed by Okumura et al. [24], using a series of measurements in and near Tokyo, and empirical equations to fit the data, valid from 150 MHz to 1.5 GHz were developed by Hata [25]. Others have since contributed to these kind of models. Physics-based analytical modeling is often difficult for complicated urban scenes. Empirical models, although based on experiments done in specific locations and therefore less flexible, can be useful in development of wireless links.

3.3. Models for Outdoor Urban Areas

Urban areas present a challenging problem in development of path loss models. A model for a scene consisting of a set of buildings arranged deterministically in a horizontal rectangular array is characteristic of an urban area. The major effects contributing to path loss are not necessarily in the vertical plane but usually require three-dimensional modeling. Typically, the direct line of sight between T and R is rarely available. The available paths are usually those involving multiple reflections between buildings and street surfaces and diffraction around buildings in both horizontal and vertical planes. A diffraction model accounting for some of these effects was developed by Walfisch and Bertoni in 1988 [26]. Three-dimensional ray-tracing models have been developed since that time to treat these kinds of urban geometries [e.g., 27–29]. With the increases in computer power that have occurred since the early 1990s, it is possible to compute the path loss in very complicated urban environments within reasonable CPU times.

3.4. Models for Indoor Propagation

With very rapid increases in the use of cell phones since the late 1990s, it has become important to understand propagation within buildings and between sites in different

buildings and to predict path losses for such situations. In this case, the important effects are (1) losses in transmission through walls; (2) multiple reflections between floors walls, and ceilings within rooms; (3) scattering from objects within a room; and (4) diffraction around objects within the direct path between transmitter and receiver. Rappaport [5, pp. 123–132] presents a particularly good summary of indoor propagation models and considerable data, much of it empirical, on losses through various materials found in buildings at various frequencies. Subsequent work has been done by many researchers on indoor propagation modeling and simulation [e.g., 30–32].

BIOGRAPHY

Harold R. Raemer received his Ph.D degree in physics from Northwestern University, Evanston, Illinois in 1959. From 1952 to 1963 he was a research engineer in industrial laboratories, performing analytical studies on problems in radio wave propagation and radar and communication systems. In 1963, he joined the faculty of the Electrical Engineering Department at Northeastern University, Boston, Massachusetts, as an associate professor. He became professor in 1966 and served as chair of the department from 1967 to 1977, and later as acting chair from 1982 to 1984. From 1984 to 1993 he was associate director of radio frequency phenomena and systems at the Center for Electromagnetics Research at Northeastern. He retired from the faculty in 1994 but has remained at the university to the present as a professor emeritus. During his career at Northeastern, he taught undergraduate and graduate courses in a number of subjects within the EE curriculum, conducted sponsored research in plasma dynamics, radio wave propagation, and later in simulation of propagation and scattering in radar systems and wireless communication systems. He is the author of two books and a number of papers in research journals and conference proceedings.

BIBLIOGRAPHY

1. D. E. Kerr, ed., *Propagation of Short Radio Waves*, McGraw-Hill, New York, 1951.
2. J. Griffiths, *Radio Wave Propagation*, McGraw-Hill, New York, 1987.
3. L. Boithias, *Radio Wave Propagation*, McGraw-Hill, New York, 1987.
4. D. Parsons, *The Mobile Radio Propagation Channel*, Wiley, New York, 1992.
5. T. Rappaport, *Wireless Communications*, Prentice-Hall, Upper Saddle River, NJ, 1996.
6. R. E. Collin, *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985, Chap. 6.
7. P. Beckmann and A. Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces*, Artech, Norwood, MA, 1987.
8. H. Raemer, *Radar Systems Principles*, CRC Press, Boca Raton, FL, 1997.
9. S. Ayasli and M. B. Carlson, SEKE: *A Computer Model for Low Altitude Radar Propagation over Irregular Terrain*, Project Report CMT-70, MIT Lincoln Laboratory, Lexington, MA, May 1, 1985.
10. K. Bullington, Radio propagation at frequencies above 30Mc, *Proc. IRE* **35**(10): 1122–1136 (Oct. 1947).
11. J. Epstein and D. W. Peterson, An experimental study of wave propagation at 850Mc, *Proc. IRE* **41**(4): 595–611 (May 1953).
12. J. Deygout, Multiple knife-edge diffraction of microwaves, *IEEE Trans. Antennas Propag.* **AP-14**(4): 480–489 (July 1966).
13. A. G. Longley and P. L. Rice, *Prediction of Tropospheric Radio Transmission Loss over Irregular Terrain, a Computer Method*, ESSA Technical Report, ERL 79-ITS67, 1968.
14. R. Edwards and J. Durkin, Computer prediction of service area for VHF mobile radio networks, *Proc. IEE* **116**(9): 1493–1500 (Sept. 1969).
15. C. L. Giovaneli, An analysis of simplified solutions for multiple knife-edge diffraction, *IEEE Trans. Antennas Propag.* **AP-32**(3): 297–301 (March 1984).
16. L. E. Vogler, An attenuation function for multiple knife-edge diffraction, *Radio Sci.* **17**(6): 1541–1546 (Nov.–Dec. 1982).
17. J. H. Whitteker, A series solution for diffraction over terrain modeled as multiple bridged knife-edges, *Radio Sci.* **28**: 487–500 (July–Aug. 1993).
18. A. G. Longley, *Radio Propagation in Urban Areas*, Office of Telecommunications (OT) Report, April 1978, pp. 78–144.
19. IEEE Vehicular Technology Society Committee on Radio Propagation, Coverage prediction for mobile radio systems operating in the 800/900 MHz frequency range, *IEEE Trans. Vehic. Technol.* **VT-37**(1): 3–72 (Feb. 1988).
20. C. E. Dadson, J. Durkin, and E. Martin, Computer prediction of field strength in the planning of radio systems, *IEEE Trans. Vehic. Technol.* **VT-24**(1): 1–7 (Feb. 1975).
21. R. J. Luebbers, Finite conductivity uniform GTD versus knife-edge diffraction in prediction of propagation path loss, *IEEE Trans. Antennas Propag.* **AP-32**: 70–76 (Jan. 1984).
22. S. Y. Tan and H. S. Tan, UTD propagation model in an urban street scene for micro-cellular communications, *IEEE Trans. Electromagn. Compat.* **EC-37**: 423–428 (Nov. 1993).
23. S. Y. Tan and H. S. Tan, Propagation model for micro-cellular communications in Ottawa city streets, *IEEE Trans. Vehic. Technol.* **VT-44**: 313–317 (May 1995).
24. Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, Field strength and its variability in the VHF and UHF and mobile radio service, *Rev. Electron. Commun. Lab.* **16**(9–10): 825–873 (Sept.–Oct. 1968).
25. M. Hata, Empirical formula for propagation loss in land mobile radio service, *IEEE Trans. Vehic. Technol.* **VT-29**(3): 317–325 (Aug. 1980).
26. S. Walfisch and H. L. Bertoni, A. theoretical model for VHF propagation in urban environments, *IEEE Trans. Antennas Propag.* **AP-36**: 1788–1796 (Oct. 1988).
27. F. Ikegami, S. Yoshida, T. Takeuchi, and M. Umehira, Propagation factors controlling mean field strength on urban streets, *IEEE Trans. Antennas Propag.* **AP-32**(8): 822–829 (Aug. 1984).
28. A. J. Rustako, N. Amitay, G. J. Owens, and R. S. Roman, Radio propagation at microwave frequencies for line-of-sight

microcellular mobile and personal communications, *IEEE Trans. Vehic. Technol.* **VT-40**(1): 203–210 (Feb. 1991).

29. G. Liang and H. L. Bertoni, A new approach to 3-D ray tracing for propagation prediction in cities, *IEEE Trans. Antennas Propag.* **AP-46**(6): 853–863 (June 1998).
30. U. Dersch and E. Zollinger, Propagation mechanisms in micro-cell and indoor environments, *IEEE Trans. Vehic. Technol.* **VT-43**: 1058–1066 (Nov. 1994).
31. C. F. Yang, B. C. Wu, and C. J. Ko, A ray-tracing method for modeling indoor wave propagation and penetration, *IEEE Trans. Antennas Propag.* **AP-46**(6): 907–919 (June 1998).
32. K. W. Chang, J. H. M Sau, and R. D. Murch, A new empirical model for indoor propagation prediction, *IEEE Trans. Vehic. Technol.* **VT-47**(3): 996–1001 (Aug. 1998).

AUTHENTICATION CODES

THOMAS JOHANSSON
Lund University
Lund, Sweden

1. INTRODUCTION

The protection of unauthorized access to sensitive information has been a prime concern throughout the centuries. Still, it was not until Shannon's work in the late 1940s [11] a theoretical model for secrecy was developed. Shannon's work was based on the concept of unconditional security, by which we mean that the enemy faced is assumed to have access to infinite computing power. Under this assumption Shannon developed some rather pessimistic results on the requirements for a cryptosystem to be secure.

More recently, we have understood that one usually needs to protect data not only against unauthorized access but also against unauthorized modifications. In a communication situation, we need to *authenticate* our transmitted messages. We need to check that they are indeed sent by the claimed sender and that they have not been modified during transmission. The threat from the enemy can be viewed as "intelligent noise," the noise taking the worst possible value for the sender and receiver. This means that error correcting codes will not help (because the noise just changes a transmitted codeword to another codeword), but we must introduce secret *keys* that are known to the sender/receiver but unknown to the enemy.

Authentication of transmitted messages can be done in (at least) three fundamentally different ways. We refer to them as *unconditionally secure authentication codes, message authentication codes*, and *digital signatures*.

Unconditionally secure authentication codes is the only solution to the authentication problem for an enemy with unlimited computing power. As this is the topic of the article, we continue the discussion in the next section.

Message authentication codes (MACs) refer to authentication techniques that use symmetric cryptographic primitives (i.e., block ciphers and hash functions) to provide authentication. As for unconditionally secure authentication codes, the sender and receiver are here assumed to share a common secret key. MACs is a very common authentication technique in, for example, banking transactions. MACs appear in many standards, and some common modes of operations for block ciphers provide MACs. We refer to Ref. 9 for more details. Comparing with unconditionally secure authentication codes, MACs are not secure against an unlimited enemy. But they have other practical advantages, such as being able to authenticate many messages without changing the key.

Finally, digital signatures is an asymmetric solution. This means that the sender has a personal secret signing key and the receiver has access to a corresponding public verification key. The sender first hashes the message to be transmitted using a cryptographic hash function. The result is then signed by a signature scheme. Common hash functions are MD5 and SHA-1, and common signature schemes are RSA and DSA. See Ref. 9 for more information. Digital signatures possess several advantages compared to the other two authentication techniques. Since it is an asymmetric technique, there is no need to distribute or establish a common secret key between the sender and the receiver. Basically, the sender generates his/her secret signing key and the corresponding public verification key. The verification key can then be presented in public. This means that anyone can verify the authenticity of a message. This leads to the second important difference, referred to as nonrepudiation. Since the sender is the only person able to generate an authentic message (the receiver cannot), we know that if a message is authentic, it must have been generated by the sender. If the receiver has received an authentic message, the sender cannot deny having sent it. This somewhat resembles a handwritten signature, once you have signed you cannot later deny having signed. There are also drawbacks. Signature schemes rely on the hardness of problems, such as factoring and taking discrete logarithms. This means that we must work with very large numbers, which make the solutions slow compared to the other techniques, especially for short messages.

2. AUTHENTICATION CODES

An unconditionally secure solution to the authentication problem first appeared in 1974 when Gilbert, MacWilliams, and Sloane published their landmark paper "Codes which detect deception" [4]. As mentioned in that paper, Simmons was independently working with the same problems. In the early 1980s Simmons published several papers on the subject that established the authentication model, [12,13,15]. Simmons work on authentication theory has a similar role as Shannons work on secrecy.

This section deals with unconditionally secure authentication codes. We provide some fundamental definition and results. We also include some common constructions. We start by presenting the mathematical model of unconditionally secure authentication due to Simmons.

The communication model for authentication includes three participants, *the transmitter, the receiver*, and *the opponent*. The transmission from the transmitter to the receiver takes place over an insecure channel. The opponent, who is the enemy, has access to the channel in the sense that it can insert a message into the channel, or alternatively, observe a transmitted message and then

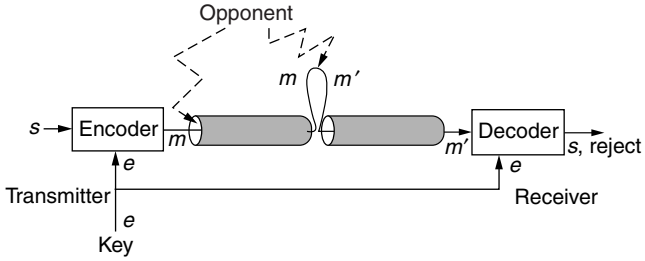


Figure 1. The authentication model.

replace it with another message. The authentication model is illustrated in Fig. 1.

The information that the transmitter wants to send is called a *source message*, denoted by s and taken from the finite set $\mathcal{S}$ of possible source messages. The source message is mapped into a (channel) *message*, denoted by m and taken from the set $\mathcal{M}$ of possible messages. Exactly how this mapping is performed is determined by the secret *key*, which is denoted by e and taken from the set $\mathcal{E}$ of possible encoding rules. The key is secretly shared between the transmitter and the receiver.

Each key determines a mapping from $\mathcal{S}$ to $\mathcal{M}$. Equivalently, the encoding process can be described by the mapping f, where

$$f: \mathcal{S} \times \mathcal{E} \to \mathcal{M}, (s, e) \mapsto m \tag{1}$$

An important property of f is that if $f(s, e) = m$ and $f(s', e) = m$, then $s = s'$ (injective for each $e \in \mathcal{E}$). Two different source messages cannot map to the same message for a given encoding rule, since then the receiver would not be able to determine which source message was transmitted. The mapping f together with the sets $\mathcal{S}, \mathcal{M}$, and $\mathcal{E}$ define an *authentication code* (A-code).

When the receiver receives a message m, it must check whether a source message s exists, such that $f(s, e) = m$. If such an s exists, the message m is accepted as authentic (m is called "valid"). Otherwise, m is not authentic and thus rejected. We can assume that the receiver checks $f(s, e)$ for all $s \in \mathcal{S}$, and if it finds $s \in \mathcal{S}$ such that $f(s, e) = m$, it outputs s; otherwise it outputs a reject signal.

The opponent has two possible attacks at its disposal: the impersonation attack and the substitution attack. The *impersonation attack* simply means inserting a message m and hoping for it to be accepted as authentic. In the *substitution attack*, the opponent observes the message m and replaces this with another message $m', m \neq m'$, hoping for m' to be valid.

We assume that the opponent chooses the message that maximizes its chances of success when performing an attack. The probability of success in each attack is denoted by P_I and P_S, respectively. They are more formally defined by[1]

$$P_I = \max_m P \quad (m \text{ is valid}) \tag{2}$$

[1] We abbreviate expressions such as $\max_{m \in \mathcal{M}}$ as $\max_m$ when no possibility of confusion occurs.

and

$$P_S = \max_{\substack{m, m' \\ m \neq m'}} P \quad (m' \text{ is valid} \mid m \text{ is valid}) \tag{3}$$

Note that this definition considers only transmission of a single message. For transmission of multiple messages, we must introduce a more general definition of the deception probabilities.

Continuing, we define the *probability of deception* P_D as $P_D = \max(P_I, P_S)$. It is convenient to define $\mathcal{E}(m)$ as the set of keys for which a message m is valid:

$$\mathcal{E}(m) = \{e \in \mathcal{E}; \exists s \in \mathcal{S}, f(s, e) = m\} \tag{4}$$

Let us now derive some basic properties for authentication codes. We see that of all the messages in $\mathcal{M}$, at least $|\mathcal{S}|$ must be authentic, since every source message maps to a different message in $\mathcal{M}$. Similarly, for the substitution attack, after the observation of one legal message, at least $|\mathcal{S}| - 1$ of the remaining $|\mathcal{M}| - 1$ messages must be authentic. Thus we have two obvious bounds.

Theorem 1. For any authentication code

$$P_I \geq \frac{|\mathcal{S}|}{|\mathcal{M}|} \tag{5}$$

$$P_S \geq \frac{|\mathcal{S}| - 1}{|\mathcal{M}| - 1} \tag{6}$$

From Theorem 1 we observe two fundamental properties of authentication codes. First, in order to ensure good protection $|\mathcal{M}|$, must be chosen much larger than $|\mathcal{S}|$. This affects the message expansion of our authentication code. For a fixed source message space, an increase in the authentication protection implies an increased message expansion. The second property is that a complete protection (i.e., $P_D = 0$) is not possible. We must be satisfied with a protection where P_D is small.

For example, let an authentication code with $\mathcal{S} = \{H, T\}$, $\mathcal{M} = \{1, 2, 3, 4\}$ and $\mathcal{E} = \{0, 1, 2, 3\}$ be described by the following table

		m			
	s	1	2	3	4
	0	H	T	—	—
e	1	T	—	H	—
	2	—	H	—	T
	3	—	—	T	H

It is easy to verify that $P_I = P_S = \frac{1}{2}$ if the keys are uniformly distributed.

We assume that the reader is familiar with the basic concepts of information theory. As usual, $H(X)$ denotes the entropy of the random variable X, and $I(X; Y)$ denotes the mutual information between X and Y. We are now ready to state the next fundamental result in authentication theory, namely, Simmons' bounds.

Theorem 2: Simmons' Bounds. For any authentication code, we obtain

$$P_I \geq 2^{-I(M;E)}, \tag{7}$$

$$P_S \geq 2^{-H(E|M)}, \quad \text{if } |\mathcal{S}| \geq 2. \tag{8}$$

The bound for the impersonation attack was first proved by Simmons in 1984 with a long and tedious proof [13]. Several new and much shorter proofs have been given since then. The bound for the substitution attack was proved by Simmons and Brickell in [2], but can also be proved in the same way as for the impersonation attack.

Simmons' bounds give a good feeling of how the authentication protection affects the system. For the impersonation attack, we see that P_I is upper-bounded by the mutual information between the message and the key. This means that for a good protection (i.e., P_I small), we must give away a lot of information about the key. On the other hand, in the substitution attack, P_S is lower-bounded by the uncertainty about the key when a message has been observed. Thus we cannot waste all the key entropy for protection against the impersonation attack, but some uncertainty about the key must remain for protection against the substitution attack.

Returning to Theorem 2, we multiply the two bounds together and get

$$P_I P_S \geq 2^{-I(M;E)-H(E|M)} = 2^{-H(E)} \tag{9}$$

From the inequality $H(E) \leq \log|\mathcal{E}|$ we then obtain the *square-root bound*.

Theorem 3: Square-Root Bound. For any authentication code, we have

$$P_D \geq \frac{1}{\sqrt{|\mathcal{E}|}} \tag{10}$$

The bound was originally proved in [4] under other conditions and slightly differently stated. The square root bound gives a direct relation between the key size and the protection that we can expect to obtain. Thus the following definitions are natural.

An authentication code for which equality holds in the square-root bound (10) is called a "perfect" A-code.[2] Furthermore, an A-code for which $P_I = P_S$ is called an "equitable" A-code.

Obviously, a perfect A-code must be equitable. If we can construct A-codes for which equality holds in the square root bound we can be satisfied, since in that case no better authentication codes exist, in the sense that P_D cannot be made smaller. This is a main topic, but also equitable A-codes that are not perfect are of interest. The reason for this is the following.

Theorem 4. The square-root bound (10) can be tight only if

$$|\mathcal{S}| \leq \sqrt{|\mathcal{E}|} + 1$$

The result was discussed in Ref. 4.

The square-root bound motivates a treatment of nonperfect A-codes, since for perfect A-codes a large source size demands a twice as large key size. This is not very practical. On the other hand, if the source size is very modest (i.e., $|\mathcal{S}| \leq \sqrt{|\mathcal{E}|} + 1$), then we will see in the sequel that perfect A-codes can be constructed for any $P_D = 1/q$, where $q = \sqrt{|\mathcal{E}|}$ is a prime power.

The most important kind of authentication code is when the source message s appear as a part of the channel message m. An A-code for which the map $f: \mathcal{S} \times \mathcal{E} \to \mathcal{M}$ can be written in the form

$$f: \mathcal{S} \times \mathcal{E} \to \mathcal{S} \times \mathcal{Z}, \quad (s, e) \mapsto (s, z) \tag{11}$$

where $s \in \mathcal{S}, z \in \mathcal{Z}$ is called a *systematic* (or *Cartesian*) *A-code*. The second part z in the message is called the "tag" (or authenticator) and is taken from the tag alphabet $\mathcal{Z}$. We see that systematic A-codes are codes that have no secrecy at all, the source message is transmitted in the clear, and we add some check symbols to it (the tag). In the sequel we study only systematic authentication codes. Systematic A-codes have the following important property [6].

Theorem 5. For any systematic A-code

$$P_S \geq P_I \tag{12}$$

This means that for systematic A-codes the square-root bound is expressed as

$$P_S \geq \frac{1}{\sqrt{|\mathcal{E}|}}$$

Finally, for systematic A-codes with uniformly distributed keys and $P_I = P_S = 1/|\mathcal{Z}|$, we have the inequality [6]

$$(|\mathcal{Z}| - 1)|\mathcal{S}| \leq |\mathcal{E}| - 1 \tag{13}$$

This bound shows that large source sizes for equitable A-codes require large key sizes, and gives the motivation for the study of nonequitable A-codes.

We next present some ways of constructing equitable A-codes. Equitable A-codes have the lowest possible probability of deception in the sense that $P_D = P_I = P_S$, but they have the disadvantage of having a source size that is quite modest. It is useful to note that the probability of success in a substitution attack can be written as

$$P_S = \max_{\substack{m,m' \\ m \neq m'}} \frac{|\mathcal{E}(m) \cap \mathcal{E}(m')|}{|\mathcal{E}(m)|} \tag{14}$$

provided that the keys are uniformly distributed.

To have some measure of how good a construction is, we introduce two fundamental definitions. An A-code with fixed parameters $|\mathcal{E}|, |\mathcal{M}|, |\mathcal{S}|$, and P_I is said to be weakly optimal if P_S is the lowest possible. A weakly optimal A-code is said to be strongly optimal if, additionally, $|\mathcal{S}|$ has the largest possible value among all the weakly optimal A-codes for fixed parameters $|\mathcal{E}|, |\mathcal{M}|, P_I$.

We start by giving the original *projective plane construction* proposed by Gilbert et al. [4]. Fix a line L in $\mathbf{PG}(2, \mathbb{F}_q)$. The points on L are regarded as source messages, the points not on L are regarded as keys, and the lines distinct from L are regarded as messages. The mapping from $\mathcal{S}$ to $\mathcal{M}$ means joining the source message s and the key e to the unique line m, which is the resulting message.

[2] The definition of the terminology perfect A-code may be different in other literature.

We can easily verify correctness of this construction. The joining of the point e outside L and the point s on L results in a unique line, called m. By running through all pairs (s, e), we find the message space as all lines except L itself. The parameters of the A-code are given by the following theorem.

Theorem 6. The projective plane construction gives parameters

$$|\mathcal{S}| = q + 1, \quad |\mathcal{M}| = q^2 + q, \quad |\mathcal{E}| = q^2$$

and the probabilities of success are $P_I = 1/q$ and $P_S = 1/q$.

The A-codes resulting from this construction are strongly optimal.

Another simple construction is the *vector space construction*. Let $|\mathcal{S}| = q^m, |\mathcal{Z}| = q^m$, and $|\mathcal{E}| = q^{2m}$. Decompose the keys as $e = (e_1, e_2)$, where $s, z, e_1, e_2 \in \mathbb{F}_{q^m}$. For transmission of source message s, generate a message $m = (s, z)$, where

$$z = e_1 + se_2$$

Theorem 7. The above construction described provides $P_I = P_S = 1/q^m$. Moreover, it has parameters $|\mathcal{S}| = q^m, |\mathcal{Z}| = q^m$, and $|\mathcal{E}| = q^{2m}$.

Authentication codes are closely related to combinatorial designs. This relation has been extensively examined in a number of papers. Brickell [2], and later Stinson [16], established a one-to-one correspondence between A-codes with given parameters and certain combinatorial designs. The designs used include transversal designs, orthogonal arrays, balanced incomplete block designs, and perpendicular arrays; see Ref. 17 for an introduction.

3. CONSTRUCTING USEFUL AUTHENTICATION CODES

Our treatment so far has not considered any results of interest for the case when $|\mathcal{S}|$ is large. This case is of great relevance since many practical problems concern very large source sizes. Examples of such problems are authentication of data files or computer programs. We now turn our attention to the problem of solving the authentication problem for sources that have a length of, say, a million bits. This means $\log |\mathcal{S}| = 10^6$.

So, a fundamental problem in authentication theory is to find A-codes such that $|\mathcal{S}|$ is large while keeping $|\mathcal{E}|$ and P_S as small as possible. In many practical situations one has limitations on $|\mathcal{E}|$ and wants P_S to be bounded by some small value. Also, one usually wants the redundancy ($|\mathcal{Z}|$) to be small, since it occupies a part of the bandwidth.

In an earlier study [6] it was shown that authentication codes have a close connection to coding theory. It is, for example, possible to construct authentication codes from error-correcting codes and vice versa. Some of the resulting constructions are illustrated next.

First, we give a construction based on Reed–Solomon codes1,6.

Construction is as follows. Let $\mathcal{S} = \{\mathbf{s} = (s_1, \ldots, s_k); s_i \in \mathbb{F}_q\}$. Define the source message polynomial to be $s(x) = s_1x + s_2x^2 + \cdots + s_kx^k$. Let $\mathcal{E} = \{e = (e_1, e_2); e_1, e_2 \in \mathbb{F}_q\}$ and $\mathcal{Z} = \mathbb{F}_q$. For the transmission of source message $\mathbf{s}$, the transmitter sends $\mathbf{s}$ together with the tag

$$z = e_1 + s(e_2)$$

Theorem 8. The construction gives systematic A-codes with parameters

$$|\mathcal{S}| = q^k, \quad |\mathcal{E}| = q^2, \quad |\mathcal{Z}| = q, \quad P_I = 1/q, \quad P_S = k/q.$$

The construction gives weakly optimal A-codes.

In order to make $|\mathcal{S}|$ really large, here is a final construction from [7] based on dual BCH codes. Let q be a power of a prime p, and let $Tr_{q^m/q}$ denote the trace function from $\mathbb{F}_{q^m}$ to $\mathbb{F}_q$.

Construction is as follows. Let the set $\mathcal{F}_D$ of polynomials of degree $D \le \sqrt{q^m}$, be defined by

$$\mathcal{F}_D = \{f(x): f(x) = f_1x + f_2x^2 + \cdots + f_Dx^D \in \mathbb{F}_{q^m}[x], f_i = 0 \text{ whenever } p \mid i\}$$

Let $\mathcal{S} = \mathcal{F}_D, \mathcal{E} = \{(e_1, e_2): e_2 \in \mathbb{F}_{q^m}, e_1 \in \mathbb{F}_q\}$, and let the tag z be generated as

$$z = e_1 + Tr_{q^m/q}(f(e_2))$$

Theorem 9. The parameters for the systematic A-code are

$$|\mathcal{S}| = q^{m(D-\lfloor D/p \rfloor)}, |\mathcal{E}| = q^{m+1}, |\mathcal{Z}| = q$$
$$P_I = \frac{1}{q}, \; P_S = \frac{1}{q} + \frac{D-1}{\sqrt{q^m}}$$

We can look at the performance of this construction. For example, consider the parameters $P_S = 2^{-19}$ and $m = 4$. By choosing q to be a large prime around 2^{20}, we get $\log |\mathcal{S}| = 20 \cdot 4 \cdot \sqrt{2^{40}} = 80 \cdot 2^{20}$. With a 100-bit key, we can protect a source message of 10 MB (mega bytes)!

Further constructions using, for instance, algebraic geometric codes have appeared. The general topic here has been to optimize the parameters of the authentication code, such as for a fixed source message size and fixed security level (P_S) to find the authentication code using the smallest key size. Several bounds on this problem have also been derived [e.g., 6].

Authentication codes are also very closely related to universal hash functions. Universal classes of hash functions were introduced by Carter and Wegman [3], and it quickly became an established concept in computer science. It found numerous applications, such as cryptography, complexity theory, search algorithms, and associative memories, to mention only a few.

We end our treatment of authentication codes by mentioning a different line of research. Instead of optimizing the code parameters, one could focus on authentication codes with a very fast implementation. An interesting approach, called "bucket hashing," was introduced by

Rogaway [10]. These techniques were eventually refined, resulting in constructions like the UMAC [8].

4. AUTHENTICATION FOR TWO NONTRUSTING PARTIES

We now move to the study of authentication codes where the transmitter and receiver do not necessarily have to trust each other. In this situation we include deceptions from the insiders, like the transmitter sending a message and then later denying having sent it, or conversely, the receiver claiming to have received a message that was never sent by the transmitter. Recalling the short discussion on authentication techniques in the introduction, this is a first step toward digital signatures in the world of unconditional security, since it includes the nonrepudiation aspect.

In order to solve possible disputes we include a fourth participant, called the arbiter. The arbiter has access to all key information and, *by definition*, does not cheat. The arbiter is only present to solve possible disputes and does not take part in any communication activities.

Simmons introduced this extended authentication model, which is called the *authentication with arbitration model* [14], or simply the *A^2-model*. In this model, protection is provided against deceptions both from an outsider (opponent) and from the insiders (transmitter and receiver). Until 2001 or so, it was not known whether it was even possible to construct such schemes in an unconditional setting. Simmons gave a positive answer to this question.

We give a brief description of the A^2-model. For more details, we refer to a paper by Simmons [14], which contains a thorough discussion of the different threats. The A^2-model includes four different participants: the *transmitter*, the *receiver*, the *opponent*, and the *arbiter*. As in conventional authentication, the transmitter wants to send some information, called a *source message*, to the receiver in such a way that the receiver can both recover the transmitted source message and verify that the transmitted message originates from the legitimate transmitter. The source message $s \in \mathcal{S}$, is encoded by the transmitter into a message $m \in \mathcal{M}$. The message m is subsequently transmitted over the channel. The mapping from $\mathcal{S}$ to $\mathcal{M}$ is determined by the transmitter's secret key e_t chosen from the set $\mathcal{E}_T$ of possible keys for the transmitter. We may assume that the transmitter uses a mapping $f: \mathcal{S} \times \mathcal{E}_T \to \mathcal{M}$ such that

$$f(s, e_t) = f(s', e_t) \Rightarrow s = s' \tag{15}$$

In other words, the source message can be recovered uniquely from a transmitted channel message. The opponent has access to the channel in the sense that it can either impersonate the transmitter and send a message, or replace a transmitted message with a different one. The receiver must decide whether a received message is valid or not. For this purpose the receiver uses a mapping, determined by its own key e_r taken from the set $\mathcal{E}_\mathcal{R}$ of possible receiver's keys, which determines whether the message is valid, and if so, also the source message. This mapping is denoted $g: \mathcal{M} \times \mathcal{E}_\mathcal{R} \to \mathcal{S} \cup \{\text{reject}\}$, where for all possible (e_t, e_r), specifically, $P(e_t, e_r) \neq 0$, we have

$$f(s, e_t) = m \Rightarrow g(m, e_r) = s \tag{16}$$

For the receiver to accept all legal messages from the transmitter and to translate them to the correct source message, property (16) must hold for all possible pairs (e_t, e_r). However, in general not all pairs (e_t, e_r) will be possible.

The arbiter solves a possible dispute in the following way. If the channel message m, received by the receiver, could have been generated by the transmitter according to its encoding rule e_t, then the arbiter decides that the message m was sent by the transmitter, and otherwise not. The arbiter is assumed to be honest.

In the A^2-model the following five types of cheating attacks are considered.

Attack I: *impersonation by the opponent* — the opponent sends a message to the receiver and succeeds if this message is accepted by the receiver as authentic.

Attack S: *substitution by the opponent* — the opponent observes a message that is transmitted and replaces this message with another. The opponent is successful if this other message is accepted by the receiver as authentic.

Attack T: *impersonation by the transmitter* — the transmitter sends a message to the receiver and then denies having sent it. The transmitter succeeds if this message is accepted by the receiver as authentic, and if this message is not one of the messages that the transmitter could have generated according to its key.

Attack R_0: *impersonation by the receiver* — the receiver claims to have received a message from the transmitter. The receiver succeeds if this message could have been generated by the transmitter according to its key.

Attack R_1: *substitution by the receiver* — the receiver receives a message from the transmitter, but claims to have received another message. The receiver succeeds if this other message could have been generated by the transmitter according to his key.

In all possible attempts to cheat it is understood that the cheating person chooses the message that maximizes his/her chances of success. For the five possible deceptions, we denote the probability of success in each attack by P_I, P_S, P_T, P_{R_0} and P_{R_1}, respectively. The formal definitions are

$$P_I = \max_m P \quad (m \text{ valid}) \tag{17}$$

$$P_S = \max_{\substack{m, m' \\ m \neq m'}} P \quad (m' \text{ valid} \mid m \text{ valid}) \tag{18}$$

$$P_T = \max_{\substack{m, e_t \\ m \notin \mathcal{M}(e_t)}} P \quad (m \text{ valid} \mid e_t) \tag{19}$$

$$P_{R_0} = \max_{m, e_r} P \quad (m \in \mathcal{M}(e_t) \mid e_r) \tag{20}$$

$$P_{R_1} = \max_{\substack{m, m', e_r \\ m \neq m'}} P \quad (m' \in \mathcal{M}(e_t) \mid m \in \mathcal{M}(e_t), e_r) \tag{21}$$

where $\mathcal{M}(e_t)$ is the set of possible messages for the transmitter's encoding rule e_t, specifically, $\mathcal{M}(e_t) = \{m; f(s, e_t) = m, s \in \mathcal{S}\}$. Furthermore, let us define $P_D = \max(P_I, P_S, P_T, P_{R_0}, P_{R_1})$.

If the source messages are uniformly distributed, we have the following lower bounds on the number of encoding rules and on the number of messages [5].

Theorem 10

$$|\mathcal{E}_\mathcal{R}| \geq (P_I P_S P_T)^{-1}$$
$$|\mathcal{E}_\mathcal{T}| \geq (P_I P_S P_{R_0} P_{R_1})^{-1}$$
$$|\mathcal{M}| \geq (P_I P_{R_0})^{-1} |\mathcal{S}|$$

In particular, if $P_D = 1/q$, then

$$|\mathcal{E}_\mathcal{R}| \geq q^3, |\mathcal{E}_\mathcal{T}| \geq q^4, |\mathcal{M}| \geq q^2 |\mathcal{S}|$$

We end this brief discussion with an example for the simplest possible nontrivial case: $P_D = \frac{1}{2}$. Assume that there are two possible source messages, $\mathcal{S} = \{H, T\}$. The Cartesian product construction due to Simmons [14] gives rise to the matrix shown in Table 1. In the key setup, the receiver chooses (or gets from the arbiter) one of the 16 rows as the key E_R. Assume, for example, that the row e_1 will be the receiver's key. Then the receiver will accept the messages m_1, m_2, m_5, and m_6 as authentic. The messages m_1, m_2 will be interpreted as the source message H, and the messages m_5, m_6 will be interpreted as the source message T. All the other messages are not authentic, and will thus be rejected.

The transmitter's encoding rule E_T tells which message corresponds to the source state H and which message corresponds to the source state T. One of the messages m_1–m_4 corresponds to H, and one of the messages m_5–m_8 corresponds to T. Hence there are 16 possibilities and $|\mathcal{E}_\mathcal{T}| = 16$. However, this choice must be made in such a way that the receiver accepts the messages as authentic and translates them to the correct source state; see (16). In this example, where $E_R = e_1$, the message that corresponds to the source message H must be m_1 or m_2, and the message corresponding to the source message T must be m_5 or m_6. When $E_R = e_1$ is given, there are four possible ways to choose E_T, namely, $\{H \mapsto m_1, T \mapsto m_5\}$, $\{H \mapsto m_1, T \mapsto m_6\}$, $\{H \mapsto m_2, T \mapsto m_5\}$, and $\{H \mapsto m_2, T \mapsto m_6\}$. Note that not all pairs (e_r, e_t) are possible.

Table 1. The Cartesian Product Construction for $|S| = 2$ and $P_D = \frac{1}{2}$

		m							
	$g(m, e_r)$	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
	e_1	H	H	—	—	T	T	—	—
	e_2	H	H	—	—	T	—	T	—
	e_3	H	H	—	—	—	T	—	T
	e_4	H	H	—	—	—	—	T	T
	e_5	H	—	H	—	T	T	—	—
	e_6	H	—	H	—	T	—	T	—
	e_7	H	—	H	—	—	T	—	T
e_r	e_8	H	—	H	—	—	—	T	T
	e_9	—	H	—	H	T	T	—	—
	e_{10}	—	H	—	H	T	—	T	—
	e_{11}	—	H	—	H	—	T	—	T
	e_{12}	—	H	—	H	—	—	T	T
	e_{13}	—	—	H	H	T	T	—	—
	e_{14}	—	—	H	H	T	—	T	—
	e_{15}	—	—	H	H	—	T	—	T
	e_{16}	—	—	H	H	—	—	T	T

We can check the probability of success for any kind of deception. Let us introduce the following notation. Let $\mathcal{E}_\mathcal{R}(m)$ denote the set of receiver's encoding rules for which m is a valid message, i.e., $\mathcal{E}_\mathcal{R}(m) = \{e_r; g(m, e_r) \in \mathcal{S}\}$. Similarly, let $\mathcal{E}_\mathcal{T}(m)$ denote the set of transmitter's encoding rules for which m can be generated, $\mathcal{E}_\mathcal{T}(m) = \{e_t; f(s, e_t) = m, s \in \mathcal{S}\}$. Let $\mathcal{M}(e_r)$ be the set of possible messages for encoding rule e_r, $\mathcal{M}(e_r) = \{m; g(m, e_r) \in \mathcal{S}\}$. Finally, let $\mathcal{E}_\mathcal{R}(e_t)$ be the set of possible e_r values for a given e_t, specifically, $\mathcal{E}_\mathcal{R}(e_t) = \{e_r; g(m, e_r) \in \mathcal{S}, \forall m \in \mathcal{M}(e_t)\}$, and let $\mathcal{E}_\mathcal{T}(e_r)$ be the set of possible e_t values for a given e_r, namely, $\mathcal{E}_\mathcal{T}(e_r) = \{e_t; f(s, e_t) \in \mathcal{M}(e_r), \forall s \in \mathcal{S}\}$.

Each column in Table 1 contains 8 entries out of 16, and thus $|\mathcal{E}_\mathcal{R}(m)| = 8$ for any m, and we have

$$P_I = \max_m \frac{|\mathcal{E}_\mathcal{R}(m)|}{|\mathcal{E}_\mathcal{R}|} = \frac{8}{16} = \frac{1}{2}$$

Any two columns have at most 4 rows for which they both have entries, and thus $|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(m')| \leq 4$. We then have

$$P_S = \max_{\substack{m,m' \\ m \neq m'}} \frac{|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(m')|}{|\mathcal{E}_\mathcal{R}(m)|} = \frac{4}{8} = \frac{1}{2}$$

If the receiver claims to have received a message corresponding to source message H, it must choose either m_1 or m_2. But only one of them is the message that the transmitter would have used. Thus $P_{R_0} = \frac{1}{2}$. Or equivalently, $|\mathcal{E}_\mathcal{T}(e_r)| = 4$ for any e_r, and $|\mathcal{E}_\mathcal{T}(e_r) \cap \mathcal{E}_\mathcal{T}(m)| \leq 2$, so

$$P_{R_0} = \max_{m, e_r} \frac{|\mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(e_r)|}{|\mathcal{E}_\mathcal{T}(e_r)|} = \frac{2}{4} = \frac{1}{2}$$

By similar reasoning, $|\mathcal{E}_\mathcal{T}(e_r) \cap \mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(m')| \leq 1$, and thus

$$P_{R_1} = \max_{\substack{m,m',e_r \\ m \neq m'}} \frac{|\mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(m') \cap \mathcal{E}_\mathcal{T}(e_r)|}{|\mathcal{E}_\mathcal{T}(m) \cap \mathcal{E}_\mathcal{T}(e_r)|} = \frac{1}{2} = \frac{1}{2}$$

when $P(e_r, m) \neq 0$. Finally, assume, for example, that the transmitter has the mapping $\{H \mapsto m_1, T \mapsto m_5\}$ as his key. To succeed in his attack it must send a message different from m_1 and m_5, which is accepted by the receiver. From its key it knows that the receiver's key is one of the four keys e_1, e_2, e_5, e_6. For the best choice of message, two keys out of four accept the message as authentic, and thus $P_T = \frac{1}{2}$. Or formally, $|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(e_t)| \leq 2$ when $m \notin \mathcal{M}(e_t)$, and $|\mathcal{E}_\mathcal{R}(e_t)| = 4$, gives

$$P_T = \max_{\substack{m,e_t \\ m \notin \mathcal{M}(e_t)}} \frac{|\mathcal{E}_\mathcal{R}(m) \cap \mathcal{E}_\mathcal{R}(e_t)|}{|\mathcal{E}_\mathcal{R}(e_t)|} = \frac{2}{4} = \frac{1}{2}$$

Thus $P_D = \frac{1}{2}$ when the encoding rules are uniformly distributed. The parameters of the A^2-code are

$$|\mathcal{S}| = 2, \quad |\mathcal{M}| = 8, \quad |\mathcal{E}_\mathcal{R}| = 16, \quad |\mathcal{E}_\mathcal{T}| = 16$$

Expressing the parameters in terms of entropy we get $H(S) = 1, H(M) = 3, H(E_R) = 4$, and $H(E_T) = 4$. We also observe that from the dependence between the keys E_R and E_T we have $I(E_R; E_T) = 2$. This is a typical property of A^2 codes.

BIOGRAPHY

Thomas Johansson was born in Ljungby, Sweden in 1967. He received the M.Sc. degree in computer science in 1990 and the Ph.D. degree in information theory in 1994, both from Lund University, Lund, Sweden. From 1995 he has held various teaching and research positions in the Department of Information Technology at Lund University. Since 2000 he has held a position as Professor of Information Theory in the same department.

His scientific interests include cryptology, error-correcting codes, and information theory. He has served on cryptographic program committees such as EUROCRYPT'98/'00/'01/'02 and FSE'01/'02. He was a recipient of the SSF-JIG (Junior Individual Grant).

BIBLIOGRAPHY

1. B. den Boer, A simple and key-economical unconditionally authentication scheme, *J. Comput. Security* **2**(1): 65–71 (1993).
2. E. F. Brickell, A few results in message authentication, *Congressus Numerantium* **43**: 141–154 (1984).
3. J. L. Carter and M. N. Wegman, Universal classes of hash functions, *Journal of Comput. Syst. Sci.* **18**(2): 143–154 (1979).
4. E. N. Gilbert, F. J. MacWilliams, and N. J. A. Sloane, Codes which detect deception, *Bell Syst. Tech. J.* **53**(3): 405–424 (1974).
5. T. Johansson, Lower bounds on the probability of deception in authentication with arbitration, *IEEE Trans. Inform. Theory* **40**(5): (Sept. 1994).
6. T. Johansson, *Contributions to Unconditionally Secure Authentication*, Ph.D. thesis, Lund Univ., 1994.
7. T. Helleseth and T. Johansson, Universal hash functions from exponential sums over finite fields and Galois rings, *Lecture Notes in Computer Science*, Vol. 1107 Springer-Verlag, Berlin, 1996, (CRYPTO'96), pp. 31–44.
8. J. Black et al., UMAC: Fast and secure message authentication, in *Advances in Cryptology — CRYPTO'99 Lecture Notes in Computer Science*, Springer-Verlag, 1999, pp. 216–233.
9. A. Menezes, P. van Oorschot, S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1997.
10. P. Rogaway, Bucket hashing and its application to fast message authentication, *Proc. CRYPTO'95*, Santa Barbara, CA, LNCS 963, Berlin: Springer-Verlag, Berlin, 1995, pp. 29–42.
11. C. E. Shannon, Communication theory of secrecy systems, *Bell Syst. Tech. J.* **28**: 269–279 (Oct. 1949).
12. G. J. Simmons, A game theory model of digital message authentication, *Congressus Numerantium* **34**: 413–424 (1982).
13. G. J. Simmons, Authentication theory/coding theory, *Proc. CRYPTO'84*, Santa Barbara, CA, 1984, LNCS 196, Springer-Verlag, Berlin, pp. 411–431.
14. G. J. Simmons, A Cartesian product construction for unconditionally secure authentication codes that permit arbitration, *J. Cryptol.* **2**(2): 77–104 (1990).
15. G. J. Simmons, A survey of information authentication, in G. J. Simmons, ed., *Contemporary Cryptology, The Science of Information Integrity*, IEEE Press, New York, 1992. pp. 379–420.
16. D. R. Stinson, The combinatorics of authentication and secrecy codes, *J. Cryptol.* **2**(1): 23–49 (1990).
17. D. R. Stinson, *Cryptography Theory and Practice*, CRC Press, 1995.

AUTOMATIC REPEAT REQUEST

THOMAS E. FUJA
University of Notre Dame
Notre Dame, Indiana

1. INTRODUCTION

Error control is a term that describes methods for protecting the integrity of digitally transmitted signals. Error control techniques are used to provide a desired level of reliability when transmitting digital information over inherently unreliable links.

Error control can be broken down into two broad approaches:

- *Forward error control* (FEC), in which redundancy is added to the digital signal prior to transmission, and the redundancy is used at the receiver to reconstruct the transmitted signal, even if it was corrupted en route.
- *Automatic repeat request* (ARQ), in which a (typically) smaller amount of redundancy is added to the digital signal prior to transmission — redundancy that is used to *detect* corruption that may have occurred during transmission. In an ARQ-based system, received signals that are judged to be corrupt are retransmitted.

Clearly, a major difference between FEC and ARQ is that ARQ requires the existence of a *feedback* (or *return*) channel from the receiver to the transmitter; this feedback channel is used to alert the transmitter when corrupted data have been received and to request retransmission. As a result, ARQ is not feasible for systems in which the return channel does not exist or for some real-time applications where it is impractical to request retransmission in a timely manner. Conversely, ARQ schemes *are* implemented in a wide variety of communications contexts, including satellite-based systems, local area networks, and the ubiquitous Internet.

This article describes the various methods by which an automatic repeat request protocol may be deployed to enhance the reliability of a digital communication system.

First, we consider means by which errors that occur during transmission can be detected. (This can be done via a particular form of *block coding*, and the reader should refer to the article on error control codes for more background.) Then, we consider three different protocols by which the sender and receiver coordinate the retransmission of corrupted information; these three protocols are described in Section 3 and their performances are analyzed in Section 4. Finally, *hybrid* ARQ — incorporating elements of both FEC and conventional ARQ — is described in Section 5.

2. ARQ AND CRC CODES IN THE OSI NETWORK ARCHITECTURE

It is possible to deploy error control methods at many different layers of the seven-layer open system interconnection (OSI) architecture [1,2]. For instance, at the physical layer (layer 1) a form of FEC called *trellis-coded modulation* (TCM) may be used improve the reliability of the virtual bit pipe embodied at the physical layer. At the other extreme, many applications (layer 7) that are used to disseminate multimedia files over the Internet employ error control methods to mitigate the effects of errors that may have "gotten past" the error control techniques in place at the lower layers.

Automatic repeat request schemes are typically deployed at the data-link layer (layer 2). It is at the data-link layer that the data packets formed at the transport layer are augmented with extra symbols at the beginning of the packet ("headers") and extra symbols at the end of the packet ("trailers") to form *frames* (see Fig. 1). What exactly goes into the headers and trailers depends on the particular data-link protocol, but typically the header may include an address and/or a packet sequencing number, while the trailer will include one or more *check bytes* used to detect errors that occur in the frame during transmission.

The contents of these check bytes (also known as *parity bits*) are based on the contents of the rest of the frame, and they are chosen so that the bits in the frame satisfy certain parity constraints — constraints that can be checked for violation at the receiver. The codes most commonly used in this way are the so-called cyclic redundancy check (CRC) codes [3,4].

CRC codes form a class of binary block codes; an (n, k) binary block code is an error control construct in which k data bits are appended with $r = n - k$ redundant bits to form an n-bit codeword. Because $k < n$, not all possible n-tuples are valid codewords; this means that errors can be detected when an invalid codeword is observed at the receiver. Cyclic redundancy check codes take their name from the fact that when n (the "blocklength" of the code) takes on its maximum value, the resulting code is *cyclic* — meaning that cyclically shifting one valid codeword yields another valid codeword.

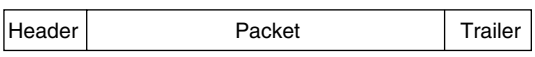


Figure 1. A *frame* consisting of a header, a packet, and a trailer.

A CRC code is defined by a generator polynomial $g(x)$. More specifically, suppose that there are n bits in a frame — call them $[a_{n-1}, a_{n-2}, \ldots, a_1, a_0]$ — and suppose that the r low-order bits (a_0 through a_{r-1}) constitute the parity check bits. [The number of parity check bits is equal to the degree of $g(x)$ — i.e., $r = \deg[g(x)]$.] Then those parity check symbols are chosen so that the polynomial $a(x) = a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ is evenly divisible by $g(x)$. [Here, $a(x)$ and $g(x)$ are both polynomials with binary coefficients, and so all operations are modulo two; moreover, the constraint $g(x)|a(x)$ — read "$g(x)$ divides $a(x)$" — is equivalent to imposing r different parity constraints on the frame.]

As a simple example, consider the generator polynomial $g(x) = x^4 + x + 1$. This is the generator for a [15,11] *Hamming code* — a cyclic code commonly used for error correction and detection. Suppose that we wish to use this code to protect a frame consisting of $n = 15$ bits — 11 "data" (i.e., non-redundant) bits and $r = \deg[g(x)] = 4$ parity bits. Suppose further that the data bits are given by [10011000110]; then the transmitted frame is given by $[10011000110a_3a_2a_1a_0]$ where the parity bits a_0 through a_3 are chosen so that the "codeword polynomial" $c(x) = x^{14} + x^{11} + x^{10} + x^6 + x^5 + a_3x^3 + a_2x^2 + a_1 + a_0$ is divisible by $g(x)$. It can be shown that the appropriate choice is $a_3 = a_1 = 1$ and $a_2 = a_0 = 0$ and so the transmitted frame is [100110001101010]; this is because the corresponding code polynomial $c(x) = x^{14} + x^{11} + x^{10} + x^6 + x^5 + x^3 + x$ is a multiple of $g(x)$ — specifically, $c(x) = g(x) \cdot (x^{10} + x^2 + x)$ — and this choice of the parity bits is the *only* choice that results in a multiple of $g(x)$.

At the receiver, the contents of the frame are checked to see if the "received polynomial" is, indeed, evenly divisible by $g(x)$. If it is, then the frame is declared error-free; if the received polynomial is not evenly divisible by $g(x)$, then the frame is declared corrupt.

CRC codes provide a mechanism through which, at a relatively small cost of overhead, the vast majority of transmission errors can be detected. There are 2^n possible binary n-tuples, but only 2^{n-r} of them satisfy the parity constraints; therefore, if the data are corrupted during transmission into something uniformly random, then the probability the received frame satisfies the constraints, resulting in an *undetected error* — is 2^{-r}. To compute the probability of undetected error for a CRC code on a binary symmetric channel with crossover probability p, we observe that CRC codes are *linear* codes, which means that an undetected error occurs if and only if the error pattern itself is a valid codeword; this means the probability of undetected error is given by $P_u = \sum_{i=1}^{n} A_i p^i (1-p)^{n-i}$, where A_i is the number of valid CRC codewords containing i ones and $n - i$ zeros (i.e., the number of CRC codewords of *Hamming weight* i). Unfortunately, computing the A_i value (known as the code's *weight enumerator*) is difficult for most large codes; however, it has been shown [3] that there exist (n, k) binary codes with an undetected error probability upper bounded by $2^{(n-k)}[1 - (1-p)^n]$.

CRC codes are used in a broad array of communication applications; for instance, a 7-bit CRC is used to detect errors in the compressed speech found in the second-generation TDMA-based digital cellular standard

Table 1. Properties of Three Standard CRC Codes

Code	$g(x)$	$n_{\max}$	$d_{\min}$
CRC-12	$x^{12}+x^{11}+x^3+x^2+x+1$	2,047	4
CRC-ANSI	$x^{16}+x^{15}+x^2+1$	32,767	4
CRC-CCITT	$x^{16}+x^{12}+x^5+1$	32,767	4

[5]. Table 1 shows the generator polynomials for three different CRC codes that have been selected as international standards for use in communication networks. In each case, the degree of $g(x)$ is equal to the number of check bits per frame required to implement the code. Also included in the table is the code's minimum distance ($d_{\min}$) — that is, the smallest number of bits in which two different valid codewords can differ — and the maximum blocklength ($n_{\max}$) that should be used; for instance, CRC-12 should be used with frames no longer than 2047 bits, meaning frames with $2047 - 12 = 2035$ nonparity bits.

3. PROTOCOLS FOR AUTOMATIC REPEAT REQUESTS

Section 2 described how redundancy can be added at the transmitter and used at the receiver to detect frames that have been corrupted during transmission. "Pure" ARQ protocols require that every frame deemed corrupt at the receiver be retransmitted. ("Hybrid" ARQ schemes use a mix of forward error control and retransmission to ensure data integrity; hybrid schemes are described in Section 5.)

This section describes three protocols for alerting the transmitter that retransmission of a frame is necessary. Each protocol is based on *acknowledgments* passed from the receiver to the transmitter: a *positive acknowledgment* ("ACK") to indicate that an error-free frame has been received and a *negative acknowledgment* ("NAK") to indicate that a corrupt frame has been detected. (Obviously, since an acknowledgment is also subject to corruption, the ACK or NAK is also often protected with a CRC code.)

Some ARQ protocols also have a provision for *timeouts*; when a timeout strategy is used, then if a frame has not been acknowledged (either positively or negatively) after a specified amount of time, the transmitter acts as if it had been acknowledged negatively. Timeouts reduce the effects of packets that get "lost" (rather than corrupted) during transmission.

Finally, we observe that the protocol descriptions that follow all assume that the frames transmitted on the forward link and the acknowledgments transmitted on the return link — at least those that are received — are received in the order in which they were sent. (Note this does not preclude the possibility that some frames and/or acknowledgments are lost.) Moreover, we are not taking into account the possibility of undetected errors on the feedback channel; for discussion and analysis of these issues, the reader is referred to Refs. 3 and 4.

3.1. Stop-and-Wait ARQ

sThe simplest ARQ protocol is the *stop-and-wait* strategy (see Fig. 2), which requires that the transmitter, on sending a frame, wait until it has been acknowledged before the next frame is transmitted.

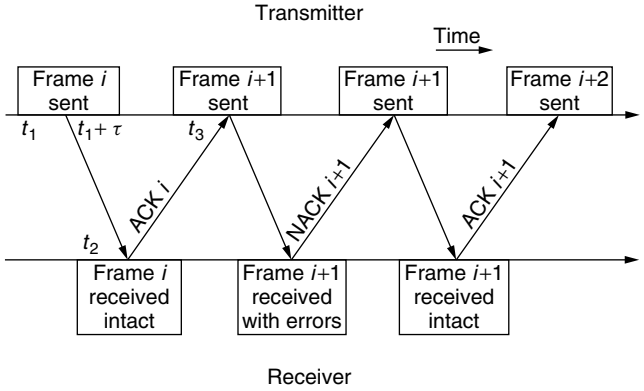


Figure 2. Timeline of stop-and-wait strategy.

The transmitter begins sending frame i at time t_1, and it is received beginning at time $t_2 > t_1$; the receiver then sends an acknowledgment (either positive or negative) through the feedback channel that is received at the transmitter beginning at time t_3. At that point, the transmitter either resends frame i (if a NAK was received) or sends frame $i+1$ (if an ACK was received). This procedure continues in a "pingpong" fashion, with the transmitter and the receiver taking turns sending frames and acknowledgments, respectively. To avoid potential confusion caused by lost frames, a sequence number is often included in the header; in a similar fashion, to avoid confusion caused by lost acknowledgments, the ACK or NAK often will be designed to include the sequence number of the next frame expected by the receiver — and so the ACK of frame i is equivalent to a request for frame $i+1$, while a NAK of frame i is equivalent to a request for frame i. These sequence numbers are incremented modulo some positive integer with each new frame; indeed, simply incrementing the sequence numbers modulo 2 — specifically, identifying the even-numbered and odd-numbered frames — is adequate for stop-and-wait ARQ.

The advantage of the stop-and-wait strategy lies in its simplicity. It does not require a full-duplex channel (i.e., it does not require simultaneous transmission in each direction). Moreover, compared with more sophisticated approaches, stop-and-wait requires minimal storage and processing at both the transmitter and the receiver.

The cost of this simplicity is *throughput* — the efficiency with which information is reliably delivered over the forward channel. Because most communication channels *are* in fact full-duplex and the stop-and-wait strategy does not exploit this capability, the result is a transmitter sitting idle, waiting for acknowledgments, when it could (in principle) be formatting and transmitting additional frames. This is seen in terms of the "idle time" between time $t_1 + \tau$ and time t_3 in Fig. 2.

3.2. Go-Back-*N* ARQ

The go-back-N protocol is designed to address the inherent inefficiency of stop-and-wait. In a communication link implementing the go-back-N strategy, the transmitter

does not have to wait for a request before sending another frame. Rather, the transmitter is permitted to send all of the frames in a "window" maintained at the transmitter. When the first (earliest) frame in the window is ACKed, the transmitter "slides" the window one position, dropping the acknowledged frame and adding a new frame that it then transmits; conversely, if a NAK is received for that first frame in the window, the transmitter "backs up" and resends all the frames in the window, beginning with the corrupt frame.

More specifically, suppose that the transmitter has received an ACK for frame i — equivalently, a request for frame $i+1$. Then, under the go-back-N protocol, the transmitter may send frames $i+1$ through $i+N$ without receiving another ACK; however, the transmitter *must* receive an ACK for frame $i+1$ before transmitting frame $i+N+1$.

Go-back-N is often referred to as a *sliding-window protocol* because the transmitter maintains a sliding window of "active" (or "outstanding") frames — a window of up to N frames that have been transmitted but have not yet been acknowledged. It should be clear that the Stop-and-Wait strategy is simply a go-back-1 strategy.

In execution, the go-back-N protocol is very simple. As long as the receiver judges its frames to be error-free, it sends ACKs to the transmitter, advancing the sliding window. If a corrupt frame is detected, then the resulting NAK tells the transmitter to back up to the corrupt frame and begin retransmitting from that point. Because the transmitter is not permitted to transmit more than N frames beyond what has been ACKed, it will never be required to back up more than N frames. [For example, in a go-back-4 system, the transmitter, having received a positive acknowledgment of (say) frame 10, is free to transmit frames 11–14, but it cannot transmit frame 15 without a positive acknowledgment of frame 11. So, in a worst case, if frame 11 is corrupted, the transmitter will have to back itself up and follow its transmission of frame 14 with a (re)transmission of frame 11.]

Figure 3 illustrates this concept. Frame $i+1$ is corrupted, so as soon as the transmitter receives a NAK for frame $i+1$, it initiates the corresponding backup. Note that, as drawn in Fig. 3, the transmitter had sent only two additional frames (frames $i+2$ and $i+3$) before it received the NAK for frame $i+1$; therefore, this figure describes an action that could have taken place in a go-back-N system for any $N \geq 3$. Moreover, as indicated in Fig. 3, the receiver discards those additional frames sent between the original (corrupt) frame and its retransmitted replica; whether those discarded frames were corrupted during transmission has no effect on the action taken by the receiver.

The appropriate value of N is determined by a number of issues, including the propagation delay of the system and the length of the frames. Obviously, if a "full window" of N frames are transmitted well before the first frame in the window is acknowledged, then the transmitter would sit idle until it gets that acknowledgment — the same kind of inefficiency that makes stop-and-wait (also known as "go-back-1") unattractive. At the opposite extreme, there's no point in making N so large that the first frame in the window is *always* acknowledged before the window reaches its full complement of N frames.

As in the stop-and-wait protocol, frame *sequence numbers* are required in the headers on the forward channel to avoid confusion caused by a lost frame, while frame *request numbers* are designed into the feedback transmission to avoid confusion caused by a lost ACK or NAK. In the stop-and-wait protocol, it was claimed that these counters need only be maintained modulo 2 (i.e., only a single bit needs to be used, indicating whether the transmitted/requested frame has an even or odd frame number). In a go-back-N protocol, the transmitter and receiver must be able to distinguish between all the frames in the window; therefore, the counters must be maintained modulo M for some integer $M > N$. This means that at least $[\log_2(N+1)]$ bits must be dedicated for sequence numbers in each header on the forward link, and the same number of bits must be dedicated for request numbers in each ACK/NAK on the return link.

The go-back-N protocol offers a reasonable balance between efficient use of the channel and reasonably simple implementation.

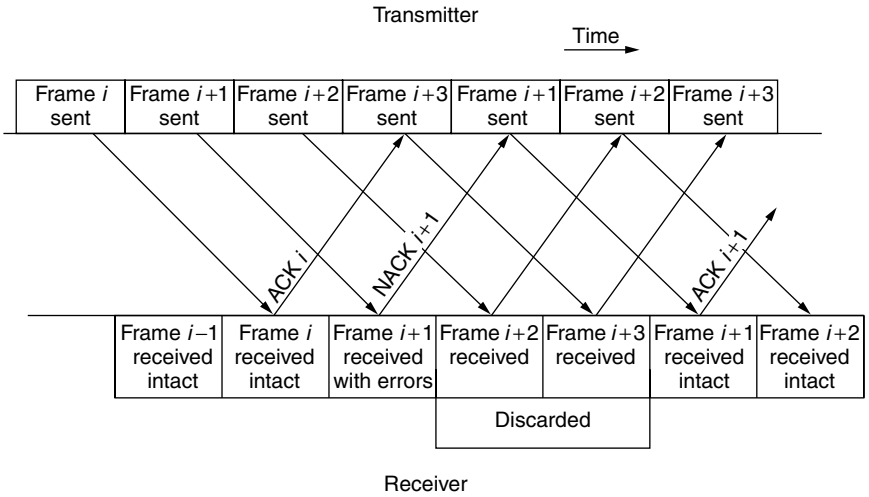


Figure 3. Timeline of go-back-N strategy.

3.3. Selective Repeat ARQ

While the go-back-N protocol clearly offers efficiency advantages relative to stop-and-wait, it is still not as efficient as it could be; when a frame is corrupted under the go-back-N protocol, all the frames that were transmitted *after* the bad frame was sent and *before* it was negatively acknowledged must be retransmitted as well — as many as $N-1$ frames in addition to the corrupt frame. These discarded frames may have, in fact, been received error-free; they are discarded only because they followed too closely on the heels of a corrupt frame.

Selective-Repeat ARQ is designed to overcome this inefficiency. In the Select-Repeat protocol, the transmitter does *not* retransmit the contents of the entire window when a NAK is received; rather, the transmitter re-sends *only* the frame "selected" as corrupt by the receiver.

Selective-repeat is typically implemented as a sliding-window protocol, in the same way as go-back-N. This means that when frame i is ACKed, frames $i+1$ through $i+N$ (but *not* frame $i+N+1$) may be transmitted before frame $i+1$ is positively acknowledged. Once frame $i+1$ is ACKed, all the "oldest" frames in the window that have been positively acknowledged are considered complete and are shifted out of the window; the same number of new frames are shifted into the window and are transmitted.

This process is illustrated in Fig. 4. As in Fig. 3, frame i is received intact and frame $i+1$ is corrupted; however, in the selective-repeat strategy, frames $i+2$ and $i+3$ — the two frames sent *after* frame $i+1$ was sent but *before* it was negatively acknowledged — are *not* discarded at the receiver (and resent by the transmitter) but instead are accepted, assuming they are not corrupt.

As with stop-and-wait and go-back-N strategies, the sequence numbers (on the forward path) and the request numbers (on the return path) used in selective-repeat ARQ are maintained modulo M. For correct operation of selective-repeat, this modulus must satisfy $M \geq 2N$, where N is the window size.

The chief advantage of selective-repeat ARQ is its efficiency; if each frame is corrupted with probability p, then selective-repeat can deliver good frames at a rate approaching $1-p$ good frames per transmitted frame — the highest possible rate. The chief disadvantage of selective-repeat is the memory and logic needed to store the arriving frames and reassemble them in the appropriate order.

4. PERFORMANCE ANALYSIS OF ARQ SYSTEMS

In analyzing the three ARQ protocols described in Section 3, we focus on two criteria — the *reliability* of the information delivered to the receiver and the *efficiency* with which that information is delivered. These two criteria are formalized in the notions of *frame error rate* and *throughput*, respectively.

4.1. Reliability Analysis of an ARQ System

When a frame is transmitted, three things can happen en route:

- It can be delivered error-free to the receiver; henceforth we assume that each packet is delivered error-free with probability P_c [e.g., if each bit in an n-bit frame is "flipped" with independent probability p, then $P_c = (1-p)^n$].
- It can be corrupted by noise in such a way that the receiver can detect the presence of the error via a CRC code parity violation. Let the probability of such an event be denoted P_d. Clearly, P_d depends on the nature of the channel and the particular CRC code under consideration.
- It can be corrupted by noise in such a way that the receiver *cannot* detect the presence of the error via a CRC code parity violation — that is, an *undetected error* occurs. Let the probability of such an event be denoted P_u. Once again, P_u depends on the particular channel and the particular CRC code being used. In Section 2 it was shown that, for a binary symmetric channel with crossover probability p, the undetected error probability is $P_u = \sum_{i=1}^{n} A_i p^i (1-p)^{n-i}$, where A_i is the number of valid codewords containing i ones and $n-i$ zeros.

As noted above, these three events are the only possibilities when a frame is transmitted, and so $P_c + P_d + P_u = 1$.

The *frame error rate* is the probability that a frame is accepted by the receiver as correct when it is, in

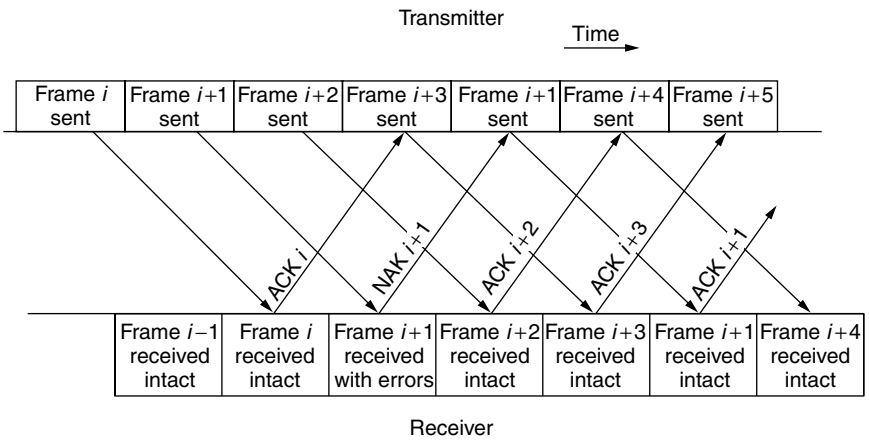


Figure 4. Timeline of selective-repeat strategy.

fact, corrupt. In an ARQ system, it is possible that a frame may be erroneously accepted the first time it is transmitted; alternatively, it could be rejected the first time but erroneously accepted the second time. Or it could be rejected twice before being erroneously accepted the third time. Continuing in this vein, and assuming there is no limit imposed on the number of retransmissions permitted, we arrive at the following expression for frame error rate:

$$\begin{aligned}\text{FER} &= P_u + P_d P_u + P_d^2 P_u + P_d^3 P_u \cdots \\ &= P_u \sum_{i=0}^{\infty} P_d^i \\ &= \frac{P_u}{1 - P_d}\end{aligned}$$

As a simple example, suppose that you were to use the [15,11] Hamming code mentioned in Section 2 to generate $r = 4$ bits of parity for $k = 11$ bits of data, requiring a frame length of $n = 15$. For this code it can be shown [4] that $A_0 = 1, A_1 = 0$, and A_i for $i = 2, 3, \ldots, 15$ can be computed via the recrusion

$$(i+1)A_{i+1} + A_i + (16 - i)A_{i-1} = \binom{15}{i}$$

With this "weight enumeration" it is simple to calculate P_u and P_d [and so $\text{FER} = P_u/(1 - P_d)$] for a binary symmetric channel; this is plotted for crossover probabilities ranging from $p = 10^{-1}$ to $p = 10^{-4}$ as seen in Fig. 5.

4.2. Efficiency Analysis of an ARQ System

The reliability analysis presented above is valid for *any* of the three ARQ protocols described in Section 3. This is because the *reliability* of any ARQ protocol depends only on the ability of the decoder to detect errors that occur during transmission — that is, it depends only on the error-detecting capability of the underlying CRC code and is independent of how any corrupted frames are subsequently retransmitted.

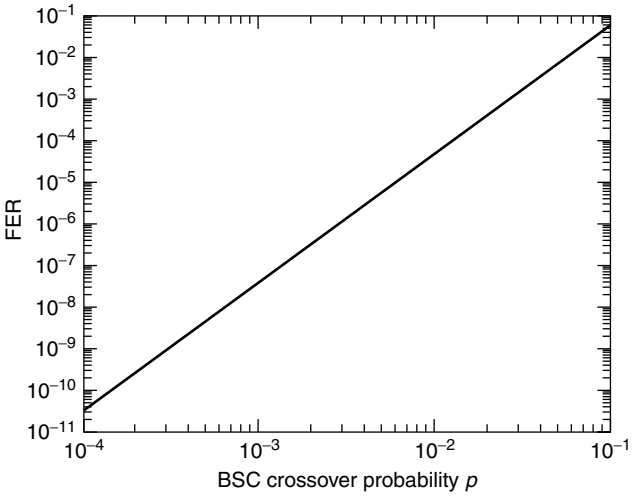


Figure 5. Frame error rate (FER) of an ARQ protocol over a binary symmetric channel assuming that the [15,11] Hamming code is used for error detection.

In contrast, the *efficiency* of an ARQ scheme is strongly tied to the mechanism by which the transmitter and receiver coordinate the retransmission of corrupt frames.

The figure of merit used to assess the efficiency of an ARQ protocol is called its *throughput*, and we denote it by η. The throughput of an ARQ scheme is defined as the ratio of the average number of information bits accepted by the receiver per unit time to the number of bits that can be transmitted over the channel per unit time. Under this definition, the throughput of a scheme employing an (n, k) CRC code (i.e., a code used in an n-bit frame, with k information bits and $n - k$ parity bits) can never be greater than $k/n < 1$.

Conceptually, selective-repeat ARQ is the simplest protocol to analyze in terms of its throughput. Each time a frame is transmitted it has probability $P \triangleq P_c + P_u$ of being accepted by the receiver; therefore, if we let T_{SR} be a random variable equal to the number of frame transmissions required until a particular frame is accepted, then $\Pr(T_{\text{SR}} = 1) = P, \Pr(T_{\text{SR}} = 2) = P(1 - P), \Pr(T_{\text{SR}} = 3) = P(1 - P)^2$, and so on; thus T_{SR} has a geometric distribution with mean $E[T_{\text{SR}}] = 1/P$. If we assume that each n-bit frame contains r bits of redundancy and $k = n - r$ bits of "data" (i.e., nonparity bits), then, under this protocol, the receiver accepts k information bits in the amount of time (on average) it takes to transmit n/P over the channel, so the throughput for selective-repeat ARQ is

$$\eta_{\text{SR}} = \frac{k}{n} P$$

The efficiency of the go-back-N protocol can be analyzed in a similar fashion; the main difference lies in the fact that, when a frame is rejected under the go-back-N protocol, the entire window of outstanding frames must be retransmitted. Once again, let $P = P_c + P_u$ be the probability that a frame is accepted, and let T_{GBN} be a random variable equal to the number of frames that must be transmitted until a specified frame is accepted; then if we assume the worst case (i.e., that a "full window" of N frames must be retransmitted every time a frame is rejected), then

$$\begin{aligned}E[T_{\text{GBN}}] &= 1 \cdot P + (N+1) \cdot (1 - P) \cdot P \\ &\quad + (2N+1) \cdot (1 - P)^2 \cdot P + \cdots \\ &= 1 + \frac{N(1 - P)}{P}\end{aligned}$$

So under the go-back-N protocol the receiver accepts k information bits in the amount of time (on average) it takes to transmit $n(1 + N(1 - P)/P)$ bits over the channel, so the throughput is

$$\eta_{\text{GBN}} = \left(\frac{k}{n}\right)\left(\frac{P}{P + N(1 - P)}\right)$$

Finally, consider stop-and-wait ARQ. In our analysis of the other two protocols, we began by asking a question: "On average, how many frames must be transmitted

before a particular frame is accepted?" In those analyses, we could measure the delay in "frame units" because it was implicitly assumed that transmission was continuous in the forward channel — that is, the frames were transmitted one after the other, with no idle time in between. (By "frame unit" we mean the amount of time required to transmit a single n-bit frame.)

For stop-and-wait, there *is* idle time between subsequent transmissions, and so to use the same approach we shall measure that idle time in terms of the number of frames that *could have been transmitted*. Specifically, let β denote the amount of time the transmitter is idle between frames divided by the amount of time it takes the transmitter to send one frame; in effect, β is the idle time measured in "frame units." [Referring to Fig. 2, $\beta = (t_3 - (t_1 + \tau))/\tau$.] So, if a frame is accepted the first time, the delay is $1 + \beta$ frame units; if a frame is rejected the first time but accepted the second, the delay is $2(1 + \beta)$ frame units. And so, if we let T_{SW} be a random variable representing the delay (in frame units) incurred from the time a frame is first sent until it accepted, we have

$$\begin{aligned} E[T_{\mathrm{SW}}] &= (1+\beta)\cdot P + 2\cdot(1+\beta)\cdot(1-P)\cdot P \\ &\quad + 3\cdot(1+\beta)\cdot(1-P)^2\cdot P + \cdots \\ &= \frac{1+\beta}{P} \end{aligned}$$

As a result, the throughput for the stop-and-wait protocol is given by

$$\eta_{\mathrm{SW}} = \left(\frac{k}{n}\right)\left(\frac{P}{1+\beta}\right)$$

The parameter β is determined by the round-trip propagation delay, the signaling rate of the system (i.e., how many bits per second are transmitted), and the length of the frames and the acknowledgments, as well as the processing delay at the transmitter and receiver.

To compare the three protocols, consider once again a system employing the [15,11] Hamming code for error detection. For the go-back-N protocol we shall set $N = 4$ and for the stop-and-wait protocol we set $\beta = 3$; these are comparable configurations, since each corresponds to a delay between the time the transmission of a frame is completed and the time the frame is acknowledged equal to 3 times the duration of one frame.

Figure 6 shows the throughput of the three protocols as a function of the crossover probability of a binary symmetric channel. We observe that, for low crossover probabilities, the throughput of both selective-repeat and go-back-N approach the rate of the error detection code: $\frac{11}{15} \approx 0.733$. By comparison, because the transmitter in stop-and-wait is sitting idle 75% of the time, its maximum throughput is $0.25 \times \frac{11}{15} \approx 0.183$.

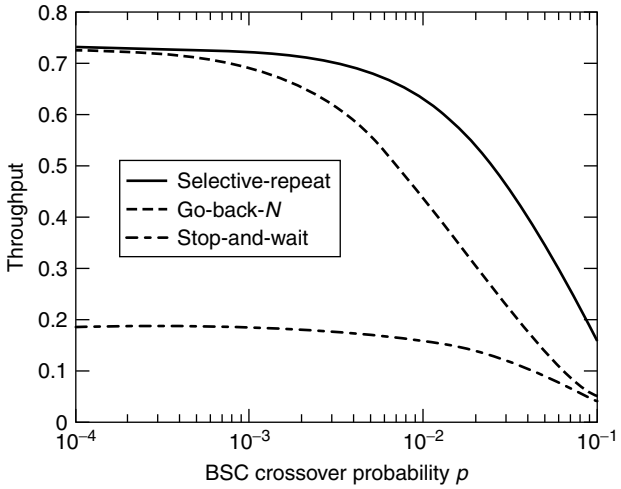


Figure 6. Throughputs of three different ARQ protocols over a binary symmetric channel assuming that the [15,11] Hamming code is used for error detection. The go-back-N protocol uses a window length of $N = 4$, and the stop-and-wait protocol assumes a parameter value of $\beta = 3$.

5. HYBRID ARQ SYSTEMS

It was claimed at the beginning of this article that error control schemes could be broadly classified into two categories: *forward error control* (FEC), in which redundancy added at the transmitter is used to recover the transmitted message even in the face of corruption, and *automatic repeat request* (ARQ), in which redundancy is used only to *detect* corruption and trigger a retransmission.

A hybrid ARQ system [6,7] employs elements of both FEC and ARQ. Hybrid ARQ schemes can be classified as either *type 1 hybrid ARQ* or *type 2 hybrid ARQ*.

The simplest implementation of a type-1 hybrid ARQ system uses two codes: a high-rate error *detection* code and a (typically) lower-rate error *correction* code. Data are first encoded using the error detection code and then encoded again using the error correction code; as a result, the frame trailer contains redundant bits from both codes. At the receiver, the decoder first attempts to reconstruct the frame using the redundancy from the error correction code; it then passes the result to the error detection unit, which checks to see if the error correction decoder was successful. If the error detection unit observes a parity violation, a retransmission is triggered; otherwise, the frame is accepted.

This implementation of a type-1 hybrid ARQ protocol basically uses the "inner code" (i.e., the error correction code) to create a more reliable "virtual" digital channel and then implements a conventional ARQ protocol over that more reliable channel.

Type-1 hybrid-ARQ can also be implemented with a single powerful code. To see how this can be done, observe that an (n, k) block code can be used to simultaneously correct t errors and detect $u > t$ errors provided its minimum distance $d_{\min}$ satisfies $d_{\min} \geq t + u + 1$. (For instance, a code with minimum distance $d_{\min} = 7$ can correct all occurrences of $t = 2$ errors while simultaneously *detecting* — i.e., neither correcting *nor* miscorrecting — three or four errors.) A decoder using such a code within the context of a type-1 hybrid ARQ would correct all error patterns affecting t or fewer bits; if more than t (but no more than u) errors occur during transmission, then a retransmission would be triggered.

Type-2 hybrid ARQ systems operate on the principle of *incremental redundancy*. While a number of type-2 hybrid protocols have been formulated, they all share in common the characteristic that, when a frame is initially rejected, what is retransmitted is *not* the same frame that was originally sent but rather a frame whose "payload" consists of parity bits based on that original frame. In this way, the original corrupted frame — kept in a buffer at the receiver — can be augmented with the newly-received parity bits to form a codeword from a longer, more powerful FEC code.

As one implementation of this approach, let C_0 be a high-rate (n, k) error-detecting code and let C_1 be a more powerful $(2n, n)$ error-correcting code. At the transmitter, k bits of data are encoded using C_0, and the resulting codeword — call it $\mathbf{c}$ — is transmitted; however, before it is sent, $\mathbf{c}$ is itself encoded with the code C_1 to form n parity bits $\mathbf{p}$ [i.e., $\mathbf{c} * \mathbf{p}$ forms a $2n$-bit codeword from C_1; here, "$*$" denotes concatenation]. These n bits of $\mathbf{p}$ are stored at the transmitter while $\mathbf{c}$ is sent. At the receiver, the received version of $\mathbf{c}$ is checked for errors and accepted if none are found; however, if $\mathbf{c}$ has been corrupted, then what is transmitted in response to the NAK is *not* the n bits of $\mathbf{c}$ but rather the n bits of $\mathbf{p}$. As a result, the decoder has a noisy version of $\mathbf{c} * \mathbf{p}$, which can be corrected using a decoder for the powerful code C_1. (If correction is impossible, the process can begin again, with retransmission of $\mathbf{c}$.)

The benefits of type-2 hybrid ARQ systems lie in the fact that they do not "waste" redundancy when the channel is good. In any reasonably designed communication link, frame errors are relatively rare, so to use a powerful error correcting code with every frame (as type-1 hybrid systems do) may be overkill; by only sending error-correcting redundancy when needed, the type-2 hybrid ARQ system makes more efficient use of the channel. As usual, the tradeoff is in buffering and logic complexity.

BIOGRAPHY

Thomas E. Fuja received his undergraduate education at the University of Michigan, obtaining a B.S.E.E. and a B.S.Comp.E. in 1981. He attended graduate school at Cornell University, Ithaca, New York, where he received an M.Eng. in 1983 and a Ph.D. in 1987, both in electrical engineering. Dr. Fuja was on the faculty of the Department of Electrical Engineering at the University of Maryland in College Park, Maryland, from 1987 until 1998; in addition, he served as the program director for communications research at the U.S. National Science Foundation, Arlington, Virginia, in 1997 and 1998. Since 1998, Fuja has been on the faculty of the University of Notre Dame in South Bend, Indiana, where he is a professor of electrical engineering. His research interests lie in channel coding and information theory — most recently focusing on issues related to wireless communications and on the interface between compression and error control. Dr. Fuja has been very active in the IEEE Information Theory Society; in 2002, he served as that organization's president.

BIBLIOGRAPHY

1. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1992.
2. A. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall PTR, Upper Saddle River, NJ, 1996.
3. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
4. S. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
5. T. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall PTR, Upper Saddle River, NJ, 1996.
6. S. Lin, D. Costello, Jr., and M. Miller, Automatic repeat request error control schemes, *IEEE Commun. Mag.* **22**: 5–16 (Dec. 1984).
7. L. Rasmussen and S. Wicker, The performance of type-I trellis coded hybrid-ARQ protocols over AWGN and slowly fading channels, *IEEE Trans. Inform. Theory* **40**(2): 418–428 (March 1994).

B

BANDWIDTH REDUCTION TECHNIQUES FOR VIDEO SERVICES

NELSON L. S. DA FONSECA
State University of Campinas
Institute of Computing
Campinas, Brazil

1. INTRODUCTION

Video-on-demand (VoD) is a client–server application which allows users to watch movies stored in remote servers. It is a critical technology for home entertainment, professional communication, and digital video libraries, to name a few uses. In recent years, it has come to be regarded as the main video application for future broadband multimedia networks.

There are two major kinds of interactiveness in VoD services, varying according to the type of video distribution and scheduling policies implemented by the server. In true video-on-demand, all requests are granted immediately if resources are available, whereas in near-video-on-demand, users may need to wait a certain time before their requests are granted.

The videostream, which constitutes the flow of bytes corresponding to the frames composing a movie, can be delivered only when adequate server resources (I/O bandwidth) as well as network bandwidth (video channels) have been reserved for this purpose. With current technology, there is a mismatch of roughly a whole order of magnitude between I/O data rates and network data rates, so techniques to utilize network bandwidth more efficiently are need.

A single videostream requires a considerable amount of bandwidth, ranging from 6 Mbps (megabits per second) for MPEG-2 streams to 20 Mbps for HDTV streams. Considering a potential population of several million viewing households, the network bandwidth demand would be beyond the present-day network capacity, approaching the order of Tbps. Such demands prevent the deployment of VoD on a large scale. Consequently, numerous techniques to reduce these demands have been developed. One of these techniques consists of the placement of several servers throughout the network so that they will be closer to the user, thus, diminishing the need for allocation of channels along long network paths [1]. In itself, however, this technique is not able to reduce the demand sufficiently, and further reductions are necessary [2]. Some sort of sharing of videostreams, either by a group of viewers (multicast) or by all users (broadcast), must be used. The choice of routing technique depends on the degree of interactiveness required. On one hand, multicasting permits true video-on-demand, but at the expense of high processing overhead costs. On the other hand, broadcasting involves lower overhead costs, but can provide only near-video-on-demand. Various techniques based on multicasting and on broadcasting have been proposed, with greater or lesser success in bandwidth reduction. These techniques will be explained here: piggybacking, patching, and batching, all of which utilize multicasting with different degrees of emphasis in the time that users will have to wait to watch a desired program, and periodic broadcasting, which is more appropriate for high request rates.

2. PIGGYBACKING

Piggybacking is based on the fact that viewers do not perceive an alteration in the display rate when it is no more than 5% of the nominal rate. In a VoD server with piggybacking, a request for viewing a video is immediately granted if a channel is available. However, when a new channel is allocated, and another exhibition of the same video is in progress for another user, the display rate of the original showing is slowed down, while the display rate of the recently admitted request is increased. When the faster streams catch up with the slower one, the two are merged, and, as a consequence, one of the video channels is released (Fig. 1) [3]. The aim of piggybacking policies is to reduce the total number of displayed frames for a set of streams, which is equivalent to reducing the bandwidth required to display this set of movies.

The change in display rate is affected by dynamically compressing or expanding the video being displayed through the insertion of additional frames produced by the interpolation of proceeding and succeeding frames, or through the deletion of frames, with neighboring frames altered to reduce the abruptness of the change. On-line alteration of the display rate has been shown to be a difficult task, however, especially for MPEG-encoded movies. Another solution would be to store copies of movies corresponding to different display rates, although this greatly increases storage demands. For a 100-min-long MPEG-2-encoded movie, for example, the storage demand is about 4.5 GB (gigabytes) per movie.

Merging two streams is possible only if the difference between the frames being displayed by the two streams is such that the faster stream will exhibit the same frame displayed by the slower one at some time prior to the end of the exhibition of the latter. In other words, a merge can occur only if the faster stream is able to catch up with the slower one. The maximum catchup window is the difference in frames for which merging is feasible. This window depends on the discrepancy of the display rate of two streams and is given by $((S_{\max} - S_{\min}) \times L)/S_{\max}$, where $S_{\min}$ and $S_{\max}$ denote the minimum and the maximum display rates, respectively and L, the movie length in frames.

Each piggybacking policy defines its own catchup window, which is not necessarily the maximum possible, because there is a tradeoff between window size and the reduction in number of frames displayed. If the window is

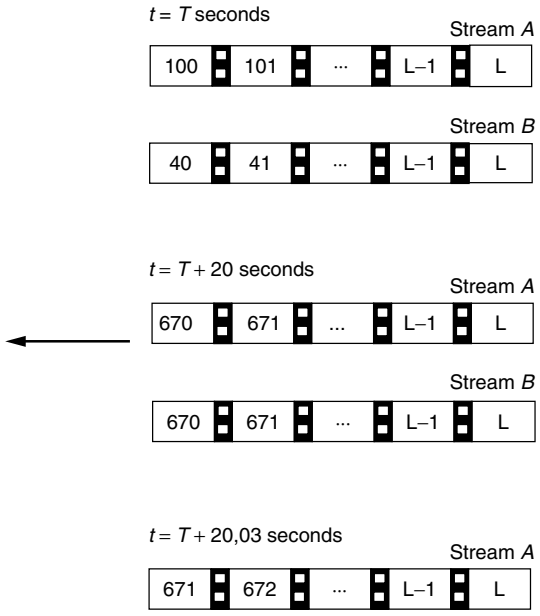


Figure 1. Stream A moves at a rate $S_{\max} = 31.5$ frames per second (FPS) while stream B moves at a rate $S_{\min} = 28.5$ FPS. At time T, the 100th frame of stream A and is the 40th frame of stream B are displayed. Twenty seconds later, at time $T + 20$ s, both streams displays the 670th frame. At $T + 20$ s a channel is released and only one stream is displayed for both viewers at a rate $S_n = 30$ frames.

large, a larger number of streams can be merged, but it is likely that the merging of these streams will occur too near the end of the movie, thus accruing less benefit. When the window is small, merges occur nearer the beginning of the movie, but fewer movies tend to be merged.

Various piggybacking policies are available. One of the simplest policies is the *odd–even* one, which tries to merge each pair of streams that arrived consecutively at the server [3]. On arrival, a stream is moving at the nominal rate, but it is then set to $S_{\min}$ if there is no stream ahead of it, or if there is one already moving at $S_{\max}$ Once a stream crosses the maximum catchup window at a rate of $S_{\min}$ and there is no stream behind it moving at $S_{\max}$, the display rate is reset to the nominal one. If there is a stream behind it moving at $S_{\max}$, however, the display rate remains at $S_{\min}$ to give the latter movie a chance to catch up.

The next policy to be discussed is the "greedy policy," which tries to merge as many videostreams as possible [3]. The greedy policy defines its catch up window based on the current frame. Whenever a merge occurs or the catchup window is crossed, a new window is computed. When a new video request is granted, the video rate is set as in the odd–even policy. Once a merge occurs, the display rate is set to the nominal one if there is no other stream in the new catchup window. Otherwise, the rate of the stream in front is set to $S_{\min}$ and the rate of the stream itself is set to $S_{\max}$. When it crosses the first catchup window, if there is another stream in front of it at the nominal rate, then the rate of this front stream is reduced to $S_{\min}$ and the rate of the following one is set to $S_{\max}$. Otherwise, therefore, if there is no stream ahead of it, the stream moves at the nominal rate.

Simple merging is another policy for merging groups; it guarantees that all streams in a group are eventually merged [3]. One stream is chosen to be the leader of the group, and all streams within this leader's maximum catchup window participate in the merging. As in the odd–even policy, on the arrival of a new stream, the rate of this new stream is set to $S_{\min}$, if there is no other stream within the maximum catchup window moving at $S_{\min}$. Otherwise, it moves at $S_{\max}$. Whenever it leaves the maximum catchup window, the rate is tuned to the nominal rate.

The *generalized simple merging* policy differs from the simple merging policy in that it computes window size to minimize the number of frames displayed by assuming that requests for video exhibition arrive according to a Poisson process with rate λ [4]. This window size is given by

$$W = -\frac{S_{\min}}{\lambda} + \sqrt{\left(\frac{S_{\min}}{\lambda}\right)^2 + 2\frac{LS_{\min}(S_{\max} - S_{\min})}{\lambda S_{\max}}}$$

The "snapshot policy" applies the generalized simple merging policy to a group of streams to form various merges in a given window [4]. At the end of this window, however, some streams may not have been merged. The rate of display of these streams must be adjusted so that further merges are possible. To do this, the stream positions will be represented by a binary tree, constructed in a bottom–up fashion using a dynamic programming solution. In such a tree, the streams are located at the leaf nodes, and merges are represented by interior nodes. The root node represents the final merge of all the streams. In this way, rates can be assigned so that all the remaining streams can merge.

The S2 policy reapplies the snapshot policy a second time to gain further reduction in the number of frames displayed [5]. It also constructs a merging tree, but in a top–down fashion, using a divide-and-conquer strategy, which allows the reduction of the computational complexity of the construction of the binary merging tree. It has been shown, however, that a third application of the snapshot policy does not bring additional benefits.

When comparing these policies, one can see that some of the policies are more efficient than others. The odd–even policy alone presents the least efficient performance, followed by the simple merging policy and by the generalized simple merging policy. The greedy policy outperforms the generalized simple merging policy, but the snapshot policy always produces even higher savings in relation to the number of frames displayed. S2, however, is the most effective of these piggybacking policies, especially since the savings on the number of frames displayed increases with the length of the movie and with the arrival rate of requests. For instance, S2 produces 9% savings over the number of frames displayed by the snapshot policy for a 30-min movie and an interarrival time of 500 s, but 90% savings for a 4 hour movie and an interarrival time of 15 s.

3. PATCHING

Patching policies exploit the client's buffer to reduce the waiting time required to watch a movie. They also

exploit the simultaneous reception of multiple channels. If a request to watch a movie is issued before a certain threshold time after the initiation of another exhibition of the same movie, the client joins the multicast group in the ongoing transmission with the current frames being stored in the client's buffer, namely, in the client's *settop box* (STB) buffer, until that client has seen the initial part recuperated from the video server. Once this has been seen, the client views the frames from the STB buffer (Fig. 2). Various patching policies have been proposed since the late 1990s.

In greedy patching, if the client's buffer is smaller than the initial part of the movie, then a new channel must be allocated for the individual display of the entire movie, with the viewer joining the multicast group of the ongoing transmission only for the storage of the final part to store in the buffer [6].

Grace patching also allocates a new channel for a transmission whenever the client's buffer is smaller than the initial clip, but this will involve multicast routing [6].

The *periodic buffer reuse with thresholding* (PBR) policy exploits the client's buffer to the maximum [7]. In this policy, the sequence of frames shown is drawn alternatively from the server and the ongoing transmission which was initiated the shortest time before the viewer's request. The buffer is initially filled with frames from the ongoing transmission while the client is watching frames from the server. The client then watches frames being drained from the buffer while it is being renewed by the server.

In the *greedy buffer reuse* (GBR) policy, parts of the video can be obtained from multiple ongoing transmissions, not only from that which was initiated the closest in time to the viewer's request [7]. It schedules the receipt of a sequence of frames as late as possible in order to utilize the client's buffer more efficiently. The buffer size and the current frame position of ongoing transmissions determine from where the sequence of frames will be fetched.

As with piggybacking, different patching policies result in different advantages and disadvantages. Greedy patching results in less data sharing than grace patching, since the latter increases the chances that a new request joins a multicast group. By making efficient use of the client's buffer, both PBR and GBR can provide greater bandwidth savings than grace patching, however, especially with large buffers, with GBR providing more savings than PBR. Under high loads, PBR may demand the display of 60% more frames per viewer than GBR. Such savings do not imply the involvement of a large number of channels, as no more than three channels are typically used during the exhibition of a single movie.

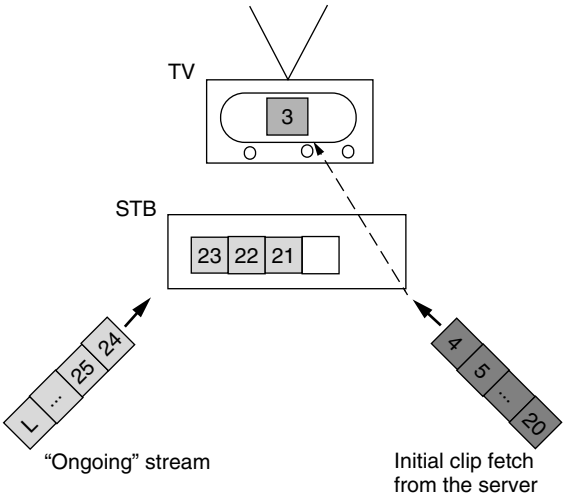


Figure 2. The initial part of the movie is fetched from the server while the frames from an ongoing transmission are stored in the STB for exhibition after the initial part is displayed.

4. BATCHING

In a VoD server with batching, requests are not granted as soon as they arrive. They are delayed so that several requests for the same film within a certain interval can be collected [8]. A single videostream is then allocated to the whole batch of requests (Fig. 3). If, on one hand, batching increases the server throughput (i.e., the rate of granted requests), on the other hand, users may not be willing to wait for long periods of time, and may cancel their requests (reneging).

Given that in batching users share the entire sequence of frames of the whole video, policies are compared by the number of users admitted into the system as well as by the number of users who renege.

Batching policies can be classified according to users' reneging behavior. Policies that do not consider reneging are first-come first-served (FCFS), maximum queue length (MQL), and maximum factor queue length (MFQL).

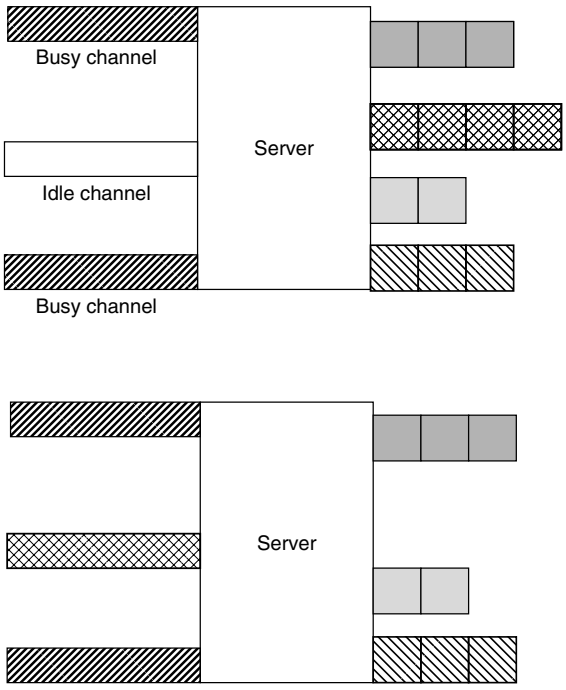


Figure 3. The figure in the top shows a server with 4 batches of requests waiting to be served and all channels busy except one, which was just released. The idle channel is then assigned to the largest batch.

FCFS serves requests according to their arrival order; it gives poor throughput, but treats all video requests equally. MQL allocates a video channel to the longest queue as soon as it becomes available; it produces considerably higher throughput than does FCFS. MFQL, a variation of the MQL policy, assigns a weighting factor to each queue, and allocates an available channel to the queue with the highest weighted length [9].

The second group of policies takes reneging into account. There are two batching schemes in this group: the Max_Batch scheme and the Min_Idle scheme. In the Min_Idle scheme, videos are classified as either "hot" or "cold" according to their popularity. Only hot videos are subject to batching. Moreover, hot videos have higher priority than cold videos for channel allocation. Two sets are defined: H and C. Hot videos, which have at least one request pending that exceeds a certain delay threshold, belong to the set H. Cold videos belong to the set C. Whenever a channel becomes available, a video in H is scheduled, either according to the longest queue criterion (IMQ) or to the highest expected number of losses (IML) criterion. If H is empty, a video in C is scheduled, regardless of how long the requests have been in queue. A cold video may migrate to the set H if any of its pending requests exceeds a certain threshold.

In the Max_Batch scheme, whenever a channel becomes available, a decision is made as to which queue (batch of requests) the channel should be allocated. A channel is allocated to a queue if and only if at least one of the enqueued requests exceeds a certain delay threshold. Two Max_Batch policies have been defined: the Max_Batch maximum queue length (BMQ) and the Max_Batch with minimum loss (BML) policies. BMQ allocates the available channel to the longest queue, whereas BML allocates the channel to the queue with the highest expected number of losses up to the next time a channel will become available.

The *look-ahead-maximize-batch* (LAMB) policy is a variant of the Max_Batch scheme. LAMB considers all videos in a server eligible for batching. Any channel will be allocated on demand, according to the expected number of losses [10].

LAMB considers a queue eligible for channel assignment only if one of its head-of-the-line (HoL) user is about to exceed his/her delay tolerance. In other words, a channel is allocated to a queue if and only if the HoL user is about to leave the system without being served. Moreover, instead of minimizing the number of losses expected by the next scheduling point, as is done in BML and IML, LAMB minimizes the losses in a batching window. This batching window is lower bounded by the current scheduling time and upper bounded by the most distant reneging time of a pending request.

LAMB maximizes the number of users admitted by considering all potential losses in the batching window. Whenever a user is about to leave the system, a decision is made about whether to allocate a channel to his/her queue. If the number of queues is less than the number of available channels, a channel is automatically allocated to the about-to-leave user's queue. Otherwise, an analysis of the implications of such an allocation, at the current scheduling time is made in relation to the admission of an overall larger number of users during the batching window. In other words, it is verified whether allocating a channel at the current scheduling time will cause a shortage of channels, which are associated with longer queues, at future scheduling times. Note that whenever a channel is allocated to a queue, all users in that queue are served at once. Otherwise, only the about-to-leave user is lost.

To maximize the number of users admitted during the batching window, it is necessary to determine when a channel should be allocated to a queue by considering all the information available at the current scheduling time, including the reneging time of all users, and the time when each channel will be released at the end of an exhibition.

Batching gives much better results than piggybacking. Allocating channels on demand to single users, even if temporarily, may result in a future shortage of channels, thus leading to a long-term rejection of a high number of users. Nevertheless, piggybacking produces fair systems, because it does not provide differentiated services for hot movies. If, on one hand, piggybacking makes it unnecessary for users to wait for a channel when it is available, on the other hand, batching significantly increases the server throughput, which is of paramount importance when deploying VoD services on a large scale.

LAMB overperforms all other existing batching policy, taking into consideration the number of admitted users (throughput) and the reneging probability, i.e., percentage of users who give up watching a movie. This trend is more striking for high loads (high arrival rate of requests) and in servers with a large number of video channels. For instance, LAMB admits the greatest number of users, 20% more than those admitted by MBQ, and the lowest reneging probabilities, 0.1 lower than MQL.

Although both batching and piggybacking furnish a single videostream to a group of viewers, they represent a clear trade-off between minimizing the delay to serve a request (piggybacking), and maximizing the server throughput (batching). One approach to enhance the throughput provided by batching is to merge streams in exhibition, which increases the number of channels available for new batches — in other words, to use a combination of batching and piggybacking. A system with both batching and piggybacking admits 20% more users than a system with batching only and produces reneging probabilities 0.05 lower than when only batching is used.

5. PERIODIC BROADCASTING

The top 10 or 20 most popular movies will be responsible for most of the requests for viewing. One possibility for coupling with bandwidth demands generated by these requests is to exhibit them from time to time (Fig. 4), taking a proactive approach rather than a reactive one (on demand), as done in techniques based on multicast [11].

Conventional broadcasting allocates a certain number of channels for showing a given video, staggering the beginning of each session evenly across the channels. The major drawback of such broadcasting is the number of channels needed to provide a low waiting time [12].

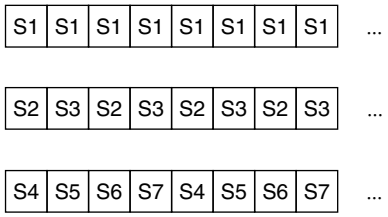


Figure 4. In periodic broadcasting, the video is divided into segments that are periodically broadcast in different channels.

One option is *periodic broadcasting*, which divides the video into a series of segments and broadcasts these periodically on dedicated channels. While the user is watching one segment, the following one is being transmitted so that it arrives just in time to assure continuous playback.

Three basic type of broadcasting protocols have been proposed. The first involves protocols dividing the video into increasing-size segments that are broadcast in channels of the same capacity. In the *pyramid broadcasting* (PB) scheme, for example, the segment size follows a geometric series, where the nth segment of each video is transmitted sequentially in each channel [13]. Although a low waiting time can be assured, high transmission rates are required, and this implies in high I/O bandwidth demand of and large STB buffers.

Permutation-based pyramid broadcasting tries to overcome the major drawbacks of the classical pyramid broadcasting protocol by dividing each channel into a specific number of subchannels. The substreams are staggered with each other to perform the same kind of timing as in PB, although the transmission rate is lower. Since the same geometrical series is used to divide the video, the client buffer cannot be reduced because the last segment is quite large [14].

Skyscraper broadcasting (SB) is similar except that the size of the segments is determined by a recursive function which generates the following sequence [1, 2, 2, 5, 5, 12, 12, 25, 25, 52, 52, ...], where the first segment is the unit size for all following segments. Each segment is broadcast periodically in a specific channel, and the client will need to download from at most two streams at the same time. The client's buffer size must still be maintained to accommodate the final film segment [15].

Fast broadcasting is similar to other pyramid schemes in that equal capacity channels are used, but the segments broadcast are of equal size, where group of segments are transmitted together. Groups of $2i$ contiguous segments are transmitted in the ith channel. One advantage of this protocol is that no buffer is required at the client [16].

In the second family of protocols, videos are divided into equal sized segments transmitted on channels of decreasing capacity. The first of these to be considered here is *harmonic broadcasting* (HB), in which the ith segment is divided into i subsegments. The first segment is repeatedly broadcast in the first channel and contiguous subsegments of the other segments are periodically transmitted on the channel dedicated to these segments; each channel has a broadcasting capacity inversely proportional to the sequential order of the segment. The main idea of HB is that when the client is ready to receive the ith segment he will already have received the $i - 1$ subsegment, and the last subsegment will be received while the client retrieves the first segments from the buffer. The bandwidth demand of HB increases harmonically as a function of the number of segments of the video, and the storage demand is about 40% of the entire video. However, HB does not always deliver data on time [17].

All the variants of HB overcome this timing problem. In *caution harmonic broadcasting* (CHB), for example, the first segment is transmitted as in HB, whereas the second and the third are transmitted in a second channel, while the remaining segments are transmitted in other channels with a capacity inversely proportional to the segment order minus one [18].

Another option for harmonic broadcasting is *quasiharmonic broadcasting* (QHB). Again, the first segment is transmitted as for HB and CHB. The remaining segments are divided in such a way that the ith segment is divided into $im - 1$ subsegments, where m is a positive ith integer, but the subsegments are not transmitted in order. The first $i - 1$ subsegments of the segment are transmitted at the end of the segment slot and the remaining $i(m - 1)$ subsegments are transmitted according to a specific rule, so that the client always has the first $i - 1$ subsegments stored in his buffer. Although QHB demands more bandwidth than do HB demands, the overhead tends to be compensated for an increase in the number of subsegments [18].

The third group consists of protocols which are a hybrid of pyramid-based and harmonic protocols. Like harmonic protocols, *pagoda broadcasting* partitions the video into equal-sized segments, but unlike them, these segments are broadcast at the same rate, although with different periodicity [19]. The effect of channel dedication is achieved by time-division multiplexing. The main advantage of this protocol is that it avoids the problems due to low transmission rates, although the determination of proper segment-to-channel mapping and periodicity is critical. The *new pagoda broadcasting* protocol uses a more sophisticated segment-to-stream mapping than that used by pagoda to further reduce the bandwidth demands [20].

6. CONCLUSIONS

Video-on-demand has been considered "the" video application for the future broadband multimedia networks. Considerable effort has been invested in making it efficient, since the huge amounts of bandwidth needed to serve a large population preclude the deployment of VoD on a large scale. Moreover, the implementation of VoD services has to be competitive with traditional video rental and pay-per-view. In this article, various proactive and reactive bandwidth reduction techniques have been described, techniques that can be used jointly with the distribution of additional servers throughout the network. Their effectiveness depends on viewers' behavior, since at high demand rates schemes based on multicasting tend to transmit videostreams with the same periodicity as schemes based on broadcasting. On the

other hand, periodic broadcasting wastes bandwidth if the rate of request is not high. The most appropriate way for implementing video-on-demand can be determined only when the dimensions of use have been established, and will be understood only when services are available at large. Moreover, providing interactiveness (VCR capability) implies additional costs in terms of careful dimensioning and signaling [21,22].

BIOGRAPHY

Nelson Fonseca received his Electrical Engineer (1984) and M.Sc. in Computer Science (1987) degrees from The Pontifical Catholic University at Rio de Janeiro, Brazil, and the M.Sc. (1993) and Ph.D. (1994) degrees in Computer Engineering from The University of Southern California (Los Angeles). Since 1995 he has been affiliated to the Institute of Computing of the State University of Campinas, Brazil, where is currently an Associate Professor.

He is the recipient of Elsevier Editor of the Year 2000, the 1994 USC International Book Award, and the Brazilian Computing Society First Thesis and Dissertations Award. Mr. Fonseca is listed in Marqui's *Who's Who in the World* and *Who's Who in Science and Engineering*.

He served as Editor-in-Chief for the *IEEE Global Communications Newsletter* from 1999 to 2002. He is an Editor for *Computer Networks*, an Editor for the *IEEE Transactions on Multimedia*, an Associate Technical Editor for the *IEEE Communications Magazine*, and an Editor for the *Brazilian Journal on Telecommunications*.

BIBLIOGRAPHY

1. J. P. Nussbaummer, B. V. Patel, F. Schaffa, and J. P. G. Sterbenz, Networking requirements for interactive video on demand, *IEEE J. Select. Areas Commun.* 779–787 (June 1995).
2. N. L. S. Fonseca, C. M. R. Franco, and F. Schaffa, Network design for the provision of distributed home theatre, *Proc. IEEE Int. Conf. Communications*, 1997, pp. 816–821.
3. L. Golubchik, J. C. S. Lui, and R. Muntz, Adaptive piggybacking: A novel technique for data sharing in video-on-demand storage servers, *Multimedia Syst.* **4**(3): 140–155 (1996).
4. C. C. Aggarwal, J. Wolf, and Philip S. Yu, On optimal piggybacking merging policies for video-on-demand systems, *Proc. ACM Sigmetrics* **24**: 200–209 (1996).
5. R. A. Façanha, N. L. S. Fonseca, and P. J. Rezende, The S2 piggybacking policy, *Multimedia Tools Appl.* **8**(3): 371–383 (May 1999).
6. K. Hua, Y. Cai, and S. Sheu, Patching: A multicast technique for true video-on-demand services, *Proc. 6th ACM Int. Multimedia Conf.*, 1998, pp. 191–200.
7. S. Sen, L. Gao, J. Rexford, and D. Towsley, Optimal patching schemes for efficient multimedia streaming, *Proc. IEEE NOSSDAV*, 1999.
8. H. Shachnai and P. S. Yu, Exploring wait tolerance in effective batching for video-on-demand scheduling, *Multimedia Syst. J.* **6**(6): 382–394 (Dec. 1998).
9. A. Dan, D. Sitaram, and P. Shahabuddin, Dynamic batching policies for an on-demand video server, *Multimedia Syst.* **4**: 112–121 (1996).
10. N. L. S. Fonseca and R. A. Façanha, The look-ahead-maximize-batch batching policy, *IEEE Trans. Multimedia* **4**(1): 1–7 (2002).
11. A. Hu, Video-on-demand broadcasting protocols: A comprehensive study, *Proc. IEEE InfoCOM*, 2001.
12. K. Almeroth and M. Ammar, The use of multicast delivery to provide a scalable and interactive video-on-demand service, *IEEE J. Select. Areas Commun.* **14**(5): 1110–1122 (Aug. 1996).
13. S. Viswanathan and T. Imielinski, Pyramid Broadcasting for video on demand service, *IEEE Multimedia Computing and Networking Conf.*, San Jose, CA, 1995, Vol. 2417, pp. 66–77.
14. C. Aggarwal, J. Wolf, and P. Yu, A permutation-based pyramid broadcasting scheme for video-on-demand systems, *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, 1996.
15. K. Hua and S. Sheu, Skyscraper broadcasting: A new broadcasting scheme for metropolitan video-on-demand systems, *Proc. ACM SIGCOMM'97*, 1997, pp. 89–100.
16. L. Juhn and L. Tseng, Fast data broadcasting and receiving scheme for popular video service, *IEEE Trans. Broadcast.* **44**(1): 100–105 (March 1998).
17. L. Juhn and L. Tseng, Harmonic broadcasting for video-on-demand service, *IEEE Trans. Broadcast.* **43**(3): 268–271 (Sept. 1997).
18. J. Paris, S. Carter, and D. Long, Efficient broadcasting protocols for video on demand, *Proc. 6th Int. Symp. Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, July 1998, pp. 127–132.
19. J. Paris, S. Carter, and D. Long, A hybrid broadcasting protocol for video on demand, *Proc. 1999 Multimedia Computing and Networking Conf.*, 1999, pp. 317–326.
20. J. Paris, A simple low-bandwidth broadcasting protocol for video-on-demand, *Proc. 8th Int. Conf. Computer Communications and Networks* (IC3N'99), 1999, pp. 118–123.
21. N. L. S. Fonseca and H. K. Rubinszjtein, Dimensioning the capacity of interactive video server, *Proc. Int. Teletraffic Congress 17*, 2001, pp. 383–394.
22. J. K. Dey-Sircar, J. D. Salehi, J. F. Kurose, and D. Towsley, Providing VCR capabilities in large-scale video servers, *Proc. 2nd ACM Int. Conf. Multimedia*, 1994, pp. 25–36.

BCH CODES — BINARY*

Arnold M. Michelson
Allen H. Levesque
Marlborough, Massachusetts

1. INTRODUCTION

Bose–Chaudhuri–Hocquenghem (BCH) codes are a broad class of cyclic block codes used for detection and correction

* Preparation of this article supported in part by the Raytheon Corporation.

of transmission errors in digital communications systems. This article describes binary BCH codes while the succeeding article treats *nonbinary BCH codes*. These codes were originally described in two papers, the first by A. Hocquenghem in 1959 [1] and the second by R. C. Bose and D. K. Ray-Chaudhuri in 1960 [2]. It would therefore be more accurate to say "HBR-C codes," but the commonly used abbreviation is BCH.

Error control coding is a field finding wide application in modern digital communications. Described in simple terms, error control coding involves adding redundancy to transmitted data to provide a means for detecting and correcting errors that inevitably occur in any real communication process. Coding can be used to provide a desired level of accuracy in data transmitted over a noisy communication channel and delivered to a user. There are, however, other ways to achieve accurate transmission of data.

For example, an alternative to the use of coding is to provide sufficient signal energy to ensure that uncoded information is delivered with the required accuracy. The energy needed might be achieved by setting signal power to a sufficiently high level or, if power limitations prevail, by using some form of diversity transmission and reception. However, error control coding may provide the required accuracy with less energy than uncoded operation and can be the economically preferred solution in spite of an increase in system complexity. Cost savings through the use of coding can be dramatic when very high accuracy is needed and power is expensive. Furthermore, in some applications, the saving in signal power is accompanied by important reductions in size and weight of the communication equipment.

To describe BCH codes, it is first necessary to provide some background in finite fields, extension fields, and polynomials defined on finite fields. This is done in Section 2. Binary BCH codes are described as cyclic codes in Section 3, and the design of generator polynomials, encoders, and decoders are also covered. In the succeeding article, nonbinary BCH codes are treated. Reed–Solomon (RS) codes are the most widely used class of nonbinary BCH codes, and the design, encoding, and decoding of RS codes are treated there.

2. MATHEMATICAL BACKGROUND

This section describes the essential mathematics needed for understanding the design and implementation of BCH codes. BCH codes are cyclic codes and are conveniently represented as polynomials with coefficients in a *finite field*. We begin with a discussion of finite fields.

2.1. Finite Fields

A *field* is a set of elements with two operations defined, addition (+) and multiplication (·). Two other operations, subtraction and division, are implied by the existence of inverse elements under the defining operations. Stated more completely, the elements in a field F, taken together with the operations + and ·, must satisfy the following conditions:

1. F is closed under the two operations, that is, the sum or product of any two elements in F is also in F.
2. For each operation, the associative and commutative laws of ordinary arithmetic hold, so that for any elements u, v, and w in F:

$$(u+v)+w = u+(v+w)$$

$$u+v = v+u$$

$$(u\cdot v)\cdot w = u\cdot(v\cdot w)$$

$$u\cdot v = v\cdot u$$

3. Connecting the two operations, the distributive law of ordinary arithmetic holds, so that

$$u\cdot(v+w) = u\cdot v + u\cdot w$$

for any u, v, and w in F.

4. F contains a unique additive identity element 0 and a unique multiplicative identity, different from 0 and written as 1, such that

$$u+0 = u$$

$$u\cdot 1 = u$$

for any element u in F. The two identity elements are the minimum elements that any field must contain. We call these two elements zero and unity.

5. Each element u in the field has a unique additive inverse, denoted by $-u$, such that

$$u+(-u) = 0$$

and, for $u \neq 0$, a unique multiplicative inverse, denoted by u^{-1}, such that

$$u\cdot u^{-1} = 1$$

From these observations, the inverse operations subtraction (−) and division (÷) are defined by

$$u - v = u + (-v), \quad \text{any } u, v \text{ in } F$$

$$u \div v = u\cdot(v^{-1}), \quad v \neq 0$$

where $-v$ and v^{-1} are the additive and multiplicative inverses, respectively, of v.

Thus a field provides four elementary operations and the familiar rules of ordinary arithmetic. Common examples of fields are the set of real numbers and the set of rational numbers under ordinary addition and multiplication. The set of real numbers equal to or greater than zero does not constitute a field under the rules of ordinary arithmetic, since the set does not include additive inverses for nonzero numbers. Similarly, the set of integers under ordinary arithmetic is not a field, since integers other than 1 do not have multiplicative inverses in the set.

The number of elements in a *field*, called the *order* of the field, may be finite or infinite, but we consider only

fields having a finite number of elements. A field having a finite number of elements is called a *finite field* and is denoted by GF(q), where q is the number of elements in the field. The notation is related to the designation *Galois field*, which is used interchangeably with "finite field" in the literature. A finite field GF(p^m) exists for any p^m, where p is a prime and m is an integer. The simplest example of a finite field is a *prime field*, GF(p), consisting of the set of all integers modulo p, where p is a prime number greater than 1 and the addition and multiplication operations are addition and multiplication modulo p. The simplest prime field is GF(2), which contains only the zero and unity elements 0 and 1. As another example, the addition and multiplication tables for GF(5) are shown below.

GF(5) Addition

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

GF(5) Multiplication

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

2.2. The Primitive Element

An important property of finite fields is that every finite field GF(q) contains at least one *primitive element*, called α, which has the property that the $q-1$ powers of α are the $q-1$ nonzero elements of the field. This means that the nonzero field elements can be represented as $\alpha, \alpha^2, \ldots, \alpha^{q-1}$.

If we take an arbitrary nonzero element β in the field and raise it to successive powers, we eventually arrive at some exponent e such that $\beta^e = 1$. For an arbitrary β in the field, the smallest positive integer e such that $\beta^e = 1$ is called the *order of the element*. (This is not to be confused with the *order of the field*, defined as the number of elements in the field, which is equal to q in the present discussion.) In the generation of the nonzero elements of GF(q) as powers of a primitive element α, we always find that $\alpha^{q-1} = \alpha^0 = 1$, but no smaller power of α equals 1, so that the order of a primitive element is $q-1$. In general, the various elements of the field can have different orders, but there is a theorem (due to Lagrange), stating that the order of an arbitrary element must be either $q-1$ or a divisor of $q-1$.

For example, consider the prime field GF(5), which consists of the integers modulo 5. Since $q-1=4$, we anticipate that the orders of various elements can be 1, 2, or 4. The element 1 has order 1. Taking successive powers of 2, we find $2^1 = 2, 2^2 = 4, 2^3 = 8 = 3 \bmod 5$, $2^4 = 6 = 1 \bmod 5$, and thus 2 has order 4. One also finds that 3 has order 4, and 4 has order 2. Therefore 2 and 3 are primitive elements of GF(5), while 1 and 4 are *nonprimitive* elements.

Since all the nonzero field elements can be expressed as the first $q-1$ powers of a primitive element α, we note that we can represent the field elements in terms of their exponents. The exponents are in effect *logarithms to the base* α. As in ordinary arithmetic, the logarithm of zero is undefined, although for convenience, the notation $0 = \alpha^{-\infty}$ is often used. Below we show the logarithm tables for GF(5) formed with $\alpha = 2$ and $\alpha = 3$:

β	$\log_2 \beta$	$\log_3 \beta$
0	$-\infty$	$-\infty$
1	0	0
2	1	3
3	3	1
4	2	2

Just as in ordinary arithmetic, multiplication of field elements can be done by adding logarithms. For example, in GF(5), using $\alpha = 2$, we can multiply 2 times 4 by adding the logarithms $(1+2=3)$ and looking up the resulting element (3 mod 5) in an antilogarithm table. In coding implementations, finite-field multiplications are often done with logarithm and antilogarithm tables.

2.3. Vectors of Field Elements and Polynomials Defined on Finite Fields

To encode and decode BCH codes, we need to find an algebraic system for doing calculations with vectors, or m-tuples, of finite-field elements and a representation for field elements that is convenient for implementation in a digital machine. First note that we can enumerate all the m-tuples of elements in a field GF(q), q^m in number, and note that they in fact constitute an m-dimensional vector space over GF(q). Thus we can add and subtract vectors, using vector (element-by-element) addition and subtraction in GF(q), and the result in every case is another vector in the vector space. However, we shall also want to do multiplication and division of vectors. To accomplish this, we associate each vector with a polynomial having coefficients corresponding to the elements in the vector. For example, the set of four 2-tuples on GF(2) can be represented by 0, 1, x, and $x+1$, corresponding to 00, 01, 10, and 11, respectively. Clearly, we can do term-by-term addition of the polynomials just as we would add the vectors. All we have done is replace the set of all 2-tuples defined on GF(2) with the set of all degree-1 polynomials defined on GF(2).

Just as we have closure with addition of vectors, we must also have closure under multiplication. In fact, if we can find a way to multiply the polynomials that conforms to all the properties of multiplication in a finite field, we will have constructed a finite field with q^m elements. First, we want the product of any two polynomials in the set to be another polynomial in the set (closure). This is no problem if the product is a polynomial of degree $(m-1)$ or less. But, what do we do with a polynomial product of degree m or greater? Clearly, we can reduce the product by taking its remainder with respect to a fixed polynomial of degree m. The remainder will always be of degree $m-1$ or less, and closure is achieved. However, we need to know what sort of polynomial to use in this reduction so that the other properties of a finite field are assured.

We can gain some insight into this question by observing that the product of any two nonzero field elements must be nonzero. For example, let two nonzero elements α^i and α^j be represented by $a(x)$ and $b(x)$, respectively, each of degree $m-1$ or less. Then, assuming

a *reduction polynomial* $p(x)$ of degree m, we can write the product $\alpha^i\alpha^j$ as

$$\alpha^i\alpha^j = a(x)b(x) \bmod p(x)$$

Now, let us set this product equal to zero and see what type of reduction polynomial would allow this to happen. That is, we write

$$a(x)b(x) \bmod p(x) = 0$$

or equivalently

$$a(x)b(x) = c(x)p(x) \tag{1}$$

which says that the left-hand side of Eq. (1) must be evenly divisible by $p(x)$. Now, if $p(x)$ is *factorable*, that is, expressible as the product of two or more polynomials of degree $m-1$ or less, there may well be polynomials $a(x)b(x)$ that are evenly divisible by $p(x)$. However, if $p(x)$ is chosen to be a degree-m *irreducible polynomial* (a polynomial that cannot be factored), then $p(x)$ must be a factor of either $a(x)$ or $b(x)$. We can readily see that neither factoring is possible, since the polynomials $a(x)$ and $b(x)$ are each of degree $m-1$ or less and $p(x)$ is of degree m. We, therefore, conclude that if $p(x)$ is chosen to be an irreducible polynomial of degree m, the equality in Eq. (1) cannot be satisfied unless $a(x)$ or $b(x)$ equals zero, in which case $c(x) = 0$. By similar arguments, we could show that the requirements for uniqueness of the products $a(x)b(x)$, and hence the uniqueness of the inverse for each polynomial, again results in choosing the reduction polynomial $p(x)$ to be an irreducible degree-m polynomial in GF(q).

Returning now to the simple example of the four 2-tuples defined on GF(2), we can use $p(x) = x^2 + x + 1$, since $x^2 + x + 1$ cannot be factored into any lower-degree polynomials on GF(2). (The only candidates for factors are x and $x+1$, and it is easily verified that none of the products of these two polynomials equals $x^2 + x + 1$.) With $p(x)$ chosen, we can now write the multiplication table for the degree-1 binary polynomials as follows:

$\cdot$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	$x+1$	1
$x+1$	0	$x+1$	1	x

We see from the multiplication table that each nonzero polynomial has a unique multiplicative inverse, x being the inverse of $x+1$ and vice versa, while 1 is its own inverse, as always. Thus, we have defined a representation of a finite field with four elements, which we denote by GF(4).

We complete this example by describing GF(4) in terms of a primitive element. We can test for a primitive element simply by taking a nonzero element other than 1 and raising it to successive powers until we find its order. For example, testing x, we have $x^1 = x, x^2 = x+1, x^3 = x^2 + x = (x+1) + x = 1$, where calculation of x^2 and x^3 required reduction modulo $x^2 + x + 1$. We see therefore that x has order $e = 3$, and since $q - 1 = 3$, x is a primitive element. It can be seen that the polynomial $x+1$ is primitive as well. Having found two primitive elements in GF(4), we can now use either one to generate a list of the nonzero field elements as powers of α. This is shown here with a table of field elements for each primitive element. The table also shows a representation of the four elements in GF(4) that is convenient for implementation in a digital machine. With each polynomial, we associate a binary 2-tuple, for example, $0 = 00$ and $x + 1 = 11$. Addition of the digital representations of field elements is then conveniently implemented with the exclusive OR operation:

REPRESENTATIONS FOR FIELD ELEMENTS

$\alpha = x$					$\alpha = x+1$				
$\alpha^{-\infty}$	=	0	=	00	$\alpha^{-\infty}$	=	0	=	00
α^0	=	1	=	01	α^0	=	1	=	01
α^1	=	x	=	10	α^1	=	$x+1$	=	11
α^2	=	$x+1$	=	11	α^2	=	x	=	10

Thus, we have a complete representation for elements in GF(4) and a consistent set of operations for addition and multiplication of elements. For multiplication, the appropriate logarithm tables can be used. In the next section, we generalize these results and define somewhat more formally the properties of fields constructed with m-tuples of field elements.

2.4. Extension Fields and Primitive Polynomials

In general, a finite field GF(p^m) exists for any number p^m, where p is a prime and m is a positive integer. For $m = 1$, we have the prime number fields GF(p). The fields GF(p^m) for $m > 1$ are commonly called *prime-power fields*, where p is the *characteristic of the field*. That is, p is the smallest integer such that

$$\sum_{i=1}^{p} \alpha^0 = 0$$

where α^0 is the multiplicative identity element. For fields of characteristic 2, each element is its own additive inverse and a minus sign is unnecessary.

The relationship between GF(p) and GF(p^m) is such that GF(p) is a *subfield* of GF(p^m); that is, the elements of GF(p) are a subset of the elements in GF(p^m), the subset itself having all the properties of a finite field. Equivalently, GF(p^m) is called an *extension field*, or simply an *extension*, of GF(p).

The procedure followed previously for constructing GF(4) from GF(2) serves as an example of how one constructs an extension field from a subfield. The procedure generalizes in a straightforward way to any extension field GF(p^m). That is, we represent elements in GF(p^m) as the p^m polynomials of degree $m-1$ or less with coefficients in GF(p). Polynomials are added by adding coefficients of corresponding powers of x, addition being done in GF(p). To define multiplication, a degree-m irreducible polynomial over GF(p) is selected and a primitive element α for GF(p^m) is found. Then the polynomials corresponding to the $p^m - 1$ distinct powers of

α are constructed. We see that the irreducible polynomial $p(x)$ provides the key link between the addition and multiplication tables and thus fixes the structure that allows us to define the two arithmetic operations and their inverses in a consistent way. Thus, we can say that the set of all polynomials in GF(p) reduced with respect to a degree-m irreducible polynomial over GF(p) forms the field GF(p^m). The role of the irreducible polynomial is seen to be directly analogous to the use of a prime number p to define the finite field GF(p).

Note that a polynomial $p(x)$ of degree m with coefficients in GF(p) is said to be irreducible if it is not divisible by any polynomial with coefficients in GF(p) of degree less than m and greater than zero. For example, consider the polynomial $p(x) = x^3 + x + 1$ having degree 3 and coefficients in GF(2). We can quickly convince ourselves that $x^3 + x + 1$ is not factorable in GF(2), as follows. If it is factorable, it must have at least one factor of degree 1. Of course x is not a factor of $p(x)$, since the lowest-order term in $p(x)$ is $x^0 = 1$. Thus, the only candidate is $x + 1$, but if this were a factor, then $x = 1$ would be a root of $p(x)$. It is easily verified that this is not the case, since $p(x)$ has an odd number of terms, and therefore, $p(x)$ evaluated at $x = 1$ sums to 1 mod 2. Therefore, $p(x) = x^3 + x + 1$ is irreducible in GF(2).

Therefore, we are able to generate the 3-tuples representing elements of GF(2^3) simply by listing all 2^3 polynomials of the form $a(x) = a_2x^2 + a_1x + a_0$ and taking each 3-tuple as the vector of coefficients a_2, a_1, a_0. It is convenient to list the polynomials $a(x)$ in a sequence that automatically provides a consecutive ordering by logarithms. This can be done here by using the polynomial $a(x) = x$ as the primitive element, multiplying repeatedly by x, and reducing the result modulo $p(x)$. This is shown in Table 1. We see from the table that by forming successive powers of x, reduced modulo $x^3 + x + 1$, we obtain all the polynomials defining the nonzero elements of GF(2^3). In order for the procedure to generate the full list of 3-tuples, it is necessary that x be a primitive element, which is clearly the case in this example.

Although $p(x) = x^3 + x + 1$ is an irreducible binary polynomial and consequently has no roots in GF(2), it does have roots defined in an extension field. In fact, it is a simple matter to find one of its roots, since from Table 1, we see that we could as easily have generated the table using powers of α letting $\alpha = x$ and $\alpha^3 = \alpha + 1$, and therefore α is a root of $p(x)$. An irreducible polynomial having a primitive element as a root is called a *primitive irreducible polynomial* or simply a *primitive polynomial*. While an irreducible polynomial with coefficients in GF(p) has no roots in GF(p), it has roots in the extension field GF(p^m). In fact, the degree-m polynomial $p(x)$ must have exactly m roots in the extension field GF(p^m).

Table 1. A Representation of GF(2^3) Generated from $x^3 + x + 1$

Zero and Powers of x		Polynomials Over GF(2)		Vectors Over GF(2)
0	=	0	=	000
x^0	=	1	=	001
x^1	=	x	=	010
x^2	=	x^2	=	100
x^3	=	$x + 1$	=	011
x^4	=	$x^2 + x$	=	110
x^5	=	$x^2 + x + 1$	=	111
x^6	=	$x^2 + 1$	=	101

It is important to note that not all irreducible polynomials are primitive and both can be used to generate a representation for a finite field. However, it is convenient to use a primitive polynomial since the field elements can be generated as powers of x. As a practical matter, tables of irreducible polynomials with primitive polynomials identified are available in the literature [3].

In summary, to construct a representation of GF(p^m), we go to a table of irreducible polynomials on GF(p), and find a polynomial $p(x)$, preferably primitive, of degree m. We then generate the list of p^m polynomials modulo $p(x)$ and take the vectors of polynomial coefficients as m-tuples representing the elements of GF(p^m). Consistent addition and multiplication tables can then be constructed for GF(p^m). The addition table is formed by adding corresponding elements in m-tuples, modulo p. The multiplication table can be formed by adding exponents of α. The addition and multiplication tables for GF(2^3), formed with the use of Table 1, are shown in Table 2. Note that we constructed the addition and multiplication tables using powers of α although we expressed field elements using polynomials in x in Table 1. However, since $x^3 + x + 1$ is a primitive polynomial, it has α as a root, so that $\alpha^3 + \alpha + 1 = 0$. Thus, Table 1 might as easily have been written with α replacing x, as was observed earlier.

It should be noted that while the multiplication table is most easily constructed by addition of exponents of α, it can also be constructed by multiplying the polynomial representations of two elements and reducing the result modulo $p(x)$. For example, we can use Table 1 to calculate

$$\begin{aligned}\alpha^2\alpha^4 &= x^2(x^2 + x) \bmod x^3 + x + 1\\ &= x^4 + x^3 \bmod x^3 + x + 1\\ &= x^2 + x + x + 1\\ &= x^2 + 1\\ &= \alpha^6\end{aligned}$$

This is analogous to generating the multiplication table for a prime field GF(p) by multiplying integers and reducing the product modulo p.

2.5. Key Properties of Irreducible Polynomials

In Section 2.3, we utilized certain properties of irreducible polynomials to provide a consistent set of rules for performing addition and multiplication in a finite field GF(p^m). It is now necessary to present further details on the properties of these polynomials, which form the basis for describing the structure of cyclic codes. In our discussion of binary codes, we confine attention to fields of characteristic 2, GF(2^m).

Our discussion concentrates on polynomials that are irreducible in GF(2), that is, degree-m binary polynomials

Table 2. Addition and Multiplication Tables for GF(2^3)

$+$	0	1	α	α^2	α^3	α^4	α^5	α^6
0	0	1	α	α^2	α^3	α^4	α^5	α^6
1	1	0	α^3	α^6	α	α^5	α^4	α^2
α	α	α^3	0	α^4	1	α^2	α^6	α^5
α^2	α^2	α^6	α^4	0	α^5	α	α^3	1
α^3	α^3	α	1	α^5	0	α^6	α^2	α^4
α^4	α^4	α^5	α^2	α	α^6	0	1	α^3
α^5	α^5	α^4	α^6	α^3	α^2	1	0	α
α^6	α^6	α^2	α^5	1	α^4	α^3	α	0

$\cdot$	0	1	α	α^2	α^3	α^4	α^5	α^6
0	0	0	0	0	0	0	0	0
1	0	1	α	α^2	α^3	α^4	α^5	α^6
α	0	α	α^2	α^3	α^4	α^5	α^6	1
α^2	0	α^2	α^3	α^4	α^5	α^6	1	α
α^3	0	α^3	α^4	α^5	α^6	1	α	α^2
α^4	0	α^4	α^5	α^6	1	α	α^2	α^3
α^5	0	α^5	α^6	1	α	α^2	α^3	α^4
α^6	0	α^6	1	α	α^2	α^3	α^4	α^5

that have no factors of degree less than m and greater than 0. It has already been stated that every degree-m polynomial $f(x)$ on GF(2) has m roots (as in ordinary arithmetic), and if $f(x)$ is irreducible all m roots are in the extension field GF(2^m). The properties of these roots in extension fields are of central importance in the theory of cyclic codes, and thus we summarize the key points required in the subsequent discussion. For convenience of presentation, certain points made earlier are repeated in this summary.

2.5.1. Properties of Polynomials Defined on Finite Fields

1. Given a polynomial $f(x)$ with coefficients in GF(2), we say that β is a root of $f(x)$ if and only if $f(\beta) = 0$, where β is an element of GF(2), or some extension GF(2^m). The multiplications and additions required for the evaluation of the polynomial can be performed in the consistent arithmetic system GF(2^m), since GF(2) is contained in any of its extensions.
2. Every polynomial of degree m has exactly m roots, some of which may be repeated.
3. For any m, there is at least one degree-m polynomial on GF(2) that is irreducible.
4. If $f(x)$ is a degree-m irreducible polynomial ($m \geq 2$) on GF(2), it has no roots in GF(2), but all its roots lie in some extension of GF(2). If $f(x)$ has a root that is a primitive element of GF(2^m), $f(x)$ is called a *primitive irreducible polynomial*, or simply a *primitive polynomial*. Since it can be shown that all the roots of an irreducible polynomial are of the same order, all the roots of a primitive polynomial are primitive. For any m, there is at least one irreducible polynomial on GF(2) that is primitive.
5. For every element β in an extension field GF(2^m), there is a polynomial on GF(2), called the *minimal polynomial* of β, which is the lowest-degree *monic* (the highest-order term has coefficient 1) polynomial having β as a root. Of course, all polynomials defined on GF(2) are monic. Minimal polynomials, sometimes called *minimum functions*, have an important place in the design of cyclic codes, and we shall have more to say about them.
6. If $f(x)$ is an irreducible degree-m polynomial on GF(2) and has a root β, then $\beta, \beta^2, \beta^4, \beta^8, \ldots, \beta^{2^{m-1}}$ are all the roots of $f(x)$. This is an important property relating to the structure of cyclic codes.

Associated with every element β in an extension field GF(2^m) is its minimal polynomial $m_\beta(x)$ with coefficients in GF(2). There is a minimal polynomial for every element in the field, even if the element lies in GF(2) itself. The important properties of minimal polynomials are summarized as follows.

2.5.2. Properties of Minimal Polynomials. The minimal polynomial $m_\beta(x)$ of any field element β must be irreducible. If this were not the case, one of the factors of $m_\beta(x)$ would have β as a root and would be of lower degree than $m_\beta(x)$ and contradict the definition. In addition

1. The minimal polynomial of β is unique, that is, for every β there is one and only one minimal polynomial of β. However, different elements of GF(2^m) can have the same minimal polynomial. (See property 6 of polynomials defined on finite fields.)
2. For every element in GF(2^m), the degree of the minimal polynomial over GF(2) is at most m.
3. The minimal polynomial of a primitive element of GF(2^m) has degree m and is a primitive polynomial.

Consider again the case of the extension field GF(2^2), which we represent as the polynomials of degree 1 or less, modulo the irreducible polynomial $y^2 + y + 1$. We have

$$\beta_0 = 0$$

$$\beta_1 = 1$$

$$\beta_2 = y$$

$$\beta_3 = y + 1$$

The minimal polynomials of β_0 and β_1 are simply

$$m_{\beta_0}(x) = x \quad \text{and} \quad m_{\beta_1} = x + 1$$

To find the minimal polynomial of $\beta_2 = y$, we use property 6 in Section 2.5, which tells us that the irreducible degree-2 polynomial having β_2 as a root has β_2 and as β_2^2 roots, and no others. Therefore, we can write

$$\begin{aligned} m_{\beta_2}(x) &= (x - y)(x - y^2) \\ &= (x + y)(x + y + 1) \\ &= x^2 + xy + x + yx + y^2 + y \\ &= x^2 + x + 1 \end{aligned}$$

where we use $y^2 = y + 1$ to reduce powers of y greater than unity. Similarly

$$\begin{aligned} m_{\beta_3}(x) &= (x + y + 1)(x + y^2 + 1) \\ &= x^2 + xy^2 + x + yx + y^3 + y + x + y^2 + 1 \\ &= x^2 + x + 1 \end{aligned}$$

Thus, we see that β_2 and β_3 have the same minimal polynomial. (We could have shown this directly by noting that $\beta_2^2 = \beta_3$.) Sets of elements having this property are called *conjugates*.

This example is given only to provide a clearer explanation of the concept of a minimal polynomial. We shall see that, fortunately, it is not necessary to derive minimal polynomials in most cases of binary code design, since they are available in published lists.

We conclude our description of minimal polynomials with an important property of polynomials that have minimal polynomials as factors. Let $\beta_1, \beta_2, \ldots, \beta_L$ be elements in some extension field of GF(2), and let the minimal polynomials of these elements be $m_{\beta_1}(x), m_{\beta_2}(x), \ldots, m_{\beta_L}(x)$. Then the smallest degree monic polynomial with coefficients from GF(2) having $\beta_1, \beta_2, \ldots, \beta_L$ as roots, say $g(x)$, is given by

$$g(x) = \text{LCM}[m_{\beta_1}(x), m_{\beta_2}(x), \ldots, m_{\beta_L}(x)]$$

where LCM denotes the *least common multiple*.

We might well refer to $g(x)$ as the minimal polynomial of the set of elements $\beta_1, \beta_2, \ldots, \beta_L$. If the minimal polynomials of these elements are distinct (recall that different field elements can have the same minimal polynomial), then $g(x)$ is simply

$$g(x) = \prod_{i=1}^{L} m_{\beta_i}(x)$$

3. BINARY BLOCK CODES

BCH codes are block codes that form a subclass of a broad class called *cyclic codes*. These codes have a well-defined algebraic structure that has led to the development of efficient encoding and decoding schemes. BCH codes have proven useful in practical applications because, over certain ranges of code parameters, good performance is achieved, and the encoders and decoders have reasonable complexity. We begin with a brief description of binary block codes.

For a binary block code, the information bits to be transmitted are first grouped into k-bit blocks. An encoding rule is applied that associates r redundant check bits to each k bit information set. The resulting group of $n = k + r$ encoded bits forms an n-bit codeword that is transmitted on the channel. Since the r check bits represent overhead in the transmission, we say that the code rate $R = k/n$. The block length of the code is n and the notation (n, k) is used to represent a code with block length n containing k information bits and $r = n - k$ check bits per block. Clearly, an (n, k) code comprises the set of codewords representing all possible k-bit information sets.

There are several ways to represent the codewords in an (n, k) code. For example, codewords may be represented as n-bit vectors or as polynomials with degree up to $n - 1$. For binary codes, the vector components or the polynomial coefficients are the elements of GF(2), namely, 0 and 1. An important property of block codes is linearity. We say that a code is linear if sums of codewords are codewords. Codewords are added by forming bit-by-bit (bitwise) modulo 2 sums of the corresponding vector positions or polynomial coefficients.

An important attribute of a block code is its minimum distance d. The *minimum distance* of a binary block code is the smallest number of codeword bit positions in which an arbitrary pair of codewords differ. A code with minimum distance d provides the capability to correct all error patterns containing $t \leq (d - 1)/2$ errors for d odd, and $t \leq (d/2) - 1$ errors for d even.

Binary block codes may also be systematic or nonsystematic. A systematic code has the feature that the k information bits appear unaltered in each codeword. With nonsystematic codes, they do not. Systematic codes are generally preferred, but we consider both.

3.1. Cyclic Block Codes

A binary block code is said to be cyclic if the following two properties hold:

1. The code is linear.
2. Any cyclic ("end around") shift of a codeword is also a codeword.

The first property means that sums of codewords are codewords, and the second means that if $c = (c_0, c_1, c_2, \ldots, c_{n-1})$ is a codeword, then so are all cyclic shifts, that is, $(c_{n-1}, c_0, c_1, c_2, \ldots, c_{n-2})$, $(c_{n-2}, c_{n-1}, c_0, c_1 \ldots, c_{n-3})$, and so forth.

In describing the structure of cyclic codes, the polynomial representation of codewords is more convenient than the vector representation. We note that for the linearity property, two codeword polynomials are summed by adding in GF(2) coefficients of corresponding terms of each power of x, and the second property means that if $c(x)$ is a codeword polynomial, then

$$x^j c(x) \bmod x^n - 1$$

is also a codeword polynomial for any cyclic shift j. This is true since multiplication of the codeword polynomial by x^j, setting $x^n = 1$, is equivalent to a cyclic shift of the codeword. [We use $x^n - 1$ instead of $x^n + 1$ since this is the general form applicable with polynomials over any finite field.]

To see how cyclic codes are constructed, consider the codeword polynomial that has the smallest degree, $g(x)$. The degree of $g(x)$ is r and it is straightforward to see that all codeword polynomials can be represented as linear combinations of $g(x)$ and cyclic shifts of $g(x)$. Consequently, all codeword polynomials are divisible evenly by $g(x)$, and

we call $g(x)$ the *generator polynomial* of the code. Thus, a cyclic code with block length n can be represented as all the polynomials of the form

$$c(x) = a(x)g(x) \bmod x^n - 1$$

We next show that a cyclic code of block length n is formed from any polynomial $g(x)$ that divides $x^n - 1$; that is, the generator polynomial must be such that

$$x^n - 1 = g(x)h(x)$$

We can verify that a code generated in this manner is cyclic, as follows. We wish to prove that a cylic shift of a codeword

$$x^j c(x) \bmod x^n - 1 = x^j a(x)g(x) \bmod x^n - 1$$

is also a codeword. If it is, the polynomial $x^j c(x) \bmod x^n - 1$ must be divisible by $g(x)$, that is, we must have

$$[x^j a(x)g(x) \bmod x^n - 1] \bmod g(x) = 0$$

Now, if $g(x)$ is a factor of $x^n - 1$, then $[b(x) \bmod x^n - 1] \bmod g(x)$ is simply $b(x) \bmod g(x)$, so that we can write

$$[x^j a(x)g(x) \bmod x^n - 1] \bmod g(x) = x^j a(x)g(x) \bmod g(x) = 0$$

showing that the cyclically shifted codeword is divisible by $g(x)$ and is therefore itself a codeword.

Since r is the degree of the generator polynomial and the generator divides $x^n - 1$, it is a simple matter to show that the resulting cyclic code has 2^k codewords, where $k = n - r$. This follows from the fact that all polynomials $a(x)$ of degree less than k produce distinct codewords, since the products $g(x)a(x)$ must have degree less than n, which in turn means that each product modulo $x^n - 1$ will simply be the polynomial $a(x)g(x)$ itself. Since there are 2^k distinct polynomials of degree $k - 1$ or less, there must be 2^k distinct codewords in the code. Therefore, using the notation adopted previously, we say that the vectors of coefficients of the codeword polynomials generated by $g(x)$, where $g(x)$ divides $x^n - 1$, form an (n, k) cyclic block code. For convenience, the term *codeword* is used interchangeably with *codeword polynomial*.

Cyclic codes may be encoded using the property that $g(x)$ divides all codewords evenly. Let $i(x)$ be the degree $k - 1$ polynomial representing a set of k information bits; then the corresponding codeword polynomial for a nonsystematic code is

$$c(x) = i(x)g(x)$$

and the systematic code may be encoded using

$$c(x) = x^r i(x) + [x^r i(x)] \bmod g(x)$$

A linear feedforward shift register, a polynomial multiplication circuit, may be used to encode the nonsystematic code, and a linear feedback shift register, a polynomial division circuit, may be used to encode the systematic code.

Since each codeword in a cyclic code contains $g(x)$ as a factor, each code polynomial will have roots [from solution of $c(x) = 0$] that must include the roots of $g(x)$. It then follows that since the cyclic code is completely described by $g(x)$, we may define the code by specifying the roots of $g(x)$.

The foregoing discussion of cyclic codes constructed from generator polynomials illustrates the usefulness of polynomial algebra in representing block codes. The factorizations of $x^n - 1$ provide a number of generator polynomials for cyclic codes with block length n, the degree r of the generator determining the number of parity check bits, and $k = n - r$ the number of information bits in the code. The description of codewords as multiples of the generator polynomial provides a characterization of the codewords as the set of polynomials whose roots are the roots of the generator polynomial. The great value of this approach in describing cyclic codes has been to enable coding theorists to draw upon the extensive body of mathematical theory on the algebra of polynomials and their roots in finite fields.

3.2. Binary BCH Codes

The *Bose–Chaudhuri–Hocquenghem codes*, usually referred to as *BCH codes*, are an infinite class of cyclic block codes that have capabilities for *multiple-error detection and correction* [1,2]. For any positive integers m and $t < n/2$, there exists a binary BCH code with block length $n = 2^m - 1$, and minimum distance $d \geq 2t + 1$ having no more than mt parity check bits. Each such code can correct up to t random errors per codeword, and thus is a *t-error-correcting code*.

A BCH code being cyclic can be defined in terms of its generator polynomial $g(x)$. Let α be a primitive element of the extension field GF(2^m). The generator polynomial for a *t-error-correcting* BCH code is chosen so that $2t$ consecutive powers of α, such as $\alpha, \alpha^2, \alpha^3, \ldots, \alpha^{2t}$, are roots of the generator polynomial and consequently are also roots of each codeword.

This defines the subclass of *primitive BCH codes* because the roots are specified to be consecutive powers of a primitive element of GF(2^m). The block length of a BCH code is the order of the element used in defining the consecutive roots. Since α is a primitive element in GF(2^m), the block length of a primitive BCH code is $2^m - 1$. To generalize the definition, if $2^m - 1$ is factorable, $2t$ consecutive powers of some nonprimitive element β of GF(2^m) may instead be specified as roots of the codewords. The resulting code is a *nonprimitive BCH code* and will have a block length that divides $2^m - 1$.

In general, the definition of a BCH code allows the powers of the roots to range over any interval of consecutive values, say, $m_0, m_0 + 1, \ldots, m_0 + 2t - 1$. The parameter m_0 is usually chosen to be zero or one. For the present discussion, $m_0 = 1$.

Since the BCH codes are cyclic, codewords are assured of having the desired set of roots by choosing the generator polynomial so that it has $\alpha, \alpha^2, \ldots, \alpha^{2t}$ as roots. This is done by letting $g(x)$ be the least common multiple of the minimal polynomials of $\alpha, \alpha^2, \ldots, \alpha^{2t}$; that is, we write

$$g(x) = \text{LCM } [m_{\alpha^1}(x), m_{\alpha^2}(x), \ldots, m_{\alpha^{2t}}(x)] \qquad (2)$$

We note from property 6 in the earlier discussion of irreducible polynomials that if a binary irreducible polynomial has β as a root, where β is an element of an extension field of GF(2), then it also has β^2 as a root. Therefore, in Eq. (2), each even-power element α^{2i} and the corresponding element α^i are roots of the same minimal polynomial, $m_{\alpha^i}(x)$, and we can condense the sequence of minimal polynomials and write instead

$$g(x) = \text{LCM}[m_{\alpha^1}(x), m_{\alpha^3}(x), \ldots, m_{\alpha^{2t-1}}(x)]$$

Since we know from property 3 of minimal polynomials that the minimal polynomial of an element in GF(2^m) will have degree no greater than m, we know that $g(x)$ will have degree no greater than mt, and thus the number of parity bits $r = n - k$ will be $\leq mt$. For high-rate codes, $n - k$ is exactly equal to mt, and as t is increased $n - k$ can be smaller than mt. The single-error-correcting primitive BCH codes, $n = 2^m - 1$, $n - k = m$, are the Hamming codes. The generator polynomial for a Hamming code is the minimal polynomial of the primitive element, $m_\alpha(x)$.

The quantity $2t + 1$ used in specifying the generator polynomial of a BCH code is called the *design distance* of the code, but the true minimum distance will in some cases be greater, that is, $d \geq 2t + 1$. The true minimum distance for an arbitrary BCH code cannot be readily given, as this general problem is as yet unsolved. For a great many cases of practical interest, the true minimum distance is equal to the design distance, and the number of check bits is mt.

Tables of generator polynomials for many codes are available in the literature. A simple example is included here to show how a generator is obtained.

Consider a primitive three-error-correcting BCH code with block length $n = 31$ and $m_0 = 1$. The generator polynomial has α, α^3, and α^5 as roots, where α is a primitive element of GF(32). Therefore, $g(x)$ is obtained by forming the product of the minimal polynomials of α, α^3, and α^5 in GF(32). A table of polynomials set up for our purpose is given in Ref. 3. (This table was originally published by Peterson [4] in the first book devoted to the subject of error-correcting codes. It also appears in Peterson and Weldon [5], which is an expanded and updated edition of the original text. For convenience, we refer to the table as the *Peterson table*). From the table, we find

$$m_\alpha(x) = 45_8 = x^5 + x^2 + 1$$

$$m_{\alpha^3}(x) = 75_8 = x^5 + x^4 + x^3 + x^2 + 1$$

$$m_{\alpha^5}(x) = 67_8 = x^5 + x^4 + x^2 + x + 1$$

We now enumerate all the roots of each of these minimal polynomials to determine whether any one polynomial has more than one of the required three roots. This is done with the aid of property 6 of minimal polynomials. The roots of the three minimal polynomials are as follows:

$$\text{Roots of } m_\alpha(x): \quad \alpha, \alpha^2, \alpha^4, \alpha^8, \alpha^{16}$$

$$\text{Roots of } m_{\alpha^3}(x): \quad \alpha^3, \alpha^6, \alpha^{12}, \alpha^{24}, \alpha^{48} = \alpha^{17}$$

$$\text{Roots of } m_{\alpha^5}(x): \quad \alpha^5, \alpha^{10}, \alpha^{20}, \alpha^{40} = \alpha^9, \alpha^{18}$$

From this enumeration, we see that α, α^3, and α^5 are roots of three distinct polynomials, and therefore, the required generator polynomial is simply the product of the three minimal polynomials just found:

$$\begin{aligned} g(x) &= (x^5 + x^2 + 1)(x^5 + x^4 + x^3 + x^2 + 1) \\ &\quad \times (x^5 + x^4 + x^2 + x + 1) \\ &= x^{15} + x^{11} + x^{10} + x^9 + x^8 + x^7 + x^5 + x^3 + x^2 + x + 1 \end{aligned}$$

From the enumeration of all the roots of $m_{\alpha^1}(x)$, $m_{\alpha^3}(x)$, and $m_{\alpha^5}(x)$, it can be verified that each of the three minimal polynomials must be of degree 5, as we found directly by use of the table of polynomials. This distance-7 code, therefore, has 15 check bits and 16 information bits in each codeword.

A table of generator polynomials for primitive BCH codes of block length up to 255 has been published by Stenbit [6], and a table of generator polynomials for BCH codes with lengths up to 1023 is given in a text by Lin and Costello [7].

A number of codes widely used for error detection with long data packets and files in communication network and computer applications are called the *cyclic redundancy check* (CRC) codes [8]. Some of the standardized CRC codes are distance-4 binary BCH codes in which the generator polynomial is formed by multiplying the generator polynomial of a Hamming code by $x + 1$. In these cases, the generator polynomial for the CRC code is of the form $(x + 1)m_\alpha(x)$ and has consecutive roots α^0, α, and α^2. In some applications, CRC codes are modified so that the all-zeros information set is not associated with the all-zeros check set in order to detect certain types of hardware failures.

A generalization of the BCH codes mentioned previously permits specifying a consecutive sequence of roots that can be powers of any element of GF(2^m). That is, with $m_0 = 1$, the sequence of roots can be selected as

$$\beta^1, \beta^2, \beta^3, \ldots, \beta^{2t}$$

where β need not be a primitive element. If $2^m - 1$ is not prime, some of the elements of GF(2^m) will be nonprimitive, and the order of each such element divides $2^m - 1$. For example, GF(2^4) contains elements of order 3, 5, and 15; GF(2^6) contains elements of order 3, 7, 9, 21, and 63; and so on. BCH codes generated from nonprimitive field elements are called *nonprimitive BCH codes*. Each such code, defined as having roots that are $d - 1$ consecutive powers of an element β, will have design distance d, and block length equal to the order of β.

An important example of a nonprimitive code is the (23,12) Golay code. This three-error-correcting code may be constructed as a nonprimitive BCH code with roots in GF(2^{11}). Since $2^{11} - 1 = 23 \times 89$, GF($2^{11}$) has elements of order 23 and 89 as well as 2047. By selecting $\beta = \alpha^{89}$, a code of length 23 can be constructed. Consider the design of a single-error-correcting code, which we specify as having roots β and β^2. Using property 6 of minimal polynomials once again, the roots of the minimal polynomial of β are enumerated as

$$\beta, \beta^2, \beta^4, \beta^8, \beta^{16}, \beta^9, \beta^{18}, \beta^{13}, \beta^3, \beta^6, \beta^{12}$$

where $\beta^{23} = \alpha^{2047} = 1$ is used to reduce powers of β greater than 22. Notice that the sequence is found to include four consecutive powers of β, namely, β, β^2, β^3, and β^4. Therefore, we see that by using the minimal polynomial $m_\beta(x)$ as a BCH code generator polynomial, with the intention of designing a distance-3 code, we have "discovered" two additional consecutive roots and thus have actually constructed a code with design distance 5. The code has 11 check bits, and its generator polynomial can be found in [3], listed as 89 5343B, which yields

$$g(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$$

Although we do not show it here, the true minimum distance is actually 7 rather than 5, and the code is the three-error-correcting (23,12) code originally described by Golay [9].

In applications of binary block codes, it is often necessary to provide a code with a block length that does not correspond exactly to one of the strict-sense BCH codes. This can usually be accomplished by choosing a code with block length greater than the required length and *shortening* the code by an appropriate amount. The shortening is most readily done by setting a number of the information bits equal to zero. The number of codewords that can be generated is reduced accordingly, and since the reduced set of codewords is a subset of the codewords in the unshortened code, the minimum distance of the shortened code must be at least as great as that of the unshortened code. Depending on the amount of the shortening and which particular bits are omitted, the minimum distance may be unchanged or it may increase.

In general, a shortened BCH code may or may not be cyclic, depending on which particular information bits are omitted. There is no general theory available to give guidance about which bits are best omitted for a required amount of shortening. Typically, the shortening is done in the most convenient manner, which is to set a string of consecutive information bits equal to zero, usually the high-order bits in the codeword.

Another commonly used code modification is the *extension* of a code of odd minimum distance by addition of a single overall parity check. Since this modification causes the weight of any odd-weight codeword in the original code to be increased by 1, the minimum distance of the original code is also increased by 1. It should be noted that this modification is not the same as inserting the factor $x + 1$ into the generator polynomial, although even-valued minimum distance $2t + 2$ results in both cases. In the earlier case, while the parity set was increased by one bit, the information set was simultaneously decreased by one bit, so that the block length was unchanged. With the extension being described here, the information set remains unchanged and the block length is increased by one bit. Furthermore, the code obtained by this one-bit extension is not a cyclic code.

A frequently used extended code is the (24,12) distance-8 code, called the *extended Golay code*, which is obtained by appending an overall parity check bit to the (23,12) distance-7 Golay code. The extended code is attractive partly because the rate k/n is exactly equal to 0.5.

As we have pointed out, a BCH code or any cyclic code can be encoded by using the generator polynomial $g(x)$ in the manner indicated by the basic definition of a cyclic code:

$$c(x) = i(x)g(x)$$

Thus, we associate a polynomial $i(x)$ of degree $k - 1$ with the set of k information bits to be transmitted and multiply by the degree-r polynomial $g(x)$ forming the degree-$(n - 1)$ code polynomial $c(x)$. However, this results in a nonsystematic code structure, and it is preferred instead to form codewords using

$$c(x) = [x^r i(x) \bmod g(x)] + x^r i(x)$$

It is seen that this encoding operation places the k information bits in the k highest-order terms of the code polynomial, while the parity check bits, represented by $x^r i(x) \bmod g(x)$, are confined to the r lowest-order terms.

The encoding of any binary cyclic code can be done in a straightforward manner using a linear feedback shift register. An encoding circuit using a shift register with $r = n - k$ stages is shown in Fig. 1. Each box in the circuit is a binary storage device. The additions indicated are done modulo 2, and the tap connections are specified by the coefficients of the generator polynomial. The operation of the encoder is as follows:

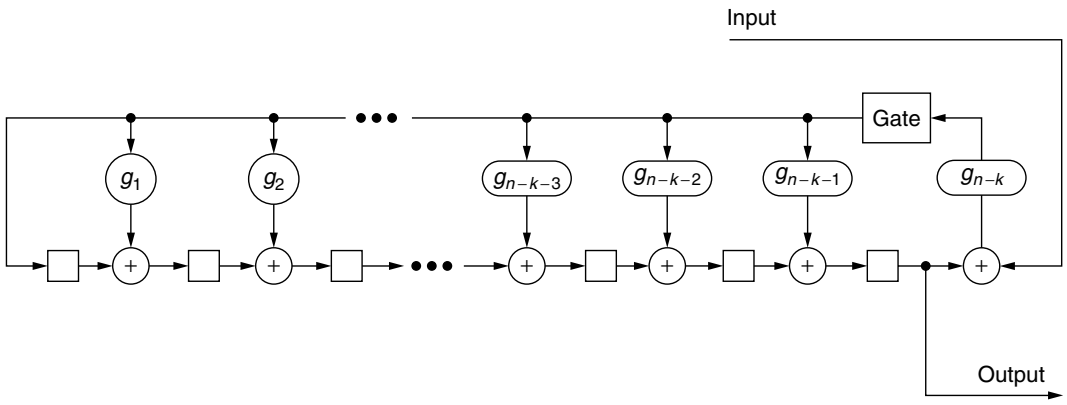


Figure 1. Encoder for a binary BCH code generated by $g(x)$.

1. Shift the k information bits into the encoder and simultaneously into the channel. As soon as the k information bits have entered the shift register, the $r = n - k$ bits in the register are the check bits.
2. Disable the feedback circuit.
3. Shift the contents of the register into the channel.

As an example, consider an encoder for the length-7 Hamming code generated by $g(x) = x^3 + x + 1$. Using a shift register of the form shown in Fig. 1, the feedback tap connections are $g_1 = g_3 = 1$, and $g_2 = 0$. The feedback circuit accomplishes division by $x^3 + x + 1$ in that it sets $x^3 = x + 1$ at each shift of the circuit in step 1.

The encoding circuit that uses an r-stage feedback shift register whose connections are given by the generator polynomial is most convenient for high-rate codes where $k > r$. For low-rate codes, a more convenient encoder realization employs a k-stage feedback shift register whose tap connections are given by $h(x) = (x^n - 1)/g(x)$. For the Hamming code considered, we have $h(x) = x^4 + x^2 + x + 1$, and the circuit shown in Fig. 2 may be used. The operation of the encoder is as follows:

1. With the feedback circuit disabled, shift the $k = 4$ information bits into the k-stage register and simultaneously into the channel.
2. When the k information bits have entered the encoding register, cycle the register $r = 3$ times with the input disabled. The $r = 3$ bits obtained at the output are the encoded parity bits. The parity bits are shifted into the channel.

3.3. Decoding Binary BCH Codes

The problem of decoding a binary BCH code can be described succinctly as follows. For each received word, first determine whether any errors have occurred during transmission. If errors have occurred, determine the most likely locations of the errors in the received word, and make the appropriate corrections. A brute-force approach would be to change one bit at a time, then 2 bits, in all combinations, and so on, until a valid codeword is found. This, of course, is impractical for any but extremely simple codes. Therefore, much work has been devoted to finding efficient algorithms for implementing error correction.

Here we describe the more important decoding techniques developed to date. First, we describe an approach based on solving a set of nonlinear algebraic equations over a finite field, a discussion that leads to the Berlekamp iterative algorithm. Error-trapping decoders are mentioned, and the important example of the Kasami decoder for the Golay code is referenced.

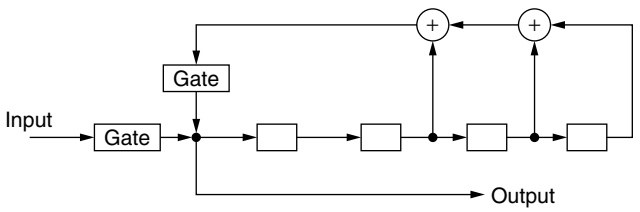


Figure 2. Encoder for the (7,4) Hamming code generated by $h(x) = x^4 + x^2 + x + 1$.

Each of these decoders operates on inputs that consist of hard-bit decisions. In systems where quality measures can be obtained for the received bits, soft-decision decoding techniques can be utilized to increase the power of a code beyond that achievable with algebraic hard-decision decoding. Several such techniques exist, ranging from the simplest forms of *erasure filling* to a set of algorithms termed *channel measurement decoding*.

3.3.1. The Syndrome Equations. Decoding a BCH code begins by calculating a quantity called the *syndrome*. We represent a codeword as the polynomial $c(x)$, the received word as the polynomial $r(x)$, and the corresponding error pattern as $e(x)$, and we can write

$$r(x) = c(x) + e(x)$$

To compute the syndrome values S_k for the received word, we simply substitute the roots of the code generator polynomial into $r(x)$:

$$S_k = r(\alpha^k) = r_0(\alpha^k)^0 + r_1(\alpha^k)^1 + r_2(\alpha^k)^2 + \cdots + r_{n-1}(\alpha^k)^{n-1}$$

which can be evaluated using

$$r(\alpha^k) = \{\cdots[(r_{n-1}\alpha^k + r_{n-2})\alpha^k + r_{n-3}]\alpha^k + \cdots\}\alpha^k + r_0$$

Note that

$$\begin{aligned} S_k &= c(\alpha^k) + e(\alpha^k) \\ &= e(\alpha^k), \quad k = 1, 3, \ldots, 2t-1 \end{aligned} \tag{3}$$

since $c(\alpha^k) = 0$, $k = 1, 3, \ldots, 2t - 1$. That is, each element S_k of the syndrome is simply the error pattern polynomial $e(x)$ evaluated at $x = \alpha^k$ and thus is some element in the extension field GF(2^m). Let us now assume that there are t errors in the received word, so that $e(x)$ has t nonzero coefficients. (To avoid unduly complicated notation, we are letting the actual number of errors be equal to the maximum number correctable by the code, i.e., the value t such that the minimum distance of the code is $d = 2t + 1$. For error patterns having fewer than t errors, one can think of the appropriate subset of the assumed t errors having values equal to 0 rather than 1.) If the ith error $(1 \le i \le t)$ occurs in received symbol $r_j (0 \le j \le n - 1)$, then we define its *error locator* to be $X_i = \alpha^j$, which is an element of GF(2^m). We thus refer to GF(2^m) as the *locator field*. Since we are considering a binary code, all *error values* are 0 or 1, and we can write for any k

$$e(\alpha^k) = \sum_{i=1}^{t} X_i^k \tag{4}$$

where t is the number of errors in the received word.

To make these points clearer, let us say, for example, that there are three errors in the received word, in the first, second, and last bit positions. Then the error polynomial evaluated at each root is simply

$$\begin{aligned} e(\alpha^k) &= c_0(\alpha^k)^0 + e_1(\alpha^k)^1 + e_{n-1}(\alpha^k)^{n-1} \\ &= (\alpha^0)^k + (\alpha^1)^k + (\alpha^{n-1})^k \\ &= X_1^k + X_2^k + X_3^k \end{aligned}$$

where X_1, X_2, and X_3, are the three error locators and the $e_j = \alpha^0 = 1$ are the error values.

From Eqs. (3) and (4) we see that

$$S_k = \sum_{i=1}^{t} X_i^k, \quad k = 1, 3, \ldots, 2t-1 \tag{5}$$

The decoding problem then is simply to find the error locators X_i from the syndrome values $S_1, \ldots, S_{2t-1}$. Note, however, that Eq. (5) represents t nonlinear coupled algebraic equations over the finite field GF(2^m). Direct solution of such equations is generally avoided, and an indirect approach is used instead. To this end, we introduce the polynomial

$$\sigma(x) = \prod_{i=1}^{t}(x + X_i) = x^t + \sigma_1 x^{t-1} + \cdots + \sigma_t \tag{6}$$

having the error locators as roots, and which we therefore call the *error locator polynomial*. The coefficients σ_i are seen to be given by the *elementary symmetric functions* of the error locators [5]:

$$\begin{aligned} \sigma_1 &= \sum_i X_i \\ \sigma_2 &= \sum_{i<j} X_i X_j \\ \sigma_3 &= \sum_{i<j<k} X_i X_j X_k \\ &\vdots \\ \sigma_t &= X_1 X_2 X_3 \cdots X_t \end{aligned}$$

[*Note*: Some authors define the error locator polynomial $\sigma(x)$ as a polynomial with factors of the form $(1 + X_i x)$, so that the roots of $\sigma(x)$ are the reciprocals of the error locators. We find the notation used here to be more convenient for purposes of exposition. However, the reciprocal-root formulation of $\sigma(x)$ will be used in later discussions.]

Several approaches can be taken to decoding a BCH code, each having relative advantages and disadvantages that depend largely on the number of errors the code is designed to correct. Several of the important techniques that are used can be broadly summarized for binary codes as follows:

Step 1. Calculate the syndrome values $S_k = r(\alpha^k), k = 1, 3, \ldots, 2t-1$.

Step 2. Determine the elementary symmetric functions, that is, the coefficients of the error locator polynomial $\sigma(x)$, from the syndrome values.

Step 3. Solve for the roots of $\sigma(x)$, which are the error locators.

Step 4. Correct the errors in the positions indicated by the error locators.

In general, the most difficult part of this procedure is step 2, determination of the coefficients of $\sigma(x)$ from the syndrome values, and it is in this step that the most prominent algorithms differ.

3.3.2. Peterson's Direct Solution Method. We saw in the previous discussion that the syndrome values $S_1, S_3, \ldots, S_{2t-1}$ are the constants in a set of simultaneous nonlinear equations in which the unknowns are the error locators $X_1, X_2, \ldots, X_t$. We now describe a method, due to Peterson [10], for direct solution of these nonlinear equations. In order to describe this method, we write the full set of syndrome values $S_1, S_2, \ldots, S_{2t}$ as

$$\begin{aligned} S_k &= r(\alpha^k) \\ &= c(\alpha^k)) + e(\alpha^k) \\ &= \sum_{i=1}^{t} X_i^k, \quad k = 1, 2, \ldots, 2t \end{aligned}$$

which gives the equations

$$\begin{aligned} X_1 + X_2 + \cdots + X_t &= S_1 \\ X_1^2 + X_2^2 + \cdots + X_t^2 &= S_2 \\ &\vdots \qquad\qquad \vdots \quad \vdots \\ X_1^{2t} + X_2^{2t} + \cdots + X_t^{2t} &= S_{2t} \end{aligned} \tag{7}$$

We call these the *syndrome equations*. The syndrome values $\{S_k\}$ are computed from the received word, and Eq. (7) is to be used to obtain the error locators $\{X_i\}$.

Rather than solving this set of nonlinear equations directly, we convert the equations into linear equations that can be solved in conjunction with the error locator polynomial $\sigma(x)$. This is accomplished by first noting that $\sigma(x)$ evaluated at each error locator value equals zero:

$$\sigma(X_i) = X_i^t + \sigma_1 X_i^{t-1} + \cdots + \sigma_t = 0, \quad i = 1, 2, \ldots, t \tag{8}$$

Clearly, we can multiply Eq. (8) through by any power of X, and the equality is preserved. In particular, let us multiply by, X_i^j, so that we have

$$X_i^{t+j} + \sigma_1 X_i^{t+j-1} + \cdots + \sigma_t X_i^j = 0, \quad i = 1, 2, \ldots, t \tag{9}$$

Now, letting j remain general, we sum Eq. (9) over $i = 1, 2, \ldots, t$, and using the syndrome equations, Eq. (7), we can write

$$S_{t+j} + \sigma_1 S_{t+j-1} + \cdots + \sigma_t S_j = 0 \tag{10}$$

The equations defined by Eq. (10), with t general, are called *Newton's identities*, which for a binary code can be shown to be equivalent to

$$\begin{aligned} S_1 + \sigma_1 &= 0 \\ S_3 + S_2\sigma_1 + S_1\sigma_2 + \sigma_3 &= 0 \\ S_5 + S_4\sigma_1 + S_3\sigma_2 + S_2\sigma_3 + S_1\sigma_4 + \sigma_5 &= 0 \\ &\vdots \end{aligned} \tag{11}$$

In principle, to decode a code of any given minimum distance, we need only truncate Eq. (11) in an appropriate manner and solve a set of linear equations for the $\{\sigma_i\}$ in terms of the given syndrome values.

For example, in decoding a single-error-correcting code, there is only one syndrome value, S_1, and the first line of Eq. (11) gives

$$S_1 + \sigma_1 = 0$$

so that we have

$$\sigma_1 = S_1$$

For $t = 1$, the error locator polynomial, Eq. (6), is simply $x + \sigma_1$, having the trivial root $x = \sigma_1$, which we have just found to be equal to S_1. Thus for a single-error-correcting BCH code, we have the very simple result that the error locator is equal to the syndrome S_1.

For a two-error-correcting code, two syndrome values are computed, S_1 and S_3, and the first two lines of Eq. (11) (with $\sigma_3 = 0$) can be written in matrix form as

$$\begin{bmatrix} 1 & 0 \\ S_2 & S_1 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_3 \end{bmatrix}$$

These simultaneous linear equations are solved using the methods of ordinary algebra except that multiplications, divisions, and additions are done using the rules for GF(2^m). We note here that while the S_k values for k even need not be calculated for a binary BCH code, these even-indexed syndrome values appear in the given formulation of the decoding problem. They are readily obtained since it is easy to show that for binary codes, $S_{2k} = S_k^2$ for any k. That is, for elements A and B_i in a field of characteristic 2, if

$$A = \sum_i B_i$$

then the square of A is simply

$$A^2 = \sum_i \sum_j B_i B_j = \sum_i B_i^2$$

so that we have

$$S_1^2 = \sum_{i=1}^{t} X_1^2 = S_2 \tag{12}$$

Similarly, $S_4 = S_2^2 = S_1^4, S_6 = S_3^2$, and so forth. Thus, in solving the simultaneous equations, the solutions can be expressed in terms of only the odd-indexed syndrome values. For the case of two-error correction, we have

$$\sigma_1 = S_1 \text{ and } \sigma_2 = \frac{S_3 + S_1^3}{S_1} \tag{13}$$

Using standard techniques to solve sets of simultaneous linear equations, direct solutions for the coefficients of the error locator polynomial can be found for any error-correction limit t. The results of such solutions for $t = 3$ and 4 are as follows:

Three-Error Correction	**Four-Error Correction**
$\sigma_1 = S_1$	$\sigma_1 = S_1$
$\sigma_2 = \dfrac{S_1^2 S_3 + S_5}{S_1^3 + S_3}$	$\sigma_2 = \dfrac{S_1(S_7 + S_1^7) + S_3(S_1^5 + S_5)}{S_3(S_1^3 + S_3) + S_1(S_1^5 + S_5)}$
$\sigma_3 = (S_1^3 + S_3) + S_1\sigma_2$	$\sigma_3 = (S_1^3 + S_3) + S_1\sigma_2$
	$\sigma_4 = \dfrac{(S_5 + S_1^2 S_3) + (S_1^3 + S_3)\sigma_2}{S_1}$

In general, however, with use of a t-error-correcting code, any error pattern with fewer than t errors is also correctable, and we do not know at the outset of decoding how many errors there actually are. To use these formulas knowing that the actual number of errors in a received word may be less than t, we start by determining whether the first t lines in Eq. (11) can be solved for $\sigma_1, \sigma_2, \ldots, \sigma_t$. This is done by using the determinant test

$$\det \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ S_2 & S_1 & 1 & 0 & 0 & \cdots & 0 \\ S_4 & S_3 & S_2 & S_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ S_{2t-4} & S_{2t-5} & S_{2t-6} & S_{2t-7} & S_{2t-8} & \cdots & S_{t-3} \\ S_{2t-2} & S_{2t-3} & S_{2t-4} & S_{2t-5} & S_{2t-6} & \cdots & S_{t-1} \end{bmatrix} \overset{?}{\neq} 0$$

It can be shown [10] that if there are t or $t-1$ errors in the received word, the determinant will be nonzero. Given this outcome, we proceed with the formulas for t-error correction. If there are actually t errors, the solutions found for $\sigma_1, \sigma_2, \ldots, \sigma_t$ define a degree-t error locator polynomial. If there are only $t-1$ errors, $\sigma_t = 0$ and thus, $\sigma(x)$ has degree $t-1$.

If the determinant shown above is found to be zero, two rows and columns of the matrix are removed, and the determinant of the resulting $(t-2) \times (t-2)$ matrix is tested in the same manner. This procedure is repeated until a nonzero determinant is found and the error locator polynomial coefficients are determined.

The final steps in decoding a binary BCH code are to find the roots of the error locator polynomial $\sigma(x)$, which are the error locators, and to correct the errors. A procedure called the *Chien search* [11] accomplishes these two processes without explicitly solving $\sigma(x)$. This can be done with the circuit shown in Fig. 3. The circuit steps sequentially through all possible error locator values and corrects the corresponding bits as the locators are found. To see how the circuit operates, consider the error locator

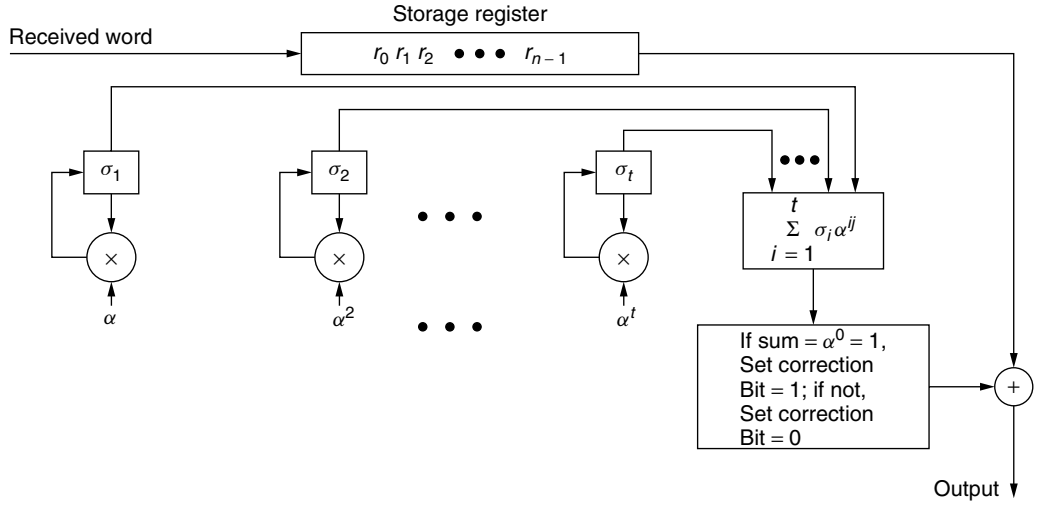


Figure 3. Circuit for implementing Chien search.

polynomial as given by Eq. (6) and divide through by x^t, which gives

$$\frac{\sigma(x)}{x^t} = 1 + \sigma_1 x^{-1} + \sigma_2 x^{-2} + \cdots + \sigma_t x^{-t}$$

The values of x that satisfy $\sigma(x) = 0$ consequently satisfy the equation

$$\sigma_1 x^{-1} + \sigma_2 x^{-2} + \cdots + \sigma_t x^{-t} = 1$$

Assuming the convention of transmitting codewords high-order bits first, it is convenient to apply the root test to locator α^{n-1} first. Note that evaluation of a term x^{-i} at α^{n-1} yields α^{-in+i}, which equals α^i if we are using a full-length BCH code, since we then have $\alpha^n = 1$ and thus $\alpha^{-in} = 1$. Therefore, we see that testing α^{n-1} as a possible root of $\sigma(x)$ is the same as testing for

$$\sigma_1 \alpha + \sigma_2 \alpha^2 + \cdots + \sigma_t \alpha^t \stackrel{?}{=} 1$$

and, in general, testing for α^{n-j} as an error locator is equivalent to finding whether or not α^j satisfies

$$\sum_{i=1}^{t} \sigma_i a^{ij} = a^0 = 1, \quad j = 0, 1, 2, \ldots, n-1$$

3.3.3. The Berlekamp Algorithm. For correction of more than about six errors in a binary BCH codeword, Peterson's direct method of solving for the coefficients of $\sigma(x)$ from the syndrome values becomes cumbersome and inefficient, since the number of finite-field multiplications required increases approximately with the square of the number of errors to be corrected. Instead, it is preferable to use an iterative algorithm developed by Berlekamp [12] for solution of Newton's identities. In contrast with the direct solution method, the *Berlekamp algorithm* has a computational complexity that grows only linearly with the number of errors to be corrected. Another version of this algorithm was given by Massey [13]. The Massey formulation is presented in the succeeding article on nonbinary BCH codes.

In the use of the Berlekamp algorithm, the sequence of calculated syndrome values, $S_1, S_2, \ldots, S_{2t}$, is represented by the polynomial

$$S(z) = S_1 z + S_2 z^2 + \cdots + S_{2t} z^{2t}$$

As a convenience in the algorithm, the error locator polynomial is replaced with an equivalent polynomial $C(z)$ whose *reciprocal roots* are the error locators $X_i, i = 1, 2, \ldots, t$; that is, $C(z)$ is defined by

$$C(z) = \prod_{i=1}^{t} (1 + X_i z)$$

so that $C(z)$ has roots at $z = Z_i$, where $Z_i = 1/X_i$, $i = 1, 2, \ldots, t$. We call the polynomial $C(z)$ the *reciprocal error locator polynomial* to distinguish it from the error locator polynomial $\sigma(x)$. Now writing $C(z)$ in its expanded form, we have

$$C(z) = 1 + \sigma_1 z + \sigma_2 z^2 + \cdots + \sigma_t z^t$$

where the coefficients $\{\sigma_i\}$ are again seen to be the elementary symmetric functions of the error locators $X_1, X_2, \ldots, X_t$.

The Berlekamp algorithm is an efficient iterative procedure for finding the minimum-degree reciprocal error locator polynomial $C(z)$ whose coefficients, taken together with the syndrome values, satisfy all t equations in Newton's identities, Eq. (11). The algorithm begins by constructing the polynomial $C^{(1)}(z)$ of least degree satisfying the first line in Eq. (11), and then setting $C^{(2)}(z) = C^{(1)}(z)$ and determining whether $C^{(2)}(z)$ satisfies the second line in Eq. (11). It is easy to see that these first steps consist in simply letting $C^{(1)}(z) = 1 + S_1 z$, so that $\sigma_1 = S_1$, and then testing (second line) the relationship $S_3 = S_2 \sigma_1$, where $S_2 = S_2^1$ from Eq. (12). This is equivalent to testing $S_3 = S_1^3$, the relationship that must hold, by Eq. (7), if there is only a single error in the received

word. If the test of the second line succeeds, then we set $C^{(3)}(z) = C^{(2)}(z)$ and test the third line of Eq. (11). If the test of the second line of Eq. (11) fails, $C^{(2)}(z)$ is modified by adding a correction term that changes $C^{(2)}(z)$ to a minimum-degree polynomial satisfying the first two lines of Eq. (10). With the corrected form of $C^{(2)}(z)$, we let $C^{(3)}(z) = C^{(2)}(z)$, and then test the third line of Eq. (11), and so forth. The iteration continues until a reciprocal error locator polynomial $C^{(l)}(z), l \leq t$, is found that satisfies all t lines in Eq. (11).

The efficiency of the Berlekamp algorithm is due largely to its provision for constructing the correction term at the ith iteration, if needed, so that the $i-1$ previous lines in Eq. (11) do not have to be retested. It can be shown that if the number of errors in the received word is t or less, the Berlekamp algorithm will end with the correct reciprocal error locator polynomial. We shall simply provide a brief description of the algorithm here. A detailed discussion of the algorithm and a rigorous proof of its error-correction properties can be found in the literature [5,12]. The algorithm as described below is actually a simplification for use with binary BCH codes. There is a more general version of the algorithm applicable to nonbinary codes as well.

The steps in the Berlekamp algorithm are described below. The initialized polynomial $C^{(0)}(z)$ fixes 1 as the leading term in $C(z)$, while $T^{(0)}(z)$ is the initialized correction polynomial. The quantity $\Delta^{(2k)}$ is the discrepancy found when an interim version of $C(z)$ constructed at one line in Eq. (11) fails to satisfy the next line. Superscripts are used to index the steps in the iteration.

The Berlekamp Algorithm for Decoding Binary BCH Codes

1. Initialize: $k = 0, C^{(0)}(z) = 1, T^{(0)}(z) = 1$.
2. If S_{2k+1} is not given, stop. Otherwise, define $\Delta^{(2k)}$ as the coefficient of z^{2k+1} in the product $[1 + S(z)]C^{(2k)}(z)$. Let

$$C^{(2k+2)}(z) = C^{(2k)}(z) + \Delta^{(2k)} z T^{(2k)}(z)$$

where

$$T^{(2k+2)} = \begin{cases} z^2 T^{(2k)}(z) & \text{if } \Delta^{(2k)} = 0, \text{ or if } \deg C^{(2k)}(z) > k \\ \dfrac{z C^{(2k)}(z)}{\Delta^{(2k)}} & \text{if } \Delta^{(2k)} \neq 0 \text{ and } \deg C^{(2k)}(z) \leq k \end{cases}$$

3. Set $k = k + 1$ and return to step 2.

Note that the multiplications and additions indicated are all in the locator field GF(2^m).

3.3.4. Other Decoding Algorithms. At times, other decoding algorithms for binary BCH codes are useful. For example, there are decoders that use "error trapping" to find and correct channel errors. These decoders evaluate the syndrome by calculating $s(x)$ where

$$s(x) = r(x) \bmod g(x)$$

The polynomial $s(x)$ has degree up to $r-1$ and is zero when no channel errors occur. Note that if t or fewer channel errors occur, and all the errors are confined to the check bit position, $s(x)$ contains at most t terms, the error pattern unmodified. It can be shown that, if there is at least one error in an information bit position, the number of terms in $s(x)$ is greater than t. Therefore, if t or fewer terms are found in $s(x)$, the channel error pattern is determined and can be corrected.

This property may be used to decode cyclic codes since if $s(x)$ has more than t terms, the received word may be shifted cyclically by one bit, and a second syndrome $s'(x)$ computed. If the error pattern has been shifted into the check bit positions, $s'(x)$ now contains t or fewer terms, and the error pattern has been successfully trapped in the check bit positions. In total, $n-1$ cyclic shifts of the received word may be tried in this way.

This decoding procedure will succeed when the channel error pattern spans at most r bit positions in the received word. This is not always the case, but for some codes, it is possible to specify additional tests that may be used to detect and correct the other correctable error patterns. The best example of this type of error trapping decoder is the Kasami decoder for the Golay code [14]. The Kasami decoder can be implemented with a simple linear feedback shift register and logic circuits.

3.3.5. Soft-Decision Decoding Techniques. Up to this point, we have discussed the decoding problem as one of finding the number and locations of errors in a received word. It has been assumed that at the receiving end of the communication circuit, a definite binary decision is made on each received digit after demodulation and prior to decoding, that is, a hard binary decision. However, it is sometimes possible to provide for quality or confidence estimates for demodulated data. In the simplest example of such schemes, we might test the demodulator output against a preselected magnitude threshold and erase each digit that falls below the threshold. The decoder is then presented with a sequence consisting of definite zeros and ones as well as erasures, and, given that sequence, the decoder has the task of deciding which of the valid codewords is most likely to have been transmitted. We call this decoding task one of *errors-and-erasures decoding*.

It has been noted that a block code having minimum distance d is capable of correcting any pattern of t or fewer errors, where $d = 2t+1$ or $2t+2$, for d odd or even, respectively. We now state that a distance-d code is capable of correcting any pattern of l errors and s erasures such that $2l + s < d$, and we show a very simple procedure that demonstrates that this is true. Assuming that we have at our disposal a decoder for correcting up to t errors, where $2t+1 = d$, we decode for s erasures and an unknown number of errors as follows:

1. Set all s erased bits in the received word equal to 0, and perform error correction of up to t errors. Note the number of errors corrected if decoding can be completed.
2. Next, set all s erased bits equal to 1, and decode the received word again, noting the number of errors corrected if decoding can be completed.

3. If only one decoding succeeds, accept that output. If both decoding attempts succeed but produce different codewords, accept the decoding result that required correction of the smaller number of errors.

The errors-and-erasures decoding procedure just described is an example of a general class of algorithms that are usually referred to as *soft-decision decoding* techniques. The simplest of such techniques is Wagner coding. In this scheme, encoding is done by appending a single overall parity check to a block of k information bits. The decoding procedure can be described as follows. On reception of each received digit r_i, the a posteriori probabilities $p(0|r_i)$ and $p(1|r_i)$ are calculated and saved, and a hard-bit decision is also made on each of the $k+1$ digits. Overall parity is checked, and if it is satisfied, the k information bits are accepted as first decoded. If parity fails, the received digit having the smallest difference between its two a posteriori probabilities is inverted before the k information bits are accepted. It is seen that this technique is in fact the simplest application of the errors-and-erasures decoding procedure described in the previous section, where here only a single erasure may be filled but no errors corrected, since the minimum distance of the single-parity-check code is only $d = 2$.

A generalization of Wagner coding applicable to any multiple-error-correcting (n, k) code is a scheme called *forced-erasure decoding*. Here we assume that the demodulator, in addition to making a hard binary decision on each received digit, also measures relative reliability; we denote the set of reliability measures by $p_1, p_2, \ldots, p_n$. For many communication channels, the probability of correct bit demodulation is monotonically related to the magnitude of the detected signal, and, in such cases, the detected signal strength can be taken as a measure of reliability for each bit.

Several decoding strategies come under the heading of forced-erasure decoding. They share the feature that decoder performance is improved by use of multiple hard-decision decoding attempts where the order of the decode attempts is based on the bit reliability information. The schemes that permit the largest number of decoding trials provide the best performance but they are complex and cumbersome. Finding efficient soft-decision decoding techniques for BCH codes is still an open problem.

BIOGRAPHIES

Arnold M. Michelson received his BSEE degree from the Johns Hopkins University, his MSEE from the University of Rochester, New York, and he did further graduate work at the Polytechnic Institute of Brooklyn, New York. In 1968, Mr. Michelson joined Sylvania Electric Products, which later became GTE Government Systems. At Sylvania and GTE he worked on the development and implementation of advanced communication techniques, including error-control coding for military applications. That work focused primarily on long-wave communications. Since 2000, he has been with the Raytheon Company where he is involved in the development of high-performance coding techniques for military and commercial satellite applications. In 1997, Mr. Michelson received GTE's Leslie H. Warner Technical Achievement Award, and, in 2002, Raytheon's Excellence in Technology Award, Distinguished Level.

Allen H. Levesque received his BSEE degree from Worcester Polytechnic Institute in 1959 and his MSEE and PhDEE degrees from Yale University in 1960 and 1965, respectively. Following completion of his graduate studies, he joined the GTE Corporation, where, over a 36-year career, he worked on and led a variety of digital communications research and development projects, with application to both defense and commercial systems. Much of his early work concerned applications of error-control coding techniques in radio networks. For the past decade, his work has concentrated on mobile and wireless communications networks. In early 1999, he retired from GTE Laboratories to begin and independent consulting practice and to take a part-time teaching and research position at WPI. He currently teaches graduate courses in modulation and coding and is a member of WPI's Center for Wireless Information Network Studies. His areas of research interest include communication theory and techniques, communication networks, wireless communications, spread-spectrum, secure communications, and digital signal processing. He has published numerous journal and conference papers, and coauthored two books, as well as chapters in several communications handbooks. He is a life fellow of the IEEE and a Registered Professional Engineer in the Commonwealth of Massachusetts.

BIBLIOGRAPHY

1. A. Hocquenghem, Codes correcteurs d'erreurs, *Chiffres* **2**: 147–156 (1959).
2. R. C. Bose and D. K. Ray-Chaudhuri, On a class of error-correcting binary group codes, *Inform. Control* **3**: 68–79 (1960).
3. A. M. Michelson and A. H. Levesque, *Error-Control Techniques for Digital Communications*, Wiley, New York, 1985.
4. W. W. Peterson, *Error Correcting Codes*, MIT Press, Cambridge, MA, 1961.
5. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2nd ed., MIT Press, Cambridge, MA, 1972.
6 J. P. Stenbit, Table of generators for Bose-Chaudhuri codes, *IEEE Trans. Inform. Theory* **IT-10**: 390–391 (1964).
7. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
8. A. Leon-Garcia and I. Widjaja, *Communication Networks*, McGraw-Hill, New York, 2000.
9. M. J. E. Golay, Notes on digital coding, *Proc. IRE* **37**: 657 (1949).
10. W. W. Peterson, Encoding and error-correction procedures for the Bose-Chaudhuri codes, *IRE Trans. Inform. Theory* **IT-6**: 459–470 (1960).

11. R. T. Chien, Cyclic decoding procedures for Bose-Chaudhuri-Hocquenghem codes, *IEEE Trans. Inform. Theory* **IT-10**: 357–363 (1964).

12. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

13. J. L. Massey, Shift-register synthesis and BCH decoding, *IEEE Trans. Inform. Theory* **IT-15**: 122–127 (1969).

14. T. Kasami, A decoding procedure for multiple-error correcting cyclic codes, *IEEE Trans. Inform. Theory* **IT-10**: 134–138 (1964).

BCH CODES — NONBINARY AND REED-SOLOMON*

ARNOLD M. MICHELSON
ALLEN H. LEVESQUE
Marlborough, Massachusetts

1. INTRODUCTION

Bose–Chaudhuri–Hocquenghem (BCH) cyclic block codes include both binary and nonbinary codes. The preceding article describes the fundamental properties and structure of *binary BCH codes*. This article treats nonbinary BCH codes and a closely related class of nonbinary codes called *Reed–Solomon codes*. Certain important and useful modifications of BCH and Reed–Solomon codes are also discussed. The approach to describing the structure of nonbinary codes closely parallels that used for binary codes in the preceding article, and reference to that discussion is made where appropriate.

The description of a nonbinary cyclic code follows directly from the binary case; that is, an (n, k) cyclic code defined on the Galois Field GF(q) can be generated as the set of all polynomials of the form $a(x)\ g(x)$, where $a(x)$ is any polynomial of degree $k-1$ or less with coefficients in GF(q) and the generator polynomial $g(x)$ divides $x^n - 1$ and has coefficients in GF(q). As in the binary case, we shall see that the design of a nonbinary code rests upon selection of a generator polynomial having prescribed roots in a field that is an extension of GF(q), say, GF(q^m).

1.1. Nonbinary BCH Codes

The binary BCH codes are a special case of a class of cyclic codes that can be constructed for any symbol alphabet defined on a finite field, say, GF(q), which can be a prime field or some extension of a prime field. As a generalization of the binary case, a t-error-correcting BCH code on GF(q) is a cyclic code, and all the codewords have roots that include $2t$ consecutive powers of some element β contained in GF(q^m), an extension field of GF(q). It will be convenient to distinguish between the two fields by calling GF(q) the *symbol field* and GF(q^m), the *locator field*. As with the binary codes, BCH codes on GF(q) can be primitive or nonprimitive, depending on whether a primitive or nonprimitive element of GF(q^m) is used to specify the consecutive roots of the codewords. For the present discussion, attention is restricted to the case of primitive codes, so that the code is specified to be a set of code polynomials whose roots include the elements $\alpha, \alpha^2, \ldots, \alpha^{2t}$, where α is a primitive element of GF(q^m). The design distance is one greater than the number of consecutive roots, and the true minimum distance can be equal to or greater than the design distance. The generator polynomial of a BCH code on GF(q) is defined as the least common multiple of the minimal polynomials of $\alpha, \alpha^2, \ldots, \alpha^{2t}$:

$$g(x) = \mathrm{LCM}[m_{\alpha^1}(x), m_{\alpha^2}(x), \ldots, m_{\alpha^{2t}}(x)]$$

The block length of the code is the order of the element chosen to prescribe the consecutive roots, and therefore, for the primitive codes, where we choose a primitive element of GF(q^m), the block length is $n = q^m - 1$.

In general, a t-error-correcting code may have either odd or even minimum design distance, given by $d = 2t + 1$, or $d = 2t + 2$, respectively. Furthermore, the sequence of powers of α can begin with an arbitrary power, say, m_0, so that we can specify the roots as $\alpha^{m_0}, \alpha^{m_0+1}, \ldots, \alpha^{m_0+d-2}$, that is, $d-1$ consecutive powers of α. Similarly, we can define the generator polynomial as

$$g(x) = \mathrm{LCM}[m_{\alpha^{m_0}}(x), m_{\alpha^{m_0+1}}(x), \ldots, m_{\alpha^{m_0+d-2}}(x)]$$

As mentioned previously, nonprimitive nonbinary BCH codes can be defined on GFq) as well. If $q^m - 1$ is factorable, a nonprimitive code with design distance d can be formed by specifying its roots to be $d-1$ consecutive powers of β, some nonprimitive element of GF(q^m). The block length n of the code is the order of β, that is, n divides $q^m - 1$. However, the most widely used nonbinary BCH codes are the Reed–Solomon codes, which we discuss next.

1.2. Reed–Solomon Codes

An important subclass of nonbinary BCH codes is obtained by choosing the locator field to be the same as the symbol field. These codes are called *Reed–Solomon codes* [1], often abbreviated as *RS codes*.

Specifically, an RS code on GF(q) with minimum distance d has as roots $d-1$ consecutive powers of α, a primitive element of GF(q). The minimal polynomial over GF(q) of any element γ in GF(q) is just $x - \gamma$. This means that the generator polynomial $g(x)$ for a design-distance-d RS code is

$$g(x) = (x - \alpha^{m_0})(x - \alpha^{m_0+1}) \cdots (x - \alpha^{m_0+d-2}) \qquad (1)$$

where m_0 is an arbitrary integer, usually chosen as 0 or 1. Since the order of α is $q-1$, the block length of an RS code is $q-1$. For any BCH code, the design distance is one greater than the number of consecutive roots in the locator field, and since from Eq. (1) the number of check symbols is always equal to the number of prescribed roots, we have for any RS code

$$d = n - k + 1$$

* Preparation of this article supported in part by the Raytheon Corporation.

where n is the block length and k is the number of information symbols in each block. An important property of any RS code is that the true minimum distance is equal to the design distance. No (n, k) linear block code can have minimum distance greater than $n - k + 1$, and a code for which the minimum distance equals $n - k + 1$ is called a *maximum-distance-separable* (MDS) code, or simply a *maximum code* [2]. Therefore, every RS code is an MDS code. Furthermore, shortening the block length of an RS code by omitting information symbols cannot reduce its minimum distance, and, therefore, we can state that any shortened RS code is also an MDS code.

2. ENCODING NONBINARY BCH CODES AND RS CODES

The formation of codewords in an RS code on GF(q) from its generator polynomial $g(x)$ extends directly from the binary case. Thus, the words in an (n, k) code correspond to the set of all polynomials over GF(q) of degree $n - 1$ or less that are divisible by $g(x)$, where the degree of $g(x)$ is $r = n - k$.

The codewords can be generated by multiplying all polynomials over GF(q) having degree $k - 1$ or less by $g(x)$. As was seen in the binary case, this will not produce a systematic code and is generally avoided. As in the case of the binary codes, systematic structure can be provided by forming codewords as

$$c(x) = x^r i(x) \bmod g(x) + x^r i(x)$$

where $i(x)$ denotes the k information symbols on GF(q) to be encoded represented as a polynomial of degree $k - 1$ or less.

Encoding can be implemented with a polynomial division circuit of the form described previously for binary BCH codes (see Fig. 1 in the article on binary BCH codes). However, multiplications and additions are now done in GF(q). As an example, we consider the (63,57) $d = 7$ RS code defined on GF(64). Assuming $m_0 = 1$, and letting α be the primitive element of GF(64), we have

$$g(x) = \prod_{i=1}^{6} (x + \alpha^i)$$

If we use the primitive binary polynomial $p(x) = x^6 + x + 1$ to provide a representation of GF(2^6), it is straightforward to show that

$$g(x) = x^6 + \alpha^{59}x^5 + \alpha^{48}x^4 + \alpha^{43}x^3 + \alpha^{55}x^2 + \alpha^{10}x + \alpha^{21}$$

An encoder for the (63,57) $d = 7$ RS code defined GF(64) is shown in Fig. 1. Each stage of the register is a 64-ary storage device, and the feedback lines require multiplication in GF(64). The feedback weights are the coefficients of the generator polynomial. The circuit shown in Fig. 1 operates as follows:

1. Enable gates G_1, G_2, and G_3. Disable gate G_4. Clock the information symbols to be encoded into the feedback shift register and simultaneously into the codeword buffer.
2. Disable gates G_1, G_2, and G_3, and enable gate G_4. The six parity symbols are now contained in the six storage elements of the feedback shift register. Clock these six symbols into the buffer to complete formation of the codeword.
3. All stages of the feedback shift register are now reset to zero, and the encoded word is shifted out while the next information set to be encoded is shifted in. Return to step 1.

3. DECODING RS CODES

We now describe algorithms for decoding RS codes that are generalizations of the bounded-distance decoding algorithms presented for binary BCH codes. Given the set of syndrome values calculated for the received word, the decoding task is to find the most likely error pattern, within the error-correction limit of the code, which produces the observed syndrome values. Therefore, as in the binary case, decoding is viewed as a problem of solving a set of simultaneous syndrome equations, but one where the set of unknowns now includes the error values or *error magnitudes* in addition to the error locators.

We use the notation adopted for the binary case, letting a transmitted codeword be represented by a polynomial

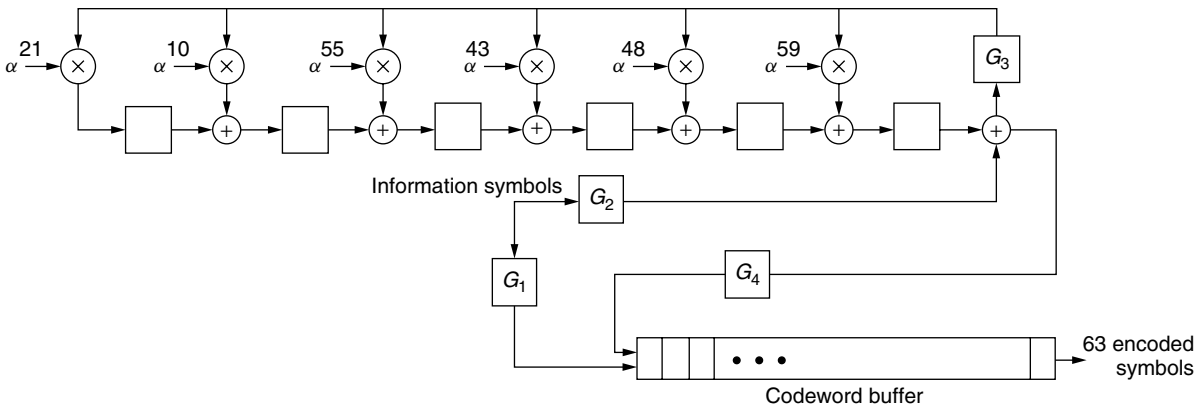


Figure 1. Encoder for the (63,57) RS code on GF(64).

$c(x)$, where here the coefficients of $c(x)$ are elements in GF(q). Similarly, a received error pattern is represented by a polynomial $e(x)$, again with coefficients in GF(q), and the received word is represented by $r(x)$, where

$$r(x) = c(x) + e(x)$$

The syndrome values are obtained by evaluating $r(x)$ at the prescribed roots of the generator polynomial:

$$\begin{aligned} S_k &= r(\alpha^k) \\ &= c(\alpha^k) + e(\alpha^k) \\ &= e(\alpha^k), \quad k = m_0, m_0+1, \ldots, m_0+d-2 \end{aligned} \tag{2}$$

where we let the roots of the code be any arbitrary sequence of consecutive powers of α, although m_0 is usually chosen to be 0 or 1.

The error polynomial $e(x)$ has nonzero terms only in those positions where errors have occurred, so that if there are t errors in the received word, we can write the syndrome values as

$$S_k = \sum_{i=1}^{t} Y_i X_i^k, \quad k = m_0, m_0+1, \ldots, m_0+d-2 \tag{3}$$

where X_i is the error locator for the ith error and Y_i is its value. Therefore, the decoding task is, given the S's, find the X and Y values. In a generalization of the procedure outlined for binary codes, syndrome decoding of an RS code proceeds as follows:

1. Calculate the syndrome values S_k, $k = m_0, m_0+1, \ldots, m_0+d-2$.
2. Determine the error locator polynomial $\sigma(x)$ from the syndrome values.
3. Solve for the roots of $\sigma(x)$, which are the error locators.
4. Given the error locators, calculate the error values.
5. Correct the indicated errors.

The fundamental difference between this sequence of steps and the procedure outlined for the binary case is step 4, calculation of the error values. However, once the error locations have been determined, finding the error values is straightforward, since, given the S and X values, Eq. (3) is simply a set of simultaneous linear equations having the error values as unknowns. As in the binary case, the most difficult part of the procedure is usually step 2, determination of the error locator polynomial $\sigma(x)$ from the syndrome values.

3.1. Peterson's Direct Solution Method

Peterson's direct solution method for finding the coefficients of the error locator polynomial $\sigma(x)$, generalizes in a straightforward way to the case of nonbinary codes, although there are a few important differences. The set of simultaneous nonlinear (in the X values) syndrome equations, Eq. (3), can be converted into a set of linear equations to be solved in conjunction with $\sigma(x)$. To begin, exactly as we did in Eqs. (6–9) in the article on binary BCH coding, we can operate repeatedly on $\sigma(x)$ and invoke the syndrome equations, Eq. (3), to establish the relationship

$$S_{t+j} + \sigma_1 S_{t+j-1} + \cdots + \sigma_t S_j = 0, \quad \text{for all } j \tag{4}$$

where the σ terms are coefficients of the error locator polynomial, $\sigma(x)$:

$$\sigma(x) = x^t + \sigma_1 x^{t-1} + \cdots + \sigma_t \tag{5}$$

The equations defined by Eq. (4) are *Newton's identities*.

Let us consider a t-error-correcting nonbinary BCH or RS code, for which we have computed $2t$ syndrome values $S_1, S_2, \ldots, S_{2t}$. From Eq. (4), we can construct t simultaneous equations, linear in coefficients of $\sigma(x)$, by letting j range from 1 through t. To illustrate this with an example, we consider the case of a three-error-correcting code so that we have

$$\begin{aligned} S_1\sigma_3 + S_2\sigma_2 + S_3\sigma_1 &= -S_4 \\ S_2\sigma_3 + S_3\sigma_2 + S_4\sigma_1 &= -S_5 \\ S_3\sigma_3 + S_4\sigma_2 + S_5\sigma_1 &= -S_6 \end{aligned} \tag{6}$$

The three equations have been written in a form suggesting their use, that is, as a set of simultaneous linear equations, with coefficients and constants that are the syndrome values. These equations are then solved for the three coefficients of $\sigma(x)$ when three errors are assumed to have occurred.

The reader should compare Eq. (6) with Eq. (10) in the article on binary BCH codes and note certain differences. First, unlike the binary case, Eq. (6) includes equations beginning with the even-indexed syndrome values. This is because the relationships $S_j^2 = S_{2j}$ are specific to the binary case and do not hold for nonbinary codes. Second, the simpler forms of the uppermost lines in Eq. (10) for the binary case are also specific to the binary case and do not apply here. Finally, negative signs are retained when the constants S_4, S_5, and S_6 are moved from the left side in Eq. (4) to the right side in Eq. (6), since addition and subtraction are identical only when the field is of characteristic 2.

As with binary codes, determining the locations of a given number of errors is done by constructing an appropriate set of simultaneous equations of the form given by Eq. (6) and solving the equations for the σ's in terms of the syndrome values $\{S_k\}$. The number of equations to be used is equal to the actual number of errors in the received code block, which must be determined as part of the decoding operation. This is done by testing determinants of various sizes corresponding to the possible numbers of errors. The equations for the σ terms in the three-error case are given in Eq. (6). We now write the sets of equations for the one-error and two-error cases, in the more compact matrix form, as follows:

$$[S_1][\sigma_1] = [-S_2] \tag{7}$$

$$\begin{bmatrix} S_1 S_2 \\ S_2 S_3 \end{bmatrix} \begin{bmatrix} \sigma_2 \\ \sigma_1 \end{bmatrix} = \begin{bmatrix} -S_3 \\ -S_4 \end{bmatrix} \tag{8}$$

Define D_2 as the determinant of the coefficient matrix in Eq. (8), that is $S_1S_3 - S_2^2$, and D_3 as the determinant of the 3×3 coefficient matrix in Eq. (6), which is

$$S_1S_3S_5 + S_2S_3S_4 + S_2S_3S_4 - S_3^3 - S_1S_4^2 - S_2^2S_5$$

Now, tests of D_2 and D_3 can be used to determine how many errors have occurred, and therefore, which set of equations should be used to solve for the σ terms. For example, if only one error has occurred, D_2 and D_3 will equal zero, and therefore, the Eqs. (6) and (8) will be indeterminate, and Eq. (7) is to be used.

Once the σ values have been determined, the error locator polynomial $\sigma(x)$ is formed and its roots obtained. The Chien search, already described for binary codes, can be used. The roots of $\sigma(x)$ are the error locator values, the X values. Once the X values have been determined, they are inserted into the syndrome equations, Eq. (3), which are then solved as linear equations for the error values, the Y terms. The steps in a direct solution decoding algorithm are described next using an example.

We describe the use of Peterson's direct solution method for decoding the (63,57) RS code defined on a 64-ary alphabet in more detail. Since the code has distance 7, it can be used to correct up to three errors in a received word or to correct combinations of l errors and s erasures such that $2l + s < 7$. In this discussion, however, we confine our attention to error-correction decoding. Combined errors-and-erasures decoding is treated later.

Since for an RS code, the symbol field and the locator field are the same, all computations for decoding are done in the field GF(64). Furthermore, since GF(64) is a field of characteristic 2, addition and subtraction are identical operations, which means, for example, that in determining the coefficients of $\sigma(x)$ and calculating error values, the minus signs can be replaced with plus signs. The 64 elements of the field may be represented conveniently as binary 6-tuples. Addition is then implemented with modulo-2 addition, applied bit by bit. For implementation in a processor, finite field multiplication and division are conveniently done with logarithm and antilogarithm tables and table lookup routines.

An error-correction decoder for the (63,57) RS code can be implemented as follows. Let the polynomial $r(x)$ represent the received word, where the high-order terms correspond to the information symbols and the low-order terms to the check symbols. The steps in the decoding process are

1. Compute the syndrome values S_k, $1 \le k \le 6$, where

$$S_k = r(\alpha^k) = \{\ldots[(r_{62}\alpha^k + r_{61})\alpha^k + r_{60}]\alpha^k + \cdots\}\alpha^k + r_0, \quad 1 \le k \le 6$$

2. Determine the number of errors in the received word:
   a. If $S_k = 0$, $1 \le k \le 6$, the received word is a codeword, and no further processing is necessary.
   b. If $D_3 = S_1S_3S_5 + S_1S_4^2 + S_2^2S_5 + S_3^3 \neq 0$, assume that three errors are present.
   c. If $D_3 = 0$ and $D_2 = S_1S_3 + S_2^2 \neq 0$, assume that two errors are present.
   d. If $D_2 = D_3 = 0$ and $S_1 \neq 0$, assume that one error is present.
3. Compute the coefficients of the error-locator polynomial:
   a. If three errors are present, compute

$$\sigma_1 = \frac{1}{D_3}[S_1S_3S_6 + S_1S_4S_5 + S_2^2S_6 + S_2S_3S_5 + S_2S_4^2 + S_3^2S_4]$$

$$\sigma_2 = \frac{1}{D_3}[S_1S_4S_6 + S_1S_5^2 + S_2S_3S_6 + S_2S_4S_5 + S_3^2S_5 + S_3S_4^2]$$

$$\sigma_3 = \frac{1}{D_3}[S_2S_4S_6 + S_2S_5^2 + S_3^2S_6 + S_4^3]$$

   b. If two errors are present, compute

$$\sigma_1 = \frac{1}{D_2}[S_1S_4 + S_2S_3]$$

$$\sigma_2 = \frac{1}{D_2}[S_2S_4 + S_3^2]$$

   c. If one error is present, compute

$$\sigma_1 = X_1 = \frac{S_2}{S_1}$$

4. If three errors are indicated in step 3, find (using the Chien search) the roots of the polynomial $\sigma(x)$, where

$$\sigma(x) = x^3 + \sigma_1x^2 + \sigma_2x + \sigma_3$$

   If two errors are indicated, find the roots of

$$\sigma(x) = x^2 + \sigma_1x + \sigma_2$$

   Of course, in the case of three errors, three distinct roots of $\sigma(x)$ must be found, and for the two-error case, $\sigma(x)$ must have two distinct roots. If the correct number of roots is not found, error detection is announced.
5. After the error locators are determined, the error values are obtained by solving the syndrome equations.
   a. *One-error case*:

$$Y_1 = \frac{S_1^2}{S_2}$$

   b. *Two-error case*:

$$Y_1 = \frac{S_1X_2 + S_2}{X_1X_2 + X_1^2}$$

$$Y_2 = \frac{S_1X_1 + S_2}{X_1X_2 + X_2^2}$$

c. *Three-error case*:
Let

$$C = X_1X_2^2X_3^3 + X_1^3X_2X_3^2 + X_1^2X_2^3X_3 + X_1^3X_2^2X_3 + X_1X_2^3X_3^2 + X_1^2X_2X_3^3$$

Then

$$Y_1 = \frac{1}{C}\left[S_1X_2^2X_3^3 + S_2X_2^3X_3 + S_3X_2X_3^2 + S_1X_2^3X_3^2 + S_2X_2X_3^3 + S_3X_2^2X_3\right]$$

$$Y_2 = \frac{1}{C}\left[S_1X_3^2X_1^3 + S_2X_1X_3^3 + S_3X_1^2X_3 + S_1X_1^2X_3^3 + S_2X_1^3X_3^3 + S_3X_1X_3^2\right]$$

$$Y_3 = \frac{1}{C}\left[S_1X_1^2X_2^3 + S_2X_1^3X_2 + S_3X_1X_2^2 + S_1X_1^3X_2^2 + S_2X_1X_2^3 + S_3X_1^2X_2\right]$$

It should be noted that when the denominators in the expressions for Y_1, Y_2, and Y_3 are written out, the expressions can be simplified.

6. Correct the received word by adding the computed error values to the corresponding symbols received in positions identified as error locations.
7. Compute the syndrome of the corrected word, and if it is not zero, announce error detection.

The correction of both errors and erasures is discussed in Section 3.3. We first present an efficient iterative decoding algorithm for correction of errors in nonbinary BCH and RS codes.

3.2. The Massey–Berlekamp Algorithm

For correction of moderate to large numbers of errors with a nonbinary BCH or RS code, Peterson's direct method of solving for the coefficients of $\sigma(x)$ from the syndrome values becomes cumbersome and inefficient due to the large number of multiplications and divisions that must be performed. Instead, it is preferable to use either of two algorithms developed by Berlekamp [3] and by Massey [4] for solution of Newton's identities. The two algorithms are closely related and are often referred to as one procedure, the *Massey–Berlekamp algorithm*. The approach used by Massey in presenting the technique is particularly instructive, and thus we shall follow Massey closely here. Berlekamp's formulation with simplifications applicable to decoding binary codes was described for binary codes. Both Massey's and Berlekamp's versions of the algorithm can be used for binary and nonbinary codes.

We let $m_0 = 1$ and return to the error-locator polynomial $\sigma(x)$ in Eq. (5). Then, substituting an error locator X_j for x, we obtain

$$X_j^t + \sigma_1 X_j^{t-1} + \cdots + \sigma_t = 0, \quad j = 1, 2, \ldots, t \tag{9}$$

Multiplying Eq. (9) by X_j^k and summing for $j = 1, 2, \ldots, t$, we obtain

$$S_{k+t} + \sigma_1 S_{k+t-1} + \cdots + \sigma_t S_k = 0, \quad k = 1, 2, \ldots \tag{10}$$

which are again Newton's identities. Letting $j = k + t$, we obtain

$$S_j + \sigma_1 S_{j-1} + \cdots + \sigma_t S_{j-t} = 0, \quad j = t+1, t+2, \ldots \tag{11}$$

With Newton's identities written in this form, one can recognize that they describe the operation of a linear feedback shift register (FSR) with initial states $S_1, S_2, \ldots, S_t$ and tap connections given by $C_i = \sigma_i$. A diagram of a linear FSR is shown in Fig. 2. From the figure, it is seen that the FSR implements the equations.

$$S_j = -C_1S_{j-1} - C_2S_{j-2} - \cdots - C_tS_{j-t}, \quad j = t+1, t+2, \ldots \tag{12}$$

or

$$S_j + C_1S_{j-1} + \cdots + C_tS_{j-t} = 0, \quad j = t+1, t+2, \ldots \tag{13}$$

With $C_i = \sigma_i$, the correspondence with Eq. (11) is immediate.

Recognizing the relationship just obtained, Massey established the equivalence between the problem of determining the coefficients of the error locator polynomial from the syndrome values and that of synthesizing an FSR with minimum length that generates the given sequence of syndromes. We shall provide a rationale for this in the following.

We define the *connection polynomial* as a convenient representation for the coefficients of the syndrome values in Eq. (13):

$$C(x) = 1 + C_1x + C_2x^2 + \cdots + C_tx^t \tag{14}$$

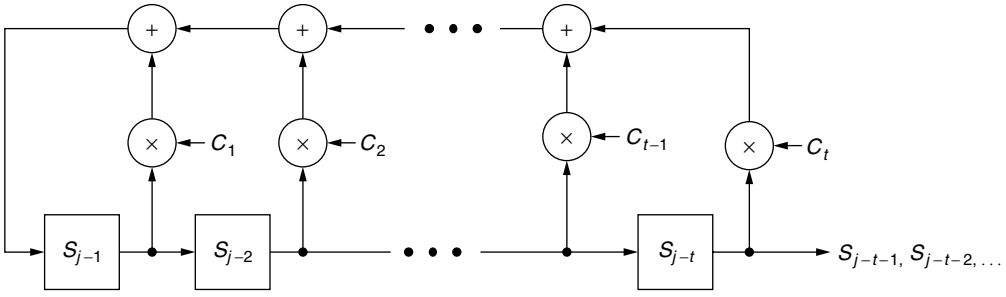


Figure 2. Linear feedback shift register for generating a sequence of syndrome values.

We now state that the problem of determining the error locator polynomial $\sigma(x)$ is equivalent to that of determining a connection polynomial $C(x)$ for a linear FSR that generates the syndrome values $S_{t+1}, S_{t+2}, \ldots$, given that the FSR is initialized with $S_1, S_2, \ldots, S_t$.

Note that in Eqs. (11)–(14), as well as in Fig. 2, the assumed length of the FSR is t stages, where t is the error correction limit of the code. However, the iterative algorithm is designed to correct l errors, where $l \leq t$. The number of errors l is not known at the start of decoding and is determined as part of the decoding procedure.

Without delving into the details of the properties of FSRs and the sequences that they generate, we simply point out that for a given sequence of syndrome values, a determinable number of connection polynomials of various lengths will generate the syndromes. This corresponds directly to the fact that, in general, a number of error patterns can account for a given set of syndrome values. However, the task of bounded-distance decoding is to find the lowest-weight error pattern corresponding to the given syndrome. Therefore, in the FSR synthesis problem, we seek the lowest degree connection polynomial $C(x)$ that generates the syndrome. In his 1969 paper [4], Massey described an algorithm that finds the minimal-length FSR. He further showed that given an error pattern of weight $l \leq t$, the algorithm yields the connection polynomial that uniquely corresponds to the correct error-locator polynomial. Massey's algorithm is often called the *FSR synthesis algorithm*. Shift register sequences are also described in [5].

Before describing the FSR synthesis algorithm in detail, we outline the procedure as follows. The FSR algorithm synthesizes the minimal-length shift register with an iterative routine that begins by postulating the shortest possible shift register and then attempts to generate the entire sequence of given syndrome values in order. The actual syndrome sequence is repeatedly compared with the output of the postulated FSR until either the entire sequence of given syndrome values is reproduced or a discrepancy is encountered. At the first discrepancy, the postulated FSR is modified with a specified rule, and the sequence generation is restarted and continued until all the remaining syndromes are reproduced or another discrepancy is encountered, and so forth. The modification rule is designed to ensure that for a correctable error pattern, the FSR eventually settles into the correct configuration. The FSR synthesis algorithm is described in detail below.

The Massey FSR Synthesis Algorithm

0. *Compute syndrome values*: S_n, $1 \leq n \leq d-1$.
1. *Initialize algorithm variables*:

   Let $C(x) = 1 \quad D(x) = x$

   $L = 0 \qquad n = 1$

2. *Take in new syndrome value and compute discrepancy*:

$$\delta = S_n + \sum_{i=1}^{L} C_i S_{n-i}$$

3. *Test discrepancy*. If $\delta = 0$, go to step 8; otherwise, go to step 4.
4. *Modify connection polynomial*. Let $C^*(x) = C(x) - \delta D(x)$.
5. *Test register length*. If $2L \geq n$, go to step 7 (i.e., do not extend register); otherwise, go to step 6.
6. *Change register length and update correction term*. Let $L = n - L$ and $D(x) = C(x)/\delta$.
7. *Update connection polynomial*. Let $C(x) = C^*(x)$.
8. *Update correction term*. Let $D(x) = xD(x)$
9. *Update syndrome counter*. Let $n = n + 1$.
10. *Test syndrome count*. If $n < d$, go to step 2; otherwise, stop.

In the algorithm, $C(x)$ is the FSR connection polynomial. The algorithm is designed to expediently build up the polynomial $C(x)$ of lowest degree that generates the given sequence of syndromes, $S_1, S_2, \ldots, S_{d-1}$. The connection polynomial is first initialized to its simplest possible form, $C(x) = 1$, and is subsequently modified as needed to correctly reproduce the syndrome values in sequence. The other polynomial formed in the algorithm, $D(x)$, is a correction term that is used to modify $C(x)$ at each iteration in which a discrepancy is encountered between a generated value and the corresponding syndrome value. The syndromes are examined by the algorithm in sequence, one in each iteration. At each iteration, the discrepancy δ, the difference between the newly entered syndrome value and the value generated by the FSR in the corresponding sequence position, is computed, using the connection polynomial $C(x)$ as it was structured at the end of the previous iteration. Note that δ is defined in such a way that at the first entry into step 2, it is given the value of the first syndrome S_1, even though there are no previous syndrome values from which S_1 could be generated. At each appearance of a nonzero value for δ, the connection polynomial is modified using the computed value of δ and the correction term (step 4). The formation and use of the correction term is the most important part of the algorithm. One reason is that in addition to zeroing out the encountered discrepancy, the modification of $C(x)$ is such that the new $C(x)$ also correctly generates all the previous syndrome values. This obviates the necessity of having to reexamine previous syndromes each time $C(x)$ is modified, and provides an algorithm in which the number of computations required per decoding is a linear function of l rather than some geometric function. Another important characteristic of the polynomial modification procedure is that it accomplishes the needed sequence modification with the smallest possible increase in the degree of the connection polynomial. The other variable used in the algorithm is L, which is the current length of the FSR. If the algorithm terminates with an FSR connection polynomial of degree greater than t, that is, $2L > d-1$, then we are not assured that the corresponding error locator polynomial is correct, and error detection is announced.

As an example, consider the case of the (31,25) RS code on GF(32), with $m_0 = 1$, for which codewords all

have the six consecutive roots $\alpha, \alpha^2, \ldots, \alpha^6$. Let the all-zeros codeword be transmitted, and assume the received word

$$000\alpha^7 00000000\alpha^3 0000000\alpha^{22} 0000000000$$

Thus, we have

$$r(x) = \alpha^7 x^3 + \alpha^3 x^{12} + \alpha^{22} x^{20}$$

To represent elements in GF(32), we use the primitive polynomial $p(x) = x^5 + x^2 + 1$. Then the syndrome values are as follows:

$$\begin{aligned} S_1 &= r(\alpha) = \alpha^{29} \\ S_2 &= r(\alpha^2) = \alpha^{28} \\ S_3 &= r(\alpha^3) = \alpha^9 \\ S_4 &= r(\alpha^4) = \alpha^4 \\ S_5 &= r(\alpha^5) = \alpha^{24} \\ S_6 &= r(\alpha^6) = \alpha^{19} \end{aligned}$$

We next use the FSR synthesis algorithm to find the shortest connection polynomial $C(x)$ that generates the six syndrome values in order. The iterative solution is summarized as follows:

n	S_n	$C(x)$	δ	L
1	α^{29}	1	α^{29}	0
2	α^{28}	$1 + \alpha^{29}x$	α^{14}	1
3	α^9	$1 + \alpha^{30}x$	α^{10}	1
4	α^4	$1 + \alpha^{30}x + \alpha^{12}x^2$	α^{11}	2
5	α^{24}	$1 + \alpha^4 x + \alpha^{23}x^2$	α^{10}	2
6	α^{19}	$1 + \alpha^4 x + \alpha^{12}x^2 + \alpha^{30}x^3$	α^{19}	3
7		$1 + \alpha^6 x + \alpha^{30}x^2 + \alpha^4 x^3$	(STOP)	

Thus, the minimal-length connection polynomial is found to be

$$C(x) = 1 + \alpha^6 x + \alpha^{30}x^2 + \alpha^4 x^3$$

and with $\sigma_i = C_i$, we can write the error locator polynomial as

$$\sigma(x) = x^3 + \alpha^6 x^2 + \alpha^{30} x + \alpha^4$$

The three roots of $\sigma(x)$, the error locator numbers, are found to be

$$X_1 = \alpha^3, \quad X_2 = \alpha^{12}, \quad X_3 = \alpha^{20}$$

which point to errors in the 4th, 13th, and 21st symbol positions.

The error magnitudes, Y_1, Y_2, and Y_3 are now computed using the equations shown previously for the three-error case. The reader may verify that the computations yield

$$Y_1 = \alpha^7, \quad Y_2 = \alpha^3, \quad Y_3 = \alpha^{22}$$

Finally, error correction is completed by subtracting [or adding, in GF(32)] the error values from the corresponding received symbols, which yields the all-zeros word as the corrected codeword.

3.3. Errors-and-Erasures Decoding

If some procedure is being used to erase unreliable symbols in a received word, then the function of the decoder is to fill in the proper values of the erasures and at the same time locate and correct any unknown errors. We recall from earlier discussions that a code of minimum distance d is capable of correcting any pattern of l errors and s erasures as long as $2l + s < d$. We now outline an efficient procedure for simultaneous errors-and-erasures decoding, which was suggested by Forney [6] for nonbinary BCH codes. First, we define the *erasure locator polynomial* $\sigma'(z)$, which is the polynomial of degree s whose roots are the erasure locators:

$$\begin{aligned} \sigma'(z) &= \prod_{i=1}^{s} (z + Z_i) \\ &= \sigma_0' z^s + \sigma_1' z^{s-1} + \cdots + \sigma_s' \end{aligned} \tag{15}$$

where Z_i gives the location of the ith erasure.

It should be noted that $\sigma'(z)$ is written in much the same form as the error locator polynomial, Eq. (5), except that for notational convenience we have given the term of highest degree the coefficient σ_0' even though it always has value 1. Since, by definition, the erasure location values are known, the coefficients of $\sigma'(z)$ may be computed directly. We also assume use of a primitive code with $m_0 = 1$.

Combined errors-and-erasures decoding begins, as does error correction decoding, with calculation of the syndrome values, which are

$$S_k = \sum_{i=1}^{n} r_i \alpha^{ik}, \quad 1 \le k \le d - 1$$

where we denote $d - 1$ rather than $2t$ syndrome values, to allow for both odd and even values of d. The reader may well ask what values should be assigned to the erasures for the syndrome calculation, but it will be seen shortly that these values are immaterial to the decoding procedure. As a practical matter, it is usually advantageous to assign zeros for all the erasure values.

To take account of the known erasure-location information in forming the syndromes, Forney introduced a linear transformation on the syndromes:

$$T_i = \sum_{j=0}^{s} \sigma_j' S_{i+s+1-j}, \quad 0 \le i \le d - s - 2 \tag{16}$$

The T values are called the *modified syndromes*. Notice that there are s fewer T than S symbols. Thus, if one symbol is erased, the $d - 1$ original syndromes are transformed by Eq. (16) into $d - 2$ modified syndromes, and so forth. We shall see how this transformation lets us establish a useful recursion among the T symbols.

Let us assume the presence of l errors and s erasures. Let the errors be at locations $X_1, X_2, \ldots, X_l$ and have values

$Y_1, Y_2, \ldots, Y_l$. Let the known erasure locations be denoted by $Z_1, Z_2, \ldots, Z_s$, and let $D_1, D_2, \ldots, D_s$ designate the erasure-discrepancy values, that is, the difference between the correct symbol values and the values arbitrarily assigned before the syndromes are computed. We can now express the syndromes as

$$S_k = \sum_{m=1}^{l} Y_m X_m^k + \sum_{n=1}^{s} D_n Z_n^k, \; 1 \le k \le d-1$$

From Eq. (16), we write the modified syndromes as

$$T_i = \sum_{j=0}^{s} \sigma_i' \left[\sum_{m=1}^{l} Y_m X_m^{i+s+1-j} + \sum_{n=1}^{s} D_n Z_n^{i+s+1-j} \right], \quad 0 \le i \le d-s-2$$

or

$$T_i = \sum_{m=1}^{l} Y_m X_m^{i+1} \sum_{j=0}^{s} \sigma_j' X_m^{s-j} + \sum_{n=1}^{s} D_n Z_n^{i+1} \sum_{j=0}^{s} \sigma_j' Z_n^{s-j} \quad (17)$$

However, from Eq. (15), we see that the second summation in the last term of Eq. (17) is the erasure locator polynomial evaluated at a root Z_n, which equals zero. Further, we recognize from Eq. (15) that the second summation in the first term of the right-hand side of Eq. (17) is simply the erasure locator polynomial evaluated at the error location X_m, which we write as $\sigma'(X_m)$. Therefore, if we define a new quantity E_m as

$$E_m = Y_m X_m \sigma'(X_m) \quad (18)$$

we can rewrite Eq. (17) as

$$T_i = \sum_{m=1}^{l} E_m X_m^i, \quad 0 \le i \le d-s-2 \quad (19)$$

What is important to note here is that Eq. (19) defines the modified syndrome values in a manner essentially the same as that in which the ordinary syndrome values are defined for l-error correction, for example, by Eq. (3). Thus, we see that for the simultaneous decoding of l errors and s erasures, the transformation in Eq. (16) has the effect of folding the known erasure locators into the original syndromes in such a way that we preserve the form of the syndrome equations in terms of the error locators. Now, by starting with Eq. (19) as the formulation of a new decoding problem, where l error locators X_m are to be determined, we can perform decoding with much the same overall procedure as is used for the case of ordinary error correction. That is, we first find the l error locators from the T values, and then compute the values of code symbols in the $l+$ s error and erasure locations.

If Peterson's direct solution method is used, error locator polynomial coefficients are computed from the T values in the same way as they are computed from the S values in the earlier discussion of errors-only decoding. After solving for the roots of $\sigma(x)$, any $l+s$ of the syndrome equations can be used as a set of simultaneous linear equations to solve for the Y and D values.

Alternatively, the Massey FSR synthesis technique may be applied in almost the same way as for ordinary error correction. That is, using Eq. (19) instead of Eq. (3), we treat the relationship of the T values to the σ values in a manner that exactly parallels the discussion in Eqs. (9)–(14), developing along the way a recursion relationship for the T values equivalent to Eq. (11):

$$T_j + \sigma_1 T_{j-1} + \cdots + \sigma_l T_{j-l} = 0, \quad j = l, l+1, \ldots$$

Thus, the problem of finding the coefficients of the error locator polynomial can be formulated again as an FSR synthesis problem, where the FSR must now be synthesized to generate a given sequence of modified syndrome values $\{T_j\}$ rather than original syndrome values $\{S_k\}$.

Once the error-locator polynomial is obtained, the roots are found efficiently using the Chien search. The l error locators, taken together with the s known erasure locators, in effect constitute $l+s$ erasures whose values are to be computed from the original syndromes. This can be done by solving $l+s$ syndrome equations, as in the direct method. However, a more efficient method of determining the erasure values has also been given by Forney [6]. Although we mention this method as part of the errors-and-erasures decoding procedure, it is also applicable in the case of errors-only decoding, since it is applied at the point in decoding where all of the unknown errors have been located. The suggested erasure-filling procedure is now described.

Let us denote the given erasure locators and computed error locators together by $Z_1, Z_2, \ldots, Z_{l+s}$. Now consider deleting Z_1 from the set of $l+s$ erasure locators, and forming the erasure locator polynomial ${}_1\sigma(z)$, which has as roots the remaining $l+s-1$ locators. Next, we calculate the coefficients of

$${}_1\sigma(z) = {}_1\sigma_0 z^{l+s-1} + {}_1\sigma_1 z^{l+s-2} + \cdots + {}_1\sigma_{l+s-1}$$

Then the erasure correction value, to be subtracted from the received or assigned value in the location Z_1, is given by

$$D_1 = \frac{\sum_{k=0}^{l+s-1} {}_1\sigma_k S_{l+s-k}}{\sum_{j=0}^{l+s-1} {}_1\sigma_j Z_1^{l+s-j}}$$

or in general by

$$D_i = \frac{\sum_{k=0}^{l+s-1} {}_i\sigma_k S_{l+s-k}}{\sum_{j=0}^{l+s-1} {}_i\sigma_j Z_i^{l+s-j}}, \quad 1 \le i \le l+s \quad (20)$$

By deleting one erasure at a time, all erasure values are calculated in turn by Eq. (20). Another procedure requiring even fewer computations can also be used. If,

after computing D_1, the syndrome values are modified using

$$S'_k = S_k + D_1 Z_1^k$$

it is only necessary to form an $(l+s-2)$-order erasure locator polynomial, with coefficients ${}_2\sigma_1, {}_2\sigma_2, \ldots, {}_2\sigma_{l+s-2}$, in order to find D_2 from Eq. (20), and so forth.

We now summarize the procedure for errors-and-erasures decoding, assuming use of the FSR synthesis algorithm, as follows:

1. Inspect the received word for erasures, assign erasure values (e. g., all 0s) and compute the syndrome.
   a. If $s > d-1$, declare the word undecodable.
   b. Otherwise, compute the syndrome values $S_1, S_2, \ldots, S_{d-1}$. If all syndromes are zero, the received word is a valid codeword, and no further processing is to be done.
2. If no symbols have been erased ($s = 0$), follow the procedure for errors-only decoding.
3. Compute the modified syndrome (if necessary).
   a. If $s = d-1$, go to step 6.
   b. If $0 < s < d-1$, compute the modified syndrome values using Eq. (16).
4. Determine the number of errors in the received word.
   a. If all $T_i = 0$, $0 \le i \le d-s-2$, assume that no errors are present and go to step 6.
   b. If some $T_i \neq 0$, use the FSR synthesis algorithm to find $\sigma_1, \sigma_2, \ldots, \sigma_l$, the coefficients of the error locator polynomial $\sigma(x)$.
5. Determine the error locators, the roots of $\sigma(x)$, using the Chien search. Put the l computed error locators together with the given s erasure locators to make up the new set of erasure locators $Z_1, Z_2, \ldots, Z_{l+s}$.
6. Compute the $l+s$ erasure magnitudes, using Eq. (20) or the more efficient procedure discussed immediately following Eq. (20).

4. FINAL COMMENTS

Errors and erasures decoding is the simplest form of *soft-decision decoding*. More complex techniques have been proposed for decoding Reed Solomon codes, for example, generalized minimum distance (GMD) decoding [7]. GMD is a technique that uses a sequence of decode attempts to find the most likely transmitted codeword. Each time a word is received, a trial decode list is executed. First, errors-only decoding is attempted and the outcome is recorded. Then errors and erasures decoding attempts are executed where we first erase the least reliable received symbol, the one with the smallest matched-filter output, and execute a decode attempt. Next, we erase the three least reliable symbols, then five, and so on up to a decode attempt in which the $d-1$ least reliable received symbols are erased where d is the minimum distance of the code. Clearly, this procedure can result in more than one decoded output, which we resolve by choosing the candidate codeword with highest likelihood.

Other soft-decision decoding schemes for RS codes are described by Cooper [8]. However, efficient soft-decision decoding of Reed–Solomon codes is still considered an open issue. As with the binary codes, what is needed is a way to decode beyond the code's minimum distance with an efficient soft-decision algorithm.

In this and the article on binary BCH coding, binary BCH codes and Reed–Solomon codes have been considered and the commonly used encoding and decoding algorithms have been described. The presentation closely follows the early developments as they were originally published. More recent treatments have proved useful as well, providing new insights into the underlying fundamentals. In some cases, more efficient encoding and decoding algorithms have resulted. A good example is Blahut's description of binary and nonbinary BCH codes in terms of Fourier transforms [9].

To conclude this article, we note again that binary BCH codes and RS codes have been incorporated into many communication systems. The wide ranges of block lengths and error correction power afforded by these codes have enabled designers to tailor solutions for particular applications, and provide significant performance gains relative to uncoded transmission. Given the continuing growth and advancement of the digital communications field, the number of applications for BCH and RS codes will certainly increase.

BIOGRAPHIES

Arnold M. Michelson received his BSEE degree from the Johns Hopkins University, his MSEE from the University of Rochester, New York, and he did further graduate work at the Polytechnic Institute of Brooklyn, New York. In 1968, Mr. Michelson joined Sylvania Electric Products, which later became GTE Government Systems. At Sylvania and GTE he worked on the development and implementation of advanced communication techniques, including error-control coding for military applications. That work focused primarily on long-wave communications. Since 2000, he has been with the Raytheon Company where he is involved in the development of high-performance coding techniques for military and commercial satellite applications. In 1997, Mr. Michelson received GTE's Leslie H. Warner Technical Achievement Award, and, in 2002, Raytheon's Excellence in Technology Award, Distinguished Level.

Allen H. Levesque received his BSEE degree from Worcester Polytechnic Institute in 1959 and his MSEE and PhDEE degrees from Yale University in 1960 and 1965, respectively. Following completion of his graduate studies, he joined the GTE Corporation, where, over a 36-year career, he worked on and led a variety of digital communications research and development projects, with application to both defense and commercial systems. Much of his early work concerned applications of error-control coding techniques in radio networks. For the past decade, his work has concentrated on mobile and wireless communications networks. In early 1999, he retired from GTE Laboratories to begin and independent consulting

practice and to take a part-time teaching and research position at WPI. He currently teaches graduate courses in modulation and coding and is a member of WPI's Center for Wireless Information Network Studies. His areas of research interest include communication theory and techniques, communication networks, wireless communications, spread-spectrum, secure communications, and digital signal processing. He has published numerous journal and conference papers, and coauthored two books, as well as chapters in several communications handbooks. He is a life fellow of the IEEE and a Registered Professional Engineer in the Commonwealth of Massachusetts.

BIBLIOGRAPHY

1. I. S. Reed and G. Solomon, Polynomial codes over certain finite fields, *J. SIAM* **8**: 300–304 (1960).
2. R. C. Singleton, Maximum distance q-nary codes, *IEEE Trans. Inform. Theory* **IT-10**: 116–118 (1964).
3. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
4. J. L. Massey, Shift register synthesis and BCH decoding, *IEEE Trans. Inform. Theory* **IT-15**: 122–127 (1969).
5. S. W. Golomb, *Shift Register Sequences*, Holden Day, San Francisco, 1967 (revised ed., Aegean Park Press, Laguna Hills, CA, 1982).
6. G. D. Forney, Jr., On decoding BCH codes, *IEEEE Trans. Inform. Theory* **IT-9**: 64–74 (1963).
7. G. D. Forney, Jr., Generalized minimum distance decoding, *IEEE Trans. Inform. Theory* **12**: 125–131 (1966).
8. S. B. Wicker and V. K. Bhargava, eds., *Reed–Solomon Codes and Their Applications*, IEEE Press, New York, 1994.
9. R. E. Blahut, *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, MA, 1983.

BISDN (BROADBAND INTEGRATED SERVICES DIGITAL NETWORK)

ENDER AYANOGLU
University of California
Irivne, California

NAIL AKAR
Bilkent University
Ankara, Turkey

1. BROADBAND ISDN

We first outline the history of the BISDN vision and then move on to the ATM technology that is envisioned to fulfill this vision.

1.1. History of BISDN

Shortly after the invention of the telephone by A. G. Bell in 1876, means to interconnect or network telephones at different locations were devised. Within only 2 years, the first switching center was built [4]. In the United States during the twentieth century, a public company, the Bell System and its parent, AT&T, emerged as the national provider of telephony services. The fundamental principle, formulated by AT&T president T. Vail in 1907, was that the telephone would operate most efficiently as a monopoly providing universal service, by nature of its technology. The U.S. government accepted this principle in 1913. The Bell System made steady progress toward its goal of universal service, which came in the 1920s and 1930s to mean that everyone should have a telephone. Percentage of American households with telephone service reached 50% in 1945, 70% in 1955, and 90% in 1969. This network was based on analog technology for transmission, signaling, and switching.

The Bell System studied digital telephony, first starting from its theoretical principles during the late 1940s. Most of the principles of digital telephony, such as theory of sampling, theory of quantization, and fundamental limits in information transfer, were invented or perfected by Bell System scientists such as H. Nyquist, J. R. Price, S. P. Lloyd, and C. E. Shannon in the late 1940s. Parallel with this progress in theory was a fundamental breakthrough in device technology known as the *transistor*, introduced, again by the Bell System, in 1948. The transistor would make the digital telephony revolution possible, while many years later, powerful integrated circuits would spark the dream of BISDN.

Digitization of the telephony network was useful since it provided a number of advantages:

- Ease of multiplexing
- Ease of signaling
- Integration of transmission and switching
- Increased noise tolerance
- Signal regeneration
- Accommodation of other services
- Performance monitoring
- Ease of encryption

The first deployment of digital transmission was in 1962 by the Bell System, while the first digital commercial microwave system was deployed in Japan in 1968. Research on digital switching was initiated in 1959 by Bell Labs. The first deployment of a digital switch in the public network was in 1970 in France while in the United States, the Bell System deployed an electronic switch known as 4ESS in 1976 [4].

CCITT (Comité Consultatif International de Télégraphique et Téléphonique, or Consultative Committee for International Telegraph and Telephone), is a committee of the International Telecommunications Union (ITU), which is a specialized agency of the United Nations. ITU was originally established after the invention of telegraphy in 1865 and became a specialized agency of the United Nations in 1947, shortly after the formation of the United Nations. Similar to ITU, CCITT was originally established as a standardization organization in the field of telegraphy, in 1925. In 1993, standardization aspects of CCITT and those of the sister radio standardization committee, CCIRR, were unified

under the name ITU-T (International Telecommunications Union — Telecommunication Standardization Sector). Members of ITU-T are governments. ITU-T is currently organized into 13 study groups that prepare standards, called Recommendations. There are 25 Series of Recommendations (A–V, X–Z). Work within ITU-T is conducted in 4-year cycles.

In 1968, CCITT established Special Study Group D to study the use of digital technology in the telephone network. This study group established 4-year study periods beginning with 1969. The first title of the group was "Planning of Digital Systems." By 1977, the emphasis of the study group was on overall aspects of integrated digital networks and integration of services. As of 1989, the title of the study shifted to "General Aspects of Integrated Services Digital Networks." The concept of an integrated services digital network was formulated in 1972 as one in which "the same digital services and digital paths are used to establish for different services such as telephony and data" [29]. The first ISDN standard was published in 1970, under the title "G.705 Integrated Services Digital Network (ISDN)." Although this first document of an ISDN standard is in the Series G Recommendations, most of the ISDN standards are in the Series I Recommendations, with some also in G, O, Q, and X Series Recommendations.

Three types of ISDN services are defined within the ISDN Recommendation I.200:

- Bearer services
- Teleservices
- Supplementary services

Bearer services (I.140) provide the means to convey information in the form of speech, data, video, and other forms of communication between users. There is a common transport rate for bearer services: it is the 64 kbps (kilobits per second) rate of digital telephony. Various bearer services are defined as multiples of this basic 64-kbps service, for example, 64, 2×64, 384, 1536, and 1920 kbps [29]. Teleservices cover user applications and are specified in I.241 as telephony, teletex, telefax, mixed mode, videotex, and telex. Supplementary services are defined in I.250. These services are related to number identification (such as calling line identification), call offering (such as call transfer, call forwarding, and call deflection), call completion (such as call waiting and call hold), multiparty (such as conference or three-party calling), community of interest (such as a closed user group), charging (such as credit card charging), and additional information transfer (such as the use of the ISDN signaling channel for user-to-user data transfer).

Toward the end of 1980s and almost two decades after the first study group on ISDN was formed at the CCITT, ISDN was still not being deployed by service providers at a commercial scale, especially in the United States. It is important to review the reasons for this absence of activity. ISDN required digitization of both the telephony network and the subscriber loop (connection between a residence or a business and the central office of the telecommunications service provider). While the network was becoming digital, and doing so involved economies of scale (and thus was relatively inexpensive), making the subscriber loop digital required replacement of the subscriber front end equipment at the central office. This was a labor-intensive, expensive process. In addition, there was no compelling push from consumers demanding ISDN. With the network becoming digital, the quality and reliability advantages of voice transmission were achieved. In addition, it was possible to offer supplementary services (such as caller ID and call waiting) as defined by ISDN Recommendations without making the subscriber loop digital. At the time, modem technology enabled data transmission over the subscriber loop at rates up to ~30 kbps, and that was sufficient for most of the available residential data services available (which were text-based). Business data communications needs were restricted to large businesses. These needs were being served with dedicated digital lines (T1 lines) at speeds of 1.5 Mbps in the United States. Although these lines were very expensive, the market for them was relatively small. In addition, it was becoming clear that in order to serve any future ISDN service needs, ISDN transmission speeds would not suffice and packet switching was going to become necessary. At the time, some overestimates were made as to the needed transmission speeds. For example, it was considered that entertainment video was one of the services that service providers would offer on such an integrated network and that the required transmission speeds for these services were in excess of 100 Mbps. ISDN was certainly insufficient to provide these speeds, and its packet switching recommendations were not yet developed.

In 1988, CCITT issued a set of Recommendations for ISDN, under the general name of "broadband aspects of ISDN" (I.113: *Vocabulary of Terms for Broadband Aspects of ISDN*, and I.121: *Broadband Aspects of ISDN*). This was a time when packet switching was proved in the Internet (although the Internet was still a research network), there was increased activity in video coding within the contexts of HDTV (high-definition television) and MPEG (video coding specification by the Moving Picture Experts Group), voice compression was beginning to achieve acceptable voice quality at rates around 8 kb/s, and first residential data access applications were appearing in the context of accessing the office computer and electronic bulletin boards. Consequently, telecommunications industry representatives came to the conclusion that a need for broadband services in the telecommunication network was imminent. Since ISDN was not capable of answering high-speed and packet-based service needs of such services, the concept of BISDN was deemed necessary. Aiding in this process was the availability of high-speed transmission, switching, and signal processing technologies. It became clear that even higher processing speeds would become available in the near future (e.g., the fact that the speed of processing doubles every 1.5 years, also known as *Moore's law*). CCITT considered these signs so important that the usual 4-year cycle of a study group to issue Recommendations was considered too long and an interim set of broadband ISDN (BISDN) Recommendations were

first issued in 1990. It should be emphasized at this point that for the telecommunications industry, and specifically for the service providers, the vision of B-ISDN involves the integration of voice, video, and data services *end-to-end* and with quality-of-service (QoS) guarantees.

1.2. ATM Fundamentals

The concept of asynchronous transfer mode (ATM) was first unveiled in an international meeting in 1987 by J. P. Coudreuse of CNET, France [9]. The basic goal of ATM was to define a networking technology around the basic idea of fast packet switching. In doing so, it was recognized that integration of services is desirable, but requires true packet switching in order to be effective and economical. Since new services were expected to operate at multimegabit rates, a fast packet-switching technology was desired. This implied a number of choices (made for simplification purposes):

- Fixed packet size (known as *cells*)
- Short packet size
- Highly simplified headers
- No explicit error protection
- No link flow control

Since ATM was an effort to define BISDN by telephone equipment vendors and service providers, voice was a major part of the BISDN effort from the onset. In fact, the decision on short cell size (53 bytes total, with a 48-byte payload) was made with considerations of echo cancellation for voice. For 64-kbps voice, the use of echo cancellation equipment becomes necessary if packetization delay is more than 32 bytes (4 ms). Although the public telephone network in the United States has echo cancellers installed, smaller European countries do not. To avoid echo cancellation equipment, European countries proposed that the payload for ATM be 32 bytes. The U.S. proposal was 64 bytes and 48 bytes were chosen as a compromise. The maximum tolerable overhead due to the header was considered 10%, and thus the 5-byte header was chosen.

1.3. ATM Protocol Reference Model

The protocol reference model for ATM is shown in Fig. 1. This model is different from that of ISDN. In this reference model, the ATM layer is common to all services. Its function is to provide packet (cell) transfer capabilities. The ATM adaptation layer (AAL) is service-dependent. The AAL maps higher-layer information into ATM cells. The protocol reference model makes reference to three separate planes:

- *User plane* — information transfer and related controls (flow and error control)
- *Control plane* — call control and connection control
- *Management plane* — management functions as a whole, coordination among all planes, and layer management

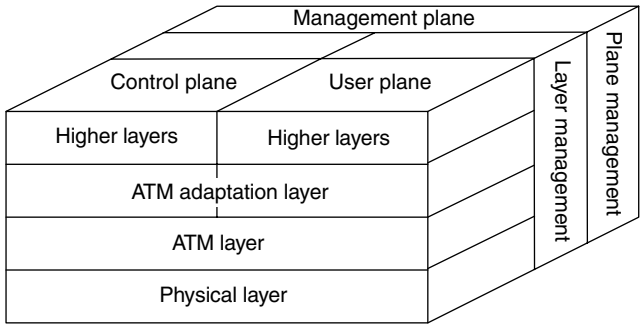


Figure 1. BISDN protocol reference model.

1.4. ATM Layer

We will first describe the ATM layer. ATM headers are very simple by design. The cell header has a different structure at the user–network interface (UNI) and at the network–network interface (NNI) (Figs. 2 and 3). Routing in ATM is achieved by an identifier field. It is the contents of this field that drives the fast hardware switching of an ATM cell. This field consists of two parts: the virtual circuit identifier (VCI) and the virtual path identifier (VPI). VCI is simply an index to a *connection* [14]. This "connection" is known as a *virtual circuit* (VC). A number of VCs are treated as a single entity known as a *virtual path* (VP). Thus, inside the network, cell switching can be performed on the basis of VPI alone. The VPI field is 8 bits at the UNI and 12 bits at the NNI. The VCI field is 16 bits long at both interfaces. It should be noted that VCIs and VPIs are not addresses. They are explicitly assigned at each segment (link between ATM nodes) of a connection when a connection is established, and they remain so for the duration of the connection. Using the VCI/VPI, the ATM layer can asynchronously interleave (multiplex) cells from multiple connections. As a historical remark, we would like to note that origins of the VPI/VCI concept can be traced back to the Datakit virtual circuit switch, developed by A. Fraser of Bell Labs during the 1970s [13,14]. Datakit was a product manufactured by AT&T for the data transmission needs of local exchange carriers.

HEC is an error check field, based on an 8-bit cyclic redundancy code (CRC), restricted to the cell header only. Three bits in the header (PT) are used to define the payload type. One bit (CLP) is reserved to indicate cell loss priority. This allows an ATM network to drop packets in case of congestion with the recovery mechanism provided by higher layers; such dropped cells will be detected by the higher layers of networking protocols (such as TCP/IP) and will be retransmitted. In passing, we would like to note that some earlier design decisions for ATM were later criticized when ATM was used to carry data belonging to the TCP/IP protocol. The most common type of IP packets carried in an IP network are TCP acknowledgment packets. Those packets are 44 bytes long and constitute about half of the packets carried in an IP network. Therefore, about half of IP packets are carried in an ATM network at an efficiency loss of about 10%. This inefficiency was later criticized by service providers in deploying IP-over-ATM networks and was termed *cell tax*.

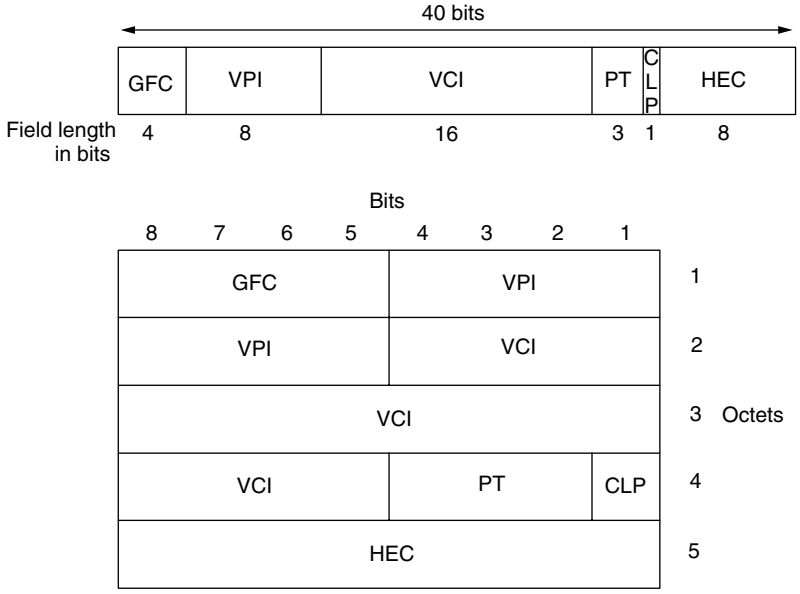


Figure 2. UNI cell header.

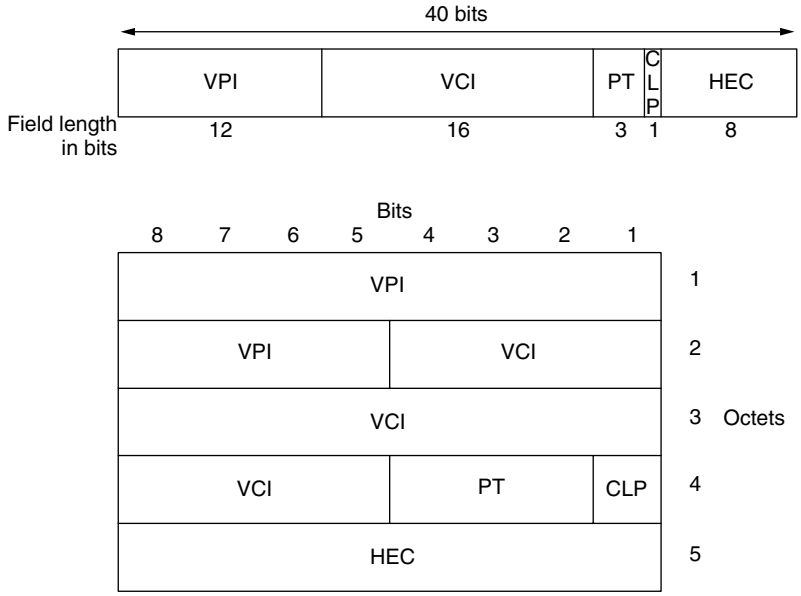


Figure 3. NNI cell header.

We stated above that VCI is an index to a connection. Thus, by this concept, ATM networks emulate connections between two points in a network and therefore are termed as *connection-oriented*. VP identifiers do not have to be universal in a network, they can be mapped from a set of values to another at the subnetwork boundary, albeit at a hardware cost. Virtual connections (consisting of VPs and VCs together) can be established permanently or on a per need basis. Permanent VCs (PVCs) are established once and all and are simple to work with. For bursty applications, switched VCs (SVCs) are designed. At a network node, SVCs can be established (added to the VC list) and removed from the VC list on a per need basis. Although this is a desirable property since not all connections in a network can be known in advance and the goal of developing the technology is indeed provided for bursty, unpredictable traffic, the hardware cost of incorporating this capability is high. In particular, this flexibility of being able to support bursty connections was later criticized since it limits the scalability of the ATM concept because of the difficulty of its implementation for high-speed backbone networks.

A PVC is not signaled by the endpoints. Both of the endpoint VC values are manually provisioned. The link-by-link route through the network is also manually provisioned. If any equipment fails, the PVC is down, unless the underlying physical network can reroute below ATM. A soft PVC also has manually provisioned endpoint VC values, but the route through the network can be automatically revised if there is a failure. Failure of a link causes a soft PVC to route around the outage and remain available. A switched VC (SVC) is established

by UNI signaling methods. So an SVC is a connection initiated by user demand. If a switch in the path fails, the SVC is broken and would have to be reconnected. The difference between an SVC and a soft PVC is that an SVC is established on an "as needed" basis through user signaling. With a soft PVC, the called party cannot drop the connection.

1.5. ATM Adaptation Layer

In order for ATM to support many kinds of services with different traffic characteristics and system requirements, it is necessary to adapt the different classes of applications to the ATM layer. This function is performed by AAL, which is service-dependent. Four types of AAL were originally recommended by CCITT. Two of these have now been merged into one, and a new one is added, making the total four once again. The four AALs are now described briefly:

- *AAL1* — supports connection-oriented services that require constant bit rates and have specific timing and delay requirements. Examples are constant bit rate services such as DS1 (1.5 Mbps) or DS3 (45 Mbps) transport.
- *AAL2* — a method for carrying voice-over ATM. It consists of variable size packets with a maximum of 64 bytes encapsulated within the ATM payload. AAL2 was previously known as "composite ATM" or "AAL-CU." The ITU specification, which describes AAL2 is called "ITU-T I.363.2."
- *AAL3/4* — a method intended for both connectionless and connection-oriented variable-bit-rate services. Two original distinct adaptation layers AAL3 and 4 have been merged into AAL3/4.
- *AAL5* — supports connection-oriented variable-bit-rate data services. Compared with AAL3/4, AAL5 is substantially lean at the expense of error recovery and built-in retransmission. This tradeoff provides a smaller bandwidth overhead, simpler processing requirements, and reduced implementation complexity. AAL5 has been proposed for use with both connection-oriented and connectionless services.

There is an additional AAL, AAL0, which normally refers to the case where the payload is directly inserted into a cell. This typically requires that the payload can always be fitted into a single cell so that the AAL is not needed for upper-layer PDU delineation when the upper-layer PDU bridges several cells.

1.6. ATM Traffic Management

During the early 1990s, the computer networking community was looking for a replacement of the 10-Mbps Ethernet standard at higher speeds of 100 Mbps and beyond. ATM, as specified by CCITT, was considered a viable alternative. Various companies active in this field formed an industry consortium, known as the *ATM Forum*. The ATM Forum later made quick progress in specifying and modifying the ATM specifications. ATM Forum defined the following traffic parameters for describing traffic that is injected into the ATM network at the UNI [2]:

- *Peak cell rate* (PCR) — maximum bit rate that may be transmitted from the source
- *Cell delay variation tolerance* (CDVT) — tolerance controlled by the network provider on how the actual peak rate deviates from the PCR
- *Sustainable cell rate* (SCR) — upper limit for the average cell rate that may be transmitted from the source
- *Maximum burst size* (MBS) — maximum number of cells for which the source may transmit at the PCR
- *Minimum cell rate* (MCR) — minimum cell rate guaranteed by the network

The ATM Forum defined different service classes to be supported by ATM. The classes are differentiated by specifying different values for the following QoS parameters defined on a per-connection basis:

- *Maximum Cell Transfer Delay* (*maxCTD*). CTD is the delay experienced by a cell between the transmission of the first bit by the source and the reception of the last bit of the cell by the destination. The CTD of a cell is smaller than the maxCTD QoS parameter of the connection with which it is carried within with a large probability, or equivalently, maxCTD is the $(1 - \alpha)$ quantile of CTD for a small α.
- *Peak-to-peak Cell Delay Variation* (*ppCDV*). The ppCDV is the difference between the $(1 - \alpha)$ quantile of the CTD and the fixed CTD that could be experienced by any delivered cell on a connection during the entire connection holding time.
- *Cell Loss Ratio* (*CLR*). This ratio indicates the percentage of cells that are lost in the network due to error or congestion and are not received by the destination.

The QoS parameters are defined for all conforming cells of a connection, where conformance is defined with respect to a generic cell rate algorithm (GCRA) described in the ATM Forum *User-Network Interface Specification 3.1* [2]. The input to this algorithm is the traffic parameters described above.

The proposed service categories by the ATM Forum are then described as follows [1]:

- *CBR* (*Constant Bit Rate*). The CBR service class is intended for real-time applications, namely, those requiring tightly constrained delay and delay variation, as would be appropriate for voice and video applications. The consistent availability of a fixed quantity of bandwidth is considered appropriate for CBR service. Cells that are delayed beyond the value specified by maxCTD are assumed to be of significantly less value to the application. For the service class CBR, the attributes PCR, CDVT, maxCTD, ppCDV, and CLR are specified.

- *VBR-rt* (*Variable Bit Rate — Real-Time*). The real-time VBR service class is intended for real-time applications, that is, those requiring minimal loss and tightly constrained delay and delay variation, as would be appropriate for voice and video applications. Sources are expected to transmit at a rate that varies with time. Equivalently, the source can be described as "bursty." VRB-rt expects a bound on the cell loss rate for cells conforming to the associated GCRA. Cells delayed beyond the value specified by maxCTD are assumed to be of significantly less value to the application. Real-time VBR service may support statistical multiplexing of real-time sources, or may provide a consistently guaranteed QoS. For VBR-rt, the ATM attributes PCR, CDVT, SCR, MBS, maxCTD, ppCDV, and CLR are specified.
- *VBR-nrt* (*Variable Bit Rate — Non-Real-Time*). The non-real-time VBR service class is intended for non-real-time applications that have "bursty" traffic characteristics and can be characterized in terms of a generic cell rate algorithm (GCRA). VRB-nrt expects a bound on the cell loss rate for cells conforming to the associated GCRA. Bounds for cell transfer delay and cell delay variation are not provided for VBR-nrt. Similar to VBR-rt, non-real-time VBR service also supports statistical multiplexing of connections. For non-real-time VBR, ATM attributes PCR, CDVT, SCR, MBS, and CLR are supported.
- *UBR* (*Unspecified Bit Rate*). The UBR service class is intended for delay-tolerant or non-real-time applications — those that do not require tightly constrained delay and delay variation, such as traditional computer communications applications. Sources are expected to transmit noncontinuous bursts of cells. UBR service supports a high degree of statistical multiplexing among sources. UBR service includes no notion of a per VC allocated bandwidth resource. Transport of cells in UBR service is not necessarily guaranteed by mechanisms operating at the cell level. However, it is expected that resources will be provisioned for UBR service in such a way as to make it usable for some set of applications. UBR service may be considered as interpretation of the common term "best-effort service." For UBR, only PCR and CDVT are specified as a traffic attribute.
- *ABR* (*Available Bit Rate*). Many applications have the ability to reduce their information transfer rate if the network requires them to do so. Likewise, they may wish to increase their information transfer rate if there is extra bandwidth available within the network. There may not be deterministic parameters because the users are willing to live with unreserved bandwidth. To support traffic from such sources in an ATM network will require facilities different from those for peak cell rate of sustainable cell rate traffic. The ABR service is designed to fill this need. The traffic attributes PCR, CDVT, MCR, and CLR are specified for the ABR service class.

There are other service categories proposed by the ITU: ABT (ATM block transfer) and CCT (controlled cell transfer). However, these categories have not gained much acceptance.

2. IP NETWORKS

In the 1990s, while ATM technology was being developed to integrate voice, data, and video, pure data services embraced the TCP/IP protocol, or the IP technology. What made the IP technology attractive is its universal adoption, due mainly to the popularity of the global Internet and the unprecedented growth rates the Internet has reached. Initially, IP was not designed for the integration of voice, data, and video to the end user. Developed under the U.S. Department of Defense (DoD) funding, IP was built around reliability and redundancy so as to allow communication to continue between nodes in case of a failure.

2.1. History of IP Networks

There was a perceived need for survivable command and control systems in the United States during the 1960s. To fulfill this need, early contributors were drawn from the ranks of defense contractors, federally funded think tanks, and universities: the RAND Corporation, Lincoln Laboratories, MIT, UCLA, and Bolt Beranek and Newman (BBN), under DoD funding.

P. Baran of RAND Corporation postulated many of the key concepts of packet-switching networks that were implemented in the ARPANET, the research network Advanced Research Projects Agency (ARPA) of DOD funded in 1967. Baran's motivation was to use novel approaches to build survivable communications systems. The traditional telephone system is based on a centralized switching architecture and the concept of connection or a "circuit" that must be established between the parties of a communications session using the centralized switches. If a link or a switch is broken (or destroyed) during a connection in this architecture, the communications session will fail, which is unacceptable for survivability purposes. Baran's work was built around the replacement of centralized switches with a larger number of distributed routers, each with multiple (potentially redundant) connections to adjacent routers. Messages then would be divided into parts (*blocks* or *packets*), and the packets would then be routed independently. This packet-switching concept allows bursty data traffic to be statistically multiplexed over available communications paths and makes it possible to adapt to changing traffic demands and to use existing resources more efficiently without a need for a priori reservation.

ARPANET was proposed by ARPA as an ambitious program to connect many host computers at key research sites across the country, using point-to-point telephone lines and the packet-switching concept. The idea of using separate switching computers, rather than the hosts, was proposed to serve as the routing elements of network, thereby offloading this function from the timesharing hosts. BBN received the contract to build the interface message processors (IMPs) in this newly

proposed architecture. The engineers at BBN developed the necessary host-to-IMP and IMP-to-IMP protocols, the original flow control algorithms, and the congestion control algorithms. In the BBN model, hosts communicate with each other via messages. When a host sends a message, it is broken down into packets by the source IMP (which is the IMP directly attached to the host). The IMP then routes each packet, individually through the network of IMPs, to the destination IMP. Each packet will be sent along the path that is estimated to be the shortest, and the path taken by each packet may be different. The destination IMP, on receiving all packets for a message, will reassemble an exact replica of the original message and forward the message on to the destination host. On the basis of the implementation of BBN, the ARPANET started to emerge with its first four nodes at UCLA, UCSB, Stanford Research Institute (SRI), and the University of Utah in 1969. ARPANET's purpose was to provide a fast and reliable communication between heterogeneous host machines. The goal of the computer network was for each computer to make every local resource available to any computer in the network in such a way that any program available to local users can be used remotely without much degradation.

In 1969, N. Abramson, motivated by the poor telephone lines in the Hawaiian Islands, launched the Aloha Project at the University of Hawaii, a project funded by ARPA. In this project, the principles underlying a packet-switched network based on fixed-site radio links were investigated. The ALOHA project developed a new technology for contention-based media access, the "ALOHA protocols," and applied these techniques to satellites as well as radio systems. R. Metcalfe at Xerox PARC built on this work, leading to the development of the Ethernet protocols for access to a shared wirelined medium as a local-area networking technology. In 1972, L. G. Roberts and R. Kahn launched the ARPA Packet Radio Program: packet switching techniques on the mobile battlefield. ARPA also created a packet-switched experimental satellite network (SATNet), with work done by Comsat, Linkabit, and BBN. All this work motivated the need for a technology to link these independent networks together in a true "network of networks," the so-called Internet.

In 1973, R. Kahn and V. Cerf developed the concept of a network gateway (or a software packet switch), as well as the initial specifications for the Transmission Control Protocol (TCP). With this breakthrough concept, transmission reliability is shifted from the network to end hosts, thus allowing the protocol to operate no matter how unreliable the underlying link is. This paradigm shift was based on the "end-to-end argument," which states that the underlying network is only as strong as its weakest link and therefore improving the reliability of a single link or even an entire subnetwork may have only a marginal effect on the end-to-end reliability. With this paradigm change, the architecture internal to the network was significantly simplified. V. Cerf then joined ARPA to complete the design of the Internet Protocol (IP) Suite, overseeing the separation of the routing portions of the protocols (IP) from the transport-layer issues (TCP), and the transition of the new protocols into ARPANET.

The global Internet began around 1980, when ARPA started converting machines attached to its research networks to the new TCP/IP protocols. ARPANET, already in place, quickly became the backbone of the new Internet and was used for many of the early experiments with TCP/IP. In 1983, the Defense Communications Agency (DCA) split ARPANET into two separate networks: one for future research and one for military communications, with the research part retaining the name ARPANET. At around the same time, most university computer science departments were running a version of the UNIX operating system available from the University of California's Berkeley software distribution, commonly known as Berkeley UNIX or BSD UNIX. By funding BBN to implement its TCP/IP protocols for use with BSD UNIX, and funding University of California Berkeley to integrate the protocols with its software distribution, ARPA was able to reach over 90% of university computer science departments in the United States. A large number of hosts subsequently connected their networks to ARPANET, thus creating the "ARPA Internet."

By 1985, ARPANET was heavily used and congested. The National Science Foundation (NSF), which needed a faster network, initiated the development of NSFNET in the mid-1980s. NSF selected the TCP/IP protocol suite used in ARPANET. However, as opposed to a single core backbone used in ARPANET, the earliest form of NSFNET in 1986 used a three-tiered architecture that consisted of universities and other research organizations that are connected to regional networks, which are then interconnected to a major backbone network using 56-kbps links. The link speeds were then upgraded to T1 (1.5 Mbps) in 1988 and later in 1991 to T3 (45 Mbps). In the early 1990s, the NSFNET was still reserved for research and educational applications. At this time, government agencies, commercial users, and the general public began demanding access to NSFNET. With the success of private networks using IP technology, NSF decided to decommission the NSFNET backbone in 1995. Commercial Internet providers then took over the role of providing Internet access. These providers have connection points called *point of presence* (PoP). Customers of these service providers are connected to the Internet via these PoPs. The collection of PoPs and the way they are interconnected form the provider's network. Providers may be regional, national, or global, depending on the scope of their networks. Today's Internet architecture is based on a distributed architecture operated by multiple commercial providers rather than a single core network (NSFNET) that are interconnected via major network exchange points. Historically, commercial Internet providers exchange traffic at network access points (NAP) and the metropolitan-area exchanges (MAEs) through a free exchange relationship called *bilateral public peering*. Two connectivity models have emerged as a result of increasing congestion in the major exchange points: (1) private peering among the largest backbone providers and (2) more recently, private transit

connections to multiple backbone providers, which are favored by specialized ISPs.

2.2. IP Fundamentals

The Internet provides three sets of services [8]. At the lowest level, one has a connectionless delivery service. The other two services (transport services and application services) lie on top of this connectionless delivery service. The protocol that defines the unreliable, connectionless delivery mechanism is called the Internet Protocol and is commonly referred to by its initials, IP. IP defines the basic data unit of data transfer and it also performs the routing function. Therefore, IP is also referred to as the *layer 3 protocol* in the Internet suite as it corresponds to the layer 3 (network layer) of the OSI model. Layer 4 protocols such as TCP and UDP run on IP and provide an appropriate higher level platform on which the applications depend.

In addition to internetwork routing, IP provides error reporting and fragmentation and reassembly of information units called *datagrams*. Datagrams of different size are used by IP for transmission over networks with different maximum data unit sizes. IP addresses are globally unique, 32-bit numbers assigned by the network information center. Globally unique addresses permit IP networks anywhere in the world to communicate with each other.

An IP address is divided into three parts. The first part designates the network address, the second part designates the subnet address, and the third part designates the host address. Originally IP addressing supported three different network classes. Class A networks were intended mainly for use with a few very large networks, because they provided only 8 bits for the network address field. Class B networks allocated 16 bits, and class C networks allocated 24 bits for the network address field. Because Internet addresses were generally assigned only in these three sizes, there were many wasted addresses. In the early 1990s only 3% of the assigned addresses were actually being used and the Internet was running out of unassigned addresses. A related problem was the size of the Internet global routing tables. As the number of networks on the Internet increased, so did the number of entries in the routing tables. By this time, Internet standards were being specified by an organization known as the Internet Engineering Task Force (IETF). IETF selected classless interdomain routing (CIDR) [15,24] to be a much more efficient method of assigning addresses and address aggregation to address these two critical issues. CIDR is a replacement for the old process of assigning class A, B, and C addresses with a generalized network prefix. Instead of being limited to network identifiers (or "prefixes") of 8, 16, or 24 bits, CIDR currently uses prefixes anywhere from 13 to 27 bits. This allows for address assignments that much more closely fit an organization's specific needs and therefore avoids address waste. The CIDR addressing scheme also enables route aggregation in which a single high-level route entry can represent many lower-level routes in the global routing tables.

In the 1990s, there were also significant developments in IP routing. There are two main routing infrastructures: flat routing and hierarchical routing. In a flat routing infrastructure, each network ID is represented individually in the routing table. The network IDs have no network/subnet structure and cannot be summarized. In a hierarchical routing infrastructure, groups of network IDs can be represented as a single routing table entry through route summarization. The network IDs in a hierarchical internetwork have a network/subnet/subsubnet structure. A routing table entry for the highest level (the network) is also the route used for the subnets and sub-subnets of the network. Hierarchical routing infrastructures simplify routing tables and lower the amount of routing information that is exchanged, but they require more planning. IP implements hierarchical network addressing, and IP internetworks can have a hierarchical routing structure.

In very large internetworks, it is necessary to divide the internetwork into separate entities known as *autonomous systems*. An autonomous system (AS) is a portion of the internetwork under the same administrative authority. The AS may be further divided into regions, domains, or areas that define a hierarchy within the AS. The protocols used to distribute routing information within an AS are known as *interior gateway protocols* (IGPs). The protocols used to distribute routing information between ASs are known as *exterior gateway protocols* (EGPs). In today's Internet, link-state protocols such as OSPF version 2 [21] and IS-IS [23] are used as IGPs, whereas the path vector protocol BGP-4 [25] is used as a exterior gateway protocol.

With the changes to IP address structure and address summarization with CIDR, and the development of efficient hierarchical routing infrastructures, IP networks have scaled up to the level of universal connectivity today. This has made the Internet a global medium in such a way that any two hosts can communicate with each other as long as they are attached to the Internet. However, currently a packet is transported in the Internet without any guarantees to its delay or loss. Because of this "best effort" forwarding paradigm, the Internet cannot provide integrated services over this infrastructure. As we described previously, the BISDN vision requires end-to-end QoS guarantees for different services. The IETF is working on several QoS models that may potentially realize the BISDN vision using IP. Using IP as opposed to ATM to realize the BISDN vision is a new approach made popular by the widespread use of IP.

2.3. QoS Models in IP Networks

Several QoS architectures are proposed by the IETF for IP networks to enable the support of integrated services over IP networks. We will briefly overview these models below.

2.3.1. Integrated Services (Intserv) Model. The integrated services architecture [6] defines a set of extensions to the traditional best-effort model of the Internet so as to provide end-to-end (E2E) QoS commitments to certain applications with quantitative performance requirements. Two services are defined: guaranteed service [28] and controlled load [31] services. Guaranteed service provides an assured level of bandwidth, a firm end-to-end delay bound,

and no loss due to queueing if the packets conform to an a priori negotiated contract. It is intended for applications with stringent real-time delivery requirements such as audio and video applications with playback buffers. A packet arriving after its playback time is simply discarded by the receiver. In the case of controlled load service, the network will commit to a flow a service equivalent to that seen by a best-effort flow on a lightly loaded network. This service is intended for adaptive real-time applications that can tolerate a certain amount of loss and delay provided it is kept to a reasonable level. The integrated services architecture assumes some explicit setup mechanism such as RSVP (Resource Reservation Protocol) [7]. This setup or signaling mechanism will be used to convey QoS requirements to IP routers so that they can provide requested services to flows that request them. On receiving per flow resource requirements through RSVP, the routers apply Intserv admission control to signaled requests. The routers also enable traffic control mechanisms to ensure that each admitted flow receives the requested service independent of other flows. These mechanisms include the maintenance of per flow classification and scheduling states. One impediment to the deployment of integrated services with RSVP is the use of per flow state and per flow processing, which typically exceeds the flow-handling capability of today's core routers. This is known as the *scalability problem* in RSVP or in Intserv.

The integrated services architecture is similar to the ATM SVC architecture in which ATM signaling is used to route a single call over an SVC that provides the QoS commitments of the associated call. The fundamental difference between the two architectures is that the former typically uses the traditional hop-by-hop IP routing paradigm, whereas the latter uses the more sophisticated QoS source routing paradigm.

2.3.2. Aggregate RSVP Reservations Model. This QoS model attempts to address some of the scalability issues arising in the traditional Intserv model. In the traditional intserv model, each E2E reservation requires a significant amount of message exchange, computation, and memory resources in each router along the way. Reducing this burden to a more manageable level via the aggregation of E2E reservations into one single aggregate reservation is addressed by the IETF [3]. Although aggregation reduces the level of isolation between individual flows belonging to the aggregate, there is evidence that it may potentially have a positive impact on delay distributions if used properly and aggregation is required for scalability purposes.

In the aggregation of E2E reservations, we have an aggregator router, an aggregation region, and a deaggregator. Aggregation is based on hiding the E2E RSVP messages from RSVP-capable routers inside the aggregation region. To achieve this, the IP protocol number in the E2E reservation's Path, PathTear, and ResvConf messages is changed by the aggregator router from RSVP to RSVP-E2E-IGNORE on entering the aggregation region, and restored to RSVP at the deaggregator point. Such messages are treated as normal IP datagrams inside the aggregation region, and no state is stored. Aggregate Path messages are sent from the aggregator to the deaggregator using RSVP's normal IP protocol number. Aggregate Resv messages are then sent back from the deaggregator to the aggregator, via which an aggregate reservation with some suitable capacity will be established between the aggregator and the deaggregator to carry the E2E flows that share the reservation. Such establishment of a smaller number of aggregate reservations on behalf of a larger number of E2E flows leads to a significant reduction in the amount of state to be stored and the amount of signaling messages exchanged in the aggregation region.

Aggregation of RSVP reservations in IP networks is very similar in concept to the virtual path in ATM networks. In this framework, several ATM virtual circuits can be tunneled into one single ATM VP for manageability and scalability purposes. A virtual path identifier (VPI) in the ATM cell header is used to classify the aggregate in the aggregation region (VP switches), and the virtual channel identifier (VCI) is used for aggregation/deaggregation purposes. A VP can be resized through signaling or management.

2.3.3. Differentiated Services (Diffserv). In contrast to the per flow nature of integrated services, differentiated services (Diffserv) networks classify packets into one of a small number of aggregated flows or "classes" based on the Diffserv Codepoint (DSCP) written in the differentiated services field of the packet's IP header [22]. This is known as *behavior aggregate* (BA) classification. At each Diffserv router in a Diffserv domain (DS domain), packets receive a per hop behavior (PHB), which is invoked by the DSCP. Differentiated services are extended across a DS domain boundary by establishing a service-level agreement (SLA) between an upstream network and a downstream DS domain. Traffic classification and conditioning functions (metering, shaping, policing, remarking) are performed at this boundary to ensure that traffic entering the DS domain conforms to the rules specified in the *traffic conditioning agreement* (TCA), which is derived from the SLA. A PHB then refers to the packet scheduling, queueing, policing, or shaping behavior of a node on any given packet belonging to a BA, as configured by a SLA or a policy decision. Four standard PHBs are defined:

- *Default PHB* [22] — provides a best-effort service in a Diffserv-compliant node.
- *Class-selector PHB* [22] — to preserve backward compatibility with any IP precedence scheme currently in use on the network, Diffserv defines a certain DSCP for class selector code points. The PHB associated with a class selector code point is a class selector PHB. Eight class selector code points are defined.
- *Assured forwarding* (*AF*) *PHB* [16] — the AF PHB group defines four AF classes: AF1, AF2, AF3, and AF4. Each class is assigned a specific amount of buffer space and interface bandwidth, according to the SLA with the service provider or a policy decision. Within each AF class, three drop precedence values are assigned. In the case of a congestion indication or

equivalently if the queue occupancy for the AF class exceeds a certain threshold, packets in that class with lower drop precedence values will be dropped. With this description, assured forwarding PHB is similar to the controlled load service available in the integrated services model.

- *Expedited forwarding* (*EF*) *PHB* [10] — EF PHB provides a guaranteed bandwidth service with low loss, delay, and delay jitter. EF PHB can be implemented with priority queuing and rate limiting on the behavior aggregate. EF PHB can be used to provide virtual leased line or premium services in Diffserv networks similar to the guaranteed service in Intserv networks and the CBR service in ATM networks.

Since Diffserv eliminates the need for per flow state and per flow processing, it scales well to large-core networks.

2.3.4. Hybrid Intserv–Diffserv [5]. In this QoS model, intserv and diffserv are employed together in a way that end-to-end, quantitative QoS is provided by applying the Intserv model end- to- end across a network containing one or more Diffserv regions. The Diffserv regions of the network appear to the Intserv-capable routers or hosts as virtual links. Within the Diffserv regions of the network, routers implement specific PHBs (aggregate traffic control) on the basis of policy decisions. For example, one of the AF PHBs can be used to carry all traffic using E2E reservations once an appropriate amount of bandwidth and buffer space is allocated for that AF class at each node. The total amount of traffic that is admitted into the Diffserv region that will receive a certain PHB may be limited by policing at the edge. The primary benefit of Diffserv aggregate traffic control is its scalability. The hybrid Intserv–Diffserv model is closely related to the RSVP reservation aggregation model.

2.3.5. Multiprotocol Label Switching (MPLS). MPLS introduces a new forwarding paradigm for IP networks in that a path is first established using a signaling protocol. A label in the IP header rather than the destination IP address is used for making forwarding decisions throughout the MPLS domain [26]. Such paths are called *label-switched paths* (LSPs), and routers that support MPLS are called *label-switched routers* (LSRs). In this architecture, edge ingress LSRs place IP packets belonging to a certain forwarding equivalence class (FEC) in an appropriate LSP. The core LSRs forward packets only according to the label in the header, and the egress edge LSRs remove the labels and forward these packets as regular IP packets. The benefits of this architecture include but are not limited to

- *Hierarchical Forwarding*. MPLS provides a forwarding hierarchy with arbitrary levels as opposed to the two-level hierarchy in ATM networks. Using this flexibility and the notion of nested labels, several level 1 LSPs can be aggregated into one level 2 LSP, and several level 2 LSPs can be aggregated into one level 3 LSP, and so on. One immediate benefit of this is that the transit provider need not know about the global routes, which makes it very scalable [11] for transit providers.
- *Traffic Engineering*. The mapping of traffic trunks (an aggregation of traffic belonging to the same class) onto a given network topology for optimal use of network resources is called the *traffic engineering problem*. In MPLS networks, traffic trunks are mapped to the network topology through the selection of routes and by establishing LSPs with certain attributes using these routes. A combination of a traffic trunk and the LSP is called an *LSP tunnel*. In its simplest application, in the case of congestion arising from suboptimal routing, LSP tunnels may be rerouted for better performance.
- *Virtual Circuit Emulation*. Another benefit is that other connection-oriented networks may be emulated by MPLS. The advantage is that a single integrated datagram network can provide legacy services such as frame relay and ATM to end customers while maintaining a single infrastructure.

2.3.6. Summary of QoS Models for IP Networks. For elastic applications that can adapt their rates to changing network conditions (e.g., data applications using TCP), a simple QoS model such as "Diffserv" will be suitable. For inelastic applications such as real-time voice and video with stringent delay and loss requirements, end-to-end Intserv is a better fit. The need for per flow maintenance in RSVP capable routers is known to lead to a scalability problem especially in core networks. Therefore, several novel QoS models have been introduced to attack this scalability problem. From the perspective of a network, both models rely on eliminating the per flow maintenance requirement by either aggregating E2E reservations into one single reservation at the border nodes of this network or carrying all E2E reservations in one preprovisioned Diffserv class. However, these architectures pose a burden on the border routers, and their success remains to be seen in the commercial marketplace. MPLS, on the other hand, is promising traffic-engineered backbones with routing scalability for all these QoS models.

3. BISDN AND THE WORLD WIDE WEB

In this section we describe the development of IP versus ATM as the underlying networking technology of BISDN.

The development of ATM realized full progress at ITU-T during 1989/90. This effort was led by telecommunications service providers as well as telecommunications equipment manufacturers. The main goal was to develop the switching and networking technology for BISDN. As cooperation and contributions from telecommunications industry leaders were at a very substantial level, the vision of an integrated wide-area network (WAN) using ATM seemed very likely to happen. This development in the WAN sparked interest in other networking platforms. The first affected was the computer communications industry, specifically the local-area networking (LAN) community. At the time, available LANs (mainly the Ethernet) had a

top speed of 10 Mbps. The technology had improved from coax to twisted pair and from shared media to switched (1991). However, as user needs increased, the top speed of 10 Mbps became insufficient and the industry began to search for a replacement at significantly higher speeds of 100–150 Mbps. At this time, the ATM effort at ITU-T defined a basic transport rate of 155 Mbps. This speed was very convenient for the LAN community. In addition, adopting the same switching and networking technology with the WAN was attractive from the viewpoint of simplifying the WAN gateway. This led to an industry standardization organization known as the *ATM Forum*. The goal was to define a set of specifications common to the member companies, primarily for the LAN. An additional goal was to speed up the standardization process, which, at ITU-T, required long study periods and consensus from national representatives.

Another development related to ATM was the emergence of the ADSL (Asymmetric Digital Subscriber Line) technology in the 1990s [19]. At the time, invoking Shannon's capacity formula, the highest transmission rate for a voiceband modem over a subscriber loop, without changing any equipment at the central office, was considered to be about 30 kbps. The ADSL technology replaces central office channel banks to exploit frequencies above 4 kHz. In addition, it employs sophisticated methods that limit near-end crosstalk and therefore substantially expand the transmission potential of the subscriber loop. As a result, it can operate at rates that are orders of magnitude higher than those of voiceband modems. The ADSL access network includes terminations both within the home and the public network (ATU-R and ATU-C, respectively). The ATU-R is commonly called a "DSL modem," and the ATU-C is commonly called a "DSLAM" (DSL access multiplexer). ATM is used as layer 2 in this "residential broadband" architecture. ADSL provides up to 1.5 Mbps (downlink) rate. It may be used to extend the ATM network, and therefore QoS properties of ATM, all the way to the residential or corporate desktop. In this model [19], the ATM user-to-network interface (UNI) is tunneled through an ADSL link. By having desktop applications talk directly to the ATM network, bandwidth can be allocated end-to- end across the network that was thought to facilitate the deployment of isochronous, delay-sensitive applications such as voice and videoconferencing [17]. In fact, this was the intent of BISDN from the onset. The effort to employ ADSL to provide integrated services to the home was led by potential application service providers [20]. At this time, PC (personal computer) operating systems did not yet include a networking stack as part of the kernel, and beyond computer terminal emulation, there were not yet any major residential networking applications available. With the arrival of the World Wide Web (WWW) and the concept of a Web browser, the need for an IP stack in PCs became apparent. At the time the most popular PC operating system was Windows version 3.1 from Microsoft. As this operating system did not have an IP stack, it was added to the operating system manually by the user. Later, Windows 95 became the first PC operating system to include an IP stack. With this development, the IP stack became an inevitable option in residential broadband networking. Consequently, the original concept of residential ATM was later modified as IP over ATM over DSL [20]. This could have been a cosmetic change, however, and by this time, the vision of BISDN using ATM still seemed likely to happen, with a form of IP over ATM being used mainly for best effort data transfer.

A number of developments that took place in the second part of the 1990s have changed the outlook for ATM as the underlying networking technology of BISDN:

1. IEEE 802.3 Working Group made rapid progress to define a newer version of the Ethernet standard to operate at 100 Mbps over twisted-pair and switched media. This LAN standard did not have any QoS guarantees, but the solution satisfied a much sought-after need for a LAN operating around 100 Mbps. This solution was quickly adopted by the marketplace, and the 100-Mbps Ethernet quickly became a commodity product. The absence of a compelling need for QoS in LANs virtually stopped the local ATM activity. With this development, the ATM Forum lost a major thrust.
2. The development of the WWW and the Web browser, as well as the commercialization of the Internet quickly made Web browsing using a PC a household activity. This development stalled or perhaps even stopped the concept of residential ATM.
3. A possible application of ATM was in digital cable access systems. By making extensions to the coaxial or hybrid fiber/coaxial cable TV plant so that duplex transmission becomes possible (providing amplification in both directions), and using digital technology so that compression can be used to transmit hundreds of TV channels, ATM was under consideration as a potential service offering. Adding data services to this potential offering was attractive. A multiaccess control algorithm was needed to share the uplink channel. A standardization activity was initiated under the IEEE 802.14 Working Group. While this group was working on an access system based on ATM and provide QoS guarantees for delivering a multitude of services, and while some progress was made, cable service providers decided to pursue their own standardization effort. They named this activity the *multimedia cable network system* (MCNS). The main reason for this secession was to make the process of standardization faster. As PC operating systems were beginning to offer IP stacks, MCNS chose IP technology as the basis of their own access system. The resulting system specification is known as the *Data over Cable Service Interface Specifications* (DOCSIS). Although version 1.0 of these specifications was for best-effort data service only, in its version 1.1, DOCSIS supports some QoS guarantees, specifically designed for voice over IP (VoIP). DOCSIS is currently the de facto worldwide standard for digital cable access, while IEEE 802.14 has stopped its activities. With this development, IP, rather than ATM became the underlying technology for digital cable access systems.
4. A significant advantage of ATM was its fast switching property. ATM was designed to be a simple switching technology so that scalable switches at total

throughput values approaching hundreds of gigabits per second could be built. This vision is by and large correct (although segmentation and reassembly at edge routers can become difficult at higher speeds). However, there was a surprising development in this period — throughput values of routers increased substantially. Today the maximum throughput values of core IP routers compete with those of core ATM switches. Implementation of algorithms for IP address lookup and memory manipulation for variable-length packet switching in ASICs is largely responsible for this development.

5. The ATM Forum was founded as an industry organization with the premise of fast standardization. As we noted above, ITU-T requires long study periods and consensus among national representatives. It was thought that the ATM Forum would move faster in reaching a standard. Although that was partially achieved, the industry perception is that signaling became too complex in the ATM Forum.

6. In the 1990s, a number of developments took place in optical transport systems that altered network switching in a major way: (a) invention of Erbium-doped fiber amplification made long-distance optical transmission without intermediate electrical conversions possible; (b) development of wavelength-division multiplexing (WDM) or dense WDM (DWDM) made transport of a large number of wavelengths in a single fiber possible — the number of wavelengths approached >100, while transmission speeds on individual wavelengths approached 10 Gbps; and (3) wavelength routing or wavelength cross-connects made it possible to demultiplex individual wavelengths from a single fiber and multiplex wavelengths from different fibers into a single fiber. The result is a wavelength switch with total throughput in the range of tens of terabits per second. As a result, wavelength routing provided an alternative to electronic switching at the network core, thus making the scalability argument of ATM switching less attractive.

7. A major advantage of ATM was its QoS capabilities. However, as described in the previous section, IP community developed a set of QoS capabilities. Although there are questions and uncertainties about the realization of these capabilities, there is some established confidence in IP QoS. We would like to note that ATM actually was never deployed for the end-to-end QoS vision. The reason for this is the complexity in signaling and the needed per flow queuing. The multiclass and aggregate IP QoS model may indeed be more scalable.

8. IP embraced ATM's VP concept. MPLS essentially implements VPs. Various tunneling mechanisms introduced into IP make switching aggregated traffic in IP possible. Furthermore, the endpoints of a VP implemented by MPLS do not need to be routing peers, which significantly reduces the number of peerings in the network, and therefore routing scalability.

9. Because of its scalable fast switching nature, ATM switches were used to carry and switch IP traffic. However, over time, other solutions were developed that avoid the ATM layer in between. For example, at one point, service providers deployed IP over ATM over SONET over WDM. IP was employed since applications required it, ATM was employed for high-speed packet switching, SONET was employed because of its fast restoration capability via SONET rings, and WDM was employed for higher transmission speeds in a single fiber. The industry sought for ways to simplify this complicated hierarchy. As a result, IP extended to assume many of the functionalities of ATM and even some of those of SONET (e.g., resilient packet rings).

10. We described the 10% inefficiency that results in carrying TCP/IP traffic over ATM, known as the *cell tax*, above. Several service providers claimed this inefficiency was too high. In reality, with IP extending to assume many of ATM's functionalities, the need for IP over ATM was alleviated and the cell tax became irrelevant.

11. In the 1980s there were several attempts made to build private networks for multiple-location enterprises. These typically employed nailed-up leased lines, used voice compression to reduce voice rates, and combined voice and data. Such networks, called *private networks*, were the precursors of integration of services, albeit on a small scale. As discussed above, first ISDN and then BISDN had the vision of integrated services. In an integrated public packet network, security, by means of proper authentication and encryption, enables construction of a virtual private network (VPN). A VP is very useful in the construction of a VPN since it simplifies processing of data belonging to a particular VPN in the network. Thus ATM is a natural way to implement VPNs. However, as described above, solutions were developed to embrace the same concept in the IP world. Examples of such protocols are the Layer 2 Tunneling Protocol (L2TP) [30], IPSec [18], and GRE [12]. MPLS, on the other hand, makes it possible to build provider-provisioned scalable VPNs also making use of BGP4 for routing and label information distribution [27]. Thus ATM is no longer a unique method to implement VPNs.

12. Another aspect of the aggregation property of VPs is the traffic engineering potential it provides. For example, one possibility in integrated networks is to use different routes based on QoS properties of different flows, such as those belonging to the same source and destination pair. There are tools, such as the concept of equivalent bandwidth, that enable traffic engineering for integrated networks. Then, VPs become very useful tools to implement the desired property. Obviously, with the development of VP-like concepts in IP networks, the superiority of ATM in this regard is no longer valid.

To summarize, from the discussion above it appears that two related events stalled the development of BISDN:

1. The appearance of the WWW made IP protocol instantly ubiquitous. Common PC operating systems quickly adopted an IP stack. A similar ATM stack was not needed because there was no immediate application tied to ATM in the way that the WWW was tied to IP.
2. IP quickly extended to assume the advantageous properties of ATM, at least in theory. As a result, ATM lost its role as the underlying technology that glues BISDN all together.

Therefore, it is safe to say that BISDN is not likely to occur as it was originally designed at the ITU-T, frequently described by the acronym BISDN/ATM.

Having said that, we must reiterate that integration of services is certainly useful for the consumer. Furthermore, there appears to be an increasing (albeit at a smaller rate than expected) demand for broadband services. Thus, in the near future some form of a service offering that unifies voice, broadband data, and video can be expected (in fact, it currently exists in digital cable). Whether this offering will eventually become a ubiquitous service such as expected of BISDN/ATM depends on many factors and it is difficult to predict today (in mid-2002). It is clear, however, that voice-over IP (VoIP) will be used to carry some voice traffic, especially in traffic-engineered enterprise networks. The degree of voice compression available for VoIP ($\sim$8 vs. 64 kbps, although with VoIP overhead, this ratio of $\frac{1}{8}$ becomes bigger), statistical multiplexing advantages, and the capability to combine with data in VPNs is an attractive value proposition. Adding the Public Switched Telephone Network and video services to this value proposition successfully in the marketplace in the short term, however, is a taller order.

BIOGRAPHIES

Ender Ayanoglu received his B.S. degree in 1980 from the Middle East Technical University and M.S. and Ph.D degrees in 1982 and 1986, respectively, from Stanford University, all in electrical engineering. He was with the Communications Sciences Research Laboratory of Bell Laboratories (AT&T and Lucent Technologies) during 1986–1999. From 1999 to summer 2002 he was with Cisco Systems. Currently he is with University of California, Irvine, he was the Chairman of the IEEE Communications Society Communication Theory Technical Committee during 1999–2001. He was the Chairman of the IEEE-ISTO Broadband Wireless Internet Forum during 2000/01. Currently he serves as an Editor for the *IEEE Transactions on Communications*. He is the recipient of two best paper awards from the IEEE Communications Society and is an IEEE Fellow.

Nail Akar received the B.S degree in 1987 from Middle East Technical University, Ankara, Turkey, and the M.S. and Ph.D degrees from Bilkent University, Ankara, Turkey, in 1989 and 1994, respectively, all in electrical and electronics engineering. He joined the Computer Science Telecommunications Program in 1994 at the University of Missouri — Kansas City as a Visiting Scholar and was a Visiting Assistant Professor in the same program in 1996. At UMKC, he worked on the development of computational algorithms for the performance analysis of communication networks. Dr. Akar joined the Technology Planning and Integration group at Sprint Long Distance Division in 1996 and was a senior member of technical staff in the same group in 1998–2000. While at Sprint, he worked on ATM traffic management and routing, IP Qos, virtual private networking architectures, and pricing. Since 2000, he has been with the Electrical and Electronics Engineering Department, Bilkent University, Turkey as an assistant professor. His areas of interest include quality of service in IP networks, network design and engineering, and queueing systems.

BIBLIOGRAPHY

1. ATM Forum, *ATM Forum Traffic Management Specification Version 4.0*, 1996.
2. ATM Forum, *ATM User-Network Interface Specification Version 3.1*, 1994.
3. F. Baker, C. Iturralde, F. L. Faucheur, and B. Davie, *Aggregation of RSVP for IPv4 and IPv6 Reservations*, RFC 3175, 2001.
4. J. C. Bellamy, *Digital Telephony*, Wiley, New York, 1991.
5. Y. Bernet et al., *A Framework for Integrated Services Operation over Diffserv Networks*, RFC 2998, 2000.
6. R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, 1994.
7. R. Braden et al., *Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification*, RFC 2205, 1997.
8. D. E. Comer, *Internetworking with TCP/IP: Principles, Protocols, and Architectures*, Prentice-Hall, 2000.
9. J. P. Coudreuse and M. Servel, Prelude: An asynchronous time-division switched network, *ICC Proc.*, Seattle, 1987.
10. B. Davie et al., *An Expedited Forwarding PHB (Per-Hop Behavior)*, RFC 3246, 2002.
11. B. Davie and Y. Rekhter, *MPLS Technology and Applications*, Academic Press, 2000.
12. D. Farinacci et al., *Generic Routing Encapsulation (GRE)*, RFC 2784, 2000.
13. A. G. Fraser, Early experiments with asynchronous time division networks, *IEEE Network* **7**: 12–26 (1993).
14. A. G. Fraser, Towards a Universal Data Transport System, *IEEE J. Select. Areas Commun.* **1**: 803–815 (1983).
15. V. Fuller, T. Li, J. Yu, and K. Varadhan, *Classless Inter-Domain Routing (CIDR): An Address Assignment and Aggregation Strategy*, RFC 1518, 1993.
16. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, *Assured Forwarding PHB Group*, RFC 2597, 1999.
17. M. Humphrey and J. Freeman, How xDSL supports broadband services to the home, *IEEE Network* **11** 14–23 (1997).
18. S. Kent and R. Atkinson, *Security Architecture for the Internet Protocol*, RFC 2401, 1998.
19. T. Kwok, *ATM: The New Paradigm for Internet, Intranet & Residential Broadband Services and Applications*, Prentice-Hall, 1998.
20. T. Kwok, A vision for residential broadband services: ATM-to-the-home, *IEEE Network* 14–28 (1995).
21. J. Moy, *OSPF Version 2*, RFC 2178, 1997.
22. K. Nichols, S. Blake, F. Baker, and D. Black, *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers*, RFC 2474, 1998.
23. D. Oran, *OSI IS-IS Intra-domain Routing Protocol*, RFC 1142, 1990.
24. Y. Rekhter and T. Li, *An Architecture for IP Address Allocation with CIDR*, RFC 1518, 1993.

25. Y. Rekhter and T. Li, *A Border Gateway Protocol 4* (*BGP-4*), RFC 1771, 1995.

26. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, 2001.

27. E. C. Rosen, BGP/MPLS VPNs, *<draft-ietf-ppvpn-rfc2547bis-01.txt>*, 2002.

28. S. Shenker, C. Partridge, and R. Guerin, *Specification of Guaranteed Quality of Service*, RFC 2212, 1997.

29. W. Stallings, *ISDN and Broadband ISDN*, Macmillan, New York, 1992.

30. W. Townsley et al., *Layer Two Tunneling Protocol "L2TP,"* RFC 2661, 1999.

31. J. Wroclawski, *Specification of the Controlled-Load Network Element Service*, RFC 2212, 1997.

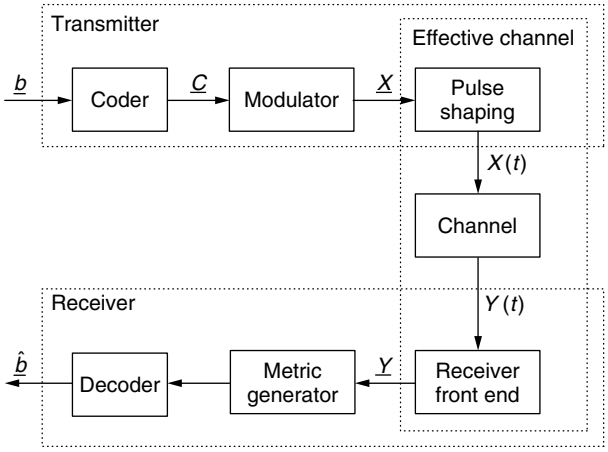


Figure 1. Block diagram of a communication system for conveying the information sequence $\underline{b}$ across a communication channel.

BIT-INTERLEAVED CODED MODULATION

DENNIS L. GOECKEL
University of Massachusetts
Amherst, Massachusetts

1. INTRODUCTION

Bit-interleaved coded modulation (BICM) has emerged as a promising method for transmitting information robustly over many types of communication channels; in particular, BICM has proven to be particularly attractive for the types of channels often found in wireless communication systems. The modern version of BICM is attributed to a 1992 paper of Zehavi [1], but it was the significant investigation of Caire and colleagues published in 1998 [2] that led to its widespread popularity. BICM marks a significant departure from the trend set in coded modulation roughly from 1978 to 1998 [2] and represents a return, at least structurally if not philosophically, to the types of coded modulation employed prior to 1980, which could be decidably suboptimal for the communication system applications to which they were applied at that time.

The goal of the error control coding and modulation, which are often referred to as one unit with the term "coded modulation," are to efficiently convey a sequence of information bits $\underline{b} = (b_0, b_1, b_2, \ldots)$ reliably across a channel, where the channel is defined as the physical entity connecting the transmitter to the receiver as shown in Fig. 1. Although the channel generally accepts the waveform $X(t)$ and produces the waveform $Y(t)$, the coded modulation is generally designed for an effective channel, which includes the transmitter pulseshaping, channel, and the sampled receiver front end. The input to this effective channel is the sequence of complex values $\underline{X}$, and its output is the sequence of complex values $\underline{Y}$. In 1948, Claude Shannon published his seminal work on information theory [3], which introduced the notion of capacity — the maximum rate (in bits per symbol) at which information can be reliably transmitted across this effective channel. Coded modulation strives to obtain this limit in a practical manner.

The ability of a practical coded modulation scheme to operate over the effective channel is often measured by the bit error rate (BER), which is a long-term average of the fraction of bits in the decoder output $\hat{\underline{b}}$ that do not agree with the transmitted sequence $\underline{b}$. Thus, a good coded modulation scheme will assign sequences of transmitted symbols $\underline{X}$ to information bit sequences in such a way that the decoder is able to ascertain which information bit sequence was transmitted with very high probability from the received sequence $\underline{Y}$. If the coding and modulation is performed separately, as is portrayed in Fig. 1, the information bit stream $\underline{b}$ is encoded to produce a coded bit stream $\underline{C}$ that contains carefully introduced redundancy in order to protect against errors that may be introduced by the channel. These coded bits are then taken individually or in groups by the modulator to choose from a number of possible complex values (termed the "constellation of possible signal points") to determine each entry of the sequence $\underline{X}$.

The field of coded modulation has progressed rapidly since the early 1970s. Prior to the late 1970s, error control coding and modulation were considered separable, as has been represented in Fig. 1. At that time, the channel type most often considered in the communication community was the additive white Gaussian noise (AWGN) channel, for which the received signal is that which was transmitted plus Gaussian noise attributed to background radiation and thermal noise in the receiver. For such a channel, the structure shown in Fig. 1 is suitable if the system is employing relatively simple modulation schemes, generally with no more than four signal points in the constellation. However, by the late 1970s, there had been significant work to increase the bandwidth efficiency of communication systems, which, assuming that the pulseshaping is left unchanged, is done by increasing the rate in information bits per symbol sent across the effective channel. To achieve this gain in rate, the rate of the convolutional code or the number of signal points used in the constellation is increased. When the structure of Fig. 1 was employed for constellations with larger numbers of

signal points, it was no longer desirable in many cases. This was demonstrated in the late 1970s and early 1980s by Ungerboeck's seminal work [4], which showed that separating the error control coding and modulation as depicted in Fig. 1 can be decidably suboptimal for the AWGN channel. Ungerboeck's construction that led to this conclusion was termed *trellis-coded modulation* (TCM), the structure of which is shown in Fig. 2. Note that certain information bits go through the encoder while others do not, thus making it impossible to separate coding and modulation. In addition, the method for choosing a signal point is jointly designed with the coder. Trellis-coded modulation with Ungerboeck's basic structure and his rules for building schemes on that structure, often termed "Ungerboeck set partitioning" after one of the key aspects of the rules, soon became an indispensable tool in the communication engineer's toolbox, and, hence, it was readily apparent that coding and modulation were thereafter inseparable.

By the late 1980s, many TCM schemes had been developed for the AWGN channel. However, wireless communication systems, which had been studied with mild intensity over the previous decades, became of increasing importance as the potential of a huge commercial cellular telephony market loomed. The AWGN channel model does not generally represent the wireless communications channel well, because, in wireless systems, the signal reflects off of many objects (automobiles, buildings, mountains, etc.), which leads to the superposition of many replicas of the transmitted signal in the environment. At a given point in the environment, these many waves can add constructively or destructively, depending on the relative phase of the replicas at that point. Thus, depending on the location of the receiver, the received signal power can be significantly more or less than would be expected without the presence of such multiple copies; because of its cause, this fluctuation of the signal power is termed *multipath fading*. From the physics of the problem, it is important to note that the received power of the signal can vary greatly with time because of the movement of the receiver to a different place in the environment or in response to changes in the reflections when objects in the environment move.

Multipath fading can have a significant impact on the performance of a communication system. When the received power drops too low, a burst of bit errors can occur, and such bursts tend to dominate the error probability — even if the occurrence of such system power drops is relatively unlikely. To combat this effect, an effective method is to make it such that possible symbol sequences for $\underline{X}$ that correspond to different information bit sequences are distinguishable even when the received power drops significantly a portion of the time. However, most coded modulation schemes produce many pairs of sequences whose distinguishing characteristics are limited to sequence elements very near each other, which are transmitted across the channel at nearly the same time under the architecture of Fig. 1; thus, a single drop in the received power, which can last many symbol periods, will make such sequences virtually indistinguishable at the receiver. A method of combating such an effect and spreading out the distinguishing characteristics between sequences in time is to reorder the symbols out of the coded modulator before they are transmitted as shown in Fig. 3. This reordering is generally done through a process termed *interleaving*, which essentially permutes the ordering of the symbols. Since symbols in $\underline{X}$ that were close to each other at the input of the interleaver are now separated significantly at its output and thus transmitted across the channel at times relatively far apart, the distinguishing characteristics between transmitted sequences associated with different information bit sequences are unlikely to be lost as a result of a single drop in the received power. This process, which is referred to as achieving "time diversity," greatly improves system performance.

There was significant work in the late 1980s and early 1990s to develop TCM schemes for wireless communication systems based on the architecture given in Fig. 3 with the coded modulation paradigm of Fig. 2. It became readily apparent that different criteria [5,6] were required for multipath fading channels relative to their AWGN counterparts; in particular, because of the importance of achieving time diversity on the multipath fading channel, the ability to distinguish between two possible sequences output from the coded modulation is not determined by the standard Euclidean distance between those sequences, which is appropriate for the AWGN channel, but rather by the number of elements in the two sequences that are different. Thus, an entire new line of coded modulation schemes employing Ungerboeck's construction was developed under this new criterion, and, although this new line of coded modulation schemes mandated the

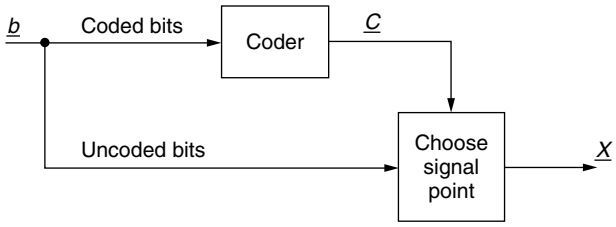


Figure 2. Ungerboeck's structure for coded modulation. Note that the method for choosing the signal point is a critical portion of the architecture, which Ungerboeck specified by a particular method of signal set partitioning.

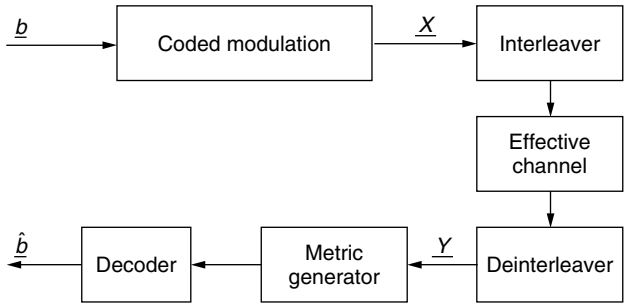


Figure 3. Block diagram of a communication system with interleaving, which is employed to obtain time diversity when conveying the information sequence $\underline{b}$ across a wireless communication channel.

removal of the uncoded bits in Fig. 2, the coding and modulation were still designed jointly, generally following Ungerboeck's rules.

Zehavi's 1992 paper [1] exploited the fact that the metric for sequence distinguishability for multipath fading channels is different from that for AWGN channels; thus, perhaps the structure of the coded modulation should change as well. Thus, bit-interleaved coded modulation was introduced, which employs the coded modulation of Fig. 1 in conjunction with the interleaver of Fig. 3, resulting in the structure shown in Fig. 4. As will be shown below, such a construction immediately leads to a larger number of differences between two possible coded modulation output sequences for a given code complexity. Analytical results and simulation results [2] have confirmed the desirability of such a structure, not only for wireless communication systems but also potentially for a variety of other channels, and it has even been suggested that BICM may be a strong contender for implementation over AWGN channels when powerful codes are employed. In addition, the ability of BICM to protect each information bit with coding has made it a robust choice for related applications [7].

The remainder of this article is organized as follows. In Section 2, the mathematical models for the AWGN channel and the interleaved multipath fading channel, as introduced above, are presented, along with the metrics for determining the quality of a given coded modulation scheme operating on that channel. In Section 3, the various coded modulation constructions discussed above are briefly reviewed. These constructions will motivate the structure of BICM, which is discussed in detail in Section 4. In particular, Section 4 presents encoding and decoding methods for BICM, and discusses its performance versus state-of-the-art TCM schemes for both the multipath fading channel and the AWGN channel. Finally, Section 5 summarizes this article and suggests further reading.

2. CHANNEL MODELS

2.1. The AWGN Channel

The additive white Gaussian noise (AWGN) channel is the most classical of communication channels and the channel studied most often before the explosion in wireless communications research in the late 1980s. The AWGN channel is a channel with additive distortion, and thus the ith element of the output $\underline{Y}$ is given by

$$Y_i = X_i + N_i$$

where X_i is the ith element of the input sequence $\underline{X}$. The sequence of random variables N_i, which correspond to additive noise encountered in the channel and the front end of the receiver, are modeled as independent random variables, each of which is Gaussian with mean zero and variance $N_0/2$. Since the N_i are independent, the statistical characterization of the channel output sequence $\underline{Y}$ given the input sequence $\underline{X} = \underline{x}$ factors into the conditional probability density functions of the individual components:

$$p_{\underline{Y}|\underline{X}}(\underline{y} \mid \underline{x}) = \prod_i p_{Y_i|X_i}(y_i \mid x_i) = \prod_i \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{(y_i - x_i)^2}{N_0}\right)$$

Under this channel model, the maximum-likelihood (ML) detector, which chooses the most likely information sequence given the received sequence $\underline{Y} = \underline{y}$, will choose from among the possible transmitted sequences the one that is closest in squared Euclidean distance to $\underline{y}$, where the squared Euclidean distance between the two sequences $\underline{y}$ and $\underline{u}$ is defined as

$$d_E^2(y, u) = |y - u|^2 = \sum_i (y_i - u_i)^2$$

The performance of the system can be characterized by the probability that the ML detector chooses the wrong information sequence (i.e., one other than the one input to the transmitter). In particular, through the use of familiar union bounding techniques, measures for the performance of the communication system (e.g., bit error rate or frame error rate) are generally a linear combination of elements from the set (over all distinct $\underline{x}$ and $\tilde{\underline{x}}$) of probabilities that the received sequence $\underline{Y}$ is closer to the possible transmitted sequence $\tilde{\underline{x}}$ than to the actual transmitted sequence $\underline{x}$. For the AWGN channel, this probability is given by

$$P(\underline{x} \to \tilde{\underline{x}}) = Q\left(\frac{d_E(\underline{x}, \tilde{\underline{x}})}{(2N_0)^{1/2}}\right) \tag{1}$$

where $Q(x) \triangleq (1/\sqrt{2\pi}) \int_x^\infty e^{-(u^2/2)}\, du$. In particular, of these pairwise error events, the one that is most likely, which is the one corresponding to the possible sequences $\underline{x}$ and $\tilde{\underline{x}}$ that are closest in Euclidean distance, dominates the system performance at high signal-to-noise ratios (SNR), where systems are generally designed. Thus, the goal of coded modulation over the AWGN channel is to map information sequences $\underline{b}$ to transmitted sequences $\underline{X}$ in such a way that the minimum Euclidean distance between any two possible transmitted sequences corresponding to different information sequences is maximized.

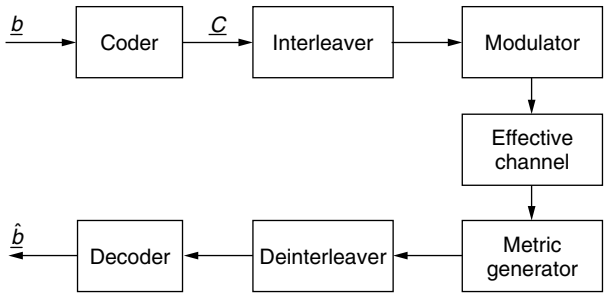


Figure 4. Block diagram of a bit-interleaved coded modulation system.

2.2. Multipath Fading Channels

Per Section 1, interest in the wireless communications channel motivated the study of a channel model very different from the classical AWGN channel. In particular, the following non-frequency-selective slowly fading channel model, which is appropriate for the traditional narrowband communication system or a subcarrier of a modern orthogonal frequency-division multiplexing (OFDM) system, is generally assumed:

$$Y_i = \alpha_i X_i + N_i \tag{2}$$

where N_i is defined in the same manner as for the AWGN model, and $\underline{\alpha}$ is a sequence of attenuations of the signal strength, where each element comes from a common distribution, which depends on the channel. This common distribution is generally assumed to be a Rayleigh, Ricean, or Nakagami-m distribution. The Rayleigh distribution will be employed throughout this paper, which implies that the probability density function of each α_i is given by

$$p_{\alpha_i}(x) = 2xe^{-x^2}, \qquad x \geq 0$$

where the average power due to this multipath fading has been normalized to unity. If the system architecture in Fig. 1 is assumed, then the elements of $\underline{\alpha}$ are correlated with each other, since the mobile moves across the interference pattern set up in the environment slowly relative to the rate at which symbols are sent across the channel. However, if the system architecture of Fig. 3 is assumed and a reasonable amount of system latency is allowed, then the entries of the sequence $\underline{\alpha}$ can be modeled as independent, since the places where $\underline{x}$ and $\underline{\tilde{x}}$ differ following the impact of a given information bit b_i can be assumed to be separated by a time sufficient to render the multipath fading affecting these places independent. Thus, the model of (2) will be adopted throughout this work with the elements of $\underline{\alpha}$ assumed to be independent. Similarly to the AWGN channel, the system performance measures of interest are linear combinations of elements from the set (over all distinct $\underline{x}$ and $\underline{\tilde{x}}$) of probabilities that the sequence $\underline{\tilde{x}}$ is chosen when sequence $\underline{x}$ was sent, which is given by

$$P(x \to \tilde{x}) = E_\alpha \left[Q\left(\frac{\alpha_i d_E(\underline{x}, \underline{\tilde{x}}))}{(2N_0)^{1/2}} \right) \right] \tag{3}$$

$$\leq \prod_{i \in D(\underline{\tilde{x}}, \underline{x})} \frac{1}{1 + \dfrac{|\tilde{x}_i - x_i|^2}{4N_0}} \tag{4}$$

where $D(\underline{\tilde{x}}, \underline{x})$ is the set of indices corresponding to locations where $\underline{\tilde{x}}$ and $\underline{x}$ differ. For high SNRs, the first term in the denominator can be ignored, which leads to

$$P(\underline{x} \to \underline{\tilde{x}}) \approx \frac{1}{\displaystyle\prod_{i \in D(\underline{\tilde{x}}, \underline{x})} \frac{|\tilde{x}_i - x_i|^2}{4N_0}} \tag{5}$$

It can be readily observed that a plot of (5) on a logarithmic scale versus the SNR in decibels will exhibit a slope at high SNRs that is the negative of the size of the set $D(\underline{\tilde{x}}, \underline{x})$. Furthermore, the minimum size of $D(\underline{\tilde{x}}, \underline{x})$ over all possible transmitted sequences $\underline{\tilde{x}}$ and $\underline{x}$ will characterize this same slope for the bit error rate for the code, and thus this is a critical parameter for good performance at high SNRs. It determines the diversity of the system, which is indeed defined as the negative of the slope of the information bit error rate versus SNR at high SNRs. Whereas the size of $D(\underline{\tilde{x}}, \underline{x})$ sets the slope of the curve, the magnitudes of the values $|\tilde{x}_i - x_i|$ for $i \in D(\underline{\tilde{x}}, \underline{x})$ set the horizontal positioning of the curve. This establishes the two criteria for designing coded modulation on the perfectly interleaved Rayleigh fading channel:

1. *Primary* — maximize the minimum number of elements in $D(\underline{\tilde{x}}, \underline{x})$.
2. *Secondary* — maximize the minimum product distance $\displaystyle\prod_{i \in D(\underline{\tilde{x}}, \underline{x})} |\tilde{x}_i - x_i|$.

3. CODED MODULATION

In this section, the evolution of coded modulation for AWGN and Rayleigh fading channels that led to the introduction of BICM is briefly reviewed.

3.1. Coding and Modulation

3.1.1. Encoder. The traditional approach to coding and modulation is shown in Fig. 1. If the overall coded modulation is trellis-based, as will be assumed throughout this article, a convolutional code is employed. In a convolutional code, the information bits are coded with a shift register circuit, an example of which is given in Fig. 5a. For this example, at each timestep, an information bit is input to the left side of the shift register; this input bit, along with the last 2 information bits, each of which is contained in one of the 1-bit memory elements, is used to determine the current pair of outputs by the modulo-2 summations. The memory elements are then clocked, which causes each information bit to move one memory location to the right, thus causing the encoder to reside in a new memory state, and a new information bit to be input to the system from the left. In general, a trellis code is denoted an (n, k) code with memory m, where n is the number of output coded bits for each clock cycle, k is the number of input information bits for each clock cycle, and m is the number of 1-bit memory elements in the circuit. Thus, the convolutional code of Fig. 5a would be denoted a $(2, 1)$ code with memory 2.

Recall that good coded modulation schemes produce, for distinct input sequences, outputs that are as distinct as possible. The number of places where two binary sequences differ is termed the *Hamming distance* between those two sequences. For a convolutional code, the minimum possible Hamming distance between output sequences corresponding to distinct information sequence inputs is

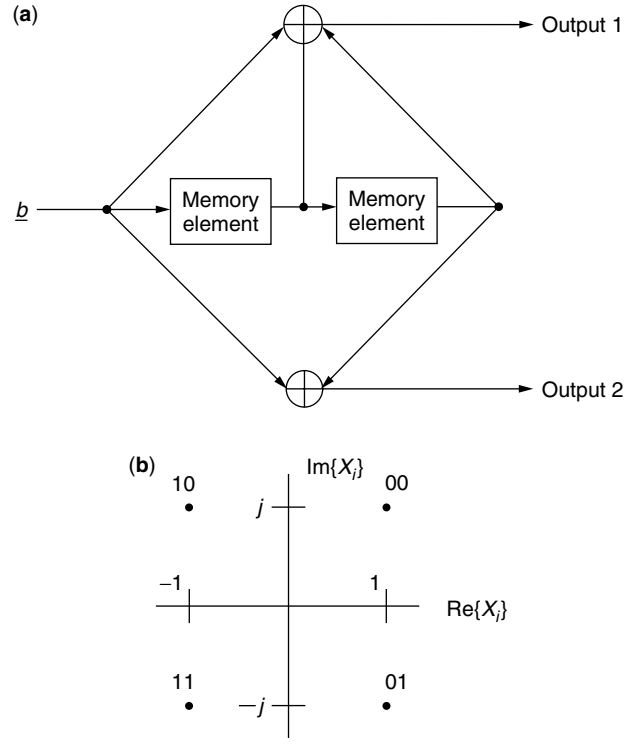


Figure 5. A simple example of coded modulation: (**a**) encoder for a $(2, 1)$ convolutional code with memory 2, (**b**) a possible constellation labeling for QPSK.

termed the *free distance*, which will be denoted by d_f. This free distance, along with other distance properties of the code, depend on which memory elements contribute to which output (i.e., the connections to the modulo-2 summers in Fig. 5a). Because of the importance of convolutional codes, an enormous amount of research has gone into finding convolutional codes of maximal free distance for a given number of memory elements. The connections for binary convolutional codes of maximal free distance can be found from standard references [8, pp. 330–331]; for example, the code shown in Fig. 5a corresponds to the recipe given in the first line of Table 11.1(c) of Ref. 8.

The sequence of bits output from the convolutional coder must then be assigned to a transmitted sequence $\underline{X}$, whose elements are drawn from the space of complex numbers. Assuming that there are M possible signal points in the constellation from which the value for each entry of $\underline{X}$ is drawn, the modulator takes $\log_2 M$ output bits from the output of the convolutional coder and uses these to select one of the M constellation points. The mapping of the $\log_2 M$ bit sequence to a constellation point is termed the "constellation labeling," because it can be specified by labeling each signal point with a $\log_2 M$ bit sequence. A simple example of this is shown in Fig. 5b for a quadrature phase-shift-keyed (QPSK) constellation.

3.1.2. Decoder. A convenient representation of a coded modulation scheme for characterizing the performance and for the decoding of convolutional codes is a trellis diagram, as shown in Fig. 6a for the coded modulation scheme of Fig. 5. Here, each branch of the trellis has been labeled with the output of the coded modulation scheme when that branch is traversed in the convolutional coder. Assuming that the encoder of Fig. 5 starts with a zero in each of its memory elements, any possible path through the trellis enumerates a possible sequence output from the coded modulator, and, any possible sequence output from the coded modulator corresponds to a path through the trellis.

Per the observations discussed above, for performance analyses, interest lies in the distinction between any two possible coded modulation sequences output from the coded modulator, or, equivalently, the distinction between any two different paths through the trellis. From Fig. 6a, we see that there are two coded modulation sequences, $\underline{x}_1 = (+1+j, +1+j, +1+j, \ldots)$ and $\underline{x}_2 = (-1-j, -1+j, -1-j, \ldots)$, which split at time $t = 0$, rejoin at time $t = 3$, and are identical after $t = 3$. The probability for choosing $\underline{X} = \underline{x}_2$ when $\underline{X} = \underline{x}_1$ was sent is then easily found from (1) or (4) for the AWGN or perfectly interleaved Rayleigh fading channel, respectively. Likewise, the pairwise error probabilities for all possible sequence differences can be found using the trellis of Fig. 6a.

The trellis is also employed for decoding of the coded modulation scheme. Recall that the goal of decoding is generally to find the most likely transmitted sequence $\underline{x}$ given the received sequence $\underline{Y} = \underline{y}$. Since each transmitted sequence is represented by a path in the trellis, this reduces to finding the sequence through the trellis that is the most likely given $\underline{Y} = \underline{y}$. At any time step t, a path in

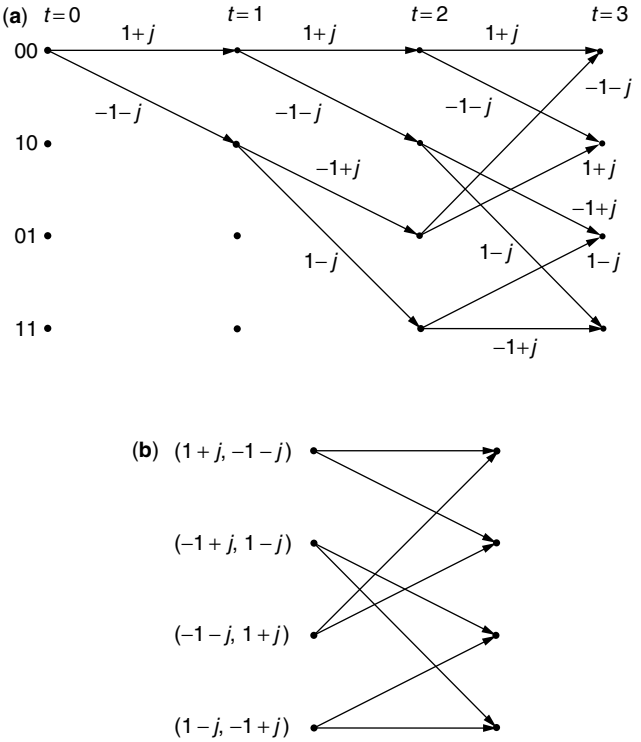


Figure 6. Trellis representation for coded modulation schemes: (**a**) full trellis for decoding and performance analyses, (**b**) compact trellis for specifying the coded modulation scheme.

the trellis is associated with its likelihood

$$p_{Y|X}(y \mid x) = \prod_{i=1}^{t} p_{Y_i|X_i}(y_i \mid x_i)$$

$$= \begin{cases} \displaystyle\prod_{i=1}^{t} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{(y_i - x_i)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{(y_i - \alpha_i x_i)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

where the fact that both the AWGN channel and the perfectly interleaved Rayleigh fading channel considered here are memoryless, has been exploited. In addition, coherent reception, where the carrier phase is available at the receiver, has been assumed, as has perfect estimation of the channel state information (CSI) α_i for the Rayleigh fading channel. Since the natural logarithm is a monotonic function, taking the natural logarithm of the likelihood of each path implies that it is equivalent to search for the path $\underline{x}$ that exhibits the minimum value of

$$\mu(\underline{y}, \underline{x}) = \begin{cases} \displaystyle\sum_{i=1}^{t} (y_i - x_i)^2, & \text{AWGN channel} \\ \displaystyle\sum_{i=1}^{t} (y_i - \alpha_i x_i)^2, & \text{Rayleigh channel} \end{cases} \tag{6}$$

which is easy to evaluate and can often be simplified further as discussed in standard digital communication texts.

Thus, at each time stage t in the trellis, there is a likelihood associated with each possible path, and it is clear from (6) that the metric for each of the paths at time $t + 1$ can be calculated from the metrics of the paths at time t. However, for the example of Fig. 5, there are 2^t possible paths at time t, implying that some sort of simplification is required if the receiver is to be implementable. The solution to this problem is provided by the celebrated Viterbi algorithm. Note from Fig. 6a that two paths merge into each channel state at time $t = 3$. Now, recalling that the receiver's goal is only to find the single path $\underline{x}$ with the *highest* likelihood given the received sequence $\underline{Y} = \underline{y}$, the likelihood for the two paths entering the same state can be compared with one another and only the best path retained without loss of optimality. The reasoning for this is as follows. Consider the supposition that the path with the smaller likelihood entering a given state is the prefix for the path through the entire trellis with the eventual highest likelihood. Then, if one takes the suffix of this "best" path and concatenates it to the path with the higher likelihood entering the given state, a path through the entire trellis with higher likelihood results, thus proving the supposition false. Hence, one can always disregard all but one of the paths entering a given state at a given time. Note that this trims the number of paths retained at time t significantly, from 2^t to the steady-state number 2^m, where $m = 2$ for the example of Fig. 5.

Notationally, it is inconvenient to specify the trellis shown in Fig. 6a for coded modulation schemes, particularly for systems with large numbers of branches. Thus, two notation simplifications are generally employed: (1) only the steady-state trellis is generally manifested, as shown in Fig. 6b, rather than the startup stages; and (2) rather than placing the label actually on each branch, which leads to significant confusion, particularly in high-rate systems, the labels are listed to the left of a given state, with the understanding that they correspond to the branches leaving that state at any given time in order from top to bottom. Figure 6b illustrates this compact notation for the coded modulation scheme of Fig. 5.

3.2. Trellis-Coded Modulation for AWGN Channels

Per the observations mentioned above, the goal of coded modulation design for AWGN channels is to separate distinct possible transmitted sequences (or paths through the trellis) $\underline{x}$ and $\tilde{\underline{x}}$ by a Euclidean distance $|\underline{x} - \tilde{\underline{x}}|^2$ that is as large as possible. In particular, the minimum of all such differences for distinct paths is critical. If the constellation employed by the modulator is the binary antipodal set $\{-1, +1\}$, often implemented as binary phase-shift keying (BPSK), or quadrature phase shift keying $\{-1 - j, -1 + j, +1 + j, +1 - j\}$, the coded modulation structure shown in Fig. 1 with a convolutional code of maximal free distance is a good solution.

If the size of the constellation employed by the modulator is larger than BPSK and QPSK, the best method of coded modulation is not so clear. For example, suppose that one would like to send 2 information bits per symbol. One possibility would be to employ QPSK with no coder, but this is often a poor choice. Before the late 1970s, the structure of Fig. 1 would have been retained, and a convolutional code of maximal free distance would have been employed. Using the rule of thumb for the constellation size of coded modulation systems operating on AWGN channels, which states that the constellation size M should be such that $\log_2 M - 1$ is the number of bits per symbol, results in the choice of a rate-$\frac{2}{3}$ convolutional code followed by modulation with a 8-ary phase shift keyed (8-PSK) signal set. A reasonable choice for the labeling of coded bits to constellation points is to employ Gray labeling, whereby the label for each signal point differs by exactly 1 bit from each of its nearest neighbors. With this labeling, 3-bit sequences separated by large Hamming distances correspond to signal points separated by large Euclidean distances.

Per Section 1, Ungerboeck [4] revolutionalized the art of coded modulation by introducing the concept of trellis-coded modulation, as shown in Fig. 2. Two key philosophical concepts became readily apparent: (1) coding and modulation could no longer be considered separately in high-rate systems, and (2) the constellation labeling plays a key role. Note that some bits in the Ungerboeck scheme do not go through the convolutional encoder; these "uncoded bits" show up in the trellis representations in Fig. 6 as parallel branches, which originate and end at the same state with a single transition. Although two sequences that differ only on these parallel branches are allowable codewords, the

large Euclidean distance generally obtained between parallel branches keeps these codewords from greatly restricting the minimum distance of the coded modulation scheme.

3.3. Trellis-Coded Modulation for Fading Channels

The concept of trellis-coded modulation produced a large number of good coded modulation schemes for the AWGN channel during the 1980s. As the Rayleigh fading channel rose in importance late in the 1980s, it was natural to extend the idea of coded modulation to this environment. However, as noted in Section 2.2, the criteria for the distance between two paths is quite different for a Rayleigh fading channel. In particular, the parallel branches of a scheme employing uncoded bits as in Fig. 2, which result in parallel paths in the trellis of Fig. 6 and, hence, sets $D(\underline{\tilde{x}}, \underline{x})$ with only a single entry, result in poor performance on the perfectly interleaved Rayleigh fading channel. Essentially, the uncoded information bit is not protected from deep signal fades since its impact is concentrated on only a single modulated symbol, and thus it is decoded in error with very high probability.

Thus, it was soon recognized [5,6] that trellis-coded modulation schemes for Rayleigh fading channels should avoid the uncoded bits often employed on the AWGN channel. One of the later instantiations of trellis-coded modulation schemes for fading channels was the I-Q TCM scheme of Al-Semari and Fuja [9], which cleverly performed trellis-coded modulation on two 4-ary amplitude shift-keyed (4-ASK) $\{-3, -1, +1, +3\}$ streams and then combined these together by using one 4-ASK stream as the in-phase component and one 4-ASK stream as the quadrature component to get a stream of 16-ary quadrature amplitude modulation (QAM) symbols. When rate-$\frac{1}{2}$ encoders are used, the resulting scheme conveys 2 information bits per symbol. Such an I-Q TCM scheme employing a four-state code is shown in Fig. 7. Note the diversity increase from 1 for a TCM scheme employing uncoded bits per Fig. 2 to 3 for the scheme of Fig. 7, which, per Section 2.2, will greatly improve performance on the perfectly interleaved Rayleigh fading channel.

4. BIT-INTERLEAVED CODED MODULATION

4.1. Motivation

The schemes discussed in Section 3 are all built on the structure shown in Fig. 2, often without the uncoded bits in the case of the multipath fading channel. In particular, all the schemes discussed in Section 3 have assumed that each set of n bits along a branch in the convolutional coder impacts a single output symbol. Under such a paradigm, the maximum diversity for a coded modulation scheme is upper-bounded by the symbolwise Hamming distance of the coded modulation, which is defined for the schemes of Section 3 as the number of branches that differ between two possible paths through the trellis. Thus, for the convolutional encoder shown in Fig. 5, the I-Q TCM of Fig. 7 achieves the upper bound of three on the diversity under such a construction paradigm.

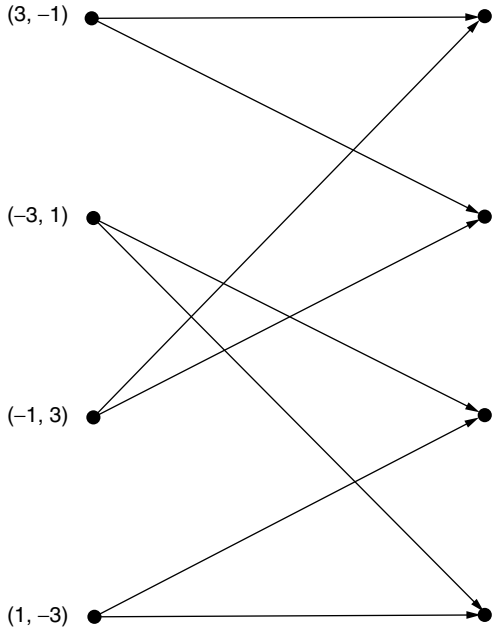


Figure 7. Specification for an I-Q TCM scheme using the encoder of Fig. 5 mapped to a 4-ASK signal set via Gray labeling.

However, the bitwise Hamming distance between codewords for the encoder of Fig. 5 is 5 instead of 3, thus motivating the method pictured in Fig. 4, which is termed *bit-interleaved coded modulation*. The goal of such a construction is to increase the diversity of the coded modulation scheme from 3 to 5 by interleaving at the bit level so that the five symbols that carry the difference between the two possible sequences out of the encoder are temporally separated. These five symbols will then see roughly independent fading, and thus a diversity of 5 should be achieved.

The BICM construction breaks from the standard coded modulation in a number of ways. Prominent among these is that the coded bits that choose a constellation point for a given modulated symbol no longer come sequentially from the encoder; in fact, to achieve the maximum diversity, it is desirable that they are drawn from places in the encoder sequence that are as distant as possible. Thus, in both encoding and decoding, the bits joining the bit on the trellis branch of interest will generally be unknown, which is why a given bit in BICM is often said to be randomly modulated. Because of this essential loss of control of the way that a given code bit affects a transmitted symbol, new rules for the design of such a system and new analysis techniques must be developed relative to those considered in Section 3.

4.2. Encoder

The generic BICM system is shown in Fig. 4. In general, assuming a constellation of M signal points, $\log_2 M$ signal points are taken from the output of the interleaver and used to choose a constellation point. However, to make this discussion more concrete, consider the system shown in Fig. 8a, which carries 2 information bits per 16-QAM

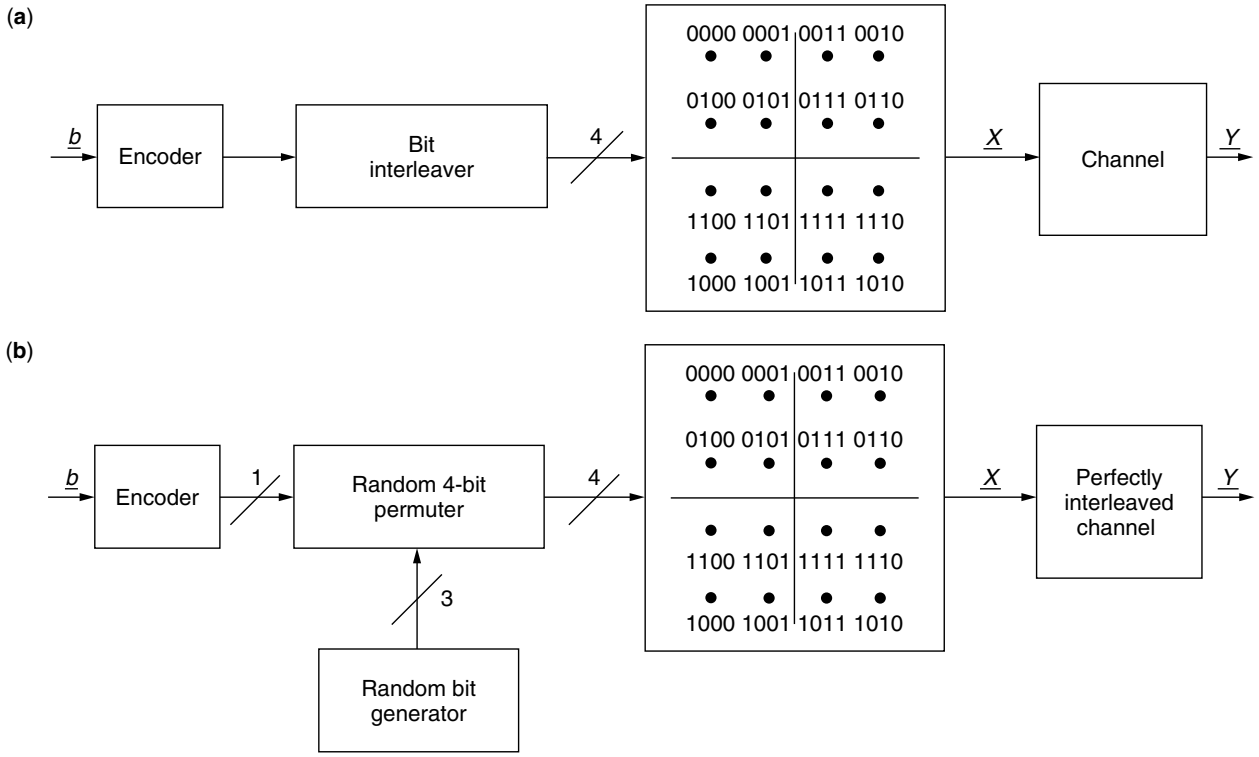


Figure 8. The BICM block diagram: (**a**) a BICM scheme that transmits 2 information bits per channel symbol; (**b**) the equivalent representation for such a scheme assuming a perfect bit interleaver.

channel symbol and can thus be compared to the sample systems of Section 3. Assuming a perfect interleaver, which separates the bits of the input sequence essentially infinitely far apart in time at the output (and thus suppressing interleaver design questions), two questions arise for the design of the encoder:

1. What are desirable properties of the convolutional code?
2. How should the bits taken at the output of the bit interleaver 4 at a time be mapped to the 16-QAM constellation points, or, equivalently, what should be the 4-bit labels of the 16-QAM constellation points?

To help answer these questions, consider the equivalent system shown in Fig. 8b for a BICM system under the assumption of perfect interleaving.

First, consider the question of how to choose the convolutional encoder. The minimum Euclidean distance squared between the two sequences caused by a given information bit is proportional to the free distance of the convolutional code (note that this result does not depend on the assumption of independent fading). More importantly, the diversity of the BICM scheme is equal to the minimum bitwise Hamming distance of the convolutional code. Thus, it is straightforward to conclude that a convolutional code of maximal free distance is a good choice.

The best mapping from each set of $\log_2 M$ bits taken at the output of the bit interleaver to one of the M-ary signal points is not as obvious. It seems reasonable to assume that sets of $\log_2 M$ bits that are separated at large Hamming distances at the output of the bit interleaver should be separated at large Euclidean distances at the output of the constellation mapper. Thus, the technique of Gray labeling, whereby the label for each signal point differs by exactly one bit from each of its nearest neighbors, is a logical choice. In fact, Gray labeling has been shown to have some optimality properties [2], although it has not been shown to be optimal for practical systems in general. Indeed, as discussed in Section 4.4, very different labelings can be desirable for certain applications and certain types of decoders.

Thus, the standard convention for BICM is to employ a convolutional code of maximal free distance in conjunction with Gray labeling of the constellation to achieve good performance. Note that this is an advantage of BICM — its design is relatively straightforward and thus less of an art than in previous TCM schemes.

4.3. Decoder

4.3.1. The Optimal Decoder. Unlike the TCM schemes discussed earlier, it is far too complex to decode BICM optimally, because the bit interleaver intertwines the bits on a given branch in the trellis with those from other branches far removed in the trellis, thus essentially requiring the complicated soft-decision decoding of a block code with length greater than the depth of the interleaver. Since a large interleaver is generally employed to achieve independent fading on the coded bits for a given branch,

this complexity is not feasible for current receivers. Since iterative methods can approach the performance of maximum-likelihood (ML) decoding as described in Section 4.4, it is unlikely that ML decoding will ever be considered for BICM.

4.3.2. The Optimal Metric Generator. Since decoding over the entire interleaving depth is not viable, it is common to treat the bits that join a bit from the branch of interest in the convolutional encoder as random as shown in Fig. 8b. Note that not only are these randomly generated bits joined with the bit of interest but also that the bit of interest is randomly assigned to one of the $\log_2 M$ label positions. The decoder knows the bit position where the bit of interest is located on a given symbol, but it does not know the value of the other bits that combine with that bit for that symbol.

As with the TCM schemes discussed in Section 3.1.2, the decoder for the BICM scheme attempts to calculate the most likely bit sequence in the trellis given the received sequence $\underline{Y} = \underline{y}$. Assuming each of the transmitted sequences is equally likely, this is equivalent to finding the bit sequence for which the likelihood of the received vector given that bit sequence is maximized. Assuming an AWGN channel or a perfectly interleaved Rayleigh fading channel, the likelihood that $\underline{Y} = \underline{y}$ given the bit sequence $\underline{C} = \underline{c}$ was transmitted is given by

$$p_{Y|C}(y \mid c) = \prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i)$$

where the received symbol Y_i corresponds to the channel symbol on which the ith bit was placed. Because of the random modulation caused by the other random bits that join C_i on a given symbol, calculating this likelihood in BICM is a bit more involved than it is for TCM. Let $j_i \in \{0, 1, \ldots, \log_2 M - 1\}$ be the position in the signal set label where C_i was mapped by the bit permuter, and, following [2], define χ_b^j to be the set of signal points in S such that the jth label position is equal to b. Then, using the law of total probability, we obtain

$$\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) = \prod_{i=1}^{t} \frac{1}{|\chi_{c_i}^{j_i}|} \sum_{x \in \chi_{c_i}^{j_i}} p_{Y_i|X_i}(y_i \mid x)$$

$$= \begin{cases} \displaystyle\prod_{i=1}^{t} \frac{1}{|\chi_{c_i}^{j_i}|} \sum_{x \in \chi_{c_i}^{j_i}} \frac{1}{\sqrt{\pi N_0}} \\ \qquad \times \exp\left(-\dfrac{(y_i - x)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \frac{1}{|\chi_{c_i}^{j_i}|} \sum_{x \in \chi_{c_i}^{j_i}} \frac{1}{\sqrt{\pi N_0}} \\ \qquad \times \exp\left(-\dfrac{(y_i - \alpha_i x)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

which can be simplified by removing common terms to

$$\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) \sim \begin{cases} \displaystyle\prod_{i=1}^{t} \sum_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - x)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \sum_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - \alpha_i x)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases} \tag{7}$$

and thus it is readily apparent that the metric generation process can be quite a bit more complicated than that for the standard TCM schemes in Section 3.1.2, particularly for large signal sets. It requires a summation over half of the signal set of a nonlinear function of the distance to form the metric contribution for a single bit, whereas the metric contribution for an entire branch can be calculated for the TCM scheme with only simple addition and multiplication involving one signal point.

4.3.3. A Suboptimal Metric Generator. As noted above, the metric given in (7) is much more complicated than that for the standard TCM schemes. Thus, in this section, a suboptimal bit metric suggested by Caire et al. [2] is reviewed. The suboptimal bit metric relies on the fact that the sum of a number of quantities that are quite disparate in magnitude is well approximated by the maximum of those quantities. Thus, the summations in (7) are replaced by maximums to yield

$$\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) \sim \begin{cases} \displaystyle\prod_{i=1}^{t} \max_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - x)^2}{N_0}\right), \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\prod_{i=1}^{t} \max_{x \in \chi_{c_i}^{j_i}} \exp\left(-\frac{(y_i - \alpha_i x)^2}{N_0}\right), \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

and, taking the natural logarithm of the quantities on the right and recognizing the monotonicity of the logarithm, yields that it is equivalent to minimize

$$-\prod_{i=1}^{t} p_{Y_i|C_i}(y_i \mid c_i) \sim \begin{cases} \displaystyle\sum_{i=1}^{t} \min_{x \in \chi_{c_i}^{j_i}} (y_i - x)^2, \\ \qquad\qquad \text{AWGN channel} \\ \displaystyle\sum_{i=1}^{t} \min_{x \in \chi_{c_i}^{j_i}} (y_i - \alpha_i x)^2, \\ \qquad\qquad \text{Rayleigh channel} \end{cases}$$

which requires only that the point in $\chi_0^{j_i}$ and $\chi_1^{j_i}$ that is closest to the received symbol Y_i be found, which is relatively simple for small constellations. For larger or multidimensional constellations, this can be accomplished though a technique known as *sphere decoding* [10].

4.4. Iterative Decoding

As discussed in Section 4.3.1, the optimal decoder for BICM is not practically implementable, due to the interlacing of different bits from different portions in the trellis onto the channel at the same time, since soft-decision decoding would essentially have to be done across the entire interleaver depth. Thus, as shown in Section 4.3.2, the typical decoder assumption is that the other bits that were used with the bit of interest to choose a constellation point are randomly generated. However, the sets $\chi_0^{j_i}$ and $\chi_1^{j_i}$ associated with the coded bit i could be reduced to a single point if the values of these other bits were known, which should substantially improve decoder performance. This notion was exploited by Li and Ritcey in a series of papers (see Ref. 11 and references cited therein) that introduced bit-interleaved coded modulation with iterative decoding. In this technique, the BICM scheme is decoded in a manner similar to that described in Section 4.3.2, but now the decoder generates "soft" information, which tells not if the bit is a 0 or a 1 but instead the probability of such. The decoding is then repeated many times, but this time with the soft information from the previous decoding as an additional input. If the bit error rate of the original BICM scheme is reasonably low, one expects that the estimates of the soft information for the bits joining the coded bit of interest on a given symbol will be very accurate much of the time, and thus the sets $\chi_0^{j_i}$ and $\chi_1^{j_i}$ will be effectively reduced to single points, as desired.

The use of iterative decoding also motivates a change in the encoding. As illustrated in Fig. 8b, bit i is often said to be randomly modulated in BICM in the sense that its effective channel varies depending on the values of the other $\log_2 M - 1$ bits that join with a given coded bit to choose a signal. Note from Fig. 8b that, regardless of the position to which the coded bit i gets mapped (i.e., $j_i = 1, 2, 3, 4$), there always exists 3 bit values such that the minimum distance between the point corresponding to a bit value 0 and that point corresponding to a bit value 1 that are separated by the minimum distance of the signal set. For example, if the coded bit of interest is mapped to the first label location, the fact that the values of the other 3 bits are 1,0, and 0, respectively, implies that the effective signaling points for the binary channel for this bit are labeled 0100 and 1100, which is at the minimum distance of the signal set. Although such a phenomenon occurs only some fraction of the time, it dominates the error probability, particularly at high SNRs and on the AWGN channel. Thus, iterative decoding suggests various relabelings of the constellation points, as described by Li et al. [11].

Finally, note that iterative decoding of BICM is reminiscent of one of the greatest advances in coding theory: concatenated codes with iterative decoding. In fact, the block diagram in Fig. 4 is very similar to that of a serial concatenated Turbo code [12], which would suggest that, if the modulator is viewed as a rate $\log_2 M$ inner code, iterative decoding may provide significant gains. As will be discussed below, iterative detection does provide significant gains, although the analogy to serial concatenated codes must be done very carefully, because there are significant differences.

4.5. Performance

In this section, the performance of BICM is compared to that of other popular coded modulation schemes. From Refs. 2 and 13, it can be inferred that the performance of BICM employing the suboptimal metric of Section 4.3.3 is generally only slightly inferior to the performance when the optimal metric generator of Section 4.3.2 is employed; thus, the suboptimal metric generation of Section 4.3.3 is most often used. In Fig. 9, the BER performance of BICM operating over a perfectly interleaved Rayleigh fading channel is shown for the case where the metric of Section 4.3.3 is employed at the receiver. If the curves in Fig. 9 are compared to a 2 bits/symbol I-Q TCM (see Fig. 6 of Ref. 9), which is a good method of coded modulation based on the paradigm of Fig. 2 in the architecture of Fig. 3, it can be observed that BICM shows a significant performance gain. Thus, when employing low-complexity noniterative decoders, BICM is an effective method of communication on the perfectly interleaved Rayleigh fading channel that motivated its development .

A number of papers have characterized the performance of BICM in a more theoretical fashion, which will allow the consideration of its potential performance with future coding techniques. In particular, Caire et al. [2] showed that the BICM structure incurs only a small loss in Shannon capacity versus coded modulation on the perfectly interleaved Rayleigh fading channel, and, somewhat surprisingly, on the AWGN channel. To attempt to capture the effects of codes of finite complexity, those authors [2] then investigated the system cutoff rate, which indicated that the BICM structure, while incurring a slight loss in the cutoff rate versus coded modulation on AWGN channels, shows significant gains versus coded modulation for the perfectly interleaved Rayleigh fading channel. This supports the gains that BICM shows over coded modulation schemes for practical coded schemes on the perfectly interleaved Rayleigh fading channel. Through a coding exponent analysis, Wachsmann, et al. [14] also investigated the performance of BICM and concluded that, on the AWGN channel, it was inferior to multilevel coding with multistage decoding, particularly for small blocklengths (including trellis codes). This supported the results of Schramm [15], which indicated that BICM is slightly inferior to multilevel techniques when employing trellis codes on the AWGN channel or the Rayleigh fading channel. Wachsmann et al. [16] then investigated the performance of BICM versus multilevel codes when Turbo codes [17] are employed, which are long block codes. Simulation results indicate that BICM is a very effective structure in such a situation, particularly for the Rayleigh fading channel, and that it possesses some universality properties in the sense that it performs very well on a variety of channels.

When standard convolutional codes are employed, numerical results from the iterative decoding of BICM employing a signal set that is not Gray-labeled as described in Section 4.4 indicate a significant performance improvement through such iteration over Gray-labeled BICM with noniterative decoding. In fact, when employing such a labeling and iterative decoding, BICM significantly outperforms standard trellis-coded modulation on the AWGN channel. This seems to conflict with the results

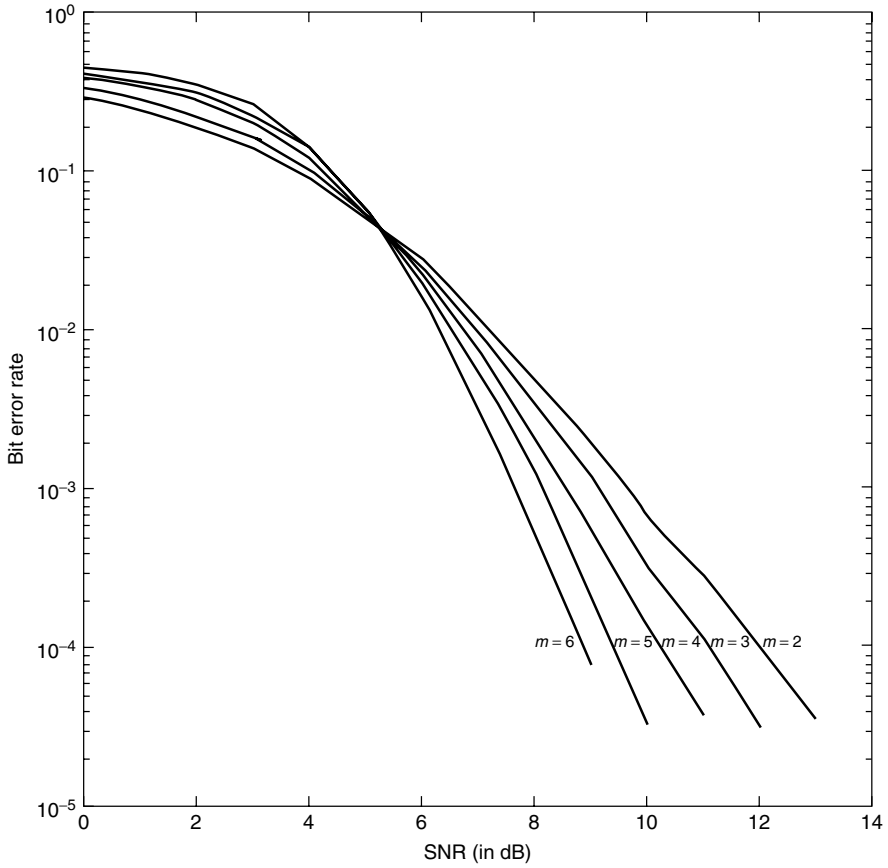


Figure 9. Simulated bit error rate for a BICM system that transmits 2 bits per symbol, which is obtained by employing a rate-$\frac{1}{2}$ convolutional encoder of maximal free distance in conjunction with a 16-QAM modulator, for various encoder memory sizes m.

of Caire et al. [2], which suggest that it is a Gray labeling of the signal points that maximizes the capacity of BICM. However, there are two explanations for this apparent conflict: (1) the model of Fig. 8b, which, as assumed by Caire et al. [2], does not take into account the constraint imposed by 4 bits of the encoder being combined for a single symbol, which is exploited by the iterative decoder; and (2) the signal-to-noise ratios (SNRs) at which the iterative decoder improves performance are quite a bit larger than capacity, indicating that capacity arguments might not be pertinent here.

It is tempting to equate iteratively decoded BICM with iteratively decoded serial concatenated codes, but one should be careful. In violation of the design rules for a serial concatenated code, the inner code in the BICM system is not recursive. Hence, the interleaver gain observed in various types of Turbo codes is not observed in iteratively decoded BICM. Perhaps a better analogy is to a little-known class of codes known as feedback-decoding trellis codes [18]. Much like in the case of BICM, feedback-decoding trellis codes use the knowledge of bits already decoded to aid in the demodulation of a constellation symbol combining bits from different sections of the same trellis.

5. SUMMARY AND SUGGESTED LITERATURE

In this article, the development of bit-interleaved coded modulation has been motivated from a history of coded modulation for the AWGN and perfectly interleaved Rayleigh fading channels. The designs for the encoder and various decoders for BICM have been described. Simulation results indicate that BICM is a strong competitor to traditional coded modulation techniques on perfectly interleaved Rayleigh fading channels and displays a sort of universality in the sense that it works well for a variety of different channels. In addition, iteratively decoded BICM can compete with Turbo code techniques for implementation on AWGN or Rayleigh fading channels.

For further information on this subject, the reader is encouraged to read in detail the work of Caire et al. [2]. The work of Li and Ritcey [11] is the authoritative work on the iterative decoding of BICM, but it is useful to understand the similar idea included in the concept of feedback-decoding trellis codes [18]. The work of Wachsmann et al. [14] on multilevel coding includes the useful extension of including BICM in that framework.

Acknowledgment

The author is indebted to Prof. Rick Wesel of UCLA, whose conversations and collaboration on topics of coded modulation have greatly contributed to the author's knowledge of the subject. In addition, the author would like to thank Prof. Jim Ritcey of the University of Washington for discussions on BICM with iterative decoding and for providing an advanced version of Ref. 11.

BIOGRAPHY

Dennis Goeckel split time between Purdue University and Sundstrand Corporation from 1987 to 1992, receiving his B.S.E.E. (with highest honors) from Purdue in 1992. From 1992 to 1995, he was a National Science Foundation Graduate Fellow at the University of Michigan, where he received his M.S.E.E. in 1993 and his Ph.D. in 1996, both in Electrical Engineering with a speciality in communications systems. In September 1996, he joined the Electrical and Computer Engineering Department at the University of Massachusetts, where he is currently an Associate Professor. Dr. Goeckel is the recipient of a 1999 CAREER Award from the National Science Foundation, and he is an Editor for the *IEEE Transactions on Wireless Communications*. His research interests are in the design of digital communication systems, particularly for wireless communication applications.

BIBLIOGRAPHY

1. E. Zehavi, 8-PSK trellis codes for a Rayleigh channel, *IEEE Trans. Commun.* **40**: 873–884 (May 1992).
2. G. Caire, G. Taricco, and E. Biglieri, Bit-interleaved coded modulation, *IEEE Trans. Inform. Theory* **44**: 927–945 (May 1998).
3. C. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (July 1948); **27**: 623–656 (Oct. 1948).
4. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **28**: 55–67 (Jan. 1982).
5. D. Divsalar and M. Simon, The design of trellis-coded MPSK for fading channels: Set partitioning for optimum code design, *IEEE Trans. Commun.* **36**: 1004–1011 (Sept. 1988).
6. C. Schlegel and D. Costello, Jr., Bandwidth efficient coding for fading channels: Code construction and performance analysis, *IEEE J. Select. Areas Commun.* **7**: 1356–1368 (Dec. 1989).
7. P. Örmeci, X. Liu, D. L. Goeckel, and R. D. Wesel, Adaptive bit-interleaved coded modulation, *IEEE Trans. Commun.* **49**: 1572–1581 (Sept. 2001).
8. S. Lin and D. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
9. S. Al-Semari and T. Fuja, I-Q TCM: reliable communication over the Rayleigh fading channel close to the cutoff rate, *IEEE Trans. Inform. Theory* **43**: 250–262 (Jan. 1997).
10. U. Fincke and M. Phost, Improved methods for calculating vectors of short length in a lattice, including a complexity analysis, *Math. Comput.* **44**(170): 463–471 (April 1985).
11. X. Li, A. Chindapol, and J. Ritcey, Bit-interleaved coded modulation with iterative decoding and 8PSK modulation, *IEEE Trans. Commun.* (in press).
12. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* 909–926 (May 1998).
13. D. L. Goeckel and G. Ananthaswamy, On the design of multidimensional signal sets for OFDM, *IEEE Trans. Commun.* **50**: 442–452 (March 2002).
14. U. Wachsmann, R. Fischer, and J. Huber, Multilevel codes: Theoretical concepts and practical design rules, *IEEE Trans. Inform. Theory* **45**: 1361–1391 (July 1999).
15. P. Schramm, Multilevel coding with independent decoding on levels for efficient communication on static and interleaved fading channels, *Proc. Personal, Indoor, and Mobile Radio Conf.* 1997, pp. 1186–1200.
16. U. Wachsmann, J. Huber, and P. Schramm, Comparison of coded modulation schemes for the AWGN and the Rayleigh fading channel, *Proc. Int. Symp. Information Theory*, 1998, p. 5.
17. C. Berrou and A. Glavieux, Near optimum limit error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* **44**: 1261–1271 (Oct. 1996).
18. G. Hellstern, Coded modulation with feedback decoding trellis codes, *Proc. IEEE Conf. Communications*, 1993, pp. 1071–1075.

BLIND EQUALIZATION TECHNIQUES

JITENDRA K. TUGNAIT
Auburn University
Auburn, Alabama

1. INTRODUCTION

Two major sources of impairments of digital communications signals as they propagate through analog channels (such as telephone, cable, and wireless radio) are multipath propagation and limited bandwidth, causing (linear) channel and signal distortions. Linear channel distortion leads to intersymbol interference (ISI) at the receiver which, in turn, may lead to high error rates in symbol detection. Equalizers are designed to compensate for these channel distortions. One may directly design an equalizer given the received signal, or one may first estimate the channel impulse response and then design an equalizer based on the estimated channel. The received signals are sampled at the baud (symbol) or higher (fractional) rate before processing them for channel estimation and/or equalization. Depending on the sampling rate, one has either a single-input/single-output (SISO) (baud rate sampling), or a single-input/multiple-output (SIMO) (fractional sampling), complex discrete-time equivalent baseband channel.

Traditionally, a training sequence (known to the receiver) is transmitted during startup (acquisition mode). In the operational stage, the receiver switches to a decision-directed mode where the previously equalized and detected symbols are used as a (pseudo)training sequence together with the received data to update the channel or the equalizer coefficients. The various issues involved and the tradeoffs among various competing approaches (linear, decision feedback, maximum-likelihood sequence estimation, least mean-square vs. recursive least-squares, baud rate vs. fractional rate, etc.) are fairly well understood and documented; see the well-known text by Proakis [21] and references cited therein. More recently, there has been much interest in blind (self-recovering) channel estimation and blind equalization where no training sequences

are available or used and the receiver starts up without any (explicit) cooperation from the transmitter. In point-to-multipoint networks, whenever a link from the server to one of the tributary stations is interrupted, it is clearly not feasible (or desirable) for the server to start sending a training sequence to reestablish a particular link. In broadcast applications such as FTTC (fiber-to-the-curb) and DSL (digital subscriber line), it is not desirable to require the transmitter to pause to train each client as it comes online. Transmission of periodic training sequences may incur costly overhead by diluting the transmission rate of the revenue-bearing content. It has also been argued [40] that a blind startup is more straightforward to implement than a startup that requires a training sequence; this eases interoperability issues among different manufacturers. In digital communications over fading/multipath channels, a restart is required following a temporary path loss due to a severe fade. During online transmission impairment monitoring, the training sequences are obviously not supplied by the transmitter.

As in the trained case, various approaches to blind channel estimation and equalization have been developed. When sampled at the baud rate, the received signal is discrete-time stationary and typically non-minimum-phase. When sampled at higher than baud rate (typically an integer multiple of baud rate), the signal is discrete-time scalar cyclostationary and equivalently, it may be represented as a discrete-time vector stationary sequence with an underlying SIMO model. With baud rate sampling, one has to exploit the higher-order statistics (HOS) of the received signal either implicitly (as in Refs. 12 and 26, where direct design of equalizers is considered) or explicitly (as in Refs. 13 and 33–36, where the focus is on first estimating the channel impulse response using higher-order cumulants of the received signal). Higher-order statistics provide an incomplete characterization of the underlying non-Gaussian process. Joint channel and data estimation using maximum-likelihood and related approaches (see Refs. 16 and 24 and references cited therein) exploit a complete (non-Gaussian) probabilistic characterization of the noisy signal. Computational complexity of these algorithms (explicit HOS and joint channel data estimation) is large when the ISI spans many symbols (as in telephone channels) but they are relatively simple when ISI span is short (as in mobile radio channels). However, they may suffer from local convergence problems.

When there is excess channel bandwidth, baud rate sampling is below the Nyquist rate leading to aliasing and depending on the symbol timing phase, in certain cases, causing deep spectral notches in sampled, aliased channel transfer function [11]. This renders the equalizer performance quite sensitive to symbol timing errors. Initially, in the trained case, fractional sampling was investigated to robustify the equalizer performance against timing error. More recently, in the blind context, it was discovered (see Ref. 29 and references cited therein) that oversampling provides some new information regarding the channel that can be exploited for blind channel estimation and equalization provided some technical conditions are satisfied (the "no common subchannel zeros" condition, also called channel disparity, for the underlying equivalent SIMO model). A similar SIMO model results if multiple sensors are used with or without fractional sampling. The work of Tong et al. [29] has spawned intense research activity in the use of second-order statistics for blind identification and equalization. It should be noted that the requisite technical conditions for applicability of these approaches are not always satisfied in practice; some examples are given in Ref. 34.

In this article, we will present a tutorial review of various approaches to single-user blind channel equalization and estimation. Our emphasis is on linear time-invariant channels; linear time-varying, as well as nonlinear channels are outside the scope of this article. The article is organized as follows. In Section 2 we present the relevant channel models and equalizer structures used later for discussion of blind equalization and channel estimation techniques. In Section 3, combined channel and symbol estimation methods are presented. Direct equalization and symbol estimation approaches without first or concurrently estimating the channel impulse response, are discussed in Section 4. In Section 5 various channel estimation approaches are presented. Commercial applications of blind equalization reported in the literature are briefly discussed in Section 6.

2. SYSTEMS MODELS

In this section we first describe the models that are used to characterize the wireless and mobile communications channels. Then we turn to a brief discussion of the various equalizer structures that are used to undo the signal distortions caused by the channel.

2.1. Channel Models

After some processing (e.g., matched filtering), the continuous-time received signals are sampled at the baud (symbol) or higher (fractional) rate before processing them for channel estimation and/or equalization. It is therefore convenient to work with an equivalent baseband discrete-time white-noise channel model [21, Sect. 10.1]. For a baud-rate sampled system, the equivalent baseband channel model is given by

$$y_k = \sum_{n=0}^{L} f_n I_{k-n} + w_k \tag{1}$$

where $\{w_k\}$ is a white Gaussian noise sequence with variance σ^2; $\{I_k\}$ is the zero-mean, independent and identically distributed (i.i.d.), information (symbol) sequence, possibly complex, taking values from a finite set; $\{f_k\}$ is an FIR (finite impulse response) linear filter (with possibly complex coefficients) that represents the equivalent channel; and $\{y_k\}$ is the (possibly complex) equivalent baseband received signal. A tapped delay line structure for this model is shown in Fig. 1.

The model (1) results in a single-input/single-output (SISO) complex discrete-time baseband-equivalent channel model. The output sequence $\{\hat{I}_k\}$ in Eq. (1) is discrete-time stationary. When there is excess channel

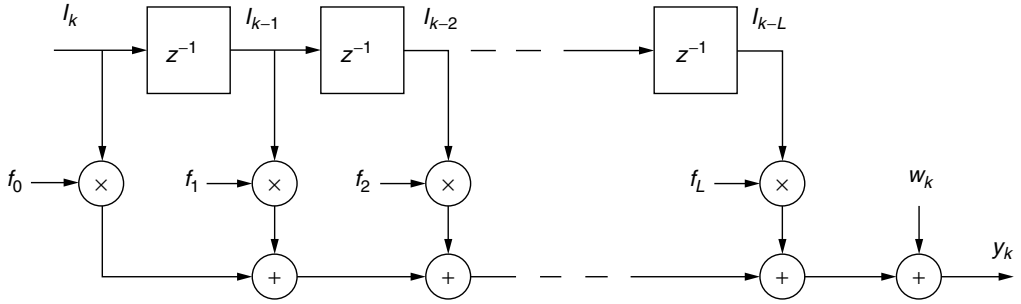


Figure 1. Tapped delay line model of the baud-rate channel.

bandwidth [bandwidth $> \frac{1}{2} \times$ (baud rate)], baud rate sampling is below the Nyquist rate leading to aliasing and depending on the symbol timing phase, in certain cases, causing deep spectral notches in sampled, aliased channel transfer function [11]. Linear equalizers designed on the basis of the baud-rate sampled channel response, are quite sensitive to symbol timing errors. Initially, in the trained case, fractional sampling was investigated to robustify the equalizer performance against timing errors. The model (1) does not apply to fractionally spaced samples, namely, when the sampling interval is a fraction of the symbol duration. The fractionally sampled digital communications signal is a cyclostationary signal [7] that may be represented as a vector stationary sequence using a time-series representation (TSR) ([7, Sec. 12.6]). Suppose that we sample at P times the baud rate with signal samples spaced T/P seconds apart where T is the symbol duration. Then a TSR for the sampled signal is given by

$$y_{ik} = \sum_{n=0}^{L} f_{in} I_{k-n} + w_{ik}; \quad (i = 1, 2, \ldots, P) \tag{2}$$

where now we have P samples every symbol period, indexed by i. Notice, however, that the information sequence I_k is still one "sample" per symbol. It is assumed that the signal incident at the receiver is first passed through a receive filter whose transfer function equals the square root of a raised-cosine pulse, and that the receive filter is matched to the transmit filter. The noise sequence in (2) is the result of the fractional rate sampling of a continuous-time, filtered white Gaussian noise process. Therefore, the sampled noise sequence is white at the symbol rate, but correlated at the fractional rate. Stack P consecutive received samples in the nth symbol duration to form a P vector $\mathbf{y}_k$ satisfying

$$\mathbf{y}_k = \sum_{n=0}^{L} \mathbf{f}_n I_{k-n} + \mathbf{w}_k \tag{3}$$

where $\mathbf{f}_n$ is the vector impulse response of the SIMO equivalent channel model given by

$$\mathbf{f}_n = [f_{1n} \quad f_{2n} \quad \cdots \quad f_{Pn}]^T \tag{4}$$

$$\mathbf{y}_k = [y_{1k} \quad y_{2k} \quad \cdots \quad y_{Pk}]^T \tag{5}$$

$$\mathbf{w}_k = [w_{1k} \quad w_{2k} \quad \cdots \quad w_{Pk}]^T \tag{6}$$

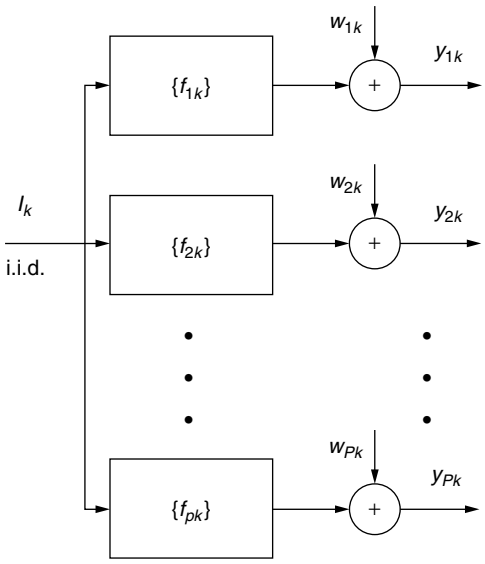


Figure 2. Block diagram of the fractionally sampled ($P \times$ baud-rate) channel.

[When $P = 2$, one way to look at the TSR model is to note that y_{1k} are "odd-numbered" fractionally spaced samples, y_{2k} are the "even-numbered" fractionally spaced samples and k indexes the baud (symbol); similarly for f_{ik}.] A block diagram of model (2) is shown in Fig. 2.

2.2. Equalizer Structures

The most common channel equalizer structure is a linear transversal filter. Given the baud-rate sampled received signal [see Eq. (1)] $\hat{I}_k$, the linear transversal equalizer output $\hat{I}_k$ is an estimate of I_k, given by

$$\hat{I}_k = \sum_{n=-N}^{N} c_n y_{k-n} \tag{7}$$

where $\{c_n\}_{n=-N}^{n=N}$ are the $(2N+1)$ tap-weight coefficients of the equalizer; see Fig. 3. As noted earlier, linear equalizers designed on the basis of the baud-rate sampled received signal, are quite sensitive to symbol timing errors [11]. Therefore, fractionally spaced linear equalizers (typically with twice the baud-rate sampling: oversampling by a factor of 2) are quite widely used to mitigate sensitivity to symbol timing errors. A fractionally

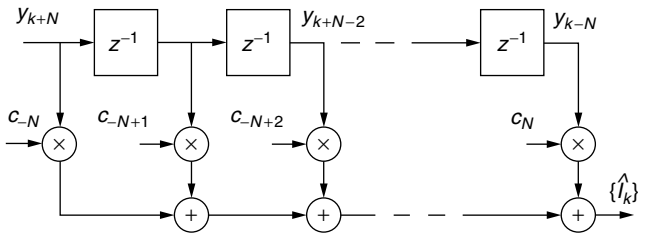


Figure 3. Structure of a baud-rate linear transversal equalizer.

spaced equalizer (FSE) in the linear transversal structure has the output

$$\hat{I}_k = \sum_{n=-N}^{N} \mathbf{c}_n^T \mathbf{y}_{k-n} = \sum_{n=-N}^{N} \left(\sum_{i=1}^{P} c_{in} y_{i(k-n)} \right) \tag{8}$$

where we have P samples per symbol, $\mathbf{y}_k$ and $\mathbf{c}_k$ are P-column vectors [cf. Eq. (3)], $\{\mathbf{c}_k\}$ are the $(2N+1)$ vector tap (or $P(2N+1)$ scalar tap) weight coefficients of the FSE, and the superscript T denotes the transpose operation. Note that the FSE outputs data at the symbol rate. Various criteria and cost functions exist to design the linear equalizers in both batch and recursive (adaptive) form; these are discussed later in this article. Figure 4 shows a block diagram of a generic FSE. [The transfer functions $F_i(z)$ and $C_i(z)$ are defined later in Eqs. (29) and (30).]

Linear equalizers do not perform well when the underlying channels have deep spectral nulls in the passband. Several nonlinear equalizers have been developed to deal with such channels. Two effective approaches are

- *The decision-feedback equalizer* (DFE), which is a nonlinear equalizer that employs previously detected symbols to eliminate the ISI due to the previously detected symbols on the current symbol to be detected. The use of the previously detected symbols makes the equalizer output a nonlinear function of the data. DFE can be symbol-spaced or fractionally spaced. Figure 5 is a block diagram of a DFE.
- *A maximum-likelihood sequence detector*, which estimates the information sequence to maximize the joint probability of the received sequence conditioned on the information sequence.

A detailed discussion may be found in the text by Proakis [21].

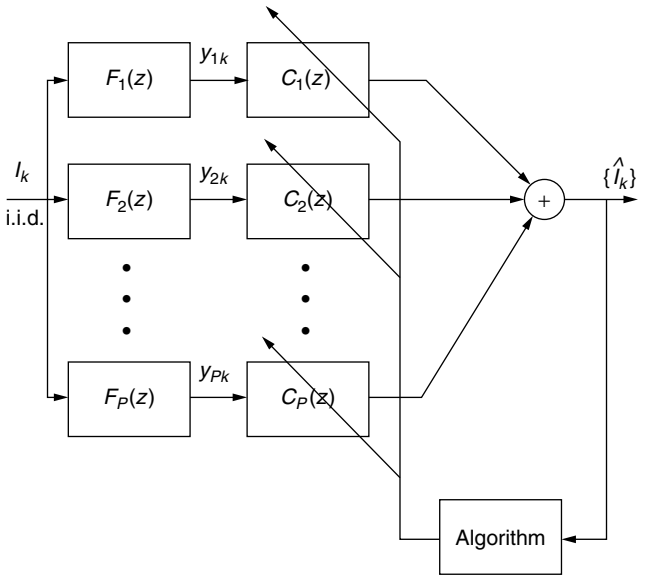


Figure 4. Block diagram of a fractionally spaced equalizer.

3. COMBINED CHANNEL AND SYMBOL ESTIMATION

In general, one of the most effective and popular parameter estimation algorithms is the maximum-likelihood (ML) method. The ML estimators can be derived in a systematic way. Perhaps more importantly, the class of ML estimators are optimal asymptotically. Not surprisingly, this class of algorithms has been applied to the blind equalization problem.

Let us consider the P-vector channel model given in Eq. (3). Suppose that we have collected M samples of the observation $Y = [\mathbf{y}_{M-1}^T, \ldots, \mathbf{y}_0^T]^T$. We then have the following linear model:

$$Y = \begin{pmatrix} I_{M-1}\mathcal{I}_P & I_{M-2}\mathcal{I}_P & \cdots & I_{M-L-1}\mathcal{I}_P \\ \vdots & \text{block} & \text{Hankel} & \text{matrix} \\ I_0\mathcal{I}_P & I_{-1}\mathcal{I}_P & \cdots & I_{-L}\mathcal{I}_P \end{pmatrix} \begin{pmatrix} \mathbf{f}_0 \\ \vdots \\ \mathbf{f}_L \end{pmatrix} + \begin{pmatrix} \mathbf{w}_{M-1} \\ \vdots \\ \mathbf{w}_0 \end{pmatrix} = \mathcal{H}(\mathbf{I})_{[MP]\times[P(L+1)]} F + W \tag{9}$$

where $\mathcal{I}_P$ is a $P \times P$ identity matrix; $\mathbf{I}$ and W are vectors consisting of samples of the input sequence $\{I_k\}$ and noise $\{\mathbf{w}_k\}$, respectively; F is the vector of the channel parameters; and a block Hankel matrix has identical block entries on its block antidiagonals.

Let θ be the vector of unknown parameters that may include the channel parameters F and possibly the entire or part of the input vector $\mathbf{I}$. Given the probability space that describes jointly the noise vector W and possibly the input data vector $\mathbf{I}$, we can then obtain, in principle, the probability density function (PDF) of the observation Y. As a function of the unknown parameter θ, the PDF of the observation $f(Y|\theta)$ is referred to as the *likelihood function*. The maximum likelihood estimator is defined by the following optimization

$$\hat{\theta} = \arg\max_{\theta \in \Theta} f(Y|\theta) \tag{10}$$

where Θ defines the domain of the optimization.

While the ML estimator is conceptually simple, and it usually has good performance when the sample size is sufficiently large, the implementation of ML estimator is sometimes computationally intensive. Furthermore, the optimization of the likelihood function in Eq. (10) is often hampered by the existence of local maxima. Therefore, it is desirable that effective initialization techniques are used in conjunction with the ML estimation. The simultaneous

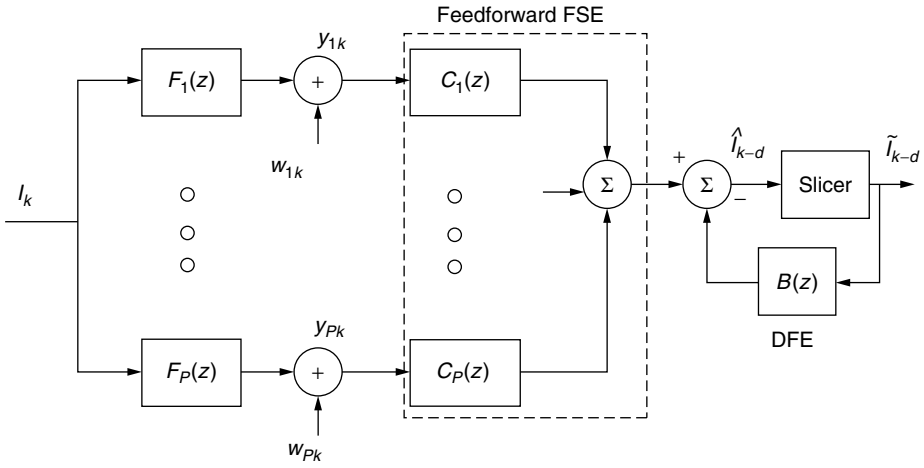


Figure 5. Feedforward and decision-feedback channel equalization filters.

estimation of the input vector and the channel appears to be ill-posed; how is it possible that the channel and its input can be distinguished using only the observation? The key in blind channel estimation is the utilization of qualitative information about the channel and the input. To this end, we consider two different types of maximum likelihood techniques based on different models of the input sequence.

3.1. Stochastic Maximum-Likelihood Estimation

While the input vector $\mathbf{I}$ is unknown, it may be modeled as a random vector with a known distribution. In such a case, the likelihood function of the unknown parameter $\theta = F$ can be obtained by

$$f(Y|F) = \int f(Y|\mathbf{I}, F) f(\mathbf{I}) d\mathbf{I} \tag{11}$$

where $f(\mathbf{I})$ is the marginal PDF of the input vector and $f(Y|\mathbf{I}, F)$ is the likelihood function when the input is known. Assume, for example, that the input data symbol I_k takes, with equal probability, a finite number of values. Consequently, the input data vector $\mathbf{I}$ also takes values from the signal set $\{\mathbf{I}_1, \ldots, \mathbf{I}_K\}$. The likelihood function of the channel parameters is then given by

$$\begin{aligned} f(Y|F) &= \sum_{i=1}^{K} f(Y|\mathbf{I}_i, F) \text{ Prob}(\mathbf{I} = \mathbf{I}_i) \\ &= C \sum_{i=1}^{K} \exp\left\{ -\frac{\|Y - \mathcal{H}(\mathbf{I}_i)F\|^2}{2\sigma^2} \right\} \end{aligned} \tag{12}$$

where C is a constant, $\|Y\|^2 := Y^H Y$, Y^H is the complex conjugate transpose of the complex vector Y, and the stochastic MLE is given by

$$\hat{F} = \arg\min_F \sum_{i=1}^{K} \exp\left\{ -\frac{\|Y - \mathcal{H}(\mathbf{I}_i)F\|^2}{2\sigma^2} \right\} \tag{13}$$

The maximization of the likelihood function defined in (11) is in general difficult because $f(Y|\theta)$ is nonconvex. The expectation–maximization (EM) algorithm can be applied to transform the complicated optimization to a sequence of quadratic optimizations. Kaleh and Vallet [16] first applied the EM algorithm to the equalization of communication channels with input sequence having finite alphabet property. By using a *hidden Markov model* (HMM), they developed a batch (offline) procedure that includes the so-called forward and backward recursions. Unfortunately, the complexity of this algorithm increases exponentially with the channel memory.

To relax the memory requirements and facilitate channel tracking, "online" sequential approaches have been proposed [17] for input with finite alphabet properties under a HMM formulation. Given the appropriate regularity conditions and a good initialization guess, it can be shown that these algorithms converge to the true channel value.

3.2. Deterministic Maximum-Likelihood Estimation

The deterministic ML approach assumes no statistical model for the input sequence $\{I_k\}$. In other words, both the channel vector F and the input source vector $\mathbf{I}$ are parameters to be estimated. When the noise is zero-mean Gaussian with covariance $\sigma^2 I$, the ML estimates can be obtained by the nonlinear least squares optimization

$$\{\hat{F}, \hat{\mathbf{I}}\} = \arg\min \|Y - \mathcal{H}(\mathbf{I})F\|^2. \tag{14}$$

Joint minimization of the likelihood function with respect to both the channel and the source parameter spaces is difficult. Fortunately, the observation vector Y is linear in both the channel and the input parameters individually. In particular, we have

$$Y = \mathcal{H}(\mathbf{I})F + W = \mathcal{T}(F)\mathbf{I} + W \tag{15}$$

where

$$\mathcal{T}(F) = \begin{pmatrix} \mathbf{f}_0 & \cdots & \mathbf{f}_L & & \\ & \ddots & & \ddots & \\ & & \mathbf{f}_0 & \cdots & \mathbf{f}_L \end{pmatrix} \tag{16}$$

is the so-called filtering matrix. We therefore have a separable nonlinear least squares problem that can be

solved sequentially:

$$\{\hat{F}, \hat{\mathbf{I}}\} = \arg\min_{\mathbf{I}}\{\min_{F} \|Y - \mathcal{H}(\mathbf{I})F\|^2\} \tag{17}$$

$$= \arg\min_{F}\{\min_{\mathbf{I}} \|Y - \mathcal{T}(F)\mathbf{I}\|^2\} \tag{18}$$

If we are interested only in estimating the channel, the minimization described above can be rewritten as

$$\hat{F} = \arg\min_{F} \| \underbrace{(I - \mathcal{T}(F)\mathcal{T}^{\dagger}(F))}_{\mathcal{P}(F)} Y\|^2 = \arg\min_{F} \|\mathcal{P}(F)Y\|^2 \tag{19}$$

where $\mathcal{P}(F)$ is a projection transform of Y into the orthogonal complement of the range space of $\mathcal{T}(F)$, or the noise subspace of the observation, and $\mathcal{T}^{\dagger}(F)$ denotes the pseudoinverse of $\mathcal{T}(F)$. Tong and Perreau have discussed algorithms of this type [28].

Similar to the HMM for statistical maximum-likelihood approach, the finite alphabet properties of the input sequence can also be incorporated into the deterministic maximum-likelihood methods. These algorithms, first proposed by Seshadri [24] and Ghosh and Weber [9], iterate between estimates of the channel and the input. At iteration k, with an initial guess of the channel $F^{(k)}$, the algorithm estimates the input sequence $\mathbf{I}^{(k)}$ and the channel $F^{(k+1)}$ for the next iteration by

$$\mathbf{I}^{(k)} = \arg\min_{\mathbf{I}\in\mathcal{S}} \|Y - \mathcal{T}(F^{(k)})\mathbf{I}\|^2 \tag{20}$$

$$F^{(k+1)} = \arg\min_{F} \|Y - \mathcal{H}(\mathbf{I}^{(k)})F\|^2 \tag{21}$$

where $\mathcal{S}$ is the (discrete) domain of $\mathbf{I}$. The optimization in (21) is a linear least-squares problem whereas the optimization in (20) can be achieved by using the Viterbi algorithm [21]. Seshadri [24] presented blind trellis search techniques. Reduced-state sequence estimation was proposed by Ghosh and Weber [9]. Raheli et al. proposed a per survivor processing technique [22].

The convergence of such approaches is not guaranteed in general. Interesting examples have been provided [5] in which two different combinations of F and $\mathbf{I}$ lead to the same cost $\|Y - \mathcal{H}(\mathbf{I})F\|^2$.

4. DIRECT EQUALIZATION AND SYMBOL ESTIMATION

In this section, we describe several types of approaches to the problem of direct input signal recovery under linear time-invariant channels. We first outline the basic principle of blind adaptive equalization based on implicit HOS criteria. Next, we explain the principle of some simple algorithms for blind symbol estimation exploiting second-order statistics. Finally, we discuss the method of symbol estimation via iterative least square criterion and some variations.

4.1. SISO Blind Equalization Based on HOS

In this subsection we consider baud-rate data and equalizers. In the case of known training sequence transmission, the linear equalizer tap $\mathbf{c}_n$ values are chosen to minimize the cost $E\{|\hat{I}_k - I_k|^2\}$ where $\{I_k\}$ is the training sequence. In the blind case, there is no training sequence. The key to designing a blind equalizer is to design rules of equalizer parameter adjustment. With the lack of training sequence, the receiver does not have access to the desired equalizer output I_k to adopt the traditional minimum mean-square-error criterion. Evidently, blind equalizer adaptation needs to minimize some special, non-mean-square-error (MSE)-type cost function, which implicitly involves higher order statistics of the channel output signal. The design of the blind equalizer thus translates into defining a *mean-cost function* $E\{\Psi(\hat{I}_k)\}$, where $\Psi(x)$ is a scalar function. Thus, the stochastic gradient descent minimization algorithm is easily determined by the derivative function $\psi(x) := \Psi'(x) := d\Psi(x)/dx$. Hence, a blind equalizer can either be defined by the cost function $\Psi(x)$, or equivalently, by its derivative $\psi(x)$ function. Ideally, the function $\Psi(\cdot)$ should be selected such that local minima of the mean cost correspond to a significant removal of ISI in the equalizer output $\hat{I}_k$.

Let

$$\mathbf{C} := [\,\mathbf{c}_{-N}^T \quad \mathbf{c}_{-N+1}^T \quad \cdots \quad \mathbf{c}_N^T\,]^T \tag{22}$$

Let $\mathbf{C}^{(k)}$ denote the value of $\mathbf{C}$ at the kth iteration. Then a stochastic gradient algorithm for the adaptation of $\mathbf{C}$ is given by

$$\mathbf{C}^{(k+1)} = \mathbf{C}^{(k)} - \alpha\nabla_{\mathbf{C}^{(k)}}\Psi(\hat{I}_k) \tag{23}$$

where $\nabla_{\mathbf{C}}\Psi$ denotes the gradient of Ψ with respect to the tap vector $\mathbf{C}$ and $\alpha > 0$ is the step-size parameter [21].

We now summarize several blind adaptation algorithms designed for feedforward equalizers.

4.1.1. Decision-Directed Algorithm. The simplest blind equalization algorithm is the decision-directed algorithm without training sequence. It minimizes the mean-square error between equalizer output $\hat{I}_k$ and the slicer output $\tilde{I}_{k-d}$. The performance of decision-directed algorithm depends on how close the initial parameters are to their optimum settings. The closer they are, the more accurate the slicer output is to the true channel input I_{k-d}. On the other hand, local convergence is highly likely if initial parameter values cause significant number of slicer errors [19].

4.1.2. Sato Algorithm and Some Generalizations. The first truly blind algorithm was introduced by Sato [23]. For M-level PAM channel input, this is defined by

$$\psi(x) = x - R_1\text{sgn}(x), \qquad \text{where} \qquad R_1 := \frac{E|I_k|^2}{E|I_k|} \tag{24}$$

The Sato algorithm was extended by Benveniste et al. [1] into a class of error functions given by

$$\psi_b(\hat{I}_k) = \psi_a(\hat{I}_k) - R_b\text{sgn}(\hat{I}_k), \qquad \text{where}$$
$$R_b := \frac{E\{\psi_a(I_k)I_k\}}{E|I_k|} \tag{25}$$

The generalization uses an odd function $\psi_a(x)$ whose second derivative is nonnegative for $x \geq 0$.

4.1.3. Constant-Modulus Algorithm and Extensions. The best known blind algorithms were presented elsewhere [12,31] with cost functions

$$\Psi_q(x) = \frac{1}{2q}\left(|x|^q - R_q\right)^2, \qquad \text{where}$$

$$R_q := \frac{E|I_k|^{2q}}{E|I_k|^q}, \quad q = 1, 2, \ldots \tag{26}$$

This class of *Godard algorithms* is indexed by the positive integer q. Using the stochastic gradient descent approach, equalizer parameters can be adapted accordingly.

For $q = 2$, the special Godard algorithm was developed as the *constant-modulus algorithm* (CMA) independently by Treichler and co-workers [31] using the philosophy of property restoral. For channel input signal that has a constant modulus $|I_k|^2 = R_2$, the CMA equalizer penalizes output samples $\hat{I}_k$ that do not have the desired constant modulus characteristics. The modulus error is simply $e_k = |\hat{I}_k|^2 - R_2$, and the squaring of this error yields the constant-modulus cost function that is the identical to the Godard cost function with $q = 2$.

This modulus restoral concept has a particular advantage in that it allows the equalizer to be adapted independent of carrier recovery. A carrier frequency offset of Δ_f causes a possible phase rotation of the equalizer output. Because the CMA cost function is insensitive to the phase of $\hat{I}_k$, the equalizer parameter adaptation can occur independently and simultaneously with the operation of the carrier recovery system. This property also allows CMA to be applied to analog modulation signals with constant amplitude such as those using frequency or phase modulation [31]. Practical implementations and theoretical properties of the CMA equalizers are discussed in the literature [15,32,40]. Blind DFE has also been considered [4]; in the absence of reliable initialization, blind DFEs can be unstable and can easily misconverge.

The methods of Shalvi and Weinstein [25] generalize CMA and are explicitly based on higher-order statistics of the equalizer output. Define the kurtosis of the equalizer output signal $\hat{I}_k$ as

$$K_{\hat{I}} := E|\hat{I}_k^4| - 2E^2|\hat{I}_k^2| - |E\{\hat{I}_k^2\}|^2 \tag{27}$$

The Shalvi–Weinstein algorithm maximizes $|K_{\hat{I}}|$ subject to the power constraint $E\{|\hat{I}_k|^2\} = E\{|I_k|^2\}$. Superexponential iterative methods have been presented [26] in which a superexponential convergence rate in the absence of noise has been established for the linear equalizer. A deconvolution-based approach can be also be found [2]. Werner et al. [40] discuss modifications of CMA, called the *multimodulus algorithm* (MMA) and generalized MMA, for high-order QAM (quadrature amplitude modulation) and CAP (carrierless amplitude and phase) signals.

4.2. SIMO Equalization and Symbol Estimation

We now consider fractionally sampled data and equalizers. Any adaptive blind equalization algorithm can be easily adopted for linear SIMO equalizers [18]. SIMO blind equalization may offer a convergence advantage given the subchannel diversity [15]. While algorithms such as CMA in SISO equalization may suffer from local convergence [15], CMA and the superexponential method [26] are shown to converge to complete ISI removal under noiseless channels [18]. Furthermore, there is a close relationship between CMA and the nonblind minimum mean-square-error (MMSE) equalizer [42].

Consider the FSE shown in Fig. 4. If the baud-rate "subchannel" transfer functions $F_i(z)$, $1 \le i \le P$, have no common zeros [i.e., there exists no complex number ρ for which $F_i(\rho) = 0$ for every $i, i = 1, 2, \ldots, P$], then there exist FIR "subequalizers" $C_i(z)$s such that

$$\sum_{i=1}^{P} C_i(z)F_i(z) = z^{-d} \tag{28}$$

where

$$C_i(z) := \sum_{n=-N}^{N} c_{in}z^{-n}, \quad 2N \ge L - 1 \tag{29}$$

$$F_i(z) := \sum_{n=0}^{N} f_{in}z^{-n} \tag{30}$$

and d is an integer $\ge -N$. This relation implies perfect equalization (i.e., complete removal of ISI), in the absence of noise, using FIR equalizers, which is not possible in the SISO (baud-rate) case.

If Eq. (28) is not satisfied, then perfect equalization is not possible (using FIR equalizers) and there may be local convergence problems [15].

4.3. Iterative Blind Symbol Estimation

The iterative channel and symbol estimation method, as summarized in Section 3.2, also allows direct channel input estimation. Iterative least-squares with enumeration (ILSE) and Iterative least-squares with projection (ILSP) methods both exploit the finite alphabet nature of the channel input signals. Given that elements in $\mathbf{I}$ come from $\mathcal{S}$, the task of implementing

$$\min_{F,\mathbf{I}\in\mathcal{S}} \|\mathcal{H}(Y) - \mathcal{T}(F)\mathcal{H}(\mathbf{I})\|^2 \tag{31}$$

can be iteratively implemented to improve the estimate in each step, as in Eqs. (20) and (21). ILSP simply replaces the complex symbol estimation step of (20) by a simpler projection [27]

$$\mathbf{I}^{(k)} = \text{proj}_{\mathcal{S}}\left(\mathcal{T}(F^{(k)})^{\dagger}Y\right). \tag{32}$$

5. BLIND CHANNEL ESTIMATION

Although the ML channel estimator discussed in Section 3 usually provides better performance, the computation complexity and the existence of local optima are the two major difficulties. Therefore, "simpler" approaches have also been investigated.

5.1. SISO Channel Estimation

For baud-rate data, second-order statistics of the data do not carry enough information to allow estimation of the channel impulse response as a typical channel is nonminimum-phase. On the other hand, higher-order statistics (in particular, fourth-order cumulants) of the baud-rate (or fractional rate) data can be exploited to yield the channel estimates to within a scale factor.

Given the mathematical model (1), there are two broad classes of approaches to channel estimation, the distinguishing feature among then being the choice of the optimization criterion. All of the approaches involve (more or less) a least-squares-error measure. The error definition differs, however, as follows:

- *Fitting Error*. Match the model-based higher-order (typically fourth-order) statistics to the estimated (data-based) statistics in a least-squares sense to estimate the channel impulse response, as in Refs. 35 and 36, for example. This approach allows consideration of noisy observations. In general, it results in a nonlinear optimization problem. It requires availability of a good initial guess to prevent convergence to a local minimum. It yields estimates of the channel impulse response. The estimator obtained by minimizing Eq. (44) is a fitting error estimate.
- *Equation Error*. This technique is based on minimizing an "equation error" in some equation that is satisfied ideally. The approaches of Refs. 13 and 39 (among others) fall in this category. In general, this class of approaches results in a closed-form solution for the channel impulse response so that a global extremum is always guaranteed provided the channel length (order) is known. These approaches may also provide good initial guesses for the nonlinear fitting error approaches. Quite a few of these approaches fail if the channel length is unknown. The estimator in Eq. (38) is an equation error estimate.

Further details may be found in Ref. 38 and references cited therein.

We now briefly consider the approach of Ref. 35 to illustrate the basic ideas.

5.1.1. Cumulant Matching. We wish to estimate the channel impulse response via a fitting error approach using fourth (and second)-order cumulants of the noisy data. Our main objective is to minimize the cost (44) discussed later in this section. Since this optimization problem requires a good initial guess, we first discuss a simple equation error approach coupled with an model order selection procedure.

Denote the fourth (joint) cumulant of the complex random variables $y^*_{k+\tau_1}$, $y_{k+\tau_2}$, $y^*_{k+\tau_3}$, and y_k as $C_4(\tau_1, \tau_2, \tau_3)$ given by (the superscript $*$ denotes the complex conjugation operation)

$$\begin{aligned} C_4(\tau_1, \tau_2, \tau_3) = {} & E\{y_k y^*_{k+\tau_1} y_{k+\tau_2} y^*_{k+\tau_3}\} - E\{y_k y^*_{k+\tau_1}\} \\ & \times E\{y_{k+\tau_2} y^*_{k+\tau_3}\} - E\{y_k y_{k+\tau_2}\} E\{y^*_{k+\tau_1} y^*_{k+\tau_3}\} \\ & - E\{y_k y^*_{k+\tau_3}\} E\{y_{k+\tau_2} y^*_{k+\tau_1}\} \end{aligned} \tag{33}$$

Then it can be shown that [γ_I = fourth cumulant (kurtosis) of I_k] for model (1) we have

$$C_4(\tau_1, \tau_2, \tau_3) = \gamma_I \sum_{k=0}^{L} f^*_{k+\tau_1} f_{k+\tau_2} f^*_{k+\tau_3} f_k \tag{34}$$

In particular, we have

$$C_4(L, \tau, \tau_1) = \gamma_I f^*_L f_\tau f^*_{\tau_1} f_0 \tag{35}$$

It then follows that

$$C_4(L, 0, \tau_1) f_\tau = C_4(L, \tau, \tau_1) \qquad \text{for} \qquad 0 \le \tau_1 \le L \tag{36}$$

Assuming that $f_L \neq 0$, this immediately leads to the least-squares solution

$$f_\tau = \frac{\sum_{\tau_1=0}^{L} C^*_4(L, 0, \tau_1) C_4(L, \tau, \tau_1)}{\sum_{\tau_1=0}^{L} |C_4(L, 0, \tau_1)|^2} \quad \text{for } 1 \le \tau \le L \tag{37}$$

In practice, true cumulants of the data are unknown. Therefore, we replace them with their consistent estimates. Let $\hat{C}_{4N}(i, j, k)$ denote an estimate of $C_4(i, j, k)$ obtained from Eq. (33) by replacing the moments in (33) by their respective sample averages, based on the N data samples; see Ref. 33 for more details. Then we have

$$\hat{f}_\tau = \frac{\sum_{\tau_1=0}^{L} \hat{C}^*_{4N}(L, 0, \tau_1) \hat{C}_{4N}(L, \tau, \tau_1)}{\sum_{\tau_1=0}^{L} |\hat{C}_{4N}(L, 0, \tau_1)|^2} \quad \text{for } 1 \le \tau \le L \tag{38}$$

In general, there is no guarantee that $f_L \neq 0$ in (1). If $f_L = 0$ in (1) (if, e.g., we overfit), then $C_4(L, i, k) = 0$ for every $0 \le i, k \le L$ rendering the estimates (37–38) useless. Therefore, we perform a search over all possible values of true order L of the FIR model (1); that is, we search over the range $0 \le L \le \overline{L}$, where $\overline{L}$ is an upper bound on the model order such that the true order is known to be less than or equal to $\overline{L}$. Denote the ith coefficient of an MA(L) model (moving-average model of order L) as $f_{i,L}$ so that $f_{0,L} := 1$ for every L and $f_{i,L} := 0$ for $L + 1 \le i \le \overline{L}$ for $L < \overline{L}$. Estimate $f_{i,L}$ by $\hat{f}_{i,L}(N)$ as

$$\hat{f}_{i,L}(N) = \frac{\sum_{\tau_1=0}^{L} \hat{C}^*_{4N}(L, 0, \tau_1) \hat{C}_{4N}(L, \tau, \tau_1)}{\sum_{\tau_1=0}^{L} |\hat{C}_{4N}(L, 0, \tau_1)|^2 + \Delta}, 1 \le i \le L \tag{39}$$

where $\Delta > 0$ is a "small" number. Define a correlation coefficient as

$$\rho_{L,\overline{L}}(N) = \frac{|\sum_{i=0}^{\overline{L}} \sum_{k=0}^{i} \sum_{l=0}^{\overline{L}} \hat{C}_{4N}(i, l, k) \tilde{C}^*_4(i, l, k|\theta_L)|}{\hat{P} P_{\theta_L}} \tag{40}$$

where the lags in (40) range over the nonredundant lag region for complex MA($\overline{L}$) models, $\theta_L := [\hat{f}_{1,L}(N), \ldots, \hat{f}_{L,L}(N)]^T$, the normalized ($\gamma_I = 1$) theoretical fourth-order cumulants corresponding to the parameter vector θ_L are given by

$$\tilde{C}_4(i,l,k|\theta_L) = \sum_{m=0}^{L} \hat{f}_{m,L}(N)\hat{f}^*_{m+i,L}(N)\hat{f}_{m+l,L}(N)\hat{f}^*_{m+k,L}(N) \quad (41)$$

$$P_{\theta_L} := \sqrt{\sum_{i=0}^{\overline{L}} \sum_{k=0}^{i} \sum_{l=0}^{\overline{L}} |\tilde{C}_4(i,l,k|\theta_L)|^2 + \Delta} \quad (42)$$

and

$$\hat{P} := \sqrt{\sum_{i=0}^{\overline{L}} \sum_{k=0}^{i} \sum_{l=0}^{\overline{L}} |\hat{C}_{4N}(i,l,k)|^2} \quad (43)$$

Thus it is easy to see that the preceding correlation coefficient is a measure of fit between the data-based cumulants and the theoretical cumulants obtained from the fitted model.

The estimation of the FIR parameters proceeds as follows. Perform the computations (37)–(43) for $0 \le L \le \overline{L}$. Pick that value of L as the correct FIR order that leads to a maximum correlation coefficient (40); denote it by $\hat{L}_{\overline{L}}$. Repeat (39) with $L = \hat{L}_{\overline{L}}$ and $\Delta = 0$ yielding the desired FIR parameter estimates noting that $\hat{f}_{i,\hat{L}_{\overline{L}}}(N) := 0$ for $\hat{L}_{\overline{L}} + 1 \le i \le \overline{L}$. To justify this procedure asymptotically, let L_0 be the true order such that $0 \le L_0 \le \overline{L}$. As $N \to \infty$, it follows from Ref. 33 that $\hat{C}_{4N}(j,i,k) \to C_4(j,i,k)$ (with probability 1) for any i, j, k. If $L < L_0$, then clearly $\rho_{L,\overline{L}}(N) < 1$ for large N because $f_{L_0,L_0} \ne 0$ by assumption whereas $f_{L_0,L} = 0$, also by assumption. If $L > L_0$, then $\hat{C}_{4N}(L,i,k) \to 0$ w.p.1 as $N \to \infty$ so that $\hat{f}_{i,L}(N) \to 0$ w.p.1 for $1 \le i \le L$, leading to $\rho_{L,\overline{L}}(N) \to \delta < 1$. When $L = L_0$, then $\hat{f}_{i,L}(N) \to \tilde{f}_{i,L}$ as $N \to \infty$, such that as $\Delta \to 0$, $\tilde{f}_{i,L} \to f_{i,L_0}$. Therefore, for Δ small enough, $\rho_{L_0,\overline{L}}(N) \approx 1$ for large N. Thus, asymptotically, correct model order will be selected.

Using the preceding estimates as the initial guess, the next step is to refine the channel estimates by minimizing a quadratic cumulant matching criterion [33]. Let $\overline{L}$ denote the upper bound on the FIR model order as before. Let θ denote the vector of all unknown system parameters given by $\theta := (f_0, f_1, \ldots, f_{\overline{L}}, \gamma_I, \sigma_I)$, $f_{i_0} := 1$ for some $0 \le i_0 \le \overline{L}$, where γ_I and σ_I^2 are the fourth cumulant and the second cumulant (variance) of the information sequence I_k (where one of the channel coefficients has been arbitrarily fixed at 1.0). Choose θ to minimize

$$\sum_{\tau_1=0}^{\overline{L}} \sum_{\tau_3=0}^{\tau_1} \sum_{\tau_2=0}^{\overline{L}} |\hat{C}_{4N}(\tau_1,\tau_2,\tau_3) - C_4(\tau_1,\tau_2,\tau_3|\theta)|^2 + \lambda \sum_{\tau=0}^{\overline{L}} |\hat{R}_N(\tau) - R(\tau|\theta)|^2 \quad (44)$$

where $\hat{C}_{4N}(-)$ denotes the data-based cumulant estimates, $C_4(-|\theta)$ denotes the theoretical cumulants obtained from the hypothesized model, $\hat{R}_N(-)$ denotes the data-based correlation estimates, $R(-|\theta)$ denotes the theoretical correlations obtained from the hypothesized model, and the weighting factor λ is designed to make the cost function invariant to any scale changes. The factor λ is chosen as [33]

$$\lambda := \lambda_0 \frac{\sum_{\tau_1=0}^{\overline{L}} \sum_{\tau_3=0}^{\tau_1} \sum_{\tau_2=0}^{\overline{L}} |\hat{C}_4(\tau_1,\tau_2,\tau_3)|^2}{\sum_{\tau=0}^{\overline{L}} |\hat{R}(\tau)|^2} \quad (45)$$

where $\lambda_0 > 0$ determines the relative weighting between the correlations and the fourth-order cumulants.

The initial guess for minimization of (44) is obtained from the linear estimator. If the selected order in the linear approach is $\hat{L}_{\overline{L}}$, then we set $m_0 = \lfloor (\overline{L} - \hat{L}_{\overline{L}})/2 \rfloor$, $f_n = 0$ for $0 \le n \le m_0 - 1$, $f_{i+m_0} = \hat{f}_{i,\hat{L}_{\overline{L}}}(N)$ for $0 \le i \le \hat{L}_{\overline{L}}$, and $f_n = 0$ for $1 + m_0 + \hat{L}_{\overline{L}} \le n \le \overline{L}$; that is, we "center" the result of the linear approach to obtain the initial guess.

The estimators of the correlation and the fourth-order cumulant functions obtained via appropriate sample averaging of data are strongly consistent [33]. By the preceding results [such as (38)] and the cost function (44) (see also Ref. 33), it follows that the parameter estimator minimizing (44) is strongly consistent provided that $\overline{L} \ge$ true length of the channel.

5.2. SIMO Channel Estimation

Here we concentrate on second-order statistical methods. For single-input (SIMO) multiple-output vector channels the autocorrelation function of the observation is sufficient for the identification of the channel impulse response up to an unknown constant [29,34], provided the various subchannels have no common zeros. This observation led to a number of techniques under both statistical and deterministic assumptions of the input sequence [28]. By exploiting the multichannel aspects of the channel, many of these techniques lead to a constrained quadratic optimization

$$\hat{F} = \arg \min_{\|F\|=1} F^H Q(Y) F \quad (46)$$

where $Q(Y)$ is a positive definite matrix constructed from the observation. Asymptotically (either as the sample size increases to infinity or the noise variance approaches to zero), these estimates converge to true channel parameters.

5.2.1. The Cross-Relation Approach. Here we present a simple yet informative approach [41] that illustrates the basic idea. Suppose that we have only two channels with finite impulse responses f_{1n} and f_{2n}, respectively. If there is no noise, the received signals from the two channels satisfy

$$y_{1n} = f_{1n} * I_n, y_{2n} = f_{2n} * I_n \quad (47)$$

where $*$ is the linear convolution. Consequently, we must have

$$y_{1n} * f_{2n} = y_{2n} * f_{1n} \quad (48)$$

Since the convolution operation is linear with respect to the channel and y_{in} $(i = 1, 2)$ are available, Eq. (48) is equivalent to solving a homogeneous linear equation

$$R\tilde{F} = 0 \tag{49}$$

where R is constructed from the M received data samples

$$R = Y_2 - Y_1 \tag{50}$$

$$Y_j := \begin{pmatrix} y_{jL} & y_{j(L-1)} & \cdots & y_{j0} \\ y_{j(L+1)} & y_{jL} & \cdots & y_{j1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{j(M-1)} & y_{j(M-2)} & \cdots & y_{j(M-L-1)} \end{pmatrix} \tag{51}$$

$$\tilde{F} := (f_{10} \quad f_{11} \quad \cdots \quad f_{1L} \quad f_{20} \quad \cdots \quad f_{2L})^T \tag{52}$$

It can be shown that under certain identifiability conditions [28] (which include knowledge of L and no common subchannel zeros), the null space of R has dimension 1, which means that the channel can be identified up to a constant. When there is noise, the channel estimator can be obtained from a constrained quadratic optimization

$$\hat{\tilde{F}} = \arg\min_{\|\tilde{F}\|=1} \tilde{F}^H R^H R \tilde{F} \tag{53}$$

which implies that $\hat{\tilde{F}}$ is the eigenvector corresponding to the smallest eigenvalue of $Q = R^H R$.

Hua [14] has shown that the cross-relation method combined with the ML approach offers performance close to the Cramer–Rao lower bound. The main problem with this method is that the channel length L needs to be accurately known (in addition to the no-common-subchannel-zeros condition).

5.2.2. Noise Subspace Approach. Alternatively, one can also exploit the subspace structure of the filtering matrix. We now consider a method proposed by Moulines et al. [20]. Define the $M \times [M + L]$ filtering matrix

$$\mathcal{T}_{M+L}(\mathbf{f}_l) = \begin{pmatrix} f_{l0} & \cdots & f_{lL} & & \\ & \ddots & & \ddots & \\ & & f_{l0} & \cdots & f_{lL} \end{pmatrix} \tag{54}$$

and the $[PM] \times [M + L]$ multichannel filtering matrix

$$\mathcal{T}_{M+L}(F) = \left(\mathcal{T}^T_{M+L}(\mathbf{f}_1) \quad \mathcal{T}^T_{M+L}(\mathbf{f}_2) \quad \cdots \quad \mathcal{T}^T_{M+L}(\mathbf{f}_P)\right)^T \tag{55}$$

Define $(M \geq L)$

$$\mathbf{Y}_n = \left(\mathbf{Y}^T_{1n} \quad \mathbf{Y}^T_{2n} \quad \cdots \quad \mathbf{Y}^T_{Pn}\right)^T$$

where

$$\mathbf{Y}_{in} = \left(y_{in} \quad y_{i(n-1)} \quad \cdots \quad y_{i(n-M+1)}\right)^T \tag{57}$$

Then the correlation matrix $\mathbf{R} = E\{\mathbf{Y}_n \mathbf{Y}_n^H\}$ has an eigenvalue decomposition (EVD)

$$\mathbf{R} = \sum_{k=1}^{PM} \lambda_k \, \mathbf{q}_k \, \mathbf{q}_k^H \tag{58}$$

where λ_ks are in the descending order of magnitude. It can be shown that the range space of $\mathbf{R}$ (signal subspace), also the range space of $\mathcal{T}_{M+L}(F)$, is spanned by the eigenvector $\mathbf{q}_k$ values for $k = 1, 2, \ldots, L + M$ whereas the noise subspace (orthogonal complement of the range space) is spanned by the remaining $\mathbf{q}_k$ values for $k = L + M + 1, L + M + 2, \ldots, PM$.

Define $\mathbf{g}_k = \mathbf{q}_{L+M+k+1}$ for $k = 0, 1, \ldots, PM - L - M - 1$. It then follows that

$$\mathcal{T}^H_{M+L}(F)\mathbf{g}_k = 0 \quad k = 0, 1, \ldots, PM - L - M - 1 \tag{59}$$

The vectors $\mathbf{g}_K$ values can be estimated from data via estimated correlation matrix $\mathbf{R}$ and its EVD. Partition the $PM-$ vector $\mathbf{g}_k$ as

$$\mathbf{g}_k = \left(\mathbf{g}^T_{1k} \quad \cdots \quad \mathbf{g}^T_{Pk}\right)^T \tag{60}$$

to conform to $\mathcal{T}_{M+L}(F)$, where $\mathbf{g}_{ik}$ is $M \times 1$. For a given k, define the $[L + 1] \times [L + M]$ matrix $\mathcal{T}_{M+L}(\mathbf{g}_{lk})$ just as $\mathcal{T}_{M+L}(\mathbf{f}_l)$ in (54) except for replacing $\mathbf{f}_l$ with $\mathbf{g}_{lk}$, and similarly define $\mathcal{T}_{M+L}(\mathbf{g}_k)$ by mimicking $\mathcal{T}_{M+L}(F)$ in (55). It has been shown by [20] that

$$\mathcal{T}^H_{M+L}(F)\mathbf{g}_k = 0 = \mathcal{T}^H_{M+L}(\mathbf{g}_k)F \tag{61}$$

It has been further shown [20] that under the knowledge of L and no common subchannel zeros, the channel F can be estimated (up to a scale factor) by the optimization problem

$$\hat{F} = \arg\min_{\|F\|=1} F^H \mathcal{Q} F \quad \text{where}$$

$$\mathcal{Q} := \sum_{k=0}^{PM-L-M-1} \mathcal{T}_{M+L}(\mathbf{g}_k)\mathcal{T}^H_{M+L}(\mathbf{g}_k) \tag{62}$$

The solution is given by the eigenvector corresponding to the smallest eigenvalue of $\mathcal{Q}$.

As with the cross-relation approach, the noise subspace method requires that the channel length L be accurately known in addition to the channel satisfying the no-common-subchannel-zeros condition. A detailed development of this class of methods may be found in Ref. 10, Chaps. 3 and 5.

5.2.3. Multistep Linear Prediction. More recently, the problem of blind channel identification has been formulated as problems of linear prediction [6,8,37] and smoothing [30]. Define the signal (noise-free) part of (3) as

$$\mathbf{s}_k = \sum_{n=0}^{L} \mathbf{f}_n I_{k-n} \tag{63}$$

with s_{ik} denoting the i-th component of $\mathbf{s}_k$. By (28) with $d = -N$, there exists a causal FIR filter of length $M \leq L-1$ such that

$$I_k = \sum_{n=0}^{M} \sum_{i=1}^{P} \tilde{c}_{in} s_{i(k-n)} \tag{64}$$

Using (63) and (64), we have

$$\mathbf{s}_k = \mathbf{e}_{k|k-1} + \hat{\mathbf{s}}_{k|k-1} \tag{65}$$

where

$$\mathbf{e}_{k|k-1} := \mathbf{f}_0 I_k \tag{66}$$

and

$$\hat{\mathbf{s}}_{k|k-1} := \sum_{n=1}^{L} \mathbf{f}_n I_{k-n} = \sum_{i=1}^{L_e} \mathbf{A}_i \mathbf{s}_{k-i} \tag{67}$$

such that

$$E\{\mathbf{e}_{k|k-1}\mathbf{s}_{k-l}^H\} = 0 \,\forall\, l \geq 1 \tag{68}$$

That is, by the orthogonality principle, $\hat{\mathbf{s}}_{k|k-1}$ is the one-step ahead linear prediction (of finite length) of $\mathbf{s}_k$, and $\mathbf{e}_{k|k-1}$ is the corresponding prediction error (linear innovations). Existence of $L_e \leq L-1$ in (67) can be established. The predictor coefficient $\mathbf{A}_i$ values can be estimated from data (after removal of noise effects); therefore, one can calculate $E\{\mathbf{e}_{k|k-1}\mathbf{e}_{k|k-1}^H\}$ from data-based correlation estimates. By (66), we obtain

$$E\{\mathbf{e}_{k|k-1}\mathbf{e}_{k|k-1}^H\} = E\{|I_k|^2\}\mathbf{f}_0\mathbf{f}_0^H \tag{69}$$

a rank 1 matrix. Equation (69) allows estimation of $\mathbf{f}_0$ up to a scale factor (the estimate equals the eigenvector of $E\{\mathbf{e}_{k|k-1}\mathbf{e}_{k|k-1}^H\}$ corresponding to the largest eigenvalue). Once we have a scaled estimate of $\mathbf{f}_0$, we can estimate the remaining channel coefficients using (63) with $\{\mathbf{s}_k\}$ as output and $\|\mathbf{f}_0\|^{-2}\mathbf{f}_0^H\mathbf{e}_{k|k-1}(= I_k e^{j\alpha})$ as input (where α is arbitrary).

The approach described above can be extended by using multistep linear prediction. It can be shown that

$$\mathbf{s}_k = \mathbf{e}_{k|k-2} + \hat{\mathbf{s}}_{k|k-2} \tag{70}$$

where

$$\mathbf{e}_{k|k-2} := \mathbf{f}_0 I_k + \mathbf{f}_1 I_{k-1} \tag{71}$$

and

$$\hat{\mathbf{s}}_{k|k-2} := \sum_{n=2}^{L} \mathbf{f}_n I_{k-n} = \sum_{i=2}^{L_e+1} \mathbf{A}_{2i} \mathbf{s}_{k-i} \tag{72}$$

such that

$$E\{\mathbf{e}_{k|k-2}\mathbf{s}_{k-l}^H\} = 0 \,\forall\, l \geq 2 \tag{73}$$

By the orthogonality principle, $\hat{\mathbf{s}}_{k|k-2}$ is the two-step-ahead linear prediction (of finite length) of $\mathbf{s}_k$, and $\mathbf{e}_{k|k-2}$ is the corresponding prediction error. Define

$$\mathbf{E}_k := \left((\mathbf{e}_{k+1|k-1} - \mathbf{e}_{k+1|k})^T \quad \mathbf{e}_{k|k-1}^T\right)^T \tag{74}$$

so that we have

$$\mathbf{E}_k = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_0 \end{pmatrix} I_k \tag{75}$$

By (75), we have

$$E\{\mathbf{E}_k\mathbf{E}_k^H\} = E\{|I_k|^2\} \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_0 \end{pmatrix} \begin{pmatrix} \mathbf{f}_1^H & \mathbf{f}_0^H \end{pmatrix} \tag{76}$$

a rank 1 matrix. That is, we can estimate $\mathbf{f}_0$ and $\mathbf{f}_1$ simultaneously up to the same scale factor. By adding larger step predictors, one can estimate the entire channel impulse response simultaneously. An advantage over one-step predictor approach is that the results are not unduly influenced by any estimation errors in estimating the leading coefficient $\mathbf{f}_0$.

The multistep linear prediction approach was proposed by Ding [6] in a different form and by Gesbert and Duhamel [8] in the form given above. Both of them assumed FIR channels with known channel length and no common subchannel zeros. Tugnait [37] extended the approach of Gesbert and Duhamel [8] by allowing common subchannel zeros, IIR (infinite impulse response) channels and unknown channel length. It has been shown [37] that minimum-phase common subchannel zeros pose no problems for the multistep linear prediction approach, and in the presence of nonminimum-phase common subchannel zeros, the multistep linear prediction approach yields a minimum-phase equivalent version of these zeros. It is also worth noting that linear prediction approaches (both single-step and multistep) are robust against overdetermination of channel length, unlike the cross-relation and noise subspace approaches.

6. COMMERCIAL APPLICATIONS

The commercial implementations and applications of blind equalizers reported in the literature are all based on CMA/Godard FIR equalizers (typically FSEs with twice the baud-rate sampling) and its variations in the acquisition stage followed by a decision-directed implementation in the operational stage. Treichler et al. [32] describe a variety of digitally implemented demodulators ranging from digital signal processor (DSP) chip-based designs used for voiceband modems (modulation types up to 128-QAM, baud rate up to 3500 baud) to very-large-scale-integration (VLSI)-based designs for digital microwave radio (modulation types up to 128-QAM, symbol rates up to 40 Mbaud). The intended applications of Treichler's designs [32] include high-speed voiceband modems, digital cable modems, and high-capacity digital microwave radios.

Werner et al. [40] describe successful laboratory experimental results with a 51.84-Mbp/s 16-CAP (12.92-Mbaud) transceiver prototype used for FTTC and VDSL (very-high-rate DSL).

Reports of the performance of other blind equalizers and channel estimators are based on computer simulations (or "controlled real data"). Promising simulation results have been reported on the application of blind equalization in the popular wireless GSM cellular system [3] using a

higher-order statistical deconvolution method [2] where the estimated channel is used in conjunction with MLSE (ML sequence estimator) for symbol estimation. Boss et al. [3] report that their HOS-based approach, using only 142 data samples per frame, incurs an SNR loss of 1.2-1.3 dB only while it saves the 22% overhead in the GSM data rate caused by the transmission of training sequences. (Thus, on the average, the Boss et al. HOS-based approach [2] requires 1.2–1.3 dB higher SNR than the conventional GSM system to achieve the same bit error rate.)

Acknowledgments
This work was prepared in part under the support of the National Science Foundation under Grant CCR-9803850.

BIOGRAPHY

Jitendra K. Tugnait received his B.Sc.(Hons.) degree in electronics and electrical communication engineering from the Punjab Engineering College, Chandigarh, India, in 1971, M.S. and the E.E. degrees from Syracuse University, Syracuse, New York, and a Ph.D. degree from the University of Illinois, Urbana-Champaign Urbana, Illinois, in 1973, 1974, and 1978, respectively, all in electrical engineering. From 1978 to 1982 he was an assistant professor of Electrical and Computer Engineering at the University of Iowa, Iowa City, IA. He was with the Long Range Research Division of the Exxon Production Research Company, Houston, TX, from 1982 to 1989 working on geophysical signal processing problems. He joined the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, in September 1989 as a professor. His research interests are in statistical signal processing, wireless and wireline digital communications, blind channel estimation and equalization for single and multiuser systems, and system identification. Dr. Tugnait has published over 95 journal and 120 conference articles. He was elected a fellow of IEEE in 1994. He is a past associate editor of the *IEEE Transactions on Signal of Processing* and of the *IEEE Transactions on Automatic Control*.

BIBLIOGRAPHY

1. A. Benveniste, M. Goursat, and G. Ruget, Robust identification of a nonminimum phase system: blind adjustment of a linear equalizer in data communications, *IEEE Trans. Autom. Control* **AC-25**: 385–399 (June 1980).
2. D. Boss, B. Jelonek, and K. D. Kammeyer, Eigenvector algorithm for blind MA system identification, *Signal Process.* **66**: 1–26 (April 1998).
3. D. Boss, K.-D. Kammeyer, and T. Petermann, Is blind channel estimation feasible in mobile communication systems? A study based on GSM, *IEEE J. Select. Areas Commun.* **SAC-16**: 1480–1492 (Oct. 1998).
4. R. A. Casas et al., Current approaches to blind decision feedback equalization, in G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 1: *Trends in Channel Estimation and Equalization*, Prentice-Hall, Upper Saddle River, NJ, 2001, Chap. 11, pp. 367–415.
5. K. M. Chugg, Blind acquisition characteristics of PSP-based sequence detectors, *IEEE J. Select. Areas Commun.* **SAC-16**: 1518–1529 (Oct. 1998).
6. Z. Ding, Matrix outer-product decomposition method for blind multiple channel identification, *IEEE Trans. Signal Process.* **45**: 3054–3061 (Dec. 1997).
7. W. A. Gardner, *Introduction to Random Processes: With Applications to Signals and Systems*, 2nd ed., McGraw-Hill, New York, 1989.
8. D. Gesbert and P. Duhamel, Robust blind channel identification and equalization based on multi-step predictors, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processes*, Seattle, WA, April 1997, pp. 3621–3624.
9. M. Ghosh and C. L. Weber, Maximum-likelihood blind equalization, *Opt. Eng.* **31**: 1224–1228 (June 1992).
10. G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 1: *Trends in Channel Estimation and Equalization*, Prentice-Hall, Upper Saddle River, NJ, 2001.
11. R. D. Gitlin and S. B. Weinstein, Fractionally-spaced equalization: An improved digital transversal equalizer, *Bell Syst. Tech. J.* **60**: 275–296 (Feb. 1981).
12. D. N. Godard, Self-recovering equalization and carrier tracking in two-dimensional data communication systems, *IEEE Trans. Commun.* **COM-28**: 1867–1875 (Nov. 1980).
13. D. Hatzinakos and C. L. Nikias, Blind equalization using a tricepstrum based algorithm, *IEEE Trans. Commun.* **COM-39**: 669–681 (May 1991).
14. Y. Hua, Fast maximum likelihood for blind identification of multiple FIR channels, *IEEE Trans. Signal Process.* **SP-44**: 661–672 (March 1996).
15. C. R. Johnson, Jr., et al., Blind equalization using the constant modulus criterion: A review, *Proc. IEEE* **86**: 1927–1950 (Oct. 1998).
16. G. K. Kaleh and R. Vallet, Joint parameter estimation and symbol detection for linear or non linear unknown dispersive channels, *IEEE Trans. Commun.* **COM-42**: 2406–2413 (July 1994).
17. V. Krishnamurthy and J. B. Moore, On-line estimation of hidden Markov model parameters based on Kullback-Leibler information measure, *IEEE Trans. Signal Process.* **SP-41**: 2557–2573 (Aug. 1993).
18. Y. Li and Z. Ding, Global convergence of fractionally spaced Godard adaptive equalizers, *IEEE Trans. Signal Process.* **SP-44**: 818–826 (April 1996).
19. O. Macchi and E. Eweda, Convergence analysis of self-adaptive equalizers, *IEEE Trans. Inform. Theory* **IT-30**: 162–176 (March 1983).
20. E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue, Subspace-methods for the blind identification of multichannel FIR filters, *IEEE Trans. Signal Process.* **SP-43**: 516–525 (Feb. 1995).
21. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
22. R. Raheli, A. Polydoros, and C. K. Tzou, Per-survivor processing: A general approach to MLSE in uncertain environments, *IEEE Trans. Commun.* **COM-43**: 354–364 (Feb.–April 1995).

23. Y. Sato, A method of self-recovering equalization for multilevel amplitude modulation, *IEEE Trans. Commun.* **COM-23**: 679–682 (June 1975).

24. N. Seshadri, Joint data and channel estimation using fast blind trellis search techniques, *IEEE Trans. Commun.* **COM-42**: 1000–1011 (March 1994).

25. O. Shalvi and E. Weinstein, New criteria for blind deconvolution of nonminimum phase systems (channels), *IEEE Trans. Inform. Theory* **IT-36**: 312–321 (March 1990).

26. O. Shalvi and E. Weinstein, Super-exponential methods for blind deconvolution, *IEEE Trans. Inform. Theory* **IT-39**: 504–519 (March 1993).

27. S. Talwar, M. Viberg, and A. Paulraj, Blind separation of synchronous co-channel digital signals using an antenna array — Part I: Algorithms, *IEEE Trans. Signal Process.* **SP-44**: 1184–1197 (May 1996).

28. L. Tong and S. Perreau, Multichannel blind channel estimation: From subspace to maximum likelihood methods, *Proc. IEEE* **86**: 1951–1968 (Oct. 1998).

29. L. Tong, G. Xu, and T. Kailath, A new approach to blind identification and equalization of multipath channels, *IEEE Trans. Inform. Theory* **IT-40**: 340–349 (March 1994).

30. L. Tong and Q. Zhao, Joint order detection and blind channel estimation by least squares smoothing, *IEEE Trans. Signal Process.* **SP-47**: 2345–2355 (Sept. 1999).

31. J. R. Treichler and M. G. Agee, A new approach to multipath correction of constant modulus signals, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-31**: 349–472 (April 1983).

32. J. R. Treichler, M. G. Larimore, and J. C. Harp, Practical blind demodulators for high-order QAM signals, *Proc. IEEE* **86**: 1907–1926 (Oct. 1998).

33. J. K. Tugnait, Identification of linear stochastic systems via second- and fourth-order cumulant matching, *IEEE Trans. Inform. Theory* **IT-33**: 393–407 (May 1987).

34. J. K. Tugnait, On blind identifiability of multipath channels using fractional sampling and second-order cyclostationary statistics, *IEEE Trans. Inform. Theory* **IT-41**: 308–311 (Jan. 1995).

35. J. K. Tugnait, Blind estimation and equalization of digital communication FIR channels using cumulant matching, *IEEE Trans. Commun.* **COM-43**(Pt. III): 1240–1245 (Feb.–April 1995).

36. J. K. Tugnait, Blind equalization and estimation of FIR communications channels using fractional sampling, *IEEE Trans. Commun.* **COM-44**: 324–336 (March 1996).

37. J. K. Tugnait, Multistep linear predictors-based blind equalization of FIR/IIR single-input multiple-output channels with common zeros, *IEEE Trans. Signal Process.* **SP-47**: 1689–1700 (June 1999).

38. J. K. Tugnait, Channel estimation and equalization using higher-order statistics, in G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 1: *Trends in Channel Estimation and Equalization*, Prentice-Hall, Upper Saddle River, NJ, 2001, Chap. 1, pp. 1–39.

39. J. Vidal and J. A. R. Fonollosa, Adaptive blind equalization using weighted cumulant slices, *Int. J. Adapt. Control Signal Process.* **10**(2–3): 213–238 (March–June 1996).

40. J.-J. Werner, J. Yang, D. D. Harman, and G. A. Dumont, Blind equalization for broadband access, *IEEE Commun. Mag.* **37**: 87–93 (April 1999).

41. G. Xu, H. Liu, L. Tong, and T. Kailath, A least-squares approach to blind channel identification, *IEEE Trans. Signal Process.* **SP-43**: 2982–2993 (Dec. 1995).

42. H. Zeng, L. Tong, and C. R. Johnson, Jr., Relationships between CMA and Wiener receivers, *IEEE Trans. Inform. Theory* **IT-44**: 1523–1538 (July 1998).

BLIND MULTIUSER DETECTION

XIAODONG WANG
Columbia University
New York, New York

1. INTRODUCTION

Code-division multiple access (CDMA) implemented with direct-sequence spread-spectrum (DSSS) modulation is emerging as a popular multiple-access technology for personal, cellular, and satellite communication services. Multiuser detection techniques can substantially increase the capacity of CDMA systems. Since the early 90s, a significant amount of research has addressed various multiuser detection schemes [33]. Considerable attention has been focused on adaptive multiuser detection [10]. For example, methods for adapting the linear decorrelating detector that require the transmission of training sequences during adaptation have been proposed [5,20,21]. An alternative linear detector, the linear minimum mean-square error (MMSE) detector, however, can be adapted either through the use of training sequences [1,18,19,24], or in the blind mode, with the prior knowledge of only the signature waveform and timing of the user of interest [9,38]. Blind adaptation schemes are especially attractive for the downlinks of CDMA systems, since in a dynamic environment, it is very difficult for a mobile user to obtain the accurate information of other active users in the channel, such as their signature waveforms; and the frequent use of training sequence is certainly a waste of channel bandwidth. There are primarily two approaches to blind multiuser detection, namely, the direct matrix inversion (DMI) approach and the subspace approach. In this article, we present batch algorithms and adaptive algorithms under both approaches. We first treat the simple synchronous CDMA channels and present the main techniques for blind multiuser detection. We then generalize these methods to the more general asynchronous CDMA channels with multipath effects.

2. LINEAR RECEIVERS FOR SYNCHRONOUS CDMA

2.1. Synchronous CDMA Signal Model

We start by introducing the most basic multiple-access signal model, namely, a baseband, K-user, time-invariant, synchronous, additive white Gaussian noise (AWGN) system, employing periodic (short) spreading sequences, and operating with a coherent BPSK (binary phase shift keying) modulation format. The waveform received by a given user in such a system can be modeled as

$$r(t) = \sum_{k=1}^{K} A_k \sum_{i=0}^{M-1} b_k[i] s_k(t - iT) + n(t) \qquad (1)$$

where M is the number of data symbols per user in the data frame of interest; T is the symbol interval; A_k, $\{b_k[i]\}_{i=0}^{M-1}$, and $s_k(t)$ denote, respectively, the received complex amplitude, the transmitted symbol stream, and the normalized signaling waveform of the kth user; and $n(t)$ is the baseband complex white Gaussian ambient channel noise with power spectral density σ^2. It is assumed that for each user k, $\{b_k[i]\}_{i=0}^{M-1}$ is a collection of independent equiprobable ± 1 random variables and that the symbol streams of different users are independent. The user signaling waveform is of the form

$$s_k(t) = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} c_{j,k}\psi(t - jT_c), \qquad 0 \le t < T \tag{2}$$

where N is the processing gain; $\{c_{j,k}\}_{j=0}^{N-1}$ is a signature sequence of ± 1 values assigned to the kth user; and $\psi(\cdot)$ is a chip waveform of duration $T_c = T/N$ and with unit energy, that is, $\int_0^{T_c} \psi(t)^2 dt = 1$.

At the receiver, the received signal $r(t)$ is filtered by a chip-matched filter and then sampled at the chip rate. The sample corresponding to the jth chip of the ith symbol is given by

$$r_j[i] \triangleq \int_{iT+jT_c}^{iT+(j+1)T_c} r(t)\psi(t - iT - jT_c)\, dt \tag{3}$$
$$j = 0, \ldots, N-1; \quad i = 0, \ldots, M-1$$

The resulting discrete-time signal corresponding to the ith symbol is then given by,

$$\boldsymbol{r}[i] = \sum_{k=1}^{K} A_k b_k[i]\boldsymbol{s}_k + \boldsymbol{n}[i] \tag{4}$$
$$= \boldsymbol{S}\boldsymbol{A}\boldsymbol{b}[i] + \boldsymbol{n}[i] \tag{5}$$

with

$$\boldsymbol{r}[i] \triangleq \begin{bmatrix} r_0[i] \\ r_1[i] \\ \vdots \\ r_{N-1}[i] \end{bmatrix}, \quad \boldsymbol{s}_k \triangleq \frac{1}{\sqrt{N}} \begin{bmatrix} c_{0,k} \\ c_{1,k} \\ \vdots \\ c_{N-1,k} \end{bmatrix},$$
$$\boldsymbol{n}[i] \triangleq \begin{bmatrix} n_0[i] \\ n_1[i] \\ \vdots \\ n_{N-1}[i] \end{bmatrix}$$

where $n_j[i] = \int_{jT_c}^{(j+1)T_c} n(t)\psi(t - iT - jT_c)dt \sim \mathcal{N}_c(0, \sigma^2)$ is a complex Gaussian random variable with independent real and imaginary components; $\boldsymbol{n}[i] \sim \mathcal{N}_c(\boldsymbol{0}, \sigma^2\boldsymbol{I})$; $\boldsymbol{S} \triangleq [\boldsymbol{s}_1 \cdots \boldsymbol{s}_K]$; $\boldsymbol{A} \triangleq \mathrm{diag}(A_1, \ldots, A_K)$; and $\boldsymbol{b}[i] \triangleq [b_1[i] \cdots b_K[i]]^T$.

2.2. Linear MMSE Detector

Suppose that we are interested in demodulating the data bits of a particular user, say user 1, $\{b_1[i]\}_{i=0}^{M-1}$, based on the received waveforms $\{\boldsymbol{r}[i]\}_{i=0}^{M-1}$. A *linear* receiver for this purpose is a vector $\boldsymbol{w}_1 \in \mathbb{C}^N$, such that the desired user's data bits are demodulated according to

$$z_1[i] = \boldsymbol{w}_1^H \boldsymbol{r}[i] \tag{6}$$

$$\hat{b}_1[i] = \mathrm{sign}\,\{\Re(A_1^* z_1[i])\} \tag{7}$$

In case that the complex amplitude A_1 of the desired user is unknown, we can resort to differential detection. Define the differential bit as

$$\beta_1[i] \triangleq b_1[i]b_1[i-1] \tag{8}$$

Then, using the linear detector output (6), the following differential detection rule can be applied:

$$\hat{\beta}_1[i] = \mathrm{sign}\,\{\Re(z_1[i]z_1[i-1]^*)\} \tag{9}$$

Substituting (4) into (6), the output of the linear receiver $\boldsymbol{w_1}$ can be written as

$$z_1[i] = A_1(\boldsymbol{w}_1^H\boldsymbol{s}_1)b_1[i] + \sum_{k=2}^{K} A_k(\boldsymbol{w}_1^H\boldsymbol{s}_k)b_k[i] + \boldsymbol{w}_1^H\boldsymbol{n}[i] \tag{10}$$

In (10), the first term contains the useful signal of the desired user; the second term contains the signals from other undesired users — the so-called multiple-access interference (MAI); and the last term contains the ambient Gaussian noise. The simplest linear receiver is the conventional matched-filter, where $\boldsymbol{w}_1 = \boldsymbol{s}_1$. It is well known that such a matched-filter receiver is optimal only in a single-user channel (i.e., $K = 1$). In a multiuser channel (i.e., $K > 1$), this receiver may perform poorly since it makes no attempt to ameliorate the MAI, a limiting source of interference in multiple-access channels.

The linear minimum mean-square error (MMSE) detector is designed to minimize the total effect of the MAI and the ambient noise at the detector output. Specifically, it is given by the solution to the following optimization problem:

$$\boldsymbol{w}_1 = \arg\min_{\boldsymbol{w}\in\mathbb{C}^N} E\left\{\|A_1 b_1[i] - \boldsymbol{w}^H\boldsymbol{r}[i]\|^2\right\} \tag{11}$$

Denote $|\boldsymbol{A}| \triangleq \mathrm{diag}(|A_1|, \ldots, |A_K|)$ and $\boldsymbol{R} \triangleq \boldsymbol{S}^T\boldsymbol{S}$. The solution to (11) is given by [33]

$$\boldsymbol{w}_1 = \boldsymbol{S}\left(\boldsymbol{R} + \sigma^2|\boldsymbol{A}|^{-2}\right)^{-1}\boldsymbol{e}_1 \tag{12}$$

where $\boldsymbol{e}_1$ denotes the first unit vector in $\mathbb{C}^K$.

3. BLIND MULTIUSER DETECTION: DIRECT METHODS

It is seen from (12) that the linear MMSE detector $\boldsymbol{w}_1$ is expressed in terms of a linear combination of the signature sequences $\boldsymbol{S}$ of all K users. Recall that for the matched-filter receiver, the only prior knowledge required is the

desired user's signature sequence $\boldsymbol{s}_1$. In the downlink of a CDMA system, the mobile receiver typically has knowledge only of its own signature sequence, and not of those of the other users. Hence it is of interest to consider the problem of *blind* implementation of the linear detector, that is, without the requirement of knowing the signature sequences of the interfering users. This problem is relatively easy for the linear MMSE detector. To see this, consider again the definition (11). Directly solving this optimization problem, we obtain the following alternative expression for the linear MMSE detector:

$$\begin{aligned}\boldsymbol{w}_1 &= \arg\min_{\boldsymbol{w}\in\mathbb{C}^N} \boldsymbol{w}^H \underbrace{E\left\{\boldsymbol{r}[i]\boldsymbol{r}[i]^H\right\}}_{\boldsymbol{C}_r}\boldsymbol{w} - 2\boldsymbol{w}^H\Re\{A_1^* \underbrace{E(\boldsymbol{r}[i]b_1[i])}_{A_1\boldsymbol{s}_1}\} \\ &= |A_1|^2\boldsymbol{C}_r^{-1}\boldsymbol{s}_1 \end{aligned} \tag{13}$$

where by (5)

$$\boldsymbol{C}_r \triangleq E\left\{\boldsymbol{r}[i]\boldsymbol{r}[i]^H\right\} = \boldsymbol{S}|\boldsymbol{A}|^2\boldsymbol{S}^T + \sigma^2\boldsymbol{I} \tag{14}$$

is the autocorrelation matrix of the receiver signal. Note that $\boldsymbol{C}_r$ can be estimated from the received signals by the corresponding sample autocorrelation. Note also that the constant $|A_1|^2$ in (13) does not affect the linear decision rule (7) or (9). Hence (13) leads straightforwardly to the following blind implementation of the linear MMSE detector — the so-called direct matrix inversion (DMI) blind detector.

- *Compute the detector*:

$$\hat{\boldsymbol{C}}_r \triangleq \frac{1}{M}\sum_{i=0}^{M-1}\boldsymbol{r}[i]\boldsymbol{r}[i]^H$$

$$\hat{\boldsymbol{w}}_1 = \hat{\boldsymbol{C}}_r^{-1}\boldsymbol{s}_1$$

- *Perform differential detection*:

$$z_1[i] = \hat{\boldsymbol{w}}_1^H\boldsymbol{r}[i]$$

$$\hat{\beta}_1[i] = \text{sign}\,\{\Re(z_1[i]z_1[i-1]^*)\}, \quad i = 1,\ldots,M-1$$

This algorithm is a *batch* processing method; that is, it computes the detector only once on the basis of a block of received signals $\{\boldsymbol{r}[i]\}_{i=0}^{M-1}$; and the estimated detector is then used to detect all the data bits of the desired user $\{b_1[i]\}_{i=0}^{M-1}$ contained in the same signal block. In what follows, we consider the *adaptive* implementation of the blind linear MMSE detector.

The idea is to perform sequential (i.e., online) blind detector estimation and data detection; that is, suppose that at time $(i-1)$, a detector $\boldsymbol{w}_1[i-1]$ is used for detecting $b_1[i-1]$. At time i, a new signal $\boldsymbol{r}[i]$ is received and is then used to update the detector to obtain $\boldsymbol{w}_1[i]$. The updated detector is used to detect the data bit $b_1[i]$. Hence the blind detector is sequentially updated at the symbol rate. In order to develop such an adaptive algorithm, we need an alternative characterization of the linear MMSE detector. Consider the following constrained optimization problem:

$$\boldsymbol{w}_1 = \arg\min_{\boldsymbol{w}\in\mathbb{C}^N} E\left\{\|\boldsymbol{w}^H\boldsymbol{r}[i]\|^2\right\}, \quad \text{s.t.} \quad \boldsymbol{w}^H\boldsymbol{s}_1 = 1 \tag{15}$$

The solution to (15) is given by

$$\boldsymbol{w}_1 = \alpha\boldsymbol{C}_r^{-1}\boldsymbol{s}_1 \tag{16}$$

where $\alpha = (\boldsymbol{s}_1^T\boldsymbol{C}_r^{-1}\boldsymbol{s}_1)^{-1}$. Comparing this solution with (13), it is seen that the two differ only by a positive scaling constant. Since such a scaling constant will not affect the linear decision rule (7) or (9), (15) constitutes an equivalent definition of the linear MMSE detector. We next consider the adaptive implementation of the linear MMSE detector basis of the on the least mean-square (LMS) algorithm. Note that $\boldsymbol{w}_1$ can be decomposed into two orthogonal components

$$\boldsymbol{w}_1 = \boldsymbol{s}_1 + \boldsymbol{x}_1 \tag{17}$$

with

$$\boldsymbol{x}_1 \triangleq \boldsymbol{P}\boldsymbol{w}_1 = \boldsymbol{P}\boldsymbol{x}_1 \tag{18}$$

where $\boldsymbol{P} \triangleq \boldsymbol{I} - \boldsymbol{s}_1\boldsymbol{s}_1^T$ is a projection matrix that projects any signal in $\mathbb{C}^N$ onto the orthogonal space of $\boldsymbol{s}_1$. Using this decomposition, the constrained optimization problem (15) can then be converted to the following unconstrained optimization problem:

$$\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\mathbb{C}^N} E\left\{\|(\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x})^H\boldsymbol{r}[i]\|^2\right\} \tag{19}$$

The LMS algorithm for adapting the weights $\boldsymbol{x}_1$ based on the cost function (19) is then given by [8]

$$\boldsymbol{x}_1[i+1] = \boldsymbol{x}_1[i] - \frac{\mu}{2}g(\boldsymbol{x}_1[i]) \tag{20}$$

where μ is the step size, and the stochastic gradient $g(\boldsymbol{x}_1[i])$ is given by

$$\begin{aligned} g(\boldsymbol{x}_1[i]) &\triangleq \frac{d}{dx}\|(\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x})^H\boldsymbol{r}[i]\|^2|\boldsymbol{x} = \boldsymbol{x}_1[i] \\ &= 2[\boldsymbol{r}[i] - (\boldsymbol{s}_1^T\boldsymbol{r}[i])\boldsymbol{s}_1][(\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x}_1[i])^H\boldsymbol{r}[i]]^* \end{aligned} \tag{21}$$

Substituting (21) into (20), we obtain the following LMS implementation of the blind linear MMSE detector. Suppose that at time i, the estimated blind detector is $\boldsymbol{w}_1[i] = \boldsymbol{s}_1 + \boldsymbol{x}_1[i]$. The algorithm performs the following steps for data detection and detector update:

- *Compute the detector output*:

$$z_1[i] = (\boldsymbol{s}_1 + \boldsymbol{P}\boldsymbol{x}_1[i])^H\boldsymbol{r}[i]$$

$$\hat{\beta}_1[i] = \text{sign}\,\{\Re(z[i]z[i-1]^*)\}$$

- *Update*:

$$\boldsymbol{x}_1[i+1] = \boldsymbol{x}_1[i] - \mu z[i]^*[\boldsymbol{r}[i] - (\boldsymbol{s}_1^T\boldsymbol{r}[i])\boldsymbol{s}_1]$$

This algorithm is initialized as $\boldsymbol{x}_1[0] = \mathbf{0}$. The adaptive approach outlined above was first proposed in [9], and is termed the *minimum-output-energy* (MOE) detector.

4. BLIND MULTIUSER DETECTION: SUBSPACE METHODS

In this section, we discuss another approach to blind multiuser detection, which is based on estimating the signal subspace spanned by the user signature waveforms. This approach, first proposed in [38], offers a number of advantages over the direct methods discussed in the previous section.

Assume that the spreading waveforms $\{\boldsymbol{s}_k\}_{k=1}^K$ of K users are linearly independent. The eigendecomposition of the signal autocorrelation matrix $\boldsymbol{C}_r$ in (14) can be written as

$$\boldsymbol{C}_r = \boldsymbol{U}_s \boldsymbol{\Lambda}_s \boldsymbol{U}_s^H + \sigma^2 \boldsymbol{U}_n \boldsymbol{U}_n^H \tag{22}$$

where $\boldsymbol{\Lambda}_s = \text{diag}(\lambda_1, \ldots, \lambda_K)$ contains the largest K eigenvalues of $\boldsymbol{C}_r$; $\boldsymbol{U}_s = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]$ contains the K orthonormal eigenvectors corresponding to the largest K eigenvalues in $\boldsymbol{\Lambda}_s$; $\boldsymbol{U}_n = [\boldsymbol{u}_{K+1}, \ldots, \boldsymbol{u}_N]$ contains the $(N-K)$ orthonormal eigenvectors corresponding to the smallest eigenvalue σ^2 of $\boldsymbol{C}_r$. It is easy to see that range $(\boldsymbol{S})$ = range $(\boldsymbol{U}_s)$. The column space of $\boldsymbol{U}_s$ is called the *signal subspace* and its orthogonal complement, the *noise subspace*, is spanned by the columns of $\boldsymbol{U}_n$. The linear MMSE detector can be expressed in terms of the signal subspace parameters $\boldsymbol{U}_s$ and $\boldsymbol{\Lambda}_s$ as [38]

$$\boldsymbol{w}_1 = \alpha \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \boldsymbol{s}_1 \tag{23}$$

with $\alpha = (\boldsymbol{s}_1^T \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \boldsymbol{s}_1)^{-1}$.

Since the decision rules (7) and (9) are invariant to a positive scaling, the subspace linear multiuser detector given by (23) can be interpreted as follows. First the received signal $\boldsymbol{r}[i]$ is projected onto the signal subspace to get $\boldsymbol{y}[i] \triangleq \boldsymbol{U}_s^H \boldsymbol{r}[i] \in \mathbb{C}^K$, which clearly is a sufficient statistic for demodulating the K users' data bits. The spreading waveform $\boldsymbol{s}_1$ of the desired user is also projected onto the signal subspace to obtain $\boldsymbol{p}_1 \triangleq \boldsymbol{U}_s^H \boldsymbol{s}_1 \in \mathbb{C}^K$. The projection of the linear multiuser detector in the signal subspace is then a signal $\boldsymbol{c}_1 \in \mathbb{C}^K$ such that the detector output is $z_1[i] \triangleq \boldsymbol{c}_1^H \boldsymbol{y}[i]$, and the data bit is demodulated as $\hat{b}_1[i] = \text{sign}\,\{\Re(A_1^* z_1[i])\}$ for coherent detection, and $\beta_1[i] = \text{sign}\,\{\Re(z_1[i] z_1[i-1]^*)\}$ for differential detection. According to (23), the projection of the linear MMSE detector in the signal subspace is given by

$$\boldsymbol{c}_1 = \begin{bmatrix} \frac{1}{\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_K} \end{bmatrix} \boldsymbol{p}_1 \tag{24}$$

Thus, it is obtained by projecting the spreading waveform of the desired user onto the signal subspace, followed by scaling the kth component of this projection by a factor of $1/\lambda_k$.

Since the autocorrelation matrix $\boldsymbol{C}_r$, and therefore its eigencomponents, can be estimated from the received signals, we see that the abovementioned subspace method indeed leads to a blind implementation of the linear MMSE detector. Finally we summarize the subspace blind multiuser detector as follows:

- *Compute the detector*:

$$\begin{aligned} \hat{\boldsymbol{C}}_r &\triangleq \frac{1}{M} \sum_{i=0}^{M-1} \boldsymbol{r}[i] \boldsymbol{r}[i]^H \\ &= \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H + \hat{\boldsymbol{U}}_n \hat{\boldsymbol{\Lambda}}_n \hat{\boldsymbol{U}}_n^H \\ \hat{\boldsymbol{w}}_1 &= \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H \boldsymbol{s}_1 \end{aligned}$$

- *Perform differential detection*:

$$\begin{aligned} & z_1[i] = \hat{\boldsymbol{w}}_1^H \boldsymbol{r}[i], \\ & \hat{\beta}_1[i] = \text{sign}\,\{\Re(z_1[i] z_1^*[i-1])\}, \quad i = 1, \ldots, M-1 \end{aligned}$$

It is seen from the discussion above that the linear MMSE detector is obtained as long as the signal subspace components are identified. The classic approach to subspace estimation is through batch eigenvalue decomposition (ED) of the sample autocorrelation matrix, or batch singular value decomposition (SVD) of the data matrix, which is computationally too expensive for adaptive applications. Modern subspace tracking algorithms are recursive in nature and update the subspace in a sample-by-sample fashion. Various subspace tracking algorithms exist in the literature [e.g., 4,6,7,28,32,41], with different computational complexity and tracking performance. Among the low-complexity subspace tracking algorithms are the PASTd algorithm [41], and the more recently developed NAHJ algorithm [25,26]. Both algorithms have a complexity of $O(NK)$ when tracking K subspace components in a N-dimensional space; but the NAHJ algorithm has a far superior performance.

The adaptive blind multiuser detector based on subspace tracking sequentially estimates the signal subspace components, and forms the closed-form detector from these estimates. Specifically, supposedly at time $(i-1)$, the estimated signal subspace rank is $K[i-1]$ and the components are $(\boldsymbol{U}_s[i-1], \boldsymbol{\Lambda}_s[i-1])$. Then at time i, the adaptive detector performs the following steps to update the detector and to estimate the data:

- *Update the signal subspace*: Using a particular signal subspace tracking algorithm, update the signal subspace rank $K[i]$ and the signal subspace components $(\boldsymbol{U}_s[i], \boldsymbol{\Lambda}_s[i])$.
- *Form the detector and perform differential detection*:

$$\begin{aligned} \boldsymbol{w}_1[i] &= \boldsymbol{U}_s[i] \boldsymbol{\Lambda}_s[i]^{-1} \boldsymbol{U}_s[i]^H \boldsymbol{s}_1, \\ z_1[i] &= \boldsymbol{w}_1[i]^H \boldsymbol{r}[i], \\ \hat{\beta}_1[i] &= \text{sign}\,\{\Re(z_1[i] z_1[i-1]^*)\} \end{aligned}$$

Simulation Example

This example compares the performance of the subspace blind adaptive multiuser detector using the NAHJ

subspace tracking algorithm [26], with that of the LMS MOE blind adaptive multiuser detector. It assumes a synchronous CDMA system with seven users ($K = 7$), each employing a gold sequence of length 15 ($N = 15$). The desired user is user 1. There are two 0-dB and four 10-dB interferers. The performance measure is the output signal-to-interference-plus-noise ratio (SINR). The performance is shown in Fig. 1. It is seen that the subspace blind detector significantly outperforms the LMS MOE blind detector in terms of both convergence rate and steady-state SINR.

5. BLIND MULTIUSER DETECTION IN MULTIPATH CHANNELS

In the previous sections, we have focused primarily on the synchronous CDMA signal model. In a practical wireless CDMA system, however, the user signals are asynchronous. Moreover, the physical channel exhibits dispersion due to multipath effects that further attenuate the user signals. In this section, we address blind multiuser detection in such channels. As will be seen, the principal techniques developed in the previous section can be applied to this more complicated situation as well.

5.1. Multipath Signal Model

We now consider a more general multiple-access signal model where the users are asynchronous and the channel exhibits multipath distortion effects. Let the multipath channel impulse response of the kth user be

$$g_k(t) = \sum_{l=1}^{L} \alpha_{l,k}\delta(t - \tau_{l,k}) \tag{25}$$

where L is the total number of paths in the channel; $\alpha_{l,k}$ and $\tau_{l,k}$ are, respectively, the complex path gain and the delay of the kth user's lth path, $\tau_{1,k} < \tau_{2,k} < \cdots < \tau_{L,k}$. The received continuous-time signal in this case is given by

$$\begin{aligned} r(t) &= \sum_{k=1}^{K}\sum_{i=0}^{M-1} b_k[i]\{s_k(t - iT) \star g_k(t)\} + n(t) \\ &= \sum_{k=1}^{K}\sum_{i=0}^{M-1} b_k[i] \sum_{l=1}^{L} \alpha_{l,k} s_k(t - iT - \tau_{l,k}) + n(t) \end{aligned} \tag{26}$$

where $\star$ denotes convolution.

At the receiver, the received signal $r(t)$ is filtered by a chip-matched filter and sampled at the chip rate. Let

$$\iota \triangleq \max_{1\le k\le K} \left\{ \left\lceil \frac{\tau_{L,k} + T_c}{T} \right\rceil \right\} \tag{27}$$

be the maximum delay spread in terms of symbol intervals. Denote $r_q[i] \triangleq \int_{iT+qT_c}^{iT+(q+1)T_C} r(t)\psi(t - iT - qT_c)dt$ and $n_q[i] = \int_{iT+qT_c}^{iT+(q+1)T_c} n(t)\psi(t - iT - qT_c)dt$, for $q = 0, \ldots, N-1; i = 0, \ldots, M-1$. Denote further

$$\underbrace{\underline{r}[i]}_{P\times 1} \triangleq \begin{bmatrix} r_0[i] \\ \vdots \\ r_{N-1}[i] \end{bmatrix}, \quad \underbrace{\underline{b}[i]}_{K\times 1} \triangleq \begin{bmatrix} b_1[i] \\ \vdots \\ b_K[i] \end{bmatrix},$$

$$\underbrace{\underline{n}[i]}_{P\times 1} \triangleq \begin{bmatrix} n_0[i] \\ \vdots \\ n_{N-1}[i] \end{bmatrix}$$

$$\underbrace{\underline{H}[m]}_{N\times K} \triangleq \begin{bmatrix} h_1[mN] & \cdots & h_K[mN] \\ \vdots & \vdots & \vdots \\ h_1[mN+N-1] & \cdots & h_K[mN+N-1] \end{bmatrix},$$

$$m = 0, \ldots, \iota$$

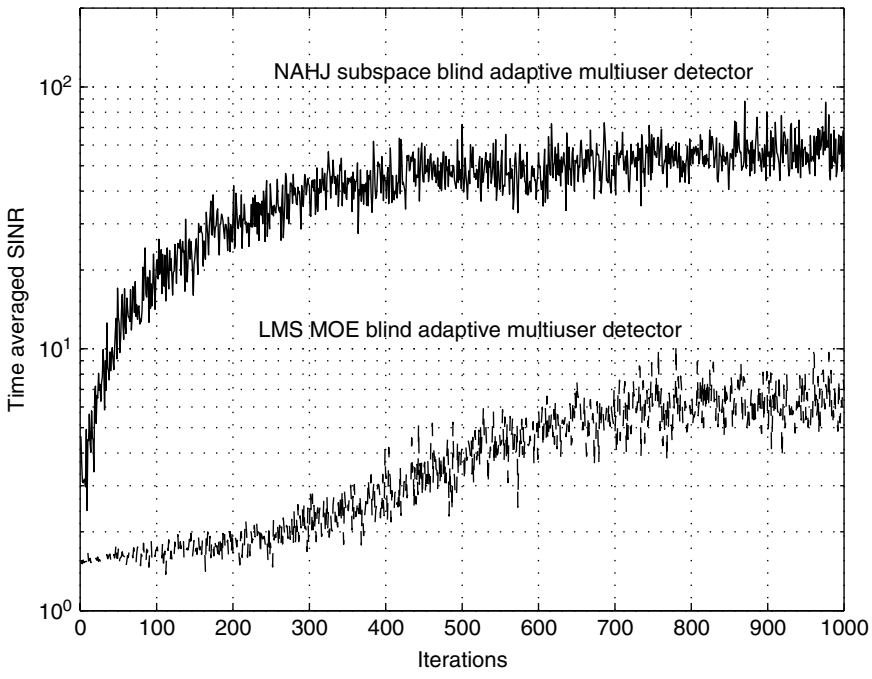


Figure 1. Performance comparison between the subspace blind adaptive multiuser detector using the NAHJ subspace tracking algorithm, and the LMS MOE blind adaptive multiuser detector.

where $\{h_k[j]\}_j$ is the composite signature waveform of the kth user, which will be discussed later. We then have the following discrete-time signal model [37]

$$\underline{r}[i] = \underline{H}[i] \star \underline{b}[i] + \underline{n}[i]. \tag{28}$$

By stacking Q successive sample vectors, we further define the quantities

$$\underbrace{\boldsymbol{r}[i]}_{NQ\times 1} \triangleq \begin{bmatrix} \underline{r}[i] \\ \vdots \\ \underline{r}[i+Q-1] \end{bmatrix}, \quad \underbrace{\boldsymbol{n}[i]}_{NQ\times 1} \triangleq \begin{bmatrix} \underline{n}[i] \\ \vdots \\ \underline{n}[i+Q-1] \end{bmatrix},$$

$$\underbrace{\boldsymbol{b}[i]}_{K(Q+\iota)\times 1} \triangleq \begin{bmatrix} \underline{b}[i-\iota] \\ \vdots \\ \underline{b}[i+Q-1] \end{bmatrix}$$

$$\underbrace{\boldsymbol{H}}_{NQ\times K(Q+\iota)} \triangleq \begin{bmatrix} \underline{H}[\iota] & \cdots & \underline{H}[0] & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \underline{H}[\iota] & \cdots & \underline{H}[0] \end{bmatrix}$$

where Q, called the "smoothing factor," is given by $Q = \lceil (N+K)/(N-K) \rceil \iota$; Note that for such Q, the matrix $\boldsymbol{H}$ is a "tall" matrix: $NQ \geq K(Q+\iota)$. We can then write (28) in a matrix forms as

$$\boldsymbol{r}[i] = \boldsymbol{H}\, \boldsymbol{b}[i] + \boldsymbol{n}[i] \tag{29}$$

5.2. Linear MMSE Multiuser Detector

Suppose that we are interested in demodulating the data bits of user 1. We can then write (28) as

$$\underline{r}[i] = \underline{H}^1[0] b_1[i] + \sum_{m=1}^{\iota} \underline{H}^1[m] b_1[i-m] + \sum_{k=2}^{K} \sum_{m=0}^{\iota} \underline{H}^k[m] b_k[i-m] + \underline{n}[i] \tag{30}$$

where $\underline{H}^k[m]$ denotes the kth column of $\underline{H}[m]$. In (30), the first term contains the data bit of the desired user at time i; the second term contains the previous data bits of the desired user, namely, intersymbol interference (ISI); and the last term contains the signals from other users, namely, multiple-access interference (MAI). Hence compared with the synchronous model considered in the previous sections, the multipath channel introduces ISI, which, together with MAI, must be contended with at the receiver. It is seen that the augmented signal model (29) is very similar to the synchronous signal model (5). We proceed to develop the linear receiver for this system.

A linear receiver for user 1 is a (NQ)-dimensional complex vector $\boldsymbol{w}_1 \in \mathbb{C}^{NQ}$, which is correlated with the received signal $\boldsymbol{r}[i]$ in (29), to compute the ith bit of this user, according to the following rule:

$$z_1[i] = \boldsymbol{w}_1^H \boldsymbol{r}[i] \tag{31}$$

$$\hat{\beta}_1[i] = \text{sign}\,\{\Re(z_1[i] z_1[i-1]^*)\} \tag{32}$$

The linear MMSE detector has the form of (32) with the weight vector chosen to minimize the output mean-square error (MSE):

$$\boldsymbol{w}_1 = \arg \min_{\boldsymbol{w}\in\mathbb{C}^{NQ}} E\left\{ \|b_1[i] - \boldsymbol{w}^H \boldsymbol{r}[i]\|^2 \right\} = \boldsymbol{C}_r^{-1} \overline{\boldsymbol{h}}_1 \tag{33}$$

where

$$\boldsymbol{C}_r = E\{\boldsymbol{r}[i]\boldsymbol{r}[i]^H\} = \boldsymbol{H}\boldsymbol{H}^H + \sigma^2 \boldsymbol{I} \tag{34}$$

$$\overline{\boldsymbol{h}}_1 \triangleq E\{\boldsymbol{r}[i] b_1[i]\} = \boldsymbol{H}[:, KQ+1] \tag{35}$$

$$= [\underbrace{h_1[0], \ldots, h_1[N-1], \ldots, h_1[\iota N], \ldots, h_1[\iota N + N - 1]}_{\boldsymbol{h}_1^T}, \underbrace{0, \ldots, 0}_{[N(Q-\iota-1)]0\text{s}}]^T$$

($\boldsymbol{H}[:, m]$ denotes the mth column of $\boldsymbol{H}$.)

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{NQ}$ be the eigenvalues of $\boldsymbol{C}_r$ in (34). Assuming that the matrix $\boldsymbol{H}$ has full column rank $r \triangleq K(Q+\iota)$, the signal component of the covariance matrix $\boldsymbol{C}_r$, namely, $(\boldsymbol{H}\boldsymbol{H}^H)$, has rank r. Therefore we have

$$\lambda_i > \sigma^2 \quad \text{for} \quad i = 1, \ldots, r$$
$$\lambda_i = \sigma^2 \quad \text{for} \quad i = r+1, \ldots, NQ$$

By performing an eigendecomposition of the matrix $\boldsymbol{C}_r$, we obtain

$$\boldsymbol{C}_r = \boldsymbol{U}_s \boldsymbol{\Lambda}_s \boldsymbol{U}_s^H + \sigma^2 \boldsymbol{U}_n \boldsymbol{U}_n^H \tag{36}$$

where $\boldsymbol{\Lambda}_s = \text{diag}(\lambda_1, \ldots, \lambda_r)$ contains the r largest eigenvalues of $\boldsymbol{C}_r$ in descending order and $\boldsymbol{U}_s = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_r]$ contains the corresponding orthonormal eigenvectors; $\boldsymbol{U}_n = [\boldsymbol{u}_{r+1} \cdots \boldsymbol{u}_{NQ}]$ contains the $(NQ-r)$ orthonormal eigenvectors that correspond to the eigenvalue σ^2. It is easy to see that range $(\boldsymbol{H})$ = range $(\boldsymbol{U}_s)$. The column space of $\boldsymbol{U}_s$ is the signal subspace and the noise subspace is spanned by the columns of $\boldsymbol{U}_n$. The linear MMSE detector given by (33) can be expressed in terms of the abovementioned signal subspace components as

$$\boldsymbol{w}_1 = \alpha \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \overline{\boldsymbol{h}}_1 \tag{37}$$

with $\alpha \triangleq (\overline{\boldsymbol{h}}_1^H \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^H \overline{\boldsymbol{h}}_1)^{-1}$.

5.3. Blind Channel Estimation

It is seen from the preceding discussion that, unlike the synchronous case where the linear MMSE detector can be written in closed form once the signal subspace components are identified, in multipath channels, the composite signature waveform of the desired user, $\overline{\boldsymbol{h}}_1$, is needed to form the blind detector. It is essentially the channel distorted original spreading waveform $\boldsymbol{s}_1$. We next address the problem of blind channel estimation. It can be shown [37] that for each k, $1 \leq k \leq K$

$$h_k[n] = \sum_{j=0}^{N-1} c_{j,k} f_k[n-j], \quad n = 0, 1, \ldots, (\iota+1)N - 1 \tag{38}$$

with

$$f_k[m] \triangleq \frac{1}{\sqrt{N}} \sum_{l=1}^{L} \alpha_{l,k} \int_0^{T_c} \psi(t)\psi(t - \tau_{l,k} + mT_c), \quad m = 0, 1, \ldots, \mu - 1 \tag{39}$$

where the length of the discrete-time channel response $\{f_k[m]\}$ satisfies

$$\mu = \left\lceil \frac{\tau_{L,k}}{T_c} \right\rceil = \left\lceil \frac{\tau_{L,k}}{T} \cdot \frac{T}{T_c} \right\rceil \le \iota N \tag{40}$$

Denote

$$\boldsymbol{h}_k = \begin{bmatrix} h_k[0] \\ \vdots \\ h_k[(\iota+1)N-1] \end{bmatrix}_{(\iota+1)N\times 1}, \quad \boldsymbol{f}_k = \begin{bmatrix} f_k[0] \\ \vdots \\ f_k[\mu-1] \end{bmatrix}_{\mu\times 1}$$

and

$$\boldsymbol{\Xi}_k = \begin{bmatrix} c_{0,k} & & & \\ c_{1,k} & c_{0,k} & & \\ \vdots & c_{1,k} & \ddots & \\ \vdots & \vdots & \ddots & c_{0,k} \\ c_{N-1,k} & \vdots & & c_{1,k} \\ & c_{N-1,k} & & \vdots \\ & & \ddots & \vdots \\ & & & c_{N-1,k} \end{bmatrix}_{(\iota+1)N\times\mu}$$

Then (38) can be written in a matrix form as

$$\boldsymbol{h}_k = \boldsymbol{\Xi}_k \boldsymbol{f}_k \tag{41}$$

Recall that when the ambient channel noise is white, through an eigendecomposition on the autocorrelation matrix of the received signal, the signal subspace and the noise subspace can be identified. In order to estimate the desired user's composite signature waveform, it suffices to estimate the corresponding channel $\boldsymbol{f}_1$, which in turn can be estimated by exploiting the orthogonality between the signal subspace and the noise subspace [3,15,29,37]. Specifically, since $\boldsymbol{U}_n$ is orthogonal to the column space of $\boldsymbol{H}$, and $\overline{\boldsymbol{h}}_1$ is in the column space of $\boldsymbol{H}$ [cf. (35)], we have

$$\boldsymbol{U}_n^H \overline{\boldsymbol{h}}_1 = \boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1 = \boldsymbol{0} \tag{42}$$

where

$$\overline{\boldsymbol{h}}_1 = \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{0}_{(Q-\iota-1)N\times 1} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\Xi}_1 \\ \boldsymbol{0}_{(Q-\iota-1)N\times\mu} \end{bmatrix}}_{\overline{\boldsymbol{\Xi}}_k} \boldsymbol{f}_1 \tag{43}$$

From this relationship, we can obtain an estimate of the desired user's channel response $\boldsymbol{f}_1$, by computing the minimum eigenvector of the matrix $(\overline{\boldsymbol{\Xi}}_1^T \boldsymbol{U}_n \boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_1)$. The condition for the channel estimate obtained in such a way to be unique is that the matrix $(\boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_k)$ has rank $(\mu - 1)$, which necessitates this matrix to be tall: $[Nm - K(m + \iota)] \ge \mu$. Since $\mu \le \iota N$ [cf. (40)], we therefore choose the smoothing factor Q to satisfy

$$NQ - K(Q + \iota) \ge \iota N \ge \mu \tag{44}$$

That is, $Q = \lceil [(N-K)/(N+K)] \cdot \iota \rceil$. On the other hand, the condition (44) implies that for fixed Q, the total number of users that can be accommodated in the system is $\lceil [(Q-\iota)/(Q+\iota)] \cdot N \rceil$. Moreover, if the received signal is sampled at a multiple (p) of the chip rate, the total number of users that can be accommodated in the system is $\lceil [(Q-\iota)/(Q+\iota) \cdot Np \rceil$ [34].

Finally, we summarize the batch algorithm for blind linear multiuser detection in multipath CDMA channels as follows:

- *Estimate the signal subspace*:

$$\hat{\boldsymbol{C}}_r \triangleq \frac{1}{M} \sum_{i=0}^{M-1} \boldsymbol{r}[i]\boldsymbol{r}[i]^H = \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H + \hat{\boldsymbol{U}}_n \hat{\boldsymbol{\Lambda}}_n \hat{\boldsymbol{U}}_n^H$$

- *Estimate channel*:

$$\boldsymbol{f}_1 = \text{min-eigenvector } (\overline{\boldsymbol{\Xi}}_1^T \boldsymbol{U}_n \boldsymbol{U}_n^H \overline{\boldsymbol{\Xi}}_1)$$
$$\overline{\boldsymbol{h}}_1 = \overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1$$
$$\hat{\boldsymbol{w}}_1 = \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\boldsymbol{U}}_s^H \overline{\boldsymbol{h}}_1$$

- *Perform differential detection*:

$$z_1[i] = \hat{\boldsymbol{w}}_1^H \boldsymbol{r}[i]$$
$$\hat{\beta}_1[i] = \text{sign } \{\Re(z_1[i] z_1[i-1]^*)\}, \quad i = 1, \ldots, M-1$$

5.4. Adaptive Receiver Structure

We next consider an adaptive blind multiuser receiver in multipath CDMA channels based on the subspace linear MMSE detector. First, we address an adaptive implementation of the blind channel estimator discussed above. Suppose that the signal subspace $\boldsymbol{U}_s$ is known. Denote by $\boldsymbol{z}[i]$ the projection of the received signal $\boldsymbol{r}[i]$ onto the noise subspace:

$$\boldsymbol{z}[i] \triangleq \boldsymbol{r}[i] - \boldsymbol{U}_s \boldsymbol{U}_s^H \boldsymbol{r}[i] \tag{45}$$
$$= \boldsymbol{U}_n \boldsymbol{U}_n^H \boldsymbol{r}[i] \tag{46}$$

Since $\boldsymbol{z}[i]$ lies in the noise subspace, it is orthogonal to any signal in the signal subspace. In particular, it is orthogonal to $\overline{\boldsymbol{h}}_1 = \overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1$. Hence $\boldsymbol{f}_1$ is the solution to the following constrained optimization problem:

$$\min_{\boldsymbol{f}_1 \in \mathbb{C}^\mu} E\left\{ \|(\overline{\boldsymbol{\Xi}}_1 \boldsymbol{f}_1)^H \boldsymbol{z}[i]\|^2 \right\}, \quad \text{s.t.} \quad \|\boldsymbol{f}_1\| = 1 \tag{47}$$

Standard adaptive algorithms can be used to sequentially update $\boldsymbol{f}_1$ on the basis of the proceeding optimization criterion [26].

The block diagram of the subspace blind adaptive receiver is shown in Fig. 2. The received signal $\boldsymbol{r}[i]$ is

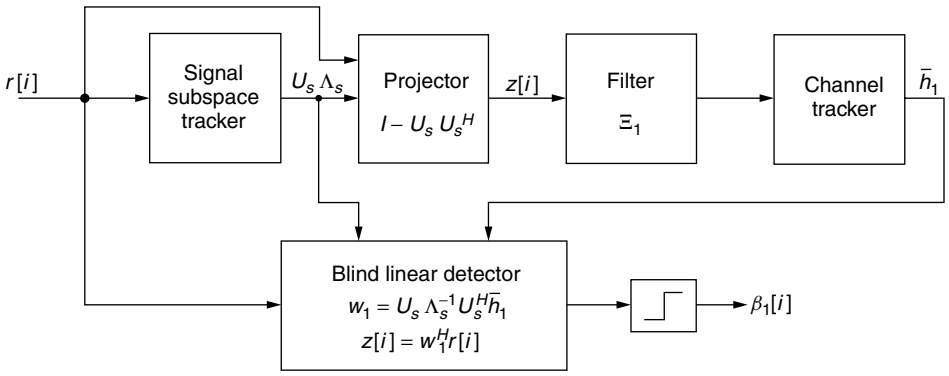


Figure 2. Diagram of a subspace blind adaptive receiver in multipath CDMA channels.

fed into a subspace tracker that sequentially estimates the signal subspace components ($\boldsymbol{U}_s$, $\boldsymbol{\Lambda}_s$). The signal $\boldsymbol{r}[i]$ is then projected onto the noise subspace to obtain $\boldsymbol{z}[i]$, which is in turn passed through a linear filter that is determined by the signature sequence $\boldsymbol{s}_1$ of the desired user. The output of this filter is fed into a channel tracker that estimates the channel state $\boldsymbol{f}_1$. Finally, the blind linear MMSE detector $\boldsymbol{w}_1$ is constructed in closed form, based on the estimated signal subspace components and the channel state. The adaptive receiver algorithm is summarized as follows. Suppose that at time $(i-1)$, the estimated signal subspace rank is $K[i-1]$ and the components are ($\boldsymbol{U}_s[i-1]$, $\boldsymbol{\Lambda}_s[i-1]$). The estimated channel vector is $\boldsymbol{f}_1[i-1]$. Then at time i, the adaptive detector performs the following steps to update the detector and to estimate the data:

- *Update the signal subspace*: Using a particular signal subspace tracking algorithm, update the signal subspace rank $K[i]$ and the subspace components ($\boldsymbol{U}_s[i]$, $\boldsymbol{\Lambda}[i]$).
- *Update the channel*: Using a particular adaptive algorithm, update the channel estimate $\boldsymbol{f}_1[i]$.
- *Form the detector and perform differential detection*:

$$\boldsymbol{w}_1[i] = \boldsymbol{U}_s[i]\boldsymbol{\Lambda}_s[i]^{-1}\boldsymbol{U}_s[i]^H\overline{\boldsymbol{\Xi}}_1\boldsymbol{f}_1[i],$$
$$z_1[i] = \boldsymbol{w}_1[i]^H\boldsymbol{r}[i],$$
$$\hat{\beta}_1[i] = \text{sign}\left\{\Re(z_1[i]z_1[i-1]^*)\right\}.$$

Simulation Example. We next give a simulation example on the performance of the blind adaptive receiver in an asynchronous CDMA system with multipath channels. The processing gain $N = 15$ and the spreading codes are Gold codes of length 15. Each user's channel has $L = 3$ paths. The delay of each path $\tau_{k,l}$ is uniform on $[0, 10T_c]$. Hence, the maximum delay spread is one symbol interval, namely, $\iota = 1$. The fading gain of each path in each user's channel is generated from a complex Gaussian distribution and is fixed for all simulations. The path gains in each user's channel are normalized so that each user's signal arrives at the receiver with the same power. The smoothing factor is $Q = 2$. The received signal is sampled at twice the chip rate ($p = 2$). Hence, the total number of users that this system can accommodate is 10. Figure 3 is shows the

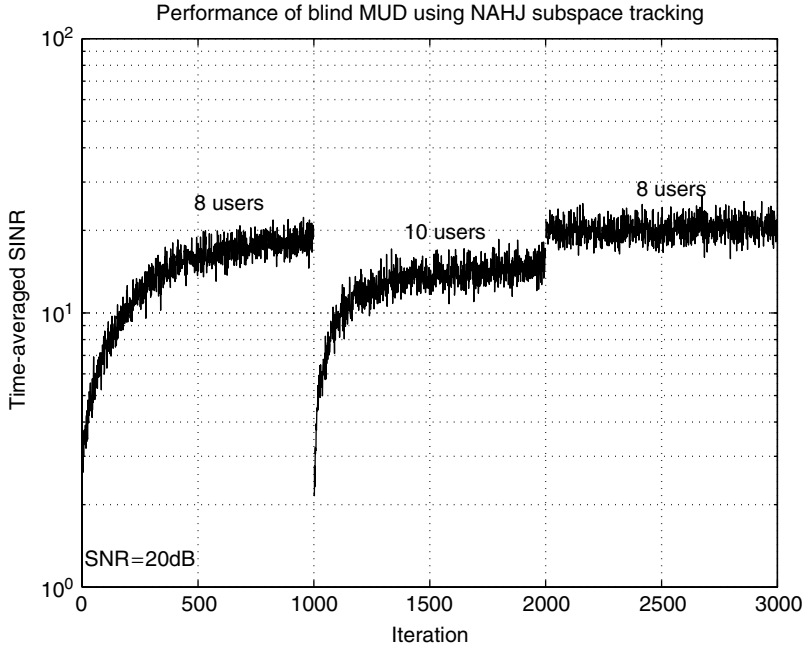


Figure 3. Performance of the subspace blind adaptive multiuser detector (MUD) in an asynchronous CDMA system with multipath.

performance of the subspace blind adaptive receiver using the NAHJ subspace tracking algorithm [26], in terms of SINR. During the first 1000 iterations there are 8 total users. At iteration 1000, 2 new users are added to the system. At iteration 2000, one additional known user is added and three existing users vanish. We see that this blind adaptive receiver can closely track the dynamics of the channel.

6. CONCLUDING REMARKS

In this article, we have presented signal processing techniques for blind multiuser detection in CDMA systems. The main objective is to perform interference suppression and signal detection in a CDMA downlink environment, with the prior knowledge of only the spreading waveform of the desired user. We have presented two approaches to blind multiuser detection — the direct matrix inversion (DMI) method and the subspace method. Note that in addition to what we discussed here, both approaches have been extended to address a number of other channel impairments. For example, under the DMI framework, techniques have been proposed to combat narrowband interference [22,23], channel dispersion [30,31] fading channels [2,14,35,42], and synchronization [16,17]. Moreover, within the subspace framework, extensions have been made to fading channels [27,36] and antenna array spatial processing [38], for blind adaptive joint channel/array response estimation, multiuser detection, and equalization. Another salient feature of the subspace approach is that it can be combined with the M-regression techniques to achieve blind adaptive robust multiuser detection in non-Gaussian ambient noise channels [39]. Furthermore, an analytical performance assessment of the DMI blind detector and the subspace blind detector is given in the literature [11–13,40]. Finally, we remark that in the CDMA uplink, typically the base station receiver has the knowledge of the spreading waveforms of all users within its cell, but not that of the users from other cells. *Group-blind* multiuser detection techniques have been developed [34] to address such scenarios.

Acknowledgments

This work is supported in part by the NSF grant CAREER CCR-9875314 and the NSF grant CCR 9980599. The author would like to thank Dr. Daryl Reynolds for providing the two simulation examples.

BIOGRAPHY

Xiaodong Wang received a B.S. degree in electrical engineering and applied mathematics (with the highest honor) from Shanghai Jiao Tong University, Shanghai, China, in 1992; an M.S. degree in electrical and computer engineering from Purdue University, West Lafayette, Indiana, in 1995; and a Ph.D degree in electrical engineering from Princeton University, New Jersey, in 1998. From July 1998 to December 2001 he was an assistant professor in the Department of Electrical Engineering, Texas A&M University. In January 2002, he joined the Department of Electrical Engineering, at Columbia University, New York, as an assistant professor.

Dr. Wang's research interests fall in the general areas of computing, signal processing, and communications. He has worked in the areas of digital communications, digital signal processing, parallel and distributed computing, nanoelectronics and quantum computing, and has published extensively in these areas. His current research interests include multiuser communications theory and advanced signal processing for wireless communications. He received the 1999 NSF CAREER Award, and the 2001 IEEE Communications Society and Information Theory Society Joint Paper Award. He currently serves as an associate editor for the *IEEE Transactions on Communications*, the *IEEE Transactions on Signal Processing*, and the *IEEE Transactions on Wireless Communications*.

BIBLIOGRAPHY

1. A. Abdulrahman, D. D. Falconer, and A. U. Sheikh, Decision feedback equalization for CDMA in indoor wireless communications, *IEEE J. Select. Areas Commun.* **12**(4): 698–706 (May 1994).
2. A. N. Barbosa and S. L. Miller, Adaptive detection of DS/CDMA signals in fading channels, *IEEE Trans. Commun.* **COM-46**(1): 115–124 (Jan. 1998).
3. S. E. Bensley and B. Aazhang, Subspace-based channel estimation for code-division multiple-access communication systems, *IEEE Trans. Commun.* **COM-44**(8): 1009–1020 (Aug. 1996).
4. C. H. Bischof and G. M. Shroff, On updating signal subspaces, *IEEE Trans. Signal Process.* **40**(1): 96–105 (Jan. 1992).
5. D.-S. Chen and S. Roy, An adaptive multiuser receiver for CDMA systems, *IEEE J. Select. Areas Commun.* **12**(5): 808–816 (June 1994).
6. P. Comon and G. H. Golub, Tracking a few extreme singular values and vectors in signal processing, *Proc. IEEE* **78**(8): 1327–1343 (Aug. 1990).
7. R. D. DeGroat, Noniterative subspace tracking, *IEEE Trans. Signal Process.* **40**(3): 571–577 (March 1992).
8. S. Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice-Hall, 1996.
9. M. Honig, U. Madhow, and S. Verdú, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **IT-41**(4): 944–960 (July 1995).
10. M. Honig and H. V. Poor, Adaptive interference suppression, in H. V. Poor and G. W. Wornell, eds., *Wireless Communications: A Signal Processing Perspective*, Prentice-Hall, Upper Saddle River, NJ, 1998, pp. 64–128.
11. A. Høst-Madsen and X. Wang, Performance of blind and group-blind multiuser detection, *IEEE Trans. Inform. Theory* **48**(6): (June 2002).
12. A. Høst-Madsen and X. Wang, Performance of blind and group-blind multiuser detectors, *Proc. 38th Annual Allerton Conf. Communications, Computing and Control*, Monticello, IL, Oct. 2000.
13. A. Høst-Madsen and X. Wang, Performance of blind multiuser detectors, *Proc. 10th Int. Symp. Information Theory and Its Applications* (*ISITA'00*), Honolulu, HI, Nov. 2000.

14. H. C. Huang and S. Verdú, Linear differentially coherent multiuser detection for multipath channels, *Wireless Pers. Commun.* **6**(1–2): 113–136 (Jan. 1998).

15. H. Liu and G. Xu, A subspace method for signal waveform estimation in synchronous CDMA systems, *IEEE Trans. Commun.* **COM-44**(10): 1346–1354 (Oct. 1996).

16. U. Madhow, Blind adaptive interference suppression for the near-far resistant acquisition and demodulation of direct-sequence CDMA signals, *IEEE Trans. Signal Process.* **45**(1): 124–136 (Jan. 1997).

17. U. Madhow, Blind adaptive interference suppression for CDMA, *Proc. IEEE* **86**(10): 2049–2069 (Oct. 1998).

18. U. Madhow and M. Honig, MMSE interference suppression for direct-sequence spread-spectrum CDMA, *IEEE Trans. Commun.* **COM-42**(12): 3178–3188 (Dec. 1994).

19. S. L. Miller, An adaptive direct-sequence code-division multiple-access receiver for multiuser interference rejection, *IEEE Trans. Commun.* **COM-43**(2–4): 1556–1565 (Feb.–April 1995).

20. U. Mitra and H. V. Poor, Adaptive receiver algorithms for near-far resistant CDMA, *IEEE Trans. Commun.* **COM-43**(2–4): 1713–1724 (Feb.–April 1995).

21. U. Mitra and H. V. Poor, Analysis of an adaptive decorrelating detector for synchronous CDMA channels, *IEEE Trans. Commun.* **COM-44**(2): 257–268 (Feb. 1996).

22. H. V. Poor and X. Wang, Code-aided interference suppression in DS/CDMA communications. Part II: Parallel blind adaptive implementations, *IEEE Trans. Commun.* **COM-45**(9): 1112–1122 (Sept. 1997).

23. H. V. Poor and X. Wang, Blind adaptive suppression of narrowband digital interferers from spread-spectrum signals, *Wireless Pers. Commun.* **6**(1–2): 69–96 (Jan. 1998).

24. P. B. Rapajić and B. S. Vučetić, Adaptive receiver structures for asynchronous CDMA systems, *IEEE J. Select. Areas Commun.* **12**(4): 685–697 (May 1994).

25. D. Reynolds and X. Wang, Adaptive group-blind multiuser detection based on a new subspace tracking algorithm, *IEEE Trans. Commun.* (in press).

26. D. Reynolds and X. Wang, Group-blind multiuser detection based on subspace tracking, *Proc. 2000 Conf. Information Sciences and Systems*, Princeton, NJ, March 2000.

27. Y. Song and S. Roy, Blind adaptive reduced-rank detection for DS-CDMA signals in multipath channels, *IEEE J. Select. Areas Commun.* **17**(11): 1960–1970 (Nov. 1999).

28. G. W. Stewart, An updating algorithm for subspace tracking, *IEEE Trans. Signal Process.* **40**(6): 1535–1541 (June 1992).

29. M. Torlak and G. Xu, Blind multiuser channel estimation in asynchronous CDMA systems, *IEEE Trans. Signal Process.* **45**(1): 137–147 (Jan. 1997).

30. M. K. Tsatsanis, Inverse filtering criteria for CDMA systems, *IEEE Trans. Signal Process.* **45**(1): 102–112 (Jan. 1997).

31. M. K. Tsatsanis and G. B. Giannakis, Blind estimation of direct sequence spread spectrum signals in multipath, *IEEE Trans. Signal Process.* **45**(5): 1241–1252 (May 1997).

32. D. W. Tufts and C. D. Melissinos, Simple, effective computation of principal eigenvectors and their eigenvalues and application to high resolution estimation of frequencies, *IEEE Trans. Acoust. Speech Signal Process.* **34**(5): 1046–1053 (Oct. 1986).

33. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, Cambridge, UK, 1998.

34. X. Wang and A. Høst-Madsen, Group-blind multiuser detection for uplink CDMA, *IEEE J. Select. Areas Commun.* **17**(11): 1971–1984 (Nov. 1999).

35. X. Wang and H. V. Poor, Adaptive joint multiuser detection and channel estimation in multipath fading CDMA channels, *ACM/Baltzer Wireless Networks*, 1998, pp. 453–470.

36. X. Wang and H. V. Poor, Blind adaptive multiuser detection in multipath CDMA channels based on subspace tracking, *IEEE Trans. Signal Process.* **46**(11): 3030–3044 (Nov. 1998).

37. X. Wang and H. V. Poor, Blind equalization and multiuser detection for CDMA communications in dispersive channels, *IEEE Trans. Commun.* **COM-46**(1): 91–103 (Jan. 1998).

38. X. Wang and H. V. Poor, Blind multiuser detection: A subspace approach, *IEEE Trans. Inform. Theory* **44**(2): 677–691 (March 1998).

39. X. Wang and H. V. Poor, Robust multiuser detection in non-Gaussian channels, *IEEE Trans. Signal Process.* **47**(2): 289–305 (Feb. 1999).

40. X. Wang, J. Zhang, and A. Høst-Madsen, Blind and group-blind multiuser detection: Effect of estimation error and large system performance. Invited talk at the 2001 IEEE Communication Theory Workshop, Borrego Springs, CA, April 2001.

41. B. Yang, Projection approximation subspace tracking, *IEEE Trans. Signal Process.* **44**(1): 95–107 (Jan. 1995).

42. L. J. Zhu and U. Madhow, Adaptive interference suppression for direct sequence CDMA over severely time-varying channels, *Proc. IEEE Globecom'97*, Nov. 1997.

BLUETOOTH RADIO SYSTEM

JAAP C. HAARTSEN
Ericsson Technology
Licensing AB
Emmen, The Netherlands

1. INTRODUCTION

Progress in microelectronics and VLSI technology has fostered the widespread use of computing and communication devices for commercial usage. The success of consumer products such as notebooks, laptops, personal digital assistants (PDAs), cell phones, cordless phones, and their peripherals has been based on continuous cost and size reduction. Information transfer between these devices has been cumbersome mainly relying on cables or infrared. Although infrared transceivers are inexpensive, they have limited range, are sensitive to directions and to objects in the propagation path, and can in principle be used only between two devices. By contrast, radio transceivers have much larger range, can propagate through various materials and around objects, and can connect many devices simultaneously. A new universal radio interface has been developed enabling electronic devices to communicate wirelessly via *short-range radio* connections. The *Bluetooth technology* — which has gained the support from leading manufacturers in the telecom, PC

and consumer industry — eliminates the need for wires, cables, and the corresponding connectors, and paves the way for completely new devices and applications. The Bluetooth technology provides a solution for access to information and personal communication by enabling connectivity between devices in proximity of each other, allowing each device to keep its inherent function based on its user interface, form factor, cost, and power constraints. Radio technology will allow this connectivity to occur without any explicit user interaction. The Bluetooth technology has been optimized with respect to low-power, small-size, and low-cost, enabling singe-chip radios that can be embedded into these personal devices.

2. AD HOC COMMUNICATIONS

Most radio systems in use today rely on a fixed infrastructure. Cellular phone systems like GSM, IS136, or IS95 [1] obtain regional coverage by applying a wired backbone network using a multitude of base stations placed at strategic positions to provide local cell coverage. The mobile users apply mobile terminals to access this public land mobile network (PLMN); the terminals maintain a connection to the network via a radio link to the base stations. There is a strict separation between the fixed base stations and the mobile terminals. Once registered to the mobile network, the terminals remain locked to the control channels on the radio interface and connections can be established and released according to the control protocols. Channel access, channel allocation, traffic control, and interference minimization is taken care of by intelligent centers in the network that also coordinate the activity of the base stations. The basic architecture of a *mobile system* is shown in Fig. 1. The mobile network, which handles all mobility issues, is strictly separated from the fixed public switched telephone network (PSTN). Gateways between the PLMN and PSTN provide a smooth connection between the mobile and fixed telephony world. Alternatively, the base stations can directly be connected to the public switched network as shown in Fig. 2. In this case, the radio interface forms a *wireless extension* of the wired network. Because of a lack of coordination between the base stations, radio functions like channel access, channel allocation, traffic control, and interference mitigation must now be dealt with by the base stations independently. Still in the wireless extensions, there

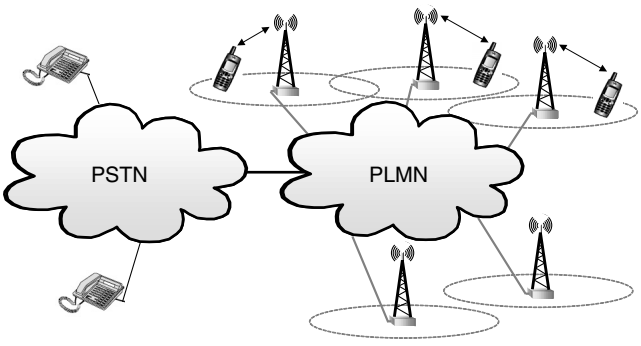


Figure 1. Mobile system architecture.

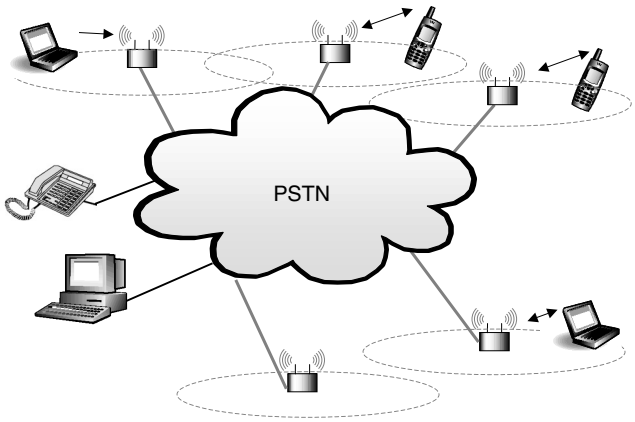


Figure 2. Wireless extension architecture.

are cells defined by the base stations and there is a strict separation between the fixed base stations and the mobile terminals. Mutual interference can reasonably be controlled by proper selection of channels. Residential and office cordless phones for example based on DECT [2] provide wireless extensions to the PSTN, whereas, for example, WLAN 802.11 or Hiperlan2 [3] provide wireless extensions to the Ethernet or ATM networks.

In *ad hoc systems*, there is no distinction between radio units; that is, there are no separate base stations or terminals. Ad hoc connectivity is based on *peer* communications. There is no wired infrastructure to support the connectivity between the portable units; there is no central controller for the units to rely on for making connections nor is there support for coordination of communications. Some WLAN systems do have an ad hoc mode where terminals can make connections without the intervention of a base station. However, in these ad-hoc scenarios, a single channel is created to which all units in range are connected as illustrated in Fig. 3. Base-station-like functions are shared among the mobile terminals. By contrast, Bluetooth is based on device-to-device connections where only two or a few mobile units share the same channel. For the scenarios envisioned by Bluetooth, it is highly likely that a large number of independent connections coexist in the same area without any mutual coordination; that is, tens of ad hoc links must share the same radio spectrum at the same location in an uncoordinated fashion. This will be indicated as a *scatter ad-hoc environment* (see Fig. 4).

Ad hoc radio systems have been in use for some time, for example, walkie-talkie systems used by the military, the police, the fire brigade and by rescue teams in general. However, the Bluetooth system is the first commercial ad-hoc radio system envisioned to be used on a large scale and widely available to the public.

3. SPECTRAL COEXISTENCE

3.1. Unlicensed Radio Band

The lack of a geographically fixed infrastructure (i.e., ad hoc networks can be considered floating) necessitates

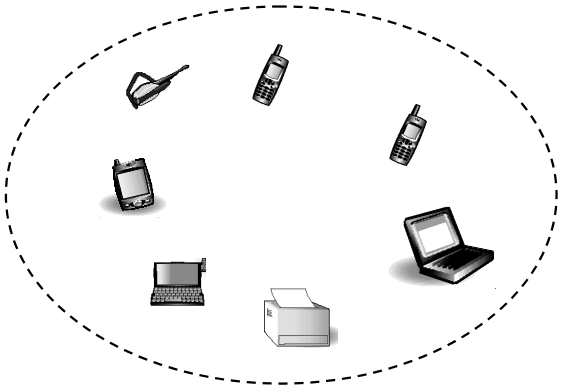


Figure 3. Single ad hoc network.

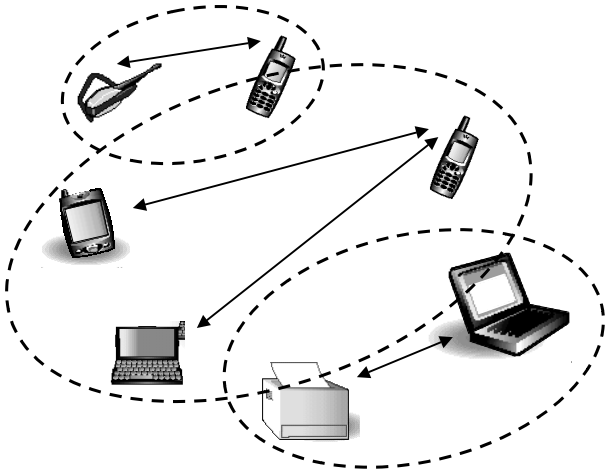


Figure 4. Scatter ad hoc network.

the deployment of a radioband that is globally available. Indeed, a user can setup an ad hoc connection anywhere in the world without the interaction of an operator or another third party. In addition, consumer-targeted applications necessitate the deployment of a radioband that is unlicensed. The most suitable band for these ad hoc applications is the industrial–medical–scientific (ISM) band ranging from 2400 to 2483.5 MHz. This band was formerly reserved for some professional user groups but has recently been opened worldwide for commercial use. The operating rules have been set by regulatory bodies such as the FCC in the United States [4], the CEPT in Europe [5], and ARIB in Japan [6]. Although the rules per region may differ slightly, their scope is to enable a fair access to the radioband by any user. The regulations generally specify the spreading of the transmitted signal energy and the maximum allowable transmit power. For a system to operate globally, a radio concept has to be found that satisfies all regulations simultaneously. The Bluetooth standard, therefore, satisfies the minimum denominator of all the requirements.

Radio propagation at 2.45 GHz provides reasonable coverage with relatively low transmit power. Current state-of-the-art radio technology allows highly integrated radio transceivers to operate with low current consumption which can be manufactured at low cost, a prerequisite for *embedded radio systems*.

3.2. Spectrum Sharing

The consequence of an unlicensed and open band is the abundance of different (radio) systems encountered in this band. Applications range from garage-door openers to microwave ovens. Also the wireless LAN systems based on 802.11 can be found in this band. The extent and nature of the interference in the 2.45-GHz ISM band cannot be predicted. With high probability, the different systems sharing the same band will not be able to communicate. Coordination is, therefore, not possible. A larger problem pose the high-power transmitters covered by the FCC part 18 rules that, for example, include microwave ovens and lighting devices. These devices fall outside the power and spreading regulations of Part 15 but still coexist in the 2.45-GHz ISM band. In addition to interference from external sources, couser interference resulting from other Bluetooth users in close proximity must be taken into account in the scatter ad hoc scenario.

Interference resistance can be obtained by interference suppression or interference avoidance. Suppression can be obtained by coding or by direct-sequence spreading. However, the dynamic range of the interfering and intended signals in a scatter ad hoc environment can be huge. Taking into account the distance ratios and the power differences of uncoordinated transmitters, near : far ratios in excess of 50 dB are no exception. With the desired user rates in the order of 1 Mbps (megabits per second) and beyond, practically attained coding and processing gains are inadequate. Instead, interference avoidance is more attractive as the desired signal is transmitted at locations in frequency and or time where interference is low or absent. Avoidance in time can be an alternative if the interference concerns a pulsed jammer and the desired signal can be interrupted. This requires coordination in time though. Avoidance in frequency is more practical. Since the 2.45-GHz band provides about 80 MHz of bandwidth and most radio transmissions are band-limited, with high probability parts of the radio spectrum can be found where there is no dominant interference. Filtering in the frequency domain provides the suppression of the interferers at other parts of the radio band. The filter suppression can easily arrive at 50 dB or more.

3.3. Frequency Hop Spread Spectrum

The selection of the multiple access scheme for ad hoc radio systems is driven by the lack of coordination and by the regulations in the ISM band. FDMA is attractive for ad hoc systems since channel orthogonality relies only on the accuracy of the crystal oscillators in the radio units. Combined with an adaptive channel allocation scheme, interference can be avoided. Unfortunately, pure FDMA does not fulfill the spreading requirements set in the ISM band by the FCC rules Part 15 [4]. TDMA requires a strict timing synchronization for channel orthogonality. For multiple collocated ad hoc connections, maintaining a common timing reference

becomes rather cumbersome. CDMA offers the best properties for ad hoc radio systems since it provides spreading and can deal with uncoordinated systems. DSCDMA is less attractive because of the *near–far* problem, which requires coordinated power control or excessive processing gain. In addition, as in TDMA, direct-sequence orthogonality requires a common timing reference. For higher user rates, DSCDMA requires rather high chip rates which are less attractive because of the wide bandwidth (interference resistance) and the higher current consumption. FHCDMA (Frequency-hopped CDMA) combines a number properties, which makes it the best choice for ad hoc radio systems. On average the signal can be spread over a wide frequency range, but instantaneously only a small bandwidth is occupied avoiding most of the potential interference in the ISM band. The hop carriers are orthogonal in frequency, and the interference on adjacent hops can effectively be suppressed by filtering. The hop sequences will not be orthogonal, though, but narrowband and couser interference is experienced as short interruptions that can be overcome with measures at higher-layer protocols. Frequency hopping was originally introduced in World War II for the remote control of torpedoes. The robustness of FH made it an ideal candidate for reliable transmission in a hostile environment. It was a Hollywood actress, Hedy Lamarr, who invented the procedure together with her pianist George Antheil [7].

Frequency hopping enables low-cost and low-power radio implementations. Since the instantaneous bandwidth is relatively narrow, conventional radio technology can be used. Also, truly single-chip solutions become feasible [8].

4. BLUETOOTH RADIO INTERFACE

4.1. Bluetooth Protocol Stack

The Bluetooth protocol stack [9] has been drafted along the OSI layered architecture but renamed (see Fig. 5). The four lower layers are the Bluetooth-specific protocols. At the RF layer, all radio-related operations are defined like modulation, frequency generation, filtering, spectral shaping, and so on. At the baseband layer, operations on the packet level are defined, such as error correction, encryption, and retransmissions. The Link Manager Protocol (LMP) takes care of control functions like authentication, setup of connections, traffic scheduling, link supervision, and power management tasks. The Logical Link Control and Adaptation Protocol (L2CAP) is an intermediate layer between the Bluetooth-specific protocols and more general protocols. It handles the multiplexing of higher-layer protocols and the segmentation and reassembly of large packets. Real-time traffic such as audio bypasses the L2CAP and LMP layers and streams into the baseband layer directly. Yet, the audio stream is controlled (non-real-time) by the LM. For more information about application oriented profiles, the user is referred to Ref. 10 or 11.

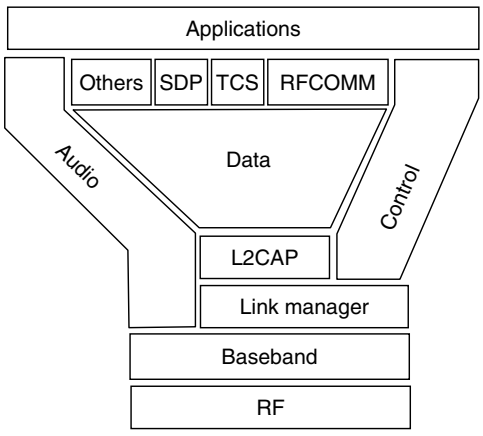


Figure 5. The Bluetooth protocol stack.

4.2. RF Layer

Bluetooth deploys FHCDMA. Each channel makes use of a different hop code or hop pattern. The radios hop over 79 carriers. These carriers have been defined at a 1-MHz spacing in the 2.45-GHz ISM band. The nominal dwell time is 625 μs. The Bluetooth radios hop with a nominal rate of 1600 hops/s. In the time domain, the channel is divided into time slots. The dwell time of 625 μs corresponds to a single slot. To simplify implementation, full-duplex communications is achieved through time-division duplex (TDD). This means that a unit alternately transmits and receives. Separation of transmission and reception in time effectively prevents crosstalk between the transmit and receive operations in the radio transceiver, which is essential if a one-chip implementation is desired. Since transmission and reception take place at different time slots, transmission and reception also take place at different hop carriers. Figure 6 illustrates the FHTDD channel applied in Bluetooth. Note that different ad hoc links will make use of different hopping sequences and will have misaligned slot timing.

The instantaneous bandwidth of the transmitted spectrum is limited to 1 MHz. For robustness, a binary modulation scheme is used. With the abovementioned bandwidth restriction, the data rates are limited to about 1 Mbps. For FH systems and support for bursty data traffic, a noncoherent detection scheme is most appropriate. Bluetooth uses a Gauss-shaped FSK modulation with a nominal modulation index of $h = 0.3$. Logical ones are sent as positive frequency deviations, logical zeros as negative

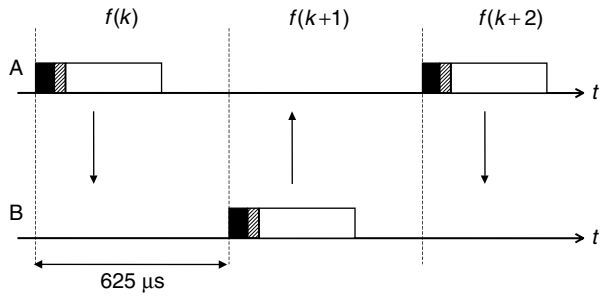


Figure 6. The slotted FH/TDD channel in Bluetooth.

frequency deviations. GFSK provides a constant envelope modulation, which is insensitive to nonlinear operations. Transmission can be achieved with a class C amplifier, while demodulation can simply be accomplished by a limiting FM discriminator. GFSK allows the implementation of low-cost and low-power radio units.

4.3. Baseband Layer

The baseband layer handles all crucial communication procedures such as connection setup, hop pattern selection and hop synchronization, traffic scheduling, and all operations at the packet level. Since most of the baseband operations occur in real time, the baseband processes have been optimized to enable dedicated hardware implementations, thus preserving power consumption. Basically, for a Bluetooth unit in idle mode, no CPU has to be activate.

4.3.1. Connection Setup. A critical design issue in ad hoc radio systems is the connection establishment. How do units find each other, and how do they make connections? In Bluetooth, three elements have been defined to support the connection establishment: *scan*, *page*, and *inquiry*. A unit that is in idle mode wants to "sleep" most of the time to save power. However, in order to allow connections to be made, the unit frequently has to listen as to whether other units want to connect. In ad hoc systems, there is no common control channel that a unit can lock to in order to listen for page messages as is common in conventional (cellular) radio systems. In Bluetooth, a unit periodically "wakes up" to listen for a page message. This page message consists of an access code which is derived from the unit's identity. When a Bluetooth receiver wakes up to scan, it opens a sliding correlator that is matched to the access code derived from its own identity. The scan window is a little longer than 10 ms. Every time the unit wakes up, it scans at a different hop carrier. This is required by the regulations that do not permit a fixed wakeup frequency; it also provides the necessary interference immunity. The Bluetooth wakeup hop sequence covers only 32 carriers and is cyclic. All 32 carriers in the wakeup sequence are unique and they span about 64 MHz of the 80 MHz available. The wakeup sequence is pseudorandom and unique for each Bluetooth device; like the access code, the sequence is derived from the unit's identity. The phase in the sequence is determined by a free-running native clock in the unit. It will be understood that a tradeoff has to be made between idle mode power consumption and the response time; increasing the sleep time T will reduce power consumption but prolong the time before an access can be made. The unit that wants to connect has to resolve the frequency–time uncertainty; it does not know when the idle unit will wake up and on what carrier frequency. The burden of resolving this uncertainty is deliberately placed at the paging unit because this will require power consumption. Since a radio unit will most of the time be in idle mode, the paging unit should take the power burden. First we assume that the paging unit knows the identity of the unit that it wants to connect to. It then knows the wakeup sequence and can also generate the access code that serves as the page message. The paging unit then transmits the access code repeatedly at different frequencies; every 1.25 ms, the paging units transmits two access codes and listens twice for a response (see Fig. 7). The access code is transmitted consecutively on different carrier frequencies selected from the wakeup sequence. In a 10-ms period, 16 different frequencies are visited, which covers half of the wakeup sequence. The paging unit transmits the access code on these 16 frequencies cyclically for the duration of the sleep period T of the idle unit. If the idle unit wakes up on any of these 16 frequencies, it will receive the access code and a connection setup procedure follows. However, since the paging unit does not know the clock that the idle unit is using, the idle unit can equally well wake up in any of the 16 remaining frequencies in the 32-hop wakeup sequence. Therefore, if the paging unit does not receive a response from the idle unit after a time corresponding to the sleep time T, it will transmit the access code repeatedly; the response time therefore amounts to twice the sleep time T. When the idle unit receives the page message, it notifies the paging unit by returning a message that again is the access code derived from the idle unit's identity. Thereafter the paging unit transmits a control packet that contains all of the pager's information (e.g., identity and clock). This information is then used by both the paging unit and the idle unit to establish a FH channel.

The above-described paging process assumed that the paging unit had no knowledge at all of the clock in the idle unit. However, if the units have met before, the paging unit will have an estimate of the clock in the idle unit. When units connect, they exchange their clock information and the time offset between their free-running native clocks is stored. This offset is only accurate during the connection; when the connection is released, the offset information becomes less reliable as a result of clock drifts. The reliability of the offset is inversely proportional to the time elapsed since the last connection. Yet, the paging unit can exploit the offset information to estimate the clock of the idle unit. Suppose, that the clock estimate of the idle unit in the paging unit is k'. If $f(m)$ represents the frequency in the wakeup sequence at time m, the paging unit will assume the idle unit will wake up in $f(k')$. But since in 10 ms it can cover 16 different frequencies, it will also transmit the access code on a few frequencies before and after $f(k')$ or $f(k'-8), f(k'-7), \ldots, f(k'), f(k'+1), \ldots, f(k'+7)$. As a result, the clock estimate in the paging unit can be off by -8 or $+7$ while it still covers the wake-up frequency of the unit in idle mode. With a free-running clock accuracy of ± 250 ppm, the clock estimate k' is still useful at least 5 h after the last connection. In that case, the average response time is reduced to half the sleep time T.

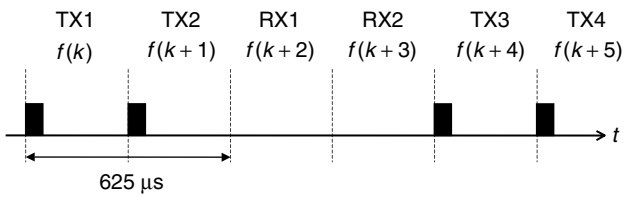


Figure 7. Transceiving routine for the paging unit.

To establish a connection, the identity of the recipient is required to determine the page message and the wakeup sequence. If the identity is not known, a unit that desires to make a connection may broadcast an inquiry message that induces recipients to return their identity and clock information. With the inquiry procedure, the inquirer can determine which units are within range and what their characteristics are. The inquiry message is again an access code, but derived from a reserved identity (the inquiry address). Idle units also listen to the inquiry message according to a 32-hop inquiry sequence. Units that receive the inquiry message return a control packet that includes their identity and clock information. For the return of the control packet a random backoff mechanism is used to prevent multiple recipients to transmit simultaneously. (In determining the frequencies of the second half of the sequence, the paging unit takes into account that the clock in the idle unit also progresses. The remaining half will therefore have one frequency in common with the first half.)

4.3.2. Hop Selection Mechanism. To allow for many overlapping hop channels (scatter ad hoc networking), a large number of pseudorandom hopping sequences have been defined. Note that no extra effort has been taken to make the sequences orthogonal. With only 79 carriers to hop over, the number of orthogonal sequences is rather limited. Bluetooth applies a special hop selection mechanism. The hop selection mechanism uses an identity and clock as inputs, and a carrier frequency as output (see Fig. 8). The identity selects a particular hop sequence while the clock points at a particular phase in this hop pattern. As the clock progresses at a rate of 1600 ticks/s, 1600 hops/s are produced by the output. By changing either the identity or the clock, instantaneously another carrier frequency is produced. The selection box contains only combinatorial logic and is memoryless. Prestored sequences are not feasible because of the large number of sequences required. The hop sequence is very long, and at a rate of 1600 hops/s, it takes more than 23 h to arrive at the exact phase in the sequence again. This feature prevents repetitions in the interference pattern when several hopping channels are collocated. Repetitive interference is detrimental for real-time services such as voice. Any segment of 32 consecutive hops in the sequence spans about 64 MHz of spectrum. By spreading as much as possible over a short time interval, maximal interference immunity is obtained. This is most important for real-time services.

We will now have a closer look at the selection scheme in Fig. 8. In the first block, the identity selects a 32-hop subsequence with pseudorandom properties. The least significant part of the clock hops through this sequence with a rate of 1600 hops/s. The first block thus provides an index in a 32-hop segment. The segments are mapped on the 79-hop carrier list. The carrier list is constructed in such a fashion that even-numbered hops are listed in the first half of the list, whereas the odd-numbered hops are listed in the second half of the list. An arbitrary segment of 32 consecutive list elements spans about 64 MHz. For the paging and inquiry procedures, the mapping of the 32-hop

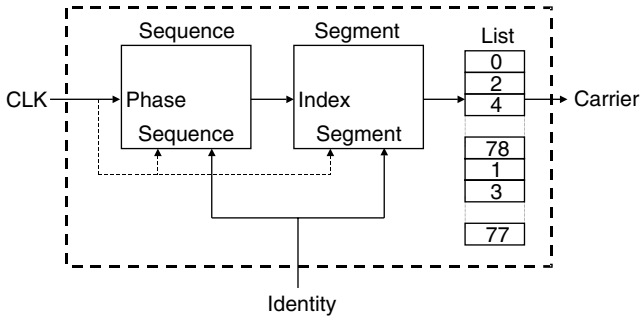


Figure 8. Carrier selection mechanism. Dashed line indicates that more significant clock part is used in connection mode only.

segment on the carrier list is fixed. When the clock runs, the same 32-hop sequence and the same 32 hop carriers will repeatedly be used. However, different identities will map to different segments and different sequences. So the wakeup hop sequences of different units are well randomized. During the connection, the more significant part of the clock affects both the sequence selection and the segment mapping; after 32 hops (one segment) the sequence is altered, and the segment is shifted in the forward direction by half its size (16 hops). Segments, each 32 hops in length, are concatenated, and the random selection of the index changes for each new segment; the segments slide through the carrier list and on average, all carriers are visited with equal probability. Changing the clock and/or identity will directly change the sequence and the segment mapping.

4.3.3. Packet-Based Communications. The Bluetooth system uses packet-based transmission; the information stream is fragmented into packets. In each TDD slot, only a single packet can be sent. All packets have the same format, starting with an access code, followed by a packet header and ending with the user payload (see Fig. 9). The access code has pseudorandom properties. All packets exchanged on the channel are preceded by the same access code. Only if the access code matches the access code corresponding to the channel will the packet be accepted by the recipient. This prevents packets sent in one FH channel falsely being accepted by units of another FH channel that happens to land on the same hop carrier. In the receiver, the access code is matched against the anticipated code in a sliding correlator. The packet header contains link control information: a 3-bit MAC address to separate the units on the channel, a 1-bit ACK/NAK for the retransmission scheme, a 4-bit packet type code to define 16 different payload types, and an 8-bit header error check (HEC) code which is a cyclic redundancy check (CRC). The packet header is limited to 18 information bits in order to restrict the overhead.

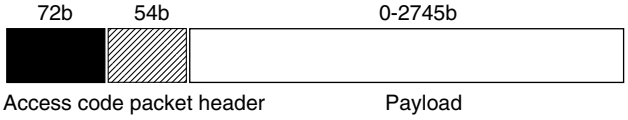


Figure 9. Format of packets exchanged on the FH/TDD channel.

The header is further protected by a rate-$\frac{1}{3}$ forward error control (FEC) coding.

Bluetooth defines four control packets:

1. *The ID or identification packet* — this packet only consists of the access code and is used for signaling.
2. *The NULL packet* — this packet only has an access code and a packet header and is used if link control information carried by the packet header has to be conveyed.
3. *The POLL packet* — this packet is similar to the NULL packet, and can be used in a polling procedure.
4. *The FHS packet* — this is a FH synchronization packet and is used to exchange real-time clock and identity information between the units. The FHS packet contains all the information to get two units hop synchronized.

The remaining 12 type codes are used to define packets for synchronous and asynchronous services. These 12 types are divided into 3 segments; segment 1 specifies packets that fit into a single slot, segment 2 specifies 3-slot packets, and segment 3 specifies 5-slot packets. Multislot packets have been introduced to increase the user data rate. They are sent on a single carrier frequency (see also Fig. 10). Note that packets can cover only an odd number of TDD slots, which guarantees that the TX/RX timing is maintained. The payload length is variable and depends on the amount of user data available. However, the maximum length is limited by the minimum switching time between RX and TX, which is specified at 200 μs. This switching time seems large, but allows the use of open-loop VCOs for direct modulation and provides time for packet processing between RX and TX.

4.3.4. Physical Links. The Bluetooth link supports both synchronous services such as voice traffic and asynchronous services such as bursty data traffic. Two physical link types have been defined:

1. The synchronous connection-oriented (SCO) link
2. The asynchronous connection-less (ACL) link

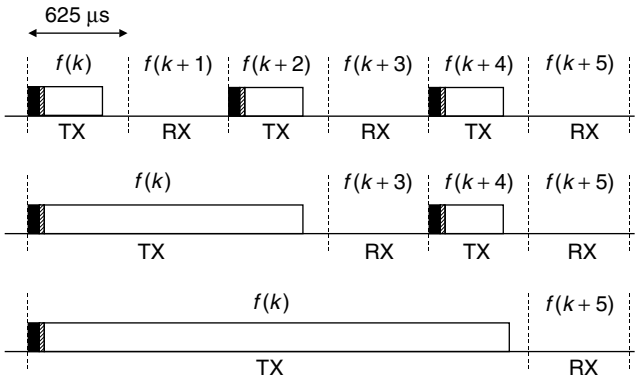


Figure 10. Characteristics of single-slot, three-slot, and five-slot packets.

The *SCO link* is a point-to-point link between two units on the FH channel. The link is established by reservation of duplex slots at regular intervals. This link can be used to carry synchronous, real-time traffic and provides quality of service. The *ACL link* is a point-to-multipoint link which is used to carry asynchronous, best-effort traffic.

Different packet types have been defined for the SCO link and the ACL link. The ACL links support payloads with or without a FEC coding scheme. In addition, on these links single-slot, 3-slot, and 5-slot packets are available. The maximum user rate that can be obtained over the ACL link is 723.2 kbps. In that case, a return link of 57.6 kbps can still be supported. If propagation conditions change, link adaptation can be applied on the ACL link by changing the packet length and the FEC coding. For the SCO link, only single-slot packets have been defined. The payload length is fixed and may choose between two different FEC schemes and no FEC. The SCO link supports a full-duplex link with a user rate of 64 kbps in both directions.

4.3.5. Error Correction Schemes. In Bluetooth, two different FEC schemes are supported. The rate-$\frac{1}{3}$ FEC code merely uses a 3-bit repeat coding with majority decisions at the recipient. With the repeat coding, extra gain is obtained because of the reduction in the instantaneous bandwidth. As a result, intersymbol interference (ISI) introduced by the receive filtering is decreased. This code is used for the packet header, and can additionally be applied on the payload of the SCO link packets. For the rate-$\frac{2}{3}$ FEC code, a (15,10) shortened Hamming code is used. Error trapping can be applied for decoding. This code can be used in the payload of both SCO link and ACL link packets. The applied FEC codes are very simple and fast to encode and decode, which is a requirement because of the limited processing time between RX and TX.

On the ACL link, an automatic retransmission query (ARQ) scheme is applied. In this scheme, a packet retransmission is carried out if the reception of the previous packet is not acknowledged. Each ACL payload contains a CRC to check for errors. To minimize complexity, overhead, and wasteful retransmissions, Bluetooth has implemented a fast-ARQ scheme where the sender is notified of the packet reception in the first packet in the return path after the transmission. The ACK/NAK information is piggybacked in the packet header of the return packet. There is only the RX/TX switching time for the recipient to determine the correctness of the received packet and creating the ACK/NAK field in the header of the return packet. In addition, the ACK/NAK field in the header of the packet received indicates whether the previously sent payload was correctly received, and thus determines whether a retransmission is required or the next packet can be sent. This process is illustrated in Fig. 11. Due to the short processing time, decoding is preferably carried out on the fly while the packet is received. The simplicity of the FEC coding schemes speed up the processing.

4.3.6. Low-Power Modes. In the Bluetooth system, special attention has been paid to the reduction of current consumption. In the idle mode, the unit scans for only about 10 ms every T seconds where T can range from 1.28

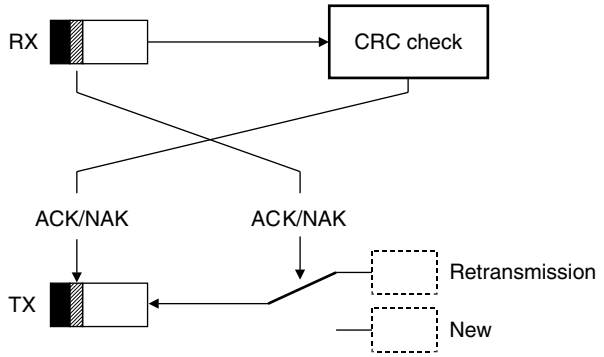


Figure 11. ARQ routine: received ACK/NAK information decides on retransmission; received payload determines returned ACK/NAK.

to 3.84s. So the duty cycle is well below 1%. Additionally, a PARK mode has been defined where the duty cycle can be reduced even more. However, the PARK mode can be applied only after the channel has been established. Units can then be parked; they listen to the channel at only a very low duty cycle. The unit only has to listen to the access code and the packet header (126 μs, excluding some guard time to account for drift) to resynchronize its clock and to decide whether it can return to sleep. Since there is no uncertainty in time and frequency (the parked unit remains locked to the channel, in much the same way as cordless and cellular phones are locked to their base stations), a much lower duty cycle is obtainable. Another low-power mode during connection is the SNIFF mode, in which a unit now and then listens to the channel, but is still considered to be active. Finally, a HOLD mode has been defined. A unit can enter the HOLD mode for a predefined time duration. During this time duration, it is not active on the channel. When the HOLD timeout expires, the unit automatically returns to the channel and is active again.

4.4. Link Manager

The link manager is involved in non-real-time, supervisory, and control operations. It takes care of attachment and detachments of units (note that real-time connection setup routines like page and scan are carried out at the baseband layer), sets up SCO and ACL links, and initiates low-power modes such as HOLD, SNIFF, and PARK. The LMP also takes care of security operations; it runs authentication procedures to prevent unauthorized usage of the connections and initiates the encryption routines. In addition, it is responsible for the key distribution. After the unit has been attached and been authenticated and logical links have been established, the link manager is involved in link supervision; it checks whether the device is still in coverage range of other units, and is responsible for selecting the proper packet type depending on the link quality and the required quality of service. Also adaptive power control is handled at the LMP level. The link manager is responsible for traffic scheduling. Finally, the link manager is used to configure (optional) parameters in the baseband layer both at connection setup and during the connection.

LMP messages are control messages and fit into the payload of a single-slot baseband packet using a rate-$\frac{2}{3}$ FEC. The LMP PDU consists of a header and a body. The header contains an 7-bit opcode specifying the LMP message, and a 1-bit transaction ID. The body may contain additional parameters used by the LMP message.

4.5. Logical Link and Adaptation Protocol

The Logical Link Control and Adaptation Protocol (L2CAP) forms an interface layer between the Bluetooth LMP layer on one hand and Bluetooth-independent higher layers on the other hand. The L2CAP handles the multiplexing of different logical links over the ACL link. It also effects segmentation and reassembly of datagram packets provided by the higher layers, for example, IP packets, onto the baseband packets. The logical links are connection-oriented with an 16-bit destination address in the L2CAP header.

5. BLUETOOTH NETWORKING

Bluetooth has been optimized to allow a large number of uncoordinated communications to take place in the same area. Unlike other ad hoc solutions where all units in the same range share the same channel, Bluetooth has been designed to allow a large number of independent channels, each channel serving only a limited number of participants. With the considered narrowband modulation scheme, a single FH channel in the ISM band only supports a gross bit rate of 1 Mbps. This capacity has to be shared among all participants on the channel. In the user scenarios targeted by Bluetooth, it is highly unlikely that all units in range need to share all information among all of them. By using a large number of independent 1-Mbps channels to which only the units are connected that really want to exchange information, the 80 MHz is exploited much more effectively.

In Bluetooth, a FH channel is designated as a *piconet*. On a single piconet, up to eight units can participate. Each piconet has its own, unique hop sequence. The particular sequence is determined by the identity of the unit that controls the FH channel, which is called the "master." The native clock of the master unit defines the phase in the hopping sequence. All other participants on the hopping channel are "slave"; they use the master identity to select the same hop sequence. Each Bluetooth radio unit has a free-running system or native clock. There is no common timing reference, but when a piconet is established, the slaves add an offset to their native clocks to synchronize to the master. These offsets are released again when the piconet is cancelled, but can be stored for later use. Different channels have different masters and therefore also different hop sequences and phases. Bluetooth is based on peer communications. The master/slave role is only attributed to a unit for the duration of the piconet. When the piconet is released, the master and slave roles are cancelled. Each unit can become a master or a slave. By definition, the unit that establishes the piconet becomes the master.

The piconet can be considered as a small network. In addition to defining the piconet, the master also controls

the traffic on the piconet and takes care of access control. The access code preceding the packets exchanged on the channel is derived from the master's identity and must match in the slave before it accepts any packet. Channel access is completely contention-free. The short dwell time of 625 μs allows the transmission of only a single packet. The master implements a centralized control in a star configuration; communication only between the master and one or more slaves is possible. The time slots are alternately used for master transmission and slave transmission. In the master transmission, the master includes the MAC address of the slave for which the information is intended. In order to prevent collisions on the channel due to multiple slave transmissions, the master applies a polling technique; for each slave-to-master slot the master decides which slave is allowed to transmit. This decision is performed at a slot-per-slot basis: only the slave addressed in the master-to-slave slot directly preceding the slave-to-master slot is allowed to transmit in the slave-to-master slot. If the master has information to send to a specific slave, this slave is polled implicitly and can return information. If the master has no information to send, it has to poll the slave explicitly with a short poll packet. Since the master schedules the traffic both in forward and return link, intelligent scheduling algorithms have to be used that take into account the slave's traffic characteristics. The master control prevents collisions among participants of the same piconet. Independent, collocated piconets may interfere when they occasionally use the same hop carrier. A type of ALOHA is applied — information is transmitted without checking for a clear carrier (listen before talk). If the information is received incorrectly, it is retransmitted at the next transmission opportunity at a different carrier frequency (for ACL links only). Because of the short dwell time, collision avoidance schemes are less appropriate for the FH radio. For each hop, different contenders are encountered. Backoff mechanisms would therefore be less efficient.

In the piconet, the master can support SCO and ACL links. The SCO links form point-to-point links between a master and a single slave. SCO links are defined by a forward and return slot pair reserved at a fixed interval. On the remaining slots, the master can address each slave via an ACL link. The traffic over the ACL link is scheduled by the master via the polling mechanism. The slotted structure of the piconet channel allows an effective mixing of the synchronous (SCO) and asynchronous (ACL) links. However, it is a centralized control and all traffic has to flow via the master. An example of a channel with SCO and ACL links is illustrated in Fig. 12.

A multiple of piconets colocated in the same area is indicated as a *scatternet*. These piconets all operate independently, each controlled by its own master. Because Bluetooth uses packet-based communication over slotted links, it is possible to interconnect different piconets. This means that units can participate in different piconets. However, since the radio can only tune to a single-carrier frequency at one time, at any instant in time a unit can communicate in one piconet only. But a unit can jump from one piconet to another piconet by adjusting the piconet channel parameters (i.e., the master identity and the master clock), thus applying time-division multiplexing (TDM) to be virtually present in several piconets. A unit can also change role when jumping from one piconet to another piconet. For example, a unit can be master in one piconet at one instant in time, and be a slave in a different piconet at another instant in time. A unit can also be slave in different piconets. However, by definition, a unit cannot be master in different piconets, since the master parameters specify the piconet FH channel. The hop selection mechanism has been designed

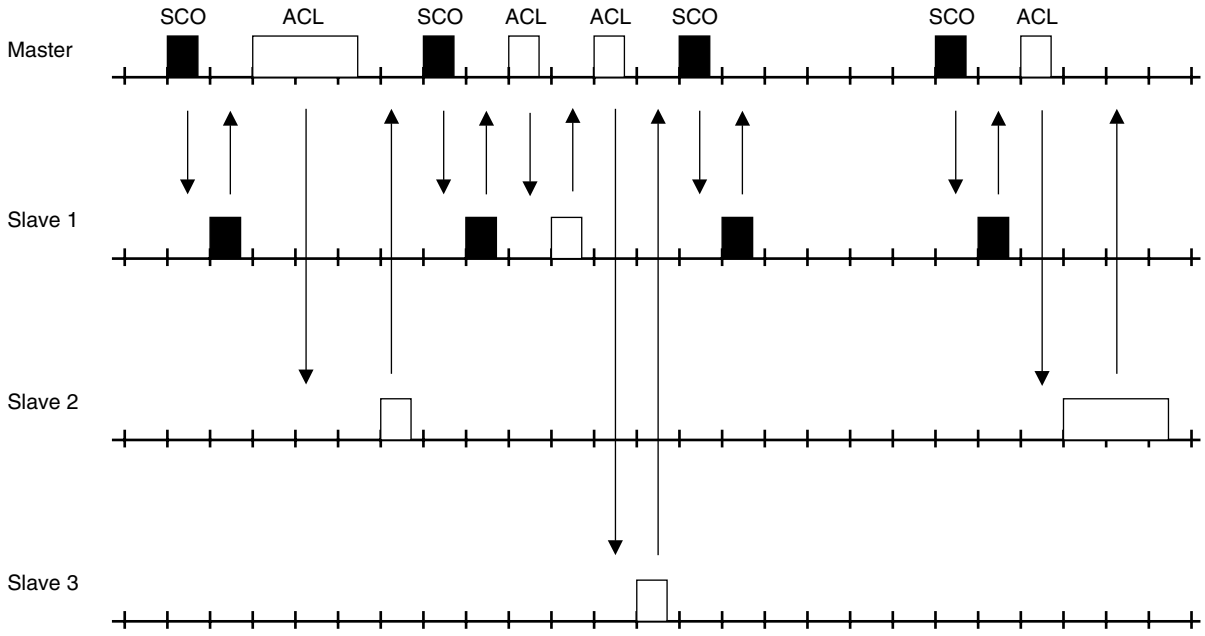


Figure 12. Master traffic scheduling and mixing SCO and ACL links.

to allow for interpiconet communications; by changing the identity and clock input to the selection mechanism, instantaneously a new frequency for the new piconet is selected. In order to make the jumps between the different piconets feasible, guard time has to be included in the traffic scheduling to account for the slot misalignment of the different piconets. In Bluetooth, the low-power modes HOLD and SNIFF can be applied to temporary leave one piconet and visit another piconet. Since all piconets are hopping independently without coordination of hopping or timing, traffic scheduling and routing in a scatternet with interpiconet communications becomes a challenge (12).

6. SECURITY

Although Bluetooth is mainly intended for short-range connectivity between personal devices, some basic security elements are included to prevent unauthorized usage and eavesdropping. At connection establishment, an authentication process is carried out to verify the identities of the units involved. The authentication process uses a conventional challenge–response routine. The claimant transmits its claimed 48-bit identity to the verifier. The verifier returns a challenge in the form of a 128-bit random number. This random number, the claimant address, and a 128-bit common secret key form the inputs to a computational secure hash function based on SAFER+ that produces a 32-bit signed response. The signed response produced by the claimant is sent to the verifier which compares this result with its own signed response. Only if the two calculated responses are the same will the challenger continue with the connection establishment. The authentication can be uni- or bidirectional.

In addition to authentication, encryption is applied. Although the pseudorandomness of the FH channel gives some protection against a casual eavesdropper, it provides no privacy in the cryptographic sense. Therefore, the payload of each packet is encrypted. Encryption is based on stream ciphering; the payload bits are modulo-2 added to a binary key stream. The binary key stream is generated by a second hash function that is based on linear feedback shift registers (LFSRs). When encryption is enabled, the master sends a random number to the slave. Before the transmission of each packet, the LFSR is initialized by a combination of this random number, the master identity, an encryption key, and the slot number. Since the slot number changes for each new packet, the initialization is new for each packet.

The central element in the security process is the 128-bit key. This key is a secret key residing in the Bluetooth hardware and is not accessible by the user. The secret key is generated during an initialization phase, also called *pairing*. Two units that want to authenticate each other and establish secure links have to be paired; that is, they have to be provided with the same secret key. An initialization phase initiated by the user is required to pair two devices. To authorize pairing, the user has to enter an identical PIN in both devices. For devices without a user interface (e.g., headsets), initialization is possible only during a short time window, such as after the user has pressed an initialization key. Once the pairing has been carried out, the 128-bit key resides in the devices and can be used for automatic authentication without user interaction [13].

Bluetooth provides a limited number security elements at the lowest level. More advanced security procedures (public keys, certificates to name only a few) can be implemented at higher layers, but are not part of the protocol.

7. CONCLUSION

The Bluetooth technology is a new radio technology for providing short-range connectivity between personal devices. FHCDMA is applied to allow many independent connections to coexist within the same area, efficiently sharing the unlicensed ISM band at 2.45 GHz. The FH channel forms a piconet between a master and one or more slaves. The hopping and the traffic scheduling in this FH channel is controlled by the master. Key elements in the Bluetooth technology are low cost, low power, and small size, enabling single-chip radio transceivers to be embedded in millions of personal devices in the near future.

BIOGRAPHY

Jaap C. Haartsen received his M.S. and Ph.D. degrees with honors in electrical engineering from the Delft University of Technology, the Netherlands, in 1986 and 1990, respectively. He joined Ericsson in 1991 and has worked in the area of wireless technology at Ericsson sites in the United States, Sweden, and the Netherlands. In Sweden, he laid the foundations for the Bluetooth radio concept. He played an active role in the creation of the Bluetooth Special Interest group and served as chair for the SIG air protocol specification group from 1998 till 2000. In April 2001, he became chief scientist of Ericsson Technology Licensing AB, an Ericsson company, fully dedicated to Bluetooth IP. In May 2000, he was appointed as adjunct professor at the Twente University of Technology, the Netherlands, in the area of mobile radio communications. He has authored numerous papers and holds over 40 patents in the area of wireless communications. His areas of interest are wireless system architectures, radio technology, ad-hoc radio communications, and short-range communications.

BIBLIOGRAPHY

1. D. J. Goodman, *Wireless Personal Communications Systems*, Addison-Wesley, Reading, MA, 1997.
2. ETSI Radio Equipment and Systems (RES), *Digital European Cordless Telecommunications (DECT), Common interface Part 1: Overview*, ETS 300 175-1, 1996.
3. R. van Nee et al., New high-rate wireless LAN standards, *IEEE Commun. Mag.* **37**: 82–88 (1999).
4. Federal Communications Commission, CFR47, *Part 15—Radio Frequency Devices*, 1999.
5. ETSI Radio Equipment and Systems (RES), *Wideband Data Transmission Systems: Technical Characteristics and Test*

Conditions for Data Transmission Equipment Operating in the 2.4 GHZ ISM Band and using Spread Spectrum Modulation Techniques, ETS 300 328, 1994.

6. ARIB Std., *Low Power Data Communication/Wireless LAN System*, RTC STD-33, 1998.
7. H. Howe Jr., A starlet's secret life as inventor, *Microwave J.* **42**: 70–74 (1999).
8. J. C. Haartsen and S. Mattisson, Bluetooth—A new low-power radio interface providing short-range connectivity, *Proc. IEEE* **88**: 1651–1661 (2000).
9. *Specification of the Bluetooth System*, Version 1.1, Bluetooth Special Interest Group, *www.bluetooth.com*.
10. B. A. Miller and C. Bisdikian, *Bluetooth Revealed*, Prentice-Hall, Upper Saddle River, NJ, 2001.
11. N. J. Muller, *Bluetooth Demystified*, McGraw-Hill, New York, 2001.
12. M. Frodigh, P. Johansson, and P. Larsson, Wireless ad-hoc networking—the art of networking without a network, *Ericsson Rev.* **77**: 248–263 (2000).
13. J. Persson and B. Smeets, Bluetooth security—an overview, *Information Security Technical Report*, Elsevier Advanced Technology, 2000, Vol. 5, pp. 32–43.

BROADBAND WIRELESS ACCESS

HIKMET SARI
Juniper Networks
Paris, France

1. INTRODUCTION

The telecommunications network has undergone major improvements to fulfill the ever-increasing need of end users for higher data rates. The multigigabit routers and optical transmission lines, which are now installed in the core and edge networks, have tremendously increased the data rates that can be transmitted. The speed bottleneck, which sets a limit on the services that can be offered to the end users, is the access network that connects them to the edge and core networks. The best-known access network is the twisted-pair copper cable, which serves virtually all homes and businesses. These cables were traditionally used to carry voice services and low-speed data communications using voiceband modems. They are now used to offer digital subscriber line (DSL) services, which are available in different forms. High-speed DSL (HDSL) uses two or three twisted pairs to offer symmetric 2-Mbps (megabits per second) data services, while the more recently developed asymmetric DSL (ADSL) technology offers a 6–8-Mbps data rate downstream and several hundreds of kbit/s upstream. Higher data rates will be offered in the near future using very high-speed DSL (VDSL) as the fiber nodes get closer to the end users and twisted-pair portion of the traditional telephone network gets shorter and shorter.

Similarly, coaxial cable networks were traditionally used for broadcast TV services only. Since the access network was opened to competition in the early 1990s, cable operators upgraded their cable plants and turned them into bidirectional networks in order to offer high-speed data services to the subscribers. Numerous cable operators today offer a variety of services using either proprietary technologies or industry standards like the data-over-cable system interface specification (DOCSIS). Upstream transmission in cable networks employs the frequency spectrum of 5–42 MHz in the DOCSIS standard and 5–65 MHz in its European version known as Euro-DOCSIS.

In countries with a well-developed telecommunications infrastructure, there has been little need in the past for fixed wireless access. This type of systems were essentially deployed in developing countries with a large population that is not served by the twisted-pair telephone cable. Those wireless access systems, however, are narrowband, and can only carry telephony and low bit-rate data services. The emergence of broadband wireless access (BWA) is very closely related to the more recent deregulation of the world telecommunications market. This deregulation has created a new environment in which new operators can compete with incumbent operators that often are former state-owned monopolies. Wireless access networks are very appealing to new operators without an existing wired infrastructure, because they not only are rapidly deployed but also involve a low initial investment, which is determined by the initial customer base. Once in place, wireless networks are easily upgraded to accommodate additional subscribers as the customer base grows. This is a very attractive feature with respect to wired networks where most of the investment needs to be made during the initial deployment phase.

Most frequencies available for BWA are at millimeter-wave frequencies between 20 and 45 GHz. Dedicated frequency bands for these applications have recently become available in Europe, North America, Asia Pacific, and other regions. After an extensive field trial period, BWA systems operating at millimeter-wave frequencies are currently in the field, but their commercial deployment remains small scale. These cellular radio networks, which are commonly referred to as *local multipoint distribution service* (LMDS) networks, are intended to offer integrated broadband services to residential and business customers. LMDS networks are particularly suited for urban or suburban areas with high user density, because the cell capacity is typically in the range of the STM-1 data rate (155 Mbps) and the cell coverage is only 2–5 km. This characteristic makes them attractive for business customers, whereas DSL and cable access systems are essentially for residential subscribers.

Although not as much as in the millimeter-wave frequency range, there are also some frequency bands below 11 GHz for BWA applications. These include the 2.5-GHz microwave multipoint distribution service (MMDS) band in the United States, the 3.5-GHz band in Europe, and the 10-GHz band in a limited number of countries. Below 11 GHz, there are also some license-exempt frequency bands that may be used for wireless access. These frequency bands are not specific to BWA and will not be specifically covered here.

The article gives a state-of-the-art review of broadband wireless access systems and discusses the current trends

and ongoing standardization work. First, in the next section, we give a brief introduction to millimeter-wave BWA (LMDS) networks. In Section 3, we present an analysis of the intercell interference that limits the frequency reuse in these networks. Next, in Section 4, we outline the current standardization work by the IEEE 802.16 and the ETSI BRAN groups for next generation BWA systems at millimeter-wave bands. Finally, in Section 5, we discuss BWA at microwave frequencies below 11 GHz, which is essentially focused on residential applications. Some conclusions are given in Section 6.

2. AN OVERVIEW OF LMDS NETWORKS

LMDS was originally used to designate the 28-GHz band in the United States, but it is often used today to designate BWA systems operating at all millimeter-wave frequencies above 20 GHz. LMDS networks are cellular, each cell serving a number of fixed subscribers located in its coverage area, which has a radius of 2–5 km. The base station (BS) is connected to the backbone network through a backhaul point-to-point link, which can be a fiber or a radio link.

The network topology resembles that of mobile cellular radio systems, but LMDS systems have several distinctive features. The main of those is that since users are at fixed locations, each user is assigned to a predetermined BS (typically the nearest BS). Furthermore, fixed wireless access systems employ narrowbeam directional subscriber antennas pointed to the serving BS during installation. The increased gain in the direction of the BS reduces network interference and increases cell coverage.

Another major difference concerns propagation. While mobile radio systems are subjected to shadowing and severe multipath propagation, LMDS systems are based on clear line-of-sight (LoS) between the BS and the fixed users, and are virtually free of multipath propagation. Signal attenuation during normal propagation conditions is proportional to the square of the distance, and what truly limits the cell coverage is "rain fading," which further attenuates the transmitted signal by several dB or several tens of dB per kilometer. Because of this phenomenon and the limited power that can be generated at low cost at millimeter-wave frequencies, the cell radius in LMDS networks is in the range of 2–5 km depending on the climatic zone, the available transmit power, and the required availability objectives.

Although LMDS systems can be based on hexagonal cell patterns that are commonly used in mobile radio systems [1], rectangular cell patterns with 90° cell sectoring have become very popular in LMDS network design [2,3] and will be considered throughout this article. Each sector in this cell pattern is served by a 90° sector antenna, and different channels are assigned to the different sectors. The channel bandwidth differs from region to region; In Europe and other countries that follow the CEPT channeling, the channel spacing is of the form $112/2^n$ MHz, where n is an integer. The typical channel spacing for BWA in this region is 28 MHz, unless the operator does not have sufficient bandwidth allocation, in which case 14 or 7 MHz channels are used. In North America, the channel bandwidth is of the form $80/2^n$ MHz, with a typical bandwidth of 20 MHz.

With a simple quaternary phase shift keying (QPSK) modulation, a 28-MHz CEPT channel is sufficient to transmit a useful data rate of 16×2 Mbps. The total bit rate per cell is then 64×2 Mbps and can be used to serve for example 64 business customers with a 2-Mbps leased line each. This example is only to give an idea of the cell capacity. The number of subscribers per cell may be several hundreds or several thousands, and such a large number of users can still be accommodated by dynamically sharing the available resources between them. Since some users may be idle while some users are requesting high instantaneous data rates, there is a substantial statistical multiplexing gain, which makes it possible to accommodate a large number of subscribers.

3. NETWORK INTERFERENCE ISSUES

Because of the lack of industry standards, first-generation BWA systems are based on proprietary solutions. In fact, technical specifications for LMDS systems were developed by the Digital Video Broadcasting (DVB) Project [4] and the Digital AudioVisual Council (DAVIC) [5] in the 1994–1996 time period, but these specifications were primarily intended for digital TV broadcasting and two-way communications with low interactivity. Most proprietary systems today use pieces of the DVB/DAVIC standards, but they do not follow these standards completely. Also, most of them, as well as the DVB and DAVIC specifications, are based on time-division multiplexing (TDM) on the downstream channel (from BS to subscribers), time-division multiple access (TDMA) on the upstream channel (from subscribers to BS), and frequency-division duplexing (FDD). This means that separate channels are used for downstream and upstream transmissions.

Since bandwidth is a limited and costly resource, the frequency reuse factor and the achievable cell capacity are crucial to the deployment of LMDS networks. These factors are strongly impacted by intercell interference. In this article, we discuss intercell interference, assuming that the network is based on rectangular cell pattern with 90° sectoring and that a separate channel is assigned to each sector, which means that four channels are used to cover each cell. But the same channels are reused in all cells as shown in Fig. 1. Note that channel assignment between neighboring cells follows a mirror-image rule in the horizontal, vertical, and diagonal directions.

The BSs, which are designated by heavy dots in Fig. 1, are located on a rectangular grid. The solid lines represent the sector borders, and the dotted lines indicate the cell boundaries. The labels A, B, C, and D represent the channels used in different sectors. As it is indicated in Refs. 2 and 3, the mirror-image assignment of these channels eliminates interference between adjacent cells.

However, the second-nearest cells have the same channel assignment as the cell at hand and interfere with it. Let 2Δ designate the distance between two adjacent BSs in the horizontal and vertical directions. Now suppose that a user is located on the border of two sectors at a distance

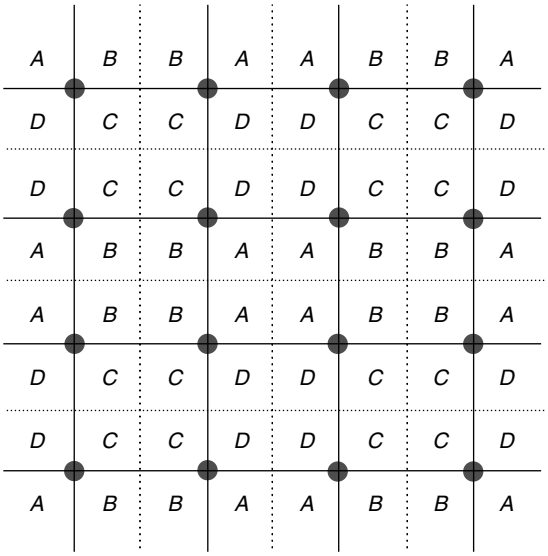


Figure 1. Rectangular cell pattern with 90° sectors.

δ from the serving BS. This user's antenna will also be pointed toward a second-nearest BS that is at a distance $4\Delta + \delta$. Assuming that all BS's transmit the same signal power and that the signal attenuation is proportional to the squared distance, which is a common assumption in line-of-sight microwave and millimeter-wave radio systems, the downstream signal-to-interference ratio (SIR) for this user is

$$\mathrm{SIR(dB)} = 20 \cdot \log\left(\frac{4\Delta + \delta}{\delta}\right) \tag{1}$$

This expression, which is valid for $0 < \delta \leq \Delta$, achieves its minimum value for $\delta = \Delta$. The corresponding SIR is 14 dB. In writing (1), we have assumed that BSs further than the second-nearest BS are not in clear LoS with the user of interest; specifically, their signals are blocked by buildings, trees, or other obstacles.

As it is shown in Ref. 2, the worst-case SIR of 14 dB is also valid for the upstream channel when automatic transmit power control is used. But the similarity of downstream and upstream channels in terms of interference is limited to the value of the worst-case SIR. On the downstream channel, the SIR is a function of the user position, and only in a very small part of the cell, the users are subjected to strong interference. Using a common subscriber antenna radiation diagram with a beamwidth of 5°, we have plotted in Fig. 2 the SIR distribution within a sector, where the BS is located in its upper left corner. Specifically, the figure shows the boundaries of the regions corresponding to an SIR higher than a given value. Notice that only in very small regions located at the other 3 corners, the SIR is lower than 15 dB. Furthermore, the SIR is higher than 30 dB in virtually half of the cell.

Figure 2 indicates that if the system design requires an SIR value higher than 15 dB, three small regions will not be covered. Coverage will be even smaller if the system design requires an SIR higher than 20 or 25 dB. This means that a bandwidth-efficient modulation scheme that

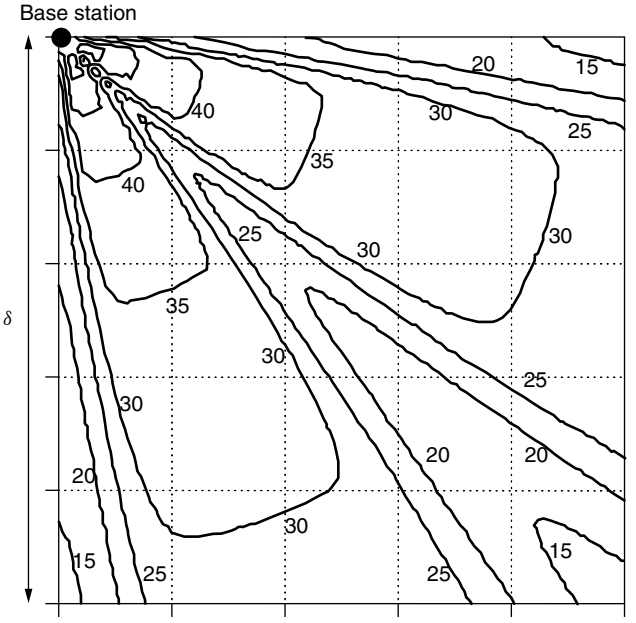


Figure 2. SIR distribution within a sector corresponding to a subscriber antenna beamwidth of 5°.

requires a high SIR value will not be usable if full cell coverage is required. But the figure also suggests that while users at unfavorable positions (regions of low SIR values) must use a low-level modulation scheme such as QPSK, users in more favorable locations can use higher-level quadrature amplitude modulation (QAM) schemes such as 16-QAM or 64-QAM, at least during normal propagation conditions. This adaptive modulation and coding concept is now used in international standards and will be discussed in Section 4.

The situation is quite different on the upstream channel, because in this direction, all users get the same amount of interference; that is, the SIR is not a function of the user position. Consequently, the user-dependent modulation concept makes little sense for the upstream channel, but the adaptive modulation concept can still be used to adapt the modulation to propagation conditions.

4. CURRENT STANDARDIZATION

While first-generation LMDS systems are today in the field, standardization activities are now at a very advanced stage at both the IEEE and the ETSI to define technical specifications for future BWA systems. The groups that are carrying out this work for millimeter-wave frequency bands are the IEEE 802.16.1 task group and the HIPERACCESS group of ETSI BRAN. Specification work by both groups covers the physical (PHY) layer and the medium access control (MAC) layer functions. At the time of this writing, the IEEE 802.16 group has already issued its draft technical specifications [6] for BWA systems at frequencies between 11 and 60 GHz. As for the HIPERACCESS group of the ETSI, which started its specification work later than the IEEE 802.16.1 task group, it intends to complete its specifications by mid-2002.

There is a significant level of commonality between the IEEE 802.16.1 and the ETSI BRAN HIPERACCESS (draft) specifications concerning basic choices for PHY layer functions. This includes the following [7]:

- The transmission technique is based on single-carrier transmission. The reason for this is that BWA systems at millimeter-wave frequencies suffer very little multipath propagation because of the small cell size and directive subscriber antennas used. This does not give much motivation for using orthogonal frequency-division multiplexing (OFDM) which is appealing for strong intersymbol interference (ISI) channels [8]. In addition, the strong sensitivity of OFDM to oscillator phase noise and transmit power amplifier nonlinearity makes this technique rather undesirable for systems operating at millimeter-wave frequencies, where high transmit power and low phase noise incur substantial cost for the outdoor radio unit.
- As in the earlier DVB/DAVIC specifications, TDM and TDMA have been adopted for the downstream channel and the upstream channel, respectively. This choice can be justified by the relative maturity of TDMA with respect to code-division multiple access (CDMA) that has been adopted in third-generation digital mobile radio standards [9].
- To increase cell capacity with respect to pure QPSK, the IEEE and ETSI specifications include adaptive modulation and coding. The idea is to use the most bandwidth-efficient modulation and coding schemes that are compatible with the signal-to-noise ratio (SNR) and the interference level affecting user signals. This is a function of the user position on the one hand (on the downstream channel), and the instantaneous fade level on the other hand. The candidate modulation schemes are 4-QAM (QPSK), 16-QAM, and 64-QAM for the downstream channel, and 4-QAM and 16-QAM for the upstream channel. To have an adaptation with a finer granularity in terms of signal-to-interference plus noise ratio (SINR), both specifications also allow adaptively changing the coding scheme.

Adaptive modulation and coding substantially increase the cell capacity for a given level of quality of service. Assuming that the SIR required is 12 dB for QPSK, 19 dB for 16-QAM, and 25 dB for 64-QAM, and using a subscriber antenna beamwidth of 6°, it was shown [10] that an adaptive modulation that combines these three signal formats on the downstream channel achieves an increase of cell capacity by a factor of 2.7 with respect to QPSK. Since all users are subjected to the same level of interference on the upstream channel, it was proposed [10] that the channel be split in two parts and each subchannel be assigned to a specific region of the sector of interest. This assignment can be done in such a way that the level of interference is significantly reduced for some subscribers. Using this subchanneling concept along with an adaptive modulation involving the QPSK and the 16-QAM signal formats, a capacity improvement by a factor of 1.4 was achieved on the upstream channel. These results indicate that adaptive modulation substantially increases the cell capacity, although to a lesser extent on the upstream channel.

One way to increase capacity on the upstream channel is to use an adaptive antenna at the BS. Indeed, if the BS employs a steered narrowbeam antenna, only the users near the sector borders in the horizontal and vertical directions and those near the diagonal will be subjected to strong upstream interference, and the situation becomes similar to that on the downstream channel. Users located outside these regions will be subjected to a smaller level of interference and can use a 16- or 64-QAM modulation. The upstream cell capacity then becomes similar to that of the downstream channel. One difficulty in applying this concept is that adaptive antenna technology is not yet mature for microwave and millimeter-wave frequencies. Adaptive antennas appear as an option in current standards, for future evolutions.

5. BWA AT LOWER FREQUENCY BANDS

Both the IEEE 802.16 Group and ETSI BRAN first put a priority on the definition of system specifications for BWA systems operating at millimeter-wave frequencies, but they later turned their attention toward licensed frequency bands between 2 and 11 GHz. The respective task groups of the IEEE and the ETSI that are in charge of defining technical specifications for BWA at frequencies below 11 GHz are the IEEE 802.16.3 task group and the HIPERMAN group of ETSI BRAN, respectively. The IEEE 802.16.3 group is already at an advanced technical specifications phase, and the ETSI HIPERMAN group has recently completed the functional requirements phase and entered the technical specifications phase.

In many aspects, BWA at lower frequencies is quite similar to LMDS, but it also has two basic distinctive features. The first concerns the traffic model. Whereas LMDS systems are essentially intended for small-business applications, frequencies below 11 GHz are primarily for residential subscribers where the major application is high-speed Internet access. The implication of this is that traffic at lower frequency bands is highly asymmetric, and most of the traffic is on the downlink from the BS to subscribers. This feature has a strong impact on both the PHY layer and the data-link (DLC) layer. The second distinctive feature is that due to larger cell sizes, smaller subscriber antenna directivity, and non-LoS propagation, lower-frequency bands are subjected to multipath propagation (and a significant level of ISI), which must be compensated.

One solution for BWA at lower microwave frequencies is to use a single-carrier transmission technology as for millimeter-waves. The only additional requirement in this case is to use an adaptive equalizer that is capable of handling the multipath propagation encountered in this kind of network. Another solution consists of using the OFDM technology, which has been adopted in the IEEE 802.11a and ETSI HIPERLAN/2 specifications for wireless local area networks (wireless LANs) at 5 GHz [11,12].

At the time of this writing, The ETSI BRAN group is still examining proposals and has not made a final decision regarding the technology to use for fixed BWA systems at lower microwave frequencies, but the IEEE 802.16.3 task group of the IEEE has already made major decisions and released a baseline document for the physical (PHY) layer. Failing to agree on a single standard, this group decided to include both an OFDM-based PHY and a single-carrier PHY layer specifications. Furthermore, the OFDM-based PHY comprises an OFDM/TDMA mode and an orthogonal frequency-division multiple access (OFDMA) [13] mode, which means that the forthcoming IEEE 802.16a actually include three different transmission and multiple access technologies. In the following subsections, we will briefly describe them.

5.1. Single-Carrier Transmission

To operate on channels that suffer from strong multipath propagation, single-carrier systems must use an adaptive equalizer. When OFDM was first proposed for the Digital Audio Broadcasting (DAB) and Digital Video Broadcasting (DVB) in Europe in the late 1980s and the early 1990s, it was assumed that single-carrier transmission does not give adequate performance on difficult radio channels, particularly for mobile reception. In the 1993–1995 time period, a series of articles were published by the present author [e.g., 8] suggesting that the common perception that single-carrier transmission does not give adequate performance on difficult radio channels is a result of constraining them to use a time-domain equalizer. After making the observation that single-carrier systems with a time-domain equalizer have an inherent limitation due to convergence and tracking problems when the number of taps is large, it was next suggested that a single-carrier system with frequency-domain equalization (SC/FDE) closely resembles an OFDM system while avoiding its well-known problems:

1. Its high peak-to-average power ratio (PAPR), which makes it very sensitive to the transmit high-power amplifier (HPA) nonlinearity
2. Its high sensitivity to the local oscillator phase noise

A schematic block diagram of the basic transmit and receive functions in OFDM and SC-FDE is given in Fig. 3. As can be seen in this figure, there is a strong resemblance between an OFDM system and an SC/FDE system. The frequency-domain equalizer in the latter system gives it the possibility to compensate for long channel impulse responses without facing the convergence problems that are inherent to single-carrier systems with time-domain equalization (SC/TDE). Indeed, under the minimum mean-square error (MMSE) criterion, the optimum coefficients of a linear transversal time-domain equalizer are the solution of the matrix equation

$$C = A^{-1}V \tag{2}$$

where A is the autocorrelation matrix of the input signal vector X_k, and V is the cross-correlation of the input signal vector X with the transmitted symbol a_k [14]. The conventional least mean squares algorithm for coefficient adaptation at time k is

$$C_{k+1} = C_k - \alpha X_k^* e_k \tag{3}$$

where α is the step-size parameter that controls convergence, and e_k is the equalizer output error at time k [14]. Without any mathematics, it can easily be seen that the equalizer coefficients do not converge independently of each other, because their adaptation is driven by the same error signal e_k and also the components of the vector X_k are not independent.

Now consider a frequency-domain equalizer with N taps. The DFT operator that forms the first stage of the equalizer gives N signal samples denoted $(Y_1, Y_2, \ldots, Y_N)$. These samples are the inputs to a complex multiplier bank whose coefficients are denoted $(D_1, D_2, \ldots, D_N)$. The coefficient values which minimize signal distortion are given by

$$D_n = \frac{H_n^*}{|H_n|^2} \tag{4}$$

where H_n denotes the channel transfer function at frequency f_n. A better criterion is the MMSE criterion, which minimizes the combined effect of channel distortion and additive noise. The optimum coefficients in the MMSE sense are

$$D_n = \frac{H_n^*}{|H_n|^2 + \gamma} \tag{5}$$

where γ is the inverse of the signal-to-noise ratio (SNR). Clearly, the optimum coefficients of a frequency-domain equalizer are independent of each other, and therefore convergence occurs at speeds much higher than is possible in time-domain equalizers. The consequence of this is that a frequency-domain equalizer can employ a large number of taps and compensate for long impulse response channels while converging fast and tracking rapid channel variations.

An important feature of the SC/FDE system concept proposed [8] is that it employs a cyclic prefix (similar to OFDM) so as to make the linear convolution of the channel look like the circular convolution performed by the frequency-domain equalizer. The articles published by this author in which SC/FDE was shown to be an attractive alternative to OFDM stimulated further research on the subject and led to the rebirth of frequency-domain equalization which had long been ignored in the literature.

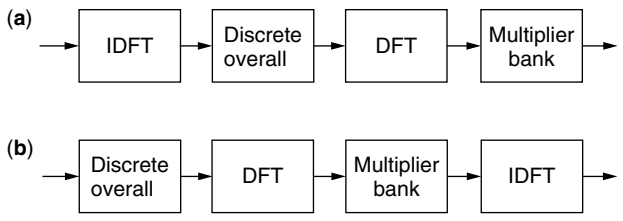


Figure 3. Transmit and receive block diagram in OFDM (**a**) and SC-FDE (**b**).

5.2. OFDM

The basic idea in OFDM is to split the channel bandwidth into a large number of subchannels such that the channel frequency response is essentially flat over the individual subchannels. This is performed using an inverse discrete Fourier transform (DFT) at the transmitter and a forward DFT at the receiver. More specifically, the transmitter of an OFDM system with N carriers includes a serial-to-parallel (S/P) converter, an inverse DFT operator of size N, and a parallel-to-serial (P/S) converter which serializes the DFT output before sending it to subsequent filtering and frequency upconversion stages.

The way OFDM compensates for frequency-selective fading is substantially different from the way single-carrier transmission handles this phenomenon. Since the N symbols of each DFT block are transmitted at different frequencies and the individual subchannels are very narrow, the symbols transmitted at faded frequencies (located on a deep notch of the channel frequency response) cannot be detected reliably. OFDM systems must therefore resort to channel coding in order to protect the symbols transmitted at faded frequencies, whereas single-carrier systems can operate on frequency-selective channels without channel coding. Operation of OFDM systems is best explained using a simple example. Suppose that the channel impulse response has a strong attenuation at K frequencies. Then, the K symbols per DFT block transmitted at these frequencies will be in error with a large probability. A block code whose length is equal to the DFT block length and error correction capability exceeds K symbols will correct these errors, and the resulting OFDM system will be efficient on that channel. OFDM systems can also use convolutional coding. In that case, the code must have a large Hamming distance (the minimum length of error events) and an interleaver must be included in order to distribute the effect of fading on transmitted symbols.

A cyclic prefix is inserted between OFDM symbols at the transmitter so that the linear convolution of the channel becomes a circular convolution for the transmitted symbols. This requires that the cyclic prefix be larger than the channel impulse response length. The two important parameters of an OFDM system are the number of carriers and the length of the cyclic prefix. The prefix represents overhead, and its length is dictated by the maximum length of the channel impulse response to be compensated. In order to limit the loss in throughput due to the cyclic prefix, the DFT block length (the number of carriers) must be increased. But increasing the number of carriers increases complexity on the one hand and the sensitivity to timing variations of the channel on the other hand. In the IEEE 802.16a specifications, the number of carriers is 256, and the prefix can have up to 64 samples (a quarter of an OFDM symbol length).

5.3. OFDMA

OFDM transmission on multiple access channels is often used with TDMA, and the resulting combination is referred to as OFDM/TDMA. (This is the case in the IEEE 802.11a and HIPERLAN/2 standards.) In this scheme, the base station assigns time slots to different users, and the signal transmitted within a time slot is an OFDM signal. For convenience, a time slot is an integer multiple of an OFDM symbol.

The third transmission mode included in the IEEE 802.16a specifications is OFDMA [13]. In this technique, the N symbols per DFT block are not all assigned to the same user, but instead they are partitioned into M subsets of N/M symbols, and resource assignment is performed subset by subset. This means that resources can be allocated to M users during the same OFDM symbol period.

OFDMA has several interesting features with respect to OFDM/TDMA. First, it reduces the granularity of the bursts allocated to different users thereby increasing the efficiency of the MAC protocol. Next, it increases the cell range in the upstream direction by concentrating the power available from the transmit amplifier on a subset of carriers. (Every division by a factor of 2 of the number of carriers used per subscriber is equivalent to increasing the transmit amplifier power by 3 dB.) This means that an OFDMA system based on splitting the total number of carriers N by 16 and allocating a single subset to users will increase the cell coverage by as much as 12 dB. The cell range can also be increased in the downstream direction by allocating a transmit power to each set of carriers that is function of the distance to the user to which this set is allocated.

6. CONCLUSIONS

Its ease of deployment and the low initial investment involved makes BWA the most attractive broadband access technology for new operators without an existing infrastructure. Millimeter-wave BWA (LMDS) is mostly suited for business subscribers in high-density urban or suburban areas, and BWA at lower microwave frequencies is essentially suited for residential subscribers. After briefly discussing the cell capacity, frequency planning, and interference issues in current LMDS networks based on proprietary technologies, this article summarizes the current status of standardization work by the ETSI BRAN and the IEEE 802.16 Groups for both millimeterwave frequencies and lower microwave frequencies between 2 and 11 GHz. Because of the LoS propagation that characterizes this type of networks, standards for BWA at millimeterwave frequencies are based on single-carrier transmission. But BWA below 11 GHz is subjected to strong multipath propagation, and the IEEE 802.16a standard for this type of networks has three different modes: SC/FDE, OFDM, and OFDMA.

BIOGRAPHY

Hikmet Sari received his Diploma (M.S.) and Ph.D. in telecommunications engineering from the ENST, Paris, France, and the *Habilitation* degree from the University of Paris XI, France. From 1980 to 2000 he held research and management positions at the Philips Research Laboratories, SAT, and Alcatel Paris, France. In May 2000, he joined Pacific Broadband Communications (PBC)

as chief scientist. He is now with Juniper Networks, which acquired PBC in December 2001. He has published over 130 technical papers and holds over 25 patents. In 1995, he was elevated to the IEEE fellow grade and received the Andre Blondel Medal from the SEE (France). He was an editor of the IEEE Transactions on Communications from 1987 to 1991, a guest editor of the European Transactions on Telecommunications (ETT) in 1993, and a guest editor of the *IEEE JSAC* in 1999. Presently, he is an associate editor of the *IEEE Communications Letters* and a distinguished lecturer of the IEEE Communications Society.

BIBLIOGRAPHY

1. T. S. Rappaport, *Wireless Communications: Principles and Practice*, IEEE Press, New York, and Prentice-Hall, Englewood Cliffs, NJ, 1996.
2. H. Sari, Broadband radio access to homes and businesses: MMDS and LMDS, *Comput. Networks* **31**: 379–393 (Feb. 1999).
3. G. LaBelle, LMDS: A broadband wireless interactive access system at 28 GHz, in M. Luise and S. Pupolin, eds., *Broadband Wireless Communication*, Springer-Verlag, Berlin, 1998, pp. 364–377.
4. ETS 300 748, *Digital Video Broadcasting* (*DVB*)*: Framing Structure, Channel Coding, and Modulation for MVDS at 10 GHz and above*, ETSI, October 1996.
5. DAVIC 1.1 Specifications, Part 8, *Lower-Layer Protocols and Physical Interfaces*, Revision 3.3, Geneva, September 1996.
6. *Air Interface for Fixed Broadband Wireless Access Systems*, IEEE 802.16.3 task group, Sept. 2000.
7. ETSI Website: *www.etsi.org*
8. H. Sari, G. Karam, and I. Jeanclaude, Transmission techniques for digital terrestrial TV broad-casting, *IEEE Commun. Mag.* **33**: 100–109 (Feb. 1995).
9. F. Adachi, M. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communications systems, *IEEE Commun. Mag.* **36**(9): 56–69 (Sept. 1998).
10. J. P. Balech and H. Sari, Advanced modulation techniques for broadband wireless access systems, *Proc. 7th Eur. Conf. Fixed Radio Systems and Networks* (ECRR 2000), Dresden, Germany, Sept. 2000, pp. 159–164.
11. P802.11a/D6.0, *LAN/MAN Specific Requirements*, Part 2: *Wireless MAC and PHY Specifications — High Speed Physical Layer in the 5 GHz Band*, IEEE 802.11, May 1999.
12. DTS/BRAN030003-1, *Broadband Radio Access Networks HIPERLAN Type 2 Functional Specification*, Part 1: *Physical Layer*, ETSI, Sophia Antipolis, Sept. 1999.
13. H. Sari and G. Karam, Orthogonal frequency-division multiple access and its application to CATV networks, *Eur. Trans. Telecommun. Related Technol.* (*ETT*) **9**(6): 507–516 (Nov.–Dec. 1998).
14. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

C

CABLE MODEMS

DONALD G. MCMULLIN
Broadcom Corporation
Irvine, California

1. INTRODUCTION

Since the inception of the Internet as a high-speed data connection between universities in the early 1970s, the search for a low-cost broadband last-mile delivery system has been pursued. The fiberoptic backbone is capable of sustaining terabits of data throughput, but the last mile connection has traditionally been limited to, at best, about 28 kb (kilobits) and more recently 56 kbits. Realizing that the needed bandwidth for these high-speed data links could be supplied by the television cable plant, in the mid 1970s the FCC mandated that all new cable television trunk lines and drop lines be installed as two-way-ready. Two-way amplifiers were installed allowing both downstream and upstream data traffic to occupy selected frequency spectra on a single coaxial cable. New head-end cable equipment was installed, and the cable operators began to deploy broadband Internet access over the cable infrastructure (Fig. 1a,b). In the event that a cable plant had not been upgraded for two-way operation, the Telco modem (or cable downstream and telephone upstream) has been successfully deployed. Typical bandwidth usage models require a broadband downstream channel, since users nominally request large amounts of data from the Internet server. The return path (or upstream) bandwidth can be reduced, since users rarely transmit large amounts of data upstream. In fact, the limited upstream traffic has allowed for further bandwidth efficiency by utilization of a time-division multiple access (TDMA) scheme for two-way cable modem implementation. This method allows multiple users to transmit data on the same IF carrier frequency, but at different times. This is known as *burst-mode transmission*, and is contrasted to the subscriber modem receiver downstream data, which are supplied as a continuous bitstream. Exceptions to this limited upstream bandwidth are applications requiring two-way videoconferencing, and these have just recently (at the time of writing) been addressed in new specifications.

In 1995 efforts were made by the newly formed Multimedia Cable Network Systems (MCNS) organization and the IEEE 802.14 committee to define and establish standards for transmission of IP data over existing cable lines. Both of these bodies eventually dissolved into what is known today as the Data over Cable Service Interface Specification (DOCSIS) standard. The lower four layers of the data protocol are primarily what DOCSIS 1.0/1.1 defines [1] and are outlined as follows:

Layer 1 — PHY (physical layer): defines upstream and downstream modulation schemes, 64/256-QAM downstream and QPSK/16-QAM upstream

Layer 2 — MPEG2: defines the data packet organization and FEC (forward error correction) codes

Layer 3 — MAC (media access control): defines the data processing protocols between cable modem (CM) at the customer premise, and the head-end (HE) equipment, also known as the cable modem termination system (CMTS) residing at the central office

Layer 4 — BPI (Baseline PrIvacy): sets the key codes for encryption to provide security on the shared cable network

The structure of the downstream payload data has a unique packet ID (PID), service ID (SID), and destination address (DA) embedded in the data packets. The PID is used to identify a "data type" packet as opposed to digital video information. The downstream SID identifies the security association of each packet and the DA identifies packets that belong to a particular user. Packets are framed in standard MPEG-2 format. This allows the data channels to occupy the already defined digital video channel spacing and decoder technology. MPEG-2 defines what is specified as "well known" packet identifiers, and for cable modem data traffic this hex value is 0x1FFE.

Thus, as the packet parser contained in the cable modem MAC looks at each PID inserted in each MEG packet received, it will proceed to the next level of decoding of the SID only if it finds a PID indicating that this is a data channel. If there are no payload data (actual data to receive), then a "null packet" will be transmitted consisting of the hexadecimal (hex) value 0xFF for all payload data bytes, enabling the downstream to remain locked to the QAM channel and decoding MPEG packets at all times. A brief description of the MPEG-2 packet structure will be presented later in this article. A block diagram of the cable modem is shown in Fig. 2 and will be discussed in detail later in this article.

For both upstream and downstream data to coexist on a single cable, a means to separate the frequency spectra is necessary. For the North American standard, the downstream data services reside with the already established downstream video channels occupying the 54–860-MHz band (using 6-MHz channel spacing). The upstream data are placed in the unused frequency bands from 5 to 42 MHz (Fig. 3). A diplex filter is used to mitigate crosstalk between the respective frequency allocations. The diplex filter consists of a HI-PASS section for the downstream channels and a LO-PASS section for the upstream channels. As mentioned earlier, upstream return channels are burst mode and symbol rates are assigned during the logon process on the basis of requested/available bandwidth. Thus, the head end can allocate bandwidth in accordance with the demands as more users logon and require more channel capacity. This results in a slow degradation in system performance, in contrast to a telephone modem, whereby when no more switch ports are available, the user cannot establish a connection at all. Additionally, when

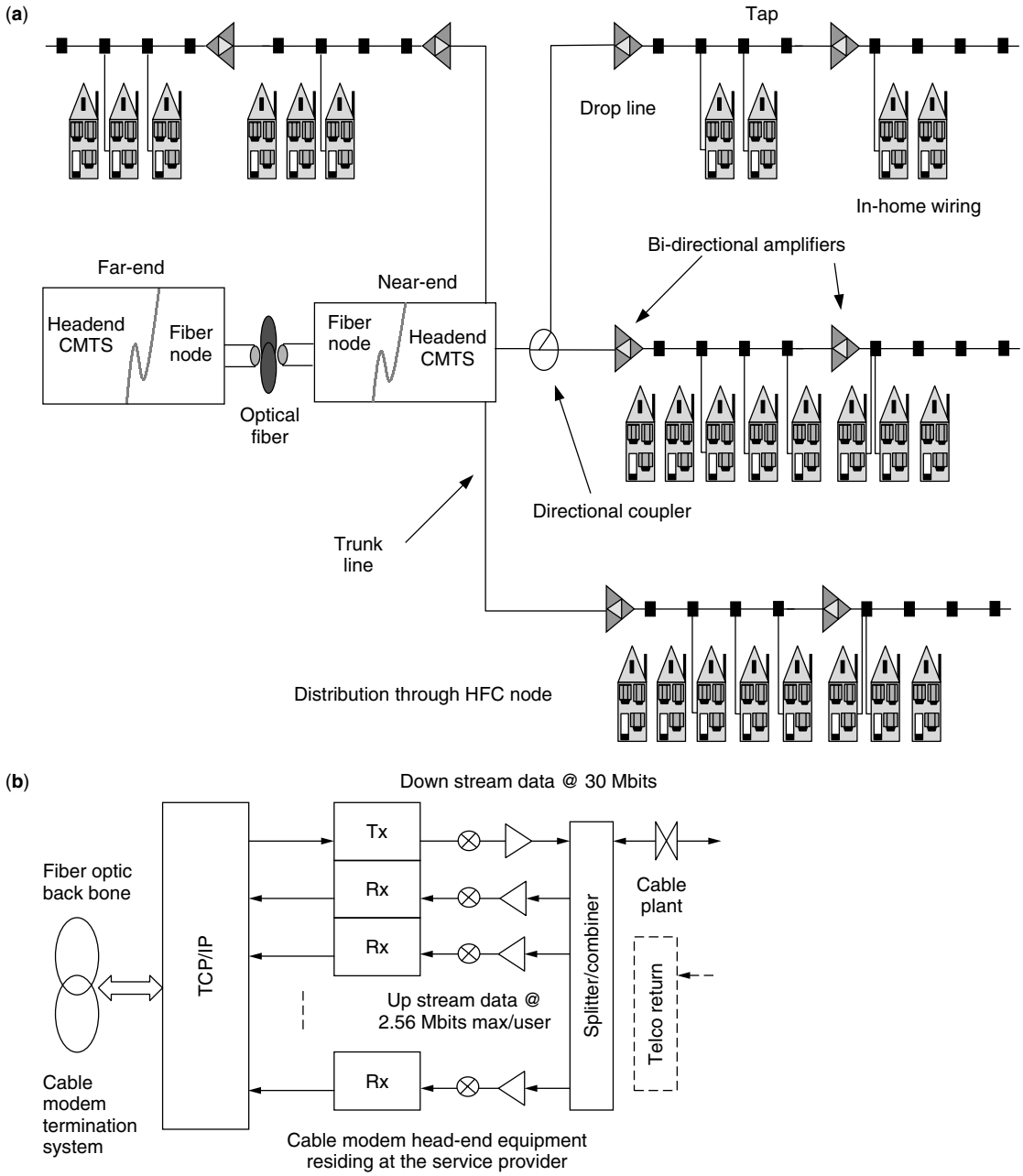


Figure 1. Block diagrams of (**a**) HFC plant and (**b**) CMTS.

the cable system bandwidth reaches an unacceptable level of performance, a new RF downstream or IF upstream frequency is assigned to some of the users, and the system data throughput can be restored dynamically without disruption of service or any knowledge by the users. Typical loading currently is about 200 or so users per downstream channel and optimum channel loading has been established by historical usage models for telephone lines.

2. PHYSICAL-LAYER: MODULATION FORMAT

Why use quadrature amplitude modulation (QAM) for transmission and reception? The search for a compact efficient means to transmit and receive data has led to the implementation of QAM for cable applications. The simplest form of QAM is called quadrature phase shift keying (QPSK) and, as the term implies, this form of modulation takes advantage of the orthogonal nature of an I (in phase) and Q (quadrature — 90° phase shift) coordinate system. As has been shown by Euler, Parseval, and others, an orthogonal system allows the encoding of two distinct and independent sets of information that can be combined, transmitted, and demodulated without interaction or distortion to each other. This allows a second degree of freedom, raising to the power of two the capacity of any transmission system. In its simplest form, QPSK has been used for satellite communications for many years, dating back to the early 1950s. Only until relatively

recent advances in semiconductor technology and system integration have higher-order QAM modulation formats become commonplace. Still, in a high-noise environment a constant vector magnitude modulation scheme such as QPSK or 8-PSK is far preferred. For this type of constellation, all symbols (or data points) lie on a circle; thus the magnitude is constant and it is only necessary to detect the phase difference between each transmitted symbol to complete the demodulation process. This facilitates robust reception of data even in environments with high levels of noise.

The transmission medium drives the choice of modulation scheme and for a "wired" or cable system, the design constraints are very different from those in a wireless system. The fundamental specification driving the choice of modulation is the obtainable system SNR, and secondary to this are the expected multipath reflections, which can be stationary (cable) or time-varying (wireless). The well-known Shannon–Hartley capacity theorem [2] states

$$C = W \log_2 \left(1 + \frac{S}{N}\right)$$

where C = system capacity
S = signal power
W = system bandwidth
N = noise power

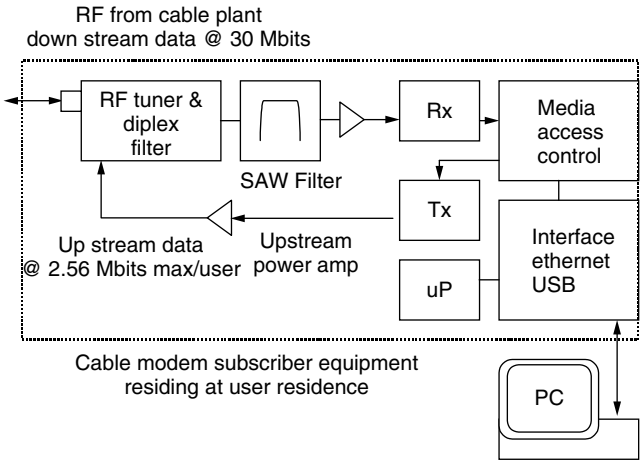


Figure 2. Block diagram of CM.

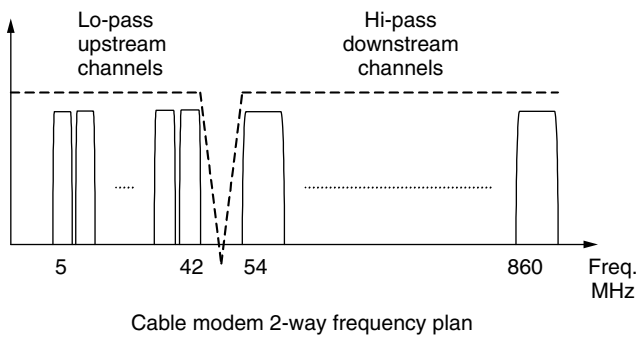


Figure 3. Cable modem frequency plan.

Rearrangement and normalizing yields the channel capacity C/W in bits per second per hertz, which defines the maximum number of bits per symbol that can be transmitted for a given SNR (Fig. 4). Although the Shannon–Hartley equation does not explicitly set a limit for the error probability, achievable SNR has a large effect in determining the QAM receiver bit error rate (BER). This, in turn, will dictate the reliability and absolute data rate of the communications link. NTSC analog video requires an SNR of ~50 dB (peak signal voltage/Rms noise voltage) and linearity, HD2 and HD3, need to be suppressed below −60 dBc. Cable plants with achievable SNRs of 40 dB (RMS power/RMS noise) allow for feasible deployment of modulation orders as high as 256-QAM (8 bits per symbol). For a symbol rate of 5 Mbaud, this would correspond to a bit rate of 40 Mbps (megabits per second). Today most North American cable operators routinely deploy downstream QAM orders of 64-QAM at approximately 5 Mbaud for a downstream data rate of 30 Mbps.

3. PHYSICAL LAYER DOWNSTREAM: RF AND AFE REQUIREMENTS

The downstream RF front end for a cable modem begins with a TV tuner designed to downconvert and filter the unwanted adjacent channels of the RF frequency spectrum of 54–860 MHz to an IF frequency of 43.75 MHz for NTSC systems or 36.125 MHz for European PAL systems. Cable service providers usually place the digital video and data channels together above 400 MHz, although they can reside at any frequency in the spectrum mentioned earlier. Traditional TV tuners (so-called single-conversion tuners) take the RF input and mix it with an LO (local oscillator) offset by the IF frequency to place the incoming RF signal at the desired IF (43.75 MHz in the NTSC case). Since these tuners were designed for off-air reception, the LO leakage back into the antenna was of little concern because of the limited range of radiation. However, this LO leakage was found to be of great concern when the RF was connected to a cable plant. The normal worst case would be for all 200 modem users to tune to the same RF channel, and since these LOs are at the same frequency but not correlated, they would add in an root sum of squares (RSS) sense at a frequency offset of 43.75 MHz

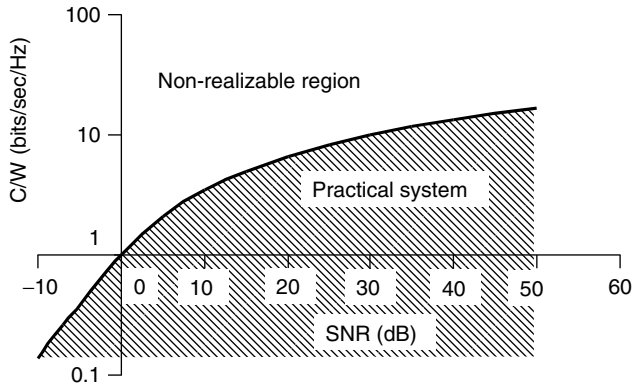


Figure 4. Channel capacity.

from the tuned RF channel. This would result in significant LO power leaking to the adjacent RF channel, which might be in use by other viewers of analog television or digital video/data, which could result in corruption of that signal. To circumvent the LO leakage problem, designers have increased the reverse isolation of the low-noise amplifier (LNA) at the very front end of the tuner or switched to what is commonly called a "double-conversion tuner architecture." Basic operation of a double-conversion tuner is to first upconvert the entire spectrum to a much higher frequency (typically 1–2 GHz), process the channel selection using a filter and then downconvert the selected channel back to the required 43.75-MHz IF. In this way, the frequency plan places the mixer LOs out of the desired frequency spectrum and thus mitigating the LO leakage issue altogether. As with any RF system, the first-stage NF (noise figure) will dominate the obtainable receiver system SNR. For RF input levels of −15 to +15 dBmV, as specified by DOCSIS, the tuner must have a NF of 10 dB or less to meet this specification with a reasonable margin of 2.5 dB above the FEC limit (Fig. 5). In addition to the noise and input level, the QAM receiver carrier recovery loop is sensitive to the phase noise contribution of the tuner, and typical values of −85 dBc at a 10-kHz offset are required. Single conversion tuners have been accepted into cable plant operations due to the low phase noise, low cost, and good NF. However, more recent advances in silicon tuner technology have introduced inexpensive double-conversion tuners in standard bulk CMOS process technology, which promise the possibility of integration onto the cable modem chip, thus further simplifying the cable modem design.

Following the downconversion to the 43.75-MHz IF, a surface acoustic wave (SAW) bandpass filter is placed in the signal path to eliminate any residual power from the adjacent channels that may not have been attenuated by the tuner and also to band-limit the noise. Stop band attenuation, passband ripple and group delay variation are all important design constraints for this filter. Virtually all QAM receivers employ an adaptive equalizer as part of the digital processing, and this can relax some the SAW filter requirements. Equalizers are unable to compensate any signal that is not correlated to the input such as AWGN. However, any symbol-spaced equalizer can correct for correlated linear distortions produced by the cable link or RF/AFE/ADC as long as they occur within the equalizer length time interval and the equalizer has sufficient dynamic range to compensate for them. This is a very powerful result, and we will see how this will affect the design requirements of the ADC and the AFE. In a typical QAM system, symbol rates of 5 Mbaud equate to symbol periods of 200 ns, and equalizer dynamic ranges of ≥10 dB allow the use of inexpensive SAW filters, which typically have passband ripple as large as 2 dB and the group delays of ≥50 nS. The equalizer will correct for these distortions, which create intersymbol interference (ISI), resulting in degraded QAM performance, and will attempt to produce a flat channel response. Generally, these inexpensive SAW filters have significant insertion loss, as much as 20 dB, and thus a fixed-gain amplifier is needed to compensate for this attenuation. The driving specification for this amplifier is low noise. Any distortion that produces ISI will again be compensated by the equalizer. An automatic gain control (AGC) amplifier is included in the tuner RF front end and is closed around this IF amplifier; thus gain drift will be automatically compensated. These two facts allow for consideration of a low-cost open-loop design for the IF amplifier; however, a significant gain is required on the order of 34–36 dB. In current implementations, this amplifier drives the input of the single-chip cable modem QAM receiver directly. An internal analog programmable gain amplifier (PGA) provides additional AGC range and the ADC samples the 43.75-MHz IF in a subsample mode; thus sample rate less than the input IF frequency usually in the range of 20–30 megasamples per second. Since the input frequency is slewing faster than the sample clock, aperture jitter of the sample hold is an important consideration.

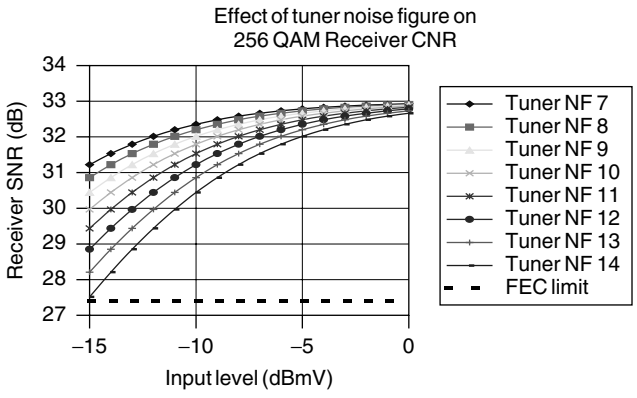


Figure 5. SNR versus input level for 256-QAM.

Traditional specifications for the ADC refer to the effective number of bits (ENOB) as the figure of merit. But for communication systems, more information is needed to optimize QAM receiver performance and reduce ADC complexity. ENOB is a metric of the combined SNR and distortion components (linearity) of an ADC. As has been shown in much communication literature, the SNR for a given ADC sampling at sub-Nyquist (nonoversampled) can be found from the following equation:

$$\text{SNR} = 6.02 \times N + 1.76\,\text{dB}$$

where N is the number of bits for the ADC.

Thus, for a 10-bit ADC, the maximum theoretical SNR that can be expected is 61.96 dB. Typical values of 59–60 dB SNR (peak signal/RMS noise) are common place for integrated ADCs in a bulk CMOS process. For a QAM receiver application, the obtainable SNR is the most important component of the ADC performance; however, the required distortion is the more interesting ADC performance parameter and must be broken into the two constituent parts [integral nonlinearity (INL) and differential nonlinearity (DNL)]. DNL can be described as the worst-case code-to-code variation in an ADC. What this means is that, as the input signal transverses the quantization levels of the ADC that contains DNL, there will be instantaneous "spikes" or distortions at the code

transitions that are not linear. Since a QAM signal has already been randomized (a scrambler is requisite for reliable transmission), these instantaneous code errors are pseudo-non-correlated and will produce a "white Gaussian noise–like" distribution. In the case of INL it is quite a different matter. INL can be described as the average nonlinear component composed of all the cells in the ADC. Generally, as the input signal swings from minimum to maximum full scale, this type of distortion will produce a second-, third-, or higher-order correlated component, and for a QAM-modulated signal the result can introduce ISI. As we will see later in this article, the ISI component can be compensated by an adaptive equalizer, and thus the DNL contributes more to degrade the performance of an ADC in a QAM system than does the INL, as mentioned earlier. An additional ADC parameter that will have an effect on receiver performance, especially when operated in the subsample mode, is the aperture jitter of the sample hold. Aperture jitter is the instantaneous amplitude error produced by imperfect sample time periods of the ADC. Typical methods for specifying this value have been based on single-tone analysis and assuming that zero crossing of a SIN wave is the worst-case slew, using the slope here for the worst-case measurement. Observing QAM constellations currently in use, it can be seen that there are no symbol decisions at zero crossing, and thus this measurement is far too pessimistic for these types of systems. The proper method for measurement of this parameter is to look at the slope of the slew on an eye diagram at the symbol decision points as shown in Fig. 6, and calculate the worst-case aperture error in this region.

Additional spreading of integral nonlinearities and distortion can be found by examination of the QAM signal itself. As mentioned earlier, in order to transmit and receive a QAM-modulated signal in a robust manner, a guaranteed level of randomization must exist in the signal; that is, the distribution of the modulated symbols must be equally likely. This is accomplished by inserting a pseudo-random bit sequence (PRBS) logically xORed (exclusive-ORed) with the data on the transmitter side, and an identical PRBS pattern on the receiver side to reverse the process. The effect this has on the original signal, and more importantly the distortion products, is to distribute them over the symbol rate bandwidth which makes them appear as additive noise components (Figs. 7 and 8). This is a very important result for a QAM system and again illustrates how important the noise contribution is to system performance and how integral distortion products are actually spread and less significant. Of particular importance related to distortion are the intermodulation distortion (IMD) products because these will appear as sidebands and can behave as adjacent-channel interference to the desired signal of the receiver. In a real HFC cable plant, many impairments inhibit the maximum SNR and BER that can be achieved and must be minimized or compensated in order to maintain a reliable data link. Figures 9–13 serve to illustrate the types of impairments that will be encountered in a typical cable environment and the effect that these imperfections will have on the receiver performance.

After the input signal has been properly downconverted and sampled by the ADC, the QAM receiver performs a complex (real/imaginary) digital down conversion to baseband (DC) for carrier recovery and symbol rate conversion. The most common form of quadrature direct digital frequency synthesis (QDDFS) is performed by using separate SIN and COS lookup tables (stored in read-only-memory) and what is known as a numerically controlled oscillator (NCO) as a mixer LO. The digitally sampled data from the ADC are split into two paths, and each set of data is input to separate mixers using the appropriate SIN (Q component) and COS (I component) driving the LO. At this time, a slight frequency offset can be added to each of these LO's to remove any gross systematic carrier frequency offset induced from components in the RF/AFE or any source that contributes a constant frequency offset. Following the quadrature downconversion, baseband processing begins with lowpass filtering to remove the image created from this conversion and then Nyquist matched filtering and timing recovery. It is well known in communications theory that in order to maximize the AWGN performance and minimize ISI, a set of identically matched filters at the transmitter and receiver must be used. The necessity of this filter is to ensure that the slew-rate-dependent properties (both amplitude and phase) of the modulation and transmission link are compensated. The convolution of the transmit and receive filters, ignoring the transmission link for the moment, gives the overall system response, which has the Nyquist property of zero ISI at symbol-spaced sampling instants (as seen in an eye diagram). Each matched filter is thus referred to as "square-root Nyquist." In addition, finite excess bandwidth must be provided beyond the ideal filter responses. The excess bandwidth damps the time-domain response of the filter and reduces sensitivity to timing recovery errors. Since it takes a finite amount of time to transverse from one symbol to the next, an additional amount of system bandwidth is required to ensure that each symbol can be transmitted and decoded properly. The most common filter mask (frequency response) for the matched filter is referred to as the *square-root raised-cosine filter*, and the mathematical form of the impulse response can be found from the following equation:

$$g(t) = \left[\frac{\sin(\pi t/T)}{(\pi t/T)}\right]\left[\frac{\cos(\alpha\pi t/T)}{(1-4(\alpha t/T)^2)}\right]$$

where α is excess bandwidth and T is 1/symbol rate. Note that when $\alpha = 0$, representing 0 excess bandwidth, this equation collapses to the following familiar form:

$$g(t) = \frac{\sin(\pi t/T)}{(\pi t/T)} = \text{sinc}\left(\frac{\pi t}{T}\right)$$

After the I (in-phase) and Q (quadrature) samples have been filtered, they are passed to the timing recovery loop, sometimes called the *baud / symbol loop*. The simplest form of timing recovery uses an I and Q zero-crossing detector to drive a simple integrator that controls a variable oscillator, and thus can lock and track any instantaneous changes in the input data timing. Generally this loop can be modeled as a phase-locked loop (PLL) with a second-order loop filter (integral and linear terms) and including an additional constant offset term equal to the desired symbol rate

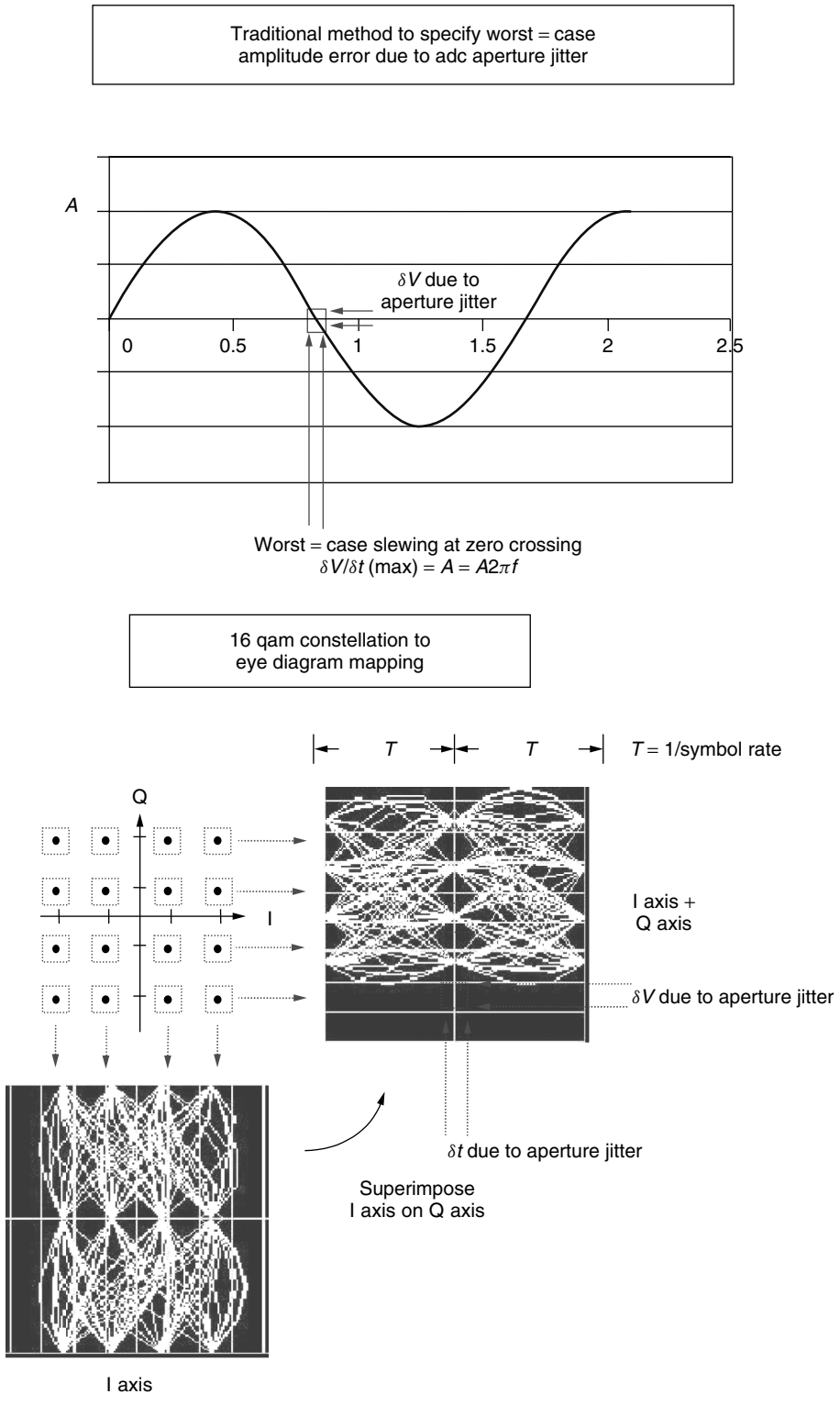


Figure 6. Aperture jitter in ADC applications for QAM.

of the receiver. If a difference term were included, this would form the popular proportional, integral, differential (PID) controller found in many modern control systems. Once the symbol timing has been recovered, the basic I/Q constellation will be formed but will need to be rotationally stabilized. This process is completed using the derotator or carrier recovery loop. Since the exact location of each ideal constellation point is known for a given QAM constellation, and the I/Q information from the receiver has been decoded, it is a relatively simple task to compute the phase difference from the received points and add it back in to compensate any rotational errors which have been introduced. Again this is done using a second order PLL structure similar to the timing recovery loop. An adaptive

equalizer, usually consisting of feedforward (FFE) and decision feedback (DFE) taps and implemented using a least-mean square (LMS) algorithm compensates for channel distortions and coax cable multipath reflections (Fig. 14). In many cases the output of this equalizer is used as an input to the carrier recovery loop, thus providing a corrected soft decision to drive the convergence of that loop, which dramatically improves the performance under impaired channel conditions.

At this point the QAM demodulation is complete and all that is left is to slice and demap the constellation points (soft decisions) back into a bitstream, derandomize and decode the forward error correction (FEC) blocks, and resolve the MPEG framing. A concatenated FEC consisting of inner trellis-coded modulation followed by an outer Reed–Solomon code are specified by DOCSIS. Also specified are various packet interleaving options that distribute any clustered errors over a number of packets, providing immunity to burst noise. MPEG-2 framing defines each packet to consist of 188 bytes with the first byte (or sync byte) to be the hex value 0x47. In addition, every 8th packet shall have the sync byte inverted (hex value 0xB8) in order to facilitate acquisition and lock retention of the packet stream.

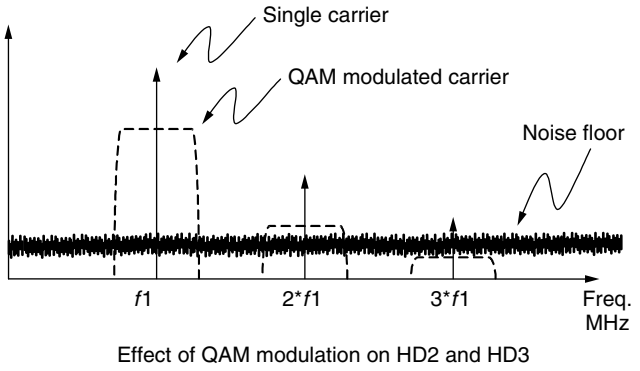


Figure 7. Distortion in a QAM modulation system.

4. PHYSICAL LAYER UPSTREAM: IF REQUIREMENTS

The upstream burst modulator consists of digital I/Q data stream that utilizes quadrature direct digital frequency synthesis (QDDFS) to upconvert to the desired IF frequency for transmission from the CM back to the head end (CMTS) and the Internet server. A high-speed DAC (with sample rate typically $\geq$200 MHz) is used to convert this digital IF to an analog voltage. The modulation format is either QPSK or 16-QAM for all current modems, with 64-QAM already defined in the next generation of the DOCSIS 2.0 specification. Since the upstream data are transmitted in a burst mode, a means for packet synchronization is necessary. This is accomplished by the addition of a preamble at the beginning of each packet, which allows the receiver (residing at the head-end equipment) to synchronize before the actual payload data

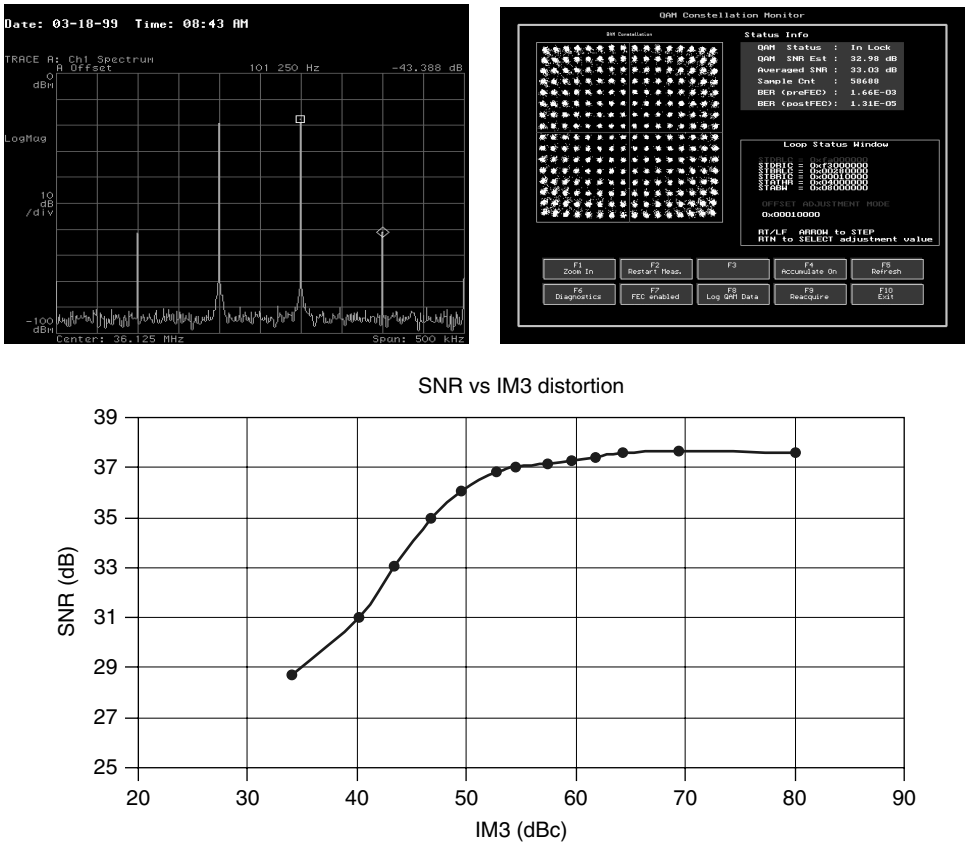


Figure 8. Effects of IMD distortion on 256-QAM.

No added C/N

Added C/N = 21 dB

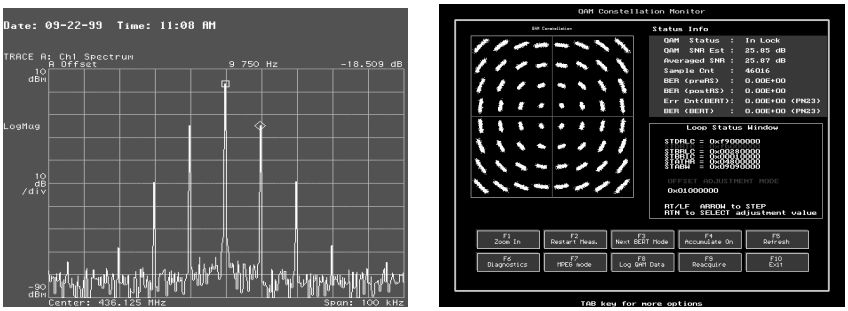
Sources ofAWGN (additive white Gaussian noise)

- Tuner noise figure
- Broadband noise in amplifiers
- ADC sample/hold aperture jitter
- Round-off errors in digital truncation
- Nyquist filter mismatch (appears like AWGN)

Methods to improve AWGN performance

- Reduce tuner noise figure to <9 dB
- Low noise amplifiers in RF/IF signal chain
- Direct clocking of ADC sample/hold
- Ensure matched α for

Figure 9. Effect of broadband AWGN on 64-QAM performance.



Added Φ_n = 10 kHz FM, 5 kHz Deviation
Carrier loop BW = 10 kHz

Sources of Φ_n

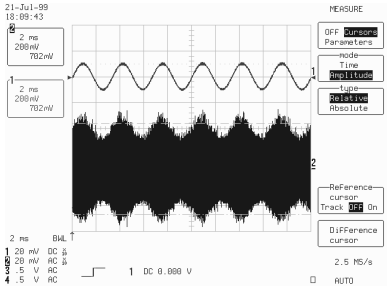
- Tuner LO
- Phase modulation of ADC sample/hold clock
- Poor supply decoupling of RF and AFE

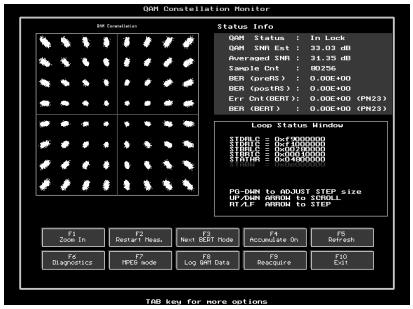
Methods to improve Φ_n performance

- Tuner LO Φ_n < –85 dBc/Hz @ 10 kHz offset
- Optimize carrier recovery loop BW (tradeoff between AWGN and phase noise effects)
- Low-noise narrowband PLL's
- Optimize power supply decoupling

Figure 10. Effects of phase noise (ϕ_n) on 64-QAm performance.

are demodulated. A simple pattern of 0xCCCCC0D is appended to the data, which corresponds to I/Q zero crossings (CCCCC) plus a unique word (0D) and has been determined to be adequate for locking a quadrature system (Fig. 15). Unlike a continuous receiver, if the burst receiver is not able to lock to the preamble, then the entire packet is lost, creating packet errors that are much more severe than bit errors. The basic burst modulator functional block diagram begins with a set of first-in first-out (FIFO) data buffers, allowing the front-end digital data for the next burst event to be loaded asynchronously while the analog IF output is transmitting the current burst, thus forming





Added AM: 20% modulation (200 mVpp/1Vpp); frequency = 130 Hz

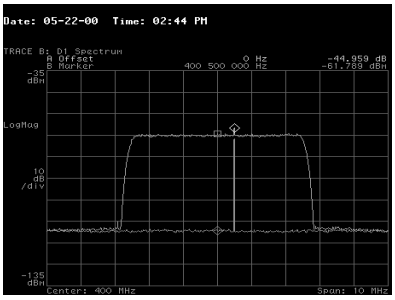
Sources of AM

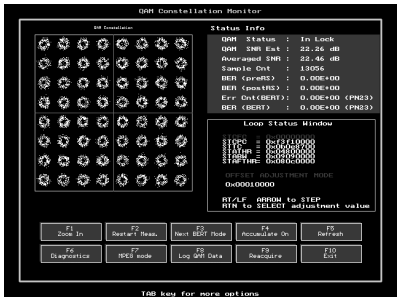
- 2nd harmonic of 60 Hz power line (120 Hz)
- Low frequency power supply ripple
- AGC loop bandwidth set incorrectly
- Incorrectly terminated grounding for RF/AFE

Methods to improve AM

- Improve filtering for power supply
- LC or ferrite in DC supply
- AGC loop BW and dominant pole set higher than frequency of AM impairment

Figure 11. Effects of AM on 64-QAM performance.





Added RFI = –24 dBc @ 401.0 MHz; 400 MHz
center frequency of RF input

Sources of RFI

- Fixed tone from digital clocks/oscillators
- Mixing products of adjacent NTSC channels
- CSO and CTB from loading (130 RF channels)
- Incorrectly terminated grounding for RF/AFE
- Ingress into cable

Methods to improve RFI

- Improve filtering for RF/AFE/digital power supply
- Change FFE main tap location in equalizer
- Narrowband notch filters
- Add high frequency ferrites to isolate grounds

Figure 12. Effects of RFI on 64-QAM performance.

an effective pipelined transmission. As data are pulled out of this FIFO, they are passed to the randomizer and the FEC, which scramble the symbol bits and encode them for transmission. Current implementations for the FEC are a programmable Reed–Solomon (RS) code with various Galois field selections and T values (number of correctable bytes) ranging from 1 to 10. This block calculates the FEC code parity bytes that are appended to the end of each burst and used by the receiver to correct for errors generated in the transmission link. After the preamble is

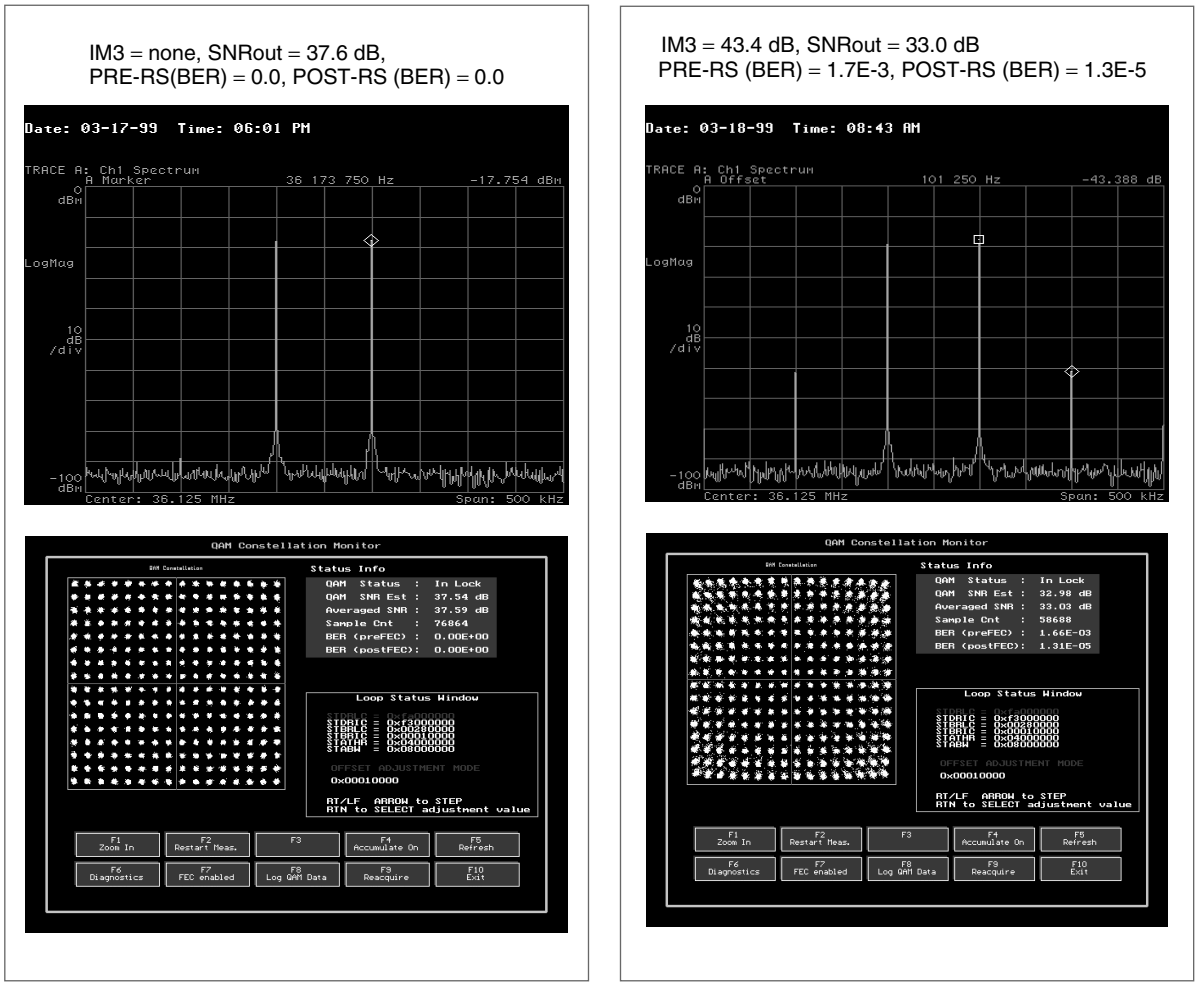
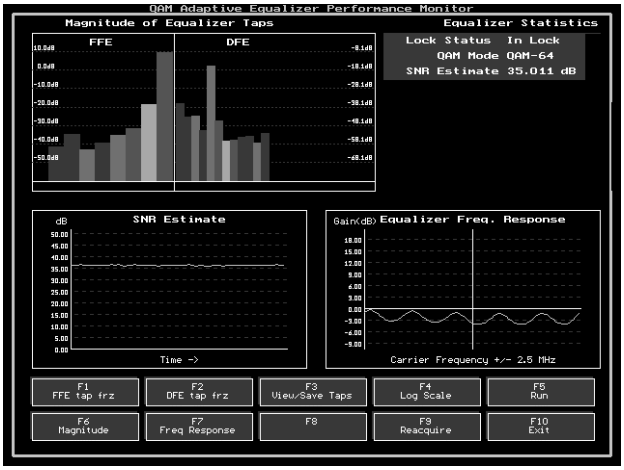


Figure 13. Effects of IM3 on 256-QAM performance.

inserted to indicate the beginning of a burst, the symbol mapping block performs a bitwise grouping into separate I and Q bitstreams. Each of these data paths is passed to matched square-root Nyquist filters that shape the necessary excess bandwidth to ensure that ISI remains at a minimum. Currently used alpha values are 0.25 (25%) for the DOCSIS upstream. In addition, on the head-end side, the receiver operates in the burst mode. To minimize preamble overhead, the receiver is not required to converge an adaptive equalizer on each burst, since each burst may come from a different transmitter having different channel characteristics. Thus an alternate means to provide echo cancellation must be derived. The DOCSIS 1.1 specification addresses this problem by defining a complex (real and imaginary) preequalizer residing in the upstream modulator that can effectively predistort the transmitted signal to cancel any echoes that will be generated in the transmission path for that particular modem. This preequalizer is located just prior to the Nyquist filtering. Coefficients for each modem's preequalizer are sent via the downstream channel from the head-end receiver on the basis of the received channel response for each particular modem. Following the Nyquist filters, the I/Q symbols are processed by a variable interpolating filter that upsamples the signal from the symbol rate to the DAC sample rate. From there the data enters the QDDFS, which consists of a structure similar to the downstream receiver whereby SIN and COS lookup tables are used as mixer LOs. These digital mixers upconvert and combine the digital data that are subsequently input to the DAC to create the analog IF output frequency.

The analog portion of the upstream design begins with an image reject filter following the QDDFS and DAC. The DAC will produce an image of the desired IF frequency (sample rate−IF frequency) and this image must be attenuated so as to not over drive the input to the power amplifier or leak into the upstream output. Typically, a high-order analog Chebyshev or elliptical filter is used for this purpose. The filter specifications can be relaxed with higher-frequency sample rates resulting in a higher-image frequency. Noise and distortion are very important for the upstream IF path because they must not be allowed to interfere with the lower downstream channels beginning at 54 MHz. With this in mind, a fully differential topology is preferred, allowing the cancellation of HD2, leaving HD3 to contend with. There is some help from the diplex filter at the output to attenuate the third-order product since it must have a sharp stopband attenuation to keep

64 QAM DFE real tap #5



64 QAM, 5 Mbaud, delay = 1uS, Phase = 320 degree, attenuation 10 dB

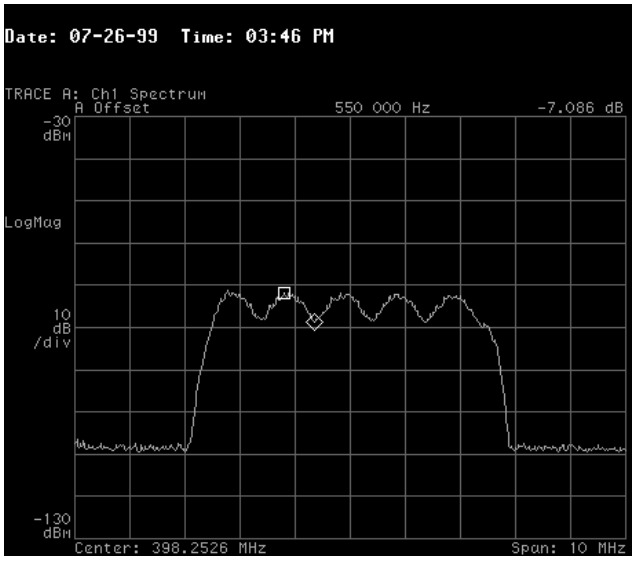


Figure 14. Multipath reflections.

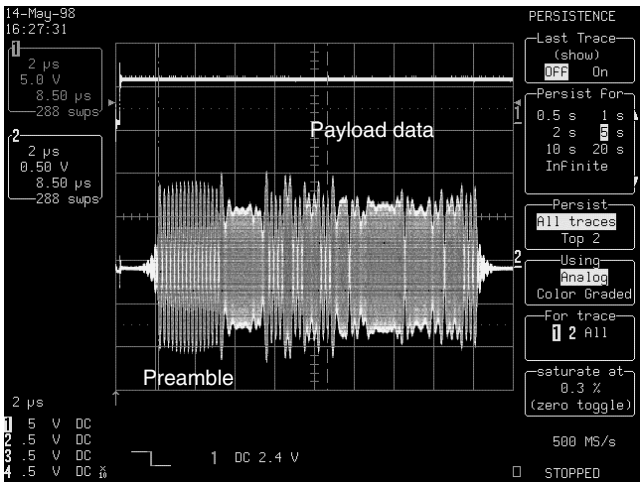


Figure 15. Burst upstream data packet.

any spurious and noise out of the downstream channels. Additionally, as with the downstream receiver, the intermodulation products must be kept to a minimum. The power amplifier that drives the upstream data back through 75-Ω cable must supply +8 to +58 dBmV of signal level implying a variable attenuation of 50 dB. Common practice is to design the burst modulator with an analog programmable gain amplifier (PGA) as fine gain control (25 dB in 0.4-dB increments) and design the power amplifier to supply the remaining 25 dB in coarse steps (6 dB). In addition to the needed large signal and low distortion, the amplifier must have very low noise (DOCSIS spec is −59 dBmV/Hz), which necessitates disabling it in between data bursts. This in turn leads to the potential turn-on/off glitch, which has been specified to be less than 100 mV integrated over 200 ns. To make matters even more difficult, the on/off impedance match must maintain a 75-Ω termination to ensure proper cable termination and return loss under all conditions, and it is desirable to use a single +3.3-V-DC supply or common voltage required for the modem chip. Taking all of this into consideration makes for a very challenging design (Fig. 16). The leading cable modem chip designers have integrated this amplifier onto the cable modem chip, thus extending the state of the art and providing further cost reductions and simplicity for modem product designers.

5. MAC LAYER: DATA PROTOCOLS

The cable modem MAC acts as the data parser and decoder to enable the DOCSIS point-to-multipoint communication system. As pointed out earlier in this text, DOCSIS employs a continuous downstream signal and a TDMA burst upstream signal, and the MAC acts as the basic controller between the modem and head-end equipment residing at the service provider. The DOCSIS specification defines many different packet types and usage codes, called interval usage codes (IUCs), but by far the three most essential MAC messages are the SYNC (synchronization), upstream channel descriptor (UCD), and minislot usage information (MAP).

The basic process flow of channel acquisition begins with the downstream receiver scanning all RF channels and obtaining QAM and FEC lock in search for MPEG packets containing the well-known PID for a DOCSIS data channel (0x1FFE). Once a DOCSIS channel has been

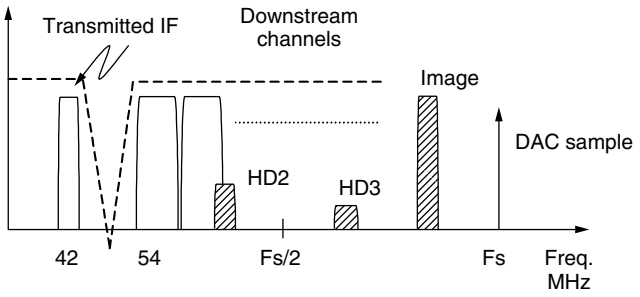


Figure 16. Burst upstream frequency spectrum.

found, the cable modem (CM) begins looking for the three MAC messages that are regularly sent from the head end (CMTS). The first necessary step is to synchronize the CM with the CMTS and all other modems in the system. This is accomplished by the CMTS, which sends a periodic SYNC message containing a 32-bit timestamp over the downstream channel. The CM receives the SYNC message and locks the frequency of its local clock so that it matches the time stamp. This process may require many SYNC messages before the CM's local clock is adequately tracking the CMTS reference clock. For upstream TDMA data transfers, the concept of minislots (a convenient partitioning of time) is used, instead actual number of bytes of information to transmit, which facilitates bandwidth allocation when switching modulation types. Once the modem has determined the common time reference, the next message required is the UCD. The UCD instructs the CM to adjust a number of upstream parameters such as the transmitter frequency, modulation type, symbol rate, minislot size, preamble pattern, and selection of a burst profile to use for further communication. This is the initial setup needed to establish basic two-way connectivity with the CMTS. The final step in channel acquisition is the processing of the bandwidth allocation in the MAP message, which corresponds to the upstream described in the UCD. This message designates the minislot information and is used to establish at what time and for how long the modem can transmit, with the SYNC message providing the time reference for these transmissions. The MAP messages assign burst type (via IUCs) and burst duration (via minislots) to upstream SIDs. The upstream SIDs are used for bandwidth allocation and security associations. The initial signon process uses a special time slot called "initial maintenance," denoted by IUC 3. At this point in the process, the modem has established (1) a time reference, (2) initial upstream transmission configuration, and (3) knowledge of when and for how long to transmit.

While the previous steps provide the modem with a notion of relative time, the CM still needs to know the exact time. The SYNC messages that provide the CM with its notion of time incur propagation delays as they travel from the CMTS to the CM. This propagation delay will vary depending on the position of each CM on the cable plant. Thus, each CM has a relative notion of time through frequency locking to the SYNC messages, but not an exact notion due to propagation delay. A process called "ranging" adjusts each CMs notion of time to be the exact notion required for TDMA operation with the CMTS. The ranging process begins with the modem sending the head end a ranging request. A number of problems may prevent the CMTS from issuing a ranging response message acknowledging the modem. The CM ranging request could collide with another modem which is initiating the logon process or the transmit power of the CM may be too low for the head end to receive it. Therefore, the ranging request will be repeated with appropriate time backoff and power adjustment until eventually the head end will acknowledge with a ranging response and will send the CM a dedicated ranging opportunity called "station maintenance," denoted IUC 4 in a new bandwidth allocation MAP. The CM will now automatically transmit a ranging request in any station maintenance slots reserved for it. At this point the CM and the CMTS enter an interactive mode whereby fine adjustments are made to the transmit frequency (must be with ± 10 Hz of commanded), transmit power (must be within ± 2 dB of commanded), time offset (must be within 1 μs of commanded) and multipath reflections (preequalizer tap adjustments). This may take many fine adjustments with the final outcome of calibrating out any time and amplitude variations for each modem's roundtrip data. Once the CMTS detect the CM is properly ranged, the CMTS sends a ranging response message with a ranging complete notation. The CM then uses request regions denoted IUC 1 to send up a request for the bandwidth required to transmit its nonranging packet. The head end will respond with the bandwidth allocation MAP, granting the modem the bandwidth requested, and the CM MAC can now send its first nonranging information to the CMTS.

Next the CM needs to establish IP layer connectivity. This is accomplished by use of the dynamic host configuration protocol (DHCP), which will assign the CM an IP address and form the IP link between the modem and the DHCP sever. When the modem has terminated connectivity, the IP address will be relinquished back to the pool, and DHCP can reallocate it to another IP user. Registration of the modem begins with the CM downloading a configuration file using trivial file transfer protocol (TFTP) and establishment of a service identification (SID). Only after a number of file checks and authorization confirmation will the CM be allowed to transmit "real" data onto the cable system. At this point the modem is able to transmit and receive data, but one final step in basic connectivity must be performed. Since the cable protocol is a shared medium, a means to protect and secure data transfers is necessary. This is accomplished by what is known as baseline privacy (BPI). Each modem is uniquely identified with a 48-bit MAC address, which can only obtain BPI encryption keying information it is authorized to access. BPI uses the Cipher Block Chaining mode of the data encryption standard (DES) algorithm to encrypt data in both upstream and downstream data paths. The CM uses RSA, a public-key encryption algorithm (proprietary to RSA Data Security, Inc.) to obtain authorization and encryption keys from the head end and to support periodic encryption key changes. The cable operators determine how often new encryption keys are sent. This final step relinquishes the cable modem to "surf the Net" at typical downstream data rates of 1–2 Mbps.

6. CONCLUSIONS

This article presents an overview of the cable modem system and descriptions of some of the key components found in cable modem equipment. A detailed discussion of the four DOCSIS layers (PHY, MPEG, MAC, BPI) is examined. Current video delivery to most homes in the United States is via cable, and as more interactive services are offered, there will be increasing emphasis on providing simultaneous high-speed data available to these users. The bandwidth is available from the existing cable

plants to provide this growth. Cable modems have shown increases in speed of 1000 times over telephone modems, and nearly all housing developments in the United States have a cable infrastructure already in place. Increased levels of integration have dramatically reduced the cost of cable modems, enabling explosive growth and accelerated deployment for the near future.

Acknowledgments
The author wishes to acknowledge the contributions of Dr. Charles Reames, Lisa Denney, Bruce Currivan, Dr. Thomas Kolze, and Dr. Henry Samueli for valuable advice and criticism of this text.

BIOGRAPHY

Donald McMullin (M'87) received his B.S. in electrical engineering from California State University Northridge. He joined the Electro-optical and Data Systems Group of Hughes Aircraft Company, El Segundo CA, in 1988 where he worked on forward looking infrared night vision systems and Aided Target Recognition computers. In 1992 he joined the Advanced Circuit Technology Center of Hughes Aircraft and specialized in full custom analog chip design for high performance data converters. In 1996 Mr. McMullin joined Broadcom Corp. and focused his attention on QAM receivers/modulators and broadband cable data transmission systems. He is currently the manager of hardware development for cable products at Broadcom Corp. Donald holds 3 patents for amplifier design topologies intended for data converter applications and has 4 patents pending in the area of communication design.

BIBLIOGRAPHY

1. DOCSIS (Data-over-Cable Service Interface Specification), *Radio Frequency Interface Specification*, SP-RFIv1.1-I08-020301.
2. B. Sklar, *Digital Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
3. J. Min and H. Samueli, Frequency-agile TDMA system for upstream cable-modem applications and B. Currivan, Cable modem physical layer specification and design, in *Cable Modems: Current Technologies and Applications*, International Engineering Consortium, Chicago, 1999.

CARRIERLESS AMPLITUDE–PHASE MODULATION

BURTON R. SALTZBERG
Middletown, New Jersey

Carrierless amplitude–phase modulation (CAP) is a variation of quadrature amplitude modulation (QAM), in which explicit modulation and demodulation is omitted. In virtually every aspect, the performance, analysis, and most of the implementation techniques of QAM are applicable to CAP. While QAM has been the preferred modulation technique for a wide variety of applications for many decades, CAP has been used for digital communications only in recent years, largely for transmission over wire-pair channels.

To illustrate conceptually how QAM may evolve into CAP, we first show a standard QAM system in Fig. 1. A datastream of user bits is first assembled into pairs of symbols, each of which is chosen from an alphabet representing some number of bits. The symbol pair may be considered to be a two-dimensional symbol, or a complex quantity. The mapper chooses a point in two-dimensional, or complex, space for each possible symbol value. The set of such points is the signal constellation. The real and imaginary values of those points are denoted as the I and Q components. Each component is lowpass-filtered, as in pulse amplitude modulation (PAM), and input to a modulator that multiplies that component by one of two sinusoidal carriers. The carriers are at the same frequency and differ in phase by 90°. The modulated signals are added and presented to the transmission channel. At the receiver, the line signal is demodulated by the same two carriers to form a pair of baseband signals. An essential component of the receiver is a means of reconstructing the carriers. The baseband filters may include an equalizer. The detection process is then completed and the bit stream reconstituted.

It has long been recognized that the order of filtering and modulation may be interchanged in either the transmitter, receiver, or both. In fact, passband filtering and equalization are quite common in QAM receivers. Figure 2 shows the case in which this interchange is done in both the transmitter and the receiver. Each pair of lowpass filters is replaced by a Hilbert pair of bandpass filters. The lowpass filter with impulse response $f(t)$ is replaced by the following pair:

$$f_1(t) = f(t)\cos\omega_c t$$
$$f_2(t) = f(t)\sin\omega_c t$$

The filters form a Hilbert pair, in that

$$f_2(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f_1(u)}{u-t}\,du$$

or, more understandably in terms of the Fourier transforms, as

$$F_2(f) = j\,\mathrm{sgn}(f)F_1(f)$$

which says the two frequency responses are equal in amplitude and differ in phase by 90° over the entire frequency range. The important property in this application is the orthogonality of the Hilbert pair:

$$\int_{-\infty}^{\infty} f_1(t)f_2(t)\,dt = 0$$

Examination of Fig. 2 reveals that the only function served by the modulator is to multiply the complex data symbols by $e^{j2\pi f_c t}$, and the demodulator is to multiply by $e^{-j2\pi f_c t}$. This amounts to a rotation and a counterrotation by the same quantity. The final and key step in forming the CAP system is simply to eliminate this rotation and counterrotation as shown in Fig. 3. If $f_c T$ is an integer,

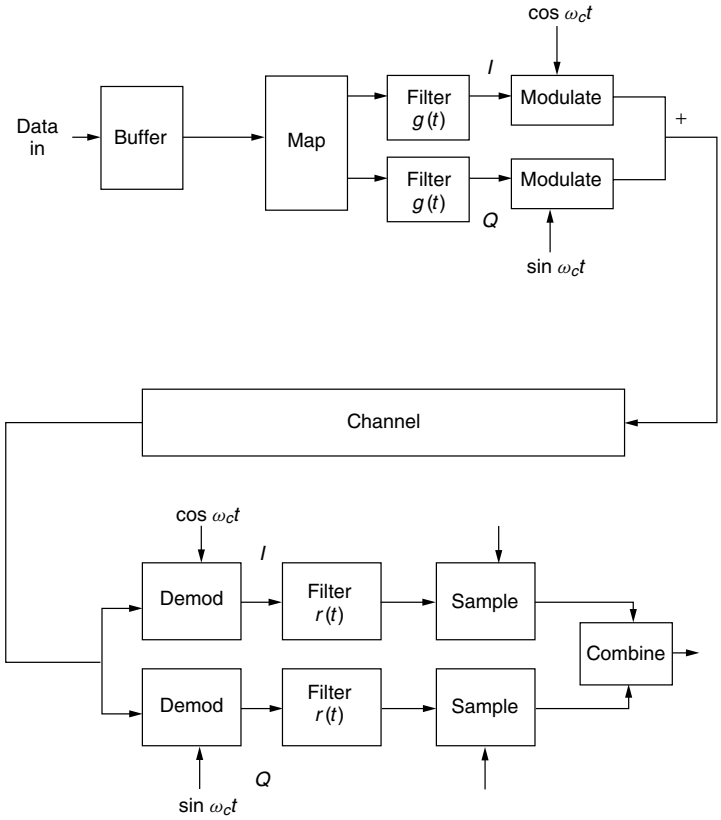


Figure 1. A standard QAM system.

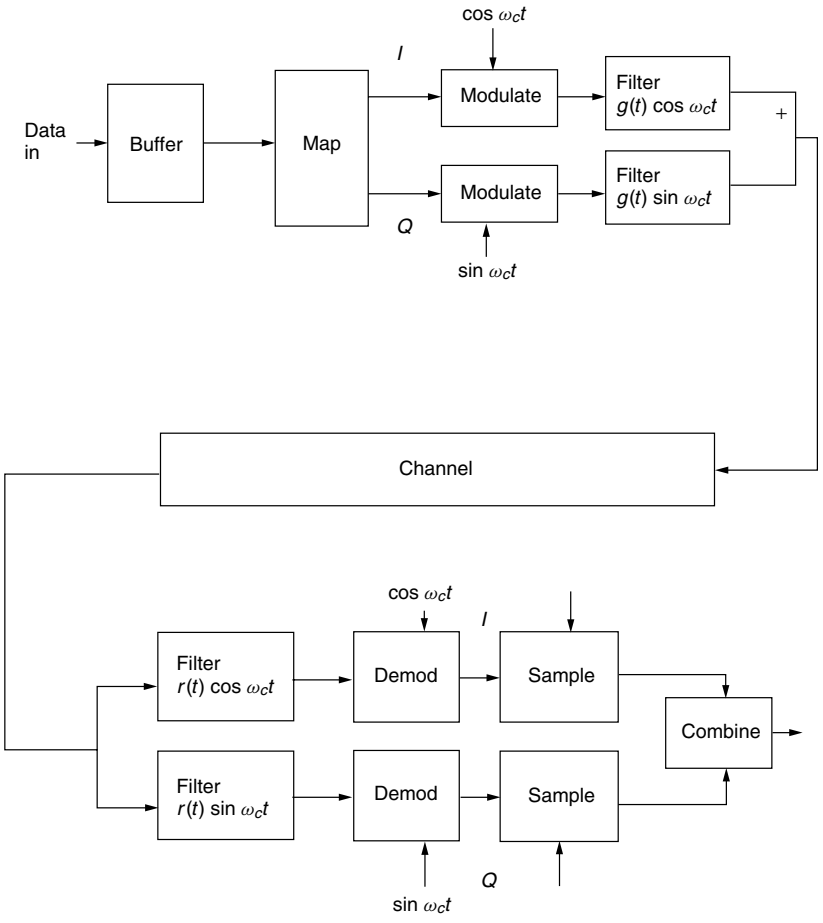


Figure 2. A QAM system with bandpass filtering.

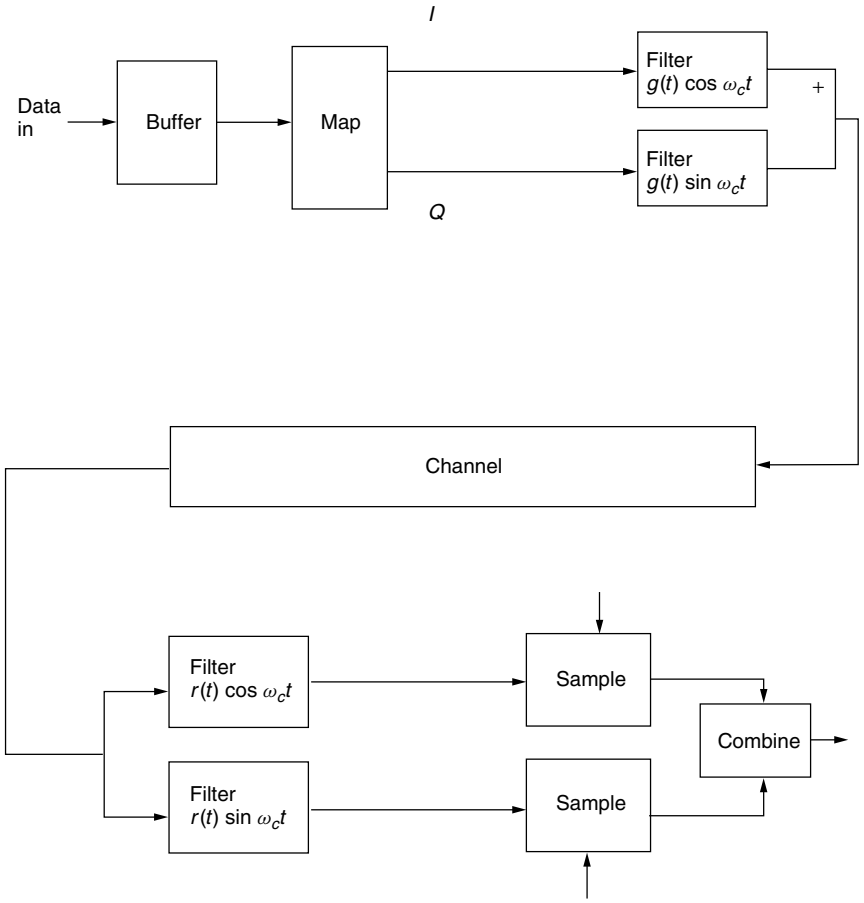


Figure 3. A CAP system.

where T is the symbol duration, then the QAM and CAP systems are totally identical. If f_cT is an integral multiple of $\frac{1}{4}$, and the constellation is symmetric, then they are again identical except for coding of the symbols. In general, for any value of f_cT, the corresponding QAM and CAP systems are virtually identical except for a unitary rotation of the constellation.

CAP modulation may be explained in terms of PAM rather than QAM. Consider a PAM system using bandpass filtering in place of the usual lowpass filtering. Such a system would require twice the bandwidth of a standard PAM system. However, if two such systems are superimposed on the same frequency band, bandwidth efficiency is achieved in that the same bandwidth is required for the same bit rate. The two passband signals may be superimposed and separated at the receiver if the bandpass filters form a Hilbert pair. This is illustrated by Fig. 4.

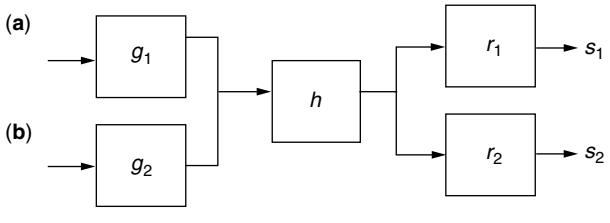


Figure 4. Basic model of a CAP system.

Let the convolution of each of the transmit filters, the channel, and each of the receive filters be denoted by an overall response

$$z_{ij}(t) = g_i(t) * h(t) * r_j(t).$$

Then the received signals are

$$s_1(t) = \sum_k a_k z_{11}(t - kT) + \sum_k b_k z_{21}(t - kT)$$

$$s_2(t) = \sum_k a_k z_{12}(t - kT) + \sum_k b_k z_{22}(t - kT)$$

Intersymbol interference is eliminated if the usual Nyquist condition is met for each subchannel:

$$z_{11}(t - kT), z_{22}(t - kT) = 1 \quad \text{for} \quad k = 0$$

$$\text{and 0 for all integers} \quad k \neq 0$$

Interchannel interference is eliminated if

$$z_{12}(t - kT), z_{21}(t - kT) = 0 \quad \text{for all integer } k$$

Under these conditions, $s_1(kT) = a_k$ and $s_2(kT) = b_k$, as desired.

A standard two-dimensional equalizer, such as is used in QAM systems, can adapt $r_1(t)$ and $r_2(t)$ so that these conditions are met. As in a passband QAM system, the

sampling rate must be at least twice the highest frequency in the line signal spectrum. Unlike the QAM equalizer, the error signal used for adaptation is not remodulated, since no modulation is present.

All techniques used in QAM systems may be directly applied in CAP systems. These include constellation shaping, trellis coding in various dimensions, decision feedback equalization, Tomlinson filtering, and sequence detection.

For channels that introduce frequency offset or phase noise, due to effects such as Doppler shift and additional stages of modulation and demodulation, some subsystem similar to carrier recovery is required at the receiver. The simplification in CAP over QAM is therefore no longer present. For this reason CAP has been applied primarily over channels in which frequency offset and phase noise are not present, in particular the wire-pair channel. CAP has been widely applied to various digital subscriber line (DSL) and local-area network systems.

BIOGRAPHY

Burton R. Saltzberg received a Sc.D. degree from New York University in 1964. He is a consultant in digital communications with several companies, and presents short courses to international audiences. He was with Bell Laboratories from 1957 through early 1996. His most recent position there, which he held for several years, was technical manager of the Data Theory Group. For most of his career, Dr. Saltzberg was engaged in research in communication theory and in analysis and initiation of new data communications offerings over a wide variety of channels. He has published extensively in this field, and was issued 29 U.S. patents. He is a fellow of the IEEE and received the IEEE Communications Society Armstrong Achievement Award in 1991.

BIBLIOGRAPHY

1. T. Starr, J. M. Cioffi, and P. Silverman, *Understanding Digital Subscriber Line Technology*, Prentice-Hall, Upper Saddle River, NJ, 1999.
2. A. K. Aman, R. L. Cupo, and N. A. Zervos, Combined trellis coding and DFE through Tomlinson precoding, *IEEE J. Select. Areas Commun.* **9**: 876–884 (1991).
3. G. H. Im et al., 51.84 Mb/s 16-CAP ATM LAN standard, *IEEE J. Select. Areas Commun.* **13**: 620–633 (1995).
4. G. H. Im and J. J. Werner, Bandwidth-efficient digital transmission over unshielded twisted-pair wiring, *IEEE J. Select. Areas Commun.* **13**: 1643–1655 (1995).

CARRIER-SENSE MULTIPLE ACCESS (CSMA) PROTOCOLS

LEONIDAS GEORGIADIS
Aristotle University of Thessaloniki
Thessaloniki, Greece

1. INTRODUCTION

Communication of information between two or more parties takes place over a variety of physical media called *channels*. Such channels can be simple twisted pair cables, coaxial and optical cables, or free space. Sometimes the channel is dedicated to a specific transmitter–receiver pair. This may be the case when two parties establish a telephone conversation over a dedicated cable. Channels of this type are called *point-to-point*. There are situations, however, where multiple users need to have access to the same channel. The most familiar one is human speech communication. When a number of humans are located in the same room, they all use the same channel, the atmosphere, for their conversation exchange. Computer local area networks (LANs) is another example: a common approach in this case is to attach a number of computers to the same cable as in Fig. 1. Hence, each computer can listen to the transmission of every other computer attached to the same cable. For a third example, consider Satellite communication. As shown in Fig. 2, a number of terminals need to communicate between each other but because of physical obstacles they cannot all listen to each other directly. Instead, each terminal first sends the information to the satellite through the upstream channel. The satellite listens to the upstream channel, receives the information, and then retransmits to the downstream channel, to which all terminal can listen. Hence the upstream channel needs to be accessed by all terminals. Channels of this type are called *multiple-access*.

If the terminals in a multiple-access channel are left unchecked so that they can transmit information whenever they need to do so, then it may be possible for two or more terminals to attempt to use the channel at the same time. In such a situation, the concurrently transmitted messages interfere with each other and generally cannot be received correctly by the intended receivers. Hence, a fundamental issue in multiple-access channel communication is how to coordinate the transmitting terminals in order to avoid or recover from the interference that may result because of concurrent transmissions. The mechanisms by which this is achieved are termed *multiple-access protocols*.

The simplest way to address the coordination problem in multiple-access communication is to avoid concurrent transmissions altogether. To be more specific, we must make certain assumptions about the manner in which transmission of information takes place. First, as is very common today, we assume that all information, whether sound, picture, or text, is transformed to a sequence of bits, 0 or 1, and that each terminal needs to transmit this sequence of bits to the receiving terminal — the receiver knows how to recover the original information from the received sequence of bits. Let us assume further that the

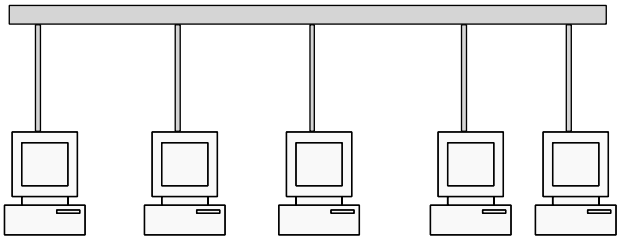


Figure 1. A local area network.

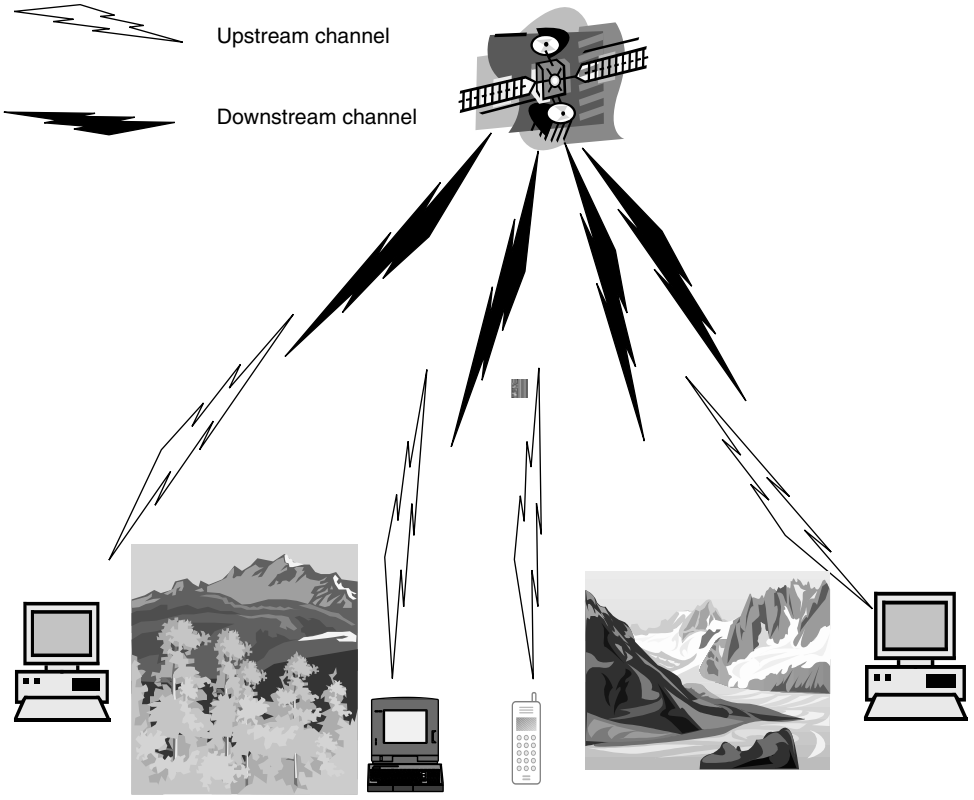
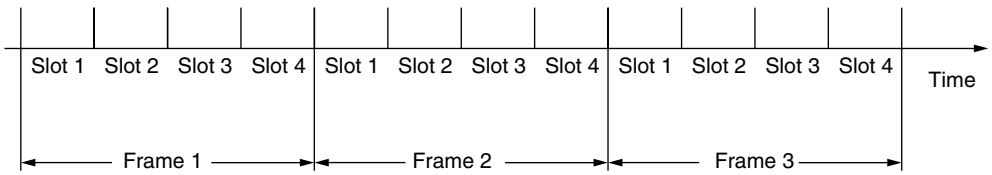


Figure 2. Satellite communications.

sequence of bits is subdivided into groups called *packets*, and that the transmitter needs to transmit one packet at a time to the receiver. All packets contain the same number of bits B. If bits can be transmitted over the channel at a rate of C bits per second (bits/s or bps), then each packet takes $T = B/C$ seconds to be transmitted. We refer to T as the "length" of the packet.

We are now ready to describe the protocol by which access to the channel is free of concurrent transmissions. We divide time into fixed intervals of length T called time *slots* (see Fig. 3). Hence each slot fits exactly one packet. Let the number of terminals that can potentially use the channel be n. We group the time slots into *frames*, where each frame contains n consecutive time slots. Terminal i is allowed to transmit in the ith time slot of each frame. The protocol just described is called *time-division multiple-access* (TDMA) protocol. Since slots are allocated exclusively to each user, no interference occurs and packets are transmitted successfully. Note that there is still a possibility that the packet may be received in error because of noise that naturally exists on the channel, but this is a lower-level issue that is addressed by methods that are beyond the scope of the current discussion.

The TDMA protocol in effect divides the channel into n point-to-point channels. While simple, the protocol has some serious disadvantages. First, if a terminal does not have packets to transmit, the slots allocated exclusively to it cannot be used by anybody else, even if other terminals have a large number of packets to transmit and could use these slots. The second disadvantage is related to packet delays. Since the time interval between two successive slots during which terminal i can transmit is n time slots, a packet generated randomly at a terminal will take on the average $n/2$ time slots to be transmitted, a delay that can be very large if the number of terminals in the system



The channel is accessed by $n = 4$ terminals. Terminal i may transmit in slot i of each frame.

Figure 3. The TDMA protocol.

is large. This will happen regardless of whether the rest of the terminals have packets to transmit.

The disadvantages of TDMA are due to the fact that a terminal can transmit only during the slots allocated to it, even if other terminals are inactive. What if we dispense with the idea of allocating slots exclusively to transmitters? In fact, what if we take the exact opposite approach and allow a terminal to transmit in any slot when it has packets to transmit? In this case, if no other terminal has packets to transmit, then the given terminal can transmit a large number of packets with very low delay. However, if more than one terminals wish to send packets in the same slot, then a "collision" will occur and no message will be received correctly. In this case one must specify how the terminals will react and cooperate in order to make sure that the packets are eventually delivered to their intended destinations. The simplest idea is to instruct the terminals to retransmit their collided packets. However, if two or more terminals pick again and again the same slot for retransmission, the packets will continue to collide and will never be transmitted successfully. There are various methods to avoid this situation. We will concentrate on the most prevalent method encountered in practice: randomized retransmissions. If collisions occur, then the terminals whose packets collided pick randomly some future time slot for retransmission. Hence, while collision may again occur, it is hoped that eventually the transmitting terminals will each pick different slots for transmission and thus their packets will be delivered successfully to their intended destination.

The algorithm just described comes by the name ALOHA protocol and will be described in more detail in Section 2. This algorithm constitutes the basis for the development of carrier-sense multiple-access (CSMA) protocol, which takes advantage of certain channel features and transmitter capabilities in order to provide improved performance. The CSMA protocol will be described in Section 3.

2. THE ALOHA PROTOCOL

The ALOHA protocol was designed by Abramson [1] to provide radio communication between several terminals scattered at various places over the islands of Hawaii. The terminals were sending their data packets to a central station over a common channel (the upstream channel). The central station was then retransmitting the packets to another channel (the downstream channel) that could be listened to by all the terminals. The situation is similar to the one described in Fig. 2. Collisions could occur at the upstream channel if two or more terminals were attempting to transmit their packets. If this happened, the central computer was informing all the terminals that a collision had occurred.

There are two versions of the ALOHA protocol: slotted and unslotted. Slotted ALOHA requires time to be divided in time slots and terminals to transmit their packets at the beginning of each slot. Unslotted ALOHA permits the stations to transmit their packets at any time. The retransmission policy in case of collision is essentially the same for both protocols. In the next two sections we examine these two variants of the ALOHA protocol. Unslotted ALOHA was the precursor of slotted, but it will be more instructive and simpler to concentrate on the slotted ALOHA first.

2.1. Slotted ALOHA

Let us provide a model for this protocol. As in Section 1, the channel is divided into time slots. Terminals are synchronized to transmit their packets at the beginning of a time slot. At the end of each time slot, terminals that transmitted their packets during that slot are informed whether there was a successful transmission or a collision in the slot. If the packet that a terminal transmitted collides with another packet, then the terminal attempts a retransmission in the next slot with probability p and defers for the end of the next slot with probability $1 - p$. In case of deferral, at the end of the next slot the terminal reattempts transmission with probability p and defers with probability $1 - p$. Figure 4 shows an example of the operation of the ALOHA protocol. At slot 1 three terminals attempt to transmit their packets and there is a collision. Hence all three terminals will attempt to retransmit their packets. No terminal chooses to retransmit at slot 2, which is thus idle. Terminals a and b attempt to retransmit at slot 3, and hence there is again a collision. Terminal b is the only one attempting retransmission at slot 4 and its transmission is successful. The transmissions from terminals a and c collide again in slot 6, but they eventually pick different slots for retransmission and their packets are transmitted successfully in slots 8 and 9. Note that other terminals may become active (i.e., they may generate a new packet for transmission) while the retransmission process takes place. These terminals may cause additional collisions. For example, if terminal d generates a new packet and attempts to transmit it in slot 8, an additional collision will occur. In the network designed by Abramson, all terminals (not only those that transmitted their packets) can listen to the downstream channel and hence can be

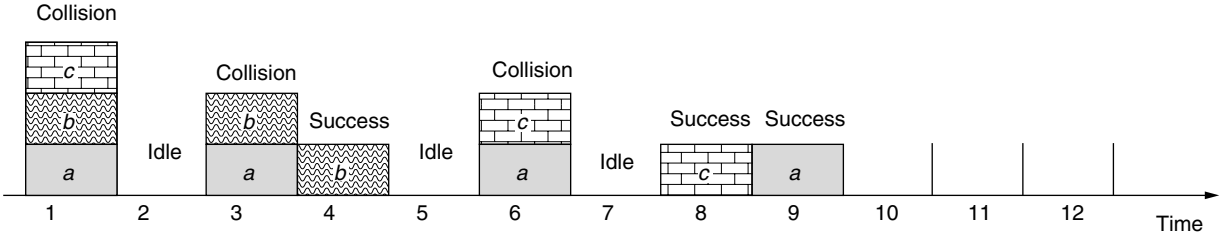


Figure 4. The operation of the ALOHA protocol.

informed about the status of the transmission at the end of the current slot, that is, whether there was no transmission, a successful transmission, or a collision during the slot. However, the ALOHA protocol does not make use of this extra information that a terminal can have.

The protocol just described has the desirable property that packets are not delayed at all if only one terminal needs to transmit at a given time slot. What happens, however, when two or more terminals attempt transmission? As the example in Fig. 4 shows, in this case collisions occur that are followed by retransmission attempts. This results in two inefficiencies; slots may (1) be wasted because of collisions or (2) remain idle even though some terminals have packets to transmit; the latter will happen if all packets that attempt retransmission are deferring in the current slot and no new packets are generated. It is therefore important to know the useful information that can come out of the channel. An appropriate measure for this information is the average number of successfully transmitted packet, S, per slot. We refer to S as the *throughput* of the channel.

Next we provide a method for evaluating S. We need to first make an assumption regarding the statistics of new packets generation process: the number of new packets, K, generated for transmission during a time slot, is a Poisson random variable with rate λ packets/slot. That is, the probability that $K = k$ is given by

$$\Pr(K = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$

This model of packet generation is called the "infinite population model" because it implies that the number of terminals in the system is potentially infinite (the probability that K is any large number is nonzero) and that each terminal generates packets infrequently, so that packet queues are not formed at the terminals. It is used because it is simple, a good approximation when the number of terminals is large, and provides some important insights.

Two sets of terminals may attempt transmission at the beginning of a time slot: (1) those that generate new packets and (2) those whose generated packets have collided in some previous slot and attempt retransmission. In case 2 we say that the packets are "backlogged." Assume that the system can reach steady state and let M be the random number of packets (newly generated and backlogged) transmitted in a given slot in steady state. Denote by G packets/slot the average value of M. Since M includes both newly generated and backlogged packets, it follows that $G > \lambda$. Observe that a successful transmission occurs only when $M = 1$. Indeed, if $M \neq 1$, then either the slot will be idle (if $M = 0$) or there will be a collision in the slot (if $M \geq 2$). Therefore, by the definition of S we have

$$S = 0\Pr(M \neq 0) + 1\Pr(M = 1) = \Pr(M = 1)$$

Hence, if we knew the statistics of M, then we would be able to evaluate S. The exact evaluation of the statistics of M is complicated. To simplify the situation, we make the additional assumption that M is a Poisson random variable. Since the rate of M is G, we have from (1)

$$S = \Pr(M = 1) = e^{-G}G \tag{2}$$

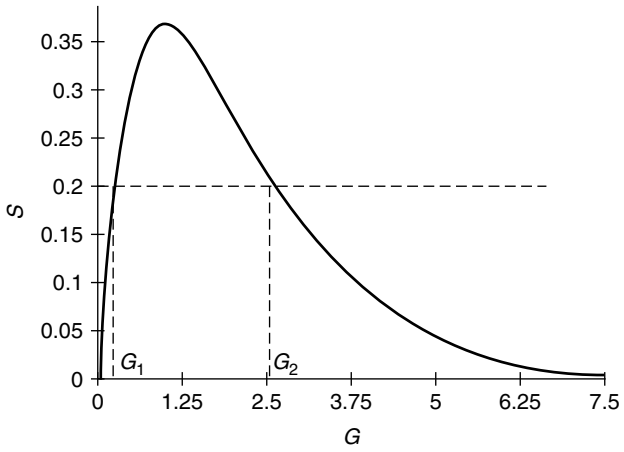


Figure 5. The throughput of the ALOHA protocol.

In Fig. 5 we plot S as a function of G given by Eq. (2). It can be shown that the maximum value of S is $1/e \approx 0.368$ and is obtained at $G = 1$. Hence the maximum channel throughput of the slotted ALOHA protocol is 0.368 packets/slot. A conspicuous feature of the plot in Fig. 5 is that a given channel throughput is achieved for *two* values of G: a small, G_1 and a large G_2. The small value implies that the number of backlogged packets is small while for the large value this number is large. Clearly we would prefer to operate the system at the value G_1, but why do two values appear and what is their meaning?

There are two flaws with the analysis presented above: (1) the existence of steady state is assumed and (2) the probability distribution of M is assumed to be Poisson. For the infinite Poisson model, both these assumptions turn out to be invalid! However, the derived bound on the achievable throughput is still correct. A more detailed analysis of the system for a finite number of users, which is beyond the scope of this presentation, reveals that indeed the throughput of the system is at most $1/e$. Moreover it can be shown that the system behaves qualitatively as follows. There are long periods of time during which the number of backlogged packets in the system remains small and the system operates well, inducing small packet delays. However, from time to time a large increase in the number of backlogged packets in the system will occur and system performance in terms of throughput and delay will degrade. Fortunately, it can also be shown that the time interval for the transition from the "good" state to the "bad" state is generally very large. Hence this instability phenomenon of transiting from good to bad states is seldom a severe problem in real systems.

2.2. Unslotted ALOHA

In the previous section we assumed that the terminals are all synchronized to begin transmission of their packets at the beginning of each slot. If this feature is unavailable, the protocol can be easily modified to still operate. Indeed, the users can be allowed to transmit their new packets at packet generation time. If a collision occurs, then the terminal attempts a retransmission at a later randomly chosen time.

Let us evaluate the performance of the unslotted ALOHA system. We adopt the infinite population model and the notation of Section 2.1. Taking into account the cautionary statements at the end of Section 2.1, let us assume the existence of steady state and that $M(\gamma)$, the number of terminals that attempt transmission in any time interval of length γT, is a Poisson random variable with rate γG. If terminal a begins transmission at time t (see Fig. 6), its transmission will be successful if no other packet begins transmission in the interval $[t-T, t+T]$. Since this interval has length $2T$, the probability that no packet (other than terminal a's packet) is transmitted in the interval $[t-T, t+T]$ is $P_s = P(M(2)=0) = e^{-2G}$. We can interpret P_s as the proportion of attempted packet transmissions that are successful. Now, the rate (average number of packets per time T) by which packet transmissions are attempted is G, and a proportion P_s of these transmissions are successful. Hence the rate of successful transmissions is

$$S = GP_s = Ge^{-2G} \tag{3}$$

where it can be seen that the maximum throughput is $1/(2e)$ and is obtained for $G = \frac{1}{2}$.

We see that the throughput of the unslotted ALOHA is half the throughput of the slotted one. However, unslotted ALOHA does not require terminal synchronization. In any case, from the previous discussion we see that the throughput of both systems is much lower than one. Throughput 1 could be achieved if the terminals could be scheduled for transmission so that collisions are avoided. On the other hand, we have seen that the ALOHA protocol is very simple and distributed in the sense that the terminals operate independently of each other and require very small amount of feedback information to make their decisions. Moreover, the protocol induces very small packet delays when the system is lightly loaded. The question arises as to whether the throughput of the ALOHA protocol can be improved while maintaining its desirable features. These considerations lead to the development of CSMA protocols, which we discuss in the next section.

3. CSMA PROTOCOLS

In this section we present the versions of CSMA protocols that have found wide application. One can think of the CSMA protocol as an evolution of ALOHA where certain terminal capabilities are exploited in order to attain improved performance. It turns out that in real systems the required terminal capabilities depend on the transmission media, that is, whether communication takes place over wires — twisted pair, coaxial, or optical — or through radiowaves in the atmosphere — wireless communication. Accordingly, we first discuss the CSMA and CSMA/CD protocols that are appropriate for wired communications and next examine the CSMA/CA protocol, which is designed for wireless communications.

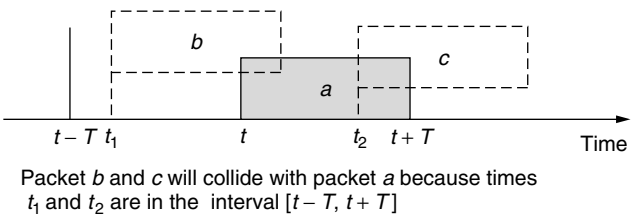


Figure 6. Possibility of collisions in the unslotted ALOHA protocol.

3.1. The CSMA and CSMA/CD Protocols

As we saw in the previous sections, the throughput loss of the ALOHA protocol is due to the fact that slots are wasted due to collisions or remain idle while there are terminals having packets ready for transmission. Let us see whether we can improve this situation while maintaining the desirable features of the ALOHA system. The throughput of the system can be improved if

1. The likelihood of a collision is reduced.
2. The time wasted transmitting garbled data when a collision occurs is reduced.

Consider the possibility of reducing collisions first. Let us assume that a terminal is able to listen to the channel and detect possible ongoing transmissions — busy channel. The ALOHA protocol can then be modified as follows. In case the terminal finds the channel busy, it defers transmission for a random time, or else it transmits its own packet. The protocol just described is called *carrier-sense multiple-access protocol*. The term "carrier sense" signifies the capability of the terminal to listen to the channel and ascertain whether it is busy.

At first sight it seems that with CSMA we succeed in avoiding collisions altogether. Indeed, if all terminals transmit their packets only when the channel is not busy and pick a random retransmission time if they find the channel busy, then it seems that a collision will occur only when two or more terminals begin transmission simultaneously, an event that is quite unlikely. However, the situation is not as rosy as it seems, due to the finite time it takes for a signal to propagate from one terminal to another. Consider the example in Fig. 7. Assume that it

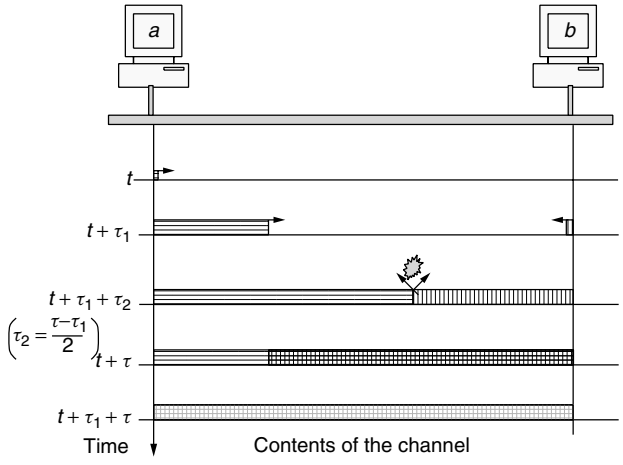


Figure 7. Collision occurrence in CSMA protocol.

takes τ seconds for a signal to be transferred from terminal a to b and vice versa. At time t terminal a senses that the channel is free and starts transmitting a packet. At time $\tau_1 < \tau$ terminal b senses the channel and finds it also free, although the packet from terminal a is well on its way on the channel. Terminal b starts transmitting its own packet, and $(\tau - \tau_1)/2$ seconds later the two packets begin to collide.

From the previous discussion we see that collisions will still occur with the CSMA protocol. However, we expect that the likelihood of a collision will indeed be reduced if the maximum signal propagation delay between two terminals in the system is small relative to the length of a packet. Indeed, this is the case. It can be shown that the throughput of the CSMA protocol is approximately, for small τ/T:

$$S_{\text{CSMA}} \approx \frac{1}{1 + 2(\tau/T)^{1/2}} \tag{4}$$

When $\tau \ll T$, the previous formula shows that S approaches one successful packet per packet duration time, that is, the maximum possible.

Let us examine Eq. (4) more closely. If the length of the packet is B bits and the transmission rate at the channel is C bps, then $T = B/C$. Therefore, we can rewrite (4) as

$$S_{\text{CSMA}} \approx \frac{1}{1 + 2(\tau C/B)^{1/2}} \tag{5}$$

The channel propagation time, τ, is independent of C and B. Therefore, if the network is extended to cover a wider area and as a result τ increases, then the throughput will be reduced. Assume next that we upgrade the channel to a higher transmission rate while maintaining the same arrangement of terminals (i.e., keep τ the same). What will happen to the channel throughput? We need to be careful here since throughput has been defined as the average number of successful packet transmissions per packet length T, and T changes as C varies and B remains constant. An appropriate measure in this case is the average number of successfully transmitted bits per second. This latter measure S^U_{CSMA} is simply related to S_{CSMA}, namely

$$S^U_{\text{CSMA}}(\text{bps}) = \frac{SB}{T} = S_{\text{CSMA}}C \approx \frac{C^{1/2}}{1/C^{1/2} + 2(\tau/B)^{1/2}} \tag{6}$$

where we see that the channel throughput in bits per second increases with C; however, the increase is proportional to $C^{1/2}$ and not C. In fact, *the throughput per channel transmission rate*, S^U_{CSMA}/C, is equal to S_{CSMA}, which decreases as C increases. Also, as seen from (6), for constant C, S^U_{CSMA} increases as the packet length B increases. These considerations should be taken into account when deploying networks operating with the CSMA protocol.

We now turn our attention to the possibility of reducing the time wasted to collisions. Assume that a terminal is able to continue listening to the channel while it transmits its own packet. In case it detects that collision occurred, it interrupts its own transmission and attempts retransmission at a later time. Hence, in general, if a collision occurs, a time interval smaller that the packet duration time will be wasted. In the example of Fig. 7, terminals b and a will detect the collision at times $t + \tau$ and $t + \tau_1 + \tau$, respectively. The CSMA protocol where nodes are interrupting their transmissions when a collision is detected comes by the term *CSMA/CD protocol* (where CD stands for collision detection). The throughput of the CSMA/CD protocol for τ/T small is given approximately by

$$S_{\text{CSMA/CD}} \approx \frac{1}{1 + 5(\tau/T)} \tag{7}$$

Figure 8 shows the throughput of the CSMA and CSMA/CD protocols for various values of $\beta = \tau/T$. We see that both protocols can achieve much higher throughput than the original ALOHA system when β is small. In fact, the throughput can be close to 1. We also see that for the same β, CSMA/CD can achieve significantly better throughput that CSMA. This improvement if due, of course, to the fact that less time is wasted in collisions in CSMA/CD systems than in CSMA.

Up to now we have specified that in case a terminal encounters a collision, it attempts a retransmission at some later random time. What is a good method of selecting such a random time? We discuss here one method that has found wide application. Intuitively, the random retransmission time, R, should depend on the number of backlogged users — the larger the number of backlogged users, the more spread out the distribution of R should be so that the likelihood of avoiding new collisions is reduced. Of course, R should not be too spread out because then terminals will attempt retransmissions rarely and a large portion of time will be left unused. In fact this intuition is correct and can be shown that if the number of backlogged terminals is known and the choice of R is based on this number, the instabilities of the CSMA protocol can be eliminated. However, in real systems the number of backlogged users is seldom known. As an alternative, a terminal may try to obtain an estimate of the number of backlogged users based on its retransmission history. This estimate should increase as the number of collisions encountered during the attempt to transmit a packet increases. Hence the distribution of R should become more spread out as the number of such collisions increases.

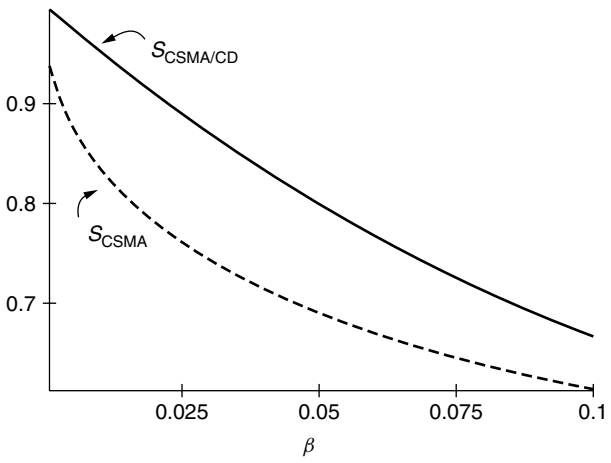


Figure 8. Comparison of CSMA and CSMA/CD protocols.

The previous discussion justifies the following retransmission strategy. If a terminal encounters k collisions during the attempt to transmit a packet, then it attempts a retransmission at time R that is uniformly distributed in the time interval $(0, A2^k)$, where A is a constant. There are various variants of this strategy; however, the main characteristic of all of them is that the "spreading" of R increases exponentially with k. For this reason, this retransmission strategy is known as *exponential backoff*.

3.1.1. Applications of CSMA/CD Protocol. The foremost application of the CSMA protocol is in the technology that connects computer terminals located within a company, an institution, university campus, or other facility using wires. Such a technology is known as *local-area network* (LAN) technology. Several LAN technologies have appeared, but the first and by far the most prevalent one is the Ethernet technology, also known as the IEEE 802.3 LAN technology.

The Ethernet technology was developed in the mid-1970s by Bob Metcalfe and David Boggs. Since then, although it faced challenges by several alternative LAN technologies (Token Ring, FDDI, ATM), it still dominates the marketplace. One of the reasons for this success is that the hardware required for its deployment became very cheap, which, in turn, is due to the large production volume and to the simplicity of the multiple-access protocol used for communication, which is the CSMA/CD protocol with exponential backoff. Moreover, the Ethernet technology proved capable of adapting itself to user demands for increased transmission rates. Currently, Ethernet LANs run at speeds of 10 Mbps, 100 Mbps, and even 1 Gbps.

3.2. The CSMA/CA Protocol

The distributed nature of the CSMA protocol and the low delays it induces when the number of active terminals is small make it a very attractive candidate for wireless communication. However, certain restrictions in such an environment do not permit the direct implementation of the protocol.

Let us recall that in order to be able to implement the CSMA/CD protocol, each terminal needs to be able to perform the following functions:

1. The terminal must be able to listen to the channel and hear whether one or more of the other terminals in the channel are attempting a transmission–carrier sensing capability.
2. The terminal must be able to listen to the channel while transmitting and detect whether its transmission collided with the transmission of some other terminals — collision detection capability.

The collision detection capability implies that a terminal must be able to transmit and receive at the same time, which in a wireless environment can be expensive and is often avoided. Hence, the transmitting terminal may not able to even ensure the correct delivery of its packet. Moreover, as we will see below, even if the collision detection capability exists, it is still possible that a transmitting station does not detect a collision while it is transmitting a packet, but the transmission collides at the receiver. This lack of collision detection capability can be remedied by having the receiver inform the transmitter that the transmitted packet has been correctly received. To do this, the receiving terminal, on correct reception of a packet, sends a short acknowledgment packet back to the transmitter. This packet is referred to as the *ACK message*.

Regarding the carrier-sensing capability of the terminals, while possible, it is not always sufficient to ensure with high probability that the channel is free of transmissions. To understand this problem, we must expand on the special restrictions imposed in a wireless environment. A characteristic of wireless transmission is that terminal a can deliver reliably information to b only if b is within a specified distance from a. Now consider the situation in Fig. 9, where we assume that transmissions are symmetric in the sense that if terminal a can deliver information to b, then b can deliver information to a. The transmission from terminal a can reach b but not c. The transmission from c can reach b but not a. Using the standard CSMA protocol in this environment, certain collisions can still be avoided by sensing the channel. For example, if b is transmitting to a, c can sense the ongoing transmission. However, assume that while a transmits to b, c receives a packet for transmission. If c listens to the channel, it will not hear a's transmission and therefore, if the standard CSMA protocol is employed, a collision will occur. This problem is known as the "hidden terminal" problem. Note that in this case, even if a is able to detect collisions, it will not be able to realize that a collision occurred since it cannot hear c's transmission — as we saw, the latter problem is remedied by the use of the ACK message. Because of the retransmission policy of the basic CSMA protocol, the system can still operate in this environment in spite of the increased number of collisions; however, system throughput may decrease dramatically if packet sizes are large. In fact, plain carrier sensing is not always desirable in this environment. To clarify this point, consider again the situation in Fig. 9. Suppose that b is sending data to a and c wishes to send data to terminal d. If c senses the channel, it will find it busy and therefore will defer transmission. However, since c's transmission cannot reach a, c could in fact deliver its packet to d without colliding with b's

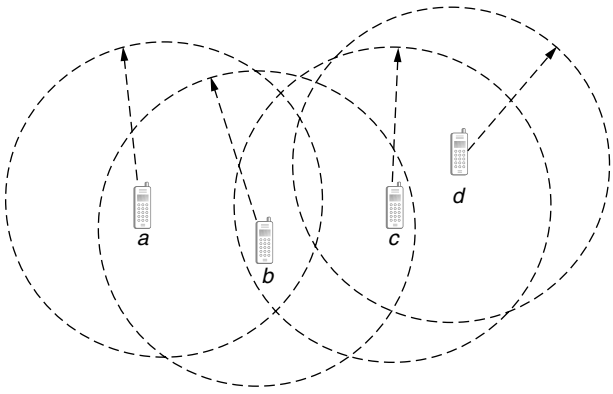


Figure 9. The hidden- and exposed-terminal problems.

transmission. As a result, plain carrier sensing in this case results in reduced utilization of the system. This problem is known as the "exposed terminal" problem.

We next provide a mechanism to address the above-mentioned problems. Two control signals are introduced. These control signals are short messages (compared to packet sizes) that are exchanged between the transmitter and the receiver before the initiation of packet transmission. The first control signal is sent by the transmitter to the receiver and indicates that the transmitter is "*r*equesting *t*o *s*end (RTS)" a packet. The receiver, on correct reception of the RTS, replies that "it is *c*lear *t*o *s*end (CTS)" the packet. Both RTS and CTS signals include a field indicating how long the packet transmission and the accompanied ACK message will last. The terminals now act as follows:

- If a terminal listens to a CTS signal, it waits until the end of the ongoing transmission; this is known since it is included in the CTS signal. It then waits for a random amount of time and attempts to initiate its own transmission process.

Let us see how this rule resolves the hidden-terminal problem. Assume for the moment that the transmission of the CTS and RTS signals is instantaneous, and let us return to the situation in Fig. 9, where a needs to transmit a packet to b. Terminal a sends an RTS to b, and b replies with a CTS signal. Terminal c receives the CTS signal and knows that a transmission has been initiated, so it defers its own transmission. Hence the hidden-terminal problems is alleviated. In effect, the exchange of CTS and RTS messages act as a virtual carrier sensing mechanism.

In fact, the RTS and CTS signals can also be used to address the exposed-terminal problem. Assume that we add the following rule.

- If a terminal listens to an RTS signal but not a CTS signal, it goes ahead with its own transmission, if any.

In Fig. 9 assume that b sends an RTS to a and a replies with a CTS. Terminal c hears the RTS from b but not the CTS from a, and so it knows that its own transmission will not interfere with the b-to-a transmission. Hence it can start its own transmission at any time. Therefore the exposed-terminal problem is avoided.

We assumed above that CTS and RTS signals are instantaneous. Of course, as described in Section 3.1, in a real system transmissions do not take place instantaneously and therefore one cannot assume that the RTS and CTS signals will be received correctly always and free of collisions. However, by now we know that by imposing appropriate retransmission rules the system can deal with occasional loss of RTS or CTS signals. RTS and CTS signals are useful if packet sizes are large. For small packet sizes, it is preferable to go ahead with the packet transmission rather than incurring the overhead of RTS–CTS message exchange.

The modified CSMA system whose principles of operation were described above, comes by the name CSMA/CA, where CA stands for *collision avoidance*. The acronym signifies that collisions are sought to be avoided, but, as we saw, they are not avoided altogether. Because of the retransmission policy of the CSMA system, collisions that may occur are not detrimental; in case of collision, the ACK message or RTS CTS messages will not be received and the transmitting terminal will defer its transmission for a later time. However, if the propagation delays are relatively large and the system is heavily loaded, collisions may degrade the performance of the system.

3.2.1. Applications of CSMA/CA Protocol. The principles of the CSMA/CA protocol have been applied to the specification of the MAC protocol for wireless local-area networks (WLANs), known as the IEEE 802.11 standard.[1] Originally the transmission rates of IEEE 802.11 were 1 and 2 Mbps. The IEEE 802.11b extension to this standard specified 5.5 and 11 Mbps transmission rates, while there is ongoing work that will increase the rate to 20 Mbps. There is currently a great interest in the development of WLAN technologies that support not only high data rates but also multimedia communication such as video, audio, and videoconference communication. The support for multimedia communication imposes additional requirements to the network, such as low packet delays and low packet loss. Networks that are able to provide such support are said to provide quality of service (QoS). CSMA networks were not designed originally to provide QoS. There is a large amount of ongoing works that either attempt to adapt the CSMA protocol to these additional requirements or investigate the feasibility of other approaches.

4. TO PROBE FURTHER

The literature on the ALOHA and the various variants of the CSMA protocols is huge and is still expanding. We do not attempt to provide a detailed account of all the works that contributed to the development of these protocols. Instead we provide some key references to which the interested reader may turn either for a more in-depth study or for a more comprehensive account of related work.

The book by Rom and Sidi [2] provides an in-depth analysis of the ALOHA, CSMA, CSMA/CD, and various other multiple-access protocols. A nice and detailed exposition of the subject can also be found in the book by Bertsekas and Gallager [3]. Very readable accounts of the protocols can be found in the books by Tanenbaum [4] and Kurose and Ross [5]. Information on the IEEE 802.3 and IEEE 802.11 standards and related activities can be found in their Website [6].

BIOGRAPHY

Leonidas Georgiadis received the Diploma degree in electrical engineering from Aristotle University, Thessaloniki, Greece, in 1979, and his M.S. and Ph.D. degrees both in electrical engineering from the University of

[1] Currently the standard does not incorporate a mechanism for dealing with the exposed terminal problem.

Connecticut, in 1981 and 1986, respectively. From 1981 to 1983 he was with the Greek army.

From 1986 to 1987 he was research assistant professor at the University of Virginia, Charlottesville. In 1987, he joined IBM T. J. Watson Research Center, Yorktown Heights, as a research staff member. Since October 1995, he has been with the Telecommunications Department of Aristotle University, Thessaloniki, Greece. His interests are in the area of high-speed networks, scheduling, congestion control, mobile communications, modeling, and performance analysis.

Professor Georgiadis is a senior member of IEEE Communications Society. In 1992, he received the IBM Outstanding Innovation Award for his work on goal-oriented scheduling for multi-class systems.

BIBLIOGRAPHY

1. N. Abramson, The Aloha system — another alternative for computer communications, *Proc. Fall Joint Comput. Conf. AFIPS Conf.*, 1970, p. 37.
2. R. Rom and M. Sidi, *Multiple Access Protocols Performance and Analysis*, Springer-Verlag, 1990.
3. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 2nd ed., 1992.
4. A. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, 1996.
5. J. F. Kurose and K. W. Ross, *Computer Networking, A Top-Down Approach Featuring the Internet*, Addison-Wesley, 2001.
6. *http://standards.ieee.org/getieee802/*.

CDMA/IS95

JHONG SAM LEE
J.S. Lee Associates, Inc.
Rockville, Maryland

LEONARD E. MILLER
Wireless Communications Technologies Group, NIST
Gaithersburg, Maryland

The IS95 cellular telephone standard [1] was designed as a second-generation (digital) system. Like the U.S. analog (first-generation) cellular system, known as the *Advanced Mobile Phone System* (AMPS) [2], the IS95 system uses frequency-division duplexing (FDD), with a 25-MHz bandwidth in each direction over the frequency allocations shown in Fig. 1. The cellular band is further divided equally between two service providers, known as the "A" (wire) and the "B" (nonwire) carriers, as illustrated in Fig. 1. In AMPS, each channel occupies 30 kHz of bandwidth in a frequency-division multiple access (FDMA) system, using analog frequency modulation waveforms. The frequencies that are used in one cell area are reused in another cell area at a distance such that mutual interference gives a carrier-to-interference power ratio of no less than 18 dB. Given this performance requirement and the fact that in the mobile radio environment the attenuation of carrier power usually is proportional to the fourth power of the distance from the emitter to a receiver, the analog cellular system utilizes seven-cell clusters, implying a frequency reuse factor of 7 [2]. The resulting capacity of AMPS is then just one call per 7×30 kHz $= 210$ kHz of spectrum in each cell, and in the total of 12.5 MHz allocated there can be no more than 60 calls per cell.

In 1988, the Cellular Telecommunications Industry Association (CTIA) stipulated requirements for the second-generation digital cellular system technology. The key requirements included a 10-fold increase in call capacity over that of AMPS, a means for call privacy, and compatibility with the existing analog system. The compatibility requirement arose from the fact that the second-generation system must operate in the same band as AMPS.

Proposed in 1989, the first U.S. standard for a second-generation cellular system was published in 1992 as IS54 [3] and adopted a time-division multiple access (TDMA) technology. The IS54 TDMA digital cellular system employs digital voice produced at 10 kbps (8 kbps plus overhead) and transmitted with $\pi/4$ differentially encoded quadrature phase shift keying ($\pi/4$ DQPSK) modulation. Because the IS54 system permits 30 kHz/10 kbps $= 3$ callers per 30-kHz channel spacing, the increase of capacity over AMPS is only a factor of 3 (180 calls per cell).

In 1990, Qualcomm, Inc. proposed a digital cellular telephone system based on code-division multiple access (CDMA) technology [4], which was adopted in 1993 as a second U.S. digital cellular standard, designated IS95 and later known as cdmaOne. Using spread-spectrum (SS) transmission techniques, the IS95 system provides a very high capacity. The full title of the IS95 CDMA standard is *Mobile Station-Base Station Compatibility Standard for*

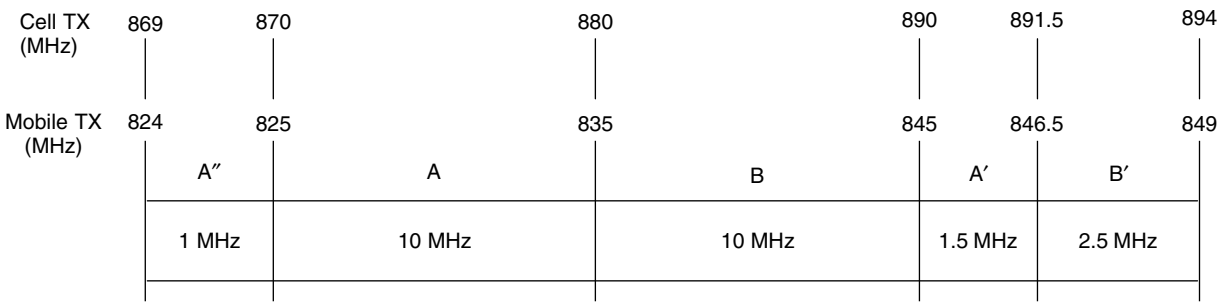


Figure 1. Cellular bands in the United States.

Dual-Mode Wideband Spread Spectrum Cellular System, indicating that the document is a common air interface (CAI) that does not specify the details of how the system is to be implemented.

1. WHAT IS CDMA?

Spread-spectrum techniques involve the transmission of a signal in a bandwidth substantially greater than the information bandwidth to achieve a particular operational advantage. How SS signals can be processed to yield gains for interference rejection can be understood by calculating the jamming margin for a SS system. Let the following parameters be defined:

S = received power for the desired signal in watts
J = received power for undesired signals in watts (jamming, other multiple access users, multipath, etc.)
$R = 1/T_b$ = data rate (data signal bandwidth in Hz)
W = spread bandwidth in Hz
E_b = received energy per bit for the desired signal in W · s (watt-seconds)
N_0 = equivalent noise spectral power density in W/Hz

Then the ratio of the equivalent "noise" power J to S is

$$\frac{J}{S} = \frac{N_0 W}{E_b/T_b} = \frac{WT_b}{E_b/N_0} = \frac{W/R}{E_b/N_0}$$

When E_b/N_0 is set to the value required for acceptable performance of the communications system, then the ratio J/S bears the interpretation of a jamming margin:

$$J/S = \text{tolerable excess of interference over desired signal power} = \frac{W/R}{(E_b/N_0)_{\text{req}}}$$

or

$$\text{Margin (dB)} = \frac{W}{R}(\text{dB}) - \left(\frac{E_b}{N_0}\right)_{\text{req}} (\text{dB})$$

The quantity W/R is called the SS *processing gain* (PG). For example, if the information bandwidth is $R = 9600$ Hz, corresponding to digital voice, the transmission bandwidth is $W = 1.2288$ MHz, and the required SNR is 6 dB, then the PG equals $128 = 21.1$ dB and the jamming margin equals $32 = 15.1$ dB. The communicator can achieve an SNR of at least 6 dB even in the face of interference (jamming) power in excess of 32 times the signal power, due to the PG. In a CDMA communications system, the cochannel communicators, occupying the same bandwidth simultaneously, account for the interference (jamming) power. If every user supplies the identical amount of signal power to the base station antenna through a perfect power control scheme, regardless of location, then for this example 32 other multiple-access users can be accommodated by a CDMA system.

The capacity of a CDMA system is proportional to the PG of the system. This fact may be illustrated as follows, assuming first that the system in question is isolated from all other forms of outside interference (i.e., a single cell): The carrier power equals $C = S = R \times E_b$ and, analogous to the jamming power, the interference power at the base station receiver may be defined as $I = W \times N_0$, where W is the transmission bandwidth and N_0 is the interference power spectral density. Thus a general expression for the carrier-to-interference power ratio for a particular mobile user at the base station is given by

$$\frac{C}{I} = \frac{RE_b}{WN_0} = \frac{E_b/N_0}{W/R}$$

where the PG is W/R. Let M denote the number of mobile users. If power control is used to ensure that every mobile has the same received power at the base station, then the interference power caused by the $M - 1$ interferers is $I = C(M - 1)$. Substituting for I in the previous equation, neglecting thermal noise, and solving for M, the capacity for a CDMA system is found to be

$$M = \frac{W/R}{E_b/N_0} + 1 \approx \frac{W/R}{E_b/N_0}$$

Thus, the capacity of a CDMA system is proportional to the PG. This PG is based on the fact that in the CDMA receiver the (multiple) interfering users' signals continue to have the bandwidth W, while the (single) selected user's signal is despread by the removal of the spreading code. The receiver then can perform filtering at the selected user's despread (baseband) bandwidth, which admits only a small fraction of the energy received from the interfering user signals.

It is possible in a digital telephone system to exploit pauses in speech to reduce the transmission rate or to use intermittent transmissions with an effective duty cycle of 40–50%. If the duty cycle of a speech traffic channels in the CDMA system is denoted by the variable α, then effectively the data rate is αR instead of R.

If the base station employs directional antennas that divide the cell into sectors, each antenna will receive only a fraction of the interference. In practice, the coverage areas of the receiving antennas overlap by approximately 15%. Standard implementations divide the cell into three sectors, providing an effective capacity increase of $G = 3 \times 0.85 = 2.55$.

Signals originating in other cells must be taken into account when determining the capacity of a particular "home" cell; such interference is, of course, diminished by the attenuation incurred by the interferers in propagating to the home cell. Simulations performed by Qualcomm have shown that interference from other cells accounts for only about 35% of that received at the base station [5–7]. On the basis of this information, the equation for CDMA capacity may be modified to include a reuse efficiency F_e to account for other-cell interference.

Taking into account voice duty cycle, antenna gain, and other-cell interference, the equation for CDMA capacity becomes

$$M \approx \frac{W/R}{E_b/N_0} \times \frac{GF_e}{\alpha}$$

For example, using realistic parameters for the IS95 system ($W/R = 128$, $E_b/N_0 = 7$ dB, $\alpha = 0.5$, $G = 2.55$, and $F_e = 0.65$), the capacity of the system is 85 users per cell. The achievement of this capacity over the mobile radio channel in practice depends on many factors [5,8,9].

2. THE IS95 SYSTEM

In the IS95 system, the mobile stations communicate with base stations over "forward" (base-to-mobile) and "reverse" (mobile-to-base) radio links, also sometimes called "downlink" and "uplink," respectively. As suggested in Fig. 2, the radiocommunications over these links are organized into different channels: *pilot, synchronization, paging*, and *traffic* channels for the forward link; and *access* and *traffic* channels for the reverse link.

Unlike an FDMA cellular system, a CDMA cellular system does not require the use of "clusters" of cells to enforce a minimum reuse distance between cells using the same frequency channels in order to control the amount of cochannel interference. Each CDMA cell uses the identical spectrum and employs pseudorandom noise (PN) code SS modulation and utilizes the resulting PG to overcome interference. The PN code signaling rate, known as the "chip rate," was initially chosen by Qualcomm to be 1.2288 megachips per second (Mchips/s), which is an integer multiple (128) of the maximum digital voice bit rate of 9600 bps, thereby requiring a spectrum occupancy of about 1.23 MHz. The particular choice of the chip rate in IS95 was presumably dictated in part by a desire to operate a single CDMA channel in the 1.5-MHz band designated as "A" in Fig. 1; Gilhousen argues that the bandwidth selected is a good choice in terms of the characteristics of the mobile channel (such as coherence bandwidth) and the complexity of a "RAKE" multipath receiver designed for that channel [10].

2.1. System Synchronization

Each base station of the IS95 system is required to maintain a clock that is synchronized to Universal Coordinated Time (UTC), indirectly through synchronization to GPS time signals. The known beginning of the GPS timecount (time 00:00:00 on Jan. 6, 1980) is traceable to UTC and is required to be aligned in a particular fixed way with the PN codes used by every base station in the CDMA system — two "short" PN codes having 26.66-ms periods, and a "long" PN code that has a period that is over 41 days long, all running at the 1.2288-Mchip/s rate. The synchronization of time standards among the CDMA base stations is necessary because each base station transmits on the same center frequency and uses the same two short PN codes to spread its (forward-link) waveform, the different base station signals being distinguished at a mobile receiver only by their unique short PN code starting positions (phase offsets), as illustrated in Fig. 3.

The system time reference for a particular mobile is established when it acquires a certain base station signal, usually that from the nearest base station, affiliates with that base station, and reads the synchronization message broadcast by that base station. The message contains information that enables the mobile unit to synchronize its long PN code and time reference with those of that particular base station.

2.2. Forward-Link Summary

The forward-link channel structure consists of the transmission of up to 64 simultaneous, distinct channels that are orthogonally multiplexed onto the same RF carrier. One of these channels is a pilot signal that is transmitted continuously, to be received by the mobiles as a coherent phase reference. Another of these channels is a continuously transmitted synchronization channel that is used to convey system information to all users in the cell. Up to

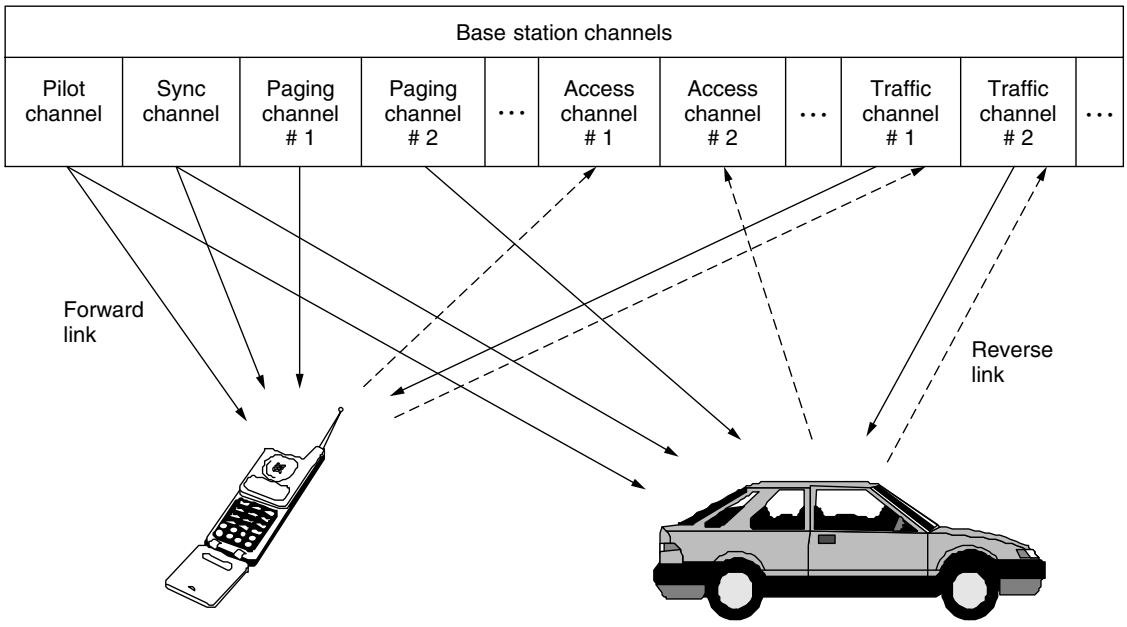


Figure 2. IS95 forward- and reverse-link channels.

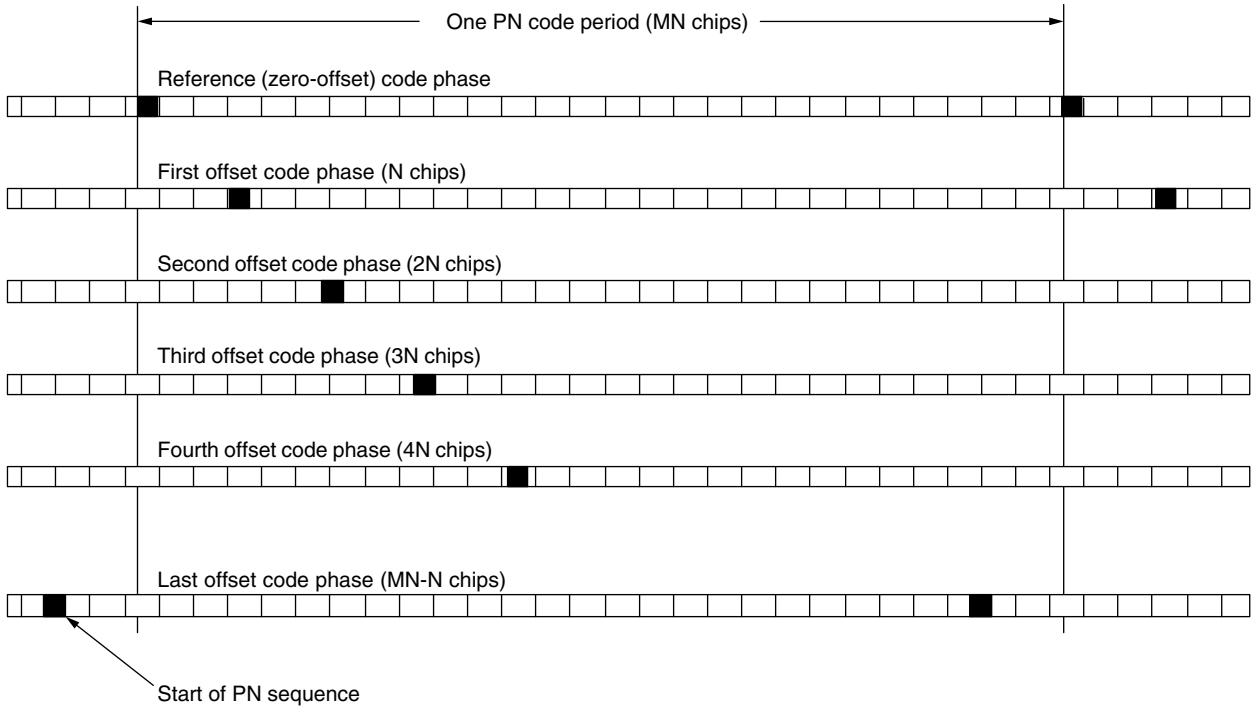


Figure 3. Short PN code offsets are assigned to different base stations.

seven paging channels are used to signal incoming calls to mobiles in the cell and to convey channel assignments and other signaling messages to individual mobiles. The remainder of the channels are designated as traffic channels, each transmitting voice and data to an individual mobile user.

The major features of the IS95 forward-link CAI are as follows:

- *Multiplexing*. The forward-link channelization is based on an orthogonal code-division multiplexing scheme using an orthogonal set of "subcarrier" digital waveforms known as *Walsh functions* [5,11].
- *Interference Rejection*. The forward-link waveform is modulated by direct-sequence PN code SS techniques to isolate the signals received from a particular base station and to discriminate against signals received from other base stations.
- *Modulation*. The forward-link waveform features modulation of I (cosine) and Q (sine) RF carriers by different PN-coded bipolar ($\pm$) baseband digital data streams, thereby generating a form of quaternary phase shift keying (QPSK).
- *Pulseshaping*. The shape of the baseband digital pulses in the I and Q output channels is determined by a finite impulse response (FIR) filter that is designed to control the spectrum of the radiated power for minimal adjacent-frequency interference [5].
- *PN Chip Rate*. The PN code chip rate, which is 1.2288 Mchips/s, is 128 times the maximal source data rate of 9.6 kbps, thereby providing a PG of 21 dB.
- *Effective Bandwidth*. For the PN chip rate and FIR filter spectrum control specified, the energy of the IS95 forward-link signal is almost entirely contained in a 1.25-MHz bandwidth.
- *Voice Coding*. A variable-rate vocoder is specified, with data rates 1200, 2400, 4800, and 9600 bps depending on voice activity. In 1996, a Personal Communications Services (PCS) version of the system specified in IS95 was released [12] with data rates of 14.4 kbps and fractions thereof; this set of rates also became available in the cellular standard revision known as IS95A.
- *Error Control Coding*. The forward link uses rate $\frac{1}{2}$ constraint length 9 convolutional coding, with Viterbi decoding.
- *Interleaving*. To protect against burst error patterns (a distinct possibility on the fading mobile communications channel), the forward link interleaves code symbols before transmission, using a 20-ms span.

The baseband data rate from each channel being multiplexed varies; the highest rate is 19.2 kilosymbols per second (ksps). The polarity of each channel's baseband data symbol multiplies an assigned 64-chip Walsh sequence that is repeated at the 19.2-ksps rate. Thus, the orthogonally multiplexed combination of forward-link channels forms a baseband data stream with a rate of 64×19.2 ksps $= 1.2288$ Mchips/s (see also Table 1). The orthogonal multiplexing operations on the forward link are shown in Figs. 4 and 5. Each channel is modulated by a channel-specific Walsh sequence, denoted $H_i, i = 0, 1, \ldots, 63$. The IS95 standard assigns H_0 for the pilot channel, H_{32} for the synchronization channel, H_1–H_7 for

Table 1. Forward Traffic Channel Modulation Parameters

Parameter	Value				Units
Data rate	9600	4800	2400	1200	bps
PN chip rate	1.2288				Mchips/s
Code rate	$\frac{1}{2}$				Bits/code symbol
Code repetitions	1	2	4	8	Modulation symbol/code symbol
Modulation symbol rate	19,200				sps
Code symbol energy	$E_b/2$	$E_b/4$	$E_b/8$	$E_b/16$	
PN chips/modulation symbol	64				N/A
PN chips/bit	128	256	512	1,024	

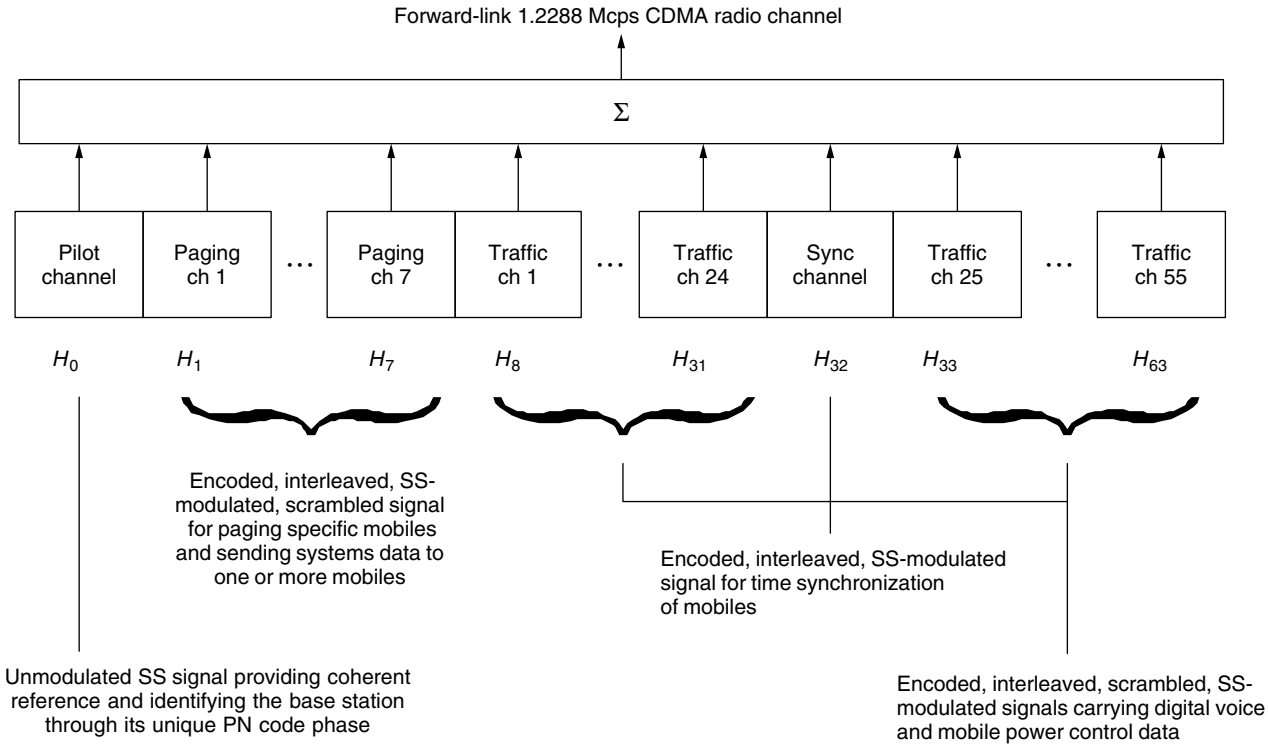


Figure 4. Forward-link channel assignments.

the paging channels, and the remainder of the H_i to the traffic channels. The multiplexed data stream for each channel is combined separately with two different short PN codes that are identified with I- and Q-quadrature carrier components.

The I- and Q-channel PN codes may be denoted by $\mathrm{PN}_I(t, \theta_i)$ and $\mathrm{PN}_Q(t, \theta_i)$, respectively, and are generated by 15-stage linear feedback shift registers (LFSRs). The parameter θ_i denotes the PN code offset phase assigned to a particular base station. Thus, unlike conventional QPSK, which assigns alternate baseband symbols to the I and Q quadratures, the IS95 system assigns the same data to both quadrature channels, each of which pseudorandomly preserves or inverts the data polarity. It is common to speak of these operations as "quadrature spreading." The two short distinct PN codes are maximum-length sequences and are lengthened by the insertion of one chip per period in a specific location in the PN sequence. Thus, these modified short PN codes have periods equal to the normal sequence length of $2^{15} - 1 = 32{,}767$ plus one chip, or 32,768 chips. At a rate of 1.2288 Mchips/s, the I and Q sequences repeat every 26.66 ms, or 75 times every 2 s.

The synchronization channel is demodulated by all mobiles (mobile units) and contains important system information conveyed by the sync (synchronized) channel message, which is broadcast repeatedly. This signal identifies the particular transmitting base station and conveys long PN code synchronization information, at a rate of 1.2 kbps, or 32 sync channel data bits per 26.66-ms sync channel frame and 96 bits per 80-ms sync channel "superframe." The sync channel frame length is equal to one period of the short PN codes, and the sync channel frames are in fact aligned with those periods so that, once the mobile has acquired a particular base station's pilot signal, the mobile automatically knows the basic timing structure of the sync channel.

After synchronization has been accomplished, the mobile can receive the paging channel and transmit on

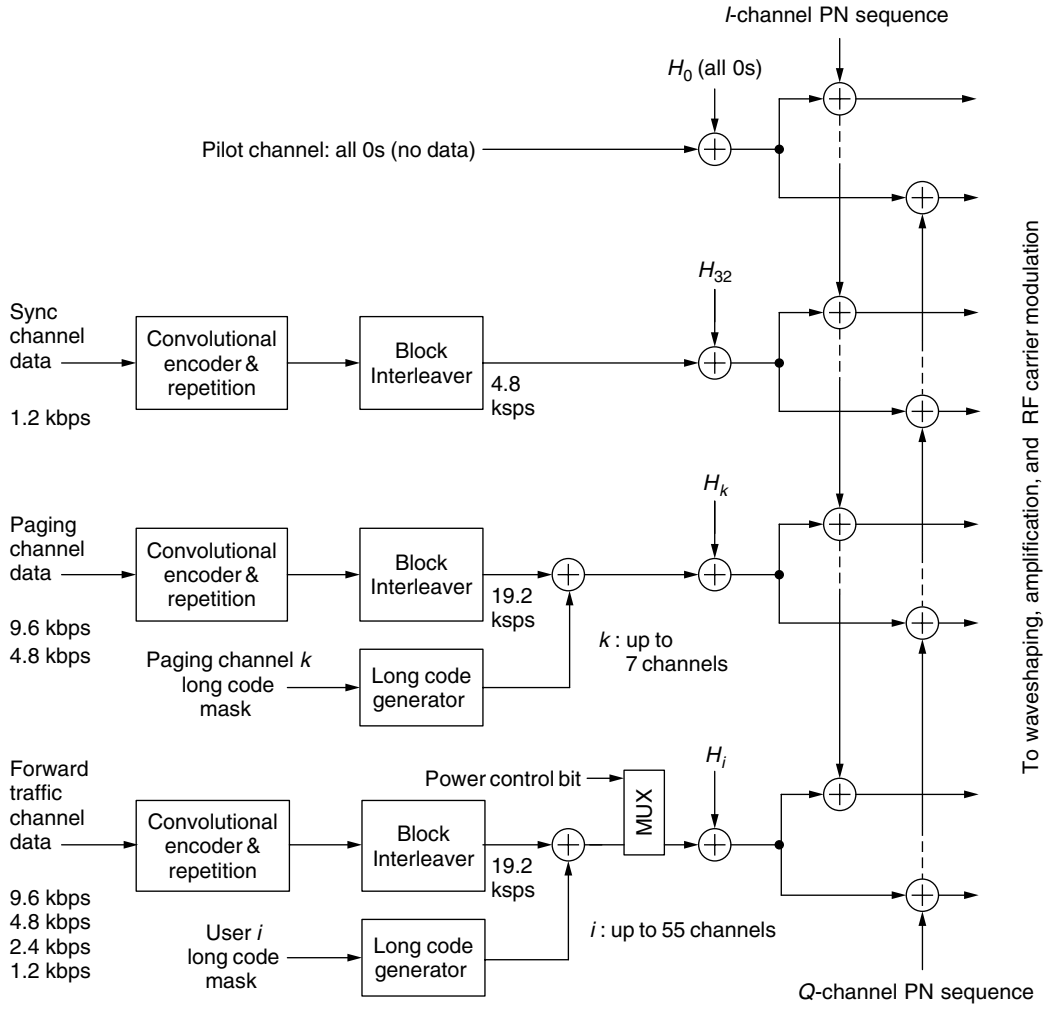


Figure 5. Forward-link multiplexing operations.

the access channel. Because all channels except the pilot and sync channels are scrambled using the long PN code, the sync channel message is necessary to correctly demodulate any other channel. The beginning of the paging and traffic channel frames coincide with the start of a sync channel superframe.

The paging channels are used to alert the mobile to incoming calls, to convey channel assignments. Paging information is generated at either 9.6 or 4.8 kbps. The information is convolutionally encoded, repeated, and interleaved. The repetition of the code symbols is adapted to the data rate to fix the rate of symbols being interleaved at 19.2 ksps. Other than for the sync channel, the data frames for the IS95 channels are 20-ms in length, and the interleaving is performed on a frame-by-frame basis. All the paging channel modulation symbols are transmitted at the same power and baseband data rate for a given CDMA system.

Unlike the sync channel, the encoded and interleaved paging channel symbols are scrambled (multiplied by a random sequence at the same rate) with a 42-stage long-code PN sequence running at 1.2288 Mchips/s that is decimated to a 19.2-ksps rate by sampling every 64th PN code chip. The long PN code is generated by a 42-stage shift register, with a period of $2^{42} - 1 \approx 4.4 \times 10^{12}$ chips (lasting over 41 days at 1.2288 Mchips/s). A phase offset of the original long PN code sequence that is unique to the particular paging channel and base station is obtained by combining the outputs of the shift register stages selected by a 42-bit mask. Details of the use of masks to shift PN codes and the effect of decimation on the codes are given in Ref. 5.

In IS95, each active paging channel has a number of periodically recurring message slots (e.g., 2048 slots) available for transmitting pages and other base-to-mobile messages. When a message is queued up for a particular mobile, the base station, using a hash function, pseudorandomly selects one of the paging channels and pseudorandomly selects one of the message slots in that paging channel for transmission to the particular mobile. The mobile knows exactly which paging channel and message slot to monitor for possible messages because the pseudorandom selection is based on its own identification number and known system parameters. The purpose of the hash function is to distribute the message traffic evenly among the paging channels and message slots. Details of the hash functions used in IS95 are given in Ref. 5.

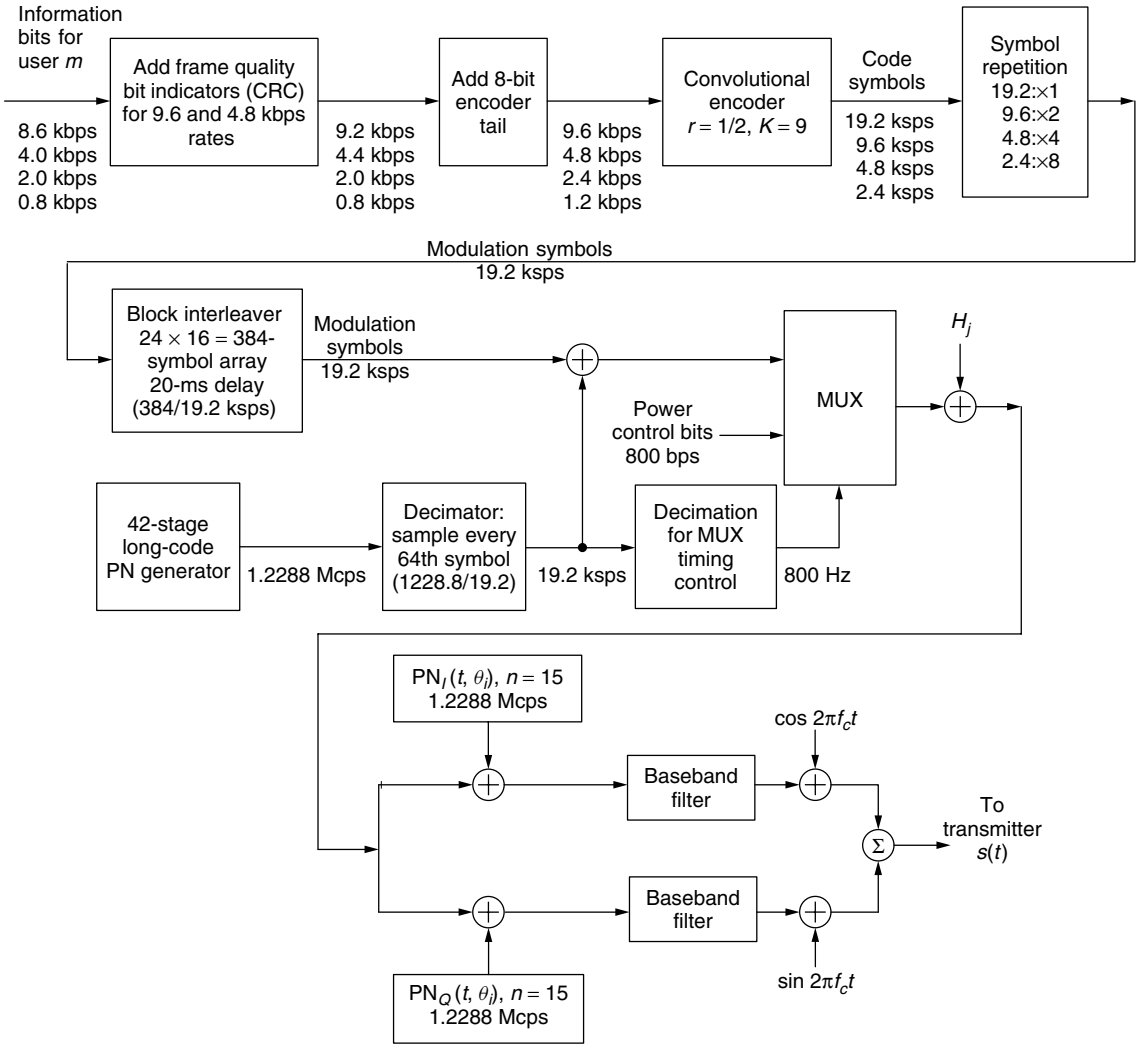


Figure 6. Forward traffic channel modulation.

A block diagram for the forward traffic channel modulation is given in Fig. 6. As shown in the diagram, voice data for the mth user is encoded on a frame-by-frame basis using a variable-rate voice coder, which generates data at 8.6, 4.0, 2.0, or 0.8 kbps depending on voice activity, corresponding respectively to 172, 80, 40, or 16 bits per 20-ms frame. A cyclic redundancy check (CRC) error-detecting code calculation is made at the two highest rates, adding 12 bits per frame for the highest rate and 8 bits per frame at the second highest rate. At the mobile receiver, which one of the possible voice data rates is being received is determined in part from performing similar CRC calculations, which also provide frame error reception statistics for forward-link power control purposes.

In anticipation of convolutional coding on a block basis (code symbols in one frame not affecting those in adjacent frames), a convolutional encoder "tail" of 8 bits is added to each block of data to yield blocks of 192, 96, 48, or 24 bits per frame, corresponding to the data rates of 9.6, 4.8, 2.4, and 1.2 kbps going into the encoder. Convolutional encoding is performed using a rate $\frac{1}{2}$, constraint length 9 code, resulting in coded symbol rates of 19.2, 9.6, 4.8, and 2.4 ksps.

Coded symbols are repeated as necessary to give a constant number of coded symbols per frame, giving a constant symbol data rate of 19.2 ksps (i.e., 19.2 ksps × 1, 9.6 ksps × 2, 4.8 ksps × 4, 1.2 ksps × 8). The 19.2 ksps × 20 ms = 384 symbols within the same 20-ms frame are interleaved to combat burst errors due to fading.

Each traffic channel's encoded voice or data symbols are scrambled to provide voice privacy by a different phase offset of the long PN code, decimated to yield 19.2 kchips/s. The scrambled data are punctured (overwritten) at an average rate of 800 bps by symbols that are used to control the power of the mobile station.

Note from Fig. 6 that, regardless of the data rate, the modulated channel symbol rate must be 19.2 ksps. This is accomplished by means of code symbol repetition for rates less than the 9.6-ksps data rate.

2.3. Reverse-Link Summary

The IS95 reverse link channel structure consists of access channels and traffic channels. To reduce interference and

save mobile power, a pilot channel is not transmitted on the reverse link. A mobile transmits on either an access or a traffic channel but never both at the same time.

The major features of the IS95 reverse link CAI are as follows:

- *Multiple Access*. The reverse-link channelization is based on a conventional SS PN CDMA scheme in which different mobile users are distinguished by distinct phase offsets of the 42-stage-long PN code, which serve as user addresses.
- *Quadrature Spreading*. In addition to the long PN code, the reverse-link datastream is direct-sequence modulated in quadrature by the same two short PN codes as on the forward link; each mobile station in each cell uses the reference or zero-offset phases of these two codes.
- *Modulation*. The reverse link waveform features 64-ary orthogonal modulation using sequences of 64 chips to represent six binary data symbols. The quadrature modulation of I (cosine) and Q (sine) RF carriers by the two different PN-coded bipolar ($\pm$) baseband digital datastreams, with the Q-quadrature stream delayed by half a PN chip, generates a form of offset quaternary phaseshift keying (OQPSK).
- *Pulseshaping*. The shape of the baseband digital pulses in the I and Q output channels is determined by a FIR filter that is designed to control the spectrum of the radiated power for minimal adjacent-frequency interference.
- *PN Chip Rate*. The PN code chip rate, which is 1.2288 Mchips/s, is 128 times the maximal source data rate of 9.6 kbps.
- *Acquisition*. The base station's acquisition and tracking of mobile signals is aided by the mobile's transmission of a preamble containing no data.
- *Voice Coding*. A variable-rate vocoder is specified, with data rates 1200, 2400, 4800, and 9600 bps depending on voice activity in a particular 20-ms frame. The transmission duty cycle of the reverse link signal during a call is proportional to the data rate.
- *Error Control Coding*. The reverse-link uses rate $\frac{1}{3}$ constraint length 9 convolutional coding, with Viterbi decoding.
- *Interleaving*. To protect against possible burst-error patterns, the reverse link interleaves code symbols before transmission, using a 20-ms span.

There is at least one reverse-link access channel for every paging channel on the forward link, with a maximum of 32 access channels per paging channel. The access channels are used for the mobile to initiate a call or respond to a page or information request from the base station. The number of active reverse-link traffic channels is equal to the number of active forward-link traffic channels. Each reverse link channel is distinguished by a distinct phase offset of the same 42-stage long-code PN sequence used on the forward link. The channels as received at the base station are illustrated in Fig. 7, where n (the number of paging channels) and m (the number of traffic channels) are limited by interference.

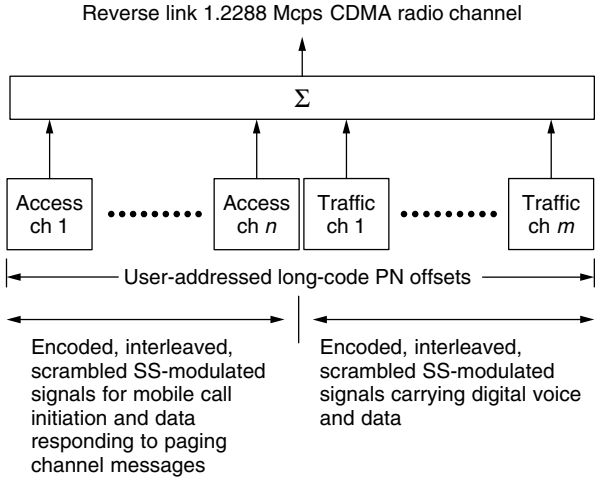


Figure 7. Reverse-link channel assignments at the base station.

The reverse-link transmitter consists of a convolutional encoder and modulator, and a quadrature PN-spreading modulator. The quadrature modulator for the reverse link is different from that used on the forward link in that a half-chip delay is inserted in the quadrature-phase (Q) channel to achieve a form of offset-QPSK (OQPSK) modulation. The one-half chip offset eliminates phase transitions through the origin to provide a modulation scheme that gives a relatively constant envelope. The transmission of the same data by means of a two-quadrature modulation scheme is a form of diversity. Its analytical justification in Refs. 5 and 13 shows that the QPSK CDMA system has a 3-dB advantage over BPSK CDMA system in terms of intersymbol interference performance and also has cochannel interference advantages.

Figure 8 is a block diagram of traffic channel processing. A variable rate vocoder is used to generate a digital voice signal at a rate varying from 0.8 to 8.6 kbps in a given 20-ms traffic channel frame. Depending on the data rate, the data frame is encoded with a CRC block code to enable the base station receiver to determine whether the frame has been received with error. An 8-bit encoder tail is added to the frame to ensure that the convolutional encoder, which follows, is reset to the all-zero state at the end of the frame. These operations result in data rates of 9600 (full rate), 4800 (half rate), 2400 ($\frac{1}{4}$ rate), or 1200 ($\frac{1}{8}$ rate) bps with, respectively, 192, 96, 48, or 24 bits per frame. The frame is then convolutionally encoded at a $\frac{1}{3}$ rate, resulting in $3 \times 192 = 576$ code symbols per frame at full rate, or 28.8 ksps. For other voice data rates, the code symbols are repeated as necessary to cause each rate to input the same number of code symbols to the interleaver in a frame.

Each group of six consecutive encoded symbols out of the interleaver is used to select a 64-chip Walsh sequence for orthogonal modulation, with a chip rate of $28.8 \times 64/6 = 307.2$ kchips/s. Because of the way that the symbols are read out from the interleaver array, these modulation symbols occur in alternating groups of six modulation symbols and $6(n-1)$ repeated modulation symbols, where n is the order of repetition. Altogether in a frame interval there are

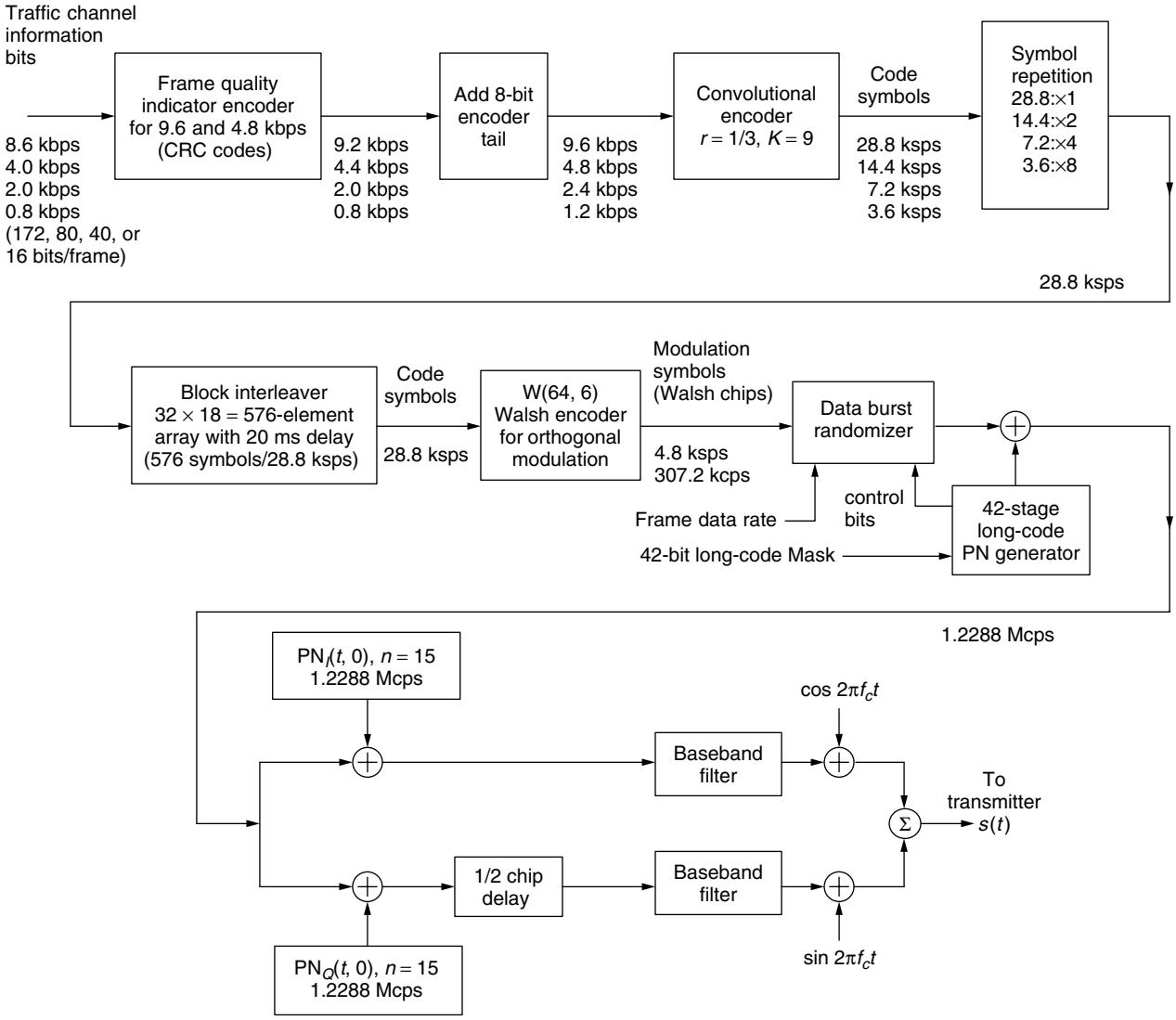


Figure 8. Reverse traffic channel modulation.

$96 \div 6 = 16$ groups of six orthogonal modulation symbols, each composed of $6 \times 64 = 384$ Walsh chips.

To reduce the average amount of reverse link interference and thereby increase user capacity, on reverse link transmissions, repeated symbols are gated off. A "data burst randomizer" is used to select in a pseudorandom manner which of the groups of six symbols are transmitted, based on "control bits" that are values of long PN code sequence bits at a certain time.

The user-distinct offset of the 42-stage long PN code is used to further spread the signal and ensure that the channels can be distinguished. The offset is implemented using a mask that depends on the electronic serial number (ESN) of the mobile. The masks and offsets on both forward and reverse traffic channels are identical for a given mobile user. The modulation parameters of the reverse traffic channel are summarized in Table 2.

Transmissions on the reverse traffic channel begin with a preamble of all-zero data frames to aid the base station in acquiring the signal. Signaling messages from the mobile to the base station may be sent on the reverse traffic channel as well as the access channel. When a message is to be sent, it can be sent in a "dim and burst" mode during periods of active speech, in which a portion of the voice data in a frame is overwritten by the message data, or in a "blank and burst" mode during periods of speech inactivity, in which all the data in the frame are message data.

2.4. Diversity Features of the IS95 System Design

Special features of the design of IS95 and of its common implementations are described in Ref. 5. Here we mention that the digital SS design of the IS95 forward- and reverse-link waveforms permits the use of several forms of diversity in addition to the time diversity inherent in the repetition, encoding, and interleaving of the data symbols. These forms of diversity include multipath diversity and base station diversity; the latter is available with or without the prospect of handing off the call to a different base station.

Table 2. Reverse Traffic Channel Modulation Parameters

Parameter	Value				Units
Data rate	9600	4800	2400	1200	bps
PN chip rate	1.2288				Mchips/s
Code rate	1/3				bits/code symbol
Transmit duty cycle	100	50	25	12.5	%
Code symbol rate	28,800				sps
Modulation rate	6				Code sym/mod sym
Mod symbol rate	4800				sps
Walsh chip rate	307.2				Kchips/s
Mod symbol duration	208.33				μs
PN chips/code symbol	42.67				—
PN chips/modulation symbol	256				—
PN chips/Walsh chip	4				—

Because the IS95 waveform for a particular channel is a dual-quadrature direct-sequence SS signal, it is possible to employ SS correlation techniques to isolate a single multipath component of that channel's signal and to discriminate not only against other channels' signals but also against multipath components of the same channel's received signal. Using the RAKE technique [14], in which the receiver uses several parallel receiver "fingers" to isolate multipath components, on the forward link it is possible to extract several multipath components from the total received signal and to align them for optimal combining.

If, at a particular mobile station, another cell site pilot signal becomes significantly stronger than the current pilot signal, the control processor initiates handoff procedures during which the forward links of both cell sites transmit the same call data to that mobile, which uses different fingers to process the two base station signals. With both sites handling the call, additional space diversity is obtained. When handoff is not contemplated, in a cell site diversity mode the strongest paths from multiple cell sites are determined by a search receiver, and the digital data receivers in the RAKE fingers are assigned to demodulate these paths. The data from multiple digital receivers are combined for improved resistance to fading.

Soft handoff methods have several advantages over conventional hard handoff methods [15]. Contact with the new base station is made before the call is switched, which prevents the mobile from losing contact with the system if the handoff signal is not heard or incorrectly interpreted. Diversity combining is used between multiple cell sites, allowing for additional resistance to fading.

2.5. System Evolution Toward Third Generation

The IS95B standard was published in 1999 [16]. This revision of the IS95 CAI features new multiplexing options that provide for transmission of up to eight simultaneous (parallel) full-rate "code channels" to constitute the forward- or reverse-link traffic channels. On the forward link, active users are assigned one *forward fundamental code channel* (including power control bit subchannel) with variable rate when only one code channel is operative, and the full rate (9600 or 14,400 bps) when multiple code channels are used. For data users needing rates greater than 9600 or 14,400 bps, from 0 to 7 *forward supplemental code channel* at full rate can used to transmit data in parallel on orthogonally multiplexed forward code channels, as illustrated in Fig. 9.

Multiple 64-chip Walsh functions are allocated as needed to implement orthogonal Walsh function multiplexing of the forward-code channels — the designation "code channel" thus refers to Walsh functions on the forward link. When multiple Walsh code channels are used

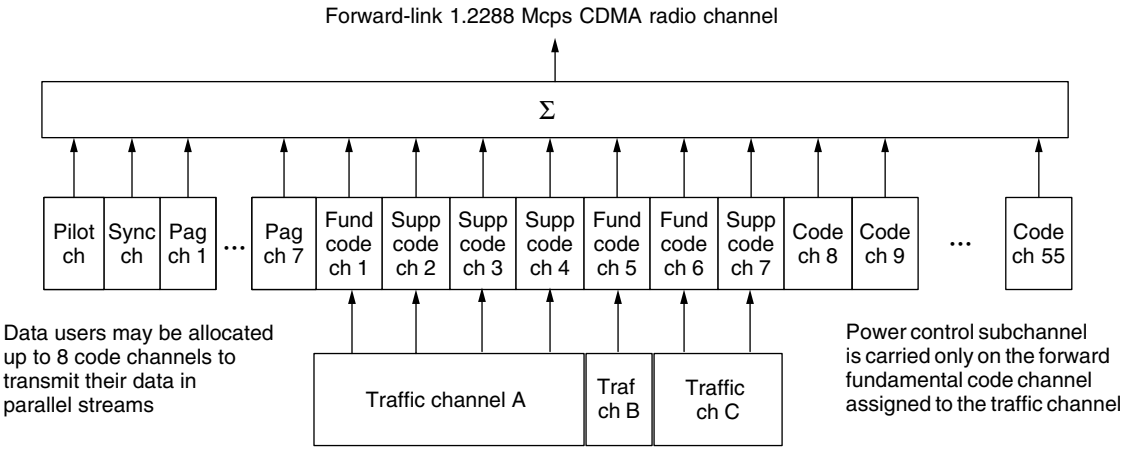


Figure 9. Forward-link channel aggregation under IS95B.

for a traffic channel, the scrambling in each code channel uses the same long code mask. The maximum bit rate for a traffic channel using eight code channels is 76.8 kbps for rate set 1 (1928 bits per 20-ms frame) and 115.2 kbps for rate set 2 (2888 bits per 20-ms frame).

The maximum information bit rate for a forward traffic channel (excluding overhead bits) using eight code channels is 68.8 kbps for rate set 1 (1728 bits per 20-ms frame) and 106.8 kbps for rate set 2 (2678 bits per 20-ms frame). Each mobile receiver "finger" must be able to demodulate up to eight forward code channels (and, by implication, reassemble them into the original single, high-speed bitstream). In 1998, Qualcomm announced the availability of the fifth-generation single-chip Mobile Station Modem (MSM3000) that uses a new "superfinger" demodulator architecture to support simultaneous demodulation in each finger of up to eight 9600-bps forward-link channels (76.8 kbps total) and up to six 14,400-bps channels (86.4 kbps total). In 1999, Qualcomm brought out the sixth generation of the chip, MSM3100. Features include voice recognition, echo cancellation, and GPS processing for position location. However, as of mid-2001, the SuperFinger is still limited to 86.4 kbps for a single user.

In a similar manner, the IS95B reverse link provides each active user with a reverse fundamental code channel with variable rate and random burst transmission when only one code channel is operative, and the full rate (9600 or 14,400 bps) when multiple code channels are used. For data users needing rates greater than 9600 or 14,400 bps, from 0 to 7 reverse supplemental code channels at full rate can be used to transmit data in parallel over PN code-division-multiplexed reverse code channels. When multiple code channels are allocated to a particular mobile user, the long PN code masks of the code channels are indexed to the code channel numbers for code-division multiplexing of the reverse code channels — the mask starts with 11xxx11000, where xxx = 000, 001, and so on. In addition, supplemental code channels, as added, are offset in carrier phase by certain increments of $\pi/4$ ($\pi/2$, $\pi/4$, $3\pi/4$, 0, $\pi/2$, $\pi/4$, $3\pi/4$, in that order) that provide at least partial carrier phase orthogonality to the supplemental code channels. Provision is made for intermittent adding or dropping of supplemental channels, with preambles to aid acquisition.

An advanced version of IS95, known as *cdma2000* and described in the TIA interim standard IS2000 [17], was proposed to the International Telecommunications Union as a candidate for the international third-generation cellular system. It was accepted as one of several technologies for future development and possible deployment.

BIOGRAPHIES

Leonard E. Miller received his B.E.E. degree in 1964 from Rensselaer Polytechnic Institute, Troy, New York; a M.S.E.E. degree in 1966 from Purdue University, Lafayette, Indiana; and a Ph.D. degree in electrical engineering from Catholic University of America, Washington, D.C., in 1973. He joined the Naval Ordnance Laboratory, Silver Spring, Maryland (later called Naval Surface Weapons Center) in 1964 as an electronics engineer. At NOL he designed instrumentation for underwater weapons and systems, and he developed algorithms for processing sonar signals. In 1978, he joined J. S. Lee Associates, Inc., Rockville, Maryland, where he performed R&D related to military surveillance and communication systems and also studied digital cellular telephone technology. Since May 2000, he has been with the Wireless Communication Technologies Group of the Advanced Network Technologies Division at the National Institute of Standards and Technology, Gaithersburg, Maryland. At NIST he is involved in R&D related to wireless ad-hoc networks. Dr. Miller, a senior member of IEEE, is coauthor of CDMA Systems Engineering Handbook (Artech House, 1998) as well as many journal and conference publications.

Jhong Sam Lee received his B.S. degree in electrical engineering in 1959 from the University of Oklahoma and his M.S.E. and D.Sc. degrees in electrical engineering from the George Washington University, Washington, D.C., in 1961 and 1967, respectively.

From 1959 to 1964 he worked in the industry in radar systems and microwave components development. From 1965 to 1968 he was an assistant professor at the George Washington University, Washington, D.C., where he taught courses in digital communication, information, and coding theories. From 1968 to 1969 he was an advisory engineer at IBM Corporation. From 1969 to 1973 he was an associate professor of electrical engineering at the Catholic University of America, Washington, D.C. From 1965 to 1973 he was a consultant at the U.S. Naval Research Laboratory in the areas of underwater signal processing and spread spectrum communication systems. In 1976, he founded J.S. Lee Associates, Inc. (JSLAI), for development of techniques in satellite and electronic warfare systems for DoD (Department of Defense) and its component services. He also founded in 2000 Advanced Technology Systems, Inc., in Seoul, Korea, for development (and for manufacturing) of interference cancellation systems for applications in the CDMA wireless transmission networks. Dr. Lee coauthored (with Dr. L.E. Miller) a book, CDMA Systems Engineering Handbook (1200pp), Artech House (1998), which is also translated into the Chinese language and was published in the People's Republic of China in 2001. He holds several patents in the area of CDMA wireless communications. Dr. Lee is a fellow of the Institute of Electrical and Electronics Engineers, Inc. (IEEE).

BIBLIOGRAPHY

1. *Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System*, TIA/EIA Interim Standard 95 (IS95), Telecommunications Industry Assoc., Washington, DC, July 1993 (amended as IS95A in May 1995).
2. V. H. MacDonald, The cellular concept, *Bell Syst. Tech. J.* **58**(1): 15–41 (Jan. 1979).
3. *Cellular System Dual-Mode Mobile Station–Base Station Compatibility Standard*, TIA/EIA Interim Standard 54 (IS54B), Telecommunications Industry Assoc., Washington, DC, April 1992.

4. U.S. Patent 5,103,459 (April 7, 1992), K. S. Gilhousen et al., System and method for generating signal waveforms in a CDMA cellular telephone system.
5. J. S. Lee and L. E. Miller, *CDMA Systems Engineering Handbook*, Artech House, Boston, 1998.
6. R. Padovani, Reverse link performance of IS95 based cellular systems, *IEEE Pers. Commun. Mag.* 28–34 (3rd quarter 1998).
7. K. I. Kim, CDMA cellular engineering issues, *IEEE Trans. Vehic. Technol.* **42**: 345–350 (Aug. 1993).
8. C. Wheatley, Trading coverage for capacity in cellular systems: A system perspective, *Microwave J.* **38**: 62–79 (July 1995).
9. K. S. Gilhousen et al., On the capacity of a cellular CDMA system, *IEEE Trans. Vehic. Technol.* **40**: 303–312 (May 1991).
10. K. Gilhousen, On the "optimum" bandwidth for spread spectrum, *Proc. 2nd Int. Conf. Personal, Mobile, and Spread Spectrum Communications*, Beijing, 1994, pp. 202–210.
11. H. Harmuth, A generalized concept of frequency and some applications, *IEEE Trans. Inform. Theory* **IT-14**: 375–382 (May 1968).
12. *Personal Station-Base Station Compatibility Requirements for 1.8 to 2.0 GHz Code Division Multiple Access (CDMA) Personal Communications Systems*, ANSI J-STD-008, Telecommunications Industry Assoc., Washington, DC, 1996.
13. A. J. Viterbi, *Principles of Spread Spectrum Multiple Access Communication*, Addison-Wesley, New York, 1995.
14. R. Price and P. E. Green, Jr., A communication technique for multipath channels, *Proc. Inst. Radio Engineers*, March 1958, Vol. 47, pp. 555–570.
15. A. J. Viterbi, A. M. Viterbi, K. S. Gilhousen, and E. Zehavi, Soft handoff extends CDMA cell coverage and increases reverse link capacity, *IEEE J. Select. Areas Commun.* **12**(8): 1281–1288 (Oct. 1994).
16. *Mobile Station-Base Station Compatibility Standard for Wideband Spread Spectrum Cellular Systems* (ANSI/TIA/EIA-95-B-99), Telecommunications Industry Assoc., Washington, DC, Feb. 1999.
17. *Physical Layer Standards for cdma2000 Spread Spectrum Systems*, TIA/EIA/IS-2000-1a, March 2000.

v

cdma2000

GIOVANNI EMANUELE CORAZZA
ALESSANDRO VANELLI-CORALLI
University of Bologna
Bologna, Italy

1. OVERVIEW

cdma2000 is the multicarrier wideband code-division multiple access (CDMA) radio interface included in the International Mobile Telecommunications 2000 (IMT-2000) family of standards. As such, cdma2000 shares with various other standardized radio interfaces many features that are best described with an introduction to IMT-2000. IMT-2000, developed under the auspices of the International Telecommunications Union (ITU), is the third-generation (3G) mobile telecommunications standard system. The objectives pursued by IMT-2000 are global service capability, standardized radio interfaces, flexible/seamless service provision, advanced multimedia services and applications, integrated terrestrial and satellite networks, and finally improved operational efficiency with respect to second-generation (2G) standards. At the present state of the standardization effort, IMT-2000 includes five terrestrial and six satellite radio interfaces [1]. The terrestrial radio interfaces are organized in two groups: CDMA interfaces and TDMA interfaces. The CDMA interfaces are IMT-DS (W-CDMA, FDD), IMT-MC (cdma2000, FDD), and IMT-TC (TD-SCDMA, TDD), whereas the TDMA interfaces are IMT-SC (UWC-136, FDD) and IMT-FT (DECT, TDD). All IMT-2000-compliant systems must support voice and data services. The latter can be symmetric or asymmetric, circuit- or packet-switched, with data rates ranging from 16 kbps to 2 Mbps. The IMT-2000 spectrum allocation was set in the works of three World Radio Conferences (WRC92, WRC95, and WRC00) and includes the following frequency bands: 806–960 MHz, 1710–1885 MHz, 1885–2025 MHz, 2110–2200 MHz, and 2500–2690 MHz.

The cdma2000 radio interface has been developed within the Third-Generation Partnership Project 2 (3GPP2) [2]. 3GPP2 is a partnership of standards development organizations, aiming at defining a 3G system based on the ANSI/TIA/EIA-41 core network. 3GPP2 partners are ARIB (Japan), CWTS (China), TIA (USA), TTA (Korea), and TTC (Japan). 3GPP2 works are organized in four main committees: Organizational Partners Committee, Steering Committee, Technical Specification Groups (TSGs), and ad hoc groups. The technical specification groups are TSG-A (access network interface), TSG-C (cdma2000), TSG-N (Intersystem operations), TSG-P (wireless packet data networking), and TSG-S (services and systems aspects). The cdma2000 standard is presently under development within 3GPP2. To date, releases 0, A, and B of the technical specifications have been published, whereas release C is in progress. To provide the most up-to-date information, this article discusses the release C specifications, which build strongly on releases A and B while adding a further radio configuration for high-speed packet data transmission, identified as 1XEVolved high-speed integrated Data and Voice (1X EV-DV) [3]. The 1X EV-DV radio configuration is also known as the high-data-rate (HDR) system.

A description of a few notable cdma2000 features is in order to conclude this introductory overview. The multicarrier structure pertains strictly to the forward link, where N ($N = 1, 3$, and optionally 6, 9, 12) adjacent direct-sequence spread RF carriers are used, while in the reverse link a single direct-spread RF carrier is adopted, with flexible spreading factor. This multicarrier flexibility is instrumental in guaranteeing backward compatibility with the 2G standard TIA/EIA-IS95B [4], also known as *cdmaOne*, thus easing the transition and coexistence between the two standards. Code-division multiple access is achieved adopting spread-spectrum techniques with long spreading codes. Notably, the time epoch offsets of these codes identify base stations and users in the forward and reverse

links, respectively. This time epoch management is made possible by the fact that all base stations in the network are synchronized to a common time reference. The time reference system takes advantage of the Global Positioning System (GPS) for clock synchronization. In the forward link, the chip rate amounts to 1.2288 Mcps on each carrier, which adds up to 3.6864 Mcps when three carriers are used. In the return link both 1.2288 and 3.6864 Mcps chip rates are imposed on a single carrier by varying the spreading factor. These two spreading rate modes are referred to as *spreading rate 1* (SR1) and *spreading rate 3* (SR3). SR1 and SR3 are commonly referred to as "1X" and "3X," respectively. Finally, the cdma2000 core network specifications are based on the evolved ANSI 41 and IP networks. Additionally, to maximize customer roaming capabilities the cross-mode operation with the GSM-MAP core network, identified as MC-MAP, is also supported [5].

2. THE cdma2000 AIR INTERFACE STANDARD

The cdma2000 core air interface standard is reported in 3GPP2 specifications C.S0001–C.S0006 [6–10]. An additional specification [11] is provided to support analog operations for dual-mode mobile stations (MSs) and base stations (BSs).

The protocol architecture of the air interface has been developed with reference to the ISO/OSI model, and is reported in Fig. 1. The ISO/OSI layer 1, or physical layer [7], provides for transmission and reception of radio signals between the base station and the mobile station. The physical layer services are offered to the upper layers through physical channels (represented in Fig. 1 by dotted lines with uppercase labels), which are the means for information transport over the air. The physical channels are characterized by the coding technique and rate, the spreading factor, and the digital modulation scheme. The precise parameters of each physical channel are defined in a set of radio configurations (RCs). There are 10 RCs for the forward link and 6 RCs for the return link, which collectively form the FDD MC-CDMA 1X/3X air interface. The ISO/OSI layer 2 (data-link layer) provides for delivery of signaling messages generated by layer 3 (network layer), making use of the services provided by layer 1. It has been subdivided into a medium access control (MAC) layer [8], and a link access control (LAC) layer [9]. The MAC layer is further subdivided into the multiplexing and QoS entity, the radio link protocol (RLP) entity, the signaling radio burst protocol (SRBP) entity, and the packet data channel control function (PDCHCF) entity. The MAC services are provided to the upper layers through logical channels (shown by solid lines with lowercase labels in Fig. 1), which are identified by the carried information typology, and are mapped onto physical channels. The LAC layer provides signaling, packet data voice, and data service transportation for the upper layers. Finally, the upper signaling layer [10] makes use of the services provided by layer 2 to support a wide range of radio interface signaling alternatives, namely, the native cdma2000 upper-layer signaling, backward-compatible TIA/EIA-IS95B signaling, and other existing or future upper-layer signaling entities. In layer 3 signaling messages between BS and MS are originated and terminated.

2.1. The Physical Layer

The physical layer (layer 1) offers information transfer between the mobile station and the base station to MAC and higher layers by means of physical channels. Physical layer specifications are defined in Ref. 7. Spreading rates, data rates, modulation parameters, forward error correction (FEC) schemes, puncturing, repetition rates, interleaving, and channel structures for the forward- and reverse-link signals are specified by RC1–RC10 and RC1–RC6, respectively. RC1 and RC2 are backward-compatible with the TIA/EIA-IS95B standard.

2.1.1. Forward Link. In the forward link, RC1–RC5 plus RC10 employ SR1, whereas RC6–RC9 correspond to SR3. Data rates for the different radio configurations are up to 9.6 kbps for RC1, 14.4 kbps for RC2, 153.6 kbps for RC3, 307.2 kbps for RC4, 230.4 kbps for RC5, 307.2 kbps for RC6, 614.4 kbps for RC7, 460.8 kbps for RC8, and 1.0368 Mbps for RC9. In RC10 data rates per subpacket range from 81.6 to 3.0912 Mbps.

The forward-link physical channels are shown in Fig. 2. They are the *pilot channels*, used for channel estimation and power-level measurements; the *common power control channel*, which carries as many power control bits as the number of active *reverse traffic*, *common control*, *acknowledgment*, or *channel quality indicator channels*; the *common assignment channel*, which is used for quick assignment of a *reverse-link channel* for random-access packets; the *common control channel*, which carries MS-specific messages; the *synchronization channel*, used to aid the initial time synchronization procedure; the *broadcast control channel*, used to transmit BS-specific, systemwide information and MS-specific messages; the *paging channel*, used in SR1 to transmit system overhead information and MS specific messages; the *quick paging channel*, used to inform MSs in idle state and slotted mode, as to whether they should receive the *forward common control* or the *paging channels* in the next slot; the *packet data control channels* and the *traffic channels*, which are used to transmit signaling and user information to a specific MS during a call.

The *traffic channel* includes the *fundamental channel*, which carries user and signaling information during a call; the *packet data channel*, which is used to transmit user packet data in RC10 with SR1; the *dedicated control channel*, which is employed to send user and signaling information during a call; the *power control subchannel*, which is used to transmit power control messages in association with a fundamental channel or a forward dedicated channel; the *supplemental* and *supplemental code channels*, which are used to transmit user information to a specific MS during a call in RC3–RC9 and RC1–RC2, respectively. There can be up to two supplemental channels and up to seven supplemental code channels for each traffic channel.

The pilot channels are a set of unmodulated spread-spectrum signals consisting of the *forward pilot channel*, which provides a phase reference for coherent

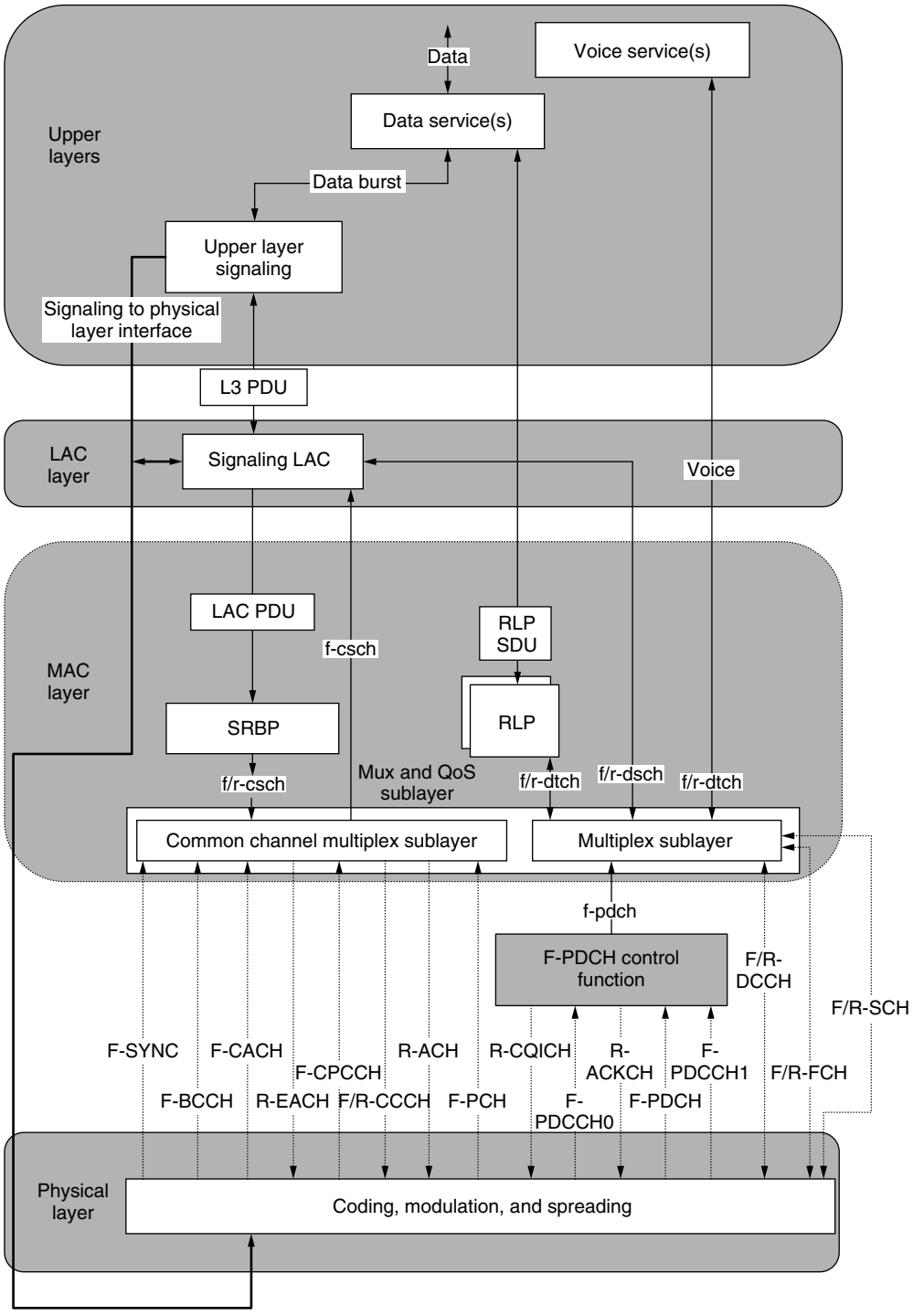


Figure 1. cdma2000 general radio interface protocol architecture [8]. Reprinted with permission.

demodulation and is used for signal strength comparison for the handoff procedure; the *transmit diversity pilot channel*, which is transmitted whenever transmission diversity is applied; the *auxiliary pilot*, and the *auxiliary transmit diversity pilot channels*.

Physical channels are processed as reported in the simplified block diagrams in Figs. 3 and 4. The actual block diagrams are channel- and RC-dependent, and are reported in Ref. 7. Information coming from the higher layers enters the forward error correction (FEC) block, then undergoes repetition and/or puncturing, interleaving, scrambling, complex modulation mapping, and gain control. The encoded and modulated output symbols are then demultiplexed in N ($N = 1$ or 3) I(cosine)/Q(sine) pairs ($I_i, Q_i, i = 1$ or 3) that are fed to the spreading, filtering, and upconversion sections. The RF output is finally transmitted on N adjacent carriers. Normally, a single antenna is used [non–transmit diversity—(NTD)

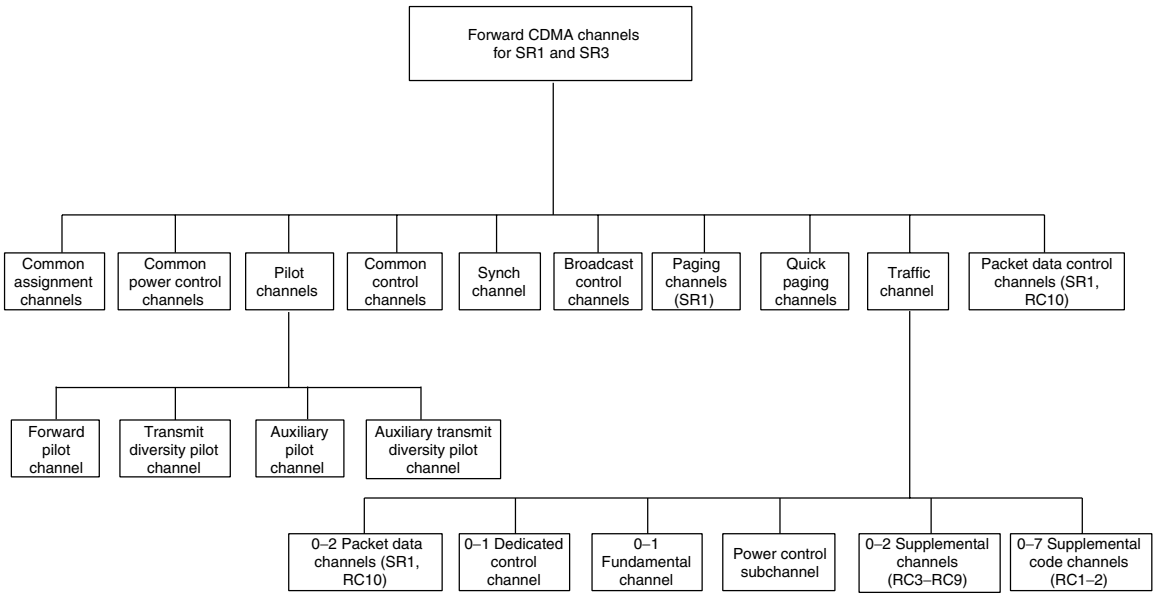


Figure 2. cdma2000 forward channels [7]. Reprinted with permission.

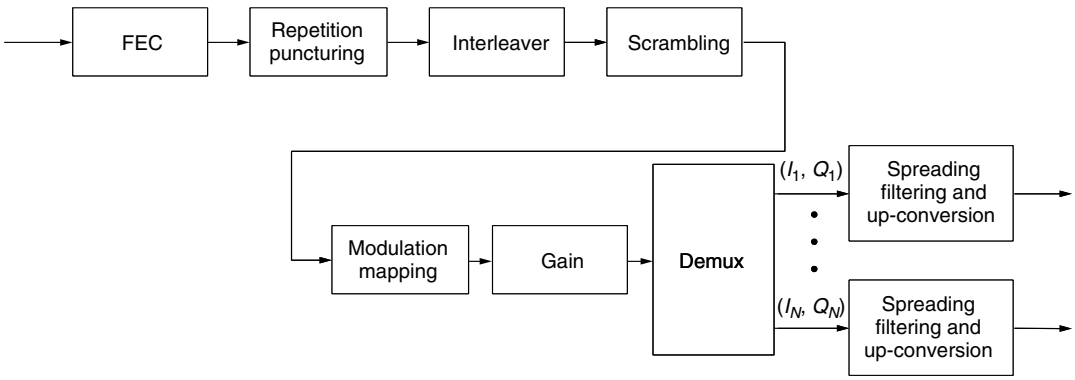


Figure 3. Simplified forward link transmitter block diagram.

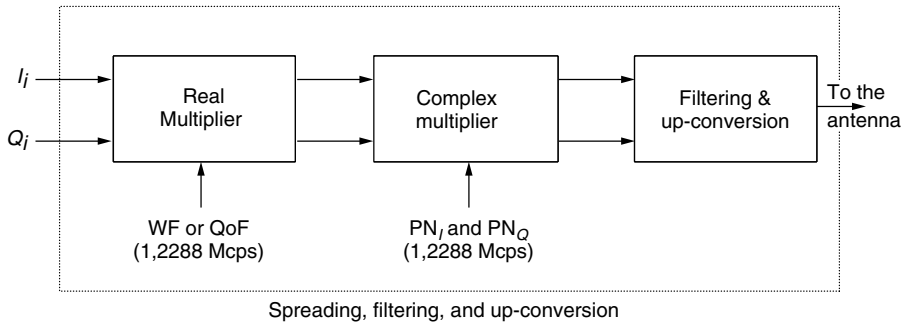


Figure 4. Simplified forward link spreading, filtering, and upconversion.

mode]. As an alternative, orthogonal TD (OTD) can be adopted, whereby two and three transmitting antennas are employed for SR1 and SR3, respectively. In these cases, the *transmit diversity pilot* and the *transmit diversity pilot channels* are also transmitted.

Forward error correction is accomplished through convolutional or turbo encoding. The possible coding rates are $R = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, and $\frac{1}{6}$ for the convolutional encoder, and $R = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, and $\frac{1}{5}$ for the turbo encoder. The coding schemes and rates for each physical channel are RC-dependent. All the convolutional encoders have constraint length $K = 9$. The rate $\frac{1}{2}$ convolutional encoder has generator polynomials $g_0 = 753$ and $g_1 = 561$. The rate $\frac{1}{3}$ convolutional encoder has generator polynomials $g_0 = 577, g_1 = 663$, and $g_2 = 711$. The rate $\frac{1}{4}$ convolutional encoder has generator polynomials $g_0 = 457$, $g_1 = 671$, $g_2 = 513$, and $g_3 = 473$. The rate $\frac{1}{6}$ convolutional encoder has generator polynomials $g_0 = 765$, $g_1 = 755$, $g_2 = 551$, $g_3 = 637$, $g_4 = 625$, and $g_5 = 727$. The turbo encoder consists of two identical recursive convolutional encoders, parallel concatenated,

with a turbo interleaver preceding the second convolutional encoder. The transfer function for the recursive convolutional code used for all coding rates is $G(D) = [1 + n_0(D)/d(D) + n_1(D)/d(D)]$, where $d(D) = 1 + D^2 + D^3$, $n_0(D) = 1 + D + D^3$, and $n_1(D) = 1 + D + D^2 + D^3$. The encoder output is punctured and repeated to obtain the desired coding rate. No coding is applied to the *quick paging* and *common power control channels*.

Data scrambling is obtained by means of a long pseudonoise sequence (PNS), and is applied to all physical channels, with the exception of the common power control, pilot, synchronization, quick paging, and packet data control channels. In RC10, data scrambling on the traffic channel is obtained by means of a different scrambling sequence, produced by a 17-tap linear feedback shift register with generator polynomial $h(D) = D^{17} + D^{14} + 1$.

The modulation schemes are BPSK in RC1–RC2 and QPSK in RC3–RC9. In RC10, QPSK, 8-PSK, or 16-QAM are chosen adaptively depending on the radio propagation channel conditions. Orthogonal spreading is used to ensure separation between channels on each carrier (Fig. 4). The modulated symbols (I_i, Q_i) are spread to a chip rate of 1.2288 Mcps by way of Walsh functions (WFs) in RC1–RC2 and RC10, and by way of WF or alternatively quasiorthogonal functions (QOFs) in RC3–RC9. QOF sequences are obtained using a nonzero sign multiplier and a nonzero rotate enable WF to enlarge the set of orthogonal codes because, depending on the particular deployment and operating environment, the system capacity may result to be limited by the number of Walsh codes. Walsh sequences are indicated as W_n^K, n tagging the nth row of a $K \times K$ Hadamard matrix. A Hadamard matrix is recursively constructed as

$$H_{2K} = \begin{bmatrix} H_K & H_K \\ H_K & H_K^{'} \end{bmatrix}$$

where K is a power of 2, $H_K^{'}$ is the binary complement of H_K, and $H_1 = 0$.

Following orthogonal spreading, the quadrature pairs are chipwise complex-multiplied with an overlay quadrature spreading sequence. The quadrature spreading PNS sequence is formed by a couple of extended maximum-length shift register (MLSR) sequences of length 32768 chips (a 0 is inserted after the run of 14 consecutive zeros) at a chip rate of 1.2288 Mcps. The PNS sequence period is 26.66 ms. Following filtering, upconversion is obtained by way of harmonic in-phase and quadrature modulation.

The MS receiver chain performs complementary operations. Multiple propagation paths can be usefully collected and combined by using a rake receiver with multiple fingers. The rake receiver is also instrumental in implementing the soft handoff procedure.

2.1.2. Reverse Link. In the reverse link, RC1–RC4 use SR1, whereas RC5 and RC6 correspond to SR3. Data rates for RCs are up to 9.6 kbps for RC1, 14.4 kbps for RC2, 307.2 kbps for RC3, 230.4 kbps for RC4, 614.4 kbps for RC5, and 1.0368 Mbps for RC6.

Figure 5 represents the reverse link physical channel organization. The physical channels used by the mobile station to communicate with the base station are the *pilot channel*, used to aid BS operation in detecting a MS transmission and that includes the *reverse power control subchannel* for RC3–RC6; the *access* and the *enhanced access channels*, used to initiate communications or to respond to a message received on the FL; the *reverse common control channel*, employed to transmit user and signaling information when the *reverse traffic channels* are not active; and the *reverse traffic channel*, used to transmit user information and signaling during a call.

The reverse traffic channel includes the *reverse fundamental channel*, aimed at transmitting user and signaling information during a call; the *reverse supplemental code channel*, used in RC1–RC2 to carry user information during a call; the *reverse dedicated control channel*, aimed at transmitting user and signaling information during a call in RC1–RC2; the *reverse supplemental channel*, aimed at transporting user information during a call in RC3–RC6; the *reverse acknowledgment channel*, which provides control for the *forward packet data channel*; and the *reverse quality indicator channel*, which is used to indicate to the BS the channel quality measurements of the serving sector. The reverse acknowledgment and the reverse quality indicator channels are used only in combination with the forward packet data channel (i.e., RC10).

The reverse link physical channels are processed similarly to the forward-link case. Information coming from higher layers undergoes FEC, repetition and/or puncturing, interleaving, and complex modulation mapping. Differently from the forward link, the I and Q channels are used separately to transmit independent information.

FEC is accomplished by means of convolutional, turbo, or block coding. The admissible coding rates are $R = \frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$ for the convolutional code; $R = \frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, and $\frac{1}{5}$ for the turbo code; and $\frac{1}{3}$ for the block code. The convolutional and the turbo encoder schemes and their generator polynomials are identical to those used in the forward link. The block code is (12,4) and is used only for the reverse channel quality indicator channel. As for the forward link, the association between coding schemes and physical channels is specified by RCs. The modulation schemes employed in the reverse link are the backward-compatible TIA/EIA IS95B 64-ary orthogonal modulation in RC1–RC2 and dual-BPSK modulation in RC3–RC6.

Figure 6 shows the simplified spreading, filtering, and upconverting subsystem block diagram for RC3–RC6. Multiple physical channels of a user are separated by spreading with orthogonal Walsh functions, whereas users are separated by means of different PNS sequence offsets. The physical channel symbols are first spread through WF to a chip rate of $N \times 1.2288$ Mcps ($N = 1$ for SR1 and $N = 3$ for SR3), and then complex-multiplied with the long PN sequence at the corresponding rate. Because of the direct spread nature of the reverse link, the RF signal at the output of the upconversion is always transmitted on a single carrier. The block diagram for RC1–RC2 is identical to that of the TIA/EIA-IS95B system [15].

The BS receiver chain performs complementary operations. Multiple propagation paths can be usefully collected and combined by using a rake receiver with multiple fingers.

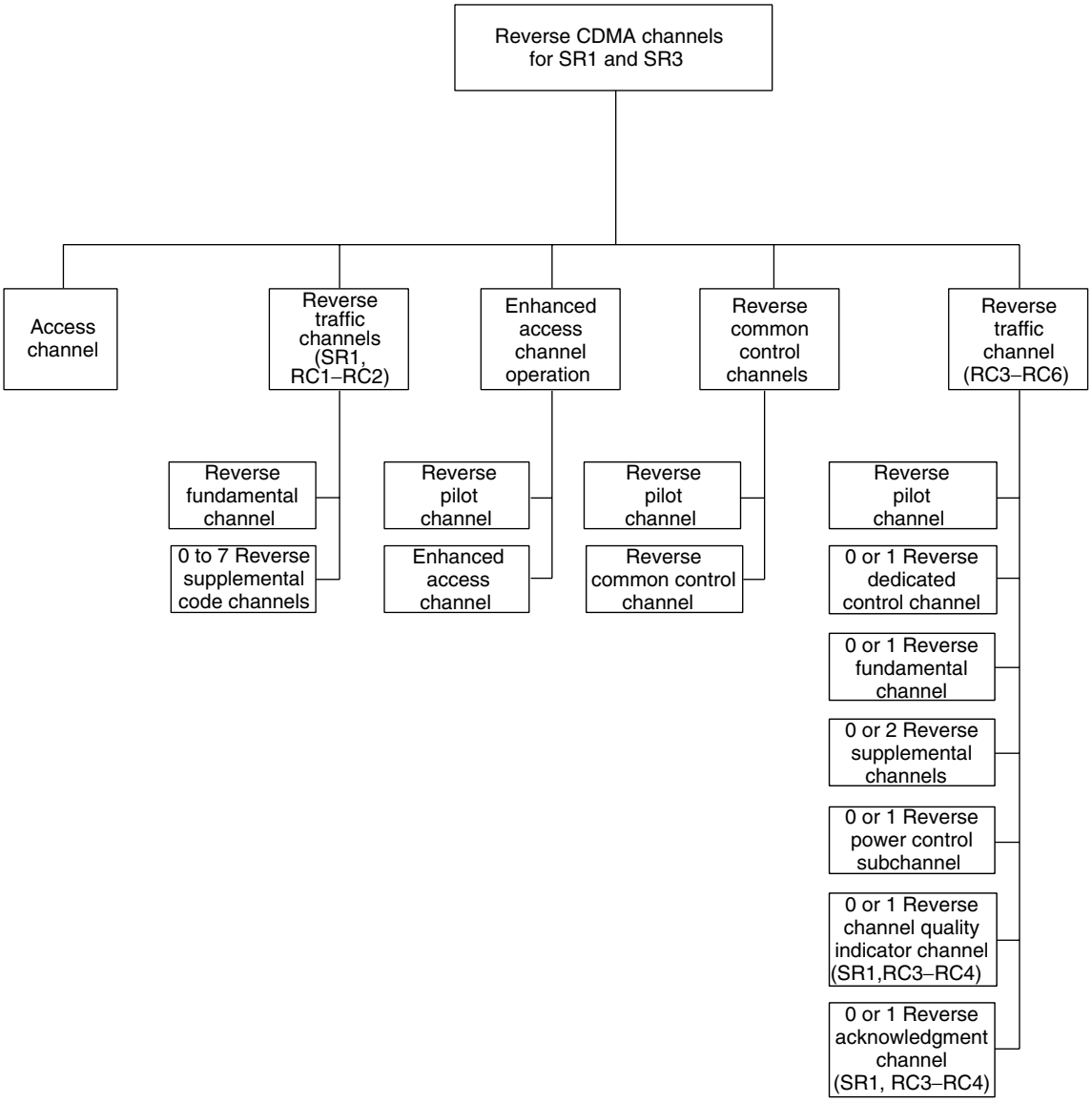


Figure 5. Reverse physical channels [7]. Reprinted with permission.

2.2. The Medium Access Control (MAC) Layer

The MAC layer is the interface toward the physical layer of the ISO/OSI reference model link layer, and it is described in Ref. 8. It provides the two important functions of best-effort delivery and of multiplexing and QoS control. However, when backward compatibility with TIA/EIA-IS95B is adopted, that is, when encoded voice data are transported directly by the physical layer, the MAC services are null.

The MAC services are provided by means of the logical channels and the MAC entities. The logical channels are connections between peer entities and they are defined by the information they carry. Logical channel categories are the *common signaling channel* (csch), *dedicated signaling channel* (dsch), *dedicated traffic channel* (dtch), and, for the forward link only, the *packet data channel* (pdch). Logical channels are associated with physical channels. Association between logical and physical channels can be (1) exclusive and permanent, (2) exclusive but temporary, or (3) shared. Information on how to perform this mapping is contained into the logical-to-physical mapping (LPM) table.

The MAC entities are the RLP, SRBP, multiplexer–QoS delivery, and the PDCHCF entity. The multiplexer–QoS delivery entity has both transmitting and receiving functions. The transmitting function combines information (LAC signaling, data services, voice services) into multiplex protocol data units (MuxPDUs), which in turn are mapped onto physical layer service data units (SDUs) and PDCHCF SDUs. The receiving function separates the physical-layer SDUs and the PDCHCF SDUs and directs information to the appropriate entity. The multiplexer entity may operate in two different modes: mode A for RC1–RC2 and mode B for RC $>$ 2. In mode A, a single MuxPDU is used to form a physical-layer SDU, while in mode B the additional flexibility of mapping one or more

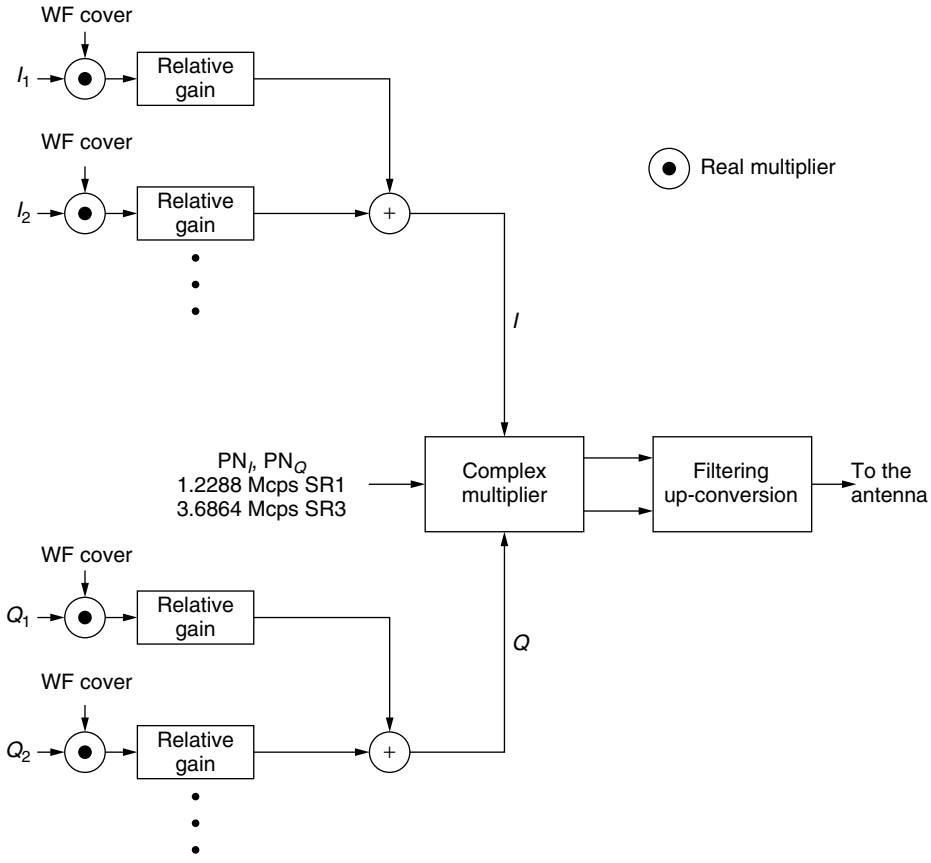


Figure 6. Reverse link I and Q spreading for RC3–RC6.

MuxPDUs into an SDU is provided. The multiplexer–QoS delivery entity determines the relative priority between information coming from higher entities. However, the precise use of priority information to guarantee a required QoS over the air is purposely not specified in the standard.

The SRBP entity manages the synchronization, paging, access, enhanced access, common assignment, broadcast channel, forward common control channel, and the return common control channel procedures. At the base station side the SRBP entity performs also the generation of the channel identifier.

The RLP is described in Ref. 14 and is used with a traffic channel to support connection-oriented negative-acknowledge-based data traffic delivery; that is, the receiver does not acknowledge correct reception and decoding of a data frame, but only requires the retransmission of erroneously received data frames. RLP is used only with RCs > 2. RLP supports both encrypted and nonencrypted data transport modes.

The PDCHCF entity is used in RC10 with the packet data channel to ensure the delivery of encoder data packets from BS to MS. In particular, four independent ARQ channels and code-division multiplexing (CDM) of encoder subpackets are adopted to enhance packet data transmission performance. All physical channels associated with the packet data channel, namely the packet data control channel, the acknowledgment channel, the channel quality indicator channel, and the packet data channel originate and terminate in the PDCHCF entity.

2.3. The Link Access Control (LAC) Layer

The LAC layer corresponds to the "upper" portion (i.e., above the MAC Layer) of the ISO/OSI Reference Model link layer. LAC provides for the correct transport and delivery of layer 3, signaling messages by implementing a data-link protocol. LAC offers the framework and the necessary support for point-to-point transmission over the air for signaling services, circuit data service provision (optionally), and transportation of encoded voice in the form of packet data or circuit data traffic. When backward compatibility with TIA/EIA-IS95B is enforced, then the LAC services are null.

The LAC layer is organized in the interface with the lower (MAC) and upper (L3 signaling) layers, protocol sublayers, and logical channels (Fig. 7). The LAC interfaces are identified as service access points (SAPs). On the L3-LAC SAP, LAC sends and receives SDUs and interface control primitives in the form of message control status blocks (MCSBs). At the LAC-MAC SAP, LAC exchanges LAC protocol data units (PDUs). The received L3-SDUs are serially processed by the protocol sublayer to form LAC PDUs, which in turn are serially processed to reconstruct L3-SDUs. The protocol sublayers are the *authentication-and-message integrity* sublayer, which performs the MS identification; the *ARQ sublayer*, which implements the ARQ protocol; the *addressing sublayer*, which takes care of the PDU addressing; the *utility sublayer*, which assembles and validates well-formed PDUs; and the *segmentation-and-reassembly sublayer*,

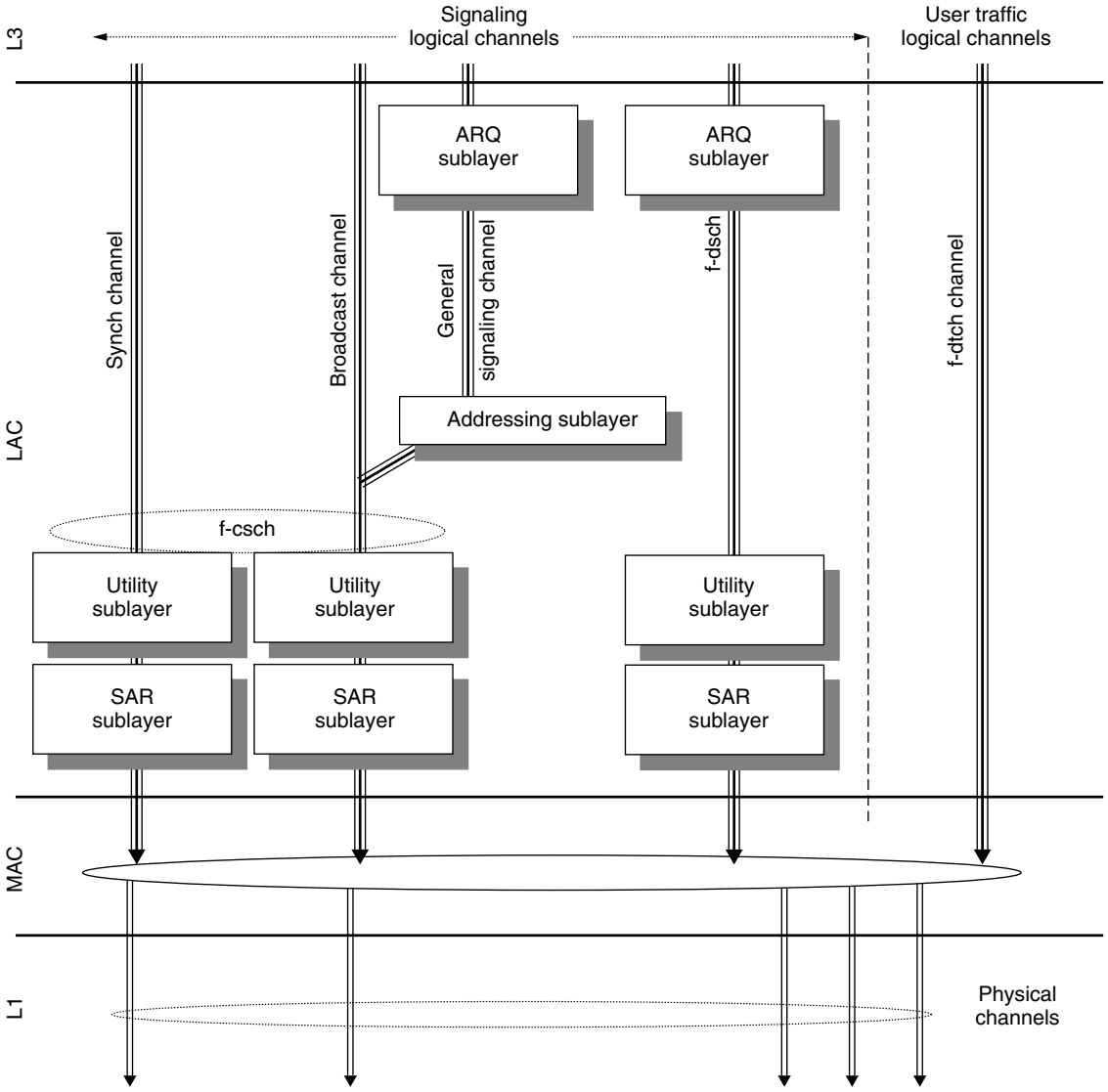


Figure 7. Architecture of the forward logical channels seen by the LAC sublayer [9]. Reprinted with permission.

which segments PDUs into PDU fragments or reassembles PDU fragments into PDUs.

Logical channels are the means for information transport from L3 and LAC to MAC. Logical channels allow the LAC to disregard details of the radio air interface characteristics. Only forward and reverse common signaling channels (f/r-csch), and forward and reverse dedicated signaling channels (f/r-dsch) are employed by LAC for access, synchronization, broadcast, and general or dedicated signaling.

2.4. Upper-Layer (Layer 3) Signaling

Upper-layer (layer 3) signaling corresponds to layer 3 (L3) and above of the ISO/OSI Reference Model. The L3 signaling layer consists of the protocol layer, which sends and receives L3 PDUs to and from lower layers, SAPs, and the communication primitives between L3 and lower layers. The L3 signaling layer is described in Ref. 10.

From the BS point of view, L3 signaling consists of the *pilot* and *synchronization channel processing*, used to transmit the pilot and synchronization channels to the MSs in the initialization state to achieve synchronization; *the common channel processing* used to transmit the paging, forward common control, and broadcast control channels, which are monitored by the MS in either idle or system access state; *the access channel and enhanced access channel processing*, employed to monitor those channels that are used by a MS in the system access state to send messages; and the *traffic channel processing*, used to communicate with a MS in the *control on the traffic channel state*. From the MS point of views, L3 signaling consists of the *initialization*, *idle*, *system access*, and *control on the traffic channel states*.

The *initialization state* is subdivided into the *system determination substate*, in which the MS selects which system to use; the *pilot channel acquisition substate*, in which the MS acquires the *pilot channel* of a CDMA

system; the *synchronization channel acquisition substate*, in which the MS obtains system configuration and timing information; and the *timing change substate*, in which the MS synchronizes its timing to that of the acquired system.

In the *idle state*, the MS monitors the *paging channel*, the *quick paging channel*, the *forward common control channel*, and the *primary broadcast control channel*. MSs in this state can receive messages or incoming calls, can cancel a *priority access and channel assignment* (PACA) call, and can initiate a registration or a message transmission or a call.

In the *system access state*, the MS sends messages to a BS on the r-csch and receives messages from a BS on the f-csch. This state consists of the update overhead information, origination attempt, page response, order/message response, registration access, message transmission, and PACA cancel substates.

In the *control on the traffic channel state*, the MS communicates with a BS using f/r-dsch and f/r-dtch. It consists of the *traffic channel initialization substate*, in which the MS verifies that it can receive the forward traffic channel and begins to transmit on the reverse traffic channel; the *traffic channel substate*, in which the MS exchanges traffic channel frames with a BS in accordance with the current service configuration; and the *release substate*, in which the MS disconnects the calls and the physical channels.

2.5. System Procedures

2.5.1. Power Control Procedures. There are three independent power control methods envisaged in the cdma2000 specifications: open-loop power control, closed-loop power control, and code channel gain adjustment.

In the *open-loop power control*, the MS sets the transmitted power according to the measured ratio of the received energy per chip E_c to the total (i.e., interference plus noise) received power spectral density I_0. The MS measures the E_c value on the received pilot signal and the I_0 level over the entire $N \times 1.25$-MHz bandwidth, considering all rake receiver fingers.

The *closed-loop power control* algorithm consists of two loops: the *outer loop*, which sets the target ratio of the energy per bit E_b to the effective noise power spectral density N_t, according to the desired QoS (i.e., frame error rate) and the *inner loop*, which estimates the received E_b/N_t ratio on the traffic channel, compares it with the target value, and accordingly sets the value of the power control subchannel bit. Power control commands are sent every 1.25 ms performing a fast 800-Hz rate power control. Nominal step values for each power control command are 1, 0.5, or 0.25 dB. Differently from the TIA/EIA-IS95 system, in the cdma2000 forward-link power control, the MS sends directly power control bits to the BS instead of reporting the frame error rate. As an alternative, the *erasure indicator bits* and the *quality indicator bits* can be used to inform the BS about the erroneous reception of a frame.

The *code channel gain adjustment* consists in setting the relative channel gains to maintain the ratio of the mean code channel output power to the mean reverse pilot channel output power within predefined limits [7].

2.5.2. Paging Procedure. The paging procedure is used by a BS to inform a MS of incoming calls or signaling messages, namely, information needed to operate with this BS. The procedure relies on the use of the *paging*, *forward common control*, and *quick paging* channels. These channels are divided into 80-ms slots. When a MS monitors every slot, it is said to operate in nonslotted mode, whereas when it monitors only some preassigned slots, it is said to operate in slotted mode. In the slotted mode, the MS can stop or reduce its processing for power conservation. The nonslotted mode cannot be operated in the idle state.

The *quick paging channel* is used in slotted mode for the transmission of paging, broadcast, and configuration change indicators for a MS. Two paging indicators are reserved for a MS in its preassigned quick paging channel slot. A BS sets the paging indicators for a MS when it needs to start receiving the paging channel or forward common control channel.

2.5.3. Handoff Procedures. The handoff procedure is used to transfer the communication from one BS to another. Depending on the MS state, different handoff procedures are possible. When a MS is in the control on the traffic channel state, the handoff procedures are the *soft handoff*, the *hard handoff*, and the *CDMA-to-analog handoff*. When a MS is in the idle state, only the *idle handoff* is possible.

During a *soft handoff*, the MS starts communications with a new BS without interrupting the communications with the previous one. This procedure applies only to CDMA channels having the same frequency assignments. Notably, soft handoff provides path diversity for the forward traffic and reverse traffic channels at the boundaries between BS coverage. By means of the *hard handoff* procedure, a MS transits between disjoint sets of BSs, band classes, and frequency assignments. The hard handoff is characterized by a temporary disconnection of the traffic channel. The *CDMA-to-analog handoff* procedure is used whenever a MS is directed from a CDMA traffic channel to an analog voice channel. In this case temporary disconnection occurs. The *idle handoff* procedure applies when a MS in the idle state detects a pilot channel signal that is sufficiently stronger than that of the serving BS.

To perform handoff a MS maintains a list of available pilot channels. The pilot channels are grouped into sets describing their status with regard to pilot searching, on the base of their relative offset to the zero-offset pilot PNS sequence. The pilot channel sets are: the *active set*, consisting of the pilot channels corresponding to the paging channel or the forward common control channel currently monitored; the *neighbor set*, which consists of all the pilot channels that are likely candidates for the idle handoff and that are specified by broadcast messages on the paging–broadcast control channel; the *remaining set*, which consists of all possible pilot channels in the current system excluding those already included in the preceding two sets; and the *private neighbor set*, which consists of all the pilot channels available for the private systems and that are specified in a dedicate broadcast list.

2.5.4. Access Procedure. The *access procedure* is a power ramping slotted ALOHA procedure performed by

a MS aiming at sending a message to a BS and receiving an acknowledgment for that message. The access procedure is based on *access attempts*. Each attempt consists of a sequence of one or more *access subattempts*, in turn made up of a sequence of one or more *access probe* sent on the access channels.

The *access probe* consists of a *preamble part* and a *message part*. The preamble part is a sequence of all-zero frames sent at the 4800 bps rate and is the actual instrument for the power ramping procedure. The message part includes the message body, length field, and cyclic redundancy check (CRC).

3. cdma2000 KEY FEATURES

The most notable cdma2000 air interface characteristics are summarized here:

Core Network Compatibility. cdma2000 has been developed with reference to the evolved ANSI-41 and all IP core networks, identified as native MC-41 mode. However, cross-modes, namely, MC-MAP and DS-41, are developed under the auspices of the Operator Harmonization Group (OHG) [5] and supported [12,13] to extend user roaming capability. In particular, in the MC-MAP cross-mode L1, MAC, LAC, and radio resource control of the cdma2000 standard are combined with the connection management and mobility management layers of the UMTS W-CDMA FDD standard.

Backward Compatibility with TIA/EIA-IS95B. Backward compatibility with the TIA/EIA-IS95B system is fully enforced by the cdma2000 standard [6]. Backward compatibility ensures that any TIA/EIA-IS95B MS can place and receive calls in any cdma2000 system, and that any cdma2000 MS can place and receive calls in any TIA/EIA-IS95B system. In the latter case, the cdma2000 MS is limited to the IS95B service capabilities (i.e., only SR1 can be used). The compatibility between cdma2000 and TIA/EIA-IS95B systems involves also the handoff procedures. A cdma2000 system in fact supports handoff of voice, data, and other supported services from and toward a TIA/EIA-IS95B network. In particular, handoffs between cdma2000 and TIA/EIA-IS95B networks can occur at cell boundaries or in the same cell, either in the same or between different frequency bands.

Overlay Capabilities with TIA/EIA-IS95B. cdma2000 supports different channel bandwidths, 1.25 MHz channel bandwidth in SR1, and 3.75 MHz channel bandwidth in SR3. This enables cdma2000 to be deployed as an overlay of a TIA/EIA-IS95B system with many different configurations. For example, combining different forward- and reverse-link RCs the following deployments are possible: 1X forward and reverse links, 3X forward link and 1X reverse link, 3X forward and reverse links, 1X forward link and 3X reverse link. Therefore, a seamless and flexible transition from TIA/EIA-IS95B to cdma2000 networks and exploitation of the existing TIA/EIA-IS95 network coverage during cdma2000 deployment is possible.

Fast Power Control. cdma2000 supports fast closed-loop power control (800 Hz) in both forward and reverse links. In particular, in the reverse link following signal to interference plus noise measurements, the MS itself sends power control commands to the BS.

Forward-Link Transmit Diversity. cdma2000 offers transmission diversity capabilities by means of the OTD mode. In the case OTD is employed, two or three transmitting antennas are used for SR1 and SR3, respectively. In the SR3 case, the multicarrier feature is exploited to transmit a carrier per antenna. This can be shown to provide a very efficient form of diversity. Auxiliary pilot and auxiliary transmit diversity pilot physical channels are supplied for each antenna to ease MS synchronization, channel estimation, and power-level measurements.

Coherent Reverse Link. Carrier phase estimation for coherent reception in the reverse link is made possible by means of the transmission of reverse pilot physical channels whenever a reverse traffic channel is assigned.

Enhanced Channel Structure. cdma2000 employs 5, 10, 20, 40 and 80 ms frame lengths, providing a means for trading off overhead and delay.

Turbo Codes. cdma2000 forward- and reverse-link RCs employ high-performance turbo codes.

Synchronous Base Stations. The cdma2000 system architecture is characterized by synchronous base stations. The system time is synchronous with the Universal Coordinated Time (UTC) and it is derived from the GPS system.

Multiple Access Spreading Codes and Code Planning. cdma2000 uses two levels of spreading. The first level is used to separate different physical channels in a CDM flux transmitted by either a BS or a user. Separation is achieved by means of orthogonal Walsh functions or quasiorthogonal Walsh functions. The second level of spreading aims at separating different CDM fluxes. Separation is achieved by means of complex long PNS sequences. Exploiting the BS synchronism, different CDM fluxes are associated to different code phases of the same PNS sequence. This is a significant difference and advantage with respect to the code planning necessity of the W-CDMA standard. In fact, since different BSs are identified by a code offset, cell planning simplifies to the association between code offsets and BSs. To relax the requirements associated with timing, among all the possible phase-shifted codes, those with minimum phase distance are avoided.

Initial Synchronization. Since a MS needs only to search for different phases of the unique PNS sequence, initial synchronization is significantly simpler in cdma2000 than in the other 3G CDMA air interfaces.

High Data Rate (HDR) Packet Transmission. cdma-2000 envisages a high-data-rate packet transmission (RC10) using the 1X mode, 1X EV-DV, which employs adaptive coding and modulation, fast retransmission of erroneously received frames, multiple (up to 4) time multiplexed ARQ channels for each MS, best serving BS selection driven by the MS, and megadiversity via sector selection. The MAC layer employs the new PDCHCF entity to support 1X EV-DV.

BIOGRAPHIES

Giovanni Emanuele Corazza received the Dr. Ing. degree (summa cum laude) in Electronic Engineering in 1988 from the University of Bologna (Italy), and a Ph.D. in 1995 from the University of Rome "Tor Vergata" (Italy). He is currently a Full Professor at DEIS, University of Bologna. He holds the chair for Telecommunications inside the Faculty of Engineering, and he is responsible for the area of Wireless Communications inside the Advanced Research Centre for Electronic Systems (ARCES). He is Vice-Chairman of the Advanced Satellite Mobile Systems Task Force (ASMS-TF), a European forum on satellite communications with more than 40 industrial partners. He visited ESA/ESTEC (Noordwjik, NL) as a Research Fellow, the University of Southern California (Los Angles, CA) as a Visiting Professor, and Qualcomm Inc. (San Diego, CA) as a Principal Engineer. He is associate Editor on spread spectrum for the *IEEE Transactions on Communications*. He received the Marconi International Fellowship Young Scientist Award in 1995 and two Best Paper Awards at IEEE Conferences. Professor Corazza has research interests in the areas of communication theory, wireless communications systems (including cellular, satellite, and fixed systems), spread-spectrum techniques, and synchronization. He is author or co-author of more than 70 papers published in international journals and conference proceedings.

Alessandro Vanelli-Coralli received the Dr. Ing. Degree (summa cum laude) in electronics Engineering and the Ph.D. in Electronics and Computer Science from the University of Bologna (Italy) in 1991 and 1996, respectively. Since 1996, he has been with the Department of Electronics, Computer Science and Systems (DEIS) at the University of Bologna where he is currently a Research Associate. Since 1995 he has held courses of Digital Communications at the Faculty of Engineering of the University of Bologna. He has been a research consultant on source coding and audio compression and has been involved in several national and international research projects. Dr. Vanelli-Coralli's research interests are in the area of digital communication systems addressing, in particular, satellite communications, spread-spectrum and CDMA systems, synchronization techniques, and digital signal processing. Dr. Vanelli-Coralli is a reviewer for IEEE journals and conferences, and has chaired sessions at IEEE Conferences. Dr. Vanelli-Coralli is co-recipient of the Best Paper Award at the IEEE ICT 2001 Conference. He is co-author of papers published in national and international journals and conference proceedings.

ACRONYMS

1X	Single carrier (i.e., spreading rate 1)
1X EV-DV	1XEVolved high-speed integrated Data and Voice
3G	Third generation
3GPP	Third-Generation Partnership Project
3X	Three carriers (i.e., spreading rate 3)
ANSI	American National Standard Institute
ARIB	Association of Radio Industries and Businesses
ARQ	Automatic repeat request
CDMA	Code-division multiple access
CRC	Cyclic redundancy check
csch	Common signaling channel
CWTS	China Wireless Telecommunication Standard Group
DECT	Digital enhanced cordless telecommunications
dsch	Dedicated signaling channel
dtch	Dedicated traffic channel
EIA	Electronic Industry Alliance
FDD	Frequency-division duplex
FEC	Forward error correction
GSM-MAP	Global System for Mobile communications — Mobile Application Part
IMT-2000	International Mobile Telecommunication 2000
IMT-DS	IMT direct spread
IMT-FT	IMT frequency time
IMT-MC	IMT multicarrier
IMT-SC	IMT single carrier
IMT-TC	IMT time code
ISO/OSI	International Standard Organization/Open System Interconnection
ITU	International Telecommunication Union
kbps	kilobit per second
LAC	Link access control
LPM	Logical-to-physical mapping
MAC	Medium access control
Mbps	megabits per second
MC	Multicarrier
MC-MAP	Multicarrier using GSM MAP
Mcps	Megachips per second
MCSB	Message control status block
MLSR	Maximum-length shift register
MS	Mobile station
PACA	Priority access and channel assignment
pdch	Packet data channel
PDCHCF	Packet data channel control function
PDU	Protocol data unit
PNS	PN sequence
QOF	Quasiorthogonal function
RC	Radio configuration
RF	Radiofrequency
RLP	Radio Link Protocol
SAP	Service access point
SDO	Standard Development Organization

SDU	Service data unit
SR1	Spreading rate 1, corresponding to 1X
SR3	Spreading rate 3 corresponding to 3X
SRBP	Signaling Radio Burst Protocol
TDD	Time-division multiplex
TD-SCDMA	Time-division single carrier CDMA
TIA	Telecommunication Industries Association
TTA	Telecommunications Technology Association
TTC	Telecommunication Technology Committee
UTC	Universal Coordinated Time
UMTS	Universal Mobile Telecommunication System
UWC-136	Universal Wireless Communication — 136
W-CDMA	Wideband CDMA
WF	Walsh function

BIBLIOGRAPHY

1. ITU-R, M.1457, *Detailed specifications of the radio interfaces of International Mobile Telecommunications-2000* (*IMT-2000*), draft revision, Doc 8/BL/6-E, April 17, 2001.
2. For further information on how to obtain 3GPP2 Technical Specifications and Technical Reports, please visit *http://www.3GPP2.org*
3. 3GPP2, S.R0026, *High-Speed Data Enhancements for cdma2000 1x — Integrated Data and Voice — Stage 1 Requirements*, version 1.0, *http://www.3gpp2.org*, Oct. 2000.
4. TIA/EIA-IS-95-B, *Mobile Station-Base Station Compatibility Standard for Wideband Spread Spectrum Cellular Systems*, Feb. 1999.
5. Operators Harmonization Group (OHG), *Specification framework for ITU IMT-2000 CDMA proposal*, *http://www.itu.int/imt/2-datio-dev/ohg/index-es.html*, Jan. 1999.
6. 3GPP2, C.S0001, *Introduction to cdma2000 Standards for Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
7. 3GPP2, C.S0002-C, *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
8. 3GPP2, C.S0003-C, *Medium Access Control* (*MAC*) *Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
9. 3GPP2, C.S0004-C, *Signaling Link Access Control* (*LAC*) *Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
10. 3GPP2, C.S0005-C, *Signalling Standard for cdma2000 Spread Spectrum Systems*, release C, version 1.0, *http://www.3gpp2.org*, May 2002.
11. 3GPP2, C.S0006-C, *Analog Signaling Standard for cdma2000 Spread Spectrum Systems*, release C, Version 1.0, *http://www.3gpp2.org*, May 2002.
12. 3GPP2, C.S0007-0, *Direct Spread Specification for Spread Spectrum Systems on ANSI-41* (DS-41) (Upper Layers Air Interface), *http://www.3gpp2.org*, June 2000.
13. 3GPP2, C.S0008-0, *Multi-carrier Specification for Spread Spectrum Systems on GSM MAP* (MC-MAP) (Lower Layers Air Interface), *http://www.3gpp2.org*, June 2000.
14. 3GPP2, C.S0017-0-2, *Data Service Options for Spread Spectrum Systems*, addendum 2, version 2.0, *http://www.3gpp2.org*, Aug. 2000.
15. V. K. Garg, *IS-95 CDMA and cdma2000 Cellular/PCS Systems Implementation*, Prentice-Hall Communications Engineering and Emerging Technologies Series, Englewood Cliffs, NJ, 2000.

CELL PLANNING IN WIRELESS NETWORKS

KAI ROHRBACHER
Head, Department of Mobile Communication Software
LStelcom Lichtenau, Germany

JÜRGEN KEHRBECK
Head, Division of e-Commerce and Mobile Communications
LStelcom Lichtenau, Germany

WERNER WIESBECK
Director, Institute for High Frequency Technology and Electronics Karlsruhe University, Germany

1. FOCUS OF THIS ARTICLE

This article is intended to give an overview of planning and simulating strategies for wireless networks, from the past to the present. The focus is on the most popular system technologies such as GSM900/1800, GPRS/EDGE, and the upcoming so-called third-generation systems (3G) with IMT-2000 framework like UMTS-FDD in Europe, UMTS-TDD, CDMA2000, and TD-SCDMA. The authors present examples and methodologies based on network technologies used in Europe and also discuss general aspects of network planning and optimization.

While most of the licenses for the 3G mobile systems have been granted and research on the rollout of 3G networks is still under way, research projects for even the next generation (4G) has begun.

2. WIRELESS NETWORKS: FROM THE BEGINNING TO THE FOURTH GENERATION

The development of mobile networks can be traced back to the early 1960s. The main target was to offer voice conversation at any point in the country at any time without being bound to fixed network lines. Widespread commercial deployment of the first analog systems begun in the 1980s [1]. The introduction of second-generation systems (2G), especially the GSM system in the late 1980s, then ignited the ultimate success story of mobile networks. In Europe, the year 2000 marked the milestone where in some countries, the number of mobile subscribers has topped the number of fixed (landline) telephone sets.

Customers increasingly, tend to see mobile networks as a natural mobile "extension" if not even a replacement of their fixed-line networks. This puts additional pressure on mobile network developments. This rising demand in quantity and quality of the networks in parallel with stiffering operator competition led to a dramatic reduction in the rollout time of new networks to a few months instead

of 2 or 3 years as in the early years of GSM. But also existing network operators are facing problems — rapidly growing user figures and demand for new services is increasingly pushing their networks to the limit. However, "bandwidth" is a physical constraint.

All these factors together made it increasingly necessary to have powerful computer systems that support the network rollout and optimization process.

Various tools have been developed since 1990, starting with simple propagation algorithms, which led to highly sophisticated computer and database systems that support the complete planning process from green field layouts to live network optimization.

2.1. 3G: From "Evolution" to "Revolution"

Having said that, the "mobile evolution" now turns into a "mobile revolution" with the upcoming third-generation mobile systems. These systems do not simply extend existing 2G, they are completely new and indeed, represent a paradigm shift in how to plan them. With their inherent vast additional level of complexity, it may become virtually impossible to plan 3G (let alone 4G) networks without massive computational support and guidance. This forces planning tool suppliers as well as network operators to use a whole new set of algorithms and planning processes to create and engineer these systems.

3. PLANNING 1G, 2G, 2.5G, 3G, 4G: A SHORT SURVEY

It is interesting to see how the method of planning cellular networks changed over time and what particular demands influenced (and do influence) that development.

The main driving factor of the development was the increasing demand for capacity on the air interface. This always has been a very challenging topic, because, as seen from a scientific viewpoint, earth's atmosphere is a poor medium for electronic signal transmissions. This led to the effect that increasing effort and complexity went into the development of the air interface whereas the backbone network remained relatively stable over time. Other parameters also affected the evolution path of cellular systems, which are also discussed here.

It is common usage to speak about "generations" of cellular networks. This is due to the fact that progress in cellular networks was a process that "warped" at certain times and then continued in a rather smooth evolution up to the next "warp"; it's these intervals that we call "generations."

3.1. 1G Systems: When It All Started

First-generation systems are characterized mostly by

- Analog transmissions
- Coarse site placements
- High power transmitters
- Simple modulation schemes
- Nationwide, incompatible systems
- Voice-only systems

The cellular systems are surprisingly old; it was only 33 years after German physicist Heinrich Rudolf Hertz had discovered the electromagnetic waves that the first public cellular phone was introduced in Germany by the Deutsche Reichsbahn on the Zossen-Berlin railway in 1918.

AT&T launched a commercial cellular network in 1946 in the city area of St. Louis: It used six fixed FM channels and manual switching by an operator. A year later, the Bell Labs obtained a patent on a frequency re-usage scheme deploying a regular cell grid.

The German "A network" was implemented in 1957 and is a classic example of a 1G system — It did provide nationwide coverage, but was divided into 136 areas. Each area used 37 frequency pairs (for up/downlink) and call connection could be done only by an operator.

Such 1G systems still were very similar to the radiobroadcast systems from which they originated; as with broadcast stations, transmitters for 1G systems were placed on high mountains or hills to cover large areas.

Spectrum efficiency was not a major issue; the "A network," for example, never exceeded 11.000 users, while occupying a frequency range of 156–174 MHz (=18 MHz bandwidth). "Planning" these networks consisted merely in finding a few sites on a hilltop and installing the system technology.

Later systems enabled self-establishment of calls, automatic call routing, and handover of outgoing calls to neighboring cells. The networks became able to track and handle the mobile user's position and establish a call to him/her automatically. This did not become possible before the wider availability of digital computers in the 1980s due to the required process automation.

Also, transmitter density increased, not only for better coverage, but also to decrease transmission power and improve spectrum efficiency.

One example of such late 1G systems was the German "C network," which worked with 287 channels in a frequency range of 450–465.74 MHz.

3.2. 2G Systems: Digital Technology

Key characteristics of 2G systems are

- Digital transmissions
- Dense site placements
- Low power transmitters
- Enhanced modulation schemes
- TDMA/FDMA access structures
- Semicompatible systems
- Voice focus, but first data service support
- Step-by-step planning: coverage before QoS (quality of service)

Typical representatives are GSM, DCS1800, PHS, IS-95, IS-136, PCS, and DECT.

In the early 1980s, it had became obvious that even with reuse patterns and denser networks, analog transmission couldn't cope with the increasing user figures any longer.

The C network reached its theoretical limit at about 1 million users only — not much for a potential 80 million German users.

Fortunately, the advances in computer technology provided a new solution for the capacity demands in the form of digital transmissions.

Once human voice is digitized, it can be sent as *compressed* data over the air interface. Exploiting the fact that the human perception system is quite forgiving in slight deviations of spoken voice, even more effective *lossy* compression algorithms could be applied. Enhanced modulation schemes such as GMSK and BPSK (Gaussian minimum shift keying and binary phase shift keying) added to this development. The density of the networks could increase further as time and frequency diversity were combined (GSM) or CDMA principles were used (IS95). This led to lower battery consumption and — together with advances in miniaturization — gave rise to truly small handsets. All this added to the capacity of the networks so that today, for example, a GSM-based network could serve theoretically about 100 million voice users.

Also, globalization effects started to drive standardization. GSM (standardized in 1982, officially released in 1990) became an especially impressive pan-European success.

Even if 2G systems began to offer some basic data services, they are still networks for mobile *voice* traffic only.

This one-service-only property simplifies the planning process to a large extent, as it does not necessitate consideration of the time-domain parameter. As we will see later in more detail, it suffices without significant loss of accuracy to work on *averaged* data and separate the planning process in distinct phases not impacting each other.

3.3. 2.5G and 3G Systems: Two Steps Instead of One

Key characteristics of 2.5G systems are

- Data and voice services
- Mixed services
- Circuit- packet-switched traffic

Key characteristics of 3G systems are

- Mainly data services
- A broad range of very differing services, multi media support
- Mainly packet-switched data
- Enhanced multiple-access schemes
- International standardization
- All-in-one planning: coverage *and* QoS
- Coexistence with 2G systems

In the late 1990s it was realized once again that the air interface would soon become a bottleneck, this time not because of the number of *voice* users (which could be handled by the existing 2G systems), but because of the growing demand for *data* services. The success of the (fixed-network) Internet is expected to lead to a similar demand for wireless Internet services. However, data services require large bandwidths, and, of course, only lossless compression schemes can be used applied, instead of the (more efficient) lossy algorithms used for voice traffic. For example, state-of-the art voice codecs can transmit human voice at acceptable speech quality with a data rate of less than 4 kbps, whereas an "acceptable" WWW session today seems to be in the range of ~64 kbps with demands rising, as multimedia contents begin to overtake WWW contents.

The required bit rate of a typical multimedia service seems to grow by nearly 50% per year. This puts pressure on the capabilities of the air interface, as customers come to expect the same behavior in their mobile service, that they have in their Fixed-network (landline) connections.

3G systems therefore focus mainly on data transmission. This, however, constitutes a complete paradigm change; even the most sophisticated 2G systems are still *circuit-switched* networks — using exactly the same principle as Philipp Reis did with world's first "fixed-network phone" experiment back in 1861! Data services traffic, on the other hand, is usually very bursty in nature, of a *packet-switched* nature. Assigning a dedicated (high-capacity) transmission channel for the entire connection time, although the channel is used for only a small fraction of time by some data packets, thus constitutes a waste of potential: This can't be afforded for the scarce air interface resource.

Similarly, international harmonization efforts for a truly worldwide 3G standard are ongoing. However, existing but still mutually incompatible 2G systems strongly dominated the market and lobbied local interests to such an extent that the outcome of the standardization process was a very complex and demanding one, and also reflected the political situation. Emphasis in the resulting five official air interfaces was placed on CDMA technology, which again represented is a giant transition for most of the existing 2G networks.

Even for the existing CDMA networks (e.g., IS95), the simultaneous support of the mixed traffic scenarios poses severe problems.

3.4. 4G Systems: True Multimedia

Although 3G networks are not even in commercial operation, R&D work on a successor technology, commonly called "4G," has already started.

Key characteristics seem to be

- Very high bandwidths and data rates
- Very asymmetric uplink/downlink traffic
- Data services clearly predominating
- Spotty coverage
- Adaptive antennas

After the initial hype (hyperbole) about 3G possibilities had cooled down to reality, it became obvious that the theoretical limits of this technology couldn't be realized practically.

The required financial requirements would simply be prohibitive. As of today, more and more incumbent UMTS operators have started to reduce customer expectations from 2 Mbps to 384 kbps and even 144 kbps in most for "typical" urban environments. However, the expected demand for true multimedia services is ~20 Mbps.

Also, the inherently assumed asymmetry in the network load of 2–1 for the downlink–uplink ratio does not fit the reality of multimedia applications; a ratio of 5–1 or even 10–1 seems more realistic. These discrepancies indicate the necessity for a new technology.

4. PLANNING OF WIRELESS NETWORKS

The process of planning wireless networks always entails a set of parameters that must be optimized simultaneously. Not surprisingly, some of these "global" parameters interfere each other. For example, good network coverage can be achieved not only with a large number of base stations and low network interference but also with a smaller number of sites — for the price of a resulting higher average interference. These "global" goals must thus be guided by some "local" planning goals.

As this brief example demonstrates, there is no single "optimal" planning solution but a *set* of "equally optimal" solutions. Seen from a scientific point of view, the task of planning a mobile network is a so-called multidimensional optimization problem with Pareto optimality criteria [19].

It can be shown that even subproblems of this task are computationally NP-complete. Without discussing of theoretical computer sciences in detail, this means that the effort to find such optimal network plans explodes exponentially in the size of the network and thus can be solved only *approximately* in reality and/or by simplifying the planning process partially.

Fortunately, the global conditions could be divided into several independent steps, forming the conventional-approach of network planning for 2G and 2.5G networks as was done with manual and PC-based planning systems:

The first classical step of radio network planning is to achieve the necessary coverage area of the desired area (and/or customers). In 2G and 2.5G systems site location and placement planning could be optimized in terms of coverage analysis alone and can be separated from the further QoS (quality-of-service) and GoS (grade-of-service) evaluations. In 3G, however, this will substantially change. Even in the first step of planning, the effects of the CDMA physics on coverage issues (e.g., cell breathing) will dramatically affect the planning strategy.

Basically the traffic, which represents a mix of services (bit rate, whether packet- or circuit-switched, etc.), will result in additional noise in a (W)CDMA cell and thus change the cell's size. This dependency between load in a cell and cell size will cause the cell edges to "float" dynamically and thereby lead to dropped users, if the network is designed poorly. This dependency leads to a situation in which, the well-known separation between coverage and QoS planning as used in 2G and 2.5G will no longer work in 3G systems, which is a challenge for network operators and planning system suppliers. The following sections discuss the conventional 2G and 2.5G planning approaches. We will also outline a more advanced planning algorithm that leads the way to 3G planning. In addition, we sketch how the previously mainly manually performed network planning and optimization task can be automated.

We then address true 3G systems to in a separate section. The results, illustrations, and other aspects are derived using state-of-the-art computer systems and are already used by network operators [2–4].

5. SITE LOCATION AND PLACEMENT

A conventional site is a base station with one or more antennas, where each antenna normally, corresponds to one cell of the network. Several antenna configurations have shown up in the past, while antenna space diversity is a popular configuration in GSM and also in 3G CDMA networks. However, for the planning task itself the antenna configuration is not that interesting, while the antenna characteristics (antenna patterns, azimuth, mechanico-electrical downtilt and gain) are important parameters for controlling coverage and interference [5].

In the 1990s the first planning systems involving a Geographic Information System (GIS) came into professional use. The amount of memory, hard-disk capacity, and processor speed as well as the availability and price of digital terrain models (DTMs) so far prevented an intensive use of digital maps in combination with PC based systems. Nowadays, powerful radio network planning (RNP) tools in combination with GIS and relational database systems are common.

Network planning starts by defining the sites and sectors and defining them in the tool. Multitechnology, multiband planning systems require combining different system technologies (e.g., GSM900, GSM1800, TETRA450) into one database–simulation scenario.

5.1. The Conventional Approach

Sites are still placed manually one, based on a high level of expertise and skill by the planner. A "green field" layout starts by using a regular hexagonal grid based on flat-earth propagation assumptions. This can be done by using a reference site and distributing sites of this type in a user-defined area with a specified distance between the site locations. Even in that simplified approach, defining a cellular site involves at least the following:

- Minimum site parameters
  - Site coordinates
  - Site identifier
  - Number of sectors (BTS)
- Minimum sector parameters
- Antenna parameters
  - Type
  - Azimuth
  - Mechanicoelectron downtilt
  - Antenna height
- Other parameters
  - System technology
  - Radiated power.

Modern tools [2] support the user in doing this with high-efficiency user interfaces of RNP tools and automatically factoring in terrain and other environmental information. Graphical interaction within the GIS system and the database enables the user by easy click and drop mechanisms to define the sites. The user can select from

a template of sites and place them on the geographic location, move, copy, re-arrange, create, or delete sites and/or cells and quickly evaluate these scenarios.

Figure 1 shows an example of a German city (Munich) after automatically placing a regular hexagonal site grid without optimization as the first network scenario. The next step of the radio network planner is then a manual refinement of the grid in order to account for additional conditions such as street orientations and high traffic areas, sectorization). Such a manual refinement of the network elements, taking clutter structure into account, leads to the layout depicted in Fig. 2.

In the scene shown in Fig. 2, the responsible engineer has optimized the grid layout, changing site positions, the number of sectors, antenna azimuth and distance between sites to accommodate the topology and terrain types within metropolitan Munich. The tools support this task by quick analysis of expected cell field strengths, line-of-sight (LoS) checks, and other methods see Fig. 3. To check the overall improvements resulting from his site placement, the planner will recalculate the combined coverage of the cells he/she created. Modern planning tools have different algorithms for wave propagation analysis, which is accompanied by calibration features from the network planning tool. The different propagation models used for the different cell types are discussed in Section 6.

Figure 4 shows the coverage of the Munich network (75 regular cells) based on a regular grid without optimization. One can see that in the city center the coverage level is poor because of the high degree of attenuation by building

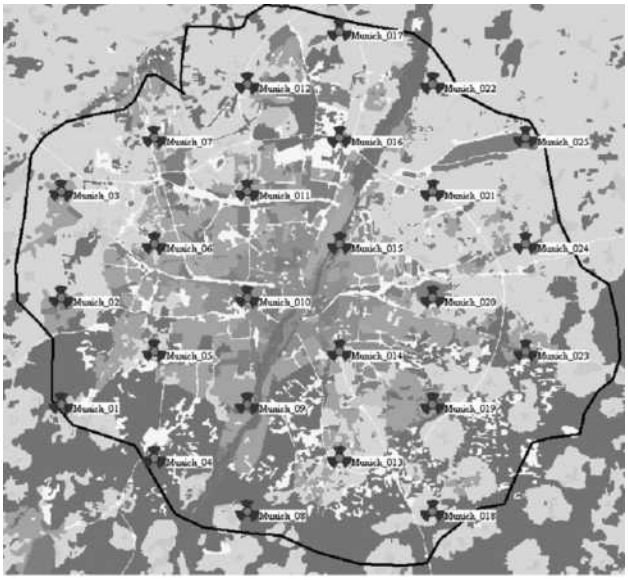


Figure 1. Regular grid layout (Region of Munich, Germany) with network area borders.

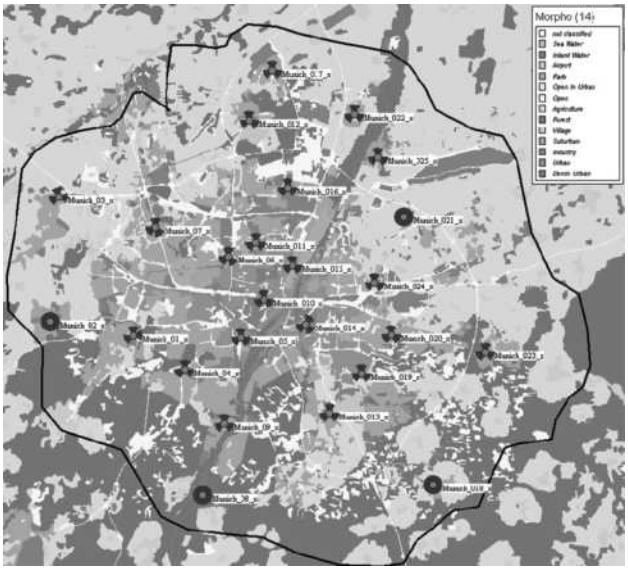


Figure 2. Manually optimized site locations taking into account morphology structure to accommodate possible high traffic areas.

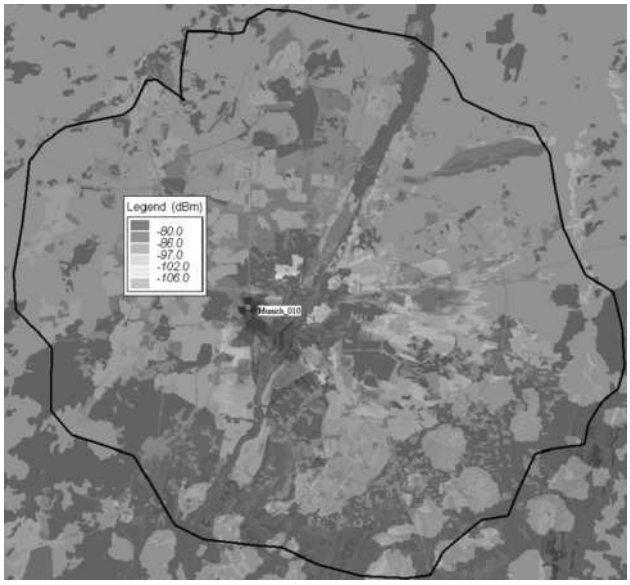


Figure 3. Single-cell field strength transparently overlaid on clutter background map.

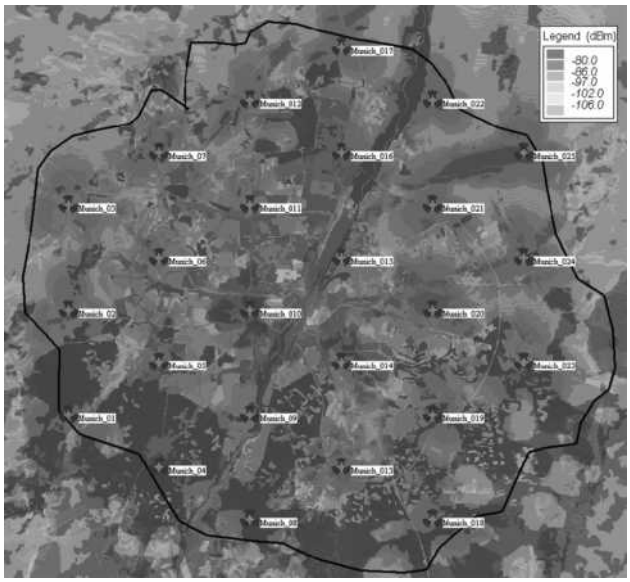


Figure 4. Networkwide coverage for regular grid layout.

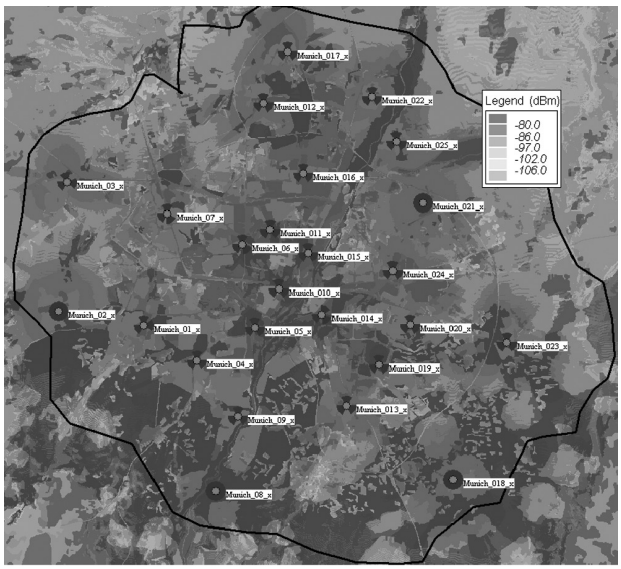


Figure 5. Coverage areas optimized for expected cell traffic especially in the city center.

structures. In Fig. 5 the coverage is enhanced according to the clutter structure and expected service areas with a lower number of cells (66 instead of 75) achieved by manual optimization of the site locations as well as changing of the site configuration (OMNI sites in lower traffic areas). Obviously, the expertise of the experienced planner paid off by a higher coverage, although nine sites less are required than in the first hexagonal grid scenario.

5.2. Automatic Site Placement

From the above paragraph it is obvious that the radio network planner's expertise and experience can be used to optimize the problem. Although the outcome seems to be fine and the number of base stations may appear to have reached a minimum, there is still plenty of room for optimization. This is because, as mentioned above, this is a multidimensional optimization problem with Pareto optimality criteria and NP-complete subproblems. Translated into simple words, this means that it is impossible for a human planner to actually find an optimal network solution. Watching how experienced network planners tackle the planning job, it seems that most of them aggressively optimize for coverage by using "promising" cell candidates first — and sticking with these, even if giving up on one cell for one or two other candidates nearby leads to better network solutions. None of them was able to plan for coverage and interference *in parallel*. The multidimensioning problem was "solved" by assuming that a low number of cells also automatically corresponds to a low interference (a heuristic that may not be the case, however). A new method has therefore been developed to overcome the time-consuming, yet suboptimal manual cell selection (choosing optimal sites from a given set of candidates), cell placement (finding the set of optimal new candidates in a "green field" situation), and cell dimensioning (finding the set of optimal parameters for a given set of cells) problem [6].

An initial network situation is iteratively improved. Several steps involving placement, selection, and parameter dimensioning are alternated and repeated. All of this is steered by a special "genetic algorithm" to break down the huge search domain into important, but practically searchable subspaces. An algorithm based on the technique of cell splitting is also used. This is intended to accommodate for traffic issues.

The boundary conditions for the local and global conditions for the objective function to be minimized are *area coverage rate*, *traffic coverage rate*, *spectral costs*, *geographic functions*, and *financial cost functions*. Here, the operator can guide the tool by defining which of the (Pareto-) equivalent solutions is preferred.

Again, the regular grid layout (Fig. 6) is used as a starting point for cell placement, while a mode exists to use the algorithm for cell selection as well (finding the optimum cells of a given set of cells) (see also Fig. 7).

For example, the coverage rate is calculated using a propagation model based on DEM data (topography and clutter) and used to calculate the influence of neighboring cells using the assignment probability of each cell in the selected network DEM (Digital Elevation Model)

$$A_{\text{cov}}(S) = \sum_{x=0}^{n-1}\sum_{y=0}^{m-1} \delta \cdot P_{\text{cov}}(x, y, S)$$

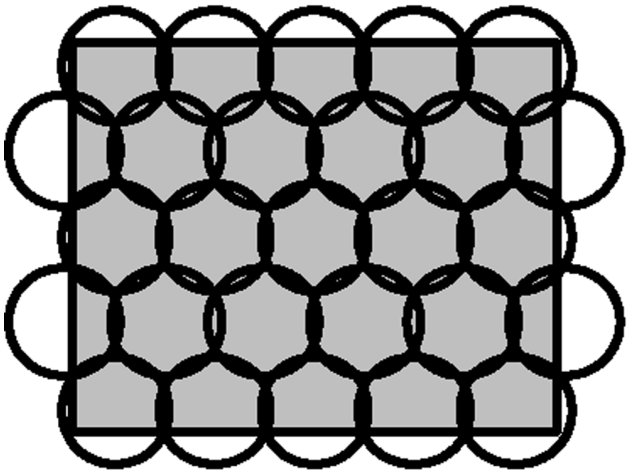


Figure 6. Regular cell grid.

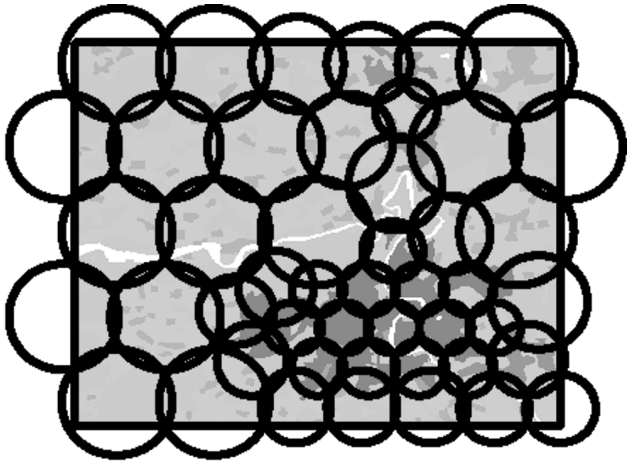


Figure 7. Cell refinement.

where δ is the pixel size and $p_{\text{cov}}(x, y, S)$ is the probability of a mobile at (x, y) being served by *any* cell in the admissible cell set S. This takes into account the handover between cells. The area coverage rate is then calculated as

$$f_1(S) = \frac{A_{\text{cov}}(S)}{A_s}$$

where A_s is the service area size. For the other costs, similar objective functions can be defined and used in a multiobjective optimization algorithm. A hierarchical workflow of optimization is shown in Fig. 8 (see also Fig. 9).

Until now one cell per site is considered and the cell size is as large as possible. What happens if the traffic demand increases? What has to be done if the existing cells cannot handle the scaled-up traffic? The method commonly used in practice is cell splitting, that is artificially reducing the existing cell sizes and adding new cells in between. The problem is then to find optimal base station (BS) sites for new cells and optimize dimensions of both original cells and additional cells so that the growing demand can be met effectively, while offering minimum disturbance to the existing network structure. Nominal cell splitting is illustrated in Fig. 10. Initially, the largest possible cell sizes are used, one cell per site. In the next step, a cell is divided into a number of sectors. Here only the three-sector case is considered. Each sector is served by a different set of channels and illuminated by a directional antenna. The sector can therefore be considered as a new cell. As a consequence, there are three cells per site using the original BS sites. BSs are located at the corner of cells, as shown in Fig. 10b. Now the number of sites is still the same, but the number of cells is 3 times higher than before. The following step is to do further cell splitting, specifically, reducing the size

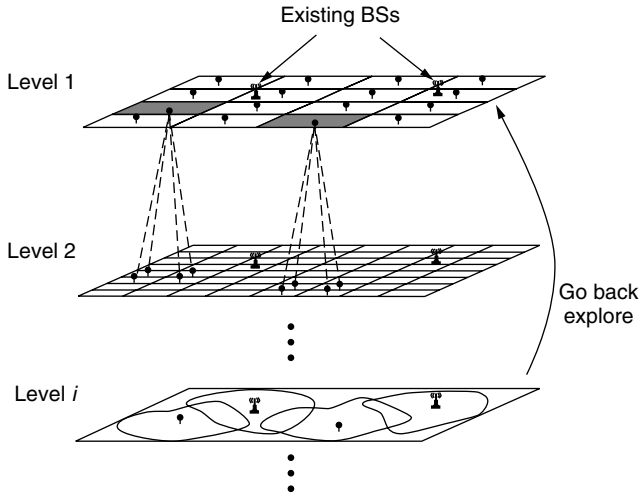


Figure 9. Structure of the hierarchical approach.

of existing cells and adding new cells. As can be seen in Fig. 10c, the former sites are still used in the new cell plan, but additional sites are now required for serving new cells.

Such algorithms will be integrated into modern network planning tools in order to accelerate network optimization. These are now the edge of modern research and will influence the future planning process.

6. PROPAGATION MODELING

Propagation modeling is the key issue for proper network planning. All further results such as coverage probability, interference, and frequency assignment depend directly on the quality of propagation prediction. Various models

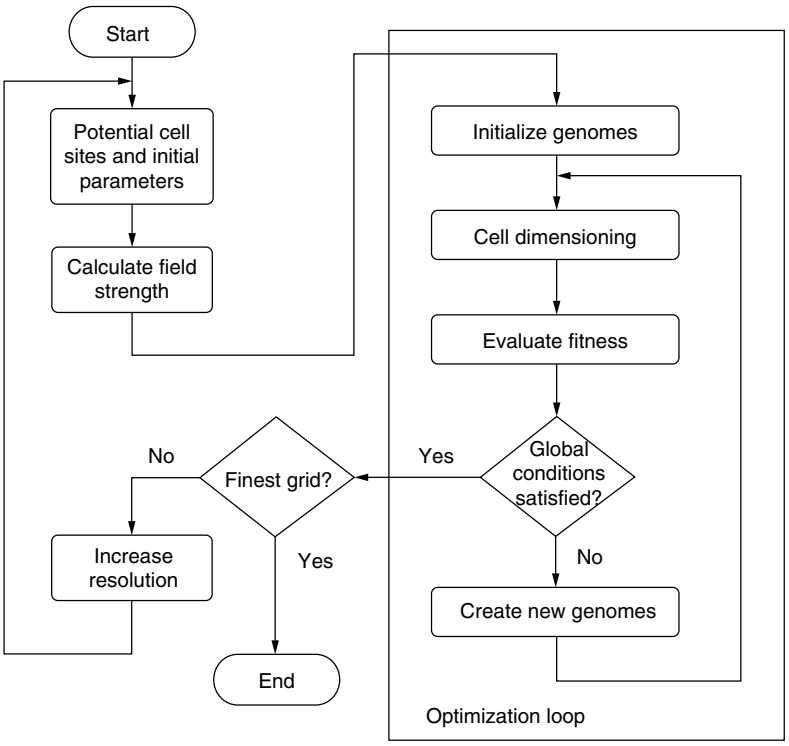


Figure 8. Hierarchical optimization process.

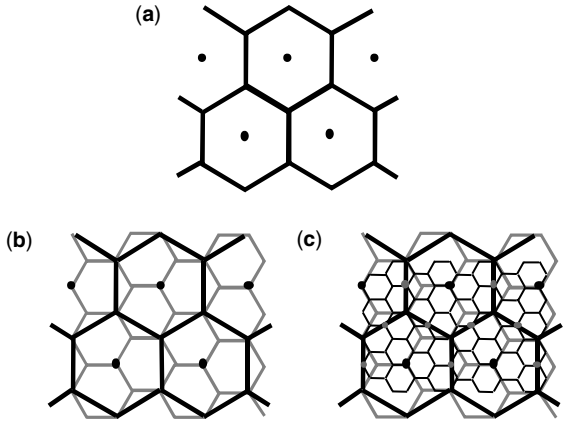


Figure 10. Nominal cell splitting: (**a**) initial cell plan; (**b**) phase 1 — each cell is divided into three cells, using original sites; (**c**) phase 2 — old cells are reduced, new cells are added, requiring additional sites.

have been developed for different frequency ranges and services. In the area of cellular (mobile) communications, the best known models are the Okumura–Hata and Walfish–Ikegami models, which have both predominated in the technology of 2G planning tools since 1980 or so.

6.1. Macrocell–Minicell–Microcell

Modern RNP tools can accommodate the various types of models for different cell layers. The macrocell layer still represents about 90% of the whole network planning; the rest is divided into microcells (cell ranges of ~100 m) and minicells (cell ranges ~500 m–2 km). The most important propagation calculation to be discussed in this section is the macrocell and its calibration, as this model is still the most widely used one in order to reduce computation time and cost for complex building data.

6.2. Model Calibration

The most widely used propagation model for macrocells is a parametric one such as the Okumura–Hata (OH) model. Many investigations and measurement campaigns have been performed in different frequency ranges and environments in order to adapt those simple models to the desired services. Because of the parametric nature of these models, calibration is quite simple and serves as the starting point for the network planning. Model calibration always compares measured drives with the predicted values. Manual, semiautomatic as well as fully automatic calibration techniques are used. The quality of calibration depends on the measurements taken for specific cells and on the quality and resolution of the topological and clutter data used in the tool used for calibration. Care must also be taken to avoid using only the transmitter point only, but the whole profile path to the receiver for conducting the calibration. The parameters of the OH model are well known from various study groups [14] for the frequency ranges around 900 and 1800 MHz (see also Fig. 11). The parameters to be optimized are mainly the clutter correction data, the gain, and the height of different land-use classes (see Fig. 12), and, to some extent, the other

Parameter	Nom. Value	min	max
a1	46.30	40	50
a2	33.90	28	40
a3	–13.80	–20	–8
b1	44.90	35	55
b2	–6.50	fix	fix

Figure 11. OH parameters and their nominal ranges for calibration (1800 MHz).

Clutter Class	Gain dB	Range (+/–)	Height (m)	Range (+/–)
not classified	0	0	0	0
Sea Water	28	4	0	0
Inland Water	27	4	0	0
Airport	6	4	5	2
Park	6	3	7	3
Open In Urban	8	8	0	0
Open	23	2	0	0
Agriculture	8	6	3	2
Forest	20	6	12	6
Village	15	8	7	3
Suburban	12	4	10	5
Industry	2	2	25	10
Urban	8	4	14	7
Dense Urban	6	3	17	9

Figure 12. Clutter parameters and their nominal ranges for calibration in Europe (1800 MHz).

parameters of the Hata equation. Typical parameters to be calibrated for a macrocell model such as OH for one frequency range (e.g., GSM1800):

$$L = a_1 + a_2 * \log_{10}(f) - a_3 * \log(h_{\text{eff}}) + (b_1 - b_2) * \log_{10}(h_{\text{eff}})(\log_{10} d)$$

where f is the frequency of operation and h_{eff} is the effective height. Several methods exist for the determination of effective height between BS (base station), and MS (mobile station):

- Height of BS antenna above ground
- Height of BS antenna above mean sea level
- Height of BS antenna in relation to the effective terrain height according to ITU
- Height of BS antenna in relation to the effective terrain height over the entire profile
- Height of BS antenna in relation to the MS antenna
- Effective height/distance ratios by rotation of the terrain against its ascent

Measurement drives are imported into modern RNP tools [2] where multiscreen, coupled cursor optimization features are used to calibrate the propagation model/clutter parameters for the averaged measured values (average using time or space windows smoothing the measured drive) (see Figs. 13–14).

The typical outcome for calibrating macrocell models are a mean value of around 0 dB of the difference and a standard deviation of 5–10 dB.

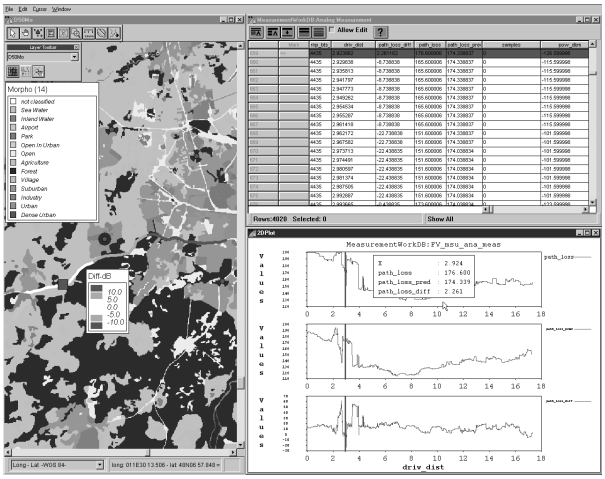


Figure 13. Measurement evaluation screens in modern RNP tools.

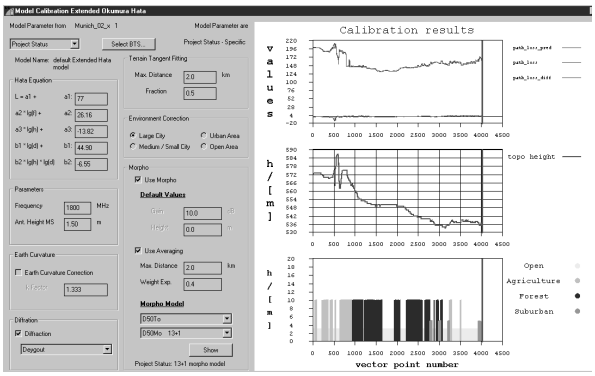


Figure 14. Calibration engine GUI, (graphical user interface), example.

7. 2G NETWORKS (GSM)

Let us summarize the strategies for 2G network planning first. The planning process, conducted either manually or automatically, consists of a number of *separate* planning steps:

- Determine a (an initial) cell selection
- Issue a coverage predictions (physical propagation modeling)
- Derive the "best server" areas
- Determine the traffic load per cell
- Calculate the number of frequencies needed
- Do a frequency plan assignment
- Analyze the resulting interference situation
- Repeat the loop for network optimization

We address some of the more important results required for this process in the next section. Please note the critical prerequisite that these steps are assumed to be *independent* of each other to a large extent; that is, if the traffic load exceeds a cell's capacity, it can be extended by adding another transceiver to the cell. If the required frequency is chosen from a previously calculated "candidate" set, the whole network can be assumed to not have changed significantly. This assumption will not hold true for 3G networks (or 2G, CDMA-based networks) and thus, require additional effort.

The strategy of applying planning tools to solve the different tasks can be generally divided into two basic approaches, a so-called "deterministic approach" used by most RNP tools and a more enhanced "probability based" approach. The following sections describe the different strategies.

7.1. Deterministic Approach

Traditional RNP tools deal with coverage aspects only, where the propagation model calculates for each possible mobile station (MS) location the power level according to the appropriate model. In the deterministic approach (see Fig. 15), only these cell specific coverage results are used to evaluate networkwide coverage and interference. An advantage of the deterministic approach is the simple calculation dependencies as the only inputs for networkwide evaluations are the cell-specific power files. Another advantage is the rapid computation time even for networks with thousands of base stations.

The input files (the cell power results) are combined with a network engine to achieve the following results described.

7.1.1. Coverage-Based Results. The following results are typically used to plan or to optimize the coverage of the network:

Input parameters — network access level [e.g., −98 dBm and maximum allowed timing advance (TA) for TDMA systems].

Generated outputs — only pixels that have a power level higher than the network access level and are inside the TA of the corresponding TDMA system (GSM: TA ∼35 km) are considered.

7.1.1.1. Maximum Server. For each pixel in the calculation area shown in Fig. 16, the transmitter produces the strongest power level compared to all others at that pixel.

7.1.1.2. Networkwide Coverage. Figure 17, shows the strongest power level in dBm for each pixel in the calculation area, indicating the maximum value of all contributing signal sources.

7.1.1.3. Strongest Interferer at Strongest Server. Figure 18 shows for each pixel in the calculation area the transmitter that causes the highest interference relative to the serving cell.

7.1.2. Interference-Based Results. The following results are typically used to plan or optimize the interference (service) of the network: *input parameters* — network access level (e.g., −98 dBm), maximum TA (e.g., 35 km in standard GSM; i.e., maximum signal delay that can be compensated by the system to remain synchronized) and

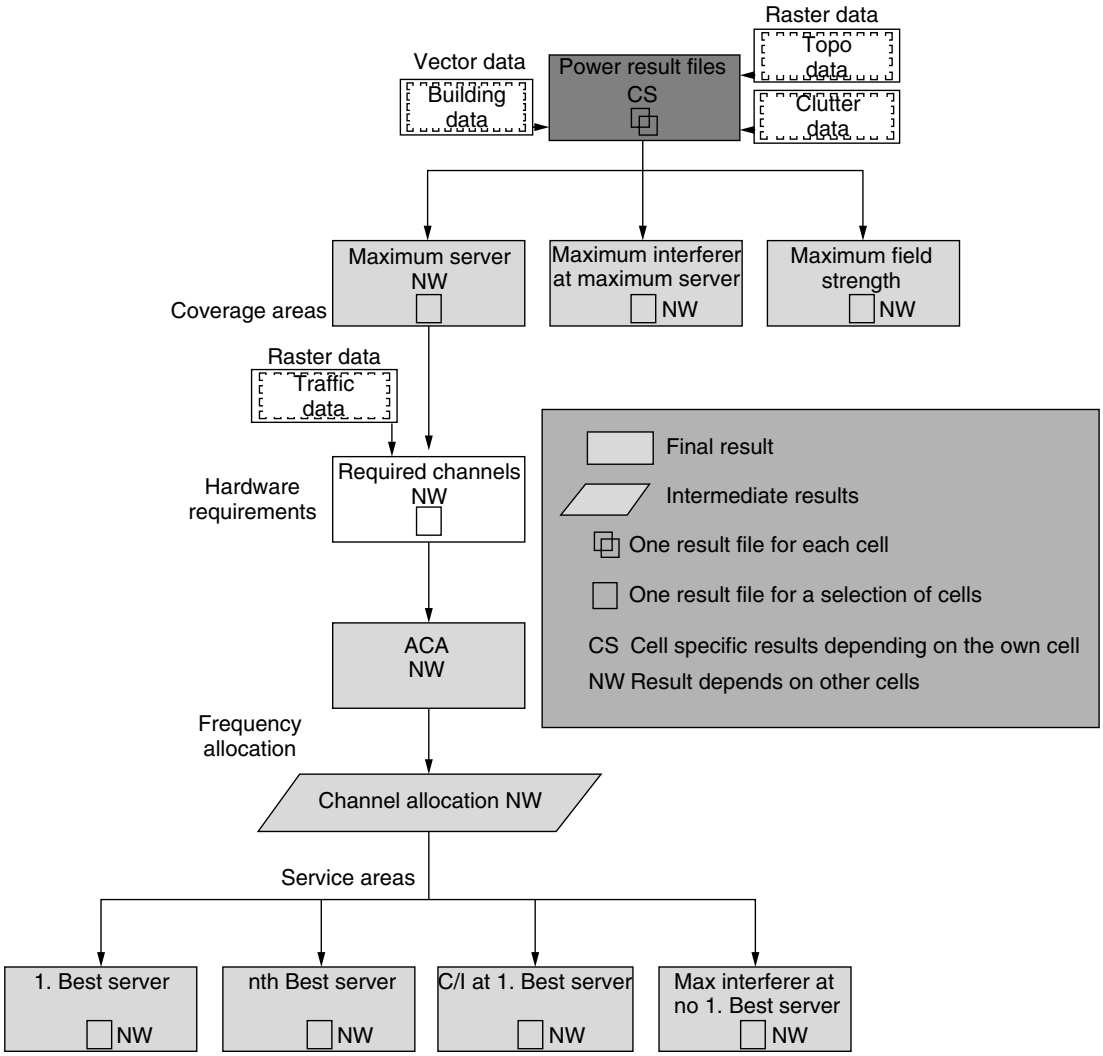


Figure 15. Calculation workflow in the deterministic approach.

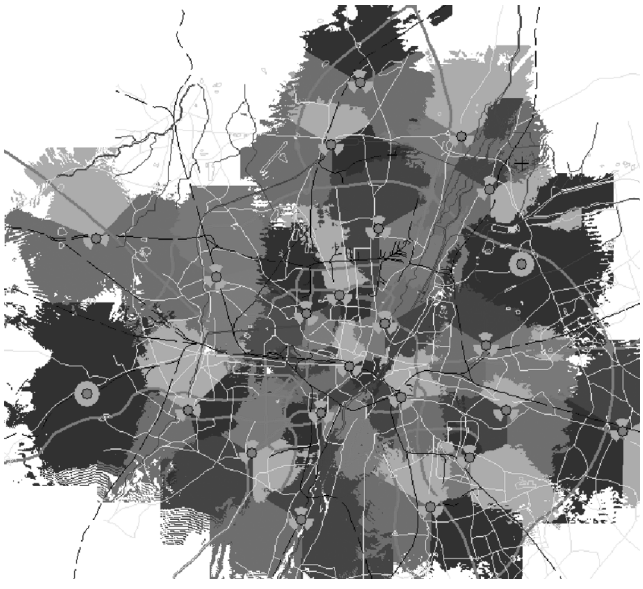


Figure 16. Maximum server result.

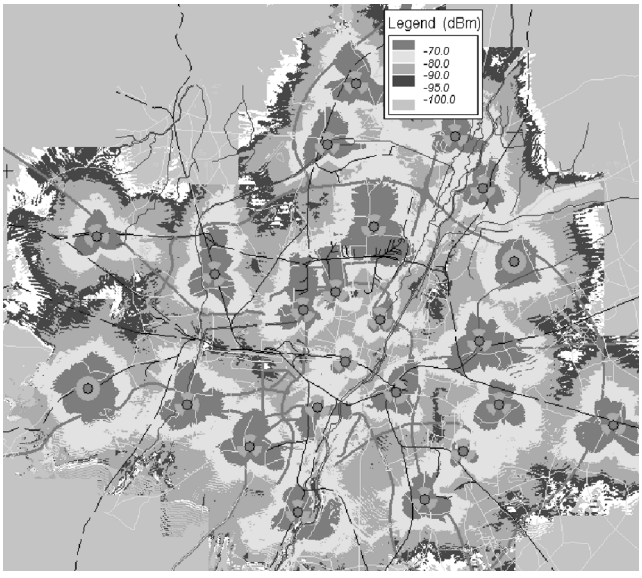


Figure 17. Networkwide coverage result.

	Co	1. Adjacent	2. Adjacent	3. Adjacent
Channel distance	0	1	2	3
Ratio/dB	9	–9	–41	–49

Figure 18. Protection ratios for GSM system.

protection ratios for co- and adjacent channel interference. For the GSM system, the values are listed in Fig. 18.

The table in Fig. 18 should be read as follows. In case the server produces at a pixel in the calculation area a power level of −65 dBm (say), then this server is assumed to be interference-free and can provide service if a potential interferer at the same channel ("co channel") is not stronger than −74 dBm. A potential interferer having two channels' distance ("2 adjacent") is allowed to produce a level of −14 dBm. Those protection ratios are also used as a quality criterion for the frequency planning.

*7.1.2.1. **Best Server/Best Server.*** This type of result shows for each pixel in the calculation area the transmitter that causes the first or nth strongest field strength at that pixel *and* is not disturbed by interference (according to the protection ratios). For this, only pixels with a serving field strength above a certain threshold *and* lying inside the timing advance range (relative to the serving transmitter) are considered. For example, if $n = 3$, the user will get three results. The first is the first best server, second best server, and so on.

*7.1.2.2. **Carrier-to-Interference Ratio.*** Figure 19 shows a QoS (quality-of-service) type of result. A high carrier-to-interferer ratio (C/I in decibels) ensures a high quality of the connection. The C/I at best server determines the carrier-to-interferer ratio C/I (in dB) for each pixel that fulfills the best server criterion in the calculation area. Therefore, in the result window all pixels are colored (served) according to the C/I ratio in dB. QoS is an important criterion for network operators as high-quality voice and data connection is a key issue to attract potential customers to a specific operator.

*7.1.2.3. **GoS (Grade of Service).*** Another important point is the so-called grade of service (GoS) of the network. Even if the network is excellent in terms of interference, a shortfall of system equipment can dramatically reduce the performance of the cells. The expected number of users producing the air traffic is a further key input to network planning. From marketing surveys an estimated load is extracted and a traffic map can be generated. Load of a (circuit-switched) telephone network is described using Erlang's [10,11] formulation for blocking and queuing systems (Fig. 20). GSM is a typical blocking system, where TETRA, for instance, is a queuing system [15–17]. The dependency of Erlang B (blocking) and Erlang C (queuing) and the equations can be found, for example, in Ref. 11.

The Erlang B formula expresses the relation between the expected traffic in a cell and its hardware, and in a GSM system, the number of time slots necessary to carry the traffic in the cell.

$$P_{\text{block}} = \frac{\dfrac{A^n}{n!}}{\displaystyle\sum_{i=0}^{n} \frac{A^i}{i!}}$$

This equation describes the relationship between three variables: the blocking probability P_{block}, the traffic load A (in erlangs), and the number of channels n. Obviously, the blocking probability increases with the traffic load and decreases with the number of channels, but *not* linearly, as shown graphically in Fig. 21. Simply stated, this nonlinearity means that an operator runs out of the last 10% (say) much network capacity a much more quickly than he/she would for any previous 10%, making capacity

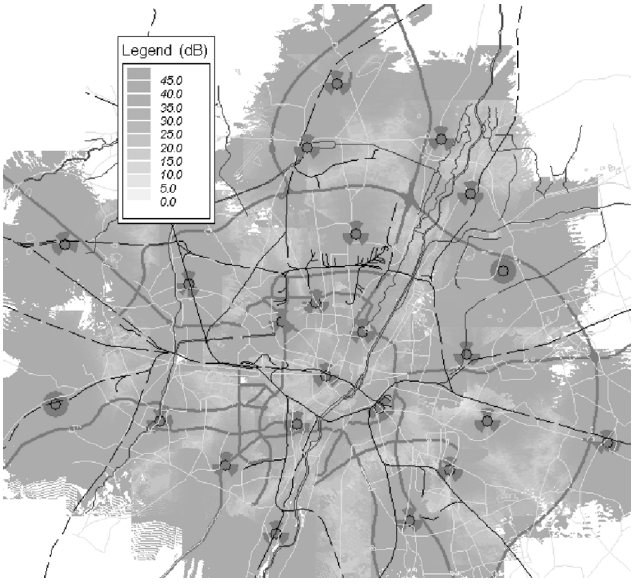


Figure 19. C/I result for QoS optimization.

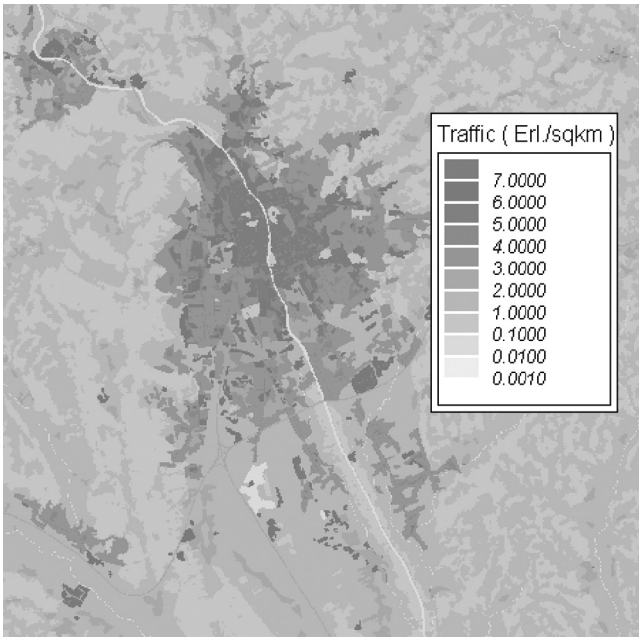


Figure 20. Traffic layer in Erlang Formula (in km^2).

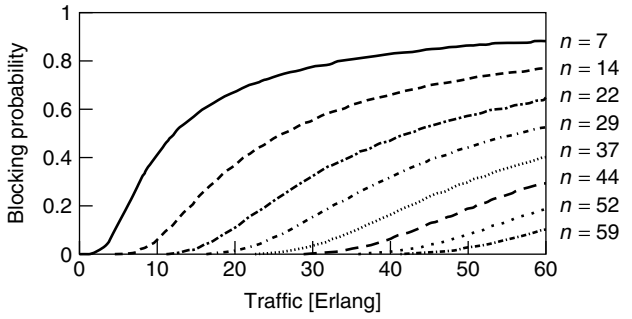


Figure 21. Example of Erlang B curves.

planning a crucial aspect of network planning. Because of its importance, the formula has been widely tabulated. Readers are referred to Lee [7] for a complete Erlang B table. By knowing any two of the three variables (A, n, and P_{block}), one can derive the third.

Assuming that the average conversation time is T seconds and the number of calls per subscriber at the busy hour is λ, the traffic produced by an average subscriber is

$$a = \frac{\lambda \cdot T}{3600}$$

and can be interpreted as the fraction of time that each user occupies a channel. A typical mobile European customer is loading a cell with about 30 millierlangs, where this figure strongly depends on the region and habit of the subscribers. The load of a subscriber varies between roughly 10 and 80 millierlangs throughout the world.

The number of time slots in a TDMA system is correlated with the number of transmitters (TRXs) for a given sector.

Typically network operators will equip their networks mostly with predefined site configurations to minimize the materials management and to overcome the problem of getting suitable marketing — and therefore traffic information. Typical site configurations for an initial GSM network (eight time slots per physical channel) are listed in Fig. 22.

7.2. Probability Approach

The convention approach described in the last section is a bit simplistic — a pixel is said to be either served (completely) by a cell or not be served by a cell (at all). This assumption does not hold true in reality, especially for pixels at the very edge of its best server area. Replacing the digital yes/no "being served" information by a *probability* of being served translates the whole planning process into the fields of probability theory and allows more detailed insight into the planning process. The major difference compared to the deterministic approach is the conversion of deterministic cell power results into assignment probabilities. Assignment probability results are still cell-specific (with one result file per cell) but dependent on the power level of each cell *and* other cells *and* the neighborhood relations and parameters between cells. One new important point addressed is the handover simulation (which has been completely neglected in the old deterministic approach). Cellular systems make it necessary to hand users from one cell to the next cell by handover (handoff) strategies in order to continuously serve their moving users. The probability approach uses the same strategy a mobile station would use, measuring and reporting all received field strength from the neighboring cells and the base station, deciding whether the mobile is handed to one cell or to the other. The dependencies of calculation for the probability approach are shown in Fig. 23.

The calculation of the assignment probability uses the following equations to convert the cell power results to an assignment probability. The assignment probability of a cell b_i at location (x, y), $p_{\text{ass}}(b_i, x, y)$, is defined as the probability of a MS at (x, y) being served by the cell b_i. It is obvious if the MS is served only by one cell that the probability will then be 100%:

$$p_{\text{ass}}(b_i, x, y) = \frac{F(b_i, x, y) - F(b_1, x, y) + DEF\_HO\_MARGIN}{\sum_{j=1}^{N} (F(b_j, x, y) - F(b_1, x, y) + DEF\_HO\_MARGIN)} \cdot P_{\text{tot}}$$

where
$p_{\text{ass}}(b_i, x, y)$ = assignment probability of cell b_i at coordinates x, y
$F_{\text{ass}}(b_i, x, y)$ = Power level of cell b_i at coordinates x, y
p_{tot} = Sum of all assignment probabilities of all cells in the selection
DEF_HO_MARGIN = handover margin for neighborhood relations

This is the basic difference between the deterministic and the probability approach. The cell-specific assignment takes into account the handover margins and therefore simulates a more realistic behavior of the mobile in the network, especially at cell edges. The following figures show examples for the cell-specific assignment. In Fig. 24 a single cell is calculated using an omnidirectional antenna on a flat terrain showing 100% probability that a mobile inside the red area is served by this cell (because no other cell is serving that area). The edge of the red area is in this case determined by the minimum access level of the MS which was set to −95 dBm.

Figures 25–27 show the assignment of each cell in a small network. These more detailed results are used for

	Sectors	Antenna	TRXs/Cell	TRXs(total)	Capacity	Subscribers/Site
Rural 1	1	1 × 360 deg	2	2	8 Erl.	250
Rural 2	3	3 × 120 deg	2	6	35 Erl.	1100
Urban	3	3 × 90 deg	3	9	55 Erl.	1800
Dense Urban	4	4 × 60 deg	3	12	80 Erl.	2600

Figure 22. GSM site configurations.

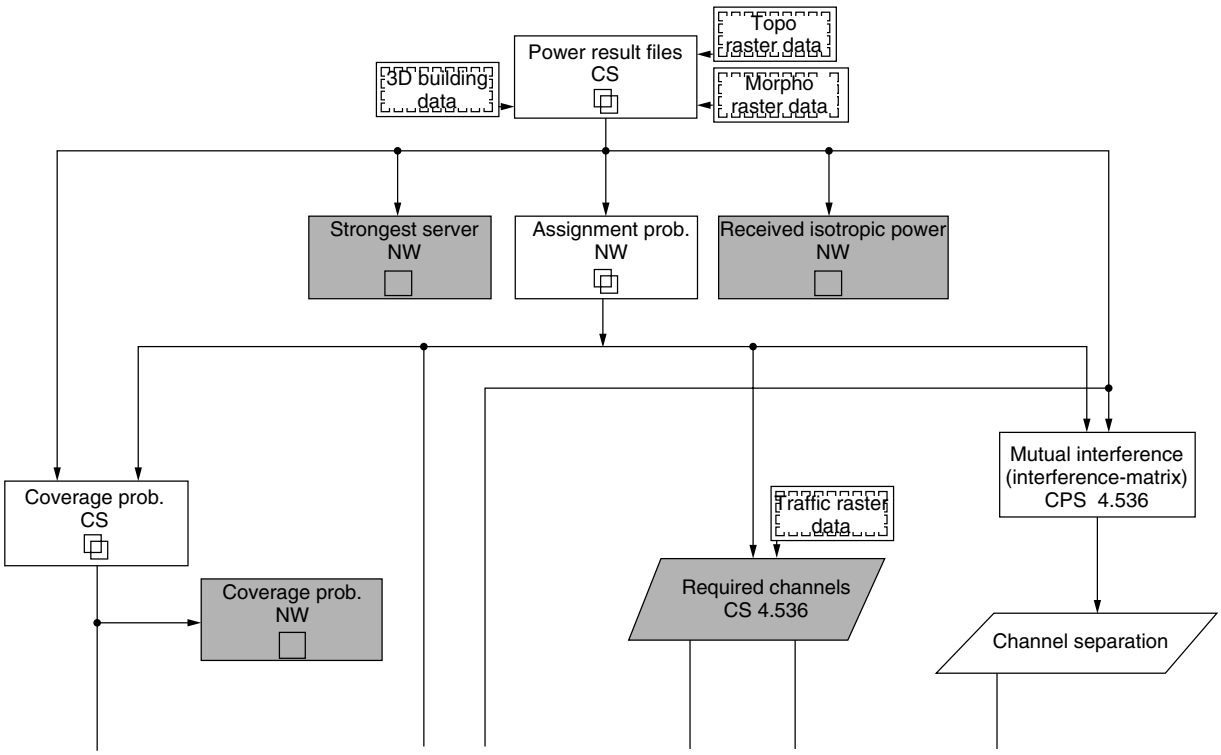


Figure 23. Dependencies of calculation; major differences to deterministic approach.

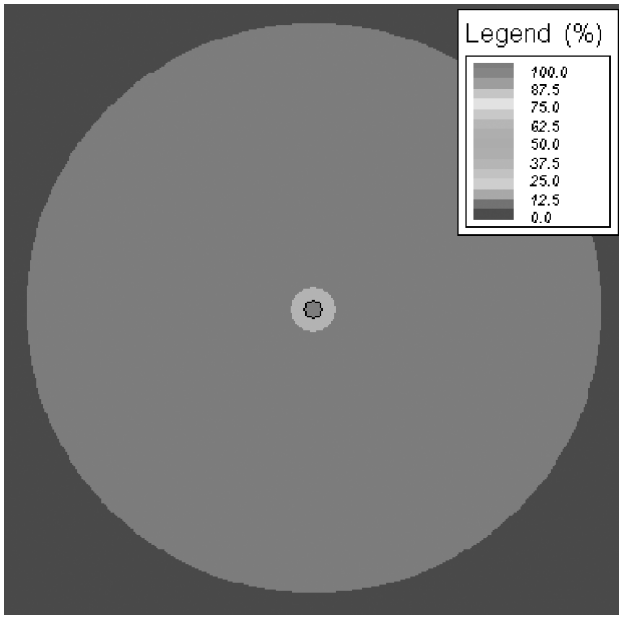


Figure 24. Assignment of a single cell.

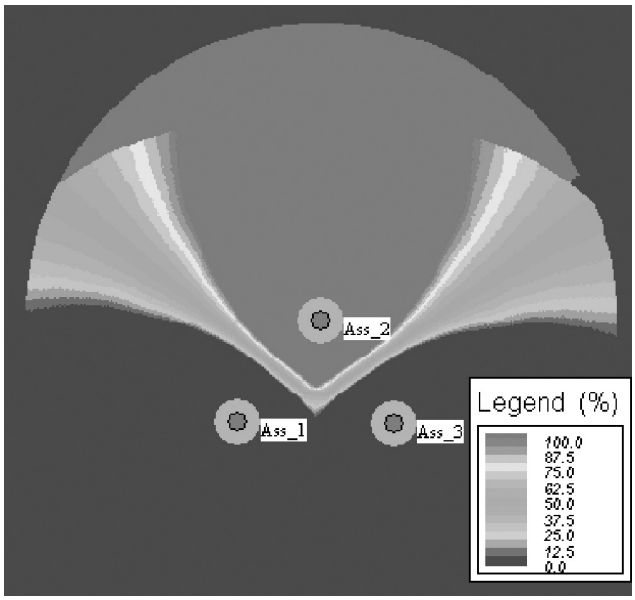


Figure 25. Assignment of cell Ass_2 while Ass_1 and Ass_3 are also serving.

further evaluations-such as coverage probability, mutual interference, and channel separation.

The coverage probability is determined by

$$P_{\text{cov}}(x,y) = \text{erf}\left(\frac{F(x,y) - F_{\text{thr}}}{\sigma(x,y)}\right)$$

where $P_{\text{cov}}(x,y)$ = coverage probability
erf = error function
$\sigma(x,y)$ = standard deviation
F_{thr} = power threshold
$F(x,y)$ = power level at x, y

Figure 28 shows the networkwide coverage probability of this simple network. The coverage probability is defined as the probability that the field strength of the signal from the BTS is greater than a given threshold.

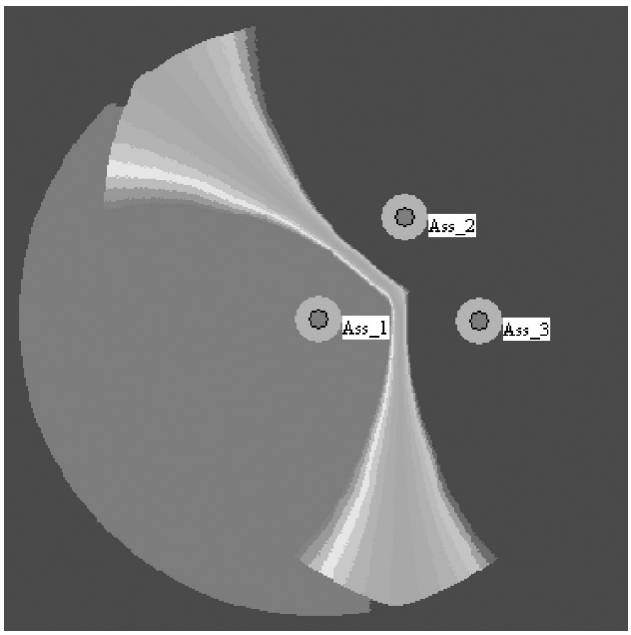


Figure 26. Assignment of cell *Ass_1* while *Ass_2* and *Ass_3* are also serving.

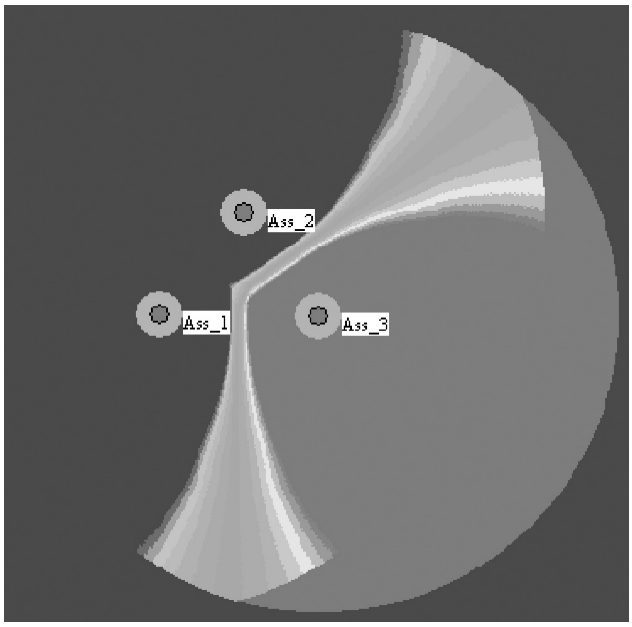


Figure 27. Assignment of cell *Ass_3* while *Ass_1* and *Ass_2* are also serving.

In a realistic network the coverage probability will resemble Fig. 29 [where F_{thr} was set to -95 dBm, $\sigma(x, y)$ to 6 dB].

7.3. Frequency Assignment and Optimization

Worldwide research activities have been done on the channel assignment problem (CAP). It is also called frequency assignment problem (FAP) in some literature [8] and has been shown to be NP-complete for subproblems. Once the cell sites and dimensions are determined, there is a lower-bound, which the minimum frequency spectrum

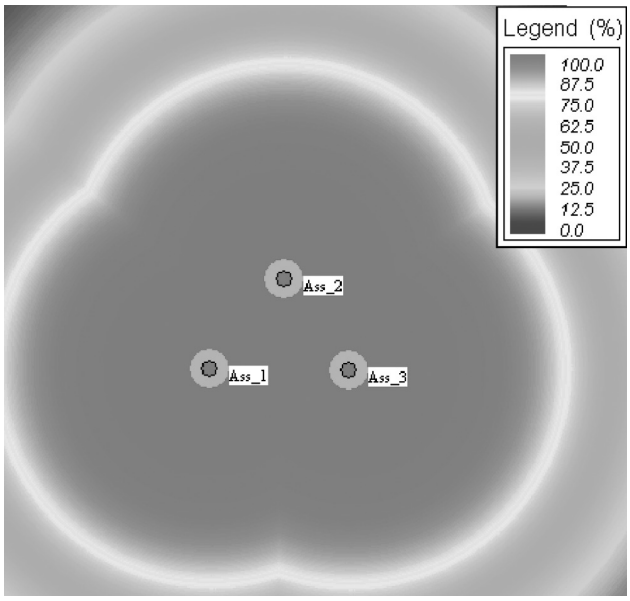


Figure 28. Coverage probability.

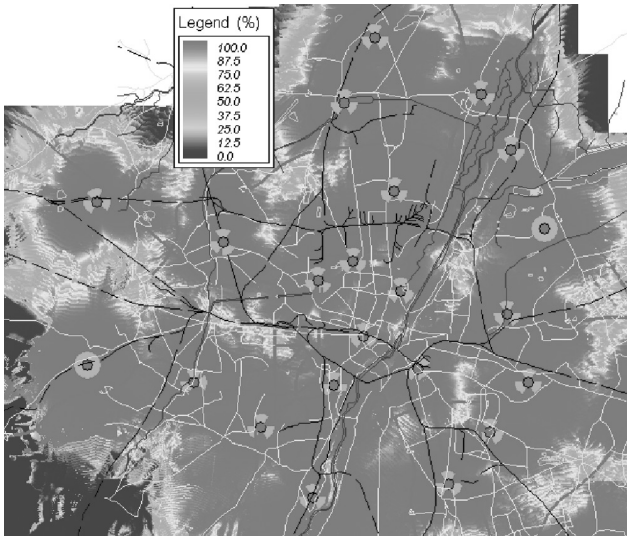


Figure 29. Coverage probability of Munich network.

required for assigning each cell a sufficient number of channels [9]. The base for solving the CAP is the so-called channel separation matrix, where the minimum channel separation between all cells in the network under evaluation are stored. This is a pair-to-pair relationship as each cell is considered as a server and all remaining cells considered as potential interferers. The boundary conditions for a channel assignment problem are: (see also Fig. 32).

7.3.1. Global Conditions. These are as follows:

Allowed frequency band (channels) — for example, GSM 900 (1-124), GSM1800 (512-885).

Channel types — some operators additionally divide their assigned spectrum by channel types (traffic or control channel to further enhance the quality

as control channels are separated from the traffic channels), for instance, to protect the more "valuable" control channels (BCCH) compared to the regular traffic channels.

Figure 30 shows a typical division of the frequency band in the GSM 900 range for one of two GSM 900 operators in a country.

7.3.2. Network-Specific Conditions

Channel separation between all cells can either be calculated or set manually.

Neighborhood relations — cells in the handover list typically will have additional channel separation requirements

Number of required TRXs: these are either calculated (by Erlang equations and assumed spatial traffic load) or manually set; see Section 7.1.2.3.

Neighboring countries — coordination issues, frequencies that are not allowed to be used at country borders, international or mutual agreements between foreign regulators or operators (Vienna agreement [18]). In a planning tool coordination issues can be set up by forbidden channels for cells radiating toward the border.

The outcome of a successful channel assignment is a channel/frequency plan that fulfills the boundary conditions or minimizes cost functions. The channel plans are stored in the RNP tools database as shown in Fig. 33. The allowed frequency spectrum for this GSM 1800 example network was from 600 to 725 and from 800 to 860 as defined in the RNP tool as global conditions (see Figs. 30 and 31 as cell-specific conditions).

The results in Figs. 34 and 35 show a test case of a C/I calculation for the network of Munich in the theoretical case that all TRXs would be operating on the same frequency compared to the C/I after an automatical frequency assignment.

8. 2.5G NETWORK EXTENSIONS (HSCSD/GPRS/EDGE)

GPRS (general packet radio service), EDGE (enhanced data rates for GSM evolution) and HSCSD (high-speed circuit-switched data) have been designed primarily as upgrades to the well-known and heavily deployed GSM standard. The same applies to IS136+ and IS136 HS in the case of the IS136 standard. In the starting phase of GSM and IS136 systems, data transmission issues were of minor importance compared to voice transmission. Beside this fact, the maximum transmission speed of 9.6 kbps that plain GSM and IS136 offered, appeared to be sufficient and was comparable with analogue wireline modem speed at the time when the white papers were drafted. In the 1990s, in particular with the increasing usage of the internet, higher data rates were provided on the fixed modem lines while GSM and IS136 still stuck with the 9.6–14.4 kbps.

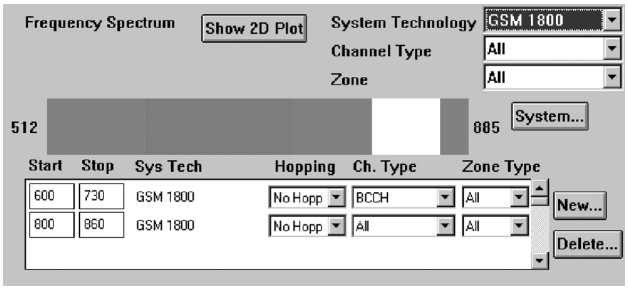


Figure 30. Global conditions for CAP.

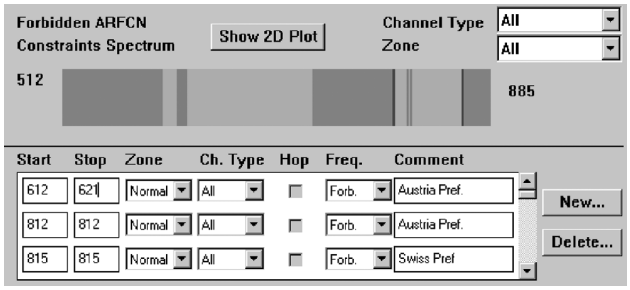


Figure 31. Cell-specific boundary conditions for channel assignment.

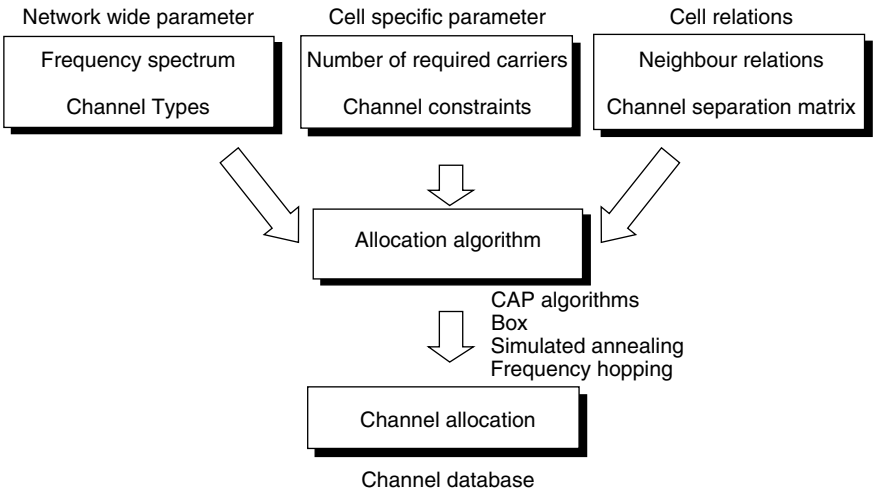


Figure 32. CAP inputs.

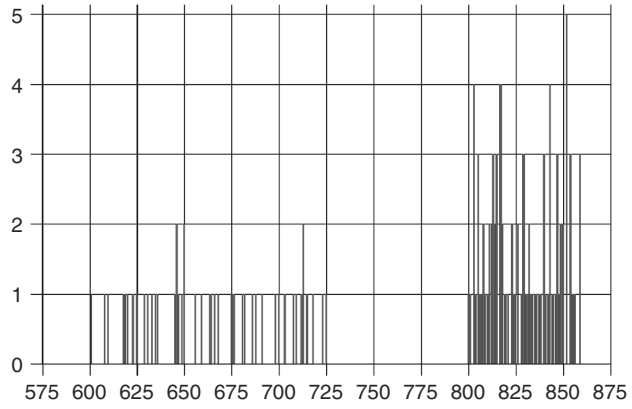


Figure 33. After successful channel assignment for a GSM 1800 system.

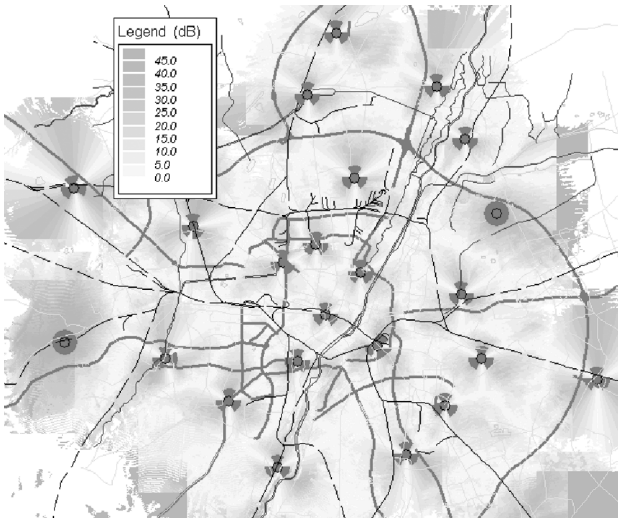


Figure 34. C/I for Munich network before channel assignment.

An optimistic timeline for deployment of data services is shown in Fig. 36.

Because of the complex change of migrating 2G networks to the 3G world, there were strong demands from the market to enable higher data rates in the already existing 2G infrastructure. Pressure increased not to wait for the "mobile Internet" up to the year 2005, where UMTS services are currently expected to be really deployed and ready for widespread use, including handset availability. The success of the possible 2.5G technologies depends largely on the support from system and handset manufacturers and the real-time schedule for 3G availability. Commercial amounts of the first 2.5G handsets

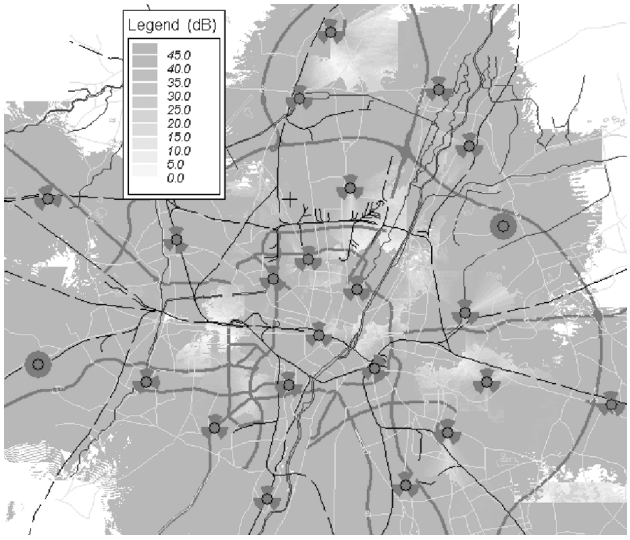


Figure 35. C/I for Munich area after channel assignment.

are currently (summer 2001) sold on the market. The first network operator in Germany (D1) has officially launched GPRS services, but the data rate offered is still about only half the speed of an analog modem of the fixed network side and available only in specific areas. EDGE in this respect is a consequent step further using existing GPRS equipment for higher data rates (up to theoretically 384 kbit/s). Simply put, GPRS is a key enabler technology to use the voice and circuit-switched GSM infrastructure for packet-switched data services. The major point for enabling this data service is an adaptation of the GSM backbone network for the coexistence of both technologies. The major advantage for all network operators is that the main parts of the existing infrastructure, especially site installations, can be used further on and reduce investment costs to a minimum. Possible data rates offered depend on the number of packet-switched users in the cell and on different modulation types, so-called coding schemes (CSs). Depending on the hardware suppliers, different QoS ratings are necessary to enable a specific data rate. However, the limits (maximum possible data rates of GPRS and EDGE [also referred as enhanced GPRS (EGPRS)] concerning the data rates are shown in Figs. 37–39.

EGPRS is expected to introduce nine new coding schemes, where the higher coding schemes use different modulation in the same timeslots. The applied modulation for higher EGPRS data rates is 8-PSK (phase shift keying) and contains phase and amplitude modulation. Thus, 8-PSK-modulated signals in EGPRS need to be transmitted with a smaller power than in GMSK

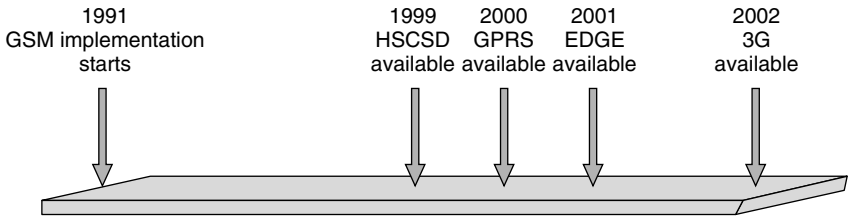


Figure 36. Optimistic view of enhanced data service deployment.

	1 Timeslot	2 Timeslot	8 Timeslots
CS-1	9.2 kBit/s	18.4 kBit/s	73.6 kBit/s
CS-2	13.55 kBit/s	27.1 kBit/s	108.4 kBit/s
CS-3	15.75 kBit/s	31.5 kBit/s	126 kBit/s
CS-4	21.55 kBit/s	43.1 kBit/s	172.4 kBit/s

Figure 37. Possible coding schemes and data rates for GPRS.

Coding Scheme	Modulation	Throughput/Timeslot
MCS-1	GMSK	8.8 kbit/s
MCS-2	GMSK	11.2 kbit/s
MCS-3	GMSK	14.8 kbit/s
MCS-4	GMSK	17.6 kbit/s
MCS-5	8-PSK	22.4 kbit/s
MCS-6	8-PSK	29.6 kbit/s
MCS-7	8-PSK	44.8 kbit/s
MCS-8	8-PSK	59.2 kbit/s
MCS-9	8-PSK	59.2 kbit/s

Figure 38. Possible data rates for EDGE/EGPRS.

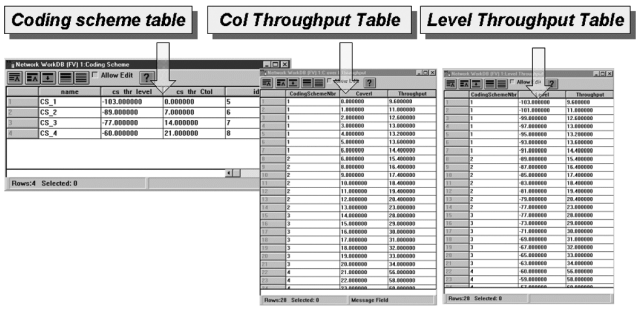


Figure 39. Hardware-dependent GPRS parameters.

(Gaussian minimum shift keying) as is currently used for GPRS and GSM. Otherwise, the respective output power amplifier would be driven to nonlinear operation, which would then result in a garbled signal. Also, 8-PSK is more error-prone than is GMSK. Therefore, the coverage of a cell running EGPRS will shrink and this fact must be accounted for in the network design.

What do the tables in Figs. 37–39 mean? For reasonable data rates for a single user, it is necessary to combine (reserve) at least two time slots in each cell used for GPRS. The coding scheme that will be used during the connection is adapted according to the QoS currently measured for this connection. Therefore a high C/I enables a higher coding scheme and thus a considerably higher data rate. However, even for high-quality networks this will increase efforts in deploying new TRXs on existing sites in order to assure a data rate at least comparable to analog fixed-line modems. As mentioned at the beginning of this section, the coding scheme that can be applied strongly depends on the hardware performance. So, for a RNP tool, it is mandatory to scope for this flexibility in the database. A possible solution is to have a database allowing the user to input lookup tables for different hardware vendors.

The possible data rates for the different coding schemes (throughput in kbps) are stored as a function of power level and C/I for each hardware unit used in the network. The calculation is divided into two streams: a maximum possible throughput depending on the power level of each cell and a best possible based on C/I calculations of the current fully loaded network.

Note, however, that the theoretical bit rate limits of GPRS, EDGE, and EGPRS are far from reality because of the transmission errors, necessary amount of error correction overhead, handset restrictions, and sharing the capacity between all users. In reality, roughly 10–20% of these maximum values can be expected.

Results of both streams are the coding scheme and throughput possible in the network deployed (see Figs. 40–42).

9. 3G NETWORKS (UMTS)

The following sections cover the most important 3G standards.

9.1. FDD/TDD/CDMA2000-TDSCDMA

The so-called "3G" standard has become a mixture of several allowed air interfaces that do not harmonize with

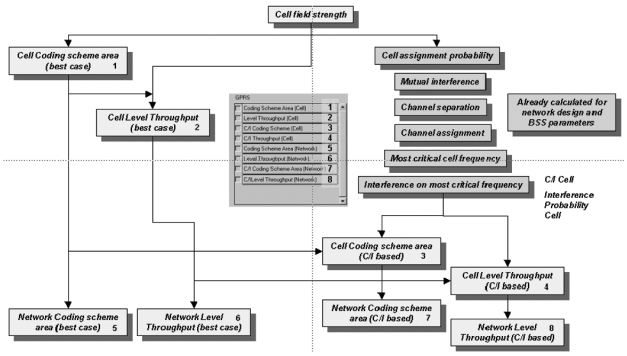


Figure 40. GPRS calculation implementation in a RNP tool.

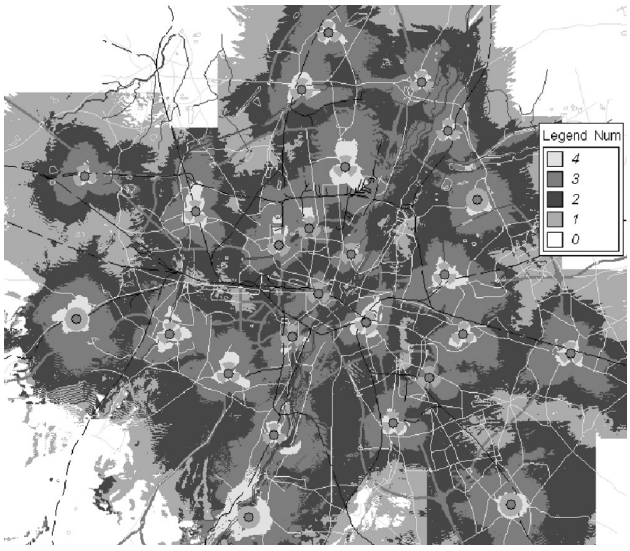


Figure 41. Possible coding schemes in the Munich network.

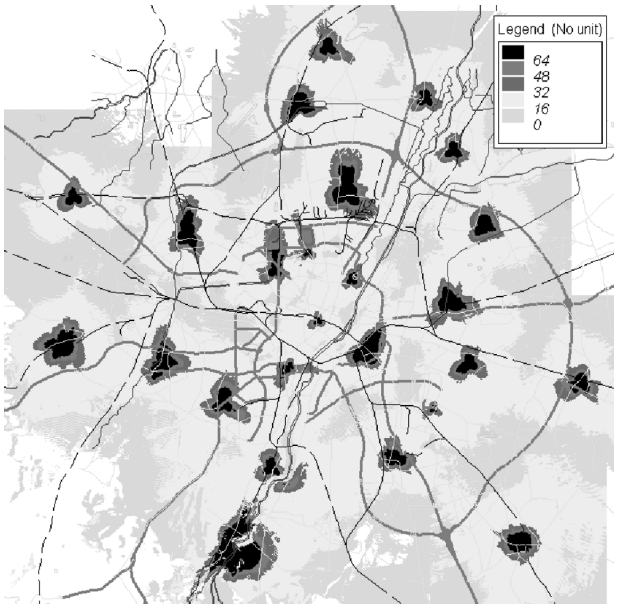


Figure 42. Throughput in kilobits per second.

each other. The framework of this 3G standard(s) is IMT-2000 and has several family members (see Fig. 43).

The process of obtaining a 3G standard was driven by political and commercial interests more than of inventing only one real worldwide standard. Each part of the world has different 2G systems running, which determined the development of a next generation into different directions [12,13]. Therefore, the official 3G standard now consists of several different technologies, all more or less incompatible to each other. "3G" in Europe means for operators and network engineers mostly "UMTS" (wideband CDMA, FDD mode, and TDD mode), where during the first rollout phase operators will stick to the FDD mode to be deployed for macro cellular approach (see Fig. 44). In China (already today world's biggest cellular market), TDSCDMA might dominate the market: TDSCDMA is a special development for the Chinese market and driven by mostly Siemens and Chinese companies, but looks promising for urban situations in general. The following sections will deal with the most popular 3G technologies in Europe: the UMTS-WCDMA FDD and TDD mode used for urban and suburban environments.

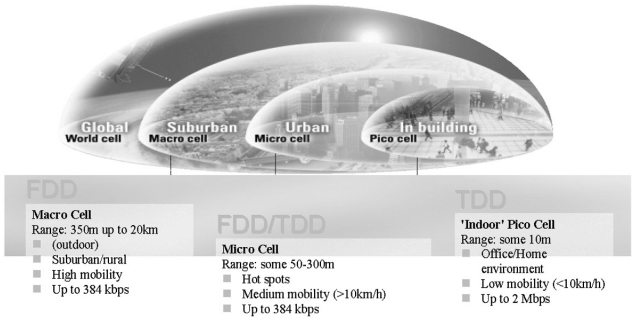


Figure 44. Hierarchical rollout strategy.

9.2. WCDMA Planning Aspects

Especially in Europe, with its dominance of GSM, CDMA is a completely new technology. Perhaps the reader of this article are familiar with UMTS planning, which is the one that most people are familiar with today. It is a wideband CDMA technology, but still, it is a CDMA technology as we have encountered already in IS95 networks. There seem to be some differences compared to IS95 (as in the mixed-traffic scenarios) and also compared to GSM (e.g., the cell breathing effects), but all in all, it seems not that terribly different from planning a 2G/2.5G network. But this is not true. . . .

9.3. UMTS: The Impact of the Service Mixes

One key feature of the UMTS system is its inherent flexibility regarding data rates and service types. The network attempts to satisfy the service requirements with a mixture of

- Spreading factors
- Forward error correction (FEC) coding types
- Signal-to-interference ratio (SIR) targets
- Numbers of spreading sequences
- Code puncturing rates

These all interact, which makes an "all in one" planning process necessary.

A given UMTS "service" offered to any user may be achieved by the allocation of a spreading sequence or a

Standard	UTRA-FDD	UTRA-TDD	TD-SCDMA	CDMA2000	UWCC136	DECT
Freq. band	Paired	Unpaired	Unpaired	Paired	Paired	Unpaired
IMT-2000	IMT-DS	IMT-TD	IMT-TD	IMT-MC	IMT-SC	IMT-FT
	IMT-2000 CDMA DS (direct spread)	Other mode of IMT-2000 CDMA TDD (⇨TD-SCDMA)	One mode of IMT-2000 CDMA TDD (⇨UTRA-TDD)	IMT-2000 CDMA MC (multi carrier)	IMT-2000 TDMA SC (single carrier)	IMT-2000 FDMa/TDMA
Core network compatibility	GSM MAP	GSM MAP	GSM MAP	ANSI-41	ANSI-41	ISDN
Primary standardisation bodies	3GPP	3GPP	CWTS 3GPP	3GPP2	TIA (U.S.)	ETSI

Figure 43. IMT-2000 family members.

number of sequences depending on orthogonal variable spreading factor (OVSF) sequence availability. The network may achieve the target bit error rate (BER) through higher transmit power or higher SIR target setting, or through use of more powerful forward error correction (FEC) coding depending on the delay constraints requested for the service.

Unfortunately, these settings cannot be optimized *per service* but only *per service mix* as a UMTS base station ("node *B*" for historical reasons[1]) must serve all users and their specific services within its cell simultaneously. But as it has only one hardware transceiver unit, the mixture of services offered by a UMTS network means any and all resources and mixes of parameters must be employed to satisfy competing user requirements.

The complexity of the air interface thereby rules out many of the conventional "set and forget" approaches used in planning systems such as GSM or IS95.

For example, assume that you have a stable UMTS network in a city area and suddenly, a bus of tourists arrives, step out and switch on their multimedia 3G handsets to send live movies of their voyage back home. What will be the impact on your network? Is a planning tool capable of simulating that situation?

In GSM, network planning assumed *static* cell boundaries (so-called best server areas), more or less given by power constraints only; traffic issues were addressed by additional TRXs. In 2G CDMA networks (and GSM), there was only *one* service, so cointerference caused by other services simply did not exist. For example, although the famous "cell breathing" already did exist in IS95, it wasn't such a big issue, as it affected only the one and only "voice" service. Given a maximum traffic load for the service, the resulting interference, and thus the amount of cell breathing could be estimated.

In UMTS, however, not only are there a vast number of services, but these services are *very* different from each other: Not only in their data rates, but also their traffic types and QoS demands. So what impact does that one new 2-Mbps packet-switched data user in a cell have on the 20 (circuit-switched) voice users and the one 384-kbps packet-switched data user currently logged on? Must you drop the latter and/or some of your voice users?

Given the mix of services and parameters, the fact is simply that the old-fashioned "static" or even empirical models of the GSM and IS95 world won't work anymore and completely new simulation and analysis strategies must be implemented in an "UMTS RNP tool."

9.4. UMTS Planning Strategy

Another bad habit is to assume the use of conventional RAKE receiver technology from narrowband CDMA networks such as IS95 for UMTS rollout.

Some of these problematic assumptions are

- Gaussian nature of interference
- A small number of multipath components
- Long spreading code analysis of RAKE receiver performance
- Homogeneous network traffic (low data rate or voice)

Given these assumptions, many of the physical-layer performance issues can be "characterized" and abstracted to higher planning levels; this makes IS95 planning simpler to some extent, but this can't be assumed for UMTS any longer. The mixed-traffic, mixed-quality, wideband nature of UMTS invalidate these assumptions. It is well known that RAKE analysis assumptions have significant weaknesses when

- Mobile signals are not tightly power-controlled
- Short spreading sequences are employed
- The number of RAKE fingers is finite and must be shared between cells in the users active set

Nevertheless, many "UMTS planning" tools still use the same old propagation-based network analysis concept.

This approach has significant disadvantages, and many generalizing assumptions must then be made, such as

- The performance of the receiver employed
- The nature of multipath channel
- The nature of the interference produced by the competing users
- The impact of active sets
- The nature and performance of advanced FEC coding schemes

While these assumptions allow a simple analysis of the network, they are so wide-sweeping that they render the results highly inaccurate. Even worse, they give *unduly optimistic* planning results and are far off from reality, the more traffic mixes occur and the more high-data-rate services are involved.

Research results show that

- The conventional "propagation only" approach may serve only as a quick first glance at the network situation. It is fast but way too optimistic for realistic network planning.
- The "static Monte Carlo" method is quite popular today, but again falls short for detailed analysis, especially for mixed-traffic scenarios and involvement of high-bit-rate services.
- An enhanced dynamic method is needed to account for the impact of user mobility and the dynamics that occur with high-data-rate users.

It seems feasible to use at least a static Monte Carlo (MC) analysis for the macrocell environments and address high-traffic environments (e.g., microcells, urban areas, hotspots) and high-mobility environments (e.g., streets and railways) with dynamic analysis. Even the static MC should be replaced by a "quasidynamic" approach, as described in one of the next sections.

[1] The strange term "node *B*" was initially used as a temporary working item during the standardization process but then never changed.

9.5. UMTS Network Design Process

Extensive R&D in WCDMA hardware and software has shown that it is necessary to use a UMTS planning tool that addresses all the issues mentioned before.

A major strength of a strong UMTS planning tool [3] is its ability to model users, service characteristics, and network features in great detail. A powerful network simulation engine provides the designer with the flexibility to explore true dynamic UMTS *real-time* scenarios in detail or examine the "achieved versus designed" network performance.

The design process can be extended beyond the approach used for purely propagation prediction based planning by allowing users and services to be modeled at either an SINR level or even a chip level. This allows the operator to examine

- Link-level performance (dropping, blocking, achieved QoS, BER)
- Mixed-traffic types and their impact on each other
- Handover (active set size change) regions to be examined in detail
- Realistic interference rather than doing simplistic interference modeling

While there are many features and parameters in the proposed UMTS standard, it is likely that only a subset of these parameters will be of interest to the network planner.

A state-of-the-art UMTS planning tool offers three modes of simulation and analysis to the user to tackle the different UMTS planning problems:

- *Static-mode analysis*, which requires only propagation prediction results and network configuration information. The results are service independent. Analysis results are based on link level calculations only. The results available include
  - Least path loss
  - Best server
  - Delay spread
  - RMS delay spread
  - Received CPiCH power
  - Handover regions (active set size changes)
- *Quasidynamic mode*, which requires propagation prediction results plus additional information about the services being offered and the average traffic load in the network. Analysis results are based on link-level calculations and iterative attempts to solve power control equations. The results partly consider the time domain by showing averaged network behavior and should at least include
  - Service coverage [uplink and downlink (UL/DL)] for each service
  - Number of served, blocked, or stolen mobiles for each cell
  - Power used (UDL) for each service
  - OVSF tree utilization
- *Dynamic mode*, which allows the user to examine the link-level performance of the real-time network behavior that includes all the dynamic characteristics of the UMTS air interface. It also allows many of the assumptions associated with ideal RAKE performance to be removed. Results produced by the dynamic simulation mode include
  - Dropping and blocking rates (hard and soft blocking)
  - BER calculations
  - Achieved SIR and SIR target setting performance
  - Dynamic active set utilization
  - Dynamic transmit and receive power requirements

The usual way to plan a network consists of

- RF coverage planning
- Capacity and quality maximization
- Network optimization and maturation

Initially, the designer's task is to provide a rapid RF rollout solution that satisfies the RF coverage design rules and provides acceptable levels of capacity in the planned regions. In 2G, this is done by static analysis of link budget calculations for received pilot signal power, BCCH, and so on. As outlined, this can be done for a first, optimistic indication about coverage, but it is not a suitable way to determine the network capacity for UMTS since there are potentially many different services with different quality requirements competing for the same radio resources. The related issue of capacity and coverage in UMTS can only be addressed by some form of dynamic or quasidynamic analysis that considers the solution to the limited power budget, intercell interference and the competing quality requirements. This task is made more difficult with microcell or "hotspot" environments. In particular, many of the assumptions that are required to be made to determine capacity and coverage begin to break down in heavily loaded regions or when high data rates are considered.

The nature of mixed-traffic-type CDMA also means that the quality of the network is of a highly variable nature. While quasidynamic simulation and analysis techniques will report average performance characteristics, many of the users in the network will experience significantly worse conditions for a significant proportion of the time leading to unacceptable levels of dropping and quality of service.

The final significant task area is to optimize and improve the quality of network. This includes determining the impact of new network equipment, considering the impact of particular dynamic traffic scenarios (e.g., at train stations, highways, or ferry terminals), and improving quality of service through operator-controlled parameters (e.g., SIR target settings). The nature of mixed-traffic CDMA means that many of the physical layer abstractions which are possible in 2G can't be done for UMTS any more; without link-level simulation taking into account the many features of UMTS, many of the optimization tasks are just not possible.

Figure 45 lists some typical tasks for the network planner and the type of analysis tool that is suitable

Design Task	Relative Complexity	Static	Quasi-Dynamic	Dynamic
RF Coverage	Low	✓		
Calculating Service Regions	Medium	✗	✓	✓
"Hot Spot" Region Planning	Medium / High	✗	✓	✓
"What IF" Traffic Profile Scenario	High	✗	✗	✓
Evaluation of New Equipment Features	High	✗	✗	✓
Calculating Blocking / Dropping Rates	High	✗	✗	✓
Determining Network Capacity	High	✗	✓	✓
Determining Network Link Quality	High	✗	✗	✓

Figure 45. Network planning and design tasks.

for carrying out these tasks as well as some relative level of complexity of the task.

For many tasks, quasidynamic analysis is suitable. With all the implicit assumptions that form part of this approach, however, the quasidynamic mode (let alone the simple Monte Carlo approaches) is only able to indicate where support of the desired services is *at all* possible in the region of interest *in the average*. But it will not allow the planner to determine whether the desired service can be supported with the required level of quality at a specific moment of time.

9.6. UMTS Network Planning Examples

The following section shows examples of a fictitious UMTS network in the southeast Sydney, Australia. It consists of 30 sites with 90 sectors.

9.6.1. Propagation-Based Results: Static Approach. The results of importance and achievable with the *static* approach based on path-loss calculations and traffic and interference assumptions are

- Least path loss (Fig. 46)
- Best server (Fig. 47)
- RX CPiCH power level (pilot received Power level)
- Handover regions (active set size changes) (Fig. 48)

The results of the static approach are easily derived and can be adapted from 2G calculations, as all influences from traffic mixes, link level, and OVSF utilization are considered by fixed margin during calculation. The propagation-based approach deals with the parameters listed in Fig. 49.

"Static" in this context means that the time domain is completely ignored. Traffic is considered only indirect by one single margin to be added as "noise." No service mixes can be considered by this approach, let alone specific service requirements.

9.6.2. Propagation-Based Results: Quasidynamic Approach. This approach adds static traffic definitions and

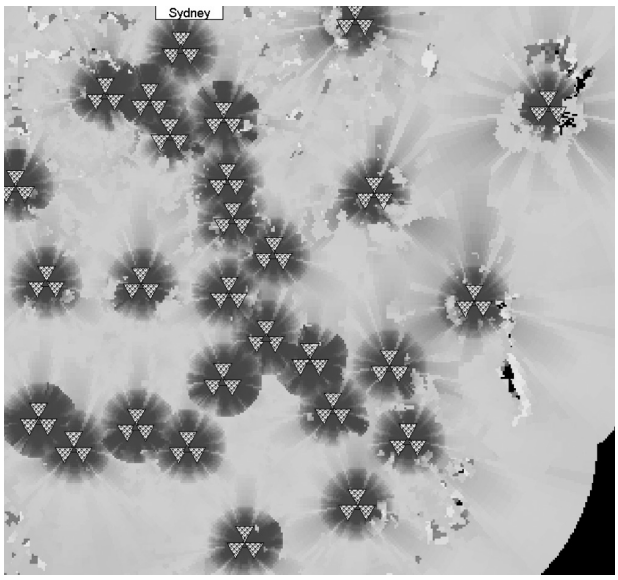


Figure 46. Least path loss.

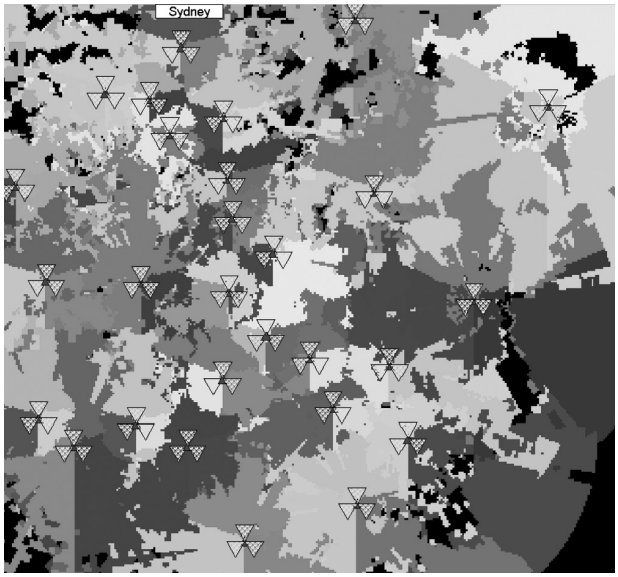


Figure 47. Best server.

traffic mixes to the analysis. The traffic load can be defined on a per cell basis as shown in Fig. 50.

The next result types are examples of the output of the quasidynamic network simulation illustrating the influence of load in a network. The calculation is done by solving the power control equations with a Monte Carlo simulation until the network is in its equilibrium. Then the last mobile unit served in a specific service of the network is analyzed. This shows the cell breathing effect, which depends directly on the load added to the network.

- Test mobile connected at voice (Fig. 51)
- Test mobile connected at 384 kbps (Fig. 52)

Comparing Figs. 51 and 52, it is obvious that the presence of a served user at 384 kbps will dramatically

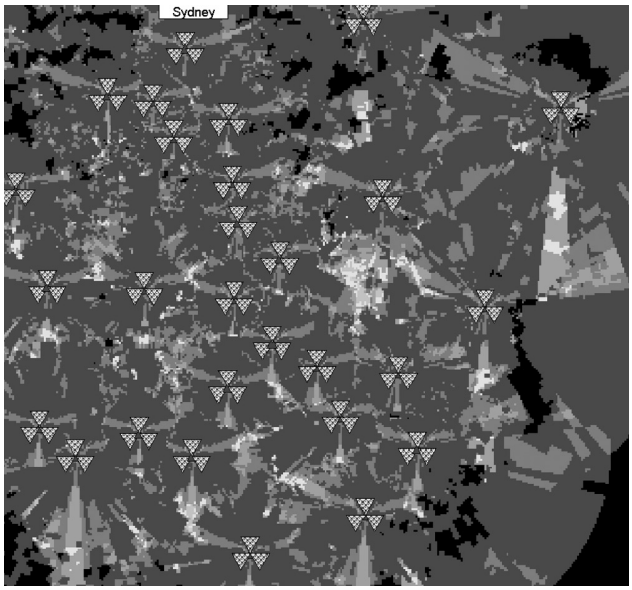


Figure 48. Handover regions.

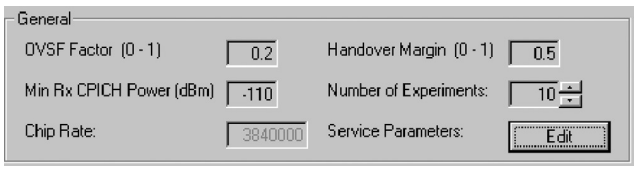


Figure 49. Parameters for static analysis.

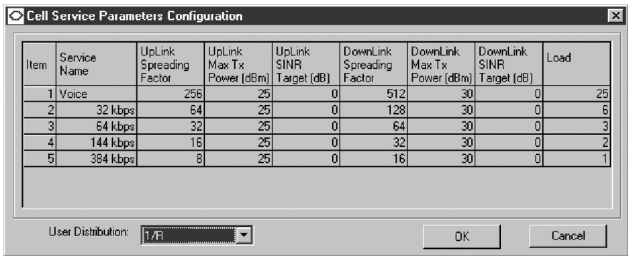


Figure 50. Additional service parameters for quasidynamic simulation.

raise the interference and force all mobiles in the network to raise their transmit power. This is exactly the cell breathing effect. Depending on the maximum TX power level of the mobiles, the "black holes" in the network become larger and larger relative to the load. The term *quasidynamic* is used here because the time domain is taken into consideration by taking several "snapshots" of the network dynamics and building an average out of this. The hope is that if enough snapshots are taken, statistical confidence in the "average" network behavior is reached. Note that this is true only for circuitswitched services and that the number of snapshots (i.e., number of iterations) depends mainly on the services taking part in the traffic mix and the network size itself.

9.6.3. Dynamic Simulation. This approach enables the user to examine in great detail the relations in the network. Instead of traffic definitions on a per cell basis, the users are defined by a statistical function or along deterministic paths. The probability density function is to describe the "move and turn" behavior of the mobile users (see Figs. 53 and 54). Also the number of users and their specific traffic model can be defined on a per user group basis.

The full dynamic approach is to solve detailed problems in the network and can be used to simulate algorithms for

- any new network features / new hardware
- competing / mixed vendor equipment
- advanced UMTS features (smart antennas, transmit diversity, non-RAKE receivers to name but a few)
- operator configurable parameters (SIR target setting, call admission)

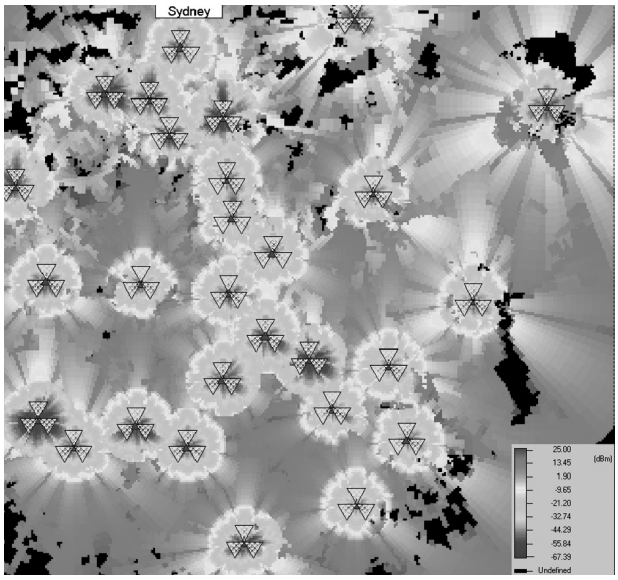


Figure 51. MS TX power at voice connection.

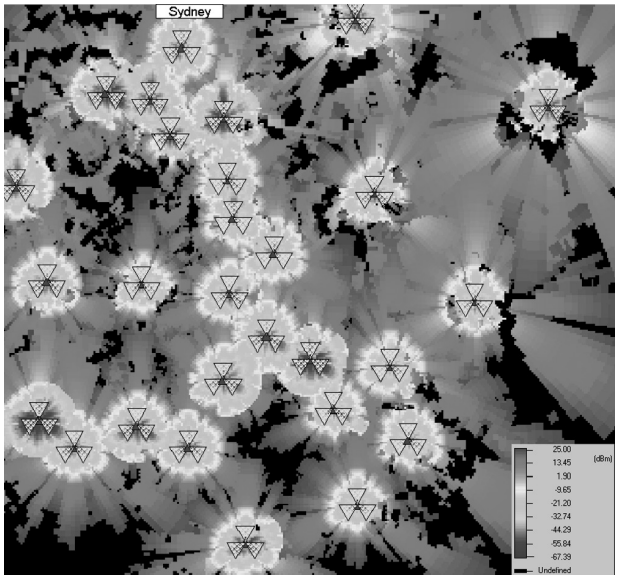


Figure 52. MS TX power for test mobile with 384 kbps.

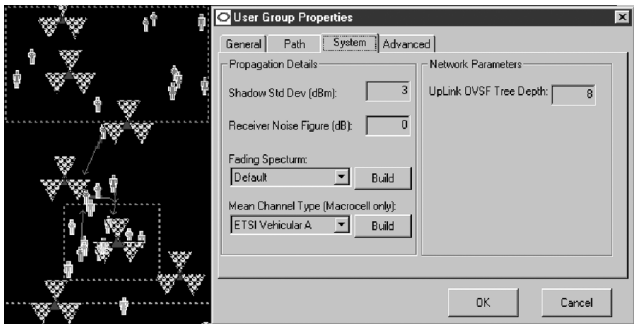


Figure 53. User group properties.

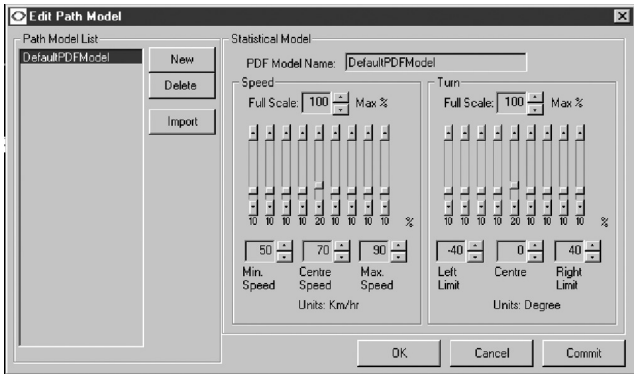


Figure 54. PDF model for a statistical user group.

As it is able to actually *simulate* the UMTS system on a chip level (so each chip which is transmitted by each mobile in up and downlink is considered) dynamic results as the SINR versus time for the different sent and received chips are calculated (see Figs. 55 and 56).

In particular, this truly dynamic simulation approach allows one to actually trace the behavior of 3G systems in dynamic situations, for example, how a UMTS system in equilibrium will behave when suddenly, a high-bit-rate user demands service.

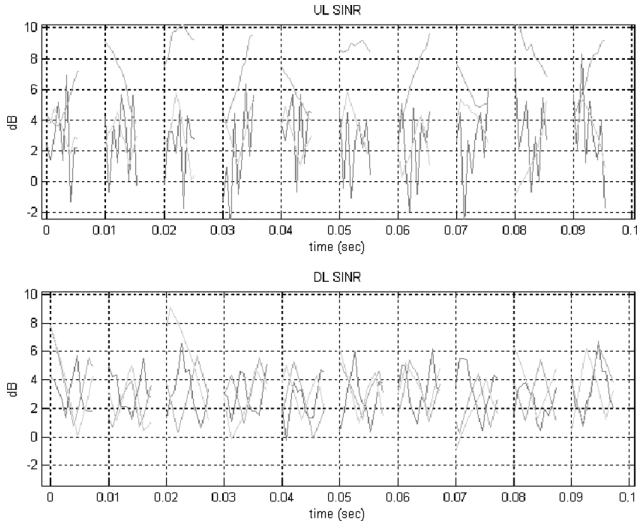


Figure 55. Signal-to-noise ratio versus time for uplink and downlink for different user groups.

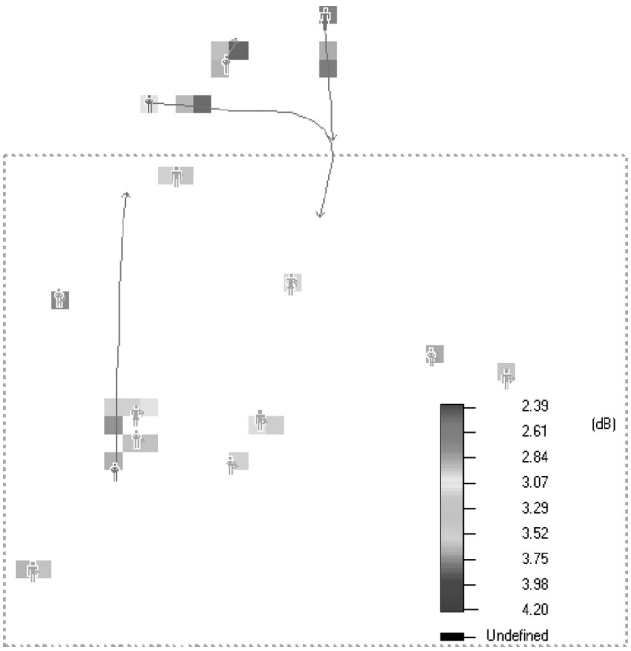


Figure 56. Spatial SINR result for the analyzed user groups. The red lines are the result of mobility model applied.

This method is truly "dynamic" as it completely considers the time domain. As there is no more averaging effect, very detailed simulations are possible. However, the price for accuracy is a lot of computational effort, which currently restricts this method to model evaluation, hotspot analysis, and planning.

9.7. UMTS — Is It Good or Bad?

This question is hard to answer. UMTS is a powerful, flexible standard that allows many new services and offers great promise for the future of mobile communications. The planning and optimization of a UMTS network is made more difficult by this same flexibility. Although it is clear that simple static simulations may be able to predict RF coverage based on pilot powers, in order to tackle the closely coupled problem of coverage and capacity analysis in a mixed-service environment, quasidynamic or dynamic analysis of the network must be performed.

Many planning jobs from 2G networks reappear in UMTS networks also, but often more delicate (e.g., neighborhood planning) or simultaneously (e.g., coverage and capacity analysis). New problems add to this, rendering the network design a very challenging process.

With the fierce competition for subscribers and the high prices paid for UMTS licenses, squeezing the best out of the network is critical. Operators who can support the required services, with the demanded quality of service, will have a clear advantage in the UMTS race.

10. WHERE DO WE GO WITH 4G?

As depicted in the beginning, 4G research and development has just begun. This should not surprise, as for 2G and 3G, the time from first R&D to commercial availability also has been around 10 years. 4G is thus to be expected around 2010.

The necessity for a new mobile generation arises from the progress in fixed network services: Customers increasingly demand the same services and accompanying bandwidths that are available at home (with landline service) when they are using their mobiles.

So, things that we foresee in fixed networks will be driving for development for mobile communication as well. As stated in this article, 20 Mbps in the downlink and roughly 2 Mbps in uplink are to be expected. Having said that, again the question arises how to realize networks that are able to deliver such data rates. If we can't improve propagation predictions, we have to concentrate the energy to the communication path between transmitter and receiver. Logically, we will encounter very small cells, enhanced by adaptive antenna arrays and beamforming. As this can't be deployed nationwide, it is also clear that 4G will live in close internetwork roaming conditions with its predecessor technologies; this can also be a path for a truly global 4G standard.

The backbone network will surely be IPv6 to cater the high demand of IP addresses and bring the mobile world in line with the fixed network world. Higher bit rates also mean higher bandwidth demand in higher frequency ranges: Japan's DoCoMo (which claims to dedicate currently about 80% of its R&D already to 4G) sees 4G in the frequency range of 3–8 GHz. This will need massive worldwide coordination effort, but fortunately, ITU started discussion about such 4G systems that are beyond the current scope of IMT-2000 in the WP-8F initiative in November 1999 and WRC2000 framework. These high-frequency ranges mean worse propagation conditions and higher power demands. Taking the power restrictions of handsets into account, this is also a clear indication for very small cells. There are also ideas of "pumping" networks, such as the example of a highway-bridge-mounted transceiver that covers only a few meters of the highway, but "pumps" that many data with 150 Mbps++ into a car, and that this will yield enough data received until they reach the next serving cell.

Dr. Nobuo Nakajima, former Senior Vice President of DoCoMo Wireless Laboratories, sees an increase in total mobile traffic of 2200% by 2015 compared to that of 1999. He assumes a required bandwidth for 4G of ~1350 MHz.

This all depicts that with 4G, again a "revolution" in system technology will occur, again requiring a whole new planning approach.

BIOGRAPHIES

Jürgen Kehrbeck was born in Karlsruhe, Germany, in 1961. He received the Dr. Ing degree in high-frequency electronics engineering at the University of Karlsruhe, Germany in 1993. His main research activities from 1989 to 1995 were small-vehicle radar sensors in the high gigahertz range. Since 1995 he has been responsible for the development of mobile network planning tools at LS telcom. Currently he is head of the Division of e-Commerce and Mobile Communications at LS telcom.

Kai Rohrbacher was born in Karlsruhe, Germany, in 1966. He received Dipl.-Inform. degree in computer sciences at the University of Karlsruhe, Germany in 1993. He worked for major German computer and telecommunications magazines (and still does so as a sideline). After 4 years of work for a German GSM operator, he joined LS telcom in 1998. Currently, he is head of department mobile communication software at LS telcom.

Werner Wiesbeck (SM'87, F'94) received the Dipl.-Ing. (M.S.E.E.) and the Dr.-Ing. (Ph.D.E.E.) degrees from the Technical University Munich, Germany in 1969 and 1972, respectively. From 1972 to 1983 he was with AEG-Telefunken in various positions, including that of head of R&D of marketing director Receiver and Direction Finder Division, Ulm. During this period he had product responsibility for mm-wave radars, receivers, direction finders, and electronic warfare systems. Since 1983 he has been director of the Institut für Höchstfrequenztechnik und Elektronik (IHE) at the Universität Karlsruhe (TH), Germany.

His research topics include radar, remote sensing, wave propagation, and antennas. In 1989 and 1994, respectively, he spent a 6-month sabbatical at the Jet Propulsion Laboratory, Pasadena. He is a member of the IEEE GRS-S AdCom (1992–2000), chairman of the GRS-S Awards Committee (1994–1998), executive vice president IEEE GRS-S (1998–1999), president IEEE GRS-S (2000–2001), associate editor *IEEE-AP Transactions* (1996–1999), and past treasurer of the IEEE German Section.

He has been general chairman of the 1988 Heinrich Hertz Centennial Symposium, the 1993 Conference on Microwaves and Optics (MIOP'93), and he has been a member of scientific committees of many conferences. For the Carl Cranz Series for Scientific Education, he serves as a permanent lecturer in radar system engineering and for wave propagation. He is a member of an Advisory Committee of the EU–Joint Research Centre (Ispra/Italy), and he is an advisor to the German Research Council (DFG), to the Federal German Ministry for Research (BMBF), and to industry in Germany.

BIBLIOGRAPHY

1. I. S. Redl, M. Weber, and O. Malcolm, *An Introduction to GSM*, Artech House, Boston, 1995.
2. CHIRplus_M, *Computer Based Planning System for 2G and 2.5G Cellular Networks*, LS telcom AG, Lichtenau, Germany, 1999.
3. UTRAplan, *Computer Based Planning System 3G WCDMA FDD/TDD Cellular Networks*, LS telcom AG, 77839 Lichtenau, Germany, 1899.
4. www.placeAbase.com, *Web Based Site Marketplace*, LS telcom AG, 77839 Lichtenau, Germany, 2000.
5. Bureau de Développement des Télécommuni-cations, *Manual on Mobile Communication Development*, ITU-Geneva, 1997.
6. X. Huang, *Automatic Cell Planning for Mobile Network Design: Optimization Models and Algorithms*, Ph.D. thesis, Faculty of Electrical Engineering, Univ. of Karlsruhe, Germany, May, 2001.
7. W. C. Y. Lee, Mobile Cellular Telecommunications Systems, McGraw-Hill, New York, 1998.
8. A. M. C. A. Koster, *Frequency Assignment—Models and Algorithms*, Ph.D. thesis, Maastricht Univ., The Netherlands, 1999.

9. A. Gamst, Some lower bounds for a class of frequency assignment problems, *IEEE Trans. Vehic. Technol.* **35**(1): 8–14.
10. D. Minoli, Broadband Network Analysis and Design, Artech House, Boston.
11. J. S. Lee and L. E. Miller, *CDMA Systems Engineering Handbook*, Artech House, Boston.
12. T. Ojanperä and R. Prasad, *Wideband CDMA for Third Generation Communications*, Artech House, Boston.
13. H. Holma and A. Toskala, *WCDMA for UMTS*, Artech House, Boston.
14. COST 231, *Digital Mobile Radio Towards Future Generation Systems* (1989–1996), EUR 18957.
15. ETSI Technical Report, *Terrestrial Trunked Radio*, ETR 086-1, Jan. 1994.
16. ETSI Technical Report, *Terrestrial Trunked Radio*, ETR 300-1, May 1997.
17. ETSI Technical Report, *Terrestrial Trunked Radio*, ETR 300-2, May 1997.
18. Agreement between the telecommunications authorities of Austria, Belgium, the Czech Republic, Germany, France, Hungary, the Netherlands, Croatia, Italy, Lithuania, Luxembourg, Poland, Romania, the Slovak Republic, Slovenia and Switzerland, on the coordination of frequencies between 29.7 MHz and 43.5 GHz for fixed services and land mobile services, Vienna, June 30, 2000.
19. E. Zitzler and L. Thiele, Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach, *IEEE Trans. Evol. Comput.* **3**(4): 257–271.

CELLULAR COMMUNICATIONS CHANNELS

Aarne Mämmelä
Pertti Järvensivu
VTT Electronics
Oulu, Finland

1. INTRODUCTION

This article gives an overview of cellular radio or wireless channels for mobile digital communications [1]. The main emphasis is in channel models for radiowave propagation in terrestrial outdoor mobile cellular systems between a base station and a mobile station in the downlink and uplink, in either microcells or macrocells. Such systems typically work in the frequency range from about 1 to 2 GHz with the corresponding wavelengths between 0.3 and 0.15 m. The bandwidth of the transmitted signals is in the order of 100 kHz–1 MHz.

The location of the base station antenna has a significant effect on channel modeling. In microcells the cell radius is about 0.1–1 km, and the base station antenna is below the rooftop level of the surrounding buildings. On the other hand, in macrocells the base station antenna is above the rooftop level and the cell radius is about 1–30 km. The area types are usually divided into urban, suburban, and rural, each of which may be nonhilly or hilly.

A radio channel is almost always linear. Because of its mobility, the channel is also time-variant. It is thus fully described by its impulse response $h(\tau, t)$, where τ is the delay parameter and t is the time. The complex impulse response $h(\tau, t)$ is a lowpass equivalent model of the actual real bandpass impulse response. Equivalently, the channel is characterized by its transfer function $H(f, t) = \int_{-\infty}^{\infty} h(\tau, t) \exp(-j2\pi f\tau) d\tau$, which is the Fourier transform of the impulse response with respect to the delay parameter.

The magnitude $|H(f, t)|$ of the transfer function at a given frequency f is changing randomly in time, and we say that the mobile radio channel is a fading channel. The phase $\arg H(f, t)$ is also a random function of time. Fading is caused mainly by multipath propagation due to reflection, scattering, and diffraction of the radiowaves from nearby objects such as buildings, vehicles, hills, and mountains. With respect to a stationary base station, multipath propagation creates a stochastic standing-wave pattern, through which the mobile station moves. Additional fading is caused by shadowing when the propagation environment is changing significantly, but this fading is typically much slower than the multipath fading. Modem design is affected mainly by the faster multipath fading, which can be normally assumed to be locally wide-sense stationary (WSS). Early important work on WSS fading multipath models in a more general framework is due to Turin, Kailath, and Bello in the 1950s and 1960s [2,3].

Some modern systems use directive antennas to amplify the desired signal and to reject the interfering signals. Conventionally only horizontal directions are taken into account. In such systems the direction of arrival of the received signals as well as the azimuthal power gain of the antenna are important issues, and the models are two-dimensional (2D). The two dimensions are the delay and the azimuth whereas in one-dimensional (1D) models the only dimension (in addition to time) is the delay. Important early work on 2D models was done by Clarke in the 1960s [4].

The channel models are used for performance analysis and simulations of mobile systems. The models can also be used for measurements in a controlled environment, to guarantee repeatability and to avoid the expensive measurements in the field. However, any model is only an approximation of the actual propagation in the field. For measurements, the average received signal-to-noise ratio must be defined. It is estimated by making a link power budget, which includes the transmitter power, distance-dependent attenuation of the channel, antenna gains in the transmitter and receiver, and various loss factors and margins. It depends on the system designer whether a margin for fading is taken into account or whether the performance simulations or measurements with the channel model will include fading. The power of additive noise is also estimated for modeling purposes. Usually only the thermal noise with a certain noise figure in the receiver is considered. The additive noise is assumed to be white Gaussian noise (WGN) within the signal bandwidth. Unless otherwise stated, we will exclude any other noise sources.

The organization of this chapter is as follows. In Section 2 we give the statistical description of one-dimensional and two-dimensional channel models. In Section 3 we summarize the methods by which the channel is measured. In Section 4 widely available simulation models are described. Finally, in Section 5 some more recent trends are noticed. More extensive reviews are included for the one-dimensional models [5,6] and two-dimensional models [4]. Models for indoor communications are summarized [7]. Much of the theory is valid also for outdoor communications. Our list of references is not exhaustive, and some very important work has been left out because of space limitations. Additional references can be found from the papers cited.

2. STATISTICS OF THE TIME-VARIANT IMPULSE RESPONSE

The most important propagation phenomena to be included in a channel model are shadowing and multipath fading, and the model is either one-dimensional or two-dimensional. We will first consider 1D models, which are characterized by the time-variant impulse response and transfer function. If the transmitted signal is denoted by $z(t)$, the received signal $w(t)$, without noise, is given by $w(t) = \int_{-\infty}^{\infty} z(t-\tau)h(t,\tau)d\tau$. Alternatively, the received signal is $w(t) = \int_{-\infty}^{\infty} Z(f)H(f,t)df$, where $Z(f)$ is the Fourier transform of $z(t)$.

If the transmitted signal has a bandwidth of W, the delay resolution of the measurement is approximately $1/W$, which means that the receiver cannot resolve delay differences smaller than $1/W$. We define such unresolved multipath components as clusters on the delay axis [3]. The receiver can resolve multipath components whose delay differences are larger than $1/W$. We will apply the central-limit theorem for the clusters. The impulse response has the general form $h(\tau,t) = \sum_{l=0}^{L-1} h_l(t)\delta(\tau-\tau_l)$, where L is the number of resolvable clusters whose amplitudes and delays are $h_l(t)$ and τ_l, respectively. Since the channel is random, we need a stochastic description for it. The delays of the clusters are usually assumed to be constant in channel models, but it must be noted that fading is caused mainly by the randomly changing delays, which change the relative phase shift between the multipath components within the clusters.

If several multipath signals due to scattering with approximately equal amplitudes, or alternatively with random amplitudes, are added at random phases, the resultant has a complex Gaussian distribution with a zero mean. The amplitude of such a cluster is Rayleigh distributed and the phase is uniformly distributed. The channel is then said to be a Rayleigh fading channel. Alternatively, if in addition to the scattered components, the received signal includes a strong component, which is a line-of-sight (LOS) signal coming either directly from the transmitter or from a specular reflection, the impulse response at that delay will have a Gaussian distribution with a nonzero mean and the amplitude will be Rice distributed. The channel is in this case a Rice fading channel. In both Rayleigh and Rice fading channels, only the first- and second-order statistics, including the mean and autocorrelation functions, are needed to fully describe them. A more general description is the covariance matrix of a discretized impulse response.

For multipath fading, a widely used model is a wide-sense stationary uncorrelated scattering (WSSUS) model. It is WSS with respect to the time variable. Uncorrelated scattering (US) means that the autocorrelation function of the WSS Rayleigh fading impulse response has the form $E\{h^*(\tau,t)h(\tau+\Delta\tau,t+\Delta t)\} = P_h(\tau,\Delta t)\delta(\Delta\tau)$, or there is no correlation on the τ axis, but some correlation may exist on the time axis. The function $P_h(\tau,\Delta t)$ is the autocorrelation of the impulse response at the delay τ with the time difference Δt. The impulse response is nonstationary white noise in the delay variable. It can be shown that in a WSSUS channel the transfer function is WSS also with respect to the frequency variable [2].

The Fourier transform of $P_h(\tau,\Delta t)$ with respect of the time difference Δt is the scattering function $S(\tau,\lambda)$ of the channel, or $S(\tau,\lambda) = \int_{-\infty}^{\infty} P_h(\tau,\Delta t)\exp(-j2\pi\lambda\Delta t)d\Delta t$, where λ is the Doppler shift variable. The scattering function is a measure of the average power output as a function of the time delay τ and the Doppler shift variable λ. The delay power spectrum is $P_h(\tau) = \int_{-\infty}^{\infty} S(\tau,\lambda)d\lambda$. Equivalently, the delay power spectrum is $P_h(\tau) = E\{|h(\tau,t)|^2\}$. The Doppler power spectrum is $S_H(\lambda) = \int_{-\infty}^{\infty} S(\tau,\lambda)d\tau$.

The width of the delay power spectrum is referred to as the *delay spread*, and the width of the Doppler power spectrum is referred to as the *Doppler spread*. A suitable engineering definition is used for the width. The channel is frequency-nonselective or flat fading if the signal bandwidth W is smaller than the inverse of the delay spread, or the coherence bandwidth; otherwise the channel is frequency-selective. The Doppler spread and its inverse, or the coherence time, are measures of the rapidity of fading.

A typical approximation for the delay power spectrum is exponential. A typical approximation for the Doppler power spectrum is $S_H(\lambda) = (1/\pi f_m)[1-(\lambda/f_m)^2]^{1/2}$, where $f_m = (v/c)f_0$ is the Doppler frequency, v is the velocity of the mobile station, c is the velocity of the radiowaves, and f_0 is the carrier frequency. This Doppler power spectrum is based on the assumption that the multipath components arrive the omnidirectional antenna uniformly from all horizontal directions. It is often referred to as Jakes's Doppler power spectrum [1] even though it was derived earlier by Clarke [5].

In addition to Rayleigh and Rice distributions, a useful amplitude distribution for the multipath fading is Nakagami m distribution, which is in fact a form of the generalized Rayleigh distribution. When selecting a suitable distribution, one should note that for system performance, the most notable effect has the distribution at small amplitudes [8].

Shadowing is essentially frequency-nonselective fading, much slower than multipath fading, and it is usually described by the lognormal distribution; thus, the received

power in decibels is normally distributed. The lognormal distribution is also based on the central-limit theorem [9]. The product of several random variables may be approximated as being lognormally distributed. The product comes from the various attenuation factors due to the obstacles between the transmitter and the receiver.

The WSSUS model described above can be generalized to the 2D case as follows. The azimuth or the angle of arrival relative to the velocity vector of the mobile station is denoted by α. As previously, the multipath components are combined into clusters in space with a delay resolution of $1/W$ and with an angular resolution $\Delta\alpha$ of the receiver antenna. Each cluster has a Rayleigh or Rice fading amplitude. Each complex gain of the impulse response has now the general form $h_l(t) = \sum_{n=0}^{N-1} a_{ln} \exp[j\varphi_{ln} + 2\pi f_m t \cos(\alpha_{ln})]$ where α_{ln} is the azimuth and φ_{ln} is the phase of the nth component in the lth delay and N is the number of components in the model at the lth delay. As a generalization of the delay power spectrum, we can define an azimuthal delay power spectrum that shows the distribution of the received power versus azimuth and delay. For a given velocity and direction of the mobile station, and for a given azimuthal power gain of the antenna, the azimuthal delay power spectrum corresponds to a certain scattering function of the WSSUS channel (see Fig. 1). The scattering function is an aliased form of the azimuthal delay power spectrum since two different azimuths α_{ln} and $-\alpha_{ln}$ of arriving clusters create the same Doppler shift due to the cosine function in $h_l(t)$. Given a uniform distribution $p(\alpha)$ for the received power, we obtain Clarke's Doppler power spectrum given earlier.

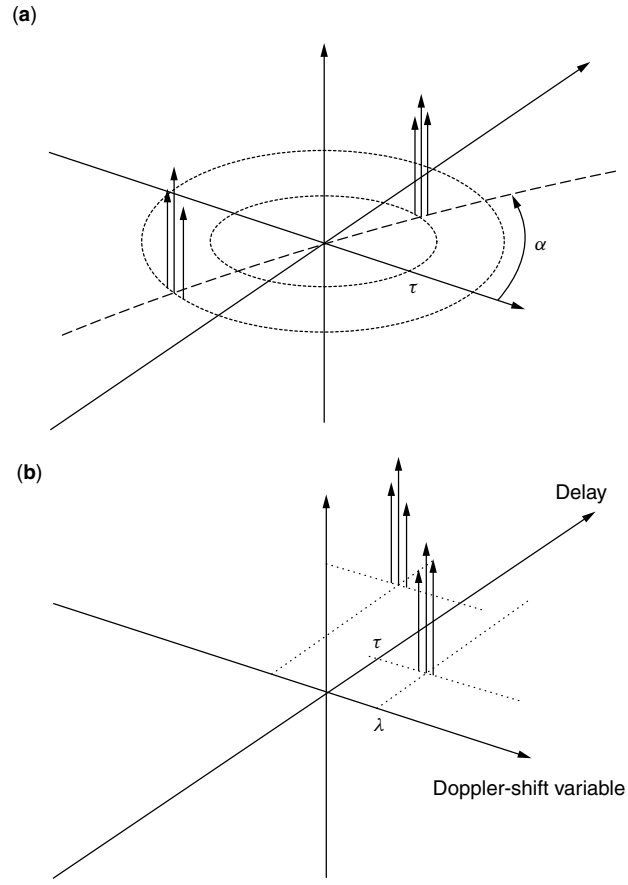


Figure 1. Azimuthal delay power spectrum (**a**) and scattering function (**b**) of the channel.

3. MEASUREMENT OF LINEAR TIME-VARIANT CHANNELS

Channel measurements can be divided into narrowband and wideband measurements. Wideband measurements use measurement signals, which have about the same bandwidth as the intended information signal. Unlike narrowband measurements, which use a single unmodulated carrier as a measurement signal, wideband measurements provide information on the multipath propagation as well as frequency selectivity of the channel. Therefore, only wideband measurements are discussed here. The measurements can also be divided into measurement of [1] instantaneous values of the impulse response and [2] the average parameters of the channel. The average parameters include first/second-order statistics, or the mean and the autocorrelation function of the impulse response, and the scattering function of the channel. Several ways to perform systematic measurements have been listed in Ref. 6.

The problem of the measurement of system functions of time-variant channels differs from that of the time-invariant case. Even in the absence of noise the system function of a time-variant channel may be unmeasurable. The condition on the measurability of a linear time-varying WSSUS channel was first presented by Kailath and later extended by Bello [10]. It turns out that the channel measurability depends on the value of the area spread factor, which in effect is the area of region where the scattering function is effectively nonzero. If the area spread factor is less than or equal to a threshold, the channel impulse response could be measured unambiguously [10]. The value of the threshold is of the order of unity. The channels for which the area spread factor fulfills the criterion mentioned above are called *underspread*; otherwise they are called *overspread*. If the channel is overspread, it is not possible to measure the instantaneous values of the impulse response, even in the absence of the noise. Fortunately, most physical channels are underspread. The average parameters can be determined either from the instantaneous values of the channel impulse response or by cross-correlation methods. Since the statistical averages contain much less information than the instantaneous values, the channel need not always be underspread before the average parameters could be measured.

Various channel measurement techniques have been proposed [11]. However, two practical methods of measuring the impulse response of the underspread cellular channel can be identified. One method is to transmit a very short impulselike pulse to the channel and observe the multiple pulses received. In order to follow the time variation of the channel, the pulses need to be transmitted periodically. The short pulses result in a high ratio of peak to average transmission power, which could be undesirable. Another, more efficient, method to measure the impulse

response is the use of direct sequence spread-spectrum signals [6]. A pseudonoise (PN) sequence is used to modulate the carrier. Maximal-length sequences (m sequences) are widely used because of their excellent periodic autocorrelation properties. The receiver is based on a correlation principle. It can be implemented by a matched filter or a sliding correlator [5]. Nowadays, high-speed digital signal processing techniques can be employed to implement real-time matched filter channel sounders. In a WSSUS channel, time averaging can be used to obtain the autocorrelation function of the measured impulse response. An estimate of the scattering function can then be calculated by the Fourier transform [12,13]. Angle-of-arrival measurements can be conducted by using directional or array antennas at the receiver [4]. To some extent, angle of arrival can be deduced from the measured scattering function [12].

One of the earliest measurement results of the impulse response of the cellular channel in urban and suburban areas by using short pulses was reported by Turin in 1972 [3]. The delay, amplitude, and phase of multipath components were measured at three different frequencies simultaneously. It was found that spatial correlation distances of these variables at neighboring geographic points vary considerably. They ranged from less than a wavelength for the phases, through tens of wavelengths for the amplitudes and delays, to hundreds of wavelengths for the means and variances, or powers, of the amplitudes.

More recent measurement studies have used almost exclusively the direct-sequence spread-spectrum signals with m sequences to measure the channel. Wideband macrocell measurements conducted at 1 GHz show that in typical urban areas the delay spread from 1 to 2 μs is characteristic. In suburban areas delay spreads from 10 to 20 μs are typical. The longest delay spreads occur in mountainous environments, where delay spreads from 100 to 150 μs have been encountered. In open areas the delay spread is practically nonexistent and the received signal consists of the directly propagated component only [12]. Wideband measurements in urban microcell environments at 2 GHz have been reported [13]. The delay power spectrum, average normalized correlation functions, and scattering functions have been calculated from the measured impulse response. Under LOS conditions, the direct component with unresolvable specular reflections dominated the propagation. Some resolvable specular reflections existed at delays of up to 1.5 μs. In non-LOS situations, the powers of the strongest received signals were more than 15 dB below the LOS components that resulted in nearby locations. The propagation process was found to be dominated by multiple reflections and scattering along the streets, and not by diffraction (13).

Relatively few results on spatial channel measurements have been published. Ertel et al. have summarized some of the results [4]. Measurements conducted at 2 GHz with a 10-MHz bandwidth, by using a rotating azimuth beam directional antenna at the receiver, have shown that delay and angle-of-arrival spreads are small in rural, suburban, and even many urban environments. Measurements in urban areas have shown that most of the major features of the delay angle of arrival spectra can be accounted for by considering the large buildings in the environment. Finally, variations in the spatial characteristics with both time and frequency have been measured. The results indicate that the uplink spatial characteristics cannot be directly applied for downlink beamforming in most of the present cellular and personal communication systems that have 45-MHz and 80-MHz separations between the uplink and downlink frequencies, respectively [4].

4. SIMULATION MODELS

The evolution of channel simulation models has been parallel to that of cellular systems. The early models considered only the signal amplitude-level distributions and Doppler spreads of the received signals. A delay spread information was later added to the channel models. In addition to those, modern channel models also include such concepts as angle-of-arrival and adaptive array antenna geometries. The signal parameters that need to be simulated in these models for each multipath component are the amplitude, carrier phase shift, time delay, angle of arrival, and Doppler shift. All of these parameters are in general time-varying, causing Doppler spread in addition to delay spread [4].

The channel simulators can be categorized into three classes according to the way the channel impulse response is modeled: stored channel impulse response, ray tracing models, and stochastic parametric models for the channel impulse response. The stored channel impulse responses are based on selected measurements, which are then stored for later use. Although this method provides actual information from the channel, the proper selection criterion for the measurements may be difficult to identify. Also, the large amount of data needed to store the measurement results could be difficult to handle. However, some models have been proposed [4,14]. The ray tracing models are deterministic. They are based on geometric propagation theory and reflection, diffraction, and scattering models. Accurate channel models are possible using this method. However, the high computational burden and lack of detailed terrain and building databases make these models difficult to use [4]. By far the most popular channel simulation models are stochastic parametric models. In this approach, the channel impulse response is characterized by a set of deterministic and random parameters. The values of the parameters and the probability distributions governing their behavior are selected according to measurements. The remaining challenge is to develop models, that exhaustively reproduce the propagation scenarios accounted in reality [14]. A recommended summary of stochastic channel models can be found in Ref. 14. Spatial, or 2D, stochastic channel models are summarized in Ref. 4.

Usually, discrete-time channel impulse responses in the form of transversal filters are used in the stochastic channel simulators. The transversal filter model allows the simulators to be implemented either by software or hardware. The time-varying complex coefficients and delays of the transversal filter are generated according to the statistics associated with the different parameters. The amplitudes are usually assumed to be Rayleigh or Rice distributed as a result of multipath fading. A uniform and

Poisson distribution is usually assumed for the phases and delays, respectively. However, in a Rice fading channel the phase is concentrated around the phase of the strong component unless there is a Doppler shift in it. As mentioned earlier, the delay power spectrum is typically approximated by an exponential function. There are several methods to simulate the angles of arrival. For example, in Ref. 15 they are modeled as normally distributed random variables. Other methods have been summarized in Ref. 4. Rayleigh and Rice processes needed to simulate the amplitudes can be generated by using colored Gaussian noise processes. A well-known method to produce colored Gaussian noise processes is to filter WGN with a filter having a transfer function of the square root of the Doppler power spectrum. Typically, the Doppler power spectrum by Clarke is used. Another method is based on Rice's sum of sinusoids. In this case, a colored Gaussian noise process is approximated by a finite sum of weighted and properly designed sinusoids [16]. The long-term variations in the channel impulse response can be modeled by making the delays drift with time and by using an attenuation filter to model the lognormal fading caused by shadowing or transitions between different environments [14]. An exponential function is used to approximate the autocorrelation of the shadowing as a function of distance. A correlation distance of 20 m is typically used for urban environments. For suburban environments, much larger correlation distances should be used. In a hardware implementation, digitally controllable attenuators can be used to simulate the attenuation caused by the shadowing [15].

Different standardization organizations are actively defining channel models as a part of specification of new mobile cellular systems. Their motivation is to specify the operational environment of the system and to provide test parameters for manufacturers. The channel models for second-generation digital advanced mobile phone system (DAMPS) and global system for mobile communications (GSM) mobile cellular systems were specified by the Telecommunications Industry Association (TIA) in the United States and the European Telecommunications Standards Institute (ETSI) in Europe. For the third-generation cellular systems a global standard has been defined as the International Mobile Telecommunications (IMT-2000) proposal by the International Telecommunication Union (ITU). For a more thorough discussion, see Ref. 6. The standardization work of third-generation systems has now shifted to the international 3rd Generation Partnership Project (3GPP).

The channel models have also been developed by different research consortia in international research programs. In Europe, particularly Cooperation in the Field of Scientific and Technical Research (COST) projects have been extremely influential when GSM and digital communication system at 1800 MHz (DCS 1800) systems were developed. The achievements in COST partly stimulated the Universal Mobile Telecommunications System (UMTS) Code-Division Testbed (CODIT) project within the Research and Development in Advanced Communication Technologies in Europe (RACE-II) program. The CODIT 2D channel models seem to be state-of-the-art. A summary of European research programs can be found in Ref. 14.

5. MORE RECENT TRENDS

Channel modeling for cellular communications is a rapidly changing area, and the models are becoming increasingly accurate. Some of the more recent trends are summarized here. Higher frequencies of up to ~60 GHz will be used in the future. The cell size is made smaller since the channel attenuation is larger at higher frequencies. Frequencies above ~10 GHz are also affected more by air molecules and rain. Consequently, the highest frequencies can be used only in indoor environments. Also, with the increasing data rates, the bandwidths are becoming larger, approaching 10–100 MHz. 3D models are important in macrocells, for example, in urban and mountainous areas where the base station antenna is much higher than the mobile station antenna. The 3D models take into account the elevation angle of the arriving waves, in addition to the azimuthal angle and the delay. Various nonstationary models are often used. In addition to the lognormal distribution, shadowing effects are modeled with birth–death processes, where some delays suddenly appear and disappear, simulating rapid changes as in street corners and tunnels. In some models the delays are time-variant to test the delay tracking ability of the receiver. Furthermore, the models are becoming more comprehensive in the sense that they will have multiple inputs and outputs. In this way the models can be used to simulate diversity systems with many users. Even handoffs between base stations should be simulated. Correlation and crosstalk between the multiple channels are important effects in such systems. An example of crosstalk is the cochannel and adjacent-channel interference between the various users of the same frequency band in the same geographic region.

BIOGRAPHIES

Aarne Mämmelä was born in Vihanti, Finland, in 1957. He received the degrees of M.Sc. (Eng), Lic.Tech., and Ph.D. (all with distinction) from the Department of Electrical Engineering, University of Oulu, Finland, in 1983, 1988, and 1996, respectively. His doctoral thesis was on diversity receivers in fast fading multipath channels. From 1982 to 1993 he was with the Telecommunication Laboratory at the University of Oulu and researched adaptive algorithms in spread-spectrum systems. In 1990–1991 he visited the University of Kaiserslautern, Germany. In 1993 he joined VTT, Computer Technology Laboratory, which was merged to VTT Electronics in 1994. In 1996–1997 he was on leave as a postdoctoral fellow at the University of Canterbury, Christchurch, New Zealand. Since 1996 he has been a research professor of digital signal processing at VTT Electronics. His research area is the design of digital transmitter-receivers in wireless communications. In addition, since 2000 he has been a docent or adjunct professor of receiver signal processing at the Helsinki University of Technology, Espoo, Finland. He

is especially interested in synchronization and estimation problems in wireless digital communications, both in single- and multi-carrier systems.

Pertti Järvensivu was born in Laukaa, Finland, in 1966. He received a M.Sc. degree in electrical engineering from the University of Oulu, Finland, in 1992. From June 1988 to December 1999 he was in various teaching and research positions at the University of Oulu. Since January 2000 he has been with VTT Electronics as research scientist in digital signal processing. His current research interests are channel estimation and wireless adaptive radio systems.

BIBLIOGRAPHY

1. W. C. Jakes, ed., *Microwave Mobile Communications*, Wiley, New York, 1974.
2. P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* **CS-11**: 360–393 (1963).
3. G. L. Turin, Introduction to spread-spectrum antimultipath techniques and their application to urban digital radio, *Proc. IEEE* **68**: 328–353 (1980).
4. R. B. Ertel et al., Overview of spatial channel models for antenna array communication systems, *IEEE Pers. Commun.* **5**: 10–22 (1998).
5. D. Parsons, *The Mobile Radio Propagation Channel*, Pentech Press, London, 1992.
6. K. Pahlavan and A. H. Levesque, *Wireless Information Networks*, Wiley, New York, 1995.
7. H. Hashemi, The indoor radio propagation channel, *Proc. IEEE* **81**: 943–968 (1993).
8. S. Stein, Fading channel issues in system engineering, *IEEE J. Select. Areas Commun.* **SAC-5**: 68–89 (1987).
9. A. J. Coulson, A. G. Williamson, and R. G. Vaughan, A statistical basis for lognormal shadowing effects in multipath fading channels, *IEEE Trans. Commun.* **46**: 494–502 (1998).
10. P. A. Bello, Measurement of random time-variant linear channels, *IEEE Trans. Inform. Theory* **IT-15**: 469–475 (1969).
11. A. Hewitt and E. Vilar, Selective fading on LOS microwave links: Classical and spread-spectrum measurement techniques, *IEEE Trans. Commun.* **36**: 789–796 (1988).
12. W. R. Braun and U. Dersch, A physical mobile radio channel model, *IEEE Trans. Vehic. Technol.* **40**: 427–482 (1991).
13. U. Dersch and E. Zollinger, Physical characteristics of urban micro-cellular propagation, *IEEE Trans. Antennas Propag.* **42**: 1528–1539 (1994).
14. B. H. Fleury and P. E. Leuthold, Radiowave propagation in mobile communications: An overview of European research, *IEEE Commun. Mag.* **34**: 70–81 (1996).
15. J. J. Olmos, A. Gelonch, F. J. Casadevall, and G. Femenias, Design and implementation of a wide-band real-time mobile channel emulator, *IEEE Trans. Vehic. Technol.* **48**: 746–764 (1999).
16. M. Pätzold, U. Killat, F. Laue, and Y. Li, On the statistical properties of deterministic simulation models for mobile fading channels, *IEEE Trans. Vehic. Technol.* **47**: 254–269 (1998).

CHANNEL MODELING AND ESTIMATION

LANG TONG
Cornell University
Ithaca, New York

1. INTRODUCTION

One of the objectives of receiver design for digital communications is to minimize the probability of detection error. In general, the design of the optimal detector requires certain prior knowledge of the channel characteristics, which are usually estimated through the use of pilot symbols. A typical example is the voiceband modem where, on establishing the initial connection, a signal known to the receiver is transmitted through the telephone channel. The receiver is then tuned to compensate for the distortions caused by the channel. The process of using pilot symbols to estimate the channel, or directly, the receiver coefficients is referred to as *training*.

If there is a sufficient amount of training time, and the channel does not vary significantly, the problem of channel estimation can be formulated either as the classical point estimation or as the Bayesian estimation. Techniques such as maximum-likelihood estimation and methods of least squares are readily applicable, and they generally offer good performance. For these applications, the choice of algorithm is often determined by the complexity of implementation.

The explosive growth in wireless communications and the increasing emphasis on packet-switched transmissions present a new set of challenges in channel estimation and receiver design. Although critical to reliable communications, channel acquisition and tracking are difficult because of the rapid variations caused by multipath fading and the mobility of users. The use of training is no longer straightforward, as the receiver must be trained repeatedly for time-varying channels. Since the time used for training is the time lost for transmitting information, there is a tradeoff between the quality of channel estimation and the efficiency of channel utilization.

This article presents an overview of the modeling and estimation of channels for digital transmission of single carrier linearly modulated signals. In Section 2, we present the complex baseband representation of intersymbol interference channels. A discrete-time linear model is obtained that relates the received data samples with the channel coefficients and the transmitted pilot and data symbols. Within the framework of parametric estimation, we present techniques and performance analysis of various channel estimation problems in Sections 3 and 4. A brief bibliography note is provided at the end of this article.

2. THE BASEBAND MODEL OF BAND-LIMITED CHANNELS

In this section, we consider the baseband model, in both continuous and discrete time, for linearly modulated signals transmitted over band-limited passband channels.

2.1. The Continuous-Time Model

Figure 1 illustrates the passband transmission of linearly modulated baseband signals. To transmit a sequence of

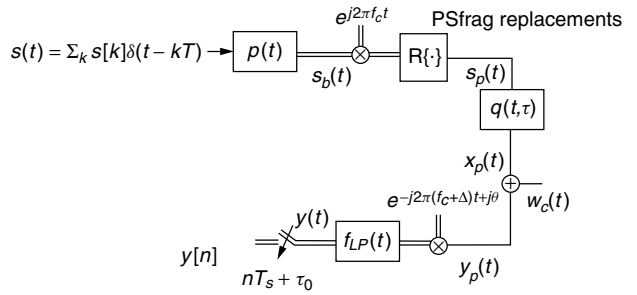


Figure 1. The transmission of linearly modulated baseband signal through a band-limited passband channel. Double lines are paths for complex signals and single lines for real signals. The operator $\Re\{\cdot\}$ takes the real part of its argument. The filter $f_{LP}(t)$ is the impulse response of an ideal lowpass filter.

information-carrying symbols $\{s[k]\}$, the baseband signal $s_b(t)$ is formed as

$$s_b(t) = \sum_k s[k]p(t - kT) \tag{1}$$

where $p(t)$ is the baseband pulse, T the symbol interval, and $1/T$ the symbol rate. For passband transmissions, where the transmitted signals do not contain any DC component, the symbol sequence $\{s[k]\}$ may be complex in general. If the transmission is at the baseband, then $s[k]$ is real. In this article, we will assume that $\{s[k]\}$ is a complex sequence and the results are also valid when $\{s[k]\}$ is a real.

If $s_b(t)$ is transmitted over a band-limited channel, the Nyquist criterion for choosing $p(t)$ needs to be satisfied. In particular, $p(t)$ should be such that

$$\frac{1}{T}\sum_{i=-\infty}^{\infty}\left|P\left(f + \frac{i}{T}\right)\right|^2 = 1 \tag{2}$$

where $P(f)$ is the Fourier transform of $p(t)$. A usual choice is from the class of square-root raised-cosine pulses. The minimum bandwidth pulse is the ideal lowpass filter with bandwidth $1/2T$. In practical implementations, the actual bandwidth is between $1/2T$ and $1/T$ for narrowband transmissions and much greater than $1/T$ for spread spectrum transmissions.

To transmit $s_b(t)$ through a particular frequency band, the baseband signal is converted to the passband signal $s_p(t)$ by (quadrature) amplitude modulation. Hence, the transmitted signal is represented as

$$\begin{aligned} s_p(t) = \Re\{s_b(t)e^{j2\pi f_c t}\} &= \Re\{s_b(t)\}\cos(2\pi f_c t) \\ &\quad - \Im\{s_b(t)\}\sin(2\pi f_c t), \end{aligned} \tag{3}$$

where the operator $\Re\{\cdot\}$ takes the real part of its argument and $\Im\{\cdot\}$ the complex part of its argument. The (real) passband signal is transmitted through a linear, possibly time-varying, propagation channel $q(t, \tau)$ whose output $x_p(t)$ is given by

$$\begin{aligned} x_p(t) &= \int q(t,\tau)s_p(t-\tau)\,d\tau \\ &= \Re\left\{e^{j2\pi f_c t}\int q(t,\tau)e^{-j2\pi f_c \tau}s_b(t-\tau)\,d\tau\right\} \\ &= \Re\left\{e^{j2\pi f_c t}\int q_b(t,\tau)s_b(t-\tau)\,d\tau\right\} \end{aligned} \tag{4}$$

where we note that the passband propagation channel $q(t, \tau)$ is always real and denote the baseband propagation channel as

$$q_b(t,\tau) \triangleq q(t,\tau)e^{-j2\pi f_c \tau} \tag{5}$$

The received *passband* signal $y_p(t)$ is corrupted by noise $w_c(t)$ assumed to be zero mean, white, and Gaussian. The passband signal is then converted back to the baseband signal by frequency downshifting and lowpass filtering:

$$\begin{aligned} y(t) &= f_{LP}(t) * [y_p(t)e^{-j2\pi(f_c+\Delta)t+j\theta}] \\ &= e^{j(2\pi\Delta t+\theta)}\int q_b(t,\tau)s_b(t-\tau)d_\tau + w(t) \end{aligned} \tag{6}$$

where Δ is the frequency offset and θ the phase offset, and $w(t)$ is zero mean complex Gaussian with constant power spectrum density within the spectral range of the signal. Substituting (1) into Eq. (6), we obtain

$$y(t) = e^{j2\pi\Delta t}\sum_k h_k(t)s[k] + w(t) \tag{7}$$

where

$$h_k(t) = e^{j\theta}\int q_b(t,\tau)p(t - kT - \tau)d\tau \tag{8}$$

Note that the transmitted signal is distorted by two major factors: the propagation channel $q_b(t, \tau)$ and carrier-phase synchronization errors Δ and θ. Typically, carrier synchronization is performed separately from channel estimation. Assuming that the frequency offset Δ has been corrected before channel estimation, we can let $\Delta = 0$ and combine the phase error θ with baseband channel $h_k(t)$.

The problem of channel estimation can then be formulated as estimating $h_k(t)$, which combines the propagation channel $q(t, \tau)$, the signal waveform $p(t)$, and the phase error θ. Notice that $p(t)$ is known in general, and that it can be exploited to improve channel estimation.

If the channel can be modeled as time-invariant within the interval that channel estimation is performed, we then have $q_b(t, \tau) = q_b(\tau)$, and

$$y(t) = \sum_k s[k]h(t - kT) \tag{9}$$

where

$$h(t) = e^{j\theta}\int q_b(\tau)p(t-\tau)d_\tau \tag{10}$$

is called the *composite baseband channel*, which is, in general, complex.

2.2. The Discrete-Time Model

For band-limited transmissions, the baseband signal $y(t)$ can be sampled without loss of information if the sampling rate $f_s = 1/T_s$ exceeds the Nyquist rate. This implies that the sampling rate should be at least the symbol rate. If the transmitted pulse $p(t)$ has the bandwidth that exceeds the minimum bandwidth of $1/2T$, the sampling rate should be higher than the symbol rate. We will assume that the received signal is "over" sampled at a rate G times the symbol rate, namely, $T = GT_s$. For narrowband

transmissions, $G = 2$ satisfies the Nyquist rate, whereas, for spread-spectrum communications, G should be greater than or equal to the spreading gain of the system.

Denote the sampled discrete-time baseband signals as

$$y[n] \triangleq y(nT_s + \tau_0), \quad w[n] \triangleq w(nT_s + \tau_0) \tag{11}$$

where τ_0 is the sampling phase. For time-invariant channels with synchronized carrier as defined in (9), the received data samples satisfy

$$y[n] = \sum_k h(nT_s - kGT_s + \tau_0)s[k] + w[n] \tag{12}$$

$$= \sum_k h[n - kG]s[k] + w[n], \quad h[n] \triangleq h(nT_s + \tau_0) \tag{13}$$

Note that, when the input sequence $s[n]$ is stationary, $y[n]$ is not stationary unless $G = 1$. In general, the oversampled signal $y[n]$ is cyclostationary.

2.2.1. The SIMO Model. A convenient model for the oversampled discrete-time channel is the vectorized single-input multioutput (SIMO) model shown in Fig. 2. This model is obtained by noting that, if the received signal is sampled G times faster than the symbol rate $1/T$, there are G samples per symbol period, and $y[n]$ can be split into G subsequences. Specifically, for $i = 1, \ldots, G$, denote

$$y_i[n] \triangleq y[nG + i - 1], \mathbf{y}[n] = [y_1[n], \ldots, y_G[n]]^T \tag{14}$$

$$w_i[n] \triangleq w[nG + i - 1], \mathbf{w}[n] = [w_1[n], \ldots, w_G[n]]^T \tag{15}$$

$$h_i[n] \triangleq h[nG + i - 1], \mathbf{h}[n] = [h_1[n], \ldots, h_G[n]]^T \tag{16}$$

We then have the SIMO model given by

$$\mathbf{y}[n] = \sum_{k=0}^{L} \mathbf{h}[k]s[n - k] + \mathbf{w}[n], \tag{17}$$

where L, referred to as the channel order, is such that $k > L$ and $k < 0$. We assume that L is finite.[1]

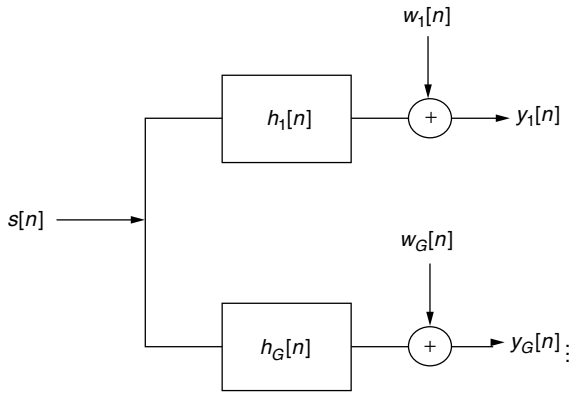


Figure 2. The SIMO vector channel model.

[1] For strictly band-limited signals, $L = \infty$. On the other hand, for strictly time-limited signals, L is always finite but, unfortunately, the Nyquist sampling frequency is ∞. In practice, L can be chosen large enough so that the model is sufficiently accurate.

If we collect all the transmitted symbols in a vector $\mathbf{s}$, all the received data samples in $\mathbf{y}$, and all channel parameters in $\mathbf{h}$

$$\mathbf{y} \triangleq \begin{pmatrix} \mathbf{y}[N-1] \\ \vdots \\ \mathbf{y}[0] \end{pmatrix}, \mathbf{h} \triangleq \begin{pmatrix} \mathbf{h}[0] \\ \vdots \\ \mathbf{h}[L] \end{pmatrix}, \mathbf{s} \triangleq \begin{pmatrix} s[N-1] \\ \vdots \\ s[-L] \end{pmatrix}$$

with noise vector $\mathbf{w}$ similarly defined, we have the following model equations:

$$\mathbf{y} = \mathcal{H}(\mathbf{h})\mathbf{s} + \mathbf{w} = \mathcal{F}(\mathbf{s})\mathbf{h} + \mathbf{w} \tag{18}$$

where $\mathcal{H}(\mathbf{h})$ is a block Toeplitz matrix generated from the channel $\mathbf{h}$ and $\mathcal{F}(\mathbf{s})$ a block Hankel matrix generated from the input $\mathbf{s}$ with dimensions matched to those of $\mathbf{y}$ and $\mathbf{s}$

$$\mathcal{H}(\mathbf{h}) = \begin{pmatrix} \mathbf{h}[0] & \cdots & \mathbf{h}[L] & & \\ & \ddots & & \ddots & \\ & & \mathbf{h}[0] & \cdots & \mathbf{h}[L] \end{pmatrix} \tag{19}$$

$$\mathcal{F}(\mathbf{s}) = \begin{pmatrix} s[N-1]\mathbf{I}_G & \cdots & s[N-L-1]\mathbf{I}_G \\ \vdots & \text{Block Hankel} & \vdots \\ s[0]\mathbf{I}_G & \cdots & s[-L]\mathbf{I}_G \end{pmatrix} \tag{20}$$

$$= \begin{pmatrix} s[N-1] & \cdots & s[N-L-1] \\ \vdots & \text{Hankel} & \vdots \\ s[0] & \cdots & s[-L] \end{pmatrix} \otimes \mathbf{I}_G, \tag{21}$$

where the operator $\otimes$ is the Kronecker product and $\mathbf{I}_G$ is the $G \times G$ identity matrix.

2.2.2. The MIMO Model. The SIMO model can be easily extended to incorporate systems that involve multiple users and multiple transmitting and receiving antennas. A general schematic is shown in Fig. 3, where there are K users, each transmitting a sequence of symbols $s_i[n]$ using a particular waveform. The signals $\{y_i(t)\}$ received by M receivers are distorted by noise, their corresponding propagation channels, and cross-interference. Let $\mathbf{y}_j[k]$ and $\mathbf{w}_j[n]$ be the received signal vector and the additive noise at the jth antenna, respectively, and $\mathbf{h}_{ji}[k]$ be the channel

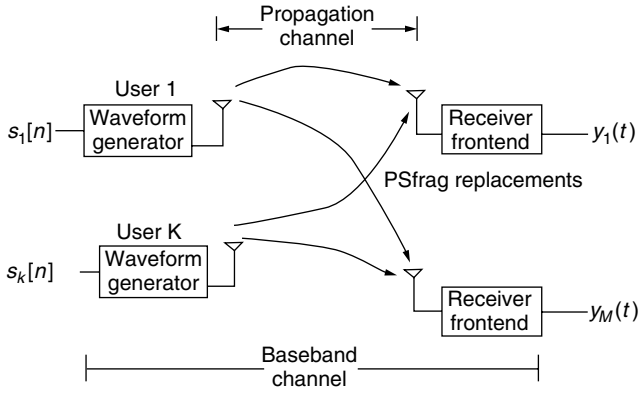


Figure 3. A general multiuser communication system.

between the ith user and the jth antenna. We then have the multiinput multioutput (MIMO) channel model

$$\mathbf{y}_j[n] = \sum_{i=1}^{K} \sum_{k} s_i[k]\mathbf{h}_{ij}[n-k] + \mathbf{w}_j[n], \quad j = 1, \ldots, M$$

Stacking data from all antennas as

$$\mathbf{y}[k] \triangleq \begin{pmatrix} \mathbf{y}_1[k] \\ \vdots \\ \mathbf{y}_M[k] \end{pmatrix}, \mathbf{h}_j[k] \triangleq \begin{pmatrix} \mathbf{h}_{1j}[k] \\ \vdots \\ \mathbf{h}_{Mj}[k] \end{pmatrix}, \quad j = 1, \ldots, M$$

one obtains the MIMO model

$$\mathbf{y}[k] = \sum_{i=1}^{M} \sum_{k} s_i[k]\mathbf{h}_i[n-k] + \mathbf{w}[n]$$

where $\mathbf{w}[n]$ is the noise vector similarly defined, and $\mathbf{h}_j[k]$ is the (vector) channel impulse response from the jth user to all receiving antennas. Again collecting all received data in a single vector $\mathbf{y}$ and transmitted symbols in $\mathbf{s} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T]^T$, we obtain the (batch) MIMO equation

$$\mathbf{y} = [\mathcal{H}(\mathbf{h}_1), \ldots, \mathcal{H}(\mathbf{h}_K)]\mathbf{s} + \mathbf{w} = [\mathcal{F}(\mathbf{s}_1), \ldots, \mathcal{F}(\mathbf{s}_K)]\mathbf{h} + \mathbf{w} \tag{22}$$

While we shall restrict our discussion to the single-user case multiple-antenna systems, many results apply directly to the estimation of MIMO channels described in Eq. (22).

3. CHANNEL ESTIMATION: GENERAL CONCEPTS

The objective of channel estimation is to infer channel parameters from the received signal. The function that maps the received signal and prior knowledge about the channel and pilot symbols is called the *estimator*. In this section, we discuss the formulation of the estimation problem. The development of specific estimators will be considered in Section 4.

3.1. Channel Estimation Techniques

3.1.1. Transmissions with Embedded Pilot Symbols. The development of channel estimation algorithms depends on the format of the transmitted symbols. Traditionally, the transmission is divided into two phases: the training phase and the transmission phase. In the training phase, pilot symbols known to the receiver are transmitted so that channel parameters can be estimated. The estimated channel is then used in the design of the receiver. When a feedback channel is available, the estimated channel can also be utilized to design an optimal transmitter.

For packet transmissions, especially in a wireless environment, it may be necessary that the channel be estimated separately for each packet. For example, the base station in a cellular system receives packets from different users, each with a different propagation channel. Therefore, pilot symbols need to be inserted into every data packet. The presence (or the absence) of pilot symbols, the number of pilot symbols, and the placement of pilot symbols all affect the parametric model from which channel estimators are derived. Figure 4 shows a typical packet format for the transmitted symbol $s[n]$ and its corresponding received signal $\mathbf{y}[n]$. The shaded area in the transmitted data frame is where pilot symbols are located. In general, there may be multiple pilot clusters $\{\mathbf{s}_{pi}\}$ whose location and values are known to the receiver.

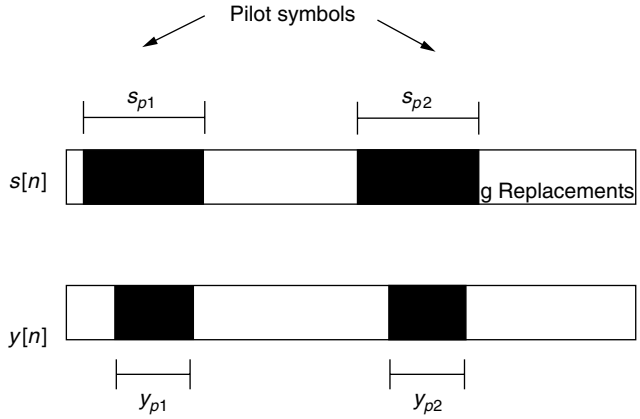


Figure 4. Signal frame structure. The shaded areas in the transmitted data frame are where pilot symbols are located. The shaded areas in the received signal frame are where samples corresponding to only pilot symbols are located.

3.1.2. Training-Based, Semiblind, and Blind Estimators. If the channel is memoryless, that is, if $L = 0$ in (17), then any received sample, say, $y[k]$, is a function of either a data symbol or a pilot symbol. However, if the channel has memory, $L > 0$, a received sample $\mathbf{y}[k]$ may be a function of (1) the (unknown) data symbols only, (2) the pilot symbols only, or (3) both data and pilot symbols. As illustrated by the shaded parts in Fig. 4, let $\mathbf{y}_{pi}$ be the cluster of received samples corresponding only to pilot cluster $\mathbf{s}_{pi}$. In other words, every sample in $\mathbf{y}_{pi}$ is a function of the pilot symbols in $\mathbf{s}_{pi}$ only. With this in mind, we consider three types of channel estimators, depending on the information that is used by the estimator.

- *Training-Based Channel Estimators*. A training-based channel estimator only uses data that correspond to the pilot symbols. In the case illustrated in Fig. 4, the estimator takes $\{\mathbf{y}_{p1}, \mathbf{s}_{p1}\}$ and $\{\mathbf{y}_{p2}, \mathbf{s}_{p2}\}$ as its input and produces an estimate of the channel. If $\mathbf{h}$ is the vector containing all channel parameters, $\mathbf{s}_p$ the vector containing all pilot clusters, and $\mathbf{y}_p$ the vector containing all received data corresponding to $\mathbf{s}_p$, a training-based channel estimator can be written as

$$\hat{\mathbf{h}} = G_T(\mathbf{s}_p, \mathbf{y}_p). \tag{23}$$

  Notice that although other received data also contain information about the channel, they are not utilized by the estimator. Training-based channel estimators are commonly used in practice. These estimators are easy to derive and analyze because the unknown data are not part of the observation. On the other hand, there needs to be a sufficient number of pilot

symbols present for good performance, and there are restrictions on how they should be placed in the data packet. For example, the size of pilot clusters must be at least $L+1$ in order to have one received data sample that is related to pilot symbols only.

- *Blind Channel Estimator*. When there are no pilot symbols available, the channel estimator is called "blind." In this case, the channel estimator uses the received signal $\mathbf{y}$ to estimate the channel

$$\hat{\mathbf{h}} = G_B(\mathbf{y}) \tag{24}$$

where the estimator $G_B(\cdot)$ is derived on the basis of certain qualitative information about the model. For example, although the input data are not known to the receiver, their statistical properties may be known. Other techniques include the exploitation of the finite-alphabet property of the source and certain parametric models of the channel. It is not obvious that blind channel estimation is even possible as neither the input nor the channel is known to the receiver. Indeed, such estimation is possible only under certain identifiability conditions, and the identification can be achieved only up to a scaling factor. In some applications such as terrestrial broadcasting of high-definition television (HDTV) where the requirement of efficient bandwidth utilization is stringent, pilot symbols may be so scarce that the receiver must acquire the channel without training. In such cases, blind channel estimation is necessary.

- *Semiblind Channel Estimator*. Between training-based and blind estimators is the class of semiblind channel estimators that utilize not only that part of signal corresponding to the training symbols but also the part corresponding to data symbols. In particular, a semiblind channel estimator takes $\{\mathbf{s}_{p1}, \mathbf{s}_{p2}, \mathbf{y}\}$ to generate a channel estimate. A semiblind channel estimator can be expressed as

$$\hat{\mathbf{h}} = G_{SB}(\mathbf{s}_p, \mathbf{y}) \tag{25}$$

By fully exploiting the information about the channel contained in the entire data record, semiblind channel estimation may provide considerable gain over training-based algorithms as shown in Section 4.2.

3.2. Performance Measure and Performance Bound

In estimating the channel, we can model $\mathbf{h}$ as a deterministic vector or a random vector with a certain probability distribution. If $\mathbf{h}$ is deterministic, we have the problem of point estimation whereas, when $\mathbf{h}$ is random, the estimation problem is Bayesian. In the following discussion, we will restrict ourselves to the case when $\mathbf{h}$ is deterministic but unknown.

The problem of channel estimation is to find an estimator $\hat{\mathbf{h}}$ that is close to the true channel under a certain performance measure. Typically, we will be concerned about the bias and covariance of the estimator defined by

$$\mathcal{B}(\hat{\mathbf{h}}) \triangleq E\{\hat{\mathbf{h}} - \mathbf{h}\}, \mathcal{V}(\hat{\mathbf{h}}) \triangleq E\{[\hat{\mathbf{h}} - E(\hat{\mathbf{h}})][\hat{\mathbf{h}} - E(\hat{\mathbf{h}})]^H\} \tag{26}$$

where both, in general, are functions of $\mathbf{h}$. If $\mathcal{B}(\mathbf{h}) = 0$ for all possible $\mathbf{h}$, then the estimator is unbiased. The performance of an estimator can also be measured by the covariance matrix of the estimation error

$$\mathcal{M}(\hat{\mathbf{h}}) \triangleq E(\hat{\mathbf{h}} - \mathbf{h})(\hat{\mathbf{h}} - \mathbf{h})^H \tag{27}$$

from which we obtain the *mean-square error*

$$E(\|\hat{\mathbf{h}} - \mathbf{h}\|^2) = tr\{\mathcal{M}(\hat{\mathbf{h}})\} \tag{28}$$

These definitions also apply to random channels.

In assessing the performance of the estimator, it is often useful to compare the covariance of the estimation error with the *Cramér–Rao bound* (CRB), which is a lower bound on the MSE of any unbiased estimator. Given a deterministic channel $\mathbf{h}$, assume that we have a well-defined probability density function $\mathbf{f}(\mathbf{y}; \mathbf{h})$. Viewed as a function of $\mathbf{h}$, $\boldsymbol{f}(\mathbf{y}; \mathbf{h})$ is the likelihood function of the channel parameter $\mathbf{h}$. The complex Fisher information matrix (FIM) $\boldsymbol{I}(\mathbf{h})$ is defined by

$$\mathbf{I}(\mathbf{h}) \triangleq E\{[\nabla_{\mathbf{h}^*} \ln f(\mathbf{y}; \mathbf{h})][\nabla_{\mathbf{h}^*}^H \ln f(\mathbf{y}; \mathbf{h})]\} \tag{29}$$

where the complex gradient operator applied to a real function $g(\mathbf{x})$ with complex argument $\mathbf{x} \in C^K$ is defined by

$$\nabla_{\mathbf{x}^*} g(\mathbf{y}) \triangleq \frac{1}{2}\begin{pmatrix} \dfrac{\partial g(\mathbf{x})}{\partial \Re\{x_1\}} \\ \vdots \\ \dfrac{\partial g(\mathbf{x})}{\partial \Re\{x_K\}} \end{pmatrix} + \frac{j}{2}\begin{pmatrix} \dfrac{\partial g(\mathbf{x})}{\partial \Im\{x_1\}} \\ \vdots \\ \dfrac{\partial g(\mathbf{x})}{\partial \Im\{x_K\}} \end{pmatrix} \tag{30}$$

Under regularity conditions [19,34], the MSE of any unbiased estimator $\hat{\mathbf{h}}$ is lower-bounded by $\mathbf{I}^{-1}$ $(\mathbf{h})$ and MSE by $tr\{\mathbf{I}^{-1}(\mathbf{h})\}$:

$$E(\hat{\mathbf{h}} - \mathbf{h})(\hat{\mathbf{h}} - \mathbf{h})^H \geq \mathbf{I}^{-1}(\mathbf{h}), \tag{31}$$

$$E(\|\hat{\mathbf{h}} - \mathbf{h}\|^2) \geq tr\{\mathbf{I}^{-1}(\mathbf{h})\} \tag{32}$$

An unbiased estimator that achieves the CRB is called *efficient*. The same expression also holds when $\mathbf{h}$ is random except that the expectation in (29) is taken over $\mathbf{y}$ and the channel vector $\mathbf{h}$.

3.3. Estimation Techniques

3.3.1. The Maximum-Likelihood Methods. One of the most popular parameter estimation algorithms is the maximum-likelihood (ML) method. The ML estimator can usually be derived in a systematic way by maximizing the likelihood function

$$\hat{\mathbf{h}}_{ML} = \arg\max_{\mathbf{h}\in\Theta} \boldsymbol{f}(\mathbf{y}; \mathbf{h}) \tag{33}$$

where Θ is the set of channels that satisfy certain constraints.

The ML estimator has a number of attractive properties. If an efficient estimator exists, it must be an ML estimator; it can be shown that the class of maximum-likelihood estimators are asymptotically efficient [23],

although examples exist that the ML estimator may perform poorly when the data size is small.

While the ML estimator is conceptually simple, the implementation of the ML estimator is sometimes computationally intensive. Furthermore, the optimization of the likelihood function in (33) is often hampered by the existence of local maxima. Therefore, it is desirable that effective initialization techniques are used in conjunction with ML estimation.

3.3.2. The Moment Methods. For some applications, the knowledge of the model is incomplete and the likelihood function cannot be specified explicitly. In such cases, the method of moments may be applied. Suppose that we know the explicit form that the ith moments $\mathbf{M}_i(\mathbf{h})$ of $\mathbf{y}$ relate to the channel parameter. The moment estimator is then given by matching the moment functions $\mathbf{M}_i(\mathbf{h})$ with moments estimated from the data. Often, simple estimators can be obtained from solving for $\hat{\mathbf{h}}$ directly from

$$\mathbf{M}_i(\mathbf{h}) = \hat{\mathbf{M}}_i \tag{34}$$

The matching of moments can also be done using least-squares techniques.

4. ESTIMATION OF SIMO CHANNELS

We now apply the performance bound and general estimation techniques to the estimation of the SIMO channel model (17) developed in Section 2. Training-based estimation is presented first followed by blind and semiblind estimation.

4.1. Training-Based Channel Estimation Algorithms

Training-based estimators use only those parts of the received signal corresponding to pilot symbols. If there is a single cluster of training symbols, we can use the model given in (18) assuming that all symbols $s[n]$ are known to the receiver, and

$$\mathbf{y} = \mathcal{F}(\mathbf{s})\mathbf{h} + \mathbf{w},\ \mathcal{F}(\mathbf{s}) = \mathcal{F}_1(\mathbf{s}) \otimes \mathbf{I}_G \tag{35}$$

where the noise vector $\mathbf{w}$ is zero mean, Gaussian with covariance $\sigma^2\mathbf{I}$, and

$$\mathcal{F}_1(\mathbf{s}) \triangleq \begin{pmatrix} s[N-1] & \cdots & s[N-L-1] \\ \vdots & \text{Hankel} & \vdots \\ s[0] & \cdots & s[-L] \end{pmatrix}$$

If there are multiple clusters, the preceding equations remain valid with $\mathcal{F}(\mathbf{s})$ replaced by a stack of $\mathcal{F}(\mathbf{S}_{pi})$, each corresponding to one cluster of pilot symbols.

4.1.1. Performance Bound and Identifiability. The likelihood function $f(\mathbf{y}, \theta)$ for the parameter $\boldsymbol{\theta} = \begin{pmatrix} \mathbf{h} \\ \sigma^2 \end{pmatrix}$ is given by

$$f(\mathbf{y};\theta) = \frac{1}{(\pi\sigma^2)^N} \exp\left\{-\frac{1}{\sigma^2}\|\mathbf{y} - \mathcal{F}(\mathbf{s})\mathbf{h}\|^2\right\} \tag{36}$$

The (complex) Fisher information matrix is given by

$$I(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} \mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}) & \mathbf{0} \\ \mathbf{0} & \dfrac{N}{\sigma^2} \end{pmatrix} \tag{37}$$

and the CRB for the training-based channel estimators is

$$CRB_T(\mathbf{h}) = \sigma^2[\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})]^{-1} \tag{38}$$

assuming the inverse exists.

The assumption that $\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})$ is invertible is significant. If this condition is not satisfied, the channel is not identifiable. Specifically, if $\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})$ is not invertible, columns of $\mathcal{F}(\mathbf{s})$ are linearly dependent. Hence, there exists a vector $\Delta\mathbf{h}$ such that

$$\mathcal{F}(\mathbf{s})\Delta\mathbf{h} = \mathbf{0}$$

which implies that

$$\mathbf{y} = \mathcal{F}(\mathbf{s})(\mathbf{h} + \gamma\,\Delta\mathbf{h}) + \mathbf{w}$$

for any γ. In other words, the estimation error can be arbitrarily large.

4.1.2. Design of Pilot Sequence. The condition of identifiability imposes certain constraints on the training sequence. To ensure that $\mathcal{F}(\mathbf{s})$ has full column rank, it is necessary and sufficient that $\mathcal{F}_1(\mathbf{s})$ have full column rank. An equivalent condition is that the *linear complexity*[2] [5] of the pilot sequence $s[n]$ should be greater than L. This implies that, to estimate a set of parallel channels of order L, the minimum number of training symbols must be greater than $2L$.

Note that the CRB in (38) is not a function of the channel parameter. It is, however, a function of the transmitted (pilot) symbol vector $\mathbf{s}$. Since the CRB can be achieved by the ML estimator described below, the training sequence should be designed to minimize the CRB. Specifically, we may choose the training sequence with constant amplitude σ_s according to the following optimization:

$$\min_{\mathbf{s}} tr\{(\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}))^{-1}\}$$

It can be shown that among all possible transmitted symbols with constant amplitude σ_s, the one that minimizes the CRB satisfies the orthogonality condition

$$\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}) = N\sigma_s^2\mathbf{I}$$

The sequence that gives the minimum CRB can be chosen from points on the circle with radius σ_s in the complex plane [6].

[2] The linear complexity of a sequence $s[n]$ is defined as the smallest number c such that there exist α_i such that $s[n] = \sum_{i=1}^{c} \alpha_i s[n-i]$ for all $n \geq c$.

4.1.3. THE ML ESTIMATOR

The maximum-likelihood (ML) estimator of the channel is given by

$$\mathbf{h}_{ML} = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathcal{F}(\mathbf{s})\mathbf{h}\|^2 \tag{39}$$

$$= [\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s})]^{-1}\mathcal{F}^H(\mathbf{s})\mathbf{y}$$

$$= (\{[\mathcal{F}_1^H(\mathbf{s})\mathcal{F}_1(\mathbf{s})]^{-1}\mathcal{F}_1^H(\mathbf{s})\} \otimes \mathbf{I}_G)\mathbf{y} \tag{40}$$

assuming again the inverse exists. It is easily verified that

$$E(\hat{\mathbf{h}}_{ML}) = \mathbf{h},\ \mathcal{V}(\hat{\mathbf{h}}_{ML}) = \sigma^2(\mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}))^{-1} \tag{41}$$

In other words, the ML channel estimator is efficient. When the optimal training sequence is used, from (41), we obtain

$$\mathcal{V}(\hat{\mathbf{h}}_{ML}) = \frac{1}{N} \xrightarrow{N\to\infty} 0 \tag{42}$$

Hence, the estimator is consistent and the estimation error decreases to zero at the rate of $1/N$.

The implementation of the ML estimator can be simplified by treating the G subchannels in Fig. 2 separately. Let $\mathbf{h}^{(i)}$ be the channel vector containing the impulse response of the ith subchannel, and $\mathbf{y}^{(i)}$ be the observation corresponding to the ith subchannel. Because the assumption that the noise samples in $\mathbf{w}$ are independent, we have

$$\hat{\mathbf{h}}^{(i)} = [\mathcal{F}_1^H(\mathbf{s})\mathcal{F}_1(\mathbf{s})]^{-1}\mathcal{F}_1^H(\mathbf{s})\mathbf{y}^{(i)} \tag{43}$$

which involves the inversion of a $(L_i+1)\times(L_i+1)$ Hermitian matrix where L_i is the order of the ith channel. In contrast, the inversion of a $G(L+1)\times G(L+1)$ matrix is involved in (40).

4.1.4. Recursive Least Squares. The ML channel estimator, under the assumption that $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is zero mean, Gaussian with covariance $\sigma^2\mathbf{I}$, is also the least-squares estimator defined by (40). In addition to reducing computation complexity by avoiding direct matrix inverse, the recursive least-squares (RLS) algorithm computes the channel estimate recursively, allowing updates as more data become available.

Since all sub-channels can be estimated independently, without loss of generality, we assume that $G = 1$. Suppose that we have already obtained the LS estimate $\hat{\mathbf{h}}_n$ using all observation $\mathbf{y}_n$ up to time n and their corresponding input symbols $\mathbf{s}_n$ defined by

$$\mathbf{y}_n = \begin{pmatrix} y[n] \\ \vdots \\ y[0] \end{pmatrix},\ \mathbf{s}_n = \begin{pmatrix} s[n] \\ \vdots \\ s[-L] \end{pmatrix} \tag{44}$$

From (40), we have

$$\hat{\mathbf{h}}_n \triangleq \underbrace{[\mathcal{F}^H(\mathbf{s}_n)\mathcal{F}(\mathbf{s}_n)]}_{\mathbf{R}_n}{}^{-1}\ \underbrace{\mathcal{F}^H(\mathbf{s}_n)\mathbf{y}_n}_{\mathbf{r}_n}$$

$$= \mathbf{R}_n^{-1}\mathbf{r}_n = \mathbf{P}_n\mathbf{r}_n \tag{45}$$

where $\mathbf{P}_n \triangleq \mathbf{R}_n^{-1}$. Letting $\mathbf{s}_{L+1}[n] \triangleq [s[n], \ldots, s[n-L]]^T$, we note that $\mathbf{R}_n$ can be computed recursively

$$\mathbf{R}_n = \mathbf{R}_{n-1} + \mathbf{s}_{L+1}^*[n]\mathbf{s}_{L+1}^T[n]$$

Recursive relations also hold for $\mathbf{P}_n$ and $\mathbf{r}_n$

$$\mathbf{P}_n \triangleq \mathbf{R}_n^{-1} = \mathbf{P}_{n-1} - \frac{\mathbf{P}_{n-1}\mathbf{s}_{L+1}^*[n]\mathbf{s}_{L+1}^T[n]\mathbf{P}_{n-1}^H}{1 + \mathbf{s}_{L+1}^T[n]\mathbf{P}_{n-1}\mathbf{s}_{L+1}^*[n]} \tag{46}$$

$$\mathbf{r}_n \triangleq \mathcal{F}^H(\mathbf{s}_n)\mathbf{y}_n = \mathbf{r}[n-1] + y[n]\mathbf{s}_{L+1}^*[n] \tag{47}$$

Suppose now that we are made available the next pilot $s[n+1]$ and the corresponding observation $y[n+1]$. Then, the ML estimator using data up to time $n+1$ can be updated from the previous estimate by

$$\hat{\mathbf{h}}_{n+1} \triangleq \hat{\mathbf{h}}_n + \mathbf{g}_{n+1}\varepsilon[n+1] \tag{48}$$

where $\varepsilon[n+1]$ is the error of the predicted observation using $\hat{\mathbf{h}}[n]$

$$\varepsilon[n+1] = y[n+1] - \mathbf{s}_{L+1}^T[n+1]\hat{\mathbf{h}}_n, \tag{49}$$

and $\mathbf{g}_{n+1}$ is the gain vector

$$\mathbf{g}_{n+1} = \frac{\mathbf{P}_n\mathbf{s}_{L+1}^*[n+1]}{1 + \mathbf{s}_{L+1}^T[n+1]\mathbf{P}_n\mathbf{s}_{L+1}^*[n+1]} \tag{50}$$

It is interesting to note that the amount of update in the channel estimate depends on the prediction error by the channel estimate. The RLS algorithm reduces the computation of the batch ML estimation to $\mathcal{O}((L+1)^2)$.

4.1.5. The LMS Algorithm. For its simplicity, the LMS algorithm, originally proposed by Widrow and Hoff [36], is perhaps the most widely used adaptive estimation algorithm. It resembles the RLS update and may be expressed as follows

$$\hat{\mathbf{h}}_{n+1} \triangleq \hat{\mathbf{h}}_n + \mu\mathbf{s}_{L+1}^*[n+1]\varepsilon[n+1] \tag{51}$$

where the computationally more expensive gain vector $\mathbf{g}_{n+1}$ in RLS is replaced by the readily available input vector and a constant-step-size μ for the update. The derivation of the LMS algorithm does not have a direct connection to the ML estimation. Under the assumption that the input sequence $s[n]$ is random, the LMS can be viewed as the stochastic gradient implementation of the minimization of the average prediction error

$$\min_{\mathbf{h}} E\left\{|y[n] - \sum_i h[i]s[n-i]|^2\right\} \tag{52}$$

where the expectation is taken over the noise and the input process.

4.2. Blind and Semiblind Channel Estimation Algorithms

We now consider the semiblind and blind channel estimation problem where the channel estimator uses the entire data record.

The input sequence can be partitioned into two parts: the pilot symbols $\mathbf{s}_p$ and data symbols $\mathbf{s}_d$. With the corresponding partition in the channel matrix, we have

$$\mathbf{y} = \mathcal{H}_p(\mathbf{h})\mathbf{s}_p + \mathcal{H}_d(\mathbf{h})\mathbf{s}_d + \mathbf{w} \tag{53}$$

where $\mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is again assumed to be zero mean Gaussian with covariance $\sigma^2\mathbf{I}$. The estimation problem can now be formulated on the basis of two models of the input data vector $\mathbf{s}_d$. The *deterministic model* assumes that $\mathbf{s}_d$ is a deterministic, unknown nuisance parameter in the channel estimation. The *stochastic model*, on the other hand, assumes that $\mathbf{s}_d$ is random with certain distribution. These two models lead to different likelihood functions and therefore, different performance bounds and estimation algorithms. The choice of modeling depends naturally on the application. Typically, if the channel is to be estimated using a small number of samples, the deterministic model is more effective.

4.2.1. The Performance Bound. Under the deterministic model, the likelihood function of the unknown parameter $\boldsymbol{\theta} = [\mathbf{h}^T \mathbf{s}_d^T \sigma^2]^T$ is given by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \frac{1}{\pi^N \sigma^{2N}} \exp\left\{-\frac{1}{\sigma^2}\|\mathbf{y} - \mathcal{H}_p(\mathbf{h})\mathbf{s}_p - \mathcal{H}_d(\mathbf{h})\mathbf{s}_d\|^2\right\} \tag{54}$$

The complex Fisher information matrix is given by

$$I(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} \mathcal{F}^H(\mathbf{s})\mathcal{F}(\mathbf{s}) & \mathcal{F}^H(\mathbf{s})\mathcal{H}_d(\mathbf{h}) & \mathbf{0} \\ \mathcal{H}_d^H(\mathbf{h})\mathcal{F}(\mathbf{s}) & \mathcal{H}_d^H(\mathbf{h})\mathcal{H}_d(\mathbf{h}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dfrac{N}{\sigma^2} \end{pmatrix} \tag{55}$$

Using the block matrix inversion formula, we obtain the CRB for any unbiased semiblind channel estimator $\hat{\mathbf{h}}$

$$\mathrm{CRB}_{SB}(\mathbf{h}) = \sigma^2[\mathcal{F}^H(\mathbf{s})\mathcal{P}^{\perp}_{\mathcal{H}_d(\mathbf{h})}\mathcal{F}(\mathbf{s})]^{-1} \tag{56}$$

where

$$\mathcal{P}^{\perp}_{\mathcal{H}_d(\mathbf{h})} \triangleq \mathbf{I} - \mathcal{H}_d(\mathbf{h})[\mathcal{H}_d^H(\mathbf{h})\mathcal{H}_d(\mathbf{h})]^{-1}\mathcal{H}_d(\mathbf{h}) \tag{57}$$

is the projection matrix onto the null space of $\mathcal{H}_d(\mathbf{h})$. It is clear from the above expressions and (38) that

$$\mathrm{CRB}_{SB}(\mathbf{h}) \le \mathrm{CRB}_T(\mathbf{h})$$

with equality when all symbols are known.

The derivation for the CRB under the stochastic model is more complicated unless one assumes that the input sequence is Gaussian. Details can be found in Ref. 7.

4.2.2. The Design of Pilot Symbols and Their Placement. The design of training for semiblind channel estimation involves the joint design of the number of pilot symbols, the pilot symbols, and their placements. The CRB can be used as the performance measure for this design.

It is natural to expect that more pilot symbols lead to better performance. On the other hand, increasing the number of pilot symbols reduces the number of data symbols transmitted in a packet. The gain in performance can be evaluated against the percentage of known symbols in a packet using (56). Figure 5 shows an example of the relation between the MSE and the percentage η of known symbols in the data packet for a multipath channel of order $L = 5$ at 10 dB SNR. Notice that $\eta = 100\%$ corresponds to the performance of the training-based ML algorithm. It can be seen that the gain of using all pilot symbols ($\eta = 100\%$) over that of using 1% pilot symbols is about 45 dB, and about 40 dB of gain can already be achieved at $\eta = 20\%$.

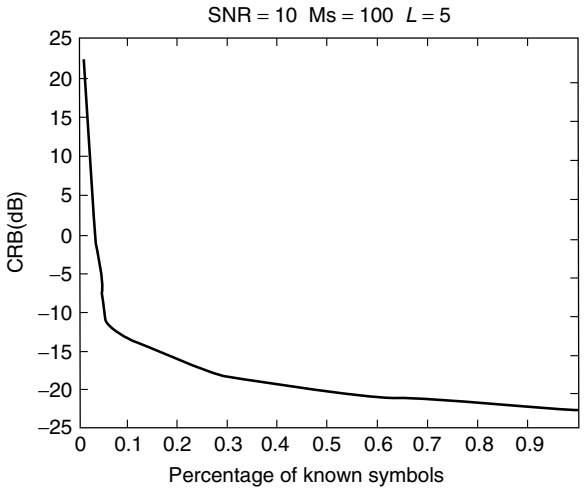


Figure 5. CRB for a multipath channel with $L = 5$; SNR = 10 dB.

Given the percentage of pilot symbols in a data packet, the problem of pilot design can be formulated as minimizing $tr\{\mathrm{CRB}_{SB}(\mathbf{h})\}$ among choices of pilot symbols and their placement. This optimization, unfortunately, depends on the channel coefficients. For random channels, however, the minimization of CRB does lead to the optimal design of pilot symbols and their placement, which are independent of the channel [9].

4.2.3. The ML Estimation. When some or all of the input symbols are unknown, the likelihood function of the channel parameters and the unknown symbols depends on the model assumed for the unknown symbols. We then have the so-called deterministic maximum likelihood (DML) where the unknown input symbols are deterministic, and the stochastic maximum likelihood (SML), where the unknown symbols are assumed to be random with some known distribution.

4.2.3.1. The SML Estimation. While the input vector $\mathbf{s}$ is unknown, it may be modeled as a random vector with known distribution. In such a case, the likelihood function of the channel parameter $\mathbf{h}$ can be obtained by

$$f(\mathbf{y}; \mathbf{h}) = \int f(\mathbf{y}|\mathbf{s}_d; \mathbf{h}) f(\mathbf{s}_d) d\mathbf{s}_d \tag{58}$$

where $f(\mathbf{s}_d)$ is the marginal pdf of the unknown data vector and $f(\mathbf{y}|\mathbf{s}_d; \mathbf{h})$ is the likelihood function of $\mathbf{h}$ for a particular choice of $\mathbf{s}_d$. If the input symbols $\{s[k]\}$ take, with equal probability, a finite number of values, the data vector $\mathbf{s}_d$ also takes values from the signal set $\{\mathbf{v}_1, \ldots, \mathbf{v}_Q\}$ with equal probability.

The likelihood function of the channel parameter is then given by

$$f(\mathbf{y};\mathbf{h}) = \sum_{i=1}^{Q} f(\mathbf{y}|\mathbf{s}_d = \mathbf{v}_i;\mathbf{h})\Pr(\mathbf{s}_d = \mathbf{v}_i)$$
$$= C\sum_{i=1}^{Q} \exp\left\{-\frac{\|\mathbf{y} - \mathcal{F}(\mathbf{v}_i)\mathbf{h} - \mathcal{F}(\mathbf{s}_p)\mathbf{h}\|^2}{\sigma^2}\right\} \quad (59)$$

where C is a constant. The stochastic maximum-likelihood estimator is given by

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \sum_{i=1}^{Q} \exp\left\{-\frac{\|\mathbf{y} - \mathcal{F}(\mathbf{v}_i)\mathbf{h} - \mathcal{F}(\mathbf{s}_p)\|^2}{\sigma^2}\right\} \quad (60)$$

The maximization of the likelihood function defined in (58) is in general difficult. The expectation–maximization (EM) algorithm [2,8] can be applied to transform the complicated optimization to a sequence of quadratic optimizations. Kaleh and Vallet [18] first applied the EM algorithm to the equalization of communication channels with the input sequence having the finite-alphabet property. By using a *hidden Markov model* (HMM), the authors of Ref. 18 developed a batch (offline) procedure that includes the so-called forward and backward recursions [27]. Unfortunately, the complexity of this algorithm increases exponentially with the channel memory. To relax the memory requirements and facilitate channel tracking, "online" sequential approaches have been proposed [30,31,35] for a general input, and for an input with finite alphabet properties under a HMM formulation [21]. Given the appropriate regularity conditions [30] and a good initialization, it can be shown that these algorithms converge (almost surely and in the mean square sense) to the true channel value.

4.2.3.2. DML Estimation. When the noise is zero-mean Gaussian with covariance $\sigma^2 I$, the DML estimator can be obtained by the nonlinear least-squares optimization

$$\{\hat{\mathbf{h}}, \hat{\mathbf{s}}_d\} = \arg\min_{\mathbf{h},\mathbf{s}_d} \|\mathbf{y} - \mathcal{H}_p(\mathbf{h})\mathbf{s}_p - \mathcal{H}_d(\mathbf{h})\mathbf{s}_d\|^2 \quad (61)$$

The joint minimization of the likelihood function with respect to both $\mathbf{h}$ and $\mathbf{s}_d$ is also difficult in general. However, for the general estimation model (18) considered here, the observation vector $\mathbf{y}$ is linear in both the channel and the input parameters individually. We therefore have a separable nonlinear least-squares problem that can be solved sequentially:

$$\{\hat{\mathbf{h}}, \hat{\mathbf{s}}_d\} = \arg\min_{\mathbf{s}_d}\{\min_{\mathbf{h}} \|\mathbf{y} - \mathcal{F}(\mathbf{s})\mathbf{h}\|^2\} \quad (62)$$
$$= \arg\min_{\mathbf{h}}\{\min_{\mathbf{s}_d} \|\mathbf{y} - \mathcal{H}(\mathbf{h})\mathbf{s}\|^2\} \quad (63)$$

If we are interested only in estimating the channel, the preceding minimization can be rewritten as

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \| \underbrace{(\mathbf{I} - \mathcal{H}(\mathbf{h})\mathcal{H}^{\dagger}(\mathbf{h}))}_{\mathcal{P}(\mathbf{h})} \mathbf{y}\|^2 = \arg\min_{\mathbf{h}} \|\mathcal{P}(\mathbf{h})\mathbf{y}\|^2 \quad (64)$$

where $\mathcal{P}(\mathbf{h})$ is a projection transform of $\mathbf{y}$ into the orthogonal complement of the range space of $\mathcal{H}(\mathbf{h})$, or the noise subspace of the observation. Discussions of algorithms of this type can be found in an earlier study [32].

Similar to the hidden Markov model (HMM) for the statistical maximum-likelihood approach, the finite-alphabet properties of the input sequence can also be incorporated into the deterministic maximum-likelihood methods. These algorithms, first proposed by Seshadri [28] and Ghosh and Weber [12], iterate between estimates of the channel and the input. At iteration k, with an initial guess of the channel $\mathbf{h}^{(k)}$, the algorithm estimates the input sequence $\mathbf{s}_d^{(k)}$ and the channel $\mathbf{h}^{(k+1)}$ for the next iteration by

$$\mathbf{s}_d^{(k)} = \arg\min_{\mathbf{s}_d \in S} \|\mathbf{y} - \mathcal{H}(\mathbf{h}^{(k)})\mathbf{s}\|^2 \quad (65)$$

$$\mathbf{h}^{(k+1)} = \arg\min_{\mathbf{h}} \|\mathbf{y} - \mathcal{H}(\mathbf{s}^{(k)})\mathbf{h}\|^2 \quad (66)$$

where $\mathcal{S}$ is the (discrete) domain of $\mathbf{s}_d$. The optimization in (66) is a linear least-squares problem whereas the optimization in (65) can be achieved by using the Viterbi algorithm [10]. The convergence of such approaches is not guaranteed in general.

Although the ML channel estimator usually provides better performance, the computation complexity and the existence of local optima are the two major impediments. Next we present two classes of suboptimal techniques that avoid the problem of local optima with significantly reduced computation complexity.

4.2.4. Moment Techniques: The Subspace Algorithms. Subspace techniques convert the problem of blind or semiblind channel estimation to the identification of a one-dimensional subspace that contains the channel vector. By exploiting the multichannel aspects of the channel, many of these techniques lead to a constrained quadratic optimization

$$\hat{\mathbf{h}} = \arg\min_{\|\mathbf{h}\|=1} \mathbf{h}^H Q(\mathbf{y}, \mathbf{s}_p)\mathbf{h} \quad (67)$$

where $Q(\mathbf{y}, \mathbf{s}_p)$ is a positive-definite matrix constructed from the observation and pilot symbols. The solution to the preceding optimization is then given by the eigenvector of $Q(\mathbf{y}, \mathbf{s}_p)$ associated with the minimum eigenvalue.

A simple yet informative approach [37] illustrates the basic idea in a noiseless two-channel scenario. From Fig. 2, if there is no noise, the received signals from the two channels satisfy the relation

$$y_1[n] = h_1[n] * s[n], y_2[n] = h_2[n] * s[n] \quad (68)$$

where $*$ is the linear convolution. Consequently, we have

$$y_1[n] * h_2[n] = y_2[n] * h_1[n] \quad (69)$$

Since the convolution operation is linear with respect to the channel and $y_i[n]$ is available, the above equation is equivalent to solving a homogeneous linear equation

$$\mathbf{Rh} = \mathbf{0} \quad (70)$$

where $\mathbf{R}$ is a matrix made of observations from the two channels. It can be shown that under certain identifiability conditions [32], the null space of $\mathbf{R}$ has dimension 1, which means that the channel can be identified up to a constant. When there is noise, the channel estimator can be obtained from a constrained quadratic optimization

$$\hat{\mathbf{h}} = \arg \min_{\|\mathbf{h}\|=1} \mathbf{h}^H \mathbf{R}^H \mathbf{R} \mathbf{h}, \tag{71}$$

which implies that $\hat{\mathbf{h}}$ is the eigenvector corresponds to the smallest eigenvalue of $\mathbf{Q} = \mathbf{R}^t \mathbf{R}$.

Some insight into the identifiability condition can be gained in the frequency domain. Equation (68) implies that

$$\frac{y_1(z)}{y_2(z)} = \frac{h_1(z)}{h_2(z)}$$

It is clear that if the two subchannels have common zeros, it is not possible to obtain all the zeros from the observation $\{y_1(z), y_2(z)\}$.

4.2.5. Projection Algorithms. The subspace algorithms are batch algorithms, and they are not easily amenable to adaptive forms. The projection based techniques [1,11,29,33,39], on the other hand, convert the problem of channel estimation to the classic problem of linear prediction or smoothing. As a result, these estimators can be implemented adaptively in time and recursively with respect to the delay spread of the channel. The first projection-based algorithm was proposed by Slock [29], where linear prediction is used to obtain the subspace of the channel matrix. Subsequent development [1,11] based on linear predictions assumed that the input sequence is a white sequence. Under the deterministic model, a least-squares smoothing (LSS) technique was developed [33,39] that offers finite sample convergence property in the absence of noise. It also allows a lattice filter implementation that is recursive both in time and the delay spread of the channel. In fact, both the channel and the input sequence can be obtained simultaneously by using oblique projections [38].

1.5. BIBLIOGRAPHY NOTES

The modeling of linearly modulated signals can be found in standard textbooks [22,26]. For fading dispersive channels encountered in wireless communications, earlier treatments can be found in Refs. 3, 16, and 20 and more recent developments in Ref. 4. The general theory of parameter estimation, including the Cramér-Rao bound, the maximum-likelihood estimation, and the moment methods, are presented in many books. See, for example, Lehmann [23] for the mathematical treatment of the subject and Refs. 19, 24, 25, and 34 from engineering application perspectives. The estimation of complex parameters is discussed by Kay [19].

The training-based channel estimation under additive white Gaussian noise is a form of a linear least-squares problem, which is discussed extensively in Ref. 17. See also Ref. 15 for various adaptive implementations. A survey of blind channel estimation algorithms can be found in Ref. 32. The problem of semiblind channel estimation is discussed in detail in Ref. 7. Articles about more recent trends in channel estimation and equalization can be found in Refs. 13 and 14.

BIOGRAPHY

Lang Tong received his B.E. degree from Tsinghua University, Beijing, China, in 1985, and M.S. and Ph.D. degrees in electrical engineering in 1987 and 1990, respectively, from the University of Notre Dame, Indiana. He was a postdoctoral research affiliate at the Information Systems Laboratory, Stanford University, in 1991. Currently, he is an associate professor in the School of Electrical and Computer Engineering, Cornell University, Ithaca, New York.

Dr. Tong received Young Investigator Award from the Office of Naval Research in 1996, and the Outstanding Young Author Award from the IEEE Circuits and Systems Society. His areas of interest include statistical signal processing, adaptive receiver design for communication systems, signal processing for communication networks, and information theory.

BIBLIOGRAPHY

1. K. Abed-Meraim, E. Moulines, and P. Loubaton, Prediction error method for second-order blind identification, *IEEE Trans. Signal Process.* **SP-45**(3): 694–705 (March 1997).
2. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.* **41**: 164–171 (1970).
3. P. A. Bellow, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* 360–393 (Dec. 1963).
4. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: Information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* **44**(4): (Oct. 1998).
5. R. E. Blahut, *Algebraic Methods for Signal Processing and Communications Coding*, Springer-Verlag, New York, 1992.
6. D. C. Chu, Polyphase codes with good periodic correlation properties, *IEEE Trans. Inform. Theory* **3**(4): 531–532 (July 1972).
7. E. de Carvalho and D. T. M. Slock, Semi-blind Methods for FIR multichannel estimation, G. Giannakis, Y. Hua, P. Stoica, and L. Tong, eds., *Signal Processing Advances in Wireless & Mobile Communications: Trends in Channel Estimation and Equalization*, Prentice-Hall, Englewood Cliffs, NJ, 2001.
8. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Stat. Soc.* **39**(Ser. B): (1977).
9. M. Dong and L. Tong, Optimal design and placement of pilot symbols for channel estimation, *Proc. ICASSP'2001*, 2001 (an extended journal submission to the *IEEE Trans. Signal Process.* is available from *http://www.ece.cornell.edu/~ltong/pubj.html*).
10. G. D. Forney, The Viterbi algorithm, *IEEE Proc.* **61**: 268–278 (March 1972).

11. D. Gesbert and P. Duhamel, Robust blind identification and equalization based on multi-step predictor, *Proc. IEEE Int. Conf. Acoustics. Speech Signal Processing*, Munich, Germany, April 1997, Vol. 5, pp. 2621–2624.

12. M. Ghosh and C. L. Weber, Maximum-likelihood blind equalization, *Opt. Eng.* **31**(6): 1224–1228 (June 1992).

13. G. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal Processing Advances in Wireless Communications: Trends in Channel Estimation and Equalization*, PTR Prentice-Hall, Englewood Cliffs, NJ, 2001.

14. G. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal Processing Advances in Wireless Communications: Trends in Single- and Multi-User Systems*, PTR Prentice-Hall, Englewood Cliffs, NJ, 2001.

15. S. Haykin, *Array Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

16. T. Kailath, Channel characterization: Time-variant dispersive channels, in E. Baghdadi ed., *Lectures on Communication Theory*, McGraw-Hill, New York, Chap. 6.

17. T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 2000.

18. G. K. Kaleh and R. Vallet, Joint parameter estimation and symbol detection for linear or non linear unknown dispersive channels, *IEEE Trans. Commun.* **42**(7): 2406–2413 (July 1994).

19. S. Kay, *Modern Spectral Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

20. R. S. Kennedy, *Fading Dispersive Communication Channels*, Wiley-Interscience, New York, 1969.

21. V. Krishnamurthy and J. B. Moore, On-line estimation of hidden Markov model parameters based on Kullback-Leibler information measure, *IEEE Trans. Signal Process.* **41**(8): 2557–2573 (Aug. 1993).

22. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer, Norwell, MA, 1988.

23. E. L. Lehmann, *Theory of Point Estimation*, Chapman & Hall, New York, 1991.

24. H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, 1994.

25. B. Porat, *Digital Processing of Random Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

26. J. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, 2001.

27. L. Rabiner, A tutorial on hidden Markov Models and selected applications in speech recognition, *IEEE Proc.* **77**(2): 257–285 (Feb. 1989).

28. N. Seshadri, Joint data and channel estimation using fast blind trellis search techniques, *Proc. Globecom'90*, 1991, pp. 1659–1663.

29. D. Slock, Blind fractionally-spaced equalization, perfect reconstruction filterbanks, and multilinear prediction, In *Proc. ICASSP'94 Conf.*, Adelaide, Australia, April 1994.

30. D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley Series in Probability and Mathematical Statistics, New York, 1985.

31. D. M. Titterington, Recursive parameter estimation using incomplete data, *J. Roy Stat. Soc. B* **46**(2): 257–267 (1984).

32. L. Tong and S. Perreau, Multichannel blind channel estimation: From subspace to maximum likelihood methods, *IEEE Proc.* **86**(10): 1951–1968 (Oct. 1998).

33. L. Tong and Q. Zhao, Joint order detection and blind channel estimation by least squares smoothing, *IEEE Trans. Signal Process.* **47**(9): (Sept. 1999).

34. H. L. Van Trees, *Detection, Estimation and Modulation Theory*, Vol. 1, Wiley, New York, 1968.

35. E. Weinstein, M. Feder, and A. Oppenheim, Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure, *IEEE Trans. Signal Process.* **SP-38**(9): 1652–1654 (Sept. 1990).

36. B. Widrow and Jr. M. E. Hoff, Adaptive switching circuits, *IRE WESCON Conf. Rec.*, 1960, Vol. 4, pp. 96–104.

37. G. Xu, H. Liu, L. Tong, and T. Kailath, A Least-squares approach to blind channel identification, *IEEE Trans. Signal Process.* **SP-43**(12): 2982–2993 (Dec. 1995).

38. Z. Yu and L. Tong, Joint channel and symbol estimation by oblique projections, *IEEE Trans. Signal Process.* **49**(12) (Dec. 2001).

39. Q. Zhao and L. Tong, Adaptive blind channel estimation by least squares smoothing, *IEEE Trans. Signal Process.* **47**(11) (Nov. 1999).

CHANNEL TRACKING IN WIRELESS COMMUNICATION SYSTEMS

GREGORY E. BOTTOMLEY
HÜSEYIN ARSLAN
Ericsson Inc.
Research Triangle Park, North Carolina

1. INTRODUCTION

In digital wireless communication systems, information is transmitted to a receiver, as illustrated in Fig. 1. The transmitted information reaches the receiver after passing through a radio channel, which can be represented as an unknown, time-varying filter. For conventional, coherent receivers, the effect of the channel on the transmitted signal must be estimated to recover the transmitted information. For example, with binary phase shift keying (BPSK), binary information is represented as $+1$ and -1 symbol values. The radio channel can apply a phase shift to the transmitted symbols, possibly inverting the symbol values. As long as the receiver estimates what the channel did to the transmitted signal, it can accurately recover the information sent.

Channel estimation is a challenging problem in wireless communications. Transmitted signals are typically reflected and scattered, arriving at the receiver along multiple paths. How these signals interact depends on

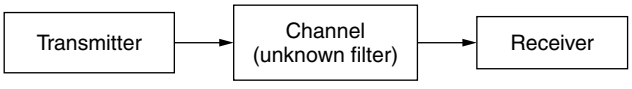


Figure 1. A wireless communication system.

their relative delays. When the relative delays are small compared to the transmitted symbol period, then different "images" of the *same* symbol arrive at the same time, adding either constructively or destructively. The overall effect is a random, fading channel response. When the relative path delays are on the order of a symbol period or more, then images of *different symbols* arrive at the same time. For example, when a particular symbol arrives at the receiver along one path, the previous symbol is arriving along another, delayed path. This is analogous to an acoustic echo and results in a more complicated channel response. Finally, because of the mobility of the transmitter, the receiver, or the scattering objects, the channel response can change rapidly with time.

This article provides an overview of channel tracking approaches commonly applied to digital cellular communication systems. Related work can be found in the study of system identification [1] and in high-frequency (HF) modem design [2].

As shown in Fig. 1, the channel impulse response is modeled as an unknown filter. Specifically, a finite-impulse-response (FIR) filter with discrete filter delays and coefficients is used. We focus on estimation of the channel coefficients, given a set of delays. Also, we focus on conventional demodulation approaches and single-antenna receivers.

In typical digital cellular systems, some part of the transmitted signal is known. In one approach, the transmitter periodically provides known *pilot symbols*, as illustrated in Fig. 2a, which can be used for channel estimation [3,4]. This approach is used in one of the downlink slot structures of the Telecommunications Industry Association/Electronics Industry Association/Interim Standard 136 (TIA/EIA/IS-136 or simply IS-136) system. In this time-division multiple-access (TDMA) system, information is transmitted in time slots. Within each time slot, clusters of known pilot symbols are provided to assist in channel estimation.

The pilot symbol approach has also been used in direct-sequence code-division multiple-access (DS-CDMA) systems. With these systems, each information symbol is represented by a sequence of "chip" symbols. This results in a spreading of the bandwidth (spread-spectrum), allowing multiple information signals to be transmitted in parallel at the same time. For convenience, we will refer to DS-CDMA systems as *wideband* systems and TDMA systems as *narrowband* systems. Pilot symbols are available in the following DS-CDMA systems: the IS-2000 system (uplink, mobile to base station) and the wideband CDMA (WCDMA) system (uplink and downlink).

In a second approach, a *pilot channel* is provided for channel estimation [5], as illustrated in Fig. 2b. This approach is related to the pilot tone approach, developed for narrowband systems [6]. The pilot channel approach is used for the downlink in the IS-95, IS-2000, and WCDMA systems. Usually the pilot channel is shared by many users and is stronger in power than an information channel.

A third approach is to provide a *training sequence* during part of the transmission, which can be used to provide an initial channel estimate. In this case, the channel must be tracked over the data portion of the signal using this signal in some way. This approach, illustrated in Fig. 2c, is used in one of the slot formats of the IS-136 system.

A training sequence is also used in the global system for mobile communications (GSM). This is a TDMA system with time slots that have short duration relative to the maximum rate of channel variation. As a result, the initial channel estimate obtained from the training sequence can be used to demodulate the data in the slot, without having to track the channel. Approaches for channel estimation in this situation are given in a separate article in this encyclopedia and elsewhere [7].

The article is organized as follows. In Section 2, a baseband-equivalent system model is given, including a model for the time-varying channel. Approximate channel models commonly used to develop tracking approaches are given in Section 3. In Section 4, filtering approaches to channel tracking are presented, based on either periodic pilot symbols or a pilot channel. Recursive approaches are presented in Section 5. Data-directed channel tracking is considered in Sections 6 and 7. Section 8 concludes the article.

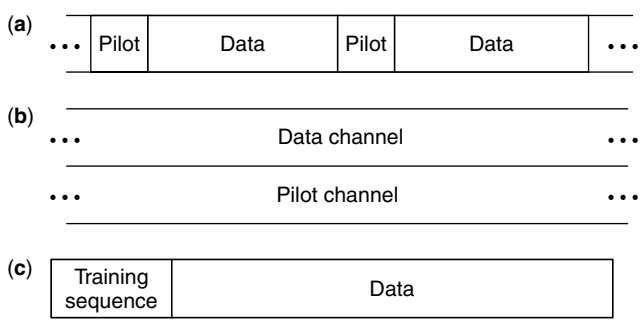


Figure 2. Systems with pilot information: (**a**) pilot symbols, (**b**) pilot channel, and (**c**) training sequence.

2. SYSTEM MODEL

A narrowband system model is presented and then extended to a wideband system model.

2.1. Narrowband System

A complex, baseband-equivalent system model is given in Fig. 3. The complex values correspond to in-phase (cosine) and quadrature (sine) components of the radio signal. At the transmitter, a sequence of digital symbols are transmitted using pulse shaping, giving

$$x(t) = \sum_k b_k f(t - kT) \tag{1}$$

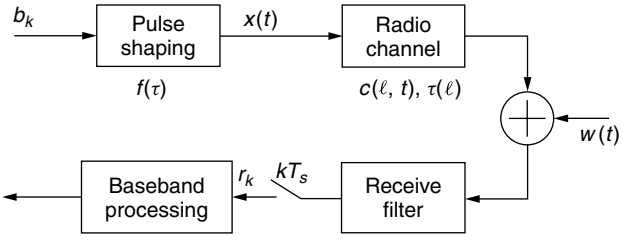


Figure 3. System model.

where b_k corresponds to the sequence of symbols, $f(\tau)$ is the pulse shape as a function of delay τ, and T is the symbol period. We assume either BPSK or quadrature phase shift keying (QPSK) modulation, so that all possible symbol values have the same amplitude ($|b_k|^2 = 1$).

The transmitted signal passes through a radio channel that can be modeled as a linear FIR filter. The resulting signal is received in the presence of noise, giving

$$y(t) = \sum_{\ell=0}^{L-1} c(\ell, t)x(t - \tau(\ell)) + w(t) \tag{2}$$

where $c(\ell, t)$ and $\tau(\ell)$ are the ℓth complex, time-varying channel coefficient and ℓth delay, respectively, and L is the number of channel taps. The noise term $w(t)$ is assumed to be white, complex (circular) Gaussian noise.

The delays are often assumed to be equally spaced; specifically, $\tau(\ell) = \ell T/M$, where M is an integer, and the spacing (T/M) for accurate modeling depends on the bandwidth of the system [8]. Typically M is 1 (symbol-spaced channel modeling) or 2 (fractionally-spaced channel modeling). For simplicity, symbol-spaced channel modeling is assumed, although extension to fractionally-spaced channel modeling is possible.

The coefficients represent the result of different multipath signal images adding together, constructively or destructively. They are well modeled as random variables. Specifically, they are modeled as uncorrelated, zero-mean complex Gaussian random variables [9]. This corresponds to "Rayleigh" fading, in that channel tap magnitudes (amplitudes) are Rayleigh-distributed. Also, the phases of the channel taps are uniformly distributed.

As the mobile transmitter or receiver moves, the phases of all multipath signal images change. This changes how the multipath images add together, so that the channel coefficient varies with time. This time variation is characterized by an autocorrelation function. The Jakes model [10], which is commonly used, assumes that the Gaussian channel coefficients have the following autocorrelation function:

$$R_\ell(\tau) = \mathrm{E}\{c^*(\ell, t)c(\ell, t + \tau)\} = \sigma_\ell^2 J_0(2\pi f_D \tau) \tag{3}$$

where the asterisk superscript denotes complex conjugation, index ℓ denotes the ℓth channel coefficient, τ is the autocorrelation delay, σ_ℓ^2 is the mean-square value of the channel coefficient, f_D is the Doppler spread, and $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind. The corresponding power spectrum of the fading process is shown in Fig. 4. This spectrum shows that the fading process has different frequency components, corresponding to different rates of change. Most of the energy is near the maximum frequency component, the Doppler spread (f_D).

The Doppler spread is proportional to the radio carrier frequency and the speed of the transmitter or receiver. For the examples given, the carrier frequency is either ~900 MHz or ~2 GHz. At a high vehicle speed of 100 km/h, these carrier frequencies correspond to Doppler spread values of 83 and 185 Hz, respectively. The ability to track channel variation depends on how rapidly the channel changes from symbol to symbol. In all the examples given, the symbol rate is much higher than the Doppler spread, so that the channel coefficient value is highly correlated

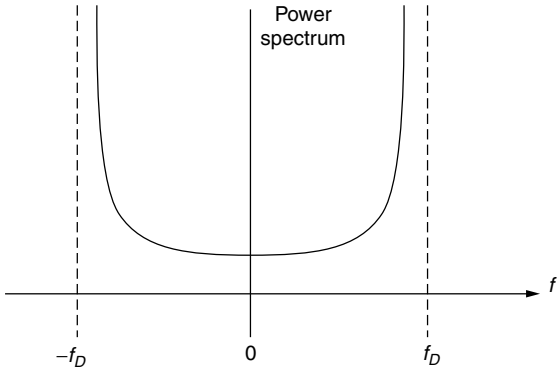


Figure 4. Spectrum of fading process.

from symbol to symbol, making channel tracking possible. Also, the Doppler spread is considered constant, because the speed of the transmitter or receiver changes slowly relative to the transmission rate. However, it is possible to model the time variation of the Doppler spread [11].

At the receiver, $y(t)$ is passed through a filter matched to the pulse shape and sampled, giving received samples

$$r_k = \int f^*(\tau)y(\tau + kT_s)\, d\tau, \quad k = 0, 1, \ldots \tag{4}$$

where $T_s = T/M = T$ is the sampling period. Because the fading varies slowly from symbol to symbol, it can be approximated as constant over the pulse shape. With this approximation, substituting (2) into (4) gives

$$\boxed{r_k = \sum_{j=0}^{J-1} h_k(j)b_{k-j} + z_k} \tag{5}$$

where $h_k(j)$ is the jth composite channel coefficient, reflecting the influence of the transmit filter, the radio channel, and the receive filter

$$h_k(j) = \sum_{\ell=0}^{L-1} c(\ell, kT)R_{ff}(jT - \ell T) \tag{6}$$

where $R_{ff}(\tau)$ is the pulse shape autocorrelation function. Note that z_k corresponds to a sequence of complex Gaussian noise samples, which may be correlated depending on the pulse shape autocorrelation function. Observe that at the receiver, there is intersymbol interference (ISI), as the received samples contain the current symbol b_k as well as interference from previous symbols. The number (J) of composite coefficients needed to accurately model r_k depends on L and on the shape of $R_{ff}(\tau)$. For the special case of root-Nyquist pulse shaping, $J = L$, $h_k(j) = c(j, kT)$, and the noise samples z_k are uncorrelated.

When developing certain channel tracking approaches, it is convenient to formulate the tracker in terms of the conjugate of $h_k(j)$, i.e., $g_k(j) = h_k^*(j)$. This gives the alternative formulation

$$\boxed{r_k = \mathbf{g}_k^H \mathbf{b}_k + z_k} \tag{7}$$

where superscript H denotes Hermitian (conjugate) transpose, $\mathbf{g}_k = [g_k(0)\ g_k(1)\ \cdots\ g_k(J-1)]^T$ is a vector of channel coefficients, $\mathbf{b}_k = [b_k, \ldots, b_{k-J+1}]^T$ is a vector of symbols, and superscript T denotes transpose.

2.2. Simple Narrowband Demodulator Example

Here, a simple coherent receiver for the case of a one-tap channel is given. Using the formulation in (5), we obtain

$$r_k = h_k(0)b_k + z_k \tag{8}$$

Assuming that the information symbol b_k is either $+1$ or -1 (BPSK), we can recover the information using

$$\hat{b}_k = \text{sign}(\text{Re}\{\hat{h}_k^*(0)r_k\}) \tag{9}$$

where $\hat{h}_k(0)$ is an estimate of the channel coefficient and $\text{Re}\{\cdot\}$ denotes the real part of a complex number. When quadrature phase shift keying (QPSK) is used, each symbol represents two bit values. One of the bit values is recovered using (9), and the other bit value is recovered using a similar expression in which the imaginary part is taken instead of the real part.

Multiplying the received value by the conjugate of the channel coefficient estimate removes the phase rotation introduced by the channel and weights the value proportional to how strong the channel coefficient is, which is important when soft information [sign operation omitted in (9)] is used in subsequent forward error correction (FEC) decoding.

2.3. Wideband System

A similar system model can be used for DS-CDMA systems. For these systems, the pulse shape $f(\tau)$ is replaced with the convolution of the chip sequence and a chip pulse shape. Basically, each symbol is represented by a sequence of N_c chips, so that $T = N_c T_c$, where T_c is the chip period. We refer to N_c as the *spreading factor*, which is typically a large integer (e.g., 64, 128) for speech applications. In the DS-CDMA examples given, the chip sequence changes each symbol period, so that the overall symbol pulse shape is time-dependent [$f(\tau)$ is replaced by $f_k(\tau)$]. Also, channel tap delays are typically modeled on the order of the chip period T_c, not the symbol period T. Thus, $\tau(\ell) = \ell T_c/M$. As in the narrowband case, we assume $M = 1$.

As in the narrowband case, the receiver correlates to the symbol pulse shape. For each symbol period k, it produces a "despread" value for each of the L channel taps:

$$r_{k,\ell} = \int f_k^*(\tau)y(\tau + \tau(\ell))\,d\tau \quad \ell = 1, \ldots, L \tag{10}$$

In practice, this involves filtering matched to the chip pulse shape followed by correlation (despreading) using the chip sequence for symbol k. Assuming the spreading factor is large enough, the contribution from adjacent symbols (ISI) can be ignored. Thus, for symbol period k, the despread value for the ℓth channel tap can be modeled as

$$r_{k,\ell} \approx b_k \sum_{\ell=0}^{L-1} c(\ell, kT)R_{f_k f_k}(kT_c - \ell T_c) + z_{k,\ell} \tag{11}$$

Typically $R_{f_k f_k}(iT_c) \approx \delta(i)$ (e.g., when N_c is large), so that

$$r_{k,\ell} \approx c(\ell, kT)b_k + z_{k,\ell} = h_k(\ell)b_k + z_{k,\ell} \tag{12}$$

Comparing (12) to (5), we see that the DS-CDMA case can be treated as L separate, one-tap channels.

The RAKE receiver [9] combines signal energy from each signal image to form a decision variable. For BPSK modulation, this gives the detected bit value

$$\hat{b}_k = \text{sign}\left(\text{Re}\left\{\sum_{\ell=1}^{L} \hat{h}_k^*(\ell, kT)r_{k,\ell}\right\}\right) \tag{13}$$

Observe that the despread values are weighted by the conjugates of the channel coefficient estimates, then added together. Thus, channel estimates are needed to combine the signal images properly.

3. MODELS FOR CHANNEL TRACKING

Channel tracking approaches are often developed from a model of the channel, such as the Rayleigh fading model described in Section 2. However, to obtain reasonable complexity, the model used can be an approximate, simpler model. In this section, such models are described.

There are basically two types of models used to develop channel trackers. The type that is used most often is a *stochastic model*, in which the channel coefficient is modeled as a random process. The Jakes model is an example of a stochastic model. These models are fairly robust, as they allow for random fluctuations in the channel coefficient.

The second type of model is a *deterministic model*, in which the channel coefficient variation in time is represented by a function with parameters. This model is useful when representing the channel variation over a limited period of time, for which the channel variation fits well to a particular functional form. It is particularly useful in predicting future coefficient values when the channel is varying rapidly but the model parameters are varying slowly.

3.1. Stochastic Models

With stochastic models, the channel coefficients are modeled as stochastic random processes. The most commonly used models can be described using the Kalman state-space model [12] given in Fig. 5. This model is fairly general and can well approximate the Jakes model given in Section 2.

With the Kalman model, the $N_s \times 1$ state vector $\mathbf{s}_k$ includes the channel coefficients. The updated state value $\mathbf{s}_{k+1}$ depends on the previous value $\mathbf{s}_k$ through a state transition matrix $\mathbf{F}$ as well as the plant noise $\mathbf{u}_k$ through a gain matrix $\mathbf{G}$. The plant noise is assumed to be a zero-mean, complex white Gaussian process with covariance $\mathbf{I}$ (the identity matrix). This leads to Gaussian channel coefficients, which is consistent with the Rayleigh fading assumption.

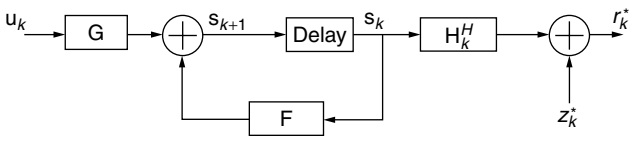


Figure 5. Kalman signal model.

The state is mapped to the output (observation) through a measurement matrix $\mathbf{H}_k$, which includes the symbol values. The output is observed in the presence of zero-mean, complex white Gaussian measurement noise z_k with mean-square value σ_z^2. The observation is defined as the conjugate of the received samples (r_k^*) so that standard expressions can be used. Mathematically, the system is described by the following *process* and *measurement* equations:

$$\mathbf{s}_{k+1} = \mathbf{F}\mathbf{s}_k + \mathbf{G}\mathbf{u}_k \quad \text{(process)} \tag{14}$$

$$r_k^* = \mathbf{H}_k^H \mathbf{s}_k + z_k^* \quad \text{(measurement)} \tag{15}$$

In general, these equations can model an autoregressive moving-average (ARMA) process [13].

One of the simplest channel models is the random-walk model [12]. With this model, the state vector is the conjugate channel coefficient vector ($\mathbf{s}_k = \mathbf{g}_k$), the measurement matrix is the symbol vector ($\mathbf{H}_k = \mathbf{b}_k$), $\mathbf{F} = \mathbf{I}$, and $\mathbf{G} = \sigma_g \mathbf{I}$, so that

$$\mathbf{g}_{k+1} = \mathbf{g}_k + \mathbf{u}_k \tag{16}$$

$$r_k^* = \mathbf{b}_k^H \mathbf{g}_k + z_k^* \tag{17}$$

Note that σ_g is a parameter related to tap strength.

A slightly more general channel model is a first-order autoregressive (AR) process (AR1) [14,15]. This process is similar to the random-walk process, except that $\mathbf{F} = \beta \mathbf{I}$, where β is a parameter between 0 and 1. The choice of β can be related to Doppler spread [14,16]. The *process* equation becomes

$$\mathbf{g}_{k+1} = \beta \mathbf{g}_k + \mathbf{u}_k \tag{18}$$

The random-walk and AR1 models are fairly simple, first-order models that are useful when the channel variation is relatively slow. Note that with these models, as with many other models, the channel coefficients corresponding to different taps are assumed to be uncorrelated (the off-diagonal elements of $\mathbf{F}$ and $\mathbf{G}$ are zero).

When channel variation is relatively fast, a second-order model is used, such as the second-order autoregressive (AR2) channel model [17,18]. Intuitively, such an approach can more accurately model the two spectral peaks that occur at plus and minus the Doppler spread (see Fig. 4). For the simple case of a one-tap channel model ($J = 1$), the model quantities become [17]

$$\mathbf{s}_k = \begin{pmatrix} g_k \\ -a_2 g_{k-1} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} -a_1 & 1 \\ -a_2 & 0 \end{pmatrix},$$
$$\mathbf{G} = \begin{pmatrix} G_1 \\ 0 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} b_k \\ 0 \end{pmatrix} \tag{19}$$

where G_1 is a model parameter related to tap strength. In Refs. 19 and 20, higher-order AR modeling is used, including adaptive estimation of the AR parameters.

Higher-order modeling can also be obtained by tracking different order derivatives of the channel coefficients. For example, the channel tap and its derivative are tracked in the second-order integrated random-walk (IRW) channel model [17,21]. For the simple case of a one-tap channel model ($J = 1$), the model quantities become [17]

$$\mathbf{s}_k = \begin{pmatrix} g_k \\ \dot{g}_k \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 0 \\ G_2 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} b_k \\ 0 \end{pmatrix} \tag{20}$$

where $\dot{g}_k$ denotes the time derivative of g_k and G_2 is a model parameter. This approach can be generalized to include higher-order derivatives [22].

3.2. Deterministic Models

The time evolution of channel coefficients can also be expressed as a deterministic function of time with parameters. Once the function's parameters have been estimated, the channel coefficient as a function of time is determined. The parameters are assumed to vary more slowly than the channel coefficients, which is helpful when the channel varies rapidly.

One model is a polynomial function of time [23,24]. For example, quadratic variation of the channel can be represented as

$$g_k = d_0 + d_1 k + d_2 k^2 \tag{21}$$

The model parameters can be estimated and tracked using a least-squares approach. Model order can be determined by the fading rate, signal-to-noise ratio (SNR), and the duration of the received data. Linear and quadratic functions are most commonly used.

Another deterministic model is the complex sinusoidal model [25]. With this model, a tap is modeled as

$$g_k = \sum_{n=1}^{N_e} A_n e^{j(2\pi f_n k T_s + \phi_n)} \tag{22}$$

where A_n, f_n, and ϕ_n are the amplitude, frequency, and phase of the nth exponential, respectively, and N_e is the number of exponentials (typically <9).

4. FILTERING APPROACHES FOR CHANNEL TRACKING

When there is a sufficiently strong pilot channel or sufficient pilot symbols, the channel can be tracked by filtering channel measurements obtained from the pilot information. The filter smooths the noisy measurements over time and works best when the channel estimate is based on future as well as past channel measurements. Specifically, for a given filter, channel estimation performance depends on the pilot information, fading channel characteristics, and noise level. Pilot information, in terms of how much energy and how often it is available, is a tradeoff between minimizing overhead and optimizing channel estimation performance. For example, with pilot symbols, how often symbols must be sent depends on how rapidly the channel is changing.

In this section, we first examine filtering approaches based on continuous measurements of the channel, which can be obtained from a pilot channel. The simple moving-average filter is examined, as well as the more advanced Wiener filter. Learning the filter weights adaptively,

using adaptive filter techniques, is also discussed. We then examine filtering approaches based on discontinuous measurements of the channel, which can be obtained from pilot symbols. Simple linear interpolation is discussed, as well as Wiener interpolation. To simplify the discussion, the example of a one-tap channel model ($J = 1$) is used throughout this section.

4.1. Estimation Using a Pilot Channel

When a continuous pilot channel is available, a sequence of *channel measurements* can be obtained, as illustrated in Fig. 6. From (8) or (12), measurements of $h_k(0)$ can be obtained by multiplying r_k by b_k^* (assuming $|b_k|^2 = 1$). To estimate the conjugate of $h_k(0)$, which is $g_k(0)$ or simply g_k, we use $r_k^* b_k$ for the channel measurements. These measurements are filtered, giving

$$\hat{g}_k = \sum_{n=N_1}^{N_2} w_{n,k}^* [r^*(k-n)b(k-n)] = \mathbf{w}_k^H \tilde{\mathbf{g}} \tag{23}$$

where $\hat{g}_k$ is the channel estimate at the kth sample position, $\mathbf{w}_k$ represents the vector filter weights as a function of time k, and $\tilde{\mathbf{g}}$ is a vector of measurements. Typically the filter "slides" across the measurements, providing a continuous sequence of channel estimates. The filter weights change slowly in time (k), due to changes in the radio environment.

The parameters N_1 and N_2 are integers ($N_2 \geq N_1$). If only the past samples are used for estimation ($N_1 > 0$), then this operation is called "prediction." Often one-step prediction is used ($N_1 = 1$). If future samples are used ($N_1 < 0$), then this is called "smoothing." In the subsequent subsections, different choices for the filter weights are given.

4.1.1. Moving-Average Filter. The moving-average or sliding-rectangular-window approach is commonly used [26–28]. The channel estimate at time k is obtained by averaging measurements from time $N_1 = k - N$ through $N_2 = k + N$ [i.e., $w_{k,n} = 1/(2N+1)$ in (23)]. The sliding-window approach implies a channel model in which the channel is constant over the averaging window period. The choice of window size is a tradeoff between tracking ability and noise suppression [26]. Thus, it depends on Doppler spread and signal-to-noise ratio (SNR) [28].

4.1.2. Wiener Filtering. With Wiener filtering, the filter weights are designed to minimize the mean-square error (MSE) between the channel estimate and the true channel coefficient [29,30]. The vector of filter weights is given by solving the Wiener–Hopf equation [12]

$$\mathbf{w}_k = \mathbf{R}_{\tilde{g}}^{-1} \mathbf{p} \tag{24}$$

where $\mathbf{R}_{\tilde{g}}$ is the correlation matrix of the measurement vector ($\mathbf{R}_{\tilde{g}} = \mathrm{E}\{\tilde{\mathbf{g}}\tilde{\mathbf{g}}^H\}$) and $\mathbf{p}$ is the correlation vector between the measurement vector and the true channel coefficient ($\mathbf{p} = \mathrm{E}\{\tilde{\mathbf{g}} g_k^*\}$). These quantities typically change slowly in time.

The expressions for $\mathbf{R}_{\tilde{g}}$ and $\mathbf{p}$ depend on the model for the channel coefficient and which channel measurements are used. Consider a simple example, in which the channel is to be estimated at time kT using measurements at times $(k-1)T$, kT, and $(k+1)T$. Also, assume that the channel follows the model given in (3) and has a mean-square value of 1. Then, the Wiener filter quantities are given by

$$\mathbf{R}_{\tilde{g}} = \begin{bmatrix} 1+\sigma_z^2 & J_0(2\pi f_D T) & J_0(2\pi f_D 2T) \\ J_0(2\pi f_D T) & 1+\sigma_z^2 & J_0(2\pi f_D T) \\ J_0(2\pi f_D 2T) & J_0(2\pi f_D T) & 1+\sigma_z^2 \end{bmatrix},$$

$$\mathbf{p} = \begin{bmatrix} J_0(2\pi f_D T) \\ 1 \\ J_0(2\pi f_D T) \end{bmatrix} \tag{25}$$

where σ_z^2 is the noise power. In this example, the Wiener filter design requires knowledge of Doppler spread (f_D) and noise power (σ_z^2). The filter design can be obtained on the basis of the worst-case expected Doppler spread value [31]. Alternatively, the Wiener quantities $\mathbf{R}_{\tilde{g}}$ and $\mathbf{p}$ can be estimated [29].

Approximations to the Wiener filter can be used. A simple approximation is to use a lowpass filter with a cutoff frequency greater than or equal to the maximum expected Doppler frequency [30,32]. In Ref. 33, the channel is approximately modeled as having constant amplitude and linearly varying phase, and other approximations are also made.

4.1.3. Adaptive Filter. Adaptive filtering approaches can be used to "learn" the filter weights [34,35]. The filter output for time k is compared to the channel measurement at time k to generate an error signal, which is used to update the filter weights. Updating approaches are given in Section 5. This approach is often used for prediction.

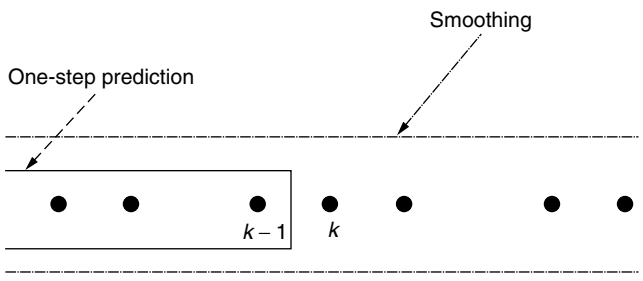


Figure 6. Prediction and smoothing of channel measurements.

4.2. Estimation Using Pilot Symbol Clusters

The channel can be estimated by periodically inserting one or more pilot symbols into the stream of data. Like the pilot channel case, channel measurements can be obtained at the pilot symbol locations. When there is a cluster of pilot symbols, it is often assumed that the channel is approximately constant over the cluster, so that these measurements can be added to give one measurement. In the case of very long pilot clusters, the cluster can be divided into smaller segments [36].

For the case of a multitap channel model, a time-invariant approach [7] can be used to obtain a channel measurement using a cluster of pilot symbols [37]. Another

approach is to run a recursive channel tracker over the pilot cluster [38].

4.2.1. Linear Interpolation. One of the simplest forms of channel estimation using pilot symbols is linear interpolation [4]. With linear interpolation, the channel estimate at a certain time period is a linear combination of the two "nearest" channel measurements. For example, suppose that there are measurements from pilot symbols at times $k = 0$ and $k = M$, denoted $\tilde{g}_0$ and $\tilde{g}_M$, respectively. Then, the channel estimate at time k, $0 < k < M$, is given by

$$\hat{g}_k = w_{0,k}\tilde{g}_0 + w_{M,k}\tilde{g}_M \tag{26}$$

where $w_{0,k}$ and $w_{M,k}$ are given as

$$w_{0,k} = \frac{M-k}{M}, \quad w_{M,k} = \frac{k}{M} \tag{27}$$

Linear interpolation can be viewed as applying a filter with symbol-spaced taps to the channel measurements, which contain zeros at the unknown data symbol points, as illustrated in Fig. 7.

Other simple interpolation filters include lowpass filters [3,38], Gaussian filters [4], and truncated Nyquist interpolation [37], as well as other filter forms [39]. Sometimes there is a tradeoff between using a simpler filter but requiring more closely spaced pilot symbols [4].

4.2.2. Wiener Interpolation. The Wiener filter described previously can also be applied to discontinuous channel measurements [36,40–42]. Continuing with the example given previously, suppose that the channel is to be estimated at time kT using pilot symbol measurements at times $(k-2)T$ and $(k+5)T$. Then, the Wiener filter quantities would be given by

$$\mathbf{R}_{\tilde{g}} = \begin{bmatrix} 1+\sigma_z^2 & J_0(2\pi f_D 7T) \\ J_0(2\pi f_D 7T) & 1+\sigma_z^2 \end{bmatrix}, \quad \mathbf{p} = \begin{bmatrix} J_0(2\pi f_D 2T) \\ J_0(2\pi f_D 5T) \end{bmatrix} \tag{28}$$

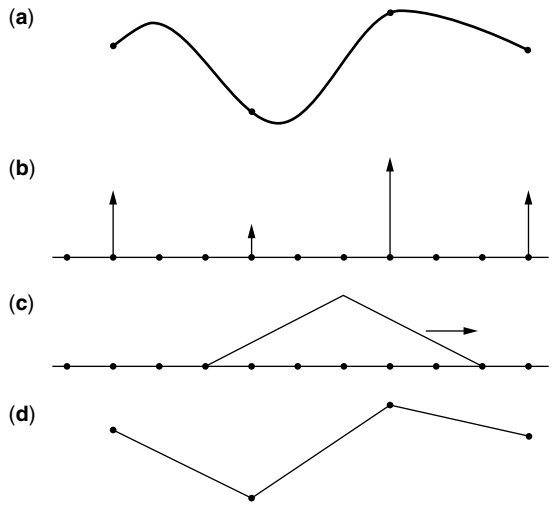


Figure 7. Linear interpolation filtering using pilot symbols: (**a**) fading channel, (**b**) measurements at pilot locations, (**c**) interpolation filter, (**d**) estimated channel.

Observe that the Wiener filter quantities will be different for time $(k+1)T$, requiring the filter weights to change. In general, Wiener interpolation is more complex than the previously described filtering approaches, but its performance is usually better, depending on the accuracy of the Wiener filter quantities.

4.3. Summary

Filtering approaches track the channel by filtering (smoothing) measurements of the channel. Simple filters, with fixed filter coefficients, work well when the channel variation is slow. When the channel variation is rapid, Wiener filtering works better. However, Wiener filtering requires knowledge of the statistics of the fading process and the statistics of the measurement noise process. The fading process statistics can be related to parameters of a channel model, such as Doppler spread and average channel coefficient power. The measurement noise process statistics are usually represented as a noise power or signal-to-noise ratio (SNR). When such information is unavailable, it is possible to "learn" good filter weights using adaptive filter techniques.

5. RECURSIVE APPROACHES FOR CHANNEL TRACKING

With recursive approaches, channel measurements are used one at a time to update channel estimates. For example, recursive approaches are often used in conjunction with data-directed tracking, in which channel estimates are used to detect a data symbol. The detected symbol value is then used to form a channel measurement for updating the channel estimates before detecting the next symbol. Data-directed tracking will be discussed in detail in Sections 6 and 7. Recursive approaches can also be used with pilot channel measurements as an alternative to filtering approaches.

To get started, an initial channel estimate value is needed. The simplest method is to set the initial value to zero, which means that there is a delay before the channel estimate becomes reliable. Another common method is to initially estimate the channel from a training sequence.

In this section, recursive channel tracking approaches are developed assuming a continuous sequence of known symbols (e.g., pilot channel). We start with two simple, related approaches, the least-mean-square (LMS) algorithm and the exponential filter. We then continue with more complex approaches, the recursive-least-squares (RLS) and Kalman filtering approaches. An approximation to the Kalman filter, the Kalman LMS approach, is then discussed.

5.1. Least-Mean-Square (LMS) Algorithm

The least-mean-square (LMS) algorithm [12,43] is one of the simplest approaches to channel tracking. LMS channel tracking is performed according to

$$\hat{\mathbf{g}}_{k+1} = \hat{\mathbf{g}}_k + \mu \mathbf{b}_k e_k^* \tag{29}$$

$$e_k = r_k - \hat{\mathbf{g}}_k^H \mathbf{b}_k \tag{30}$$

In essence, at symbol period k, an error e_k between what is received and what is modeled is formed. This error is used to update the channel coefficient estimate for the next symbol period. The step size μ is a parameter.

For a time-invariant channel, LMS channel tracking can be interpreted as an iterative, stochastic gradient approach for finding the channel coefficients that minimize the mean-square error (MSE) between the *received samples* and the model of the received samples [43]. For two noncomplex (real) channel coefficients, the MSE as a function for the two channel coefficients has a bowl shape [43]. At symbol period k, the LMS algorithm forms a noisy estimate of the slope or gradient at the place on the bowl corresponding to the current channel coefficient estimates. It then updates the taps in the direction of the negative gradient, so as to find the bottom of the bowl. Selection of the step size μ is a tradeoff between rate of convergence and how noisy the model is at convergence (misadjustment noise). For a time-varying channel, the step size μ trades tracking ability for misadjustment noise.

The LMS algorithm is a popular approach for channel tracking [44–46]. It can be derived from the Kalman filter (see Section 5.4), assuming a random-walk model for the channel coefficients [1]. The "leaky" LMS algorithm [12] [obtained by multiplying the first term on the right-hand side of (29) by leakage factor β] can be derived from the Kalman filter assuming an AR1 model [16]. The LMS structure has been further generalized [47], introducing additional parameters whose values are based on first-order or higher-order models of the channel coefficients. Conventional LMS and leaky LMS are special cases of this general, "Wiener LMS" approach.

Different step sizes can be used for different taps depending on the average tap power [16,48], which can also be derived from a Kalman filter formulation [16]. Note that the step size can be adaptive in time [49,50].

5.2. Exponential Filtering (Alpha Tracker)

Exponential filtering, also referred to as an *alpha tracker*, is commonly used with one-tap channels [51,52]. The channel estimate is updated using

$$\hat{g}_{k+1} = \alpha \hat{g}_k + (1-\alpha)\left\{r_k^* b_k\right\} \tag{31}$$

This approach can be derived from the LMS tracker by assuming that the symbols have constant magnitude ($|b_k| = 1$) [51]. It can also be interpreted as a filtering approach employing a first-order infinite-impulse-response (IIR) filter.

5.3. Recursive Least-Squares (RLS)

Recursive least-squares (RLS) [12,43] has also been applied to channel tracking, due to its rapid convergence properties. Conventional RLS channel tracking is performed according to

$$\hat{\mathbf{g}}_{k+1} = \hat{\mathbf{g}}_k + \left(\frac{\mathbf{A}_k}{\lambda + \mathbf{b}_k^H \mathbf{A}_k \mathbf{b}_k}\right) \mathbf{b}_k e_k^* \tag{32}$$

$$\mathbf{A}_{k+1} = \frac{1}{\lambda}\left(\mathbf{A}_k - \frac{\mathbf{A}_k \mathbf{b}_k \mathbf{b}_k^H \mathbf{A}_k}{\lambda + \mathbf{b}_k^H \mathbf{A}_k \mathbf{b}_k}\right) \tag{33}$$

where e_k is given in (30). Compared to LMS, RLS updates a second quantity, the matrix $\mathbf{A}_k$. This quantity is typically initialized to a diagonal matrix with large diagonal entries. The term λ is the "forgetting factor" and, like the LMS step size, determines convergence/tracking rate and misadjustment noise properties. Typically, λ is chosen to be slightly less than 1.

RLS channel tracking can be viewed as finding the set of channel coefficients that minimizes a deterministic weighted squared error between the received samples and the modeled samples [43]:

$$E = |r_k - \mathbf{g}_{k+1}^H \mathbf{b}_k|^2 + \lambda |r_{k-1} - \mathbf{g}_{k+1}^H \mathbf{b}_{k-1}|^2 + \lambda^2 |r_{k-2} - \mathbf{g}_{k+1}^H \mathbf{b}_{k-2}|^2 + \cdots \tag{34}$$

As errors in modeling past received samples are weighted less, the RLS approach tracks the channel by trying to accurately model the most recent data. Because of the exponential weighting, this form of RLS algorithm is also referred to as *exponentially windowed RLS* (EW-RLS) [53,54]. Conventional EW-RLS has been used to track wireless channels [55,56]. It has also been extended to track the channel tap and its derivative [57,58]. Alternatively, a sliding-window RLS (SW-RLS) approach can be used [54], in which the window can be tapered based on statistical knowledge of the fading channel and SNR.

RLS channel tracking can be related to Kalman filtering, based on a first-order AR process with zero plant noise [12,59]. Improved tracking approaches have been developed by considering nonzero plant noise and a time-varying state transition matrix [60].

5.4. Kalman Filtering

Kalman filtering [12,13] provides a recursive form of MMSE filtering. For the Kalman signal model given previously, the corresponding one-step prediction Kalman filter is given by

$$\hat{\mathbf{s}}_{k+1} = \mathbf{F}\hat{\mathbf{s}}_k + \mathbf{K}_k e_k^* \tag{35}$$

$$\mathbf{P}_{k+1} = \mathbf{F}\left(\mathbf{P}_k - \frac{\mathbf{P}_k \mathbf{H}_k \mathbf{H}_k^H \mathbf{P}_k}{\mathbf{H}_k^H \mathbf{P}_k \mathbf{H}_k + \sigma_z^2}\right)\mathbf{F}^H + \mathbf{G}\mathbf{G}^H \tag{36}$$

where

$$\mathbf{K}_k = \frac{\mathbf{F}\mathbf{P}_k \mathbf{H}_k}{\mathbf{H}_k^H \mathbf{P}_k \mathbf{H}_k + \sigma_z^2} \tag{37}$$

and e_k is given by (30). Note that $\mathbf{K}_k$ is a $N_s \times 1$ vector and $\mathbf{P}_k$ is a $N_s \times N_s$ matrix. Also, the Kalman filter requires knowledge of $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}_k$.

Kalman filtering has been applied to channel tracking using a variety of channel models, such as the random-walk model [1,61]. Using the random-walk expressions (16) and (17), Eqs. (35) through (37) simplify to

$$\hat{\mathbf{g}}_{k+1} = \hat{\mathbf{g}}_k + \mathbf{K}_k \mathbf{e}_k^* \tag{38}$$

$$\mathbf{P}_{k+1} = \left(\mathbf{P}_k - \frac{\mathbf{P}_k \mathbf{b}_k \mathbf{b}_k^H \mathbf{P}_k}{\mathbf{b}_k^H \mathbf{P}_k \mathbf{b}_k + \sigma_z^2}\right) + \sigma_g^2 \mathbf{I} \tag{39}$$

$$\mathbf{K}_k = \frac{\mathbf{P}_k \mathbf{b}_k}{\mathbf{b}_k^H \mathbf{P}_k \mathbf{b}_k + \sigma_z^2}. \tag{40}$$

Kalman filtering has been used with the AR1 model [14], higher-order AR models [19], and ARMA models [62]. When model parameters are unknown, extended Kalman filtering can be used to estimate the channel and the unknown parameters [11,63].

To reduce complexity, an approximate form of the Kalman filter can be used, which decouples the tracking of the different channel taps [15]. Another form of approximation is given in Section 5.5, as follows.

5.5. Kalman LMS

In Ref. 17, a series of approximations are applied to the Kalman filter to obtain a lower complexity tracking approach similar to LMS. The key approximation is to average out the effect of the time-varying symbol vector $\mathbf{b}_k$ (part of $\mathbf{H}_k$), which causes the Kalman gain $\mathbf{K}_k$ to vary with time. The resulting "Kalman LMS" (KLMS) approach was applied to two second order models: the IRW model and the AR2 model.

For a single channel tap, the KLMS tracking expressions for these two models are given by

$$\hat{\mathbf{s}}_{k+1} = \mathbf{F}\hat{\mathbf{s}}_k + \boldsymbol{\mu} b_k e_k^* \tag{41}$$

where $\boldsymbol{\mu}$ is a 2×1 vector of step sizes, and $\hat{\mathbf{s}}_k$ and $\mathbf{F}$ are defined earlier for the two models. The two elements in $\boldsymbol{\mu}$ are related by design formulas [17]. The KLMS AR2 form is particularly useful for rapid fading channels [18].

The IRW Kalman LMS form, which tracks the channel coefficient and its derivative, can alternatively be developed from a form of least-squares prediction [22]. This approach was extended to track acceleration (second derivative) as well.

5.6. Summary

When the channel varies slowly, simple approaches such as the LMS algorithm and the exponential filter work well. If rapid convergence at initialization is a concern, the RLS approach can be used. However, these three approaches are based on a first-order model of the channel coefficient. For rapid fading, a higher-order model is needed. In this case, the Kalman filter can be used, as it allows for higher-order models. To reduce complexity, the Kalman LMS approach is useful.

6. DATA-DIRECTED TRACKING, SEPARATE TRACKING

Data-directed tracking is used when there is insufficient pilot information. For systems with only an initial training sequence, the channel may vary significantly over the data portion of the received signal. For other systems, which employ a pilot channel or pilot symbols, data-directed tracking may be beneficial when the power allocated to the pilot information is low, so the channel estimates are very noisy.

In this section, we consider examples in which each channel coefficient is tracked separately. This includes narrowband systems in which there is only one channel tap as well as wideband systems. Joint tracking of multiple coefficients is addressed in Section 7.

6.1. Decision Feedback

With decision feedback, previously detected symbols are used to update the channel estimate before detecting the next symbol [51,52,64]. The detected symbol values fed back can be tentative decisions, with final decisions made using the updated channel estimates [11,22]. Decision feedback has been used with exponential filtering [52], polynomial modeling [65], linear prediction [33,66], and Kalman-based prediction [11].

Before making final symbol decisions, the channel can be estimated again, so that both channel estimation and symbol detection end up being performed twice [66,67]. In the first stage, conventional decision feedback is applied, in which channel estimates are based only on past decisions. In the second stage, channel estimates are based on first-stage decisions of future symbols as well as second-stage decisions of past symbols. The two-stage approach can be extended to multiple stages, further refining the channel estimate and data decisions [67,68]. Depending on the modulation, the first stage may not require channel estimation, as symbols can be detected noncoherently [68].

One problem with decision feedback is that decisions errors can cause the channel estimate to be phase-rotated from the true channel. For example, with BPSK modulation, decision errors correspond to a 180° rotation in the symbol values. To compensate for this, the channel estimates become rotated 180° with respect to the true channel coefficients. This is because the received signal can be equally modeled by an inverted symbol sequence and rotated channel coefficients, sometimes referred to as the *phase ambiguity problem*. Differential modulation can be used to mitigate this problem [51].

6.2. Per-Survivor Processing

Decision errors can also be mitigated by keeping multiple channel estimates. Ideally, a channel estimate should be formed for each possible sequence of symbols [64]. In practice, multiple channel estimates can be formed, corresponding to all possible values for the previous K symbols [29,69–71]. A cost function can be used to decide which channel estimate to keep. This approach can be interpreted as a form of per-survivor processing (PSP), which will be discussed in Section 7.

6.3. Data-Directed Tracking with Pilot Information

When the pilot information is weak, a combination of reference-assisted and data-directed channel tracking can be used. Similar to the case without pilot information, a multistage approach can be used [72,73]. In the first stage, a channel estimate is obtained from the pilot information and used to make tentative symbol decisions. In the second stage, these tentative decisions provide more channel measurements, which are used with the pilot information to produce a refined channel estimate and new symbol values. Note that in the first stage, it is possible to use a mixture of pilot information and decision feedback [34] or a PSP-based approach [29].

A PSP-based approach can be used with pilot information in a single-stage approach as well. Channel measurements from a weak pilot channel can be combined with

PSP-based data-directed channel measurements to form improved channel estimates for demodulation [70,71]. Alternatively, periodic pilot symbols can be used to simply constrain the symbol hypotheses of the PSP process [29,69]. This helps resolve phase ambiguity problem [69] as well as prevent the channel estimator from breaking down because of a high level of decision errors [29].

7. DATA-DIRECTED TRACKING, JOINT TRACKING

Data-directed tracking can also be used when multiple channel coefficients must be estimated together. In this situation, data symbols interfere with one another (ISI) and must be detected together using some form of equalization. Here we consider only equalization approaches that rely on channel estimates, focusing mostly on decision feedback equalization (DFE) and maximum-likelihood sequence estimation (MLSE). There is also the situation in which there may be only one channel coefficient, but the transmitter causes ISI due to the modulation (e.g., partial response pulse shaping).

First, the basic principles for DFE and MLSE are reviewed. Initial channel estimation using a training sequence is then discussed. Various data-directed tracking approaches are presented, showing the tradeoffs between tracking delay, symbol value accuracy, and complexity.

7.1. Equalization

Adaptive equalization has an extensive history [74,75]. In narrowband wireless communication systems, DFE and MLSE are two commonly used forms of equalization. To understand the basic principles of these approaches, consider the simple case of BPSK symbols ($b_k = \pm 1$) and a two-tap channel model, so that

$$r_k = g_k^*(0)b_k + g_k^*(1)b_{k-1} + z_k \tag{42}$$

The DFE, shown in Fig. 8, consists of a feedforward filter, a feedback filter, and a decision device [9]. Conceptually, the feedforward filter tries to collect all signal energy for b_k (which appears in both r_k and r_{k+1}) while suppressing intersymbol interference (ISI) from subsequent symbols (e.g., b_{k+1}). The feedback filter removes ISI from previous symbols (e.g., b_{k-1}). Notice that the feedforward filter introduces delay between when the first image of a symbol arrives and when that symbol is decided.

With MLSE, the likelihood of the received data samples, conditioned on the symbol values, is maximized [9]. The conditional loglikelihood of the kth data sample, assuming that z_k is Gaussian and uncorrelated in time, is related to the following metric or cost function:

$$M(\hat{b}_k, \hat{b}_{k-1}) = |r_k - (g_k^*(0)\hat{b}_k + g_k^*(1)\hat{b}_{k-1})|^2 \tag{43}$$

For different symbol sequence hypotheses ("paths"), this metric is accumulated, generating a path metric. Once all the data samples have been processed, the sequence corresponding to the smallest path metric determines the detected sequence. Intuitively, the metric indicates how well the model of the received data fits the actual received data.

A brute-force search of all possible sequences would require high computational complexity. However, it is possible to determine the smallest metric sequence through a process of path pruning known as the Viterbi algorithm [76]. This involves defining a set of "states" corresponding to a set of paths. Tentative decisions can be made after some delay D. The path "history" (sequence of symbol values) corresponding to the state with the best path metric after processing the kth sample can be used to determine the $(k - D)$th symbol value.

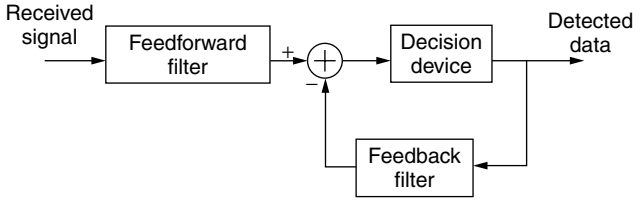


Figure 8. DFE receiver.

7.2. Initialization

A time-invariant approach [7] can be used to initially estimate the channel with a training sequence [77]. Recursive channel tracking can also be used to obtain an initial estimate [78] or to refine an estimate obtained by a time-invariant approach [18].

7.3. Decision Feedback

As in Section 6, decision feedback can be used with MLSE [44]. To obtain reliable tentative symbol decisions, a certain delay (D) is needed. To compensate for this delay, D-step channel prediction can be used, such as linear prediction [78] or a Kalman-based approach [18]. However, channel prediction becomes less reliable with larger D values. Thus, there is still a tradeoff between accuracy of symbol values and tracking delay.

There are several other issues related to decision feedback tracking and MLSE:

1. The "best" path can suddenly change, so that the detected symbol sequence does not correspond to any one path history. Regularization can be used to improve channel tracking in this situation [79].
2. As discussed in Section 6, the channel estimate can be phase rotated [80]. Bidirectional channel tracking can be used to resolve this problem [81].
3. Overmodeling the finite-impulse-response (FIR) channel can cause equalizer timing divergence after fading dips in fast-fading channels [18]. Figure 9 shows an example where a two-tap model is used when only one tap is needed. During recovery from a deep fade, the position of the nonzero tap might change (false lock or time slip). When this happens, the channel is, in fact, being tracked, but with the wrong delay. This problem can be mitigated by using

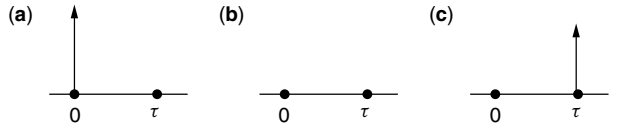


Figure 9. Time-slip problem, channel estimates (**a**) before, (**b**) during, and (**c**) after fade.

different recursive tracking step sizes for different taps [18] or by employing bidirectional tracking [82].

4. Finally, as in Section 6, a multistage approach can be used [29,83].

Decision feedback for channel estimation can also be used with DFE [14,45,48]. Similar to MLSE, delay can be addressed with prediction [14,48]. Decision feedback has also been applied to a form of linear equalization that employs channel estimates [84].

7.4. Per-Survivor Processing

For MLSE, tracking delay can be eliminated by keeping a different channel model for each state [85,86]. In the pruning process, the path that is kept also determines which channel model is updated for the new state. Such an approach is a form of *per-survivor processing* (PSP) [86]. While the channel estimate is usually updated after pruning, it is possible to update the channel estimate before pruning [87].

To reduce complexity, the number of channel models can be less than the number of states [88,89]. Conversely, to improve performance, the size of the state space can be increased [90] to create more channel models.

When the channel is estimated, Viterbi pruning does not necessarily lead to the smallest metric sequence [91]. This leads to keeping channel models for each sequence hypothesis (path) [90]. The M-algorithm [92] can be used to prune paths, keeping the M best paths [93].

8. CONCLUSION

Channel tracking is an important part of receiver design in digital wireless communication systems. Tracking approaches are based on a model of how the channel changes in time. First-order models are often used for slowly varying channels, whereas higher-order models are used for rapidly varying channels.

Reference-assisted channel estimation, using either pilot symbols or a pilot channel, is commonly used to estimate the time-varying channel impulse response. When reference information is unavailable or insufficient, data-directed tracking can be used. Obtaining reliable symbol decisions and reducing tracking delay are the main challenges. The tracking delay can be minimized by keeping multiple channel models, corresponding to different hypotheses of the data symbol values.

Currently, third-generation digital wireless communication systems are being deployed. As modulation formats for these systems require coherent reception, channel estimation will continue to be a key element in receiver design.

Acknowledgment

The authors would like to thank A. Khayrallah, K. Molnar, J. G. Proakis, and Y.-P. E. Wang for reviewing a draft of this article. The authors gratefully acknowledge the help of others in identifying references. Finally, the authors wish to thank their colleagues for many helpful discussions on channel estimation.

BIOGRAPHIES

Gregory E. Bottomley received his B.S. and his M.S. degrees from Virginia Polytechnic Institute and State University, Blacksburg, Virginia, in 1983 and 1985, respectively, and his Ph.D. degree from North Carolina State University, Raleigh, in 1989, all in electrical engineering.

From 1985 to 1987 he was with AT&T Bell Laboratories, Whippany, NJ, working in the area of sonar signal processing. In 1990, he was a visiting lecturer at North Carolina State University, Raleigh. Since 1991, he has been with Ericsson Inc., Research Triangle Park, NC, where he is currently a member of the Advanced Development and Research Department. He is listed as an inventor on over 30 patents in wireless communications and was a recipient of Ericsson's Inventor of the Year Award in 1997.

Dr. Bottomley is an associate member of Sigma Xi and a senior member of The Institute of Electrical and Electronics Engineers, Inc. (IEEE). In 1998, he was a recipient of the IEEE Eastern North Carolina Section Outstanding Engineer Award. He served as associate editor (1997–2000) and currently serves as editor for the *IEEE Transactions on Vehicular Technology*. His research interests are in baseband signal processing for wireless communications, including equalization, RAKE reception, and interference suppression.

Huseyin Arslan (eushura@rtp.ericsson.se) was born in Nazilli, Turkey, in 1968. He received a B.S. degree from Middle East Technical University, Ankara, Turkey, and M.S. and Ph.D. degrees from Southern Methodist University, Dallas, Texas, in 1992, 1994, and 1998, respectively, all in electrical engineering. Since January 1998, he has been at Ericsson research at RTP, North Carolina. His research interests are in baseband signal processing for mobile communications, including interference cancellation, channel estimation, modulation, demodulation, and equalization.

BIBLIOGRAPHY

1. L. Ljung and S. Gunnarsson, Adaptation and tracking in system identification: A survey, *Automatica* **26**: 7–21 (1990).
2. A. P. Clark, *Adaptive Detectors for Digital Modems*, Pentech, London, 1989.
3. M. L. Moher and J. H. Lodge, TCMP—a modulation and coding strategy for Rician fading channels, *IEEE J. Select. Areas Commun.* **7**: 1347–1355 (Dec. 1989).
4. S. Sampei and T. Sunaga, Rayleigh fading compensation for QAM in land mobile radio communications, *IEEE Trans. Vehic. Technol.* **42**: (May 1993).

5. K. S. Gilhousen et al., On the capacity of a cellular CDMA system, *IEEE Trans. Vehic. Technol.* **40**: 303–312 (May 1991).
6. H. W. Li and J. K. Cavers, An adaptive filtering technique for pilot-aided transmission systems, *IEEE Trans. Vehic. Technol.* **40**: 532–545 (Aug. 1991).
7. H. Arslan and G. E. Bottomley, Channel estimation in narrowband wireless communication systems, *Wireless Commun. Mobile Comput. J.* **1**: 201–219 (April/June 2001).
8. H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Radar-Sonar Signal Processing and Gaussian Signals in Noise*, Krieger, Malabar, FL, 1992.
9. J. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
10. W. C. Jakes, ed., *Microwave Mobile Communications*, IEEE Press, Piscataway, NJ, 1993.
11. A. Aghamohammadi, H. Meyr, and G. Ascheid, Adaptive synchronization and channel parameter estimation using an extended Kalman filter, *IEEE Trans. Commun.* **37**: 1212–1218 (Nov. 1989).
12. S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.
13. B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.
14. M. Stojanovic, J. G. Proakis, and J. A. Catipovic, Analysis of the impact of channel estimation errors on the performance of a decision-feedback equalizer in fading multipath channels, *IEEE Trans. Commun.* **43**: 877–885 (Feb.–April 1995).
15. M. E. Rollins and S. J. Simmons, Simplified per-survivor Kalman processing in fast frequency-selective fading channels, *IEEE Trans. Commun.* **45**: 544–552 (May 1997).
16. W. Liu, Performance of joint data and channel estimation using tap variable step size LMS for multipath fast fading channel, *Proc. IEEE Globecom Conf.*, San Francisco, CA, 1994, pp. 973–978.
17. L. Lindbom, Simplified Kalman estimation of fading mobile radio channels: High performance at LMS computational load, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Minneapolis, MN, 1993, pp. 352–355.
18. K. Jamal, G. Brismark, and B. Gudmundson, Adaptive MLSE performance on the D-AMPS 1900 channel, *IEEE Trans. Vehic. Technol.* **46**: 634–641 (Aug. 1997).
19. H. Zamiri-Jafarian and S. Pasupathy, Adaptive MLSD receiver with identification of flat fading channels, *Proc. IEEE Vehicular Technology Conf.*, Phoenix, AZ, May 4–7, 1997, pp. 695–699.
20. L. M. Davis, I. B. Collings, and R. J. Evans, Coupled estimators for equalization of fast fading mobile channels, *IEEE Trans. Commun.* **46**: 1262–1265 (Oct. 1998).
21. S. Gazor, Prediction in LMS-type adaptive algorithms for smoothly time varying environments, *IEEE Trans. Signal Process.* **47**: 1735–1739 (June 1999).
22. A. P. Clark, Adaptive channel estimator for an HF radio link, *IEEE Trans. Commun.* **37**: 918–926 (Sept. 1989).
23. W. D. Rummler, R. P. Coutts, and M. Liniger, Multipath fading channel models for microwave digital radio, *IEEE Commun. Mag.* **24**: 30–42 (Nov. 1986).
24. D. K. Borah and B. D. Hart, Frequency-selective fading channel estimation with a polynomial time-varying channel model, *IEEE Trans. Commun.* **47**: 862–873 (June 1999).
25. A. Duel-Hallen, S. Hu, and H. Hallen, Long-range prediction of fading signals, *IEEE Signal Process. Mag.* **17**: 62–75 (May 2000).
26. V.-P. Kaasila and A. Mämmelä, The adaptive rake matched filter in a time-variant two-path channel, *Proc. IEEE Int. Symposium on Personal Indoor and Mobile Radio Communications* (*PIMRC*), Boston, MA, Oct. 19–21, 1992, pp. 441–445.
27. U. Fawer, A coherent spread-spectrum diversity-receiver with AFC for multipath fading channels, *IEEE Trans. Commun.* **42**: 1300–1311 (Feb.–April 1994).
28. M. Benthin and K.-D. Kammeyer, Influence of channel estimation on the performance of a coherent DS-CDMA system, *IEEE Trans. Vehic. Technol.* **46**: 262–268 (May 1997).
29. A. N. D'Andrea, A. Diglio, and U. Mengali, Symbol-aided channel estimation with non-selective Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **44**: 41–49 (Feb. 1995).
30. F. Ling, Optimal reception, performance bound, and cut-off rate analysis of reference-assisted coherent CDMA communications with applications, *IEEE Trans. Commun.* **47**: 1583–1592 (Oct. 1999).
31. P. Schramm, Differentially coherent demodulation for differential BPSK in spread spectrum systems, *IEEE Trans. Vehic. Technol.* **48**: 1650–1656 (Sept. 1999).
32. G. Chen, X.-H. Yu, and J. Wang, Adaptive channel estimation and dedicated pilot power adjustment based on the fading-rate measurement for a pilot-aided CDMA systems, *IEEE J. Select. Areas Commun.* **19**: 132–139 (Jan. 2001).
33. L. Bin and P. Ho, Data-aided linear prediction receiver for coherent DPSK and CPM transmitted over Rayleigh flat-fading channels, *IEEE Trans. Vehic. Technol.* **48**: 1229–1236 (July 1999).
34. Y. Liu and S. D. Blostein, Identification of frequency non-selective fading channels using decision feedback and adaptive linear prediction, *IEEE Trans. Commun.* **43**: 1484–1492 (Feb.–April 1995).
35. R. J. Young and J. H. Lodge, Detection of CPM signals in fast Rayleigh flat-fading using adaptive channel estimation, *IEEE Trans. Vehic. Technol.* **44**: 338–347 (May 1995).
36. S. A. Fechtel and H. Meyr, Optimal parametric feedforward estimation of frequency-selective fading radio channels, *IEEE Trans. Commun.* **42**: 1639–1650 (Feb.–April 1994).
37. N. W. K. Lo, D. D. Falconer, and A. U. H. Sheikh, Adaptive equalization and diversity combining for mobile radio using interpolated channel estimates, *IEEE Trans. Vehic. Technol.* **40**: 636–645 (Aug. 1991).
38. A. Aghamohammadi, H. Meyr, and G. Ascheid, A new method for phase synchronization and automatic gain control of linearly modulated signals on frequency-flat fading channels, *IEEE Trans. Commun.* **39**: 25–29 (Jan. 1991).
39. H. Andoh, M. Sawahashi, and F. Adachi, Channel estimation using time multiplexed pilot symbols for coherent rake combining for DS-CDMA mobile radio, *Proc. IEEE Int. Symposium on Personal Indoor and Mobile Radio Communications* (*PIMRC*), Helsinki, Finland, Sept. 1–4 1997, pp. 954–958.
40. J. K. Cavers, An analysis of pilot symbol assisted modulation for Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **40**: 686–693 (Nov. 1991).

41. W.-Y. Kuo and M. P. Fitz, Designs for pilot-symbol-assisted burst-mode communications with fading and frequency uncertainty, *Int. J. Wireless Inform. Networks* **1**: 239–252 (1994).
42. C. D'Amours, M. Moher, and A. Yongaçoğlu, Comparison of pilot symbol-assisted and differentially detected BPSK for DS-CDMA systems employing RAKE receivers in Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **47**: 1258–1267 (Nov. 1998).
43. S. T. Alexander, *Adaptive Signal Processing*, Springer-Verlang, New York, 1986.
44. F. R. Magee and J. G. Proakis, Adaptive maximum-likelihood sequence estimation for digital signaling in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 120–124 (Jan. 1973).
45. P. K. Shukla and L. F. Turner, Channel-estimation-based adaptive DFE for fading multipath radio channels, *IEE Proc.-I Communications, Speech and Vision* **138**: 525–543 (1991).
46. M.-C. Chiu and C.-C. Chao, Analysis of LMS-adaptive MLSE equalization on multipath fading channels, *IEEE Trans. Commun.* **44**: 1684–1692 (Dec. 1996).
47. L. Lindbom, M. Sternad, and A. Ahlén, Tracking of time-varying mobile radio channels—Part I: The Wiener LMS algorithm, *IEEE Trans. Commun.* **49**: 2207–2217 (Dec. 2001).
48. S. A. Fechtel and H. Meyr, An investigation of channel estimation and equalization techniques for moderately rapid HF-channels, *Proc. IEEE Int. Conf. Communications*, Denver, CO, June 23–26 1991, pp. 768–772.
49. H. Shiino, N. Yamaguchi, and Y. Shoji, Performance of an adaptive maximum-likelihood receiver for fast fading multipath channel, *Proc. IEEE Vehicular Technology Conf.*, Denver, CO, May 1992, pp. 380–383.
50. S. Denno and Y. Saito, Orthogonal-transformed variable-gain least mean squares (OVLMS) algorithm for fractional tap-spaced adaptive MLSE equalizers, *IEEE Trans. Commun.* **47**: 1151–1160 (Aug. 1999).
51. K. Pahlavan and J. W. Matthews, Performance of adaptive matched filter receivers over fading multipath channels, *IEEE Trans. Commun.* **38**: 2106–2113 (Dec. 1990).
52. G. J. R. Povey, P. M. Grant, and R. D. Pringle, A decision-directed spread-spectrum rake receiver for fast-fading mobile channels, *IEEE Trans. Vehic. Technol.* **45**: 491–502 (Aug. 1996).
53. E. Eleftheriou and D. D. Falconer, Tracking properties and steady state performance of RLS adaptive filter algorithms, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**: 1097–1110 (Oct. 1986).
54. J. Lin, J. G. Proakis, F. Ling, and H. Lev-Ari, Optimal tracking of time-varying channels: A frequency domain approach for known and new algorithms, *IEEE J. Select. Areas Commun.* **13**: 142–154 (Jan. 1995).
55. P. Newson and B. Mulgrew, Adaptive channel identification and equalization for GSM European digital mobile radio, *Proc. IEEE Int. Conf. Communications*, Denver, CO, June 23–26 1991, pp. 23–27.
56. H.-N. Lee and G. J. Pottie, Fast adaptive equalization/diversity combining for time-varying dispersive channels, *IEEE Trans. Commun.* **46**: 1146–1162 (Sept. 1998).
57. A P. Clark and S. Hariharan, Efficient estimators for an HF radio link, *IEEE Trans. Commun.* **38**: 1173–1180 (Aug. 1990).
58. N. Zhou and N. Holte, Least squares channel estimation for a channel with fast time variations, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1992, pp. 165–168.
59. A. H. Sayed and T. Kailath, A state-space approach to adaptive RLS filtering, *IEEE Signal Process. Mag.* **11**: 18–60 (July 1994).
60. S. Haykin et al., Adaptive tracking of linear time-variant systems by extended RLS algorithms, *IEEE Trans. Signal Process.* **45**: 1118–1128 (May 1997).
61. G. E. Bottomley and K. J. Molnar, Adaptive channel estimation for multichannel MLSE receivers, *IEEE Commun. Lett.* **3**: 40–42 (Feb. 1999).
62. Q. Dai and E. Shwedyk, Detection of bandlimited signals over frequency selective Rayleigh fading channel, *IEEE Trans. Commun.* **42**: 941–950 (Feb.–April 1994).
63. R. A. Iltis and A. W. Fuxjaeger, A digital DS spread-spectrum receiver with joint channel and Doppler shift estimation, *IEEE Trans. Commun.* **39**: 1255–1267 (Aug. 1991).
64. R. Haeb and H. Meyr, A systematic approach to carrier recovery and detection of digitally phase modulated signals on fading channels, *IEEE Trans. Commun.* **37**: 748–754 (July 1989).
65. Y. Sanada, A. Kajiwara, and M. Nakagawa, Adaptive rake system for mobile communications, *Proc. IEEE Int. Conf. Selected Topics in Wireless Communication* (*ICWC'92*), Vancouver, BC, Canada, 1992, pp. 227–230.
66. G. Auer, G. J. R. Povey, and D. I. Laurenson, Mobile channel estimation for decision directed RAKE receivers operating in fast fading radio channels, *Proc. IEEE Int. Symp. Spread Spectrum Techniques and Applications* (*ISSSTA*), Sun City, South Africa, Sept. 2–4, 1998, pp. 576–579.
67. P. Y. Kam, Optimal detection of digital data over the non-selective Rayleigh fading channel with diversity reception, *IEEE Trans. Commun.* **39**: 214–219 (Feb. 1991).
68. B. H. Park, K. J. Kim, S.-Y. Kwon, and K. C. Whang, Multi-stage decision-directed channel estimation scheme for DS-CDMA system with M-ary orthogonal signaling, *IEEE Trans. Vehic. Technol.* **49**: 43–49 (Jan. 2000).
69. G. M. Vitetta and D. P. Taylor, Maximum likelihood decoding of uncoded and coded PSK signal sequences transmitted over Rayleigh flat-fading channels, *IEEE Trans. Commun.* **43**: 2750–2758 (Nov. 1995).
70. J. Choi, Multipath CDMA channel estimation by jointly utilising pilot and traffic channels, *IEE Proc. Commun.* **146**: 312–318 (Oct. 1999).
71. S.-C. Hong, J.-S. Joo, and Y. H. Lee, Per-survivor processing sequence detection for DS/CDMA systems with pilot and traffic channels, *IEEE Commun. Lett.* **5**: 346–348 (Aug. 2001).
72. G. T. Irvine and P. J. McLane, Symbol-aided plus decision-directed reception for PSK/TCM modulation on shadowed mobile satellite fading channels, *IEEE J. Select. Areas Commun.* **10**: 1289–1299 (Oct. 1992).
73. S. Min and K. B. Lee, Pilot and traffic based channel estimation for DS/CDMA systems, *IEE Electron. Lett.* **34**: 1070–1071 (May 1998).
74. R. W. Lucky, A survey of the communication theory literature: 1968–1973, *IEEE Trans. Inform. Theory* **19**: 725–739 (Nov. 1973).
75. J. G. Proakis, Adaptive equalization for TDMA digital mobile radio, *IEEE Trans. Vehic. Technol.* **40**: 333–341 (May 1991).

76. G. D. Forney, The Viterbi algorithm, *Proc. IEEE* **61**: 268–277 (March 1973).

77. R. D'Avella, L. Moreno, and M. Sant'Agostino, An adaptive MLSE receiver for TDMA digital mobile radio, *IEEE J. Select. Areas Commun.* **7**: 122–129 (Jan. 1989).

78. E. Dahlman, New adaptive Viterbi detector for fast-fading mobile-radio channels, *IEE Electron. Lett.* **26**: (Sept. 1990).

79. M. Martone, Optimally regularized channel tracking techniques for sequence estimation based on cross-validated subspace signal processing, *IEEE Trans. Commun.* **48**: 95–105 (Jan. 2000).

80. P. K. Shukla and L. F. Turner, Examination of an adaptive DFE and MLSE/near-MLSE for fading multipath radio channels, *IEE Proc.-I Communications, Speech and Vision* **139**: 418–428 (Aug. 1992).

81. H. Arslan, R. Ramesh, and A. Mostafa, Interpolation and channel tracking based receivers for coherent Mary-PSK modulations, *Proc. IEEE Vehicular Technology Conf.*, Houston, Tex, May 1999, pp. 2194–2199.

82. Y.-J. Liu, M. Wallace, and J. W. Ketchum, A soft-output bidirectional decision feedback equalization technique for TDMA cellular radio, *IEEE J. Select. Areas Commun.* **11**: 1034–1045 (Sept. 1993).

83. D. K. Borah and B. D. Hart, Receiver structures for time-varying frequency-selective fading channels, *IEEE J. Select. Areas Commun.* **17**: 1863–1875 (Nov. 1999).

84. P. Butler and A. Cantoni, Noniterative automatic equalization, *IEEE Trans. Commun.* **COM-23**: 621–633 (June 1975).

85. H. Kubo, K. Murakami, and T. Fujino, An adaptive maximum-likelihood sequence estimator for fast time-varying intersymbol interference channels, *IEEE Trans. Commun.* **42**: 1872–1880 (Feb.–April 1994).

86. R. Raheli, A. Polydoros, and C.-K. Tzou, Per-survivor processing: A general approach to MLSE in uncertain environments, *IEEE Trans. Commun.* **43**: 354–364 (Feb.–April 1995).

87. H. Zamiri-Jafarian and S. Pasupathy, Adaptive MLSDE using the EM algorithm, *IEEE Trans. Commun.* **47**: 1181–1193 (Aug. 1999).

88. R. Raheli, G. Marino, and P. Castoldi, Per-survivor processing and tentative decisions: What is in between?, *IEEE Trans. Commun.* **44**: 127–129 (Feb. 1996).

89. M. J. Bradley and P. Mars, Application of multiple channel estimators in MLSE detectors for fast time-varying and frequency selective channels, *IEE Electron. Lett.* **32**: 620–621 (March 1996).

90. N. Seshadri, Joint data and channel estimation using blind trellis search techniques, *IEEE Trans. Commun.* **42**: 1000–1011 (Feb.–April 1994).

91. K. M. Chugg, The condition for the applicability of the Viterbi algorithm with implications for fading channel MLSD, *IEEE Trans. Commun.* **46**: 1112–1116 (Sept. 1998).

92. J. B. Anderson and S. Mohan, Sequential coding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**: 169–176 (Feb. 1984).

93. P. Castoldi, R. Raheli, and G. Marino, Efficient trellis search algorithms for adaptive MLSE on fast Rayleigh fading channels, *Proc. IEEE Globecom Conf.*, San Francisco, CA, Nov. 1994, pp. 196–200.

CHAOS IN COMMUNICATIONS

KUNG YAO
CHI-CHUNG CHEN
University of California
Los Angeles, California

1. INTRODUCTION

In 1963, Lorenz used a digital computer to study the numerical solutions of nonlinear dynamical systems modeling convection in the atmosphere. He found that even a very small difference in the initial conditions can lead to solutions that can grow rapidly apart with time [1]. The deterministic but unpredictable behaviors of certain classes of nonlinear dynamical systems have been called *chaos*. Later, Lorenz presented a talk with the title "Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" Thus, the essence of the sensitivity of initial conditions to the long-term solutions of these chaotic systems has been colorfully called the "butterfly effect." Since that time, chaos has become a well-developed branch of mathematics, with applications to physics, biology, medicine, engineering, economics, and other fields. Chaos was also shown to be related to the work of fractal by Mandelbrot [2]. Researchers in nonlinear circuits and systems have found large numbers of fairly simple nonlinearities that can induce quite complicated chaotic solutions [3]. Since the early 1960s, thousands of papers and hundreds of books have been written on various aspects of chaos. Since then, a small-scale intellectual industry has been formed in exploiting chaos. Gleick's book, *Chaos: The Amazing Science of the Unpredictable,* which introduced chaos to the general public, is extremely readable, and was a bestseller when it was published [4]. Quoting Gleick's book (pp. 5–6), "The most passionate advocates of this new science go so far as to say that twentieth-century science will be remembered for just three things: relativity, quantum mechanics, and chaos." Chaos, they contend, has become the century's third great revolution in the physical sciences. Like the first two revolutions, chaos cuts away at the tenets of Newton's physics. As one physicist put it: "Relativity eliminated the Newtonian illusion of absolute space and time; quantum mechanics eliminated the Newtonian dread of a controllable measurement process; and chaos eliminates the Laplacian fantasy of "deterministic predictability." Of the three, the revolution in chaos applies to the universe as we see and touch, to objects at human scale. It remains to be seen whether the long-range impacts of chaos both to theory and applications are as profound as some advocates have claimed. In any case, for communication engineers not familiar with chaos, some useful books include: a very readable account of basic chaotic concepts by Williams [5], a medium-level graduate text on chaos by Hilborn [6], and an advanced treatise on stochastic aspects of dynamics and chaos by Lasota and Mackey [7]. Various introductory surveys, overviews, and special issue papers on chaotic communications have appeared elsewhere in the literature [3,8–14].

In the early 1990s, Pecora and Carroll [15] showed that despite the broadband nature of the spectrum, noiselike

behavior, and sensitivity to initial conditions inherented in chaotic systems, two chaotic systems can be synchronized. This rather remarkable discovery has caused much interest and created the research field of "chaotic communication." Many papers motivated by this work have been published since the early 1990s. The drive–response synchronization configuration that Pecora and Carroll proposed is shown in Fig. 1. We will use the Lorenz system example unless mentioned otherwise. If a chaotic system can be decomposed into two subsystems, a drive system x and a conditionally stable response system $\{y, z\}$ as shown in this example, then the identical chaotic system at the receiver can be synchronized when driven with a common signal. The reason for this is simple. In the absence of noise, the output signals y_r and z_r will follow the signals y and z since it is a stable subsystem. For more discussion on chaotic synchronization, see the paper by Pecora et al. [16].

On the basis of the self-synchronization properties of chaotic systems, various chaotic communication systems using chaos as carrier have been proposed. We summarize many of these systems in Section 2. Specifically, in Section 2.1, we deal with the chaotic masking modulation/demodulation scheme of Cuomo et al. [17], which was one of the earliest proposed methods to perform chaotic communication. In order to circumvent an inadequacy of chaotic masking modulation/demodulation due to the small amplitude of the information signal, the dynamical feedback modulation (DFM) scheme of Milanovic and Zaghloul [18] was proposed and treated in Section 2.2. In Section 2.3, we consider the chaotic switching modulation (CSM) scheme of Kocarev et al. [19]. In all three of these schemes, since the communication system performances strongly depend on the synchronization capability of the chaotic systems, the robustness ability of self-synchronization to the white noise needs to be explored [17,20]. However, it is generally difficult to obtain the analytic solutions to the nonlinear stochastic systems, in particular, the chaotic systems. The use of Monte Carlo (MC) simulation is a standard method for the evaluation of complicated communication systems over noisy channels. However, because of the nonlinear stochastic dynamical system modeling of these chaotic communication systems, standard deterministic MC simulation method turns out to be inadequate. In Section 2.4, we show the need to use Ito–Stratonovich stochastic integrations for the performance evaluation of continuous-time chaotic communication systems. Specifically, we show that the use of conventional Runge–Kutta numerical integration method can yield incorrect results. Furthermore, the precise evaluation of the performance of many of these chaotic communication systems showed the required SNRs (signal-to-noise ratios) to achieve nominal acceptable error probabilities are significantly higher than those conventional (nonchaotic) communication systems.

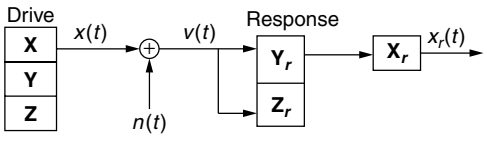


Figure 1. Drive–response self-synchronization scheme.

In order to mitigate the poor system performance of the chaotic modulation schemes due to the ill effect of the self-synchronizing aspects of these schemes, Dedieu et al. [21] first proposed the chaos shift-keyed modulation(CSK), then Kennedy and Kolumban [22–24] proposed the differential shift-keyed modulation (DCSK), and later the frequency modulation DCSK (FM-DCSK) schemes [25,26]. In Section 2.5, we discuss these different forms of CSK schemes and show in particular the FM-DCSK scheme to be competitive to conventional digital modulation schemes in performance. This is to be contrasted to earlier chaotic modulation schemes, which are intellectually interesting but not practical for implementation on a real communication system.

The symbolic dynamic model is one tool that can be used to analyze a complex chaotic dynamical system. Based on the symbolic dynamics of a chaotic system and using chaos control technique, Hayes et al. [27,28] proposed that the information message can be embedded into the chaotic dynamics for transmission to the receiver. This approach is considered in Section 3. Since it is usually difficult to estimate a chaotic signal, Papadopoulos and Wornell [29] developed a maximum-likelihood estimator for the tent map. By setting the information data into the initial condition, the information data can be encoded using a chaotic dynamical system, and can be retrieved by estimating the initial condition by Chen and Wornell [30]. These issues are considered in Section 4.

Instead of transmitting the chaotic waveform, a chaotic pulse position modulation (CPPM) scheme was proposed [31]. This proposed system is similar to an ultra-wide-bandwidth impulse radio [32] that offers a very promising communication platform, especially in a severe multipath environment. A pulse position method is used to modulate the binary information onto the carrier. The separation between the adjacent pulses is chaotic because of the dynamical system with irregular behavior. These methods are discussed in Section 5.

Since the 1960s, first for military, then later for commercial applications, code-division multiple access (CDMA) communication techniques have been used extensively. Early special issues on CDMA appeared [33,34], but extensive paper publications and books appeared later. More recently, the use of CDMA for cellular telephone communication has provided explosive worldwide interests in CDMA systems. System performance of a direct-sequence CDMA (DSCDMA) system critically depends on the auto- and cross-correlation properties of the *spreading sequences*. The chaotic signals have low non-zero-shift autocorrelation and all cross-correlation properties due to the intrinsic broad spectrum and sensitivity to initial conditions. The sequences generated by a logistic map can also be used for a DSCDMA communication system as first proposed [49]. The use of the correlation function characteristics of some specific chaotic sequences and its comparison to m sequences/Gold sequences have been discussed [50,51,53,54,56,57,59–61]. Such optimum sequences derived and generated on chaos-based concepts can support about 15% more users than previously known sequences generated by deterministic methods. These issues are discussed in Section 6. Finally, two

distinct applications of chaos-related stochastic processes to optical communication based on chaos in semiconductor lasers [65] and modeling of radar and radio propagation effects [66,74] are considered in Section 7.

2. COMMUNICATIONS USING CHAOS AS CARRIER

The use of modulating an aperiodic chaotic waveform, instead of a periodic sinusoidal signal, for carrying information messages has been proposed, in particular, chaotic masking, dynamical feedback, chaotic switching, chaos shift keying, and inverse approach modulations [10,17–19,21,22,35,36]. In this section, we summarize several of these chaotic modulation schemes and the impact of self-synchronization in the demodulation process.

2.1. Chaotic Masking Modulation

The basic idea of a chaotic masking modulation (CMM) scheme is based on chaotic signal masking and recovery. As shown in Fig. 2, we add a noiselike chaotic signal to the information signal at the transmitter, and at the receiver the masking signal is removed. The received signal, consisting of masking and information signals, is used to regenerate the masking signal at the receiver. The masking signal is then subtracted from the received signal to retrieve the information signal. The regeneration of the masking signal can be done by synchronizing the receiver chaotic system with the transmitter chaotic system. This communication system could be used for analog and digital information data. Cuomo et al. [17] have built a Lorenz system circuit and demonstrated the performance of a CMM system with a segment of speech from the sentence "He has the bluest eyes." The communication system performance truly relies on the synchronization ability of chaotic system. The masking properties of this scheme work only when the amplitudes of the information signals are much smaller than the masking chaotic signals.

2.2. Dynamical Feedback Modulation

To avoid the small-amplitude restriction of the information signal, another modulation scheme, called *dynamical feedback modulation* (DFM), has been proposed by Milanovic and Zaghloul [18] and is shown in Fig. 3. The basic idea is to feed back the information signal into the chaotic transmitter in order to ensure the identical input signal for the chaotic transmitter and the receiver. Specifically, the transmitted signal $v(t) = x(t) + m(t)$, consisting of the information signal $m(t)$ and the chaotic signal $x(t)$, is communicated to the receiver which is identical to the chaotic

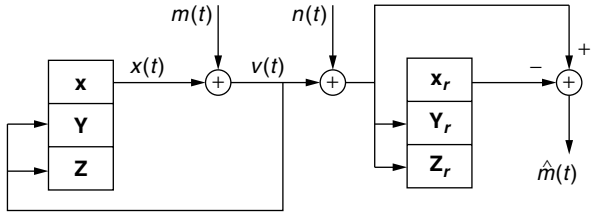


Figure 3. Dynamical feedback modulation.

transmitter. Since the reconstructed signal $x_r(t)$ will be identical to $x(t)$ in the absence of noise $n(t)$, the information signal $m(t)$ can be decoded from the received signal by using $\hat{m}(t) = x(t) + m(t) - x_r(t)$.

This *analog* communication technique can be also applied to binary data communication by setting $m(t) = A$ if binary information data are one, and $m(t) = -A$ if binary data are zero. The sufficient statistic η_d of detection is the average of the error signal at the receiver after discarding a transient period before synchronization, and is given by

$$\eta_d = \frac{1}{T}\int_{t_0}^{t_0+T} e_d(t)\,dt \tag{1}$$

where the error signal $e_d(t)$ is defined by

$$e_d(t) = v(t) + n(t) - x_r(t) \tag{2}$$

Since the feedback information will affect the chaotic property, the information level A should be scaled carefully to make the transmitter still chaotic to maintain the desired communication security.

2.3. Chaotic Switching Modulation

In contrast to DFM, the chaotic switching modulation (CSM) communication system, as illustrated in Fig. 4, does not suffer the above scaling problem as discussed by Kocarev et al. [19]. The basic idea is to encode the binary data $m(t)$ with different chaotic attractors by modulating the transmitter parameters and then transmitting the chaotic drive signal $x_m(t)$. At the receiver, the parameter modulation will produce a synchronization error between the received drive signal and the regenerated drive signal with an error amplitude that depends on the modulation. Using the synchronization error, the binary data can be detected.

In Fig. 4, the parameter b of the transmitter is modulated by the binary data $m(t)$. Assume that the communication link is an AWGN channel; the synchronization

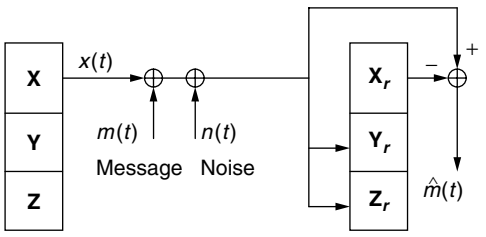


Figure 2. Chaotic masking modulation.

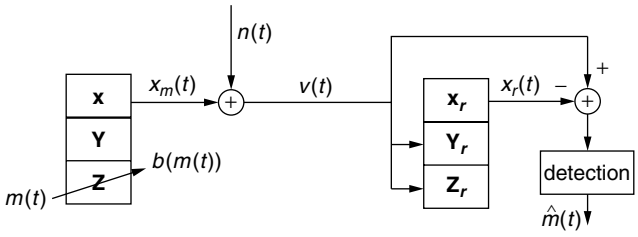


Figure 4. Chaotic switching modulation.

error $e_s(t)$ is defined in Eq. (3) as the difference between the noisy received signal $r(t) = x_m(t) + n(t)$ and the regenerated signal $x_r(t)$ at the receiver, and is given by

$$e_s(t) = r(t) - x_r(t) \tag{3}$$

For this binary hypothesis problem, the sufficient statistic η_s is the squared mean of the synchronization error after discarding some transient data and is defined by

$$\eta_s = \frac{1}{T}\int_{t_0}^{t_0+T} e_s^2(t)\,dt \tag{4}$$

Synchronization of the chaos signal is a common feature of the above three chaotic modulation–demodulation schemes, and system performance depends crucially on this synchronization capability. When the channel condition is so poor that it is impossible to achieve chaotic synchronization, different chaotic modulation techniques for digital communication, based on variations of chaos shift keying modulation (CSK) (see Fig. 5, have been introduced. The details of this approach will be discussed in Section 2.5.

2.4. Numerical Algorithm and Performance Evaluation of Chaotic Communications Based on Stochastic Calculus

Because of inherent nonlinearity of chaotic systems, the analytic performance evaluation of a chaotic communication system using chaos as a carrier is in general very difficult, and thus a numerical simulation approach is needed as shown by Chen and Yao [20]. It is known that commercial numerical computational packages using the standard Euler or Runge–Kutta (RK) algorithm designed for a deterministic differential equation to approximate the solution of a nonlinear stochastic differential equation (SDE) will incur significant errors [38]. This is particularly true for a nonlinear chaotic dynamical system modeling the transmitter inputting into an AWGN channel and followed by another nonlinear chaotic dynamical system.

We use the stochastic calculus approach to perform the integration algorithm for the sample functions of nonlinear dynamic systems excited by the stochastic white noise. Depending on the precise interpretation of the white noise, there are two different solutions to the SDE based on the Stratonovich or Ito integral [40–42]. Using the conversion between them, a correct numerical integration algorithm in the Ito sense is introduced. With this algorithm, the correct error probability of the robust self-synchronization

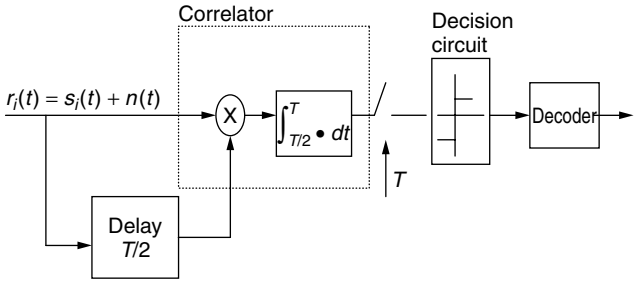


Figure 5. Differential chaos shift keying receiver.

Lorenz communication system with AWGN perturbation can be evaluated numerically. In the following, we present the numerical problem when evaluating the synchronization ability of the Lorenz system. Details regarding the error probabilities of CSM and DFM systems are given in Ref. 20.

2.4.1. Problem Description. A modified Lorenz system [17] is given by

$$\begin{aligned}\frac{dx}{d\tau} &= \sigma(y(t) - x(t))\\ \frac{dy}{d\tau} &= rx(t) - y(t) - 20x(t)z(t)\\ \frac{dz}{d\tau} &= 5x(t)y(t) - bz(t)\end{aligned} \tag{5}$$

where σ, r, and b are system parameters, and $\tau = t/K$, in which K is a timescaling factor. To characterize the robust ability of synchronization to white noise, the modified Lorenz system can be interpreted as the drive system, the signal $v(t)$ is the received waveform at the response system as defined by

$$\begin{aligned}\frac{dx_r}{d\tau} &= \sigma(y_r(t) - x_r(t))\\ \frac{dy_r}{d\tau} &= rv(t) - y_r(t) - 20v(t)z_r(t)\\ \frac{dz_r}{d\tau} &= 5v(t)y_r(t) - bz_r(t)\end{aligned} \tag{6}$$

where $v(t) = x(t) + n(t)$, and $n(t)$ is white Gaussian noise with zero mean and power spectrum density σ_n^2. The chosen coefficients are $\sigma = 16$, $r = 45.6$, and $b = 4$.

We define the vector $\mathbf{X} = [x, y, z, x_r, y_r, z_r]^T$. The entire system composed of the drive subsystem and the response subsystem can be viewed as a nonlinear system with an external white-noise input, and has the following standard form

$$\dot{\mathbf{X}}_i = f_i(\mathbf{X}) + g_i(\mathbf{X})n(t),\ i = 1, 2, \ldots, 6 \tag{7}$$

where f_1, f_2, f_3 are the modified Lorenz system equations, and

$$\begin{aligned}g_1 &= g_2 = g_3 = g_4 = 0\\ f_4 &= \sigma(y_r - x_r)\\ g_5 &= r - 20z_r\\ f_5 &= rx - y_r - 20xz_r\\ g_6 &= 5y_r\\ f_6 &= 5xy_r - bz_r\end{aligned} \tag{8}$$

and $n(t)$ is the WGN with zero mean and unity variance. The evolution of (7) is then given by

$$\mathbf{X}_i(t_0 + h) = \mathbf{X}_i(t_0) + \int_{t_0}^{t_0+h} f_i(\mathbf{X})\,dt + \int_{t_0}^{t_0+h} g_i(\mathbf{X})n(t)\,dt \tag{9}$$

A commonly made mistake is to treat the third term of (9) using the deterministic ordinary calculus method and apply the standard Euler or RK integration algorithm. Thus, assuming $g_i(\mathbf{X})$ is a smooth function, the integration result can be approximated by

$$\int_{t_0}^{t_0+h} g_i(\mathbf{X})n(t)\,dt \approx g_i(\mathbf{X}(t_0))Y_1 h \tag{10}$$

where Y_1 is a Gaussian random variable with zero mean and unity variance.

In order to illustrate this issue clearly, we use the preceding algorithm to characterize the robust self-synchronization ability of a Lorenz system by numerical computation as considered in Ref. 17. The simulation results are shown in Fig. 6 with the middle dashed curve, which is consistent with that in Ref. 17. The definitions of the input SNR and output SNR quantities in Fig. 6 are defined by input $\text{SNR} = 10\log_{10}(\sigma_x^2/\sigma_n^2)$, and output $\text{SNR} = 10\log_{10}(\sigma_x^2/\sigma_e^2)$, where σ_x^2 is the power of transmitted signal $x(t)$, and σ_e^2 is the power of the synchronization error $e(t) = x(t) - x_r(t)$. Clearly, we note the output SNR varies with the integration step size using the standard Euler/RK integration algorithm. Furthermore, the output SNR decreases by 3 dB as the step size is doubled. This is not a reasonable consequence for a given system which is excited by a stationary external white noise.

2.4.2. Numerical Algorithm for SDE. For the above-described chaotic system, the corresponding SDE in the sense of Ito is

$$dX_i = f_i(\mathbf{X})\,dt + g_i(\mathbf{X})\,dw(t), \mathbf{X}(t_0) = \mathbf{X}_0 \tag{11}$$

where $w(t)$ is the one-dimensional Wiener process or Brownian motion, and $\mathbf{X}_0$ is the initial conditions.

Mannella and Palleschi [39] have derived an accurate integration algorithm in the sense of Stratonovich [40], which treats the white noise as the limiting behavior of band-limited white noise, and the approximate results are summarized by the following algorithm

$$X_i(h) - X_i(0) = \delta X_i^{1/2} + \delta X_i^1 + \delta X_i^{3/2} + \delta X_i^2 + \cdots \tag{12}$$

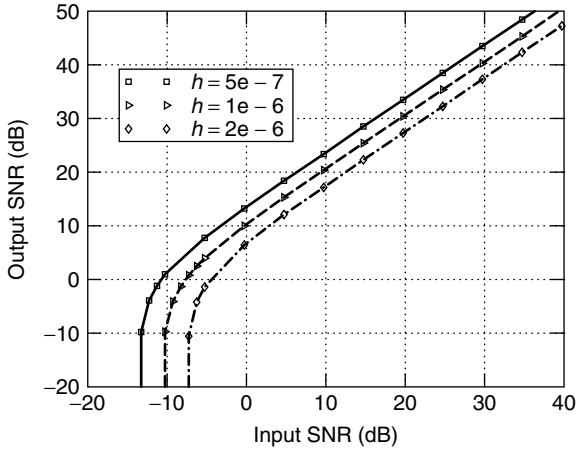


Figure 6. Robust ability for incorrect algorithm simulation with $K = 2505$.

where h is the integration step size, and $\delta X_i^{1/2} = g_i \int_0^h dw(t) = \sqrt{h} g_i Y_1$, where Y_1 is a Gaussian random variable with zero mean and unity variance, while the remaining terms are given in Eq. (6) of Ref. 39.

According to Stratonovich [40–42], the integral in the sense of Stratonovich can be converted into an Ito integral by adding one correction term. That is, if the SDE is modified and is given in the sense of Stratonovich as

$$dX_i = \left[f_i(\mathbf{X}) - \frac{1}{2}\sum_J \frac{\partial g_i(\mathbf{X})}{\partial \mathbf{X}_J} g_J(\mathbf{X}) \right] dt + g_i(\mathbf{X})\,dw(t) \tag{13}$$

the system evolution of (13) by using the above numerical algorithm is statistically equivalent to the evolution of (11) in the sense of Ito [41,42], which is desired here because the stochastic term $n(t)$ is modeled as a true white noise. We use this algorithm to resimulate the robust self-synchronization ability for white noise, and the simulation results are shown Fig. 7. Now, simulation results are consistent using different integration step sizes, which is necessary for a valid system modeling. [20] also provided the error probabilities of CSM and DFM communication systems using the appropriate stochastic integration algorithm as described above (see Fig. 8).

2.5. CSK, DCSK, and FM-DCSK

As seen in Fig. 8, the extremely poor performances of antipodal DFM and CSM systems motivate a fundamental consideration of the use of chaotic waveforms for digital communications. Since the chaotic waveform used in a coherent receiver, obtained by self-synchronization, is known to be extremely sensitive to channel noise, one approach to improve the system performance is to consider the use of different versions of CSK modulations. For an antipodal CSK transmitter, let $x(t)$ be a chaotic waveform and the modulated bits 0 and 1 be given by $s_0(t) = x(t),\ 0 \le t \le T$, and $s_1(t) = -x(t),\ 0 \le t \le T$, respectively. Of course, at the receiver, in order to perform coherent correlation operation, the chaotic waveform $x(t)$ is not available if the self-synchronization property of the chaotic waveform is

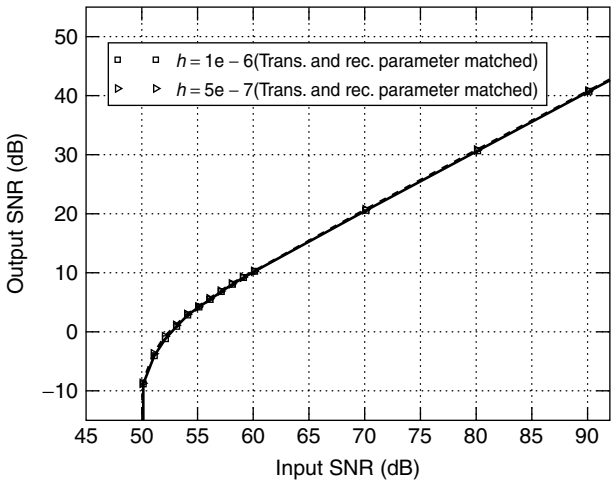


Figure 7. Robust ability for Mannella algorithm simulation with $K = 2505$.

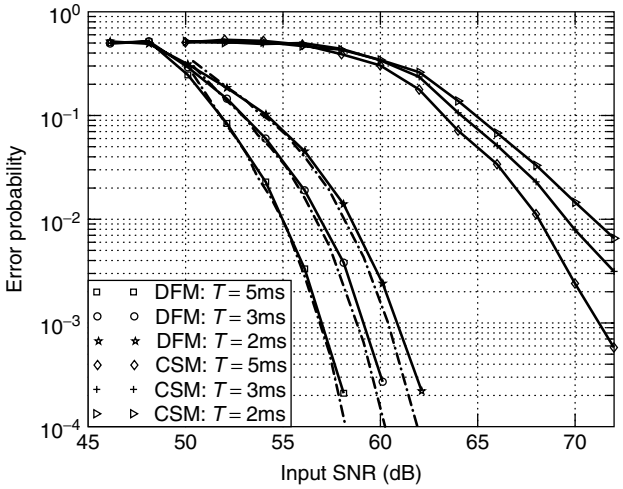


Figure 8. Comparison of BER versus input SNR for DFM and CSM systems.

not used. One way to circumvent this problem is to use the differential CSK (DCSK) modulation scheme as advocated by Kolumban [22]. For an antipodal DCSK transmitter, the two waveforms are given by

$$s_0(t) = \begin{cases} x(t), 0 \le t \le \dfrac{T}{2}, \\ x\left(t - \dfrac{T}{2}\right), \dfrac{T}{2} \le t \le T \end{cases}$$

$$s_1(t) = \begin{cases} x(t), 0 \le t \le \dfrac{T}{2} \\ -x\left(t - \dfrac{T}{2}\right), \dfrac{T}{2} \le t \le T \end{cases} \tag{14}$$

Then the received waveform is given by

$$r(t) = s_i(t) + n(t) = \begin{cases} x(t) + n(t), 0 \le t \le \dfrac{T}{2}; x\left(t - \dfrac{T}{2}\right) \\ \qquad +n(t), \dfrac{T}{2} \le t \le T, H_0 \\ x(t) + n(t), 0 \le t \le \dfrac{T}{2}; -x\left(t - \dfrac{T}{2}\right) \\ \qquad +n(t), \dfrac{T}{2} \le t \le T, H_1 \end{cases} \tag{15}$$

where $n(t)$ is the AWGN channel noise. For a differential coherent detection scheme, the first portion of the received chaotic waveform is used as a reference to perform coherent integration of the second portion of the received waveform. Then the output of the integrator is given by

$$\eta = \int_0^{T/2} r(t) r\left(t + \frac{T}{2}\right) dt = \pm \int_0^{T/2} x^2(t)\, dt \pm \int_0^{T/2} x\left(t + \frac{T}{2}\right) n(t)\, dt + \int_0^{T/2} x(t) n\left(t + \frac{T}{2}\right) dt + \int_0^{T/2} x\left(t + \frac{T}{2}\right) n(t)\, dt \tag{16}$$

In a conventional differential coherent BPSK system, the output of the integrator has the same form as that given in (16) except for the first term on the right-hand side (RHS). In a conventional system, $x(t)$ is a known deterministic waveform, and thus the first term is a known constant. However, for a chaotic $x(t)$ waveform that first term is not a constant, but only the expectation of that term (i.e., $E\{\int_0^{T/2} x^2(t)\, dt\}$) is a constant [22]. The nonconstancy of this term causes an additional system performance degradation of the DCSK system relative to the differential coherent BPSK system. We also note that, in the DCSK scheme, since the first portion of the waveform over a duration of $T/2$ seconds carries no information, the transmission rate of the system is one-half that of a nondifferential CSK system.

One possible remedy to eliminate the nonconstancy problem is to frequency-modulate the chaotic $x(t)$ waveform, which results in still another chaotic waveform but has constant energy. This leads to the FM-DCSK modulation scheme [24,25]. The demodulation of the FM-DCSK waveforms is performed in the same manner as that of the DCSK system, except now FM chaotic waveforms are used. Performances of non-band-limited BPSK, FSK, noncoherent FSK, and FM-DCSK are compared in Fig. 9. Some advantages of the FM-DCSK system [25] include the following:

1. There is no need to use the self-synchronization property of the chaotic waveform, which yields low performance in the presence of noise; it is relatively insensitive to channel distortions, and thus nonlinear amplifier can be used.
2. Since the FM-DCSK waveforms are wideband, they are relatively immune to frequency-selective fading in multipath scenarios.
3. Unlike the case in conventional DS-CDMA systems, since chaotic waveforms are used here, no distinct spectral lines are present; multiuser capability also exists.

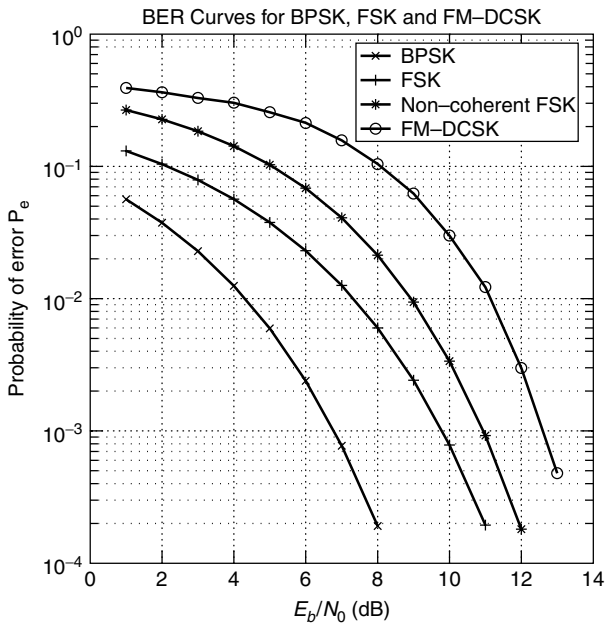


Figure 9. Comparison of BER versus E_b/N_0 of BPSK, FSK, noncoherent FSK, and FM-DCSK systems.

FM-DCSK is probably one of the more practical communication systems using chaos-based carrier modulations.

Besides frequency modulation, the basic DCSK scheme can be extended in various directions. Just as QPSK and QAM are extensions of BPSK in order to transmit more information data per baud, multilevel quadrature CSK scheme has also been proposed by Galias and Maggio [37].

3. SYMBOLIC DYNAMICS

A quite different chaotic communication technique using symbolic dynamics was originally proposed by Hayes et al. [27]. This approach attempts to provide a bridge between the theory of the chaotic system and information theory to design a communication system using chaotic dynamics. From the formalism of *symbolic dynamics* [63], Hayes et al. considered a chaotic system to be a natural digital information source with some constraint, denoted as the *grammar*. They showed that the symbolic dynamics of a chaotic oscillator can be made to follow a desired symbol sequence by using small perturbations. The mechanism behind the chaos control can be explained in an abstract but simple way [28]. Consider the Bernoulli map given by $x_{n+1} = 2x_n \bmod 1$. If x is represented by a binary fraction with finite precision, say, $x_n = .10101010$, then either $x_{n+1} = .01010101$ or $x_{n+1} = .01010100$ can be obtained by changing the eighth significant bit, representing a change of about 0.004 in base 10. By repeatedly changing the eighth bit, the change will show up in the most significant bit, which determines whether $x \geq \frac{1}{2}$ or $x < \frac{1}{2}$, a large-scale and easily observable signal attribute. Therefore, any message that can be encoded in a sequence of bits can be transmitted by controlling the symbolic dynamics of the chaotic system. The receiver is a simple level threshold detector.

For the continuous-time and continuous-state system, which is described by a set of ordinary differential equations, the symbolic dynamics of the system can be constructed using the Poincaré section concept by running the system without control [27]. Then we associate each binary sequence generated when the corresponding trajectory, starting from the initial condition, crosses the Poincaré surface with a real number, denoted by the *coding function*. The information message can be embedded into the chaotic dynamics by applying control pulses according to the coding function. The receiver observes the transmitted chaotic waveform and make a decision by observing the points on the Poincaré surface to see which side of the surface the crossing point lies in.

As stated earlier, in a DCSK system, the first portion of the chaotic waveform is used for reference purpose, carries no information, and yields an inefficient use of the channel bandwidth. Maggio and Galias [43] encoded some information onto the first portion of the chaotic waveform through the use of symbolic dynamics and increased the effective throughput of the system. Ciftci and Williams [44] proposed the use of optimum Viterbi estimation and channel equalization algorithms to estimate sequences encoded using symbolic dynamics over a channel with distortion. Other advanced symbolic dynamics and related analytic tools relevant to chaotic communications include those described in Refs. 45–47.

4. ANALOG CHANNEL ENCODING AND ESTIMATION

The use of chaotic dynamics as a channel encoder was proposed by Chen and Wornell [30]. A novel analog code based on the tent map dynamics and having a fast decoding algorithm was proposed for use on unknown, multiple, and time-varying channels with different SNRs. These practical chaotic codes having recursive receiver structures and important performance advantages over conventional codes were demonstrated. A convolutional encoder and multiresolution codes using chaotic systems are also developed in this article.

The basic idea of this chaotic encoder is to encode the analog message into the initial condition of the chaotic system, and an estimation technique was used to retrieve the message. Optimal state estimation for chaotic sequences is in general a difficult problem. Papadopoulos and Wornell [29] derived the maximum-likelihood (ML) estimator for the tent map sequences in stationary AWGN and showed that it can be implemented by a forward recursive filter followed by a backward recursive smoother. The forward recursive filter developed by Papadopoulos and Wornell [29] is identical to the Kalman filter.

5. CHAOTIC PULSE POSITION MODULATION

Instead of transmitting the chaos waveform, a chaotic pulse position modulation (CPPM) scheme was proposed Sushchik et al. [31]. This proposed system is similar to an ultra-wide-bandwidth impulse radio system of Win [32] that offers a promising communication method, especially in a severe multipath environment. A pulse position method is used to modulate binary information onto the carrier. The separation between the adjacent pulses is chaotic, arising from a dynamical system with irregular behavior.

The communication scheme is built around a chaotic pulse regenerator (CPRG), as shown in Fig. 10. With a pulse train of interpulse intervals T_i as its input to the CPRG, the nth incoming pulse is produced after a delay time given by $\Delta T_n = F(T_{n-1}, \ldots, T_{n-k})$ where $F(.)$ is a chaotic map. Thus, the system is expected to generate a pulsetrain with chaotic interpulse intervals. The binary information message is modulated at the output of the CPRG by delaying the pulse of a fixed time if binary 1

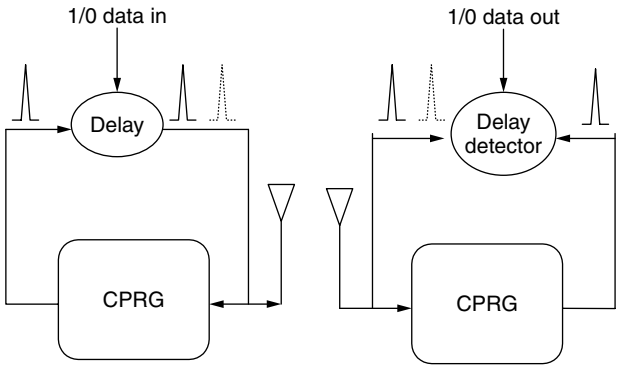


Figure 10. Chaotic pulse position modulation.

is being transmitted, or leaving unchanged if binary 0 is being transmitted. The received signal is fed into the identical CPRG at the receiver. Since the inputs to both CPRGs at the transmitter and receiver are identical, the outputs of CPRGs are expected to be identical. By computing the timing difference between the received signal and the output signal of the CPRG at the receiver, the embedded message can be retrieved.

This communication system may have a lower probability of intercept due to the aperiodicity of the chaos signal. This system performs well compared to other chaos-based convert communication schemes. Rulkov et al. have further analyzed the CPPM system with application to multiuser communication [64].

6. CHAOTIC SPREAD-SPECTRUM SEQUENCES

There has been an increasing interest in spread-spectrum communications, particularly in code-division multiple-access (CDMA) format. Some of the operations, concepts, and advantages of direct-sequence CDMA (DSCDMA) have been described [48]. System performance of DSCDMA critically depends on the auto- and cross-correlation properties of the *spreading sequences*. The use of chaotic sequence/waveform as spread sequences for DSCDMA communication systems has been proposed [49–51,53–58]. The chaotic signals have low nonzero shift autocorrelation and all cross-correlation properties due to the intrinsic broad nature of the spectrum and sensitivity to initial conditions. The use of the correlation function characteristics of some specific chaotic sequences and its comparison to *m* sequences/Gold sequences has been studied.

The chaotic sequence, generated by a logistic map, was first proposed for DSCDMA in 1992 [49]. The correlation properties of these logistic sequences are similar to random white noise. By simulations, the performance of chaotic sequences in the DSCDMA system is shown to be similar to that of PN sequences. Furthermore, due to their real values instead of binary values, chaotic sequences outperform PN sequences in low probability of intercept (LPI). Umeno et al. [51] used Chebyshev sequences for a synchronous CDMA system and analyzed the system performance using ergodic theory [7,52].

The statistical properties of binary sequences generated by a class of ergodic maps with some symmetric properties have been discussed [53]. Simple methods were used to generate a sequence of i.i.d. binary random sequence. The correlation functions of various types of chaotic binary sequences have been evaluated exactly by ensemble-averaging technique based on the Perron–Frobenius operator theory [7]. They also obtained a sufficient condition for a binary function to produce a sequence of i.i.d. binary random variables.

Mazzini et al. [54,55] proposed chaotic complex spreading sequences for asynchronous DSCDMA. They provided rigorous analysis of DSCDMA system performance bounds using chaotic complex spreading sequences. The simulation results in these papers show that systems based on chaotic spreading sequences perform generally better than the *Gold sequences*. Moreover, by treating the spreading sequences as random processes, Mazzini et al. [57] have shown the optimal *ensemble-averaged* autocorrelation of spreading sequences with minimum achievable interference variance decays nearly exponentially. They also proposed a chaotic map to generate the sequences with nearly exponential autocorrelation function. Without the assumption of independence and stationarity of the spreading sequences required by Mazzini et al. [57], Chen and Yao [60] provided a methodology to derive the general results on the partial autocorrelation function of the spreading sequences to minimize interference variance as well as a real-valued spreading sequence implementation that is at the same time optimal and practical by using the ergodic theory. For an asynchronous DSCDMA system, results in Refs. 57 and 60 show that these sequences generated based on chaos and ergodic theory concepts, can support approximately 15% more users for a fixed amount of interferences compared to Gold codes. Equivalently, for a fixed number of users, these sequences produce ~15% fewer interferences. In Fig. 11, there are nine BER curves for sequences of length 64 in AWGN. But there are essentially only three sets of curves, with each set having about the same performance. The lowest set of curves include the two optimal sequences for asynchronous CDMA operation (induced from optimal filtering of either a second- or third-order Chebyshev sequence); the middle set of curves include the Gold code and the original unfiltered second- or third-order Chebyshev sequence in asynchronous CDMA operation; and the upper set of curves represent operations under the chip-synchronized CDMA operations. There is an approximate 15% improvement in the lower set of curves compared to the middle set of curves. Various other related issues on sequences generated based on chaos theory and their implementations have also appeared in the literature [59,61,62].

7. CHAOS IN LASER COMMUNICATION AND MODELING OF RADAR AND RADIO PROPAGATION EFFECTS

There are many applications of chaotic nonlinear dynamics to various system problems. We consider only two such

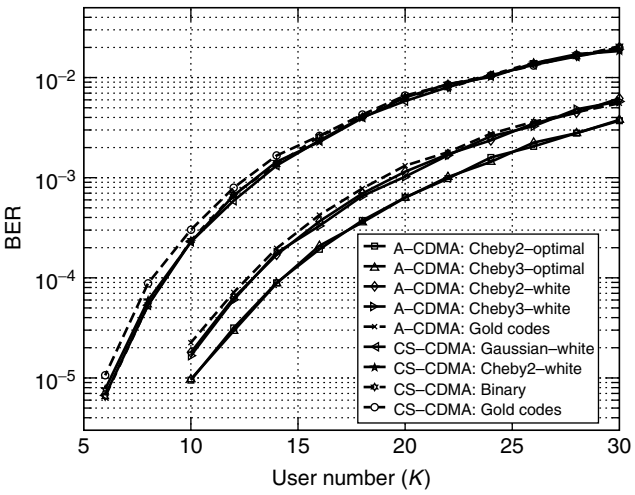


Figure 11. BER versus user number for nine chaotic spreading sequence asynchronous DSCDMA scenarios.

applications in communications. Semiconductor lasers are the most important light sources for optical communications. Unlike the previously considered electronic chaotic communication systems, where the nonlinearities are often introduced intentionally to create chaos, single-mode semiconductor lasers operate with various complex intrinsic nonlinearities due to the physical behaviors of the devices. Thus, chaos may or may not be avoidable in such devices. Many papers have dealt with the chaotic behaviors of lasers. The paper by Liu et al. [65] exploits the chaotic behaviors of the lasers and deals with the dynamics, synchronization, and message encoding and decoding of two optical laser communication systems. One system uses optical injection and the other uses delayed optoelectronic feedback. From these numerical simulation and experimental measurement works, the basic concept of using chaotic optical communication to transmit and receive hundreds of gigabit rate data has been demonstrated as feasible and practical.

Radar and radiofrequency propagation effects due to scattering, reflection, and shadowing are known in various scenarios to severely limit the performance of these systems. The modeling and understanding of these propagation effects are of great theoretical and practical interest. In this section we consider two physical problems. The first problem considers sea clutter as the backscattered returns from a patch of the sea surface illuminated by a transmitted radar pulse. Sea clutter waveforms are quite complicated. Traditionally, they have been modeled by statistical means such as through the marginal densities of the amplitude of the waveform. Often there is little insight or justification for these statistical characterizations such as lognormal, or k distribution. Since the sea clutter waveforms are functions of sometimes turbulent wave motions on the sea surface, it is not unreasonable to conjecture that perhaps nonlinear dynamics may be in operation. Perhaps the apparent seemingly randomness of the sea clutter may be modeled by deterministic chaos. Haykin and colleagues [66,67] have collected considerable amount of sea clutter data and used the method of Grassberger and Procaccia [68] and found nonintegral fractal dimensions and positive Lyapunov exponents from these data. Thus, they conjectured that under certain conditions, sea clutter may be modeled as deterministic chaos. But it is also known that a nonintegral fractal dimension together with a positive largest Lyapunov exponent obtained by computational means is not a sufficient condition. Specifically, the $1/f$ fractal random process (which is not a deterministic chaos) may have a nonintegral fractional dimension and also a positive Lyapunov exponent. Gao and Zheng [69,70] provided a more stringent test for chaos by showing that for a chaotic sequence, a plot of the time-dependent exponent $\Lambda(k)$–time index k forms a common envelope over different shells. The slope of these envelope is the largest Lyapunov exponent. This common envelope property is shown in Fig. 12 for the well-known chaotic Lorenz sequence. On the other hand, for a nonchaotic sequence, such as that for a white sequence (Fig. 13), the common envelope property does not hold. For the 130 s of sea clutter, Haykin et al. [71] obtained the results shown in Fig. 14, which does not manifest the common

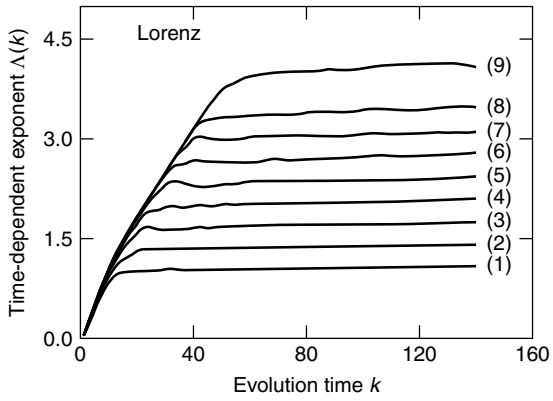


Figure 12. Exponent versus evolution time for different shells of a chaotic Lorenz sequence.

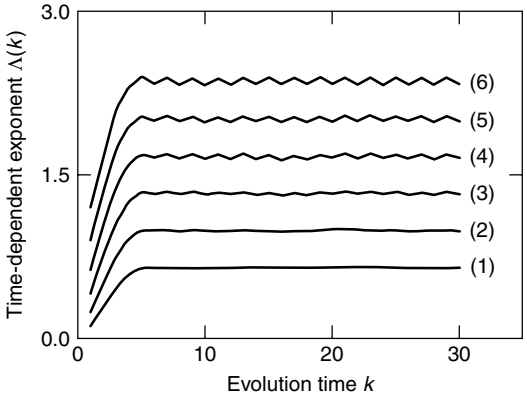


Figure 13. Exponent versus evolution time for different shells of a white-noise sequence.

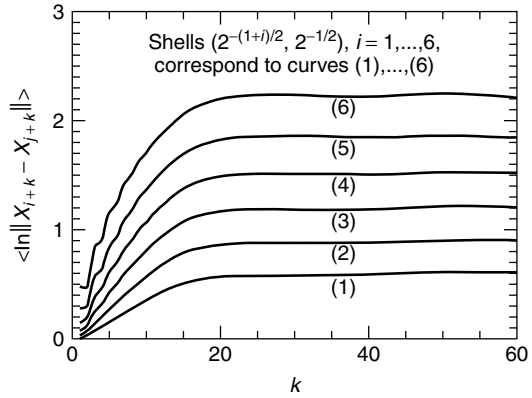


Figure 14. Exponent versus evolution time for different shells of a sea clutter sequence.

envelope property. This would seem to indicate that sea clutter data are not deterministic chaos. To obtain deeper insight from these sea clutter data, a multifractal analysis was performed [71]. Without going into detail, random multiplicative multifractal theory [76,77] shows that if a sequence of data has infinitely many power-law scaling relationships, then plots of $\log_2 M_q(\varepsilon)$ versus $-\log_2 \varepsilon$ for different values of q should form a straight line and the amplitudes must be lognormally distributed. Here the

moment $M_q(\varepsilon) = \sum_i w_i^q$, $\varepsilon = 2^{-N}$ at stage N, where q is a real number, and the weights $w_i, i = 1, \ldots$ are obtained directly from the measured data. For this set of sea clutter data, Fig. 15 shows the amplitude and particularly the envelopes form essentially sets of straight lines. Furthermore, Fig. 16 shows the amplitude and particularly the envelope indeed satisfies the lognormal distribution. Indeed, lognormal distribution for radar clutter amplitude has been known for many years, although no known justification has been given until now. It is also interesting that some work by Cowper and Mulgrew [72] and even Haykin et al. [73] shows reservations about the proper modeling of sea clutter as deterministic chaos. Clearly, more data should be collected and analysis performed in order to construct a valid propagation model for radar clutter returns.

The second problem considered in this section deals with the modeling of the fading radio propagation phenomena in wireless communication. There have been many propagation measurements in different frequency bands, and various statistical models have been proposed. Tannous et al. [74] applied the method of Grassberger and Procaccia [68] to some indoor 0.915-GHz radio propagation data, and found nonintegral fractal dimensions together with positive largest Lyapunov exponents and concluded that the propagation channel may be modeled by chaos. However, due to the limited data length and highly nonstationary character of their measured data, it is not obvious that a deterministic chaos conclusion can be made from these data. Gao et al. [75] reported on short timescaled analysis of some indoor 1-GHz propagation data using the method due to Gao and Zheng [69,70]. Figure 17 shows a

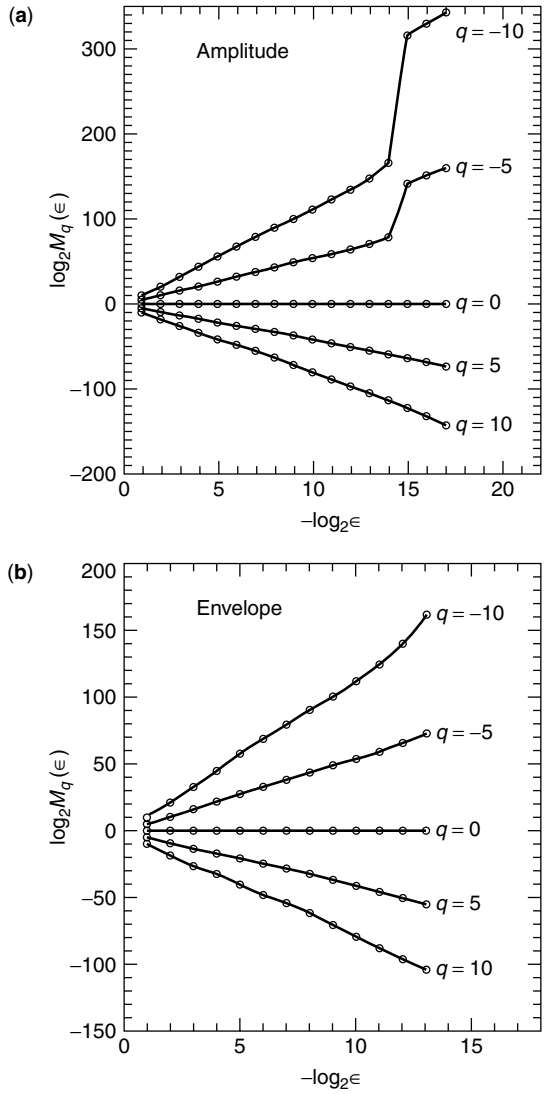


Figure 15. Multifractal scaling law for amplitude and envelope of a sea clutter sequence.

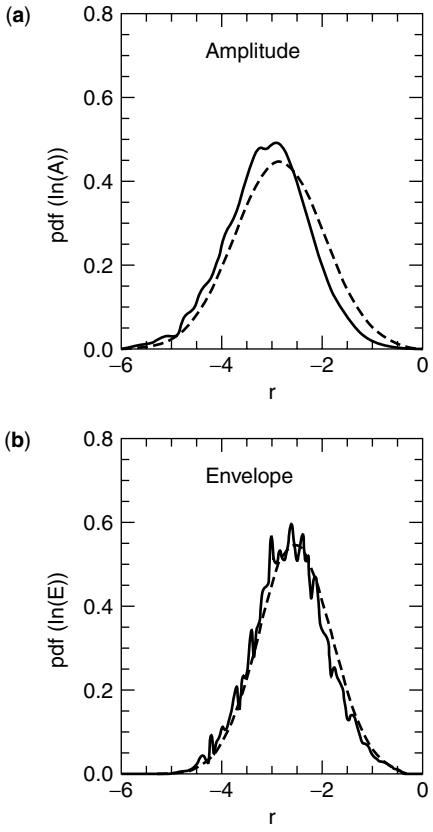


Figure 16. Probability distribution functions of log of amplitude and envelope versus values.

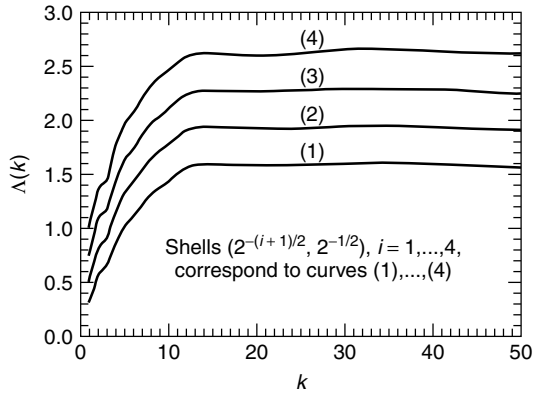


Figure 17. Exponent versus evolution time for different shells of a measured radio propagation data.

typical plot of the time-dependent exponent $\Lambda(k)$ versus time index k and reveals no common envelope property. From the earlier comments made with regard to Figs. 12 and 13, we may conclude that at least for these measurements, deterministic chaos were not present. Further analysis based on fractional Brownian motion process (FBMP) [78] were reported [75]. A FBMP is a Gaussian process with zero-mean and stationary increments. Its variance has the form of t^{2H}, and its power spectral density has the form of $f^{-(2H+1)}$, where H is the Hurst parameter. For $0.5 < H < 1$, the process has long-range dependence. For $H = 0.5$, the process becomes the standard BMP (whose formal derivative is the standard white Gaussian noise used in communication system analysis). For $0 < H < 0.5$, the adjacent values of the process have highly negative correlations and the process fluctuates widely. Details are omitted here, but if the data are from an FBMP, then the set of variance–time $\log_2 F_q(m)/\log_2 m$ curves in Fig. 18 should form straight lines. For two sets of measured data, the estimated Hurst parameters have almost constant values of about 0.4 and 0.45. For these limited number of measurements, the data may suggest a fractal FBM-like process modeling. However, since the RF band around 1 GHz is crowded with various radio transmitters and microwave ovens, drawing definitive conclusion about the true nature of the observed data is delicate. It is interesting that researchers in electromagnetics [79] have also proposed the modeling of RF scattering based on fractal theory. Clearly, more careful data collection and advanced tools for analyzing these data must be performed before valid models can be established and these results can be applied to practical wireless communication applications.

8. CONCLUSIONS

In this article, we have summarized various aspects of deterministic chaos to the analysis, design, and modeling of communication systems and channels. We first provided some background on the history of chaos. Then early chaotic modulation techniques utilizing self-synchronization of the chaotic waveform were discussed. Evaluation of the error probability of these systems cannot be performed analytically or even by conventional numerical simulation. A numerical simulation method for a nonlinear stochastic system is discussed, and an accurate stochastic integration algorithm is provided. Because of the sensitivity of the self-synchronization process to channel distortion and noise, the performances of these early chaotic communication systems were poor. The FM-DCSK modulation scheme was proposed and shown to be practical for implementation. It is competitive to conventional communication system performance and may possess certain additional desirable properties. Chaotic pulse position modulation particularly with application to UWB systems may also be practical and useful. More recently, various chaotic nonlinear dynamics and ergodic theory concepts have been proposed to create CDMA sequences that can perform better than known sequences in the asynchronous CDMA mode. Preliminary investigations show these results are practical for implementation.

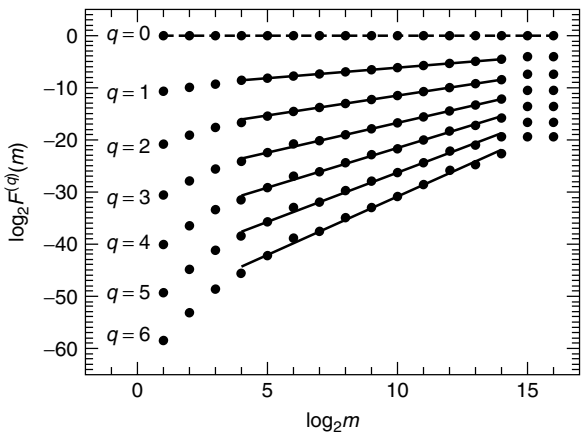


Figure 18. Power versus time for different scalings for a measured radio propagation data.

It is quite clear that many of the researchers in "chaotic communications" are more interested in exploiting various complex and often interesting properties of chaotic nonlinear dynamics rather than using them for explicit communication purpose. Ultimately, "chaotic communication" schemes must be compared to comparable conventional communication schemes with respect to bandwidth, data rate, energy per bit, error probability, and complexity of the transmitter and receiver. It remains a challenge to exploit the complexity and richness of chaos and related analytic tools to understand, analyze, design, and model communication systems and channels.

Acknowledgment

This work is partially supported by an ARO-MURI grant on "chaotic communication," a UC-CoRe grant sponsored by ST Microelectronics, and NASA/Dryden grant NCC4-153.

BIOGRAPHIES

Kung Yao received the B.S.E. (Highest Honors), M.A., and Ph.D. degrees in electrical engineering, all from Princeton University, Princeton, New Jersey. He was a NAS-NRC Post-Doctoral Research Fellow at the University of California, Berkeley. Presently, he is a Professor in the Electrical Engineering Department at UCLA. In 1969, he was a Visiting Assistant Professor at the Massachusetts Institute of Technology. In 1985–1988, he served as an Assistant Dean of the School of Engineering and Applied Science at UCLA. His research and professional interests include sensor array systems, digital communication theory and systems, smart antenna and wireless radio systems, chaos communications and system theory, digital and array signal and array processing, systolic and VLSI algorithms, architectures and systems, radar systems, and simulation. He has published over 250 journal and conference papers. Dr. Yao received the IEEE Signal Processing Society's 1993 Senior Award in VLSI Signal Processing. He was the co-editor of a two-volume series of an IEEE reprint book entitled *High Performance VLSI Signal Processing*, IEEE Press, 1997. He was a Guest Associate Editor of a Special Issue on "Applications of Chaos in Modern Communication Systems" of the *IEEE Transactions on Circuits and Systems* — Part I, December 2001. He is a Fellow of IEEE.

Chi-Chung Chen was born in Taiwan in 1970. He received the B.S. and M.S. degrees in control engineering from National Chiao Tung University, Taiwan, in 1993 and 1995, respectively, and the Ph.D. degree in electrical engineering from the University of California at Los Angeles in 2000. He was employed as an Associate Researcher on the design of controller in MIRL, ITR1, Taiwan from 1995 through 1996. From 1997 to 2000, he was a Research Assistant at UCLA. His current research interests include chaotic communications, spread-spectrum COMA systems, pseudorandom sequences, adaptive systems, and wireless communication systems. Dr. Chen is a member of the Phi Tau Phi Scholastic Honor Society. Since 2001, he has been with Accton Technology Corporation in Tainan, Taiwan.

BIBLIOGRAPHY

1. E. Lorenz, Deterministic nonperiodic flow, *J. Atm. Sci.* **20**: 130–141 (1963).
2. B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, 1982.
3. *IEEE Trans. Circuits Syst.* (Special Issue on Chaos in Nonlinear Electronic Circuits) **40**(10): (1993).
4. J. Gleick, *Chaos: The Amazing Science of the Unpredictable*, Random House, 1988.
5. G. P. Williams, *Chaos Theory Tamed*, Joseph Henry Press, 1997.
6. R. C. Hilborn, *Chaos and Nonlinear Dynamics*, Oxford Univ. Press, 1994.
7. A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, 2nd ed., Springer-Verlag, 1994.
8. L. M. Pecora, Overview of chaos and communications research, *SPIE* **2038**: (1993).
9. M. Hasler, Synchronization of chaotic systems and transmission of information, *Int. J. Bifurc. Chaos* **8**(4): 647–659 (1998).
10. A. V. Oppenheim, G. W. Wornell, S. H. Isabelle, and K. M. Cuomo, Signal processing in the context of chaotic signals, *Proc. IEEE ICCASP* **4**: 117–120 (March 1992).
11. M. P. Kennedy, R. Rovatti, and G. Setti, *Chaotic Electronics in Telecommunications*, CRC Press, 2000.
12. *IEEE Trans. Circuits Syst. — I* (Special Issue in Noncoherent Chaotic Communications) **47**: (Dec. 2000).
13. *IEEE Trans. Circuits Syst. — I* (Special Issue on Applications of Chaos in Modern Communication Systems **48**: (Dec. 2001).
14. *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: (May 2002).
15. L. M. Pecora and T. L. Carroll, Synchronization in chaotic systems, *Phys. Rev. Lett.* **64**: 821–824 (Feb. 1990).
16. L. M. Pecora, T. L Carroll, G. A. Johnson, and D. J. Mar, Fundamentals of synchronization in chaotic systems, concepts, and applications, *Int. J. Bifurc. Chaos* **7**: 520–543 (1997).
17. K. M. Cuomo, A. V. Oppenheim, and S. H. Strogatz, Synchronization of Lorenz-based chaotic circuits with applications to communications, *IEEE Trans. Circuits Syst. — II* **40**: 626–633 (1993).
18. V. Milanovic and M. E. Zaghloul, Improved masking algorithm for chaotic communications systems, *IEE Electron. Lett.* **32**: 11–12 (1996).
19. L. Kocarev, K. Halle, K. Eckert, and L. Chua, Experimental demonstration of secure communication via chaotic synchronization, *Int. J. Bifurc. Chaos* **2**: 709–713 (Sept. 1992).
20. C. C. Chen and K. Yao, Stochastic-calculus-based numerical evaluation and performance analysis of chaotic communication systems, *IEEE Trans. Circuits Syst. — I* **47**: 1663–1672 (Dec. 2000).
21. H. Dedieu, M. Kennedy, and M. Hasler, Chaos shift keying: modulation and demodulation of a chaotic carrier using self-synchronizing Chua's circuits, *IEEE Trans. Circuits Syst. — II* **40**: 634–642 (Oct. 1993).
22. G. Kolumban, B. Vizvari, W. Schwarz, and A. Abel, Differential Chaos shift keying: A robust coding for chaotic communication, *Proc. NDES*, 1996, pp. 87–92.
23. G. Kolumban, M. P. Kennedy, and L. O. Chua, The role of synchronization in digital communications using chaos — Part II: Chaotic modulation and chaotic synchronization, *IEEE Trans. Circuits Syst. — I: Fund. Theory Appl.* **45**: 1129–1140 (1998).
24. M. P. Kennedy and G. Kolumban, Digital communication using chaos, *Signal Process.* **80**: 1307–1320 (2000).
25. G. Kolumban, G. Kis, Z. Jako, and M. P. Kennedy, FM-DCSK: A robust modulation scheme for chaotic communication, *IEICE Trans. Fund. Electron. Commun. Comput. Sci.* **E81-A**: 1798–1802 (Oct. 1998).
26. G. Kolumban, M. P. Kennedy, Z. Jako, and G. Kis, Chaotic communications with correlator receiver: Theory and performance limits, *Proc. IEEE* **90**: 711–732 (May 2002).
27. S. Hayes, C. Grebogi, and E. Ott, Communicating with chaos, *Phys. Rev. Lett.* **70**: 3031–3034 (May 1993).
28. S. Hayes, *Communicating with Chaos: A Physical Theory for Communication via Controlled Symbolic Dynamics*, Ph.D. thesis, Univ. Maryland, 1994.
29. H. C. Papadopoulos and G. W. Wornell, Maximum likelihood estimation of a class of chaotic signals, *IEEE Trans. Inform. Theory* **41**: 312–317 (1995).
30. B. Chen and G. W. Wornell, Analog error-correcting codes based on chaotic dynamical systems, *IEEE Trans. Commun.* **46**: 881–890 (1998).
31. M. Sushchik et al., Chaotic pulse position modulation: A robust method of communicating with chaos, *IEEE Commun. Lett.* **4**: 128–130 (April 2000).
32. M. Z. Win and R. A. Scholtz, Impulse radio: How it works, *IEEE Commun. Lett.* **2**: 360–363 (1998).
33. *IEEE Trans. Commun.* (Special Issue on Spread Spectrum Communications) **25**(8): (1977).
34. *IEEE Trans. Commun.* (Special Issue on Spread Spectrum Communications) **30**(5): (1982).
35. K. S. Halle, C. W. Wu, M. Itoh, and L. O. Chua, Spread spectrum communication through modulation of chaos, *Int. J. Bifurc. Chaos* **3**: 409–477 (1993).
36. G. Kolumban, M. P. Kennedy, and L. O. Chua, The role of synchronization in digital communications using chaos — Part I: Fundamentals of digital communications, *IEEE Trans. Circuits Syst. — I: Fund. Theory Appl.* **44**(10): 927–936 (1997).
37. Z. Galias and G. M. Maggio, Quadrature chaos-shift keying: theory and performance analysis, *IEEE Trans. Circuits*

Syst. — I (Special Issue on Applications of Chaos in Modern Communication Systems) **48**: 1510–1518 (Dec. 2001).

38. N. J. Kasdin, Runge-Kutta algorithm for the numerical integration of stochastic differential equations, *J. Guidance, Control, Dynamics* **18**: 114–120 (1995).
39. R. Mannella and V. Palleschi, Fast and precise algorithm for computer simulation of stochastic differential equations, *Phys. Rev. A* **40**: 3381–3386 (Sept. 1989).
40. R. L. Stratonovich, A new representation for stochastic integrals and equations, *SIAM J. Control* **4**: 362–371 (1966).
41. R. E. Mortensen, Mathematical problems of modeling stochastic nonlinear dynamic systems, *J. Stat. Phys.* **1**: 271–296 (1969).
42. L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Wiley, 1973.
43. G. M. Maggio and Z. Galias, Enhanced differential shift keying using symbolic dynamics, *Proc. IEEE GLOBECOM*, Nov. 2001, Vol. 2, pp. 1157–1161.
44. M. Ciftci and D. B. Williams, Optimal estimation and sequential channel equalization algorithms for chaotic communication systems, *EURASIP J. Appl. Signal Process.* **2001**: 249–256 (Dec. 2001).
45. D. Lind and B. Marcus, *Introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press, 1995.
46. T. Kohda, Information sources using chaotic dynamics, *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: 641–661 (May 2002).
47. G. Setti, G. Mazzini, R. Rovatti, and S. Callegari, Statistical modeling of discrete-time chaotic processes-basic finite-dimensional tools and applications, *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: 662–690 (May 2002).
48. A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, 1992.
49. G. Heidari-Bateni and C. D. McGillem, Chaotic sequences for spread spectrum: An alternative to PN-sequences, *Proc. IEEE Int. Conf. Selected Topics in Wireless Communications*, 1992, pp. 437–440.
50. G. Heidari-Bateni and C. D. McGillem, A chaotic direct-sequence spread-spectrum communication system, *IEEE Trans. Commun.* **422**(2–4): 1524–1527 (1994).
51. K. Umeno and K. I. Kitayama, Improvement of SNR with chaotic spreading sequences for CDMA, *Proc. IEEE Information Theory Workshop*, South Africa, June 1999, p. 106.
52. R. L. Adler and T. J. Rivlin, Ergodic and mixing properties of Chebyshev polynomials, *Proc. Am. Math. Soc.* **15**: 794–796 (1964).
53. T. Kohda and A. Tsuneda, Statistics of chaotic binary sequences, *IEEE Trans. Inform. Theory* **43**: 104–112 (1997).
54. G. Mazzini, G. Setti, and R. Rovatti, Chaotic complex spreading sequences for asynchronous DS-CDMA — Part I: System modeling and results, *IEEE Trans. Circuits Syst. — I* **44**: 937–947 (Oct. 1997).
55. R. Rovatti, G. Setti, and G. Mazzini, Chaotic complex spreading Sequences for asynchronous DS-CDMA — Part II: Some theoretical performance bounds, *IEEE Trans. Circuits Syst. — I* **44**: 937–947 (Oct. 1997).
56. R. Rovatti and G. Mazzini, Interference in DS-CDMA systems with exponentially vanishing autocorrelations: Chaos-based spreading is optimal, *IEE Electron. Lett.* **34**: 1911–1913 (October 1998).
57. G. Mazzini, R. Rovatti, and G. Setti, Interference minimization by auto-correlation shaping in Asynchronous DS-CDMA systems: Chaos-based spreading is nearly optimal, *IEE Electron. Lett.* **35**: pp. 1054–1055 (June 1999).
58. T. Yang and L. O. Chua Chaotic digital code-division multiple access (CDMA) communication systems, *Int. J. Bifurc. Chaos* **7**: 2789–2805 (1997).
59. L. Cong and L. Shaoqian, Chaotic spreading sequences with multiple access performance better than random sequences, *IEEE Trans. Circuits Syst. — I* **47**: 394–397 (March 2000).
60. C. C. Chen, E. Biglieri, and K. Yao, Design of spread spectrum sequences using chaotic dynamical systems and ergodic theory, *IEEE Trans. Circuits Syst. — I* **48**: 1110–1113 (Sept. 2001).
61. G. Mazzini, R. Rovatti, and G. Setti, Chaos-based asynchronous DS-CDMA systems and enhanced Rake receivers: Measuring and improvements, *IEEE Trans. Circuits Syst. — I* **48**: 1445–1454 (Dec. 2001).
62. C. C. Chen, K. Yao, and E. Biglieri, Optimal spread spectrum sequences — constructed from Gold codes, *Proc. IEEE GLOBECOM*, Nov. 2000, pp. 867–871.
63. B.-L. Hao, *Elementary Symbolic Dynamics and Chaos in Dissipative Systems*, World Scientific, Singapore, 1989.
64. N. Rulkov, M. Sushchik, L. Tsimring, and A. Volkvskii, Digital communication using chaotic pulse position modulation, *IEEE Trans. Circuits Syst. — I* (Special Issue on Applications of Chaos in Modern Communication Systems) **48**: 1436–1444 (Dec. 2001).
65. J. M. Liu, H. F. Chen, and S. Tang, Optical communication systems based on chaos in semiconductor lasers, *IEEE Trans. Circuits Syst. — I* (Special Issue on Applications of Chaos in Modern Communication Systems) **48**: 1475–1483 (Dec. 2001).
66. S. Haykin and S. Puthusserypady, Chaotic dynamics of sea clutter, *Int. J. Bifurc. Chaos* **7**: 777–802 (1997).
67. S. Haykin, *Chaotic Dynamics of Sea Clutter*, Wiley, New York, 1999.
68. P. Grassberger and I. Procaccia, Characterization of strange attractors, *Phy. Rev. Lett.* **50**: 346 (1983).
69. J. B. Gao and Z. M. Zheng, Local exponent divergence plot and optimal embedding of a chaotic time series, *Phys. Lett. A* **181**: 153–158 (1993).
70. J. B. Gao and Z. M. Zheng, Direct dynamical test for deterministic chaos and optimal embedding of a chaotic time series, *Phys. Rev. E* **49**: 3807–3814 (1994).
71. J. B. Gao and K. Yao, Multifractal features of sea clutter, *Proc. 2002 IEEE Radar Conf.*, April 2002, pp. 500–505.
72. M. R. Cowper and B. Mulgrew, Nonlinear processing of high resolution radar sea clutter, *Proc. IJCNN*, July 1999, Vol. 4, pp. 2633–2638.
73. S. Haykin, R. Bakker, and B. W. Currie, Uncovering nonlinear dynamics — the case study of sea clutter, *Proc. IEEE* (Special Issue on Applications of Nonlinear Dynamics to Electronics and Information Engineering) **90**: 860–881 (May 2002).
74. C. Tannous, R. Davies, and A. Angus, Strange attractors in multipath propagation, *IEEE Trans. Commun.* **38**: 629–631 (May 1991).

75. J. B. Gao et al., Can sea clutter and indoor radio propagation be modeled as strange attractors? *Proc. 7th Experimental Chaos Conf.*, Aug. 2002.

76. J. F. Gouyet, *Physics and Fractal Structure*, Springer-Verlag, 1995.

77. B. B. Mandelbrot, *Fractals and Scaling in Finance*, Springer-Verlag, 1997.

78. B. B. Mandelbrot and J. W. Van Ness, Fractional Brownian motions, fractional noises and applications, *SIAM Rev.* **10**: 422–437 (Oct. 1968).

79. D. L. Jaggard, On fractal electrodynamics, in H. N. Kritikos and D. L. Jaggard, eds., *Recent Advances in Electromagnetic Theory*, Springer-Verlag, 2001, pp. 183–224.

CHARACTERIZATION OF OPTICAL FIBERS

GILBERTO M. CAMILO
OmniGuide Communications
Cambridge, Massachusetts

1. INTRODUCTION

At present there are many practical applications of optical fibers of different technologies. Optical fibers are used primarily (90% of total applications) in telecommunications because optical signals can be transmitted over long distances at high speed in optical fibers. Other applications are also important, such as military, medical, and industrial. The economic advantage of the optical fiber technology and its exploding implementation has pushed the manufacturing process from 60 m/min in 1978 to 1800 m/min in 2002.

Different glasses and coating materials have been developed since the early 1980s, but most of the optical fibers installed in the field are silica glass optical fibers protected by a polymeric coating. The coating is applied during the manufacturing process and is used to protect the surface of the glass from defects caused by abrasion, thus avoiding premature breaks. The high purity of the silica glass being used can guarantee very low attenuation of the light radiation being transmitted and also good mechanical qualities.

Other optical fiber technologies not based in silica glass are under study or already in use as all-plastic optical fibers [1], used in illumination systems and short-distance data communication; and infrared optical fibers [2,3], manufactured using chalcogenide glasses, or crystal glasses, which are used in sensor systems and telecommunication systems.

The optical and mechanical characterization methods used in optical fiber qualification are in general the same for all optical fiber types. Depending on which optical fibers are being tested, the appropriate light wavelength is used during the measurement process. Optical fibers can operate from the visible range (600 nm) to the infrared range (10,000 nm).

Measurement methods for silica glass optical fibers, coated by a primary and a secondary polymeric material, with operating range 1300–1550 nm, are considered here. Multimode silica optical fibers will be discussed briefly.

A silica glass optical fiber has an overall polymeric coating diameter of 250 μm and a centered glass with diameter around 125 μm. The glass part is composed of a core and a cladding, one glass cylinder inside another with a slight difference in refraction index. In the lightpulse that travels through the single-mode optical fiber, the core guides only the fundamental mode. The diameter of the core in these optical fibers is around 9 μm. In multimode optical fibers, hundreds of modes propagate at the same time in the lightpulse, and the optical fiber core is around 50 μm in diameter.

Presently the most widely used coating polymers are ultraviolet-cured acrylates. For mechanical and optical reasons, two polymer layers with distinct mechanical properties are applied to the glass surface. The soft inner polymer is in direct contact with the glass and absorbs small mechanical stresses. The harder external polymer has much higher elastic modulus compared to the internal polymer and is intended to resist the environment and abrasion.

2. CHARACTERIZATION METHODS

The National Standard Committees formed by industries, universities, and government standard institutions propose and discuss various characterization methods. These methods are used to specify and qualify optical fiber systems, including optical fibers as a telecommunication product. Fiberoptic test procedures (FOTPs) are developed to provide a uniform plan of action during the tests of optical fiber system components. The Telecommunications Industry Association and the Electronic Industries Alliance (TIA/EIA) specify the optical fiber and optical fiber cable standards in the United States. Many other countries use these standards as well. In Europe, Japan, and other foreign markets the International Electrotechnical Commission (IEC) determine alternative telecommunication standards studies.

Among these organizations, in the United States and Europe, a homogenization process of procedures and specifications for optical fibers and optical fiber cables is currently under way.

The characterization methods and techniques discussed here are in accordance with national and international standard procedures. The optical fiber characterization can be divided in three different aspects:

- Geometric characterization
- Transmission characterization
- Mechanical characterization

2.1. Geometric Characterization

An optical fiber is an electromagnetic waveguide operating at a very short wavelength, around 1300–1550 nm. The micrometer optical fiber geometric characteristics are essential to maintain the integrity of the optical signal, which carries the information. According to the standard specifications, many parameters need to be checked and must be maintained inside narrow ranges; these include:

- Diameters of the core and of the cladding
- Core–cladding concentricity

- Ellipticity of the core and of the cladding
- Length of the optical fiber
- Numerical aperture
- Primary and secondary coating diameters

Numerical aperture is the cone angle light acceptance in front of the optical fiber. If the incident light radiation is inside this cone angle, and this light consists of a wavelength compatible with the geometric guidance characteristics of the optical fiber, this radiation will be coupled in the core and transmitted through the length of the fiber. If it is outside the cone angle, the radiation will be reflected by the glass in the extremity of the optical fiber or refracted and spread in the cladding, and from the cladding to the outside of the optical fiber.

Different techniques can be used to measure these optical fiber geometric parameters. The most common are

- Refracted near-field method
- Transmitted near-field method
- Transmitted far-field method
- Microscopy method

These techniques use the radiation pattern that is refracted or transmitted at a very short distance in the extremity of the optical fiber in order to measure the geometric parameters. In the refracted near-field method [4], a laser coupled to a very precise stepper motor is used to focus a lightbeam at different angles in the core of the optical fiber. By analyzing the cladding refracted light, it is possible to measure precisely the dimensions of the core–cladding and cladding–coating interfaces.

In the *transmitted field method*, after the light is transmitted by a small piece of optical fiber, very sensitive optical detectors, or charge-coupled device (CCD) cameras, measure its dimensions.

In the microscopic method, a microscope with a photographic machine, or a videocamera, inspects directly the transmitted radiation in the extremity of a small piece of optical fiber and performs the geometric measurement.

To measure the length of the optical fiber, it is most common to use optical time-domain reflectometry (OTDR) [5]. This technique was developed as a modification of a similar technique that was used during decades in metallic telecommunication cables. The great advantage of this technique is that it is necessary to have access to only one extremity of the optical fiber. In this method, modulated laser light is injected in the core of the optical fiber and during its propagation through the optical fiber length part of the radiation is reflected and spread by defects in the glass structure. A small backscattered portion of the reflected radiation returns to the extremity and brings back the information necessary to localize precisely (to within a few centimeters), those defects. By capturing the light that reflects in the other extremity it is possible to determine the length of the optical fiber.

In an OTDR, which is the equipment that uses such a technique in optical fibers, a small percentage of the incident radiation returns to the detector and defines the maximum length range to be analyzed. This technique can still be used to measure the attenuation of the optical signal after it has been transmitted through the optical fiber length. Actually, a commercial OTDR can measure the optical attenuation, the length, and localize defects to within a range of a few meters over 200 km of an optical fiber length.

It must be pointed out that some of these methods are still being refined, and the commercial apparatus that use them have significantly improved since the early 1980s.

2.2. Transmission Characterization

The transmission characteristics are related to capacity of the optical fiber and its ability to maintain the optical signal integrity after being conducted through the length of the fiber. Different characteristics of the optical fiber can affect its capacity to perform well at long distances. The most important factors that cause power depletion can be subsumed in what is called the attenuation of the optical fiber.

The optical fiber materials absorb part of the light radiation when the lightpulse travels through the glass and a part is spread in glass imperfections, causing attenuation. Most of the spread light goes from the core into the cladding, and from there it is lost. Absorption and spreading are caused by several different phenomena, including nonuniformity of the glass composition, particles and gas bubbles trapped in the glass structure, cracks and other mechanical defects in the glass, nonuniformity of the glass geometry, and defective coatings. The majority of these imperfections originate during the optical fiber manufacturing process and others result from installation and use of the optical fiber cables.

The methods most commonly used to measure the attenuation of the optical fiber consist of measuring the power loss, using power detectors. Another important technique is based on optical time-domain reflectometry (OTDR). OTDR measuring apparatus are compact and versatile devices, easily transported to the field, and have the capability to measure very long lengths of optical fiber. Attenuation can be measured for a specific operating wavelength of the optical fiber, as in OTDRs, or in a broader wavelength range, called *spectral attenuation*.

Figure 1 shows the schematics of an apparatus to measure the optical fiber spectral attenuation by power

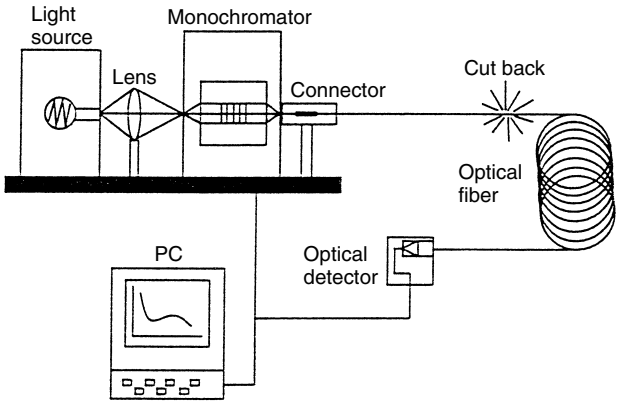


Figure 1. Spectral attenuation apparatus schematics.

measurement. In this test the wavelength is selected using a computer-controlled monochromator, and the light source consists of a white-colored, powerful device. After the monochromator the monochromatic selected light is injected into the optical fiber extremity. This extremity of the optical fiber must be cleaned and cut carefully in order to guarantee that the maximum light is injected in the optical fiber core. The other extremity of the optical fiber, clean and with a good quality cut, is in contact with the power detector. For silica glass optical fibers, the first power measurement series is done with the monochromator scanning at each few nanometers in the range 600–1600 nm. After the first power scanning, the optical fiber is cut at approximately 2 m from the launching extremity and inserted again in the power detector. Another power measurement is performed as described above, and the attenuation at each wavelength is the quotient between the second power measurement and the first power measurement divided by the length of the optical fiber. It is convenient to use the common logarithm of the power quotient to express the attenuation in decibels per kilometer (dB/km).

Because the cut is made close to the power source, this technique is also known as the *cutback method*. This method is precise and reliable and is used as a reference for other attenuation measurements methods.

In single-mode optical fibers the loss of light radiation that affects the intensity and spreading of the lightpulse can be caused by other phenomena. Light sources, such as lasers and light-emitting diodes (LEDs), are designed to emit a specific wavelength, but they really emit light in a wavelength range. A lightpulse generated by those sources has many components with different wavelengths. In the core of the optical fiber different wavelengths travel at different velocities causing a time spreading of the pulse. The phenomenon in which the velocity of propagation of an electromagnetic wave is wavelength-dependent is called *dispersion*, in which different wavelengths are connected with different colors. This phenomenon is thus called *chromatic dispersion*.

At least two methods are used to measure chromatic dispersion: the spectral group delay method and the phase-shift method [6,7]. These methods measure the time dispersion of the lightpulse that occurs after the pulse has traveled the length of the optical fiber.

The optical fiber core does not have a perfectly symmetric cylindrical shape; the purpose of this design is to avoid *polarization mode dispersion* (PMD). In practice, the core diameter and shape vary slightly in a random fashion during the optical fiber manufacturing process. In PMD, internal stresses induced by thermal expansion and external forces induced by the environment through handling and cabling adds more stress fields inside the optical fiber core. Those perturbations induce the two orthogonally polarized modes that travel at different group velocities in a single-mode optical fiber, and the lightpulse is broadened and distorted. At frequency transmissions above 10 gigabits per second (Gb/ps), PMD is a limiting factor for lightwave transmission in optical fiber systems.

Nowadays systems operating at very high transmission rates are commonly installed in practice. The PMD limitation has motivated many efforts to understand and quantify the phenomenon. Since the early 1990s, at least six methods have been proposed to measure PMD; they are divided in two groups: (1) methods that involve performing measurements in the time domain and (2) methods involving measurements in the frequency domain. In the time domain, three methods are used: the modulation phase-shift method, the pulse delay method, and the interferometric method. In the frequency domain the methods are the fixed analyzer method; the Poincare arc method, which is also called the *Muller matrix* method, and the Jones matrix method.

The fixed analyzer method is the most commonly used method [8] (Fig. 2). In this method polarized light is injected into an optical fiber and then the optical power transmitted through the optical fiber and then through a polarizer, as a function of the wavelength, is measured. As the wavelength is changed, the power transmitted through the polarizer goes up and down. By counting the number of maximums and minimums or counting the number of zero crossings it is possible to determine the average PMD. When these zero crossings are counted, the test actually consists in measuring the rate at which the output state of polarization changes with the wavelength.

It is possible to use an OTDR to measure the PMD in optical fibers by interferometric methods [9].

In multimode optical fibers a wide range of wavelengths constitutes the lightpulse and hundreds of modes travel together. The electromagnetic interference of those mode components reduces the power intensity being transmitted and, because the modes all travel at different velocities, the pulse spreads in time. If the transmission rate of pulses being transmitted increases, the pulse spreading causes adjacent pulses to overlap in time, resulting in interpulse interference and errors in signal detection. The maximum frequency at which it is still possible to recover the information is called the *bandwidth* of the multimode optical fiber. Measurement of this parameter is done in two domains; (1) the time domain, where the pulse spreading time is measured; and (2) the frequency domain, where the maximum frequency is measured according to the power loss in system detection [10].

2.3. Mechanical Characterization

Ensuring adequate integrity of optical fibers is complex because of the wide range conditions of temperature, humidity, and mechanical stresses that are present in

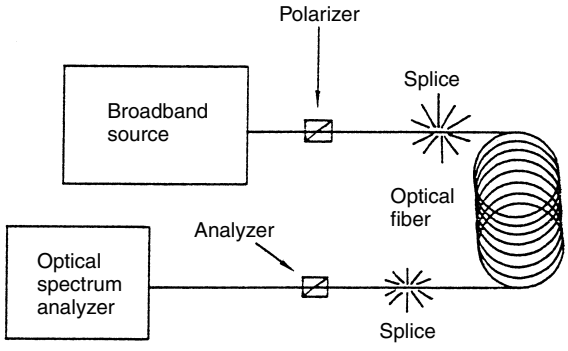


Figure 2. PMD measurement using the fixed analyzer method.

the manufacturing and in the optical fiber field. It is safe to state that the only way to determine the effectiveness of the optical fiber mechanical quality and how long it survives in the presence of moisture in the field is to wait until it breaks. But the lifetime can be very long, and predictability is needed. A short-term mechanical characterization can be useful to qualify the optical fiber, and a strength degradation model is necessary in order to predict failures.

After the manufacturing process it is necessary to embed the optical fibers in stronger structures in order to be installed and reliably operated during a minimum calculated period of time. These structures are called *optical fiber cables*.

Currently hundreds of optical fiber cable structures have been proposed for a large variety of applications, including submarine transoceanic cables, aerial cables, and underground cables. A common optical fiber cable design consists of a few basic elements. A strong tensile member of stranded steel wires, which are used to pull the cable during the installation process; a multilayer polymeric and metallic cover used to protect the optical fiber from contamination due to chemicals from the surrounding environment; and polymeric tubes with optical fibers inside it.

The silica glass chemically reacts with water, causing mechanical degradation of the glass optical fiber in a well-known phenomenon studied since the 1950s [11]. In the presence of stress and humidity a glass surface flaw can accelerate its growth and subsequently cause a fracture. Mechanical laboratory tests have shown that in a free water environment, as in tests performed with the sample immersed in liquid nitrogen, the strength of the optical fiber is at least 3 times higher compared with tests in a humid environment. Some authors found that this water glass corrosion is still possible in a stress-free situation for optical fibers coated with different polymeric materials [12,13].

Flaws in the surface of the glass are classified in two groups. The first group, *extrinsic defects*, represents a serious danger to the optical fiber. These flaws are many micrometers in length and in general originated during the manufacturing process, or were introduced in the glass surface by mechanical abrasion and handling. These defects can usually be identified after the optical fiber breaks, using microscopic techniques to analyze the broken surface [14]. The second group of flaws, called *intrinsic defects*, whose lengths are of the order of the silica glass structure (a few nanometers), cannot be observed using conventional microscopes, but studies using atomic force microscopy have revealed how they can affect the strength degradation process [15].

To mechanically qualify the optical fiber immediately after the manufacturing process, it is possible to do an optical fiber length tensile test in a proof testing machine [16]. The objective of this test is to eliminate large flaws that cause breaks during the optical cable manufacturing process, or in the cable lifetime. The test consists in applying a controlled tensile stress to the entire length of the optical fiber, which is done by using a system with pulleys and belts driven by electric motors (Fig. 3). In this test, when a large flaw is present in the length of the optical fiber that passes through the tensile proof test region, the applied stress activates the crack growth, causing a break in the fiber. The minimum stress level guaranteed for the survival of the optical fiber pieces after this destructive test is a function of the constant stress used in the tensile region, of the crack growth during the unloading time, and of the functional characteristics of the machine, such as the velocity of the fiber passing through the system.

The complete analysis of the proof testing must take into account the additional crack growth during the applied proof stress. Some of these flaws can be taken to their critical fracture size during the test causing breaks, and others can be very close to the critical size after the test. The proof test must be performed at the highest velocity of the fiber passing through the machine, in order to minimize the mechanical strength degradation.

After the proof stress area, the optical fiber is unloaded in the last pulley and an additional crack growth occurs, and this additional flaw growth during the unloading time

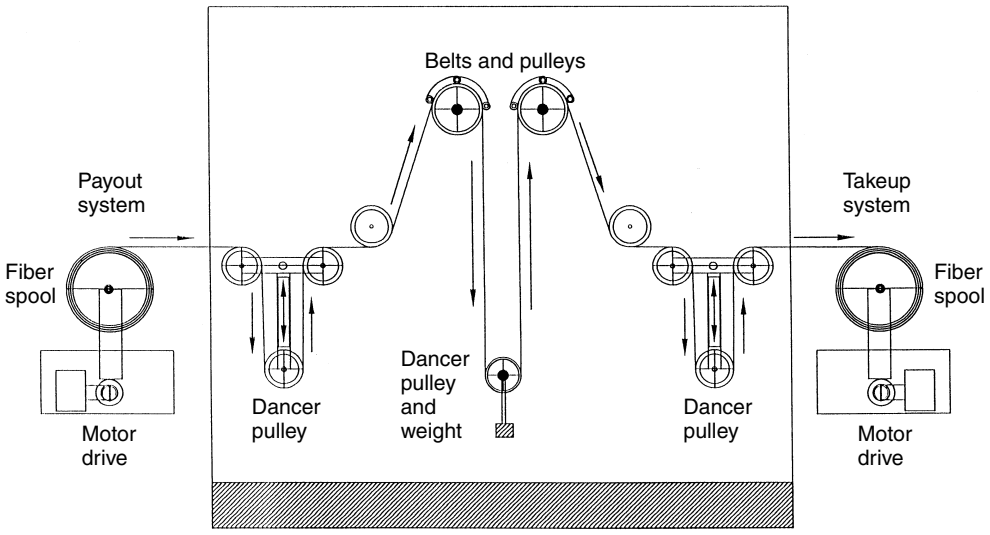


Figure 3. Optical fiber proof test machine.

reduces the guaranteed minimum strength to a level that is under the stress level applied to the tensile region. It is possible to calculate the minimum guaranteed strength by using the crack growth theory, the proof test specifications, and the velocity of the optical fiber passing through the system. The minimum strength after the proof test is a very useful parameter in order to guarantee a reliable cable manufacturing process, the optical fiber integrity during the installation of the cable, and the use of the cable.

After the proof test, additional mechanical characterization consists of measuring the strength of the optical fibers under tensile or bending stresses and their susceptibility to environmental changes. These tests, applied to a few samples removed from one extremity of the optical fiber, consist of applying a crescent force until the break. By assuming that the maximum force is connected with the geometric and physical parameters of the optical fiber, such as its diameter and length, and the elastic modulus of the silica glass, it is possible to calculate the maximum stress in the moment of the fracture. This is called the *strength* of the optical fiber. The results for these few samples removed from one extremity of the fiber are extrapolated to the entire length, assuming that the optical fiber presents the same mechanical characteristics in its entire length. The flaws distributed in the glass surface length can be compared to weak links in a chain, and a specific fracture probability distribution can be developed to describe the fracture event. This distribution is called the *Weibull fracture probability distribution* [17]. The Weibull model is applied to the strength of the tested samples from which the mean strength and the variability of the strength, referred as the m Weibull parameter, are calculated. These parameters are very useful for comparison of distinct optical fiber mechanical qualities.

In the past, mechanical apparatus used to test metal wires and pieces of plastics were adapted to test the tensile strength of pieces of optical fibers. Currently, mechanical apparatus have been developed specifically to test multiple and longer samples simultaneously. Laboratory tensile apparatus normally use a few meters of optical fibers, around 24 samples of 0.5 m each, tested one by one, to characterize many kilometers of an optical fiber (Fig. 4). Once the extremities of the optical fiber sample are securely held, a crescent force is applied until it breaks; this is known as the *dynamic fatigue test*. It is necessary to apply a high intensity force to hold the extremities of an optical fiber sample until it breaks. To avoid fractures to the extremities held by grip devices, it is common to wrap two or three turns around cylindrical pieces of metal, called *mandrels*. The force is applied by pulling the mandrels mechanically attached to the tensile machine. To guarantee no slippage of the optical fiber sample in the mandrel surface, it is necessary to use a double-face tape in the surface of the mandrel. Today this is the most commonly used method.

The problem with this approach is that it is not possible to know exactly what length of fiber is being tested, and part of the force is absorbed by the piece of optical fiber glued to the mandrel surface. In order to do fracture probability extrapolation to the untested piece of optical fiber using the Weibull model, it is essential to know the tested

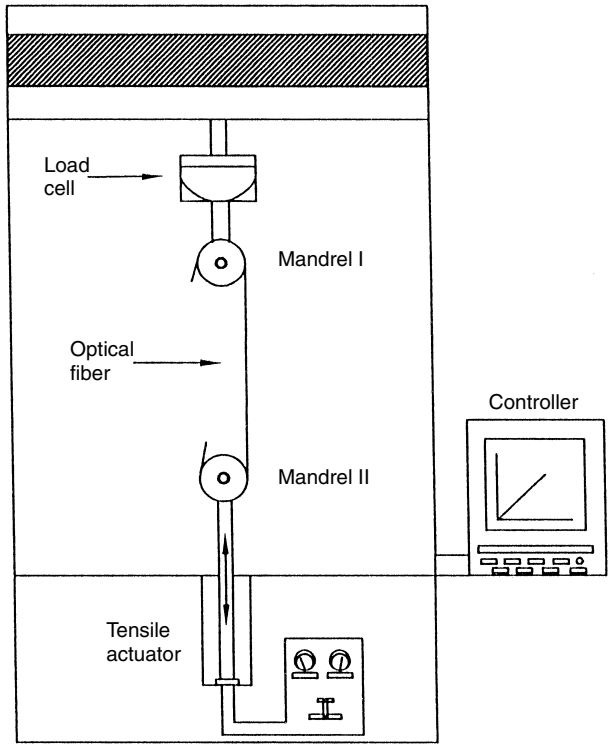


Figure 4. Optical fiber tensile tester.

length. To minimize this problem, one possibility is to test long-length samples of ($\geq$10 m). Another advantage with long length tests is that it gives more knowledge about the extrinsic defects present in the optical fiber [18].

Bending tests are performed in very small pieces of optical fiber samples, around 5 cm in length. In this test an optical fiber sample is introduced between two grooved steel plates. When one plate is pushed against the other, the curvature radius decreases until the optical fiber breaks. This is called the *bending dynamic fatigue test* (Fig. 5). By measuring the optical fiber curvature radius in the moment of the fracture, it is possible to know the maximum bending stress at the fracture (bending strength). This test can be performed on one by one sample or in multigrooved plates, up to 24 samples. Acoustic detectors can be used to precisely measure the distance

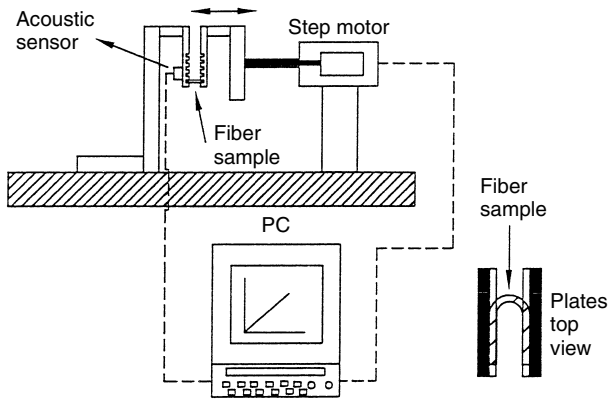


Figure 5. Two-point bending tester with acoustic detector.

of the plates that is connected with the curvature radius during the breaks [19].

The maximum bending stress is applied to a few square micrometers of the glass surface. The probability of finding an extrinsic defect in such areas is small, assuming that the optical fiber is of reasonable quality. This test is intended to measure the intrinsic strength of the glass surface. The test can be used to check how much the coating covers the glass, or the glass corrosion when the optical fiber is under chemical attack in harsh environments.

Another test in bending consists of holding the steel plates at a calculated distance, maintaining the bending stress constant, until the samples break. This is called a *static fatigue in bending*. In such a case, the survival time of the optical fiber under bending stress is measured. This bending test can be easily performed in harsh environments but cannot be done so easily under tension. In order to perform the static fatigue test for tensile, strength it is necessary to have a lot of space in order to accommodate the samples and the expensive equipment that is required to maintain constant environmental conditions in the area of the test.

The dimensions of the optical fiber core and its cladding refractive-index contrast confine the light being transmitted, but slight perturbations in geometry can cause a loss of power. Small mechanical forces acting on the surface of the optical fiber coating can be mechanically transmitted to the core, causing small perturbations in diameter. These perturbations are of the order of the fiber core diameter or less. Pressing a piece of fiber in a rough surface can be sufficient to change the optical fiber attenuation. This phenomenon is called *microbending sensitivity*. In cable design, this parameter must be minimized, taking in to account the maximum and minimum temperature fluctuations that can affect the format of the fiber inside the cable, the number of optical fibers in contact with each other, and the compatibility of the optical fiber with other cable structures. In order to measure the microbending sensitivity, attenuation measurements are used during susceptibility tests, such as temperature fluctuations and mechanical tests, with optical fiber cables.

It is still possible to measure the microbending sensitivity of a piece of optical fiber by monitoring the attenuation in a test where the optical fiber is compressed at known loads between two rough sandpaper sheets. This it can be done for a specific light wavelength using a power detector or for a broader wavelength range, using a spectral attenuation apparatus. Actually there does not exists standard procedure for this test.

When a piece of optical fiber is bent in a radius larger than that of the fiber, a small part of the light radiation can propagate through the core and leak to the cladding, causing the light to be lost. This phenomenon is called *macrobending attenuation*. Using metallic mandrels at different diameters and wrapping the same length of fiber around them, it is possible to measure relatively accurately the relationship between the fiber bent diameter and the optical attenuation. This measurement can be done for a specific wavelength or for a wide wavelength range, using a spectral attenuation apparatus. This is useful information in cable design and installation of optical fiber systems.

Silica glass is susceptible to environmental humidity, and polymer-coated optical fibers reach equilibrium with the environment in less than an hour. The plastic polymer coating works as a net on the surface of the glass, as it is permeable to the water molecules. Environmental humidity fluctuations, around 10% relative humidity (RH), can affect the optical fiber mechanical tests. Other liquids and gases can affect the strength performance of the fiber. Basic substances are more supportive of glass corrosion. All mechanical tests must be performed in a humidity-controlled environment, after the optical fiber is in equilibrium with the environment. Environmental recommendations are present in all standardized mechanical tests.

The glass susceptibility to the environment can be measured by performing strength tests at different stress rates. Stress rates are related to the force variation in time applied to a sample. Using different stress rates, it is possible to isolate the strength degradation caused by the presence of humidity from the strength degradation caused by the stress alone. The parameter that describes how rapidly the strength degradation occurs in the presence of humidity is called *optical fiber fatigue n* [20]. This parameter is obtained using at least four groups of samples tested at four different stress rates. In terms of fiber reliability, this is the most important parameter. The n number is the power that will be used to calculate the optical fiber lifetime.

An important aspect of the optical fiber mechanical degradation is related to the survival time of the optical fiber in the field. This is called the *optical fiber reliability* [21]. For this calculation it is necessary to know well the *fiber stress history* (FSH), which *FSH* is the accumulation of all the stresses applied to the optical fiber during the different transmission system construction and use: optical fiber manufacturing, proof testing, cabling, installation, and application. The FSH parameters using time fracture probability to estimate the fiber lifetime after installation are the minimum guaranteed strength after the proof test, as described above; the optical fiber fatigue parameter n; and the low−stress break flaw distribution, measured using the tensile test For long lengths of the proof tested optical fibers. To complete the necessary calculations, it is necessary to use a crack growth model when the optical fiber is under low stress in the presence of humidity. Normally, a *power-law crack growth model* is used, which assumes a power relation between the crack growth velocity, the parameters of the material, and the applied stress. The fatigue parameter n is the power variable in this model. Other crack growth models were proposed in the last decade but the power law is the most reliable and treatable model. The power-law model is used in national and international standards.

To complete the mechanical characterization, it is necessary to measure the qualities of the optical fiber coatings. The most important characteristics of the fiber coatings are (1) the primary coating must have a good chemical reactivity with the glass surface, to protect the glass from moisture; (2) the primary polymer coating must have lower elastic modulus, to absorb external small stress that causes attenuation by microbending sensitivity; and

(3) the secondary polymer coating must have a higher elasticity modulus to improve the resistance to the abrasion.

It is possible to join the extremities of optical fibers by keeping the attenuation and the mechanical qualities under control. This operation is called *optical fiber splice*. Splices are common in optical fiber cable installation and in optical fiber cable repairs. An important step in the splice procedure consists in removing the coating of the fiber extremities. The glass must be completely clean and well cleaved to perform the thermal fusion of the extremities. If the primary coating is overconnected with the glass, the operation will be difficult and can affect the splice quality. If the primary polymer coating is loose on the surface of the fiber, or not completely cured, it does not promote the necessary protection [22].

Standard procedures to measure coating quality and how it affects splice performance are under study by standards committees and include methods to measure the simplicity of removing the coating of the optical fiber, strip-force and pullout force techniques, and methods to measure the elasticity modulus of the primary coating.

BIOGRAPHY

Gilberto Camilo received his D.Sc. degree from Campinas University, Sao Paulo, Brazil, in 1991. He was employed by PIRELLI Optical Fibers, Sao Paulo, Brazil, between 1987 and 1991. He taught physics and mechanical engineering at Goias and CEFET-Parana Federal University, Brazil, during 1991–1998. He was a Post-Doctoral Fellow at Rutgers University, and worked on the Ceramic and Materials Engineering — Optical Fibers Project during 1994–1995. He was with Furukawa Optical Fiber Cables, Parana State, Brazil, during 1997–1998 and with Alcatel, in charge of Optical Fibers Mechanical Reliability at Claremont, North Carolina, during 1998–2001. Since 2001 he has been with OmniGuide Communications, at Cambridge, Massachusetts, as Optical Fiber Reliability Specialist. He has been active in the TIA/EIA Optical Fibers Standard Committees since 1998. He is a member of the OSA, IEEE, and SPIE. Dr. Camilo has published over 50 papers in the area of optical fiber mechanical Characterization. His main interest areas are optical and mechanical characterization and reliability of optical fibers and optical cables.

BIBLIOGRAPHY

1. T. Kaino, Plastic optical fibers, *Proc. SPIE* **CR63**: 164–187 (1996).
2. J. A. Harrington, Infrared optical fibers, in *Handbook of Optics*, Optical Society of America, McGraw-Hill, 2001.
3. S. G. Johnson et al., Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers, *Optics Express* **9**: 748–779 (2001).
4. W. J. Stewart, A new technique for measuring the refractive index profiles of graded optical fibers, *Proc. IOOC Tech. Digest* 395–398 (1977).
5. R. Girbig and M. Hoffart, Highly accurate backscatter measurement in the quality control of the cabling of single-mode fibers, *Proc. IWCS* **38**: 480–485 (1989).
6. A. J. Barlow, R. S. Jones, and K. W. Forsyth, Technique for direct measurement of single-mode fiber chromatic dispersion, *J. Lightwave Technol.* **LT-5**: 1207–1213 (1987).
7. B. Costa, M. Puleo, and E. Vezzoni, Phase-shift technique for the measurement of chromatic dispersion single-mode fibers using LED's, *Electron. Lett.* **19**: 1074–1076 (1983).
8. C. D. Poole and D. L. Favin, Polarization-mode dispersion measurements based on transmission spectra through a polarizer, *J. Lightwave Technol.* **LT-12**: 917–922 1994.
9. A. J. Rogers, Polarization-optical time domain reflectometry: A technique for the measurement of field distributions, *Appl. Opt.* **20**: 1060–1074 (1981).
10. A. H. Hartog et al., Comparison of measured and predicted bandwidth of graded-index multimode fibers, *J. Lightwave Technol.* **QE-18**: 825–838 (1982).
11. R. J. Charles, Static fatigue of glass I and II, *J. App. Phys.* **29**: 1549–1560 (1958).
12. N. Evanno, M. Poulain, and A. Gouronnec, Optical fiber lifetime in harsh conditions, *Proc. SPIE* **3848**: 70–76 (1999).
13. P. Regio, P. Motta, and S. Apone, Influence of the coating in mechanical behavior of aged optical fiber, *Proc. Eurocable 97*, 1997.
14. L. K. Baker and G. S. Glaesemann, Break source analysis: alternate mirror measurement method, *Proc. IWCS* **47**: 933–937 (1998).
15. G. Camilo, C. Turnbill, and B. Overton, Glass corrosion in commercial optical fibers with defective coatings, *Proc. IWCS* (in press).
16. T. A. Hanson, Analysis of the proof test with power law assumptions, *Proc. SPIE* **2074**: 108–119 (1994).
17. W. Weibull, A statistical theory of the strength of materials, *Proc. Royal Swed. Inst. Eng. Res.* **151**: 1–45 (1939).
18. W. Griffioen, Mechanical lifetime of optical fibers, *Proc. European Fibre Optic Communications and Networks*, 1994, pp. 164–168.
19. M. J. Matthewson, C. R. Kurkjian, and S. T. Gulati, Strength measurement of optical fiber by bending, *J. Am. Cer. Soc.* **69**: 815–821 (1986).
20. V. V. Rondinella and M. J. Matthewson, Ionic effects on silica optical fiber strength and models for fatigue, *Proc. SPIE* **1366**: 1–8 (1990).
21. W. Griffioen, *Optical Fiber Mechanical Reliability*, Eindhoven Univ. Technology, 1994.
22. G. Camilo and B. Overton, Evolution of fiber strength after draw, *Proc. NFOEC* **17**: 143–153 (2001).

CHIRP MODULATION

DIRK DAHLHAUS
Communication Technology Laboratory
Zurich, Switzerland

1. INTRODUCTION

Chirp modulation (CM) represents a special type of spread-spectrum signaling where a carrier signal is modulated in two ways. The primary modulation is carried

out in the complex baseband and constitutes the usual formats such as phase shift keying (PSK), pulse position modulation (PPM), or binary orthogonal keying (BOK). The primary modulation is combined with a secondary modulation for spectrum spreading. For a data rate T^{-1} with T denoting the symbol duration, the occupied Fourier bandwidth B exceeds T^{-1} considerably; that is, in general, the time-bandwidth product $TB \gg 1$. The spreading is advantageous in frequency-selective fading channels often encountered in wireless or mobile radio systems. If the occupied spectrum is larger than the coherence bandwidth of the channel, the transmission is more robust against the fading because of the resulting frequency diversity. In addition, as shown by Berni and Gregg [1], CM is resistant to the Doppler effect arising in time-variant scenarios typically encountered in mobile radio applications. CM signals have been first proposed by Winkler [2] for their high robustness against distortions and different types of interference.

In most cases, "chirp modulation" refers to a sinusoidal signal of duration T whose instantaneous frequency changes linearly in time t between the lower frequency $f_1 = f_0 - B/2$ and the upper frequency $f_2 = f_0 + B/2$, where f_0 denotes the carrier frequency of the signal. In its simplest form, CM is used in combination with binary signaling as primary modulation. To transmit a logical 0 using a binary CM signal, an "upchirp" is used, which corresponds to a linear frequency sweep from f_1 to f_2. A logical 1 is transmitted correspondingly as a "downchirp," a linear frequency sweep from f_2 to f_1. For sufficiently large values of the time-bandwidth product, the upchirp and downchirp signals transmitted in a common band constitute the aforementioned BOK. The name *chirp* has been given to such signals by Bell Telephone Laboratories because of the resemblance to a sound heard in nature [2]. CM signals have their roots in radar applications where one of the most important observations states that range resolution and accuracy are functions of the signal bandwidth, and not of the transmitted pulsewidth [3].

In Section 2, important properties of CM signals are discussed including the form of linear frequency-modulated (FM) signals, the signal spectrum, the matched-filter (MF) characteristics, measures for sidelobe reduction, and modulation schemes. In Section 3, the performance of the different schemes is analyzed. Section 4 describes different ways to implement CM systems including surface acoustic wave (SAW) devices as well as digital baseband techniques. Other aspects related to CM in communication systems are discussed in Section 5.

2. PROPERTIES OF CM SIGNALS

2.1. Time and Frequency Representation

The general form of FM bandpass signals to be considered is given by

$$s(t) = a(t)\cos(\omega_0 t + \theta(t)), \qquad -\frac{T}{2} < t < \frac{T}{2} \tag{1}$$

where $f_0 = \omega_0/2\pi$ and $\theta(t)$ denote the carrier frequency and the signal phase, respectively. The envelope $a(t)$ can be used as a weighting function to improve the autocorrelation properties of $s(t)$ as discussed in Section 2.3. Here, it is first assumed a rectangular function over the interval $[-T/2, T/2]$. The instantaneous frequency is defined by

$$f(t) = \frac{1}{2\pi}\left(\frac{\omega_0 + d\theta}{dt}\right) \tag{2}$$

For linear FM signals, we obtain

$$f(t) = f_0 + \mu t, \qquad -\frac{T}{2} < t < \frac{T}{2} \tag{3}$$

with $\mu \in \mathcal{R}$ denoting the dispersive slope or rate of the chirp signal. From (3), $f(t)$ varies between the lower (resp. upper) frequencies

$$f_1 = f_0 - |\mu|\frac{T}{2}$$

$$f_2 = f_0 + |\mu|\frac{T}{2}$$

over a range $B = |\mu|T$[1], where $\mu > 0$ and $\mu < 0$ denote an upchirp and (resp. downchirp) signal, as shown in Fig. 1. The signal phase results in

$$\theta(t) = \pi\mu t^2 + \theta_0, \qquad -\frac{T}{2} < t < \frac{T}{2} \tag{4}$$

with a suitable initial phase value θ_0. For the usually valid narrowband assumption $f_0 \gg |\mu|T$ [3] and $\theta_0 = 0$, the spectrum of an upchirp $S(\omega) = \int_{\mathcal{R}} s(t)\exp[-\jmath\omega t]\,d\omega$ of $s(t)$ is given by [3]

$$S(\omega) = \frac{1}{2\sqrt{2\mu}}\exp\left[-\jmath\frac{(\omega-\omega_0)^2}{4\pi\mu}\right] \times \left[\mathcal{C}(X_+) + \jmath\mathcal{S}(X_+) + \mathcal{C}(X_-) + \jmath\mathcal{S}(X_-)\right]$$

where

$$\mathcal{C}(X) = \int_0^X \cos\left(\frac{\pi y^2}{2}\right)dy, \qquad \mathcal{S}(X) = \int_0^X \sin\left(\frac{\pi y^2}{2}\right)dy$$

are Fresnel integrals and the integral limits are given by

$$X_\pm = \frac{\pi\mu T \pm (\omega - \omega_0)}{\pi\sqrt{2\mu}}$$

On substituting

$$\mu = \frac{B}{T}, \qquad \omega - \omega_0 = n\pi B$$

with the normalized frequency n, we obtain

$$X_\pm = \frac{1 \pm n}{\sqrt{2}}\sqrt{TB}$$

[1] The bandwidth occupied by $s(t)$ is larger than B, but approaches B for $TB \to \infty$.

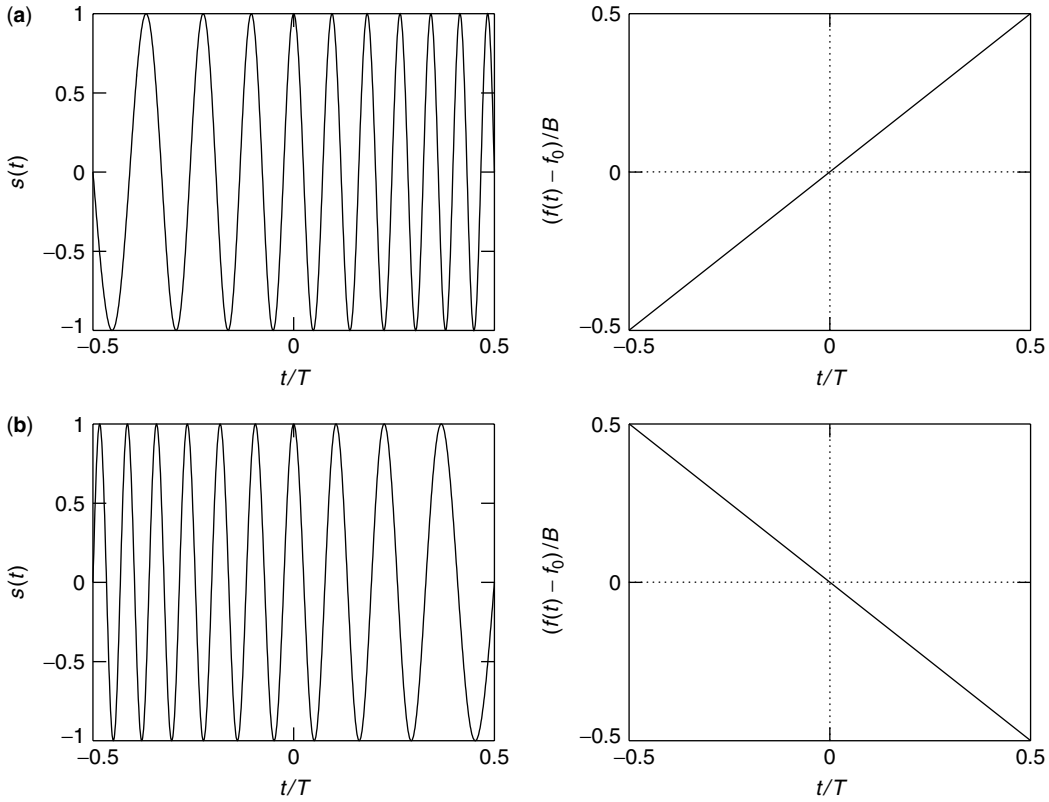


Figure 1. Chirp signals and corresponding instantaneous frequencies: (**a**) upchirp; (**b**) downchirp.

specifically, the amplitude spectrum of $s(t)$

$$|S(\omega)| = \frac{1}{2\sqrt{2\mu}}\sqrt{\left[\mathcal{C}(X_+) + \mathcal{C}(X_-)\right]^2 + \left[\mathcal{S}(X_+) + \mathcal{S}(X_-)\right]^2} \tag{5}$$

in Fig. 2 depends solely on the time–bandwidth product TB. The spectrum characteristics in the band center are mainly determined by the so-called Fresnel ripples arising from the Fresnel integrals. The height of the ripples increase for decreasing TB while the spectrum is asymptotically rectangular for $TB \to \infty$. It can be shown that the amplitude spectrum in (5) is valid also for a downchirp if μ is replaced by $|\mu|$ in all terms containing the dispersive slope variable. More information on the phase spectrum for upchirp and downchirp signals can be found in the treatise by Cook and Bernfeld [3]. For $TB \gg 1$, the described binary modulation is sometimes termed BOK since the normalized cross-correlation of upchirp and downchirp signals is almost zero. The correlation properties of the chirp signals are considered in the next section.

2.2. Matched Filtering

The MF providing a sufficient statistic for symbol detection in additive white Gaussian noise (AWGN) and maximizing the signal-to-noise ratio (SNR) for a perfectly synchronized receiver is termed a *compression filter* in linear FM systems. For an upchirp signal in (1), the impulse response of the MF is a downchirp given by

$$h(t) = ks(-t) = k\cos\left(\omega_0 t - \pi\mu t^2\right), \qquad -\frac{T}{2} < t < \frac{T}{2}$$

where $k = 2\sqrt{\mu}$ is chosen for a unity gain of the MF at $f = f_0$. Here, the MF output is considered for a channel with a Doppler shift f_D. This situation studied extensively in radar applications [3] arises in mobile communications, such as in time-varying line-of-sight channels. The MF output signal in the absence of thermal receiver noise is given by

$$\begin{aligned} g(t, f_D) &= \int_{\mathcal{R}} s(\tau)\, h(t-\tau)\, d\tau \\ &= 2\sqrt{\mu}\int_a^b \cos\left[(\omega_0 + 2\pi f_D)\,\tau + \pi\mu\tau^2\right] \\ &\quad \times \cos\left[\omega_0(t-\tau) - \pi\mu(t-\tau)^2\right] d\tau \\ &= \begin{cases} \dfrac{\sqrt{\mu}}{\pi}\,\dfrac{\sin\left(\pi\left(f_D + \mu t\right)\left(T - |t|\right)\right)}{f_D + \mu t} \\ \quad \times \cos\left(2\pi\left(\dfrac{f_0 + f_D}{2}\right)t\right), & -T < t < T \\ 0 & |t| \geq T \end{cases} \end{aligned} \tag{6}$$

with

$$a = -\frac{T}{2} + t, \quad b = \frac{T}{2} \qquad \text{for } t \geq 0$$

$$a = -\frac{T}{2}, \qquad b = \frac{T}{2} + t \quad \text{for } t < 0$$

The frequency shift of $f_D/2$ in (6) can be easily understood from considering the spectra of the input and MF signals [3]. Figure 3 shows the envelope of $g(t) =$

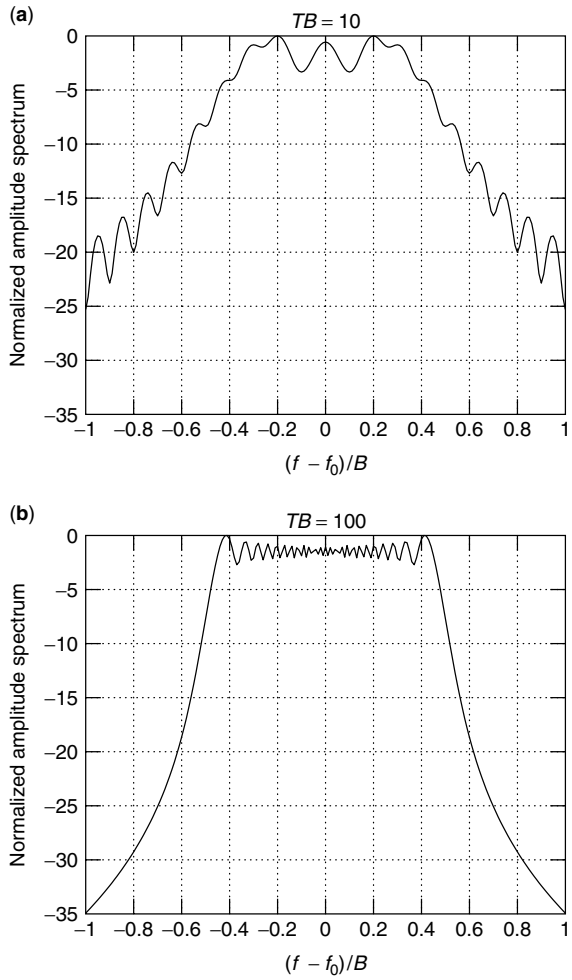


Figure 2. Amplitude spectra for different time–bandwidth product values: (**a**) $TB = 10$; (**b**) $TB = 100$.

$g(t, f_D = 0)$ for different values of TB. As can be concluded from (6), the height of the correlation peak is $T\sqrt{\mu} = \sqrt{TB}$. Thus, the *compression gain*, defined as the ratio of the peak value to the chirp amplitude, is $G = 10\log(TB)$ dB. For $TB \gg 1$, it can be shown that the width of the mainlobe is $2/B$ while the minimum sidelobe suppression, expressed by the ratio G_{SL} of the mainlobe and the first sidelobe values, is approximately 13.3 dB. This ratio essentially determines the system robustness against intersymbol interference (ISI) in frequency-selective fading channels arising from multipath propagation of the transmitted signal. Measures for reducing the sidelobes are discussed in the next section. Figure 4 shows the shape of $g(t, f_D)$ for different values f_D/B. Obviously, the correlation peak is shifted, attenuated, and spread as compared to Fig. 3. Proceeding as in (6), it is readily shown that the cross-correlation for $f_D = 0$ and $t = 0$ between upchirp and downchirp equals $(C\sqrt{TB})$, which approaches $\frac{1}{2}$ for large TB. In this case, the ratio of the cross-correlation and autocorrelation equals $1/(2\sqrt{TB})$ which justifies the assumption of BOK for large TB.

2.3. Sidelobe Reduction

One way to increase the dynamic range of the pulse compression is to modify the signal spectrum of $g(t)$

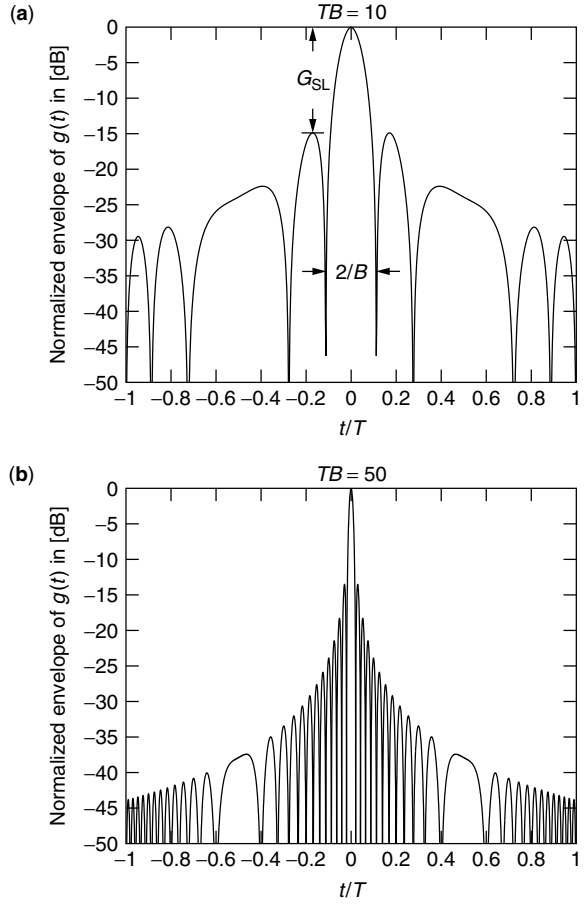


Figure 3. Normalized envelope of the MF output signal $g(t)$ for different time–bandwidth product values: (**a**) $TB = 10$ (**b**) $TB = 50$.

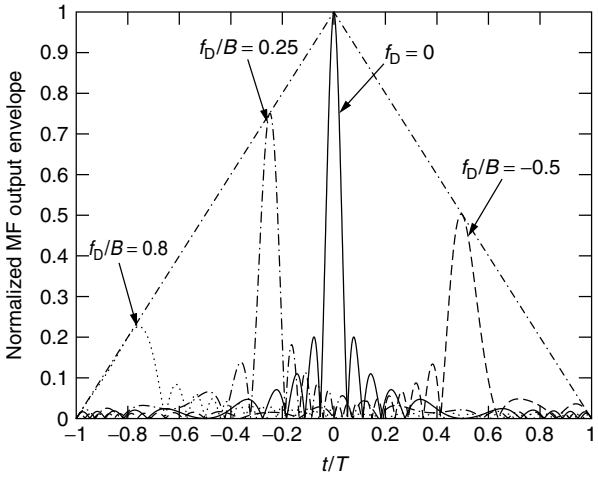


Figure 4. MF output signal $g(t, f_D)$ for different values f_D/B and $TB = 20$.

for decreasing the sidelobe levels. The corresponding weighting can be implemented in the frequency or time domain. The optimum distribution function can be derived from an analogous problem in antenna theory treated by Dolph [4] and Van der Mass [5] that targets the narrowest beamwidth of a broadside antenna array for a desired

sidelobe level. However, the resulting Dolph–Chebyshev weighting function has infinite power and is thus physically not realizable. Approximations to the optimum solution are provided by the Taylor functions and modified Taylor functions [3]. The latter contain the so-called Hamming function

$$W_H(f) = 0.08 + 0.92\cos^2\left(\pi\frac{f-f_0}{B}\right), \qquad |f-f_0| \le \frac{B}{2}$$

as a special case. It can be shown that this weighting results in a sidelobe suppression of −42.8 dB, a spreading of the mainlobe by a factor of 1.47 and a loss in compression gain of 1.34 dB. The influence of the Fresnel ripples on Hamming weighted chirp compression is treated by Kowatsch and others [6,7] who considered both time and frequency weighting as well as Doppler shifts. A considerable reduction of the Fresnel ripples can be achieved by using a Tukey window, where the undesirable mainlobe width increase is only moderate [6].

2.4. Modulation Schemes

Although the spectrum spreading in CM can in principle be combined with any baseband signaling, the choice of the primary modulation in a practical system is restricted by, For example, performance requirements, an efficient implementation, and component imperfections. Here, some modulation schemes are described that can be used in CM and are discussed in Section 3 in terms of the achievable bit error rate (BER). As in Proakis' text [8], linear modulation is distinguished from nonlinear modulation with memory.

2.4.1. Linear Modulation. The chirp pulse in (1) can be implemented efficiently by a SAW device (see Section 4). The properties of this technology rule out certain modulation schemes, such as amplitude shift keying (ASK). The problem with the latter scheme is the high dependence of the output power on the rising time of the broadband pulse exciting the filter. This problem is usually circumvented if PSK, BOK, or PPM is employed [9,10]. El-Khamy and Shaaban [11], match μ to the dispersion parameters of the communication channel. They show that for a channel with a second-order polynomial phase spectrum and a partially coherent detection, μ should be chosen according to $TB = 2.65$, which minimizes the BER. The aforementioned BOK requires upchirp and downchirp filters at both the transmitter and the receiver. Clearly, the bit error performance depends critically on the sidelobe level and the cross-correlation properties of the employed chirp signals as well as the delay dispersion of the channel. Some of the problems can be solved if PPM signals are employed using only one chirp signal type (e.g., an upchirp). In PPM [9], the binary signals are orthogonal, the cross-correlation problem does not arise, and only one chirp filter has to be implemented in the transmitter and the receiver. In case of a logical 1, the chirp is sent Δt before the reference clock, while for a logical 0, the chirp is delayed by Δt. The system performance is determined by the value Δt and the channel delay dispersion. Another standard modulation method is differentially encoded quaternary PSK (DQPSK), which allows for a differential demodulation without carrier phase estimation. Usually, $\pi/4$ DQPSK offering reduced envelope fluctuations as compared to ordinary DQPSK is employed. Again, the performance limiting factors are the sidelobe levels and the channel delay dispersion. In general, the data rate can be increased by applying overlapping signal pulses that can be resolved at the receiver for sufficient sidelobe reduction capabilities of the employed MF output signals.

2.4.2. Nonlinear Modulation with Memory. CM for binary signaling Hirt and Pasupathy combined with full-response (FR) continuous phase modulation and termed FR *continuous-phase chirp modulation* (FR-CPCM). The idea of FR-CPCM is the improvement of the independent bit-by-bit detection in conventional CM systems by observing the phase-constrained received signal over two or more bit intervals prior to bit detection. In CPCM, instead of (4), the transmitted signal phase for $t \ge 0$ is given by

$$\theta_k(t) = a_k\psi(t-kT) + \pi q\sum_{r=0}^{k-1} a_r + \theta_{-1}, \quad kT \le t \le (k+1)T \tag{7}$$

where $a_i = \pm 1$ denotes the binary data, $k = 0, 1, \ldots$ and the phase function is defined as

$$\psi(t) = \begin{cases} 0, & t \le 0, t > T \\ \pi\left(h\dfrac{t}{T} - w\left(\dfrac{t}{T}\right)^2\right), & 0 \le t \le T \\ \pi q = \pi(h-w), & t = T \end{cases}$$

where h and w represent the normalized initial peak-to-peak frequency deviation and the frequency sweep width, respectively. For the case of coherent detection, it has been shown [12] that a receiver with an observation of 2 bits provides a good compromise between signal-to-noise ratio (SNR) gain and system complexity.

As conjectured by Hirt and Pasupathy [12], the system performance can be further improved by considering multimode continuous phase systems. Raveendra [13] investigated the approach of varying the modulation of a continuous-phase FSK. Here, the phase in (7) is replaced by

$$\theta(t) = \sum_{i=1}^{n} a_i\psi_i(t-(i-1)T), \qquad 0 \le t \le nT \tag{8}$$

where the phase functions in (8) depend now on the symbol interval i according to

$$\psi_i(t) = \begin{cases} 0, & t \le 0 \\ \pi\left(h_i\dfrac{t}{T} - w_i\left(\dfrac{t}{T}\right)^2\right), & 0 \le t \le T \\ \pi q_i = \pi(h_i - w_i), & t \ge T \end{cases}$$

While in conventional monomode continuous-phase chirp transmission $q_i = q$ and $w_i = w$ for $i = 1, 2, \ldots$, the (q_i, w_i) now form a sequence of sets with period K, specifically, $(q_i, w_i) = (q_{i+K}, w_{i+K})$.

Fonseka [14] employed partial-response CPCM (PR-CPCM) signals in an attempt to increase the minimum distance $d_{\min}$ of the signals that determines the system performance. At the same time, the spectrum is to be kept flat so that the system is robust against jammers in the transmission band. The increase in the number of states in PR-CPCM as compared to FR-CPCM is an important issue. If in the latter $(h - w)$ is expressed as the ratio of two relatively prime integers as $h - w = \ell/m$, the m possible states during any interval can be represented by m evenly spaced phase states. In PR-CPCM with LT denoting the support of the baseband frequency pulse, the number of phase states depends on the individual values of h and w. When h and w are expressed as ratios of integers $h/L = \ell_1/m$ and $w/L^2 = \ell_2/m$ with the smallest common denominator m, the number of phase states during any interval is m. In view of the symbol states arising from the $(L - 1)$ previous symbols, the total number of states in PR-CPCM is $m2^{(L-1)}$.

2.4.3. Multiple Access. The aforementioned modulation schemes are designed for the case where receiver thermal noise (and possibly some narrowband interfering signals) represent the only disturbances in the bandwidth occupied by the signal. If the spectrum is shared among M simultaneously transmitting users, the resulting multiple-access interference (MAI) among the users increases the BER as compared to the single-user system. To avoid complex signal processing schemes for interference mitigation at the receiver, the signal formats have to be chosen carefully in order to limit the MAI. Different approaches based on CM have been proposed and are discussed below.

Takeuchi and Yamanouchi [15] assigned consecutive bits transmitted by CM with DPSK to different users; thus, time-division multiple access (TDMA) is applied here. The sidelobe level suppression and processing gain are 30 and 19 dB, respectively, and the system is implemented using SAW devices. Nonlinear CM is applied in order to obtain a flat amplitude spectrum within B.

A more sophisticated approach [16] for multiple access assigns $2M$ different instantaneous frequency functions $f(t)$ [cf. Eq. (2)] to M users employing binary signaling. The chirp duration T is split into two intervals of length $T/2$, and each of the proposed multiuser chirp signals of duration T is characterized by two different slopes in the two intervals. This approach is a straightforward extension of the chirp signals used in BOK to a M-user system where all signals occupy a common bandwidth B. Since the detection of the M symbols requires $2M$ different chirp matched filters, however, the complexity of the receiver is relatively high as compared to that reported by Takeuchi and Yamanochi [15].

Frequency-hopped code-division multiple access (FH-CDMA) is employed in another study [17]. Using basically the multiuser chirp signals of Ref. 16 in each time–frequency (TF) hop, the collision of different FH-CDMA user signals containing binary frequency-shift keying (FSK) symbols in the same TF hop can be resolved. Still, however, $2M$ different chirp matched filters are required and all chirp filters have to be changed after a new user has entered the system.

Improved flexibility and a simple receiver design are the main objectives of the multiple-access scheme in another paper [18], where transmission from a base station to multiple users over a multipath channel is considered. Here, the advantages of synchronous binary direct-sequence (DS) CDMA are combined with chirp signaling. The positive and negative chip pulses $\pm c(t)$ normally used in DS-CDMA for spreading the signal spectrum are replaced by the upchirp and downchirp signals in Section 2.1. It is shown that the quasiorthogonality of these signals allows for a noncoherent detection followed by a postdetection integrator (PDI) whose output is sampled to provide estimates of the superimposed code sequences of the different users. Finally, the information bit of a certain user is estimated from the correlation of the sample sequence with the user code. Only two different chirp filters are required, and the multiple access can be fully controlled by assigning codes in the digital domain.

3. PERFORMANCE ANALYSIS

For the case of linear modulation, the BER of a BOK system has been evaluated [10] for nonoverlapping chirp signals with $T = 2\,\mu\text{s}$, $B = 80\,\text{MHz}$, and a sidelobe suppression of $G_{SL} = 42\,\text{dB}$, resulting in a bit rate of 500 kbps (kilobits per second). To increase the data rate to 2 Mbps, the chirp pulses are allowed to overlap. Figure 5 shows the resulting BER as a function of the SNR $\gamma = E_b/N_0$ in an AWGN channel where E_b and N_0 denote the bit energy and the noise power spectral density, respectively. It has been pointed out [10] that multipath propagation limits the BER performance where sequences of consecutive upchirp or downchirp signals are to be avoided. BER results corresponding to Fig. 5 for $\pi/4$-DQPSK can be found in Gugler's thesis [9].

Hirt and Pasupathy [12] investigated the BER performance of FR-CPCM for coherent detection. It has been shown that a receiver with an observation of 2 bits provides a good compromise between SNR gain and system complexity. In this case, the optimum choice $(q, w) = (0.28, 1.85)$ gives an SNR gain of 1.75 dB over the optimum coherent 1-bit chirp receiver with $(q, w) = (0.35, 1.55)$.

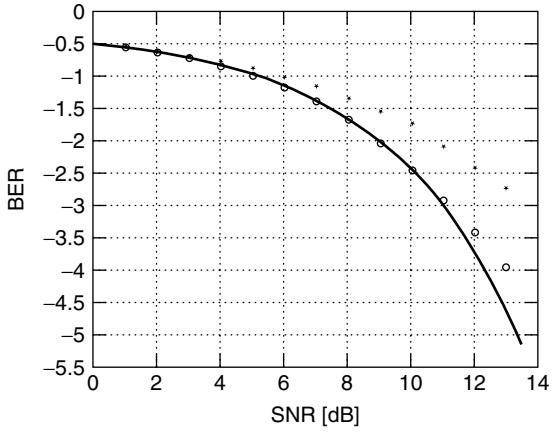


Figure 5. BER for BOK and different data rates: lower bound for orthogonal signals [8] (−), nonoverlapping 2-μs-long chirp signals (○), and overlapping 2-μs-long chirp signals with a data rate of 2 Mbps (∗) (figure taken from Ref. 10).

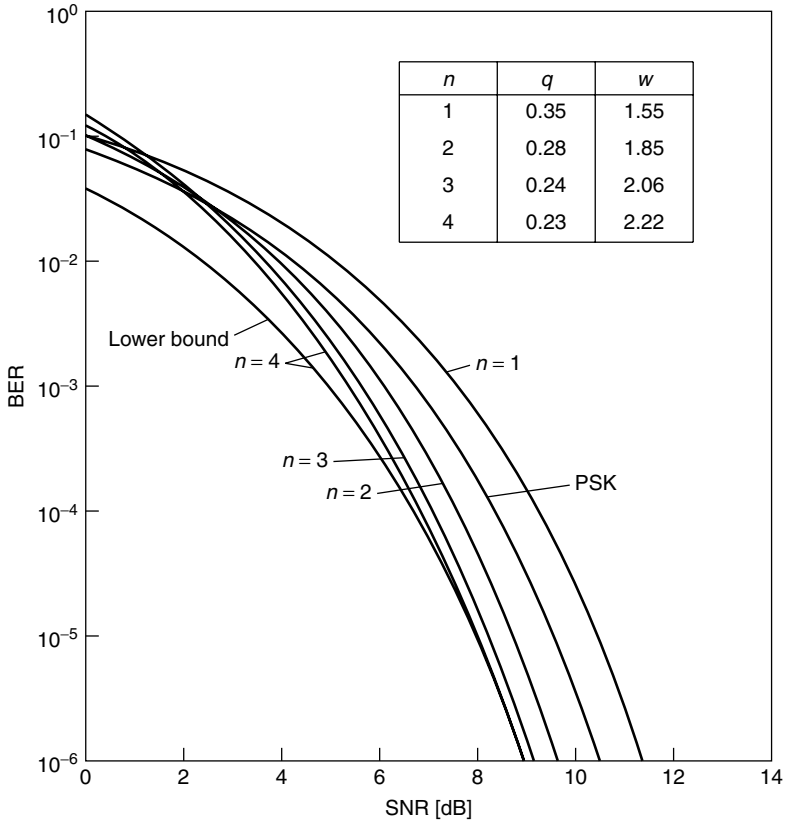


Figure 6. BER bounds for a coherent FR-CPCM receiver (figure taken from Ref. 12).

Upper BER bounds for different values of the observation length n together with the optimal values for q and w are shown in Fig. 6. The lower bound for $n = 4$ indicates the tightness of the bounds for increasing SNR values. In the context of a possible implementation, a simple suboptimum average matched filter (AMF) is shown to provide binary PSK performance for an optimum 2-bit observation. In another study, Hirt and Pasupathy [19] investigated the noncoherent detection case for FR-CPCM and showed the 3-bit noncoherent AMF receiver is to yield a 3-dB SNR gain over a wide range of signal parameters.

An investigation of multimode transmission performance [13] reveals that for an observation interval with $n = 5$, the optimum dual-mode chirp system, namely, $K = 2$, with $(q_1, w_1) = (0.3, 1.68)$ and $(q_2, w_2) = (0.5, 1.68)$ outperforms the optimum coherent 1-bit chirp receiver with $(q, w) = (0.35, 1.55)$ by 3.4 dB.

For Fonseka's [14] PR-CPCM with $L = 4$ and 24 states, the value of $d^2_{\min}$ can be increased by a factor of 2.15 as compared to the case $L = 1$ [14]. Furthermore, the PR-CPCM signals, are shown to have better spectral variations than conventional CM signals which indeed leads to the required robustness against jamming.

A CM TDMA scheme with DPSK [15] shows a BER that increases is only marginally for an increasing number of simultaneous users M. It is shown in simulations that the SNR loss for $M = 9$ as compared to $M = 1$ is only about 1 dB, where the latter case is about 2 dB worse than the lower BER bound for DPSK transmission.

The CM multiple-access scheme with the $2M$ different instantaneous frequency functions has been analyzed [16] for an AWGN channel using upper BER bounds for different values of M. As in the case of the CM TDMA scheme with DPSK, the bounds are relatively robust against different values of M. For $TB = 500$, the SNR loss of $M = 16$ as compared to $M = 1$ is only about 1 dB, and the BER decreases for increasing values of TB.

In another study [17], FH-CDMA with multirate chirp rate (MRC) signals is compared with a FH-CDMA scheme with FSK for an AWGN channel. It is observed that MRC-FH-CDMA is at least 2 dB better than the FSK-FH-CDMA scheme. For a BER of 10^{-1}, an increase of M to $M + 5$ results in a loss of 0.6 dB for MRC-FH-CDMA and 1.5 dB for FSK-FH-CDMA, respectively.

Kocian and Dahlhaus [18] observed bounds on the BER performance of the CDMA scheme described at the end of Section 2.4.3 with PDI and derived a square-law envelope detector in a non-frequency-selective channel (NFSC). The BER in a frequency-selective channel (FSC) with Rayleigh fading and an exponential power delay profile is depicted in Fig. 7 for $M = 1$ and $M = 16$ users, respectively, for a CDMA codelength of $N_c = 16$. For comparison, the BER of the optimum noncoherent detector (ND) has been included in Fig. 7. If the data rate is increased, the decision variable is corrupted by ISI and MAI, and the BER starts to saturate for increasing γ.

The preceding analysis has assumed a perfect channel state information at the receiver. Partially coherent detection of CPCM signals is considered in another study [20], while parameter estimation of chirp signals has been treated by Kay and others [21,22].

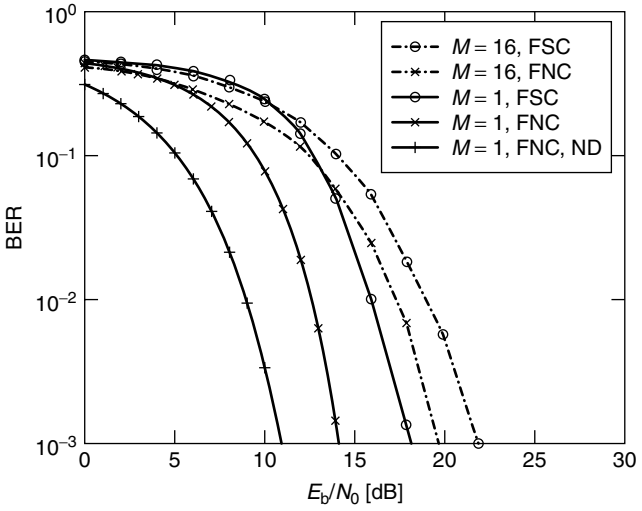


Figure 7. Mean BER for $N_c = 16$ in a multipath fading channels for different numbers M of users (figure taken from Ref. 18).

4. IMPLEMENTATION ISSUES

There are different ways to implement communication systems using CM. The two most prominent ones, namely SAW filters and direct digital frequency synthesizers (DDFS), are outlined below. Approaches based on voltage-controlled oscillators with appropriate function generators as well as excitation of a conjugate MF network with an impulse are described in another treatise [3].

In SAW filters there is no need for complex digital baseband signal processing. SAW filters are well suited for today's wireless communications because of their high performance, small size, and low cost [23]. On the other hand, since the CM parameters specify the form of the filter, SAW devices are not flexible and cannot be applied in systems where the parameters are subject to changes. SAW filters operate at an intermediate frequency (IF), and a mixer is used to upconvert the signal to the radiofrequency (RF). For the system reported by Koller et al. [23], IF = 348.8 MHz, RF = 2.45 GHz, $B = 80$ MHz, and $T = 0.5\,\mu$s. For the $\pi/4$-DQPSK modulation described by Gugler [9], a suitable pulse for exciting the filter is located at the IF center frequency and has a rectangular shape, the length of which equals four periods of the IF. Unlike in conventional systems where the IF is modulated by a $\pi/4$-DQPSK signal, the IF pulse exciting the SAW filter is modulated. Because of the sensitivity of the output power on the rising time of the broadband pulse exciting the filter, ASK is not suited for CM with SAW filters. Other examples of CM systems based on SAW filters can be found in the literature [15,24,25].

Unlike SAW filters, DDFS are highly flexible since the parameters of the CM signal can be set by a digital controller. Salous et al. [26] have presented a digital chirp sounder for mobile radio applications. In particular, the time and frequency resolutions are fully programmable. This is important for the different multiple-access schemes in Section 2.4.3, where the CM signal format depends on the number of simultaneous users. Clearly, the computational effort is large, but it is expected for decreasing costs of digital components such as digital–analog converters and dedicated signal processors that the DDFS will be preferred to the SAW approach in the future. A commercially available digital channel sounder with CM is described in Ref. 27 and a digital local oscillator for CM generation, in Ref. 28. Allen et al. [29] have direct digital synthesis of CM signals for a light detection and ranging application.

5. OTHER ASPECTS RELATED TO CM IN COMMUNICATION SYSTEMS

CM has been described in Section 2.4 as a means of transmitting information in form of a spread-spectrum signal over a linear channel. In semiconductor lasers (SCL), frequency chirping arises as an undesired effect in fiberoptic transmission using current modulation. As pointed out in another study [30], when the device current is modulated at frequencies approaching a few gigahertz, the dynamic response of SCL leads to an increased linewidth of an individual longitudinal mode where the line broadening is proportional to the linewidth enhancement factor (also termed the *antiguiding parameter*) β_c. The resulting chirp has its origin in the carrier-induced refractive-index change that accompanies any gain change in SCL. Agrawal and Dutta [30] described several measures that lead to light emission of SCL predominantly in a single longitudinal mode even under high-speed modulation.

Concerning the use of CM in communication systems, there are presently only very few applications. One example is the Consumer Electronic Bus (CEBus) EIA/IS60 powerline communications standard where CM can be used with a data rate of 10 kbps in an unlicensed frequency band, 100–450 kHz, for home networking. Although proposals have been made to use CM in wireless communication systems, especially in wireless local-area networks, standardization bodies have preferred other types of modulation to CM. With the advent of more advanced CM systems based on DDFS in combination with software defined radio concepts, however, CM might be an interesting alternative modulation format for future communication systems operating in frequency-selective channel environments.

BIOGRAPHY

Dirk Dahlhaus received the Dipl.-Ing. degree in electrical engineering from Ruhr-Universität Bochum, Germany, in 1992, and the Ph.D. degree from Swiss Federal Institute of Technology (ETH) Zurich, Switzerland, in 1998. Since April 1999, he has been assistant professor for mobile radio systems at the Communication Technology Laboratory of ETH Zurich. He was president of the 2002 International Zurich Seminar on Broadband Communications. His main research interests include radio channel modelling, digital signal processing and link adaptation in multiuser wireless and mobile radio communication systems.

BIBLIOGRAPHY

1. A. J. Berni and W. D. Gregg, On the utility of chirp modulation for digital signaling, *IEEE Trans. Commun.* **COM-21**: 748–751 (1973).

2. M. R. Winkler, Chirp signals for communications, *Proc. Western Electronic Show and Convention* (*WESCON*), Los Angeles, Aug. 21–24, 1962, Vol. 14.2.

3. C. E. Cook and M. Bernfeld, *Radar Signals*, Artech House, Norwood, MA, 1993.

4. C. L. Dolph, A current distribution for broadside arrays which optimizes the relationship between beamwidth and sidelobe level, *Proc. IRE* **34**: 335–348 (1946).

5. G. J. Van der Maas, A simplified calculation for Dolph-Tchebycheff arrays, *J. Appl. Phys.* **25**: 121–124 (1954).

6. M. Kowatsch, *Codierte Nachrichtenübertragung mit Chirp-Modulation*, Ph.D. thesis (in German), Technical Univ. Vienna, Vienna, Austria, 1981.

7. M. Kowatsch, H. R. Stocker, F. J. Seifert, and J. Lafferl, Time sidelobe performance of low time-bandwidth product linear FM pulse compression systems, *IEEE Trans. Sonics Ultrasonics* **28**(4): 285–288 (July 1981).

8. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

9. W. Gugler, *Untersuchung von hochratigen OFW-basierten Chirp-Übertragungssystemen*, Ph.D. thesis (in German), J. Kepler Univ. Linz, Linz, Austria, 2000.

10. A. Springer et al., A robust ultra-broad-band wireless communication system using SAW chirped delay lines, *IEEE Trans. Microwave Theory Tech.* **46**(12): 2213–2219 (Dec. 1998).

11. S. E. El-Khamy and S. E. Shaaban, Matched chirp modulation: Detection and performance in dispersive communication channels, *IEEE Trans. Commun.* **36**(4): 335–348 (April 1988).

12. W. Hirt and S. Pasupathy, Continuous phase chirp (CPC) signals for binary data communication — Part I: Coherent detection, *IEEE Trans. Commun.* **COM-29**(6): 836–847 (June 1981).

13. K. V. Raveendra, Digital transmission using multimode phase-continuous chirp signals, *IEE Proc. Commun.* **143**(2) (April 1996).

14. J. P. Fonseka, Partial response continuous phase chirp modulation, *IEE Electron. Lett.* **35**(6): 448–449 (March 1999).

15. Y. Takeuchi and K. Yamanouchi, A chirp spread spectrum DPSK modulator and demodulator for a time shift multiple access communication system by using SAW devices, *Microwave Symp. Digest*, 1998 IEEE MTT-S International, 1998, Vol. 2, pp. 507–510.

16. S. E. El-Khamy, S. E. Shaaban, and E. A. Thabet, Efficient multiple access communications using multi-user chirp modulation signals, *Proc. IEEE 4th Int. Symp. Spread Spectrum Techniques and Applications* (*ISSSTA'96*), Mainz, Germany, 1996, Vol. 3, pp. 1209–1213.

17. C. Gupta and A. Papandreou-Suppappola, Wireless CDMA communications using time-varying signals, *Proc. 6th Int. Symp. Signal Processing and Its Applications*, 2001, Vol. 1, pp. 242–245.

18. A. Kocian and D. Dahlhaus, Downlink performance analysis of a CDMA mobile radio system with chirp modulation, *Proc. 49th IEEE Vehicular Technology Conf.* (*VTC'99*) *Spring*, Houston, TX, 1999, Vol. 1, pp. 238–242.

19. W. Hirt and S. Pasupathy, Continuous phase chirp (CPC) signals for binary data communication — Part II: Noncoherent detection, *IEEE Trans. Commun.* **COM-29**(6): 847–858 (June 1981).

20. S. E. El-Khamy, S. E. Shaaban, and E. A. Thabet, Partially coherent detection of continuous phase signals, *Proc. 13th Nat. Radio Science Conf.*, Cairo, Egypt, 1996, pp. 1–11.

21. P. M. Djuric and S. M. Kay, Parameter estimation of chirp signals, *IEEE Trans. Acoustics Speech Signal Process.* **38**(12): 2118–2126 (Dec. 1990).

22. S. Saha and S. M. Kay, Maximum likelihood parameter estimation of superimposed chirps using Monte Carlo importance sampling, *IEEE Trans. Signal Process.* **50**(2): 2118–2126 (Feb. 2002).

23. R. Koller et al., A SAW based high-speed spread-spectrum WLAN using chirp $\pi/4$-DQPSK modulation, *Proc. IEEE 2000 Ultrasonics Symp.*, 2000, pp. 367–370.

24. J. Q. Pinkney, A. B. Sesay, S. Nichols, and R. Behin, A robust high speed indoor wireless communications system using chirp spread spectrum, *Proc. 1999 IEEE Canadian Conf. Electrical and Computer Engineering*, Edmonton, Alberta, Canada, May 1999, Vol. 1, pp. 84–89.

25. Y. R. Tsai and J. F. Chang, The feasibility of combating multipath interference by chirp spread spectrum techniques over Rayleigh and Rician fading channels, *Proc. IEEE 3rd Int. Symp. Spread Spectrum Techniques and Applications* (*ISSSTA'94*), 1994, Vol. 1, pp. 282–286.

26. S. Salous, N. Nikandrou, and N. F. Bajj, Digital techniques for mobile radio chirp sounders, *IEE Proc. Commun.* **145**(3): 191–196 (June 1998).

27. *http://www.gage applied.com/resource/newslett/07.3/Real-World.htm* (2002).

28. *http://www.spectrumsignal.com/support/_and_training/3_manuals/tim-ddc.pdf* (2002).

29. C. Allen, Y. Cobanoglu, S. K. Chong, and S. Gogineni, Performance of a 1319nm laser radar using RF pulse compression, *Proc. 2001 Int. Geoscience and Remote Sensing Symp.* (*IGARSS '01*), July 2001.

30. G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*, Kluwer, Boston, 1993.

COCHANNEL INTERFERENCE IN DIGITAL CELLULAR TDMA NETWORKS

Savo G. Glisic
Pekka Pirinen
University of Oulu
Oulu, Finland

1. INTRODUCTION

In cellular TDMA networks cochannel interference is generated in surrounding cells using the same carrier frequency. For this reason a careful planning of sectors and surrounding layers allowed to reuse the same frequency is required. In addition to sectorization (three sectors per cell), narrower antenna lobes can be used to further reduce the angular sectors of the receiving antennas so that the interference can be spatially filtered.

Usually none of these measures are efficient enough to warrant additional action to deal with the interference by using different cancellation techniques in either time,

frequency, or spatial domain. Having this in mind, we can represent the residual interference signal power as

$$I_r(r,\theta,f,t) = (1-C_f)(1-C_p)(1-C_\theta)(1-C_t)I(r,\theta,f,t)$$
$$= (1-C_r)(1-C_\theta)(1-C_t)I(r,\theta,f,t) \quad (1)$$

where C_f, C_p, C_θ, and C_t are frequency, propagation (distance + shadowing + fading), angle (space), and time isolation coefficients, respectively. $I(r,\theta,f,t)$ is the interference signal power without any suppression techniques. For perfect isolation, at least one of these coefficients is equal to one and the interference has no influence on the received signal. In practice, it is rather difficult and economically impractical to reach the point where $C_i = 1$. Instead, the product $(1-C_r)(1-C_\theta)(1-C_t)$ depending on these coefficients should be kept as low as possible with an affordable effort measured by cost, power consumption, and physical size of the hardware required for the solution.

Coefficient C_f is related to frequency assignment in the cellular network, while coefficient C_p is related to the propagation conditions. $C_f = 1$ if the interfering signal frequency is different from the frequency of the useful signal. $C_p = 1$ if, as a result of propagation losses, the interfering signal cannot reach the site of the useful reference signal. In general, the same frequency can be used in two cells only if the propagation losses between the two cells are high enough that the interfering signals are attenuated to the acceptable level. This will be characterized by the frequency reuse coefficient C_r defined as $(1-C_r) = (1-C_f)(1-C_p)$ and will be discussed in Section 2. Coefficient C_θ is related to antenna beamforming, and possibilities of reducing the interference level by spatial filtering are discussed in Section 3. Finally, interference cancellation and equalization in time domain, which is included in coefficient C_t, will be discussed in Section 4.

2. NETWORK PLANNING AND FREQUENCY REUSE

Depending on the cell size, three different categories of cellular networks can be distinguished. *Macrocells* are the largest, with a cell radius of 1 km up to 35 km or more. Base station antennas are located well above the rooftop level. The commonly used macrocellular modeling structure assumes a uniform grid of hexagonal cells [1]. Part of the hexagonal cellular layout is illustrated in Fig. 1. In Fig. 1a frequencies are reused in each cluster of seven cells and in Fig. 1b the cluster size is 3.

The hexagonal grid is optimal in the sense that there is no overlap between cells. In addition, hexagons closely approximate circles. This kind of modeling is highly theoretical since effective cell coverage areas vary considerably depending on factors such as terrain, buildings, weather, and time. Cellular models can be further classified according to base station antenna directivity. In the case of omnidirectional antennas, base stations can be located in the cell centers as illustrated in Fig. 2a. When directional antennas are used, cells can be divided into widebeam sectors as shown in Fig. 2b. Directional antenna patterns can also be modeled by corner-illuminated base stations with three narrow antenna lobes per base station. One advantage of this approach is lower cost. Fewer base stations are required over a certain geographical area than with direct sectorization. Corner-illuminated cells or the so-called "three leaf clover" structure is illustrated in Fig. 2c.

Microcells are smaller than macrocells with a typical cell radius of 20–300 m. In this scenario base station antennas are usually below the mean rooftop level. In urban areas microcells are often characterized as having a Manhattan type of grid, where the base stations are in the crossings of linear streets as shown in Fig. 1c. *Picocells* or indoor cells usually cover indoor areas (rooms, halls) with typical cell radius of 5–30 m. These scenarios are not covered in this article.

Frequency reuse is an essential element in cellular networks. It means that the same frequencies are reused in the system within a certain distance depending on the reuse factor. This reuse factor can be represented as a cluster size, which includes the group of cells where all different available channels are used. Regular cluster sizes K [1] can be calculated according to

$$K = i^2 + ij + j^2 \quad (2)$$

where i and j are nonnegative integers. Equation (2) leads to balanced cluster sizes $K = 1, 3, 4, 7, 9, 12, \ldots$. If D is defined as the distance between the closest cochannel centers and R is the cell radius (see Fig. 1a), the frequency reuse factor D/R and the cluster size K are related as [1]

$$\frac{D}{R} = (3K)^{1/2} \quad (3)$$

In order to increase network capacity, cluster size must be reduced. The more aggressive reuse (smaller the cluster size), the higher the level of cochannel interference that will be generated and vice versa. For these reasons frequency reuse has been studied extensively in the literature [1–4]. The problem becomes even more challenging if macrocells and microcells are overlaid [5].

2.1. Cochannel Interference Distributions

From the previous discussion one can see that no matter what cluster size is chosen, a certain level of cochannel interference (CCI) can not be avoided. For the analysis of cochannel interference statistics, the CCI distribution function is required. Cochannel interference can be seen as a superposition of distance-dependent attenuation (path loss), short-term fluctuations, and long-term variations. The long-term or large-scale signal variation (shadowing, slow fading) can be characterized by the lognormal distribution. The short-term signal variation (fast fading), on the other hand, may fit to some other distributions such as Rayleigh, Rice, or Nakagami. An overview of fading distributions related to CCI can be found in the paper by Yacoub [6]. In the sequel, the lognormal shadowing is assumed.

The total interference power is often accumulated from several cochannel signals. Unfortunately, there is no known closed-form expression for the distribution of

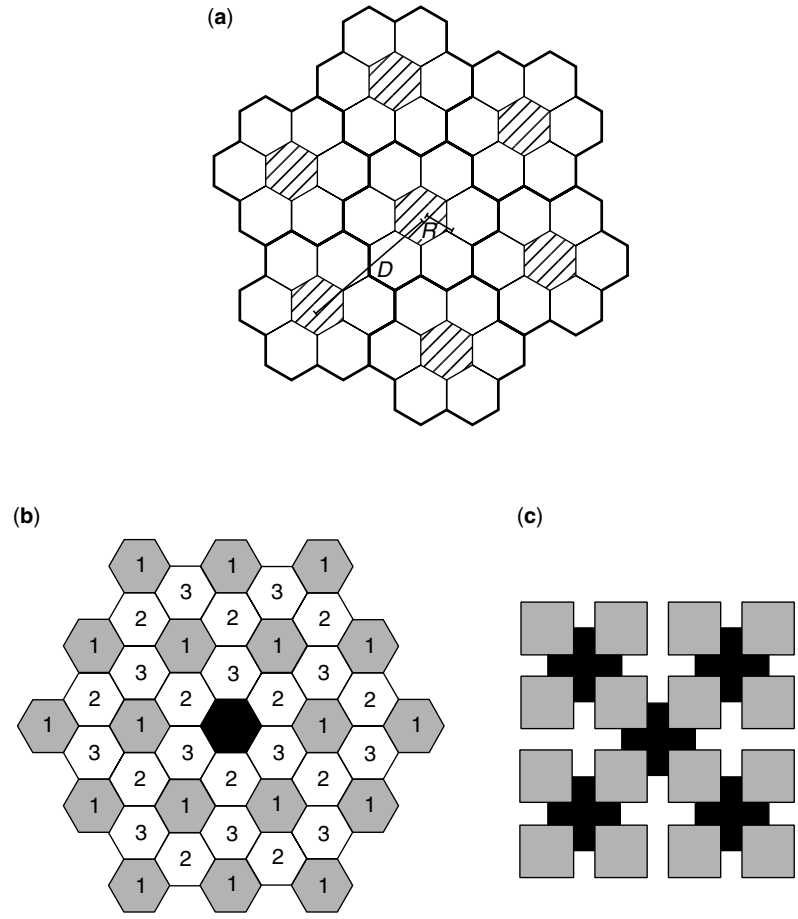


Figure 1. Cellular layouts: (**a**) uniform hexagonal cellular layout with reuse 7; (**b**) macrocell layout with reuse 3; (**c**) street microcell layout with reuse 2.

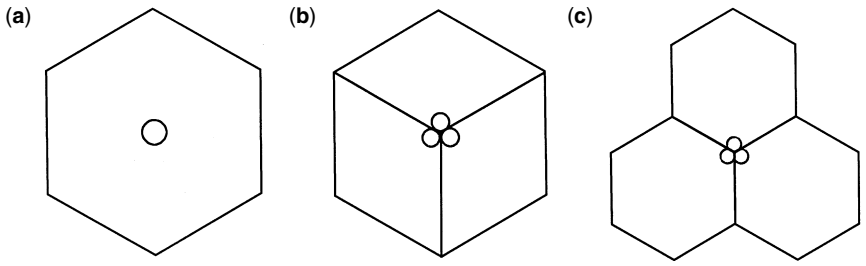


Figure 2. Cell types considered: (**a**) omnicell; (**b**) sectored cell; (**c**) corner-illuminated cells.

the sum of lognormally distributed random variables. However, several approximations have been derived. A common feature for all these approximations is that they estimate the sum of lognormal random variables by another lognormally distributed random variable [7]. This can be represented as

$$L = \sum_{i=1}^{n} e^{y_i} = \sum L_i \cong e^z \tag{4}$$

where y_i represents Gaussian random variables. In the Fenton–Wilkinson (FW) approximation [7–10], the mean m_z and the standard deviation σ_z of z are derived by matching the first two moments of the both sides of Eq. (4). If the first moment of $(L_1 + L_2 + \cdots + L_n)$ is denoted by u_1 and the second by u_2, the following expression is obtained after moment matching [9]

$$m_z = 2 \ln u_1 - \tfrac{1}{2} \ln u_2 \tag{5}$$

$$\sigma_z^2 = \ln u_2 - 2 \ln u_1. \tag{6}$$

The Fenton–Wilkinson approach is applicable when the standard deviations of the lognormal components are lower than 4 dB for uncorrelated signal components [11]. For higher deviation values, this approximation tends to underestimate the mean and overestimate the variance of the sum distribution. When there is correlation between the components, the FW approximation is quite accurate for higher deviation values ($\leq$12 dB), too [9].

The Schwartz–Yeh (SY) method [7,9–11] is also based on the assumption that the power sum is lognormally

distributed. The SY approximation is different in the use of the exact expressions for the first two moments of the sum of two lognormal random variables. Nesting and recursion techniques are then used to extend the approach to a larger number of cumulative random variables. Originally, the SY method was developed for the sum of independent lognormal random variables. However, it has been extended to the case of correlated lognormal random variables with some modifications [9].

The Schwartz–Yeh approximation can be best applied when the range of the standard deviation is $4 \leq \sigma \leq 12$ dB. If all components in the summation are identically distributed, this approximation tends to underestimate the variance in the resulting signal distribution. The error increases as a function of the number of added components.

In addition to the Fenton–Wilkinson and Schwartz–Yeh approaches, there are some other approximations for the sum of lognormal components. For example, Farley's approximation is a strict lower bound for the cumulative distribution function (CDF) of a sum of independent lognormal random variables [7]. For further studies on lognormal sum approximations the reader is referred to the additional reading listed at the end of this article.

2.2. Cochannel Interference and Outage Probabilities

Following Refs. 10 and 12, cochannel interference probability is defined as

$$P(I_c) = \sum_n P(I_c|n)P_n(n) \tag{7}$$

where $P_n(n)$ is the probability of n cochannel interferers being active and $P(I_c|n)$ is the corresponding conditional CCI probability.

The conditional CCI probability can be defined as

$$P(I_c|n) = P\left(\frac{C}{I} < \alpha\right) \tag{8}$$

where C is the instantaneous power of the desired signal (carrier), I is the joint interference power from n active cochannel users, and α is the specified cochannel interference protection ratio.

$P_n(n)$ can be represented by the binomial distribution in terms of carried traffic per channel

$$P_n(n) = \binom{N}{n} a_c^n (1 - a_c)^{N-n} \tag{9}$$

where N is the number of effective cochannel interferers ($N = 6$ if only the closest ring cochannel interferers are taken into account) and $a_c = m_1/m_t$ is carried traffic per channel (erlangs per channel). Parameters m_1 and m_t are discussed in more detail in the next Section 2.3. It is assumed that the number of traffic channels is equal for all cells.

The outage probability P_{out} for the desired user can be defined as the probability of failing to achieve a bit error probability P_e lower than a fixed threshold P_{e0}, namely

$$P_{\text{out}} = P(P_e > P_{e0}) \tag{10}$$

If only the effects of cochannel interference are taken into account, the received carrier-to-interference ratio C/I is the key parameter. If the minimum required carrier-to-interference ratio is α and it corresponds to the bit error probability $P_e = P_{e0}$, the outage probability is the same as the conditional CCI probability defined by (8).

Following the procedure in [9], the outage probability of the lognormally distributed signals can be represented in the form

$$P_{\text{out}} = P(I_c|n) = 1 - Q\left(\frac{\ln\alpha - \ln\xi_d + m_{z_n}}{(\sigma_d^2 + \sigma_{z_n}^2 - 2r_{yz}\sigma_d\sigma_{z_n})^{1/2}}\right) \tag{11}$$

where ξ_d is the area mean desired signal power, m_{z_n} is the area mean joint interference power of n interferers, σ_d is the standard deviation of the desired signal, σ_{z_n} is the standard deviation of the joint interference from n interferers, and r_{yz} is the correlation coefficient of the desired signal and joint interference. The initial mean single interferer power in the worst geometric case can be approximated by

$$m_{z_{1w}} = \ln[((3K)^{1/2} - 1)^{-\beta}] \tag{12}$$

In the average geometric case, the exact interferer power is of the form

$$m_{z_{1a}} = \ln[(3K)^{-\beta/2}] \tag{13}$$

The desired signal area mean power ξ_d can be represented as

$$\xi_d = \left(\frac{r}{R}\right)^{-\beta} \tag{14}$$

In Eqs. (12)–(14), β denotes the path loss exponent, K is the cluster size, and $r/R \in (0, 1]$ is the normalized distance between the desired mobile station and the base station.

The combined effect of frequency allocation and propagation conditions, characterized implicitly by the parameter $(1 - C_r)$, is illustrated in Fig. 3. The figure shows the outage probability defined by (11) with the maximum number of first-tier interferers as a function of cluster size (worst and average case geometries) with variable path loss exponents β. The standard deviation of each lognormal component is 6 dB. All signals are uncorrelated. The Fenton–Wilkinson method has been used for the mean and variance approximations.

It can be noted that in free-space propagation conditions ($\beta = 2$), cochannel interference can be very severe even for large cluster sizes. On the other hand, in dense urban areas, where the path loss attenuation slope is steep, small cluster sizes can be supported. That allows larger system capacity for highly populated cities where it is the most desirable. The outage probability is very sensitive to the changes in the propagation environment.

Figure 4 shows the strong impact of normalized mobile distance (14) on conditioned full load CCI probability (outage probability) in the presence of lognormal shadowing ($\sigma = 6$ dB, FW approximation). It can be seen that without power control, outage events are more likely near the cell edges. Larger cluster sizes guarantee lower outage probabilities. The gap between worst and average

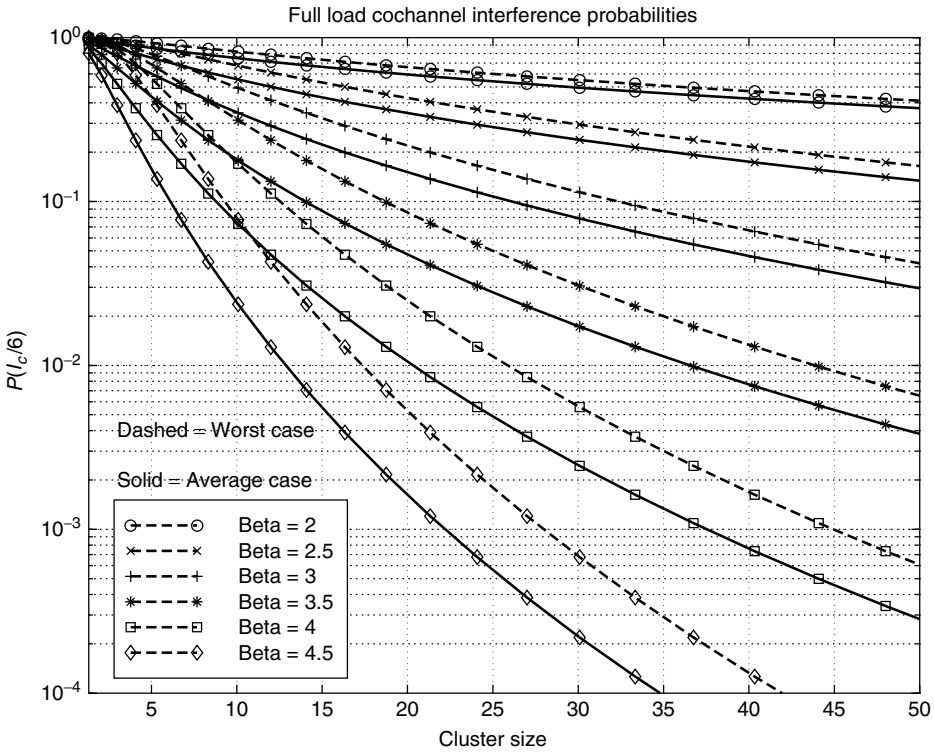


Figure 3. Effect of path loss exponent variation to the outage probability.

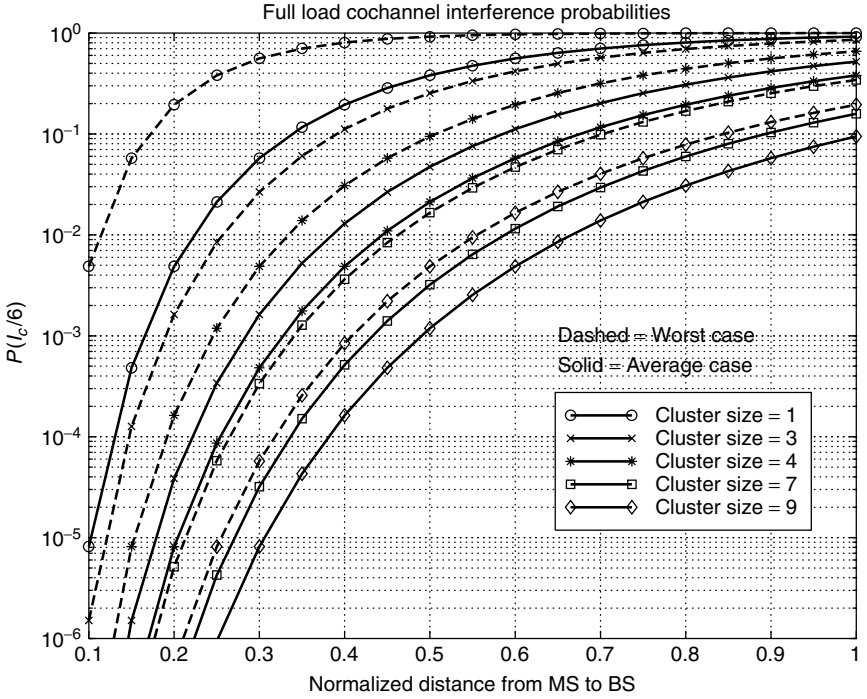


Figure 4. Full-load outage probabilities at variable cluster sizes.

case interference geometries diminishes as the cluster size increases.

2.3. Spectrum Efficiency

Spectrum efficiency describes how effectively a system can utilize limited frequency resources. In general, spectrum efficiency can be seen as a ratio between benefit (number of traffic channels, data rate) and cost (bandwidth) [13]. Spectrum efficiency is usually measured in units related to system capacity. Problems may arise if system capacities of different systems have been calculated with different assumptions or if they are represented in different units. A fair comparison of different systems is often cumbersome. Falciasecca et al. [13] discuss the effect of some working assumptions on spectrum efficiency, and ways to allow a fair comparison.

The system capacity of a voice-oriented network is related to the grade of service by the Erlang-B formula

$$P_B = \frac{{m_1}^{m_t}/m_t!}{\sum_{n=0}^{m_t} {m_1}^{m_t}/n!} = \Im(m_t, m_1) \tag{15}$$

where P_B is blocking probability, m_1 is the offered traffic (capacity) (in erlangs), and m_t is the total number of traffic channels. The blocking probability P_B refers to the probability that a new call attempt will not find an available channel in a trunk of channels and is dropped. Thus, there is no queueing in the system. The Erlang formula was originally developed for wired telephone traffic. It is not strictly applicable to cellular systems, because it does not take into account the handover traffic. An additional assumption of this model is that the total offered traffic is constant, which is not valid in the cell where the traffic is time-varying as a result of moving subscribers. Despite the limitations of the Erlang formula, it can be used for relative comparison purposes; however, one must be careful in interpreting the absolute values.

The spectrum efficiency and capacity evaluation are based on the radio capacity m_t introduced by Lee [14]. The radio capacity of the omnicellular TDMA system is defined as

$$m_t = \frac{N_s B_t}{B_c \left(\frac{2}{3}\left(\frac{C}{I}\right)_s\right)^{1/2}} = \frac{M_t}{\left(\frac{2}{3}\left(\frac{C}{I}\right)_s\right)^{1/2}} = \frac{M_t}{K} \quad \text{(frequency channels/cell)} \tag{16}$$

where B_t is the total allocated spectrum for the system, B_c is the channel bandwidth, $(C/I)_s$ is the minimum required carrier-to-interference ratio, M_t is the total number of frequency channels multiplied by the number of TDMA slots N_s, and K is the cluster size.

Radio capacity can be represented in different units as presented in Lee's paper [14]. These new measures can be derived from the general radio capacity definition and depend on issues such as service area, call duration, number of calls, and total bandwidth. Other commonly used units for spectrum efficiency are erlangs per MHz/km^2 and erlangs per cell/MHz.

For the system with parameters given in Table 1, maximum capacity obtained from Eq. (15) is presented in Table 2. By using (15) and (16), Fig. 5 illustrates real Erlang capacities for TDMA omnicellular uplink with

Table 1. Essential Parameters for the Capacity Evaluation of TDMA System

B_t (MHz)	R_C (kHz)	B_c (kHz)	α	$M = B_t/B_c$	M_t
10	270.8	200	9 dB	50	400

Table 2. Maximum Radio Capacities of Compared Cluster Sizes ($P_B = 0.02$)

TDMA (K)	m_t (Traffic Channels)	m_1 (erlangs/cell)	$a_c = m_1/m_t$
TDMA(7)	57	46.8	0.821
TDMA(9)	44	34.7	0.789
TDMA(12)	33	24.6	0.745
TDMA(21)	19	12.3	0.679

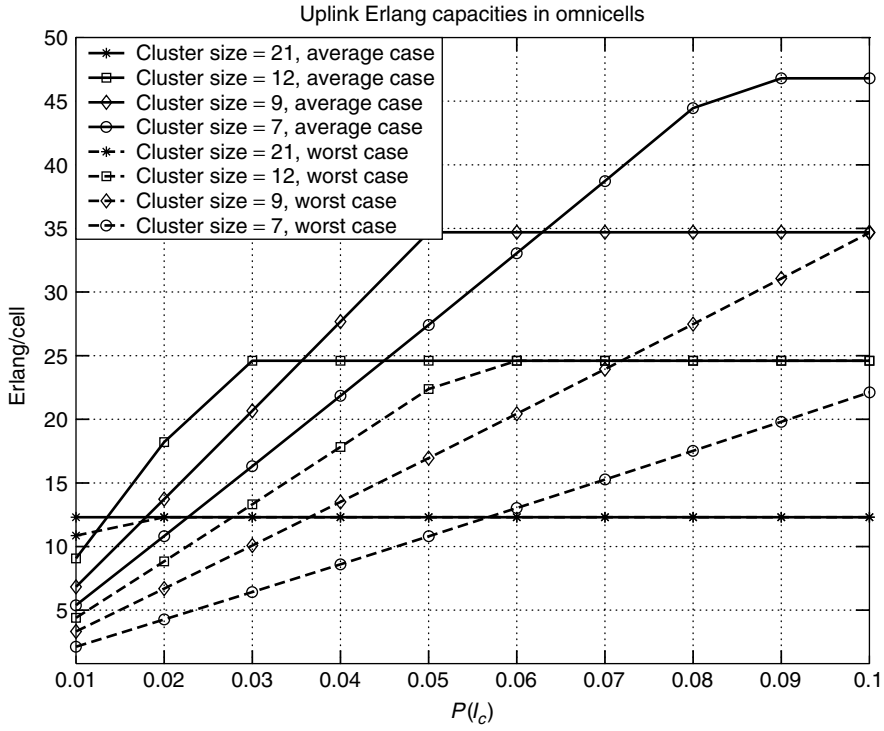


Figure 5. Uplink Erlang capacities in omnicells (Rayleigh fading only).

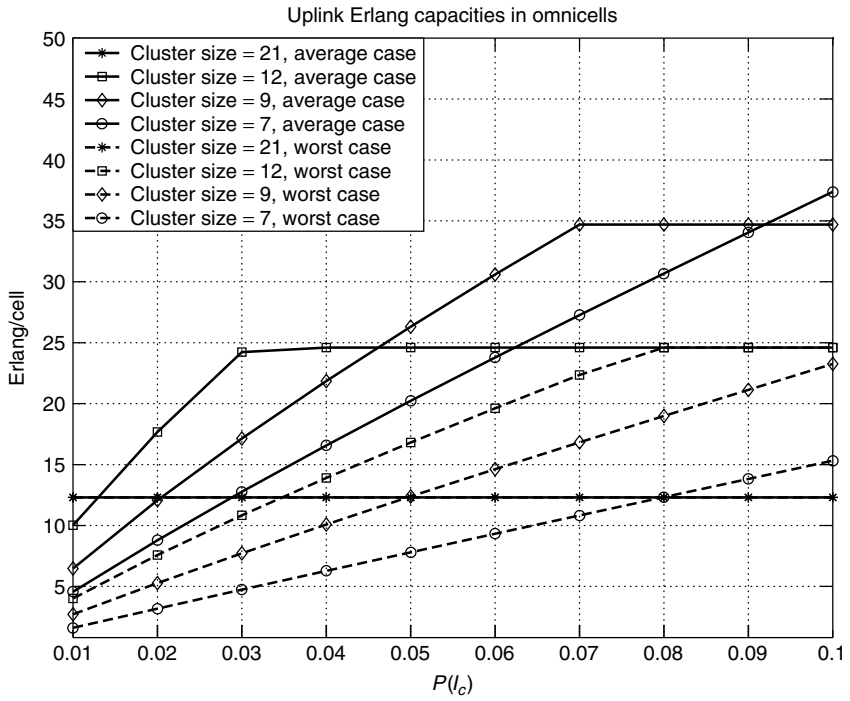


Figure 6. Uplink Erlang capacities in omnicells (lognormal shadowing only, $\sigma = 6$ dB).

several cluster sizes in a Rayleigh fading environment. Figure 6 presents uplink Erlang capacities when both the desired signal and cochannel interferers are lognormally shadowed with standard deviation $\sigma = 6$ dB.

Curves in Fig. 6 show that the performance is very close to the purely Rayleigh case. For larger cluster sizes, lognormal shadowing only ($\sigma = 6$ dB) will give more optimistic results than the purely Rayleigh case. Horizontal parts of the curves indicate that the maximum capacity limit m_1, for the particular cluster size, has been reached (hard capacity limit). Elsewhere, the Erlang capacity is softly limited by the highest tolerated CCI probability.

Alouini and Goldsmith consider a slightly different definition of spectrum efficiency, the area spectral efficiency (ASE) [15]. It is better suited for variable rate data transmissions, where the total throughput is of interest. The measure for average area spectral efficiency is the sum of the maximum average data rates divided by the bandwidth and the unit area for the system: bandwidth per second per Hz/m^2. The analytical framework provides theoretical limits for area spectral efficiency versus reuse distance for adaptive data rate cellular systems, where the rate adaptation depends on fading and interference conditions. Users' random locations, impact of propagation parameters, cell size, carrier frequency, and sectorization in both macrocells and microcells are taken into account. Furthermore, the loading in the cells is varied. Best and worst case analytical results are verified via average case Monte Carlo simulations.

Results based on the worst-case interference geometry show that the optimal reuse distance is close to 4. The area efficiency decreases as an exponential of a fourth-order polynomial relative to the cell size. Shadowing and fading reduce the absolute ASE, but do not affect the relative behavior as the function of reuse distance. Moreover, it is noted that the fading parameters of the desired user have stronger contribution on the ASE than do the fading parameters of the interferers [15].

3. SPATIAL FILTERING

Spatial domain processing included in the term $(1 - C_\theta)$ can be used to combat cochannel interference. At least at the base station there is a possibility to steer radiation/reception to the desired directions: (1) spatially directed transmission can enhance signal coverage and quality and (2) interference coming outside from the antenna main lobe is suppressed significantly in the reception.

3.1. Sectorization

One conventional way to improve cellular system capacity is cell splitting, that is, subdividing the coverage area of one base station to be covered by several base stations (smaller cells) [1]. Another simple and widely applied technique to reduce interference spatially is to divide cells into sectors, for example, three 120° sectors. These sectors are covered by one or several directional antenna elements. Effects of sectorization to spectrum efficiency have been studied [16]. Chan [16] concluded that sectorization reduces cochannel interference and improves signal-to-noise ratio of the desired link at the given cluster size. However, at the same time the trunking efficiency is decreased. Because of the improved link quality, a tighter frequency reuse satisfies the performance criterion in comparison to the omnicellular case. Therefore, the net effect of sectorization is positive at least for large cells and high traffic densities.

3.2. Adaptive Antennas

By using M-element antenna arrays at the base station the spatial filtering effect can be further improved.

The multiple beam adaptive array would not reduce the network trunking efficiency unlike sectorization and cell splitting [17]. These adaptive or "smart" antenna techniques can be divided into switched-beam, phased-array, and purely adaptive antenna systems. Advanced adaptive systems are also called *spatial division multiple access* (SDMA) systems. Advanced SDMA systems maximize the gain toward the desired mobile user and minimize the gain toward interfering signals in real time.

According to Winters [18], by applying a four-element adaptive array at the TDMA uplink, frequencies can be reused in every cell (three-sector system) and sevenfold capacity increase is achieved. Correspondingly, a four-beam antenna leads to reuse of 3 or 4 and doubled capacity at small angular spread.

Some practical examples of the impact of the use of advanced antenna techniques on the existing cellular standards have been described [19,20]. Petrus et al. [19] use the AMPS reference system and Mogensen et al. [20] use GSM. The Petrus et al. analysis [19] uses ideal and flat-top beamformers. The mainlobe of the ideal beamformer is flat and there are no sidelobes whereas the flat-top beamformer has a fixed sidelobe level. The ideal beamformer can be seen as a realization of the underloaded system; that is, there are less interferers than there are elements in the array. The overloaded case is better modeled by the flat-top beamformer because all interferers cannot be nulled and sidelobe level is increased. Performance results show that a reuse factor of 1 is not feasible in AMPS, but reuse factors of 4 and 3 can be achieved with uniform linear arrays (ULA) with five and eight elements, respectively.

Mogensen et al. [20] concentrate on the design and performance of the frequency-hopping GSM network using conventional beamforming. Most of the results are based on simulated and measured data of eight-element ULA. The simulated C/I improvement follows closely the theoretical gain at low azimuth spreads. In urban macrocells the C/I gain is reduced from the theoretical value 9 dB down to approximately 5.5–7.5 dB. The designed direction of arrival (DoA) algorithm is shown to be very robust to cochannel interference. The potential capacity enhancement is reported to be threefold in a $\frac{1}{3}$ reuse FHGSM network for an array size of $M = 4$–6.

4. INTERFERENCE CANCELLATION IN TIME DOMAIN

For the purpose of this section, the overall received signal, which is a superposition of N cochannel components received through M antennas, can be represented in the simplified case, when all signals are received bit synchronously as

$$\begin{aligned} \boldsymbol{r} &= (r^{(1)}, r^{(2)}, r^{(3)}, \ldots, r^{(M)})^{-1} \\ r^{(m)} &= \sum_{n=1}^{N} a^{(n)} h^{(m,n)} + n^{(m)} \\ h^{(n,m)} &= c^{(m,n)} * g \end{aligned} \tag{17}$$

where $a^{(n)}$ is data of user n, g is the pulseshape, and $c^{(m,n)}$ is the channel transfer function between the cochannel signal source n and receiving antenna m. The corresponding vector representation is

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{a} + \boldsymbol{n} \tag{18}$$

where $\boldsymbol{a}$ and $\boldsymbol{n}$ are vectors with components $a^{(n)}$ and $n^{(m)}$, respectively, and $\boldsymbol{H}$ is the matrix with elements $h^{(n,m)}$. In a real situation the received cochannel signals are not bit synchronous, and (18) should be modified to include two additional components representing the impact of the previous and subsequent bits accordingly, similar to asynchronous CDMA detectors [21]. Details may be found in the paper by Valenti and Woerner [22]. In accordance with Eq. (14), the system capacity can be increased by using more advanced demodulation techniques that provide the same quality of service (QoS) with lower $(C/I)_s$. A number of algorithms have been presented in the literature.

When used for DSCDMA systems, the objective of multiuser detection (MUD) is primarily to jointly detect signals that originate from the same cell (intracell interference) because the most critical interference comes from there. This is quite the opposite of the situation in a TDMA network where the interest is focused on the interference coming from the adjacent cells.

An optimum detector has been introduced [23] as a joint demodulator of cochannel signals. This detector is based on already known solutions for optimum single-user detection with intersymbol interference and Gaussian noise for *M-input/M-output* (MIMO) systems [24,25] and joint maximum-likelihood sequence estimation (MLSE) [26]. A similar approach is used for CDMA systems [27]. Practical results for the Japanese PDC system [28] and for GSM [29] have been shown. The joint MLSE for a hybrid CDMA/TDMA based on the GSM system has also been presented [30].

MLSE-type detectors are so complex and impractical that a number of blind detector algorithms must be considered. These algorithms do not require knowledge of the other signal parameters. See the "Further Reading" section for further information about blind CCI cancellation.

The latest results include cochannel interference suppression with successive cancellation [31], which is a technique already well established in CDMA systems. Performance results show that the cancellation succeeds poorly when the signal levels are comparable. Timing differences can be used for initial signal separation in order to improve performance. Soft subtraction provides further improvement.

When M signals $r^{(m)}$ from Eq. (17) are combined by using maximal ratio combining, then a reliable estimate of the channel coefficients is required. One of the latest references dealing with this problem is that by Grant and Cavers [32].

The most recent development in Turbo decoding has inspired research in the field of iterative multiuser detection, macrodiversity combining, and decoding for the TDMA cellular uplink [22]. In this approach, as the first step, each base station (BS) in a cluster of cochannel cells performs soft-output multiuser detection of the desired signal (originating from its cell) and the interfering

signals (originating from other cells in the cluster). So the multiuser detector will produce a loglikelihood ratio (LLR) for each mobile in the cluster. These LLRs for each user are then summed up across the BS cluster, which in effect produces a diversity combining signal. After that the signal may be deinterleaved and decoded. If the decoder also produces soft outputs, this may be reinterleaved and fed back to the multiuser detector to be used as a priori information in the next iteration. Once again one should be aware that the system places an additional burden on the backhaul links. Since soft information is now shared among BSs more capacity is needed on the links between BSs and the base station controller (BSC).

BIOGRAPHIES

Savo Glisic is Professor of Electrical Engineering at the University of Oulu, and Director of Globalcomm Institute for Telecommunications. He was a Visiting Scientist at Cranfield Institute of Technology, Cranfield, England (1976/77) and the University of California at San Diego (1986/87). He has been active in the field of spread-spectrum and wireless communications for 20 years and has published a number of papers and five books. He is doing consulting in this field for industry and government. He has served as Technical Program Chairman of The Third IEEE ISSSTA'94, The 8 IEEE PIMRC'97, and IEEE ICC'01. Dr Glisic was Director of IEEE ComSoc MD programs.

Pekka Pirinen received the M.S. and Lic.Tech. degrees in electrical engineering from the University of Oulu, Oulu, Finland, in 1995 and 1998, respectively. He started his career as Research Assistant in the Telecommunication Laboratory, University of Oulu, in 1994. Since 1995, he has been with the Telecommunication Laboratory and Centre for Wireless Communications, University of Oulu, as a Research Scientist in various wireless communication research projects, including the European ACTS project FRAMES. His research interests are focused on multiple access techniques, radio network planning, modeling, capacity, and performance evaluation issues. Mr. Pirinen is currently pursuing a doctoral degree in electrical engineering at the University of Oulu.

BIBLIOGRAPHY

1. V. H. MacDonald, The cellular concept, *Bell Syst. Tech. J.* **58**: 15–41 (1979).
2. I. Katzela and M. Nagshineh, Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey, *IEEE Pers. Commun.* **3**: 10–31 (1996).
3. S. W. Halpern, Reuse partitioning in cellular systems, *IEEE Trans. Vehic. Technol.* **32**: 322–327 (1983).
4. P. M. Blair, G. C. Polyzos, and M. Zorzi, Plane cover multiple access: A new approach to maximizing cellular system capacity, *IEEE J. Select. Areas Commun.* **19**: 2131–2141 (2001).
5. R. Coombs and R. Steele, Introducing microcells into macrocellular networks: A case study, *IEEE Trans. Commun.* **47**: 568–576 (1999).
6. M. D. Yacoub, Fading distributions and co-channel interference in wireless systems, *IEEE Antennas Propag. Mag.* **42**: 150–159 (2000).
7. G. L. Stüber, *Principles of Mobile Communication*, Kluwer, Norwell, MA, 1996.
8. A. A. Abu-Dayya and N. C. Beaulieu, Outage probabilities in the presence of correlated lognormal interferers, *IEEE Trans. Vehic. Technol.* **43**: 164–173 (1994).
9. L. F. Fenton, The sum of log-normal probability distributions in scatter transmission systems, *IRE Trans. Commun.* **CS-8**: 57–67 (1960).
10. R. Prasad and A. Kegel, Improved assessment of interference limits in cellular radio performance, *IEEE Trans. Vehic. Technol.* **40**: 412–419 (1991).
11. S. Schwartz and Y. S. Yeh, On the distribution function and moments of power sums with log-normal components, *Bell Syst. Tech. J.* **61**: 1441–1462 (1982).
12. R. Muammar and S. C. Gupta, Cochannel interference in high-capacity mobile radio systems, *IEEE Trans. Commun.* **30**: 1973–1978 (1982).
13. G. Falciasecca, C. Caini, G. Riva, and M. Frullone, General approach for the comparison of spectrum efficiency of digital mobile radio systems, *Eur. Trans. Telecommun.* **5**: 77–83 (1994).
14. W. C. Y. Lee, Spectrum efficiency in cellular, *IEEE Trans. Vehic. Technol.* **38**: 69–75 (1989).
15. M.-S. Alouini and A. J. Goldsmith, Area spectral efficiency of cellular mobile radio systems, *IEEE Trans. Vehic. Technol.* **48**: 1047–1066 (1999).
16. G. K. Chan, Effects of sectorization on the spectrum efficiency of cellular radio systems, *IEEE Trans. Vehic. Technol.* **41**: 217–225 (1992).
17. S. C. Swales, M. A. Beach, D. J. Edwards, and J. P. McGeehan, The performance enhancement of multibeam adaptive base station antennas for cellular land mobile radio systems, *IEEE Trans. Vehic. Technol.* **39**: 56–67 (1990).
18. J. H. Winters, Smart antennas for wireless systems, *IEEE Pers. Commun.* **5**: 23–27 (1998).
19. P. Petrus, R. B. Ertel, and J. H. Reed, Capacity enhancement using adaptive arrays in an AMPS system, *IEEE Trans. Vehic. Technol.* **47**: 717–727 (1998).
20. P. E. Mogensen et al., Performance of adaptive antennas in FH-GSM using conventional beamforming, *Wireless Pers. Commun.* **14**: 255–274 (2000).
21. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press (1998).
22. M. C. Valenti and B. D. Woerner, Iterative multiuser detection, macrodiversity combining, and decoding for the TDMA cellular uplink, *IEEE J. Select. Areas Commun.* **19**: 1570–1583 (2001).
23. W. van Etten, Maximum likelihood receiver for multiple channel transmission systems, *IEEE Trans. Commun.* **COM-24**: 276–283 (1976).
24. D. Forney, Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363–378 (1972).
25. G. Ungerboeck, Adaptive maximum likelihood receiver for carrier modulated data transmission systems, *IEEE Trans. Commun.* **COM-22**: 624–636 (1974).

26. K. Giridhar et al., Nonlinear techniques for the joint estimation of cochannel signals, *IEEE Trans. Commun.* **COM-45**: 473–484 (1997).

27. S. Verdú, Minimum probability of error for asynchronous Gaussian multiple access channels, *IEEE Trans. Inform. Theory* **IT-32**: 85–96 (1986).

28. H. Yoshino, K. Fukawa, and H. Suzuki, Interference canceling equalizer (ICE) for mobile radio communication, *IEEE Trans. Vehic. Technol.* **46**: 849–861 (1997).

29. S. W. Wales, Technique for cochannel interference suppression in TDMA mobile radio systems, *IEE Proc. Commun.* **142**: 106–114 (1995).

30. J. Blanz, A. Klein, M. Nasshan, and A. Steil, Performance of a cellular hybrid C/TDMA mobile radio system applying joint detection and coherent receiver antenna diversity, *IEEE J. Select. Areas Commun.* **12**: 568–579 (1994).

31. H. Arslan and K. Molnar, Cochannel interference suppression with successive cancellation in narrow-band systems, *IEEE Commun. Lett.* **5**: 37–39 (2001).

32. S. J. Grant and J. K Cavers, Multiuser channel estimation for detection of cochannel signals, *IEEE Trans. Commun.* **49**: 1845–1855 (2001).

FURTHER READING

Cochannel Interference Distributions

Abu-Dayya A. A. and N. C. Beaulieu, Outage probabilities of cellular mobile radio systems with multiple Nakagami interferers, *IEEE Trans. Vehic. Technol.* **40**: 757–768 (1991).

Cardieri P. and T. S. Rappaport, Statistical analysis of co-channel interference in wireless communications systems, *Wireless Commun. Mobile Comput.* **1**: 111–121 (2001).

French R. C., The effect of fading and shadowing on channel reuse in mobile radio, *IEEE Trans. Vehic. Technol.* **VT-28**: 171–181 (1979).

Ho C.-L., Calculating the mean and variance of power sums with two log-normal components, *IEEE Trans. Vehic. Technol.* **44**: 756–762 (1995).

Lee C.-C. and R. Steele, Signal-to-interference calculations for modern TDMA cellular communication systems, *IEE Proc. Commun.* **142**: 21–30 (1995).

Prasad R. and A. Kegel, Effects of Rician faded and log-normal shadowed signals on spectrum efficiency in microcellular radio, *IEEE Trans. Vehic. Technol.* **42**: 274–281 (1993).

Punt J. B. and D. Sparreboom, Summing received signal powers with arbitrary probability density functions on a logarithmic scale, *Wireless Pers. Commun.* **3**: 215–224 (1996).

Safak A., Statistical analysis of the power sum of multiple correlated log-normal components, *IEEE Trans. Vehic. Technol.* **42**: 58–61 (1993).

Schleher D., Generalized Gram-Charlier series with application to the sum of lognormal variates, *IEEE Trans. Inform. Theory* **23**: 275–280 (1977).

Outage Probability

Caini C., G. Immovilli, and M. L. Merani, Outage probability for cellular mobile radio systems: Simplified analytical evaluation and simulation results, *Electron. Lett.* **28**: 669–671 (1992).

Caini C., G. Immovilli, and M. L. Merani, Outage probability in FDMA/TDMA mobile communication networks, *Eur. Trans. Telecommun.* **5**: 59–68 (1994).

Immovilli G. and M. L. Merani, Simplified evaluation of outage probability for cellular mobile radio systems, *Electron. Lett.* **27**: 1365–1367 (1991).

Linnartz J.-P. M. G., Exact analysis of the outage probability in multiple-user mobile radio, *IEEE Trans. Commun.* **40**: 20–23 (1992).

Sowerby K. W. and A. G. Williamson, Outage probability calculations for multiple cochannel interferers in cellular mobile radio systems, *Proc. IEE Commun.* **135**: 208–215 (1988).

Sowerby K. W. and A. G. Williamson, Outage probability calculations for mobile radio systems with multiple interferers, *Electron. Lett.* **24**: 1073–1075 (1988).

Sowerby K. W. and A. G. Williamson, Outage probabilities in mobile radio systems suffering cochannel interference, *IEEE J. Select. Areas Commun.* **10**: 516–522 (1992).

Yeh Y.-S. and S. C. Schwartz, Outage probability in mobile telephone due to multiple log-normal interferers, *IEEE Trans. Commun.* **32**: 380–388 (1984).

Spectrum Efficiency

Clark M. V., V. Erceg, and L. J. Greenstein, Reuse efficiency in urban microcellular networks, *IEEE Trans. Vehic. Technol.* **46**: 279–288 (1997).

Nagata Y. and Y. Akaiwa, Analysis for spectrum efficiency in single cell trunked and cellular mobile radio, *IEEE Trans. Vehic. Technol.* **35**: 100–113 (1987).

Prasad R. and J. C. Arnbak, Comments on "Analysis for spectrum efficiency in single cell trunked and cellular mobile radio," *IEEE Trans. Vehic. Technol.* **37**: 220–222 (1988).

Spatial Filtering

Au W. S., R. D. Murch, and C. T. Lea, Comparison between the spectral efficiency of SDMA systems and sectorized systems, *Wireless Pers. Commun.* **16**: 15–67 (2001).

Godara L. C., Applications of antenna arrays to mobile communications, Part I: Performance improvement, feasibility, and system considerations, *Proc. IEEE* **85**: 1031–1060 (1997).

Howitt I. and Y. M. Hawwar, Evaluation of outage probability due to cochannel interference in fading for a TDMA system with a beamformer, *Proc. IEEE Vehicular Technology Conf.*, 1998, 520–524.

Litva J. and T. K.-Y. Lo, *Digital Beamforming in Wireless Communications*, Artech House, Boston, (1996).

Paulraj A. J. and C. B. Papadias, Space-time processing for wireless communications, *IEEE Signal Process. Mag.* **14**: 49–83 (1997).

Wang L.-C., K. Chawla, and L. J. Greenstein, Performance studies of narrow-beam trisector cellular systems, *Int. J. Wireless Inform. Networks* **5**: 89–102 (1998).

Winters J. H., Optimum combining in digital mobile radio with cochannel interference, *IEEE Trans. Vehic. Technol.* **VT-33**: 144–155 (1984).

Zetterberg P. and B. Ottersten, The spectrum efficiency of a base station antenna array system for spatially selective transmission, *IEEE Trans. Vehic. Technol.* **44**: 651–660 (1995).

Zetterberg P., A comparison of two systems for downlink communication with base station antenna arrays, *IEEE Trans. Vehic. Technol.* **48**: 1356–1370 (1999).

Interference Cancellation in Time Domain

Batra A. and J. R. Barry, Blind cancellation of co-channel interference, *Proc. IEEE Global Telecommunications Conf.*, 1995, pp. 157–162.

Berangi R. and P. Leung, Indirect cochannel interference cancelling, *Wireless Pers. Commun.* **19**: 37–55 (2001).

Fukawa K. and H. Suzuki, Blind interference canceling equalizer for mobile radio communications, *IEICE Trans. Commun.* **E77-B**: 580–588 (1994).

Grant S. J. and J. K. Cavers, Performance enhancement through joint detection of cochannel signals using diversity arrays, *IEEE Trans. Commun.* **46**: 1038–1049 (1998).

Lo B. C. W. and K. B. Letaief, Adaptive equalization and interference cancellation for wireless communication systems, *IEEE Trans. Commun.* **47**: 538–545 (1999).

Ranta P. A., A. Hottinen, and Z.-C. Honkasalo, Co-channel interference cancelling receiver for TDMA mobile systems, *Proc. IEEE Int. Conf. Communications*, 1995, pp. 17–21.

Ranta P. A., Z.-C. Honkasalo, and J. Tapaninen, TDMA cellular network application of an interference cancellation technique, *Proc. IEEE Vehicular Technology Conf.*, 1995, pp. 296–300.

CODE-DIVISION MULTIPLE ACCESS

Branimir R. Vojčić
Raymond L. Pickholtz
George Washington University
Washington, District of Columbia

1. INTRODUCTION

Multiple-access communications is a means by which many individual, geographically dispersed users access a shared medium and/or resources in order to transmit/receive information. Multiple access is used for local-area networks (LAN), satellite and cellular terrestrial radio networks, and other applications. Conventional multiple access schemes include random access such as ALOHA and its successor CSMA/CD, which is used in LANs, and structured orthogonal signals multiple access such as *frequency-* and *time-division multiple access* (FDMA, TDMA), which are used in satellite systems and terrestrial cellular radio. *Code-division multiple access* (CDMA) is a multiuser communications method, which uses spread-spectrum signals with uniquely addressable signature waveforms that permit the separation of each signal at the receiver. The main paradigm change in spread spectrum is that all users use the entire available spectrum simultaneously. It is possible, in some instances, to arrange for the signature waveforms to be orthogonal at the receiver, so that this separation can be affected by means of a linear correlator, wherein the desired users' signal is extracted while the other users' signals are completely suppressed. However, even when orthogonality is not perfect, for design or practical reasons, the spread-spectrum processing gain causes undesired multiuser signals to be significantly suppressed. In conventional CDMA receivers, such linear correlators are used and, to the extent that the *multiple-access interference* (MAI) is suppressed by the processing gain, it is tolerated. Naturally, this effect results in the characteristic "MAI-limited" channel whose capacity in terms of number of users is thereby limited according to what performance objective is specified. The two principal CDMA approaches are based on *direct-sequence* (DS) spreading and *frequency hopping* (FH). The FH transmission is conceptually very similar to conventional narrowband schemes, except that the carrier frequency is changed pseudorandomly over the spread spectrum bandwidth. In addition, it appears that DSCDMA is a less costly approach for commercial applications. Consequently, in our discussion we will mainly address DSCDMA and provide only a brief discussion of FHCDMA.

A major shortcoming of conventional detection CDMA systems is that, since each user contributes interference in proportion to their received power level, users that either generate excessive power, or whose power is received as larger than the desired signal (e.g., by virtue of being close to the receiver), degrade performance. This effect, sometimes called the "near–far problem," is a major impediment to practical CDMA using conventional detectors. The near–far problem is usually mitigated by exercising tight, closed-loop power control on all users so that all the received signals are of equal power at the receiver.

In Section 2 we first address the fundamental principles of CDMA using the notion that spread spectrum may be viewed as a way of embedding a signal into a high-dimensionality signal space and show how conventional CDMA receivers deal with MAI and its effects on both performance and user capacity. Next, in Section 3, we introduce some well-established performance measures by which, in addition to *signal-to-noise-and-interference ratio* (SNIR) and *bit-error probability* (BER), we can assess the behavior of multiuser systems. In Section 4, we examine the effect of the near–far problem and demonstrate, by an example, the consequences of imperfect power control. The obvious question is that since all the "signature" waveforms are presumable known at the receiver, why must we tolerate the MAI as if they were not known? Indeed, the optimum, maximum likelihood multiuser detector attempts to demodulate all received signals *jointly*. In Section 5, we will examine both the performance and complexity of this optimum approach and subsequently examine several suboptimal schemes, which exhibit significantly reduced complexity, and their performance. It appears that some of the multiuser detection schemes are practically feasible and, as such, can be exploited to improve the CDMA capacity, especially in situations in which tight power control is not achievable.

2. DIMENSIONALITY, PROCESSING GAIN, AND MULTIPLE ACCESS

A fundamental issue in CDMA is how this technique affords multiple simultaneous transmissions using a common bandwidth. The underlying principle is that of distributing relatively low-dimensional data signals in a high-dimensional environment. This is accomplished by means of spreading codes (signature waveforms), unique for each user so that all multiple-access signals are mutually nearly orthogonal. The idea of using quasiorthogonal signature waveforms (noiselike waveforms) for multiple access is originally due to Claude Shannon [2,3]. He perceived it as a democratic way of sharing the frequency spectrum: "If more people (signals) were there (in the

crowded radio spectrum), gradually the noise level would increase on each channel. But everyone could still talk, even though it might be a pretty noisy 'cocktail party' by that time" (this is now called a graceful degradation in CDMA).

In the "standard" problem of digital transmission, the set of M signaling waveforms $\{s_i(t), 0 \le t \le T, 1 \le i \le M\}$, known to both transmitter and receiver, is used to transmit $(\log_2 M)/T$ bps (bits per second). If, for example, $s_i(t)$ is sent, the received signal is $r(t) = s_i(t) + n_w(t)$, $0 \le t \le T$, where T denotes the symbol duration and $n_w(t)$ is additive, white Gaussian noise (AWGN) with two-sided power spectral density $N_0/2$ W/Hz. It is well known that the signal set can be completely specified by a linear combination $D \le M$ orthogonal basis functions. The dimensionality D, of the signal waveforms, is approximately equal to $2B_DT$, where B_D is the total (approximate) spectral occupancy of the employed signal set [4]. If the total *available* bandwidth is B_N, corresponding to an N-dimensional signal space, the maximum number of simultaneously active users, each one using D dimensionality, with orthogonal multiplexing is $K = N/D$. With quasiorthogonal multiplexing, however, it is possible to accommodate more than K users in the same bandwidth, but with some mutual multiple-access interference. In addition to sharing the bandwidth, the quasiorthogonal users share interference as well. The quasiorthogonal multiplexing is usually accomplished by means of *pseudonoise* (PN) spreading (signature) sequences, which have desired cross-correlation properties. At least one sequence is available to the cooperating transmitter and receiver, which may or may not know the PN sequences employed by other transmitter/receiver pairs.

A general model, which conveys the idea of CDMA, is as follows. Consider K simultaneous binary antipodal ($D = 1$) transmissions embedded in an N-dimensional signal space. Assuming jointly synchronous transmission, the aggregate of all transmitted signals can be represented by

$$x(t) = \sum_{i=1}^{K} x_i(t), \quad 0 \le t \le T \tag{1}$$

where the transmitted signal of the ith user is

$$x_i(t) = \sqrt{W_i} b_i s_i(t), \quad 0 \le t \le T \tag{2}$$

where W_i represents the signal energy per bit, b_i is the bit value (± 1) and $s_i(t)$ is the unit energy signature waveform of the ith user, defined as

$$s_i(t) = \sum_{k=1}^{N} s_{ik} \varphi_k(t), \quad 0 \le t \le T \tag{3}$$

where

$$s_{ik} = \int_0^T s_i(t) \phi_k(t)\, dt \tag{4}$$

and where $\{\phi_k(t), 1 \le k \le N\}$ is an orthonormal basis spanning the space:

$$\int_0^T \phi_l(t) \phi_m(t)\, dt = \delta_{lm} = \left\{ \begin{array}{ll} 1, & l = m \\ 0, & l \ne m \end{array} \right\} \tag{5}$$

Note that each binary antipodal signal requires one dimension ($D = 1$) for transmission, while N dimensions are employed to generate K distinct signaling waveforms. The factor N is usually called *spreading factor* or *processing gain*. The term spreading factor reflects the fact that the actual transmission bandwidth B_N (Fourier bandwidth), is N times larger than the Shannon bandwidth[1] of the modulated signal. The latter term, processing gain, stems from the capability of the spread signal to suppress interference by exploiting spectral redundancy and will be discussed subsequently.

In general, the PN sequence of the ith user, $\{s_{il}, \ldots, s_{iN}\}$, is chosen so as to have minimal possible cross-correlation with PN sequences of other users $\{s_{jl}, \ldots, s_{jN}\}, j = 1, \ldots, K$ and $j \ne i$. Here we assume, for the time being, that the sequences are random such that

$$E[s_{ij}] = 0, \quad \forall i, j, \quad i = 1, \ldots, K, \quad j = 1, \ldots, N \tag{6}$$

$$E[s_{il} s_{im}] = \frac{1}{N} \delta_{im}, \quad \forall i \tag{7}$$

$$E[s_{il} s_{jm}] = 0, \quad i \ne j, \quad \forall l, m \tag{8}$$

Although the spreading sequences are generated randomly, they are known to the communicators (at least to communicating pairs).

Consider next the conventional DSCDMA receiver, where the received signal is given by $r(t) = x(t) + n_w(t)$. The output of the ith correlation receiver is given by

$$\begin{aligned} U_i &= \int_0^T r(t) s_i(t)\, dt \\ &= \sum_{k=1}^{n} \left(\sqrt{W_i} b_i + \sum_{\substack{i=1 \\ l \ne i}}^{K} \sqrt{W_l} b_l s_{lk} s_{ik} + s_{ik} n_k \right), \end{aligned} \tag{9}$$

where

$$n_k = s_{ik} \int_0^T n_w(t) \phi_k(t)\, dt. \tag{10}$$

In conventional CDMA receivers the multiple access interference (MAI) at the output of correlation receiver is tolerated and a decision is made according to $b_i = \text{sgn}(U_i)$. Under these circumstances, a measure of performance which is monotonically related to the bit error rate (BER) is the SNIR, which can be expressed as

$$\begin{aligned} \text{SNIR}_i &= \frac{E^2[U_i \mid b_i]}{Var[U_i \mid b_i]} \\ &= \left[\frac{N_0}{2W_i} + \frac{1}{NW_i} \sum_{\substack{i=1 \\ l \ne i}}^{K} W_l \right]^{-1} \end{aligned} \tag{11}$$

The first term is due to thermal noise, the second term is due to MAI. It can be seen from the second term that

[1] By Shannon bandwidth we mean one-half the minimum number of orthogonal functions per T seconds that are required in a basis for a signal space in which signal can be represented [12].

MAI is suppressed by the factor N (spreading factor), which explains why the spreading factor is often called the processing gain. If we let $W_i/N_0 \to \infty$ (MAI dominates) and assume $W_i = W_l, \forall i, l$, then

$$\lim_{W_i/N_0 \to \infty} \text{SNIR}_i = \frac{N}{K-1} \tag{12}$$

This last result reveals the fundamental difference between orthogonal and quasiorthogonal multiplexing when conventional, single-user detection is employed; even for vanishingly small noise, the SNIR is finite. Hence, in a "cocktail party" of this example, to use Shannon's words, the number of simultaneous transmissions is $K \leq N/\text{SNIR}_{\text{REQ}} (K \gg 1)$, where SNIR_{REQ} corresponds to the desired transmission quality. Depending upon whether $\text{SNIR}_{\text{REQ}} < 1$ or $\text{SNIR}_{\text{REQ}} > 1$, more than N or less than N users, respectively, can transmit simultaneously.

Consider now the probability of error when the sign decision on U_i is employed. For large N we can invoke the central-limit theorem and assume that the MAI is Gaussian. Then it is easy to see that the conditional probability of error for the ith user is given by

$$P(i)\Big|_{\rho_{il}} = \frac{1}{2^{K-1}} \sum_{\text{all } b_l} Q\left[\sqrt{\frac{2W_i}{N_0}}\left(1 - \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} b_l \rho_{il}\right)\right] \tag{13}$$

where the first summation is over all possible combinations of data bits of interfering users and

$$\begin{aligned} \rho_{il} &= \int_0^T s_i(t) s_l(t)\, dt \\ &= \sum_{k=1}^{n} s_{ik} s_{lk} \end{aligned} \tag{14}$$

is the (random) cross-correlation between signature waveforms of the ith and lth user signature waveforms. For vanishingly small N_0, the probability of error is dominated by the term corresponding to the worst combination of interfering bits, so that

$$P(i)_{\text{wc}}\Big|_{\rho_{il}} = \frac{1}{2^{K-1}} Q\left[1 - \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} |\rho_{il}|\right]. \tag{15}$$

Since random spreading sequences were assumed, we obtain the following, after averaging over all signature waveforms, according to Jensen's inequality:

$$P(i)_{\text{wc}} \geq \frac{1}{2^{K-1}} Q\left[\sqrt{\frac{2W_i}{N_0}}\left(1 - \frac{1}{N} \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}}\right)\right]. \tag{16}$$

Hence, we can see that if

$$\left(1 - \frac{1}{N} \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}}\right) \leq 0$$

$P(i)_{\text{wc}}$ does not vanish as $W_i/N_0 \to \infty$. This phenomenon is not inherently a CDMA characteristic, but is rather a consequence of suboptimum detection. We will return to this in a subsequent section.

In a synchronous CDMA system, it is possible to choose spreading sequences, which are mutually orthogonal, as long as the number of signals does not exceed the dimensionality of the signal space. In that case we have orthogonal multiplexing (MAI free) and the probability of error is given by

$$P(i)_{\text{ort}} = Q\left(\sqrt{\frac{2W_i}{N_0}}\right) \tag{17}$$

Again with the Gaussian assumption[2] on MAI which we will relate below. Moreover, this is possible to achieve as long as the spreading sequences of K users are linearly independent [11]. At this point we abandon the analysis tool of random signature sequences and merely assume that they have *known* cross-correlations.

We now show that the signals can be processed at both transmitter and receivers so that the $\text{sgn}(U_i)$ is optimal and individual probabilities of error are given by Eq. (17).

Consider the vector of sufficient statistics (vector of correlation receivers' outputs in (9) for demodulation of all K signals jointly), given by

$$\mathbf{U} = \mathbf{RWb} + \mathbf{N}_w \tag{18}$$

where $\mathbf{R} = \{\rho_{ij}\}_{K \times K}$ is the cross-correlation matrix of signature waveforms $\mathbf{W} = \text{diag}(\sqrt{W_i})$ is a diagonal $K \times K$ matrix of signal amplitudes, $\mathbf{b}$ is the $K \times 1$ data vector, and $\mathbf{N}_w$ is a $K \times 1$ zero-mean Gaussian noise vector with covariance matrix $\mathbf{R}_N = \mathbf{R}N_0/2$. Hence, each component of the vector $\mathbf{U}$ contains the desired signal component, multiuser interference and a Gaussian noise component. When the matrix $\mathbf{R}$ is positive-definite, which is equivalent to the linear independence of K signature waveforms, there exist a unique Cholesky decomposition [18] of the matrix $\mathbf{R}$, such that $\mathbf{R} = \mathbf{G}^T\mathbf{G}$, where $\mathbf{G}$ is an upper triangular matrix. Consider a linear transformation $\mathbf{G}^{-1}$ at the transmitter such that the transmitted signal $x(t)$ is given by

$$x(t) = \mathbf{s}(t) T \mathbf{G}^{-1} \mathbf{W} \mathbf{b} \tag{19}$$

where $\mathbf{s}(t)$ is $K \times 1$ vector of signature waveforms. Then the vector of correlation receiver outputs is given by

$$\mathbf{U} = \mathbf{G}^T \mathbf{W} \mathbf{b} + \mathbf{N}_w \tag{20}$$

and by applying the linear transformation $(\mathbf{G}^T)^{-1}$ on the vector $\mathbf{U}$ in the receiver, we obtain

$$\mathbf{U}_0 = \mathbf{W}\mathbf{b} + \mathbf{Z} \tag{21}$$

where $\mathbf{Z}$ is a $K \times 1$ zero-mean Gaussian noise vector with covariance matrix $\mathbf{R}_Z = \mathbf{I}N_0/2$ and where $\mathbf{I}$ is the identity

[2] The additive noise is always assumed to be Gaussian in this regard.

matrix. Hence, by means of a pair of linear transformations $\mathbf{G}^{-1}$ and $(\mathbf{G}^T)^{-1}$, in the transmitter and the receiver, respectively, K users are decoupled and the resultant noise vector $\mathbf{Z}$ is white. Thus, individual sign decisions on the components of $\mathbf{U}_0$ will be optimal. Moreover, it is easy to see that the total transmit energy per bit interval is independent of the vector $\mathbf{b}$ and is equal to $W_{\text{tot}} = \sum_{i=1}^{K} W_i$, as it would be if orthogonal multiplexing were used in the first place. Indeed, the linear transformation $\mathbf{G}^{-1}$ in the transmitter gives rise to K orthogonal signature waveforms $\mathbf{p}(t)^T = \mathbf{s}(t)^T\mathbf{G}^{-1}$ and, similarly, by employing a bank of corresponding matched filters in the receiver, an orthogonal *multiuser communication system* results. A block diagram of the resulting end-to-end system is shown in Fig. 2. Actually, one could use the signature waveforms $\mathbf{p}(t)$ in the first place, or another set of orthogonal waveforms, in many multiple access scenarios, but there may be advantages to forcing this condition. Other decompositions of the correlation matrix, such as $\mathbf{R} = \mathbf{R}^{1/2}\mathbf{R}^{1/2}$ and $\mathbf{R} = \mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix, yield the same result as Cholesky decomposition, but the Cholesky decomposition is preferred due to its numerical stability. An adaptive scheme for joint transmitter–receiver optimization suitable for asynchronous channels is described elsewhere [38].

The described method of coordinated linear transformations in the transmitter and receiver in synchronous multiuser channels orthogonalizes the multiple access users, with no transmit power or noise enhancement penalties whatsoever, when users are linearly independent. It will be always possible to choose $\mathbf{s}(t)$ to have linearly independent users when the number of users is such that $K \leq N$. For $K > N$, this will not be possible for then the matrix $\mathbf{R}$ is singular so that the described coordinated transmitter/receiver linear transformations are not feasible. However, the use of joint transmitter/receiver processing with other criteria is not precluded.

At this point we would like to generalize the assertion presented above to the case where modulated signals may have different Shannon bandwidths but occupy the same Fourier bandwidth, such that the spreading factor of the ith user is $\gamma_i, i = 1, \ldots, K$. This generalization is succinctly summarized by the following proposition [12].

Proposition 1. Suppose that K users send their modulated signals to a single receiver, using a common Fourier bandwidth. Then zero interuser interference (IUI) is possible at the receiver only if the users transmit spread-spectrum signals whose spreading factors satisfy

$$\sum_{i=1}^{K} \frac{1}{\gamma_i} \leq 1 \tag{22}$$

It should be noted that in many practical multiple access channels, characterized by multipath propagation and loss of synchronization, for example, the single-user performance (no IUI) may not be achievable even when the conditions stated above are satisfied.

3. PERFORMANCE MEASURES

The SNIR and BER, introduced in the previous section, are the most common performance measures in digital communications. In multiuser communications, these performance measures very often do not admit analytic evaluation, whereas some asymptotic performance measures (corresponding to vanishingly small noise) may be readily found. To facilitate the comparison of different schemes in subsequent sections, we introduce the *asymptotic multiuser efficiency* (AME) and *near–far resistance*, originally proposed by Verdú [17].

Definition 1. The asymptotic multiuser efficiency of a multiuser detector, characterized by the probability of error for the ith user equal to $P(i)$, is given by

$$\eta_i = \sup\left\{ 0 \leq r \leq 1 : \lim_{N_0 \to 0} \frac{P(i)}{Q\left(\sqrt{r\frac{2W_i}{N_0}}\right)} < +\infty \right\} \tag{23}$$

or equivalently

$$\eta_i = \lim_{N_0 \to 0} \frac{W_{eff}(i)}{W_i} \tag{24}$$

where $W_{eff}(i)$ represents the effective signal energy of the ith user, reduced by the presence of MAI, such that the corresponding probability of error can be expressed as $P(i) = Q(\sqrt{2W_{eff}(i)N_0})$.

Definition 2. The near–far resistance of a multiuser detector for the ith user represents the minimum asymptotic efficiency over the relative energies of all other users:

$$\overline{\eta}_i = \inf_{\substack{W_j > 0 \\ j \neq i}} \eta_i \tag{25}$$

The AME measures the slope at which $P(i)$ goes to 0 in the high signal-to-noise ratio region for a given set of signal amplitudes of the desired and interfering users. On the other hand, the near–far resistance represents the AME for the worst-case combination of interfering signal amplitudes relative to the desired signal amplitude.

In the next sections, we use the AME and the near–far resistance to demonstrate the relative performance of conventional CDMA compared to optimal, or near-optimal, CDMA detectors.

4. CDMA, NEAR–FAR EFFECT, AND POWER CONTROL

Consider again the synchronous CDMA system of Section 2, whereby users employ *deterministic* spreading sequences with the cross-correlation between the ith and the jth spreading waveform given by ρ_{ij}. The probability of error for the ith user, when the conventional correlation receiver is employed, similarly as in (13), is given by

$$P(i) = \frac{1}{2^{K-1}} \sum_{\text{all } b_l} Q\left[\sqrt{\frac{2W_i}{N_0}}\left(1 - \sum_{\substack{l=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} b_l \rho_{il}\right)\right] \tag{26}$$

The corresponding AME is found as

$$\eta_i = \left[\max \left(1 - \sum_{\substack{i=1 \\ l \neq i}}^{K} \sqrt{\frac{W_l}{W_i}} |\rho_{il}| \right) \right]^2 \tag{27}$$

It can be easily seen from (27) that the AME of the conventional detector can take the value 0 for a user with relatively small W_i. Indeed, the near–far resistance of the; conventional receiver is 0 if at least one $\rho_{il} \neq 0, i \neq l, il = 1, \ldots, K$. The situation in which a strong interferer overwhelms the desired signal is usually referred to as the near–far effect. To provide the same performance to all receivers, power control is used so as to have all received signal energies the same ($W_i = W_l, \forall i, l$). The effect of imperfect power control on conventional CDMA, when the aggregate MAI was approximated by equivalent Gaussian noise,[3] was thoroughly analyzed in an earlier study [19]. To illustrate the near–far effect and the effect of power control error, consider a two-user example.

Let the cross-correlation between signature waveforms of users 1 and 2 be equal to ρ. Then, the probability of error for user 1 is given by

$$P(1) = \frac{1}{2} Q \left[\sqrt{\frac{2W_1}{N_0}} \left(1 - \sqrt{\frac{W_2}{W_1}} \rho \right) \right] + \frac{1}{2} Q \left[\sqrt{\frac{2W_1}{N_0}} \left(1 + \sqrt{\frac{W_2}{W_1}} \rho \right) \right]$$

and the corresponding AME is given by

$$\eta_1 = \left[\max \left(1 - \sqrt{\frac{W_2}{W_1}} \rho \right) \right]^2$$

Hence, we can see that $\eta_1 = 0$ for $W_2/W_1 \geq 1/\rho^2$. To provide the same performance to both users, we need $W_1 = W_2$, in which case $\eta_i = (1 - \rho)^2, i = 1, 2$.

To illustrate the effect of power control error, without examining specific power control mechanisms, we consider a simplified model of power control error. Let us assume that the received energy per bit for the ith user is $\alpha_i W_i, i = 1, 2$, where α_1 and α_2 are independent and identically distributed with the probability mass function given by

$$f_\alpha(\alpha_l) = \left\{ \begin{array}{ll} \Lambda, & \text{with probability } \dfrac{1}{4} \\ 1, & \text{with probability } \dfrac{1}{2} \\ \dfrac{1}{\Lambda}, & \text{with probability } \dfrac{1}{4} \end{array} \right\}, \quad i = 1, 2$$

[3] It should be noted that the Gaussian approximation for the MAI yields satisfactory performance estimation accuracy for the conventional correlation receiver, for low to moderate values of W_i/N_0 and when the number of interfering users and/or the processing gain are relatively large. This should not be mistaken for near optimality of the correlation receiver in the presence of MAI, which is in fact non-Gaussian.

Then, it is easy to calculate the AME in the presence of the power control error, by comparing the two-user performance with the single-user performance with the same distribution of the power control error; it is given by

$$\eta_1 = \left[\max \left(0, 1 - \sqrt{\frac{W_2}{W_1}} \rho \Lambda \right) \right]^2$$

The detrimental effect of power control error is apparent. Essentially, as can be seen from the formula for the AME, the power control error has an equivalent effect on the performance as an increase in the correlation coefficient between signature waveforms or/and an increase in the number of users; thus, it eats up the available capacity. In the next section we will see that multiuser detectors do not exhibit such a sensitivity to the imbalance in received signal amplitudes, and that the performance/capacity can be significantly improved, even compared to the conventional detector with perfect power control. Moreover, some multiuser detectors achieve optimal performance when the received energies are quite dissimilar.

5. MULTIUSER DETECTION AND INTERFERENCE CANCELLATION

As indicated in the previous section, the near–far effect is detrimental to conventional CDMA. This is a consequence of suboptimum detection, which ignores the interference from other users. When interfering signals are accounted for in the detection process, the adverse near–far problem can be eliminated. Moreover, these schemes, which we refer to as *multiuser detection* or *interference cancellation*, provide better performance than does a conventional correlation receiver, even when all signals arrive at the same power level at the common receiver (no near–far effect). The correlation receiver (optimum for AWGN channels) was often considered near optimum, based on the conjecture that the aggregate multiple access interference is approximately Gaussian, which may not be a good approximation for a finite-user population with dissimilar power levels. The non-Gaussian nature of the MAI is what enables optimum or near-optimum multiuser detectors to outperform, in some cases significantly, the conventional single-user correlation receiver. To obtain some insight into possible performance improvements and the incurred complexity, we will discuss some characteristic multiuser detection schemes. More detailed surveys of multiuser detection and interference cancellation schemes can be found in the literature [13,29,42].

Without loss of generality, we continue to use the synchronous model introduced in Section 2. The likelihood function (conditional probability density function of the channel output given the binary data vector $\mathbf{b}$ from all the K users) is given by

$$f_{r(t)}[r(t), t \in [0, T] | \mathbf{b}] = C \exp \left\{ -\frac{1}{N_0} \int_0^T \left[r(t) - \sum_{i=1}^{K} \sqrt{W_i} b_i s_i(t) \right]^2 dt \right\}. \tag{28}$$

The maximum-likelihood multiuser detector selects the most likely hypothesis given observation, choosing $\hat{b}$, which maximizes the likelihood function:

$$\hat{b} = \arg \min_{b \in \{-1,1\}^K} \int_0^T \left[r(t) - \sum_{i=1}^{K} \sqrt{W_i} b_i s_i(t) \right]^2 dt$$
$$= \arg \min_{b \in \{-1,1\}^K} (2\mathbf{U}^T - \mathbf{b}^T \mathbf{W}\mathbf{R})\mathbf{W}\mathbf{b} \tag{29}$$

To make a decision, 2^K hypotheses must be examined. Hence, the optimum multiuser detector has exponential complexity in the number of users, which may be restrictive for moderate and large values of K. In the asynchronous case, the optimum detector is the maximum-likelihood sequence detector operating on the sequences of MF outputs, again characterized with exponential complexity with respect to the number of users. It is the complexity of the optimum detector that has motivated research for suboptimum multiuser detectors with polynomial complexity.

One of the best-known multiuser detectors with linear complexity in the number of users is the *decorrelator* detector, first proposed in 1979 [16], but later thoroughly analyzed and correctly characterized [17]. The decorrelator detector follows immediately from the vector of sufficient statistics in (20). By applying the linear transformation $\mathbf{T} = \mathbf{R}^{-1}$ on the vector of sufficient statistics, we obtain

$$\mathbf{U}_0 = \mathbf{R}^{-1}\mathbf{U} = \mathbf{W}\mathbf{b} + \mathbf{Z} \tag{30}$$

where the transformed noise vector has the covariance matrix $\mathbf{R}_Z = \mathbf{R}^{-1}N_0/2$ and the MAI is decoupled. The receiver is completed by applying the sign rule, $\hat{\mathbf{b}} = \text{sgn}(\mathbf{U}_0)$, which is not optimal because the noise vector is not white. The MAI is completely eliminated at the expense of noise enhancement. It should be noted that this detector does not need the knowledge of received amplitudes, unlike the optimum detector, and achieves optimum near–far resistance. Moreover, the decorrelator is the maximum likelihood solution when signal amplitudes are not known. The probability of error for the decorrelator detector has the simplest form of all multiuser detectors and is given by

$$P_d(i) = Q\left(\sqrt{\frac{2W_i}{N_0} \frac{1}{R_{ii}^1}} \right) \tag{31}$$

where R_{ii}^1 represents the ith diagonal element of $\mathbf{R}^{-1}$. The corresponding AME is

$$\eta_d(i) = \frac{1}{R_{ii}^1} \tag{32}$$

It can be shown that the AME of the decorrelator detector is bounded, when the signature waveforms are linearly independent, according to [15]

$$\frac{4\lambda_{\max}\lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2} \le \frac{1}{R_{ii}^1} \le 1 \tag{33}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ represent the minimum and maximum eigenvalues of $\mathbf{R}$, respectively. Hence, as the system becomes heavily loaded and/or eigenvalue spread increases, the lower bound on the asymptotic efficiency decreases. When the matrix R becomes singular (loss of linear independence of signature waveforms), the near–far resistance of the decorrelator detector becomes equal to 0.[4] A generalization of the decorrelator detector to the asynchronous case can be found elsewhere [20].

For the two-user case of Example 1, the AME of the decorrelator detector is $\eta_d = 1 - \rho^2$ while the AME of the optimum detector was obtained [17] as $\eta_{ml} = \min[1, 1 + W_2/W_1 - 2\rho\sqrt{W_2/W_1}]$. In Fig. 1 we compare the AMEs of these two multiuser detectors with that of the conventional receiver; in all cases, perfect power control was assumed and $\rho = 0.7$. The advantage of multiuser detectors over the conventional receiver is apparent, even in the absence of the near–far effect and power control error ($W_1 = W_2$). It is a simple exercise to see that the AME of the decorrelator detector remains unchanged in the presence of power control error, and hence, the near–far resistance of the optimum detector does not degrade, either.

Another linear multiuser detector that is closely related to decorrelator detector is the *minimum mean-square-error* (MMSE) detector, originally proposed in the context of multiuser detection for asynchronous channels [21]. For the synchronous case, it is defined by Proposition 2:

Proposition 2. The MMSE detector for the synchronous multiuser system corresponding to the vector of sufficient statistics in (18) is defined by

$$\mathbf{T}^* = \left(\mathbf{R} + \frac{N_0}{2}\mathbf{W}^{-2} \right)^{-1} \tag{34}$$

$$\hat{\mathbf{b}} = \mathbf{T}^*\mathbf{U} \tag{35}$$

Proof: The MMSE criterion leading to the desired estimator is given by

$$\mathbf{T}^* = \arg \min_{\mathbf{T} \in R^{K*K}} E_{\mathbf{b},\mathbf{N}} \|\mathbf{T}\mathbf{U} - \mathbf{b}\|^2 \tag{36}$$

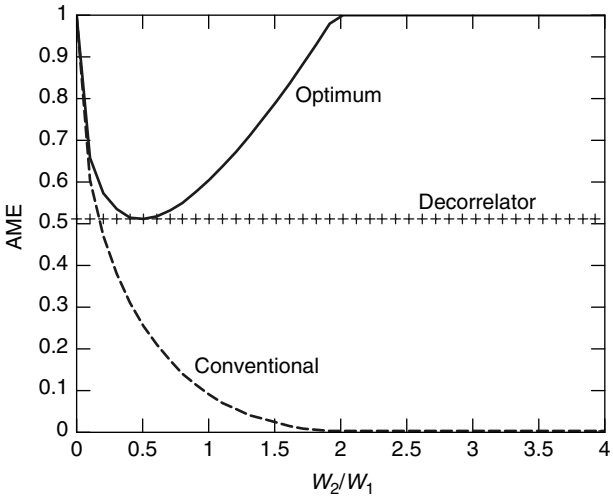


Figure 1. Comparison of AME for decorrelator, conventional, and optimum detectors.

[4] When signature waveforms are not linearly independent, no multiuser detector is near–far-resistant, in the synchronous case [17].

By applying the orthogonality principle, the optimum detector is obtained from

$$E[(\mathbf{TU} - \mathbf{b})\mathbf{U}^T] = 0 \tag{37}$$

from which we obtain

$$\mathbf{T}^* = \mathbf{W}^{-1}\left(\mathbf{R} + \frac{N_0}{2}\mathbf{W}^{-2}\right)^{-1} \tag{38}$$

and since the sign decision on $\mathbf{T}^*\mathbf{U}$ suffices, the factor $\mathbf{W}^{-1}$ is irrelevant and (34) follows. It should be noted that for vanishingly small noise the MMSE detector tends to the decorrelator detector, that is $\lim_{N_0 \to 0} \mathbf{T}^* = \mathbf{R}^{-1}$.

The most important feature of the MMSE detector is its suitability for adaptive implementation, whereby no information about interfering signals is required. Only the timing of the desired signal and a training sequence is required for the adaptive receiver to converge to its optimum setting. This adaptive MMSE detector was analyzed, in various forms, in Refs. 22–24 and references cited therein. Since the adaptive MMSE receiver does not need the knowledge of the signal attributes of interfering users, it represents a natural choice for mobile receivers in cellular or packet radiocommunications. Its performance in the steady state is superior to that of a conventional CDMA receiver, especially in near–far scenarios. Even in the presence of perfect power control, this receiver achieves roughly 2 times larger communication capacity than the conventional receiver [23].

A blind adaptive multiuser detector, closely related to the adaptive MMSE receiver was proposed [25]. The main advantage of this blind receiver is in that the training mode is not required and only an approximate knowledge of the signature waveform of the desired user is needed. The latter characteristic is important in mobile channels in which transmitted waveforms usually suffer from distortion. For a survey of adaptive multiuser detection schemes, the reader is referred to Verdu [26].

Finally, we would like to discuss an important class of multiuser detectors that is based on the cancellation of the estimated MAI in a feedback fashion [27–29]. Although nonlinear in structure, these detectors are characterized by linear complexity in the number of users. A simplified block diagram of one such detector, proposed by Varanasi and Aazhang, is shown in Fig. 2; for simplicity the first two detection stages are shown completely only for user 1.

This multistage detector employs the decorrelator receiver in the first stage to get an initial estimate of interfering bits. These bit estimates are multiplied by the corresponding correlation coefficients and amplitudes of interfering signals to reconstruct the MAI. The estimated MAI is subtracted from the matched filter output of the corresponding user, user 1, and a new decision is made on the thus obtained decision statistics. Since the matched filter output is used to obtain the decision statistics for the second stage, the noise enhancement effect of the decorrelator detector is not explicitly present in later stages. However, the effect of noise enhancement in the first stage propagates into subsequent stages through the tentative first stage decisions. The original version of the multistage detector employed conventional receivers in the first stage [27]. The version with the decorrelator

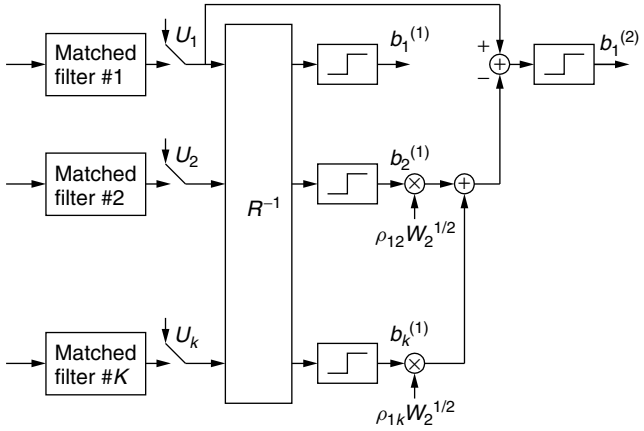


Figure 2. Block diagram of a two-stage detector with the decorrelator receiver in the first stage.

detector in the first stage achieves better performance and admits simpler analysis. The tentative decisions in the first stage are obtained by applying the sign decision on the vector of decision statistics defined in (30), that is, $\mathbf{b}_{(l)} = \text{sgn}(\mathbf{R}^{-1}\mathbf{U}_0)$. The vector of decision statistics for the second stage is formed as

$$\mathbf{Y} = \mathbf{Wb} + (\mathbf{R} - \mathbf{I})\mathbf{W}(\mathbf{b} - \hat{\mathbf{b}}) \tag{39}$$

The second-stage decisions are made according to $\mathbf{b}^{(2)} = \text{sgn}(\mathbf{Y})$. It is apparent that this detector results in the isolated transmission performance in the second stage when the first-stage tentative decisions are perfect. However, when the decision errors are made in the first stage, the corresponding interference terms double, thus adversely affecting the decisions in the second stage.

Returning to our two-user example, it is easy to show, by exploiting the results in Ref. 28, that the probability of error for user 1 in the second stage is given by

$$\begin{aligned} P(1)^{(2)} = {} & Q\left(\sqrt{\frac{2W_1}{N_0}}\right) \times \left[1 - Q\left(\sqrt{\frac{2W_2(1-\rho^2)}{N_0}}\right)\right] \\ & + \frac{1}{2}Q\left(\sqrt{\frac{2W_2(1-\rho^2)}{N_0}}\right) \\ & \times \left\{Q\left[\sqrt{\frac{2W_1}{N_0}}\left(1 - 2\rho\sqrt{\frac{W_2}{W_1}}\right)\right]\right. \\ & \left. + Q\left[\sqrt{\frac{2W_1}{N_0}}\left(1 + 2\rho\sqrt{\frac{W_2}{W_1}}\right)\right]\right\} \end{aligned}$$

It can be shown that the AME in this case is

$$\eta_{\text{ms}} = \min(\alpha, 1)$$

where

$$\alpha = (1-\rho^2)\frac{W_2}{W_1} + \left[\max\left(0, 1 - 2\rho\sqrt{\frac{W_2}{W_1}}\right)\right]^2$$

To minimize the effect of doubling the interference when wrong decisions are made, soft interference cancellation

can be employed. The idea here is to make soft tentative decisions at the decorrelator output in such a way that decisions are somehow weighted according to their reliability. It was shown [30] that two-stage detection with soft interference cancellation can significantly outperform its hard-interference cancellation counterpart, and in the two-user case achieves optimum AME. Specifically, for $K = 2$, it was shown that a linear clipper as a soft weighting nonlinearity achieves the optimum AME when the threshold of the clipper is chosen according to

$$\delta\left(\rho, \sqrt{\frac{W_2}{W_1}}\right) = \sqrt{W_1}\max\left[0, \frac{|\rho|\left(\sqrt{\frac{W_1}{W_2}} - |\rho|\right)}{1-\rho^2}\right] \quad (40)$$

Hence, when the decorrelator output is larger than δ, a hard decision is made, otherwise the decorrelator output is scaled proportionally to the linear part of the clipper, before feedback cancellation. In the K user case the soft limiting nonlinearity is optimized on a pairwise basis according to (40). For numerical results on possible improvements with soft interference cancellation, the reader is referred to Ref. 30. Similar performance improvements for the asynchronous case have been demonstrated [32].

BIOGRAPHIES

Branimir R. Vojčić is a professor in, and chairman of, the Department of Electrical and Computer Engineering at the George Washington University, Washington, D.C. He has received his D.Sc. degree from the University of Belgrade, Yugoslavia. His current research interests are in the areas of communication theory, performance evaluation and modeling mobile and wireless networks, code division multiple access, multiuser detection, adaptive antenna arrays and space-time coding and ad-hoc networks. He has also been an industry consultant in these areas and has published and lectured extensively in these areas. Dr. Vojcic is a senior member of IEEE, was an associate editor for IEEE Communications Letters and a receipient of 1995 National Science Foundation CAREER Award.

Raymond L. Pickholtz is a professor in, and former chairman of, the Department of Electrical and Computer Engineering at The George Washington University, Washington, D.C., and received his Ph.D. in electrical engineering from the Polytechnic Institute of Brooklyn, New York. He was an editor of the *IEEE Transactions on Communications*, and guest editor for special issues on computer communications, military communications spread spectrum systems and social impacts of technology. He is currently the coeditor of chief of the *Journal of Communications and Networks*. He has published scores of papers (several award winning), acts as a consultant to industry, and holds six U.S. patents.

Dr. Pickholtz is a fellow of the IEEE, AAAS, and the Washington Academy of Sciences. He was elected president of the IEEE Communications Society in 1991. He received the Donald W. McLellan Award in 1994. He was a visiting Erskine fellow at the University of Canterbury, Christchurch, New Zealand, 1997. He was awarded the IEEE Third Millenium Medal in 2000.

BIBLIOGRAPHY

1. A. Viterbi, *CDMA Principles of Spread Spectrum Communications*, Addison-Wesley, Reading, MA, 1995.
2. J. R. Pierce, A conversation with Claude Shannon, *IEEE Commun. Mag.* **22**: 123–126 (May 1984).
3. J. Costas, Poisson, Shannon and the radio amateur, *IEEE Proc.* **47**: 2058–2068 (Dec. 1959).
4. J. M. Wozencraft and I. M. Jacobs, *Principles of Communications Engineering*, Waveland Press, Prospect Heights, IL, 1990.
5. R. Dixon, *Spread Spectrum Systems*, Willey-Interscience, New York, 1984.
6. M. Simon, J. Omura, R. Scholtz, and B. Levitt, *Spread Spectrum Communications Handbook*, McGraw-Hill, New York, 1995.
7. R. Peterson, R. Ziemer, and D. Borth, *Introduction to Spread Spectrum Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
8. J. Proakis, *Digital Communications*, McGraw-Hill, New York, 1995.
9. R. Pickholtz, D. Schilling, and L. Milstein, Theory of spread spectrum communications — a tutorial, *IEEE Trans. Commun.* **COM-30**(5): (May 1982).
10. R. Pickholtz, D. Schilling, and L. Milstein, Theory of spread spectrum communications — a tutorial (revisions), *IEEE Trans. Commun.* **COM-32**(2): (Feb. 1984).
11. B. R. Vojcic and R. L. Pickholtz, Joint transmitter receiver optimization in synchronous multiuser communications, Information Theory Workshop, Rydzyna, Poland, 1995.
12. J. L. Massey, Spectrum spreading and Multiple accessing, Information Theory Workshop, Rydzyna, Poland, 1995.
13. S. Verdu, Recent progress in multiuser detection, in N. Abramson, ed., *Multiple Access Communications*, IEEE Press, New York, 1993.
14. A. Duel-Hallen, J. Holtzman, and Z. Zvonar, Multiuser detection for CDMA systems, *IEEE Pers. Commun.* **2**: 46–58 (April 1995).
15. R. Lupas, *Near-Far Resistant Linear Multiuser Detection*, Ph.D. thesis, Princeton Univ., Princeton, NJ, 1989.
16. K. S. Schneider, Optimum detection of code division multiplexed signals, *IEEE Trans. Aerospace Electric Syst.* **AES-15**: 181–185 (Jan. 1979).
17. R. Lupas and S. Verdu, Linear multiuser detectors for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **IT-34**: (1988).
18. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1983.
19. B. R. Vojcic, R. L. Pickholtz, and L. B. Milstein, Performance of DS-CDMA with imperfect power control operating over a low earth orbiting satellite link, *IEEE J. Select. Areas Commun.* **12**: (May 1994).
20. R. Lupas and S. Verdu, Near–far resistance of multiuser detectors in asynchronous channels, *IEEE Trans. Commun.* **COM-38**: 496–508 (April 1990).

21. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multiuser communications, *IEEE J. Select. Areas Commun.* 683–690 (May 1990).

22. P. B. Rapajic and B. S. Vucetic, Adaptive receiver structures for asynchronous CDMA systems, *IEEE J. Select. Areas Commun.* 685–697 (May 1994).

23. S. L. Miller, An adaptive direct-sequence code-division multiple-access receiver for multiuser interference rejection, *IEEE Trans. Commun.* 1746–1755 (Feb.–April 1995).

24. U. Madhow and M. L. Honig, MMSE interference suppression for direct-sequence spread-spectrum CDMA, *IEEE Trans. Commun.* 3178–3188 (Dec. 1994).

25. M. Honig, U. Madhow, and S. Verdu, Blind adaptive multiuser detector, *IEEE Trans. Inform. Theory* 944–960 (July 1995).

26. S. Verdu, Adaptive multiuser detection, *Proc. IEEE Int. Symp. Spread Spectrum Theory and Applications*, Oulu, Finland, July 1994.

27. M. Varanasi and B. Aazhang, Multistage detection in asynchronous code-division multiple-access communications, *IEEE Trans. Commun.* **COM-38**(4): (April 1990).

28. M. Varanasi and B. Aazhang, Near-optimum detection in Synchronous CDMA systems, *IEEE Trans. Commun.* **COM-39**: (May 1991).

29. A. Duel-Hallen, Decorrelating decision-feedback multiuser detector for asynchronous code-division multiple-access channel, *IEEE Trans. Commun.* **COM-41**: 285–290 (1993).

30. V. Vanghi and B. Vojcic, Soft interference cancellation in multiuser communications, *Int. J. Wireless Pers. Commun.* (Special Issue on Signal Separation and Interference Cancellation for Personal, Indoor and Mobile Radio Communications), **3**: 111–128 (1996).

31. P. Patel and J. Holtzman, Performance comparison of a DS/CDMA system using a successive interference cancellation (IC) scheme and a parallel IC scheme under fading, *Int. Conf. Communication*, New Orleans, 1994, pp. 510–514.

32. X. Zhang and D. Brady, Soft-decision multistage detection for asynchronous AWGN channels, *Proc. 31st Annual Allerton Conf. Communication, Control and Computing*, Allerton House, Urbana-Champaign, IL, Oct. 1993.

33. B. R. Vojcic, Transmitter precoding for synchronous multiuser communications, Workshop on Mobility Management, George Mason University, Fairfax, VA, Oct. 1994.

34. B. R. Vojcic, Transmitter precoding in multiuser communications, *Proc. 1995 IEEE IT Workshop on Information Theory, Multiple Access and Queueing*, St. Louis, April 1995.

35. Z. Tang and S. Cheng, Interference cancellation for DS-CDMA systems over flat fading channels through predecorrelating, *Proc. PIMRC'94*, Hague, The Netherlands, 1994.

36. Y. Yasuda, K. Kashiki, and Y. Hirata, High-rate punctured convolutional codes for soft decision Viterbi decoding, *IEEE Trans. Commun.* **COM-32**: 315–319 (March 1984).

37. B. Vojcic and R. Pickholtz, Spectral shaping in DS CDMA on a satellite link, *Proc. AIAA Conf.*, Washington, DC, Feb. 1996.

38. P. Rapajic and B. Vucetic, Linear adaptive transmitter-receiver structures for asynchronous CDMA systems, *Eur. Trans. Telecommun.* **6**: 21–27 (Jan.–Feb. 1995).

39. W. M. Jang and B. Vojcic, Transmitter precoding in synchronous multiuser communications over multipath channels, *Proc. Symp. Interference Rejection and Signal Separation in Wireless Communications*, New Jersey Institute of Technology, Newark, NJ, March 1996.

40. J. Hui, Throughput analysis for code division multiple accessing of the spread spectrum channel, *IEEE J. Select. Areas Commun.* **SAC-2**: 482–486 (July 1984).

41. M. Pursley, Performance evaluation of phase-coded spread spectrum multiple-access communication — system analysis, *IEEE Trans. Commun.* **COM-25**: 795–799 (Aug. 1977).

42. S. Verdu, *Multiuser Detection*, Cambridge Univ. Press, Cambridge, UK, 1998.

CODING FOR MAGNETIC RECORDING CHANNELS

LIH-JYH WENG
Maxtor Corporation
Shrewsbury, Massachusetts

1. INTRODUCTION

Since 1990, the areal magnetic recording density for rigid disks has been growing at a rate of 60% annually and the trend is accelerating [1]. This is one of the most important factors for the magnetic recording capacity of more recent rigid disks to double every nine months. To maintain this pace of density increase and to meet the stringent requirement of today's digital storage systems in both the data integrity and recording density increase, error-correcting codes (ECCs) become both indispensable in achieving the low postdecoding bit error rate and effective in overall recording density optimization. An ECC encoder first encodes the user data into ECC codewords; these code words are then mapped, using a modulation code, into a form suitable for the write circuit to record the encoded data to the disk surfaces. During a read, the process is reversed; the readback signal is first processed to recover the codewords for the modulation code; the role of the ECC is to correct all errors that may have occurred during the entire read/write process. Figures 1 and 2 give the logical flow of the entire process. The main purpose of the ECC is to ensure that the user data in Fig. 1 are identical to the user data of Fig. 2. In more conventional communication systems, the main emphasis is to communicate reliably

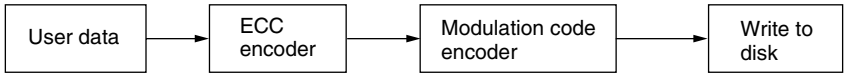


Figure 1. Logic flow of writing data to disks.

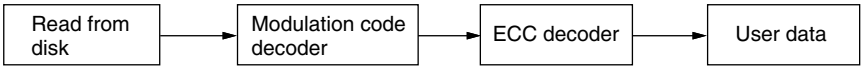


Figure 2. Logic flow of retrieving data from disks.

from one point to another with an acceptable amount of delay. On the other hand, disk and tape drives serve the purpose of storing data at one time instant and reading them back reliably at another instant.

Some of the modern modulation codes have the capability of performing a limited amount of error correction by themselves [2]. However, the error rate at the output of the modulation code decoder, in most cases, does not achieve the stringent low error rates required of present storage systems. An ECC is a simple and effective way to bring the error rate at the output of the modulation code decoder to the level specified by the user. Furthermore, with the separation of the modulation code and ECC, each component can be more effectively designed to achieve the overall optimality of the entire recording system.

2. REED–SOLOMON CODES: ORIGIN AND BASIC CHARACTERISTICS

In the early days, the recording density was relatively low, and, hence, the main concern was a single burst in a sector. The single-burst correcting codes and codes that correct a small number of errors were the dominant ECCs at the time. Reference 2 is a good source for early ECC implementations. When the recording density increases, the most likely errors are not confined to single bursts; therefore, codes that can deal with both random and burst errors are necessary to adequately protect the written data. Reed–Solomon codes have such characteristics and also possess efficient encoding and decoding algorithms, which can be readily implemented. Consequently, they are presently the most widely used codes for magnetic disk drives. All the techniques discussed here are applicable to rigid disk drives and tape drives as well as optical drives. When complexity is a main concern, tape drives also use other coding techniques. On tape drives, the error bursts tend to be much longer than those observed in disk drives. Some structure is usually introduced among code blocks to deal with long bursts. Specifically, tape drives often employ so-called two-dimensional codes. This can be pictured with data arranged in a two-dimensional array; both the rows and the columns of the array are protected by ECCs, or some sort of redundancy is introduced both vertically and horizontally. Furthermore, redundancies can be introduced diagonally or along any line of a given slope in the array. The main reason for using codes along more than two directions is to use very simple codes, such as single-parity-check binary codes along each direction to provide sufficient protection. Array codes [4] provide such advantages.

The Reed–Solomon codes employed in recording systems need to satisfy some special constraints, which may be different from those in other applications. In the following discussion, several important considerations for applying Reed–Solomon codes in magnetic recording systems are addressed. First, a simple example is presented to introduce some terminology commonly used in Reed–Solomon codes as well as more generally in ECCs. Other related issues concerning the code applications to magnetic recording such as implementation, block synchronization, interleaving, performance, and error detection are discussed. Finally, some remarks concerning tape drives, RAID (reductant array of independent disks) systems, soft decoding and increasing the sector size are also addressed.

2.1. An Example of a Reed–Solomon Code

A Reed–Solomon (10,3,8) code over Galois Field code GF(2^4) is selected for illustration purposes. The meaning of the parameters 10, 3, and 8 will be given shortly. First, the field is a Galois field, commonly denoted as GF(2^m), where m is 4 in this example. This field is used to perform all arithmetic operations needed for the Reed–Solomon code. Therefore, the definition and essential properties for the GF(2^m) should be given first. GF(2^m) is often referred to as an *extension field* of GF(2) = {0, 1}. A more concrete way of viewing a GF(2^m) field is to list the field elements as all possible m-bit binary representation of the integer $0, 1, 2, \ldots, 2^m - 1$. For this example, $m = 4$, the field elements of GF(2^4) are the set {0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111}. A field needs two basic operations: addition (+) and multiplication (∗) (symbol ∗ used to indicate Galois field multiplication). The addition of two elements (a, b, c, d) and (e, f, g, h) are defined to be $(a + e, b + f, c + g, d + h)$, where a, b, c, d, e, f, g, h are either one or zero, the elements of GF(2) = {0, 1}. The same symbol + is used for addition over GF(2) and addition over GF(2^m). The addition over GF(2) is modulo-2 addition, which follows the rule that $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, and $1 + 1 = 0$. For example, $(1010) + (1100) = (0110)$. The multiplication rule over GF(2) is defined by $0 * 0 = 0 * 1 = 1 * 0 = 0$, and $1 * 1 = 1$. In more familiar engineering terms, the addition is equivalent to an EXOR (exclusive OR) operation and the multiplication is the same as the AND operation commonly seen in logic. To define the multiplication rules for GF(2^m), an irreducible binary polynomial of degree m is needed. A degree m binary polynomial, where the coefficients are either 0 or 1, is said to be irreducible if it is not divisible by any binary polynomial of degree lower than m except the trivial degree 0 polynomial, specifically, the constant 1. In the case of $m = 4$, $p(x) = x^4 + x + 1$ is an irreducible polynomial. Each element of GF(2^m) is associated with a unique polynomial of degree $m - 1$ or lower. The elements of GF(2^4) can be expressed in two convenient ways: GF(2^4) = {0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, …} = $\{0, x^0, x, x + 1, x^2, x^2 + 1, x^2 + x, x^2 + x + 1, x^3, x^3 + 1, \ldots\}$. Let $a(x)$ and $b(x)$ be two elements of GF(2^m); then the multiplication rule is given by $c(x) = a(x) * b(x) \bmod p(x)$. Since the degree of $p(x)$ is m, the degree of $c(x)$, which is the remainder of $a(x) * b(x)$ divided by $p(x)$, is less than m. Next an element of GF(2^m) is selected whose powers generate every possible nonzero element of the field by multiplying itself many times. In this selected field, any of the following nonzero elements can be used as such a generating element: $x, x^2, x + 1, x^2 + 1, x^3 + 1, x^3 + x + 1, x^3 + x^2 + 1$, and $x^3 + x^2 + x$. Let the generating element be x, then $x^0 = 1, x^1 = x, x^2 = x^2, x^3 = x^3, x^4 = x + 1, x^5 = x * x^4 = x * (x + 1) = x^2 + x, x^6 = x * x^5 = x * (x^2 + x) = x^3 +$

$x^2, x^7 = x * x^6 = x * (x^3 + x^2) = x^4 + x^3 = x + 1 + x^3 = x^3 + x + 1, x^8 = x * x^7 = x * (x^3 + x + 1) = x^4 + x^2 + x = x + 1 + x^2 + x = x^2 + 1, x^9 = x * x^8 = x * (x^2 + 1) = x^3 + x,$ .... Expressed as powers of the generating element, the field elements are written as $\mathrm{GF}(2^4) = \{0000, 0001, 0010, 0100, 1000, 0011, 0110, 1100, 1011, 0101, 1010, \ldots\} = \{0, x^0, x^1, x^2, x^3, x^4, x^5, x^6, x^7, x^8, x^9, \ldots\}$. In this case, all the nonzero elements of the field can be expressed as distinct powers of x. There are exactly $2^m - 1$ distinct powers $x^j \bmod p(x)$ for $j = 0, 1, 2, \ldots, 2^m - 2$. The class of polynomials, powers of whose roots generate all nonzero field elements, is collectively called *primitive polynomials*. A nonprimitive polynomial can also be used to form the field. For example, the polynomial $p'(x) = x^4 + x^3 + x^2 + x + 1$ is irreducible but not primitive. This can be easily checked by the fact that $x^4 = x^3 + x^2 + x + 1 \bmod p'(x)$ and $x^5 = x * x^4 = x * (x^3 + x^2 + x + 1) = x^4 + x^3 + x^2 + x = (x^3 + x^2 + x + 1) + x^3 + x^2 + x = 1 \bmod p'(x)$. Since x is not a primitive element for the field defined by $p'(x)$, a different element must be used. Let $\alpha = x + 1$, then the successive powers of $\alpha^j \bmod p'(x)$ for $j = 0, 1, 2, \ldots, 2^m - 2$, are all distinct. Therefore, another $\mathrm{GF}(2^4)$ can be generated by $p'(x)$ using α^j as a nonzero element. It is convenient to represent $\mathrm{GF}(2^m)$ as $\{0, \alpha^0, \alpha^1, \alpha^2, \alpha^3, \ldots\}$, where α is a primitive element. For the primitive irreducible $p(x)$ shown above, $\alpha = x$. For all primitive polynomials, the generating element α can be selected to be x. For the remaining discussion, the binary primitive polynomial $p(x)$ is used. The generating element α is sometimes referred to as the *primitive element*.

With α^j defined for $j = 0, 1, 2, \ldots, 2^m - 2$, the product of two elements α^i and α^j can be obtained in at least two ways: (1) express α^i and α^j as polynomials, find the product as the polynomial multiplication of $\alpha^j * \alpha^j \bmod p(x)$; (2) using the simple exponent addition rule, namely $\alpha^i * \alpha^j = \alpha^k$ with $k = i + j \bmod 2^m - 1$. There are $2^m - 1$ nonzero distinct elements, each corresponding to a distinct power $0, 1, 2, \ldots, 2^m - 2$. It should be noted that $\alpha^s = \alpha^0$ if $s = 2^m - 1$. The identity element for $\mathrm{GF}(2^m)$ multiplication is α^0. The identity element for $\mathrm{GF}(2^m)$ addition is the zero element. In $\mathrm{GF}(2^m)$, the subtraction operation is the same as the addition operation and the "division" a/b can be considered as $a * (b^{-1})$ for any nonzero element b. The element b^{-1} can be obtained from b by examining the exponent of b; namely, if $b = \alpha^j$, then $b^{-1} = \alpha^k$ such that $j + k = 0 \bmod 2^m - 1$. For example, in the example of $\mathrm{GF}(2^4)$, to find the inverse of $x^3 + x$, which is α^9, the exponent is 9; it is a simple matter to determine that $9 + 6 = 15 = 0 \bmod 2^4 - 1$. Therefore, $(\alpha^9)^{-1} = \alpha^6$. References 5 and 6 provide more detailed properties of Galois fields.

In software or firmware implementations, the most common method of computing the inverse of a field element is by making use of a logarithm table and an antilogarithm table. The logarithm table associates every nonzero element with an exponent; for example, in the $\mathrm{GF}(2^4)$ above, the logarithm table gives the elements 0100 and 1001 their respective exponents 3 and 14. On the other hand, the antilogarithm table takes the exponents 3 and 14 as inputs and produces the elements 0100 and 1001, respectively. The approach of finding the inverse is very similar to the real-number computation using a logarithm table and an antilogarithm table [6]. In a hardware implementation, the logarithm table and the antilogarithm table are seldom provided. To find the inverses, a special algorithm is employed; the algorithm is often dependent on the symbol size m selected and sometimes depends on the irreducible polynomial generating the field [7,8].

2.2. Reed–Solomon Encoder

A Reed–Solomon code can now be defined over the field $\mathrm{GF}(2^m)$. One way to specify a Reed–Solomon code is to define the roots of the generator polynomial of the code. The generator polynomial can be written as

$$g(x) = (x + \alpha^L) * (x + \alpha^{L+1}) * (x + \alpha^{L+2}) * \cdots * (x + \alpha^{L+R-1}) \tag{1}$$

This is the generator polynomial of a Reed–Solomon (n, k, d) code, where n is the code length, k is the number of information symbols, and d is the minimum distance (or Hamming distance) among all possible codewords, where the distance between two codewords is equal to the number of symbols at which these two codewords differ. The code rate of a code is defined as the ratio k/n, which is a number between 0 and 1. In magnetic recording, high-rate codes are frequently employed. The value L can be selected arbitrarily; n is at most $2^m - 1$ for easy hardware implementation; and $d = n - k + 1 = R + 1$, where R is equal to the number of redundant symbols of the code or the degree of the generator polynomial. In other words, there are n symbols in a codeword, among which k symbols can be assigned arbitrarily as information symbols or data symbols. The degree of the generator polynomial is equal to $R = n - k$. The one requirement for the roots is that they must be consecutive roots $\alpha^L, \alpha^{L+1}, \alpha^{L+2}, \ldots, \alpha^{L+R-1}$. The choice of the value L does not change the code minimum distance. Let $n = 10, k = 3, R = 7$ or $d = 8$, and set $L = 11$. Then the generator polynomial is given by

$$\begin{aligned} g(x) &= (x + \alpha^{11}) * (x + \alpha^{12}) * (x + \alpha^{13}) * (x + \alpha^{14}) \\ &\quad * (x + \alpha^0) * (x + \alpha^1) * (x + \alpha^2) \\ &= \alpha^0 * x^7 + \alpha^1 * x^6 + \alpha^3 * x^5 + \alpha^{12} * x^4 + \alpha^{11} * x^3 \\ &\quad + \alpha^0 * x^2 + \alpha^{11} * x + \alpha^8 \end{aligned} \tag{2}$$

All the codewords of this Reed–Solomon (10,3,8) code can be expressed in polynomial form such as

$$c(x) = c_9 * x^9 + c_8 * x^8 + c_7 * x^7 + c_6 * x^6 + c_5 * x^5 + c_4 * x^4 + c_3 * x^3 + c_2 * x^2 + c_1 * x + c_0 \tag{3}$$

The requirements for $c(x)$ to be a codeword are (1) its degree is no higher than 9 and (2) it must be a multiple of the generator polynomial [i.e., $c(x) = m(x) * g(x)$]. To simplify the notation, the polynomial is often written as a vector:

$$c(x) = (c_9, c_8, c_7, c_6, c_5, c_4, c_3, c_2, c_1, c_0) \tag{4}$$

For example, $g(x) = (0, 0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8)$ and $x * g(x) = (0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0)$. It is clear that both $g(x)$ and $x * g(x)$ are codewords because they are both multiples of the generator polynomial $g(x)$. Another way of specifying a code is by a generator matrix. The generator matrix of this (10,3,8) code is given as

$$G = \begin{bmatrix} 0, 0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8 \\ 0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0 \\ \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0, 0 \end{bmatrix} \quad (5)$$

With the help of the generator matrix, a codeword can be expressed as

$$(c_9, c_8, c_7, c_6, c_5, c_4, c_3, c_2, c_1, c_0) = (m_2, m_1, m_0) * G \quad (6)$$

where m_2, m_1, m_0 are arbitrary elements of GF(2^m). It should be noted that not all codes possess generator polynomials or generator matrices. A class of codes, called *cyclic codes*, can be specified by generator polynomials. Another class of codes, called *linear codes*, have generator matrices. The Reed–Solomon codes are both cyclic and linear; therefore, they have rich algebraic properties and structure, which provide the foundation of being well studied and understood. As a result, Reed–Solomon codes are widely employed. As mentioned previously, the distance between two codewords is the number of symbols at which these two codewords differ. For example, the distance between $g(x)$ and $x * g(x)$ is 9 (symbols.) If all possible codewords of this Reed–Solomon (10,3,8) code are listed, every pair of distinct codewords differs in at least eight (8) symbols. Therefore, any combination of three or fewer symbol errors will not change a codeword closer to another codeword. Consequently, this code can correct all possible combinations of three or fewer errors. In addition, any error pattern of four symbol errors cannot change a codeword closer to another codeword; in the worst case, the codeword corrupted by a four-error pattern can be at the same distance from many codewords. As a result, the decoder cannot decode the erroneous code word to a unique codeword. In this case, the errors are detected but not corrected. In general, a distance d Reed–Solomon code is capable of correcting any combination of t symbol errors per code block if $t < or = \lfloor d/2 \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x ($\lfloor x \rfloor$ is referred to as the "floor function" of x). A more interesting way of employing a distance d Reed–Solomon code is to use a distance d code to correct t or fewer errors and detect with certainty if there are $t+1, t+2, \ldots, e$ symbol errors per code block provided $t + e < d$ and $t < e$ [6,9–11]. In this application, it is very important that the value t be used to determine the postdecoding symbol or block error rate and the value e be designed to ensure that the miscorrection probability meets the stringent requirement for data storage applications. It should also be noted that for highly redundant Reed–Solomon codes, the value of e can be set to be very close to or the same as t. In other words, very little or no additional symbol detection is needed to achieve the level of miscorrection probability demanded by the data storage systems.

In the last paragraph, the distance is measured by symbols. Because every symbol contains m bits, it is also possible to measure the distance in bits. For example, the Reed–Solomon (10,3,8) code over GF(2^4) is also a binary (40,12,10) code. The minimum distance of the binary code is now measured in bits. This binary code can correct four (4) random bit errors and detect five (5) random bit errors with certainty. The minimum symbol distance of a Reed–Solomon code can be easily determined as $d = n - k + 1$. However, the minimum distance of its binary expansion version cannot be determined easily except that the lower bound of the minimum distance is d, the same as the symbol minimum distance. This lower bound may not be very tight, as can be seen from the example presented above.

The encoding process of a Reed–Solomon (n, k, d) code is often achieved through a division process, which ensures that the codeword $c(x)$ is a multiple of the generator polynomial $g(x)$. It is possible to achieve this with the additional condition that the data symbols of $c(x)$ are unaltered. This class of codes is called *systematic codes*.

2.3. Reed–Solomon Decoder

In decoding, the first step is also by division. For error-detection purposes, a corrupted codeword $c'(x)$ is divided by $g(x)$. If the remainder of the division is zero, the decoder assumes that the codeword is error-free; otherwise, the decoder assumes the codeword is corrupted. Instead of dividing by the generator polynomial $g(x)$, an equivalent method is to divide the corrupted codeword by each factor $x + \alpha^{L+j}$ of $g(x)$. The results of the divisions are called syndromes S_{L+j}. Therefore, for an uncorrupted codeword, all the syndromes S_{L+j} must be zero. The syndrome polynomial $S(x)$ is a degree $R-1$ polynomial defined as

$$S(x) = S_L + S_{L+1} * x + S_{L+2} * x^2 + S_{L+3} * x^3 + \cdots + S_{L+R-1} * x^{R-1} \quad (7)$$

In most algebraic coding textbooks [6,9–11], the starting point of decoding Reed–Solomon codes is the syndrome polynomial, which contains all the necessary information to find error locations and error values. In Reed–Solomon codes over GF(2^m), the symbol positions within a code block are denoted by $\alpha^0, \alpha^1, \alpha^2, \ldots, \alpha^{n-2}, \alpha^{n-1}$, with α^p corresponding to symbol position P. The first data symbol occupies position $n-1$, and the last redundant symbol occupies position 0. The common decoding algorithms for high-rate (n, k, d) Reed–Solomon codes are divided into the following four major steps:

1. Compute the syndrome polynomial $S(x)$ from a corrupted codeword $c'(x)$.
2. Solve the key equations $\sigma(x) * S(x) = \omega(x) \bmod x^R$, where $R = n - k$ and $\sigma(x)$ is the error locator polynomial containing all the error location information and $\omega(x)$ is the error evaluator polynomial, which can be used in conjunction with the error locator polynomial to find the error values.
3. Find the roots of $\sigma(x) = (\alpha^{a1} * x + 1) * (\alpha^{a2} * x + 1) * \cdots * (\alpha^{at} * x + 1)$. The following conditions can be used to abort the decoding process: (a) if the number of roots found is less than the degree of $\sigma(x)$, (b) any

roots found does not correspond to a location between 0 and $n-1$, and (c) if repeated roots are found.

4. For each root α^{ai} of $\sigma(x)$ find the error values with the help of $\sigma(x)$, $\omega(x)$ using the Forney formula. The error value at location α^{ai} is given by $\omega(x) * x^{(L-1)} * /\sigma'(x)$ evaluated at $x = \alpha^{ai}$, where $\sigma'(x)$ is the formal derivative of $\sigma(x)$. Over GF(2^m), if $\sigma(x). = \sigma_0 + \sigma_1 * x + \sigma_2 * x^2 + \sigma_3 * x^3 + \sigma_4 * x^4 + \sigma_5 * x^5 + \sigma_6 * x^5 + \sigma_7 * x^7 + \cdots$, then $\sigma'(x). = \sigma_1 + \sigma_3 * x^2 + \sigma_5 * x^4 + \sigma_7 * x^6 + \cdots$. In other words, $\sigma'(x)$ contains only even-powered terms of x.

To complete the decoding process, the errors values computed in step 4 should be added to the corrupted code symbol indicated by the respective decoded error locations.

The most time-consuming steps are steps 1 and 3. The most difficult step to understand is step 2. Again, most coding textbooks provide detailed information about solving the key equations [6,9–11]. Two competing algorithms for solving the key equations are the Berlekamp–Massey algorithm and the Euclidean algorithm. The former may use fewer gates to implement, but the latter can be understood more easily for first-time readers. In addition, there are other decoding algorithms, which can be found in Refs. 12 and 13.

The most frequently used technique for finding the roots of the error locator polynomial is the well-known Chien search [6,9–11], which is a systematic and efficient trial-and-error root testing technique. Every possible root, one root for each possible error location, is tested as a possible solution to the error locator polynomial $\sigma(x)$. Therefore, this is usually a very time-consuming process.

It may be helpful to continue the example with errors introduced. Let the codeword be $c(x) = x * g(x) = (0, \alpha^0, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, 0)$. Let the two introduced errors be one at position 0 and one at position 8. The error value at position 0 is α^3 and the error value at position 8 is α. Then the corrupted codeword $c'(x)$ becomes $c'(x) = x * g(x) = (0, \underline{\alpha}^4, \alpha^1, \alpha^3, \alpha^{12}, \alpha^{11}, \alpha^0, \alpha^{11}, \alpha^8, \underline{\alpha}^3)$. The underlined symbols are erroneous. The first step of the decoding process is to compute the syndrome polynomial, which is $S(x) = \alpha^0 + \alpha^4 * x + \alpha^{14} * x^2 + \alpha^{13} * x^3 + \alpha^9 * x^4 + \alpha * x^5 + \alpha^6 * x^6$. Either the Berlekamp–Massey or the Euclidian algorithm will produce the error locator polynomial $\sigma(x) = 1 + \alpha^2 * x + \alpha^8 * x^2$ and the error evaluator polynomial $\omega(x) = 1 + \alpha^{10} * x$. It can be checked that $\sigma(x) * S(x) = \omega(x) \bmod x^R$. The next step is to find the roots of $\sigma(x)$, which is equivalent to the complete factorization of $\sigma(x)$. The result of this process is $\sigma(x) = (x+1) * (\alpha^8 * x + 1)$. The two roots α^0 and α^{-8} correspond to errors at symbol locations 0 and 8, respectively. To find the error values, the Forney formula is used, as follows. At symbol location i, the error value is given by $Y_i = \omega(x)x^{L-1}/\sigma'(x)$ evaluated at $x =$ root i of $\sigma(x)$. In this case, the results are $Y_0 = \alpha^3$ and $Y_1 = \alpha^1$.

2.4. Comparison of Decoding Algorithms

The Reed–Solomon codes employed in most magnetic storage devices are high-rate codes. As mentioned previously, the major steps involved in the decoding algorithms are (1) syndrome computation (2) error locator polynomial and error evaluator polynomial determination, and (3) error value computation. Step 2 has the most variations. When the maximum number of errors to be corrected is small — typically, four or fewer errors — the most efficient way of obtaining the error locator polynomial is to compute all the coefficients of the error locator polynomial directly from the syndromes using a special and simple technique [14]. For higher numbers of errors, usually the well-known Berlekamp–Massey algorithm or the Euclidean algorithm is used. Most of the algebraic coding books give detailed descriptions of these two algorithms [6,9–11]. These two algorithms are both iterative and equivalent but they have two interesting opposite characteristics, which may be the determining consideration in algorithm selection. In the Berlekamp–Massey algorithm, the degrees of the polynomials used in the iteration increases as the procedure continues and the total number of iterations is a fixed number. On the other hand, in the Euclidean algorithm, the maximum polynomial degree decreases from iteration to iteration and the number of iterations to complete the procedure is variable. Another notable thing is that the Berlekamp–Massey algorithm can be used to find either the error locator polynomial alone or both the error locator polynomial and the error evaluator polynomial. On the other hand, in the Euclidean algorithm, both of these polynomials are computed at the same time. It is not very efficient to use the Euclidean algorithm to compute the error locator polynomial alone.

3. OTHER ASPECTS OF REED–SOLOMON CODING

3.1. Hardware Versus Firmware Implementations

The entire encoding and/or decoding process can be implemented either by microprocessor firmware or by hardware. Typically, the encoding is done in hardware to improve the speed of the process of writing to disk. Also, the complexity of an all-hardware encoder is much lower than that of the hardware decoder. Disk drives have the luxury of rereads, which are equivalent to requests for retransmission in communication systems. When hardware gates were expensive, the ECCs employed in magnetic recording systems made use of rereads to reduce the cost of hardware implementation. The hardware complexity is a function of the number of symbol errors corrected. Therefore, one way to save gates or hardware complexity is to correct only a smaller number of symbol errors than the designed error-correcting capability in hardware; when the hardware decoder fails, the firmware decoder is then used to correct all the errors. For example, when the ECC is designed to correct six errors per interleave, the hardware decoder can be designed to correct two or fewer errors and the firmware decoder to correct three, four, five, and six errors per interleave. Data throughput is affected every time the firmware algorithm is employed. The hardware encoder can also be designed to perform the syndrome computation with simple modification [15]. This design further reduces the total complexity of the hardware decoding.

There is a mode of error correction that may be different from other communication systems. Occasionally, a small particle may get in contact with the read head resulting in an increase in temperature. The heat produced by the contact may cause a long burst of errors to occur. This phenomenon is called a thermal asperity. Fortunately, the location of a thermal asperity can be detected by the signal processing system prior to ECC decoding. As a result, a more powerful error–erasure decoding algorithm can be used to deal with the long burst in the event that the normal error-only decoding algorithm fails to correct the errors. This decoding algorithm is capable of correcting e erasures and t symbol errors when $2 * t + e < d$, where an erasure is an error with known location. This algorithm is also given in most coding textbooks [6,9–11].

3.2. Block Missynchronization Detection

As can be seen from the example, a Reed–Solomon code word is likely decoded to a different codeword if a shift in the symbol position occurs. As mentioned previously, the Reed–Solomon code is a cyclic code, which means that a cyclic rotation of a nonshortened codeword is also a codeword. If $(c_{n-1}, c_{n-2}, c_{n-2}, \ldots, c_2, c_1, c_0)$ is a codeword of a nonshortened Reed–Solomon code over GF(2^m), that is, $n = 2^m - 1$, then both cyclic shifted versions $(c_{n-2}, c_{n-2}, \ldots, c_2, c_1, c_0, c_{n-1})$ and $(c_0, c_{n-1}, c_{n-2}, c_{n-2}, \ldots, c_2, c_1)$ are also codewords of the same code. As a result, Reed–Solomon codes are susceptible to miscorrection when symbol synchronization errors occur. To minimize this possibility, block synchronization must be assured for each sector prior to ECC decoding. This is accomplished with a specially designed sequence called an *address mark*. In other words, the data of a sector are preceded by an address mark. The address mark is designed to ensure correct block synchronization before the ECC attempts error correction. There is a preamble preceding the address mark. The main purpose of the preamble is for training the phase-locked loop (PLL) to acquire the bit synchronization. The secondary purpose of the preamble is to assist the address mark to establish a distinct position in the bitstream, which, in turn, enables the system to find the beginning of the first symbol of the ECC block. To facilitate the PLL training, a repeated pattern is used as the preamble. This pattern is often a long sequence of identical bits, such as a sequence of ones. In the following discussion, let a long sequence of ones (1s) serve as the preamble, which precedes the address mark. Let the address mark be a 12-bit sequence 000010100110 with the bit on the left preceding the bit on the right in time. The pattern, including the preamble and address mark, recorded on the disk is $\cdots$ 111111111111000010100110, where the underlined bits are the address mark. As soon as the bit synchronization is established by the phase-locked loop, a circuit compares every 12 consecutive bits with the address mark. The Hamming distance between the address mark and the 12 ones is 8 (bits), because there are 8 zeros in the address mark. Let the distance from the address mark to an out-of-phase sequence be $d_a(s)$. An out-of-phase sequence consists of s ones of the preamble followed by the first 12 s bits of the address mark. It can be checked that $d_a(s)$ is 7,8,9,8,8,7,6,7,7,7,7,0 for $s = 11,10,9,8,7,6,5,4,3,2,1,0$, respectively. Other than the case of $s = 0$, the minimum distance for this address mark from the out-of-phase sequences is 7. Therefore, this address mark can tolerate three errors. In other words, with three or fewer errors in any span of 12 consecutive bits prior to the last bit of the address mark, the position of the address mark can be correctly established. This address mark is called a "3-error-tolerance address mark." Let t be any threshold with t less than or equal to 3. Then the correct address mark is assumed found if any span of 12 consecutive bits differs from the correct address mark by t or fewer bits. By varying the value t, the performance of the address mark can be controlled to a certain degree. There are two important probabilities associated with an address mark for a specified threshold t. The probability of failure to synchronize is the probability that the system fails to find the correct synchronization position; this occurs when there are more than t errors on the address mark retrieved from the disk. This probability for the case of independent errors is given by the dominant term of the probability

$$P_{\text{failure\_to\_synchronize}} \approx C(L, t+1)p^{(t+1)}(1-p)^{(L-t-1)} \tag{8}$$

where L is the length of the address mark in bits, p is the raw bit error rate, t is the threshold, and $C(L, t+1)$ is the binomial coefficient of $x^{(t+1)}$ in the expansion of $(x+1)^L$. For most cases of interest, p is a small number. Therefore, the $1-p$ term approaches 1. Consequently, the probability of failure to synchronize is decreased as the threshold is increased. The other important probability is the probability of false synchronization. This is the probability that a span of L bits prior to the correct address location is mistakenly identified as the address mark. This probability is lower bounded by

$$P_{\text{false\_synchronization}} \approx C(d_a, d_a - t)p^{d'}(1-p)^t \tag{9}$$

where d_a is the minimum distance between the address mark and any span of L bits prior to the correct address mark position and $d' = d_a - t$. As can be seen, the probability is proportional to $p^{d'}$, and this probability increases as the threshold decreases.

Therefore, these two probabilities put conflicting requirement on the threshold t. One obvious solution is to increase the length of the address mark L, which leads to higher values of error tolerance; as a result, there are more values of threshold to select to meet the demand for both probabilities. Unfortunately, a longer address mark means larger overhead and more complicated circuits for detecting the address mark. The other choice is not to increase the length of the address mark and rely on the ECC to help in detecting the false synchronization. Using the 12-bit address mark with a raw bit error rate of $p = 1.0 \times 10^{-4}$ and $t = 3$ as an example, the two probabilities are $P_{\text{failure\_to\_synchronize}} \approx 9.92 \times 10^{-21}$ and $P_{\text{false\_symchronization}} \approx 7.00 \times 10^{-15}$. The probability of failure to synch is acceptable but the probability of false synchronization is

too high for the data storage application. Reed–Solomon codes over GF(2^m) are susceptible to synchronization errors when the bit slippage is a multiple of m, the symbol size. Therefore, the miscorrection probability is of the order of $P_{\text{false\_symchronization}}/m$. With a false synchronization probability of 7.00×10^{-15}, the miscorrection probability is of the order of 1.00×10^{-18}, which is too high. A solution to this situation is to use a coset instead of the code itself for recording purposes.

A coset consists of the set $\text{cl}(x) + c(x)$, where $c(x)$ is any codeword of the Reed–Solomon code and $\text{cl}(x)$ is a fixed sequence or vector of length n, called the coset leader. During recording, a member of the coset is recorded in the disk. When a sector is retrieved, the coset leader is added to the sequence resulting in $c(x)$, the codeword, assuming no errors and perfect synchronization. When out of synchronization, the retrieved sector contains $\text{cl}'(x) + \text{c}'(x)$, where $\text{cl}'(x)$ and $\text{c}'(x)$ are respectively shifted versions of $\text{cl}(x)$ and $\text{c}(x)$ with proper truncation and padding under error-free conditions. With $\text{cl}(x)$ added before decoding, the decoder sees $c'(x)$, a codeword plus possible errors due to truncation and padding, with $\text{cl}(x) + \text{cl}'(x)$. Therefore, with proper selection of $\text{cl}(x)$, $\text{cl}(x) + \text{cl}'(x)$ contains many nonzero terms. Each of these nonzero terms appears as an error to the ECC decoder. A well-designed $\text{cl}(x)$ results in more errors than the ECC correcting capability most of the time. Therefore, the decoder cannot correct these "errors." Consequently, the miscorrection probability is substantially lower than the value $P_{\text{false\_symchronization}}/m$. In fact, it is given by $\{P_{\text{false\_symcronization}}/m\} * \{\text{probability of decoding } \text{cl}(x) + \text{cl}'(x) + c'(x) \text{ as a codeword}\}$. With proper design of $\text{cl}(x)$ and sufficiently long redundant symbols, the probability of decoding $\text{cl}(x) + \text{cl}'(x) + \text{c}'(x)$ as a codeword can be made very small. Therefore, the miscorrection probability is significantly reduced. Reference 16 provides a way to find the coset leaders. In the event that the bit slippage is not a multiple of the symbol size m, the miscorrection probability caused by the false synchronization is given by $\{(1-m) * P_{\text{false\_synchronization}}/m\} * \{\text{probability of decoding } \text{cl}(x) + \text{cl}'(x) + \text{c}'(x) \text{ as acodeword }\}$; however, in this case, $c'(x)$ looks like a random sequence to the decoder, and so does the sequence $\text{cl}(x) + \text{cl}'(x) + \text{c}'(x)$. Therefore, the probability that a randomlike sequence is decoded as a codeword can be easily computed or approximated [17] to be a small number. Consequently, with a properly designed coset leader, the threshold t for the address mark can be simply set according to the requirement for $P_{\text{failure\_to\_synchronize}}$ alone. From a different point of view, a shorter length for address mark can be used with the help of the coset leader. Another similar approach for avoiding false synchronization is to initialize the encoder to a nonzero state prior to the encoding. This approach is more appropriate for cases where that the decoder needs to compute the remainder of the corrupted codeword as the first decoding step. More detailed information can be found in Ref. 18. Either approach will enhance the performance of the address mark, because the address mark deals mainly with one probability, the probability of failure to synchronize.

3.3. Interleaving Versus Noninterleaving

Let the symbols in a sector be orderly numbered as 0,1,2,3,4,5,6,7,8,9,.... Assuming that there are three codewords in a sector, there are many ways to associate the symbols with code symbols: $a[0], a[1], a[2], \ldots, b[0], b[1], b[2], \ldots, c[0], c[1], c[2], \ldots$, where $a[k], b[k]$, and $c[k]$ are kth symbols of code a, code b, and code c, respectively. Let the code length of each code be L. Then the first way of interleaving is $a[0], a[1], a[2], \ldots, a[L-1], b[0], b[1], b[2], \ldots, b[L-1], c[0], c[1], c[2], \ldots, c[L-1]$. Namely, $a[0], a[1], \ldots, a[L-1]$ are the symbols $0, 1, \ldots, L-1$ of the sector, respectively; $b[0], b[1], \ldots, b[L-1]$ are the symbols $L, L+1, \ldots, 2L-1$, of the sector, respectively; and $c[0], c[1], \ldots, c[L-1]$ are the symbols $2L, 2L+1, \ldots, 3L-1$ of the sector, respectively. This arrangement of three codewords has the main advantage that the single decoder can be used sequentially. However, the error-correction power of Reed–Solomon coding is not "enhanced." The more commonly used arrangement is $a[0], b[0], c[0], a[1], b[1], c[1], a[2], b[2], c[2], \ldots$. This arrangement is called *interleaving*. The number of code words involved is denoted by I and called the *depth of interleaving*. The main advantage of this arrangement is that a burst of I symbols can affect at most one symbol for each codeword. Therefore, from the entire sector point of view, the I code words in a sector form a t-burst correction code if each codeword can correct t-symbol errors.

A primary consideration for code selection involves the decision for determination of the degree of interleaving. The length of a Reed–Solomon code over GF(2^m) is limited to $2^m - 1$ symbols or $(2^m - 1) * m$ bits. Let the degree of interleaving be I. Then the minimum depth of interleaving is given by the equation $I * (2^m - 1) * m >$ or $= 4096 +$ number of redundant bits per sector. For $m = 8$, the minimum value of I is 3. The smallest value of m for I as 1 (the case of noninterleaving) is 9. In addition to code length, the ease of decoding sometimes plays an important role in determination of interleaving depth. For example, for the case of $m = 9$ and $I = 2$, each interleave needs to correct t errors when there are t bursts, where each burst corrupts no more than two 9-bit symbols. This design is much simpler than the design of using a single code to correct $2t$-symbol errors, especially when t is a small value such as 1, 2, 3, or 4 [14]. It should be noted that the ECC can correct any combination of $2t$ symbol errors for the noninterleaved case and that the interleaved case can correct only the "well behaved" $2t$ or fewer symbol errors. By a "well-behaved" pattern, the $2t$ or fewer errors are distributed in such a way that each interleave sees t or fewer symbol errors. As a result, the noninterleaved case is more powerful in terms of ECC correction power. However, the interleaved implementation may be the right design if the errors are frequently in bursts and complexity is the main concern.

3.4. Performance

The main purpose of using an ECC is to bring the raw bit error rate at the output of the modulation code decoder to an acceptable level. The raw bit error rate at the output of the modulation code decoder is often targeted at 1.0×10^{-6}

or higher. The required bit error rate at the output of the ECC is usually below 1.0×10^{-15}. Traditionally, the error rate specification is given in terms of bit error rate. However, for Reed–Solomon codes, it is easier to perform the error rate computation on the basis of the symbol error rate and the code block error rate. Therefore, the definitions and relationships among various error rates should first be clarified.

3.5. Error Rate Definitions

The commonly used definition of bit error rate is as follows:

$$\text{Bit error rate} = \frac{\text{total number of bits in error}}{\text{total number of bits observed}} \tag{10}$$

Similarly, symbol error rate is defined as

$$\text{Symbol error rate} = \frac{\text{total number of symbols in error}}{\text{total number of symbol observed}} \tag{11}$$

$$\text{Block error rate} = \frac{\text{total number of blocks in error}}{\text{total number of blocks observed}} \tag{12}$$

The denominators of these three definitions have very simple relations:

$$\text{Total number of bits observed} = m * (\text{total number of symbols observed}) \tag{13}$$

$$\text{Total number of symbols observed} = n * (\text{total number of blocks observed}) \tag{14}$$

$$\text{Total number of bits observed} = n * m * (\text{total number of blocks observed}) \tag{15}$$

where m is the number of bits per symbol and n is the number of symbols per code block. If the relationship among the numerators can be made as simple, then the conversion from one error rate to another becomes straightforward. With different modulation codes and signal processing techniques, the relationships among the numerators may not be simple. However, bounding techniques and approximations can be used to obtain estimated results, which often provide sufficient information from a designer's point of view. The relationships among the various error rates at the output of the modulation code decoder can be obtained more readily. The error rates at the output of the modulation code decoder are often measurable quantities as the total number of bits to be observed or counted is no more than a few million if the bit error rate is 1.0×10^{-6} or worse. The relationships among various error rates at the output of an ECC system are often difficult to quantify. Therefore, the following discussion deals with the decoded error rates.

When a t-symbol error-correcting code is employed, the most likely errors for which the ECC fails to correct are error patterns containing $(t+1)$-symbol errors. Therefore, the following relationship can be established:

$$\text{The total number of symbols in error} \approx (t+1)(\text{the total number of blocks in error})$$

The number of bits in error in an erroneous symbol can be anywhere from 1 to m, where m is the total number of bits per symbol. In fact, this number depends on the modulation code used. If this number can be obtained from the modulation code, it should be a number between 1 and m. For a completely random system the number is approximately $m/2$. A good modulation code tends to make this number less than $m/2$. Therefore, the following relationship can be established:

$$\begin{aligned}\text{Total number of bits in error} &= am(\text{total number of symbols in error}) \\ &\approx am(t+1)(\text{the total number of blocks in error})\end{aligned}$$

(with $0 < a \leq 1$, for random data $a = 0.5$). With these approximations, it is straightforward to show that

$$\begin{aligned}\text{Bit error rate} &\leq a * (\text{symbol error rate}) \quad (\text{with } 0 < a \leq 1) \\ \text{Symbol error rate} &\approx \frac{t+1}{n}(\text{block error rate}) \\ \text{Bit error rate} &\leq a\frac{t+1}{n}(\text{block error rate}) \quad (\text{with } 0 < a \leq 1).\end{aligned}$$

For the commonly used symbol sizes of $m = 8$ and 10, whether the true value of a or its upper bound is used in the preceding relationship, the results differ less than one order of magnitude. Therefore, in the following discussion, the symbol error rate and the block error rate are used; the results can be readily converted to the desired bit error rate with either the accurate estimate of a or its upper bound. The starting point is the symbol error rate.

Let the raw symbol error rate be P_s at the input of the t-symbol-correcting ECC decoder. For a system in which symbol errors are statistically independent, the decoded block error rate P_b is given by

$$P_b = \sum C(n,j)P_s^j(1-P_s)^{(n-j)} \tag{16}$$

where the summation limits are from $j = t+1$ to $j = n$ and $C(n,j)$ is the binomial coefficient for expanding the polynomial $(1+x)^n = \sum C(n,j)x^j$. Knowing the decoded block error rate, we can obtain the decoded symbol error rate. In fact, the decoded symbol error rate P_d can be computed directly from the formula

$$P_d \cong \frac{1}{n}\sum C(n,j)jP_s^j(1-P_s)^{(n-j)} \tag{17}$$

where the summation limits are from $j = t + 1$ to $j = n$. The results in Eqs. (16) and (17) can be obtained using combinatorial arguments. The reason Eq. (17) is an approximation is that when there are more than t-symbol errors, the decoder may introduce additional errors. Reference 17 provides the formula for the exact expression for Reed–Solomon codes under the condition that the symbol errors are statistically independent. Equations (16) and (17) both indicate that the decoded error rate is proportional to $P_s^{(t+1)}$. This can be seen from the observation that every term in the summation contains $P_s^{(t+1)}$ as a factor. Therefore, as t increases, the decoded error rate decreases. However, using more powerful ECC results in a lower code rate, which has the effect of making P_s worse. As a result, an ECC system cannot arbitrarily increase the correcting power to gain better overall decoded error rates. For the cases where the symbol errors are not statistically independent, the preceding arguments also hold. For more detailed information, see Ref. 19, where a soft bit error rate (SBER) is defined as (total number of error events)/(total bits read that are protected by the ECC system); therefore, it is equal to 1/b times the bit error rate defined by Eq. (10), where b is the average number of bits in error per error event.

3.6. Separate EDC or Embedded EDC

When a disk drive or a tape drive is used to store data, the miscorrection probability must be many orders of magnitude better than the probability that the ECC cannot decode the code block. There are two common ways of achieving this goal. The first approach is to use a separate error-detecting code (EDC) and let the error-correcting code decode to the limit of its error correcting capability. In this case, an odd-distance error-correcting code is often employed and the code corrects t errors, where t is equal to $(d-1)/2$, where d is the minimum distance of the code. The second approach is to correct t errors only with $t < (d-1)/2$. Let the number of bits in the EDC for the first approach be r_1 and the relationship for the second approach be $r_2 = (d - 1 - 2t) * m$, where m is the symbol size. Then r_2 can be considered as the effective number of bits in the second approach reserved for error detection. When $r_1 = r_2$, and t is the same for both approaches, the miscorrection probabilities for both approaches are about the same. The main consideration as to which approach to adopt is listed below:

1. Approach 1 is more flexible. The value of r_1 can be more or less arbitrarily selected to achieve any desirable miscorrection probability. For approach two, the value for r_2 needs to be a multiple of m, the symbol size.
2. Approach 1 needs the additional step for checking the correctness of EDC after the error-correction procedure. This verification process is often time-consuming and complex. For approach 2, the EDC is automatically satisfied at the end of the decoding algorithm if the algorithm is properly designed and employed.
3. When it is required to provide the data block to the host computer, which may require that the data block have its own EDC, it may be mandatory to use approach 1.

In practice, both approaches are sometimes employed simultaneously. Both embedded EDC and separated EDC are used. The embedded EDC assures the low probability of miscorrection, while the separate EDC is used to communicate with the host computer. Often, the separate EDC itself is covered by the ECC with embedded EDC. As a result, the separate EDC is not checked as the embedded EDC provides the necessary detection power.

The separate EDC has another advantage in that it may not be affected by cyclic shifts of symbols. Therefore it can reduce the miscorrection probability due to symbol shift. As a result, the selection of the coset leader for missynchronization detection may be easier if the EDC is of sufficient length.

3.7. Tape Drive ECC

All the coding techniques employed in disk drives can be readily applied to tape drives. Tape drives support multiple tracks; in other words, a read head reads several tracks simultaneously. This provides the opportunity to apply a code across the tracks. Usually, the ECC along a single track corrects a limited number of symbol errors or none at all, but with sufficient redundancy to make sure that the miscorrection or mis-detection probability is extremely small. Whenever a track fails to correct, the errors are detected. The ECC across the track then corrects the errors indicated by the tracks whose ECC fail to correct. In other words, the symbols on the tracks, which are known to contain erroneous symbols as indicated by the ECC, are treated as erasures for the ECC across the tracks. An erasure is an error with known location. The more powerful error–erasure decoding algorithm is then used to correct both errors and erasures across the tracks. An example for this approach can be found in Ref. 20.

Another way of protecting data on tape drives is to use array codes [4]. This can be understood again by visualizing the two-dimensional structure of tape recording. Simple codes, preferably the single-parity-check codes, are introduced on both the longitudinal direction (the direction following the head movement) and the direction across the tracks as well as the directions that are at constant angle with respect to the longitudinal direction. Therefore, a symbol is protected by several simple parity-check equations, one for each direction. When one or more of these directions contains one or no errors, the symbol can be correctly recovered [4].

3.8. Codes for RAID

In redundant array of independent disks (RAID), inter-relations are introduced to several disk drives to form redundancy among sectors, one sector from each drive. The codes are in fact so-called concatenated codes. These are also two-dimensional codes. The code in one dimension

is the Reed–Solomon code presently used in each sector. The code used across the sectors is usually a very simple single-parity-check code. Namely, if the information symbols are $c[n-1], c[n-2], \ldots, c[2], c[1]$, then the redundant symbols are given by $c[0] = c[1] + c[2] + \cdots + c[n-1]$. The reason this is used is because the recovery of a failed drive can be achieved simply. The only limitation for this simple scheme is that it can protect against the case that a single drive fails. More sophisticated distance d Reed–Solomon codes can also be used to protect against the case that $d-1$ or fewer drives fail simultaneously. However, the reconstruction of the failed sectors often involves many reads and writes of the related sectors from different drives; also, the data recovery from the failed sectors needs more complicated operations than the simple exoring. Therefore, there is no reason not to employ more powerful Reed–Solomon codes across the drives. The time delays and extensive read and write operations may be the deciding factor for future adoption of more powerful codes.

3.9. Decoding Beyond $(d-1)/2$ for Reed–Solomon Codes

The decoding algorithms for Reed–Solomon codes presently employed in magnetic recording are the so-called hard-decoding algorithm and the error–erasure algorithm. In a hard-decoding algorithm, the channel outputs a bit-stream to the ECC decoder, where each bit is either 1 or 0. Each bit can be either right or wrong. The symbols at the input to the ECC decoder are symbols with predetermined "hard" values. In the error–erasure decoding algorithm, the channel also provides the locations of symbols, which are erasures whose values may not be correct. Therefore, in the case of error–erasure decoding, each symbol is associated with a reliability information, that is, with reliability information of 1 bit. For example, a one indicates that the symbol is reliable and a zero indicates that the symbol is not reliable; therefore, it is an erasure.

The present "hard" error-only decoding algorithm for a minimum distance d Reed–Solomon codes attempts to correct all possible error patterns, which are at a distance no more than $(d-1)/2$ symbols away from a codeword. New algorithms have been designed to provide a list of codewords, which are no more than $n*(1-r^{1/2})$ symbols away from the n symbol sequence to be decoded [21]. (Here $r = k/n$ is the code rate.) Conventional hard decoding can correct $(d-1)/2 = (n-k)/2 = n*(1-r)/2$ symbols, which is smaller than $n*(1-r^{1/2})$ symbols. The advantage of this new decoding algorithm is more significant when the code rate r is low. With a properly selected algorithm and code rate, this approach has the potential to enhance the performance for Reed–Solomon codes. More information can be found in Refs. 21 and 22.

Another direction is soft-decision decoding. Instead of providing a symbol with one bit of reliability information like the erasure information, many bits of reliability information are associated with a symbol. The decoding algorithm makes use of this reliability information to further enhance code performance. There are many encouraging results for soft-decision decoding algorithms employed with binary codes. References 23 and 24 are examples for soft-decision decoding for nonbinary codes. The performance gain for soft-decision decoding in Reed–Solomon codes over hard-decision decoding is an interesting and important area of research in magnetic recording. If it proves to provide a significant improvement, the algorithm no doubt will soon be adopted in magnetic recording applications.

3.10. Larger Sector Size

Strictly from a coding point of view, the sector size for hard-disk drives should be larger than the present 512 bytes. For the same code rate, code performance improves as block size increases. For example, the Reed–Solomon (440,410,31) code over GF(2^{10}) can correct 15 symbol errors among 440 symbols, while the (880,820,61) code can correct 30 symbols among 880. By Eq. (16), the shorter code has the decoded block error rate proportional to P_s^{16}, while the longer code has the decoded block error rate proportional to P_s^{31}. It can be easily seen that for a very large range of P_s, the longer code gives lower decoded block error rate for the same raw symbol error rate P_s. In addition, there are other overhead savings such as preambles and address marks. Unfortunately, many operating systems assume a 512-byte sector size. A change in disk drive sector size would require corresponding modification to these operating systems. At present, the 512-bytes data constitute the one sector size that the disk industry must provide.

BIBLIOGRAPHY

1. H. J. Richter, Longitudinal recording at 10 to 20 gbit/inch^2 and beyond, *IEEE Trans. Magn.* **35**(5): 2790–2795 (1999).
2. J. Lee and V. K. Madisetti, Error correcting run-length limited codes for magnetic recording, *IEEE Trans. Magn.* **31**(6): 3084–3086 (1995).
3. N. Glover, *Practical Error Correction Design for Engineers*, Data Systems Technology Corp., Broomfield, CO, 1982.
4. M. Blaum and R. M. Roth, New array codes for multiple phased burst correction, *IEEE Trans. Inform. Theory* **39**(1): 66–77 (1993).
5. R. J. McEliece, *Finite Fields for Computer Scientists and Engineers*, Kluwer, Norwell, MA, 1987.
6. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
7. U.S. Patent 4,975,867 (1990), L. J. Weng, Apparatus for dividing elements of a Galois field of GF(2$\hat{Q}$M).
8. U.S. Patent 6,044,389 (2000), L. J. Weng and B. A. Shen, System for computing the multiplicative inverse of a field element for Galois field without using tables.
9. R. E. Blahut, *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, MA, 1983.

10. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2nd ed., MIT Press, Cambridge, MA, 1972.

11. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

12. U.S. Patent 4,633,470 (1986), L. R. Welch and E. R. Berlekamp, Error correction for algebraic block codes.

13. U. Sorger, A new Reed–Solomon decoding algorithm based on Newton's interpolation, *IEEE Trans. Inform. Theory* **39**: 358–365 (1993).

14. U.S. Patent 5,710,782 (1998), L. J. Weng, System for correction of three and four errors.

15. G. Fettweis and M. Hassner, A combined Reed–Solomon encoder and syndrome generator with small hardware complexity, *1992 IEEE Int. Symp. Circuit Syst.* **4**: 1871–1874 (1992).

16. U.S. Patent 5,528,607 (1996), L. J. Weng, B. Leshay, and D. Langer, Method and apparatus for protection of data from mis-synchronization.

17. Z. McC. Huntoon and A. M. Michelson, On the computation of the probability of post-decoding error events for block codes, *IEEE Trans. Inform. Theory* **23**: 399–403 (1977).

18. U.S. Patent 4,989,211 (1991), L. J. Weng, Sector mis-synchronization detection method.

19. C. M. Riggle and S. G. McCarthy, Design of error correction systems for disk drives, *IEEE Trans. Magn.* **34**(4): 2062–2371 (1998).

20. U.S. Patent 5,136,592 (1992), L. J. Weng, Error detection and correction system for long burst errors.

21. V. Guruswami and M. Sudan, Improved decoding of Reed–Solomon and algebraic-geometric codes, *IEEE Trans. Inform. Theory* **45**: 1755–1764 (1999).

22. M. Sudan, Decoding of Reed–Solomon codes beyond the error correction bound, *J. Complexity* **12**: 180–193 (1997).

23. E. R. Berlekamp, Bounded distance+1 soft-decision Reed–Solomon decoding, *IEEE Trans. Inform. Theory* **42**: 704–721 (1996).

24. G. D. Forney, Jr., Generalized minimum distance decoding, *IEEE Trans. Inform. Theory* **12**: 125–131 (1966).

COMMUNICATION SATELLITE ONBOARD PROCESSING

STEVEN BERNSTEIN
MIT Lincoln Laboratory*
Lexington, Massachusetts

1. INTRODUCTION TO ONBOARD PROCESSING

The primary purpose of communication satellites is to receive signals from one or more sources and relay them to intended recipients. With the exception of early passive on-orbit reflectors, the relay function is accomplished by some form of active processing carried out on board the satellite. The most common form of active processing is the "translating repeater" or "transponder," which simply amplifies whatever is received in a given uplink frequency band and retransmits it in a different downlink frequency band. While this could be called "onboard processing," we will reserve this term for techniques that manipulate the signals passing through a satellite in more complex ways.

Digital circuit and signal processing technology has progressed to the point that it is feasible to consider very ambitious onboard processing functions. In this article we will concentrate on concepts, confident that technology will soon enable virtually all the ambitious techniques to be described.

1.1. Conventional (Nonprocessing) Communication Satellites

Figure 1 is a block diagram of a translating repeater, the workhorse of conventional (nonprocessing) communication satellites. This is also sometimes called a "bent pipe" because it simply takes in a band of frequency spectrum on its uplink and bends it back to earth. An onboard oscillator and mixer is used to translate the uplink band

* This work was sponsored by the Air Force under Air Force Contract FI9628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Air Force.

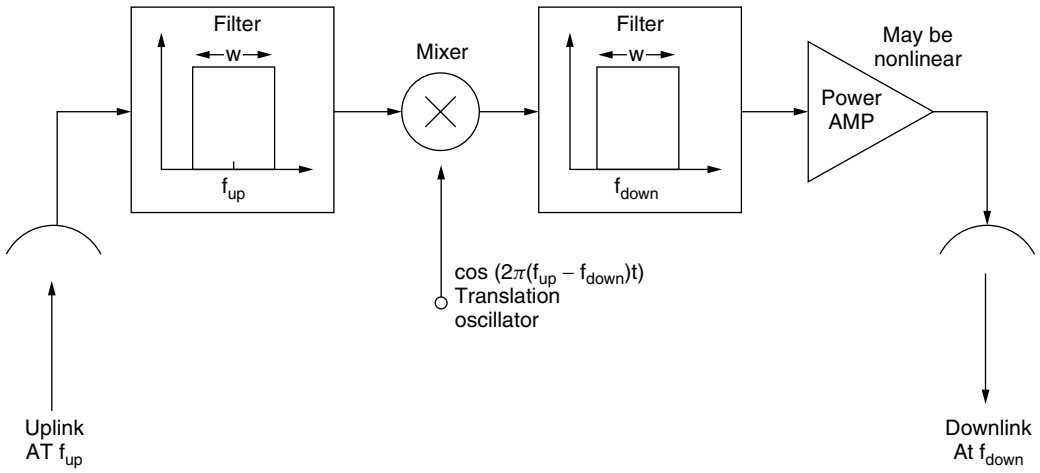


Figure 1. Translating repeater.

to a different downlink band in order to isolate the input and output communication signals. For example, many C-band commercial satellites operate with uplink signals around 6 GHz and downlink signals around 4 GHz. Most such satellites are built with a number of translating repeaters, each operating at a slightly different frequency.

The amplifier shown may be linear or nonlinear. Linear amplifiers are often used in order to minimize crosstalk due to intermodulation products when there are multiple signals within the operating band. Automatic gain control (AGC) is sometimes used to keep the amplifier in a linear region. Nonlinear amplifiers, often operating in a saturated mode, are used when it is desired to maximize prime-to-RF power efficiency. In the nonlinear case, using signals that rely on amplitude modulation may not be feasible; constant envelope phase modulated signals would be more appropriate.

Also shown is a representation of a single antenna beam on the uplink and a single one on the downlink. The antenna beam patterns may be quite complex (e.g., shaped to cover a single country), but all signals in each direction share a common beam. (The uplink and downlink beam shapes and positions, however, may be different.)

1.2. Limitations of Conventional Satellite Architecture

While appealingly simple (and useful), the translating repeater has a number of limitations. Many of these stem from the fact that the power of the single RF amplifier must be shared.

1.2.1. Multiple Access. If a number of communication signals are being relayed simultaneously as in an FDMA (frequency division multiple access) system, then care must be taken to ensure that each user signal arrives at its downlink destination with sufficient signal-to-noise ratio. In order to achieve this, coordination is required among the uplink transmitting stations, especially with regard to power control. For example, an uplink signal that is 3 dB stronger than necessary may capture twice as much transponder power as it entitled to, thus acting as though it were two users. (In this case, even if the amplifier were operating in a linear mode, the amplifier gain, and consequently the downlink power transmitted, would need to be reduced.)

In a system where the user population is heterogeneous, such as operating with different rates and with different terminal antenna sizes, the power control problem becomes quite complex and real-time system monitoring and control is used. This is summarized in Fig. 2.

An approach that reduces the power control complexity is to use TDMA (time-division multiple access). In this design only one user transmits through the transponder at a time and hence may use full uplink power. However, power control in the form of burst data rate and duty cycle is still needed so that each downlink signal obtains at least the minimum required signal-to-noise ratio at the receiving terminal.

1.2.2. Uplink Interference. In some systems uplink interference (intentional or natural) is a major factor to consider. In order to illustrate this, consider the simple scenario consisting of a single user signal and uplink interference that is equivalent to flat Gaussian noise over the system bandwidth, W Hz. Assume that the transponder signal power is shared proportionately between the communication signal and the interference.

Define other needed parameters as

Total satellite power received at the downlink terminal = Pr

Background noise density at the downlink terminal = N_0

Uplink signal power received at the satellite = S

Uplink interference power received at the satellite (which would include the satellite's own receiver noise) = I

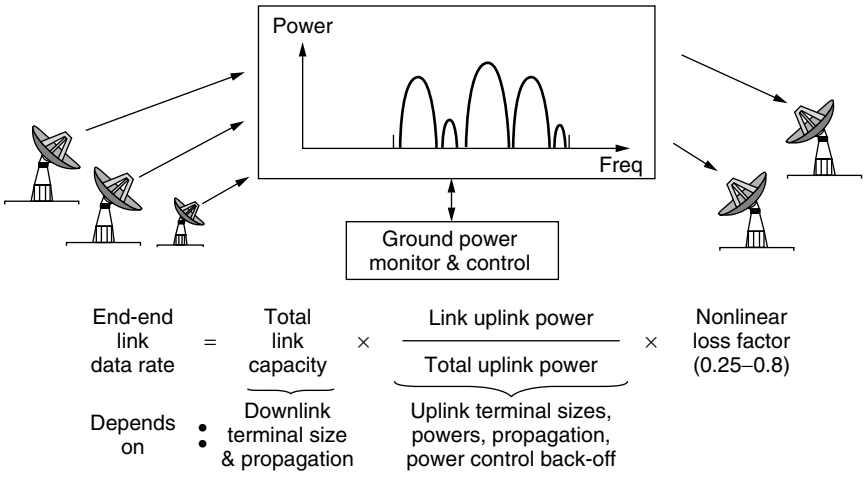


Figure 2. Transponder, multiple access.

Required signal-energy-to-noise-ratio per bit determined by the modulation/coding scheme used $= (E_b/N_0)_{\text{req}}$

The useful signal power received at the downlink terminal is $(S/(S+I))\text{Pr}$, and the interference power retransmitted by the satellite will be $(I/(S+I))\text{Pr}$. The total noise density at the downlink terminal will be the sum of the background and retransmitted amounts $N_0 + (I/(S+I))\text{Pr}/W$. Hence the effective carrier power : noise density ratio will be

$$\left(\frac{\text{Pr}}{N_0}\right)_{\text{eff}} = \frac{\dfrac{S}{S+I}\dfrac{\text{Pr}}{N_0}}{N_0 + \dfrac{I}{S+I}\dfrac{\text{Pr}}{W}}$$

From this we can find the data rate that can be supported as

$$R = \frac{\left(\dfrac{\text{Pr}}{N_0}\right)_{\text{eff}}}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}$$

$$= \frac{W}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}\frac{S}{I}\frac{\dfrac{\text{Pr}}{N_0 W}}{1+\dfrac{\text{Pr}}{N_0 W}} \quad \text{bps}$$

This relationship is shown qualitatively in Fig. 3 as a function of the total downlink signal-to-noise ratio, $\text{Pr}/N_0 W$. We assume that $I \gg S$.

We observe that when $\text{Pr}/N_0 W \gg 1$, the supportable data rate is given by

$$R = \frac{W}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}\frac{S}{I} \quad \text{bps}$$

which is not a function of the downlink signal to noise ratio; however, it is a function of system bandwidth, W. This is called the *bandwidth-limited case* because increasing the bandwidth W would directly improve performance; there would be no improvement if the satellite power were increased. This is a goal sought after in designing systems to counter uplink interference, such as jamming. It will be shown later that onboard processing can approach this goal.

On the other hand, if $\text{Pr}/N_0(W) \ll 1$, then the supportable data rate is

$$R = \frac{\left(\dfrac{\text{Pr}}{N_0}\right)}{\left(\dfrac{E_b}{N_0}\right)_{\text{req}}}\frac{S}{I} \quad \text{bps}$$

which is not a function of the system bandwidth, W. This is called the *power-limited* case because the only satellite parameter we can change to improve performance is to increase the downlink power. We see that in this case the interference is working by an effect called "power robbing." This is a serious limitation of transponder-based systems. We will return to this example when discussing onboard processing techniques.

It should be noted that in either case, a nonlinear transponder can degrade performance by an additional 1 dB through a small-signal suppression effect. If the interference had a constant envelope, the degradation could reach 6 dB.

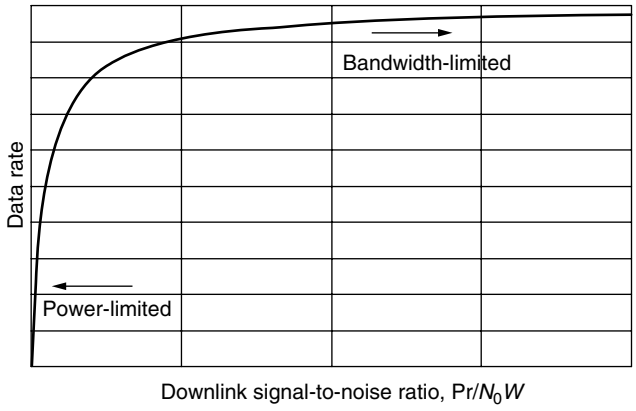


Figure 3. Power- and bandwidth-limited performance of transponder with uplink interference.

1.2.3. User Interconnection. Many satellite systems have multiple antenna beams on both the uplink and downlink. Multiple beams can provide more gain in the direction of ground terminals and permit the reuse of the same frequencies in different parts of the earth. Observe, however, from Fig. 1 that no provision in a simple translating repeater is made for anything more than a simple beam-to-beam connection. Although this is often acceptable, it is a significant limitation in other scenarios.

1.3. What Is Onboard Processing?

Onboard processing includes a wide variety of techniques: analog and digital, RF and baseband, circuit-oriented and packet-oriented. Following the discussion above it is instructive to relate these techniques to the limitations of the model of nonprocessing, the translating repeater.

Table 1 lists some of the main onboard processing techniques that address the limitations, which will be discussed further below. (Note that not all systems with onboard processing necessarily employ all of these techniques.)

2. ONBOARD PROCESSING TECHNIQUES

This section describes some of the principal onboard processing techniques.

2.1. Antenna Beam Switching

Many satellites are deployed with multiple antenna beams. A principal reason for doing so is to increase capacity by taking advantage of the high gain of narrow beams. This is illustrated in Fig. 4, which shows typical link capacities for several diameter geosynchronous satellite

Table 1. Onboard Processing Techniques

Translating Repeater Limitation	Onboard Processing Technique
Limited means of sharing resources without significant user cooperation	Demodulation–remodulation
	Access control
	Circuit multiplexing and switching
	Packet switching
Significant vulnerability to uplink interference	Demodulation–remodulation
	Adaptive antennas
	Despreading of spread-spectrum signals
Fixed interconnectivity	Antenna beam switching and crosslinks
	Adaptive antennas
	Demodulation–remodulation
	Access control
	Circuit multiplexing and switching
	Packet switching

spot beams as a function of ground terminal diameter. (It is assumed that the link is limited by signal-to-noise ratio, not bandwidth.) The link capacity varies inversely with the square of the diameter of the beam on the earth. The motivation for using spot beams should be clear.

A second reason for using spot beams is to reuse frequency assignments in different (usually nonneighboring) beams. The smaller the beam size, the greater number of times the frequency can be reused in a region of coverage.

The presence of multiple beams raises the question of how to interconnect users in different beams to each other. One way to do this is with RF crossbar switches as illustrated in Fig. 5. This is the architecture used in a portion of the NASA ACTS satellite [1]. The crossbar switch can be controlled quasistatically from the ground or, as shown, can respond dynamically to a signal structure such as TDMA. The beam–beam switch can also be combined with frequency filters to switch bands of the frequency spectrum to designated beams.

While uplink and downlink beams are being used for illustration, the same technique can be combined with satellite-to-satellite crosslinks.

2.2. Adaptive Antennas

While many satellites are launched with antenna systems that form carefully designed beam shapes to cover a specific country or region, few can adapt their patterns to meet changing user needs. Pattern adaptation can take several forms, including

- Steerable narrow beams for both uplink and downlink
- Formation of shaped uplink and downlink beams to provide enhanced gain to one or more specific regions
- Formation of spatial uplink "nulls" to reduce the effect of interference sources

The steering of individual narrow beams is usually implemented with mechanically steered dishes. The implementation of more complex shapes can be achieved with

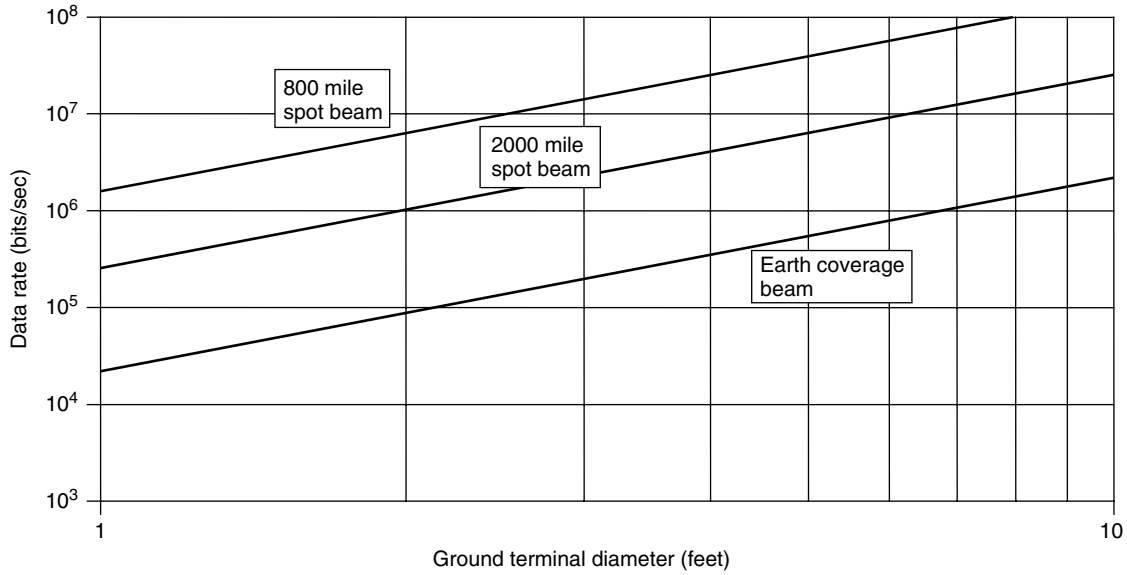


Figure 4. Typical total link capacity for geosynchronous satellites.

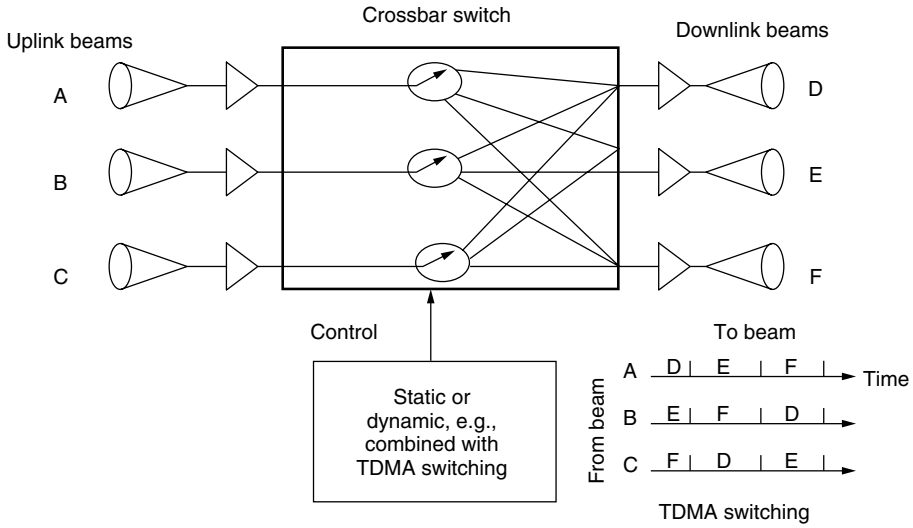


Figure 5. Interconnecting spot beams.

microwave lens antennas or phased arrays [2]. The control of the antenna patterns is usually performed on the ground, but some systems are autonomous in space. More recent approaches to adaptive designs often use *digital beamforming*, whereby the output of each antenna element is downconverted and sampled for subsequent digital signal processing.

It should be noted that adaptive antenna processing could be applied to both analog and digital communication systems.

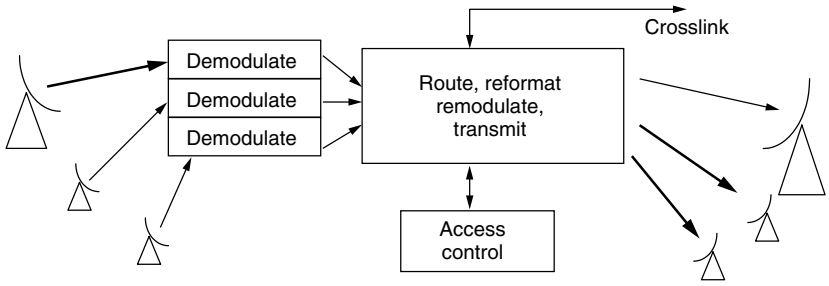
2.3. Demodulation–Remodulation

Demodulation–remodulation is one of the most powerful onboard processing techniques. This is illustrated in Fig. 6, which shows each user uplink being demodulated to a bit stream. The user bit streams are then processed by a digital switching subsystem that can route and reformat the streams and finally remodulate them onto one or more downlinks. (Specific system designs might take advantage of only a subset of these functions. There are also a number of intermediate degrees of processing, such as demodulating user datastreams for error-control coding of parity symbols without decoding to information bits.)

The advantages of this approach are numerous, particularly compared with those of translating repeaters:

1. *Satellite transmitter power is used more efficiently*. Satellite processing permits the renormalization of downlink power sharing. For example, an uplink signal that is received at a power level 3 dB higher than that of all the other signals would capture 3 dB more downlink power than its "fair share" in a translating repeater system. However, with a processor on board, the downlink signal can be renormalized to its fair value. (Of course, some degree of uplink power control must be utilized to guarantee that the uplink signal-to-noise : interference ratio on the uplink permits onboard demodulation.) In addition, an onboard processor can adjust the amount of power devoted to each downlink stream to match the capabilities of each downlink terminal's receiver. Thus a broad system capacity maximization can be approached with minimal complexity imposed on individual terminals.

2. *Downlink power is not wasted on retransmitted noise*. Uplink demodulation, in effect, strips off uplink noise that may come from natural sources, uplink interference, or the satellite's receiver front end. Translating repeaters would



- Reduces need to power balance uplink
- Doesn't waste downlink power retransmitting uplink noise
- Allocates downlink power where needed
- Connects users in different narrow beams
- Optimizes uplink and downlink resources independently

Figure 6. Onboard demodulation–remodulation.

repeat such sources of noise, wasting downlink power. Uplink signals that are below a specified quality measure, such as bit error rate, can be rejected by the satellite processor and not transmitted on the downlink. This property is advantageous for multiple access systems and for systems countering uplink interference such as jamming. The effect is to bring the system into the *bandwidth-limited* regime discussed previously since the power-robbing effect is virtually eliminated. This advantage in power sharing is illustrated in Fig. 7.

3. *Users in different antenna beam ranges can be interconnected*. The routing capability of an onboard processor permits uplink and downlink users in different satellite antenna beams to interconnect. This gives the system designer significant degrees of freedom in the design of antenna patterns. For example, a satellite with numerous high-gain narrow beams can be utilized without limiting connectivity to users to in the same beam. Various interconnectivity structures can also be accommodated, including point-to-point, broadcast, multicast, and many-to-one connections.

4. *Uplink and downlink signal structures can be independently optimized*. Onboard processing effectively permits the uplink and downlink signal structures (and antenna designs) to be optimized independently. This is a very powerful degree of freedom for the system designer.

A generic example that combines several of the concepts discussed above is shown in Fig. 8. In this example a number of simultaneous uplink antenna beams are used to permit small user terminals to transmit at low power into high-gain satellite receive beams. Uplink user signals are prevented from interfering with each other by a combination of FDMA (frequency-division multiple access)

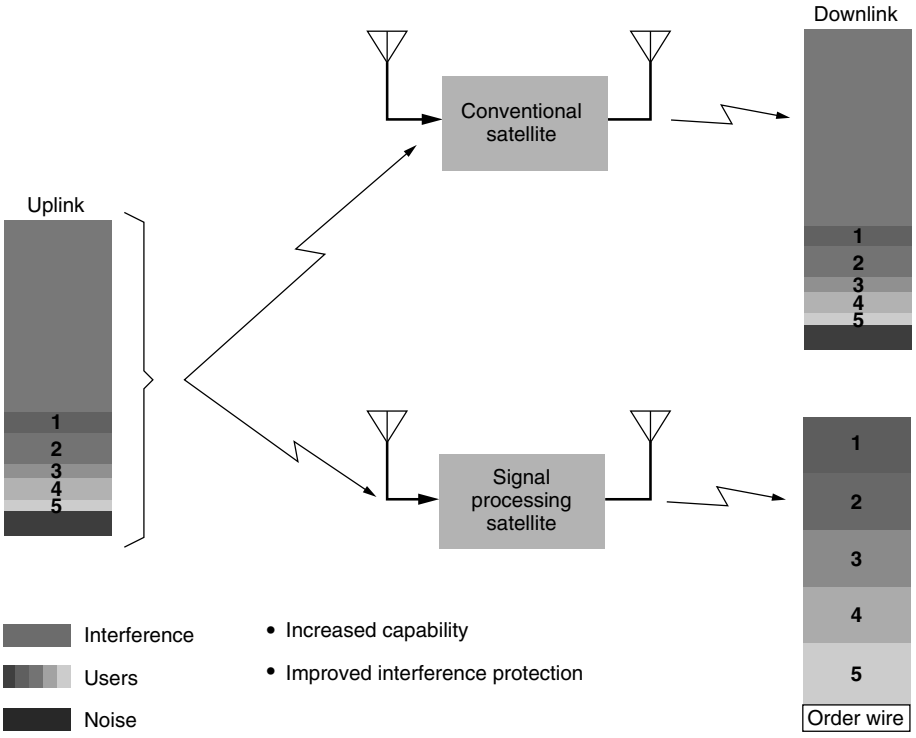


Figure 7. Power sharing advantage of onboard signal processing.

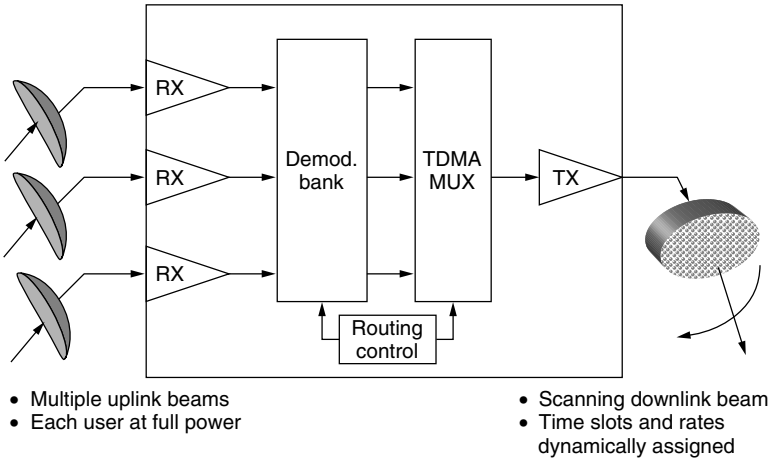


Figure 8. Onboard demodulation–remodulation example with beam switching.

for users in the same beam and low beam-to-beam sidelobe ratios. The satellite processor includes a bank of receivers and demodulators. The downlink consists of a single high-rate TDMA (time-division multiple access) datastream that carries all the downlink data. The burst rate of the data destined for an individual user can be adjusted in accordance with the receive capability of the user terminal. The TDMA downlink beam position is rapidly switched in synchronism with the TDMA stream to place downlink power on the location of the intended user. A downlink satellite transmitter can be used at high efficiency if the TDMA stream uses constant envelope modulation.

2.4. Despreading of Spread-Spectrum Signals

Additional techniques related to demodulation-remodulation can also be used to advantage. One is the onboard despreading of spread-spectrum signals, without complete demodulation.

Consider, for example, a frequency-hopped spread-spectrum system utilizing system bandwidth W Hz. An onboard processor consisting of a synchronized frequency-hopped local oscillator followed by a filter could be used to despread the uplink signal to its baseband bandwidth, approximately equal to the data rate, R bps. The resulting (analog) signal could then be transmitted on the downlink. Note that any broadband uplink noise or interference would be reduced by the ratio R/W, thus reducing the amount of retransmitted noise by the same factor. Although not quite as effective as complete demodulation, this would greatly reduce the "power robbing" caused by uplink interference. In a similar manner, a direct-sequence spread-spectrum system could be designed with an onboard despreader.

2.5. Access Control

Onboard access control places some of or all the functions of user system admission, such as interconnection control or usage monitoring, on the satellite. It provides the designer with a number of degrees of freedom to split these necessary functions between space and ground implementations. Its advantages lie with such factors as system security and survivability, system response times, and efficient use of control links. While some satellite systems require little real-time access control, such as those servicing television broadcasting, others must respond in real-time to user demands (e.g., telephone voice calls). A fully implemented spaceborne access control system would be a "switchboard in the sky."

System security and survivability can be enhanced by exposing fewer links that reach to ground control centers. An autonomous space system reduces the need for such links. In a hostile environment this could be an important factor.

System response times are reduced by not requiring a lot of control data traffic to make two ground-to-space-to-ground round trips between users and ground control sites. Whether this is a important consideration depends on the response time needed by the user community.

Efficient use of control links would reduce the requirements for uplink and downlinks devoted mainly to control. Virtually all satellite systems require control links for satellite on-orbit health maintenance. However, systems that must respond in real time to user demands have greater need for control dataflow. Placing control function on board reduces the data load on the control links, which could be advantageous in some system designs.

2.6. Circuit Multiplexing and Switching

Circuit switching was discussed in the context of onboard demodulation–remodulation of datastreams. Here we want to elaborate on the point that onboard processing can also serve as an *add–drop multiplexer*, a standard subsystem of the ground telecommunications industry. An uplink datastream from an individual user terminal might actually consist of a number of multiplexed substreams. Each of these substreams could have a different ground terminal as its destination. The onboard processors could demultiplex the uplink substreams, switch each substream to its intended downlink, and multiplex all the downlink substreams intended for a given terminal before transmission. By matching each substream to its intended downlink beam and terminal, satellite resources are used most efficiently.

2.7. Packet Switching

In data transmission technology, there is a strong trend toward the use of packet transmission protocols, particularly the IP (Internet Protocol) suite. Onboard processing is directly applicable to this approach. Indeed, aside from specific points made regarding data circuits or datastreams in the preceding sections, the basic advantages of onboard processing apply, and even more strongly to packet-switched systems.

Consider the following advantages to packet switching onboard the satellite:

- Packets can be routed to downlink users and satellite beams as needed, thus making efficient use of downlink resources without requiring dedicated circuits that might be used only in bursts.
- Multicast routing (one-to-many) is easily accomplished.
- With sufficient caching of packets on board, latency can be reduced by halving the user-to-user satellite transmission time.
- Data-link-layer protocols can be utilized to mitigate satellite link impairments, including path blockage due to the motion of ground terminals.

A major decision that needs to be made in the design of an onboard packet-switched system is the choice of protocol layer or layers to be implemented. If only the *physical* and *data-link layers* (layers 1 and 2 of the OSI Reference Model) are implemented, the satellite system is analogous to the connectivity of a LAN (local-area network). The system could directly interconnect satellite users to each other, but would rely on ground-based gateways for connectivity to users that are external to the system and possibly to satellite users that are connected to different satellites of a global system.

Onboard processing systems that include routers that operated at the next higher protocol layer, namely, the *network layer* (OSI layer 3), would be able to interconnect users in more capable ways within and external to the system in accordance with global networking standards. It would also be more effective in routing traffic between users within a global system that are attached to different satellites.

Operating the onboard processor at higher protocol layers could also be considered. For example, operation at the *transport layer* (OSI layer 4) would permit the caching of TCP packets that could be retransmitted with only a single round trip to the satellite to reduce latency. Operation at even higher layers, for example, the *application layer* (OSI layer 7), might also be considered with email and Web servers on board. (Store-and-forward messaging is a simplified version of this.)

2.8. Ground Processing Tradeoffs

The point of view taken in this article has been to illuminate the advantages and capabilities of onboard processing. However, it must be acknowledged that onboard processing comes with a price, namely, the need to place hardware on the satellites with the corresponding burdens of additional complexity, power, and weight. Whether this tradeoff is worthwhile depends on the system application. For example, satellites used for wide-area television broadcasting would gain little from processing; high-power translating repeaters generally suffice.

Another major tradeoff issue is where to place the processing functions that may be desired — in space or on the ground? Figure 9 illustrates some of these tradeoffs. Strong feeder links could, in principle, be used to relay all signals (from all antenna beams) received at the satellite in analog (or sampled) form to the ground. Virtually all processing functions (demodulation–remodulation, antenna beamshaping, packet routing, etc.) could then be done on the ground and relayed on strong links back to the satellite for subsequent retransmission, although at the cost of additional latency.

Designers of different systems may come to different conclusions regarding the use of processing and where to implement it. The next section presents selected system implementations that represent a range of processing applications and the range of design choices made.

3. EXAMPLES OF ONBOARD PROCESSING SYSTEMS

This section provides examples of satellite systems that have included onboard processing to varying degrees. Each will be described briefly to relate it to the types of processing described in the preceding sections.

3.1. Government Systems

A number of pioneering onboard processing systems were developed under U.S. government sponsorship:

- **Lincoln Laboratory Experimental Satellites 8 and 9 (LES-8 and 9)**, shown in Fig. 10, launched in 1976, were among the first to include onboard demodulation–remodulation and spread-spectrum despreading. These feature were included to demonstrate satellite-to-satellite crosslinks, UHF–Ka-band frequency crossbanding, and antijamming protection for a variety of military platforms. A more complete description can be found in Ward's study [3].
- The **Defense Satellite Communication System (DSCS)** [4], operating in the SHF region of the spectrum since the early 1970s, is an example of a system that utilizes flexible antenna patterns both to shape coverage areas and reject interference. Microwave lens antennas and steerable parabolic dishes are used. The signal processing that supports the beamforming is shared between the satellite and the ground.
- The NASA **Advanced Communication Technology Satellite (ACTS)** [1] has several onboard processing features. It includes a number of narrow beams, some of which can be rapidly steered. It also carries a signal processor very much like that shown in Fig. 5.
- The **Fleetsat EHF Packages (FEPs)** built by Lincoln Laboratory were launched in 1986 and 1989. The FEPs took jamming-resistant onboard processing a significant step forward with the

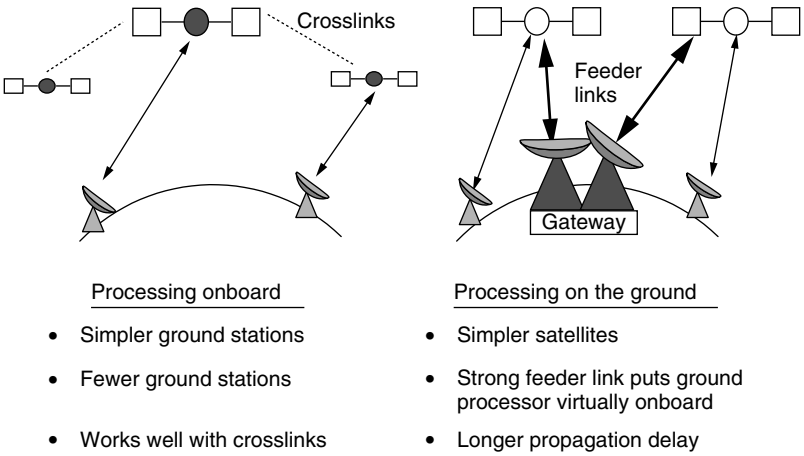


Figure 9. Processing onboard compared to processing on the ground.

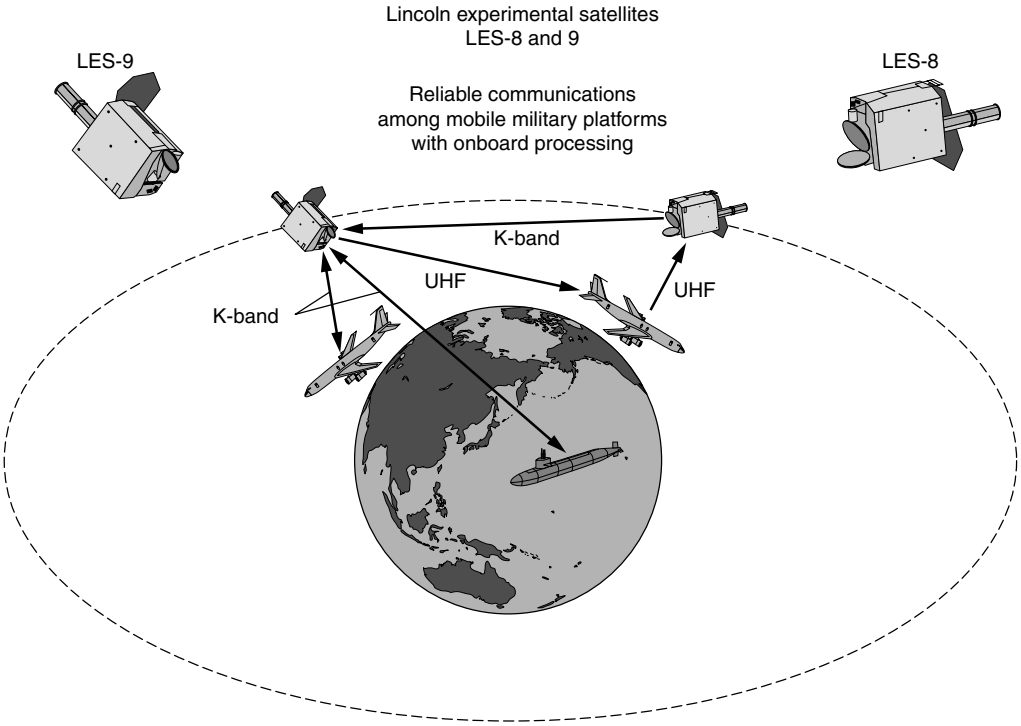


Figure 10. Lincoln laboratory experimental satellites 8 and 9 (LES-8 and 9).

inclusion of extensive demodulation–remodulation of multiple-user onboard access control realizing a true "switchboard in the sky" [5] and a steerable antenna. The FEPs can operate as a self-contained demand-assigned circuit-switched system. The FEPs are the forerunners of the Milstar [6] system, which includes more antennas, more channels, higher data rates, and satellite-to-satellite crosslinks.

3.2. Commercial Systems

A growing number of commercial system are using or are expected to use onboard processing.

Iridium provides worldwide digital voice and paging services to handheld user terminals [7]. The system makes extensive use of onboard processing. The spaceborne part of the system consists of 66 low-orbit (780-km-altitude) satellites in six orbital planes. Each satellite projects 48 narrow beams onto the earth; each beam is about 30 mi in diameter. In order to provide worldwide connectivity, the satellites are crosslinked to each other. A user signal (at L band) emanating from a handheld terminal is demodulated on board and crosslinked (at K band) around the Iridium constellation until it is downlinked to a gateway terminal connected to the terrestrial telecommunication infrastructure.

The **Thuraya** system provides voice and data services to most of Europe and portions of Asia and Africa. Each of its geosynchronous satellites includes a 12.25-m antenna that is used to form 250–300 spot beams. Its onboard processing is described as [8] follows:

- On-board digital signal processing (DSP) to facilitate interconnectivity between the common feeder link coverage and the spot beams to make effective use of the feeder link band and to facilitate mobile to mobile links between any spot beams
- Digital beamforming capability that will allow Thuraya to reconfigure beams in the coverage area, to enlarge beams, and to activate new beams. It also allows the system to maximize coverage of "hot spots," or those areas where excess capacity is required
- The flexibility to allocate 20% of the total power to any spot beam
- The flexibility to reuse the spectrum up to 30 times and therefore use the spectrum efficiently

The **Teledesic** [9] and **Astrolink** [10] systems have been under development since the mid-1990s. They both aim at providing broadband data services to small terminals around the world. As originally conceived, the Teledesic would have included a constellation of numerous low-orbiting satellites with extensive onboard demodulation–remodulation, routing, and crosslinking. Astrolink uses geosynchronous satellites with onboard ATM (asynchronous transfer mode) switches. Both systems' satellites have multibeam antennas.

It is too soon to predict the commercial success or failure of any these ventures. However, they are all pioneers in the technology of onboard processing, which will inevitably become a key part of the global information infrastructure.

BIOGRAPHY

Steven Bernstein received the S.B. and S.M. Degrees in Electrical Engineering from the Massachusetts Institute of Technology in 1964 and the Degree of Electrical Engineer from MIT in 1966. In 1966 he joined the Communication Division in of MIT Lincoln Laboratory, where he has

worked on and led a wide variety of communication and networking projects. His early assignments included work on a pioneering Air Force frequency-hopping UHF satellite communication system and the development of an ELF system for communication to submerged submarines. Following these staff assignments, he managed a succession of satellite communication projects at UHF and EHF for the U.S. Navy, Army, and Air Force. Following a one-year leave of absence to teach graduate level courses in digital communication at ENST (Paris, France), he returned to Lincoln and was given the responsibility to lead the development of a new satellite operations center. He also has managed the Optical Communication Technology Group, which developed novel techniques for optical ground and space communication. He is currently the Associate Leader of the Applied Communications and Information Technology Group. His areas of technical interest are satellite communication systems, channel coding/decoding, optical transmission, and applying these technologies to practical problems.

BIBLIOGRAPHY

1. F. M. Naderi and S. J. Campanella, NASA's Advanced Communications Technology Satellite (ACTS)—an overview of the satellite, the network, and the underlying technologies, *AIAA Int. Communication Satellite Systems Conf.*, 1988, pp. 204–224; AIAA paper 88-0797.
2. L. J. Ricardi, Communication satellite antennas, *Proc. IEEE* 336–369 (March 1977).
3. W. W. Ward, *Developing, testing and operating Lincoln Experimental Satellites 8 and 9* (*LES-8/9*), Lincoln Laboratory Technical Note 1079-3, Jan. 16, 1979.
4. R. Donovan, R. Kelley, and K. Swimm, Evolution of the DSCS Phase III satellite through the 1990's, *IEEE Int. Conf. Communication*, 1983, Vol. 2, pp. 611–619.
5. M. D. Semprucci, The first "Switchboard in the Sky": An autonomous satellite-based access/resource controller, *Lincoln Lab. J.* **1**(1): 5–18 (1988).
6. L. F. Kwiaykowski et al., The Milstar system, *AIAA Int. Communications Satellite Systems Conf.*, 1994, pp. 744–748; AIAA paper 94-1013.
7. Iridium, World Wide Web (WWW) Website *http://www.iridium.com*.
8. Thuraya, Website *http://www.thuraya.com*.
9. Teledesic, Website *http://www.teledesic.com*.
10. Astrolink, Website *http://www.astrolink.com*.

COMMUNICATION SYSTEM TRAFFIC ENGINEERING

APOSTOLOS K. KAKAES
Cosmos Communications
Consulting Corporation
Centreville, Virginia

1. INTRODUCTION AND MOTIVATION

At its core, the traffic engineering problem is very easy. At least stating the problem is! Indeed, traffic engineering falls into the relatively small set of problems that happen to be relatively easy to articulate, but whose solutions provide a window to a world that is new and in many ways surprisingly complex. Even though our current emphasis is on communication systems and networks, the problems and associated solutions have much broader applicability. We will occasionally refer to these other areas of applications, as they can also aid in the understanding of some of the underlying principles.

In a somewhat abstract sense, the problem can be formulated as follows (see Fig. 1). To a given pool of resources, consisting of N "servers," "customers" arrive with the intention (or need) to use one of the servers for a certain amount of time. Indeed, if one of the resources is available, it is "held" or "occupied" by the arriving customer. If no resources are available, the arriving customer is blocked. Thus we observe that the entire "offered load" is "split" into what becomes "served load" and "blocked load." The basic question then appears to be simple and final: "What is the probability that the system will be full, and thus will be unable to serve a potential customer, that is, how much of the offered load is served and how much is blocked?" Once this relationship is fully understood, it can be exploited to design, or size, a given system to meet certain performance objectives, or at least determine what the performance may be for a given set of conditions.

Note that the set of resources can be, for example

- Channels between points A and B (as is often the case in communication networks)
- Tellers at the bank
- Check-in counters at an airport

Also note that the language is not always helpful in uniquely identifying the problem! Arriving "customers" can be telephone calls or passengers who need to board a plane. Similarly, one would not ordinarily state that "a customer at a bank held (or occupied) the bank teller." In an abstract sense, however, these statements all refer to the same situation, where the resource is being used by an arriving entity, thus is unavailable for use by others, until, of course, it is "released." Once again, in this article, as a rule, we will refer to arrivals as *calls* that attempt to use one *channel* for a certain amount of time, called the *holding time*. Similarly, the terms "resource," "channel," and "server" will be used interchangeably, as indeed they are so used in the real world.

As we will see in detail later, one of the most important quantities that arise in the analysis of such a system,

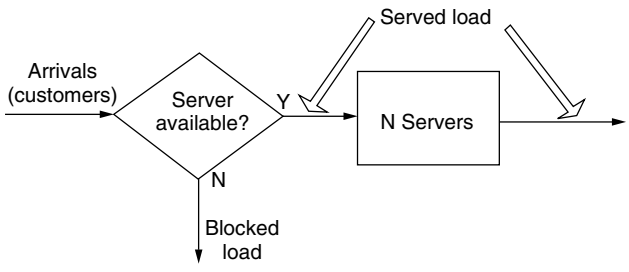


Figure 1. Schematic diagram of a service facility.

is the product of the average call arrival rate, commonly denoted by λ, and the average holding time, also commonly denoted by $1/\mu$. We thus define the *offered load*, A, by

$$A = \lambda \times \frac{1}{\mu} = \frac{\lambda}{\mu}$$

What are the units of the offered load? λ has the units of time^{-1} ("calls" is dimensionless), whereas $1/\mu$ clearly has the units of time. Thus A is dimensionless.[1]

It is in honor of the Danish mathematician A. K. Erlang, the father of traffic engineering, that a load A is referred to as "A erlangs." The basic problem can be stated very concisely as follows. An offered load of A erlangs is offered to a pool of N servers. What is the blocking probability, that is, the probability that an arbitrary arrival finds no server available? As it turns out, the process of answering this seemingly simple question entails more complexity than one might think! It is intuitively clear that, for a given value of N, the blocking probability will increase with increasing offered load, A, thus with increased values of the arrival rate and/or the average service time. What is not necessarily as obvious, is that the *arrival statistics* play a vital role. One can easily appreciate this by considering two particularly simple examples. These examples will also allow us to introduce a number of concepts used later in this article. We will assume that the

- Number of channels, $N = 1$
- Average arrival rate, $\lambda = 1$ arrival/min
- Average holding time, $1/\mu = 1$ min, and it is *constant* for each and every call

Examples 1 and 2 differ only in the way in which calls arrive to our one and only channel, as follows. We should emphasize that for now, we will not worry about how realistic (or unrealistic) these examples are. This point will be amply discussed later when we discuss modeling arrival processes in realistic situations.

Example 1. Calls arrive every minute on the minute. One might think of this as being "organized" by an external entity that coordinates the call arrivals to occur in this fashion: on the minute every minute.

Since each arrival occupies the channel for precisely one minute, it departs just as the next call arrives, and the new call occupies the channel.

We can easily see that the fraction of calls that find the system unavailable is 0, that is, that the blocking probability is 0.

Similarly, it is just as easy to see that the system utilization is 1, that is, the system is fully utilized, or as the Chief Financial Officer would say "we are making money at 100% of the potential to make money." From a practical perspective, the question of utilization is often of importance, as it is the *utilization* of the system that generates revenue!

This is perfection! No user is disappointed (blocked) and we (the network provider) make as much money as possible.

Example 2. All 60 calls arrive within the first minute following the first call's arrival, say, at the beginning of an hour. Once again, one might think of this as being "organized" by an external entity that coordinates the call arrivals to occur in this fashion: a whole bunch of them close together and then nothing for the remaining hour.[2]

It is clearly seen that the first arrival in the hour occupies the channel, and all remaining 59 calls find the channel occupied, thus are blocked. Thus the fraction of calls that find the system unavailable is $\frac{59}{60}$, with a blocking probability of almost 1.

Similarly, it is just as easy to see that the system utilization is $\frac{1}{60}$, that is, we are making money a mere 1.67% of the potential to make money. This is about as bad as it can get! Almost all users are disappointed (blocked), and we (the network provider) make essentially no money — this is one way to get into bankruptcy!

We can summarize Examples 1 and 2 in a manner that will also prove useful later, as shown in Fig. 2. Note that this is a two-dimensional space representing all the possible values of blocking probability, in the range [0,1] and utilization, also in the range [0,1]. Any point (x, y) in this space represents a system for which the blocking probability is x and the utilization is y. As a result, the points (0,1) and (1,0) represent the limiting cases discussed in the two previous examples, respectively. Clearly, in general, we would like to be operating at a point (x, y) with x as small as possible and y as large as possible, specifically, in the upper left quadrant of the unit square, as shown in Fig. 2.

What is the *only* difference in the two examples? The *arrival process*. Thus it becomes important that we undertake a study of the arrival process statistics in order to investigate the traffic engineering problem. We do so

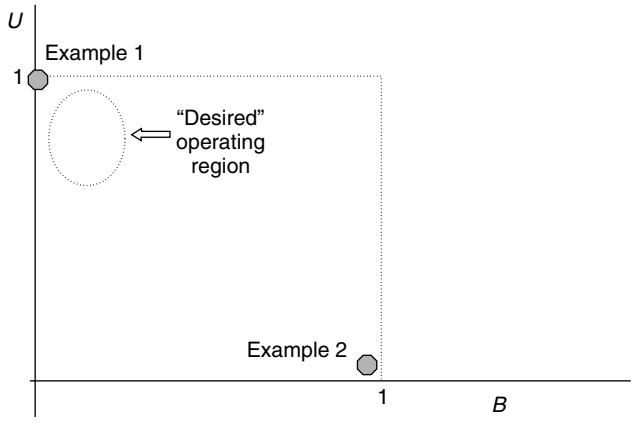


Figure 2. Utilization and blocking for different arrival statistics.

[1] It is worthwhile to note that, as we will see later, μ is called the "service rate" and then the load becomes the ratio of two rates: the arrival rate and the service rate.

[2] Note that the average arrival rate is still 1 call per minute and the average holding time is still 1 minute, thus the offered load is still 1 erlang, exactly as in Example 1.

in Section 2. In Section 3 we analyze the most common models used in traffic engineering: the Erlang B model, followed by the Erlang C and the Poisson models. The fundamental similarity and difference of these models lies in the way in which blocked arrivals are treated: In Erlang B it is assumed that blocked calls depart the system unserved. In Erlang C, it is assumed that arrivals that find the system full, thus are blocked, enter a queue, which is served in a first-in first-out (FIFO) manner. However, it is important to note early on that the Erlang C model assumes that blocked arrivals are willing to wait an unbounded amount of time for service. Thus, one might consider the Erlang B and the Erlang C as diametrically opposite models: One assumes that the arrivals are willing to wait 0 time, while the other assumes that the blocked arrivals are willing to wait as long as it takes. Clearly, there ought to be a model for some sort of in-between case! Indeed, the Poisson model is precisely such a model. We will elaborate on these three models, individually and collectively, in Section 3.

Finally, in Section 4 we discuss the problem of random access, which is of a somewhat different flavor, as articulated in time; while often not associated with conventional traffic engineering problems, the random access problem is indeed a traffic engineering problem and thus needs to be discussed herein.

2. MODELING THE ARRIVAL PROCESS

Clearly, the tools of statistics and random processes are needed in order to characterize the arrival process. This can get very technical and, while these technicalities are important, one can miss the essence of the problem. Our present discussion is limited to an intuitive level and the technicalities can be found in any number of books on probability theory and on random processes, such as the classic books by Feller [1,2]. We discuss the more critical concepts, in particular as they impact traffic engineering issues.

2.1. Stationarity

As noted earlier, load is measured in erlangs, a dimensionless quantity. First and foremost, defining the load has little to do with a period of time. Many practicing engineers, unfortunately, think of one Erlang as being "one call occupying one channel for one full hour." While this statement can be interpreted correctly (we will not attempt to do that here), it is very open to misinterpretation, which unfortunately happens too often. This is one reason why we will not make such usage here and we will use the definition of load in a way that leaves no room for misinterpretation. However, to do that, we need the concept of stationarity. Indeed, implicit in most of traffic engineering is the assumption that the problem is stationary.

Although the mathematical definition of stationarity is not too difficult, we will avoid such technical issues, appealing to one's intuition instead. Simply put, we require that the statistical nature of the underlying problem remain the same over time. As a trivial but helpful example is that of a sequence of flipping the same coin. While each outcome is in general different from the other outcomes, stationarity implies that the probability distributions that govern these outcomes remain the same. Note that we are *not* saying that it is a fair coin — it can be any coin. Stationarity merely says that whatever the statistical behavior is, it remain the same over time.

In a way the question is simple: Is the statistical nature of calls arriving to our resources (say a switch) the same at all times? Clearly, from a purist's perspective the answer must be "no," since things do change, people behave differently over time, and so on. Thus we do not have stationarity. But from a practical perspective, we can ask a slightly different question: Is it realistic to assume that over some nontrivial amount of time, the statistical nature of the problems remains constant? If so, then we need to ensure that such a period of time is long enough so that enough events take place, and yet not too long, so that the underlying statistical dynamics of the problem remain constant. In general, how long is long enough and not too long? While the technical definition of stationarity could be called on, we will appeal more to one's intuition. Clearly, what happens at 3 A.M. is different than what happens at 9 A.M. But how about 8 A.M. and 9 A.M. — is it sufficiently the same, from a statistical nature point of view, or is it changed and thus needs to be considered separately? The traditional answer to the question has been to not pay too much attention to it! In the early days of developing telecommunication networks (when the voice network was the prime example) the computing power to do a more elaborate analysis was simply not available, thus the technical community used a one-hour period as being long enough and not too long to be considered as a period of stationarity. Clearly, then, the hourly division of the 24-h day made some sense. Although some of the original premises are no longer valid, we still tend to use hourly data. To some extent, this has become a chicken-and-egg problem: switch manufacturers tend to default their software programs and so on, into one-hour periods because practicing engineers are used to that. Practicing engineers, in turn, tend to use hourly data because the switch provides these data! Even though in many of today's switches there are options of keeping and reporting other statistics, most engineers do not know that such options even exist, let alone how they might use it. So that brings us to the traditional hourly engineering. We, however, will not make any such assumptions. We will treat the problems from a slightly more abstract point of view, and simply assume that stationarity exists and not be concerned over how long it lasts for.

2.2. Busy-Hour Engineering

The discussion above naturally brings us to the notion of *busy-hour engineering*. As discussed above, there is nothing particular or important about a one-hour period. It was just convenient at one time. Second, the definition of *busy hour* is straight forward, albeit with its own difficulties. If we accumulate load on an hourly basis, then there are 24 measurements for a day, and the highest one is the busy hour. Assuming that the statistical nature of the problem does not change over days (which may or may

not be true), we can aggregate the same-hour data over a number of days and thus generate the concept of "busy hour" over periods of many days, weeks, or even months. Once again, stationarity can play an important part here: In an area serving ski resorts, clearly the statistics change from winter to summer months, so we need to be aware of the stationarity issues both in the small and the larger scale of time constants! As we indicated earlier, we will follow the usual approach and assume that the problem is stationary. A detailed analysis of the complexities arising from considering the *day-to-day variation* can be found in the book by Ash [3].

Two different fundamental questions are

- *Why* do we do busy-hour engineering?
- *Should* we do busy-hour engineering?

The answer as to *why* lies in history! Without getting into a deep historical retrospective, suffice it to say that in the (traditional) regulated monopoly environment, there were both regulatory and business reasons that were pointing toward taking a busy hour engineering approach to the network design problem. It is intuitively obvious that such an approach resulted in a network that was more expensive than it might otherwise be. This observation notwithstanding, the combination of business and regulatory issues in a regulated monopoly provided sufficient justification to indeed adopt such practices quite widely.

The question of continuing to do busy hour engineering in the new days of deregulation of telecommunications on a worldwide basis is a harder one to answer! Indeed, we will not answer it, as it becomes an economic analysis. Any such analysis must be interdisciplinary by nature. In this article we provide the traffic engineering tools needed to carry out such an analysis, but the point should be raised as too often methodologies are adopted only because "that's how it has always been done."

2.3. Definition of Load

As indicated earlier, the proper measure of load is the product of the average arrival rate times the average holding time. This product is referred to as "*A* erlangs." Even though on its way out, an old-fashioned measure still exists, so one should be aware of it: 100 call-seconds per hour (CCS). (The acronym is derived as follows: C, from the Latin initial for hundred; C, call; S, seconds; per hour, by common agreement or convention.) Some, especially older, systems still use CCS, but its usage is strongly discouraged. As will become apparent later, one can easily convert CCS to erlangs: 36 CCS = 1 erlang. Furthermore, the relationship of CCS to erlangs is analogous to that of degrees to radians in measuring angles; the first is arbitrary and based on some agreed-on convention. (Why does a complete rotation correspond to an angle of 360°? Why 360 and not, say 480? The number 360 is just as arbitrary as 480 or any other number! It only "feels" better because we have been used to it!) Much as radians are dimensionless measures of angles and defined as the ratio of two lengths, the erlang is dimensionless and is the ratio of two rates, which is an item we will return to shortly, For now, suffice it to say, that we will use only erlangs from here on, but one must be aware of the possibility of encountering CCS!

2.4. The Poisson Arrival Process

One of the most commonly made assumptions is that the arrivals occur according to a *Poisson process*. What does that mean? What are the practical implications of such an assumption? What are the practical and theoretical implications if this assumption does not hold? Before we attempt to answer some of these questions, we must develop at least some understanding of this concept, which is indeed of paramount importance.

Consider an arrival process of, for example, calls to a switch that satisfies one of the following three properties. We will consider them one at a time, and then use them to define the arguably most fundamental modeling assumption in traffic engineering: The Poisson process.

Property 1: The Infinitesimal Generator. Suppose that the arrival process is such that over an infinitesimal period of time h, the probability of exactly one arrival and exactly zero arrivals are given respectively by[3]

$$P(1 \text{ arrival}) = \lambda h + o(h)$$

$$P(0 \text{ arrivals}) = 1 - \lambda h + o(h)$$

Note that, in a very general sense, for arbitrarily small amount of time h, specifically, in the limit as h goes to 0, the probability of 1 arrival goes to 0 and the probability of no arrivals goes to 1! What we are saying here is much more profound, albeit not obviously so! We are saying that, in the limiting sense as time goes to 0, P (1 arrival) goes to 0, *linearly* with time, with constant of proportionality λ. Similar statement can be made about the probability of zero arrivals.

It is a direct consequence of these assumptions that the probability of two or more arrivals is $o(h)$. In other words, the probability of two or more arrivals over a small interval of time *goes to 0 very rapidly*. As it turns out, from a practical point of view $o(h)$ functions can be disregarded without any harm.

Thus, we are stipulating that over an arbitrarily small interval of time h (and thus the term "infinitesimal generator"), "essentially" (i.e., we ignore $o(h)$ terms) the following happen:

- One arrival occurs with probability λh.
- No arrivals occur with probability $1 - \lambda h$.
- Two or more arrivals occur with probability 0.

[3] The notation $o(h)$ (read as "little o of h") is common and is used to designate any function that satisfies the following limiting property:

$$\lim_{x \to 0} \frac{f(x)}{x} = 0$$

i.e., the function $f(x)$ approaches 0 (as x goes to 0) **faster** than x itself goes to 0.

We further assume that the arrivals in disjoint intervals are independent events, thus the number of such arrivals are independent random variables.

This "infinitesimal" property of arrivals may or may not appeal to one's intuition as being a good model for our problems. We will table further discussion of this model/property until we discuss two other options and then come back and compare all three of them.

Property 2: Poisson Distribution of Number of Arrivals. Here we consider an arbitrary period of time T — no longer an infinitesimal one — and ask the question of how many arrivals occur in it. Let's suppose that the number of arrivals K in the time period T has the Poisson distribution with parameter, λ. Recall then, that the probability distribution of the number of arrivals, K, is given by

$$P(K = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \quad k = 0, 1, 2, \ldots$$

We further assume that the number of arrivals in disjoint intervals of times T_1 and T_2 are independent random variables.

Property 3: Exponentially Distributed Interarrival Times. Here we assume yet another property of some arrival process: The time between two successive arrivals (called the *interarrival time*), X are independent random variables, exponentially distributed with some parameter λ. Recall that X has the exponential distribution if its probability density function is given by

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for} \qquad x \geq 0$$

Among other natural and useful questions to ask, from the engineering perspective, are the followings:

- Which of these three models, if any, is a realistic model for traffic arrivals to a switch or to a set of channels?
- How would one go about confirming or rejecting the modeling assumption that one of these models is indeed a good one?
- If one of these models is not quite good enough, could another one of them be better (or worse for that matter), but harder (or easier) to work with?
- If none of these three models is valid in a given practical situation, and we have established methodology to discover that fact, how do we go about finding a better one?
- Continuing with the previous item, how much error in our performance evaluation and/or design are we making by using the wrong model? It may be worthwhile to use an easier model, even if it is not quite right, if the deviations from the more accurate one are small.
- Assuming that we do find a better model, what will the performance be?

This may come as somewhat of a surprise, and it is intuitively not obvious at all, but the following is a fundamental theorem with many practical implications:

Theorem 1. If an arrival process satisfies any one of properties 1, 2, or 3, then it satisfies *all* of them and the arrival process is called a *Poisson arrival process*, or simply a *Poisson process*. The constant of proportionality λ is the same in all three defining properties and is called the average arrival rate of the arrival process. It is also referred to as the *arrival rate*, or even just the *rate*, or the *parameter* of the Poisson process.

In this article we will forego all proofs, but if this is the first time one sees this, it is highly recommended that one try to prove the equivalence of these three properties, as indeed they form the defining properties of the Poisson process.

While we could spend a large amount of time in investigating the deeper properties of the Poisson process, in the interest of brevity, we will not; however, we will briefly point out some of the more important salient points:

2.4.1. "Random" Arrivals. As it turns out, if a large population generates "arrivals" in a manner that involves *no external* control, the overall resulting arrivals will constitute a Poisson arrival process. This is precisely the reason why the Poisson arrival assumption has been used in telephony so much: when a user A makes a call is not "controlled" by any entity. Thus, precisely when user B makes a call is neither influencing nor is being influenced by user's A actions. This is also why some refer to the Poisson process as being one in which arrivals occur "at random." This, however, can be misleading, since "random" arrivals can obey any number of statistical properties and does not necessarily have to be Poisson. We must also emphasize that in practice the external controls *may* be there implicitly, thus "destroying" what might have otherwise been a Poisson arrival process! The requirement is that there be *no controls*, explicit or implicit ones!

2.4.2. Merging of Poisson Processes

Theorem 2. If n independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_n$ are merged, then the resulting process is Poisson with rate $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. In other words, merging several (independent) Poisson processes preserves the Poisson property.

2.4.3. Splitting a Poisson Process. How about the converse? Does splitting the arrivals of a Poisson process into two (or more) subprocesses preserve the Poisson property? As we are about to discover, the converse problem is a bit more complicated, the results are a bit more surprising, and the practical applications a bit more intriguing.

Suppose that a Poisson process with rate λ is "split", or bifurcated, into two (or more) sub processes. The question then is whether each subprocess is or is not Poisson (see Fig. 3). The mechanism labeled "load regulator" routes each arrival to the set of resources (channels) labeled

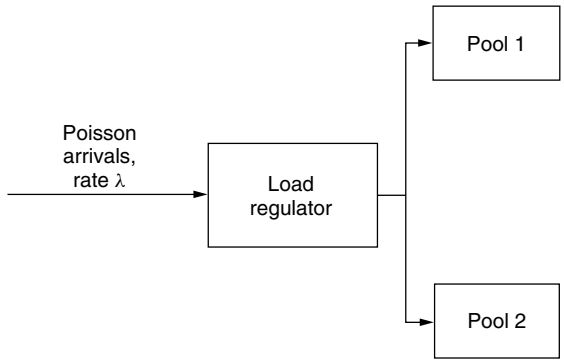


Figure 3. Bifurcation of Poisson arrivals.

"pool 1" or to the one labeled "pool 2." The mechanism by which it decides where to route each call is critical and we will be elaborating on this in the next few paragraphs.

As a first step, assume that the load regulator routes each call to the two pools in an alternating fashion. For convenience, we may assume that the first call is routed to, say, pool 1. Once again, assuming that the original process is Poisson with rate λ, the question is what can be said about each subprocess, i.e., the arrivals as seen by each pool of channels. The answer is provided in the next theorem.

Theorem 3. A stream of Poisson arrivals with rate λ is split into n substreams. Each call is routed to substream i with probability p_i, *independently* of all other calls, or the state of the system (thus of prior arrivals, etc.) Then each sub stream is also Poisson with rate $p_i\lambda$. Conversely, if the routing decision is made in some dependent fashion, including based on the state of the system, then the sub processes are *not* Poisson.

This is a good illustration of the fact that the Poisson process is a "delicate" entity. It is very easy to destroy it! Two practical examples will illustrate the point.

Example 3. Consider a network, a portion of which is as shown in Fig. 4. The direct route between nodes A and B is preferable, so the "load regulator" sends calls to the direct route, unless all channels are occupied in which case it sends the call via the alternative route, namely, via node C. Then, from a practical perspective, it is important to ask if the set of channels on the alternative path, links A–C and C–B are offered a Poisson arrival stream.[4] Since the decision to send the call to the A–B link or the A–C–B one is based on the state of the system, the substream of offered calls to the alternative route is not Poisson, even though the original traffic arrivals is assumed to have satisfied the Poisson assumption.

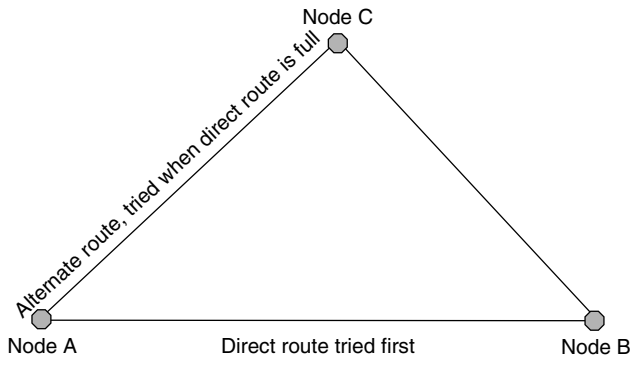


Figure 4. A simple 3-node network.

Example 4. In modern mobile communications networks, it is not uncommon for a given region to be capable of being served by two (or even more) cells. This is an example of the well-established concept of underlay/overlay. We will not get into the pros and cons of such designs or the radio engineering issues surrounding them, but the concept is simple. A given geographic region is divided into small cells as well as large ones (sometimes called "umbrella cells"). A given call is first attempted on the small cell, and if all channels are busy, it is then attempted on the large one. It is obvious that this is a problem equivalent to the one above. The point is the same — the offered load to the alternative set of resources is no longer Poisson.

As one can see, the mechanics and perhaps the reasons by which the load is "split" is rather irrelevant. However, destroying the Poisson property is not all that difficult to do! Both from a theoretical and a practical vantage point, we must be aware of that possibility and take appropriate measures.

2.4.4. Modeling Assumptions. From either a theoretical or a practical perspective, it is often the case that one needs to confirm or refute the assumption that arrivals form a Poisson process. It is often easy to make such an evaluation using one of the three properties of the Poisson process, whereas using the other two properties may be very difficult and/or impractical. We emphasize that these three properties are equivalent and one should avail oneself of whichever one is easier to work with. Furthermore, if one of these properties is shown not to hold, the others do not hold either and the assumption of the arrival process being Poisson is not valid. From the practical perspective this can have very significant implications to which we will return shortly.

2.4.5. If Not Poisson, Then What? If an arrival process is not Poisson, is there anything we can say about it? Yes, indeed there is. The Poisson process is merely one particular point in a continuum of possibilities. What is particular about it can be summarized in one sentence — its peakedness is 1. Peakedness is defined as the ratio of the variance of the number of arrivals to the mean number of arrivals. If the arrivals are Poisson, then we recall that the number of

[4] Clearly in a network environment this is only a portion of the problem. We are specifically and narrowly concerned with this particular portion of the traffic. In network designs, we obviously need to integrate this into a considerably larger set of issues. The reader is referred to Ash's treatise [3] for a comprehensive analysis of the issues in such network designs.

arrivals over any time of duration T satisfies the following conditions:

- Mean number of arrivals = λT.
- Variance of the number of arrivals = λT.
- Thus peakedness = variance/mean = $\lambda T/\lambda T = 1$.

Arrivals for which the peakedness $Z > 1$ is called "peaked" traffic, whereas if the peakedness $Z < 1$ we say that the traffic is "smooth."

We observe that the arrival processes discussed at the outset of this article satisfied:

- When arrivals arrive on the minute every minute, then variance = 0; thus peakedness $= 0 \ll 1$, thus we say that the traffic is very smooth.
- When arrivals came all "bunched up," the variance was large (with the same mean), thus peakedness $Z \gg 1$; thus, we had very peaked traffic.

Recalling our earlier discussion of Poisson traffic being often referred to as "random" traffic, we can make the additional observations that

- If no controlling mechanism exists, then arrivals will be in accordance with the Poisson assumption: $Z = 1$.
- If a controlling mechanism (implicit or explicit) "coordinates" arrivals to occur on a "regular" basis (thus lowering the variance), then the traffic exhibits some smoothness, where $Z = 0$ is the limiting case where there is no variability in the arrival stream.
- If, on the other hand, the controlling mechanism tends to "bunch up" traffic, then it is peaked traffic and the peakedness Z is arbitrarily large, according to the degree to which the "bunching up" takes place.

2.4.6. Time versus Call Congestion. Finally, Poisson versus smooth versus peaked traffic can be seen to have one more impact that needs to be discussed. Two related concepts are those of

- *Call congestion* (CC)—the fraction of call arrivals that find the system full. This is indeed what we have been referring to as probability of blocking.
- *Time congestion* (TC)—The fraction of time that the system is fully occupied and thus can not admit a new arrival.

Note that there is no a priori reason that these two measures of congestion should be equal to each other. Indeed, if the arrivals do form a Poisson process, they are. Otherwise they are not. Figure 5 summarizes the relationship where the time congestion and the call congestion occupy the two sides of a balance beam. In Fig. 5a arrivals are assumed to be Poisson ($Z = 1$); thus CC = TC. Figure 5b illustrates smooth traffic ($Z < 1$). In that case TC > CC. Note that this, in principle, is desirable, as it implies that utilization is higher while blocked arrivals are fewer (than would be the case with $Z = 1$). Of course, the converse is shown if Fig. 5c, where

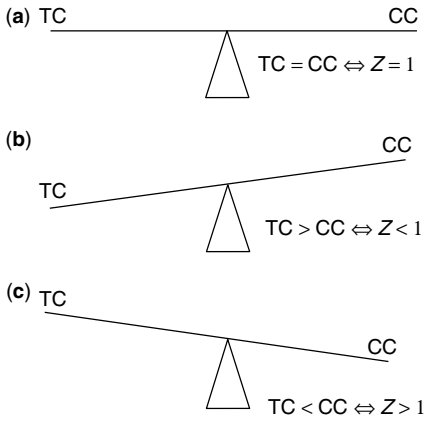


Figure 5. Time congestion versus call congestion.

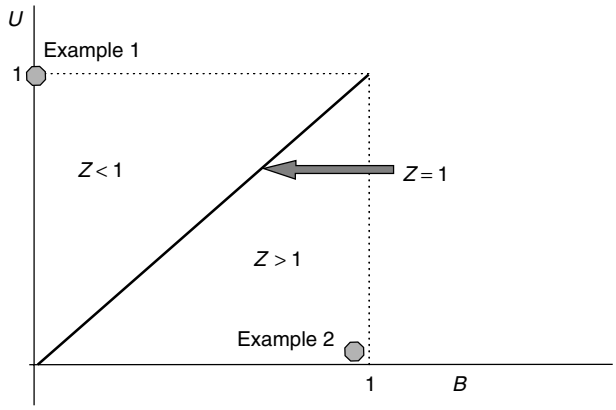


Figure 6. Utilization versus blocking for Poisson arrivals and $N = 1$.

$Z > 1$. Indeed, see Fig. 6, which is a generalization of what we discussed earlier in Fig. 2 for the case that $N = 1$. As the offered load ranges from 0 to infinity, one can see that a path is traced in the unit square of (B, U). The precise path depends on the peakedness. If $Z = 1$, the path is on the "45°" line, as indeed $U = B$ (for the case $N = 1$; we will shortly return to the more interesting case where $N > 1$). This line separates the unit square into two regions: One for smooth traffic ($Z < 1$) and one where traffic is peaked ($Z > 1$) as illustrated in the figure. While the straight line is a direct result of the system having just one channel ($N = 1$), the general concept is valid and will be generalized shortly, after we obtain the so-called result of the Erlang B model.

3. THE ERLANG *B* MODEL

As mentioned earlier, the most common model used for determining the quantitative relationship between the offered load A, the number of channels N, and the blocking probability $B(A, N)$ is the Erlang B model.

The Danish mathematician A. K. Erlang developed what is now known as the Erlang B model early this century, just as telephony was being born.

3.1. Modeling Assumptions and Associated Notation

The Erlang B model is used so widely, often without adequate understanding of the underlying assumptions.

Although a variety of modifications to the basic model are possible, and we will point to some of them, the basic model makes the following fundamental assumptions:

- The "service facility" or the "system" consists of a fixed number of "servers" or channels, N.
- Calls arrive to the system according to a Poisson process with an average arrival rate of λ calls per second.
- If a server, or channel, is available on a call's arrival, the call seizes a server. If more than one server is available, any one of them is seized — it makes no difference which one, as all servers are equivalent in all senses.
- A call occupies the server for an amount of time that is exponentially distributed with parameter, or rate μ, and independent of all other calls, arrivals, system state, or anything else. Thus the average service time, or average holding time is $1/\mu$. It is precisely the desire to express the service statistics in terms of the service rate (rather than the average service time itself) that early on we had the somewhat surprising notation of $1/\mu$ for the average holding time.
- If all servers are occupied at a call's arrival instant, the call departs the system unserved. In particular, it does *not* attempt to be served "a little bit later," try again, or any variant thereof.

Before we proceed, a brief notational comment. The system that we just described is often referred to as an $M/M/N/N$ queueing system. The notation is straight forward, and we present it in a somewhat more general context in the next paragraph.

In general, the notation $F_1/F_2/F_3/F_4/F_5$ implies the following service facility:

- In field F_1, a letter designates the interarrival times, namely, the arrival process. So, for example, M implies that the arrival process is Poisson. Maybe it should be P, you say! It is M, because a Poisson arrival process contributes to the system being a *Markov* chain, thus the letter M. In general, one can have other letters, For example, E_r, implying that the interarrival times have the r-stage Erlangian distribution. G stands for "general" distribution, and so on.
- In field F_2, once again a letter designates the service time distribution. So, for example, M implies that the service time is exponentially distributed. Once again, one might think that it should be E! It is M, because exponential service time is the other component that contributes to the Markovian behavior of the system. In general, one can have other letters; For example, D stands for deterministic (i.e., all service times are the same, etc.).
- Field F_3 is a numeric field, indicating the number of servers available at the service facility.
- Field F_4 is also a numeric field, indicating the "capacity" of the service facility. The term "capacity" is often used in many different ways, so we need to be careful! Field F_4 indicates the maximum number of customers that can be in the service facility, not necessarily being served. So, for example, $M/M/10/15$ represents a system to which arrivals (calls) occur according to a Poisson process, service times are exponentially distributed, there are 10 servers (channels), but there is room for 5 $(15 - 10 = 5)$ arrivals (calls) to be "waiting" for service. In other words, the maximum number of arrivals or users or calls in the system is indeed 15, but only 10 can be in service at any one time.
- Finally, field F_5 is also numeric and it represents the overall population from which arrivals can occur. In our work, and whenever it is much greater than the value in F_4, we just assume it is infinite.
- Also, if any of the numeric fields is missing, it is assumed to be infinite, so in our example above $M/M/10/15$, indeed we assumed that the population base is infinite, indeed a good approximation as long as the population base is much larger than 15.

Note that the $M/M/N/N$ system implies that there is no "waiting room," which is consistent with our assumptions. Indeed, this is an assumption that is often in conflict with human nature. For instance, if I try to make a call and do not get through, I am more likely than not to try again, and that very action violates the assumption at hand. We will return to this point, but for now we'll just say that this assumption, which is somewhat in conflict with human nature, is a weakness of the model and we have to take the model at its face value, else we can not proceed!

3.2. State-Space and State-Transition Diagrams

The mathematical definition of *state space* and the resulting *state-transition diagram* are nontrivial and involve a fair amount of mathematical formalism. The reader is referred to the excellent presentations in Refs. 4 and 5. Here we present a summary of how these tools become useful in analyzing a variety of traffic engineering problems. We accomplish this through the analysis of the $M/M/N/N$ system, which in turn leads us to the development of the so-called Erlang B formula.

As it turns out, for a Markovian system [1,2,4], the number of customers in the system is a state.[5] For non-Markovian systems, the interested reader is referred to the treatise by Akimaru and Kawashinia [6]. The set of states constitutes the state space, as illustrated in Fig. 7. Once all states have been identified, they can be represented in any convenient manner. Traditionally, we simply draw a circle, with the state label inside it, as in Fig. 7a. However, in general, one can think of software tools and therefore other representations of the state space.

The next concept is critical: The *state-transition diagram*. Transitions from one state to another are clearly possible, as indicated on a state-transition diagram (Fig. 7b). The transitions are indicated by an arrow going from or to the respective states. The arrows are labeled

[5] Why this is not necessarily so in general is a difficult and deep question beyond our scope. While not obvious at all, it is the fact that all underlying times (interarrival times and service times) are independent and exponentially distributed that makes this statement possible. For a deeper understanding, the interested reader is strongly encouraged to pursue this in the references.

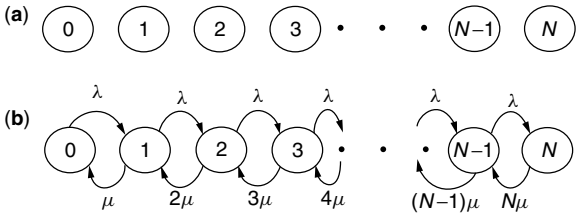


Figure 7. (**a**) State space; (**b**) state-transition diagram for the M/M/N/N system.

with the rate at which a "departure" from a given state to another state occurs. Once this set of states and associated transitions are identified, the entire "thing" is called the state-transition diagram. The state-transition diagram becomes useful in finding the steady-state probability distribution of being in each state. Note that although the state space consists of $N + 1$ states in this example, in general the state space can be finite or infinite. Similarly, drawing the state-transition diagram may be easy or not viable at all, depending on the complexities of the state space and associated transitions.

3.3. Flow Conservation Principle and the Erlang *B* Formula

The fundamental property that makes the state space a useful definition is the notion of flow conservation. As do other conservation laws in physics (conservation of energy, conservation of momentum, etc.), they tend to be rather technical in their proof, but once we accept them, their utility becomes very clear. How many readers have read the proof of conservation of energy? We all learned about it though! Similarly, here we will take the law of conservation of flow as given and work with it. Once again, the interested reader is referred to the references for further reading, specifically the volumes by Kleinrock [4,5].

What is the law of conservation of flow? First, let's define "flow." Referring to Fig. 8, consider a closed surface encompassing state i. Let P_i be the steady-state probability that the system is in state i. Consider a transition from state i to some other state with some rate λ as shown by the arrow (without necessarily indicating which state the transition is into). As it turns out, the natural definition of flow is very easy and quite intuitive: Flow out of state i is the product λP_i.

The flow conservation law, as you might guess, simply states that the flow out of a state is equal to the flow into the state. In fact, it is more general! It states that across any closed surface, containing any set of states, the flow into the closed surface equals the flow out of it. The choice of which states to include in the closed surface is entirely

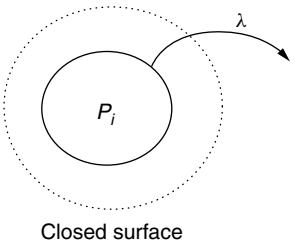


Figure 8. Definition of flow.

up to us! Judicious choices will result in a set of equations that can be solved for the probabilities P_i, i.e., the steady state probability distribution of the various states can be obtained and then applied in a variety of ways.

Once again, the state-transition diagram for the $M/M/N/N$ system is shown in Fig. 7b. Note that transitions from any state $i < N$ to state $(i + 1)$ occur with rate λ, as that is the arrival rate of new calls. Similarly, if the system is in state 1, then the one existing call departs with rate μ, so the transition to state 0 occurs with rate μ. On the other hand, if the system is in state 2, then there are two ongoing calls in the system. Departure of either one of the two, moves the system to state 1. But each call occupies its channel for an independent, exponentially distributed time, with parameter μ. Thus, the departure from state 2 to state 1 occurs with exponential time, with rate $\mu + \mu = 2\mu$. Similarly, if the system is in state $i \leq N$ it departs for state $(i - 1)$ with rate $i\mu$.

The Erlang B formula is essentially one step away! Let us consider the sequence of closed surfaces that enclose

1. State 0
2. States 0 and 1
3. States 0,1, and 2
4. ....
5. States 0,1,2,... $(N - 1)$

Applying the flow conservation principle to these surfaces, one gets the following system of N equations, with the probabilities $P_i (i = 0, 1, 2, \ldots, N)$ as the $(N + 1)$ unknowns:

$$\lambda P_0 = \mu P_1 \tag{1}$$

$$\lambda P_1 = 2\mu P_2 \tag{2}$$

$$\cdots \tag{3}$$

$$\lambda P_{N-1} = N\mu P_N \tag{4}$$

As the reader can see, an easy inductive argument will show that

$$P_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0$$

Finally, the last equation comes from the normalization step, that is, from the fact that all probabilities must sum to 1 which allows us to compute P_0:

$$P_0 = \frac{1}{\sum_{n=0}^{N} \frac{A^n}{n!}}$$

Note that this last normalization step is often one that is quite difficult to perform in more complex systems. Nevertheless, for our problem, we finally get that the probability of being in the blocking state (state N), which is indeed the blocking probability is simply

$$B(A, N) = \frac{\frac{A^N}{N!}}{\sum_{n=0}^{N} \frac{A^n}{n!}}$$

where we have obviously given the term λ/μ the name it deserves. Load A, which as discussed earlier is dimensionless and indeed the ratio of two rates: the arrival rate and the service rate. This is the famous Erlang B formula, a cornerstone of traffic engineering.

3.4. Important Points and Issues

Several points need to be made about the Erlang B formula/result:

1. In Erlang's days, computing power was simply not available. Thus, in the interest of making his result easily applicable and accessible to the engineering community, he developed the *Erlang B tables*, which is a tabulation of the Erlang B formula relating the offered load, the number of available channels, and the resulting blocking probability.
2. For a given number of channels (N) and a desired maximum blocking probability, the table provides the maximum offered load that would result in the said probability of blocking. An illustration of the Erlang B table is presented in Table 1. It should be noted that the granularity of neither N nor $B(A, N)$ is fixed. Some tables provide finer granularity than others, but they are all referred to as "the Erlang B table."
3. While until a few years ago, the Erlang B table was indispensable to a traffic engineer, these days, there are much more efficient numerical techniques that can be implemented in a few lines of code that work much more accurately, for instance, than a table lookup approach. One such approach is the iterative formula is

$$\frac{1}{B(A,N)} = 1 + \frac{N}{A}\frac{1}{B(A,N-1)}$$

It should be noted that there are other approaches, but they are beyond our scope. Notably, the Erlang B formula can be generalized to noninteger values of N, in particular when dimensioning networks. The interested reader is referred to Girard's treatise [7], where, in addition, the dimensioning problem is treated comprehensively.

4. It is clear that the relationship between N, A, and $B(A, N)$, has certain intuitively obvious monotonic behavior e.g., for a fixed number of channels, as the load goes up, the blocking probability goes up. However, the relationship is anything but linear! Interpolations and extrapolations are very dangerous, as the nonlinear behavior of the blocking probability can result in significant errors.
5. What is meant by *capacity*? As is often in engineering problems, this term must be defined in relationship to some performance objective. This is where the commonly made definition of N as capacity is rather inaccurate and possibly misleading. In traffic engineering terms, *capacity* is defined as the maximum offered load that a system can sustain at a given level of performance, that is, at a given blocking probability. For, example, referring to Table 1, we see that a system consisting of five channels and a desired blocking probability of no more than 0.02 (2%) has capacity of 1.66 erlangs. If the number of available channels were to be increased to 10 (i.e., doubled), the capacity would become 5.08 erlangs, about 3.1 times more than with five channels. If one were thinking that the capacity is doubled, the error made would be significant. So one needs to be careful when referring to capacity one must distinguish between the amount of resources (N) and the maximum offered load at a given performance level.[6] Indeed, this nonlinearity is often referred to as *economies of scale*, or *trunking efficiency*, a point to which we return in item 7 below.
6. Unfortunately, increasing the number of channels by a certain factor does not uniformly increase the capacity by the same, or even a constant, factor. The reader should consider what happens if the number of channels is increased from, say, 20 to 40, again a factor of 2, as in the previous example. Similarly, these capacity increases are different, depending on what value we have selected for the blocking probability.
7. It is worth our effort to return to a more general version of Fig. 6, shown in Fig. 9. Recall that as the offered load increases from 0 to infinity, the point (x, y) representing the blocking and utilization of the system traces some path in the $B \times U$ unit square from the point (0,0) to the point (1,1). That

Table 1. The Erlang *B* Table

B(A,N) *N*	1%	2%	5%	10%	30%	50%
5	1.36	1.66	2.22	2.88	5.19	8.44
10	4.46	5.08	6.22	7.51	12.0	18.3
15	8.11	9.01	10.6	12.5	18.9	28.2
20	12.0	13.2	15.2	17.6	25.9	38.2
25	16.1	17.5	20.0	22.8	33.0	48.1
30	20.3	21.9	24.8	28.1	40.0	58.1
35	24.6	26.4	29.7	33.4	47.1	68.1
40	29.0	31.0	34.6	38.8	54.2	78.1
45	33.4	35.6	39.6	44.2	61.3	88.1
50	37.9	40.3	44.5	49.6	68.5	98.1
55	42.4	44.9	49.5	55.0	75.6	108.1
60	46.9	49.6	54.6	60.4	82.7	118.1
65	51.5	54.4	59.6	65.8	89.8	128.1
70	56.1	59.1	64.7	71.3	96.9	138.1
75	60.7	63.9	69.7	76.7	104.1	148.0
80	65.4	68.7	74.8	82.2	111.2	158.0
85	70.0	73.5	79.9	87.7	118.3	168.0
90	74.7	78.3	85.0	93.1	125.5	178.0
95	79.4	83.1	90.1	98.6	132.6	188.0
100	84.1	88.0	95.2	104.1	139.7	198.0

[6] This problem has been quite prevalent in the mobile communications community, where, depending on technology used, and other variables, the number of available channels is increased by a certain factor (e.g., by a factor of 3). It is widely reported, then, that the capacity is increased threefold. As we have seen that is not quite accurate, and as traffic engineers we should not make such oversimplifications.

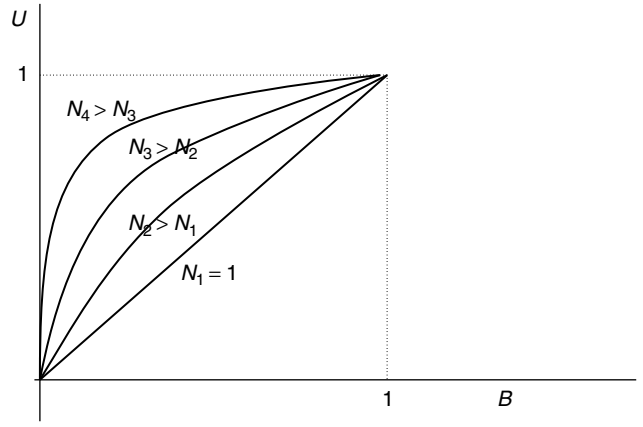


Figure 9. Utilization versus blocking for Poisson arrivals and arbitrary number of channels, N.

"curve" is shown in Fig. 9 for a system in which the number of channels is $N_1 = 1$ (as we did in Fig. 6) to progressively larger values $N_2, N_3, \ldots$. Although not illustrating a precise quantitative relationship, one can see why the larger system is said to be more efficient; for a given level of blocking, the utilization is higher. We ought to note that the curves in Fig. 9, also partition the unit square into two parts: one above and one below the "curve" that joins (0,0) with (1,1) for any given N. Points that lie on that curve represent the situation where arrivals occur according to a Poisson process. As before, the region above that curve represents what would happen if the peakedness is less than 1 and the region below the curve represents the situation where traffic is peaked.

8. Since a fraction of the offered load is blocked, we can now easily introduce the concept of carried load, $C(A, N)$:

$$C(A, N) = A[(1 - B(A, N))]$$

9. Although not necessarily obvious, it is relatively easy to prove that the carrier load is always less than the number of channels N. In fact, it is also easy to show that the ratio $C(A, N)/N$ is a measure of utilization of the system and is indeed a number in the range [0,1). We note that if the offered load A is Poisson, A = 1 erlang and $N = 1$, then $C(1, 1) = \frac{1}{2}$ and the utilization is also $\frac{1}{2}$. Compare this result with what happened when the same 1 erlang of load had peakedness $Z = 0 \ll 1$ and when $Z \gg 1$.
10. Solving the system of equations resulting from the flow conservation equations is not always an easy task. Numeric techniques can be employed, and the interested reader is referred to the treatise by Robertazzi [8], where the problem of a network of such queues is analyzed and numerical techniques are also presented.

3.5. The Erlang *C* Model and Blocked Calls Delayed

While the terminology "blocked calls delayed" and "blocked calls held" is not sufficiently clear and differential, it has been widely adopted by the community at large, so one has to be aware of the differences of the Erlang C model to be discussed here and the Poisson model to be discussed later.

Analysis of the Erlang C model is quite straightforward. Once we realize that the basic assumptions are the same as before (in the Erlang B model), with only one exception, the process is strikingly similar. The exception is that we assume that arrivals that find the system full, simply enter a queue and are served by the service facility on a first-come, first-served basis, else known as *FIFO* (first-in, first-out). We thus have an $M/M/N$ system.[7]

The state-transition diagram is shown in Fig. 10. Obtaining the steady-state probability distribution is quite straightforward as before, but with a couple of important observations:

1. The state space is infinite. This is a direct consequence of assuming an unbounded queueing room capacity.
2. The transition rates are just as before up to state N. Once the system is full, only one of the N customers that are in service can depart, so from that point on, the departure rate (from states k to state $k - 1$, for $k > N$) is always $N\mu$.
3. The fact that the state space is infinite implies that the sum of all probabilities, must converge. As we will see in a moment, the convergence requirement translates into the requirement that the offered load A must satisfy $A < N =$ number of channels in the system.
4. The term "blocking probability" has quite a different meaning in the Erlang B and the Erlang C models. Note that when blocked calls are cleared, they depart the system without being served, so indeed the term "blocking probability" is reflective of what happens. In the Erlang C model, *no* arrival is *blocked* in the sense that *all* arrivals do get served. It is only a question of being served immediately on arrival or having to first enter the queue for some time $T > 0$. Thus when one refers to blocking probability in the Erlang C model, one refers to the probability that an arbitrary arrival has to enter the queue, and thus be subjected to some delay, before being served. By the way, this is also the reason why this model is often referred to as *blocked-calls delayed model*.
5. A natural and important question arises from consideration of paragraph 4. Not only is the "blocking probability" important; probability of the

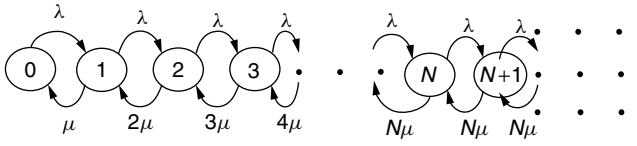


Figure 10. State-transition diagram: Erlang C model.

[7] Recall that the convention is that if a numeric field is absent it is assumed to be infinity. Thus we are making the explicit assumption that the waiting room is infinite.

delay exceeding some threshold may be even more important and critical. Thus we need to quantify the delay distribution, namely, the probability that the waiting time T exceeds some value t, $P(T > t)$. In what follows we show both the probability of finding the system full (blocking probability, or queueing probability) as well as the distribution of the delay. Which of these two expressions is *the* Erlang C formula is somewhat unclear. Some authors refer to one of these expressions — others, to the other one of these expressions — as being the Erlang C formula.

Application of the flow conservation equations reveals that the steady-state probabilities P_k, if they exist, must satisfy

$$P_k = \begin{cases} P_0 \dfrac{A^k}{k!} & \text{if } k \le N \\ P_0 \dfrac{A^k N^{N-k}}{N!} & \text{if } k > N \end{cases}$$

The normalization step requires that the infinite sum of probabilities be summable to 1; in particular, it must converge. As is easy to see, the infinite sum converges if and only if $A < N$. Indeed, from here on we will assume that the load A satisfies $A < N$, that is, that there exists a steady-state solution. While the mathematics forces us directly into making that assumption, it is worthwhile contemplating the practical aspects of this requirement. What would happen if $A > N$? After all, the offered load is whatever it is, and there is no reason why it could not be greater than the number of channels. The answer is strikingly simple — the system never reaches steady state, and thus it becomes unstable. The state-transition diagram of Fig. 10 is helpful in understanding what happens in that case. Recall that we can consider the flows as a "tendency" toward certain states. $A > N$ is equivalent to $\lambda > \mu N$. Note that the arrival rate λ, tends to "push" the system into the higher states. On the other hand, the service rate μ tends to push it into the lower states. In steady state ($\lambda < \mu N$) there is a balance of sorts; the systems drifts up and down and reaches all states, each one with a given probability, precisely the steady state probability of the given state. If, on the other hand, λ is large enough, then μ looses the battle, and the drifting toward the higher states is overwhelming. In such a case, the queue builds up without bound, and every arrival is assured of finding the system full — indeed, the queue is growing without bound as time evolves. The only way to stabilize the system is to reduce the offered load. (or, of course, increase the number of servers N).

Finally, one can show that as long as the system is stable ($A < N$), the complimentary cdf of the waiting time T is given by

$$P(T > t) = \frac{NB(A,N)}{N - A[1 - B(A,N)]} e^{-t\mu(N-A)}$$

where A = offered load
N = number of channels (and recall $N > A$)
μ = service rate (i.e., $1/\mu$ is the average holding time)
$B(A,N)$ = Erlang B blocking for offered load A and system of N channels

It follows trivially that the "Erlang C blocking probability," or as it is sometimes called, the "queueing probability," is given by

$$P_Q(A,N) = P(T > 0) = \frac{NB(A,N)}{N - A[1 - B(A,N)]}$$

Note that it is a simple algebra problem to show that $P_Q(A,N) > B(A,N)$ for all values of A and N. Although this is true, one should be aware of the discussion earlier regarding the direct comparison of these probabilities. Also, depending on the author, either of these two formulas is referred to as the *Erlang C formula*. Since the second one is a special case of the first one, it is the preference of some to refer to the first of these formulas as the Erlang C formula, but the terminology is by no means consistently used.

Note that, technically speaking, if $A < N$ is violated, nothing can be said, since the system is unstable. However, it is often stated that if $A > N$, the blocking probability is 1, thus, in that case, effectively all arrivals enter the queue. In any case, one should also note that as A approaches N, the waiting time grows without bound.

What is the capacity of such a system, operating in accordance with the Erlang C model? Recall that in the Erlang B model, capacity is defined as the maximum offered load that can be sustained at a given blocking probability. Thus, referring to Table 1, a system of 20 channels in which the maximum blocking probability is 2% has capacity of 13.2 erlangs. If the acceptable blocking probability were to be 5%, then the system capacity would be increased to 15.2 erlangs. Capacity is intimately tied to performance. What is the corresponding performance metric in the Erlang C case? Unfortunately, the answer is not as clear-cut. Just as there is some confusion about which formula is *the* Erlang C formula, there is plenty of confusion what performance objectives one should adopt. Some opt to just require that the blocking probability be no more than some value, say, 2%. Others require that some delay threshold be met most of the time; for instance the probability that the delay exceeds $\frac{1}{2}$ normalized units[8] be no more than 2%. As was the case with the Erlang B model, the more relaxed performance requirements we pose, the more capacity the system has. However, if one examines the Erlang C formula a little closer, one will observe that as the delay objective is relaxed from 0 to some positive value, the capacity increases rather sharply. But then, as the delay requirement is relaxed further, the marginal capacity improvement is small. Conversely, one could make the observation that at some point, even small increases of the load, result in large increases in the delay threshold, for a given probability of exceeding that threshold.

[8] Rather than using seconds, minutes, or whatever unit of time one might find convenient in a given situation, the uniformly most convenient unit is one that is normalized to the average service time. Thus a normalized delay of $\frac{1}{2}$ simply refers to the fact that the wait time is $\frac{1}{2}$ of the average service time. Clearly, such a unit is dimensionless.

Table 2. The Erlang *C* Table

Number of Channels	Offered Load	Blocking Probability (%) for Erlang *B*	Blocking Probability (%) for Poisson	Blocking Probability (%) for Erlang *C*
5	2.5	6.97	10.88	13.04
	3.0	11.01	18.47	23.62
	3.5	15.41	27.46	37.78
	4.0	19.91	37.12	55.41
	4.5	24.30	46.79	76.25
	5.0	28.49	55.95	100.00
	5.5	32.41	64.25	100.00
	6.0	36.04	71.49	100.00
10	5.0	1.84	3.18	3.61
	6.0	4.31	8.39	10.13
	7.0	7.87	16.95	22.17
	8.0	12.17	28.34	40.92
	9.0	16.80	41.26	66.87
	10.0	21.46	54.21	100.00
	11.0	25.96	65.95	100.00
	12.0	30.19	75.76	100.00
20	10.0	0.19	0.35	0.37
	12.0	0.98	2.13	2.41
	14.0	3.00	7.65	9.36
	16.0	6.44	18.78	25.61
	18.0	10.92	34.91	55.08
	20.0	15.89	52.97	100.00
	22.0	20.90	69.40	100.00
	24.0	25.71	81.97	100.00

For the sake of an example, we have tabulated a small portion of the "Erlang *C* table" in Table 2. Note that once again, table lookup was a shortcut when the numerical calculations were not readily available. As is easily seen today one could easily make the calculations, even in a basic spreadsheet, so the need for a table is obsolete.

As noted earlier, the Erlang *B* and the Erlang *C* models are diametrically opposite; one assumes that the blocked user waits no time, the other model assume that such a blocked user is willing to wait an unbounded amount of time. Clearly, from a practical perspective, we'd like an "in between" modeling approach. Indeed the Poisson model, examined next, is precisely that.

3.6. The Poisson Model and Blocked Calls Held

This is the $M/M/\infty$ model. It assumes an infinite waiting room, as in the Erlang *C* model, but it assumes that arrivals that have entered the queue can depart unserved. Before we proceed, the reader should be alerted that the term "Poisson model" can be misleading. As Poisson was very prolific, many things have been named after him possibly causing confusion! In all three models, we should remind ourselves, it is assumed that arrivals occur according to a Poisson process. This assumption is no more or no less true in the Poisson model. In many ways, it would have been much simpler if this had been termed the "Erlang *D*" model, but such is not the case! Or maybe the "Erlang *H* model," *H* for *held*, as this is also referred to the situation where blocked calls are "held" until a server becomes available, or they give up and depart unserved. Had we picked such a terminology, it would be clearer that all models make the same fundamental assumptions. They differ only as to how a blocked call is handled. Be that as it may, the acceptable terminology is "Poisson model" and we are not about to try to change that.

However, we have not completely specified the model as yet. What is the statistical nature of the waiting time? In other words, what assumptions does the Poisson model make about the willingness of blocked customers to wait in the queue for service? Indeed, that is a critical component of the Poisson model. Recall that in all models, it is assumed that the service time distribution is the exponential distribution with some parameter μ. The Poisson model assumes that the waiting time has the same exponential distribution as the underlying service time. Intuitively speaking, the Poisson model assumes that if a given arrival were to keep the channel for a certain amount of time t drawn from the exponential distribution with mean $1/\mu$, the amount of time that the customer would be willing to wait in the queue for service would be drawn from the same distribution. Of course, once the service begins, it all "starts fresh" and the holding time is whatever it is.[9]

Since any customer, whether in service or not, is allowed to depart, the departure rates are not quite the same as they were in the Erlang *C* model. They are shown as part of the state-transition diagram shown in Fig. 11.

As before, one can estimate the steady-state probability distribution. From that we obtain the blocking probability,

[9] This is not as strange as one might think. Actually, it is a direct consequence of the memoryless property of the exponential distribution, which itself is somewhat counterintuitive, until one develops a deeper understanding of the exponential distribution. The interested reader is referred to good books on probability theory [e.g., 1,4].

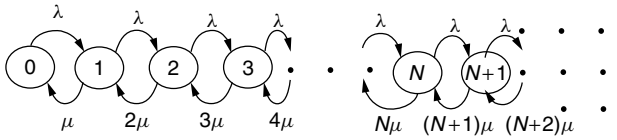


Figure 11. State-transition diagram: Poisson model.

that is, the probability that the system will be found full by an arbitrary arrival and thus it has to enter the queue:

$$P(\text{system full}) = 1 - e^{-A} \sum_{n=0}^{N-1} \frac{A^n}{n!}$$

Note that this is valid, that is, the infinite sum converges, for all values of offered load A. Unlike in the Erlang C case, we do not need to make the assumption of the load being $A < N$. Why is this so? From a mathematical perspective the answer is trivial — the infinite sum of steady-state probabilities converges, regardless of A. From a practical/intuitive perspective, the fact that departures occur with progressively higher rates as the system finds itself in the higher states, creates a stabilization force of sorts. Indeed, if we compare the state-transition diagrams of the Erlang C and the Poisson models (Fig. 10 and 11), we see that their fundamental difference was that the departure rate was the same (μN) for all states k after state N for the Erlang C model, whereas in the Poisson model the rates continue to increase indefinitely; specifically, for sate k, the rate is μk, for all values of k.[10]

As was the case with the Erlang C model, there is a semantic difficulty with respect to the term "blocking probability," but there is an even more pronounced difficulty here. Whereas in both the Erlang B and the C models there are two "classes of citizens," here there are three. Indeed, on arrival, here are the set of possibilities:

- Erlang B model
  - The call is served immediately and departs after its service is completed.
  - The call departs immediately, without being served (blocked calls).
- Erlang C model
  - The call is served immediately and departs after its service is completed.
  - The call first enters a queue, waits as long as necessary and then is served, leaving the system when service is completed (blocked calls).
- Poisson model
  - The call is served immediately and departs after its service is completed.
  - The call first enters a queue, waits for some time and departs unserved.
  - The call enters a queue, waits for some time and a server becomes available, at which time it starts its service. It departs on completion of its service.

[10] One might also note that there were no such considerations in the Erlang B model, as the state space there is finite, so there are no convergence issues.

Although it is clear that unserved calls should be counted as "blocked calls," what about those that had to wait but eventually were served? There are rational arguments that one can make for counting those calls with either the "blocked" ones or the complement, that is, the nonblocked ones. Which of these arguments may be more persuasive is irrelevant! The Poisson model allows us to (relatively easily) find the probability that the system is full and anything above and beyond that can get rather difficult. As such, the blocking probability refers to the probability that an arbitrary arrival finds the system full, thus has to enter the queue. What fraction of these "blocked" calls eventually is served is not part of the modeling question under consideration! Thus one ought to be careful, when interpreting data, that the statement that the blocking probability is 5% in fact implies that 95% of the calls were served immediately on their arrival and some fraction of the 5% were also served, but we do not know how big of small that fraction may be.

Although the term "blocking probability" has a different interpretation in each of the three models, we often find it convenient to compare the results. See Table 2 for such a comparison. Indeed, it is easy to see that the intuitive placement of the Poisson model as "in between"; the Erlang B Erlang C differentiation makes sense. From a practical perspective, we can make an even stronger statement. In practice, it is very unlikely that either extreme occurs. Namely, it is not likely that:

- Arrivals wait for no time at all. In fact, some systems provide for explicit queues, but even if no such queueing capability is explicitly provided for, customers tend to "redial," or try again in some way, thus creating somewhat of a virtual queue.[11]
- Arrivals that find the system full, "wait forever," that is, an unbounded amount of time.

Indeed, assuming some waiting time is a reasonable modeling assumption. Assuming whether the specific distribution of the Poisson model is the right one or not is almost irrelevant. If for no other reason, it is rarely the case that we have a high confidence in our assumptions as to customer behavior. Thus it is unlikely that our model will be an accurate one to start with.

Given the points outlined above, what is one to do? Very simply, if one looks at these three models as simply three data points in a continuum of possibilities, one can get a lot of insight into the problem. Indeed, in an intuitive sense, system performance is a continuous function of the waiting-time distribution. Thus by making certain assumptions about the waiting time distribution, one can check the predicted performance versus what one

[11] It is important to note that unlike in Erlang C, this virtual queue is not served in a FIFO manner. The most appropriate model for this type of situation is *random service order* (RSO); the Erlang C delay distribution needs to be modified to account for this. However, as it turns out, the probability of the system being full is the same, so we can use the Erlang C results for the blocking probability, even on this case where the service discipline is not FIFO.

is measuring in a real situation and use that information in a feedback loop to better calibrate the model of the problem at hand.

Thus it is often unreasonable to ask "Which model is the best?" More often than not, it is the case that we need to understand all these models and see which one best describes the underlying dynamics of the problem at hand.

4. OTHER RELATED TOPICS

A number of problems and applications emerge from the fundamental concepts presented thus far. They are far reaching and their fair treatment is not possible in this article. However, we will briefly discuss some of them.

4.1. Random-Access Protocols

In "conventional" traffic engineering, it is assumed that the need for a channel is made known to the system by some other means. For example, in traditional telephony, when one picks up the telephone receiver, a dial tone is allocated, which in effect activates the existing line between the telephone and the central office. The traffic engineering problem in its core concentrates on whether a line between your central office and that of the called party exists. The lines to/from one's telephone are "always" there, waiting to be activated by the dial tone or by the incoming call. The need for a channel between points A and B is known to the system by the dialed digits. In many other applications, however, the need for a channel is not known by any such external mechanism. A good example is found in the exploding area of mobile communications, even though the problem had been identified much earlier within the domain of computer networks. The terms used might be different but the underlying concepts are identical. For instance, in Fig. 12, a "central entity," labeled node A, is "in charge" of the system. It may be a base station in a mobile communications network or a host computer in a data network. It allocates idle channels to entities that request one. These entities may be mobile or peripheral devices, such as terminals, printers, or card readers — this does go back to the days of punchcard readers! The problem is very simple to state: how does the base station (node A) become aware of the need for a channel by the mobile units?[12] Note that any "central control" is precluded as they are impractical and are made possible only with unacceptably high overhead. Thus we are forced into what are called *random-access protocols*. Each mobile transmits whenever it feels like! As a moment's reflection makes clear, such transmissions are subject to collisions, as they must use the same mechanism for accessing the base station. These protocols were first analyzed in 1970 by N. Abramson of the University of Hawaii, and the simplest of a family of such protocols is called ALOHA, appropriately enough. The interested reader is referred to Rom and Sidi's study [9] for more detailed analysis and discussion of these problems. Here we present some of the most basic ideas.

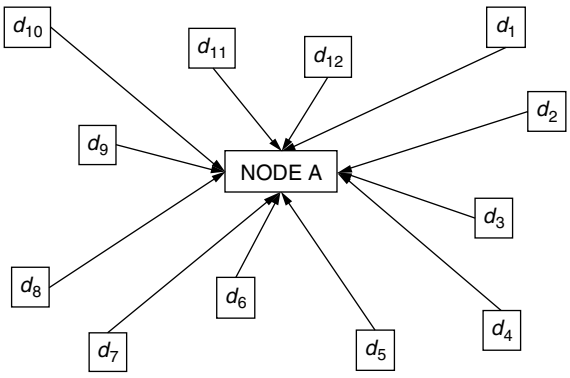


Figure 12. Abstract representation of a base station in a mobile communications network.

In the ALOHA protocol, each mobile would, quite literally, transmit whenever it feels like transmitting. The base station is assumed to receive the information successfully if and only if there is only one ongoing transmission. Any overlap of two (or more) transmissions renders all of them worthless, as even partially received data are ultimately discarded as being corrupted. The fundamental question is then *not* how many channels are available, but what is the throughput of such a system. Throughput is defined as the fraction of time that the channel is used successfully. Idle time and time spent in a colliding state count against throughput. As one can readily conjecture, the throughput of such systems is not great. If the offered load is low, most time is spent in the idle mode. As traffic increases, more collisions occur, thus more of the time is spent in the colliding state, leaving a small fraction for success! Thus the graph of Fig. 13, showing the throughput as a function of offered load, makes intuitive sense. We will return to this graph momentarily.

However, quantifying this intuitively obvious result is not trivial. We will only make some key observations here. The concept of "load" must be defined differently for this class of traffic-engineering-related problems. It should be obvious that holding time is irrelevant. What matters is the number of attempts being made per unit time. It has become common practice in these problems to normalize

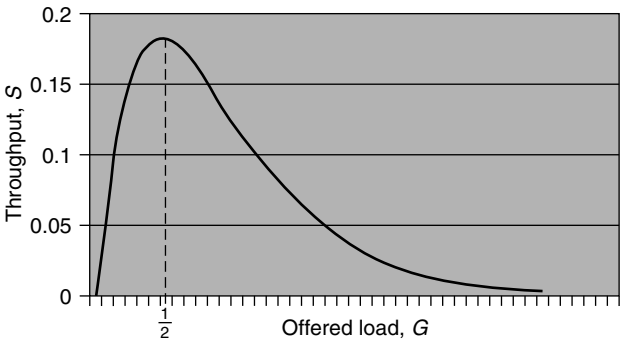


Figure 13. Throughput, S, versus offered load, G, in the ALOHA system.

[12] We will word the rest of this discussion within the framework of a hypothetical mobile communications network, as this simplifies the discussion and reflects the growing area of mobile communications.

the unit time to the average length of the transmission. Owing to the birthplace of these problems in packet switching, this is often called "packet time." Thus one talks about arrivals per packet time. Further, since the transmission is often "a packet" the arrival rate is often referred to as G packets per packet time, or G arrivals per packet time. Clearly, if the world were perfect and ultimate control and coordination were possible, the maximum value of G would be $G = 1$. In such a hypothetical case, the coordinating process would ensure that there were no collisions and that time was constantly utilized for the successful transmission of arriving packets, resulting in a throughput, S, having the ideal value of $S = 1$. Clearly this is not the case, as we realistically have to account for collisions.

Whatever the arrival rate is, collisions will occur. Furthermore, it is common to assume that arrivals form a Poisson process with rate G arrivals per packet time. What happens to colliding arrivals? Once again, the usual models assume that such arrivals are retransmitted at some future instance in time. Note that this creates an instability in the system. Colliding attempts appear at a later time, increasing the probability of future collisions, thus the need for additional transmissions, thus higher probability of collisions, and so on. This is a classic example of positive feedback, thus unstable behavior. If a system reaches such a state, then from a practical perspective it has simply collapsed. Naturally, such catastrophic events must be prevented from occurring. We will return to this subject momentarily.

Under a number of technical conditions, one can show that the throughput is expressed as a function of the offered load by

$$S = Ge^{-2G}$$

This expression is graphed in Fig. 13. However, this simple relationship can be misleading, so we need to clarify a few points:

1. As one can see by elementary calculus, it appears that the maximum throughput is $1/(2e)$, or about 18% and it occurs when $G = 0.5$ arrivals/packet time.
2. However, as G approaches 0.5 from below, the delay, D, associated with retransmissions that have been necessitated approaches infinity. The exact relationship of delay to the offered load G is too complex, requires a number of assumptions and is beyond the scope here. What is important to recognize is that regardless of the specific details, the delay does tend to infinity. Thus from a realistic perspective, one needs to back away to values of G below 0.5.
3. The region of the graph for which $G > 0.5$ is somewhat misleading. If G exceeds 0.5, the system is unstable. If G were to be larger than 0.5, it would drift to infinity as a result of retransmissions, the collisions would dominate, and the throughput would tend to 0, that is, the system would collapse. As a result, authors often refer to the region of $G > 0.5$ as being an "unstable region."
4. While this is correct, it often is interpreted as meaning that the region $G < 0.5$ as being "stable." Unfortunately, this is not the case, in the strict sense of the word "stable." Once again, under a broad set of technical assumptions, sooner or later any system will go unstable, for any value of $G > 0$.[13] However, the amount of time it will take the system to "exit" this so-called "stable region" and enter the doomed instabilities increases rapidly with the distance $0.5 - G$ [for G in the range (0,0.5)], that is, as we back away from $G = 0.5$.[14]
5. One is then entitled to pose the reasonable question of what is the highest value that G should be allowed to be. While this question has a number of interesting theoretical components [e.g., 9], no firm practical answers are agreed to by the practitioners of the trade! Apparently, it is more of a question of how much risk and what kind of safety valves have been put in place than anything else. Conservative designers want to keep G below about 0.1. More risk taking (with appropriate safety mechanisms in place, one would assume) would allow one to be at about $G = 0.2$. However, these values are debatable, subject to a number of conditions, thus should be treated as a rule of thumb rather than precise guidelines! Finally, one sees that at $G = 0.2$, $S = 0.13$, thus one often also talks about the throughout S as being no more than some value in the range of 0.1–0.2.

In conclusion, we see that the throughout of such channels operating in the ALOHA mode is quite small. One way to improve it is slotted ALOHA. In such systems, time is organized into "slots," as shown in Fig. 14. Any mobile station that at time $t = t_0$ decides that it wants to transmit, must wait until the beginning of the next slot. Clearly, collisions are still possible, though less likely, as the "vulnerability period" is cut in half, from two packet transmission times to one packet transmission time. As a result, the offered load–throughput relationship is changed to

$$S = Ge^{-G}$$

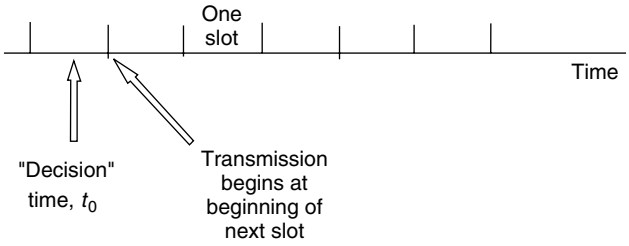


Figure 14. "Slotting" of time for Slotted ALOHA.

[13] Mathematically, the point $G = 0$ is stable, but from a practical perspective this is a useless piece of information. It simply states that a system which has no traffic is stable!

[14] Under certain conditions, this time increases exponentially with the amount by which G is less than 0.5.

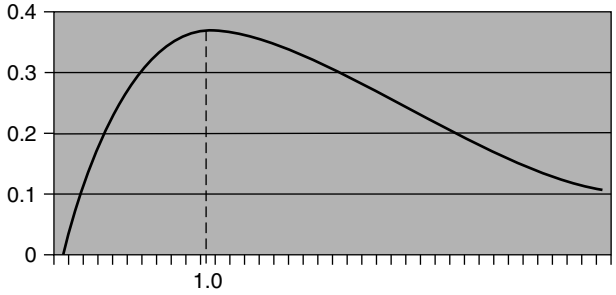


Figure 15. Throughput, S, versus offered load, G, in the Slotted ALOHA system.

This is plotted in Fig. 15. As we can see, this is very reminiscent of the behavior of the ALOHA system, except that the maximum occurs when $G = 1$ and it is $1/e$ or about 36%. While this is double of what it is in the ALOHA case, the entire discussion of the ALOHA system issues, relating to stability, delays, and other factors carries over here as well. Once again, there is no uniform agreement on what acceptable values of G (and thus of S) are, but it is certain that they are higher than in the ALOHA case. This increase in utilization comes at the expense of needing to maintain synchronization information. Whereas in some systems synchronization is necessary because of a number of other independent issues (and thus the incremental cost is virtually zero), in other systems, this additional requirement of synchronization may be prohibitive and the increased throughput may not justify the additional complexity and cost. Indeed, these are some of the problems that both theoretical and applied research addresses in specific situations.

4.2. Network Design

The concepts presented here form the basis for analyzing more realistic systems, such as large networks, consisting of nodes and links joining these nodes. These links have varying capacities, and calls can be routed trough such networks according to a number of algorithms. While space and scope limitations do not allow us to diverge into these topics, the interested reader will find the topic of network design discussed in a number of references [e.g., 3,6,7,10]. In particular, Ash has discussed the problems related to routing algorithms in circuit-switched networks in detail [3]. Similarly, Bertsekas and Gallager have discussed the problem of routing and capacity assignments in packet-switched networks [10].

In a network environment, where routing decisions are influenced by the state of the system, what might originally be Poisson traffic can be offered to various resources as non-Poisson traffic. The traffic engineering problem then increases in complexity and a number of approximate techniques can be used. Both from the theoretical and the practical perspective, these problems become quite complex and there are excellent treatments of these problems in the literature [3–5,7,8].

5. CONCLUSION

In summary, we note that traffic engineering involves a fair amount of mathematical detail that, if not accounted for, may lead to significant errors in the analysis and design of networks. Allocation of resources thus becomes a multidisciplinary problem involving many aspects of engineering, economics, and social behavior, to mention only a few.

Although the problem is quite easy to state in its most elementary form, the solutions to the real problems in a networked environment become quite more complex. Furthermore, the random-access problem, albeit originally a traffic engineering problem, presents us with quite a different set of challenges and ways of looking at a problem.

BIOGRAPHY

Apostolos K. Kakaes received his B.S. and M.S. degrees in applied mathematics with a minor in electrical engineering in 1978 and 1980, respectively, from the University of Colorado. He received a Ph.D. degree in EE from the Polytechnic University in 1987. He had joined AT&T Bell Laboratories in 1980, where he stayed until 1987 working on traffic engineering and network design in both circuit and packet switched networks. In 1987, he joined the faculty of the Department of Electrical Engineering and Computer Science of the George Washington University where his interests in both the physical layer issues and networking problems lead him to work on wireless networks. In 1996, he founded and has since been running Cosmos Communications Consulting, an independent consulting company specializing in mobile communications.

He has published, conducted IEEE tutorials, and lectured extensively around the world in all aspects of both fixed and mobile communications, including traffic engineering issues in current as well as emerging high-speed fixed and mobile communications networks.

BIBLIOGRAPHY

1. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, Wiley, New York, 1970.
2. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley, New York, 1971.
3. G. R. Ash, *Dynamic Routing in Telecommunications Networks*, McGraw-Hill, New York, 1997.
4. L. Kleinrock, *Queueing Systems*, Vol. 1, *Theory*, Wiley, New York, 1975.
5. L. Kleinrock, *Queueing Systems*, Vol. 2, Computer applications, Wiley, New York, 1976.
6. H. Akimaru and K. Kawashima, *Teletraffic Theory and Applications*, Springer-Verlag, 1992.
7. A. Girard, *Routing and Dimensioning in Circuit-Switched Networks*, Addison-Wesley, Reading, MA, 1990.
8. T. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*, Springer-Verlag, 1990.
9. R. Rom and M. Sidi, *Multiple Access Protocols; Performance and Analysis*, Springer-Verlag, 1989.
10. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

COMMUNICATIONS FOR INTELLIGENT TRANSPORTATION SYSTEMS

ORESTE ANDRISANO
GIANNI PASOLINI
ROBERTO VERDONE
University of Bologna
DEIS, Italy

MASAO NAKAGAWA
Keio University
Japan

1. INTRODUCTION

Intelligent transportation systems (ITSs) have been investigated for many years in Europe, North America, and Japan, with the aim to provide new technologies capable of improving the safety and efficiency of road transport [1–5]. In this context several technologies and applications have been investigated and demonstrated with reference to onboard communication networks, and short- and long-range communication systems [6].

Going back to the evolution of guidance support, the first efforts date back several decades (from the time of writing) and were essentially aimed at diffusing traffic information by means of road signs and FM broadcasting. An amazingly significant step in the direction of improving transport safety, featured the adoption of a millimeter-wave radar sensor installed in the front grille of vehicles in order to promptly detect obstacles and to warn drivers when they reach an unsafe distance. This technology has been improved and adopted by several car manufacturers that integrated the radar with an onboard cruise control [7]. This solution offers the possibility, for instance, to automatically maintain an optimal following distance behind cars traveling ahead, thus reducing the risk of collisions as well as the burden on the driver during long trips.

However, as soon as the first ITS systems were implemented, it was clear that their effectiveness would be greatly improved by the possibility to establish a bidirectional communication link between the different entities (cars, pedestrians, ITS service providers, etc.) acting in the road transport scenario. In this regard, the communication-based systems can play a fundamental role in the context of ITS, since they can overcome the main limitations of self-sufficient systems (those relying on passive sensors, radars, videocameras, etc.) that are solely based on the unilateral perception of the environment surrounding the vehicle. To provide advanced ITS services, localization systems will also be suitably exploited in conjunction with mobile communications.

Different technologies were investigated in Europe at the beginning of the ITS research activity, during the 1980s: infrared communications, millimeter-wave communications, and mobile radio. The first two appeared to be the most interesting at that time, and were oriented to the implementation of short-range mobile communication networks, whereas since the early 1990s, some projects have developed the concept of using cellular radio for providing ITS services [e.g., 8]; it is to be pointed out that, at that time, only second-generation (2G) systems were available, such as the Global System for Mobile communications (GSM), characterized by a limited ability to provide advanced data services with high quality levels and flexibility.

In the near future third-generation (3G) systems should be available, enabling advanced data services to be provided to mobile users in the 2000-MHz frequency band, in both Japan, where they have been under development, and Europe, where their introduction is planned for the year 2003 [9]; these systems will be able to provide different bit rates [$\leq$2 Mbps (megabits per second)], different levels of quality of service (QoS), and much more flexibility than that offered by 2G networks.

However, the evolution from 2G to 3G systems will not be sharp, and is based on the implementation of some 2.5G solutions, such as the GPRS (General Packet Radio Service) in Europe, providing intermediate bit rates (up to $\sim$100 kbps) in the GSM band (900 and 1800 MHz), based on packet-oriented techniques and flexible operating modes. Therefore, GPRS is thought to be a suitable candidate to offer services to the ITS context, too.

The increasing interest on wireless personal-area networks (WPANs) and the consequent penetration of low-cost portable devices equipped with the emerging Bluetooth technology [10], has suggested the possibility of adopting WPAN devices for the provision of short-range ITS services as well.

In this article, following a brief introduction to the typical ITS services, we first show some of the results of the research carried out in Italy and Japan concerning the role of communications in ITS; then we discuss the suitability of 2.5G and 3G systems and Bluetooth for ITS applications.

2. ITS SERVICES

The ITS services based on communication devices rely, essentially, on three kinds of techniques: *roadside-to-vehicle communication* (RVC), based on the use of an infrastructure network covering the service area, intervehicle communication (IVC), establishing a direct communication link among automobiles, and target-to-vehicle communication (TVC), based on proper active devices mounted on vehicles with the aim of detecting the presence of the unprotected road user (the target, hereafter) through suitable communication interfaces. Hereafter we provide a synthetic overview of the services that can be provided by the systems described above.

Although sometimes at an early stage, RVC based services have already been introduced in Europe, Japan, and the United States; we can mention, for example, the following applications:

- *Automatic tolling*, probably nowadays the communication-based ITS service most widely used by road users; it proved to be amazingly effective in reducing traffic congestion and improving driver convenience by cashless payment. It is based on RVC techniques.
- The *services for guidance support*, providing the user with information related to traffic, weather

conditions, and so on; they are still at a very early stage of development, and in many cases are based on traditional techniques such as road signs or broadcasting through FM stations (RVC techniques).

- The *services for traffic control*, based on the joint concepts of navigation and communication (RVC techniques).

Still based on RVC systems and not yet introduced, the *services for driving safety* should play a very important role in the future, providing the driver with real-time information concerning possible emergency situations such as those due to fog (banks), traffic accidents, or road hazards. These services are the most difficult to provide because of the stringent requirements of the application; this is the case, for instance, of the "emergency warning" service, which should be based on the possibility to inform all the vehicles in the vicinity of a dangerous situation within a short amount of time from its occurrence, and this requires a prompt system response. Consequently the introduction of such services is still seen as a long-term goal; nevertheless, we show the suitability of GPRS for this application, even if with limitations on the system response promptness.

A significant step in the evolution of ITS should be the introduction of the *cooperative driving* service, which belongs to the class of *services for driving safety* and is based on IVC systems. It consists in forming groups (platoons) of vehicles exchanging data on their status (speed, acceleration, position, etc.) in order to improve the safety and efficiency of the vehicles flow by keeping the intervehicle distances under control in all situations (sudden change of speed, position, etc.). The data exchanged among the cooperating vehicles can be processed by automatic agents and used to control the onboard actuators (brakes, etc.), or presented to the driver through suitable (e.g., vocal) interfaces. Each vehicle of the platoon must be equipped with a communication device, and direct intervehicle communication (IVC) is the only way to provide the requested promptness.

TVC-based services are finally a challenging issue for researchers since the communication link involves pedestrians, cyclists, and others whose communication devices have to be lightweight, small, low-power-consuming, and, possibly, cheap; by means of TVC systems it will be possible to prevent car accidents involving pedestrians and cyclists by means of onboard warning devices. In Europe PROTECTOR, an ITS project, developed research in this area.

3. WIRELESS STANDARDS FOR ITS: DEDICATED SOLUTIONS

3.1. Past Activities

Since the early 1980s remarkable research activity has been carried out in the United States within the context of ITS. One of the most important actors in this scenario is the Californian PATH [11], a joint venture of the University of California, the California Department of Transportation (CALTRANS), and private industry, established in 1986 to develop more efficient transit and highway systems. The goal of PATH is to increase the capacity of the busiest highways and to decrease traffic congestion, air pollution, accident rates, and fuel consumption and to perform field operational tests.

PATH participation in U.S. DoT (Federal Department of Transportation) ITS programs includes several projects within the Intelligent Vehicle Initiative (IVI) program [12], developing research in the fields of

- Sensor-friendly vehicle and roadway systems
- Forward-collision warning systems
- Rear-collision warning systems
- Automotive collision avoidance systems (ACASs)

In Europe and Japan some major research programs (e.g. DRIVE, PROMETHEUS, VASCO) in Europe, AHS and JSK programs in Japan [13–23] have been involved in the definition and testing of telecommunication systems for the field of road transportation, such as systems for automatic tolling, fleet management, traffic control, and cooperative driving; most of them were based on dedicated solutions, explicitly designed for ITS applications. Suitable frequency allocations have been proposed in order to avoid more congested bands. On the other hand, in some cases this represents an obstacle to the introduction of these systems, due to the initial high costs of dedicated solutions when entering the mass market.

For both IVC and the RVC systems, the 60–64-GHz band was chosen among the possible candidates within the DRIVE and PROMETHEUS projects in Europe for reasons related to the size of RF devices, the amount of available bandwidth, and the advantages provided by oxygen absorption, which causes spatial filtering and frequent channel reuse, suitable for short-range systems such as those based on IVC.

The results of all these programs were based on dedicated radio interfaces; we leave the interested reader to refer to the literature [13–23]. In the following we simply provide, as a synthetic view, an example of the results obtained within the TELCO project (1995–1997), which was funded in Italy by Consiglio Nazionale delle Ricerche, after the conclusion of PROMETHEUS with the aim of investigating on the possible use of the millimeter-wave band, and we mention some of the experimental test beds realized in Japan.

3.1.1. RVC at 60–64 GHz: Research in Italy. The 60–64-GHz band is characterized by a peak of oxygen absorption. To highlight the main features of the use of this band, let's consider a simple monodimensional scenario, in urban context, with beacons serving the area and separated by a distance $2R$ [18]. The multiple-access method employed is TDMA. Let's denote by N and Br the channel reuse factor and the overall bit rate, respectively. Figure 1 shows the outage probability, for a given transmission system, as a function of the cell radius with the reuse factor and the bit rate as parameters. The *outage probability* is defined as the probability that the signal-to-noise and the signal-to-interference ratios are

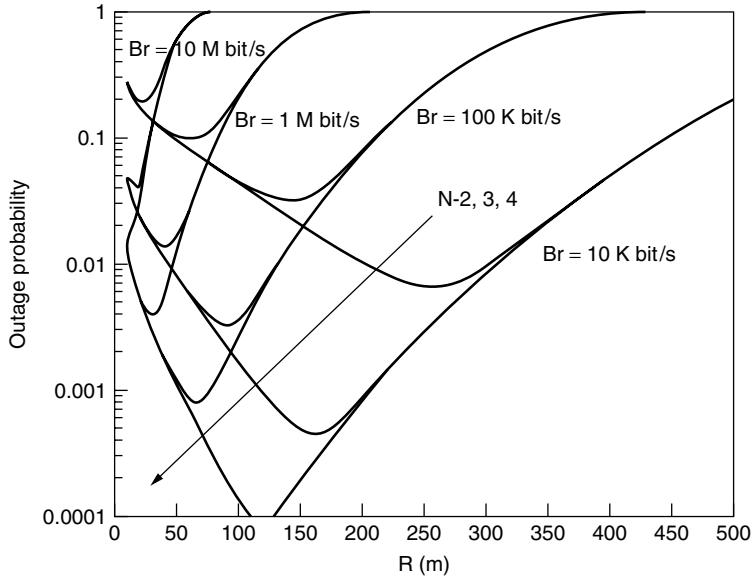


Figure 1. The outage probability as a function of the cell radius, with the reuse factor and the bit rate as parameters (© IEEE, 2000).

below specified thresholds that are defined as functions of the transmission techniques implemented. A noncoherent MSK unprotected system is considered here, as a simple reference, and omnidirectional antennas are assumed. The figure shows that a minimum outage probability is obtained for a given value of R, as a compromise between the effects of noise and interference. This is typical of the 60–64-GHz band, due to the effects of oxygen absorption: no optimum value of R would be found in a frequency band unaffected by the additional attenuation due to oxygen, or rain. As can be expected the optimum outage probability decreases when N increases or Br decreases. It is also interesting to note that, with the system parameters chosen, the values of R are $\sim$100–250 ms, thus giving a large number of beacons per kilometer and suggesting the implementation of these systems in urban areas where the number of users can be large. Suitable scaling of the bit rates and of the optimum cell size can be obtained by employing more sophisticated transmission techniques, directional antennas, or other means. In any case, the typical cell size remains below 1 km.

3.1.2. IVC at 60–64 GHz: Research in Italy. The advantages of using the 60-GHz band for IVC have been shown in several studies [13,15,17,19,20], and the results of these studies provided the design of suitable multiple-access techniques for cooperative driving applications. We let the interested reader refer to these papers.

3.1.3. IVC Experiments in Japan. In Japan, JSK carried out some IVC experiments based on infrared technology in 1996, and some new experiments were performed during the year 2000 at 5.8 GHz. The AHS project is now extending its research studies from RVC to road sensing; the RVC frequency should be more than 5 GHz due to the lack of frequency bands at lower frequencies. Road sensing uses infrared, 60-, and 90-GHz technologies.

Infrared technology was initially selected among all candidates for IVC, for economical and practical reasons, in spite of its weather-dependent characteristics. JSK carried out an experiment in 1996 to verify the possibility of networking between traveling vehicles, as a preliminary testbed in order to plan a new experiment based on microwave technology [16].

Figure 2 shows the experimental vehicles used, equipped with two infrared radiators separated by a 1-m distance on the roofs, a photodetector, a videocamera, and a processor. The distance between forward and backward vehicles was measured by the videocamera, which could record the images of the two radiators on its CCD film. These radiators not only enabled distance measurement but also sent information to the vehicle behind. The test demonstrated the feasibility of IVC, while pointing out that some features, such as network aspects, deserve thorough consideration [16].

3.1.4. CDMA or TDMA for IVC: Research in Japan. The controversy about CDMA and TDMA in cellular communications in the early 1990s concluded CDMA to be preferred to TDMA for channel capacity reasons and its ability to counteract interference. What about both access methods in IVC? Few research works about this matter have been published. Michael and Nakagawa [21] show that analysis of the interference from oncoming vehicles reveals CDMA to be better than TDMA. Since the main information from each vehicle is related to control data of the vehicle, it is characterized by regular occurrence. Consequently each vehicle should transmit periodic data bursts to other vehicles. The interval of the data bursts should be the same for all vehicles. However, if the data bursts of an

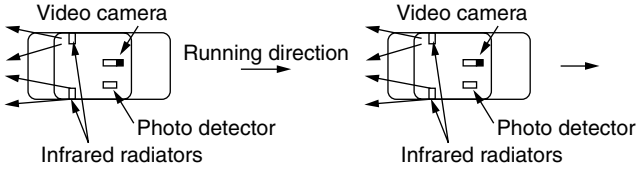


Figure 2. Top view of vehicles in the infrared experiment performed by JSK in Japan in 1996 (© IEEE, 2000).

oncoming vehicle group collide with those of the ongoing vehicle group, the collision is kept for a while and a large amount of data are lost if a TDMA scheme is used. CDMA can save the data from this type of collision; this is the main reason for the selection of CDMA for applications based on IVC. On the other hand, if different carriers were used for communications in the two directions, the problem discussed above would vanish, but this would require a centralized allocation of carriers to the vehicles.

3.1.5. Spread-Spectrum Radar and FM-CW Radar: Research in Japan. FM-CW radar was standardized for a radar system to be mounted on vehicles to detect the presence of other vehicles. When there are a number of vehicles on a road, interference between the radar signals from all the vehicles should be considered. Shiraki et al. [22] show simulation results for the FM-CW radar and spread-spectrum (SS) radar. The SS radar shows much better performance than FM-CW radar as reported in Fig. 3, in terms of signal-to-interference ratio (SIR).

This result also suggests the investigation of the possible integration of radar and communication systems, based on a common frequency band (e.g., the millimeter-wave band) and processing (e.g., SS) techniques.

3.2. Current Activities

A brief overview of the current ITS applications and testbeds is presented in order to explain to the reader which is the state of the art in this field [7].

3.2.1. Services for Guidance Support. The next generation of communication technology for guidance support is being developed behind the steering wheel and, more recently, on the wireless handset. The emerging trend in this field is to offer location-based voice and data communication tools that provide "smart" information, tailored to where the customers are and to what they are doing,

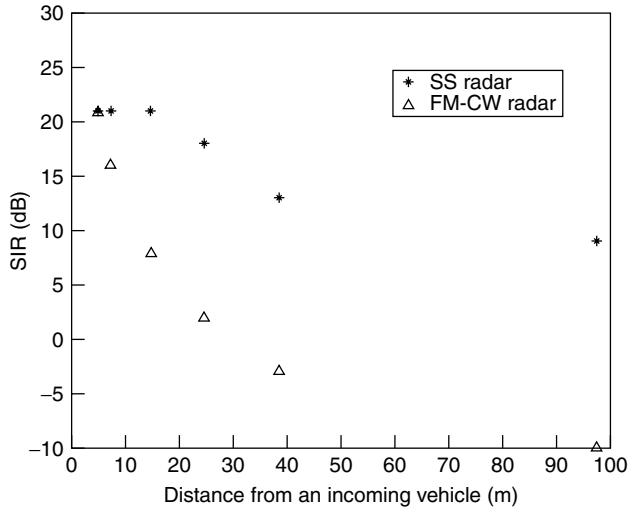


Figure 3. The ratio between the useful received power from the target and the radar interference power received from an oncoming vehicle at the same distance (© IEEE, 2000).

providing enhanced security, navigation, and convenience to mobile consumers.

The adoption of wireless terminals in this currently deeply investigated context is proving to be extremely effective, greatly improving the usefulness of ITS services; in the greater metropolitan Paris area, on the Boulevard périphérique and the Paris city roads, for instance, the time needed to travel the distance between two points is displayed by variable message signs (VMSs). Thanks to portable terminals, this information is available in the car (the terminal can be easily installed) and to people who are not on the road (they can take the terminal along). In this manner route planning is improved by conveying information on travel times before the journey commences.

3.2.2. Automatic Tolling System. The well-known system for Automatic Tolling of the Italian Autostrade S.p.A., namely, Telepass, is the first system in the world for toll collection where drivers do not need to stop at toll stations [7].

3.2.3. Services for Traffic Control. Congestion is steadily worsening on the streets of Europe's cities as more people choose to travel by car. Buses, trams and trains take less space, thus increasing their share of trips can help reduce congestion. Unfortunately, many public transport services get caught in traffic and are reputed for poor reliability. Public transport performance can be improved by better controlling and managing traffic generally. The City of Turin, as an example, has implemented a system — named 5T — that integrates nine ITS subsystems under the co-ordination of a tenth system, the "traffic and transport supervisor," which monitors and controls all the other subsystems. Among the 5T subsystems, public transport vehicle location and urban traffic control subsystems provide priority to public transport vehicles at traffic signals, particularly to those that are running late. Similar priority can be given to emergency service vehicles [7].

3.2.4. Services for Driving Safety. In the case of an accident on the motorway, significant danger is caused by vehicles that block the road (cars involved in the accident, emergency services, a tailback or bottleneck of cars that cannot pass the location). Secondary accidents, which are often more serious than the original one, can be avoided if the accident is detected immediately and the scene of the accident is cordoned off right away. In Germany, an advanced warning system that automatically detects incidents and provides immediate warning to drivers has been proposed. Beacons equipped with a light unit and additional electronics are installed along the emergency lanes.

The beacons are interconnected by cable and linked to a traffic computer center, where an operator views the accident on a computer screen, via an appropriate monitoring interface. The light units can be activated from the control centre, to flash warnings until the police arrive to cordon off the area. The light flashes, initiated within seconds of the accident, warn approaching traffic of problems downstream, and invite the drivers to slow down.

3.2.5. Dedicated Short-Range Communications. Dedicated short-range communications (DSRC) systems aim at providing one-way or two-way high-speed radio links between a fixed roadside unit and onboard equipment: ITS-related information is exchanged within the communication area, made up of the roadside antenna, therefore configuring a RVC system. DSRC systems have been deeply studied in the United States, Europe, Korea, and Japan, and several proposals have been presented so far [24–27]. The various solutions adopt different modulation formats as well as different data rates, ranging from hundreds of kilobits per second to a few megabits per second, and operate in different bands within the microwave region.

4. WIRELESS STANDARDS AND ITS

The abovementioned considerations on the difficult penetration of ITS dedicated systems in the mass market suggested that researchers investigate the possibility of providing ITS services by means of already standardized systems such as the 2.5G [28] and 3G cellular systems as well as Bluetooth [29]. The capability of these standards to offer packet-switched services will be studied in order to show how they can be exploited for the provision of services for driving safety, such as "emergency warning," and for traffic control and guidance support as well.

4.1. A GPRS-Based System for Emergency Warning, Traffic Control, and Guidance Support

GPRS supports packet switched services, based on a radio interface almost identical to that of GSM, at different rates, roughly 9–100 kbps, which can be obtained by assigning multiple GSM slots per frame to the same user. On the other hand, if the user bit rate is very low, GPRS can allocate the radio resource depending on real needs, thus preventing the inefficient use of the radio spectrum and allowing the user to pay for the amount of data actually transferred.

Let's consider a typical European highway trunk, consisting of three plus three lanes, as a reference scenario to describe the system investigated (see Fig. 4). If we assume that the whole trunk is covered by the public GPRS network, we can exploit its capability to support packet-switched services and priority-based scheduling [30].

To provide a prompt "emergency warning" (EW) service, in fact, each vehicle has to establish a connection with the network at the entrance of the trunk, in order to reserve some radio resource units to be used when needed (e.g., if a dangerous situation is monitored by the onboard computer, which also controls airbag deployment, sudden braking, etc); this event is considered to be a very infrequent but significant situation. The connection is relinquished at the output of the trunk. The packet transmission mode controlled by the network should also rely on a suitable scheduling procedure, taking the priority of services into account. In this manner, once the different vehicles have allocated their radio resource units, the system could also support other services characterized by lower priority.

Therefore, we assume that each vehicle has allocated a minimum radio resource every T_{cycle} milliseconds; the minimum radio resource should be given by a RLC (radio-link control) block, which in the GPRS is transmitted over four GSM bursts; with this choice, the packet size (at network level) is around a few hundreds of bits, but this is enough to provide an EW message.

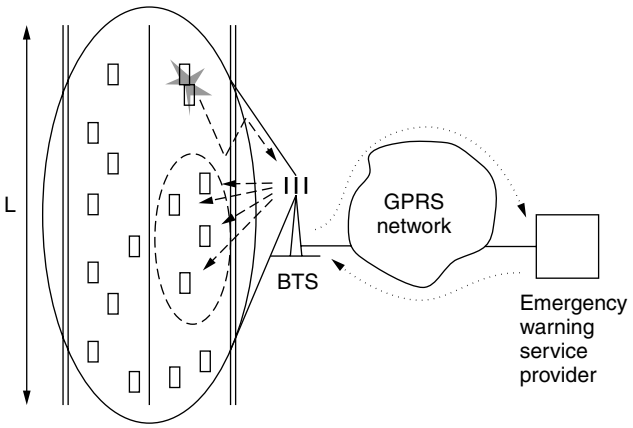


Figure 4. An ITS scenario based on GPRS: the service provider receives the Warning message from a mobile, and sends it to all the vehicles that can be involved in the information, considering the location, speed, and direction of the mobiles, the traffic conditions and the requirements of the application (© IEEE, 2000).

When a packet is generated by a mobile, it is transmitted through the GPRS network to a fixed node (representing the serving entity introduced by the service provider) that retransmits the message in multicast mode to all the vehicles in the vicinity of the source (see Fig. 4). The subset of vehicles to be addressed by the warning message has to be carefully determined taking different aspects into account, such as the speed of mobiles, the traveling direction of the source (it can be assumed that only the vehicles behind that involved in the dangerous event should be warned), and the traffic density; all these data can be known at the service provider if all vehicles periodically inform the network with their locations, speed, and other specifics. Therefore, communication, location and traffic management aspects have to be considered at the centralized node.

The implementation of a centralized control on the network side has the main advantage that the message generated on the vehicle has a fixed destination address; therefore, the determination of the subset of vehicles that should receive the warning message is controlled by the service provider, and this can be efficiently performed if all vehicles periodically inform the network of their locations. On the other hand, this choice implies a double connection, and double transfer delay, so special attention should be paid to this assessment.

The time spent in the network by the packet is subject to the delays introduced by the network entities and the time needed for radio access. The choice of T_{cycle} is of relevance, since it is a measure of the maximum delay that the message can suffer due to the radio access: in any case, the value of T_{cycle} should be small, in order to avoid a large contribution to the overall message delivery delay. On the other hand, the number of vehicles that can be served is proportional to T_{cycle}; hence, T_{cycle} should be

fixed as a tradeoff between the network capacity and the desired promptness.

It is worth noting, however, that the number of vehicles that can be served is also proportional to the number m of GPRS carriers used for the service, hence, having fixed T_{cycle}, the service coverage can be improved increasing the value of m.

Let's assume, in the six lanes of our scenario, an average intervehicle distance, d_{iv}; consequently the number of vehicles present in a trunk of length L served by the same BTS, is given by $N_{\text{veh}} = 6L/d_{\text{iv}}$; let's assume, furthermore, that the percentage of vehicles equipped with the GPRS terminal is 100 k, where k is the penetration factor.

Figure 5 shows the maximum value of the cell size L that is compatible with a traffic of kN_{veh} vehicles as a function of the penetration k; both cases of one and two GPRS carrier(s) are considered, and the intervehicle distance is fixed at 25 m.

Figure 5 shows that, for $m = 1$, at the first stage of service provision (when the penetration is low), cell sizes of around tens of kilometers; that is, the typical cell sizes of the GSM network in suburban environments, can be used, thus allowing the utilization of the existing sites. On the other hand, once the penetration becomes much larger (e.g., $k = 0.25$), new investments should be introduced, either by adding new GPRS carriers or by reducing the cell sizes; however, this would be the case of a very mature service used by millions of users, such as by a successful business.

Finally, it is worth noting that, when no packets with highest priority (i.e., EW packets) have to be transmitted, the radio resources can be dynamically reallocated to other services; therefore, the GPRS carrier(s) is (are) not dedicated to the EW service, since it (they) can be shared by other services having lower priority. For instance, the provision of information concerning traffic and weather conditions (on subscription, or on demand) or the transmission of roadmaps could be offered to the user through the same physical channels. All these services,

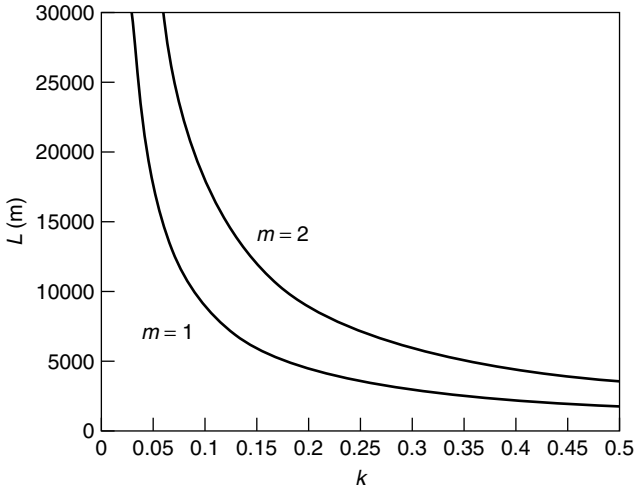


Figure 5. The maximum cell size (having fixed $T_{\text{cycle}} = 500$ ms) as a function of the penetration of users, with $m = 1, 2$ GPRS carriers and for an average intervehicle distance fixed at 25 m (© IEEE, 2000).

together with the selection of vehicles to be included in the multicast group to which the warning message should be delivered, require the knowledge (with different degrees of precision) of the vehicle locations. A mobile terminal can solve the problem of self-localization in a number of ways (through the capabilities of GSM, or through a GPS receiver, etc); in any case, the information concerning the vehicle position should be transmitted periodically to the network and this could be done by means of the allocated radio resource units.

4.2. The Role of 3G Networks

GPRS is expected to become a successful standard in the near future; however, it should only represent an evolutionary path toward more sophisticated systems such as UMTS (Universal Mobile Telecommunication System) [9], the 3G cellular network, which should replace the previous 2G/2.5G systems. It seems reasonable to assume that in many countries the 3G core network will be initially based on, or will also embrace, the GSM/GPRS core network; the main difference compared to the previous standards will be on the air interface, which in 3G systems will rely on CDMA techniques, unlike those for GSM and GPRS. However, the use of multistandard (multimode) terminals will determine a potential integration of the different air interfaces (e.g., GPRS or wideband CDMA). Therefore, the most revolutionary concept introduced by 3G solutions, that is, the possibility of having a full-coverage cellular system based on packet switching, thus allowing the extension of many services based on IP (Internet Protocol) and the adoption of a "pay what you send" billing method to the wireless context, will be introduced with GPRS. In this regard, it should be emphasized that UMTS will stimulate this concept by providing larger bandwidths.

The previous example of potential application of GPRS in the field of ITS shows some limitations; for instance, the limited bandwidth of 2.5G systems affects the promptness of the "warning message" service, as well as the delays introduced by the fixed network.

The portion of spectrum allocated to 3G standards is larger than that dedicated to 2G (and 2.5G) systems; therefore, the provision of EW services through GPRS can be seen as a first step toward a more sophisticated solution, which could be provided by means of 3G networks. In fact, the delays introduced by the GPRS network depend partly on the limited bandwidth of the GPRS system, as shown before. Within this context, the main role that 3G systems will be able to play, with respect to GPRS, will be given by the larger bandwidths that could affect the quality of the services offered.

The larger the overall bit rate, the smaller the radio access delay, the smaller the delay of message delivery, and the higher the spatial resolution of the warning message. With GPRS, taking its network delays into account, it can be expected that, at a speed of 120 km/h, a vehicle can be alerted to avoid an accident that occurred about 150–200 ms ahead. This is not enough to elude all possible events. On the other hand, this could be sufficient to avoid the catastrophic highway "chain accidents" (occurring sometimes in the presence of huge banks of fogs,

typical of different areas in Italy) that in some cases involve many cars, separated by distances much greater than 100 m. A rough EW service based on GPRS could save some human lives. The larger bandwidth of 3G standards could do even better.

However, this does not reduce the importance of the investigation of dedicated systems for ITS, which could provide even more complex services, such as "cooperative driving," which should be considered in the long term to be a very important target for increasing the safety and efficiency of road transport.

4.3. The Role of Bluetooth

The main disadvantage of the adoption of cellular communication (GSM, GPRS, or UMTS) for ITS service provision relies on the fact that direct communication between the interacting entities would not be possible; the exchange of data between the two would be performed via the cellular network. This would require suitable integration of the cellular service with localization aspects, would make the service provision prone to the network malfunction, and would increase the packet transfer delay.

An alternative solution is represented by the adoption of devices with the capability, on one hand, to automatically detect the presence of other similar devices located nearby and, on the other hand, to establish a direct communication link (hence without the need of an infrastructure network).

The Bluetooth technology, the most recent development in the field of WPANs (wireless personal-area networks), fulfills both the abovementioned requirements; furthermore, Bluetooth-based devices are expected to be lightweight, small, and economic, therefore representing a perfect candidate not only for RVC systems but also for TVC systems.

The aim of the Bluetooth technology is to allow the establishment of effortless, wireless, instant, and low-cost connections between various communication devices. The Bluetooth radio is built into a small microchip, which is estimated to cost just a few dollars, and operates in a globally available frequency band [the ISM (industrial–scientific–medical)-band at 2.4 GHz] thus ensuring communication compatibility worldwide.

Two or more units communicating one with the other(s) form a *piconet*, where one unit acts as a "master," controlling the whole traffic in the piconet and the other units act as "slaves."

The master implements a centralized control; only communications between the master and one or more slaves are possible, and there is a strict alternation between master and slave transmissions.

A simple binary, Gaussian-shaped FSK modulation scheme (GFSK) is applied in order to reduce costs and device complexity and a symbol rate of 1 Mbps can be achieved. Three different power classes are defined within Bluetooth, as reported in Table 1, where P_0 is the maximum transmitted power.

4.3.1. Estimation of the Maximum Range. An important aspect to be investigated is the maximum distance between the transmitting and receiving devices that ensures possibility of communicating.

Table 1. Bluetooth-Defined Power Classes

Power Class	Maximum Transmitter Power
1	$P_0 = 100$ mW
2	$P_0 = 2.5$ mW
3	$P_0 = 1$ mW

A precise assessment of the maximum range allowed, by the Bluetooth technology under nonideal conditions requires an experimental testbed; hence we carried out a measurement campaign with the aim to evaluate the performance of a Bluetooth link in an outdoor scenario and in different propagation conditions.

A data communication link between two commercial class 1 Bluetooth devices separated by a distance d was established and the time needed to perform a 1.57-MB (megabyte) FTP (File Transfer Protocol) transmission measured. The scope of this test was to evaluate the limit distance that determines a significant increase in the file transfer delay (T_d) due to the retransmissions requested by the ARQ strategy adopted by Bluetooth; this experiment was performed considering both a "free-space-like" [i.e., Line-of-Sight (LoS)] and a partially obstructed LoS environment.

In the first case we ensured that the volume enclosed in the ellipsoid defined by the first Fresnel zone was free from obstacles; in the second case the electromagnetic propagation was partially obstructed by the presence of the terrain (we considered the transmitter and the receiver placed at a height of 45 cm from the terrain).

In Fig. 6 the file transfer delay, T_d, is reported as a function of the distance d between the two Bluetooth devices in the case of "free-space-like" propagation conditions; as we can observe, T_d is not affected by the value of d until the distance between the two devices becomes larger than about 110 m, after which T_d rapidly increases. Ten transfers were performed.

Consequently, under the conditions considered, the maximum communication range can be estimated to be

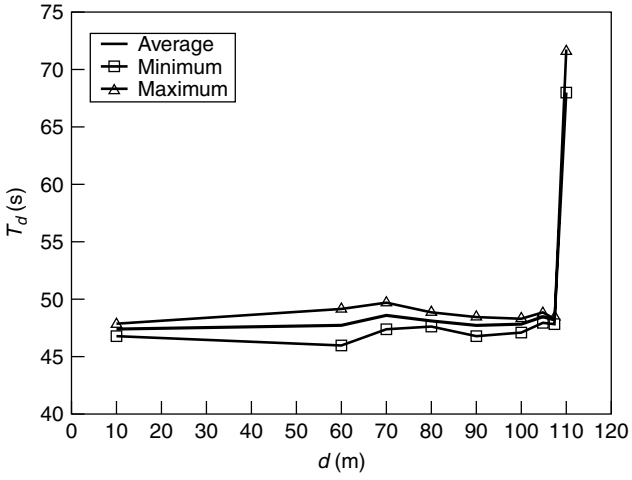


Figure 6. File transfer delay between two Bluetooth devices: "free-space-like" condition.

around 100 m for RVC systems, which are typically experiencing LOS conditions, owing to the possibility to place the beacon in suitable positions.

The experiment described above was repeated assuming partially obstructed LoS conditions: the transmitter and the receiver were placed at distances 45 cm above the terrain, made up of an asphalted surface, representing an obstruction for the first Fresnel ellipsoid when the distance d is >7 ms. This scenario could be representative of an IVC system for, example.

In Fig. 7 the result of this measurement campaign is shown. As we can observe, when the distance between the transmitting and receiving devices is lower than 60 m, the file transfer delay is not affected by the presence of the terrain, which later determines a rapid increase of T_d. Consequently, under the conditions considered, the threshold distance value that marks the transition between LoS and non-LoS conditions is significantly larger than the theoretical value of 7 ms; this observation leads to the conclusion that the reflections produced by the road can improve communication by extending the coverage distance with respect to what is predicted by simple modeling.

Analytic estimation of the maximum range, as in every wireless system, requires the assessment of a suitable propagation law, which is complex because of the presence of many vehicles and other obstacles, thus requiring the specification of the operating environment.

In this section we merely give an approximate evaluation of the maximum range, taking only free-space loss into consideration.

The minimum level of the received power, on the input to the receiver, required for a bit error rate equal to 10^{-3} is $P_r = -70$ dBm [31]; therefore, the maximum loss $L_{\max}$ (dB) $= P_0$ (dB) $- P_r$ (dB) depends on the power class and is reported in the second column of Table 2.

Knowing the $L_{\max}$ value, we can assess the maximum allowable distance $d_{\max}$; let's assume free-space loss (optimistic conditions), so that, assuming the value of carrier frequency equal to 2.4 GHz, we obtain

$$L_{\max} = 40 + 20\log(d_{\max}) - G_e - G_r$$

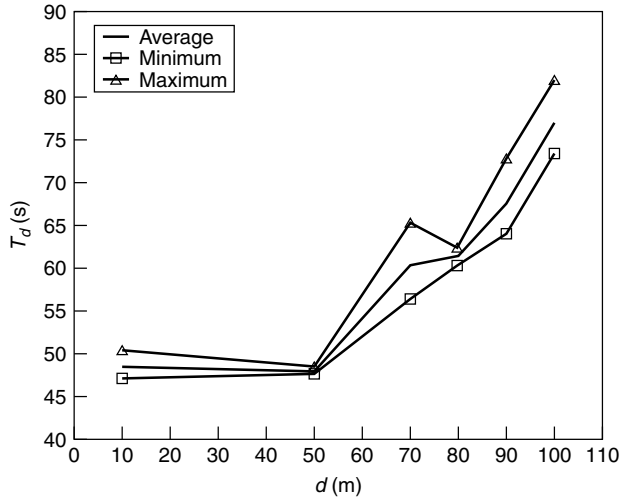


Figure 7. File transfer delay between two Bluetooth devices: "non-LoS" condition.

Table 2. Minimum Received Power Required for BER = 10^{-3}

Power Class	$L_{\max}$ (dB)	$d_{\max}$ (m) ($G_e = G_r = 0$ dB)
1	90	316
2	74	50
3	70	32

where G_e and G_r are the transmitter and receiver antenna gains, respectively.

Under the simple assumption of omnidirectional antennae ($G_e = G_r = 0$ dB), with free-space loss, we have

$$L = 40 + 20\log(d_{\max})$$

which gives the values reported in the third column of Table 2.

The coverage range calculated with our very simplified approach, which doesn't take nonideality typical of a real system into account, is in the case of class 1 devices, around a few hundreds of meters, not too different from the measured values reported above, which were obtained with commercial products, probably not transmitting at the maximum allowed level of power.

However, free-space loss is a rather optimistic assumption that in many cases could not be fulfilled in realistic conditions, especially in urban scenarios. The assessment of a loss formula taking multipath into account should be considered, but is beyond the scope of this work, as it would depend on the particular environment (urban, suburban), including the traffic conditions and the number of scatterers (cars, buildings, etc.).

4.3.2. Effects of Vehicle Speed and Multipath. We note that the Bluetooth technology was designed for static indoor environments (communication between computers, printers, etc.); the possibility of using it in an outdoor scenario, with moving terminals at different speeds, must be carefully investigated by taking the protocol and physical aspects of Bluetooth technology into account and considering the typical aspects related to the propagation of electromagnetic waves in outdoor urban environments, characterized by fading and multipath effects.

This analysis requires the consideration of the consequences of the Doppler effect (viz., the effects of the speed of terminals) on the modulation/demodulation format and the protocols.

Concerning signal dispersion, the following preliminary considerations lead to the conclusion that the dispersion due to multipath should not represent a serious drawback: Bluetooth is a short-range communication system; hence the coherence bandwidth is expected to be around tens of megahertz, that is, well beyond the signal bandwidth (1 MHz), and this renders the performance insensitive to the amount of signal dispersion. A more precise consideration of this aspect would be very complex because of the characteristics of the Bluetooth technology.

As far as the effects of speed are concerned, an analytic model has been exploited to assess the bit error rate

of a Bluetooth device under the conditions stated above (neglecting the benefits of the channel code). It is worth noting that the modulation format is nonlinear and its performance is evaluated in the presence of flat fading.

In Fig. 8 the BER (for a Bluetooth terminal neglecting the effects of channel coding, which will further improve the performance) as a function of average signal-to-noise ratio W_m, is reported where the average is made with respect to fading statistics, considered to be flat Rayleigh. The computation takes the effect of the speed of terminals into account (viz., the Doppler effect, for a relative speed equal to 110 km/h, considering a vehicle and a target running in opposite directions, one toward the other, i.e., on a "head-on collision course"). Since the implementation of the receiver characteristics is not fixed by the standard a limiter–discriminator device is assumed, as a reference, with a 4-pole Butterworth filter having normalized bandwidth equal to 1. It is expected that many implementations of the Bluetooth technology will be based on limiter–discriminator detection, which is known to be simple and efficient under hostile propagation conditions. The choice related to the filter has proved not to have a deep impact on the results.

The results show that an error floor is found, due to the speed of the terminal, at BER $= 10^{-6}$. We assume that this level of BER is sufficiently large to fulfill the performance requirement in terms of false-alarm rate. Therefore, the movement of terminals does not seem to introduce limiting effects on the performance (let's bear in mind that the effect of channel coding was not considered, and it will further lower the error floor).

In order to verify the results of our analysis, we carried out experimental activities to test the Bluetooth transmission reliability in a dynamic scenario. A communication link was established between a static class 1 Bluetooth transmitter placed on the pavement beside an urban road, at 50 cm above the terrain, and a Bluetooth receiver placed on a vehicle traveling at a constant speed and passing in front of the transmitter at a minimum distance of 3 ms.

The experiment was repeated assuming different speeds of the mobile device, and we ensured that the transmission was performed correctly even when traveling at 100 km/h, therefore confirming our expectation on the sturdiness of Bluetooth communications.

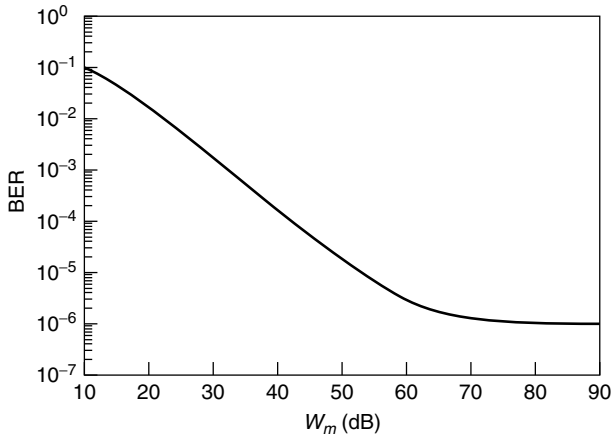


Figure 8. Bluetooth link: the effect of mobility.

As a consequence of the previous observations, we can state that link performance is not limited by the speed of the vehicle but is determined by the amount of signal-to-noise ratio and, consequently, by the transmitted power. On the other hand, it is well known that Bluetooth uses a polling technique at the MAC (media access control) level, and this means that, in the presence of many vehicles, the MAC protocol can cause a performance degradation.

5. CONCLUSIONS

This article presented an overview of the evolution of research in the field of communications for ITS. Starting from the activities oriented to dedicated solutions, the discussion then focused on the assessment of existing relevant standards in mobile communications to define and activate new ITS services. As far as dedicated solutions are concerned, this article showed some results related to radars as well as 60-GHz and infrared communication systems specifically designed for RVC and IVC. The issue of ITS service provision by means of an already standardized system was then addressed, and, as an example, the feasibility of an "emergency warning" service based on GPRS was considered. Finally, the performance of Bluetooth-based RVC and TVC links was addressed by both analytically and experimentally.

ACRONYMS

AHS	Automated highway system
ARQ	Automatic Repeat reQuest; automatic repeat request (generic)
BER	Bit error rate
BTS	Base transceiver station
CCD	Charge-coupled device
CDMA	Code-division multiple access
DSRC	Dedicated short-range communication
FM	Frequency modulation
FM-CW	Frequency-modulated continuous wave
FSK	Frequency shift keying
FTP	File Transfer Protocol
GFSK	Gaussian frequency shift keying
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
IP	Internet Protocol
ISM	Industrial–scientific–medical
ITS	Intelligent transportation system
IVC	Intervehicle communications
JSK	Association of Electronic Technology for Automobile Traffic and Driving (Japan)
LOS	Line of sight
MAC	Medium access control
MSK	Minimum shift keying
RF	Radiofrequency
RLC	Radio link control
RVC	Roadside-to-vehicle communication
SIR	Signal-to-interference ratio
SS	Spread spectrum
TDMA	Time-division multiple access

TVC	Target-to-vehicle communication
UMTS	Universal Mobile Telecommunication System
USDOT	United States Department of Transportation
VMS	Variable message sign
WPAN	Wireless personal-area network

BIOGRAPHIES

Oreste Andrisano was born in Bologna, Italy, on February 14, 1952. He received the Dr. Ing. degree in electronic engineering cum laude from the University of Bologna, Bologna, Italy, in 1975. In the same year he joined the University of Bologna, where he became a Professor of Electrical Engineering in 1985. Since 1992 he has been the Director of CSITE (Centro di studio per l'Informatica e i Sistemi di Telecomunicazioni), University of Bologna and Consiglio Nazionale delle Ricerche, Roma. Since 2000 he has been the Director of the Laboratorio Nazionale di Comunicazioni Multimediali, Napoli (Consorzio Nazionale Interuniversitario per le Telecomunicazioni, CNIT). In the period 1996-2001 he was an Editor of the IEEE Transactions on Communications (Modulation for fading channels). His research activity has been concerned with different fields in the digital communication area, such as digital signal processing, data transmission for satellite and fixed radio links applications, local wireless and mobile radio networks. He has also been active in the Intelligent Transportation Systems (ITS) area, with reference to vehicle-to-vehicle and vehicle-to-infrastructure communication systems. In 1987 and 1988 he cooperated in the definition phase of PROMETHEUS (EUREKA), as a European coordinator for the transmission systems research area. Then, he was in the Steering Committee of Project DACAR (Data Acquisition and Communication Techniques and their Assessment for Road Transport) in the framework of DRIVE I, ECC, 1988-1991. In the period 1991-1997 he was responsible for the ITS communication activities in Italy (coordination of PROCOM and TELCO). Since 1998 he is the national coordinator of the project Multimedia Systems funded at national level by MIUR and CNR (Roma). Oreste Andrisano is a member of the IEEE Communication and Vehicular Technology Societies and of the IEEE Radiocommunication Committee.

Masao Nakagawa was born in Tokyo, Japan, in 1946. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1969, 1971 and 1974 respectively. Since 1973, he has been with the Department of Electrical Engineering, Keio University, where he is now Professor. His research interests are in CDMA, consumer communications, mobile communications, ITS (Intelligent Transport Systems), wireless home networks, and optical communication with lighting. He received 1989 IEEE Consumer Electronics Society Paper Award, 1999-Fall Best Paper Award in IEEE VTC, IEICE Achievement Award in 2000, IEICE Fellow Award in 2001. He was the executive committee chairman of International Symposium on Spread Spectrum Techniques and Applications in 1992 and the technical program committee chairman of ISITA (International Symposium on Information Theory and its Applications) in 1994. He is an editor of Wireless Personal Communications and was a guest editor of the special issues on "CDMA Networks I, II, III and IV" published in IEEE JSAC in 1994(I and II) and 1996(III and IV). He chairs the Wireless Home Link subcommittee in MMAC (Multimedia Mobile Access Communication Promotion Committee).

Gianni Pasolini received his Laurea degree in telecommunications engineering from the University of Bologna, Italy, in March 1999. In May 1999 he joined CSITE-CNR (Centre for Studies in Computer Science and Telecommunication Systems of the National Research Council), and he is currently working toward his PhD. His research activity is concerned with Wireless Local Area Networks, Digital Communications and Radio Resource Management. He is a student member of IEEE.

Roberto Verdone was born in Bologna, Italy, in 1965. He received his Laurea degree in Electronic Engineering (with honours) and his Ph.D. in Electronic Engineering and Computer Science from the University of Bologna, Italy, in March 1991 and October 1995, respectively. In April 1996 he became a researcher at CSITE-CNR (Centre for Studies in Computer Science and Telecommunication Systems of the National Research Council) in Telecommunications. Since November 2001 he is a Full Professor in Telecommunications at the University of Bologna. His research activity is concerned with Digital Transmission, Cellular and Mobile Radio Systems, Wireless Local Area Networks, Digital Broadcasting and Intelligent Transportation Systems. In 1992 he was involved in the European research program PROMETHEUS. Since 1995 to 1997 he worked in the context of the National research program TELCO (TELecommunications network for COoperative Driving). He is also responsible for the activities of CSITE in projects funded by ESA and MIUR (Italy). Since 1997 to 2000 he has participated to the COST259 Action. Since 2001 he is chairman of the WG on Network Aspects within the follow-on Action COST273. He is a member of IEEE.

BIBLIOGRAPHY

1. *DRIVE Workplan*, April 26, 1988.
2. *PROMETHEUS PRO-COM White Book, Definition Phase*, Feb. 19, 1988.
3. *IEEE Commun. Mag.* (Half Special Issue on IVHS), **34**(10) (Oct. 1996).
4. *IEEE Trans. Vehic. Technol.* (Special issue on Intelligent Vehicle Highway Systems), **40**(1) (Feb. 1991).
5. *Proc. 7th World Congress on Intelligent Transportation Systems*, Turin, Italy, Nov. 6–9, 2000.
6. *Final demonstration of PROMETHEUS project*, Paris, France, Oct. 1994.
7. ERTICO homepage, *http://www.ertico.com/*.
8. I. Catling and R. Harris, SOCRATES — progress towards commercial implementation, *Proc. Vehicle Navigation Information Systems Conf.* (*VNIS'95*), Seattle, WA, July 30–Aug. 2, 1995.
9. *IEEE Trans. Vehic. Technol.* (Special Issue on the UMTS), **47**(4) (Nov. 1998).

10. Bluetooth homepage, *http://www.bluetooth.com/*.

11. PATH homepage, *http://www.path.berkeley.edu/*.

12. *California PATH Annual Report 2000*, *http://www.path.berkeley.edu/*.

13. O. Andrisano, M. Chiani, V. Tralli, and R. Verdone, Impact of cochannel interference on vehicle-to-vehicle communications at millimetre waves, *Proc. IEEE Int. Conf. Communication Systems* (*ICCS'92*), Singapore, Nov. 16–20, 1992, Vol. 2, pp. 924–928.

14. W. Kremer et al., Computer-aided design and evaluation of mobile radio local area networks in RTI/IVHS environments, *IEEE J. Select. Areas Commun.* **11**(3): 406–421 (April 1993).

15. O. Andrisano et al., Millimetre wave short range communications for advanced transport telematics, *Eur. Trans. Telecommun.* (July–Aug. 1993).

16. H. Fujii, O. Hayashi, and N. Nagakata, Experimental research on inter-vehicle communication using infrared, *Proc. IEEE Intelligent Vehicles Symp.*, 1996, pp. 266–271.

17. R. Verdone, Communication systems at millimetre waves for ITS applications, *Proc. IEEE Vehicular Technology Conf. 1997* (*VTC'97*), Phoenix, AZ, May 5–7, 1997, Vol. 2, pp. 914–918.

18. R. Verdone, Outage probability analysis for short range communication systems at 60 GHz in ATT urban environments, *IEEE Trans. Vehic. Technol.* **46**(4): 1027–1039 (Nov. 1997).

19. R. Verdone, Multi-hop R-ALOHA for inter-vehicle communications at millimetre waves, *IEEE Trans. Vehic. Technol.* **46**(4): 992–1005 (Nov. 1997).

20. M. Chiani and R. Verdone, A TDD-TCDMA radio interface at millimetre waves for ITS applications, *Proc. IEEE Vehicular Technology Conf.* (*VTC'99—Fall*), Amsterdam, The Netherlands, Sept. 19–22, 1999, Vol. 2, pp. 770–774.

21. L. Michael and M. Nakagawa, Interference characteristics in inter-vehicle communication from oncoming vehicles, *Proc. IEEE Vehicular Technology Conf.* (*VTC'99–Fall*), Amsterdam, The Netherlands, Sept. 19–22, 1999, Vol. 2, pp. 753–757.

22. Y. Shiraki et al., Evaluation of interference reduction effect of SS radar, *Proc. 1999 IEICE General Conf.*, A-17–22, 1999, p. 421.

23. CSITE homepage, *http://www-csite.deis.unibo.it/htcsite/prometheus/Indice.html*.

24. C. Cseh, Architecture of the dedicated short-range communications (DSRC) protocol, *Proc. 48th IEEE Vehicular Technology Conf.*, 1998, Vol. 3, pp. 2095–2099.

25. CEN TC 278 WG 9 homepage, *Dedicated Short-Range Communications*, *http://www.comnets.rwth-aachen.de/~ftp-wg9/*.

26. R. Yuan, North American dedicated short range communications (DSRC) standards, *Proc. IEEE Conf. Intelligent Transportation System, 1997* (*ITSC '97*), pp. 537–542.

27. O. Hyunseo, Y. Chungil, A. Donghyon, and C. Hanberg, 5.8 GHz DSRC packet communication system for ITS services, *Proc. 50th IEEE Vehicular Technology Conf., 1999* (*VTC 1999—Fall*), 1999, Vol. 4, pp. 2223–2227.

28. I. Catling, R. Harris, L. James, and N. Simmons, ITS services in Wales using the wireless application protocol (WAP), *Proc. Int. Conf. Advanced Driver Assistance Systems, 2001* (*ADAS*), pp. 73–75.

29. R. Nusser and R. M. Pelz, Bluetooth-based wireless connectivity in an automotive environment, *Proc. 52nd Vehicular Technology Conf., 2000.* (*IEEE-VTS—Fall VTC*), 2000, Vol. 4, pp. 1935–1942.

30. O. Andrisano, R. Verdone, and M. Nakagawa, Intelligent transportation systems: The role of third generation mobile radio networks, *IEEE Commun. Mag.* **38**(9): 144–151 (Sept. 2000).

31. *Specification of the Bluetooth System*, Core, version 1.0B.

COMMUNITY ANTENNA TELEVISION (CATV) (CABLE TELEVISION)

ROGER FREEMAN*
Independent Consultant
Scottsdale, Arizona

1. INTRODUCTION

The principal thrust of community antenna television (CATV) is entertainment. Lately, CATV has taken on some new dimensions. It is indeed a broadband medium, providing up to 1 GHz of bandwidth at customer premises. It was originally a unidirectional system, from the point of origin, which we call the *headend*, toward customer premises. It does, though, have the capability of being a two-way system by splitting the band, say, from 5 to 50 MHz for upstream traffic (i.e., toward the headend), and the remainder is used for downstream traffic (i.e., from the headend to customer premises). CATV is certainly a major and viable contender for *last-mile communications*.

We will briefly discuss one approach to provide capability for two-way traffic, usually voice and data. First, however, conventional CATV will be described, and it includes the concept of supertrunks and HFC (hybrid fiber–coaxial cable systems). We will involve the reader with such topics as wideband amplifiers in tandem, optimum amplifier gain, IM (intermodulation) noise, beat noise, and cross-modulation products. System layout, hubs, and last-mile or last-100-ft considerations will also be covered. There will also be a brief discussion of the conversion to a digital system using some of the compression techniques now employed on modern cable television systems. The section includes overviews of the three important competing standards for broadband hybridcoax (HFC) CATV systems. These provide the user, besides conventional TV downstream, upstream, and downstream connectivity for megabit Internet, IP intranet, various data services, still-image transmission, and POTS* telephony.

* Roger Freeman took an early retirement from the Raytheon Company, Equipment Division, in 1991 where he was principal engineer to establish *Roger Freeman Associates*, Independent Consultants in Telecommunications. He has been writing books on various telecommunication disciplines for John Wiley & Sons, since 1973. Roger has seven titles that he keeps current, including *Reference Manual, Inc. for Telecommunication Engineers* now in its third edition. His Website *www.rogerfreeman.com*, and his email address is *rogerf@pcslink.com.*

* POTS-plain old telephone service.

2. THE EVOLUTION OF CATV

2.1. The Beginnings

Broadcast television, as we know it, was in its infancy around 1948. Fringe-area problems were much more acute in that period. By "fringe area," we mean areas with poor or scanty signal coverage. A few TV users in fringe areas found that if they raised their antennas high enough and improved antenna gain characteristics, they could receive an excellent picture. Such users were the envy of the neighborhood. Several of these people who were familiar with RF signal transmission employed signal splitters so that their neighbors could share the excellent picture service. It was soon found that there was a limit on how much signal splitting could be done before signal levels got so low that they were snowy or unusable.

Remember that each time a signal splitter (even power split) is added, neglecting insertion losses, the TV signal dropped by 3 dB. Then someone got the bright idea of amplifying the signal before splitting. Now some real problems arose. One-channel amplification worked fine, but two channels from two antennas with signal combining became difficult. Now we are dealing with comparatively broadband amplifiers. Among the impairments we can expect from broadband amplifiers and their connected transmission lines (coaxial cable) are

- Poor frequency response. Some part of the received band had notably lower levels than other parts. This is particularly true as the frequency increases. In other words, there was fairly severe amplitude distortion; thus equalization became necessary.
- The mixing of two or more RF signals in the system caused intermodulation products and "beats" (harmonics), which degraded reception.
- When these TV signals carried modulation, cross-modulation (Xm) products further degraded or impaired reception.

Several small companies were formed to sell these "improved" television reception services. Some of the technicians working for these companies undertook ways of curing the ills of broadband amplifiers. What these service companies did was to install a reasonably high tower in an appropriate location. Comparatively high gain antennas were installed such that a clear, line-of-sight signal could be received from TV emitters within range. The high-tower receiving site was called a *headend*. The headend had RF amplifiers, TV line amplifiers, signal combiners, translators, and all that was necessary to process and distribute multiple TV signals to CATV subscribers. The distribution system was entirely based on coaxial cable. To keep the signal strength to a usable level, broadband amplifiers were installed at convenient intervals along the distribution line.

A subscriber's TV set was connected to the distribution system, and the signal received looked just the same as if it was taken off the air with its own antenna. In fringe areas signal quality, however, was much better than own antenna quality. The key to everything was that no changes were required in the users' TV set. This was just an extension of his/her TV set antenna. Such a simple concept is shown in Fig. 1.

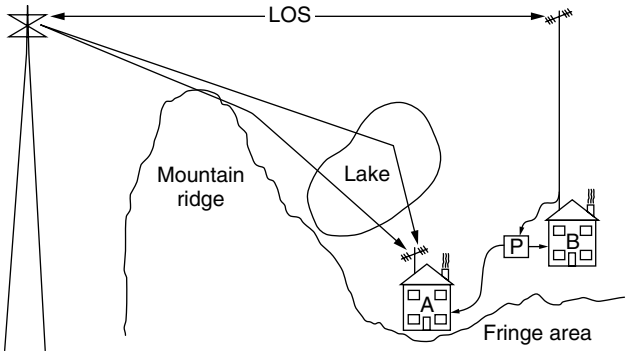


Figure 1. CATV initial concept (P = power split).

Note in Fig. 1 that home A is in the shadow of a mountain ridge and receives a weakened diffracted signal off the ridge and a reflected signal off a lake. Here is the typical multipath scenario resulting in ghosts in A's TV screen. The picture is also snowy, meaning it is noisy, as a result of a poor carrier-to-noise ratio. Home B extended the height of the antenna to be in line of sight of the TV transmitting antenna. Its antenna is of higher gain; thus it is more discriminating against unwanted reflected and diffracted signals. Home B has an excellent picture without ghosts. Home B shares its fine signal with home A by use of a 3-dB power split (P).

2.2. Early System Layouts

Figure 2 illustrates an early CATV distribution system (ca. 1968). Taps and couplers (power splits) are not shown. These systems provided 5–12 channels. A microwave system brought in channels from distant cities (50–150 mi). We had direct experience with the Atlantic City, NJ system where channels were brought in from Philadelphia and New York by microwave (MW). A 12-channel system resuited which occupied the entire assigned VHF band (i.e., channels 2–13).

As UHF TV stations began to appear, a new problem arose for the CATV operator. It was incumbent on that operator to keep the bandwidth as narrow as possible. One approach was to convert UHF channels to vacant VHF channel allocations at the headend.

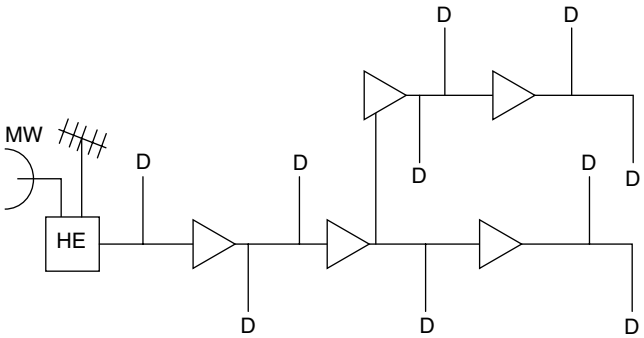


Figure 2. An early CATV distribution system (HE = headend; D = drop wire to residence; MW = microwave connectivity).

Satellite reception at the headend doubled or tripled the number channels that could be available to the CATV subscriber. Each satellite has the potential of adding 24 channels to the system. Note how the usable cable bandwidth is "broadened" as channels are added. We assume contiguous channels across the band, starting at 55 MHz. For 30 channels, we have 55–270 MHz; 35 channels, 55–300 MHz, 40 channels, 55–330 MHz; 62 channels, 55–450 MHz; and 78 channels, 55–550 MHz. These numbers of channels were beyond the capability of many TV sets of the day. Settop converters were provided that converted all channels to a common channel, an unoccupied channel, usually channel 2, 3 or 4 to which the home TV set is tuned. This approach is still very prevalent today.

In the next section we discuss impairments and measures of system performance. In Section 4, hybrid fiber–coaxial cable systems are addressed. The replacement of coaxial cable trunk by fiberoptic cable made a major stride to improved performance and system reliability/availability.

3. SYSTEM IMPAIRMENTS AND PERFORMANCE MEASURES

3.1. Overview

A CATV headend places multiple TV and FM broadcast (from 30 to 125) carriers on a broadband coaxial cable trunk and distribution system. The objective is to deliver a signal-to-noise ratio (S/N) of 42–45 dB at a subscriber's TV set. If the reader has background in the public switched telecommunications network (PSTN), he/she would expect such impairments as the accumulation of thermal and intermodulation noise. We find that CATV technicians use the term *beat* to mean intermodulation (IM) products. For example, there is triple beat distortion, defined by Bill Grant [6], as "spurious signals generated when three or more carriers are passed through a non-linear circuit (such as a wideband amplifier). The spurious signals are sum and difference products of any three carriers, sometimes referred to as 'beats.' Triple beat distortion is calculated as a voltage addition."

The wider the system bandwidth is and the more RF carriers transported on that system, the more intermodulation distortion, "triple beats" and cross-modulation we can expect. We can anticipate combinations of all the above such as *composite triple beat* (CTB), which represents the pile up of beats at or near a single frequency.

Bill Grant [6] draws a dividing line at 21 TV channels. On a system with 21 channels or less one must expect cross-modulation (Xm) to predominate. Above 21 channels, CTB will predominate.

3.2. dBmV and Its Applications

The value 0 dBmV is defined as 1 millivolt (mV) across an impedance of 75 Ω. The 75-Ω value is the standard impedance used in CATV. From the power law

$$P_w = E^2/R$$

$$P_w = 0.001^2/75 \tag{1}$$

$$0 \text{ dBmV} = 0.0133 \times 10^{-6} \text{ watts or } 0.0133 \ \mu\text{W}$$

By definition, then, 0.0133 *watts* (W) is +60 dBmV.

If $0 \text{ dBmV} = 0.0133 \times 10^{-6}$ W and 0 dBm = 0.001 W, and gain in $\text{dB} = 10\log(P_1)/(P_2)$, or in this case $10\log(0.001/0.0133 \times 10^{-6})$, then 0 dBm = +48.76 dBmV.

Remember that when working with decibels in the voltage domain, we are working with the E^2/R relationship, where $R = 75\ \Omega$. With this in mind, the definition of dBmV is

$$\text{dBmV} = 20\log(\text{voltage in millivolts})/(1 \text{ mV}) \tag{2}$$

If a signal level is 1 volt (V) at a certain point in a circuit, what is the level in dBmV?

$$\text{dBmV} = 20\log(1000)/1 = +60 \text{ dBmV}$$

If we are given a signal level of +6 dBmV, what voltage level does this correspond to?

$$+6 \text{ dBmV} = 20\log(X_{\text{mV}})/1 \text{ mV}$$

Divide through by 20:

$$\frac{6}{20} = \log(X_{\text{mV}})/1 \text{ mV}$$

$$\text{Antilog}\left(\frac{6}{20}\right) = X_{\text{mV}}$$

$$X_{\text{mV}} = 1.995 \text{ mV or } 2 \text{ mV or } 0.002 \text{ V}.$$

These signal voltages are RMS (root mean square) volts. For peak voltage, divide by 0.707. If we are given peak signal voltage and wish the RMS value, multiply by 0.707.

3.3. Thermal Noise in CATV Systems

The lowest noise levels permissible in a CATV system: at antenna output terminals, at repeater (amplifier) inputs or at a subscriber's TV set, without producing snowy pictures, are determined by thermal noise.

Consider the following, remembering that we are in the voltage domain. Any resistor or source that appears resistive over the band of interest, including antennas, amplifiers, and long runs of coaxial cable, generates thermal noise. In the case of a resistor, the noise level can be calculated on the basis Fig. 3.

To calculate the noise voltage, e_n, use the following formula:

$$e_n = (4RBk)^{1/2} \tag{3}$$

where V = an electronic voltmeter measuring the noise voltage
e_n = RMS noise voltage
R = resistance (Ω)
B = bandwidth of the measuring device (electronic voltmeter) (Hz.)

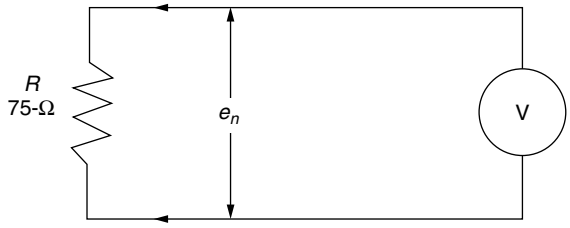


Figure 3. Resistor model for thermal noise voltage, e_n.

k = a constant equal to 40×10^{-16} at standard room temperature.

Letting the bandwidth, B, be equal to that of an NTSC TV signal be rounded to 4 MHz, the open-circuit noise voltage for a 75-Ω resistor is

$$e_n = (4 \times 75 \times 4 \times 10^{-16})^{1/2}$$
$$= 2.2\ \mu\text{V RMS}$$

Figure 4 shows a 2.2 μV noise-generating source (resistor) connected to a 75-Ω (noiseless) load. Only half of the voltage (1.1 μV) is delivered to the load. Thus the noise input to 75 Ω is 1.1 μV RMS or −59 dBmV. This is the basic noise level, the minimum that will exist in any part of a 75-Ω CATV system. The value, −59 dBmV, will be used repeatedly below.

The noise figure of typical CATV amplifiers ranges between 7 and 9 dB [4].

3.4. Signal-To-Noise Ratio (S/N) Versus Carrier-To-Noise (C/N) Ratio in CATV Systems

S/N (signal-to-noise ratio) and C/N (carrier-to-noise ratio) are familiar parameters in telecommunication transmission systems. In CATV systems S/N has a slightly different definition: *This relationship is expressed by the "signal-to-noise ratio," which is the difference between the signal level measured in dBmV, and the noise level, also measured in dBmV, both levels being measured at the same point in the system* [3].

S/N can be related to C/N on CATV systems as

$$\text{C/N} = \text{S/N} + 4.1\ \text{dB} \tag{4}$$

This is based on work by Carson [5], where the basis is "noise just perceptible" by a population of TV viewers, NTSC 4.2-MHz bandwidth TV signal. Here the S/N is 39 dB and the C/N is 43 dB.

Adding noise weighting improvement (6.8 dB), we obtain

$$\text{S/N} = \text{C/N} + 2.7\ \text{dB} \tag{5}$$

It should be noted that S/N is measured where the signal level is peak to peak and the noise level is RMS. For C/N, both the carrier and noise levels are rms. These values are based on a VSB-AM TV signal with 87.5% modulation index.

For comparison, consider another series of tests conducted by the Television Allocations Study Organization (TASO) and published in their report to the U.S. FCC in 1959. Their ratings, corrected for a 4-MHz bandwidth, instead of the 6 MHz bandwidth they used, are shown below:

TASO picture rating	S/N ratio (dB)
1. Excellent, no perceptible noise	45
2. Fine (snow just perceptible)	35
3. Passable (snow definitely perceptible, but not objectionable)	29
4. Marginal (snow somewhat objectionable)	25

Once a tolerable noise level is determined, the levels required in a CATV system can be specified. If the desired S/N has been set at 43 dB at a subscriber TV set, the minimum signal level required at the first amplifier would be −59 dBmV + 43 dB or −16 dBmV, considering thermal noise only. Actual levels would be quite a bit higher because of the noise generated by subsequent amplifiers in cascade.

It has been found that the optimum gain of a CATV amplifier is about 22 dB. When the gain is increased, intermodulation/cross-modulation products become excessive. For gains below this value, thermal noise increases, and system length is shortened or the number of amplifiers must be increased, neither of which is desirable.

There is another rule-of-thumb we should be cognizant of. Every time the gain of an amplifier is increased 1 dB, intermodulation products and "beats" increase their levels by 2 dB. And the converse is true; every time gain is decreased 1 dB, IM products and beat levels are decreased by 2 dB.

With most CATV systems, coaxial cable trunk amplifiers are identical. This, of course, eases noise calculations. We can calculate the noise level at the output of one trunk amplifier. This is

$$N_v = -59\ \text{dBmV} + \text{NF}_{\text{dB}} \tag{6}$$

where NF is the noise figure of the amplifier in decibels.

In the case of two amplifiers in cascade (tandem), the noise level (voltage) is

$$N_v = -59\ \text{dBmV} + \text{NF}_{\text{dB}} + 3\ \text{dB} \tag{7}$$

If we have M identical amplifiers in cascade, the noise level (voltage) at the output of the last amplifier is:

$$N_v = -59\ \text{dBmV} + \text{NF}_{\text{dB}} + 10 \log M \tag{8}$$

This assumes that all system noise is generated by the amplifiers, and none is generated by the intervening sections of coaxial cable (3).

Example 1. A CATV system has 30 amplifiers in tandem, and each amplifier has a noise figure of 7 dB. Assume that the input of the first amplifier is terminated in 75 Ω resistive. What is the thermal noise level (voltage) at the last amplifier output?

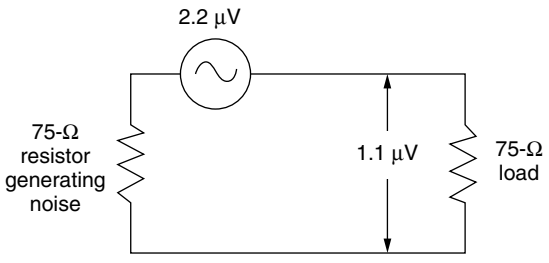


Figure 4. Minimum noise model.

Use Eq. (6):

$$N_v = -59 \text{ dBmV} + 7 \text{ dB} + 10 \log 30$$
$$= -59 \text{ dBmV} + 7 \text{ dB} + 14.77 \text{ dB}$$
$$= -37.23 \text{ dBmV}$$

For carrier-to-noise ratio (C/N) calculations, we can use the following procedures. To calculate the C/N at the output of one amplifier:

$$\text{C/N} = 59 - \text{NF}_{\text{dB}} + \text{input level (dBmV)} \quad (9)$$

Example 2. If the input level of a CATV amplifier were +5 dBmV and its noise figure were 7 dB, what would the C/N a the amplifier output be?

Use Eq. (9):

$$\text{C/N} = 59 - 7 \text{ dB} + 5 \text{ dBmV}$$
$$= 57 \text{ dB}$$

With N cascaded amplifiers, we can calculate the C/N at the output of the last amplifier, assuming all the amplifiers were identical, by the following equation:

$$\text{C/N}_L = \text{C/N(single amplifier)} - 10 \log N \quad (10)$$

Example 3. Determine the C/N at the output of the last amplifier with a cascade (in tandem) of 20 amplifiers, where the C/N of a single amplifier is 62 dB.

Use Eq. (10):

$$\text{C/N}_L = 62 \text{ dB} - 10 \log 20$$
$$= 62 \text{ dB} - 13.0 \text{ dB}$$
$$= 49 \text{ dB}$$

Another variation for calculating C/N is when we have disparate C/N ratios in different parts of a CATV system. CATV systems are usually of a tree topology where the headend is the base of the tree. There is a trunk branching to limbs and further branching out to leaf stems supporting many leaves. In the case of CATV, there is the trunk network, bridger, and line extenders with different gains, and possibly different noise figures. To solve this problem we must use a relationship of several C/N values in series:

$$\text{C/N}_{\text{sys}} = 1/[1/\text{C/N}_1) + (1/\text{C/N}_2) - - - + (1/\text{C/N}_n)] \quad (11)$$

Example 4. Out of a cascade of 20 trunk amplifiers the C/N was 49 dB; out of a bridger, 63 dB and two line extenders, 61 dB, calculate the C/N at the end of the system described.

Use Eq. (11), but first convert each decibel value to a value in decimal units and then take its inverse ($1/X$):

$$49 \text{ dB:antilog}(\tfrac{49}{10}) = 79,433\tfrac{1}{X} = 12,589 \times 10^{-9}$$
$$63 \text{ dB:antilog}(\tfrac{63}{10}) = 2,000,000\tfrac{1}{X} = 500 \times 10^{-9}$$
$$61 \text{ dB:antilog}(\tfrac{61}{10}) = 1,258,925\tfrac{1}{X} = 794 \times 10^{-9}$$

We can now sum the inverses because they all have the same exponent. Sum $= 13,883 \times 10^{-9}$. Take the inverse: 72,030. $10 \log_{(72,030)} = 48.57$ dB. Remember that the "sum" of C/N series values must be something less than the worst value of the series. That is a way of self-checking.

3.5. The Problem of Cross-Modulation (Xm)

Many specifications for TV picture quality are based on the judgment of a population of viewers. One example was the TASO ratings for picture quality given above. In the case of cross-modulation (X-mod or Xm) and CTB (composite triple beat), acceptable levels are −51 dB for Xm and −52 dB for CTB. These are good guideline values [6].

Cross-modulation is a form of third-order distortion so typical of a broadband, multicarrier system. Xm varies with the operating level of an amplifier in question and the number of TV channels being transported. Xm is derived from the amplifier manufacturer specifications. The manufacturer will specify a value for Xm (in dB) for several numbers of channels and for a particular level. The level in the specification may not be the operating level of a particular system. To calculate Xm for an amplifier to be used in given system, using manufacturer's specifications, the following formula applies:

$$\text{Xm}_a = \text{Xm}_{(\text{spec})} + 2(\text{OL}_{\text{oper}} - \text{OL}_{\text{spec}}) \quad (12)$$

where Xm_a = Xm for the amplifier in question
$\text{Xm}_{(\text{spec})}$ = Xm specified by the manufacturer of the amplifier
OL_{oper} = desired operating output signal level (dBmV)
OL_{spec} = manufacturer's specified output signal level

We spot the "2" multiplying factor and relate it to our earlier comments, namely, increase the operating level 1 dB, third-order products increase 2 dB, and the contrary for reducing signal level. And as we said, Xm is a form of third-order product.

Example 5. Suppose a manufacturer tells us that for an Xm of −57 dB for a 35-channel system, the operating level should be +50.5 dBmV. We want a longer system and use an operating level of +45 dBmV, what Xm can we expect under these conditions.

Use Eq. (12):

$$\text{Xm}_a = -57 \text{ dB} + 2(+45 \text{ dBmV} - 50.5 \text{ dBmV})$$
$$\text{Xm}_a = -68 \text{ dB}$$

CATV trunk systems have numerous identical amplifiers. To calculate Xm for N amplifiers in cascade (tandem), our approach is similar to that of thermal noise:

$$\text{Xm}_{\text{sys}} = \text{Xm}_a + 20 \log N \quad (13)$$

where N is the number of identical amplifiers in cascade, Xm_a is the Xm for one amplifier, and Xm_{sys} is the Xm value at the end of the cascade.

Example 6. A certain CATV trunk system has 23 amplifiers in cascade where Xm_a is -88 dB, what is Xm_{sys}?

Use Eq. (13):

$$\begin{aligned} Xm_{sys} &= -88 \text{ dB} + 20 \log 23 \\ &= -88 + 27 \\ &= -61 \text{ dB} \end{aligned}$$

To combine unequal Xm values, we turn to a technique similar to Eq. (11), but now, because we are in the voltage domain, we must divide through by 20 rather than 10 when converting logarithms to equivalent numerics.

Example 7. At the downstream end of our trunk system the Xm was -58 dB and taking the bridger/line extender system alone, their combined Xm is -56 dB. Now, from the headend through the trunk and bridger/line extender system, what is the Xm_{sys}? Convert -56 and -58 dB to their equivalent decimal numerics and invert ($1/X$):

$$\begin{aligned} -56 \text{ dB:antilog} - \tfrac{56}{20} &= 0.001584 \\ -58 \text{ dB:antilog} - \tfrac{58}{20} &= 0.001259 \\ \text{Sum} &: 0.002843 \end{aligned}$$

Take 20 log this value.

$$Xm_{sys} = -50.9 \text{ dB}$$

3.6. Gains and Levels for CATV Amplifiers (3)

Setting both gain and level settings for CATV broadband amplifiers is like walking a tightrope. If levels are set too low, thermal noise will limit system length (i.e., number of amplifiers in cascade). If the level is set too high, system length will be limited by excessive CTB and cross-modulation (Xm). On trunk amplifiers available gain is between 22 and 26 dB [6]. Feeder amplifiers will usually operate at higher gains, trunk systems at lower gains. Feeder amplifiers usually operate in the range of 26–32 dB gain with output levels in the range of +47 dBmV. Trunk amplifiers have gains of 21–23 dB, with output levels in the range of +32 dBmV. If we wish to extend the length of the trunk plant, we should turn to using lower loss cable. employing fiberoptics in the trunk plant is even a better alternative (see Section 4).

The gains and levels of feeder systems are purposefully higher. This is the part of the system serving customers through taps. These taps are passive and draw power. Running the feeder system at higher levels improves tap efficiency. Because feeder amplifiers run at higher gain and with higher levels, the number of these amplifiers in cascade must be severely limited to meet CTB and cross-modulation requirements at the end user.

3.7. The Underlying Coaxial Cable System

The coaxial cable employed in CATV plant is nominally 75 Ω. A typical response curve for such cable ($\frac{7}{8}$-inch, air dielectric) is illustrated in Fig. 5. This frequency response of coaxial cable is called "tilt" in the CATV industry.

For 0.5-inch cable, the loss per 100 ft at 50 MHz is 0.52 dB; for 550 MHz, 1.85 dB. Such cable systems require equalization. The objective is to have a comparatively "flat" frequency response across the entire system. An equalizer is a network that presents a mirror image of the frequency response curve, introducing more loss at the lower frequencies and less loss at the higher frequencies. These equalizers are often incorporated with an amplifier.

Equalizers are usually specified for a certain length of coaxial cable, where length is measured in dB at the highest frequency of interest. Grant [6] describes a 13-dB equalizer for a 300-MHz system, which is a corrective unit for a length of coaxial cable having 13 dB loss at 300 MHz. This would be equivalent to approximately

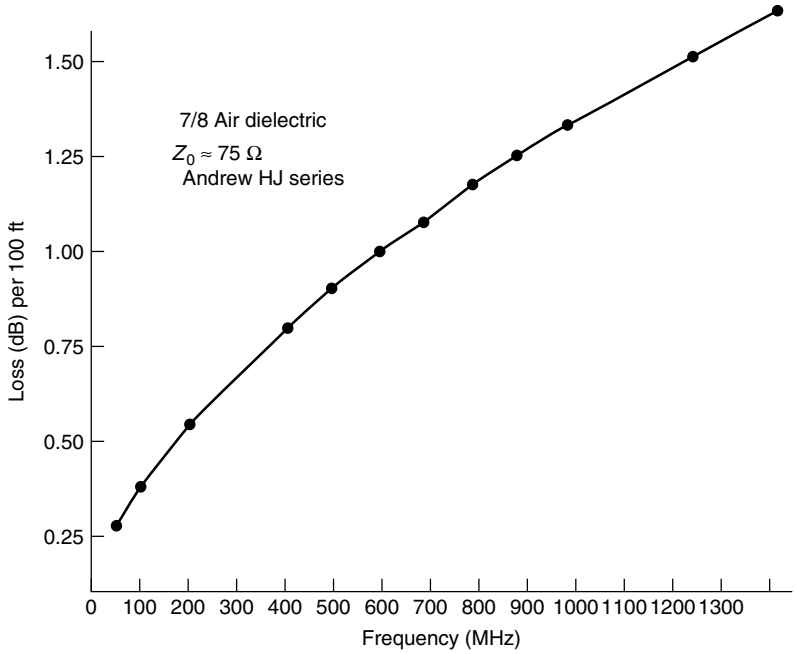


Figure 5. Attenuation–frequency response for $\frac{7}{8}$-in. coaxial cable, air dielectric, $Z_0 = 75$, Andrew HJ series heliax.

1000 ft of $\frac{1}{2}$-inch coaxial cable. Such a length of cable would have 5.45 dB loss at 54 MHz and 13 dB loss at 300 MHz. The equalizer would probably present a loss of 0.5 dB at 300 MHz and 8.1 dB at 54 MHz.

3.8. Taps

A *tap* is similar to a directional coupler. It is a device inserted into a coaxial cable that diverts a predetermined amount of its input energy to one or more tap outputs for the purpose of feeding a TV signal into subscriber drop cables. The remaining balance of the signal energy is passed on down the distribution system to the next tap or distribution amplifier. The concept of the tap and its related distribution system is shown in Fig. 6.

Taps are available to feed 2, 4, or 8 service drops from any one unit. Many different types of taps are available to serve different signal levels that appear along a CATV cable system. Commonly, taps are available in 3-dB increments. For 2-port taps, the following tap losses may be encountered: 4, 8, 11, 14, 17, 20, 23 dB. The insertion loss for the lower value tap loss may be in the order of 2.8 dB, and once the tap loss exceeds 26 dB, the insertion is 0.4 dB and remains so as tap values increase. Another important tap parameter is isolation. Generally, the higher the tap loss, the better the isolation. With 8 dB tap loss, the isolation may only be 23 dB, but with 29 dB tap loss (2-port taps), the isolation can be as high as 44 dB. Isolation in this context is the isolation between the two tap ports to minimize undesired interference from a TV set on one tap to the TV set on the other tap. For example, a line voltage signal level is +34.8 dBmV entering a tap. The tap insertion loss is 0.4 dB, so the level of the signal leaving the tap to the next tap or extender amplifier is +34.4 dBmV. The tap is 2-port. We know we want at least a +10.5 dBmV at the port output. Calculate +34.8 dBmV $-X$dB = +10.5 dBmV. Then $X = 24.3$ dB, which would be the ideal tap loss value. Taps are not available off-the-shelf at that loss value; the nearest value is 23 dB. Thus the output at each tap port will be +34.8 dBmV −23 dB = 11.8 dBmV.

4. HYBRID FIBER-COAX SYSTEMS (HFC)

The following advantages accrue by replacing the coaxial cable trunk system with optical fiber:

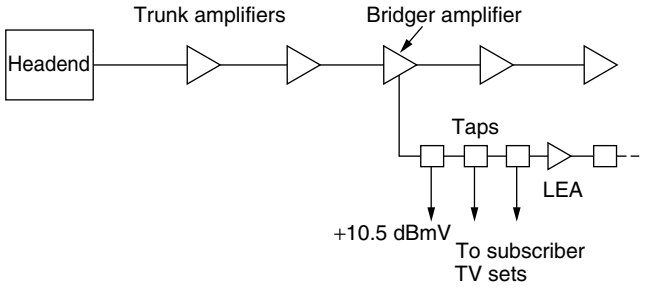


Figure 6. A simplified layout of a CATV system showing its basic elements. The objective is to provide a +10.5-dBmV signal level at the drops (tap outputs) (LEA = line extender amplifier).

- Reduces the number of amplifiers required per unit distance to reach the fartherest subscriber
- Results in improved C/N, reduced CTB and Xm levels
- Also results in improved reliability (i.e., by reducing the number of active components)
- Has the potential to greatly extend a particular CATV serving area.

Figure 7 shows the basic concept of an HFC system. One disadvantage is that a second fiber link has to be installed for reverse direction, or a form of WDM is needed, when two-way operation is required and/or for the CATV management system (used for monitoring the health of the system, amplifier degradation or failure). Figure 8 illustrates an HFC system where there are no more than three amplifiers to a subscriber tap. Note that with this system layout there cannot be a catastrophic failure. For the loss of an amplifier, only $\frac{1}{16}$ of the system is affected, worst case; with the loss of a fiber link, the worst case would be $\frac{1}{6}$ of the system.

4.1. Design of the Fiberoptic Portion of an HFC System

There are two approaches to fiberoptic transmission of analog CATV signals. Both approaches take advantage of the intensity modulation characteristics of the fiberoptic source. Instead of digital modulation of the source, amplitude modulation (analog) is employed. The most common method takes the entire CATV spectrum as it would appear on a coaxial cable and uses that as the modulating signal. The second method also uses analog amplitude modulation, but the modulating signal is a grouping of subcarriers that are each frequency-modulated. One off-the-shelf system multiplexes in a broad FDM configuration, 8 television channels, each on a separate subcarrier. Thus, a 48-channel CATV system would require six fibers, each with eight subcarriers (plus 8 or 16 audio subcarriers).

4.1.1. Link Budget for an AM System. Assume a model using a distributed feedback (DFB) laser with an output of +5 dBm coupled to the pigtail. The receiver is a PINFET, where the threshold is −5 dBm. This threshold will derive approximately 52 dB S/N in a video channel. Compared to digital operation, the C/N is around 49.3 dB, assuming that the S/N value is noise-weighted (see Section 3.4). This is a very large C/N value and leaves only 10 dB to be allocated to fiber, splice loss, and margin. If we allocated 2 dB for the link margin, only 8 dB is left for fiber/splice loss.

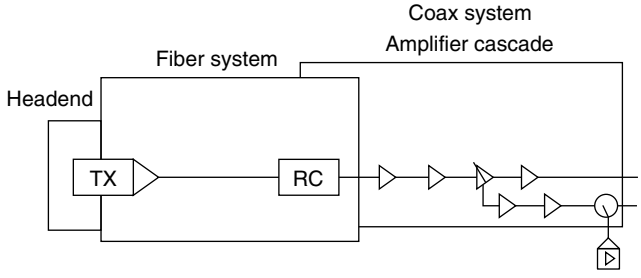


Figure 7. The concept of a hybrid fiber–coaxial cable CATV system (TX = fiberoptic transmitter, RC = fiberoptic receiver).

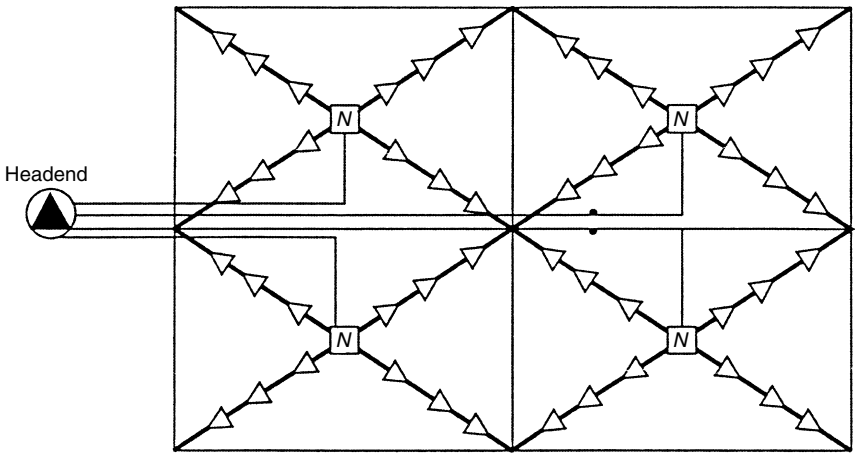


Figure 8. An HFC system layout for optimal performance (one-way). (N = node fiber) interface with coaxial cable).

At 1550-nm operation, assuming a conservative 0.4 dB/km fiber/splice loss, the maximum distance to the coax hub or fiberoptic repeater is only 8/0.4 or 20 km. Of course if we employ a EDFA (erbium-doped fiber amplifier) with, say, only a 20-dB gain, this distance can be extended by 20/0.4 or an additional 50 km. Figure 9 illustrates a typical laser diode transfer characteristic showing the amplitude-modulated input drive.

Representative design goals for the video/TV output of a fiber optics trunk are

$$\text{CNR} = 50 \text{ dB}$$

$$\text{CSO } \textit{products} = -62 \text{ dBc}$$

$$\text{CTB} = -65 \text{ dBc}$$

(where CSO = composite second order). One common technique used on HFC systems is to employ optical couplers where one fiber trunk systems feeds several hubs. A *hub* is a location where the optical signal is converted back to an electrical signal for transmission on coaxial cable. Two applications of optical couplers are illustrated in Fig. 10. Keep in mind that a signal split includes not only splitting the power but also the insertion loss of the coupler. The values shown in parentheses in Fig. 10 give the loss in the split branches (e.g., 5.7 dB and 2.0 dB).

4.1.2. FM Systems. FM systems are more expensive than their AM counterparts, but provide improved performance. EIA-250C, a well-known and respected standard for television transmission, specifies a signal-to-noise ratio of 67 dB for short-haul systems. With AM systems it is impossible to achieve this S/N, whereas a well-designed FM system can conform to EIA-250C. AM systems are degraded by dispersion on the fiber link; FM systems, much less so. FM systems can also be extended further. FM systems are available with 8, 16, or 24 channels, depending on the vendor.

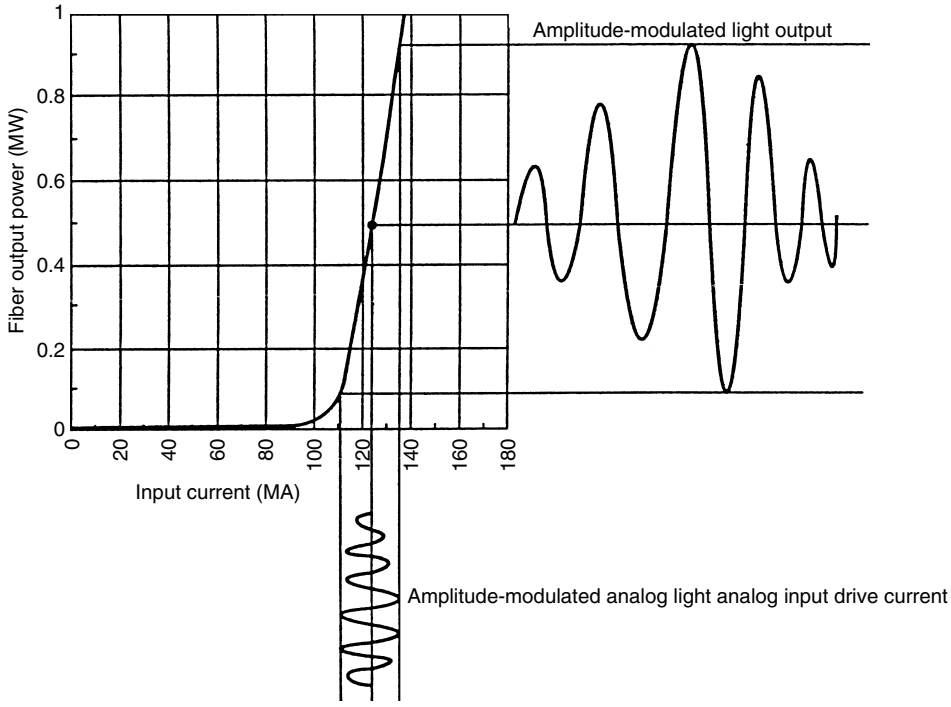


Figure 9. Laser transfer characteristics.

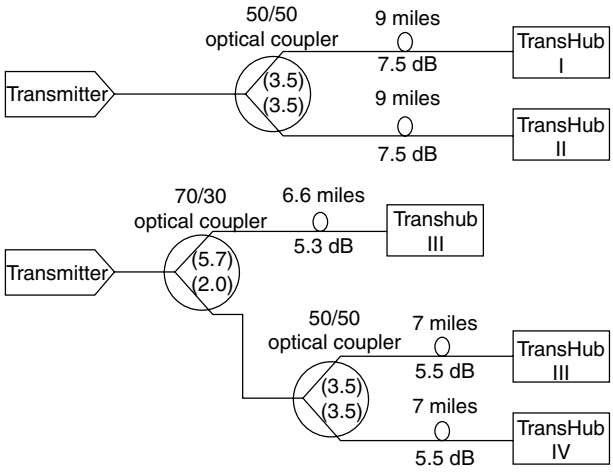


Figure 10. Two-way and three-way splits of a light signal transporting TV.

Figure 11 shows an 8-channel per fiber frequency plan, and Fig. 12 is a transmit block diagram for the video portion of the system. Figure 13 is a layout of a typical fiber hub, and Fig. 14 is a plot of optical receiver input threshold power (dBm) versus signal-to-noise ratio of individual channels derived from an FM HFC system.

Figure 12 is an FM system model. At the headend, each video and audio channel is broken out separately. And as shown in Fig. 11, each channel FM modulates a subcarrier. Note that there is a similar but separate system for the associated audio (aural) channels with 30 MHz spacing starting at 70 MHz and these audio channels may be multiplexed before transmission. Each video carrier occupies a 40 MHz slot. These RF carriers, audio and video, are combined in a passive network. The composite RF signal intensity-modulates a laser diode source. Figure 13 shows a typical fiber/FM hub.

4.1.2.1. Calculation of Video S/N for an FM System. Given the C/N (CNR) for a particular FM system, the S/N of a TV video channel may be calculated as shown in Example 8.

$$\mathrm{SNRw} = K + \mathrm{CNR} + 10\log\frac{B_{\mathrm{IF}}}{B_{\mathrm{F}}} + 20\log\frac{1.6\Delta F}{B_{\mathrm{F}}}$$

where K = a constant ($\sim$23.7 dB) made of weighting network, deemphasis, and rms to p-p conversion factors
CNR = carrier-to-noise ratio in the IF bandwidth

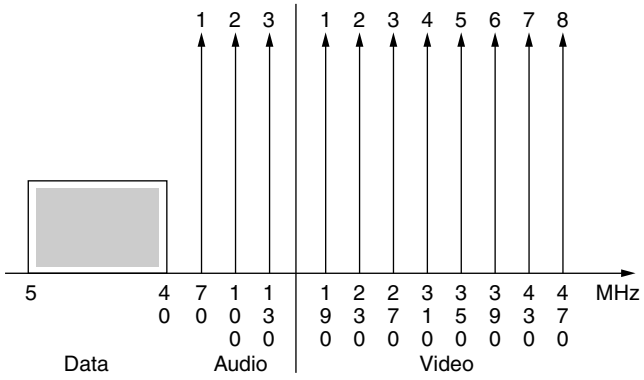


Figure 11. Eight-TV-channel frequency plan for an FM system.

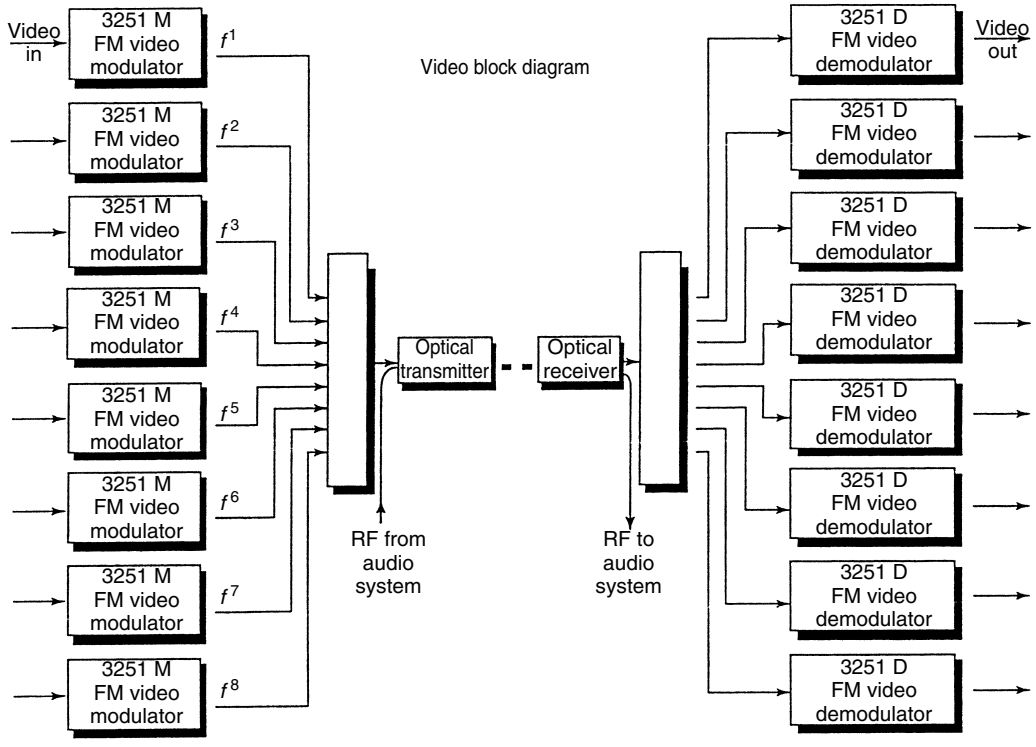


Figure 12. FM system model block diagram for video transmission subsystem. (Courtesy of Catel Corp.)

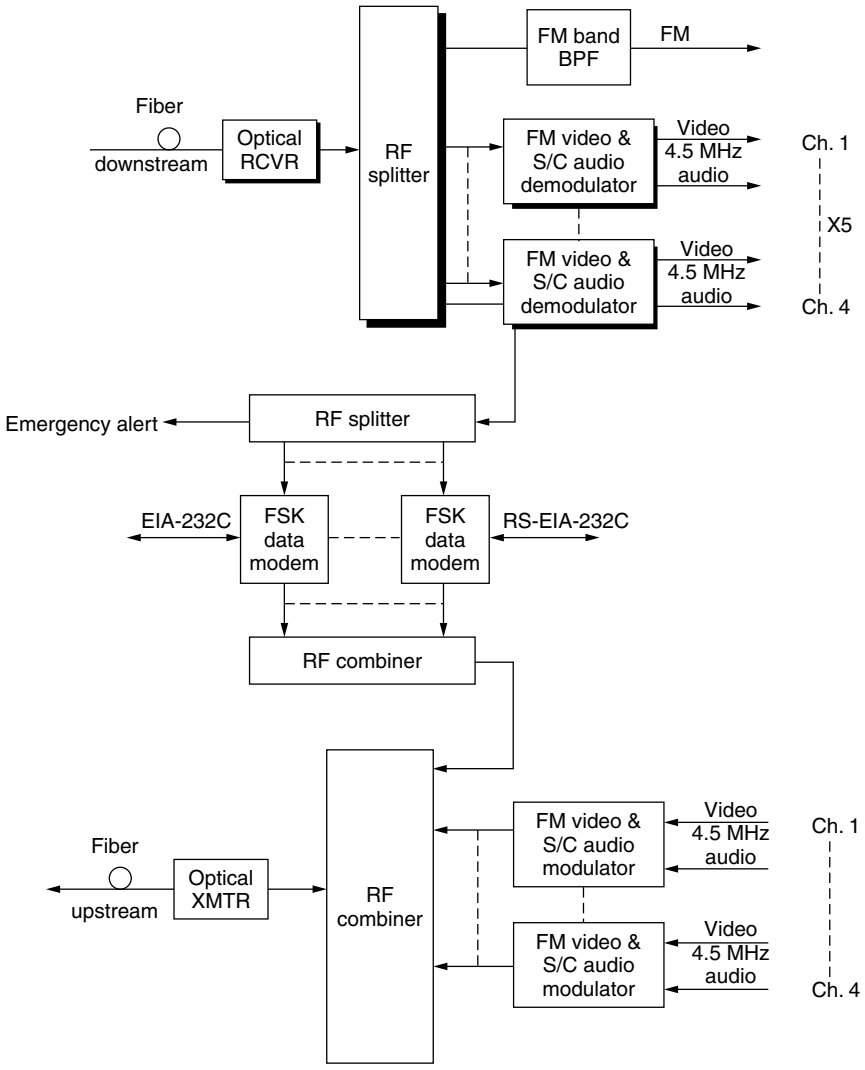


Figure 13. A typical FM/fiber hub.

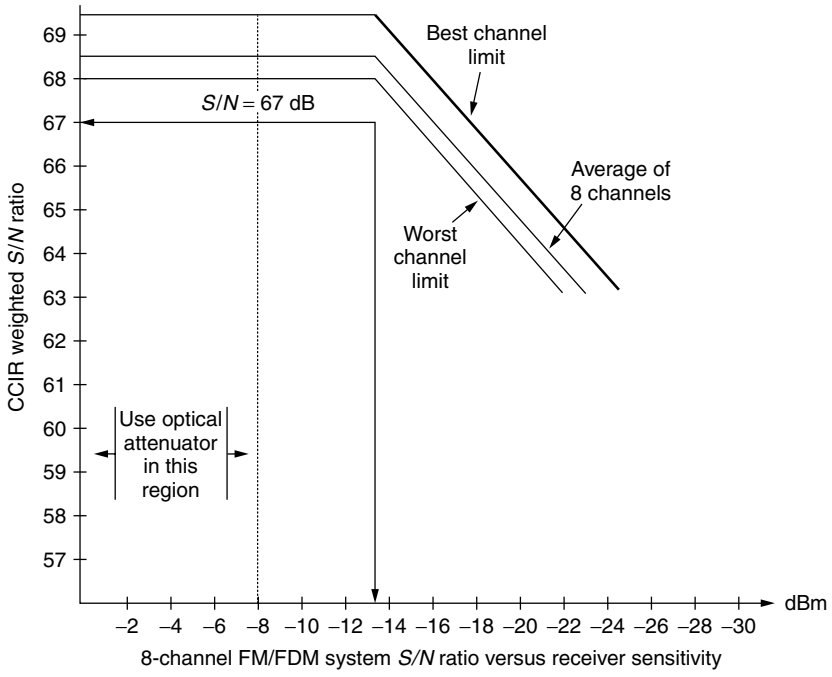


Figure 14. Link performance of an FM system. (Courtesy of Catel Corp.)

B_{IF} = IF bandwidth
B_F = baseband filter bandwidth
ΔF = sync tip-to-peak white (STPW) deviation

With $\Delta F = 4$ MHz, $B_{IF} = 30$ MHz, and $B_F = 5$ MHz, the SNRw is improved by approximately 34 dB above CNR.

Example 8. If the C/N on an FM fiber link is 32 dB, what is the S/N for a TV video channel using the values given above?

Use Eq. (14):

$$\begin{aligned} \text{S/N} &= 23.7\text{ dB} + 32\text{ dB} + 10\log(\tfrac{30}{5}) + 20\log(1.6 \times \tfrac{4}{5}) \\ &= 23.7 + 32 + 7.78 + 2.14 \\ &= 65.62\text{ dB} \end{aligned}$$

Figure 14 illustrates the link performance of an FM fiberoptic system for video channels.

Table 1 shows typical link budgets for an HFC AM system.

5. DIGITAL TRANSMISSION OF CATV

5.1. Approaches

There are two approaches to digitally transmit TV, both audio and video. The first is to transport raw, uncompressed video. The second method is to transport compressed video. Each method has advantages and disadvantages. Some advantages and disadvantages are application-driven. For example, if the objective is digital to the residence or office, compressed TV may be the most advantageous. In either case the ability to regenerate is a distinct advantage.

5.2. Transmission of Uncompressed Video on CATV Trunks

Video is an analog signal. It is converted to a digital format using techniques with some similarity to the 8-bit PCM so widely employed in the PSTN. A major difference is in the sampling. Broadcast quality TV is generally *oversampled*. Here we mean that the sampling rate is greater than the Nyquist rate. The Nyquist rate, as we remember, requires the sampling rate to be twice the highest frequency of interest. In our case this is 4.2 MHz, the bandwidth of a TV signal. Thus, the sampling rate is greater than 8.4×10^6 samples per second.

Typically, the sampling rate is based on the frequency of the color subcarrier. For NTSC television, the color subcarrier is at 3.58 MHz and we call this frequency f_{sc}. In some cases the sampling rate is set at three times this frequency ($3f_{sc}$) and in other cases four times the sampling rate ($4f_{sc}$). Thus, for NTSC color television, the sampling rate for the A/D converter is either $3 \times 3.58\text{ MHz} = 10.74 \times 10^6\text{ s}^{-1}$ or $4 \times 3.58\text{ MHz} = 14.32 \times 10^6$/s. For PAL television, the color subcarrier is 4.43 MHz and the sampling rate then may be $3 \times 4.43 = 13.29 \times 10^6$ MHz or $4 \times 4.43\text{ MHz} = 17.72 \times 10^6\text{ s}^{-1}$.

A major advantage of digital transmission is the regeneration capability just as it is in PSTN 8-bit PCM. As a result, there is no noise accumulation on the digital portion of the network. These digital trunks can be extended hundreds or more miles. The complexity is only marginally greater than an FM system. The 10-bit system can easily provide an S/N at the conversion hub of 67 dB in a video channel and an S/N of 63 dB with an 8-bit system. With uncompressed video, BER requirements are not very stringent because video contains highly redundant information.

5.3. Compressed Video

MPEG types of compression are widely used today. A common line bit rate for MPEG[2] is 1.544 Mbps. Allowing 1 bit per Hz of bandwidth, BPSK modulation and a cosine rolloff of 1.4, the 1.544 TV signal can be effectively transported in a 2 MHz bandwidth. Certainly 1000 MHz coaxial cable systems are within the state of the art. With simple division we can see that 500-channel CATV systems are technically viable. If the modulation scheme utilizes 16-QAM (4 bits per Hz theoretical), three 1.544 Mbps compressed channels can be accommodated in a 6-MHz slot. We select 6 MHz because it is the current RF bandwidth assigned for one NTSC TV channel.

6. TWO-WAY CATV SYSTEMS

6.1. Introduction

Panels (a) and (b) of Fig. 15 are two views of the CATV spectrum as it would appear on coaxial cable. Of course, with conventional CATV systems, each NTSC television channel is assigned a 6 MHz slot just as it is done with conventional broadcast television. In Fig. 15a, only 25 MHz is assigned for upstream services. Not all of this bandwidth

[2] MPEG = Motion Picture Experts' Group, a standardization agency, most know for motion picture (television) compression schemes.

Table 1. Typical Link Budgets for an AM Fiber Link

Distance (mi)	Distance (km)	Fiber Loss/km	Total Fiber Loss	Splice Loss/2 km	Total Splice Loss	Total Path Loss	Link Budget	Link Margin
Mileage, Losses, and Margins — 1310 nm								
12.40	19.96	0.5 dB	9.98	0.1 dB	1.00	10.98	13.00	2.02
15.15	24.38	0.4 dB	9.75	0.1 dB	1.22	10.97	13.00	2.03
17.00	27.36	0.35 dB	9.58	0.1 dB	1.37	10.94	13.00	2.06
Mileage, Losses, and Margins — 1550 nm								
22.75	36.61	0.25 dB	9.15	0.1 dB	1.83	10.98	13.00	2.02

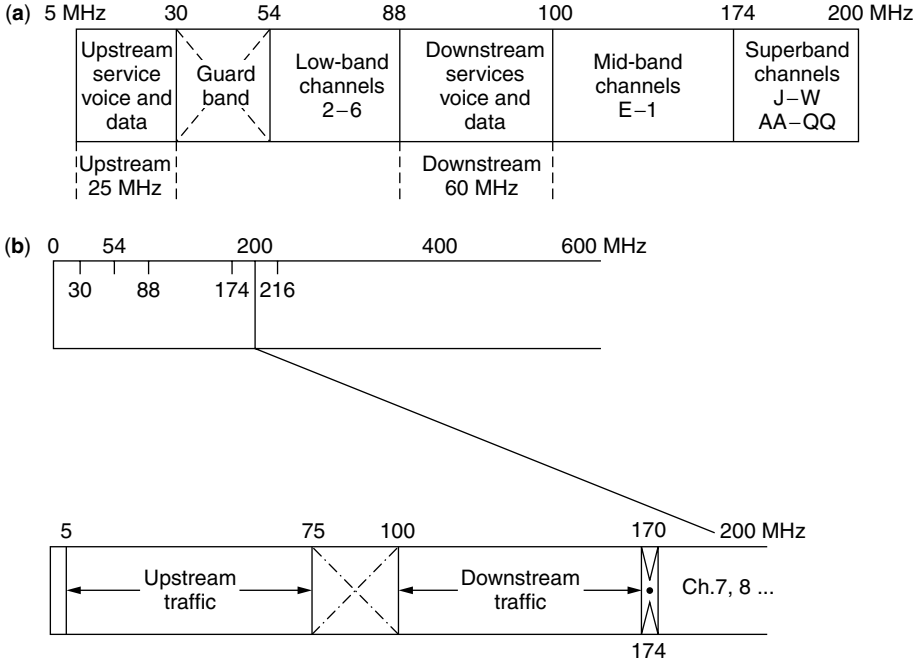


Figure 15. (**a**) CATV spectrum based on Grant [6] showing additional upstream and downstream services — note the bandwidth imbalance between upstream and downstream; (**b**) CATV spectrum with equal upstream and downstream bandwidths for other services [part (a) based on Ref. 6; part (b) prepared by the author]. On the other hand, downstream has 60 MHz assigned. In this day of the Internet, this would be providential, for the majority of the traffic would be downstream.

may be used for voice and data. A small portion should be set aside for upstream telemetry to monitor active CATV equipment. On the other hand, downstream has 60 MHz assigned. In this day of the internet, this would be providential, for the majority of the traffic would be downstream.

6.2. Comments on Fig. 15

Large guard bands isolate upstream from downstream TV and other services, with 24 MHz in A and 25 MHz in B. A small guardband was placed in the slot from 170 to 174 MHz to isolate downstream data and voice signals from conventional CATV television. We assume that the voice service will be "POTS" (plain old telephone service) and that both the data and voice would be digital.

In another approach, downstream voice, data and special video are assigned the band 550 to 750 MHz, which is the highest frequency segment portion of this system. (Ref. 12) In this case, we are dealing with a 750-MHz system.

The optical fiber trunk terminates in a node or hub. This is where the conversion from optical to the standard CATV coaxial cable format occurs. Let a node serve four groupings of subscribers, each with a coaxial cable with the necessary amplifiers, line extenders, and taps. Such subscriber groups consist of 200–500 terminations (TV sets). Assume that each termination has upstream service using the band 5–30 MHz (Fig. 15a). In our example, the node has four incoming 5–30-MHz bands, one for each coaxial cable termination. It then converts each of these bands to a higher-frequency slot 25 MHz wide in a frequency-division configuration for backhaul on a return fiber. In one scheme, at the headend, each 25-MHz slot is demultiplexed and the data and voice traffic are segregated for switching and processing.

An interesting exercise is to divide 25 MHz by 500. This tells us that we can allot each user 50 kHz full period. By taking advantage of the statistics of calling (usage), we could achieve 4–10 times bandwidth multiplier by using forms of concentration. However, upstream video, depending on the type of compression, might consume a large portion of this spare bandwidth.

There are many other ways for a subscriber can gain access, such as by TDMA, FDMA, and contention. Several protocols use combinations of time division and frequency division based on the concept of the minislot.

6.3. Impairments Peculiar to Upstream Service

6.3.1. More Thermal Noise Upstream than Downstream. Figure 16 shows a hypothetical layout of amplifiers in a CATV distribution system for two-way operation. In the downstream direction, broadband amplifiers point outward, down trunks and out distribution cables. In the upstream direction, the broadband amplifiers point inward

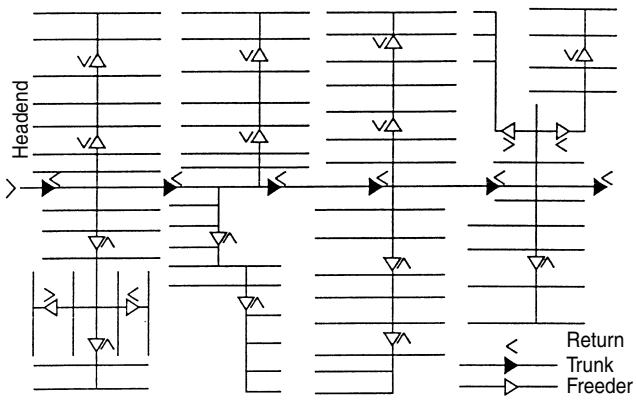


Figure 16. Trunk/feeder system layout for two-way operation. (From Grant [6]. Reprinted with permission.)

toward the headend, and all their thermal noise accumulates and concentrates at the headend. This can account for 3–20 dB additional noise upstream at the headend where the upstream demodulation of voice and data signals takes place. Fortunately, the signal-to-noise ratio requirements for good performance for voice and data are much less stringent than for video, which compensates to a certain extent for this additional noise.

6.3.2. Ingress Noise. This noise source is peculiar to CATV system. It basically derives from the residence/office TV sets that terminate the system. Parts 15.31 and 15.35 of the *FCC Rules and Regulations* govern such unintentional radiators. These rules have not been rigidly enforced.

One problem that the 75-Ω impedance match between the coaxial cable and the TV set is poor. Thus not only all radiating devices in the TV set but other radiating devices nearby in residences and office buildings couple back through the TV set into the CATV system in the upstream direction. This type of noise is predominant in the lower frequencies, that band from 5 to 30 MHz that carries the upstream signals. As frequency increases, ingress noise intensity decreases. Fiberoptic links in an HFC configuration provide some isolation, but it still can be a major problem.

6.4. Data Over Cable Service Interface Specification (DOCSIS)

DOCSIS is a complete specification for transmitting data across a cable television system. The intended service allows transparent bidirectional transfer of IP[5] traffic between the headend and customer facilities, over an all-coaxial or hybrid fiber/coax (HFC) cable network. The concept is illustrated in Fig. 17.

The specification for the cable modem (CMTS) is described in detail in DOCSIS [9,10]. A brief overview of the modulation and coding is given in the following section, which treats digital video transmission.

7. DIGITAL VIDEO TRANSMISSION STANDARD FOR CABLE TELEVISION

Based on document ANSI/SCTE 07 2000, issued Oct. 25, 1996 [9] Courtesy of Dr. Ted Woo, Technical Director, SCTE.

7.1. Introduction

This section describes the framing structure, channel coding, and channel modulation for a digital multiservice television distribution system that is specific to a cable channel. The system can be used transparently with the distribution from a satellite channel, in that many cable systems are fed directly from satellite links. The specification covers both 64- and 256-QAM waveforms. Most features of the two modulation schemes are the same. Where there are differences, the specific details of each modulation scheme are covered in the DOCSIS specification.

The design of the modulation, interleaving and coding is based on test and characterization of cable systems in North America. The modulation is quadrature amplitude modulation (QAM) with a 64-point signal constellation (64-QAM) or with a 256-point signal constellation (256-QAM), transmitter selectable. The forward error correction (FEC) is based on a concatenated coding approach that produces high coding gain at moderate complexity and overhead. Concatenated coding offers improved performance over a block code, at a similar overall complexity. The system FEC is optimized for quasi-error-free operation at a threshold output error event rate of one error event per 15 minutes.

The data format input to the modulation and coding is assumed to be MPEG-2[3] transport. However, the method used for MPEG synchronization is decoupled from the FEC synchronization. For example, this enables the system to carry asynchronous transfer mode (ATM) packets without interfering with ATM synchronization. In fact, ATM synchronization may be performed by defined ATM synchronization mechanisms.

Two modes are supported by this standard. Mode 1 has a symbol rate of 5.056 Msps, and mode 2 has a symbol rate of 5.361 Msps (megasymbols per second). Typically, mode 1 is used for 64-QAM and mode 2 is used for 256-QAM. The system is compatible with future implementations of higher-data-rate schemes employing higher order QAM extensions.

7.2. Cable System Concept

Channel coding and transmission are specific to a particular medium or communication channel. The expected channel error statistics and distortion characteristics are

[5] IP-Internet Protocol

[3] MPEG — Motion Picture Experts Group.

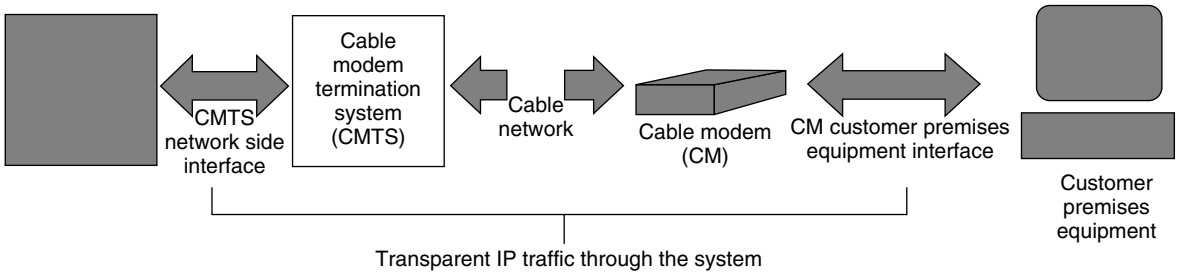


Figure 17. Transparent IP traffic through the data-over-cable system (from DOCSIS [10] Fig. 1.1, p. 2).

critical in determining the appropriate error detection and demodulation. The cable channel, including optical fiber, is regarded primarily as a bandwidth-limited linear channel with a balanced combination of white noise, interference, and multipath distortion. The quadrature amplitude modulation (QAM) technique, together with adaptive equalization and concatenated coding, is well suited to this application and channel.

The basic layered block diagram of cable transmission processing is shown in Fig. 18. The following subsections define these layers from the "outside" in, and from the perspective of the transmit side.

7.3. MPEG-2 Transport Framing

The transport layer for MPEG-2 data is composed of packets having 188 bytes, with 1 byte for synchronization purposes, 3 bytes of header containing service identification, and scrambling and control information, followed by 184 bytes of MPEG-2 or auxiliary data.

The MPEG transport framing is the outermost layer of processing. It is provided as a robust means of delivering MPEG packet synchronization to the receiver output. This processing block receives an MPEG-2 transport datastream consisting of a continuous stream of fixed-length 188-byte packets. This datastream is transmitted in serial fashion, MSB (most significant bit) first. The first byte of a packet is specified to be a sync byte having a constant value of 47_{HEX}.

7.4. Forward Error Correction

The forward error correction (FEC) definition is composed of four processing layers, as illustrated in Fig. 19. There are no dependencies on input data protocol in any of the FEC layers. FEC synchronization is fully internal and

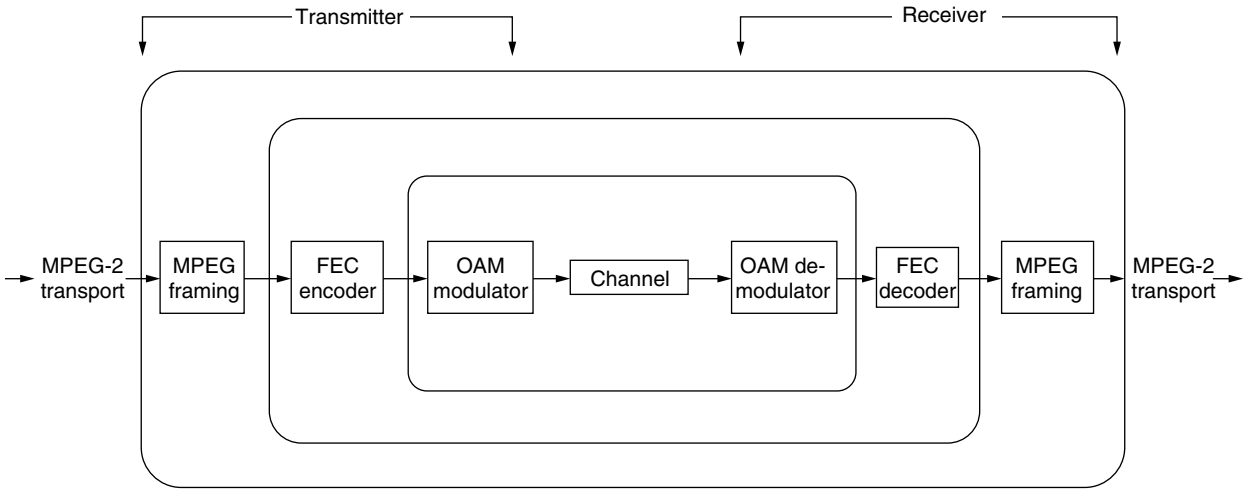


Figure 18. Cable transmission block diagram.

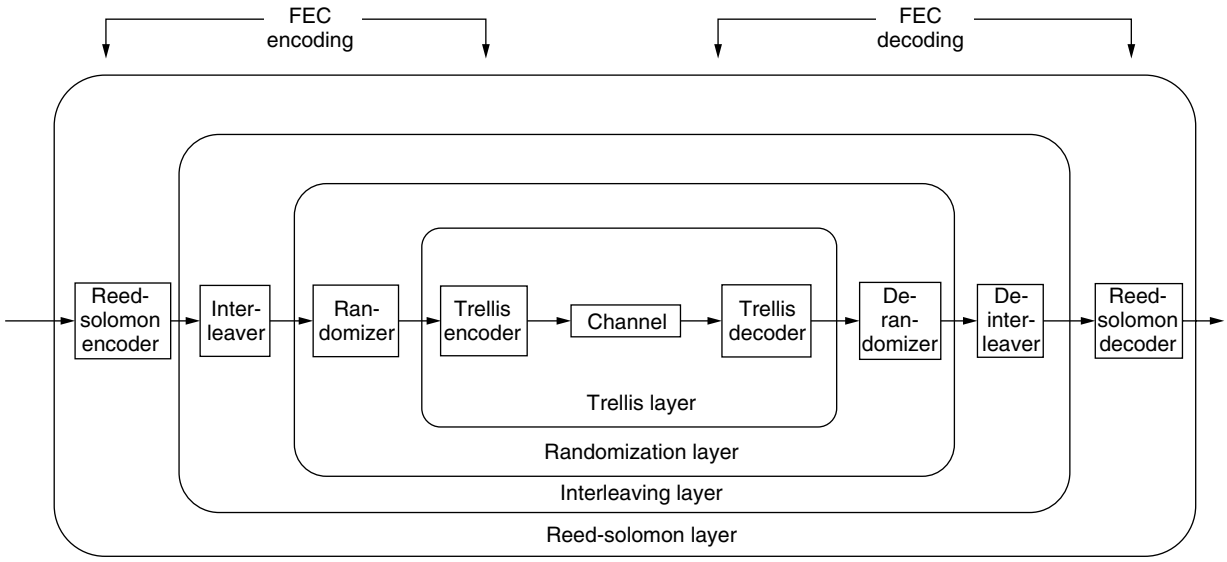


Figure 19. Layers of processing in the FEC.

transparent. Any data sequence will be delivered from the encoder input to the decoder output.

The FEC section uses various types of error-correcting algorithms and interleaving techniques to transport data reliably over the cable channel.

Reed–Solomon (RS) coding — provides block encoding and decoding to correct up to three symbols within an RS block

Interleaving — evenly disperses the symbols, protecting against a burst of symbol errors from being sent to the RS decoder

Randomization — randomizes the data on the channel to allow effective QAM demodulator synchronization

Trellis coding — provides convolutional encoding and with the possibility of using soft-decision trellis decoding of random channel errors

The following subsections define these four layers.

7.4.1. Reed–Solomon Coding. The MPEG-2 transport stream is Reed–Solomon (RS) encoded using a (128, 122) code over GF(128). This code has the capability of correcting up to $t = 3$ symbol errors per R–S block. The same R–S code is used for both 64-QAM and 256-QAM.

7.4.2. Interleaving. Interleaving is included in the modem between the RS block coding and the randomizer to enable the correction of burst noise induced errors. In both 64-QAM and 256-QAM, a convolutional interleaver is employed. The interleaver consists of a single fixed structure for 64-QAM, along with a programmable structure for 256-QAM.

7.4.3. Randomization. The randomizer is the third layer of processing in the FEC block diagram. The randomizer provides for even distribution of the symbols in the constellation, which enables the demmodulator to maintain proper lock. The randomizer adds a pseudorandom noise (PN) sequence of 7 bits over Galois Field GF(128) (i.e., bitwise exclusive-OR) to the symbols within the FEC frame to assure a random transmitted sequence.

7.4.4. Trellis-Coded Modulation. As part of the concatenated coding scheme, trellis coding is employed for the inner code. It allows introduction of redundancy to improve the threshold signal-to-noise ratio (SNR) by increasing the symbol constellation without increasing the symbol rate. As such, it is more properly termed *trellis-coded modulation*.

7.4.5. 64-QAM Modulation. For 64-QAM, the input to the trellis-coded modulator is a 28-bit sequence of four, 7-bit RS symbols, which are labeled in pairs of "A" and "B" symbols. A block diagram of a 64-QAM trellis-coded modulator is shown in Fig. 20. All 28 bits are assigned to a trellis group, where each trellis group forms 5 QAM symbols.

Of the 28 input bits that form a trellis group, each of two groups of 4 bits of the differentially precoded bitstreams in a trellis group are separately encoded by a binary convolutional coder.

The differential precoder allows the information to be carried by the change in phase, rather than by the absolute phase. For 64-QAM the third and sixth bits of the 6-bit symbols are differentially encoded, and for 256-QAM the fourth and eighth bits are differentially encoded.

7.4.6. Binary Convolutional Coder. The trellis-coded modulator includes a punctured rate-$\frac{1}{2}$ binary convolutional encode that is used to introduce the redundancy into the LSBs (least significant bits) of the trellis group. The convolutional encoder is a 16-state nonsystematic rate-$\frac{1}{2}$ encoder with the generator: G1 =

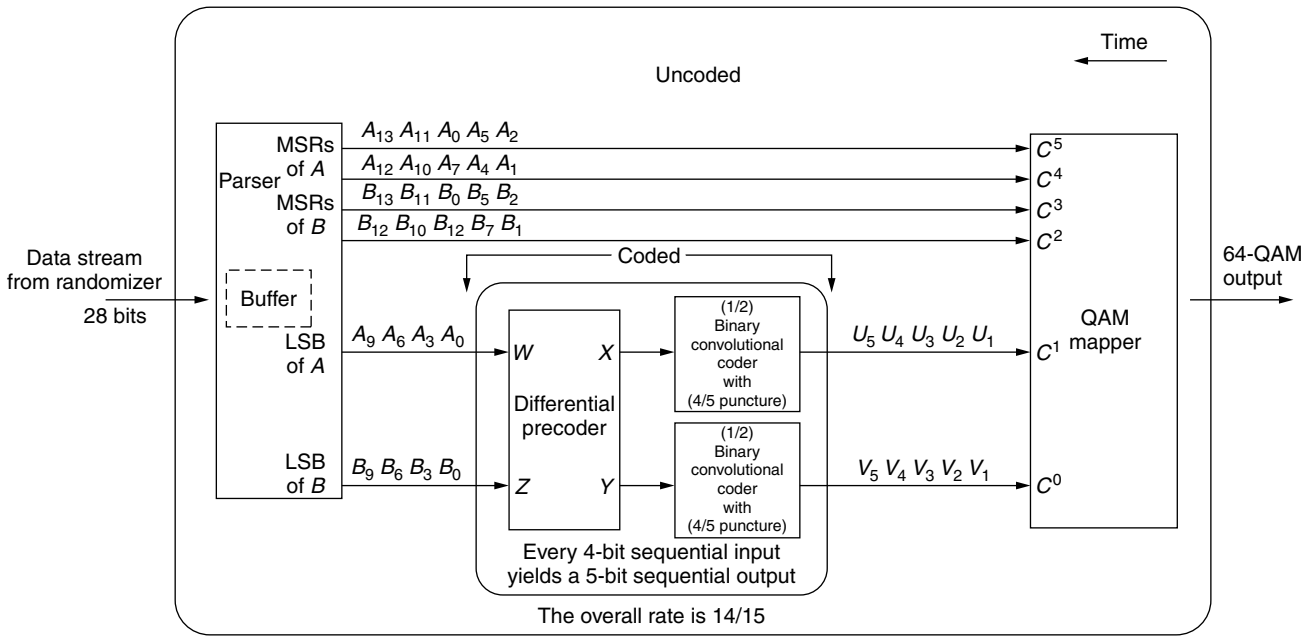


Figure 20. 64-QAM trellis-coded modulator block diagram.

010101, G2 = 011111(15, 37_{octal}), or equivalently the generator matrix[$1 \oplus D^2 \oplus D^4$, $1 \oplus D \oplus D^2 \oplus D^3 \oplus D^4$].

8. CONCLUSION

This concludes our brief description of the digital video transmission standard for CATV. A detailed description of this standard is contained in Ref. 9.

BIOGRAPHY

Roger Freeman has over 50 years experience in telecommunications including a stint in the US Navy and radio officer on merchant vessels. He attended Middlebury College and has two degrees from New York University. He has had assignments with the Bendix Corporation in Spain and North Africa which was followed by five years as a member technical staff for ITT Communications Systems. Roger then became manager of microwave systems for CATV extension at Jerrold Electronics Corporation followed by assignments at Page Communications Engineers in Washington, DC where he was a project engineer on earth stations and on various data communication programs. During this period he was assigned by the ITU as Regional Planning Expert for northern Sourth America based in Quito, Ecuador. From Quito he took a position with ITT at their subsidiary in Madrid, Spain where he did consulting in telecommunication planning. In 1978 he joined the Raytheon Company as principal engineer in their Communication Systems Directorate where he held design positions on military communications such as on the AN/TRC-170, MILSTAR, AN/ASC-30 and on wideband HF. At the same time he taught various telecommunication courses in the evenings at Northeastern University and 4-day seminars at the University of Wisconsin. These seminars were based on his several textbooks on telecommunications published by John Wiley & Sons, New York. He also gives telecommunication seminars (in Spanish) in Monterrey, Mexico City and Caracas. Roger is a contributor and guest editor (Desert Storm edition) of the IEEE Communications magazine and was advanced by the IEEE to senior life member in 1994. He served on the board of directors of the Spain Section of the IEEE and was its secretary for four years. In 1991 Roger took early retirement from the Raytheon Company and organized Roger Freeman Associates, Independent Consultants in Telecommunications. The group has undertaken over 50 assignments from Alaska to South America.

Roger may be reached at *rogerf67@cox.net*; his website is *www.rogerfreeman.com*. Also of interest would be *www.telecommunicationbooks.com* where the reader may subscribe to the on-line Reference Manual for Telecommunication Engineering, 3rd ed, updated quarterly.

BIBLIOGRAPHY

1. *How to Characterize CATV Amplifiers Effectively*, Application Note 1288–4, Hewlett-Packard Co., Palo Alto, CA, 1997.
2. R. L. Freeman, *Telecommunication Transmission Handbook*, 4th ed., Wiley, New York, 1998.
3. K. Simons, *Technical Handbook for CATV Systems*, 3rd ed, Jerrold Electronics Corp., Hatboro, PA, 1968.
4. E. R. Bartlett, *Cable Television Technology and Operations*, McGraw-Hill, New York, 1990.
5. D. N. Carson, CATV amplifiers: figure of merit and coefficient system, 1966 *IEEE International Convention Record,* Part I, *Wire and Data Communications*, IEEE, New York, March 1966, pp. 87–97.
6. W. O. Grant, *Cable Television*, 3rd ed., GWG Associates, Schoharie, NY, 1994.
7. *System Planning, Product Specifications and Services*, Catalog no. 36, Andrew Corp., Orland Park, IL, 1994.
8. *Cable Television Channel Identification Plan*, EIA-542, Electronic Industries Alliance, Washington, DC, April 1997.
9. *Digital video transmission standard for cable television*, ANSI/SCTE 07 2000, Society of Cable Telecommunications Engineers, Exton, PA, Oct. 1996.
10. *Data over Cable Service Interface Specification*, subtitled *Radio Frequency Interface Specification*, SP-RFI-103-980202, SCTE 22, Part 1, SCTE, Exton, PA, 1998.
11. R. L. Freeman, *Reference Manual for Telecommunications Engineering*, 3rd ed., Wiley, New York, 2002.
12. *Buyers Guide*, Lightwave, CMP Publications, Nashua, NY, Oct. 2000.

COMPANDERS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

The word *compander* comes from the words *comp*ressor and exp*ander*. Companders are widely used in communication systems to compress the dynamic range of an information-bearing signal, such as a speech signal, prior to digitizing the signal. In the transmission of analog signal waveforms digitally, the analog signal is usually sampled at some nominal rate that exceeds the Nyquist rate in order to avoid aliasing of frequency components. For example, in digital transmission of speech signals, the analog speech waveform at the transmitter is lowpass filtered to some nominal bandwidth, say, 3.5 kHz, and then sampled at a rate of 8 kHz. Each sample is quantized to one of a set of quantization levels and, then, represented by a sequence of bits that are transmitted via binary modulation to the receiver. The receiver reconstructs the quantized values of the samples from the received binary sequence and synthesizes the analog signal by passing the sampled values through a digital-to-analog converter. In such a system, the compressor is used at the transmitter to compress the dynamic range of the signal samples being quantized and the reverse process is performed at the receiver, where the dynamic range of the compressed signal values is inversely expanded.

Companding will be described below in the context of pulse code modulation (PCM), which is widely used to digitally encode analog signal waveforms.

2. PULSE CODE MODULATION

Pulse code modulation (PCM) is a very simple method for converting an analog signal waveform into a sequence of binary digits. The operations performed at the encoder of a PCM system are illustrated by the functional block diagram shown in Fig. 1.

The sampling is performed at a rate higher than the Nyquist rate to avoid aliasing of high-frequency signal components. The compression of the signal dynamic range is embedded in the quantizer. If the quantizer is selected to be uniform, then no signal compression is performed. On the other hand, if a nonuniform quantizer is selected that maps a signal sample x_n into a value $g(x_n)$, where $g(x_n)$ is some nonlinear function of x_n, then $g(x)$ determines the characteristics of the compressor. It is instructive to consider a uniform quantizer whose input are samples in the range$[-x_{\max}, +x_{\max}]$ and the number of quantization levels N is a power of 2, $N = 2^v$. From this, the length of each quantization region is given by

$$\Delta = \frac{2x_{\max}}{N} = \frac{x_{\max}}{2^{v-1}} \tag{1}$$

The quantized values are chosen to be midpoints of the quantization regions and, therefore, the error $\tilde{x} = x - Q(x)$ is a random variable taking values in the interval $\left(-\frac{\Delta}{2}, +\frac{\Delta}{2}\right]$. In ordinary PCM applications, the number of levels (N) is usually high and the range of variations of the input signal (amplitude variations $x_{\max}$) is small. This means that the length of each quantization region (Δ) is small and, under these assumptions, in each quantization region the error $\tilde{X} = X - Q(X)$ can be well approximated by a uniformly distributed random variable on $\left(-\frac{\Delta}{2} + \frac{\Delta}{2}\right]$. The distortion introduced by quantization (quantization noise) is therefore

$$E[\tilde{X}^2] = \int_{-(\Delta/2)}^{+(\Delta/2)} \frac{1}{\Delta}\tilde{x}^2\, d\tilde{x} = \frac{\Delta^2}{12} = \frac{x_{\max}^2}{3N^2} = \frac{x_{\max}^2}{3 \times 4^v} \tag{2}$$

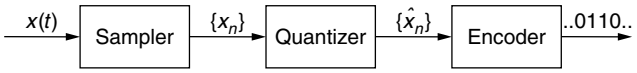


Figure 1. Block diagram of a PCM system.

where v is the number of bits per source sample. The SQNR ratio then becomes

$$\text{SQNR} = \frac{\overline{X^2}}{\overline{\tilde{X}^2}} = \frac{3 \times N^2\overline{X^2}}{x_{\max}^2} = \frac{3 \times 4^v\overline{X^2}}{x_{\max}^2} \tag{3}$$

If we denote the normalized X by $\breve{X}$, that is, $\breve{X} = \dfrac{X}{x_{\max}}$, then

$$\text{SQNR} = 3 \times N^2\overline{\breve{X}^2} = 3 \times 4^v\overline{\breve{X}2} \tag{4}$$

Note that by definition $|\breve{X}| \leq 1$ and, therefore, $\overline{\breve{X}^2} \leq 1$. This means that $3N^2 = 3 \text{ x } 4^v$ is an upperbound to the SQNR for a uniform quantizer. This also means that SQNR in a uniform quantizer deteriorates as the dynamic range of the source increases because an increase in the dynamic range of the source results in a decrease in $\overline{\breve{X}^2}$.

Expressing SQNR in decibels, one obtains

$$\text{SQNR}|_{\text{dB}} = P_{\breve{X}}|_{\text{dB}} + 6v + 4.8 \tag{5}$$

It is seen that each extra bit (increase in v by one) increases the SQNR by 6 dB.

3. COMPANDING

As long as the statistics of the input signal are close to the uniform distribution, a uniform quantizer works fine. However, in coding of certain signals such as speech, the input distribution is far from being uniformly distributed. For a speech waveform, in particular, there exists a higher probability for smaller amplitudes and lower probability for larger amplitudes. Therefore, it makes sense to design a quantizer with more quantization regions at lower amplitudes and less quantization regions at larger amplitudes. The resulting quantizer will be a nonuniform quantizer having quantization regions of various sizes.

The usual method for performing nonuniform quantization is to first pass the samples through a nonlinear element that compresses the large amplitudes (reduces dynamic range of the signal) and then perform a uniform quantization on the output. At the receiving end, the inverse (expansion) of this nonlinear operation is applied to obtain the sampled value. This techniques is called *companding* (compressing–expanding). A block diagram of this system is shown in Fig. 2.

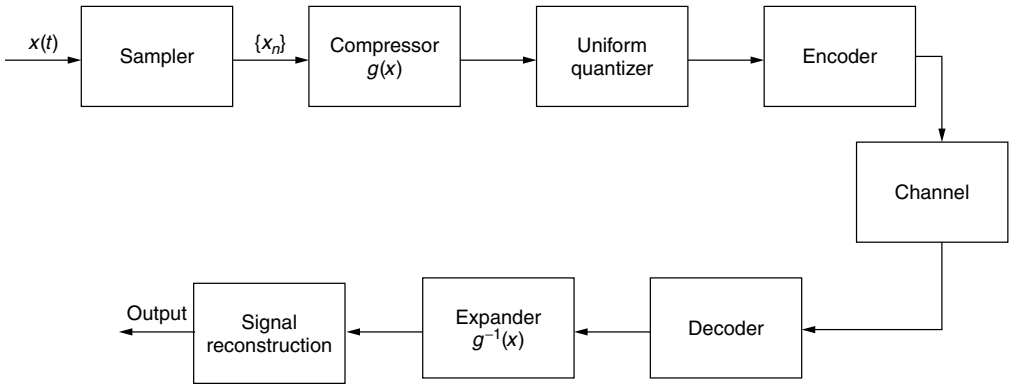


Figure 2. Block diagram of a PCM system employing a compander.

Two types of companders are widely used for speech coding. The μ-law compander used in the United States and Canada employs the logarithmic function at the transmitting side, where $|x| \leq 1$:

$$g(x) = \frac{\log(1 + \mu|x|)}{\log(1 + \mu)} \operatorname{sgn}(x) \tag{6}$$

The parameter μ controls the amount of compression and expansion. The standard PCM system in the United States and Canada employs a compressor with $\mu = 225$, followed by a uniform quantizer with 128 levels (7 bits per sample). Use of a compander in this system improves the performance of the system by ~30 dB. This means that the compander has implicitly provided an additional 5 bits of precision to that obtained by a uniform quantizer. A plot of the μ-law compander characteristics is shown in Fig. 3.

The second widely used logarithmic compressor is the A-law compander. The characteristic of this compander is given by

$$g(x) = \frac{1 + \log A|x|}{1 + \log A} \operatorname{sgn}(x) \tag{7}$$

where A is chosen to be 87.56. The performance of this compander is comparable to the performance of the μ-law compander. The characteristics of this compander are shown in Fig. 4.

4. CONCLUDING REMARKS

In digital coding of analog signals, such as speech signals, that have a nonuniform amplitude distribution, a nonuniform quantizer should be employed in order to reduce the number of bits per sample for a specified signal fidelity. Companding used in conjunction with a uniform quantizer results in a nonuniform quantization of the signal. In the case of speech signals, the μ-law or the A-law compander used in conjugation with a 7-bit/sample uniform quantizer result in an SQNR that is comparable to that obtained with a 12-bit/sample uniform quantizer without a compander. Thus, the use of a compander in the coding of speech signals via PCM has resulted in a significant reduction in the bit rate (by 5 bits per sample) that is to be transmitted over the channel.

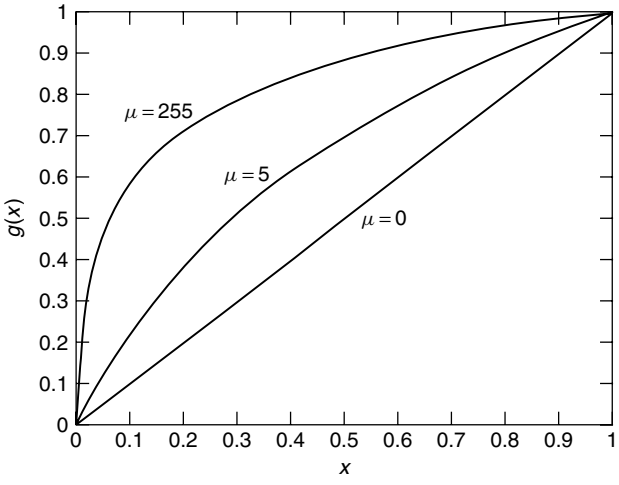


Figure 3. μ-law compander characteristics.

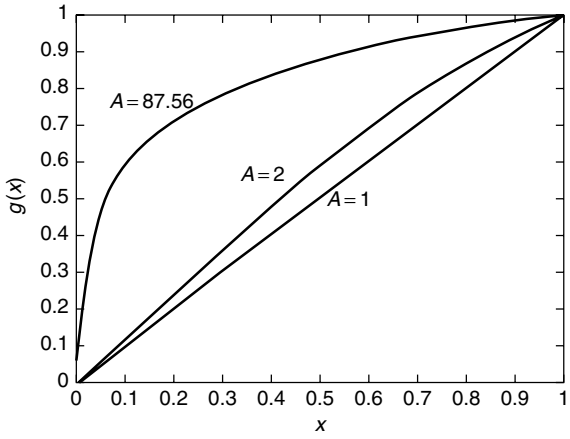


Figure 4. A-law compander characteristics.

For a detailed treatment of digital coding of signal waveforms, the reader may refer to the treatise by Jayant and Noll [1].

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing*

Laboratory (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. N. S. Jayant and P. Noll, *Digital Coding of Waveforms–Principles and Applications of Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

COMPENSATION OF NONLINEAR DISTORTION IN RF POWER AMPLIFIERS

SEKCHIN CHANG
EDWARD J. POWERS
University of Texas at Austin
Austin, Texas

1. INTRODUCTION

In all wireless communication systems, whether terrestrial or satellite, an RF power amplifier (PA) is required at the transmitter to ensure the signal will be received at the receiver with sufficient signal-to-noise ratio (SNR). For small signal level inputs, the PA input–output relationship is linear. However, as the input signal level is increased, the output power eventually saturates, causing the input–output relationship to become nonlinear [1].

On one hand, it is desirable to operate a PA near saturation because of increased RF power output (important in, e.g., satellite communications) and increased DC-to-RF power conversion efficiency (important in battery-powered RF devices, e.g., mobile phones). However, operating the PA near saturation results in a number of undesirable nonlinear effects, including amplitude and phase distortion and the generation of in-band and out-of-band intermodulation frequencies. These phenomena associated with nonlinearity lead to an increased bit error rate (BER), in-band interference, and adjacent-channel interference (ACI).

One approach to mitigating the undesirable effects associated with saturation is to back off from saturation into a more linear region, but at the expense of reducing the available RF power output and efficiency. For example, if the RF power output is backed off 10 dB (which is not unreasonable), the available RF output power will be reduced to one-tenth of the maximum output power. For these reasons there is a significant need to mitigate the effects of nonlinearities and, thereby, reduce the amount of backoff required.

To compensate for the deleterious effects of PA nonlinearities, one might try to compensate the distorted signal received at the receiver. Even when this can be done in an efficient and practical matter with some type of equalizer, it does not eliminate the undesirable effects occurring at the transmitter associated with the generation of intermodulation frequencies and the resultant in-band and out-of-band interference. Thus we focus on nonlinear compensation at the transmitter.

To mitigate the efforts of nonlinearities at the transmitter, it is customary to predistort the signal to be transmitted in a way that is equal and opposite to the distortion introduced by the PA. In such a case, the PA output should ideally correspond to the output of an ideal linear amplifier (i.e., one that is linear right up to the saturation point). Of course, no predistortion technique is perfect, so some backoff is required. The next question is how much backoff is required and how one addresses the tradeoff between large backoff (linear operation, but low RF power output) and small backoff (nonlinear operation, but high RF power output). In this article we will address the tradeoff issue via the well-known concept of total degradation versus output backoff in order to determine the optimum output power backoff.

Generally speaking, the deleterious effects of nonlinearities are less severe for constant-amplitude modulation schemes such as various forms of PSK (phase shift keying) (BPSK, QPSK, etc.) versus more bandwidth-efficient multiamplitude modulation schemes such as QAM (quadrature amplitude modulation). Thus in this article we will demonstrate the performance of predistorters using the more challenging case of QAM modulation.

Another factor to be considered involves the use of single-carrier (frequency) versus multicarrier systems. Multicarrier systems, such as OFDM (orthogonal frequency-division multiplexing) are increasingly being used in such systems as digital audiobroadcasting and future-generation personal communication systems because of their robustness to impulse noise and multipath fading. However, such multicarrier systems are characterized by high peak-to-average power ratios (PAR) because the multicarrier signals will occasionally constructively interfere. This constructive interference leads to high peak power levels that drive the PA into the saturation region with all its negative nonlinear consequences. This suggests that, in general, multicarrier systems require more output power backoff than does a single-carrier system. Thus the utilization of predistorters in multicarrier systems is the more challenging case and is, therefore, why we choose to demonstrate the use of predistorters using OFDM in this article.

In the next section we overview some of the nonlinear characteristics of power amplifiers. In Section 3 we consider the sensitivity of OFDM systems to nonlinear distortion, and in Section 4 provide an overview of various approaches to predistorters with emphasis on Volterra-based predistorters. In particular, the ability of the Volterra-based predistorter to reduce BER, total degradation, and output backoff is demonstrated via a simulation experiment using 16-QAM-based OFDM system. Conclusions are stated in Section 5.

2. NONLINEAR CHARACTERISTICS OF POWER AMPLIFIERS

The characteristics of PAs are usually expressed by amplitude modulation–amplitude modulation (AM/AM) and amplitude modulation–phase modulation (AM/PM) conversions, which represent the amplitude and phase distortions, respectively, depending on the input signal magnitude. If $x(n)$ is an input signal to the PA, it can be defined by

$$x(n) = r(n)e^{j\theta(n)} \tag{1}$$

where n denotes discrete time and $r(n)$ and $\theta(n)$ are the amplitude and the phase of the input signal $x(n)$, respectively. Let $A[\cdot]$ and $\Phi[\cdot]$ be AM/AM and AM/PM conversions, respectively. Therefore, if $s(n)$ is the output signal of the PA, it can be expressed as

$$s(n) = A[r(n)]e^{j\{\Phi[r(n)]+\theta(n)\}} \tag{2}$$

After the input signal $x(n)$ is amplified as in Eq. (2), the amplified signal $s(n)$ is transmitted into the wireless channel.

PAs can be broadly classified into traveling-wave tube amplifiers (TWTAs) and solid-state power amplifiers (SSPAs). TWTA PAs usually exhibit higher output power, but lower reliability and more severe nonlinearities than do SSPAs. Therefore, TWTAs are still dominant in applications requiring high levels of RF output power [2]. For TWTAs, the AM/AM and AM/PM conversions are modeled by the following equations [3]

$$A[(r(n)] = \frac{\alpha_a r(n)}{1+\beta_a r^2(n)} \qquad \text{(AM/AM conversion)} \tag{3}$$

$$\Phi[(r(n)] = \frac{\alpha_p r^2(n)}{1+\beta_p r^2(n)} \qquad \text{(AM/PM conversion)} \tag{4}$$

where α_a, β_a, α_p, and β_p are constants, and the subscripts a and p denote amplitude and phase, respectively. Figure 1 illustrates the curves of the AM/AM and AM/PM conversions for the typical values of constants: $\alpha_a = 2.0$, $\beta_a = 1.0$, $\alpha_p = \pi/3$, and $\beta_p = 1.0$ [3]. Note that the input and output amplitudes are normalized to their respective saturation values.

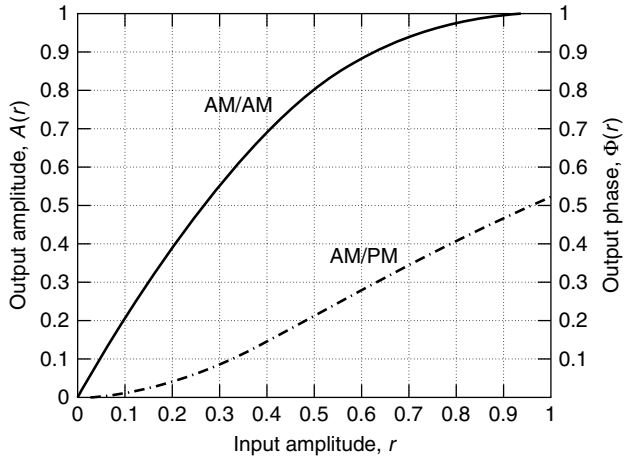


Figure 1. The AM/AM and AM/PM conversions of a traveling-wave tube amplifier (TWTA).

Unlike phase distortions in TWTAs, those in SSPAs are usually smaller [4]. In an SSPA AM/AM conversion is modeled as follows [4]:

$$A[(r(n)] = \frac{v_k r(n)}{\left(1+\left(\frac{v_k r(n)}{A_0}\right)^{2p_k}\right)^{1/2p_k}} \qquad \text{(AM/AM conversion)} \tag{5}$$

where v_k, A_0, and p_k are constants. TWTAs exhibit higher power gain and more severe nonlinearities than do SSPAs [5]. In the ideal case of either TWTAs or SSPAs, the AM/AM conversion curve should be a straight line with a constant slope, and the AM/PM conversion curve should be a constant.

Regardless of the kind of PA, the transmitted signal experiences some nonlinear distortion as indicated in Eq. (2). If the input signal $x(n)$ utilizes a simple modulation scheme such as binary phase shift keying (BPSK) or quadrature phase shift keying (QPSK), the nonlinear distortion has mild effects on the transmitted signal since these modulation schemes exhibit constant amplitude. However, if the input signal corresponds to a channel-efficient modulation scheme such as 16- or 64-QAM (quadrature amplitude modulation) where the amplitude varies, the nonlinear distortion will degrade the system performance severely. Moreover, multicarrier systems such as OFDM systems are more sensitive to nonlinear distortion as explained in Sections 1 and 3.

The degree of PA nonlinear distortion depends on the backoff amount. Backoff may be divided into input backoff (IBO) and output backoff (OBO) which are defined (in decibels) as Eqs. (6) and (7), respectively:

$$\text{IBO} = 10\log_{10}\frac{P_{i,\max}}{P_i} \tag{6}$$

$$\text{OBO} = 10\log_{10}\frac{P_{0,\max}}{P_0} \tag{7}$$

In Eq. (6), $P_{i,\max}$ represents the maximum input power of the PA at saturation and P_i denotes the mean input power of the signal at the PA input. Similarly, in Eq. (7) $P_{0,\max}$ represents the maximum output power of the PA at saturation and P_0 denotes the mean output power of the signal at the PA output. Figure 2 depicts the relationship between IBO and OBO using the AM/AM characteristics of a TWTA PA. Therefore, the operating point can be placed in the linear region of PA by backing off the PA from saturation, which leads to an effective reduction of nonlinear distortion. However, this scheme reduces the power efficiency of the PA and its output power. Therefore, a tradeoff between backoff and nonlinear distortion must be considered.

3. SENSITIVITY OF OFDM SYSTEMS TO PA NONLINEAR DISTORTION

Currently, OFDM is utilized for asymmetric digital subscriber line (ADSL), digital audiobroadcasting (DAB), and wireless local-area networks (WLANs) [6,7]. In

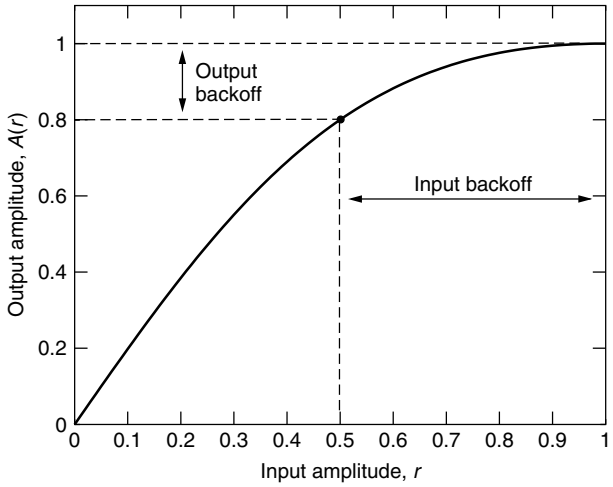


Figure 2. Input and output backoffs for a TWTA.

addition, OFDM is expected to be a strong candidate for wireless multimedia systems. Therefore, OFDM is believed to be a promising technique in wireless as well as wireline systems for high-rate data transmission [8]. OFDM utilizes a multicarrier modulation scheme, and exhibits many advantages over single-carrier modulation. Because of the increased symbol length, OFDM is robust to the effects of impulse noise and severe multipath fading [8,9]. In addition, the effects of the frequency selective fading can be greatly reduced without a high-cost equalizer by using a cyclic prefix.

However, multicarrier systems such as OFDM show great sensitivity to nonlinear distortion introduced by the PA. OFDM and other systems usually have a PA at the RF stage on the transmitter side as shown in Fig. 3 to increase the link fading margin of OFDM signals in a wireless channel. In Fig. 3, the incoming symbol input is a serial stream such as M-ary quadrature amplitude modulation (QAM) signals. After the input symbols are converted from serial to parallel (S/P) to form a vector of N M-ary symbols, the N symbols are modulated onto N subcarriers by the inverse fast Fourier transform (IFFT) and subsequently converted back from parallel to serial (P/S) data symbols. As mentioned by Karam and Sari [10], digital radio systems usually employ baseband pulseshaping at the transmitter. Therefore, the linear filter in Fig. 3 indicates the pulse shaping filter.

Since the input data to the IFFT are generally independent and identically distributed (i.i.d.) in OFDM systems, the modulated signals can be considered random signals with a zero-mean Gaussian distribution according to central limit theorem. Therefore, the signals of OFDM systems exhibit higher peak-to-average power ratio (PAR) than do those of single-carrier systems in the time domain. Therefore, OFDM systems are more sensitive to PA nonlinear distortion than are single-carrier systems because of their larger PAR. In other words, more frequent utilization of the PA's high-power region is inevitable in OFDM systems in order to produce an average output power comparable to that of single-carrier systems, thus resulting in severe performance degradation due to nonlinear distortion introduced by the PA.

4. COMPENSATION METHODS OF NONLINEAR DISTORTION

Nonlinear distortion can be reduced by backing off the PA from saturation. However, as stated earlier, back-off reduces the PA output power. Thus, some form of nonlinear compensation is desirable to mitigate nonlinear distortion while at the same time reducing the amount of backoff required. For the efficient compensation of nonlinear distortion, predistorters have been widely utilized, where the predistorter is placed in front of the PA. The predistorter distorts the input signal in such a way as to compensate for the nonlinear distortion introduced by the PA. Moreover, predistorters may be designed to be adaptive. The adaptation property is very desirable because the characteristics of PAs are time variant due to temperature variation and aging [3].

4.1. Some General Predistortion Schemes

In general, predistorters can be classified into several types according to the predistorter structure and position. Usually, predistorters belong to a minimum mean-square error (MMSE) type or amplitude/phase (AP) type based on predistorter structure. According to the location where the predistorter is placed, it is commonly called an analog or digital (or data) predistorter.

Figure 4 shows an MMSE predistorter. As depicted in this figure, the coefficients of the predistorter are trained to minimize the mean square error between output $y(n)$ of PA and input $x(n)$ to the predistorter. An MMSE predistorter has been proposed [11] and its performance analyzed in OFDM systems [12]. In these articles, the input and output responses were related with third-order polynomials. Generally, the MMSE predistorter can achieve a global compensation of the nonlinear distortion

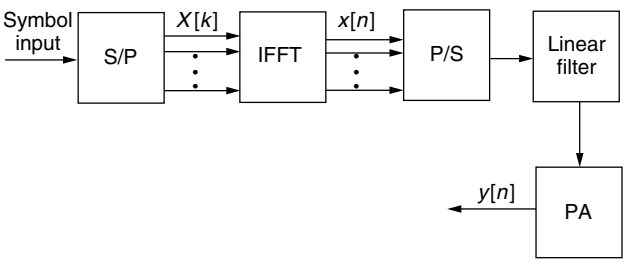


Figure 3. An OFDM system with PA.

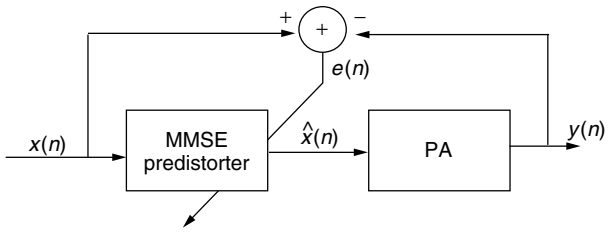


Figure 4. The general structure of a MMSE predistorter.

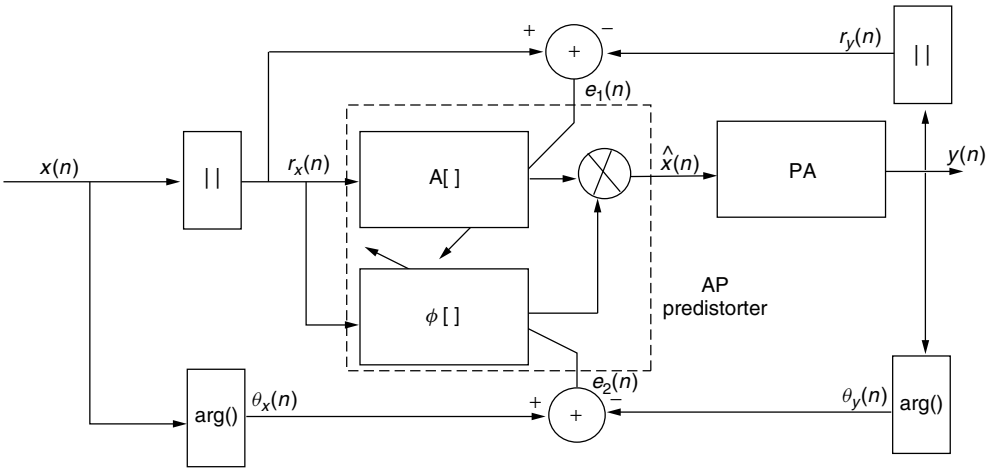


Figure 5. The general structure of an AP predistorter.

of the PA if a proper learning architecture is employed. On the other hand, the MMSE approach requires complex-valued coefficients in the predistorter to compensate for amplitude and phase distortion.

The amplitude/phase (AP) predistorter is illustrated in Fig. 5, which shows that the predistorter consists of amplitude $A[\cdot]$ and phase $\Phi[\cdot]$ coefficients. In other words, the coefficients of each type are separately trained so as to minimize the amplitude and the phase errors between output $y(n)$ of the PA and input $x(n)$ of the predistorter. The AP predistorter was introduced by D'Andrea et al. [13] and its performance was measured in OFDM systems [14]. In these articles, the linearizing functions $A[\cdot]$ and $\Phi[\cdot]$ of the predistorter were separately approximated as polynomials. Usually, each polynomial has real-valued coefficients. Therefore, the AP predistorter can exhibit a faster convergence rate than does the MMSE predistorter in training the predistorter coefficients. However, the predistorter requires additional separation and combination modules for the amplitude/phase of the input signal $x(n)$ because it must obtain separate optimal solutions to compensate for the amplitude and phase distortions of the PA.

Figure 6 depicts general structures of analog and digital (or data) predistorters. As shown in this figure, the analog predistorter is placed after the pulseshaping filter in the RF (or possibly IF) band. On the other hand, the digital predistorter is placed before the pulseshaping filter at baseband. Therefore, while the analog predistorter just compensates the memoryless nonlinearity of the PA, the digital predistorter is required to compensate a nonlinearity with memory due to the series combination of the linear pulseshaping filter, which provides memory, and the PA, which is essentially a memoryless nonlinearity. However, the digital predistorter exhibits more flexibility in determining the predistorter coefficients than the analog predistorter, since the learning algorithm is programmable in the digital predistorter. This suggests that the digital predistorter can be more adaptive to time-variant PAs.

In the following sections, we describe several predistorters that have shown good performance in compensating nonlinear distortion introduced by the PAs. We will emphasize the Volterra-based predistorter, which has been utilized for compensation of nonlinear distortion in OFDM systems. The Volterra-based predistorter belongs to the MMSE and digital class of predistorters. As indicated above, such a predistorter type has a global solution and training flexibility, but also demands a lot of learning time. Therefore, we will also indicate an efficient learning architecture for the predistorter.

4.2. Predistortion Using a Lookup Table

This efficient predistortion scheme has been applied to QAM radio systems with good results [10]. The scheme utilized a lookup table or random-access memory (RAM), which included the predistorter candidates. If an input sequence to the predistorter is given, the candidate that minimizes the error between the input data to the predistorter and the output data of the PA is selected as the output of the predistorter. Each candidate is updated until the error converges to a desired value. Since the input signals exhibit a finite number of data levels regardless of the modulation scheme used in single-carrier systems, the lookup table predistortion scheme can be easily adapted to various PAs. However, it is probably not feasible to

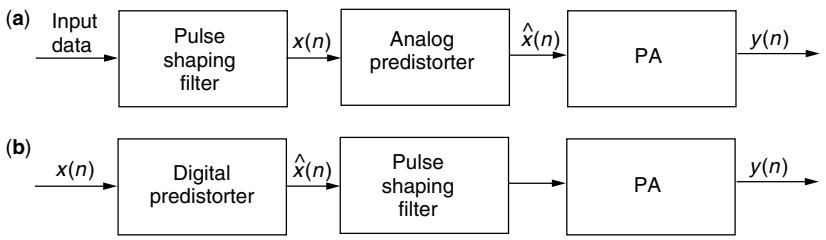


Figure 6. The general structures of analog and digital predistorters for (**a**) analog predistorter and (**b**) digital predistorter.

utilize this scheme in multicarrier systems because these systems usually show many different data levels due to the constructive and destructive interference of the multicarrier components.

4.3. Volterra-Based Predistorter

This adaptive predistorter essentially consists of a Volterra series because the Volterra series has shown excellent performance in modeling and compensating nonlinear phenomena with memory. Thus, it is particularly suitable for use as a digital predistorter. In discrete time, a third-order Volterra series for a causal, finite memory system becomes

$$y[n] = \sum_{k=0}^{N_1-1} h_k^{(1)} x[n-k] + \sum_{k=0}^{N_2-1}\sum_{l=0}^{N_2-1} h_{k,l}^{(2)} x[n-k]x^*[n-l] + \sum_{k=0}^{N_3-1}\sum_{l=0}^{N_3-1}\sum_{m=0}^{N_3-1} h_{k,l,m}^{(3)} x[n-k]x[n-l]x^*[n-m] + e[n] \tag{8}$$

where N_1, N_2, and N_3 respectively denote the memory duration of the first-order, the second-order, and the third-order terms; $x[n]$ and $y[n]$ are the complex input and output, respectively; $h_k^{(1)}$, $h_{k,l}^{(2)}$, and $h_{k,l,m}^{(3)}$ are the complex discrete time domain Volterra kernels of order 1, 2, 3, respectively; and $*$ and $e[n]$ denote the complex conjugate and the modeling error, respectively. In Fig. 6b, the PA is preceded by a linear filter. In digital communication systems, the combination of the PA and linear filter may be regarded as a nonlinear system with memory that is to be compensated [15]. Since the Volterra series may be regarded as a Taylor series with memory, a Volterra-based predistorter exhibits a structure suitable for compensation of such a system.

We may represent the third-order Volterra series of Eq. (8) in matrix form as

$$\hat{d}[n] = \mathbf{h}\mathbf{x}^T[n] \tag{9}$$

where $\hat{d}[n]$ is the estimated output of the Volterra predistorter, the superscript T denotes the transpose of the matrix, $\mathbf{h}$ is the Volterra kernel vector, and $\mathbf{x}[n]$ is the input vector, which are defined by

$$\mathbf{h} = [h_0^{(1)}, h_1^{(1)}, \ldots, h_{N_1-1}^{(1)}, h_{000}^{(3)}, h_{001}^{(3)}, h_{002}^{(3)}, \ldots, \times h_{klm}^{(3)}, \ldots, h_{(N_3-1)(N_3-1)(N_3-1)}^{(3)}] \tag{10}$$

$$\mathbf{x}[n] = [x[n], x[n-1], \ldots, x[n-N_1+1], |x[n]|^2x^*[n], \times |x[n]|^2x^*[n-1], |x[n]|^2x^*[n-2], \ldots, \times |x[n-N_3+1]|^2x^*[n-N_3+1]] \tag{11}$$

where N_1 and N_3 are the memory durations of the first-order term and the third-order term, respectively. In Eqs. (10) and (11), the absence of the second-order term is due to the fact that even-order intermodulation components do not interfere with the in-band signal for a bandpass nonlinear channel [16], since they lie out of band. However, odd-order nonlinearities generate intermodulation frequency components that lie both out of band and in band. Because of this latter fact, odd-order terms are retained in the Volterra series.

Figure 7 shows why only the odd-terms contribute to the bandpass channel. In communication systems, the PA is placed at the RF stage. Therefore, the signal $x(t)$ is upconverted to the carrier frequency f_0. The upconverted signal $r(t)$ is inputted to the nonlinear PA, thus various harmonics of the carrier frequency f_0 appear in the output. As Fig. 7 indicates, only the odd-order nonlinearities contribute to in-band components centered at f_0.

It is well known that Volterra kernels can be assumed symmetric without any loss of generality [17]. Therefore, the third term in Eq. (8) can be rewritten as follows:

$$y_3[n] = \sum_{k=0}^{N_3-1}\sum_{l=k}^{N_3-1}\sum_{m=0}^{N_3-1} h_{k,l,m}^{(3)} x[n-k]x[n-l]x^*[n-m] \tag{12}$$

Taking into account symmetry, the number of kernel coefficients for each order term will be

$$K_1 = N_1 \tag{13}$$

$$K_3 = \frac{N_3^2(N_3+1)}{2} \tag{14}$$

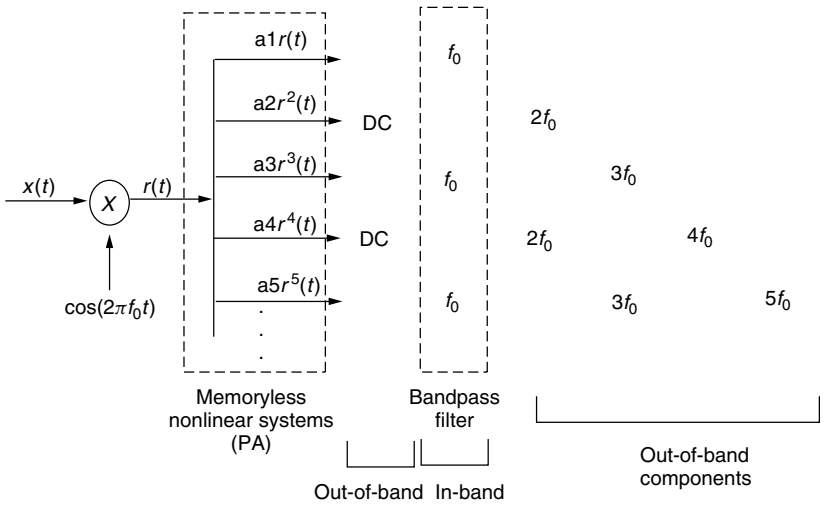


Figure 7. Abscence of even-order terms in a bandpass channel.

and the total number of Volterra coefficients for the predistorter can be represented as

$$K_T = K_1 + K_3 \tag{15}$$

The next challenge is to determine the numerical values of the Volterra coefficients for a given PA.

4.4. Learning Architecture for Volterra-Based Predistorter

The Volterra-based predistorter belongs to the MMSE and digital classes of predistorters. A particular challenge associated with this predistorter (and indeed most predistorters) is to determine the Volterra coefficients, since the desired output of the predistorter is not readily known beforehand. Thus, the predistorter design requires an efficient learning architecture to train the predistorter coefficients rapidly.

A relatively new and efficient training architecture [18] composed of both the indirect and direct learning algorithms is depicted in Fig. 8. In earlier work [15], only the indirect learning scheme was utilized. As seen in Fig. 8, there are two identical models: the actual predistorter and another for training. The two models share the same predistorter coefficient vector **h**. As stated previously, the "PA with memory" in Fig. 8 represents the PA preceded by a linear filter. In the learning structure of Fig. 8, at each iteration the predistorter coefficients are first updated using the indirect learning algorithm, which makes $\alpha[n]$ approach zero, and are then updated using the direct learning algorithm, which makes $\beta[n]$ approach zero. As $\alpha[n]$ and $\beta[n]$ approach zero, $y[n]$ approaches $x[n]$, which implies complete compensation of the PA's nonlinear distortion. These updates continue until the coefficients converge. Since the coefficients are updated twice in each iteration, this training scheme exhibits rapid convergence. Both the indirect and direct learning algorithms utilize the recursive least-squares (RLS) technique [19] for fast updates of the predistorter coefficients. A particular advantage of the approach described here is that one does not first require a Volterra model of the system to be compensated.

Figure 9 shows the learning curves (mean squared error (MSE) vs the number of iterations) for two Volterra-based adaptive predistorters, one of which was trained by the new learning scheme, direct and indirect architecture (DIA), and the other by the old learning scheme, indirect architecture-only (IAO) methods. From this figure, it is seen that the DIA scheme achieves an improvement of up to 5 dB or more in the MSE performance over the IAO scheme for 2000 or more iterations. This result signifies that the new learning algorithm is effective in increasing the convergence rate of the Volterra-based adaptive predistorter.

4.5. Simulation Experiments

In this section we carry out a simulation experiment to give the reader some feeling for the type of advantages one can achieve using predistorters.

An OFDM system with 128 subcarriers and 16-QAM symbols is considered. In this simulation, it is assumed that the nonlinear degree is 3 and the memory length of the linear filter shown in Fig. 3 is 3. Therefore, a third-order Volterra-based predistorter is used, and the first-order memory length (N_1) and the third-order memory length (N_3) for the Volterra predistorter are set to 3 in this simulation. From Eq. (15), the total number of the predistorter coefficients is 21. As mentioned in Section 2, the TWTA is known to be more nonlinear than the SSPA. Thus, we utilize a TWTA in this simulation because it poses a more severe challenge for the predistorter.

Received 16-QAM constellations of the OFDM system with and without the Volterra predistorter are shown in Fig. 10, where E_b/N_0 (energy per bit divided by the

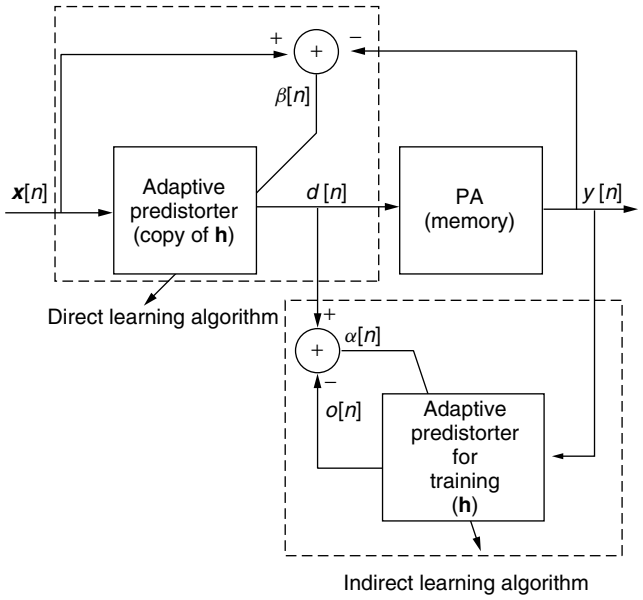


Figure 8. The learning architecture for a Volterra-based predistorter. ([18], © 2000 IEEE).

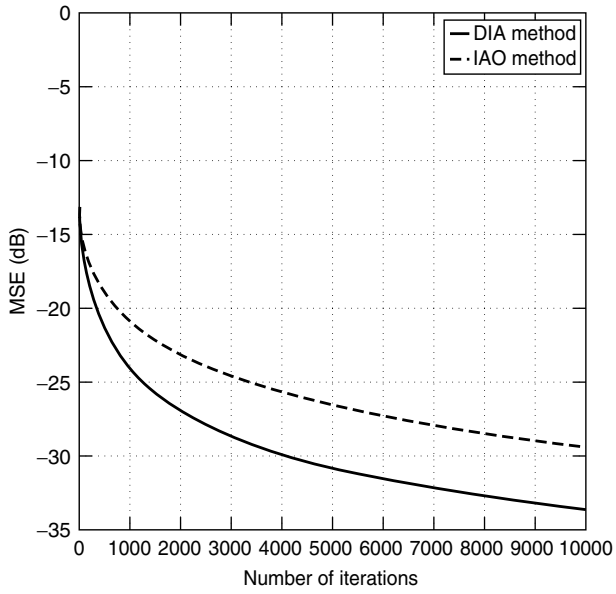


Figure 9. Learning curves of predistorters based on DIA and IAO ([18], © 2000 IEEE).

noise power spectral density) is assumed to be ∞. In the figure captions, the normalized mean-squared error (NMSE) values are also given for numerical comparisons of distorted or compensated constellations. Figures 10a,c show the distorted received signal constellations (no predistortion compensation) at output backoffs (OBOs) of 8.5 and 4.7 dB, respectively. Figures 10b,d show the received signal constellations when compensated by the proposed Volterra predistorter for OBOs of 8.5 and 4.7 dB, respectively. Comparison of Figs. 10a and 10b indicates that the signal distortion is almost perfectly compensated by the Volterra predistorter when the OBO is 8.5 dB. Figure 10d also shows some compensation of the signal distortion seen in Fig. 10c for an OBO of 4.7 dB. The relatively large clustering of Fig. 10d is due to the fact that the predistorter is driven into the saturation region by the large envelope fluctuations characteristic of multicarrier systems since the OBO of 4.7 dB is relatively small.

Therefore, we can see from this simulation result that the performance of the Volterra-based predistorter in compensating nonlinear distortion of PA is dependent on the OBO values. As a performance measure to investigate the tradeoff between nonlinear distortion and OBO in the PA, the total degradation (TD) is usually utilized and defined in decibels by

$$\mathrm{TD} = \left[\frac{E_b}{N_{0_{(\mathrm{NLPA})}}} - \frac{E_b}{N_{0_{(\mathrm{LPA})}}}\right] + \mathrm{OBO} \quad (\mathrm{dB}) \tag{16}$$

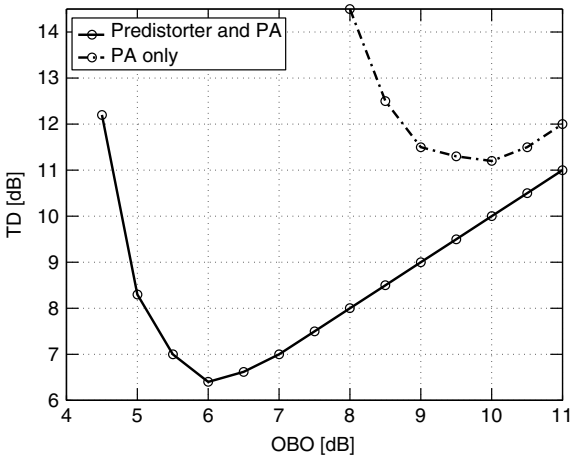


Figure 11. Total degradation versus OBO at a BER of 10^{-4} for the PA with predistorter and the PA-only ([18], © 2000 IEEE).

where $E_b/N_{0_{(\mathrm{NLPA})}}$ represents the required E_b/N_0 to obtain a specific BER when the nonlinear PA is used, and $E_b/N_{0_{(\mathrm{LPA})}}$ denotes the required E_b/N_0 to maintain the same BER, assuming that the amplifier exhibits no nonlinearities. In Fig. 11, the TD is shown for various OBO values for a BER of 10^{-4}. This figure illustrates the cases for the PA with predistorter and for the PA-only. The OBO which minimizes TD is called the *optimum OBO*. As seen in the figure, the minimum TD for the PA-only is 11.2 dB at an OBO of 10 dB, and the minimum TD of the PA with

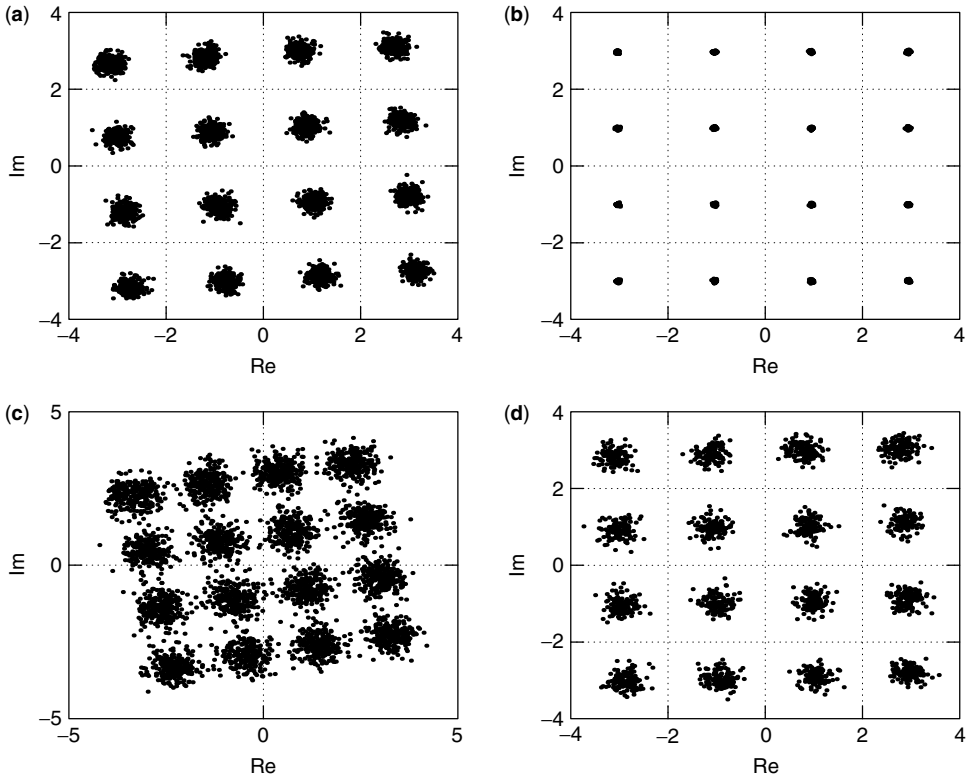


Figure 10. Received 16-QAM constellations for (**a**) PA-only (NMSE = 0.00913) at OBO = 8.5 dB, (**b**) predistorter and PA (NMSE = 0.000071) at OBO = 8.5 dB, (**c**) PA-only (NMSE = 0.0533) at OBO = 4.7 dB, and (**d**) predistorter and PA (NMSE = 0.00825) at OBO = 4.7 dB, respectively. ([18], © 2000 IEEE).

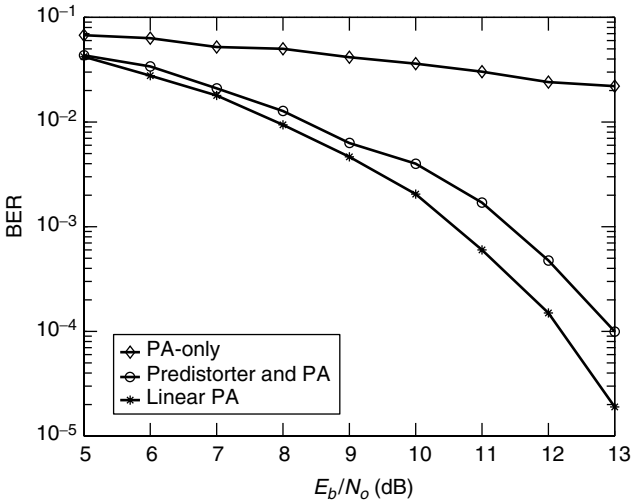


Figure 12. BER performance versus E_b/N_0 at OBO of 6.0 dB for the PA-only, the PA with Volterra-based adaptive predistorter, and the linear PA.

predistorter is 6.4 dB at an OBO of 6.0 dB. Therefore, the PA with predistorter achieves an output power increase of about 4 dB (10 dB − 6 dB) over PA-only case. Furthermore, the TD is reduced by 4.8 dB (11.2 dB − 6.4 dB).

Figure 12 compares the bit error rate (BER) performances for the cases of the PA without any predistortion and with the Volterra-based predistorter, and the ideal case without any nonlinear distortion of the PA (linear PA) at an OBO of 6.0 dB. In Fig. 12, it is shown that the BER performance is severely degraded as a result of nonlinear distortion in the case without predistortion; that is, increasing E_b/N_0 has a relatively small effect on reducing the BER. On the other hand, the BER performance of the combination of the Volterra-based predistorter and PA is fairly close to that of an ideal linear PA, thereby demonstrating the efficacy of the predistorter.

5. CONCLUSION

In this brief article the advantages of using predistorters to mitigate the effects of nonlinear distortion introduced by PAs and to reduce the amount of output backoff have been stressed and demonstrated via a predistorter example using a Volterra-based predistorter. In conclusion, it should be reiterated that there are many approaches to predistorters, some of which were mentioned in this article, and others of which are cited in the bibliography.

BIOGRAPHIES

Edward J. Powers received his B.S., M.S., and Ph.D. degrees from Tufts University, Massachusetts Institute of Technology, and Standford University in 1957, 1959, and 1965, respectively. All degrees were in electrical engineering. From 1959 to 1965 Dr. Powers was employed by Lockheed Missiles and Space Company in Sunnyvale and Palo Alto, California. In 1965, he joined the University of Texas at Austin, where he subsequently became Chair of the Department of Electrical and Computer Engineering from 1981 to 1989. He is currently the Texas Atomic Energy Research Foundation Professor in Engineering and Director of the Telecommunications and Signal Processing Research Center. His current professional interests include applications of higher-order statistical signal processing to detect, analyze, and model time series data associated with nonlinear physical phenomena; and applications of wavelets and time-frequency techniques to detect and classify various transient events in physical systems. He was elected a Fellow of IEEE in 1983.

Sekchin Chang received the B.S. and the M.S. degrees in electronics engineering from Korea University, Seoul, Korea in 1991 and 1993, respectively, and the Ph.D. degree in electrical engineering from the University of Texas at Austin in 2001. From 1993 to 1998, he was with Electronics and Telecommunications Research Institute (ETRI), Taejon, Korea, where he worked on the design and development of IS95 CDMA systems. In 2000, he joined Motorola, Austin, Texas, where he contributed to the design of modems for WCDMA systems, and is currently involved in the development of WLAN systems. His research interests include OFDM, WCDMA, adaptive predistortion, carrier and timing recovery, RAKE receivers, blind equalizers, and MIMO systems.

BIBLIOGRAPHY

1. S. Benedetto, E. Biglieri, and V. Castellani, *Digital Transmission Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
2. N. Escalera, W. Boger, P. Denisuk, and J. Dobosz, Ka-band, 30 watts solid state power amplifier, *Proc. IEEE MTT-S Int. Microwave Symp.*, Boston, June 2000, Vol. 2, pp. 561–563.
3. A. A. M. Saleh, Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers, *IEEE Trans. Commun.* **29**(11): 1715–1720 (Nov. 1981).
4. E. Bogenfeld, R. Valentin, K. Metzger, and W. Sauer-Greff, Influence of nonlinear HPA on trellis-coded OFDM for terrestrial broadcasting of digital HDTV, *Proc. 1993 IEEE Global Telecommun. Conf.*, Houston, TX, Nov. 1993, pp. 1433–1438.
5. M. Obermier and E. J. Powers, The effects of nonlinear high power amplifiers on space based phased array antenna patterns, *Proc. IEEE Int. Conf. Phased Array Systems and Technology*, Dana Point, CA, May 2000, pp. 45–48.
6. *Asymmetric Digital Subscriber Line* (*ADSL*) *Metallic Interface*, ANSI T1.413, 1995.
7. M. Alard and R. Lassalle, Principles of modulation and channel coding for digital broadcasting for mobile receivers, *EBU Rev. Techn.* **224**: 168–190 (Aug. 1987).
8. L. J. Cimini, Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing, *IEEE Trans. Commun.* **33**(7): 665–675 (July 1985).
9. J. A. C. Bingham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* **28**: 5–14 (May 1990).
10. G. Karam and H. Sari, A data predistortion technique with memory for QAM radio systems, *IEEE Trans. Commun.* **39**(2): 336–344 (Feb. 1991).

11. M. G. Di Benedetto and P. Mandarini, A new analog predistortion criterion with application to high efficiency digital radio links, *IEEE Trans. Commun.* **43**: 2966–2974 (Dec. 1995).

12. M. G. Di Benedetto and P. Mandarini, An application of MMSE predistortion to OFDM systems, *IEEE Trans. Commun.* **44**: 1417–1420 (Nov. 1996).

13. N. A. D'Andrea, V. Lottici, and R. Reggiannini, RF power amplifier linearization through amplitude and phase predistortion, *IEEE Trans. Commun.* **44**: 1477–1484 (Nov. 1996).

14. N. A. D'Andrea, V. Lottici, and R. Reggiannini, Nonlinear predistortion of OFDM signals over frequency-selective fading channels, *IEEE Trans. Commun.* **49**: 837–843 (May 2001).

15. C. Eun and E. J. Powers, A predistorter design for memoryless nonlinearity preceded by a dynamic linear system, *Proc. 1995 IEEE Global Telecommunications Conf.*, Singapore, Nov. 1995, pp. 152–156.

16. S. Benedetto and E. Biglieri, Nonlinear equalization of digital satellite channels, *IEEE J. Select. Areas Commun.* **SAC-1**(1): 57–62 (Jan. 1983).

17. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, Wiley, New York, 1980.

18. S. Chang and E. J. Powers, A compensation scheme for nonlinear distortion in OFDM systems, *Proc. 2000 IEEE Global Telecommunications Conf.*, San Francisco, Nov. 27–Dec. 1, 2000, pp. 736–740.

19. S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.

COMPUTER COMMUNICATIONS PROTOCOLS

EMMANOUEL VARVARIGOS
University of Patras
Patras, Greece

THEODORA VARVARIGOU
National Technical University
Patras, Greece

1. WHAT IS A COMMUNICATION PROTOCOL?

In the "Information Age," concepts such as "communications," "information sharing," and "networking" sometimes seem to monopolize the attention of not only the information technology (IT) experts but also the ever-growing IT end-user community. In this context, "networking" denotes the information exchange between computers or like systems, which can be very different from one another and may be scattered over wide geographic areas.

Outside the network are the computers, databases, terminals, and other devices that need to exchange information. Messages originate at these external hosts, pass into the network, pass from node to node on the communication links, and finally pass out to the external recipients. The nodes of the network, usually computers in their own right, serve primarily to forward the messages through the network.

In order for such external systems to communicate, a common "language" has to be established among them. At the human level, when two people want to communicate (effectively), they have to agree first on a language that both understand, and then on a set of rules that both have to follow; for instance, they must not speak simultaneously (or nobody will be heard), and they have to adjust their voice volume to the environmental conditions (so that voice will not fade away before it reaches the other's ears). All these rules, conventions, and distributed algorithms that communicating parts have to follow are referred to, in total, as a "protocol." Without an agreed-on protocol, communication may be hard or impossible to accomplish.

Similarly, every computer communication is governed by a certain set of rules that have to be agreed on by all the partners involved, prior to communication establishment. These rules make up the protocol of a computer communication and, typically, include items such as the data format to be used, the order in which messages are exchanged, the actions to be taken on receipt of a message, the error detection and handling techniques to be employed, and so on.

From the real world, we already know that a communication task may be rather complex. For example, imagine the case where the commander of the "BLUE" army in a battlefield has to communicate an "attack plan" to the commander of the allied "GREEN" army, who has a different nationality, speaks a different language, probably has a different perception of operations, and is located at the other side of the battlefield, with their common enemy (the "RED" army) operating in between them. To ensure that the "attack order" is correctly received and perceived by the GREEN commander, the BLUE commander has to undertake a number of subtasks. For example:

1. Use a language that the recipient of the message will understand. This can include the use of a mutually accepted language (e.g., English) and the use of a standard terminology to reflect a certain military concept of operations.
2. Use an agreed-on format to compose the message (i.e., use a certain military messaging structured format).
3. Encrypt the message with an appropriate code, so that nobody can read the original contents except for the recipient of the message, who has the proper code to decrypt it.
4. Print the encrypted message on a piece of paper and then put the paper in a stiff envelope, to protect it from environmental conditions, and probably write the recipient's name on it.
5. Give the envelope to a messenger who will carry it to its destination. The messenger must find the appropriate route to follow to arrive at the destination and avoid ambushes while crossing the hostile (RED) ground. En route to the destination, the messenger may have to modify the route, due to an unpredicted roadblock, and thus must know how to use a map, compass, or other device to reorientate.
6. The messenger, on reaching the allied campus, must find the qualified person to hand over the envelope. This person may be the GREEN commander in person, or the commander's authorized secretariat.
7. The GREEN commander or this secretariat signs an appropriate document (a "receipt"), acknowledging

receipt of the envelope, and hands it over (enclosed in a suitable envelope) to the messenger soldier.

8. The messenger then must carry the receipt back to the home campus, again crossing the RED ground, probably following a different return path.
9. The messenger, on reaching the home campus, hands over the receipt envelope to the commander (or to qualified personnel and is then dismissed).

To make things a little closer to reality, suppose now that the messenger is killed while crossing the enemy lines to carry the attack plan, and that if the two allied armies attack simultaneously, the allies win, but if they attack separately, the enemy wins. The two allied army commanders would therefore like to synchronize their attack at some given time, but each of them is unwilling to attack unless assured with certainty that the other will also attack. Thus, the first commander might send a message saying, "Let's attack on Saturday at noon; please acknowledge if you agree." The second commander, hearing such a message, might send a return message saying, "We agree; send an acknowledgment if you receive our message." It is not hard to see that this strategy leads to an infinite sequence of messages, where the last commander who sends a message is unwilling to attack until obtaining a commitment from the other side. What is more surprising is that it can be shown that no strategy (protocol) exists for allowing the two armies to synchronize with certainty. If the conditions are relaxed somewhat so as to require a high probability of simultaneous attack, the problem is solved (e.g., the first army decides on the time of the attack and sends many messengers simultaneously to the other army).

This scenario illustrates the difficulties that arise in the design of protocols, where distributed decisions based on distributed information must be made. We see how complex a communication task may get, how many subtasks have to be considered, and that protocols can be proved to be correct only in a specific, well-defined sense. Computer communications follow, in general terms, principles similar to those described in the previous real-life communication example. Those principles are enforced by certain rules and conventions that make up the various computer communications protocols and are implemented by suitably designed hardware and software components located anywhere between the communicating partners (including themselves).

2. THE OSI MODEL

Since a communication task across a computer network can be too complicated to be efficiently controlled as a whole, it is reasonable to split it up into several simpler, manageable, and cohesive subtasks and associate each subtask with a number of protocols. In this direction, the International Organization for Standardization (ISO) has developed a reference model, the *Open Systems Interconnection* (*OSI*) model, which defines seven communication subtasks arranged in a layered (hierarchical) structure, as shown in Fig. 1. Therefore, each building-block layer of

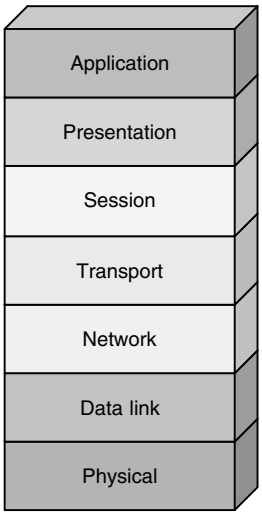


Figure 1. The seven layers of the OSI model.

the OSI model logically groups functions associated with the implementation of a specific communication subtask.

At the bottom of the OSI model, the *physical layer* is responsible for the transmission of the raw bit stream it receives from the next higher layer over the physical transmission medium. This function is usually performed by a *modem* (*mo*dulator–*dem*odulator), which maps a bit stream into a signal suitable for transmission over the underlying physical medium according to the electrical, mechanical, functional, and procedural characteristics of the medium and, at the receiving end, it does the inverse mapping and delivers the bit stream to the higher layer, via an appropriate interface. The physical layer thus provides its upper layers with a *virtual point-to-point bit pipe* and hides from them all the physical medium complexities. The design of the physical layer is usually considered as part of the job done by communication engineers as opposed to network engineers and is outside the scope of this article.

The next layer in the hierarchy, the *data-link control* (*DLC*) *layer*, provides the higher layers a *reliable point-to-point packet pipe over a single link*. It does so by organizing the exchanged bit stream into *frames* (data packets with header/trailer overhead control bits) and providing functions such as error control, retransmissions, and speed matching between sender and receiver.

The *network layer,* one layer above, is very important in internetworking, since it provides the means for end systems to communicate across a collection of communication networks. It is present at every intermediate network node (e.g., router) that interconnects communication networks, and its main purpose is to hide from its upper layers the underlying network technology and topology, providing a *virtual end-to-end packet pipe* between end systems. In order to transfer packets from the source to the destination, the network layer has to implement functions such as routing, addressing, and flow control. These functions convert the link packet pipe provided by the DLC layer to the network layer into an end-to-end packet pipe provided by the network layer to the higher layers. The network layer is considered the most complex layer of

the OSI model, since it requires the cooperation of many nodes (instead of the two end nodes, as is the case for the physical and the DLC layers of a point-to-point link).

The next layer in the hierarchy is the *transport layer,* which ensures end-to-end, reliable message transfer in a transparent way to its upper layers. Its services include message fragmentation and reassembly, and error recovery and flow control between endpoints; however, its functions and complexity vary among different network implementations and depend primarily on the reliability of the underlying network and of the lower-layer services.

The *session layer* establishes, manages, and ends the dialog between applications in end systems, thus providing a virtual session service to the higher layer. Typical functions at this layer include authentication, data grouping, and billing. It is generally considered as a rather "thin" layer.

The next higher layer is the *presentation layer, w*hich deals with the syntax of exchanged data and provides a common data representation to the user applications. Important services here include data encryption, data compression, and code conversion.

The uppermost layer of the OSI model is the *application layer*, which provides the interface of the OSI environment to user applications. It does all the remaining work that is not done by other layer. This layer provides end-application protocols like email, WWW applications (browsers), teleconferencing, FTP, Telnet, and chat services.

The layers of the OSI model are organized hierarchically so that each layer provides services to its next upper layer and is based on services provided to it by its next lower layer. Entities at the same layer between different systems ("peer" entities) communicate on the basis of a mutually agreed-on protocol. Peer entities do not communicate directly; instead, they pass their messages to their next lower layer, which then forwards them to the next lower layer, and so on, until they reach the physical layer, where they are transmitted to the interconnected system. When the messages are received at the other end of the communication medium, they travel all the way up the recipient's lower layers (*demultiplexing*), until they reach the recipient peer entity. This procedure is controlled by headers that are attached successively to each packet (*encapsulation*) by each layer as the packet crosses on its way downward (toward the physical layer) and then stripped off successively by each layer as the packet crosses on its way upward (toward the peer entity of the communication originator), as outlined in Fig. 2. Each layer has to perform its function based only on information included in the header of that layer and cannot look in the body of the message or on the headers added by other layers. This is helpful because if we need to replace the implementation of one layer, this will not affect the operations of other layers.

It is worth noting that the OSI model is not really a communication standard, but rather a framework, or guideline, for developing standards to enable the interconnection of heterogeneous computing systems. Two systems that adhere to some standard developed according to the OSI model will have the same number of layers implementing the same communication functions (although maybe in a different way) and follow a common protocol.

3. THE TCP/IP PROTOCOL SUITE

While the OSI architecture can be considered a guideline for developing communications standards, another firmly

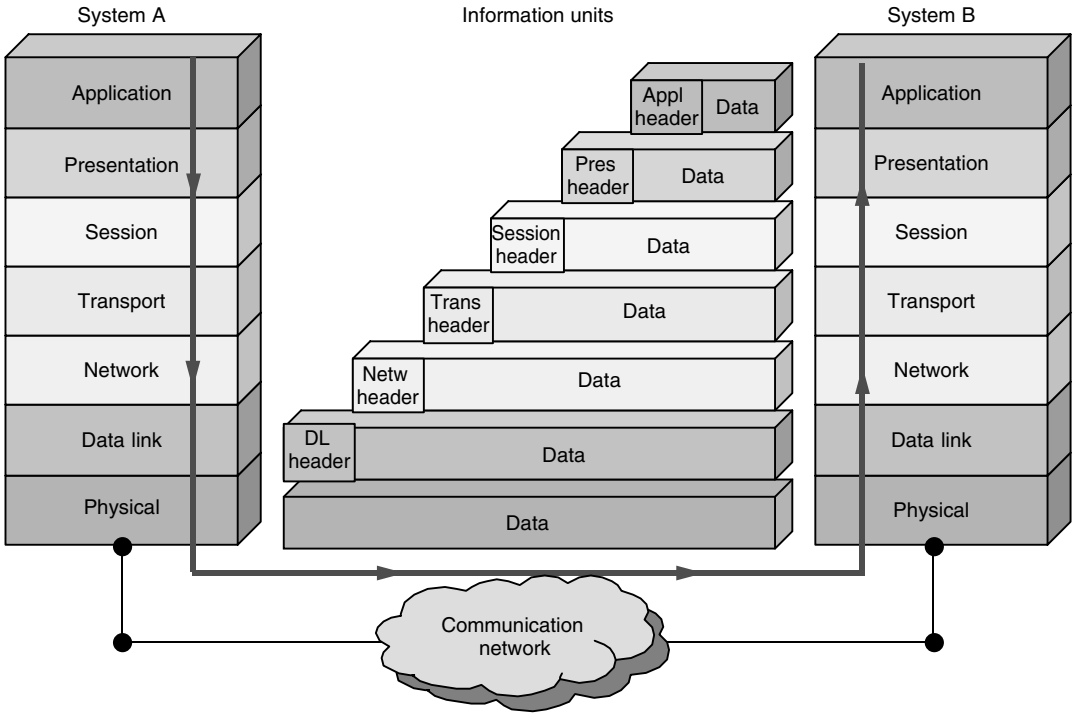


Figure 2. Computer communication through the OSI model.

established architecture has been extensively employed in computer communications. This architecture originates in the U.S. Department of Defense (DoD) efforts to support communications transparently among heterogeneous computer systems over heterogeneous physical communication networks, thus forming a "network of networks" that has nowadays evolved to what is best known as the "Internet." This architecture and its associated standards are collectively known as "TCP/IP."

The acronym *"TCP/IP"* (Transmission Control Protocol/Internet Protocol) does not imply just a pair of communications protocols, but rather refers to a large collection of interrelated protocols and applications (a *protocol suite*); the TCP and IP protocols are the more significant (and widely used) constituent parts of it. Because of its effectiveness and simplicity, the TCP/IP protocol suite has now become the dominant computer communications architecture and is almost synonymous with the terms "Internet" and "computer network."

3.1. TCP/IP Layering

The TCP/IP protocol suite architecture follows a reasoning similar to the one reflected in the OSI model — the whole communication task is divided into a number of smaller and more manageable subtasks, each subtask implemented by one or more protocols. These protocols can be organized conceptually as a hierarchical set of layers (a *protocol stack*), wherein each layer builds on its lower layer, adding new functionality to it. The TCP/IP protocol suite involves four such layers:

1. The application layer
2. The transport layer
3. The internetwork layer
4. The network interface layer

While the top layer (application layer) deals only with application details, the next three lower layers deal with the communication details of an application. Figure 3 illustrates the four layers of the TCP/IP protocol suite, as well as their functionally equivalent layers, in rough terms, in the context of the OSI Reference Model. The

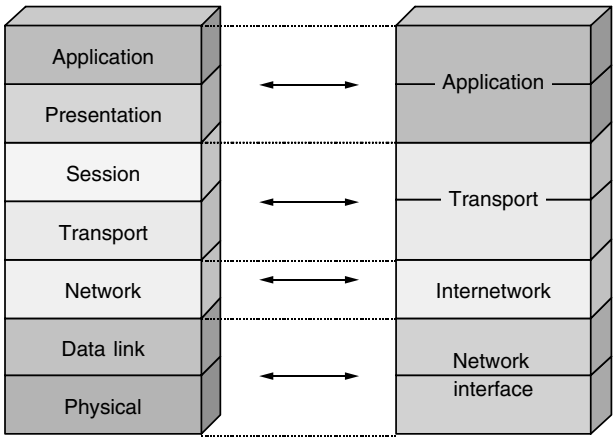


Figure 3. The TCP/IP protocol suite and its correspondence with the OSI Reference Model.

strict use of all layers is not mandated, and the layers are not explicitly stated in the standards.

The lowest layer in the hierarchy (*network interface layer* or *link layer*) forms the interface to the underlying network hardware and hides the network implementation details from the layers above. Protocols at this layer depend on the actual communication network to which the end system is attached and may provide services such as error control.

The purpose of the *internetwork* (or *network*) *layer* is to route data appropriately to the destination host, hiding from its higher layers the internetwork architecture layout between the hosts. The most important protocol at this layer is the Internet Protocol (IP), which provides connectionless services for end systems, without assuming reliability from lower layers.

The *transport layer* provides end-to-end data transfer between peer processes on different hosts. The most important protocol at this layer is the Transmission Control Protocol (TCP), which provides reliable connection-oriented services between peer processes.

Finally, the *application layer*, on top, contains network applications and services, as well as protocols for resource sharing and remote access that are needed to support the user applications. Important protocols at this layer are the Telnet protocol, the File Transfer Protocol (FTP), and the Simple Mail Transfer Protocol (SMTP).

Figure 4 shows a typical TCP/IP communications example between two application processes located at different hosts (end systems) and over two different networks interconnected with a router (an intermediate system).[1] Note in this figure that, while application- and transport-layer protocols are end-to-end, network as well as network interface-layer protocols are hop-by-hop protocols; that is, they are used between end systems and intermediate systems, or between intermediate systems across a collection of communication networks.

TCP/IP employs a two-level addressing scheme; at the low level, every host on an internetwork is assigned a unique (global) address (*IP address*), so that data can be routed to the correct host, while at the high level, every process residing on a host is assigned a unique (within the host) address (*port number*), so that data can be routed to the correct process.[2] Moreover, TCP/IP implements encapsulation and demultiplexing techniques similar to the ones used in the OSI model. On the transmit side, TCP accepts the bytestream from an application, segments it into small data blocks, and appends to each of them a header containing items such as the destination port, sequence number, and checksum for error detection. The resulting data unit, known as a *TCP segment*, is sent down to IP, which appends additional control information relevant to IP functionality (e.g., the destination host address). The resulting

[1] A *router* is a system that attaches to two or more (usually) different physical networks and forwards data packets between the networks. It consists of a network-layer protocol as well as a number of network interface protocols, depending on the actual communication networks it has to interconnect.

[2] The combination of an IP address and a port number is called a "socket" and can uniquely identify a service running on a host.

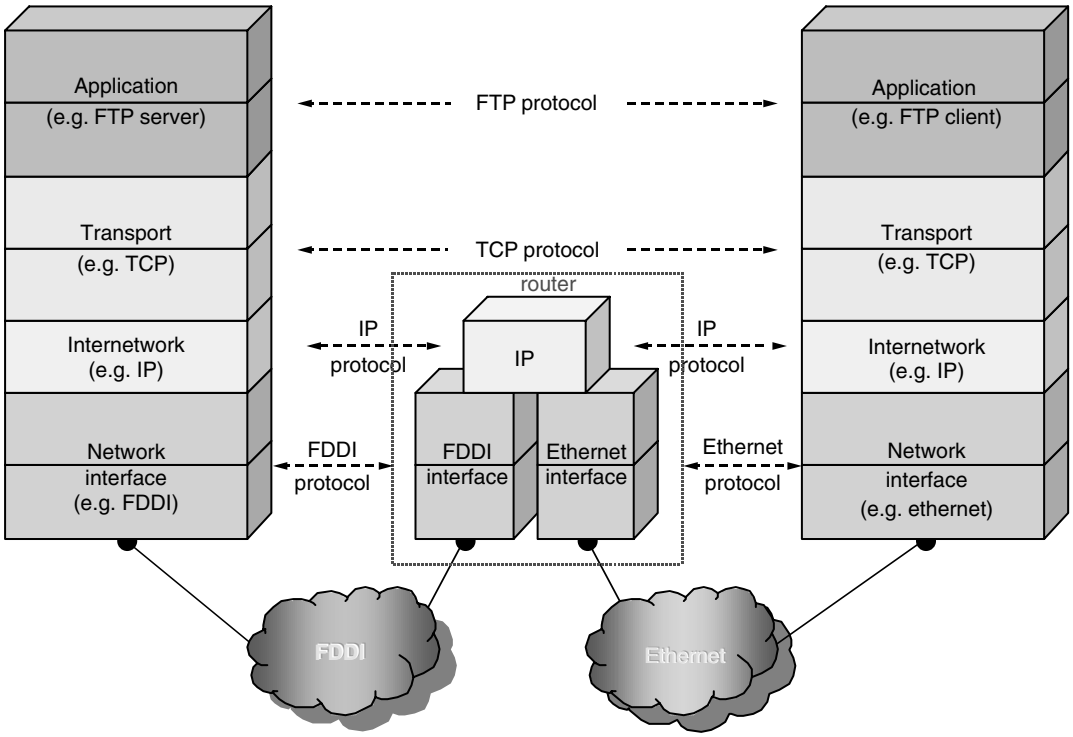


Figure 4. TCP/IP communications over two networks.

data unit, known as an *IP datagram*, is handed over to the network interface layer, which in turn adds its own specific header and transmits the resulting *frame* (which has now the structure shown in Fig. 5) across its attached communication network to the appropriate router. On the receive side, the reverse process takes place; headers are removed and used by the appropriate layers to control the communication procedure within the scope of their layer, until the original bytestream is delivered intact to the recipient application.

3.2. The Internet Protocol

The *Internet Protocol* (IP) is the most widely used internetworking protocol and provides connectionless and unreliable packet delivery services. "Connectionless" indicates that there is no predetermined route through the network between the endpoints but, rather, each packet works its way through the network independently and without any prior coordination. As a result, each packet of a session may follow a different path to reach the same destination, and therefore packets may be delivered out of order; additional services such as sequencing, if required, are dealt with at higher layers (e.g., by TCP). "Unreliable" suggests that IP makes a best effort to get a packet to its destination, but without providing any guarantee of actual delivery. Again, if reliability in the communication is required, it has to be addressed properly at higher layers.

An IP datagram consists of the IP header and the data. Its format is shown in Fig. 6, where the fields have the following meaning:

- *Version* (4 bits)—indicates the IP version number.
- *Header length* (4 bits)—indicates the length of the header (including options) in 32-bit words (i.e., rows in Fig. 6). Its normal value is 5, corresponding to a header length of 20 bytes (no options are used).
- *Type of service* (8 bits)—specifies delay, throughput, reliability, and precedence parameters to request a particular quality of service for the datagram.
- *Total length* (16 bits)—indicates the length of the total IP datagram in bytes. Thus, the maximum IP datagram size is 65,535 bytes (although, in practice, much smaller IP datagrams are used).
- *Identification* (16 bits)—a sequence number that uniquely identifies (along with the source address,

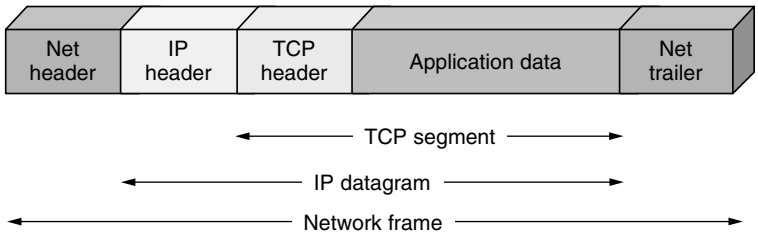


Figure 5. Encapsulation in a TCP/IP frame.

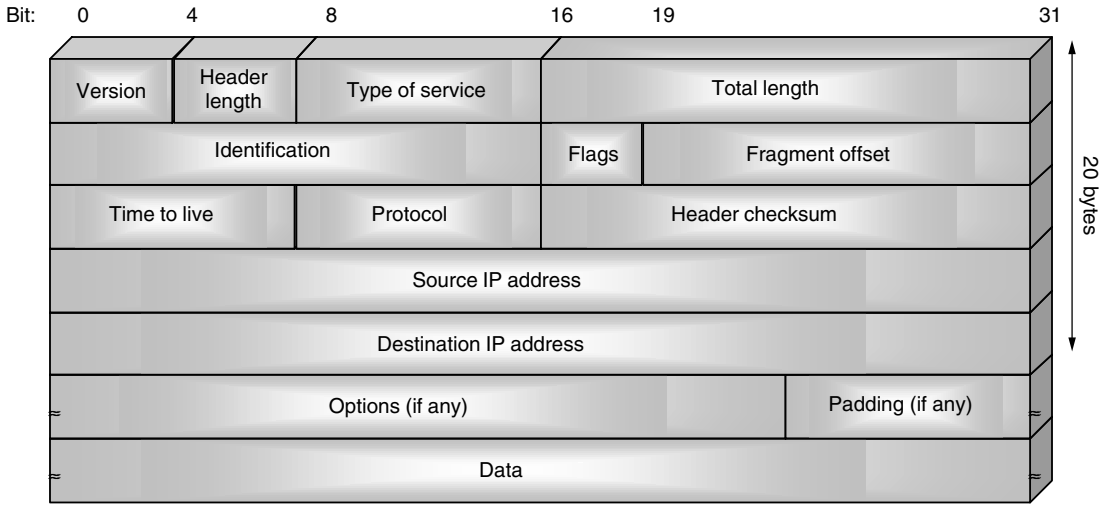


Figure 6. The format of an IP datagram.

the destination address and the protocol used) each datagram transmitted by a sender. It is used for the fragmentation and reassembly of an IP datagram.

- *Flags* (3 bits)—only 2 bits are defined. When an IP datagram is fragmented, the "more" bit is set for each fragment (except the last one). When the "don't fragment" bit is set, the IP datagram will not be fragmented.
- *Fragment offset* (13 bits)—when fragmentation is used, it indicates the offset (in 8-byte units) of this fragment from the beginning of the original IP datagram.
- *Time to live* (8 bits)—is used to prevent IP datagrams from getting caught in routing loops, by specifying the amount of time that a datagram can stay in an internetwork. Practically it has a value of 32 or 64 and is equivalent to a hop count (see Section 5.2 below).
- *Protocol* (8 bits)—indicates the higher-level protocol that has to receive the IP datagram after demultiplexing at the receiver.
- *Header checksum* (16 bits)—is used for error detection and covers only the IP header. It is maintained at each router the datagram comes through.
- *Source address* (32 bits) and *destination address* (32 bits)—indicate the IP addresses of the originator and the recipient(s) of the IP datagram (see Section 5.1.2).
- *Options* (variable length)—indicates the options requested by the sender (e.g., security restrictions, timestamping, route recording). This value is padded by 0s as necessary, to ensure that the header length is a multiple of 32 bits.

3.3. The Transmission Control Protocol

The next widely used protocol of the TCP/IP suite is the *Transmission Control Protocol* (TCP), which builds on the services provided by IP and provides additional functionality to application processes, such as reliable data transfer, error control, and flow control. TCP is implemented in edge systems, and its task is to transform the unreliable delivery service provided by IP into a reliable "connection-oriented" data transmission system, suitable for building network applications. "Reliable" suggests that packets are guaranteed to reach their destination, error-free and in the correct order, while "connection-oriented" indicates that a logical connection is established between the endpoints prior to any data transfer between them and lasts for the duration of a session. Since the actual data packets are transferred by the underlying IP protocol in a connectionless mode, TCP does not in fact set up predefined paths across which all data flow during a given session but, instead, establishes a tight relationship between the end hosts, which can be logically considered as a "virtual circuit." When communication is desired, the initiating TCP first sends a special connection request segment, and awaits a connection response. When it arrives, the initiating TCP confirms connection establishment and begins the reliable communications described earlier.

A TCP segment consists of the TCP header and the data. Its format is shown in Fig. 7, where the fields have the following meaning:

- *Source port* (16 bits) and *destination port* (16 bits)—indicate the TCP port numbers of the source and destination applications.
- *Sequence number* (32 bits)—since two applications exchange a stream of bytes across a TCP connection, the byte in the stream the first data byte in this segment represents is identified by the *sequence number* field.
- *Acknowledgment number* (32 bits)—indicates the sequence number of the next data byte that the sender application is ready to receive (implying that all data bytes with previous sequence numbers were successfully received).
- *Header length* (4 bits)—indicates the length of the header (including options) in 32-bit words (rows in Fig. 7).

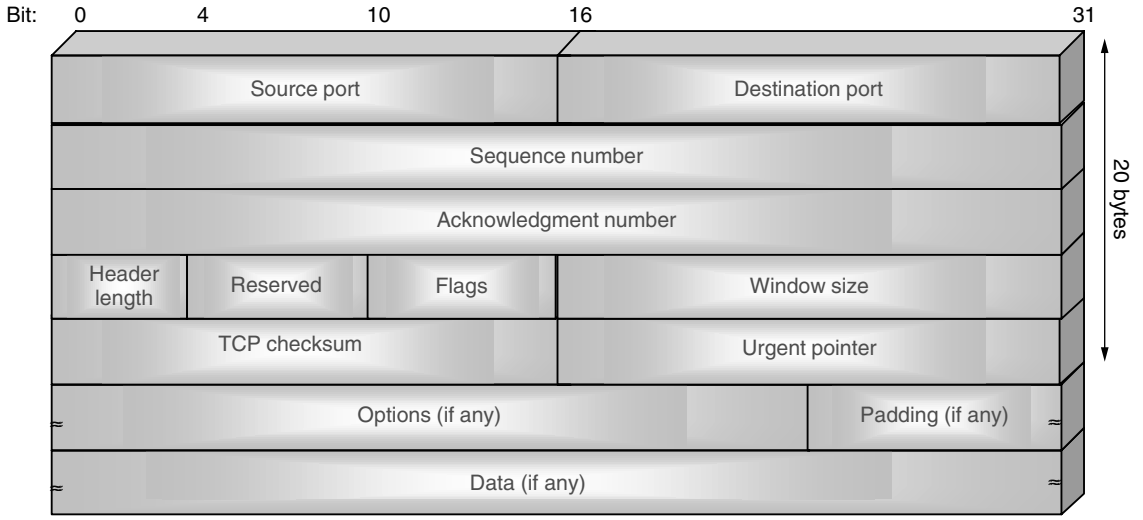


Figure 7. The format of a TCP segment.

- *Reserved* (6 bits) — these bits are reserved for future use.
- *Flags* (6 bits) — there are 6 flag bits (viz., URG, ACK, PSH, RST, SYN, FIN) that are used to control various functions in a TCP connection.
- *Window size* (16 bits) — used for the *flow control* function and indicates the number of data bytes that the sender is willing to accept (starting with the one specified by the acknowledgment number field).
- *TCP checksum* (16 bits) — used for error detection and covers both the TCP header and the TCP data. This checksum is calculated by the sender and verified by the receiver.
- *Urgent pointer* (16 bits) — the sender can transmit urgent data to the receiver, and this field points to the last byte of such data, delimiting in this way the urgent bytes in the bytestream of the segment.
- *Options* (variable length) — indicates the options requested by the sender (e.g., maximum segment size). This value is padded by 0s as necessary, to ensure that the header length is a multiple of 32 bits.

In TCP/IP, the network is simple, and doesn't guarantee anything, except a high probability of packet delivery. The complexity is in TCP, which exists only in edge systems. The edge systems themselves are powerful computers — sufficiently powerful to run TCP. We can say that the end user provides the complexity, while the Internet provides a basic service. This is in contrast to the telephone network, which was designed to provide an expensive (circuit-switched) and reliable service, but whose end systems are extremely simple (telephones).

3.4. The User Datagram Protocol and Some Basic Applications

Another important protocol of the TCP/IP suite is the *User Datagram Protocol* (UDP) at the transport layer. UDP is an alternative to TCP host-to-host protocol that adds no reliability, sequencing, error control, or flow control to IP. These functions, instead, are accommodated by the applications on top of UDP. UDP is connectionless, and its header consists of four 16-bit fields: the source and destination ports, the total length of the UDP segment, and the checksum applied to the entire UDP segment. UDP is used mainly for the exchange of self-contained data packets (or datagrams) between application processes and is extensively employed in real-time applications, like VoIP (Voice over IP) and Net audio.

The following are some basic applications that are part of the TCP/IP suite and exploit the characteristics of TCP and UDP:

1. Telnet (Telecommunications Network), a client/server application that supports terminal emulation and remote logon capability over a TCP connection.
2. FTP (File Transfer Protocol), a client/server application that enables the exchange of text and binary files between host computers, under user command. It uses two TCP connections: one for control messages and the other for the actual file transfer.
3. SMTP (Simple Mail Transfer Protocol), which supports basic electronic mail in text form over a TCP connection. MIME (Multipurpose Internet Mail Extension) is an SMTP extension that provides support for the attachment of other file forms, including audio, graphics, and video.
4. SNMP (Simple Network Management Protocol), which enables the exchange of network management information between host computers (agents) and an SNMP manager, over UDP.

4. DATA-LINK CONTROL PROTOCOLS

The physical layer of the OSI model, as we have already discussed, is responsible for the transmission of a raw bit stream over the physical communication medium. However, to manage this raw bit stream and render it useful, some form of control has to be exercised on top of the physical layer, to account for issues such as link

frame synchronization, link flow control, error control, and retransmissions over a link. This level of control is referred to as *data-link control* and is implemented by protocols residing functionally in the *data-link* layer of the OSI model.

4.1. Flow Control

Since computers that communicate across a network can vastly vary in terms of capabilities, a mechanism has to be set up to account for such inequalities. For example, a fast (high-speed) sender should be prevented from overwhelming with data a slow receiver; otherwise the latter (and maybe the network between them) will become congested. *Flow control* is the function that assumes responsibility for such congestion occurrences and there are two common mechanisms in this direction, known as *stop-and-wait* and *sliding-window*. Both mechanisms are based on the principle that the sender may transmit only when the receiver permits so.

Stop-and-wait is a simple mechanism, according to which a sender, after transmitting a frame, stops transmission and waits until the receiver "acknowledges reception" of the frame. In this way, a slow receiver can avoid getting flooded with frames by sending back reception acknowledgments only after having processed the received frame. In cases, however, where the propagation time over the communication medium is high (e.g., a satellite link), an extended version of the stop-and-wait mechanism is used, the sliding window mechanism, according to which the sender requires an acknowledgment from the receiver after a certain number of frames (determined by the *window size*) have been transmitted. This way, several frames can be in transit at the same time, increasing the utilization of the communication link.

4.2. Error Control

The error control function deals with the detection and, in some cases, correction of errors that occur during the transmission of frames (resulting in lost or damaged frames). Errors are detected at the cost of increasing the frame length by adding some extra information at transmission, which is used by the receiver to either correct the error by itself (*forward error correction*), or ask for retransmission of the frames found in error (*backward error correction*). In most cases, backward error correction is more efficient and there are three common mechanisms in this direction (collectively known as *automatic repeat request* (*ARQ*)): *stop-and-wait ARQ*, *go-back-N ARQ*, and *selective-reject ARQ*. All these mechanisms are based on the flow control techniques described previously.

Stop-and-wait ARQ is based on the principle of the stop-and-wait flow control technique, where the sender transmits a single frame and then waits to get back an acknowledgment for this frame. Only a single frame can be in transit at any one time. If the sender receives either a negative acknowledgment from the receiver or no acknowledgment at all after a certain time period, it retransmits the same frame.

The go-back-*N* ARQ improves the efficiency of the stop-and-wait ARQ but increases its complexity. It is based on the sliding-window technique, where the sender buffers the transmitted frames (up to a number determined by the window size) until an acknowledgment is received. The acknowledgment is cumulative; that is, an ACK for frame *n* acknowledges reception of all previously transmitted frames. If the sender receives a negative acknowledgment for one frame, or no answer at all after a certain time period, then it retransmits all buffered frames from the not-acknowledged frame on.

In the selective-reject ARQ, the go-back-*N* ARQ principle is further refined, by limiting retransmission to only the not-acknowledged (i.e., rejected or timed out) frames. In this way, the amount of retransmission is decreased, at the cost, however, of increasing the receiver's buffer size and both sender's and receiver's complexity.

4.3. Synchronization and Framing

For successful data exchange between computer systems, some form of control has to be imposed on the sequence of bits transmitted over a communication medium, in terms of determining the boundaries between successive frames (*framing*) and maintaining common timing parameters (duration, rate, spacing) for the frame bits between the sender and the receiver (*synchronization*). Synchronization at the data-link layer can be achieved by either asynchronous or synchronous transmission schemes.

In *asynchronous transmission*, data are transmitted one character (5–8 bits) at a time. The beginning and the end of a character are indicated by a start and a stop bit, respectively, while a parity bit before the stop bit is often added for error detection purposes (see Fig. 8). When no character is being transmitted, the communication link is in idle state and, thus, the receiver can resynchronize at each start bit (the beginning of a new character). Asynchronous transmission is widely used (e.g., at PC interfaces with low-volume and low-speed transmissions) because it is simple to implement and the relative equipment is inexpensive. Nevertheless, by using many extra bits per character, it increases the communication overhead, wasting about 20–30% of the available medium bandwidth.

Synchronous transmission, on the other hand, deals with blocks of data, avoiding the overhead of the start/stop bits around each character and providing alternate means (*pre-* and *postamble* bit patterns) to delimit a frame at both ends. In this case, the frame is treated as either a

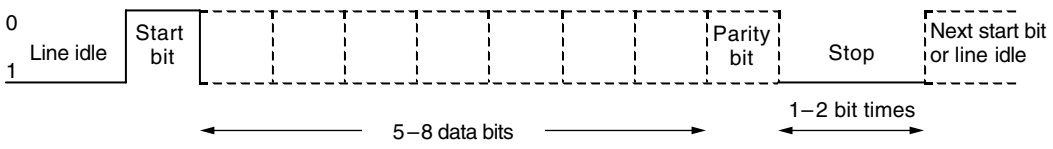


Figure 8. Character format in asynchronous transmission.

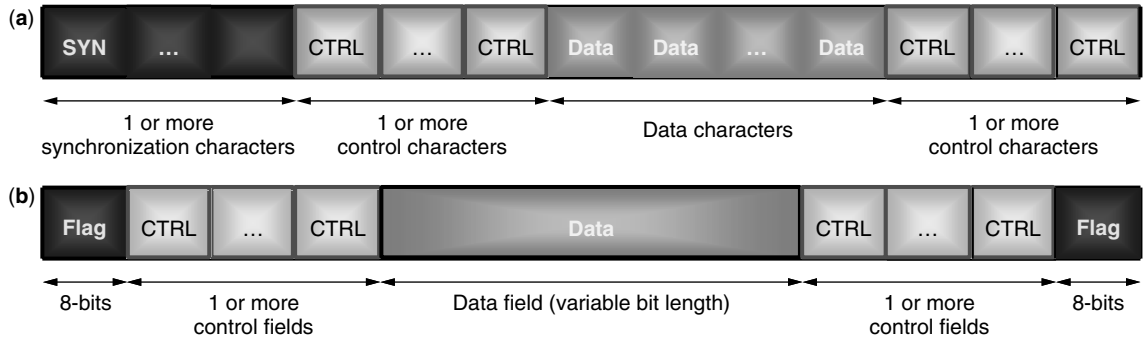


Figure 9. (**a**) Character-oriented and (**b**) bit-oriented frame format in synchronous transmission.

sequence of characters (*character-oriented* transmission) or a sequence of bits (*bit-oriented* transmission), as shown in Fig. 9. Bit-oriented protocols do not depend on specific codes (e.g., ASCII) for line control — as it is required with the character-oriented protocols — and hence they are nowadays widespread. Because of its low overhead (typically ~5%), synchronous transmission is generally used for large-volume and high-speed transmissions.

4.4. The High-Level Data Link Control (HDLC) Protocol

We pause for a moment to examine the most important Data Link Control protocol, the *High-level Data Link Control protocol* (HDLC), which has been standardized by ISO and constitutes the basis for almost all the bit-oriented protocols at the data-link layer that are in common use today: LAP-B (in the X.25 packet-switching network protocol suite), LAP-D (in the ISDN protocol suite), SDLC (in the SNA protocol suite), and IEEE 802.2 LLC (in the IEEE 802 LAN protocol suite). HDLC uses synchronous transmission and supports both full- and half-duplex communications in point-to-point and multipoint link configurations.

To organize data communications between hosts, HDLC defines two basic types of stations: the primary and the secondary stations. The *primary* station is responsible for controlling the data link (e.g., error/flow control) and transmits frames called *commands*. The secondary station responds to commands from a primary station by transmitting frames called *responses*. In a conversation, there is one primary and one or more secondary stations involved. A third type of station, the *combined* station, can also be defined, which combines the features of the primary and the secondary station.

HDLC operates in three data transfer modes:

- *Normal response mode* (NRM), in which secondary stations can transmit only in response to a poll from the primary station. This mode is used in point-to-point or multipoint link configurations (e.g., one host with many terminals attached).
- *Asynchronous response mode* (ARM), in which a secondary station may transmit without explicit permission from the primary station. This mode is used in special cases.
- *Asynchronous balanced mode* (ABM), in which all stations have equal status and may initiate transmission without receiving permission from the other station. This mode is used mainly on full-duplex point-to-point links.

Communication between primary and secondary stations in HDLC is achieved by exchanging frames (commands and responses) that fall in one of the following categories:

- *Information frames*, which are sequentially numbered and contain the user data and, sometimes, piggybacked error and flow control data (using the ARQ mechanism)
- *Supervisory frames*, which contain error and flow control data (not piggybacked as before)
- *Unnumbered frames*, which do not have sequence numbers, but are used for miscellaneous link control and management functions

An HDLC frame consists of a three-field header, the payload, and a two-field trailer, and has the general structure shown in Fig. 10. The *flag* is a special 8-bit sequence (binary 01111110) that identifies the start and the end of the frame (the end flag of one frame can also constitute the start flag of the next consecutive frame). Since the flag is essential to achieve frame synchronization, it is necessary to ensure that the flag sequence does not accidentally appear anywhere in the frame between start and

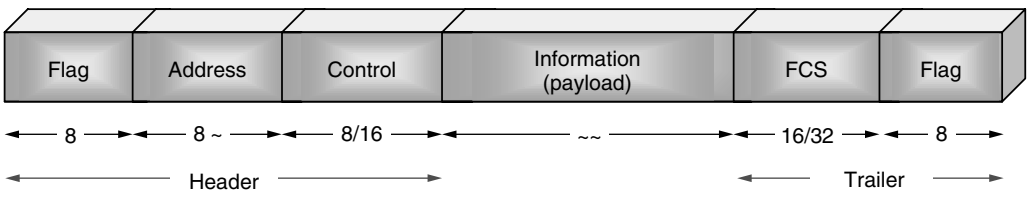


Figure 10. HDLC frame structure and fields' length (in bits).

end flags. To that end, a technique known as *bit stuffing* is used that inserts at transmitter (and deletes at receiver), an additional 0 (zero) bit, after each occurrence of five consecutive 1s in the data between the start and end flags of a frame. The *address* field is usually 8 bits long, but an extended format may also be possible. Depending on the station type of the issuer of the frame, the address field may indicate the sender (in responses) or intended receiver (in commands) of the frame. The *control* field is usually 8 bits long, but an extended format may also be possible. It is used to indicate the type of the frame (information, supervisory, or unnumbered) and it also contains some control information, according to the frame type (e.g., send sequence number for information frames).

The *information* field is the "payload" of the frame that contains the actual user data (any sequence of bits) coming from the higher layers. Its length, though undefined, is usually a multiple of 8 bits. Finally, the *frame-check sequence* (FCS) field goes after the information field and contains a 16- or 32-bit cyclic redundancy check (CRC) value that is used by the receiver to check the address, control, and information fields of the frame for errors. Practically, 16-bit CRC (2 bytes) is used for frames up to 4 KB (kilobytes).

4.5. Local-Area-Network Protocols

Due to the widespread use of local-area-networks (LANs),[3] a great number of standards and protocols have emerged that relate to LAN technology and belong functionally to the data-link layer of the OSI Reference Model. In this context, the data-link layer is usually divided into two sublayers: the *logical link control* and the *medium access control* sublayers.

Protocols in the logical link control sublayer are similar to typical data-link control protocols, like the HDLC. The most commonly used protocol at this level is LLC, described in the IEEE 802.2 standard. LLC supports both connectionless and connection-oriented services and provides mechanisms for addressing and data-link control. Moreover, it is independent of the underlying network topology, medium access control technique, and transmission medium.

Since all devices interconnected via a LAN share a common physical communication medium with a fixed capacity, some procedure has to be established to control access to the underlying medium and provide for an efficient usage of its capacity. Such procedures and techniques are provided by protocols at the medium-access control (MAC) sublayer, which, apparently, are closely dependent to the transmission medium and the network topology used. Common MAC protocols are the *carrier sense multiple access with collision detection* (CSMA/CD), the *token bus* and the *token ring,* each specified in a number of widely used standards, such as IEEE 802.3, IEEE 802.4, and FDDI.

[3] A LAN is a high-speed data network, optimized for use over small geographic areas (like buildings), that usually interconnects personal computers, peripheral devices, and other equipment over a common (shared) physical communication medium.

CSMA/CD falls under the category of contention techniques, in which all communicating parts "contend" for access to network resources. The precursor of CSMA/CD is ALOHA, which was developed in the 1970s for packet radio networks. According to ALOHA, every station transmits its packet on the air whenever it is ready to do so, without any prior coordination with the other stations in the network. In this fashion, collisions between packets of different stations that are transmitted simultaneously are quite common, especially as the number of stations increases. In case of a packet collision, the transmitting station waits for a time interval and then attempts retransmission of the packet. ALOHA is a very simple to implement technique, with a limited efficiency, however. CSMA improves this situation by requiring that each station has to listen to the medium first ("sense the carrier"), and if no other station is transmitting at that time, it can go on transmitting its own packet, or else it has to wait. CSMA can be further improved by having the transmitting station listen to the medium while transmitting, to notice early any possible collision ("collision detection") and stop its transmission (and thus free network resources). The CSMA/CD medium-access technique is used in Ethernet/IEEE 802.3 LANs, over coaxial cable or unshielded twisted pair as transmission media and on bus, tree, or star network topologies.

In the token-passing schemes, network devices access the physical medium based on the possession of a token that circulates when stations are idle. A token is like a permission to transmit, and whoever has the token can transmit or pass it to the next node. Examples of LANs that use such techniques are the *Token Bus* (IEEE 802.4), over coaxial cable or optical fiber and on bus, tree, or star network topologies; the *Token Ring* (IEEE 802.5), over shielded/unshielded twisted-pair cable and on ring topology; and the *Fiber Distributed Data Interface* (FDDI), over optical fiber and on ring topology.

5. INTERNETWORKING PROTOCOLS

In an internetworking environment, where a number of communication networks are interconnected to provide data transfer among the hosts attached to them, an addressing scheme has to be established so that all communicating entities are uniquely identified in the internetwork, to allow data to be directed (or *routed*) to the intended destination. In this section we focus on the protocols and mechanisms behind these two fundamental functions in internetworking: *addressing* and *routing*.

5.1. Addressing and Naming

In the global postal system, a postal carrier knows exactly where to deliver a letter, based on an agreed-on addressing scheme, according to which a house location can be uniquely determined, by specifying some positional parameters (country, city, street, number, etc.) that altogether constitute the recipient's "home address," which is hierarchical, and is written on the envelope of the letter. Addressing in an internetworking environment follows a similar concept; each entity in an internetwork can be reached by means of a unique (global) address, which is usually a number like a telephone number in telephony. Since, however,

humans cannot easily remember numbers, especially in the binary format that computers understand, translation schemes between symbolic names and addresses are practically used that match high-level human-intelligible names to machine-intelligible internetwork addresses and vice versa. This translation process, called *address resolution*, is usually performed in a distributed fashion by special network servers, called *name servers*. In the Internet context, address resolution is performed by the *Domain Name System* (DNS), which is a global network of name servers implementing a distributed database.

5.1.1. The Domain Name System. According to DNS, symbolic names are grouped in domains that are hierarchically organized, like the chain of command reflected in the organization chart of a big company. Within each domain there is one primary (*authoritative*) name server that has the responsibility of performing address resolution through the maintenance of a local database. It is, however, common for domain name servers to delegate authority for their subdomains to other name servers, therefore defining subareas of responsibility. The domain (and subdomains, if any) for which a name server is authoritative is referred to as "zone of authority." Consequently, DNS logically interconnects all name servers into a hierarchical tree of domains and uses a hierarchical naming structure, like the one used in telephone numbers, in which names consist of a sequence of fields [typically a top-level domain,[4] a domain, subdomain(s), and the host name] that jointly identify the entity. For example, *talos.telecom.ntua.gr* refers to the host named *talos* in the Telecommunications Laboratory (*telecom*) subdomain of the National Technical University of Athens (*ntua*) domain, in Greece (*gr*).[5]

Since DNS is a distributed database, a name server doesn't need to know all the names and addresses of hosts in the Internet; its area of knowledge and responsibility is usually confined to its own zone of authority. It does need to know, though, the name servers who are responsible for other domains. When, for instance, we type *www.cis.ohio-state.edu* on our Internet browser, a module called *resolver* tries to look up the requested address locally, if a local name cache is kept. Otherwise, it sends a DNS query to the local name server. The local name server will then follow these steps:

1. Check to see if it already knows (from its database or from a previous query) the address of *www.cis.ohio-state.edu*. If so, it finds the address in its database or cache and replies to the query. The browser then uses this (IP) address to request a connection to the host containing the desired Webpage.
2. If the local name server cannot resolve the address by itself, it will query one of the "root" servers, at the top of the DNS hierarchy, whose addresses are definitely known to all name servers, for the address of *www.cis.ohio-state.edu*.
3. Today, there are 13 root servers on the Internet, and each of them knows the IP addresses of the servers for all the top-level domains (*.com*, *.edu*, *.uk*, etc.). So, ultimately, the root server contacted will refer our name server to (i.e., give the address of) a list of *.edu* name servers.
4. Our name server will query one of the *.edu* servers for the address of *www.cis.ohio-state.edu*.
5. The *.edu* server queried will refer our name server to a list of name servers for the domain *ohio-state.edu*.
6. Our name server will then query one of the *ohio-state.edu* name servers for the address of *www.cis.ohio-state.edu*.
7. If the *ohio-state.edu* name server queried knows the address of *www.cis.ohio-state.edu*, it returns that address to our name server. If there is another name server responsible for the *cis* (delegated) subdomain, the address of that name server will be returned to our name server, which will in turn query that server for the address of *www.cis.ohio-state.edu*.
8. Finally, our browser will be given the requested (IP) address or an error if the query could not be answered.

This is a worst-case scenario that can take many seconds to complete, but actually things are simpler, since, as it was mentioned in the above process, each name server caches for a certain time period all the information it retrieves this way. This way, a huge load is removed from the root and top-level servers.

5.1.2. IP Addresses. IP addresses are 32 bits in length. They are typically written in a format known as "dotted decimal notation," according to which each of the 4 address bytes is expressed in its decimal equivalent value (0–255) and the four values are separated by periods, for example, "147.102.105.18."

IP addresses generally consist of two parts:

- The *network identifier* identifying the TCP/IP subnetwork to which the host is connected
- The *host identifier* identifying a specific host within a subnetwork

To account for networks of different sizes, IP defines several *address classes,* characterized by the length of the network identifier. Class B addresses, for instance, have a 14-bit network identifier and a 16-bit host identifier and thus can address up to 65,536 (2^{16}) hosts per network; therefore, class B addresses are used for moderate-sized networks. Five address classes are defined (A through E), while only classes A, B, and C are used for host addressing.[6]

5.1.3. Address Resolution Protocol (ARP). The *address resolution protocol* is used in specific network implementations where the Internet Protocol (IP) is applied

[4] For instance, *com*, *edu*, *org*, *gov*, or two-letter country codes such as *uk* (United Kingdom) and *it* (Italy).

[5] It is not, however, necessary for a host to be physically located at the country specified by the country code.

[6] Class D addresses are used for IP multicasting, while class E addresses are reserved for future use.

over Ethernet or token ring local-area networks (LANs). Addressing in such LANs is twofold; at the IP level, every entity in the network is assigned a unique IP address (which is contained in the IP datagram) and, at the medium-access control (MAC) level, every entity in the LAN has a unique MAC (hardware) address (6 bytes in length, which is placed in the MAC frame).

An entity's IP address is different from its MAC address. Therefore the address resolution protocol is used to translate between the two types of address, so that a sender's IP process can communicate with the intended receiver's IP process on the same network, when knowing only the receiver's IP address. The process is relatively simple: (1) the sender transmits an ARP Request message using the hardware broadcast address (so that the intended receiver will surely receive the request); (2) the ARP Request advertises the destination IP address and requests the associated MAC address; (3) the intended receiver recognizes its own IP address and forms an ARP Response that contains its own MAC address; (4) since the original Request also includes the sender's MAC address, this address is used by the receiver to unicast its ARP Response to the original sender. To reduce the number of ARP requests in a LAN, hosts maintain a cache of recently resolved addresses. ARP cache entries timeout at regular intervals (typically 20 min), and new queries are made so that changes to the network topology are properly handled.

5.2. Routing

An internetwork can be considered as a collection of communication networks. Attached to each communication network we find devices (usually computers) that support end-user applications or services. These devices are called *end systems*, as opposed to the *intermediate systems* that are used to interconnect the communication networks (and thus form an internetwork), so that end systems attached to different networks can communicate across the internetwork, share information, and provide services to each other.

On their way toward their intended recipient, data packets travel across multiple networks and reach several intermediate systems, called *routers,* that forward them to the next intermediate system/communication network. The process of moving data from source to destination across an internetwork is known as *routing*. Routing includes two basic functions: *optimal path determination* and *switching*.

The optimal path toward an internetwork destination is determined by special algorithms, called *routing algorithms*. These algorithms perform calculations based on multiple metrics, such as pathlength, packet delay, and communication cost. The results of these calculations are used to populate *routing tables* that are maintained within each router and list the next "hop" (i.e., the next router) to which a data packet should be sent on the (optimal) way to its destination. Therefore, routers forward a received data packet according to the packet's destination address and the association for this address that the routing table indicates. The packet's physical address is changed to the next router's physical address, and the packet is transmitted. This way, the packet is switched hop by hop through the internetwork, until it finally reaches its destination.

This routing process, however, can be processor-intensive. In a more efficient alternative, called *multiprotocol label switching* (MPLS), optimum end-to-end paths through the network are calculated in advance (at the network edge) and appropriate routing information is appended as a *label* between the layer 2 and layer 3 packet headers;[7] then, routers along the path use the information in this label to simply switch the packet to the next hop. MPLS is also employed in traffic engineering (i.e., establishing traffic patterns that balance overall network resources utilization) and quality-of-service (QoS) routing (i.e., selecting routes that provide a desirable level of service, e.g., bandwidth, latency, priority requirements). MPLS supports many protocols of the network layer (including IPv4, IPv6, and AppleTalk), as well as the link layer (e.g., Ethernet, token ring, ATM).

5.2.1. Communication Between Routers. Routing algorithms calculate optimal paths on the basis of the routing information available at the router at a certain point in time. This routing information reflects the status of the networks (reachability, traffic delays, etc.) and is exchanged between routers by certain protocols, called *routing protocols*. We can distinguish two broad categories of routing protocols: *interior router protocols* (IRPs) and *exterior router protocols* (ERPs), if they are used between routers within the same or in different *autonomous systems*,[8] respectively. Practically, IRPs exchange a wealth of routing information to help determine optimal paths within an AS, while ERPs exchange only summary reachability information between ASs and, therefore, are simpler than IRPs. Examples of IRPs are the *Routing Information Protocol* (RIP) and the *Open Short Path First* (OSPF), while *Border Gateway Protocol* (BGP) is a typical ERP.

RIP belongs to the distance-vector class of protocols, according to which neighboring routers send all or a portion of their routing tables by exchanging routing update messages periodically and when network topology changes occur. Whenever a router receives a routing update message, it updates its routing table in light of the new information and subsequently transmits new routing update messages (using UDP/IP) to propagate the network changes across the network. RIP uses the "hop count" metric to measure the pathlength between a source and a destination. The maximum number of hops in a path is limited to 15; a value of 16 implies that the host is unreachable. RIP is still used efficiently today in small ASs.

OSPF is a nonproprietary protocol and belongs to the link-state class of protocols, according to which a router broadcasts to all routers only the portion of its routing table that describes the status of its own links. Such updates (using IP directly) produce minimum traffic and

[7] Or, in the virtual path identifier/virtual channel identifier (VPI/VCI) cell header fields, in ATM networks (see Section 6).

[8] An *autonomous system* (*AS*) is a logical portion (i.e., a collection of routers and subnetworks) of a larger internetwork that constitutes a distinct routing domain (and usually corresponds to commercial or administrative entities). An AS is managed by a single authority and may connect to other AS, as well as other public or private networks.

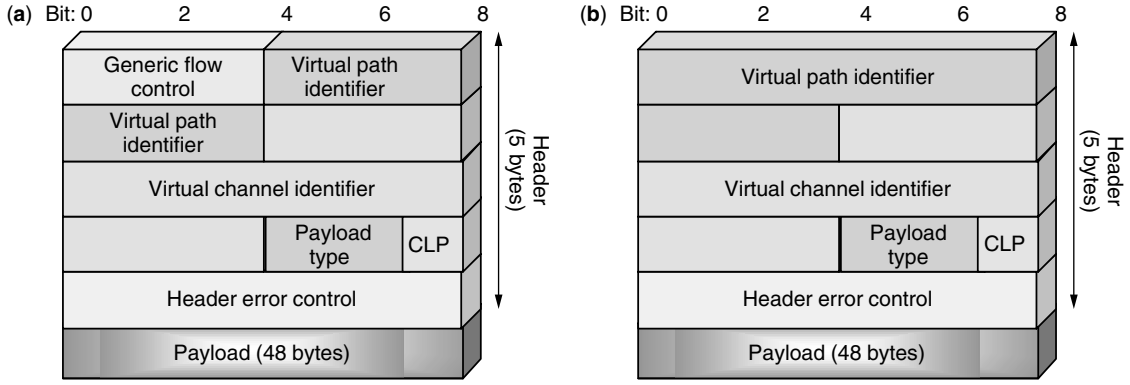


Figure 11. ATM cell format: (**a**) at the user–network interface; (**b**) at the network–network interface.

hence require less network bandwidth. Unlike RIP, OSPF is hierarchical; an AS is partitioned in a number of smaller groups of networks, called *areas*, and an area's topology is invisible to entities outside it. Routing between areas (interarea routing, as opposed to intraarea routing, which takes place within the same area) is handled by an OSPF backbone of routers. The backbone topology is invisible to all intraarea entities. OSPF uses the "shortest path first" or Dijkstra's algorithm to determine optimal paths. OSPF is more efficient and more scalable than RIP and for this reason it is widely used in TCP/IP networks and the Internet.

BGP is a scalable protocol used for the exchange of routing information between different ASs and is widely used on the Internet between the Internet service providers. BGP is a kind of distance vector protocol, with the difference that BGP neighbors don't send routing updates periodically, but exchange full routing information at the beginning (using TCP) and then send updates only about the routes that have changed. Also, a router's BGP table stores the networks that it can reach and the best route to each reachable network. BGP version 4 supports also policy-based routing based on security, legal, economic, and other issues rather than purely technical ones.

6. ASYNCHRONOUS TRANSFER MODE

Asynchronous transfer mode (ATM), also known as *cell relay*, is a high-speed packet transfer technology that tries to combine the benefits of packet switching (efficiency and flexibility for data transfer) with those of circuit switching (guaranteed bandwidth and transmission delay). ATM organizes data in packets of fixed size, called *cells*; this way the processing overhead at switching nodes is minimized. Also, since ATM takes advantage of the inherent dependability of modern communication systems, it uses minimal error control, thus further reducing the processing overhead of ATM cells and allowing for very high data transfer speeds.

The ATM cell consists of a 5-byte header and a 48-byte information (data) field, for a total cell length of 53 bytes. The cell header format is slightly different at the user–network interface and inside the ATM network, as shown in Fig. 11 parts (a) and (b), respectively. The header fields have the following meaning:

- *Generic flow control* (4 bits)—this field is present only at the user–network interface and provides some form of local cell flow control.
- *Virtual path identifier* (VPI) (8/12 bits)—identifies the virtual path a virtual channel belongs to and is used as a routing field in an ATM network. The VPI field is longer (12 bits) at the network–network interface, allowing for a large number of virtual paths to be supported inside the ATM network.
- *Virtual channel identifier* (VCI) (16 bits)—identifies a virtual channel inside a virtual path and is used as a routing field in an ATM network.
- *Payload type* (3 bits)—indicates the type of information contained in the cell's payload (control or user data) and provides some additional control information.
- *Cell loss priority* (CLP) (1 bit)—indicates the priority of the cell in case of network congestion. A cell with a CLP value of 1 is considered of low priority and subject to discard if required by network conditions.
- *Header error control* (8 bits)—used for error detection and single-bit error correction; it covers only the (first 4) bytes of the cell header.

ATM is *connection-oriented*, that is, a logical connection (called *virtual circuit* or, in ATM terminology, *virtual channel*) is established between two end stations prior to data transfer, and all packets follow the same preplanned route in the network and arrive at their destinations in sequence. ATM also defines another type of logical connection, the *virtual path*, which is essentially a bundle of virtual channels that share a large part of their path. Cell routing in an ATM network is based on both the virtual path and the virtual channel identifiers (VPI and VCI) contained in the cell header.

During a virtual channel connection (VCC) setup, a user can specify a set of parameters relating to the desired *quality of service* (QoS) and the input *traffic characteristics* of the VCC. QoS parameters include the *cell loss ratio* (CLR) (i.e., the percentage of cells that are lost in the network—due to error or congestion—to

the total transmitted cells); the *cell transfer delay* (CTD) (i.e., the delay experienced by a cell throughout the ATM network); and the *Cell Delay Variation* (CDV) (i.e., the variance of CTD). Input traffic characteristics parameters that can be negotiated between the user and the network include the *peak cell rate* (PCR), which is the maximum rate at which the user will transmit; the *sustained cell rate* (SCR), which is the average transmission rate measured over a long interval; and the *burst tolerance* (BT) and the *maximum burst size* (MBS), which define the burstiness of the sender.

On the basis of these parameters, five service categories are defined that characterize the different types of traffic that can be transferred by an ATM network:

- *Constant bit rate* (CBR), intended for applications that require a fixed data rate and an upper bound on transfer delay (e.g., videoconferencing, telephony).
- *Real-time variable bit rate* (rt-VBR), intended for applications that transmit at a variable rate and are time-sensitive (i.e., require tight upper bounds on cell transfer delay and delay variation, e.g., interactive compressed video).
- *Non-real-time variable bit rate* (nrt-VBR), intended for applications that transmit at a variable rate but are not time-sensitive (e.g., banking transactions).
- *Available bit rate* (ABR), used by normal (bursty) applications with relaxed delay and cell loss requirements (such as file transfer and email) and expected to be the most commonly used service category. ABR sources use explicit network feedback to control their cell rate.
- *Unspecified bit rate* (UBR), used by applications that are not sensitive to delay and cell loss and provide best-effort services by taking advantage of network capacity not allocated to CBR, VBR, or ABR traffic (e.g., file transfer).

6.1. ATM Congestion Avoidance and Control

ATM was designed so as to minimize the processing and transmission overhead inside the network. The only factor that can lead to cell delay variation is network congestion. ATM handles network congestion with functions that fall within two general categories: congestion avoidance and congestion control.

Congestion avoidance is concerned mainly with (1) establishing *traffic contracts* with the users, which specify the traffic parameters for a connection and the QoS parameters the network will support for that connection; and (2) enforcing the agreed-on traffic contracts, that is, watching that the agreed-on restrictions are met (*traffic policing*). Sometimes, however, congestion avoidance actions may not be effective, in which case some nodes may become congested, and congestion control functions, which are based primarily on network feedback, are brought into play.

The QoS required by CBR and VBR traffic is supported by congestion avoidance techniques based on traffic contracts and policing; cells that violate a traffic contract are discarded or tagged as low-priority cells. Traffic policing can be combined with *traffic shaping* to achieve better network efficiency, fair allocation of resources, and reduced average delays, while meeting the QoS objectives. A common mechanism used for traffic shaping is the "leaky bucket." This mechanism can smooth out a bursty cell flow by buffering arriving cells and then serving the queue (i.e., transmitting the cells) at a constant service rate of r cells per second (see Fig. 12). The leaky bucket example shown can be controlled by adjusting two parameters: the bucket capacity and the bucket leaking rate. As long as the bucket is not empty, the cells are transmitted with the constant rate of r cells per second. In case the bucket capacity is exceeded, a bucket overflow is caused and the excessive cells can be either discarded, or tagged as low priority cells.

These techniques prevent congestion buildup but get no feedback from the network concerning the congestion conditions and, therefore, are called *open-loop control techniques*. Open-loop control may be insufficient in cases where the bandwidth requirements of applications are not known at connection setup time. With ABR traffic, however, dynamic load management is possible by taking advantage of network feedback. In this case, feedback techniques are employed (*closed-loop control* techniques) that support the lossless transport of ABR traffic and the fair capacity allocation.

There are two main closed-loop control techniques: the credit-based and the rate-based. *Credit-based* schemes are based on a window flow control mechanism; each intermediate node on a session's path sends information (credit) to the previous (upstream) node and does so on a per link and per VC basis. A traffic source can transmit on a VC only when it obtains credit from the next node for this connection, which implies that the next node has adequate buffer and can accommodate the number of cells specified by the credit it grants to the source. In *rate-based* schemes, the network sends appropriate information to the user, specifying the bit rate at which the user could transmit, and the feedback control loop may extend end-to-end across the network. The rate-based approach is less expensive in terms of implementation complexity and hardware cost, but it doesn't handle bursty traffic well. The credit-based approach, on the other hand, is well suited for bursty traffic (under ideal conditions, zero cell loss can be guaranteed), but it requires complex bookkeeping at

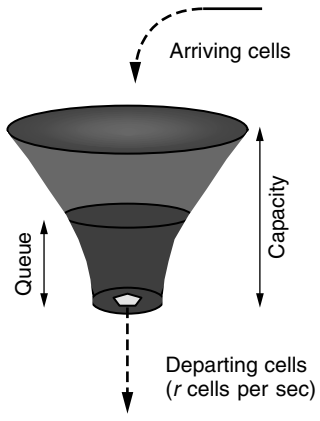


Figure 12. The leaky-bucket mechanism.

the network nodes on a per session (i.e., per VC) basis. The need for per session queuing limits the flexibility of the designer and is one of the main reasons why the ATM Forum has selected rate-based schemes for ABR traffic in ATM networks.

The rate-based scheme adopted for use in ATM uses the following mechanism. The source of an ABR flow inserts among the data cells of the flow some *resource management* (RM) cells, one RM cell after a certain number of data cells. Each RM cell contains three fields that can provide congestion feedback to the source: the *congestion indication* (CI) bit, the *no increase* (NI) bit, and the *explicit cell rate* (ER) field. The values of these fields can be changed by either the destination of the flow or an intermediate ATM switch, to reflect experienced congestion effects. Finally RM cells return to their source, where their fields are examined and corrective actions (rate increase or decrease) are taken to respond to the network or destination conditions. In particular, the source at any time is allowed to transmit cells at any rate between zero and a value called *allowed cell rate* (ACR). ACR is adjusted dynamically according to network feedback and has a lower and an upper limit, called *minimum cell rate* (MCR) and *peak cell rate* (PCR), respectively. If the source gets an RM cell with the CI bit set (signaling congestion), then it reduces ACR by an amount proportional to its current ACR (but down to MCR). If neither CI nor NI is set, then the source increases ACR by an amount proportional to the PCR (but up to PCR). Therefore, rate increases are linear, but rate decreases are exponential, so that sources respond drastically to congestion. Finally, if ACR is bigger than the value contained in the ER field (which is used to explicitly dictate a cell rate), then ACR is reduced to ER.

6.2. Connection Establishment Control Protocols for High-Speed Networks

The rapid developments in optoelectronics technology have substantially increased system transmission rates in optical communication networks since the first systems were installed, in the early–mid-1980s. Having, however, communication links of multigigabit transmission rates does not necessarily result in a communication network of the same effective capacity. An important (but not the only) issue is related to the protocols and algorithms used to perform network control. These protocols should allow full utilization of the network resources in a way that is fair to all users; they should be capable of providing delay and packet loss guarantees to the users (QoS), in the presence of node and link failures, and they should impose small processing requirements on the switches.

6.2.1. Protocols for CBR Traffic and for Traffic Consisting of Long Bursts. A sizable portion of traffic in future multigigabit-per-second networks will involve high-speed transfer of traffic at nearly constant rates (CBR traffic) and would require guaranteed lossless delivery and an explicit reservation of bandwidth. Clearly, the bandwidth–delay product being very large can result in the discarding of substantial amounts of data and retransmissions, unless bandwidth reservations are made in advance, or substantial buffer space is provided. Also, for high-speed file-transfer-type applications, long burst transmissions can easily overload the network, unless they have prenegotiated at least a minimum bandwidth with the network. Therefore, from the point of view of both transmission integrity and network efficiency, traffic of this type should be transferred only after a specific and explicit allocation precedes each data burst. This is especially true for the case of all-optical networks, where buffering has to be very limited because of technological constraints.

A key to efficiently utilizing the large bandwidth of emerging gigabit networks is to devise protocols that can overcome the problems posed by increased propagation latency of such networks. In most reservation protocols [e.g., the FRP/DT protocol, the fast bandwidth reservation schemes and the fast resource management (FRM) protocols], a setup packet is sent to the destination to make the appropriate reservations, and the capacity required by a session at an intermediate node is reserved starting at the time the setup packet arrives at that node. An obvious inefficiency in all these schemes arises because the capacity reserved for the session is not needed immediately, but it is actually needed at least one round-trip delay after arrival of the setup packet at the node. This is because the setup packet has to travel from the intermediate node to the destination, an acknowledgment has to be sent back to the source, and the first data packet of the session has to arrive from the source to the intermediate node (see Fig. 13). Over long transmission distances, the round-trip delay may be comparable to, or even larger than, the holding time of a session. In particular, if a typical session requests capacity r bps (bits per second), and transfers a total of M bits over a distance of L kilometers, the maximum percentage of time that the capacity is efficiently used is $e = (M/r)/[2Ln/c) + (M/r)]$, where c/n is the propagation speed in the fiber. Typical values of these parameters for multigigabit networks may be $r = 10$ Gbps, $M = 0.2$ Gbit, and $L = 1500$ km, which yields $e = 0.57$. This efficiency factor e decreases as r or L increase, or M decreases.

The *efficient reservation virtual circuit* (*ERVC*) protocol was designed to overcome these limitations. It is suitable for sessions that require an explicit reservation of bandwidth, and it does not suffer from the inefficiencies of the reservation protocols mentioned above. The ERVC protocol keeps track of sessions (or burst) durations and

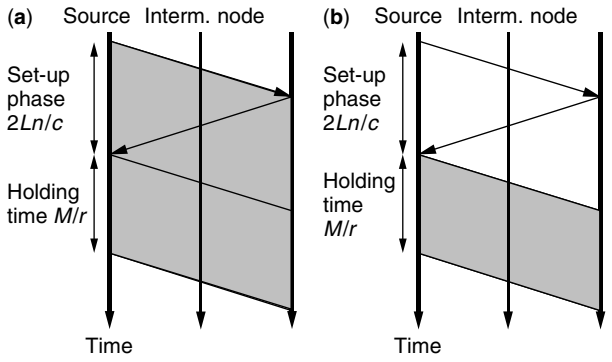


Figure 13. Comparison between (**a**) ERVC and standard reservation (SR) (**b**) protocols (shaded areas represent the time periods during which the capacity is reserved for a particular session).

reserves capacity only for the duration of a session (or burst), thus eliminating the inefficiency that results in existing schemes from holding capacity idle for a round-trip delay before it is actually used by data packets. In the ERVC protocol, session durations (or burst durations) are recorded, and each node keeps track of the utilization profile $r^l(t)$ of each outgoing link l, which describes the amount of residual capacity available on link l as a function of time t. This feature allows capacity to be reserved only for the duration of the session (or burst), starting at the time it is actually needed. Correct timing is crucial in ensuring that data transmission starts *after* all reservations are made and terminates *before* any intermediate node releases the reserved capacity. The ERVC protocol uses capacity on a demand basis, leading to more efficient utilization and a lower blocking probability than previous reservation protocols. It also has the "reservation ahead" feature that allows a node to calculate the time at which the requested capacity will become available and reserve it in advance (provided it is available within the QoS requirements of the session), avoiding in this way the wasteful repetition of the call setup phase. The protocol uses an asynchronous, distributed algorithm that allows the nodes along a session's path to collaborate when reserving capacity and to maintain timing consistency. This ensures that adequate outgoing capacity is available to service the data packets when they arrive at a link, so that the transmission is loss-free. Processing requirements at a node are minimized by using efficient update mechanisms and simple data structures that store a compact representation of the utilization profile of an outgoing link. The information required by the protocol (rates and session durations) can be recorded and processed using a simple linked list structure. The protocol is robust to link and node failures, and it allows soft recovery from processor failures. The efficiency factor e for the ERVC protocol can be as large as $e = 1$, independently of the parameters r, L, and M and efficiency is maintained even for traffic that consists of sporadic bursts of data. Also, the performance of ERVC does not depend on the round-trip delay. This is because for a single link, a different round-trip delay means only that the arrivals of sessions on link l are translated in time by a different amount; therefore, the picture in terms of load (and consequently the blocking) as seen by newly arriving sessions remains the same, irrespective of the round-trip delay.

6.2.2. The Virtual Circuit Deflection Protocol. Traffic in high-speed networks can be switched either optically or electronically. Optical switching has advantages for circuit switching, but substantial disadvantages for packet switching, because effective packet switching requires packet storage at each switch, which is difficult to achieve with current optical technology. Despite this drawback, it is believed that optical switching may open new dimensions in future networking, provided appropriate protocols that take into account its constraints are developed.

To eliminate the need for buffering (but without making advance bandwidth reservations, which requires a round-trip pretransmission delay), a variation of deflection routing, called *virtual circuit deflection* (VCD) protocol can be used. The VCD protocol is a combination of virtual circuit switching and deflection routing, and is appropriate for sessions that simultaneously require minimal pretransmission delay and lossless communication. VCD is a "tell and go" (or "immediate transmission") type of protocol, and does not therefore use end-to-end reservations. In the VCD protocol, a path (called "preferred path") is selected for a new session, based on (possibly outdated) topology and link utilization information available at its source at the time. A setup packet is sent to the destination to establish the connection, followed after a short delay (much shorter than the end-to-end round-trip delay required by reservation protocols) by the data packets. This delay should be large enough to permit the electronic processing of the setup packet, without being overpassed by the data packets. If the available capacity on a preferred link of a session is inadequate, the session may have to follow a different, longer path; we then say that the session is deflected. When the total incoming link capacity is equal to the total outgoing link capacity of a node, as is usually the case in most data networks, it can be shown that there is always adequate available capacity on the outgoing links of an intermediate node to accommodate a new session. This, however, may happen at the expense of interrupting (preempting) an existing session that originates at that node, and/or splitting the new session into two or more smaller subsessions that are routed through different paths (session splitting). Deflection or splitting of sessions at intermediate nodes is infrequent in the VCD protocol, and can happen only when the topology or link utilization information at the source is outdated and the network is congested. Resequencing of packets, which is the major drawback of conventional (datagram) deflection schemes, is much simpler to accomplish in the VCD protocol. If a session is split, a few blocks of data packets (each of which is ordered) will have to be resequenced; this is a considerably easier task to perform than the resequencing of millions of individual packets that are out of order, as is the case in conventional deflection schemes.

Even though the effective utilization of idle links is an advantage, the increase of the number of used links per call is a disadvantage of the VCD protocol. An important performance measure is the *inefficiency ratio* $\eta(\lambda)$, defined as the ratio $\eta(\lambda) = D(\lambda)/D(0)$ of the average path length $D(\lambda)$ taken by a session for a given arrival per node rate λ, over the average shortest pathlength $D(0)$ of a given network topology. Results obtained from experiments on a Manhattan street network topology, indicate that the VCD protocol can be very efficient for high-speed networks, where link capacities are big and links are shared by a large number of small sessions.

7. TCP CONGESTION CONTROL

As we discussed in Section 6.1, TCP implements a credit-based flow control mechanism, using the *window size* field of the TCP header. This way, a destination entity avoids buffer overflow by limiting the dataflow from the source entity. This mechanism, however, can be further enhanced to provide network congestion control.

A source entity using TCP maintains a queue that holds the transmitted but not yet acknowledged segments

of its datastream. If a segment is not acknowledged after a certain time period, then this segment is considered lost and is retransmitted. A critical issue here is the determination of the retransmission timeout period. This is accomplished, in most TCP implementations, on the basis of *round-trip time* (RTT) and *RTT variance* estimates for the transmitted segments. Early TCP implementations calculated RTTs (and checked for timeouts) at "clock ticks" of 500 ms. This clock coarseness, however, led to long timeout intervals and, subsequently, large delays in retransmissions. This was solved by having retransmissions occur when, aside from timeouts (a number of), duplicate ACKs are received. A receiver transmits a duplicate ACK when it cannot acknowledge a segment because an earlier segment is lost. Therefore, when a certain number of duplicate ACKs (usually 3) are received, the source is warned that the segment after the one acknowledged is lost, and so it triggers retransmission. Another parameter that can be managed by congestion control techniques is the size of the TCP window, which can be adapted dynamically to the changing network conditions. A common mechanism for doing this is the *slow-start* mechanism, according to which at the beginning of a transmission or retransmission only one segment is transmitted, and then, for each ACK received, an extra segment is transmitted (in addition to the one acknowledged in the ACK) up to a maximum value. In this way the source probes the network with a small amount of data and increases exponentially its flow up to a certain threshold, after which the flow increases linearly.[9] Increase goes on until segments are lost, which implies that the available bandwidth is exceeded; then the source responds by decreasing its window size. In other words, TCP finds the available bandwidth by congesting the network and causing own fragment losses.

The mechanisms described in the previous paragraph are found in the early TCP implementation known as *Reno*. However, as this field is the focus of extensive research, several alternative and more effective mechanisms came up (and still do!) to enhance the functionality of TCP. Several such modifications were incorporated into the well-known *TCP Vegas* implementation. Vegas is reported to achieve 37–71% better throughput, with one-fifth to one-half the losses as compared to Reno. Specifically, Vegas features a new retransmission mechanism, a congestion avoidance mechanism, and an improved slow-start mechanism. The new retransmission mechanism detects lost segments much sooner than did Reno (and without the need for a second or third duplicate ACK) by using the fine-grain system clock to calculate RTTs.[10] Consequently, when a duplicate ACK is received, the more accurate RTT estimate is compared against the timeout value, and if it is found to be larger, the source retransmits the segment without having to wait for n duplicate ACKs. Moreover, Vegas checks the first couple of nonduplicate ACKs received after a retransmission, and if the calculated (over these ACKs) RTTs exceed the timeout value, then the corresponding segments are retransmitted. In this way any other segment that was lost prior to the retransmission is detected without having to wait for a duplicate ACK. Vegas also implements a congestion avoidance (*proactive*) mechanism, in contrast to the inherently *reactive* congestion detection mechanism of Reno. This mechanism performs a comparison between the *Expected throughput* of the connection [calculated by dividing the size of the current (congestion) window by the minimum of all the measured RTTs] and the measured *Actual throughput* (calculated by dividing the actual number of bytes transmitted during the RTT of a segment by this RTT). The result of this comparison (i.e., *expected throughput* minus *actual throughput,* which is always nonnegative) is used to adjust the congestion window size, and in particular, if the difference is below a low threshold (α), the actual and expected throughput values are too close and therefore the window size is increased linearly to catch up with the available bandwidth. If, on the other hand, the difference is above a high threshold (β), the actual and expected throughput values are too distant and therefore the window size is decreased linearly to react to the network congestion observed. The window size remains constant if the difference is between the two threshold values. The two threshold values practically correspond to the number of network buffers the connection occupies; thus, Vegas detects network congestion and responds to it by trying to limit the number of occupied buffers. Finally, Vegas uses a modified slow-start mechanism, which is integrated with the congestion avoidance mechanism described previously. At the beginning of the connection, this mechanism tries to find the connection's available bandwidth, without incurring the segment losses that Reno does. This is done by allowing the (exponential) growth of the congestion window only every other RTT. In the meantime, the window size remains constant, so that a comparison between the expected and actual rates can signal congestion and trigger a switch to the linear increase/decrease mode, implying that connection's available bandwidth is reached.

[9] The exact value of the threshold is not known during the initial slow start; when, however, a retransmit timeout occurs (indicating congestion), the threshold is set to half the current window size.

[10] The RTT of a segment is calculated by subtracting the segment's transmission time from the corresponding ACK reception time (by the segment source).

BIOGRAPHIES

Emmanouel (Manos) Varvarigos was born in Athens, Greece, in 1965. He received a Diploma in Electrical and Computer Engineering from the National Technical University of Athens in 1988, and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, Cambridge, in 1990 and 1992, respectively. In 1990 he was a researcher at Bell Communications Research, Morristown, New Jersey. From 1992 to 1998 he was an Assistant and later an Associate Professor at the department of Electrical and Computer Engineering at the University of California, Santa Barbara. In 1998/99 he was an Associate Professor at the Electrical Engineering Department at Delft University of Technology, the Netherlands. In 1999 he became a Professor in the Department of Computer Engineering

and Informatics at the University of Patras, where he is currently Director of the Hardware and Computer Architecture Division and Head of the Data Transmission and Networking Lab. His research activities are in the areas of protocols and algorithms for high-speed networks, all-optical networks, high-performance switch architectures, parallel and distributed computing, interconnection networks, VLSI layout design, performance evaluation, and ad hoc networks.

Theodora A. Varvarigou received the B.Tech. degree from the National Technical University of Athens, Athens, Greece in 1988, the M.S. degrees in Electrical Engineering (1989) and in Computer Science (1991) from Stanford University, Stanford, California in 1989 and the Ph.D. degree from Stanford University as well in 1991.

She worked at AT&T Bell Labs, Holmdel, New Jersey between 1991 and 1995. Between 1995 and 1997 she worked as an Assistant Professor at the Technical University of Crete, Chania, Greece. Since 1997 she has been working as an Assistant Professor at the National Technical University of Athens.

Her research interests include parallel algorithms and architectures, fault-tolerant computation, and parallel scheduling on multiprocessor systems.

FURTHER READING

J. S. Ahn et al., Evaluation of TCP Vegas: Emulation and experiment, *Proc. SIGCOMM '95*, Aug. 1995.

D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

U. Black, *Data Link Protocols*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

F. Bonomi and K. Fendick, The rate-based flow control framework for the available bit rate ATM service, *IEEE Network* (March/April 1995).

L. Brakmo and L. Peterson, TCP Vegas: End to end congestion avoidance on a global Internet, *IEEE J. Select. Areas Commun.* (Oct. 1995).

J. Carlo et al., *Understanding Token Ring Protocols and Standards*, Artech House, Boston, 1999.

E. Carne, *Telecommunications Primer*, Prentice-Hall, Upper Saddle River, NJ, 1995.

T. Chen, S. Liu, and V. Samalam, The available bit rate service for data in ATM networks, *IEEE Commun. Mag.* (May 1996).

D. Clark, The design philosophy of the DARPA Internet protocols, *Proc. SIGCOMM '88, Computer Communication Review*, Aug. 1988.

D. Clark, S. Shenker, and L. Zhang, Supporting real-time applications in an integrated services packet network: Architecture and mechanism, *Proc. SIGCOMM '92*, Aug. 1992.

D. Comer and D. Stevens, *Internetworking with TCP/IP,* Vol. II: *Design Implementation, and Internals*, Prentice-Hall, Upper Saddle River, NJ, 1999.

D. Comer, *Internetworking with TCP/IP,* Vol. I: *Principles, Protocols, and Architecture*, Prentice-Hall, Upper Saddle River, NJ, 1995.

B. Dorling, P. Pieters, and E. Valenzuela, *IBM Frame Relay Guide*, IBM Publication SG24-4463-01, 1996.

A. Eckeberg, D. Luan, and M. Lucantoni, An approach to controlling congestion in ATM networks, *Int. J. Digital Analog Commun. Syst.* **3**(2): 1990.

A. Gersht and K. Lee, A congestion control framework for ATM networks, *IEEE J. Select. Areas Commun.* (Sept. 1991).

W. Goralski, *Introduction to ATM Networking*, McGraw-Hill, New York, 1995.

F. Halsall, *Data Communications, Computer Networks, and Open Systems*, Addison Wesley, Reading, MA, 1996.

R. Handel, N. Huber, and S. Schroder, *ATM Networks: Concepts, Protocols, Applications*, Addison Wesley, Reading, MA, 1994.

S. Haykin, *Communication Systems*, Wiley, New York, 1995.

C. Huitema, *Routing in the Internet*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

C. Huitema, *Ipv6: The New Internet Protocol*, Prentice-Hall, Upper Saddle River, NJ, 1998.

Internetworking Technologies Handbook, 3rd ed., Cisco Press, 2000.

V. Jacobson, Congestion avoidance and control, *Proc. SIGCOMM '88, Computer Communication Review*, Aug. 1988.

V. Jacobson, Berkeley TCP evolution from 4.3 Tahoe to 4.3-Reno, *Proc. 18th Internet Engineering Task Force*, Sept. 1990.

R. Jain, Congestion control in computer networks: Issues and trends, *IEEE Network Mag.* (May 1990).

R. Jain, Myths about congestion management in high-speed networks, *Internetworking: Research and Experience*, Vol. 3, 1992.

R. Jain et al., Source behavior for ATM ABR traffic management: An explanation, *IEEE Commun. Mag.* (Nov. 1996).

R. Jain, Congestion control and traffic management in ATM networks: Recent advances and a survey, *Computer Networks ISDN Syst.* **28**(13): (Oct. 1996).

P. Karn and C. Partridge, Improving round-trip estimates in reliable transport protocols, *ACM Trans. Comput. Syst.* (Nov. 1991).

N. Kavak, Data communication in ATM networks, *IEEE Network* (May/June 1995).

H. T. Kung, T. Blackwell, and A. Chapman, A credit-based flow control scheme for ATM networks: Credit update protocol, adaptive credit allocation, and statistical multiplexing, *Proc. SIGCOMM '94*, Aug./Sept. 1994.

H. T. Kung and R. Morris, Credit-based flow control for ATM networks, *IEEE Network* (March 1995).

S. Low, L. Peterson, and L. Wang, Understanding TCP Vegas: A duality model, *Proc. SIGMETRICS '01*, June 2001.

D. Mc Dysan and D. Spohn, *ATM: Theory and Applications*, McGraw-Hill, New York, 1999.

S. Miller, *IPv6: The Next Generation Internet Protocol*, Digital Press, Bedford, MA, 1998.

J. Mo et al., Analysis and comparison of TCP Reno and Vegas, *Proc. IEEE Infocom*, March 1999.

G. Moshos, *Data Communications: Principles and Problems*, West Publishing, New York, 1989.

M. Murhammer et al., *TCP/IP: Tutorial and Technical Overview*, Prentice-Hall, Upper Saddle River, NJ, 1998.

M. Naugle, *Local Area Networking*, McGraw-Hill, New York, 1996.

P. Newman, ATM local area networks, *IEEE Commun. Mag.* (March 1994).

H. Oshaki et al., Rate-based congestion control for ATM networks, *Comput. Commun. Rev.* (April 1995).

L. Peterson and B. Davie, *Computer Networks: A Systems Approach*, Morgan Kaufmann, San Francisco, 1996.

J. Pitts and J. Schormans, *Introduction to ATM Design and Performance*, Wiley, New York, 1996.

J. Proakis and M. Salehi, *Communication Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

M. Prycker, *Asynchronous Transfer Mode: Solutions for Broadband ISDN*, Ellis Horwood, New York, 1996.

A. Rodriguez et al., *TCP/IP Tutorial and Technical Overview*, 7th ed., IBM Publication GG24-3376-06, 2001.

K. Sato, S. Ohta, and I. Tokizawa, Broadband ATM network architecture based on virtual paths, *IEEE Trans. Commun.* (Aug. 1990).

M. Schwartz, *Computer-Communication Network Design and Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

A. Shah and G. Ramakrishnan, *FDDI: A High-Speed Network*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

D. Spohn, *Data Network Design*, McGraw-Hill, New York, 1994.

J. Spragins, J. Hammond, and K. Pawlikowski, *Telecommunication Protocols and Design*, Addison-Wesley, Reading, MA, 1991.

W. Stallings, *High-Speed Networks: TCP/IP and ATM Design Principles*, Prentice-Hall, Upper Saddle River, NJ, 1998.

W. Stallings, *Local and Metropolitan Area Networks*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2000.

W. Stallings, *Data and Computer Communications*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2000.

M. Steenstrup, *Routing in Communications Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

S. Steinke, IP addresses and subnet masks, *LAN Mag.* (Oct. 1995).

W. Stevens, *TCP/IP Illustrated,* Vol. 1: *The Protocols*, Addison-Wesley, Reading, MA, 1994.

A. Tanenbaum, *Computer Networks*, Prentice-Hall, Upper Saddle River, NJ, 1996.

E. Varvarigos, Control protocols for multigigabit-per-second networks, *IEICE Trans. Commun.* (Feb. 1998).

W. Weiss, QoS with differentiated services, *Bell Labs Tech. J.* (Oct.–Dec. 1998).

P. White and J. Crowcroft, The integrated services in the Internet: State of the art, *Proc. IEEE*, Dec. 1997.

G. Wright and W. Stevens, *TCP/IP Illustrated,* Vol. 2: *The Implementation*, Addison-Wesley, Reading, MA, 1995.

X. Xiao and L. Ni, Internet QoS: A big picture, *IEEE Network* (March/April 1999).

C. Yang and A. Reddy, A taxonomy for congestion control algorithms in packet switching networks, *IEEE Network* (July/Aug. 1995).

L. Zhang, Why TCP timers don't work well, *Proc., SIGCOMM '86 Symp.*, Aug. 1986.

H. Zhang, Service disciplines for guaranteed performance service in packet switching networks, *Proc. IEEE*, Oct. 1995.

CONCATENATED CONVOLUTIONAL CODES AND ITERATIVE DECODING

WILLIAM E. RYAN
University of Arizona
Tucson, Arizona

1. INTRODUCTION

Turbo codes, first presented to the coding community in 1993 [1,2], represent one of the most important breakthroughs in coding since Ungerboeck introduced trellis codes in 1982 [3]. A turbo code encoder, comprises a concatenation of two (or more) convolutional encoders, and its decoder consists of two (or more) "soft" convolutional decoders that feed probabilistic information back and forth to each other in a manner that is reminiscent of a turbo engine. This chapter presents a tutorial exposition of parallel and serial concatenated convolutional codes (PCCCs and SCCCs), which we will also call *parallel* and *serial turbo codes*. Included here are a simple derivation for the performance of these codes and a straightforward presentation of their iterative decoding algorithms. The treatment is intended to be a launching point for further study in the field and to provide sufficient information for the design of computer simulations. This article borrows from some of the most prominent publications in the field [4–12].

The article is organized as follows. Section 2 describes details of the parallel and serial turbo code encoders. Section 3 derives a truncated union bound on the error rate of these codes under the assumption of maximum-likelihood decoding. This section explains the phenomenon of interleaver gain attained by these codes. Section 4 derives in detail the iterative (turbo) decoder for both PCCCs and SCCCs. Included in this section are the BCJR (Bahl–Cocke–Jelinek–Raviv) decoding algorithm for convolutional codes and soft-in/soft-out decoding modules for use in turbo decoding. The decoding algorithms are presented explicitly to facilitate the creation of computer programs. Section 5 contains a few concluding remarks.

2. ENCODER STRUCTURES

Figure 1 depicts a parallel turbo encoder. As seen in the figure, the encoder consists of two binary rate 1/2 convolutional encoders arranged in a so-called parallel concatenation, separated by a K-bit pseudorandom interleaver or permuter. Also included is an optional puncturing mechanism to obtain high code rates [13]. Clearly, without the puncturer, the encoder is rate $\frac{1}{3}$, mapping K data bits to $3K$ code bits. With the puncturer, the code rate $R = K/(K+P)$, where P is the number of parity bits remaining after puncturing. Observe that the constituent encoders are recursive systematic convolutional (RSC) codes. As will be seen below, recursive encoders are necessary to attain the exceptional performance (attributed to "interleaver gain") provided by

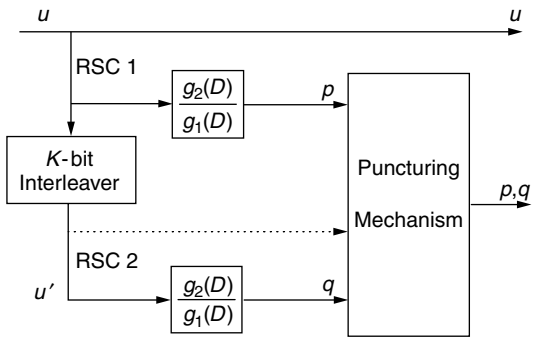


Figure 1. PCCC encoder diagram.

turbo codes. Without any essential loss of generality, we assume that the constituent codes are identical.

Figure 2 depicts a serial turbo encoder. As seen in the figure, the serially concatenated convolutional encoders are separated by an interleaver, and the inner encoder is required to be an RSC code, whereas the outer encoder need not be recursive [6]. However, RSC inner and outer encoders are often preferred since it is convenient to puncture only parity bits to obtain high code rates [14]. Further, an RSC outer code will facilitate our analysis below. The code rate for the serial turbo encoder is $R = R_o \cdot R_i$ where R_o and R_i are the code rates for the outer and inner codes, respectively.

For both parallel and serial turbo codes, the codeword length is $N = K/R$ bits, and we may consider both classes to be (N, K) block codes.

We now discuss in some detail the individual components of the turbo encoders.

2.1. The Recursive Systematic Encoders

Whereas the generator matrix for a rate $\frac{1}{2}$ nonrecursive convolutional code has the form $G_{NR}(D) = [g_1(D)\ g_2(D)]$, the equivalent recursive systematic encoder has the generator matrix

$$G_R(D) = \left[1\ \frac{g_2(D)}{g_1(D)}\right]$$

Observe that the code sequence corresponding to the encoder input $u(D)$ for the former code is $u(D)G_{NR}(D) = [u(D)g_1(D)\ \ u(D)g_2(D)]$, and that the identical code sequence is produced in the recursive code by the sequence $u'(D) = u(D)g_1(D)$, since in this case the code sequence is $u(D)g_1(D)G_R(D) = u(D)G_{NR}(D)$. Here, we loosely call the pair of polynomials $[u(D)g_1(D)\ u(D)g_2(D)]$ a *code sequence*, although the actual code sequence is derived from this polynomial pair in the usual way.

Observe that, for the recursive encoder, the code sequence will be of finite weight if and only if the input sequence is divisible by $g_1(D)$. We have the following corollaries of this fact, which we shall use later.

Fact 1. A weight 1 input will produce an infinite weight output for such an input is never divisible by a (nontrivial) polynomial $g_1(D)$. (In practice, "infinite" should be replaced by "large" since the input length is finite.)

Fact 2. For any nontrivial $g_1(D)$, there exists a family of weight 2 inputs of the form $D^j(1 + D^p)$, $j \geq 0$, which produce finite weight outputs, i.e., which are divisible by $g_1(D)$. When $g_1(D)$ is a primitive polynomial of degree m, then $p = 2^m - 1$. More generally, p is the length of the pseudorandom sequence generated by $g_1(D)$.

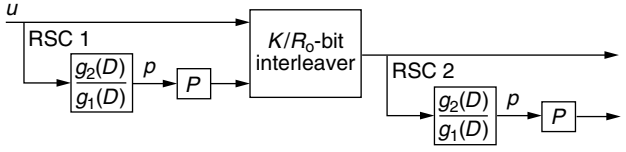


Figure 2. SCCC encoder diagram with RSC component codes. "P" signifies possible puncturing of parity bits.

Proof: Because the encoder is linear, its output due to a weight 2 input $D^j(1 + D^t)$ is equal to the sum of its outputs due to D^j and D^jD^t. The output due to D^j will be periodic with period p since the encoder is a finite-state machine (see Example 1 and Fig. 3); the state at time j must be reached again in a finite number of steps p, after which the state sequence is repeated indefinitely with period p. Now letting $t = p$, the output due to D^jD^p is just the output due to D^j shifted by p bits. Thus, the output due to $D^j(1 + D^p)$ is the sum of the outputs due to D^j and D^jD^p, which must be of finite length and weight since all but one period will cancel in the sum.

In the context of the code's trellis, fact 1 says that a weight-1 input will create a path that diverges from the all-zeros path, but never remerges. Fact 2 says that there will always exist a trellis path that diverges and remerges later, which corresponds to a weight 2 data sequence.

Example 1. Consider the code with generator matrix

$$G_R(D) = \left[1\ \frac{1 + D^2 + D^3 + D^4}{1 + D + D^4}\right].$$

Thus $g_1(D) = 1 + D + D^4$ and $g_2(D) = 1 + D^2 + D^3 + D^4$ or, in octal form, $(g_1, g_2) = (31, 27)$. Observe that $g_1(D)$ is primitive so that, for example, $u(D) = 1 + D^{15}$ produces the finite-length code sequence $(1 + D^{15}, 1 + D + D^3 + D^4 + D^7 + D^{11} + D^{12} + D^{13} + D^{14} + D^{15})$. Of course, any delayed version of this input, say, $D^7(1 + D^{15})$, will simply produce a delayed version of this code sequence. Figure 3 gives one encoder realization for this code. We remark that, in addition to elaborating on Fact 2, this example serves to demonstrate the conventions generally used in the literature for specifying such encoders.

2.2. The Interleaver

The function of the interleaver is to take each incoming block of bits and rearrange them in a pseudorandom fashion prior to encoding by the second encoder. For the PCCC, the interleaver permutes K bits and, for the SCCC, the interleaver permutes K/R_o bits. Unlike the classical interleaver (e.g., block or convolutional interleaver), which

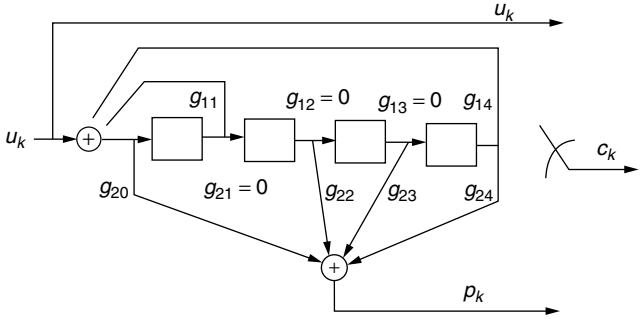


Figure 3. RSC encoder with $(g_1, g_2) = (31, 27)$.

rearranges the bits in some systematic fashion, it is crucial that this interleaver sort the bits in a manner that lacks any apparent order, although it might be tailored in a certain way for weight 2 and weight 3 inputs as will be made clearer below. The $\mathcal{S}$ random interleaver [8] is quite effective in this regard. This particular interleaver ensures that any two inputs bits whose positions are within $\mathcal{S}$ of each other are separated by an amount greater than $\mathcal{S}$ at the interleaver output. $\mathcal{S}$ should be selected to be as large as possible for a given value of K. Also, as we shall see, performance increases with K, and so $K \geq 1000$ is typical.

2.3. The Puncturer

The role of the turbo code puncturer is identical to that of its convolutional code counterpart, that is, to delete selected bits to reduce coding overhead. We have found it most convenient to delete only parity bits, but there is no guarantee that this will maximize the minimum codeword distance. For example, to achieve a rate of $\frac{1}{2}$, one might delete all even parity bits from the top encoder and all odd parity bits from the bottom one.

3. PERFORMANCE WITH MAXIMUM-LIKELIHOOD DECODING

As will be elaborated upon in the next section, a maximum-likelihood (ML) sequence decoder would be far too complex for a turbo code, due to the presence of the permuter. However, the suboptimum iterative decoding algorithm to be described there offers near-ML performance. Hence, we shall now estimate the performance of an ML decoder on a binary input AWGN channel with power spectral density $N_0/2$ (analysis of the iterative decoder is much more difficult).

Armed with the preceding descriptions of the components of the turbo encoders of Figs. 1 and 2, we can easily conclude that it is linear since its components are linear. The constituent codes are certainly linear, and the interleaver is linear since it may be modeled by a permutation matrix. Further, the puncturer does not affect linearity since all the constituent codewords share the same puncture locations. As usual, the importance of linearity in evaluating the performance of a code is that one may choose the all-zeros sequence as a reference. Thus, we shall assume that the all-zeros codeword was transmitted. The development below holds for both parallel and serial turbo codes.

Now consider the all-zeros codeword (the 0th codeword) and the kth codeword, for some $k \in \{1, 2, \ldots, 2^K - 1\}$. The ML decoder will choose the kth codeword over the 0th codeword with probability $Q(\sqrt{2d_k RE_b/N_0})$, where d_k is the weight of the kth codeword and E_b is the energy per information bit. The bit error rate for this two-codeword situation would then be

$$\begin{aligned} P_b(k|0) &= w_k \text{ (bit errors/cw error)} \\ &\quad \times \frac{1}{K} \text{ (cw/ data bits)} \\ &\quad \times Q(\sqrt{2Rd_kE_b/N_0}) \text{ (cw errors/cw)} \\ &= \frac{w_k}{K} Q\left(\sqrt{\frac{2Rd_kE_b}{N_0}}\right) \text{ (bit errors/data bit)} \end{aligned}$$

where w_k is the weight of the kth data word and "cw" = codeword. Now including all of the codewords and invoking the usual union bounding argument, we may write

$$\begin{aligned} P_b &= P_b(\text{choose any } k \in \{1, 2, \ldots, 2^K - 1\} \mid 0) \\ &\leq \sum_{k=1}^{2^K-1} P_b(k \mid 0) \\ &= \sum_{k=1}^{2^K-1} \frac{w_k}{K} Q\left(\sqrt{\frac{2Rd_kE_b}{N_0}}\right) \end{aligned}$$

Note that every nonzero codeword is included in the above summation. Let us now reorganize the summation as

$$P_b \leq \sum_{w=1}^{K} \sum_{v=1}^{\binom{K}{w}} \frac{w}{K} Q\left(\sqrt{\frac{2Rd_{wv}E_b}{N_0}}\right) \tag{1}$$

where the first sum is over the weight w inputs, the second sum is over the $\binom{K}{w}$ different weight w inputs, and d_{wv} is the weight of the v^{th} codeword produced by a weight w input. We emphasize that (1) holds for any linear code.

Consider now the first few terms in the outer summation of (1) in the context of parallel and serial turbo codes. Analogous to the fact that the top encoder in the parallel scheme is recursive, we shall assume that the outer encoder in the serial scheme is also recursive. By doing so, our arguments below will hold for both configurations. Further, an RSC outer code facilitates the design of high-rate serial turbo codes as mentioned above.

$w = 1$. From Fact 1 and associated discussion above, weight 1 inputs will produce only large weight codewords at both PCCC constituent encoder outputs since the trellis paths created never remerge with the all-zeros path. (We ignore the extreme case where the single 1 occurs at the end of the input words for both encoders for this is avoidable by proper interleaver design.) For the SCCC, the output of the outer encoder will have large weight because of Fact 1, and its inner encoder output will have large weight since its input has large weight. Thus, for both cases, each d_{1v} can be expected to be significantly greater than the minimum codeword weight so that the $w = 1$ terms in (1) will be negligible.

$w = 2$. Suppose that, of the $\binom{K}{2}$ possible weight 2 encoder inputs, $u_*(D)$ is the one of least degree that yields the minimum turbo codeword weight, $d_{2,\min}$, for weight-2 inputs. In the presence of the pseudorandom interleaver, the encoder is not time-invariant, and only a small fraction of the inputs of the form $D^j u_*(D)$ (there are approximately K of them) will also produce turbo codewords of weight $d_{2,\min}$. (This phenomenon has been called *spectral thinning* [10].) Denoting by n_2 the number of weight 2 inputs that produce weight $d_{2,\min}$ turbo codewords, we may conclude that $n_2 \ll K$. (For comparison, $n_2 \simeq K$ for a single RSC code as shifts of some worst-case input merely shifts the encoder output, thus maintaining a constant output weight.) Further, the overall minimum codeword weight, $d_{\min}$, is likely to be equal or close to $d_{2,\min}$ since low-degree, low-weight input words tend to produce low-weight

codewords. (This is easiest to see in the parallel turbo code case which is systematic.)

$w \geq 3$. When w is small (e.g., $w = 3$ or 4), an argument similar to the $w = 2$ case may be made to conclude that the number of weight w inputs, n_w, that yield the minimum turbo codeword weight for weight-w inputs, $d_{w,\min}$, is such that $n_w \ll K$. Further, we can expect $d_{w,\min}$ to be equal or close to $d_{\min}$. No such arguments can be made as w increases beyond about 5.

To summarize, by preserving only the dominant terms, the bound in (1) can be approximated as

$$P_b \simeq \sum_{w=2}^{3} \frac{w n_w}{K} Q\left(\sqrt{\frac{2Rd_{w,\min}E_b}{N_0}}\right) \tag{2}$$

where $w \geq 4$ terms may be added in the event that they are not negligible (more likely to be necessary for SCCCs). We note that n_w and $d_{w,\min}$ are functions of the particular interleaver employed. Since $w = 2$ or 3 in (1) and $n_w \ll K$ with $K \geq 1000$, the coefficients out in front of the Q function are much less than unity. (For comparison, the coefficient for a convolutional code can be much greater than unity [10].) This effect, called *interleaver gain*, demonstrates the necessity of large interleavers. Finally, we note that recursive encoders are crucial elements of a turbo code since, for nonrecursive encoders, division by $g_1(D)$ (nonremergent trellis paths) would not be an issue and (2) would not hold [although (1) still would].

When $K \simeq 1000$, it is possible to exhaustively find via computer the weight spectra $\{d_{2v}: v = 1, \ldots, \binom{K}{2}\}$ and $\{d_{3v}: v = 1, \ldots, \binom{K}{3}\}$ corresponding to the weight 2 and 3 inputs. In this case, an improved estimate of P_b, given by a truncation of (1), is

$$P_b \simeq \sum_{w=2}^{3} \sum_{v=1}^{\binom{K}{w}} \frac{w}{K} Q\left(\sqrt{\frac{2Rd_{wv}E_b}{N_0}}\right) \tag{3}$$

We remark that if codeword error rate, P_{cw}, is the preferred performance metric, then an estimate of P_{cw} may be obtained from (2) or (3) by removing the factor w/K from these expressions. That this is so may be seen by following the derivation above for P_b.

Example 2. We consider in this example a PCCC and an SCCC code, both rate $\frac{8}{9}$ with parameters $(N, K) = (1152, 1024)$. We use identical 4-state RSC encoders in the PCCC encoder whose generators polynomials are, in octal form, $(g_1, g_2) = (7, 5)$. To achieve a code rate of $\frac{8}{9}$, only one bit is saved in every 16-bit block of parity bits at each encoder output. The outer constituent encoder in the SCCC encoder is this same 4-state RSC encoder, and the inner code is a rate-1 differential encoder with transfer function $\frac{1}{1 \oplus D}$. A rate of $\frac{8}{9}$ is achieved in this case by saving one bit in every 8-bit block of parity bits. The PCCC interleaver is a 1024-bit pseudorandom interleaver with no constraints added (e.g., no $\mathcal{S}$-random constraint). The SCCC interleaver is a 1152-bit pseudorandom interleaver with no constraints added.

Figure 4 presents performance results for these codes based on computer simulation using the iterative (i.e., non-ML) decoding algorithm of the next section. Simulation results for both bit error rate P_b (BER in the figure) and

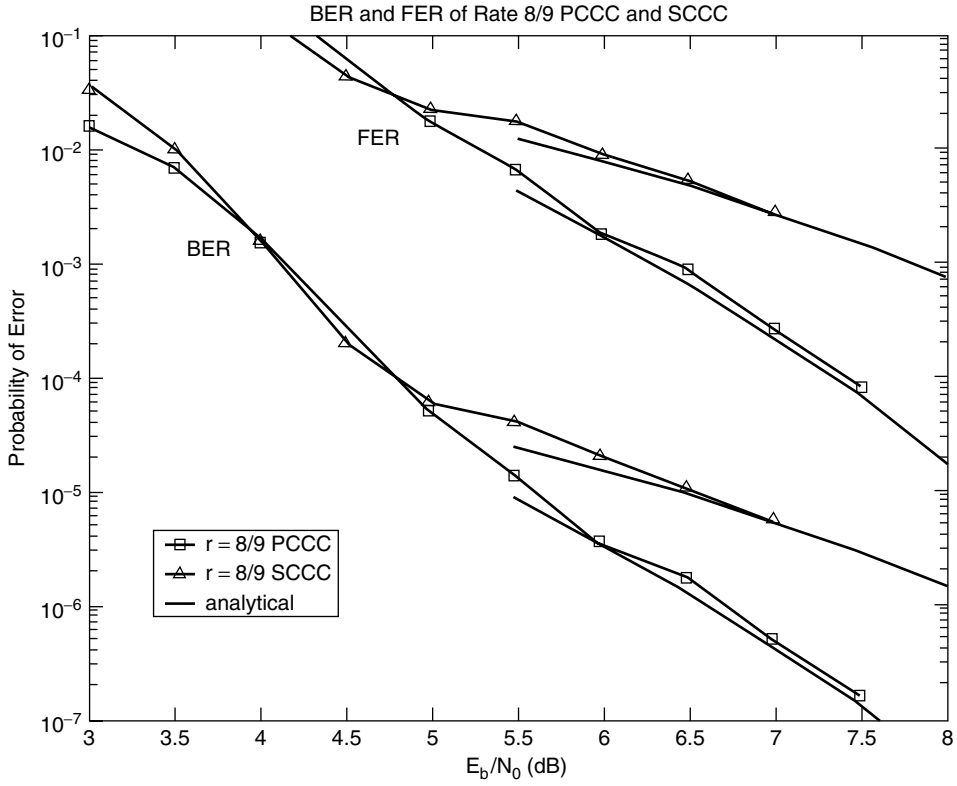


Figure 4. PCCC and SCCC bit error rate (BER) and frame error rate (FER) simulation results together with analytical result in (3).

frame or codeword error rate P_{cw} (FER in the figure) are presented. Also included in the figure are analytic performance curves for ML decoding using the truncated union bound in (3). (P_{cw} is obtained by removing the factor w/K in (3) as indicated above.) We see the close agreement between the analytical and simulated results in this figure.

In addition to illustrating the use of the estimate (3), this example helps explain the "flooring" effect of the error rate curves: it may be interpreted as the usual Q-function shape for a signaling scheme with a modest $d_{\min}$, "pushed down" by the interleaver gain $w^* n_{w^*}/K$, where w^* is the value of w corresponding to the dominant term in (2) or (3).

We comment on the fact that the PCCC in Fig. 4 is substantially better than the SCCC whereas it is known that SCCCs generally have lower floors [6]. We attribute this to the fact that the outer RSC code in the SCCC has been punctured so severely that $d_{\min} = 2$ for this outer code (although $d_{\min}$ for the SCCC is a bit larger). The RSC encoders for the PCCC is punctured only half as much, and so $d_{\min} > 2$ for each of these encoders. We also attribute this to the fact that we have not used an optimized interleaver for this example. In support of these comments, we have also simulated rate $\frac{1}{2}$ versions of this same code structure so that no puncturing occurs for the SCCC and much less occurs for the PCCC. In this case, $(N, K) = (2048, 1024)$ and $\mathcal{S}$-random interleavers were used ($\mathcal{S} = 16$ for PCCC and $\mathcal{S} = 20$ for SCCC). The results are presented in Fig. 5, where we observe that the SCCC has a much lower error rate floor, particularly for the FER curves. Finally, we remark that $w \geq 4$ terms in (2) are necessary for an accurate estimate of the floor level of the SCCC case in Fig. 5.

4. THE ITERATIVE DECODERS

4.1. Overview of the Iterative Decoder

Consider first an ML decoder for a rate $\frac{1}{2}$ convolutional code (recursive or not), and assume a data word of length $K \geq 1000$. Ignoring the structure of the code, a naive ML decoder would have to compare (correlate) 2^K code sequences to the noisy received sequence, choosing in favor of the codeword with the best correlation metric. Clearly, the complexity of such an algorithm is exorbitant. Fortunately, as we know, such a brute-force approach is simplified greatly by the Viterbi algorithm, which permits a systematic elimination of candidate code sequences.

Unfortunately, we have no such luck with turbo codes, for the presence of the interleaver immensely complicates the structure of a turbo code trellis. A near-optimal solution is an iterative decoder (also called a *turbo decoder*) involving two soft-in/soft-out (SISO) decoders that share probabilistic information cooperatively and iteratively. The goal of the iterative decoder is to iteratively estimate the *a posteriori* probabilities (APPs) $\Pr(u_k \mid \mathbf{y})$ where u_k is the kth data bit, $k = 1, 2, \ldots, K$, and $\mathbf{y}$ is the received codeword in noise, $\mathbf{y} = \mathbf{c} + \mathbf{n}$. In this equation, we assume the components of $\mathbf{c}$ take values in the set $\{\pm 1\}$ (and similarly for $\mathbf{u}$) and $\mathbf{n}$ is a noise word whose components

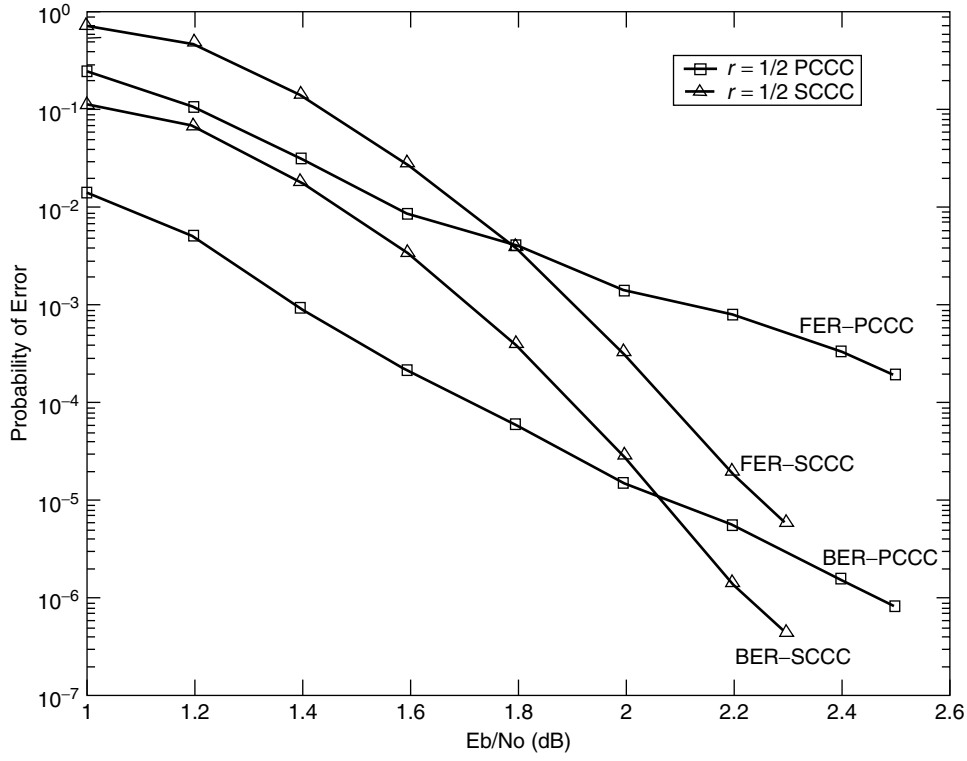


Figure 5. Rate $\frac{1}{2}$ PCCC and SCCC bit error rate (BER) and frame error rate (FER) simulation results.

are AWGN samples. Knowledge of the APPs allows for optimal decisions on the bits u_k via the maximum *a posteriori* (MAP) rule[1]

$$\frac{P(u_k = +1 \mid \mathbf{y})}{P(u_k = -1 \mid \mathbf{y})} \underset{-1}{\overset{+1}{\gtrless}} 1$$

or, more conveniently

$$\hat{u}_k = \text{sign } [L(u_k)]$$

where $L(u_k)$ is the log *a posteriori* probability (log-APP) ratio defined as

$$L(u_k) \triangleq \log\left(\frac{P(u_k = +1 \mid \mathbf{y})}{P(u_k = -1 \mid \mathbf{y})}\right) \tag{4}$$

We shall use the term *log-likelihood ratio* (LLR) in place of log-APP ratio for consistency with the literature.

The component SISO decoders that jointly estimate the LLRs $L(u_k)$ for parallel and serial turbo codes compute the LLRs for component code inputs (u_{ik}), component code outputs (c_{ik}), or both. Details on the SISO decoders will be presented below. For now, we simply introduce the convention that, for PCCCs, the top component encoder is encoder 1 (denoted E1) and the bottom component decoder is encoder 2 (denoted E2). For SCCCs, the outer encoder is encoder 1 (E1) and the inner encoder is encoder 2 (E2). The SISO component decoders matched to E1 and E2 will be denoted by D1 and D2, respectively. Because the SISO decoders D1 and D2 compute $L(u_{ik})$ and/or $L(c_{ik}), i = 1, 2$, we will temporarily use the notation $L(b_k)$ where b_k represents either u_{ik} or c_{ik}.

From Bayes' rule, the LLR for an arbitrary SISO decoder can be written as

$$L(b_k) = \log\left(\frac{P(\mathbf{y} \mid b_k = +1)}{P(\mathbf{y} \mid b_k = -1)}\right) + \log\left(\frac{P(b_k = +1)}{P(b_k = -1)}\right) \tag{5}$$

with the second term representing *a priori* information. Since $P(b_k = +1) = P(b_k = -1)$ typically, the *a priori* term is usually zero for conventional decoders. However, for *iterative* decoders, each component decoder receives *extrinsic* or *soft* information for each b_k from its companion decoder, which serves as *a priori* information. The idea behind extrinsic information is that D2 provides soft information to D1 for each b_k using only information not available to D1, and D1 does likewise for D2. For SCCCs, the iterative decoding proceeds as D2 $\rightarrow$ D1 $\rightarrow$ D2 $\rightarrow$ D1 $\rightarrow \ldots$, with the previous decoder passing soft information along to the next decoder at each half-iteration. For PCCCs, either decoder may initiate the chain of component decodings or, for hardware implementations, D1 and D2 may operate simultaneously.

This type of iterative algorithm is known to converge to the true value of the LLR $L(u_k)$ for the concatenated code provided that the graphical representation of this code contains no loops [15–17]. The graph of a turbo code does in fact contain loops [17], but the algorithm nevertheless provides near-optimal performance for virtually all turbo codes. That this is possible even in the presence of loops is not fully understood.

This section provided an overview of the turbo decoding algorithm in part to motivate the next section on SISO decoding of a single RSC code using the BCJR algorithm [18]. Following the description of the SISO decoder for a single RSC code will be sections that describe in full detail the iterative PCCC and SCCC decoders that utilize slightly modified SISO decoders.

4.2. The BCJR Algorithm and SISO Decoding

4.2.1. Probability Domain BCJR Algorithm for RSC Codes. Before we discuss the BCJR algorithm in the context of a turbo code, it is helpful to first consider the BCJR algorithm applied to a single rate $\frac{1}{2}$ RSC code on an AWGN channel. Thus, as indicated in Fig. 3, the transmitted codeword $\mathbf{c}$ will have the form $\mathbf{c} = [c_1, c_2, \ldots, c_K] = [u_1, p_1, u_2, p_2, \ldots, u_K, p_K]$ where $c_k \triangleq [u_k, p_k]$. The received word $\mathbf{y} = \mathbf{c} + \mathbf{n}$ will have the form $\mathbf{y} = [y_1, y_2, \ldots, y_K] = [y_1^u, y_1^p, y_2^u, y_2^p, \ldots, y_K^u, y_K^p]$, where $y_k \triangleq [y_k^u, y_k^p]$, and similarly for $\mathbf{n}$. As above, we assume our binary variables take values from the set $\{\pm 1\}$.

Our goal is the development of the BCJR algorithm [18] for computing the LLR

$$L(u_k) = \log\left(\frac{P(u_k = +1 \mid \mathbf{y})}{P(u_k = -1 \mid \mathbf{y})}\right)$$

given the received word $\mathbf{y}$. In order to incorporate the RSC code trellis into this computation, we rewrite $L(u_k)$ as

$$L(u_k) = \log \frac{\sum_{U+} p(s_{k-1} = s', s_k = s, \mathbf{y})}{\sum_{U-} p(s_{k-1} = s', s_k = s, \mathbf{y})} \tag{6}$$

where s_k is encoder state at time k, U^+ is set of pairs (s', s) for the state transitions $(s_{k-1} = s') \rightarrow (s_k = s)$, which correspond to the event $u_k = +1$, and U^- is similarly defined. To write (6) we used Bayes' rule, total probability, and then canceled $1/p(\mathbf{y})$ in the numerator and denominator. We see from (6) that we need only compute $p(s', s, \mathbf{y}) = p(s_{k-1} = s', s_k = s, \mathbf{y})$ for all state transitions and then sum over the appropriate transitions in the numerator and denominator. We now provide the crucial results that facilitate the computation of $p(s', s, \mathbf{y})$.

Result 1. The probability density function (pdf) $p(s', s, \mathbf{y})$ may be factored as

$$p(s', s, \mathbf{y}) = \alpha_{k-1}(s') \cdot \gamma_k(s', s) \cdot \beta_k(s) \tag{7}$$

where

$$\alpha_k(s) \triangleq p(s_k = s, \mathbf{y}_1^k)$$

$$\gamma_k(s', s) \triangleq p(s_k = s, y_k \mid s_{k-1} = s')$$

$$\beta_k(s) \triangleq p(\mathbf{y}_{k+1}^K \mid s_k = s)$$

and where $\mathbf{y}_a^b \triangleq [y_a, y_{a+1}, \cdots, y_b]$.

[1] It is well known that the MAP rule minimizes the probability of bit error. For comparison, the ML rule, which maximizes the likelihoods $P(\mathbf{y} \mid \mathbf{c})$ over the codewords $\mathbf{c}$, minimizes the probability of codeword error.

Proof: By several applications of Bayes' rule, we have

$$\begin{aligned} p(s', s, \mathbf{y}) &= p(s', s, \mathbf{y}_1^{k-1}, y_k, \mathbf{y}_{k+1}^K) \\ &= p(\mathbf{y}_{k+1}^K \mid s', s, \mathbf{y}_1^{k-1}, y_k) p(s', s, \mathbf{y}_1^{k-1}, y_k) \\ &= p(\mathbf{y}_{k+1}^K \mid s', s, \mathbf{y}_1^{k-1}, y_k) \\ &\quad \times p(s, y_k \mid s', \mathbf{y}_1^{k-1}) \cdot p(s', \mathbf{y}_1^{k-1}) \\ &= p(\mathbf{y}_{k+1}^K \mid s) \cdot p(s, y_k \mid s') \cdot p(s', \mathbf{y}_1^{k-1}) \\ &= \beta_k(s) \cdot \gamma_k(s', s) \cdot \alpha_{k-1}(s') \end{aligned}$$

where the fourth line follows from the third because the variables omitted on the fourth line are conditionally independent.

Result 2. The probability $\alpha_k(s)$ may be computed in a "forward recursion" via

$$\alpha_k(s) = \sum_{s'} \gamma_k(s', s) \alpha_{k-1}(s') \tag{8}$$

where the sum is over all possible encoder states.

Proof: By several applications of Bayes' rule and the theorem on total probability, we have

$$\begin{aligned} \alpha_k(s) &\triangleq p(s, \mathbf{y}_1^k) \\ &= \sum_{s'} p(s', s, \mathbf{y}_1^k) \\ &= \sum_{s'} p(s, y_k \mid s', \mathbf{y}_1^{k-1}) p(s', \mathbf{y}_1^{k-1}) \\ &= \sum_{s'} p(s, y_k \mid s') p(s', \mathbf{y}_1^{k-1}) \\ &= \sum_{s'} \gamma_k(s', s) \alpha_{k-1}(s') \end{aligned}$$

where the fourth line follows from the third due to conditional independence of $\mathbf{y}_1^{k-1}$.

Result 3. The probability $\beta_k(s)$ may be computed in a "backward recursion" via

$$\beta'_{k-1}(s') = \sum_{s} \beta_k(s) \gamma_k(s', s) \tag{9}$$

Proof: Applying Bayes' rule and the theorem on total probability, we have

$$\begin{aligned} \beta'_{k-1}(s') &\triangleq p(\mathbf{y}_k^K \mid s') \\ &= \sum_{s} p(\mathbf{y}_k^K, s \mid s') \\ &= \sum_{s} p(\mathbf{y}_{k+1}^K \mid s', s, y_k) p(s, y_k \mid s') \\ &= \sum_{s} p(\mathbf{y}_{k+1}^K \mid s) p(s, y_k \mid s') \\ &= \sum_{s} \beta_k(s) \gamma_k(s', s) \end{aligned}$$

where conditional independence led to the omission of variables on the fourth line.

The recursion for the $\{\alpha_k(s)\}$ is initialized according to

$$\alpha_0(s) = \begin{cases} 1, & s = 0 \\ 0, & s \neq 0 \end{cases}$$

which makes the reasonable assumption that the convolutional encoder is initialized to the zero state. The recursion for the $\{\beta_k(s)\}$ is initialized according to

$$\beta_K(s) = \begin{cases} 1, & s = 0 \\ 0, & s \neq 0 \end{cases}$$

which assumes that "termination bits" have been appended at the end of the data word so that the convolutional encoder is again in state zero at time K.

All that remains at this point is the computation of $\gamma_k(s', s) = p(s, y_k \mid s')$. Observe that $\gamma_k(s', s)$ may be written as

$$\begin{aligned} \gamma_k(s', s) &= \frac{P(s', s)}{P(s')} \cdot \frac{p(s', s, y_k)}{P(s', s)} \\ &= P(s \mid s') p(y_k \mid s', s) \\ &= P(u_k) p(y_k \mid u_k) \end{aligned} \tag{10}$$

where the event $'u_k$ corresponds to the event $s' \to s$. Note $P(s \mid s') = P(s' \to s) = 0$ if s is not a valid state from state s' and $P(s' \to s) = \frac{1}{2}$ otherwise (since we assume binary input encoders with equiprobable inputs). Hence, $\gamma_k(s', s) = 0$ if $s' \to s$ is not valid and, otherwise

$$\gamma_k(s', s) = \frac{P(u_k)}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\|y_k - c_k\|^2}{2\sigma^2} \right] \tag{11}$$

$$= \frac{1}{2\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_k^u - u_k)^2 + (y_k^p - p_k)^2}{2\sigma^2} \right] \tag{12}$$

where $\sigma^2 = N_0/2$.

In summary, we may compute $L(u_k)$ via (6) using (7), (8), (9), and (12). This "probability domain" version of the BCJR algorithm is numerically unstable for long and even moderate codeword lengths, and so we now present the stable "log domain" version of it.

4.2.2. Log-Domain BCJR Algorithm for RSC Codes. In the log-BCJR algorithm, $\alpha_k(s)$ is replaced by the *forward metric*

$$\begin{aligned} \tilde{\alpha}_k(s) &\triangleq \log(\alpha_k(s)) \\ &= \log\left(\sum_{s'} \alpha_{k-1}(s') \gamma_k(s', s) \right) \\ &= \log\left(\sum_{s'} \exp(\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s)) \right) \end{aligned} \tag{13}$$

where the *branch metric* $\tilde{\gamma}_k(s', s)$ is given by

$$\begin{aligned} \tilde{\gamma}_k(s', s) &= \log \gamma_k(s', s) \\ &= -\log(2\sqrt{2\pi}\sigma) - \frac{\|y_k - c_k\|^2}{2\sigma^2} \end{aligned} \tag{14}$$

We will see that the first term in (14) may be dropped. Note that (13) not only defines $\tilde{\alpha}_k(s)$ but also gives its recursion. These log-domain forward metrics are initialized as

$$\tilde{\alpha}_0(s) = \begin{cases} 0, & s = 0 \\ -\infty, & s \neq 0 \end{cases} \tag{15}$$

The probability $\beta_{k-1}(s')$ is replaced by the *backward metric*

$$\begin{aligned} \tilde{\beta}_{k-1}(s') &\triangleq \log(\beta_{k-1}(s')) \\ &= \log\left(\sum_s \exp \tilde{\beta}_k(s) + \tilde{\gamma}_k(s', s)\right) \end{aligned} \tag{16}$$

with initial conditions

$$\tilde{\beta}_K(s) = \begin{cases} 0, & s = 0 \\ -\infty, & s \neq 0 \end{cases} \tag{17}$$

under the assumption that the encoder has been terminated.

As before, $L(u_k)$ is computed as

$$\begin{aligned} L(u_k) &= \log \frac{\sum_{U^+} \alpha_{k-1}(s')\gamma_k(s', s)\beta_k(s)}{\sum_{U^-} \alpha_{k-1}(s')\gamma_k(s', s)\beta_k(s)} \\ &= \log\left[\sum_{U+} \exp(\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s))\right] \\ &\quad - \log\left[\sum_{U^-} \exp(\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s))\right] \end{aligned} \tag{18}$$

It is evident from (18) that the constant term in (14) may be ignored since it may be factored all the way out of both summations. At first glance, Eqs. (13)–(18) do not look any simpler than the probability domain algorithm, but we use the following results to obtain the simplification.

Result 4.

$$\max(x, y) = \log\left(\frac{e^x + e^y}{1 + e^{-|x-y|}}\right)$$

Proof: Without loss of generality, when $x > y$, the right-hand side equals x.

Now define

$$\max^*(x, y) \triangleq \log(e^x + e^y) \tag{19}$$

so that from Result 4, we obtain

$$\max^*(x, y) = \max(x, y) + \log(1 + e^{-|x-y|}) \tag{20}$$

This may be extended to more than two variables. For example

$$\max^*(x, y, z) \triangleq \log(e^x + e^y + e^z)$$

which may be computed pairwise according to the following result.

Result 5.

$$\max^*(x, y, z) = \max^*[\max^*(x, y), z]$$

Proof:

$$\begin{aligned} \text{RHS} &= \log[e^{\max^*(x,y)} + e^z] \\ &= \log[e^{\log(e^x + e^y)} + e^z] \\ &= \log[e^x + e^y + e^z] \\ &= \text{LHS} \end{aligned}$$

Given the function $\max^*(\cdot)$, we may now rewrite (13), (16), and (18) as

$$\tilde{\alpha}_k(s) = \max_{s'}{}^*[\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s)] \tag{21}$$

$$\tilde{\beta}_{k-1}(s') = \max_{s}{}^*[\tilde{\beta}_k(s) + \tilde{\gamma}_k(s', s)] \tag{22}$$

and

$$\begin{aligned} L(u_k) = &\max_{U^+}{}^*[\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s)] \\ &- \max_{U^-}{}^*[\tilde{\alpha}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k(s)] \end{aligned} \tag{23}$$

Figure 6 illustrates pictorially the trellis-based computations that these last three equations represent.

We see from Eqs. (21)–(23) how the log-domain computation of $L(u_k)$ is vastly simplified relative to the probability-domain computation. From (20), implementation of the $\max^*(\cdot)$ function involves only a two-input $\max(\cdot)$ function plus a lookup table for the "correction term" $\log(1 + e^{-|x-y|})$. Robertson et al. [11] have shown that a table size of 8 is usually sufficient.

Note that the correction term is bounded as

$$0 < \log(1 + e^{-|x-y|}) \leq \log(2) \simeq 0.693$$

so that $\max^*(x, y) \simeq \max(x, y)$ when $|\max(x, y)| \geq 7$. When $\max^*(x, y)$ is replaced by $\max(\cdot)$ in Eqs. (21) and (22), these recursions become forward and reverse Viterbi algorithms, respectively. The performance loss associated with this approximation in turbo decoding depends on the specific turbo code, but a loss of about 0.5 dB is typical [11].

Finally, observe the sense in which the BCJR decoder is a SISO decoder: the decoder input is the "soft decision" (unquantized) word $\mathbf{y} \in \mathbb{R}^{2K}$ and its outputs are the soft outputs $L(u_k) \in \mathbb{R}$ on which final hard decisions may be made according to (4). Alternatively, in a concatenated code context, these soft outputs may be passed to a companion decoder.

4.2.3. Summary of the Log-Domain BCJR Algorithm. We assume as above a rate $\frac{1}{2}$ RSC encoder, a data block $\mathbf{u}$ of length K, and that encoder starts and terminates in the zero state (the last m bits of $\mathbf{u}$ are so selected, where m is the encoder memory size). In practice, the value $-\infty$ used in initialization is simply some large-magnitude negative number.

Initialize $\tilde{\alpha}_0(s)$ and $\tilde{\beta}_K(s)$ according to Eqs. (15) and (17).
for $k = 1{:}K$
— get $y_k = [y_k^u, y_k^p]$
— compute $\tilde{\gamma}_k(s', s) = -\|y_k - c_k\|^2/2\sigma^2$ for all allowable state transitions $s' \to s$ (note $c_k = c_k(s', s)$ here)[2]
— compute $\tilde{\alpha}_k(s)$ for all s using the recursion (21)
end

[2] We may alternatively use $\tilde{\gamma}_k(s', s) = u_k y_k^u/\sigma^2 + p_k y_k^p/\sigma^2$. (See next section.)

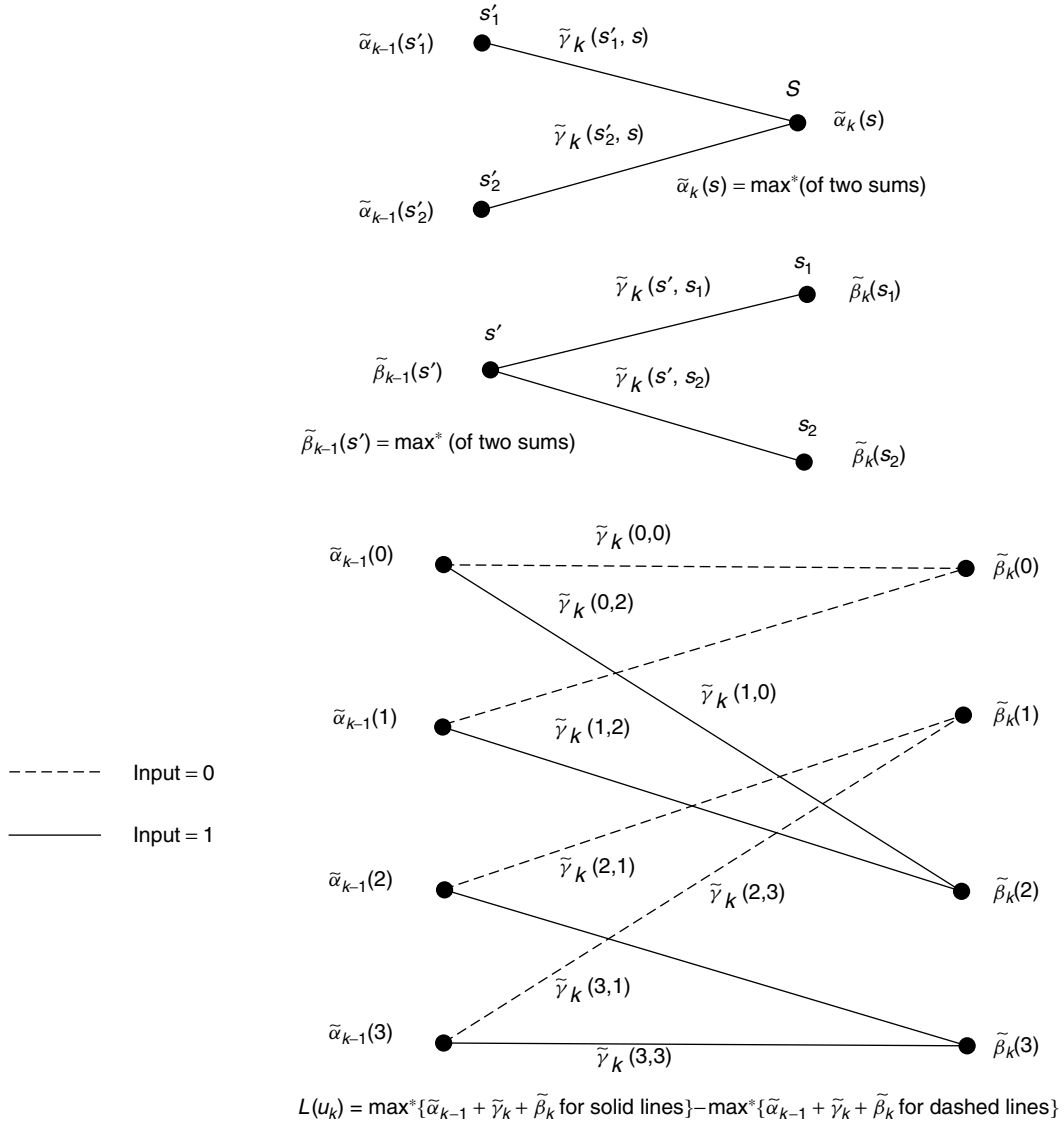


Figure 6. The top diagram depicts the forward recursion in (21), the middle diagram depicts the backward recursion in (22), and the bottom diagram depicts the computation of $L(u_k)$ via (23).

for $k = K: -1 : 2$
— compute $\tilde{\beta}_{k-1}(s')$ for all s' using (22)
end
for $k = 1{:}K$
— compute $L(u_k)$ using (23)
— compute hard decisions via $\hat{u}_k = \text{sign}\left[L(u_k)\right]$
end

4.3. The PCCC Iterative Decoder

We present in this section the iterative decoder for a PCCC consisting of two component rate $\frac{1}{2}$ RSC encoders concatenated in parallel. We assume no puncturing so that the overall code rate is $\frac{1}{3}$. Block diagrams of the PCCC encoder and its iterative decoder with component SISO decoders are presented in Fig. 7. As indicated in Fig. 7a, the transmitted codeword $\mathbf{c}$ will have the form $\mathbf{c} = [c_1, c_2, \ldots, c_K] = [u_1, p_1, q_1, \ldots, u_K, p_K, q_K]$ where $c_k \stackrel{\Delta}{=} [u_k, p_k, q_k]$. The received word $\mathbf{y} = \mathbf{c} + \mathbf{n}$ will have the form $\mathbf{y} = [y_1, y_2, \ldots, y_K] = [y_1^u, y_1^p, y_1^q, \ldots, y_K^u, y_K^p, y_K^q]$, where $y_k \stackrel{\Delta}{=} [y_k^u, y_k^p, y_k^q]$, and similarly for $\mathbf{n}$. We denote the codewords produced by E1 and E2 by, respectively, $\mathbf{c}_1 = [c_1^1, c_2^1, \ldots, c_K^1]$ where $c_k^1 \stackrel{\Delta}{=} [u_k, p_k]$ and $\mathbf{c}_2 = [c_1^2, c_2^2, \ldots, c_K^2]$ where $c_k^2 \stackrel{\Delta}{=} [u'_k, q_k]$. Note that $\{u'_k\}$ is a permuted version of $\{u_k\}$ and is not actually transmitted (see Fig. 7a). We define the noisy received versions of $\mathbf{c}_1$ and $\mathbf{c}_2$ to be $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively, having components $y_k^1 \stackrel{\Delta}{=} [y_k^u, y_k^p]$ and $y_k^2 \stackrel{\Delta}{=} [y_k^{u'}, y_k^q]$, respectively. Note that $\mathbf{y}_1$ and $\mathbf{y}_2$ can be assembled from $\mathbf{y}$ in an obvious fashion (using an interleaver to obtain $\{y_k^{u'}\}$ from $\{y_k^u\}$). By doing so, the component decoder inputs are the two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ as indicated in the Fig. 7b.

In contrast to the BCJR decoder of the previous sections whose input was $\mathbf{y} = \mathbf{c} + \mathbf{n}$ and whose output was $\{L(u_k)\}$ (or $\{\hat{u}_k\}$), the SISO decoders in Fig. 7b possess two inputs and two outputs. The SISO decoders are essentially the BCJR decoders discussed above, except that the SISO

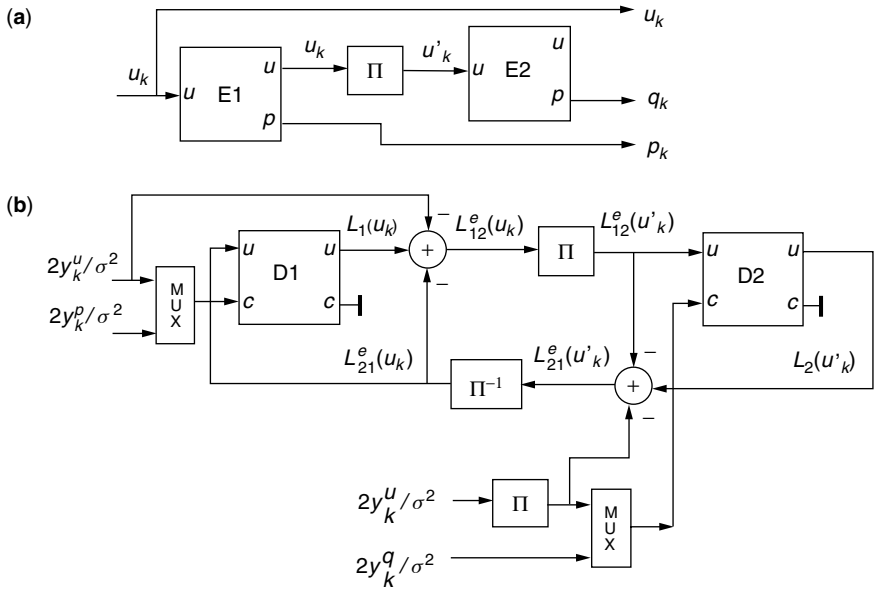


Figure 7. Block diagrams for the PCCC encoder (**a**) and iterative decoder (**b**).

decoders have the ability to accept from a companion decoder "extrinsic information" about its encoder's input (SISO input label '*u*') and/or about its encoder's output (SISO input label '*c*'). The SISO decoders also have the ability to produce likelihood information about its encoder's input (SISO output label '*u*') and/or about its encoder's output (SISO output label '*c*'). Note that the SISO decoder is to be interpreted as a decoding module not all of whose inputs or outputs need be used [7]. (Note that the RSC encoders in Fig. 7a have also been treated as modules.) As we will see, the SISO modules are connected in a slightly different fashion for the SCCC case.

Note from Fig. 7b that the extrinsic information to be passed from D1 to D2 about bit u_k, denoted $L_{12}^e(u_k)$, is equal to the LLR $L_1(u_k)$ produced by D1 minus the channel likelihood $2y_k^u/\sigma^2$ and the extrinsic information $L_{21}^e(u_k)$ that D1 had just received from D2. The idea is that $L_{12}^e(u_k)$ should indeed be extrinsic (and uncorrelated with) the probabilistic information already possessed by D2. As we will see, $L_{12}^e(u_k)$ is strictly a function of received E1 parity $\{y_k^p\}$ which is not directly sent to D2. Observe that $\{L_{12}^e(u_k)\}$ must be interleaved prior to being sent to D2 since E2 and D2 operate on the interleaved data bits u'_k. Symmetrical comments may be made about the extrinsic information to be passed from D2 to D1, $L_{21}^e(u_k)$ (e.g., it is a function of E2 parity and deinterleaving is necessary).

We already know from the previous section how the SISO decoders process the samples from the channel, $\mathbf{y}_i$ $(i = 1, 2)$, to obtain LLR's about the decoder inputs. We need now to discuss how the SISO decoders include the extrinsic information in their computations. As indicated earlier, the extrinsic information takes the role of *a priori* information in the iterative decoding algorithm [see (5) and surrounding discussion]:

$$L^e(u_k) \triangleq \log\left(\frac{P(u_k = +1)}{P(u_k = -1)}\right). \tag{24}$$

The *a priori* term $P(u_k)$ shows up in (11) in an expression for $\gamma_k(s', s)$. In the log domain, (11) becomes[3]

$$\tilde{\gamma}_k(s', s) = \log P(u_k) - \log(\sqrt{2\pi}\sigma) - \frac{\|y_k - c_k\|^2}{2\sigma^2} \tag{25}$$

Now observe that we may write

$$\begin{aligned} P(u_k) &= \left(\frac{\exp[-L^e(u_k)/2]}{1 + \exp\left[-L^e(u_k)\right]}\right) \cdot \exp\frac{u_k L^e(u_k)}{2} \\ &= A_k \exp\frac{u_k L^e(u_k)}{2} \end{aligned} \tag{26}$$

where the first equality follows since it equals

$$\left(\frac{\sqrt{P_-/P_+}}{1 + P_-/P_+}\right)\sqrt{P_+/P_-} = P_+ \text{ when } u_k = +1$$

$$\left(\frac{\sqrt{P_-/P_+}}{1 + P_-/P_+}\right)\sqrt{P_-/P_+} = P_- \text{ when } u_k = -1$$

where we have defined $P_+ \triangleq P(u_k = +1)$ and $P_- \triangleq P(u_k = -1)$ for convenience. Substitution of (26) into (25) yields

$$\tilde{\gamma}_k(s', s) = \log\left(\frac{A_k}{\sqrt{2\pi}\sigma}\right) + \frac{u_k L^e(u_k)}{2} - \frac{\|y_k - c_k\|^2}{2\sigma^2} \tag{27}$$

where we will see that the first term may be ignored.

Thus, the extrinsic information received from a companion decoder is included in the computation through the branch metric $\tilde{\gamma}_k(s', s)$. The rest of the BCJR/SISO algorithm proceeds as before using Eqs. (21)–(23).

[3] For the time being, we will discuss a generic SISO decoder so that we may avoid using cumbersome superscripts until it is necessary to do so.

Upon substitution of (27) into (23), we have

$$L(u_k) = L^e(u_k) + \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{u_k y_k^u}{\sigma^2} + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{u_k y_k^u}{\sigma^2} + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] \quad (28)$$

where we have applied the fact that

$$\|y_k - c_k\|^2 = (y_k^u - u_k)^2 + (y_k^p - p_k)^2 = (y_k^u)^2 - 2u_k y_k^u + u_k^2 + (y_k^p)^2 - 2p_k y_k^p + p_k^2$$

and only the terms dependent on U^+ or U^-, $u_k y_k^u/\sigma^2$ and $p_k y_k^p/\sigma^2$, survive after the subtraction. Now note that $u_k y_k^u/\sigma^2 = y_k^u/\sigma^2$ under the first $\max^*(\cdot)$ operation in (28) (U^+ is the set of state transitions for which $u_k = +1$) and $u_k y_k^u/\sigma^2 = -y_k^u/\sigma^2$ under the second $\max^*(\cdot)$ operation. Using the definition for $\max^*(\cdot)$, it is easy to see that these terms may be isolated out so that

$$L(u_k) = \frac{2y_k^u}{\sigma^2} + L^e(u_k) + \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k(s) \right] \quad (29)$$

The interpretation of this new expression for $L(u_k)$ is that the first term is likelihood information received directly from the channel, the second term is extrinsic likelihood information received from a companion decoder, and the third "term" ($\max_{U^+}{}^* - \max_{U^-}{}^*$) is extrinsic likelihood information to be passed to a companion decoder. Note that this third term is likelihood information gleaned from received parity not available to the companion decoder. Thus, specializing to decoder D1, for example, on any given iteration, D1 computes

$$L_1(u_k) = \frac{2y_k^u}{\sigma^2} + L_{21}^e(u_k) + L_{12}^e(u_k)$$

where $L_{21}^e(u_k)$ is extrinsic information received from D2, and $L_{12}^e(u_k)$ is the third term in (29) which is to be used as extrinsic information from D1 to D2.

4.3.1. Summary of the PCCC Iterative Decoder. The algorithm given below for the iterative decoding of a parallel turbo code follows directly from the development above. The constituent decoder order is D1, D2, D1, D2, and so on. Implicit is the fact that each decoder must have full knowledge of the trellis of the constituent encoders. For example, each decoder must have a table (array) containing the input bits and parity bits for all possible state transitions $s' \to s$. Also required are interleaver and de-interleaver functions (arrays) since D1 and D2 will be sharing reliability information about each u_k, but D2's information is permuted relative to D1. We denote these arrays by $P[\cdot]$ and $Pinv[\cdot]$, respectively. For example, the permuted word $\mathbf{u}'$ is obtained from the original word $\mathbf{u}$ via the pseudo-code statement: for $k = 1{:}K$, $u'_k = u_{P[k]}$, end.

We next point out that knowledge of the noise variance $\sigma^2 = N_0/2$ by each SISO decoder is necessary. Also, a simple way to obtain higher code rates via (simulated) puncturing is, in the computation of $\gamma_k(s', s)$, to set to zero the received parity samples, y_k^p or y_k^q, corresponding to the punctured parity bits, p_k or q_k. (This will set to zero the term in the branch metric corresponding to the punctured bit.) Thus, puncturing need not be performed at the encoder for computer simulations. We mention also that termination of encoder E2 to the zero state can be problematic due to the presence of the interleaver (for one solution, see Ref. 19). Fortunately, there is only a small performance loss when E2 is not terminated. In this case, $\beta_K(s)$ for D2 may be set to $\alpha_K(s)$ for all s, or it may be set to a nonzero constant (e.g., $1/S_2$, where S_2 is the number of E2 states).

Finally, we remark that some sort of iteration stopping criterion is necessary. The most straightforward criterion is to set a maximum number of iterations. However, this can be inefficient since the correct codeword is often found after only two or three iterations. Another straightforward technique is to utilize a carefully chosen outer error detection code. After each iteration, a parity check is made and the iterations stop whenever no error is detected. Other stopping criteria are presented in [9] and elsewhere in the literature.

4.3.1.1. Initialization

D1:

$\tilde{\alpha}_0^{(1)}(s) = 0$ for $s = 0$
$= -\infty$ for $\neq 0$

$\tilde{\beta}_K^{(1)}(s) = 0$ for $s = 0$
$= -\infty$ for $s \neq 0$

$L_{21}^e(u_k) = 0$ for $k = 1, 2, \ldots, K$

D2:

$\tilde{\alpha}_0^{(2)}(s) = 0$ for $s = 0$
$= -\infty$ for $s \neq 0$

$\tilde{\beta}_K^{(2)}(s) = \tilde{\alpha}_K^{(2)}(s)$ for all s (set once after computation of $\{\tilde{\alpha}_K^{(2)}(s)\}$ in the *first* iteration)

$L_{12}^e(u_k)$ is to be determined from D1 after the first half-iteration and so need not be initialized

4.3.1.2. The nth Iteration

D1:

for $k = 1{:}K$

— get $y_k^1 = [y_k^u, y_k^p]$

— compute $\tilde{\gamma}_k(s', s)$ for all allowable state transitions $s' \to s$ from (27) which simplifies to [see discussion following (27)]

$$\tilde{\gamma}_k(s', s) = \frac{u_k L_{21}^e(u_{Pinv[k]})}{2} + \frac{u_k y_k^u}{\sigma^2} + \frac{p_k y_k^p}{\sigma^2}$$

[u_k (p_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$]

— compute $\tilde{\alpha}_k^{(1)}(s)$ for all s using (21)

end

for $k = K{:}-1:2$

— compute $\tilde{\beta}_{k-1}^{(1)}(s)$ for all s using (22)

end

for $k = 1{:}K$

— compute $L_{12}^e(u_k)$ using[4]

$$L_{12}^e(u_k) = \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k^{(1)}(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k y_k^p}{\sigma^2} + \tilde{\beta}_k^{(1)}(s) \right]$$

end

D2:

for $k = 1{:}K$

— get $y_k^2 = [y_{P[k]}^u, y_k^q]$

— compute $\tilde{\gamma}_k(s', s)$ for all allowable state transitions $s' \to s$ from

$$\tilde{\gamma}_k(s', s) = \frac{u_k L_{12}^e(u_{P[k]})}{2} + \frac{u_k y_{P[k]}^u}{\sigma^2} + \frac{q_k y_k^q}{\sigma^2}$$

[u_k (q_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$]

— compute $\tilde{\alpha}_k^{(2)}(s)$ for all s using (21)

end

for $k = K{:}-1:2$

— compute $\tilde{\beta}_{k-1}^{(2)}(s)$ for all s using (22)

end

for $k = 1{:}K$

— compute $L_{21}^e(u_k)$ using

$$L_{21}^e(u_k) = \max_{U^+}{}^* \left[\tilde{\alpha}_{k-1}^{(2)}(s') + \frac{q_k y_k^q}{\sigma^2} + \tilde{\beta}_k^{(2)}(s) \right] - \max_{U^-}{}^* \left[\tilde{\alpha}_{k-1}^{(2)}(s') + \frac{q_k y_k^q}{\sigma^2} + \tilde{\beta}_k^{(2)}(s) \right]$$

end

[4] Note here we are computing $L_{12}^e(u_k)$ directly rather than computing $L_1(u_k)$ and then subtracting $2y_k^u/\sigma^2 + L_{21}^e(u_k)$ from it to obtain $L_{12}^e(u_k)$ as in Fig. 5(b). We will do likewise in the analogous step for D2.

4.3.1.3. After the Last Iteration

for $k = 1{:}K$

— compute

$$L_1(u_k) = \frac{2y_k^u}{\sigma^2} + L_{21}^e(u_{p_{inv}[k]}) + L_{12}^e(u_k)$$

— $\hat{u}_k = \text{sign}\,[L(u_k)]$

end

4.4. The SCCC Iterative Decoder

We present in this section the iterative decoder for an SCCC consisting of two component rate $\frac{1}{2}$ RSC encoders concatenated in series. We assume no puncturing so that the overall code rate is $\frac{1}{4}$. Higher code rates are achievable via puncturing and/or by replacing the inner encoder with a rate 1 differential encoder with transfer function $\frac{1}{1 \oplus D}$. It is straightforward to derive the iterative decoding algorithm for other SCCC codes from the special case that we consider here.

Block diagrams of the SCCC encoder and its iterative decoder with component SISO decoders are presented in Fig. 8. We denote by $\mathbf{c}_1 = [c_1^1, c_2^1, \ldots, c_{2K}^1] = [u_1, p_1, u_2, p_2, \ldots, u_K, p_K]$ the codeword produced by E1 whose input is $\mathbf{u} = [u_1, u_2, \ldots, u_K]$. We denote by $\mathbf{c}_2 = [c_1^2, c_2^2, \ldots, c_{2K}^2] = [v_1, q_1, v_2, q_2, \ldots, v_{2K}, q_{2K}]$ (with $c_k^2 \stackrel{\Delta}{=} [v_k, q_k]$) the codeword produced by E2 whose input $\mathbf{v} = [v_1, v_2, \ldots, v_{2K}]$ is the interleaved version of $\mathbf{c}_1$, that is, $\mathbf{v} = \mathbf{c}_1'$. As indicated in Fig. 8a, the transmitted codeword $\mathbf{c}$ is the codeword $\mathbf{c}_2$. The received word $\mathbf{y} = \mathbf{c} + \mathbf{n}$ will have the form $\mathbf{y} = [y_1, y_2, \ldots, y_{2K}] = [y_1^v, y_1^q, \ldots, y_{2K}^v, y_{2K}^q]$ where $y_k \stackrel{\Delta}{=} [y_k^v, y_k^q]$, and similarly for $\mathbf{n}$.

The iterative SCCC decoder in Fig. 8b employs two SISO decoding modules (described in the previous section). Note that unlike the PCCC case, these SISO decoders share extrinsic information on the code bits $\{c_k^1\}$ (equivalently, on the input bits $\{v_k\}$) in accordance with the fact that these are the bits known to both encoders. A consequence of this is that D1 must provide likelihood information on E1 *output* bits whereas D2 produces likelihood information on E2 *input* bits as indicated in Fig. 8b. However, because LLRs must be obtained on the original data bits u_k so that final decisions may be made, D1 must also compute likelihood information on E1 input bits. Note also that, because E1 feeds no bits directly to the

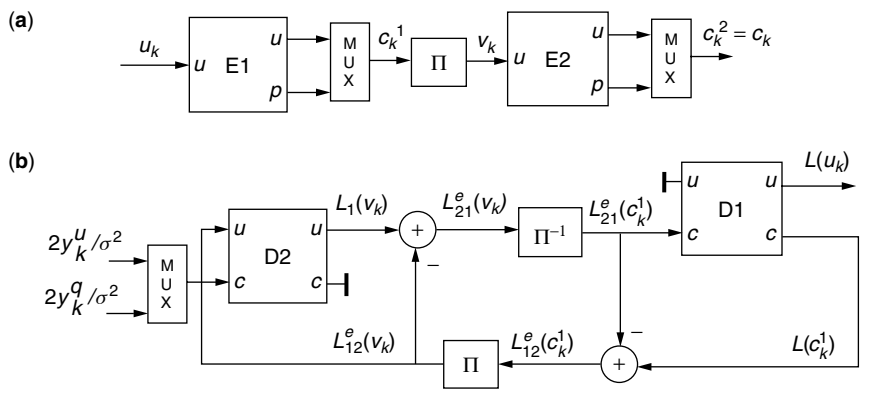


Figure 8. Block diagrams for the SCCC encoder (**a**) and iterative decoder (**b**).

channel, D1 receives no samples directly from the channel. Instead, the only input to D1 is the extrinsic information it receives from D2.

Thus, the SISO module D1 requires two features that we have not discussed in any detail to this point. The first is providing likelihood information on the encoder's input *and* output. However, since we assume the component codes are systematic, we need only compute LLRs on the encoder's output bits $[u_1, p_1, u_2, p_2, \ldots, u_K, p_K]$. Doing this is a simple matter of modifying the summation indices in (6) to those relevant to the output bit of interest. For example, the LLR $L(p_k)$ for the E1 parity bit p_k is obtained via

$$L(p_k) = \log \frac{\sum_{P^+} p(s_{k-1} = s', s_k = s, \mathbf{y})}{\sum_{P^-} p(s_{k-1} = s', s_k = s, \mathbf{y})} \tag{30}$$

where P^+ is set of state transition pairs (s', s) corresponding to the event $p_k = +1$, and P^- is similarly defined. (A trellis-based BCJR/SISO decoder is generally capable of decoding either the encoder's input or its output, whether or not the code is systematic. This is evident since the trellis branches are labeled by both inputs and outputs.)

The second feature is required by D1 is decoding with only extrinsic information as input. In this case the branch metric is simply modified as [cf. (27)]

$$\tilde{\gamma}_k(s', s) = \frac{u_k L^e_{21}(u_k)}{2} + \frac{p_k L^e_{21}(p_k)}{2} \tag{31}$$

Other than these modifications, the iterative SCCC decoder proceeds much like the PCCC iterative decoder and as indicated in Fig. 8b.

4.4.1. Summary of the SCCC Iterative Decoder. Essentially all of the comments mentioned for the PCCC decoder hold here as well and so we do not repeat them. The only difference is that the decoding order is D2, D1, D2, D1, and so on.

4.4.1.1. Initialization

D1:

$$\tilde{\alpha}^{(1)}_0(s) = 0 \text{ for } s = 0$$
$$= -\infty \text{ for } s \neq 0$$

$$\tilde{\beta}^{(1)}_K(s) = 0 \text{ for } s = 0$$
$$= -\infty \text{ for } s \neq 0$$

$L^e_{21}(c^1_k)$ is to be determined from D2 after the first half-iteration and so need not be initialized

D2:

$$\tilde{\alpha}^{(2)}_0(s) = 0 \text{ for } s = 0$$
$$= -\infty \text{ for } s \neq 0$$

$\tilde{\beta}^{(2)}_{2K}(s) = \tilde{\alpha}^{(2)}_{2K}(s)$ for all s (set after computation of $\{\tilde{\alpha}^{(2)}_{2K}(s)\}$ in the *first* iteration)

$L^e_{12}(v_k) = 0$ for $k = 1, 2, \ldots, 2K$

4.4.1.2. The nth Iteration

D2:

for $k = 1{:}\,2K$

— get $y_k = [y^v_k, y^q_k]$

— compute $\tilde{\gamma}_k(s', s)$ for all allowable state transitions $s' \to s$ from

$$\tilde{\gamma}_k(s', s) = \frac{v_k L^e_{12}(v_k)}{2} + \frac{v_k y^v_k}{\sigma^2} + \frac{q_k y^q_k}{\sigma^2}$$

[v_k (q_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$; $L^e_{12}(v_k)$ is $L^e_{12}(c^1_{P[k]})$, the interleaved extrinsic information from the previous D1 iteration.]

— compute $\tilde{\alpha}^{(2)}_k(s)$ for all s using (21)

end

for $k = 2K{:}\,-1 : 2$

— compute $\tilde{\beta}^{(2)}_{k-1}(s)$ for all s using (22)

end

for $k = 1{:}\,2K$

— compute $L^e_{21}(v_k)$ using

$$L^e_{21}(v_k) = \max^*_{V^+}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}^{(2)}_k(s)\right]$$
$$-\max^*_{V^-}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}^{(2)}_k(s)\right] - L^e_{12}(v_k)$$
$$= \max^*_{V^+}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + v_k y^v_k/\sigma^2 + q_k y^q_k/\sigma^2 + \tilde{\beta}^{(2)}_k(s)\right]$$
$$-\max^*_{V^-}\left[\tilde{\alpha}^{(2)}_{k-1}(s') + v_k y^v_k/\sigma^2 + q_k y^q_k/\sigma^2 + \tilde{\beta}^{(2)}_k(s)\right]$$

where V^+ is set of state transition pairs (s', s) corresponding to the event $v_k = +1$, and V^- is similarly defined.

end

D1:

for $k = 1{:}\,K$

— for all allowable state transitions $s' \to s$ set $\tilde{\gamma}_k(s', s)$ via

$$\tilde{\gamma}_k(s', s) = \frac{u_k L^e_{21}(u_k)}{2} + \frac{p_k L^e_{21}(p_k)}{2}$$
$$= \frac{u_k L^e_{21}(c^1_{2k-1})}{2} + \frac{p_k L^e_{21}(c^1_{2k})}{2}$$

[u_k (p_k) in this expression is set to the value of the encoder input (output) corresponding to the transition $s' \to s$; $L^e_{21}(c^1_{2k-1})$ is $L^e_{21}(v_{Pinv[2k-1]})$, the de-interleaved extrinsic information from the previous D2 iteration, and similarly for $L^e_{21}(c^1_{2k})$].

— compute $\tilde{\alpha}^{(1)}_k(s)$ for all s using (21)

end

for $k = K: -1 : 2$

— compute $\tilde{\beta}_{k-1}^{(1)}(s)$ for all s using (22)

end

for $k = 1:K$

— compute $L_{12}^e(u_k) = L_{12}^e(c_{2k-1}^1)$ using

$$L_{12}^e(u_k) = \max_{U^+}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - \max_{U^-}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - L_{21}^e(c_{2k-1}^1)$$

$$= \max_{U^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k L_{21}^e(p_k)}{2} + \tilde{\beta}_k^{(1)}(s)\right] - \max_{U^-}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{p_k L_{21}^e(p_k)}{2} + \tilde{\beta}_k^{(1)}(s)\right]$$

— compute $L_{12}^e(p_k) = L_{12}^e(c_{2k}^1)$ using

$$L_{12}^e(p_k) = \max_{P^+}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - \max_{P^-}{}^*[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)] - L_{21}^e(c_{2k}^1)$$

$$= \max_{P^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{u_k L_{21}^e(u_k)}{2} + \tilde{\beta}_k^{(1)}(s)\right] - \max_{P^-}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \frac{u_k L_{21}^e(u_k)}{2} + \tilde{\beta}_k^{(-1)}(s)\right]$$

end

4.4.1.3. After the Last Iteration

for $k = 1:K$

— for all allowable state transitions $s' \to s$ set $\tilde{\gamma}_k(s', s)$ via

$$\tilde{\gamma}_k(s', s) = \frac{u_k L_{21}^e(c_{2k-1}^1)}{2} + \frac{p_k L_{21}^e(c_{2k}^1)}{2}$$

— compute $L(u_k)$ using

$$L(u_k) = \max_{U^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)\right] - \max_{U^+}{}^*\left[\tilde{\alpha}_{k-1}^{(1)}(s') + \tilde{\gamma}_k(s', s) + \tilde{\beta}_k^{(1)}(s)\right]$$

— $\hat{u}_k = \text{sign}\left[L(u_k)\right]$

end

5. CONCLUSION

We have seen in this article the how and why of both parallel and serial turbo codes. That is, we have seen how to decode these codes using an iterative decoder, and why they should be expected to perform so well. The decoding algorithm summaries should be sufficient to decode any binary parallel and serial turbo codes, and can in fact be easily extended to the iterative decoding of any binary hybrid schemes. It is not much more work to figure out how to decode any of the turbo trellis-coded modulation (turbo TCM) schemes that appear in the literature. In any case, this article should serve well as a starting point for the study of concatenated codes (and perhaps graph-based codes) and their iterative decoders.

Acknowledgments

The author would like to thank Rajeev Ramamurthy and Bo Xia for producing Figs. 4 and 5, and Steve Wilson, Masoud Salehi, and John Proakis for helpful comments. He would also like to thank Cheryl Drier for typing the first draft.

BIOGRAPHY

William E. Ryan received his B.S. in electrical engineering degree from Case Western Reserve University, Cleveland, Ohio, in 1981, and his M.S. and Ph.D. degrees in electrical engineering from the University of Virginia, Charlottesville, in 1984 and 1988, respectively. He has been with The Analytic Sciences Corporation, Ampex Corporation, and Applied Signal Technology prior to his positions in academia. From 1993 to 1998 he was with the Electrical and Computer Engineering Department faculty at New Mexico State University, Las Cruces. Since August 1998, he has been with the Electrical and Computer Engineering Department at the University of Arizona, Tucson, where he is an associate professor. He is an associate editor for the *IEEE Transactions on Communications for Coding, Modulation, and Equalization*. His research interests are in coding and signal processing for data transmission and storage.

BIBLIOGRAPHY

1. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: turbo codes, *Proc. 1993 Int. Conf. Communications*, pp. 1064–1070.
2. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: turbo-codes, *IEEE Trans. Commun.* **44**: 1261–1271 (Oct. 1996).
3. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**: 55–67 (Jan. 1982).
4. S. Benedetto and G. Montorsi, Unveiling turbo codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* **42**: 409–428 (March 1996).
5. S. Benedetto and G. Montorsi, Design of parallel concatenated codes, *IEEE Trans. Commun.* **44**: 591–600 (May 1996).
6. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* **44**: 909–926 (May 1998).
7. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, *A Soft-Input Soft-Output Maximum a Posteriori (MAP) Module to Decode Parallel and Serial Concatenated Codes*, TDA Progress Report 42-127, Nov. 15, 1996.
8. D. Divsalar and F. Pollara, *Multiple turbo codes for Deep-Space Communications*, JPL TDA Progress Report, 42-121, May 15, 1995.
9. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**: 429–445 (March 1996).

10. L. Perez, J. Seghers, and D. Costello, A distance spectrum interpretation of turbo codes, *IEEE Trans. Inform. Theory* **42**: 1698–1709 (Nov. 1996).

11. P. Robertson, E. Villebrun, and P. Hoeher, A comparison of optimal and suboptimal MAP decoding algorithms operating in the log domain, *Proc. 1995 Int. Conf. Communications*, pp. 1009–1013.

12. A. Viterbi, An intuitive justification and a simplified implementation of the MAP decoder for convolutional codes, *IEEE JSAC* **16**: 260–264 (Feb. 1998).

13. O. Acikel and W. Ryan, Punctured turbo codes for BPSK /QPSK channels, *IEEE Trans. Commun.* **47**: 1315–1323 (Sept. 1999).

14. O. Acikel and W. Ryan, Punctured high rate SCCCs for BPSK/QPSK channels, *Proc. 2000 IEEE Int. Conf. Communications*, Vol. 1, pp. 434–439.

15. J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

16. B. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.

17. N. Wiberg, *Codes and Decoding on General Graphs*, Ph.D. dissertation, Univ. Linkoping, Sweden, 1996.

18. L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).

19. D. Divsalar and F. Pollara, turbo codes for PCS applications, *Proc. 1995 Int. Conf. Communications*, pp. 54–59.

CONSTRAINED CODING TECHNIQUES FOR DATA STORAGE

WIM M. J. COENE
HENK D. L. HOLLMANN
Philips Research Laboratories
Eindhoven, The Netherlands

1. INTRODUCTION

Modulation codes are one of the key elements in digital communication or storage systems such as a CD or DVD recorder, a hard-disk drive (HDD) in a computer, a modem, or a fax. A schematical form of a storage system is shown in Fig. 1. Here, two parts can be distinguished: the transmitting part, including the write channel in which one user stores data on the recording medium; and the receiving part of the system, including the read-channel in which the same or another user tries to restore the original information by reading out the data written on the medium.

In order to realize a sufficiently high level of reliability, the data are first encoded before being stored. This *channel encoding* typically comprises an error-correcting code (ECC) and a modulation code (MC). The channel encoder at the transmitting end consists of the error correction encoder and the modulation encoder, usually cascaded one after the other in that order.

Located at the receiving end of the channel are the physical signal detection with the read head scanning the information on the medium, followed by the bit detection module, which attempts to derive the written bits (also called *channel* bits) from the measured signals as reliably as possible. These blocks precede the channel decoding, which consists of the respective counterparts of the modules at the transmitting end, with first the MC decoder, followed by the ECC decoder.

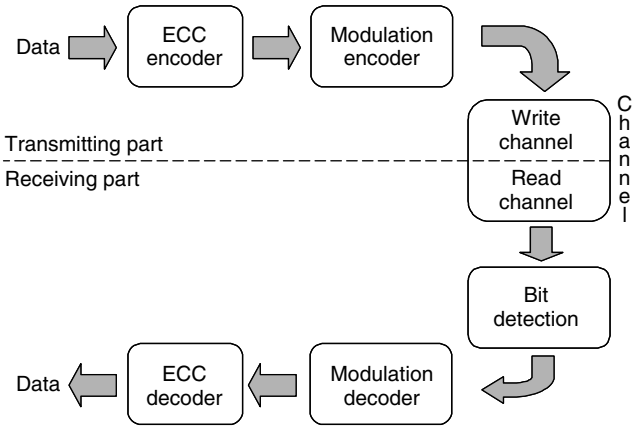


Figure 1. Schematic form of a digital storage system.

The ECC adds redundancy in the form of parity symbols, which makes it possible to restore the correct information in the presence of channel imperfections such as random errors and/or burst errors that may occur during readout from the medium. The modulation code serves to transform arbitrary (binary) sequences into sequences that possess certain "desirable" properties. Note the difference from error-correcting codes, which are used to ensure that *pairs* of encoded sequences have certain properties (i.e., being "very different"), while a modulation code serves to ensure certain properties of *each individual* encoded sequence. Which properties are desirable strongly depends on the particular storage or communication system for which the code is designed. For example, in most digital magnetic or optical recording systems, the stored sequences preferably contain neither very short nor very long runs of successive zeros or ones. The reason for this originates in how a stored sequence is read from the storage medium. The explanation is as follows.

In a storage system, the modulation of the physical signals is concerned with two physical conditions or states of the medium: (1) for magnetic recording, it is the magnetisation of magnetic domains in one of two opposite directions; (2) for optical recording, as shown in Fig. 2, it is the level of reflectivity (high and low) of the marks (or *pits*) and spaces (or *lands*) on the medium. One physical state can be associated with a channel bit (binary) 1, the other state with a channel bit (binary) 0. This representation, where the value of the channel bit represents the physical state (mark or space), is commonly known as *non-return-to-zero inverse* (NRZI), an old term originating from telegraphy. An equivalent representation of a channel bit stream is the *non-return-to-zero* (NRZ) notation, where a 1 bit indicates the start of a new mark or space on the medium, that is, a change of the physical state, and a 0 bit indicates the continuation of a mark or space.

A channel bit stream in NRZI notation can be partitioned into a sequence of phrases or *runs*, where each run consists of a number of consecutive channel bits of the

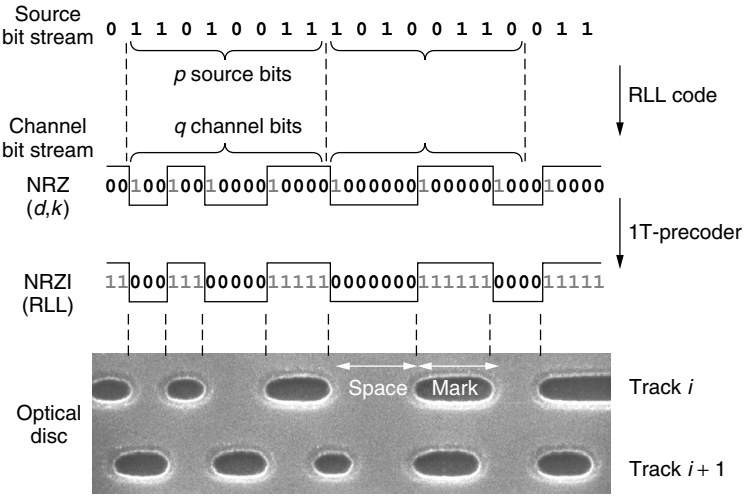


Figure 2. RLL coding in optical recording.

same type. So each run is associated with the physical area of a mark or space on the medium. The number of bits in a run is called the *run length*. As an example of a storage medium, we shall consider an optical disc in a little more detail. On the disk, data are organized in a single spiral; a small part of two successive revolutions of the spiral can be seen in Fig. 2. Along the track, physical marks and the spaces between them alternate. The marks and spaces have lengths that are multiples of the channel bit length T, and a mark or space of length nT represents a run with a run length of n bits.

Very short runs lead to small-signal amplitudes in the readout by the physical detection, and are therefore more prone to errors in the bit detection module (which is positioned directly after the read channel in Fig. 1). This is explained in more detail in Section 2. Moreover, very long runs lead to inaccuracies in the *timing recovery*, which is dealt with by a device called a *phase-locked loop* (PLL). The PLL regenerates the internal "clock" that is matched to the length of the bits on the medium; the bit clock is adjusted at each occurrence of a transition. Areas on the medium with too few transitions may cause "clock drift."

Avoiding very short and/or very long runs is achieved by using a run-length-limited (RLL) code, which typically constrains the allowable minimum and maximum run lengths that occur in the channel bit stream. The RLL constraints are described in terms of two parameters, d and k, and stipulate that the minimum and maximum run lengths are equal to $d+1$ and $k+1$, respectively. Note that the uncoded case corresponds to $d=0$ and $k=\infty$. In NRZ notation, a run of length $m+1$ is represented by a 1 bit followed by m 0 bits, that is, by 10^m in shorthand notation. Hence the (d,k) constraint in NRZ notation requires that the number of 0 bits between two successive 1 bits be at least d and at most k. Sequences that obey this constraint are called (d,k) *sequences*; a code for this constraint is referred to as a (d,k) *code*. Most RLL codes are constructed for a bit stream in NRZ notation. Subsequent transformation from NRZ to NRZI is required to obtain the channel bits that are actually written on the medium; such a transformation is formally carried out by a so-called 1T precoder, which is essentially an integrator modulo 2 (see Fig. 2).

A run-length constraint forms an example of a constraint that is specified by the absence of a finite number of "forbidden patterns." For example, a (d,k) constraint with $d>0$ can be specified in terms of the forbidden patterns $11, 101, \ldots, 10^{d-1}1$, and 0^{k+1}. Such a constraint is commonly referred to as a *constraint of finite type*. Many constraints that occur in practice are of this kind. Forbidding certain specific patterns implies that a sequence of source bits must be translated into a longer sequence of channel bits; the ratio of the length of the original sequence and the length of the encoded sequence is called the *rate* of the code.

Run-length-limited codes originated already in the 1960s through the work of Franaszek [1], Tang and Bahl [2], and others. Since then, various mechanisms for the construction of RLL and other modulation codes have been devised. A very elegant method is the ACH state-splitting algorithm [3], which will be discussed in Section 5. A detailed overview of RLL codes and their construction is given in Immink's book [4]; for other review articles, see, for example, Refs. 5 and 6.

The remainder of this survey is organised as follows. In Section 2, some practical aspects of the use of an RLL code are considered; Section 3 discusses the maximal coding rate (the capacity) of a modulation code; encoder and decoder structures are considered in Section 4; various code construction methods are described in Section 5; and finally Section 6 presents a collection of research topics and more recent trends in modulation coding.

2. PRACTICAL CONSIDERATIONS FOR THE USE OF AN RLL CODE

High-capacity storage applications employ such small bit sizes that the readout signal waveform generated by the physical detection for a given bit location does depend not only on that single bit but also on a limited number of neighboring bits. This bit-smearing effect is better known as *intersymbol interference* (ISI). For a simple read channel with linear readout characteristics, the ISI is

characterized by the impulse response of the channel, or, equivalently, by its Fourier transform, which yields the modulation transfer function (MTF) of the channel [7, Chap. 3.2]. The MTF indicates the (amplitude) response of the channel for each frequency in the system.

For storage systems, the MTF typically has a lowpass character: for instance, in optical recording, the MTF has an almost linear rolloff up to the cutoff frequency of the channel (see Fig. 3). Therefore, short run lengths in the channel bit stream, which lead to high-frequent signals, suffer most from ISI and are thus more prone to errors during read-out. One of the purposes of runlength-limited coding is to impose constraints that do not allow these high-frequency bit sequences; by doing so, the spectral content of the RLL-coded sequences is shaped to have a more lowpass character.

To illustrate this principle, we discuss the effect of employing three different d constraints, for $d = 0$ (uncoded), $d = 1$, and $d = 2$, while maintaining the same density of source bits on the storage medium in all three cases for fair comparison. So let T denote the common physical size of a source bit. Using a d-constrained code at a rate R_d, the physical channel bit size T_d will necessarily satisfy $T_d = R_d T$. Figure 4 shows the respective channel bit lengths and the highest frequency in the system (which correspond to an alternation of runs of minimum run length). Here, we of course have $R_0 = 1$ in the uncoded case. Furthermore, we assume that practical codes are

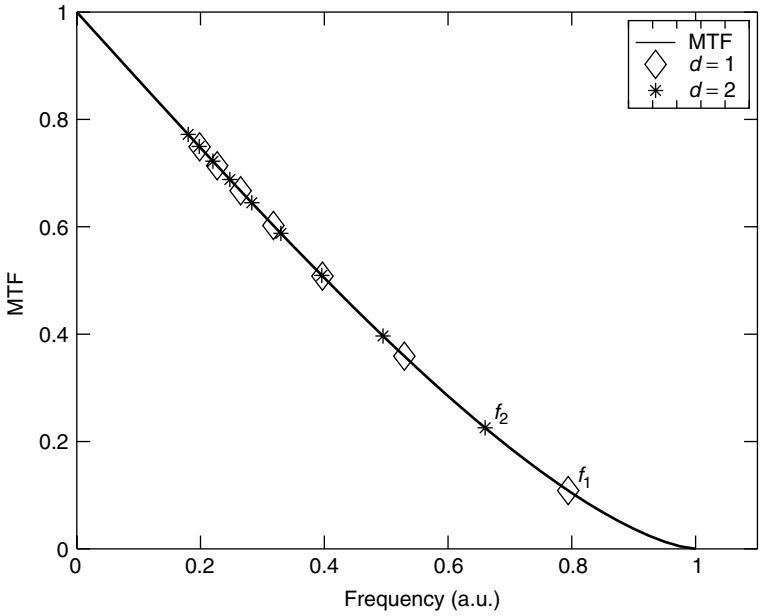


Figure 3. MTF for the optical recording channel as a function of frequency (in arbitrary units) with the frequencies of the pure tones $\ldots |nT_d \mid nT_d| \ldots$ superimposed.

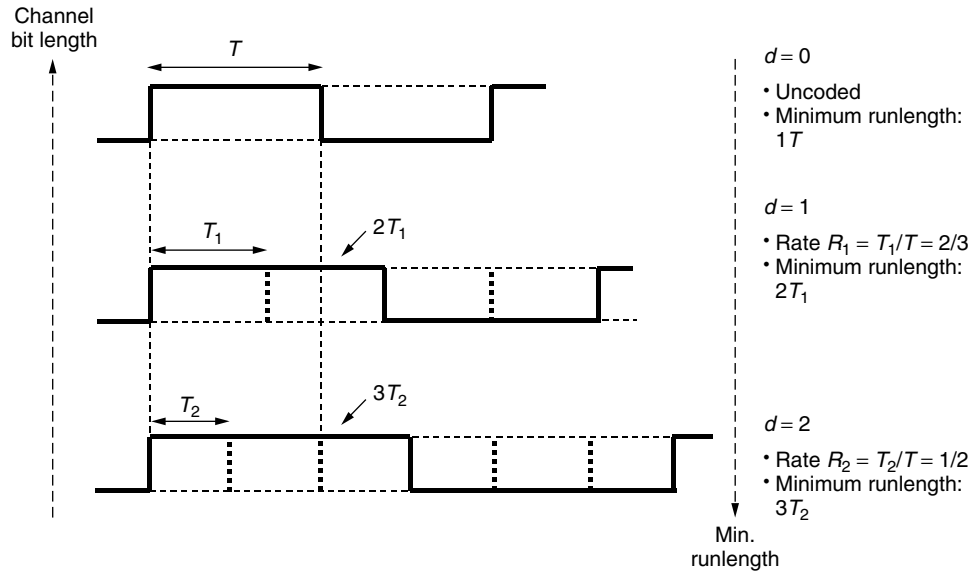


Figure 4. Channel bit length and minimum runlength for different d constraints at the same recording capacity in a storage channel.

used that have rates $R_1 = \frac{2}{3}$ and $R_2 = \frac{1}{2}$. These are reasonably close to the maximal achievable code rates of 0.6942 and 0.5515, respectively (see Section 3). For these rates, the minimum run length for $d = 1$ amounts to $2T_1 = 4/3T$, which is larger than the minimum run length T for $d = 0$; also, the minimum run length for $d = 2$ amounts to $3T_2 = 3/2T$, which is larger than the minimum run length for $d = 1$. Consequently, the highest frequencies f_d in the system for $d = 0, d = 1$, and $d = 2$ are

$$f_0 = \frac{1}{2T} > f_1 = \frac{1}{4R_1T} = \frac{3}{8T} > f_2 = \frac{1}{6R_2T} = \frac{1}{3T}$$

This relation reveals the increasing lowpass character of the code for increasing d constraint, which is the major attractiveness of RLL coding. This can also be observed from Fig. 3, where we have drawn the MTF (as a function of frequency) for the optical recording channel, with the frequencies of the pure tones $\ldots |nT_d \mid nT_d \mid nT_d| \ldots$ for $n = d + 1, d + 2, \ldots$ superimposed.

However, note that the channel bit length (or *timing window*, also known as *detection window*) decreases for increasing d constraint, which leads to a greater sensitivity with respect to *jitter* or *mark-edge noise* in the system. This counteracting effect favours the use of a *lower* d constraint.

The practical choice for a certain d constraint is a compromise between all pros and cons, and depends on many aspects, such as the actual bit detector used in the receiver (e.g., threshold detection, or some form of partial-response maximum-likelihood bit detection [7, Chap. 6]), the characteristics of the write channel, and the characteristics of the various noise sources in the system such as media noise and electronic noise.

3. CODING RATE AND CAPACITY

In the foregoing we have discussed various technical reasons for putting constraints on the sequences that we want to store or to send over a channel. Sequences that satisfy the constraints at hand are called *constrained sequences* or *codewords*, and the collection of all these sequences will be called the *constrained system*. A *modulation code* for a given constrained system consists of an *encoder* to translate arbitrary sequences into constrained sequences, and a *decoder* to recover the original sequence from the encoded sequence. The bits making up the original sequence and the encoded sequence are usually referred to as *source bits* and *channel bits*, respectively.

To achieve encoding, there is a price to be paid. Indeed, since there are more arbitrary sequences of a given length than there are constrained sequences of the same length, the encoding process will necessarily lead to an *increase* in the number of bits in the channel bit stream. This increase is measured by a number called the *rate* of the code. If, on the average, p source bits are translated into q channel bits, then the rate R of the code is $R = p/q$.

All other things being equal, we would, of course, like our code to have the highest possible rate. However, for all practical constraints there is a natural barrier for the code rate, called the *capacity* of the constraint, beyond which no encoding is possible. This discovery by Shannon [8] has been of great theoretical and practical importance. Indeed, once we know the capacity C of a given constrained system, then, on the one hand, we know that the best encoding rate that could possibly be achieved is bounded from above by C; on the other hand, once we have actually constructed a code with a rate R, the number R/C, called the *efficiency* of the code, serves as a benchmark for our engineering achievement.

In what follows, we shall sketch a derivation of Shannon's results. Consider a constrained system $\mathcal{L}$, and let N_n denote the number of constrained sequences of length n. Suppose that we can encode at a rate R. By definition, this means that, for large n, approximately Rn source bits are translated into n channel bits. There are 2^{Rn} distinct source sequences of length Rn, all of which need to be translated into distinct constrained sequences of length n, of which there are only N_n. Therefore, we necessarily have that $2^{Rn} \leq N_n$, or, equivalently, that $R \leq n^{-1} \log N_n$. (Here and in the sequel, all logarithms will be to the base 2.)

We now follow Shannon and *define* the capacity $C(\mathcal{L})$ of the constrained system $\mathcal{L}$ as

$$C(\mathcal{L}) = \lim_{n\to\infty} \frac{\log N_n}{n} \tag{1}$$

provided this limit exists.[1]

Intuitively, the capacity represents the *average amount of information* (measured in bits) *per bit of the constrained sequence*. The reasoning described above shows that encoding at a rate R is possible only if $R \leq C(\mathcal{L})$ and, moreover, suggests the possibility of encoding at rates arbitrarily close to $C(\mathcal{L})$, for example, by using long codewords glued together by a few suitably chosen merging bits.

It turns out that the kind of systems encountered in practice can typically be described in terms of a finite labeled directed *graph*. Here, a graph $G = (V, A)$ consists of a collection V of vertices or *states*, and a collection A of labeled arcs or *transitions* between states. Some authors refer to a "labeled directed graph" as defined above as a *finite-state transition diagram* (FSTD). The constrained system $\mathcal{L}(G)$ *presented* by G consists of all sequences that can be obtained by reading off the labels along a (directed) *path* in the graph, or, as we shall see, the sequences that are *generated* by paths in G. We shall refer to the graph G as a *presentation* of the system $\mathcal{L}(G)$.

Constrained systems that can be presented by some finite labeled directed graph as explained above are called *sofic systems*. Sofic systems are of great theoretical and practical importance. It turns out that about every constrained system encountered in practical applications is in fact a sofic system. (This is less remarkable than it

[1] For *subword-closed* systems, the fact that this limit exists immediately follows from Fekete's lemma [9, p. 233; 10]; namely, if the non-negative numbers a_n are such that $a_{m+n} \leq a_m + a_n$ for all $n, m \geq 0$, then $a^* = \lim_{n\to\infty} a_n/n$ exists and $a^* \leq a_n/n$ for all n. Indeed, if N_n is the number of sequences of length n contained in a subword-closed system, then obviously $N_{m+n} \leq N_m N_n$; hence an application of this lemma with $a_n = \log N_n$ shows that subword-closed systems have a well-defined capacity.

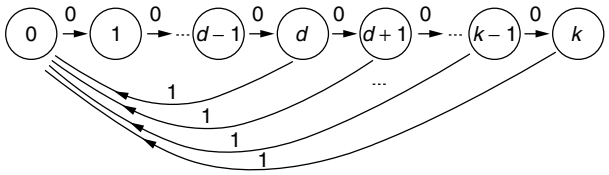


Figure 5. Presentation of a (d, k)-constrained system.

seems once we realize that about any digital device that we build is actually a *finite-state device*, whose possible output sequences necessarily constitute some sofic system.)

For example, a (d, k)-constrained system can be presented by the graph in Fig. 5. Here, the state label indicates the number of terminal zeros of a sequence generated by a path ending in that state. In the case where $k = \infty$, the constraint can be presented by a similar $(d + 1)$-state graph, where state d now indicates *at least* d terminal zeros.

We say that a presentation G is *irreducible* if there is a path in G from any state to any other state; G is said to be *primitive* if moreover all these paths may be chosen to have a fixed length h for some positive integer h.

It can be shown that each presentation can be broken down into irreducible parts. For coding purposes, it is then sufficient to consider only the "richest" component, that is, the component that presents the system with the largest capacity (which is then the capacity of the entire system). An irreducible presentation that is not primitive is in fact s-partite for some integer $s > 1$. That is, the set of states V can be partitioned into sets $V_0, \ldots, V_{s-1}$ such that each arc that starts in some V_i terminates in V_{i+1} (or in V_0 if $i = s - 1$). In that case, the sth power of the presentation (the presentation that generates s labels per arc; see Section 5) consists of s disjoint primitive components, which all have the same capacity sC, where C is the capacity of the original system.

So, for coding purposes we may in general assume that the presentation of the constrained system at hand is primitive. Note that primitivity of the presentation is precisely the property needed to assure that a *fixed* number of h merging bits, for some h, can always be used to glue any two codewords together; the merging bits can be read off from a path of length h connecting the terminal state of a path generating the first codeword to the initial state of the path generating the second codeword [see the discussion following Eq. (1) above]. For further details and proofs, the interested reader is referred, for example, to Ref. 10.

The (labeling of a) presentation is called *lossless* if any two paths with the same initial state and terminal state generate different sequences. This is an important property in connection with capacity computations. Indeed, let $G = (V, A)$ be a presentation, with $m = |V|$ states, say. If G is lossless, then at most m^2 paths can generate a given sequence, so the number P_n of paths of length n in G and the number N_n of sequences of length n presented by G are related by

$$P_n/m^2 \leq N_n \leq P_n \tag{2}$$

Hence the numbers P_n and N_n exhibit the same growth rate.

The growth rate of the numbers P_n can be computed from a matrix describing the underlying graph of the presentation, defined as follows. The adjacency matrix D of G is an $m \times m$ matrix where the entry $D(s, t)$ counts the number of arcs going from state s to state t. Note that the (s, t) entry of the nth power D^n of D counts the number of paths of length n in the graph from state s to state t. So the sum of the entries in D^n equals the number P_n of paths of length n in the graph:

$$P_n = \sum_{s,t \in V} D^n(s, t) \tag{3}$$

Now we can use a result from the classical Perron–Frobenius theory for nonnegative matrices (see e.g., Refs. 11 and 12; note that the adjacency matrix D is of this type), which states that the largest real "Perron–Frobenius" eigenvalue λ_D of a nonnegative matrix D equals its spectral radius and determines the growth rate of the entries of D^n; in fact we have [e.g., 10]

$$\lim_{n \to \infty} \frac{1}{n} \log \sum_{s,t \in V} D^n(s, t) = \log \lambda_D \tag{4}$$

Combining Eqs. (1)–(4), we conclude that the capacity $C(\mathcal{L})$ of a sofic system $\mathcal{L}$ with lossless presentation G is given by

$$C(\mathcal{L}) = \log \lambda_D \tag{5}$$

where λ_D is the largest real eigenvalue of the adjacency matrix D of G. This eigenvalue can be computed, for instance, as the largest real root of the *characteristic equation*

$$\det(\lambda I - D) = 0 \tag{6}$$

Alternatively, the capacity can be obtained by setting up linear recurrence relations for the numbers $P_n(s)$ of paths of length n starting in state s. The theory of such recurrences implies that these numbers grow as $c_s \lambda^n$, where λ is the largest real zero of the characteristic equation associated with this recurrence relation, which turns out to be the same as (6). For further details, see, for example, the book by Immink [13].

For example, the presentation of the (d, k)-constrained system in Fig. 5 is lossless. So we can apply the abovementioned method, and it turns out that the maximum code rate (the capacity) of a (d, k) code is given by the logarithm of the largest real root of the equation

$$x^{k+2} - x^{k+1} - x^{k+1-d} + 1 = 0 \tag{7}$$

if k is finite or

$$x^{d+1} - x^d - 1 = 0 \tag{8}$$

if $k = \infty$.

There are various properties of (the labeling of) a presentation that imply losslessness. All these properties are in fact concerned with the number of paths that generate a given sequence and with their localization in the graph. For example, a presentation has *finite local anticipation* if there is a number a such that any two paths of length

$a+1$ with the same initial state and generating the same sequence have the same initial arc. Similarly, the presentation has *finite local coanticipation* if the presentation obtained by reversing the direction of all arcs has finite local anticipation.

A presentation is *of finite type* if there exists a pair of numbers (m, a) (referred to as the *memory* and *anticipation* of the type) such that all paths that generate a given constrained sequence are equal, with the possible exception of at most m initial and a terminal arcs. Thus, given a (long) constrained sequence, we can reconstruct the path used to generate the sequence, up to a few arcs at the beginning and the end of the path. It turns out that a constrained system is of finite type (i.e., it can be described in terms of a finite number of "forbidden patterns") if and only if it has *some* presentation of finite type [14].

For future reference, we also introduce a slightly weaker property. A labeling is of almost finite type if it has both finite local anticipation and finite local coanticipation. As suggested by this terminology, it can be shown that a labeling of finite type is of almost finite type. A constrained system is said to be of almost finite type if it can be presented by some system of almost finite type.

A labeling with local anticipation 0 has the property that for any state, the outgoing arcs carry distinct labels. Such a labeling is called *deterministic*. This is an important property, for several reasons. Most "natural" presentations of constrained systems [e.g., the presentation of (d, k)-constrained systems in Fig. 5] are already deterministic. Moreover, any presentation G can be transformed into a deterministic presentation G^* by using a variant of the well-known *subset construction* for finite automata (see, e.g., Ref. 15 or 16). Here, G^* has as states all nonempty subsets of states of G, and for each such subset S and each labeling symbol a, the presentation G^* will have an arc $S \xrightarrow{a} T$, where T consists of all states t for which G has an arc $s \xrightarrow{a} t$ for some s in S. It is not difficult to see that indeed G and G^* present *the same* sofic system.

At this point, it is important to realize that a given sofic system can have (infinitely) many different presentations. Fortunately, every *irreducible* sofic system (i.e., one that has an irreducible presentation) has a unique *minimal* (in terms of the number of states) *deterministic* presentation, called the *Shannon cover*. It can be constructed from any irreducible deterministic presentation by an operation called *state merging*, in the following way. The set of sequences that can be generated by paths that start in a given state is called the *follower set* of that state. Now, if two states have the same follower set, then for sequence generation purposes these states are equal and can therefore be combined or *merged* into one state. Now, we repeatedly apply state merging until the follower sets in the states are all distinct. For further details about this construction, we refer to the book by Lind and Marcus [10].

The Shannon cover is important because it is "small" (hence very suitable for capacity computations), and also because many properties of a sofic system can be determined directly from its Shannon cover. For example, a sofic system is of finite type (respectively, of almost finite type) if and only if the labeling of its Shannon cover is of finite type (respectively, is of almost finite type). As a consequence, the Shannon cover is the natural starting point of many code construction methods.

4. ENCODERS AND DECODERS

An *encoder* for a given constrained system $\mathcal{L}$ is a device that transforms arbitrary binary sequences of *source bits* into sequences contained in $\mathcal{L}$. Commonly, the encoder is realized as a *synchronous finite-state device*. Such a device takes *source symbols*, groups of p consecutive source bits, as its input, and translates these into q-bit codewords, where the actual output depends on the input and possibly on the *internal state*, the (necessarily finite) content of an internal memory, of the device. The rate of such an encoder is then $R = p/q$. (To stress the individual values of p and q, we sometimes speak of a "rate $p \to q$" encoder.) Obviously, each codeword must itself satisfy the given constraint. Moreover, the encoder needs to ensure that the bit stream made up of successive codewords produced by the encoder also satisfies the constraint.

The use of the word "code" implies that it should be possible to recover or *decode* the original sequence of p-bit source symbols from the corresponding sequence of q-bit codewords. In practice, a stronger requirement is needed. Since modulation codes are typically used in the context of a noisy channel, it is important that the decoder limits the propagation of input errors. Therefore, we commonly require that the modulation code can be decoded by a *sliding-block decoder*.

A sliding-block decoder for a rate $p \to q$ encoder takes a sequence of q-bit words y_n as its input, and produces a sequence of p-bit symbols x_n as its output, where each output symbol x_n depends only on a corresponding sequence $y_{n-m}, \ldots, y_{n+a}$ of $w = m + 1 + a$ consecutive inputs, for some fixed numbers m and a, $m \le a$. We will refer to the number w as the *window size* of the decoder. (The numbers m and a are referred to as the *memory* and *anticipation* of the decoder.) The name "sliding-block decoder" refers to the image of the decoder sliding a "window" of width w over the sequence to be decoded. In Fig. 6 we depict a sliding-block decoder with $m = 2$, $a = 1$, and

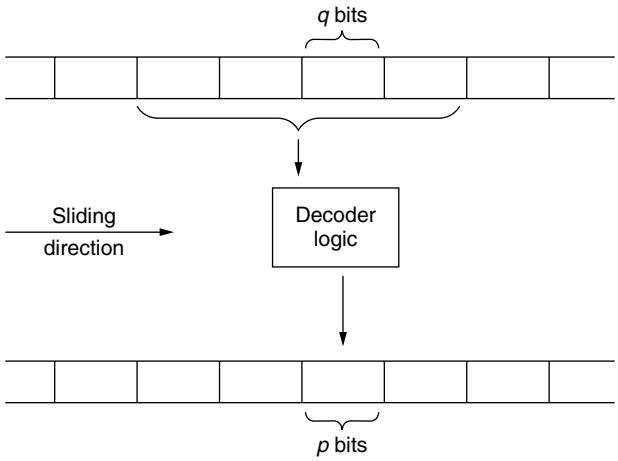


Figure 6. A sliding-block decoder.

window size $w = 4$. Note that an error in the encoded sequence will cause at most w symbol errors in the original sequence. So the *error propagation*, the amount of erroneous bits in the decoder output caused by a single input error, is limited to at most wp bits.

The window size of the decoder is an important parameter of a code: it provides an upper bound on the error propagation of the code, and also gives a good indication of the *size* of the decoder, the amount of hardware necessary to implement the decoder (i.e., it provides an upper bound on the *complexity* of the decoding operation).

Codes with a window size of one codeword (i.e., for which $w = 1$) are called *block-decodable*. For such codes, decoding a given codeword or *block* can be achieved without any knowledge of preceding or succeeding codewords. In many present-day applications, modulation codes are used in combination with symbol-error-correcting codes based on Reed–Solomon codes over some finite (Galois) field GF(2^p). In that situation, the use of a rate $p \to q$ block-decodable modulation code becomes especially attractive, since then a channel-bit error affects at most one p-bit symbol of a Reed–Solomon codeword. Note that this cannot be guaranteed if a non-block-decodable code is used, even if its decoding window size is smaller than q bits.

Note that decoding of a code requires a form of *synchronisation* between encoder and decoder in order to enable the decoder to identify the codeword boundaries in the encoded sequence. In practice this is usually achieved by using *frame-based* codes. Here, codewords are grouped into *frames* and unique *frame headers* are used to signal the beginning of a new frame. For further details, we again refer to the treatise by Immink [13].

5. CODE CONSTRUCTION METHODS

In the simplest case, the encoder translates each source symbol into a unique corresponding codeword, according to some code table. Of course, this will produce an admissible sequence only if any concatenation of these codewords also satisfies the given constraint. Codes of this type are called *block codes* or *block-encodable codes*. Decoding then consists of simply translating the codewords back into their corresponding source symbol, so these codes are block-decodable.

For example, consider the $(1, \infty)$-constrained system, in which the pattern 11 is not allowed to occur in the sequence. It is easily seen that the code where the source symbol 0 is encoded into the codeword 00 and 1 is encoded into 01 is a rate-$\frac{1}{2}$ block-decodable $(1, \infty)$ code. This simple code is called the *frequency modulation* (FM) or *biphase* code.

A slight improvement of this code is the *modified frequency modulation* (MFM) or *Miller* code. This is a rate $1 \to 2$ $(1, 3)$ code that encodes by inserting a *merging bit* between each two consecutive source bits. Here, the merging bit is set to 0 except when both surrounding bits are 0, in which case it is set to 1. Note that the FM code is obtained if the merging bit is always set to 0. Although the MFM code is not a block code, it is still block-decodable. In fact, decoding consists of simple deletion of the merging bits.

The use of merging bits is a well-established technique for constructing block-decodable (d, k) codes. They are often employed in combination with $(dklr)$ sequences, that is, (d, k)-constrained sequences in which the number of leading and trailing zeros is restricted to be at most l and r, respectively [17]. Here, p-bit source symbols are uniquely translated into $(dklr)$ sequences of a fixed length q'; in addition, a fixed number of $q - q'$ merging bits, chosen according to a set of *merging rules*, is employed to ensure that the resulting bit stream is a valid (d, k) sequence. Possible freedom in choosing these merging bits can then be used, for instance, to limit the DC content of the resulting sequence. These ideas can be found, for example, in the EFM recording code for the compact disk (see also Section 6). For some additional information on these methods, we refer to the paper by Weber and Abdel–Ghaffar [18]. A similar idea has been applied [19] to construct *almost-block-decodable* (d, k) codes.

Next, we shall discuss a number of code construction methods that employ an *approximate eigenvector* to guide the construction. We shall first explain this concept and motivate its importance for code construction.

Consider a given constrained system $\mathcal{L}$, presented by some (irreducible deterministic) graph G. Suppose that we wish to design a code with rate $p \to q$, where we assume that $p/q \leq C(\mathcal{L})$. Since we are interested in admissible sequences of q-bit codewords, it is convenient to consider the qth power graph G^q of G. This graph has the same states as G, and has an arc for each path of length q in G, labeled with the q-bit sequence generated by this path. Obviously, G^q essentially generates the same sequences as G, but does so with q bits or one codeword at the time. Note that if D is the adjacency matrix of G, then the adjacency matrix of G^q is given by D^q.

The desired encoder encodes arbitrary sequences of p-bit symbols into sequences of codewords; we shall refer to the collection of all these codeword sequences as the *code system* of the code. Note that each such sequence corresponds to one or more *encoding paths* in G^q.

Intuitively, we would expect that the number $\phi_s^{(n)}$ of encoding paths of length n in G^q that start in state s grows exponentially in 2^p, the number of source symbols. Indeed, under certain assumptions, it can be shown [20] that for each state s the limit

$$\phi_s = \lim_{n \to \infty} \frac{\phi_s^{(n)}}{2^{pn}} \tag{9}$$

exists and takes on an *integer value*. We will write ϕ for the vector with as entries the numbers ϕ_s.

Since an encoding path of length n starting in s consists of a transition in G^q from s to some state t, say, followed by an encoding path of length $n - 1$ starting in t, the vector $\phi^{(n)}$ with the numbers $\phi_s^{(n)}$ as entries satisfies

$$\phi^{(n)} \leq D^q \phi^{(n-1)} \tag{10}$$

(Note that we have an inequality here since not each path of length n obtained in this way needs to be an encoding path.) If we now combine Eqs. (9) and (10), we conclude that

$$2^p \phi \leq D^q \phi \tag{11}$$

A nonnegative integer vector ϕ that satisfies the inequality (11) is called a $[D^q, 2^p]$-*approximate eigenvector*, or a $[G^q, 2^p]$-*approximate eigenvector* if we wish to stress the connection with the presentation. We think of the numbers ϕ_s as weights associated with the states. In terms of these weights, the inequalities state that the sum of the weights of the (terminal states of) arcs leaving a state is at least 2^p times the weight of this state. The discussion above makes precise the idea that in code construction, the weight of a state is an indicator for the *relative encoding power* of that state.

It can be shown [12] that a $[G^q, 2^p]$-approximate eigenvector exists if and only if $p/q \leq C$, the capacity of the system presented by G. The following simple algorithm, first introduced by Franaszek [21], produces such an approximate eigenvector with components not larger than a given number M, provided such a vector exists. In this algorithm, we initially set $\phi_s = M$ for all states s, and then repeatedly perform the operation

$$\phi \leftarrow \min\{\phi, \lfloor 2^{-p} D^q \phi \rfloor\} \tag{12}$$

(where both the rounding operation $\lfloor\ \rfloor$ and taking the minimum are performed componentwise), until either $\phi = 0$ (in which case no such approximate eigenvector exists) or the vector ϕ remains unchanged (in which case ϕ is the desired approximate eigenvector).

Approximate eigenvectors were first introduced by Franaszek. In a series of pioneering papers [1,21–24], he developed a number of code construction methods that all employ approximate eigenvector in an essential way.

For example, the *principal-state method* is based on the observation that the existence of a *binary* $[D^q, 2^p]$-approximate eigenvector is equivalent to the existence of a set of *principal states* with the property that from each principal state there are (at least) 2^p distinct paths of length q in G (i.e., 2^p arcs in G^q) ending in another principal state. (The principal states are the states with weight one.) If it exists, a binary approximate eigenvector can be obtained by the recursive elimination procedure (12) by taking $M = 1$.

Given such a collection of paths in G, we can immediately construct a rate $p \rightarrow q$ encoder as follows. In each principal state, assign each of the 2^p possible source symbols to a q-bit codeword label of a path leaving this state. Now, the encoder moves from one principal state to another, using, for example, an encoding table in each principal state to translate input symbols into codewords. Usually, the encoding tables are implemented by using *enumerative methods* [13,25]. Such codes are called *fixed-length principal-state codes*.

If the constraint is of finite type, then any assignment of source symbols to codewords will lead to a sliding-block decodable code: in that case, the sequence of principal states traversed by the encoder can be reconstructed from the sequence of codewords.

The principal-state method may lead to prohibitively large values of p and q. (Note that these values have a large impact both on the complexity of the source-to-codeword assignment and on the size of the encoding tables.) For example, the minimum codeword lengths of a fixed-length principal-state rate-$\frac{2}{3}(1, 7)$ code and a rate-$\frac{1}{2}(2, 7)$-code are 33 and 34, respectively. Sometimes this problem can be avoided if we allow the coding paths between principal states to have varying lengths. As an example, we consider the design of a rate-$\frac{1}{2}(2, \infty)$ code. (The forbidden patterns are 11 and 101.) The minimum codeword length of a *fixed-length* rate-$\frac{1}{2}$ code for this constraint is 14 [e.g., 13]. The constraint can be presented by a three-state graph G with states named 0, 1, and 2, (see Fig. 7). Here, sequences ending in state 0 or 1 have precisely 0 or 1 terminal zeros, respectively, and a sequence ending in state 2 has at least 2 terminal zeros. The second-power graph G^2 of G needed in the construction of a rate $1 \rightarrow 2$ code is depicted in Fig. 8.

Let the set of principal states consist of the single state 2. Now consider the possible encoding paths, that is, paths in G^2 starting and ending in state 2. By inspection of G^2, we obtain the three paths

$$2 \xrightarrow{00} 2, \quad 2 \xrightarrow{10} 1 \xrightarrow{00} 2, \quad 2 \xrightarrow{01} 0 \xrightarrow{00} 2$$

from which a code can now be designed. The encoding rules are specified as

$$0 \rightarrow 00, \quad 10 \rightarrow 10.00, \quad 11 \rightarrow 01.00$$

The code can be decoded with a sliding-block decoder that has a window size of two codewords.

In general, for this method to succeed we need a (finite) collection $\mathcal{P}$ of paths π in G^q between principal states such that in each principal state s, we have that

$$\sum 2^{-|\pi|} \geq 1 \tag{13}$$

where the sum is over all paths π starting in s (and ending in the same or another principal state) and where $|\pi|$ denotes the *length* (number of transitions in G^q) of π. This condition, the Kraft–McMillan inequality for prefix codes [e.g., 26] is necessary and sufficient for the existence of a

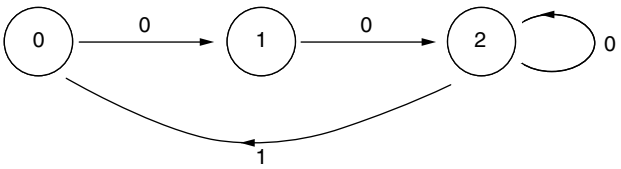


Figure 7. Presentation of the $(2, \infty)$ constraint.

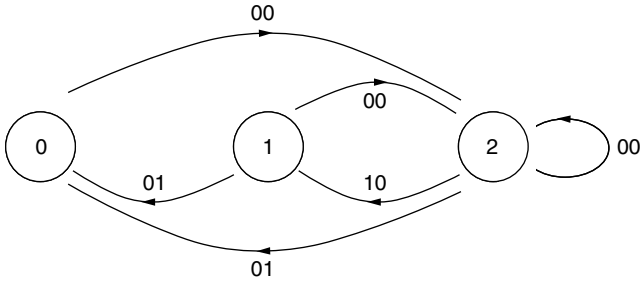


Figure 8. Presentation of G^2.

prefix code (such as the source words 0, 10, and 11 in the preceding example) with wordlengths equal to the path lengths. A similar but more complicated example of the above method is the rate-$\frac{1}{2}(2, 7)$ code [27].

Codes for constraints of finite type found by this method are always sliding-block-decodable. Such codes are called *variable-length principal-state codes* or (synchronous) *variable-length codes*. A special case of this method is the *substitution method*. Here, we set up some simple encoding rules, and then try to remove violations of the constraints by suitable substitutions. For example, consider again the design of a rate-$\frac{1}{2}(2, \infty)$ code. Suppose that we try to use the simple encoding rules $0 \to 00$ and $1 \to 10$. This works fine except when the source sequence contains two consecutive 1s, which produces 1010. These violations can be removed by replacing, from left to right, each occurrence of 1010 in the encoded stream by 0100. Note that we have obtained the same code as the one constructed earlier. A similar but more complicated example is the rate-$\frac{2}{3}(1, 7)$ code ("Jacoby code") [28].

Franaszek [23,24] and Lempel and Cohn [29] also introduced the class of *bounded-delay encodable codes* and the *bounded-delay method* for constructing such codes. Bounded-delay encodable codes involve a finite-state encoder where the encoding of the current source symbol may depend on the present state, on a bounded number of upcoming source symbols (so *lookahead* may be employed), and on a bounded number of previous states (the *history*) in the encoding path. The bounded-delay construction method for such codes employs an approximate eigenvector in an essential way. The idea is to construct in each state a suitable collection of *independent path sets*, each consisting of a set of encoding paths for this state, by exhaustive search up to a fixed maximum pathlength. Although this is a powerful construction method (in fact, much later it was shown [30] to be as powerful as the ACH method to be mentioned next, see also Refs. 31–33, and Ref. 34, Chaps. 4 and 5), there is no easy way to turn this method into an algorithm.

A technique called *state combination* [35] uses an approximate eigenvector with all components equal to 0, 1, or 2 to construct block-decodable codes that can be encoded by employing one-symbol lookahead, a special case of bounded-delay encodable codes. State combination is especially suited to the construction of (d, k) RLL codes. The method has been further developed and extended [32,36].

A breakthrough in code construction was achieved by the discovery of the state-splitting method or *ACH algorithm* [3]. This method employs an approximate eigenvector to construct an encoder, and does so in a number of steps bounded by the sum of the components of the approximate eigenvector. It can be used to prove the following theorem.

Theorem 1. For any given constraint of finite type and for any given rate $p \to q$ for which $p/q \leq C$, the capacity of the constraint, there exists a sliding-block decodable code for that constraint, with a synchronous finite-state encoder that encodes binary data at a constant rate $p \to q$.

Starting point in this method is a pair (H, ϕ), where $H = G^q$ is the qth power of an (irreducible) graph G presenting our constraint and ϕ is a $[H, 2^p]$-approximate eigenvector. (Here we may assume without loss of generality that H is irreducible and that $\phi > 0$.) The algorithm repeatedly transforms the current pair (H, ϕ) into another such pair (H', ϕ') by an operation called *weighted state splitting*. (This operation is called *ϕ-consistent state splitting* in Ref. 37, to which we refer for an excellent overview of the method.) This transformation is guaranteed to succeed except when all nonzero weights are equal. It has the property that the new weights are in some sense "smaller" than the original weights.

So with each transformation the approximate eigenvector gets "smaller" until finally a pair (G, ϕ) is reached where all nonzero components of ϕ are equal. Then the approximate eigenvector inequalities for ϕ show that in each state with a nonzero weight there are at least 2^p transitions leading to other such states, that is, we have obtained an encoder for our constraint.

State splitting, on which the transformation is based, can be understood intuitively as follows. While generating a sequence in the graph, each state encountered stands for a collection of future opportunities (represented by the arcs leaving the state) from which we may choose one. Now subdivide or *split* the state, this collection of opportunities, into *nonempty* parts called *substates* (to each of which we assign the corresponding part of the original arcs). Before this splitting operation, we could move from another state to this state without worrying about which opportunity (which arc) to utilize later, but now we have to choose first, before moving, from *which* of the parts we wish to pick our next opportunity. We do not lose opportunities, but we have to choose earlier.

So if s is a state and A_s is the collection of arcs leaving this state, then to split this state, we proceed as follows. First we partition the set A_s into *nonempty* parts $A_{s_1}, \ldots, A_{s_r}$. Then, in the graph G we replace the state s by states $s_1, \ldots, s_r$ (the substates of s), we replace each arc α in part A_{s_i}, $i = 1, \ldots, r$, by an arc from s_i with the same label and terminal state as α, and then we replace each arc β ending in s by r arcs $\beta_1, \ldots, \beta_r$, with the same label and initial state as β, with β_i ending in s_i, $i = 1, \ldots, r$.

It should be evident from the preceding discussion that the new graph obtained by splitting a state presents the same sequences as the original graph. Moreover, it is not difficult to see that if the original graph is of finite type, then the new graph is again of finite type, with the same memory and with an anticipation that has increased by at most one. Note that if the final graph has memory m and anticipation a, then the encoder obtained from this graph will have a decoding window of size at most $m + 1 + a$.

Weighted state splitting is another type of simple state splitting where we also distribute the weight of the state that is split over the substates (where each substate should receive a *nonzero* amount), and in such a way that the resulting weights constitute an approximate eigenvector for the new graph. A state that allows weighted state splitting can always be found among the states with maximum weight [3], provided not all nonzero weights are equal, and an algorithm is given to find such a state.

Example 1. Consider again the presentation $H = G^2$ in Fig. 8. Take $\phi = (1, 1, 2)$. Obviously, ϕ is a $[G^2, 2]$-approximate eigenvector. We shall split state 2 into two states, 2_1 and 2_2, according to the partition $A_{2_1} = \{2 \xrightarrow{00} 2\}, A_{2_2} = \{2 \xrightarrow{01} 0, 2 \xrightarrow{10} 1\}$. We also distribute the weight of state 2 over the two descendents of state 2 by assigning weight 1 to both 2_1 and 2_2. Since state 2_1 has successor state 2 (i.e., both 2_1 and 2_2) of total weight 2, and since state 2_2 has as successors the states 0 and 1, also of total weight 2, the new weights again form an approximate eigenvector, so the state split is ϕ-consistent. The new approximate eigenvector $\phi' = (1, 1, 1, 1)$ is constant, hence the resulting graph (depicted in Fig. 9) includes an encoder graph. By choosing a suitable assignment of 1-bit source symbols to the arcs (also shown in Fig. 9), we (essentially) obtain the rate-$\frac{1}{2}(2, \infty)$ code constructed earlier.

Ashley and Marcus [31], have shown that the ACH algorithm is *universal* in the sense that, given a sliding-block decoder, a matching finite-state encoder can be obtained by weighted state splitting (see also Refs. 32 and 33 and Ref. 34, Chaps. 4 and 5). However, precisely for that reason, the algorithm offers a large amount of freedom, first in the choice of the approximate eigenvector, then both in the choice of the states to be split and in the actual splits themselves, and it is not clear how to choose in order to obtain a good code.

It is usually best to choose the *smallest possible* approximate eigenvector. This works well in practice; however, pathological cases are known where the best possible code can be produced only by arbitrarily large approximate eigenvectors [20].

Also, some heuristics have been developed to guide the state-splitting process [e.g., 37] in order to minimize the number of encoder states of the resulting encoder. However, in practical applications we are often especially interested in codes with a small *decoding window*. This problem is addressed in Ref. 32, where weighted state splitting is combined with ideas from the bounded-delay method of Franaszek to obtain heuristics to guide the choices of which states to split. For completeness, we mention that Theorem 1 holds even for constraints of almost finite type [14].

Which code construction method to use strongly depends on the structure of the power graph G^q presenting the constraint at the desired rate, and in particular on the maximum number of arcs leaving a state. If this number is relatively *small*, then use an ACH-type method. (An attractive choice is the variant described by Hollmann [32].) On the other hand, if this number is relatively *large*, then such methods are less suitable because of the increasing number of choices for the state-splitting steps. In that case, use a method based on merging bits, or a method for (almost) block-decodable codes such as described in Ref. 35 or 33, or any other more heuristic method that does the job.

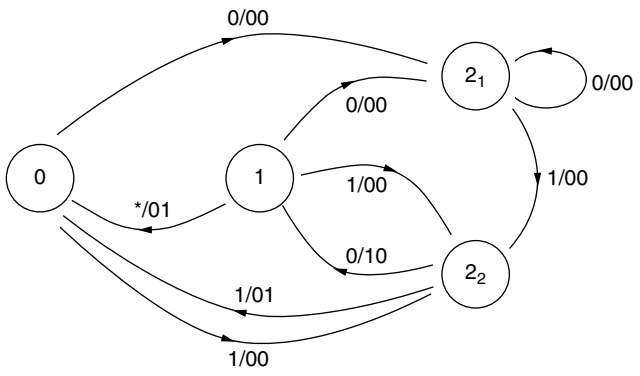


Figure 9. Presentation after state splitting.

Extremely large values of p and q are encountered in the design of high-rate codes, or when the efficiency of the code should be very high. In such cases, methods as *guided scrambling* [38], or a promising variant of enumerative encoding (see Refs. 39 and 40, and Ref. 34, Chap. 1) should be considered. Here, the extremely large block-length imposes a special system architecture to limit error propagation. For further details, we refer to the paper by Immink [40].

6. SELECTION OF RELATED TRENDS AND TOPICS

6.1. DC Control in RLL Codes for Optical Recording

Run-length-limited (RLL) codes have been successfully applied in optical storage, as witnessed by the well-known examples of EFM [41] for CD and EFMPlus [42] for DVD. The EFM and EFMPlus codes both employ a $(2, 10)$ constraint. EFM stands for *8-to-14 modulation*: the EFM code uses a single code table that maps a byte to a 14-bit channel word. Successive channel words are concatenated via the insertion of three merging bits, so the EFM code has rate $\frac{8}{17}$. The EFMPlus code maps a byte to a 16-bit channel word. EFMPlus is an ACH-type sliding-block code (see Section 5) based on a 4-state encoder graph. Both EFM and EFMPlus are byte-oriented, which is desirable for formats with the error correction coding (ECC) based on 8-bit symbols. An overview of the EFM and EFMPlus codes is given by Immink [43]. More recently, the *Blu-Ray Disc* (sic) or BD (formerly known as the DVR system [44]), which is the third generation of optical recording after CD and DVD, employs a code called "17PP" that uses a $(1, 7)$ constraint.

All RLL codes used in optical recording are *DC-free*; that is, they have almost no content at low frequencies. This property is an example of a *frequency-domain constraint*. Here, restrictions are enforced on the energy content per time unit of the sequence at certain frequencies, that is, on the *power spectral density function* of the sequence. (Constraints like run-length constraints are called *time-domain* constraints.) Most of these constraints belong to the family of *spectral null constraints*, where the power density function of the sequence must have a zero of a certain order at certain specific frequencies. The constraint that specifies a zero at DC, the zero frequency, is referred to as the *DC-free* constraint. We shall represent the NRZI channel bits by the bipolar values ± 1. A sequence $x_1, x_2, \ldots$ is called *DC-free* if its *running digital sum* (RDS)

$$\mathrm{RDS}_i = x_1 + \cdots + x_i$$

takes on only finitely many different values. In that case, the power spectral density function vanishes at DC.

One common way to ensure this is to constrain the code sequence to be *N-balanced*; we allow only sequences whose RDS takes on values between $-N$ and N. Note that such a constraint cannot be specified in terms of a finite collection of forbidden patterns; that is, it is not of finite type.

The DC-free property is needed in optical recording for a number of reasons: (1) it is necessary to separate the data signal from low-frequency disk noise such as fingerprints, dust, or defects; (2) DC-free coding is needed for control of the slicer level in the case of nonlinearities in the physical signals like pit–land asymmetries [45]; and (3) servo systems used for tracking of the laser spot position typically require a DC-free data signal.

We shall now discuss a general method to achieve DC control in RLL sequences. As discussed above, DC control is performed via control of the running digital sum (RDS). A very useful concept herein is the *parity*, the number of ones modulo 2, of a sequence of bits. Recall that an NRZ 1 bit indicates the start of a new run in the (bipolar) NRZI bit stream. Hence, because of the 1T precoder between NRZ and NRZI channel bit streams (see Section 1), each 1 bit in the NRZ bit stream changes the polarity in the corresponding NRZI bit stream. Consequently, an *odd* number of ones in a segment of the NRZ bit stream *reverses* the NRZI polarity after that segment while an *even* number of ones leaves the polarity unchanged.

This observation can be used for DC control as follows (see also Fig. 10). Suppose that for a certain segment of the NRZ bit stream, we can choose between two candidate sequences, one with parity 0 and the other with parity 1. Then the part of the NRZI bit stream *after* this segment will have a contribution to the RDS where the *sign* but not the magnitude depends on which of the two sequences is chosen. The *best* choice is, of course, the one that keeps the value of the RDS as close to zero as possible. For obvious reasons, we shall refer to these segments as *DC-control segments*.

In order to realize DC control, we have to insert DC-control segments at regular positions in the bit stream. Such positions are referred to as *DC-control points*. This is the basic mechanism for DC control in RLL codes.

Two types of DC-control segments can be distinguished. One type just serves the primary purpose of parity selection; another type additionally encodes some data bits. Further, we can differentiate between two types of DC-control: one type providing *guaranteed* DC control where parity selection is possible at each DC-control point, and another type providing *stochastic* DC-control where the possibility of parity selection depends on the data that is to be encoded.

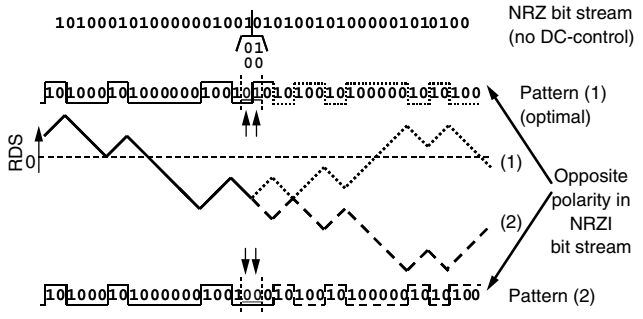


Figure 10. Principle of DC control via insertion of merging bits in a valid $(1, \infty)$ bit stream.

We shall now review several practical design methods for DC control that use the above mechanism in one form or another.

1. Insertion of a DC-control segment of N channel bits in a NRZ bit stream already satisfying a (d, k) constraint. At each DC-control point, the original bit stream is cut open, and the segment is inserted. If we require that both parities (0 and 1) can be selected without violation of the (d, k) constraint, then the minimum possible value of N is $N = 2(d + 1)$ if $2d + 1 \leq k < \infty$ and $N = d + 1$ if $k = \infty$. This can be seen as follows. At a DC-control point, the situation looks like $\ldots 10^i \cdot 0^{r-i}1\ldots$, for some i and r with $0 \leq i \leq r$ and $d \leq r \leq k$. First, let $k = \infty$. A possible segment of odd parity is $0^{d-i}10^i$ if $i \leq d$ or 10^d if $i \geq d$; if $i = 0$, no shorter segment is possible. Of course, the segment 0^{d+1}, of even parity, can always be inserted. Next, suppose that $2d + 1 \leq k < \infty$. Now, a possible segment of even parity is $0^{d-i}10^d10^i$ if $i \leq d$ or 10^d10^d if $i \geq d$; again, if $i = 0$, no shorter segment is possible. Also, a possible segment of odd parity is $0^{2d+1-i}10^i$ if $i \leq 2d + 1$ or 10^{2d+1} if $i \geq 2d + 1$. The case where $k = \infty$ and $d = 1$, which requires only two merging bits, is illustrated in Fig. 10. Here, following the 1 bit, two segments 01 and 00, of opposite parity, can be merged into the sequence; for both choices, the respective RDS traces are drawn. Since the RDS for the first pattern remains close to zero, whereas the RDS trace of the second pattern drifts away, obviously the first pattern is the best choice.

2. Insertion of merging bits between successive NRZ channel words, as is used in the EFM code. The function of the merging bits in EFM is twofold; they (a) serve to prevent violations of the $(2, 10)$ run-length constraint and (b) provide a means for DC control via the available freedom in the choice of the merging bits. Note that whether parity selection is possible depends on the two EFM channel words at hand; that is, the DC control of the EFM code is of a stochastic nature. For example, when the previous EFM word ends in a 1 and the current EFM word starts with 1 or 01, then the only valid merging bit pattern (i.e., not violating the $d = 2$ constraint) is 000, so that no DC control is possible.

3. Use of a substitution table for DC control. Here a code is used where certain source words can be encoded into two NRZ channel words (a *standard* word and a *substitute* word) of opposite parity. This mechanism is used, for example, in the EFMPlus code. Here, code construction via the ACH algorithm yields a 4-state encoder graph. In each state, the main encoding table contains 256 entries; because of the presence of *surplus* words, 88 of these entries have a substitution entry. (Additional DC control in EFMPlus is achievable via occasional swapping of encoder state [43].) Note that the DC-control mechanism via substitution tables (as in EFMPlus) is of a stochastic nature; whether a byte can be used as a DC-control point depends on whether the encoding table has a substitute entry at that location.

4. DC control via the use of a *parity-preserving* code [46]. Such a code preserves the parity on RLL

encoding; that is, the parity of a source word is identical to the parity of the corresponding channel word. The major difference with the previous methods is that single DC-control bits are inserted in the *source* bit stream. Changing a DC-control bit from 0 to 1 changes the parity in the source bit stream and hence also in the NRZ channel bit stream; this property enables the selection of the polarity of the NRZI channel bit stream, and thus allows for DC control.

5. DC control via the use of *combicodes* [47]. This is a typical example of a method where some data bits are encoded in the DC-control segment. A combicode for a given constraint consists of a set of at least two codes for that constraint, possibly with different rates, where the encoders of the various codes share a *common* set of encoder states. As a consequence, after each encoding step the encoder of the current code may be replaced by the encoder of any other code in the set, where the new encoder has to start in the ending state of the current encoder. Typically, one of the codes, called the *standard* code or *main* code, is an efficient code for standard use; the other codes serve to realize certain additional properties of the channel bit stream. Sets of sliding-block decodable codes for a combicode can be constructed via the ACH algorithm; here the codes are jointly constructed starting with suitable presentations derived from the basic presentation for the constraint and using the *same* approximate eigenvector [47].

Coene has used the combicode idea [47] for DC-control as follows. The *main* code has a single channel word for each source word. As a second code, of a lower rate, a *substitution code* is used. This code has for *every* source word a set of two possible channel words, of opposite parity and ending in the same encoder state.

In a practical format, the sequence of uses of the various codes of the combicode needs to be chosen beforehand. This choice will be the result of a tradeoff between the overall rate and the required property (e.g., the amount of DC control) to be realized. Obviously, DC control by use of these combicodes is of the *guaranteed* type.

6.2. Time-Varying Codes

The concept of time-varying codes [48] is related to a generalization of the state-splitting (ACH) algorithm. In this generalization, the starting point is a *periodic* presentation of the constraint where the codeword labels may have different lengths in different phases. Thus, the set V of vertices can be partitioned into subsets $V_0, V_1, \ldots, V_{s-1}$, say, with the property that an arc starting in some part V_i ends in V_{i+1} (or in V_0 if $i = s - 1$); moreover, labels on arcs that start in the same part have the same length. An encoder derived from such a periodic presentation operates cyclically in s phases, where the length of the codeword produced by the encoder depends only on the current phase.

The framework of time-varying codes is useful for the design of efficient codes with reduced error propagation. Note that a time-varying code may be considered as a pair of codes where the order in which the codes have to be used is fixed beforehand; this in contrast with combicodes, where the order in which the various codes can be used is completely free. The concept of multiple phases in a time-varying code [48] is strongly related to the idea of representing a bit by a number of (virtual) *fractional* bits, as used by Coene [47] to generate a highly efficient DC-free combicode for $d = 1$.

6.3. RLL Parity-Check Coding

In the standard digital storage system as sketched in Fig. 1, random channel errors that may occur in the channel bit stream after bit detection are repaired by the ECC error correction decoder situated after the MC decoder. In the early 1990s, it was realized that a *combination* of error correction coding and modulation coding may be quite advantageous in terms of overall efficiency and performance [e.g., 49–52]. RLL parity-check coding focuses on the most prominent error patterns that are left by the bit detector in the receiver. For magnetic recording, the most simple parity-check coding schemes aim at peak shift errors in the NRZ bit stream, where the 1 bits are shifted (to the left or right). For optical recording, where the bit detector regenerates the NRZI bit stream, the most prominent random errors are transition shifts that cause the runs at the left and right sides of the transition to become one or more bits longer or shorter. Because of the 1T precoder, the transition error in the NRZI bit stream corresponds to a peak shift error in the NRZ bit stream. In parity-check coding, the channel bits are partitioned into fixed-length *segments*, and a parity check is generated for each segment. (Subsequently, these parity checks can be handled in various ways.) As a consequence, inspection of the NRZ bit stream after detection will reveal the violation of the parity check in the case that a channel error occurs, thus enabling error detection; for error correction, some side information from the signal waveform is needed. One scheme [52] considers merging of parity blocks into the original NRZ bit stream in such a way that these blocks become an integral part of the corresponding parity-check segment. Another scheme [53] is concatenated parity-check coding, where parity bits that are generated on the NRZ bit stream are separately RLL-encoded. The merging scheme has a low coding efficiency, but the advantage is its simplicity and lack of error propagation. The concatenated scheme has a high efficiency but suffers from error propagation. A promising route is to use a parity-check coding scheme based on a combination of codes [54], like the *combicodes* used for DC control. Apart from a *main* RLL code, a second RLL code, the *parity-check-enabling code*, is required. The latter code is used to set the parity-check constraint of a segment of NRZ bits to a predetermined value. For DC control together with parity-check control, a third code is needed: the *substitution* code. All three codes are jointly constructed, so that the channel words of these codes can be freely concatenated.

6.4. MTR Constraint for Magnetic Recording

The *maximum transition run* (MTR) constraint [55] specifies the maximum number of consecutive 1 bits in the NRZ bit stream. Equivalently, in the NRZI bit stream, the MTR constraint limits the number of successive 1T runs. The MTR constraint can also be combined with a d constraint, in which case the MTR constraint limits the

number of consecutive *minimum runlengths*. An application for $d = 1$ can be found in the paper by Narahara et al. [44]. The basic idea behind the use of MTR codes is to eliminate the *dominant* error patterns, that is, those patterns that would cause most of the errors in the *partial-response maximum-likelihood* (PRML) sequence detectors used for high-density recording. A highly efficient rate $16 \rightarrow 17$ MTR code limiting the number of consecutive transitions to at most two has been described [56].

6.5. (0, G/I) Constraint for Magnetic Recording

In magnetic hard-disk drives using the PR4 or *class 4* partial-response maximum-likelihood (PRML) sequence detectors [57] with response equal to $1 - D^2$, it is advantageous to use modulation coding with so-called $(0, G/I)$ constraints [58,59]. The PR4 detector can be partitioned into two $1 - D$ sequence detectors, where each detector operates at half the bit rate on either the even- or odd-indexed bits. The 0 in $(0, G/I)$ stands for $d = 0$, and the G and the I stand for the *global* and *interleaved* constraint on the maximum number of consecutive zeros in the *joint* and in both *interleaved* bit streams. The global constraint is just a k constraint and is meant for the purpose of timing recovery; the interleaved constraint is introduced in order to limit the effects of truncation of the path memory depths in the separate Viterbi detectors for each of the $1 - D$ responses [7, Chaps. 6 and 7].

6.6. Two-Dimensional Constraints

Two-dimensional (2D) modulation codes have received considerable attention. This research effort is based on the recognition that 2D coding might lead to significant coding gains, especially in technologies such as holographic data storage [60] with a 2D page-based readout mechanism. Just like their 1D counterparts (as explained in Section 2), one of the aims of 2D modulation codes is to combat intersymbol interference (ISI), which is achieved by forbidding certain patterns that lead to high spatial frequencies. 2D lowpass filtering constraints are typically defined in terms of restrictions on neighbouring bits for 2D bit arrays on a rectangular lattice; some examples can be found in Ref. 61 (where they are called "checkerboard codes") and in Ref. 62. Another class of 2D codes are the *multitrack codes* [63], designed for systems where multiple tracks are encoded, written and read in parallel; the d constraint is maintained for each track independently as in the 1D case (in view of ISI), but the k constraint used for timing recovery (or clocking) is defined *jointly* for a set of tracks, since synchronisation is also carried out jointly. Other 2D constraints with potential practical interest are 2D (d, k) constraints, where the 1D (d, k) constraint has to be satisfied in both horizontal and vertical directions. In general, unlike the case for 1D RLL coding, only bounds can be derived for the capacity of 2D (d, k) constraints. This area has been the subject of intensive research (see Refs. 64 and 65 and references cited therein). A practical proposal to use 2D $d = 1$ modulation coding in optical recording can be found in Ref. 66.

Concerning the construction of 2D codes, 2D bit arrays can be encoded in one-dimensional strips containing a number of bit rows (or tracks) as proposed in Refs. 61 and 63. In this way, 2D code construction can be reduced to a 1D problem for which code construction methods such as the ACH sliding-block codes discussed in Section 5 can be used.

Acknowledgments

With great pleasure we thank our colleagues A. H. J. Immink, L. M. G. Tolhuizen, and J.H. van Lint for their many helpful comments made during the preparation of this text.

BIOGRAPHIES

Wim Coene received a Ph.D. in physics from the Center for High-Voltage Electron Microscopy (University of Antwerp) for work on computational modeling of electron diffraction and image formation in a transmission electron microscope (TEM). In 1988 he joined Philips Research Laboratories, where he first worked on signal processing for ultra-high-resolution TEM, in particular on phase retrieval methods used for digital correction of aberration artifacts. In 1996, after one year of work on MPEG-2 video coding, he started to work on signal processing for optical storage in the field of bit detection algorithms and channel modulation codes and techniques. Currently, his research activities are focused on coding and signal processing for next-generation (optical) storage channels.

Henk Hollmann was born in Utrecht, the Netherlands, on March 10, 1954. He received the master's degree in mathematics in 1982, with a thesis on association schemes, and the Ph.D. degree in 1996, with a thesis on modulation codes, both from Eindhoven University of Technology. In 1997, he was awarded the SNS bank prize for the best Ph.D. thesis in fundamental research of this university. In 1982 he joined CNET, Issy-les-Moulineaux, France, where he worked mainly on number-theoretic transforms. Since 1985 he has been with Philips Research Laboratories, Eindhoven, the Netherlands. His research interests include discrete mathematics and combinatorics, information theory, cryptography, and digital signal processing.

BIBLIOGRAPHY

1. P. A. Franaszek, Sequence-state coding for digital transmission, *Bell Syst. Tech. J.* **47**: 143–157 (1968).
2. D. T. Tang and L. R. Bahl, Block codes for a class of constrained noiseless channels, *Inform. Control* **17**: 436–461 (1970).
3. R. L. Adler, D. Coppersmith, and M. Hassner, Algorithms for sliding block codes. An application of symbolic dynamics to information theory, *IEEE Trans. Inform. Theory* **IT-29**(1): 5–22 (Jan. 1983).
4. K. A. S. Immink, *Codes for Mass Data Storage Systems*, Shannon Foundation Publishers, The Netherlands, 1999.
5. B. H. Marcus, R. M. Roth, and P. H. Siegel, Constrained systems and coding for recording channels, in V. S. Pless and W. C. Huffman, eds., *Handbook of Coding Theory II*, Elsevier, Amsterdam, 1998, 1635–1764.

6. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorders, *IEEE Trans. Inform. Theory* (Special Commemorative Issue) 2260–2299 (1998).
7. J. W. M. Bergmans, *Digital Baseband Transmission and Recording*, Kluwer, Amsterdam, 1996.
8. C. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (1948).
9. M. Fekete, Uber die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit Ganzzahligen Koeffizienten, *Math. Zeitschr.* **17**: 228–249 (1923).
10. D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press, Cambridge, MA, 1995.
11. H. Minc, *Nonnegative Matrices*, Wiley, New York, 1988.
12. E. Seneta, *Non-negative Matrices and Markov Chains*, (2nd ed.,) Springer, New York, 1981.
13. K. A. S. Immink, *Coding Techniques for Digital Recorders*, Prentice-Hall, Englewood Cliffs, NJ, 1991.
14. R. Karabed and B. H. Marcus, Sliding-block coding for input-restricted channels, *IEEE Trans. Inform. Theory* **IT-34**(1): 2–26 (1988).
15. J. E. Hopcroft and J. D. Ullmann, *Formal Languages and Their Relation to Automata*, Addison-Wesley, Reading, MA, 1969.
16. J. E. Hopcroft and J. D. Ullmann, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA, 1979.
17. G. F. M. Beenker and K. A. S. Immink, A generalized method for encoding and decoding runlength-limited binary sequences, *IEEE Trans. Inform. Theory* **IT-29**(5): 751–754 (1983).
18. J. H. Weber and K. A. Abdel-Ghaffar, Cascading runlength-limited sequences, *IEEE Trans. Inform. Theory* **IT-39**(6): 1976–1984 (1993).
19. K. A. S. Immink, Constructions of almost block-decodable runlength-limited codes, *IEEE Trans. Inform. Theory* **IT-41**(1): 284–287 (Jan. 1995).
20. H. D. L. Hollmann, On an approximate eigenvector associated with a modulation code, *IEEE Trans. Inform. Theory* **IT-43**(5): 1672–1678 (1997).
21. P. A. Franaszek, A general method for channel coding, *IBM J. Res. Dev.* **24**: 638–641 (1980).
22. P. A. Franaszek, On future-dependent block coding for input-restricted channels, *IBM J. Res. Dev.* **23**: 75–81 (1979).
23. P. A. Franaszek, Synchronous bounded delay coding for input restricted channels, *IBM J. Res. Dev.* **24**: 43–48 (1980).
24. P. A. Franaszek, Construction of bounded delay codes for discrete noiseless channels, *IBM J. Res. Dev.* **26**: 506–514 (1982).
25. T. M. Cover, Enumerative source coding, *IEEE Trans. Inform. Theory* **IT-19**: 73–77 (1973).
26. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
27. J. S. Eggenberger and P. Hodges, Sequential encoding and decoding of variable word length, fixed rate data codes, U.S. Patent 4,115,768, (1978).
28. G. Jacoby and R. Kost, Binary two-thirds rate code with full word look-ahead, *IEEE Trans. Magn.* **MAG-20**(5): 709–714 (1984).
29. A. Lempel and M. Cohn, Lookahead coding for input-restricted channels, *IEEE Trans. Inform. Theory* **IT-28**: 933–937 (1982).
30. P. A. Franaszek, Coding for constrained channels: A comparison of two approaches, *IBM J. Res. Dev.* **33**: 602–608 (1989).
31. J. Ashley and B. H. Marcus, Canonical encoders for sliding block decoders, *Siam J. Disc. Math.* **8**(4): 555–605 (1995).
32. H. D. L. Hollmann, On the construction of bounded-delay encodable codes for constrained systems, *IEEE Trans. Inform. Theory* **IT-41**(5): 1354–1378 (1995).
33. H. D. L. Hollmann, Bounded-delay encodable, block-decodable codes for constrained systems, *IEEE Trans. Inform. Theory* Special Issue on Codes and Complexity **IT-42**(6): 1957–1970 (1996).
34. H. D. L. Hollmann, *Modulation Codes*, doctoral thesis, Eindhoven Univ. Technology, Eindhoven, The Netherlands, 1996.
35. K. A. S. Immink, Block-decodable runlength-limited codes via look-ahead technique, *Philips J. Res.* **46**(6): 293–310 (1992).
36. H. D. L. Hollmann, Bounded-delay encodable codes for constrained systems from state combination and state splitting, *Proc. 14th Benelux Symp. Information Theory*, Veldhoven, 1993, pp. 80–87.
37. B. H. Marcus, P. H. Siegel, and J. K. Wolf, Finite-state modulation codes for data storage, *IEEE J. Select. Areas Commun.* **10**(1): 5–37 (1992).
38. I. J. Fair, W. D. Gover, W. A. Krzymien, and R. I. Macdonald, Guided scrambling: A new line coding technique for high bit rate fiber optic transmission systems, *IEEE Trans. Commun.* **COM-39**(2): 289–297 (1991).
39. L. Pátrovics and K. A. S. Immink, Encoding of *dklr*-sequences using one weight set, *IEEE Trans. Inform. Theory* **IT-42**(5): 1553–1554 (1996).
40. K. A. S. Immink, A practical method for approaching the channel capacity of constrained channels, *IEEE Trans. Inform. Theory* **IT-43**(5): 1389–1399 (1997).
41. K. A. S. Immink and H. Ogawa, Method for encoding binary data, U.S. Patent 4,501,000 (1985).
42. K. A. S. Immink, EFMPlus: The coding format of the multimedia compact disc, *IEEE Trans. Consum. Electron.* **41**(3): 491–497 (1995).
43. K. A. S. Immink, A survey of codes for optical disk recording, *IEEE J. Select. Areas Commun.* **19**: 756–764 (2001).
44. T. Narahara et al., Optical disc system for digital video recording, *Jpn. J. Appl. Phys.* **39**(2B)(Part 1): 912–919 (2000).
45. A. F. Stikvoort and J. A. C. van Rens, An all-digital bit detector for compact disc players, *IEEE J. Select. Areas Commun.* **10**(1): 191–200 (1992).
46. J. A. H. Kahlman and K. A. S. Immink, Device for encoding/decoding N-bit source words into corresponding M-bit channel words, and vice versa, U.S. Patent 5,477,222 (1995).
47. W. Coene, Combi-codes for DC-free runlength-limited coding, *IEEE Trans. Consum. Electron.* **46**(4): 1082–1087 (2000).

48. J. J. Ashley and B. H. Marcus, Time-varying encoders for constrained systems: an approach to limiting error propagation, *IEEE Trans. Inform. Theory* **IT-46**: 1038–1043 (2000).

49. H. M. Hilden, D. G. Howe, and E. J. Weldon, Shift error correcting modulation codes, *IEEE Trans. Magn.* **27**: 4600–4605 (1991).

50. Y. Saitoh, I. Ibe, and H. Imai, Peak-shift and bit error-correction with channel side information in runlength-limited sequences, *Proc. 10th Int. Symp. Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, (AAECC), 1993, Vol. 10, pp. 304–315.

51. A. V. Kuznetsov and A. J. H. Vinck, A coding scheme for single peak-shift correction in (d, k)-constrained channels, *IEEE Trans. Inform. Theory* **IT-39**: 1444–1450 (1993).

52. P. Perry, M.-C. Lin, and Z. Zhang, Runlength-limited codes for single error-detection with mixed type errors, *IEEE Trans. Inform. Theory* **IT-44**: 1588–1592 (1998).

53. S. Gopalaswamy and J. Bergmans, Modified target and concatenated coding for $d = 1$ constrained magnetic recording channels, *Proc. IEEE Int. Conf. Communications*, New Orleans, LA, 2000, pp. 89–93.

54. W. M. J. Coene, H. P. Pozidis, and J. W. M. Bergmans, Runlength limited parity-check coding for transition-shift errors in optical recording, *Proc. Global Telecommunications Conf. 2001*, GLOBECOM'01; *IEEE* **5**: 2982–2986 (2001).

55. J. Moon and B. Brickner, Maximum transition run codes for data storage systems, *IEEE Trans. Magn.* **32**(5): 3992–3994 (1996).

56. T. Nishiya et al., Turbo-EEPRML: An EEPRML channel with an error correcting post-processor designed for 16/17 rate quasi MTR code, *Proc. Globecom'98*, Sydney, 1998, pp. 2706–2711.

57. H. Kobayashi and D. T. Tang, Application of partial-response channel coding to magnetic recording systems, *IBM J. Res. Dev.* **14**: 368–375 (1970).

58. B. Marcus and P. Siegel, *Constrained Codes for PRML*, IBM Report RJ 4371, 1984.

59. P. H. Siegel and J. K. Wolf, Modulation and coding for information storage, *IEEE Commun. Mag.* **29**(12): 68–86 (1991).

60. J. Ashley et al., Holographic data storage, *IBM J. Res. Dev.* **44**(3): 341–368 (2000).

61. W. Weeks and R. E. Blahut, The capacity and coding gain of certain checkerboard codes, *IEEE Trans. Inform. Theory* **IT-44**(3): 1193–1203 (1998).

62. J. J. Ashley and B. M. Marcus, Two-dimensional low-pass filtering codes, *IEEE Trans. Commun.* **46**(6): 724–727 (1998).

63. M. W. Marcellin and H. J. Weber, Two-dimensional modulation codes, *IEEE J. Select. Areas Commun.* **10**(1): 254–266 (1992).

64. A. Kato and K. Zeger, On the capacity of two-dimensional runlength constrained channels, *IEEE Trans. Inform. Theory* **IT-45**(5): 1527–1540 (1999).

65. R. M. Roth, P. H. Siegel, and J. K. Wolf, Efficient coding schemes for the hard-square model, *IEEE Trans. Inform. Theory* **IT-47**(3): 1166–1176 (2001).

66. S. Taira et al., *Study of Recording Methods for Advanced Optical Disks*, Technical Report of IEICE, MR2001-117, 2002, pp. 57–64.

CONTINUOUS-PHASE-CODED MODULATION

CARL-ERIK W. SUNDBERG
iBiquity Digital Corp.
Warren, New Jersey

JOHN B. ANDERSON
Lund University
Lund, Sweden

1. INTRODUCTION

Transmission cost is often proportional to bandwidth and the radiofrequency spectrum is a limited resource. The motivation for searching for spectrally efficient modulation is thus clear. The available transmitter power is also limited for many applications. For satellite and land-mobile radio applications, modulation with a constant RF (radiofrequency) envelope is advantageous because transmitters use more efficient nonlinear amplifiers. The combination of these factors dictates a coded modulation system based on constant-amplitude sinusoids. In this article we outline the power and spectrum properties of a large class of such signals. This class, continuous-phase modulation (CPM), can be viewed as a generalization of minimum shift keying (MSK) containing such schemes as tamed frequency modulation (TFM) and Gaussian MSK (GMSK). We review the performance of the CPM class with optimum reception and show how the choice of system parameters affects its error performance and the signal bandwidth. We also show how these two trade off against each other. Transmitters and simplified receivers are also discussed.

CPM coding was the first coded modulation class to be extensively studied and marked the advent of a combined energy and bandwidth view in practical coding schemes. It had its origins in the invention of MSK [1] and a related scheme called *fast frequency shift keying* [2]. It gained impetus by several papers [3–5] that introduced continuous-phase frequency shift keying (CPFSK) and its optimal detection in the early 1970s. With Miyakawa et al. [6] and Anderson and Taylor [7] in the mid-1970s, more sophisticated codinglike ideas were introduced; the second paper introduced the modern concepts of distance calculation, trellis decoding, and simultaneous consideration of code energy and bandwidth. CPM reached its full flower as a coded modulation method in 1981 with the basic papers of Aulin, Sundberg and others [8–11].

MSK has attracted new study since the 1970s as well. We will outline methods to improve on MSK while maintaining a constant amplitude. By *improvement* is meant a narrower power spectrum, lower spectral sidelobes, cheaper implementation, better error probability, or all the above. The main part of the article considers a number of methods for constructing constant-amplitude signals that significantly outperform MSK in either energy, bandwidth, or both. An important issue is at what level of complexity these improvements are obtained.

2. THE CPM SIGNAL CLASS

A large class of constant-amplitude modulation schemes is defined by the signal

$$s(t) = \left[\frac{2E}{T}\right]^{1/2} \cos(2\pi f_0 t + \phi(t, \boldsymbol{a})) \tag{1}$$

where the transmitted information is contained in the phase

$$\phi(t, \boldsymbol{a}) = 2\pi h \sum_{i=-\infty}^{\infty} a_i q(t - iT) \tag{2}$$

with $q(t) = \int_{-\infty}^{t} g(\tau)\, d\tau$. Normally the function $g(t)$ is a smooth pulseshape over a finite time interval $0 \le t \le LT$ and zero outside. Thus the parameter L is the length of the pulse (per unit T) and T is the symbol time; E is the energy per symbol, f_0 is the carrier frequency and h is the modulation index. The M-ary data symbols a_i take values $\pm 1, \pm 3 \cdots \pm(M-1)$. M is normally selected to be a power of 2, and we will mainly consider binary, quaternary, and octal systems ($M = 2,4,8$). From the definition above we note that the pulse $g(t)$ defines an instantaneous frequency and its integral $q(t)$ is the phase response. The precise shape of $g(t)$ determines the smoothness of the transmitted information carrying phase. The rate of change of the phase (or instantaneous frequency) is proportional to the parameter h, which is the *modulation index*. The pulse $g(t)$ is normalized in such a way that $\int_{-\infty}^{\infty} g(t)\, dt$ is $\frac{1}{2}$. This means that for schemes with positive pulses of finite length, the maximum phase change over any symbol interval is $(M-1)h\pi$.

By choosing different pulses $g(t)$ and varying the parameters h and M, a great variety of CPM schemes can be obtained. Some of the more popular pulseshapes are listed in Table 1. These include CPFSK, tamed frequency modulation (TFM) [12], generalized TFM (GTFM) [13], Gaussian MSK (GMSK) [14], duobinary FSK (2REC) [15], raised cosine (LRC), and spectrally raised cosine (LSRC). In the table we use the notation LXX for a pulse of length L symbol intervals; thus 3RC is a raised cosine pulse of length $3T$. For spectral raised cosine (LSRC), the main time lobe has width LT. The pulse of length $1T$, namely, 1REC, is another name for CPFSK. The 2REC pulse is also called duobinary.

MSK is obtained as a special case of the signals defined in Eq. (1) by selecting the pulse 1REC ($L = 1$) from Table 1 and using binary ($M = 2$) data with $h = \frac{1}{2}$. The CPM signal can be viewed as phase modulation or as frequency modulation, but for understanding the optimum coherent receiver, it is advantageous to view the signal as phase modulation.

Memory is introduced into the CPM signal by means of its continuous phase. Each information-carrying phase function $\phi(t, \boldsymbol{a})$ is continuous at all times for all combinations of data symbols. Further memory can be built into the CPM signal by choosing a $g(t)$ pulse with $L > 1$. These schemes have overlapping pulse shaping and called *partial-response* techniques. CPM signals with $L \le 1$ are *full-response* schemes. In this case the memory is in the continuous phase only.

Although the CPM signals in Eq. (1) are in principle conceivable for any value of the modulation index h, a key to the design of practical maximum-likelihood detectors is to consider CPM schemes with rational values of h. For $h = 2k/p$ where k and p have no common factors, the phase

Table 1. Definition of the Frequency Pulse Function $g(t)$

LRC	$g(t) = \begin{cases} \dfrac{1}{2LT}\left[1 - \cos\left(\dfrac{2\pi t}{LT}\right)\right]; & 0 \le t \le LT \\ 0 \quad ; & \text{otherwise} \end{cases}$
	L is the pulse length, e.g., 3RC has $L = 3$
TFM[a]	$g(t) = \dfrac{1}{8}[Ag_0(t-T) + Bg_0(t) + Ag_0(t+T)]; \quad A = 1; B = 2$
	$g_0(t) \approx \sin\left(\dfrac{\pi t}{T}\right)\left[\dfrac{1}{\pi t} - \dfrac{2 - \dfrac{2\pi t}{T}\cot\left(\dfrac{\pi t}{T}\right) - \dfrac{\pi^2 t^2}{T^2}}{\dfrac{24\pi t^3}{T^2}}\right]$
LSRC	$g(t) = \dfrac{1}{LT} \cdot \dfrac{\sin\left(\dfrac{2\pi t}{LT}\right)}{\dfrac{2\pi t}{LT}} \cdot \dfrac{\cos\left(\beta \cdot \dfrac{2\pi t}{LT}\right)}{1 - \left(\dfrac{4\beta}{LT} \cdot t\right)^2}; \quad 0 \le \beta \le 1$
GMSK	$g(t) = \dfrac{1}{2T}\left[Q\left(2\pi B_b \dfrac{t - \dfrac{T}{2}}{\sqrt{\ell n 2}}\right) - Q\left(2\pi B_b \dfrac{t + \dfrac{T}{2}}{\sqrt{\ell n 2}}\right)\right]; \quad 0 \le B_b T < \infty$
	$Q(t) = \int_t^{\infty} \dfrac{1}{\sqrt{2\pi}} e^{-\tau^2/2} d\tau$
LREC	$g(t) = \begin{cases} \dfrac{1}{2LT}; & 0 \le t \le LT \\ 0 \quad ; & \text{otherwise} \end{cases}$
	$L = 1$ yields CPFSK

[a] The class of GTFM pulses is obtained by varying A, B, and $g_0(t)$.

$\phi(t, \boldsymbol{a})$ during interval $nT \le t \le (n+1)T$ can be written

$$\phi(t, \boldsymbol{a}) = 2\pi h \sum_{i=n-L+1}^{n} a_i q(t - iT) + \theta_n = \theta(t, \boldsymbol{a}) + \theta_n \qquad (3)$$

where $\theta_n = \left[h\pi \sum_{i=-\infty}^{n-L} a_i \right]$ modulo 2π has only p different values. Thus the total number of states that (at most) is needed to describe the signal in Eq. (1) is $S = pM^{(L-1)}$, where a state is defined as the vector $(\theta_n, a_{n-1}, a_{n-2}, \ldots, a_{n-L+1})$. The state vector consists of the phase state θ_n and $M^{(L-1)}$ *correlative states* for partial response systems. For a full-response system the number of states is p. The finite state description for CPM signals allows the use of a finite Viterbi decoder.

As an example of a CPM scheme, we choose binary 3RC; thus the pulse $g(t)$ is a raised-cosine pulse of length 3 symbol intervals. Figure 1a shows the pulse $g(t)$ and phase response $q(t)$ for 3RC. For comparison, the corresponding functions are also shown for 1REC (CPFSK). The information-carrying phase function $\phi(t, \boldsymbol{a})$ is illustrated both for 1REC and 3RC in Fig. 1b for a particular data sequence. Note that all changes in the 3RC phase take longer time than in the CPFSK scheme. Figure 1c shows all phase functions starting at the arbitrary phase 0° and

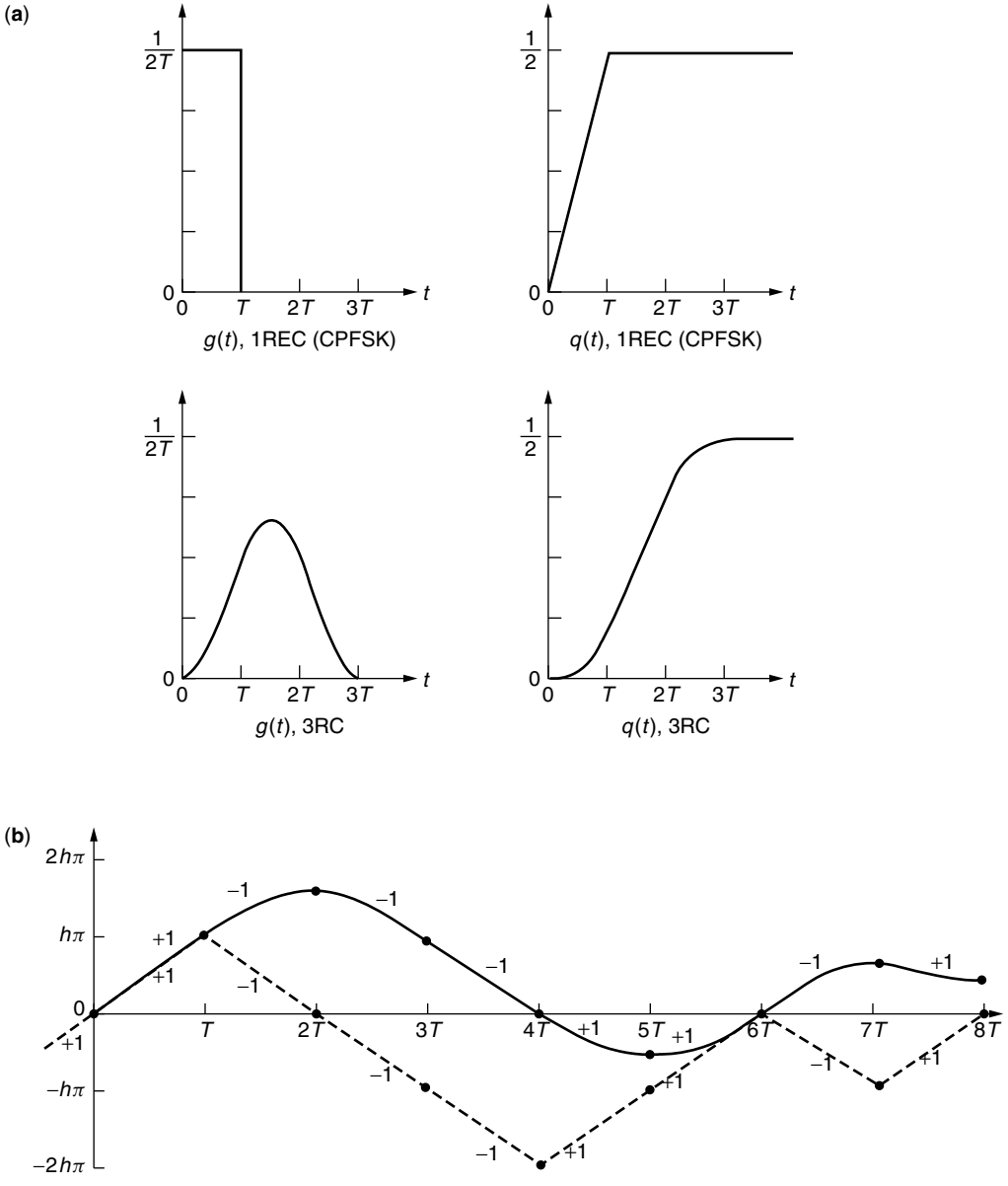


Figure 1. (**a**) Pulseshapes $g(t)$ and phase responses $q(t)$ for the full-response 1REC (CPFSK) and partial response 3RC CPM schemes; (**b**) examples of the phase function $\phi(t, \boldsymbol{a})$ for 1REC (dashed) and 3RC (solid) for the data sequence +1, −1, −1, −1, +1, +1, −1, +1; (**c**) phase tree for binary 3RC with $h = \frac{2}{3}$. The state description of the signal is also given. The transitions with arrows are also shown in Fig. 2a.

time 0, where the two previous data symbols are +1,+1. It is obvious from Fig. 1c that all phase changes are very smooth. The corresponding phase tree for MSK has linear phase changes with sharp corners when the data change. The phase tree in Fig. 1c also displays the state vector at each node.

The finite-state description of the CPM signal implies a trellis description, and this finiteness stems from the modulo-2π property of the phase. By plotting $I = \cos[\phi(t, \boldsymbol{\alpha})]$ and $Q = \sin[\phi(t, \boldsymbol{\alpha})]$ versus time in a three-dimensional plot, all signals appear on the surface of a cylinder. Such a phase cylinder is shown in Fig. 2a for the parameters used in Fig. 1c. This scheme has 12 states. Contrary to Fig. 1c, where restrictions were imposed for $t < 0$, the phase cylinder shows all signals over three symbol intervals. The phase nodes and some of the state vectors are also shown. To clarify the connection between the tree in Fig. 1c and the trellis in Fig. 2a, we have marked three identical transitions with arrows in both figures.

While the appearance of the phase tree does not change with h, the number of phase states does and so does the phase cylinder. Figure 2b shows the simplicity of the phase cylinder for $h = \frac{1}{2}$ and Fig. 2c shows how the complexity has grown for $h = \frac{3}{4}$. It is seen in Fig. 2b that with a proper phase offset, I and Q exhibit quite open-amplitude eye diagrams. We will see that this property leads to simple linear receivers and this is the reason for the popularity of the binary $h = \frac{1}{2}$ schemes.

An interesting generalization of the CPM class is to let the modulation index vary cyclically with time. This gives a so called multi-h scheme. These systems have better performance than the fixed h schemes. Typically most of the available improvement is obtained with two or three different h values.

3. CPM POWER SPECTRA

A large number of methods are available in the literature for calculating the power spectrum of CPM; for a detailed treatment, see, for instance, Ref. 16. Computer simulations can also be employed to estimate the power spectra. As a measure of signal bandwidth in what follows, we will generally take the frequency band around the carrier frequency containing 99% of the signal power. Figure 3 shows the power spectral density of some binary CPM schemes. The term *GMSK4* here means that the GMSK pulse in Table 1 is truncated symmetrically to 4 symbol intervals. The CPM schemes 3RC, GMSK4 with $B_bT = 0.25$, and 3SRC6 have comparable power spectra. The corresponding pulses $g(t)$ are also quite similar as are their detection properties — the rate of reduction of the spectral sidelobes is determined by the smoothness of the pulse $g(t)$. For most applications, the raised-cosine pulses probably have sufficient smoothness. Figure 3 also shows the MSK and 2REC (duobinary) schemes. Notice the lower spectral sidelobes of the smooth partial response schemes.

In comparison, Fig. 4 shows power spectra for some four-level schemes ($M = 4$). Figures 3 and 4 illustrate that a longer pulse $g(t)$ narrows the power spectra for fixed h and M. TFM has a power spectrum similar to binary 3.7 RC and 3.7 SRC. GMSK with $B_bT = 0.18$ corresponds approximately to 4RC and GMSK with $B_bT = 0.2$ to TFM.

The width of the main spectral lobe decreases with increasing L but increases with increasing h and M. The relationship between bandwidth or fractional out of band power and pulseshape $g(t)$ is rather complicated.

4. DETECTION AND ERROR PROBABILITY

Coherent maximum likelihood sequence detection (Viterbi detection) can be performed for all CPM schemes that

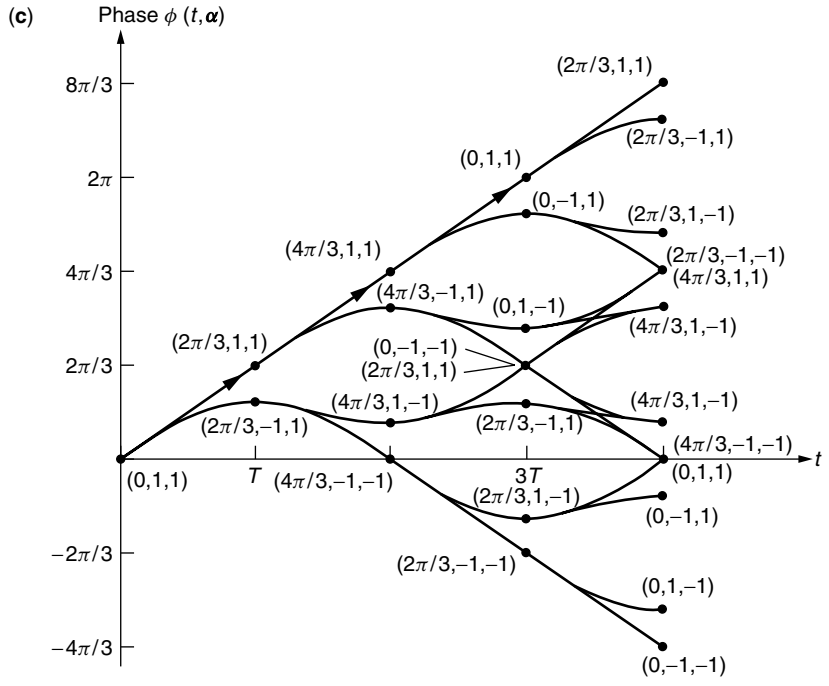


Figure 1. (*Continued*)

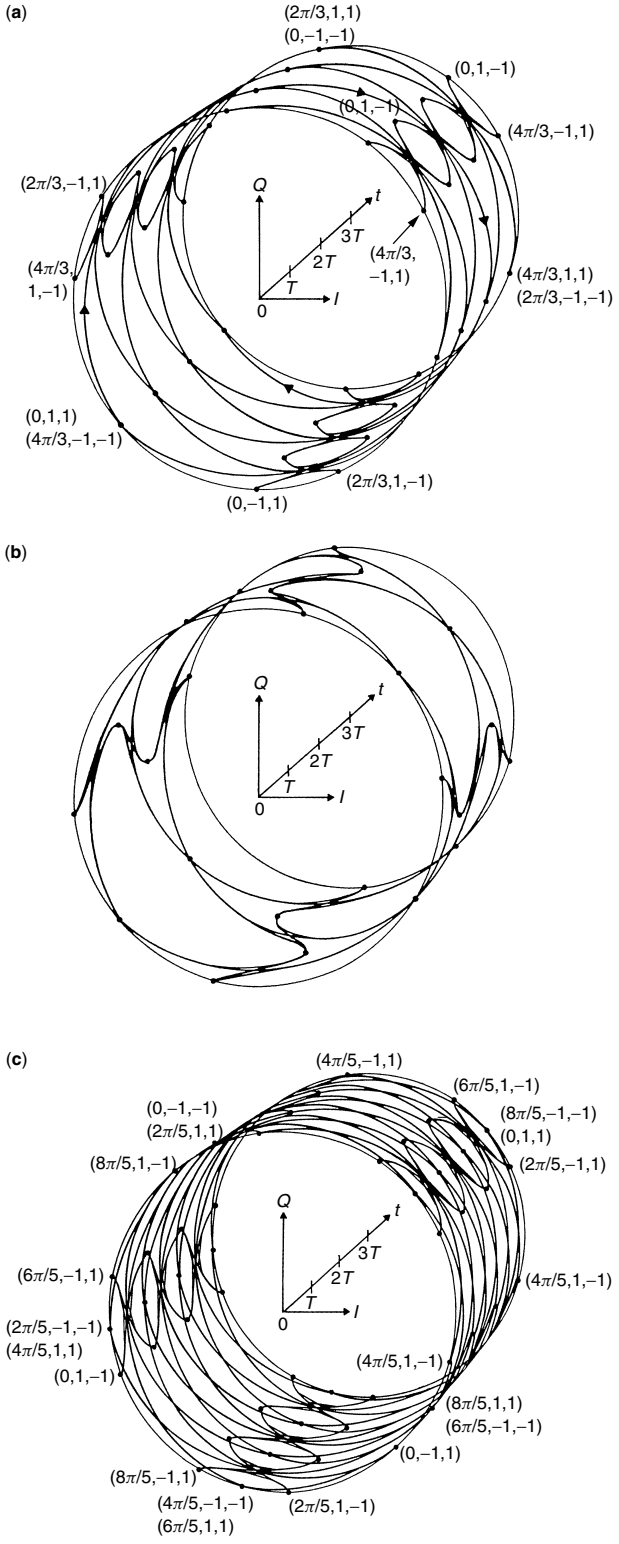


Figure 2. (**a**) Phase cylinder for 3RC with $h = \frac{2}{3}$ and $M = 2$. Compare the phase tree in Fig. 1c. Note the arbitrary phase offset between the phase in the tree in Fig. 1c and the cylinder, in Fig. 2a. Also note that the tree is plotted over T and the cylinder over $3T$. Notice the transitions with arrows; these are also shown in the tree in Fig. 1c. (**b**) phase cylinder for 3RC, $h = \frac{1}{2}$, $M = 2$. (**c**) phase cylinder for $M = 2$, 3RC with $h = \frac{4}{5}$ (10 states).

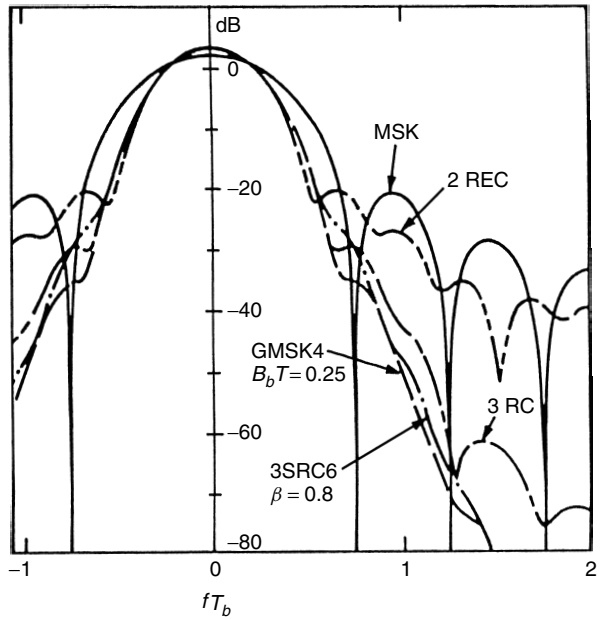


Figure 3. Average power spectrum for some CPM schemes with $h = \frac{1}{2}$ and $M = 2$. See Table 1 for details.

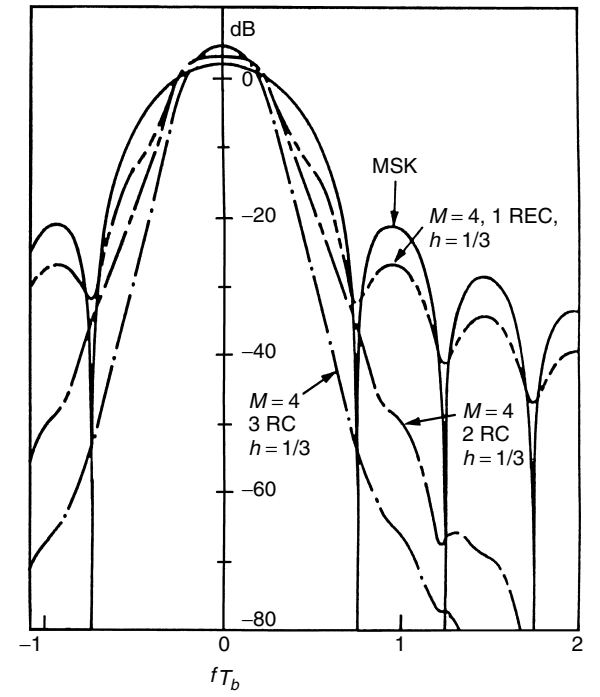


Figure 4. Average power spectra for MSK ($M = 2$, 1REC, $h = \frac{1}{2}$) and the $h = \frac{1}{3}$, $M = 4$ schemes with 1REC, 2RC, and 3RC pulses.

can be described by the finite state and trellis description given above. Although the structure of the optimum ideal coherent receiver for CPM is known [16], it is difficult to evaluate its bit error probability performance. Simulations are required for low channel signal-to-noise ratios. The most convenient and useful parameter for describing the error probability of CPM schemes with maximum-likelihood sequence detection is the minimum Euclidean

distance between all pairs of signals

$$D^2_{\min} = d^2_{\min} \cdot 2E_b = \min\left\{2E_b \log_2(M)\frac{1}{T} \times \int_0^{NT} [1 - \cos[\phi(t, \boldsymbol{a}) - \phi(t, \boldsymbol{b})]]dt\right\} \quad (4)$$

where E_b is the signal energy per bit given by $E_b \log_2 M = E$ and NT is length of the receiver observation interval. When N is sufficiently large, the largest obtainable distance, called the *free distance*, is reached.

For ideal coherent transmission over an additive white Gaussian noise (AWGN) channel, the bit error probability for high signal-to-noise ratios E_b/N_0 is approximately

$$P_b \approx Ce^{-d^2_{\min} E_b/N_0} \quad (5)$$

where C is a constant.

Efficient algorithms are available for computing the minimum distance for different $g(t)$, L, h, and M [16,17]. Figure 5 shows an upper bound d_B^2 to the free distance d_f^2 as a function of h for binary LRC codes; d_B is found by considering certain error events [16]. The term d_B^2 equals the free distance d_f^2 for almost all h values in Fig. 5. Note that the distance grows with L, at least for larger h. The vertical axis of Fig. 5 also has a decibel (dB) scale that gives the gain in E_b/N_0 relative to MSK (which has $d_f^2 = 2$).

The comparisons in Fig. 5 are somewhat artificial, since the bandwidth of a CPM scheme also changes with L and h. Another comparison of CPM schemes is given in the scatterplot in Fig. 6, where each point

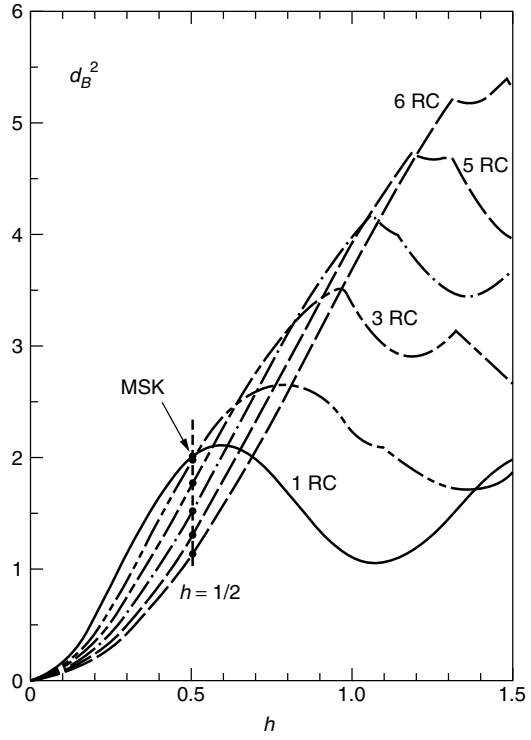


Figure 5. Upper bound d_B^2 on the distance for the binary CPM schemes 1 RC, 2 RC, ..., 6 RC.

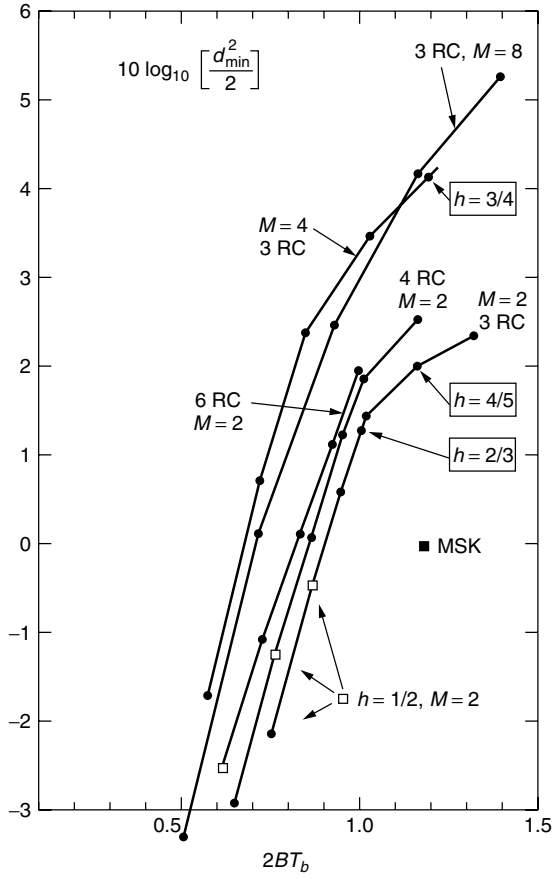


Figure 6. Power–bandwidth tradeoff for CPM schemes using raised-cosine pulses.

represents a system with its 99% power RF bandwidth $2BT_b$ (where $T_b = T/\log_2 M$) and the required signal-to-noise ratio (SNR) relative to MSK in dB at high E_b/N_0. Thus schemes on the same vertical line (for example, through the MSK point) have the same bandwidth at equal data rates; schemes on the same horizontal line have similar error probability for high SNRs. It is evident that larger L and larger M yield more energy- and bandwidth-efficient systems. Not surprisingly, the system complexity increases in the same direction. We have marked some binary $h = \frac{1}{2}$ schemes, some binary 3RC schemes and some quaternary 3RC schemes in Fig. 6. The $h = \frac{2}{3}$ and $\frac{4}{5}$ binary 3RC schemes correspond to the phase cylinders shown above in Fig. 2a,c. For these two schemes, the number of states is 12 and 20, respectively.

The bandwidth and energy of CPM signaling may also be compared to that of other coded modulations. As a class, CPM lies in a middle range of bandwidth and energy; it achieves a moderate error rate ($\sim 10^{-5}$) in 0.5–1.5 Hz·s/data bit and 4–12 dB for E_b/N_0, depending on the code parameters and complexity. Ordinary parity-check coding together with a simple modulation like QPSK work at wider bandwidth and lower energy, and this coding partially overlaps the CPM range on the wideband/low-energy side. In the area of overlap, CPM has the advantage of constant RF envelope. Trellis-coded modulation (TCM) is another coded modulation method

based on set partitioning. It works in a region of narrower bandwidth and higher energy than CPM and partially overlaps it on the other side. In the overlap region, TCM needs either half the bandwidth or 5 dB less energy than CPM. TCM and CPM, however, are not directly comparable because TCM demands a linear channel with accurate amplitude response and CPM does not. Transmitter amplifiers with sufficient linearity for TCM are 2–4 dB less efficient than nonlinear amplifiers of the sort that can be used with CPM signals [18].

5. GENERATING CPM SIGNALS

A conceptual general transmitter structure is shown in Fig. 7. This is based on Eq. (1). This structure demonstrates the fact that CPM is "digital FM"; that is, it is ordinary linear pulsetrain modulation, with pulseshaper $g(t)$, but applied to an FM modulator instead of an amplitude modulator. However, the structure is not easily converted into hardware for coherent systems. The reason is that an exact relation between the symbol rate and the modulation index is required and this requires control circuitry.

The most general and straightforward way of implementing a robust CPM transmitter is to use stored lookup tables. This is seen by rewriting the normalized CPM waveform $S_0(t, \boldsymbol{a}_n) = S(t, \boldsymbol{a}_n)/[2E/T]^{1/2}$ as

$$S_0(t, \boldsymbol{a}_n) = I(t)\cos(2\pi f_0 t) - Q(t)\sin(2\pi f_0 t) \qquad (6)$$

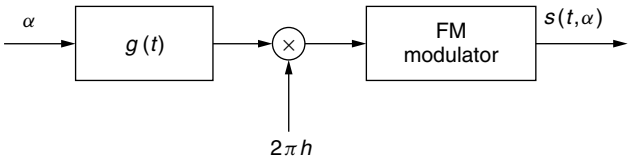


Figure 7. Conceptual modulator for CPM based on Eq. (1).

where $I(t) = \cos(\theta(t, \boldsymbol{a}_n) + \theta_n)$ and $Q(t) = \sin(\theta(t, \boldsymbol{a}_n) + \theta_n)$. The subscript n on $\boldsymbol{a}_n$ indicates that we are considering the data symbol a_n and sufficiently many of the previous symbols. Figure 8 shows a transmitter based on Eq. (6) where the two read only memories contain sampled and quantized versions of $I(t)$ and $Q(t)$ for each data symbol a_n, correlative state vector (the $L-1$ previous data symbols) and phase-state value. The address field for the ROM is roughly $L\log_2 M + [\log_2 p] + 1$ bits and the ROM size is $p \cdot M^L \cdot m \cdot m_q$ bits, where m is the number of samples per symbol time and m_q is the number of bits per quantized sample. The transmitter in Fig. 8 also contains a small sequential machine with a phase state lookup table for calculating the next phase state, given the previous one and the incoming data symbol. The transmitter also contains two D/A converters.

For binary 3RC with $h = \frac{2}{3}$, for example, the ROM (read-only memory) address length is 5-bits and the ROM size is 1024 bits. For a wide range of CPM parameters, the ROM size is manageable. Alternative transmitter structures are possible that have reduced lookup tables.

Several special modulator structures have been devised for MSK. Because MSK is a quadrature-multiplexed modulation scheme, it can be optimally detected by coherently demodulating its in-phase and quadrature components in parallel. The quadrature channels of the modulator and demodulator must be time synchronized, amplitude balanced, and in phase quadrature. The serial method [19] is an alternative approach to parallel modulation and demodulation of MSK which avoids some of these problems.

6. RECEIVERS

Receivers for coherent CPM are an active area of research. For a general CPM scheme with a rational modulation index h and a pulse of finite length L, ideal

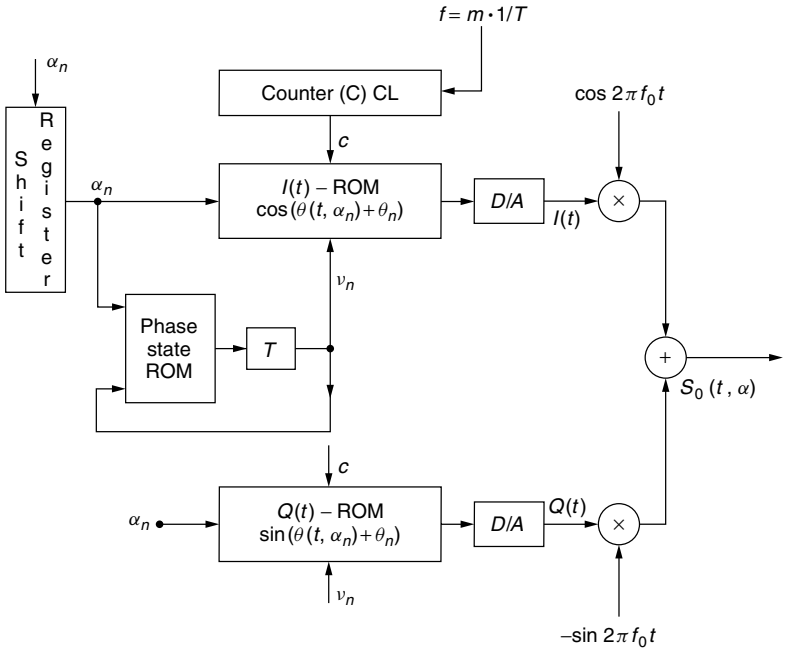


Figure 8. General CPM transmitter with the lookup table principle.

optimum coherent detection can be performed by means of the Viterbi algorithm. The state and trellis description discussed earlier is used. The metric is calculated in a bank of linear filters that are sampled every symbol interval. Figure 9 shows the conceptual diagram for this general receiver. The path memory in the trellis processor causes a delay of N_T symbol intervals. The N_T needed is related to the growth of the minimum distance with the observation interval length. It should be sufficiently large that the free distance is obtained between all paths.

Although there are no special theoretical problems in constructing a receiver based on the principles illustrated in Fig. 9 there are several practical ones. As with all Viterbi detectors, the complexity grows exponentially with signal memory. The limiting factors are the number of states $S = pM^{L-1}$ and the number of filters $F = 2M^L$ for calculating the metrics. For many cases with long smoothing pulses, the optimum receiver can be approximated by a receiver based on a shorter and simpler pulseshape $g_R(t)$ of length $L_R < L$. Thus the complexity of the suboptimum receiver is reduced by a factor of M^{L-L_R} for both the number of states and the number of filters in the filter bank. The simpler pulse shape can be optimized (for large SNRs) for a given transmitter pulseshape and modulation index. The loss in error performance can be very small.

For some cases of CPM it is not necessary to use the Viterbi detector. Much work has been devoted to the so-called MSK-type receiver, which is based on the structure given in Fig. 10. The circuit is inspired by the parallel MSK receiver; it has only two filters and just a small amount of processing. The receiver makes single

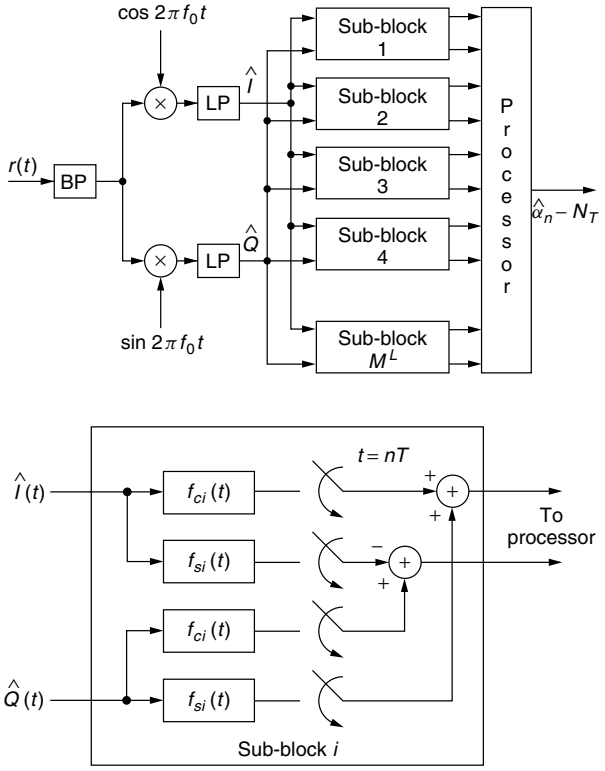


Figure 9. A general receiver structure for CPM based on the Viterbi algorithm. There are $4M^L$ linear filters.

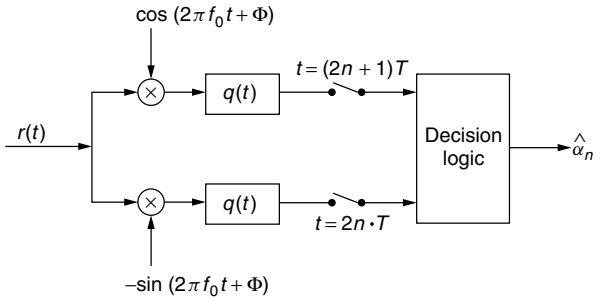


Figure 10. Receiver structure for a parallel MSK-type receiver for binary $h = \frac{1}{2}$ CPM.

symbol decisions. This simplified receiver is suboptimum but it works well for binary modulations with modulation index $h = \frac{1}{2}$. The decisions are made every $2T$ in alternate quadrature arms. Various ideas for selecting the receiver filters are analyzed in the early papers on this subject and an optimum filter has been derived for various correlative FSK schemes (i.e., ones with piecewise linear phase functions) [20] and for smooth pulses [21]. The performance for this type of receiver is almost equal to the optimum Viterbi receiver for schemes with a moderate degree of smoothing, that is, overlapping frequency pulses of length L up to three to four symbol intervals, like 3RC, 4RC, TFM, and some GMSK schemes.

A large literature exists on linear receiver simplifications for CPM. Further details are given in Ref. 16 and the survey article [22].

7. ADVANCED TOPICS

Several topics on more advanced levels are suggested:

Combinations of Convolutional Codes and CPM. The CPM-coded modulations just described may be preceded by an ordinary convolutional code with a rate such as $\frac{1}{2}$ or $\frac{2}{3}$. Pioneering work appeared in [23]. The result is properly considered a concatenated code. Modern techniques of iterative decoding and soft information transfer between the two decoders may be applied, but the joint code trellis is simple enough that outright maximum-likelihood decoding is practical. Convolutional plus CPM code constructions have a wider bandwidth than CPM coding alone, but they have very good minimum distance.

Partially Coherent Detection. A problem of practical interest is how to detect CPM signals when the phase reference is not completely known. Only the derivative of phase (i.e., frequency) may be known, or the phase reference may be stable but subject to an unknown constant offset. In the latter case the minimum distance of CPM is often unaffected.

Reduced-Search Receivers. CPM codes are trellis codes, and as such they can be detected with trellis search algorithms that view only a reduced part of the code trellis. Reduced searching for CPM in fact works well, better than it does for ordinary convolutional codes. Schemes such as the M algorithm need only extend two to four paths

through a complicated trellis. Indicative results are presented in Ref. 24.

Shannon Theory of CPM. An information-theoretic channel that models CPM is unfortunately one with memory, and the problem of computing Shannon information rates for CPM channels is difficult. However, methods exist to compute the cutoff rate, an underbound to capacity that is generally tight at moderate to high signal energy. Cutoff rate studies have been performed by a number of authors, beginning with Ref. 11.

Filtered CPM. An extensive literature exists on CPM that has undergone channel filtering. Early research is discussed in Ref. 16. In general, it can be said that removal of spectral sidelobes by filtering has little effect on CPM detection. More narrow filtering can have a severe effect unless special receivers are used, in which case detection losses are much reduced and can sometimes be removed completely [25].

8. FURTHER READING

A monograph devoted to CPM is Ref. 16. General textbooks that include an introductory chapter about CPM include, for example, Ziemer and Peterson [26] and Proakis [27]. Useful tutorial articles in the literature include Refs. 18 and 22.

BIOGRAPHY

Carl-Erik W. Sundberg received the M.S.E.E. and the Dr. Techn. degrees from the University of Lund, Lund, Sweden in 1966 and 1975 respectively. From 1977 to 1984, he was a Research Professor (Docent) in Telecommunication Theory, University of Lund. From 1984 to 2000, he was a Distinguished Member of Technical Staff (DMTS) at Bell Laboratories, Murray Hill, New Jersey, and during 2001 he was a DMTS at Agere Systems, Murray Hill. He currently is a Senior Scientist at iBiquity Digital Corp., Warren, New Jersey. His research interests include source and channel coding, digital modulation, fault-tolerant systems, digital mobile radio, digital audio broadcasting, spread-spectrum, digital satellite systems, and optical communications. He has published more than 95 journal papers and contributed over 140 conference papers. He has 67 patents, both granted and pending. He is a coauthor of *Digital Phase Modulation*, (New York: Plenum, 1986), *Topics in Coding Theory*, (New York: Springer-Verlag, 1989) and *Source-Matched Digital Communications* (New York: IEEE Press, 1996). In 1986 he received the IEEE Vehicular Technology Society's Paper of the Year Award, and in 1989 he was awarded the Marconi Premium Proc. IEE Best Paper Award. Two of his papers were selected for inclusion in the IEEE Communications Society 50th Anniversary Journal Collection, Volume 2002. He is a Fellow of the IEEE and is listed in *Marquis Who's Who in America*.

John B. Anderson was born in New York State in 1945 and received the Ph.D. degree in electrical engineering from Cornell University in 1972. He has been a faculty member at McMaster University in Canada, Rensselaer Polytechnic Institute in Troy, New york, and since 1998 has held the Ericsson Chair in Digital Communication at Lund University, Lund, Sweden. His researcg work is in coding and digital commuication, bandwidth-efficient coding, and practical application of these. Dr. Anderson has served as president of the IEEE Information Theory Society and editor-in-chief of IEEE Press. He is the author of five textbooks, including the forthcorning *Coded Modulation Systems* (Plenum, 2002). He is fellow of the IEEE (1987) and received the Humboldt Research Prize in 1991 and the IEEE Third Millenmium Medalj in 2000.

BIBLIOGRAPHY

1. U.S. Patent 2,917,417 (March 28, 1961), M. L. Doelz and E. H. Heald, Minimum shift data communication system.
2. R. de Buda, Coherent demodulation of frequency-shift keying with low deviation ratio, *IEEE Trans. Commun.* **COM-20**(3): 429–436 (June 1972).
3. M. G. Pelchat, R. C. Davis, and M. B. Luntz., Coherent demodulation of continuous phase binary FSK signals. *Proc. Int. Telemetering Conf.*, Washington, DC, Nov. 1971, pp. 181–190.
4. W. P. Osborne and M. B. Luntz, Coherent and noncoherent detection of CPFSK, *IEEE Trans. Commun.* **COM-22**(8): 1023–1036 (Aug. 1974).
5. T. A. Schonhoff, Symbol error probabilities for M-ary CPFSK: Coherent and noncoherent detection, *IEEE Trans. Commun.* **COM-24**(6): 644–652 (June 1976).
6. H. Miyakawa, H. Harashima, and Y. Tanaka, A new digital modulation scheme — multimode binary CPFSK, *Proc. 3rd Int. Conf. Digital Satellite Communications*, Kyoto, Japan, Nov. 1975, pp. 105–112.
7. J. B. Anderson and D. P. Taylor, A bandwidth-efficient class of signal space codes, *IEEE Trans. Inform. Theory* **IT-24**(6): 703–712 (Nov. 1978).
8. T. Aulin, *Three Papers on Continuous Phase Modulation* (*CPM*), Ph.D. thesis (on telecommunication theory), Univ. Lund, Lund, Sweden, Nov. 1979.
9. T. Aulin and C.-E. Sundberg, Continuous phase modulation — Part I: Full response signaling, *IEEE Trans. Commun.* **COM-29**(3): 196–209 (March 1981).
10. T. Aulin, N. Rydbeck, and C.-E. Sundberg, Continuous phase modulation — Part II: Partial response signaling, *IEEE Trans. Commun.* **COM-29**(3): 210–225 (March 1981).
11. J. B. Anderson, C.-E. Sundberg, T. Aulin, and N. Rydbeck, Power-bandwidth performance of smoothed phase modulation codes, *IEEE Trans. Commun.* **COM-39**(3): 187–195 (March 1981).
12. F. de Jager and C. B. Dekker, Tamed frequency modulation, a novel method to achieve spectrum economy in digital transmission, *IEEE Trans. Commun.* **COM-26**(5): 534–542 (May 1978).
13. K. S. Chung, General tamed frequency modulation and its application for mobile radio communication, *IEEE J. Select. Areas Commun.* **SAC-2**(4): 487–497 (July 1984).
14. K. Murota and K. Hirade, GMSK modulation for digital mobile telephony, *IEEE Trans. Commun.* **COM-29**(7): 1044–1050 (July 1981).
15. G. S. Deshpande and P. H. Wittke, Correlative encoded digital FM, *IEEE Trans. Commun.* **COM-29**(2): 156–162 (Feb. 1981).

16. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.

17. S. G. Wilson and M. G. Mulligan, An improved algorithm for evaluating trellis phase codes, *IEEE Trans. Inform. Theory* **IT-30**(6): 846–851 (Nov. 1984).

18. J. B. Anderson and C.-E. Sundberg, Advances in constant envelope coded modulation, *IEEE Commun. Mag.* **30**(12): 36–45 (Dec. 1991).

19. F. Amoroso and J. A. Kivett, Simplified MSK signal technique, *IEEE Trans. Commun.* **COM-25**(4): 433–441 (April 1977).

20. P. Galko and S. Pasupathy, Optimal linear receiver filters for binary digital signals, *Proc. Int. Conf. Communication*, Philadelphia, PA, June 1982, pp. 1H.6.1–1H.6.5.

21. A. Svensson and C.-E. Sundberg, Optimum MSK-type receivers for CPM on Gaussian and Rayleigh fading channels, *IEE Proc., Part F, Commun., Radar, Signal Process* **131**(8): 480–490 (Aug. 1984).

22. C.-E. Sundberg, Continuous phase modulation, *IEEE Commun. Mag.* **24**(4): 25–38 (April 1986).

23. G. Lindell, *On Coded Continuous Phase Modulation*, Ph.D. thesis (on telecommunication theory), Univ. Lund, Lund, Sweden, May 1985.

24. S. J. Simmons and P. H. Wittke, Low complexity decoders for constant envelope digital modulations, *IEEE Trans. Commun.* **COM-31**(12): 290–295 (Dec. 1983).

25. N. Seshadri and J. B. Anderson, Asymptotic error performance of modulation codes in the presence of severe intersymbol interference, *IEEE Trans. Inform. Theory* **IT-34**: 1203–1216 (Sept. 1988).

26. R. E. Ziemer and R. L. Peterson, *Digital Communications and Spread Spectrum Systems*, Macmillan, New York, 1985.

27. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.

CONTINUOUS PHASE FREQUENCY SHIFT KEYING (CPFSK)

THOMAS A. SCHONHOFF
Titan System Corporation
Shrewsbury, Massachusetts

1. INTRODUCTION

Continuous phase frequency shift keying (CPFSK) is a modulation that, as its name implies, can be characterized as a traditional frequency shift keyed (FSK) signal constrained to maintain continuous phase at its symbol time boundaries. This constraint offers two important advantages from a communication point of view:

1. The continuous phase at the symbol boundaries essentially "smooths" the waveform, thereby offering a signal bandwidth that can be considerably smaller than conventional modulations such as FSK or phase-shift-keying (PSK). The spectral characteristics are presented in Section 3.
2. The waveform during each symbol period is dependent on the data and waveforms during previous symbol periods (i.e., the signal waveform contains memory). This memory can be used to improve error rate performance relative to more conventional modulations. The error rate performance is discussed in Section 4.

Historically, CPFSK is a generalization of minimum shift keying (MSK) [1] and, as shown below, MSK is indeed one form of CPFSK. In turn, CPFSK has been generalized to continuous phase modulation (CPM) [2]. Some specific CPM techniques and their relationship to CPFSK are given in Section 2.3.

2. DEFINITION OF CPFSK

During the ith symbol period, the transmitted CPFSK waveform can be written as

$$s(t) = \sqrt{\frac{2E}{T}} \cos\left(2\pi f_c t + \frac{d_i \pi h[t-(i-1)T]}{T} + \pi h \sum_{j=i-1} d_j + \phi\right) \quad (1)$$

where E is the transmitted signal energy during symbol period T and f_c is the carrier frequency. d_i represents the digital data during the ith symbol; for M-ary signaling, $d_i = \pm 1, \ldots, \pm(M-1)$. h is referred to as the deviation ratio. Its importance is made evident in succeeding sections. The starting phase at the beginning of the i^{th} symbol is seen to be $\pi h \sum_{j=i-1} d_j$. This term shows that previous symbols have an effect on the transmitted waveform. ϕ is the initial starting phase.[1]

2.1. Phase Trajectories of CPFSK

From the preceding equation, the phase term is $\vartheta(t) = \frac{d_i \pi h[t-(i-1)T]}{T} + \pi h \sum_{j=i-1} d_j$. This is known as the phase trajectory, and an example for quaternary CPFSK is shown in Fig. 1.

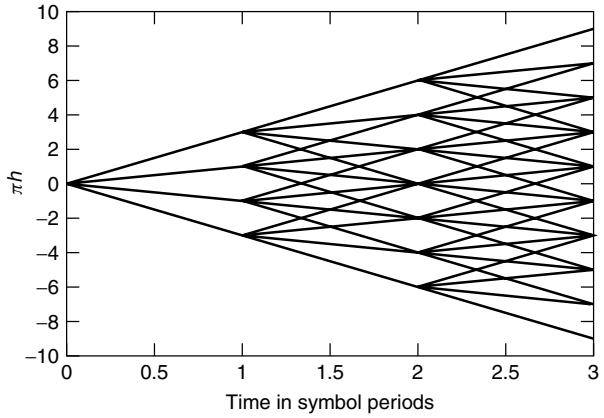


Figure 1. Quaternary CPFSK phase trajectories.

[1] It is assumed that starting phase at time zero is 0.

For the specific case of $h = \frac{1}{2}$, the signal is known as minimum shift keying (MSK) and is widely used in radio systems for which the transmitted frequency is 500 kHz or less.[2]

2.2. Frequency modulation Interpretation of CPFSK

Since the instantaneous transmitted frequency of the signal is the time derivative of the phase, it is clear from Eq. (1) that, during every symbol periods, one of M possible transmitted frequencies are sent, namely $2\pi f_c + \dfrac{d_i \pi h}{T}$, where $d_i = \pm 1, \ldots, \pm(M-1)$.

2.3. Relationship to General CPM

Continuous phase modulation (CPM) is a generalization of CPFSK where the generalization uses the two criteria of performance improvements explained in the Introduction. For example, one approach for CPM is to make the modulation even "smoother" (i.e., make not only the phase continuous but also higher derivatives of the phase continuous). The second approach is to introduce more memory into the modulation by using phase and/or frequency pulse shapes that extend over more than one symbol period. In this case, memory is introduced not only by having continuous phase, frequency, etc. at the symbol boundaries, but also by the pulse shapes themselves. This latter approach has come to be called partial-response CPM.

Adapting the notation of Ref. 4, a general CPM signal can be written in terms of its phase as

$$\vartheta(t;\vec{d}) = 2\pi \sum_{i=-\infty}^{n} d_i h_i q(t - iT), \quad nT \le t \le (n+1)T \qquad (2)$$

In this article, we simply identify the parameters in this equation. The interested reader is directed to Ref. 2 or 4, This equation shows that the phase is determined, in general, by the vector of all past data $\vec{d}$, which represent a sequence of independent M-ary symbols taken from the set $\{\pm 1, \pm 3, \ldots, \pm(M-1)\}$.

Another item of note in Eq. (2) is that, in general, a different value of the modulation index h_i can be used for each symbol period. At present, a popular digital modulation candidate for military UHF satellite systems is a form of this multi-h quaternary modulation. The normalized phase shape $q(t)$ extends over a general L symbol periods and, in general, is constrained only by its end regions. These constraints can be written as

$$q(t) = \begin{cases} 0, & t < 0 \\ \frac{1}{2}, & t \ge LT \end{cases} \qquad (3)$$

3. TRANSMITTED SPECTRAL PROPERTIES OF CPFSK

The general derivation of the power spectral density of CPFSK was first given in Salz [3], although the derivation and terminology of Proakis [4] is used herein. The power spectral density of CPFSK is shown in Ref. 4 to be

$$\Phi(f) = T\left[\frac{1}{M}\sum_{n=1}^{M} A_n^2(f) + \frac{2}{M^2}\sum_{n=1}^{M}\sum_{m=1}^{M} B_{nm}(f) A_n(f) A_m(f)\right] \qquad (4)$$

where

$$A_n(f) = \frac{\sin\pi\left[fT - \frac{1}{2}(2n-1-M)h\right]}{\pi\left[ft - \frac{1}{2}(2n-1-M)h\right]}$$

$$B_{nm}(f) = \frac{\cos(2\pi fT - \alpha_{nm}) - \varphi\cos\alpha_{nm}}{1 + \varphi^2 - 2\varphi\cos 2\pi fT}$$

$$\alpha_{nm} = \pi h(m + n - 1 - M) \qquad (5)$$

and

$$\varphi = \frac{\sin M\pi h}{M \sin \pi h}$$

Figure 2 shows a plot of the one-sided power spectral density of binary CPFSK for three values of the deviation ratio h, namely, MSK, for which h is $\frac{1}{2}$, the value of h that gives the best binary error rate, namely, $h = 0.715$, and an intermediate value of $h = 0.6$.

Figure 3 shows a comparable one-sided power spectral density for three selected values of h for quaternary CPFSK.

4. RECEIVER STRUCTURES AND ERROR RATE PERFORMANCE OF CPFSK

Much of this section is based on the original developments of Refs. 5 and 6. The initial work in Ref. 5 developed the

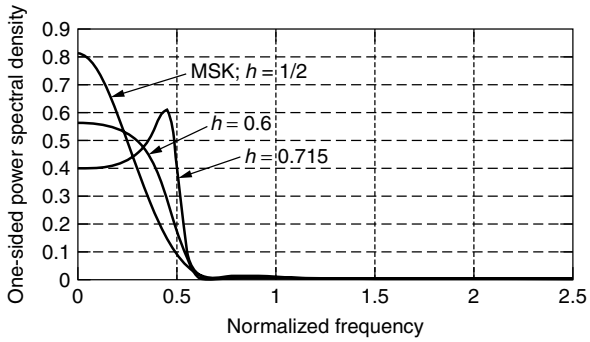


Figure 2. One-sided power spectral density of binary CPFSK.

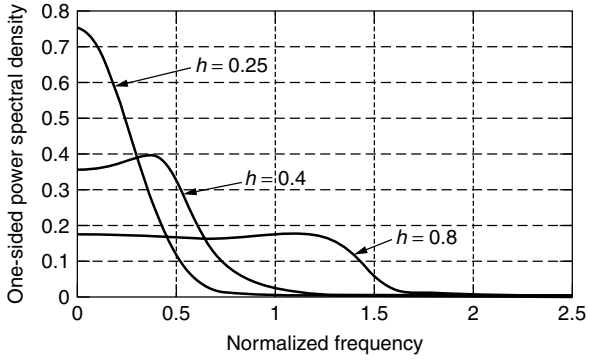


Figure 3. One-sided power spectral density of quaternary CPFSK.

[2] Unfortunately, when $h = \frac{1}{2}$, the memory benefits of CPFSK are significantly reduced so that the error rate performance is theoretically identical to antipodal PSK.

theory and structures for binary signaling, and this was generalized and expanded upon in Ref. 6. Two receiver structures associated with coherent and noncoherent processing, respectively, are presented in Sections 4.1 and 4.2, whereas a compilation of alternative structures is presented in Section 4.3.

4.1. Coherent Structures and Performance

A shorthand notation for the received CPFSK signal can be written as

$$r(t) = \alpha s(t, d_1, D_k) + n(t) \tag{6}$$

where α corresponds to the received signal attenuation, $s(t, d_1, D_k)$ is a compact representation of the transmitted signal of Eq. (1), d_1 is the first symbol on which we wish to make a decision, $D_k = \{d_2, \ldots, d_n\}$ corresponds to the next n symbols, and $n(t)$ is a narrowband zero-mean white Gaussian noise process with a double-sided power spectral density of $N_0/2$. We wish to observe n symbols and make a decision on the first.

It follows that the M likelihood parameters can be written as

$$l_1 = \iiint_{n-\text{fold}} \exp\left(\frac{2}{N_0}\right) \int_0^{nT} r(t)s(t, 1, D_k) f(D_k)\, dD_k$$

$$l_2 = \iiint_{n-\text{fold}} \exp\left(\frac{2}{N_0}\right) \int_0^{nT} r(t)s(t, -1, D_k) f(D_k)\, dD_k$$

$$\vdots$$

$$l_M = \iiint_{n-\text{fold}} \exp\left(\frac{2}{N_0}\right) \int_0^{nT} r(t)s(t, -(M-1), D_k) \times f(D_k)\, dD_k \tag{7}$$

where $f(D_k)$ is the discrete pdf of the $(n-1)$-tuple. Evaluating over these n integrals results in the n decision variables

$$U_1 = \sum_{j=1}^{m} \exp\left(\frac{2}{N_0} \int_0^{nT} r(t)s(t, 1, D_j)\, dt\right)$$

$$U_2 = \sum_{j=1}^{m} \exp\left(\frac{2}{N_0} \int_0^{nT} r(t)s(t, -1, D_j)\, dt\right)$$

$$\vdots$$

$$U_M = \sum_{j=1}^{m} \exp\left(\frac{2}{N_0} \int_0^{nT} r(t)s(t, -(M-1), D_j\, dt\right) \tag{8}$$

where $m = M^{n-1}$ and corresponds to all of the possible sequences of the $(n-1)$-tuple D_j.

Although this set of decision variables is very complicated, it allows us to develop an optimal and a suboptimal receiver structure for coherent CPFSK. From Eq. (8), we can see that an optimum receiver structure can be depicted as shown in Fig. 4. $m = M^{n-1}$ correlators (or matched filters) are used for each of the M possible received symbols. All m correlators associated with one of the symbols (e.g., d_i are added to produce the decision

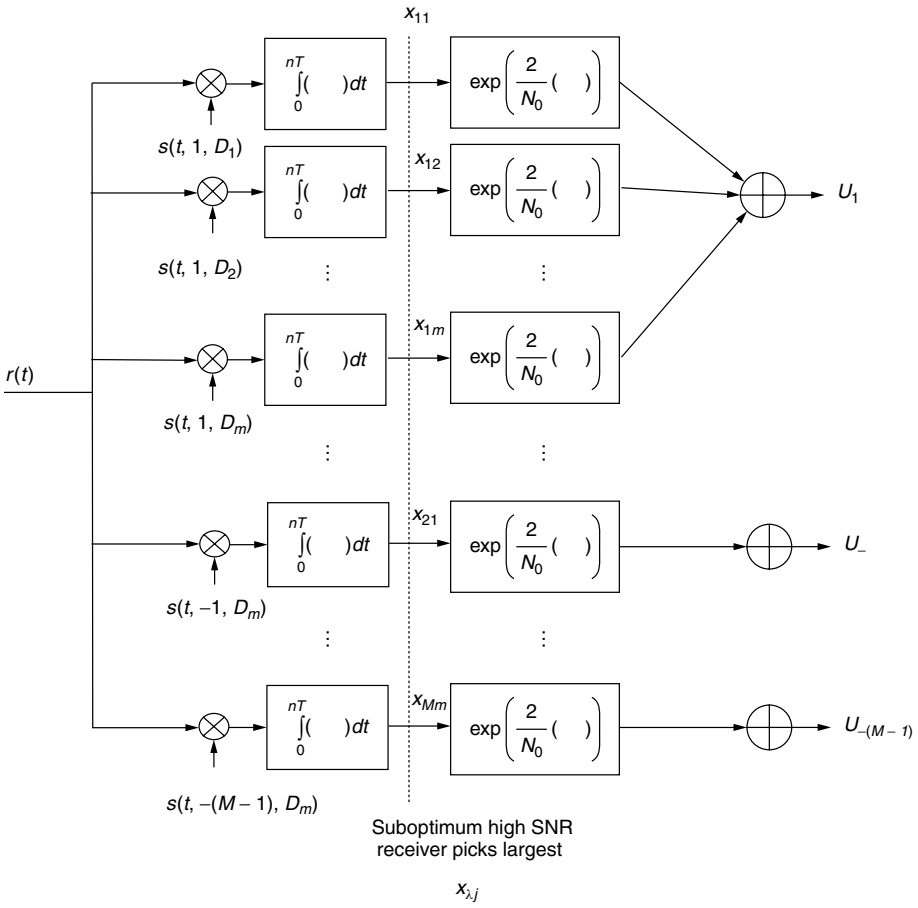


Figure 4. Optimal and suboptimal coherent CPFSK receivers.

variable U_i), and the maximum decision variable is used to estimate which symbol was transmitted during the i^{th} symbol period.

Because of the nonlinear nature of the optimum decision variables, it is analytically impossible to determine the error rate performance of this optimum receiver. Nonetheless, both low snr and high snr bounds can be used to estimate its performance [5,6]. In particular, the high snr approximation uses the fact that the exponential function is monotonic. Thus, if we truncate the receiver structure along the dotted line in Fig. 4, we determine the mM correlator outputs $x_{\lambda j}$, pick the largest, and then base our decision on the subvariable λ. That is, we still make one decision every symbol period, but we use correlators or matched filters that span the last n symbols.

Since the noise is Gaussian, it is clear that each of the correlator outputs is Gaussian and the performance can be estimated using either a union bound as indicated in Refs. 5 and 6 or the minimum distance of all the possible sequences as used in Ref. 2. At high snr, both approaches give comparable estimates. Figures 5 and 6 give binary and quaternary high snr receiver structure error rate performances respectively. These graphs were extracted from the same data as that of [6].

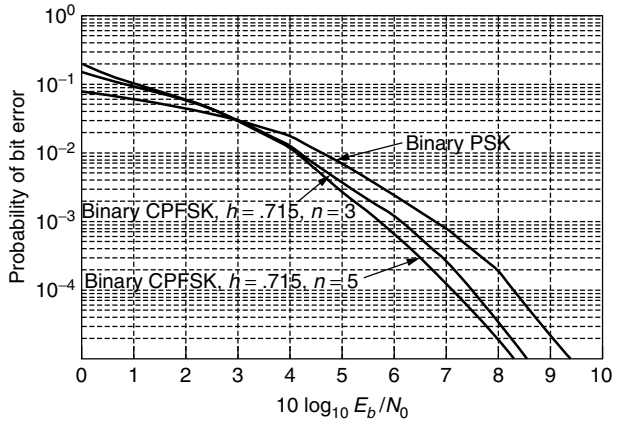


Figure 5. Probability of bit error for selected coherent binary modulations.

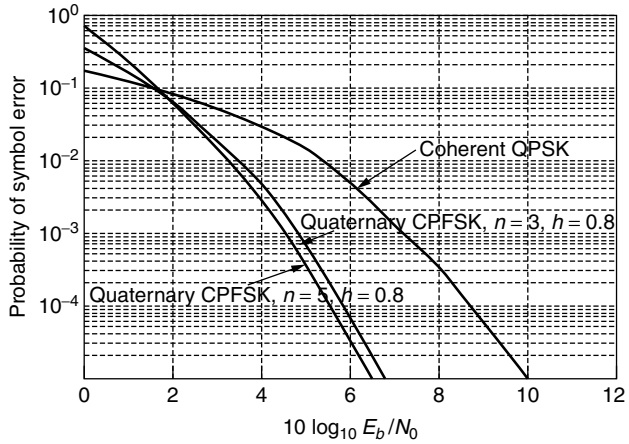


Figure 6. Probability of symbol error for selected coherent quaternary modulations.

For the binary error rate results of Fig. 5, it is known that the optimal value of the deviation ratio is $h = 0.715$. It can be seen from the figure that observing the signal for $n = 3$ or $n = 5$ bits both result in performance improvements over antipodal PSK. For $n = 5$, the improvement is approximately 0.8 dB at an error rate of 10^{-5}.

The results of Fig. 6 show that the improvements of quaternary CPFSK are much more impressive than those of binary CPFSK. Indeed, at an error rate of 10^{-5}, up to 3.5 dB improvement in SNR is possible. The values of h in the last two figures are the optimal values found; however, for quaternary CPFSK in particular, other values of h also result in improved performance. The interested reader is directed to Refs. 2 and 6.

4.2. Noncoherent Structures and Performance

The results of this section again parallel those of the developments in [6]. For noncoherent detection, the initial phase ϕ of Eq. (1) is assumed unknown with a uniform pdf from $(0, 2\pi)$. We use a slightly amended shorthand notation from [6] and model the received signal as

$$r(t) = s(t, d_{n+1}, \Delta_k, \phi) + n(t) \tag{9}$$

where we are observing $2n + 1$ symbols and making a decision on the middle symbol $d_{n+1}.\Delta_k$ is a $2n$-tuple consisting of the symbols before and after the decision symbol $d_{n+1}.\Delta_k$ can then be written as $\Delta_k = \{d_1, d_2, \ldots, d_n, d_{n+2}, \ldots, d_{2n+1}\}$. This implies that the subscript k progresses over $\mu = M^{2n}$ different values. The optimum noncoherent receiver structure is derived from the M likelihood parameters

$$l_1 = \int_\phi \iiint_\Delta \exp\left(\frac{2}{N_0}\int_0^{(2n+1)T} r(t)s(t, 1, \Delta_k, \phi)\,dt\right) \times f(\Delta)f(\phi)\,d\phi\,d\Delta$$

$$l_2 = \int_\phi \iiint_\Delta \exp\left(\frac{2}{N_0}\int_0^{(2n+1)T} r(t)s(t, -1, \Delta_k, \phi)\,dt\right) \times f(\Delta)f(\phi)\,d\phi\,d\Delta$$

$$\vdots$$

$$l_M = \int_\phi \iiint_\Delta \exp\left(\frac{2}{N_0}\int_0^{(2n+1)T} r(t)s(t, -(M-1), \Delta_k, \phi)\,dt\right) \times f(\Delta)f(\phi)\,d\phi\,d\Delta \tag{10}$$

These likelihood ratios can be seen to be similar to those of the coherent receiver structure given by Eq. (7) except for the averaging over the initial phase ϕ which results in a Bessel function of order zero $I_0(x)$. Performing the averages of Eq. (10) results in the M noncoherent decision variables

$$U_1 = \frac{1}{\mu}\sum_{k=1}^{\mu} I_0\left(\frac{2}{N_0}\chi_{1k}\right)$$

$$U_2 = \frac{1}{\mu}\sum_{k=1}^{\mu} I_0\left(\frac{2}{N_0}\chi_{2k}\right)$$

$$\vdots$$

$$U_M = \frac{1}{\mu}\sum_{k=1}^{\mu} I_0\left(\frac{2}{N_0}\chi_{Mk}\right) \tag{11}$$

where χ_{Nk} is a Rician statistical variable defined as

$$\chi_{Nk} = \sqrt{x_{Nk}^2 + y_{Nk}^2} \tag{12}$$

and x_{Nk} and y_{Nk} are the in-phase and quadrature variables, respectively, defined in terms of our shorthand notation as

$$x_{Nk} = \begin{cases} \displaystyle\int_0^{(2n+1)T} r(t)s(t, N, \Delta_k, 0)\,dt & Nodd \\ \displaystyle\int_0^{(2n+1)T} r(t)s(t, \_(N-1), \Delta_k, 0)\,dt & Neven \end{cases} \tag{13}$$

and

$$y_{Nk} = \begin{cases} \displaystyle\int_0^{(2n+1)T} r(t)s\left(t, N, \Delta_k, \frac{\pi}{2}\right)\,dt & Nodd \\ \displaystyle\int_0^{(2n+1)T} r(t)s\left(t, \_(N-1), \Delta_k, \frac{\pi}{2}\right)\,dt & Neven \end{cases} \tag{14}$$

Figure 7 shows the functional block diagram of the optimal noncoherent CPFSK receiver structure, which is developed from Eqs. (11) through (14). As can be seen, a total of M^{2n+1} noncoherent correlators or matched filters are used to make a decision on the middle symbol.

The performance of one example of noncoherently detected CPFSK is shown in Fig. 8. This figure shows that significant improvement is still possible even when the input starting phase is not estimated.

4.3. Other Receiver Structures

The phase trellis, shown in the example in Fig. 1, leads to another receiver structure based on the Viterbi algorithm (VA). This was first identified by Forney in Ref. 7 and explored specifically for CPFSK in Ref. 8. The VA is most useful when the deviation ratio h is a convenient fraction. For example, for binary CPFSK, if $h = \frac{2}{3}$, which is close to the optimum value of 0.715, the phase trellis can be reduced to a phase state diagram as shown in Fig. 9. The VA uses a traditional state history and state metric to determine the most likely path through the trellis or phase-state diagram. References 2 and 8 show that the error rate performance of the VA is virtually indistinguishable from that of the fully implemented multicorrelator receiver as derived in Sections 4.1 and 4.2.

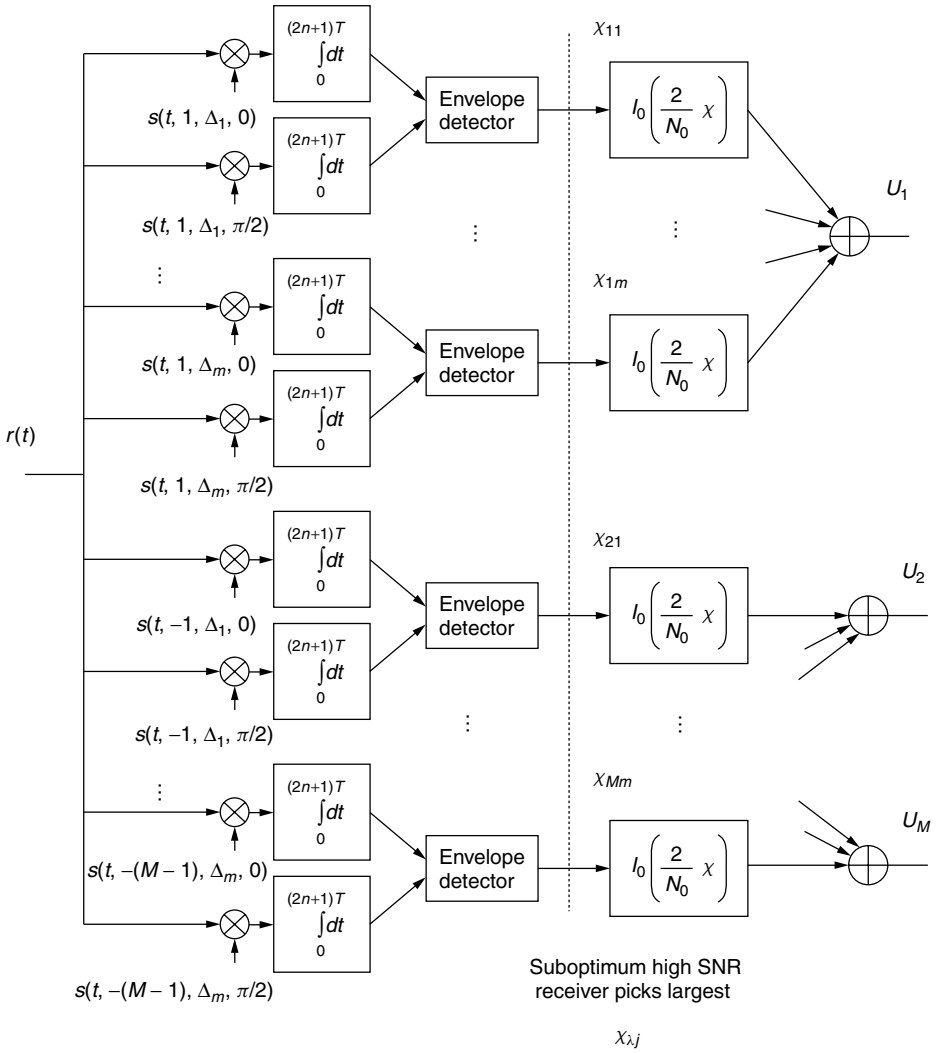


Figure 7. Optimal and high SNR noncoherent CPFSK receiver structure.

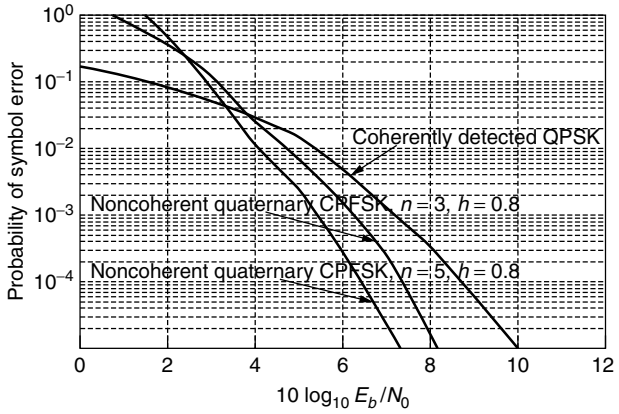


Figure 8. Probability of symbol error for quaternary noncoherent CPFSK.

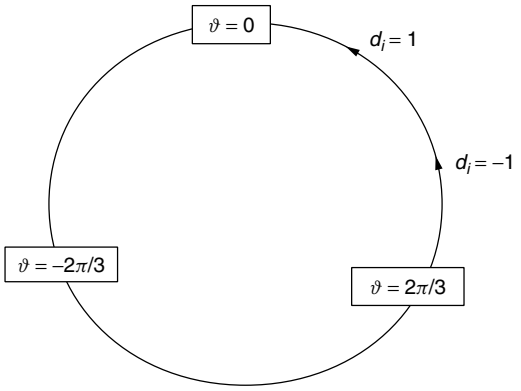


Figure 9. Phase-state diagram for binary CPFSK with deviation ratio of $\frac{2}{3}$.

5. CONCLUSION

This article has presented the concept of continuous phase frequency shift keying (CPFSK). Its transmitted spectra is given and examples are presented. Optimal receiver structures for coherent and noncoherent detection are derived, and examples of symbol error rate are presented. Finally, an alternate receiver structure, based on the maximum likelihood sequence estimator (MLSE) or Viterbi algorithm is outlined.

BIOGRAPHY

Thomas A. Schonhoff received his bachelor's degree from M.I.T., his master's degree from Johns Hopkins University, and his Ph.D. from Northeastern University. He has worked at six different corporations, although he has been with LinCom Corporation (now Titan System Corporation, Communication and Software Solutions Division) since 1985. For the past 21 years, Dr. Schonhoff has also taught graduate courses as an adjunct at Worcester Polytechnic Institute.

BIBLIOGRAPHY

1. Rudi de Buda, Coherent demodulation of frequency shift keying with low deviation ratio, *IEEE Trans. Commun.* 429–435 (1972).
2. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
3. R. R. Anderson and J. Salz, Spectra of digital FM, *Bell System Tech. J.* 1165–1189 (1965).
4. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
5. W. P. Osborne and M. B. Luntz, Coherent and noncoherent detection of CPFSK, *IEEE Trans. Commun.* 1023–1036 (1974).
6. T. A. Schonhoff, Symbol error probabilities for M-ary CPFSK: coherent and noncoherent detection, *IEEE Trans. Commun.* 644–652 (1976).
7. G. D. Forney, Jr., The Viterbi algorithm, *Proc. IEEE* March 268–278 (1973).
8. T. A. Schonhoff, H. Nichols, and H. Gibbons, Use of the MLSE algorithm to demodulate CPFSK, *1978 International Conference on Communications*, Toronto, June 1978.

CONVOLUTIONAL CODES

RICHARD D. WESEL
University of California at Los Angeles
Los Angeles, California

1. INTRODUCTION

Convolutional codes represent one technique within the general class of channel codes. Channel codes (also called *error-correction codes*) permit reliable communication of an information sequence over a channel that adds noise, introduces bit errors, or otherwise distorts the transmitted signal. Elias [1,2] introduced convolutional codes in 1955. These codes have found many applications, including deep-space communications and voiceband modems. Convolutional codes continue to play a role in low-latency applications such as speech transmission and as constituent codes in Turbo codes. Two reference books on convolutional codes are those by Lin and Costello [3] and Johannesson and Zigangirov [4].

Section 2 introduces the shift-register structure of convolutional encoders including a discussion of equivalent encoders and minimal encoders. Section 3 focuses on the decoding of convolutional codes. After a brief mention of the three primary classes of decoders, this section delves deeply into the most popular class, Viterbi decoders. This discussion introduces trellis diagrams, describes the fundamental add-compare-select computation, compares hard and soft decoding, and describes the suboptimal (but commonly employed) finite traceback version of Viterbi decoding.

Section 4 defines the free distance of a convolutional code and describes how free distance may be computed by a specialized application of the Viterbi algorithm. This procedure also yields an analytic lower bound on the decision depth that should be used for finite-traceback decoding. Catastrophic encoders are also discussed in this section. Section 5 describes the generating function that enumerates all the paths associated with error events in the decoder trellis. This section then gives union bounds

on bit error rate that are computed from the generating function. Section 6 provides some final remarks regarding the effective blocklength of convolutional codes and their role today.

2. ENCODER STRUCTURE

As any binary code, convolutional codes protect information by adding redundant bits. A rate-k/n convolutional encoder processes the input sequence of k-bit information symbols through one or more binary shift registers (possibly employing feedback). The convolutional encoder computes each n-bit symbol ($n > k$) of the output sequence from linear operations on the current input symbol and the contents of the shift register(s). Thus, a rate k/n convolutional encoder processes a k-bit input symbol and computes an n-bit output symbol with every shift register update. Figures 1 and 2 illustrate feedforward and feedback encoder implementations of a rate-$\frac{1}{2}$ code. Section 2.1 explores the similarities and differences between feedforward and feedback encoders by examining their state diagrams.

2.1. Equivalent Encoders

Convolutional encoders are finite-state machines. Hence, state diagrams provide considerable insight into their behavior. Figures 3 and 4 provide the state diagrams for the encoders of Figs. 1 and 2, respectively. The states are labeled so that the least significant bit is the one residing in the leftmost memory element of the shift register. The branches are labeled with the 1-bit (single-bit) input and the 2-bit output separated by a comma. The most significant bit (MSB) of the two-bit output is the bit labeled MSB in Figs. 1 and 2.

If one erases the state labels and the single-bit input labels, the remaining diagrams for Figs. 3 and 4 (labeled

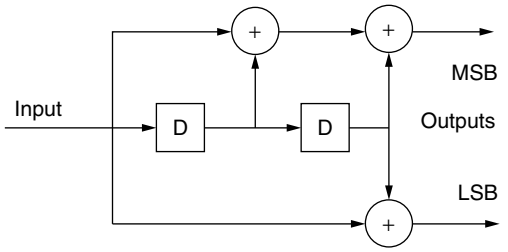


Figure 1. Rate-$\frac{1}{2}$ feedforward convolutional encoder with two memory elements (four states). MSB and LSB refer to the most and least significant bits, respectively.

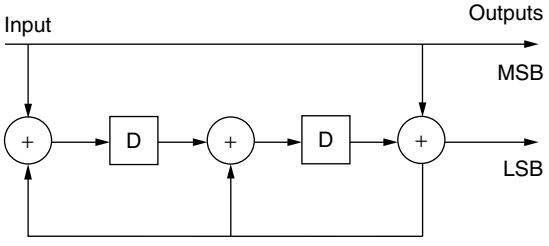


Figure 2. Rate-$\frac{1}{2}$ feedback convolutional encoder with two memory elements (four states).

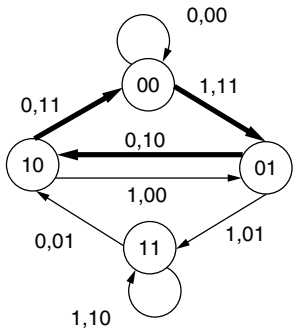


Figure 3. State diagram for rate-$\frac{1}{2}$ feedforward convolutional encoder of Fig. 1.

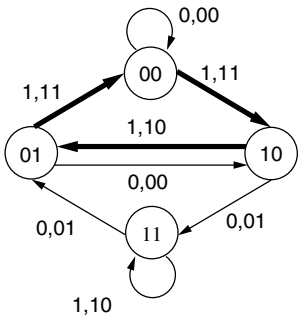


Figure 4. State diagram for rate-$\frac{1}{2}$ feedback convolutional encoder of Fig. 2.

with only the 2-bit outputs) would be identical. This illustrates that the two encoders are equivalent in the sense that both encoders produce the same set of possible output sequences (or codewords). Strictly speaking, a code refers to the list of possible output sequences without specifying the mapping of inputs sequences to output sequences. Thus, as in the above example, two equivalent encoders have the same set of possible output sequences, but may implement different mappings from input sequences to output sequences. In the standard convolutional coding application of transmission over additive white Gaussian noise (AWGN) with Viterbi decoding, such encoders give similar BER performance, but the different mappings of inputs to outputs do lead to small performance differences.

The three-branch paths emphasized with thicker arrows in Figs. 3 and 4 are each the shortest nontrivial (i.e., excluding the all-zeros self-loop) loop from the all-zeros state back to itself. Notice that for Fig. 3, the state diagram corresponding to the feedforward encoder, this loop requires only a single nonzero input. In contrast, for the state diagram corresponding to Fig. 4, this loop requires three nonzero inputs. In fact, for Fig. 4 no nontrivial loop from the all-zeros state to itself requires fewer than two nonzero inputs. Thus the feedforward shift register has a finite impulse response, and the feedback shift register has an infinite impulse response.

This difference is not particularly important for convolutional codes decoded with Viterbi, but it is extremely important to convolutional encoders used as constituents in Turbo codes, which are constructed by concatenating

convolutional codes separated by interleavers. Only feedback encoders (with infinite impulse responses) are effective constituents in Turbo codes. Thus, equivalent encoders can produce dramatically different performance as constituents in Turbo codes, depending on whether or not they meet the requirement for an infinite impulse response.

2.2. Minimal Encoders

A practical question to ask about a convolutional encoder is whether there is an equivalent encoder with fewer memory elements. This question may be answered by performing certain diagnostic computations on the encoder matrix. Furthermore, if the encoder is not minimal, it may be easily "repaired" yielding an encoder that is equivalent but requires fewer memory elements. Forney's classic paper [5] treats this fundamental area of convolutional coding theory elegantly. More recently, Johannesson and Wan [6] extended Forney's results by taking a linear algebra approach. This fascinating area of convolutional code theory is important to convolutional code designers, but less so for "users." Any code published in a table of good convolutional codes will be minimal.

3. DECODING CONVOLUTIONAL CODES

Convolutional code decoding algorithms infer the values of the input information sequence from the stream of received distorted output symbols. There are three major families of decoding algorithms for convolutional codes: sequential, Viterbi, and maximum a posteriori (MAP). Wozencraft proposed sequential decoding in 1957 [7]. Fano in 1963 [8] and Zigangirov in 1966 [9] further developed sequential decoding. See the book by Johannesson and Zigangirov [4] for a detailed treatment of sequential decoding algorithms. Viterbi originally described the decoding algorithm that bears his name in 1967 [10]. See also Forney's work [11,12] introducing the trellis structure and showing that Viterbi decoding is maximum-likelihood in the sense that it selects the sequence that makes the received sequence most likely.

In 1974, Bahl et al. [13] proposed MAP decoding, which explicitly minimizes bit (rather than sequence) error rate. Compared with Viterbi, MAP provides a negligibly smaller bit error rate (and a negligibly larger sequence error rate). These small performance differences require roughly twice the complexity of Viterbi, making MAP unattractive for practical decoding of convolutional codes. However, MAP decoding is crucial to the decoding of Turbo codes. For the application of MAP decoding to Turbo codes, see the original paper on Turbo codes by Berrou et al. [14] and Benedetto et al.'s specific discussion of the basic turbo decoding module [15].

When convolutional codes are used in the traditional way (not as constituents in Turbo codes), they are almost always decoded using some form of the Viterbi algorithm, and the rest of this section focuses on describing Viterbi. The goal of the Viterbi algorithm is to find the transmitted sequence (or codeword) that is closest to the received sequence. As long as the distortion is not too severe, this will be the correct sequence.

3.1. Trellis Diagrams

The state diagrams of Figs. 3 and 4 illustrate what transitions are possible from a particular state regardless of time. In contrast, trellis diagrams use a different branch for each different symbol time. As a result, trellis diagrams more clearly illustrate long trajectories through the states. Figure 5 shows one stage (one symbol time) of the trellis diagram associated with the rate-$\frac{1}{2}$ feedforward encoder of Figs. 1 and 3. Each column of states in the trellis diagram includes everything in the original state diagram. All the branches emanating from states in a particular column are incident on the states in the adjacent column to the right. In other words, each state transition in the trellis moves the trajectory one stage to the right.

To avoid crowding in Fig. 5, branch labels appear at the left and right of the trellis rather than on the branch itself. For each state, the top label belongs to the top branch emanating from or incident to that state. Figure 6 uses thick arrows to show the same path emphasized in the state diagram of Fig. 3. However, in Fig. 3 the beginning and end of the path were not clear. In Fig. 6 the path clearly begins in state 00 and then travels through 01 and then 10 before returning to 00.

3.2. The Basic Viterbi Algorithm

The Viterbi algorithm uses the trellis diagram to compute the accumulated distances (called the *path metrics*) from the received sequence to the possible transmitted sequences. The total number of such trellis paths grows

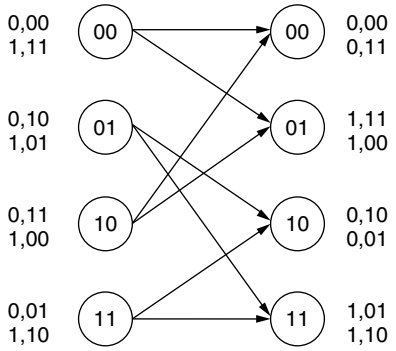


Figure 5. One stage of the trellis diagram for rate-$\frac{1}{2}$ feedforward convolutional encoder of Figs. 1 and 3.

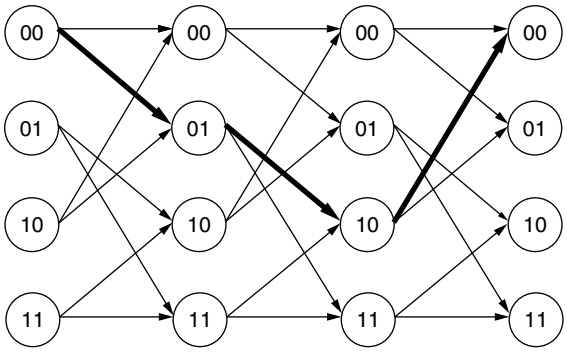


Figure 6. Trellis diagram for the path emphasized in Fig. 3.

exponentially with the number of stages in the trellis, causing potential complexity and memory problems. However, the Viterbi algorithm takes advantage of the fact that the number of paths truly in contention to have the minimum distance is limited to the number of states in a single column of the trellis, assuming that ties may be arbitrarily resolved.

As an example of the Viterbi algorithm, consider transmission over the binary symmetric channel (bit error channel) where the probability of a bit error is less than $\frac{1}{2}$. On such a channel, maximum likelihood decoding reduces to finding the output sequence that differs in the fewest bit positions (has the minimum Hamming distance) from the received sequence. For this example, assume the encoder of Fig. 1 with the state diagram of Fig. 3 and the trellis of Fig. 5. For simplicity, assume that the receiver knows that the encoder begins in state 00.

Figure 7 illustrates the basic Viterbi algorithm for the received sequence 01 01 10. Beginning at the far left column, the only active state is 00. The circle representing this state contains a path metric of zero, indicating that as yet, the received sequence differs from the possible output sequences in no bit positions. Follow the two branches leaving the first column to the active states in the second column of the trellis. Branch metrics label each branch, indicating the Hamming distance between the received symbol and the symbol transmitted by the encoder when traversing that branch. The two hypothetical transmitted symbols are 00 for the top branch and 11 for the bottom (see Fig. 5). Since both differ in exactly one bit position from the received symbol 01, both branch labels are one.

The path metric for each destination state is the sum of the branch metric for the incident branch and the path metric at the root of the incident branch. In the second column, both path metrics are one since the root path metric is zero. These equal path metrics indicate that no path is favored at this point. Now follow the branches from the second column to the third. Exactly one branch reaches each state in the third column. Once again, adding the branch metric and the associated root path metric produces the new path metric.

When following branches from the third column to the fourth, two branches are incident on each state. Only the path with the minimum path metric needs to survive. For example, state 00 (the top state) in the fourth column has a path incident from state 00 in the third column with a path metric of $2 + 1 = 3$. It also has a path incident from state 10 in the third column with a path metric of $3 + 1 = 4$. Only the path with the smaller path metric needs to survive. Figure 7 shows the incident branches of survivor paths with thicker arrows than nonsurvivor paths. Each state in the fourth column has exactly one survivor path, and the values shown indicate the path metrics of the survivor paths.

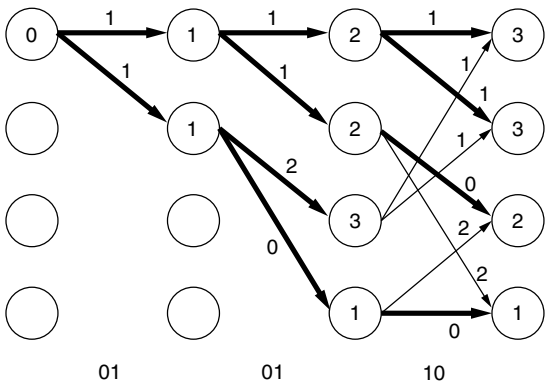


Figure 7. Illustration of the basic Viterbi algorithm on the bit error channel. This is also the trellis for hard Viterbi decoding on the AWGN channel in contrast to the soft Viterbi decoding shown in Fig. 8.

After all received symbols have been processed, the final step in decoding is to examine the last column and find the state containing the smallest path metric. This is state 11, the bottom state, in the fourth column. Following the survivor branches backward from the minimum-path-metric state identifies the trellis path of the maximum likelihood sequence. Reference to Fig. 5 reveals that the maximum likelihood path is the state trajectory $00 \rightarrow 01 \rightarrow 11 \rightarrow 11$. This state trajectory produces the output symbol sequence 11 01 10, which differs in exactly one bit position from the received sequence as indicated by its path metric. The input information sequence is decoded to be 1 1 1.

In this short example, only one trellis stage required path selection. However, once all states are active, path selection occurs with each trellis stage. In fact, if the initial encoder state is not known, path selection occurs even at the very first trellis stage. The basic computational module of the Viterbi algorithm is an add-compare-select (ACS) module. Adding the root path metric and incident branch metric produces the new path metric. Comparing the contending path metrics allows the decoder to select the one with minimum distance. When there is a tie (where two incident paths have the same path metric), a surviving path may be arbitrarily selected. In practice, ties are not uncommon. However, ties usually occur between paths that are ultimately destined to be losers.

3.3. Hard Versus Soft Decoding

The integer branch and path metrics of the binary error channel facilitate a relatively simple example of the Viterbi algorithm. However, the AWGN channel is far more common than the bit error channel. For the AWGN channel, binary phase shift keying (BPSK) represents binary 1 with 1.0 and binary 0 with -1.0. These two transmitted values are distorted by additive Gaussian noise, so that the received values will typically be neither 1.0 nor -1.0. A novice might choose to simply quantize each received value to the closest of 1.0 and -1.0 and assign the appropriate binary value. This quantization would effectively transform the AWGN channel to the bit error channel, facilitating Viterbi decoding exactly as described above. This method of decoding is called hard decoding, because the receiver makes a binary (hard) decision about each bit before Viterbi decoding.

Hard decoding performs worse by about 2 dB than a more precise form of Viterbi decoding known as *soft decoding*. Soft decoding passes the actual received values to the Viterbi decoder. These actual values are called soft values because hard decisions (binary decisions) have

not been made prior to Viterbi decoding. Soft Viterbi decoding is very similar to hard decoding, but branch and path metrics use squared Euclidean distance rather than Hamming distance. Figure 8 works an example analogous to that of Fig. 7 for the case where 1.0 and −1.0 are transmitted over the AWGN channel and soft Viterbi decoding is employed. A fixed-point implementation with only a few bits of precision captures almost all the benefit of soft decoding.

3.4. Finite Traceback Viterbi

The maximum likelihood version of Viterbi decoding processes the entire received sequence and then selects the most likely path. Applications such as speech transmission can conveniently process relatively short data packets in this manner. However, stream-oriented applications such as a modem connection cannot wait until the end of the received sequence before making any decisions about the information sequence. In such cases, a suboptimal form of Viterbi decoding is implemented in which decisions are made about transmitted bits after a fixed delay. This fixed delay is called the traceback depth or decision depth of the Viterbi decoder. The exact choice of the traceback depth is usually determined by simulation, but there is an analytic technique that identifies a good lower bound on what the traceback depth should be. We will discuss this "analytic traceback depth" in the next section, since its computation is a natural by-product of computing the free distance of a convolutional code.

Figures 9 and 10 illustrate finite traceback Viterbi decoding. Figure 9 shows the path metrics and survivor paths (indicated by thick arrows) for a soft Viterbi decoder in steady state. Actually, the selection of survivor paths and path metrics is the same for maximum-likelihood Viterbi decoding and finite-traceback Viterbi decoding. Figure 10 shows the distinguishing behavior of finite-traceback Viterbi decoding. Rather than wait until the end of the received sequence, each k-bit input symbol is decoded after a fixed delay. In Fig. 10 this delay is three symbols. After each update of the path metrics, the path with the smallest metric (identified in Fig. 10 by a thick circle) is traced back three branches and the k-bit input symbol associated with the oldest branch is decoded. Notice that the paths selected by this algorithm do not have to be consistent with each other. For example, the two paths traced back in Fig. 10 could not both be correct, but this inconsistency does not force the decoded bits to be incorrect.

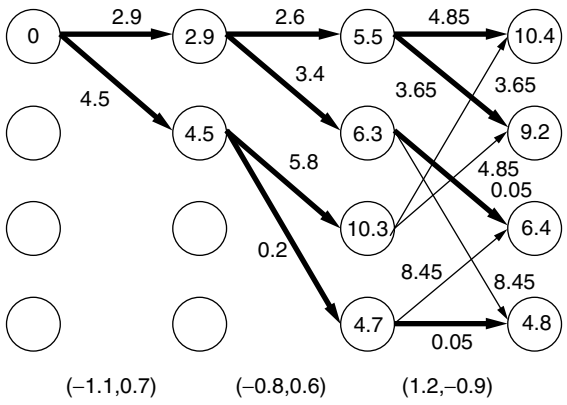


Figure 8. Illustration of soft Viterbi decoding.

4. FREE DISTANCE

The ultimate measure of a convolutional code's performance is the bit error rate (BER) or block error rate (BLER) of the code as a function of signal-to-noise ratio (SNR). However, free distance gives a good indication of convolutional code performance. The free distance of a convolutional code is the minimum distance (either Hamming or Euclidean) between two distinct valid output sequences. Unlike algebraic block codes, which are designed to have specific distance properties, good convolutional codes are identified by an exhaustive search over encoders with a given number of memory elements. Often free distance is the metric used in these searches.

For simplicity, we will restrict the discussion of free distance to free Hamming distance. For BPSK, this restriction imposes no loss of generality since the Hamming and squared Euclidean free distances are related by a constant.

4.1. Computation of Free Distance

The set of distances from a codeword to each of its neighbors is the same for all codewords. Hence, the free distance

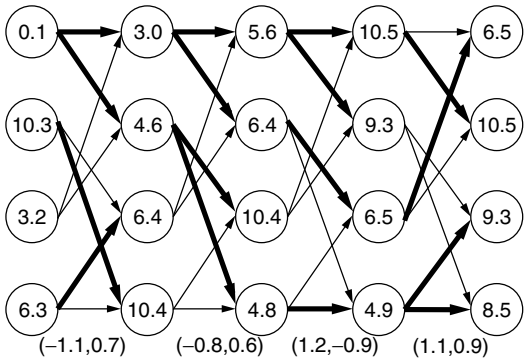


Figure 9. Steady state soft Viterbi state updates. Survivor paths are shown as thick arrows.

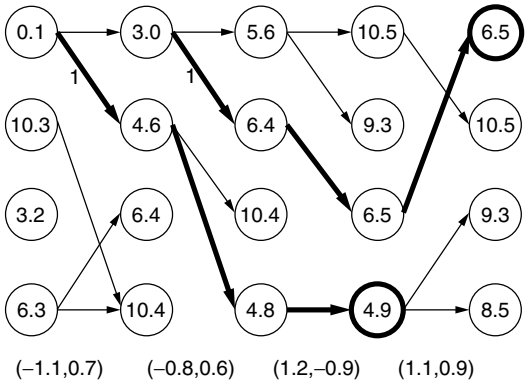


Figure 10. Finite traceback soft Viterbi decoding with a traceback depth of 3. Only the survivor paths of Fig. 9 are shown. Each traceback operation decodes only $k = 1$ bit. Thick arrows identify two such traceback paths.

is the distance from the all-zeros output sequence to its nearest-neighboring codeword. A Viterbi decoding operation with some special restrictions efficiently performs this computation. Viterbi decoding is performed on the undistorted all-zeros received sequence, but the first trellis branch associated with the correct path is disallowed. Thus prevented from decoding the correct sequence, the Viterbi algorithm identifies the nearest-neighbor sequence. Since the received sequence is noiseless, the path metric associated with the decoded sequence is the distance between that sequence and the all-zeros sequence, which is the free distance.

Figure 11 illustrates the computation of free Hamming distance using the Viterbi algorithm for the encoder described in Figs. 1, 3, and 5. The disallowed branch is shown as a dashed line. Only survivor branches are shown, and the thick branches indicate the minimum distance survivor path. Below each column is the minimum survivor path metric, which is called the *column distance*. The *free distance* is formally defined as the limit of the column distance sequence as the survivor pathlength tends to infinity. This limit is 5 in Fig. 11.

For noncatastrophic feedforward convolutional encoders, the free distance is equivalent to the minimum distance of a path that returns to the zero state. In general, the minimum distance path need not be the shortest path. For encoders with more states than the simple example of Fig. 11, there are typically several such paths having the same minimum distance. The number of minimum-distance paths is the number of nearest-neighbor output sequences. This is sometimes called the *multiplicity* of the free distance. If two codes have the same free distance, the code with the smaller multiplicity is preferred.

4.2. Analytic Decision Depth

As mentioned in Section 3.4, the specific decision depth used in finite traceback Viterbi is usually determined by simulation. However, a good estimate of the decision depth helps designers know where to begin simulations. For noncatastrophic feedforward encoders, Anderson and Balachandran [16] compute a useful lower bound on the required decision depth as a by-product of the free-distance computation.

This analytic decision depth is the pathlength at which the survivor path incident on the zero state has a path metric that is the unique minimum distance in the column. In other words, the path metric of the survivor path to the zero state is the only distance in the column equal to the column distance. In Fig. 11, this analytic decision depth is 8; after the eighth branch the path metric of the survivor path to the zero state is the only distance in the column equal to the column distance of 5. For noncatastrophic feedforward encoders, the Viterbi decoding procedure for computing free distance may be stopped when the analytic decision depth is identified. The column distance remains fixed thereafter.

When using this analytic decision depth, finite traceback decoding gives performance consistent with the first-order metrics of free distance and multiplicity. The asymptotic performance (in the limit of high SNR) is the same as maximum-likelihood Viterbi. In practice, a somewhat larger decision depth is often used to capture some additional performance at SNRs of interest by improving second-order metrics of performance (i.e., distances slightly larger than the minimum distance). For example, the analytic decision depth of the standard rate-$\frac{1}{2}$ 64-state feedforward convolutional encoder is 28, but simulation results show that a decision depth of 35 gives a noticeable performance improvement over 28. Decision depths larger than 35 give only negligible improvement.

4.3. Catastrophic Encoders

A convolutional encoder is catastrophic if a finite number of errors in the output sequence can cause an infinite number of errors in the input sequence. With such an encoder, the consequences of a decoding error can be truly catastrophic. Catastrophic encoders are certainly undesirable, and they are never used in practice. In fact, they are easily avoided because they are not minimal encoders. Hence if an encoder is catastrophic it also uses more memory elements than does a minimal equivalent encoder, which is not catastrophic.

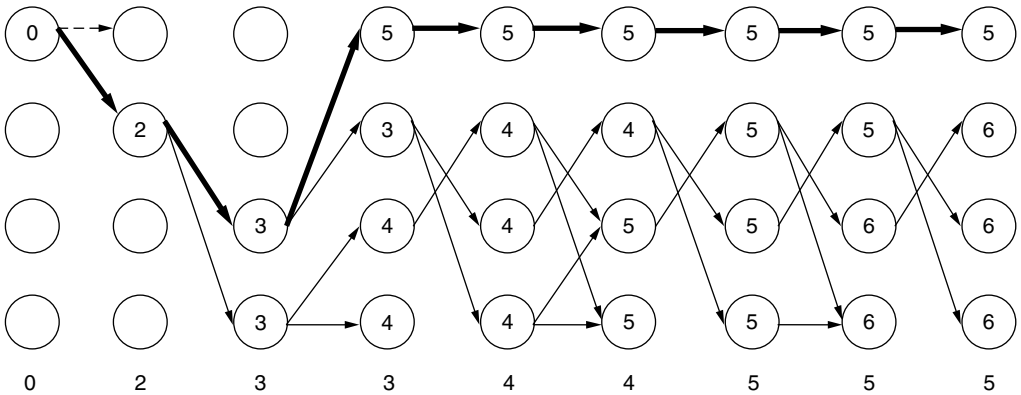


Figure 11. Application of the Viterbi algorithm to identify the free Hamming distance of the code described by Figs. 1, 3, and 5. The column distances are shown below each column. The disallowed branch is shown as a dashed line. Only survivor paths are shown, and the minimum distance path is shown with thick arrows.

An encoder is catastrophic if and only if its state diagram has a loop with zero output weight and nonzero input weight. Catastrophic encoders still have a free distance as defined by the limit of the column distance, but this free distance is seldom equal to the minimum survivor path metric to the zero state. Usually, some of the survivor path metrics for nonzero states never rise above the minimum survivor path metric to the zero state. An additional stopping rule for the Viterbi decoding computation of free distance resolves this problem: If the column distance does not change for a number of updates equal to the number of states, the free distance is equal to that column distance.

Noncatastrophic encoders may also require this additional stopping rule if they have a nontrivial zero-output-weight loop. Such a loop does not force catastrophic behavior if it is also a zero-input-weight loop. Such a situation only occurs with feedback encoders since feedforward encoders do not have loops with zero output weight and zero input weight except the trivial zero-state self-loop. In cases where this stopping criterion is required, the analytic decision depth of Anderson and Balachandran is not well defined. However, a practical place to start simulating decision depths is the pathlength at which the Viterbi computation of free distance terminates.

Because nontrivial zero-output-weight loops indicate a nonminimal encoder, their free distance is not often computed. However, there are circumstances where computation of the free distance is still interesting. As described by Fragouli et al. [17], these "encoders" arise not from poor design but indirectly when severe erasures in the channel transform a minimal encoder into a weaker, nonminimal encoder. An alternative to the additional stopping rule is simply to compute the free distance of an equivalent minimal encoder.

5. BOUNDS ON BIT ERROR RATE

As mentioned at the beginning of Section 4, BER and BLER as functions of SNR are the ultimate metrics of convolutional code performance. Monte Carlo simulation plays an important role in the characterizing BER and BLER performance. However, accurate characterization by Monte Carlo simulation at very low BER or BLER, say, less than 10^{-10}, is not computationally feasible with today's technology. However, analytic upper bounds on BER are very accurate below BER 10^{-5}. Thus, the use of bounds in conjunction with Monte Carlo simulation for high BER provides a good overall performance characterization.

5.1. The Generating Function

To facilitate the bound, a generating function or transfer function enumerates in a single closed-form expression all paths (including nonsurvivors) that return to the zero state in an infinite extension of the trellis of Fig. 11. The bound itself is analogous to the moment generating function technique for computing expectations of random variables. Figure 12 shows an altered version of the state diagram of Fig. 3 where the zero state has been split into a beginning zero state and an ending zero state. This new

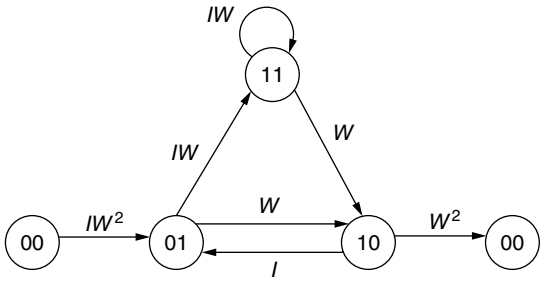


Figure 12. Split-state diagram for the encoder described by Figs. 1, 3, and 5. The exponent of W indicates the Hamming weight of the output error symbol. The exponent of I indicates the Hamming weight of the input error symbol.

diagram is called a *split-state diagram*. For the bound, only the Hamming weights of input and output symbols are needed. These Hamming weights are given as exponents of I and W, respectively. These values appear as exponents, so that when the labels along any path from the beginning zero state to the ending zero state are multiplied, the result is a single expression I^iW^w, where i is the overall input Hamming weight of the path and w is the overall output Hamming weight of the path.

Let A be the matrix of branch labels for all branches that neither begin nor end in a zero state. Each column of A represents a beginning state, and each row represents an ending state for a branch. A zero indicates no branch between the corresponding beginning and ending states. For Fig. 12

$$A = \begin{bmatrix} 0 & I & 0 \\ W & 0 & W \\ IW & 0 & IW \end{bmatrix} \tag{1}$$

Let b be the column of branch labels for all branches that begin in the zero state. For Fig. 12

$$b = \begin{bmatrix} IW^2 \\ 0 \\ 0 \end{bmatrix} \tag{2}$$

Let c be the row of branch labels for all branches that end in the zero state. For Fig. 12

$$c = \begin{bmatrix} 0 & W^2 & 0 \end{bmatrix} \tag{3}$$

The shortest path from the beginning zero state to the ending zero state has three branches; it is the path shown by thick arrows in Fig. 11. The product of the labels for this path may be computed as $cAb = IW^5$. Note that the exponent of 5 is consistent with the path metric of 5 in Fig. 11, which is also the free distance. There is one four-branch path and its label product is $cA^2b = I^2W^6$. There are two five-branch paths and their label products are $cA^3b = I^3W^7 + I^2W^6$. In general, $cA^{L-2}b$ gives the label products of all L-branch paths. Thus, the equation

$$T(W, I) = \sum_{L=3}^{\infty} cA^{L-2}b \tag{4}$$

$$= c(I - A)^{-1}b \tag{5}$$

enumerates the label products of all paths from the beginning zero state to the ending zero state. Note that I is

the input weight indeterminate in Eq. (4) but the identity matrix in Eq. (5). $T(W, I)$ is the generating function (or transfer function) of the convolutional encoder.

5.2. Union Bounds

Manipulation of $T(W, I)$ produces upper bounds on the BER for both the bit error channel and the AWGN channel. These upper bounds compute the sum over all error events e

$$\sum_e i_e P_e \tag{6}$$

where an error event is simply a path from the beginning zero state to the ending zero state. The input Hamming weight i_e associated with the error event e counts the total number of bit errors that all shifts of this error event can induce on a fixed symbol position. P_e is the probability that this path is closer to the received sequence than the transmitted (all-zeros) sequence. Note that (6) is an upper bound because P_e does not subtract probability for situations where more than one path is closer to the received sequence than the all-zeros sequence. For the bit error channel with bit error probability p,

$$\text{BER} \le \frac{1}{k} \left\{ \frac{\partial T(W, I)}{\partial I} \right\}_{I=1, W=2(p-p^2)^{1/2}} \tag{7}$$

For BPSK transmission of $\pm E_s^{1/2}$ over the AWGN channel with noise variance $N_0/2$, we obtain

$$\text{BER} \le \frac{1}{k} Q\left[\left(\frac{2 d_{\text{free}} E_s}{N_0} \right)^2 \right] e^{d_{\text{free}} E_s / N_0} \times \left\{ \frac{\partial T(W, I)}{\partial I} \right\}_{I=1, W=e^{-E_s/N_0}} \tag{8}$$

$$\le \frac{1}{2k} \left\{ \frac{\partial T(W, I)}{\partial I} \right\}_{I=1, W=e^{-E_s/N_0}} \tag{9}$$

where the tighter bound of Eq. (8) requires knowledge of the free distance d_{free}, but the looser bound of Eq. (9) does not.

6. FINAL REMARKS

Although a block code has a well-defined blocklength, convolutional codes do not. Convolutional codes are sometimes considered to have infinite blocklength, and this perspective is valuable for certain derivations, such as the derivation of union bounds on BER presented in Section 5. However, Sections 3.4 and 4.2 demonstrate that most of the useful information for decoding a particular input symbol lies within a relatively small interval of output symbols called the *decision depth*. In two important senses of blocklength, the latency required for decoding and the general strength of the code, the (properly chosen) decision depth is a good indicator the effective blocklength of a convolutional code. The decision depth of standard convolutional codes is small, certainly less than 50 for the standard rate-$\frac{1}{2}$ code with six memory elements in a single shift register.

Shannon's channel capacity theorem [18] (see also the treatise by Cover and Thomas [19]) computes the maximum rate that can be sent over a channel (or the maximum distortion that can be tolerated for a given rate). This theorem applies only as blocklength becomes infinite; in general it is not possible to achieve the performance promised by Shannon with small-blocklength codes. Indeed, convolutional code performance is hampered by their relatively small effective blocklength. For a bit error rate (BER) of 10^{-5}, they typically require about 4 dB of additional signal-to-noise ratio (SNR) beyond the Shannon requirement for error free transmission in the presence of AWGN. In contrast, for a BER of 10^{-5} Turbo codes and low-density parity-check codes, which both typically have blocklengths on the order of 10^3 or 10^4, require less than 1 dB of additional SNR beyond the Shannon requirement for error-free transmission in AWGN.

On the other hand, the performance of convolutional codes is actually quite good, given their short blocklengths. Applications such as speech transmission that require very low latency continue to employ convolutional codes because they provide excellent performance for their low latency and may be decoded with relatively low complexity. Furthermore, since Turbo codes contain convolutional encoders as constituents, a good understanding of convolutional codes remains essential even for long-blocklength applications.

BIOGRAPHY

Richard D. Wesel received both B.S. and M.S. degrees in electrical engineering from MIT in 1989 and the Ph.D. degree in Electrical Engineering from Stanford University in 1996. From 1989 to 1991 he was with AT&T Bell Laboratories, where he worked on nonintrusive measurement and adaptive correction of analog impairments in AT&T's long-distance network and the compression of facsimile transmissions in packet-switched networks. He holds patents resulting from his work in both these areas.

From July 1996 to July 2002 he was an Assistant Professor in the Electrical Engineering Department of the University of California, Los Angeles. Since July 2002 he is an Associate Professor at UCLA. His research is in communication theory with particular interests in the topics of channel coding the distributed communication. In 1998 he was awarded a National Science Foundation CAREER Award to pursue research on robust and rate-compatible coded modulation. He received an Okawa Foundation Award in 1999 for research in information and telecommunications, and he received the 2000 TRW Excellence in Teaching Award from the UCLA School of Engineering and Applied Science. Since 1999 he has been an Association Editor for the *IEEE Transactions on Communications* in the area of coding and coded modulation.

BIBLIOGRAPHY

1. P. Elias, Coding for noisy channels, *Proc. IRE Conv. Rec. part 4* 37–46 (1955) (this paper is also available in Ref. 2).
2. E. R. Berlekamp, ed., *Key Papers in the Development of Coding Theory*, IEEE Press, 1974.

3. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, 1983.

4. R. Johannesson and K. Sh. Zigangirov, *Fundamentals of Convolutional Coding*, IEEE Press, 1999.

5. G. D. Forney, Jr., Convolutional codes I: Algebraic structure, *IEEE Trans. Inform. Theory* **16**(6): 720–738 (Nov. 1970).

6. R. Johannesson and Z. Wan, A linear algebra approach to minimal convolutional encoders, *IEEE Trans. Inform. Theory* **39**(4): 1219–1233 (July 1993).

7. J. M. Wozencraft, Sequential decoding for reliable communication, *Proc. IRE Conv. Rec. part 2* 11–25 (1957).

8. R. M. Fano, A heuristic discussion of probabilistic decoding, *IEEE Trans. Inform. Theory* **9**: 64–74 (April 1963).

9. K. Sh. Zigangirov, Some sequential decoding procedures, *Probl. Peredachi Inform.* **2**: 13–25 (1966) (in Russian).

10. A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Trans. Inform. Theory* **13**: 260–269 (April 1967).

11. G. D. Forney, Jr., The Viterbi algorithm, *Proce. IEEE* **61**: 268–278 (March 1973).

12. G. D. Forney, Jr., Convolutional codes II: Maximum likelihood decoding, *Inform. Control* **25**: 222–266 (July 1974).

13. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**(2): 248–287 (March 1974).

14. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error correcting coding and decoding: Turbo-codes, *Proc. Int. Conf. Communication*, May 1993, pp. 1064–1070.

15. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, A soft-input soft-output APP module for the iterative decoding of concatenated codes, *IEEE Commun. Lett.* **1**(1): 22–24 (Jan. 1997).

16. J. B. Anderson and K. Balachandran, Decision depths of convolutional codes, *IEEE Trans. Inform. Theory* **35**(2): 455–459 (March 1989).

17. C. Fragouli, C. Komninakis, and R. D. Wesel, Minimality under periodic puncturing, *IEEE Int. Conf. Communication*, Helsinki, Finland, June 2001, pp. 300–304.

18. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (1948).

19. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

CRYPTOGRAPHY

IAN F. BLAKE
University of Toronto
Toronto, Ontario, Canada

1. INTRODUCTION

The need for secure communications has existed for centuries, and many ciphers such as substitution, transposition, and other types were in use by the Middle Ages. An excellent account of the historical development of cryptography is given in the comprehensive book of Kahn [3].

The first and most influential contribution of the modern era is the work of Claude Shannon [8], where fundamental notions of secrecy systems, including the principles of diffusion and confusion, unicity distance, and an information-theoretic approach to secrecy, were introduced. The notions of diffusion and confusion are very much in evidence in the design of the Data Encryption Standard, introduced in 1977 and used worldwide to this time. Only relatively recently has a replacement been announced, the Advanced Encryption Standard. These two block ciphers, as well as other so-called symmetric key ciphers, including stream ciphers, are considered in the next section.

It is the paper by Diffie and Hellman [1] that has ushered in the modern era of cryptography and decisively changed the landscape. It proposed the notions of asymmetric key (public key) cryptography, one-way functions, trap-door one-way functions, and digital signatures (called *one-way authentication* there) that have revolutionized secure communications in a networked world. Perhaps the most elegant incarnation of their ideas is that of RSA [6] (standing for Rivest, Shamir, and Adleman, the inventors), which includes the first realization of a digital signature. The notion of a cryptographic protocol, a sequence of steps to achieve a given purpose, arose out of these works and continues as a vital area of research.

This article overviews the subject of modern cryptography in a manner that those with a technical background will hopefully find useful, while omitting mathematical proofs.

Three excellent reference works on cryptography are those by Schneier [7], which gives a comprehensive and readable account of the subject and those by Menezes et al. [5] and Stinson [9], which give a more detailed and mathematical account that those interested in current research directions and implementation will find invaluable. Both of these last two references will be referred to liberally in this article.

The Website of the National Institute of Standards and Technology (NIST) contains an organized, authoritative, and up-to-date reference on standards, documentation and ongoing activity on the theory and practice of cryptography that is invaluable. It publishes the standards in documents referred to as *Federal Information Processing Standards* (FIPS), which invariably become de facto worldwide standards. These will be referred to throughout the article.

This article attempts to cover many of the important aspects of modern cryptography. A guide to this has been the list provided by NIST on its Website, which lists the following items in its cryptographic toolkit:

- Encryption
- Digital signatures
- Prime-number generation
- Modes of operation
- Authentication
- Random-number generation
- Secure hashing
- Key management

Each category contains references and links to standards and further materials, as appropriate. It is a reference list

for proper cryptographic practice in all of its important aspects. The list is used as a guide for this article. Authentication here is taken to include both entity authentication and data integrity.

Cryptography has among its goals to achieve confidentiality, data integrity, authentication, identification, and nonrepudiation of data transmission, storage, and transactions, and this article will outline how cryptography is used to achieve these goals. Formally, *cryptography* is the study of "secret writing," and *cryptanalysis* is the study of breaking a cryptographic technique. Together they are referred to as *cryptology*.

The following section first considers symmetric key encryption systems, including both block and stream ciphers, where the transmitter and receiver must have a common key. Before proceeding to examine public key cryptosystems, we briefly consider the mathematical problems and their complexity on which such systems are based. The main functions that public key systems are used for, including key exchange, encryption, digital signatures, and authentication techniques, are then presented. The article concludes with a brief look at prime-number and random-number generation and some indication of future directions of cryptography.

2. SYMMETRIC KEY ENCRYPTION SYSTEMS

Symmetric key encryption systems require a common key at both the transmitter and receiver. In a network environment this might mean that each pair of users requires a unique shared key, implying that each user in a community of n users is required to store $\binom{n}{2}$ keys, often a prohibitively large number. An efficient solution to this problem will be considered below.

The set of symmetric key systems is divided into block and stream systems. In a block system the input data are divided into blocks of equal length and produces equal length ciphertext blocks of output, usually of the same length as the input blocks. A stream cipher encrypts one symbol at a time using a system that produces a long sequence of symbols in some complex manner. They are typically generated by hardware, tend to be faster than block ciphers and require less complex hardware and storage. They find particular application in constrained devices such as wireless systems.

2.1. Symmetric Key Ciphers

As noted, a block cipher, represented in Fig. 1, operates on a block of plaintext P bits with a key K to produce a block of ciphertext bits, C. The decryption process uses the same key as the encryption process.

The best known and most widely used block cipher has been the Data Encryption Standard (DES). Originally introduced in 1977, it was the result of a solicitation by NIST (then the National Bureau of Standards) for an encryption method for computer data and was derived from a submission by IBM. A detailed description of its operation is given in FIPS 46 (1977) (and modifications given in FIPS 46-1, 1988 and FIPS 46-2, 1993).

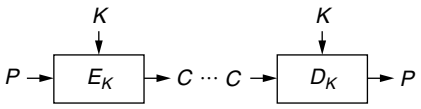


Figure 1. Basic block encryption.

The operation of DES is briefly described, with some details omitted. It operates on input plaintext data blocks of length 64 bits, producing ciphertext blocks of the same length. While its key length is 64 bits, only 56 bits are used in the algorithm itself; every 8th bit is a parity bit that is ignored in producing ciphertext. The first step of the algorithm is a fixed permutation of the block held in a register. After this step, the data are divided into left and right halves, each of 32 bits. The algorithm proceeds in 16 rounds with a typical round represented in Fig. 2.

At each of the 16 rounds, a different 48 bit subkey, K_i, is derived from the 56 bit key K in a simple and deterministic manner, not given here. The function f takes the 32 bit R_i together with the 48-bit subkey K_i to produce a 32-bit output. The block R_i is first expanded to 48 bits, by repeating certain of the bits, and XORed with the subkey K_i to produce 8 blocks of 6 bits each. Each subblock addresses one of eight so-called S boxes, which are 4×16 arrays containing integers 0–15. The first and last bits of the 6-bit subblock address the row and the middle 4 bits address a column, producing an integer representing 4 bits. Encryption concludes with the inverse of the initial permutation. The decryption process is very similar to the encryption process, with certain parts of the algorithm inverted or run backward as appropriate.

The algorithm is actually used in one of four modes (referred to as the *modes of operation*). The basic technique described is referred to as *electronic codebook* (ECB). The other modes are called *cipher feedback* (CFB), *output feedback* (OFB), and *cipher block chaining* (CBC). These modes introduce feedback and hence dependence between cipher blocks, useful in avoiding certain types of repetition attacks, among other uses. The four modes of operation can actually be used with any block cipher.

It has been observed for many years that the 56-bit key is too short to withstand a brute-force attack on current computing equipment, and indeed DES is now viewed as insecure, and its use is no longer recommended. It is possible to cascade DES ciphers, using double or triple encryption, thereby expanding the key length to 112 or

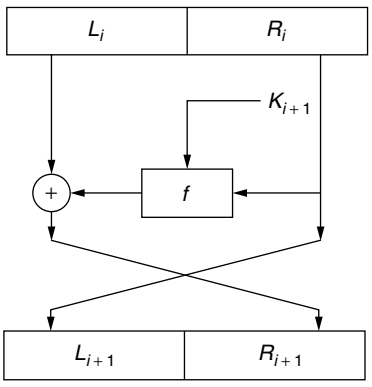


Figure 2. One round of DES.

168 bits, although here also, care must be taken. These so-called double and triple DES versions will likely be in use for many more years as they are in widespread applications in both hardware and software. DES has proved to be a remarkably successful block cipher, far outliving its planned lifetime with no inherent weaknesses discovered.

Recognizing the need for stronger encryption, NIST solicited the cryptographic community for algorithms to replace DES. Of the many submissions received, an algorithm from Belgium, Rijndael, was chosen, after careful consideration of its security, speed in both software and hardware, and suitability on a variety of platforms. The algorithm as adopted by NIST is referred to as the Advanced Encryption Standard (AES), and it is described in detail in FIPS 197 (Dec. 2001). An overview of the algorithm is given here to contrast it with the DES algorithm described above. The algorithm supports key lengths of 128, 192, and 256 bits. While the original Rijndael algorithm was designed to support block lengths 128, 192, and 256, only the data block length 128 is supported in the current standard.

The algorithm is byte- and word-oriented (4 bytes to a word) and involves two types of arithmetic. For arithmetic on bytes, the 8-bit sequence $(a_0, a_1, \ldots, a_7)$ is interpreted as the polynomial $a(x) = a_0 + a_1x + a_2x^2 + \cdots + a_7x^7$. Arithmetic in the finite field with 256 elements, $\mathbb{F}_{2^8}$, is taken as the arithmetic of polynomials of degree <8 over $\mathbb{F}_2$, modulo the irreducible polynomial $m(x) = 1 + x + x^3 + x^4 + x^8$. In particular, the inverse of a byte is defined for all nonzero bytes (the inverse of the all-zero byte, when called for in the algorithm, is taken as the all-zero byte). The second type of arithmetic involves arithmetic on polynomials with byte coefficients. The array of bytes $\{b_0, b_1, b_2, b_3\}$ is equated with the polynomial $b(x) = b_0 + b_1x + b_2x^2 + b_3x^3$. Arithmetic on such polynomials is taken modulo the (reducible) polynomial $M(x) = x^4 + 1$.

Using these arithmetics, the encryption algorithm is briefly described. Assume a key and block length of 128 bits; the algorithm for the other key sizes is easily derived from this. A *state* matrix is first defined as a 4×4 matrix with each entry a byte, and this is initially set to the input block to be encrypted, the matrix filled in by the data bytes down columns. The state matrix is continually modified at each step of the algorithm, with the final state matrix containing the output ciphertext block. The algorithm uses four basic operations (three described here and the fourth, in the next paragraph): (1) for each byte in the state matrix a *SubBytes* operation is defined as an affine transformation that replaces the byte $b = (b_0, b_1, \ldots, b_7)$ by $b' = (b'_0, b'_1, \ldots, b'_7)$, where $b' = Ab + c$, where A is a circulant matrix with first row [10001111] (as bits) and where the vector c is [11000110]; (2) a *ShiftRows* operation on the state matrix is defined as shifting row i of the matrix i positions to the left, cyclically, for rows $i = 0, 1, 2, 3$; and (3) the last operation, *MixColumns*, is defined on the columns of the state matrix by replacing a column that has a polynomial representation $a(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ (a_i a byte) and multiplying it by the (fixed) polynomial $c(x) = c_0 + c_1x + c_2x^2 + c_3x^3$ modulo $M(x)$ where, in hexadecimal notation $c_0 = 0 \times$ '02', $c_1 = 0 \times$ '01', $c_2 = 0 \times$ '01', $c_3 = 0 \times$ '03' (e.g., $c_3 = [00000011]$).

Finally the encryption process requires the derivation of a number of rounds $N_r + 1$ of a key schedule $K[i]$, $i = 0, 1, \ldots, N_r$. The number of rounds required depends on the key size and is 10, 12, and 14 depending on whether the key size is 4, 6, or 8 words in length. The precise generation of the key rounds from the original key is mechanical and is not described here. Suffice it to say that at the ith round a 4×4 matrix of bytes, $K[i]$, is generated and this matrix is added to the state matrix under byte addition (XOR), in the operation referred to as *AddRoundKey*.

With the operations described, the encryption algorithm is simply described as follows; recall that the initial state matrix is the 128 bits of the data to be encrypted fed in byte-wise down columns:

```
AddRoundKey (state, K[0])
for (i = 1 to Nr − 1) do {
   SubBytes (state);
   ShiftRows (state);
   MixColumns (state);
   AddRoundKey (state, K[i]);
   }
SubBytes (state);
ShiftRows (state);
AddRoundKey (state, K[ Nr]);
```

The encrypted output block is then the state matrix. The decryption process is similar in structure.

The AES will be the block encipherment algorithm of choice for the foreseeable future. Because of the varying block and key lengths, the modes of operation for AES are still under consideration and a *counter* mode has been added to the four standard modes mentioned previously. As with any block cipher, the last block is padded in an appropriate manner to round the data sequence out to an integral number of blocks.

There are numerous other block ciphers in use, some in the public domain and others proprietary [7]. While some may have specific advantages such as speed, use of AES will likely dominate the future of block encryption.

2.2. Stream Ciphers

From an information-theoretic point of view, the only secure cipher is a one-time pad, where purely random bits are recorded and given to the transmitter and receiver. These bits may then be XORed to the bits of the message to form the ciphertext. The plaintext can then be recovered at the receiver by XORing the ciphertexts with the one-time pad bits. Stream ciphers attempt, in some sense, to emulate this situation. More complicated operations than XORing might also be used. A general additive stream cipher is shown in Fig. 3. Here the key K may be the initial

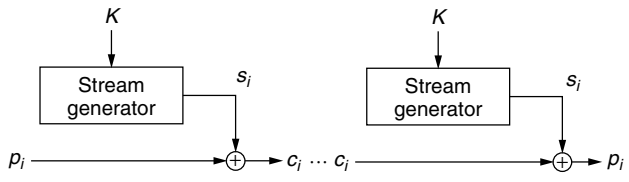


Figure 3. Configuration for an additive stream cipher.

state of the circuit used to generate the bit stream. Clearly the machines at the transmitter and receiver must be initially in the same state to generate the same sequence, allowing the plaintext to be recovered at the receiver.

There is a wealth of literature on the generation of suitable sequences. Many, if not most, of these techniques use maximum-length shift register sequences in some manner. As shown in Fig. 4, these sequences are generated by a shift register of length n, say, where the linear feedback connections are determined by a certain type of polynomial, a primitive polynomial, over the field of two elements, $\mathbb{F}_2$. Tables of such polynomials exist to quite high degrees. Clearly the sequences, coming from a deterministic circuit with a finite number of states, must be periodic. For a given shift register length, these sequences have the maximum length possible, $2^n - 1$ for a register of length n. Such sequences themselves are not secure — it is known the feedback connections of the register may be determined from knowledge of approximately $2n$ bits. However, the outputs of several such registers may be combined in a highly nonlinear Boolean function to produce a binary sequence of much greater complexity more suitable for cryptographic applications.

As noted, such stream ciphers are attractive in some applications for their high speed and relatively low circuit complexity. They have been incorporated in some standards, although many systems use proprietary generation techniques.

3. THE COMPLEXITY OF CERTAIN MATHEMATICAL PROBLEMS

Public key cryptography depends very much on the computational complexity of certain mathematical problems using the currently best known algorithms to solve them, which gives a notion of computational security. A few of the most important such problems, in terms of their use in public key cryptography, are reviewed here. Specifically these will be the problems of integer factorization, modular discrete logarithms, and modular square roots. Only the discrete logarithm problem will be considered in any detail. Informally we classify a problem as being *computationally feasible* or easy, if it is likely that a presumed attacker will have the resources to solve the system in a reasonable amount of time. Otherwise we refer to a problem as being *computationally infeasible*.

A *one-way function* f [5] is one for which it is "easy" to compute $f(x)$ for all elements in its domain X but for a randomly selected value $y \in \mathrm{Im}(f)$ it is computationally infeasible to find a value $x \in X$ such that $f(x) = y$. A *trapdoor one-way function* is a one-way function for which, given some extra information, it becomes computationally feasible to find such a value of $x \in X$, where the trapdoor information is independent of x.

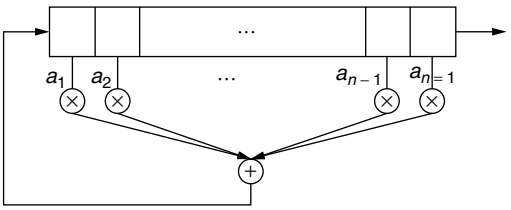


Figure 4. An LFSR for the polynomial $f(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n$.

A measure of the computational complexity of the problems that will be of interest for the problems considered, is

$$L_n(a, c) = O(\exp(c + o(1)[(\log(n))^a (\log\log(n))^{1-a}]))$$

where $O(\cdot)$ and $o(\cdot)$ are the standard complexity notation. This function is referred to as *subexponential* in $\log(n)$ since if $a = 1$ it is exponential in $\log(n)$ and if $a = 0$ it is polynomial in $\log(n)$.

The difficulty of factoring an integer plays a central role in many public key systems. It might be argued the hardest integers of a given size to factor are integers that are the product of two large primes of approximately the same size, $n = pq$, where p and q are primes on the order of $\sqrt{n}$. The most efficient factoring algorithm currently, to factor a general integer (one with no special structure), is the general number field sieve (NFS). It has a conjectured complexity [5] of $L_n\left(\frac{1}{3}, c\right)$, where $c = \left(\frac{64}{9}\right)^{1/3}$ in both time and space. The NFS continues to be developed.

The problem of determining square roots in $\mathbb{Z}_n^*$, the group of units of $\mathbb{Z}_n$, where $n = pq$ is the product of two odd primes, is formally equivalent to factoring n of such form. It can also be shown there are efficient algorithms for determining square roots modulo a prime number and for primes of the form $p \equiv 3 \pmod 4$ they are particularly so. In fact a solution to the equation $x^2 \equiv a \pmod p$ for $p \equiv 3 \pmod 4$, when a is a square $(\mathrm{mod}\ p)$, is given by [5]

$$u \equiv \pm a^{(p+1)/4} \pmod p$$

An efficient nondeterministic algorithm for finding such square roots exists for any prime. Determining square roots modulo the product of two primes is also easily accomplished if the two primes are known. For instance, if $n = pq$, p and q primes, to solve the equation $x^2 \equiv a \pmod n$, one first solves it modulo p, then modulo q and combines the solutions by use of the Chinese remainder theorem to determine the (four) solutions modulo n, assuming that the equation has solutions both modulo p and modulo q. Thus determining solutions modulo n is a simple computation if the factorization of n is known. It is surprising, then, that the problem is equivalent to factoring n when the factorization of n is not known, since the factoring problem is a known difficult problem for integers of the assumed form. This is an example of a trapdoor one-way function.

To discuss the discrete logarithm problem (DLP) in prime fields, consider the multiplicative group of the integers modulo a large prime p, $\mathbb{F}_p^*$ and let $\alpha \in \mathbb{F}_p^*$ be a primitive element, namely, an element of order $p - 1$. The discrete logarithm problem in $\mathbb{F}_p^*$ is then

$$\text{DLP: Given } \alpha,\ p,\ \text{and } y = \alpha^x \pmod p,\ \text{find } x \pmod{p-1}$$

While the most efficient algorithm currently available to solve the DLP is an adaptation of the number field sieve

algorithm mentioned previously for factoring integers, a simpler algorithm, referred to as the *index calculus method*, is briefly discussed here. In fact, similar algorithms can be applied to the discrete logarithm problem in any algebraic structure that possesses a norm. In the case of the integers modulo a prime, $\mathbb{F}_p^*$, the first phase of the algorithm considers a *factor base*, $\mathcal{D}$, consisting of all the prime numbers less than some suitably chosen bound, often several hundred thousand. It attempts to establish a sufficient number of random relations between the discrete logarithms of primes in $\mathcal{D}$. One method, for example, might be to choose random powers of the primitive element α, say, $\alpha^u \in \mathbb{F}_p^*$, for randomly chosen u, and determine whether this integer factors in $\mathcal{D}$. If it does, an equation is obtained that relates u with the discrete logs of elements in $\mathcal{D}$. Such an integer would be called *smooth* with respect to the bound on $\mathcal{D}$. If a sufficient number of such relations are found (at least $|\mathcal{D}|$), then matrix reduction techniques should determine the logarithms of all elements in $\mathcal{D}$.

In the second phase of the algorithm, one attempts to find the log of the given element y by multiplying it successively by a random power of α (to yield, say, $w = \alpha^v y \in \mathbb{F}_p^*$) and determine whether it factors entirely over the factor base. If it does, the logarithm of w and hence of y can be found.

This simple idea has turned out to be a powerful one and is the basis of most of the current most effective algorithms for both the DLP and integer factoring. It turns out that most algorithms designed to factor integers can be modified to find discrete logarithms. The current most efficient algorithm to determine logarithms is the number field sieve and it has a conjectured complexity in both time and space of $L_n\left(\frac{1}{3}, c\right)$, where $c \approx \left(\frac{64}{9}\right)^{1/3} \approx 1.923$ the same as the integer factorization problem.

The discrete logarithm problem can take place in any Abelian group and typically, for reasons of security and efficiency, it is possible to reduce it to being in a cyclic group of prime order.

There is an additive version of the DLP, which is briefly described since it is finding increasing favor for cryptography on small devices. The set of solutions to an elliptic curve can be shown to be an Abelian group under point addition. In the case of a curve over a prime field $\mathbb{F}_p$ such a curve can, without loss of generality, be taken to be of the form

$$y^2 = x^3 + ax + b$$

and the Abelian group consists of the points (x, y) satisfying this equation, together with a point at infinity. The operation of point addition is very natural and is discussed here only for the case of $\mathbb{F}_p$.

Another type of finite field, one of characteristic two, is also popular for cryptographic applications and the equations of the curve and for point addition are slightly different for that case. The addition stems from the observation that if a straight line intersects the curve in at least two points, there is a unique third point of intersection. The situation is depicted in the Fig. 5.

For the last equation above, let $E_{a,b}(\mathbb{F}_p) = \{(x, y) \mid y^2 = x^3 + ax + b\}$ and let $P_1 = (x_1, y_1)$, $-P_1 = (x_1, -y_1)$ and $P_3 = (x_3, y_3) = P_1 + P_2$. If $P_1 \neq P_2$ define $\lambda = (y_2 - y_1)/(x_2 - x_1)$,

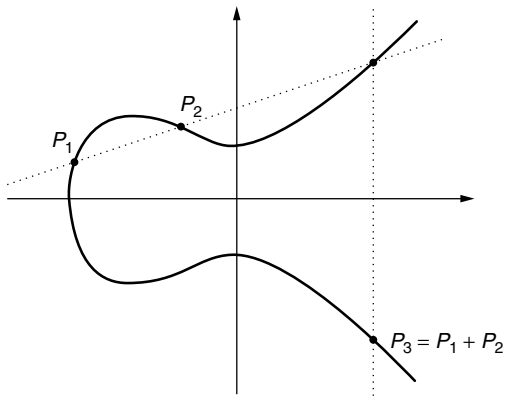


Figure 5. Elliptic curve addition.

$x_1 \neq x_2$, then $x_3 = \lambda^2 - x_1 - x_2$ and $y_3 = (x_1 - x_3)\lambda - y_1$. Similar equations hold if $P_1 = P_2$, in which case the line through the points is a tangent and the addition is referred to as a *point doubling*.

With this notion of point addition, for a given point $P = (x, y)$ one defines a general point multiple $k \cdot P = Q$ and the DLP in this setting is, given Q and P (as well as p and the curve equation), find k. This will be referred to as the ECDLP. The interesting aspect of this problem, for carefully chosen curves, is that no notion of smoothness has been found and hence we have been unable to formulate an index calculus attack on such a problem. As a consequence, the complexity of this ECDLP problem (on carefully chosen curves) is $O(\sqrt{p})$, which is considerably greater than the equivalent DLP in $\mathbb{F}_p^*$. Hence the ECDLP on an elliptic curve over $\mathbb{F}_p$ appears to be a considerably more difficult problem than the DLP in $\mathbb{F}_p^*$ for the same size prime. An implication of this is that the ECDLP on an elliptic curve over $\mathbb{F}_p$ can be used with a much smaller field size, for the same level of security as the DLP in $\mathbb{F}_p^*$.

The subject of elliptic curves has been of interest to mathematicians for over a century and is a deep and fascinating area of research. To use elliptic curves for cryptography it is necessary to know the order of the cyclic subgroup to be used. This requires determining the exact number of points on the elliptic curve and the factorization of this number, from which the order of the subgroup can be found. An important result in this direction is the Hasse–Weil theorem which says that $\#E_{a,b}(\mathbb{F}_p) \in (p + 1 \pm 2\sqrt{p})$. Once the order of the group is known, it is often straightforward to determine a generating point for the cyclic subgroup to be used.

In the following sections the three problems noted here will be used in various cryptographic protocols. The next section discusses the important problem of key exchange.

4. KEY EXCHANGE

In a network environment with a large user community, the use of a symmetric key cryptosystem such as AES described earlier, introduces the problem of establishing unique common keys between each pair of users. A solution to this problem, now in almost universal use in one form or another, is the Diffie–Hellman (DH) key exchange introduced in their 1976 paper [1], based on the discrete

logarithm problem. A version of the protocol in $\mathbb{F}_p$ is as follows. Given public information p, a prime, and $\alpha \in \mathbb{F}_p$, a primitive element, user A generates a random $a \in [1, p-1]$ and computes α^a which is sent to user B. User B chooses $b \in [0, p-1]$ at random and sends α^b to A. Both parties are now able to compute α^{ab}, which is now used to form the common key. Typically this might be used to encrypt larger messages with a faster symmetric key cryptosystem.

An eavesdropper observing the information transmissions in this protocol sees α^a, α^b and would like to compute α^{ab} and this is referred to as the *Diffie–Hellman problem* (DHP). Certainly if one is able to compute discrete logarithms in $\mathbb{F}_p$, then this can be accomplished. The more general question as to whether the DHP and DLP are computationally equivalent problems remains open, although some progress has been reported in special cases.

It is clear that the DH protocol can take place in any group, for example, the additive group of the points on an elliptic curve. In this structure the DH protocol (ECDH) is, for a fixed curve and point P on the curve of known order, all public information, user A transmits aP to A and receives bP from which the common key abP is computed by both users. The previous comments on the security of the system, specifically, the complexity of the ECDLP problem, as compared to the DLP in $\mathbb{F}_p^*$, apply to this case as well. For security reasons, the cyclic group is always taken to have prime order since any small factors of the group order tend to weaken the system for the given bit length.

A weakness of the DH key exchange protocol is the *person-in-the-middle attack*, where a user E inserts herself in the middle; users A and B believe they are exchanging keys with each other but in fact are each exchanging keys with E in the middle. The remedy for such an attack is to use an *authenticated key exchange*, where some public and private information as exists with a public key system, discussed in the next section, is incorporated in the key exchange to authenticate the user identity.

The key management problem extends far beyond the simple key exchange protocol noted above. In a large user community of users, the use, distribution, and management of keys is a critical aspect of system operation. Questions of the privileges associated with keys, the revocation and updating of keys, the generation and storage of keys, and so on are all vital aspects of the problem.

5. ASYMMETRIC OR PUBLIC KEY CRYPTOGRAPHY

The notions of public key cryptography, where the encrypting and decrypting keys are different, were introduced in the landmark paper of Diffie and Hellman mentioned previously. The paper included applications as to what might be achieved with such an asymmetric system under certain assumptions. These included the notion of one-way authentication and digital signatures as well as the use of the discrete logarithm for DH key exchange discussed earlier.

Apart from the results in that paper, perhaps the most important public key system is the RSA system. To introduce this system let $\mathbb{Z}_n$ denote the set of integers modulo the positive integer n, $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$. Let $\mathbb{Z}_n^*$ denote the set of invertible elements in $\mathbb{Z}_n$, namely, those elements in $\mathbb{Z}_n$ that have a gcd with n of 1:

$$\mathbb{Z}_n^* = \{x \in \mathbb{Z}_n \mid \gcd(x, n) = 1\}$$

Such a set forms a multiplicative group, the group of units of $\mathbb{Z}_n$ and has order $\phi(n)$, where $\phi(\cdot)$ is the Euler phi function. For an arbitrary positive integer n with prime factorization $n = \prod_i p_i^{e_i}$, e_i a positive integer, $\phi(n) = \prod_i p_i^{e_i - 1}(p_i - 1)$. Euler's theorem then states that:

$$a^{\phi(n)} \equiv 1 \pmod{n}, \qquad a \in \mathbb{Z}_n^*$$

It is actually true that $a^r \equiv a^s \pmod{n}$ for all $r \equiv s \pmod{\phi(n)}$ and all $a \in \mathbb{Z}_n$. The case of interest for the RSA system is when $n = pq$, where p and q are large primes on the order of $\sqrt{n}$. An encryption exponent $e \in \mathbb{Z}_{\phi(n)}^*$ is chosen and a decryption exponent d such that $ed \equiv 1 \pmod{\phi(n)}$ is computed. A message m is interpreted as an element in $\mathbb{Z}_n$ via a suitable embedding from the message space to the integers. The encryption of m is then

$$c \equiv m^e \pmod{n}$$

By Euler's theorem the decryption of c is

$$c^d \equiv m^{ed} \equiv m^{k\phi(n)+1} \equiv m \pmod{n}$$

for some integer k, and the message m is recovered. To use such a system, user A places in a public directory the information n_A, e_A, their RSA modulus and encrypting modulus, respectively. The user retains as secret information the factorization of n_A, p_A, q_A, and the decryption exponent d_A. Another user wishing to send a message to user A, sends $c_A \equiv m^{e_A} \pmod{n_A}$. As only user A knows the decryption exponent d_A, only they are able to decrypt the message.

The security of this system is believed, in general, to be equivalent to the difficulty in factoring the modulus n. Certainly if the decryption exponent can be found, the system is broken. If one is able to find $\phi(n)$, one can determine the decryption exponent and, indeed, factor the modulus. The system is an example of a trapdoor one-way function in that determining m from c is computationally infeasible for a properly chosen modulus and encryption exponent, but knowing the factorization of the modulus makes it easy.

Another public key encryption system based on the difficulty of finding square-roots modulus a composite number is the Rabin encryption scheme. Again, assume that user A publishes a modulus $n_A = p_A q_A, p_A, q_A$, primes, where n_A is placed in a public directory with user A's identity and the factorization is kept secret. Another user wishing to send $m \in \mathbb{Z}_n^*$ to A sends $c \equiv m^2 \pmod{n_A}$. User A, on receiving c, computes the square roots of c modulo the factors p_A and q_A, which is a simple task. The (four) square roots of $c \pmod{n}$ are then found by computing the square roots modulo p_A and modulo q_A and combining them via the CRT noted earlier. The correct square root is then found by context (it is unlikely that more than one of the square roots makes sense), or else some prearranged type

of message padding is used prior to encryption. As noted previously, the operation of finding modular square roots is a trapdoor one-way function in that finding square roots modulo n, a product of two primes, is formally equivalent to factoring if the factorization is not known, yet simple if the factorization is known.

The final public key encryption system discussed is that of ElGamal [2], based on the DLP. The DLP is a one-way function with no trapdoor and a little more work is needed to make a cryptosystem (i.e., an encryption system) out of it. User A's public key consists of a prime p_A, a primitive element α (or else a generator of a prime order cyclic subgroup) in $\mathbb{F}_{pA}$, and an element $y = \alpha^a \pmod{p_A}$ for a secret exponent a. For another user to send to user A an encrypted message m, assuming an element of $\mathbb{F}_p^*$, they first choose a random (one-time) integer $x \in [1, p-2]$ (where $[1, p-2]$ denotes the range of integers $\{1, 2, \ldots, p-2\}$) and compute $\gamma \equiv \alpha^k \pmod{p_A}$ and $\delta \equiv my^k \pmod{p_A}$. The ciphertext is then the pair of elements (γ, δ). Since user A knows the secret exponent a, they are able to compute $\gamma^{-a} \equiv \alpha^{-ak} \pmod{p_A}$ and then $\delta\alpha^{-ak} \equiv m \pmod{p_A}$. Notice the system has message expansion since two elements are passed to convey one message element.

These public key systems have uses well beyond the encryption functions outlined here. For example, while the Diffie–Hellman key exchange of the previous section requires two passes, with the notion of a public directory containing user public keys, a user can transport the key to be used in a symmetric system by encrypting with the intended users public key in a single pass. Other, more symmetric, protocols use the notion of public keys for authenticated key exchange to defeat the person-in-the-middle attack noted earlier. The public key systems also have critical properties that can be used for more general authentication and the establishment of trust in a network.

6. DIGITAL SIGNATURES AND HASH FUNCTIONS

The notion of a written signature is central to transactions of all kinds, especially to financial ones. Such signatures have characteristics in terms of forgeability and detection of forgeries that are central to many legalities. The idea of a digital signature is, in a sense, even stronger in that it is composed by a machine and constructing a forgery requires the solution of a computationally infeasible problem. It has obtained a legal standing in several countries in terms of acceptability in court actions.

To illustrate the notion of a digital signature, suppose that a message m can be embedded in some unique way into $\mathbb{Z}_n$. Let n_A, e_A, and d_A be the public and private parameters of user A's RSA system. For user A to sign the message m, the quantity $s_A(m) \equiv m^{d_A} \pmod{n_A}$ is computed as A's signature for m. Any other user may verify the signature by computing $s_A(m)^{e_A} \equiv m \pmod{n_A}$, thereby recovering the message at the same time. It is clear that only user A could have created the signature and that the signature is bound with the message m; thus, for example, the signature is invalid when associated with any other message. This is referred to as a *signature with message recovery* scheme since the message is recovered in the process of authenticating the signature.

This system has two disadvantages in that it required the message to be embedded into $\mathbb{Z}_n^*$, and so had to be relatively short to achieve this. In addition, the system described above is susceptible to an *existential forgery attack*, meaning that it is possible for an adversary to produce another bit stream, which is likely not intelligible as a real message, that has the same signature. To overcome both of these problems the notion of hash functions is introduced. A *hash function* is a mapping from binary strings of arbitrary length to strings of fixed length (usually much shorter than the original length, hence providing compression)

$$h: \{0,1\}^* \to \{0,1\}^n$$

$$x \mapsto h(x)$$

with certain properties. The function should be computationally efficient to compute on long strings. It should also have the property of being one-way in the sense that for any given hash value $y = h$ it should be computationally infeasible to determine any x, such that $y = h(x)$, since that would allow forgeries, for example, when the hash function is used for signature generation. This property is often referred to as *preimage resistance*. Furthermore it should have the property that, given x and $y = h(x)$, it should be computationally infeasible to find a second value x' so that $y = h(x) = h(x')$, sometimes referred to as *second preimage resistance*. A slightly different notion is that h is said to be *collision-resistant* if it is computationally infeasible to find any two inputs x, x' that hash to the same value.

The design of hash functions with these properties requires careful attention. There are many such hash functions, but perhaps the most widely used is the SHA-1 (secure hash algorithm) (FIPS 180-1, April 1995), a modification of the earlier FIPS 180. This produces a hash of length 160 bits and is mandated for use in the DSA (digital signature algorithm, to be discussed below). A new family of hash functions is currently under consideration, to be called SHA-256, SHA-384, and SHA-512 with the suffix indicating hash size, to be FIPS 180-2, currently in draft form. These increased size hash functions will be used in an updated DSA in the future.

To return to digital signatures, for user A to sign a long message m using RSA, one might first hash the message to produce $h(m)$, embed this hash value into the integers modulo n_A, and then sign by using as the signature $h(m)^{d_A} \pmod{n_A}$. In such a system, the message is no longer recovered when verifying the signature and so must be transmitted, usually in the clear, along with the message; that is the pair $(m, s_A(h(m)))$ is produced as the signed message by A and verified by any other user. Such a scheme is referred to [5] as a *signature with appendix*, where the signature is an appendix to the message.

Other signature schemes are also of importance. In particular the use of the DLP to produce signatures originated with the work of El Gamal [2]. Since the DLP is simply a one-way function with no trapdoor, more work is required to achieve a signature. In this scheme user A generates a large prime p and finds an α that generates $\mathbb{F}_p^*$. A random integer a is chosen, $y = \alpha^a \pmod{p}$ computed,

and the information (p, α, y) made public with a kept secret. To sign a message m, a random integer k is chosen in $[1, p-2]$ such that $\gcd(k, p-1) = 1$ and let k^{-1} be its inverse modulo $p-1$. The element $r = \alpha^k \in \mathbb{F}_p^*$ is formed and the integer $s \equiv (h(m) - ar)k^{-1} \pmod{p-1}$ computed. The signature for the message m is then the pair (r, s). The scheme is clearly a signature with appendix since it requires the message m to be transmitted in order for the signature to be verified.

To verify the signature, a user retrieves the public information of user A (p, α, y) from the public directory and, using the signature (r, s) and the message m computes

$$u \equiv y^r r^s \equiv \alpha^{ar}\alpha^{ks} \equiv \alpha^{ar+k(h(m)-ar)k^{-1}} \equiv \alpha^{h(m)} \pmod{p}.$$

The hash value $h(m)$ is computed independently and if $v \equiv \alpha^{h(m)}$ is equal to u, the signature is accepted.

This scheme is the forerunner of the Digital Signature Algorithm (Standard) (FIPS 186-2, Feb. 2000) to be described now. For the DSA each user chooses a prime number p with a number of bits $512 + 64\ell$, where ℓ is an integer between 1 and 8. A second prime q with 160 bits is chosen so that $q \mid p-1$ and α is chosen as a generator of the unique subgroup $\mathcal{G}$ of order q in $\mathbb{F}_p^*$. An integer a is chosen and $y = \alpha^a \pmod{p}$. User A's public information is then (p,q,α,y), and secret information is the integer a. From this point the DSA is very similar to the El Gamal signature scheme. A random integer k is chosen in $[1, q-1]$ and $r \equiv (\alpha^k \pmod{p}) \pmod{q}$; thus the element is first taken modulo p and then modulo q. The signature for message m is then the pair (r, s), where $s \equiv k^{-1}(h(m) + ar) \pmod{q}$. The verification is very much as with El Gamal. The point of this system, first suggested by Schnorr, is that even though the cyclic group in which the signature takes place, $\mathcal{G}$ is of order q, the arithmetic in the subgroup is in $\mathbb{F}_p^*$; that is, it is modulo p arithmetic. Any attempt to break this system would also take place in modulo p arithmetic, which is more difficult since p is very much larger than q.

There is also a version of the DSA, in the FIPS 186-2, which uses elliptic curves, referred to as ECDSA. In this system, restricting attention to curves over $\mathbb{F}_p$ as before, let E be an elliptic curve and $P = (x, y)$ a point of prime order q with both the curve parameters, where point P and q represent public information. User A chooses a random integer $a \in [1, q-1]$ and computes $Q = aP$, which is user A's public key; a is maintained secret. To sign a message m a random integer k in $[1, q-1]$ is chosen and $kP = (x_1, y_1)$ computed. Let $x \equiv x_1 \pmod{q}$, recalling that $x_1 \in \mathbb{F}_p$. The signature for the message m is then (r, s), where $s \equiv k^{-1}(h(m) + xr) \pmod{q}$ and if either r or s are 0, the procedure is run again.

To verify the signature, the user obtains A's public key Q and from the accompanying cleartext message m computes $h(m)$. With r, s and $h(m)$ as integers modulo q, the quantities $u \equiv h(m)s^{-1} \pmod{q}$ and $v \equiv rs^{-1} \pmod{q}$ are computed and the point multiple $uP + vQ = s^{-1}(h(m) + rx)P = k \cdot P = (x_2, y_2)$ found. If $x_2 \equiv r \pmod{q}$, the signature is accepted.

The secret keys a and k in the El Gamal, DSA, and ECDSA schemes are referred to as "static keys" and "ephemeral keys," respectively. In particular it is important that the ephemeral keys k be chosen independently and randomly for each message to be signed.

7. AUTHENTICATION AND IDENTIFICATION

The problem of establishing trust in a network environment is central to the application of public key cryptography. The trust includes not only the integrity of a datastream but also the identity of the person communicating. A very large number of techniques are available to address these and similar problems. This section is a brief introduction to some of these.

A simple *manipulation detection code* (MDC) is typically a short appendix to the message formed by an unkeyed hash function. Its only purpose is to detect whether any changes have occurred in the message portion of the transmission, that is, to determine message integrity. If any changes have occurred then with very high probability the MDC computed at the receiver will differ from the appendix attached. A *message authentication code* (MAC) is similar except that the hash function used is keyed, thereby allowing for the verification of the sender, since it is assumed the sender and receiver are in possession of a common key and that it is computationally infeasible for anyone without the key to compute the hash appendix.

The constructions of either type of code begins with a hash function, such as SHA-1 mentioned earlier. A MAC, however, requires the incorporation of a key in some manner and experience has shown that this must be done very carefully — several such hash functions have been broken when subtle faults in their construction were found. A particular type of MAC is the *keyed hash MAC* or HMAC, currently in draft for a FIPS. It recommends the following construction, using any FIPS-approved hash function. For an input block size of B bytes to the hash function (e.g., for SHA-1 $B = 64$), construct an *ipad* of B bytes consisting of the byte (in hexadecimal) of $0 \times$ '36' and an *opad* consisting of the byte $0 \times$ '5c' each repeated B times. If K is the secret shared key, let K_0 be K with zeros appended to form a B byte key. The recommended formation of the HMAC is then the leftmost t bytes of the hash value

$$H((K_0 \oplus opad) \| H((K_0 \oplus ipad) \| text))$$

where $\|$ indicates catenation.

The subject of MACs is much wider than touched on here with many such constructions and concepts used. However, most of them share common features with the above.

Most approaches to the establishment of trust in a network involve public key concepts and the use of a *central authority* (CA) [these are also referred to as a *trusted third party* (TTP) or *trusted authority* (TA)]. The idea here is for the CA to have a public and a private key, with the public key known to all subscribers of the network. The CA establishes in some secure manner the identity of a user and their public key and binds the two together by encrypting the pair with the CA's private key. This is a simple example of a certificate — typically such certificates include other information such as level of authority

given to the individual and time limit for validity of the certificate. The crucial point here is that the CA has verified the information relating the identity of the individual and their public key and bound it together with the CA's private key, as for a signature. Anyone who has received the certificate can verify the information by use of the CA's public key and can assume that the information is as reliable as the CA. The construction of certificates and their maintenance and issuance to clients is the idea behind a *public key infrastructure* (PKI). A client of the PKI may request a certificate for any user in the system thereby assuring identity.

Many cryptographic protocols rely on the existence of a CA, although sometimes the requirements of the protocol are less than the need for a full CA. As an example, consider the Diffie–Hellman key exchange described earlier. It was noted it is susceptible to the person-in-the-middle attack where a third party injects themselves in between users A and B, who end up communicating with each other through the third party, C, who controls the flow of information. A remedy for this situation is the use of a CA where each user obtains sufficient information from the CA and the common key is established using both private and public information of both users. Such a scheme is termed an authenticated key exchange protocol, noted earlier.

Identification protocols [5] usually involve the notions of what a person knows [e.g., a PIN (personal identification number) for an ATM card], what they have (the ATM card with their account number and name on it) and a physical attribute (such as fingerprint, ocular iris characteristics, facial features, and even DNA). Two identification protocols are briefly described here for the interesting features they introduce. Both require the use of a CA although not explicitly the notions of certificates.

In the Schnorr identification protocol ([5,9]), the CA chooses a large prime q such that q is a large divisor of $p-1$ and α is an element in $\mathbb{Z}_p$ of order q. The CA has a public and private key, which allows the creation of signatures for information, $s_{CA}(\cdot)$, and the creation of certificates, $C_{CA}(\cdot)$. Each user A chooses a random integer $a \in [1, q-1]$, a private key, and computes $v \equiv \alpha^{-a} \pmod{p}$ and v is sent to the CA, who creates the certificate $C_{CA} = (I(A), v, s_{CA}(I(A), v))$. For user A to identify his/herself to B, A chooses a random number $k \in [1, q-1]$ and sends $x \equiv \alpha^k \pmod{p}$ to B, along with his/her certificate. User B verifies the certificate, which has bound the public key of A, v, to his/her identity. User B then chooses a random integer $r \in [1, 2^t]$ for some suitably large integer t, say 2^{40} (t is referred to as the security parameter of the scheme), and sends it to user A. User A sends back $y \equiv k + ar \pmod{q}$, which is sent to B. User B then verifies that $\gamma \equiv \alpha^y v^r \pmod{p}$. The protocol uses, in an essential way, the notion of certificates and a CA to establish identity in a reliable manner.

The Fiat–Shamir identification protocol introduces the notions of a *zero-knowledge proof* and a *statistical, interactive proof* which consists of multiple rounds of a three-pass challenge–response protocol. It operates as follows (basic idea only). For user A to prove his/her identity to user B, the CA first chooses the (secret) primes p and q and publishes the modulus $n = pq$. Each user chooses a secret s relatively prime to n, $s \in [1, n-1]$ and registers $v \equiv s^2 \pmod{n}$ with the CA as its public key. For A to identify his/herself to B, the following three steps are performed (and repeated t times):

$$A \to B: x \equiv r^2 \pmod{n}$$

$$A \leftarrow B: e \in \{0, 1\}$$

$$A \to B: y \equiv rs^e \pmod{n}$$

The purpose of user B choosing e in the second step of the protocol is to prevent cheating by A in the sense that if user B always chose $e = 1$, then A could compute $x \equiv r^2/v$ and answer the challenge $e = 1$ with the correct answer $y \equiv r \pmod{n}$; thus, any user knowing s can always answer either challenge while another user can only answer the one question — hence an impostor has a probability of $\frac{1}{2}$ of answering a given question. Repeating this basic protocol t times reduces the probability of success by the impostor to less than $(\frac{1}{2})^t$.

This basic protocol, which can be made much more efficient [e.g., 5,9], illustrates the two notions of a statistical interactive proof and a zero-knowledge proof. The term *zero knowledge* refers to the fact that the protocol reveals no knowledge about the factorization of the modulus n or the value of a users secret key s except for the fact that the public modulus v is in fact a square modulo n.

Many more sophisticated zero-knowledge protocols exist with the preceding example a small introduction to the area.

8. PRIME-NUMBER GENERATION AND PSEUDORANDOMNESS

Much of public key cryptography depends on the difficulty of certain number theoretic problems, such as factoring, discrete logarithms, and taking modular square roots. The setup procedure for these problems invariably involves the generation of large primes, often of several hundred digits. Many standards suggest the use of probabilistic methods to achieve this, and perhaps the most commonly used of these techniques and the most efficient is the *Miller–Rabin* method, which is briefly described [e.g., 5].

In this method, as with many others, an integer n is chosen at random and tested for primality. By the *prime-number theorem*, if $\pi(x)$ is the number of primes less than x, then for large values of x this quantity is well approximated by $x/\ln x$. This gives an estimate of the number of trials that must be made to locate a prime; thus, in the neighborhood of n the average spacing between primes is approximately $\ln(n)$. To find a *probable prime* choose an integer n at random. It is often convenient to trial-divide n for a stored table of small primes to avoid work on integers that are thus easily dismissed. Let $n - 1 = 2^s r$, where $2 \nmid r$. Let a be a randomly chosen integer in $[1, n-2]$ and compute $y \equiv a^r \pmod{n}$. If $y \equiv \pm 1 \pmod{n}$, declare n to be prime. Otherwise if $y \not\equiv \pm 1 \pmod{n}$, then successively square $y \pmod{n}$ up to $s-1$ times. If the result of the squaring never results in $y \equiv \pm 1 \pmod{n}$, then n is declared composite. The procedure rests on the observation that if p is a prime, then for $p - 1 = 2^s r$, r odd, either $a^r \equiv 1 \pmod{n}$ or $a^{2^j r} \equiv -1 \pmod{n}$ for some $j \in [1, s-1]$.

An integer a that fails the test proves conclusively that the integer n is composite and is called a "witness" to the compositeness of n. It can be shown that the probability the test declares a composite number to be prime on any given trial is less than $\frac{1}{4}$. If the test is run t times the probability the test always declares a composite number a prime is less than $(1/4)^t$. Thus for $t = 40$, this probability is $1/2^{80}$, which is usually the recommended level of certainty in standards that suggest the use of this technique, for example, the DSA (FIPS 186-2). Primes produced with this test are termed *probable primes*.

The Miller–Rabin test produces a probably prime with a probability of failing so low that it is regarded as reliable and finds wide use in practice. Maurer [e.g., 5] has devised a test, however, that produces a *provable* prime in almost the time it takes to run one iteration of the Miller–Rabin test. It is based on a modification of Pocklington's theorem of computational number theory. While Maurer's technique on provable primes removes the uncertainty of primality inherent in the Miller–Rabin test, this latter test is still widely used, and with the number of iterations proposed for this test, it yields very acceptable results.

Another frequently needed facility in any cryptographic application is the ability to generate random or pseudorandom numbers. These are used to initialize many functions. Indeed, the security of a system might often be equated to the degree of uncertainty in the random seed. While pure random generators are to be preferred, these are usually only found in hardware devices and chips where the outputs of so-called noisy diodes are often used. Such systems invariably include self-checking devices to ensure that the system has not failed and is still producing sequences that pass recognized statistical tests. In software, it is desirable to use as many different sources of randomness available, such as time between keystrokes, and to put the catenation of such information through a hash function to produce a random seed for a random-number generator. This is still often the weakest point of security of the system. Many systems in use are proprietary, making an assessment of their security difficult.

A commonly used technique to produce pseudorandom numbers is the linear congruential technique. Typically these produce sequences of numbers from an initial seed x_0, often kept secret, with the recursion

$$x_n \equiv ax_{n-1} + b \pmod{m}$$

where a and b are fixed parameters and m is a modulus. Sequences derived from such a system are in fact predictable in that given a part of the sequence, the parameters can often be derived and the entire sequence generated. However, variations on these sequences are sometimes used. Typically it is required that the seed value be sufficiently large that an exhaustive search over all possibilities be infeasible. In addition it is required that the sequence pass statistical properties of randomness in that failure to do so may lead to ways of breaking the sequence.

Other pseudorandom bit generators can be formulated that depend on the use of one-way functions, for example, the Blum–Blum–Shub generator, which depends on modular square roots for its security. While very secure, such generators tend to be very inefficient.

9. COMMENTS

Current cryptographic research goes beyond the standard techniques discussed here and is concerned with applications that can be achieved with cryptographic techniques. A few examples of these are given. One such area is that of zero-knowledge proofs, noted briefly previously, where one user proves to a verifier that they are in possession of knowledge and proves to another that this is so without revealing any information. Protocols exist showing, for example, that a prover knows the discrete logarithm of a given element $y \in \mathbb{F}_p^*$ to some known base α, p also known, without revealing what the logarithm is. Similarly, it is possible to prove to a verifier that a given RSA modulus is indeed a product of two primes without revealing the factorization. Another such protocol shows how to compute an RSA modulus and an encryption exponent, among a number of parties without any of the parties knowing the factorization. At the same time each party obtains a share of the decryption exponent, which must be combined with other user shares to either create a signature or decrypt an encryption. Such a protocol is an example of a secure multiparty distributed computation, and many other such examples exist.

Electronic cash is a payment technique whereby a user is able to spend electronic cash obtained from a bank and spent at a store. The protocol preserves anonymity from the bank as to how the e-cash was spent and prevents double-spending of it by either the individual or the merchant. It involves the notion of a blind signature that also arises in the context of electronic voting. A distributed voting scheme has among its requirements the anonymity, the ability for any participant to verify the tally and the ability to ensure that no one votes more than once.

An example of the key exclusion problem is where a pay-TV provider encrypts a movie and provides sufficient information for those users who paid to decrypt the movie but those who did not pay cannot. Of course, the set of users who view each movie changes frequently, and the challenge is to derive an efficient scheme to achieve this. There is also a problem of "traitor tracing," whereby users who collaborate to form new valid keys can be detected and identified. The research on protocols shows a power and elegance that can be exploited to achieve interesting goals.

Perhaps the single topic of greatest interest to public key cryptography is that of quantum computation. While a description of the idea of a quantum computer is beyond the scope of this article, a few of the implications of the existence of such a computer might be mentioned, and for this the article by Gottesman and Lo [4] is used as a guide. The state of a quantum computer, rather than being a deterministic state as in a classical computer, is actually a superposition of exponentially many basis states, and each of these corresponds to a state of a classical computer of the same size. The result of this is that such a computer would take a very small amount of time to do tasks that are not possible to contemplate on a classical computer. In a celebrated result, Peter Shor of AT&T Laboratories has shown that the two problems of central interest in

cryptography, integer factoring and discrete logarithms, have a complexity in the quantum computing model that is polynomial in the integer lengths, rather than the subexponential complexity, $L(\frac{1}{3}, c)$, mentioned earlier, for the conventional computing model. In a similarly spectacular result, Grover of Bell Laboratories (Lucent Technologies), has shown that a searching algorithm to find one of N items has a classical complexity of order N but a quantum algorithm of complexity of order $\sqrt{N}$. This fact could, for example, be used to search for keys for a block cipher. While a quantum computer has not yet been constructed, informed opinion considers them likely and indeed several corporations are in the process of attempting to construct them. The implications for cryptography are quite clear. Without the notions of one-way functions, the field would have to reconstitute itself in a very different direction.

In spite of the results noted above, it has also been shown that quantum key distribution using quantum systems is entirely practical, and indeed such systems have been successfully constructed and demonstrated by several research groups around the world. While such systems have been shown to be theoretically secure, it is not yet clear how they might withstand practical attacks.

Other aspects of quantum cryptography have been investigated, and some, like quantum bit commitment, have been shown to be insecure. It is clear that quantum computing will have an important impact on cryptography although at this point it is not quite clear precisely what this impact will be.

BIOGRAPHY

Ian F. Blake received his undergraduate education at Queen's University in Kingston, Ontario and his Ph.D. at Princeton University in New Jersey. From 1967 to 1969 he was a Research Associate with the Jet Propulsion Laboratories in Pasadena, California. From 1969 to 1996 he was with the Department of Electrical and Computer Engineering at the University of Waterloo, in Waterloo, Ontario, where he was Chairman from 1978 to 1984 and Director of the Institute of Computer Research from 1990 to 1994. He is currently with the Department of Electrical and Computer Engineering at the University of Toronto, where he is Director of the Bell University Labs program.

His research interests are in the areas of cryptography, algebraic coding theory, digital communications, and spread-spectrum systems.

BIBLIOGRAPHY

1. W. Diffie and M. Hellman, New directions in Cryptography, *IEEE Trans. Inform. Theory* **22**: 644–654 (1976).
2. T. El Gamal, A public key cryptosystem and signature scheme based on discrete logarithms, *IEEE Trans. Inform. Theory* **31**: 469–472 (1985).
3. D. Kahn, *The Codebreakers*, Macmillan New York, 1967.
4. D. Gottesman and H.-K. Lo, From quantum cheating to quantum security, *Physics Today* **53**(11): 22–27 (Nov. 2000).
5. A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, FL, 1996.
6. R. L. Rivest, A. Shamir, and L. M. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Commun. ACM* **21**: 120–126 (1978).
7. B. Schneier, *Applied Cryptography*, 2nd, ed., Wiley, New York, 1996.
8. C. E. Shannon, Communication theory of secrecy systems, *Bell Syst. Tech. J.* **28**: 656–715 (1949).
9. D. Stinson, *Cryptography: Theory and Practice*, CRC Press, Boca Raton, FL, 2002.

CYCLIC CODES

STEPHEN B. WICKER
Cornell University
Ithaca, New York

1. INTRODUCTION

Cyclic codes are a class of highly structured algebraic block error control codes. This class includes Golay, BCH (Bose–Chaudhuri–Hocquenghem), and Reed–Solomon codes — codes that have arguably seen greater application than any other error control codes save the simple parity check. A Reed–Solomon decoder, for example, can be found in every compact-disk player. Golay, BCH, and Reed–Solomon codes have also been used in paging, mobile data systems, and deep-space telecommunications. Such cyclic codes have become an important basis for practical error control for a variety of reasons. First, they can be implemented using shift-register-based encoders and decoders, implementations that are of particular importance in high-speed data communication systems. Also, nonbinary cyclic codes, such as Reed–Solomon codes, provide a form of error trapping that is highly effective on fading channels. For this reason alone, Reed–Solomon codes have been used extensively in wireless data applications. In this article I will describe the structure of cyclic codes, paying particular attention to the special cases of Golay, BCH, and Reed–Solomon codes. I will mention several theoretical results, but will provide no proofs — the mathematically inclined reader is referred to Wicker [1] or Lin and Costello [2]. The true disciple is referred to MacWilliams and Sloane [3].

I begin with an historical overview. The general class of cyclic codes was first discussed in a series of technical notes and reports published from 1957 to 1959 by E. Prange at the Air Force Cambridge Research Labs [4–6]. In his work Prange identified cyclic codes with an algebraic structure called an "ideal," a connection that would lead to the development of BCH codes and a reinterpretation of Golay and Reed–Solomon codes a few years later.

Prange himself introduced a class of cyclic codes whose construction is based on quadratic residues. Quadratic residue (QR) codes are linear cyclic codes that generally have rates close to $\frac{1}{2}$ and have large minimum distances. A great deal of recent work on algebraic decoding algorithms for quadratic residue codes has increased their utility in a number of applications [e.g., 7,8].

Prange's work on QR codes showed that they include a small but very important group of codes — the Golay codes — that had been discovered almost 10 years earlier. In his 1948 paper that, among other things, gave birth to the field of information theory, Shannon gave a brief description of Hamming's perfect [7,4] binary code [9]. A code is said to be perfect if it has the maximum possible number of code words for a given error-correcting capability. After reading Shannon's paper, a number of people began searching for other perfect codes, with varying degrees of success. Golay, an engineer at the Signal Corps Engineering Laboratories in Fort Monmouth, New Jersey, published one of the first follow-up papers in June, 1949 [10]. In what is almost certainly the best short paper ever written (it fits on one side of an $8\frac{1}{2} \times 11$-in. sheet of paper, with room to spare), Golay extended the (7,4) Hamming code to a general class of p-ary codes of length $(p^n - 1)/(p - 1)$, where p is a prime.[1] In this same paper, Golay went on to describe a binary triple-error-correcting code and a ternary double-error-correcting code, both of which are perfect. Golay deduced the existence of the Golay codes through an examination of Pascal's triangle and the recognition of the relationship between the triangle's entries and the parameters of perfect codes. He then proceeded to find what we now call the Golay codes through a "limited search" of the triangle. The resulting code has served as fodder for specialists in abstract algebra and combinatorics ever since. On the practical side, Golay codes have seen frequent application in the United States space program, most notably with the *Voyager* I and II spacecraft. The extended binary Golay code served as the primary *Voyager* error control system, providing clear color pictures of Jupiter and Saturn between 1979 and 1981.[2] Golay codes were also used as the basis for a paging standard (called, not surprisingly, Golay).

Several efficient decoding algorithms for the Golay codes have been discovered. The most important was discovered in 1964, when Kasami described a shift register-based "error trapping" decoder [11]. The error trapping decoder is an extension of the shift register decoding techniques that will be discussed in this article.

The next subclass of cyclic codes that I will consider in detail is the BCH codes. Two independent research teams conducted the fundamental work on BCH codes and published their results at roughly the same time. A. Hocquenghem discussed binary BCH codes as "a generalization of Hamming's work" in a 1959 paper entitled "Codes correcteur d'erreurs" [12]. This was followed in March and September 1960 by Bose and Ray-Chaudhuri's publications on error-correcting binary group codes [13,14]. Given their simultaneous discovery of these codes, all three have given their name to what are now called BCH codes. Shortly after these initial publications, Peterson proved that BCH codes were cyclic and presented a moderately efficient decoding algorithm [15]. Gorenstein and Zierler then extended BCH codes from the binary field to arbitrary fields of size p^m, where p is a prime number [16].

Reed-Solomon codes were first described in a June 1960 paper entitled "Polynomial codes over certain finite fields," published in the *SIAM Journal on Applied Mathematics* by Irving Reed and Gus Solomon [17]. Through the work of Gorenstein and Zierler it was later discovered that Reed–Solomon codes and BCH codes are closely related, and that Reed–Solomon codes can be described as nonbinary BCH codes.

In 1960 Peterson provided the first explicit description of a decoding algorithm for binary BCH codes [15]. His "direct solution" algorithm is quite useful for correcting small numbers of errors, but becomes computationally intractable as the number of errors increases. Reed and Solomon discussed a decoding algorithm in their original paper on Reed–Solomon codes [17], but that algorithm was also inefficient for large codes and large numbers of errors corrected. Peterson's algorithm was improved and extended to nonbinary codes by Gorenstein and Zierler [16], Chien [21], and Forney [22], but it was not until 1967 that Berlekamp introduced the first truly efficient decoding algorithm for both binary and nonbinary BCH codes [18]. In 1969 Massey showed that Berlekamp's algorithm was a general solution to the problem of synthesizing the shortest linear feedback shift register capable of generating a given sequence [23]. Massey then demonstrated a fast shift-register-based decoding algorithm for BCH and Reed–Solomon codes that is equivalent to Berlekamp's algorithm.

In 1975 Sugiyama et al. showed that Euclid's algorithm can be used to decode BCH and Reed–Solomon codes [24]. Reed et al. then showed in 1978 that a related technique based on continued fractions and Fermat-theoretic transforms resulted in a fast decoding algorithm for Reed–Solomon codes [25]. Completing the decoding picture, a frequency domain approach to decoding BCH codes was introduced by Gore in 1973 [26]. Blahut provided a more general discussion of spectral decoding techniques in 1979 [27].

In the remainder of this article, I will provide a brief overview of the general theory of cyclic codes, and then turn to the three important classes (the Golay codes, BCH, and Reed–Solomon codes). I will close with a discussion of several important applications.

2. GENERAL THEORY

A code is said to be cyclic if it satisfies a very simple property–every codeword must be the right cyclic shift of another codeword. More formally, an (n, k)[3] code $\boldsymbol{C}$ is said to be cyclic if for every codeword $\mathbf{c} = (c_0, c_1, \ldots, c_{n-2}, c_{n-1}) \in \boldsymbol{C}$, there is also a codeword $\mathbf{c}' = (c_{n-1}, c_0, \ldots, c_{n-3}, c_{n-2}) \in \boldsymbol{C}$. Since the codeword $\mathbf{c}$ in this definition has been

[1] It is a virtual certainty that this result was already known to Hamming [18]. A more detailed discussion of the general class of codes now known as *Hamming codes* is included in Hamming's 1950 paper [19], which was probably delayed due to patent considerations.

[2] A secondary error control system based on Reed–Solomon codes was substituted for the Golay system for the *Voyager* 2 flybys of Uranus and Neptune in the mid-1980s [20].

[3] In this article the usual conventions are adopted — an (n, k) code over GF(q) is a collection of vectors of length n that form a vector space of dimension k over the Galois field GF(q).

arbitrarily selected from among all the codewords in $\boldsymbol{C}$, it follows that all n of the distinct cyclic shifts of $\mathbf{c}$ must also be codewords in $\boldsymbol{C}$. To see this, replace $\mathbf{c}$ with $\mathbf{c}'$ and apply the definition again.

The key to the underlying structure of cyclic codes lies in the association of a *code polynomial* $c(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{n-1}x^{n-1}$ with every codeword $\mathbf{c} = (c_0, c_1, \ldots, c_{n-2}, c_{n-1}) \in \boldsymbol{C}$. If C is an (n, k) code over Galios Field GF(q), it follows that the set of code polynomials associated with $\boldsymbol{C}$ form a vector space as well. The terms *codeword* and *code polynomial* are henceforth used interchangeably. When we look at the definition of cyclic in terms of the code polynomials, some interesting structure comes to light. If the code word $\mathbf{c}'$ is the right cyclic shift of the code word $\mathbf{c} \in \boldsymbol{C}$, then $c'(x) = x \cdot c(x)$ modulo $(x^n - 1) \in \boldsymbol{C}$. This can be seen as follows.

Continuing along this line, we can show that two right cyclic shifts of a code word are equivalent to the multiplication modulo $(x^n - 1)$ of the associated code polynomial by x^2

$$\begin{aligned} x \cdot c(x) \bmod (x^n - 1) &= (c_0x + c_1x^2 + \cdots + c_{n-1}x^n) \bmod (x^n - 1) \\ &\equiv c_{n-1} + c_0x + \cdots + c_{n-2}x^{n-1} \\ &= c'(x) \end{aligned}$$

where three right cyclic shifts are equivalent to multiplication modulo $(x^n - 1)$ by x^3, and so on. Let the cyclic shifts of $\mathbf{c}$ and the associated polynomials be represented as follows.

$$\begin{aligned} \mathbf{c} &= (c_0, c_1, \ldots, c_{n-1}) \leftrightarrow c(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1} \\ \mathbf{c}' &= (c_{n-1}, c_0, \ldots, c_{n-2}) \\ &\leftrightarrow c'(x) = c_{n-1} + c_0x + \cdots + c_{n-2}x^{n-1} \\ \mathbf{c}'' &= (c_{n-1}, c_0, \ldots, c_{n-2}) \\ &\leftrightarrow c''(x) = c_{n-2} + c_{n-1}x + \cdots + c_{n-3}x^{n-1} \\ &\vdots \\ \mathbf{c}^{(n-1)} &= (c_{n-1}, c_0, \ldots, c_{n-2}) \\ &\leftrightarrow c^{(n-1)}(x) = c_1 + c_2x + \cdots + c_0x^{n-1} \end{aligned}$$

Let $a(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}$ be an arbitrary polynomial in GF(q)[x]/$(x^n - 1)$ [the ring of polynomials with coefficients in GF(q) and maximum degree $n - 1$]. The product $a(x)c(x)$ is a linear combination of cyclic shifts of $\mathbf{c}$. Since $\boldsymbol{C}$ forms a vector space, $a(x)c(x)$ must be a valid code polynomial. This result shows that the space formed by the code polynomials of $\boldsymbol{C}$ has an interesting structure — it is an *ideal* in the ring of polynomials of degree n or less with coefficients in GF(q) (normally written as GF(q)[x]/$(x^n - 1)$). Several interesting, practical properties follow immediately from this result.

2.1. The Basic Properties of Cyclic Codes

Let $\boldsymbol{C}$ be a (n, k) linear cyclic code over GF(q):

- Within the set of code polynomials in $\boldsymbol{C}$ there is a unique monic polynomial $g(x)$ with minimal degree $r < n$. $g(x)$ is called the *generator polynomial* of $\boldsymbol{C}$.
- Every code polynomial $c(x)$ in $\boldsymbol{C}$ can be expressed uniquely as $c(x) = m(x)g(x)$, where $g(x)$ is the generator polynomial of $\boldsymbol{C}$ and $m(x)$ is a polynomial of degree less than $(n - r)$ in GF(q)[x].
- The generator polynomial $g(x)$ of $\boldsymbol{C}$ is a factor of $(x^n - 1)$ in GF(q)[x].

The proof for these statements follows from the definition of ideal; the interested reader is referred to Wicker [1]. or MacWilliams [3]. The last of the three properties leads to some very interesting structural issues for cyclic codes, but to see this, we have to introduce some additional abstract algebra.

Let β be an element in the Galois field GF(q^m). The *conjugates of β with respect to the subfield GF(q)* are the elements $\beta, \beta^q, \beta^{q2}, \beta^{q3}$, and so on. Note that, since the field is finite, this series of elements has to start repeating at some point. These conjugates form a set called a *conjugacy class*. The elements in a Galois field can be partitioned into conjugacy classes with respect to a subfield of the Galois field. For example, the elements in the field GF(8) can be partitioned into conjugacy classes with respect to GF(2) in the following way. Let α be a primitive element in GF(8) (a "primitive" element is an element whose powers generate all the nonzero elements in the field). The various powers of α fall into three conjugacy classes with respect to GF(2): $\{\alpha^0 = 1\}$, $\{\alpha, \alpha^2, \alpha^4\}$, and $\{\alpha^3, \alpha^5, \alpha^6\}$. The fourth and final conjugacy class is $\{0\}$.

Rather than deal with a given conjugacy class itself, it is often easier to focus on the exponents of the powers of α that form the conjugacy class. The result is a *cyclotomic coset*. The cyclotomic cosets associated with the nonzero conjugacy classes in the above example are $\{0\}$, $\{1, 2, 4\}$, and $\{3, 5, 6\}$. If we want to include the zero element, we can adopt a symbol, say, an asterisk $\{^*\}$, to formally represent $\log_\alpha 0$, but this will not be necessary for what follows. If we ignore the zero element, the cyclotomic cosets modulo n with respect to GF(q) are a partitioning of the integers $\{0, 1, \ldots, n - 1\}$ into sets of the form $\{a, aq, aq^2, aq^3, \ldots, aq^{d-1}\}$.

With a bit of effort, the preceding development and some Galois field mathematics yield the following key result: the roots of a polynomial with coefficients in GF(q) must be the union of one or more conjugacy classes with respect to GF(q) (see, e.g., Wicker [1]). To see what this means, consider a generator polynomial $g(x)$ for a binary cyclic code of length 7. According to our properties of cyclic codes, $g(x)$ must be a divisor of $x^7 - 1$. The key result says that any binary polynomial that is a divisor of $x^7 - 1$ must have roots that are the union of one of the conjugacy classes with respect to GF(2) that we listed above. Since zero is not a root of $x^7 - 1$, we have to focus on the three nonzero conjugacy classes. Each of these classes is associated with a *minimal polynomial*, a polynomial whose roots are the elements of a single conjugacy class.

Conjugacy Class	Associated Minimal Polynomial
$\{\alpha^0 = 1\}$	$M_0(x) = (x - 1) = x + 1$
$\{\alpha, \alpha^2, \alpha\}$	$M_1(x) = (x - \alpha)(x - \alpha^2)(x - \alpha^4)$ $= x^3 + x + 1$
$\{\alpha^3, \alpha^6, \alpha^5\}$	$M_3(x) = (x - \alpha^3)(x - \alpha^6)(x - \alpha^5)$ $= x^3 + x^2 + 1$

We conclude that the only generator polynomials for binary, cyclic codes of length 7 are $M_0(x)$, $M_1(x)$, $M_3(x)$, $M_0(x)M_1(x)$, $M_0(x)M_1(x)$, and $M_1(x)M_3(x)$. In general, the binary polynomials that are available for use as generator polynomials for cyclic codes are the products of one or more minimal polynomials.

2.2. Systematic Encoding

An encoding for a given error control code is said to be *systematic* if the codeword consists of a set of parity symbols followed by the message itself. This greatly facilitates decoding, for if a parity check indicates that there are no errors in the received word, the parity symbols can be deleted and the remaining message forwarded to the application. By exploiting the algebraic structure of cyclic codes, it is possible to develop a simple systematic encoder.

Consider an (n, k) cyclic code $\boldsymbol{C}$ with generator polynomial $g(x)$. A k-symbol message $\mathbf{m} = (m_0, m_1, \ldots, m_{k-1})$ is encoded as follows. Multiply the corresponding message polynomial $m(x)$ by x^{n-k}, obtaining $x^{n-k}m(x) = m_0x^{n-k} + m_1x^{n-k+1} + \cdots + m_{k-1}x^{n-1}$. This product is associated with an n-symbol block $(0, 0, \ldots, 0, m_0, m_1, \ldots, m_{k-1})$ whose first $(n-k)$ coordinates are zero. Now divide $x^{n-k}m(x)$ by $g(x)$ to obtain $x^{n-k}m(x) = q(x)g(x) + d(x)$, where $d(x)$ is the remainder. Since $c(x) = [x^{n-k}m(x) - d(x)] = q(x)g(x)$ is a multiple of $g(x)$, it must be a valid code polynomial. Now note that the remainder $d(x)$ has a degree less than $(n-k)$, the degree of the generator polynomial $g(x)$. The term $-d(x)$ can be associated with an n-symbol block whose last k coordinates are zero: $-d(x) \leftrightarrow (-d_0, -d_1, \ldots, -d_{n-k-1}, 0, 0, \ldots, 0)$. The codeword associated with the code polynomial $c(x) = [x^{n-k}m(x) - d(x)]$ thus has the form

$$c(x) = [x^{n-k}m(x) - d(x)] \leftrightarrow (-d_0, -d_1, \ldots, -d_{n-k-1}, m_0, m_1, \ldots, m_{k-1})$$

where $m(x)$ has been systematically mapped to a codeword. The encoding algorithm is summarized below:

1. Multiply the message polynomial $m(x)$ by x^{n-k}.
2. Divide the result of step 1 by the generator polynomial $g(x)$. Let $d(x)$ be the remainder.
3. Set $c(x) = x^{n-k}m(x) - d(x)$.

As an example, consider the systematic encoding of a (7,3) binary code with generator polynomial $g(x) = x^4 + x^3 + x^2 + 1$; we will encode the message block (101):

1. $x^{n-k}m(x) = x^4(x^2 + 1) = x^6 + x^4$
2. $$\begin{array}{r|l} x^4+x^3+x^2+1 & x^2+x+1 = q(x) \\ & \underline{x^6+x^4} \\ & \underline{x^6+x^5+x^4+x^2} \\ & \underline{x^5+x^2} \\ & \underline{x^5+x^4+x^3+x} \\ & \underline{x^4+x^3+x^2+1} \\ & x+1 = d(x) \end{array}$$
3. $c_m(x) = x^{n-k}m(x) - d(x) = 1 + x + x^4 + x^6$ $\leftrightarrow \boldsymbol{c}_m = (1100101)$

2.3. Shift Register Encoders and Decoders for Cyclic Codes

Data rates in the hundreds or even thousands of megabits per second are common in many applications. Unfortunately, such data rates severely limit the device technologies that can be used to implement error control systems, and within a given technology, the complexity of the circuits. It is extremely important to note that encoders and decoders for cyclic codes can be implemented using simple exclusive-OR gates, switches, shift registers, and, in the case of nonbinary encoders and decoders, finite-field adder and multiplier circuits. Shift registers are among the simplest of digital circuits, consisting of a collection of flipflops connected in series. They are operable at speeds quite close to the maximum speed possible for a single gate using a given device technology.

The systematic encoding procedure described above has a simple shift register implementation, as shown in Fig. 1. The first step — multiplication of the message polynomial — is implemented by inserting the message into the shift register circuit on the right side [this is equivalent to padding the message block with an initial $(n-k)$ zeros]. Polynomial division is then performed through the use of a linear feedback shift register (LFSR).

Encoding proceeds as follows. During the first step of the encoding operation the three switches are placed in position X and the k message symbols are fed into the

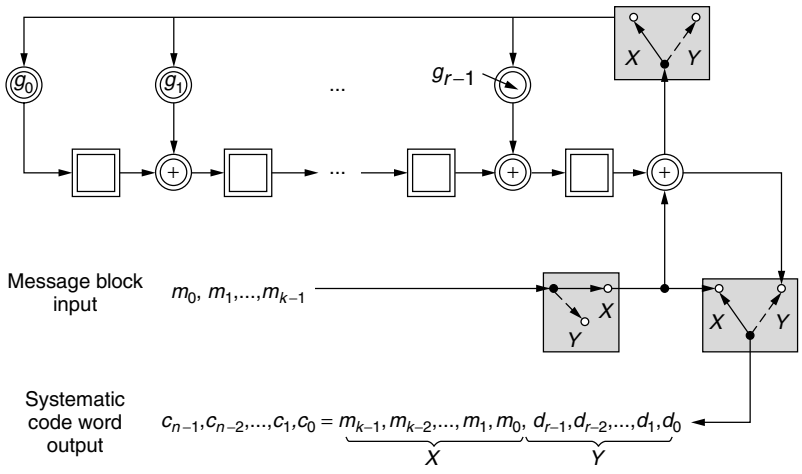


Figure 1. Systematic shift register encoding circuit for cyclic codes [1].

encoder in order of decreasing index. The k symbol bits are simultaneously sent to the transmitter, for they represent the last k coordinates of the systematic codeword. This, by the way, is the primary rationale for placing the message symbols at the end of a systematic code word. After the kth message symbol has been fed into the shift register, the switches are moved to position Y. At this point the shift register cells contain the remainder generated by the division operation. These symbols are then shifted out of the shift register and to the transmitter, where they constitute the remaining systematic codeword coordinates.

Cyclic codes allow for a number of highly convenient techniques for detecting errors using shift register circuits. For example, consider the following:

1. Since the transmitted codeword was systematically encoded, we can construct an estimated message block $\mathbf{m}'$ and estimated remainder block $\mathbf{d}'$ using the values in the message and parity positions of the received word $\mathbf{r}$.
2. Encode $\mathbf{m}'$ using an encoder identical to that used by the transmitter and obtain an estimated remainder block $\mathbf{d}''$.
3. Compare $\mathbf{d}'$ to $\mathbf{d}''$. If they are not the same, then $\mathbf{r}$ is not a valid codeword, indicating the presence of errors in the received word.

This approach to error detection has a significant advantage. The encoder and error detection circuits are essentially identical, and the design process is correspondingly simplified.

The error detection technique described above can be used as the first steps in an error-correcting algorithm. The difference $\mathbf{s} = \mathbf{d}' - \mathbf{d}''$ is called the *syndrome* for the received word $\mathbf{r}$. Maximum-likelihood error correction is performed by computing the syndrome for a received vector, determining the most likely error pattern among those associated with the syndrome, and subtracting this error pattern from the received vector. The primary drawback to this approach is usually the size of the syndrome lookup table. For an arbitrary (n, k) q-ary code, the syndrome table must contain $q^{n-k}n$-tuples — one for each possible syndrome. Cyclic codes have an interesting property that allows us to cut the size of the syndrome table to $1/n$th its original size.

> Let $s(x)$ be the syndrome polynomial corresponding to a received polynomial $r(x)$. Let $r^{(1)}(x)$ be the polynomial obtained by cyclically shifting the coefficients of $r(x)$ once to the right. Then the remainder obtained when dividing $xs(x)$ by $g(x)$ is the syndrome $s^{(1)}(x)$ corresponding to $r^{(1)}(x)$.

Let's consider this result in terms of the shift register syndrome circuit. Given a received vector $\mathbf{r}$, the corresponding syndrome $\mathbf{s}$ is obtained by entering $\mathbf{r}$ into a shift register division circuit. When the last symbol of $\mathbf{r}$ has been shifted into the circuit, the shift register cells contain the syndrome. The input to the circuit of an additional zero at this point is equivalent to multiplying s(x) by x and dividing by $g(x)$. The remainder $s^{(1)}(x)$ is now in the shift register cells. According to the result given above, $s^{(1)}(x)$ is the syndrome for $r^{(1)}(x)$. This process can be repeated n times, bringing us back to the starting point with the original syndrome in the shift register cells. We need only store one syndrome $\mathbf{s}$ for an error pattern $\mathbf{e}$ and all cyclic shifts of $\mathbf{e}$. A syndrome decoder with a reduced-size lookup table is used as follows.

2.4. Syndrome Decoder for Cyclic Codes

1. Set a counting variable j to the value 0. Compute the syndrome $\mathbf{s}$ for a received vector $\mathbf{r}$.
2. Look for the error pattern $\mathbf{e}$ corresponding to $\mathbf{s}$ in the syndrome lookup table. If there is such a value, go to step 7.
3. Increment the counter j by 1 and enter a zero into the shift register circuit at the input, computing $\mathbf{s}^{(j)}$.
4. Look for the error pattern $\mathbf{e}^{(j)}$ corresponding to $\mathbf{s}^{(j)}$ in the syndrome lookup table. If there is such a value, go to step 6.
5. Go to step 3.
6. Determine the error pattern $\mathbf{e}$ corresponding to $\mathbf{s}$ by cyclically shifting $\mathbf{e}^{(j)}j$ times to the left.
7. Subtract $\mathbf{e}$ from $\mathbf{r}$, obtaining the codeword $\mathbf{c}$.

So in general, there are fast, simple encoding and decoding circuits for cyclic codes of modest size. To get truly powerful error control with limited complexity, however, it is necessary to turn to one of the special cases.

3. QUADRATIC RESIDUE CODES AND GOLAY CODES

The nonzero squares modulo p, p a prime, are called the quadratic residues modulo p. Quadratic residues can be found by simply squaring every integer modulo p. For example, note that, using modulo 7 arithmetic, $1^2 = 6^2 = 1$, $2^2 = 5^2 = 4$ and $3^2 = 4^2 = 2$. So 1, 2, and 4 are the three quadratic residues modulo 7. In showing how quadratic residues can lead to some interesting codes, we have to indulge in some abstract algebra. The reader interested solely in applications and other practical matters can safely skip the next few paragraphs.

The set of integers modulo p, p a prime, form the field GF(p) under modulo p addition and multiplication. Let Q be the set of quadratic residues modulo p and N the set of corresponding nonresidues. Since GF(p) is a Galois field, there must exist at least one primitive element $\gamma \in \mathrm{GF}(p)$ that generates all the elements in Q and N. It follows that γ must be a quadratic nonresidue; otherwise there exists some element $\sqrt{\gamma} \in \mathrm{GF}(p)$ that generates $2(p-1)$ distinct elements in GF(p), contradicting the order of GF(p). One can then see that $\gamma^e \in Q$ if and only if e is even; otherwise, $\gamma^e \in N$. We conclude that all the elements in Q correspond to the first $(p-1)/2$ consecutive powers of γ^2, and that Q is a cyclic group under modulo p multiplication.

Now consider a field GF(s^m) that contains a primitive pth root of unity. Such a field exists for a given s, m, and p whenever $p|s^m$-1. We add the further restriction that s must be a quadratic residue modulo p. This can be somewhat restrictive; for example, if $s = 2$, then p must be of the form $p = (8k \pm 1)$. Since Q is a cyclic group,

multiplication of any element in Q by any other element in Q must result in an element in Q. It follows that the conjugates with respect to GF(s) of any element in Q must also be in Q. We conclude that Q is the union of one or more cyclotomic cosets modulo p with respect to GF(s).

Let α be primitive in GF(s^m). The results above show that the following polynomials have coefficients in the subfield GF(s):

$$q(x) = \prod_{i \in Q} (x - \alpha^i)$$

$$n(x) = \prod_{i \in N} (x - \alpha^i)$$

The *quadratic residue codes* of length p are defined by the generator polynomials $q(x)$, $(x-1)q(x)$, $n(x)$, and $(x-1)n(x)$, respectively.

The *binary Golay code* $\mathbf{G}_{23}$ is the (23,12,7) quadratic residue code with $p = 23$ and $s = 2$. The construction of this code proceeds as follows. The quadratic residues modulo 23 are $Q = 1, 2, 3, 4, 6, 8, 9, 12, 13, 16, 18$. Let β be a primitive 23rd root of unity [22 different β values can be found in GF(2^{11}) and its extensions]. The distinct powers of β form two cyclotomic cosets modulo 23 with respect to GF(2):

$$C_1 = \{1, 2, 3, 4, 6, 8, 9, 12, 13, 16, 18\}$$

$$C_5 = \{5, 7, 10, 11, 14, 15, 17, 19, 20, 21, 22\}$$

The term $x^{23} + 1$ factors into three binary irreducible polynomials:

$$x^{23} + 1 = (x + 1)(x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1) \times (x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1)$$

Depending on the selection of β, there are two possible generator polynomials for $\mathbf{G}_{23}$:

$$g_1(x) = x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1$$

$$g_2(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$$

Using either of these generator polynomials, the resulting code can be shown to be triple-error correcting. Given that $\mathbf{G}_{23}$ has dimension 12, it can be shown that $\mathbf{G}_{23}$is perfect. Each codeword is associated with a decoding sphere containing all vectors that are Hamming distance ≤ 3 from the codeword. Since the vectors have length 23, the decoding spheres have cardinality:

$$\begin{aligned} V_2(23, 3) &= \binom{23}{0} + \binom{23}{1} + \binom{23}{2} + \binom{23}{3} \\ &= 1 + 23 + 253 + 1771 \\ &= 2^{11} \end{aligned}$$

Through the addition of a parity-check bit, $\mathbf{G}_{23}$ can be extended to form $\mathbf{G}_{24}$, the triple-error-correcting, quadruple-error-detecting code that was used on the *Voyager* spacecraft.

The ternary [i.e., defined over GF(3)] Golay code ternary Golay code $\mathbf{G}_{11}$ is the quadratic residue code with $p = 11$ and $s = 3$. An analysis similar to that for $\mathbf{G}_{23}$ yields the factorization of $x^{11} - 1$ into three irreducible polynomials in GF(3)[x]:

$$x^{11} - 1 = (x - 1)(x^5 + x^4 - x^3 + x^2 - 1) \times (x^5 - x^3 + x^2 - x - 1)$$

Again there are two possible generator polynomials:

$$g_1(x) = x^5 + x^4 - x^3 + x^2 - 1$$

$$g_2(x) = x^5 - x^3 + x^2 - x - 1$$

$\mathbf{G}_{11}$ has length 11, dimension 6, and minimum distance 5, and is perfect.

4. BCH AND REED–SOLOMON CODES

When constructing an arbitrary cyclic code, there is no guarantee as to the resulting minimum distance. Given a generator polynomial $g(x)$, we must conduct a computer search of all corresponding nonzero code words $c(x)$ to determine the minimum-weight codeword and thus the minimum distance of the code. BCH codes, on the other hand, take advantage of a useful result that ensures a minimum "design distance" given a particular constraint on the generator polynomial. This result is known as the *BCH bound*.

4.1. The BCH Bound

Let $\boldsymbol{C}$ be a q-ary (n, k) cyclic code with generator polynomial $g(x)$. A *primitive* nth *root of unity* is defined to be an element β such that $\beta^n = 1$, but there is not smaller nonzero integer j such that $\beta^j = 1$. If GF(q^m) is the smallest extension field of GF(q) that contains a primitive nth root of unity, then we say that m is the multiplicative order of q modulo n. The BCH bound developed as follows. Let α be a primitive nth root of unity. Select $g(x)$ to be the minimal degree polynomial in GF(q)[x] such that $g(\alpha^b) = g(\alpha^{b+1}) = g(\alpha^{b+2}) = \cdots = g(\alpha^{b+d-2}) = 0$ for some integers $b \geq 0$ and $d \geq 1$. This $g(x)$ has $(d-1)$ consecutive powers of α as zeros. The BCH bound states that the code $\boldsymbol{C}$ defined by such a $g(x)$ has minimum distance $d_{\min} \geq d$.

We can use the BCH bound to construct a t-error-correcting q-ary BCH code of length n in the following way:

1. Find a primitive nth root of unity α in a field GF(q^m), where m is minimal.
2. Select $(d-1) = 2t$ consecutive powers of α, starting with α^b for some nonnegative integer b.
3. Let $g(x)$ be the least common multiple of the minimal polynomials for the selected powers of α with respect to GF(q). (Each minimal polynomial should appear only once in the product.)

Step 1 follows from our design procedure for general cyclic codes. Steps 2 and 3 ensure, through the BCH bound,

that the minimum distance of the resulting code equals or exceeds d and that the generator polynomial has the minimal possible degree. Since $g(x)$ is a product of minimal polynomials with respect to GF(q), $g(x)$ must be in GF(q)[x] and the corresponding code is q-ary with $d_{\min} \geq d$.

If $b = 1$, then the BCH code is said to be *narrow-sense*. If $n = q^m - 1$ for some positive integer m, then the BCH code is said to be *primitive*, for the nth root of unity α is a primitive element in GF(q^m). Let's consider a few binary BCH codes of length 31 to see how the design rule derived from the BCH bound fits in with the earlier discussion of conjugacy classes. Let α be a root of the primitive polynomial $x^5 + x^2 + 1$. It is thus a primitive element in the field GF(32). Since 31 is of the form $2^m - 1$, our BCH codes in this example are primitive. We begin by determining the cyclotomic cosets modulo 31 with respect to GF(2) and the associated minimal polynomials.

Cyclotomic Cosets	Minimal Polynomials
$C_0 = \{0\}$	$\leftrightarrow M_{(0)}(x) = x + 1$
$C_1 = \{1, 2, 4, 8, 16\}$	$\leftrightarrow M_{(1)}(x) = x^5 + x^2 + 1$
$C_3 = \{3, 6, 12, 24, 17\}$	$\leftrightarrow M_{(3)}(x) = x^5 + x^4 + x^3 + x^2 + 1$
$C_5 = \{5, 10, 20, 9, 18\}$	$\leftrightarrow M_{(5)}(x) = x^5 + x^4 + x^2 + x + 1$
$C_7 = \{7, 14, 28, 25, 19\}$	$\leftrightarrow M_{(7)}(x) = x^5 + x^3 + x^2 + x + 1$
$C_{11} = \{11, 22, 13, 26, 21\}$	$\leftrightarrow M_{(11)}(x) = x^5 + x^4 + x^3 + x + 1$
$C_{15} = \{15, 30, 29, 27, 23\}$	$\leftrightarrow M_{(15)}(x) = x^5 + x^3 + 1$

Recall that if a code $\mathbf{C}$ is to be a binary cyclic code, then it must have a generator polynomial $g(x)$ that is the product one or more of the minimal polynomials listed above. According to the BCH bound, if $\mathbf{C}$ is to be t-error correcting BCH code, then $g(x)$ must have as zeros $2t$ consecutive powers of α.

- *One-Error-Correcting Narrow-Sense Primitive BCH Code*. Since the code is to be narrow-sense and single-error-correcting, $b = 1$ and $\delta = 3$. The generator polynomial must thus have α and α^2 as zeros. $M_{(1)}(x)$ is the minimal polynomial of both α and α^2. The generator polynomial is thus

$$g(x) = \text{LCM}(M_{(1)}(x)M_{(2)}(x)) = M_{(1)}(x) = x^5 + x^2 + 1$$

  Since the degree of the generator polynomial $g(x)$ is 5, the dimension of the resulting code is $31 - 5 = 26$. Thus $g(x)$ defines a (31,26) binary single-error-correcting BCH code.
- *Two-Error-Correcting Narrow-Sense Primitive BCH Code*. Again $b = 1$, but δ has been increased to 5. $g(x)$ must thus have as roots α, α^2, α^3, and α^4.

$$\begin{aligned} g(x) &= LCM(M_{(1)}(x), M_{(2)}(x), M_{(3)}(x), M_{(4)}(x)) \\ &= M_{(1)}(x)M_{(3)}(x) \\ &= (x^5 + x^2 + 1)(x^5 + x^4 + x^3 + x^2 + 1) \\ &= x^{10} + x^9 + x^8 + x^6 + x^5 + x^3 + 1 \end{aligned}$$

  Since the degree of $g(x)$ is 10, it defines a (31,21) binary double-error-correcting code.

There are a number of ways to define Reed–Solomon codes. Reed and Solomon's initial definition focused on the evaluation of polynomials over the elements in a finite field [17]. Reed–Solomon codes can also be viewed as a natural extension of BCH codes. Simply put, a Reed–Solomon code is a q^m-ary BCH code of length $q^m - 1$.

Consider the construction of a t-error-correcting Reed–Solomon code of length $(q^m - 1)$. The first step is to note that the required primitive $(q^m - 1)$st root of unity α can be found in GF(q^m) (every finite field of size q contains an element with order $q - 1$). Since the code symbols are to be from GF(q^m), the next step is to construct the cyclotomic cosets modulo $(q^m - 1)$ with respect to GF(q^m). This is a trivial task, for $(s \cdot q^m) \equiv s$ modulo $(q^m - 1)$. The cyclotomic cosets are singleton sets of the form $\{s\}$ and the associated minimal polynomials are of the form $(x - \alpha^s)$.

The BCH bound indicates that $2t$ consecutive powers of a are required as zeros of the generator polynomial $g(x)$ for a t-error-correcting Reed-Solomon code. The generator polynomial is the product of the associated minimal polynomials:

$$g(x) = (x - \alpha b)(x - \alpha^{b+1})(x - \alpha^{b+2}) \cdots (x - \alpha^{b+2t-1}).$$

Consider a two-error-correcting 8-ary Reed–Solomon code of length 7. Let α be a root of the primitive binary polynomial $x^3 + x + 1$ and thus a primitive 7th of unity. The Galois field GF(8) can be represented as consecutive powers of α:

$$\begin{aligned} \alpha &= \alpha & \alpha^5 &= \alpha^2 + \alpha + 1 \\ \alpha^2 &= \alpha^2 & \alpha^6 &= \alpha^2 + 1 \\ \alpha^3 &= \alpha + 1 & \alpha^7 &= 1 \\ \alpha^4 &= \alpha^2 + \alpha & 0 &= 0 \end{aligned}$$

If the resulting code is to be double-error-correcting, it must have $2t = 4$ consecutive powers of a as zeros. A narrow-sense generator polynomial is constructed as follows:

$$\begin{aligned} g(x) &= (x - \alpha)(x - \alpha 2)(x - \alpha 3)(x - \alpha 4) \\ &= x^4 + \alpha^3 x^3 + x^2 + \alpha x + \alpha^3 \end{aligned}$$

Since the generator polynomial has degree 4, the (7,3) Reed–Solomon code it defines has dimension 3 over GF(8) and thus $8^3 = 512$ codewords.

Reed–Solomon codes have a number of interesting properties that are not shared by the other BCH codes. Recall that the minimum distance for BCH codes is in general lower-bounded by the design distance, but in many cases the actual minimum distance exceeds the design distance. One of the most significant properties of Reed–Solomon codes is the fact that an (n, k) Reed–Solomon code always has minimum distance exactly equal to $(n - k + 1)$.

BCH and Reed–Solomon codes can be decoded in a number of different ways. Perhaps the most efficient technique is Berlekamp's algorithm. In this article we provide a

brief description of Berlekamp's algorithm. Readers interested in a detailed exposition are referred to the source in Ref. 18.

The definition of a Reed–Solomon generating polynomial requires that, for some b and t, $g(\alpha^b) = g(\alpha^{b+1}) = \cdots = g(\alpha^{b+2t-1}) = 0$. A binary vector $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})$ is a codeword if and only if its associated polynomial $c(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}$ has as zeros these same $2t$ consecutive powers of α. Now consider a received polynomial $r(x)$, which can be expressed as the sum of the transmitted code polynomial $c(x)$ and an error polynomial $e(x) = e_0 + e_1 x + \cdots + e_{n-1}x^{n-1}$. A series of syndromes is obtained by evaluating the received polynomial at the $2t$ zeros. To minimize the complexity of our notation, it is assumed henceforth that all codes under discussion are narrow-sense ($b = 1$). The syndromes are computed as follows:

$$\begin{aligned} S_j &= r(\alpha^j) = c(\alpha^j) + e(\alpha^j) = e(\alpha^j) \\ &= \sum_{k=0}^{n-1} e_k (\alpha^j)^k, \qquad j = 1, 2, \ldots, 2t \end{aligned}$$

The computations in this expression are performed in GF(2^m), the field containing the primitive nth root of unity. Now assume that the received word $\mathbf{r}$ has ν errors in positions $i_1, i_2, \ldots, i_\nu$. We will assume that the code is binary, so the errors in these positions have value $e_{ij} = 1$. The syndrome sequence can be reexpressed in terms of these error locations:

$$S_j = \sum_{l=1}^{v} e_{i_l} (\alpha^j)^{i_l} = \sum_{l=1}^{v} (\alpha^{i_l})^j = \sum_{l=1}^{v} X_l^j, \qquad j = 1, \ldots, 2t$$

The $\{X_l\}$ are *error locators*, for their values indicate the positions of the errors in the received word. Expanding this equation we obtain a sequence of $2t$ algebraic *syndrome equations syndrome equations* in the ν unknown error locations:

$$\begin{aligned} S_1 &= X_1 + X_2 + \cdots + X_v \\ S_2 &= X_1^2 + X_2^2 + \cdots + X_v^2 \\ S_3 &= X_1^3 + X_2^3 + \cdots + X_v^3 \\ &\vdots \\ S_{2t} &= X_1^{2t} + X_2^{2t} + \cdots + X_v^{2t} \end{aligned}$$

Equations of this form are called *power-sum symmetric functions*. Since they form a system of nonlinear algebraic equations in multiple variables, they are somewhat difficult to solve in a direct manner.[4] Peterson showed, however, that the BCH syndrome equations can be translated into a series of linear equations that are much easier to work with [15]. Let $\Lambda(x)$ be the *error locator polynomial* that has as its roots the inverses of the ν error locators $\{X_l\}$:

$$\Lambda(x) = \prod_{l=1}^{v} (1 - X_l x) = \Lambda_v x^v + \Lambda_{v-1} x^{v-1} + \cdots + \Lambda_1 x + \Lambda_0$$

This equation can be used to express the coefficients of $\Lambda(x)$ directly in terms of the $\{X_l\}$. The resulting expressions are the *elementary symmetric functions* of the error locators. Power-sum symmetric functions and elementary symmetric functions are related by *Newton's identities*, which are generally expressed as follows for polynomials over arbitrary fields:

$$\begin{aligned} S_1 + \Lambda_1 &= 0 \\ S_2 + \Lambda_1 S_1 + 2\Lambda_0 &= 0 \\ S_3 + \Lambda_1 S_2 + \Lambda_2 S_1 + 3\Lambda_3 &= 0 \\ &\vdots \\ S_v + \Lambda_1 S_{v-1} + \Lambda_2 S_{v-2} + \cdots + \Lambda_{v-1} S_1 + v\Lambda_v &= 0 \\ S_{v+1} + \Lambda_1 S_v + \Lambda_2 S_{v-1} + \cdots + \Lambda_v S_1 &= 0 \\ &\vdots \\ S_{2t} + \Lambda_1 S_{2t-1} + \Lambda_2 S_{2t-2} + \cdots + \Lambda_v S_{2t-v} &= 0 \end{aligned}$$

If we assume that the codes in question are binary, we can reduce these expressions to the following:

$$\begin{aligned} S_1 + \Lambda_1 &= 0 \\ S_3 + \Lambda_1 S_2 + \Lambda_2 S_1 + \Lambda_3 &= 0 \\ S_5 + \Lambda_1 S_4 + \Lambda_2 S_3 + \Lambda_3 S_2 + \Lambda_4 S_1 + \Lambda_5 &= 0 \\ &\vdots \\ S_{2t-1} + \Lambda_1 S_{2t-2} + \Lambda_2 S_{2t-3} + \cdots + \Lambda_t S_{t-1} &= 0 \end{aligned}$$

Now suppose, for a moment, that we had an infinite number of syndromes available. We could then define an infinite-degree syndrome polynomial as follows:

$$S(x) = S_1 x + S_2 x^2 + \cdots + S_{2t} x^{2t} + S_{2t+1} x^{2t+1} + \cdots$$

Clearly we do not know all of the coefficients of $S(x)$, but fortunately the first $2t$ coefficients are entirely sufficient. $S(x)$ is made into an infinite degree polynomial so that it can be treated as a *generating function*. Define a third polynomial as follows:

$$\begin{aligned} \Omega(x) &\triangleq [1 + S(x)]\Lambda(x) \\ &= (1 + S_1 x + S_2 x^2 + \cdots)(1 + \Lambda_1 x + \Lambda_2 x^2 + \cdots) \\ &= 1 + (S_1 + \Lambda_1)x + (S_2 + \Lambda_1 S_1 + \Lambda_2)x^2 \\ &\quad + (S_3 + \Lambda_1 S_2 + \Lambda_2 S_1 + \Lambda_3)x^3 + \cdots \\ &= 1 + \Omega_1 x + \Omega_2 x^2 + \cdots \end{aligned}$$

where $\Omega(x)$ is called the *error magnitude polynomial*, and is useful in nonbinary decoding. For now we will simply

[4] The general problem of finding solutions to systems of algebraic equations in multiple variables is NP-hard, and is the basis for a number of nice cryptosystems, including the Data Encryption Standard.

note that if the syndrome and error locator polynomials are to satisfy this expression, then the odd-indexed coefficients of $\Omega(x)$ must be zero (see the Newton's identities for the binary case presented above). Given that we know only the first $2t$ coefficients of $S(x)$, the decoding problem then becomes one of finding a polynomial $\Lambda(x)$ of degree less than or equal to t that satisfies

$$[1 + S(x)]\Lambda(x) \cdots (1 + \Omega_2 x^2 + \Omega_4 x^4 + \cdots + \Omega_{2t} x^{2t}) \bmod x^{2t+1}$$

Berlekamp's algorithm proceeds iteratively by breaking this equation) down into a series of smaller problems of the form

$$[1 + S(x)]\Lambda^{(2k)}(x) \cdots (1 + \Omega_2 x^2 + \Omega_4 x^4 + \cdots + \Omega_{2k} x^{2k}) \bmod x^{2k+1}$$

where k runs from 1 to t. A solution $\Lambda^{(0)}(x) = 1$ is first assumed and tested to see if it works for the case $k = 1$. If it does work, we proceed to $k = 2$; otherwise, a correction factor correction factor is computed and added to $\Lambda^{(0)}$, creating a new solution $\Lambda^{(2)}(x)$. The genius of the algorithm lies in the computation of the correction factor. It is designed so that the new solution will work not only for the current case but for all previous values of k as well. We first consider the binary case.

4.2. Berlekamp's Algorithm for Decoding Binary BCH Codes

1. Set the initial conditions: $k = 0$, $\Lambda^{(0)}(x) = 1$, $T^{(0)} = 1$.
2. Let $\Delta^{(2k)}$ be the coefficient of x^{2k+1} in the product $\Lambda^{(2k)}(x)[1 + S(x)]$.
3. Compute

$$\Lambda^{(2k+2)}(x) = \Lambda^{(2k)}(x) + \Delta^{(2k)}[x \cdot T^{(2k)}(x)]$$

4. Compute

$$T^{(2k+2)}(x) = \begin{cases} x^2 T^{(2k)}(x) & \text{if } \Delta^{(2k)} = 0 \text{ or if } \deg[\Lambda^{(2k)}(x)] > k \\ \dfrac{x\Lambda^{(2k)}(x)}{\Delta^{(2k)}} & \text{if } \Delta^{(2k)} \neq 0 \text{ and } \deg[\Lambda^{(2k)}(x)] \leq k \end{cases}$$

5. Set $k = k + 1$. If $k < t$, then go to 2.
6. Determine the roots of $\Lambda(x) = \Lambda^{(2t)}(x)$. If the roots are distinct and lie in the right field, then correct the corresponding locations in the received word and *stop*.
7. Declare a decoding failure and *stop*.

For an example, consider a narrow-sense double-error-correcting code of length 31 with generator polynomial $= 1 + x^3 + x^5 + x^6 + x^8 + x^9 + x^{10}$. Let the received vector and associated polynomial be as follows:

$$\mathbf{r} = (0010000110011000000000000000000)$$

$$\updownarrow$$

$$r(x) = x^2 + x^7 + x^8 + x^{11} + x^{12}$$

A bit of number crunching in GF(32) yields the following syndrome polynomial:

$$S(x) = \alpha^7 x + \alpha^{14} x^2 + \alpha^8 x^3 + \alpha^{28} x^4$$

Applying Berlekamp's algorithm, we obtain the following sequence of solutions.

k	$\Lambda^{(2k)}(x)$	$T^{(2k)}(x)$	$\Delta^{(2k)}$
0	1	1	α^7
1	$1 + \alpha^7 x$	$\alpha^{24} x$	α^{22}
2	$1 + \alpha^7 x + \alpha^{15} x^2$	—	—

$\Lambda^{(4)}(x) = 1 + x\alpha^7 x + \alpha^{15} x^2$ is the error locator polynomial. The error locators are $X_1 = \alpha^5$ and $X_2 = \alpha^{10}$, indicating errors at the fifth and tenth coordinates of $\mathbf{r}$. The corrected word, with the corrected positions underlined, is

$$\mathbf{c} = (0010010110\underline{1}11\underline{0}00000000000000000)$$

$$\updownarrow$$

$$c(x) = x^2 + x^5 + x^7 + x^8 + x^{10} + x^{11} + x^{12} = x^2 g(x)$$

For the nonbinary case, we first note that the syndromes are now a function of the magnitude of the errors as well as their locations. Assuming that some ν errors have corrupted the received word, the syndromes are as follows:

$$S_j = e(\alpha^j) = \sum_{k=0}^{n-1} e_k (\alpha^j)^k = \sum_{l=1}^{v} e_{i_l} X_l^j$$

This expression defines a series of $2t$ algebraic equations in 2ν unknowns:

$$\begin{aligned} S_1 &= e_{i_1} X_1 + e_{i_2} X_2 + \cdots + e_{i_v} X_v \\ S_2 &= e_{i_1} X_1^2 + e_{i_2} X_2^2 + \cdots + e_{i_v} X_v^2 \\ S_3 &= e_{i_1} X_1^3 + e_{i_2} X_2^3 + \cdots + e_{i_v} X_v^3 \\ &\vdots \\ S_{2t} &= e_{i_1} X_1^{2t} + e_{i_2} X_2^{2t} + \cdots + e_{i_v} X_v^{2t} \end{aligned}$$

We can reduce this system of equation to a set of linear functions in the unknown quantities. We first assume an error locator polynomial $\Lambda(x)$ whose zeros are the inverses of the error locators $\{X_i\}$:

$$\Lambda(x) = \prod_{l=1}^{v} (1 - X_l x) = \Lambda_v x^v + \Lambda_{v-1} x^{v-1} + \cdots + \Lambda_1 x + \Lambda_0$$

It follows that for some error locator X_l

$$\Lambda(X_l^{-1}) = \Lambda_v X_l^{-v} + \Lambda_{v-1} X_l^{-v+1} + \cdots + \Lambda_1 X_l^{-1} + \Lambda_0 = 0$$

Since the expression sums to zero, we can multiply through by a constant:

$$e_{i_1} X_l^j (\Lambda_v X_l^{-v} + \Lambda_{v-1} X_l^{-v+1} + \cdots + \Lambda_1 X_l^{-1} + \Lambda_0) =$$

$$e_{i_1} (\Lambda_v X_l^{-v+j} + \Lambda_{v-1} X_l^{-v+j+1} + \cdots + \Lambda_1 X_l^{j-1} + \Lambda_0 X_l^j) = 0$$

Now sum both sides of over all indices l, obtaining an expression from which Newton's identities can be constructed:

$$\sum_{l=1}^{v} e_{i_1}(\Lambda_v X_l^{j-v} + \Lambda_{v-1} X_l^{j-v+1} + \cdots + \Lambda_1 X_l^{j-1} + \Lambda_0 X_l^j)$$

$$= \Lambda_v \sum_{l=1}^{v} e_{i_1} X_l^{j-v} + \Lambda_{v-1} \sum_{l=1}^{v} e_{i_1} X_l^{j-v+1} + \cdots + \Lambda_1 \sum_{l=1}^{v} e_{i_1} X_l^{j-1}$$

$$+ \Lambda_0 \sum_{l=1}^{v} e_{i_1} X_l^j$$

$$= \Lambda_v S_{j-v} + \Lambda_{v-1} S_{j-v+1} + \cdots + \Lambda_1 S_{j-1} + \Lambda_0 S_j = 0$$

From the earlier expressions it is clear that Λ_0 is always one. The preceding equation can thus be reexpressed as

$$\Lambda_v S_{j-v} + \Lambda_{v-1} S_{j-v+1} + \cdots + \Lambda_1 S_{j-1} = -S_j$$

This expression shows that the syndrome S_j can be expressed in recursive form as a function of the coefficients of the error locator polynomial $\Lambda(x)$ and the earlier syndromes $S_{j-1}, \ldots, S_{j-v}$. In 1969 Massey showed that this expression can be given a physical interpretation through the use of a linear feedback shift register (LFSR) [23]. The double-lined elements in Fig. 2 denote storage of and operations on nonbinary field elements.

Part of the problem of decoding BCH and Reed–Solomon codes can be reexpressed as follows. Find an LFSR of minimal length such that the first $2t$ elements in the LFSR output sequence are the syndromes $S_1, S_2, \ldots, S_{2t}$. The taps of this shift register provide the desired error locator polynomial $\Lambda(x)$.

Let $\Lambda^{(k)}(x) = \Lambda_k x^j + \Lambda_{k-1} x^{k-1} + \cdots + \Lambda_1 x + 1$ be the *connection polynomial* of length k whose coefficients specify the taps of a length-k LFSR. Massey's construction of Berlekamp's algorithm starts by finding $\Lambda^{(1)}$ such that the first element output by the corresponding LFSR is the first syndrome S_1. The second output of this LFSR is then compared to the second syndrome. If the two do not have the same value, then the *discrepancy* between the two is used to construct a modified connection polynomial. If there is no discrepancy, then the same connection polynomial is used to generate a third sequence element, which is compared to the third syndrome. The process continues until a connection polynomial is obtained that specifies an LFSR capable of generating all $2t$ elements of the syndrome sequence.

Massey showed that, given an error pattern of weight $\leq t$, the connection polynomial resulting from the Berlekamp algorithm uniquely specifies the correct error locator polynomial.

The algorithm has four basic parameters: the connection polynomial $\Lambda^{(k)}(x)$, the correction polynomial $\mathrm{T}(x)$, the discrepancy $\Delta^{(k)}$, the length L of the shift register, and the indexing variable k. The algorithm proceeds as follows.

4.3. The Berlekamp–Massey Shift Register Synthesis Decoding Algorithm

1. Compute the syndrome sequence $S_1, \ldots, S_{2t}$ for the received word.
2. Initialize the algorithm variables as follows:

$$k = 0, \Lambda^{(0)}(x) = 1, \quad L = 0, \quad \text{and} \quad \mathrm{T}(x) = x$$

3. Set $k = k + 1$. Compute the discrepancy $\Delta^{(k)}$ by subtracting the kth output of the LFSR defined by $\sigma^{(k-1)}(x)$ from the kth syndrome:

$$\Delta^{(k)} = S_k - \sum_{i=1}^{L} \Lambda_i^{(k-1)} S_{k-i}$$

4. If $\Delta^{(k)} = 0$, then go to step 8.
5. Modify the connection polynomial: $\Lambda^{(k)}(x) = \Lambda^{(k-1)}(x) - \Delta^{(k)}\mathrm{T}(x)$
6. If $2L^3 k$, then go to step 8.
7. Set $L = k - L$ and $T(x) = \Lambda^{(k-1)}(x)/\Delta^{(k)}$.
8. Set $T(x) = x \cdot T(x)$
9. If $k < 2t$, then go to step 3.
10. Determine the roots of $\Lambda(x) = \Lambda^{(2t)}(x)$. If the roots are distinct and lie in the right field, then determine the error magnitudes, correct the corresponding locations in the received word, and *stop*.
11. Declare a decoding failure and *stop*.

The Berlekamp–Massey algorithm allows us to find the error locator polynomial, but there remains the problem of finding the error magnitudes. Forney [22]. showed that the following expression will do the trick:

$$e_{i_k} = \frac{-X_k \Omega(X_k^{-1})}{\Lambda'(X_k^{-1})}$$

Consider the following example of double-error correction [1]. using the Berlekamp–Massey algorithm and a (7,3) Reed–Solomon code (from Wicker [1]). Let the received polynomial be $r(x) = \alpha^2 x^6 + \alpha^2 x^4 + x^3 + \alpha^5 x^2$, giving the syndrome sequence $S_1 = \alpha^6$, $S_2 = \alpha^3$, $S_3 = \alpha^4$, $S_4 = \alpha^3$. The algorithm generates the following set of connection polynomials, discrepancies, and correction polynomials [the last column, $T(x)$, Follows at conclusion of step 8:

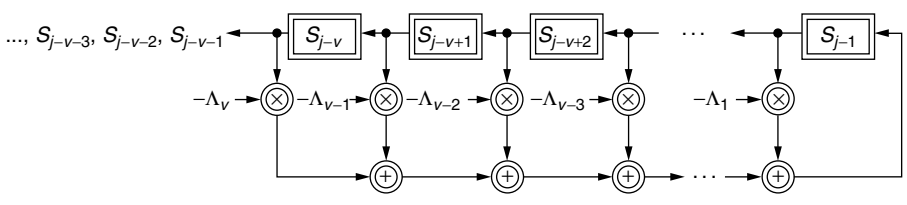


Figure 2. LFSR interpretation.

k	S_k	$\Lambda^{(k)}(x)$	$\Delta^{(k)}$	L	$T(x)$
0	—	1	—	0	x
1	α^6	$1+\alpha^6 x$	$S_1 - 0 = \alpha^6$	1	αx
2	α^3	$1+\alpha^4 x$	$S_2 - \alpha^5 = \alpha^2$	1	αx^2
3	α^4	$1+\alpha^4 x+\alpha^6 x^2$	$S_3 - 1 = \alpha^5$	2	$\alpha^2 x+\alpha^6 x^2$
4	α^3	$1+\alpha^2 x+\alpha x^2$	$S_4 - \alpha^4 = \alpha^6$	—	—

We obtain the error locator polynomial $\Lambda(x) = 1 + \alpha^2 x + \alpha x^2$.

The LFSRs corresponding to the connection polynomials $\Lambda^{(1)}(x)$ through $\Lambda^{(4)}(x)$ are drawn in Fig. 3, along with the initial conditions and the generated output sequence. Consider the LFSR with connection polynomial $\Lambda^{(3)}(x)$. This LFSR correctly generates the first three syndromes, but its fourth output, α^4, is not equal to $S_4 = \alpha^3$. The discrepancy $\Delta^{(4)} = \alpha^6$ is the difference between the two values, as shown in Fig. 3. This discrepancy is used to determine the connection polynomial $\Lambda^{(4)}(x)$, which, as shown in Fig. 3, correctly generates all four syndromes.

The syndrome sequence provides the syndrome polynomial $S(x) = \alpha^6 x + \alpha^3 x^2 + \alpha^4 x^3 + \alpha^3 x^4$. We used the Berlekamp–Massey algorithm to obtain the error locator polynomial $\Lambda(x) = 1 + \alpha^2 x + \alpha x^2$. We can now compute the error magnitude polynomial:

$$\begin{aligned}\Omega(x) &\equiv \Lambda(x)[1 + S(x)] \bmod x^{2t+1} \\ &\equiv (1 + \alpha^{2x} + \alpha^{x^2})(1 + \alpha^{6x} + \alpha^{3x^2} + \alpha^{4x^3} + \alpha^{3x^4}) \bmod x^5 \\ &\equiv (1 + x + \alpha^{3x^2}) \bmod x^5\end{aligned}$$

The errors locators were found to be $X_1 = \alpha^3$ and $X_2 = \alpha^5$ in the first part of this example. Using the Forney algorithm, the error magnitudes are found to be

$$\begin{aligned}e_{i_x k} &= \frac{-X_k \Omega(X_k^{-1})}{\Lambda'(X_k^{-1})} = \frac{-X_k[1 + X_k^{-1} + \alpha^3 X_k^{-2}]}{\alpha^2} \\ &= \alpha^5 X_k + \alpha^5 + \alpha X_k^{-1}\end{aligned}$$

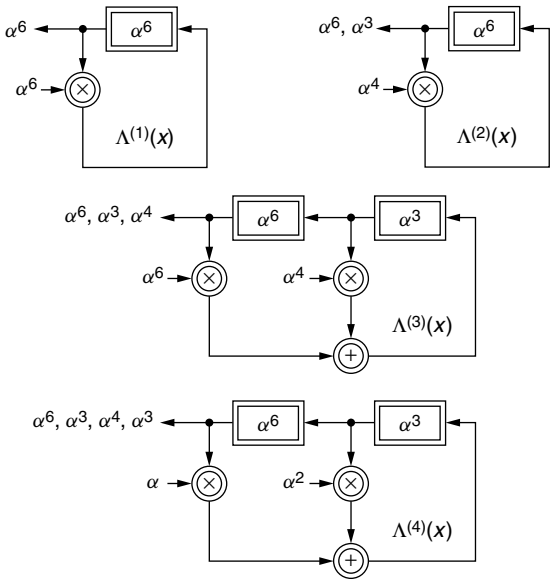


Figure 3. Several different LFSR syndromes.

$$e_3 = \alpha^5 \alpha^3 + \alpha^5 + \alpha\alpha^4 = \alpha$$

$$e_5 = \alpha^5 \alpha^5 + \alpha^5 + \alpha\alpha^2 = \alpha^5$$

The error polynomial is thus $e(x) = \alpha x^3 + \alpha^5 x^5$.

5. APPLICATIONS

The applications of cyclic codes are legion. In this final section we will focus on two of the more interesting applications: digital audio and deep-space telecommunications.

5.1. The Compact-Disk (CD) Player

The most ubiquitous application of cyclic codes (or possibly any error control codes) lies in the CD player. The channel in a CD playback system consists of a transmitting laser, a recorded disk, and a photodetector. Assuming that the player is working properly, the primary contributor to errors on this channel is the contamination of the surface of the disk (e.g., fingerprints and scratches). As the surface contamination surface contamination affects an area that is usually quite large compared to the surface used to record a single bit, channel errors occur in bursts when the disk is played. As we shall see, the CD error control system handles bursts through cross-interleaving and through the burst error-correcting capability of Reed–Solomon codes.

Figure 4 shows the various stages through which music is processed on its way to being recorded on a disk. Each channel is sampled 44,100 times per second, allowing accurate reproduction of all frequencies up to 22 kHz. Each sample is then converted into digital form by a 16-bit analog-to-digital converter.

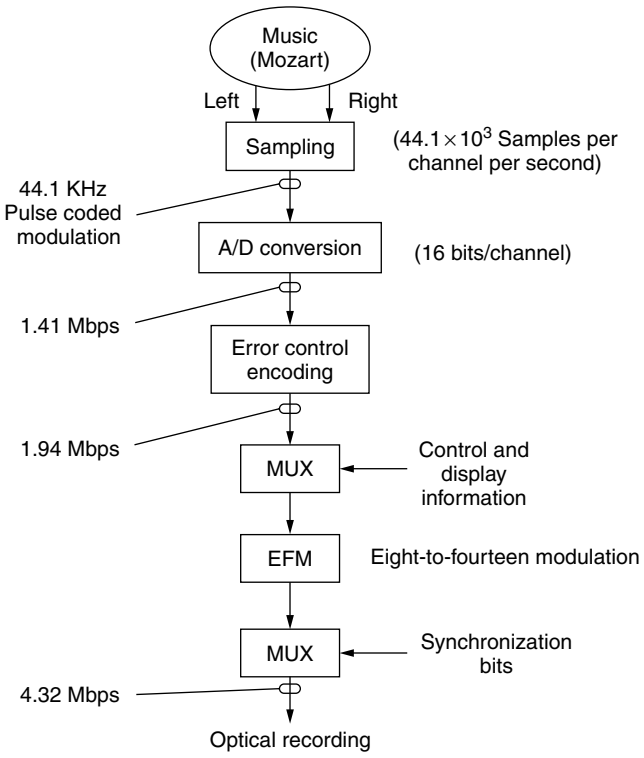


Figure 4. Data processing in the creation of a compact disk [1].

The output of the A/D converter forms a 1.41-Mbps (megabits per second) datastream, which is passed directly to the CIRC encoder. The CIRC encoder, as shown in Fig. 4, uses two shortened Reed–Solomon codes, $\mathbf{C}_1$ and $\mathbf{C}_2$. Both codes uses 8-bit symbols from the code alphabet GF(256). This provides for a nice match with the 16-bit samples emerging from the A/D converter. The "natural" length of the RS code over GF(256) is 255, which would lead to 2040-bit codewords and a relatively complicated decoder. It should be remembered that the decoder will reside in the retail player, and it is extremely important that its cost be minimized. The codes are thus shortened significantly: $\mathbf{C}_1$ is a (32,28) code and $\mathbf{C}_2$ is a (28,24) code. Both have redundancy 4 and minimum distance 5.

Each 16-bit sample is treated as a pair of symbols from GF(256). The samples are encoded 12 at a time by the $\mathbf{C}_2$ encoder to create a 28-symbol codeword. The 28 symbols in each $\mathbf{C}_2$ codeword are then passed through a cross-interleaver ($m = 28$, $D = 4$ symbols) before being encoded once by the $\mathbf{C}_1$ encoder. The resulting 32-symbol $\mathbf{C}_1$ codewords are then processed as shown in the figure above.

The CIRC encoding process for the CD system is standard; no matter where you buy your CDs (standard size), they will play on any CD player. However, the CIRC decoding process is not standardized and can vary from player to player [28]. This was done intentionally to allow the manufacturers to experiment with various designs and to speed the player to market. The basic building blocks of the decoder are shown in the figure. The $\mathbf{C}_1$ decoder is followed by a cross-deinterleaver and a $\mathbf{C}_2$ decoder. Since both codes have minimum distance 5, they can be used to correct all error patterns of weight

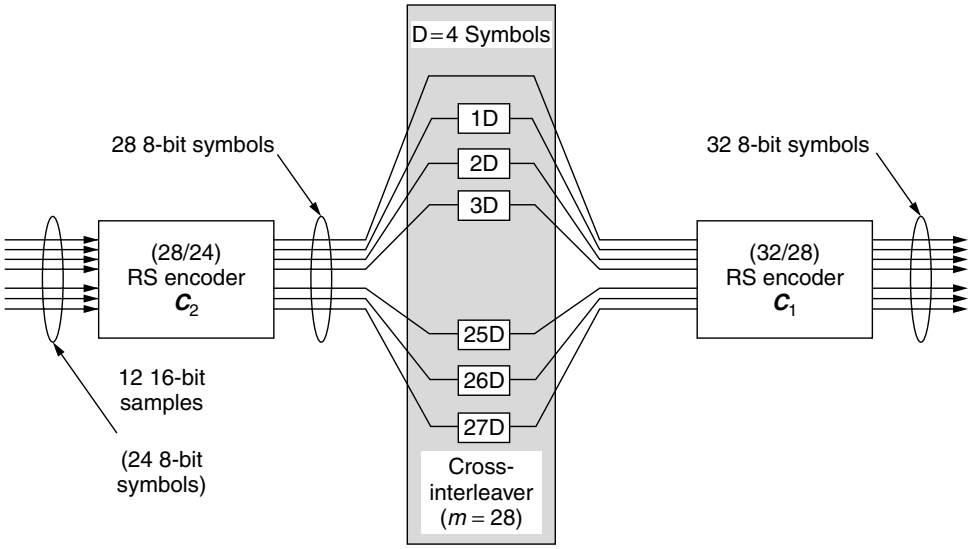


Figure 5. Cross-interleaved encoding [1].

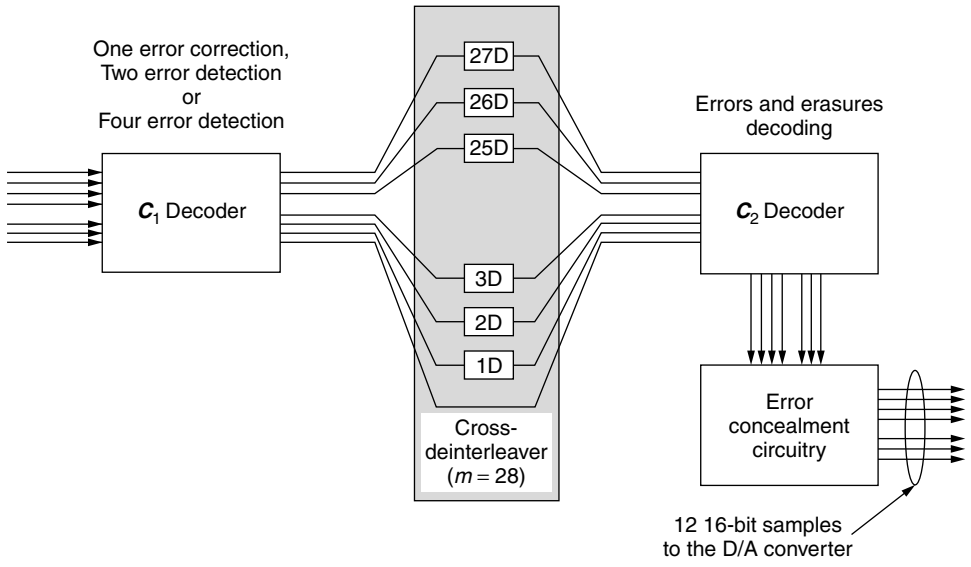


Figure 6. Basic structure of a CIRC decoder [1].

2 or less. Even when the full error correcting capacity of an RS code is used, there remains a significant amount of error detecting capacity. It is also the case that some of the error correction capacity of the RS code can be exchanged for an increase in error detection capacity and a substantial improvement in reliability performance. Most CIRC decoders take advantage of both of these principles. The $\mathbf{C}_1$ decoder is set to correct all single-error patterns. When the $\mathbf{C}_1$ decoder sees a higher-weight error pattern (double-error patterns and any pattern causing a decoder failure), the decoder outputs 28 erased symbols. The cross-deinterleaver spreads these erasures over 28 $\mathbf{C}_2$ codewords. $\mathbf{C}_1$ may also be set to correct double error patterns, or, at the other extreme, may simply be used to declare erasures (the least expensive implementation).

The $\mathbf{C}_2$ decoder can correct any combination of e errors and s erasures, where $2e + s < 5$. It is generally designed to decode erasures only (again, an inexpensive solution) due to the small probability of a $\mathbf{C}_1$ decoder error. Whenever the number of erasures exceeds 4, the $\mathbf{C}_2$ decoder outputs 24 erasures, which corresponds to 12 erased music samples. The error concealment circuitry error concealment circuitry responds by muting these samples or by interpolating values through the use of correct samples adjacent to the correct samples. A number of additional interleaving and delay operations are included in the encoding and decoding operations in order to enhance the operation of the error concealment circuitry. For example, samples adjacent in time are further separated by additional interleavers to improve the impact of interpolation [28].

5.2. Deep-Space Telecommunications

The earliest use of a cyclic code in deep-space telecommunications was in conjunction with the *Mariner* mission in 1971. *Mariner* included an infrared interferometer spectrometer (IRIS) [20,29]. The data from this instrument required a bit error rate on the order of 10^{-5}, and thus needed protection beyond that provided by the then standard Reed–Muller-based system. It was decided to precode the IRIS data using a [6,4] Reed-Solomon code with symbols from GF(2^6), thus creating the first concatenated system designed for use in deep-space telecommunications. The RS outer code was applied only to the IRIS data, thus keeping the overall code rate of the telemetry channel quite close to that of the Reed–Muller system by itself. This use of concatenated schemes to provide selective error protection was revisited in the development of the *Voyager* mission.

The *Voyager* 1 and 2 spacecraft were launched toward Saturn and Jupiter, respectively, in the summer of 1977. They carried a number of scientific instruments and imaging systems for a mission that was to be the most successful in the history of the exploration of deep space. The *Voyager* spacecraft were originally intended to explore Jupiter, Saturn, and their moons. The distances involved were substantial: Jupiter and Saturn are 483 million miles and 870 million miles from the sun, respectively. During the Jupiter flyby, *Voyager* 2 transmitted a 21.3-W signal through an antenna with $G_T = 6.5 \times 10^4$ (48.1 dB). By the time the signal was recovered by a 64-m dish on the earth, the signal measured 3.04×10^{-16} W. The telemetry signal consisted of two basic types of data: imaging and general science and engineering (GSE). The color images were reconstructed from three (800×800) arrays of 8-bit pixels, giving a total of 15,360,000 bits per image [29]. The uncompressed images required a nominal bit error rate of 5×10^{-3}. It was found that the rate-$\frac{1}{2}$ Planetary Standard was sufficient to provide the desired level of reliability. The GSE data, however, required a much lower bit error rate than the imaging data. The configuration in Fig. 7 was adopted. The GSE data was first encoded using the extended (24,12) Golay code discussed earlier in this article. The GSE and imaging data were then both encoded using the rate-$\frac{1}{2}$ Planetary Standard. At a 2.3 dB

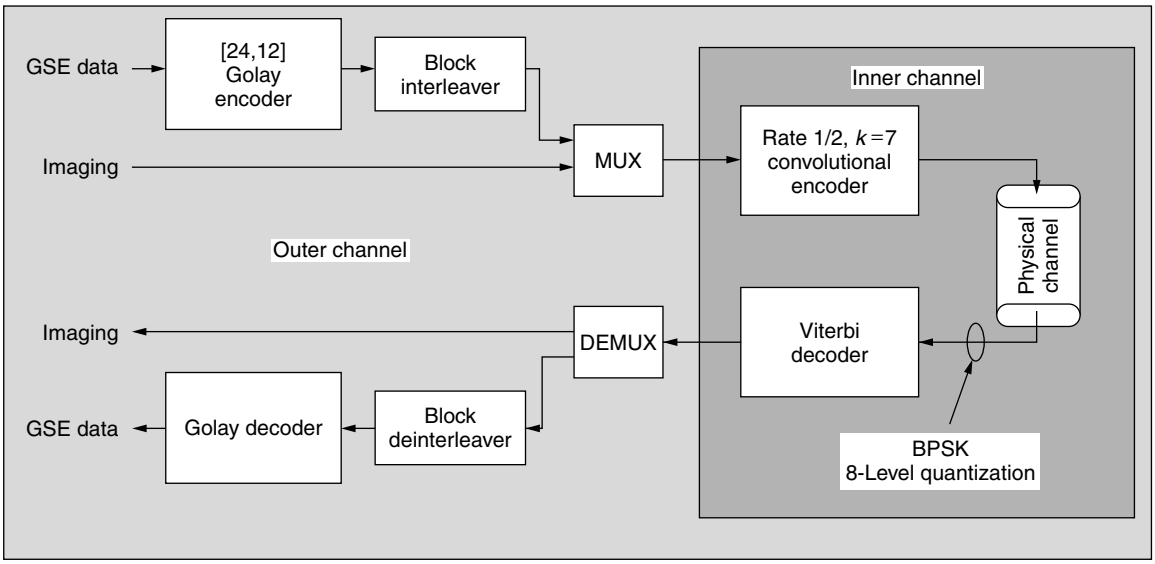


Figure 7. Convolutional/Viterbi and Golay concatenated system for the *Voyager* spacecraft (Jupiter and Saturn flybys) [20].

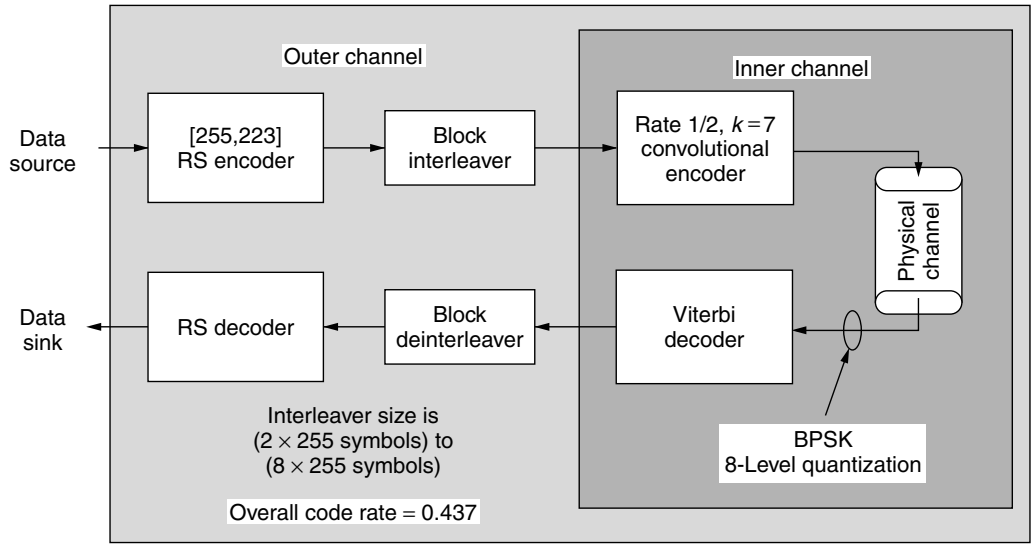


Figure 8. CCSDS Standard for deep-space telemetry links [20].

E_b/N_0, the convolutional code provided an inner channel bit error rate of 5×10^{-3}. The Golay code provides an outer channel bit error rate of less than 1×10^{-5}.

When the *Voyager* mission was extended to cover several of the outer planets, it was necessary to increase the level of error protection. As with the GSE data in the Jupiter and Saturn flybys, the additional reliability requirement was handled through the use of a concatenated system. In this case, however, the extremely low bit error rate requirement and the sensitivity of the link budget to a reduction in code rate pointed towards the use of Reed–Solomon technology for the outer code. The resulting concatenated system later became the basis for the CCSDS standard for deep-space telemetry links [30]. This standard has been used extensively by both NASA and the European Space Agency. In the CCSDS standard, the rate-$\frac{1}{2}$ Planetary Standard is joined by a (255,223) RS outer code with symbols from GF(2^8), as shown in Fig. 8.

BIBLIOGRAPHY

1. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
2. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
3. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1977.
4. E. Prange, *Cyclic Error-Correcting Codes in Two Symbols*, Air Force Cambridge Research Center Report TN-57-103, Cambridge, MA, Sept. 1957.
5. E. Prange, *Some Cyclic Error-Correcting Codes with Simple Decoding Algorithms*, Air Force Cambridge Research Center Report TN-58-156, Cambridge, MA, April 1958.
6. E. Prange, *The Use of Coset Equivalence in the Analysis and Decoding of Group Codes*, Air Force Cambridge Research Center Report TR-59-164, Cambridge, MA, 1959.
7. I. S. Reed (1990).
8. I. S. Reed (1992).
9. Shannon (1948).
10. Golay (1949).
11. Kasami (1964).
12. A. Hocquenghem, Codes correcteurs d'erreurs, *Chiffres* **2**: 147–156 (1959).
13. R. C. Bose and D. K. Ray-Chaudhuri, On a class of error correcting binary group codes, *Inform. Control* **3**: 68–79 (March 1960).
14. R. C. Bose and D. K. Ray-Chaudhuri, Further results on error correcting binary group codes, *Inform. Control* **3**: 279–290 (Sept. 1960).
15. W. W. Peterson, Encoding and error-correction procedures for the Bose–Chaudhuri codes, *IRE Trans. Inform. Theory* **IT-6**: 459–470 (Sept. 1960).
16. D. Gorenstein and N. Zierler, A class of error correcting codes in p^m symbols, *J. Soc. Indust. Appl. Math.* **9**: 207–214 (June 1961).
17. I. S. Reed and G. Solomon, Polynomial codes over certain finite fields, *SIAM J. Appl. Math.* **8**: 300–304 (1960).
18. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968 (rev. ed. Aegean Park Press, Laguna Hills, CA, 1984).
19. Hamming (1950).
20. S. B. Wicker, Deep space applications, in V. Pless and W. C. Huffman, eds., *Handbook of Coding Theory*, Elsevier, Amsterdam, 1998.
21. R. T. Chien, Cyclic decoding procedure for the Bose–Chaudhuri–Hocquenghem codes, *IEEE Trans. Inform. Theory* **IT-10**: 357–363 (Oct. 1964).
22. G. D. Forney, On decoding BCH codes, *IEEE Trans. Inform. Theory* **IT-11**: 549–557 (Oct. 1965).
23. J. L. Massey, Shift register synthesis and BCH decoding, *IEEE Trans. Inform. Theory* **IT-15**(1): 122–127 (Jan. 1969).

24. Y. Sugiyama, Y. Kasahara, S. Hirasawa, and T. Namekawa, A method for solving key equation for Goppa codes, *Inform. Control* **27**: 87–99 (1975).

25. Reed (1978).

26. W. C. Gore, Transmitting binary symbols with Reed–Solomon Codes, *Proc. Princeton Conf. Information Science and Systems*, Princeton, NJ, 1973, pp. 495–497.

27. R. E. Blahut, Transform techniques for error control codes, *IBM J. Res. Devel.* **23**: 299–315 (1979).

28. K. A. S. Immink, RS codes and the compact disc, in S. B. Wicker and V. K. Bhargava, eds., *Reed Solomon Codes and Their Applications*, IEEE Press, Piscataway, NJ, 1994.

29. R. J. McEliece and L. Swanson, Reed–Solomon codes and the exploration of the solar system, in S. B. Wicker and V. K. Bhargava, eds., *Reed–Solomon Codes and Their Applications*, IEEE Press, Piscataway, NJ, 1994, pp. 25–40.

30. Consultative Committee for Space Data Systems, *Recommendations for Space Data Systems Standard: Telemetry Channel Coding*, Blue Book Issue 2, CCSDS 101.0-B2, Jan. 1987.

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 2

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 2

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at
http://www.mrw.interscience.wiley.com/eot

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor.
p. cm.
includes index.
ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G.
TK5102 .W55 2002
621.382′03 — dc21

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

D

DATA COMPRESSION

JOHN KIEFFER
University of Minnesota
Minneapolis, Minnesota

1. INTRODUCTION

A modern-day data communication system must be capable of transmitting data of all types, such as text, speech, audio, image or video data. The block diagram in Fig. 1 depicts a data communication system, consisting of *source, encoder, channel*, and *decoder*:

The source generates the data sequence that is to be transmitted through the data communication system. The encoder converts the data sequence into a binary codeword for transmission through the channel. The decoder generates a reconstructed data sequence that may or not be equal to the original data sequence. The encoder/decoder pair in Fig. 1 is the *code* of the data communication system. In Fig. 1, the source and channel are fixed; the choice of code is flexible, in order to accomplish the twin goals of *bandwidth efficiency* and *reliable transmission*, described as follows:

1. *Bandwidth efficiency* — the portion of the available channel bandwidth that is allocated in order to communicate the given data sequence should be economized.
2. *Reliable transmission* — the reconstructed data sequence should be equal or sufficiently close to the original data sequence.

Unfortunately, these are conflicting goals; less use of bandwidth makes for less reliable transmission, and conversely, more reliable transmission requires the use of more bandwidth. It is the job of the data communication system designer to select a code that will yield a good tradeoff between these two goals. Code design is typically done in one of the following two ways.

1. *Separated Code Design*. Two codes are designed, a *source code* and a *channel code*, and then the source code and the channel code are cascaded together. Figure 2 illustrates the procedure. The source code is the pair consisting of the source encoder and source decoder; the channel code is the (channel encoder, channel decoder) pair. The source code achieves the goal of bandwidth efficiency: The source encoder removes a large amount of redundancy from the data sequence that can be restored (or approximately restored) by the source decoder. The channel code achieves the goal of reliable transmission: The channel encoder inserts a small amount of redundancy in the channel input stream that will allow the channel decoder to correct the transmission errors in that stream that are caused by the channel.

2. *Combined Code Design*. One code (as in Fig. 1) is designed to accomplish the twin goals of bandwidth efficiency and reliable transmission. Clearly, combined code design is more general than separated code design. However, previously separated codes were preferred to combined codes in data communication system design. There were two good reason for this: (a) Claude Shannon showed that if the data sequence is sufficiently long, and if the probabilistic models for the source and the channel are sufficiently simple, then there is no loss in the bandwidth versus reliability tradeoff that is achievable using separated codes of arbitrary complexity instead of combined codes of arbitrary complexity; and (b) the code design problem is made easier by separating it into the two decoupled problems of source code design and channel code design. For the communication of short data sequences, or for the scenario in which the complexity of the code is to be constrained, there can be an advantage to using combined codes as opposed to separated codes; consequently, there much attention has focused on the combined code design problem since the mid-1980s. At the time of the writing of this article, however, results on combined code design are somewhat isolated and have not yet been combined into a nice theory. On the other hand, the two separate theories of source code design and channel code design are well developed. The purpose of the present article is to provide an introduction to source code design.

In source code design, one can assume that the communication channel introduces no errors, because the purpose of the channel code is to correct whatever channel errors occur. Thus, we may use Fig. 3 below, which contains no channel, as the conceptual model guiding source code design.

The system in Fig. 3 is called a *data compression system* — it consists of the source and the source code consisting of the (source encoder, source decoder) pair. The data sequence generated by the source is random and is denoted X^n; the notation X^n is a shorthand for the following random sequence of length n:

$$X^n = (X_1, X_2, \ldots, X_n) \tag{1}$$

The X_i values $(i = 1, 2, \ldots, n)$ are the individual *data samples* generated by the source. In Fig. 3, B^K is a

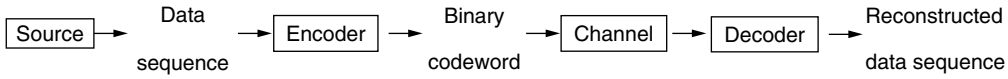


Figure 1. Block diagram of data communication system.

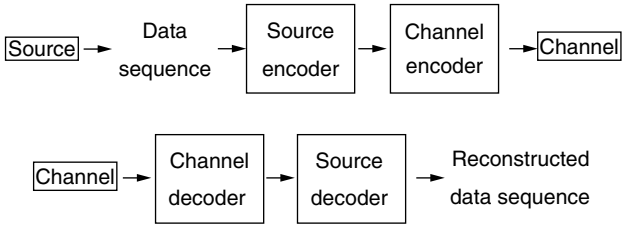


Figure 2. Data communication system with separated code.

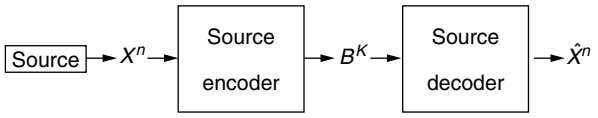


Figure 3. Block diagram of data compression system.

binary codeword generated by the source encoder as a deterministic function of X^n; B^K is random, and its distribution can be computed from the distribution of X^n. Our notational convention means that B^K is a shorthand for

$$B^K = (B_1, B_2, \ldots, B_K) \tag{2}$$

where each $B_i (i = 1, 2, \ldots, K)$, is a *code bit* belonging to the binary alphabet $\{0, 1\}$. The fact that K is capitalized means that K is random; that is, a *variable-length* codeword is used. The reconstructed data sequence generated by the source decoder in Fig. 3 is of length n and has been denoted $\hat{X}^n$. The sequence $\hat{X}^n$ is a deterministic function of B^K, and therefore also a deterministic function of X^n; $\hat{X}^n$ is random and its distribution can be computed from the distribution of X^n. Notationally

$$\hat{X}^n = (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n) \tag{3}$$

where each $\hat{X}_i (i = 1, 2, \ldots, n)$ is an approximation to X_i.

1.1. Mathematical Description of Source

We need to give a formal mathematical model describing the probabilistic nature of the source in Fig. 3. Accordingly, a source is defined to be a triple $[n, A_n, P_n]$, where

- n is a positive integer that is the length of the random data sequence X^n generated by the source $[n, A_n, P_n]$.
- A_n is the set of all sequences of length n which are realizations of X^n. (The set A_n models the set of all possible deterministic sequences that could be processed by the data compression system driven by the source $[n, A_n, P_n]$.)
- P_n denotes the probability distribution of X^n; it is a probability distribution on A_n. We have

$$\Pr[X^n \in S_n] = P_n(S_n), S_n \subset A_n$$
$$\Pr[X^n \in A_n] = P_n(A_n) = 1$$

The *alphabet* of the source $[n, A_n, P_n]$ is the smallest set A such that $A_n \subset A^n$, where A^n denotes the set of all sequences of length n whose entries come from A. For a fixed positive integer n, a source $[n, A_n, P_n]$ shall be referred to as an nth-*order source*.

1.2. Memoryless Source

The most common type of source model is the *memoryless source*. In an nth-order memoryless source, the data samples $X_1, X_2, \ldots, X_n$ are taken to be independent, identically distributed random variables. Therefore, the joint probability density function $f(x_1, x_2, \ldots, x_n)$ of the memoryless source output X^n factors as

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_1(x_2) \cdots f_1(x_n)$$

where f_1 is a fixed probability density function.

1.3. Markov Source

The second most common type of source model is the *stationary Markov source*. For an nth-order source, the stationary Markov source assumption means that the joint probability density function $f(x_1, x_2, \ldots, x_n)$ of the source output X^n factors as

$$f(x_1, x_2, \ldots, x_n) = \frac{f_2(x_1, x_2) f_2(x_2, x_3) \cdots f_2(x_{n-1}, x_n)}{f_1(x_2) f_1(x_3) \cdots f_1(x_{n-1})}$$

where f_2 is a fixed 2D (two-dimensional) probability density function, and f_1 is a 1D probability density function related to f_2 by

$$f_1(x_1) = \int_{-\infty}^{\infty} f_2(x_1, x_2)\, dx_2 = \int_{-\infty}^{\infty} f_2(x_2, x_1)\, dx_2$$

1.4. Lossless and Lossy Compression Systems

Two types of data compression systems are treated in this article: *lossless compression systems* and *lossy compression systems*. In a lossless compression system, the set A_n of possible data sequence inputs to the system is finite, the encoder is a one-to-one mapping (this means that there is a one-to-one correspondence between data sequences and their binary codewords), and the decoder is the inverse of the encoder; thus, in a lossless compression system, the random data sequence X^n generated by the source and its reconstruction $\hat{X}^n$ at the decoder output are the same:

$$\Pr[X^n = \hat{X}^n] = 1$$

In a lossy compression system, two or more data sequences in A_n are assigned the same binary codeword, so that

$$\Pr[X^n \neq \hat{X}^n] > 0$$

Whether one designs a lossless or lossy compression system depends on the type of data that are to be transmitted in a data communication system. For example, for textual data, lossless compression is used because one typically wants perfect reconstruction of the transmitted text; on the other hand, for image data, lossy compression would be appropriate if the reconstructed image is required only to be perceptually equivalent to the original image.

This article is divided into two halves. In the first half, we deal with the design of lossless codes, namely, source codes for lossless compression systems. In the second half,

design of lossy codes (source codes for lossy compression systems) is considered.

2. LOSSLESS COMPRESSION METHODOLOGIES

In this section, we shall be concerned with the problem of designing lossless codes. Figure 4 depicts a general lossless compression system. In Fig. 4, the pair consisting of source encoder and source decoder is called a *lossless code*. A lossless code is completely determined by its source encoder part, since the source decoder is the inverse mapping of the source encoder.

Let $[n, A_n, P_n]$ be a given source with finite alphabet. When a lossless code is used to compress the data generated by the source $[n, A_n, P_n]$, a lossless compression system S_n results as in Fig. 4. The effectiveness of the lossless code is then evaluated by means of the figure of merit

$$R(S_n) \triangleq n^{-1} \sum_{x^n \in A_n} P_n(x^n) K(x^n)$$

where $K(x^n)$ is the length of the codeword assigned by the lossless code to the data sequence $x^n \in A_n$. The figure of merit $R(S_n)$ is called *compression rate* and its units are "code bits per data sample." An efficient lossless code for compressing data generated by the source $[n, A_n, P_n]$ is a lossless code giving rise to a compression system S_n for which the compression rate $R(S_n)$ is minimized or nearly minimized. In this section, we put forth various types of lossless codes that are efficient in this sense.

In lossless code design, we make the customary assumption that the codewords assigned by a lossless code must satisfy the *prefix condition*, which means that no codeword is a prefix of any other codeword. If $K_1, K_2, \ldots, K_j$ are the lengths of the codewords assigned by a lossless code, then *Kraft's inequality*

$$2^{-K_1} + 2^{-K_2} + \cdots + 2^{-K_j} \leq 1 \tag{4}$$

must hold. Conversely, if positive integers $K_1, K_2, \ldots, K_j$ obey Kraft's inequality, then there exists a lossless code whose codewords have these lengths. In this case, one can build a rooted tree T with j leaves and at most two outgoing edges per internal vertex, such that $K_1, K_2, \ldots, K_j$ are the lengths of the root-to-leaf paths; the codewords are obtained by labeling the edges along these paths with 0s and 1s. The tree T can be found by applying the Huffman algorithm (covered in Section 2.2) to the set of probabilities $2^{-K_1}, 2^{-K_2}, \ldots, 2^{-K_j}$.

There are two methods for specifying a lossless code for the source $[n, A_n, P_n]$: (1) an encoding table can be given which lists the binary codeword to be assigned to each sequence in A_n (decoding is then accomplished by using the encoding table in reverse), or (2) encoding and decoding algorithms can be given that indicate how to compute the binary codeword for each sequence in $x^n \in A_n$ and how to compute x^n from its codeword. We will specify each lossless code discussed in this section using either method 1 or 2, depending on which method is more convenient. Method 1 is particularly convenient if the codeword lengths are known in advance, since, as pointed out earlier, a tree can be constructed that yields the codewords. Method 2 is more convenient if the data length n is large (which makes the storing of an encoding table impractical).

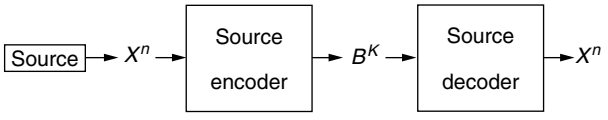


Figure 4. Lossless data compression system.

2.1. Entropy Bounds

It is helpful to understand the entropy upper and lower bounds on the performance of lossless codes. With these bounds, one can determine before designing a lossless code what kind of performance it is possible for such a code to achieve, as well as what kind of performance it is not possible to achieve.

The entropy of the source $[n, A_n, P_n]$ is defined by

$$H_n \triangleq \sum_{x^n \in A_n} (-\log_2 P_n(x^n)) P_n(x^n)$$

Suppose that the random data sequence generated by the source $[n, A_n, P_n]$ is compressed by an arbitrary lossless code and let S_n be the resulting lossless compression system (as in Fig. 4). Then, the compression rate is known to satisfy the relationship

$$R(S_n) \geq \frac{H_n}{n} \tag{5}$$

Conversely, it is known that there exists at least one lossless code for which

$$R(S_n) \leq \frac{H_n + 1}{n} \tag{6}$$

Assume that the data length n is large. We can combine the bounds (5) and (6) to assert that a lossless code is efficient if and only if the resulting compression rate satisfies

$$R(S_n) \approx \frac{H_n}{n}$$

The quantity H_n/n is called the *entropy rate* of the source $[n, A_n, P_n]$. Our conclusion is that a lossless code is efficient for compressing source data if and only if the resulting compression rate is approximately equal to the entropy rate of the source. For the memoryless source and the Markov source, this result can be sharpened, as the following discussion shows.

2.1.1. Efficient Lossless Codes for Memoryless Sources.

Assume that the given source $[n, A_n, P_n]$ is a memoryless source; let A be the finite source alphabet. There is a probability mass function $[p(a): a \in A]$ on A such that

$$P_n(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i),\ (x_1, x_2, \ldots, x_n) \in A_n \tag{7}$$

Let H_0 be the number

$$H_0 \triangleq \sum_{a \in A} (-\log_2 p(a)) p(a) \tag{8}$$

It is easy to show from (7) that the source entropy satisfies

$$H_n = nH_0$$

and therefore H_0 is the entropy rate of the source $[n, A_n, P_n]$. Assuming that the data length n is large, we conclude that a lossless code for compressing the data generated by the memoryless source $[n, A_n, P_n]$ is efficient if and only if the resulting compression rate is approximately equal to H_0.

2.1.2. Efficient Lossless Codes for Markov Sources. Now assume that the source $[n, A_n, P_n]$ is a stationary Markov source; let A be the finite source alphabet. There is a probability mass function $[p(a): a \in A]$ on A, and a nonnegative matrix $[\pi(a_1, a_2): a_1, a_2 \in A]$ whose rows each sum to one such that both of the following are true:

$$p(a_2) = \sum_{a_1 \in A} p(a_1)\pi(a_1, a_2), \ a_2 \in A \tag{9}$$

$$P_n(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} \pi(x_{i-1}, x_i), \ (x_1, x_2, \ldots, x_n) \in A_n \tag{10}$$

Let H_0 be the number (8) and H_1 be the number

$$H_1 \triangleq \sum_{a_1 \in A} \sum_{a_2 \in A} (-\log_2 \pi(a_1, a_2)) p(a_1)\pi(a_1, a_2)$$

It can be shown from these last two equations that

$$H_n = H_0 + (n-1)H_1$$

Thus, for large n, the entropy rate H_n/n of the source is approximately equal to H_1. Assuming that the data length n is large, we conclude that a lossless code for compressing the data generated by the stationary Markov source $[n, A_n, P_n]$ is efficient if and only if the resulting compression rate is approximately equal to H_1.

In the rest of this section, we survey each of the following efficient classes of lossless codes:

- Huffman codes
- Enumerative codes
- Arithmetic codes
- Lempel–Ziv codes

2.2. Huffman Codes

Fix a source $[n, A_n, P_n]$ with A_n finite; let S_n denote a lossless compression system driven by this source (see Fig. 4). In 1948, Claude Shannon put forth the following monotonicity principle for code design for the system S_n: The length of the binary codeword assigned to each data sequence in A_n should be inversely related to the probability with which that sequence occurs. According to this principle, data sequences with low probability of occurrence are assigned long binary codewords, whereas data sequences with high probability of occurrence are assigned short binary codewords. In Shannon's 1948 paper [42], a code called the *Shannon–Fano code* was put forth that obeys the monotonicity principle; it assigns a codeword to data sequence $(x_1, x_2, \ldots, x_n) \in A_n$ of length

$$\lceil -\log_2 P_n(x_1, x_2, \ldots, x_n) \rceil$$

The compression rate $R(S_n)$ resulting from the use of the Shannon–Fano code in system S_n is easily seen to satisfy

$$\frac{H_n}{n} \le R(S_n) \le \frac{H_n + 1}{n} \tag{11}$$

However, the Shannon–Fano code does not yield the minimal compression rate. The problem of finding the code that yields the minimal compression rate was solved in 1952 by David Huffman [22], and this code has been named the "Huffman code" in his honor.

We discuss the simplest instance of Huffman code design, namely, design of the Huffman code for a first-order source $[1, A, P]$; such a code encodes the individual letters in the source alphabet A and will be called a *first-order Huffman code*. If the letters in A are $a_1, a_2, \ldots, a_j$ and K_i denotes the length of the binary codeword into which letter a_i is encoded, then the Huffman code is the code for which

$$P(a_1)K_1 + P(a_2)K_2 + \cdots + P(a_j)K_j$$

is minimized. The Huffman algorithm constructs the encoding table of the Huffman code. The Huffman algorithm operates recursively in the following way. First, the letters a_i, a_j with the two smallest probabilities $P(a_i), P(a_j)$ are removed from the alphabet A and replaced with a single "superletter" $a_i a_j$ of probability $P(a_i) + P(a_j)$. Then, the Huffman code for the reduced alphabet is extended to a Huffman code for the original alphabet by assigning codeword $w0$ to a_i and codeword $w1$ to a_j, where w is the codeword assigned to the superletter $a_i a_j$.

Table 1 gives an example of the encoding table for the Huffman code for a first-order source with alphabet $\{a_1, a_2, a_3, a_4\}$. It is easy to deduce that the code given by Table 1 yields minimum compression rate without using the Huffman algorithm. First, notice that the codeword lengths $K_1, K_2, \ldots, K_j$ assigned by a minimum compression rate code must satisfy the equation

$$2^{-K_1} + 2^{-K_2} + \cdots + 2^{-K_j} = 1 \tag{12}$$

Table 1. Example of a Huffman Code

Source Letter	Probability	Codeword
a_1	$\frac{1}{2}$	0
a_2	$\frac{1}{5}$	10
a_3	$\frac{3}{20}$	110
a_4	$\frac{3}{20}$	111

[Kraft's inequality (2.4) is satisfied; if the inequality were strict, at least one of the codewords could be shortened and the code would not yield minimum compression rate.] Since the source has only a four-letter alphabet, there are only two possible solutions to (12), and therefore only two possible sets of codeword lengths for a first-order code, namely, 1,2,3,3 and 2,2,2,2. The first choice yields compression rate (or, equivalently, expected codeword length) of

$$1\tfrac{1}{2} + 2\tfrac{1}{5} + 3\tfrac{3}{20} + 3\tfrac{3}{20} = 1.8$$

code bits per data sample, whereas the second choice yields the worse compression rate of 2 code bits per data sample. Hence, the code given by Table 1 must be the Huffman code. One could use the Huffman algorithm to find this code. Combining the two least-probable letters a_3 and a_4 into the superletter a_3a_4, the following table gives the Huffman encoder for the reduced alphabet:

Source Letter	Probability	Codeword
a_1	$\frac{1}{2}$	0
a_3a_4	$\frac{6}{20}$	11
a_2	$\frac{1}{5}$	10

(This is immediate because for a three-letter alphabet, there is only one possible choice for the set of codeword lengths, namely, 1,2,2.) Expanding the codeword 11 for a_3a_4 into the codewords 110 and 111, we obtain the Huffman code in Table 1.

2.3. Enumerative Codes

Enumerative coding, in its present state of development, is due to Thomas Cover [11]. Enumerative coding is used for a source $[n, A_n, P_n]$ in which the data sequences in A_n are equally likely; the best lossless code for such a source is one that assigns codewords of equal length. Here is Cover's approach to enumerative code design:

Step 1. Let N be the number of sequences in A_n, and let A be the source alphabet. Construct the rooted tree T with N leaves, such that the edges emanating from each internal vertex have distinct labels from A, and such that the N sequences in A_n are found by writing down the labels along the N root-to-leaf paths of T.

Step 2. Locate all paths in T that visit only unary vertices in between and that are not subpaths of other such paths. Collapse each of these paths to single edges, labeling each such single edge that results with the sequence of labels along the collapsed path. This yields a tree T^* with N leaves (the same as the leaves of T). Label each leaf of T^* with the sequence obtained by concatenating together the labels on the root-to-leaf path to that leaf; these leaf labels are just the sequences in A_n.

Step 3. Assign an integer weight to each vertex v of T^* as follows. If v has no siblings or is further to the left than its siblings, assign v a weight of zero. If v has siblings further to the left, assign v a weight equal to the number of leaves of T^* that are equal to or subordinate to the siblings of v that are to the left of v.

Step 4. To encode $x^n \in A_n$, follow the root-to-leaf path in T^* terminating in the leaf of T^* labeled by x^n. Let I be the sum of the weights of the vertices along this root-to-leaf path. The integer I is called the *index* of x^n, and satisfies $0 \le I \le N-1$. Encode x^n into the binary codeword of length $\lceil \log_2 N \rceil$ obtained by finding the binary expansion of I and then padding that expansion (if necessary) to $\lceil \log_2 N \rceil$ bits by appending a prefix of zeros.

For example, suppose that the source $[n, A_n, P_n]$ satisfies

$$A_n = \{aaa, aba, abb, abc, baa, bba, caa, cba, cbb, cca\} \tag{13}$$

Then steps 1–3 yield the tree T^* in Fig. 5, in which every vertex is labeled with its weight from step 3. [The 10 leaves of this tree, from left to right, correspond to the 10 sequences in (13), from left to right.] The I values along the 10 paths are just the cumulative sums of the weights, which are seen to give $I = 0,1,2,3,4,5,6,7,8,9$. The codeword length is $\lceil \log_2 10 \rceil = 4$, and the respective codewords are

0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001

Sometimes, the tree T^* has such a regular structure that one can obtain an explicit formula relating a data sequence and its index I, thereby dispensing with the need for the tree T^* altogether. A good example of this occurs with the source $[n, A_n, P]$ in which A_n is the set of all binary sequences of length n and having a number of ones equal to m (m is a fixed positive integer satisfying

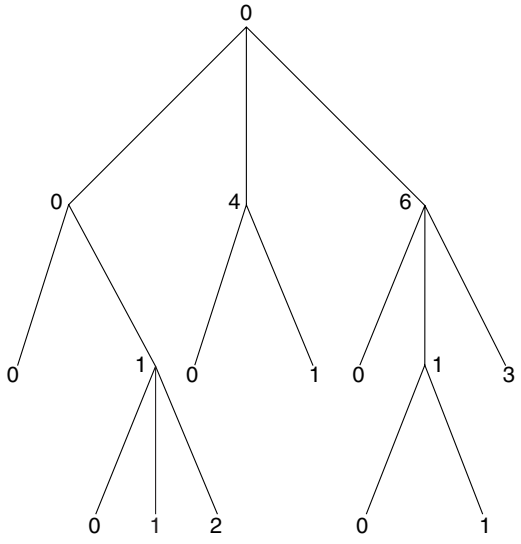


Figure 5. Weighted enumerative coding tree T^*.

$0 \le m \le n$). For a data sequence $(x_1, x_2, \ldots, x_n) \in A_n$, Cover showed that its index I is computable as

$$I = \sum_{j=1}^{n} x_j \binom{j-1}{x_1 + x_2 + \cdots + x_j} \tag{14}$$

There are $\binom{n}{m}$ sequences in A_n. Therefore, Eq. (14) provides a one-to-one correspondence between these sequences and the integers

$$0, 1, 2, 3, \ldots, \binom{n}{m} - 1$$

Having obtained I from $(x_1, x_2, \ldots, x_n)$ via this equation, $(x_1, x_2, \ldots, x_n)$ is encoded by expanding integer I in binary out to $\lceil \log_2 \binom{n}{m} \rceil$ bits. Conversely, $(x_1, x_2, \ldots, x_n)$ is decoded from its bit representation by first finding I, and then finding the unique expansion of I in (14) given by the right-hand side. To illustrate, suppose that $n = 8$ and $m = 4$. Then, A_n contains $\binom{8}{4} = 70$ sequences and I can be any integer between 0 and 69, inclusively. Suppose $I = 52$. To decode back into the data sequence in A_n that gave rise to the index $I = 52$, the decoder finds the unique integers $0 \le j_1 < j_2 < j_3 < j_4 < 8$ satisfying

$$52 = \binom{j_1}{1} + \binom{j_2}{2} + \binom{j_3}{3} + \binom{j_4}{4}$$

The solution is $j_1 = 1, j_2 = 4, j_3 = 5$, and $j_4 = 7$. The data sequence we are looking for must have ones in positions $j_1 + 1 = 2, j_2 + 1 = 5, j_3 + 1 = 6, j_4 + 1 = 8$; this is the sequence (0,1,0,0,1,1,0,1).

2.4. Arithmetic Codes

Arithmetic codes were invented by Peter Elias in unpublished work around 1960, but his schemes were not practical. In the 1970s, other people put Elias' ideas on a practical footing [33,36,37]. This section gives an introduction to arithmetic codes.

Arithmetic coding presents a whole new philosophy of coding. As the data samples in a data sequence $(x_1, x_2, \ldots, x_n)$ are processed from left to right by the arithmetic encoder, each data sample x_i is not replaced with a string of code bits as is done in conventional encoding—instead, each x_i is assigned a subinterval I_i of the unit interval [0, 1] so that

$$I_1 \supset I_2 \supset \cdots \supset I_n \tag{15}$$

and so that I_i is recursively determined from I_{i-1} and $x_i (i \ge 2)$. When the final interval I_n is determined, then the binary codeword $(b_1, b_2, \ldots, b_k)$ into which $(x_1, x_2, \ldots, x_n)$ is encoded is chosen so that the number

$$\frac{b_1}{2} + \frac{b_2}{4} + \frac{b_3}{8} + \cdots + \frac{b_k}{2^k} \tag{16}$$

is a point in I_n, where the codeword length k is approximately equal to $\log_2$ of the reciprocal of the probability assigned by the source to $(x_1, x_2, \ldots, x_n)$.

2.4.1. Precise Description. Arithmetic codes can be constructed for any source. For simplicity, we assume a memoryless source $[n, A_n, P_n]$ in which A_n consists of all sequences of length n whose entries come from the set $\{0, 1, 2, \ldots, j-1\}$, where j is a fixed positive integer. Then, for each data sequence $(x_1, x_2, \ldots, x_n) \in A_n$, the probability assigned by the source is

$$P_n(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p_{x_i}$$

where $p_0, p_1, \ldots, p_{j-1}$ are given nonnegative numbers that sum to one. We specify the arithmetic code for this source by describing algorithms for encoding and decoding. Let $a_0 = 0, a_1 = 1$. Arithmetic encoding of $(x_1, x_2, \ldots, x_n) \in A_n$ takes place according to the following three-step algorithm:

Encoding Step 1. For each $i = 1, 2, \ldots, n$, a subinterval $I_i = [a_i, b_i]$ of [0, 1] is recursively determined according to the formula

$$\begin{aligned} I_i &= [a_{i-1}, a_{i-1} + (b_{i-1} - a_{i-1})p_0], x_i = 0 \\ &= [a_{i-1} + (p_0 + \cdots + p_{x_i - 1})(b_{i-1} - a_{i-1}), a_{i-1} \\ &\quad + (p_0 + \cdots + p_{x_i})(b_{i-1} - a_{i-1})], x_i > 0 \end{aligned}$$

By construction, the last interval I_n will have length equal to $P_n(x_1, x_2, \ldots, x_n)$.

Encoding Step 2. The integer

$$k = \lceil -\log_2 P_n(x_1, x_2, \ldots, x_n) \rceil + 1$$

is determined. This integer will be the length of the codeword assigned by the arithmetic encoder to $(x_1, x_2, \ldots, x_n)$.

Encoding Step 3. The midpoint M of the interval I_n is computed. The codeword $(b_1, b_2, \ldots, b_k)$ assigned to $(x_1, x_2, \ldots, x_n)$ consists of the first k digits in the binary expansion of M.

The following arithmetic decoding algorithm is applied to the codeword $(b_1, b_2, \ldots, b_k)$ in order to reclaim the data sequence $(x_1, x_2, \ldots, x_n)$ that gave rise to it:

Decoding Step 1. Compute the point $\hat{M}$ given by the expression (16). By choice of k, the point $\hat{M}$ will lie in the interval I_n.

Decoding Step 2. There are j possibilities for I_1, depending on what x_1 is. Only one of these possibilities for I_1 contains $\hat{M}$. Using this fact, the decoder is able to determine I_1. From I_1, the decoder determines x_1.

Decoding Step 3. For each $i = 2, \ldots, n$, the decoder determines from I_{i-1} (determined previously) what the j possibilities for I_i are. Since only one of these possibilities for I_i contains $\hat{M}$, the decoder is able to determine I_i. From I_i, the decoder is able to determine x_i.

Example 1. Take the source alphabet to be $\{0, 1\}$, take the probabilities defining the memoryless source to be

$p_0 = \frac{2}{5}, p_1 = \frac{3}{5}$, and take the source to be $[5, \{0,1\}^5, P_5]$ (known to both encoder and decoder). Suppose that the data sequence to be arithmetically encoded is (1, 0, 1, 1, 0). We need to recursively determine the intervals I_1, I_2, I_3, I_4, I_5. We have

$$I_1 = \text{right } \frac{3}{5}\text{ths of } [0,1] = \left[\frac{2}{5}, 1\right]$$

$$I_2 = \text{left } \frac{2}{5}\text{ths of } I_1 = \left[\frac{2}{5}, \frac{16}{25}\right]$$

$$I_3 = \text{right } \frac{3}{5}\text{ths of } I_2 = \left[\frac{62}{125}, \frac{16}{25}\right]$$

$$I_4 = \text{right } \frac{3}{5}\text{ths of } I_3 = \left[\frac{346}{625}, \frac{16}{25}\right]$$

$$I_5 = \text{left } \frac{2}{5}\text{ths of } I_4 = \left[\frac{346}{625}, \frac{1838}{3125}\right]$$

The length of the interval I_5 is $\frac{108}{3125}$. Therefore, the length of the binary codeword must be

$$k = \left\lceil \log_2 \left(\frac{3125}{108} \right) \right\rceil + 1 = 6$$

The midpoint of the interval I_5 is $M = \frac{1784}{3125}$. Expanding this number in binary, we obtain

$$\frac{1784}{3125} = .100100\ldots$$

The binary codeword is therefore $(1,0,0,1,0,0)$. Given this codeword, how does the decoder determine the data sequence that gave rise to it? Since the decoder knows the source description, it knows that the data sequence to be found is of the form $(x_1, x_2, x_3, x_4, x_5)$, where the x_i terms are binary. To decode, the decoder first computes

$$\hat{M} = \frac{1}{2} + \frac{1}{16} = \frac{9}{16}$$

The decoder knows that

$$I_1 = \left[0, \tfrac{2}{5}\right] \quad \text{or} \quad I_1 = \left[\tfrac{2}{5}, 1\right]$$

Since $\frac{9}{16}$ is in the right interval, the decoder concludes that $I_1 = \left[\frac{2}{5}, 1\right]$ and that $x_1 = 1$. At this point, the decoder knows that

$$I_2 = \left[\frac{2}{5}, \frac{16}{25}\right] \text{ or } I_2 = \left[\frac{16}{25}, 1\right]$$

Since $\frac{9}{16}$ is in the left interval, the decoder concludes that $I_2 = \left[\frac{2}{5}, \frac{16}{25}\right]$ and that $x_2 = 0$. The decoder now knows that

$$I_3 = \left[\frac{2}{5}, \frac{62}{125}\right] \text{ or } I_3 = \left[\frac{62}{125}, \frac{16}{25}\right]$$

Since $\frac{9}{16}$ lies in the right interval, the decoder determines that $I_3 = \left[\frac{62}{125}, \frac{16}{25}\right]$, and that the third data sample is $x_3 = 1$. Similarly, the decoder can determine x_4 and x_5 by two more rounds of this procedure.

The reader sees from the preceding example that the arithmetic code as we have prescribed it requires ever greater precision as more and more data samples are processed. This creates a problem if there are a large number of data samples to be arithmetically encoded. One can give a more complicated (but less intuitive) description of the arithmetic encoder/decoder that uses only finite precision (integer arithmetic is used). A textbook [41] gives a wealth of detail on this approach.

2.4.2. Performance. In a sense that will be described here, arithmetic codes give the best possible performance, for large data length. First, we point out that for any source, an arithmetic code can be constructed. Let the source be $[n, A_n, P_n]$. For $(x_1, x_2, \ldots, x_n) \in A_n$, one can use conditional probabilities to factor the probability assigned to this sequence:

$$P_n(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i \mid x_1, x_2, \ldots, x_{i-1}).$$

The factors on the right side are explained as follows. Let $X_1, X_2, \ldots, X_n$ be the random data samples generated by the source. Then

$$p(x_1) = \Pr[X_1 = x_1]$$

$$p(x_i \mid x_1, x_2, \ldots, x_{i-1}) = \Pr[X_i = x_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}]$$

To arithmetically encode $(x_1, x_2, \ldots, x_n) \in A_n$, one constructs a decreasing sequence of intervals $I_1, I_2, \ldots, I_n$. One does this so that I_1 will have length $p(x_1)$ and for each $i = 2, \ldots, n$, the length of the interval I_i will be $p(x_i \mid x_1, x_2, \ldots, x_{i-1})$ times the length of the interval I_{i-1}. The rest of the encoding and decoding steps will be as already described for the memoryless source. Suppose a lossless data compression system S_n is built with the given source $[n, A_n, P_n]$ and the arithmetic code we have just sketched. It can be shown that the resulting compression rate satisfies

$$\frac{H_n}{n} \le R(S_n) \le \frac{H_n + 2}{n}$$

where H_n is the entropy of the source $[n, A_n, P_n]$. For large n, we therefore have $R(S_n) \approx H_n$. This is the best that one can possibly hope to do. If the source $[n, A_n, P_n]$ is memoryless or Markov, and n is large, one obtains the very good arithmetic code compression rate performance just described, but at the same time, the arithmetic code is of low complexity. As discussed in Section 2.2, for large n, the compression system S_n built using the Huffman code for the source $[n, A_n, P_n]$ will also achieve the very good compression rate performance $R(S_n) \approx H_n$, but this Huffman code will be very complex. For this reason, arithmetic codes are preferred over Huffman codes in many data compression applications. Two notable successes of arithmetic coding in practical

applications are the PPM text compression algorithm [8] and IBM's Q-coder [34], used for lossless binary image compression.

2.5. Lempel–Ziv Codes

Lempel–Ziv codes are examples of *dictionary codes*. A dictionary code first partitions a data sequence into variable-length phrases (this procedure is called *parsing*). Then, each phrase in the parsing is represented by means of a pointer to that phrase in a dictionary of phrases constructed from previously processed data. The phrase dictionary changes dynamically as the data sequence is processed from left to right. A binary codeword is then assigned to the data sequence by encoding the sequence of dictionary pointers in some simple way. The most popular dictionary codes are the Lempel–Ziv codes. There are many versions of the Lempel–Ziv codes. The one we discuss here is LZ78 [55]. Another popular Lempel–Ziv code, not discussed here, is LZ77 [54]. Two widely used compression algorithms on UNIX systems are Compress and Gzip; Compress is based on LZ78 and Gzip is based on LZ77.

In the rest of this section, we discuss the parsing technique, the pointer formation technique, and the pointer encoding technique employed in Lempel–Ziv coding.

2.5.1. Lempel–Ziv Parsing. Let $(x_1, x_2, \ldots, x_n)$ be the data sequence to be compressed. Partitioning of this sequence into variable-length blocks via *Lempel–Ziv parsing* takes place as follows. The first variable-length block arising from the Lempel–Ziv parsing of $(x_1, x_2, \ldots, x_n)$ is the single sample x_1. The second block in the parsing is the shortest prefix of $(x_2, x_3, \ldots, x_n)$ that is not equal to x_1. Suppose that this second block is $(x_2, \ldots, x_j)$. Then, the third block in Lempel–Ziv parsing will be the shortest prefix of $(x_{j+1}, x_{j+2}, \ldots, x_n)$ that is not equal to either x_1 or $(x_2, \ldots, x_j)$. In general, suppose that the Lempel–Ziv parsing procedure has produced the first k variable-length blocks $B_1, B_2, \ldots, B_k$ in the parsing, and $x^{(k)}$ is that part left of $(x_1, x_2, \ldots, x_n)$ after $B_1, B_2, \ldots, B_k$ have been removed. Then the next block B_{k+1} in the parsing is the shortest prefix of $x^{(k)}$ that is not equal to any of the preceding blocks $B_1, B_2, \ldots, B_k$. [If there is no such block, then $B_{k+1} = x^{(k)}$ and the Lempel-Ziv parsing procedure terminates.]

By construction, the sequence of variable-length blocks $B_1, B_2, \ldots, B_t$ produced by the Lempel–Ziv parsing of $(x_1, x_2, \ldots, x_n)$ are distinct, except that the last block B_t could be equal to one of the preceding ones. The following example illustrates Lempel–Ziv parsing.

Example 2. The Lempel–Ziv parsing of the data sequence

$$(1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1) \tag{17}$$

is

$$B_1 = (1)$$
$$B_2 = (1, 0)$$
$$B_3 = (1, 1)$$
$$B_4 = (0)$$
$$B_5 = (0, 0)$$
$$B_6 = (1, 1, 0) \tag{18}$$
$$B_7 = (1) \tag{19}$$

2.5.2. Pointer Formation. We suppose that the alphabet from which the data sequence $(x_1, x_2, \ldots, x_n)$ is formed is $A = \{0, 1, \ldots, k-1\}$, where k is a positive integer. After obtaining the Lempel–Ziv parsing $B_1, B_2, \ldots, B_t$ of $(x_1, x_2, \ldots, x_n)$, the next step is to represent each block in the parsing as a pair of integers. The first block in the parsing, B_1, consists of a single symbol. It is represented as the pair $(0, B_1)$. More generally, any block B_j of length one is represented as the pair $(0, B_j)$. If the block B_j is of length greater than one, then it is represented as the pair (i,s), where s is the last symbol in B_j and B_i is the unique previous block in the parsing that coincides with the block obtained by removing s from the end of B_j.

Example 3. The sequence of pairs corresponding to the parsing (19) is

$$(0, 1), (1, 0), (1, 1), (0, 0), (4, 0), (3, 0), (0, 1) \tag{20}$$

For example, $(4, 0)$ corresponds to the block $(0, 0)$ in the parsing. Since the last symbol of (0,0) is 0, the pair (4,0) ends in 0. The 4 in the first entry refers to the fact that $B_4 = (0)$ is the preceding block in the parsing, which is equal to what we get by deleting the last symbol of $(0, 0)$.

For our next step, we replace each pair $(i,\ s)$ by the integer $ki + s$. Thus, the sequence of pairs (20) becomes the sequence of integers

$$I_1 = 2 * 0 + 1 = 1$$
$$I_2 = 2 * 1 + 0 = 2$$
$$I_3 = 2 * 1 + 1 = 3$$
$$I_4 = 2 * 0 + 0 = 0 \tag{21}$$
$$I_5 = 2 * 4 + 0 = 8$$
$$I_6 = 2 * 3 + 0 = 6$$
$$I_7 = 2 * 0 + 1 = 1$$

2.5.3. Encoding of Pointers. Let $I_1, I_2, \ldots, I_t$ denote the integer pointers corresponding to the blocks $B_1, B_2, \ldots, B_t$ in the Lempel–Ziv parsing of the data sequence $(x_1, x_2, \ldots, x_n)$. To finish our description of the Lempel–Ziv encoder, we discuss how the integer pointers $I_1, I_2, \ldots, I_t$ are converted into a stream of bits. Each integer I_j is expanded to base 2, and these binary expansions are "padded" with zeros on the left so that the overall length of the string of bits assigned to I_j is $\lceil \log_2(kj) \rceil$. The reason why this many bits is necessary and sufficient is seen by examining the largest that I_j can possibly be. Let (i,s) be the pair associated with I_j. Then the largest that i can be is $j-1$ and the largest that s can be is $k-1$. Thus the largest that I_j can be is $k(j-1) + k - 1 = kj - 1$, and the number

of bits in the binary expansion of $kj-1$ is $\lceil \log_2(kj) \rceil$. Let W_j be the string of bits of length $\lceil \log_2(kj) \rceil$ assigned to I_j as described in the preceding text. Then, the Lempel–Ziv encoder output is obtained by concatenating together the strings $W_1, W_2, \ldots, W_t$.

To illustrate, suppose that the data sequence $(x_1, x_2, \ldots, x_n)$ is binary (i.e., $k=2$), and has seven blocks $B_1, B_2, \ldots, B_7$ in its Lempel–Ziv parsing. These blocks are assigned, respectively, strings of code bits $W_1, W_2, W_3, W_4, W_5, W_6, W_7$ of lengths $\lceil \log_2(2) \rceil = 1$, $\lceil \log_2(4) \rceil = 2$, $\lceil \log_2(6) \rceil = 3$, $\lceil \log_2(8) \rceil = 3$, $\lceil \log_2(10) \rceil = 4$, $\lceil \log_2(12) \rceil = 4$, and $\lceil \log_2(14) \rceil = 4$. Therefore, any binary data sequence with seven blocks in its Lempel–Ziv parsing would result in an encoder output of length $1+2+3+3+4+4+4=21$ code bits. In particular, for the data sequence (2.17), the seven strings $W_1, \ldots, W_7$ are [referring to (21)]

$$W_1 = (1)$$
$$W_2 = (1, 0)$$
$$W_3 = (0, 1, 1)$$
$$W_4 = (0, 0, 0)$$
$$W_5 = (1, 0, 0, 0)$$
$$W_6 = (0, 1, 1, 0)$$
$$W_7 = (0, 0, 0, 1)$$

Concatenating, we see that the codeword assigned to data sequence (17) by the Lempel–Ziv encoder is

$$(1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1) \quad (22)$$

We omit a detailed description of the Lempel–Ziv decoder. However, it is easy to see what the decoder would do. For example, it would be able to break up the codeword (22) into the separate codewords for the phrases, because, from the size k of the data alphabet, it is known how many code bits are allocated to the encoding of each Lempel–Ziv phrase. From the separate codewords, the decoder recovers the integer representing each phrase; dividing each of these integers by k to obtain the quotient and remainder, the pairs representing the phrases are obtained. Finally, these pairs yield the phrases, which are concatenated together to obtain the original data sequence.

2.5.4. Performance. Let $[n, A_n, P_n]$ be any data source with alphabet of size k. Let S_n be the lossless data compression system driven by this source that employs the Lempel–Ziv code. It is known that there is a positive constant C_k (depending on k but not on n) such that

$$\frac{H_n}{n} \le R(S_n) \le \frac{H_n}{n} + \frac{C_k}{\log_2 n}$$

where H_n is the source entropy. Thus, the Lempel–Ziv code is not quite as good as the Huffman code or the arithmetic code, but there is an important difference. The Huffman code and arithmetic code require knowledge of the source. The preceding performance bound is valid regardless of the source. Thus, one can use the same Lempel–Ziv code for all sources — such a code is called a *universal code* [13].

In practical compression scenarios, the Lempel–Ziv code has been superseded by more efficient modern dictionary codes, such as the YK algorithm [51].

3. LOSSY COMPRESSION METHODOLOGIES

In this section, we shall be concerned with the problem of designing lossy codes. Recall from Fig. 3 that a lossy compression system consists of source, noninvertible source encoder, and source decoder. Figure 6 gives separate depictions of the source encoder and source decoder in a lossy compression system.

The same notational conventions introduced earlier are in effect here: X^n [Eq. (1)] is the random data sequence of length n generated by the source, $\hat{X}^n$ (1.3) is the reconstructed data sequence, and B^K (2) is the variable-length codeword assigned to X^n. The source decoder component of the lossy compression system in Fig. 3 is the cascade of the "quantizer" and "lossless encoder" blocks in Fig. 6a; the source decoder component in Fig. 3 is the lossless decoder of Fig. 6b. The quantizer is a many-to-one mapping that converts the data sequence X^n into its reconstruction $\hat{X}^n$. The lossless encoder is a one-to-one mapping that converts the reconstructed data sequence $\hat{X}^n$ into the binary codeword B^K from which $\hat{X}^n$ can be recovered via application of the lossless decoder to B^K.

It is the presence of the quantizer that distinguishes a lossy compression system from a lossless one. Generally speaking, the purpose of the quantizer is alphabet reduction — by dealing with a data sequence from a reduced alphabet instead of the original data sequence over the original alphabet, one can hope to perform data compression using fewer code bits. For example, the "rounding-off quantizer" is a very simple quantizer. Let the data sequence

$$X^n = (1.1, 2.6, 4.4, 2.3, 1.7) \quad (23)$$

be observed. Then, the rounding-off quantizer generates the reconstructed data sequence

$$\hat{X}^n = (1, 3, 4, 2, 2) \quad (24)$$

It takes only 10 code bits to compress (24), because its entries come from the reduced alphabet $\{1, 2, 3, 4\}$; it

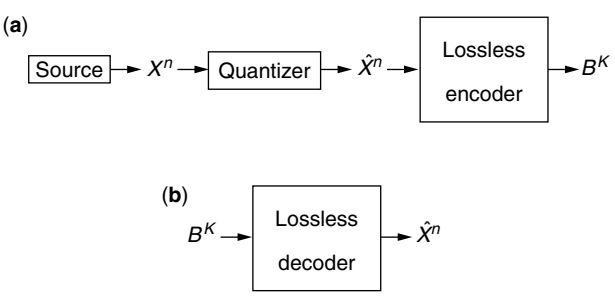


Figure 6. Source encoder (**a**) and decoder (**b**) in lossy compression system.

would take many more code bits to compress the original sequence (23).

As indicated in Fig. 6, a lossy code consists of quantizer, lossless encoder, and lossless decoder. When a lossy code is used to compress the data generated by a source $[n, A_n, P_n]$, a lossy compression system S_n results. The effectiveness of the lossy code is then evaluated by means of two figures of merit, the (compression) rate $R(S_n)$ and the distortion $D(S_n)$, defined by

$$R(S_n) \triangleq \frac{E[K]}{n}$$

$$D(S_n) \triangleq n^{-1} \sum_{i=1}^{n} E[d(X_i, \hat{X}_i)]$$

In the preceding, E denotes the expected value operator, K is the random codeword length of the codeword B^K assigned to X^n in Fig. 5a, and d is a fixed distortion function mapping pairs of source letters into nonnegative real numbers. The distortion function d is typically one of the following two types:

1. *Squared-error distortion*:

$$d(X_i, \hat{X}_i) = (X_i - \hat{X}_i)^2$$

2. *Hamming distortion*:

$$d(X_i, \hat{X}_i) = \begin{cases} 0, & X_i = \hat{X}_i \\ 1, & \text{otherwise} \end{cases}$$

Squared-error distortion is typically used for an infinite source alphabet and Hamming distortion is typically used for a finite source alphabet. When squared-error distortion is used, distortion is sometimes measured in decibels as the figure

$$[D(S_n)]_{\text{dec}} = 10 \log_{10} \left(\frac{n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} x^2 f_{X_i}(x)\, dx}{D(S_n)} \right)$$

where $X_i (1 \leq i \leq n)$ represents the ith data sample generated by the source $[n, A_n, P_n]$ and f_{X_i} denotes the probability density function of X_i; note that small $D(S_n)$ would correspond to a large decibel measure of distortion.

Suppose that two different lossy codes have been designed to compress the random data generated by a given source $[n, A_n, P_n]$, resulting in lossy compression systems S_n^1 and S_n^2, respectively. Then, one can declare that the lossy code giving rise to system S_n^1 is better than the lossy code giving rise to system S_n^2 if $R(S_n^1) < R(S_n^2)$ and $D(S_n^1) < D(S_n^2)$. However, it may be that neither lossy code is better than the other one in this sense, since the inverse relation between rate and distortion precludes the design of a lossy code for which rate and distortion are simultaneously small. Instead, for a given source, the design goal should be to find a lossy code that yields the smallest rate for a fixed distortion, or the smallest distortion for a fixed rate. The theory detailing the rate–distortion tradeoffs that are possible in lossy code design is called *rate–distortion theory*. Section 3.1 gives an introduction to this subject.

The quantizer employed in a lossy code can be one of two types, either a scalar quantizer or vector quantizer. A *scalar quantizer* quantizes one data sample at a time, whereas for some $m > 1$, a *vector quantizer* quantizes m data samples at a time. Lossy codes that employ scalar quantizers are called "SQ-based codes" and are covered in Section 3.2; lossy codes that employ vector quantizers are called "VQ-based codes" and are covered in Section 3.3. Subsequent sections deal with two other important lossy coding techniques, trellis-based coding, and transform coding.

3.1. Distortion Bounds

In designing a lossy code for a source $[n, A_n, P_n]$ to produce a lossy compression system S_n, the usual approach is the *fixed-rate approach*, in which one attempts to find a lossy code that minimizes or approximately minimizes the distortion $D(S_n)$ among all lossy codes satisfying the rate constraint $R(S_n) \leq R$, where $R > 0$ is a fixed constant. We adopt the fixed-rate approach for the rest of this article.

In this section, we will give upper and lower bounds on the distortion performance of lossy codes for a fixed source, subject to the constraint that the compression rate be no more than R code bits per data sample, on average. With these bounds, one can determine before designing a lossy code what types of performance are and are not possible for such a code to achieve. The upper and lower bounds on distortion performance that shall be developed in this section are expressible using Claude Shannon's notion of *distortion–rate function* [43], discussed next.

3.1.1. Distortion–Rate Function. The concept of *mutual information* will be needed in order to define the distortion–rate function. Let X, Y be two random variables. Let $f(x)$ be the probability density function of X, and let $g(y \mid x)$ be the conditional probability density function of Y given $X = x$. Then, the mutual information $I(X;Y)$ of X, Y is the number

$$I(X;Y) \triangleq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) g(y \mid x) \log_2 \times \left[\frac{g(y \mid x)}{\int_{-\infty}^{\infty} f(u) g(y \mid u) du} \right] dx\, dy$$

We are now ready to define the concept of distortion–rate function. For simplicity, we restrict ourselves to the memoryless source $[n, A_n, P_n]$. We suppose that the source alphabet is a subset of the real line; let X be a random variable such that the random data samples $X_i (1 \leq i \leq n)$ generated according to the memoryless source $[n, A_n, P_n]$ are independent copies of X. The distortion–rate function $D(R)$ of the memoryless source $[n, A_n, P_n]$ (for a given nonnegative distortion function d such as Hamming distortion or squared-error distortion) is then defined for each $R > 0$ by

$$D(R) \triangleq \min\{E[d(X,Y)] : I(X;Y) \leq R\} \qquad (25)$$

where Y denotes any random variable jointly distributed with X, and we assume that the minimum in (25) exists and is finite.

Example 4. One important type of memoryless source for which the distortion–rate function has a closed-form expression is the *Gaussian memoryless source* $[n, A_n, P_n]$, which is the memoryless source whose independent random data samples are generated according to a Gaussian distribution with mean 0 and variance σ^2. For the Gaussian memoryless source with squared-error distortion

$$D(R) = \sigma^2 2^{-2R} \tag{26}$$

Sources such as this one in which $D(R)$ can be computed explicitly are rare. However, in general, one can use the Blahut algorithm [6] to approximate $D(R)$ arbitrarily closely.

3.1.2. Distortion Lower Bound. Fix a memoryless source $[n, A_n, P_n]$ with distortion–rate function $D(R)$. Fix $R > 0$, and fix any lossy code for $[n, A_n, P_n]$ for which the resulting compression system S_n satisfies the rate constraint $R(S_n) \le R$. Then, the following well-known [4] distortion lower bound is valid:

$$D(S_n) \ge D(R) \tag{27}$$

Example 5. Assume the nth-order Gaussian memoryless source model and squared-error distortion. Substituting the expression for $D(R)$ in (26) into (27), one concludes that any lossy code with compression rate $\le R$ yields distortion in decibels no greater than $(20 \log_{10} 2)R \approx 6R$. Thus, a Gaussian memoryless source cannot be encoded to yield a better distortion performance than "6 decibels per bit."

3.1.3. Distortion Upper Bound. Let f be a probability density function. For each $n = 1, 2, \ldots$, let $[n, A_n, P_n]$ be the memoryless source in which n independent random data samples are generated according to the density f. Assume finiteness of the distortion–rate function $D(R)$ for these sources [since the distortion–rate function depends only on f, all of these sources will have the same distortion-rate function $D(R)$]. Fix $R > 0$. It is well known [35, 53] that there is a positive constant C such that for every $n = 2, 3, \ldots$, there is a lossy code for $[n, A_n, P_n]$ such that the rate and distortion for the resulting compression system S_n satisfy

$$\begin{aligned} R(S_n) &\le R \\ D(S_n) &\le D(R) + \frac{C \log_2 n}{n} \end{aligned} \tag{28}$$

Combining the distortion upper bound (28) with the distortion lower bound (27), if n is large, there must exist a code for an nth order memoryless source that yields rate $\le R$ and distortion $\approx D(R)$; that is, the distortion–rate function $D(R)$ does indeed describe the distortion performance of the most efficient lossy codes. But, in general, it is not known how to find codes that are this efficient. This represents a clear difference between lossless code and lossy code design; it is known how to construct efficient lossless codes, but it is a computationally difficult problem to find efficient lossy codes [18]. For example, for large n, the nth-order Gaussian memoryless source can be encoded to yield squared-error distortion of roughly "6 decibels per bit" (the distortion–rate function performance), but only since 1990 has it been discovered how to find such codes for this simple source model [29].

3.2. SQ-Based Codes

Let R be a fixed positive integer. A 2^R-*bit scalar quantizer* for quantizing real numbers in the interval $[a, b]$ is a mapping Q from $[a, b]$ into a finite subset of $[a, b]$ of size $N = 2^R$. Let $I_1, I_2, \ldots, I_N$ be subintervals of $[a, b]$ that form a partition of $[a, b]$ (the I_j values are called the *quantization intervals* of Q). Let $L_1, L_2, \ldots, L_N$ be points in $[a, b]$ chosen so that $L_j \in I_j$ for each $j = 1, 2, \ldots, N$ (L_j is called the *quantization level* for interval I_j). The quantizer Q accepts as input any real number x in the interval $[a, b]$. The output $Q(x)$ generated by the quantizer Q in response to the input x is the quantization level L_j assigned to the subinterval I_j of $[a, b]$ containing x. In other words, the 2^R-bit quantizer Q is a nondecreasing step function taking 2^R values.

Let $[n, A_n, P_n]$ be a source (such as a memoryless source) in which each randomly generated data sample has the same probability density function f. Let the alphabet for $[n, A_n, P_n]$ be the interval of real numbers $[a, b]$. Let Q be a 2^R-bit scalar quantizer defined on $[a, b]$. We describe a lossy code for the source $[n, A_n, P_n]$ induced by Q. Referring to Fig. 6a, we must explain how the lossy code quantizes source sequences of length n and losslessly encodes the quantized sequences. The lossy code quantizes each source sequence $(x_1, x_2, \ldots, x_n) \in A^n$ into the sequence $(Q(x_1), Q(x_2), \ldots, Q(x_n))$; this makes sense because each entry of each source sequence belongs to the interval $[a, b]$ on which Q is defined. Assign each of the 2^R quantization levels of Q an R-bit binary address so that there is a one-to-one correspondence between quantization levels and their addresses; then, the lossy code losslessly encodes $(Q(x_1), Q(x_2), \ldots, Q(x_n))$ by replacing each of its entries with its R-bit address, yielding an overall binary codeword of length nR. Let S_n be the lossy compression system in Fig. 6 arising from the lossy code just described, and let d be the distortion function that is to be used. It is not hard to see that

$$\begin{aligned} R(S_n) &= R \\ D(S_n) &= \int_a^b d(x, Q(x)) f(x)\, dx \end{aligned}$$

If in the preceding construction we let Q vary over all 2^R-bit scalar quantizers on $[a, b]$, then we obtain all possible SQ based lossy codes for the source $[n, A_n, P_n]$ with compression rate R.

Let R be a positive integer. Consider the following problem: Find the 2^R-bit scalar quantizer Q on $[a, b]$ for which

$$\int_a^b d(x, Q(x)) f(x)\, dx$$

is minimized. This quantizer Q yields the SQ based lossy code for the source $[n, A_n, P_n]$, which has compression rate R and minimal distortion. We call this scalar quantizer the *minimum distortion* 2^R-bit scalar quantizer.

3.2.1. Lloyd–Max Quantizers. We present the solution given in other papers [17,28,30] to the problem of finding the minimum distortion 2^R-bit scalar quantizer for squared-error distortion. Suppose that the probability density function f satisfies

$$\int_a^b x^2 f(x)\,dx < \infty$$

and that $-\log_2 f$ is a concave function on $[a, b]$. Let Q be a 2^R-bit scalar quantizer on $[a, b]$ with quantization levels $\{L_j\}_{j=1}^N$ and quantization intervals $\{I_j\}_{j=1}^N$, where $N = 2^R$. Let

$$y_0 < y_1 < \cdots < y_{N-1} < y_N$$

be the points such that interval I_j has left endpoint y_{j-1} and right endpoint $y_j (j = 1, \ldots, N)$. We call Q a *Lloyd–Max quantizer* if

$$y_j = (\tfrac{1}{2})[L_j + L_{j+1}],\ j = 1, \ldots, N-1 \tag{29}$$

and

$$L_j = \frac{\displaystyle\int_{y_{j-1}}^{y_j} xf(x)\,dx}{\displaystyle\int_{y_{j-1}}^{y_j} f(x)\,dx},\ j = 1, 2, \ldots, N. \tag{30}$$

There is only one 2^R-bit Lloyd–Max scalar quantizer, and it is the unique minimum distortion 2^R-bit scalar quantizer.

Example 6. One case in which it is easy to solve the Lloyd–Max equations (29) and (30) is the case in which $R = 1$ and f is an even function on the whole real line. The unique Lloyd–Max quantizer Q is then given by

$$Q(x) = \begin{cases} 2\displaystyle\int_0^\infty xf(x)\,dx, & x < 0 \\ -2\displaystyle\int_0^\infty xf(x)\,dx, & x \geq 0 \end{cases}$$

Example 7. Assume that $[a, b]$ is a finite interval. Let the source $[n, A_n, P_n]$ be memoryless with each random data sample uniformly distributed on the interval $[a, b]$. The unique 2^R-bit Lloyd–Max scalar quantizer Q for this source is obtained by partitioning $[a, b]$ into 2^R equal subintervals, and by assigning the quantization level for each interval to be the midpoint of that interval [the Lloyd–Max equations (29) and (30) are easily verified]. The SQ-based lossy code for the source $[n, A_n, P_n]$ induced by Q yields rate R and distortion in decibels equal to $(20 \log_{10} 2)R \approx 6R$. This is another "6 decibels per bit" result. Computation of the distortion–rate function for the memoryless source $[n, A_n, P_n]$ [24] reveals that for large n, $[n, A_n, P_n]$ can be encoded at rate $R = 1$ bit per sample at a distortion level of nearly 6.8 decibels. This clearly cannot be achievable using SQ-based lossy codes — more sophisticated lossy codes would be required.

In general, it may not be possible to solve the Eqs. (29) and (30) to find the Lloyd–Max quantizer explicitly. But, a numerical approximation of it can be found using the LBG algorithm covered in the next section.

3.3. VQ-Based Codes

Fix a positive integer m. Let $\mathcal{R}$ denote the set of real numbers, and let $\mathcal{R}^m$ denote m-dimensional Euclidean space, that is, the set of all sequences $(x_1, x_2, \ldots, x_m)$ in which each entry x_i belongs to $\mathcal{R}$. An *m-dimensional vector quantizer* Q is a mapping from $\mathcal{R}^m$ onto a finite subset $\mathcal{C}$ of $\mathcal{R}^m$. The set $\mathcal{C}$ is called the *codebook* of the m-dimensional vector quantizer Q; the elements of $\mathcal{C}$ are called *codevectors*, and if $x \in \mathcal{R}^m$, then $Q(x)$ is called the *codevector for* x. We define a mapping to be a *vector quantizer* if it is an m-dimensional vector quantizer for some m. The scalar quantizers can be regarded as special cases of the vector quantizers (they are the one-dimensional vector quantizers); the codebook of a scalar quantizer is just the set of its quantization levels, and a codevector for a scalar quantizer is just one of its quantization levels.

Let Q be an m-dimensional vector quantizer with codebook $\mathcal{C}$. We call Q a *nearest-neighbor quantizer* if Q quantizes each $x \in \mathcal{R}^m$ into a codevector from codebook $\mathcal{C}$ that is at least as close to x in Euclidean distance as any other codevector from $\mathcal{C}$. (We use squared-error distortion in this section; therefore, only nearest-neighbor quantizers will be of interest to us.)

Fix a memoryless source $[n, A_n, P_n]$ whose alphabet is a subset of the real line, such that n is an integer multiple of m; let f_m be the common probability density function possessed by all random vectors of m consecutive samples generated by this source. Let Q be a fixed m-dimensional nearest-neighbor vector quantizer whose codebook is of size 2^j. We describe how Q induces a lossy code for the source $[n, A_n, P_n]$ that has compression rate $R = j/m$. Referring to Fig. 6, we have to explain how the induced lossy code quantizes source sequences generated by $[n, A_n, P_n]$, and how it losslessly encodes the quantized sequences. Let $(x_1, x_2, \ldots, x_n) \in A_n$ be any source sequence. Partitioning, one obtains the following n/m blocks of length m lying in m-dimensional Euclidean space $\mathcal{R}^m$:

$$(x_1, x_2, \ldots, x_m)$$

$$(x_{m+1}, x_{m+2}, \ldots, x_{2m})$$

$$\ldots$$

$$(x_{n-m+1}, x_{n-m+2}, \ldots, x_n)$$

The induced lossy code quantizes $(x_1, x_2, \ldots, x_n)$ into $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, where

$$(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m) = Q(x_1, x_2, \ldots, x_m)$$

$$(\hat{x}_{m+1}, \hat{x}_{m+2}, \ldots, \hat{x}_{2m}) = Q(x_{m+1}, x_{m+2}, \ldots, x_{2m})$$

$$\ldots$$

$$(\hat{x}_{n-m+1}, \hat{x}_{n-m+2}, \ldots, \hat{x}_n) = Q(x_{n-m+1}, x_{n-m+2}, \ldots, x_n)$$

Each codevector in $\mathcal{C}$ can be uniquely represented using a j-bit binary address. The induced lossy code losslessly encodes the quantized sequence $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ into the binary codeword obtained by concatenating together the binary addresses of the codevectors above that were used to form the quantized sequence; this binary codeword is of fixed length $(n/m)j = nR$, and so, dividing by n, the compression rate R code bits per data sample has been achieved. Let $\mathcal{S}_n$ be the lossy compression system that arises when the lossy code just described is used to compress the data sequences generated by the source $[n, A_n, P_n]$. The resulting rate and distortion performance are given by

$$R(\mathcal{S}_n) = R = \frac{j}{m}$$
$$D(\mathcal{S}_n) = m^{-1} \int_{\mathcal{R}^m} \min_{y \in \mathcal{C}} \|x - y\|^2 f_m(x)\, dx \qquad (31)$$

where $\|x - y\|^2$ denotes the square of the Euclidean distance between vectors $x = (x_i)$ and $y = (y_i)$ in $\mathcal{R}^m$:

$$\|x - y\|^2 = \sum_{i=1}^{m} (x_i - y_i)^2$$

If we let Q vary over all nearest-neightbor m-dimensional vector quantizers whose codebooks are of size 2^j, then the lossy codes for the source $[n, A_n, P_n]$ induced by these are the VQ-based codes of rate $R = j/m$. Obviously, of this large number of lossy codes, one would want to choose one of minimal distortion; however, it is typically an intractable problem to find such a minimum distortion code.

Example 8. Suppose that the source $[n, A_n, P_n]$ for which a VQ-based code is to be designed is a Gaussian memoryless source with mean 0 and variance 1. Suppose that the desired compression rate is 1.5 code bits per data sample. Since $1.5 = \frac{3}{2}$, we can use a two-dimensional vector quantizer with codebook of size $2^3 = 8$. The table below gives one possibility for such a vector quantizer. The left column gives the codevectors in the codebook; the right column gives the binary address assigned to each codevector. The codevectors in this codebook can be visualized as eight equally spaced points along the unit circle in the plane, which is of radius 1 and centered at the origin (0, 0).

Codevector	Address
(1, 0)	000
$(\cos(\pi/4), \sin(\pi/4))$	001
(0, 1)	010
$(\cos(3\pi/4), \sin(3\pi/4))$	011
$(-1, 0)$	100
$(\cos(5\pi/4), \sin(5\pi/4))$	101
$(0, -1)$	110
$(\cos(7\pi/4), \sin(7\pi/4))$	111

The lossy code induced by this 2D vector quantizer, when used to compress data generated by the source $[n, A_n, P_n]$, gives rise to a distortion figure that can be obtained using symmetry considerations. The 2D plane can be partitioned into eight congruent regions over which the integrals of the integrand in (31) are identical. One of these regions is

$$\mathcal{S} = \left\{(x_1, x_2) : x_1 \geq 0, -\tan\frac{\pi}{8}x_1 \leq x_2 \leq \tan\frac{\pi}{8}x_1\right\}$$

The region S is the set of points in the plane $\mathcal{R}^2$ that are closest in Euclidean distance to the codevector (1, 0). Therefore, the distortion is

$$4 \iint_S [(x_1 - 1)^2 + x_2^2] \left(\frac{1}{2\pi}\right) \exp\left(\frac{-(x_1^2 + x_2^2)}{2}\right) dx_1\, dx_2$$

This integral is easily evaluated via conversion to polar coordinates. Doing this, one obtains distortion of 5.55 dB. One can improve the distortion to 5.95 dB by increasing the radius of the circle around which the eight codevectors in the codebook are distributed. Examining the distortion–rate function of the Gaussian memoryless source, we see that a distortion of about 9 dB is best possible at the compression rate of 1.5 code bits per data sample. Hence, we can obtain a >3-dB improvement in distortion by designing a VQ-based code that uses a vector quantizer of dimension >2.

3.3.1. LBG Algorithm. The LBG algorithm [27] is an iterative algorithm for vector quantizer design. It works for any dimension m, and employs a large set T of "training vectors" from $\mathcal{R}^m$. The training vectors, for example, could represent previously observed data vectors of length m. An initial codebook $\mathcal{C}_0$ contained in $\mathcal{R}^m$ of some desired size is selected (the size of the codebook is a reflection of the desired compression rate). The LBG algorithm then recursively generates new codebooks

$$\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \ldots$$

as follows. Each codebook $\mathcal{C}_i (i \geq 1)$ is generated from the previous codebook $\mathcal{C}_{i-1}$ in two steps:

Step 1. For each $v \in T$, a closest vector to v in $\mathcal{C}_{i-1}$ is found (with respect to Euclidean distance), and recorded. Let x_v denote the vector that is recorded for $v \in T$.

Step 2. For each vector $y \in \{x_v : v \in T\}$ recorded in step 1, the arithmetic mean of all vectors v in T for which $x_v = y$ is computed. The set of arithmetic means forms the new codebook $\mathcal{C}_i$.

The LBG codebooks $\{\mathcal{C}_i\}$ either eventually coincide, or else eventually keep cycling periodically through a finite set of codebooks; in either case, one would stop iterations of the LBG algorithm at a final codebook $\mathcal{C}_i$ as soon as one of these two scenarios occurs. The final LBG codebook, if used to quantize the training set, will yield smaller distortion on the training set than any of the previously generated codebooks (including the initial codebook). The LBG algorithm has been used extensively since its discovery for VQ-based code design in both

theoretical and practical scenarios. The one-dimensional LBG algorithm can be used to find a good approximation to the unique Lloyd–Max quantizer, if the training set is big enough. However, in higher dimensions, the LBG algorithm has some drawbacks: (1) the final LBG codebook may depend on the initial choice of codebook or (2) the final LBG codebook may not yield minimum distortion. Various stochastic relaxation and neural network–based optimization techniques have been devised to overcome these deficiencies [38, Chap. 14;52].

3.3.2. Tree-Structured Vector Quantizers. In tree-structured vector quantization [19, Chap. 12;32], the 2^j codevectors in the VQ codebook are placed as labels on the 2^j leaf vertices of a rooted binary tree of depth j. Each of the $2^j - 1$ internal vertices of the tree is also labeled with some vector. To quantize a vector x, one starts at the root of the tree and then follows a unique root-to-leaf path through the tree by seeing at each intermediate vertex of the path which of its two children has its label closer to x, whereupon the next vertex in the path is that child; x is quantized into the codevector at the terminus of this path. A tree-structured vector quantizer encodes speedily, since the time for it to quantize a vector using a codebook of size 2^j is proportional to j instead of 2^j. The minimum distortion vector quantizer for a probabilistic source or a training set is typically not implementable as a tree-structured vector quantizer, but there are some scenarios in which a tree-structured vector quantizer yields close to minimum distortion.

3.3.3. Lattice Vector Quantizers. A lattice quantizer is a vector quantizer whose codebook is formed from points in some Euclidean space lattice. Lattice quantizers are desirable because (1) there are fast algorithms for implementing them [9], and (2) for high compression rates, lattice quantizers yield nearly minimum distortion among all vector quantizers for the memoryless source in which the data samples have a uniform distribution. A monograph [10, Table 2.3] tabulates the best-known m-dimensional lattices in terms of distortion performance, for various values of m between $m = 1$ and $m = 24$. For example, for dimension $m = 2$, one should use a hexagonal lattice consisting of the centers of hexagons that tile the plane.

3.4. Trellis-Based Codes

Suppose that one has a finite-directed graph G satisfying the following properties:

- There are a fixed number of outgoing edges from each vertex of G, and this fixed number is a power of two. (Let this fixed number be 2^j.)
- Each edge of G has a label consisting of a sequence of fixed length from a given data alphabet. A. (Let this fixed length be m.)
- The graph G is connected (i.e., given any two vertices of G, there is a finite path connecting them).

Let v^* be any fixed vertex of G (since G is connected, it will not matter what v^* is). Let n be a fixed positive integer that is an integer multiple of m. Let $\mathcal{P}$ be the set of all paths in G that begin at v^* and consist of n/m edges. Then

- Each path in $\mathcal{P}$ gives rise to a sequence of length n over the alphabet A if one writes down the labels on its edges in the order in which these edges are visited.
- Let $R = j/m$. There are 2^{nR} paths in $\mathcal{P}$, and it consequently takes nR bits to uniquely identify any one of these paths.

A *trellis* is a pictorial representation of all of the paths in $\mathcal{P}$. The *trellis-based lossy code* for quantizing/encoding/decoding all the data sequences in A^n works in the following way:

Quantization Step. Given the data sequence $(x_1, x_2, \ldots, x_n)$ from A^n, the encoder finds a "minimal path" in $\mathcal{P}$, which is a path in $\mathcal{P}$ giving rise to a sequence $(y_1, y_2, \ldots, y_n)$ in A^n for which

$$\sum_{i=1}^{n} d(x_i, y_i)$$

is a minimum, where d is the distortion measure to be used. [The sequence $(y_1, y_2, \ldots, y_n)$ is the output of the quantizer in Fig. 6a.]

Encoding Step. Let $R = j/m$. Having found the minimal path in $\mathcal{P}$ for the data sequence $(x_1, x_2, \ldots, x_n)$, the encoder transmits a binary codeword of length nR in order to tell the decoder which path in $\mathcal{P}$ was the minimal path.

Decoding Step. From the received binary codeword of length nR, the decoder follows the minimal path and writes down the sequence $(y_1, y_2, \ldots, y_n)$ to which that path gives rise. That sequence is the reconstruction sequence $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ for the data sequence $(x_1, x_2, \ldots, x_n)$.

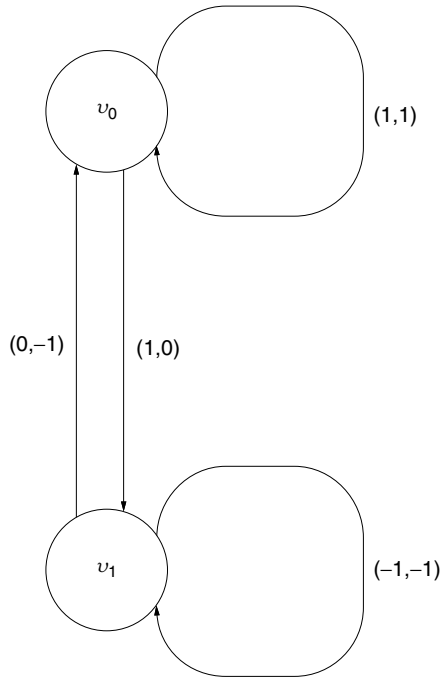
Computation of the minimal path in the quantization step is a dynamic programming problem that can be efficiently solved using the *Viterbi algorithm*.

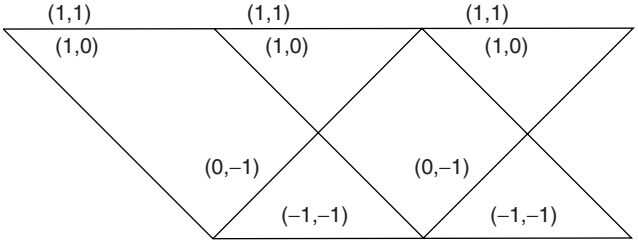
The trellis-based lossy code just described will work for any source $[n, A_n, P_n]$ in which $A_n \subset A^n$. The resulting compression rate is $R = j/m$; the resulting distortion depends on the source model — there is no simple formula for computing it. Trellis-based lossy coding is quite unlike SQ-based or VQ-based lossy coding because of its feedback nature. The best way to understand it is through an extended example, which follows.

Example 9. The graph G is taken as follows:

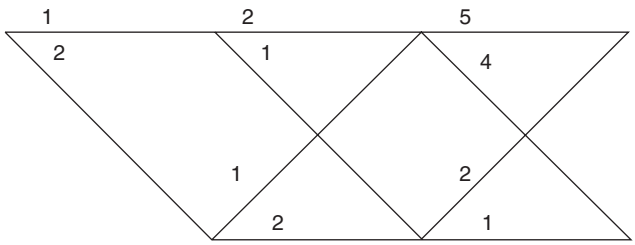
There are two vertices v_0, v_1, four edges with two outgoing edges per vertex ($j = 1$), and the edge labels are of length 2 ($m = 2$). The compression rate for the trellis-based code based on G is $R = j/m = 0.5$ code bits per data sample. Suppose that we wish to encode the data sequence (0, 1, 0, 0, −1, 0) of length $n = 6$, using squared-error distortion. Taking the vertex v_0 as our distinguished vertex v^*, the set of $2^{nR} = 8$ paths $\mathcal{P}$ in

G of length $n/m = 3$ that start at v_0 are then given the trellis representation:

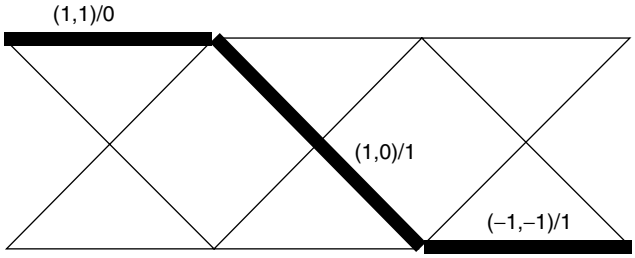




The trellis comes in three stages from left to right, which designate stage 1, stage 2, and stage 3, respectively. In stage 1, one replaces each label with the square of the Euclidean distance between that label and the pair (0,1) consisting of the first two of the six given data samples. In stage 2, one replaces each label with the square of the Euclidean distance between that label and the pair (0,0) consisting of the next two data samples. In stage 3, one replaces each label with the square of the Euclidean distance between that label and the pair $(-1, 0)$ consisting of the final two data samples. This gives us the following trellis whose edges are weighted by these squared Euclidean distances:

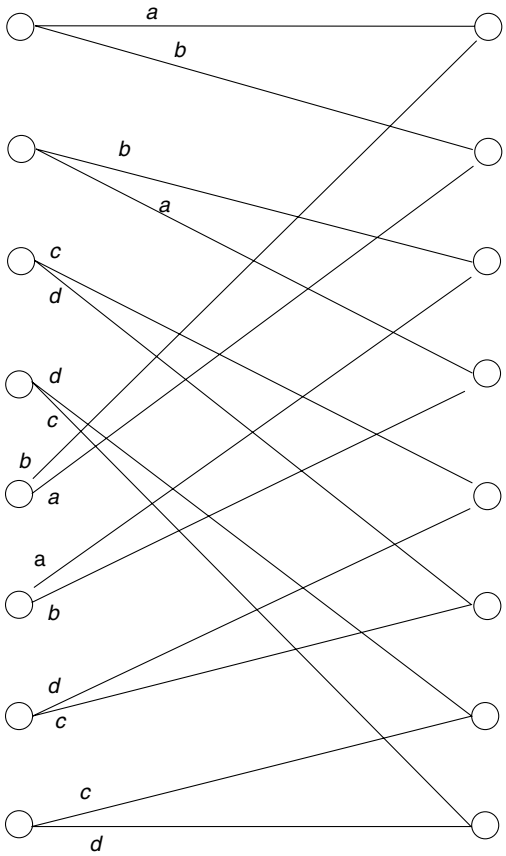


The minimal path is the path whose sum of weights is the smallest; this is the path marked in bold in the following figure:



We have added an additional label to each edge of the minimal path to denote the bit sent to the decoder (bit = 0 means upper branch chosen; bit = 1 means lower branch chosen).

Example 10. Consider the memoryless source $[n, A_n, P_n]$ with alphabet $\{a, b, c, d\}$, in which each random data sample is equidistributed over this alphabet. Suppose that $[n, A_n, P_n]$ is encoded using the trellis-based code, one stage of which is the following:



This code has a compression rate of 1 code bit per data sample. Let Hamming distortion be used; for large n, simulations show that the Hamming distortion is slightly less than 0.25. Hence, for a large number of data samples, this simple code can reconstruct over 75% of the data

samples at the decoder, on average, while transmitting only half of the bits that would be necessary for perfect reconstruction (which would require a compression rate of 2 bits per sample).

3.4.1. Marcellin–Fischer Codes. Marcellin and Fischer [29] developed an important class of trellis-based codes. They use DeBrujn graphs as the basis for their codes — labeling of the edges is done using heuristic symmetry rules motivated by Ungerboeck's work on trellis-coded modulation [48]. Marcellin and Fischer report quite good distortion performance when their trellis-based codes are used to encode Gaussian and uniform memoryless sources.

3.5. Transform Codes

In certain applications, the data samples to be compressed are not one-dimensional quantities; that is, for some $m > 1$, each data sample is a point in m-dimensional Euclidean space.

Example 11. One may wish to compress a large square image by compressing the sequence of 8×8 blocks that are obtained from partitioning the image. Each data "sample" in this case could be thought of as a point in m-dimensional Euclidean space with $m = 64$. (The JPEG image compression algorithm takes this point of view.)

For simplicity, let us assume that the dimension of the data samples is $m = 2$. Let us write the data samples as random pairs

$$\begin{bmatrix} X_i^{(1)} \\ X_i^{(2)} \end{bmatrix}, \quad i = 1, 2, \ldots, n \tag{32}$$

Suppose that one is committed to lossy codes that employ only scalar quantization. The sequence

$$X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)}$$

might have one type of model that would dictate that some scalar quantizer Q_1 be used to quantize these samples. On the other hand, the sequence

$$X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}$$

might have another type of model that would dictate the use of a different scalar quantizer Q_2. Let us assume that Q_1 is a 2^{R_1}-bit quantizer and that Q_2 is a 2^{R_2}-bit quantizer. The sequence resulting from quantization of (3.32) would then be

$$\begin{bmatrix} \hat{X}_i^{(1)} \\ \hat{X}_i^{(2)} \end{bmatrix}, \quad i = 1, 2, \ldots, n \tag{33}$$

where $\hat{X}_i^{(1)} = Q_1(X_i^{(1)})$ and $\hat{X}_i^{(2)} = Q_2(X_i^{(2)})$. The sequence (33) is transmitted by the encoder to the decoder using $nR_1 + nR_2$ bits, and is reconstructed by the decoder as the decoder's estimate of the original sequence (32). Let us use squared-error distortion. Then, the compression rate R and distortion D for this lossy code are

$$R = R_1 + R_2$$

$$D = n^{-1} \sum_{i=1}^{n} E[(X_i^{(1)} - \hat{X}_i^{(1)})^2] + n^{-1} \sum_{i=1}^{n} E[(X_i^{(2)} - \hat{X}_i^{(2)})^2]$$

Suppose that the compression rate is to be kept fixed at R. Then, to minimize D, one would optimize the lossy code just described by choosing integers R_1 and R_2 so that $R = R_1 + R_2$, and so that

$$\min_{Q_1} \sum_{i=1}^{n} E[(X_i^{(1)} - Q_1(X_i^{(1)}))^2] + \min_{Q_2} \sum_{i=2}^{n} E[(X_i^{(2)} - Q_2(X_i^{(2)}))^2]$$

is a minimum, where in the first minimization Q_1 ranges over all 2^{R_1}-bit quantizers, and in the second minimization Q_2 ranges over all 2^{R_2}-bit quantizers. Finding the best way to split up R into a sum $R = R_1 + R_2$ is called the *bit allocation problem* in lossy source coding.

We may be able to obtain a smaller distortion by *transforming* the original data pairs from $\mathcal{R}^2$ into another sequence of pairs from $\mathcal{R}^2$, and then doing the quantization and encoding of the transformed pairs in a manner similar to what we did above for the original pairs. This is the philosophy behind *transform coding*. In the following, we make this idea more precise.

Let m be a fixed positive integer. The data sequence to be compressed is

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \tag{34}$$

where each $\mathbf{X}_i$ is an $\mathcal{R}^m$-valued random column vector of the form

$$\mathbf{X}_i = \begin{bmatrix} X_i^{(1)} \\ X_i^{(2)} \\ \vdots \\ X_i^{(m)} \end{bmatrix}$$

We write the data sequence in more compact form as the $m \times n$ matrix $M(X)$ whose columns are the $\mathbf{X}_i$ values:

$$M(X) = [\mathbf{X}_1 \; \mathbf{X}_2 \; \cdots \; \mathbf{X}_n]$$

Let A be an $m \times m$ invertible matrix of real numbers. The matrix A transforms the matrix $M(X)$ into an $m \times n$ matrix $M(Y)$ as follows:

$$M(Y) = AM(X)$$

Equivalently, A can be used to transform each column of $M(X)$ as follows:

$$\mathbf{Y}_i = A\mathbf{X}_i, \quad i = 1, 2, \ldots, n$$

Then, $M(Y)$ is the matrix whose columns are the $\mathbf{Y}_i$ values:

$$M(Y) = [\mathbf{Y}_1 \; \mathbf{Y}_2 \; \cdots \; \mathbf{Y}_n]$$

Let $\mathbf{Y}^{(j)}$ denote the jth row of $M(Y)(j = 1, 2, \ldots, m)$. Then, the matrix $M(Y)$ can be partitioned in two different ways:

$$M(Y) = [\mathbf{Y}_1\ \mathbf{Y}_2\ \cdots\ \mathbf{Y}_n] = \begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \\ \vdots \\ \mathbf{Y}^{(m)} \end{bmatrix}$$

The row vectors $\mathbf{Y}^{(j)}(j = 1, 2, \ldots, m)$ are the *coefficient streams* generated by the transform code; these streams are separately quantized and encoded for transmission to the decoder, thereby completing the "front end" of the transform code, which is called the *analysis stage* of the transform code and is depicted in Fig. 7a.

Note in Fig. 7a that the rates at which the separate encoders operate are $R_1, R_2, \ldots, R_m$, respectively. The separate quantizers in Fig. 7a are most typically taken to be scalar quantizers (in which case the R_i terms would be integers), but vector quantizers could be used instead (in which case all R_i would be rational numbers). If we fix the overall compression rate to be the target value R, then the bit allocation must be such that $R = R_1 + R_2 + \cdots + R_m$. The m separate bit streams generated in the analysis stage are multiplexed into one rate R bit stream for transmission to the decoder, who then obtains the separate bit streams by demultiplexing — this multiplexing/demultiplexing part of the transform code, which presents no problems, has been omitted in Fig. 7a.

The decoder must decode the separate bit streams and then combine the decoded streams by means of an inverse transform to obtain the reconstructions of the original data samples; this "back end" of the system is called the *synthesis stage* of the transform code, and is depicted in Fig. 7b.

In Fig. 7b, $\hat{\mathbf{Y}}^{(j)}$ is the row vector of length n obtained by quantizing the row vector $\mathbf{Y}^{(j)}(j = 1, 2, \ldots, m)$. The matrix $M(\hat{Y})$ is the $m \times n$ matrix whose rows are the $\hat{\mathbf{Y}}^{(j)}$ terms. The matrix $M(\hat{X})$ is also $m \times n$, and is formed by

$$M(\hat{X}) = A^{-1}M(\hat{Y})$$

Let us write the columns of $M(\hat{X})$ as

$$\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \ldots, \hat{\mathbf{X}}_n.$$

These are the reconstructions of the original data sequence (34).

The distortion (per sample) D resulting from using the transform code to encode the source $[n, A_n, P_n]$ is

$$D = n^{-1} \sum_{i=1}^{n} E[\|\mathbf{X}_i - \hat{\mathbf{X}}_i\|^2] \tag{35}$$

With the compression rate fixed at R, to optimize the transform code, one must select $R_1, R_2, \ldots, R_m$ and quantizers at these rates so that $R = R_1 + R_2 + \cdots + R_m$

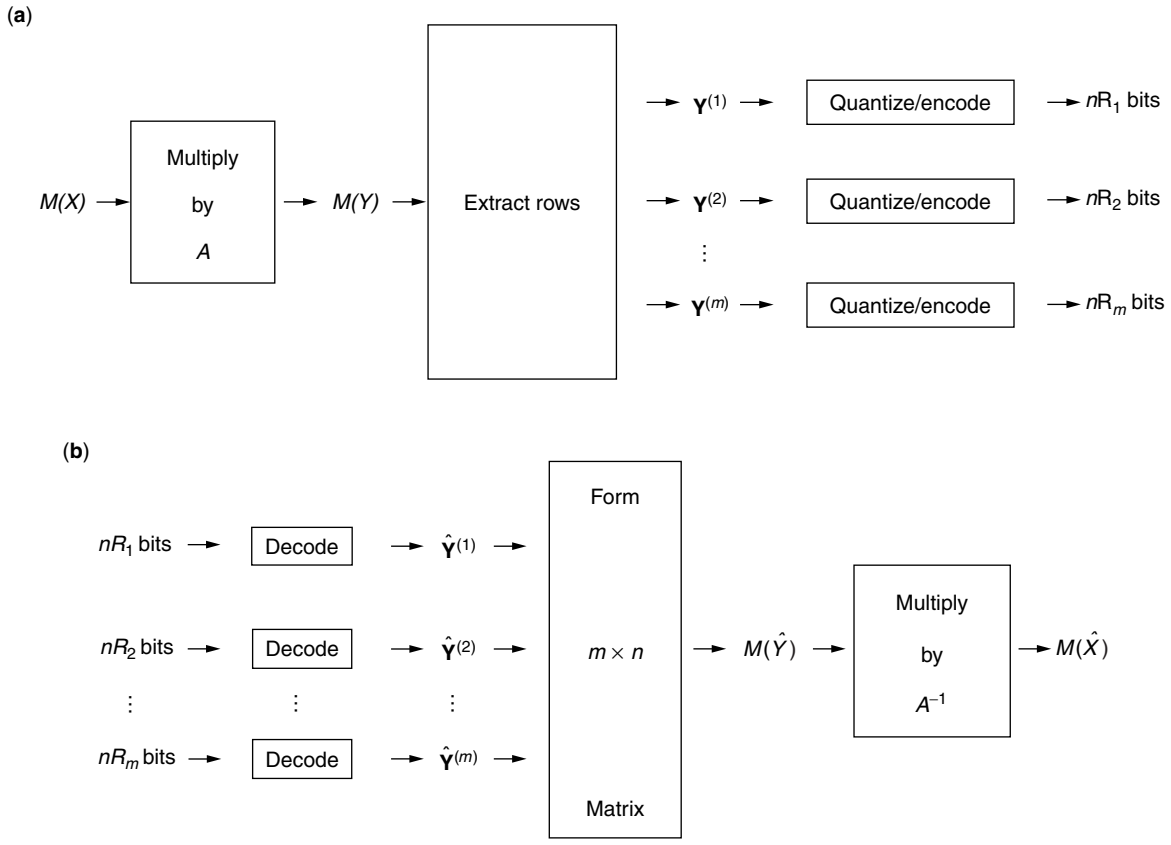


Figure 7. (**a**) Analysis stage and (**b**) synthesis stage of transform code.

and D is minimized. One can also attempt to optimize the choice of the transformation matrix A as well, although, more typically, the matrix A is fixed in advance. If A is an orthogonal matrix (i.e., the transformation is unitary), then the right side of (35) is equal to

$$n^{-1}\sum_{i=1}^{n} E[\|\mathbf{Y}_i - \hat{\mathbf{Y}}_i\|^2]$$

which makes the problem of optimally designing the transform code much easier. However, A need not be orthogonal.

Various transforms have been used in transform coding. Some commonly used transforms are the discrete cosine transform, the discrete sine transform, and the Karhunen–Loeve transform — a good account of these transforms may be found in Ref. 41, Chap. 12. Special mention should be made of the wavelet transform, which is becoming ubiquitous in state of the art compression methods. Various image compression methods employ the wavelet transform, including the EZW method [44], the SPIHT method [39], and the EBCOT method [46]; the wavelet transform is the backbone of the JPEG 2000 image compression standard [47]. The wavelet transform has also been applied to fingerprint compression, speech, audio, electrocardiogram, and video compression — a good account of these applications may be found in Ref. 45, Chap. 11.

Subband coding may be regarded as a special case of transform coding. It originally arose in the middle 1970s as an effective technique for speech compression [12]; subsequently, its applications have grown to include the compression of recorded music, image compression, and video compression. In subband coding, one filters a data sequence of length n by both a highpass filter and a lowpass filter and then throws every other filtered sample away, yielding a stream of $n/2$ highpass-filtered samples and a stream of $n/2$ lowpass-filtered samples; these two streams are separately quantized and encoded, completing the analysis stage of a subband coding system. The highpass and lowpass filters are not arbitrary; these must be complementary in the sense that if no quantization and encoding takes place, then the synthesis stage must perfectly reconstruct the original sequence of data samples. Further passes of highpass and lowpass filterings can be done after the first pass in order to yield several substreams with frequency content in different subbands; in this way, the subband coding system can become as sophisticated as one desires — one can even obtain the wavelet transform via a certain type of subband decomposition. Subband coding has received extensive coverage in several recent textbooks on multirate and multiresolution signal processing [1,49,50].

We have concentrated on lossy transform codes, but there are lossless transform codes as well. A lossless transform code is similar in concept to a lossy transform code; the main difference is that the transformed data sequence is not quantized in a lossless transform code. Good examples of lossless transform codes are (1) the transform code for text compression based on the *Burrows–Wheeler transform* [7]; and (2) grammar-based codes [25,31], which are based on *grammar transforms*.

4. NOTES ON THE LITERATURE

This article has provided an introduction to basic data compression techniques. Further material may be found in one textbook [41] that provides excellent coverage of both lossless and lossy data compression, and in another textbook [40] that provides excellent coverage of lossless data compression. There have been several excellent survey articles on data compression [5,15,20].

Various data compression standards employ the basic data compression techniques that were covered in this article. Useful material on data compression standards may be found in a book by Gibson et al. [21], which covers the JPEG and JBIG still-image compression standards, the MPEG audio and MPEG video compression standards, and multimedia conferencing standards.

4.1. Speech Compression

Speech is notoriously hard to compress — consequently, a specialized body of techniques have had to be developed for speech compression. Since these specialized techniques are not applicable to the compression of other types of data, and since this article has focused rather on general compression techniques, speech compression has been omitted; coverage of speech compression may be found in the textbook by Deller et al. [14].

4.2. Fractal Image Compression

Fractal image compression has received much attention since 1987 [2,3,16,23]. A fractal image code compresses an image by encoding parameters for finitely many contractive mappings, which, when iterated, yield an approximation of the original image. The limitations of the fractal image compression method arise from the fact that it is not clear to what extent natural images can be approximated in this way.

4.3. Differential Coding

Differential codes (such as *delta modulation* and *differential pulsecode modulation*), although not covered in this article, have received coverage elsewhere in this encyclopedia. Differential codes have become part of the standard body of material covered in a first course in communication systems [26, Chap. 6].

BIOGRAPHY

John C. Kieffer received the B.S. degree in applied mathematics in 1967 from the University of Missouri, Rolla, Missouri, and the M.S. and Ph.D. degrees in mathematics from the University of Illinois Champaign — Urbana in 1968 and 1970, respectively. He joined the University of Missouri Rolla (UMR) in 1970 as an Assistant Professor. At UMR he worked on ergodic theory and information theory. Since 1986, he has been a Professor at the University of Minneapolis Twin Cities Department of Electrical

and Computer Engineering, where he has been working on data compression. Dr. Kieffer has 70 MathSciNet publications. He was named a Fellow of the IEEE in 1993. His areas of interest are grammar-based codes, trellis-coded quantization, and the interface between information theory and computer science.

BIBLIOGRAPHY

1. A. Akansu and R. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, Wavelets*, Academic Press, San Diego, 1992.
2. M. Barnsley, *Fractals Everywhere*, Academic Press, Boston, 1988.
3. M. Barnsley and L. Hurd, *Fractal Image Compression*, Wellesley, AK Peters, Ltd., MA, 1993.
4. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
5. T. Berger and J. Gibson, Lossy source coding, *IEEE Trans. Inform. Theory* **44**: 2693–2723 (1998).
6. R. Blahut, Computation of channel capacity and rate-distortion functions, *IEEE Trans. Inform. Theory* **18**: 460–473 (1972).
7. M. Burrows and D. Wheeler, *A block-sorting lossless data compression algorithm*, unpublished manuscript, 1994.
8. J. Cleary and I. Witten, Data compression using adaptive coding and partial string matching, *IEEE Trans. Commun.* **32**: 396–402 (1984).
9. J. Conway and N. Sloane, Fast quantizing and decoding algorithms for lattice quantizers and codes, *IEEE Trans. Inform. Theory* **28**: 227–232 (1982).
10. J. Conway and N. Sloane, *Sphere Packings, Lattices, and Groups*, 2nd ed., Springer-Verlag, New York, 1993.
11. T. Cover, Enumerative source encoding, *IEEE Trans. Inform. Theory* **19**: 73–77 (1973).
12. R. Crochiere, S. Webber and J. Flanagan, Digital coding of speech in subbands, *Bell Syst. Tech. J.* **56**: 1056–1085 (1976).
13. L. Davisson, Universal noiseless coding, *IEEE Trans. Inform. Theory* **19**: 783–795 (1973).
14. J. Deller, J. Proakis and J. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, Englewood Cliffs, NJ, 1993.
15. D. Donoho, M. Vetterli, R. DeVore and I. Daubechies, Data compression and harmonic analysis, *IEEE Trans. Inform. Theory* **44**: 2435–2476 (1998).
16. Y. Fisher, ed., *Fractal Image Compression: Theory and Application*, Springer-Verlag, New York, 1995.
17. P. Fleischer, *Sufficient conditions for achieving minimum distortion in a quantizer*, *IEEE Int. Conv. Record*, 1964, Part I, Vol. 12, pp. 104–111.
18. M. Garey, D. Johnson and H. Witsenhausen, The complexity of the generalized Lloyd-Max problem, *IEEE Trans. Inform. Theory* **28**: 255–256 (1982).
19. A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer, Boston, 1992.
20. R. Gray and D. Neuhoff, Quantization, *IEEE Trans. Inform. Theory* **44**: 2325–2383 (1998).
21. J. Gibson et al., *Digital Compression for Multimedia: Principles and Standards*, Morgan-Kaufmann, San Francisco, 1998.
22. D. Huffman, A method for the construction of minimum redundancy codes, *Proc. IRE* **40**: 1098–1101 (1952).
23. A. Jacquin, Image coding based on a fractal theory of iterated contractive image transformations, *IEEE Trans. Image Proc.* **1**: 18–30 (1992).
24. N. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
25. J. Kieffer and E. Yang, Grammar-based codes: A new class of universal lossless source codes, *IEEE Trans. Inform. Theory* **46**: 737–754 (2000).
26. B. Lathi, *Modern Digital and Analog Communications Systems*, 3rd ed., Oxford Univ. Press, New York, 1998.
27. Y. Linde, A. Buzo and R. Gray, An algorithm for vector quantizer design, *IEEE Trans. Commun.* **28**: 84–95 (1980).
28. S. Lloyd, Least squares quantization in PCM, *IEEE Trans. Inform. Theory* **28**: 129–137 (1982).
29. M. Marcellin and T. Fischer, Trellis coded quantization of memoryless and Gauss-Markov sources, *IEEE Trans. Commun.* **38**: 82–93 (1990).
30. J. Max, Quantizing for minimum distortion, *IRE Trans. Inform. Theory* **6**: 7–12 (1960).
31. C. Nevill-Manning and I. Witten, Compression and explanation using hierarchical grammars, *Comput. J.* **40**: 103–116 (1997).
32. A. Nobel and R. Olshen, Termination and continuity of greedy growing for tree-structured vector quantizers, *IEEE Trans. Inform. Theory* **42**: 191–205 (1996).
33. R. Pasco, *Source Coding Algorithms for Fast Data Compression*, Ph.D. thesis, Stanford Univ., 1976.
34. W. Pennebaker, J. Mitchell, G. Langdon, and R. Arps, An overview of the basic principles of the Q-coder adaptive binary arithmetic coder, *IBM J. Res. Dev.* **32**: 717–726 (1988).
35. R. Pilc, The transmission distortion of a source as a function of the encoding block length, *Bell Syst. Tech. J.* **47**: 827–885 (1968).
36. J. Rissanen, Generalized Kraft inequality and arithmetic coding, *IBM J. Res. Dev.* **20**: 198–203 (1976).
37. J. Rissanen and G. Langdon, Arithmetic coding, *IBM J. Res. Dev.* **23**: 149–162 (1979).
38. H. Ritter, T. Martinez and K. Schulten, *Neural Computation and Self-Organizing Maps*, Addison-Wesley, Reading, MA, 1992.
39. A. Said and W. Pearlman, A new fast and efficient coder based on set partitioning in hierarchical trees, *IEEE Trans. Circuits Syst. Video Tech.* **6**: 243–250 (1996).
40. D. Salomon, *Data Compression: The Complete Reference*, Springer-Verlag, New York, 1998.
41. K. Sayood, *Introduction to Data Compression*, 2nd ed., Morgan Kaufmann, San Francisco, 2000.
42. C. Shannon, A mathematical theory of communication, *Bell System Tech. J.* **27**: 379–423, 623–656 (1948).
43. C. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE Nat. Conv. Record*, 1959, Part 4, pp. 142–163.
44. J. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. Signal Process.* **41**: 3445–3462 (1993).

45. G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Univ. Press, Wellesley, MA, 1996.

46. D. Taubman, High performance scalable image compression with EBCOT, *IEEE Trans. Image Process.* **9**: 1158–1170 (2000).

47. D. Taubman and M. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards and Practice*, Kluwer, Hingham, MA, 2000.

48. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **28**: 55–67 (1982).

49. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

50. M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

51. E. Yang and J. Kieffer, Efficient universal lossless data compression algorithms based on a greedy sequential grammar transform. I. Without context models, *IEEE Trans. Inform. Theory* **46**: 755–777 (2000).

52. K. Zeger, J. Vaisey and A. Gersho, Globally optimal vector quantizer design by stochastic relaxation, *IEEE Trans. Signal Process.* **40**: 310–322 (1992).

53. Z. Zhang, E. Yang and V. Wei, The redundancy of source coding with a fidelity criterion. I. Known statistics, *IEEE Trans. Inform. Theory* **43**: 71–91 (1997).

54. J. Ziv and A. Lempel, A universal algorithm for data compression, *IEEE Trans. Inform. Theory* **23**: 337–343 (1977).

55. J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Trans. Inform. Theory* **24**: 530–536 (1978).

DESIGN AND ANALYSIS OF A WDM CLIENT/SERVER NETWORK ARCHITECTURE

WUSHAO WEN
BISWANATH MUKHERJEE
University of California, Davis
Davis, California

1. INTRODUCTION

The client/server architecture has been one of the driving forces for the development of modern data communication networks. Most of data services, such as distributed database systems, web applications, interactive multimedia services, and so on make use of client-server network architecture. In such an architecture, clients are connected to the server via data networks. The data networks can be local-area networks, such as Ethernet, or can be wide-area networks. The required bandwidth for the client–server network is increasing with the increasing penetration of interactive multimedia and World Wide Web (WWW) applications. In a traditional client–server network, the server bandwidth is limited to several hundred megabits per second (Mbps) at the maximum. Such a bandwidth is sufficient to provide conventional data service. However, with the increasing deployment of multimedia applications and services such as network games, video on demand, virtual reality applications, and similar, the system must provide more bandwidth. For example, a video-on-demand system that provides 1000 videostreams of MPEG2 quality needs about 5–6 Gbps of bandwidth at the server side. Such high-bandwidth requirement cannot be satisfied in conventional client/server systems. The most recent technological advances in optical communication, especially wavelength-division multiplexing (WDM) technology, as well as computer hardware, make the design of such a high-bandwidth server possible. However, what kind of network architecture can we use to construct such high-bandwidth network? How well is the performance of the proposed network architecture? We describe and study a client/server WDM network architecture and its performance. This architecture is suitable for multimedia services, and is unicast-, multicast-, and broadcast-capable.

The remainder of this article is organized as follows. We outline the enabling technologies in Section 2. The WDM networking technology is summarized in Section 3. In Section 4, we describe the system architecture and connection setup protocol. Section 5 describes the system performance and analysis for unicast and multicast services. We discuss the scheduling policy for user requests in Section 6. Finally, Section 7 provides concluding remarks.

2. WDM CLIENT–SERVER NETWORK-ARCHITECTURE-ENABLING TECHNOLOGIES

Optical fiber communication plays a key role to enable high-bandwidth data connection. Now, a single fiber strand has a potential bandwidth of 50 THz corresponding to the low-loss region shown in Fig. 1. To make good use of the potential bandwidth of the fiber, WDM technology is widely used. In WDM, the tremendous bandwidth of a fiber (up to 50 THz) is divided into many nonoverlapping wavelengths, called *WDM channels* [1]. Each WDM channel may operate at whatever possible speed independently, ranging from hundreds of Mbps to tens of Gbps. WDM technology is now deployed in the backbone networks [1,2]. OC-48 WDM backbone networks have already

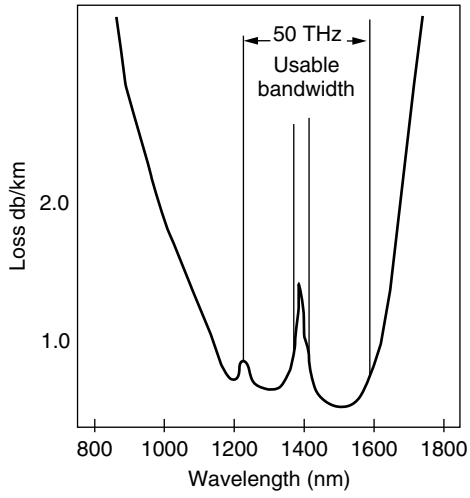


Figure 1. The low-loss region in a single-mode optical fiber.

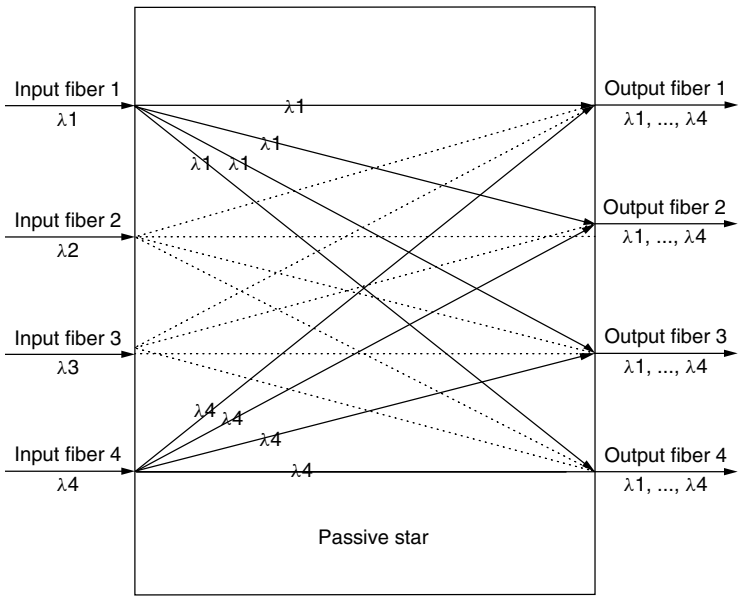


Figure 2. A 4×4 passive star.

been widely implemented, and the OC-192 WDM backbone network is being tested and implemented. The enabling technologies of the WDM client/server network architecture include tunable/nontunable wavelength transmitter and receiver, optical multiplexer, optical tap, and high-performance computer in additional to WDM technology. We outline the basic functions for each optical component.

- *Transmitter and Receiver*. An optical transmitter [1] is used to convert an electrical signal into an optical signal and inject it to the optical transmission media, namely, the fiber. An optical receiver actually consists of an optical filter [1] and an opticoelectrical signal transformer. The transmitter and receiver can be classified as tunable and nontunable. A nontunable transmitter/receiver transmits or receives signal only from a fixed frequency range, implying that only fixed wavelengths can be used. For a tunable transmitter/receiver, the passband can be tuned to different frequencies, which makes it possible to transmit or receive data for different wavelengths at different times. In a WDM system, the bandwidth of an optical fiber is divided into multiple channels (or wavelengths). Communication between two nodes is possible only when the transmitter of the source node and receiver of the destination node are tuned to the same channel during the period of information transfer. There are four types of configuration between two communication end nodes. They are fixed transmitter/fixed receiver, fixed transmitter/tunable receiver, tunable transmitter/fixed receiver, and tunable transmitter/tunable receiver. A tunable transmitter/receiver is more expensive than a fixed transmitter/receiver.
- *Optical Multiplexer*. An optical multiplexer combines signals from different wavelengths on its input ports (fiber) onto a common output port (fiber) so that a single outgoing fiber can carry multiple wavelengths at the same time.
- *Passive Star*. A passive star (see Fig. 2) is a "broadcast" device. A signal that is inserted on a given wavelength from an input fiber port will have its power equally divided among (and appear on the same wavelength on) all output ports. As an example, in Fig. 2, a signal on wavelength λ_1 from input fiber 1 and another on wavelength λ_4 from input fiber 4 are broadcast to all output ports. A "collision" will occur when two or more signals from the input fibers are simultaneously launched into the star on the same wavelength. Assuming as many wavelengths as there are fiber ports, an $N \times N$ passive star can route N simultaneous connections through itself.
- *Optical Tap*. An optical tap is similar to a one-input, two-output passive star. However, the output power in one port is much higher than that in the other. The output port with higher power is used to propagate a signal to the next network element. The other output port with lower power is used to connect to a local end system. The input fiber can carry more than one wavelength in an optical tap.

3. WDM NETWORKING ARCHITECTURE

3.1. Point-to-Point WDM Systems

WDM technology is being deployed by several telecommunication companies for point-to-point communications. This deployment is being driven by the increasing demands on communication bandwidth. When the demand exceeds the capacity in existing fibers, WDM is a more cost-effective alternative compared to laying more fibers.

WDM mux/demux in point-to-point links is now available in product form. Among these products, the maximum number of channels is 160 today, but this number is expected to increase.

3.2. Broadcast-and-Select (Local) Optical WDM Network

A local WDM optical network may be constructed by connecting network nodes via two-way fibers to a passive

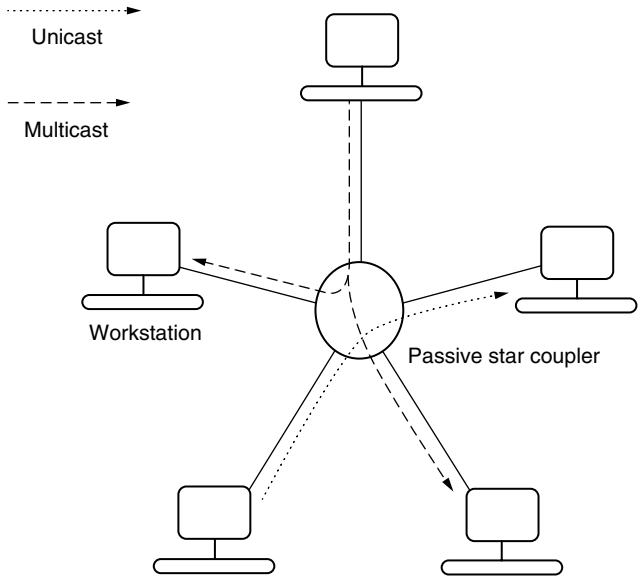


Figure 3. A passive-star-based local optical WDM network.

star, as shown in Fig. 3. A node sends its transmission to the star on one available wavelength, using a laser that produces an optical information stream. The information streams from multiple sources are optically combined by the star and the signal power of each stream is equally split and forwarded to all the nodes on their receive fibers. A node's receiver, using an optical filter, is tuned to only one of the wavelengths; hence it can receive the information stream. Communication between sources and receivers may follow one of two methods: (1) *single-hop* [3] or (2) *multihop* [4]. Also, note that, when a source transmits on a particular wavelength λ_1, more than one receiver can be tuned to wavelength λ_1, and all such receivers may pick up the information stream. Thus, the passive star can support "*multicast*" services.

Detailed, well-established discussions on these network architectures can be found elsewhere, [e.g., 1,3,4].

4. A WDM CLIENT/SERVER NETWORK ARCHITECTURE

4.1. Architecture

In our client/server architecture, the server is equipped with M fixed WDM transmitters that operate on M different wavelengths, and a conventional bidirectional Ethernet network interface that operates on a separate WDM channel is used as the control signaling channel. Every workstation (client) is equipped with a tunable WDM receiver (TR) and an Ethernet network interface. The Ethernet client and server interfaces are connected together and form the signaling channel. All control signals from the clients to the server or vice versa will be broadcast on the control channel by using the IEEE 802.3 protocol [5]. The control channel can be used as a data channel between the clients, or between a client and the server. In the normal data transmission from the server to the clients, the server uses the M data channels to provide data service to the clients. A client can tune its receiver to any of the M server channels and receive data from the server. We show in Fig. 4 a typical client/server WDM system described above. This system can be implemented using a passive star coupler, which connects the server and clients using two-way fibers [1].

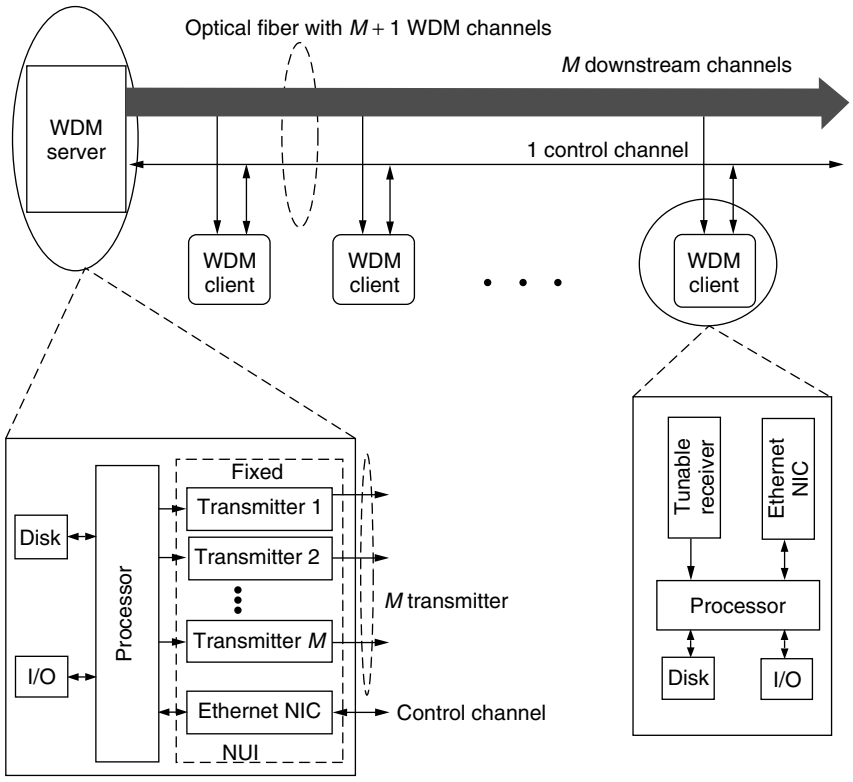


Figure 4. A typical client/server WDM network.

Since a WDM data channel can have a data rate as high as several gigabits per second, a WDM client/server network can be used for those applications that require a high volume of data transfer or those multimedia applications that require high bandwidth. However, a fiber can carry only several hundred wavelengths today, and this may not be enough if we assign a whole wavelength for a single user request. To make good use of the channel capacity, time-division multiplexing (TDM) technology is used to carve a WDM channel into multiple subchannels in time domain. The server can then assign a TDM subchannel to serve the user request. Clients must use some filtering mechanism to get the data from the subchannel. However, the client and the server must communicate with each other to negotiate on which subchannel they plan to use.

This network is also broadcast- and multicast-capable because every client can tune its receiver to any of the M data channels, and more than one client may tune their receivers to the same channel at any given time.

4.2. Operations and Signaling

In a client/server network, the server is able to provide service to clients on request. In such a network, the upstream data traffic volume (from clients to server) is generally much smaller than the downstream data traffic (from server to clients). Therefore, in a WDM client/server network, it is possible to use a single Ethernet channel to fulfill the signaling tasks between the clients and the server. All the clients' upstream traffic and the server's control traffic share the common signaling channel via statistical multiplexing. Figure 5 shows a connection setup procedure of a client's request in the WDM client/server architecture.

Figure 5 indicates that, to setup a connection for a client's request, the network must perform the following operations:

Step 1. A client sends a request message to the server via the common control channel. The message should include information such as the identification of the client, the identification of the requested-file, and the service profile. On receiving the user request, the server will respond with a request–acknowledgment message to the client. This message includes information on the request's status, the server status, and so on.

Step 2. The server processes the client's request. If no data channel or server resource is available, the request is put in the server's request queue and it waits for the server's channel scheduler (SCS) to allocate a data channel or server resource for it. If the request waits for a long time, the server should send keep-alive messages periodically to the client. As soon as the SCS knows that a channel is available, the server's scheduler will select the request (if any) with highest priority and reserve the channel for it. Then, a setup message will be sent to the client via the control channel. The setup message includes information such as the target client's identification information, request ID, the data channel that will be used, the information of the file to be sent out, and the service profile granted for the request.

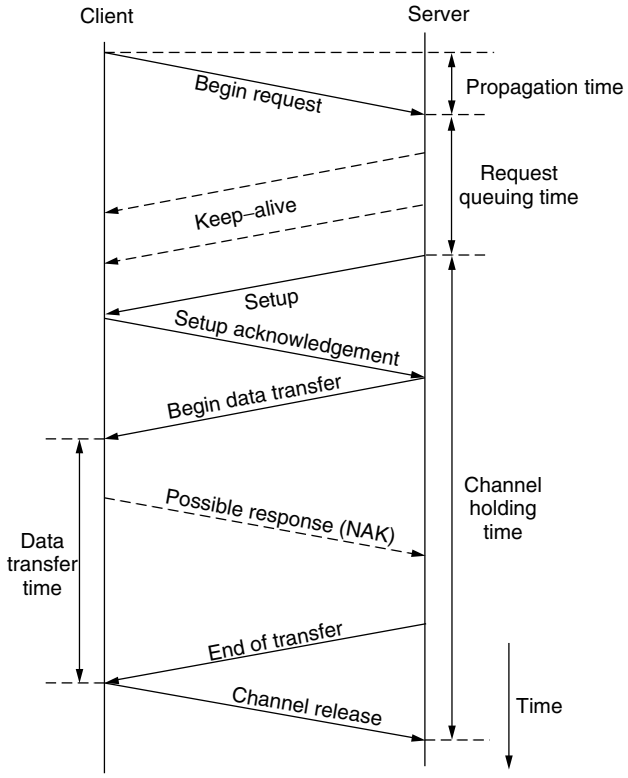


Figure 5. The request setup procedure in a WDM client/server system.

Step 3. On receiving the setup message, the target client's tunable receiver tunes to the assigned channel and sends back a setup acknowledgment message. Now, the client's receiver is waiting for the requested data.

Step 4. When the server receives the setup acknowledgment message, the connection is set up. Then, the server begins to send out the data as soon as the channel is available. If no setup acknowledgment message is received from the client within certain timeout duration, the reserved channel is released and the client's request is discarded.

Step 5. Now the connection is set up and data transfer begins. As we know, the optical channel's bit error rate (BER) is very low, on an order of 10^{-9}–10^{-12}, so it is safe to assume that the data channels have very low packet error rates. Therefore, the system uses NAK signals to indicate any error packets. No acknowledgment signal will be sent out if a packet is correctly received. However, it is possible and helpful for a client to send out some control message to synchronize the data transmission.

Step 6. After the server finishes sending the client's requested data, it sends an end-of-connection

message. The client then responds with a channel-release message to the server. Then, the server tears down the connection and frees the channel.

In this client/server WDM network, the server may receive three kinds of messages: connection request messages, acknowledgment messages, and data transmission control messages. The server should give higher priority to acknowledgment and control messages so that the system can process clients' responses promptly and minimize the channel idle-holding time.[1] It is possible that the requests from the clients come in a bursty manner. Therefore, sometimes, the server may not have enough channels to serve the requests immediately. In that case, the request is buffered. When a channel is available again, there may be more than one client waiting for a free channel. Therefore, the server should have some mechanism to allocate a free channel when it is available. The mechanism used to decide which request should be assigned with the free channel when there are multiple requests waiting for it is called the (SCS) mechanism [6].

4.3. Internetworking Technology

The WDM client/server network architecture is easy to internetwork. In this network architecture, the server has very high bandwidth; therefore, it is not cost-effective to connect the server with the outside network directly. Instead, the clients are connected to outside networks. Figure 6 shows the internetworking scheme where the clients act as proxy servers for outside networks. A client now has two major tasks: handling user requests and caching data. An end user who wants to access some data service, connects to the nearest WDM client. This can be done by a standard client/server service, which implies that no modification is necessary for the end user. On receiving a service request from the end user, the WDM client then analyzes and processes the request. Batching[2] may be used to accommodate requests for the same service from different users. If the user's requested file or program is stored at the WDM client, the request can be served directly without referring to the WDM server; otherwise, the WDM client will contact the server for the data. However, no matter whether the service is provided by the WDM client or the WDM server, it is transparent to the end users.

This internetworking architecture is suitable for multimedia service providers. A service provider can deploy the WDM clients in different cities to provide service to end users. Because the system is broadcast- and multicast-capable, it has very good scalability.

5. SYSTEM PERFORMANCE ANALYSIS

To analyze the system's performance, we consider an example system that uses fixed-length cells (k bytes/cell) to send data from the server to the clients. Every cell occupies exactly one time slot in a WDM channel. The control messages between a client and the server also use

[1] A channel's idle holding time is the time that the channel is reserved for a client but not used for data transfer.

[2] Batching makes use of multicast technology. It batches multiple requests from different users who ask for the same service and use a single stream from the server to serve the users at the same time by using multicasting [6].

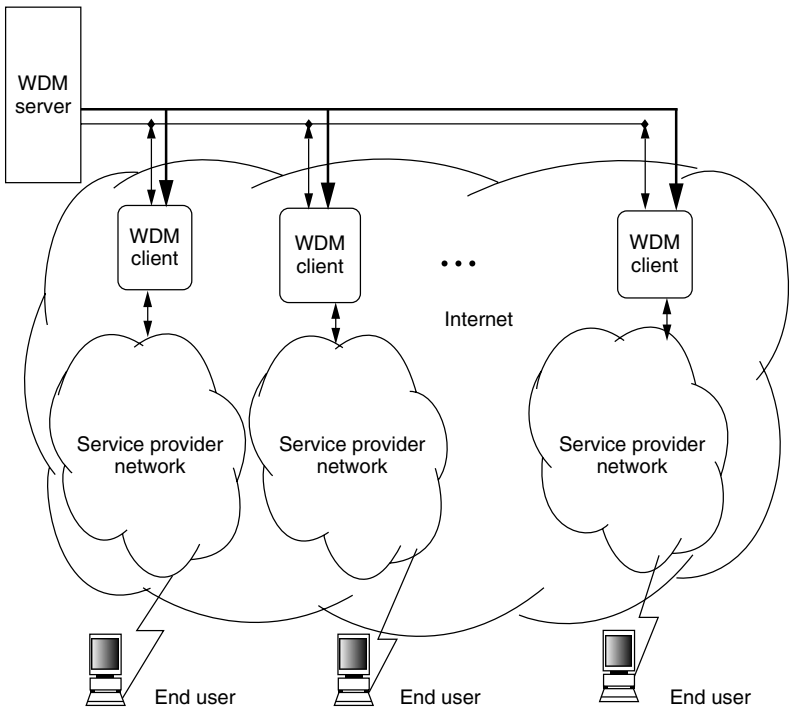


Figure 6. WDM client/server internetworking.

fixed-length cells (k_1 bytes/cell). We use E_1 to represent the bit-error rate (BER) for an optical channel.

5.1. Bottleneck Analysis

5.1.1. Unicast Service. We assume that the system uses negative acknowledgement (NAK) to ask the server to resend the packet that is received with error. It is clear that the expected number of NAK signals per second (denoted as A_1) that will be sent to the control channel is

$$A_1 = \frac{MB}{k}(1 - (1 - E_1)^k) \tag{1}$$

where M is the number of channels and B is the bandwidth for a channel. The BER for an optical channel (E_1) is very small, on the order of 10^{-11}, so Eq. (1) can be simplified as $A_1 = MBE_1$.

Now, suppose the requests from all stations form a Poisson process with average request rate λ_r requests per second. We also assume that the server responds to a user's request by sending back an acknowledgement signal after a random delay. The delay is negative exponential distributed with mean $1/\lambda_r$ minutes. As indicated by the connection setup procedure, the system exchanges five messages via the control channel for every connection in a normal situation.[3] With the additional assumption that the NAK messages are also generated as a Poisson process with average arrival rate λ_n requests/second, the number of cells sent to the Ethernet channel is a Poisson process with total arrival rate λ, ($\lambda = E_1MB + 5\lambda_r$) at the maximum.[4]

Suppose that the mean file size is $\overline{f}$. Then to serve all requests with limited delay, the system should satisfy the condition $\lambda_r * \overline{f} < MB$. So $\lambda_r < MB/\overline{f}$. Therefore, the bandwidth requirement for the control channel should satisfy

$$B_e = \lambda k_1 < \frac{K_1}{p}\left(\frac{E_1MB + 5MB}{f}\right) = \frac{MBk_1}{p}\left(\frac{E_1 + 5}{\overline{f}}\right) \tag{2}$$

where p is the maximum offered load allowed on the control channel. In this above equation, $E_1 \ll 5/\overline{f}$ in the actual system. Figure 7 shows the plot of expected control channel bandwidth (B_e) requirement vs. the average file size ($\overline{f}$) in the server with default parameter $p = 0.5, E_1 = 10^{-8}, M = 10, B = 1$ Gbps, and $k_1 = 40$ bytes. From the plot, we notice that, when $\overline{f}$ is small, B_e must be very high. For example, when $\overline{f} = 1$ kbyte, the system requires a bandwidth of 500 Mbps for the control. However, with $f = 1$ Mbyte, the required bandwidth B_e is only about 0.5 Mbps. Clearly, with such a low bandwidth requirement, the control channel is not a bottleneck. When the average file size $\overline{f}$ is between 10 and 100 kbytes, which is a practical file size, the required bandwidth B_e is between 5 and 50 Mbps. From Fig. 7, we can conclude that the WDM client/server network architecture is useful to provide those services that need to transfer a lot of large files. If the system needs to support small files, then the control channel's bandwidth must be increased.

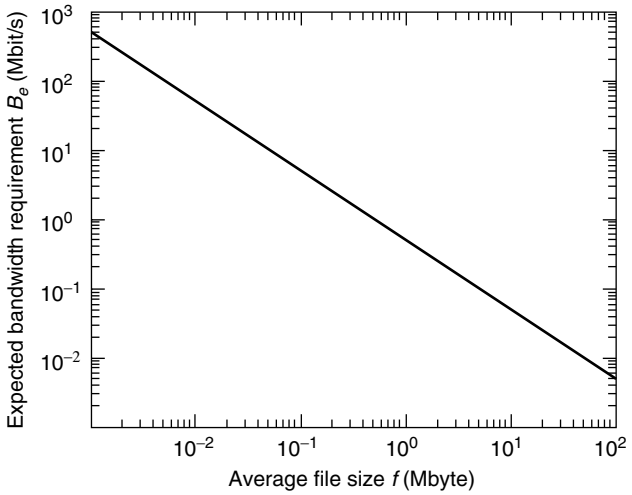


Figure 7. Expected bandwidth requirement of the control channel versus the average file size in the server.

5.1.2. Multicast Service. We have stated that the WDM client-server network is multicast- and broadcast-capable. If the system provides on-demand multicast services, then the server can group together user requests that ask for the same file coming in a certain period and use a single-server channel to multicast the requested file to all users. In an actual implementation, a WDM channel may be carved into several subchannels, and any multicast group will only use a subchannel instead of a whole WDM channel. A multicast service system scales very well even with a limited number of data channels. For example, if a server has 10-Gbps data bandwidth, it can provide about 1800 MPEG2 or 6700 MPEG1 video channels at the same time. A client/server system providing video-on-demand service, which only provides several hundred different video programs at the same time, can easily support any number of user requests by using batching. However, all the request messages and control messages are sent via the Ethernet channel, and the Ethernet channel capacity is limited. Therefore, it is possible that the Ethernet may be a bottleneck. However, in an actual system, this is unlikely to happen. Let us assume, on the average, that every request needs to exchange m 40-byte messages between a client and the server. If we assume $m = 10$ in the multicast system, a control channel with 100 Mbps can accommodate a request rate as high as $\lambda_r = 31{,}250$ requests/s. Such a high request rate is unlikely to happen in any actual system. Therefore, the control channel should not be a bottleneck.

For broadcast service, there is not a lot of interaction between clients and the server. The server notifies the clients what kind of file or program is sent in each channel. A client then chooses its interested channel to receive the

[3] We omit the keep-alive messages in this discussion.

[4] The aggregation of several Poisson processes is also a Poisson process. The average arrival rate of the aggregated process is the sum of the arrival rates of all the original processes.

data. Therefore, bottleneck analysis does not apply in this service mode.

5.2. System Performance Illustrative Study

This subsection studies the system performance in terms of delay, throughput, and average queuing length. Let us first define several keywords:

- *A Request's System Time*. A request's system time is measured from the time that a user sends out the request message to the time that the user receives all data. Let us denote the time as S_t. Then, from Fig. 5, we can express S_t as

$$S_t = 2R + t_q + t_d \tag{3}$$

  where R is the round-trip time between the client and the server; t_q is the request's queuing time, defined as the time from the instant a request reaches the server to the time the request starts to be served by the server; and t_d is the data transfer time. For every client's request, the round-trip time and the data transfer time are fixed. The only optimizable factor is the queuing time for a given system. Therefore, in order to reduce the response time, we should have a good scheduling algorithm in the server so that we can minimize the average request delay.
- *A Request's Channel Holding Time*. The channel holding time for a request, denoted as H, can be expressed by

$$H = 2R + t_d = D - t_q \tag{4}$$

  Therefore, for a given request, the channel holding time is almost nonoptimizable, because the round-trip time and data transfer time are all nonoptimizable for a given request in a specific system.
- *System Throughput*. When the system operates in the stable state, it should be able to serve all user requests. So the system throughput should equal the average request rate multiplied by the average file size. The maximum reachable throughput in the system, denoted as $\theta_{\max}$, can be express as

$$\theta_{\max} = MB\frac{\overline{f}/B}{\overline{f}/B + 2\overline{R}} \tag{5}$$

where M is total number of data channels, B is the channel capacity, $\overline{R}$ is the average round-trip time from the clients to the server, and $\overline{f}$ is the average file size. Here, $2\overline{R}$ is the average overhead conceived by the server to setup a connection.

We analyze a system's performance based on the following assumptions:

- *Clients' requests* form a Poisson process. The request rate is λ_r requests/s. So the interarrival times are negative exponentially distributed with mean $1/\lambda_r$ minutes.
- There are M data channels altogether. Each channel's capacity is B bps.
- *The file sizes* that users request from the server are exponentially distributed with average file size $\overline{f}$. We also assume the average file size to be relatively large. Therefore, the data transfer time is much larger than the round-trip time and the client's receiver tuning time, and we can omit the round-trip time and client's receiver-tuning time in our analysis.
- *The scheduler uses first-come, first-served* (*FCFS*) policy.

Using on the above assumptions, we analyze the system as follows. For FCFS scheduling policy, we can model the system as an $M/M/m$ queuing system. In this system, a data channel is regarded as a server in the model. The service time is equal to the channel holding time. We have assumed that the file sizes that users requested from the server are exponentially distributed with average $\overline{f}$. Therefore, the channel holding time at each channel is negative exponential distributed with mean $\mu = B/\overline{f}$. Let us define $\rho = \lambda_r/M\mu$. Clearly the system should require $0 \le \rho < 1$. If $\rho \ge 1$, then the system's service rate will be smaller than the arrival rate, which means that the waiting queue will be built up to infinity, and the system will become unstable. On the basis of these parameters and the $M/M/m$ model, we have the following results [7]:

- *The average number of busy channels* U_b in the system is given by

$$E[U_b] = M\rho = \frac{\lambda_r}{\mu} \tag{6}$$

- *The throughput* of the system is given by $\lambda_r\overline{f}$. This is because the average number of the served requests and the arrival requests should be in balance.
- *The average queue length* Q_n of the system is given by

$$Q_n = M\rho + \frac{(M\rho)^M}{M!}\frac{\rho}{(1-\rho)^2}\pi_0 \tag{7}$$

  where π_0 is the stationary zero queue-length probability and is given by

$$\pi_0 = \left[1 + \sum_{i=1}^{M-1}\frac{(M\rho)^i}{i!} + \frac{(M\rho)^M}{(M)!}\frac{1}{1-\rho}\right]^{-1} \tag{8}$$

- *The average queuing time* is given by the average system time minus the service time (we omit the

round-trip time here). Therefore, we can calculate the average queuing time from the expression

$$E[t_q] = \frac{1}{\mu}\frac{(M\rho)^M}{M!}\frac{\pi_0}{M(1-\rho)^2} \tag{9}$$

6. SCHEDULING POLICIES

In Section 4, we pointed out that the delay for a client's request can be decreased by reducing the queuing time with a proper scheduler. However, the FCFS scheduling policy does not account for a request's channel holding time in allocating an available channel. Therefore, it is possible that some long-time tasks will occupy the server channels for very long time and the short-time tasks have to wait for very long time to be served. Several scheduling policies, such as priority queues, and feedback priority queues, are available to solve this problem. We discuss these policies below.

Priority Queues (PQ) Scheduling Policy. In a priority-queue scheduling policy, the server maintains several queues with different priorities. When no channel is available, the requests for small files will be assigned a higher priority and placed in the higher-priority queue in FCFS manner, and requests for larger files will be put in lower-priority queue. When a channel is available, it is first allocated to a request in the higher-priority queue. Requests in a lower-priority queue cannot be allocated a channel unless all higher-priority queues are empty. The PQ method can greatly reduce the average queuing time. However, there is a starvation problem in this policy. When the request rate is high, requests that ask for large files may never be served because they remain in the lower-priority queue forever and the channels are always allocated to the requests in the higher-priority queue.

Priority Feedback Queues (PFQ) Scheduling Policy. To avoid the starvation problem in the PQ scheduling policy, we can use the priority feedback queues (PFQ) scheduling policy, which allows a request to move from a lower-priority queue to a higher-priority queue if it waits too long. One possible method is to use a threshold function $T_r = f(i, f, t)$. It is the function of i (the current queue it belongs to), the request's file size f, and the waiting time t. Let $T_h(i)$ be the threshold value for the current queue i. Then, we can set the scheduling policy as follows. If $T_r < T_h(i)$, then the request will remain in the same queue; otherwise, it will move up to the higher-priority queue. A scheduler will always select the request with highest priority and allocate the channel to it.

The PFQ scheduling policy solves the starvation problem, but the implementation is more complex.

7. CONCLUSION

We described a WDM client/server network architecture based on a passive-star-coupler-based broadcast-and-select network. In this architecture, all downstream data traffic uses WDM data channels and all upstream requests, upstream control messages, and downstream control messages use an Ethernet channel (called the *control channel*). This architecture is broadcast and multicast capable. We described a detailed point-to-point connection setup procedure for this architecture. An internetworking scheme was also provided. This system architecture is appropriate for an asymmetric data transmission system in which the downstream traffic volume is much higher than the upstream traffic volume. The system's performance was analyzed in terms of whether the control channel or the data channels are the bottleneck. We also analyzed the request delay, request's channel holding time, and system throughput following the model description. We concluded that the control channel is not a bottleneck in the system. When the user request rate is high, the server channel scheduler is the most important adjustable factor to reduce the user response time. An illustrative analysis of FCFS scheduling policy for the system was also provided. To explore WDM technology further, the literature please refer to [1,10,11].

BIBLIOGRAPHY

1. B. Mukherjee, *Optical Communication Networks*, McGraw-Hill, New York, 1997.
2. *http://www.ipservices.att.com/backbone/*.
3. B. Mukherjee, WDM-based local lightwave networks — Part I: Single-hop systems, *IEEE Network Mag.* **6**: 12–27 (May 1992).
4. B. Mukherjee, WDM-based local lightwave networks — Part II: Multiple-hop systems, *IEEE Network Mag.* **6**: 20–32 (July 1992).
5. Digital Equipment Corporation, INTEL corporation, and XEROX Corporation, *The Ethernet: A local Area Network Data Link Layer and Physical Layer Specification*, Sept. 1980.
6. W. Wen, G. Chan, and B. Mukherjee, Token-tray/weighted queuing-time (TT/WQT): An adaptive batching policy for near video-on-demand, *Proc. IEEE ICC 2001*, Helsinki, Finland, June 2001.
7. C. G. Cassandras, *Discrete Event Systems — Modeling and performance Analysis*, Aksen Associates, 1993.
8. P. E. Green, Optical networking update, *IEEE J. Select. Areas Commun.* **14**: 764–779 (June 1996).
9. W. J. Goralski, *SONET*, McGraw-Hill, New York, 2000.
10. R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Perspective*, Morgan-Kaufmann, San Fransciso, 1998.
11. T. E. Stern and K. Bala, *Multiplewavelength Optical Networks, a Layered Approach*, Addison-Wesley, Reading, MA, 1999.

DESIGN AND ANALYSIS OF LOW-DENSITY PARITY-CHECK CODES FOR APPLICATIONS TO PERPENDICULAR RECORDING CHANNELS

EROZAN M. KURTAS
ALEXANDER V. KUZNETSOV
Seagate Technology
Pittsburgh, Pennsylvania

BANE VASIC
University of Arizona
Tucson, Arizona

1. INTRODUCTION

Low-density parity-check (LDPC) codes, first introduced by Gallager [19], are error-correcting codes described by sparse parity-check matrices. The recent interest in LDPC codes is motivated by the impressive error performance of the Turbo decoding algorithm demonstrated by Berrou et al. [5]. Like Turbo codes, LDPC codes have been shown to achieve near-optimum performance [35] when decoded by an iterative probabilistic decoding algorithm. Hence LDPC codes fulfill the promise of Shannon's noisy channel coding theorem by performing very close to the theoretical channel capacity limit.

One of the first applications of the Gallager LDPC codes was related to attempts to prove an analog of the Shannon coding theorem [47] for memories constructed from unreliable components. An interesting scheme of memory constructed from unreliable components was proposed by M. G. Taylor, who used an iterative threshold decoder for periodic correction of errors in memory cells that degrade steadily in time [52]. The number of elements in the decoder per one memory cell was a priori limited by a capacity-like constant C as the number of memory cells approached infinity. At the time, the Gallager LDPC codes were the only class of codes with this property, and therefore were naturally used in the Taylor's memory scheme. For sufficiently small probabilities of the component faults, using Gallager LDPC codes of length N, Taylor constructed a reliable memory from unreliable components and showed that the probability of failure after T time units, $P(T, N)$, is upper-bounded by $P(T, N) < A_1 T N^{-\alpha}$, where $0 < \alpha < 1$ and A is a constant. A similar bound with a better constant α was obtained [26]. In fact, Ref. 26 report even the tighter bound $P(T, N) < A_2 T \exp\{-\gamma N^{\beta}\}$, with a constant $0 < \beta < 1$. Using a slightly different ensemble of LDPC and some other results [60], Kuznetsov [27] proved the existence of LDPC codes that lead to the "pure" exponential bound with a constant $\beta = 1$ and decoding complexity growing linearly with code length. A detailed analysis of the error correction capability and decoding complexity of the Gallager-type LDPC codes was done later by Pinsker and Zyablov [61].

Frey and Kschischang [18] showed that all compound codes including Turbo codes [5], classical serially concatenated codes [3,4], Gallager's low-density parity-check codes [19], and product codes [20] can be decoded by Pearl's belief propagation algorithm [41] also referred to as the *message-passing algorithm*. The iterative decoding algorithms employed in the current research literature are suboptimal, although simulations have demonstrated their performance to be near optimal (e.g., near maximum likelihood). Although suboptimal, these decoders still have very high complexity and are incapable of operating in the >1-Gbps (gigabit-per-second) regime.

The prevalent practice in designing LDPC codes is to use very large randomlike codes. Such an approach completely ignores the quadratic encoding complexity in length N of the code and the associated very large memory requirements. MacKay [35] proposed a construction of irregular LDPC codes resulting in a deficient rank parity-check matrix that enables fast encoding since the dimension of the matrix to be used in order to calculate parity bits turns out to be much smaller than the number of parity check equations. Richardson et al. [44] used the same encoding scheme in the context of their highly optimized irregular codes. Other authors have proposed equally simple codes with similar or even better performance. For example, Ping et al. [42] described concatenated tree codes, consisting of several two-state trellis codes interconnected by interleavers, that exhibit performance almost identical to turbo codes of equal block length, but with an order of magnitude less complexity.

An alternative approach to the random constructions is the algebraic construction of LDPC codes using finite geometries [24,25]. Finite geometry LDPC codes are quasicyclic, and their encoders can be implemented with linear shift registers with feedback connections based on their generator polynomials. The resulting codes have been demonstrated to have excellent performance in AWGN, although their decoder complexities are still somewhat high. Vasic [54–56] uses balanced incomplete block designs (BIBDs) [6,11] to construct regular LDPC codes. The constructions based on BIBD's are purely combinatorial and lend themselves to low complexity implementations.

There have been numerous papers on the application of turbo codes to recording channels [14–16,22] showing that high-rate Turbo codes can improve performance dramatically. More recent work shows that similar gains can be obtained by the use of random LDPC and Turbo product codes [17,28–30]. Because of their high complexity, the LDPC codes based on random constructions cannot be used for high speed applications (>1 Gbps) such as the next generation data storage channels.

In this article we describe how to construct LDPC codes based on combinatorial techniques and compare their performance with random constructions for perpendicular recording channels. In Section 2 we introduce various deterministic construction techniques for LDPC codes. In Section 3 we describe the perpendicular channel under investigation which is followed by the simulation results. Finally, we conclude in Section 4.

2. DESIGN AND ANALYSIS OF LDPC CODES

In this section we introduce several deterministic constructions of low-density parity-check codes. The constructions are based on combinatorial designs. The

codes constructed in this fashion are "well-structured" (cyclic and quasicyclic) and, unlike random codes, can lend themselves to low-complexity implementations. Furthermore, the important code parameters, such as minimum distance, code rate and the graph girth, are fully determined by the underlying combinatorial object used for the construction. Several classes of codes with a wide range of code rates and minimum distances are considered First, we introduce a construction using general Steiner system (to be defined later), and then discuss projective and affine geometry codes as special cases of Steiner systems. Our main focus are regular Gallager codes [19] (LDPC codes with constant column and row weight), but as it will become clear later, the combinatorial method can be readily extended to irregular codes as well. The first construction is given by Vasic [54–56] and exploits resolvable Steiner systems to construct a class of high-rate codes. The second construction is by Kou et al. [24,25], and is based on projective and affine geometries.

The concept of combinatorial designs is well known, and their relation to coding theory is profound. The codes as well as their designs are combinatorial objects, and their relation does not come as a surprise. Many classes of codes including extended Golay codes and quadratic residue (QR) codes can be interpreted as designs (see Refs. 8 and 48 and the classical book of MacWilliams and Sloane [37]). The combinatorial designs are nice tools for code design in a similar way as bipartite graphs are helpful in visualizing the message passing decoding algorithm. Consider an (n, k) linear code C with parity-check matrix H [31]. For any vector $x = (x_v)_{1 \le v \le n}$ in C and any row of H

$$\sum_v h_{c,v} \cdot x_v = 0, 1 \le c \le n - k \tag{1}$$

This is called the *parity-check equation*. To visualize the decoding algorithm, the matrix of parity checks is represented as a bipartite graph with two kinds of vertices [38,58,50]. The first subset (B), consists of bits, and the second is a set of parity-check equations (V). An edge between a bit and an equation exists if the bit is involved in the check. For example, the bipartite graph corresponding to

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

is shown in Fig. 1.

2.1. Balanced Incomplete Block Design (BIBD)

Now we introduce some definitions and explain the relations of 2-designs to bipartite graphs. A *balanced incomplete block design* (BIBD) is a pair (V,B), where V is a v-element set and B is a collection of $b \cdot k$-subsets of V, called blocks, such that each element of V is contained in exactly r blocks and any 2-subset of V is contained in exactly λ blocks. A design for which every block contains the same number k of points, and every point is contained in the same number r of blocks is called a *tactical configuration*. Therefore, BIBD can be considered as a special case of a tactical configuration. The notation BIBD(v, k, λ) is used for a BIBD on v *points, block size* k, and index λ. The BIBD with a block size $k = 3$ is called a *Steiner triple system*. A BIBD is called *resolvable* if there exists a nontrivial partition of its set of blocks B into *parallel classes*, each of which in turn partitions the point set V. A resolvable Steiner triple system with index $\lambda = 1$ is called a *Kirkman* system. These combinatorial objects originate from Kirkman's famous schoolgirl problem posted in 1850 [23] (see also Refs. 49 and 43). For example, collection $B = \{B_1, B_2, \ldots, B_7\}$ of blocks $B_1 = \{0, 1, 3\}$, $B_2 = \{1, 2, 4\}$, $B_3 = \{2, 3, 5\}$, $B_4 = \{3, 4, 6\}$, $B_5 = \{0, 4, 5\}$, $B_6 = \{1, 5, 6\}$ and $B_7 = \{0, 2, 6\}$ is a BIBD(7,3,1) system or a Kirkman system with $v = 7$, and $b = 7$.

We define the block-point incidence matrix of a (V, B) as a $b \times v$ matrix $A = (a_{ij})$, in which $a_{ij} = 1$ if the jth element of V occurs in the ith block of B, and $a_{ij} = 0$ otherwise. The point-block incidence matrix is A^T.

The block-point matrix for the BIBD(7,3,1) described above is

$$A = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and the corresponding bipartite graph is given in Fig. 2.

Each block is incident with the same number of points k, and every point is incident with the same number r of blocks. If $b = v$, and hence $r = k$, the BIBD is called *symmetric*. The concepts of a symmetric BIBD(v, k, λ) with $k \ge 3$ and of a finite projective plane are equivalent (see Ref. 34, Chap. 19). If one thinks of points as parity-check equations and of blocks as bits in a linear block code, then the A^T defines a matrix of parity checks H of a Gallager code [19,36]. The row weight of A is k, column weight is r, and the code rate is $R = [b - \text{rank}(H)]/b$.

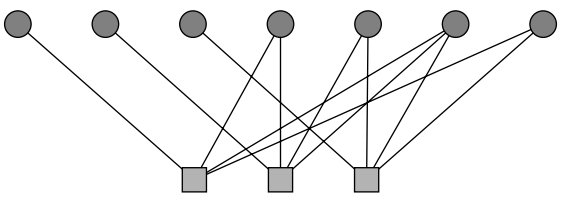


Figure 1. An example of a bipartite graph.

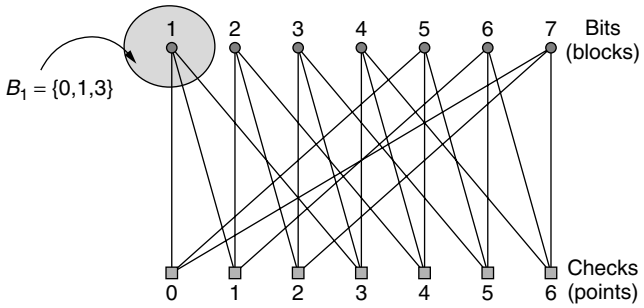


Figure 2. The bipartite graph representation of the Kirkman (7,3,1) system.

It is desirable to have each bit "checked" in as many equations as possible, but because of the iterative nature of the decoding algorithm, the bipartite graph must not contain short cycles. In other words, the graph *girth* (the length of the shortest cycle) must be large [7,51]. These two requirements are contradictory, and the tradeoff is especially difficult when we want to construct a code that is both short and has a high rate. The girth constraint is related to the constraint that every t-element subset of V is contained in as few blocks as possible. If each t-element subset is contained in exactly λ blocks, the underlying design is known as t design. An example of a 5-design is the extended ternary Golay code [8]. However, the codes based on t designs ($t > 2$) have short cycles, and therefore we will restrict ourselves to the 2-designs (i.e., BIBD) or more specifically to the designs with the index $\lambda = 1$. The $\lambda = 1$ constraint means that no more than one block contains the same pair of points or, equivalently, that there are no cycles of length four in a bipartite graph. The lower bound on a rate of a 2-(v, k, λ)-design-based code is given by

$$R \geq \frac{\lambda \dfrac{v(v-1)}{k(k-1)} - v}{\lambda \dfrac{v(v-1)}{k(k-1)}} \tag{2}$$

This bound follows from the basic properties of designs (Ref. 9, Lemma 10.2.1 and Corollary 10.2.2, pp. 344–345). By counting ones in H across the rows and across the columns, we conclude that

$$b \cdot k = v \cdot r \tag{3}$$

If we fix the point u, and find the number of pairs (u, w), $u \neq w$, we arrive to the relation

$$r \cdot (k-1) = \lambda \cdot (v-1) \tag{4}$$

Since the point-block incidence matrix (matrix of parity checks) H has v rows and b columns ($v \leq b$), the rank of A cannot be larger than v, and the code rate, which is $R = [b - \text{rank}(A^T)]/b$ cannot be smaller than $(b-v)/b$. Dividing (3) and (4) yields (2). A more precise characterization of code rate can be obtained by using the rank (and p rank) of the incidence matrix of 2-designs as explained by Hamada [21]. In the case of t-designs, $t > 1$ the number of blocks is $b = \lambda\binom{v}{t}/\binom{k}{t}$ (see Ref. 34, p. 191).

We have shown that the concept of BIBD offers a useful tool for designing codes. Let us now show a construction of Steiner triple systems using difference families of Abelian groups. Let V be an additive Abelian group of order v. Then $t \cdot k$-element subsets of $V, B_i = \{b_{i,1}, \ldots, b_{i,k}\} 1 \leq i \leq t$ form a (v, k, λ) *difference family* (DF) if every nonzero element of V can be represented exactly λ ways as a difference of two elements lying in a same member of a family, that is, it occurs λ times among the differences $b_{i,m} - b_{i,n}, 1 \leq i \leq t, 1 \leq m, n \leq k$. The sets B_i are called *base blocks*. If V is isomorphic with Z_v, a group of integers modulo v, then a (v, k, λ) DF is called a *cyclic difference family* (CDF). For example, the block $B_1 = \{0, 1, 3\}$ is a base block of a (7,3,1) CDF. To illustrate this, we create an array $\Delta^{(1)} = (\Delta_{i,j})$, of differences $\Delta^{(1)}{}_{i,j} = b_{1,i} - b_{1,j}$

$$\Delta^{(1)} = \begin{bmatrix} 0 & 6 & 4 \\ 1 & 0 & 5 \\ 3 & 2 & 0 \end{bmatrix}$$

Given base blocks B_j, $1 \leq j \leq t$, the orbits $B_j + g$ can be calculated as a set $\{b_{j,1} + g, \ldots, b_{j,k} + g\}$, where $g \in V$. A construction of a BIBD is completed by creating orbits for all base blocks. If the number of base blocks in the difference families is t, then the number of blocks in a BIBD is $b = tv$. For example, it can be easily verified (by creating the array Δ) that the blocks $B_1 = \{0, 1, 4\}$ and $B_2 = \{0, 2, 7\}$ are the base block of a (13,3,1) CDF of a group $V = Z_{13}$. The two parallel classes (or orbits) are as given below:

$$B_1 + g = \left\{ \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 0 \\ 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 0 & 1 & 2 & 3 \end{matrix} \right\}$$

$$B_2 + g = \left\{ \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 0 & 1 \\ 7 & 8 & 9 & 10 & 11 & 12 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \right\}$$

The corresponding matrix of parity checks is

$$H = \begin{bmatrix} 1&0&0&0&0&0&0&0&0&1&0&0&1&1&0&0&0&0&0&1&0&0&0&0&1&0 \\ 1&1&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&1&0&0&0&0&1 \\ 0&1&1&0&0&0&0&0&0&0&0&1&0&1&0&1&0&0&0&0&0&1&0&0&0&0 \\ 0&0&1&1&0&0&0&0&0&0&0&0&1&0&1&0&1&0&0&0&0&0&1&0&0&0 \\ 1&0&0&1&1&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0 \\ 0&1&0&0&1&1&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0 \\ 0&0&1&0&0&1&1&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0 \\ 0&0&0&1&0&0&1&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0&0&0&0&0 \\ 0&0&0&0&1&0&0&1&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&1&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0&0&0 \\ 0&0&0&0&0&0&1&0&0&1&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0&0 \\ 0&0&0&0&0&0&0&1&0&0&1&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0 \\ 0&0&0&0&0&0&0&0&1&0&0&1&1&0&0&0&0&0&1&0&0&0&0&1&0&1 \end{bmatrix}$$

This matrix contains only columns of weight 3. Of course, since the order of the group in our example is low, the code rate is low as well ($R \geq \frac{1}{2}$). Generally, given a (v, k, λ) CDF, a $t \cdot k$-element subset of Z_v, with base blocks $B_i = \{b_{i,1}, \ldots, b_{i,k}\}, 1 \leq i \leq t$, the matrix of parity checks can be written in the form

$$H = [H_1 \, H_2 \, \cdots \, H_t] \tag{5}$$

where each submatrix is a circulant matrix of dimension $v \times v$. Formula (5) indicates that the CDF codes have a quasicyclic structure similar to the Townsend and Weldon [53] self-orthogonal quasicyclic codes and Weldon's [57] difference set codes. Each orbit in the design corresponds to one circulant submatrix in a matrix of parity checks of quasicyclic codes.

The codes based on Z_v are particularly interesting because they are conceptually extremely simple and have a structure that can be easily implemented in hardware. Notice also that for a given constraint (v, k, λ) the CDF-based construction maximizes the code rate because for a given v the number of blocks is maximized. The code

rate is independent of the underlying group. Other groups different than Z_v may lead to similar or better codes.

2.2. Constructions for Cyclic Difference Families

As we have shown, it is straightforward to construct a BIBD once the CDF is known. However, finding the CDF is a much more complicated problem and solved only for some values of v, k, and λ. Now we describe how to construct a CDF.

2.2.1. The Netto's First Construction. This scheme (see Refs. 10 and 12, p. 28) is applicable if $k = 3$, and v is a power of a prime, such that $v \equiv 1 \pmod 6$. Since v is a power of a prime, then Z_v is a Galois field [GF(v)]. Let Ψ be a multiplicative group of the field, and let ω be its generator [a primitive element in GF(v)] [39]. Write v as $v = 6t + 1, t \geq 1$, and for d, a divisor of $v - 1$, denote by Ψ^d the group of dth powers of ω in GF($6t + 1$), and by $\omega^i\Psi^d$ the coset of dth powers of ω^i. Then the set $\{\omega^i\Psi^{2t} \mid 1 \leq i \leq t\}$ defines a $(6t + 1, 3, 1)$ difference family [40,59]. In the literature the base blocks are typically given in the form $\{0, \omega^i(\omega^{2t} - 1), \omega^i(\omega^{4t} - 1)\}$ rather than as $\{\omega^i, \omega^{i+2t}, \omega^{i+4t}\}$.

An alternative combinatorial way of constructing a CDF is proposed by Rosa [33]. Rosa's method also generates a $(6t + 3, 3, 1)$ CDF. In Ref. 53 a list of constructions for rate-$\frac{1}{2}$ codes is given. The details are not given here, but the parameters are the same as of Netto construction. In Ref. 12 [p. 28] Colbourn noticed that this construction is perhaps wrongly associated with Netto.

2.2.2. The Netto's Second Construction. This construction [40] can be used to create a cyclic difference family when the number of points v is a power of a prime, and $v \equiv 7 \pmod{12}$. Again let ω be a generator of the multiplicative group and let $v = 6t + 1, t \geq 1$, and Ψ^d the group of dth powers in GF(v), then the set $\{\omega^{2i}\Psi^{2t} \mid 1 \leq i \leq t\}$ defines base blocks of the so called Netto triple system. The Netto systems are very interesting because of the following property. Netto triple systems on Z_v, v power of a prime and $v \equiv 19 \pmod{24}$ are Pasch-free [32,45,46] (see also Ref. 12, p. 214, Lemma 13.7), and as shown in Ref. 56 achieve the upper bound on minimum distance. For example, consider the base blocks of the Netto triple system difference family on $Z_v(\omega = 3)$ where $B_1 = \{0, 14, 2\}$, $B_2 = \{0, 40, 18\}$, $B_3 = \{0, 16, 33\}$, $B_4 = \{0, 15, 39\}$, $B_5 = \{0, 6, 7\}$, $B_6 = \{0, 11, 20\}$, and $B_7 = \{0, 13, 8\}$. The resulting code is quasi-cyclic, has $d_{\min} = 6$, length $b = 301$ and $R \geq 0.857$.

2.2.3. Burratti Construction for $k = 4$ and $k = 5$. For $k = 4$, Burratti's method gives CDFs with v points, provided that v is a prime and $v = 1 \bmod 12$. The CDF is a set $\{\omega^{6i}B{:}\, 1 \leq i \leq t\}$, where base blocks have the form $B = \{0, 1, b, b^2\}$, where ω is a primitive element in GF(v). The numbers $b \in$ GF($12t + 1$) for some values of v are given in Ref. 10. Similarly, for $k = 5$, the CDF is given as $\{\omega^{10i}B{:}\, 1 \leq i \leq t\}$, where $B = \{0, 1, b, b^2, b^3\}, b \in$ GF($20t + 1$). Some Buratti designs are given in Ref. 10.

2.2.4. Finite Euclidean and Finite Projective Geometries. The existence and construction of short designs ($b < 10^5$) is an active area of research in combinatorics. The handbook edited by Colbourn and Dinitz [12] is an excellent reference. The Abel and Greg Table 2.3 in Ref. 12 (pp. 41–42) summarizes the known results in existence of short designs. However, very often the construction of these design is somewhat heuristic or works only for a given block size. In many cases such constructions give a very small set of design with parameters of practical interests. An important subclass of BIBDs are so-called infinite families (Ref. 12, p. 67). The examples of these BIBDs include projective geometries, affine geometries, unitals, Denniston designs and some geometric equivalents of 2-designs (see Refs. 2 and 12, VI.7.12). The known infinite families of BIBD are listed in Table 1 [12,13]. Since q is a power of a prime, they can be referred as to *finite Euclidean* and *finite projective geometries*.

As we explained above, resolvable BIBDs are those BIBDs that possess parallel classes of blocks. If the blocks are viewed as lines, the analogy of BIBD and finite geometry becomes apparent. It is important to notice that not all BIBDs can be derived from finite geometries, but this discussion is beyond the scope of the discussion here. The most interesting projective and affine geometries are those constructed using Gallois fields [GF(q)]. In a m-dimensional projective geometry PG(m, q) a point $\boldsymbol{a}$ is specified by a vector $a = (a_j)_{0 \leq j \leq m}$, where $a_j \in$ GF(q). The vectors $a = (a_j)_{0 \leq j \leq m}$, and $\lambda a = (\lambda a_j)_{0 \leq j \leq m}$, [$\lambda \in$ GF(q), $\lambda \neq 0$] are considered to be representatives of the same point. There are $q^{(m+1)} - 1$ nonzero tuples, and λ can take $q - 1$ nonzero values, so that there are $(q - 1)$ tuples representing the same point. Therefore, the total number of points is $(q^{m+1} - 1)/(q - 1)$. The line through two distinct points $a = (a_j)_{0 \leq j \leq m}$ and $b = (b_j)_{0 \leq j \leq m}$ is incident with the points from the set $\{\mu a + vb\}$, where μ and v are not both zero. There are $q^2 - 1$ choices for μ and v, and each point appears $q - 1$ times in line $\{\mu a + vb\}$, so that the number of points on a line is $(q^2 - 1)/(q - 1) = q + 1$. For example, consider the nonzero elements of GF(2^2) = $\{0, 1, \omega, \omega^2\}$ that represent the points of PG(2,4). The triples $a = (1, 0, 0), \omega a = (\omega, 0, 0)$, and $\omega^2 a = (\omega^2, 0, 0)$, as we ruled, all represent the same point. The line incident with a and $b = (0, 1, 0)$ is incident with all of the following distinct points: a, $b, a + b = \{1, 1, 0\}, a + \omega b = \{1, \omega, 0\}, a + \omega^2 b = (1, \omega^2, 0)$. In the expression $\{\mu a + vb\}$, there are $4^2 - 1 = 15$ combinations of μ and v that are not both zero, but each point has $4 - 1 = 3$ representations, resulting in $\frac{15}{3}$ points on each line.

A *hyperplane* is a subspace of dimension $m - 1$ in PG(m, q); that is, it consists of the points satisfying a homogenous linear equation $\sum_{0 \leq j \leq m} \lambda_j \cdot a_j = 0, \lambda_j \in$ GF(q). A

Table 1. Known Infinite Families of 2-($v, k, 1$) Designs

k	v	Parameter	Name
q	q^n	$n \geq 2$, -power of a prime	Affine geometries
$q + 1$	$(q^n - 1)/(q - 1)$	$n \geq 2$, -power of a prime	Projective geometries
$q + 1$	$q^3 + 1$	q-power of a prime	Unitals
2^m	$2^m(2^s + 1) - 2^s$	$2 \leq m < s$	Denniston designs

projective plane of order q is a hyperplane of dimension $m = 2$ in PG(m, q). It is not difficult to see that the projective plane of order q is a BIBD$[(q^3 - 1)/(q - 1), q + 1, 1]$. The Euclidean (or affine) geometry EG(q, m) is obtained by deleting the points of a one hyperplane in PG(m, q). For example, if we delete a hyperplane $\lambda_0 a_0 = 0$, that is, the hyperplane with $a_0 = 0$, then the remaining q^m points can be labeled by the m-tuples $a = (a_j)_{1 \le j \le m}$. Euclidean or affine geometry EG(q, m) is a BIBD$(q^2, q, 1)$ [37]. For more details, see Kou et al. [25].

2.2.5. Lattice Construction of 2-(v, k,1) Designs. In this section we address the problem of construction of BIBDs of large block sizes. As shown, the Buratti-type CDFs and the projective geometry approach offer a quite limited set of parameters and therefore a small class of codes. In this section we give a novel construction of 2-$(v, k,1)$ designs with arbitrary block size. The designs are lines connecting points of a rectangular integer lattice. The idea is to trade a code rate and number of blocks for the simplicity of construction and flexibility of choosing design parameters. The construction is based on integer lattices as explained below.

Consider a rectangular integer lattice $L = \{(x, y): 0 \le x \le k - 1, 0 \le y \le m - 1\}$, where m is a prime. Let $l: L \to V$ be an one-to-one mapping of the lattice L to the point set V. An example of such mapping is a simple linear mapping $l(x, y) = m \cdot x + y + 1$. The numbers $l(x, y)$ are referred to as *lattice point labels*. For example, Fig. 3 depicts the rectangular integer lattice with $m = 7$ and $k = 5$.

A *line with slope* s, $0 \le s \le m - 1$, *starting at* the point (x, a), contains the points $\{(x, a + sx \bmod m): 0 \le x \le k - 1\}$, $0 \le a \le m - 1$. For each slope, there are m classes of parallel lines corresponding to different values. In our example, the lines of slope 1 are the points $\{1, 9, 17, 25, 33\}$, $\{2, 10, 18, 26, 34\}$, $\{3, 11, 19, 27, 35\}$, and so on. We assume that the lattice labels are periodic in vertical (y) dimension. The slopes $2, 3, \ldots, m$ can be defined analogously.

A set $\boldsymbol{B}$ of all k-element sets of V obtained by taking labels of points along the lines with different slopes s, $0 \le s \le m - 1$ is a BIBD. Since m is a prime, for each lattice point (x, y) there is exactly one line with slope s that go through (x, y). For each pair of lattice points, there is exactly one line that passes through both points. Therefore, the set $\boldsymbol{B}$ of lines is a 2-design. The block size is k, number of blocks is $b = m^2$ and each point in the design occurs in exactly m blocks. We can also show [56] that the $(m, k = 3)$ lattice BIBD is Pasch-free. Consider a periodically extended lattice. The proof is based on the observation that it is not possible to draw the quadrilateral (no sides with infinite slope are allowed) in which each point occurs twice. Figure 4 shows one such quadrilateral. Without loss of generality, we can assume that the ting point of two lines is (0,0). The slopes of four lines in Fig. 4 are $s, p, p + a/2$, and $s - a/2$. The points (0,0), (0,a), (2,2s), and $(2, a + 2p)$ are all different, and each occupies two lines. As long as the remaining four points are concerned, they will be on two lines in one of these three cases: (1) $s = p + a/2$ and $s + a/2 = a + p$; (2) $s = s + a/2$ and $p + a/2 = p + a$; and (3) $s = p + a$ and $p + a/2 = s + a/2$ (all operations are modulo m). Case 1 implies $s - p = a/2$, which means that points (2,2s) and $(2, a + 2p)$ are identical, which is a contradiction. If both equalities in cases 2 and 3 were satisfied, then a would have to be 0, which would mean that two leftmost points are identical, again a contradiction.

By generalizing this result to $k > 3$, we can conjecture that the (m, k) lattice BIBDs contain no generalized Pasch configurations. The lattice construction can be extended to nonprime vertical dimension m, but the slopes s and m must be coprime.

Figure 5 shows the growth of the required code length with a lower bound on a minimum distance as a parameter.

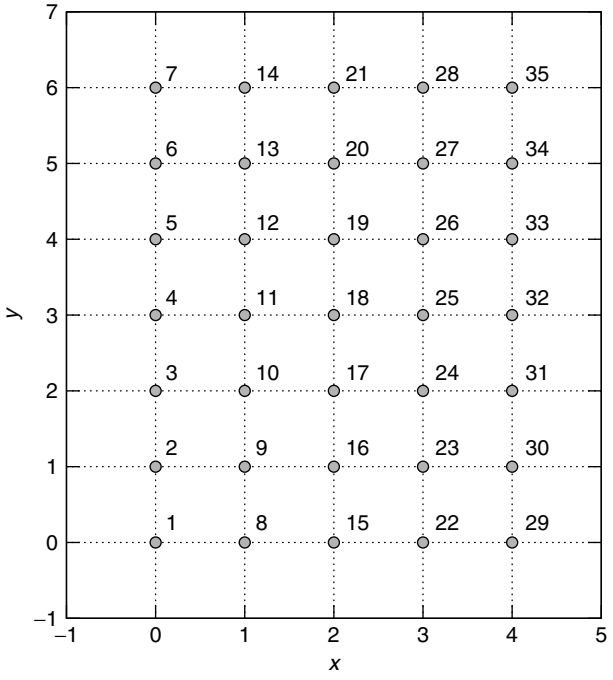


Figure 3. An example of the rectangular grid for $m = 7$ and $k = 5$.

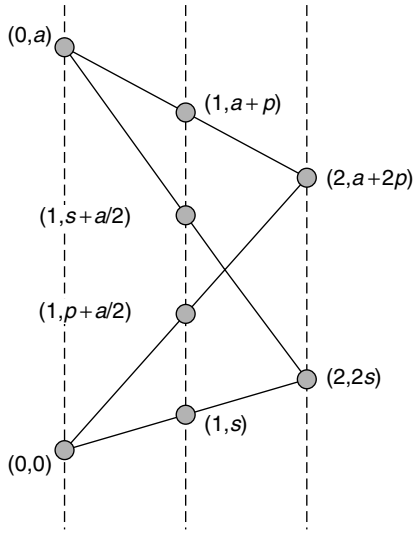


Figure 4. Quadrilateral in a lattice finite geometry.

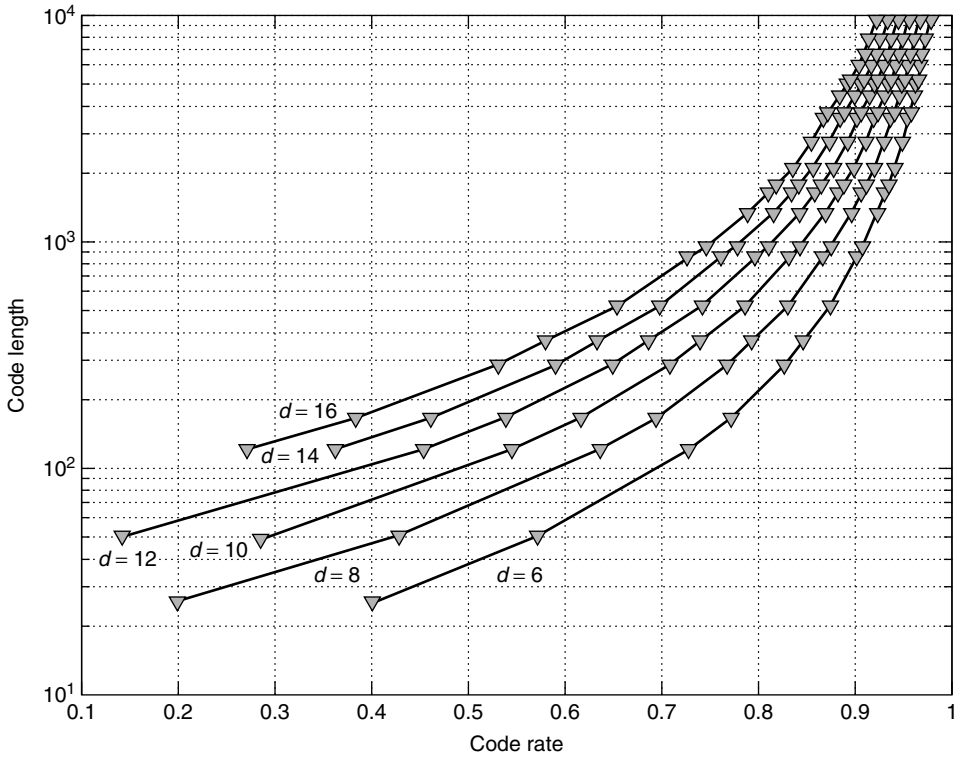


Figure 5. The rate–length curve for lattice designs with the minimum distance as a parameter.

3. APPLICATIONS OF LDPC CODES TO PERPENDICULAR RECORDING SYSTEMS

In this section we consider the performance of LDPC codes based on random and deterministic constructions in perpendicular recording systems. First we give a brief description of the perpendicular recording channel model.

3.1. Perpendicular Recording System

The write head, the magnetic medium, and the read head constitute the basic parts of a magnetic recording system. The binary data are written onto the magnetic medium by applying a positive or negative current to the inductive write head, creating a magnetic field that causes the magnetic particles on the media to align themselves in either of two directions. During playback, the alignment of the particles produces a magnetic field rising that produces a voltage in the read head when it passes over the media. This process can be approximated in the following simplified manner.

If the sequence of symbols $a_k \in C \subseteq \{\pm 1\}$ is written on the medium, then the corresponding write current can be expressed as

$$i(t) = \sum_k (a_k - a_{k-1}) u(t - KT_c)$$

where $u(t)$ is a unit pulse of duration T_c seconds. Assuming the read-back process is linear, the induced read voltage, $V(t)$, can be approximated as

$$V(t) = \sum_k a_k g(t - kT_c)$$

where $g(t)$ is the read-back voltage corresponding to an isolated positive going transition of the write current (transition response). For perpendicular recording systems under consideration $g(t)$ is modeled as

$$g(t) = \frac{2}{\sqrt{\pi}} \int_0^{St} e^{-x^2}\, dx = \operatorname{erf}(St)$$

where $S = 2\sqrt{1n2/D}$ and D is the normalized density of the recording system. There are various ways one can define D. In this work we define D as $D = T_{50}/T_c$, where T_{50} represents the width of impulse response at a half of its peak value.

In Figs. 6 and 7, the transition and dibit response $[g(t + T_c/2) - g(t - T_c/2)]$ of a perpendicular recording system are presented for various normalized densities.

In our system, the received noisy read-back signal is filtered by a lowpass filter, sampled at intervals of T_c seconds, and equalized to a partial response (PR) target. Therefore, the signal at the input to the detector at the kth time instant can be expressed as

$$z_k = \sum_j a_j f_{k-j} + w_k$$

where w_k is the colored noise sequence at the output of the equalizer and $\{f_n\}$ are the coefficients of the target channel response. In this work we investigated PR targets of the form $(1 + D)^n$ with special emphasis on $n = 2$, named PR2 case. The input signal-to-noise-ratio (SNR) is defined as $\mathrm{SNR} = 10 \times \log_{10}(V_p^2/\sigma^2)$, where V_p is the peak amplitude of the isolated transition response which is assumed to be unity.

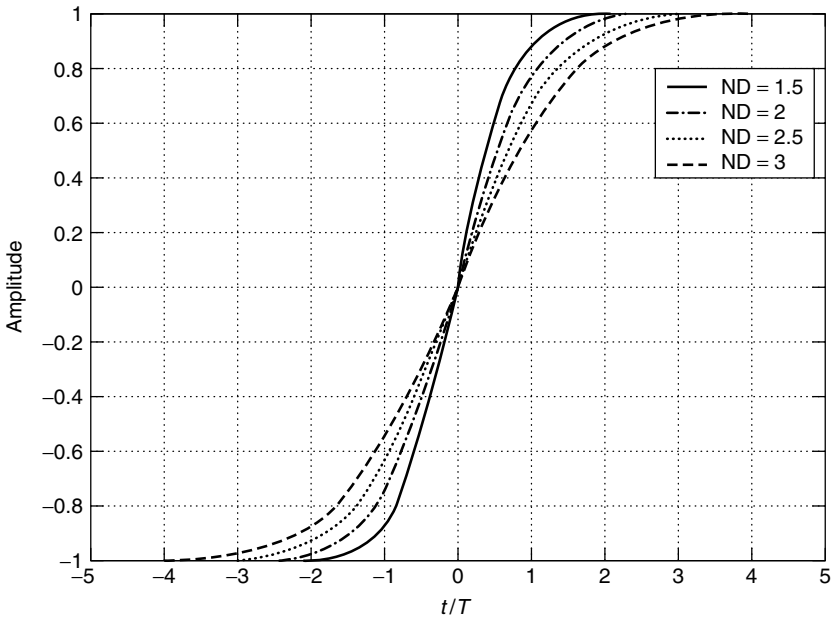


Figure 6. Transition response of a perpendicular recording channel.

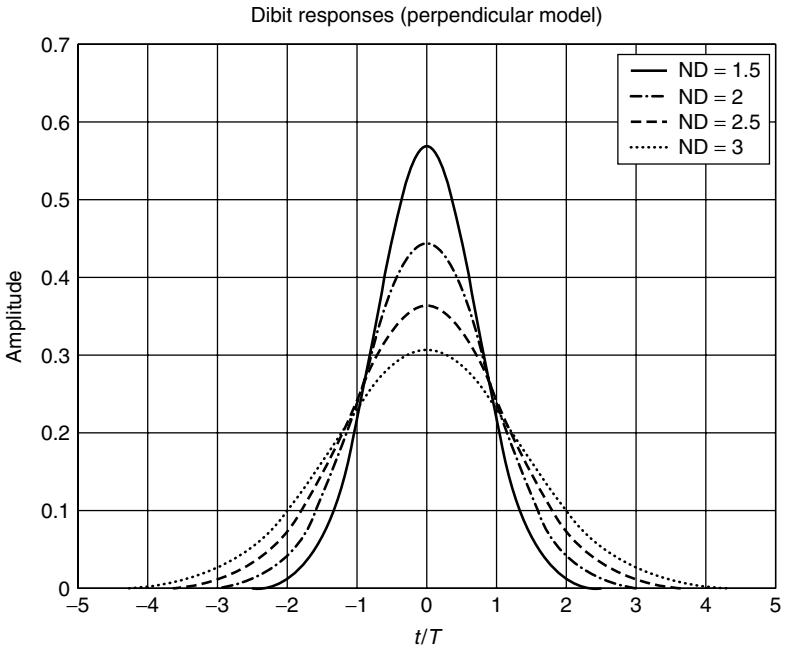


Figure 7. Dibit response of a perpendicular recording channel.

3.2. Simulation Results

We consider an iterative decoding scheme in which the BCJR algorithm [1] is used on the channel trellis, and the message-passing algorithm (MPA) is used for the decoding LDPC codes. The decoding is established by iterating between channel decoder and the outer LDPC decoder. LDPC decoder performs four inner iterations prior to supplying channel decoder with extrinsic information [17].

Throughout our simulations the normalized density for the uncoded system was 2.0 and the coded system channel density was obtained by adjusting with rate, namely, channel density $= D/R$, where R is the rate of the code.

In Fig. 8 we compare the performance of randomly constructed LDPC codes for different column weights, J, and rates. Random LDPC with $J = 3$ outperforms the other two and does not show any error flooring effect.

In Fig. 9 we compare Kirkman codes $(J = 3, v = 121, n = 2420)$ and $(J = 3, v = 163, n = 4401)$ and the random constructions with $J = 3$ and comparable block lengths and rates. It is clear that the LDPC codes based on Kirkman constructions show an earlier error floor than do their random counterparts. However, the lower error floor and less than 0.5 dB superior performance of random LDPC codes over Kirkman types come at a much larger implementation complexity penalty.

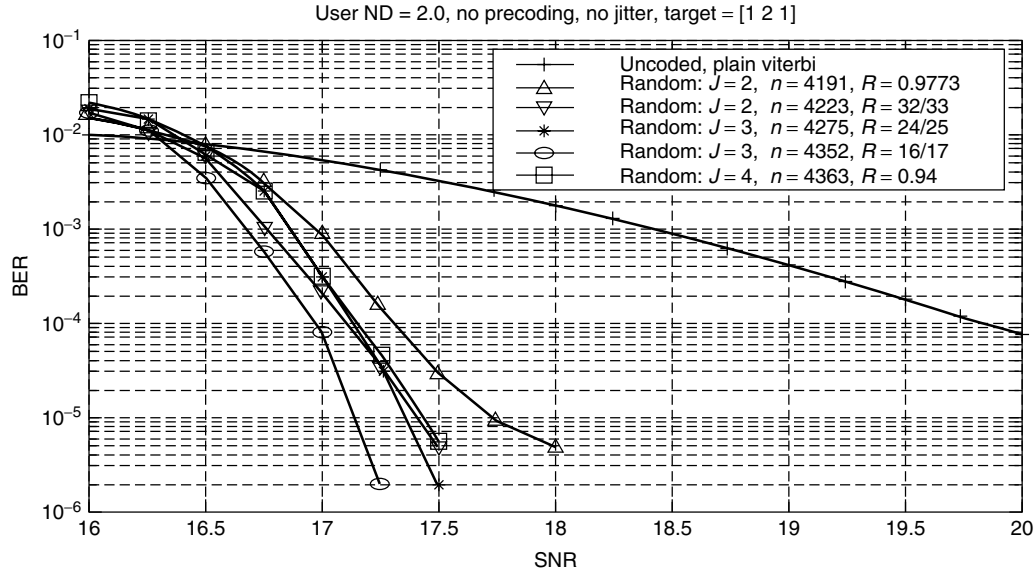


Figure 8. Random LDPC codes with different weights.

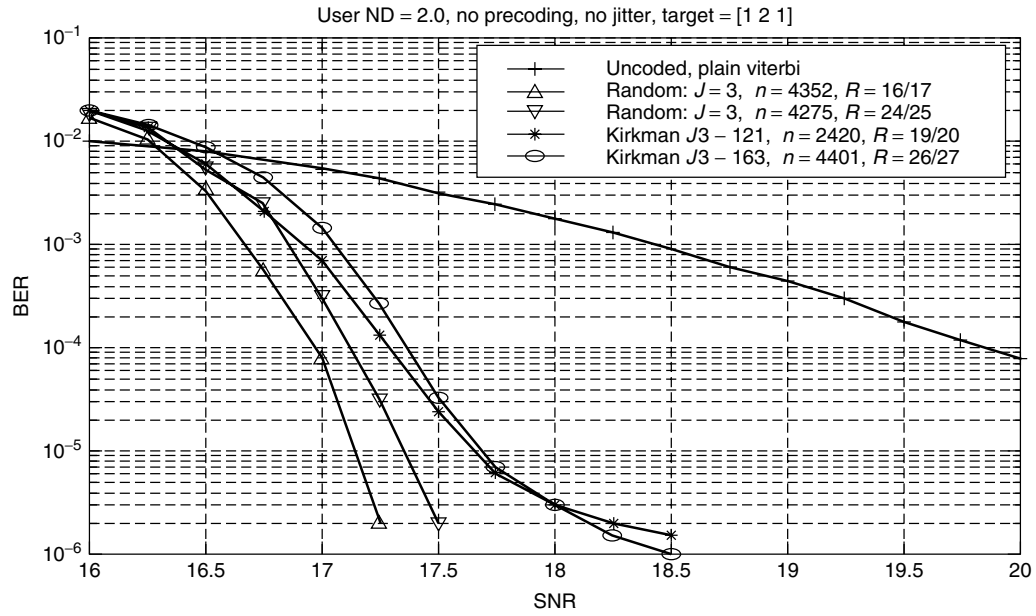


Figure 9. Random versus Kirkman LDPC codes.

Figure 10 makes a similar comparison between random LDPC codes of weights 3 and 4, and lattice designs with weights 4 and 5. As in Kirkman constructions, lattice designs have a higher error floor than do their random counterparts.

4. CONCLUSION

In this work we have presented various designs for LDPC code constructions. We have shown that LDPC codes based on BIBD designs can achieve very high rates and simple implementation compared to their random counterparts. We have analyzed the performance of these codes for a perpendicular recording channel. We have shown the tradeoff between BER performance, error floor, and complexity via extensive simulations.

Acknowledgment
The authors wish to thank our editor, John G. Proakis.

BIOGRAPHIES

Erozan M. Kurtas received his B.S. degree in electrical and electronics engineering in 1991 from Bilkent University, Ankara, Turkey, and an M.S. and Ph.D. degree in

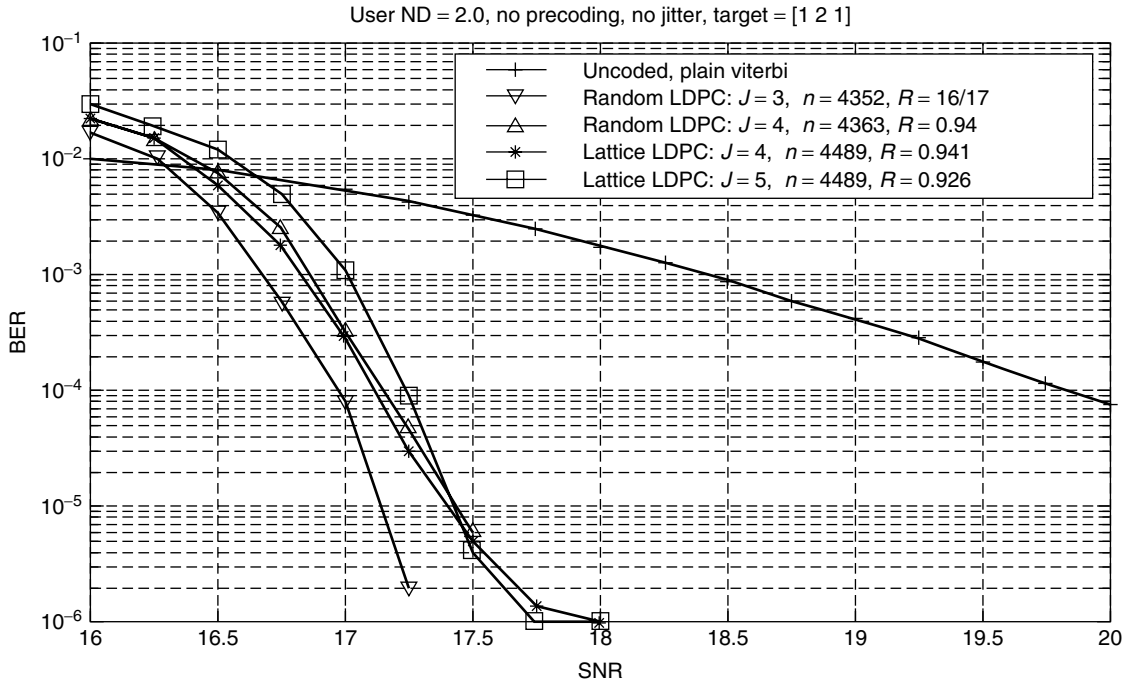


Figure 10. Random versus lattice LDPC codes.

electrical and computer engineering from the Northeastern University, Boston, Massachusetts, in 1994 and 1997, respectively. He joined Quantum Corporation (now Maxtor) in 1996 as a senior design engineer and worked on coding and signal processing for read-channel architectures employed in data storage devices. Since 1999, he has been with Seagate Technology where he is the director of channels research working on future data coding and recovery systems for storage applications. Dr. Kurtas has over 50 papers in various journals and conference proceedings in the general field of digital communications technology. His research interests span information theory, coding, detection and estimation, synchronization, and signal processing techniques. Dr. Kurtas is the recipient of the 2001 Seagate Technology Outstanding Technical Contribution and Achievement Award. Dr. Kurtas has five pending patent applications.

Alexander V. Kuznetsov received his Diploma with excellence in radio engineering and a Ph.D. degree in information theory from the Moscow Institute of Physics and Technology, Moscow, Russia, and the degree of doctor of engineering sciences from the Institute for Information Transmission Problems IPPI, Russian Academy of Sciences, in 1970, 1973, and 1988, respectively. In 1973, he joined the Russian Academy of Sciences as a scientific fellow, where for 25 years he worked on coding theory and its applications to data transmission and storage. On leave from IPPI, he held different research positions at the Royal Institute of Technology and Lund University in Sweden, Concordia University in Canada, Eindhoven Technical University in the Netherlands, Rand African University in RSA, Osaka University and Nara Advanced Institute of Sciences and Technology in Japan, Institute of Experimental Mathematics of Essen University in Germany, CRL of Hitachi, Ltd., in Tokyo, Japan. In 1998–1999, he was a member of technical staff of the Data Storage Institute (DSI) in Singapore, and worked on coding and signal processing for hard disk drives. Currently, he is a research staff member of the Seagate Technology in Pittsburgh, Philadelphia.

Dr. Vasic received his B.S., M.S., and Ph.D., all in electrical engineering from University of Nis, Serbia. From 1996 to 1997 he worked as a visiting scientist at the Rochester Institute of Technology, and Kodak Research, Rochester, New York, where he was involved in research in optical storage channels. From 1998 to 2000 he was with Lucent Technologies, Bell Laboratories. He was involved in research in read channel architectures and iterative decoding and low-density parity check codes. He was involved in development of codes and detectors for five generations of Lucent (now Agere) read channel chips. Presently, Dr. Vasic is a faculty member of the University of Arizona, Electrical and Computer Engineering Department. Dr. Vasic is an author of a more than fifteen journal articles, more than fifty conference papers, and one book chapter "Read channels for magnetic recording," in *CRC Handbook of Computer Engineering*. He is a member of the editorial board of the *IEEE Transactions on Magnetics*. He will serve as a technical program chair for IEEE Communication Theory Workshop in 2003, and as a coorganizer of the Center for Discrete Mathematics and Theoretical Computer Science Workshop on Optical/Magnetic Recording and Optical Transmission in 2003. His research interests include: coding theory, information theory, communication theory, digital communications, and recording.

BIBLIOGRAPHY

1. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (1974).
2. L. M. Batten, *Combinatorics of Finite Geometries*, London, Cambridge Univ. Press, 1997.
3. S. Benedetto, G. Montorsi, D. Divsalar, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *Telecommun. Data Acquis. Progr. Rep.* **42**: 1–26 (Aug. 1996).
4. S. Benedetto and G. Montorsi, Unveiling turbo codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* **42**: 409–428 (March 1996).
5. G. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. IEEE Int. Conf. Communications* (ICC'93), Geneva, Switzerland, May 1993, pp. 2.1064–2.1070.
6. T. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, Cambridge Univ. Press, 1986.
7. R. A. Beezer, The girth of a design, *J. Combinat. Math. Combinat. Comput.* (in press) (also online, *http://buzzard.ups.edu/pubs.html).*
8. V. K. Bhargava and J. M. Stein, (v, k, λ) configurations and self-dual codes, *Inform. Control* **28**: 352–355 (1975).
9. R. A. Brualdi, *Introductory Combinatorics*, Prentice-Hall, Upper Saddle River, NJ, 1999.
10. M. Buratti, Construction of (q, k, l) difference families with q a prime power and $k = 4, 5$, *Discrete Math.* **138**: 169–175 (1995).
11. P. J. Cameron and J. H. van Lint, *Graphs, Codes and Designs*, London Math. Soc. Lecture Note Series 23, Cambridge Univ. Press, 1980.
12. C. J. Colbourn and J. H. Dinitz, eds., *The Handbook of Combinatorial Designs*, CRC Press, Boca Raton, FL, 1996.
13. C. Colbourn and A. Rosa, *Steiner Systems*, Oxford Univ. Press (Oxford Mathematical Monographs), London, 1999.
14. T. M. Duman and E. Kurtas, Performance of Turbo codes over magnetic recording channels, *Proc. MILCOM*, Nov. 1999.
15. T. M. Duman and E. Kurtas, Comprehensive performance investigation of Turbo codes over high density magnetic recording channels, *Proc. IEEE GLOBECOM*, Dec. 1999.
16. T. M. Duman and E. Kurtas, Performance bounds for high rate linear codes over partial response channels, *Proc. IEEE Int. Symp. Information Theory* (Sorrento, Italy), IEEE, June 2000, p. 258.
17. J. Fan, A. Friedmann, E. Kurtas, and S. W. McLaughlin, Low density parity check codes for partial response channels, Allerton Conf. Communications, Control and Computing, Urbana, IL, Oct. 1999.
18. B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.
19. R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.
20. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**(2): 439–446 (March 1996).
21. N. Hamada, On the p-rank of the incidence matrix of a balanced or partially balanced incompleted block design and its applications to error correcting codes, *Hiroshima Math. J.* **3**: 153–226 (1973).
22. C. Heegard, Turbo coding for magnetic recording, *Proc. IEEE Information Theory Workshop*, (San Diego, CA), IEEE, Feb. 1998, pp. 18–19.
23. T. P. Kirkman, Note on an unanswered prize question, *Cambridge Dublin Math. J.* **5**: 255–262 (1850).
24. Y. Kou, S. Lin, and M. Fossorier, Construction of low density parity check codes: A geometric approach, *Proc. 2nd Int. Symp. Turbo Codes*, Brest, France, Sept. 4–7, 2000.
25. Y. Kou, S. Lin, and M. P. C. Fossorier, Low density parity check codes based on finite geometries: A rediscovery and new results, *IEEE Trans. Inform. Theory* **47**(7): 2711–2736 (Nov. 2001).
26. A. V. Kuznetsov and B. S. Tsybakov, On unreliable storage designed with unreliable components, *Proc. 2nd Int. Symp. Information Theory*, 1971, Tsahkadsor, Armenia (Publishing House of the Hungarian Academy of Sciences), 1973, pp. 206–217.
27. A. V. Kuznetsov, On the storage of information in memory constructed from unreliable components, *Problems Inform. Trans.* **9**(3): 100–113 (1973).
28. J. Li, E. Kurtas, K. R. Narayanan, and C. N. Georghiades, On the performance of Turbo product codes over partial response channels, *IEEE Trans. Magn.* (July 2001).
29. J. Li, E. Kurtas, K. R. Narayanan, and C. N. Georghiades, Iterative decoding for Turbo product codes over Lorentzian channels with colored noise, *Proc. GLOBECOM*, San Antonio, TX, Nov. 2001.
30. J. Li, E. Kurtas, K. R. Narayanan, and C. N. Georghiades, On the performance of Turbo product codes and LDPC codes over partial response channels, *Proc. Int. Conf. Communications*, Helsinki, Finland, June 2001.
31. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Englewood Cliffs, NJ, Prentice-Hall, 1983.
32. A. C. Ling, C. J. Colbourn, M. J. Grannell, and T. S. Griggs, Construction techniques for anti-pasch Steiner triple systems, *J. Lond. Math. Soc.* **61**(3): 641–657 (June 2000).
33. A. C. H. Ling and C. J. Colbourn, Rosa triple systems, in J. W. P. Hirschfeld, S. S. Magliveras, M. J. de Resmini, eds., *Geometry, Combinatorial Designs and Related Structures*, Cambridge Univ. Press, 1997, pp. 149–159.
34. J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge Univ. Press, 1992.
35. D. J. C. MacKay, Good error-correcting codes based on very sparse matrices, *IEEE Trans. Inform. Theory* **45**: 399–431 (March 1999).
36. D. MacKay and M. Davey, Evaluation of Gallager codes for short block length and high rate applications (online), *http://www.cs.toronto.edu/~mackay/CodesRegular.html*.
37. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Oxford, UK, 1977.
38. R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, Turbo decoding as an instance of Pearl's "Belief propagation" algorithm, *IEEE J. Select. Areas Commun.* **16**: 140–152 (Feb. 1998).
39. R. J. McEliece, *Finite Fields for Computer Scientist and Engineers*, Kluwer, Boston, 1987.

40. E. Netto, *Zur theorie der triplesysteme, Math. Ann.* **42**: 143–152 (1893).

41. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA, 1988.

42. L. Ping and K. Y. Wu, Concatenated tree codes: A low-complexity, high performance approach, *IEEE Trans. Inform. Theory* (Dec. 1999).

43. D. K. Ray-Chaudhuri and R. M. Wilson, Solution of Kirkman's school-girl problem, *Proc. Symp. Pure Math.*, Amer. Mathematical Society, Providence, RI, 1971, pp. 187–203.

44. T. Richardson, A. Shokrollahi, and R. Urbanke, Design of provably good low-density parity check codes, *IEEE Trans. Inform. Theory* **47**: 619–637 (Feb. 2001).

45. R. M. Robinson, Triple systems with prescribed subsystems, *Notices Am. Math. Soc.* **18**: 637 (1971).

46. R. M. Robinson, The structure of certain triple systems, *Math. Comput.* **29**: 223–241 (1975).

47. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* 372–423, 623–656 (1948).

48. E. Spence and V. D. Tonchev, Extremal self-dual codes from symmetric designs, *Discrete Math.* **110**: 165–268 (1992).

49. D. R. Stinson, Frames for Kirkman triple systems, *Discrete Math.* **65**: 289–300 (1988).

50. R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **IT-27**: 533–547 (Sept. 1981).

51. R. M. Tanner, Minimum-distance bounds by graph analysis, *IEEE Trans. Inform. Theory* **47**(2): 808–821 (Feb. 2001).

52. M. G. Taylor, Reliable information storage in memories designed from unreliable components, *Bell Syst. Tech. J.* **47**(10): 2299–2337 (1968).

53. R. Townsend and E. J. Weldon, Self-orthogonal quasi-cyclic codes, *IEEE Trans. Inform. Theory* **IT-13**(2): 183–195 (1967).

54. B. Vasic, Low density parity check codes: Theory and practice, National Storage Industry Consortium (NSIC) quarterly meeting, Monterey, CA, June 25–28, 2000.

55. B. Vasic, Structured iteratively decodable codes based on Steiner systems and their application in magnetic recording, *Proc. GLOBECOM 2001*, San Antonio, TX, Nov. 2001.

56. B. Vasic, Combinatorial constructions of structured low-density parity check codes for iterative decoding, *IEEE Trans. Inform. Theory* (in press).

57. E. J. Weldon, Jr., Difference-set cyclic codes, *Bell Syst. Tech. J.* **45**: 1045–1055 (Sept. 1966).

58. N. Wiberg, H.-A. Loeliger, and R. Kötter, Codes and iterative decoding on general graphs, *Eur. Trans. Telecommun.* **6**: 513–525 (Sept./Oct. 1995).

59. R. M. Wilson, Cyclotomy and difference families in elementary Abelian groups, *J. Number Theory* **4**: 17–47 (1972).

60. V. V. Zyablov and M. S. Pinsker, Error correction properties and decoding complexity of the low-density parity check codes, 2nd Int. Symp. Information Theory, Tsahkadsor, Armenia, 1971.

61. V. V. Zyablov and M. S. Pinsker, Estimation of the error-correction complexity for Gallager low-density codes, *Problems Inform. Trans.* **11**: 18–28 (1975) [transl. from *Problemy Peredachi Informatsii* **11**(1): 23–26].

DIFFERENTIATED SERVICES

IKJUN YEOM
Korea Advanced Institute of Science and Technology
Seoul, South Korea

A. L. NARASIMHA REDDY
Texas A&M University
College Station, Texas

1. INTRODUCTION

Providing quality of service (QoS) has been an important issue in Internet engineering, as multimedia/real-time applications requiring certain levels of QoS such as delay/jitter bounds, certain amount of bandwidth and/or loss rates are being developed. The current Internet provides only *best-effort* service, which does not provide any QoS. The Internet Engineering Task Force (IETF) has proposed several service architectures to satisfy different QoS requirements to different applications.

The Differentiated Services (Diffserv) architecture has been proposed by IETF to provide service differentiation according to customers' service profile called service-level agreement (SLA) in a scalable manner [1–4]. To provide service differentiation to different customers (or flows), the network needs to identify the flows, maintain the requested service profile and conform the incoming traffic based on the SLAs.

A simple Diffserv domain is illustrated in Fig. 1. In a Diffserv network, most of the work for service differentiation is performed at the edge of the network while minimizing the amount of work inside the network core, in order to provide a scalable solution. The routers at the edge of the network, called boundary routers or edge routers, may monitor, shape, classify, and mark *differentiated services code point* (DSCP) value assigned to specific packet treatment to packets of flows (individual or aggregated) according to the subscribed service. The routers in the network core forward packets differently to provide the subscribed service. The core routers need to provide only several forwarding schemes, called *per-hop behaviors* (PHBs) to provide service differentiation. It is expected that with appropriate network engineering, *per-domain behaviors* (PDBs), and end-to-end services can be constructed based on simple PHBs.

1.1. Differentiated Services Architecture

In the Diffserv network, we can classify routers into two types, *edge or boundary* routers and *core* routers according to their location. A boundary router may act both as an *ingress router* and as an *egress router* depending on the direction of the traffic. Traffic enters a Diffserv network at an ingress router and exits at an egress router. Each type of router performs different functions to realize service differentiation. In this section, we describe each of these routers with its functions and describe how service differentiation is provided.

1.1.1. Boundary Router. A basic function of ingress routers is to mark DSCPs to packets based on SLAs so that

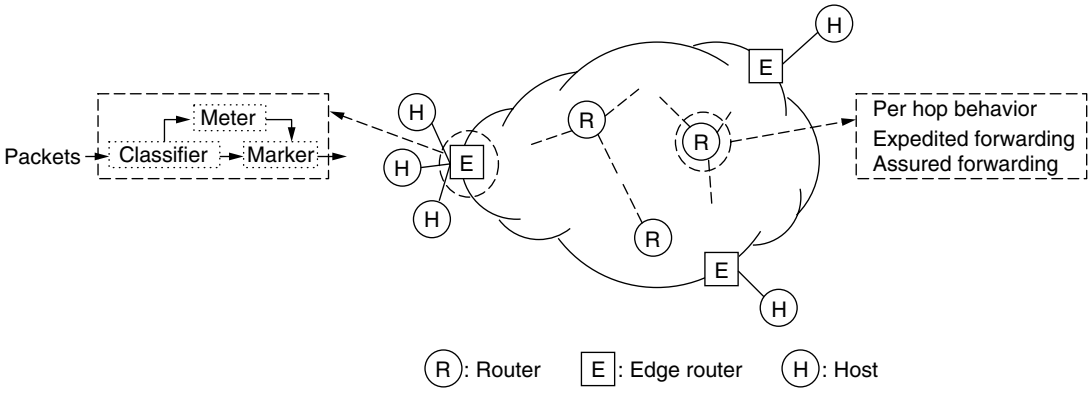


Figure 1. A Diffserv domain.

core routers can distinguish the packets and forward them differently. When a packet arrives at an ingress router, the router first identifies the flow to which the packet belongs and monitors the flow to conform to its contract. If it conforms to its contract, the packet is marked by the DSCP of the contracted service. Otherwise, the packet may be delayed, discarded, or unmarked to make injected traffic conform to its contract.

An egress router may monitor traffic forwarded to other domains and perform shaping or conditioning to follow a traffic conditioning agreement (TCA) with the other domains.

1.1.2. Core Router. In the Diffserv model, functions of core routers are minimized to achieve scalability. A core router provides requested PHB specified in the DS field of the arriving packets. Currently, *expedited forwarding* (EF) [3] and *assured forwarding* (AF) [4] PHBs have been standardized by IETF.

To provide end-to-end QoS, we may consider flows that transit over multiple network domains. While a hop-by-hop forwarding treatment in a Diffserv domain is defined by a PHB, Per-domain behavior (PDB) is used to define edge-to-edge behavior over a Diffserv domain [5]. A PDB defines metrics that will be observed by a set of packets with a particular DSCP while crossing a Diffserv domain. The set of packets subscribing to a particular PDB is classified and monitored at an ingress router. Conformant packets are marked with a DSCP for the PHB associated with the PDB. While crossing the Diffserv domain, core routers treat the packets based only on the DSCP. An egress router may measure and condition the set of packets belonging to a PDB to ensure that exiting packets follow the PDB.

1.1.3. Assured Forwarding PHB. Assured forwarding (AF) PHB has been proposed in [4,6]. In the AFPHB, the edge devices of the network monitor and mark incoming packets of either individual or aggregated flows. A packet of a flow is marked IN (in profile) if the temporal sending rate at the arrival time of the packet is within the contract profile of the flow. Otherwise, the packet is marked OUT (out of profile). Packets of a flow can be hence marked both IN and OUT. The temporal sending rate of a flow is measured using *time sliding window* (TSM) or a *token bucket* controller. IN packets are given preferential treatment at the time of congestion; thus, OUT packets are dropped first at the time of congestion.

Assured forwarding can be realized by employing RIO (RED with IN/OUT) drop policy [6] in the core routers. RIO drop policy is illustrated in Fig. 2. Each router maintains a virtual queue for IN packets and a physical queue for both IN and OUT packets. When the network is congested and the queue length exceeds minTh_OUT, the routers begin dropping OUT packets first. If the congestion persists even after dropping all incoming OUT packets and the queue length exceeds minTh_IN, IN packets are discarded. With this dropping policy, the RIO network gives preference to IN packets and provides different levels of service to users per their service contracts. Different marking policies [7–9] and correspondingly appropriate droppers have been proposed to improve the flexibility in providing service differentiation.

1.1.4. Expedited Forwarding PHB. Expedited forwarding (EF) PHB was proposed [3] as a premium service for the Diffserv network. The EFPHB can be used to guarantee low loss rate, low latency, low jitter, and assured

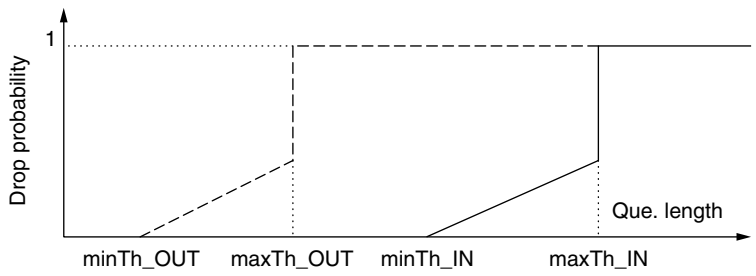


Figure 2. RED IN/OUT (RIO) drop policy.

throughput like a "virtual leased line." According to Jacobson et al. [3], the departure rate of the EF traffic measured over any time interval equal to or longer than a packet time should equal or exceed a configurable rate independent of the intensity of any other traffic.

However, there is evidence [10] that the original EFPHB configuration in Ref. 3 is too strict and hard to implement in practice. In a network where there are many connections frequently established, closed, merged and split, it is hard to maintain each node's arrival rate to be less than its departure rate at timescales required by the EFPHB. Alternatively, it has been proposed [10–12] that the EFPHB should be redefined as "a forwarding treatment for a particular Diffserv aggregate where the node offers to the aggregate a packet scale rate guarantee R with latency E, where R is a configurable rate and E is a tolerance that depends on the particular node characteristics." In a network providing packet scale rate guarantees, any EF packet arriving at a node at time t will leave the node no later than at time $t + Q/R + E$, where Q is the total backlogged EF traffic at time t. Here note that Q is zero in the original EF configuration since the arrival rate is less than the departure rate at any node at any given time.

Several types of queue scheduling schemes (e.g., a priority queue, a single queue with a weighted round-robin scheduler, and class-based queue [13]) may be used to implement the EFPHB.

It is important to understand the QoS delivered to individual applications by the extra support and functionality provided by a Diffserv network. In the next two sections, we present throughput analysis of AFPHB and delay analysis of EFPHBs.

2. TCP PERFORMANCE WITH AFPHB

Popular transport protocol TCP reacts to a packet loss by halving the congestion window and increases the window additively when packets are delivered successfully [14]. This AIMD congestion control policy makes the throughout of a TCP flow highly dependent on the dropping policy of the network. With AFPHB, service differentiation is realized by providing different drop precedences, and thus, TCP throughput with AFPHB is an interesting issue.

In Fig. 3, we present a simulation result using ns-2 [15] to show realized TCP throughput in a simple network with AFPHB. In this simulation, there are five TCP flows sharing one 4-Mbps bottleneck link. Each flow contracts bandwidth $\{0, 0.1, 0.5, 1, 2\}$ Mbps with network provider. From the figure, it is shown that flows with higher contract rates get higher throughput than flows with lower contract rates. However, it is also shown that the flows with 2 Mbps contract rate do not reach 2 Mbps while flows with 0- and 0.1-Mbps contract rates exceed their contract rates.

Similar results have been reported in the literature [9,16,17], and it has been also shown that it is difficult to guarantee absolute bandwidth with a simple marking and dropping scheme [18]. There is a clear need to understand the performance of a TCP flow with the AFPHB.

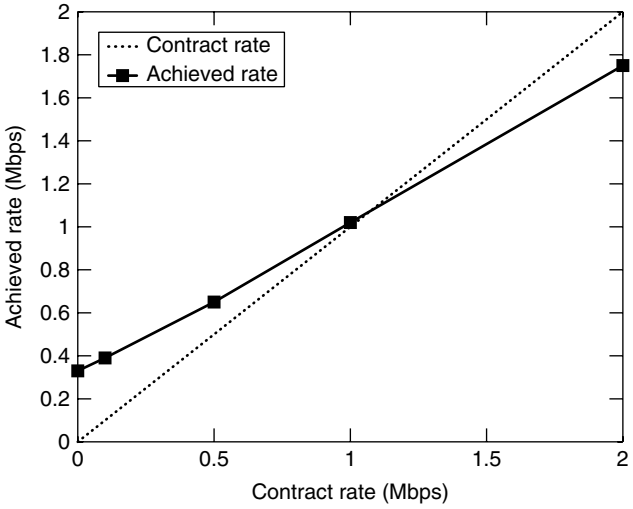


Figure 3. Realized TCP throughput with AFPHB.

In a steady state, a flow can experience different levels of congestion on the basis of its contract rate and the network dynamics. A flow that experiences no IN packet drops is said to observe an undersubscribed path. A flow that does not transmit any OUT packets either because every OUT packet is dropped or because the sending rate is less than the contract profile is said to observe an oversubscribed path. In a reasonably configured network, however, IN packets are expected to be protected, and only OUT packets may be discarded. In this section, we focus on the model for undersubscribed path.[1]

The steady-state TCP throughput, B is given by [19]

$$B = \begin{cases} \dfrac{3k}{4\mathrm{RTT}}\left(\sqrt{\dfrac{2}{p_{\mathrm{out}}}} + R\right) & \text{if } R \geq \dfrac{W}{2} \\ \dfrac{k}{2\mathrm{RTT}}\left(\sqrt{R^2 + \dfrac{6}{p_{\mathrm{out}}}} + R\right) & \text{otherwise} \end{cases} \tag{1}$$

where k is the packet size, p_{out} is the drop rate of OUT packets R is the reservation window defined as (contract rate/packet size $\times$ RTT), and W is the maximum congestion window size in steady state. From Eq. (1), B is proportional to the contract rate and inversely proportional to drop rate.

To illustrate the model described above, we present Fig. 4, where results from the model are compared with simulations. In the simulation, there are 50 TCP flows sharing a 30-Mbps link. Contract rate of each flow is randomly selected from 0 to 1 Mbps. In the figure, squares indicate simulation results, and stars indicate estimated throughput from the model. It is observed that the model can estimate TCP throughput accurately.

To discuss the interaction between contract rate and realized throughput more in depth, we define the

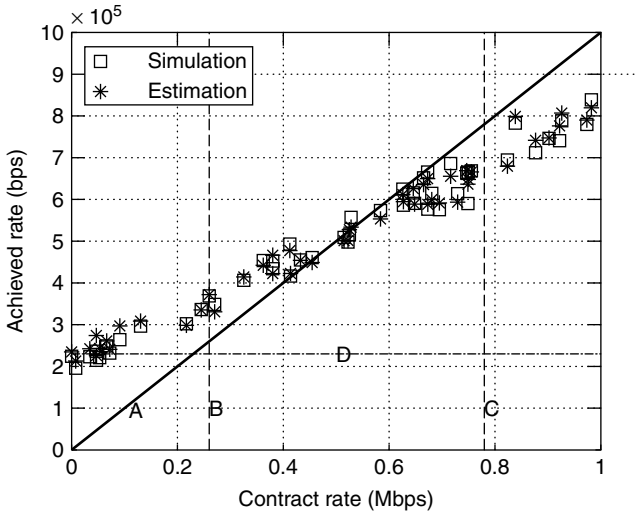
[1] To develop a model for a general situation where both IN and OUT packets are dropped, a model for an oversubscribed path is also required. For the general model, please refer to Ref. 19.

excess bandwidth B_e as the difference between realized throughput and its contract rate. It is given by

$$B_e = \begin{cases} \dfrac{k}{4\text{RTT}}\left(3\sqrt{\dfrac{2}{p_{\text{out}}}} - R\right) & \text{if } R \geq \dfrac{W}{2} \\ \dfrac{k}{2\text{RTT}}\left(\sqrt{R^2 + \dfrac{6}{p_{\text{out}}}} - R\right) & \text{otherwise} \end{cases} \tag{2}$$

If B_e of a flow is positive, this means that the flow obtains more than its contract rate. Otherwise, it does not reach its contract rate. From (2) and Fig. 4, we can observe that

- When a flow reserves relatively higher bandwidth ($R \geq \sqrt{2/p_{\text{out}}}$), B_e is decreased as the reservation rate is increased. Moreover, if R is greater than $3\sqrt{2/p_{\text{out}}}$ (see line C in Fig. 4), the flow cannot reach its reservation rate.
- When a flow reserves relatively lower bandwidth ($R < \sqrt{2/p_{\text{out}}}$; see line B in Fig. 4), it always realizes at least its reservation rate. As it reserves less bandwidth, it obtains more excess bandwidth. TCP's multiplicative decrease of sending rate after observing a packet drop results in a higher loss of bandwidth for flows with higher reservations. This explains the observed behavior.
- Equation (2) also shows that as the probability of OUT packet drop decreases, the flows with smaller reservation benefit more than do the flows with larger reservations. This points to the difficulty in providing service differentiation between flows of different reservations when there is plenty of excess bandwidth in the network.
- The realized bandwidth is observed to be inversely related to the RTT of the flow.
- For best-effort flows, $R = 0$. Hence, $B_e (= k\sqrt{6/p_{\text{out}}}/2\text{RTT}$; see line D in Fig. 4) gives the bandwidth likely to be realized by flows with no reservation.
- Comparing the above mentioned best-effort bandwidth when $R \geq \sqrt{2/p_{\text{out}}}$, we realize that the reservation rates larger than 3.5 times the best-effort bandwidth cannot be met.
- Equation (2) clearly shows that excess bandwidth cannot be equally shared by flows with different reservations without any enhancements to basic RIO scheme or to TCP's congestion avoidance mechanism.



- A: B = Contract rate
- B: $R = \sqrt{\dfrac{2}{p_{\text{out}}}}$
- C: $R = 3\sqrt{\dfrac{2}{p_{\text{out}}}}$
- D: $B = \dfrac{k}{2\text{RTT}}\sqrt{\dfrac{6}{p_{\text{out}}}}$

Figure 4. Observations from the model.

When several TCP flows are aggregated, the impact of an individual TCP sawtooth behavior is reduced, and the aggregated sending rate is stabilized. If the marker neither maintains per-flow state nor employs other specific methods for distinguishing individual flows, an arriving packet is marked IN with the probability, $p_m (= \textit{contract\_rate}/\textit{aggregated\_sending\_rate})$. In the steady state, p_m is approximately equal for all the individual flows. A flow sending more packets then gets more IN packets, and consequently, the contract rates consumed by individual flows is roughly proportional to their sending rates. We call this marking behavior *proportional marking*.

With the assumptions that all packets of aggregated flows are of the same size, k, a receiver does not employ delayed ACK, and the network is not oversubscribed, the throughput of the ith flow and the aggregated throughput B_A of n flows are given in [19] by

$$B_i = \frac{m_i}{\sum_{j=1}^{n} m_j} \cdot \frac{3r_A}{4} + \frac{3k}{4} m_i \tag{3}$$

$$B_A = \sum_{i=1}^{n} B_i = \frac{3k}{4} \sum_{i=1}^{n} m_i \tag{4}$$

where r_A is the contract rate for the aggregated flows, and m_i is $(1/\text{RTT}_i)\sqrt{(2/p_i)}$.

Equation (3) relates the realized throughput of an individual flow to the aggregated contract rate r_A and the network conditions (RTT_i and p_i) observed by various flows within the aggregation. From (4), B_e (the excess bandwidth) of aggregated flows is calculated as follows

$$B_e = \tfrac{3}{4} r_A + B_s - r_A = B_s - \tfrac{1}{4} r_A \tag{5}$$

where $B_s = \frac{3k}{4} \sum_{i=1}^{n} m_i$, and it is approximately the throughput which the aggregated flows can achieve with zero contract rate ($r_A = 0$). According to the analysis above, the following observations can be made:

- The total throughput realized by an aggregation is impacted by the contract rate. The larger the contract

rate, the smaller the excess bandwidth claimed by the aggregation.

- When the contract rate is larger than 4 times B_s, the realized throughput is smaller than the contract rate.
- The realized throughput of a flow is impacted by the other flows in the aggregation (as a result of the impact on p_m) when proportional marking is employed.
- The total realized throughput of an aggregation is impacted by the number of flows in the aggregation.

There are two possible approaches for enhancing better bandwidth differentiation with AFPHB. The first approach tries to enhance the dropping policy of the core routers [20] while the second approach tries to enhance the marking policies of the edge routers [7–9,21]. Below, we briefly outline one approach of enhancing the edge markers.

2.1. Adaptive Marking for Aggregated TCP Flows

The Diffserv architecture allows aggregated sources as well as individual sources [2]. When several individual sources are aggregated, output traffic of the aggregation is not like traffic of one big single source in the following respects: (1) each source within the aggregation responds to congestion individually, and (2) the aggregation has multiple destinations, and each source within the aggregation experiences different delays and congestion. Therefore, when we deal with aggregated sources in the Diffserv network, we need to consider not only the total throughput achieved by the aggregated sources but also the throughput achieved by individual flows.

Given individual target rates for each flow, how does one allocate a fixed aggregate contract rate among individual flows within the aggregation under dynamic network conditions? The most desirable situation is to guarantee individual target rates for all the flows. However, there exist situations in which some targets cannot be met: (1) when there is a severe congestion along the path and the current available bandwidth is less than the target and (2) when the contract rate is not enough to achieve the target. If we try to achieve the target by increasing the marking rate of an individual flow, the congestion along that flow's path may become more severe and result in wasting contracted resources of the aggregation. This is undesirable for both customers and service providers.

To solve this problem, an adaptive marking scheme has been proposed [21]. The proposed scheme achieves at least one of the following for all the flows in an aggregation:

1. Individual target rate when it is reachable.
2. Maximized throughput without IN packet loss when the current available bandwidth is less than the individual target rate.
3. Throughput achieved with a fair marking rate M/n, where M and n are the total marking rate and the number of flows within the aggregation, respectively.

These three goals correspond to (1) meeting individual flow's BW needs, (2) maximization of utility of the aggregated contract rate, and (3) fairness among the flows within the aggregation.

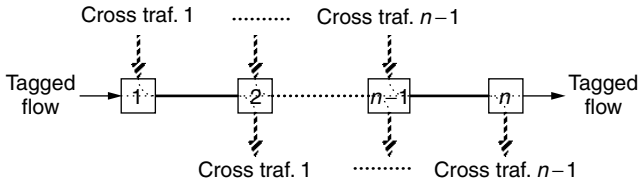


Figure 5. Multihop topology.

To observe how the marking rate is adjusted and an individual flow achieves its target rate, we conducted a set of simulations. We consider a multihop path as shown in Fig. 5. There are n routers, and cross traffic is injected to this network at the ith router and exits at the $(i+1)$th router.

In the simulation, we set the link capacity at 3 Mbps and use 10 TCP flows for cross traffic. The contract rate for each TCP flow is randomly selected from 0 to 1 Mbps, and the total contract of cross-traffic is 2.7 Mbps, so that the subscription level is 90%. The number of routers (n) is 5. For the tagged flow, we use a single TCP flow.

First, to observe path characteristics, we use a static contract rate for the tagged flow. We vary the contract rate from 0 to 0.8 Mbps. In Fig. 6, the solid line shows the achieved rate with static marking rate, and the dashed line indicates that the achieved rate is equal to 75% of the marking rate. The achieved rate increases as the marking rate increases up to 0.5 Mbps. However, the achieved rate does not increase beyond the 0.55-Mbps marking rate. In this example, the maximum achievable rate is about 0.42 Mbps.

Now the tagged flow is an individual flow within an aggregation with aggregated contract rate. The marker for the aggregation employs the adaptive marking. We vary the target rate for the tagged flow from 0.1 to 0.5 Mbps. Figure 6 shows the results. In each figure, dots indicate instantaneous marking and achieved rate, and a square shows the average. In this path, a flow gets 0.15 Mbps with zero contract rate. When the target rate is 0.1 Mbps (Fig. 6a), the marking rate stays around zero. When the target rate is achievable (<0.42 Mbps), the adaptive marking scheme finds the minimum marking rate to realize the target rate (Fig. 6a,b). In Fig. 6c, the marking rate stays below 0.55 Mbps to avoid wasting resources when the target is unachievable.

We present one other experiment to show the utility of the adaptive marker in reducing bandwidth differences due to RTTs as described by the model earlier. We use a topology in which 40 TCP flows are aggregated and compete in a 25-Mbps bottleneck link. The aggregated contract rate is 10 Mbps. The RTT (excluding queueing delay) of each flow is randomly selected from within 50–150 ms. Figure 7 shows the result. The adaptive marker effectively removes RTT bias of TCP flows and realizes QoS goals of individual flows within the aggregation.

The adaptive marker addresses the important problem of establishing a relationship between per-session

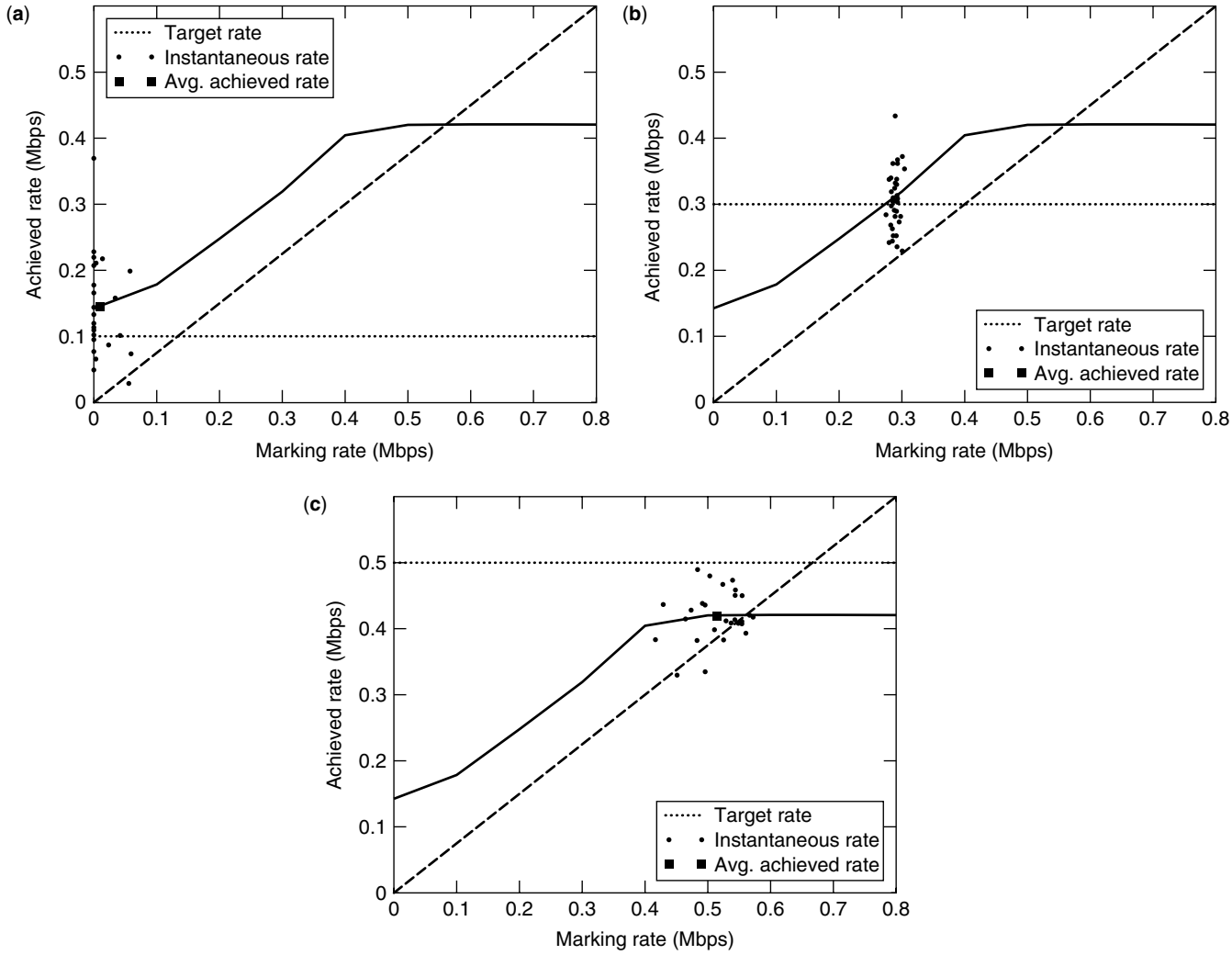


Figure 6. Achieved rates with the adaptive marking, with targets of (**a**) 0.1 Mbps, (**b**) 0.3 Mbps, and (**c**) 0.5 Mbps

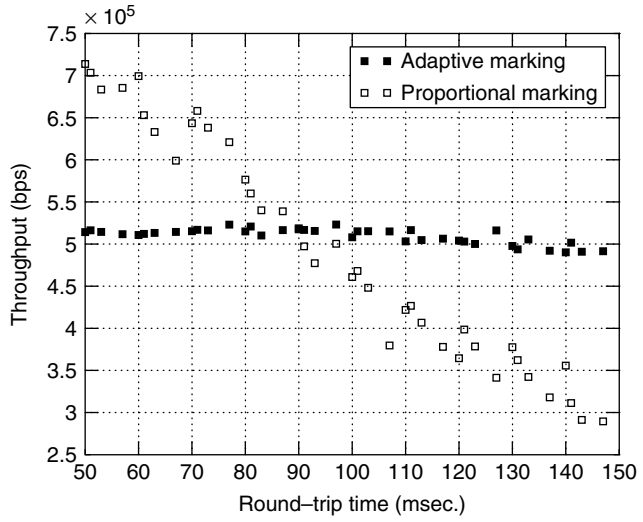


Figure 7. Throughput of flows with different round-trip times

behavior, aggregate packet marking, and packet differentiation within a differentiated services network. The adaptive marking algorithm is based on the TCP performance model within Diffserv networks. The adaptive marker enables reaching specific QoS goals of individual flows while efficiently managing the aggregate resources. Moreover, it shows that with appropriate edge devices, it is possible for applications to realize bandwidth guarantees utilizing the AFPHB.

3. DELAY ANALYSIS OF THE EFPHB

The EFPHB is proposed to provide low delay, jitter, and loss rate. Low loss rate is easily achievable with strict admission control and traffic conditioning. Regarding the EFPHB, providing low delay and jitter is an important issue and has been widely studied. In this section, we introduce studies on delay analysis of the EFPHB and present some experimental work for practical performance evaluation of the EFPHB.

3.1. Virtual Wire Service

Virtual wire (VW) service aims to provide a very low end-to-end delay jitter for flows subscribing to the EFPHB [22]. More precisely, as the term, *virtual wire* suggests, VW service is intended to "mimic, from the point of view of the originating and terminating nodes, the behavior of a hard-wired circuit of some fixed capacity [22]."

To provide VW service, the following two conditions are required: (1) each node in the domain must implement the EFPHB, and (2) conditioning the aggregate so that it's arrival rate at any node is always less than that node's departure rate. If the arrival rate is less than the virtual wire's configured rate, packets are delivered with almost no distortion in the interpacket timing. Otherwise, packets are unconditionally discarded not to disturb other traffic.

In [22], a *jitter window* is defined as the maximum time interval between two consecutive packets belonging to a flow so that the destination cannot detect delay jitter. Consider a constant-bit rate (CBR) flow with rate R. The packets of the flow arrive at an egress router through a link of bandwidth $B(= nR)$ and leave to their destination through a link of bandwidth R as shown in Fig. 8.

Let's consider two consecutive packets: P_0 and P_1. To transmit P_1 immediately after transmitting P_0 and to hide jitter, the last bit of P_1 should arrive at the node before T_1. T_1 is the time when the last bit of P_0 leaves the node and is calculated by

$$T_1 = T_0 + \frac{S}{R} \tag{6}$$

where S is the packet size. Then, the jitter window, Δ, is given by

$$\Delta = \frac{S}{R} - \frac{S}{B} = \frac{S}{R} \times \left(1 - \frac{1}{n}\right) \tag{7}$$

Generally, a packet experiences propagation, transmission, and queueing delay while traveling a network path. As long as the path is identical, propagation and transmission delay are the same. Therefore, as long as the sum of queueing delays that a packet observes in an EF domain is less than Δ, the destination does not observe jitter.

There are three possible sources of queueing delay of a VW packet: (1) non-EF packets ahead of the packet in a queue, (2) the previous VW packet of the same flow, and (3) VW packets of other flows. In a properly configured EF domain, the arrival rate of the EF traffic should be less than the departure rate at any node. Then, the queueing delay caused by the other EF traffic (cases 2 and 3) is zero.

In an EF domain, EF and non-EF packets are separately queued, and EF packets are serviced at a higher priority. Therefore, case 1 occurs only when an EF packet arrives at a node that is serving a non-EF packet. The worst-case delay occurs when the EF packet arrives immediately after the node begins to send a non-EF packet, and it is S/B, where S is the packet size. From (7), n should be at least 2 in order to satisfy the jitter window, and which means that the bandwidth assigned to EF traffic should be configured to be less than half of the link bandwidth.

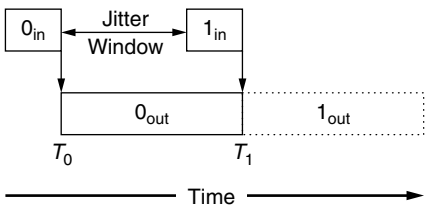


Figure 8. Jitter window.

3.2. Packet-Scale Rate Guarantee

Packet-scale rate guarantee service has been proposed to analyze and characterize the delay of the EFPHB more precisely [10–12]. A node is said to provide packet-scale rate guarantee R with latency E if, all $j \geq 0$, the jth departure time, $d(j)$, of the EF traffic is less than or equal to $F(j) + E$, where $F(j)$ is defined iteratively by

$$F(0) = 0, d(0) = 0 \tag{8a}$$

For all $j > 0$:

$$F(j) = \max[a(j), \min\{d(j-1), F(j-1)\}] + \frac{L(j)}{R} \tag{8b}$$

where $a(j)$ is the arrival time of the jth packet of length $L(j)$.

If a node provides packet-scale rate guarantee, it has been shown [10] that any EF packet arriving at a node at time t will leave that node no later than at time $t + Q/R + E$, where Q is the total backlogged EF traffic at time t. This property infers that a single-hop worst delay is bounded by $B/R + E_p$ when all input traffic is constrained by a leaky-bucket regulator with parameters (R, B), where R is the configured rate, and B is the bucket size. Note that E_p is the error term for processing individual EF packets.

If every node in a Diffserv domain regulates its EF input traffic with the maximum burst size B, then the worst case delay of each packet crossing h hops is just h times the single hop delay. However, the Diffserv architecture performs traffic conditioning only at the ingress and not in the interior. As a result, it is possible for bursty traffic larger than B to arrive at a node in the network even if the EF traffic has been regulated initially at the ingress. This problem may be solved by appropriately distributing traffic and limiting the number of hops traversed by each flow. Therefore, we need to consider topology constraints as well as ingress traffic conditioning when designing an EF service. Topology-independent utilization bounds have been obtained for providing bounded delay service that take network diameter into account [23].

3.3. QBone and Practical Performance Evaluation

QBone (Quality of Service Backbone) has been constructed by Internet2 QoS Working Group to provide an inter-domain testbed for Diffserv [24]. Currently, the EFPHB has been implemented, and the QBone Premium Service (QPS) [25] using the EFPHB is available.

To use the QPS, customers should provide their traffic information, {`source, dest, route, startTime, endTime, peakRate, MTU, jitter`}. This information

is required to configure EF nodes [3]. The parameter `route` specifies a DS-domain-to-DS-domain route. The parameters `startTime` and `endTime` denote the time duration of the traffic, `peakRate` is the peak rate of the traffic, `MTU` is the maximum transmission unit, and `jitter` is the worst-case jitter required. QPS traffic is regulated at an ingress node by a token bucket profiler with token rate of `peakRate` and bucket depth of `MTU`.

The QPS provides low loss rate, low queueing delay, and low jitter as long as the traffic arriving at the ingress node conforms to its token bucket profile. It is necessary to discard the nonconformant packets or to reshape them in order to achieve the desired QoS.

QBone enables us to evaluate and analyze the Diffserv performance more practically through large scale, interdomain experiments. An interesting experiment with video streaming applications has been conducted for observing end-to-end performance of the Diffserv network [26]. Here we present a summary of the experiment and results.

3.3.1. Video Streaming Applications in Diffserv Networks. With faster computer networks and improved compression schemes, video streaming applications are becoming popular. To improve quality of video playback on a given network, the applications deploy new technologies such as FEC (forward error correction) and layered coding to adjust to network dynamics. However, there is a limit to improving the quality without any QoS support from the network. To play back video clips seamlessly, a certain amount of bandwidth, bounded jitter, and low loss rate are required. Therefore, video streaming applications can be good target applications of a Diffserv network.

Experiments with video streaming applications in a Diffserv network with EFPHB have been performed [26]. A local network with several routers configured for EFPHB connected to the QBone network was used as a testbed. *Quality index* and frame loss rate were measured with various token rates and bucket sizes at ingress router policers. *Quality index* was measured by a variant of ITS (Institute of Telecommunication Sciences) video quality measurement (VQM) tool [27]. The VQM captures features of individual and sequence of frames from both the received and original streams, and measures the difference between them. The VQM was originally developed to measure quality of television and videoconferencing systems. The variant was modified from the VQM for measuring quality of video stream transmitted over IP networks.

From the experiments and measurements, interesting relationship between video quality and token bucket parameters of EF ingress policer have been observed: (1) the quality of video stream can be controlled by the ingress policer, (2) frame loss rate itself does not reflect the quality of the video stream, (3) the token rate should be configured to be larger than the video encoding rate for achieving acceptable quality, and (4) a small increase of bucket depth may result in substantial improvement of the video quality.

The first observation confirms that the Diffserv network can provide service differentiation to real-world applications.

4. NOVEL AND OTHER QoS

The Diffserv framework is still in development, and new services are constantly being proposed. In this section, we introduce some of the proposals receiving attention.

4.1. A Bulk Handling Per-Domain Behavior

Some applications need to transfer bulk data such as movie files or backup data over the network without any timing constraints. These transfers do not need any assurance as long as they are eventually completed. A bulk-handling (BH) per-domain behavior (PDB) has been proposed [28] for supporting such traffic. It is reasonable to exclude such traffic from best-effort traffic in order to prevent competition with other valuable traffic.

BHPDB traffic has the lowest priority in a Diffserv network. A BH packet is transferred only when there is no other packet. In the presence of other traffic, it may be discarded or delayed. To implement the BHPDB, only marking for PDB is enough. Policing or conditioning is not required since there is no configured rate, reserved resources or guarantees. BHPDB traffic is different from best-effort (BE) traffic in that there are at least some resources assigned to BE traffic.

4.2. An Assured Rate Per-Domain Behavior

An assured rate (AR) PDB has been proposed [29]. The ARPDB provides assured rate for one-to-one, one-to-few, or one-to-any traffic. Here *one-to-one traffic* means traffic entering a network at one ingress router and exiting at one egress router, and *one-to-few traffic* means traffic having more than one fixed egress routers. *One-to-any traffic* refers to traffic entering at one ingress router and exiting at multiple (any) egress routers. Assured rate can be implemented using the AFPHB.

The ARPDB is suitable for traffic requiring certain amount of bandwidth but no delay bounds. A possible example service with the ARPDB is VPN (virtual private network) services with one-to-few traffic. One-to-any service of the ARPDB can be used to deliver multicast traffic with a single source, but appropriate traffic measurement should be supported since packets in a multicast are duplicated in the interior of the network and the total amount of traffic at the egress routers may be larger than the amount of traffic at the ingress router. The ARPDB stipulates how packets should be marked and treated across all the routers within a Diffserv domain.

4.3. Relative Differentiated Services

Relative differentiated service has been proposed [30] to provide service differentiation without admission control. The fundamental idea of the relative differentiated services is based on the fact that absolute service guarantees are not achievable without admission control. In such cases, only relative service differentiation can be provided since network resources are limited.

A proportional delay differentiation model, which provides delays to each class on the basis of a proportionality constraint, has been proposed [30]. *Backlog-proportional rate* schedulers and *waiting-time priority* schedulers [31]

were proposed as candidate schedulers for such service. Both schedulers were shown to provide predictable and controllable delay differentiation independent of the variations of loads in each class. A proportional loss rate model has also been proposed and evaluated [32].

5. SUMMARY

We have presented the basic architecture of Diffserv networks and showed how service differentiation can be achieved in such networks. The AFPHB has been shown to enable us to realize service differentiation in TCP throughput with appropriate edge marking devices. The EFPHB has been proposed to provide low delay, jitter, and loss rate. While simple scheduling schemes such as priority queueing and class-based queueing are enough to implement the EFPHB in the network core, considerable work in traffic policing and shaping at the network edge is required. Initial experience on QBone points to the possibility that applications can realize the necessary QoS in Diffserv networks. Much research work is in progress in identifying and improving various aspects of a Diffserv network, particularly in traffic management, admission control and routing.

BIOGRAPHIES

A. L. Narasimha Reddy is currently an associate professor in the Department of Electrical Engineering at Texas A&M University, College Station. He received his Ph.D in computer Engineering from the University of Illinois at Urbana-Champaign in August 1990. He was a research staff member at IBM Almaden Research Center in San Jose from August 1990–August 1995.

Reddy's research interests are in multimedia, I/O systems, network QOS, and computer architecture. Currently, he is leading projects on building scalable multimedia storage servers and partial-state based network elements. His group is also exploring various issues related to network QOS. While at IBM, he coarchitected and designed a topology-independent routing chip operating at 100 MB/sec, designed a hierarchical storage management system, and participated in the design of video servers and disk arrays. Reddy is a member of ACM SIGARCH and is a senior member of IEEE Computer Society. He has received an NSF CAREER award in 1996. He received an Outstanding Professor Award at Texas A&M during 1997–98.

Ikjun Yeom received his B.S. degree in electronic engineering from Yonsei University, Seoul, Korea, in February 1995 and his M.S. and Ph.D. degrees in computer engineering from Texas A&M University, College Station, in August 1998 and May 2001, respectively. He worked at DACOM Company located in Seoul, Korea, between 1995 and 1996 and at Nortel Networks in 2000. After working as a research professor at Kyungpook National University in 2001, he has been an assistant professor in the Department of Computer Science at KAIST since January 2002. His research interests are in congestion control, network performance evaluation, and Internet QoS.

BIBLIOGRAPHY

1. K. Nichols, S. Blake, F. Baker, and D. Black, *Definition of the Differentiated Service Field (DS Field) in the IPv4 and IPv6 Headers*, RFC 2474, Dec. 1998.
2. S. Blake et al., *An Architecture for Differentiated Services*, RFC 2475, Dec. 1998.
3. V. Jacobson, K. Nichols, and K. Poduri, *An Expedited Forwarding PHB*, RFC 2598, June 1999.
4. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, *Assured Forwarding PHB Group*, RFC 2597, June 1999.
5. K. Nochols and B. Carpenter, *Definition of Differentiated Services Per Domain Behaviors and Rules for Their Specification*, RFC 3086, April 2001.
6. D. Clark and W. Fang, Explicit allocation of best-effort packet delivery service, *IEEE/ACM Trans. Network.* **6**(4): 362–373 (Aug. 1998).
7. J. Heinanen, T. Finland, and R. Guerin, *A Three Color Marker*, RFC 2697, Sept. 1999.
8. J. Heinanen, T. Finland, and R. Guerin, *A Two Rate Three Color Marker*, RFC 2698, Sept. 1999.
9. F. Azeem, A. Rao, and S. Kalyanaraman, TCP-friendly traffic marker for IP differentiated services, *Proc. IWQoS'2000*, Pittsburgh, PA, June 2000, pp. 35–48.
10. J. Bennett et al., Delay jitter bounds and packet scale rate guarantee for expedited forwarding, *Proc. Infocom*, 2001.
11. B. Davie et al., *An Expedited Forwarding PHB*, Work in Progress, April 2001.
12. A. Charny et al., *Supplemental Information for the New Definition of the EFPHB*, Work in Progress, June 2001.
13. S. Floyd and V. Jacobson, Like-sharing and resource management models for packet networks, *IEEE/ACM Trans. Network.* **3**(4): 365–386 (Aug. 1995).
14. V. Jacobson and M. Karels, Congestion avoidance and control, *Proc. SIGCOMM'88*, Stanford, CA, Aug. 1998, pp. 314–329.
15. S. McCanne and S. Floyd, *ns-LBL network simulator*; see: *http://www.nrg.ee.lbl.gov/ns/*.
16. I. Yeom and A. L. N. Reddy, Realizing throughput guarantees in a differentiated services network, *Proc. ICMCS'99*, Florence, Italy, June 1999, pp. 372–376.
17. S. Sahu et al., On achievable service differentiation with token bucket marking for TCP, *Proc. SIGMETRICS'2000*, Santa Clara, CA, June 2000, pp. 23–33.
18. I. Stoica and H. Zhang, LIRA: An approach for service differentiation in the Internet, *Proc. NOSSDAV'98*, Cambridge, UK, June 1998, pp. 345–359.
19. I. Yeom and A. L. N. Reddy, Modeling TCP behavior in a differentiated services network, *IEEE/ACM Trans. Network.* **9**(1): 31–46 (Feb. 2001).
20. S. Gopalakrishnan and A. L. N. Reddy, SACRIO: An active buffer management schemes for differentiated services networks, *Proc. NOSSDAV'01*, June 2001.
21. I. Yeom and A. L. N. Reddy, Adaptive marking for aggregated flows, *Proc. Globecom*, 2001.
22. V. Jacobson, K. Nochols, and K. Poduri, *The Virtual Wire Behavior Aggregate*, March 2000, Work in Progress.
23. S. Wang, D. Xuan, R. Bettati, and W. Zhao, Providing absolute differentiated services for real-time applications in static priority scheduling networks, *Proc. IEEE Infocom*, April 2001.

24. B. Teitelbaum, *QBone Architecture* (*v1.0*), Internet2 QoS Working Group Draft, Aug. 1999; see: *http://www.internet2.edu/qos/wg/qbArch/1.0/draft-i2-qbonearch-1.0.html*.

25. K. Nichols, V. Jacobson, and L. Zhang, *A Two-Bit Differentiated Services Architecture for the Internet*, Nov. 1997; see: *ftp://ftp.ee.lbl.gov/papers/dsarch.pdf*.

26. W. Ashmawi, R. Guerin, S. Wolf, and M. Pinson, On the impact of pricing and rate guarantees in Diff-Serv networks: A video streaming application perspective, *Proc. ACM Sigcomm'01*, Aug. 2001.

27. ITU-T Recommendation J. 143, *User Requirements for Objective Perceptual Video Quality Measurements in Digital Cable Television*, Recommendations of the ITU, Telecommunication Standardization Sector.

28. B. Carpenter and K. Nichols, *A Bulk Handling Per-Domain Behavior for Differentiated Services*, Internet Draft, Jan. 2001; see: *http://www.ietf.org/internet-drafts/draft-ietf-diffserv-pdb-bh-02.txt*.

29. N. Seddigh, B. Nandy, and J. Heinanen, *An Assured Rate Per-Domain Behavior for Differentiated Services*, Work in Progress, July 2001.

30. C. Dovrolis, D. Stiliadis, and P. Ramanathan, Proportional differentiated services: delay differentiation and packet scheduling, *Proc. SIGCOMM'99*, Aug. 1999, pp. 109–120.

31. L. Kleinrock, *Queueing Systems*, Vol. II, Wiley, 1976.

32. C. Dovrolis and P. Ramanathan, Proportional differentiated services, Part 2: Loss Rate Differentiation and Packet Dropping, *Proc. IWQoS'00*, Pittsburgh PA, June 2000.

DIGITAL AUDIOBROADCASTING

RAINER BAUER
Munich University of Technology (TUM)
Munich, Germany

1. INTRODUCTION

The history of broadcasting goes back to James Clerk Maxwell, who theoretically predicted electromagnetic radiation in 1861; and Heinrich Hertz, who experimentally verified their existence in 1887. The young Italian Guglielmo Marconi picked up the ideas of Hertz and began to rebuild and refine the Hertzian experiments and soon was able to carry out signaling over short distances in the garden of the family estate near Bologna. In the following years the equipment was continuously improved and transmission over long distances became possible. The first commercial application of radio transmission was point-to-point communication to ships, which was exclusively provided by Marconi's own company. Pushed by the promising possibilities of wireless communication the idea of broadcasting soon also came up. The first scheduled radio program was transmitted by a 100-W transmitter at a frequency around 900 kHz in Pittsburgh late in 1920.

The installation of more powerful transmitters and progress in receiver technology led to a rapidly increasing number of listeners. A milestone in broadcasting history was the invention of frequency modulation (FM) by Armstrong in the 1930s and the improvement in audio quality that came along with the new technology. Because FM needed higher transmission bandwidths, it was necessary to move to higher frequencies. While the first FM radio stations in the United States operated at frequencies around 40 MHz, today's FM radio spectrum is located in the internationally recognized VHF FM band from 88 to 108 MHz. Although the reach of the FM signals was below that of the former AM programs, due to the shorter wavelength of the FM band frequencies, broadcasters adopted the new technology after initial skepticism and the popularity of FM broadcasting grew rapidly after its introduction.

The next step toward increased perceived quality was the introduction of stereobroadcasting in the 1960s. Constant advances in audiobroadcasting technology resulted in today's high-quality audio reception with analog transmission technology. So, what is the reason for putting in a lot of time and effort into developing a new digital system?

In 1982 the compact disk (CD) was introduced and replaced the existing analog record technology within just a few years. The advantages of this new digital medium are constant high audio quality combined with robustness against mechanical impacts. The new technology also resulted in a new awareness of audio quality and the CD signal became a quality standard also for other audio services like broadcasting.

Using good analog FM equipment, the audio quality is fairly comparable to CD sound when the receiver is stationary. However, along with the development of broadcasting technology, the customs of radio listeners also changed. While at the beginning, almost all receivers were stationary, nowadays many people are listening to audiobroadcasting services in their cars, which creates a demand for high-quality radio reception in a mobile environment. This trend also reflects the demand for mobile communication services, which has experienced a dramatic boom.

Due to the character of the transmission channel, which leads to reflections from mountains, buildings, and cars, for example, in combination with a permanently changing environment caused by the moving receiver, a number of problems arise that cannot be handled in a satisfying way with existing analog systems. The problem here is termed multipath propagation and is illustrated in Fig. 1. The transmitted signal arrives at the receiver not only via the direct (line-of-sight) path but also as reflected and scattered components that correspond to different path delays and phase angles. This often results in severe interference and therefore signal distortion.

Another drawback is interference caused by transmission in neighboring frequency bands (adjacent-channel interference) or by transmission of different programs in the same frequency with insufficient spacial distance between the transmitters (cochannel interference).

The tremendous progresses in digital signal processing and microelectronics in the 1980s initiated a trend to supplement and even substitute existing analog systems with digital systems. Introduction of the compact disk was already mentioned above, but a change from analog

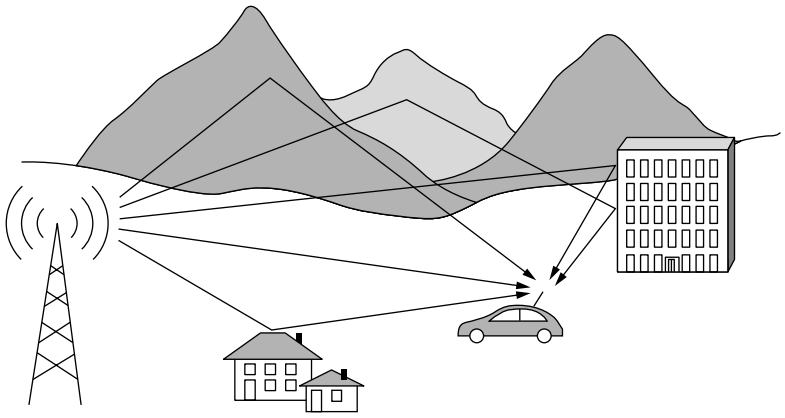


Figure 1. Scenario with multipath propagation.

to digital systems also took place in the field of personal communications. This was seen, for example, in the introduction of ISDN (Integrated Services Digital Network) in wireline communications and also in GSM (Global System for Mobile Communications) for personal mobile communications. The latter system caused a rapidly growing market for mobile communications systems and drove the development of new high-performance wireless communication techniques.

Several advantages of digital systems suggest the application of digital techniques also in broadcasting, which is still dominated by analog systems [1]. For instance, advanced digital receiver structures and transmission and detection methods are capable of eliminating the distortions caused by multipath propagation, which enables high-quality reception also in mobile environments.

To use the available frequency spectrum in an efficient way, the digitized audio signals can be compressed with powerful data reduction techniques to achieve a compression factor of up to 12 without perceptible degradation of audio quality compared to a CD signal.

Techniques such as channel coding to correct transmission errors are applicable only to digital signals. These methods allow an error free transmission even when errors occur on the transmission channel.

With advanced signal detection and channel coding techniques, a significantly lower signal to noise ratio is sufficient in producing a particular sound quality when compared to equivalent analog transmission systems. This results in lower transmission power and therefore reduced costs and also less "electromagnetic pollution," which has recently become a topic of increasing importance and relevance.

Another advantage of a digital broadcasting system is its flexibility. While conventional analog systems are restricted to audio transmission, a digital system can also provide other services besides audio transmission and is therefore open to trends and demands of the future. In combination with other digital services such as Internet or personal communications, a digital broadcasting technology further supports the convergence of information and communication techniques. Also there is a change of paradigm in the way we obtain information in everyday life. The growing importance of the Internet obliterates the boundaries between individual and mass communications. Broadcasting will have to face competition with other communication networks that have access to the consumer. To meet future demands, it is necessary to introduce digital broadcasting systems.

2. SYSTEM ASPECTS AND DIFFERENT APPROACHES

Besides technical aspects, the introduction of a digital audiobroadcasting system must also meet market issues. The transition from an analog to a digital system has to take place gradually. Users have to be convinced that digital audiobroadcasting has significant advantages compared to analog systems in order to bring themselves to buy new digital receivers instead of conventional analog ones. This can be obtained only by added value and new services. On the other hand, the situation of the broadcasters and network operators must be considered. Simultaneous transmission of digital and analog programs results in increasing costs that have to be covered. Standards have to be adopted to guarantee planning reliability to broadcasters, operators, and also manufacturers [2].

2.1. An Early Approach Starts Services Worldwide

An early approach to standardize a digital audiobroadcasting system was started in the 1980s by a European telecommunication consortium named Eureka. In 1992 the so-called Eureka-147 DAB system was recommended by the International Telecommunication Union (ITU) and became an ITU standard in 1994. The standard is aimed at terrestrial and satellite sound broadcasting to vehicular, portable and fixed receivers and is intended to replace the analog FM networks in the future. The European Telecommunications Standard Institute (ETSI) also adopted this standard in 1995 [3].

The terrestrial service uses a multicarrier modulation technique called coded orthogonal frequency-devision multiplex (COFDM), which is described later in this article along with other technical aspects of this system. With this multiplexing approach up to six high-quality audio programs (called an ensemble) are transmitted via closely spaced orthogonal carriers that allocate a total bandwidth of 1.536 MHz. The overall net data rate that can be transmitted in this ensemble is approximately 1.5 Mbps

(megabits per second) [4]. Although originally a sound broadcasting system was to be developed, the multiplex signal of Eureka-147 DAB is very flexible, allowing the transmission of various data services also. Because of the spreading in frequency by OFDM combined with channel coding and time interleaving, a reliable transmission can be guaranteed with this concept. A further advantage of the Eureka-147 system is the efficient use of radio spectrum by broadcasting in a single-frequency network (SFN). This means that one program is available at the same frequency across the whole coverage area of the network. In a conventional analog audiobroadcasting environment, each program is transmitted on a separate carrier (in FM broadcasting with a bandwidth of 300–400 kHz). If the network offers the same program outside the coverage area of a particular transmitter another frequency is usually used to avoid cochannel interference. On the other hand, adjacent channels within the coverage area of one transmitter are not used in order to reduce adjacent-channel interference. This leads to a complex network and requires a sophisticated frequency management, both of which can be avoided with a single-frequency network. The basic structures of a single-frequency network and a multiple frequency network are compared in Fig. 2. A further advantage of a single-frequency network is the simple installation of additional transmitters to fill gaps in the coverage.

To operate a system like Eureka-147 DAB dedicated frequency bands have to be reserved because it cannot coexist with an analog system in the same frequency band. In 1992 the World Administrative Radio Conference (WARC) reserved almost worldwide a 40-MHz wide portion (1452–1492 MHz) of the L band for this purpose. In several countries a part of the spectrum in the VHF band III (in Germany, 223–240 MHz) is also reserved for Eureka-147 DAB.

At present, the Eureka-147 DAB system is in operation or in pilot service in many countries worldwide [5].

2.2. A System for More Flexibility

The strategy to merge several programs into one broadband multiplex signal prior to transmission has several advantages with respect to the received signal quality. However, the ensemble format also shows drawbacks that aggravate the implementation of Eureka-147 DAB in certain countries. The broadcasting market in the United States for example is highly commercial with many independent private broadcasters. In this environment it would be almost impossible to settle on a joint transition scenario from analog to digital. Also, the reserved frequency spectrum in the L band is not available in the United States [6]. Therefore, a different system was considered by the U.S. broadcasting industry. The idea was that no additional frequency spectrum (which would cause additional costs, because spectrum is not assigned but auctioned off in the United States) should be necessary, and the broadcasters should remain independent from each other and be able to decide on their own when to switch from analog to digital. A concept that fulfills all these demands is in-band/on-channel (IBOC) digital audiobroadcasting. In this approach the digital signal is transmitted in the frequency portion to the left and the right of the analog spectrum, which is left free to avoid interference from nearby signals. The basic concept is shown in Fig. 3. The spectral position and power of the digital signal is designed to meet the requirements of the spectrum mask of the analog signal. The IBOC systems will be launched in an "hybrid IBOC" mode, where the digital signal is transmitted simultaneously to the analog one. When there is sufficient market penetration of the digital services, the analog programs can be switched off and the spectrum can be filled with a digital signal to obtain an all-digital IBOC system.

After the merger of USA Digital Radio (USADR) and Lucent Digital Radio (LDR) to iBiquity Digital Corp. there is currently a single developer of IBOC technology in the United States.

2.3. Digital Sound Broadcasting in Frequency Bands Below 30 MHz

The first radio programs in the 1920s were transmitted at frequencies around 900 kHz using analog amplitude modulation (AM). Although frequency modulation (FM), which was introduced in the late 1930s and operated at higher frequencies, allows a better audio quality, there are still many radio stations, particularly in the United States, that use the AM bands for audiobroadcasting. One advantage of these bands is the large coverage area that

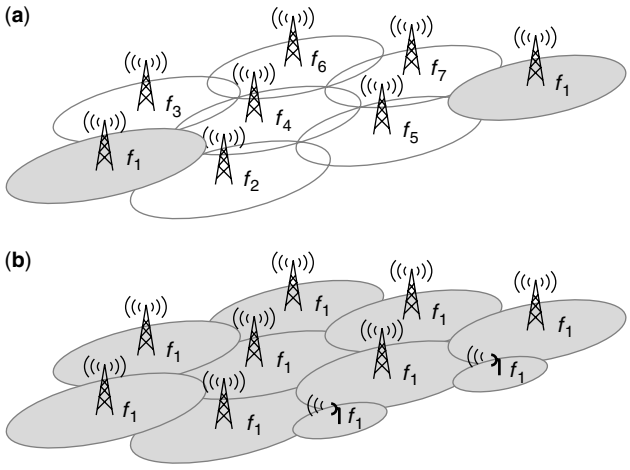


Figure 2. Conventional FM networks (**a**) and single-frequency network (**b**).

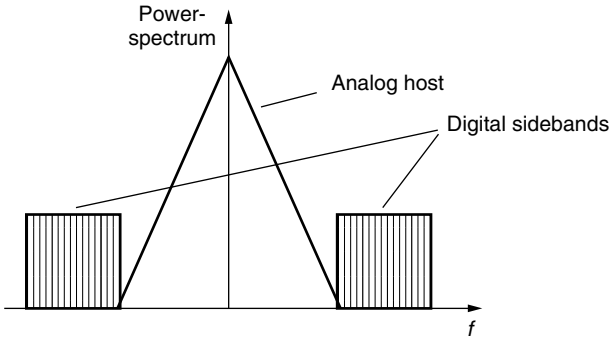


Figure 3. In-band/on-channel signaling.

can be obtained because of the propagation conditions of these frequencies. A major drawback, however, is the poor sound quality, due to the limited bandwidth available in these low-frequency bands.

On the other hand, the coverage properties still make this frequency band attractive for audiobroadcasting. With state-of-the-art audio compression technology and digital transmission methods, the sound quality can be increased to near FM quality.

In 1994 the European project (*Narrow Band Digital Broadcasting* (*NADIB*)) started to develop concepts for a digital system operating in the AM band. The international consortium Digital Radio Mondiale (DRM) [7], which was officially founded in 1998, began to develop an appropriate system. Together with a second approach proposed by iBiquity Digital Corp., the ITU finally gave a recommendation that encompasses both systems for digital sound broadcasting in the frequency bands below 30 MHz [8].

One major difference between both approaches is the strategy to deal with existing analog AM signals. While the DRM scheme is an all-digital approach that occupies all frequencies within the channels with a bandwidth of 9 or 10 kHz or multiples of these bandwidths, the iBiquity system applies the IBOC strategy described above for FM systems.

2.4. Mobile Satellite Digital Audio Radio Services (SDARS)

While the intention of terrestrial digital audio broadcasting technology is to replace existing analog radio in the AM and FM bands, new satellite platforms are emerging that are intended for mobile users. Previous satellite services such as ASTRA Digital (a proprietary standard of the satellite operator SES/ASTRA) or DVB-S, the satellite distribution channel of the European DVB (Digital Video Broadcasting) system, were designed to serve stationary users with directional satellite dishes. In 1999 the WorldSpace Corp. started its satellite service to provide digital high quality audio to emerging market regions such as Africa, Asia, and South and Central America [9]. Originally developed to serve portable receivers, the system is about to be expanded to mobile receivers as well.

Two systems that are targeting mobile users from the beginning are the two U.S. satellite systems operated by Sirius Satellite Radio Inc. and XM Satellite Radio Inc., which intended to launch their commercial service in 2001. Both systems are based on proprietary technology [10].

The three systems differ in the orbital configurations of their satellites. The complete WorldSpace network will consist of three geostationary satellites (each using three spot beams and serving one of the intended coverage areas Africa, parts of Asia, and South and Central America). The XM system uses only two geostationary satellites located to guarantee optimum coverage of the United States. A completely different approach was chosen for the Sirius system, where three satellites in a highly elliptical orbit rise and set over the coverage area every 16 h. The orbit enables the satellites to move across the coverage area at a high altitude (even higher than a geostationary orbit) and therefore also provide a high elevation angle. Two of the three satellites are always visible to provide sufficient diversity when a direct signal to one of the satellites is blocked.

To guarantee reception even in situations when the satellite signal is totally blocked, for example, in tunnels or in urban canyons, Sirius and XM use terrestrial repeaters that rebroadcast the signal.

One major advantage of a satellite system is the large coverage area. Regions with no or poor terrestrial broadcasting infrastructure can be easily supplied with high-quality audio services. On the other hand, it is difficult to provide locally oriented services with this approach.

2.5. Integrated Broadcasting Systems

Besides systems that are designed primarily to broadcast audio services, approaches emerge that provide a technical platform for general broadcasting services. One of these systems is the Japanese Integrated Services Digital Broadcasting (ISDB) approach, which covers satellite, cable, and terrestrial broadcasting as distribution channels. ISDB is intended to be a very flexible multimedia broadcasting concept that incorporates sound and television and data broadcasting in one system. The terrestrial component ISDB-T, which is based on OFDM transmission technology, will be available by 2003–2005. A second scheme that should be mentioned here is DVB-T, the terrestrial branch of the European digital video broadcasting (DVB) system, which is also capable of transmitting transparent services but is not optimized for mobile transmission.

2.6. Which Is the Best System?

To summarize the different approaches in sound broadcasting to mobile receivers, we have to consider the basic demands of the respective market (both customer and broadcaster aspects).

If the situation among the broadcasters is sufficiently homogeneous, which allows the combination of several individual programs to ensembles that are jointly transmitted in a multiplex, then the Eureka-147 system provides a framework for spectrally efficient nationwide digital audiobroadcasting. By adopting this system, a strategy for the transition from analog to digital that is supported by all parties involved must be mapped out.

A more flexible transition from analog to digital with no need for additional frequency bands is possible with the IBOC approach. Also, small local radio stations can be more easily considered using this approach.

However, if the strategy is to cover large areas with the same service, then satellite systems offer advantages over the terrestrial distribution channel.

3. THE EUREKA-147 DAB SYSTEM

Several existing and future systems have been summarized in the previous section. In this section we focus on Eureka-147 DAB as it was the first digital audiobroadcasting system in operational service.

Eureka-147 DAB is designed to be the successor of FM stereobroadcasting. It started as a proposal of a European consortium and is now in operational or pilot

service in countries around the world [5]. Although several alternative approaches are under consideration, right now Eureka-147 DAB is the first all-digital sound broadcasting system that has been in operation for years. When we refer to DAB below, we mean Eureka-147 DAB. Some features of the system are

- Data compression with MPEG-Audio layer 2 (MPEG — Moving Picture Experts Group) according to the standards ISO-MPEG 11172-3 and ISO-MPEG 1318-3 (for comparison with the well-known audio compression algorithm MP3, which is MPEG-Audio layer 3).
- Unequal error protection (UEP) is provided to the compressed audio data where different modes of protection are provided to meet the requirements of different transmission channels, namely, radiofrequency transmission for a variety of scenarios or cable transmission.
- The concept of (coded orthogonal frequency-devision multiplex (COFDM)) copes very well with the problem of multipath propagation, which is one of the major problems in mobile systems. Furthermore, this transmission scheme allows the operation of a single-frequency network (SFN).
- The DAB transmission signal carries a multiplex of several sound and data services. The overall bandwidth of one ensemble is 1.536 MHz and provides a net data rate of approximately 1.5 Mbps. The services can be combined very flexibly within one ensemble. Up to six high-quality audio programs or a mixture of audio and data services can be transmitted in one ensemble.

3.1. General Concept

The basic concept of signal generation in a DAB system is shown in Fig. 4. The input to the system may either be one or several audio programs or one or several data services together with information about the multiplex structure, service information, and so on. Audio and data services form the main service channel (MSC), while service and multiplex information are combined in the fast information channel (FIC). Every input branch undergoes a specific channel coding matched to the particular protection level required by the channel. Data compression is specified only for audio signals in the DAB standard. Data services are transparent and are transmitted in either a packet mode or a stream mode.

After additional time interleaving, which is skipped in Fig. 4, audio and data services form the MSC, which consists of a sequence of so-called common interleaved frames (CIFs) assembled by the main service multiplexer. The final transmission signal is generated in the transmission frame multiplexer, which combines the MSC and the FIC. Together with the preceding synchronization information, OFDM symbols are formed and passed to the transmitter.

In the following the main components of the block diagram (flowchart) in Fig. 4 are explained in more detail.

3.2. Audio Coding

Since the bit rate of a high-quality audio signal (e.g., a CD signal with 2×706 kbps) is too high for a spectral efficient transmission, audio coding according to the MPEG Audio layer 2 standard is applied. For a sampling frequency of 48 kHz the resulting bitstream complies with the ISO/IEC 11172-3 layer 2 format, while a sampling frequency of 24 kHz corresponds to ISO/IEC 13818-3 layer 2 LSF (low sampling frequency). The main idea behind the compression algorithms is utilization of the properties of human audio perception, which are based on spectral and temporal masking phenomena. The concept is shown in Fig. 5, where the sound pressure level is plotted as a function of frequency. The solid curve indicates the threshold in quiet, which is the sound pressure level of a pure tone that is barely audible in a quiet environment. The curve shows that a tone has to show a higher level

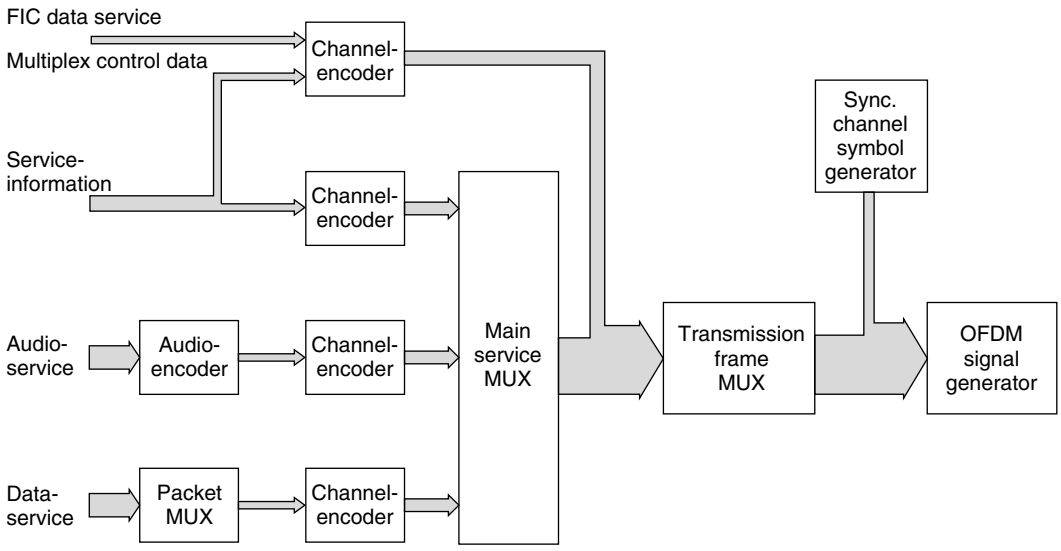


Figure 4. Generation of the transmission signal.

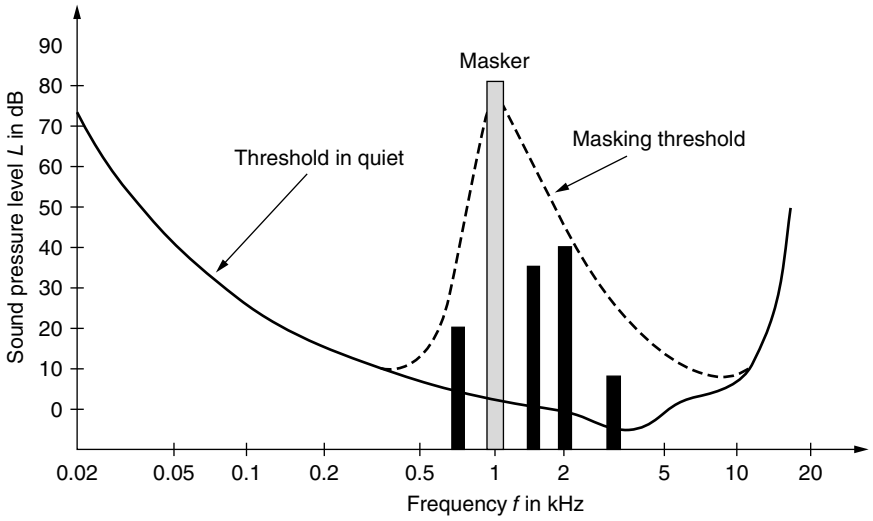


Figure 5. Example for masking.

at very low and very high frequencies to be perceivable than at medium frequencies around 3 kHz, where the human auditory system is very sensitive. Besides the threshold in quiet, each additional sound also creates a masking pattern that is also depicted in the figure. The shape of the pattern depends on the level and the frequency of the underlying sound. A general observation is that the slope of the masking pattern is steeper toward lower frequencies than in the opposite direction. All sound events that lie below this masking pattern (also indicated in Fig. 5) are not perceivable by the human ear and therefore do not have to be transmitted. Since a general audio signal consists of a more complex spectrum, the first step in audio coding is a transformation from the time domain into the frequency domain. Each frequency component creates a masking pattern by itself, and the masking pattern of the overall signal can be calculated by a superposition of the individual patterns. The signal is divided into frequency bands in the spectral domain, and each band is coded (quantized) separately in such a way that the quantization noise lies below the masking threshold in this band. By this technique the quantization noise can be shaped along the frequency axis, and the overall bit rate that is necessary to represent the signal can be reduced.

A simplified block diagram of the audio encoder is shown in Fig. 6. For processing, the sampled signal is divided into segments of length 24 or 48 ms, respectively. Each segment is then transformed from the time domain into the frequency domain by a filter bank with 32 subbands. In parallel, the masking threshold is calculated for each segment in a psychoacoustic model. The subband samples undergo a quantization process in which the number of quantization levels are controlled by the requirements given by the psychoacoustic model. Finally, the quantized samples together with the corresponding side information that is necessary to reconstruct the signal in the decoder are multiplexed into an audio frame. The frame also contains program-associated data (PAD).

The bit rates available for DAB are between 8 and 192 kbps for a monophonic channel. To achieve an audio quality that is comparable to CD quality, approximately 100 kbps are necessary per monochannel, which means a data reduction of a factor of ~7. Since the Eureka-147 DAB standard was fixed in the early 1990s, MPEG Audio layer 2 was chosen for audio coding because it combines good compression results with reasonable complexity.

With today's leading-edge audio compression algorithms such as MPEG2-AAC (advanced audiocoding) or the Lucent PAC (perceptual audiocoder) codec, compression gains of a factor of 12 can be realized without noticeable differences in a CD signal. The AAC codec, for example, is proposed for audiocoding in the Digital Radio Mondiale approach as well as in the Japanese ISDB system, while the PAC technology is applied in the Sirius and XM satellite services, for example.

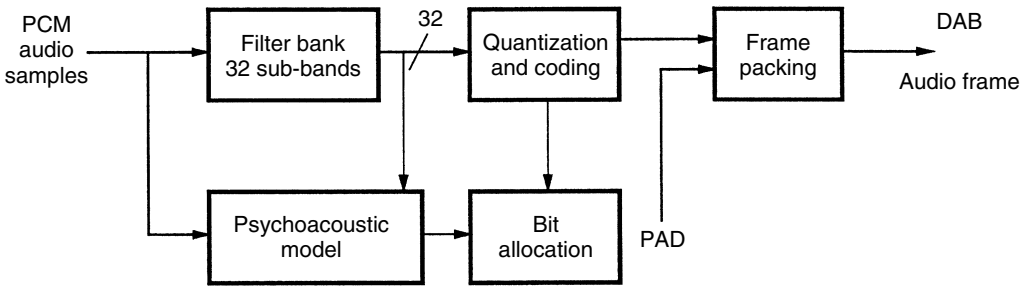


Figure 6. Simplified block diagram of an DAB audio encoder.

3.3. Channel Coding

Between the components for audioencoding and channel encoding an energy dispersal is performed by a pseudo random scrambling that reduces the probability of systematic regular bit patterns in the data stream.

The basic idea of channel coding is to add redundancy to a digital signal in such a way, that the redundancy can be exploited in the decoder to correct transmission errors. In DAB, a convolutional code with memory six and rate $\frac{1}{4}$ is applied. The encoder is depicted in Fig. 7. The code rate $\frac{1}{4}$ means that for every information bit that enters the encoder, 4 coded bits are produced. In the case of an audio signal not every part of the source coded audio frame has the same sensitivity to transmission errors. Very sensitive segments have to be protected by a strong code, while a weaker code can be applied to other parts. This concept, called unequal error protection (UEP), can easily be realized by a technique termed puncturing, which means that not every output bit of the convolutional encoder is transmitted. According to a defined rule, some of the bits are eliminated (punctured) prior to transmission. This technique is also shown in Fig. 7. A binary 1 in the puncturing pattern means that the corresponding bit at the output of the convolutional encoder is transmitted, while a binary 0 indicates that the bit is not transmitted (punctured). In the DAB specification, 24 of these puncturing patterns are defined, which allows a selection of code rates between $\frac{8}{9}$ and $\frac{8}{32}$. A code rate of $\frac{8}{9}$ means that for every 8 bits that enter the encoder, 9 bits are finally transmitted over the channel (depicted in the upper pattern in Fig. 7). On the other hand, with a code rate of $\frac{8}{32}$, all output bits are transmitted as indicated by the lower pattern in Fig. 7. To apply unequal error protection, the puncturing pattern can be changed within an audio frame. This is shown in Fig. 8. While header and side information is protected with the largest amount of redundancy, the scale factors are protected by a weaker code, and the subband samples finally have the least protection. The program-associated data (PAD) at the end of an audio frame are protected roughly by the same code rate as the scale factors. Besides the different protection classes within an audio frame, five protection levels are specified to meet the requirements of the intended transmission scenario. A cable transmission for example needs far less error protection than a very critical mobile multipath environment. Therefore, the DAB specifications contain 64 different protection profiles

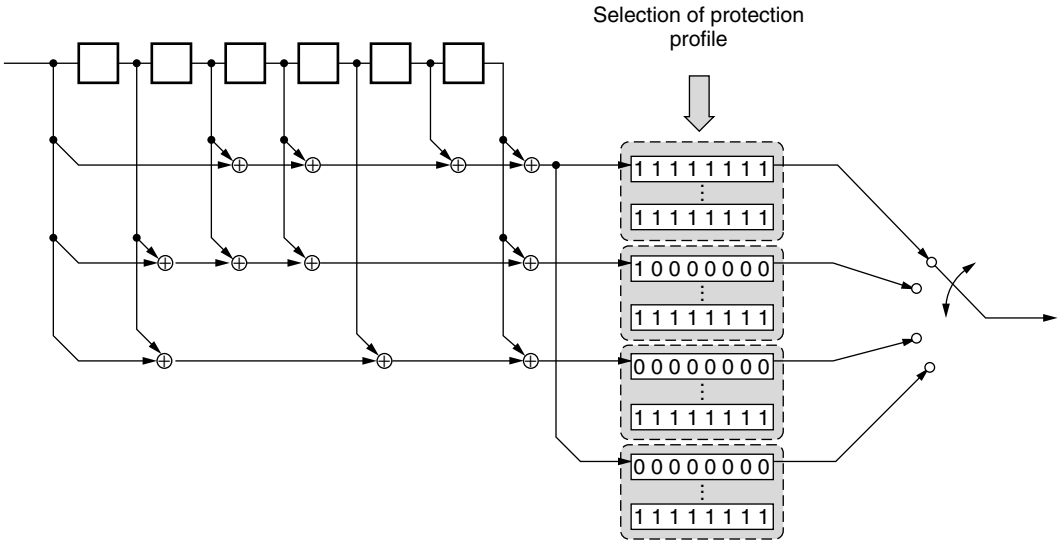


Figure 7. Channel encoder and puncturing.

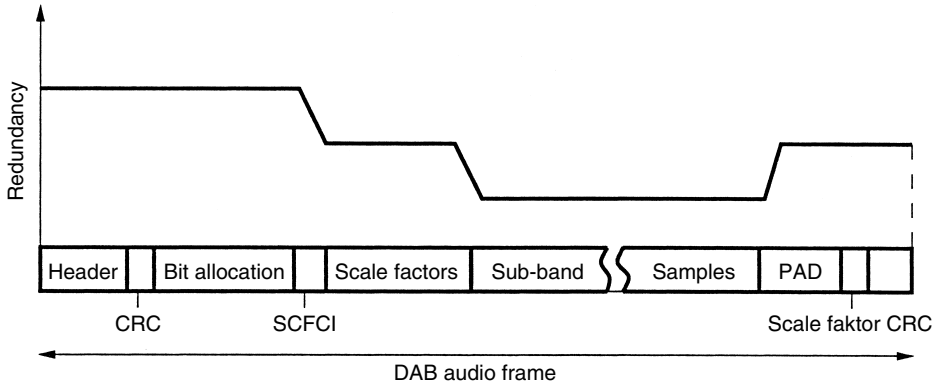


Figure 8. DAB audio frame with unequal error protection.

for different combinations of audio bit rate (32–384 kbps) and protection level.

3.4. Modulation and Transmission Format

As shown in Fig. 4, the channel coded data are multiplexed in the main service multiplexer. The structure of the frames that are finally transmitted across the air interface depends on the chosen transmission mode. DAB provides four different modes depending on the intended radiofrequency range. The duration of a transmission frame is either 24, 48, or 96 ms, and each frame consists of data from the fast information channel (FIC) and the main service channel (MSC) preceded by synchronization information.

The transmission scheme for DAB is orthogonal frequency-devision multiplex (OFDM), the motivation for which will be illustrated in a brief example.

A main problem of the mobile radio channel is multipath propagation. Let us consider a system that transmits at a symbol rate of $r = 1$ Msps (million symbols per second) (with QPSK modulation, this means a bit rate of 2 Mbps) on a single-carrier system. Hence, one symbol has a duration of $T_{sc} = 1\ \mu s$ (where subscript "sc" indicates single carrier), and the bandwidth of the system is $1/T_{sc} = 1$ MHz. We further assume a maximum channel delay τ_{max} of 80 μs. This means a difference in length between the direct path and the path with the largest delay of 24 km, which is a reasonable value for a terrestrial broadcasting channel. If we consider the relation between the symbol duration and the maximal path delay $\tau_{max}/T_{sc} = 80$, we see that this approach leads to heavy intersymbol interference since a received signal is still influenced by the 80 previously sent symbols. We can cope with this problem more easily if we do not transmit the datastream on a single carrier but multiplex the original datastream into N parallel steams (e.g., $N = 1000$) and modulate a separate carrier frequency with each individual stream. With the overall symbol rate given above, the symbol rate on each carrier is reduced to $r/N = 1000$ sps, which means a symbol duration of only $T_{mc} = 1$ ms (where subscript "mc" denotes multiple carriers). If we compare again the symbol duration with the maximal channel delay, a received symbol overlaps with only an 8% fraction of the previous symbol.

But what does this mean for the bandwidth that is necessary to transmit the large number of carriers? A very elegant way to minimize the carrier spacing without any interference between adjacent carriers is to use rectangular pulses on each subcarrier and space the resulting $\sin(x)/x$-type spectra by the inverse of the pulse (symbol) duration T_{mc}. This results in orthogonal carriers, and the overall bandwidth of the system with the parameters given above is $N \times 1/T_{mc} = 1000 \times 1\ \text{kHz} = 1\ \text{MHz}$, which is the same as for the single-carrier approach. This multicarrier approach with orthogonal subcarriers, called orthogonal frequency-devision multiplex (OFDM), is used as the transmission scheme in Eureka-147 DAB. Another advantage of OFDM is the simple generation of the transmission signal by an inverse discrete Fourier transform (IDFT), which can be implemented with low complexity using the fast fourier transform (FFT). Therefore, the modulation symbols of a transmission frame are mapped to the corresponding carriers before an IFFT generates the corresponding time-domain transmission signal. The individual carriers of the DAB signal are DQPSK-modulated, where the "D" stands for differential, meaning that the information is carried in the phase difference between two successive symbols rather than in the absolute phase value. This allows an information recovery at the receiver by just comparing the phases of two successive symbols. To initialize this process, the phase reference symbol has to be evaluated, which is located at the beginning of each transmission frame.

One essential feature of OFDM we did not mention yet is the guard interval. Since each overlapping of received symbols disturbs the orthogonality of the subcarriers and leads to a rapidly decreasing system performance, this effect has to be avoided. By introduction of a guard interval at the beginning of each OFDM symbol, this interference can be avoided. This guard interval is generated by periodically repeating the tail fraction of the OFDM symbol at the beginning of the same symbol. As long as the path delay is not longer than the guard interval, only data that belong to the actual symbol fall into the receiver window. By this technique all information of delayed path components contribute constructively to the received signal. This concept is sketched in Fig. 9.

The length of the guard interval specifies the maximal allowed path delay. In order to fix the parameters of an OFDM system, a tradeoff between two elementary properties of the mobile radio channel has to be made.

On one hand, the guard interval has to be sufficiently large to avoid interference due to multipath effects. This can be obtained by a large symbol duration. Because the carrier spacing is the inverse of the symbol duration, this results in a large number of closely spaced carriers within the intended bandwidth. On the other hand, a small carrier spacing means a high sensitivity to frequency shifts caused by the Doppler effect when the receiver is moving. This, in turn, depends on the used radiofrequency and the speed of the vehicle.

In Table 1 four sets of transmission parameters are given for the different scenarios (transmission modes). Mode I is suitable for terrestrial single frequency networks in the VHF frequency range. Mode II is suitable for smaller single-frequency networks or locally oriented conventional networks because of the rather small guard interval. Mode III has an even smaller guard interval and is designed for satellite transmission on frequencies up to 3 GHz where path delay is not the dominating problem. Mode IV is designed for single-frequency networks operating at frequencies higher than those of Mode I, and the parameters are in between those of Mode I and Mode II.

3.5. DAB Network

As mentioned before, DAB allows the operation of single-frequency networks. Several transmitters synchronously broadcast the same information on the same frequency. This becomes possible because of the OFDM transmission technique. For the receiver, it makes no difference whether

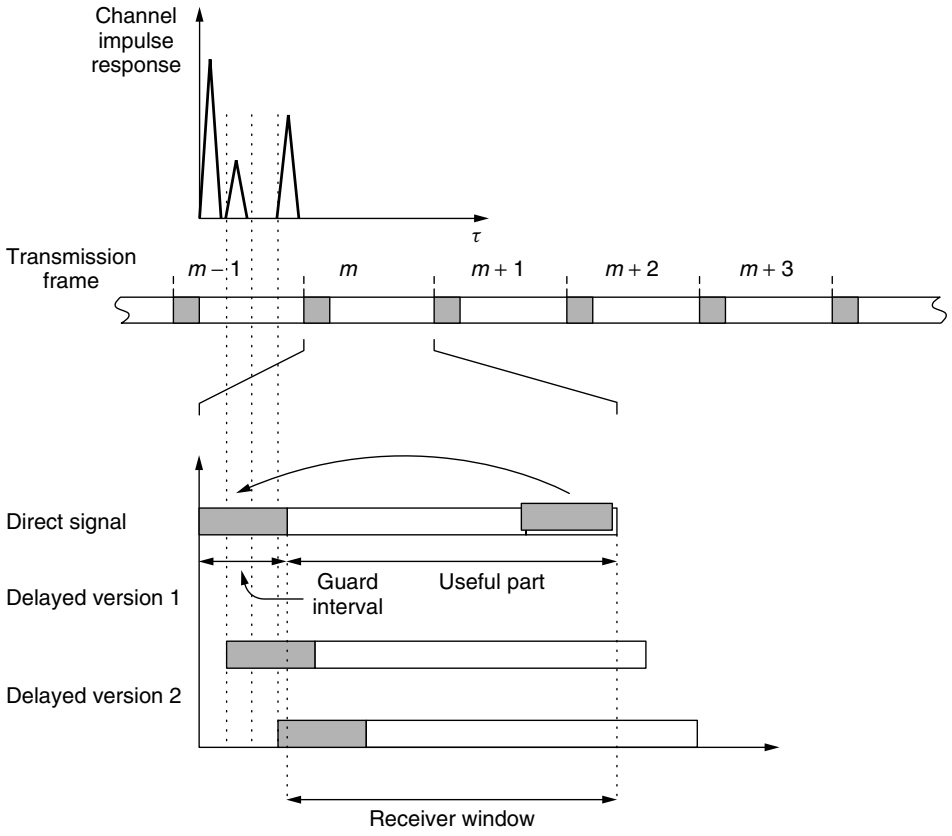


Figure 9. OFDM symbol in multipath environment.

Table 1. DAB Parameters for Transmission Modes I–IV [4]

Parameter	Mode I	Mode II	Mode III	Mode IV
Number of carriers	1536	384	192	768
Frame duration	96 ms	24 ms	24 ms	48 ms
Carrier spacing	1 kHz	4 kHz	8 kHz	2 kHz
Useful symbol duration (inverse carrier spacing)	1 ms	250 μs	125 μs	500 μs
Guard interval duration	246 μs	62 μs	31 μs	123 μs
Maximal transmitter separation	96 km	24 km	12 km	48 km
Frequency range for mobile transmission	≤375 MHz	≤1.5 GHz	≤3 GHz	≤1.5 GHz

the received signal components all come from the same transmitter or stem from different transmitters. Each component contributes constructively to the received signal as long as the path delays stay within the guard interval. Therefore, the maximal distance between the transmitters is determined by the length of the guard interval specified by the used transmission mode. To ensure synchronous transmission within the network, a time reference is necessary, which can be provided, for example, by the satellite-based Global Positioning System (GPS).

BIOGRAPHY

Rainer Bauer received his Dipl.-Ing. degree in electrical engineering and Dr.-Ing. degree from Munich University of Technology (TUM) in 1995 and 2002, respectively. Since 1995, he has been working at the Institute for Communications Engineering of TUM. His areas of interest are source-channel coding and decoding, as well as iterative decoding techniques and their application to wireless audio communications.

BIBLIOGRAPHY

1. W. Hoeg and T. Lauterbach, eds., *Digital Audio Broadcasting—Principles and Applications*, Wiley, New York, 2001.
2. WorldDAB Forum (no date), public documents (online), WorldDAB Documentation: *Thought on a transition scenario from FM to DAB*, Bayerische Medien Technik GmbH (BMT), *http://www.worlddab.org/dab/aboutdab_frame.htm* (Aug. 2001).

3. European Telecommunications Standards Institute, *Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to Mobile, Portable and Fixed Receivers*, European Standard ETSI EN 300 401 V1.3.2 (2000-09), 2000.
4. WorldDAB Forum (no date), public documents (online), *EUREKA-147 — Digital Audio Broadcasting*, *http://www.worlddab.org/dab/aboutdab_frame.htm* (Aug. 2001).
5. WorldDAB (2001), homepage (online), *http://www.WorldDAB.org*.
6. D. Lavers, IBOC — Made in America, *Broadcast Dialogue* (Feb. 2000).
7. Digital Radio Mondiale (2001), Homepage (online), *http://www.DRM.org*.
8. International Telecommunication Union, *Systems for Digital Sound Broadcasting in the Broadcasting Bands below 30 MHz*, draft new recommendation ITU-R BS.[DOC.6/63], Oct. 2000.
9. WorldSpace (2001), homepage (online), *http://www.worldspace.com*.
10. D. H. Layer, Digital radio takes the road, *IEEE Spectrum* **38**(7): 40–46 (2001).

DIGITAL FILTERS

HANOCH LEV-ARI
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

The omnipresence of noise and interference in communication systems makes it necessary to employ filtering to suppress the effects of such unwanted signal components. As the cost, size, and power consumption of digital hardware continue to drop, the superiority of digital filters over their analog counterparts in meeting the increasing demands of modern telecommunication equipment becomes evident in a wide range of applications. Moreover, the added flexibility of digital implementations makes adaptive digital filtering a preferred solution in situations where time-invariant filtering is found to be inadequate.

The design of linear time-invariant digital filters is by now a mature discipline, with methodological roots reaching back to the first half of the twentieth century. Digital filter design combines the power of modern computing with the fundamental contributions to the theory of optimized (analog) filter design, made by Chebyshev, Butterworth, Darlington, and Cauer. In contrast, the construction of adaptive digital filters is still an evolving area, although its roots can be traced back to the middle of the twentieth century, to the work of Kolmogorov, Wiener, Levinson, and Kalman on statistically optimal filtering. Adaptive filtering implementations became practical only after the early 1980s, with the introduction of dedicated digital signal processing hardware.

The practice of digital filter design and implementation relies on two fundamental factors: a well-developed mathematical theory of signals and systems and the availability of powerful digital signal processing hardware. The synergy between these two factors results in an ever-widening range of applications for digital filters, both fixed and adaptive.

1.1. Signals, Systems, and Filters

Signals are the key concept in telecommunications — they represent patterns of variation of physical quantities such as acceleration, velocity, pressure, and brightness. Communication systems transmit the information contained in a signal by converting the physical pattern of variation into its electronic analogue — hence the term "analog signal." *Systems* operate on signals to modify their shape. *Filters* are specialized systems, designed to achieve a particular type of signal shape modification. The relation between the signal that is applied to a filter and the resulting output signal is known as the *response* of the filter.

Most signals encountered in telecommunication systems are *one-dimensional* (1D), representing the variation of some physical quantity as a function of a single independent variable (usually time). Multidimensional signals are functions of several independent variables. For instance, a video signal is 3D, depending on both time and (x, y) location in the image plane. Multidimensional signals must be converted by scanning to a 1D format before transmission through a communication channel. For this reason, we shall focus here only on 1D signals that vary as a function of time.

Filters can be classified in terms of the operation they perform on the signal. Linear filters perform linear operations. Offline filters process signals after they have been acquired and stored in memory; online filters process signals instantaneously, as they evolve. The majority of filters used in telecommunication systems are online and linear. Since the input and output of an online filter are continuously varying functions of time, memory is required to store information about past values of the signal. Analog filters use capacitors and inductors as their memory elements, while digital filters rely on electronic registers.

1.2. Digital Signal Processing

Digital signal processing is concerned with the use of digital hardware to process digital representations of signals. In order to apply digital filters to real life (i.e., analog) signals, there is a need for an input interface, known as *analog-to-digital converter*, and an output interface, known as *digital-to-analog converter*. These involve *sampling* and *quantization*, both of which result in some loss of information. This loss can be reduced by increasing the speed and wordlength of a digital filter.

Digital filters have several advantages over their analog counterparts:

- *Programmability* — a time-invariant digital filter can be reprogrammed to change its configuration and response without modifying any hardware. An adaptive digital filter can be reprogrammed to change its response adaptation algorithm.
- *Flexibility* — digital filters can perform tasks that are difficult to implement with analog hardware, such as

large-scale long-term storage, linear phase response, and online response adaptation.

- *Accuracy and robustness* — the accuracy of digital filters depends mainly on the number of bits (wordlength) used, and is essentially independent of external factors such as temperature and age.
- *Reliability and security* — digital signals can be coded to overcome transmission errors, compressed to reduce their rate, and encrypted for security.
- *Cost / performance tradeoff* — the cost, size, and power consumption of digital hardware continue to drop, while its speed keeps increasing. As a result, a digital implementation often offers a better cost/performance trade-off than its analog counterpart.

Still, digital hardware is not entirely free from limitations, and the choice between digital and analog has to be made on a case-by-case basis.

1.3. Telecommunication Applications of Digital Filters

Among the many application areas of digital filters, we focus only on those applications that are directly related to communication systems. Our brief summary (Section 7) distinguishes between two main types of applications:

- *Time-invariant digital filters* — used as an alternative to analog filters in applications ranging from frequency-selective filtering and symbol detection in digital communication systems, through speech and image coding, to radar and sonar signal processing
- *Adaptive digital filters* — used in applications that require continuous adjustment of the filters response, such as channel equalization, echo cancellation in duplex communication systems, sidelobe cancellation and adaptive beamforming in antenna arrays, and linear predictive speech coding

2. FUNDAMENTALS OF DISCRETE-TIME SIGNALS AND SYSTEMS

2.1. Discrete-Time Signals

A *discrete-time signal* is a sequence of numbers (real or complex). As explained in the introduction (Section 1), such a sequence represents the variation of some physical quantity, such as voltage, pressure, brightness, and speed. Because discrete-time signals are often obtained by sampling of continuous-time signals, their elements are known as *samples*.

The samples of a discrete-time signal $x(n)$ are labeled by a discrete index n, which is an integer in the range $-\infty < n < \infty$. For instance, the signal

$$x(n) = \begin{cases} 2n+1 & 0 \le n \le 2 \\ 0 & \text{else} \end{cases}$$

consists of three nonzero samples preceded and followed by zero samples (Fig. 1).

Because signals represent physical quantities, they are subject to various constraints:

- A signal $x(\cdot)$ is called *bounded* if there exists a positive constant B such that $|x(n)| \le B$ for all n.

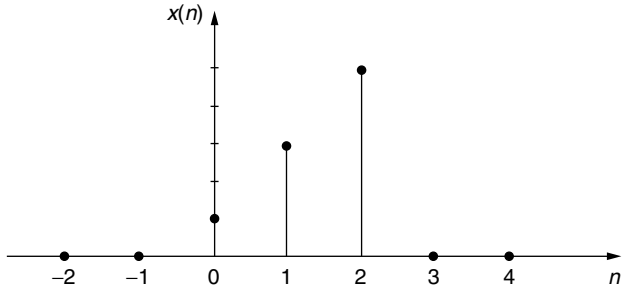


Figure 1. Graphical representation of a discrete-time signal.

- A signal $x(\cdot)$ is said to have *finite energy* if the infinite sum $\sum_{n=-\infty}^{\infty} |x(n)|^2$ converges to a finite value.
- A signal $x(\cdot)$ is said to have *finite average power* if the limit

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |x(n)|^2$$

exists and is finite.

Here, and in the sequel, we use the shorthand notation $x(\cdot)$, which is equivalent to $\{x(n); -\infty < n < \infty\}$, to refer to the entire signal sequence, rather than to a particular sample. From a mathematical standpoint, a bounded signal has finite ℓ_∞ norm, while a finite-energy signal has finite ℓ_2 norm [2].

Certain elementary signal models are essential to the characterization and analysis of discrete-time systems. These include

- The discrete-time unit impulse, also known as the Krönecker delta:[1]

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & n \ne 0 \end{cases}$$

- The discrete-time unit step

$$u(n) = \begin{cases} 1 & n \ge 0 \\ 0 & n < 0 \end{cases}$$

- A two-sided sinusoid, $x(\cdot) = \{\cos 2\pi f_o n; -\infty < n < \infty\}$.
- A one-sided exponential, $a^n u(n)$. A decaying exponential is obtained by choosing $|a| < 1$; in this case the parameter a is known as a "forgetting factor" [3].

We observe that the discrete-time unit impulse $\delta(n)$ satisfies all three constraints — boundedness, finite energy, and finite power — while the discrete-time step $u(n)$ and the sinusoid $\cos 2\pi f_0 n$ are both bounded and have finite power, but infinite energy. Finally, the exponential $a^n u(n)$ satisfies all three constraints when $|a| < 1$, but it violates all three when $|a| > 1$. In general, every finite-energy

[1] The Krönecker delta should not be confused with the continuous-time impulse, namely, the Dirac delta function.

signal is bounded, and every bounded signal must have finite power:

$$\text{Finite energy} \Rightarrow \text{boundedness} \Rightarrow \text{finite power}$$

Also, since finite-energy signals have the property $\lim_{n\to\pm\infty} x(n) = 0$, they represent transient phenomena. Persistent phenomena, which do not decay with time, are represented by finite-power signals.

2.2. Discrete-Time Systems

A discrete-time system is a *mapping* namely, an operation performed on a discrete-time signal $x(\cdot)$ that produces another discrete-time signal $y(\cdot)$ (Fig. 2). The signal $x(\cdot)$ is called the *input* or *excitation* of the system, while $y(\cdot)$ is called the *output* or *response*.

Most synthetic (human-made) systems are *relaxed,* in the sense that a zero input [i.e., $x(n) = 0$ for all n] produces a zero output. Nonrelaxed systems, such as a wristwatch, rely on internally stored energy to produce an output without an input. In general every system can be decomposed into a *sum* of two components: a relaxed subsystem, which responds to the system input, and an autonomous subsystem, which is responsible for the zero-input response (Fig. 3). Since every analog and every digital filter is a relaxed system, we restrict the remainder of our discussion to relaxed systems only.

Relaxed systems can be classified according to certain input–output properties:

- A relaxed system is called *memoryless* or *static* if its output $y(n)$ at any given instant n depends only on the input sample $x(n)$ at the same time instant, and does not depend on any other input sample. Such a system can be implemented without using memory; it maps every input sample, as it becomes available, into a corresponding output sample.
- A relaxed system is called *dynamic* if its output $y(n)$ at any given instant n depends on past and/or future samples of the input signal $x(\cdot)$. This means that memory is required to store each input sample as it becomes available until its processing has been completed, and it is no more needed by the system. The number of input samples that need to be kept in memory at any given instant is known as the order of the system, and it is frequently used as a measure of system complexity.
- A relaxed system is called *causal* if its output $y(n)$ at any given instant n does not depend on future samples of the input $x(\cdot)$. In particular, every memoryless system is causal. Causality is a physical constraint, applying to systems in which n indicates (discretized) physical time.
- A relaxed system is called *time-invariant* if its response to the time-shifted input $x(n - n_o)$ is a similarly shifted output $y(n - n_o)$, for every input signal $x(\cdot)$ and for every integer n_o, positive or negative. In the context of (digital) filters, time-invariance is synonymous with fixed hardware, while time variation is an essential characteristic of adaptive filters.
- A relaxed system is called *stable* if its response to any bounded input signal $x(\cdot)$ is a bounded output signal $y(\cdot)$. This is known as bounded-input/bounded-output (BIBO) stability, to distinguish it from other definitions of system stability, such as internal (or Lyapunov) stability.
- A relaxed system is called *linear* if its response to the input $x(\cdot) = a_1x_1(\cdot) + a_2x_2(\cdot)$ is $y(\cdot) = a_1y_1(\cdot) + a_2y_2(\cdot)$ for any a_1, a_2 and any $x_1(\cdot), x_2(\cdot)$. Here $y_1(\cdot)$ (resp. $y_2(\cdot)$) is the system's response to the input $x_1(\cdot)$ [resp. $x_2(\cdot)$].

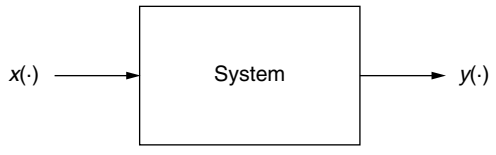


Figure 2. Block diagram representation of a discrete-time system.

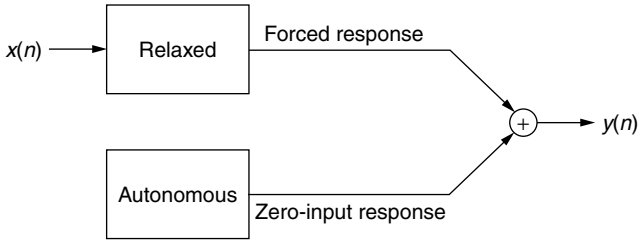


Figure 3. The decomposition of a nonrelaxed system.

As explained in the introduction, digital filters can be either fixed or adaptive. Fixed digital filters are time invariant and most often linear, while adaptive digital filters usually consist of a linear time-variant module and a nonlinear time-invariant module.

Our definitions of fundamental system properties make use of several basic building blocks of discrete-time systems in general, and digital filters in particular:

- *Unit-delay element* — delays a signal by one sample (Fig. 4).
- *Adder* — forms the sum of two signals, such that each output sample is obtained by adding the corresponding input samples (Fig. 5).
- *Signal scaler* — multiplies each sample of the input signal by a constant (Fig. 6). It is the discrete-time equivalent of an (wideband) amplifier.

These three basic building blocks — unit delay, addition, and signal scaling — are relaxed, linear, time-invariant,

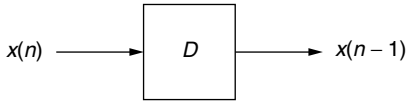


Figure 4. Block diagram representation of a unit-delay element.

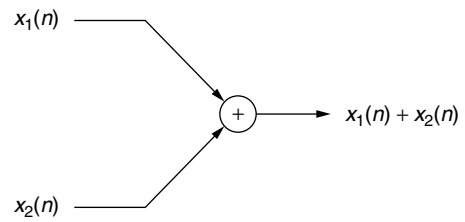


Figure 5. Block diagram representation of an adder.

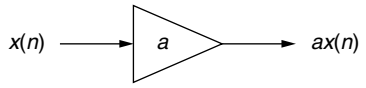


Figure 6. Block diagram representation of a signal scaler.

causal, and stable systems. In fact, every linear, time-invariant, and causal system can be implemented as a network of interconnected delays, scalers and adders (see Section 3.2). Notice that the adder and scaler are memoryless, while the unit delay is dynamic.

In reality, scalers are implemented using digital multipliers (Fig. 7). Multipliers can also be used to form the pointwise product of two signals (Fig. 8), and they serve as a fundamental building block in adaptive digital filters and other nonlinear systems.

2.3. Discrete-Time Sinusoidal Signals

A discrete-time signal of the form

$$x(n) = A\cos(2\pi f_0 n + \phi) \tag{1a}$$

is called *sinusoidal*. The *amplitude* A, *frequency* f_0, and *phase shift* ϕ are all real and, in addition

$$A > 0, \qquad 0 \le f_0 \le \tfrac{1}{2}, \qquad -\pi \le \phi \le \pi \tag{1b}$$

The *radial frequency* $\omega_0 = 2\pi f_0$ $(0 \le \omega_0 \le \pi)$ is often used as an alternative to f_0. The dimensions of f_0 are cycles per sample, and those of ω_0 are radians per sample.

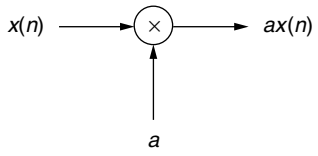


Figure 7. Implementing a signal scaler using a multiplier.

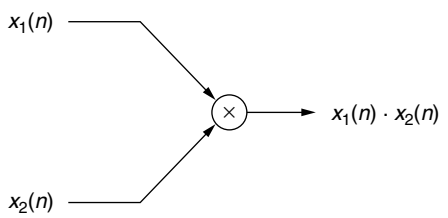


Figure 8. Block diagram representation of a signal multiplier.

The restriction imposed on the range of the frequency f_0 stands in sharp contrast to the continuous-time case, where sinusoids can have unlimited frequencies. However, in the discrete-time context this restriction is necessary in order to avoid *aliasing*.

Aliasing occurs because the two discrete-time sinusoidal signals

$$x_1(n) = \cos 2\pi f_1 n, \qquad x_2(n) = \cos 2\pi f_2 n$$

are indistinguishable when $f_1 - f_2$ is an integer. For instance, $\cos\left(\frac{3\pi}{2}n\right) = \cos\left(-\frac{\pi}{2}n\right)$, so that the frequency $f_1 = \frac{3}{4}$ cannot be distinguished from its *alias* $f_2 = -\frac{1}{4}$. From a mathematical standpoint, we observe that the sinusoidal signal of (1) is periodic in f_0 with a period of 1. In order to avoid ambiguity, we must therefore restrict f_0 to its fundamental (or principal) range $-\frac{1}{2} < f_0 \le \frac{1}{2}$. In addition, by using the symmetry property $\cos(-x) = \cos x$, we can avoid using negative frequencies altogether, which leads to the frequency range specified in Eq. (1b).

The constraint on f_0 is closely related to the well-known Nyquist condition: the range of frequencies allowed at the input of a sampler cannot exceed half the sampling rate. In particular, when the continuous-time sinusoid $\cos 2\pi f_0 t$ is sampled at a rate of F_s samples per second, the resulting sequence of samples forms a discrete-time sinusoid, $x(n) = \cos 2\pi f_0 n$, where $f_0 = f_0/F_s$. According to the Nyquist condition we must have $f_0 < F_s/2$ in order to avoid aliasing: the equivalent restriction on f_0 is $f_0 < \frac{1}{2}$. Anti-aliasing filters must be used to avoid the presence of aliased components in sampled signals.

2.4. Relaxed Linear Time-Invariant (LTI) Systems

The input–output characterization of relaxed discrete-time LTI systems relies on the notion of *impulse response*, namely the response of the system to the discrete-time unit impulse $\delta(\cdot)$ (Fig. 9).

The response of such a system to an arbitrary input $x(\cdot)$ is determined by a linear convolution between the input signal and the impulse response of the system:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) \tag{2}$$

This is usually expressed using the shorthand notation

$$y(\cdot) = h(\cdot) \circledast x(\cdot)$$

Thus, a relaxed LTI system is completely characterized by its impulse response. In particular, such a system is causal if, and only if

$$h(n) = 0 \quad \text{for} \quad n < 0$$

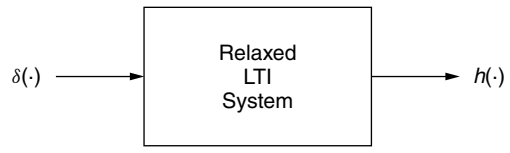


Figure 9. The impulse response of a relaxed LTI system.

As a result the convolution sum (2) for a causal system ranges only over $0 \le k < \infty$. Also, an LTI system is BIBO stable if, and only if

$$\sum_{k=-\infty}^{\infty} |h(k)| < \infty$$

The response of an LTI system to a sinusoidal input is of particular interest. The input signal $x(n) = \exp\{j\omega_0 n\}$ produces the output

$$y(n) = H(e^{j\omega_0})e^{j\omega_0 n} \equiv H(e^{j\omega_0})x(n) \tag{3a}$$

where

$$H(e^{j\omega}) = \sum_{k=-\infty}^{\infty} h(k)e^{-j\omega k} \tag{3b}$$

is known as the *frequency response* of the LTI system. We observe that the response to the complex sinusoid $x(n) = \exp\{j\omega_0 n\}$ is $H(e^{j\omega_0})x(n)$, namely, a scaled version of the input signal. From a mathematical standpoint, this means that complex sinusoids are eigenfunctions of LTI systems.

The infinite sum (3b) that defines the frequency response $H(e^{j\omega})$ converges for all ω if, and only if, $\sum_k |h(k)| < \infty$, which is also the necessary and sufficient conditions for BIBO stability. Thus, an unstable system does not have a frequency response — exciting it with a sinusoidal input produces an unbounded output signal. Unstable systems can be characterized by their response to exponentially decaying sinusoids, which leads us to the more general concept of a $\mathcal{Z}$ transform (see Section 3.1).

A real-life LTI system has a real-valued impulse response. The response of such a system to the real-valued sinusoid $x(n)\cos\omega_0 n$ is given by the real part of the response in (3)

$$y(n) = \mathrm{Re}\{H(e^{j\omega_0})e^{j\omega_0 n}\} = |H(e^{j\omega_0})|\cos[\omega_0 n + \theta(\omega_0)]$$

where $\theta(\omega_0) = \arg H(e^{j\omega_0})$. The effect of an LTI system on a real sinusoid is to scale its amplitude by $|H(e^{j\omega_0})|$, and to increase its phase by $\arg H(e^{j\omega_0})$. For this reason, $|H(e^{j\omega})|$ is known as the *magnitude response* of the system, and $\arg H(e^{j\omega})$ is known as its *phase response*.

3. TRANSFORM DOMAIN ANALYSIS OF SIGNALS AND SYSTEMS

Transform-domain analysis is a powerful technique for characterization and design of (fixed) digital filters. The role of the $\mathcal{Z}$ transform in the context of discrete-time signals and systems is similar to the role of the Laplace transform in the context of continuous-time signals and systems. In fact, the two are directly related via the process of sampling and reconstruction. Recall that interpolation, specifically the process of reconstructing a continuous-time signal from its samples, produces the continuous-time signal

$$x_r(t) = \sum_{n=-\infty}^{\infty} x(n)g(t - nT)$$

where $g(\cdot)$ is the impulse response of the interpolating filter and T is the sampling interval. The Laplace transform of this signal is

$$X_r(s) = \left[\sum_{n=-\infty}^{\infty} x(n)e^{-snT}\right] G(s)$$

namely, a product of the transfer function $G(s)$ of the interpolating filter with a transform-domain characterization of the discrete-time sequence of samples $x(\cdot)$. This observation motivates the introduction of the $\mathcal{Z}$ transform

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \tag{4}$$

so that $X_r(s) = G(s)X(z)|_{z=e^{sT}}$.

The $\mathcal{Z}$ transform converts difference equations into algebraic equations, which makes it possible to characterize every discrete-time LTI system in terms of the poles and zeros of its *transfer function*, namely, the $\mathcal{Z}$ transform of its impulse response. This is entirely analogous to the role played by the Laplace transform in the context of continuous-time LTI systems.

3.1. The $\mathcal{Z}$ Transform and the Discrete-Time Fourier Transform

From a mathematical standpoint, the $\mathcal{Z}$ transform (4) is the sum of two power series, viz., $X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} = X_+(z) + X_-(z)$, where

$$X_+(z) = \sum_{n=0}^{\infty} x(n)z^{-n}, \qquad X_-(z) = \sum_{n=-1}^{-\infty} x(n)z^{-n}$$

The $\mathcal{Z}$ transform $X(z)$ is said to exist only if both $X_+(z)$ and $X_-(z)$ converge absolutely and uniformly. This implies that the *region of convergence* (RoC) of $X(z)$ is $\{z; r < |z| < R\}$, where r is the radius of convergence of $X_+(z)$ and R is the radius of convergence of $X_-(z)$. Thus a $\mathcal{Z}$ transform exists if, and only if, $r < R$. When it does exist, it is an analytic function within its RoC.

The $\mathcal{Z}$ transform converts time-domain convolutions into transform-domain products. In particular, the input–output relation of a relaxed LTI system, specifically, $y(\cdot) = h(\cdot) \circledast x(\cdot)$ transforms into $Y(z) = H(z)X(z)$, where $X(z)$, $Y(z)$, and $H(z)$ are the $\mathcal{Z}$ transforms of $x(\cdot)$, $y(\cdot)$, and $h(\cdot)$, respectively. Thus, specifying the *transfer function*

$$H(z) = \sum_{k=-\infty}^{\infty} h(k)z^{-k}$$

along with its RoC provides the same information about the input–output behavior of a relaxed LTI system as the impulse response $h(\cdot)$. In particular

- The system is causal if, and only if, the RoC of $H(z)$ is of the form $|z| > r$ for some $r \ge 0$.
- The system is (BIBO) stable if, and only if, the unit circle is within the RoC of $H(z)$.

Thus the transfer function of a stable system is always well defined on the unit circle $\mathbf{T} = \{z; |z| = 1\}$, and we recognize $H(z)|_{z=e^{j\omega}}$ as the frequency response of the system [as defined in (3)]. More generally, if $x(\cdot)$ is a discrete-time sequence, whether a signal or an impulse response, such that $\sum_{n=-\infty}^{\infty} |x(n)| < \infty$, then the unit circle $\mathbf{T}$ is included in the RoC of the $\mathcal{Z}$ transform $X(z)$, and so $X(z)|_{z=e^{j\omega}}$ is well defined. The resulting function of ω, that is

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \tag{5a}$$

is called the *discrete-time Fourier transform* (DTFT) of $x(\cdot)$.

From a mathematical standpoint, relation (5) is a complex Fourier series representation of the periodic function $X(e^{j\omega})$, and we recognize the samples $x(n)$ as the Fourier coefficients in this representation. The standard expression for Fourier coefficients

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n} d\omega \tag{5b}$$

is known as the *inverse DTFT*. The DTFT (5) converges absolutely (and uniformly) for ℓ_1 sequences, and the resulting limit is a continuous function of ω. The DTFT can be extended to other types of sequences by relaxing the notion of convergence of the infinite sum in (5a). For instance, the DTFT of ℓ_2 sequences is defined by requiring convergence in the $\mathcal{L}^2(\mathbf{T})$ norm, and the resulting limit is a square-integrable function on the unit circle $\mathbf{T}$. A further extension to ℓ_∞ sequences (= bounded signals) results in limits that are $\mathcal{L}^1(\mathbf{T})$ functions, and thus may contain impulsive components. For instance, the DTFT of the bounded signal $x(n) = e^{j\omega_0 n}$ is $X(e^{j\omega}) = 2\pi\delta(\omega - \omega_0)$, where $\delta(\cdot)$ is the Dirac delta function. The convergence of (5) in this case is defined only in the distribution sense. Also, notice that this signal has no $\mathcal{Z}$ transform.

3.2. Transfer Functions of Digital Filters

A fixed digital filter implements a realizable discrete-time linear time-invariant system. Realizability restricts us to causal systems with *rational transfer functions*

$$H(z) = \frac{b(z)}{a(z)} \tag{6a}$$

where $a(z)$ and $b(z)$ are finite-order polynomials in z^{-1}:

$$a(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_N z^{-N} \tag{6b}$$

$$b(z) = b_o + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_M z^{-M} \tag{6c}$$

The roots of the numerator polynomial $b(z)$ are known as the *zeros* of the transfer function $H(z)$, and the roots of the denominator polynomial $a(z)$ are known as the *poles* of $H(z)$. A digital filter with $N = 0$ is known as a *finite-impulse response* (FIR) filter, while one with $N > 0$ is known as an *infinite-impulse response* (IIR) filter. In view of our earlier statements about causality and stability in terms of the region of convergence, it follows that a digital filter is stable if all its poles have magnitudes strictly less than unity. Thus, FIR filters are unconditionally stable, because all of their poles are at $z = 0$.

In view of (6), the input–output relation of a digital filter can be written as $a(z)Y(z) = b(z)X(z)$, which corresponds to a *difference equation* in the time domain

$$\begin{aligned} y(n) + a_1 y(n-1) + \cdots + a_N y(n-N) \\ = b_o x(n) + b_1 x(n-1) + \cdots + b_M x(n-M) \end{aligned} \tag{7}$$

The same input–output relation can also be represented by a block diagram, using multipliers (actually signal scalers), adders, and delay elements (Fig. 10). Such a block diagram representation is called a *realization* of the transfer function $H(z) = \dfrac{b(z)}{a(z)}$. Digital filter realizations are not unique: the one shown in Fig. 10 is known as the *direct form type 2* realization. Other realizations are described in Section 5.

A hardware implementation (i.e., a digital filter) is obtained by mapping the realization into a specific platform, such as a DSP chip or an ASIC (see Section 5). A software implementation is obtained by mapping the same realization into a computer program.

3.3. The Power Spectrum of a Discrete-Time Signal

The *power spectrum* of a finite-energy signal $x(\cdot)$ is defined as the square of the magnitude of its DTFT $X(e^{j\omega})$. Alternatively, it can be obtained by applying a DTFT to the *autocorrelation sequence*

$$C_{xx}(m) = \sum_{n=-\infty}^{\infty} x(n+m)x^*(n) = x(m) \circledast x^*(-m)$$

where the asterisk $(*)$ denotes complex conjugation. For finite-power signals the DTFT $X(e^{j\omega})$ may contain impulsive components, so that the square of its magnitude cannot be defined. Instead, the autocorrelation is defined in this case as

$$C_{xx}(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n+m)x^*(n)$$

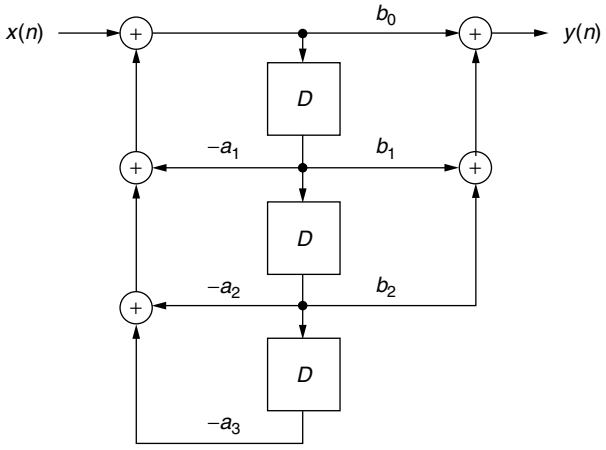


Figure 10. Direct-form type 2 realization of a digital filter with $M = 2$ and $N = 3$.

and the power spectrum is defined as the DTFT of the autocorrelation $C_{xx}(\cdot)$. Similarly, the autocorrelation of a random stationary signal is defined as

$$C_{xx}(m) = E\{x(n+m)x^*(n)\}$$

where $E\{\ \}$ denotes expectation (i.e., probabilistic mean). Thus, in all three cases the power spectrum is

$$S_{xx}(e^{j\omega}) = \sum_{m=-\infty}^{\infty} C_{xx}(m)e^{-j\omega m} \tag{8a}$$

and can be viewed as a restriction to the unit circle of the so-called *complex power spectrum*

$$S_{xx}(z) = \sum_{m=-\infty}^{\infty} C_{xx}(m)z^{-m} \tag{8b}$$

The complex power spectrum is used in the design of optimal (Wiener) filters (see Section 6.1).

Similarly, the cross-correlation of two signals is defined as

$$C_{yx}(m) = \begin{cases} \displaystyle\sum_{n=-\infty}^{\infty} y(n+m)x^*(n) \\ \qquad \text{finite-energy signals} \\ \displaystyle\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} y(n+m)x^*(n) \\ \qquad \text{finite-power signals} \\ E\{y(n+m)x^*(n)\} \\ \qquad \text{jointly stationary random signals} \end{cases} \tag{9a}$$

and the (complex) cross spectrum is the transform of the cross-correlation:

$$S_{yx}(z) = \sum_{m=-\infty}^{\infty} C_{yx}(m)z^{-m} \tag{9b}$$

When $y(\cdot)$ is the response of a stable LTI system to the input $x(\cdot)$, then

$$S_{yx}(z) = H(z)S_{xx}(z) \tag{10}$$

where $H(z)$ is the transfer function of the system, and

$$S_{yy}(z) = H(z)S_{xx}\left[H\left(\frac{1}{z^*}\right)\right]^* \tag{11}$$

In particular, by using a unit-power white noise [i.e., one with $S_{xx}(z) = 1$] input, we find that

$$S_{yy}(z) = H(z)\left[H\left(\frac{1}{z^*}\right)\right]^* \tag{12}$$

This expression is called a *spectral factorization* of the complex power spectrum $S_{yy}(z)$ and the transfer function $H(z)$ is called a *spectral factor*.

4. DESIGN OF FREQUENCY-SELECTIVE DIGITAL FILTERS

The complete process of designing a digital filter consists of seven stages:

1. *Problem analysis* — determine what the filter is supposed to accomplish.
2. *Filter specification* — select a desired (ideal) frequency response for the filter, and decide how accurately it should be approximated.
3. *Filter design* — obtain the coefficients of a realizable transfer function $H(z)$ that approximates the desired frequency response within the specified tolerance.
4. *Filter realization* — determine how to construct the filter by interconnecting basic building blocks, such as delays, adders, and signal scalers (i.e., multipliers).
5. *Implementation choices* — select the specific hardware/software platform in which the building blocks will be implemented.
6. *Performance analysis* — use the physical parameters of the selected platform (accuracy, cost, speed, etc.) to evaluate the compliance of the selected implementation with the specification of the filter.
7. *Construction* — implement the specific choices made in the design and realization stages into the selected hardware/software platform.

In the problem analysis stage the designer uses specific information about the application to determine a desired frequency response, say, $D(\omega)$, for the filter. The desired magnitude response $|D(\omega)|$ is often a classical (ideal) frequency-selective prototype — lowpass, highpass, bandpass, or bandstop — except in specialized filter designs such as Hilbert transformers, differentiators, notch filters, or allpass filters. This means that $|D(\omega)|$ is piecewise constant, so that the frequency scale decomposes into a collection of bands. The desired phase response $\arg D(\omega)$ could be linear (exactly or approximately), minimum-phase, or unconstrained.

The designer must also define the set of parameters that determine the transfer function of the filter to be designed. First a choice has to be made between FIR and IIR, considering the following facts:

- Exact linear phase is possible only with FIR filters.
- FIR filters typically require more coefficients and delay elements than do comparable IIR filters, and therefore involve higher input–output delay and higher cost.
- Currently available design procedures for FIR filters can handle arbitrary desired responses (as opposed to ideal prototypes) better than IIR design procedures.
- Finite precision effects are sometimes more pronounced in IIR filters. However, such effects can be ameliorated by choosing an appropriate realization.
- Stability of IIR filters is harder to guarantee in adaptive filtering applications (but not in fixed filtering scenarios).

Additional constraints, such as an upper bound on the overall delay (e.g., for decision feedback equalization), or a requirement for maximum flatness at particular frequencies, serve to further reduce the number of independent parameters that determine the transfer function of the filter.

On completion of the problem analysis stage, the designer can proceed to formulate a specification and determine a transfer function that meets this specification, as described in the remainder of Section 4. The remaining stages of the design process (realization, implementation, performance analysis, and construction) are discussed in Section 5.

4.1. Filter Specification

Once the desired frequency response $D(\omega)$ has been determined along with a parametric characterization of the transfer function $H(z)$ of the designed filter, the next step is to select a measure of approximation quality and a tolerance level (i.e., the highest acceptable deviation from the desired response). The most commonly used measure is the weighted Chebyshev (or $\mathcal{L}^\infty$) norm

$$\max_\omega W(\omega)|H(e^{j\omega}) - D(\omega)|$$

where $W(\omega)$ is a nonnegative weighting function. Alternative measures include the weighted mean-square (or $\mathcal{L}^2$) norm

$$\int_{-\pi}^{\pi} W^2(\omega)|H(e^{j\omega}) - D(\omega)|^2 d\omega$$

and the truncated time-domain mean-square norm

$$\sum_{n=0}^{L} |h(n) - d(n)|^2$$

where $h(n)$ is the impulse response of the designed filter and $d(n)$ is the inverse DTFT of the desired frequency response $D(\omega)$.

The mean-square measures are useful mainly when the desired response $D(\omega)$ is arbitrary, since in this case optimization in the Chebyshev norm can be quite demanding. However, the approximation of ideal prototypes under the Chebyshev norm produces excellent results at a reasonable computational effort, which makes it the method of choice in this case.

The approximation of ideal prototypes is usually carried out under the *modified* Chebyshev norm

$$\max_\omega W(\omega)||H(e^{j\omega})| - |D(\omega)||$$

The elimination of phase information from the norm expression reflects the fact that in this case phase is either completely predetermined (for linear-phase FIR filters) or completely unconstrained (for IIR filters). Furthermore, the desired magnitude response is constant in each frequency band (e.g., unity in passbands and zero in stopbands), so it makes sense to select a weighting function $W(\omega)$ that is also constant in each frequency band. As a result, constraining the Chebyshev norm to a prescribed tolerance level is equivalent to providing a separate tolerance level for each frequency band, say

$$||H(e^{j\omega})| - D_i| \le \delta_i \quad \text{for all} \quad \omega \in B_i \tag{13}$$

where B_i is the range of frequencies for the ith band, D_i is the desired magnitude response in that band, and δ_i is the prescribed level of tolerance. For instance, the specification of a lowpass filter is

$$\begin{array}{lll} \text{Passband:} & 1 - \delta_p \le |H(e^{j\omega})| \le 1 + \delta_p & \text{for } |\omega| \le \omega_p \\ \text{Stopband:} & |H(e^{j\omega})| \le \delta_s & \text{for } \omega_p \le |\omega| \le \pi \end{array}$$

as illustrated by the template in Fig. 11. The magnitude response of the designed filter must fit within the unshaded area of the template. The tolerance δ_p characterizes *passband ripple*, while δ_s characterizes *stopband attenuation*.

4.2. Design of Linear-Phase FIR Filters

The transfer function of an FIR digital filter is given by (6) with $N = 0$, so that $a(z) = 1$ and $H(z) \equiv b(z) = \sum_{i=0}^{M} b_i z^{-i}$. The design problem is to determine the values of the coefficients $\{b_i\}$ so that the phase response is linear, and the magnitude response $|H(e^{j\omega})|$ approximates a specified $|D(\omega)|$. In order to ensure phase linearity, the coefficients $\{b_i\}$ must satisfy the symmetry constraint

$$b_{M-i} = sb_i \quad \text{for} \quad 0 \le i \le M, \quad \text{where} \quad s = \pm 1 \tag{14}$$

As a result, the frequency response satisfies the constraint $H^*(e^{j\omega}) = sH(e^{j\omega})e^{jM\omega}$, so that the phase response is indeed linear: $\arg H(e^{j\omega}) = (M\omega + s\pi)/2$. Another consequence of the symmetry constraint (14) is that the zero pattern of $H(z)$ is symmetric with respect to the unit circle: $H(z_o) = 0 \leftrightarrow H(1/z_o^*) = 0$.

The design of a linear-phase FIR filter is successfully completed when we have selected values for the filter coefficients $\{b_i\}$ such that (1) the symmetry constraint (14) is satisfied and (2) the magnitude tolerance constraints (13) are satisfied. A design is considered *optimal* when the order M of the filter $H(z)$ is as small as possible under the specified constraints. It is customary to specify the magnitude tolerances in decibels — for instance, a lowpass filter is characterized by

$$R_p = 20 \log \frac{1 + \delta_p}{1 - \delta_p}, \qquad A_s = 20 \log \frac{1}{\delta_s}$$

The most popular techniques for linear-phase FIR design are

- *Equiripple* — optimal in the weighted Chebyshev norm, uses the Remez exchange algorithm to optimize filter coefficients. The resulting magnitude response is equiripple in all frequency bands (as in Fig. 11). The weight for each band is set to be inversely proportional to the band tolerance. For instance, to design a lowpass filter with passband ripple $R_p = 1$ dB and stopband attenuation $A_s = 40$ dB, we

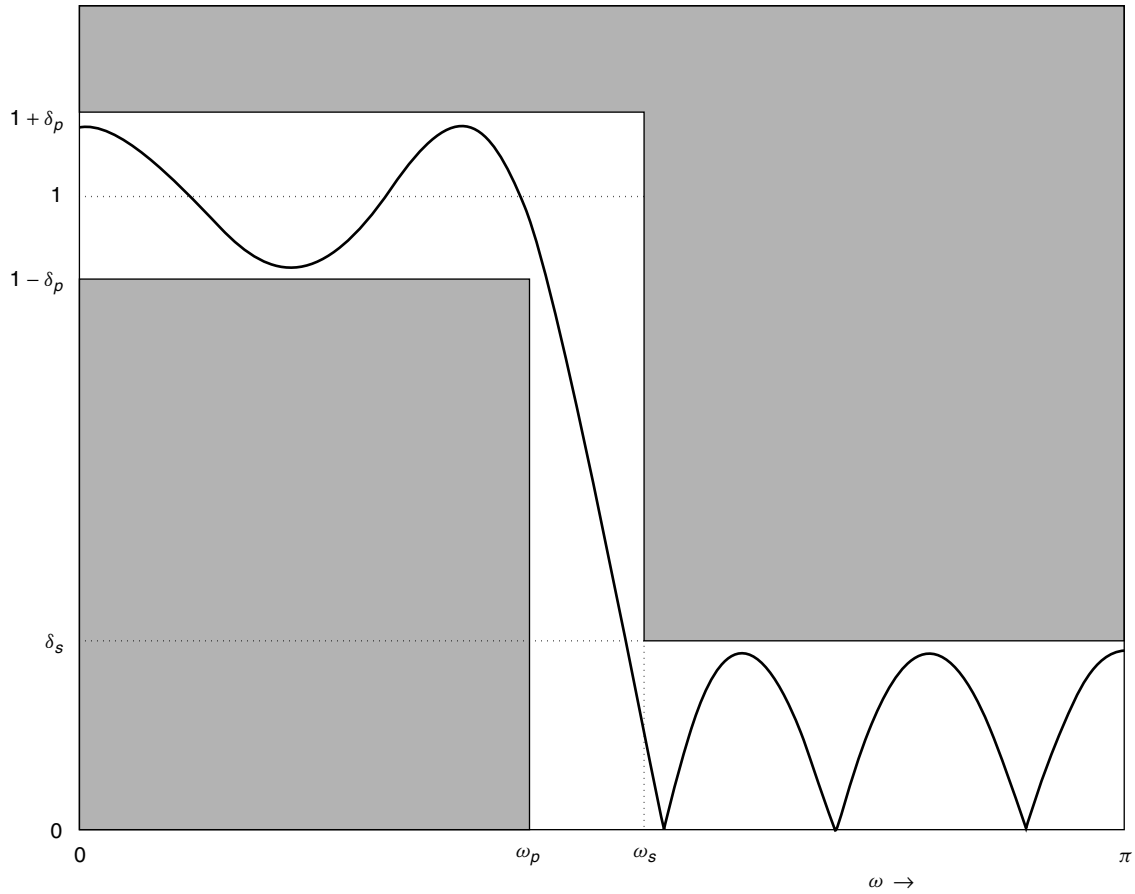


Figure 11. Specification for FIR lowpass filter design.

calculate first $\delta_p = 0.0575$ and $\delta_s = 0.01$ and then set $W_p = \delta_p^{-1} = 17.39$ and $W_s = \delta_s^{-1} = 100$. While, in principle, the Remez exchange algorithm can be applied to any desired magnitude response $|D(\omega)|$, most filter design packages (such a the *signal processing toolbox* in Matlab) accept only piecewise constant gain specifications.

- *Least squares* — optimal in the weighted mean-square norm, uses closed-form expressions for the optimal filter coefficients. As in the equiripple method, the weight for each band is set to be inversely proportional to the band tolerance. Because of the availability of a simple closed-form solution, the least-squares method is most frequently used to design filters with arbitrarily shaped magnitude responses.
- *Truncation and windowing* — optimal in the time-domain mean-square norm (when $L = M$) but not in any frequency-domain sense. The resulting filter usually has many more coefficients than the one designed by the equiripple method. The filter coefficients are obtained by (1) applying the inverse DTFT (5b) to the desired response $D(\omega)$ (which contains the appropriate linear phase term) to obtain the impulse response $d(n)$, and (2) multiplying this impulse response by a *window function* $w(n)$, which vanishes outside the range $0 \leq n \leq M$. The window function is selected to control the tradeoff between the passband attenuation of the filter, and the width of the transition band. The Kaiser window has a control parameter that allows continuous adjustment of this tradeoff. Other popular windows (e.g., Bartlett, Hamming, Von Hann, and Blackman) are not adjustable and offer a fixed tradeoff. The simplicity of the truncation and windowing method makes it particularly attractive in applications that require an arbitrarily shaped magnitude response (and possibly also an arbitrarily shaped phase response).

Specialized design methods are used to design nonstandard filters such as maximally flat or minimum-phase FIR filters, Nyquist filters, differentiators, Hilbert transformers, and notch filters [5,6].

4.3. IIR Filter Design

The transfer function of an IIR digital filter is given by (6) with $N > 0$, so that $H(z) = b(z)/a(z)$. The design problem is to determine the values of the coefficients $\{a_i\}$ and $\{b_i\}$ so that the magnitude response $|H(e^{j\omega})|$ approximates a specified $|D(\omega)|$, with no additional constraints on the phase response. The design of an IIR filter is successfully completed when we have selected values for the filter coefficients such that the magnitude tolerance constraints (13) are satisfied. A design is considered

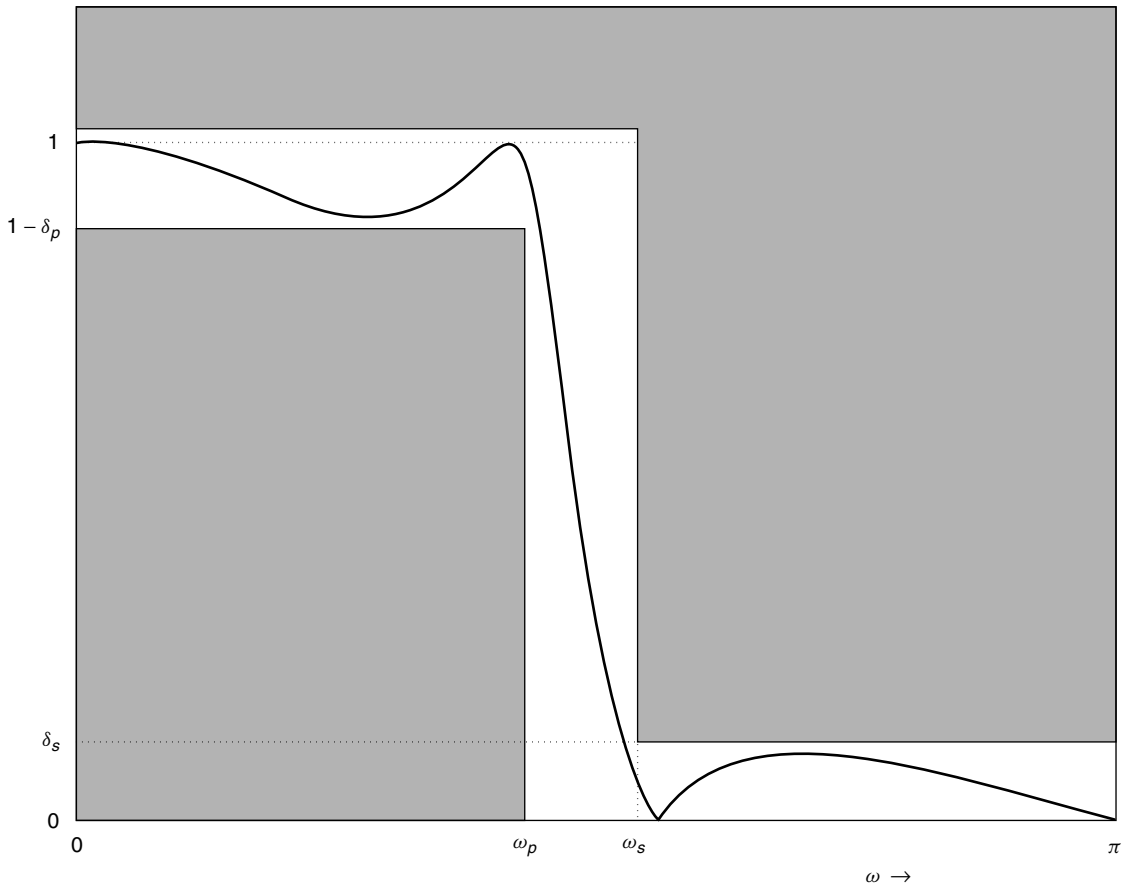


Figure 12. Specification for IIR lowpass filter design.

optimal when the filter order N is as small as possible under the specified constraints.

There exist a number of techniques for the design of IIR filters with an arbitrarily shaped $D(\omega)$, some based on the weighted (frequency-domain) $\mathcal{L}^2$ and $\mathcal{L}^\infty$ norms [5], and others on the truncated time-domain mean-square norm (with $L = M + N + 1$) [6]. Here we shall discuss only the design of classical (ideal) frequency-selective prototypes, which is the most common application for IIR digital filters. The most popular IIR design technique relies on a well-developed theory of analog filter design from closed-form (analytical) formulas. This means that the design of an IIR digital filter decomposes into three steps:

1. The specification of the desired classical frequency selective digital filter (lowpass, highpass, bandpass, or bandstop) is translated into a specification of an equivalent analog filter.
2. An analog filter $H_a(s)$ that meets the translated specification is obtained by (a) designing a lowpass analog filter $H_{\mathrm{LP}}(s)$ from closed-form formulas and (b) transforming $H_{\mathrm{LP}}(s)$ into a highpass, bandpass, or bandstop filter, as needed, by a complex variable mapping $s \rightarrow g(s)$, namely, $H_a(s) = H_{\mathrm{LP}}[g(s)]$.
3. The analog filter $H_a(s)$ is transformed into a digital filter $H(z)$ by (another) complex variable mapping.

The analog filter $H_{LP}(s)$ is one of four standard prototypes: Butterworth, Chebyshev type 1 or 2, or elliptic (also known as Chebyshev–Cauer). Since the magnitude response of such prototypes is always bounded by unity, the specification template has to be modified accordingly (see Fig. 12). As a result, the decibel scale characterization of tolerances for IIR design is also modified:

$$R_p = 20 \log \frac{1}{1-\delta_p}, \qquad A_s = 20 \log \frac{1}{\delta_s}$$

The variable transformation $s \rightarrow g(s)$ that converts the analog lowpass filter $H_{\mathrm{LP}}(s)$ into a highpass, bandpass, or bandstop filter $H_a(s)$ is described in Table 1. It involves $\Omega_{p,\mathrm{LP}}$, the passband edge frequency of the prototype $H_{\mathrm{LP}}(s)$,

Table 1. Frequency Transformations for Analog Filters

Type of Transformation	$g(s)$	Band-Edge Frequency of Target Filter
Lowpass to highpass	$\dfrac{\Omega_{p,\mathrm{LP}}\Omega_p}{s}$	Ω_p
Lowpass to bandpass	$\Omega_{p,\mathrm{LP}}\dfrac{s^2+\Omega_{p1}\Omega_{p2}}{s(\Omega_{p2}-\Omega_{p1})}$	Ω_{p1}, Ω_{p2}
Lowpass to bandstop	$\Omega_{p,\mathrm{LP}}\dfrac{s(\Omega_{p2}-\Omega_{p1})}{s^2+\Omega_{p1}\Omega_{p2}}$	Ω_{p1}, Ω_{p2}

as well as the passband edge frequencies of the target analog filter $H_a(s)$.

Several methods — including the impulse invariance, matched $\mathcal{Z}$, and bilinear transforms — can be used to map the analog transfer function $H_a(s)$, obtained in the second step of IIR design, into a digital transfer function $H(z)$. The most popular of these is the bilinear transform: the transfer function $H(z)$ is obtained from $H_a(s)$ by a variable substitution

$$H(z) = H_a\left(\alpha \frac{1 - z^{-1}}{1 + z^{-1}}\right) \tag{15a}$$

where α is an arbitrary constant. The resulting digital filter has the same degree N as the analog prototype $H_a(s)$ from which it is obtained. It also has the same frequency response, but with a warped frequency scale; as a result of the variable mapping (15a), we have

$$H(e^{j\omega}) = H_a(j\Omega)|_{\Omega=\alpha \tan(\omega/2)} \tag{15b}$$

Here we use Ω to denote the frequency scale of the analog filter $H_a(s)$ in order to distinguish it from the frequency scale ω of the digital filter $H(z)$. The frequency mapping $\Omega = \alpha \tan(\omega/2)$ governs the first step of the IIR design process — the translation of a given digital filter specification into an equivalent analog filter specification. While the tolerances R_p and A_s are left unaltered, all critical digital frequencies (such as ω_p and ω_s for a lowpass filter) are *prewarped* into the corresponding analog frequencies, using the relation

$$\Omega = \alpha \tan \frac{\omega}{2} \equiv \alpha \tan \pi f \tag{15c}$$

For instance, in order to design a lowpass digital filter with passband $|f| \le f_p = 0.2$ with ripple $R_p = 1$ dB and stopband $0.3 = f_s \le |f| \le 0.5$ with attenuation $A_s = 40$ dB, we need to design an analog filter with the same tolerances, but with passband $|\Omega| \le \Omega_p$ and stopband $|\Omega| \ge \Omega_s$, where

$$\Omega_p = \alpha \tan(0.2\pi), \qquad \Omega_s = \alpha \tan(0.3\pi)$$

Since the value of α is immaterial, the most common choice is $\alpha = 1$ (sometimes $\alpha = 2$ is used). The value of α used in the prewarping step must also be used in the last step when transforming the designed analog transfer function $H_a(s)$ into its digital equivalent $H(z)$ according to (15a).

Once the complete specification of the desired analog filter is available, we must select one of the four standard lowpass prototypes (Fig. 13):

- *Butterworth* — has a monotone decreasing passband and stopband magnitude response, which is maximally flat at $\Omega = 0$ and $\Omega = \infty$
- *Chebyshev 1* — has an equiripple passband magnitude response, and a monotone decreasing stopband response, which is maximally flat at $\Omega = \infty$
- *Chebyshev 2* — has a monotone decreasing passband magnitude response, which is maximally flat at $\Omega = 0$, and an equiripple stopband response
- *Elliptic* — has an equiripple magnitude response in both the passband and the stopband

All four prototypes have a maximum gain of unity, which is achieved either at the single frequency $\Omega = 0$ (for Butterworth and Chebyshev 2), or at several frequencies throughout the passband (for Chebyshev 1 and elliptic). When the Butterworth or Chebyshev 1 prototype is translated into its digital form, its numerator is proportional to $(1 + z^{-1})^N$. Thus only $N + 1$ multiplications are required to implement these two prototypes, in contrast to Chebyshev 2 and elliptic prototypes, which have nontrivial symmetric numerator polynomials, and thus require $N + \text{ceil}(N + 1/2)$ multiplications each.

As an example we present in Fig. 13 the magnitude response of the four distinct digital lowpass filters, designed from each of the standard analog prototypes to meet the specification $f_p = 0.23$, $f_s = 0.27$, $R_p = 1$ dB, and $A_s = 20$ dB. The order needed to meet this specification is 12 for the Butterworth, 5 for Chebyshev (both types), and 4 for the elliptic. Thus the corresponding implementation cost (=number of multipliers) is in this case 13 for Butterworth, 6 for Chebyshev 1, 11 for Chebyshev 2, and 9 for elliptic.

5. REALIZATION AND IMPLEMENTATION OF DIGITAL FILTERS

The preceding discussion of digital filters has concentrated solely on their input–output response. However, in order to build a digital filter, we must now direct our attention to the internal structure of the filter and its basic building blocks. The construction of a digital filter from a given (realizable) transfer function $H(z)$ usually consists of two stages:

- *Realization* — in this stage we construct a block diagram of the filter as a network of interconnected basic building blocks, such as unit delays, adders, and signal scalers (i.e., multipliers).
- *Implementation*–in this stage we map the filter realization into a specific hardware/software architecture, such as a general-purpose computer, a digital signal processing (DSP) chip, a field-programmable gate array (FPGA), or an application-specific integrated circuit (ASIC).

A realization provides an idealized characterization of the internal structure of a digital filter, in the sense that it ignores details such as number representation and timing. A given transfer function has infinitely many realizations, which differ in their performance, as explained in Section 5.3. The main distinguishing performance attributes are numerical accuracy and processing delay. These and other details must be addressed as part of the implementation stage, before we can actually put together a digital filter.

5.1. Realization of FIR Filters

FIR filters are always realized in the so-called transversal (also tapped-delay-line) form, which is a special case of the direct-form realization described in Section 3.2. Since

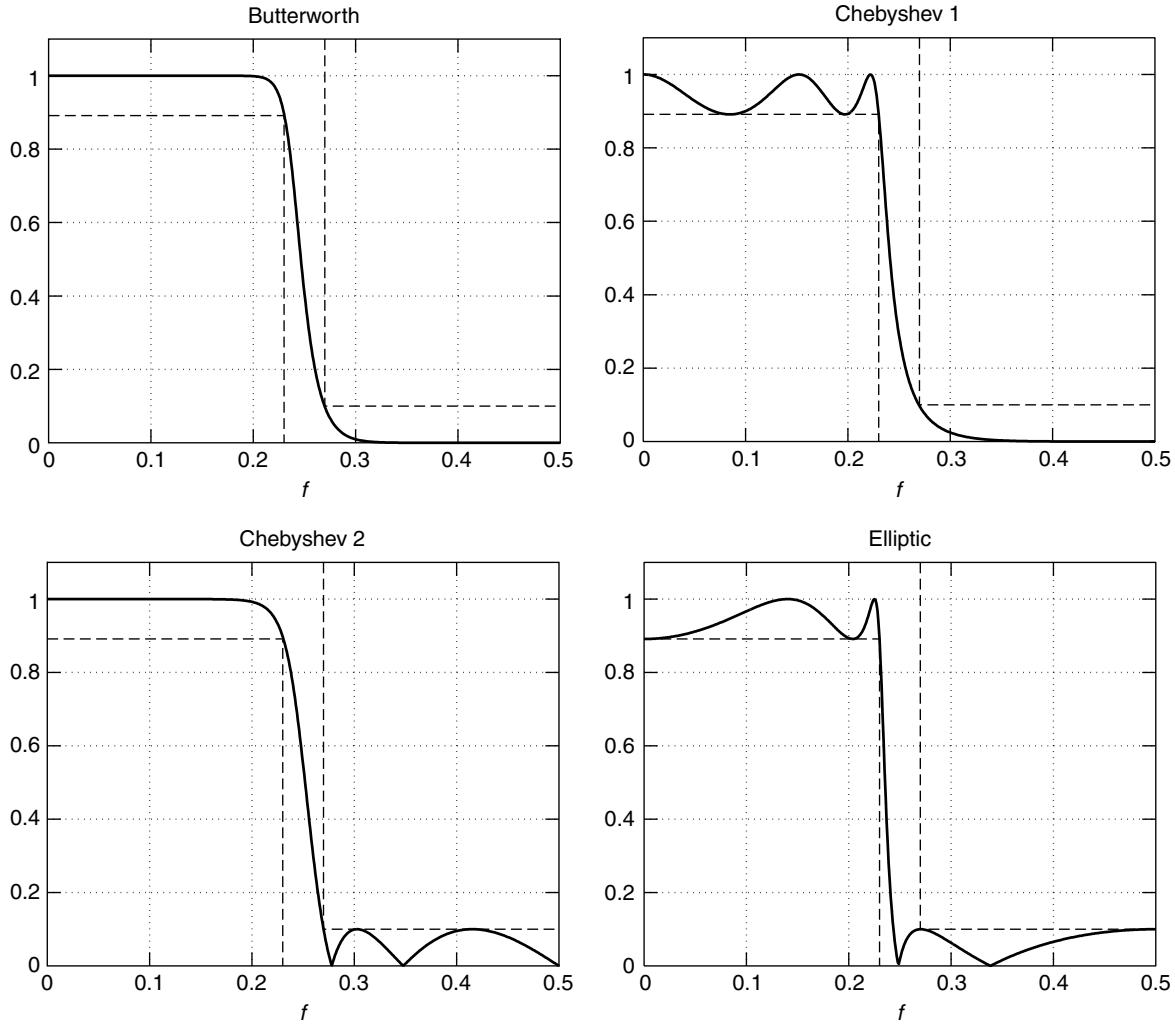


Figure 13. The four standard lowpass prototypes for IIR filter design.

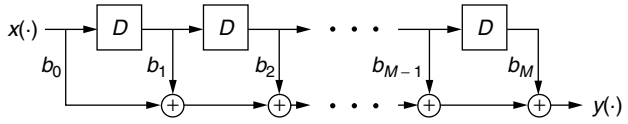


Figure 14. Transversal (direct-form) realization of an FIR digital filter.

$a(z) = 1$ for FIR filters, the direct form realization reduces to the configuration shown in Fig. 14.

The realization requires M delay elements, $M + 1$, multipliers and M adders. However, since FIR filters usually satisfy the symmetry constraint $b_i = \pm b_{M-i}$ in order to achieve linear phase, we can save about half of the number of multipliers.

5.2. Realization of IIR Filters

The transfer function of most practical IIR filters, such as those obtained from analog prototypes, have the same numerator and denominator degrees: $N \equiv \deg a(z) = \deg b(z) \equiv M$. Consequently, their direct-form realization requires N delay elements, $2N + 1$ multipliers, and N adders (see Fig. 10). We now describe two alternative realizations that require the same number of delays, multipliers, and adders as the direct-form realization.

The *parallel realization* of a rational transfer function $H(z)$ is obtained by using its partial fraction expansion

$$H(z) = A_0 + \sum_{j=1}^{N} \frac{A_j}{1 - p_j z^{-1}} \tag{16}$$

where p_j are the poles of $H(z)$ and where we assumed that (1) $M \leq N$ and (2) all poles are simple. Both assumptions hold for most practical filters, and in particular for those obtained from the standard analog prototypes.

Since the denominator polynomial $a(z)$ has real coefficients, its roots (i.e., the poles p_j) are either real or conjugate pairs. In order to avoid complex filter coefficients, the two terms representing a conjugate pair of poles are combined into a single second-order term:

$$\frac{A}{1 - p_0 z^{-1}} + \frac{A^*}{1 - p_0^* z^{-1}} = 2\frac{(\mathrm{Re}A) - (\mathrm{Re}p_0)z^{-1}}{1 - (2\mathrm{Re}p_0)z^{-1} + |p_0|^2 z^{-2}}$$

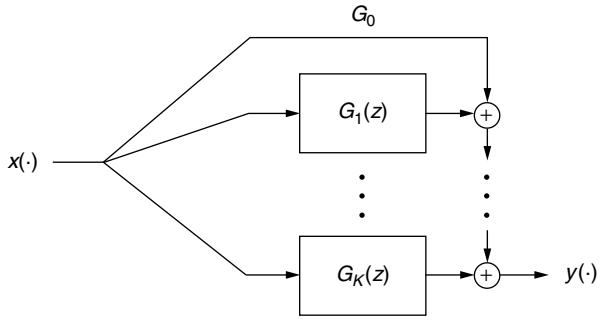


Figure 15. Parallel realization of an IIR digital filter.

Thus we obtain an additive decomposition

$$H(z) = A_0 + \sum_{k=1}^{K} G_k(z)$$

where each $G_k(z)$ is a strictly proper IIR filter or order 2 (for conjugate pairs of poles) or order 1 (for real poles). This results in a parallel connection of the subsystems $G_k(z)$ (Fig. 15). Finally, each individual $G_k(z)$ is realized in direct form.

The *cascade realization* is obtained by using the pole–zero factorization of the transfer function

$$H(z) \equiv \frac{b(z)}{a(z)} = b_0 \frac{\prod_{i=1}^{M} (1 - z_i z^{-1})}{\prod_{j=1}^{N} (1 - p_j z^{-1})} \tag{17}$$

where we recall again that the *zeros* z_i are the roots of the numerator polynomial $b(z)$ and the *poles* p_j are the roots of the denominator polynomial $a(z)$. Since these polynomials have real coefficients, their roots are either real or conjugate pairs. For each conjugate pair, the corresponding first-order factors are combined into a single second-order term:

$$(1 - z_0 z^{-1})(1 - z_0^* z^{-1}) = 1 - (2\mathrm{Re} z_0) z^{-1} + |z_0|^2 z^{-2}$$

Thus we obtain a multiplicative factorization

$$H(z) = \prod_{k=1}^{K} H_k(z) \tag{18}$$

where each $H_k(z)$ is a proper IIR filter or order 2 (for conjugate pairs of poles) or order 1 (for real poles). This results in a cascade connection of the subsystems $H_k(z)$ (Fig. 16). Again, each individual $H_k(z)$ is realized in direct form.

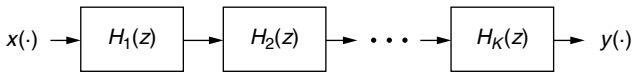


Figure 16. Cascade realization of an IIR digital filter.

There are multiple ways to form the factors $H_k(z)$ in (18), all requiring the same number of delays, multipliers, and adders as the direct-form realization. There are, however, differences in performance between these alternative factorizations, as explained in Section 5.3.

A unified algebraic framework for all possible realizations of a given finite-order transfer function $H(z)$ is provided by the theory of *factored state variable descriptions* (FSVD) [7]. The FSVD is a refinement of the well-known *state-space* description, which describes the relations between the input signal $x(\cdot)$, output signal $y(\cdot)$, and a state vector $s(n)$:

$$\begin{pmatrix} \mathbf{s}(n+1) \\ y(n) \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \mathbf{s}(n) \\ x(n) \end{pmatrix} \tag{19a}$$

The FSVD refines this characterization by specifying a multiplicative factorization

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \mathcal{F}_K \mathcal{F}_{K-1} \cdots \mathcal{F}_2 \mathcal{F}_1 \tag{19b}$$

that captures the decomposition of the realization into interconnected modules, as in the cascade or parallel realizations. Each $\mathcal{F}_k$ describes a single module, and each module is associated with a subset of the state vector $\mathbf{s}(n)$. The totality of all possible minimal realizations of a given $H(z)$ is captured by the notion of similarity transformations, combined with the added flexibility of factoring the (A, B, C, D) matrix as in (19b).

5.3. Implementation and Performance Attributes

The next stage in the design of a digital filter is *implementation*, that is, the mapping of a selected realization into a specific hardware/software architecture. Since a given transfer function has multiple implementations, which differ in their performance, the choice of a specific implementation should be based on objective performance criteria. Commonly used performance attributes for evaluating implementations include

- *Processing speed* — can be quantified in terms of two distinct attributes: (1) *throughput*, which is the number of input samples that can be processed by the digital filter in a unit of time, and (2) *input–output delay*, which is the duration between the instant a given input sample $x(n_o)$ is applied to the digital filter, and the instant the corresponding output sample $y(n_o)$ becomes available.
- *Numerical accuracy* — the level of error due to finite precision number representation.
- *Hardware cost* — the total number of each kind of basic building blocks, such as delays, multipliers, and adders.
- *Implementation effort* – the amount of work required to map a given realization into a specific hardware/software architecture. Structural attributes of the realization, such as modularity and/or regularity, contribute to the reduction of this effort.

The same attributes can be also used to evaluate the performance of realizations [6,7]. For instance, the direct realization of an IIR digital filter has much poorer numerical accuracy than does either the parallel or the cascade realizations. Similarly, the input–output delay of the parallel realization is somewhat shorter than that of the cascade or direct realizations. Such observations make it possible to optimize the selection of a realization, and to quantify the relative merits of alternative implementations.

6. OPTIMAL AND ADAPTIVE DIGITAL FILTERS

The desired response of classical frequency-selective filters is completely specified in terms of passbands, stopbands, and transition bands. In contrast, optimal filters use detailed information about the frequency content of the desired signal and the interference in achieving maximal suppression of the latter, along with minimal distortion of the former. Since the information needed to construct an optimal filter is typically not available a priori, it is usually estimated online from measurements of the available signals. When this process of frequency content estimation is carried out simultaneously with signal filtering, and the coefficients of the optimal filter are continuously updated to reflect the effect of new information, the resulting linear time-variant configuration is known as an *adaptive filter*.

Fixed frequency-selective filters are appropriate in applications where (1) the desired signal is restricted to known frequency bands (e.g., AM and FM radio, television, frequency-division multiplexing) or (2) the interfering signal is restricted to known frequency bands (e.g., stationary background in Doppler radar, fixed harmonic interference). On the other hand, in numerous applications the interference and the desired signal share the same frequency range, and thus cannot be separated by frequency selective filtering. Adaptive filters can successfully suppress interference in many such situations (see Section 7.2).

6.1. Optimal Digital Filters

Optimal filtering is traditionally formulated in a probabilistic setting, assuming that all signals of interest are *random*, and that their joint statistics are available. In particular, given the second-order moments of two discrete-time (zero-mean) random signals $x(\cdot)$ and $y(\cdot)$, the corresponding mean-square optimal filter, also known as a *Wiener filter*, is defined as the (unique) solution $h_{\text{opt}}(\cdot)$ of the quadratic minimization problem

$$\min_{h(\cdot)} E|y(n) - h(n) \circledast x(n)|^2 \tag{20}$$

where E denotes expectation. If $x(\cdot)$ and $y(\cdot)$ are jointly stationary, the Wiener filter $h_{\text{opt}}(\cdot)$ is time-invariant; otherwise it is a linear time-variant (LTV) filter. The Wiener filter can be applied in two distinct scenarios:

- *Linear mean-square estimation* — an unknown random signal $y(\cdot)$ can be estimated from observations of another random signal $x(\cdot)$. The corresponding estimate $\hat{y}(\cdot)$ is obtained by applying the observed signal $x(\cdot)$ to the input of the Wiener filter, so that $\hat{y}(\cdot) = h_{\text{opt}}(\cdot) \circledast x(\cdot)$.
- *System identification* — the impulse response of an unknown relaxed LTI system can be identified from observations of its input signal $x(\cdot)$ and output signal $y(\cdot)$. In fact, the optimal solution of the Wiener problem (20) coincides with the unknown system response, provided (1) the input signal $x(\cdot)$ is stationary, and is observed with no error and (2) the (additive) error in measuring the output signal $y(\cdot)$ is uncorrelated with $x(\cdot)$.

The unconstrained optimal solution of the Wiener problem (20) is a noncausal IIR filter, which is not realizable (see Section 3.2 for a discussion of realizability). In the case of jointly stationary $x(\cdot)$ and $y(\cdot)$ signals, realizable solutions are obtained by imposing additional constraints on the impulse response $h(\cdot)$ and/or the joint statistics of $x(\cdot), y(\cdot)$:

- When both the autospectrum $S_{xx}(z)$ and the cross-spectrum $S_{yx}(z)$ are rational functions, and the Wiener filter is required to be causal, the resulting Wiener filter is realizable; specifically, it is causal and rational.
- When $x(\cdot)$ and $y(\cdot)$ are characterized jointly by a state-space model, the resulting optimal filter can be described by a state-space model of the same order. This is the celebrated *Kalman filter*.
- Realizability can be enforced regardless of structural assumptions on the joint statistics of $x(\cdot)$ and $y(\cdot)$. In particular, we may constrain the impulse response in (20) to have finite length. This approach is common in adaptive filtering.

The classical Wiener filter is time-invariant, requiring an infinitely long prior record of $x(\cdot)$ samples, and thus is optimal only in steady state. In contrast, the Kalman filter is optimal at every time instant $n \geq 0$, requiring only the finite (but growing) signal record $\{x(k); 0 \leq k \leq n\}$. Consequently, the impulse response of the Kalman filter is time-variant and its length grows with n, asymptotically converging to the classical Wiener filter response. Finally, adaptive filters continuously adjust their estimates of the joint signal statistics, so the resulting filter response remains time-variant even in steady state, randomly fluctuating around the Wiener filter response.

6.2. Adaptive Digital Filters

In practice, the statistics used to construct an optimal filter must also be estimated from observed signal samples. Thus, the construction of an optimal filter consists of two stages:

- *Training Stage*. Finite-length records of the signals $x(\cdot)$ and $y(\cdot)$ are used to estimate the joint (first- and) second-order moments of these signals.

Subsequently, the estimated moments are used to construct a realizable optimal filter.

- *Filtering Stage*. The fixed optimal filter that was designed in the training stage is now applied to new samples of $x(\cdot)$ to produce the estimate $\hat{y}(\cdot)$.

Alternatively, we may opt to continue the training stage indefinitely; as new samples of $x(\cdot)$ and $y(\cdot)$ become available, the estimated statistics and the corresponding optimal filter coefficients are continuously updated. This approach gives rise to an adaptive filter configuration — a linear time-variant digital filter whose coefficients are adjusted according to continuously updated moment estimates (Fig. 17).

Since adaptive filtering requires ongoing measurement of both $x(\cdot)$ and $y(\cdot)$, it can be applied only in applications that involve the system identification interpretation of the Wiener filter. Thus, instead of applying $h_{\text{opt}}(\cdot)$ to $x(\cdot)$ to estimate an unknown signal $y(\cdot)$, adaptive filters rely on the explicit knowledge of both $x(\cdot)$ and $y(\cdot)$ to determine $h_{\text{opt}}(\cdot)$. Once this optimal filter response is available, it is used to suppress interference and extract desired signal components from $x(\cdot)$, $y(\cdot)$ (see Section 7.2).

The most common adaptive filters use a time-variant FIR configuration with continuously adjusting filter coefficients. Such filters take full advantage of the power of digital signal processing hardware because (1) the ongoing computation of signal statistics and optimal filter coefficients usually involves nonlinear operations that would be difficult to implement in analog hardware, and (2) modifications to the updating algorithms can be easily implemented by reprogramming. The most commonly used algorithms for updating the filter coefficients belong to the *least-mean-squares* (LMS) family and the *recursive least-squares* (RLS) family [3].

7. APPLICATIONS IN TELECOMMUNICATION SYSTEMS

Since the early 1980s, digital filtering has become a key component in a wide range in applications. We provide here a brief summary of the main telecommunication applications of digital filters, separated into two categories: (1) *time-invariant digital filters*, which are used to replace analog filters in previously known applications, and (2) *adaptive digital filters*, which enable new applications that were impossible to implement with analog hardware. A detailed discussion of numerous digital filtering applications can be found in the literature [1,4–6].

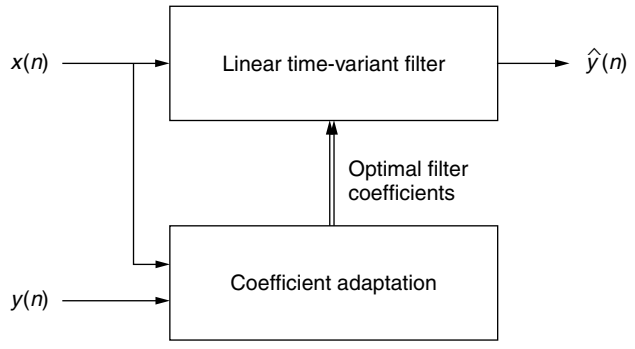


Figure 17. Adaptive filter configuration.

7.1. Time-Invariant Digital Filters

Ongoing improvements in the cost, size, speed, and power consumption of digital filters have made them an attractive alternative to analog filters in a variety of telecommunication applications that require a time-invariant filter response. The main types of such applications are:

- *Frequency-selective filters* — pass (or stop) prespecified frequency bands. Some examples include
  - RF-to-IF-to-baseband conversion in digital wireless systems
  - FDM multiplexing and demultiplexing
  - Subband decomposition for speech and image coding (e.g., MPEG)
  - Doppler radar moving-target indicator (MTI) to remove the nonmoving background
  - Digital spectrum analysis
- *Matched filters/correlators* — have an impulse response that matches a given waveform. Some examples include:
  - Matched filter for symbol detection in digital communication systems
  - Waveform-based multiplexing/demultiplexing (e.g., CDMA)
  - Front-end Doppler radar processing
- *Analog-to-digital and digital-to-analog conversion*
  - Oversampling converters (e.g., sigma–delta modulation)
  - Compact-disk recording and playing
- *Specialized filters* (such as)
  - Hilbert transformers
  - Timing recovery in digital communications (e.g., early–late gate synchronizer)

7.2. Adaptive Digital Filters

As explained in Section 6.2, adaptive filters implement the system identification scenario of the Wiener filter (Fig. 17), requiring two observed signals. Many applications use a variation of this configuration, known as the *adaptive interference canceler*, in which the received signal $y(\cdot)$ is a noisy version of an unknown desired signal $s(\cdot)$, and the so-called "reference signal" $x(\cdot)$ is a filtered version of the interference component of $y(\cdot)$ (Fig. 18).

Two assumptions must be satisfied in order for the adaptive filter to perform well as an interference canceler:

- *Lack of correlation* — the reference signal $x(\cdot)$ should be uncorrelated with the desired signal component of the received signal $y(\cdot)$.
- *FIR response* — the unknown system $H(z)$ should be FIR, and of a known order.

If these two assumptions are met, the estimated $\hat{H}(z)$ matches the true $H(z)$ perfectly, and the effect of

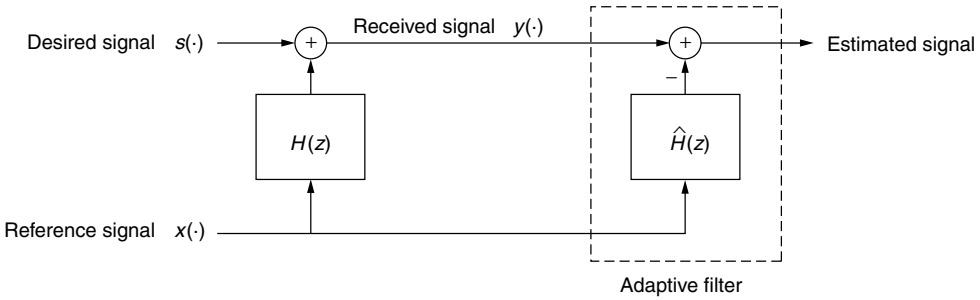


Figure 18. Adaptive interference canceller configuration.

interference is *completely cancelled*. In reality, both assumptions are met only approximately, so that only partial cancellation of the interference is achieved.

The interference cancelling configuration is useful in a variety of specific telecommunication applications, including

- *Echo cancellation* — used in duplex communication systems to suppress interference caused by signal leakage from the incoming far-end signal (the "reference signal" in Fig. 18) to the outgoing local signal (the "desired signal"). Such leakage is common in telephone lines (2W/4W converters), modems, teleconferencing systems and hands-off telephone units.
- *Decision feedback equalization* — used to reduce the effect of intersymbol interference (ISI) in a digital communication channel. Here the "reference signal" is the sequence of previously received symbols, which are assumed to be uncorrelated with the current symbol (the "desired signal"). In order for the equalizer to work correctly, the previously received symbols must be known without error. To meet this requirement, the equalizer alternates between a training phase and a tracking phase. In the training phase a prespecified "training sequence" of symbols is transmitted through the channel, and this information is used by the equalizer to determine the channel estimate $\hat{H}(z)$. In the tracking phase the equalizer uses detected previous symbols to track slow variations in the channel response: as long as the estimated response $\hat{H}(z)$ remains close to the true response $H(z)$, the cancellation of ISI is almost perfect, symbols are detected without error, and the correct "reference signal" is available to the equalizer. Since decision feedback equalization relies on previous symbols, it cannot reduce the ISI caused by the precursors of future symbols — this task is left to the feedforward equalizer.
- *Sidelobe cancellation* — used to modify the radiation pattern of a narrowbeam directional antenna (or antenna array), in order to reduce the effect of strong nearby interferers received through the sidelobes of the antenna. The "reference signal" in this case is obtained by adding an inexpensive omnidirectional antenna; if the interfering RF source is much closer than the source at which the main antenna is pointed, then the reference signal received by the omnidirectional antenna is dominated by the interference, and the lack-of-correlation assumption is (approximately) met.

The *adaptive beamformer* configuration, a variation of the sidelobe-canceling approach, is used to maintain the mainlobe of an antenna array pointed in a predetermined direction, while adaptively reducing the effect of undesired signals received through the sidelobes [3]. It can be used, for instance, to split the radiation pattern of a cellular base-station antenna into several narrow beams, each one tracking a different user.

Another common adaptive filtering configuration is the *adaptive linear predictor*, in which only one observed signal is available (Fig. 19). In this case the "reference signal" is simply a delayed version of the observed signal $y(\cdot)$, so that the Wiener problem (20) now becomes

$$\min_{h(\cdot)} E|y(n) - h(n) \circledast y(n-\Delta)|^2 \tag{21}$$

which is, a Δ-step-ahead linear prediction problem. In some applications the object of interest is the adaptively estimated linear predictor response $\hat{h}(\cdot)$, while in other applications it is the *predicted signal* $\hat{y}(n) = \hat{h}(n) \circledast y(n-\Delta)$, or the *residual signal* $y(n) - \hat{y}(n)$. Telecommunication applications of the linear predictor configuration include

- *Linear predictive coding* (LPC) — used for low-rate analog-to-digital conversion, and has become the standard speech coding method in GSM cellular systems. While the original LPC approach used only the linear predictor response $\hat{h}(\cdot)$ to represent the observed speech signal $y(\cdot)$, current LPC-based speech coders use, in addition, a compressed form of the residual signal $y(n) - \hat{y}(n)$ [6].
- *Adaptive notch filter* — used to suppress sinusoidal interference, such as narrowband jamming in spread-spectrum systems. Since spread-spectrum signals resemble broadband noise, which is almost completely unpredictable, while sinusoidal signals are

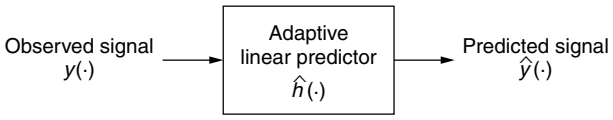


Figure 19. Adaptive linear predictor configuration.

perfectly predictable, the adaptive linear predictor response is almost entirely dominated by the sinusoidal component of the observed signal $y(\cdot)$. As a result, one can use the adaptive linear predictor configuration to suppress sinusoidal components or, with a minor modification, to enhance sinusoidal components.

- *Feedforward equalizer* — used to reduce the effect of ISI from precursors of future symbols. This is made possible by the fact that the adaptive linear predictor tends to render the equalized channel response minimum phase, thereby reducing the length of precursors. This minimum-phase property is an intrinsic characteristic of the MSE linear predictor defined by (21).

BIOGRAPHY

Hanoch Lev-Ari received the B.S. (summa cum laude) and the M.S. degrees in electrical engineering from the Technion, Israel Institute of Technology, Haifa, Israel in 1971 and 1976, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, California, in 1984. During 1985 he held a joint appointment as an Adjunct Research Professor of Electrical Engineering with the Naval Postgraduate School, Monterey, California and as a Research Associate with the Information Systems Laboratory at Stanford. He stayed at Stanford as a Senior Research Associate until 1990. Since 1990 he has been an Associate Professor with the Department of Electrical and Computer Engineering at Northeastern University, Boston, Massachusetts. During 1994–1996 he was also the Director of the Communications and Digital Signal Processing (CDSP) Center at Northeastern University. Dr. Lev-Ari has over 100 journal and conference publications. He served as an Associate Editor of *Circuits, Systems and Signal Processing*, and of *Integration, the VLSI Journal*. His areas of interest include model-based spectrum analysis and estimation for nonstationary signals, scale-recursive (multirate) detection and estimation of random signals, Markov renewal models for nonstationary signals, and adaptive linear and nonlinear filtering techniques.

BIBLIOGRAPHY

1. M. E. Frerking, *Digital Signal Processing in Communication Systems*, Van Nostrand Reinhold, New York, 1994.
2. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, 1985.
3. S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
4. V. K. Madisetti and D. B. Williams, eds., *The Digital Signal Processing Handbook*, CRC Press, 1998.
5. S. K. Mitra and J. F. Kaiser, eds., *Handbook for Digital Signal Processing*, Wiley, New York, 1993.
6. B. Porat, *A Course in Digital Signal Processing*, Wiley, New York, 1997.
7. R. Roberts and T. Mullis, *Digital Signal Processing*, Addison-Wesley, 1987.

DIGITAL OPTICAL CORRELATION FOR FIBEROPTIC COMMUNICATION SYSTEMS

JOHN E. MCGEEHAN
MICHELLE C. HAUER
ALAN E. WILLNER
University of Southern California
Optical Communications Laboratory
Los Angeles, California

1. INTRODUCTION

As bit rates continue to rise in the optical core of telecommunication networks [toward $\geq$40 Gbps (gigabits per second)], the potential inability of electronic signal processors to handle this increase is a force driving research in optical signal processing. Correlation, or matched filtering, is an important signal processing function. The purpose of a correlator is to compare an incoming signal with one that is "stored" in the correlator. At the appropriate sample time, a maximum autocorrelation peak will be produced if the input signal is an exact match to the stored one. This function is typically used to pick a desired signal out of noise, an essential requirement for radar and wireless CDMA systems. As telecommunication systems tend toward the use of optical fiber as the preferred transmission medium, optical CDMA networks are receiving greater attention and will require the use optical correlator implementations.

In present-day fiberoptic networks, data packets are converted to electrical form at each node to process their headers and make routing decisions. As routing tables grow in size, this is becoming a predominant source of network latency. The advent of optical correlation could enable packet headers to be read at the speed of light by simply passing the packets through a bank of correlators, each configured to match a different entry in the routing table. Simple decision logic could then configure a switch to route each packet according to which correlator produced a match. Thus, with the growing demand for all-optical networking functions, there is a strong push to develop practical and inexpensive optical correlators that can identify high-speed digital bit patterns on the fly and with minimal electronic control.

2. DIGITAL OPTICAL CORRELATION

The term "correlator" is typically used to describe a hardware implementation of a matched filter. Although there exist strict definitions for matched filters versus correlators, most hardware implementations that produce the same output as a matched filter, sampled at the peak autocorrelation value, are referred to as "correlators." The detailed theory of matched filters and correlation is presented in most textbooks on communication systems [1]. The aim here is to give a brief overview of the concepts that apply to the correlation of digital binary waveforms modulated onto optical carriers. This presents a unique case in that the data bits in an optical communication system are most commonly represented by the optical

power of the signal as opposed to a voltage as in electrical systems. Consequently, the digital waveforms consist of only positive values (a "unipolar" system), causing the correlation function of any two optical signals to be a completely nonnegative function. This creates some limitations for systems that transmit specially designed sets of codewords intended to have good autocorrelation properties; thus, a large autocorrelation peak when the incoming signal is synchronized with the receiver and matches the desired codeword and a low, ideally zero, level for all other codewords, for all possible shifts in time [2]. Sets of codewords with these properties are termed *orthogonal codes*. It is possible with optical phase modulation to achieve a bipolar system, but this requires a coherent receiver, which is far more complex to implement in optics than the standard direct intensity detection receivers.

Just as with electronics, prior to the development of high-speed digital signal processors, a common implementation of an optical correlator is the tapped-delay line. A basic implementation of an optical tapped-delay line is shown in Fig. 1a. The delay line is configured to match the correlation sequence 1101. Thus, the delay line requires four taps (one for each bit in the desired sequence), weighted by the factors 1, 1, 0 and 1, respectively. The weights are implemented by placing a switch in each path that is closed for weight = 1, and opened for weight = 0. The incoming optical bit stream is equally split among the four fiberoptic delay lines. Each successive delay line adds one additional bit of delay to the incoming signal before the lines are recombined, where the powers of the four signals are added together to yield the correlation output function. This function is sampled at the optimum time, T_s, and passed through a threshold detector that is set to detect a power level above 2 since the autocorrelation peak of 1101 with itself equals 3 (or more specifically, 3 times the power in each 1 bit). This threshold detection may be implemented either electronically or optically, although optical threshold detection is still a nascent research area requiring further development to become practical. Electronically, the output is detected with a photoreceiver and a simple decision circuit is used to compare the correlation output to the threshold value. The high-speed advantage of optics still prevails in this case since the correlation function is produced in the time it takes the signal to traverse the optical correlator, and the decision circuit only needs to be triggered at the sample-rate, which is often in the kilohertz–megahertz range, depending on the number of bits in each data packet or codeword. The mathematical function describing the tapped-delay-line correlator is

$$y(t) = \sum_{k=0}^{N-1} x(t - kT_{\text{bit}})h(kT_{\text{bit}}) \tag{1}$$

where N is the number of bits in the correlation sequence, T_{bit} is one bit period, $x(t - kT_{\text{bit}})$ is the input signal delayed by k bit times, and $h(kT_{\text{bit}})$ represents the k weights that multiply each of the k-bit delayed input signals. For a phase-modulated system, the same operation can be performed by replacing the switches with the appropriate optical phase shifters to match the desired codeword (e.g., $\pi\pi 0\pi$ instead of 1101). Figure 1b illustrates the delay-and-add operation of the correlator for the case when the three 4-bit words 1010, 1011, and 1101 are input to the correlator, where the second word is an exact match to the desired sequence. Since the correlation function for two 4-bit words is 7 bits long and the peak occurs during the fourth time slot, the correlation output is sampled every 4 bits and compared to a threshold as shown in Fig. 1c. As expected, the correlation output for the second word exceeds the threshold while the first and third samples produce no

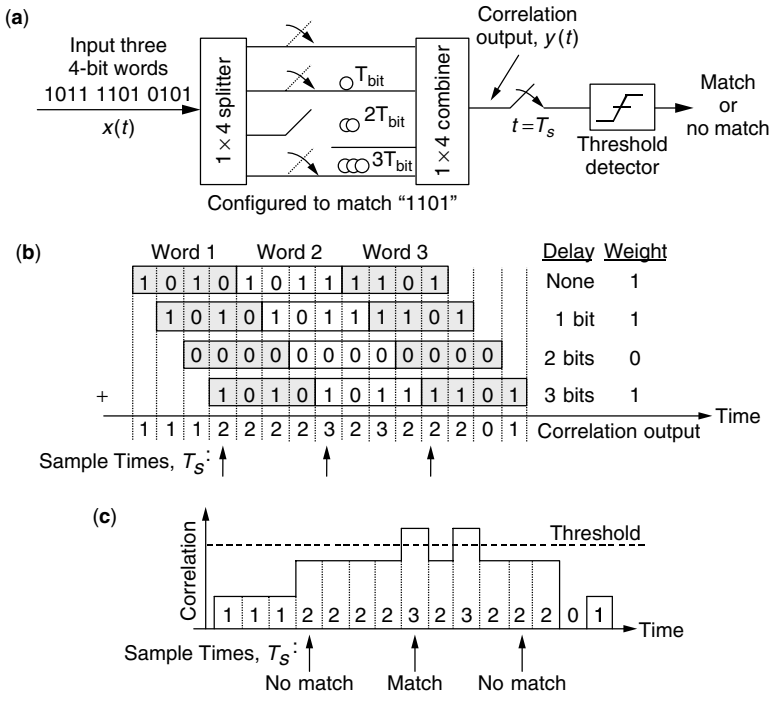


Figure 1. (**a**) A basic implementation of a fiberoptic tapped-delay-line correlator configured to produce an autocorrelation peak at the sample time, T_s, for the sequence 1101; (**b**) the weighted delay-and-add computation of the digital correlation output for three input words when correlated with the bit pattern 1101 (the three optimum sample times are labeled); (**c**) an intensity profile of the correlation output from this tapped-delay-line correlator. Only when the intensity is above the threshold at the sample time is a match signal produced.

matches. Note that for an input signal L bits long, the length of the correlation output will be $L + N - 1$ bits long.

It should also be noted that the correlator as shown will also produce a level 3 peak that is above the threshold at time T_s for a 1111 input, which is not the desired codeword. This is because the open switch in the third delay line, corresponding to the third correlation bit, does not care if the third bit is a 1 or a 0 since it does not pass any light. Thus, the correlator as shown is really configured to produce a match for the sequence $11x1$, where the x indicates a "don't care" bit that can either be 0 or 1. In many cases, such as in optical CDMA systems, the set of all possible codewords used in the system is specifically designed to always have a constant number of 1 bits so that this is not an issue. But, for cases in which the correlator must identify a completely unique sequence, the present correlator must be augmented with, for example, a second, parallel tapped-delay line that is configured in complement to the first one (the switches are closed for desired 0 bits and open otherwise), and produces a "match" signal when zero power is present at the sample time. Then, the incoming signal is uniquely identified only when both correlators produce a match. This configuration will be discussed later in further detail.

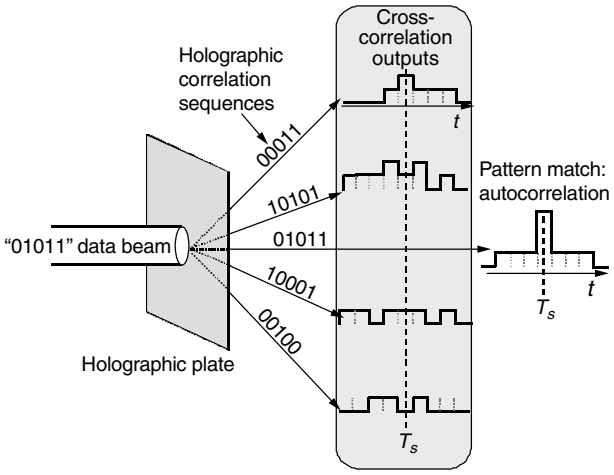


Figure 2. A collimated 01011 input beam in free space impinges on an optical angular multiplexed spectral holographic (AMSH) plate that is programmed to recognize five different correlation sequences. Each correlation sequence corresponds to a different scattering angle from the plate. Only the direction corresponding to a pattern match (01011) produces an intensity autocorrelation peak that will be above threshold at the sample time, while the rest produce cross-correlation outputs.

3. OPTICAL CORRELATOR IMPLEMENTATION

Digital optical correlators can be fabricated using free-space optics, fiber-based devices, and fiber-pigtailed crystals or semiconductors. Four varieties of optical correlators will be reviewed here, including (1) a free-space holographic correlator, (2) a fiberoptic correlator using separate fiber delay lines terminated with mirrors, (3) a single-fiber device with periodically spaced fiber Bragg gratings (FBGs) to provide time-delayed reflections, and (4) an optical phase-correlator implemented using FBGs. The first three of these correlators are explained assuming optical intensity-modulated correlation sequences. However, most can be modified to act as phase-coded correlators as well. No assumptions are made regarding any special data coding schemes.

One example of a free-space optical correlator employs spectral holography to create a correlating plate that is able to correlate an incoming bit stream with several correlation sequences simultaneously. The correlation sequences are encoded using an angular multiplexed spectral hologram (AMSH). The hologram acts upon the incoming modulated lightbeam to produce a correlation output [3]. The incoming bit pattern is spread out in the frequency domain using a grating, collimated by a lens and then sent through the AMSH plate, which is coded for a finite number of possible correlation patterns. The hologram is written using a 632.8-nm helium–neon laser and a photosensitive holographic plate and is designed to operate at signal wavelengths around 1550 nm (a standard fiberoptic communication wavelength). Each desired correlation sequence is written onto the holographic plate such that each pattern corresponds to a different diffraction angle. Figure 2 shows an input datastream of 01011 scattering off an AMSH plate. The correlation outputs corresponding to each correlation sequence are produced at different scattering angles at the output. As the number of correlation patterns increases, the difference between the output angles for each pattern decreases. At each diffraction angle, either an autocorrelation or cross-correlation output is produced, with the autocorrelation output appearing only at the diffraction angle corresponding to a matched correlation pattern, and cross-correlation outputs appearing at all other angles. A second grating can also be used to reroute any diffracted light back into an optical fiber for continued transmission. Otherwise, photodetectors may be placed at each diffraction angle to detect the correlation outputs. This free-space correlation method has a number of potential advantages over the fiber-based correlators that are detailed below, including the potential low size and cost associated with having one hologram that can correlate with many bit patterns simultaneously. However, further research is required on low-loss holographic materials at standard fiberoptic communication wavelengths, as the loss for the material used for this demonstration can exceed 80 dB due to high absorption at these frequencies.

To avoid the losses associated with exiting and reentering the fiber medium, a reconfigurable optical correlator that operates entirely within optical fibers is highly desirable for fiberoptic communication systems. In addition to the optical tapped-delay-line structure described in Fig. 1, a similar device may be constructed, except that each fiber branch is terminated with a fiber mirror (a metallic coating on the fiber end face), thereby providing a double pass through the delay lines and using the splitter also as a recombiner (see Fig. 3) [4]. The double pass requires that the relative delays be halved and an optical circulator must be added at the input to route the counterpropagating correlation output to the threshold detector. Aside from these differences, the operation of the correlator is identical to that described for Fig. 1. The

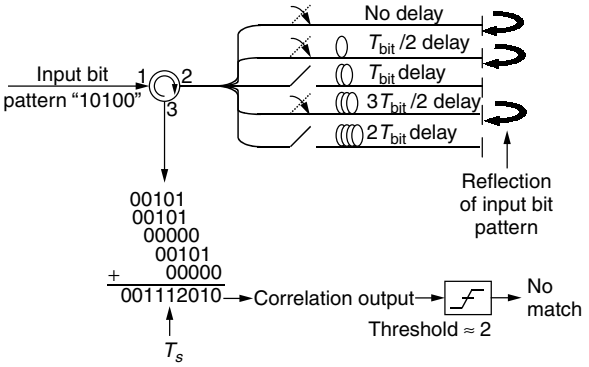


Figure 3. A fiberoptic delay line correlator in which each branch is terminated with a fiber mirror. The correlator is configured for the correlation sequence 11x1x, where the switches are closed to represent 1 bits and opened to indicate "don't care" bits, meaning that the peak of the output correlation function at time T_s will be the same regardless of whether the "don't care" bits are ones or zeros. The correlation output for an input pattern of 10100 is 001112010. The level 1 output at the sample time does not exceed the threshold and produces no match.

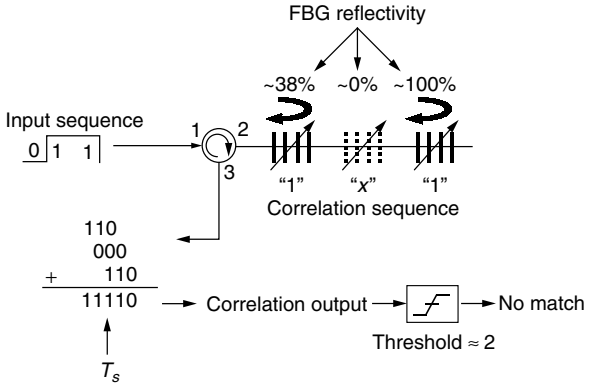


Figure 4. A fiber Bragg grating (FBG)–based optical correlator in which the correlation sequence is programmed by tuning the reflectivities of the gratings such that desired 1 bits reflect and "don't care" bits do not reflect. The cross-correlation output 11110 is the result of the correlation between the input sequence 011 and the programmed sequence 1x1. The output is sampled at time T_s and a threshold detector produces a "no match" decision.

fiber-mirror-based correlator shown in Fig. 3 is configured to recognize the correlation sequence 11010 (or more accurately, 11x1x, since the open switches represent "don't care" bits) by closing the first, second and fourth switches and setting the threshold just below a level 3. The figure shows the output cross-correlation function for the input sequence 10100, which will not produce a correlation peak above the threshold at the sample time.

A set of fiber Bragg gratings (FBGs) can also be used to construct an effective fiber-based optical correlator. An FBG is fabricated by creating a periodic variation in the fiber's index of refraction for a few millimeters to a centimeter of length along the fiber core [5]. Since optical fiber is photosensitive at ultraviolet frequencies, the grating can be written by illuminating the fiber from the side with the interference pattern of two ultraviolet laser beams. A conventional FBG acts as a reflective, wavelength-selective filter; that is, it reflects light at a particular wavelength that is determined by the spatial period of the index grating, and passes light at all other wavelengths. The bandwidth of the FBG filter depends largely on the magnitude of the index variation and is typically designed to be $<$100 GHz (0.8 nm) for communications applications. A nice feature of FBG filters is that the reflection spectrum can be adjusted by a few nanometers via heating or stretching of the grating. This allows one to alter the wavelength that the FBG reflects. The reflectivity of the grating is nearly 100% at the center of the reflected spectrum and falls off outside the grating bandwidth. By tuning the FBG so that the signal wavelength intersects with the rising or falling edge of the passband, the reflected energy at that wavelength will be reduced from 100% toward 0% as the grating's passband is tuned farther away.

Although it is possible to produce an optical correlator using FBGs to simply replace the fiber mirrors described above, the only advantage to this would be that the optical switches could be removed and the FBGs could simply be tuned *not* to reflect the incoming wavelength to represent the case of an open switch. However, the fact remains that this tapped-delay-line structure requires many separate branches of optical fiber, each precisely cut to provide delay differences as small as $\frac{1}{2}$ of a bit time, which is only 1 cm in fiber length at 10 Gbps. Using FBGs, a more compact, producible and manageable correlator can be produced [6]. The procedure for writing gratings into fibers makes it relatively simple to write several FBGs into a single fiber with precise control down to centimeter spacings (see Fig. 4). Using separate piezoelectric crystals or small heaters, each FBG can be independently stretched or heated to tune its reflection spectrum. Now the tapped-delay line may be implemented within a single piece of fiber, again with a circulator at the input to route the correlation output to the threshold detector. The FBG-based correlator shown in Fig. 4 is configured to recognize a bit pattern of 101 (or, more accurately, 1x1). This is accomplished by writing three gratings in a row, with center-to-center spacings equal to $\frac{1}{2}$ of a bit time in fiber so that the round-trip time between gratings corresponds to a 1-bit delay. The reflectivity of the third grating is ideally 100% since it is the last "mirror" in the series. The reflection spectrum of the second grating is tuned away from the incoming signal wavelength so that it will simply transmit the light and there will be no reflection for the $2T_{bit}$ delay. This is the equivalent of the "open switch" in the previous configuration. The first grating must then be tuned to only partially reflect the incoming light so that the remaining light can pass through to reflect off the third grating. The reflectivity of the first grating must be chosen carefully to guarantee that the pulses reflecting off the first grating have power equal to that of those that reflect off the third grating. In practice this can be determined by repeatedly sending a single pulse into the FBG array and observing the powers of the two, time-delayed output pulses on an oscilloscope (the two gratings will reflect the input pulse at different times, creating two pulses at the output). The first FBG can then be tuned until the two

pulses have equal power. The required reflectivities of the gratings in the array can also be calculated. Assume that there are n gratings in the array that represent 1 bits and therefore must provide some partial reflection of the incoming signal. Let the reflectivities of each grating be R_n (e.g., if $R_n = 0.4$, then the grating will reflect 40% of the incoming light and transmit the remaining 60% since $1 - R_n = 0.6$). Then the equation to determine the reflectivities needed to achieve equal output powers of all the time-delayed output pulses is

$$\frac{R_n}{(1-R_n)^2} = R_{n+1} \tag{2}$$

Thus, the reflectivity of the last grating in the array should be set equal to 1, and then the reflectivities of all the preceding gratings can be calculated using this recursive equation. For the case shown in Fig. 4, with $R_3 = 1$, we get $R_1 = (1 - R_1)$, which results in $R_1 = 38\%$. The cross-correlation output for the case of an input sequence of 011 is depicted in Fig. 4. Limitations of this method due to the increasing attenuation of the signal as it makes a double pass through the series of gratings will be discussed in the following section.

By replacing the switches or reflectivity-tuned FBGs in the previous configurations with phase shifters or phase-encoded FBGs, the fiber-based optical intensity correlators described above can be used as phase correlators instead [7]. The holographic correlator may also be adapted to correlate with phase-encoded signals. For these cases, the correlation sequences are a series of phase shifts such as $0\pi\pi00\pi0\pi$, for an 8-bit codeword. As long as the incoming codeword contains the same phase shifts, an autocorrelation peak will result. There are several methods of generating optical codewords containing such a series of phase shifts. One all-fiber method utilizes a single FBG, 4 cm long, with seven very thin wolfram wires (diameter = 25 μm) wrapped around the grating every 5 mm [8]. The 5 mm spacing provides a 50-ps round-trip delay, corresponding to 1 bit time at 20 Gbps, and the 7 wires mean that this correlator can recognize an 8-bit phase-encoded sequence. By passing a current through one of the wires, the FBG will be heated at only that point in the grating, causing the signal being reflected by the FBG to experience a phase shift at the time delay corresponding to the location of the wire. The heat-induced phase shift occurs because the index of refraction of the glass fiber is temperature-sensitive. Note that only point heating is desired here to induce the phase shifts — it is not the aim to shift the grating's reflection spectrum, which would occur if the entire grating were heated. This device may be used to both generate phase-encoded signals at the transmitter and to correlate with them at the receiver. The most common application of optical phase correlators is to construct optical CDMA encoders and decoders. One possible advantage of phase modulation over intensity modulation is the potential for multilevel coding. While intensity modulation coding primarily uses on/off keying (OOK), where a 1 bit is represented by the presence of light and a 0 bit by the absence of it, phase coding need not utilize phase shifts of only 0 and π. For example, another article demonstrated four-level encoding, using phase shifts of 0, $\pi/2$, π, and $3\pi/2$ [9].

This is merely a sampling of the available optical correlator designs, specifically concentrating on those that are easily compatible with fiberoptic telecommunication systems and provide easy-to-understand illustrations of optical correlators. However, many other novel correlator configurations have been developed in research labs. These include correlators based on semiconductor optical amplifiers (SOA) [10], erbium-doped fibers (EDFs) [11], optical loop mirrors [12], and nonlinear optical crystals [13].

4. IMPLEMENTATION CHALLENGES

While the previous section detailed some of the more common optical correlator structures, there are a number of roadblocks facing optical correlators that keep them from seeing wide use beyond research environments. However, a number of advances not only begin tackling these roadblocks but also aim to decrease the cost and increase the commercial viability of optical correlators.

A driving motivation for research in optical signal processing is the fact that electronics may at some point present a bottleneck in telecommunication networks. Optical signals in commercial networks currently carry data at 2.5 and 10 Gbps and in research environments, at 10, 40, and even 160 Gbps. At these higher speeds, either electronic circuits will be unable to efficiently process the data traffic, or it may actually become more economical to use optical alternatives. However, optical techniques also face significant hurdles in moving to 40 Gbps and beyond. In particular, optical correlators that rely on fiberoptic delays require greater fabrication precision as the bit rate rises. At 40 Gbps, a single bit time is 25 ps, corresponding to 0.5 cm of propagation in standard optical fiber. For the fiber-mirror-based correlator, the differences in length of the fiber branches must equal a $\frac{1}{2}$ of a bit time delay, or 2.5 mm — an impractically small length. The FBG-based correlator has an additional problem in that not only must the gratings have a center-to-center spacing of 2.5 mm at 40 Gbps, but the gratings themselves are often longer than 2.5 mm. While it is possible to make FBGs that are only 100s of micrometers long, the shorter length will make it difficult, but not impossible, to achieve a high-reflectivity grating with the appropriate bandwidth to reflect the higher data-rate signal. Furthermore, there must be some method to provide precise tuning of the closely spaced individual FBGs. One report [14] demonstrated a grating-based tapped-delay line correlator created from a single FBG. A periodically spaced series of thin-film metallic heaters were deposited on the surface of the FBG to effectively create several tunable FBGs from one long one. By passing individual currents through the heaters, the reflection spectrums for the portions of the FBG beneath the heaters are tuned away. Since the heaters can be deposited with essentially lithographic precision, this resolves the issue of how to precisely fabricate and tune very closely spaced FBGs. Of course, this technique will eventually be limited by how closely spaced the heaters can be before thermal crosstalk between neighboring elements becomes a significant problem. Spectral holography is perhaps one of the more promising techniques for higher bit rates, as it is currently

feasible (albeit expensive) to create holographic plates that can correlate with high-speed signals. However, the high absorption loss of current free-space holographic correlators at communication wavelengths remains a roadblock to any practical implementation.

Coherence effects can also present problems in optical correlators that utilize tapped-delay-line structures in which the light is split into several time-delayed replicas that are recombined at the correlator output. The coherence time of standard telecommunication lasers is typically tens of nanoseconds, corresponding to a coherence length in fiber of ~2 ms. When the differential time delay between two branches of the correlator is less than the coherence time of the laser (which is clearly the typical case), then the recombined signals will coherently interfere with each other, causing large power fluctuations in the correlation output function. This effect destroys the correlation output and must be mitigated or prevented in order to effectively operate the correlator. There are a number of ways that this problem can be solved. A polarization controller followed by a polarization beamsplitter (PBS) can be used at the input of each 1×2 optical splitter/combiner to ensure that the electric field polarizations between the two branches are orthogonal. This will prevent coherent interference of the recombined signals because orthogonally polarized lightbeams will not interfere with each other. Thus, for more than two branches, a tree structure of 1×2 splitters with polarization controllers and polarization beamsplitters can be used. Another, more manageable, solution is to somehow convert the coherent light into an incoherent signal before entering the correlator. One method of doing this uses cross-gain modulation in a semiconductor optical amplifier (SOA) to transfer the coherent data pattern onto incoherent light, which in this case is the amplified spontaneous emission light generated by the SOA [15].

An additional problem facing the grating-based correlator is the severe optical losses associated with multiple passes through the FBGs. This can so significantly limit the length of the correlation sequence that it can realistically correlate with before the power in the correlation output pulses are so low that they drop below the system noise floor. As explained before, the reflectivities of the gratings are determined by the need to equalize the pulses reflecting off of each grating representing a 1 bit, while recognizing that the pulses are attenuated with each pass through each grating. With four 1 bits in a correlation sequence, only ~65% of the incident light is present in the total correlation output. The rest is lost as a result of multiple reflections within the correlator, as the reflection off an internal grating can reflect again on the return trip and will no longer fall within the correlation time window. These multiple internal reflections do cause undesired replicas of the signal to add to the correlation output function, but they are so attenuated by the multiple reflections that they are not typically considered a significant impairment. As the number of 1 bits increases, the correlation sensitivity will decrease as more light is lost to these reflections. One method to mitigate this problem is to interleave the gratings between multiple fiber branches [6]. For the 4-bit sequence, two branches, each with two gratings, preceded by a splitter, can be used instead of a single row of gratings, thereby increasing the efficiency (assuming that all gratings are 1 bits) to ~75% of the incident light. This solution also alleviates the difficulty of spacing the gratings very closely together since only every other grating is on each branch, at the cost of introducing coherence effects that must be mitigated.

5. ADVANCED TOPICS

As wavelength-division-multiplexed (WDM) systems are becoming the standard in optical networks, it is becoming increasingly desirable to build modules that can act on multiple wavelength channels simultaneously. To this end, there is a growing interest in multiwavelength optical correlators. One study of a multiwavelength optical correlator uses sampled FBGs in a grating-based correlator structure. The reflection spectrum of a sampled FBG possesses multiple passbands so that it can filter several wavelength channels simultaneously [16]. When this type of FBG is stretched or heated, the entire reflection spectrum shifts, causing the reflectivity at each wavelength to experience the same variation. Thus, by replacing the standard FBGs with sampled FBGs in the correlator structure described previously, incoming signals on multiple wavelengths can simultaneously be correlated with a single correlation sequence. While it may still be necessary to demultiplex these channels prior to detection in order to check the correlation output for each channel, this technique still significantly reduces the number of components that would otherwise be required in a system that provided a separate correlator for each wavelength.

In Section 2, we mentioned that the conventional N-bit optical tapped-delay-line correlator cannot be configured to uniquely identify all 2^N possible bit patterns. This is because a 0 bit is represented by an open switch or a grating that is tuned not to reflect any light. This really means that these bit positions are considered "don't care" bits and the desired autocorrelation peak will result at the sample time whether the "don't care" bits are 1s or 0s. This is not an issue in optical CDMA systems, where the set of codewords can be specifically designed to maintain a constant number of 1 bits in each codeword. However, for applications that must be able to uniquely recognize any of the 2^N possible N-bit sequences, this situation will result in false-positive matches whenever a 1 is present where a 0 bit is desired. This is important for applications such as packet header or label recognition. A solution to this problem is to add a second correlator that is configured in complement to the first one and produces a "match" signal when *zero* power is present at the sample time. This is accomplished by placing a NOT gate at the output of the threshold detector which is set just above level zero. If the power goes above the threshold, this indicates that at least one 1 bit is present where a 0 is desired, and the NOT gate will convert the high output to a low one to indicate "no match" for this correlator. This correlator therefore correlates with the desired 0 bits in the sequence and is thus called a "zeros" correlator. Likewise, the originally described correlator is called a "ones" correlator. In the zeros correlator, the switches are closed for desired 0 bits

and open otherwise (or the FBGs reflect for desired 0 bits and are tuned away otherwise). Thus, the 1 bits are "don't care" bits in a zeros correlator. By combining the outputs of the ones and zeros correlators with an AND gate, the input sequence will only produce a final "match" signal when the input pattern uniquely matches the desired correlation sequence. An illustration of how the combination of a ones and a zeros correlator can avoid false positive matches is depicted in Fig. 5. The desired correlation sequence is 1001, meaning that the ones correlator is configured to match a 1xx1 pattern and the zeros correlator will produce a match for an $x00x$ pattern. In Fig. 5a, the incoming sequence is 1001, and so the ones and zeros correlators both produce "match" signals, resulting in a "match" signal at the output of the AND gate. In Fig. 5b, the input sequence is a 1101, causing the ones correlator to still produce a "match" signal (this would be a false-positive match if only this correlator were used). But the undesired 1 bit in the second time slot of the input sequence causes the power at the sample time in the zeros correlator to go above threshold, resulting in a "no match." The combination of the "match" and "no match" signals in the AND gate produce the correct "no match" result.

6. CONCLUSION

While optical correlators currently see limited commercial application, the frontier of all-optical networking is rapidly approaching, aided by the ever-increasing demand to transmit more bandwidth over the network core. This, in addition to the growing interest in optical CDMA networks, will push designers to develop producible optical correlators that can be dynamically adjusted to recognize very high bit-rate sequences. For applications such as header or label recognition, technologies that can implement huge arrays or banks of correlators to efficiently test the incoming signal against all possible bit sequences will be needed. Reaching these goals presents a significant engineering challenge, but research is continuing to make progress, and optical correlators, combined with the appropriate data coding techniques, offer great potential for bringing the ever-expanding field of optical signal processing to light.

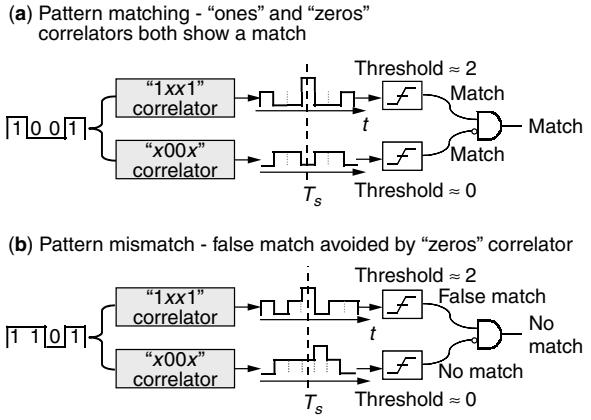


Figure 5. The concept of combining 1s and 0s correlators to uniquely recognize a bit sequence and avoid false-positive matches. The "ones" correlator tests for 1 bits in the correlation sequence and the "zeros" correlator tests for 0 bits. The correlators shown are configured to recognize a 1001 pattern when their outputs are combined in an AND gate. (**a**) The input pattern 1001 results in a match for both correlators, producing a final "match" decision at the output. (**b**) The input pattern 1101 results in a match for the "ones" correlators but a "no match" for the "zeros" correlator. The combination of these two outputs produces the correct "no match" decision at the output of the AND gate.

BIOGRAPHIES

John E. McGeehan received his B.S. and M.S. degrees in electrical engineering at the University of Southern California in Los Angeles in 1998 and 2001, respectively. He joined the Optical Communications Laboratory at the University of Southern California as a research assistant in 2001 and currently is working toward his doctorate. His research interests include the implementation of all-optical networking functions and optical signal processing as well as Raman amplification and signal monitoring. He is an author or co-author of nine technical papers.

Michelle C. Hauer received the B.S. degree in engineering physics from Loyola Marymount University, Los Angeles, California, in 1997. She currently is a research assistant in the Optical Fiber Communications Laboratory at the University of Southern California, Los Angeles, California, where she received the M.S.E.E. degree in 2000. Her doctoral research includes optical signal processing techniques for implementing all-optical networking functions. She is the author or co-author of more than 13 research papers. She is a member of the Tau Beta Pi, Eta Kappa Nu, and Sigma Pi Sigma academic honor societies in engineering, electrical engineering, and physics, respectively. She has also held a position as a systems engineer at Raytheon Company in El Segundo, California since 1997.

Alan Willner received his B.S. from Yeshiva University and his Ph.D. from Columbia University. He has worked at AT&T Bell Labs and Bellcore, and is Professor of Electrical Engineering at the University of Southern California. Professor Willner has received the following awards: the National Science Foundation (NSF) Presidential Faculty Fellows Award from the White House, the Packard Foundation Fellowship, the NSF National Young Investigator Award, the Optical Society of America (OSA) Fellow Award, the Fulbright Foundation Senior Scholars Award, the Institute of Electronic and Electrical Engineers (IEEE) Lasers and Electro-Optics Society (LEOS) Distinguished Traveling Lecturer Award, the USC/TRW Best Engineering Teacher Award, and the Armstrong Foundation Memorial Prize. His professional activities have included: editor-in-chief of the IEEE/OSA *Journal of Lightwave Technology*, editor-in-chief of the IEEE *Journal of Selected Topics in Quantum Electronics*, V.P. for Technical Affairs of the IEEE LEOS, Co-Chair of the OSA Science and Engineering Council, Elected Member of the IEEE LEOS Board of Governors, Program Co-Chair of the Conference on Lasers and Electro-Optics (CLEO), General Chair of the LEOS Annual Meeting Program, Program Co-Chair of the OSA Annual Meeting, OSA Photonics Division Chair,

General Co-Chair of the OSA Optical Amplifier Meeting, and Steering and Program Committee Member of the Conference on Optical Fiber Communications (OFC). Professor Willner has 325 publications, including one book.

BIBLIOGRAPHY

1. F. G. Stremler, *Introduction to Communications Systems*, 3rd ed., Addison-Wesley, Reading MA, 1990.
2. A. Stok and E. H. Sargent, Lighting the local area: Optical code-division multiple access and quality of service provisioning, *IEEE Network* 42–46 (Nov./Dec. 2000).
3. J. Widjaja, N. Wada, Y. Ishii, and W. Chijo, Photonic packet address processor using holographic correlator, *Electron. Lett.* **37**(11): 703–704 (2001).
4. J.-D. Shin, M.-Y. Jeon, and C.-S. Kang, Fiber-optic matched filters with metal films deposited on fiber delay-line ends for optical packet address detection, *IEEE Photon. Tech. Lett.* **8**(7): 941–943 (1996).
5. R. Kashyap, *Fiber Bragg Gratings*, Academic Press, San Diego, 1999.
6. D. B. Hunter and R. A. Minasian, Programmable high-speed optical code recognition using fibre Bragg grating arrays, *Electron. Lett.* **35**(5): 412–414 (1999).
7. A. Grunnet-Jepsen et al., Spectral phase encoding and decoding using fiber Bragg gratings, *Proc. Conf. Optical Fiber Communications* (*OFC*) *1999*, paper PD33, 1999, pp. PD33-1–PD33-3.
8. M. R. Mokhtar, M. Ibsen, P. C. Teh, and D. J. Richardson, Simple dynamically reconfigurable OCDMA encoder/decoder based on a uniform fiber Bragg grating, *Proc. Conf. Optical Fiber Communications* (*OFC*) *2002*, paper ThGG54, 2002, pp. 688–690.
9. P. C. Teh et al., Demonstration of a four-channel WDM/OCDMA system using 255-chip 320-Gchip/s quaternary phase coding gratings, *IEEE Photon. Tech. Lett.* **14**(2): 227–229 (2002).
10. P. Petruzzi et al., All optical pattern recognition using a segmented semiconductor optical amplifier, *Proc. Eur. Conf. Optical Communication* (*ECOC*) *2001*, paper We.B.2.1, 2001, pp. 304–305.
11. J. S. Wey, J. Goldhar, D. L. Butler, and G. L. Burdge, Investigation of dynamic gratings in erbium-doped fiber for optical bit pattern recognition, *Proc. Conf. Lasers and Electro-optics* (*CLEO*) *1997*, paper CThW1, 1997, pp. 443–444.
12. N. Kishi, K. Kawachi, and E. Yamashita, Auto-correlation method for weak optical short pulses using a nonlinear amplifying loop mirror, *Proc. Eur. Conf. Optical Communication* (*ECOC*) *1997*, paper 448, 1997, pp. 215–218.
13. Z. Zheng and A. M. Weiner, Spectral phase correlation of coded femtosecond pulses by second-harmonic generation in thick nonlinear crystals, *Opt. Lett.* **25**(13): (2000).
14. M. C. Hauer et al., Dynamically reconfigurable all-optical correlators to support ultra-fast internet routing, *Proc. Conf. Optical Fiber Communications* (*OFC*) *2002*, paper WM7, 2002, pp. 268–270.
15. P. Parolari et al., Coherent-to-incoherence light conversion for optical correlators, *J. Lightwave Technol.* **18**(9): 1284–1288 (2000).
16. J. McGeehan, M. C. Hauer, A. B. Sahin, and A. E. Willner, Reconfigurable multi-wavelength optical correlator for header-based switching and routing, *Proc. Conf. Optical Fiber Communications* (*OFC*) *2002*, paper WM4, 2002, pp. 264–266.

DIGITAL PHASE MODULATION AND DEMODULATION

FUQIN XIONG
Cleveland State University
Cleveland, Ohio

1. INTRODUCTION

Digital signals or messages, such as binary 1 and 0, must be modulated onto a high-frequency carrier before they can be transmitted through some communication channels such as a coaxial cable or free space. The carrier is usually an electrical voltage signal such as a sine or cosine function of time t

$$s(t) = A\cos(2\pi f_c t + \theta)$$

where A is the amplitude, f_c is the carrier frequency, and θ is the initial phase. Each of them or a combination of them can be used to carry digital messages. The total phase $\Phi(t) = 2\pi f_c t + \theta$ can also be used to carry digital messages. The process that impresses a message onto a carrier by associating one or more parameters of the carrier with the message is called *modulation*; the process that extracts a message from a modulated carrier is called *demodulation*.

There are three basic digital modulation methods: amplitude shift keying (ASK), frequency shift keying (FSK), and phase shift keying (PSK). In ASK, data 1 and 0 are represented by the presence and absence of a burst of the carrier sine wave, respectively. The burst of the carrier that lasts a duration of a data period is called a *symbol*. The frequency and phase of the carrier are kept unchanged from symbol to symbol. In FSK, data 1 and 0 are represented by two different frequencies of the carrier, respectively. The amplitude and phase of the carrier are kept unchanged from symbol to symbol. In PSK, data 1 and 0 are represented by two different phases (e.g., 0 or π radians) of the carrier, respectively. The amplitude and frequency of the carrier are kept unchanged from symbol to symbol.

Modulation schemes are usually evaluated and compared using three criteria: *power efficiency, bandwidth efficiency, and system complexity*.

The bit error probability (P_b), or bit error rate (BER), as it is commonly called, of a modulation scheme is related to E_b/N_0, where E_b is the energy of the modulated carrier in a bit duration, and N_0 is the noise power spectral density. The *power efficiency* of a modulation scheme is defined as the required E_b/N_0 for a certain bit error probability (P_b), over an additive white Gaussian noise (AWGN) channel.

The *bandwidth efficiency* is defined as the number of bits per second that can be transmitted in one hertz (1 Hz) of system bandwidth. System bandwidth requirement depends on different criteria. Assuming the system uses

Nyquist (ideal rectangular) filtering at baseband, which has the minimum bandwidth required for intersymbol-interference (ISI)-free transmission of digital signals, then the bandwidth at baseband is $0.5R_s$, where R_s is the symbol rate, and the bandwidth at carrier frequency is $W = R_s$. Since $R_s = R_b/\log_2 M$, where R_b is the bit rate, for M-ary modulation, the bandwidth efficiency is

$$\eta_B = \frac{R_b}{W} = \log_2 M \quad \text{(bps/Hz) (Nyquist)}$$

This is called the Nyquist bandwidth efficiency. For modulation schemes that have power density spectral nulls such as the ones of PSK in Fig. 2, the bandwidth may be defined as the width of the main spectral lobe. Bandwidth efficiency based on this definition of bandwidth is called null-to-null bandwidth efficiency. If the spectrum of the modulated signal does not have nulls, null-to-null bandwidth no longer exists, as in the case of continuous-phase modulation (CPM). In this case, energy percentage bandwidth may be used as a definition. Usually 99% is used, even though other percentages (e.g., 90%, 95%) are also used. Bandwidth efficiency based on this definition of bandwidth is called percentage bandwidth efficiency.

System complexity refers to the circuit implementation of the modulator and demodulator. Associated with the system complexity is the cost of manufacturing, which is, of course, a major concern in choosing a modulation technique. Usually the demodulator is more complex than the modulator. A coherent demodulator is much more complex than a noncoherent demodulator since carrier recovery is required. For some demodulation methods, sophisticated algorithms, such as the Viterbi algorithm, are required. All these are basis for complexity comparison.

In comparison with ASK and FSK, PSK achieves better power efficiency and bandwidth efficiency. For example, in terms of power efficiency, binary PSK is 3 dB better than binary ASK and FSK at high signal-to-noise ratios, while BPSK is the same as BASK and better than BFSK in terms of bandwidth efficiency. Because of the advantages of PSK schemes, they are widely used in satellite communications, fixed terrestrial wireless communications, and wireless mobile communications.

A more complex, but more efficient PSK scheme is quadrature phase shift keying (QPSK), where the initial phase of the modulated carrier at the start of a symbol is any one of four evenly spaced values, say, $(0, \pi/2, \pi, 3\pi/2)$ or $(\pi/4, 3\pi/4, 5\pi/4, 7\pi/4)$. In QPSK, since there are four different phases, each symbol can represent two data bits. For example, 00, 01, 10, and 11 can be represented by $0, \pi/2, \pi$, and $3\pi/2$, respectively. In general, assuming that the PSK scheme has M initial phases, known as M-ary PSK or MPSK, the number of bits represented by a symbol is $n = \log_2 M$. Thus each 8-PSK symbol represents 3 bits, each 16-PSK symbol represents 4 bits, and so on.

As the order (M) of the PSK scheme is increased, the bandwidth efficiency is increased. In terms of null-to-null bandwidth, which is often used for PSK schemes, the efficiency is

$$\eta_B = \frac{R_b}{W} = \frac{R_b}{2R_s} = \frac{R_b}{2R_b/\log_2 M} = \frac{1}{2}\log_2 M$$

(bps/Hz) (null-to-null)

Thus the bandwidth efficiencies of BPSK, QPSK, 8-PSK, and 16-PSK are 0.5, 1, 1.5, and 2 bps/Hz.

Using the bandwidth efficiency, the bit rate that can be supported by a system bandwidth can be easily calculated. From above we have

$$R_b = \eta_B W$$

For example, for a satellite transponder that has a bandwidth of 36 MHz, assuming that Nyquist filtering is achieved in the system, the bit rate will be 36 Mbps for BPSK, 72 Mbps for QPSK, and so on. If null-to-null bandwidth is required in the system, the bit rate will be 18 Mbps for BPSK, 36 Mbps for QPSK, and so on. Practical systems may achieve bit rates somewhere between these two sets of values.

In this article, we first describe the commonly used class of PSK schemes, that is, MPSK. Then BPSK and QPSK are treated as special cases of MPSK. Next, differential PSK schemes are introduced, which are particularly useful in channels where coherent demodulation is difficult or impossible, such as fading channels. At the end of this article, advanced phase modulation schemes, such as offset QPSK (OQPSK), $\pi/4$-DQPSK, and continuous-phase modulation (CPM), including multi-h CPM, are briefly introduced. A future trend in phase modulation is in the direction of CPM and multi-h CPM.

2. PSK SCHEMES AND THEIR PERFORMANCE

2.1. Signal Waveforms

In PSK, binary data are grouped into n bits in a group, called n-tuples or symbols. There are $M = 2^n$ possible n-tuples. When each n-tuple is used to control the phase of the carrier for a period of T, *M-ary phase shift keying* (MPSK) is formed. The MPSK signal set is defined as

$$s_i(t) = A\cos(2\pi f_c t + \theta_i), \qquad 0 \le t \le T, \qquad i = 1, 2, \ldots, M \tag{1}$$

where A is the amplitude, f_c is the frequency of the carrier, and

$$\theta_i = \frac{(2i-1)\pi}{M}$$

are the (initial) phases of the signals. Note that θ_i are equally spaced. Each signal $s_i(t)$ is a burst of carrier of period T, called a *symbol waveform*. Each symbol waveform has a unique phase that corresponds to a n-tuple; or, in other words, each n-tuple is represented by a symbol waveform with a unique phase. When the PSK signal is transmitted and received, the demodulator detects the phase of each symbol. On the basis of phase information, the corresponding n-tuple is recovered. Except for the phase difference, the M symbol waveforms in MPSK have the same amplitude (A), the same frequency (f_c), and the same energy:

$$E = \int_0^T s_i^2(t)\,dt = \frac{A^2T}{2}, \qquad i = 1, 2, \ldots, M$$

The simplest PSK scheme is *binary phase shift keying* (BPSK), where binary data (1 and 0) are represented

by two symbols with different phases. This is the case when $M = 2$ in MPSK. Typically these two phases are 0 and π. Then the two symbols are $\pm A \cos 2\pi f_c t$, which are said to be *antipodal*.

If binary data are grouped into 2 bits per group, called dibits, there are four possible dibits: 00, 01, 10, and 11. If each dibit is used to control the phase of the carrier, then we have *quadrature phase shift keying* (QPSK). This is the case when $M = 4$ in MPSK. Typically, the initial signal phases are $\pi/4, 3\pi/4, 5\pi/4$, and $7\pi/4$. Each phase corresponds to a dibit. QPSK is the most often used scheme since its BER performance is the same while the bandwidth efficiency is doubled in comparison with BPSK, as will be seen shortly.

2.2. Signal Constellations

It is well known that sine-wave signals can be represented by phasors. The phasor's magnitude is the amplitude of the sine-wave signal; its angle with respect to the horizontal axis is the phase of the sine-wave signal. In a similar way, signals of modulation schemes, including PSK schemes, can be represented by phasors. The graph that shows the phasors of all symbols in a modulation scheme is called a signal constellation. It is a very convenient tool for visualizing and describing all symbols and their relationship in the modulation scheme. It is also a powerful tool for analyzing the performance of the modulation scheme.

The PSK waveform can be written as

$$s_i(t) = A \cos\theta_i \cos 2\pi f_c t - A \sin\theta_i \sin 2\pi f_c t \tag{2}$$
$$= s_{i1}\phi_1(t) + s_{i2}\phi_2(t), \qquad 0 \le t \le T, \qquad i = 1, 2, \ldots, M$$

where

$$\phi_1(t) = \sqrt{\frac{2}{T}} \cos 2\pi f_c t, \qquad 0 \le t \le T \tag{3}$$

$$\phi_2(t) = -\sqrt{\frac{2}{T}} \sin 2\pi f_c t, \qquad 0 \le t \le T \tag{4}$$

are two orthonormal basis functions and

$$s_{i1} = \int_0^T s_i(t)\phi_1(t)\,dt = \sqrt{E}\cos\theta_i \tag{5}$$

$$s_{i2} = \int_0^T s_i(t)\phi_2(t)\,dt = \sqrt{E}\sin\theta_i \tag{6}$$

are the projections of the signal onto the basis functions.

The phase is related with s_{i1} and s_{i2} as

$$\theta_i = \tan^{-1}\frac{s_{i2}}{s_{i1}}$$

Thus PSK signals can be graphically represented by a signal constellation in a two-dimensional coordinate system with the two orthonormal basis functions in Eqs. (5) and (6), $\phi_1(t)$ and $\phi_2(t)$, as its horizontal and vertical axes, respectively (Fig. 1). Each signal $s_i(t)$ is represented by a point (s_{i1}, s_{i2}), or a vector from the origin to this point, in the two coordinates. The polar coordinates of the signal are $(\sqrt{E}, \theta_i)$; that is, the signal vector's magnitude is $\sqrt{E}$ and its angle with respect to the horizontal axis is θ_i. The signal points are equally spaced on a circle of radius $\sqrt{E}$ and centered at the origin. The bits-signal mapping could be arbitrary provided that the mapping is one-to-one. However, a method called Gray coding is usually used in signal assignment in MPSK. Gray coding assigns n-tuples with only one-bit difference in two adjacent signals in the constellation. When an M-ary symbol error occurs because of the presence of noise, it is most likely that the signal is detected as the adjacent signal on the constellation; thus only one of the n input bits is in error. Figure 1 shows the constellations of BPSK, QPSK, and 8-PSK, where Gray coding is used for bit assignment.

2.3. Power Spectral Density

The *power spectral density* (PSD) of a signal shows the signal power distribution as a function of frequency. It provides very important information about the relative strength of the frequency components in the signal, which, in turn, determines the bandwidth requirement of the communication system.

The power spectral density of a bandpass signal, such as a PSK signal, is centered about its carrier frequency. Moving the center frequency down to zero, we obtain the baseband signal PSD, which completely characterizes the signal [1–3]. The baseband PSD expression of MPSK can be expressed as [3]

$$\Psi_{\tilde{s}}(f) = A^2 T \left(\frac{\sin \pi f T}{\pi f T}\right)^2 = A^2 n T_b \left(\frac{\sin \pi f n T_b}{\pi f n T_b}\right)^2 \tag{7}$$

where T_b is the bit duration, $n = \log_2 M$. Figure 2 shows the PSDs ($A = \sqrt{2}$ and $T_b = 1$ for unit bit energy: $E_b = 1$) for different values of M where the frequency axis is normalized to the bit rate (fT_b). From the figure we can see that the bandwidth decreases with M, or in other words, the bandwidth efficiency increases with M. The Nyquist bandwidth is

$$B_{\text{Nyquist}} = \frac{1}{nT_b} = \frac{R_b}{\log_2 M}$$

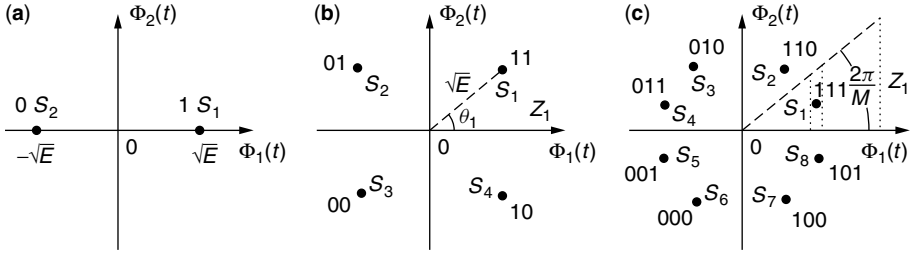


Figure 1. PSK constellations: (**a**) BPSK; (**b**) QPSK; (**c**) 8PSK.

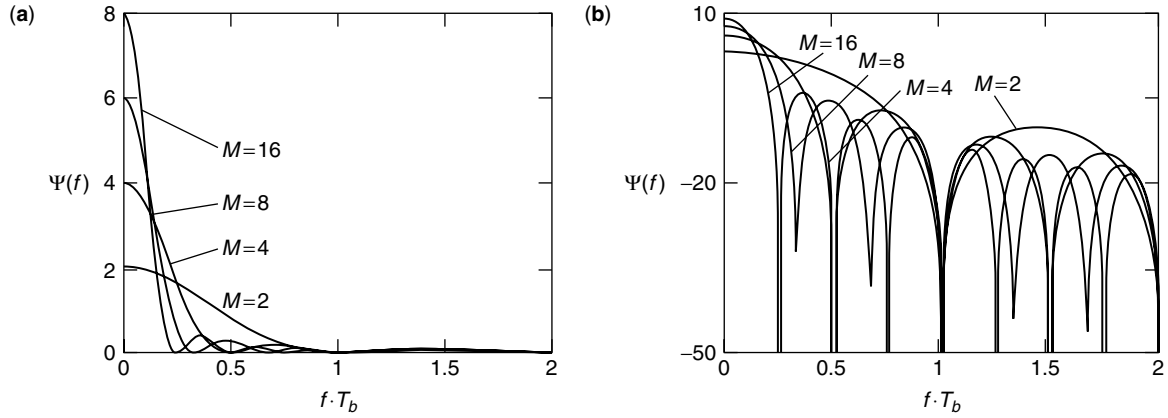


Figure 2. PSDs of MPSK: (**a**) linear; (**b**) logarithmic.

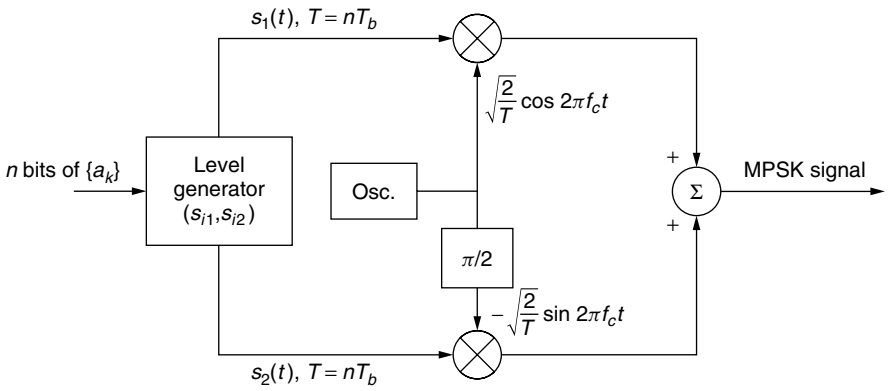


Figure 3. MPSK modulator (Osc. = oscillator). (From Ref. 3, copyright © 2000 Artech House.)

This translates to the Nyquist bandwidth efficiency of $\log_2 M$. The null-to-null bandwidth is

$$B_{null-to-null} = \frac{2}{nT_b} = \frac{2R_b}{\log_2 M}$$

This translates to a bandwidth efficiency of $0.5 \log_2 M$, which is half of that of the Nyquist bandwidth efficiency. Practical systems can achieve bandwidth efficiencies between these two values.

2.4. Modulator and Demodulator

Over the entire time axis, we can write MPSK signal as

$$s(t) = s_1(t)\sqrt{\frac{\pi}{2}} \cos 2\pi f_c t - s_2(t)\sqrt{\frac{\pi}{2}} \sin 2\pi f_c t, \quad -\infty < t < \infty \tag{8}$$

where

$$s_1(t) = \sqrt{E} \sum_{k=-\infty}^{\infty} \cos(\theta_k) p(t - kT) \tag{9}$$

$$s_2(t) = \sqrt{E} \sum_{k=-\infty}^{\infty} \sin(\theta_k) p(t - kT) \tag{10}$$

where θ_k is one of the M phases determined by the input binary n-tuple and $p(t)$ is the rectangular pulse with unit amplitude defined on $[0, T]$. Expression (8) requires that the carrier frequency be an integer multiple of the symbol timing so that the initial phase of the signal in any symbol period is θ_k.

Since MPSK signals are two-dimensional, for $M \geq 4$, the modulator can be implemented by a quadrature *modulator* (Fig. 3), where the upper branch is called the *in-phase channel* or *I channel*, and the lower branch is called the *quadrature channel* or the *Q channel*. The only difference for different values of M is the level generator. Each n-tuple of the input bits is used to control the level generator. It provides the I and Q channels the particular sign and level for a signal's horizontal and vertical coordinates, respectively.

Modern technology intends to use completely digital devices. In such an environment, MPSK signals are digitally synthesized and fed to a D/A (digital-to-analog) converter whose output is the desired phase-modulated signal.

The coherent demodulation of MPSK is shown in Fig. 4. The input to the *demodulator* is the received signal $r(t) = s_i(t) + n(t)$, where $n(t)$ is the additive white Gaussian noise (AWGN). There are two correlators, each consisting of a multiplier and an integrator. The carrier recovery (CR) block synchronizes the reference signals for the multipliers with the transmitted carrier in frequency and phase. This makes the demodulation

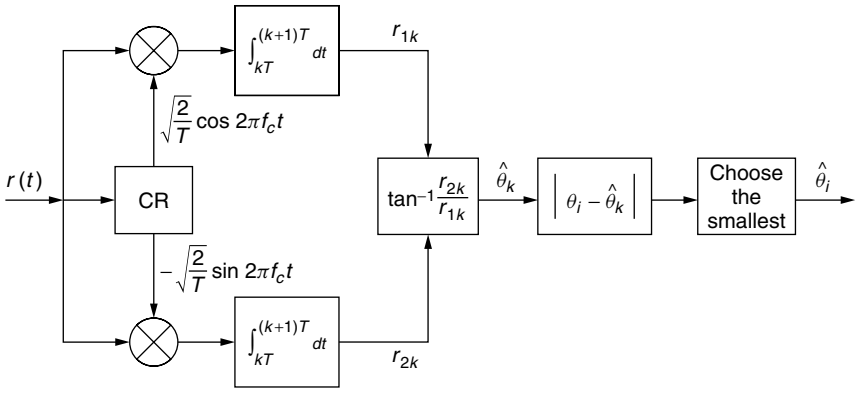


Figure 4. Coherent MPSK demodulator using two correlators. (From Ref. 3, copyright © 2000 Artech House.)

coherent and ensures the lowest bit error probability. The correlators correlate $r(t)$ with the two reference signals. Because of the orthogonality of the two components of the PSK signal, each correlator produces an output as follows:

$$r_1 = \int_0^T r(t)\phi_1(t)\,dt = \int_0^T [s(t) + n(t)]\phi_1(t)\,dt = s_{i1} + n_1$$

$$r_2 = \int_0^T r(t)\phi_2(t)\,dt = \int_0^T [s(t) + n(t)]\phi_2(t)\,dt = s_{i2} + n_2$$

where s_{i1} and s_{i2} are given as in Eqs. (5) and (6), respectively, and n_1 and n_2 are noise output. In Fig. 4 the subscript k indicates the kth symbol period.

Define

$$\widehat{\theta} \triangleq \tan^{-1}\frac{r_2}{r_1}$$

In the absence of noise, $\widehat{\theta} = \tan^{-1} r_2/r_1 = \tan^{-1} s_{i2}/s_{i1} = \theta_i$; that is, the PSK signal's phase information is completely recoverable in the absence of noise. With noise, $\widehat{\theta}$ will deviate from θ_i. To recover θ_i, the $\widehat{\theta}$ difference with all θ_i are compared and the θ_i that incurs the smallest $|\theta_i - \widehat{\theta}|$ is chosen.

The modulator and demodulator for BPSK and QPSK can be simplified from Figs. 3 and 4. For BPSK, since there is only I-channel component in the signal, the modulator is particularly simple: only the I channel is needed (Fig. 5a). The binary data (0,1) are mapped into (−1,1), which are represented by non-return-to-zero (NRZ) waveform $a(t)$, which is equivalent to Eq. (9). This NRZ waveform is multiplied with the carrier, and the result is the antipodal BPSK signal. The BPSK demodulator needs only one channel, too, as shown in Fig. 5b. In the absence of noise, the correlator output l is directly proportional to the data. Then l is compared with threshold zero, if $l \geq 0$ the data bit is 1; otherwise it is 0.

For QPSK, since it is equivalent to two parallel BPSK signals, the modulator can be simplified as shown in Fig. 6a, where $I(t)$ and $Q(t)$ are equivalent to Eqs. (9) and (10), respectively, except for a constant factor; the level generator is simply a serial-to-parallel converter. The demodulator is shown in Fig. 6b, which is simply a pair of two parallel BPSK demodulators.

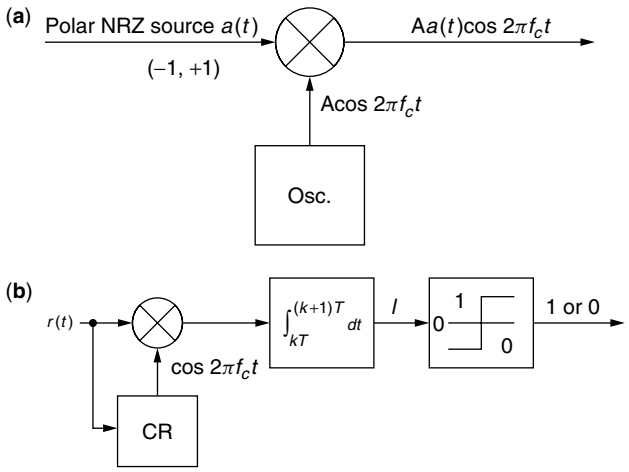


Figure 5. BPSK modulator (**a**); coherent BPSK demodulator (**b**). (From Ref. 3, copyright © 2000 Artech House.)

For $M > 4$, the general modulator and demodulator must be used.

2.5. Symbol and Bit Error Probability

In the demodulator shown in Fig. 4, a symbol error occurs when the estimated phase $\widehat{\theta}$ is such that the phase deviation $|\varphi| = |\widehat{\theta} - \theta_i| > \pi/M$ (see Fig. 1c). Thus the symbol error probability of MPSK is given by

$$P_s = 1 - \int_{-\pi/M}^{\pi/M} p(\varphi)\,d\varphi \tag{11}$$

It can be shown that $p(\varphi)$ is given by [3]

$$p(\varphi) = \frac{e^{-E/N_0}}{2\pi}\left\{1 + \sqrt{\frac{\pi E}{N_0}}(\cos\varphi)e^{(E/N_0)\cos^2\varphi} \times \left[1 + \operatorname{erf}\left(\sqrt{\frac{E}{N_0}}\cos\varphi\right)\right]\right\} \tag{12}$$

where

$$\operatorname{erf}(x) \triangleq \frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}\,du \tag{13}$$

is the error function.

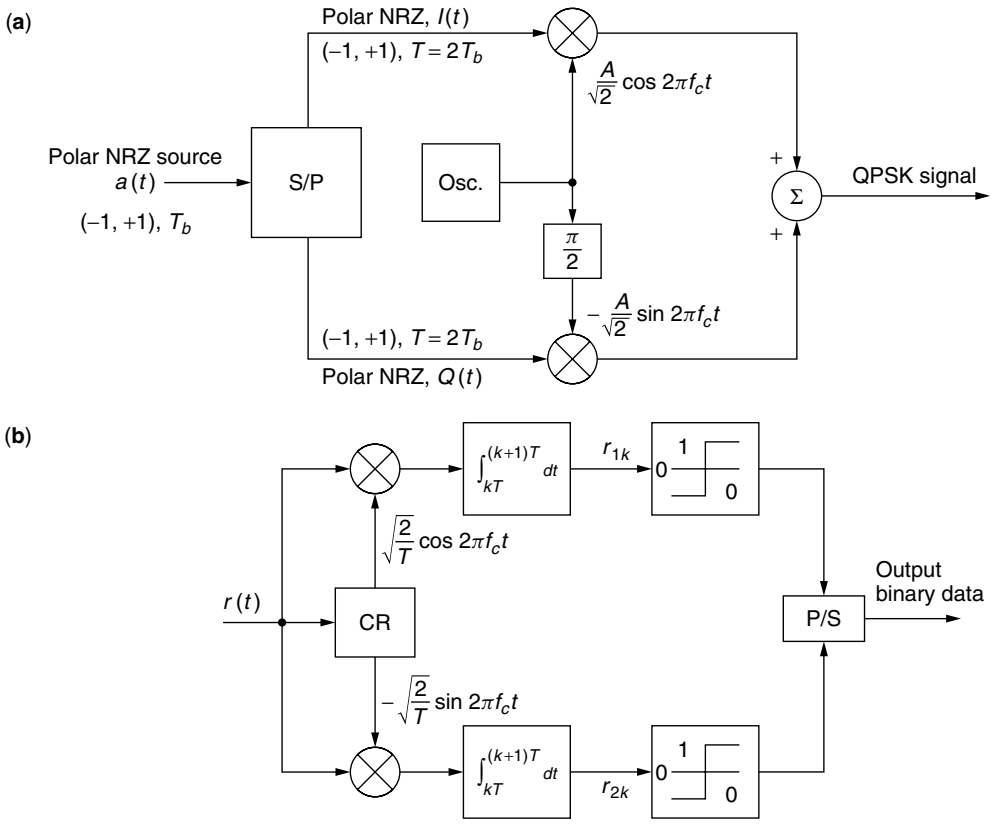


Figure 6. (**a**) QPSK modulator; (**b**) QPSK demodulator. (From Ref. 3, copyright © 2000 Artech House.)

When $M = 2$, (11) results in the formula for the symbol (and bit) error probability of BPSK:

$$P_s = P_b = Q\left(\sqrt{\frac{2E_b}{N_o}}\right) \qquad \text{(BPSK)} \tag{14}$$

where the E_b is the bit energy, which is also equal to the symbol energy E since a bit is represented by a symbol in BPSK, and

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{(-u^2/2)}\, du \tag{15}$$

is called the Q function.

When $M = 4$ [Eq. (11)] results in the formula for QPSK:

$$P_s = 2Q\left(\sqrt{\frac{2E_b}{N_0}}\right) - \left[Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\right]^2 \tag{16}$$

Since the demodulator of QPSK is simply a pair of two parallel BPSK demodulators, the bit error probability is the same as that of BPSK:

$$P_b = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \qquad \text{(QPSK)} \tag{17}$$

Recall that QPSK's bandwidth efficiency is double that of BPSK. This makes QPSK a preferred choice over BPSK in many systems.

For $M > 4$, expression (11) cannot be evaluated in a closed form. However, the symbol error probability can be obtained by numerically integrating (11).

Figure 7 shows the graphs of P_s for $M = 2, 4, 8, 16$, and 32 given by the exact expression (11). Beyond $M = 4$, doubling the number of phases, or increasing the n-tuples represented by the phases by one bit, requires a substantial increase in E_b/N_0 [or signal-to-noise ratio (SNR)]. For example, at $P_s = 10^{-5}$, the SNR difference between $M = 4$ and $M = 8$ is approximately 4 dB, the difference between $M = 8$ and $M = 16$ is approximately 5 dB. For large values of M, doubling the number of phases requires an SNR increase of 6 dB to maintain the same performance.

For $E/N_0 \gg 1$, we can obtain an approximation of the P_s expression as

$$P_s \approx 2Q\left(\sqrt{\frac{2E}{N_0}} \sin\frac{\pi}{M}\right) \tag{18}$$

Note that only the high signal-to-noise ratio assumption is needed for the approximation. Therefore (18) is good for any values of M, even though it is not needed for $M = 2$ and 4 since precise formulas are available.

The bit error rate can be related to the symbol error rate by

$$P_b \approx \frac{P_s}{\log_2 M} \tag{19}$$

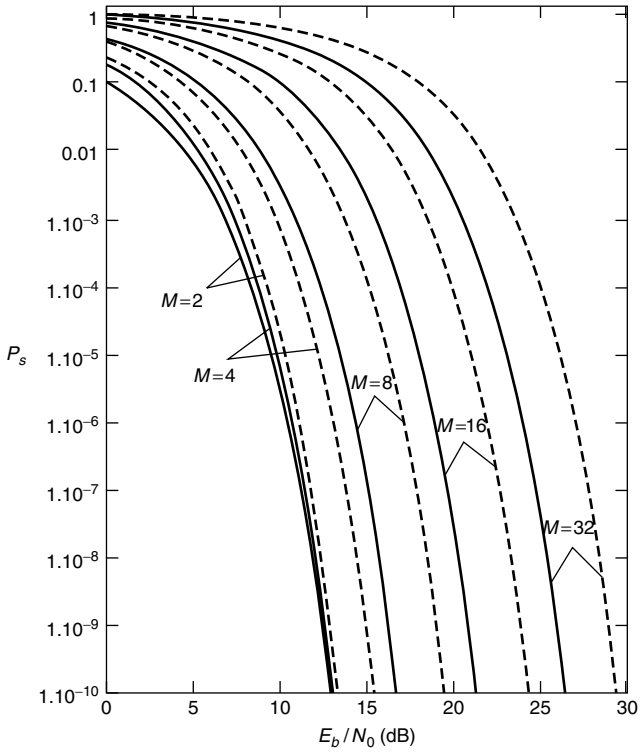


Figure 7. P_s of MPSK (solid lines) and DMPSK (dotted lines). (From Ref. 3, copyright © 2000 Artech House.)

for Gray-coded MPSK signals, since the most likely erroneous symbols are the adjacent signals that differ by only one bit.

Because of the substantial increase in SNR to maintain the same BER for larger M, higher-order PSK schemes beyond 8-PSK are not very often used. If further higher bandwidth efficiency is desired, quadrature amplitude modulation (QAM), which is a combination of phase and amplitude modulation, is a preferable choice over MPSK. For the same bandwidth efficiency, QAM delivers better power efficiency for $M > 4$ (see Section 8.7 of Ref. 3). However, QAM needs to preserve its amplitude through the transmitter stages. This can be difficult when nonlinear power amplifiers, such as traveling-wave-tube amplifiers (TWTAs) in satellite transponders, are used. QAM is widely used in telephone-line modems.

3. DIFFERENTIAL PSK SCHEMES

Differential encoding of a binary data sequence converts the original sequence into a new sequence of which each bit is determined by the difference of the current uncoded bit and the previous coded bit. Differential coding is needed in situations where coherent demodulation is difficult or phase ambiguity is a problem in carrier recovery. BPSK, QPSK, and MPSK all can be differentially coded.

3.1. Differential BPSK

We denote differentially encoded BPSK as *DEBPSK*. Figure 8a depicts the DEBPSK modulator. A DEBPSK signal can be coherently demodulated or differentially demodulated. We denote the modulation scheme that uses differential encoding and differential demodulation as *DBPSK*, which is sometimes simply called *DPSK*.

DBPSK does not require a coherent reference signal. Figure 8b shows a simple, but suboptimum, differential demodulator that uses the previous symbol as the reference for demodulating the current symbol. (This is commonly referred to as a *DPSK demodulator*. Another DPSK demodulator is the optimum differentially coherent demodulator. Differentially encoded PSK can also be coherently detected. These will be discussed shortly.) The front-end bandpass filter reduces the noise power but preserves the phase of the signal. The integrator can be replaced by a LPF (lowpass filter). The output of the integrator is

$$l = \int_{kT}^{(k+1)T} r(t)r(t-T)\,dt$$

In the absence of noise and other channel impairment

$$l = \int_{kT}^{(k+1)T} s_k(t)s_{k-1}(t)\,dt = \begin{cases} E_b & \text{if } s_k(t) = s_{k-1}(t) \\ -E_b & \text{if } s_k(t) = -s_{k-1}(t) \end{cases}$$

where $s_k(t)$ and $s_{k-1}(t)$ are the current and the previous symbols. The integrator output is positive if the current signal is the same as the previous one; otherwise the output is negative. This is to say that the demodulator makes decisions based on the difference between the two signals. Thus the information data must be encoded as the difference between adjacent signals, which is exactly what the differential encoding can accomplish. The encoding rule is

$$d_k = \overline{a_k \oplus d_{k-1}}$$

where $\oplus$ denotes modulo-2 addition. Inversely, we can recover a_k from d_k using

$$a_k = \overline{d_k \oplus d_{k-1}}$$

If d_k and d_{k-1} are the same, then they represent a 1 of a_k. If d_k and d_{k-1} are different, they represent a 0 of a_k.

The demodulator in Fig. 8 is suboptimum, since the reference signal is the previous symbol, which is noisy. The optimum noncoherent, or differentially coherent, demodulation of DEBPSK is given in Fig. 9. The derivation of this demodulator and its BER performance can be found in Ref. 3. Note that the demodulator of Fig. 9 does not require phase synchronization between the reference signals and the received signal. But it does require the reference frequency be the same as that of the received signal. Therefore the suboptimum receiver in Fig. 8b is more practical. Its error performance is slightly inferior to that of the optimum receiver.

The performance of the optimum receiver in Fig. 9 is given by

$$P_b = \frac{1}{2}e^{-E_b/N_0} \quad \text{(optimum DBPSK)} \tag{20}$$

The performance of the suboptimum receiver is given by Park [4]. It is shown that if an ideal narrowband IF

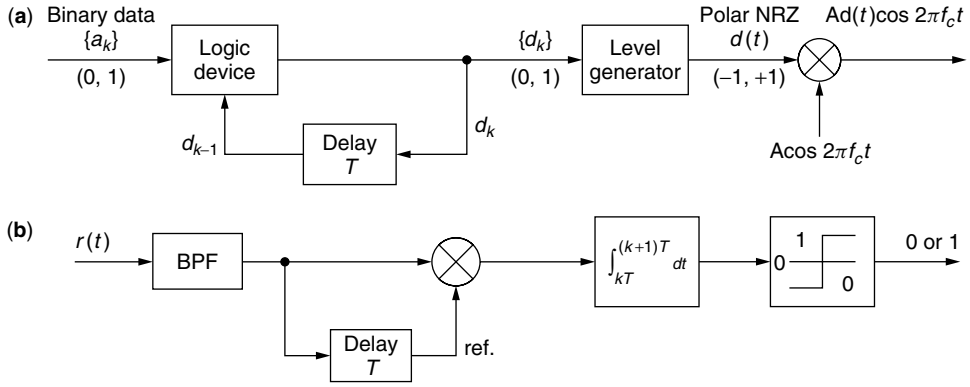


Figure 8. DBPSK modulator (**a**) and demodulator (**b**). (From Ref. 3, copyright © 2000 Artech House.)

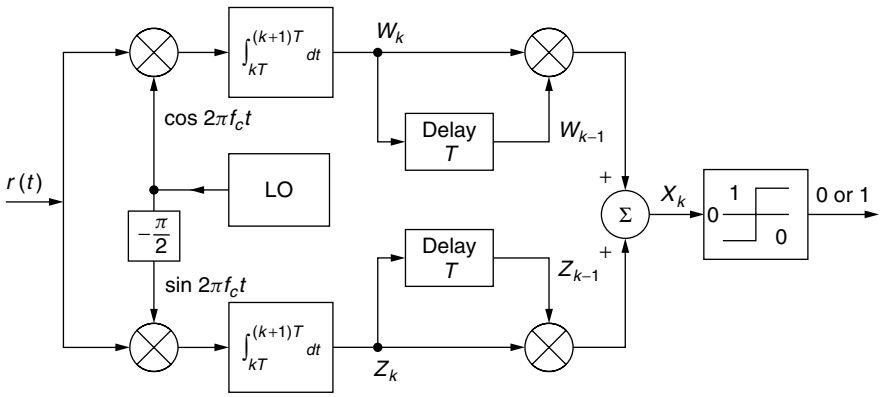


Figure 9. Optimum demodulator for DBPSK. (From Ref. 3, copyright © 2000 Artech House.)

(intermediate-frequency) filter with bandwidth W is placed before the correlator in Fig. 8b, the bit error probability is

$$P_b = \frac{1}{2}e^{-0.76E_b/N_0} \quad \text{for } W = \frac{0.5}{T} \quad \text{(suboptimum DBPSK)} \tag{21}$$

or

$$P_b = \frac{1}{2}e^{-0.8E_b/N_0} \quad \text{for } W = \frac{0.57}{T} \quad \text{(suboptimum DBPSK)} \tag{22}$$

which amounts to a loss of 1.2 and 1 dB, respectively, with respect to the optimum. If an ideal wideband IF filter is used, then

$$P_b \approx Q\left(\sqrt{\frac{E_b}{N_0}}\right) \quad \text{for } W > \frac{1}{T}$$
$$\approx \frac{1}{2\sqrt{\pi}\sqrt{E_b/2N_0}}e^{-E_b/2N_0}$$
$$\text{for } W > \frac{1}{T} \quad \text{(suboptimum DBPSK)} \tag{23}$$

The typical value of W is 1.5/T. If W is too large or too small, Eq. (23) does not hold [4]. The P_b for the wideband suboptimum receiver is about 2 dB worse than the optimum at high SNR. The bandwidth should be chosen as $0.57/T$ for the best performance. P_b curves of DBPSK are shown in Fig. 10.

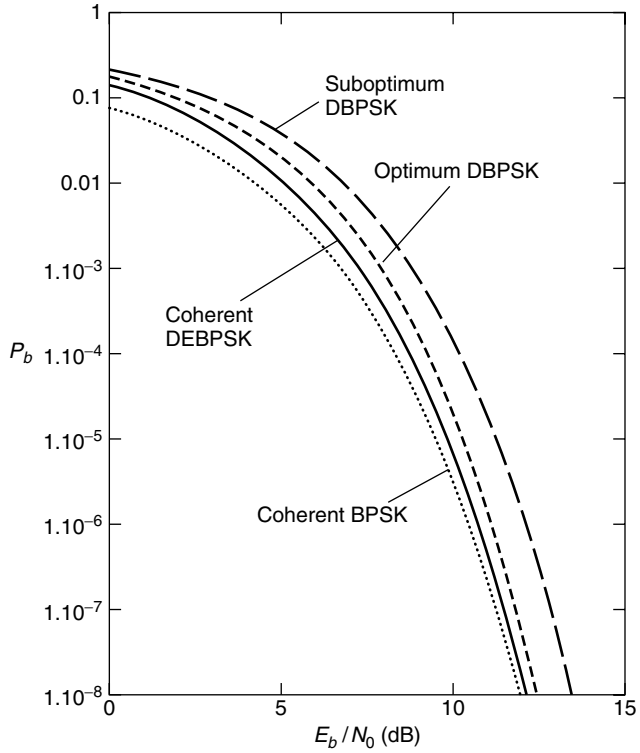


Figure 10. P_b of differential BPSK in comparison with coherent BPSK scheme. (From Ref. 3, copyright © 2000 Artech House.)

A differentially encoded BPSK signal can also be demodulated coherently (denoted as *DEBPSK*). It is used when the purpose of differential encoding is to eliminate phase ambiguity in the carrier recovery circuit for coherent PSK (see Section 4.10 of Ref. 3). This is rarely denoted by the acronym *DBPSK*. DBPSK refers to the scheme of differential encoding and differentially coherent demodulation as we have discussed above.

In the case of DEBPSK, the bit error probability (P_b) of the final decoded sequence $\{\widehat{a}_k\}$ is related to the bit error probability ($P_{b,d}$) of the demodulated encoded sequence $\{\widehat{d}_k\}$ by

$$P_b = 2P_{b,d}(1 - P_{b,d}) \tag{24}$$

(see Section 2.4.1 of Ref. 3). Substituting $P_{b,d}$ as in (14) into Eq. (24), we have

$$P_b = 2Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\left[1 - Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\right] \quad \text{(DEBPSK)} \tag{25}$$

for coherently detected differentially encoded PSK. For large SNR, this is just about 2 times that of coherent BPSK without differential encoding.

Finally we need to say a few words about the power spectral density of differentially encoded BPSK. Since the difference of differentially encoded BPSK from BPSK is differential encoding, which always produces an asymptotically equally likely data sequence (see Section 2.1 of Ref. 3), the PSD of the differentially encoded BPSK is the same as BPSK, in which we assume that its data sequence is equally likely. However, it is worthwhile to point out that if the data sequence is not equally likely, the PSD of the BPSK is not the one in Fig. 2, but the PSD of the differentially encoded PSK is still the one in Fig. 2.

3.2. Differential QPSK and MPSK

The principles of differential BPSK can be extended to MPSK, including QPSK. In *differential MPSK*, information bits are first differentially encoded. Then the encoded bits are used to modulate the carrier to produce the differentially encoded MPSK (DEMPSK) signal stream. In a DEMPSK signal stream, information is carried by the phase difference $\Delta\theta_i$ between two consecutive symbols. There are M different values of $\Delta\theta_i$; each represents an n-tuple ($n = \log_2 M$) of information bits.

In light of the modern digital technology, DEMPSK signals can be generated by digital frequency synthesis technique. A phase change from one symbol to the next is simply controlled by the n-tuple that is represented by the phase change. This technique is particularly suitable for large values of M.

Demodulation of DEMPSK signal is similar to that of differential BPSK [3]. The symbol error probability of the differentially coherent demodulator is approximated by

$$P_s \approx 2Q\left(\sqrt{\frac{2E}{N_0}}\sin\frac{\pi}{\sqrt{2}M}\right) \quad \text{(optimum DMPSK)} \tag{26}$$

for large SNR [6,7]. The exact curves are given as dotted lines in Fig. 7 together with those of coherent MPSK. Compared with coherent MPSK, asymptotically the DMPSK requires 3 dB more SNR to achieve the same error performance.

4. OTHER PSK SCHEMES

4.1. Offset QPSK

Offset QPSK (OQPSK) is devised to avoid the 180° phase shifts in QPSK [3,5]. OQPSK is essentially the same as QPSK except that the I- and Q-channel pulsetrains are staggered. The OQPSK signal can be written as

$$s(t) = \frac{A}{\sqrt{2}}I(t)\cos 2\pi f_c t - \frac{A}{\sqrt{2}}Q\left(t - \frac{T}{2}\right)\sin 2\pi f_c t, \quad -\infty < t < \infty \tag{27}$$

The modulator and the demodulator of OQPSK are basically identical to those of QPSK, except that an extra delay of $T/2$ seconds is inserted in the Q channel in the modulator and in the I channel in the demodulator.

Since OQPSK differs from QPSK only by a delay in the Q-channel signal, its power spectral density is the same as that of QPSK, and its error performance is also the same as that of QPSK.

In comparison to QPSK, OQPSK signals are less susceptible to spectral sidelobe restoration in satellite transmitters. In satellite transmitters, modulated signals must be band-limited by a bandpass filter in order to conform to out-of-band emission standards. The filtering degrades the constant-envelope property of QPSK, and the 180° phase shifts in QPSK cause the envelope to go to zero momentarily. When this signal is amplified by the final stage, usually a highly nonlinear power amplifier, the constant envelope will be restored. But at the same time the sidelobes will be restored. Note that arranging the bandpass filter after the power amplifier is not feasible since the bandwidth is very narrow compared with the carrier frequency. Hence the Q value of the filter must be extremely high such that it cannot be implemented by the current technology. In OQPSK, since the 180° phase shifts no longer exist, the sidelobe restoration is less severe.

4.2. π/4-DQPSK

Although OQPSK can reduce spectral restoration caused by a nonlinearity in the power amplifier, it cannot be differentially encoded and decoded. $\pi/4$-*DQPSK* is a scheme that not only has no 180° phase shifts like OQPSK but also can be differentially demodulated. These properties make it particularly suitable for mobile communications where differential demodulation can reduce the adversary effects of the fading channel. $\pi/4$-DQPSK has been adopted as the standard for the digital cellular telephone system in the United States and Japan.

$\pi/4$-DQPSK is a form of differential QPSK. The correspondence between data bits and symbol phase difference are shown as follows.

I_k Q_k	1 1	−1 1	−1 −1	1 −1
$\Delta\theta_k$	$\pi/4$	$3\pi/4$	$-3\pi/4$	$-\pi/4$

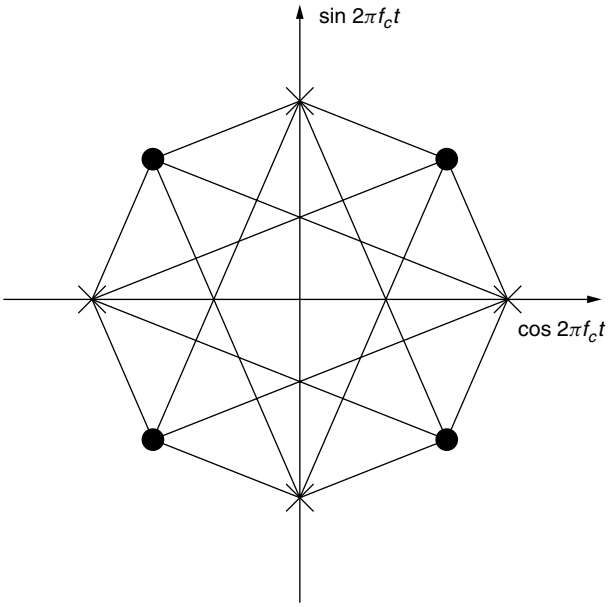


Figure 11. $\pi/4$-DQPSK signal constellation. (From Ref. 3, copyright © 2000 Artech House.)

We can see that the phase changes are confined to odd-number multiples of $\pi/4(45^\circ)$. There are no phase changes of 90° or 180°. In addition, information is carried by the phase changes $\Delta\theta_k$, not the absolute phase Φ_k. The signal constellation is shown in Fig. 11. The angle of a vector (or symbol) with respect to the positive direction of the horizontal axis is the symbol phase Φ_k. The symbols represented by $\bullet$ can become symbols represented only by $\times$, and vice versa. Transitions among themselves are not possible. The phase change from one symbol to the other is $\Delta\theta_k$.

Since information is carried by the phase changes $\Delta\theta_k$, differentially coherent demodulation can be used. However, coherent demodulation is desirable when higher power efficiency is required. There are four ways to demodulate a $\pi/4$-DQPSK signal: baseband differential detection, IF band differential detection, FM-discriminator detection, and coherent detection. The error probability of the $\pi/4$-DQPSK in the AWGN channel is about 2 dB inferior to that of coherent QPSK at high SNRs [3,8].

4.3. Continuous-Phase Modulation

The trend of phase modulation is shifting to *continuous-phase modulation* (CPM), which is a class of power- and bandwidth-efficient modulations. With proper choice of pulseshapes and other parameters, CPM schemes may achieve higher bandwidth and power efficiency than QPSK and higher-order MPSK schemes.

The CPM signal is defined by

$$s(t) = A\cos(2\pi f_c t + \Phi(t, \mathbf{a})), \quad -\infty \le t \le \infty \tag{28}$$

The signal amplitude is constant. Unlike signals of previously defined modulation schemes such as PSK, where signals are usually defined on a symbol interval, this signal is defined on the entire time axis. This is due to the continuous, time-varying phase $\Phi(t, \mathbf{a})$, which usually is influenced by more than one symbol. The transmitted M-ary symbol sequence $\mathbf{a} = \{a_k\}$ is embedded in the excess phase

$$\Phi(t, \mathbf{a}) = 2\pi h \sum_{k=-\infty}^{\infty} a_k q(t - kT) \tag{29}$$

with

$$q(t) = \int_{-\infty}^{t} g(\tau)\, d\tau \tag{30}$$

where $g(t)$ is a selected pulseshape. The M-ary data a_k may take any of the M values: $\pm 1, \pm 3, \ldots, \pm(M-1)$, where M usually is a power of 2. The phase is proportional to the parameter h which is called the modulation index. Phase function $q(t)$, together with modulation index h and input symbols a_k, determine how the phase changes with time. The derivative of $q(t)$ is function $g(t)$, which is the frequency shape pulse. The function $g(t)$ usually has a smooth pulseshape over a finite time interval $0 \le t \le LT$, and is zero outside. When $L = 1$, we have a full-response pulseshape since the entire pulse is in a symbol time T. When $L > 1$, we have a partial-response pulseshape since only part of the pulse is in a symbol time T.

The modulation index h can be any real number in principle. However, for development of practical maximum-likelihood CPM detectors, h should be chosen as a rational number.

Popular frequency shape pulses are the rectangular pulse, the raised-cosine pulse, and the Gaussian pulse. The well-known *Gaussian minimum shift keying* (GMSK) is a CPM scheme that uses the Gaussian frequency pulse:

$$g(t) = \frac{1}{2T}\left[Q\left(2\pi B_b \frac{t - \frac{T}{2}}{\sqrt{\ln 2}}\right) - Q\left(2\pi B_b \frac{t + \frac{T}{2}}{\sqrt{\ln 2}}\right)\right], \quad 0 \le B_b T \le 1 \tag{31}$$

where B_b is the 3-dB bandwidth of the premodulation Gaussian filter, which implements the Gaussian pulse effect [9]. GMSK is currently used in the U.S. cellular digital packet data system and the European GSM system.

If the modulation index h is made to change cyclically, then we obtain *multi-h CPM* (or MHPM); the phase is

$$\Phi(t, \mathbf{a}) = 2\pi \sum_{k=-\infty}^{\infty} h_k a_k q(t - kT) \tag{32}$$

where the index h_k cyclically changes from symbol to symbol with a period of K, but only one index is used during one symbol interval, that is, $h_1, h_2, \ldots, h_K, h_1, h_2, \ldots, h_K$, and so on. With proper choice of the index set and pulseshape, MHPM can be more power- and bandwidth-efficient than single-h CPM.

The major drawback of CPM and MHPM is the system complexity, especially the circuit complexity and the computational complexity of the demodulator, since optimum demodulation of CPM and MHPM requires complicated maximum-likelihood sequence estimation [3]. However, with the rapid advances in microelectronics and

digital signal processing power of the electronic devices, the practical implementation and use of CPM and MHPM is quickly emerging.

Significant contributions to CPM schemes, including signal design, spectral and error performance analysis were made by Sundberg, Aulin, Svensson, and Anderson, among other authors [10–13]. Excellent treatment of CPM up to 1986 can be found in the book by Anderson, et al. [13] or the article by Sundberg [10]. A relatively concise, but up-to-date, description of CPM and MHPM can be found in Ref. 3.

BIOGRAPHY

Fuqin Xiong received his B.S. and M.S. degrees in electronics engineering from Tsinghua University, Beijing, China, and a Ph.D. degree in electrical engineering from University of Manitoba, Winnipeg, Manitoba, Canada, in 1970, 1982, and 1989, respectively. He was a faculty member at the Department of Radio and Electronics in Tsinghua University from 1970 to 1984. He joined the Department of Electrical and Computer Engineering, Cleveland State University, Ohio, as an assistant professor in 1990, where he is currently a full professor and director of the Digital Communications. Research Laboratory. He was a visiting scholar at University of Manitoba in 1984–1985, visiting fellow at City University of Hong Kong, Kowloon, in spring 1997, and visiting professor at Tsinghua University in summer 1997. His areas of research are modulation and error control coding. He is the author of the best selling book *Digital Modulation Techniques*, published by Artech House, 2000. He is an active contributor of technical papers to various IEEE and IEE technical journals and IEEE conferences. He has been a reviewer for IEEE and IEE technical journals for years. He has been the principal investigator for several NASA sponsored research projects. He is a senior member of IEEE.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 2nd ed., McGraw-Hill, New York, 1989.
2. S. Haykin, *Digital Communications*, Wiley, New York, 1988.
3. F. Xiong, *Digital Modulation Techniques*, Artech House, Boston, 2000.
4. J. H. Park, Jr., On binary DPSK detection, *IEEE Trans. Commun.* **26**(4): 484–486 (April 1978).
5. B. Sklar, *Digital Communications, Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
6. S. Benedetto, E. Biglieri, and V. Castellani, *Digital Transmission Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
7. K. M. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques, Signal Design and Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
8. C. L. Liu and K. Feher, $\pi/4$-QPSK Modems for satellite sound/data broadcast systems, *IEEE Trans. Broadcast.* (March 1991).
9. G. Stüber, *Principle of Mobile Communication*, Kluwer, Boston, 1996.
10. C-E. Sundberg, Continuous phase modulation: A class of jointly power and bandwidth efficient digital modulation schemes with constant amplitude, *IEEE Commun. Mag.* **24**(4): 25–38 (April 1986).
11. T. Aulin and C-E. Sundberg, Continuous phase modulation — part I: Full response signaling, *IEEE Trans. Commun.* **29**(3): 196–206 (March 1981).
12. T. Aulin, N. Rydbeck, and C-E. Sundberg, Continuous phase modulation — Part II: Partial response signaling, *IEEE Trans. Commun.* **29**(3): 210–225 (March 1981).
13. J. B. Anderson, T. Aulin, and C-E. Sundberg, *Digital Phase Modulation*, Plenum, New York, 1986.

DISTRIBUTED INTELLIGENT NETWORKS

Iakovos S. Venieris
Menelaos K. Perdikeas
National Technical University of Athens
Athens, Greece

1. DEFINITION AND CHARACTERISTICS OF THE DISTRIBUTED INTELLIGENT NETWORK

The distributed intelligent network represents the next stage of the evolution of the intelligent network (IN) concept. The term corresponds to no specific technology or implementation but rather is used to refer, collectively, to a family of architectural approaches for the provisioning of telecommunication services that are characterized by: use of distributed, object oriented and, potentially, mobile code technologies, and a more open platform for service provisioning.

The intelligent network [1] first emerged around 1980 when value-added services that had previously been offered only on a private branch exchange basis involving mainly corporate users, first begun to be made available on the public network as well. The problem was that the only way that new services could be introduced using the infrastructure of the time required upgrading a large number of deployed telephone switches. In the first available programmable switches, services resided in the memory space of each switch and service view was therefore local. Thus, each installed service provided a number of additional features on the switch it was installed and so could be provided only for those telephone calls that were routed through that particular switch. A certain switch could have a number of services implemented in it, while others could have different sets or none. Also, there was no guarantee that the implementation of a service would behave in the same way, uniformly across switches coming from different vendors. This was the era of the "switch-based services."

New services therefore could not be provided nationwide simultaneously and often even when all involved switches had been upgraded, different service dialects existed as a result of the heterogeneity of the equipment and the unavoidable discrepancies between the implementation of the same service in switches provided by different vendors. These problems were further aggravated by the

need to guard against undesirable feature interaction problems: a process that even now cannot be automated or tackled with, in a systematic, algorithmic manner and usually involves manually checking through hundreds or even thousands of possible service feature combinations. Since this process had to also account for the different service dialects, each service that was added to the network increased the complexity in a combinatorial manner.

The result was that service introduction soon became so costly that new services could be provided only infrequently. The intelligent network idea sought to alleviate these problems. The key concept was to separate the resources (public telephony switches) from the logic that managed them. Once this was achieved, the logic could be installed in a central location called "service control point" and from there, operate on the switches under its control. Service introduction would consist simply in the installation of a new software component in the service control point and would therefore take effect immediately once this was completed and once involved switches were configured to recognize the new service prefixes. Once a switch identified an "intelligent network call," that is, a call prefixed with a number corresponding to an IN service (e.g., 0800 numbers), it would suspend call processing and request further instructions on how to proceed from the remotely located service or service logic program. No other logic therefore needed to be installed in the switches apart from a generic facility for recognizing such IN calls and participating in the interaction with the remotely located service logic programs. This interaction consisted of notifying the service logic programs of intelligent network calls or other important call events, receiving instructions from them, and putting these instructions into effect by issuing the appropriate signaling messages toward the terminals that requested the IN service or towards other resources involved in the provisioning of the service. Typical resources of the latter kind were intelligent peripherals and specialized resource functions where recorded service messages were and still are typically kept.

In order to implement a mechanism such as the one called for by the IN conception, three key artifacts are needed: (1) a finite state machine implementing an abstraction of the call resources of a public switch, (2) a remote centralized server where programs providing the algorithmic logic of a service are executed, and (3) a protocol for remote interaction between the switch and the service logic programs. Through this protocol (1) a limited aperture of visibility of the switch functionality — in terms of the said finite state machine — is presented to the service logic programs; and (2) hooks are provided allowing the latter to influence call processing in order to effect the desired behavior of any given service.

As noted above, prior to introduction of the intelligent network concept the last two of these three elements were not existent as service logic was embedded into every switch in order for a service to be provided. There was therefore no remote centralized server responsible for service execution and, hence, due to the locality of the interaction, no protocol needed to support the dialogue between the switch and the service logic program. Even the first of these three elements (the state machine abstracting call and connection processing operations in the switch) was not very formalized as switch-based services were using whatever nonstandardized programming handles a vendor's switch was exposing. This was, after all, why for a new service to be introduced, direct tampering with all affected switches was necessary, and in fact the implementation of the same service needed to be different in different switches. Also this accounted for the fact that, considering the heterogeneous nature of the switches comprising a certain operator's network, different service dialects were present, resulting in a nonuniform provision of a given service depending on the switch to which a subscriber was connected. Intelligent networks changed all that by defining a common abstraction for all switches and by centralizing service logic to a few (often one) easily administrated and managed servers. The abstraction of the switches was that of a state machine offering hooks for interest on certain call events to be registered, and supporting a protocol that allowed remote communication between the switches and the now centrally located service logic programs. It will be shown in the following paragraphs that this powerful and well-engineered abstraction is also vital in the distributed intelligent network.

2. THE NEED FOR DISTRIBUTED INTELLIGENT NETWORKS

The IN concept represents the most important evolution in telecommunications since the introduction of programmable switches that replaced the old electromechanic equipment. It allows for the introduction of new services, quickly, instantly, across large geographic areas, and at the cost of what is essentially a software development process as contrasted to the cost of directly integrating a new service in the switching matrix of each involved telephony center.

However, after 1998, a number of technical and socioeconomic developments have opened up new prospects and business opportunities and also have posed new demands, which traditional IN architectures seem able to accommodate only poorly: (1) use of mobile telephony became widespread, particularly in Europe and Japan; (2) the Internet came of age both in terms of the penetration it achieved and is projected to achieve, and also in terms of the business uses it is put to; and (3) deregulation seems to be the inevitable process globally setting network operators and carriers in fierce competition against each other. The import of the first two of these developments is that the public switched telephony network is no longer the only network used for voice communications. Cellular telephony networks and also telephony over the Internet are offering essentially the same services. Therefore, an architecture for service creation such as the IN that is entirely focused on the public telephony network falls short of providing the universal platform for service provisioning that one would ideally have wished: a single platform offering the same services over the wired, wireless and Internet components of a global integrated network for voice and data services. Deregulation, on the other hand, has two

primary effects; first, as noted above, it promotes competition between operators for market share making the introduction of new and appealing services a necessity for a carrier that wishes to stay in business. Since the basic service offered by any carrier is very much the same, a large amount of differentiation can be provided in the form of value-added or intelligent services that are appealing and useful to end users. The new telecommunications landscape is no longer homogenous but instead encompasses a variety of networks with different technologies and characteristics. The prospect for innovative services that focus not only on the telephone network but also Internet and mobile networks is immense. Particularly useful would be hybrid services that span network boundaries and involve heterogeneous media and access paradigms. Typical services belonging to this genre are "click to" services whereby a user can point to a hyperlink in his or hers browsers and as a result have a phone or fax (Facsimile) call being set up in the network. Also, media conversion services whereby a user who is not able to receive a certain, urgent email can have a phone call in his mobile terminal and listen to the contents of the mail using text to speech conversion. All these services cannot be offered by relying solely on an intelligent network. Other technologies and platforms would have to be combined and this would not result in a structured approach to service creation. Essentially, traditional intelligent networks cannot fulfill the new demands on rapid service creation and deployment in a converged Internet — telecommunications environment because the whole concept of this technology had at its focus the public telephony network and revolved around its protocols, mechanisms, and business models. Internet and cellular telephony, on the other hand, have their own protocols, infrastructure, and mechanisms, and these cannot be seamlessly incorporated into an architecture designed and optimized with a different network in mind. Furthermore, the business model envisaged by the traditional intelligent network is a closed one. It has to be said that the focus of the IN standardization was not to propose and enable new business models for telephony: only to expedite the cumbersome and costly process of manual switch upgrading that made introduction of new services uneconomical to the point of impeding the further growth of the industry. Intelligent network succeeded in solving this problem by removing the service intelligence from the switches and locating that intelligence in a few centralized points inside the network. However, the same organization continued to assume the role of the network operator or carrier and that of the service provider. Deregulation and the growth of Internet that adheres to a completely different business model necessitate the separation of these two roles and so require a technological basis that would support this new model. This is the problem that the next generation of the intelligent network technologies face.

The distributed intelligent network represents the next stage of the intelligent network evolution. The three main elements of the traditional intelligent network concept survive this evolution as could be inferred by the central role they have been shown to play in delivering services to the end users in the traditional IN context. Distributed IN is characterized by the use of distributed object technologies such as the Common Object Request Broker Architecture (CORBA), Java's Remote Method Invocation (RMI) or Microsoft's Distributed Component Object Model (DCOM) to support the switch — services interaction. In a traditional IN implementation, the Intelligent Network Application Protocol (INAP) information flows are conveyed by means of static, message-based, peer-to-peer protocols executed at each functional entity. The static nature of the functional entities and of the protocols they employ means that in turn the associations between them are topologically fixed. An IN architecture as defined in Ref. 2 is inherently centralized with a small set of service control points and a larger set of service switching points engaged with it in INAP dialogs. The service control points are usually the bottleneck of the entire architecture and their processing capacity and uptime in large extent determine the number of IN calls the entire architecture can handle effectively. Distributed object technologies can help alleviate that problem by making associations between functional entities less rigid. This is a by-product of the location transparencies that use of these technologies introduces in any context. More importantly, the fact that under these technologies the physical location of an entity's communicating peer is not manifested makes service provisioning much more open. This will be explained in more detail in later paragraphs. The remainder of this encyclopedia article is structured as follows: the distributed IN's conceptual model is introduced and juxtaposed with that of the traditional IN, typical distributed IN implementation issues are presented, and finally some emerging architectures that adhere to the same distributed IN principles are identified.

3. THE DISTRIBUTED INTELLIGENT FUNCTIONAL MODEL

According to the International Telecommunication Unions (ITU) standardization of IN, an intelligent network conceptual model is defined, layered in four planes depicting different views of the intelligent network architecture from the physical plane up to the service plane. Of these planes we will use as a means of comparing the IN architecture with that of distributed IN, the distributed functional plane. The distributed functional plane identifies the main functional entities that participate in the provision of any given IN service without regard to their location or mapping to physical elements of the network (the aggregation of functional entities to physical components is reflected in the physical plane). The functional models of IN and distributed IN are quite different in certain respects. First, a number of emerging distributed IN architectures like Parlay and JAIN (discussed later on) incorporate new functional entities to account for the new types of resources that can be found in a converged Internet — Telecommunications environment and which traditional IN models could not anticipate. Apart from that, a further difference emerges in the form of the more open environment for service provisioning, the distributed or remote-object interactions that characterize distributed IN and the potential use of mobile code technologies that can change the locality

of important components at runtime (and with this, the end points of control flows). Finally, the business model of distributed IN with its use of distributed object technologies explicitly enables and accommodates the separation between the roles of the network operator and the service provider.

Some of the abovementioned differences cannot be reflected in the distributed functional plane as defined in the IN's conceptual model as this is at a level of abstraction where such differences are hidden. Therefore, Fig. 1 compares the two technologies (traditional and distributed IN) by reflecting features of both their distributed and physical planes.

Figure 1a presents a limited view of the traditional IN conceptual model at the distributed and physical planes. A number of functional entities are there identified, of which the more central to this discussion will be the service creation environment function/service management function, the service control function, and the service switching function. Figure 1 also depicts a specialized resource function and an intelligent peripheral, which is were announcements are kept and digits corresponding to user's input are collected. The second part of Fig. 1 also depicts an abstract entity entitled "Internet/mobile resources," which represents functional entities present in distributed IN that correspond to resource types unique in the Internet or mobile networks components such as mail servers, media conversion servers or location servers.

Starting from the bottom up, the service switching function corresponds to the finite state machine reflecting the switch's call resources that the article drew attention to as being one of the main artifacts on which the whole IN concept is based. This functional entity is closely linked with the service control function, which is the place where service logic programs are executed. These are not depicted since they are thought of as incorporated with the latter. Finally, the service creation environment function and the service management function are the entities responsible for the creation of new service logic programs, and their subsequent injection and monitoring into the architecture and eventual withdrawal/superseding by other services or more up to date versions.

The INAP protocol is executed between the service switching and the service control functions that can be regarded as implementing a resource-listener-controller pattern. The service switching function is the resource that is monitored whereas the service control function is responsible for setting triggers corresponding to call events for which it registers an interest and also the controller that acts on these events when informed of their occurrence. A typical event for instance would be the detection of a digit pattern that should invoke an IN service (e.g., 0800). The service switching function undertakes the job of watching for these events and of suspending call processing when one is detected and delivering an event notification to the service control function for further instructions. This pattern is the same also in the distributed IN and can in fact also be identified in recent telecommunication architectures such as those articulated by the JAIN and Parlay groups.

The distributed IN conceptual model (Fig. 1b) differs by (1) explicitly enabling mobility of the service logic programs through the use of mobile code technologies and (2) replacing the message-based INAP protocol with distributed processing technologies like CORBA, RMI

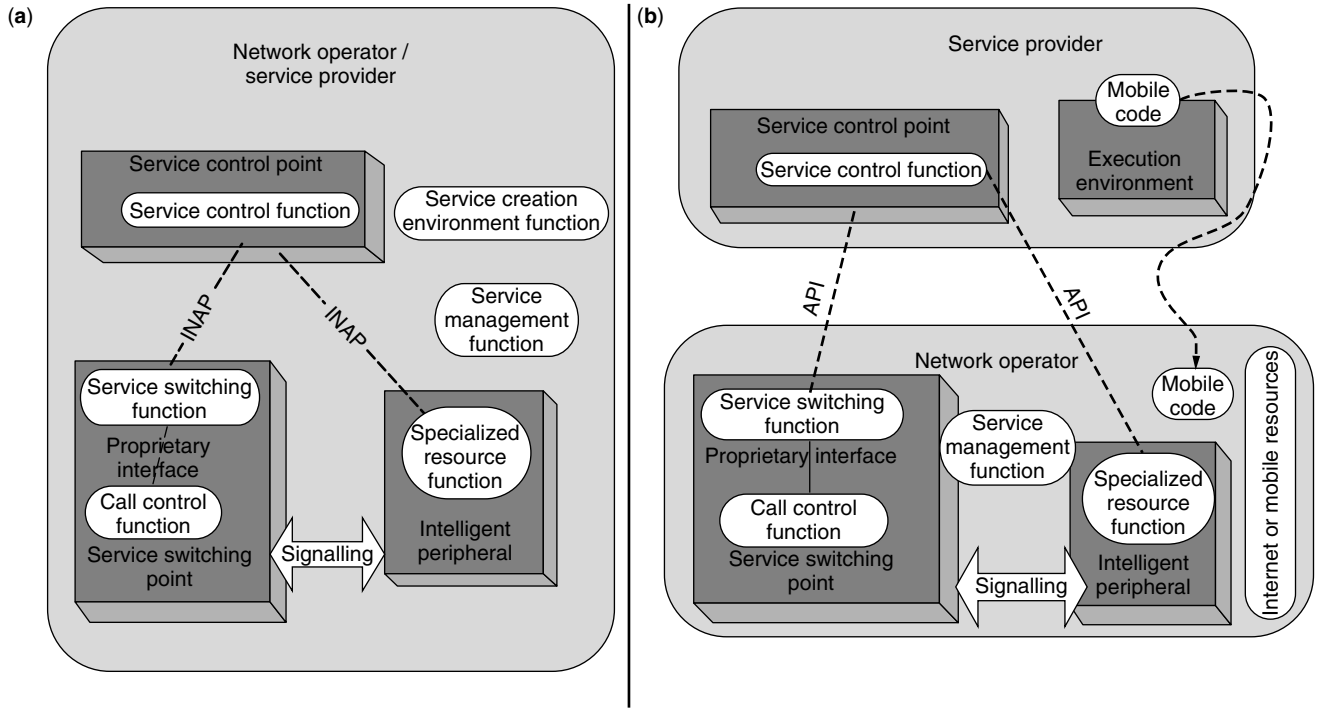


Figure 1. Traditional (**a**) and distributed (**b**) IN functional models.

or DCOM. It should be noted of course that both the traditional and the distributed IN conceptual models are populated by an additional number of functional entities which are not depicted in Fig. 1 nor enter this discussion for reasons of economy. Code mobility (often implemented using Java and/or some additional mobile code libraries) is used to allow service logic programs to reside not only in the service control function but also in the service switching function where an execution environment identical with the one that exists within the service control function can be found. Service logic programs in the form of mobile code components populate both execution environments and constitute the control part of IN service provisioning. From these execution environments, using distributed processing technologies, the switch resources, as exposed by the service switching function objects, are monitored and manipulated. This means that service logic programs in the distributed IN conceptual model are prime level entities and cannot be suitably depicted as pinned down inside the implementation of the service control function. In contrast to the functional plane of the IN conceptual model where the service control function and the service logic programs were one and the same, in the functional plane of the distributed IN conceptual model, the service control and also the service switching functions are containers (or execution environments) for service components capable of migrating to whichever environment is best suited to accommodate their execution. It is important to consider that this amount of flexibility would not be attainable were it not for the use of distributed processing technologies at the control plane. The defining characteristic of these technologies is that they abstract process, machine and network boundaries and provide to the programmer and the runtime instances of service code a view of the network as a generalized address space spanning conventional boundaries. In the same manner in which a program may hold a pointer to an object or function residing in the same local process, so it can, when distributed technologies are used, hold a pointer to an object located in a different process in the same machine, in a machine in the same local network, or in a machine located in a different network. Therefore the abstraction of a single memory space is supported and compiled code components can be mapped in an arbitrary way in processes and physical nodes without needing recompilation or rebooting of the system. This is a potent facility and it opens new capabilities which the distributed IN employs. It also has to be stated that the articulation of the traditional IN conceptual model, to an extent, anticipated a certain flexibility in the allocation of functional entities to physical ones according to the correspondences between the functional and the physical planes. This was however not materialized since the communication infrastructure used for the conveyance of the INAP information flows between the various entities of the system (and most importantly between the service switching and the service control entities) was the signaling system 7 network, which does not demonstrate the properties of a distributed processing environment. Therefore the full potential inherent in the abstract definition of the IN conceptual model was materialized only to a limited extent and mainly had to do with different configuration and aggregations of functional entities to the hardware nodes of the system. The interested reader should refer to Ref. 2 for a presentation of the various alternatives that are possible in an IN architecture. This leeway afforded to the network designer is something completely different from the ability of service logic programs to roam through the various execution environments of the system at runtime without requiring the system to suspend its operation and, under certain configuration options discussed in the following paragraph, in a manner completely automated and transparent even to the network management system.

4. ISSUES IN DISTRIBUTED IN IMPLEMENTATIONS

Given the considerations outlined above and the differences at the functional plane between traditional and distributed IN architectures, a number of approaches exist. Each of these approaches essentially answers a defining question in a different manner, and since for the most part they are orthogonal to each other, a large number of widely differing distributed IN implementations can be envisaged; all, however, share the same fundamental characteristics of increased flexibility and openness when compared with traditional IN.

The first point to examine is the set of considerations that govern the location of service logic programs inside the network. Attendant to it are the dynamics of their mobility. This is possible once employment of mobile code technologies is assumed. Service logic programs implemented as mobile components can migrate between execution environments dynamically. Considerations valid for incorporation in any given distribution algorithm are processing load, signaling load, and functionality. Given a number of execution environments present in the network's service switching and service control functions, each with its own characteristics, simple or elaborate load balancing mechanisms can be devised and implemented. These would allow service logic programs to locate themselves in such a way so that no individual execution environment's resources are strained beyond a certain point. Apparent tradeoffs exist with respect to complexity, time, and computing power spent in process or code migrations, which are in themselves costly procedures. This cost is reflected in terms of both time and processor load they temporarily create as they involve suspension or creation of new threads, allocation of objects in memory, and so on. These overheads, when also viewed in the light of the stringent performance and responsiveness demands that exist in a telecommunication system, should tilt the balance in favor of a simple and not very reactive load balancing algorithm that operates on ample time scales.

Processing load is not however the only kind of load to consider. Control load is another. Control load should not be confused with signaling load, which concerns protocol message exchanges between terminals and switches at various levels of the classic Five-layered telephony network architecture. Signaling load is for the most part transparent to the IN with the exception of those call events that invoke an IN service and for

which a service switching function has been asked to suspend call processing and wait for instructions from the service control function. Control load in this discussion is about the exchange of INAP information flows using the selected distributed processing technology between the service logic programs and the switching resources of the system. The location transparency provided by distributed technologies cannot clearly be interpreted to suggest that local or remote interactions take the same time. Therefore, when service logic programs are located in an execution environment closer to the switch they control, better performance can be expected. This means that, in general, the code distribution mechanism, should locate service logic programs at the proximity of the resources with which they are engaged if not at an execution environment collocated with those resources. There is, however, a tradeoff here between processing and control load. Because of the higher level of abstraction (method calls instead of messages) offered to service logic programs when distributed processing technologies are used, it is necessary that, under the hood, the respective middleware that is used to support this abstraction enters into some heavy processing. For every method call that is issued, it is necessary that each argument's transitive closure is calculated and then the whole structure is serialized into an array of bytes for subsequent transmission in the form of packets using the more primitive mechanisms offered by the network layer. On the recipient side, the reverse procedure should take place. This set of processes is known as marshaling/demarshaling and is known to be one of the most costly operations of distributed processing. Because of the processor intensive character of marshaling/demarshaling operations, it is possible that, under certain configurations, better performance is attained when the service logic program is located to a remote execution environment than to a local, strained, one. This in spite of the fact that the time necessary for the propagation of the serialized byte arrays in the form of packets from the invoking to the invoked party will be higher in the remote case. This interrelation between processing and control load means that no clear set of rules can be used to produce an algorithm that is efficient in all cases. The complex nature of the operations that take place in the higher software layers means that this is not an optimization problem amenable to be expressed and solved analytically or even by means of simulation, and therefore an implementor should opt for simple, heuristic designs, perhaps also using feedback or historic data. Another approach would be to rely on clusters of application servers into which service logic programs can be executed. Commercially available application servers enable process migration and can implement a fair amount of load balancing operations themselves, transparently to the programmer or the runtime instances of the service components.

The second point that can lead to differentiations in architecture design in distributed intelligent network is the question of who makes the abovementioned optimization and distribution decisions. There could be a central entity that periodically polls the various execution environments receiving historic processing and signal load data, runs an optimization algorithm and instructs a number of service logic programs to change their location in the network according to the results thus produced. The merit of this approach is that the optimization is networkwide and the distribution algorithm can take into account a full snapshot of the network's condition at any given instance of time when it is invoked. The disadvantage is the single point of failure and the cost of polling for the necessary data and issuing the necessary instructions. These communications could negatively affect traffic on the control network, depending, of course, on the frequency with which they are executed. A point to consider in this respect that can lead to more efficient implementations is whether networkwide optimization is necessary or whether locally executed optimizations could serve the same purpose with similar results and while projecting a much lower burden on the communication network. Indeed, it can be shown that it is highly unlikely for the purposes of any optimization to be necessary to instruct any given service logic program to migrate to an execution environment that is very remote to the one in which it was until that time executed. To see that this is the case, one can consider that a migration operation moving a service component to a very distant location will most likely result in an unacceptably higher propagation delay that would degrade the responsiveness of the corresponding resource–service logic link. Therefore locally carried optimizations could result in comparable performance benefits at a greatly reduced communication cost when compared to a networkwide optimization. Therefore, in each subnetwork an entity (migration manager) can be responsible for performing local optimizations and instructing service components to assume different configurations accordingly. Taking this notion to its logical conclusion, a further implementation option would be to have service logic programs as autonomous entities (mobile agents) that are responsible for managing their own lifecycle and proactively migrating to where they evaluate their optimal location to be. As before, considerations for deriving such an optimal location can be signaling or control load experienced at the execution environment where they were hosted but also, more appropriately in this case, the location of mobile users. Indeed, in this last scenario one can have a large population of service code components, each instantiated to serve a specific mobile (roaming) user. See Ref. 3 for a discussion of this approach. Service logic programs could then evaluate their optimal location in the network, also taking into consideration the physical location of the user they serve. This can, for instance, lead to migration operations triggered by the roaming of a mobile user. In this manner concepts like that of the virtual home environment for both terminal and personal mobility can be readily supported enabling one user to have access to the same portfolio of intelligent services irregardless of the network into which he/she is roaming and subject only to limitations posed by the presentation capabilities of the terminal devices he/she is using. Of course, the usefulness of autonomous service logic programs is not limited to the case of roaming or personal mobility

users, but this example provides an illustration of the amount of flexibility that becomes feasible when advanced software technologies such as distributed processing environments and mobile code are used in the context of an intelligent network architecture. Moreover, as more logic is implemented in the form of mobile code components, the supporting infrastructure can become much more generic, requiring fewer modifications and configuration changes and able to support different service paradigms. The generic character of the architecture means that stationary code components that require human individual or management intervention for their installation or modification are less likely to require tampering with, and that a greater part of the service logic can be deployed from a remote management station dynamically, at runtime, contributing to the robustness of the system and to an increased uptime. Naturally, the tradeoff is that making service logic programs more intelligent necessarily means increasing their size, straining both the memory resources of the execution environments into which they are executed and also requiring more time for their marshaling and demarshaling when they migrate from one functional entity to another. If their size exceeds a certain threshold, migration operations may become too costly and thus rarely triggered by the load balancing algorithm they implement, negating the advantages brought by their increased autonomy and making services more stationary and less responsive in changing network or usage conditions. Again, simpler solutions may be more appropriate and a designer should exercise caution in determining which logic will be implemented in the form of stationary code components, engraved in the infrastructure, and which will be deployed at runtime using mobile code.

Another major issue to consider in the implementation of a distributed intelligent network architecture is which distributed processing technology to use and how to support its interworking with the signaling system 7 network that interconnects the physical components hosting the functional entities in any intelligent network architecture. There are three main competing distributed technologies: CORBA, DCOM, and RMI. It is not the purpose of this article to examine or compare these three technologies, nor to identify their major strengths and weaknesses. However, for the purposes of an intelligent network implementation, CORBA is best suited because it is the only one of these three technologies to have interworking between it and the signaling system 7 network prescribed [4]. The alternative would be to bypass the native IN control network and install a private internet over which the traffic from DCOM or RMI could be conveyed. Since this is not a development that operators can be expected to implement quickly and for reasons of providing a solution that has the benefit of backward compatibility and also allows an evolutionary roadmap to be defined, CORBA should be the technology of choice in distributed IN implementations. A further reason is that CORBA, which is targeted more than the other two technologies for the telecommunications domain, has more advanced real time characteristics and is therefore better equipped to meet the demands of an operator. The interested reader is referred to Ref. 5 for an in-depth look of the implementation of a distributed IN architecture using CORBA.

5. A DISTRIBUTED IN ARCHITECTURE IN DETAIL

Figure 2 depicts a typical distributed IN implementation in more detail. The main components are (1) the network resources that provide the actual call control and media processing capabilities, (2) the remotely enabled distributed objects that wrap the functionality provided by the physical resources and expose it as a set of interfaces to the service layer, (3) the code components residing in the service layer that incorporate the business logic, and (4) the communication infrastructure that makes communication between the service logic programs and the network resources, via their wrappers, possible. Of course, in an actual architecture a lot more components would need to be identified, such as service management stations, and service creation systems. However, the purpose of the discussion presented here is to allow more insight, where appropriate, into the more technical aspects of a DIN implementation and not to provide a full and comprehensive description of a real system.

Notice, first, that terminals are not depicted in Fig. 2. This is because interaction with terminals at the signaling level is exactly the same as in the case of traditional IN. Signaling emitted by the terminals is used to trigger an IN session and signaling toward terminals or media resources of an IN system (like Intelligent peripherals or specialized resource functions) is used to carry out the execution of a service. However, service components do not directly perceive signaling as they interact with their peer entities on the network side by means of message- or method-based protocols. INAP is a typical method based protocol used at the control plane of IN and Parlay, JAIN, or a version of INAP based on remote application programming interfaces (APIs) are prime candidates for the control plane of distributed IN. In any case it is the responsibility of the network resources to examine signaling messages and initiate an IN session where appropriate, or, in the opposite direction, to receive the instructions sent to them by the service logic programs and issue the appropriate signaling messages to bring these instructions into effect. Therefore, since signaling is conceptually located 'below' the architecture presented in Fig. 2, it can safely be omitted in the discussion that follows. The reader who wishes to gain a fuller appreciation of the temporal relationships between the signaling and INAP messages that are issued in the course of the provisioning of an IN session can refer to Ref. 5.

Assuming a bottom–up approach, the first step would be to explore the programmability characteristics of the deployed equipment that forms the basis of a distributed IN architecture. We refer to deployed equipment as *distributed IN*, like IN before it, which has to be able to encompass and rely on equipment that is already deployed in the field if it is to succeed. Telecom operators have made huge investments in building their networks, and it would be uneconomical to replace all this equipment. At the very least, an evolutionary approach should

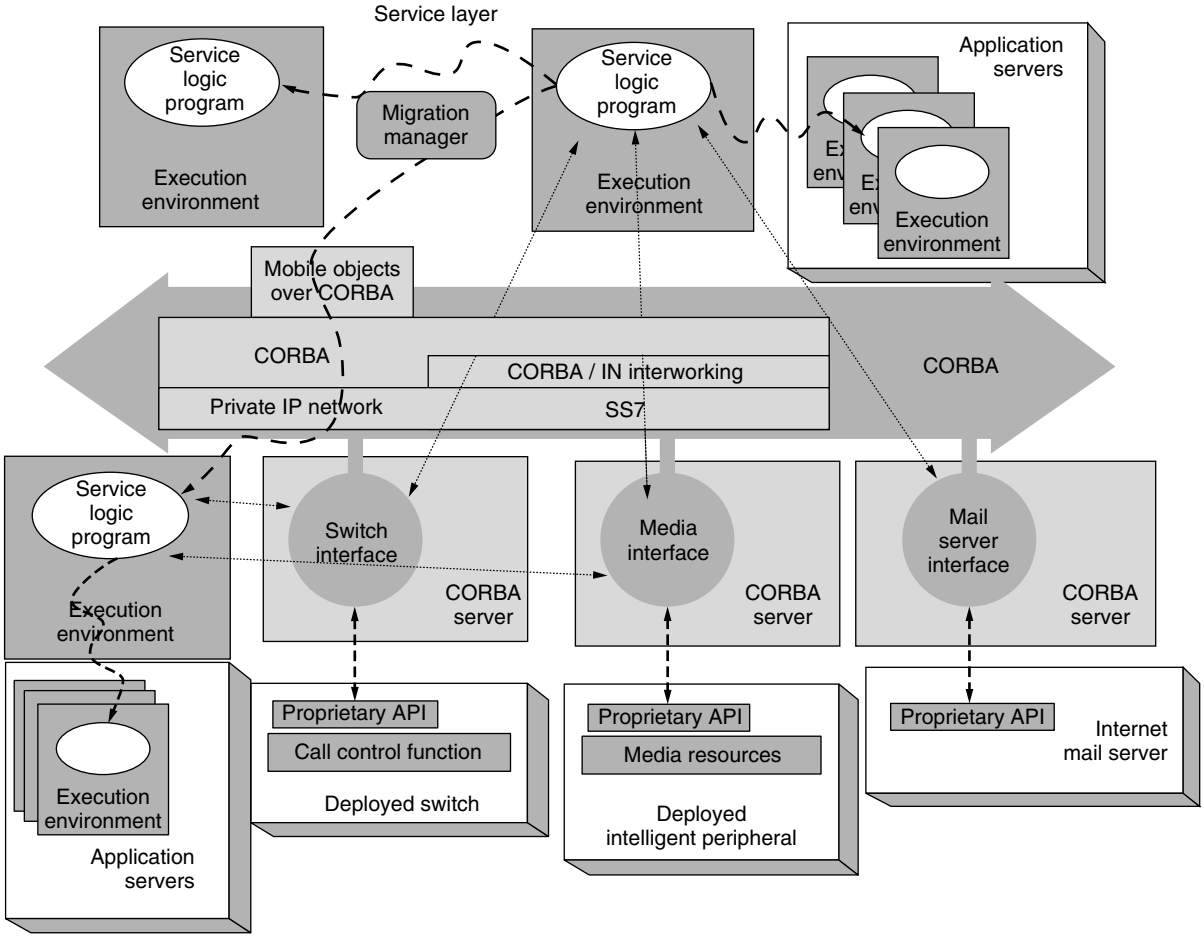


Figure 2. A distributed IN architecture in more detail.

be possible to implement where equipment would be replaced gradually over a period of years and the two technologies, IN and distributed IN, would coexist. It would be further advantageous if existing equipment could be seamlessly integrated into the new distributed intelligent network infrastructure. To address this issue we note that in general, network switches are not meant to be overly programmable. Certain proprietary APIs are always offered that provide an amount of control over how the switch responds to an incoming call, but there are few specifications defining standard interfaces that network switches should offer, and in any case there is much discrepancy between the equipment provided by different vendors. Of course, the IN itself could be regarded as a standard interface for network switches, but the purpose here is not to build the new distributed IN infrastructure over the already existing IN one. Rather, it is to build it on top of the same lower-level facilities and services that IN is using as part of its own implementation: and such lower-level facilities are accessible only to equipment vendors. An external integrator exploring the programmability features of a switch will be able to find standard interfaces only at the IN level, which, as we noted, is conceptually very elevated to be useful as the infrastructure of a distributed IN architecture (note, nevertheless, that Parlay, which is one of a group of emerging technologies that can fall under the "distributed" IN heading, also allows for an approach of using existing IN interfaces as an interim solution before it can be provided natively in switches as IN does). Below this level one, can find only proprietary APIs at different levels of abstraction.

It is therefore possible that a certain operation that a developer would wish to expose to the service components cannot be implemented because the programmability characteristics of the switch would not allow the programming of this behavior. Assuming however that the semantic gap between the operations that a network resource needs to make available for remote invocation and the programming facilities that are available for implementing these operations, can be bridged, generally accepted software engineering principles, object-oriented or otherwise could serve to provide the substrate of the distributed IN [6]. This substrate consists of set of remotely enable objects (CORBA objects "switch interface," "media interface," and "mail server interface" depicted in Fig. 2) that expose the facilities of the network resources they represent. Once this critical implementation phase has been carried out, subsequent implementation is entirely at the service layer. The "switch interface," "media

interface," and "mail server interface" objects are each responsible for exposing a remotely accessible facet of the functionality of the network resource they represent. Their implementation has then to mediate the method invocations it receives remotely, to the local proprietary API that each network resource natively provides. In the opposite direction, one has events that are detected by the network resources and have, ultimately, to reach the components residing in the service layer. This is accomplished by having the CORBA objects register themselves as listeners for these events. The process of registering an external object as a listener involves identifying the events to which it is interested and providing a callback function or remote pointer that will serve to convey the notification when an event satisfying the criteria is encountered. From that point on, the implementation of the CORBA object will itself convey the notification to the higher software layers. It is interesting to note that at this second stage the same pattern of listener and controller objects is also observed, with the exception that now the listener objects are the service logic programs or, in general, the service control logic in the network and the resource is the CORBA object itself. This observation is depicted at Fig. 3.

Through this wrapping of the native control interface to remote API-based ones two things are achieved: (1) the middleware objects residing at the CORBA servers in the middle tier of Fig. 3 can be used to implement standardized interfaces. As an example, Fig. 3 indicates INAP or Parlay API. INAP is, of course, a message-based protocol, but it is relatively straightforward to derive method — based interfaces from the original protocol specification and to express the protocol semantics in terms of method calls and argument passing instead of asynchronous exchange of messages and packets with their payloads described in Abstract Syntax Notation 1. In that sense, API-based versions of INAP can be used to implement the control plane of the distributed intelligent network, and so can, for that matter, emerging technologies such as Parlay and JAIN. The next section discusses such approaches.

6. EMERGING ARCHITECTURES

A number of architectures and most notably JAIN and Parlay have emerged that, although not classified under the caption of "distributed intelligent network," have nevertheless many similarities with it. Distributed intelligent networks make telecommunications service provisioning more open by resting on distributed object technologies and utilizing software technologies such as mobile code, which make the resulting implementation more flexible and responsive to varying network conditions. Parlay and JAIN move a step further toward this direction again by leveraging on CORBA, DCOM, or RMI to implement a yet more open service model that allows the execution of the services to be undertaken by actors different than the operator's organization, in their own premises closer to the corporate data they control and manipulate [7]. In particular, Parlay and JAIN use the same approach as in Ref. 5 of defining methods corresponding more or less to actual INAP information flows and of exposing a switch's (or, for that matter, any other network resident resources)

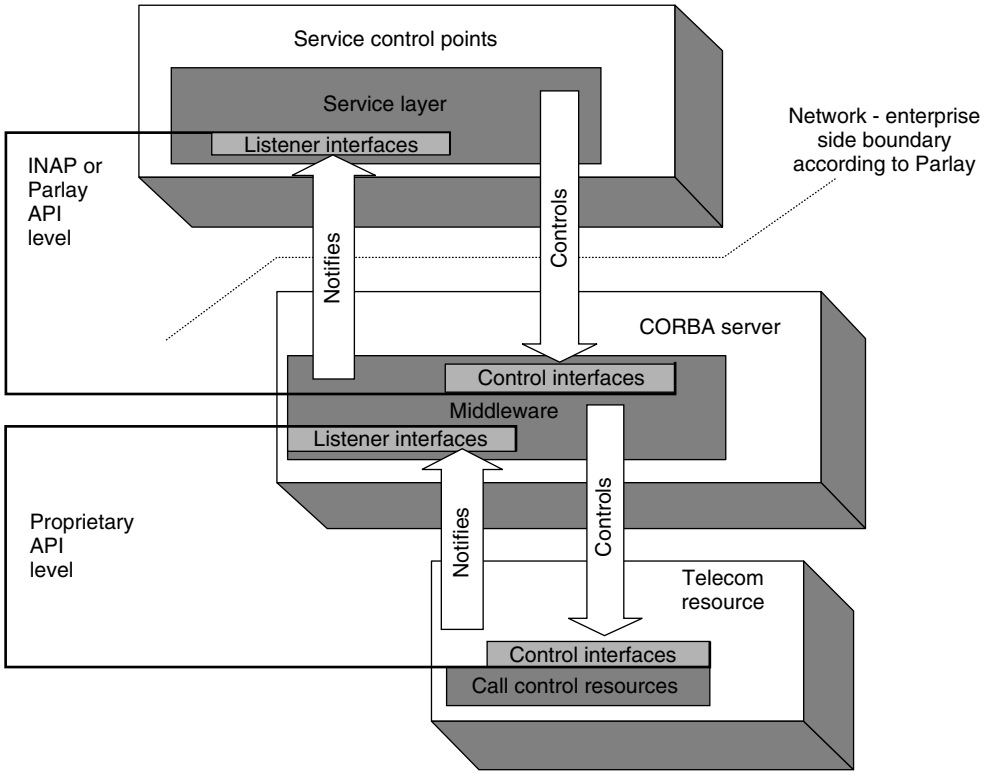


Figure 3. Parlay or other remote API implemented in middleware form over a native resource.

functionality to remotely located service code in the form of an interface of available methods. Moreover, a number of technologies such as the call processing language or telephony services expressing their interaction with the switch in the form of eXtensible Markup Language (XML) statements have appeared that can be used in the same manner as mobile code is used in a distributed IN environment. Service logic expressed in XML or Call Processing Language is inherently mobile in that it is not compiled code and can be interpreted by any appropriate execution environment or script engine. Such engines could then use INAP (message- or method-based) or Parlay and JAIN to interact with actually deployed equipment. Again, this approach enhances the open characteristics of the foreseen telecommunications environment by allowing even the service users themselves to code their services and injecting them into the network exactly in the same manner that mobile service components of the distributed IN where injected into the network. In either case, code mobility is exploited. In the call processing language or XML case, mobility accrues due to the interpreted or scripted nature of the code, in the distributed IN case, due to the semiinterpreted characteristics of the Java language and the Java virtual machine architecture. Service users could program their services themselves since the call processing language expresses telephony services logic at a very simple and abstract level relying on the actual execution engine for the translation of this logic to the appropriate set of INAP or Parlay commands. Alternatively, a graphical front end could be used where elementary building blocks are arranged in a two-dimensional canvas and then the appropriate call processing language or XML code is generated. This is not different from the service creation environments used in IN (both the traditional and the distributed ones), with the exception that such tools were very elaborate and expensive, requiring trained personnel to use them, whereas graphical front ends such as those described above can be simple and inexpensive as they need only produce a set of call processing language statements and not actual code.

7. CONCLUSIONS

The distributed intelligent network encompasses a wide range of architectures each adopting different implementation options and exhibiting different tradeoffs among the various properties that characterize such a system. The common elements in all such systems are the use of distributed processing technologies for the control plane (either INAP in the form of CORBA methods or Parlay/JAIN), mobile service code (Java components or interpreted call processing language or XML statements), and a more open service provisioning model (services executed at the premises of their users, directly managed and maintained or even coded by them).

Distributed IN thus represents the next stage in the evolution of traditional IN architectures that witnesses a new business model with the network operator concentrating on the role of the telecommunication infrastructure provider while more and more of the service provisioning role is assumed by external actors.

BIOGRAPHIES

Iakovos S. Venieris (venieris@cs.ntua.gr) was born in Naxos, Greece, in 1965. He received a Dipl.-Ing. from the University of Patras, Greece in 1988, and a Ph.D. from the National Technical University of Athens (NTUA), Greece, in 1990, all in electrical and computer engineering. In 1994 he became an assistant professor in the Electrical and Computer Engineering Department of NTUA where he is now an associate professor. His research interests are in the fields of broadband communications, Internet, mobile networks, intelligent networks, signaling, service creation and control, distributed processing, agents technology, and performance evaluation. He has over 150 publications in the above areas and has contributed to standardization bodies (ETSI, ITU-T, OMG, and IETF). He has participated in several European Union and national R&D projects. He is an associate editor of the *IEEE Communication Letters*, member of the editorial board of *Computer Communications* (Elsevier), and has been a guest editor for *IEEE Communications Magazine*. He is a reviewer for several journals and has been member of the technical program committee and session chairman of several international conferences. Dr. Venieris is the editor and coauthor of two international books on *Intelligent Broadband Networks* (Wiley 1998) and *Object oriented Software Technologies in Telecommunications* (Wiley 2000).

Dr. Ing. Menelaos Perdikeas (mperdikeas@semantix.gr) was born in Athens, Greece, in 1974 and received a Dipl.-Ing. in computer engineering and informatics (Summa Cum Laude) from the University of Patras, Greece, in 1997 and a Ph.D. in telecommunications engineering from the National Technical University of Athens in 2001. He has over 25 publications in international journals and conferences and has received a number of national and international distinctions, among which, the "D. Chorafas Award for Academic Achievement" for the year 2001. He is coauthor of the book *Object Oriented Software Technologies in Telecommunications* published by Wiley. His interests include intelligent network systems, service provisioning architectures using distributed object technologies, and the application of aspect-oriented programming methodologies and patterns in the development of telecommunications software. Dr. Perdikeas is among the founders of Semantix S. A., a Greek software company.

BIBLIOGRAPHY

1. I. S. Venieris and H. Hussman, eds., *Intelligent Broadband Networks*, Wiley, 1998.
2. ITU-T, *Distributed Functional Plane for Intelligent Network Capability Set 2*, Recommendation Q.1224.
3. M. Breugst and T. Magedanz, Impacts of mobile agent technology on mobile communications system evolution, *IEEE Pers. Commun. Mag.* **5**(4): (Aug. 1998).
4. Object Management Group Telecom Specifications, *Interworking between CORBA and Intelligent Network Systems*, *http://www.omg.org.*

5. F. Chatzipapadopoulos, M. Perdikeas, and I. Venieris, Mobile agent and CORBA technologies in the broadband intelligent network, *IEEE Commun. Mag.* **38**(6): 116–124 (June 2000).

6. I. Venieris, F. Zizza, and T. Magedanz, eds., *Object Oriented Software Technologies in Telecommunications*, Wiley, 2000.

7. M. Perdikeas and I. Venieris, Parlay-based service engineering in a converged Internet-PSTN environment, *Comput. Networks* **35**(6): 565–578 (2001).

DIVERSITY IN COMMUNICATIONS

MOHSEN KAVEHRAD
Pennsylvania State University
University Park, Pennsylvania

1. INTRODUCTION

In designing a reliable communication link, the system must be planned around the chosen transmission medium referred to as the *channel*. The disturbances of the medium must be taken into account in the process of encoding the signal at the transmitting end, and in the process of extracting the message from the received waveform at the receiving end. Once a satisfactory characterization of the anticipated channel disturbances has been made, the message encoding chosen for transmission must be designed so that the disturbances will not damage the message beyond recognition at the receiving end. With a corrupted message at hand, the receiving system must be prepared to operate continuously in the presence of the disturbances and to take maximum advantage of the basic differences between the characteristics of messages and of disturbances.

In this article, we assume that the encoded form in which the message is to be transmitted has been selected, and that the encoded form has been translated into a radiofrequency (RF) or lightwave signal by an appropriate modulation technique, such as by varying some distinguishable parameter of a sinusoidal carrier in the RF or optical frequency spectrum. Improvements in system performance can be realized only through the utilization of appropriate corrective signal processing measures. Of primary interest here will be what is widely known as *diversity techniques* [1,2] as countermeasures for combating the effects of loss of received signal energy in parts or over its entire transmission bandwidth, termed as signal fading.

In many practical situations, one seeks economical ways of either transmitting and/or receiving signals in such a way that the signal is never completely lost as a result of transmission disturbances. This has been traditionally the case, in particular, in wireless communications. Ideally, one would like to find transmission methods that are negatively correlated in the sense that the loss of signal in one channel is offset by the guaranteed presence of signal in another channel. This can occur in some diversity systems, such as those that utilize antennas at different elevations in order to minimize the received signal loss of energy. Also, in Section 4.3 of this article the same scenario (negative correlation) applies. In a way, expert investment firms, claiming to provide a diversified portfolio to investors, try to do the same. They opt for capital investment from among those economy sectors whose mutual fund return values are to some degree negatively correlated or at least fluctuate, independently from one sector to another. Consequently, over a long time period, there will be a net gain associated with the diversified portfolio.

The principles of diversity combining have been known in the radio communication field for decades; the first experiments were reported in 1927 [1]. In diversity transmission techniques, one usually settles for fluctuations of signal transmissions over each channel that are more or less uncorrelated with those in other channels and the simultaneous loss of signal will occur rarely over a number of such channels. In order to make the probability of signal loss as low as possible, an effort is made to find many channels that are either statistically independent or negatively correlated. This may be performed over the dimensions of time, frequency and space in a wireless system. For this purpose, it is occasionally possible to use two different polarizations on the receiving antennas, or receivers at several different angles of arrival for the electromagnetic wavefront, or to place antennas in several different spatial locations (spatial diversity) or to transmit the signal over several widely separated carrier frequencies or at several widely separated times (time diversity). The term "diversity improvement" or "diversity gain" is commonly employed to describe the effectiveness of various diversity configurations. There is no standard definition for the effectiveness of diversity reception techniques. One common definition is based on the significance of diversity in reducing the fraction of the time in which the signal drops below an unusable level. Thus, one may define an *outage rate* at some specified level, usually with respect to the mean output noise level of the combiner of the diversity channel outputs. In the rest of this section, the problem of correlation among such multiport channels is discussed.

Assume that a number of terminal pairs are available for different output signals $y_i(t)$ from one or more input signals $x_j(t)$. When frequency (or time) diversity is used, there is no mathematical distinction between a set of multi-terminal-pair channels, each centered on the different carriers (or different times) and a single channel whose system function encompasses all frequencies (or all times) in use. In practice, since one may use physically different receivers, the use of separate system functions to characterize the outputs from each receiver is useful. If space diversity is used, one may be concerned with the system function that depends on the spatial position as a continuous variable.

The cross-correlation function between the outputs of two diversity channels when the channels are both excited by the same signal $x_j(t)$ is fully determined by a complete knowledge of the system functions for each channel alone. In view of the random nature of the channels, the most that can be done to provide this knowledge in practice is to determine the joint statistical properties of the channels.

The signal diversity techniques mentioned above do not all lead to independent results. One must, therefore, recognize those diversity techniques that are dependent in order to avoid trying to "squeeze blood out of a bone." For example, one cannot apply both frequency *and* time diversity to the same channel in any wide sense. In fact one might argue that complete distinction between frequency and time diversity may be wholly artificial, since the signal designer usually has a given time–bandwidth product available that can be exploited in conjunction with the channel characteristics. Another pair of diversity channels that are not necessarily independent of each other are distinguished as angular diversity and space diversity channels. Consider an array of n isotropic antennas that are spaced a sufficient distance apart so that the mutual impedances between antennas can be ignored. If the transmission characteristics of the medium are measured between the transmitting-antenna terminals and the terminals of each of the n receiving antennas, then n channel system functions will result, each of which is associated with one of the spatially dispersed antennas. If the antenna outputs are added through a phase-shifting network, the resultant array will have a receiving pattern that can be adjusted by changing the phase shifting network to exhibit preferences for a variety of different angles of arrival. The problem of combining the outputs of the array through appropriate phase shifters, in order to achieve major lobes that are directed at favorable angles of arrival would be considered a problem in angular diversity, while the problem of combining the outputs of the elements in order to obtain a resultant signal whose qualities are superior to those of the individual outputs is normally considered as the problem of space diversity combining, yet both can lead to the same end result. Signal diversity techniques and methods of combining signals from a number of such channels are discussed in the next section.

2. DIVERSITY AND COMBINING TECHNIQUES

Diversity is defined here as a general technique that utilizes two or more copies of a signal with varying degrees of disturbance to achieve, by a selection or a combination scheme, a consistently higher degree of message recovery performance than is achievable from any one of the individual copies, separately. Although diversity is commonly understood to aim at improving the reliability of reception of signals that are subject to fading in the presence of random noise, the significance of the term will be extended here to cover conceptually related techniques that are intended for other channel disturbances.

The first problem in diversity is the procurement of the "diverse" copies of the disturbed signal, or, if only one copy is available, the operation on this copy to generate additional "diversified" copies. When the signal is disturbed by a combination of multiplicative and additive disturbances, as in the case of fading in the presence of additive random noise, the transmission medium can be tapped for a permanently available supply of diversity copies in any desired numbers.

Propagation media are generally time-varying in character, and this causes transmitted signals to fluctuate randomly with time. These fluctuations are usually of three types:

1. Rapid fluctuations, or fluctuations in the instantaneous signal strength, whose cause can be traced to interference among two or more slowly varying copies of the signal arriving via different paths. This may conveniently be called *multipath fading*. If the multiple paths are resolved by the receiver [2], fading is called *frequency-selective*. Otherwise, it is called *single-path* (flat) *fading*. This type of fading often leads to a complete loss of the message during time intervals that are long even when compared with the slowest components of the message. It is observed, however, that if widely spaced receiving antennas are used to pick up the same signal, then the instantaneous fluctuations in signal-to-noise ratio (SNR) at any one of the receiving sites is almost completely independent of the instantaneous fluctuations experienced at the other sites. In other words, at times when the signal at one of the locations is observed to fade to a very low level, the same signal at some other sufficiently distant site may very well be at a much higher level compared to its own ambient noise. This type of variation is also referred to as *macrodiversity*. Signals received at widely spaced time intervals or widely spaced frequencies also show almost completely independent patterns of instantaneous fading behavior. Nearly uncorrelated multipath fading has also been observed with signal waves differing only in polarization. It will be evident that by appropriate selection or combining techniques, it should be possible to obtain from such a diversity of signals a better or more reliable reception of the desired message than is possible from processing only one of the signals all the time.

2. The instantaneous fluctuations in signal strength occur about a mean value of signal amplitude that changes relatively so slowly that its values must be compared at instants separated by minutes to hours before any significant differences can be perceived. These changes in short-term (or "hourly") mean signal amplitude are usually attributable to changes in the attenuation in the medium that the signals will experience in transit between two relatively small geographic or space locations. No significant random spatial variations in the received mean signal amplitude are usually perceived in receiving localities that could be utilized for diversity protection against this attenuation fading or, as sometimes called, "fading by shadowing". However, it is possible to combat this type of fading by a feedback operation in which the receiver informs the transmitter about the level of the received mean signal amplitude, thus "instructing" it to radiate an adequate amount of power. But the usual practice is to anticipate the greatest attenuation to be expected at the design stage and counteract it by appropriate antenna design and adequate transmitter power.

3. Another type of attenuation fading is much slower than that just described. The "hourly" mean signal levels are different from day to day, just as they are from hour to hour in any single day. The mean signal level over one

day changes from day to day and from month to month. The mean signal level for a period of one month changes from month to month and from season to season, and then there are yearly variations, and so on. As in the case of the "hourly" fluctuations in paragraph 2, the long-term fluctuations are generally caused by changes in the constitution of the transmission medium, but the scale and duration of these changes for the long-term fluctuations are vastly greater than those for the "hourly" changes. Diversity techniques per se are ineffective here.

In addition to the instantaneous-signal diversity that can be achieved by seeking two or more separate channels between the transmitting and receiving antennas, certain types of useful diversity can also be achieved by appropriate design of the patterns of two or more receiving antennas placed essentially in the same location (microdiversity), or by operations in the receiver on only one of the available replicas of a disturbed signal. The usefulness of "receiver diversity" of a disturbed signal will be demonstrated in examples of the next section. The application discussed in Section 4.1 demonstrates the use of diversity (under certain circumstances) from a delayed replica of the desired signal arriving via a different path of multiple fading paths. The latter is referred to as *multipath diversity*, where the same message arrives at distinct arrival times at a receiver equipped to resolve the multipath into a number of distinct paths [3] with different path lengths. The example presented in Section 4.2 shows application of diversity for the case in which the interference from some other undesired signal source is the cause of signal distortion.

The second problem in diversity is the question of how to utilize the available disturbed copies of the signal in order to achieve the least possible loss of information in extracting the desired message. The techniques that have thus far been developed can be classified into (1) switching, (2) combining, and (3) a combination of switching and combining. These operations can be carried out either on the noisy modulated carriers (predetection) or on the noisy, extracted modulations that carry the message specifications (postdetection). In any case, if K suitable noisy waveforms described by $f_1(t), f_2(t), \ldots, f_k(t)$ are available, let the kth function be weighted by the factor a_k, and consider the sum

$$f(t) = \sum_{k=1}^{K} a_k f_k(t) \tag{1}$$

In the switching techniques only one of the a_k values is different from zero at any given time. In one of these techniques, called *scanning diversity*, the available waveforms are tried one at a time, in a fixed sequence, until one is found whose quality exceeds a preset threshold. That one is then delivered for further processing in order to extract the desired message, until its quality falls below the preset threshold as a result of fading. It is then dropped and the next one that meets the threshold requirement in the fixed sequence is chosen. In scanning diversity, the signal chosen is often not the best one available. A technique that examines the K available signals simultaneously and selects only the best one for delivery is conceptually (although not always practically) preferable. Such a technique is referred to as *optimal selection diversity*.

In the combining techniques, all the available noisy waveforms, good and poor, are utilized simultaneously as indicated in Eq. (1); the a_k values are all nonzero all the time. Of all the possible choices of nonzero a_k values, only two are of principal interest. First, on the assumption that there is no a priori knowledge or design that suggests that some of the $f_k(t)$ values will always be poorer than the others, all the available copies are weighted equally in the summation of Eq. (1) irrespective of the fluctuations in quality that will be experienced. Thus, equal mean values of signal level and equal RMS (root-mean-square) values of noise being assumed, the choice $a_1 = a_2 = \cdots = a_k$ is made, and the technique is known as *equal-weight* or *equal-gain combining*. The second possible choice of nonzero weighting factors that is of wide interest is one in which a_k depends upon the quality of $f_k(t)$ and during any short time interval the a_k values are adjusted automatically to yield the maximum SNR for the sum $f(t)$. This is known as *maximal ratio combining*.

In the alternate switching–combining technique a number of the a_k values up to $K - 1$ can be zero during certain time intervals because some of the available signals are dropped when they become markedly noisier than the others. This approach is based on the fact that the performance of an equal-gain combiner will approximate that of the maximal ratio combiner as long as the SNRs of the various channels are nearly equal. But if any of the SNRs become significantly inferior to the others, the overall SNR can be kept closer to the maximum ratio obtainable if the inferior signals are dropped out of the sum $f(t)$.

Over a single-path fading channel, implementation of selection combining does not require any knowledge about the channel, that is, no channel state information (CSI) is necessary at the receiver, other than that needed for coherent carrier recovery, if that is employed. The receiver simply selects the diversity branch that offers the maximum SNR. For equal-gain combining some CSI estimation can be helpful in improving the combiner performance. For example, in a multipath diversity receiver, the maximum delay spread of a multipath fading channel, which is indicative of the channel memory length, can guide the integration time in equal-gain combining, such that the combiner can collect the dispersed signal energy more effectively and perhaps avoid collecting noise over low signal energy time intervals [4]. Maximal ratio combining (MRC) performs optimally, when CSI estimates on both channel phase and multiplicative path attenuation coefficient are available to the combiner. MRC, by using these estimates and proper weighting of the received signal on each branch, yields the maximum SNR ratio for the sum $f(t)$, compared to the selection and equal-gain combining. In the MRC case, diversity branches that bear a strong signal are accentuated and those that carry weak signals are suppressed such that the total sum $f(t)$ will yield the maximum SNR [2]. This is similar to the philosophy that in a free-market society, by making the rich richer, the society as a whole is better off, perhaps because of

increased productivity. Needless to say, in a society as such, laws and justice must also protect the welfare of the needy in order to bring social stability.

Where there is rich scattering, for all multipath fading cases, a RAKE receiver [5] can yield MRC performance. A simplified equal-gain combiner can also be used in some cases, achieving a somewhat lesser diversity gain [4]. However, in digital transmission, if the channel maximum delay spread exceeds the duration of an information bit, intersymbol interference is introduced which needs to be dealt with.

2.1. Statistical Characterization of Fading

For analytical purposes, combined time and frequency, three-dimensional presentations of recordings of fading signals envelopes are obtained through elaborate measurements. These are usually treated as describing a random variable whose statistical properties are determined from fraction-of-time distributions and are hence intimately related to the duration of the interval of observation. The probability distribution functions of such random variables can be considered to characterize a type of stochastic process for which ergodicity theorem applies. According to this theorem, time and distribution averages of random variables described by fraction-of-time distributions are one and the same thing, and they can be used interchangeably depending on expediency.

It is important to note here that although the rate at which the envelope of a received carrier fluctuates may often appear to be high, it is usually quite slow in comparison with the slowest expected variations in the message waveform. In other words, the envelope of the carrier is usually approximately constant when observed over intervals of time that extend over the duration of the longest message element, or over a few periods of the lowest-frequency component in the message spectrum. On the timescale of the fading envelope, such time intervals are then too short for any significant changes in the envelope to occur but not so short that the details of the message waveform are perceived in averaging over the interval. The probability distribution of a fading envelope is usually determined from samples of short time duration, and the results are presented in histograms. Such histograms are invariably compared with simple mathematical curves such as the Rayleigh density and distribution functions or some other functions whose shapes resemble the appearance of the experimental presentations. The fit of the experimental distributions to the Rayleigh distribution is most often excellent for long-range SHF and UHF tropospheric transmission, quite often so for short-range UHF and for ionospheric scatter and reflection of VHF and HF. Accordingly, the Rayleigh fading model is almost always assumed in theoretical treatments, although it is well known that serious deviations from it arise in some situations. According to this model, if a sinusoid of frequency ω_c is radiated at the transmitting end, it will reach a receiver in the form:

$$R(t) = X(t)\cos[\omega_c t + \phi(t)] \tag{2}$$

where $X(t)$ is a slowly fluctuating envelope (or instantaneous amplitude) whose possible values have a probability density function (PDF)

$$p(X) = \frac{2X}{x^2}\exp\left[-\frac{X^2}{x^2}\right] \qquad \text{for} \qquad X \geq 0$$
$$= 0 \qquad \text{otherwise} \tag{3}$$

where x^2 is the mean-square value of X during the small time interval discussed earlier.

No explicit assumptions are usually made concerning the phase $\phi(t)$ beyond the fact that its fluctuations, like those of $X(t)$, are slow compared to the slowest expected variations in the message waveform. But one possible and sometimes convenient assumption to make is that $\phi(t)$ fluctuates in a random manner and can assume all values between 0 and 2π in accordance with the probability density function:

$$p(\phi) = \frac{1}{2\pi} \qquad \text{for} \quad 0 \leq \phi \leq 2\pi$$
$$= 0 \qquad \text{otherwise} \tag{4}$$

The convenience that results from the assumption of a uniformly distributed phase is due to the fact that $R(t)$ of Eq. (2) can now be viewed as a sample function of a narrowband Gaussian process with zero mean and variance $\frac{x^2}{2}$. One may envision, over a multipath channel, many unresolved scattered rays combine in order to give rise to a Gaussian envelope, $R(t)$.

A Rayleigh PDF, as presented, has a single maximum that tends to occur around small values of the random variable X. Thus, in transmitting a binary modulated signal over a Rayleigh fading channel and receiving the signal in additive white Gaussian noise (AWGN), the average bit error rate (BER) for the detected bits tends to be inversely proportional to the receiver SNR, as shown in Fig. 1. This is shown for several binary modulation techniques. This is quite a slow decrease compared to the same, transmitted over an AWGN channel that only suffers with additive white Gaussian noise. For the latter, the BER drops as a function of SNR, exponentially. Naturally, the Rayleigh fading channel model demands significantly higher transmitted power for delivering the bits as reliably as over an AWGN channel. Now, consider use of diversity combining in transmitting a binary modulated signal over a Rayleigh fading channel when the signal is received in AWGN. The average BER for the detected bits now tends to decrease inversely as a function of SNR raised to the power of diversity order, L, as shown in Fig. 2. This is quite a remarkable improvement. In a way, diversity combining process modifies the Rayleigh PDF to look more like a truncated Gaussian PDF, as the order of diversity increases. Thus, loosely speaking, the performance in BER versus SNR for binary transmission over a Rayleigh fading channel starts to resemble that of transmission over an AWGN channel when diversity combining is employed.

Another case of interest in transmission over fading channels is when the received signal in AWGN has a strong deterministic (nonrandom) component. This will contribute to moving the received signal mean away from the region of small signals to a range of rather large

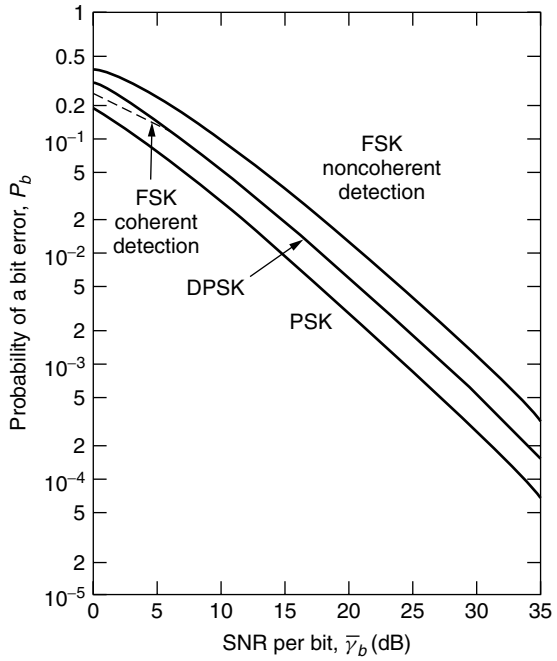


Figure 1. Performance of binary signaling on a Rayleigh fading channel (from Proakis [2]).

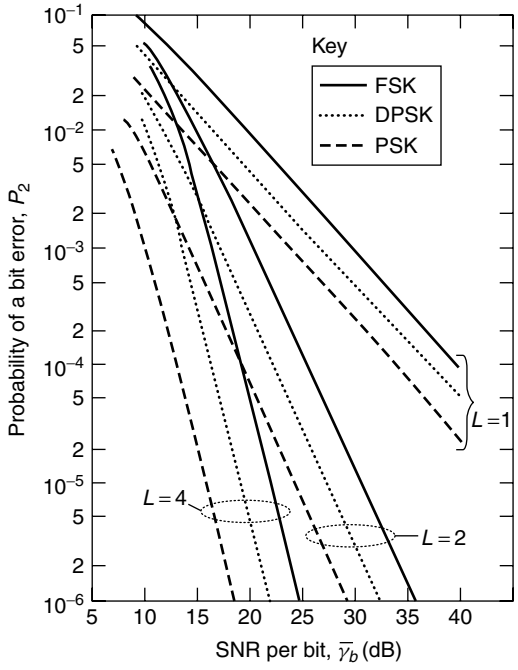


Figure 2. Performance of binary signals with diversity (from Proakis [2]).

signal values. Statistically, the Rayleigh PDF is modified to a Rice PDF [2] that is much milder in the manner by which it affects the transmitted signal. Transmission over a short (∼3 km) line-of-sight microwave channel is often subject to Rice fading. Finally, another PDF that is frequently used to describe the statistical fluctuations of signals received from a multipath fading channel is the Nakagami m distribution [2]. The parameter m is defined as the ratio of moments, called the *fading figure*. By setting $m = 1$, the Nakagami PDF reduces to a Rayleigh PDF. As m increases to values above $m = 1$, transmission performance over a Nakagami channel improves. In a way, this is similar to what happens on a Rice channel. That is, receiver is provided with a stronger signal average, as m increases above the $m = 1$ value.

3. DIVERSITY THROUGH CHANNEL CODING WITH INTERLEAVING

In general, time and/or frequency diversity techniques may be viewed as a form of trivial repetition (block) coding of the information signal. The combining techniques can then be considered as soft-decision decoding of the trivial repetition codes. Intelligent (nontrivial) coding in conjunction with interleaving provides an efficient method of achieving diversity on a fading channel, as emphasized, for example, by chase [6]. The amount of diversity provided by a code is directly related to the code minimum distance, $d_{\min}$. With soft-decision decoding, the order of diversity is increased by a factor of $d_{\min}$, whereas, if hard-decision decoding is employed, the order of diversity is increased by a factor of $\frac{d_{\min}}{2}$. Note that, although coding in conjunction with interleaving can be an effective tool in achieving diversity on a fading channel, it cannot help the signal quality received through a single stationary antenna located in a deep-fade null. The interleaving will not be helpful, since in practice, the interleaving depth cannot be indefinite.

In the next section, we present some examples illustrating the benefits of diversity combining in various communications applications.

4. APPLICATION EXAMPLES

Three examples, illustrating benefits of diversity techniques in practical communications systems, are presented in this section.

4.1. Wideband Code-Division-Multiple-Access Using Direct-Sequence Spread-Spectrum (DSSS) Communications

One application [4] treats a wireless cellular scenario. This represents the up and downstream transmission mode, from user to base station (BS) and from the base to user, of a wireless local-area network (LAN) with a star architecture. Each user has a unique DSSS code and a correlator exists for each user in a channel bank structure at the base station. The output of the bank of correlators is fed to the usual BS circuitry and call setup is handled using standard BS features. Average power control is used in this system to avoid the classical near/far problem. In wireless LAN we have a severe multipath fading problem. It is really a classical Rayleigh multipath, fading channel scenario. The role of asynchronous transmission is clear — a user transmits at random. The signal arrives at the correlator bank and is detected, along with interfering signals. Because the broad bandwidth of a DSSS signal can indeed exceed the channel coherence band (this is the channel band over which all frequency components of the transmitted signal are treated in a correlated

manner), there is inherent diversity in transmission that can be exploited as multipath diversity [3] by a correlation receiver. The psuedonoise (PN) sequences used as direct sequence spreading codes in this application are indeed trivial repeat codes. By exclusive-OR addition of a PN sequence to a data bit, the narrowband of data is spread out to the level of the wide bandwidth of PN sequence. A correlation receiver that knows and has available the matching PN sequence, through a correlation operation, generates correlation function peaks representing the narrowband information bits with a correlation function base, in time, twice the size of a square PN sequence pulse, called a *PN chip*. In this application, the correlator, or matched-filter output, will be a number of resolved replicas of the same transmitted information bit, displayed by several correlation peaks. The number of correlation peaks representing the same information bit corresponds to the diversity order, in this application. Many orders of diversity can be achieved this way.

The replicas obtained this way may now be presented to a combiner for diversity gain. A simple equal gain combiner has been adopted [4] that is by far simpler in implementation than a RAKE receiver [5]. The multipath diversity exploitation in conjunction with multiantenna space diversity [7] establishes a foundation for joint space and time coding.

4.2. Indoor Wireless Infrared (IR) Communications

The purpose of using infrared wireless communication systems in an indoor environment is to eliminate wiring. Utilization of IR radiation to enable wireless communication has been widely studied and remote-control units used at homes introduce the most primitive applications of this type of wireless systems. A major advantage of an IR system over an RF system is the absence of electromagnetic wave interference. Consequently, IR systems are not subject to spectral regulations as RF systems are. Infrared radiation, as a medium for short-range indoor communications, offers unique features compared to radio. Wireless infrared systems offer an inherent spatial diversity, making multipath fading much less of a problem. It is known that the dimension of the coherence area of a fully scattered light field is roughly of the order of its wavelength [8].

This is due to the fact that the receive aperture diameter of a photodiode by far exceeds the wavelength of an infrared waveform. Therefore, the random path phase of a fading channel is averaged over the photo-receiver surface. Hence, the signal strength fluctuations that are caused by phase cancellations in the RF domain are nonexistent in the optical domain. An optical receiver actually receives a large number (hundreds of thousands) of independent signal elements at different locations on the receiving aperture of the photodiode. This in fact provides spatial diversity, which is very similar to employing multiple, geographically separated antennae in an RF fading environment. In summary, because of the inherent diversity channels, the frequency-selective fading effect at the optical carrier frequency level is not a problem in an IR system.

Since infrared transmission systems use an optical medium for data transmission, they have an inherent potential for achieving a very high capacity level. However, in order to make this feasible, the communication system design has to offer solutions to the problems associated with IR propagation in a noisy environment. Various link designs may be employed in indoor wireless infrared communication systems. The classification is based on the degree of directionality and existence of a line-of-sight path between a transmitter and a receiver. In one configuration, instead of transmitting one wide beam, multi-beam transmitters are utilized [9]. These emit multiple narrow beams of equal intensity, illuminating multiple small areas (often called *diffusing spots*) on a reflecting surface. Each beam aims in a prespecified direction. Such a transmitting scheme produces multiple-line-of-sight (as seen by the receiver) diffusing spots, all of equal power, on an extended reflecting surface, such as a ceiling of a room. Each diffusing spot in this arrangement may be considered as a secondary line-of-sight light source having a Lambertian illumination pattern [10]. Compared to other configurations, this has the advantage of creating a regular grid of diffusing spots on the ceiling, thus distributing the optical signal as uniformly as possible within the communication cell as shown in Fig. 3. A direction diversity (also known as angle diversity) receiver that utilizes multiple receiving elements, with each element pointed at a different direction, is used, in order to provide a diversity scheme for optimal rejection of ambient noise power from sunlight or incandescent light, for example, and to substantially reduce the deleterious effects of time dispersion via multiple arrivals of reflecting light rays, causing intersymbol interference in digital transmission. The composite receiver consists of several narrow field-of-view (FoV) elements replacing a single element wide-FoV receiver. The receiver consists of more than one element in order to cover several diffusing spots, thus ensuring uninterrupted communication in case some of the transmitter beams are blocked. Additionally, a multiple-element receiver provides diversity, thus allowing combining of the output signals from different receiver elements, using optimum combining methods. An increased system complexity is the price that one has to pay to escape from restrictions of line-of-sight links, retaining the potential for a high capacity wireless communication system.

For indoor/outdoor wireless transmission systems, use of multiple antennas at both transmit and receive sides to achieve spatial diversity at RF has gained a significant amount of attention. This is motivated by the lack of

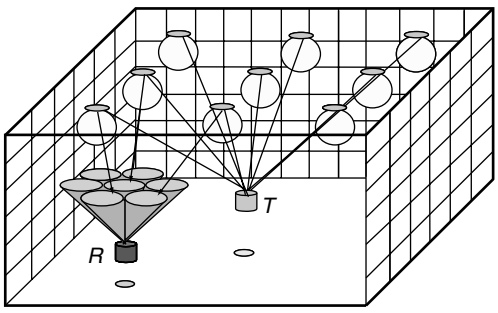


Figure 3. Multispot diffusing configuration (T — transmitter; R — receiver).

available bandwidth at the low-frequency end of the radio spectrum. The approach, in a way, is similar to the IR system described earlier. The capacity increase and spatial multiplexing for high-data-rate transmissions via multiple transmit antennas have been illustrated in Ref. 11. Earlier, transmission of orthogonally-polarized transmitted digitally modulated RF waveforms was introduced to achieve capacity increase over multipath fading channels of point-to-point high-data-rate digital radio routes [12].

4.3. Polarization-Insensitive Fiberoptic Communications

In a coherent optical receiver, receiving a modulated lightwave on a single-mode fiber, the polarization state of the received optical signal must be matched with that of the receiver local laser source signal. A mismatch reduces (perhaps drastically) the received signal strength. Unless the polarization states of the light fields are controlled carefully, receiver performance degradation is unavoidable. The problem encountered here is very similar to the flat fading on a single-path fading channel. Occasionally, the entire band of the received signal is severely attenuated. In an earlier study [13] we examined a polarization-insensitive receiver for binary frequency shift keying (FSK) transmission with discriminator demodulation. The polarization-state-insensitive discriminator receiver is shown in Fig. 4. The two branches carry horizontally and vertically polarized optical beams obtained through a polarization beamsplitter. The information-carrying optical beams are subsequently heterodyne demodulated down to FSK-modulated IF signals and are received by the discriminators for demodulation. The demodulated baseband signals are then combined; thereby a polarization-independent signal is obtained. This is yet another example of diversity combining of equal-gain type.

5. CONCLUDING REMARKS

Given proper operating condition of the equipment, the reliability of a communication system is basically determined by the properties of the signal at the receiving end. We have concentrated in this article on diversity and the effects it may have upon the signal reliability. It is established that "diversification" offers a gain in reliable signal detection. However, a wireless channel offers endless possibilities over a multiplicity of dimensions. Diversity is only one way of introducing a long-term average gain into the detection process. More recently, the availability of low-cost and powerful processors and the development of good channel estimation methods have rejuvenated an interest in adaptive transmission rate techniques with feedback. This new way of thinking is termed *opportunistic communication*, whereby dynamic rate and power allocation may be performed over the dimensions of time, frequency, and space in a wireless system. In a fading (scattering) environment, the channel can be strong sometimes, somewhere, and *opportunistic* schemes can choose to transmit in only those channel states. Obviously, some channel state information is required for an opportunistic communication approach to be successful. Otherwise, it becomes like shooting in the dark. This is in some respects similar to building a financial investment portfolio of stocks, based on some "insider's" information. Clearly, it results in more gain compared to traditional methods of building a diversified portfolio of stocks, based on long-term published trends. Similarly, one would expect a great loss, if the "insider's" information turned out to be wrong. Consequently, opportunistic communications based on wrong channel states will result in a great loss in the wireless network capacity. *Thus, in those wireless applications where reliable channel states may easily be obtained, it is possible to achieve enormous capacities, at a moderate realization complexity*.

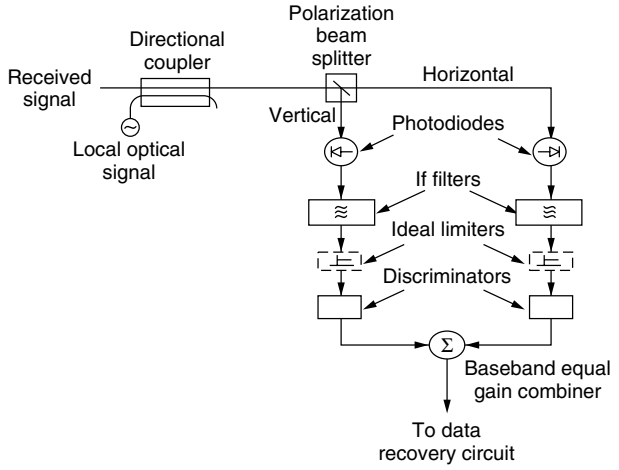


Figure 4. Basic proposed polarization-insensitive receiver.

BIOGRAPHY

Mohsen Kavehrad received his B.Sc. degree in electronics from Tehran Polytechnic, Iran, 1973, his M.Sc. degree from Worcester Polytechnic in Massachusetts, 1975, and his Ph.D. degree from Polytechnic University (Brooklyn Polytechnic), Brooklyn, New York, November 1977 in electrical engineering. Between 1978 and 1989, he worked on telecommunications problems for Fairchild Industries, GTE (Satellite and Labs.), and AT&T Bell Laboratories. In 1989, he joined the University of Ottawa, Canada, EE Department, as a full professor. Since January 1997, he has been with The Pennsylvania State University EE Department as W. L. Weiss Chair Professor and founding director of the Center for Information and Communications Technology Research. He is an IEEE fellow for his contributions to wireless communications and optical networking. He received three Bell Labs awards for his contributions to wireless communications, the 1991 TRIO feedback award for a patent on an optical interconnect, 2001 IEEE VTS Neal Shepherd Best Paper Award, 5 IEEE Lasers and Electro-Optics Society Best Paper Awards between 1991–95 and a Canada NSERC Ph.D.-thesis award in 1995, with his graduate students for contributions to wireless systems and optical networks. He has over 250 published papers, several book chapters, books, and patents in these areas. His current research interests are in wireless communications and optical networks.

BIBLIOGRAPHY

1. W. C. Jakes, Jr., *Microwave Mobile Communications*, Wiley, New York, 1974.
2. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1983.
3. G. L. Turin, Introduction to spread-spectrum anti-multipath techniques and their application to urban digital radio, *Proc. IEEE* **68**: 328–353 (March 1980).
4. M. Kavehrad and G. E. Bodeep, An experiment with direct-sequence spread spectrum and differential phase shift keying modulation for indoor, wireless communications, *IEEE J. Select. Areas Commun.* **SAC-5**(5): 815–823 (June 1987).
5. R. Price and P. E. Green, A communication technique for multipath channels, *Proc. IRE* **46**(3): 555–570 (1958).
6. D. Chase, Digital signal design concepts for a time-varying Ricean channel, *IEEE Trans. Commun.* **COM-24**: 164–172 (Feb. 1976).
7. M. Kavehrad and P. J. McLane, Performance of low-complexity channel-coding and diversity for spread-spectrum in indoor, wireless communication, *AT&T Tech. J.* **64**(8): 1927–1966 (Oct. 1985).
8. R. J. Collier et al., *Optical Holography*, Academic Press, New York, 1971.
9. G. Yun and M. Kavehrad, Spot diffusing and fly-eye receivers for indoor infrared wireless communications, *Proc. IEEE Int. Conf. Selected Topics in Wireless Communications*, Vancouver, Canada, 1992, pp. 262–265.
10. J. R. Barry, *Wireless Infrared Communications*, Kluwer, Boston, 1994.
11. G. J. Foschini and M. J. Gans, On limits of wireless communications in a fading environment using multiple antennas, *Wireless Pers. Commun.* 311–335 (June 1998).
12. M. Kavehrad, Baseband cross-polarization interference cancellation for *M*-quadrature amplitude modulated signals over multipath fading radio channels, *AT&T Tech. J.* **64**(8): 1913–1926 (Oct. 1985).
13. M. Kavehrad and B. Glance, Polarization-insensitive frequency shift keying optical heterodyne receiver using discriminator demodulation, *IEEE J. Lightwave Technol.* **LT-6**(9): 1388–1394 (Sept. 1988).

DMT MODULATION

ROMED SCHUR
STEPHAN PFLETSCHINGER
JOACHIM SPEIDEL
Institute of Telecommunications
University of Stuttgart
Stuttgart, Germany

1. INTRODUCTION

1.1. Outline

Discrete multitone (DMT) modulation as well as orthogonal frequency-division multiplex (OFDM) belong to the category of multicarrier schemes. The early ideas go back to the late 1960s and early 1970s [e.g., 1,2]. With the development of fast digital signal processors, the attraction of these techniques increased [e.g., 3–5] and advanced developments have been carried out [e.g., 6,7]. Meanwhile, DMT was introduced as a standard for digital communications on twisted-pair cables [digital subscriber line (DSL)] [8].

The basic idea of multicarrier modulation is to partition a high-rate datastream into a large number of low-rate data signals that are modulated onto different carrier frequencies and are transmitted simultaneously over parallel subchannels. Because of the partition of the datastream, the data rate on each subchannel is much lower than for the original signal. As low-rate signals are much less susceptible to channel impairments, the reception and reconstruction of the subchannel signals at the receiver side is simplified significantly. However, all subchannels have to be received in parallel and have to be processed simultaneously — a requirement that can be met in an economic way only with digital signal processing. Because of the large number of carriers, the subchannel signals can be well adapted to the channel characteristics. As a consequence, multicarrier schemes like DMT offer the ability to maximize the data throughput over frequency-selective channels, such as the telephone subscriber line.

As multicarrier modulation like DMT or OFDM can be interpreted as a further development of frequency-division multiplexing (FDM) [4], we begin with the explanation of the classical FDM principles.

1.2. Frequency-Division Multiplexing (FDM)

With FDM the available bandwidth of the transmission medium is separated into a number of frequency bands in order to transmit various signals simultaneously on the same medium. The principal block diagram is given in Fig. 1. Each band-limited baseband signal $x_\nu(t)$ is modulated onto a carrier $\cos(\omega_\nu t)$ with carrier frequency ω_ν, $\nu = 0, 1, \ldots, N-1$.

The received FDM signal is first bandpass (BP)-filtered with center frequency ω_μ, multiplied by $\cos(\omega_\mu t)$ and then lowpass (LP)-filtered to obtain the demodulated signal $\hat{x}_\mu(t)$.

The FDM signal at the transmitter output is

$$s(t) = \sum_{\nu=0}^{N-1} x_\nu(t)\cos(\omega_\nu t) \tag{1}$$

For the spectra of the transmitter signals, we have

$$s(t) \leftrightarrow S(\omega) \tag{2}$$

$$x_\nu(t)\cos(\omega_\nu t) \leftrightarrow \tfrac{1}{2}X_\nu(\omega-\omega_\nu) + \tfrac{1}{2}X_\nu(\omega+\omega_\nu) \tag{3}$$

$$s(t) \leftrightarrow S(\omega) = \frac{1}{2}\sum_{\nu=0}^{N-1} X_\nu(\omega-\omega_\nu) + X_\nu(\omega+\omega_\nu) \tag{4}$$

The term $S(\omega)$ is shown in Fig. 2 for the simplified case that all $X_\nu(\omega)$ are real-valued. In conventional FDM systems, the frequencies ω_ν have to be chosen in such a way that the spectra of the modulated signals do not overlap. If all baseband signals $X_\nu(\omega)$ have the same

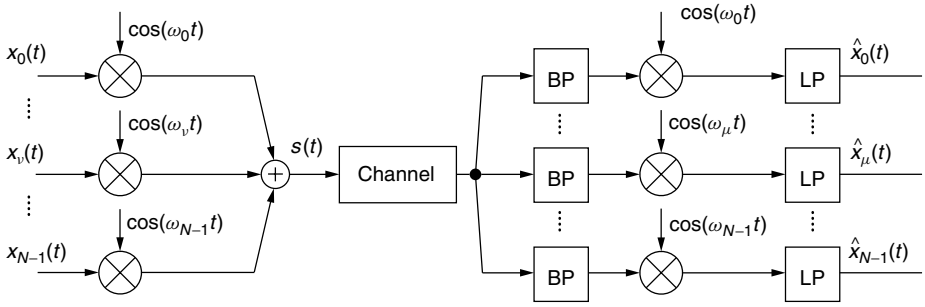


Figure 1. Analog frequency division multiplexing. The baseband signals $x_0(t), \ldots, x_{N-1}(t)$ are modulated on different carriers.

bandwidth, the carrier frequencies are normally chosen equidistantly. However, the FDM system described above wastes valuable bandwidth because (1) some space must be available for the transition between the passband and the stopband of each bandpass filter in Fig. 1; and (2), if $s(t)$ is real, $S(\omega)$ is conjugated symmetric, that is, $S(-\omega) = S^*(\omega)$, where * denotes the conjugate complex operation. As a consequence, the bandwidth of $S(\omega)$ is twice as required theoretically.

The second drawback can be solved by quadrature amplitude modulation (QAM) [9]. With QAM, two independent baseband signals are modulated onto a sine and a cosine carrier with the same frequency ω_ν which gives an unsymmetric spectrum. As a result the spectral efficiency is doubled.

The first drawback of FDM can be overcome by multicarrier modulation, such as by DMT, which allows for a certain overlap between the spectra illustrated in Fig. 2. Of course, spectral overlap in general could lead to nontolerable distortions. So the conditions for this overlap have to be carefully established in order to recover the signal perfectly at the receiver side. This will be done in the next section.

2. MULTICARRIER BASICS

2.1. Block Diagram and Elementary Impulses

Following the ideas outlined in the previous section and applying discrete-time signals $X_\nu[k]$ at the input leads us to the block diagram in Fig. 3.

The impulse modulator translates the sequence $\{\ldots, X_\nu[0], X_\nu[1], \ldots\}$ into a continuous-time function

$$x_\nu(t) = T_S \sum_{k=-\infty}^{\infty} X_\nu[k] \cdot \delta(t - kT_S), \quad \nu = 0, 1, \ldots, N-1 \tag{5}$$

where T_S is the symbol interval, $\delta(t)$ is the Dirac impulse and $\delta[k]$ is the unit impulse with

$$\delta[k] = \begin{cases} 1 & \text{for} \quad k = 0 \\ 0 & \text{for} \quad k \neq 0 \end{cases} \tag{6}$$

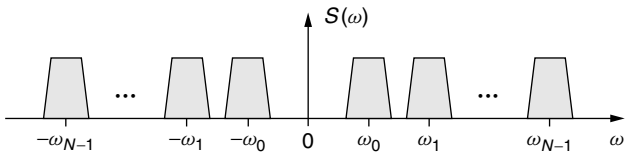


Figure 2. Spectrum $S(\omega)$ of the FDM signal $s(t)$.

The output signal $s(t)$ of the multicarrier transmitter is given by

$$s(t) = T_S \sum_{\nu=0}^{N-1} e^{j\omega_\nu t} \sum_{k=-\infty}^{\infty} X_\nu[k] g(t - kT_S) \tag{7}$$

where $g(t)$ is the impulse response of the impulse shaping filter. The carrier frequencies ω_ν are chosen as integer multiples of the carrier spacing $\Delta\omega$:

$$\omega_\nu = \nu \cdot \Delta\omega, \quad \nu = 0, 1, \ldots, N-1 \tag{8}$$

All practical multicarrier systems are realized with digital signal processing. Nevertheless, we will use the analog model of Fig. 3 in this section, because the analysis can be done more conveniently and the understanding of the principles of multicarrier modulation is easy. Section 3.1 deals with the digital implementation.

The output signal $s(t)$ can be either real or complex. If complex, $s(t)$ can be considered as the complex envelope and the channel is the equivalent baseband channel. As will be shown in Section 3.2 , DMT modulation provides a real-valued output signal $s(t)$. Nevertheless, the following derivations hold for both cases.

The receiver input signal $w(t)$ is demodulated and filtered by the receiver filters with impulse response $h(t)$, resulting in the signal $y_\mu(t)$, $\mu \in \{0, \ldots, N-1\}$. Sampling this signals at the time instants $t = kT_S$ gives the discrete-time output $Y_\mu[k]$. The signal after filtering is given by

$$\begin{aligned} y_\mu(t) &= (w(t)e^{-j\omega_\mu t}) \star h(t) \\ &= \int_{-\infty}^{\infty} w(\tau) e^{-j\omega_\mu \tau} h(t - \tau) d\tau, \quad \mu = 0, \ldots, N-1 \end{aligned} \tag{9}$$

where $\star$ denotes the convolution operation. For the moment, we assume an ideal channel. Thus $w(t) = s(t)$ and we get from (9) with (7) and (8)

$$\begin{aligned} y_\mu(t) = T_S \int_{-\infty}^{\infty} \Bigg(&\sum_{\nu=0}^{N-1} e^{j(\nu-\mu)\Delta\omega\tau} \\ &\times \sum_{k=-\infty}^{\infty} X_\nu[k] g(\tau - kT_S) h(t - \tau) \Bigg) d\tau \end{aligned} \tag{10}$$

Sampling $y_\mu(t)$ at $t = kT_S$ yields the output signal

$$\begin{aligned} Y_\mu[k] = y_\mu(kT_S) = \int_{-\infty}^{\infty} \Bigg(&\sum_{\nu=0}^{N-1} e^{j(\nu-\mu)\Delta\omega\tau} \\ &\times \sum_{l=-\infty}^{\infty} X_\nu[l] g(\tau - lT_S) h(kT_S - \tau) \Bigg) d\tau \end{aligned} \tag{11}$$

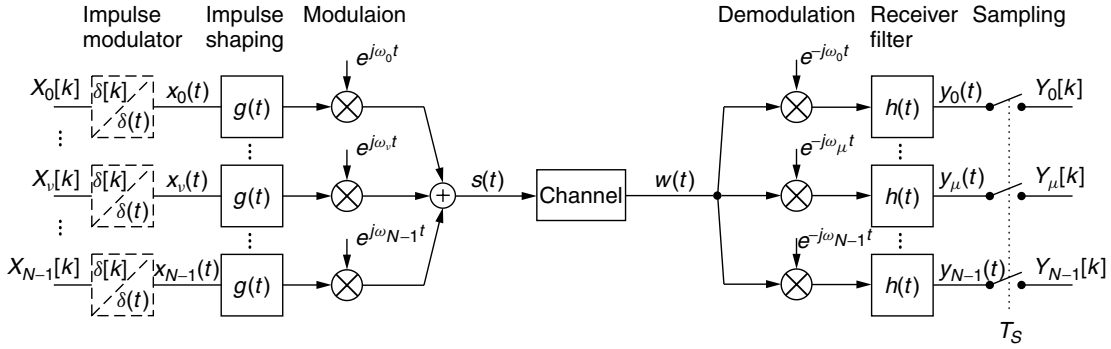


Figure 3. Block diagram of a general multicarrier system.

Now, the target is to recover the sequences $X_\nu[k]$ at the receiver side without distortion. As the given system is linear, we can restrict our analysis to the evaluation of the response to one single transmitted symbol on one subcarrier. We can then make use of the superposition theorem for the general case of arbitrary input sequences. Basically, two types of interference can be seen from Eq. (11):

1. *Intersymbol interference* (ISI), in which a symbol sent at time instant k on subcarrier μ has impact on previous or following samples $y_\mu(lT_S)$ with $l \neq k$.
2. *Intercarrier interference* (ICI), which is the result of crosstalking between different subchannels at time instants kT_S.

Without loss of generality we assume that the system model in Fig. 3 has zero delay, i.e., the filters $g(t)$, $h(t)$ are not causal. We sent one unit impulse $\delta[k]$ at the time $k = 0$ on subcarrier ν:

$$X_i[k] = \delta[\nu - i]\delta[k] \tag{12}$$

The received signal $y_\mu(t)$ is free of any interference at the sampling instants kT_S, if

$$Y_\mu[k] = \delta[\nu - \mu]\delta[k] \tag{13}$$

To gain further insight into the nature of ISI and ICI, we take a closer look at the received signals before they are being sampled. A unit impulse $X_\nu[k] = \delta[k]$ on carrier ν yields the transmitter output signal

$$s(t) = T_S\, g(t) e^{j\nu\Delta\omega t} \tag{14}$$

which is also the input signal $w(t)$ as we assume an ideal channel. We now define the elementary impulse $r_{\nu,\mu}(t)$ as the response $y_\mu(t)$ to the receiver input signal of Eq. (14).

$$\begin{aligned} r_{\nu,\mu}(t) &= T_S \left(g(t)e^{j(\nu-\mu)\Delta\omega t}\right) \star h(t) \\ &= T_S \int_{-\infty}^{\infty} g(\tau) e^{j(\nu-\mu)\Delta\omega\tau} h(t-\tau)\, d\tau \end{aligned} \tag{15}$$

Obviously, $r_{\nu,\mu}(t)$ depends only on the difference $d = \nu - \mu$. Thus we get

$$r_d(t) = T_S(g(t)e^{jd\Delta\omega t}) \star h(t) \tag{16}$$

We can now formulate the condition for zero interference as

$$r_d(kT_S) = \delta[d]\delta[k] \tag{17}$$

This can be interpreted as a more general form of the first Nyquist criterion because it forces not only zero intersymbol interference but also zero intercarrier interference [10,11]. If we set $d = 0$, Eq. (17) simplifies to the Nyquist criterion for single-carrier systems. In the context of multicarrier systems and filterbanks, (17) is also often called an *orthogonality condition* or a *criterion for perfect reconstruction*.

For DMT systems without guard interval $g(t) = h(t)$ holds and they are rectangular with duration T_S. The purpose of the guard interval will be explained in the next section. With the carrier spacing

$$\Delta\omega = \frac{2\pi}{T_S}, \tag{18}$$

we obtain from Eq. (16) the elementary impulses

$$r_0(t) = \begin{cases} 1 - \dfrac{|t|}{T_S} & |t| \leq T_S \\ 0 & \text{elsewhere} \end{cases} \tag{19}$$

$$r_d(t) = \begin{cases} \dfrac{\operatorname{sgn}(t)(-1)^d}{j2\pi d} \left(1 - e^{jd(2\pi/T_S)t}\right) & \text{for } |t| \leq T_S \\ 0 & \text{elsewhere} \end{cases} \quad d \neq 0 \tag{20}$$

These functions are shown in Fig. 4a and give us some insight into the nature of intersymbol and intercarrier interference. We clearly see that the sampling instants $t = kT_S$ do not contribute to any interference. As the elementary impulses are not always zero between the sampling instants, they cause quite some crosstalk between the subchannels. The oscillation of the crosstalk is increased with the distance d between transmitter and receiver subchannel, whereas their amplitude is decreasing proportional to $1/d$.

The impact of interference can also be seen from Fig. 4b, where the eye diagram of the real part of $y_\mu(t)$ is shown for a 16-QAM on $N = 256$ subcarriers; that is, the real and imaginary parts of the symbols $X_\nu[k]$ are taken at random out of the set $\{\pm 1, \pm 3\}$ and are modulated on all subcarriers.

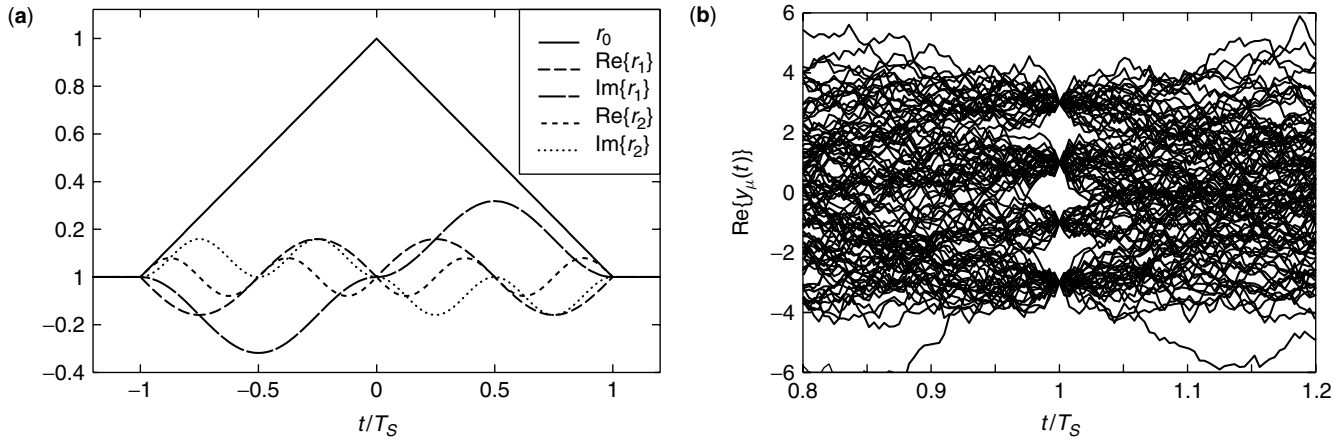


Figure 4. (**a**) Elementary impulses $r_d(t)$ of a multicarrier system with rectangular impulse shapes; (**b**) eye diagram of real part of $y_\mu(t)$ for 16-QAM and $N = 256$ carriers.

Obviously, even with an ideal channel the horizontal eye opening tends to zero, requiring very small sampling jitter and making a correct detection of the transmitted symbols extremely difficult.

2.2. Introduction of a Guard Interval

An effective solution to increase the horizontal eye opening is the introduction of a guard interval. It is introduced by choosing different impulse responses $g(t) \neq h(t)$ at transmitter and receiver. The duration of the guard interval is given as

$$T_G = T_S - T_u \geq 0 \tag{21}$$

where T_S and T_u denote the length of the impulse response of the transmitter and the receiver filter $g(t)$ and $h(t)$, respectively. Both impulse responses are rectangular and symmetric to $t = 0$.

In contrast to (18), the carrier spacing is now

$$\Delta\omega = \frac{2\pi}{T_u} \tag{22}$$

For the elementary impulses follows

$$r_0(t) = \begin{cases} \dfrac{-|t| + T_S - T_G/2}{T_S - T_G} & \text{for} \quad \dfrac{T_G}{2} < |t| < T_S - \dfrac{T_G}{2} \\ 1 & \text{for} \quad |t| < \dfrac{T_G}{2} \\ 0 & \text{for} \quad |t| > T_S - \dfrac{T_G}{2} \end{cases} \tag{23}$$

$$r_d(t) = \left\{ \begin{array}{ll} \dfrac{\text{sgn}(t)(-1)^d}{j2\pi d} \left(e^{j\text{sgn}(t)\, d\pi (T_G/T_u)}\right. & \text{for } \dfrac{T_G}{2} < |t| \\ \left. - e^{jd(2\pi/T_u)t}\right) & < T_S - \dfrac{T_G}{2} \\ 0 & \text{elsewhere} \end{array} \right\} d \neq 0 \tag{24}$$

As can be seen from Fig. 5, there are flat regions around the sampling instants $t = kT_S$, $k = 0, \pm 1, \ldots$ which prevent any interference. In many multicarrier systems, DMT as well as OFDM, the guard interval is chosen as $T_G/T_u = \frac{1}{16}, \ldots, \frac{1}{4}$. In Fig. 5b the eye diagram for a DMT system with guard interval is shown. Obviously, the horizontal eye opening is now increased

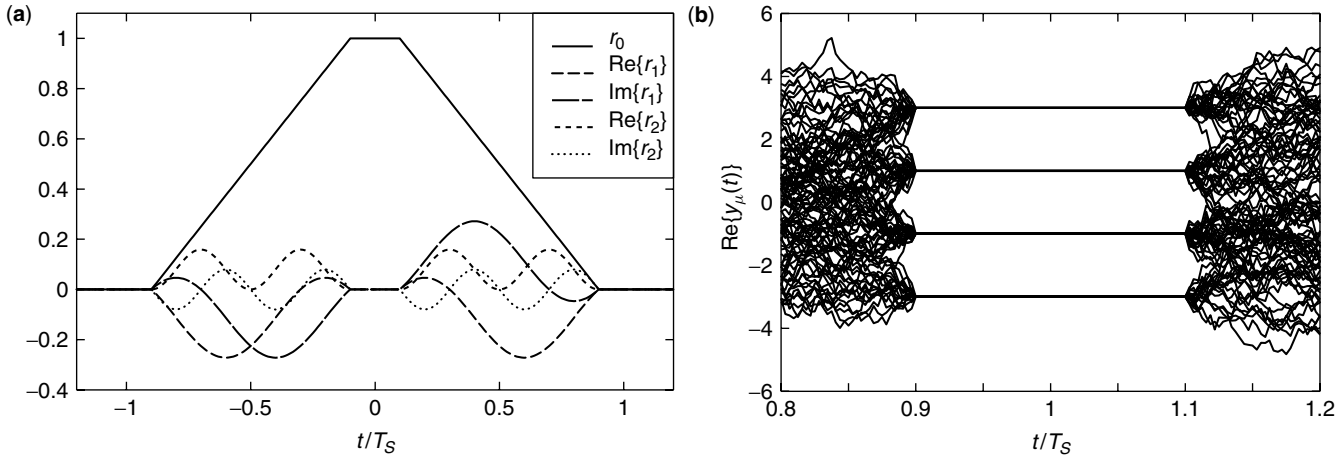


Figure 5. (**a**) Elementary impulses $r_d(t)$ with guard interval $T_G = T_u/4$; (**b**) eye diagram of real part of $y_\mu(t)$ for 16-QAM and $N = 256$ carriers and $T_G = T_u/4$.

and is as long as the guard interval. Therefore, the impact of timing jitter and interference is reduced considerably. However, DMT and OFDM systems are much more sensitive to timing jitter than are single-carrier systems [12].

During the guard interval period T_G, no information can be transmitted. Thus, the spectral efficiency is reduced by a factor T_G/T_S.

3. PRINCIPLES OF DISCRETE MULTITONE MODULATION

3.1. Implementation with Digital Signal Processing

While the continuous-time model in Fig. 3 is valuable for the analysis, the implementation of DMT systems is exclusively done using digital signal processing. Sampling the continuous-time signals in Fig. 3 with the sampling period T_A leads to the discrete-time or digital DMT system in Fig. 6.

The impulse modulators are replaced by upsamplers. They insert $N_S - 1$ zero samples between each incoming symbol $X_\nu[k]$. Thus $T_S = N_S \cdot T_A$ holds. We define

$$T_S = N_S \cdot T_A, \quad T_G = G \cdot T_A, \quad T_u = N \cdot T_A \tag{25}$$

From (21) follows

$$N_S = G + N \tag{26}$$

where G denotes the number of guard samples. For the complex carriers in Fig. 3, we obtain the following with Eqs. (8), (22), and (25) and $t = nT_A$:

$$\exp(j\omega_\nu n T_A) = \exp\left(j\frac{2\pi}{N}\nu n\right) = w^{-\nu n} \quad \text{with}$$
$$w = \exp\left(-j\frac{2\pi}{N}\right) \tag{27}$$

Of course, we have to ask whether the sampling theorem is fulfilled. As $g(t)$ and $h(t)$ are rectangular, and thus have infinite bandwidth, the sampling theorem is not met, at least not exactly. Consequently, the digital system in Fig. 6 is not an exact representation of the continuous-time system of Fig. 3. Nevertheless it is a reasonable approximation, and the concept of the elementary impulses and the generalized Nyquist criterion can be used similarly. We now adopt causal discrete-time filters:

$$g[n] = \begin{cases} \dfrac{1}{\sqrt{N}} & \text{for } n = 0, \ldots, N_S - 1 \\ 0 & \text{elsewhere} \end{cases} \tag{28}$$

The output signal of the transmitter in Fig. 6 can be written as

$$s[n] = \sum_{\nu=0}^{N-1} w^{-\nu n} \sum_{k=-\infty}^{\infty} X_\nu[k] g[n - kN_S] \tag{29}$$

We introduce the operator div for integer divisions as $n \operatorname{div} N_S = \lfloor n/N_S \rfloor$, where $\lfloor\ \rfloor$ indicates rounding toward the nearest smaller integer. With (28) we get from (29)

$$s[n] = \frac{1}{\sqrt{N}} \sum_{\nu=0}^{N-1} X_\nu[n \operatorname{div} N_S] \cdot w^{-\nu n} \tag{30}$$

Here we recognize the expression for the discrete Fourier transform (DFT; IDFT = inverse DFT) pair:

$$\text{DFT: } X_m = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n w^{nm} \quad \text{IDFT: } x_n = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} X_m w^{-nm} \tag{31}$$

Thus, we identify the input signals $X_0[k], \ldots, X_{N-1}[k]$ as a block with index k and define the blockwise IDFT:

$$x_i[k] = \frac{1}{\sqrt{N}} \sum_{\nu=0}^{N-1} X_\nu[k] \cdot w^{-i\nu} \tag{32}$$

which allows us to express (30) as

$$s[n] = x_{n \bmod N}[n \operatorname{div} N_S] \tag{33}$$

For each block k of N input samples, $N_S = N + G$ output samples are produced. The first block $k = 0$ produces the output sequence

$$\{s[0], \ldots, s[N_S - 1]\} = \{x_0[0], \ldots, x_{N-1}[0], x_0[0], \ldots, x_{G-1}[0]\}$$

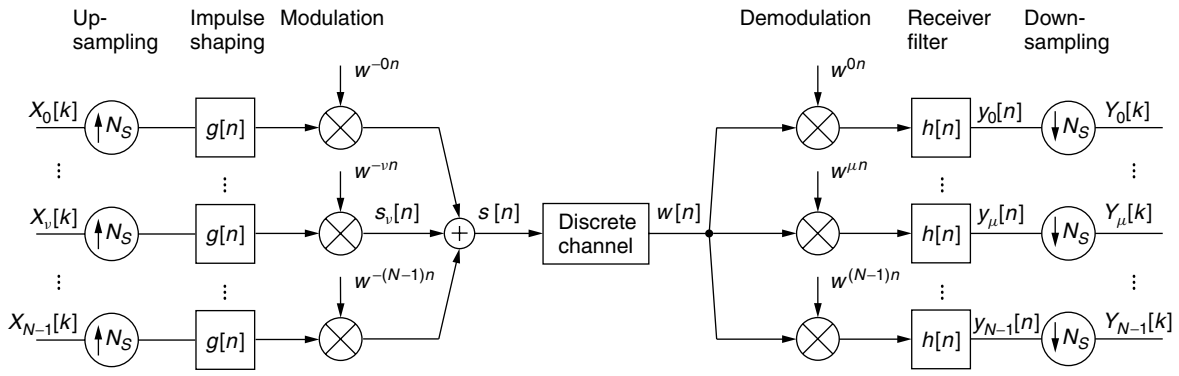


Figure 6. Transmitter and receiver for digital multicarrier transmission.

while the second block ($k = 1$) yields the output

$$\{s[N_S], \ldots, s[2N_S - 1]\}$$
$$= \{x_G[1], \ldots, x_{N-1}[1], x_0[1], \ldots, x_{2G-1}[1]\}$$

We recognize that each output block contains the symbols $x_0[k], \ldots, x_{N-1}[k]$ plus G additionally samples taken out of the same set. The calculation of $s[n]$ applying a blockwise IDFT is illustrated in Fig. 7, where a block of N input symbols is first IDFT-transformed and then parallel–serial-converted. The commutator puts out N_S symbols for each block by turning more than one revolution, resting after each block in a different position. Thus, although each block contains all transformed symbols, the ordering and the subset of doubled samples vary. To overcome this inconvenience and to facilitate the block processing at the receiver side, practically all DMT systems compute a block of N samples by inverse DFT processing and insert the additional samples at the beginning of the block. Therefore the guard interval is also called *cyclic prefix* (CP). Thus, a DMT block for index k will be ordered as follows:

$$\underbrace{x_{N-G}[k], x_{N-G+1}[k], \ldots, x_{N-1}[k]}_{G \text{ samples, cyclic prefix}}, \underbrace{x_0[k], x_1[k], \ldots, x_{N-1}[k]}_{N \text{ data samples}} \tag{34}$$

Thus, we can express the transmitter signal after insertion of the CP as

$$\tilde{x}[n] = x_{n \bmod N_S}[n \operatorname{div} N_S] \tag{35}$$

The discrete-time output signal with block processing is denoted as $\tilde{x}[n]$ in order to distinguish it from $s[n]$. At this point, the mathematical notation seems to be a little bit more tedious than in Eq. (30), but (32) and (35) describe the signals for independent block processing, which simplifies the operations in the transmitter and especially in the receiver considerably. Later we will derive a compact matrix notation that operates blockwise, taking advantage of the block independence.

3.2. Real-Valued Output Signal for Baseband Transmission

Because the DMT signal is transmitted over baseband channels, we must ensure that the output signal $\tilde{x}[n]$ and thus $x_i[k]$ are real-valued. Therefore we have to decompose $x_i[k]$ of (32) into real and imaginary parts

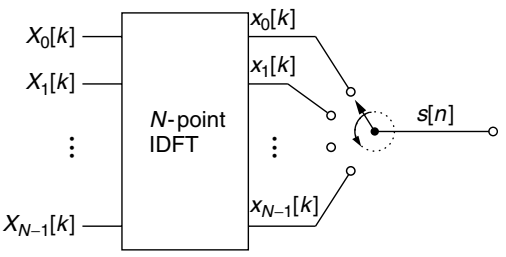


Figure 7. Inverse IDFT with parallel–serial conversion.

and set the latter to zero. After some calculations, this provides the following conditions on the input sequences $X_\nu[k]$:

$$X_0, X_{N/2} \in \mathcal{R} \tag{36}$$

$$X_\nu = X^*_{N-\nu}, \qquad \nu = 1, \ldots, \frac{N}{2} - 1 \tag{37}$$

Consequently, we get from (32) with $X_\nu[k] = X'_\nu[k] + jX''_\nu[k]$

$$x_i[k] = \frac{1}{\sqrt{N}} \left(X_0[k] + (-1)^i X_{N/2}[k] + 2 \times \sum_{\nu=1}^{\frac{N}{2}-1} X'_\nu[k] \cos\left(\frac{2\pi}{N} i\nu\right) - X''_\nu[k] \sin\left(\frac{2\pi}{N} i\nu\right) \right) \tag{38}$$

To simplify the transmitter scheme, we can choose $\tilde{X}_0 = X_0 + jX_{N/2}$ as indicated in Fig. 8. For practical implementations, this is of minor importance as usually X_0 and $X_{N/2}$ are set to zero. The reason will be discussed in Section 3.3.

It is convenient to interpret the DMT signal in (38) as a sum of $N/2$ QAM carriers where $X_\nu[k]$ modulates the νth carrier.

The detailed block diagram of a DMT transmitter as realized in practice is depicted in Fig. 8. The parallel–serial conversion and the insertion of the guard interval is symbolized by the commutator which inserts at the beginning of each block G guard samples, copied form the end of the block. A digital-to-analog converter (DAC) provides the continuous-time output $\tilde{x}(t)$. We clearly see that the guard interval adds some overhead to the signal and reduces the total data throughput by a factor $\eta = G/N_S$. Therefore, the length G of the cyclic prefix is normally chosen much smaller than the size N of the inverse DFT. We will see in Section 3.5 that the cyclic prefix allows for a rather simple equalizer.

Figure 9 depicts the corresponding DMT receiver. After analog-to-digital conversion and synchronization, the signal $\tilde{y}[n]$ is serial–parallel converted. Out of the N_S symbols of one block, the G guard samples are discarded and the remaining N symbols are fed into a DFT. Note that synchronization and timing estimation are essential receiver functions to produce reliable estimates of the transmitted data at the receiver. There have been many proposals for synchronization with DMT modulation, such as that by Pollet and Peeters [13].

Another major item to be considered with DMT modulation is the large peak-to-average power ratio (PAPR) of the output signal $\tilde{x}[n]$. Thus, a wide input range of the DAC is required. Furthermore, large PAPR can cause severe nonlinear effects in a subsequent power amplifier. The PAPR should be considered particularly for multicarrier modulation with a large numbers of subcarriers, because this ratio increases with the number of subcarriers. In order to reduce the

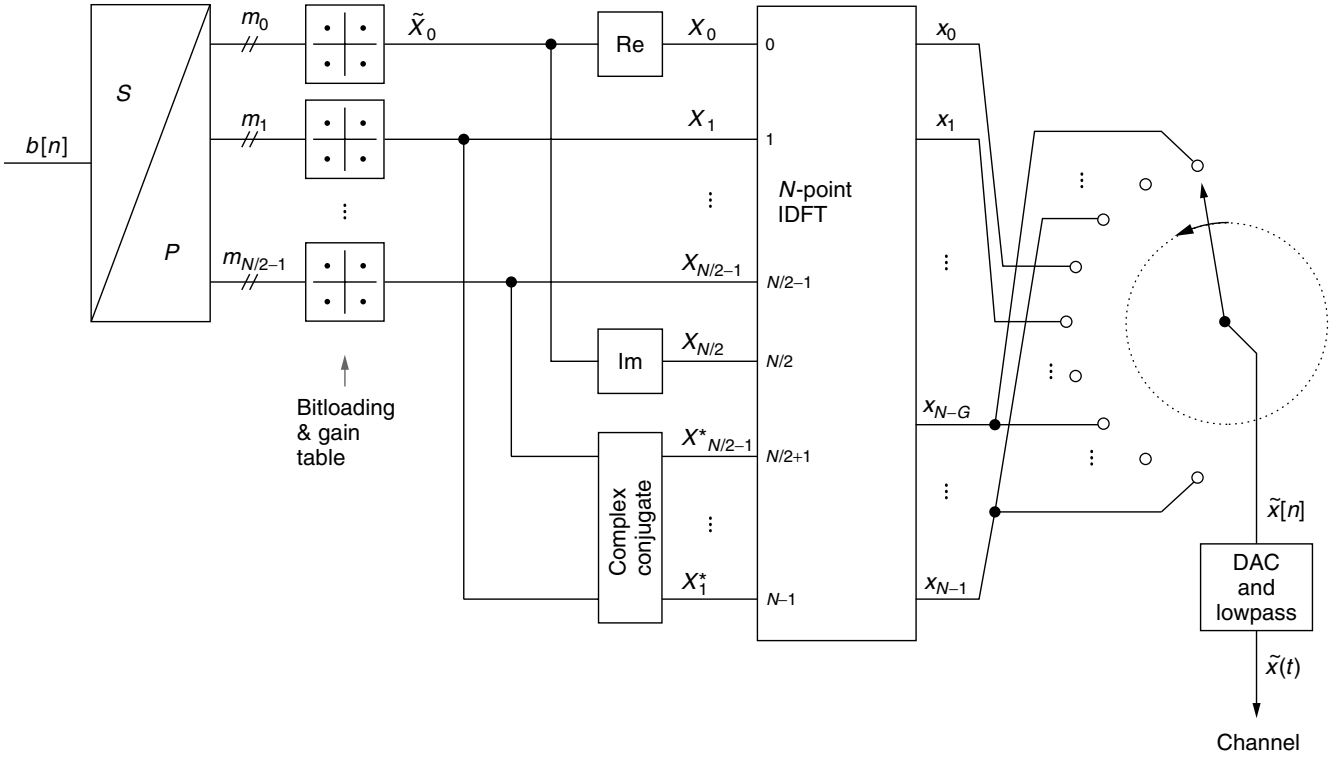


Figure 8. DMT transmitter.

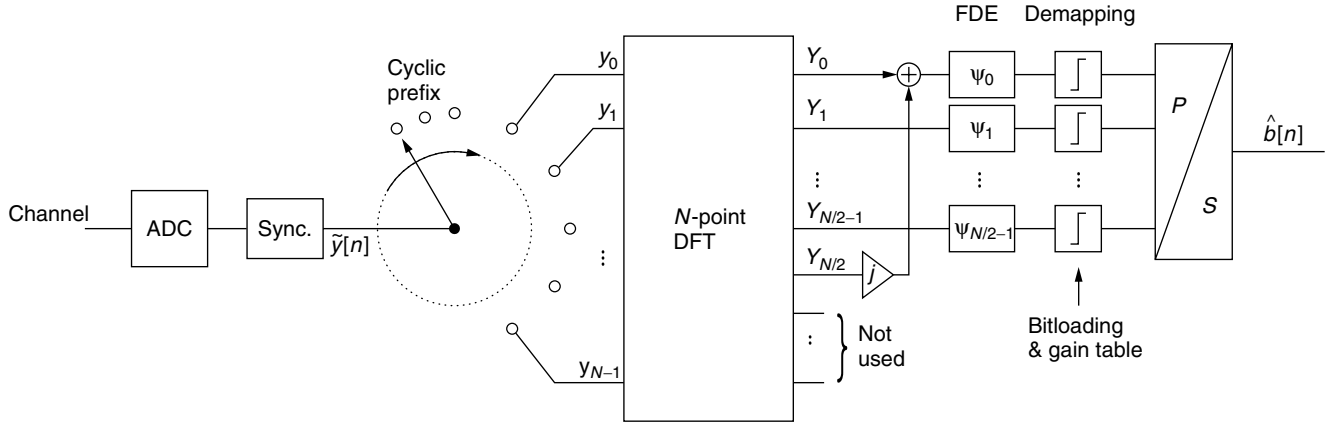


Figure 9. DMT receiver.

PAPR, several techniques have been proposed, such as "selective mapping" and "partial transmit sequence" approaches [14].

3.3. Spectral Properties

We calculate the spectrum of the output signal $s[n]$ in Fig. 6 for one modulated subchannel, namely, the input signal is given by (12). This gives

$$s[n] = g[n] \cdot w^{-\nu n} \leftrightarrow S(\omega) = \sum_{n=-\infty}^{\infty} g[n] \cdot w^{-\nu n} \cdot e^{-j\omega n T_A} \quad (39)$$

From Eqs. (22) and (25) we obtain $T_A = 2\pi/(N\Delta\omega)$. Together with (28), it follows from (39) that

$$S(\omega) = \frac{1}{\sqrt{N}} \begin{cases} \dfrac{1 - w^{(\omega/\Delta\omega - \nu)N_S}}{1 - w^{(\omega/\Delta\omega - \nu)}} & \text{for } \dfrac{\omega}{\Delta\omega} \neq \nu \\ N_S & \text{for } \dfrac{\omega}{\Delta\omega} = \nu \end{cases} \quad (40)$$

In order to obtain a real-valued output signal, according to (37), we have to sent a unit impulse on subchannel $N - \nu$, too. In Fig. 10a the magnitude of the spectrum is depicted for the case that subcarrier ν is modulated. The spectrum is quite similar to a $\sin(\omega)/\omega$-function. However,

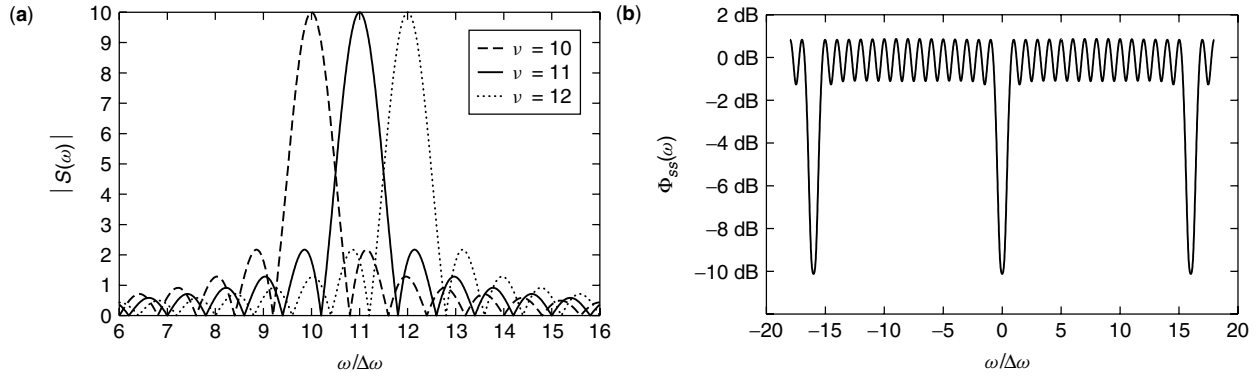


Figure 10. (**a**) Magnitude response $|S(\omega)|$ for modulated carrier ν, $G = N/8$; (**b**) power spectral density $\Phi_{ss}(\omega)$ for $N = 32$, $G = 4$. The ripple increases with the duration of the guard interval.

as Fig. 6 is a discrete-time system, $S(\omega)$ is periodic with $\omega_A = N \cdot \Delta\omega$. Note that the spectrum of a single subcarrier with carrier frequency $\nu\Delta\omega$ does *not* have zero crossings at $\omega = (\nu \pm m)\Delta\omega$, $m = 1, 2, \ldots$ if a guard interval with length $G > 0$ is used.

In order to calculate the power spectral density (PSD) of a DMT signal, we now consider stochastic input sequences $X_\nu[k]$ that are uncorrelated, have zero mean and variances σ_ν^2. Then the PSD of the output signal $s_\nu[n]$ of modulator ν in Fig. 6 becomes

$$\Phi_{s_\nu s_\nu}(\omega) = \frac{\sigma_\nu^2}{N_S} \cdot |G(\omega)|^2, \qquad \text{with}$$

$$G(\omega) = \sum_{n=-\infty}^{\infty} g[n] e^{-j\omega n T_A} = \frac{1}{\sqrt{N}} \frac{1 - w^{\frac{\omega}{\Delta\omega} N_S}}{1 - w^{\frac{\omega}{\Delta\omega}}} \tag{41}$$

From this it follows the total PSD of the transmitter

$$\Phi_{ss}(\omega) = \frac{1}{N_S} \sum_{\substack{\nu=-15 \\ \nu \neq 0}}^{15} \sigma_\nu^2 \cdot |G(\omega - \nu\Delta\omega)|^2 \tag{42}$$

which is depicted in Fig. 10b for a system with $N = 32$. All carriers are modulated with equal power $\sigma_\nu^2 = 1$, except for carriers $\nu = 0$ and $\nu = N/2$. Because of (37), the subcarriers $\nu = 17, \ldots, 31$[1] are modulated with the complex conjugate of the sequences $X_1[k], \ldots, X_{15}[k]$. Note that we can approximately add the PSD of all sequences $s_\nu[n]$ despite their pairwise correlation because their spectra overlap to only a very small extent. As the antialiasing lowpass in the DAC has a cutoff frequency of $\omega_A/2$ [15], the carriers in this frequency region cannot be modulated; therefore, at least the carrier at $\nu = N/2$ remains unused. The carrier at $\nu = 0$ is seldom used, either, because most channels do not support a DC component.

[1] Because of the periodicity of $\Phi_{ss}(\omega)$ this is equivalent to modulate the subcarriers $\nu = -15, \ldots, -1$.

3.4. System Description with Matrix Notation

We have seen that the transmitter processes the data blockwise, resulting in simple implementation without the need of saving data from previous blocks. We now consider how this block independence can be extended to the receiver when including the channel. Figure 11 shows a block diagram for the complete DMT transmission system. For the moment we focus on the signal only and consider the influence of the noise later. The receiver input signal is then given by

$$\tilde{y}[n] = \tilde{x}[n] \star c[n] = \sum_{m=-\infty}^{\infty} \tilde{x}[m]\, c[n-m] \tag{43}$$

The discrete-time channel impulse response $c[n]$ includes the influences of DAC and ADC (digital-to-analog and analog-to-digital conversion), respectively. We assume that the channel can be described by a finite-length

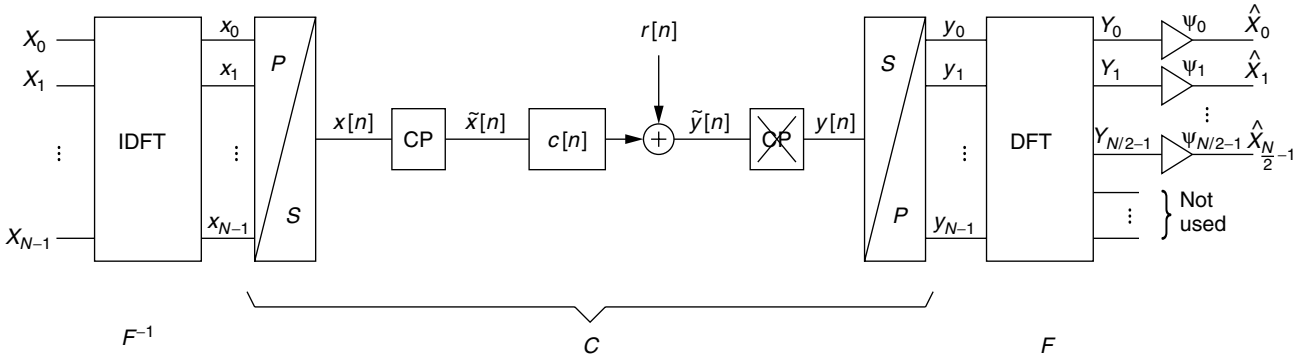


Figure 11. Discrete-time model of a DMT transmission system.

impulse response $c[n]$ with $L+1$ samples $c[0], \ldots, c[L]$. From (43) we conclude that for independence of the received blocks, it must hold that

$$L \leq G \tag{44}$$

Thus, the cyclic prefix must be at least as long as the length of the channel impulse response. If this condition is satisfied, we can adopt a vector notation for the block data:

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}, \quad \tilde{\mathbf{x}} = \begin{pmatrix} x_{N-G} \\ \vdots \\ x_{N-1} \\ x_0 \\ \vdots \\ x_{N-1} \end{pmatrix}, \quad \tilde{\mathbf{y}} = \begin{pmatrix} \tilde{y}_0 \\ \vdots \\ \tilde{y}_{G-1} \\ \tilde{y}_G \\ \vdots \\ \tilde{y}_{N+G-1} \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} \tilde{y}_G \\ \tilde{y}_{G+1} \\ \vdots \\ \tilde{y}_{N+G-1} \end{pmatrix} \tag{45}$$

where $\mathbf{y}$ is the input signal after removal of the cyclic prefix. It can be expressed as

$$\mathbf{y} = \begin{pmatrix} c[G] & c[G-1] & \cdots & c[0] & 0 & \cdots & 0 \\ 0 & c[G] & c[G-1] & \cdots & c[0] & \cdots & 0 \\ \cdots & & & & & \cdots & \\ 0 & 0 & \cdots & c[G] & c[G-1] & \cdots & c[0] \end{pmatrix} \cdot \tilde{\mathbf{x}} \tag{46}$$

or

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{G-1} \\ y_G \\ y_{G+1} \\ \vdots \\ y_{N-1} \end{pmatrix} = \underbrace{\begin{pmatrix} c[0] & 0 & \cdots & 0 & c[G] & \cdots & c[2] & c[1] \\ c[1] & c[0] & 0 & \cdots & 0 & c[G] & \cdots & c[2] \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c[G-1] & \cdots & \cdots & c[0] & 0 & \cdots & 0 & c[G] \\ c[G] & c[G-1] & \cdots & \cdots & c[0] & 0 & \cdots & 0 \\ 0 & c[G] & c[G-1] & \cdots & \cdots & c[0] & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & c[G] & \cdots & \cdots & c[1] & c[0] \end{pmatrix}}_{\mathbf{C}} \times \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{G-1} \\ x_G \\ x_{G+1} \\ \vdots \\ x_{N-1} \end{pmatrix} \tag{47}$$

The latter expression is preferred as it directly relates the signals $\mathbf{x}$ and $\mathbf{y}$ and transforms the effect of the guard interval insertion and removal into the channel matrix $\mathbf{C}$. The matrix equation (47) represents the circular convolution, which is equivalent to multiplication in the frequency domain [15]:

$$y[n] = x[n] \otimes c[n] = \sum_{m=0}^{N-1} x[m] \cdot c[(n-m) \bmod N] \tag{48}$$

From this point of view, the cyclic prefix transforms the linear convolution (43) into a circular convolution. The matrix $\mathbf{C}$ is a circulant matrix whose eigenvalues λ_μ and eigenvectors $\boldsymbol{\vartheta}_\mu$ are given by

$$\lambda_\mu = \sum_{n=0}^{N-1} c[n] w^{\mu n}, \quad \boldsymbol{\vartheta}_\mu = \frac{1}{\sqrt{N}} \begin{pmatrix} w^{-0} \\ w^{-\mu} \\ w^{-2\mu} \\ \vdots \\ w^{-(N-1)\mu} \end{pmatrix},$$

$$\mu = 0, \ldots, N-1 \tag{49}$$

This can be easily verified by checking the equation $\mathbf{C}\boldsymbol{\vartheta}_\mu = \lambda_\mu \boldsymbol{\vartheta}_\mu$. We identify the inverse DFT matrix as

$$\mathbf{F^{-1}} = (\boldsymbol{\vartheta_0}, \boldsymbol{\vartheta_1}, \ldots, \boldsymbol{\vartheta_{N-1}}) \tag{50}$$

This means that the eigenvectors of the channel matrix $\mathbf{C}$ are the columns of the inverse DFT matrix. As a consequence, we can diagonalize the channel matrix by multiplying with the IDFT and its inverse, the DFT matrix $\mathbf{F}$

$$\mathbf{FCF^{-1}} = \boldsymbol{\Lambda} \qquad \text{where } \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_\mu) \tag{51}$$

With $\mathbf{X} = (X_0[k], \ldots, X_{N-1}[k])^T$ and $\mathbf{Y} = (Y_0[k], \ldots, Y_{N-1}[k])^T$ we can write

$$\mathbf{x} = \mathbf{F^{-1}X}, \qquad \mathbf{y} = \mathbf{Cx} = \mathbf{CF^{-1}X}, \qquad \mathbf{Y} = \mathbf{Fy} = \mathbf{FCF^{-1}X} \tag{52}$$

We can now describe the input-output-relation of the whole transmission system of Fig. 11 with a single diagonal matrix: $\mathbf{Y} = \boldsymbol{\Lambda}\mathbf{X}$, or $Y_\mu[k] = \lambda_\mu \cdot X_\mu[k]$. The result shows that due to the cyclic prefix the parallel subchannels are independent, and perfect reconstruction can be realized by a simple one-tap equalizer with tap weight ψ_μ per subchannel located at the output of the receiver DFT as shown in Fig. 11. As equalization is done after the DFT, this equalizer is usually referred to as a *frequency-domain equalizer* (FDE).

Following from Eq. (49), we can interpret the eigenvalues λ_μ of the channel matrix as the DFT of the channel impulse response. If we define

$$C(\omega) = \sum_{n=0}^{N-1} c[n] \cdot e^{-j\omega n T_A} \tag{53}$$

as the discrete-time Fourier transform of $c[n]$, we see that the eigenvalues are just the values of the channel transfer function at the frequencies $\omega_\mu = \mu \cdot \Delta\omega$:

$$\lambda_\mu = C(\omega_\mu) \tag{54}$$

Because DMT is a baseband transmission scheme, the channel impulse response $c[n]$ is real-valued and therefore its spectrum shows hermitian symmetry:

$$\lambda_0, \lambda_{N/2} \in \mathcal{R}; \qquad \lambda_\mu = \lambda^*_{N-\mu} \qquad \text{for } \mu = 1, \ldots, N/2-1 \tag{55}$$

Let us summarize the results. We have shown that the insertion of the cyclic prefix translates the linear convolution (43) into a circular convolution (48) that corresponds to a multiplication in frequency domain, as long as the impulse response of the channel is not longer than the guard interval. If the guard interval is sufficiently large, no interblock and no intercarrier interference occur and each subchannel signal is weighted only by the channel transfer function at the subchannel frequency. If the channel impulse response $c[n]$ exceeds the guard interval, the described features of DMT are not valid anymore. To overcome this problem, $c[n]$ can be compressed by a so-called time-domain equalizer (TDE) to the length of the guard interval. The TDE will be applied before the guard interval is removed at the receiver [e.g., 16–18].

3.5. Frequency-Domain Equalization (FDE)

The output signal of the DFT at the receiver in Fig. 11 is given by

$$Y_\mu[k] = \lambda_\mu \cdot X_\mu[k] + q_\mu[k] \tag{56}$$

where $q_\mu[k]$ represents the noise introduced on the subchannel after DFT processing. For many practical channels, the noise $r[n]$ on the channel will be Gaussian but not white. Therefore, the power of $q_\mu[k]$ will depend on the subchannel. If N is chosen sufficiently large, the subchannel bandwidth is very small and thus the noise in each subchannel will be approximately white with variance $\sigma^2_{q,\mu}$. Further, the noise is uncorrelated with the noise on any other subchannel [5]. The equalizer coefficients ψ_μ can be derived with or without considering the noise term in Eq. (56). If the signal-to-noise ratio (SNR) is high, we approximately neglect the noise and find the optimal FDE parameters as $\psi_\mu = 1/\lambda_\mu$. With $\hat{X}_\mu[k] = \psi_\mu Y_\mu[k]$ follows $\hat{X}_\mu[k] = X_\mu[k] + q_\mu[k]/\lambda_\mu$ for the reconstructed signal at the receiver. The SNR at the output of the μ-th subchannel is then given by

$$\mathrm{SNR}_\mu = \frac{E\{|X_\mu|^2\}}{E\{|\hat{X}_\mu - X_\mu|^2\}} = \frac{\sigma^2_\mu \cdot |\lambda_\mu|^2}{\sigma^2_{q,\mu}} \tag{57}$$

where $\sigma^2_\mu = E\{|X_\mu[k]|^2\}$ is the signal power of the μth input signal at the transmitter.

However, the SNR per subchannel can be improved by considering the noise in the derivation of the equalizer coefficients ψ_μ. Considering the AWGN (additive white Gaussian noise) in the subchannels, we calculate the equalizer coefficients by minimizing the mean-square error (MSE) $E\{|\hat{X}_\mu - X_\mu|^2\}$ which can be written with $\psi_\mu = \psi'_\mu + j\psi''_\mu$ and $\lambda_\mu = \lambda'_\mu + j\lambda''_\mu$ as

$$\begin{aligned} E\{|\hat{X}_\mu - X_\mu|^2\} = {} & (\psi'^2_\mu + \psi''^2_\mu)(|\lambda_\mu|^2\sigma^2_\mu + \sigma^2_{q,\mu}) \\ & - 2\sigma^2_\mu(\psi'_\mu\lambda'_\mu - \psi''_\mu\lambda''_\mu) + \sigma^2_\mu \end{aligned} \tag{58}$$

For minimization of the MSE, we set the gradient of (58) to zero:

$$\frac{\partial}{\partial \psi'_\mu} E\{|\hat{X}_\mu - X_\mu|^2\} = 0 \quad \frac{\partial}{\partial \psi''_\mu} E\{|\hat{X}_\mu - X_\mu|^2\} = 0 \tag{59}$$

which results in

$$\begin{aligned} \psi'_\mu = \frac{\lambda'_\mu}{|\lambda_\mu|^2 + \sigma^2_{q,\mu}/\sigma^2_\mu}, \; \psi''_\mu = \frac{-\lambda''_\mu}{|\lambda_\mu|^2 + \sigma^2_{q,\mu}/\sigma^2_\mu} \\ \Rightarrow \psi_\mu = \frac{\lambda^*_\mu}{|\lambda_\mu|^2 + \sigma^2_{q,\mu}/\sigma^2_\mu} \end{aligned} \tag{60}$$

The subchannel SNR can be calculated as

$$\mathrm{SNR}_\mu = \frac{E\{|X_\mu|^2\}}{E\{|\hat{X}_\mu - X_\mu|^2\}} = \frac{\sigma^2_\mu \cdot |\lambda_\mu|^2}{\sigma^2_{q,\mu}} + 1 \tag{61}$$

which gives a better performance than the first case, especially for low SNR.

It is this amazingly simple equalization with a one-tap equalizer per subchannel that makes DMT so popular for the transmission over frequency-selective channels like the telephone subscriber line.

4. CHANNEL CAPACITY AND BIT LOADING

The channel capacity [19], or the maximum error-free bitrate, of an ideal AWGN channel is given by

$$R_{\max} = \frac{\omega_B}{2\pi} \cdot \log_2\left(1 + \frac{\sigma^2_y}{\sigma^2_r}\right) \tag{62}$$

where $\omega_B = N \cdot \Delta\omega$ is the total channel bandwidth, σ^2_y is the received signal power, and σ^2_r is the noise power. The capacity of subchannel μ with very small bandwidth $\Delta\omega$ is then

$$R_\mu = \frac{\Delta\omega}{2\pi} \cdot \log_2\left(1 + \frac{\Delta\omega\Phi_{xx}(\omega_\mu)|C(\omega_\mu)|^2}{\Delta\omega\Phi_{rr}(\omega_\mu)}\right) \tag{63}$$

where $\Phi_{xx}(\omega)$ denotes the power spectral density (PSD) of the transmitter signal $\tilde{x}[n]$, $C(\omega)$ is the channel transfer function in (53) and $\Phi_{rr}(\omega)$ is the PSD of the (colored) noise $r[n]$. The total capacity of all subchannels is

$$R = \frac{\Delta\omega}{2\pi} \cdot \sum_{\mu=0}^{N-1} \log_2\left(1 + \frac{\Phi_{xx}(\omega_\mu)|C(\omega_\mu)|^2}{\Phi_{rr}(\omega_\mu)}\right) \tag{64}$$

In the limit for $\Delta\omega \to 0$ with $\Delta\omega \cdot N = \omega_B = \text{const.}$ we get

$$R = \frac{1}{2\pi}\int_0^{\omega_B} \log_2\left(1 + \frac{\Phi_{xx}(\omega)|C(\omega)|^2}{\Phi_{rr}(\omega)}\right) d\omega \tag{65}$$

We now want to maximize the channel capacity R, subject to the constraint that the transmitter power is limited:

$$\frac{1}{2\pi}\int_0^{\omega_B} \Phi_{xx}(\omega)d\omega = P_t \tag{66}$$

Thus, we search for a function $\Phi_{xx}(\omega)$ that maximizes Eq. (65) subject to the constraint (66). This can be

accomplished by calculus of variations [20]. Therefore we set up the Lagrange function

$$L(\Phi_{xx}, \omega) = \log_2\left(1 + \Phi_{xx}(\omega)\frac{|C(\omega)|^2}{\Phi_{rr}(\omega)}\right) + \lambda \cdot \Phi_{xx}(\omega) \quad (67)$$

which must fulfill the Euler–Lagrange equation

$$\frac{\partial L}{\partial \Phi_{xx}} = \frac{d}{d\omega}\frac{\partial L}{\partial \Phi'_{xx}}, \quad \text{where } \Phi'_{xx} = \frac{d\Phi_{xx}}{d\omega} \quad (68)$$

This requires that $\Phi_{xx}(\omega) + \Phi_{rr}(\omega)/|C(\omega)|^2 = \Phi_0 = \text{const}$. Thus,

$$\Phi_{xx}(\omega) = \begin{cases} \Phi_0 - \Phi_{rr}(\omega)/|C(\omega)|^2 & \text{for } |\omega| < \omega_B \\ 0 & \text{elsewhere} \end{cases} \quad (69)$$

This represents the "water-filling solution." We can interpret $\Phi_{rr}(\omega)/|C(\omega)|^2$ as the bottom of a bowl in which we fill an amount of water corresponding to P_t. The water will distribute in a way that the depth represents the wanted function $\Phi_{xx}(\omega)$, as illustrated in Fig. 12.

The optimum solution according to (69) can be achieved by appropriately adjusting the constellation sizes and the gain factors for the bit-to-symbol mapping for each carrier.

In practice, when the bit rate assignments are constrained to be integer, the "water-filling solution" has to be modified. There have been extensive studies on the allocation of power and data rate to the subchannels, known as *bitloading algorithms* [21–24]. In the following, the connection between the channel characteristics and bitloading is presented for an asymmetric digital subscriber line (ADSL) system.

5. APPLICATIONS OF DMT

The DMT modulation scheme has been selected as standard for ADSL [8], a technique for providing high-speed data services over the existing copper-based telephone network infrastructure. Typically, these phone lines consist of unshielded twisted-pair (UTP) wire, which share the same cable with multiple other pairs. Hence, the performance of ADSL transmission depends on the noise environment and cable loss [25]. When lines are bound together in a cable, they produce crosstalk from one pair to another, at levels that increase with frequency and the number of crosstalking pairs or disturbers. For this reason, crosstalk is one of the major noise sources. Additional disturbances are given by impulse noise and possibly radiofrequency interferers. Beside that, line attenuation increases with frequency and distance. A modulation scheme that is well suited for the wide variations in the ADSL channel characteristics is provided by DMT. In particular, rate adaption is quite simple for DMT modulation to optimize the ADSL system transmission. Some aspects of an ADSL system are regarded in the remainder of this section.

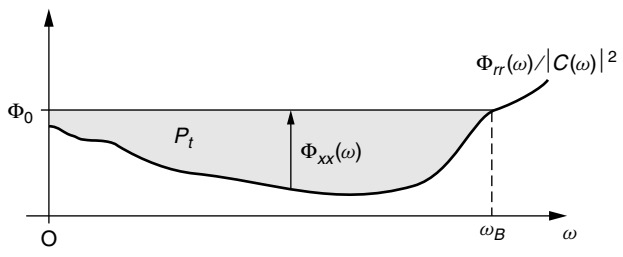


Figure 12. Optimum water-filling solution.

The parameter of the DMT modulation scheme for downstream transmission are determined by the ADSL standard [8] as follows:

- FFT size of $N = 512$. Consequently, excluding the two carriers at $\nu = 0$ and $\nu = N/2$, 255 usable parallel subchannels result.
- Sampling rate $f_A = 1/T_A = 2.208$ MHz.
- Guard interval length $G = 32 = N/16$.
- Adaptive bitloading with a maximum number of $m_{\max} = 15$ bits per subcarrier (tone).
- A flat transmit power spectral density of about -40 dBm/Hz (dBm = decibels per milliwatt). As the spectrum ranges up to $f_A/2 = 1.104$ MHz, the total transmit power is about 20 dBm.

The achievable data throughput depends on the bit allocation that is established during initialization of the modem and is given by

$$R = \frac{1}{T_S}\sum_{\nu=1}^{N/2-1} m_\nu \quad (70)$$

where m_ν denotes the number of bits modulated on the νth subcarrier. The frequency spacing $\Delta f = \Delta\omega/2\pi$ between subcarriers is given by

$$\Delta f = \frac{f_A}{N} = 4.3125 \text{ kHz} \quad (71)$$

As ADSL services share the same line with the plain old telephone service (POTS), a set of splitter filters (POTS splitter) at the customer premises and the network node separate the different frequency bands. Usually, POTS signals occupy bandwidths of up to 4 kHz. To allow a smooth and inexpensive analog filter for signal separation, ADSL services can use a frequency range from 25 kHz up to 1.1 MHz.

Figure 13a shows the variation of the SNR in the DMT subchannels at the receiver for a subscriber line of 3 km length and 0.4 mm wire diameter. DMT modems for ADSL with rate adaption measure the SNR per subchannel during an initialization period.

Of course, the POTS splitter in an ADSL system essentially removes all signals below 25 kHz. Thus, the first few channels cannot carry any data, as can be seen in Fig. 13b, which shows the bitloading corresponding to the SNR of Fig. 13a. A minimum bitload of 2 bits per carrier is stipulated in the standard. The total bit rate is about 4 Mbps. Further, attenuation can be severe at the upper end of the ADSL frequency band. As a consequence, SNR is low and the adaptive bitloading algorithm allocates fewer bits per carrier.

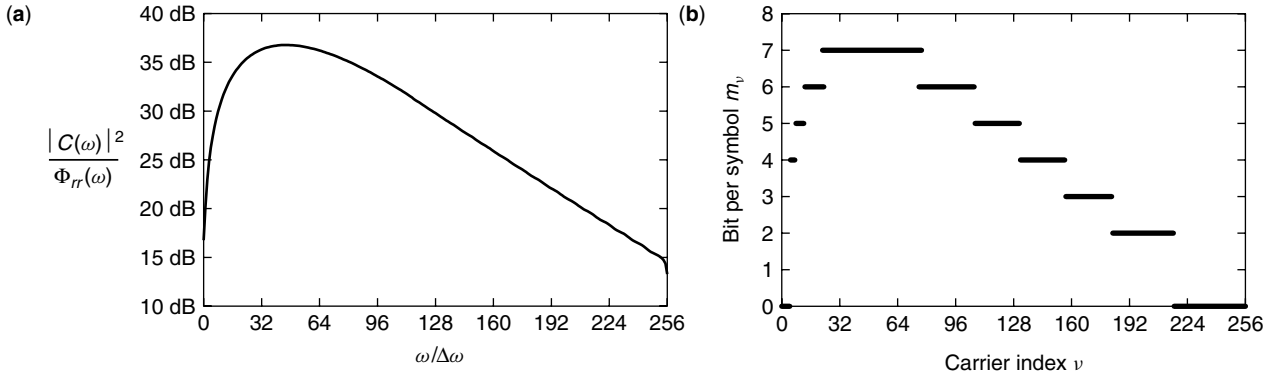


Figure 13. (**a**) SNR in the DMT subchannels at the receiver for a subscriber line of 3 km length and 0.4-mm wire diameter for an asymmetric digital subscriber line (ADSL); (**b**) adaptive bitloading for DMT modulation for transmission over a subscriber line 3 km length and 0.4 mm diameter.

BIOGRAPHIES

Romed Schur received his Dipl.-Ing. degree in electrical engineering and information technology from the University of Stuttgart, Germany, in 1997. He joined the Institute for Telecommunications at the University of Stuttgart as a research and teaching assistant in 1997 and is currently working toward his Ph.D. His interests include multicarrier modulation, xDSL transmission, as well as filterbanks and their applications to communication.

Stephan Pfletschinger studied electrical engineering and information technology at the University of Stuttgart, Germany, and received the Dipl.-Ing. degree with distinction in 1997. Since then he has been with the Institute of Telecommunications at the University of Stuttgart as a teaching and research assistant and is working toward his Ph.D. His research interests include multicarrier modulation and cable TV networks.

Joachim Speidel studied electrical engineering and information technology at the University of Stuttgart, Germany, and received his Dipl.-Ing. and Dr.-Ing. degree in 1975 and 1980, respectively, all with summa cum laude. From 1980 to 1992 he worked for Philips Communications (today Lucent Technologies Bell Labs Innovations) in the field of digital communications, ISDN, and video coding. During his industry career he has held various positions in R&D, as a member of technical staff, laboratory head, and finally as vice president. Since autumn 1992 he has been full professor at the University of Stuttgart and Director of the Institute of Telecommunications. His research areas are digital multimedia communications in mobile, optical, and electrical networks with emphasis on systems, modulation, source and channel coding.

BIBLIOGRAPHY

1. B. R. Saltzberg, Performance of an efficient parallel data transmission system, *IEEE Trans. Commun. Technol.* **COM-15**: 805–811 (Dec. 1967).
2. S. B. Weinstein and P. M. Ebert, Data transmission by frequency-division multiplexing using the discrete Fourier transform, *IEEE Trans. Commun. Technol.* **COM-19**(5): 628–634 (Oct. 1971).
3. I. Kalet, The multitone channel, *IEEE Trans. Commun.* **37**(2): 119–124 (Feb. 1989).
4. J. A. C. Bingham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* **28**(5): 5–14 (May 1990).
5. J. S. Chow, J. C. Tu, and J. M. Cioffi, A discrete multitone transceiver system for HDSL applications, *IEEE J. Select. Areas Commun.* **9**(6): 895–908 (Aug. 1991).
6. A. N. Akansu, P. Duhamel, X. Lin, and M. de Courville, Orthogonal transmultiplexers in communications: a review, *IEEE Trans. Signal Process.* **46**(4): 979–995 (April 1998).
7. J. M. Cioffi et al., Very-high-speed digital subscriber lines, *IEEE Commun. Mag.* **37**(4): 72–79 (April 1999).
8. ITU-T Recommendation G.992.1, *Asymmetric Digital Subscriber Line (ADSL) Transceivers*, June 1999.
9. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 2000.
10. G. Cherubini, E. Eleftheriou, S. Ölçer, and J. M. Cioffi, Filter bank modulation techniques for very high-speed digital subscriber lines, *IEEE Commun. Mag.* **38**(5): 98–104 (May 2000).
11. K. D. Kammeyer, U. Tuisel, H. Schulze, and H. Bochmann, Digital multicarrier-transmission of audio signals over mobile radio channels, *Eur. Trans. Telecommun.* **3**(3): 243–253 (May–June 1992).
12. B. R. Saltzberg, Comparison of single-carrier and multitone digital modulation for ADSL applications, *IEEE Commun. Mag.* **36**(11): 114–121 (Nov. 1998).
13. T. Pollet and M. Peeters, Synchronization with DMT modulation, *IEEE Commun. Mag.* **37**(4): 80–86 (April 1999).
14. L. J. Cimini, Jr. and N. R. Sollenberger, Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences, *Proc. IEEE ICC'99*, June 1999, Vol. 1, pp. 511–515.
15. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
16. P. J. W. Melsa, R. C. Younce, and C. E. Rohrs, Impulse response shortening for discrete multitone transceivers, *IEEE Trans. Commun.* **44**(12): 1662–1672 (Dec. 1996).

17. R. Schur, J. Speidel, and R. Angerbauer, Reduction of guard interval by impulse compression for DMT modulation on twisted pair cables, *Proc. IEEE Globecom'00*, Nov. 2000, Vol. 3, pp. 1632–1636.

18. W. Henkel, Maximizing the channel capacity of multicarrier transmission by suitable adaption of time-domain equalizer, *IEEE Trans. Commun.* **48**(12): 2000–2004 (Dec. 2000).

19. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (July, Oct. 1948).

20. K. F. Riley, M. P. Hobson, and S. J. Bence, *Mathematical Methods for Physics and Engineering*, Cambridge Univ. Press, Cambridge, UK, 1998.

21. U.S. Patent 4,679,227 (July, 1987), D. Hughes-Hartogs, Ensemble modem structure for imperfect transmission media.

22. P. S. Chow, J. M. Cioffi, and J. A. C. Bingham, A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels, *IEEE Trans. Commun.* **43**(2–4): 773–775 (Feb.–April 1995).

23. R. F. H. Fischer and J. B. Huber, A new loading algorithm for discrete multitone transmission, *Proc. IEEE Globecom'96*, Nov. 1996, Vol. 1; pp. 724–728.

24. R. V. Sonalkar and R. R. Shively, An efficient bit-loading algorithm for DMT applications, *IEEE Commun. Lett.* **4**(3): 80–82 (March 2000).

25. S. V. Ahamed, P. L. Gruber, and J.-J. Werner, Digital subscriber line (HDSL and ADSL) capacity of the outside loop plant, *IEEE J. Select. Areas Commun.* **13**(9): 1540–1549 (Dec. 1995).

DWDM RING NETWORKS

DETLEF STOLL
JIMIN XIE
Siemens ICN, Optisphere Networks
Boca Raton, Florida

JUERGEN HEILES
Siemens Information & Communication Networks
Munich, Germany

1. INTRODUCTION

Ring networks are well known and widely used in today's Synchronous Optical Network (SONET) or Synchronous Digital Hierarchy (SDH) transport networks [1]. They are attractive because they offer reliable and cost-efficient transport. In combination with dense wavelength-division multiplexing (DWDM), ring networks provide high capacity and transparent transport for a variety of client signals. Compared to a star and bus topology, the ring topology provides two diverse routes between any two network nodes. Ring network topologies are therefore often used in combination with protection schemes for increased reliability. Additionally, network management, operation and node design are less complex in rings than in mesh topologies.

DWDM is the key technology to provide high transmission capacity by transmitting many optical channels simultaneously over the same optical fiber. Each channel uses a dedicated optical wavelength (or, equivalently, color or frequency) [2]. Multiplexing in the frequency domain has been used for many decades in the communication industry, including radio or TV broadcast, where several channels are multiplexed and transmitted over the air or cable. The information signals are modulated at the transmitter side on carrier signals of different frequencies (wavelengths), where they are then combined for transmission over the same medium. At the receiver side, the channels are selected and separated using bandpass filters. More recent advances in optical technology allow for the separation of densely spaced optical channels in the spectrum. Consequently, more than 80 optical channels can be transmitted over the same fiber, with each channel transporting payload signals of $\geq$2.5 Gbps (gigabits per second). In combination with optical amplification of multiplexed signals, DWDM technology provides cost-efficient high-capacity long-haul (several 100 km) transport systems. Because of its analog nature, DWDM transport is agnostic to the frame format and bit rate (within a certain range) of the payload signals. Therefore, it is ideal in supporting the transportation of various high–speed data and telecommunication network services, such as Ethernet, SONET, or SDH.

These advantages fit the demands of modern communication networks. Service providers are facing the merging of telecommunication and data networks and a rapid increase in traffic flow, driven mainly by Internet applications. The level of transmission quality is standardized in the network operation business, or, commonly agreed. There are few possibilities to achieve competitive advantages for network operators. But these differences, namely, better reliability of service or flexibility of service provisioning, have a major influence on the market position. This explains why the transportation of telecommunication and data traffic is a typical volume market [3, p. 288]. As a default strategy in volume markets, it is the primary goal to provide highest data throughput at the most competitive price in order to maximize market share. As a secondary goal, network operators tend to provide flexibility and reliability as competitive differentiators. DWDM ring networks fulfill these basic market demands in an almost ideal way.

In the following sections, we will approach DWDM ring networks from three different viewpoints:

- *The Network Architecture Viewpoint*. DWDM ring networks are an integral part of the worldwide communication network infrastructure. The network design and the architecture have to meet certain rules in order to provide interoperability and keep the network structured, reliable, and manageable. The DWDM ring network architecture, the network layers, and specific constrains of DWDM ring networks will be discussed in Section 2. We will also take a look at protection schemes, which are an important feature of ring networks.
- *The Network Node Viewpoint*. Optical add/drop multiplexers (OADMs) are the dominant type of network elements in DWDM ring networks. The ring features will be defined mainly by the OADM

functionality. The basic OADM concept and its functionality is discussed in Section 3.

- *The Component Viewpoint*. Different component technologies can be deployed to realize the OADM functionality. It is the goal of system suppliers, to utilize those components that minimize the cost involved in purchasing and assembling these components and maximize the functionality of the OADM and DWDM ring network. State-of-the-art technologies and new component concepts are presented in Section 4.

2. DWDM RING NETWORK ARCHITECTURE

2.1. Basic Architecture

The basic ring architecture is independent of transport technologies such as SONET, SDH, or DWDM. Two-fiber rings (Fig. 1, *left*) use a single fiber pair; one fiber for each traffic direction between any two adjacent ring nodes. Four-fiber rings (Fig. 1, *right*) use two fiber pairs between each of the nodes. The second fiber pair is dedicated to protection or low priority traffic. It allows for enhanced protection schemes as discussed in Section 2.4.

The optical channels are added and dropped by the OADM nodes. Signal processing is performed mainly in the optical domain using wavelength-division multiplexing and optical amplification techniques. All-optical processing provides transparent and cost-efficient transport of multiwavelength signals. However, it has some constraints:

- *Wavelength Blocking*. In order to establish an optical path, either the wavelength of this path has to be available in all spans between both terminating add/drop nodes or the wavelength has to be converted in intermediate nodes along the path. Otherwise, wavelength blocking occurs. It results in lower utilization of the ring resources. It can be minimized, but not avoided completely by careful and farsighted network planning and wavelength management. It has been shown that the positive effects of wavelength conversion in intermediate nodes are limited [4].
- *"Signal Quality Supervision."* As today's services are mainly digital, the bit error ratio (BER) is the preferred quality-of-service information. All-optical performance supervision does not provide a parameter that directly corresponds to the BER of the signal.
- *Optical Transparency Length*. This is the maximum distance that can be crossed by an optical channel. The transparency length is limited by optical signal impairments, such as attenuation of optical signal power (loss), dispersion, and nonlinear effects. It depends on the data rate transported by the optical channels, the ring capacity, the quality of the optical components, and the type of fiber (e.g., standard single-mode fiber, dispersion shifted fiber). For 10-Gbps signals, transparency lengths of 300 km can be achieved with reasonable effort. This makes DWDM ring networks most attractive for metropolitan-area networks (MAN; $\leq$80 km ring circumference) and regional area networks ($\leq$300 km ring circumference).

2.2. Transparent Networks and Opaque Networks

If the optical transparency length is exceeded, full regeneration of the signal, including reamplification, reshaping, and retiming (3R regeneration) is necessary. With 3R regeneration however, the transport capability is restricted to certain bit rates and signal formats. In this case, the ring is called opaque. Specific digital processing equipment must be available for each format that shall be transported. Furthermore, 3R regeneration is very costly. The typical opaque DWDM ring topology uses all-optical processing within the ring. 3R regeneration is only performed at the tributary interfaces where the signals enter and exit the ring (i.e., at the add/drop locations). This decouples the optical path design in the ring from any client signal characteristic. Furthermore, it provides digital signal quality supervision the add/drop traffic. Support of different signal formats is still possible by using the appropriate tributary interface cards without the need for reconfiguration at intermediate ring nodes.

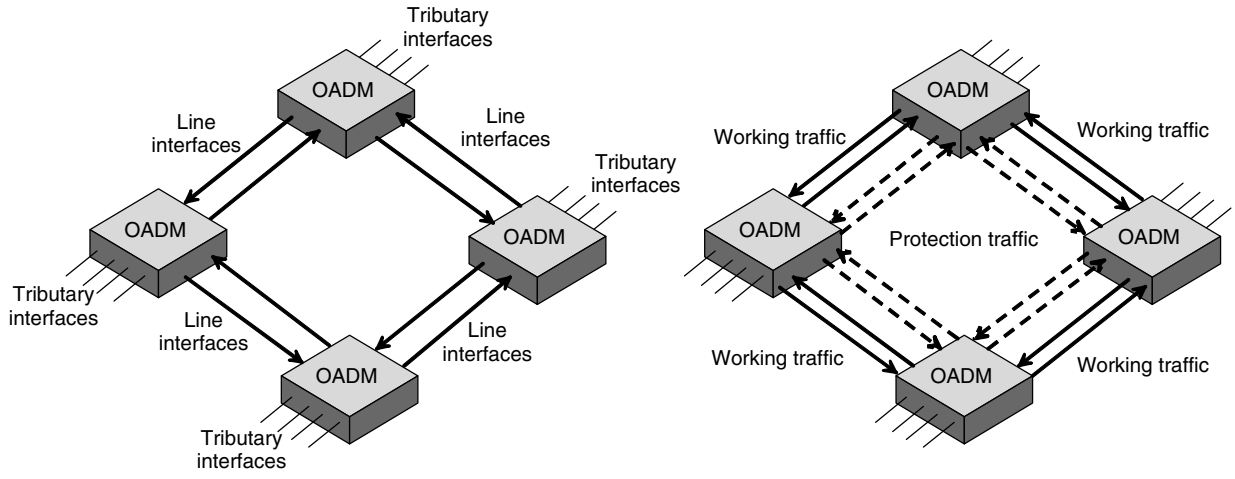


Figure 1. Basic ring architectures: *left*, two-fiber ring; *right*, four-fiber ring.

2.3. Optical-Layer Networks

The International Telecommunication Union (ITU) has standardized the architecture of the Optical Transport Network (OTN) in its recommendation ITU-T G.872 [5]. G.872 defines three optical layer networks: the optical transmission section (OTS) layer, the optical multiplex section (OMS) layer, and the optical channel (OCh) layer network. Figure 2 shows the mapping of these layer networks on a DWDM ring network.

- The OTS represents the DWDM signal transmitted between two nodes (OADM) and/or optical line amplifiers (OLA). Intermediate optical line amplifiers can be used to amplify the signal in cases of long distances between two adjacent nodes.
- The OMS represents the DWDM signal between two adjacent nodes.
- The OCh represents a single optical channel between the 3R regeneration points. For the case of the typical opaque ring architecture, the OCh is terminated at the tributary interfaces of the OADM nodes.

Supervision and management functions are defined for each network layer. Overhead that supports these functions is transported on an optical supervisory channel, which uses a dedicated wavelength for transmission between the nodes. ITU-T G.872 defines also two digital layer networks for a seamless integration of optical and digital processing. These digital layers, the optical channel transport unit (OTU) and the optical channel data unit (ODU) are further described in ITU–T recommendation G.709 [6]. They provide the following three functions: digital signal supervision, time division multiplexing for the aggregation of lower bit rate signals, and forward error correction (FEC) [7]. FEC allows for extending the optical transparency length for a given signal quality.

2.4. Protection

Ring networks support protection switching ideally, since there are always two alternate paths between any two ring nodes. Ring protection schemes are already defined for SONET and SDH [1,8]. Similar concepts can be applied to DWDM ring networks.

$1+1$ *optical channel* (*OCh*) *protection* is the most simple ring protection scheme. It belongs to the class of "dedicated protection," as it uses a dedicated protection connection for each working connection. At the add ring node, the traffic is bridged (branched) to both the working connection and the protection connection. The drop ring nodes select between the two connections based on the quality of the received signals. As a result, $1+1$ OCh protected connections always consume one wavelength per transmission direction in the entire ring, whereas unprotected connections use this wavelength on the shortest path between the terminating nodes only. The utilization of the ring capacity for working traffic can maximally reach 50%.

Optical multiplex section bidirectional self-healing ring (*OMS-BSHR*) belongs to the class of "shared protection," as it shares the protection connection between several working connections. BSHR is also known in SONET as bidirectional line switched ring (BLSR) [9]. OMS-BSHR can be used in both two- and four-fiber rings. In case of a two-fiber BSHR (Fig. 3), One-half of the wavelength channels are used for the working connections; the other half are used for protection. In order to avoid wavelength converters at the protection switches, working and protection wavelengths of opposite directions should be the same (flipped wavelength assignment). For example, in a DWDM ring of N wavelength channels, the wavelengths 1 to $N/2$ (lower band) are used for working connections whereas the wavelengths $N/2+1$ to N (upper band) are used for protection connections. This working

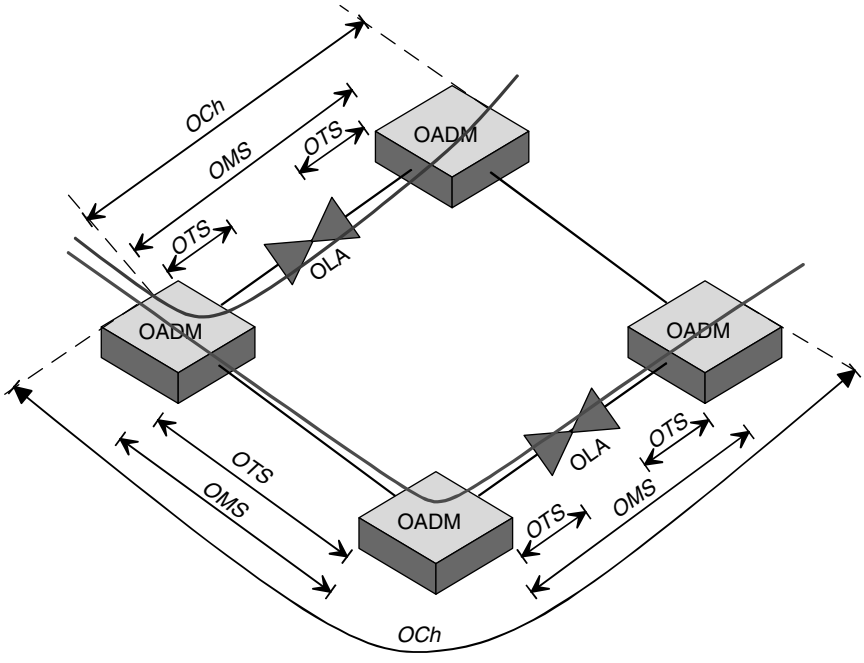


Figure 2. Optical-layer networks.

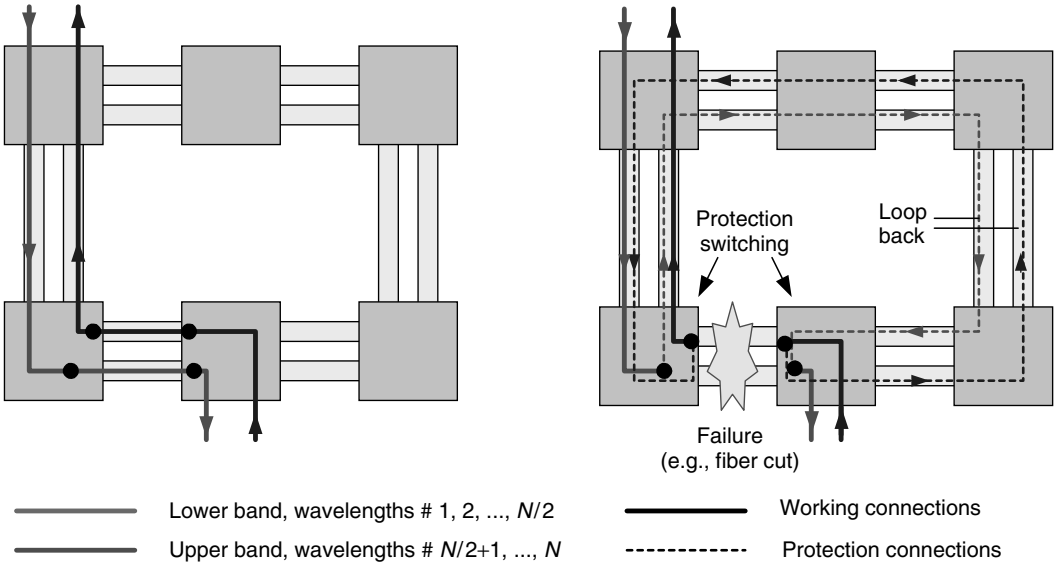


Figure 3. Two-fiber OMS-BSHR: *left*, normal operation; *right*, protection case.

and protection traffic runs in a clockwise direction. In the counter clockwise direction, the wavelengths $N/2+1$ to N are used for working connections and the wavelengths 1 to $N/2$ are used for protection. In case of a failure, the two nodes adjacent to the failure location perform ring protection switching (Fig. 3, *right*). They bridge and select the clockwise working connection to and from the counter clockwise protection connection and vise versa. The whole working traffic that passed the failure location is now looped back to the other direction of the ring.

In a four-fiber BSHR, dedicated fibers for working traffic and protection traffic are used. By this, four-fiber BSHR rings provide twice the capacity than the two-fiber BSHR rings. In addition to the ring protection of the two-fiber BSHR, the four-fiber BSHR can perform span protection. This provides protection against failures of the working fiber pair between two adjacent nodes. An "automatic protection switching" (APS) protocol is necessary for the coordination of the bridge and switch actions and for the use of the shared protection connection. Unlike in $1+1$ OCh protecting the utilization of the ring capacity for working traffic is always exactly 50%. A disadvantage of OMS-BSHR are the long pathlengths that are possible due to the loop back in the protection case. In order not to exceed the transparency length in a case of protection, the ring circumference is reduced.

The *bidirectional optical channel shared protection ring* (*OCh-SPR*) avoids the loop back problem of the OMS-BSHR. Here, ring protection is performed by the add and drop nodes of the optical channel. The add nodes switch the add channels to the working or protection connection. The drop nodes select the working or protection channels directly (Fig. 4).

OCh-SPR can be used in both the two- and four-fiber rings. An automatic protection switching protocol is required for the coordination of the bridge and switch actions; as well as, for the use of the shared protection connection. In an OCh-SPR, protection switches for each optical channel are needed, while the OMS-BSHR allows the use of a protection switch for all working/protection connections in common.

A general issue of shared protection schemes in all-optical networks is the restricted supervision capability of inactive protection connections. Failures of protection connections that are not illuminated can be detected only after activation. Usually, activation occurs due to a protection situation. In that case, the detection of failures in the protection connection comes too late. Furthermore, the dynamic activation/deactivation of protection channels on a span has an influence on the other channels of the span. It will result in bit errors for these channels if not carefully performed.

2.5. Single-Fiber Bidirectional Transmission

DWDM allows bidirectional signal transmission over a single fiber by using different wavelengths for the two signal directions. This allows for building two-fiber ring architectures by only using a single fiber between the ring nodes. It also allows for four-fiber architectures with two physical fibers. However, the two-fiber OMS-BSHR and OCh-SPR protection schemes with their sophisticated wavelength assignment are not supported.

3. OADM NODE ARCHITECTURE

The network functionalities described in the previous section are realized in the optical add/drop multiplexers. The OADM node architecture and functionality is described below.

3.1. Basic OADM Architecture

An OADM consists of three basic modules as shown in Fig. 5:

- The *DWDM line interfaces*, which physically process the ring traffic

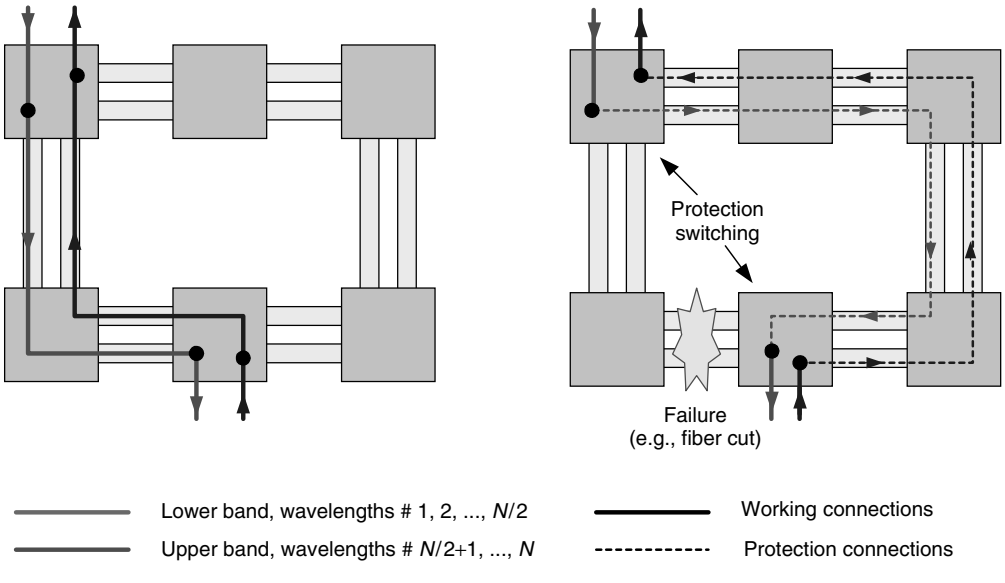


Figure 4. Two-fiber OCh shared protection: *left*, normal operation; *right*, protection case.

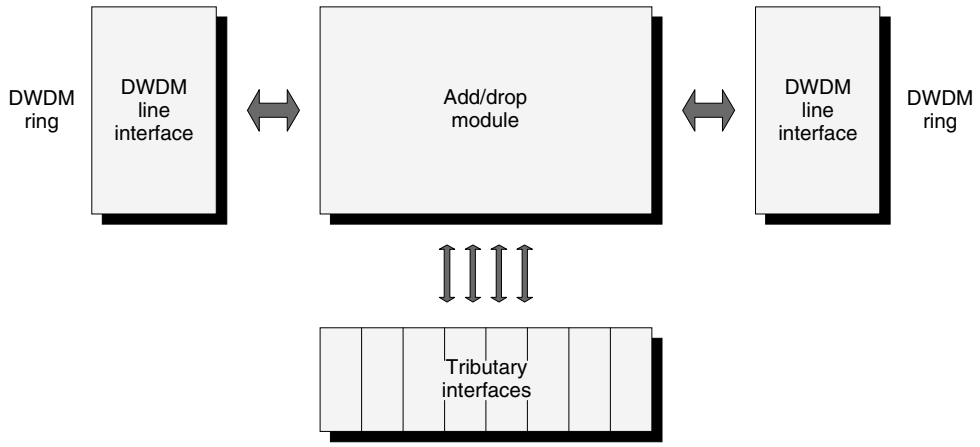


Figure 5. Basic OADM architecture.

- The *add/drop module*, which performs both through connections and the add/drop function for the optical channels
- The *tributary interfaces*, which physically process the add/drop channels

The detailed functionality of these modules depends on the OADM features.

3.2. Characteristic OADM Features

This section describes the most important features of the OADM nodes.

3.2.1. Add/Drop Capacity. The add/drop capacity is defined as the relationship between the capacity of the tributary interfaces and the capacity of both line interfaces. In an extreme case, the entire traffic entering the OADM from both sides of the ring must be dropped. Therefore, an add/drop multiplexer is considered to have 100% add/drop capacity if the capacity of the tributary interfaces equals the capacity of both line interfaces. For instance, an OADM of 100% add/drop capacity in an 80-channel DWDM ring provides 160 channels at its tributary interfaces. Designing an OADM for an appropriate add/drop capacity is an efficient approach in minimizing cost and complexity.

3.2.2. Flexibility. The selection of add/drop channels in an OADM can either be remotely configurable for channels that require frequent reconfiguration (dynamic), or it can be performed manually for "long term" connections (static). Furthermore, channels that will never be dropped can be passed through directly from one line interface to the other without add/drop capabilities (express channels). Other important OADM features are

- *Transparency* or *opacity* as discussed in Section 2.2.
- *Wavelength conversion* capability as discussed in Sections 2.

- *Protection switching* features as discussed in Section 2.4.
- *Signal supervision* capabilities can be based on optical parameters only, such as optical power and optical signal-to-noise ratio (OSNR). Additionally, in opaque rings, signal supervision can be based on bit error ratio monitoring.

3.3. The DWDM Line Interfaces

The main functionality of the DWDM line interfaces includes, primarily the compensation for physical degradation of the DWDM signal during transmission and secondarily the supervision of the DWDM signal. There are three major categories of degradation: (1) loss of optical signal power due to attenuation, (2) dispersion of the impulse shape of high-speed optical signals, and (3) impulse distortion or neighbor channel crosstalk due to nonlinear fiber-optical effects. To overcome these degradations, optical amplifiers and dispersion compensation components may be used. For metropolitan applications, moderate-gain amplifiers can be used to reduce the nonlinear fiber-optical effects, since the node distance is relatively short. Design approaches are mentioned in the references [2,10]. In the line interfaces, signal supervision is based on optical parameters (e.g., optical power, OSNR).

3.4. The Add/Drop Module

The add/drop module performs the following major functions:

- It configures the add/drop of up to 100% of the ring traffic.
- It converts the wavelengths of the add/drop and the through signals (optional).
- It protects the traffic on the OMS or OCh level (optional).

3.4.1. Optical Channel Groups. Optical channels can be classified as dynamic, static, or express traffic. In order to minimize the hardware complexity, in other words installation costs, traffic classes are assigned to optical channel groups accordingly. The optical channel groups are processed according to their specific characteristics. Figure 6 illustrates the group separation (classes) of DWDM traffic by group multiplexer and demultiplexer components. The groups are processed within the add/drop module using different techniques. Methods for defining groups and related component technologies are discussed in Section 4 based on technological boundary conditions.

For dynamic traffic, wavelength routing technologies, such as remotely controlled optical switching fabrics or tunable filters, are used. For static traffic, *manual configuration* via distribution frames (patch panels) is employed. Wavelength routing necessitates higher installation costs than manual configuration. However, it enables the reduction of operational costs due to automation. Operational costs can become prohibitive for manual configuration if frequent reconfiguration is necessary. A cost-optimal OADM processes part of the traffic via distribution frames and the other part via wavelength routing techniques. Hence, the use of costly components is focused on cases where they benefit most [11]. Designing an OADM for a specific add/drop capacity is very efficient especially in the wavelength routing part of the add/drop module. Figure 7 shows a possible architecture of a wavelength routing add/drop matrix.

The add/drop process of the dynamic traffic can be performed in two stages: the add/drop stage and the distribution stage. In order to perform wavelength conversion, wavelength converter arrays (WCA) can be used in addition. The add/drop stage is the only requirement. The distribution stage and the wavelength converters are optional.

The *add/drop-stage* filters the drop channels out of the DWDM signal and adds new channels to the ring. Single optical channels are typically processed. But it is also possible to add or to drop optical channel groups to

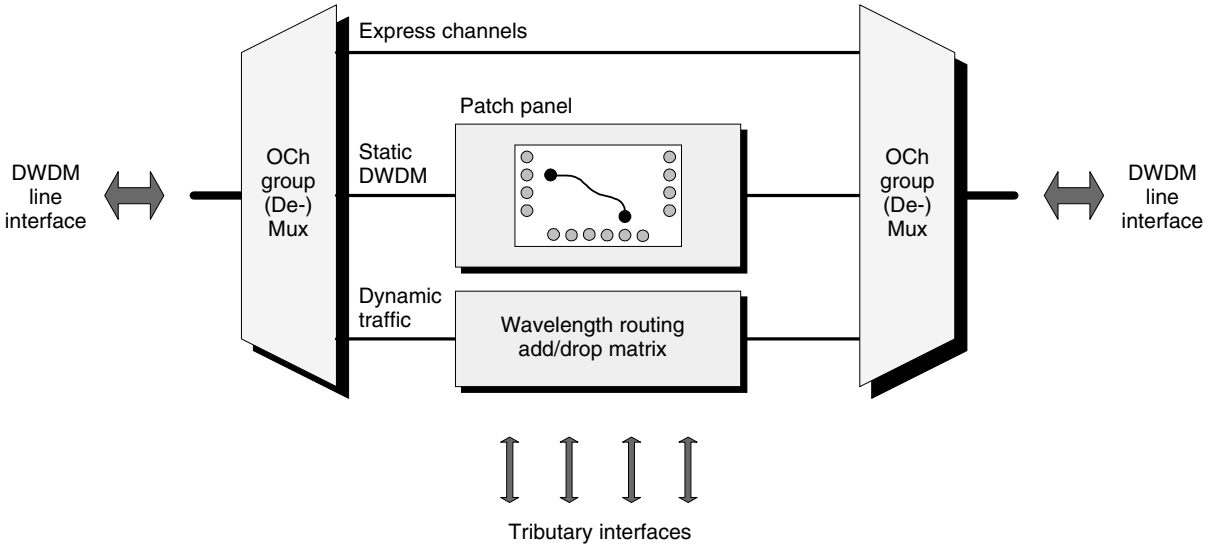


Figure 6. Add/drop module — processing according to traffic characteristics.

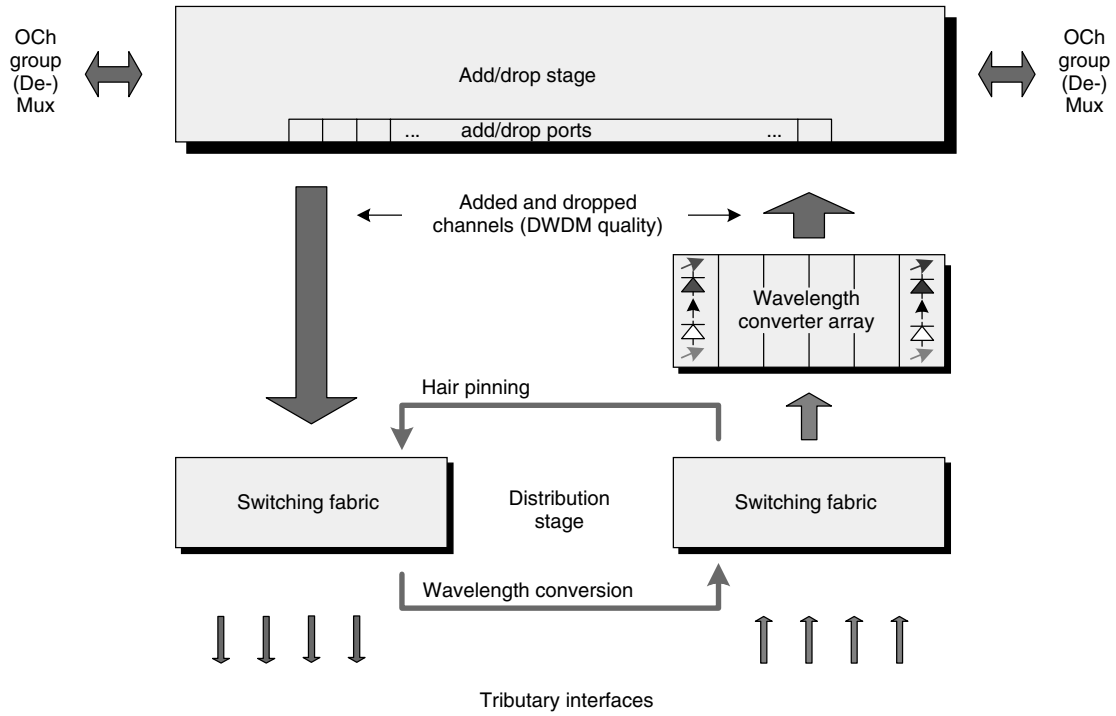


Figure 7. Wavelength routing add/drop matrix (two-stage).

a single port. Various technologies for realizing add/drop stages will be discussed in Section 4.2.

The *wavelength converter array* can be used for wavelength assignment of add channels from the tributary interfaces and for wavelength conversion of through channels. Technologies are discussed in Sections 4.3 and 4.4.

The distribution stage performs the following functions:

1. *Assignment of Channels to Specific Tributary Interfaces*. Not every tributary signal must be connected to the ring at anytime. There may be "part-time" leased lines that share the same wavelength channel such as one at business hours other one at nighttime or at weekends. For example, one customer uses the wavelength channel at the distribution stage connects tributary signals to specific add channels for insertion into the ring. For extraction, it connects them to specific drop channels.
2. *Wavelength Conversion for the Through Channels*. Dropped channels can be reinserted on another wavelength via the link "wavelength conversion" in Fig. 7. A WCA is required for this application.
3. *Hairpinning*. In some applications, add/drop multiplexers are used to route signals between tributary interfaces. This function is known as "hairpinning." It can be realized by the connection as shown in Fig. 7.
4. *Flexible Wavelength Assignment for the Add Channels*. Besides wavelength conversion, the WCA can be used to assign a specific wavelength to an incoming tributary signal. The distribution stage allows for flexible assignment of tributary signals to a WCA element of a specific wavelength. This function can be part of the tributary interfaces as well.

Techniques of switching fabrics to perform these functions are mentioned in Section 4.3.

3.5. The Tributary Interfaces

The tributary interfaces prepare the incoming signals for transmission over the DWDM ring. Incoming signals do not necessarily have DWDM quality. Furthermore, they monitor the quality of the incoming and outgoing tributary signals. It is a strong customer requirement to support a large variety of signal formats and data rates as tributary signals. In the simplest case the tributary signal is directly forwarded to the add/drop module (transparent OADM). In this case, the incoming optical signal must have the correct wavelength to fit into the DWDM line signal; otherwise, the signal has to enter the ring via wavelength converters, either as part of the tributary interface or via the WCA of the add/drop module. The simplest way of signal supervision is monitoring the optical power (loss of signal) of the incoming and outgoing tributary signals. A more advanced approach in the optical domain is the measurement of the optical signal-to-noise ratio. If digital signal quality supervision (i.e., bit error ratio measurement) is needed, 3R regeneration has to be performed. This normally requires client signal-specific processing using dedicated tributary ports for different client formats, such as SONET or Gigabit Ethernet.

4. OADM COMPONENTS

This section focuses on component technologies of the add/drop module and the tributary interfaces.

4.1. Group Filters and Multiplexer/Demultiplexer Components

A multiplexer/demultiplexer in the sense of a WDM component is a device that provides one optical fiber port for DWDM signals and multiple fiber ports for optical channels or channel groups. These multiplexers/demultiplexers are passive components that work in both directions. They can be used for either separating DWDM signals or for combining optical channels to one DWDM signal. Group filters and WDM multiplexers/demultiplexers are basically the same type of component. They just have different optical filter characteristics. A WDM multiplexer/demultiplexer combines or separates single optical channels whereas a group filter combines or separates optical channel groups. Figure 8 shows three basic realizations of wavelength multiplexer/demultiplexer filters.

A basic advantage of block channel groups as opposed to interleaved optical channel groups is the stronger suppression of neighbor group signals. Nevertheless, we will see in Section 4.2 that there are applications in which interleaved channel groups support the overall system performance better than channel blocks. Static multiplexer/demultiplexer filters are usually based on technologies such as multilayer dielectric thin-film filter (TFF), fixed fiber bragg grating (FBG), diffraction grating, arrayed waveguide grating (AWG), cascaded Mach–Zehnder interferometer, and Fabry–Perot interferometer. A detailed discussion of these technologies is given in Section 3.3 of Ref. 2. Interleaver filter technologies are described in Refs. 12 and 13.

4.2. Add/Drop Filter Technologies

Figure 9 shows two popular realization concepts of add/drop filters: (*left*) wavelength multiplexer/demultiplexer — optical switch combination and (*right*) tunable filter cascade — circulator/coupler combination.

The multiplexer/demultiplexer–optical switch solution usually demultiplexes the entire DWDM signal and provides single optical channels via switches at each add/drop port. The tunable filter cascade (Fig. 9, *right*) provides exactly the same functionality. The incoming DWDM signal is forwarded clockwise to the tunable filter cascade by a circulator. The tunable filters reflect selected channels. All other channels pass the filter cascade. The reflected channels travel back to the circulator to be forwarded to the drop port, again in a clockwise motion [14]. For adding optical channels, either a coupler

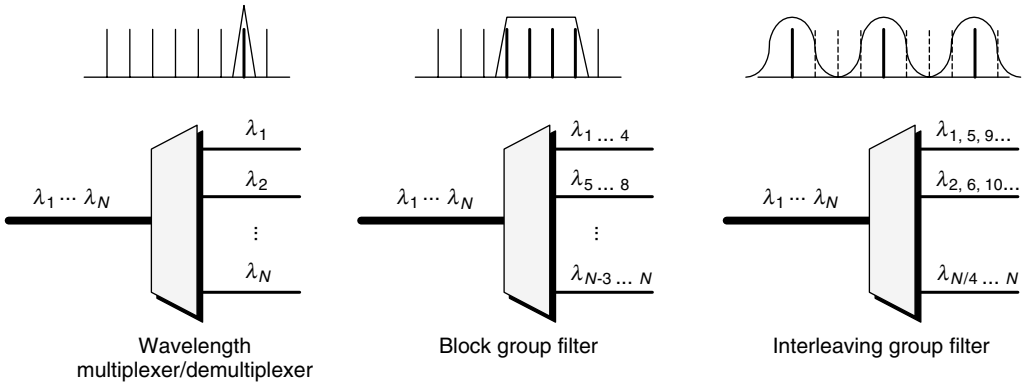


Figure 8. Static wavelength multiplexer/demultiplexer filters.

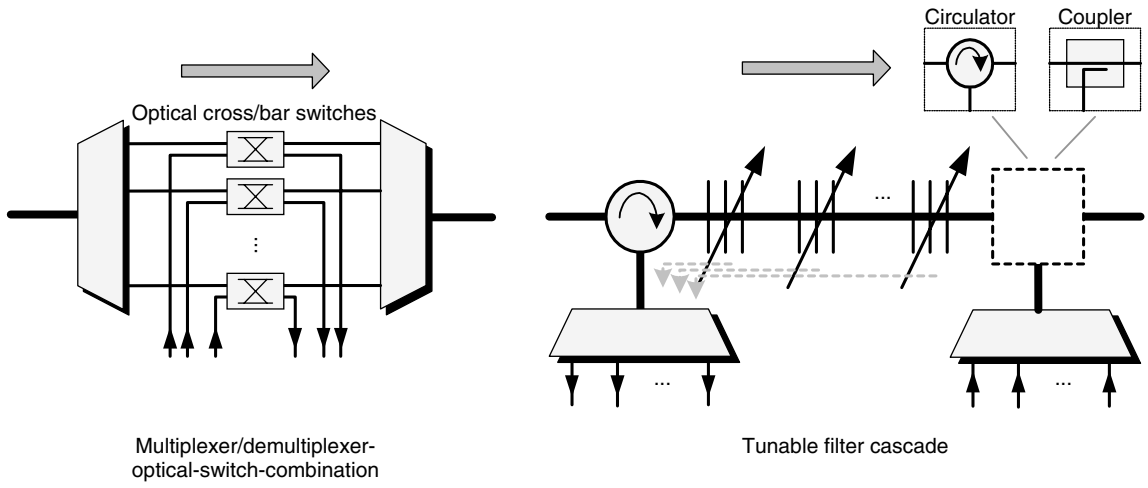


Figure 9. Add/drop filter technologies.

or a circulator can be used. A coupler provides better isolation between the add ports and the drop ports. It is also the least expensive approach. On the other hand, circulators allow for a lower transmission loss. Here, the advantage of interleaved channel groups becomes evident: the bandwidth requirements for tunable filters are lower. In tunable filter cascades, the reconfiguration process can affect uninvolved traffic. For example, if a filter is tuned from wavelength 1 to wavelength 4, it passes the wavelengths 2 and 3. Therefore, traffic running on these channels is affected by the retuning process and thus may render it unusable. The influence of the dispersion of the tunable filters is strong compared to the multiplexer/demultiplexer–optical switch solution. This is another drawback of tunable filter cascades. If tunable filters are tuned to a neutral park position between two densely spaced channels, these channels can be degraded by the dispersion of the filter. This design issue can be overcome by using interleaved channel groups. Wider spacing between the channels lowers the bandwidth requirements for the filters, and thus resulting in a lower dispersion. The basic advantage of tunable add/drop filters is the lower number of internal fiber connections and splices. However, because of the typical characteristics of fiber gratings (e.g., cladding modes), this application is limited to small and medium size channel groups. A more detailed comparison of both concepts is given in Ref. 15.

4.3. Optical Switching Fabrics

Optical switching fabrics provide cross–connection functionality. They are the technology of choice in distribution stages of optical add/drop multiplexers (see Fig. 7). From the viewpoint of OADM optical signal processing, we can distinguish between protection switching and wavelength routing applications.

Protection switching requires switching times of a few milliseconds. Switching times up to hundreds of milliseconds are sufficient for wavelength routing applications. For protection switching and wavelength routing applications, electromechanical, thermooptic, electrooptic, and acoustooptic switches can be used, see Ref. 2, Section 3.7 and Ref. 16, Chap. 10. Today, among the electromechanical switches, the MEMS technology (microelectromechanical system) is seen as the most promising technology in providing high port counts. Switches of more than 100×100 ports have been realized.

Tunable filters, as discussed in Section 4.2, are also an alternative technology to switching fabrics in order to realize cross–connection functionality. Optical circuits using tunable filters and circulators have been presented by Chen and Lee [17].

4.4. Wavelength Converter Technologies

Today, wavelength conversion is performed by O/E/O (optical/electrical/optical) conversion using photodiodes and tunable or fixed-wavelength lasers that are either modulated externally or internally (directly). Tunable lasers allow for covering a wide range of wavelengths and to minimize the number of wavelength converters that are needed. An overview about tunable laser technology is given in Ref. 18. The optical receiver and transmitter components can be connected via analog electrical amplifiers allowing for the transmission of any signal format up to a certain bit rate. If 3R regeneration is performed (digital signal processing between receiver and transmitter), a further limitation to certain signal formats may apply. In future transparent systems, the use of all-optical wavelength converters is expected. All-optical wavelength converters make use of nonlinear optical effects such as cross-gain modulation (CGM), cross phase modulation (XPM) or four-wave mixing (FWM). These effects are treated in detail in Ref. 19, Section 2.7. Semiconductor optical amplifiers (SOA) are the preferred active medium as they exhibit strong nonlinearity, wide gain bandwidth and easy integration. As a pump source, unmodulated lasers are used. Tunable pump lasers provide the capability of tunable all-optical wavelength converters.

5. SUMMARY

DWDM ring networks provide high capacity, high flexibility (multiservice integration), and high reliability (protection) at low operational costs to the operators of metropolitan and regional networks. Because of the simple ring topology, the network management is relatively less complex. The ring functions are determined mainly by the optical add/drop multiplexers. The DWDM line interface mainly determines the maximum ring circumference that can be achieved. An add/drop module, that provides manual routing capabilities, allows for low installation costs. But if frequent reconfiguration is necessary, operational costs can become prohibitive. For this type of traffic, wavelength routing capabilities that provide remotely controlled dynamic routing should be implemented. The technologies of choice for wavelength routing are integrated optical switching fabrics such as MEMS and tunable filters.

DWDM networks can either be transparent or opaque. In the transparent realization, no electronic traffic processing occurs within the ring. The transport is independent of the data format. In opaque networks, for quality-of-service supervision and management reasons, digital (electronic) processing is performed at the borders of the DWDM ring network. The transport in opaque networks may be limited to certain data formats.

The business success of network operation is driven mainly by an overall cost minimization by selling data transport in high volume. This is a general rule in markets that are characterized by low differentiation possibilities and high impact of small differences (volume markets). In the network operation business, differentiation is possible by better reliability, availability and flexibility. Therefore, DWDM ring networks are ideal in paving the way for the business success of metropolitan and regional network operators.

BIOGRAPHIES

Detlef Stoll received a Dipl.-Ing. degree (M.S.) in communication engineering in 1988 from the University of

Hannover, Germany, and a Ph.D. degree in electrical engineering in 1993 from the University of Paderborn, Germany. The subject of his Ph.D. thesis was the derivation of an integrated model for the calculation of the nonlinear transmission in single-mode optical fibers including all linear and nonlinear effects. He joined the Information & Communication Networks Division of the Siemens Corporation in 1994 as a research & development engineer. At Siemens, he worked on the advance development of broadband radio access networks and on the development of SONET/SDH systems. From 1998 to 2000 he led the advance development of a Wavelength Routing Metropolitan DWDM Ring Network for a multi-vendor field trial. Since 2000 he is managing the development of optical network solutions at Optisphere Networks, Inc. located in Boca Raton, Florida. Dr. Stoll has filed approximately 20 patents and published numerous articles in the field of optical networks and forward error correction techniques. His areas of interest are linear and nonlinear systems and wavelength routing networks.

Juergen Heiles received a Dipl.-Ing. (FH) degree in electrical engineering from the University of Rhineland-Palatinate, Koblenz, Germany, in 1986. He joined the Public Networks Division of the Siemens Corporation, Munich, Germany, in 1986 as a research & development engineer of satellite communication systems and, later on, fiber optic communication systems. Since 1998 he has been responsible for the standardization activities of the Siemens Business Unit Optical Networks. He participates in the ITU, T1, OIF and IETF on SONET/SDH, OTN, ASON, and GMPLS and is editor or coauthor of several standard documents. Juergen Heiles has over 15 years of experience in the design, development, and system engineering of optical communication systems starting from PDH systems over the first SONET/SDH systems to today's DWDM networks. He holds three patents in the areas of digital signal processing for optical communication systems. His areas of interest are network architectures and the functional modeling of optical networks.

Jimin Xie received a B.S. degree in electronics in 1986 from the Institute of Technology in Nanjing, China, and an M.S. degree in radio communications in 1988 from the Ecole Superieure d'Electricite in Paris, France, and a Ph.D. degree in opto-electronics in 1993 from the Université Paris-Sud, France. He worked on submarine soliton transmission systems for Alcatel in France from 1993 to 1995. In 1996, he worked at JDS Uniphase as a group leader in the Strategy and Research Department in Ottawa, Canada. At JDSU, he worked on the design and development of passive DWDM components, such as interleaver filters, DWDM filters, OADM modules, polarization detectors and controllers, dispersion compensators, gain flatteners, optical switches, etc. Dr. Xie has filed 12 patents in the field of optical components. Since 2001, he has been the manager of the Optical Technologies department at Optisphere Networks, Inc. located in Boca Raton, Florida. He participates in the OIF standardization activities for Siemens. His areas of interest are the new technologies of components and modules for optical networks.

BIBLIOGRAPHY

1. M. Sexton and A. Reid, *Broadband Networking: ATM, SDH, and SONET*, Artech House, Norwood, MA, 1997.
2. R. Ramaswami and K. N. Sivarajan, *Optical Networks*, Morgan Kaufmann, San Francisco, CA, 1998.
3. P. Kotler, *Marketing Management*, Prentice-Hall, Englewood Cliffs, NJ, 1999.
4. P. Arijs, M. Gryseels, and P. Demeester, Planning of WDM ring networks, *Photon. Network Commun.* **1**: 33–51 (2000).
5. ITU-T Recommendation G.872, *Architecture of the Optical Transport Networks*, International Telecommunication Union, Geneva, Switzerland, Feb. 1999.
6. ITU-T Recommendation G.709, *Interface for the Optical Transport Network* (*OTN*), International Telecommunication Union, Geneva, Switzerland, Oct. 2001.
7. I. S. Reed and X. Chen, *Error-Control Coding for Data Networks*, Kluwer, Norwell, MA, 1999.
8. ITU-T Recommendation G.841, *Types and Characteristics of SDH Network Protection Architectures*, International Telecommunication Union, Geneva, Switzerland, Oct. 1998.
9. GR-1230-CORE, *SONET Bidirectional Line-Switched Ring Equipment Generic Criteria*, Telcordia, 1998.
10. P. C. Becker, N. A. Olsson, and J. R. Simpson, *Erbium-Doped Fiber Amplifiers, Fundamentals and Technology*, Academic Press, San Diego, CA, 1997.
11. D. Stoll, P. Leisching, H. Bock, and A. Richter, Best effort lambda routing by cost optimized optical add/drop multiplexers and cross-connects, *Proc. NFOEC*, Baltimore, MD, Session A-1, July 10, 2001.
12. H. van de Stadt and J. M. Muller, Multimirror Fabry-Perot interferometers, *J. Opt. Soc. Am. A* **8**: 1363–1370 (1985).
13. B. B. Dingel and M. Izutsu, Multifunction optical filter with a Michelson-Gires-Tournois interferometer for wavelength-division-multiplexed network system applications, *Opt. Lett.* **14**: 1099–1101 (1998).
14. U.S. Patent 5,748,349 (1998), V. Mizrahi, Gratings-based optical add-drop multiplexers for WDM optical communication systems.
15. D. Stoll, P. Leisching, H. Bock, and A. Richter, Metropolitan DWDM: A dynamically configurable ring for the KomNet field trial in Berlin, *IEEE Commun. Mag.* **2**: 106–113 (2001).
16. I. P. Kaminow and T. L. Koch, *Optical Fiber Telecommunications III B*, Academic Press, Boston, MA, 1997.
17. Y.-K. Chen and C.-C. Lee, Fiber Bragg grating based large nonblocking multiwavelength cross-connects, *IEEE J. Lightwave Technol.* **16**: 1746–1756 (1998).
18. E. Kapon, P. L. Kelley, I. P. Kaminow, and G. P. Agrawal, *Semiconductor Lasers II*, Academic Press, Boston, MA, 1999.
19. G. P. Agrawal, *Fiber-optic Communication Systems*, Wiley, New York, NY, 1998.

E

EXTREMELY LOW FREQUENCY (ELF) ELECTROMAGNETIC WAVE PROPAGATION

Steven Cummer
Duke University
Durham, North Carolina

1. INTRODUCTION

Extremely low frequency (ELF) electromagnetic waves are currently the lowest frequency waves used routinely for wireless communication. The IEEE standard radio frequency spectrum defines the ELF band from 3 Hz to 3 kHz [1], although the acronym ELF is often used somewhat loosely with band limits near these official boundaries. Because of the low frequencies involved, the bandwidth available for ELF communication is very small, and the data rate of correspondingly low. Existing ELF communications systems with signal frequencies between 40 and 80 Hz transmit only a few bits per minute [2]. Despite this severe limitation, ELF waves have many unique and desirable properties because of their low frequencies. These properties enable communication under conditions where higher frequencies are simply not usable. Like all electromagnetic waves excited on the ground in the HF (3–30 MHz) and lower bands, ELF waves are strongly reflected by the ground and by the Earth's ionosphere, the electrically conducting portion of the atmosphere above roughly 60 km altitude [3]. The ground and ionosphere bound a spherical region commonly referred to as the *earth–ionosphere* waveguide in which ELF and VLF (very low frequency, 3–30 kHz) propagate. The combination of strongly reflecting boundaries and long ELF wavelengths (3000 km at 100 Hz) enables ELF waves to propagate extremely long distances with minimal attenuation. Measured ELF attenuation rates (defined as the signal attenuation rate in excess of the unavoidable geometric spreading of the wave energy) are typically only ~1 dB per 1000 km at 70–80 Hz [4–6]. The signal from a single ELF transmitter can, in fact, be received almost everywhere on the surface of the planet, even though only a few watts of power are radiated by existing ELF systems. Because the wavelength at ELF is so long, ELF antennas are very inefficient and it takes a very large antenna (wires tens of kilometers long) to radiate just a few watts of ELF energy. This makes ELF transmitters rather large and expensive. Only two ELF transmitters currently operate in the world, one in the United States and one in Russia. Besides their global propagation, the value of ELF waves results from their ability to penetrate effectively through electrically conducting materials, such as seawater and rock, that higher frequency waves cannot penetrate. These properties make ELF waves indispensable for applications that require long-distance propagation and penetration through conducting materials, such as communication with submarines [7]. Unlike many other radiobands, there are strong natural sources of ELF electromagnetic waves. The strongest of these in most locations on the earth is lightning, but at high latitudes natural emissions from processes in the magnetosphere (the highest portions of the earth's atmosphere) can also be strong. There are a variety of scientific and geophysical applications that rely on either natural or man-made ELF waves, including subsurface geologic exploration [8], ionospheric remote sensing [9], and lightning remote sensing [10].

2. THE NATURE OF ELF PROPAGATION

Communication system performance analysis and most geophysical applications require accurate, quantitative models of ELF propagation. Because the ELF fields are confined to a region comparable to or smaller than a wavelength, phenomena in ELF transmission and propagation are substantially different from those at higher frequencies. But before delving too deeply into the relevant mathematics, a qualitative description will give substantial insight into the physics of ELF transmission, propagation, and reception.

2.1. Transmission

A fundamental limit on essentially all antennas that radiate electromagnetic waves is that, for maximum efficiency, their physical size must be a significant fraction of the wavelength of the radiated energy [11]. At ELF, this is a major practical hurdle because of the tremendously long wavelengths (3000 km at 100 Hz) of the radiated waves. Any practical ELF transmission system will be very small relative to the radiated wavelength. Very large and very carefully located systems are required to radiate ELF energy with sufficient power for communication purposes, and even then the antenna will still be very inefficient.

Multikilometer vertical antennas are not currently practical. Thus a long horizontal wire with buried ends forms the aboveground portion of existing ELF antennas. The aboveground wire is effectively one half of a magnetic loop antenna, with the other half formed by the currents closing below ground through the earth, as shown in Fig. 1. The equivalent depth d_{eq} of the closing current with frequency ω over a homogeneous ground of conductivity σ_g is [12]

$$d_{\mathrm{eq}} = (\omega\sigma_g\mu_0)^{-1/2} = 2^{-1/2}\delta \tag{1}$$

where δ is the skin depth [13] of the fields in the conducting ground. A grounded horizontal wire antenna of length l is thus equivalent to an electrically small magnetic loop of area $2^{-1/2}\delta l$. This effective antenna area, which is linearly related to the radiated field strength, is inversely proportional to the ground conductivity. This implies that a poorly conducting ground forces a deeper closing current and therefore increases the efficiency of the antenna. This general effect is opposite that for vertically oriented

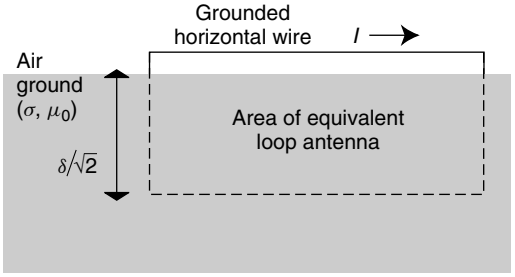


Figure 1. Typical implementation of an ELF antenna. By grounding the ends of a long, current-carrying horizontal wire, the current is forced to close through the ground at a significant depth if the ground is poorly conducting. This forms a physically large loop antenna. The U.S. Navy ELF transmitter antenna is many tens of kilometers long above ground and has an effective current closure depth of 2.6 km.

antennas, in which a good conducting ground serves to improve the radiation efficiency. Horizontal ELF antennas thus require placement over as poorly conducting ground as possible for maximum efficiency.

The U.S. Navy ELF transmitter, which transmits at frequencies between 70 and 80 Hz and is described in some detail by Friedman [2], is distributed in multiple wire antennas in Wisconsin and Michigan. This is one of only a few locations in the United States with relatively low ground conductivity; the measured conductivity is $\sim 2.4 \times 10^4$ S/m (siemens per meter) [14], giving an effective loop depth of 2.6 km. Individual horizontal wires in the antenna system range from 45 to 90 km in length. Despite their physical length, the antennas are still very short compared to the radiated wavelength, and their radiation resistance is very low. The antennas in the ELF system are driven with nearly 1 MW of total power to force 200–300 A of current through them, but the total radiated power from the two-site system is between only 2 and 8 W [15]. Such a small radiated power is still sufficient to cover the globe with a receivable ELF signal at submarine depths.

A similar 82-Hz Russian ELF transmitter became operational in the early 1990s [16]. This antenna consists of multiple 60-km wires on the Kola Peninsula in northwestern Russia. This system radiates slightly more power than does the U.S. version because of lower ground conductivity in its location.

Interestingly, lightning is a stronger radiator of ELF electromagnetic waves than are controlled artificial (human-made) sources. Cloud-to-ground lightning return strokes have typical vertical channel lengths of 5–10 km and contain current pulses that last for hundreds of microseconds with peak currents of tens of kiloamperes. An average lightning stroke radiates a peak power of 10 GW in electromagnetic waves, approximately 1% of which is spread throughout the ELF band [17]. This peak power lasts only for the duration of a lightning stroke, on the order of one millisecond. But the sum of the electromagnetic fields generated by lightning discharges over the entire globe creates a significant ELF noise background [18] that must be overcome in communications applications.

2.2. Propagation

ELF electromagnetic waves propagate along the earth's surface in a manner significantly different from waves in an unbounded medium. The main difference is that waves are bounded above and below by very efficient reflectors of electromagnetic waves. The lower boundary is either earth or water, both of which are very good electrical conductors at ELF. The atmosphere is also electrically conducting, very poorly at low altitudes but with a conductivity that increases exponentially up to ~100 km altitude, above which it continues to increase in a more complicated manner. The region above ~60 km, where the atmosphere is significantly conducting, is called the *ionosphere*. In general, the effect of electrical conductivity on electromagnetic waves is frequency-dependent; the lower the frequency, the greater the effect. The ionospheric conductivity affects ELF waves significantly above ~50–70 km, depending on the precise frequency. Higher-frequency waves can penetrate much higher, but ELF waves are strongly reflected at approximately this altitude. ELF waves are thus confined between the ground and the ionosphere, which are separated by a distance on the order of or much smaller than an ELF wavelength. This spherical shell waveguide, a section of which is shown in Fig. 2, is commonly referred to as the *earth–ionosphere waveguide*.

Because ELF electromagnetic waves are almost completely confined to this small (compared to a wavelength) region, their energy attenuates with distance from the transmitter more slowly than do higher-frequency waves. As a result of two-dimensional energy spreading, the electromagnetic fields decay with distance as $r^{-1/2}$ (producing a r^{-1} power decay), with a slight modification for long propagation distances over the spherical earth. Experimental measurements have shown that the additional attenuation with distance of ELF radiowaves due to losses or incomplete reflection in the ground or ionosphere is typically only 1 dB per 1000 km at 70–80 Hz [4–6]. In general, this attenuation through losses increases with increasing frequency. Because of their low attenuation, ELF waves can be received worldwide from a single transmitter. It is this global reach that makes ELF waves so useful in a variety of applications discussed below.

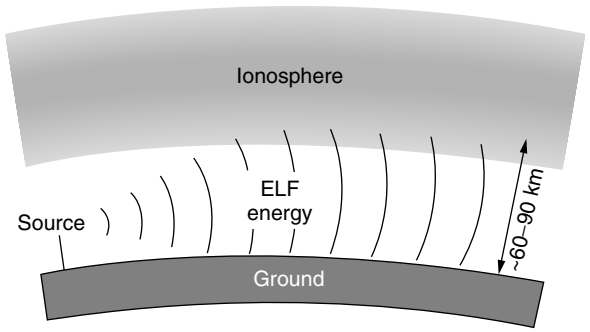


Figure 2. Qualitative ELF electromagnetic wave propagation. The ground and ionosphere (above ~60 km) are good electrical conductors that reflect ELF energy, forming the earth–ionosphere waveguide.

The low attenuation of ELF waves propagating on the surface of a sphere also produces a number of interesting propagation effects. As an ELF receiver approaches the point on the earth directly opposite the transmitter (the antipode), the signals arriving from all angles are comparable in amplitude, and therefore all contribute to the electromagnetic fields. This leads to so-called antipodal focusing, which has been theoretically investigated [12,19,20] but was experimentally verified only in 1998 [16].

At frequencies low enough that the attenuation of a fully around-the-world signal is not too severe, the multi-round-trip signals self-interfere, producing cavity resonances of the earth–ionosphere shell. These resonances are commonly called the *Schumann resonances*, after W. O. Schumann who first predicted them theoretically [21]. They can often be observed as peaks in the broadband background ELF fields in an electromagnetically quiet location and are generated primarily by steady lightning activity around the globe [22]. The frequencies of the first three resonances are approximately 8, 14, and 20 Hz [22], and in a low-noise location they can be observed up to at least 7 orders [16].

Wave propagation in the earth–ionosphere waveguide, like that in a simple parallel-plate waveguide, can be very compactly described mathematically by a sum of discrete waveguide modes that propagate independently within the boundaries. The mathematical details of the mode theory of waveguide propagation are discussed in later sections. An important consequence of this theory is that frequencies with a wavelength greater than half the waveguide height can propagate in only a single mode. For the earth–ionosphere waveguide, propagation is single mode at frequencies less than approximately 1.5–2.0 kHz, depending on the specific ionospheric conditions. This suggests a propagation-based definition of ELF, which is sometimes used in the scientific literature, as the frequencies that propagate with only a single mode in the earth–ionosphere waveguide (i.e., $f \lesssim 2$ kHz) and those for which multiple paths around the earth are not important except near the antipode (i.e., $f \gtrsim 50$ Hz, above Schumann resonance frequencies).

2.3. Reception

Receiving or detecting ELF electromagnetic waves is substantially simpler than transmitting them. Because of the long wavelength at ELF, any practical receiving antenna will be electrically small and the fields will be spatially uniform over the receiving antenna aperture. An ELF receiving antenna is thus usually made from either straight wires or loops as long or as large as possible. Both of these configurations are used as ELF receiving antennas in communication and scientific applications [12].

Again, because of the long wavelength and low transmitted field strength, the maximum practical length possible for an electric wire antenna and the maximum area and number of turns for a magnetic loop antenna are normally preferred to receive the maximum possible signal. How to deploy an antenna as big as possible is often an engineering challenge, as demonstrated by the issues involved in towing a long wire behind a moving submarine to act as an ELF antenna [23].

Designing an antenna preamplifier for an ELF system is also not trivial, especially if precise signal amplitude calibration is needed. Because the output reactance of an electrically short wire antenna is very large, the preamplifier in such a system must have a very high input impedance [12]. And even though the output impedance of a small loop antenna is very small, system noise issues usually force a design involving a stepup voltage transformer and a low-input-impedance preamplifier [24]. Another practical difficulty with ELF receivers is that in many locations, electromagnetic fields from power lines (60 Hz and harmonics in the United States, 50 Hz and harmonics in Europe and Japan) are much stronger than the desired signal. Narrowband ELF receivers must carefully filter this noise, while ELF broadband receivers, usually used for scientific purposes, need to be located far from civilization to minimize the power line noise.

The frequencies of ELF and VLF electromagnetic waves happen to overlap with the frequencies of audible acoustic waves. By simply connecting the output of an ELF sensor directly to a loudspeaker, one can "listen" to the ELF and VLF electromagnetic environment. Besides single-frequency signals from ELF transmitters and short, broadband pulses radiated by lightning, one might hear more exotic ELF emissions such as whistlers [25], which are discussed below, and chorus [26].

3. APPLICATIONS OF ELF WAVES

The unique properties of ELF electromagnetic waves propagating between the earth and the ionosphere, specifically their low attenuation with distance and their relatively deep penetration into conducting materials, make them valuable in a variety of communication and scientific applications discussed below.

3.1. Submarine Communication

The primary task of existing ELF communications systems is communication with distant submerged submarines. The ability to send information to a very distant submarine (thousands of kilometers away) without requiring it to surface (and therefore disclose its position) or even rise to a depth where its wake may be detectable on the surface is of substantial military importance. The chief difficulty in this, however, is in transmitting electromagnetic waves into highly conducting seawater. The amplitude of electromagnetic waves propagating in an electrically conducting material decays exponentially with distance. The distance over which waves attenuate by a factor of e^{-1} is commonly referred to as the "skin depth." This attenuation makes signal penetration beyond a few skin depths into any material essentially impossible. For an electromagnetic wave with angular frequency ω propagating in a material with conductivity σ and permittivity ε, if $\sigma/\omega\varepsilon \gg 1$, then the skin depth δ is very closely approximated by

$$\delta = \left(\frac{2}{\omega\sigma\mu_0}\right)^{-1/2} \tag{2}$$

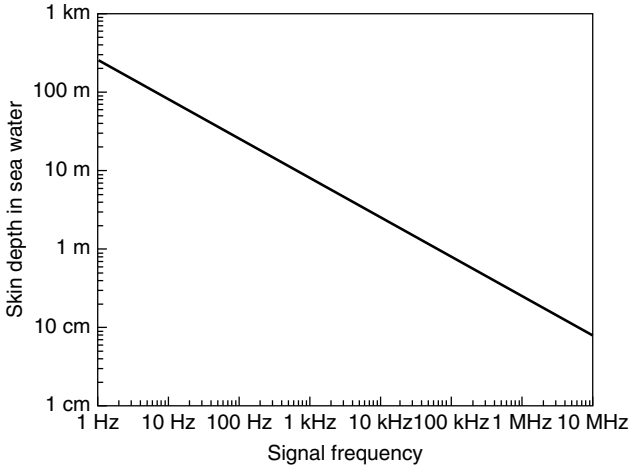


Figure 3. Skin depth (e^{-1} decay depth) of an electromagnetic wave in seawater versus frequency. Only ELF wave frequencies less than ~1 kHz can penetrate sufficiently into seawater to enable communication with submerged submarines.

This expression is usually valid in even weakly conductive materials at ELF because of the low frequency. Figure 3 shows the skin depth in seawater ($\sigma = 4$ S/m, $\varepsilon = 81\varepsilon_0$) as a function of frequency. Only ELF signals below approximately 1 kHz penetrate deeply enough to be practically useful for sending signals to submarines at a depth where they are not detectable from the surface. Because of the very low rate of data transmission on ELF (only a few bits per minute on existing systems [2]), ELF signals from transmitters are used primarily as "bell ringers" to notify the submarine to rise in order to receive more detailed information on a higher frequency. The second signal is often transmitted at ~20 kHz, a frequency low enough to/enable the submarine to remain submerged to receive the signal, using one of a number of military VLF transmitters. Submarine communication at VLF and ELF is only one-way because a transmitting system at these low frequencies is too large to install on a submarine. If higher-rate or two-way communication is needed, the submarine must extend an antenna above water to receive or transmit a signal. Because this can render the submarine detectable by radar, one-way ELF and VLF communication is often the preferred means of communication.

The same low attenuation through conducting materials that makes ELF useful for submarine communications makes it useful for communication in underground mines, where it has been used in some applications [27].

3.2. Global Communication and Navigation

Before satellites were available to relay HF and higher-frequency radio signals long distances over the horizon, global wireless communications and navigation systems depended on very long-distance coverage from a small number of transmitters. ELF and VLF signals provide a very straightforward way to do this. Besides communicating with submarines, VLF transmitters provide a one-way communication link with other military operations. These global VLF links provide a robust and jam-resistant alternative to much higher-data-rate, satellite-based communications. In the event of a global catastrophe, ordinary wireless long-distance communication channels may not be available; satellites may be disabled, and the ionosphere may be so strongly perturbed that HF signals are too strongly attenuated to be useful for long-distance communication. Even under these extreme conditions, ELF and VLF signals are still expected to propagate with low attenuation.

A VLF-based navigation system, called Omega [28], operated from 1955 until 1997. This system of eight transmitters distributed around the world and operating between 10 and 14 kHz provided global positioning accuracy to approximately 2 km. The receiver location was derived from phase differences between signals received simultaneously from multiple Omega transmitters. The satellite-based Global Positioning System (GPS) is a substantially more accurate system, leading to the decommission of the Omega system.

3.3. Geophysical Exploration

The ability of ELF electromagnetic waves to penetrate conducting materials such as rock enables their use in geophysical exploration. A standard subsurface geophysical exploration tool called *magnetotellurics* [8] relies on the fact that ELF and VLF fields on the surface of the ground are influenced by the subsurface properties because of this deep penetration. The specific source of the observed ELF and VLF fields does not matter, provided it is sufficiently far away. Both natural and artificial wave sources are commonly used. Whatever the source, the ratio of perpendicular horizontal components of the electric and magnetic field is a specific function of the subsurface electrical properties [8]. Measurements of this ratio as a function of frequency and position can be inverted into a subsurface geological map, often revealing key subsurface features such as hydrocarbon and ore-bearing deposits [29].

3.4. Scientific Applications

ELF electromagnetic waves are also used in a variety of scientific applications. Natural ELF and VLF emissions from lightning can be used very effectively for remotely sensing the ionosphere [30–32]. The lower ionosphere (~60–150 km) is one of the most inaccessible regions of the atmosphere; it is too low for satellites and too high for airplanes or balloons to probe directly, and higher-frequency incoherent scatter radar [33] used to study the higher-altitude ionosphere seldom returns a useful signal from the lower regions [34]. Because ELF and VLF waves are strongly reflected by the lower ionosphere (~60–150 km) as they propagate, they can very effectively be used to probe the lower ionosphere. And lightning is a very convenient broadband and high power source for such remote sensing.

The ELF radiation from lightning also provides a means for remotely sensing the characteristics of the source lightning discharge itself. Because this energy travels so far with minimal attenuation, a single ELF magnetic or electric field sensor can detect strong lightning discharges many thousands of kilometers away. By modeling the

propagation of the ELF signal (techniques for this are described in the later sections of this article), important parameters describing the current and charge transfer in the lightning can be quantitatively measured [10]. Applications of this technique for lightning remote sensing have led to discoveries about the strength of certain lightning processes [35] and have helped us understand the kind of lightning responsible for a variety of the most recently discovered effects in the mesosphere from strong lightning [36].

A specific kind of natural ELF–VLF electromagnetic emission led to the discovery in the 1950s that near-earth space is not empty but rather filled with ionized gas, or plasma. A portion of the ELF and VLF electromagnetic wave energy launched by lightning discharges escapes the ionosphere and, under certain conditions, propagates along magnetic field lines from one hemisphere of earth to the other. These signals are called "whistlers" because when the signal is played on a loudspeaker, the sound is a frequency-descending, whistling tone that lasts around a second. Thorough reviews of the whistler phenomenon can be found in Refs. 25 and 37. L. Storey, in groundbreaking research, identified lightning as the source of whistlers and realized that the slow decrease in frequency with time in the signal could be explained if the ELF–VLF waves propagated over a long, high-altitude path through an ionized medium [38]. The presence of plasma in most of near-earth space was later confirmed with direct measurements from satellites. The study of natural ELF waves remains an area of active research, including emissions observable on the ground at high latitudes [26,39] and emissions observed directly in space on satellites [e.g., 40] and rockets [e.g., 41].

An interesting modern ELF-related research area is ionospheric heating. A high-power HF electromagnetic wave launched upward into the ionosphere can nonlinearly interact with the medium and modify the electrical characteristics of the ionosphere [42,43]. By modulating the HF wave at a frequency in the ELF band, and by doing so at high latitudes where strong ionospheric electric currents routinely flow [44], these ionospheric currents can be modulated, forming a high-altitude ELF antenna [45]. The ELF signals launched by this novel antenna could be used for any of the applications discussed above. Currently, a major research project, the High-frequency Auroral Active Research Program (HAARP), is devoted to developing such a system in Alaska [46].

4. MATHEMATICAL MODELING OF ELF PROPAGATION

Accurate mathematical modeling of ELF propagation is needed for many of the applications described in the previous sections. It is also complicated because of the influence of the inhomogeneous and anisotropic ionosphere on the propagation. Some of the best-known researchers in electromagnetic theory (most notably J. Wait and K. Budden) worked on and solved the problem of ELF and VLF propagation in the earth–ionosphere waveguide. This work is described in many scientific articles [e.g., 47–51] and technical reports [e.g., 52,53], and is also conveniently and thoroughly summarized in a few books [19,20,54]. The approaches taken by these researchers are fundamentally similar but treat the ionospheric boundary in different ways. A solution for a simplified ELF propagation problem that provides significant insight into subionospheric ELF propagation is summarized below, followed by a treatment with minimal approximations that is capable of more accurate propagation predictions.

4.1. Simplified Waveguide Boundaries

To understand the basic principles of ELF propagation, we first consider a simplified version of the problem. Consider a vertically oriented time-harmonic short electric dipole source at a height $z = z_s$ between two flat, perfectly conducting plates separated by a distance h, as shown in Fig. 4. This simple approximation gives substantial insight into the ELF propagation problem, which is harder to see in more exact treatments of the problem. This simplified problem is, in fact, a reasonably accurate representation of the true problem because at ELF the ionosphere and ground are very good electrical conductors. The main factors neglected in this treatment are the curvature of the earth and the complicated reflection from and losses generated in a realistic ionosphere.

Wait [19] solved this problem analytically and showed that the vertical electric field produced by this short electric dipole source, as a function of height z and distance r from the source, is given by

$$E_z(r,z) = \frac{\mu_0 \omega I\, dl}{4h} \sum_{n=0}^{\infty} \delta_n S_n^2 H_0^{(2)}(kS_n r) \times \cos(kC_n Z_0) \cos(kC_n z) \tag{3}$$

where $\omega = 2\pi \times$ source frequency
$k = \omega/c$
$Idl =$ source current $\times$ dipole length
$\delta_0 = \frac{1}{2}$, $\delta_n = 1$, $n \geq 1$
$C_n = n\lambda/2h =$ cosine of the eigenangle θ_n of the nth waveguide mode
$S_n = (1 - C_n^2)^{1/2} =$ sine of the eigenangle θ_n of the nth waveguide mode
$H_0^{(2)} =$ Hankel function of zero order and second kind

If $|kS_n/r| \gg 1$, the Hankel function can be replaced by its asymptotic expansion $H_0^{(2)}(kS_n r) =$

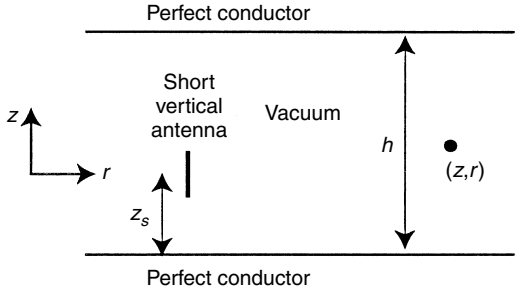


Figure 4. A simplified version of the earth–ionosphere waveguide propagation problem. A slab of free space containing a vertical antenna is bounded above and below by perfectly conducting planes.

$(2/(\pi k S_n r))^{1/2} \exp[-i(kS_n r - \pi/4)]$. This substitution, often referred to as the *far-field approximation*, is generally valid for distances more than one wavelength from the source. In this form, it is easy to see that the electric field decays with distance as $r^{-1/2}$, as it should because the propagation is confined by the waveguide to two dimensions.

This solution shows that the fields inside the perfectly conducting waveguide are the sum of the fields in an infinite number of independently propagating waveguide modes, each of which corresponds to a specific plane-wave angle θ_n of incidence on the waveguide boundaries. Each mode propagates with a different phase velocity $v_p = c/\sin(\theta_n)$, and the fields at a distance are simply the sum of the fields in each mode. This concept of waveguide modes is very general for wave propagation in bounded structures [54] and applies to light propagation in optical fibers [55], acoustic propagation in the ocean [56], and many other scenarios. A physical interpretation of the eigenangle of a waveguide mode is that the modal eigenangles are the set of incidence angles on the boundaries for which a plane wave reflected once each from the upper and lower boundaries is in phase with the incident (nonreflected) plane wave.

The summation in Eq. (3) is over an infinite number of modes. However, the equation for C_n shows that for any frequency there is a mode order $n_{\max}$ above which C_n is greater than unity. For modes of order greater than $n_{\max}$, S_n is purely imaginary, and the mode fields exponentially decay, rather than propagate, with increasing distance from the source. Such modes are called *evanescent* and, because of their exponential decay, do not contribute significantly to the total field except very close to the source. Equivalently, for a fixed mode of order n, there is a frequency below which the mode is evanescent. This frequency is called the *cutoff frequency* of the mode, which in this case is given by $f_{cn} = nc/2h$. Because at any frequency the number of propagating waves is finite, the summation can be terminated at $n_{\max}$ to a very good approximation to compute the fields beyond a significant fraction of a wavelength from the source.

The number of propagating modes $n_{\max}$ is a function of frequency. As frequency decreases, so does the number of modes required to describe the propagation of the energy. This is demonstrated in Fig. 5 with a calculation using Eq. (3). We assume a waveguide upper boundary height of 80 km, representative of the earth–ionosphere waveguide at night, and the source and receiver are on the lower boundary of the waveguide ($z = 0$). The two curves in the figure show the vertical electric field as a function of distance from a 1-A/m vertical electric dipole source at 200 Hz and 10 kHz. The difference between them is obvious and has a clear physical interpretation. At $f = 200$ Hz, only one waveguide mode propagates; all others are evanescent. This one mode is called the *transverse electromagnetic* (TEM) mode because it contains only E_z and H_ϕ fields that point transverse to the propagation direction. Thus at 200 Hz, the field decays with distance smoothly as $r^{-1/2}$ from this single mode. This represents classic, single-mode ELF propagation. But at $f = 10$ kHz, six modes propagate, each with a different phase velocity. The relative phase of these modes changes rapidly with distance, which produces a very complicated field variation with distance, as shown. The six modes interfere constructively where the amplitude is a local maximum, and interfere destructively where the amplitude is a local minimum. This complicated field pattern with distance is very typical of multimode propagation, which occurs at VLF in the earth–ionosphere waveguide. The overall amplitude difference between the two curves results from the fixed antenna length used in this calculation, which is electrically shorter (and therefore less efficient) at lower frequencies. Interesting practical issues arise in multimode propagation. The presence of deep nulls means that at certain locations and under certain waveguide conditions, the received signal level can be very low even for high transmitted power. Predicting signal coverage from military VLF transmitters and finding the locations of these nulls was one of the original motivations behind modeling the ELF–VLF propagation problem as accurately as possible.

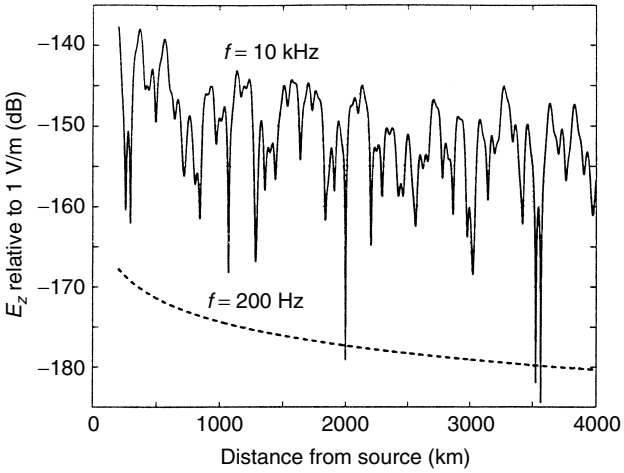


Figure 5. Vertical electric field strength versus distance from a 1-A/m electric dipole source, assuming perfectly conducting earth–ionosphere waveguide boundaries separated by 80 km, and calculated using Eq. (3). At $f = 200$ Hz, only one mode propagates, and the total field decays simply as $r^{-1/2}$. At $f = 10$ kHz, 6 modes propagate with different phase velocities. Even though the fields in each mode decay as $r^{-1/2}$, the mutual interference of all 6 modes creates the very complicated amplitude variation with distance shown here.

Equation (3) is valid for a source at any frequency. However, for frequencies high enough that the wavelength is much smaller than h, the number of propagating modes $n_{\max}$ is so great that the mode-based formulation becomes difficult to use in calculations. Ray theory [57], in which the fields at a distance are described by a sum of discrete rays that each undergo a different number of reflections from the ground and ionosphere, becomes a much more practical and compact representation of the fields at a distance. For propagation in the earth–ionosphere waveguide, this mode theory/ray theory boundary occurs around 50 kHz. At VLF and ELF frequencies, the mode theory description of the fields can very efficiently describe the propagation of the wave energy because only a few modes are propagating (i.e., are not evanescent).

There are other electromagnetic field components produced by this vertical dipole source. In the case of perfectly conducting boundaries, H_ϕ and E_r are nonzero, and expressions for these components can be found directly from the solution for E_z [19]. In general, all the field components in a single mode vary sinusoidally with altitude; thus the altitude variation of the fields is also quite complicated in the case of multimode propagation. The fields produced by a horizontal electric dipole, such as the ELF antenna described above for submarine signaling, are a similar modal series containing different field components [19].

4.2. Realistic Waveguide Boundaries

While the perfectly conducting boundary approximation demonstrates the fundamental characteristics of ELF propagation, it is rarely accurate enough for quantitatively correct simulations of earth–ionosphere waveguide propagation. The real ionosphere is not a sharp interface like that assumed above, but is a smoothly varying waveguide boundary. Nor is it a simple electrical conductor as assumed; it is a magnetized cold plasma with complicated electromagnetic properties. The key ionospheric parameters for electromagnetic wave propagation and reflection are the concentration of free electrons and ions (electron and ion density), the rate of collisions for electrons and ions with other atmospheric molecules (collision frequencies), and the strength and direction of Earth's magnetic field. Figure 6 shows representative altitude profiles of electron and ion density for midday and midnight and profiles of electron and ion collision frequencies. The index of refraction of a cold plasma like the ionosphere is a complicated function of all of these parameters [58], in which is therefore difficult to handle in analytic calculations. Also, in general, the ground is not a perfect conductor and may even be composed of layers with different electromagnetic properties, such as an ice sheet on top of ordinary ground. These realistic boundaries cannot be treated as homogeneous and sharp.

Both Budden [49] and Wait [19] derived solutions for the earth–ionosphere waveguide problem with arbitrary boundaries. Their solutions are fundamentally similar, with slight differences in the derivation and how certain complexities, such as the curved earth, are treated. We follow Wait's treatment below.

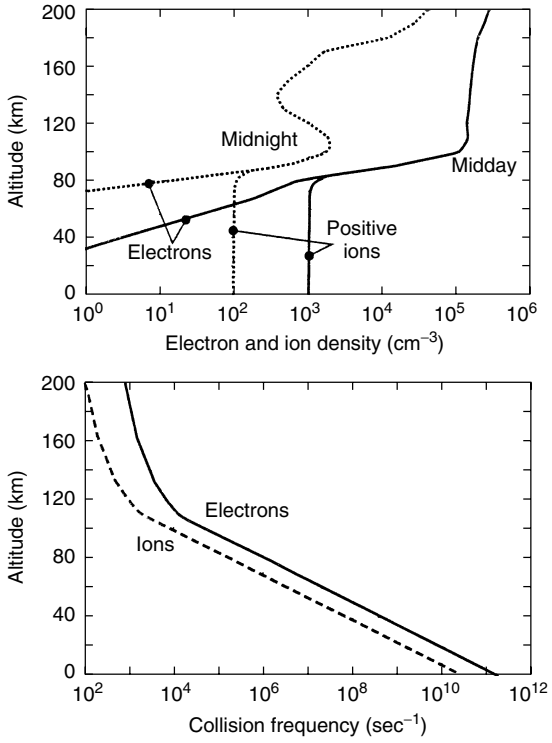


Figure 6. Representative midday and midnight ionospheric electron density, ion density, and collision frequency profiles. Negative ions are also present where needed to maintain charge neutrality.

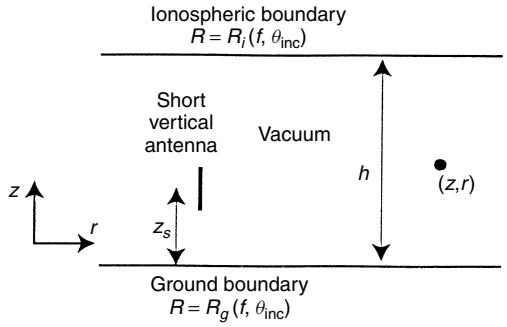


Figure 7. A generalized version of the earth–ionosphere waveguide propagation problem. A slab of free space containing a vertical antenna is bounded above and below by completely arbitrary reflecting boundaries.

4.2.1. Propagation Modeling with General Boundaries. We again consider propagation inside a free-space region between two boundaries separated by a distance h. Now, however, the boundaries at $z = 0$ and $z = h$ are completely general and are described only by their plane-wave reflection coefficients, as shown in Fig. 7. These reflection coefficients are functions of incidence angle, incident polarization, and frequency. This important generalization makes the following formulation applicable to essentially any two-dimensional planar waveguide, regardless of size or bounding material.

Ordinarily, the electromagnetic fields can be separated into transverse electric (TE) and transverse magnetic (TM) groups that propagate independently in a two-dimensional waveguide. However, since the ionosphere is a magnetized plasma and thus is anisotropic, these field groups are coupled at the upper boundary and an incident TE or TM wave produces both TE and TM reflections. Purely TE or TM fields do not exist in the earth–ionosphere waveguide. The coupling between the fields means that the reflection coefficient from the upper boundary is not a scalar quantity but is rather a 2×2 matrix, where each matrix element is one of the four different reflection coefficients for a specific incident and reflected polarization. The lower boundary of the earth–ionosphere waveguide is the ground, which is generally isotropic so that the cross-polarized reflection coefficients in the ground reflection matrix are zero.

With this in mind, we define $\mathbf{R}_I$, the reflection matrix of the ionosphere at altitude $z = h$, and $\mathbf{R}_G$, the reflection matrix of the ground at $z = 0$, as

$$\mathbf{R}_I(\theta) = \begin{bmatrix} {}_{\parallel}R^i_{\parallel} & {}_{\parallel}R^i_{\perp} \\ {}_{\perp}R^i_{\parallel} & {}_{\perp}R^i_{\perp} \end{bmatrix} \mathbf{R}_G(\theta) = \begin{bmatrix} {}_{\parallel}R^g_{\parallel} & 0 \\ 0 & {}_{\perp}R^g_{\perp} \end{bmatrix} \tag{4}$$

These reflection coefficients are implicitly functions of the angle of incidence and the frequency. The left subscript on the matrix elements denotes the incident wave polarization (parallel or perpendicular to the plane of incidence containing the wavevector and the boundary normal), and the right subscript denotes the reflected polarization.

By using the plane-wave spectrum representation of the fields from a short vertical electric dipole, by postulating a particular solution form in the presence of the waveguide, and by enforcing continuity of the tangential field components between the free-space and boundary regions, Wait [19] shows that the fields in the waveguide, in terms of the electric and magnetic Hertz vectors $\mathbf{U}$ and $\mathbf{V}$ are given by the complex contour integral

$$\begin{bmatrix} U_z \\ V_z \end{bmatrix} = -\frac{kIdl}{8\pi\omega\varepsilon_0} \int_\Gamma \mathbf{F}(C) \begin{bmatrix} 1 \\ 0 \end{bmatrix} H_0^{(2)}(kSr)dC \tag{5}$$

with

$$\mathbf{F}(C) = \frac{\begin{matrix}(\exp ikCz + \mathbf{R}_G(C)\exp -ikCz) \\ (\exp ikCh + \mathbf{R}_I(C)\exp -ikCh)\end{matrix}}{\begin{matrix}\exp ikCh(1 - \mathbf{R}_G(C) \\ \mathbf{R}_I(C)\exp -2ikCh)\end{matrix}}, \tag{6}$$

where C and S are the cosine and sine of the complex angle of incidence θ of the wave on the upper and lower boundaries, respectively, as was the case for the simplified boundaries discussed above. The symbols in Eqs. (5) and (6) are consistent with those in Eq. (3). The integrand contains poles where

$$\det(1 - \mathbf{R}_G(C)\mathbf{R}_I(C)\exp -2ikCh) = 0 \tag{7}$$

and thus the integral can be evaluated as a residue series [59]. Equation (7), commonly referred to as the *mode condition*, requires that one eigenvalue of the net reflection coefficient $\mathbf{R}_G\mathbf{R}_I \exp -2ikCh$ be unity. This is equivalent to stating that the plane wave at the given incidence angle reflected once each from the upper and lower boundaries must be in phase with and equal in amplitude to the incident plane wave. In this way, the fields in the waveguide can be thought of as the sum of contributions from the angular spectrum of plane waves at angles for which propagation in the waveguide is self-reinforcing. Each angle of incidence θ_n that satisfies the mode condition is referred to as an *eigenangle* and defines a waveguide mode at the frequency ω under consideration. Budden [49] solved the same problem in a slightly different way by summing the fields produced by an infinite number of sources, each corresponding to a different multiply reflected plane wave in the waveguide. Wait's and Budden's solutions are essentially identical.

From Eqs. (5)–(7), an explicit expression for the Hertz vectors in the free-space region between the two boundaries is given by

$$\begin{bmatrix} U_z \\ V_z \end{bmatrix} = \frac{ikIdl}{4\omega\varepsilon_0} \sum_n \frac{\exp 2ikC_nh}{\partial\Delta/\partial C\Big|_{\theta=\theta_n}} \times \begin{bmatrix} (\exp 2ikC_nh - {}_{\perp}R^g_{\perp\perp}R^i_{\perp})f_p^1(z) \\ i_{\parallel}R^g_{\parallel\parallel}R^i_{\perp}f_p^2(z) \end{bmatrix} H_0^2(kS_nr) \tag{8}$$

where $\Delta(C) = \det(\exp 2ikCh - \mathbf{R}_G\mathbf{R}_I)$. The actual electric and magnetic fields are easily derived from these Hertz vector components [19].

Each term in (8) has a physical interpretation. The leading constant is a source term that depends on the current–moment Idl of the vertical dipole source. The first term in the summation is commonly referred to as the *excitation function* for a particular mode at a given frequency, and it quantifies the efficiency with which that mode is excited by a vertical dipole on the ground. The 2×1 matrix in the summation describes the field variation with altitude, and the functions f_p^1 and f_p^2 are defined explicitly by Wait [19]. The $H_0^2(kS_nr)$ term describes the propagation of a cylindrically expanding wave, which exists because the expansion in the vertical direction is limited by the waveguide boundaries so that the mode fields spread only horizontally. For distances where $kS_nr \gtrsim 1$, we can approximate $H_0^2(kS_nr) \approx \left(\frac{2}{\pi kS_nr}\right)^{1/2} \exp[-i(kS_nr - \pi/4)]$, which more explicitly shows the $r^{-1/2}$ cylindrical spreading. In this approximation, it is also clear that each mode propagates as $\exp[-i(kS_nr)]$. The sine of the modal eigenangle thus contains all the information about the phase velocity and attenuation rate of the mode. Because the boundaries are lossy in the earth–ionosphere waveguide (due to both absorption in the boundaries and energy leakage out of the waveguide from imperfect reflection), the eigenangles and thus S_n are necessarily complex.

As was the case for the simplified, perfectly conducting waveguide, the summation in Eq. (8) is over an infinite number of modes. In practice, however, it can be limited only to the modes that contribute significantly to the fields at a distance r from the source. All modes are at least somewhat lossy with realistic waveguide boundaries, but there generally is a sharp transition between low loss and very high loss at the cutoff frequency of each mode. Often for long distances at ELF and VLF, only a few low-attenuation modes or even just one contribute significantly to the fields, leading to a very compact and efficient calculation. For very short distances, however, even highly attenuated modes can contribute to the fields and mode theory becomes less efficient and more difficult to implement.

Two factors have been neglected in this analysis of ELF propagation with realistic boundaries. One is the curvature of the earth. Treating the curvature exactly substantially increases the complexity of the modeling [19], but relatively simple corrections to Eqs. (7) and (8) can account for this curvature fairly accurately.

By introducing an artificial perturbation to the index of refraction of the free-space region between the ground and ionosphere to the flat-earth problem [19,49], the ground curvature can be handled correctly. At lower VLF ($\lesssim$5 kHz) and ELF, the flat-earth mode condition in Eq. (7) is accurate, and this correction is not needed [19]. The curved boundaries also affect the geometric decay of the fields with distance. Simply replacing the $r^{-1/2}$ term in Eq. (8) with the factor $[a \sin(r/a)]^{-1/2}$, where a is the radius of the earth, properly accounts for energy spreading over a spherical boundary [49,19].

Antipodal effects are also neglected in the analysis presented above. When the receiver approaches the antipode of the transmitter, signals that propagate around the earth in directions other than the direct great circle path are comparable in amplitude to the direct signal and thus interfere. As discussed above, this effect is most significant at ELF and has been studied theoretically [12,19,20]. Experimental results from 1998 compare favorably with predictions[16].

4.2.2. Practical Implementation. The general equations describing mode-based ELF and VLF propagation are complicated because of the complexity of the physical boundaries of the earth–ionosphere waveguide. The most difficult part of implementing calculations in Eq. (8) with realistic boundaries is solving Eq. (7) to find the modal eigenangles for a general boundary. When the reflection matrixes cannot be described analytically, such as for an arbitrary ionospheric profile, an iterative numerical procedure is needed to find the set of eigenangles that satisfy the mode condition. Such a numerical model and code was developed over a number of years by the Naval Electronics Laboratory Center (NELC), which later became the Naval Ocean Systems Center (NOSC) and is now the Space and Naval Warfare Systems Center (SSC). The details of this model are described in a series of papers [51,60] and technical reports [53,61,62]. The end result was a model called long wave propagation capability (LWPC) that calculates the eigenangles for an arbitrary ionospheric profile and arbitrary but uniform ground, and then calculates the total field at a given distance from the source. This model includes the modifications to the mode condition [Eq. (7)] and the field equation [Eq. (8)] to account for propagation over a curved earth, and it also includes the capability of accounting for sharp inhomogeneities in the ionosphere along the direction of propagation. The technique used for this calculates mode conversion coefficients at the two sides of the sharp inhomogeneity [63,64].

Using the nighttime ionospheric profiles in Fig. 6 and ground parameters representative of propagation over land ($\varepsilon_r = 15$, $\sigma = 10^{-3}$), we have used the LWPC model to calculate the vertical electric field versus distance from a 1 A m vertical electric dipole source for source frequencies of 200 Hz and 10 kHz. The results, shown in Fig. 8, are directly comparable to the simplified model results in Fig. 5. While qualitatively similar, there are significant differences. Again, the overall amplitude difference between the 200-Hz and 10-kHz signals relates to our use of a fixed antenna length in the simulation. At ELF (200 Hz), there is again only one propagating mode

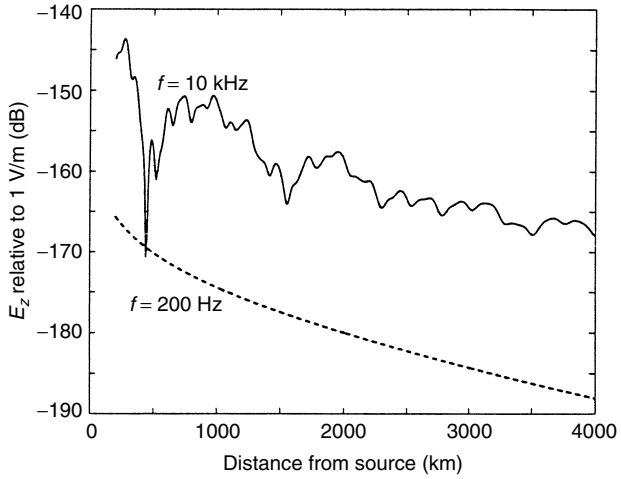


Figure 8. Model ELF and VLF calculations using LWPC with the ionospheric profiles shown in Fig. 6. At $f = 200$ Hz, only one mode propagates, and the total field decays slightly faster than $r^{-1/2}$ because of the lossy waveguide boundaries. At $f = 10$ kHz, more modes propagate and produce a more complicated amplitude variation with distance. Because the boundaries are lossy, however, the total signal is dominated by only a few modes, resulting in a simpler amplitude variation than the lossless waveguide results in Fig. 5.

that gives a smooth amplitude variation with distance. For the realistic ionosphere, however, the field attenuates faster than $r^{-1/2}$ with distance because energy is lost in the imperfectly conducting boundaries as well as through geometric spreading. At VLF (10 kHz), the amplitude varies considerably with distance, which is again a consequence of multimode propagation at the higher frequency. These variations are much less severe than in the calculation with perfectly conducting boundaries. This is also a consequence of the lossy boundaries. Certain modes are lossier than others, and these decay more rapidly with distance, so that fewer propagating modes contribute significantly to the fields at a given distance. If the calculation were extended for longer distances, there would be a point beyond which one mode dominates the fields, and the amplitude variation would become smooth. This does not happen for perfectly conducting boundaries because all modes decay with distance at the same rate. Measurements of VLF signal strength with distance made from aircraft have shown that this model can very accurately predict the variation of signal with distance, provided the correct ionosphere is used in the simulation [65].

4.3. Analytic Approximations

Analytic approximations have been developed for ELF propagation that are more realistic than the overly simplified perfectly conducting parallel-plate waveguide but that are much less complicated than the exact approach. One of the most widely used and most accurate approximations for ELF propagation was developed by Greifinger and Greifinger [66]. They recognized that because ELF wavelengths are so long, only a few key characteristics of the ionospheric altitude profile strongly

influence the characteristics of ELF propagation. These approximate solutions apply only to single-mode, ELF propagation and thus to frequencies less than 1 kHz, and are most accurate at frequencies less than a few hundred hertz.

By making a few key approximations but still accounting for the anisotropy of the ionosphere, Greifinger and Greifinger [66] show that the sine of the eigenangle of the single ELF propagating mode S_0 is approximated by

$$S_0 \approx \left(\frac{h_1(h_1 + i\pi\zeta_1)}{(h_0 - \zeta_0 i\pi/2)(h_1 + \zeta_1 i\pi/2)} \right)^{1/2} \tag{9}$$

The parameter h_0 is the altitude at which the parallel ionospheric conductivity $\sigma_\parallel(h_0) = \omega\varepsilon_0$, and ζ_0 is the parallel conductivity scale height at the altitude h_0. The parameter h_1 is a higher altitude at which $4\mu_0\omega\sigma_H(h_1)\zeta_1^2 = 1$, where ζ_1 is the Hall conductivity scale height at the altitude h_1. The parallel and Hall conductivities are functions of the ionospheric electron and ion densities, collision frequencies, and magnetic field [67] and are straightforward to calculate. With this expression for S_0, the vertical electric field at ground level radiated by a vertical electric dipole is then given by [68]

$$E_z(r, 0) = \frac{\mu_0\omega Idl}{4(h_0 - i\pi\zeta_0/2)} S_0^2 H_0^{(2)}(kS_0 r) \tag{10}$$

Equations (9) and (10) give a simple, noniterative method for calculating the ELF propagation characteristics under a realistic ionosphere. This method was compared with LWPC calculations at 50 and 100 Hz and was found to be very accurate [66]. To demonstrate, we consider 200 Hz ELF propagation under the ionospheric profiles shown in Fig. 6. After calculating the associated parallel and Hall conductivity profiles, we find for this ionosphere that $h_0 = 55$ km, $\zeta_0 = 2.5$ km, $h_1 = 74$ km, and $\zeta_1 = 7.2$ km. Plugging these numbers into Eqs. (9) and (10), the vertical electric field strength versus distance from a 1-A/m source at 200 Hz is as shown in Fig. 9. The approximate method in this section is very close to the full-wave, LWPC calculation, as shown; only the modal attenuation rate is slightly underestimated. The approximate method is also much closer to the full-wave calculations than is the perfectly conducting waveguide approximation, which, because it is lossless, drastically underestimates the modal attenuation. For realistic nighttime ionospheres, there is often complicated attenuation behavior as a function of ELF frequency because of sharp ionospheric gradients between 100 and 150 km [9] that is not be captured by this approximate method. Nevertheless, this method can accurately predict ELF field strength for realistic ionospheres and is much simpler to implement than a more exact, full-wave calculation.

4.4. Finite-Difference Methods

A completely different approach to modeling ELF propagation that has been used applies full-wave finite-difference simulations [69] to model the electromagnetic fields everywhere in the computational domain [32,70]. This brute-force but versatile approach has become usable only relatively recently because of the computing power available on modest platforms. This method is based on decomposing the volume of interest into a finite grid of discrete points at which all electromagnetic field components are computed. As a general rule, the spacing between grid points must be significantly smaller than one wavelength in order to accurately model the fields. This makes finite-difference methods well suited to low-frequency ELF and VLF propagation propagation problems in which the wavelength is quite long and the usual computational domain is a relatively small number of wavelengths in size. A major advantage of this method is that arbitrary lateral inhomogeneities, such as smooth or sharp changes in the ionospheric profile along the propagation direction, can be introduced without any increase in model complexity. Also, because the model complexity depends only on the size of the computational domain, short propagation paths are significantly easier to model than long paths. The opposite is true for mode-theory-based calculations, in which long paths are simpler because fewer modes contribute to the total fields. Mode theory and finite-difference methods are thus rather complementary. Finite-difference methods can easily handle arbitrary inhomogeneities and work especially well over short distances, while mode theory calculations can be very efficient, especially over long distances. Most importantly, the mode theory formulation of the problem also provides essential physical insight into the ELF propagation problem.

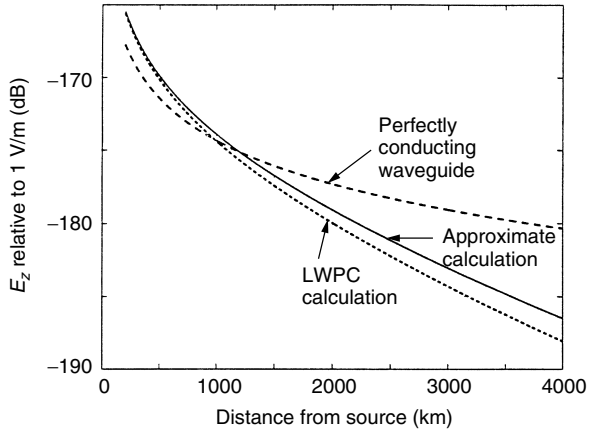


Figure 9. Comparison of 200 Hz ELF field strength with distance for a perfectly conducting waveguide, for full-wave LWPC calculations using a realistic ionosphere, and for the approximate analytical method of Greifinger and Greifinger.

BIOGRAPHY

Steven A. Cummer is an assistant professor of the Department of Electrical and Computer Engineering at Duke University Durham, North Carolina. He received his B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, California, in 1991, 1993, and 1997, respectively. He spent two years at NASA Goddard Space Flight Center in Greenbelt, Maryland, as an National Research Council postdoctoral research associate, and joined Duke University in 1999. He received a National Science Foundation CAREER award

in 2000 and a Presidential Early Career Award for Scientists and Engineers (PECASE) in 2001. His current research is in a variety of problems in ionospheric and space physics, emphasizing electromagnetic modeling and remote sensing using ground-based and satellite instruments.

BIBLIOGRAPHY

1. J. Radatz, ed., *The IEEE Standard Dictionary of Electrical and Electronics Terms*, IEEE, New York, 1997.
2. N. Friedman, *The Naval Institute Guide to World Naval Weapons Systems*, U.S. Naval Institute, Annapolis, MD, 1997.
3. K. Davies, *Ionospheric Radio*, Peter Peregrinus, London, 1990.
4. W. L. Taylor and K. Sao, ELF attenuation rates and phase velocities observed from slow tail components of atmospherics, *Radio Sci.* **5**: 1453–1460 (1970).
5. D. P. White and D. K. Willim, Propagation measurements in the extremely low frequency (ELF) band, *IEEE Trans. Commun.* **22**(4): 457–467 (April 1974).
6. P. R. Bannister, Far-field extremely low frequency (ELF) propagation measurements, *IEEE Trans. Commun.* **22**(4): 468–473 (1974).
7. T. A. Heppenheimer, Signalling subs, *Popular Sci.* **230**(4): 44–48 (1987).
8. K. Vozoff, The magnetotelluric method, in M. Nabighian, ed., *Electromagnetic Methods in Applied Geophysics*, Vol. 2, Society of Exploration Geophysics, Tulsa, OK, 1991, pp. 641–711.
9. S. A. Cummer and U. S. Inan, Ionospheric *E* region remote sensing with ELF radio atmospherics, *Radio Sci.* **35**: 1437 (2000).
10. S. A. Cummer and U. S. Inan, Modeling ELF radio atmospheric propagation and extracting lightning currents from ELF observations, *Radio Sci.* **35**: 385–394 (2000).
11. J. D. Kraus, *Antennas*, McGraw-Hill, New York, 1988.
12. M. L. Burrows, *ELF Communications Antennas*, Peter Peregrinus, Herts, UK, 1978.
13. U. Inan and A. Inan, *Engineering Electromagnetics*, Prentice-Hall, Englewood Cliffs, NJ, 1999.
14. P. R. Bannister, Summary of the Wisconsin test facility effective earth conductivity measurements, *Radio Sci.* **11**: 405–411 (1976).
15. J. C. Kim and E. I. Muehldorf, *Naval Shipboard Communications Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
16. A. C. Fraser-Smith and P. R. Bannister, Reception of ELF signals at antipodal distances, *Radio Sci.* **33**: 83–88 (1998).
17. D. R. MacGorman and W. D. Rust, *The Electrical Nature of Storms*, Oxford Univ. Press, New York, 1998.
18. D. A. Chrissan and A. C. Fraser-Smith, Seasonal variations of globally measured ELF/VLF radio noise, *Radio Sci.* **31**(5): 1141–1152 (Sept. 1996).
19. J. R. Wait, *Electromagnetic Waves in Stratified Media*, Pergamon Press, Oxford, 1970.
20. J. Galejs, *Terrestrial Propagation of Long Electromagnetic Waves*, Pergamon Press, Oxford, 1972.
21. W. O. Schumann, Über die strahlungslosen eigenschingungen einer leitenden kugel, diel von einer luftschicht und einer ionosphärenhülle umgeben ist, *Z. Naturforsch.* **7a**: 149 (1952).
22. D. D. Sentman, Schumann resonances, in H. Volland, ed., *Handbook of Atmospheric Electrodynamics*, Vol. 1, CRC Press, Boca Raton, FL, 1995, pp. 267–295.
23. C. T. Fessenden and D. H. S. Cheng, Development of a trailing-wire E-field submarine antenna for extremely low frequency (ELF) reception, *IEEE Trans. Commun.* **22**: 428–437 (1974).
24. C. D. Motchenbacher and J. A. Connelly, *Low Noise Electronic System Design*, Wiley-Interscience, Englewood Cliffs NJ, 1993.
25. V. S. Sonwalkar, Whistlers, in J. G. Webster, ed., *Wiley Encyclopedia of Electrical and Electronic Engineering*, pages Wiley, New York, 1999, pp. 580–591.
26. S. S. Sazhin and M. Hayakawa, Magnetospheric chorus emissions — a review, *Planet. Space Sci.* **50**: 681–697 (May 1992).
27. J. N. Murphy and H. E. Parkinson, Underground mine communications, *Proc. IEEE* **66**: 26–50 (1978).
28. E. R. Swanson, Omega, *Proc. IEEE* **71**(10): 1140–1155 (1983).
29. G. M. Hoversten, Papua new guinea MT: Looking where seismic is blind, *Geophys. Prospect.* **44**(6): 935–961 (1996).
30. H. G. Hughes, R. J. Gallenberger, and R. A. Pappert, Evaluation of nighttime exponential ionospheric models using VLF atmospherics, *Radio Sci.* **9**: 1109 (1974).
31. S. A. Cummer, U. S. Inan, and T. F. Bell, Ionospheric *D* region remote sensing using VLF radio atmospherics, *Radio Sci.* **33**(6): 1781–1792 (Nov. 1998).
32. S. A. Cummer, Modeling electromagnetic propagation in the earth-ionosphere waveguide, *IEEE Trans. Antennas Propag.* **48**: 1420 (2000).
33. J. V. Evans, Theory and practice of ionosphere study by thomson scatter radar, *Proc. IEEE* **57**: 496 (1969).
34. J. D. Mathews, J. K. Breakall, and S. Ganguly, The measurement of diurnal variations of electron concentration in the 60–100 km, *J. Atmos. Terr. Phys.* **44**: 441 (1982).
35. S. A. Cummer and M. Füllekrug, Unusually intense continuing current in lightning causes delayed mesospheric breakdown, *Geophys. Res. Lett.* **28**: 495 (2001).
36. S. A. Cummer and M. Stanley, Submillisecond resolution lightning currents and sprite development: Observations and implications, *Geophys. Res. Lett.* **26**(20): 3205–3208 (Oct. 1999).
37. R. A. Helliwell, *Whistlers and Related Ionospheric Phenomena*, Stanford Univ. Press, Stanford, CA, 1965.
38. L. R. O. Storey, An investigation of whistling atmospherics, *Phil. Trans. Roy. Soc. London, Ser. A* **246**: 113 (1953).
39. A. J. Smith et al., Periodic and quasiperiodic ELF/VLF emissions observed by an array of antarctic stations, *J. Geophys. Res.* **103**: 23611–23622 (1998).
40. P. Song et al., Properties of ELF emissions in the dayside magnetopause, *J. Geophys. Res.* **103**: 26495–26506 (1998).
41. P. M. Kintner, J. Franz, P. Schuck, and E. Klatt, Interferometric coherency determination of wavelength or what are broadband ELF waves? *J. Geophys. Res.* **105**: 21237–21250 (Sept. 2000).

42. A. V. Gurevich, *Nonlinear Phenomena in the Ionosphere*, Springer-Verlag, Berlin, 1978.

43. K. Papadopoulos, Ionospheric modification by radio waves, in V. Stefan, ed., *Nonlinear and Relativistic Effects in Plasmas*, American Institute of Physics, New York, 1992.

44. A. D. Richmond and J. P. Thayer, Ionospheric electrodynamics: A tutorial, in S. Ohtani, ed., *Magnetospheric Current Systems*, American Geophysical Union, Washington, DC, 2000, pp. 131–146.

45. R. Barr, The generation of ELF and VLF radio waves in the ionosphere using powerful HF transmitters, *Adv. Space Res.* **21**(5): 677–687 (1998).

46. G. M. Milikh, M. J. Freeman, and L. M. Duncan, First estimates of HF-induced modifications of the D-region by the HF active auroral research program facility, *Radio Sci.* **29**(5): 1355–1362 (Sept. 1994).

47. K. G. Budden, The propagation of very-low-frequency radio waves to great distances, *Phil. Mag.* **44**: 504 (1953).

48. J. R. Wait, The mode theory of v.l.f. ionospheric propagation for finite ground conductivity, *Proc. IRE* **45**: 760 (1957).

49. K. G. Budden, The influence of the earth's magnetic field on radio propagation by wave-guide modes, *Proc. Roy. Soc. A* **265**: 538 (1962).

50. J. R. Wait, On the propagation of E.L.F. pulses in the earth-ionosphere waveguide, *Can. J. Phys.* **40**: 1360 (1962).

51. R. A. Pappert and W. F. Moler, Propagation theory and calculations at lower extremely low frequencies (ELF), *IEEE Trans. Commun.* **22**(4): 438–451 (April 1974).

52. J. R. Wait and K. P. Spies, *Characteristics of the Earth-Ionosphere Waveguide for VLF Radio Waves*, Technical Report, NBS Technical Note 300, National Bureau of Standards, 1964.

53. D. G. Morfitt and C. H. Shellman, *MODESRCH, an Improved Computer Program for Obtaining ELF/VLF/LF Mode Constants in an Earth-Ionosphere Waveguide*, Technical Report Interim Report 77T, Naval Electronic Laboratory Center, San Diego, CA, 1976.

54. K. G. Budden, *The Wave-Guide Mode Theory of Wave Propagation*, Logos Press, London, 1961.

55. G. Keiser, *Optical Fiber Communications*, McGraw-Hill, Boston, MA, 2000.

56. F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*, American Institute of Physics, Woodbury, NY, 1994.

57. J. R. Wait and A. Murphy, The geometrical optics of VLF sky wave propagation, *Proc. IRE* **45**: 754 (1957).

58. K. G. Budden, *The Propagation of Radio Waves*, Cambridge Univ. Press, New York, 1985.

59. R. V. Churchill and J. W. Brown, *Complex Variables and Applications*, McGraw-Hill, New York, 1990.

60. R. A. Pappert and J. A. Ferguson, VLF/LF mode conversion model calculations for air to air transmissions in the earth-ionosphere waveguide, *Radio Sci.* **21**: 551–558 (1986).

61. J. A. Ferguson and F. P. Snyder, *Approximate VLF/LF Mode Conversion Model*, Technical Report, Technical Document 400, Naval Ocean Systems Center, San Diego, CA, 1980.

62. J. A. Ferguson, F. P. Snyder, D. G. Morfitt, and C. H. Shellman, *Long-wave Propagation Capability and Documentation*, Technical Report, Technical Document. 400, Naval Ocean Systems Center, San Diego, CA, 1989.

63. J. R. Wait, Mode conversion and refraction effects in the earth-ionosphere waveguide for VLF radio waves, *J. Geophys. Res.* **73**(11): 3537–3548 (1968).

64. R. A. Pappert and D. G. Morfitt, Theoretical and experimental sunrise mode conversion results at VLF, *Radio Sci.* **10**: 537 (1975).

65. J. E. Bickel, J. A. Ferguson, and G. V. Stanley, Experimental observation of magnetic field effects on VLF propagation at night, *Radio Sci.* **5**: 19 (1970).

66. C. Greifinger and P. Greifinger, On the ionospheric parameters which govern high-latitude ELF propagation in the earth-ionosphere waveguide, *Radio Sci.* **14**(5): 889–895 (Sept. 1979).

67. J. D. Huba and H. L. Rowland, Propagation of electromagnetic waves parallel to the magnetic field in the nightside Venus ionosphere, *J. Geophys. Res.* **98**: 5291 (1993).

68. C. Greifinger and P. Greifinger, Noniterative procedure for calculating ELF mode constants in the anisotropic Earth-ionosphere waveguide, *Radio Sci.* **21**(6): 981–990 (1986).

69. A. Taflove and S. C. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, Artech House, Norwood, MA, 2000.

70. M. Thevenot, J. P. Berenger, T. Monediere, and F. Jecko, A FDTD scheme for the computation of VLF-LF propagation in the anisotropic earth-ionosphere waveguide, *Ann. Telecommun.* **54**: 297–310 (1999).

EM ALGORITHM IN TELECOMMUNICATIONS

COSTAS N. GEORGHIADES
Texas A&M University
College Station, Texas

PREDRAG SPASOJEVIĆ
Rutgers, The State University of New Jersey
Piscataway, New Jersey

1. INTRODUCTION

Since its introduction in the late 1970s as a general iterative procedure for producing maximum-likelihood estimates in cases a direct approach is computationally or analytically intractable, the expectation-maximization (EM) algorithm has been used with increasing frequency in a wide variety of application areas. Perhaps not surprisingly, one of the areas that has seen an almost explosive use of the algorithm since the early 1990s is telecommunications. In this article we describe some of the varied uses of the EM algorithm that appeared in the literature in the area of telecommunications since its introduction and include some new results on the algorithm that have not appeared elsewhere in the literature.

The expectation-maximization (EM) algorithm was introduced in its general form by Dempster et al. in 1977 [1]. Previous to that time a number of authors had in fact proposed versions of the EM algorithm for particular applications (see, e.g., Ref. 4), but it was [1] that established the algorithm (in fact, as argued by some, a "procedure", rather than an algorithm) as a general tool for producing maximum-likelihood (ML) estimates in

situations where the observed data can be viewed as "incomplete" in some sense. In addition to introducing the EM algorithm as a general tool for ML estimation, Dempster et al. [1] also dealt with the important issue of convergence, which was further studied and refined by the work of Wu [5].

Following the publication of [1], the research community experienced an almost explosive use of the EM algorithm in a wide variety of applications beyond the statistics area in which it was introduced. Early application areas for the EM algorithm outside its native area of statistics include genetics, image processing, and, in particular, positron emission tomography (PET) [6,7], (an excellent treatment of the application of the EM algorithm to the PET problem can be found in Snyder and Miller [8]). In the late 1980s there followed an increasing number of applications of the EM algorithm in the signal processing area, including in speech recognition, neural networks, and noise cancellation. Admittedly somewhat arbitrarily, we will not cite references here in the signal processing area, with the exception of one that had significant implications in the area of communications: the paper by Feder and Weistein [9]. This paper will be described in somewhat more detail later, and it is important in the area of telecommunications in that it dealt with the problem of processing a received signal that is the superposition of a number of (not individually observable) received signals. Clearly, this framework applies to a large number of telecommunications problems, including multiuser detection and detection in multipath environments. The reader can find other references (not a complete list) on the use of the EM algorithm in the signal processing (and other areas) in the tutorial paper by Moon [10]. For a general introduction to the EM algorithm and its applications in the statistics area, the reader is urged to read the original paper of Dempster et al. [1], as well as a book published in 1997 [11] on the algorithm. The appearance of a book on the EM algorithm is perhaps the best indication of its widespread use and appeal.

As mentioned above, in this manuscript we are interested in a somewhat narrow application area for the EM algorithm, that of telecommunications. Clearly, the boundaries between signal processing and telecommunications are not always well defined and in presenting the material below some seemingly arbitrary decisions were made: Only techniques that deal with some aspect of data transmission are included.

In Section 2 and for completeness we first give a brief introduction of the EM algorithm. In Section 3 we present a summary of some of the main published results on the application of the EM algorithm to telecommunications. Section 4 contains some recent results on the EM algorithm, and Section 5 concludes.

2. THE ESSENTIAL EM ALGORITHM

Let $\mathbf{b} \in \mathcal{B}$ be a set of parameters to be estimated from some observed data $\mathbf{y} \in \mathcal{Y}$. Then the ML estimate $\widehat{\mathbf{b}}$ of $\mathbf{b}$ is a solution to

$$\widehat{\mathbf{b}} = \arg \max_{\mathbf{b} \in \mathcal{B}} g(\mathbf{y} \mid \mathbf{b}) \tag{1}$$

where $g(\mathbf{y} \mid \mathbf{b})$ is the conditional density of the data given the parameter vector to be estimated. In many cases a simple explicit expression for this conditional density does not exist, or is hard to obtain. In other cases such an expression may be available, but it is one that does not lend itself to efficient maximization over the set of parameters. In such situations, the expectation–maximization algorithm may provide a solution, albeit iterative (and possibly numerical), to the ML estimation problem.

The EM-based solution proceeds as follows. Suppose that instead of the data $\mathbf{y}$ actually available one had data $\mathbf{x} \in \mathcal{X}$ from which $\mathbf{y}$ could be obtained through a many-to-one mapping $\mathbf{x} \rightarrow \mathbf{y}(\mathbf{x})$, and such that their knowledge makes the estimation problem easy (for example, the conditional density $f(\mathbf{x} \mid \mathbf{b})$ is easily obtained.) In the EM terminology, the two sets of data $\mathbf{y}$ and $\mathbf{x}$ are referred to as the incomplete and complete data, respectively.

The EM algorithm makes use of the loglikelihood function for the complete data in a two-step iterative procedure that under some conditions converges to the ML estimate given in (1) [1,5]. At each step of the EM iteration, the likelihood function can be shown to be nondecreasing [1,5]; if it is also bounded (which is mostly the case in practice), then the algorithm converges. The two-step procedure at the ith iteration is

1. *E step*: Compute $Q(\mathbf{b} \mid \mathbf{b}^i) \equiv E[\log f(\mathbf{x} \mid \mathbf{b}) \mid \mathbf{y}, \mathbf{b}^i]$.
2. *M step*: Solve $\mathbf{b}^{i+1} = \arg \max_{\mathbf{b} \in \mathcal{B}} Q(\mathbf{b} \mid \mathbf{b}^i)$.

Here $\mathbf{b}^i$ is the parameter vector estimate at the ith iteration. The two steps of the iteration are referred to as the expectation (E step) and maximization (M step) steps, respectively. Note that in the absence of the data $\mathbf{x}$, which makes $\log f(\mathbf{x} \mid \mathbf{b})$ a random variable, the algorithm maximizes its conditional expectation instead, given the incomplete data and the most recent estimate of the parameter vector to be estimated.

As mentioned earlier, the EM algorithm has been shown [1,5] to result in a monotonically nondecreasing loglikelihood. Thus, if the loglikelihood is bounded (which is the case in most practical systems), then the algorithm converges. Under some conditions, the stationary point coincides with the ML estimate.

3. AN OVERVIEW OF APPLICATIONS TO TELECOMMUNICATIONS

In this section we provide a brief overview on the use of the EM algorithm in the telecommunications area.

3.1. Parameter Estimation from Superimposed Signals

One of the earliest uses of the EM algorithm with strong direct implications in communications appeared in 1988 [9]. Feder and Weinstein [9] look at the problem of parameter estimation from a received signal, $y(t)$ (it can be a vector in general), which is a superposition of a number of signals plus Gaussian noise:

$$y(t) = \sum_{k=1}^{K} s_k(t; \theta_k) + n(t) \tag{2}$$

For simplicity in illustrating the basic idea, we will assume that $y(t)$, $n(t)$, and each $\theta_k, k = 1, 2, \ldots, K$ are scalar (Feder and Weinstein [9] handle the more general case of vectors); the objective is to estimate the parameters $\theta_k, k = 1, 2, \ldots, K$ from the data $y(t)$ observed over a T-second interval. It is assumed that the $s_k(t; \theta_k)$ are known signals given the corresponding parameters θ_k, and $n(t)$ is white Gaussian noise with $\sigma^2 = E[|n(t)|^2]$. Clearly, if instead of the superimposed data $y(t)$ one had available the individual components of $y(t)$, the problem of estimating the θ_k would become much simpler since there would be no coupling between the parameters to be estimated. Thus, the complete data $\mathbf{x}(t)$ are chosen to be a decomposition of $y(t)$: $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_K(t)]'$ (where the prime sign $'$ means transpose) where

$$x_k(t) = s_k(t; \theta_k) + n_k(t), \quad k = 1, 2, \ldots, K \tag{3}$$

$$\sum_{k=1}^{K} n_k(t) = n(t) \tag{4}$$

and, thus

$$y(t) = \sum_{k=1}^{K} x_k(t)$$

The $n_k(t)$ are assumed statistically independent (for analytic convenience) and have corresponding variances $\beta_k \sigma^2$, where it is assume that $\beta_k \geq 0$ and

$$\sum_{k=1}^{K} \beta_k = 1$$

Then the E step of the EM algorithm yields [9]

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) = -\sum_{k=1}^{K} \frac{1}{\beta_k} \int_T \left| \hat{x}_k(t) - s_k(t; \theta_k) \right|^2 dt \tag{5}$$

where

$$\hat{x}_k(t) = s_k(t; \hat{\theta}_k) + \beta_k \left[y(t) - \sum_{i=1}^{K} s_i(t; \hat{\theta}_i) \right] \tag{6}$$

and $\hat{\theta}_i, i = 1, 2, \ldots, K$ are the most recent estimates. At the M step, maximization of (5) with respect to the parameter vector $\boldsymbol{\theta}$ corresponds to minimizing each term in the sum in (5) with respect to the corresponding individual parameter. Thus, the desired decoupling.

Feder and Weinstein [9] went on to apply the general results presented above to the problems of estimating the multipath delays and to the problem of multiple source location. No modulation was present in the received signals, but, as mentioned above, the technique is quite general and has over the years been used in a number of applications in telecommunications. Direct applications of the results in Ref. 9 to the problem of multiuser detection can be found in Refs. 12–16.

3.2. The Multiuser Channel

In Ref. 17, also published in 1988, Poor uses the EM algorithm to estimate the amplitudes of the user signals in a multiuser environment, in the presence of modulation. In this application, the modulation symbols (binary-valued) are treated as "nuisance" parameters. The complete data are chosen as the received (incomplete) data along with the binary modulation symbols over an observation window. The E step of the EM iteration involves the computation of conditional expectations of the binary-valued symbols, which results in "soft" data estimates in the form of hyperbolic tangents. At convergence, the algorithm yields the amplitude estimates and on slicing the soft-data estimates, also estimates of the binary modulation symbols (although no optimality can be claimed). Follow-up work that relates to Ref. 17 can be found in Refs. 18–21, in which the emphasis is not on amplitude estimation (amplitudes are assumed known) but on K-user multiuser detection. In Refs. 18, 20, and 21 the complete data are taken to be the received (baud rate) matched-filter samples along with the binary modulation symbols of $(K - 1)$ users, treated as interference in detecting a particular user. In addition to the application of the standard EM algorithm, the authors in Refs. 18, 20, and 21 also study the use of the space-alternating generalized EM (SAGE) algorithm [22,23], a variant of the EM algorithm that has better convergence properties and may simplify the maximization step. Other work in the area of spread-spectrum research that uses the SAGE algorithm in jointly detecting a single user in the presence of amplitude, phase, and time-delay uncertainties has also been presented [24,25].

3.3. Channel Estimation

In Refs. 26 and 27, the authors deal with detection in an impulsive noise channel, modeled as a class A mixture. In these papers the role of the EM algorithm is not directly in detection itself, but rather in estimating the triplet of parameters of the (Gaussian) mixture model, namely, the variances of the nominal and contaminant distributions and the probability of being under one or the other distribution. The estimate of the mixture model is then used (as if it were perfect) to select the nonlinearity to be used for detection. This particular application of the EM algorithm to estimate the Gaussian mixture model is one of its original and most typical uses. Follow-up (but more in-depth) work on the use of the EM algorithm to estimating class A noise parameters can be found in Ref. 28. Further follow-up work on the problem of detection in impulsive noise can be found in Refs. 29 and 30. As in Refs. 26 and 27, the authors in Refs. 29 and 30 use the EM algorithm to estimate parameters in the impulsive noise model, which are then used for detection in a spread-spectrum, coded environment. In Ref. 31 the EM algorithm is used to estimate the noise parameters in a spatial diversity reception system when the noise is modeled as a mixture of Gaussian distributions. Also in a spatial diversity environment, Baroso et al. [32] apply the EM algorithm to estimate blindly the multiuser array channel transfer function. Data detection in the paper is considered separately.

The EM algorithm has also been used for channel estimation under discrete signaling. In a symposium paper [33] and in a follow-up journal paper [34], Vallet

and Kaleh study the use of the EM algorithm for channel estimation modeled as having a finite impulse response, for both linear and nonlinear channels. In this work, the modulation symbols are considered as the "nuisance" parameters and the authors pose the problem as one of estimating the channel parameters. At convergence, symbol detection can be performed as well. Other work on the use of the EM algorithm to channel estimation/equalization in the presence of data modulation can be found in the literature [12,35–37]. Zamiri-Jafarian and Pasupathy present an algorithm for channel estimation, motivated by the EM algorithm, that is recursive in time.

3.4. The Unsynchronized Channel

In 1989 a paper dealing (for the first time) with sequence estimation using the EM algorithm appeared [38] with a follow-up journal paper appearing in 1991 [39]. In Refs. 38 and 39 the received data are "incomplete" in estimating the transmitted sequence because time synchronization is absent. The complete data in the paper were defined as the baud rate (nonsynchronized) matched-filter samples along with the correct timing phase. The E step of the algorithm was evaluated numerically, and the M step trivially corresponded to symbol-by-symbol detection in the absence of coding. The algorithm converged within 2–4 iterations, depending on the signal-to-noise ratio (SNR). The fast convergence of the algorithm can be attributed to the fact that the parameters to be estimated came from a discrete set. Follow-up work using the EM algorithm for the time unsynchronized channel appeared in 1994 and 1995 [40–42]. In one of Kaleh's papers [40], which in fact uses the Baum–Welch algorithm [4], a predecessor to EM, instead of posing the problem as sequence estimation in the presence of timing offset, the author poses the problem as one of estimating the timing offset (and additive noise variance) in the presence of random modulation. At convergence, symbol estimates can also be obtained. In Ref. 42, in addition to the absence of time synchronization, the authors assume an unknown amplitude. In the paper the authors use the SAGE algorithm [22] to perform sequence estimation.

3.5. The Random Phase Channel

Similar to Refs. 38 and 39, in a 1990 paper [43] the authors consider sequence estimation in the presence of random phase offset, for both uncoded and trellis-coded systems. We provide a brief description of the results here.

Let $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ be the vector containing the complex modulation symbols, $\mathbf{D}$ be a diagonal matrix with the elements of $\mathbf{s}$ as diagonal elements, and θ be the phase offset over the observation window. Then the received (incomplete) data vector $\mathbf{r}$ can be expressed as

$$\mathbf{r} = \mathbf{D}e^{j\theta} + \mathbf{N} \tag{7}$$

where $\mathbf{N}$ is a zero mean, i.i.d., complex, Gaussian noise vector. The parameter vector to be estimated is the modulation sequence $\mathbf{s}$. Let the complete data $\mathbf{x}$ consist of the incomplete data $\mathbf{r}$ along with knowledge of the random phase vector θ, namely, $\mathbf{x} = (\mathbf{r}, \theta)$. Assuming PSK modulation (the case of QAM can also be handled easily), the E step of the algorithm at the ith iteration yields

$$Q(\mathbf{s} \mid \mathbf{s}^i) = \Re\{\mathbf{r}^\dagger \mathbf{s} \cdot (\mathbf{r}^\dagger \mathbf{s}^i)^*\} \tag{8}$$

For the maximization step of the EM algorithm, we distinguish between coded and uncoded transmission. Observe from (8) that in the absence of coding, maximizing $Q(\mathbf{s} \mid \mathbf{s}^i)$ with respect to sequences $\mathbf{s}$ is equivalent to maximizing each individual term in the sum, specifically, making symbol-by-symbol decisions. When trellis coding is used, the maximization over all trellis sequences can be done efficiently using the Viterbi algorithm. Follow-up work on sequence estimation under phase offset can be found in the literature [44–47]. The use of the EM algorithm for phase synchronization in OFDM systems has also been studied [48].

3.6. The Fading Channel

Sequence estimation for fading channels using the EM algorithm was first studied in 1993 [49], with follow-up work a bit later [45,47,50]. Here the authors look at the problem as one of sequence estimation, where the random fading are the unwanted parameters. A brief summary of the EM formulation as presented in the references cited above is given below.

Let $\mathbf{D}$, $\mathbf{s}$, and $\mathbf{N}$ be as defined above and $\mathbf{a}$ be the complex fading process modeled as a zero-mean, Gaussian vector with independent real and imaginary parts having covariance matrix $\mathbf{Q}$. Then the received data $\mathbf{r}$ are

$$\mathbf{r} = \mathbf{D}\mathbf{a} + \mathbf{N} \tag{9}$$

Let the complete data $\mathbf{x}$ be the incomplete data $\mathbf{r}$ along with the random fading vector $\mathbf{a}$: $\mathbf{x} = (\mathbf{r}, \mathbf{a})$. Then, for PSK signaling, the expectation step of the EM algorithm yields

$$Q(\mathbf{s} \mid \mathbf{s}^i) = \Re[\mathbf{r}^\dagger \mathbf{D}\hat{\mathbf{a}}_i] \tag{10}$$

where

$$\hat{\mathbf{a}}_i = E[\mathbf{a} \mid \mathbf{r}, \mathbf{s}^i] = \mathbf{Q}\left[\mathbf{Q} + \frac{\mathbf{I}}{\mathrm{SNR}}\right]^{-1} (\mathbf{D}^i)^* \mathbf{r} \tag{11}$$

When the fading is static over a number of symbols, the expectation step of the algorithm becomes $Q(\mathbf{s} \mid \mathbf{s}^i) = \Re[\mathbf{r}^\dagger \mathbf{s}\hat{a}_i]$, where $\hat{a}_i = \mathbf{r}\mathbf{s}^{i\dagger}$. Again, the maximization step can be done easily and corresponds to symbol-by-symbol detection in the uncoded case, or Viterbi decoding in the trellis-coded case.

Other work on the use of the EM algorithm for sequence estimation in multipath/fading channels can be found in the literature [51–57]. In Refs. 54 and 55 the authors introduce an approximate sequence estimation algorithm motivated by the EM algorithm.

3.7. The Interference Channel

In addition to the multiuser interference channel, a number of papers applying the EM algorithm to sequence estimation in interference and/or to interference

suppression have appeared in the literature more recently. These include Refs. 58–62. An application to sequence estimation [60] is briefly presented below.

Let

$$r(t) = S(t;\mathbf{a}) + J(t) + n(t) \tag{12}$$

be the received data, where $S(t;\mathbf{a})$, $0 \le t < T$, is the transmitted signal (assuming some arbitrary modulation) corresponding to a sequence of M information symbols $\mathbf{a}$. $J(t)$ models the interference, and $n(t)$ is zero-mean, white, Gaussian noise having spectral density $N_0/2$. For generality, we will not specify the form of the signal part, $S(t;\mathbf{a})$, but special cases of interest include direct-sequence spread-spectrum (DSSS) and multicarrier spread-spectrum (MCSS) applications.

Let the complete data be the received data $r(t)$ along with the interference $J(t)$ (which is arbitrary at this point). Application of the EM algorithm yields the following expectation step ($\mathbf{a}^k$ is the sequence estimate at the kth iteration and $\langle\cdot,\cdot\rangle$ denotes inner product):

$$Q(\mathbf{a} \mid \mathbf{a}^k) = \langle r(t) - \hat{J}^k(t), S(t;\mathbf{a})\rangle \tag{13}$$

$$\hat{J}^k(t) = E[J(t) \mid \{r(\tau) : 0 \le \tau \le T\}, \mathbf{a}^k] \tag{14}$$

These expressions are quite general and can be applied to arbitrary interference models, provided the conditional mean estimates (CMEs) in (14) can be computed. For a vector-space interference model

$$J(t) = \sum_{j=1}^{N} g_j \phi_j(t) \tag{15}$$

where $\{\phi_1(t), \phi_2(t), \ldots, \phi_N(t)\}$ is an orthonormal basis set and $g_j, j = 1, 2, \ldots, N$ are zero-mean, uncorrelated random variables having corresponding variances λ_j, we have

$$\hat{J}^k(t) = \sum_{i=1}^{N} \hat{g}_i^k \phi_i(t) \tag{16}$$

where

$$\hat{g}_i^k = \left(1 + \frac{N_0}{2\lambda_i}\right)^{-1} \langle r(t) - S(t;\mathbf{a}^k), \phi_i(t)\rangle$$

4. SOME RESULTS ON APPLICATION OF THE EM ALGORITHM

We present briefly next some more recent results on the use of the EM algorithm in telecommunications.

4.1. Decoding of Spacetime-Coded Systems

The results presented here are a brief summary of those reported in Ref. 63. We consider a mobile radio system where the transmitter is equipped with N transmit antennas and the mobile receiver is equipped with M receive antennas. Data blocks of length L are encoded by a spacetime encoder. After serial-to-parallel conversion, the coded symbol stream is divided into N substreams, each of which is transmitted via one transmit antenna simultaneously. The transmitted code block can be written in matrix form as

$$\mathbf{D} = \begin{pmatrix} d_1^{(1)} & d_2^{(1)} & \ldots & d_N^{(1)} \\ d_1^{(2)} & d_2^{(2)} & \ldots & d_N^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ d_1^{(L)} & d_2^{(L)} & \ldots & d_N^{(L)} \end{pmatrix} \tag{17}$$

where the superscript in $d_n^{(l)}$ represents time index and the subscript is the space index. In other words, $d_n^{(l)}$ is the complex symbol transmitted by the nth antenna during the lth symbol time. We denote the row vectors of $\mathbf{D}$ by $\mathbf{D}^{(l)}$.

Assuming quasistatic fading (the nonstatic case has been handled as well), let

$$\mathbf{\Gamma}_j = (\,\gamma_{1j} \quad \gamma_{2j} \quad \cdots \quad \gamma_{Nj}\,)^T, j = 1, 2, \ldots, M$$

be the fading vector whose components γ_{ij} are the fading coefficients in the channel connecting the ith transmit antenna to the jth receive antenna, for $i = 1, 2, \ldots, N$. The γ_{ij} are assumed independent. Then the received data at the jth receive antenna are

$$\mathbf{Y}_j = \mathbf{D}\mathbf{\Gamma}_j + \mathcal{N}_j, j = 1, 2, \ldots, M$$

Defining the complete data as $\{\mathbf{Y}_j, \mathbf{\Gamma}_j\}_{j=1}^M$, the E step of the EM algorithm at the kth iteration yields

$$\mathbf{Q}(\mathbf{D} \mid \mathbf{D}^k) = \sum_{l=1}^{L}\sum_{j=1}^{M}\left[\Re(\overline{y_j^{(l)}}\mathbf{D}^{(l)}\hat{\mathbf{\Gamma}}_j^k) - \frac{1}{2}\mathbf{D}^{(l)}(\hat{\mathbf{\Omega}}_j)^k(\mathbf{D}^{(l)})^*\right] \tag{18}$$

$$\hat{\mathbf{\Gamma}}_j^k = \left((\mathbf{D}^k)^*\mathbf{D}^k + \frac{\mathbf{I}}{\mathrm{SNR}}\right)^{-1}(\mathbf{D}^k)^*\mathbf{Y}_j \tag{19}$$

$$\hat{\mathbf{\Omega}}_j^k = \mathbf{I} - \left((\mathbf{D}^k)^*\mathbf{D}^k + \frac{\mathbf{I}}{\mathrm{SNR}}\right)^{-1}(\mathbf{D}^k)^*\mathbf{D}^k + \hat{\mathbf{\Gamma}}_j^k(\hat{\mathbf{\Gamma}}_j^k)^* \tag{20}$$

and the M step

$$\mathbf{D}^{k+1} = \arg\max_{\mathbf{D}} \sum_{l=1}^{L}\sum_{j=1}^{M}\left[\Re(\overline{y_j^{(l)}}\mathbf{D}^{(l)}\hat{\mathbf{\Gamma}}_j^k) - \frac{1}{2}\mathbf{D}^{(l)}\hat{\mathbf{\Omega}}_j^k(\mathbf{D}^{(l)})^*\right] \tag{21}$$

For the spacetime codes designed in Ref. 64, the Viterbi algorithm can be used to efficiently perform the required maximization. Simulation results are shown in Fig. 1 for the 8-state code over QPSK introduced in Ref. 64.

Other work using the EM algorithm in the spacetime area can be found in Ref. 65. This paper uses again Feder and Weistein's results [9] in a multitransmit antenna, multipath environment in conjunction with orthogonal frequency-division multiplexing (OFDM).

4.2. SNR Estimation

Estimation of the SNR and the received signal energy is an important function in digital communications. We present

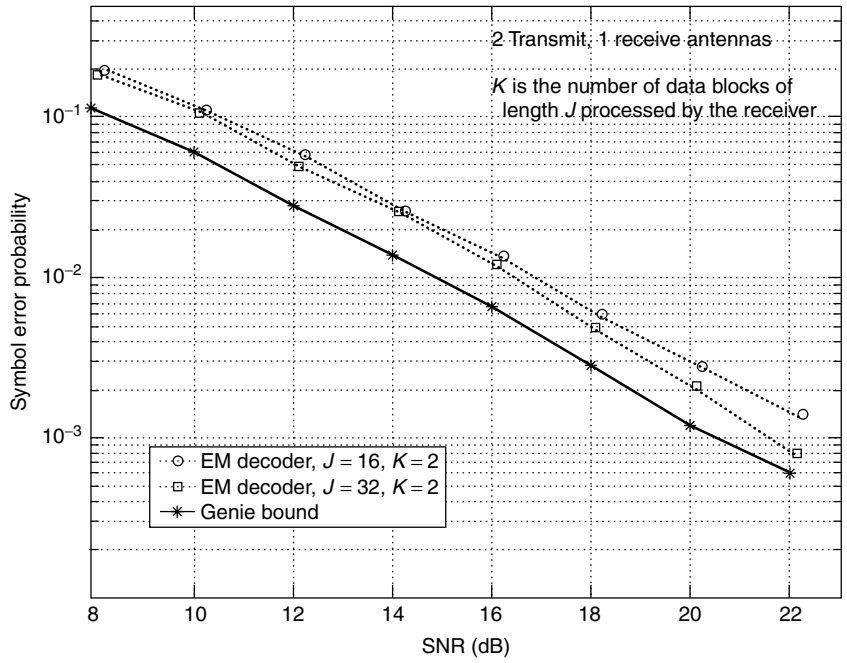


Figure 1. Symbol error probability of the "genie" and EM-based decoders as a function of block length: $N = 2$, $M = 1$, 8-state space-time code, quasistatic fading.

here an application of the EM algorithm to this problem. This is a summary of results from Ref. 66.

Let the received data be

$$r_k = \sqrt{E}d_k + \sqrt{v}n_k, k = 1, 2, \ldots, N$$

where E is the received signal energy; v is the additive noise variance; the n_k are zero-mean, unit-variance, i.i.d. Gaussian; and the d_k are binary modulation symbols (the more general case is handled as well) with $d_k \in \{1, -1\}$. We are interested in estimating the vector $\boldsymbol{\theta} = (E, v)$ from the received (incomplete) data vector $\mathbf{r}$ using the EM algorithm. Let the complete data be $\mathbf{x} = (\mathbf{r}, \mathbf{d})$. Then the E step of the EM algorithm yields

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^i) = -N\ln(v) - \frac{1}{v}\sum_{k=1}^{N} r_k^2 - \frac{N \cdot E}{v} + \frac{2\sqrt{E}}{v} \cdot \sum_{k=1}^{N} r_k \hat{d}_k^i \tag{22}$$

where

$$\hat{d}_k^i = E\left[d_k \mid \mathbf{r}, \boldsymbol{\theta}^i\right] = \tanh\left(\frac{\sqrt{E^i}}{v^i} r_k\right) \tag{23}$$

Taking derivatives w.r.t. (with respect to) E and v, the M step of the EM algorithm yields the following recursions:

$$\hat{E}^{i+1} = \left(\frac{B^i}{N}\right)^2 \tag{24}$$

$$\hat{v}^{i+1} = \frac{A}{N} - \hat{E}^{i+1} \tag{25}$$

$$\hat{\mathrm{SNR}}^{i+1} = \frac{\hat{E}^{i+1}}{\hat{v}^{i+1}} \tag{26}$$

where

$$A = \sum_{k=1}^{N} r_k^2 \tag{27}$$

$$B^i = \sum_{k=1}^{N} r_k \tanh\left(\frac{\sqrt{E^i}}{v^i} r_k\right) \tag{28}$$

Figure 2 shows results on the bias and mean-square error in estimating the SNR and the received signal energy for the EM algorithm and a popular high-SNR approximation to the ML estimator (the true ML estimator is much too complex to implement).

4.3. On the (Non) Convergence of the EM Algorithm for Discrete Parameters

For continuous parameter estimation, under continuity and weak regularity conditions, stationary points of the EM algorithm have to be stationary points of the loglikelihood function [e.g., 11]. On the other hand, any fixed point can be a convergence point of the discrete EM algorithm even though the likelihood of such a point can be lower than the likelihood of a neighboring discrete point. In this subsection we summarize some results on the (non)convergence of the discrete EM algorithm presented in Ref. 67.

Let $\mathbf{a}$ be a M-dimensional discrete parameter with values in $\mathcal{A}$. It is assumed that, along with the discrete EM algorithm, a companion continuous EM algorithm

$$\mathbf{a}_c^{k+1} = \arg\max_{\mathbf{a} \in \mathcal{C}^{\mathbf{M}}} Q(\mathbf{a} \mid \mathbf{a}^k) \tag{29}$$

where $\mathcal{C}^M$ is the M-dimensional complex space, is well defined. The companion continuous EM mapping is the map $\mathbf{M}_c : \mathbf{a}^k \to \mathbf{a}_c^{k+1}$, for $\mathbf{a}_c^{k+1} \in \mathcal{C}^M$. The Jacobian of the continuous EM mapping $\mathbf{M}_c(\mathbf{a})$ is denoted with $\mathbf{J}_c(\mathbf{a})$, and the Hessian of its corresponding objective function $Q(\mathbf{a} \mid \mathbf{a}^k)$ is denoted with $\mathbf{H}^Q(\mathbf{a})$.

Computationally "inexpensive" maximization of the objective function $Q(\mathbf{a} \mid \mathbf{a}^k)$ over $\mathcal{A}$ can be obtained for

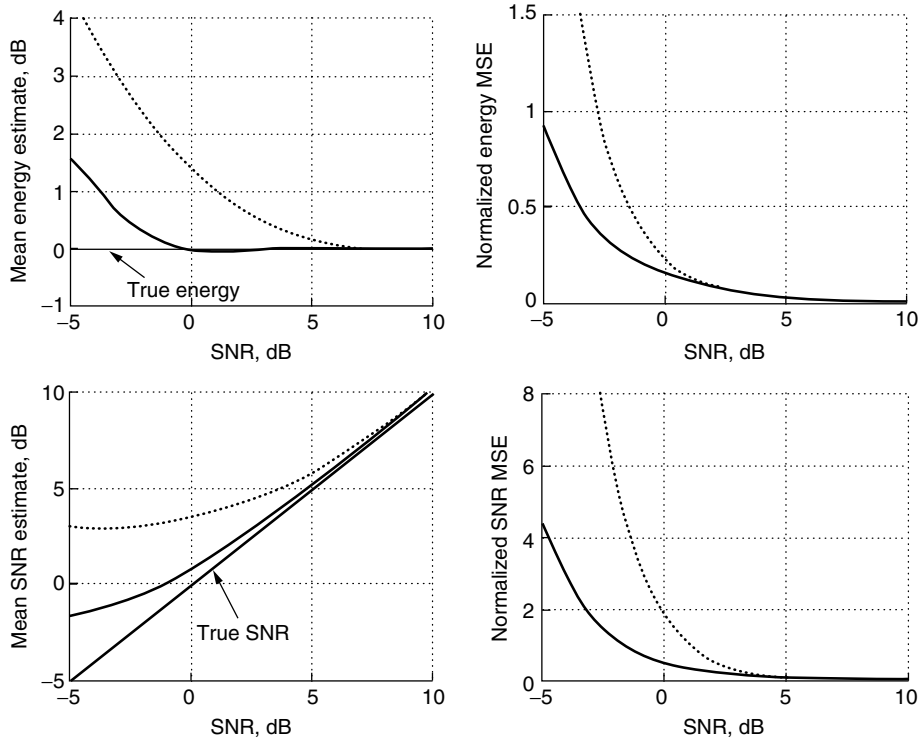


Figure 2. Comparison between the EM and high SNR ML approximation (dashed lines) for sequence length $N = 50$.

some special structures. The particular case of interest is the quadratic form

$$Q(\mathbf{a} \mid \mathbf{a}^k) = -\tfrac{1}{2}\|\mathbf{a} - \mathbf{a}_c^{k+1}\|^2 \tag{30}$$

for which the discrete maximization in the *M step* can be obtained on a parameter-by-parameter basis. Equation (30) implies that $\mathbf{a}^{k+1}$ is the discrete point closest to the continuous maximizer of $Q(\mathbf{a} \mid \mathbf{a}^k)$ over $\mathcal{C}^M$, $\mathbf{a}_c^{k+1}$. Note that $\mathbf{a}_c^{k+1}$ is (implicitly) a function of the previous iterate $\mathbf{a}^k$. As a special case of (30), $\mathbf{a}^{k+1} = \text{sign}[\mathbf{a}_c^{k+1}]$ for QPSK signaling, namely, for $a_{d,i} \triangleq [\mathbf{a}]_i \in \{\pm 1 \pm j\}$. Clearly, not only is the objective function (30) quadratic but, additionally, its Hessian is $\mathbf{H}^Q(\mathbf{a}) = -\mathbf{I}$. It is important to note that (30) holds for all the sequence estimation problems described in this overview [see Refs. 8, 10, and 13 for a linear modulation $S(t; \mathbf{a})$] except for (18), where the objective function is quadratic in the unknown symbol sequence $\{d_j^{(l)}\}$ but the Hessian is not $-\mathbf{I}$ in the general case. The arguments given below can easily be generalized for a nonidentity Hessian matrix.

Let $d_{\min}$ denote half the minimum Euclidean distance between any two discrete points from $\mathcal{A}$:

$$d_{\min} = \tfrac{1}{2} \min_{\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{A}} \|\mathbf{a}_1 - \mathbf{a}_2\|. \tag{31}$$

$d_{\min} = 1$ for uncoded binary antipodal signaling. Let $\hat{\mathbf{a}}_c$ be a fixed point of the companion continuous EM algorithm (29). One of the authors [67] has been shown, based on the first two terms of the Taylor expansion of $\mathbf{M}_c(\mathbf{a})$ in a neighborhood $U(\hat{\mathbf{a}}_c)$ of $\hat{\mathbf{a}}_c$, that all $\mathbf{a} \in \mathcal{A} \cap U(\hat{\mathbf{a}}_c)$ such that

$$\|\hat{\mathbf{a}}_c - \mathbf{a}\| < \frac{d_{\min}}{1 - \hat{\lambda}_{\min}} \tag{32}$$

where $\hat{\lambda}_{\min} \triangleq \lambda_{\min}[\mathbf{J}_c(\hat{\mathbf{a}}_c)]$ is the minimum eigenvalue of the Jacobian matrix $\mathbf{J}_c(\hat{\mathbf{a}}_c)$, are stationary points of the discrete EM algorithm. Inequality (32) follows from the fact that the eigenvalues of $\mathbf{J}_c(\hat{\mathbf{a}}_c)$ are nonnegative and smaller than one (see lemma in Ref. 2). It defines a ball of radius $r^{nc} = d_{\min}/(1 - \hat{\lambda}_{\min})$ centered at a fixed point of the companion continuous EM algorithm $\hat{\mathbf{a}}_c$.

Parameter $\hat{\lambda}_{\min}$ is, in the general case, a function of the received signal. As shown by Meng and Rubin [3] (see also Ref. 11), $\mathbf{J}_c(\hat{\mathbf{a}}_c)$ and, consequently, its eigenvalues can be estimated in a reliable manner using a continuous EM algorithm. Furthermore, when the complete data of the EM algorithm are chosen in such a way that $\mathbf{H}^Q(\hat{\mathbf{a}}_c) = -\mathbf{I}$ as in (30), then matrices $\mathbf{J}_c(\hat{\mathbf{a}}_c)$ and the Hessian of the loglikelihood function, $\mathbf{H}^l(\hat{\mathbf{a}}_c)$, have the same eigenvectors. Their eigenvalues have the following relationship: $1 - \lambda_i[\mathbf{J}_c(\hat{\mathbf{a}}_c)] = \lambda_i[-\mathbf{H}^l(\hat{\mathbf{a}}_c)]$, for all i. Thus

$$r^{nc} = \frac{d_{\min}}{\lambda_{\max}[-\mathbf{H}^l(\hat{\mathbf{a}}_c)]}$$

The term $(1 - \hat{\lambda}_{\min})$ is the largest eigenvalue of the matrix $\mathbf{I} - \mathbf{J}_c(\hat{\mathbf{a}}_c)$, often referred to as the *iteration matrix* in the optimization literature [e.g., 11], since it determines the convergence iteration steps in the neighborhood $U(\hat{\mathbf{a}}_c)$. Its smallest eigenvalue is a typical measure of convergence speed of the continuous EM algorithm since it measures the convergence step along the slowest direction. Its largest eigenvalue determines the convergence step along the fastest direction. Thus, we can say that the nonconvergence radius is determined by (it is in fact inversely proportional to) the largest convergence step of the companion continuous EM algorithm.

An important implication of (32) is that $\hat{\lambda}_{\min} = 0$ is sufficient for the nonconvergence ball to hold at most one discrete stationary point. Sequence estimation, in the presence of interference when $\hat{\lambda}_{\min} = 0$ or when this identity can be forced using rank-reduction principles, has been analyzed in Ref. 67 and in part in Ref. 70, and is presented briefly next.

In the following, the received signal model given in (12) with a linear modulation

$$S(t;\mathbf{a}) = \sum_{m=0}^{M-1} a_n h(t - mT) \tag{33}$$

where pulses $\{h(t - mT), m \in [0, M-1]\}$ are orthonormal, is assumed.

For this problem, a reduced nonconvergence ball radius has been derived [67] without having to resort to the Taylor expansion of $\mathbf{M}_c(\mathbf{a})$. The reduced nonconvergence radius is a function of only $d_{\min}$ and the statistics of noise and interference. In the special case when $N = M$ and $\text{span}\{h(t - nT), n \in [0, N-1]\} = \text{span}\{\phi_n(t), n \in [0, N-1]\}$, where $\phi_n(t)$ are defined in (15), the reduced radius and r^{nc} are equal. Both are

$$r^{nc} = \left[1 + \frac{\tilde{\lambda}_{\min}}{N_0}\right] d_{\min} \tag{34}$$

where $\tilde{\lambda}_{\min} = \min_j\{\lambda_j\}$; λ_j is the variance of the coefficient g_j in (15). For example, let's assume $M = N = 2$, interference energy $J = \lambda_1 + \lambda_2 = 2\lambda_1 = 2\tilde{\lambda}_{\min}$, binary antipodal signaling with $d_{\min} = 1$, and noise and signal realizations such that $\hat{\mathbf{a}}_c \approx \mathbf{0}$ (see Ref. 67 for details). Then, if $J/N_0 > 2\sqrt{2} - 2$ all possible discrete points in $\mathcal{A}$ will be found inside the convergence ball whose radius is given by (34). Consequently, the discrete EM algorithm will generate a sequence of identical estimates $\mathbf{a}^k = \mathbf{a}^0$ for any k and any initial point $\mathbf{a}^0 \in \mathcal{A}$.

Next, we provide symbol error rate results for a binary communications example where $N > M$ and $\text{span}\{h(t - mT),\ m \in [0, M-1]\} \subset \text{span}\{\phi_n(t),\ n \in [0, N-1]\}$ holds. Figure 3 demonstrates that, in the case of an interference process whose full spectrum covers the spectrum of the useful signal, one could significantly increase the convergence probability of the EM algorithm using a rank-reduction approach. Two second-order Gaussian processes uniformly positioned in frequency within the signal spectra are used to model interference using $N > M$ eigenfunctions. The EM algorithm obtained by modeling the interference with $P < M < N$ largest eigenfunctions $\phi_n(t)$, and thus forcing $\hat{\lambda}_{\min} = 0$ in (32), is compared to the performance of the EM algorithm that uses the full-rank model. It can be observed that the reduced-rank EM algorithm has little degradation, whereas the full-rank EM algorithm does not converge and has a catastrophic performance for large SNR.

4.4. Detection in Fading Channels with Highly Correlated Multipath Components

Here we study another example where the convergence of the discrete EM algorithm can be stymied by a large nonconvergence ball.

We assume a continuous-time L-path Rayleigh fading channel model:

$$\begin{aligned} r_1(t) &= \sum_{k=1}^{K}\sum_{l=1}^{L} \alpha_{l,k,1} S(t - \tau_{l,k,1}; \mathbf{a}_k) + n_1(t) \\ &\vdots \\ r_{N_a}(t) &= \sum_{k=1}^{K}\sum_{l=1}^{L} \alpha_{l,k,N_a} S(t - \tau_{l,k,N_a}; \mathbf{a}_k) + n_{N_a}(t) \end{aligned} \tag{35}$$

Here N_a is the number of receive antennas and K is the number of users; $\mathbf{a}_k$ is the vector of N symbols of

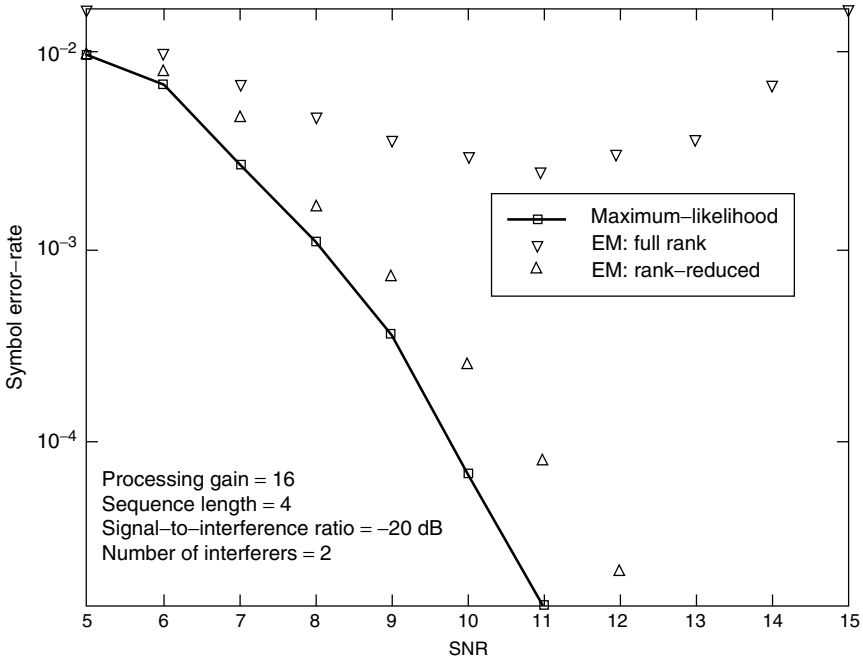


Figure 3. Comparison of ML, reduced-rank EM, and full-rank EM receivers for two second-order Gaussian interferers.

user k; $\alpha_{l,k,i}$ are zero-mean, complex circular Gaussian random variables of respective variances $\gamma_{l,k,i}$ and $\tau_{l,k,i}$ are (assumed known) path delays. We group the fading coefficients, $\alpha_{l,k,i}$, for each i into a vector $\boldsymbol{\alpha}_i$ such that $\alpha_{l,k,i} = [\boldsymbol{\alpha}_i]_{l+L\cdot(k-1)}$ is the $l + L\cdot(k-1)$th element of the vector $\boldsymbol{\alpha}_i$. The covariance matrix of $\boldsymbol{\alpha}_i$ is $\boldsymbol{\Gamma}_i$. P_s of the transmitted symbols are assumed to be known pilot symbols that allow for ambiguity resolution (67). The signal is assumed to have the following form:

$$S(t;\mathbf{a}_k) = \sum_{n=1}^{N} a_{n,k} h(t - nT)$$

where $h(t)$ is the known signaling pulse (assumed common for simplicity) modulated by the transmitted symbol sequence $\mathbf{a}$ of length N. Pulses $h(t - iT)$ are typically orthogonal to avoid ISI for frequency nonselective channels. The vector of time-delayed modulated signals is $\mathbf{s}_i(t;\mathbf{a}) = [S(t-\tau_{1,1,i};\mathbf{a}_1), S(t-\tau_{2,1,i};\mathbf{a}_1), \ldots, S(t-\tau_{L,K,i};\mathbf{a}_K)]^T$, where $\mathbf{a}$ is a NK-dimensional vector of user symbols such that its subvector of length K $[\mathbf{a}]_{(k-1)\cdot N+1}^{k\cdot N}$ is the vector of N symbols of user k.

For simplicity the receiver antennas are assumed to be sufficiently separated so that any $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ for $i \neq j$ are independent. $\mathbf{G}_i(\mathbf{a}) = \langle \mathbf{s}_i(t;\mathbf{a}), \mathbf{s}_i^H(t;\mathbf{a})\rangle$ is the Gramian of $\mathbf{s}_i(t;\mathbf{a})$ in the L_2 space.

An EM solution is nontrivially based on the EM solution for superimposed signals introduced by Feder and Weinstein described in Section 3.1 (see, also Refs. 53 and 69). The complete data vector includes not only the path plus noise components but also the fading vectors $\boldsymbol{\alpha}_i$, as

$$(\mathbf{x}, \boldsymbol{\alpha}) \triangleq ([\mathbf{x}_1, \ldots, \mathbf{x}_{N_A}], \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{N_a}) \tag{36}$$

$$\triangleq (\{x_{1,1,1}(t), \ldots, x_{L,K,N_a}(t), \quad t \in [T_i, \ldots, T_f]\}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{N_a}) \tag{37}$$

where $x_{l,k,i}(t) \triangleq \alpha_{l,k,i} S(t - \tau_{l,k,i};\mathbf{a}_k) + n_{l,k,i}(t)\cdot n_{l,k,i}(t)$ is a complex zero-mean AWGN process with variance $\beta_{l,k}N_0$ such that $E\{n_{l,k,i}(t) n^*_{j,m,i}(t)\} = 0$ for all $l \neq j$ and $k \neq m$ and $n_i(t) = \sum_{l=1}^{L}\sum_{k=1}^{K} n_{l,k,i}(t)$. Clearly, $\beta_{l,k}$ has to satisfy the constraint $\sum_{l=1}^{L}\sum_{k=1}^{K}\beta_{l,k} = 1$. In the following, $\beta_{l,k} \equiv 1/(KL)$.

The terms $x_{l,k,i}(t)$ are mutually independent given the data sequence $\mathbf{a}$ and the fading vectors $\boldsymbol{\alpha}_i$, and the EM objective function, in the case of orthogonal signaling, can be represented in the following manner:

$$Q(\mathbf{a} \mid \mathbf{a}^m) = \sum_{k=1}^{K}\sum_{n=1}^{N}\left[\Re\left\{a_{n,k}\sum_{i=1}^{N_a}\sum_{l=1}^{L}\beta_{l,k}\hat{z}_{l,n,k,i}(\mathbf{a}^m)\right\} - \frac{1}{2}|a_{n,k}|^2\sum_{i=1}^{N_a}\sum_{l=1}^{L}\beta_{l,k}\hat{\rho}_{ll,kk,i}(\mathbf{a}^m)\right] \tag{38}$$

Here $\hat{z}_{l,n,k,i}(\mathbf{a}^m) = \langle \hat{\chi}_{l,k,i}(t;\mathbf{a}^m), h(t - \tau_{l,k,i} - nT)\rangle$ is the sampled pulse-matched filter response to signal path component estimates $\hat{\chi}_{l,k,i}(t;\mathbf{a}^m)$.

The E step requires estimation of the phase and amplitude compensated signal path components

$$\hat{\chi}_{l,k,i}(t;\mathbf{a}^m) = E\{\alpha^*_{l,k,i}x_{l,k,i}(t) \mid \{r_i(t); t \in [T_i, \ldots, T_f]\}, \mathbf{a}^m\}$$
$$= \hat{\rho}_{ll,kk,i}(\mathbf{a}^m)S(t - \tau_{l,k,i};\mathbf{a}^m) + \beta_{l,k}\left[\hat{\alpha}^*_{l,k,i}(\mathbf{a}^m)r_i(t) - \sum_{q=1}^{K}\sum_{j=1}^{L}\hat{\rho}_{lj,kq,i}(\mathbf{a}^m)S(t - \tau_{j,q,i};\mathbf{a}^m)\right]$$

Signal path components are estimated by successive refinement. The refinement attempts to find those component estimates that explain the received signal with a smallest measurement error. $\hat{\alpha}_{l,k,i}(\mathbf{a}^m)$ and $\hat{\rho}_{lj,kq,i}(\mathbf{a}^m)$ are, respectively, the conditional mean estimates of the complex coefficients

$$\hat{\alpha}_{l,k,i}(\mathbf{a}^m) = E\{\alpha_{l,k,i} \mid \{r_i(t); t \in [T_i, \ldots, T_f]\}, \mathbf{a}^m\}$$
$$= [(N_0\boldsymbol{\Gamma}_i^{-1} + \mathbf{G}_i(\mathbf{a}^m))^{-1}\langle \mathbf{s}_i(t;\mathbf{a}^m), r_i^H(t)\rangle]_l$$

and their cross-correlations

$$\hat{\rho}_{lj,kq,i}(\mathbf{a}^m) = E\{\alpha^*_{l,k,i}\alpha_{j,q,i} \mid \{r_i(t); t \in [T_i, \ldots, T_f]\}, \mathbf{a}^m\}$$
$$= \left[\left(\boldsymbol{\Gamma}_i^{-1} + \frac{\mathbf{G}_i(\mathbf{a}^m)}{N_0}\right)^{-1} + \hat{\boldsymbol{\alpha}}_i(\mathbf{a}^m)\hat{\boldsymbol{\alpha}}_i^H(\mathbf{a}^m)\right]_{l+(k-1)L, j+(q-1)L} \tag{39}$$

The M step of the companion continuous EM algorithm can be expressed in closed form as

$$\hat{a}^{m+1}_{cu,n,k} = \frac{\sum_{i=1}^{N_a}\sum_{l=1}^{L}\beta_{l,k}\hat{z}_{l,n,k,i}(\mathbf{a}^m)}{\sum_{i=1}^{N_a}\sum_{l=1}^{L}\beta_{l,k}\hat{\rho}_{ll,kk,i}(\mathbf{a}^m)}, (n,k) \in \overline{\mathcal{J}}_{ps}$$
$$\hat{a}^{m+1}_{cu,n,k} = a^k_n, (n,k) \in \mathcal{J}_{ps} \tag{40}$$

where $\mathcal{J}_{ps}$ are index pairs corresponding to pilot symbols and

$$\overline{\mathcal{J}}_{ps} = \{(1,1), \ldots, (K,N)\}\backslash\mathcal{J}_{ps}$$

It combines in an optimal manner the phase/amplitude-compensated signal path components estimated in the E step. In this way, it achieves multipath diversity combining.

The M step of the (discrete) EM algorithm requires quantization, for example, taking the sign of the components of $\mathbf{a}^{m+1}_{cu}$, in case of binary signaling.

The number of fixed points (and, consequently, convergence to the ML estimate) of the discrete EM algorithm is a function of the path Gramian matrices $\mathbf{G}_i(\mathbf{a})$ whose structure defines the size of the nonconvergence ball. For highly correlated paths the nonconvergence ball will be so large that we need to use other detection methods.

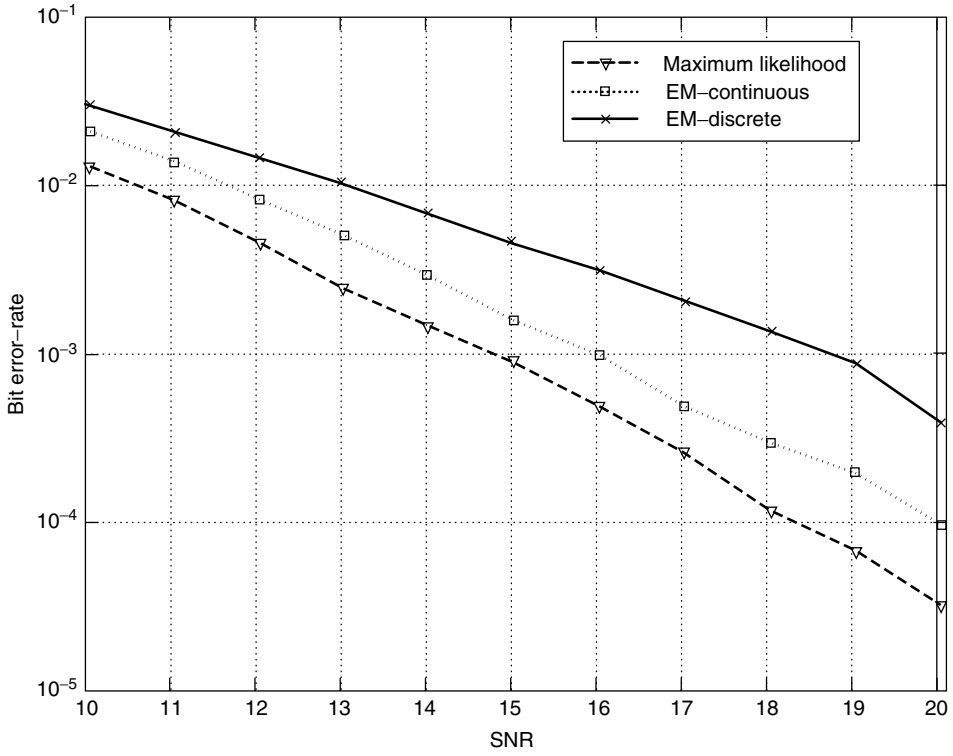


Figure 4. Performance comparison for highly correlated path components.

An alternative approach to detection termed *generalized decorrelator* (see, e.g., Refs. 18, 53, 68, and 69) allows the continuous EM algorithm to converge and quantizes the continuous maximizer. For uncoded BPSK signaling we have

$$\hat{\mathbf{a}}_c = \text{sign}[\mathbf{a}_c^{\infty}]$$

In the following example, the delays are $\tau_l/T \in \{0, \frac{1}{3}, \frac{4}{3}\}$. The number of antennas is 1, processing gain is equal to 3, the sequence length is 6 and the pilot symbol block length is 3. The paths are fading independently and their variance is $1/L$. The last two signal path components are highly correlated because of their integer symbol delay. Cross-correlation to the first component is also high as a result of the small processing gain. This particular case is detrimental to receivers that are based on the decorrelation between signal path components. The generalized decorrelator (continuous EM algorithm), as can be seen in Fig. 4, achieves a performance within 1.5–2 dB of the optimal. It manages diversity combining even for this particularly difficult situation. On the other hand, the discrete EM algorithm has a loss of 1–3 dB relative to the continuous EM detector.

5. CONCLUSION

We have provided a short overview of the use of the EM algorithm in telecommunications. Clearly, the EM algorithm has established itself as a generic tool in the design of telecommunication systems. If current work using it in various problems is any indication, the algorithm will get even more attention in the future.

BIOGRAPHIES

Costas N. Georghiades received his B.E. degree, with distinction, from the American University of Beirut in 1980, and his M.S. and D.Sc. degrees from Washington University in 1983 and 1985, respectively, all in electrical engineering. Since 1985, he has been with the Electrical Engineering Department at Texas A&M University, where he is a professor and holder of the J.W. Runyon, Jr. Endowed Professorship. He currently serves as director of the Telecommunications and Signal Processing Group in the department. Over the years he served in editorial positions with the *IEEE Transactions on Communications*, the *IEEE Transactions on Information Theory*, the *IEEE Journal on Selected Areas in Communications* and the *IEEE Communications Letters*. Dr. Georghiades is a fellow of the IEEE. His general interests are in the application of information and estimation theories to the study of communication systems, with particular interest in optimum receiver design, wireless communication, and optical systems.

Predrag Spasojevic received the Diploma of Engineering degree from the School of Electrical Engineering, University of Sarajevo, in 1990; and his M.S. and Ph.D. degrees in electrical engineering from Texas A&M University, College Station, Texas, in 1992 and 1999, respectively. From 2000 to 2001 he has been with WINLAB, Rutgers University, where he is currently an assistant professor in the Department of Electrical & Computer Engineering. His research interests are in the general areas of communication theory and signal processing.

BIBLIOGRAPHY

1. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc.* **39**: 1–17 (1977).
2. X. L. Meng and D. B. Rubin, On the global and componentwise rates of convergence of the EM algorithm, *Linear Algebra Appl.* **199**: 413–425 (1994).
3. X. L. Meng and D. B. Rubin, Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, *J. Am. Stat. Assoc.* **86**: 899–909 (1991).
4. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, A maximization technique in statistical estimation for probabilistic functions of Markov chains, *Ann. Math. Stat.* **41**: 164–171 (1970).
5. C. F. Wu, On the convergence properties of the EM algorithm, *Ann. Stat.* **11**(1): 95–103 (1983).
6. L. A. Shepp and Y. Vardi, Maximum-likelihood reconstruction for emission tomography, *IEEE Trans. Med. Imag.* **1**: 113–122 (Oct. 1982).
7. D. L. Snyder and D. G. Politte, Image reconstruction from list-mode data in an emission tomography system having time-of-flight measurements, *IEEE Trans. Nucl. Sci.* **30**(3): 1843–1849 (1983).
8. D. L. Snyder and M. I. Miller, *Random Processes in Time and Space*, Springer-Verlag, New York, 1991, Chap. 3.
9. M. Feder and E. Weinstein, Parameter estimation of superimposed signals using the EM algorithm, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-36**: 477–489 (April 1988).
10. T. Moon, The expectation-maximization algorithm, *IEEE Signal Process. Mag.* (Nov. 1996).
11. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, 1997.
12. M. Feder and J. A. Catipovic, Algorithms for joint channel estimation and data recovery-application to equalization in underwater communications, *IEEE J. Ocean. Eng.* **16**(1): 42–55 (Jan. 1991).
13. J. W. Modestino, Reduced-complexity iterative maximum-likelihood sequence estimation on channels with memory, *Proc. 1993 IEEE Int. Symp. Inform. Theory* 422–422 (1993).
14. A. Radović, An iterative near-far resistant algorithm for joint parameter estimation in asynchronous CDMA systems, *5th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, 1994, Vol. 1, pp. 199–203.
15. M. J. Borran and M. Nasiri-Kenari, An efficient decoding technique for CDMA communication systems based on the expectation maximization algorithm, *IEEE 4th Int. Symp. Spread Spectrum Techniques and Applications Proc.*, 1996, Vol. 3, pp. 1305–1309.
16. A. Chkeif and G. K. Kaleh, Iterative multiuser detector with antenna array: An EM-based approach, *Proc. Int. Symp. Information Theory*, 1998, p. 424.
17. H. V. Poor, On parameter estimation in DS/SSMA formats, *Proc. Advances in Communications and Control Systems*, Baton Rouge, LA, Oct. 1988, pp. 59–70.
18. L. B. Nelson and H. V. Poor, Soft-decision interference cancellation for AWGN multiuser channels, *Proc. 1994 IEEE Int. Symp. Information Theory*, 1994, p. 134.
19. H. V. Poor, Adaptivity in multiple-access communications, *Proc. 34th IEEE Conf. Decision and Control*, 1995, Vol. 1, pp. 835–840.
20. L. B. Nelson and H. V. Poor, EM and SAGE algorithms for multi-user detection, *Proc. 1994 IEEE-IMS Workshop on Information Theory and Statistics*, 1994, p. 70.
21. L. B. Nelson and H. V. Poor, Iterative multiuser receivers for CDMA channels: An EM-based approach, *IEEE Trans. Commun.* **44**(12): 1700–1710 (Dec. 1996).
22. J. A. Fessler and A. O. Hero, Complete-data spaces and generalized EM algorithms, *1993 IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (ICASSP-93), 1993, Vol. 4, pp. 1–4.
23. J. A. Fessler and A. O. Hero, Space-alternating generalized EM algorithm, *IEEE Trans. Signal Process.* (Oct. 1994).
24. I. Sharfer and A. O. Hero, A maximum likelihood CDMA receiver using the EM algorithm and the discrete wavelet transform, *1996 IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (ICASSP-93), 1996, Vol. 5, pp. 2654–2657.
25. C. Carlemalm and A. Logothetis, Blind signal detection and parameter estimation for an asynchronous CDMA system with time-varying number of transmission paths, *Proc. 9th IEEE SP Workshop on Statistical Signal and Array Processing*, 1998, pp. 296–299.
26. D. Zeghlache and S. Soliman, Use of the EM algorithm in impulsive noise channels, *Proc. 38th IEEE Vehicular Technology Conf.*, 1988, pp. 610–615.
27. D. Zeghlache, S. Soliman, and W. R. Schucany, Adaptive detection of CPFSK signals in non-Gaussian noise, *Proc. 20th Southeastern Symp. System Theory*, 1988, pp. 114–119.
28. S. M. Zabin and H. V. Poor, Efficient estimation of class A noise parameters via the EM algorithm, *IEEE Trans. Inform. Theory* **37**(1): 60–72 (Jan. 1991).
29. A. Ansari and R. Viswanathan, Application of expectation-maximizing algorithm to the detection of direct-sequence signal in pulsed noise jamming, *IEEE Military Communications Conf., (MILCOM '92) Conference Record*, 1992, Vol. 3, pp. 811–815.
30. A. Ansari and R. Viswanathan, Application of expectation-maximization algorithm to the detection of a direct-sequence signal in pulsed noise jamming, *IEEE Trans. Commun.* **41**(8): 1151–1154 (Aug. 1993).
31. R. S. Blum, R. J. Kozick, and B. M. Sadler, An adaptive spatial diversity receiver for non-Gaussian interference and noise, *1st IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, 1997, pp. 385–388.
32. V. A. N. Baroso, J. M. F. Moura, and J. Xavier, Blind array channel division multiple access (AChDMA) for mobile communications, *IEEE Trans. Signal Process.* **46**: 737–752 (March 1998).
33. R. Vallet and G. K. Kaleh, Joint channel identification and symbols detection, *Proc. Int. Symp. Information Theory*, 1991, p. 353.
34. G. K. Kaleh and R. Vallet, Joint parameter estimation and symbol detection for linear and nonlinear unknown channels, *IEEE Trans. Commun.* **42**: 2406–2413 (July 1994).
35. Min Shao and C. L. Nikias, An ML/MMSE estimation approach to blind equalization, *1994 IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (ICASSP-94), 1994, Vol. 4, pp. 569–572.
36. H. Zamiri-Jafarian and S. Pasupathy, Recursive channel estimation for wireless communication via the EM algorithm,

1997 IEEE Int. Conf. Personal Wireless Communications, 1997, pp. 33–37.

37. H. Zamiri-Jafarian and S. Pasupathy, EM-Based Recursive estimation of channel parameters, *IEEE Trans. Commun.* **47**: 1297–1302 (Sept. 1999).
38. C. N. Georghiades and D. L. Snyder, An application of the expectation-maximization algorithm to sequence detection in the absence of synchronization, *Proc. Johns Hopkins Conf. Information Sciences and Systems*, Baltimore, MD, March 1989.
39. C. N. Georghiades and D. L. Snyder, The expectation maximization algorithm for symbol unsynchronized sequence detection, *IEEE Trans. Commun.* **39**(1): 54–61 (Jan. 1991).
40. G. K. Kaleh, The Baum-Welch Algorithm for detection of time-unsynchronized rectangular PAM signals, *IEEE Trans. Commun.* **42**: 260–262 (Feb.–April 1994).
41. N. Antoniadis and A. O. Hero, Time-delay estimation for filtered Poisson processes using an EM-type algorithm, *IEEE Trans. Signal Process.* **42**(8): 2112–2123 (Aug. 1994).
42. I. Sharfer and A. O. Hero, Spread spectrum sequence estimation and bit synchronization using an EM-type algorithm, *1995 Int. Conf. Acoustics, Speech, and Signal Processing* (ICASSP-95), 1995, Vol. 3, pp. 1864–1867.
43. C. N. Georghiades and J. C. Han, Optimum decoding of trellis-coded modulation in the presence of phase-errors, *Proc. 1990 Int. Symp. Its Applications* (ISITA' 90), Hawaii, Nov. 1990.
44. C. N. Georghiades, Algorithms for Joint Synchronization and Detection, in *Coded Modulation and Bandwidth Efficient Transmission*, Elsevier, Amsterdam, 1992.
45. C. N. Georghiades and J. C. Han, On the application of the EM algorithm to sequence estimation for degraded channels, *Proc. 32nd Allerton Conf. Univ. Illinois*, Sept. 1994.
46. C. R. Nassar and M. R. Soleymani, Joint sequence detection and phase estimation using the EM algorithm, *Proc. 1994 Canadian Conf. Electrical and Computer Engineering*, 1994, Vol. 1, pp. 296–299.
47. C. N. Georghiades and J. C. Han, Sequence Estimation in the presence of random parameters via the EM algorithm, *IEEE Trans. Commun.* **45**: 300–308 (March 1997).
48. E. Panayirci and C. N. Georghiades, Carrier phase synchronization of OFDM systems over frequency-selective channels via the EM algorithm, *1999 IEEE 49th Vehicular Technology Conf.*, 1999, Vol. 1, pp. 675–679.
49. J. C. Han and C. N. Georghiades, Maximum-likelihood sequence estimation for fading channels via the EM algorithm, *Proc. Communication Theory Mini Conf.*, Houston, TX, Nov. 1993.
50. J. C. Han and C. N. Georghiades, Pilot symbol initiated optimal decoder for the land mobile fading channel, *1995 IEEE Global Telecommunications Conf.* (GLOBECOM '95), 1995, pp. 42–47.
51. K. Park and J. W. Modestino, An EM-based procedure for iterative maximum-likelihood decoding and simultaneous channel state estimation on slow-fading channels, *Proc. 1994 IEEE Int. Symp. Information Theory*, 1994, p. 27.
52. L. M. Zeger and H. Kobayashi, MLSE for CPM signals in a fading multipath channel, *1999 IEEE Pacific Rim Conf. Communications, Computers and Signal Processing*, 1999, pp. 511–515.
53. P. Spasojević and C. N. Georghiades, Implicit diversity combining based on the EM algorithm, *Proc. WCNC '99*, New Orleans, LA, Sept. 1999.
54. H. Zamiri-Jafarian and S. Pasupathy, Generalized MLSDE via the EM algorithm, *1999 IEEE Int. Conf. Communication*, 1999, pp. 130–134.
55. H. Zamiri-Jafarian and S. Pasupathy, Adaptive MLSDE using the EM algorithm, *IEEE Trans. Commun.* **47**(8): 1181–1193 (Aug. 1999).
56. M. Leconte and F. Hamon, Performance of /spl pi/-constellations in Rayleigh fading with a real channel estimator, *1999 IEEE Int. Conf. Personal Wireless Communication*, 1999, pp. 183–187.
57. W. Turin, MAP decoding using the EM algorithm, *1999 IEEE 49th Vehicular Technology Conf.*, 1999, Vol. 3, pp. 1866–1870.
58. Q. Zhang and C. N. Georghiades, An application of the EM algorithm to sequence estimation in the presence of tone interference, *Proc. 5th IEEE Mediterranean Conf. Control and Systems*, Paphos, Cyprus, July 1997.
59. O. C. Park and J. F. Doherty, Generalized projection algorithm for blind interference suppression in DS/CDMA communications, *IEEE Trans. Circuits Syst. II: Analog Digital Signal Process.* **44**(6): 453–460 (June 1997).
60. C. N. Georghiades, Maximum-likelihood detection in the presence of interference, *Proc. IEEE Int. Symp. Inform. Theory* 344 (Aug. 1998).
61. C. N. Georghiades and D. Reynolds, *Interference Rejection for Spread-Spectrum Systems Using the EM Algorithm*, Springer-Verlag, London, 1998.
62. Q. Zhang and C. N. Georghiades, An interference rejection application of the EM algorithm to direct-sequence signals, *Kybernetika* (March 1999).
63. Y. Li, C. N. Georghiades, and G. Huang, Iterative maximum likelihood sequence estimation of space-time codes, *IEEE Trans. Commun.* **49**: 948–951 (June 2001).
64. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**: 744–765 (March 1998).
65. Y. Xie and C. N. Georghiades, An EM-based channel estimation algorithm for OFDM with transmitter diversity, *Proc. Globecom 2001*, San Antonio, TX, Nov. 2001.
66. C. N. Georghiades and U. Dasgupta, On SNR and energy estimation, manuscript in preparation.
67. P. Spasojević, *Sequence and Channel Estimation for Channels with Memory*, dissertation, Texas A&M Univ., College Station, TX, Dec. 1999.
68. P. Spasojević and C. N. Georghiades, The slowest descent method and its application to sequence estimation, *IEEE Trans. Commun.* **49**: 1592–1601 (Sept. 2001).
69. P. Spasojević, Generalized decorrelators for fading channels, *Int. Conf. Information Technology: Coding and Computing*, April 2001, pp. 312–316.
70. C. N. Georghiades and P. Spasojević, Maximum-likelihood detection in the presence of interference, *IEEE Trans. Commun.* (submitted).

F

FADING CHANNELS

ALEXANDRA DUEL-HALLEN
North Carolina State University
Raleigh, North Carolina

1. INTRODUCTION

Radio communication channels include shortwave ionospheric radiocommunication in the 3–30-MHz frequency band (HF), tropospheric scatter (beyond-the-horizon) radio communications in the 300–3000 MHz frequency band (UHF) and 3000–30,000-MHz frequency band (SHF), and ionospheric forward scatter in the 30–300-MHz frequency band (VHF) [1,2]. These channels usually exhibit *multipath propagation* that results from *reflection, diffraction* and *scattering* of the transmitted signal. Multipath propagation and *Doppler effects* due to the motion of the transmitter and/or receiver give rise to *multipath fading* channel characterization. The fading signal is characterized in terms of *large-scale* and *small-scale* fading.

The large-scale fading model describes the average received signal power as a function of the distance between the transmitter and the receiver. Statistical variation around this mean (on the order of 6–10 dB) resulting from *shadowing* of the signal due to large obstructions is also included in the model of large-scale fading. Large-scale fading describes the variation of the received power over large areas and is useful in estimating radio coverage of the transmitter. The large scale fading model is determined by averaging the received power over many wavelengths (e.g., 1–10 m for cellular and PCS frequencies of 1–2 GHz) [3].

The small-scale fading model describes the instantaneous variation of the received power over a few wavelengths. This variation can result in dramatic loss of power on the order of 40 dB. It is due to superposition of several reflected or scattered components coming from different directions. Figure 1 depicts small-scale fading and slower large-scale fading for a mobile radiocommunication system. The figure illustrates that small-scale variation is averaged out in the large-scale fading model.

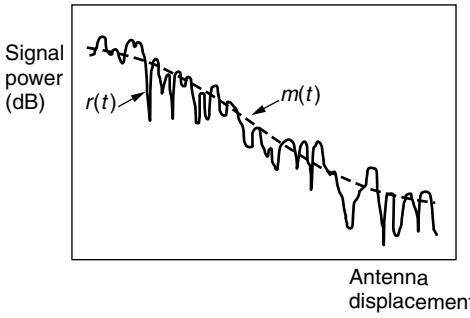


Figure 1. Large- and small-scale fading (reprinted from Ref. 2 © IEEE).

In Section 2, we review the large-scale fading models. Sections 3–5 describe small-scale fading. Multipath fading channels are characterized in Section 3. Section 4 describes useful fading models. Performance analysis for flat fading channels and diversity combining approaches are described in Section 5.

2. LARGE-SCALE FADING

Propagation path loss $L_s(d)$ is defined as the ratio of the transmitted power to the received power in a radiofrequency (RF) channel. In a free-space model, the propagation path loss is proportional to d^2, where d is the distance between the transmitter and the receiver. When the receiving antenna is isotropic, this loss is given by [2]

$$L_s(d) = \left(\frac{4\pi d}{\lambda}\right)^2 \tag{1}$$

where λ is the wavelength of the RF signal. In the mobile radio channel, the average path loss is usually more severe due to obstructions and is inversely proportional to d^n, where $2 \le n \le 6$. The average path loss is given by [2,3]

$$L_{\text{av}}(d)\ (\text{dB}) = L_s(d_0) + 10n \log_{10} \frac{d}{d_0} \tag{2}$$

where d_0 is the close-in reference distance. This distance corresponds to a point located in the far field of the antenna. For large cells, it is usually assumed to be 1 km, whereas for smaller cells and indoor environments, the values of 100 m and 1 m, respectively, are used. The value of the exponent n depends on the frequency, antenna heights, and propagation conditions. For example, in urban area cellular radio, n takes on values from 2.7 to 3.5; in building line-of-sight conditions, $1.6 \le n \le 1.8$; whereas in obstructed in building environments, $n = 4$–6 [3].

The actual path loss in a particular location can deviate significantly from its average value (2) due to *shadowing* of the signal by large obstructions. Measurements show that this variation is approximately *lognormally* distributed. Thus, the path loss is represented by the random variable

$$L_p(d)\ (\text{dB}) = L_{\text{av}}(d)\ (\text{dB}) + X_\sigma \tag{3}$$

where X_σ has Gaussian distribution (when expressed in decibels) with mean zero and standard deviation σ (also in decibels). Note that average path loss (2) corresponds to a straight line with slope $10n$ (dB) per decade when plotted on log–log scale. Thus, in practice, the values of n and σ are determined using linear regression to minimize the difference between the actual measurements (over various locations and distances between the transmitter and receiver) and the estimated average path loss in the mean-squared-error (MSE) sense. Figure 2 illustrates the actual measured data and the estimated average path loss for several transmitter-receiver separation values.

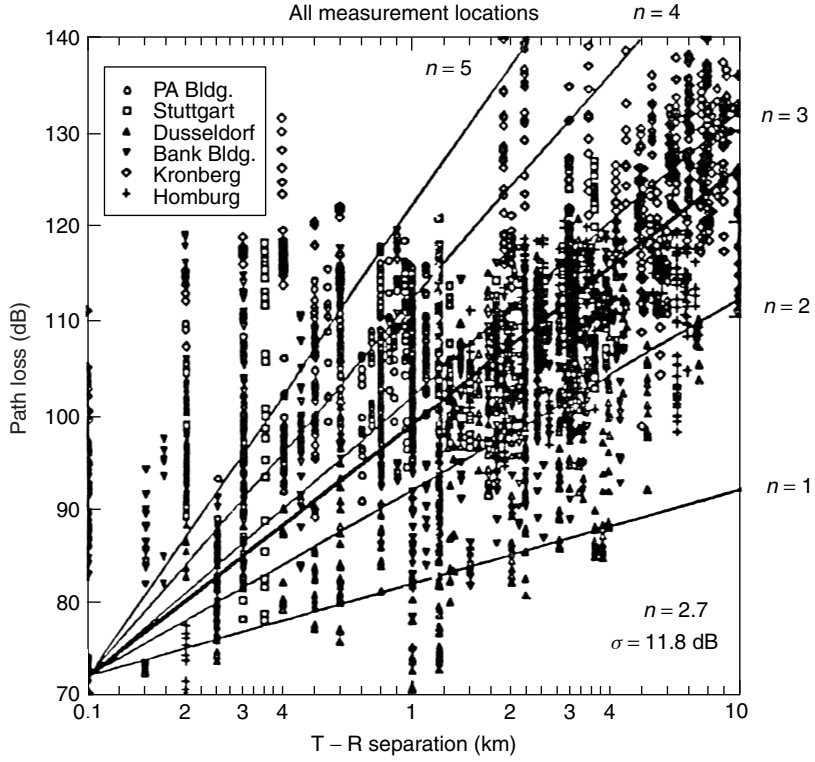


Figure 2. Scatterplot of measured data and corresponding MMSE path loss model for many cities in Germany. For these data, $n = 2.7$ and $\sigma = 11.8$ dB (reprinted from Ref. 15, © IEEE).

References 1–7 describe appropriate models and approaches to measurement and estimation of propagation path loss for various wireless channels and provide additional useful references on this subject. The rest of this article is devoted to characterization, modeling, and performance analysis of small-scale fading.

3. CHARACTERIZATION OF FADING MULTIPATH CHANNELS

A signal transmitted through a wireless channel arrives at its destination along a number of different paths (referred to as *multipath propagation*) as illustrated in Fig. 3. Multipath causes interference between *reflected or scattered* transmitter signal components. As the receiver moves through this *interference pattern*, a typical fading signal results as illustrated in Fig. 4. If an unmodulated carrier at the frequency f_c is transmitted over a fading channel, the complex envelope (the *equivalent lowpass signal*) [1] of the received fading signal is given by

$$c(t) = \sum_{n=1}^{N} A_n e^{j(2\pi f_n t + 2\pi f_c \tau_n + \phi_n)} \tag{4}$$

where N is the number of scatterers, and for the nth scatterer, A_n is the *amplitude*, f_n is the *Doppler frequency shift*, τ_n is the *excess propagation delay* (relative to the arrival of the first path), and ϕ_n is the phase. The *Doppler frequency shift* is given by [3]

$$f_n = f_c \frac{v}{c} \cos\theta = f_{dm} \cos\theta \tag{5}$$

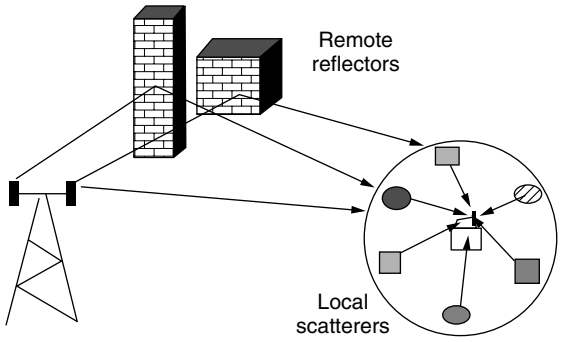


Figure 3. Typical mobile radio environment.

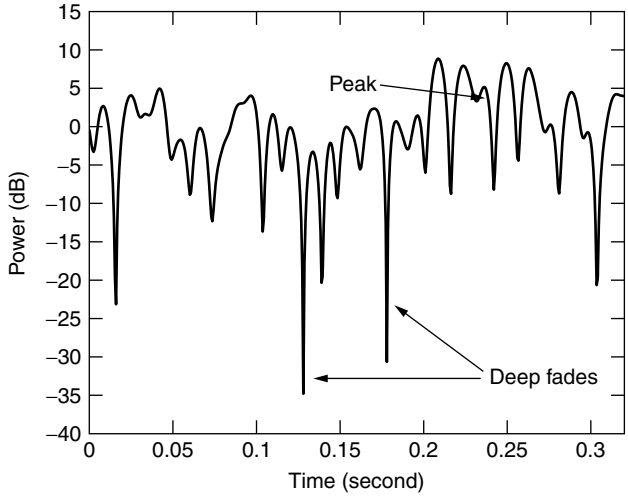


Figure 4. A typical fading signal (provided by Ericsson Inc.).

where v is the vehicle speed (assumed constant), c is the speed of light, θ is the incident radiowave angle with respect to the motion of the mobile, and f_{dm} is the *maximum Doppler frequency shift*.

The parameters A_n, f_n, τ_n, and ϕ_n are very *slowly time-variant,* and can be viewed as fixed on the timescale of a few milliseconds. Thus, the signal in (4) is a superposition of complex sinusoids with approximately constant amplitudes, frequencies, and phases, and varies in time as the mobile moves through the interference pattern. The superposition of terms in (4) can result in destructive or constructive interference, causing deep fades or peaks in the received signal, respectively, as illustrated in Fig. 4. The power of the fading signal can change dramatically, by as much as 30–40 dB. This variation can be conveniently modeled by characterizing $c(t)$ as a *stationary random process*. This statistical characterization is useful for describing time dispersion and fading rapidity of multipath fading channels [1–9].

3.1. Statistical Characterization

If we assume that the complex envelope of the transmitted signal is $s_l(t)$, the equivalent baseband signal received at the output of the fading channel is

$$r(t) = \int_{-\infty}^{\infty} c(\tau, t) s_l(t - \tau)\, d\tau \tag{6}$$

where the time-variant impulse response $c(\tau, t)$ is the response of the channel at time t to the impulse applied at time $t - \tau$ [1]. (In practice, additive Gaussian noise is also present at the output of the channel.) Expression (6) can be viewed as the superposition of delayed and attenuated copies of the transmitted signal $s_l(t)$, where the delays are given by τ_n and the corresponding complex gains have amplitudes A_n [see (4)] and time-variant phases [determined from the phase terms in (4)].

For each delay τ, the response $c(\tau, t)$ is modeled as a *wide-sense stationary* stochastic process. Typically, the random processes $c(\tau_1, t)$ and $c(\tau_2, t)$ are uncorrelated for $\tau_1 \neq \tau_2$ since different multipath components contribute to these signals (this is called *uncorrelated scattering*). For fixed delay τ, the autocorrelation function of the impulse response is defined as [1]

$$\phi_c(\tau; \Delta t) = \tfrac{1}{2} E[c^*(\tau, t) c(\tau, t + \Delta t)] \tag{7}$$

The power of $c(\tau, t)$ as a function of the delay τ is

$$\phi_c(\tau; 0) \equiv \phi_c(\tau) \tag{8}$$

This is called *multipath intensity profile* or *power delay profile* and is determined from measurements [1,3]. A "worst case" multipath intensity profile for an urban channel is shown in Fig. 5. The range of values of τ where $\phi_c(\tau)$ is nonnegligible is called the *multipath spread* of the channel and denoted as T_m. The values of multipath spread vary greatly depending on the terrain. For urban and suburban areas, typical values of multipath spread are from 1 to 10 μs, whereas in rural mountainous area, the multipath spreads are much greater with values from 10 to 30 μs [1]. The *mean excess delay* and *RMS delay spread* σ_τ are defined as the mean and the standard deviation of the excess delay, respectively, and can be determined from the multipath intensity profile [2,3]. Typical RMS delay spreads are on the order of one microsecond for urban outdoor channels, hundreds of nanoseconds for suburban channels, and tens of nanoseconds for indoor channels [3].

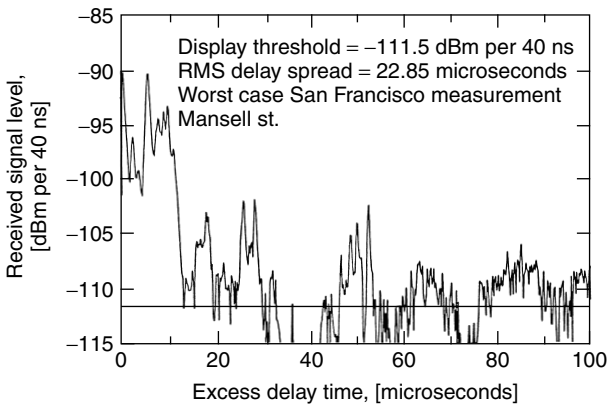


Figure 5. Measured multipath power delay profiles: from a 900-MHz cellular system in San Francisco (reprinted from Ref. 16, © IEEE).

The Fourier transform of the channel response $c(\tau, t)$ is the time-variant channel transfer function $C(f; t)$. It is also modeled as a wide-sense stationary random process. The correlation

$$\phi_c(\Delta f; \Delta t) = \tfrac{1}{2} E[C^*(f; t) C(f + \Delta f; t + \Delta t)] \tag{9}$$

is called the *spaced-frequency spaced-time autocorrelation function*. It is the Fourier transform of the autocorrelation function $\phi_c(\tau; \Delta t)$ in (7) [1]. It can be shown that this function can be factored into a product of time-domain and frequency-domain correlation functions. The latter is the *spaced-frequency correlation function* $\phi_c(\Delta f; 0) \equiv \phi_c(\Delta f)$ and is the Fourier transform of the multipath intensity profile $\phi_c(\tau)$ in (8) [1].

The complex envelope of the response $C(f; t)$ is specified by (4) with f viewed as the frequency of the unmodulated input carrier. Consider input frequency separation Δf. In the expressions for $C(f; t)$ and $C(f + \Delta f; t)$, the multipath components corresponding to the Doppler shift f_n have phase separation $\Delta\phi_n = 2\pi \Delta f \tau_n$. As Δf increases, these phase shifts result in decreased correlation between fading responses associated with two frequencies [10]. The *coherence bandwidth* $(\Delta f)_c$ provides a measure of this correlation. If the frequency separation is less than $(\Delta f)_c$, the signals $C(f; t)$ and $C(f + \Delta f; t)$ are strongly correlated, and thus fade similarly. The coherence bandwidth is inversely proportional to the multipath spread. However, the exact relationship between the coherence bandwidth and the multipath spread depends on the underlying meaning of the strong correlation of fading signals at different frequencies and varies depending on the channel model [2,3]. For example, if $|\phi_c(\Delta f)|$ is required to remain above 0.9, the corresponding coherence bandwidth is defined as $(\Delta f)_c \approx 1/(50\sigma_\tau)$, where σ_τ is the RMS delay

spread. When the frequency correlation is allowed to decrease to 0.5, greater coherence bandwidth $(\Delta f)_c \approx 1/(5\sigma_\tau)$ results.

The time variation of the channel response $c(\tau, t)$ due to the Doppler shift can be statistically characterized using the *spaced-time correlation function* determined from (9) as

$$\phi_c(0; \Delta t) \equiv \phi_c(\Delta t) \tag{10}$$

or its Fourier transform $S_c(\lambda)$ [1]. The function $S_c(\lambda)$ is called the *Doppler power spectrum* of the channel. As the maximum Doppler shift increases, the channel variation becomes more rapid [see (4)], and $S_c(\lambda)$ widens, resulting in *spectral broadening* at the receiver. The shape of the autocorrelation function depends on channel characteristics. For example, the popular Rayleigh fading channel discussed in Section 4 has the autocorrelation function

$$\phi_c(\Delta t) = J_0(2\pi f_{dm} \Delta t) \tag{11}$$

where $J_0(.)$ is the zero-order Bessel function of the first kind [11]. The Doppler power spectrum for this channel is given by

$$S_c(\lambda) = \frac{1}{\pi f_{dm}} \left[1 - \left(\frac{f}{f_{dm}} \right)^2 \right]^{1/2}, \quad |f| \le f_{dm} \tag{12}$$

These functions are plotted in Fig. 6.

Time variation of the fading channel is characterized in terms of the *Doppler spread* and the *coherence time*. The Doppler spread B_d is defined as the range of frequencies over which the Doppler power spectrum is essentially nonzero. For example, in (12), $B_d = 2f_{dm}$. The coherence time $(\Delta t)_c$ measures the time interval over which the time variation is not significant, or the samples of the fading signal are strongly correlated when their time separation is less than the coherence time, although different interpretations are used [3]. The Doppler spread and the coherence time are inversely proportional to one another. A popular rule of thumb is $(\Delta t)_c = 0.423/f_{dm}$. This definition implies that if the time separation is greater than the coherence time, the signals will be affected differently by the channel.

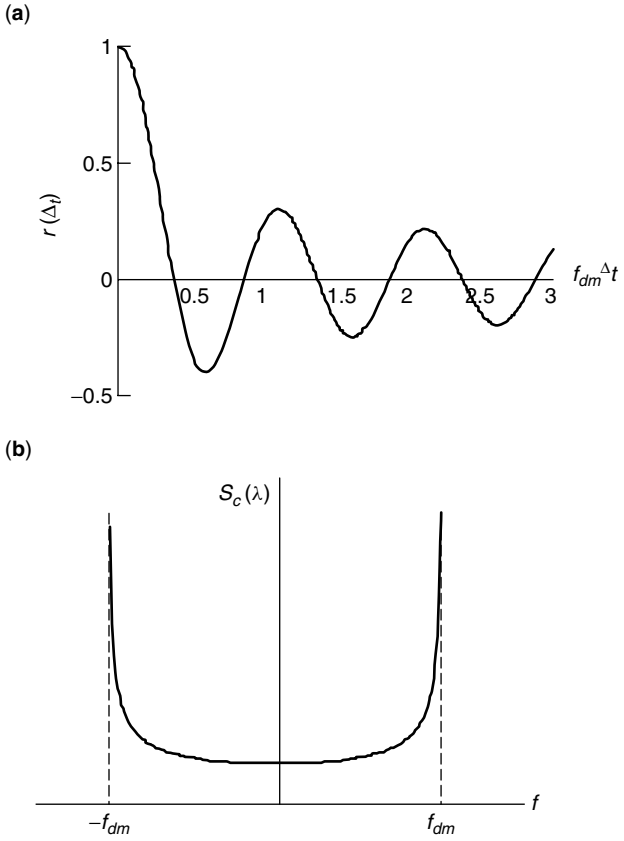


Figure 6. (**a**) The spaced-time autocorrelation function of the Rayleigh fading channel; (**b**) the Doppler power spectrum of the Rayleigh fading channel.

3.2. Relationship Between Signal Characteristics and the Fading Channel Model

Consider a linear communication system that transmits data at the symbol rate $1/T$ and employs a pulseshape with the complex envelope $s_l(t)$ and the frequency response $S_l(f)$ [1]. Assume that the bandwidth W of the transmitted signal is approximately $1/T$. The resulting output signal is characterized in terms of the response of the channel to this pulseshape as in (6), and the frequency response of the output signal is given by $C(f; t)\, S_l(f)$. The relationship between the symbol interval (or bandwidth W) and the statistical fading-channel parameters defined above dictate the choice of the underlying channel model used in analyzing performance of wireless communication systems.

First, consider the *multipath channel characterization* and *signal dispersion*. Suppose that the symbol interval is much larger than the multipath spread of the channel, $T \gg T_m$. Then all multipath components arrive at the receiver within a small fraction of the symbol interval. In this case, the coherence bandwidth significantly exceeds the bandwidth of the signal, $W \ll (\Delta f)_c$. Thus, all spectral components are affected by the channel similarly. Also, there is no *multipath-induced intersymbol interference* (ISI) in this case. This channel is modeled as complex time-varying attenuation $c(t)$ (4) [also given by $C(0; t)$], so the complex envelope of the received signal is $r(t) = c(t)\, s_l(t)$. It is called *frequency-nonselective*, or *flat fading* channel. These channels primarily occur in *narrowband* transmission systems.

If the symbol interval is smaller the multipath spread of the channel, $T < T_m$, or equivalently, the coherence bandwidth is smaller than the signal bandwidth, $W > (\Delta f)_c$, the channel becomes *frequency-selective* (it is also sometimes called the *multipath fading channel*). A common rule of thumb is that the channel is frequency-selective if $T < 10\sigma_\tau$, and flat fading if $T \ge 10\sigma_\tau$, where σ_τ is the RMS delay spread [3] With this definition of a frequency-selective channel, spectral components of the transmitted signal separated by the coherence bandwidth fade differently, resulting in frequency diversity, as discussed in Section 5. On the other hand, frequency

selectivity also causes dispersion, or ISI, since delayed versions of the transmitted signal arrive at the receiver much later (relative to the symbol interval) than components associated with small delays. This channel is often modeled using several fading rays with different excess multipath delays

$$c(t) = c_1(t)\delta(t-\tau_1) + c_2(t)\delta(t-\tau_2) + \cdots + c_L(t)\delta(t-\tau_L) \tag{13}$$

where the components $c_l(t), l = 1, \ldots, L$ are uncorrelated flat fading (e.g., Rayleigh distributed) random variables. The powers associated with these rays are determined by the multipath intensity profile (8).

Now, consider *rapidity*, or *time variation* of the fading channel. The channel is considered *slowly varying* if the channel response changes much slower than the symbol rate. In this case, the symbol interval is much smaller than the coherence time, $T \ll (\Delta t)_c$, or the signal bandwidth significantly exceeds the Doppler spread, $W \gg B_d$. If the symbol interval is comparable to or greater than the coherence time, (or the coherence bandwidth is similar to or exceeds the signal bandwidth), the channel is *fast-fading*. While most mobile radio, or PCS, channels are slowly fading, as the velocity of the mobile and the carrier frequency increase, the channel becomes *rapidly time-varying* since the Doppler shift increases [see (5)]. This rapid time variation results in time selectivity (which can be exploited as time diversity), but degrades reliability of detection and channel estimation [12–14].

4. FADING-CHANNEL MODELS

The *complex Gaussian distribution* is often used to model the equivalent lowpass flat-fading channel. This model is justified since superposition of many scattered components approximates a Gaussian distribution by the central-limit theorem. Even if the number of components in (5) is modest, experiments show that the Gaussian model is often appropriate. The *Rayleigh fading* process models fading channels without strong *line of sight* (LoS). Define $c(t) = c_I(t) + jc_Q(t)$, where the in-phase (real) and quadrature (imaginary) components are independent and identically distributed (i.i.d.) zero-mean stationary real Gaussian processes with variances σ^2. The average power of this process is $\frac{1}{2}E[c^*(t)c(t)] = \sigma^2$. The amplitude of this process has a Rayleigh distribution with the probability density function (PDF):

$$p_R(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{14}$$

and the phase is uniformly distributed:

$$p_\theta(\theta) = \frac{1}{2\pi},\ |\theta| \le \pi \tag{15}$$

The Rayleigh distribution is a special case of the *Nakagami-m* distribution that provides a more flexible model of the statistics of the fading channel. The PDF of the amplitude of the Nakagami-m distribution is [1]

$$p_R(r) = \frac{2}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m r^{2m-1}\exp\left(-\frac{mr^2}{\Omega}\right) \tag{16}$$

where $\Gamma(.)$ is the Gamma function [11], $\Omega = E(R^2)$ and the parameter m is the *fading figure* given by $m = \Omega^2/E[(R^2-\Omega)^2]$, $m \ge 1/2$. While the *Rayleigh* distribution uses a single parameter $E(R^2) = 2\sigma^2$ to match the fading statistics, the Nakagami-m distribution depends on two parameters, $E(R^2)$ and m. For $m = 1$ the density (16) reduces to Rayleigh distribution. For $\frac{1}{2} \le m < 1$, the tail of the Nakagami-m PDF decays slower than for Rayleigh fading, whereas for $m > 1$, the decay is faster. As a result, the Nakagami-m distribution can model fading conditions that are either more or less severe than Rayleigh fading. The Nakagami-m PDF for different values of m is illustrated in Fig. 7.

The Rayleigh distribution models a complex fading channel with zero mean that is appropriate for channels without the line-of-sight (LoS) propagation. When strong nonfading, or specular, components, such as LoS propagation paths, are present, a DC component needs to be added to random multipath, resulting in *Ricean* distribution. The pdf of the amplitude of Ricean fading is given by

$$p_R(r) = \frac{r}{\sigma^2}\exp -\left(\frac{r^2+s^2}{2\sigma^2}\right) I_0\left(\frac{rs}{\sigma^2}\right), \qquad r \ge 0 \tag{17}$$

where s is the peak amplitude of the dominant nonfading component and $I_0(.)$ is the modified Bessel function of the first kind and zero order [11]. The *Ricean factor K*

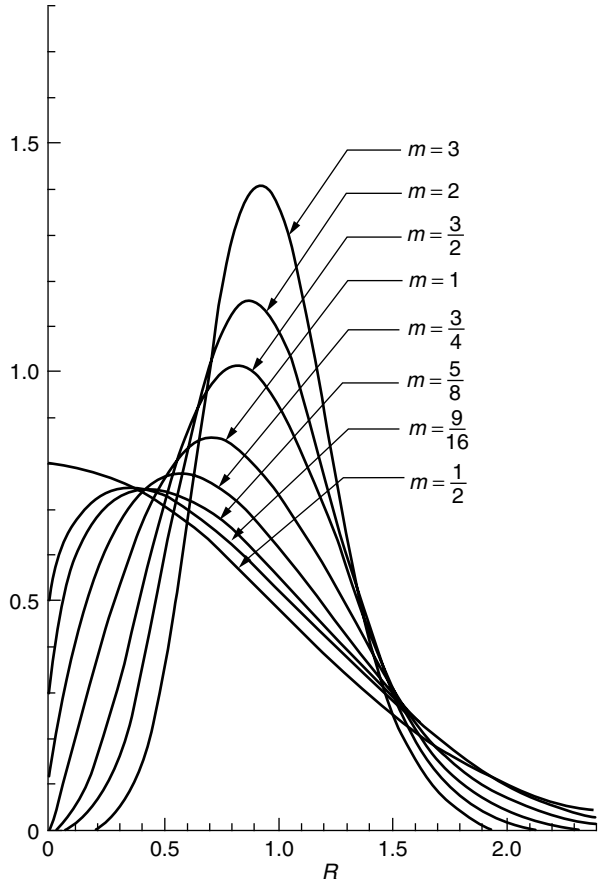


Figure 7. The PDF for the Nakagami-m distribution, shown with $\Omega = 1$, where m is the fading figure (reprinted from Ref. 17).

specifies the ratio of the deterministic signal power and the variance of the multipath:

$$K = 10 \log_{10} \left(\frac{s^2}{2\sigma^2} \right) \text{ (dB)} \tag{18}$$

As $s \to 0$ ($K \to -\infty$), the power of the dominant path diminishes, and the Ricean PDF converges to the Rayleigh PDF. Examples of Ricean fading and other LoS channels include airplane to ground communication links and microwave radio channels [1].

As an alternative to modeling the Rayleigh fading as a complex Gaussian process, one can instead approximate the channel by summing a set of complex sinusoids as in (4). The number of sinusoids in the set must be large enough so that the PDF of the resulting envelope provides an accurate approximation to the Rayleigh PDF. The Jakes model is a popular simulation method based on this principle [10]. The signal generated by the model is

$$c(t) = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} e^{j(\omega_d t \cos \alpha_n + \phi_n)} \tag{19}$$

where N is the total number of plane waves arriving at uniformly spaced angles α_n as shown in Fig. 8. The $c(t)$ can be further represented as

$$c(t) = \frac{E_0}{\sqrt{2N_0 + 1}} (c_1(t) + j c_Q(t))$$

$$c_I(t) = 2 \sum_{n=1}^{N_0} \cos \phi_n \cos \omega_n t + \sqrt{2} \cos \phi_N \cos \omega_m t$$

$$c_Q(t) = 2 \sum_{n=1}^{N_0} \sin \phi_n \cos \omega_n t + \sqrt{2} \sin \phi_N \cos \omega_m t$$

where $N_0 = \frac{1}{2}[(N/2) - 1)$, $\omega_m = 2\pi f_{dm}$, and $\omega_n = \omega_m \cos(2\pi n/N)$. In the Jakes model, the parameter $N_0 + 1$ is often referred as the *number of oscillators* and N is termed the *number of scatterers*. The Jakes model with as few as nine oscillators ($N_0 = 8$, and $N = 34$) closely approximates the Rayleigh fading distribution (14,15).

When a multipath fading channel with impulse response (13) needs to be modeled, the Jakes model can be extended to produce several uncorrelated fading components using the same set of oscillators [10]. The autocorrelation function and the Doppler spectrum of the signals generated by the Jakes model are characterized by Eqs. (11) and (12), respectively, and are shown in Fig. 6. This statistical characterization is appropriate for many channels where the reflectors are distributed uniformly around the mobile (*isotropic* scattering). Several other approaches to simulating Rayleigh fading channels based on *Clarke and Gans fading models* are described by Rappaport [3]. Moreover, *physical models* are useful [e.g., when the variation of amplitudes, frequencies, and phases in (4) is important to model as in *long-range fading prediction*] and *autoregressive* (*Gauss–Markov*) models are often utilized to approximate fading statistics in fading estimation algorithms since they result in rational spectral characterization [12,13].

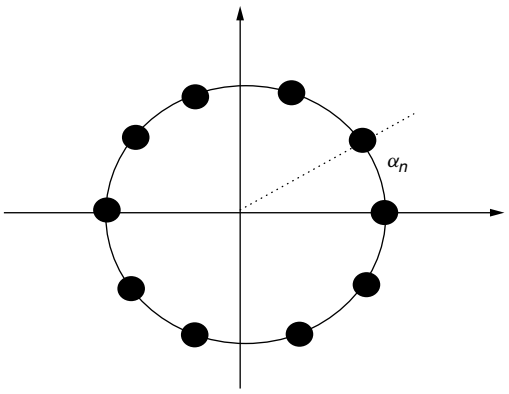


Figure 8. The Jakes model with $N = 10$ scatterers ($N_0 = 2$, 3 oscillators).

5. DIVERSITY TECHNIQUES AND PERFORMANCE ANALYSIS

5.1. Performance Analysis for Flat Fading Channels

Fading channels undergo dramatic changes in received power due to multipath and Doppler effects. When communication signals are transmitted over these channels, the bit error rate (BER) varies as a function of the signal-to-noise ratio (SNR) and is significantly degraded relative to the BER for the additive white Gaussian noise (AWGN) channel with the same average SNR. Consider the following example of transmission of binary phase-shift-keyed (BPSK) signal over the flat Rayleigh fading channel. At the output of the matched filter and sampler at the bit rate $1/T_b$ (where T_b is the bit interval), the complex envelope of the received signal is

$$r_k = c_k b_k + z_k \tag{20}$$

where c_k are the samples of the fading signal $c(t)$, b_k is the i.i.d. information sequence that takes on values $\{+1, -1\}$, and z_k is the i.i.d. complex white Gaussian noise sequence with the variance $\frac{1}{2}E[|z_k|^2] = N_0$. Since the channel is assumed to be stationary, we omit the subscript k in subsequent derivations. The equivalent passband energy is normalized as $E_b = \frac{1}{2}$, so the instantaneous SNR at the receiver is given by

$$\gamma = \frac{|c|^2}{2N_0} \tag{21}$$

Assume without loss of generality that the average power of the fading signal $E[|c_k|^2] = 1$. Then the average SNR per bit

$$\Gamma = \frac{1}{2N_0} \tag{22}$$

Suppose coherent detection of the BPSK signal is performed, and signal phase is estimated perfectly at the receiver. (This assumption is difficult to satisfy in practice for rapidly varying fading channels [12,13], so the analysis below represents a lower bound on the achievable performance.) The BER for each value of the instantaneous

SNR (21) can be computed from the BER expression for the AWGN channel [1]:

$$\text{BER}\ (\gamma) = Q[(2\gamma)^{1/2}] \tag{23}$$

where the Q function $Q(x) = 1/(2\pi)^{1/2} \int_{-\infty}^{x} \exp(-y^2/2)\, dy$. To evaluate average BER for the Rayleigh fading channel, the BER (23) has to be averaged over the distribution of the instantaneous SNR:

$$\text{BER} = \int_0^{\infty} Q[(2\gamma)^{1/2}] p(\gamma)\, d\gamma \tag{24}$$

Since $|c|$ is *Rayleigh* distributed, the instantaneous SNR γ has a *chi-square* distribution with the PDF $p(\gamma) = 1/\Gamma \exp(-\gamma/\Gamma)$, $\gamma \geq 0$. When this density is substituted in (24), the resulting BER for binary BPSK over flat Rayleigh fading channel is

$$\text{BER} = \frac{1}{2}\left[1 - \left(\frac{\Gamma}{1+\Gamma}\right)^{1/2}\right] \tag{25}$$

The BER for other fading distributions (e.g., Ricean or Nakagami-m) can be obtained similarly by averaging the BER for AWGN (23) over the fading statistics as in (24). Figure 9 illustrates performance of BPSK over a Nakagami-m fading channel for different values of m. Observe that as m increases, the fading becomes less severe, and the BER approaches that of the AWGN channel. Rayleigh fading ($m = 1$) results in significant SNR loss relative to the nonfading channel. In fact, for large SNR, the BER in (25) behaves asymptotically as $\frac{1}{4}\Gamma$, whereas the BER decreases exponentially with SNR for AWGN. The error rates of other modulation methods [e.g., coherent and noncoherent frequency shift keying (FSK), differential PSK (DPSK)] also *decrease only inversely with SNR*, causing very large power consumption.

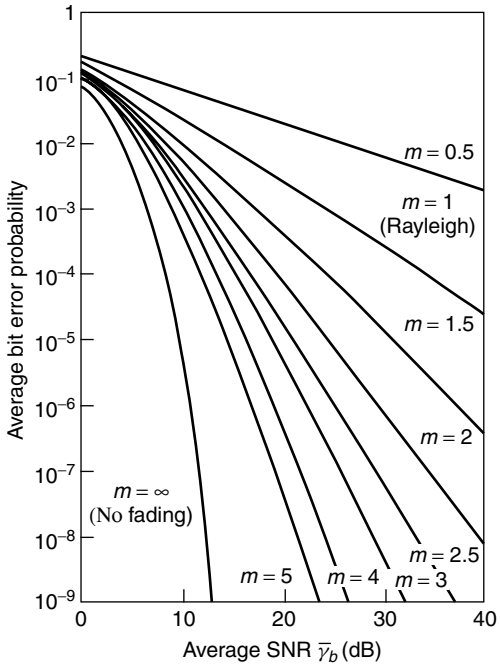


Figure 9. Average error probability for two-phase PSK with Nakagami fading (reprinted from Ref. 1).

5.2. Diversity Techniques

The poor performance of flat fading channels is due to the presence of deep fades in the received signal (Fig. 4) when the received SNR is much lower than the average SNR value. *Diversity techniques* help to combat fading by sending replicas of transmitted data over several uncorrelated (or partially correlated) fading channels. Since these channels are unlikely to go through a deep fade at the same time, higher average received SNR results when the outputs of the diversity branches are combined. Many diversity techniques are used in practice.

In *space*, or *antenna diversity* systems, several antenna elements are placed at the transmitter and/or receiver and separated sufficiently far apart to achieve desired degree of independence. (Note that the correlation function in space is analogous to that in time [10]. Typically, antenna separations of about $\lambda/2$ at the mobile station and several λ at the base station are required to achieve significant antenna decorrelation, where λ is the wavelength.)

Time diversity relies on transmitting the same information in several time slots that are separated by the coherence time $(\Delta t)_c$. Time diversity is utilized in coded systems by interleaving the outputs of the encoder.

Frequency diversity can be employed when the transmitter bandwidth is larger than the coherence bandwidth of the channel, and several fully or partially uncorrelated fading components can be resolved. The number of such uncorrelated components is determined by the ratio $W/(\Delta f)_c$ [1]. In PSK or quadrature-amplitude-modulated (QAM) systems, the transmitter bandwidth is approximately $1/T$. Because of the narrow transmission bandwidth, these channels are often frequency-nonselective. When frequency selectivity is present, it usually causes ISI since the multipath delay is significant relative to the symbol interval. Thus, *equalizers* are used to mitigate the ISI and to obtain the diversity benefit [1,12–14]. On the other hand, *direct-sequence spread-spectrum* (DSSS) systems employ waveforms with the transmission bandwidth that is much larger than the symbol rate $1/T$, and thus enjoy significant frequency diversity. DSSS signals are designed to achieve approximate orthogonality of multipath-induced components, thus eliminating the need for equalizers. Frequency diversity combining in these systems is achieved using a *RAKE correlator* [1].

In addition to the diversity techniques mentioned above, *angle of arrival* and *polarization* diversity are utilized in practice. Different diversity methods are often combined to maximize diversity gain at the receiver. To illustrate the effect of diversity on the performance of communication systems in fading channels, consider the following simple example. Suppose the transmitted BPSK symbol b_k is sent over L independent Rayleigh fading channels (we suppress the time index k below). The received equivalent lowpass samples are

$$r^i = c^i b + z^i, \qquad i = 1, \ldots, L \tag{26}$$

where c^i are i.i.d. complex Gaussian random variables with variances $E[|c^i|^2] = 1/L$ [this scaling allows us to compare performance directly with a system without diversity in (20)], b is the BPSK symbol as in Eq. (20), and z^i are i.i.d. complex Gaussian noise samples with $\frac{1}{2}E[|z_k|^2] = N_0$. Thus, the average SNR *per channel* (*diversity branch*) is $\Gamma_c = \Gamma/L$, where Γ is the SNR per bit.

There are several options for combining L diversity branches. For example, the branch with the highest instantaneous SNR can be chosen resulting in *selection diversity*. Alternatively, *equal gain combining* is a technique where all branches are weighted equally and cophased [3]. The maximum diversity benefit is obtained using *maximum ratio combining* (MRC) (this is also the most complex method). In MRC, the outputs r^i are weighted by the corresponding channel gains, cophased, and summed producing the decision variable (the input to the BPSK threshold detector):

$$U = (c^1) * r^1 + (c^2) * r^2 + \cdots + (c^L) * r^L$$

Note that if a signal undergoes a deep fade, it carries weaker weight than a stronger signal with higher instantaneous power. It can be shown that for large SNR, the BER for this method is approximately [1]

$$\text{BER} \approx \left(\frac{1}{(4\Gamma_c)}\right)^L \binom{2L-1}{L} \tag{27}$$

This BER of MRC for BPSK is illustrated in Fig. 10 for different values of L (binary PSK curve). The figure also shows performance of two less complex combining methods used with orthogonal FSK (*square-law combining*) and DPSK. The latter two techniques are noncoherent (i.e., do not require amplitude or phase estimation). From Eq. (27) and Fig. 10 we observe that the *BER for all methods decreases inversely with the Lth power of the SNR*. Thus, diversity significantly reduces power consumption in fading channels.

In addition to diversity, *adaptive transmission* is an effective tool in overcoming the effects of fading. The idea is to adjust the transmitted signal (power, rate, etc.) to fading conditions to optimize average power and bandwidth requirements. Many advances in the area of communication over fading channels, and additional sources are cited in Refs. 12 and 13.

6. SUMMARY

The fading signal was characterized in terms of large-scale and small-scale fading. The large-scale fading models that describe the average and statistical variation of the received signal power were presented. The small-scale fading channel was characterized statistically in terms of its time-variant impulse response. Time and frequency-domain interpretation was provided to describe signal dispersion and fading rapidity, and it was shown how the transmitted signal affects the choice of the fading channel model. Several statistical models of fading channels were presented, and simulation techniques were discussed. Finally, performance limitations of fading channels were revealed, and fading mitigation methods using diversity techniques were reviewed.

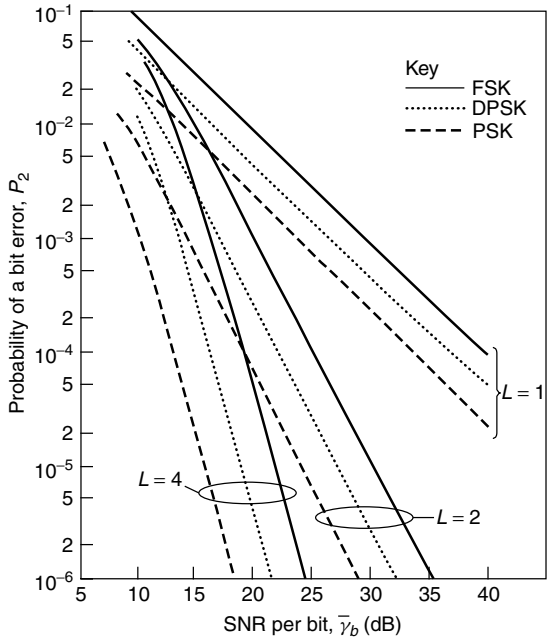


Figure 10. Performance of binary signals with diversity (reprinted from Ref. 1).

Acknowledgment

The author is grateful to Shengquan Hu, Hans Hallen, Tung-Sheng Yang, Ming Lei, Jan-Eric Berg, and Henrik Asplund for their assistance and helpful comments. This work was partially supported by NSF grant CCR-9815002 and ARO grant DAA19–01-1-0638.

BIOGRAPHY

Alexandra Duel-Hallen received the B.S. degree in Mathematics from Case Western Reserve University in 1982, the M.S. degree in Computer, Information and Control Engineering from the University of Michigan in 1983, and a Ph.D. in Electrical Engineering from Cornell University in 1987. During 1987–1990 she was a Visiting Assistant Professor at the School of Electrical Engineering, Cornell University, Ithaca, New York. In 1990–1992, she was with the Mathematical Sciences Research Center, AT&T Bell Laboratories, Murray Hill, New Jersey. She is an Associate Professor at the Department of Electrical and Computer Engineering at North Carolina State University, Raleigh, North Carolina, which she joined in January 1993. From 1990 to 1996, Dr. Duel-Hallen was Editor for Communication Theory for the *IEEE Transactions on Communications*. During 2000–2002, she has served as Guest Editor of two Special Issues on Multiuser Detection for the *IEEE Journal on Selected Areas in Communications*. Dr. Duel-Hallen's current research interests are in wireless and multiuser communications. Her 1993 paper was selected for the IEEE Communications Society 50th Anniversary Journal Collection as one of 41 key papers in physical and link layer areas, 1952–2002.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, 2001.
2. B. Sklar, Rayleigh fading channels in mobile digital communication systems, Part 1: Characterization, *IEEE Commun. Mag.* **35**(7): 90–100 (July 1997).
3. T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed., Prentice-Hall, 2002.
4. H. L. Bertoni, *Radio Propagation for Modern Wireless Systems*, Prentice-Hall, 2000.
5. W. C. Y. Lee, *Mobile Communications Engineering: Theory and Applications*, 2nd ed., McGraw-Hill Telecommunications, 1997.
6. R. Steele, *Mobile Radio Communications*, Pentech Press, 1992.
7. G. L. Stuber, *Principles of Mobile Communication*, Kluwer, 2001.
8. P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* **11**: 360–393 (1963).
9. S. Stein, Fading channel issues in system engineering, *IEEE J. Select. Areas Commun.* **5**(2): 68–69 (Feb. 1987).
10. W. C. Jakes, *Microwave Mobile Communications*, Wiley, New York, 1974.
11. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, 1981.
12. E. Biglieri, J. Proakis, and S. Shamai (Shitz), Fading channels: information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* **44**(6): 2619–2692 (Oct. 1998).
13. *IEEE Signal Process. Mag.* (Special Issue on Advances in Wireless and Mobile Communications; G. B. Giannakis, Guest Editor) **17**(3): (May 2000).
14. B. Sklar, Rayleigh fading channels in mobile digital communication systems, Part 2: Mitigation, *IEEE Commun. Mag.* **35**(7): 102–109 (July 1997).
15. S. Y. Seidel et al., Path loss, scattering, and multipath delay statistics in four European cities for digital cellular and microcellular radiotelephone, *IEEE Trans. Vehic. Technol.* **40**(4): 721–730 (Nov. 1991).
16. T. S. Rappaport, S. Y. Seidel, and R. Singh, 900 MHz multipath propagation measurements for U.S. digital cellular radiotelephone, *IEEE Trans. Vehic. Technol.* 132–139 (May 1990).
17. Y. Miyagaki, N. Morinaga, and T. Namekawa, Error probability characteristics for CPFSK signal through *m*-distributed fading channel, *IEEE Trans. Commun.* **COM-26**: 88–100 (Jan. 1978).

FEEDBACK SHIFT REGISTER SEQUENCES

HONG-YEOP SONG
Yonsei University
Seoul, South Korea

1. INTRODUCTION

Binary random sequences are useful in many areas of engineering and science. Well-known applications are digital ranging and navigation systems because of the sharp peak in their autocorrelation functions [1], spread-spectrum modulation and synchronization using some of their correlation properties [2–6], and stream ciphers, in which the message bits are exclusive-ORed with key streams that must be as random as possible [7,8]. Several randomness tests for binary sequences have been proposed in practice [8], but no universal consensus has been made yet with regard to the true randomness of binary sequences [9].

Random binary sequences can be obtained in theory by flipping an unbiased coin successively, but this is hardly possible in most practical situations. In addition, not only must the random sequence itself be produced at some time or location but also its exact replica must also be produced at remote (in physical distance or time) locations in spread-spectrum modems. This forces us to consider the sequences that appear random but can be easily reproduced with a set of simple rules or keys. We call these *pseudorandom* or *pseudonoise* (PN) sequences. It has been known and used for many years that *feedback shift registers* (FSRs) are most useful in designing and generating such PN sequences. This is due to their simplicity of defining rules and their capability of generating sequences with much longer periods [10]. Approaches using FSR sequences solve the following two basic problems in most applications: cryptographic secrecy and ease of generating the same copy over and over.

One of the basic assumptions in conventional cryptography is that the secrecy of a system does not depend on the secrecy of how it functions but rather on the secrecy of the key, which is usually kept secret [11]. Feedback shift registers are most suitable in this situation because we do not have to keep all the terms of the sequences secret. Even though its connection is revealed and all the functionality of the system is known to the public, any unintended observer will have an hard time of locating the exact phase of the sequence in order to break the system, provided that the initial condition is kept secret. The current CDMA modem (which is used in the successful second and third generation mobile telephone systems) depends heavily on this property for its privacy [12].

One previous difficulty in employing spread-spectrum communication systems was on the effort of reproducing at the receiver the exact replica of PN sequences that were used at the transmitter [2]. Store-and-replay memory wheels to be distributed initially were once proposed, and the use of a secret and safe third channel to send the sequence to the receiver was also proposed. The theory and practice of FSR sequences have now been well developed so that by simply agreeing on the initial condition and/or connection method (which requires much smaller memory space or computing time), both ends of communicators can easily share the exact same copy.

In Section 2, the very basics of feedback shift registers (FSRs) and their operations are described, following the style of Golomb [10]. We will concentrate only on some basic terminologies, state transition diagrams, truth tables, cycle decompositions, and the like. In fact, the detailed proofs of claims and most of discussions and a lot more can be found in Golomb's study [10]. Section 3

covers mostly the *linear* FSRs. The linear FSR sequences have been studied in various mathematics literature under the term *linear recurring sequences*. Lidl and Niederreiter gave a comprehensive treatment on this subject [13]. Some other well-known textbooks on the theory of finite fields and linear recurring sequences are available [14–18]. In this article, we will discuss the condition for their output sequences to have maximum possible period. The maximum period sequences, which are known as m-sequences, are described in detail, including randomness properties. Two properties of m-sequences deserve special attention: m-sequences of period P have the two-level ideal autocorrelation, which is the *best* over all the balanced binary sequences of the same period, and they have the linear complexity of $\log_2(P)$, which is the *worst* (or the smallest) over the same set. Some related topics on these properties will also be discussed. To give some further details of the ideal two-level autocorrelation property, we describe a larger family of balanced binary sequences which come from, so called, *cyclic Hadamard difference sets*. An m-sequence can be regarded as the characteristic sequence of a cyclic Hadamard difference set of the Singer type. To give some better understanding of the linear complexity property, we describe the *Berlekamp–Massey algorithm* (BMA), which determines the shortest possible linear FSR that generates a given sequence.

In Section 4, we describe some special cases of FSRs (including nonlinear FSRs) with disjoint cycles in their state diagrams. Branchless condition and balanced logic condition will be discussed. De Bruijn sequences will briefly be described. Finally, four-stage FSRs are analyzed in detail for a complete example. Section 5 gives some concluding remarks. We will restrict our treatment to only the sequences over a binary alphabet {0, 1} in this article.

2. FEEDBACK SHIFT REGISTER SEQUENCES, TRUTH TABLES, AND STATE DIAGRAMS

The operation of an FSR can best be described by its state transition diagram. Its output at one time instant depends only on the previous state. Figure 1 shows a generic block diagram of an FSR (linear or non-linear) with L stages. At every clock, the content of a stage is shifted to the left, and the connection logic or the boolean function f calculates a new value x_k

$$x_k = f(x_{k-L}, x_{k-L+1}, \ldots, x_{k-1}) \tag{1}$$

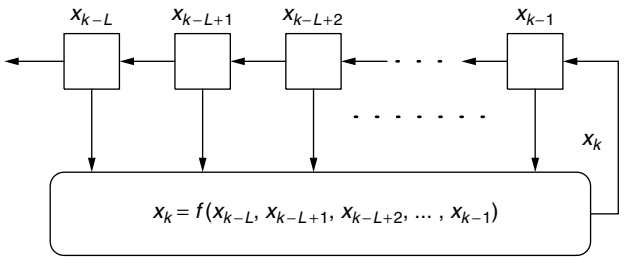


Figure 1. An L-stage FSR with a feedback (Boolean logic) function f.

to be fed back to the rightmost stage. The leftmost stage gives an output sequence in which the first L terms are in fact given as an initial condition.

A *state* of this FSR at one instant k can be defined simply as the vector $(x_{k-L}, x_{k-L+1}, \ldots, x_{k-1})$, and this will be changed into $(x_{k-L+1}, x_{k-L+2}, \ldots, x_k)$ at the next instant. An FSR is called *linear* if the connection logic is a linear function on $x_{k-L}, x_{k-L+1}, \ldots, x_{k-1}$, that is, if it is of the form

$$\begin{aligned} x_k &= f(x_{k-L}, x_{k-L+1}, \ldots, x_{k-1}) \\ &= c_L x_{k-L} \oplus c_{L-1} x_{k-L+1} \oplus \cdots \oplus c_1 x_{k-1} \end{aligned} \tag{2}$$

for some fixed constants $c_1, c_2, \ldots, c_L$. Otherwise, it is called *nonlinear*. Here, the values of x_i are either 0 or 1, and hence the sequence is said to be over a *binary alphabet*, which is usually denoted as F_2, and $c_i \in F_2$ for all i. The operation $\oplus$ can easily be implemented as an *exclusive-OR* operation and $c_i x_{k-i}$ as an AND operation both using digital logic gates. In the remaining discussion, we will simply use addition and multiplication (mod 2), respectively, for these operations. Over this binary alphabet, therefore, one can add and multiply two elements, and the subtraction is the same as addition. There is only one nonzero element (which is 1), and the division by 1 is the same as the multiplication by 1.

Another method of describing an FSR is to use its truth table, in which all the 2^L states are listed on the left column and the next bits calculated from the connection logic f are listed on the right. The state change can easily be illustrated by the state diagram in which every state is a node and an arrow indicates the beginning (predecessor) and ending (successor) states. Figures 2 and 3 show examples of three-stage FSRs, including their truth tables, and state diagrams. Note that there are exactly 2^L states in total, and every state has at most two predecessors and exactly one successor. Note also that there are 2^{2^L} different L-stage FSRs, corresponding to the number of choices for the next bit column in the truth table. Finally, note that Fig. 3 has two disjoint cycles while Fig. 2 has branches and some absorbing states. Therefore, the FSR in Fig. 2 eventually will output the all-zero sequence, while that in Fig. 3 will output a sequence of period 7 unless its initial condition is 000.

In order to investigate this situation more closely, observe that any state has a unique successor, but up to two predecessors. From this, we observe that a branch occurs at a state that has two predecessors, and this

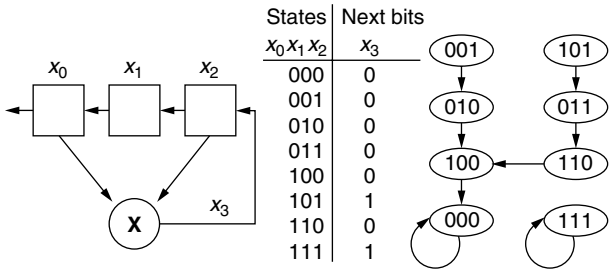


Figure 2. A three-stage nonlinear FSR with a feedback function $x_k = x_{k-1}x_{k-3}$, including its truth table and state diagram.

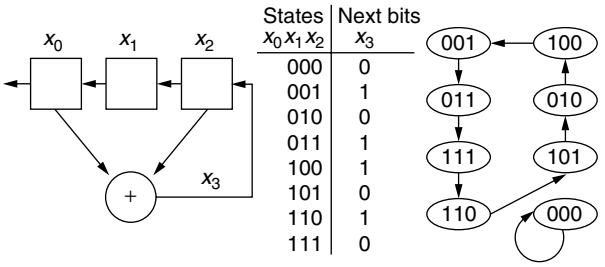


Figure 3. A three-stage linear FSR with a feedback function $x_k = x_{k-1} \oplus x_{k-3}$, including its truth table and state diagram.

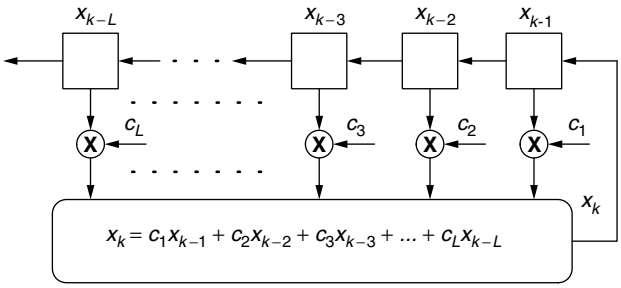


Figure 4. An L-stage LFSR with connection coefficients $c_1, c_2, \ldots, c_L$. Note that $c_L = 1$ for LFSR to have genuinely L stages.

happens in a state diagram if and only if there is a state that has no predecessor. A branch in a state diagram should be avoided since it will either seriously reduce the period of the output sequences or result in an ambiguous initial behavior. The necessary and sufficient condition for a branchless state diagram is, therefore, that no two states have the same successor. Consider any pair of states $(a_0, a_1, \ldots, a_{L-1})$ and $(b_0, b_1, \ldots, b_{L-1})$. If they are different in any other position except for the first, their successors $(a_1, a_2, \ldots, a_L)$ and $(b_1, b_2, \ldots, b_L)$ will still be different, because all the components except for the first will be shifted to the left and the difference remains. The remaining case is the pair of the form $(a_0, a_1, \ldots, a_{L-1})$ and $(a_0', a_1, \ldots, a_{L-1})$, where a_0' represents the complement of a_0. Their successors will be $(a_1, a_2, \ldots, a_{L-1}, f(a_0, a_1, \ldots, a_{L-1}))$ and $(a_1, a_2, \ldots, a_{L-1}, f(a_0', a_1, \ldots, a_{L-1}))$. For these two states to be distinct, the rightmost component must be different:

$$\begin{aligned} f(a_0', a_1, \ldots, a_{L-1}) &= f(a_0, a_1, \ldots, a_{L-1}) \oplus 1 \\ &= f'(a_0, a_1, \ldots, a_{L-1}) \end{aligned}$$

Let $g(a_1, a_2, \ldots, a_{L-1})$ be a boolean function on $L-1$ variables such that

$$f(0, a_1, \ldots, a_{L-1}) = g(a_1, a_2, \ldots, a_{L-1})$$

Then, the relation shown above can be written as

$$f(a_0, a_1, \ldots, a_{L-1}) = a_0 \oplus g(a_1, a_2, \ldots, a_{L-1}) \qquad (3)$$

This is called the *branchless condition* for an L-stage FSR. For FSRs with the branchless condition, the corresponding truth table has only 2^{L-1} independent entries, and the top half of the truth table must be the complement of the bottom half. This condition is automatically satisfied with *linear* FSRs, which are the topic of the next section.

3. LINEAR FEEDBACK SHIFT REGISTERS AND *m*-SEQUENCES

3.1. Basics

The output sequence $\{s(k) | k = 0, 1, 2, \ldots\}$ of a linear feedback shift register (LFSR) with L stages as shown in Fig. 4 satisfies a linear recursion of degree L. Given $L+1$ constants, $c_1, c_2, \ldots, c_L, b$, and the initial condition $s(0), s(1), \ldots, s(L-1)$, the terms $s(k)$ for $k \geq L$ satisfy

$$s(k) = c_1 s(k-1) + c_2 s(k-2) + \cdots + c_L s(k-L) + b \qquad (4)$$

or equivalently

$$s(k) + c_1 s(k-1) + c_2 s(k-2) + \cdots + c_L s(k-L) + b = 0 \qquad (5)$$

The recursion is called *homogeneous* linear if $b = 0$ and *inhomogeneous* linear if $b \neq 0$. We will assume that $b = 0$ in this section and consider mainly the homogeneous linear recursion.

The *characteristic polynomial* of the homogeneous linear recursion in Eq. (4) or (5) is defined as

$$f(x) = 1 + c_1 x + c_2 x^2 + \cdots + c_L x^L \qquad (6)$$

This contains all the connection coefficients, and will completely determine the operation of the LFSR provided that the initial condition is specified. Note that $c_L \neq 0$ in order for this LFSR to be genuinely with L stages. Otherwise, the recursion becomes of degree less than L, and the LFSR with less than L stages can also be used to implement the recursion.

Note that it is also the characteristic polynomial of the sequence satisfying this recursion. A given sequence may satisfy many other recursions that differ from each other. The *minimal polynomial* of a given sequence is defined as the minmum degree characteristic polynomial of the sequence. It is irreducible, it becomes unique if it is restricted to be monic, and it divides all the characteristic polynomials of the sequence. We will return to this and more later when we discuss the linear complexity of sequences.

In the state diagram of any LFSR, every state will have a unique successor and a unique predecessor, as stated at the end of the previous section. This forces the diagram to be (possibly several) disjoint cycles of states. In Fig. 3, the state diagram has two disjoint cycles, one with length 7, and the other with length 1. From this, we can easily see that the output sequence of an LFSR is *periodic*, and the period is the length of the cycle that the initial state (or the initial condition) belongs to. We can conclude, therefore, that *the output sequence of a LFSR is periodic with some period P that depends on both the initial condition and the characteristic polynomial*.

One special initial condition is the all-zero state, and this state will always form a cycle of length 1 *for any LFSR*. For any other initial state, the cycle will have a certain length ≥ 1, and this length is the period of the output sequence with the given initial condition. Certainly, the cycle with the maximum possible length must contain every not-all-zero state exactly once, and the output sequence in this case is known as the *maximal length linear feedback shift register sequence*, or the *m*-sequence, in short. Sometimes, PN sequences are used instead of *m*-sequences and vice versa, but we will make a clear distinction between these two terms. PN sequences refer to (general) pseudonoise sequences that possess some or various randomness properties, and *m*-sequences are a specific and very special example of PN sequences. For an L-stage LFSR, this gives the period $2^L - 1$, and Fig. 3 shows an example of an *m*-sequence of period 7. In fact, it shows seven different *phases* of this *m*-sequence depending on the seven initial conditions. Note also that the history of any stage is the same *m*-sequence in different phase.

The operation of an LFSR is largely determined by its characteristic polynomial. It is a polynomial of degree L over the binary alphabet F_2. How it factors over F_2 is closely related to the properties of the output sequence. In order to discuss the relation between the characteristic polynomials and the corresponding output sequences, we define some relations between sequences of the same period.

Let $\{s(k)\}$ and $\{t(k)\}$ be *arbitrary* binary sequences of period P. Then we have the following three important relations between these two sequences:

1. One is said to be a *cyclic shift* of the other if there is a constant integer τ such that $t(k - \tau) = s(k)$ for all k. Otherwise, two sequences are said to be cyclically distinct. When one is a cyclic shift of the other with $\tau \neq 0$, they are said to be in different *phases*. Therefore, there are P distinct phases of $\{s(k)\}$ that are all cyclically equivalent.
2. One is a *complement* of the other if $t(k) = s(k) + 1$ for all k.
3. Finally, one is a *decimation* (or d decimation) of the other if there are constants d and τ such that $t(k - \tau) = s(dk)$ for all k. If d is not relatively prime to the period P, then the d decimation will result in a sequence with shorter period, which is P/g, where g is the GCD of P and d.

If some combination of these three relations applies to $\{s(k)\}$ and $\{t(k)\}$, then they are called *equivalent*. Equivalent sequences share lots of common properties, and they are essentially the same sequences even if they look very different.

The necessary and sufficient condition for an L-stage LFSR to produce an *m*-sequence of period $2^L - 1$ is that the characteristic polynomial of degree L is *primitive* over F_2. This means simply that $f(x)$ is irreducible and $f(x)$ divides $x^{2^L-1} - 1$, but $f(x)$ does not divide $x^j - 1$ for all j from 1 to $2^L - 2$. The elementary theory of finite fields (or Galois fields) deals much more with these primitive polynomials, which we will not discuss in detail here. Instead, we refer the reader to some references for further exploration in theory [13–18]. See the article by Hansen and Mullen [19] for a list of primitive polynomials of degree up to a few hundreds, which will generally suffice for any practical application. There are $\phi(2^L - 1)/L$ primitive polynomials of degree L over F_2. Some primitive polynomials are shown in Table 1 for L up to 10. Here, $\phi(n)$ is the Euler ϕ function, and it counts the number of integers from 1 to n that are relatively prime to n.

In order to describe some properties of the output sequences of LFSRs, we include all the four-stage LFSRs: block diagrams, characteristic polynomials, truth tables, state transition diagrams, and the output sequences in Figs. 5–8. Note that there are 16 linear logics for four-stage LFSRs, and the condition $c_L = 1$ reduces it into half.

Figure 5 shows two LFSRs that generate *m*-sequences of period 15. The detailed properties of *m*-sequences will be described in Section 3.2. Here, we note that two characteristic polynomials $f_1(x) = x^4 + x^3 + 1$ and $f_2(x) = x^4 + x + 1$ are reciprocal to each other. That is, the coefficients are 11001 and 10011. This gives two *m*-sequences that are reciprocal to each other. In other words, one is a 14 decimation of the other. Little arrows with a dot under the output sequences indicate this fact. Note that the roots of $f_2(x)$ are the 14th power of those of $f_1(x)$. In general, if $f(x)$ and $g(x)$ are primitive of the same degree and the roots of one polynomial are dth power of the other, then the *m*-sequence from one polynomial is a d decimation of that from the other. The truth table shows only the top half since its bottom half is the complement of what is shown here, as a result of the branchless condition.

Figure 6 shows LFSRs with the characteristic polynomials that factor into smaller-degree polynomials. Note that $f_3(x) = x^4 + x^3 + x^2 + 1 = (x + 1)(x^3 + x + 1)$ and $f_4(x) = x^4 + x^2 + x + 1 = (x + 1)(x^3 + x^2 + 1)$. Since both of the degree 3 polynomials in their factorizations are primitive, the LFSR generates *m*-sequences of period 7, which could have been generated by a three-stage LFSR. Two characteristic polynomials are reciprocal, and the output

Table 1. The Number of Primitive Irreducible Polynomials of Degree L and Some Examples[a]

Degree L	$\phi(2^L - 1)/L$	Primitive Polynomial
1	1	11
2	1	111
3	2	1011
4	2	10011
5	6	100101
6	6	1000011
7	18	10000011
8	16	100011101
9	48	1000010001
10	60	10000001001

[a]The binary vector 1011 for $L = 3$ represents either $x^3 + x + 1$ or $1 + x^2 + x^3$.

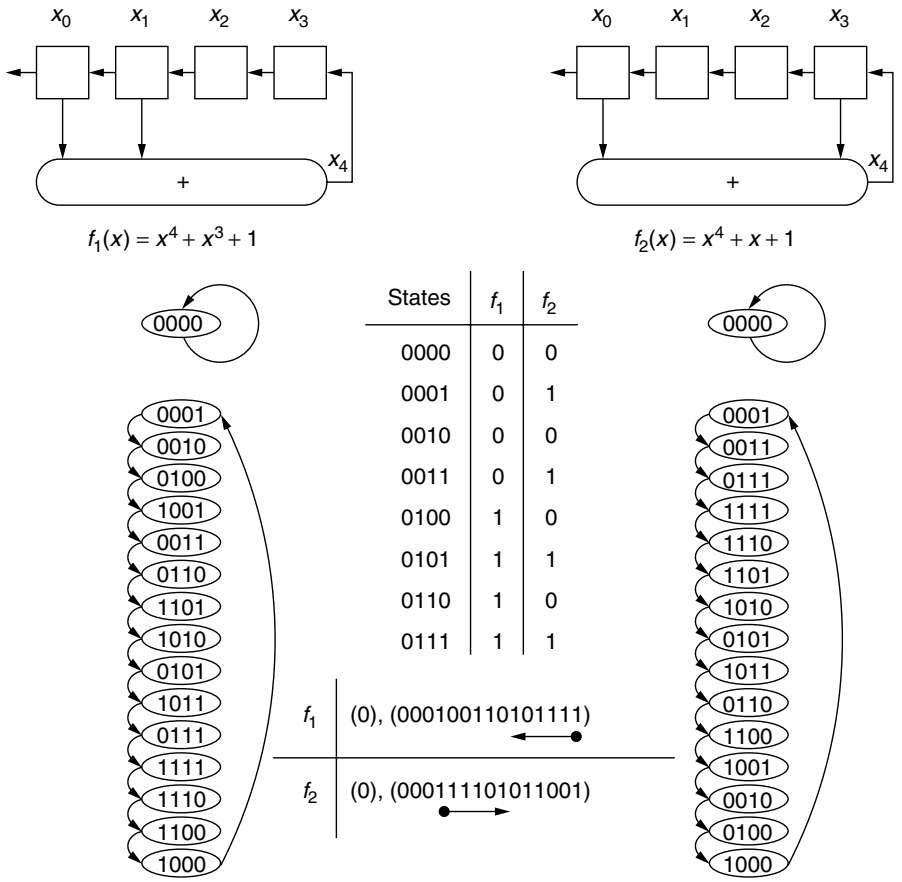


Figure 5. Block diagrams, state diagrams, truth tables, and output sequences of four-stage LFSRs that generate m-sequences: $f_1(x) = x^4 + x^3 + 1$ and $f_2(x) = x^4 + x + 1$.

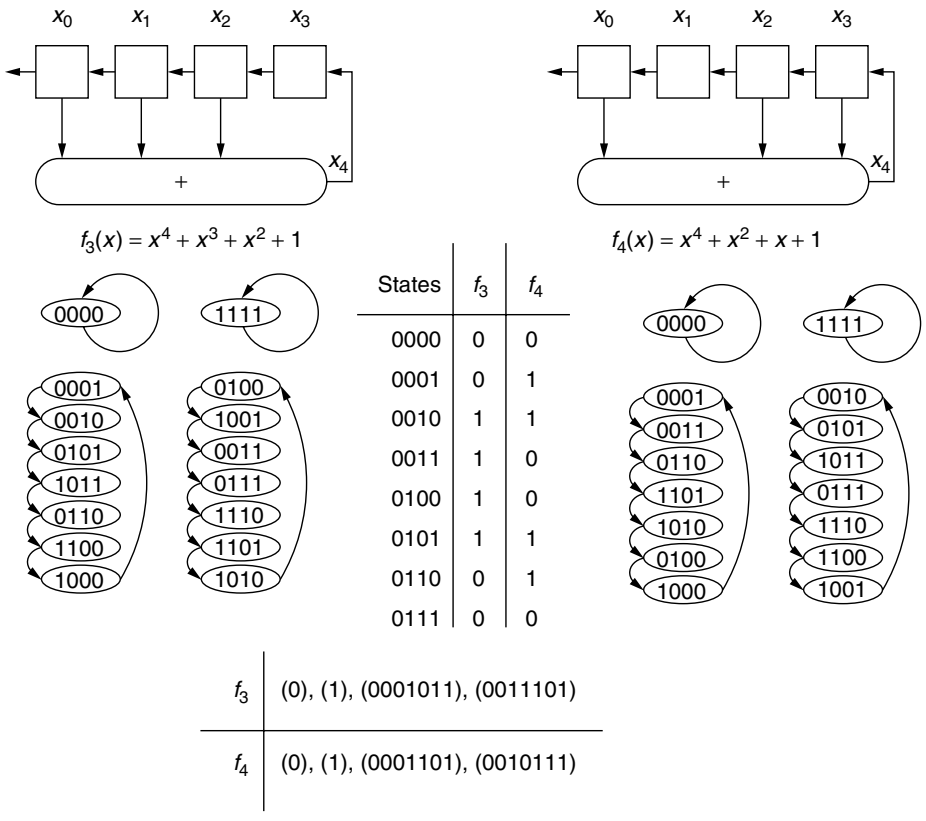
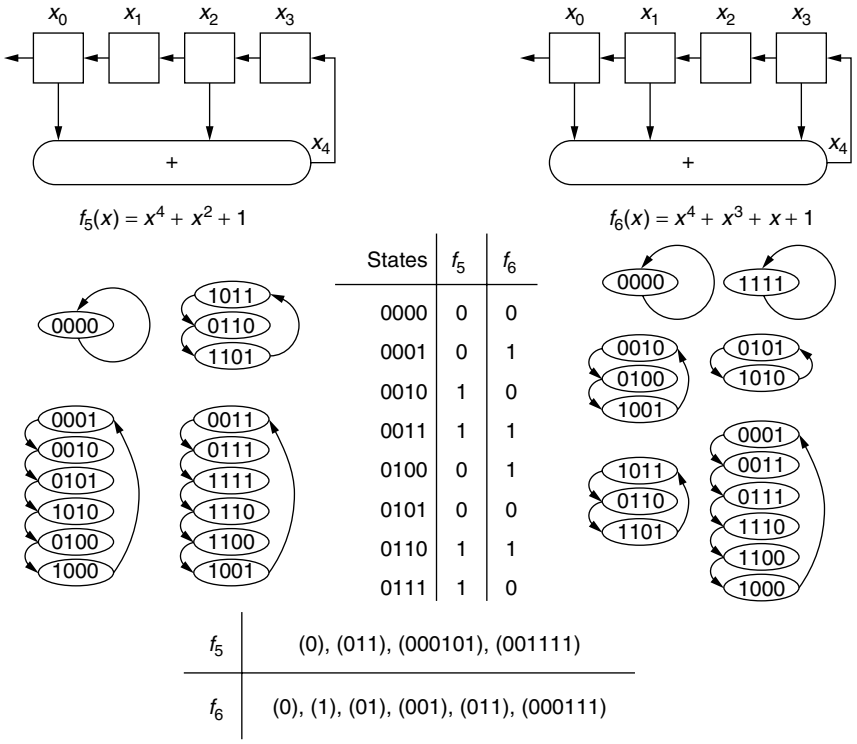
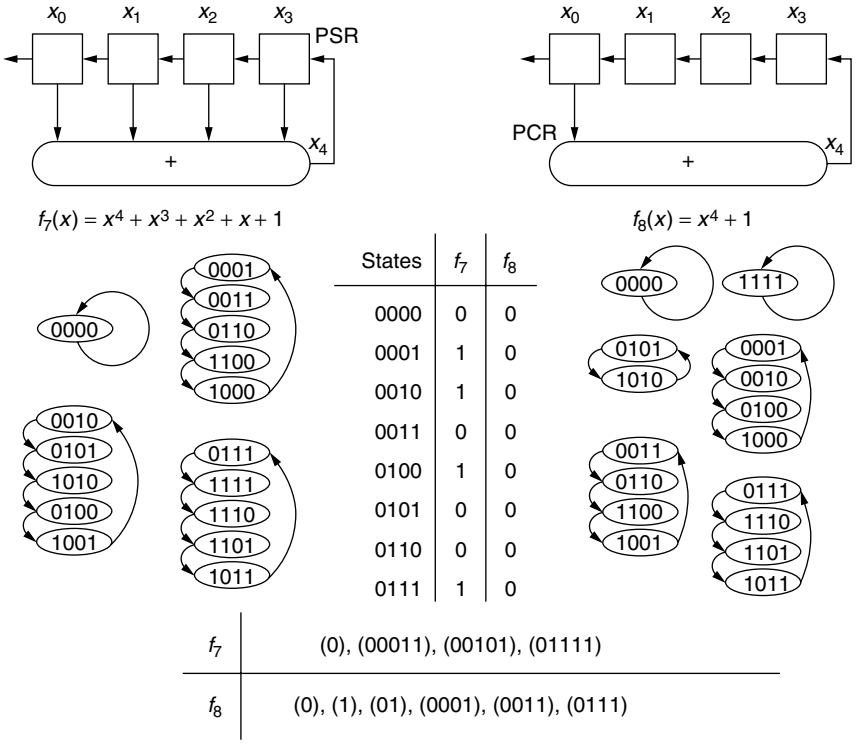


Figure 6. Block diagrams, state diagrams, truth tables, and output sequences of four-stage LFSRs with characteristic polynomials $f_3(x) = x^4 + x^3 + x^2 + 1$ and $f_4(x) = x^4 + x^2 + x + 1$.



f_5	(0), (011), (000101), (001111)
f_6	(0), (1), (01), (001), (011), (000111)

Figure 7. Block diagrams, state diagrams, truth tables, and output sequences of four-stage LFSRs with characteristic polynomials $f_5(x) = x^4 + x^2 + 1$ and $f_6(x) = x^4 + x^3 + x + 1$.



f_7	(0), (00011), (00101), (01111)
f_8	(0), (1), (01), (0001), (0011), (0111)

Figure 8. Block diagrams, state diagrams, truth tables, and output sequences of four-stage LFSRs with characteristic polynomials $f_7(x) = x^4 + x^3 + x^2 + x + 1$ (PSR) and $f_8(x) = x^4 + 1$ (PCR).

sequences are reciprocal. Figure 7 shows LFSRs with the self-reciprocal characteristic polynomials, and the output sequences are self-reciprocal also. This means that reading a sequence in the reverse direction gives a cyclically equivalent one to the original.

Figure 8 shows two special LFSRs: the pure summing register (PSR) and the pure cycling register (PCR). PSR has an irreducible but not primitive characteristic polynomial f_7. Observe that all the cycles except for the cycle containing (0000) have the same length, or the same period for its output sequences except for the all-zero sequence. This happens because the characteristic polynomial f_7 is irreducible. An irreducible polynomial, therefore, corresponds to a unique period, and it is called

the period of the irreducible polynomial. A primitive polynomial of degree L is simply an irreducible polynomial with period $2^L - 1$. Possible periods of a given irreducible polynomial of degree L are the factors of the integer $2^L - 1$, which are not of the form $2^j - 1$ for $j < L$. When $2^L - 1$ is a prime, called a *Mersenne prime*, then every irreducible polynomial of degree $2^L - 1$ must be primitive.

Details on the property of PCR and some other properties of FSRs will be given at the end of Section 4.

3.2. Properties of *m*-Sequences

Now, we will describe some basic properties of m-sequences of period $2^L - 1$, mostly without proofs. The first three properties, namely, balance, run-distribution, and ideal autocorrelation are commonly known as "Golomb's postulates on random sequences" [10]. Most of the following properties can be easily checked for the examples shown in Figs. 3 and 5.

3.2.1. Balance Property. In one period of an m-sequence, the number of 1s and that of 0s are nearly the same. Since the period is an odd integer, they cannot be exactly the same, but differ by one. This is called the balance property. When a matrix of size $(2^L - 1) \times L$ is formed by listing all the states of the maximum length cycle, then the rightmost column is the m-sequence and it will contain 2^{L-1} ones and $2^{L-1} - 1$ zeros since the rows are permutations of all the vectors of length L except for the all-zero vector.

3.2.2. Run Distribution Property. A string of the same symbol of length l surrounded by different symbols at both ends is called a "run of length l." For example, a run of 1s of length 4 looks like $\ldots 0\underline{1111}0 \ldots$. The run distribution property of m-sequences refers to the fact that a shorter run appears more often than a longer run, and that the number of runs of 1s is the same as that of 0s. Specifically, it counts the number of runs of length l for $l \geq 1$ in one period as shown in Table 2. The span property of m-sequences implies this run distribution property.

3.2.3. Ideal Autocorrelation Property. A periodic unnormalized autocorrelation function $R(\tau)$ of a binary sequence $\{s(k)\}$ of period P is defined as

$$R(\tau) = \sum_{k=0}^{P-1} (-1)^{s(k)+s(k-\tau)}, \quad \tau = 0, 1, 2, \ldots,$$

Table 2. Run Distribution Property of *m*-Sequences of Period $2^L - 1$

Length	Number of Runs of 1s	Number of Runs of 0s
L	1	0
$L-1$	0	1
$L-2$	1	1
$L-3$	2	2
$L-4$	4	4
$\vdots$	$\vdots$	$\vdots$
2	2^{L-4}	2^{L-4}
1	2^{L-3}	2^{L-3}
Total	2^{L-3}	2^{L-3}

where $k - \tau$ is computed mod P. When binary phase-shift-keying is used to digitally modulate incoming bits, we are considering the incoming bits whose values are taken from the complex values $\{+1, -1\}$. The change in alphabet between $s_i \in \{0, 1\}$ and $t_i \in \{+1, -1\}$ is commonly performed by the relation $t_i = (-1)^{s_i}$. Then we have $R(\tau) = \sum_{k=0}^{P-1} t(k)t(k-\tau)$ for each τ, and this calculates *the number of agreements minus the number of disagreements* when one period of $\{s(k)\}$ is placed on top of its (cyclically) τ-shifted version. For any integer $L \geq 2$, and for any m-sequence $\{s(k)\}$ of period $P = 2^L - 1$, the ideal autocorrelation property of m-sequences refers to the following:

$$R(\tau) = \begin{cases} 2^L - 1, & \tau \equiv 0 \pmod{2^L - 1} \\ -1, & \tau \not\equiv 0 \pmod{2^L - 1} \end{cases} \tag{7}$$

The ideal two-level autocorrelation property of an m-sequence enables one to construct a Hadamard matrix of order 2^L of, so called, *cyclic* type. A Hadamard matrix of order n is an $n \times n$ matrix with entries only of ± 1 such that any two distinct rows are orthogonal to each other [20]. When the symbols of an m-sequence are mapped onto $\{\pm 1\}$ and all the cyclic shifts are arranged in a square matrix of order $(2^L - 1)$, the relation in (7) implies that the dot product of any two distinct rows is exactly -1 over the complex numbers. Therefore, adjoining a leftmost column of all $+1$s and a top row of all $+1$s will give a cyclic Hadamard matrix of order 2^L.

Cyclic-type Hadamard matrices can be constructed from a balanced binary sequence of period $P \equiv 3 \pmod 4$ that has the ideal two-level autocorrelation function. The m-sequences are one such class of sequences. Some other well-known balanced binary sequences with period $P \equiv 3 \pmod 4$ will be described later.

3.2.4. Span Property. If two vectors, $(s(i), s(i+1), \ldots, s(i+L-1))$ and $(s(j), s(j+1), \ldots, s(j+L-1))$, of length L are distinct whenever $i \neq j$, then the sequence $\{s(k)\}$ is said to have this property. The indices of terms are considered mod P. For an m-sequence of period P, in addition, all the not-all-zero vectors of length L appear exactly once on the windows of length L. This can easily be seen by observing that an m-sequence is the sequence of the rightmost bits of states in the maximum length cycle of the state diagram. Each window of length L can then easily be identified with a state in this state diagram.

If we insert an additional 0 right after the run of 0's of length $L - 1$, the sequence will have period 2^L and the span property becomes perfect in that every window of length L shows all the vectors of length L exactly once. This is an example of a *de Bruijn sequence* of order L. The above construction has been successfully adopted [12] for use in spread-spectrum modulations, for the values of $L = 14$ and $L = 41$. In general, many algorithms are currently known for de Bruijn sequences of period 2^L [21], and the modification described above can easily be implemented, as described in Section 4.

3.2.5. Constant-on-the-Coset Property. For any m-sequence of period $2^L - 1$, there are $2^L - 1$ cyclically

equivalent sequences corresponding to the $2^L - 1$ starting points. The term *constant-on-the-coset property* refers to the fact that there exists exactly one among all these such that it is fixed with 2 decimation. An m-sequence in this phase is said to be in the *characteristic phase*. Therefore, for the m-sequence $\{s(k)\}$ in the characteristic phase, the following relation is satisfied:

$$s(2k) = s(k), \quad \text{for all } k \tag{8}$$

This relation deserves some special attention. It implies that every term in the $2k$th position is the same as the one in the kth position. This gives a set (or several sets) of positions in which the corresponding terms are the same. For example, $\{1, 2, 4, 8, 16, \ldots\}$ is one such set so that all the terms indexed by any number in this set are the same. Since the sequence is periodic with period $2^L - 1$, the preceding set is a finite set and called a *cyclotomic coset* mod $2^L - 1$. Starting from 3 gives another such set, $\{3, 6, 12, 24, \ldots\}$, and so on. In general, the set of integers mod $2^L - 1$ can be decomposed into some number of disjoint cyclotomic cosets, and now the constant-on-the-coset property describes itself clearly.

3.2.6. Cycle-and-Add Property. When two distinct phases of an m-sequence are added term by term, a sequence of the same period appears and it is a different phase of the same m-sequence. In other words, for any given constants $\tau_1 \not\equiv \tau_2 \pmod{2^L - 1}$, there exists yet another constant τ_3 such that

$$s(k - \tau_1) + s(k - \tau_2) = s(k - \tau_3), \quad k = 0, 1, 2, \ldots \tag{9}$$

This is the cycle-and-add property of m-sequences. On the other hand, if a balanced binary sequence of period P satisfies the cycle-and-add property, then P must be of the form $2^L - 1$ for some integer L and the sequence must be an m-sequence.

Golomb has conjectured that the span property and the ideal two-level autocorrelation property of a balanced binary sequence implies its cycle-and-add property [22]. This has been confirmed for L up to 10 by many others, but still awaits a complete solution.

3.2.7. Number of Cyclically Distinct *m*-Sequences of Period $2^L - 1$. For a given L, the number of cyclically distinct m-sequences of period $2^L - 1$ is equal to the number of primitive polynomials of degree L over F_2, and this is given by $\phi(2^L - 1)/L$, where $\phi(n)$ is the Euler ϕ function and counts the number of integers from 1 to n that are relatively prime to n. All these $\phi(2^L - 1)/L$ m-sequences of period $2^L - 1$ are equivalent, and they are related with some decimation of each other. Therefore, any given one m-sequence can be used to generate all the others of the same period by using some appropriate decimations.

3.2.8. Trace Function Representation of *m*-Sequences. Let q be a prime power, and let F_q be the finite field with q elements. Let $n = em > 1$ for some positive integers e and m. Then the trace function $\mathrm{tr}_m^n(\cdot)$ is a mapping from F_{2^n} to its subfield F_{2^m} given by

$$\mathrm{tr}_m^n(x) = \sum_{i=0}^{e-1} x^{2^{mi}}$$

It is easy to check that the trace function satisfies the following: (1) $\mathrm{tr}_m^n(ax + by) = a\,\mathrm{tr}_m^n(x) + b\,\mathrm{tr}_m^n(y)$, for all $a, b \in F_{2^m}$, $x, y \in F_{2^n}$; (2) $\mathrm{tr}_m^n(x^{2^m}) = \mathrm{tr}_m^n(x)$, for all $x \in F_{2^n}$; and (3) $\mathrm{tr}_1^n(x) = \mathrm{tr}_1^m(\mathrm{tr}_m^n(x))$, for all $x \in F_{2^n}$. See the literature [13–18] for the detailed properties of the trace function.

Let $q = 2^L$ and α be a primitive element of F_q. Then, an m-sequence $\{s(k)\}$ of period $2^L - 1$ can be represented as

$$s(k) = \mathrm{tr}_1^L(\lambda\alpha^k), \quad k = 0, 1, 2, \ldots, 2^L - 2 \tag{10}$$

where $\lambda \neq 0$ is a fixed constant in F_q. We just give a remark that λ corresponds to the initial condition and the choice of α corresponds to the choice of a primitive polynomial as a connection polynomial when this sequence is generated using an LFSR. Any such representation, on the other hand, gives an m-sequence [17]. When $\lambda = 1$, the sequence is in the characteristic phase, and the constant-on-the-coset property can easily be checked since $s(2k) = \mathrm{tr}_1^L(\alpha^{2k}) = \mathrm{tr}_1^L(\alpha^k) = s(k)$ for all k.

3.2.9. Cross-Correlation Properties of *m*-Sequences. No pair of m-sequences of the same period have the ideal cross-correlation. The best one can achieve is a three-level cross-correlation, and the pair of m-sequences with this property is called a *preferred pair*. Since all m-sequences of a given period are some decimations or cyclic shifts of each other, and they can all be represented as a single trace function from F_{2^L} to F_2, the cross-correlation of a pair of m-sequences can be described as

$$R_d(\tau) = \sum_{k=0}^{2^L-1} (-1)^{\mathrm{tr}_1^L(\alpha^{k+\tau}) + \mathrm{tr}_1^L(\alpha^{dk})}$$

where d indicates that the second m-sequence is a d decimation of the first, and τ represents the amount of phase offset with each other.

Many values of d have been identified that result in a preferred pair of m-sequences, but it is still unknown whether we have found them all. The most famous one that gives a Gold sequence family comes from the value $d = 2^i + 1$ or $d = 2^{2i} - 2^i + 1$ for some integer i when $L/(L, i)$ is odd. Some good references on this topic are Refs. [36–38], and also Chapter 5 of Ref. [2].

3.2.10. Linear Complexity. Given a binary periodic sequence $\{s(k)\}$ of period P, one can always construct an LFSR that outputs $\{s(k)\}$ with a suitable initial condition. One trivial solution is the *pure cycling register* as shown in Fig. 8. It has P stages, the whole period is given as its initial condition, and the characteristic polynomial (or the connection) is given by $f(x) = x^P + 1$ corresponding to $s(k) = s(k - P)$. The best one can do is to find the LFSR with the smallest number of stages, and the linear

complexity of a sequence is defined as this number. Equivalently, it is the degree of the minimal polynomial of the given sequence, which is defined as the minimum degree characteristic polynomial of the sequence.

The linear complexity of a PN sequence, in general, measures cryptographically how strong it is. It is well known that the same copy (whole period) of a binary sequence can be generated whenever $2N$ consecutive terms or more are observed by a third party where N is the linear complexity of the sequence. This forces us to use those sequences with larger linear complexity in some practice. The m-sequences are the worst in this sense because an m-sequence of period $2^L - 1$ has its linear complexity L, and this number is the smallest possible over all the balanced binary sequences of period $2^L - 1$. In the following, we describe the famous Berlekamp–Massey algorithm for determining the linear complexity of a binary sequence [23].

3.3. Berlekamp–Massey Algorithm

Suppose that we are given N terms of a sequence S, which we denote as $S^N = (s(0), s(1), \ldots, s(N-1))$. It does not necessarily mean that S has period N. The goal of the Berlekamp–Massey algorithm (BMA) is to find the minimum degree recursion satisfied by S. This minimum degree $L_N(S)$ is called the *linear complexity* of S^N. This recursion can be used to form an LFSR with $L_N(S)$ stages that generates N terms of S exactly, given the initial condition of $s(0), s(1), \ldots, s(L_N(S) - 1)$. We will denote this LFSR as LFSR$(f^{(N)}(x), L_N(S))$, where the characteristic polynomial after the Nth iteration is given by

$$f^{(N)}(x) = 1 + c_1^{(N)}x + c_2^{(N)}x^2 + \cdots + c_{L_N(S)}^{(N)}x^{L_N(S)}$$

It is not difficult to check that (1) $L_N(S) = 0$ if and only if $s(0), s(1), \ldots, s(N-1)$ are all zeros, (2) $L_N(S) \le N$, and (3) $L_N(S)$ must be monotonically nondecreasing with increasing N.

The BMA updates the degree $L_n(S)$ and the characteristic polynomial $f^{(n)}(x)$ for each $n = 1, 2, \ldots, N$. Assume that $f^{(1)}(x), f^{(2)}(x), \ldots, f^{(n)}(x)$ have been constructed, where the LFSR with connection $f^{(n)}(x)$ of degree $L_n(S)$ generates $s(0), s(1), \ldots, s(n-1)$. Let

$$f^{(n)}(x) = 1 + \sum_{i=1}^{L_n(S)} c_i^{(n)}x^i$$

The next discrepancy, d_n, is the difference between $s(n)$ and the $(n+1)$st bit generated by so far the minimal-length LFSR with $L_n(S)$ stages, and given as

$$d_n = s(n) + \sum_{i=1}^{L_n(S)} c_i^{(n)}s(n-i)$$

Let m be the sequence length before the last length change in the minimal-length register:

$$L_m(S) < L_n(S), \quad \text{and} \quad L_{m+1}(S) = L_n(S)$$

The LFSR with the characteristic polynomial $f^{(m)}(x)$ and length $L_m(S)$ could not have generated $s(0), s(1), \ldots, s(m-1), s(m)$. Therefore, $d_m \neq 0$.

If $d_n = 0$, then this LFSR also generates the first $n+1$ bits $s(0), s(1), \ldots, s(n)$ and therefore, $L_{n+1}(S) = L_n(S)$ and $f^{(n+1)}(x) = f^{(n)}(x)$.

If $d_n \neq 0$, a new LFSR must be found to generate the first $n+1$ bits $s(0), s(1), \ldots, s(n)$. The connection polynomial and the length of the new LFSR are updated by the following:

$$f^{(n+1)}(x) = f^{(n)}(x) - d_n d_m^{-1}x^{n-m}f^{(m)}(x)$$
$$L_{n+1}(S) = \max\left[L_n(S), n + 1 - L_n(S)\right]$$

The complete BM algorithm for implementations is as follows:

1. Initialization:

$$f(x) = 1, \quad g(x) = 1, \quad r = 1, \quad L = 0,$$
$$b = 1, \quad n = 0$$

2. If $n = N$, then stop. Otherwise compute

$$d = s(n) - \sum_{i=1}^{L} c_i s(n-i)$$

3. If $d = 0$, then $r = r + 1$, and go to step 6.
4. If $d \neq 0$ and $2L > n$, then

$$f(x) = f(x) - db^{-1}x^r g(x), \quad r = r + 1$$

and go to step 6.

5. If $d \neq 0$ and $2L \le n$, then

$$h(x) = f(x), \quad f(x) = f(x) - db^{-1}x^r g(x), \quad L = n + 1 - L,$$
$$g(x) = h(x), \quad b = d, \quad r = 1.$$

6. Increase n by 1 and return to step 2.

When $n = N$ and the algorithm is stopped in step (2), the quantities produced by the algorithm bear the following relations:

$$f(x) = f^{(N)}(x)$$
$$L = L_N(S)$$
$$r = N - m$$
$$d = d_{N-1}$$
$$g(x) = f^{(m)}(x)$$
$$b = d_m$$

Table 3. Example of BM Algorithm to the Sequence $(s_0, s_1, s_2, s_3, s_4, s_5, s_6) = (1, 0, 1, 0, 0, 1, 1)$ over F_2

n	L	$f(x)$	r	$g(x)$	b	s_n	d
0	0	1	1	1	1	1	1
1	1	$1+x$	1	1	1	0	1
2	1	1	2	1	1	1	1
3	2	$1+x^2$	1	1	1	0	0
4	2	$1+x^2$	2	1	1	0	1
5	3	1	1	$1+x^2$	1	1	1
6	3	$1+x+x^3$	2	$1+x^2$	1	1	0
7	3	$1+x+x^3$	3	$1+x^2$	1		

An example of BM algorithm applied to a binary sequence of length 7 is shown in Table 3.

3.4. Balanced Binary Sequences with the Ideal Two-Level Autocorrelation

In addition to m-sequences, there are some other well-known balanced binary sequences with the ideal two-level autocorrelation function. For period $P = 4n - 1$ for some positive integer n, all these are equivalent to $(4n - 1, 2n - 1, n - 1)$-cyclic difference sets [24].

In general, a (v, k, λ)-cyclic difference set (CDS) is a k-subset D of the integers mod v, Z_v, such that for each nonzero $z \in Z_v$ there are exactly λ ordered pairs (x, y), $x \in D$, $y \in D$ with $z = x - y \pmod{v}$ [24–28]. For example, $D = \{1, 3, 4, 5, 9\}$ is a $(11, 5, 2)$-CDS and every nonzero integer from 1 to 10 is represented by the difference $x - y \pmod{11}$, $x \in D$, $y \in D$, exactly twice.

The characteristic sequence $\{s(t)\}$ of a (v, k, λ)-cyclic difference set D is defined as $s(t) = 0$ if and only if $t \in D$.

For other values of t, the value 1 is assigned to $s(t)$. This completely characterizes binary sequences of period v with two-level autocorrelation function, and it is not difficult to check that the periodic unnormalized autocorrelation function is given as [29]

$$R(\tau) = \begin{cases} v, & \tau \equiv 0 \pmod{v} \\ v - 4(k - \lambda), & \tau \not\equiv 0 \pmod{v} \end{cases} \tag{11}$$

The out-of-phase value should be kept as low as possible in some practice, and this could happen when $v = 4(k - \lambda) - 1$, resulting in the out-of-phase value to be independent of the period v. The CDS with this parameter is called a *cyclic Hadamard difference set*, and this has been investigated by many researchers [24–28].

In the following, we will simply summarize all the known constructions for $(4n - 1, 2n - 1, n - 1)$ cyclic *Hadamard* difference sets, or equivalently, balanced binary sequences of period $v = 4n - 1$ with the two-level *ideal* autocorrelation function, which are also known as *Hadamard sequences* [30].

Three types of periods are currently known: (1) $v = 4n - 1$ is a prime, (2) $v = 4n - 1$ is a product of twin primes, and (3) $v = 4n - 1$ is one less than a power of 2. All these sequences can be represented as a sum of some decimations of an m-sequence. Song and Golomb have conjectured that a Hadamard sequence of period v exists if and only if v is one of the above three types [31], and this has been confirmed for all the values of $v = 4n - 1$ up to $v = 10000$ with 13 cases unsettled, the smallest of which is $v = 3439$ [30].

3.4.1. When $v = 4n - 1$ Is a Prime. There are two methods in this case [24,29]. The first corresponds to all such values of v, and the resulting sequences are called *Legendre sequences*. Here, D picks up integers mod v that are squares mod v. The second corresponds to some such values of v that can be represented as $4x^2 + 27$ for some integer x, and the resulting sequences are called *Hall's sextic residue sequences*. Here, D picks up integers mod v that are sixth powers mod v and some others.

3.4.2. When $v = 4n - 1$ Is a Product of Twin Primes. This is a generalization of the method for constructing Legendre sequences, and the resulting sequences are called *twin prime sequences*. Let $v = p(p + 1)$, where both p and $p + 2$ are prime. Then, D picks up the integers d that are (1) squares both mod p and mod $p + 2$, (2) nonsquares both mod p and mod $p + 2$, and (3) 0 mod $p + 2$.

3.4.3. When $v = 4n - 1 = 2^L - 1$. Currently, this is a most active area of research, and at least seven families are known. All the sequences of this case can best be described as a sum of some decimations of an m-sequence, or a sum of trace functions from F_{2^L} to F_2. The m-sequences for all the positive integers L and GMW sequences for all the composite integers L [32,33] have been known for many years. One important construction of a larger period from a given one is described by No et al. [34]. More recent discoveries were summarized by No et al. [35], most of which have now been proved by many others.

4. SOME PROPERTIES OF FSR WITH DISJOINT CYCLES

We now return to the basic block diagram of an L-stage FSR as shown in Fig. 1, and its truth table, and state transition diagram as shown in Figs. 2 and 3. Unless otherwise stated, the feedback connection logic $f(x_{k-L}, x_{k-L+1}, \ldots, x_{k-1})$ of all FSRs in this section satisfy the branchless condition given in Eq. (3).

The simplest FSR with L stages is the *pure cycling register* (PCR), as shown in Fig. 8 for $L = 4$. It is linear and has the characteristic polynomial $f(x) = x^L + 1$, or the feedback connection logic $x_k = x_{k-L}$. This obviously satisfies the branchless condition (3), and the state diagram consists only of disjoint cycles. In fact, one can prove that the number $Z(L)$ of disjoint cycles of an L-stage PCR is given as

$$Z(L) = \frac{1}{L} \sum_{d|L} \phi(d) 2^{L/d}, \tag{12}$$

where $\phi(d)$ is the Euler ϕ function and the summation is over all the divisors of L. It is not very surprising that this number is the same as the number of irreducible polynomials of degree L over the binary alphabet F_2. Golomb had conjectured that the number of disjoint cycles from an L-stage FSR with branchless condition satisfied

by its connection logic is *at most* $Z(L)$ given in (12), and this was confirmed by Mykkeltveit in 1972 [39].

On the other hand, the minimum number of disjoint cycles is 1, and this corresponds to de Bruijn sequences of period 2^L. Inserting a single 0 into any m-sequence right after the run of 0s of length $L-1$ gives a de Bruijn sequence of period 2^L. This can be done by the following modification to the linear logic f_{old} that generates the m-sequence

$$f_{\text{new}} = f_{\text{old}} \oplus x'_{k-1}x'_{k-2}\cdots x'_{k-L+1}, \tag{13}$$

where x' represents the complement of x. A de Bruijn sequence can best be described using Good's diagram. This is shown in Fig. 9 for $L=2$ and $L=3$. Note that any node in a Good diagram has two incoming edges as well as two outgoing edges. A de Bruijn sequence of period 2^L corresponds to a closed path (or a cycle) on the Good diagram of order L, which visits every node exactly once. It was shown earlier in 1946 by de Bruijn that the number of such cycles is given as $2^{2^{L-1}-L}$ [10].

The number of disjoint cycles of an L-stage FSR with (possibly) nonlinear logic connections is not completely determined. Toward this direction, we simply state a condition for the parity of this number for FSRs with branchless condition. *The number of disjoint cycles of an FSR with the branchless condition is even* (*or odd, respectively*) *if and only if the number of 1s in its truth table of* $g(x_{k-1}, x_{k-2}, \ldots, x_{k-L+1})$ *is even* (*or odd, respectively*) [10]. In other words, the parity of the top half of the truth table is the parity of the number of disjoint cycles. This implies that PCR has an even number of disjoint cycles $L > 2$.

In addition to PCR, there are three more degenerate cases: complemented cycling registers (CCRs), pure summing registers (PSRs), and complemented summing registers (CSRs). Note that PCR and PSR are homogeneous linear, but CCR and CSR are inhomogeneous linear:

$$\begin{aligned} x_k &= x_{k-L}, \quad \text{PCR} \\ &= 1 + x_{k-L}, \quad \text{CCR} \\ &= x_{k-1} + x_{k-2} + \cdots + x_{k-L} \quad \text{PSR} \\ &= 1 + x_{k-1} + x_{k-2} + \cdots + x_{k-L} \quad \text{CSR} \end{aligned}$$

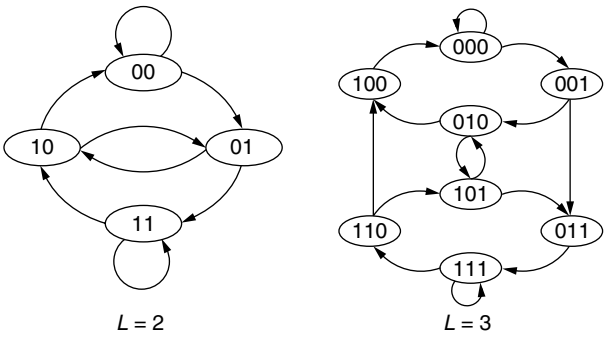


Figure 9. Good's diagrams for $L=2$ and $L=3$.

Table 4. Output Sequences from Two Degenerated FSR with $L=4$

CCR	Period	CSR	Period
(00001111)	8	(1)	1
		(00001)	5
(01011010)	8	(00111)	5
		(01011)	5

All these satisfy the branchless condition, and the output sequences from CCR and CSR for $L=4$ are listed in Table 4.

The number of L-stage FSRs with branchless condition is given as $2^{2^{L-1}}$. Another condition for the truth table of an FSR to be practically useful is the *balanced logic* condition. A truth table of f for an L-stage FSR is said to have the balanced logic condition if f has equally many 1s and 0s in its truth table. Both the branchless condition and the balanced logic condition together imply that f has equally many 1s and 0s in its top half of the truth table. The balanced logic condition guarantees that *the autocorrelation of the output sequence of period approaching* 2^L *tends to zero for* $\tau = 1, 2, \ldots, L$.

Finally, we give a detailed analysis on the branchless four-stage FSRs with all possible connections. With the branchless condition on it, there are $2^{2^3} = 2^8 = 256$ such FSRs. Among these, 8 FSRs are linear as shown in Figs. 5–8. These are collected in Table 5. Except for PCR, all the other 7 LFSRs are balanced logic FSRs. Among the 248 nonlinear FSRs, there are $2^{2^3-4} = 16$ FSRs, which generate de Bruijn sequences of period 16, as shown in Table 6. This table also shows all the 16 output sequences in four equivalent classes, denoted as A, B, C, and D. The linear complexity of each de Bruijn sequence is also shown here. The 4th and 14th sequences are modified versions from the 6th and 8th m-sequences in Table 5, respectively. The modification is the insertion of a single 0 into the m-sequence right after the run of 0s of length 3. Note that none of the connection logic satisfies the balanced logic condition. Among the 256 FSRs with the branchless condition, there are $\binom{8}{4} = 70$ FSRs that satisfy the balanced logic condition. Table 7 shows all of them. In this table, $*$ represents that it is linear and is also shown in Table 5.

5. CONCLUDING REMARKS

There is, in fact, a large amount of literature on FSRs, on FSR sequences, and their variations, generalizations, and applications.

Analysis of LFSR and LFSR sequences can also be done using at least two other standard methods not described in this article: the generating function approach and the matrix representation approach [10,13].

There are generalizations of LFSR over a nonbinary alphabet. For this, at least two operations, addition and multiplication, must be well defined over the alphabet. The well-known example of such an alphabet is a finite field with q elements. A finite commutative ring sometimes serves as an appropriate alphabet over which an LFSR is

Table 5. Truth Tables of All Four-Stage LFSRs, Including Their Output Sequences and Characteristic Polynomials

	Truth Table	Output Sequences	Characteristic Polynomials	Figures
1	00000000 11111111	0, 1, 01, 0001, 0011, 0111	x^4+1	Fig. 8
2	00111100 11000011	0, 1, 0001011, 0011101	$x^4+x^3+x^2+1$	Fig. 6
3	01011010 10100101	0, 1, 01, 001, 011, 000111	x^4+x^3+x+1	Fig. 7
4	01100110 10011001	0, 1, 0001101, 0010111	x^4+x^2+x+1	Fig. 6
5	01101001 10010110	0, 00011, 00101, 01111	$x^4+x^3+x^2+x+1$	Fig. 8
6	01010101 10101010	0, 000111101011001	x^4+x+1	Fig. 5
7	00110011 11001100	0, 011, 000101, 001111	x^4+x^2+1	Fig. 7
8	00001111 11110000	0, 000100110101111	x^4+x^3+1	Fig. 5

Table 6. The Truth Tables of All Four-Stage FSRs That Generate de Bruijn Sequences in Four Equivalent Classes[a]

	Truth Table	Output Sequence	Equivalent Class	LC
1	11110001 00001110	0000111101100101	A	15
2	10111001 01000110	0000101001111011	C	15
3	11100101 00011010	0000110010111101	D	14
4*	11010101 00101010	0000111101011001	A	15
5	10110101 01001010	0000101100111101	D	14
6	10101101 01010010	0000101111010011	D	14
7	10011101 01100010	0000100111101011	C	15
8	11111101 00000010	0000111101001011	B	12
9	11100011 00011100	0000110111100101	C	15
10	10101011 01010100	0000101001101111	A	15
11	11000111 00111000	0000110101111001	C	15
12	10100111 01011000	0000101111001101	D	14
13	11110111 00001000	0000111100101101	B	12
14*	10001111 01110000	0000100110101111	A	15
15	11101111 00010000	0000110100101111	B	12
16	10111111 01000000	0000101101001111	B	12

[a]The linear complexity of each de Bruijn sequence is also shown. The 4th and 14th sequences are modified versions of 6th 8th m-sequences in Table 5, respectively. The asterisk denotes that the FSR is linear.

operating. Integers mod 4 is the best known in this regard, due to the application of the output sequences into QPSK modulation [13,38].

There are other directions to which FSR may be generalized. For example, one can consider LFSR with inputs. One application is to use the LFSR with input as a polynomial division circuit. These are used in the decoding/encoding of Hamming codes or other channel (block) codes. Another example is to use multiple (nonlinear) FSRs on a stack so that the stages of an FSR in one layer are used to produce inputs to upper layer FSRs on top of it. These find some important applications to generating PN sequences with larger linear complexity in streamcipher systems.

All these topics are currently under active research, and one should look at journal transactions for the most recent results and applications.

BIOGRAPHY

Hong-Yeop Song received his B.S. degree in Electronics Engineering from Yonsei University in 1984 and his M.S.E.E. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1986 and 1991, respectively. After spending 2 years on the research staff in the Communication Sciences Institute at USC, he joined Qualcomm Inc., San Diego, California in 1994 as a Senior Engineer and worked in a team researching and developing North American CDMA Standards for PCS and cellular air interface. Finally, he joined the Department of Electrical and Electronics Engineering at Yonsei University, Seoul, South Korea in 1995, and is currently serving as an Associate Professor. His areas of research interest are sequence design and analysis for speed-spectrum communications and stream ciphers, theory and application of cyclic difference sets, and channel coding/decoding problems for wireless communications. He is currently visiting the University of Waterloo, Ontario, Canada, from March 2002 to Feb 2003 during his sabbatical leave.

BIBLIOGRAPHY

1. S. W. Golomb, *Digital Communications with Space Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

Table 7. Truth Tables and Cycle Length Distributions of All Four-Stage FSRs That Satisfy Both Branchless and Balanced Logic Conditions[a]

	Truth Table	Cycle Length Distribution		Truth Table	Cycle Length Distribution
1	1111000000001111	1:1, 15:1	36	1110000100011110	5:1, 11:1
2	1110100000010111	1:1, 4:1, 5:1, 6:1	37	1101000100101110	2:1, 14:1
3	1101100000100111	1:1, 2:1, 3:1, 10:1	38	1011000101001110	7:1, 9:1
4	1011100001000111	1:1, 15:1	39	0111000110001110	1:1, 15:1
5	0111100010000111	1:2, 5:1, 9:1	40	1100100100110110	2:1, 3:1, 5:1, 6:1
6	1110010000011011	1:1, 15:1	41	1010100101010110	5:1, 11:1
7	1101010000101011	1:1, 15:1	42*	0110100110010110	1:1, 5:3
8	1011010001001011	1:1, 15:1	43	1001100101100110	2:1, 14:1
9	0111010010001011	1:2, 6:1, 8:1	44	0101100110100110	1:1, 2:1, 3:1, 10:1
10	1100110000110011	1:1, 3:1, 6:2	45	0011100111000110	1:1, 15:1
11	1010110001010011	1:1, 15:1	46	1100010100111010	7:1, 9:1
12	0110110010010011	1:2, 5:1, 9:1	47	1010010101011010	4:1, 12:1
13	1001110001100011	1:1, 15:1	48	0110010110011010	1:1, 15:1
14	0101110010100011	1:2, 3:1, 11:1	49	1001010101101010	5:1, 11:1
15*	0011110011000011	1:2, 7:2	50*	0101010110101010	1:1, 15:1
16	1110001000011101	1:1, 15:1	51	0011010111001010	1:1, 15:1
17	1101001000101101	1:1, 2:1, 3:1, 10:1	52	1000110101110010	7:1, 9:1
18	1011001001001101	1:1, 3:1, 5:1, 7:1	53	0100110110110010	1:1, 3:1, 5:1, 7:1
19	0111001010001101	1:2, 3:1, 11:1	54	0010110111010010	1:1, 15:1
20	1100101000110101	1:1, 2:1, 3:1, 10:1	55	0001110111100010	1:1, 15:1
21	1010101001010101	1:1, 15:1	56	1100001100111100	2:1, 14:1
22	0110101010010101	1:2, 5:1, 9:1	57	1010001101011100	7:1, 9:1
23	1001101001100101	1:1, 2:1, 3:1, 10:1	58	0110001110011100	1:1, 15:1
24*	0101101010100101	1:2, 2:1, 3:2, 6:1	59	1001001101101100	2:1, 3:1, 5:1, 6:1
25	0011101011000101	1:2, 3:1, 11:1	60	0101001110101100	1:1, 2:1, 3:1, 10:1
26	1100011000111001	1:1, 15:1	61*	0011001111001100	1:1, 3:1, 6:2
27	1010011001011001	1:1, 15:1	62	1000101101110100	2:1, 14:1
28*	0110011010011001	1:2, 7:2	63	0100101110110100	1:1, 2:1, 3:1, 10:1
29	1001011001101001	1:1, 5:3	64	0010101111010100	1:1, 15:1
30	0101011010101001	1:2, 5:1, 9:1	65	0001101111100100	1:1, 2:1, 3:1, 10:1
31	0011011011001001	1:2, 5:1, 9:1	66	1000011101111000	5:1, 11:1
32	1000111001110001	1:1, 15:1	67	0100011110111000	1:1, 15:1
33	0100111010110001	1:2, 3:1, 11:1	68	0010011111011000	1:1, 15:1
34	0010111011010001	1:2, 6:1, 8:1	69	0001011111101000	1:1, 4:1, 5:1, 6:1
35	0001111011100001	1:2, 5:1, 9:1	70*	0000111111110000	1:1, 15:1

[a]Here, $a:b$ means that there are b cycles of length a, and * implies that the FSR is linear.

2. M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook*, Computer Science Press, Rockville, MD, 1985; rev. ed., McGraw-Hill, 1994.
3. R. L. Perterson, R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
4. J. G. Proakis and M Salehi, *Communication Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1994.
5. A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, Reading, MA, 1995.
6. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
7. R. A. Rueppel, *Analysis and Design of Stream Ciphers*, Springer-Verlag, Berlin, 1986.
8. A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, FL, 1996.
9. S. B. Volchan, What is a Random Sequence? *Am. Math. Monthly* **109**: 46–63 (2002).
10. S. W. Golomb, *Shift Register Sequences*, Holden-Day, San Francisco, CA, 1967; rev. ed., Aegean Park Press, Laguna Hills, CA, 1982.
11. D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, Boca Raton, FL, 1995.
12. TIA/EIA/IS-95, *Mobile Station–Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System*, published by Telecommunications Industry Association as a North American 1.5 MHz Cellular CDMA Air-Interface Standard, July 1993.
13. R. Lidl and H. Niederreiter, Finite fields, in *Encyclopedia of Mathematics and Its Applications*, Vol. 20, Addison-Wesley, Reading, MA, 1983.
14. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
15. W. W. Peterson and E. J. Weldon, *Error-Correcting Codes*, 2nd ed., MIT Press, Cambridge, MA, 1972.
16. F. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, 1977.
17. R. J. McEliece, *Finite Fields for Computer Scientists and Engineers*, Kluwer, 1987.

18. K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, 2nd ed., Springer-Verlag, New York, 1991.

19. T. Hansen and G. L. Mullen, Supplement to primitive polynomials over finite fields, *Math. Comput.* **59**: S47–S50 (Oct. 1992).

20. J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge Univ. Press, New York, 1992.

21. H. Fredricksen, A survey of full length nonlinear shift register cycle algorithms, *SIAM Rev.* **24**: 195–221 (1982).

22. S. W. Golomb, On the classification of balanced binary sequences of period $2^n - 1$, *IEEE Trans. Inform. Theory* **26**: 730–732 (1980).

23. J. L. Massey, Shift-register synthesis and BCH decoding, *IEEE Trans. Inform. Theory* **15**: 122–127 (1969).

24. L. D. Baumert, *Cylic Difference Sets, Lecture Notes in Mathematics*, Vol. 182, Springer-Verlag, New York, 1971.

25. M. Hall, Jr., A survey of difference sets, *Proc. Am. Math. Soc.* **7**: 975–986 (1956).

26. H. J. Ryser, *Combinatorial Mathematics*, *The Carus Mathematical Monographs* No. 14, Mathematical Association of America, 1963.

27. D. Jungnickel, Difference sets, in J. H.Dinitz and D. R. Stinson, eds., *Contemporary Design Theory*, Wiley, New York, 1992, pp. 241–324.

28. C. J. Colbourn and J. H. Dinitz, *The CRC Handbook of Combinatorial Designs*, CRC Press, New York, 1996.

29. S. W. Golomb, Construction of signals with favourable correlation properties, in A. D. Keedwell, ed., *Survey in Combinatorics*, *LMS Lecture Note Series*, Vol. 166, Cambridge Univ. Press, 1991, pp. 1–40.

30. J.-H. Kim, *On the Hadamard Sequences*, Ph.D. thesis, Yonsei Univ., South Korea, 2001.

31. H.-Y. Song and S. W. Golomb, On the existence of cyclic Hadamard difference sets, *IEEE Trans. Inform. Theory* **40**: 1266–1268 (1994).

32. B. Gordon, W. H. Mills, and L. R. Welch, Some new difference sets, *Can. J. Math.* **14**: 614–625 (1962).

33. R. A. Scholtz and L. R. Welch, GMW sequences, *IEEE Trans. Inform. Theory* **30**: 548–553 (1984).

34. J. S. No, H. Chung, K. Yang, and H. Y. Song, On the construction of binary sequences with ideal autocorrelation property, *Proc. IEEE Int. Symp. Information Theory and Its Application*, Victoria, BC, Canada, Sept. 1996, pp. 837–840.

35. J.-S. No et al., Binary pseudorandom sequences of period $2^n - 1$ with ideal autocorrelation, *IEEE Trans. Inform. Theory* **44**: 814–817 (1998).

36. D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudorandom and related sequences, *Proc. IEEE* **68**: 593–619 (1980).

37. P. Fan and M. Darnell, *Sequence Design for Communications Applications*, Research Studies Press LTD, Taunton, Somerset, England, 1995.

38. T. Helleseth and P. V. Kumar, Sequences with low correlation, in V. S. Pless and W. C. Huffman, eds., *Handbook of Coding Theory*, Elsevier Science B.V., 1998, Chap. 21.

39. J. Mykkeltveit, A proof of Golomb's conjecture on the de bruijn graph, *J. Comb. Theory Ser. B* **13**: 40–45 (1972).

FINITE-GEOMETRY CODES

M. FOSSORIER
University of Hawaii at Manoa
Honolulu, Hawaii

1. INTRODUCTION

Euclidean geometry (EG) and projective geometry (PG) codes belong to the class of algebraic block codes derived from finite geometries. These codes were introduced by Rudolph [1,2] and have been studied by many other researchers since [3–8].

An m-dimensional finite geometry is composed of a finite number of elements called *points*, or 1-flat. For $1 \leq \mu < m$, it is then possible to divide this set of points into subsets of identical structure defined by an equation with in general either μ, or $\mu + 1$ unknowns. Each subset is called a *μ-dimensional hyperplane* or μ-flat, and the ensemble of these μ-flats can be associated with a linear block code. In this article, we focus on codes associated with the value $\mu = 1$. These codes belong to the class of one-step majority-logic decodable codes, which have many nice structural properties that can exploited in their decoding. For example, an important but very limited class of such codes is that of difference-set cyclic (DSC) codes found independently by Weldon [9]. A DSC code has been used for error correction of a digital audio broadcasting system in Japan [10, Sect. 5-A].

The construction of EG and PG codes has long been motivated by their fast decoding with majority-logic decoding [11–13]. In addition, one-step majority logic decodable EG and PG codes can easily be decoded iteratively. Efficient iterative decoding methods have been devised for these codes, initially based on heuristic approaches [14,15], and more recently in conjunction with iterative decoding of the low-density parity-check (LDPC) codes [16,17]. On the basis of this relationship, many new classes of LDPC codes have also been constructed from finite geometries [17].

The construction concepts of EG and PG codes are summarized in Section 2 of this article. The subclasses of one-step majority logic decodable EG and PG codes are first considered, and the extension to other EG and PG codes is then discussed. The interested reader is referred to the literature [18–23] for more detailed expositions of finite geometries and their applications to error control coding. The link between EG and PG codes, and LDPC codes is finally briefly discussed. Several decoding methods are presented in Section 3.

2. CODE CONSTRUCTIONS

2.1. EG Codes

An m-dimensional Euclidean geometry over the finite Galois field GF(2^s) [denoted EG($m, 2^s$)] consists of the 2^{ms} m-tuples $\mathbf{p} = (p_0, p_1, \ldots, p_{m-1})$ (referred to as *points*), where for $0 \leq i \leq m-1$, each p_i is an element of GF(2^s). Two linearly independent points $\mathbf{p}_1$ and $\mathbf{p}_2$ [i.e., for $\alpha_1 \in$ GF(2^s) and $\alpha_2 \in$ GF(2^s), $\alpha_1 \mathbf{p}_1 + \alpha_2 \mathbf{p}_2 = \mathbf{0}$ if and only if

$\alpha_1 = \alpha_2 = 0$] define a unique line $L(\mathbf{p}_1, \mathbf{p}_1 + \mathbf{p}_2)$ passing through $\mathbf{p}_1$ and $\mathbf{p}_1 + \mathbf{p}_2$. The line $L(\mathbf{p}_1, \mathbf{p}_1 + \mathbf{p}_2)$ contains a total of 2^s points $\mathbf{p}_i$ which satisfy the equation

$$\mathbf{p}_i = \mathbf{p}_1 + \alpha \mathbf{p}_2 \tag{1}$$

$\alpha \in \mathrm{GF}(2^s)$. Given a point $\mathbf{p}$, the $2^{ms} - 1$ other points in $\mathrm{EG}(m, 2^s)$ can be divided into subsets of $2^s - 1$ points, each based on Eq. (1) such that each subset corresponds to a distinct line of $\mathrm{EG}(m, 2^s)$ containing $\mathbf{p}$. As a result, there are $(2^{ms} - 1)/(2^s - 1)$ lines intersecting on each point $\mathbf{p}$ of $\mathrm{EG}(m, 2^s)$ and the total number of lines in $\mathrm{EG}(m, 2^s)$ is $2^{(m-1)s}\ (2^{ms} - 1)/(2^s - 1)$.

Let us consider the incidence matrix $H = [h_{i,j}]$ whose rows are associated with the $(2^{(m-1)s} - 1)\ (2^{ms} - 1)/(2^s - 1)$ lines not containing the all-zero point $\mathbf{0}$, and whose columns are associated with the $2^{ms} - 1$ points other than $\mathbf{0}$. Then $h_{i,j} = 1$ if the jth point of $\mathrm{EG}(m, 2^s)$ is contained in the ith line of $\mathrm{EG}(m, 2^s)$, and $h_{i,j} = 0$ otherwise. This incidence matrix H defines the dual space of a linear block EG code C denoted $C_{\mathrm{EG}}(m, s)$ of length $N = 2^{ms} - 1$. In general, the dimension K and minimum distance $d_{\min}$ of $C_{\mathrm{EG}}(m, s)$ depend on the values m and s, with $d_{\min} \geq (2^{ms} - 1)/(2^s - 1)$. From the structural properties of $\mathrm{EG}(m, 2^s)$, we conclude that (1) every row of H contains exactly 2^s "ones," (2) every column of H contains exactly $(2^{ms} - 1)/(2^s - 1) - 1$ "ones," and (3) any two rows of H have at most one "one" in common. These three properties, and especially property 3, are being used in the decoding of EG codes. Another interesting property of EG codes is that with a proper ordering of the columns of H, $C_{\mathrm{EG}}(m, s)$ is a cyclic code. This property is important for simple encoding of EG codes based on linear feedback shift registers (LFSRs).

If all 2^{ms} points of $\mathrm{EG}(m, 2^s)$ are kept to build the incidence matrix H, then we obtain an extended EG code with length $N = 2^{ms}$ and $d_{\min} \geq (2^{ms} - 1)/(2^s - 1) + 1$. In that case, properties 1 and 3 remain the same while property 2 becomes: Every column of H contains exactly $(2^{ms} - 1)/(2^s - 1)$ "ones" as each position has exactly one additional checksum associated with a line passing through the origin. This extended code is no longer cyclic, but encoding of an extended cyclic code is readily achieved from the cyclic encoder by the trivial addition of an overall parity-check bit.

As an example, let us choose $m = 2$. Then the two-dimensional EG codes $C_{\mathrm{EG}}(2, s)$ have the following parameters:

$$N = 2^{2s} - 1$$
$$K = 2^{2s} - 3^s$$
$$d_{\min} = 2^s + 1$$

For $s = 2$, $\mathrm{EG}(2, 2^2)$ has 15 lines not passing through $\mathbf{0}$, each point belonging to four different lines. Then, the incidence matrix

$$H = \begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

defines the dual space of $C_{\mathrm{EG}}(2, 2)$, an $(N, K, d_{\min}) = (15, 7, 5)$ code. On the basis of H, and assuming that column i is associated with point $\mathbf{p}_i$ for $0 \leq i \leq 14$, we observe that the four lines intersecting point $\mathbf{p}_{14}$ and not passing through $\mathbf{0}$ are $\{\mathbf{p}_0, \mathbf{p}_2, \mathbf{p}_6, \mathbf{p}_{14}\}$, $\{\mathbf{p}_1, \mathbf{p}_5, \mathbf{p}_{13}, \mathbf{p}_{14}\}$, $\{\mathbf{p}_3, \mathbf{p}_{11}, \mathbf{p}_{12}, \mathbf{p}_{14}\}$, and $\{\mathbf{p}_7, \mathbf{p}_8, \mathbf{p}_{10}, \mathbf{p}_{14}\}$. Finally, the incidence matrix of the extended (16,7,6) EG code is obtained from H by adding an all-zero column in front of H and then appending to the bottom a 5×16 matrix composed of the all-one column followed by the 5×5 identity matrix repeated 3 times. The first column in the new 16×20 incidence matrix corresponds to the origin of $\mathrm{EG}(2, 2^2)$, and the five last rows correspond to the five lines of $\mathrm{EG}(2, 2^2)$ passing through $\mathbf{0}$.

The definition of a line or 1-flat given in (1) can be generalized as

$$\mathbf{p}_i = \mathbf{p}_0 + \alpha_1 \mathbf{p}_1 + \alpha_2 \mathbf{p}_2 + \cdots + \alpha_\mu \mathbf{p}_\mu \tag{2}$$

with $1 \leq \mu < m$ and for $1 \leq i \leq \mu, \alpha_i \in \mathrm{GF}(2^s)$. Equation (2) defines a μ-flat of $\mathrm{EG}(m, 2^s)$ that passes through the point $\mathbf{p}_0$. It is then straightforward to associate with $\mathrm{EG}(m, 2^s)$ an incidence matrix H whose rows are associated with the μ-flats of $\mathrm{EG}(m, 2^s)$ (possibly not containing $\mathbf{0}$), and whose columns are associated with the points of $\mathrm{EG}(m, 2^s)$ (possibly excluding $\mathbf{0}$). This incidence matrix H defines the dual space of a linear block EG code C. It follows that in general, an EG code is totally defined by the three parameters m, s, and μ. Further extensions of the definition of an EG code are possible by considering collections of flats. For example, twofold EG codes are obtained by considering pairs of parallel μ-flats, called "$(\mu, 2)$-frames" [6].

2.2. PG Codes

Since $\mathrm{GF}(2^{(m+1)s})$ contains $\mathrm{GF}(2^s)$ as a subfield, it is possible to divide the $2^{(m+1)s} - 1$ nonzero elements of $\mathrm{GF}(2^{(m+1)s})$ into $N(m, s) = (2^{(m+1)s} - 1)/(2^s - 1) = 2^{ms} + 2^{(m-1)s} + \cdots + 2^s + 1$ disjoint subsets of $2^s - 1$ elements each. Then each subset is regarded as a point in $\mathrm{PG}(m, 2^s)$, the m-dimensional projective geometry over the finite field $\mathrm{GF}(2^s)$.

Two distinct points $\mathbf{p}_1$ and $\mathbf{p}_2$ of PG$(m, 2^s)$ define a unique line $L(\mathbf{p}_1, \mathbf{p}_2)$ passing through $\mathbf{p}_1$ and $\mathbf{p}_2$. The line $L(\mathbf{p}_1, \mathbf{p}_2)$ contains a total of $(2^{2s} - 1)/(2^s - 1) = 2^s + 1$ points $\mathbf{p}_i$ satisfying the equation

$$\mathbf{p}_i = \alpha_1 \mathbf{p}_1 + \alpha_2 \mathbf{p}_2, \tag{3}$$

with α_1 and α_2 in GF(2^s), and not both zero. Given a point $\mathbf{p}$, the $N(m, s) - 1 = 2^s(2^{ms} - 1)/(2^s - 1)$ other points in PG$(m, 2^s)$ can be divided into subsets of 2^s points each based on Eq. (3) such that each subset corresponds to a distinct line of PG$(m, 2^s)$ containing $\mathbf{p}$. As a result, there are $(2^{ms} - 1)/(2^s - 1)$ lines intersecting on each point $\mathbf{p}$ of PG$(m, 2^s)$ and the total number of lines in PG$(m, 2^s)$ is $N(m, s)/(2^s + 1) \cdot (2^{ms} - 1)/(2^s - 1)$.

As for EG codes, we associate to PG$(m, 2^s)$ an incidence matrix H whose rows are associated with the $N(m, s)/(2^s + 1) \cdot (2^{ms} - 1)/(2^s - 1)$ lines of PG$(m, 2^s)$, and whose columns are associated with the $N(m, s)$ points of PG$(m, 2^s)$. This incidence matrix H defines the dual space of a linear block PG code C denoted $C_{PG}(m, s)$ of length $N = N(m, s)$. As for EG codes, the dimension K and minimum distance $d_{\min}$ of $C_{\mathrm{PG}}(m, s)$ depend on the values m and s, with $d_{\min} \geq (2^{ms} - 1)/(2^s - 1) + 1$. Also, PG codes are cyclic codes and have structural properties similar to those of EG codes, namely (1) every row of H contains exactly $2^s + 1$ "ones," (2) every column of H contains exactly $(2^{ms} - 1)/(2^s - 1)$ "ones," and (3) any two rows of H have at most one "one" in common.

For $m = 2$, the 2-dimensional PG codes $C_{\mathrm{PG}}(2, s)$ are equivalent to the DSC codes described by Weldon [9] and have the following parameters:

$$N = 2^{2s} + 2^s + 1$$
$$K = 2^{2s} + 2^s - 3^s$$
$$d_{\min} = 2^s + 2$$

For $1 \leq \mu < m$, a μ-flat of PG$(m, 2^s)$ is defined by the set of points $\mathbf{p}_i$ of the form

$$\mathbf{p}_i = \alpha_1 \mathbf{p}_1 + \alpha_2 \mathbf{p}_2 + \cdots + \alpha_{\mu+1} \mathbf{p}_{\mu+1} \tag{4}$$

with for $1 \leq i \leq \mu$, $\alpha_i \in$ GF(2^s). The dual code of a linear block PG code C is defined by the incidence matrix H whose rows and columns are associated with the μ-flats and the points of PG$(m, 2^s)$, respectively. Further extensions of this definition are also possible.

Some EG and PG codes of length $N \leq 1000$ are given in Table 1. Note that with respect to that table, for a given triplet (m, s, μ), the corresponding EG code is commonly referred to as the "$(\mu - 1, s)$th-order binary EG code," while the corresponding PG code is commonly referred to as the "(μ, s)th-order binary PG" code.

2.3. EG and PG Codes Viewed as LDPC Codes

An (J, L) LDPC code is defined by as the null space of a matrix H such that (1) each column consists of J "ones," (2) each row consists of L "ones," and (3) no two rows have more than one "one" in common [24] (note that this last property is not explicitly stated in Gallager's original definition). It is readily seen that both EG and PG codes satisfy this definition. However, compared with their counterpart LDPC codes presented by Gallager [24], EG and PG codes have several fundamental differences: (1) their values J and L are in general much larger, (2) their total number of check sums defining H is larger, and (3) they usually have a cyclic structure. On the basis of these observations, new LDPC codes can be constructed from EG and PG codes by splitting the rows or columns of H or its transposed matrix [17]. If these operations are done in a systematic and uniform way, new LDPC codes are obtained. Interestingly, these codes become quasicyclic, so that fast encoding based on LFSRs is still possible.

As an example, let us consider $C_{\mathrm{EG}}(2, 6)$, a (4095,3367) EG code with $J = L = 64$, and let us split each column of H into 16 columns of weight 4 each. Then we obtain a (65520,61425) EG extended code with $J = 4$ and $L = 64$ [17]. Note that this new code has rate $R = 0.9375 = 1 - J/L$.

Table 1. Some EG and PG Codes of Length $N \leq 1000$

			EG Codes			PG Codes		
m	s	μ	N	K	$d_{\min}$	N	K	$d_{\min}$
2	2	1	15	7	5	21	11	6
2	3	1	63	37	9	73	45	10
3	2	1	63	13	21	85	24	22
3	2	2	63	48	5	85	68	6
2	4	1	255	175	17	273	191	18
4	2	1	255	21	85	341	45	86
4	2	2	255	127	21	341	195	22
4	2	3	255	231	5	341	315	6
3	3	1	511	139	73	585	184	74
3	3	2	511	448	9	585	520	10

3. DECODING METHODS

3.1. Decoding of EG and PG Codes with $\mu = 1$

3.1.1. One-Stage Decoding. Let us assume that an information sequence $\mathbf{u} = (u_0, u_1, \ldots, u_{K-1})$ is encoded into a codeword $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$ using an EG, extended EG or PG (N, K) code associated with $\mu = 1$. This codeword is then transmitted over a noisy communications channel and at the receiver, either a hard-decision received vector $\mathbf{y} = (y_0, y_1, \ldots, y_{N-1})$, or a soft-decision estimate $\mathbf{r} = (r_0, r_1, \ldots, r_{N-1})$ is available.

For any codeword $\mathbf{v}$ of the code considered, each row $\mathbf{h}$ of H satisfies the checksum $\mathbf{v} \cdot \mathbf{h} = 0$. Based on properties 2 and 3 of extended EG and PG codes described in Section 2, each position i, $0 \leq i \leq N - 1$ is associated with a set $B(i)$ of $(2^{ms} - 1)/(2^s - 1)$ checksums orthogonal on that position [i.e., no other position than i appears more than once in $B(i)$]. As a result, each checksum can be used to provide uncorrelated information about bit i, and since all checksums have the same weight (see property 1), the same amount of information is provided by each of them. This suggests the following algorithm:

1. For $0 \le i \le N-1$, evaluate all checksums $\mathbf{y} \cdot \mathbf{h}$ in $B(i)$.
2. Let $B(i)^+$ and $B(i)^-$ represent the sets of satisfied and unsatisfied checksums in $B(i)$, respectively.
   If $|B(i)^+| \ge |B(i)^-|$, decode $\hat{v}_i = y_i$.
   Else, decode $\hat{v}_i = y_i \oplus 1$.

Since a majority vote is taken for each bit, this decoding method is known as "majority-logic decoding." It was first introduced by Reed [11] to decode Reed–Muller codes. Furthermore, since for the codes considered, all N bits can be decoded by this algorithm, these codes belong to the class of "one-step majority-logic decodable codes." Majority-logic decoding is a very simple algebraic decoding method that fits high-speed implementations since only binary operations are involved. It allows one to correct any error pattern of Hamming weight $(|B(i)| - 1)/2 = \lfloor (2^{ms-1} - 2^{s-1})/(2^s - 1) \rfloor$, which corresponds to the guaranteed error-correcting capability $t = \lfloor (d_{\min} - 1)/2 \rfloor$ of extended EG and PG codes. Since for EG codes, $|B(i)| = (2^{ms} - 1)/(2^s - 1) - 1$ is even, this algorithm has to be refined in order to consider the case where $|B(i)|/2$ checksums are satisfied and $|B(i)|/2$ checksums are unsatisfied. In that case, we always decode $\hat{v}_i = y_i$, so that the guaranteed error-correcting capability $t = |B(i)|/2 = \lfloor (d_{\min} - 1)/2 \rfloor$ of EG codes is also achieved.

The sets $B(i)$, $0 \le i \le N-1$ can also be used to evaluate the a posteriori probabilities q_i that the values y_i are in error, based on the a priori error probabilities p_i defined by the channel model considered. For example, $p_i = \varepsilon$ for a binary symmetric channel (BSC) with crossover probability ε and $p_i = e^{-|4\, r_i/N_0|}/(1 + e^{-|4\, r_i/N_0|})$ for an additive white Gaussian noise (AWGN) channel with associated variance $N_0/2$. Since the checksums in each set $B(i)$, $0 \le i \le N-1$ are orthogonal on position i, it follows that

$$q_i = \left(1 + \left(\frac{1-p_i}{p_i}\right) \prod_{j=1}^{|B(i)|} \left(\frac{1-m_{j,i}}{m_{j,i}}\right)^{\sigma_j \oplus 1} \left(\frac{m_{j,i}}{1-m_{j,i}}\right)^{\sigma_j}\right)^{-1} \tag{5}$$

where σ_j is the result of the jth checksum in $B(i)$, and

$$m_{j,i} = \left(\frac{1}{2}\right)\left(1 - \prod_{i' \in N(j)\backslash i} (1 - 2p_{i'})\right)$$

$N(j)\backslash i$ representing the set of nonzero positions corresponding to checksum σ_j, but position i. Note that $m_{j,i}$ represents the probability that the sum of the bits in $N(j)\backslash i$ mismatches the transmitted bit i. The corresponding decoding algorithm is given as follows:

1. For $0 \le i \le N-1$, evaluate all checksums $\mathbf{y} \cdot \mathbf{h}$ in $B(i)$.
2. Evaluate q_i based on Eq. (5).
   If $q_i \le 0.5$, decode $\hat{v}_i = y_i$.
   Else, decode $\hat{v}_i = y_i \oplus 1$.

This decoding method, known as "a posteriori probability (APP) decoding," was introduced by Massey [25].

3.1.2. Iterative Decoding. We observe that either majority-logic decoding, or APP decoding, provides for each of the N initial inputs a new estimate of this quantity, based on exactly the same set of constraints. This is due to the fact that all EG, extended EG and PG codes presented in Section 2 are one-step majority-logic decodable codes. Consequently, this observation suggests a straightforward heuristic approach; we may use these estimates as new inputs and iterate the decoding process. If for $0 \le i \le N-1$, $x_i^{(0)}$ represents the a priori information about position i and for $l \ge 1$, and $x_i^{(l)}$ represents the a posteriori estimate of position i evaluated at iteration l, then iterative decoding is achieved on the basis of:

$$x_i^{(l)} = x_i^{(0)} + f_i(\mathbf{x}^{(l-1)}) \tag{6}$$

where $f_i()$ represents the function used to compute the a posteriori estimate of position i from the a priori inputs and $\mathbf{x}^{(l-1)}$ represents the vector of a posteriori values evaluated at iteration $(l-1)$. We notice that (6) implicitly assumes an iterative updating of the original a priori values on the basis of the latest a posteriori estimates. Iterative majority-logic decoding (also known as "iterative bit flipping" decoding) and iterative APP decoding approaches were first proposed [24] to decode LDPC codes. Refined versions of this heuristic approach for some EG, extended EG and PG codes have been proposed [14,15].

Although these iterative decoding methods allow one to achieve significant performance improvements on the first iteration, they fall short of optimum decoding. The main reason is the introduction and propagation of correlated values from the second iteration, which is readily explained as follows. Let us assume that positions i and j contribute to the same checksum. Then, according to (6), for $l \ge 2$, $x_j^{(l-1)}$ depends on $x_i^{(0)}$ and consequently, $x_i^{(0)}$ and $f_i(\mathbf{x}^{(l-1)})$ are correlated when evaluating $x_i^{(l)}$.

This problem can be overcome by computing for each position i as many a posteriori values as checksums intersecting on that position. This method, implicitly contained in the performance analysis of Ref. 24, is also known as "belief propagation" (BP) [26]. A good presentation of BP decoding can be found in Ref. 27. BP decoding of long LDPC codes has been showed to closely approach the Shannon capacity of the BSC and AWGN channel [28,29]. For $0 \le i \le N-1$, let $x_i^{(0)}$ represent the a priori information about position i and for $1 \le j \le |B(i)|$, and $l \ge 1$ let $x_{j,i}^{(l)}$ represent the a posteriori estimate of position i computed at iteration l based on the checksums other than checksum j in $B(i)$. Then at iteration l, BP decoding evaluates

$$x_{j,i}^{(l)} = x_i^{(0)} + f_{j,i}(\mathbf{x}^{(l-1)}(i)) \tag{7}$$

where $f_{j,i}()$ represents the function used to compute the a posteriori estimate of position i from the a priori inputs other than that in checksum j and $\mathbf{x}^{(l-1)}(i)$ represents the vector of a posteriori values evaluated at iteration $(l-1)$ by discarding the check sums containing position i. BP iterative decoding achieves optimum decoding if the Tanner graph representation [30] of the code considered contains no loop.

BP decoding of EG, extended EG and PG codes has been investigated [16] and its application to the LDPC codes presented in Section 2.3 was presented in a 1999

paper [17]. Despite the presence of loops in the Tanner graph of all these codes, near-optimum performance can still be achieved.

3.2. General Decoding of EG and PG Codes

One-stage decoding of EG and PG codes in general is achieved by repeating the method described in Section 3.1.1 in at most μ steps. At each step, checksums with less and less common positions are estimated until at most only one common position between any two checksums remains, as in Section 3.1.1. A good description of majority-logic decoding of EG and PG codes can be found in Chaps. 7 and 8 of Ref. 23.

On the other hand, iterative decoding of EG and PG codes in general is not as straightforwardly generalizable. In general, a multistep decoding method is also required to obtain the a posteriori error probabilities of all N bits, which inherently complicates the application of iterative decoding.

BIOGRAPHY

Marc P.C. Fossorier was born in Annemasse, France, on March 8, 1964. He received the B.E. degree from the National Institute of Applied Sciences (I.N.S.A.) Lyon, France, in 1987, and the M.S. and Ph.D. degrees from the University of Hawaii at Manoa, Honolulu, in 1991 and 1994, all in electrical engineering.

In 1996, he joined the Faculty of the University of Hawaii, Honolulu, as an Assistant Professor of Electrical Engineering. He was promoted to Associate Professor in 1999.

His research interests include decoding techniques for linear codes, communication algorithms, combining coding and equalization for ISI channels, magnetic recording and statistics. He coauthored (with S. Lin, T. Kasami, and T. Fujiwara) the book *Trellises and Trellis-Based Decoding Algorithms*, (Kluwer Academic Publishers, 1998).

Dr. Fossorier is a recipient of a 1998 NSF Career Development award. He has served as editor for the *IEEE Transactions on Communications* since 1996, as editor for the *IEEE Communications Letters* since 1999, and he currently is the treasurer of the IEEE Information Theory Society. Since 2002, he has also been a member of the Board of Governors of the IEEE Information Theory Society.

BIBLIOGRAPHY

1. L. D. Rudolph, *Geometric Configuration and Majority Logic Decodable Codes*, M.E.E. thesis, Univ. Oklahoma, Norman, 1964.
2. L. D. Rudolph, A class of majority logic decodable codes, *IEEE Trans. Inform. Theory* **13**: 305–307 (April 1967).
3. P. Delsarte, A geometric approach to a class of cyclic codes, *J. Combin. Theory* **6**: 340–358 (1969).
4. P. Delsarte, J. M. Goethals, and J. MacWilliams, On GRM and related codes, *Inform. Control* **16**: 403–442 (July 1970).
5. S. Lin, Shortened finite geometry codes, *IEEE Trans. Inform. Theory* **18**: 692–696 (July 1972).
6. S. Lin, Multifold Euclidean geometry codes, *IEEE Trans. Inform. Theory* **19**: 537–548 (July 1973).
7. C. R. P. Hartmann, J. B. Ducey, and L. D. Rudolph, On the structure of generalized finite geometry codes, *IEEE Trans. Inform. Theory* **20**: 240–252 (March 1974).
8. S. Lin and K. P. Yiu, An improvement to multifold euclidean geometry codes, *Inform. Control* **28**: (July 1975).
9. E. J. Weldon, Difference-set cyclic codes, *Bell Syst. Tech. J.* **45**: 1045–1055 (Sept. 1966).
10. D. J. Costello, Jr., J. Hagenauer, H. Imai, and S. B. Wicker, Applications of error-control coding, *Commemorative Issue, IEEE Trans. Inform. Theory* **44**: 2531–2560 (Oct. 1998).
11. I. S. Reed, A class of multiple-error-correcting codes and the decoding scheme, *IRE Trans. Inform. Theory* **4**: 38–49 (Sept. 1954).
12. C. L. Chen, On majority-logic decoding of finite geometry codes, *IEEE Trans. Inform. Theory* **17**: 332–336 (May 1971).
13. T. Kasami and S. Lin, On majority-logic decoding for duals of primitive polynomial codes, *IEEE Trans. Inform. Theory* **17**: 322–331 (May 1971).
14. K. Yamaguchi, H. Iizuka, E. Nomura, and H. Imai, Variable threshold soft decision decoding, *IEICE Trans. Elect. Commun.* **72**: 65–74 (Sept. 1989).
15. R. Lucas, M. Bossert, and M. Breitbach, On iterative soft-decision decoding of linear binary block codes and product codes, *IEEE J. Select. Areas Commun.* **16**: 276–296 (Feb. 1998).
16. R. Lucas, M. Fossorier, Y. Kou, and S. Lin, Iterative decoding of one-step majority logic decodable codes based on belief propagation, *IEEE Trans. Commun.* **48**: 931–937 (June 2000).
17. Y. Kou, S. Lin, and M. Fossorier, Low density parity check codes based on finite geometries: A rediscovery and more, *IEEE Trans. Inform. Theory* (Oct. 1999).
18. H. B. Mann, *Analysis and Design of Experiments*, Dover, New York, 1949.
19. R. D. Carmichael, *Introduction to the Theory of Groups of Finite Order*, Dover, New York, 1956.
20. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2nd ed., MIT Press, Cambridge, MA, 1972.
21. I. F. Blake and R. C. Mullin, *The Mathematical Theory of Coding*, Academic Press, New York, 1975.
22. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland Mathematical Library, Amsterdam, 1977.
23. S. Lin and D. J. Costello, Jr., *Error Control Coding, Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
24. R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.
25. J. L. Massey, *Threshold Decoding*, MIT Press, Cambridge, MA, 1963.
26. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA, 1988.
27. D. J. C. MacKay, Good error-correcting codes based on very sparse matrices, *IEEE Trans. Inform. Theory* **45**: 399–432 (March 1999).
28. T. Richardson, A. Shokrollahi, and R. Urbanke, Design of probably good low-density parity check codes, *IEEE Trans. Inform. Theory* (in press).

29. S. Y. Chung, G. D. Forney, Jr., T. Richardson, and R. Urbanke, On the design of low-density parity-check codes within 0.0051 dB of the Shannon limit, *IEEE Commun. Lett.* (in press).

30. R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **27**: 533–547 (Sept. 1981).

FREQUENCY AND PHASE MODULATION

MASOUD SALEHI
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Analog angle modulation methods are modulation methods in which the information is carried by the phase of a sinusoidal. Angle modulation can be carried out either by frequency modulation or by phase modulation. In frequency modulation (FM) systems, the frequency of the carrier f_c is changed by the message signal, and in phase modulation (PM) systems the phase of the carrier is changed according to the variations in the message signal. Frequency and phase modulation are obviously quite nonlinear, and very often they are jointly referred to as *angle modulation methods*. As we will see, angle modulation, due to its inherent nonlinearity, is rather difficult to analyze. In many cases only an approximate analysis can be done. Another property of angle modulation is its bandwidth expansion property. Frequency and phase modulation systems generally expand the bandwidth such that the effective bandwidth of the modulated signal is usually many times the bandwidth of the message signal.[1] With a higher implementation complexity and a higher bandwidth occupancy, one would naturally raise a question as to the usefulness of these systems. As we will show, the major benefit of these systems is their high degree of noise immunity. In fact, these systems trade off bandwidth for high noise immunity. That is the reason why FM systems are widely used in high-fidelity music broadcasting and point-to-point communication systems where the transmitter power is quite limited. Another benefit of theses systems is their constant envelop property. Unlike amplitude modulation, these systems have a constant amplitude which makes them attractive choices to use with nonlinear amplification devices (class C amplifiers or TWT devices).

FM radio broadcasting was first initiated by Edwin H. Armstrong in 1935 in New York on a frequency of 42.1 MHz. At that time when FM broadcasting started. Later with the advent of TV broadcasting the 88–108-MHz frequency band was used for FM broadcasting. Commercial stereo FM broadcasting started in Chicago in 1961.

[1] Strictly speaking, the bandwidth of the modulated signal, as will be shown later, is infinite. That is why we talk about the *effective bandwidth*.

2. REPRESENTATION OF FM AND PM SIGNALS

An angle modulated signal in general can be expressed as

$$u(t) = A_c \cos(\theta(t))$$

$\theta(t)$ is the *phase* of the signal and its *instantaneous frequency* $f_i(t)$ is given by

$$f_i(t) = \frac{1}{2\pi}\frac{d}{dt}\theta(t) \tag{1}$$

Since $u(t)$ is a bandpass signal it can be represented as

$$u(t) = A_c \cos(2\pi f_c t + \phi(t)) \tag{2}$$

and therefore

$$f_i(t) = f_c + \frac{1}{2\pi}\frac{d}{dt}\phi(t) \tag{3}$$

If $m(t)$ is the message signal, then in a PM system, we have

$$\phi(t) = k_p m(t) \tag{4}$$

and in an FM system we have

$$f_i(t) - f_c = k_f m(t) = \frac{1}{2\pi}\frac{d}{dt}\phi(t) \tag{5}$$

where k_p and k_f are phase and frequency *deviation constants*. From the above relationships we have

$$\phi(t) = \begin{cases} k_p m(t), & \text{PM} \\ 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau, & \text{FM} \end{cases} \tag{6}$$

This expression shows the close relation between FM and PM. This close relationship makes it possible to analyze these systems in parallel and only emphasize their main differences. The first interesting result observed from the above is that if we phase modulate the carrier with the integral of a message, it is equivalent to frequency modulation of the carrier with the original message. On the other hand, the above relation can be expressed as

$$\frac{d}{dt}\phi(t) = \begin{cases} k_p \dfrac{d}{dt} m(t), & \text{PM} \\ 2\pi m(t), & \text{FM} \end{cases} \tag{7}$$

which shows that if we frequency-modulate the carrier with the derivative of a message the result is equivalent to phase modulation of the carrier with the message itself. Figure 1 shows the above relation between FM and PM. Figure 2 illustrates a square-wave signal and its integral, a sawtooth signal, and their corresponding FM and PM signals.

The demodulation of an FM signal involves finding the instantaneous frequency of the modulated signal and then subtracting the carrier frequency from it. In the demodulation of PM, the demodulation process is done by finding the phase of the signal and then recovering $m(t)$. The *maximum phase deviation* in a PM system is given by

$$\Delta\phi_{\max} = k_p \max[|m(t)|] \tag{8}$$

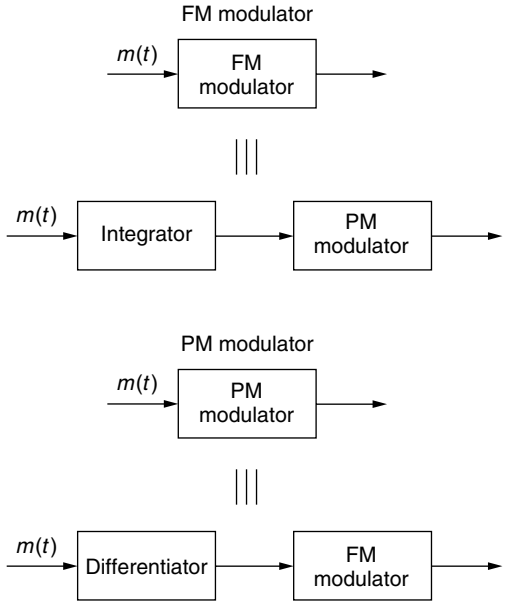


Figure 1. A comparison of frequency and phase modulators.

and the *maximum frequency deviation* in an FM system is given by

$$\Delta f_{\max} = k_f \max[|m(t)|] \tag{9}$$

We now define the *modulation index* for a general nonsinusoidal signal $m(t)$ as

$$\beta_p = k_p \max[|m(t)|] \tag{10}$$

$$\beta_f = \frac{k_f \max[|m(t)|]}{W} \tag{11}$$

where W denotes the bandwidth of the message signal $m(t)$. In terms of the maximum phase and frequency deviation $\Delta\phi_{\max}$ and $\Delta f_{\max}$, we have

$$\beta_p = \Delta\phi_{\max} \tag{12}$$

$$\beta_f = \frac{\Delta f_{\max}}{W} \tag{13}$$

2.1. Narrowband Angle Modulation

If in an angle modulation[2] system the deviation constants k_p and k_f and the message signal $m(t)$ are such that for all t we have $\phi(t) \ll 1$, then we can use a simple approximation to expand $u(t)$ as

$$\begin{aligned} u(t) &= A_c \cos 2\pi f_c t \cos\phi(t) - A_c \sin 2\pi f_c t \sin\phi(t) \\ &\approx A_c \cos 2\pi f_c t - A_c \phi(t) \sin 2\pi f_c t \end{aligned} \tag{14}$$

This last equation shows that in this case the modulated signal is very similar to a conventional AM signal of the form $A_c(1 + m(t))\cos(2\pi f_c t)$. The bandwidth of this signal is similar to the bandwidth of a conventional AM signal, which is twice the bandwidth of the message signal. Of course, this bandwidth is only an approximation to the real bandwidth of the FM signal. A phasor diagram for this signal and the comparable conventional AM signal are given in Fig. 3. Note that compared to conventional AM, the narrowband angle modulation scheme has far less amplitude variations. Of course the angle modulation system has constant amplitude and, hence, there should be no amplitude variations in the phasor diagram representation of the system. The slight variations here are due to the first-order approximation that we have used for the expansions of $\sin(\phi(t))$ and $\cos(\phi(t))$. As we will see later, the narrowband angle modulation method does not provide any better noise immunity compared to

[2] Also known as *low-index angle modulation*.

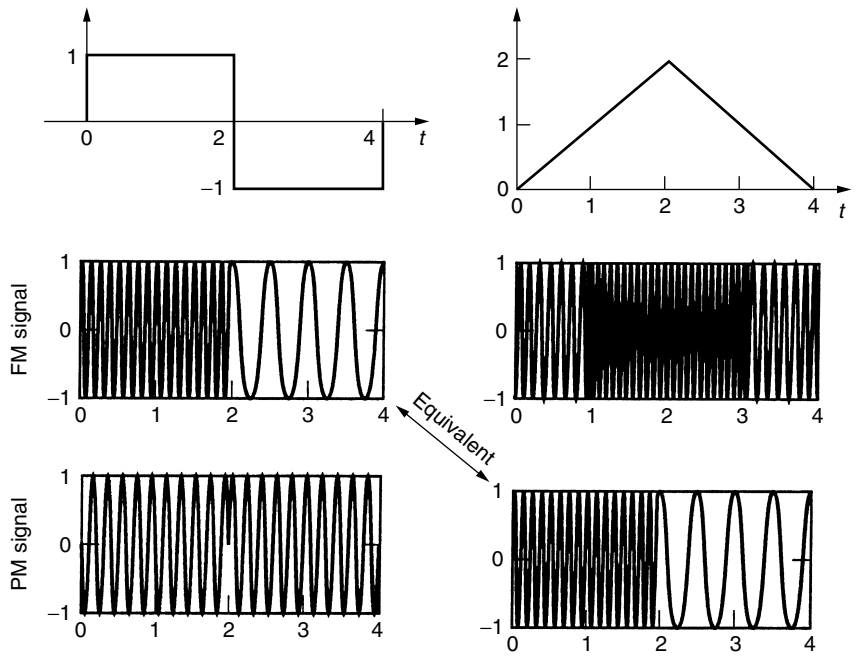


Figure 2. Frequency and phase modulation of square and sawtooth waves.

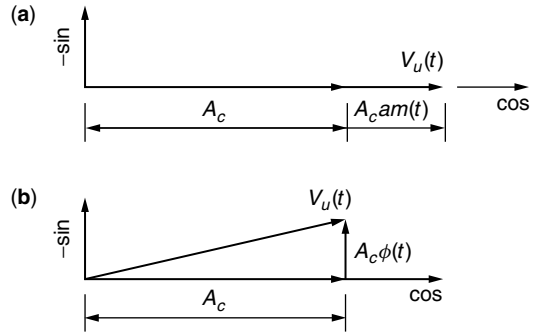


Figure 3. Phasor diagram for the conventional AM and narrowband angle modulation.

a conventional AM system, and therefore, narrowband angle modulation is not used in broadcasting. However, these systems can be used as an intermediate stage for generation of wideband angle-modulated signals, as we will discuss in Section 4.

3. SPECTRAL CHARACTERISTICS OF ANGLE-MODULATED SIGNALS

Because of the inherent nonlinearity of angle modulation systems, the precise characterization of their spectral properties, even for simple message signals, is mathematically intractable. Therefore, the derivation of the spectral characteristics of these signals usually involves the study of very simple modulating signals and certain approximations. Then, the results are generalized to the more complicated messages. We will study the spectral characteristics of an angle-modulated signal when the modulating signal is a sinusoidal signal and when the modulating signal is a general periodic signal. Then we generalize these results to an arbitrary modulating signal.

3.1. Angle Modulation by a Sinusoidal Signal

Let us begin with the case where the message signal is a sinusoidal signal $m(t) = a\cos(2\pi f_m t)$. In PM we have

$$\phi(t) = k_p m(t) = k_p a\cos(2\pi f_m t) \tag{15}$$

and in FM we have

$$\phi(t) = 2\pi k_f \int_{-\infty}^{t} m(\tau)\,d\tau = \frac{k_f a}{f_m}\sin(2\pi f_m t) \tag{16}$$

Therefore, the modulated signals will be

$$u(t) = \begin{cases} A_c\cos(2\pi f_c t + k_p a\cos(2\pi f_m t)), & \text{PM} \\ A_c\cos\left(2\pi f_c t + \dfrac{k_f a}{f_m}\sin(2\pi f_m t)\right), & \text{FM} \end{cases} \tag{17}$$

and the modulation indices are

$$\beta_p = k_p a \tag{18}$$

$$\beta_f = \frac{k_f a}{f_m} \tag{19}$$

Using β_p and β_f, we have

$$u(t) = \begin{cases} A_c\cos(2\pi f_c t + \beta_p\cos(2\pi f_m t)), & \text{PM} \\ A_c\cos(2\pi f_c t + \beta_f\sin(2\pi f_m t)), & \text{FM} \end{cases} \tag{20}$$

As shown above, the general form of an angle-modulated signal for the case of a sinusoidal message is

$$u(t) = A_c\cos(2\pi f_c t + \beta\sin 2\pi f_m t) \tag{21}$$

where β is the modulation index that can be either β_p or β_f. Therefore the modulated signal can be written as

$$u(t) = \text{Re}\,(A_c e^{j2\pi f_c t} e^{j\beta\sin 2\pi f_m t}) \tag{22}$$

Since $\sin 2\pi f_m t$ is periodic with period $T_m = \dfrac{1}{f_m}$, the same is true for the complex exponential signal

$$e^{j\beta\sin 2\pi f_m t}$$

Therefore, it can be expanded in a Fourier series representation. The Fourier series coefficients are obtained from the integral

$$\begin{aligned} c_n &= f_m \int_0^{\frac{1}{f_m}} e^{j\beta\sin 2\pi f_m t} e^{-jn2\pi f_m t}\,dt \\ &\overset{u=2\pi f_m t}{=} \frac{1}{2\pi}\int_0^{2\pi} e^{j\beta(\sin u - nu)}\,du \end{aligned} \tag{23}$$

This latter integral is a well-known integral known as the *Bessel function of the first kind of order n* and is denoted by $J_n(\beta)$. Therefore, we have the Fourier series for the complex exponential as

$$e^{j\beta\sin 2\pi f_m t} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi n f_m t} \tag{24}$$

By substituting (24) in (22), we obtain

$$\begin{aligned} u(t) &= \text{Re}\left(A_c\sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi n f_m t} e^{j2\pi f_c t}\right) \\ &= \sum_{n=-\infty}^{\infty} A_c J_n(\beta)\cos(2\pi(f_c + nf_m)t) \end{aligned} \tag{25}$$

This relation shows that even in this very simple case where the modulating signal is a sinusoid of frequency f_m, the angle-modulated signal contains all frequencies of the form $f_c + nf_m$ for $n = 0, \pm 1, \pm 2, \ldots$. Therefore, the actual bandwidth of the modulated signal is infinite. However, the amplitude of the sinusoidal components of frequencies $f_c \pm nf_m$ for large n is very small and their contribution to the total power in the signal is low. Hence, we can define an finite *effective bandwidth* for the modulated signal, as the bandwidth that contains the component frequencies that contain a certain percentage (usually 98% or 99%) of

the total power in the signal. A series expansion for the Bessel function is given by

$$J_n(\beta) = \sum_{k=0}^{\infty} \frac{(-1)^k \left(\frac{\beta}{2}\right)^{n+2k}}{k!(k+n)!} \tag{26}$$

The expansion here shows that for small β, we can use the approximation

$$J_n(\beta) \approx \frac{\beta^n}{2^n n!} \tag{27}$$

Thus for a small modulation index β, only the first sideband corresponding to $n = 1$ is of importance. Also, using the expansion in (26), it is easy to verify the following symmetry properties of the Bessel function.

$$J_{-n}(\beta) = \begin{cases} J_n(\beta), & n \text{ even} \\ -J_n(\beta), & n \text{ odd} \end{cases} \tag{28}$$

Plots of $J_n(\beta)$ for various values of n are given in Fig. 4, and a tabulation of the values of the Bessel function is given in Table 1. The underlined and doubly underlined entries for each β indicate the minimum value of n that guarantees the signal will contain at least 70% or 98% of the total power, respectively.

In general the effective bandwidth of an angle-modulated signal, which contains at least 98% of the signal power, is approximately given by the relation

$$B_c = 2(\beta + 1)f_m \tag{29}$$

where β is the modulation index and f_m is the frequency of the sinusoidal message signal.

It is instructive to study the effect of the amplitude and frequency of the sinusoidal message signal on the bandwidth and the number of harmonics in the modulated signal. Let the message signal be given by

$$m(t) = a\cos(2\pi f_m t) \tag{30}$$

The bandwidth[3] of the modulated signal is given by

$$B_c = 2(\beta + 1)f_m = \begin{cases} 2(k_p a + 1)f_m, & \text{PM} \\ 2\left(\frac{k_f a}{f_m} + 1\right)f_m, & \text{FM} \end{cases} \tag{31}$$

or

$$B_c = \begin{cases} 2(k_p a + 1)f_m, & \text{PM} \\ 2(k_f a + f_m), & \text{FM} \end{cases} \tag{32}$$

This relation shows that increasing a, the amplitude of the modulating signal, in PM and FM has almost the same effect on increasing the bandwidth B_c. On the other hand, increasing f_m, the frequency of the message signal, has a more profound effect in increasing the bandwidth of a PM signal as compared to an FM signal. In both PM and FM the bandwidth B_c increases by increasing f_m, but in PM this increase is a proportional increase and in FM this is only an additive increase which in most cases of interest, (for large β) is not substantial. Now if we look at the number of harmonics in the bandwidth (including the carrier) and denote it with M_c, we have

$$M_c = 2\lfloor\beta\rfloor + 3 = \begin{cases} 2\lfloor k_p a\rfloor + 3, & \text{PM} \\ 2\left\lfloor \frac{k_f a}{f_m} \right\rfloor + 3, & \text{FM} \end{cases} \tag{33}$$

Increasing the amplitude a increases the number of harmonics in the bandwidth of the modulated signal in both cases. However, increasing f_m, has no effect on the number of harmonics in the bandwidth of the PM signal and decreases the number of harmonics in the FM signal almost linearly. This explains the relative insensitivity of the bandwidth of the FM signal to the message frequency. On one hand, increasing f_m decreases the number of harmonics in the bandwidth and, at the same time, increases the spacing between the harmonics. The net effect is a slight increase in the bandwidth. In

[3] From now on, by bandwidth we mean effective bandwidth unless otherwise stated.

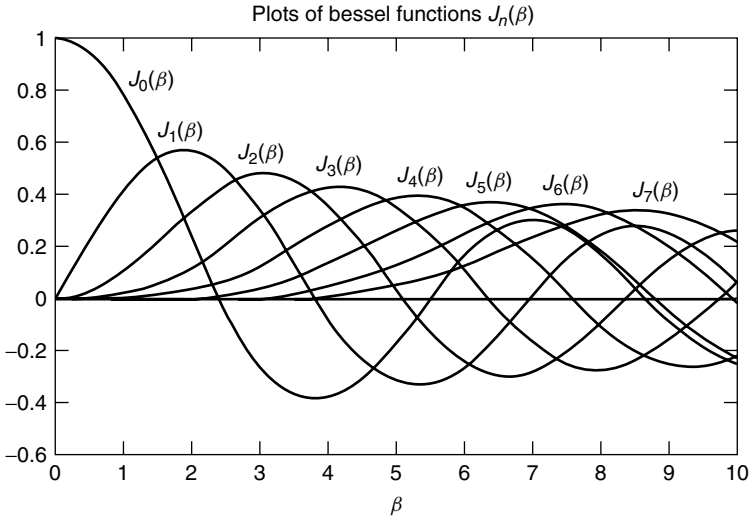


Figure 4. Bessel functions for various values of n.

Table 1. Table of Bessel Function Values

n	$\beta = 0.1$	$\beta = 0.2$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	$\beta = 5$	$\beta = 8$	$\beta = 10$
0	0.997	0.990	0.938	0.765	0.224	−0.178	0.172	−0.246
1	0.050	0.100	0.242	0.440	0.577	−0.328	0.235	0.043
2	0.001	0.005	0.031	0.115	0.353	0.047	−0.113	0.255
3	—	—	—	0.020	0.129	0.365	−0.291	0.058
4	—	—	—	0.002	0.034	0.391	−0.105	−0.220
5	—	—	—	—	0.007	0.261	0.186	−0.234
6	—	—	—	—	0.001	0.131	0.338	−0.014
7	—	—	—	—	—	0.053	0.321	0.217
8	—	—	—	—	—	0.018	0.223	0.318
9	—	—	—	—	—	0.006	0.126	0.292
10	—	—	—	—	—	0.001	0.061	0.207
11	—	—	—	—	—	—	0.026	0.123
12	—	—	—	—	—	—	0.010	0.063
13	—	—	—	—	—	—	0.003	0.029
14	—	—	—	—	—	—	0.001	0.012
15	—	—	—	—	—	—	—	0.005
16	—	—	—	—	—	—	—	0.002

Source: From Ziemer and Tranter (1990) © Houghton Mifflin, reprinted by permission.

PM, however, the number of harmonics remains constant and only the spacing between them increases. Therefore, the net effect is a linear increase in bandwidth. Figure 5 shows the effect of increasing the frequency of the message in both FM and PM.

3.2. Angle Modulation by a Periodic Message Signal

To generalize the abovementioned results, we now consider angle modulation by an arbitrary periodic message signal $m(t)$. Let us consider a PM modulated signal where

$$u(t) = A_c \cos(2\pi f_c t + \beta m(t)) \tag{34}$$

We can write this as

$$u(t) = A_c \mathrm{Re}\,[e^{j2\pi f_c t} e^{j\beta m(t)}] \tag{35}$$

We are assuming that $m(t)$ is periodic with period $T_m = \frac{1}{f_m}$. Therefore $e^{j\beta m(t)}$ will be a periodic signal with the same period and we can find its Fourier series expansion as

$$e^{j\beta m(t)} = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_m t} \tag{36}$$

where

$$c_n = \frac{1}{T_m} \int_0^{T_m} e^{j\beta m(t)} e^{-j2\pi n f_m t}\, dt \overset{u=2\pi f_m t}{=} \frac{1}{2\pi} \int_0^{2\pi} e^{j\left[\beta m\left(\frac{u}{2\pi f_m}\right) - nu\right]} du \tag{37}$$

and

$$u(t) = A_c \mathrm{Re}\left[\sum_{n=-\infty}^{\infty} c_n e^{j2\pi f_c t} e^{j2\pi n f_m t}\right] = A_c \sum_{n=-\infty}^{\infty} |c_n| \cos(2\pi(f_c + n f_m)t + \angle c_n) \tag{38}$$

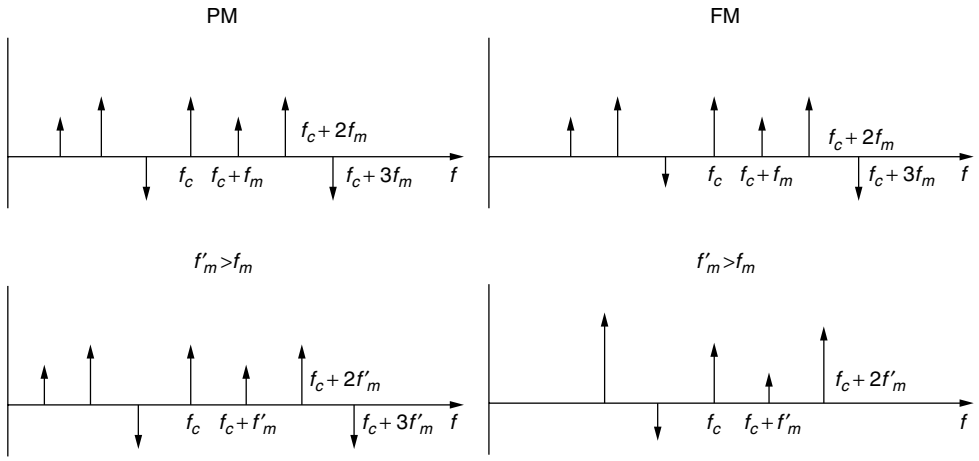


Figure 5. The effect of doubling bandwidth of the message in FM and PM.

It is seen again that the modulated signal contains all frequencies of the form $f_c + nf_m$.

The detailed treatment of the spectral characteristics of an angle-modulated signal for a general nonperiodic deterministic message signal $m(t)$ is quite involved because of the nonlinear nature of the modulation process. However, there exists an approximate relation for the effective bandwidth of the modulated signal, known as the *Carson's rule*, and given by

$$B_c = 2(\beta + 1)W \tag{39}$$

where β is the modulation index defined as

$$\beta = \begin{cases} k_p \max[|m(t)|], & \text{PM} \\ \dfrac{k_f \max[|m(t)|]}{W}, & \text{FM} \end{cases} \tag{40}$$

and W is the bandwidth of the message signal $m(t)$. Since in wideband FM the value of β is usually around ≥ 5, it is seen that the bandwidth of an angle-modulated signal is much greater than the bandwidth of various amplitude modulation schemes, which is either W (in SSB) or $2W$ (in DSB or conventional AM).

4. IMPLEMENTATION OF ANGLE MODULATORS AND DEMODULATORS

Angle modulators are in general time-varying and nonlinear systems. One method for generating an FM signal directly is to design an oscillator whose frequency changes with the input voltage. When the input voltage is zero the oscillator generates a sinusoid with frequency f_c, and when the input voltage changes, this frequency changes accordingly. There are two approaches to design such an oscillator, usually called a VCO or *voltage-controlled oscillator*. One approach is to use a *varactor diode*. A varactor diode is a capacitor whose capacitance changes with the applied voltage. Therefore, if this capacitor is used in the tuned circuit of the oscillator and the message signal is applied to it, the frequency of the tuned circuit, and the oscillator, will change in accordance with the message signal. Let us assume that the inductance of the inductor in the tuned circuit of Fig. 6 is L_0 and the capacitance of the varactor diode is given by

$$C(t) = C_0 + k_0 m(t) \tag{41}$$

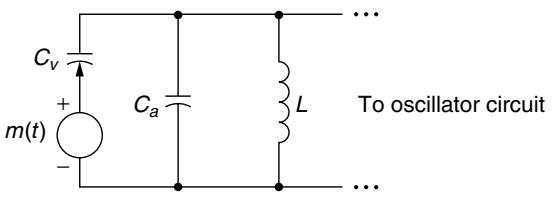


Figure 6. Varactor diode implementation of an angle modulator.

When $m(t) = 0$, the frequency of the tuned circuit is given by $f_c = \dfrac{1}{2\pi\sqrt{L_0 C_0}}$. In general, for nonzero $m(t)$, we have

$$\begin{aligned} f_i(t) &= \frac{1}{\pi\sqrt{L_0(C_0 + k_0 m(t))}} \\ &= \frac{1}{2\pi\sqrt{L_0 C_0}} \frac{1}{\sqrt{1 + \dfrac{k_0}{C_0} m(t)}} \\ &= f_c \frac{1}{\sqrt{1 + \dfrac{k_0}{C_0} m(t)}} \end{aligned} \tag{42}$$

Assuming that

$$\varepsilon = \frac{k_0}{C_0} m(t) \ll 1$$

and using the approximations

$$\sqrt{1+\varepsilon} \approx 1 + \frac{\varepsilon}{2} \tag{43}$$

$$\frac{1}{1+\varepsilon} \approx 1 - \varepsilon \tag{44}$$

we obtain

$$f_i(t) \approx f_c \left(1 - \frac{k_0}{2C_0} m(t)\right) \tag{45}$$

which is the relation for a frequency-modulated signal.

A second approach for generating an FM signal is by use of a *reactance tube*. In the reactance tube implementation an inductor whose inductance varies with the applied voltage is employed and the analysis is very similar to the analysis presented for the varactor diode. It should be noted that although we described these methods for generation of FM signals, due to the close relation between FM and PM signals, basically the same methods can be applied for generation of PM signals (see Fig. 1).

4.1. Indirect Method for Generation of Angle-Modulated Signals

Another approach for generating an angle-modulated signal is to first generate a narrowband angle-modulated signal and then change it to a wideband signal. This method is usually known as the *indirect method* for generation of FM and PM signals. Because of the similarity of conventional AM signals, generation of narrowband angle modulated signals is straightforward. In fact, any modulator for conventional AM generation can be easily modified to generate a narrowband angle modulated signal. Figure 7 is a block diagram of a narrowband angle modulator. The next step is to use the narrowband angle-modulated signal to generate a wideband angle-modulated signal. Figure 8 is a block diagram of a system that generates wideband angle-modulated signals from narrowband angle-modulated signals. The first stage of such a system is, of course, a narrowband angle modulator such as the one shown in Fig. 7. The narrowband angle-modulated signal enters a frequency multiplier that multiplies the instantaneous frequency of the input by some constant n. This is usually done by applying the

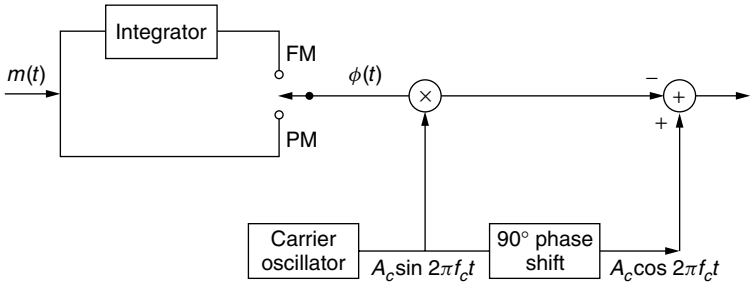


Figure 7. Generation of narrowband angle-modulated signal.

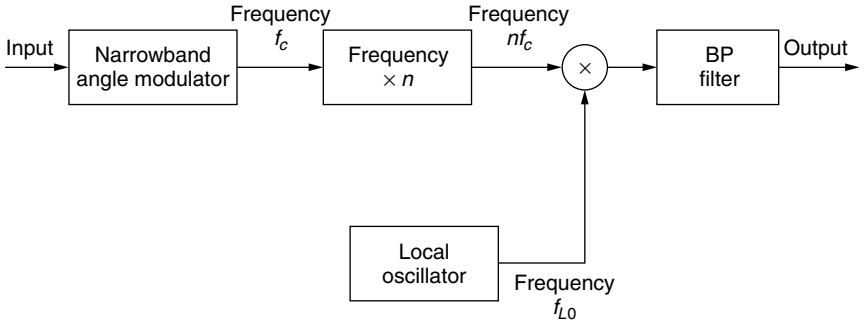


Figure 8. Indirect generation of angle-modulated signals.

input signal to a nonlinear element and then passing its output through a bandpass filter tuned to the desired central frequency. If the narrowband modulated signal is represented by

$$u_n(t) = A_c \cos(2\pi f_c t + \phi(t)) \tag{46}$$

the output of the frequency multiplier (output of the bandpass filter) is given by

$$y(t) = A_c \cos(2\pi n f_c t + n\phi(t)) \tag{47}$$

In general, this is, of course, a wideband angle-modulated signal. However, there is no guarantee that the carrier frequency of this signal, nf_c, will be the desired carrier frequency. The last stage of the modulator performs an up/down conversion to shift the modulated signal to the desired center frequency. This stage consists of a mixer and a bandpass filter. If the frequency of the local oscillator of the mixer is f_{LO} and we are using a down converter, the final wideband angle modulated signal is given by

$$u(t) = A_c \cos(2\pi (nf_c - f_{\text{LO}})t + n\phi(t)) \tag{48}$$

Since we can freely choose n and f_{LO}, we can generate any modulation index at any desired carrier frequency by this method.

4.2. FM Demodulators

FM demodulators are implemented by generating an AM signal whose amplitude is proportional to the instantaneous frequency of the FM signal, and then using an AM demodulator to recover the message signal. To implement the first step, specifically, transforming the FM signal into an AM signal, it is enough to pass the FM signal through an LTI system whose frequency response is approximately a straight line in the frequency band of the FM signal. If the frequency response of such a system is given by

$$|H(f)| = V_0 + k(f - f_c) \quad \text{for } |f - f_c| < \frac{B_c}{2} \tag{49}$$

and if the input to the system is

$$u(t) = A_c \cos\left(2\pi f_c t + 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau\right) \tag{50}$$

then, the output will be the signal

$$v_o(t) = A_c(V_0 + kk_f m(t)) \cos\left(2\pi f_c t + 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau\right) \tag{51}$$

The next steps is to demodulate this signal to obtain $A_c(V_0 + kk_f m(t))$, from which the message $m(t)$ can be recovered. Figure 9 is a block diagram of these two steps.

There exist many circuits that can be used to implement the first stage of an FM demodulator, i.e., FM to AM conversion. One such candidate is a simple differentiator with

$$|H(f)| = 2\pi f \tag{52}$$

Another candidate is the rising half of the frequency characteristics of a tuned circuit as shown in Fig. 10. Such

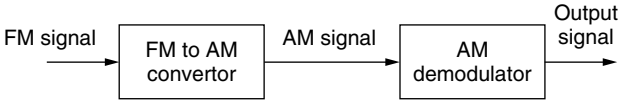


Figure 9. A general FM demodulator.

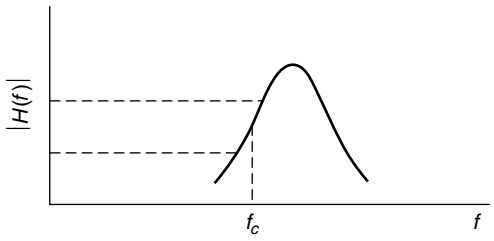


Figure 10. A tuned circuit used in an FM demodulator.

a circuit can be easily implemented but usually the linear region of the frequency characteristic may not be wide enough. To obtain a linear characteristics over a wider range of frequencies, usually two circuits tuned at two frequencies f_1 and f_2 are connected in a configuration which is known as a *balanced discriminator*. A balanced discriminator with the corresponding frequency characteristics is shown in Fig. 11.

The FM demodulation methods described above that transform the FM signal into an AM signal have a bandwidth equal to the channel bandwidth B_c occupied by the FM signal. Consequently, the noise that is passed by the demodulator is the noise contained within B_c.

4.2.1. FM Demodulation with Feedback. A totally different approach to FM signal demodulation is to use feedback in the FM demodulator to narrow the bandwidth of the FM detector and, as will be seen later, to reduce the noise power at the output of the demodulator. Figure 12 illustrates a system in which the FM discrimination is placed in the feedback branch of a feedback system that employs a voltage-controlled oscillator (VCO) path. The bandwidth of the discriminator and the subsequent lowpass filter is designed to match the bandwidth of the message signal $m(t)$. The output of the lowpass filter is the desired message signal. This type of FM demodulator is called an FM demodulator with feedback (FMFB). An alternative to FMFB demodulator is the use of a phase-locked loop (PLL), as shown in Fig. 13. The input to the PLL is the angle-modulated signal (we neglect the presence of noise in this discussion)

$$u(t) = A_c \cos[2\pi f_c t + \phi(t)] \tag{53}$$

where, for FM, we obtain

$$\phi(t) = 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau \tag{54}$$

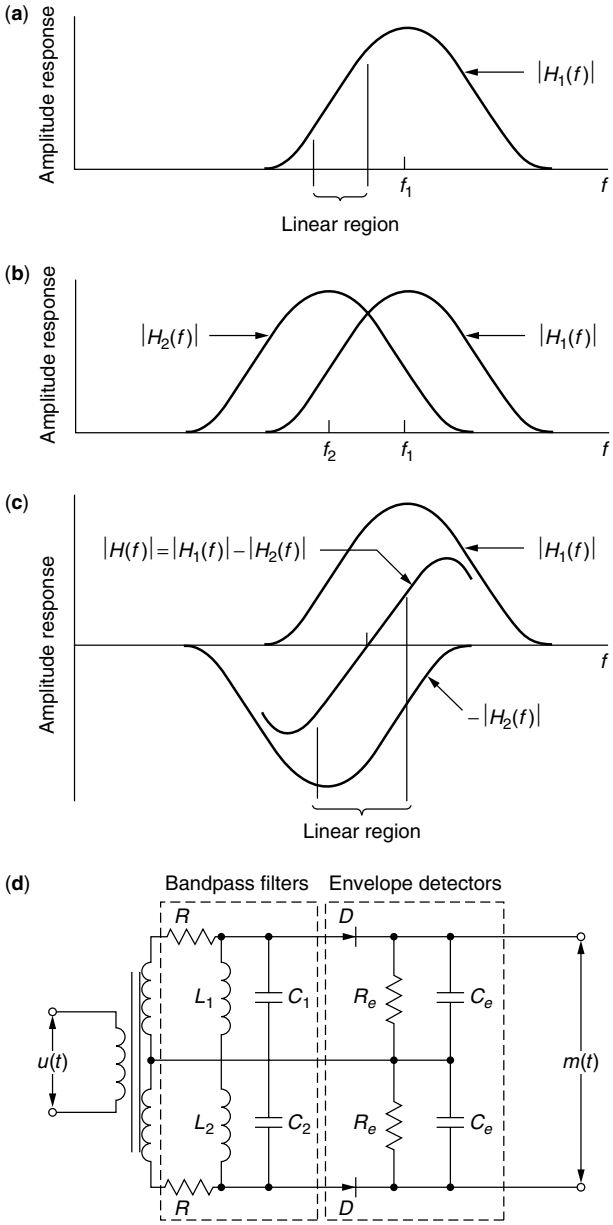


Figure 11. A ratio detector demodulator and the corresponding frequency response.

The VCO generates a sinusoid of a fixed frequency, in this case the carrier frequency f_c, in the absence of an input control voltage.

Now, suppose that the control voltage to the VCO is the output of the loop filter, denoted as $v(t)$. Then, the

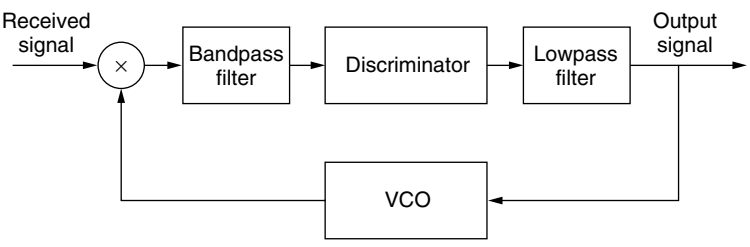


Figure 12. Block diagram of FMFB demodulator.

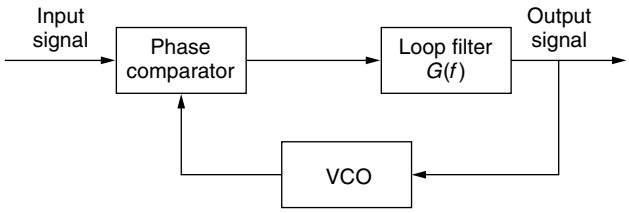


Figure 13. Block diagram of PLL FM demodulator.

instantaneous frequency of the VCO is

$$f_v(t) = f_c + k_v v(t) \tag{55}$$

where k_v is a deviation constant with units of hertz per volt (Hz/V). Consequently, the VCO output may be expressed as

$$y_v(t) = A_v \sin[2\pi f_c t + \phi_v(t)] \tag{56}$$

where

$$\phi_v(t) = 2\pi k_v \int_0^t v(\tau)\, d\tau \tag{57}$$

The phase comparator is basically a multiplier and a filter that rejects the signal component centered at $2f_c$. Hence, its output may be expressed as

$$e(t) = \tfrac{1}{2} A_v A_c \sin[\phi(t) - \phi_v(t)] \tag{58}$$

where the difference $\phi(t) - \phi_v(t) \equiv \phi_e(t)$ constitutes the phase error. The signal $e(t)$ is the input to the loop filter.

Let us assume that the PLL is in lock, so that the phase error is small. Then

$$\sin[\phi(t) - \phi_v(t)] \approx \phi(t) - \phi_v(t) = \phi_e(t) \tag{59}$$

under this condition, we may deal with the linearized model of the PLL, shown in Fig. 14. we may express the phase error as

$$\phi_e(t) = \phi(t) - 2\pi k_v \int_0^t v(\tau)\, d\tau \tag{60}$$

or, equivalently, as either

$$\frac{d}{dt}\phi_e(t) + 2\pi k_v v(t) = \frac{d}{dt}\phi(t) \tag{61}$$

or

$$\frac{d}{dt}\phi_e(t) + 2\pi k_v \int_0^\infty \phi_e(\tau) g(t-\tau)\, d\tau = \frac{d}{dt}\phi(t) \tag{62}$$

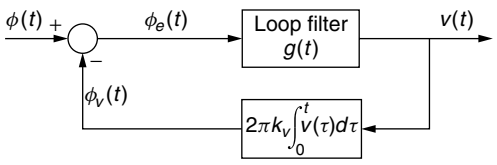


Figure 14. Linearized PLL.

The Fourier transform of the integrodifferential equation in (62) is

$$(j2\pi f)\Phi_e(f) + 2\pi k_v \Phi_e(f) G(f) = (j2\pi f)\Phi(f) \tag{63}$$

and, hence

$$\Phi_e(f) = \frac{1}{1 + \left(\dfrac{kv}{jf}\right) G(f)} \Phi(f) \tag{64}$$

The corresponding equation for the control voltage to the VCO is

$$\begin{aligned} V(f) &= \Phi_e(f) G(f) \\ &= \frac{G(f)}{1 + \left(\dfrac{k_v}{jf}\right) G(f)} \Phi(f) \end{aligned} \tag{65}$$

Now, suppose that we design $G(f)$ such that

$$\left| k_v \frac{G(f)}{jf} \right| \gg 1 \tag{66}$$

in the frequency band $|f| < W$ of the message signal. Then from (65), we have

$$V(f) = \frac{j2\pi f}{2\pi k_v} \Phi(f) \tag{67}$$

or, equivalently

$$\begin{aligned} v(t) &= \frac{1}{2\pi k_v} \frac{d}{dt} \Phi(t) \\ &= \frac{k_f}{k_v} m(t) \end{aligned} \tag{68}$$

Since the control voltage of the VCO is proportional to the message signal, $v(t)$ is the demodulated signal.

We observe that the output of the loop filter with frequency response $G(f)$ is the desired message signal. Hence, the bandwidth of $G(f)$ should be the same as the bandwidth W of the message signal. Consequently, the noise at the output of the loop filter is also limited to the bandwidth W. On the other hand, the output from the VCO is a wideband FM signal with an instantaneous frequency that follows the instantaneous frequency of the received FM signal.

The major benefit of using feedback in FM signal demodulation is to reduce the threshold effect that occurs when the input signal-to-noise-ratio to the FM demodulator drops below a critical value. We will study the threshold effect later in this article.

5. EFFECT OF NOISE ON ANGLE MODULATION

In this section we will study the performance of angle modulated signals when contaminated by additive white Gaussian noise and compare this performance with the performance of amplitude modulated signals. Recall that in amplitude modulation, the message information is contained in the amplitude of the modulated signal and

since noise is additive, the noise is directly added to the signal. However, in a frequency-modulated signal, the noise is added to the amplitude and the message information is contained in the frequency of the modulated signal. Therefore the message is contaminated by the noise to the extent that the added noise changes the frequency of the modulated signal. The frequency of a signal can be described by its zero crossings. Therefore, the effect of additive noise on the demodulated FM signal can be described by the changes that it produces in the zero crossings of the modulated FM signal. Figure 15 shows the effect of additive noise on the zero crossings of two frequency-modulated signals, one with high power and the other with low power. From the above discussion and also from Fig. 15 it should be clear that the effect of noise in an FM system is less than that for an AM system. It is also observed that the effect of noise in a low-power FM system is more than in a high-power FM system. The analysis that we present in this chapter verifies our intuition based on these observations.

A block diagram of the receiver for a general angle-modulated signal is shown in Fig. 16. The angle-modulated signal is represented as[4]

$$\begin{aligned} u(t) &= A_c \cos(2\pi f_c t + \phi(t)) \\ &= \begin{cases} A_c \cos(2\pi f_c t + 2\pi k_f \int_{-\infty}^{t} m(\tau)\, d\tau), & \text{FM} \\ A_c \cos(2\pi f_c t + k_p m(t)), & \text{PM} \end{cases} \end{aligned} \tag{69}$$

The additive white Gaussian noise $n_w(t)$ is added to $u(t)$ and the result is passed through a noise limiting filter whose role is to remove the out-of-band noise. The bandwidth of this filter is equal to the bandwidth of the modulated signal and, therefore, it passes the modulated signal without distortion. However, it eliminates the out-of-band noise and, hence, the noise output of the filter is a bandpass Gaussian noise denoted by $n(t)$. The output of this filter is

$$\begin{aligned} r(t) &= u(t) + n(t) \\ &= u(t) + n_c(t)\cos 2\pi f_c t - n_s(t)\sin 2\pi f_c t \end{aligned} \tag{70}$$

where $n_c(t)$ and $n_s(t)$ denote the in-phase and the quadrature components of the bandpass noise. Because of the nonlinearity of the modulation process, an exact analysis of the performance of the system in the presence of noise is mathematically involved. Let us make the assumption that the signal power is much higher than the noise power. Then, if the bandpass noise is represented as

$$\begin{aligned} n(t) &= \sqrt{n_c^2(t) + n_s^2(t)}\cos\left(2\pi f_c t + \arctan\frac{n_s(t)}{n_c(t)}\right) \\ &= V_n(t)\cos(2\pi f_c t + \Phi_n(t)) \end{aligned} \tag{71}$$

where $V_n(t)$ and $\Phi_n(t)$ represent the envelope and the phase of the bandpass noise process, respectively, the assumption that the signal is much larger than the noise means that

$$p(V_n(t) \ll A_c) \approx 1 \tag{72}$$

Therefore, the phasor diagram of the signal and the noise are as shown in Fig. 17. From this figure it is obvious that we can write

$$\begin{aligned} r(t) &\approx (A_c + V_n(t)\cos(\Phi_n(t) - \phi(t)))\times \\ &\quad \times \cos\left(2\pi f_c t + \phi(t)\right. \\ &\quad \left. + \arctan\frac{V_n(t)\sin(\Phi_n(t) - \phi(t))}{A_c + V_n(t)\cos(\Phi_n(t) - \phi(t))}\right) \\ &\approx (A_c + V_n(t)\cos(\Phi_n(t) - \phi(t))) \\ &\quad \times \cos\left(2\pi f_c t + \phi(t) + \frac{V_n(t)}{A_c}\sin(\Phi_n(t) - \phi(t))\right) \end{aligned}$$

The demodulator processes this signal and, depending whether it is a phase or a frequency demodulator, its output will be the phase or the instantaneous frequency of

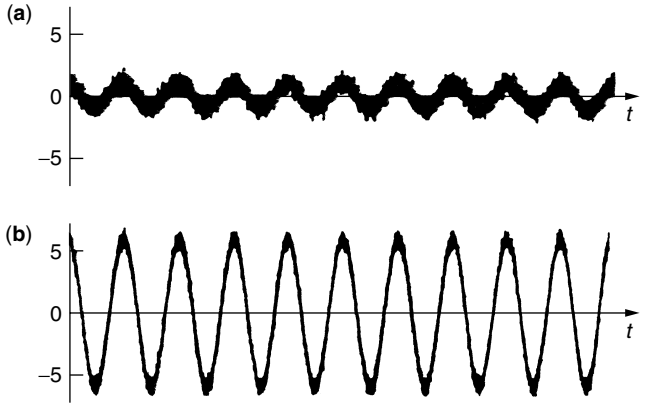


Figure 15. Effect of noise on the zero crossings of (**a**) low-power and (**b**) high-power modulated signals.

[4] When we refer to the modulated signal, we mean the signal as received by the receiver. Therefore, the signal power is the power in the received signal, not the transmitted power.

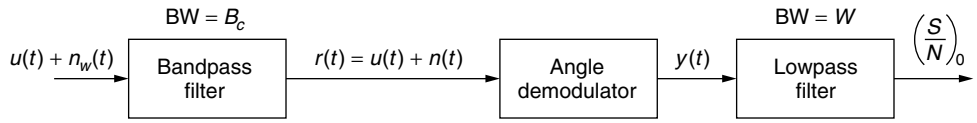


Figure 16. Block diagram of receiver for a general angle-demodulated signal.

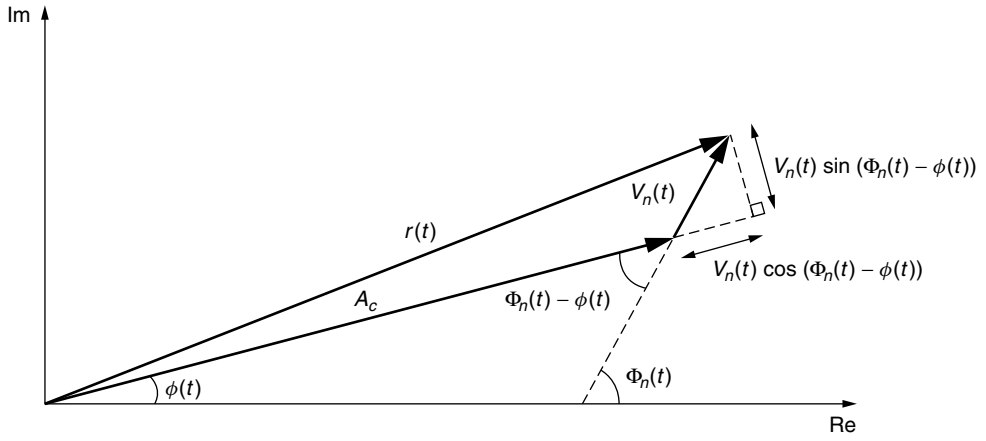


Figure 17. Phasor diagram of signal and noise in an angle-modulated system.

this signal.[5] Therefore, noting that

$$\phi(t) = \begin{cases} k_p m(t), & \text{PM} \\ 2\pi k_f \displaystyle\int_{-\infty}^{t} m(\tau)\,d\tau, & \text{FM} \end{cases} \tag{73}$$

the output of the demodulator is given by

$$\begin{aligned} y(t) &= \begin{cases} k_p m(t) + Y_n(t), & \text{PM} \\ k_f m(t) + \dfrac{1}{2\pi}\dfrac{d}{dt}Y_n(t), & \text{FM} \end{cases} \\ &= \begin{cases} k_p m(t) + \dfrac{V_n(t)}{A_c}\sin(\Phi_n(t)-\phi(t)), & \text{PM} \\ k_f m(t) + \dfrac{1}{2\pi}\dfrac{d}{dt}\dfrac{V_n(t)}{A_c}\sin(\Phi_n(t)-\phi(t)), & \text{FM} \end{cases} \end{aligned} \tag{74}$$

where we have defined

$$Y_n(t) \overset{\text{def}}{=} \frac{V_n(t)}{A_c}\sin(\Phi_n(t)-\phi(t)). \tag{75}$$

The first term in the preceding expressions is the desired signal component, and the second term is the noise component. From this expression we observe that the noise component is inversely proportional to the signal amplitude A_c. Hence, the higher the signal level, the lower will be the noise level. This is in agreement with the intuitive reasoning presented at the beginning of this section based on Fig. 15. Note also that this is not the case with amplitude modulation. In AM systems the noise component is independent of the signal component and a scaling of the signal power does not affect the received noise power.

Let us study the properties of the noise component given by

$$Y_n(t) = \frac{V_n(t)}{A_c}\sin(\Phi_n(t)-\phi(t))$$

[5] Of course, in the FM case the demodulator output is the instantaneous frequency deviation of $v(t)$ from the carrier frequency f_c.

$$\begin{aligned} &= \frac{1}{A_c}[V_n(t)\sin\Phi_n(t)\cos\phi(t) \\ &\quad - V_n(t)\cos\Phi_n(t)\sin\phi(t)] \\ &= \frac{1}{A_c}[n_s(t)\cos\phi(t) - n_c(t)\sin\phi(t)] \end{aligned} \tag{76}$$

The autocorrelation function of this process is given by

$$\begin{aligned} E[Y_n(t+\tau)Y_n(t)] &= \frac{1}{A_c^2}E[R_{n_s}(\tau)\cos\phi(t)\cos\phi(t+\tau) \\ &\quad + R_{n_c}(\tau)\sin\phi(t+\tau)\sin\phi(t)] \\ &= \frac{1}{A_c^2}R_{n_c}(\tau)E[\cos(\phi(t+\tau)-\phi(t))] \end{aligned} \tag{77}$$

where we have used the fact that the noise process is stationary and $R_{n_c}(\tau) = R_{n_s}(\tau)$ and $R_{n_c n_s}(\tau) = 0$. Now we assume that the message $m(t)$ is a sample function of a zero mean, stationary Gaussian process $M(t)$ with the autocorrelation function $R_M(\tau)$. Then, in both PM and FM modulation, $\phi(t)$ will also be a sample function of a zero mean stationary and Gaussian process $\Phi(t)$. For PM this is obvious because

$$\Phi(t) = k_p M(t) \tag{78}$$

and in the FM case we have

$$\Phi(t) = 2\pi k_f \int_{-\infty}^{t} M(\tau)\,d\tau \tag{79}$$

Noting that $\displaystyle\int_{-\infty}^{t}$ represents a linear time-invariant operation, it is seen that, in this case, $\Phi(t)$ is the output of an LTI system whose input is a zero mean, stationary Gaussian process. Consequently $\Phi(t)$ will also be a zero mean, stationary Gaussian process.

At any fixed time t, the random variable $Z(t,\tau) = \Phi(t+\tau) - \Phi(t)$ is the difference between two jointly Gaussian random variables. Therefore, it is itself a Gaussian random variable with mean equal to zero and variance

$$\begin{aligned} \sigma_Z^2 &= E[\Phi^2(t+\tau)] + E[\Phi^2(t)] - 2R_\Phi(\tau) \\ &= 2[R_\Phi(0) - R_\Phi(\tau)] \end{aligned} \tag{80}$$

Now, using this result in (5.9) we obtain

$$\begin{aligned} E[Y_n(t+\tau)Y_n(t)] &= \frac{1}{A_c^2}R_{n_c}(\tau)E\cos(\Phi(t+\tau)-\Phi(t)) \\ &= \frac{1}{A_c^2}R_{n_c}(\tau)\mathrm{Re}[Ee^{j(\Phi(t+\tau)-\Phi(t))}] \\ &= \frac{1}{A_c^2}R_{n_c}(\tau)\mathrm{Re}[Ee^{jZ(t,\tau)}] \\ &= \frac{1}{A_c^2}R_{n_c}(\tau)\mathrm{Re}[e^{-(1/2)\sigma_Z^2}] \\ &= \frac{1}{A_c^2}R_{n_c}(\tau)\mathrm{Re}[e^{-(R_\Phi(0)-R_\Phi(\tau))}] \\ &= \frac{1}{A_c^2}R_{n_c}(\tau)e^{-(R_\Phi(0)-R_\Phi(\tau))} \end{aligned} \tag{81}$$

This result shows that under the assumption of a stationary Gaussian message, the noise process at the output of the demodulator is also a stationary process whose autocorrelation function is given above and whose power spectral density is

$$\begin{aligned} \mathcal{S}_Y(f) &= \mathcal{F}[R_Y(\tau)] \\ &= \mathcal{F}\left[\frac{1}{A_c^2}R_{n_c}(\tau)e^{-(R_\Phi(0)-R_\Phi(\tau))}\right] \\ &= \frac{e^{-R_\Phi(0)}}{A_c^2}\mathcal{F}[R_{n_c}(\tau)e^{R_\Phi(\tau)}] \\ &= \frac{e^{-R_\Phi(0)}}{A_c^2}\mathcal{F}[R_{n_c}(\tau)g(\tau)] \\ &= \frac{e^{-R_\Phi(0)}}{A_c^2}\mathcal{S}_{n_c}(f) * G(f) \end{aligned} \tag{82}$$

where $g(\tau) = e^{R_\Phi(\tau)}$ and $G(f)$ is its Fourier transform.

It can be shown [1,2] that the bandwidth of $g(\tau)$ is $B_c/2$, that is, half of the bandwidth of the angle-modulated signal. For high-modulation indices this bandwidth is much larger than W, the message bandwidth. Since the bandwidth of the angle-modulated signal is defined as the frequencies that contain 98–99% of the signal power, $G(f)$ is very small in the neighborhood of $|f| = \dfrac{B_c}{2}$ and, of course,

$$\mathcal{S}_{n_c}(f) = \begin{cases} N_0, & |f| < \dfrac{B_c}{2} \\ 0, & \text{otherwise} \end{cases} \tag{83}$$

A typical example of $G(f)$, $\mathcal{S}_{n_c}(f)$, and the result of their convolution is shown in Fig. 18. Because $G(f)$ is very small in the neighborhood of $|f| = \frac{B_c}{2}$, the resulting $\mathcal{S}_Y(f)$ has almost a flat spectrum for $|f| < W$, the bandwidth of the message. From Fig. 18 it is obvious that for all $|f| < W$, we have

$$\begin{aligned} S_Y(f) &= \frac{e^{-R_\Phi(0)}}{A_c^2}S_{n_c}(f) * G(f) \\ &= \frac{e^{-R_\Phi(0)}}{A_c^2}N_0\int_{-(B_c/2)}^{B_c/2} G(f)\,df \end{aligned}$$

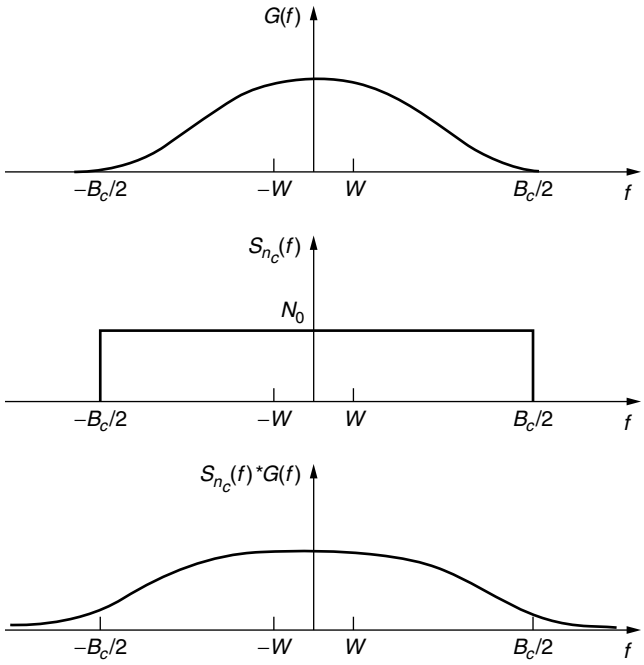


Figure 18. Typical plots of $G(f)$, $S_{n_c}(f)$ and the result of their convolution.

$$\begin{aligned} &\approx \frac{e^{-R_\Phi(0)}}{A_c^2}N_0\int_{-\infty}^{\infty} G(f)\,df \\ &= \frac{e^{-R_\Phi(0)}}{A_c^2}N_0 g(\tau)_{|_{\tau=0}} \\ &= \frac{e^{-R_\Phi(0)}}{A_c^2}N_0 e^{R_\Phi(0)} \\ &= \frac{N_0}{A_c^2} \end{aligned} \tag{84}$$

It should be noted that this relation is a good approximation for $|f| < W$ only. This means that for $|f| < W$, the spectrum of the noise components in the PM and FM case are given by

$$S_{n_o}(f) = \begin{cases} \dfrac{N_0}{A_c^2}, & \text{PM} \\ \dfrac{N_0}{A_c^2}f^2, & \text{FM} \end{cases} \tag{85}$$

where we have used the fact that in FM the noise component is given by $\frac{1}{2\pi}\frac{d}{dt}Y_n(t)$ as indicated in (74). The power spectrum of noise component at the output of the demodulator in the frequency interval $|f| < W$ for PM and FM is shown in Fig. 19. It is interesting to note that PM has a flat noise spectrum and FM has a parabolic noise spectrum. Therefore, the effect of noise in FM for higher-frequency components is much higher than the effect of noise on lower-frequency components. The noise power at the output of the lowpass filter is the noise power in the frequency range $[W, +W]$. Therefore, it is given by

$$P_{n_o} = \int_{-W}^{+W} S_{n_o}(f)\,df$$

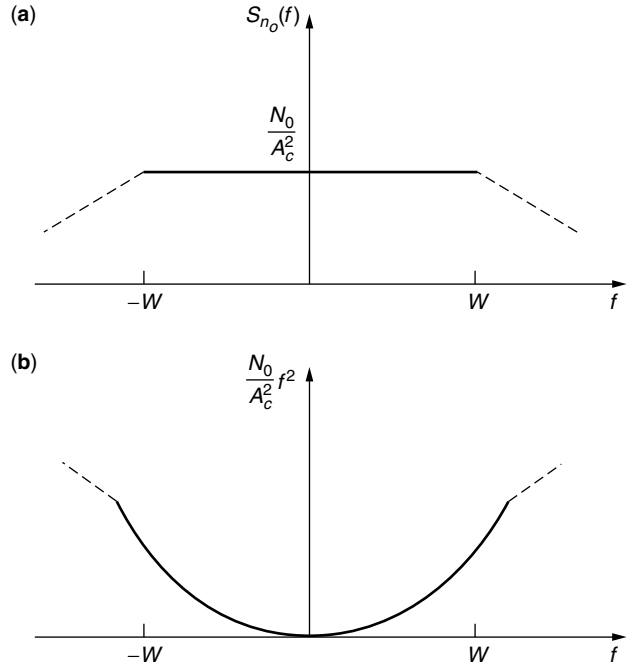


Figure 19. Noise power spectrum at demodulator output in (**a**) PM and (**b**) FM.

$$= \begin{cases} \int_{-W}^{+W} \dfrac{N_0}{A_c^2}\, df, & \text{PM} \\ \int_{-W}^{+W} f^2 \dfrac{N_0}{A_c^2}\, df, & \text{FM} \end{cases}$$

$$= \begin{cases} \dfrac{2WN_0}{A_c^2}, & \text{PM} \\ \dfrac{2N_0W^3}{3A_c^2}, & \text{FM} \end{cases} \tag{86}$$

Now we can use (74) to determine the output signal-to-noise ratio in angle modulation. First we have the output signal power

$$P_{s_o} = \begin{cases} k_p^2 P_M, & \text{PM} \\ k_f^2 P_M, & \text{FM} \end{cases} \tag{87}$$

Then, the signal-to-noise ratio, defined as

$$\left(\frac{S}{N}\right)_o = \frac{P_{s_o}}{P_{n_o}}$$

becomes

$$\left(\frac{S}{N}\right)_o = \begin{cases} \dfrac{k_p^2 A_c^2}{2} \dfrac{P_M}{N_0 W}, & \text{PM} \\ \dfrac{3k_f^2 A_c^2}{2W^2} \dfrac{P_M}{N_0 W}, & \text{FM} \end{cases} \tag{88}$$

Nothing that $\frac{A_c^2}{2}$ is the received signal power, denoted by P_R, and

$$\begin{cases} \beta_p = k_p \max|m(t)|, & \text{PM} \\ \beta_f = \dfrac{k_f \max|m(t)|}{W}, & \text{FM} \end{cases} \tag{89}$$

we may express the output SNR as

$$\left(\frac{S}{N}\right)_o = \begin{cases} P_R \left(\dfrac{\beta_p}{\max|m(t)|}\right)^2 \dfrac{P_M}{N_0 W}, & \text{PM} \\ 3P_R \left(\dfrac{\beta_f}{\max|m(t)|}\right)^2 \dfrac{P_M}{N_0 W}, & \text{FM} \end{cases} \tag{90}$$

If we denote $\frac{P_R}{N_0 W}$ by $\left(\frac{S}{N}\right)_b$, the signal-to-noise ratio of a baseband system with the same received power, we obtain

$$\left(\frac{S}{N}\right)_o = \begin{cases} \dfrac{P_M \beta_p^2}{(\max|m(t)|)^2} \left(\dfrac{S}{N}\right)_b, & \text{PM} \\ 3\dfrac{P_M \beta_f^2}{(\max|m(t)|)^2} \left(\dfrac{S}{N}\right)_b, & \text{FM} \end{cases} \tag{91}$$

Note that in this expression $\frac{P_M}{(\max|m(t)|)^2}$ is the average-to-peak-power-ratio of the message signal (or equivalently, the power content of the normalized message, P_{M_n}). Therefore

$$\left(\frac{S}{N}\right) = \begin{cases} \beta_p^2 P_{M_n} \left(\dfrac{S}{N}\right)_b, & \text{PM} \\ 3\beta_f^2 P_{M_n} \left(\dfrac{S}{N}\right)_b, & \text{FM} \end{cases} \tag{92}$$

Now using Carson's rule $B_c = 2(\beta + 1)W$, we can express the output SNR in terms of the bandwidth expansion factor, which is defined to be the ratio of the channel bandwidth to the message bandwidth and denoted by Ω:

$$\Omega = \frac{B_c}{W} = 2(\beta + 1) \tag{93}$$

From this relationship we have $\beta = \frac{\Omega}{2} - 1$. Therefore

$$\left(\frac{S}{N}\right)_o = \begin{cases} P_M \left(\dfrac{\frac{\Omega}{2} - 1}{\max|m(t)|}\right)^2 \left(\dfrac{S}{N}\right)_b, & \text{PM} \\ 3P_M \left(\dfrac{\frac{\Omega}{2} - 1}{\max|m(t)|}\right)^2 \left(\dfrac{S}{N}\right)_b, & \text{FM} \end{cases} \tag{94}$$

From (90) and (94), we observe that

1. In both PM and FM the output SNR is proportional to the square of the modulation index β. Therefore, increasing β increases the output SNR even with low received power. This is in contrast to amplitude modulation, where such an increase in the received signal-to-noise ratio is not possible.
2. The increase in the received signal-to-noise ratio is obtained by increasing the bandwidth. Therefore angle modulation provides a way to trade off bandwidth for transmitted power.
3. The relation between the output SNR and the bandwidth expansion factor, Ω, is a quadratic

relation. This is far from optimal.[6] An information-theoretic analysis shows that the optimal relation between the output SNR and the bandwidth expansion factor is an exponential relation.

4. Although we can increase the output signal-to-noise ratio by increasing β, having a large β means having a large B_c (by Carson's rule). Having a large B_c means having a large noise power at the input of the demodulator. This means that the approximation $p(V_n(t) \ll A_c) \approx 1$ will no longer apply and that the preceding analysis will not hold. In fact if we increase β such that the preceding approximation does not hold, a phenomenon known as the *threshold effect* will occur and the signal will be lost in the noise.
5. A comparison of the preceding result with the signal-to-noise ratio in amplitude modulation shows that in both cases increasing the transmitter power (or the received power), will increase the output signal-to-noise ratio, but the mechanisms are totally different. In AM, any increase in the received power directly increases the signal power at the output of the receiver. This is basically due to the fact the message is in the amplitude of the transmitted signal and an increase in the transmitted power directly affects the demodulated signal power. However, in angle modulation, the message is in the phase of the modulated signal and, consequently, increasing the transmitter power does not increase the demodulated message power. In angle modulation what increases the output signal-to-noise ratio is a *decrease in the received noise power* as seen from (86) and Fig. 15.
6. In FM the effect of noise is higher at higher frequencies. This means that signal components at higher frequencies will suffer more from noise than will the lower-frequency components. In some applications where FM is used to transmit SSB FDM signals, those channels that are modulated on higher-frequency carriers suffer from more noise. To compensate for this effect, such channels must have a higher signal level. The quadratic characteristics of the demodulated noise spectrum in FM is the basis of preemphasis and deemphasis filtering, which are discussed later in this article.

5.1. Threshold Effect in Angle Modulation

The noise analysis of angle demodulation schemes is based on the assumption that the signal-to-noise ratio at the demodulator input is high. With this crucial assumption we observed that the signal and noise components at the demodulator output are additive and we were able to carry out the analysis. This assumption of high signal-to-noise ratio is a simplifying assumption that is usually made in analysis of nonlinear modulation systems. Because of the nonlinear nature of the demodulation process, there is no reason that the additive signal and noise components at the input of the modulator result in additive signal and noise components at the output of the demodulator. In fact, this assumption is not at all correct in general, and the signal and noise processes at the output of the demodulator are completely mixed in a single process by a complicated nonlinear relation. Only under the high signal-to-noise ratio assumption is this highly nonlinear relation approximated as an additive form. Particularly at low signal-to-noise ratios, signal and noise components are so intermingled that one cannot recognize the signal from the noise and, therefore, no meaningful signal-to-noise ratio as a measure of performance can be defined. In such cases the signal is not distinguishable from the noise and a *mutilation* or *threshold effect* is present. There exists a specific signal to noise ratio at the input of the demodulator known as the *threshold SNR* beyond which signal mutilation occurs. The existence of the threshold effect places an upper limit on the tradeoff between bandwidth and power in an FM system. This limit is a practical limit in the value of the modulation index β_f.

It can be shown that at threshold the following approximate relation between $\frac{P_R}{N_0 W} = (\frac{S}{N})_b$ and β_f holds in an FM system:

$$\left(\frac{S}{N}\right)_{b,\text{th}} = 20(\beta + 1) \tag{95}$$

From this relation, given a received power P_R, we can calculate the maximum allowed β to make sure that the system works above threshold. Also, given a bandwidth allocation of B_c, we can find an appropriate β using Carson's rule $B_c = 2(\beta + 1)W$. Then, using the threshold relation given above we determine the required minimum received power to make the whole allocated bandwidth usable.

In general there are two factors that limit the value of the modulation index β. The first is the limitation on channel bandwidth that affects β through Carson's rule. The second is the limitation on the received power that limits the value of β to less than what is derived from (95). Figure 20 shows plots of the SNR in an FM system as a function of the baseband SNR. The SNR values in these curves are in decibels, and different curves correspond to different values of β as marked. The effect of threshold is apparent from the sudden drops in the output SNR. These plots are drawn for a sinusoidal message for which

$$\frac{P_M}{(\max |m(t)|)^2} = \frac{1}{2} \tag{96}$$

In such a case

$$\left(\frac{S}{N}\right)_o = \frac{3}{2}\beta^2 \left(\frac{S}{N}\right)_b \tag{97}$$

As an example, for $\beta = 5$, the preceding relation yields

$$\left(\frac{S}{N}\right)_{o|\text{dB}} = 15.7 + \left(\frac{S}{N}\right)_{b|\text{dB}} \tag{98}$$

$$\left(\frac{S}{N}\right)_{b,\text{th}} = 120 \sim 20.8 \text{ dB} \tag{99}$$

[6] By *optimal relation* we mean the maximum saving in transmitter power for a given expansion in bandwidth. An optimal system achieves the fundamental limits on communication predicted by information theory.

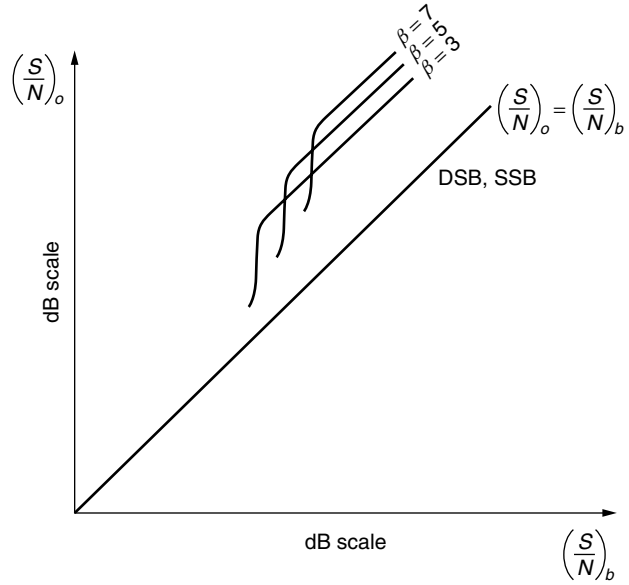


Figure 20. Output SNR versus baseband SNR in an FM system for various values of β.

On the other hand, if $\beta = 2$, we have

$$\left(\frac{S}{N}\right)_{o|\text{dB}} = 7.8 + \left(\frac{S}{N}\right)_{b|\text{dB}} \tag{100}$$

$$\left(\frac{S}{N}\right)_{b,\text{th}} = 60 \sim 17.8 \text{ dB} \tag{101}$$

From this discussion it is apparent that if, for example $\left(\frac{S}{N}\right)_b = 20$ dB, then, regardless of the available bandwidth, we cannot use $\beta = 5$ for such a system because the demodulator will not demodulate below the threshold of 20 dB. However, $\beta = 2$ can be used, which yields an SNR equal to 27.8 dB at the output of the receiver. This is an improvement of 7.8 dB compared to a baseband system.

In general, if we want to employ the maximum available bandwidth, we must choose the largest possible β that guarantees that the system operates above threshold. This is the value of β that satisfies

$$\left(\frac{S}{N}\right)_{b,\text{th}} = 20(\beta + 1) \tag{102}$$

By substituting this value in (92), we obtain

$$\left(\frac{S}{N}\right)_o = 60\beta^2(\beta + 1)P_{M_n} \tag{103}$$

which relates a desired output SNR to the highest possible β that achieves that SNR.

5.1.1. Threshold Extension in Frequency Modulation. We have already seen that the nonlinear demodulation effect in angle modulation in general results in nonadditive signal and noise at the output of the demodulator. In high received signal-to-noise ratios, the nonlinear demodulation process can be well approximated by a linear equivalent and therefore signal and noise at the demodulator output will be additive. At high noise levels, however, this approximation is not valid anymore and the threshold effect results in signal mutilation. We have also seen that in general the modulated signal bandwidth increases with the modulation index and since the power of the noise entering the receiver is proportional to the system bandwidth, higher modulation indices cause the threshold effect to appear at higher received powers.

In order to reduce the threshold — in other words, in order to delay the threshold effect to appear at lower received signal power — it is sufficient to decrease the input noise power at the receiver. This can be done by decreasing the effective system bandwidth at the receiver.

Two approaches to FM threshold extension are to employ FMFB or PLL FM (see Figs. 12 and 13) at the receiver. We have already seen in Section 4 in the discussion following FMFB and PLL FM systems that these systems are capable of reducing the effective bandwidth of the receiver. This is exactly what is needed for extending the threshold in FM demodulation. Therefore, in applications where power is very limited and bandwidth is abundant, these systems can be employed to make it possible to use the available bandwidth more efficiently. Using FMFB the threshold can be extended approximately by 5–7 dB.

5.2. Preemphasis and Deemphasis Filtering

As observed in Fig. 19, the noise power spectral density at the output of the demodulator in PM is flat within the message bandwidth whereas for FM the noise power spectrum has a parabolic shape. This means that for low-frequency components of the message signal FM performs better and for high-frequency components, PM is a better choice. Therefore, if we can design a system that for low-frequency components of the message signal performs frequency modulation and for high-frequency components works as a phase modulator, we have a better overall performance compared to each system alone. This is the idea behind preemphasis and deemphasis filtering techniques.

The objective in preemphasis and deemphasis filtering is to design a system that behaves like an ordinary frequency modulator–demodulator pair in the low frequency band of the message signal and like a phase modulator–demodulator pair in the high-frequency band of the message signal. Since a phase modulator is nothing but the cascade connection of a differentiator and a frequency modulator, we need a filter in cascade with the modulator that at low frequencies does not affect the signal and at high frequencies acts as a differentiator. A simple highpass filter is a very good approximation to such a system. Such a filter has a constant gain for low frequencies and at higher frequencies it has a frequency characteristic approximated by $K|f|$, which is the frequency characteristic of a differentiator. At the demodulator side, for low frequencies we have a simple FM demodulator and for high-frequency components we have a phase demodulator, which is the cascade of a simple FM demodulator and an integrator. Therefore, at the demodulator we need a filter that at low frequencies has a constant gain and at high frequencies

behaves as an integrator. A good approximation to such a filter is a simple lowpass filter. The modulator filter which emphasizes high frequencies is called the pre-emphasis filter and the demodulator filter which is the inverse of the modulator filter is called the *deemphasis filter*. Frequency responses of a sample preemphasis and deemphasis filter are given in Fig. 21.

Another way to look at preemphasis and deemphasis filtering is to note that due to the high level of noise in the high-frequency components of the message in FM, it is desirable to attenuate the high frequency components of the demodulated signal. This results in a reduction in the noise level but it causes the higher-frequency components of the message signal to be also attenuated. To compensate for the attenuation of the higher components of the message signal we can amplify these components at the transmitter before modulation. Therefore at the transmitter we need a highpass filter, and at the receiver we must use a lowpass filter. The net effect of these filters should be a flat frequency response. Therefore, the receiver filter should be the inverse of the transmitter filter.

The characteristics of the preemphasis and deemphasis filters depend largely on the power spectral density of the message process. In commercial FM broadcasting of music and voice, first order lowpass and highpass *RC* (resistance–capacitance) filters with a time constant of 75 microseconds (μs) are employed. In this case the frequency response of the receiver (deemphasis) filter is given by

$$H_d(f) = \frac{1}{1 + j\dfrac{f}{f_0}} \tag{104}$$

where $f_0 = \dfrac{1}{2\pi \times 75 \times 10^{-6}} \approx 2100$ Hz is the 3-dB frequency of the filter.

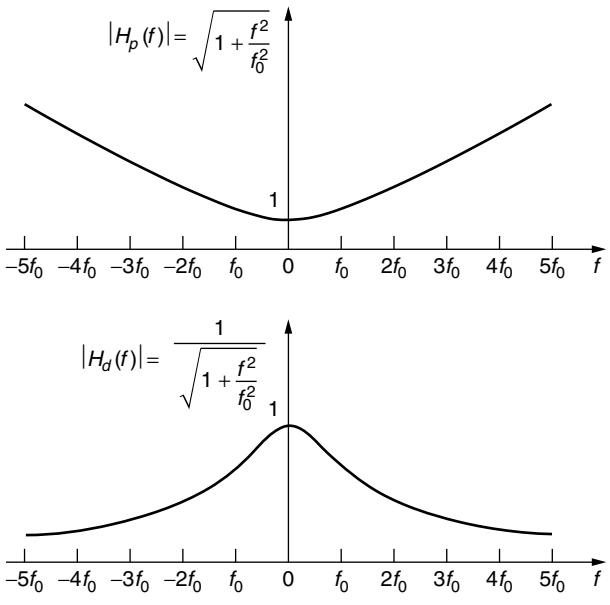


Figure 21. Preemphasis (**a**) and deemphasis (**b**) filter characteristics.

To analyze the effect of preemphasis and deemphasis filtering on the overall signal-to-noise ratio in FM broadcasting, we note that since the transmitter and the receiver filters cancel the effect of each other, the received power in the message signal remains unchanged and we only have to consider the effect of filtering on the received noise. Of course, the only filter that has an effect on the received noise is the receiver filter that shapes the power spectral density of the noise within the message bandwidth. The noise component before filtering has a parabolic power spectrum. Therefore, the noise component after the deemphasis filter has a power spectral density given by

$$\begin{aligned} S_{n_{\mathrm{PD}}}(f) &= S_{n_o}(f)|H_d(f)|^2 \\ &= \frac{N_0}{A_c^2} f^2 \frac{1}{1 + \dfrac{f^2}{f_0^2}} \end{aligned} \tag{105}$$

where we have used (85). The noise power at the output of the demodulator now can be obtained as

$$\begin{aligned} P_{n_{\mathrm{PD}}} &= \int_{-W}^{+W} S_{n_{\mathrm{PD}}}(f)\, df \\ &= \frac{N_0}{A_c^2} \int_{-W}^{+W} \frac{f^2}{1 + \dfrac{f^2}{f_0^2}}\, df \\ &= \frac{2N_0 f_0^3}{A_c^2} \left[\frac{W}{f_0} - \arctan \frac{W}{f_0} \right] \end{aligned} \tag{106}$$

Because the demodulated message signal power in this case is equal to that of a simple FM system with no preemphasis and deemphasis filtering, the ratio of the output SNRs in these two cases is inversely proportional to the noise power ratios:

$$\begin{aligned} \frac{\left(\dfrac{S}{N}\right)_{o_{\mathrm{PD}}}}{\left(\dfrac{S}{N}\right)_o} &= \frac{P_{n_o}}{P_{n_{\mathrm{PD}}}} \\ &= \frac{\dfrac{2N_0 W^3}{3A_c^2}}{\dfrac{2N_0 f_0^3}{A_c^2} \left[\dfrac{W}{f_0} - \arctan \dfrac{W}{f_0} \right]} \\ &= \frac{1}{3} \frac{\left(\dfrac{W}{f_0}\right)^3}{\dfrac{W}{f_0} - \arctan \dfrac{W}{f_0}} \end{aligned} \tag{107}$$

where we have used (86). Equation (107) gives the improvement obtained by employing preemphasis and deemphasis filtering.

In a broadcasting FM system with signal bandwidth $W = 15$ kHz, $f_0 = 2100$ Hz, and $\beta = 5$, using preemphasis and deemphasis filtering improves the performance of an FM system by 13.3 dB. The performance improvement of an FM system with no preemphasis and deemphasis filtering compared to a baseband system is 15–16 dB.

Thus an FM system with preemphasis and deemphasis filtering improves the performance of a baseband system by roughly 29–30 dB.

6. FM RADIO BROADCASTING

Commercial FM radio broadcasting utilizes the frequency band 88–108 MHz for transmission of voice and music signals. The carrier frequencies are separated by 200 kHz and the peak frequency deviation is fixed at 75 kHz. With a signal bandwidth of 15 kHz, this results in a modulation index of $\beta = 5$. Preemphasis filtering with $f_0 = 2100$ Hz is generally used, as described in the previous section, to improve the demodulator performance in the presence of noise in the received signal. The lower 4 MHz of the allocated bandwidth is reserved for noncommercial stations; this accounts for a total of 20 stations, and the remaining 80 stations in the 92–108-MHz bandwidth are allocated to commercial FM broadcasting.

The receiver most commonly used in FM radio broadcast is a superheterodyne type. The block diagram of such a receiver is shown in Fig. 22. As in AM radio reception, common tuning between the RF amplifier and the local oscillator allows the mixer to bring all FM radio signals to a common IF bandwidth of 200 kHz, centered at $f_{\text{IF}} = 10.7$ MHz. Since the message signal $m(t)$ is embedded in the frequency of the carrier, any amplitude variations in the received signal are a result of additive noise and interference. The amplitude limiter removes any amplitude variations in the received signal at the output of the IF amplifier by band-limiting the signal. A bandpass filter centered at $f_{\text{IF}} = 10.7$ MHz with a bandwidth of 200 kHz is included in the limiter to remove higher-order frequency components introduced by the nonlinearity inherent in the hard limiter.

A balanced frequency discriminator is used for frequency demodulation. The resulting message signal is then passed to the *audiofrequency amplifier*, which performs the functions of deemphasis and amplification. The output of the audio amplifier is further filtered by a lowpass filter to remove out-of-band noise and its output is used to drive a loudspeaker.

6.1. FM Stereo Broadcasting

Many FM radio stations transmit music programs in stereo by using the outputs of two microphones placed in two different parts of the stage. Figure 23 is a block diagram of an FM stereo transmitter. The signals from the left and right microphones, $m_\ell(t)$ and $m_r(t)$, are added and subtracted as shown. The sum signal $m_\ell(t) + m_r(t)$ is left as is and occupies the frequency band 0–15 kHz. The difference signal $m_\ell(t) - m_r(t)$ is used to AM modulate (DSB-SC) a 38-kHz carrier that is generated from a 19-kHz oscillator. A pilot tone at the frequency of 19 kHz is added to the signal for the purpose of demodulating the DSB SC AM signal. The reason for placing the pilot tone at 19 kHz instead of 38 kHz is that the pilot is more easily separated from the composite signal at the receiver. The combined signal is used to frequency modulate a carrier.

By configuring the baseband signal as an FDM signal, a monophonic FM receiver can recover the sum signal $m_\ell(t) + m_r(t)$ by use of a conventional FM demodulator. Hence, FM stereo broadcasting is compatible with conventional FM. The second requirement is that the resulting FM signal does not exceed the allocated 200-kHz bandwidth.

The FM demodulator for FM stereo is basically the same as a conventional FM demodulator down to the limiter/discriminator. Thus, the received signal is converted to baseband. Following the discriminator, the baseband message signal is separated into the two signals $m_\ell(t) + m_r(t)$ and $m_\ell(t) - m_r(t)$ and passed through deemphasis filters, as shown in Fig. 24. The difference signal is obtained from the DSB SC signal by means of a synchronous demodulator using the pilot tone. By taking the sum and difference of the two composite signals, we recover the two signals $m_\ell(t)$ and $m_r(t)$. These audio signals are amplified by audio band amplifiers and the two outputs drive dual loudspeakers. As indicated above, an FM receiver that is not configured to receive the

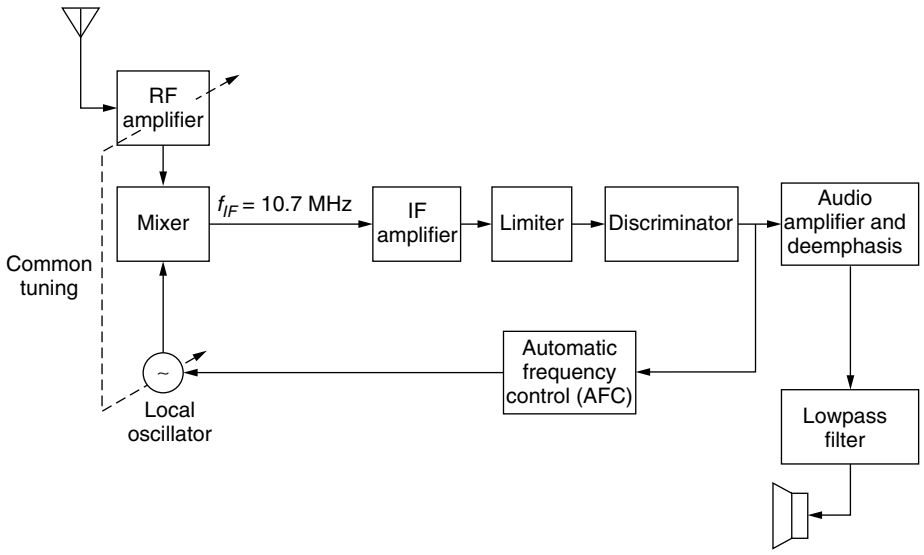


Figure 22. Block diagram of a superheterodyne FM radio receiver.

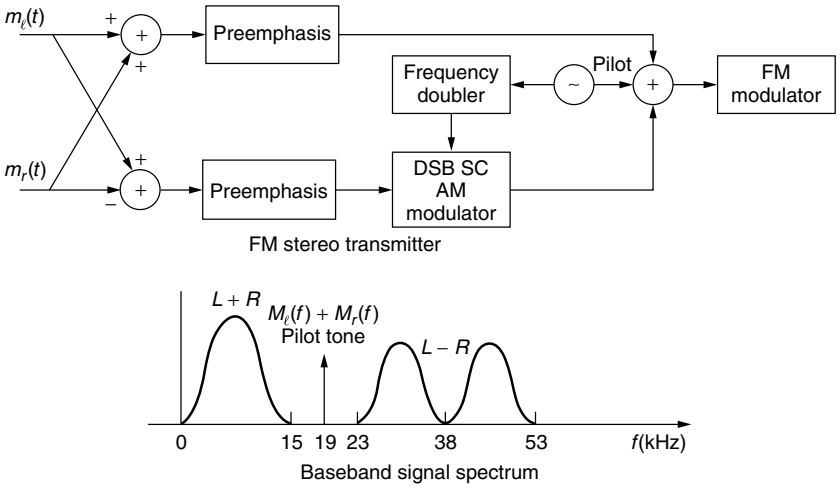


Figure 23. FM stereo transmitter and signal spacing.

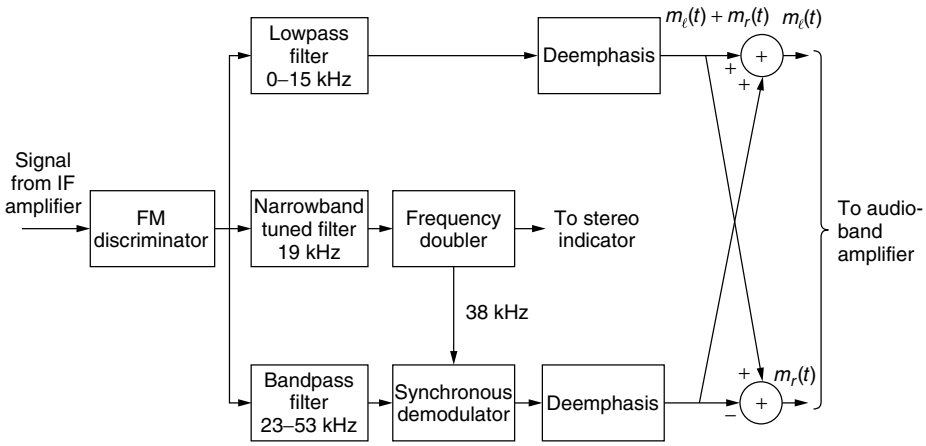


Figure 24. FM stereo receiver.

FM stereo sees only the baseband signal $m_\ell(t) + m_r(t)$ in the frequency range 0–15 kHz. Thus, it produces a monophonic output signal which consists of the sum of the signals at the two microphones.

BIOGRAPHY

Masoud Salehi received a B.S. degree from Tehran University and M.S. and Ph.D. degrees from Stanford University, all in electrical engineering. Before joining Northeastern University, he was with the Electrical and Computer Engineering Departments, Isfahan University of Technology and Tehran University.

From February 1988 until May 1989 Dr. Salehi was a visiting professor at the Information Theory Research Group, Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands, where he did research in network information theory and coding for storage media.

Professor Salehi is currently with the Department of Electrical and Computer Engineering and a member of the CDSP (Communication and Digital Signal Processing) Center, Northeastern University, where he is involved in teaching and supervising graduate students in information and communication theory. His main areas of research interest are network information theory, source-channel matching problems in single and multiple user environments, data compression, channel coding, and particularly turbo codes. Professor Salehi's research has been supported by research grants from the National Science Foundation, DARPA, GTE, and Analog Devices. Professor Salehi is the coauthor of the textbooks *Communication Systems Engineering*, published by Prentice-Hall in 1994 and 2002, and *Contemporary Communication Systems Using MATLAB*, published by Brooks/Cole in 1998 and 2000.

BIBLIOGRAPHY

1. J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, in press.
2. D. J. Sakrison, *Communication Theory: Transmission of Waveforms and Digital Information*, Wiley, New York, 1968.
3. A. B. Carlson, *Communication Systems*, 3rd ed., McGraw-Hill, New York, 1986.
4. L. W. Couch, II, *Digital and Analog Communication Systems*, 4th ed., Macmillan, New York, 1993.
5. J. D. Gibson, *Principles of Digital and Analog Communications*, 2nd ed., Macmillan, New York, 1993.

6. M. A. McMahon, *The Making of A Profession — Century of Electrical Engineering in America*, IEEE Press, 1984.
7. K. S. Shanmugam, *Digital and Analog Communication Systems*, Wiley, New York, 1979.
8. H. Taub and D. A. Schilling, *Principles of Communication Systems*, 2nd ed., McGraw-Hill, New York, 1986.
9. R. E. Ziemer and W. H. Tranter, *Principles of Communications*, 4th ed., Wiley, New York, 1995.
10. M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed., McGraw-Hill, New York, 1990.
11. F. G. Stremler, *Introduction to Communication Systems*, 3rd ed., Addison-Wesley, Reading, MA, 1990.
12. S. Haykin, *Communication Systems*, 4th ed., Wiley, New York, 2001.
13. B. P. Lathi, *Modern Digital and Analog Communication Systems*, 3rd ed., Oxford Univ. Press, New York, 1998.

FREQUENCY-DIVISION MULTIPLE ACCESS (FDMA): OVERVIEW AND PERFORMANCE EVALUATION

FOTINI-NIOVI PAVLIDOU
Aristotle University of Thessaloniki
Thessaloniki, Greece

1. BASIC SYSTEM DESCRIPTION

FDMA (frequency-division multiple access) is a very basic multiple-access technique for terrestrial and satellite systems. We recall that *multiple access* is defined as the ability of a number of users to share a common transmission channel (coaxial cable, fiber, wireless transmission, etc.). Referring to the seven-layer OSI (Open System Interconnection) model, the access methods and the radio channel (frequency, time, and space) assignment are determined by the media access control (MAC) unit of the second layer.

Historically, FDMA has the highest usage and application of the various access techniques. It is one of the three major categories of fixed-assignment access methods (TDMA, FDMA, CDMA), and since it is definitely the simplest one, it has been extensively used in telephony, in commercial radio, in television broadcasting industries, in the existing cellular mobile systems, in cordless systems (CT2), and generally in many terrestrial and satellite wireless applications [1–3].

This access method is efficient if the user has a steady flow of information to send (digitized voice, video, transfer of long files) and uses the system for a long period of time, but it can be very inefficient if user data are sporadic in nature, as is the case with bursty computer data or short-message traffic. In this case it can be effectively applied only in hybrid implementations, as, for example, in FDMA/Aloha systems.

The principle of operation of FDMA is shown in Figs. 1a–d. The total common channel bandwidth is B Hz, and K users are trying to share it. In the FDMA technique each of the K users (transmitting stations) can transmit all of the time, or at least for extended periods of time but using only a portion B_i (subchannel) of the total channel bandwidth B such that $B_i = B/K$ Hz. If the users generate constantly unequal amounts of traffic, one can modify this scheme to assign bandwidth in proportion to the traffic generated by each one.

Adjacent users occupy different carriers of B_i bandwidth, with guard channels D Hz between them to avoid interference. Then the actual bandwidth available to each station for information is $B_i = \{B - (K + 1)D\}/K$. The input of each source is modulated over a carrier and transmitted to the channel. So the channel transmits several carriers simultaneously at different frequencies. User separability is therefore achieved by separation in frequency. At the receiving end, the user bandpass filters select the designated channel out of the composite signal. Then a demodulator obtains the transmitted baseband signal.

Instead of transmitting one source signal on the carrier, we can feed a multiplexed signal on it [e.g., a pulse code modulation (PCM) telephone line]. Depending on the multiplexing and modulation techniques used, several transmission schemes can be considered. Although FDMA is usually considered to be built on the well-known FDM scheme, any multiplexing and modulation technique can be used for the processing of the baseband data, so several forms of FDMA are possible [4–6].

In Figs. 1b,c we give a very general implementation of the system. The first "user/station" transmits analog baseband signals, which are combined in a frequency-division multiplex (FDM) scheme. This multiplexed signal can modulate a carrier in frequency (FM) and then it is transmitted on the common channel, together with other carriers from other stations. If the technique used for all the stations is FDM/FM, the access method is called FDM/FM/FDMA (Fig. 1b).

If the stations transmit digital data, other modulation techniques can be used like PSK (phase shift keying) and TDM (time-division multiplex) can be applied. Then again this multiplexed signal can be transmitted on a frequency band on the common channel together with the carriers of other similar stations. This application is called TDM/PSK/FDMA.

The possible combinations of the form mux/mod/FDMA are numerous even if in practice the schemes shown in Fig. 1 are the most commonly used ones [3,6]. They are generally known as *multichannel-per-carrier* (MCPC) techniques.

Of course, for lower-traffic requirements the baseband signals can modulate directly a carrier in either analog or digital form (Fig. 1c). This concept, *single channel per carrier* (SCPC), has been extensively used in mobile satellites since it allows frequency reallocations according to increases of traffic and future developments in modulation schemes. The best-known application of SCPC/FDMA is the *single-channel-per-carrier pulse-code-modulation multiple-access demand assigned equipment* (SPADE) system applied in Intelsat systems.

FDMA carriers are normally assigned according to a fixed-frequency plan. However, in applications where

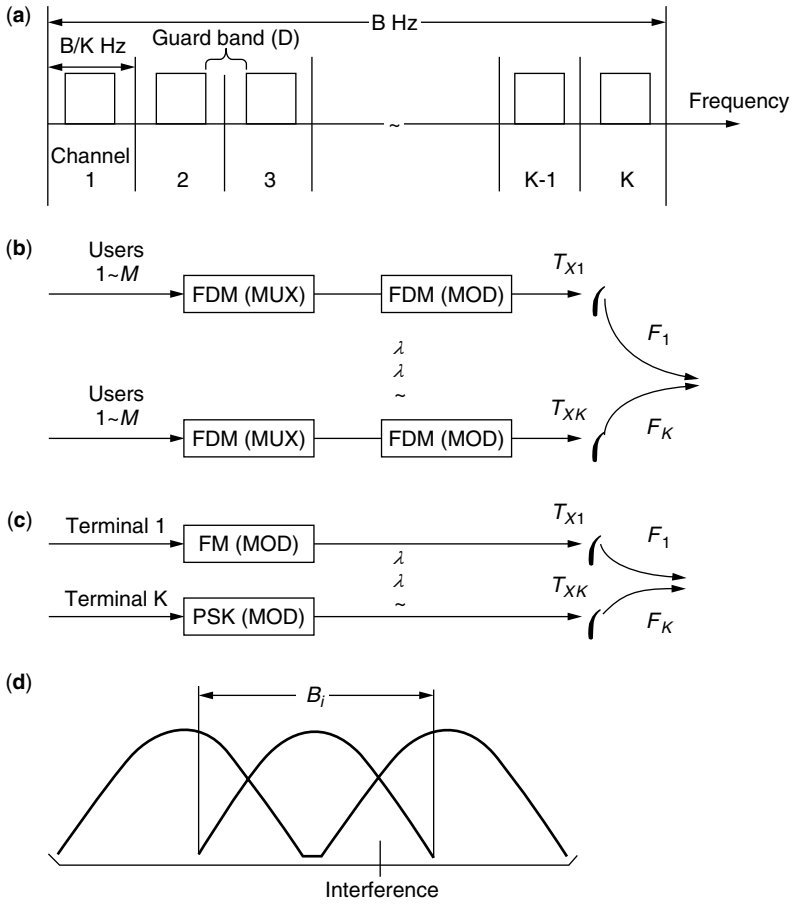


Figure 1. (**a**) FDMA frequency plan; (**b**) FDM/FM/FDMA implementation; (**c**) single-channel-per-carrier (SCPC)/FDMA implementation; (**d**) adjacent-channel interference.

limited space segment resources are to be shared by a relatively large number of users, the channel can be accessed on demand [demand assignment multiple access (DAMA)].

In wireless applications, the FDMA architecture is also known as "narrowband radio," as the bandwidth of the individual data or digitized analog signal (voice, facsimile, etc.) is relatively narrow compared with TDMA and CDMA applications.

2. SYSTEM PERFORMANCE

The major factor that determines the performance of the FDMA scheme is the received carrier-to-noise (plus interference) ratio, and for a FDM/FM/FDMA satellite system this is given by [6]

$$\frac{1}{C_T/N_T} = \frac{1}{C_u/N_u} + 1\left(\frac{C_u}{I_u}\right) + 1\left(\frac{C_d}{\mathrm{IM}}\right) + 1\left(\frac{C_d}{N_d}\right) + \frac{1}{C_d/I_d}$$

where subscripts T, u, and d denote total, uplink, and downlink factors: N gives the white noise; I denotes the interference from other systems using the same frequency; and IM is the intermodulation distortion, explained in Sections 2.1 and 2.2.

2.1. Adjacent-Channel Interference and Intermodulation Noise

As we have noted, in FDMA applications frequency spacing is required between adjacent channels. This issue has been put forward several times by opponents of FDMA to claim that the technique is not as efficient in spectrum use as TDMA or CDMA. Indeed, excessive separation causes needless waste of the available bandwidth. Whatever filters are used to obtain a sharp frequency band for each carrier, part of the power of a carrier adjacent to the one considered will be captured by the receiver of the last one. In Fig. 1d we can see three adjacent bands of the FDMA spectrum (received at a power amplifier) composed of K carriers of equal power and identically modulated. To determine the proper spacing between FDMA carrier spectra, this adjacent channel interference (crosstalk power) must be carefully calculated. Spacings can then be selected for any acceptable crosstalk level desired. Common practice is to define the guard bands equal to around 10% of the carrier bandwidth (for carriers equal in amplitude and bandwidth). This will keep the noise level below the ITU-T (CCITT) requirements [6]. Simplified equations as well as detailed analysis for crosstalk are given in Refs. 3 and 6.

In addition, when the multiple FDMA carriers pass through nonlinear systems, like power amplifiers in satellites, two basic effects occur: (1) the nonlinear

device output contains not only the original frequencies but also undesirable frequencies, that is, unwanted intermodulation (IM) products that fall into the FDMA bands as interference; and (2) the available output power decreases as a result of conversion of useful satellite power to intermodulation noise. Both of these effects depend on the type of nonlinearity and the number of simultaneous FDMA carriers present, as well as their power levels and spectral distributions. This makes it necessary to reduce the input to the amplifier from its maximum drive level in order to control intermodulation distortion. In satellite repeaters, this procedure is referred to as *input backoff* and is an important factor in maximizing the power efficiency of the repeater. So the traveling-wave tube (TWT) has to be *backed off* substantially in order to operate it as a linear amplifier. For example, Intelsat has established a series of monitoring stations to ensure that its uplink power levels are maintained. This in turn leads to inefficient usage of the available satellite power. Several nonlinear models for nonlinear power amplifiers, which have both amplitude and phase nonlinearities, have been proposed to calculate the carrier-to-IM versus input backoff.

An analysis of IM products for a satellite environment can be found elsewhere in the literature, where a detailed description and an exact calculation of intermodulation distortion are given [3,4,6–8]. Generally we have to determine the spectral distribution obtained from the mixing of the FDMA carriers. In Fig. 2a the intermodulation spectrum received by the mixing of the spectra of Fig. 1d is shown. The intermodulation power tends to be concentrated in the center of the total bandwidth B, so that the center carriers receive the most interference. Figure 2b shows the ratio of carrier power C_T to the intermodulation noise power IM, as a function of the number of carriers K and the degree of backoff.

In an ideal FDMA system the carriers have equal powers, but this is not the case in practice. When a mixture of both strong and weak carriers are present on the channel, we must ensure that the weaker carriers can maintain a communication link of acceptable quality. Some techniques have been implemented to accommodate this situation; one is known as *channelization*. In channelization carriers are grouped in *channels*, each with its own bandpass filter and power amplifier; in each channel (group) only carriers of the same power are transmitted.

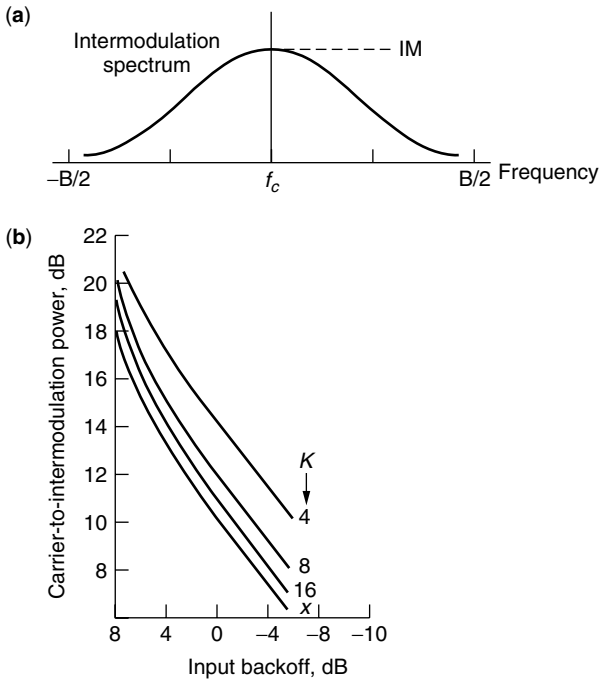


Figure 2. Intermodulation distortion.

2.2. FDMA Throughput

The throughput capability of an FDM/FM/FDMA scheme has been studied [3,6,7] as a function of the number of carriers taking into account the carrier-to-total noise (C_T/N_T) factor. The carriers are modulated by multiplexed signals of equal capacity. As the number of carriers increases, the bandwidth allocated to each carrier must decrease, and this leads to a reduction of the capacity of the modulating multiplexed signal. As the total capacity is the product of the capacity of each carrier and the total number of carriers, it could be imagined that the total capacity would remain sensibly constant. But it is not; the total capacity decreases as the number of carriers increases. This results from the fact that each carrier is subjected to a reduction in the value of C/N since the backoff is large when the number of carriers is high (extra carriers bring more IM products). Another reason is the increased need for guard bands. Figure 3 depicts the throughput of a FDMA system as a function of the number of carriers for an Intelsat transponder of 3 MHz bandwidth; it effectively shows the ratio of the total real channel capacity and the potential capacity of the channel.

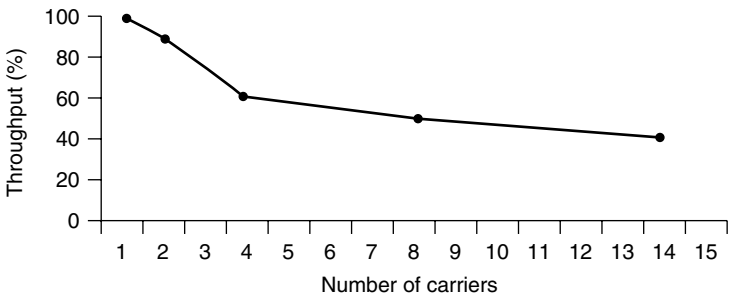


Figure 3. Throughput of FDM/FM/FDMA (channel bandwidth 36 MHz) (Dicks and Brown [4], © 1974 IEEE).

3. IMPLEMENTATION ALTERNATIVES

As we have noted, FDMA is a very basic technology that can be combined with almost all the other access techniques. Figure 4a shows FDD/FDMA for analog and digital transmission. In frequency-division duplex (FDD) there is a group of K subbands for transmission in one direction and a similar contiguous group of K subbands for transmission in the reverse direction. A band of frequencies separates the two groups. Each station is allocated a subband in both FDD bands for the duration of its call. All the first generation analog cellular systems use FDD/FDMA. FDMA can also be used with TDD (time-division duplex). Here only one band is provided for transmissions, so a timeframe structure is used allowing transmissions to be done during one-half of the frame while the other half of the frame is available to receive signals. TDD/FDMA is used in cordless communications (CT2). The TDMA/FDMA structure is used in GSM (Global System for Mobile Communication) systems (Fig. 4b), where carriers of 200 kHz bandwidth carry a frame of 8 time slots.

Further generalization of the strict fixed-assignment FDMA system is possible and has been implemented in commercial products. Frequency-hopping schemes in FDMA have been proposed for cellular systems. Hopping is based on a random sequence. In particular, slow frequency-hopped spread-spectrum (FHSS) FDMA systems, combined with power and spectrally efficient modulation techniques such as FQPSK, can have significantly increased capacity over other access methods. The GSM system supports the possibility for frequency hopping. FDMA systems in which users can randomly access each channel with an Aloha-type attempt have also been proposed.

An interesting variation of FDMA is the OFDMA (orthogonal frequency–division multiple access) scheme proposed for wideband communications. In OFDMA, multiple access is achieved by providing each user with a number of the available subcarriers, which are now

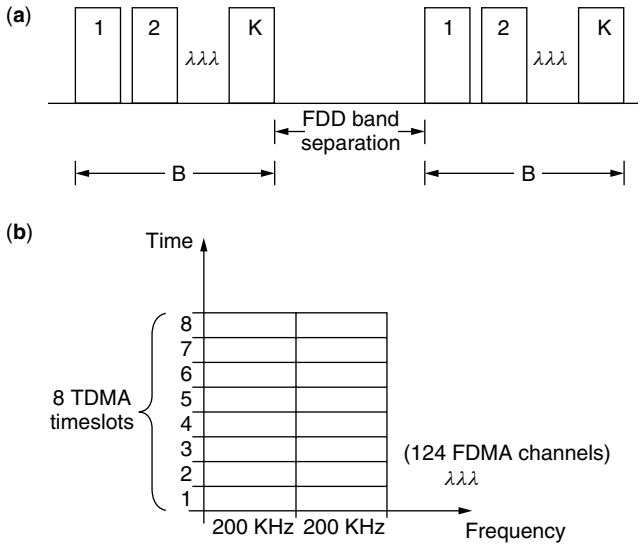
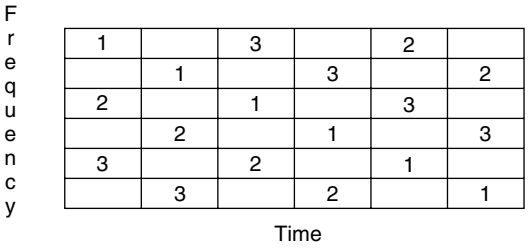


Figure 4. (**a**) FDMA/FDD arrangement; (**b**) TDMA/FDMA for GSM system.



1		3		2	
	1		3		2
2		1		3	
	2		1		3
3		2		1	
	3		2		1

Figure 5. OFDMA with frequency hopping.

orthogonal to each other. So there is no need for the relatively large guard bands necessary in conventional FDMA. An example of an OFDMA/TDMA time-frequency grid is shown in Fig. 5, where users 1, 2, and 3 each use a certain part of the available subcarriers. The part can be different for each one. Each user can have a fixed set of subcarriers, but it is relatively easy to allow hopping with different hopping patterns for each user and to result in orthogonal hopping schemes. OFDMA has been proposed for the European UMTS (Universal Mobile Telecommunications System) [9].

4. COMPARISON WITH OTHER TECHNIQUES

The comparison between the FDMA and other access methods is based on a number of performance and implementation criteria, the importance of which is varying very much depending on the type of system in which the access method is to be employed.

In digital transmission TDMA appears to be more "natural," so today most of the systems operate in this scheme or at least in combination with FDMA schemes. TDMA offers format flexibility since time-slot assignments among the multiple users are readily adjusted to provide different access rates for different users.

In some circumstances, FDMA schemes may be comparatively inefficient since they require guard bands between neighboring bands to prevent interfering phenomena, so this results in a waste of the system resources.

Another problem with FDMA systems is that they suffer from system nonlinearities; for example, they require the satellite transponder to be linear in nature, which cannot be achieved in practice. However, this is not necessarily as important in terrestrial communication systems, where the power consumption in the base station electronics is not a major design issue.

Referring to throughput performance, FDMA and TDMA should provide the same capability for carrying information over a network, and this is true with respect to bit-rate capability, if we neglect all overhead elements such as guard bands in FDMA and guard times in TDMA.

For K users generating data at a constant uniform rate and for an overall rate capability of the system equal to R bits per second, R/K bps (bits per second) is available to each user in both systems. Furthermore, both systems have the same capacity-wasting properties, because if a user has nothing to transmit, its frequency band cannot be used by another user.

However, if we examine the average delay, assuming that each user is transmitting every T seconds, the delay

is found to be $D_{\text{FDMA}} = T$ for the FDMA system while it is $D_{\text{TDMA}} = D_{\text{FDMA}} - T/2\{1 - 1/K\}$ in TDMA [1]. Therefore, for two or more users TDMA is superior to FDMA. Note that for large numbers of users the difference in packet delay is approximately $T/2$. Because of these problems, fixed assignment strategies have increasingly tended to shift from FDMA to TDMA as certain technical problems associated with the latter were overcome.

On the other hand, in transmission environments where spurious narrowband interference is a problem, the FDMA format, with a single user channel per carrier, has an advantage compared to TDMA in that a narrowband interferer can impair the performance of only one user channel.

The major advantage in the implementation of FDMA systems is that FDMA is a very mature technology and has the advantage of inexpensive terminals. Furthermore, channel assignment is simple and straightforward, and no network timing is required. This absence of synchronization problems is very attractive in fading communications channels.

Finally, in multipath fading environments, in a typical FDMA system, channel bandwidth is usually smaller than the coherence bandwidth of the transmission channel and there is no need to use an adaptive equalizer at the receiver. But at the same time, this situation removes the opportunity for the implicit frequency diversity gains that are achievable when signal bandwidth approaches the coherence bandwidth. Of course, in multipath environments CDMA is proved to be very effective. Its immunity to external interference and jamming, its low probability of intercept, and the easy integration of voice/data messages it offers establish its use in future multimedia communications systems. But a combination of FDMA/CDMA can improve the capacity of the system.

A detailed comparison of FDMA with other multiple access techniques can be found in well-known textbooks [5,6]. Schwartz et al. have provided a very fundamental comparison [10].

5. BASIC APPLICATIONS

FDMA was the first method of multiple access used for *satellite* communication systems (fixed, broadcasting, and mobile services) since around 1965 and will probably remain so for quite some time. In 1969 three satellites, Intelsat III, provided the first worldwide satellite service via analog FDM/FM/FDMA techniques for assemblies of 24, 60, or 120 telephone channels, while later, in Intelsat VI, 792 or 972 voice channels were delivered. Also, special techniques are applied for the distribution of audio and video channels on the same transponder (TV/FM/FDMA) [6]. Further, as a good example of digital operation in the FDMA mode, we can refer to Intelsat's *intermediate data rate* (IDR) carrier system described by Freeman [5].

A specific FDMA application is the single-channel-per-carrier (SCPC) system; one of the best known is the Intelsat (*SPADE*) system described above. SCPC systems are most suitable for use where each terminal is required to provide only a small number of telephone circuits, referred to as a *thin-route service*. The *satellite multiservice system* (SMS), provided by Eutelsat for business applications, has up to 1600 carriers in its 72 MHz bandwidth, which can be reallocated extremely quickly in response to changes in demand. FDMA has been used also for *VSAT* (very small-aperture terminals) satellite access, especially as DAMA method with FDMA/SCPC application. Most of today's commercial networks are based on demand assignment FDMA due to simple network control needed. Detailed VSAT implementations can be easily found in the literature.

At the present time, all *mobile satellite systems* that offer voice services have adopted a frequency–division multiple-access approach. The usual channel bandwidth is around 30 kHz, but figures of 22.5, 36, and 45 kHz are also used. The use of FDMA in conjunction with common channel signaling enables future developments in modulation and processing technologies to be easily incorporated into the system and allows a variety of communication standards to be supported. Also, as growth in the mobile terminal population occurs, incremental increases in the required system spectrum allocation can be easily assigned [6].

The modulation and multiple-access techniques used in the *IRIDIUM* system are patterned after the GSM terrestrial cellular system. A combined FDMA-TDMA access format is used along with data or vocoded voice and digital modulation techniques. Each FDMA frequency slot is 41.67 kHz wide, including guard bands, and supports 4-duplex voice channels in a TDMA arrangement [5].

All the *analog terrestrial mobile systems* were based on FDMA techniques. The band of frequencies was divided into segments, and half of the contiguous segments are assigned to outbound and the other to inbound (FDD/FDMA) cell sites with a guard band between outbound and inbound contiguous channels. A key question in these systems was the actual width of one user segment (e.g., in North American amperage, the segment width was 30 kHz).

FDMA is also applied in the contemporary cellular systems and in VHF and UHF land-mobile radio systems. The well-known *GSM* standard is based on a TDMA/FDMA combination that has been considered a very efficient scheme. The European digital cordless phone (DECT) is also based on a TDMA/TDD/FDMA format. In the United States in several FCC-authorized frequency bands, particularly those below 470 MHz, the authorized bandwidth per channel is limited to the 5–12.5-kHz range. In these narrowband mobile or cellular systems, digital FDMA could offer the most spectral- and cost-efficient solutions.

A very interesting application of the FDMA concept is found in the third-generation optical networks under the name of *wavelength-division multiple access* (WDMA) [11], meaning that the bits of the message are addressed on the basis of different wavelengths. WDMA networks have been investigated and prototyped at the laboratory level by a number of groups such as British Telecom Laboratories and AT&T Bell Laboratories.

In conclusion, we can state that FDMA as a stand-alone concept or as a basic component in hybrid implementations

is being applied and will be applied in the future in most communications systems.

BIOGRAPHY

Fotini-Niovi Pavlidou received her Ph.D. degree in electrical engineering from the Aristotle University of Thessaloniki, Greece, in 1988 and the Diploma in mechanical–electrical engineering in 1979 from the same institution.

She is currently an associate professor at the Department of Electrical and Computer Engineering at the Aristotle University engaged in teaching for the under- and postgraduate program in the areas of mobile communications and telecommunications networks. Her research interests are in the field of mobile and personal communications, satellite communications, multiple access systems, routing and traffic flow in networks, and QoS studies for multimedia applications over the Internet.

She is involved with many national and international projects in these areas (Tempus, COST, Telematics, IST) and she has been chairing the European COST262 Action on "Spread Spectrum Systems and Techniques for Wired and Wireless Communications." She has served as a member of the TPC in many IEEE/IEE conferences and she has organized/chaired some conferences like the "IST Mobile Summit 2002," the 6th "International Symposium on Power Lines Communications-ISPLC2002," the "International Conference on Communications-ICT 1998," etc.

She is a permanent reviewer for many IEEE/IEE1 journals. She has published about 60 studies in refereed journals and conferences.

She is a senior member of IEEE, currently chairing the joint IEEE VT & AES Chapter in Greece.

BIBLIOGRAPHY

1. K. Pahlavan and A. Levesque, *Wireless Information Networks*, Wiley, New York, 1995.
2. V. K. Bhargava, D. Haccoun, R. Matyas, and P. P. Nuspl, *Digital Communications by Satellite-Modulation, Multiple Access and Coding*, Wiley, New York, 1981.
3. G. Maral and M. Bousquet, *Satellite Communications Systems; Systems, Techniques and Technology*, Wiley, New York, 1996.
4. R. M. Gagliardi, *Introduction to Communications Engineering*, Wiley, New York, 1988.
5. R. L. Freeman, *Telecommunications Transmission Handbook*, Wiley, New York, 1998.
6. W. L. Morgan and G. D. Gordon, *Communications Satellite Handbook*, Wiley, New York, 1989.
7. J. L. Dicks and M. P. Brown, Jr., Frequency division multiple access (FDMA) for satellite communications systems, paper presented at IEEE Electronics and Aerospace Systems Convention, Washington, DC, Oct. 7–9, 1974.
8. N. J. Muller, *Desktop Encyclopedia of Telecommunications*, McGraw-Hill, New York, 1998.
9. OFDMA evaluation report, *The Multiple Access Scheme Proposal for the UMTS Terrestrial Radio Air Interface* (*UTRA*) *System, Part 1: System Description and Performance Evaluation*, SMG2 Tdoc 362a/97, 1997.
10. J. W. Schwartz, J. M. Aein, and J. Kaiser, Modulation techniques for multiple-access to a hard limiting repeater, *Proc. IEEE* **54**: 763–777 (1966).
11. P. E. Green, *Fiber Optic Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

FREQUENCY SYNTHESIZERS

ULRICH L. ROHDE
Synergy Microwave Corporation
Paterson, New Jersey

1. INTRODUCTION

Frequency synthesizers are found in all modern communication equipment, and signal generators, particularly wireless communication systems such as cell phones, require these building blocks [1]. On the basis of a frequency standard, the synthesizer provides a stable reference frequency for the system.

Synthesizers are used to generate frequencies with arbitrary resolution covering the frequency range from audio to millimeterwave. Today, simple frequency synthesizers consist of a variety of synthesizer chips, an external voltage-controlled oscillator (VCO), and a frequency standard. For high-volume applications, such as cell phones, cordless telephones, walkie-talkies, or systems where frequency synthesizers are required, a high degree of integration is desired. The requirements for synthesizers in cordless telephones are not as stringent as in test equipment. Synthesizers in test equipment use custom building blocks and can be modulated to be part of arbitrary waveform generators [2].

The VCO typically consists of an oscillator with a tuning diode attached. The voltage applied to the tuning diode tunes the frequency of the oscillator. Such a simple system is a phase-locked loop (PLL). The stability of the VCO is the same as the reference. There are single-loop and multiloop PLL systems. Their selection depends on the characteristic requirements. Figure 1 shows the block diagram of a single loop PLL [3].

The PLL consists of a VCO, a frequency divider, a phase detector, a frequency standard, and a loop filter [4–7].

There are limits to how high the frequency-division ratio can be. Typically, the loop where the RF frequency is divided below 1 kHz, becomes unstable. This is due to the fact that microphonic effects of the resonator will unlock the system at each occurrence of mechanical vibration. At 1 GHz, this would be a division ratio of one million. To avoid such high division ratios, either multiloop synthesizers are created, or a new breed of PLL synthesizers called *fractional-N division synthesizers* will be considered. At the same time, direct digital synthesis is being improved. Using direct digital synthesis in loops can also overcome some of the difficulties associated with high division ratios. There are also combinations of techniques, which we will refer to as *hybrid synthesizers*. They will be covered here.

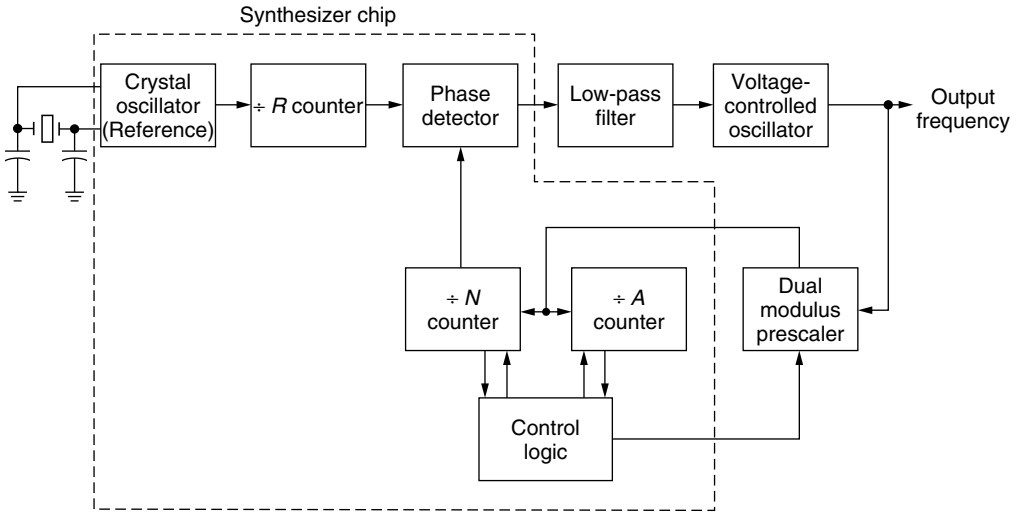


Figure 1. Block diagram of an integrated frequency synthesizer. In this case, the designer has control over the VCO and the loop filter; the reference oscillator is part of the chip. In most cases ($\leq$2.5 GHz), the dual-modulus prescaler is also inside the chip.

The quality of the system, or signal generator, is determined by the properties of the synthesizer and, of course, its building blocks. There are a variety of parameters in characterizing the synthesizer. To name a few important ones, we need to worry about the frequency stability, spurious suppression, and phase noise characteristics. The frequency resolution of the synthesizer tends to be covered by the switching speed and its spurious response. A lot of research is put into developing the ideal synthesizer, whatever ideal means. For portable applications such as cell phones, size and power consumption is a real issue, as well as the cost of the system. As mentioned, there are several competing approaches. In particular, there is a race between the fractional-N division synthesizer and the direct digital synthesis. The fractional synthesizer allows for generation of the output frequencies, which are not exact integers of the reference frequency. This results in an average frequency and problems with spurious sidebands. The direct digital synthesis uses a lookup table to construct a sine-wave, and the size of the lookup table and the sample rate determines the quality of the signal and its output frequency.

2. FREQUENCY SYNTHESIZER FUNDAMENTALS

There are several approaches to "synthesize" a frequency as we already mentioned. Probably the first and the oldest approach is called the direct frequency synthesis where a bank of crystals, as frequency standards, will be used to generate output frequencies. Such a system is called *frequency-incoherent*. There is no phase coherency between the various oscillators [8].

A simple example of direct synthesis is shown in Fig. 2. The new frequency $\frac{2}{3}f_0$ is realized from f_0 by using a divide-by-3 circuit, a mixer, and a bandpass filter. In this example $\frac{2}{3}f_0$ has been synthesized by operating directly on f_0.

Figure 3 illustrates the form of direct synthesis module most frequently used in commercial frequency

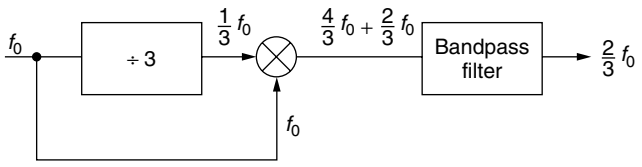


Figure 2. Direct frequency generation using the mix-and-divide principle. It requires excessive filtering.

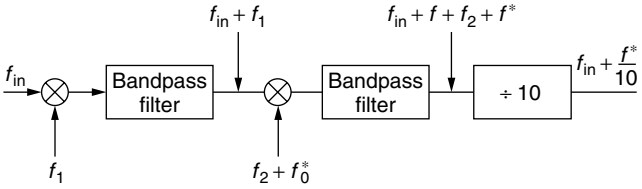


Figure 3. Direct frequency synthesizer using a mix-and-divide technique to obtain identical modules for high resolution.

synthesizers of the direct form. The method is referred to as the *double-mix-divide* approach.

An input frequency f_{in} is combined with a frequency f_1, and the upper frequency $f_1 + f_{in}$ is selected by the bandpass filter. This frequency is then mixed with a switch-selectable frequency $f_2 + f^*$ (in the following text f^* refers to any one of 10 switch-selectable frequencies). The output of the second mixer consists of the two frequencies $f_{in} + f_1 + f_2 + f^*$ and $f_{in} + f_1 - f_2 - f^*$; only the higher-frequency term appears at the output of the bandpass filter. If the frequencies f_{in}, f_1, and f_2 are selected so that

$$f_{in} + f_1 + f_2 = 10 f_{in} \tag{1}$$

then the frequency at the output of the divide by 10 will be

$$f_{out} = f_{in} + \frac{f^*}{10} \tag{2}$$

The double-mix-divide module has increased the input frequency by the switch-selectable frequency increment $f^*/10$. These double-mix-divide modules can be cascaded to form a frequency synthesizer with any degree of resolution. The double-mix-divide modular approach has the additional advantage that the frequencies f_1, f_2, and f_{in} can be the same in each module, so that all modules can contain identical components.

A direct frequency synthesizer with three digits of resolution is shown in Fig. 4. Each decade switch selects one of 10 frequencies $f_2 + f^*$. In this example the output of the third module is taken before the decade divider.

For example, it is possible to generate the frequencies between 10 and 19.99 MHz (in 10-kHz increments), using the three module synthesizer, by selecting

$$f_{in} = 1 \text{ MHz}$$

$$f_1 = 4 \text{ MHz}$$

$$f_2 = 5 \text{ MHz}$$

Since

$$f_{in} + f_1 + f_2 = 10 f_{in} \tag{3}$$

the output frequency will be

$$f_0 = 10 f_{in} = f_3^* + \frac{f_2^*}{10} + \frac{f_1^*}{100} \tag{4}$$

Since f^* occurs in 1-MHz increments, $f_1^*/100$ will provide the desired 10-kHz frequency increments.

Theoretically, either f_1 or f_2 could be eliminated, provided

$$f_{in} + f_1 (\text{or } f_2) = 10 f_{in} \tag{5}$$

but the additional frequency is used in practice to provide additional frequency separation at the mixer output. This frequency separation eases the bandpass filter requirements. For example, if f_2 is eliminated, $f_1 + f_{in}$ must equal $10 f_{in}$ or 10 MHz. If an f_1^* of 1 MHz is selected, the output of the first mixer will consist of the two frequencies 9 and 11 MHz. The lower of these closely spaced frequencies must be removed by the filter. The filter required would be extremely complex. If, instead, a 5-MHz signal f_2 is also used so that $f_{in} + f_1 + f_2 - 10$ MHz, the two frequencies at the first mixer output will (for an f_1^* of 1 MHz) be 1 and 11 MHz. In this case the two frequencies will be much easier to separate with a bandpass filter. The auxiliary frequencies f_1 and f_2 can be selected in each design only after considering all possible frequency products at the mixer output.

Direct synthesis can produce fast frequency switching, almost arbitrarily fine frequency resolution, low phase noise, and the highest-frequency operation of any of the methods. Direct frequency synthesis requires considerably more hardware (oscillators, mixers, and bandpass filters) than do the two other synthesis techniques to be described. The hardware requirements result in direct synthesizers becoming larger and more expensive. Another disadvantage of the direct synthesis technique is that unwanted (spurious) frequencies can appear at the output. The wider the frequency range, the more likely that the spurious components will appear in the output. These disadvantages are offset by the versatility, speed, and flexibility of direct synthesis.

2.1. PLL Synthesizer

The most popular synthesizer is based on PLL [9]. An example of such a PLL-based synthesizer is shown in Fig. 5. In preparing ourselves for hybrid synthesizers, it needs to be noted that the frequency standard can also be replaced by a direct digital synthesizer (DDS). The number of references for synthesizers seems endless, but the most complete and relevant ones are given at the end. Its also very important to monitor the patents. If any incremental improvements are achieved there, the inventors immediately try to protect them with a patent. The following will give some insight into the major building blocks used for PLL and hybrid synthesizers. It should be noted that most designers use a combination of available integrated circuits, and most of the high-performance solutions are due to the careful design of the oscillator, the integrated/lowpass filter, and the systems architecture. All of these items will be addressed, particularly the fractional-N synthesizer, which requires a lot of handholding in the removal of spurious products.

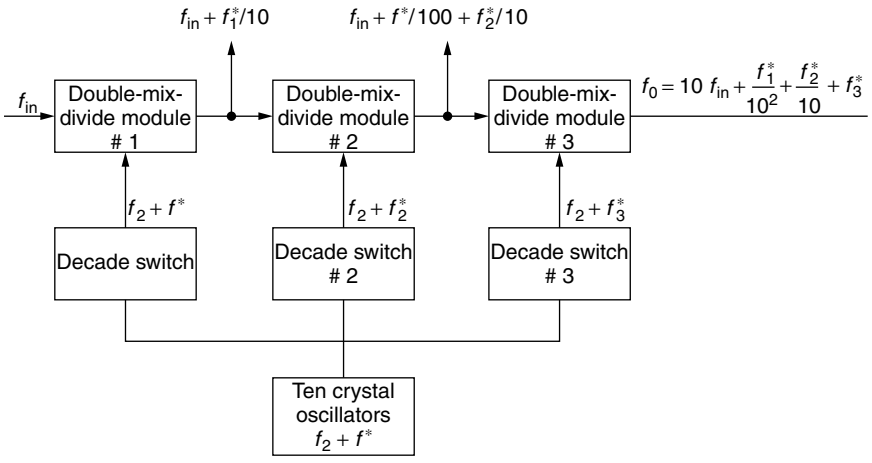


Figure 4. Phase-incoherent frequency synthesizer with three-digit resolution.

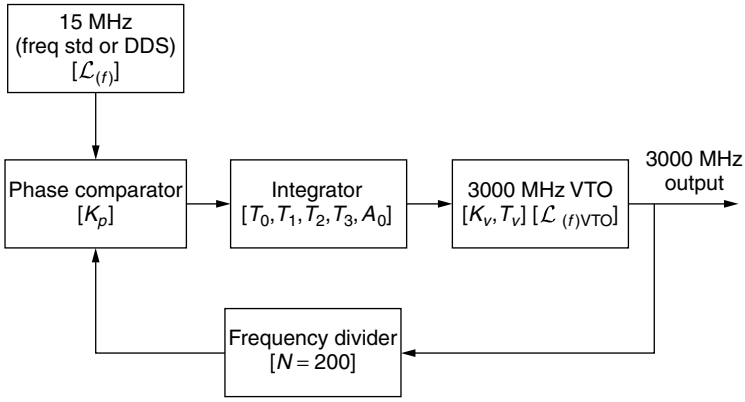


Figure 5. Block diagram of a PLL synthesizer driven by a frequency standard, DDS, or fractional-N synthesizer for high resolution at the output. The last two standards allow a relatively low division ratio and provide quasiarbitrary resolution.

How does the PLL work? According to Fig. 5, we have a free-running oscillator, which can operate anywhere from audio to millimeterwave. The output is typically sinusoidal. The VCO also is occasionally called a voltage-tuned oscillator (VTO), which drives an output stage and a pulseshaping stage to prepare the signal to drive a frequency divider chain. The frequency divider chain consists of silicon germanium or GaAs dividers to reduce the signal below 1000 MHz. At these frequencies, either silicon-based dividers or CMOS dividers can take over. The frequency divider, typically part of an integrated circuit, is a synchronous divider, which is programmable over a wide range. Division ratios as low as four and as high as one million are possible [10].

The output of the frequency divider is fed to a phase comparator, which in most cases is actually a phase-frequency detector. The phase frequency detector compares the output of the frequency divider, which typically is the same magnitude as the reference frequency, with a reference frequency, which is derived from a frequency standard. Frequency standards come as precision standards such as atomic frequency standards, followed by oven-controlled crystal oscillators, to temperature-compensated crystal oscillators. Sometimes even simple crystal oscillators will do the trick. The output from the phase comparator is a DC voltage typically between 1 and 25 V, which is applied to the tuning diode of the VTO or VCO. This tuning voltage is modulated by the differences of the two prior to lock. The frequency detector portion of the phase frequency comparator jams the voltage to one extreme charging the capacitors and integrator and acquiring frequency lock. After frequency lock is obtained, the control voltage will change the frequency at the output to have a fixed phase relationship compared to the reference frequency. The advantage of having a frequency detector in parallel to a phase detector is that the system always requires frequency lock [11,12].

2.2. Fractional-*N* Phase-Locked Loops

The principle of the fractional-N PLL synthesizer has been around for a while. In the past, implementation of this has been done in an analog system [13,14]. It would be ideal to be able to build a single-loop synthesizer with a 1.25- or 50-MHz reference and yet obtain the desired step-size resolution, such as 25 kHz. This would lead to

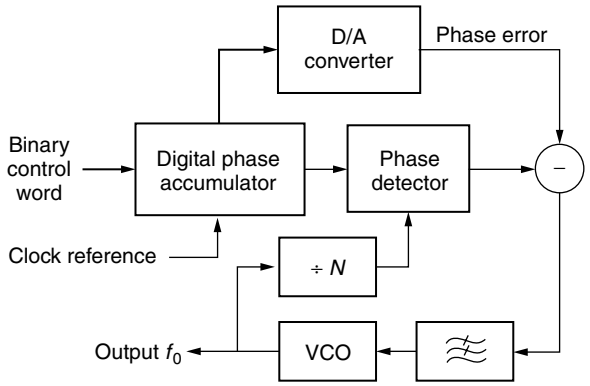


Figure 6. Simplified block diagram of an analog fractional-N synthesizer.

the much smaller division ratio and much better phase noise performance. Figure 6 shows the block of an analog fractional-N synthesizer.

An alternative would be for N to take on fractional values. The output frequency could then be changed in fractional increments of the reference frequency. Although a digital divider cannot provide a fractional division ratio, ways can be found to accomplish the same task effectively.

The most frequently used method is to divide the output frequency by $N + 1$ every M cycles and to divide by N the rest of the time. The effective division ratio is then $N + 1/M$, and the average output frequency is given by

$$f_0 = \left(N + \frac{1}{M}\right) f_r \tag{6}$$

This expression shows that f_0 can be varied in fractional increments of the reference frequency by varying M. The technique is equivalent to constructing a fractional divider, but the fractional part of the division is actually implemented using a phase accumulator. The phase accumulator approach is illustrated in Section 6. This method can be expanded to frequencies much higher than 6 GHz using the appropriate synchronous dividers. For more details, see Section 6 [15–53].

2.3. Digital Direct Frequency Synthesizer

The digital direct frequency uses sampled data methods to produce waveforms [54–69]. The digital hardware block

provides a datastream of k bits per clock cycle for digital-to-analog conversion (DAC). Ideally, the DAC is a linear device with glitch-free performance. The practical limits of the DAC will be discussed later in this article. The DAC output is the desired signal plus replications of it around the clock frequency and all of the clock's harmonics. Also present in the DAC output signal is a small amount of quantization noise from the effects of finite math in the hardware block. Figure 7 shows the frequency spectrum of an ideal DAC output with a digitally sampled sine-wave datastream at its input. Note that the desired signal, f_0 (a single line in the frequency domain), is replicated around all clock terms. Figure 8 shows the same signal in the time domain.

The DAC performs a sample-and-hold operation as well as converting digital values to analog voltages. The sample occurs on each rising edge of the clock; the hold occurs during the clock period. The transfer function of a sample-and-hold operator is a $(\sin x)/x$ envelope response with linear phase. In this case, $x = (\pi F/F_{\text{clock}})$. It should be noted that the sinc function rolloff affects the passband flatness. A 2.4 dB roll-off should be expected at 40% of F_{clock}.

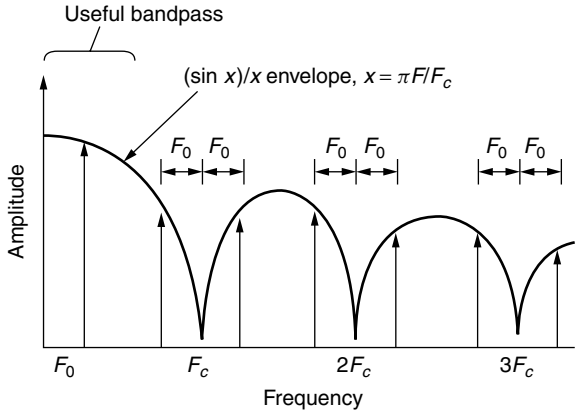


Figure 7. Ideal DAC output with F_0, a sampled-and-held sine wave, at its output. Notice the $(\sin x)/x$ envelope rolloff. As F_0 moves up in frequency, an aliased component $F_C - F_0$ moves down into the passband.

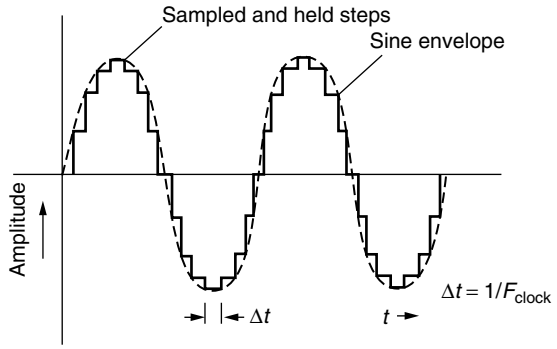


Figure 8. Samples per cycle sine wave. This is typical of a single-tone DAC output. $F_0 = F_{\text{clock}}/16$ after low pass filtering; only the sine envelope is present. The lowpass filter removes the sampling energy. Each amplitude step is held for a clock period.

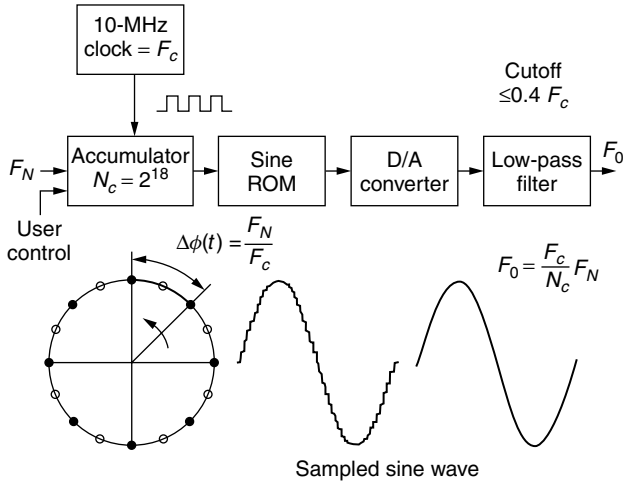


Figure 9. Direct digital frequency synthesizer.

Referring to Fig. 9, the output of the DAC is passed through a lowpass filter (LPF). With proper attention to design, an LPF may be realized that has linear phase in a flat passband with a width of $0.4F_{\text{clock}}$. With this design, the maximum available bandwidth is achieved. For example, with $F_{\text{clock}} = 125$ MHz, the useful synthesized bandwidth of about 50 MHz is attained. The LPF output is the desired signal without any sampling artifacts. Viewing the LPF strictly as a device to remove sampling energy, it is obvious why the output contains only the desired signal. It is also instructive to view the LPF from the time domain. From this point, the LPF may be seen as the perfect interpolator. It fills the space between time samples with a smooth curve to reconstruct perfectly the desired signal.

In the design of a DDS, the following guidelines apply:

- The desired frequency resolution determines the lowest output frequency f_L.
- The number of D/A conversions used to generate f_L is $N = 4k = 4f_U/f_L$, provided four conversions are used to generate $f_U (P = 4)$.
- The maximum output frequency f_U is limited by the maximum sampling rate of the DDS, $f_U \leq 1/4T$. Conversely, $T \leq 1/4f_U$.

To generate nf_L, the integer n addresses the register, and each clock cycle kn is added to the content of the accumulator so that the content of the memory address register is increased by kn. Each knth point of the memory is addressed, and the content of this memory location is transferred to the D/A converter to produce the output sampled waveform.

To complete the DDS, the memory size and length (number of bits) of the memory word must be determined. The word length is determined by system noise requirements. The amplitude of the D/A output is that of an exact sinusoid corrupted with the deterministic noise due to truncation caused by the finite length of the digital words (quantization noise). If an $(n + 1)$-bit word length (including one sign bit) is used and the output of the A/D

converter varies between ±1, the mean noise from the quantization will be

$$\rho^2 = \frac{1}{12}\left(\frac{1}{2}\right)^{2n} = \frac{1}{3}\left(\frac{1}{2}\right)^{2(n+1)} \tag{7}$$

The mean noise is averaged over all possible waveforms. For a worst-case waveform, the noise is a square wave with amplitude $\frac{1}{2}(\frac{1}{2})^n$ and $\rho^2 = \frac{1}{4}(\frac{1}{2})^{2n}$ for each bit added to the word length, the spectral purity improves by 6 dB.

The main drawback of a low-power DDS is that it is limited to relatively low frequencies. The upper frequency is directly related to the maximum usable clock frequency; today, the limit is about 1 GHz. DDS tends to be noisier than other methods, but adequate spectral purity can be obtained if sufficient lowpass filtering is used at the output. DDS systems are easily constructed using readily available microprocessors. The combination of DDS for fine frequency resolution plus other synthesis techniques to obtain higher-frequency output can provide high resolution with very rapid setting time after a frequency change. This is especially valuable for frequency-hopping spread-spectrum systems.

In analyzing both the resolution and signal-to-noise ratio (or rather signal-to-spurious performance) of the DDS, one has to know the resolution and input frequencies. As an example, if the input frequency is approximately 35 MHz and the implementation is for a 32-bit device, the frequency resolution compared to the input frequency is $35 \times 10^6 \div 2^{32} = 35 \times 10^6 \div 4.294967296 \times 10^9$ or 0.00815 Hz ≈0.01 Hz. Given the fact that modern shortwave radios with a first IF of about 75 MHz will have an oscillator between 75 and 105 MHz, the resolution at the output range is more than adequate. In practice, one would use the microprocessor to round it to the next increment of 1 Hz relative to the output frequency.

As to the spurious response, the worst-case spurious response is approximately $20 \log 2^R$, where R is the resolution of the digital/analog converter. For an 8-bit A/D converter, this would mean approximately 48 dB down (worst case), as the output loop would have an analog filter to suppress close-in spurious noise. Modern devices have a 14-bit resolution. Fourteen bits of resolution can translate into $20 \log 2^{14}$ or 80 dB, worse case, of suppression. The actual spurious response would be much better. The current production designs for communication applications, such as shortwave transceivers, despite the fact that they are resorting to a combination of PLLs and DDSs, still end up somewhat complicated. By using 10 MHz from the DDS and using a single-loop PLL system, one can easily extend the operation to above 1 GHz but with higher complexity and power consumption. This was shown in Fig 5. Figure 10 shows a multiple-loop synthesizer using a DDS for fine resolution.

3. IMPORTANT CHARACTERISTICS OF SYNTHESIZERS

The following is a list of parameters that are used to describe the performance of the synthesizer. These are referred to as figures of merit.

3.1. Frequency Range

The output frequency of a synthesizer can vary over a wide range. A synthesizer signal generator typically offers output frequencies from as low as 100 kHz to as high as several gigahertz. The frequency range is determined by the architecture of the signal generator as the system frequently uses complex schemes of combining frequencies

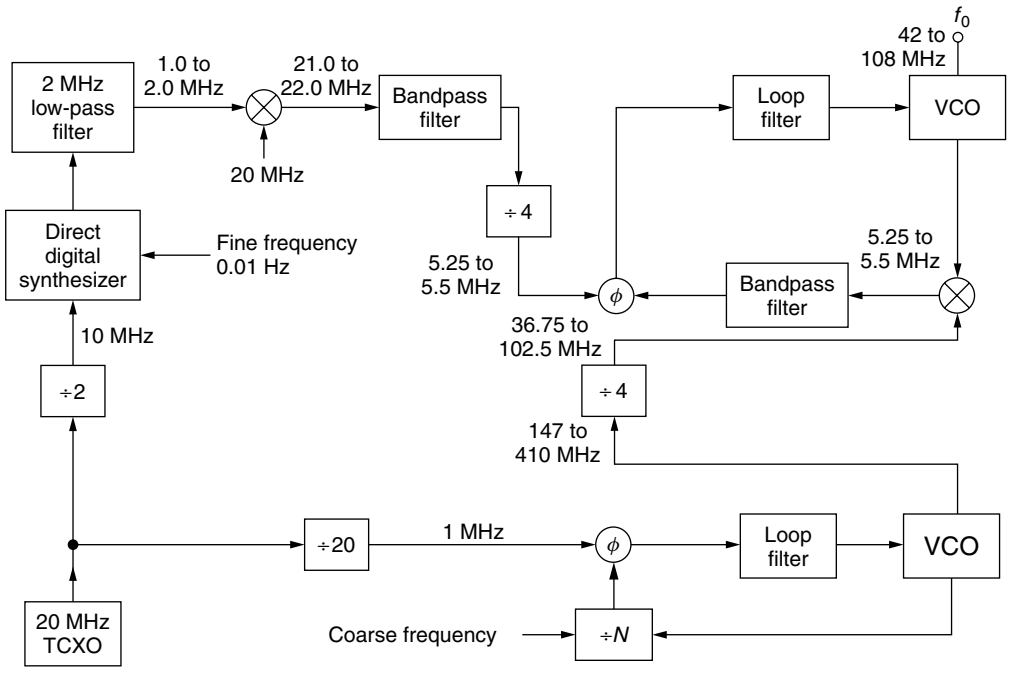


Figure 10. A multiple-loop synthesizer using a DDS for fine resolution.

in various loops. A standard loop-based synthesizer has a frequency range typically less than 1-2, as an example, 925–1650 MHz.

3.2. Phase Noise

Oscillators unfortunately are not clean, but the various noise sources in and outside of the transistor modulate the VCO, resulting in energy or spectral distribution on both sides of the carrier. This occurs via modulation and conversion. The noise, or better, FM noise is expressed as the ratio of output power divided by the noise power relative to 1 Hz bandwidth measured at an offset of the carrier. Figure 11 shows a typical phase noise plot of a synthesizer. Inside the loop bandwidth, the carrier signal is cleaned up, and outside the loop bandwidth, the measurement shows the performance of the VCO itself.

3.3. Output Power

The output power is measured at the designated output port of the frequency synthesizer. Practical designs require an isolation stage. Typical designs require one or more isolation stages between the oscillator and the output. The output power needs to be flat. While the synthesized generator typically is flat with only 0.1 dB +/− deviation, the VCO itself can vary as much as +/− 2 dB over the frequency range.

3.4. Harmonic Suppression

The VCO inside a synthesizer has a typical harmonic suppression of better than 15 dB. For high-performance applications, a set of lowpass filters at the output will reduce the harmonic contents to a desired level. Figure 12 shows a typical output power plot of a VCO.

3.5. Output Power as a Function of Temperature

All active circuits vary in performance as a function of temperature. The output power of an oscillator over a temperature range should vary less than a specified value, such as 1 dB.

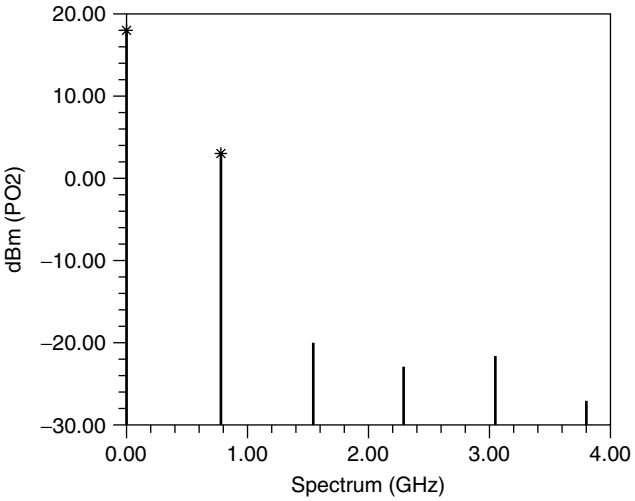


Figure 12. Predicted harmonics at the output of a VCO.

3.6. Spurious Response

Spurious outputs are signals found around the carrier of a synthesizer that are not harmonically related. Good, clean synthesizers need to have a spurious-free range of 90 dB, but these requirements make them expensive. While oscillators typically have no spurious frequencies besides possibly 60- and 120-Hz pickup, the digital electronics in a synthesizer generates a lot of signals, and when modulated on the VCO, are responsible for these unwanted output products. (See also Fig. 11.)

3.7. Step Size

The resolution, or step size, is determined by the architecture.

3.8. Frequency Pushing

Frequency pushing characterizes the degree to which an oscillator's frequency is affected by its supply voltage. For

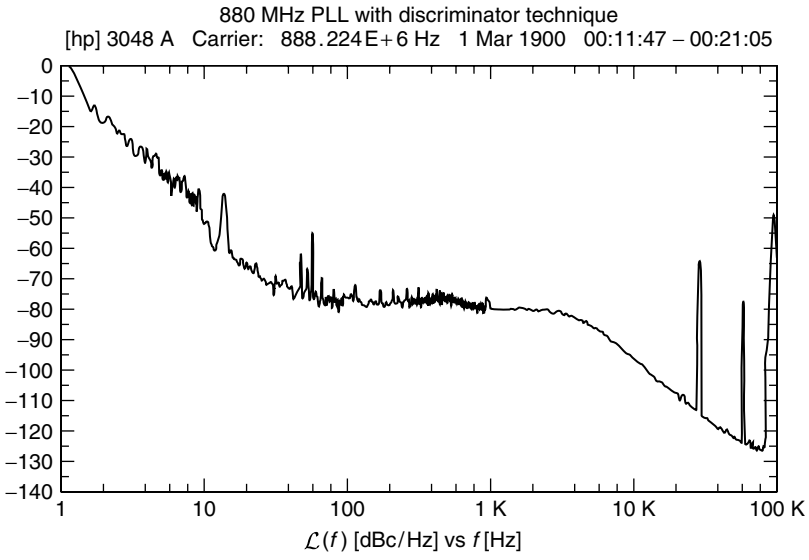


Figure 11. Measured phase noise of a 880-MHz synthesizer using a conventional synthesizer chip.

example, a sudden current surge caused by activating a transceiver's RF power amplifier may produce a spike on the VCO's DC power supply and a consequent frequency jump. Frequency pushing is specified in frequency/voltage form and is tested by varying the VCO's DC supply voltage (typically ± 1 V) with its tuning voltage held constant.

3.9. Sensitivity to Load Changes

To keep manufacturing costs down, many wireless applications use a VCO alone, without the buffering action of a high reverse-isolation amplifier stage. In such applications, frequency pulling, the change of frequency resulting from partially reactive loads, is an important oscillator characteristic. Pulling is commonly specified in terms of the frequency shift that occurs when the oscillator is connected to a load that exhibits a nonunity VSWR (such as 1.75, usually referenced to 50 Ω), compared to the frequency that results with unity-VSWR load (usually 50 Ω). Frequency pulling must be minimized, especially in cases where power stages are close to the VCO unit and short pulses may affect the output frequency. Such poor isolation can make phase locking impossible.

3.10. Tuning Sensitivity

This is a VCO parameter also expressed in frequency/voltage and is not part of a synthesizer specification.

3.11. Posttuning Drift

After a voltage step is applied to the tuning diode input, the oscillator frequency may continue to change until it settles to a final value. The posttuning drift is one of the parameters that limits the bandwidth of the VCO input.

3.12. Tuning Characteristic

This specification shows the relationship, depicted as a graph, between the VCO operating frequency and the tuning voltage applied. Ideally, the correspondence between operating frequency and tuning voltage is linear.

3.13. Tuning Linearity

For stable synthesizers, a constant deviation of frequency versus tuning voltage is desirable. It is also important to make sure that there are no breaks in tuning range, for example, that the oscillator does not stop operating with a tuning voltage of 0 V.

3.14. Tuning Sensitivity and Tuning Performance

This datum, typically expressed in megahertz per volt (MHz/V), characterizes how much the frequency of a VCO changes per unit of tuning voltage change.

3.15. Tuning Speed

This characteristic is defined as the time necessary for the VCO to reach 90% of its final frequency upon the application of a tuning voltage step. Tuning speed depends on the internal components between the input pin and the tuning diode, including the capacitance present at the input port. The input port's parasitic elements determine the VCO's maximum possible modulation bandwidth.

3.16. Power Consumption

This characteristic conveys the DC power, usually specified in milliwatts and sometimes qualified by operating voltage, required by the oscillator to function properly [70].

4. BUILDING BLOCKS OF SYNTHESIZERS

4.1. Oscillator

An oscillator is essentially an amplifier with sufficient feedback so the amplifier becomes unstable and begins oscillation. The oscillator can be divided into an amplifier, a resonator, and a feedback system [71]. One of the most simple equations describes this. It describes the input admittance of an amplifier with a tuned circuit at the output described by the term Y_L.

$$Y_{11}^{*} = Y_{11} - \frac{Y_{12} \times Y_{21}}{Y_{22} + Y_L} \tag{8}$$

The feedback is determined by the term Y_{12}, or in practical terms, by the feedback capacitor. The transistor oscillator (Fig. 13a) shows a grounded base circuit. For this type of oscillator, using microwave transistors, the emitter and collector currents are in phase. That means that the input circuit of the transistor needs to adjust the phase of the oscillation. The transistor stage forms the amplifier, the tuned resonator at the output determines the frequency of oscillation, and the feedback capacitor provides enough energy back into the transistor so that oscillation can occur. To start oscillation, the condition, relative to the output, can be derived in a similar fashion from the input:

$$Y_{22}^{*} = (Y_{22} + Y_L) - \frac{Y_{12} \times Y_{21}}{Y_{11} + Y_T} \tag{9}$$

If the second term of the equation on the right side is larger than the first term on the right of the = sign, then Real (Y_{22}^{*}) is negative and oscillation at the resonant frequency occurs.

The oscillator circuit shown in Fig 13c is good for low-frequency applications, but the Colpitts oscillator is preferred for higher frequencies. The Colpitts oscillator works by rotating this circuit and grounds the collector instead of the base. The advantage of the Colpitts oscillator is the fact that it is an emitter–follower amplifier with feedback and shows a more uniform gain of a wider frequency range (see Fig. 14). Oscillators with a grounded emitter or source have more stability difficulties over wider frequency ranges.

Depending on the frequency range, there are several resonators available. For low frequencies up to about 500 MHz, lumped elements such as inductors are useful. Above these frequencies, transmission line–based resonators are microwave resonators, which are better. Very

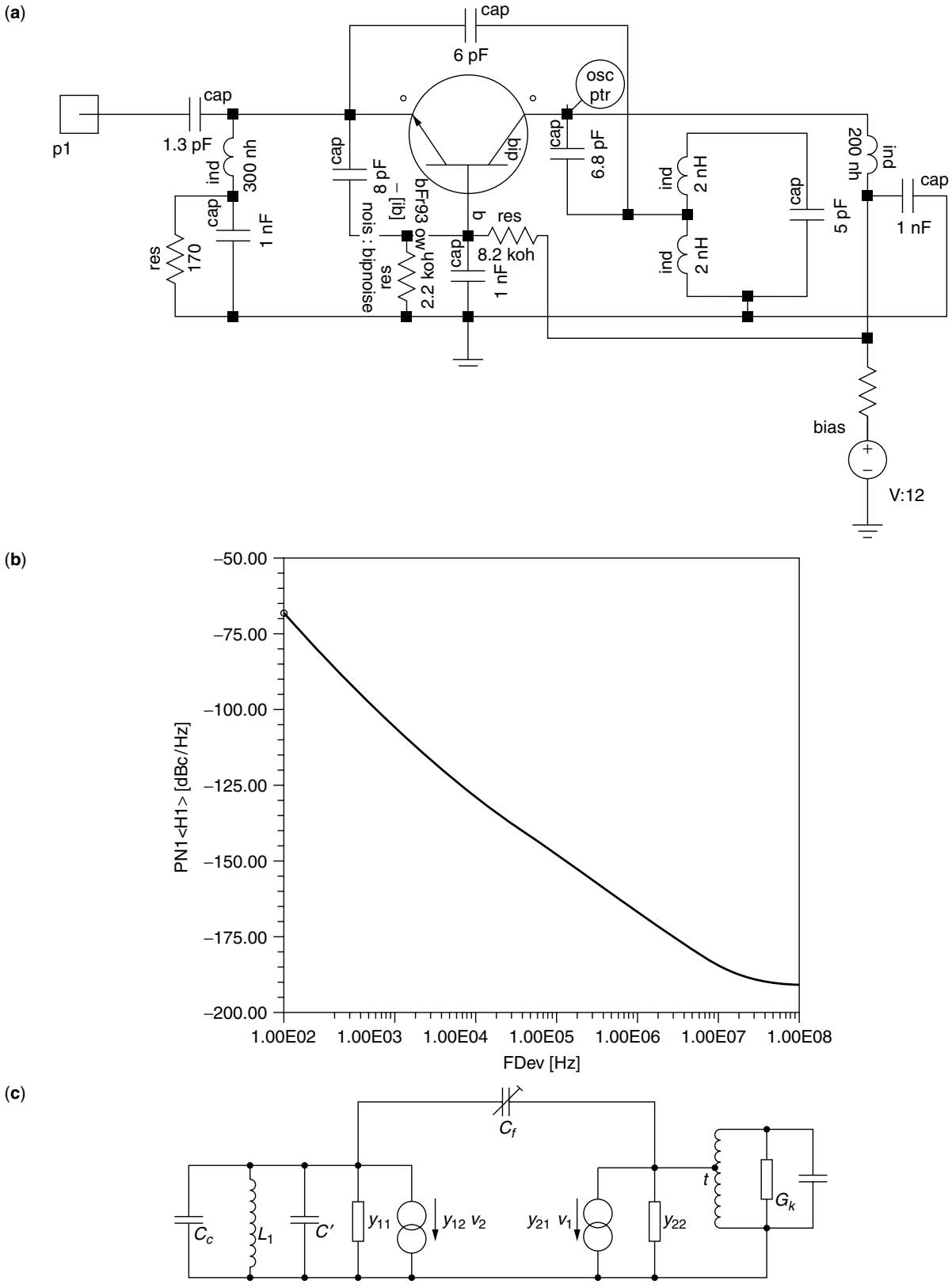


Figure 13. (**a**) A grounded base VHF/UHF oscillator with a tuned resonator tapped in the middle to improve operational Q; (**b**) the predicted phase noise of the oscillator in (**a**); **c** the equivalent circuit of the oscillator configuration of (**a**). For the grounded base configuration, emitter and collector currents have to be in phase. The phase shift introduced by C_f is compensated by the input tuned circuit.

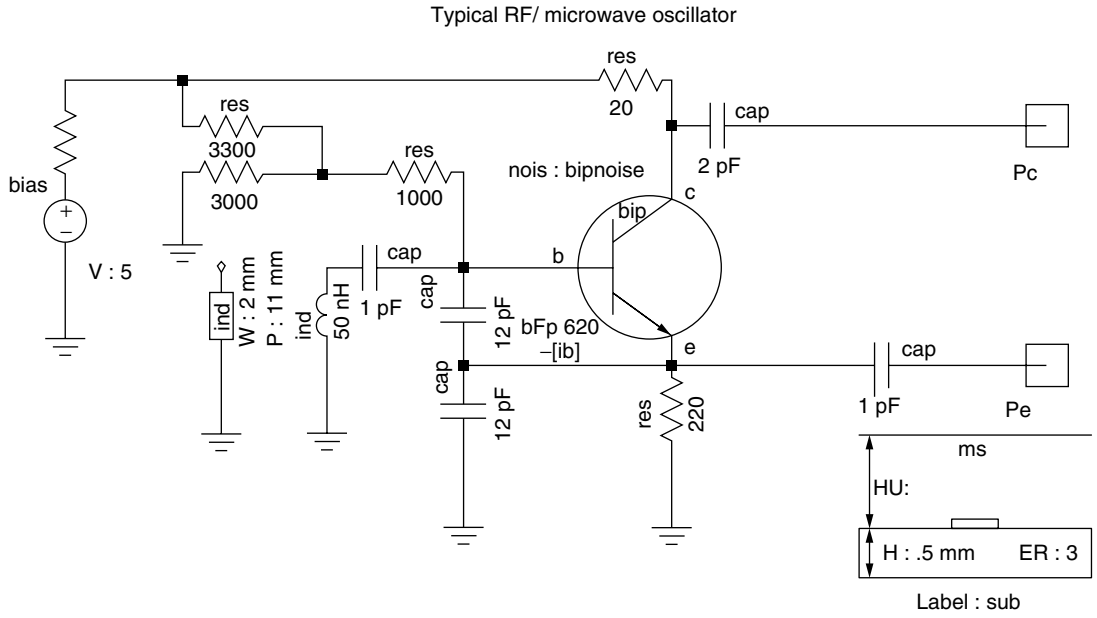


Figure 14. Colpitts oscillator using a coil inductor or a transmission line as a resonator. In this case all the physical parameters of the transmission line have to be provided.

small oscillators use printed transmission lines. High-performance oscillators use ceramic resonators (CROs), dielectric resonators (DROs), and surface acoustic wave resonators (SAWs), to name only a few. The frequency standard, if not derived from an atomic frequency standard is typically a crystal oscillator. Even the atomic frequency standards synchronize a crystal against an atomic resonance to compensate for the aging of the crystal. Figure 15a shows a circuit of a crystal oscillator with typical component values that are based on a 10-MHz third overtone AT cut crystal. Q is the ratio of stored energy divided by dissipated energy. The Q of the crystal can be as high as one million. The figure of merit Q is defined by $\omega L/R$. In our case $\omega L/R$ is

$$\frac{2\pi \times 1Hy \times 10 \times 10^6}{50} = 1.25 \text{ million}$$

Therefore, the resulting Q is 1.25 million. Typical resonator Q values for LC oscillators are 200; for structure-based resonators such as ceramic resonators or dielectric resonators, Q values of 400 are not uncommon.

An oscillator operating at 700 MHz is shown in Fig. 15a, including its schematic. CAD simulation was used to determine the resonance frequency phase noise and the harmonic contents, shown in Fig. 15b. The output power can be taken off either the emitter, at which provides better harmonic filtering, or from the collector, which provides smaller interaction between the oscillator frequency and the load. This effect is defined as *frequency pulling*. The tuned capacitor can now be replaced by a voltage-dependent capacitor. A tuning diode is a diode operated in reverse. Its PN junction capacitance changes as a function of applied voltage.

A two-port oscillator analysis will now be presented. It is based on the fact that an ideal tuned circuit (infinite Q), once excited, will oscillate infinitely because there is no resistance element present to dissipate the energy. In the actual case where the inductor Q is finite, the oscillations die out because energy is dissipated in the resistance. It is the function of the amplifier to maintain oscillations by supplying an amount of energy equal to that dissipated. This source of energy can be interpreted as a negative resistor in series with the tuned circuit. If the total resistance is positive, the oscillations will die out, while the oscillation amplitude will increase if the total resistance is negative. To maintain oscillations, the two resistors must be of equal magnitude. To see how a negative resistance is realized, the input impedance of the circuit in Fig. 16 will be derived.

If Y_{22} is sufficiently small ($Y_{22} \ll 1/R_L$), the equivalent circuit is as shown in Fig. 16. The steady-state loop equations are

$$V_{\text{in}} = I_{\text{in}}(X_{C_1} + X_{C_2}) - I_b(X_{C_1} - \beta X_{C_2}) \tag{10}$$

$$0 = -I_{\text{in}}(X_{C_1}) + I_b(X_{C_1+h_{ie}}) \tag{11}$$

After I_b is eliminated from these two equations, Z_{in} is obtained as

$$Z_{\text{in}} = \frac{V_{\text{in}}}{I_{\text{in}}} = \frac{(1+\beta)X_{C_1}X_{C_2} + h_{ie}(X_{C_1} + X_{C_2})}{X_{C_1} + h_{ie}} \tag{12}$$

If $X_{C_1} \ll h_{ie}$, the input impedance is approximately equal to

$$Z_{\text{in}} \approx \frac{1+\beta}{h_{ie}} X_{C_1}X_{C_2} + (X_{C_1} + X_{C_2}) \tag{13}$$

$$Z_{\text{in}} \approx \frac{-g_m}{\omega^2 C_1 C_2} + \frac{1}{j\omega[C_1C_2/(C_1+C_2)]} \tag{14}$$

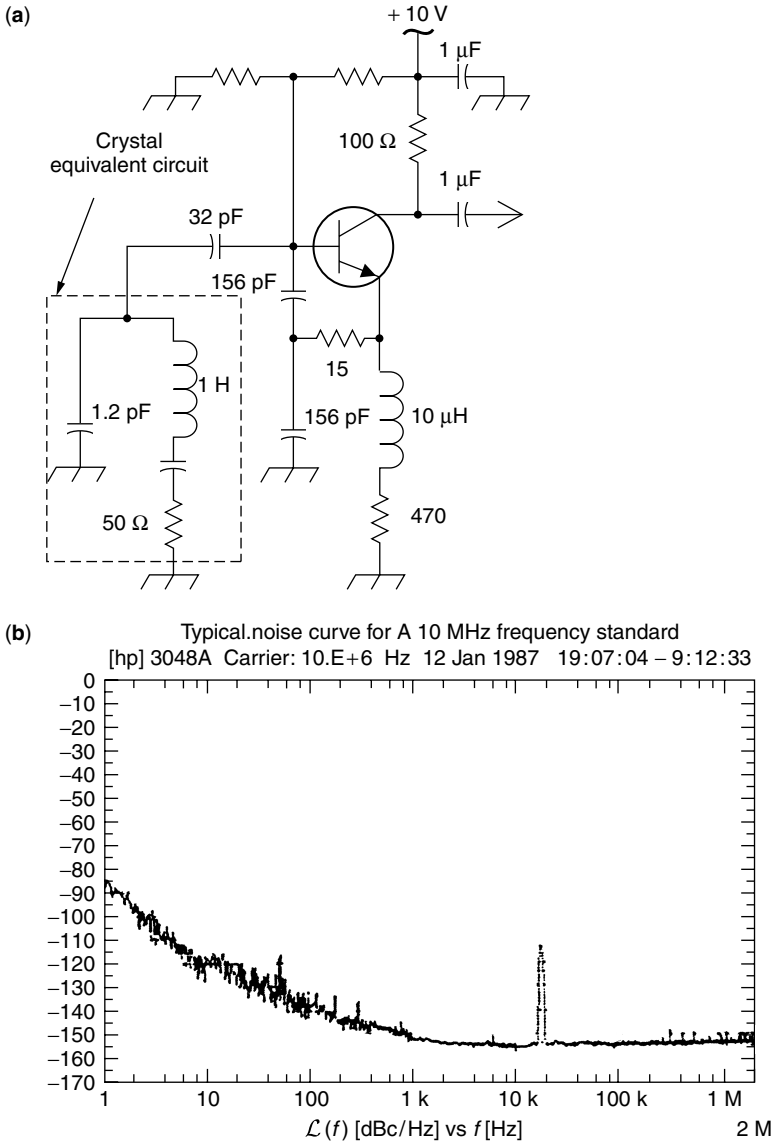


Figure 15. (**a**) Abbreviated circuit of a 10-MHz crystal oscillator; (**b**) measured phase noise for this frequency standard by HP of (**a**).

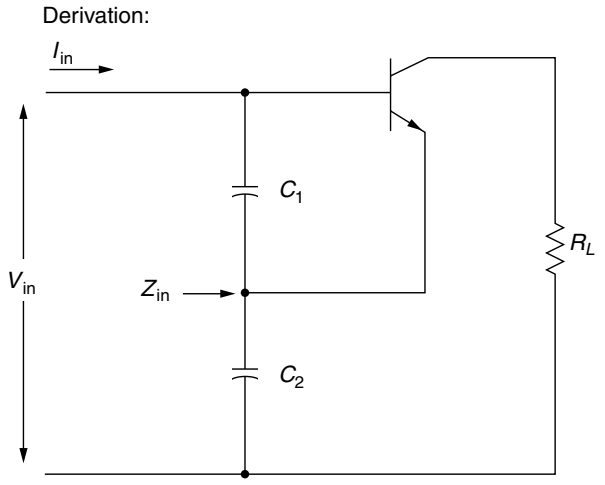


Figure 16. Calculation of input impedance of the negative-resistance oscillator.

That is, the input impedance of the circuit shown in Fig. 17 is a negative resistor

$$R = \frac{-g_m}{\omega^2 C_1 C_2} \tag{15}$$

in series with a capacitor

$$C_{\text{in}} = \frac{C_1 C_2}{C_1 + C_2} \tag{16}$$

which is the series combination of the two capacitors.

With an inductor L (with the series resistance R_S) connected across the input, it is clear that the condition for sustained oscillation is

$$R_S = \frac{g_m}{\omega^2 C_1 C_2} \tag{17}$$

and the frequency of oscillation

$$f_o = \frac{1}{2\pi\sqrt{L[C_1 C_2/(C_1 + C_2)]}} \tag{18}$$

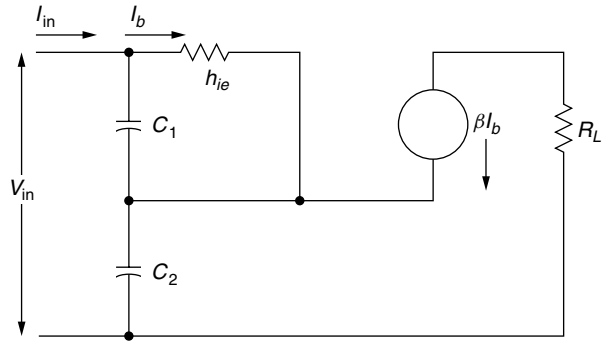


Figure 17. Equivalent small-signal circuit of Fig. 16.

This interpretation of the oscillator readily provides several guidelines that can be used in the design. First, C_1 should be as large as possible so that

$$X_{C_1} \ll h_{ie} \tag{19}$$

and C_2 is to be large so that

$$X_{C_2} \ll \frac{1}{Y_{22}} \tag{20}$$

When these two capacitors are large, the transistor base-to-emitter and collector-to-emitter capacitances will have a negligible effect on the circuit's performance. However, there is a maximum value of the capacitances since

$$r \leq \frac{g_m}{\omega^2 C_1 C_2} \leq \frac{G}{\omega^2 C_1 C_2} \tag{21}$$

where G is the maximum value of g_m. For a given product of C_1 and C_2, the series capacitance is at maximum when $C_1 = C_2 = C_m$. Thus

$$\frac{1}{\omega C_m} > \sqrt{\frac{r}{G}} \tag{22}$$

The design rule is

$$C_2 = C_1 \times \left| \frac{Y_{21}}{Y_{11}} \right| \tag{23}$$

This equation is important in that it shows that for oscillations to be maintained, the minimum permissible reactance $1/\omega C_m$ is a function of the resistance of the inductor and the transistor's mutual conductance, g_m. Figure 18 shows the resonant and oscillation condition for optimum performance. The negative real value should occur at $X = 0$!

An oscillator circuit known as the *Clapp circuit* or *Clapp–Gouriet circuit* is shown in Fig. 19. This oscillator is equivalent to the one just discussed, but it has the practical advantage of being able to provide another degree of design freedom by making C_0 much smaller than C_1 and C_2.

It is possible to use C_1 and C_2 to satisfy the condition of Eq. (20) and then adjust C_o for the desired frequency of

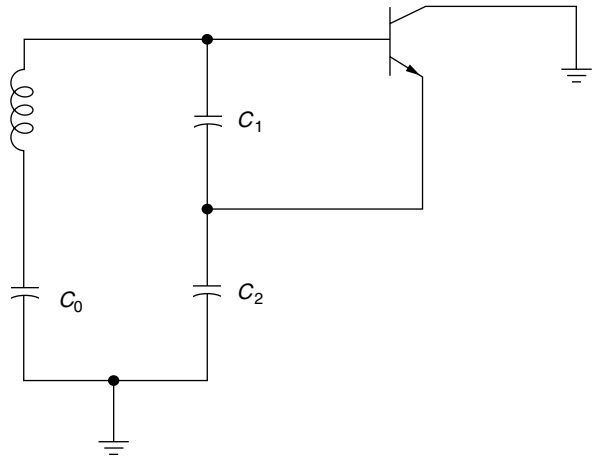


Figure 19. Circuit of a Clapp–Gouriet oscillator.

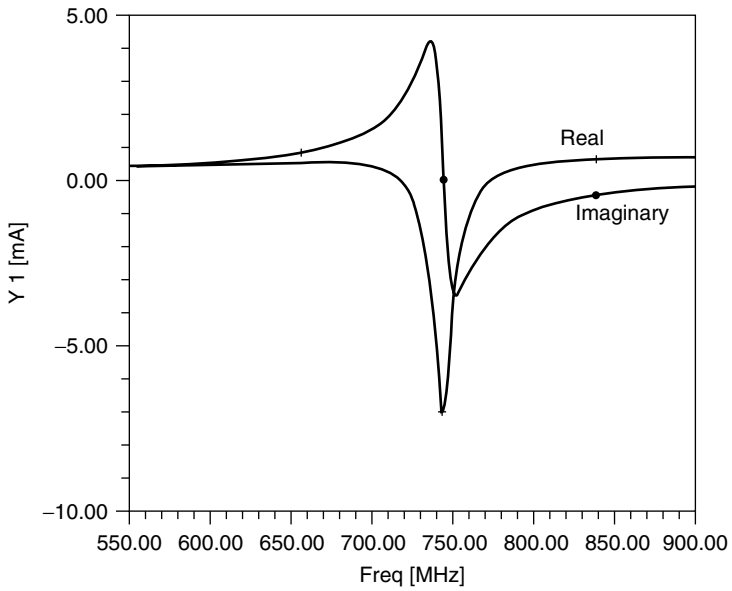


Figure 18. The CAD-based linearized currents. Conditions for oscillation are $X = 0$ and $R < 0$.

oscillation ω_o, which is determined from

$$\omega_o L - \frac{1}{\omega_o C_o} - \frac{1}{\omega_o C_1} - \frac{1}{\omega_o C_2} = 0 \tag{24}$$

Figure 20 shows the Clapp–Gouriet oscillator. Like the Colpitts, the Clapp–Gouriet obtains its feedback via a capacitive voltage divider; unlike the Colpitts, an additional capacitor series-tunes the resonator. The Pierce oscillator, a configuration used only with crystals, is a rotation of the Clapp–Gouriet oscillator in which the emitter is at RF ground [72–75].

4.1.1. Phase Noise. An estimate of the noise performance of an oscillator is as follows:

$$\mathcal{L}(\omega_m) = \frac{1}{8}\,\frac{FkT}{P_{\text{sav}}}\,\frac{\omega_0^2}{\omega_{\text{m}}^2}\left(\frac{P_{\text{in}}}{\omega_0 W_e} + \frac{1}{Q_{\text{unl}}} + \frac{P_{\text{sig}}}{\omega_0 W_e}\right)^2\left(1 + \frac{\omega_c}{\omega_m}\right)$$

Phase perturbation; Resonator Q; Flicker effect; Input power/reactive power ratio; Signal power/reactive power ratio

(25)

Equation (25) is based on work done by Dieter Scherer of Hewlett-Packard about 1978. He was the first to introduce the flicker effect to the Leeson equation by adding the AM-to-PM conversion effect, which is caused by the nonlinear capacitance of the active devices [76]. Figure 21 shows details of the noise contribution. This equation must be further expanded:

$$\pounds(f_m) = 10\log\left\{\left[1 + \frac{f_0^2}{(2f_m Q_{\text{load}})^2}\right]\left(1 + \frac{f_c}{f_m}\right) \times \frac{FkT}{2P_{\text{sav}}} + \frac{2kTRK_0^2}{f_m^2}\right\} \tag{26}$$

where $\pounds(f_m)$ = ratio of sideband power in 1-Hz bandwidth at f_m to total power in dB
f_m = frequency offset
f_0 = center frequency
f_c = flicker frequency
Q_{load} = loaded Q of the tuned circuit
F = noise factor
$kT = 4.1 \times 10^{-21}$ at 300 K_0 (room temperature)
P_{sav} = average power at oscillator output
R = equivalent noise resistance of tuning diode (typically 200 Ω–10 kΩ)
K = oscillator voltage gain

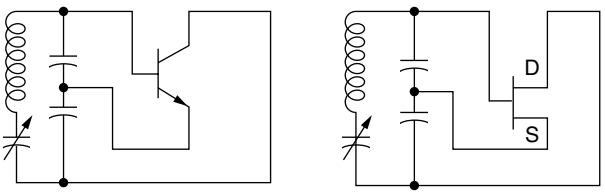
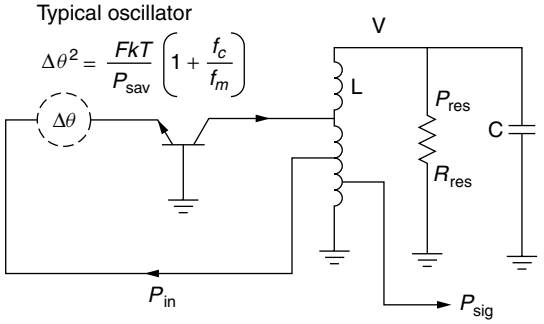


Figure 20. Clapp–Gouriet oscillator.



For $f_m < \frac{f_0}{2Q_{\text{load}}}$

$$\mathcal{L}(f_m) = \frac{1}{2}\,\frac{1}{\omega_{\text{m}}^2}\left(\frac{\omega_0}{2Q_{\text{load}}}\right)^2\,\frac{FkT}{P_{\text{sav}}}\left(1 + \frac{f_c}{f_m}\right)$$

$$Q_{\text{load}} = \frac{\omega_0 W_e}{P_{\text{diss . total}}} = \frac{\omega_0 W_e}{P_{\text{in}} + P_{\text{res}} + P_{\text{sig}}}$$

$$= \frac{\text{Reactive power}}{\text{Total dissipated power}}$$

Maximum energy in C or L : $W_e = \frac{1}{2}CV^2$

$$\mathcal{L}(\omega_m) = \frac{1}{8}\,\frac{FkT}{P_{\text{sav}}}\,\frac{\omega_0^2}{\omega_{\text{m}}^2}\left(\frac{P_{\text{in}}}{\omega_0 W_e} + \frac{1}{Q_{\text{unl}}} + \frac{P_{\text{sig}}}{\omega_0 W_e}\right)^2\left(1 + \frac{\omega_c}{\omega_m}\right)$$

Phase perturbation; Resonator Q; Flicker effect; Input power over reactive power; Signal power over reactive power

Figure 21. Diagram for a feedback oscillator showing the key components considered in the phase noise calculation and its contribution.

Table 1. Flicker Corner Frequency F_C as a Function of I_C

I_C (mA)	I_C (mA)
0.25	1
0.5	2.74
1	4.3
2	6.27
5	9.3

Source: Motorola

Table 1 shows the flicker corner frequency f_C as a function of I_C for a typical small-signal microwave BJT. $I_{C(\max)}$ of this transistor is about 10 mA.

Note that f_C, which is defined by AF and KF in the SPICE model, increases with I_C. This gives us a clue about how f_C changes when a transistor oscillates. As a result of the bias-point shift that occurs during oscillation, an oscillating BJT's average I_C is higher than its small-signal

I_C. KF is therefore higher for a given BJT operating as an oscillator than for the same transistor operating as a small-signal amplifier. This must be kept in mind when considering published f_C data, which are usually determined under small-signal conditions without being qualified as such. Determining a transistor's oscillating f_C is best done through measurement; operate the device as a high-Q UHF oscillator (we suggest using a ceramic-resonator-based tank in the vicinity of 1 GHz), and measure its close-in (10 Hz–10 kHz) phase noise–offset from the carrier. f_C will correspond to a slight decrease in the slope of the phase noise–offset curve. Generally, f_C varies with device type as follows: silicon JFETs, 50 Hz and higher; microwave RF BJTs, 1–10 kHz (as above); MOSFETs, 10–100 kHz; GaAs FETs, 10–100 MHz. Figure 22 shows the phase noise of oscillators using different semiconductors and resonators.

The additional term introduces a distinction between a conventional oscillator and a VCO. Whether the voltage- or current-dependent capacitance is internal or external makes no difference; it simply affects the frequency.

For a more complete expression for a resonator oscillator's phase noise spectrum, we can write

$$s_\phi(f_m) = \frac{\alpha_R F_0^4 + \alpha_E \left(\frac{F_0}{2Q_L}\right)^2}{f_m^3} + \frac{\left(\frac{2GFkT}{P_0}\right)\left(\frac{F_0}{2Q_L}\right)^2}{f_m^2} + \frac{2\alpha_R Q_L F_0^3}{f_m^2} + \frac{\alpha_E}{f_m} + \frac{2GFkT}{P_0} \tag{27}$$

where
G = compressed power gain of the loop amplifier
F = noise factor of the loop amplifier
k = Boltzmann's constant
T = temperature in kelvins
P_0 = carrier power level (in watts) at the output of the loop amplifier
F_0 = carrier frequency in hertz
f_m = carrier offset frequency in hertz
$Q_L (= \pi F_0 \tau_g)$ = loaded Q of the resonator in the feedback loop
α_R, α_E = flicker-noise constants for the resonator and loop amplifier, respectively

[77–86].

4.2. Frequency Divider

The output from the VCO has to be divided down to the reference frequency. The reference frequency can vary from a few kilohertz to more than 100 MHz [87–89]. A smaller division ratio provides better phase noise. Most of the frequency dividers are either off-the-shelf devices or custom devices. A typical frequency divider consists of a CMOS synchronous divider that can handle division ratios as low as 5 and as high as 1 million. The division ratio is determined by the number of dividers. Typical CMOS dividers end at 250 MHz. To extend the frequency range by using an asynchronous divider means extending the frequency range up to several gigahertz, but then the frequency resolution is compromised. This prescaler has to be a synchronized counter that has to be clocked by the main divider, but because of propagation delays, this can become difficult to achieve and can introduce phase jitter. A way around this is to use a dual-modulus prescaler, which toggles between two stages, dividing by N and dividing by $N + 1$. Dual-modulus counters are available in numbers such as $\frac{5}{6}$, $\frac{10}{11}$, and $\frac{20}{21}$.

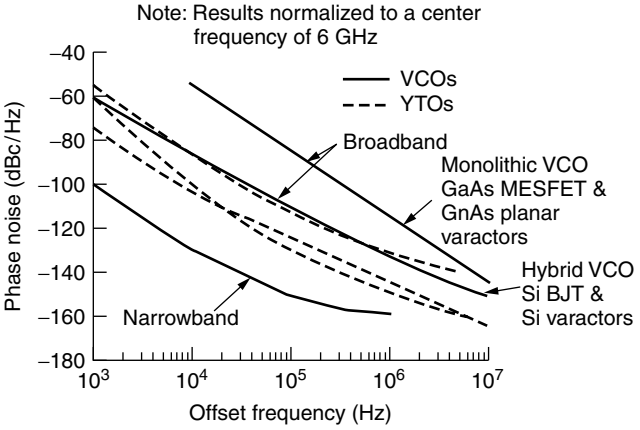


Figure 22. Phase noise of oscillators using different semiconductors and resonators.

Consider the system shown in Fig. 23. If the $P/(P+1)$ is a $\frac{10}{11}$ divider, the A counter counts the units and the M counter counts the tens. The mode of operation depends on the type of programmable counter used, but the system might operate as follows. If the number loaded into A is greater than zero, then the $P/(P+1)$ divider is set to divide by $P+1$ at the start of the cycle. The output from the $P/(P+1)$ divider clocks both A and M. When A is full, it ceases count and sets the $P/(P+1)$ divider into the P mode. Only M is then clocked, and when it is full, it resets both A and M, and the cycle repeats:

$$(M-A)P + A(P+1) = MP + A \tag{28}$$

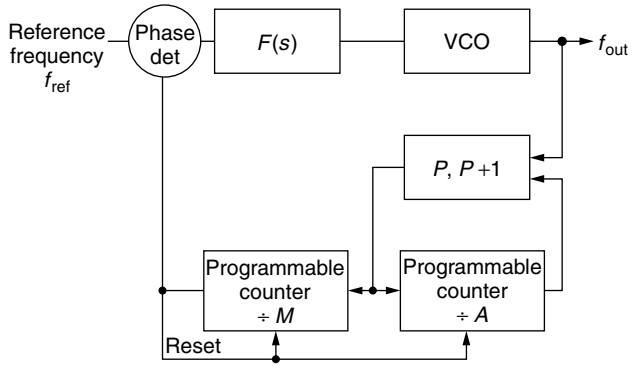


Figure 23. System using dual-modulus counter arrangement.

Therefore

$$f_{\text{out}} = (MP + A)f_{\text{ref}} \tag{29}$$

If A is incremented by one, the output frequency changes by f_{ref}. In other words, the channel spacing is equal to f_{ref}. This is the channel spacing that would be obtained with a fully programmable divider operating at the same frequency as the $P/(P+1)$ divider. For this system to work, the A counter must underflow before the M counter does; otherwise, $P/(P+1)$ will remain permanently in the $P+1$ mode. Thus, there is a minimum system division ratio, $M_{\min}$, below which the $P/(P+1)$ system will not function. To find that minimum ratio, consider the following. The A counter must be capable of counting all numbers up to and including $P-1$ if every division ratio is to be possible, or

$$A_{\max} = P - 1 \tag{30}$$

$$M_{\min} = P \text{ since } M > A \tag{31}$$

The divider chain divides by $MP+A$; therefore, the minimum systems division ratio is

$$\begin{aligned} M_{\min} &= M_{\min}(P + A_{\min}) \\ &= P(P+0) = p^2 \end{aligned} \tag{32}$$

Using a $\frac{10}{11}$ ratio, the minimum practical division ratio of the system is 100.

In the system shown in Fig. 23, the fully programmable counter, A, must be quite fast. With a 350-MHz clock to the $\frac{10}{11}$ divider, only about 23 ns is available for counter A to control the $\frac{10}{11}$ divider. For cost reasons it would be desirable to use a TTL fully programmable counter, but when the delays through the ECL-to-TTL translators have been taken into account, very little time remains for the fully programmable counter. The $\frac{10}{11}$ function can be extended easily, however, to give a $+N(N+1)$ counter with a longer control time for a given input frequency, as shown in Figs. 24 and 25. Using the $\frac{20}{21}$ system shown in Fig. 24, the time available to control $\frac{20}{21}$ is typically 87 ns at 200 MHz and 44 ns at 350 MHz. The time available to control the $\frac{40}{41}$ (Fig. 25) is approximately 180 ns at 200 MHz and 95 ns at 350 MHz.

Figure 26 is a block diagram of an advanced digital synthesizer block produced by analog devices. There are numerous manufacturers of such chips on the market. Figure 25 gives some insight into the frequency divider system. The top accepts input from a frequency standard, also referred to as a *reference signal*, which is reduced to a number between 5 kHz and 20 MHz. The use of a high-frequency reference requires a higher division ratio, but typically these reference frequencies are also used for some other mixing processes. The 24-bit input register controls both the reference divider and the frequency divider. The frequency divider uses a prescaler like $\frac{5}{6}$ or $\frac{10}{11}$, and its output is applied to the phase frequency detector. The multiplex unit on the right is doing all the housekeeping and providing information such as block and, detect. The divider typically has very little control over this portion of the synthesizer, and it's a constant battle to find better parts. Very few high-end synthesizers use custom ICs. To have these built is typically very costly and is cost-effective only if the frequency resolution required is not provided by any other chip on the market. Most of the high-end fractional-N synthesizer chips fall in this category.

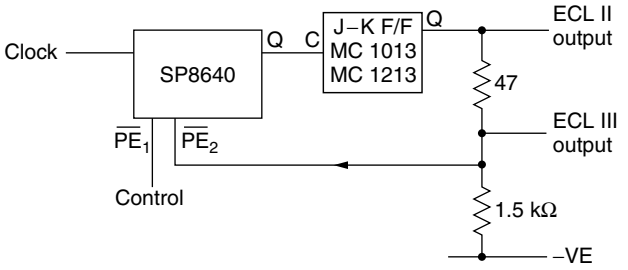


Figure 24. Level shifting information for connecting the various ECL2 and ECL3 stages.

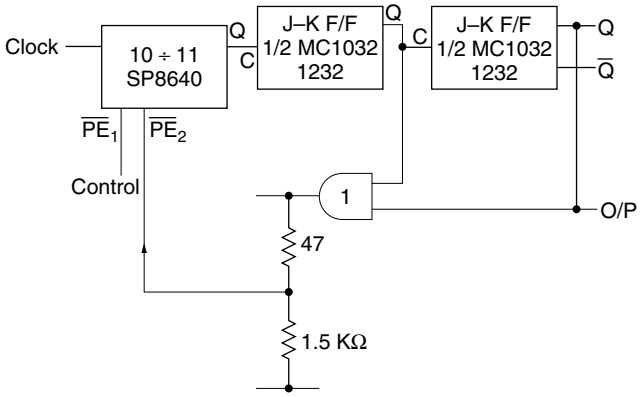


Figure 25. Level shifter diagram to drive from ECL2 and ECL3 levels.

4.3. Phase Detector

The phase detector at minimum consists of a phase-sensitive circuit such as a double balanced mixer [90–93]. Such a simple circuit has two disadvantages: (1) it's not sensitive to frequency changes, and (2) the DC output level is only 0.7 V per diode in the ring; therefore, a poor signal-to-noise ratio can be expected. Today, modern circuits use a phase discriminator with a charge pump output. The phase frequency is edge-triggered and sensitive to both phase and frequency changes. Figure 27 shows a digital phase frequency discriminator with a programmable delay. Under locked condition, the charge pump does not supply current. Under unlocked condition, the current at the point CP charges or discharges a capacitor, which is part of the integrated system and smooths the output voltage to become ripple-free. The output from the reference divider is a pulsetrain with a small duty cycle, and the input from the frequency divider(/N) is a pulsetrain with a very small duty cycle. The duty cycle typically is as short as a division ratio is high. So, for a division ratio of 1000, the duty cycle is 0.1%, but the repetition frequency is equal to the output. The charge output, therefore, is also a very narrow train of pulses that are fed to the loop filter. The phase detector has to deal with complicated issues such as zero crossings causing instability. The best phase frequency detectors are

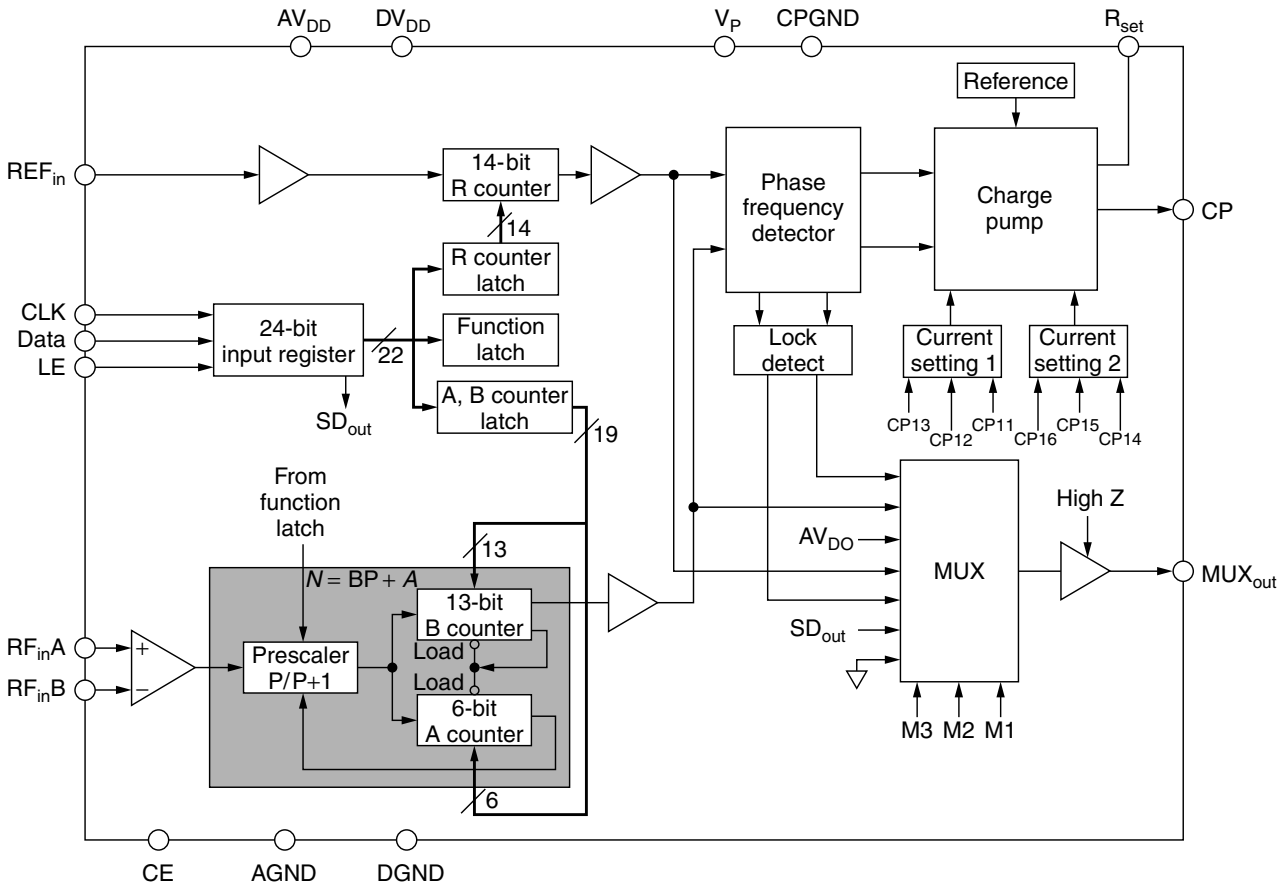


Figure 26. Block diagram of an advanced digital fractional-*N* synthesizer.

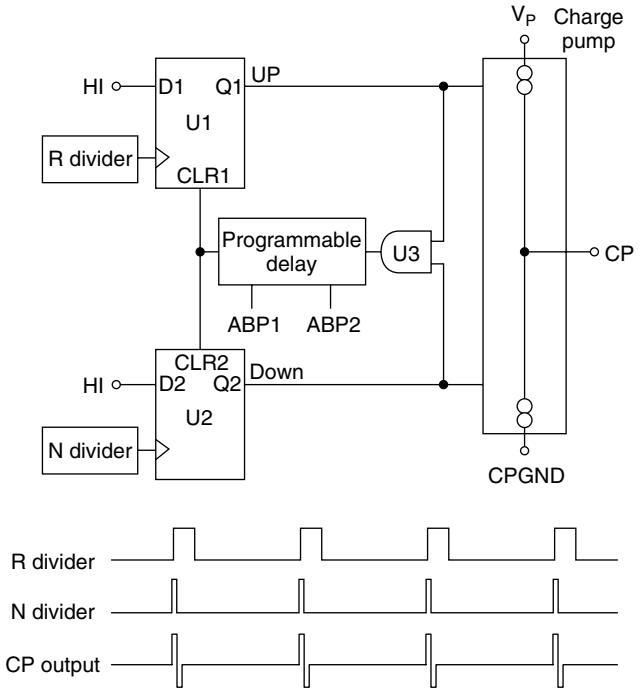


Figure 27. A digital phase frequency discriminator with a programmable delay.

either TTL or CMOS because they have a larger voltage output swing. In some cases, the charge pump is built from discrete components if a very high performance is required.

4.4. Loop Filter

Loop filters range from a simple lowpass filter to a complex arrangement of active filters. Figure 28 shows the configuration and frequency response of passive and active loop filters. Figure 29 shows an arrangement of more complex filters, including their calculations. The charge pumps can frequently lift with a purely passive filter, as seen in Fig. 30; however, the DC gain of the active filters provides better close-in phase noise and tracking. There may be a penalty if the active circuit is noise; however, the latest available operation amplifiers have sufficient performance [94].

5. PHASE-LOCKED LOOP DESIGNS

5.1. The Type 2, Second-Order Loop

The following is a derivation of the properties of the type 2, second-order loop. This means that the loop has two integrators, one being the diode and the other the operational amplifier, and is built with the order

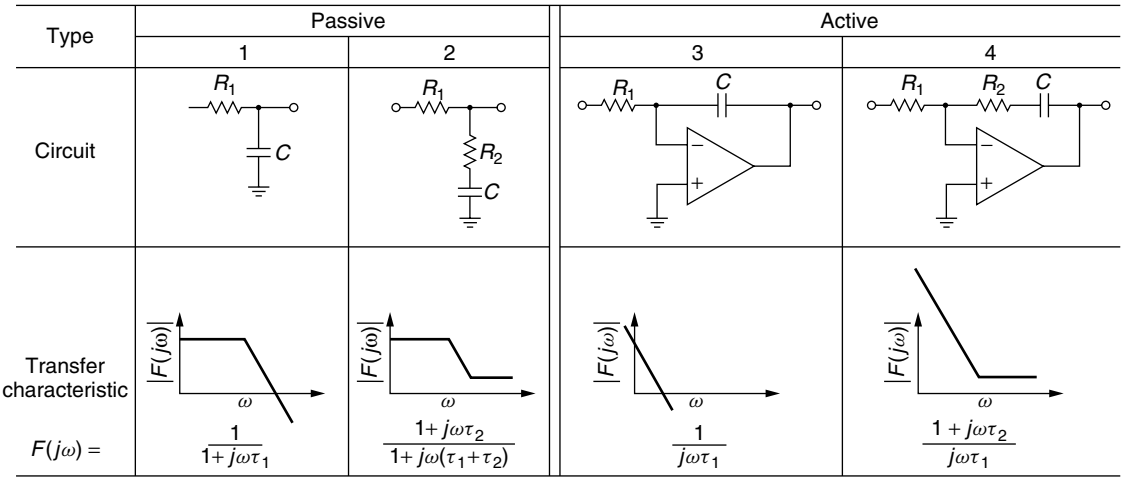


Figure 28. Circuit and transfer characteristics of several PLL filters.

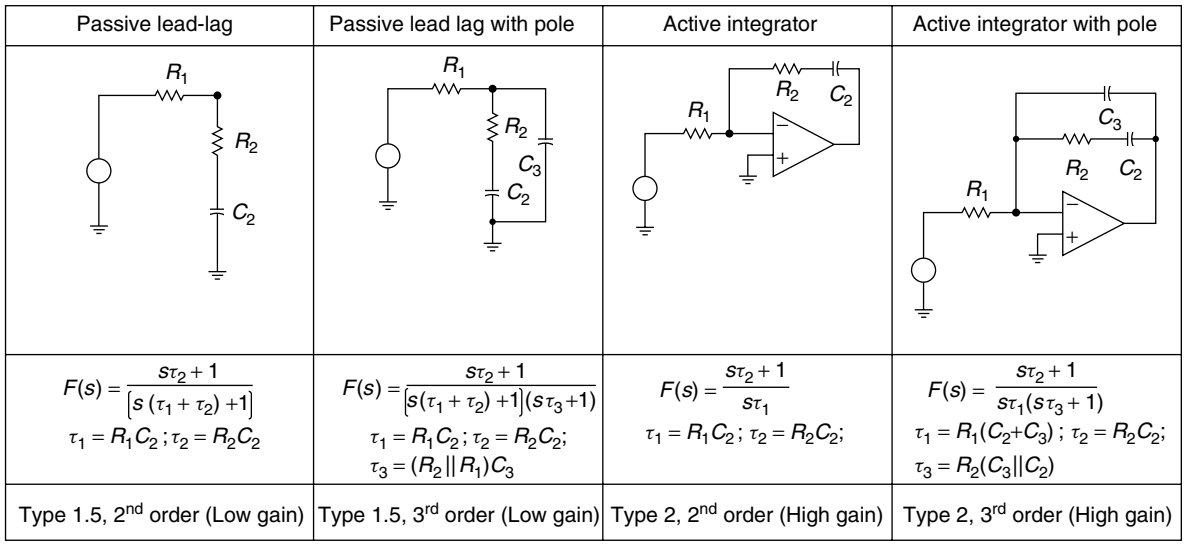


Figure 29. Implementation of different loop filters.

of 2 as can be seen from the pictures above. The basic principle to derive the performance for higher-order loops follows the same principle, although the derivation is more complicated. Following the math section, we will show some typical responses [95].

The type 2, second-order loop uses a loop filter in the form

$$F(s) = \frac{1}{s}\frac{\tau_2 s + 1}{\tau_1} \tag{33}$$

The multiplier $1/s$ indicates a second integrator, which is generated by the active amplifier. In Table 1, this is the type 3 filter. The type 4 filter is mentioned there as a possible configuration but is not recommended because, as stated previously, the addition of the pole of the origin creates difficulties with loop stability and, in most cases, requires a change from the type 4 to the type 3 filter. One can consider the type 4 filter as a special case of the type 3 filter, and therefore it does not have to be treated separately. Another possible transfer function is

$$F(s) = \frac{1}{R_1 C}\frac{1 + \tau_2 s}{s} \tag{34}$$

with

$$\tau_2 = R_2 C \tag{35}$$

Under these conditions, the magnitude of the transfer function is

$$|F(j\omega)| = \frac{1}{R_1 C \omega}\sqrt{1 + (\omega R_2 C)^2} \tag{36}$$

and the phase is

$$\theta = \arctan(\omega\tau_2) - 90^\circ \tag{37}$$

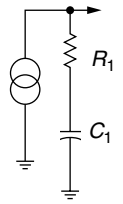
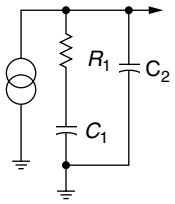
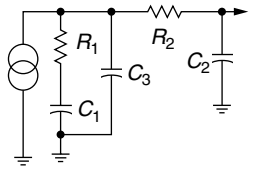
Integrator	Integrator with poles	Integrator with 2 poles
		
$F(s) = R_1 \dfrac{s\tau_1 + 1}{s\tau_1}$ $\tau_1 = R_1 C_1$	$F(s) = R_1 \dfrac{s\tau_1 + 1}{s\tau_1(s\tau_2 + 1)}$ $\tau_1 = R_1 C_1$; $\tau_2 = R_2\left(\dfrac{C_1 C_2}{C_1 + C_2}\right)$	$F(s) = R_1 \dfrac{s\tau_1 + 1}{s\tau_1(s\tau_2 + 1)(s\tau_3 + 1)}$ $\tau_1 = R_1 C_1$; $\tau_2 = R_1 \dfrac{C_1 C_3}{C_1 + C_3}$; $\tau_3 = R_2 C_2$
Type 2, 2nd order	Type 2, 3rd order	Type 2, 4th order

Figure 30. Recommended passive filters for charge pumps.

Again, as if for a practical case, we start off with the design values ω_n and ξ, and we have to determine τ_1 and τ_2. Taking an approach similar to that for the type 1, second-order loop, the results are

$$\tau_1 = \frac{K}{\omega_n} \tag{38}$$

and

$$\tau_2 = \frac{2\zeta}{\omega_n} \tag{39}$$

and

$$R_1 = \frac{\tau_1}{C} \tag{40}$$

and

$$R_2 = \frac{\tau_2}{C} \tag{41}$$

The closed-loop transfer function of a type 2, second-order PLL with a perfect integrator is

$$B(s) = \frac{K(R_2/R_1)[s + (1/\tau_2)]}{s^2 + K(R_2/R_1)s + (K/\tau_2)(R_2/R_1)} \tag{42}$$

By introducing the terms ξ and ω_n, the transfer function now becomes

$$B(s) = \frac{2\zeta\omega_n s + \omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{43}$$

with the abbreviations

$$\omega_n = \left(\frac{K}{\tau_2}\frac{R_2}{R_1}\right)^{1/2} \text{ rad/s} \tag{44}$$

and

$$\zeta = \frac{1}{2}\left(K\tau_2\frac{R_2}{R_1}\right)^{1/2} \tag{45}$$

and $K = K_\theta K_o / N$.

The 3-dB bandwidth of the type 2, second-order loop is

$$B_{3\text{ dB}} = \frac{\omega_n}{2\pi}\left[2\zeta^2 + 1 + \sqrt{(2\zeta^2 + 1)^2 + 1}\right]^{1/2} \text{ Hz} \tag{46}$$

and the noise bandwidth is

$$B_n = \frac{K(R_2/R_1) + 1/\tau_2}{4} \text{ Hz} \tag{47}$$

Again, we ask the question of the final error and use the previous error function

$$E(s) = \frac{s\theta(s)}{s + K(R_2/R_1)\{[s + (1/\tau_2)]/s\}} \tag{48}$$

or

$$E(s) = \frac{s^2\theta(s)}{s^2 + K(R_2/R_1)s + (K/\tau_2)(R_2/R_1)} \tag{49}$$

As a result of the perfect integrator, the steady-state error resulting from a step change in input phase or change of magnitude of frequency is zero.

If the input frequency is swept with a constant range change of input frequency $(\Delta\omega/dt)$, for $\theta(s) = (2\Delta\omega/dt)/s^3$, the steady-state phase error is

$$E(s) = \frac{R_1}{R_2}\frac{\tau_2(2\Delta\omega/dt)}{K} \text{ rad} \tag{50}$$

The maximum rate at which the VCO frequency can be swept for maintaining lock is

$$\frac{2\Delta\omega}{dt} = \frac{N}{2\tau_2}\left(4B_n - \frac{1}{\tau_2}\right) \text{ rad/s} \tag{51}$$

The introduction of N indicates that this is referred to the VCO rather than to the phase/frequency comparator. In the previous example of the type 1, first-order loop, we referred it only to the phase/frequency comparator rather than the VCO.

Figure 31 shows the closed-loop response of a type 2, third-order loop having a phase margin of 10° and with the optimal 45°.

A phase margin of 10° results in overshoot, which in the frequency domain would be seen as peaks in the oscillator noise–sideband spectrum. Needless to say, this is a totally undesirable effect, and since the operational amplifiers and other active and passive elements add to

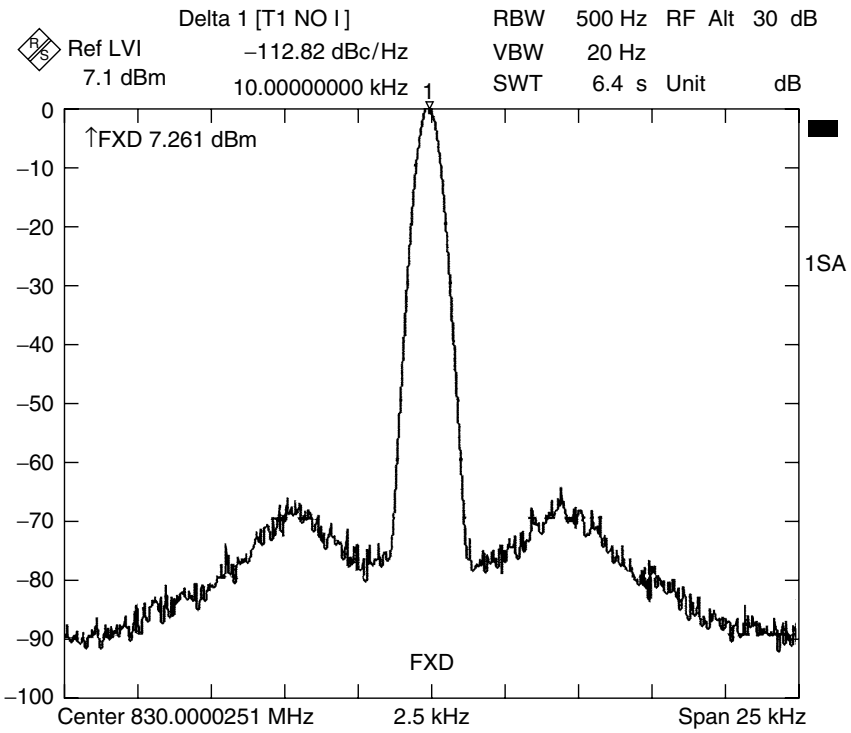


Figure 31. Measured spectrum of a synthesizer where the loop filter is underdamped, resulting in ≈10-dB increase of the phase noise at the loop filter bandwidth. In this case, we either don't meet the 45° phase margin criterion, or the filter is too wide, so it shows the effect of the up-converted reference frequency.

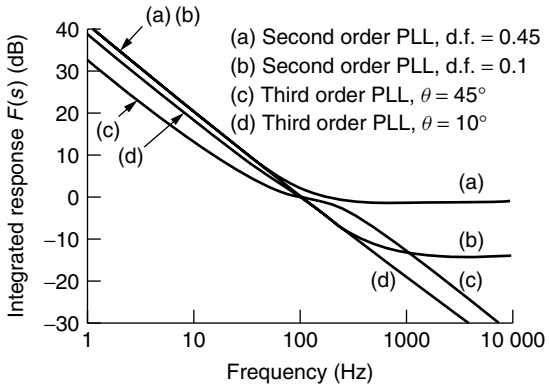


Figure 32. Integrated response for various loops as a function of the phase margin.

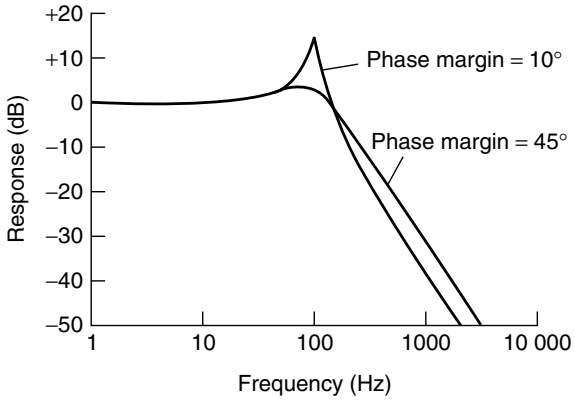


Figure 33. Closed-loop response of a type 2, third-order PLL having a phase margin of 10°.

this, the loop filter has to be adjusted after the design is finalized to accommodate the proper resulting phase margin (35°–45°). The open-loop gain for different loops can be seen in Figs. 32 and 33.

5.2. Transient Behavior of Digital Loops Using Tristate Phase Detectors

5.2.1. Pullin Characteristic. The type 2, second-order loop is used with either a sample/hold comparator or a tristate phase/frequency comparator.

We will now determine the transient behavior of this loop. Figure 34 shows the block diagram.

Very rarely in the literature is a clear distinction between pullin and lockin characteristics or frequency and phase acquisition made as a function of the digital phase/frequency detector. Somehow, all the

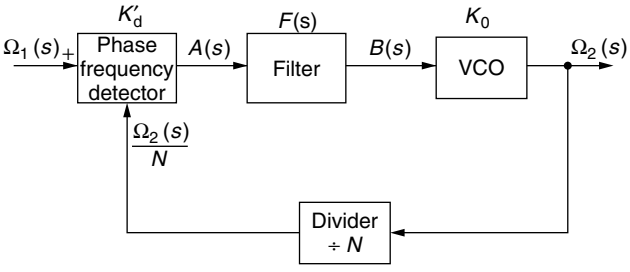


Figure 34. Block diagram of a digital PLL before lock is acquired.

approximations or linearizations refer to a sinusoidal phase/frequency comparator or its digital equivalent, the exclusive-OR gate.

The tristate phase/frequency comparator follows slightly different mathematical principles. The phase detector gain is

$$K_d' = \frac{V_d}{\omega_0} = \frac{\text{phase detector supply voltage}}{\text{loop idling frequency}}$$

and is valid only in the out-of-lock state and is a somewhat coarse approximation to the real gain which, due to nonlinear differential equations, is very difficult to calculate. However, practical tests show that this approximation is still fairly accurate.

Definitions are

$$\Omega_1(s) = \mathcal{L}[\Delta\omega_1(t)] \quad \text{(reference input to } \delta/\omega \text{ detector)}$$
$$\Omega_2(s) = \mathcal{L}[\Delta\omega_2(t)] \quad \text{(signal VCO output frequency)}$$
$$\Omega_e(s) = \mathcal{L}[\omega_e(t)] \quad \text{(error frequency at } \delta/\omega \text{ detector)}$$
$$\Omega_e(s) = \Omega_1(s) - \frac{\Omega_2(s)}{N}$$
$$\Omega_2(s) = [\Omega_1(s) - \Omega_e(s)]N$$

From the circuit described above

$$A(s) = \Omega_e(s)K_d'$$
$$B(s) = A(s)F(s)$$
$$\Omega_2(s) = B(s)K_o$$

The error frequency at the detector is

$$\Omega_e(s) = \Omega_1(s)N\frac{1}{N + K_oK_d'F(s)} \tag{52}$$

The signal is stepped in frequency:

$$\Omega_1(s) = \frac{\Delta\omega_1}{s} \quad (\Delta\omega_1 = \text{magnitude of frequency step}) \tag{53}$$

5.2.1.1. Active Filter of First Order. If we use an active filter

$$F(s) = \frac{1 + s\tau_2}{s\tau_1} \tag{54}$$

and insert this in (51), the error frequency is

$$\Omega_e(s) = \Delta\omega_1 N\frac{1}{s\left(N + K_oK_d'\dfrac{\tau_2}{\tau_1}\right) + \dfrac{K_oK_d'}{\tau_1}} \tag{55}$$

Utilizing the Laplace transformation, we obtain

$$\omega_e(t) = \Delta\omega_1\frac{1}{1 + K_oK_d'(\tau_2/\tau_1)(1/N)}\exp\left[-\frac{t}{(\tau_1N/K_oK_d') + \tau_2}\right] \tag{56}$$

and

$$\lim_{t\to 0}\omega_e(t) = \frac{\Delta\omega_1 N}{N + K_oK_d'(\tau_2/\tau_1)} \tag{57}$$
$$\lim_{t\to\infty}\omega_e(t) = 0 \tag{58}$$

5.2.1.2. Passive Filter of First Order. If we use a passive filter

$$\lim_{t\to\infty}\omega_e(t) = 0 \tag{59}$$

for the frequency step

$$\Omega_1(s) = \frac{\Delta\omega_1}{s} \tag{60}$$

the error frequency at the input becomes

$$\Omega_e(s) = \Delta\omega_1 N\left\{\frac{1}{s}\frac{1}{s[N(\tau_1+\tau_2) + K_oK_d'\tau_2] + (N + K_oK_d')} + \frac{\tau_1+\tau_2}{s[N(\tau_1+\tau_2) + K_oK_d'\tau_2] + (N + K_oK_d')}\right\} \tag{61}$$

For the first term we will use the abbreviation A, and for the second term we will use the abbreviation B:

$$A = \frac{1/[N(\tau_1+\tau_2) + K_oK_d'\tau_2]}{s\left[s + \dfrac{N + K_oK_d'}{N(\tau_1+\tau_2) + K_oK_d'\tau_2}\right]} \tag{62}$$

$$B = \frac{\dfrac{\tau_1+\tau_2}{N(\tau_1+\tau_2) + K_oK_d'\tau_2}}{s + \dfrac{N + K_oK_d'}{N(\tau_1+\tau_2) + K_oK_d'\tau_2}} \tag{63}$$

After the inverse Laplace transformation, our final result becomes

$$\mathcal{L}^{-1}(A) = \frac{1}{N + K_oK_d'} \times\left\{1 - \exp\left[-t\frac{N + K_oK_d'}{N(\tau_1+\tau_2) + K_oK_d'\tau_2}\right]\right\} \tag{64}$$

$$\mathcal{L}^{-1}(B) = \frac{\tau_1+\tau_2}{N(\tau_1+\tau_2) + K_oK_d'\tau_2} \times\exp\left(-t\frac{N + K_oK_d'}{N(\tau_1+\tau_2) + K_oK_d'\tau_2}\right) \tag{65}$$

and finally

$$\omega_e(t) = \Delta\omega_1 N[\mathcal{L}^{-1}(A) + (\tau_1+\tau_2)\mathcal{L}^{-1}(B)] \tag{66}$$

What does the equation mean? We really want to know how long it takes to pull the VCO frequency to the reference. Therefore, we want to know the value of t, the time it takes to be within 2π or less of lockin range.

The PLL can, at the beginning, have a phase error from -2π to $+2\pi$, and the loop, by accomplishing lock, then takes care of this phase error. We can make the reverse assumption for a moment and ask ourselves, as we have done earlier, how long the loop stays in phase lock. This is called the *pullout range*. Again, we apply signals to the

input of the PLL as long as the loop can follow and the phase error does not become larger than 2π. Once the error is larger than 2π, the loop jumps out of lock. When the loop is out of lock, a beat note occurs at the output of the loop filter following the phase/frequency detector.

The tristate state phase/frequency comparator, however, works on a different principle, and the pulses generated and supplied to the charge pump do not allow the generation of an AC voltage. The output of such a phase/frequency detector is always unipolar, but relative to the value of $V_{\text{batt}}/2$, the integrator voltage can be either positive or negative. If we assume for a moment that this voltage should be the final voltage under a locked condition, we will observe that the resulting DC voltage is either more negative or more positive relative to this value, and because of this, the VCO will be "pulled in" to this final frequency rather than swept in. The swept-in technique applies only in cases of phase/frequency comparators, where this beat note is being generated. A typical case would be the exclusive-OR gate or even a sample/hold comparator. This phenomenon is rarely covered in the literature and is probably discussed in detail for the first time in the book by Roland Best [9].

Let us assume now that the VCO has been pulled in to final frequency to be within 2π of the final frequency, and the time t is known. The next step is to determine the lockin characteristic.

5.2.2. Lockin Characteristic. We will now determine the lockin characteristic, and this requires the use of a different block diagram. Figure 5 shows the familiar block diagram of the PLL, and we will use the following definitions:

$$\theta_1(s) = \mathcal{L}[\Delta\delta_1(t)] \quad \text{(reference input to } \delta/\omega \text{ detector)}$$
$$\theta_2(s) = \mathcal{L}[\Delta\delta_2(t)] \quad \text{(signal VCO output phase)}$$
$$\theta_e(s) = \mathcal{L}[\delta_e(t)] \quad \text{(phase error at } \delta/\omega \text{ detector)}$$
$$\theta_e(s) = \theta_1(s) - \frac{\theta_2(s)}{N}$$

From the block diagram, the following is apparent:

$$A(s) = \theta_e(s)K_d$$
$$B(s) = A(s)F(s)$$
$$\theta_2(s) = B(s)\frac{K_o}{s}$$

The phase error at the detector is

$$\theta_e(s) = \theta_1(s)\frac{sN}{K_oK_dF(s) + sN} \tag{67}$$

A step in phase at the input, where the worst-case error is 2π, results in

$$\theta_1(s) = 2\pi\frac{1}{s} \tag{68}$$

We will now treat the two cases using an active or passive filter.

5.2.2.1. Active Filter. The transfer characteristic of the active filter is

$$F(s) = \frac{1 + s\tau_2}{s\tau_1} \tag{69}$$

This results in the formula for the phase error at the detector:

$$\theta_e(s) = 2\pi\frac{s}{s^2 + (sK_oK_d\tau_2/\tau_1)/N + (K_oK_d/\tau_1)/N} \tag{70}$$

The polynomial coefficients for the denominator are

$$a_2 = 1$$
$$a_1 = \frac{K_oK_d\tau_2/\tau_1}{N}$$
$$a_0 = \frac{K_oK_d/\tau_1}{N}$$

and we have to find the roots W_1 and W_2. Expressed in the form of a polynomial coefficient, the phase error is

$$\theta_e(s) = 2\pi\frac{s}{(s + W_1)(s + W_2)} \tag{71}$$

After the inverse Laplace transformation has been performed, the result can be written in the form

$$\delta_e(t) = 2\pi\frac{W_1e^{-W_1t} - W_2e^{-W_2t}}{W_1 - W_2} \tag{72}$$

with

$$\lim_{t\to 0}\delta_e(t) = 2\pi$$

and

$$\lim_{t\to\infty}\delta_e(t) = 0$$

The same can be done using a passive filter.

5.2.2.2. Passive Filter. The transfer function of the passive filter is

$$F(s) = \frac{1 + s\tau_2}{1 + s(\tau_1 + \tau_2)} \tag{73}$$

If we apply the same phase step of 2π as before, the resulting phase error is

$$\theta_e(s) = 2\pi\frac{[1/(\tau_1 + \tau_2)] + s}{s^2 + s\dfrac{N + K_oK_d\tau_2}{N(\tau_1 + \tau_2)} + \dfrac{K_oK_d}{N(\tau_1 + \tau_2)}} \tag{74}$$

Again, we have to find the polynomial coefficients, which are

$$a_2 = 1$$
$$a_1 = \frac{N + K_oK_d\tau_2}{N(\tau_1 + \tau_2)}$$
$$a_0 = \frac{K_oK_d}{N(\tau_1 + \tau_2)}$$

and finally find the roots for W_1 and W_2. This can be written in the form

$$\theta_e(s) = 2\pi \left[\frac{1}{\tau_1 + \tau_2} \frac{1}{(s + W_1)(s + W_2)} + \frac{s}{(s + W_1)(s + W_2)} \right] \tag{75}$$

Now we perform the inverse Laplace transformation and obtain our result:

$$\delta_e(t) = 2\pi \left(\frac{1}{\tau_1 + \tau_2} \frac{e^{-W_1 t} - e^{-W_2 t}}{W_2 - W_1} + \frac{W_1 e^{-W_1 t} - W_2 e^{-W_2 t}}{W_1 - W_2} \right) \tag{76}$$

with

$$\lim_{t \to 0} \delta_e(t) = 2\pi$$

with

$$\lim_{t \to \infty} \delta_e(t) = 0$$

When analyzing the frequency response for the various types and orders of PLLs, the phase margin played an important role. For the transient time, the type 2, second-order loop can be represented with a damping factor or, for higher orders, with the phase margin. Figure 35 shows the normalized output response for a damping factor of 0.1 and 0.47. The ideal Butterworth response would be a damping factor of 0.7, which correlates with a phase margin of 45°.

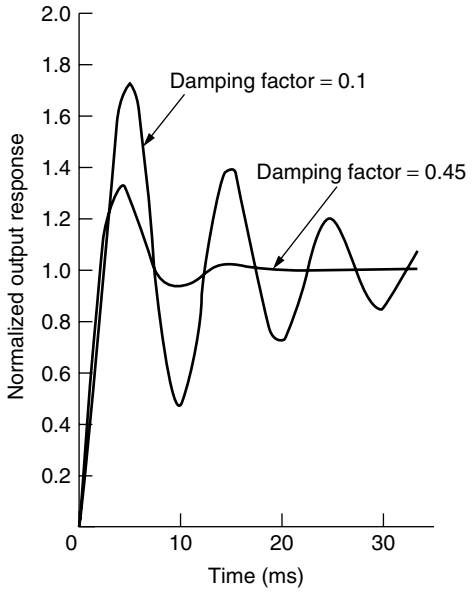


Figure 35. Normalized output response of a type 2, second-order loop with a damping factor of 0.1 and 0.05 for $\Omega_n = 0.631$.

5.3. Loop Gain/Transient Response Examples

Given the simple filter shown in Fig. 36 and the parameters as listed, the Bode plot is shown in Fig. 37. This approach can also be translated from a type 1 into a type 2 filter as shown in Fig. 38 and its frequency response is shown in Fig. 39. The lockin function for this type 2, second-order loop with an ideal damping factor of 0.707 (Butterworth response) is shown in Fig. 40. Figure 41 shows an actual settling time measurement. Any deviation from ideal damping, as we'll soon see, results in ringing (in an underdamped system) or, in an overdamped system, the voltage will crawl to its final value. This system can be increased in its order by selecting a type 2, third-order loop using the filter shown in Fig. 42. For an ideal synthesis of the values, the Bode diagram looks as shown in Fig. 43, and its resulting response is given in Fig. 44.

The order can be increased by adding an additional lowpass filter after the standard loop filter. The resulting system is called a *type 2, fifth-order loop*. Figure 45 shows the Bode diagram or open-loop diagram, and Fig. 46 shows the locking function. By using a very

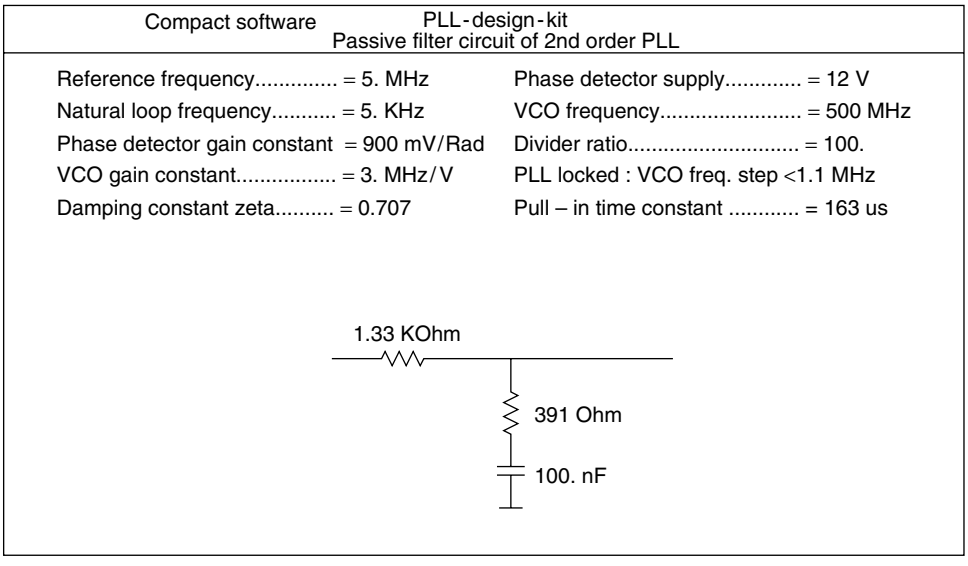


Figure 36. Loop filter for a type 1, second-order synthesizer.

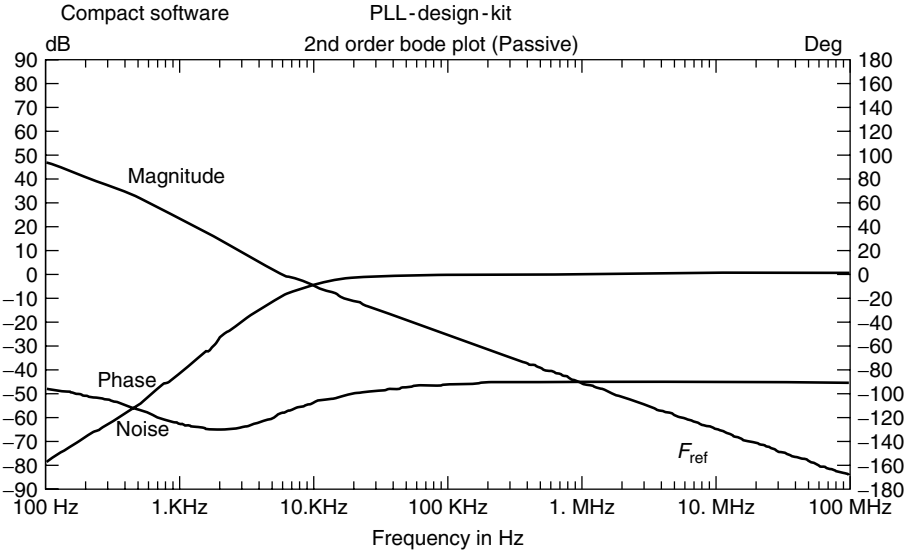


Figure 37. Type 1, second-order loop response.

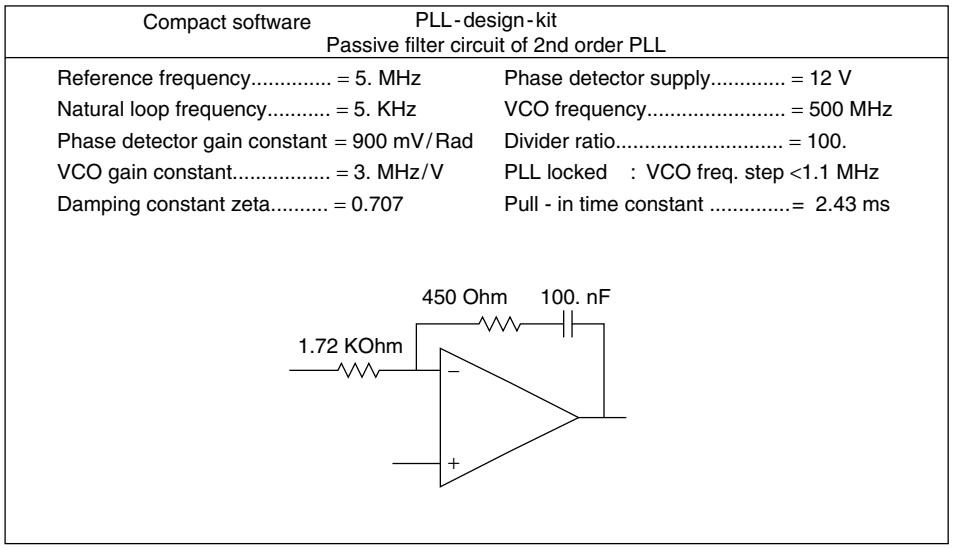


Figure 38. Loop filter for a type 2, second-order synthesizer.

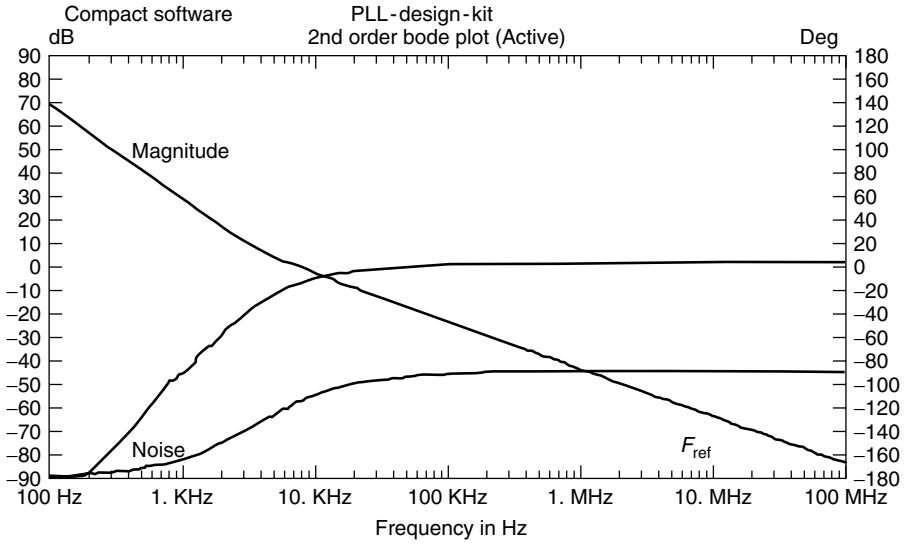


Figure 39. Response of the type 2, second-order loop.

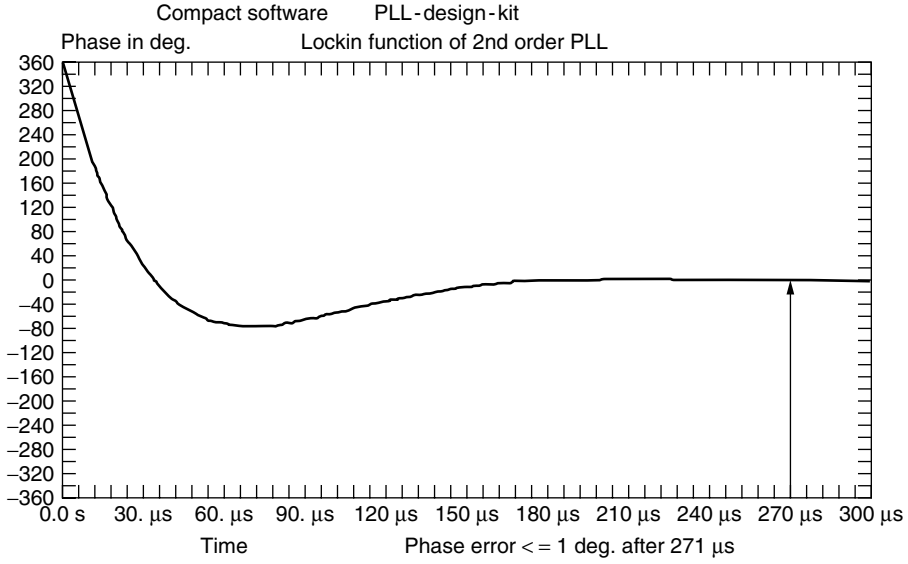


Figure 40. Lock-in function of the type 2, second-order PLL, indicating a lock time of 271 µs and an ideal response.

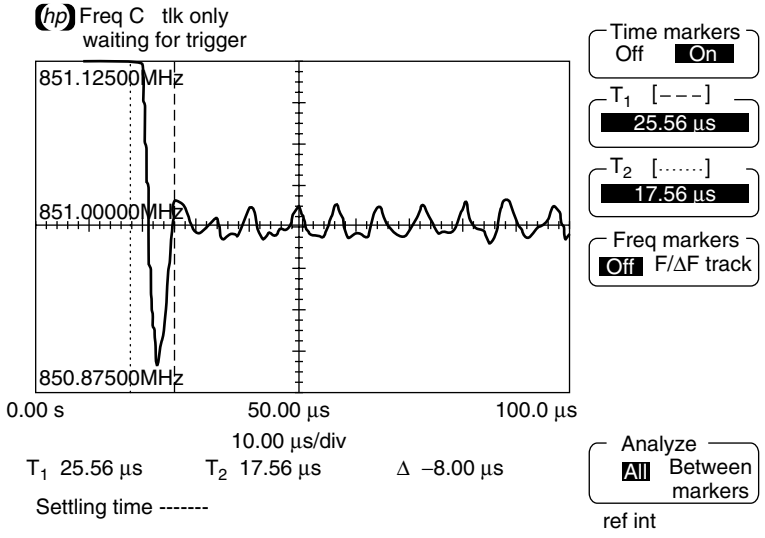


Figure 41. Example of settling time measurement.

wide loop bandwidth, this can be used to clean up microwave oscillators with inherent comparatively poor phase noise. This cleanup, has a dramatic influence on the performance.

By deviating from the ideal 45° to a phase margin of 33°, one obtains the above mentioned ringing, as is evident from Fig. 47. The time to settle has increased from 13.3 to 62 µs.

To more fully illustrate the effects of nonideal phase margin, Figs. 48, 49, 50, and 51 show the lockin function of a different type 2, fifth-order loop configured for phase margins of 25°, 35°, 45°, and 55°, respectively.

I have already mentioned that the loop should avoid "ears" (Fig. 31) with poorly designed loop filters. Another interesting phenomenon is the tradeoff between loop bandwidth and phase noise. In Fig. 52 the loop bandwidth has been made too wide, resulting in a degradation of the phase noise but provides faster settling time. By reducing the loop bandwidth from about 1 kHz to 300 Hz, only a very slight overshoot remains, improving the phase noise significantly. This is shown in Fig. 53.

5.4. Practical Circuits

Figure 54 shows a passive filter used for a synthesizer chip with constant current output. This chip has a charge pump output, which explains the need for the first capacitor.

Figure 55 shows an active integrator operating at a reference frequency of several megahertz. The notch filter at the output reduces the reference frequency considerably. The notch is about 4.5 MHz.

Figure 56 shows the combination of a phase/frequency discriminator and a higher-order loop filter as used in more complicated systems, such as fractional-division synthesizers.

Figure 57 shows a custom-built phase detector with a noise floor of better than — 168 dBc/Hz.

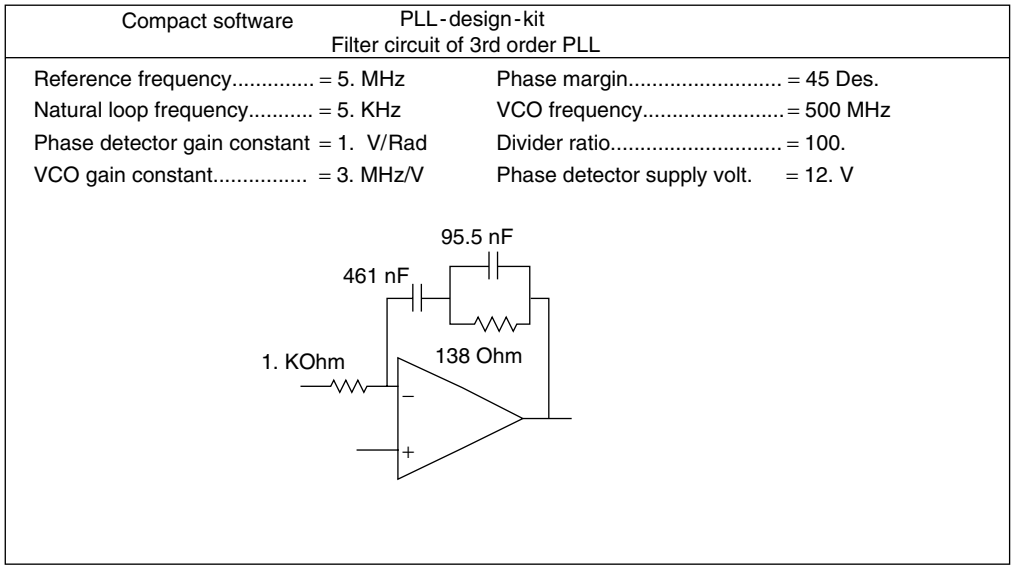


Figure 42. Loop filter for a type 2, third-order synthesizer.

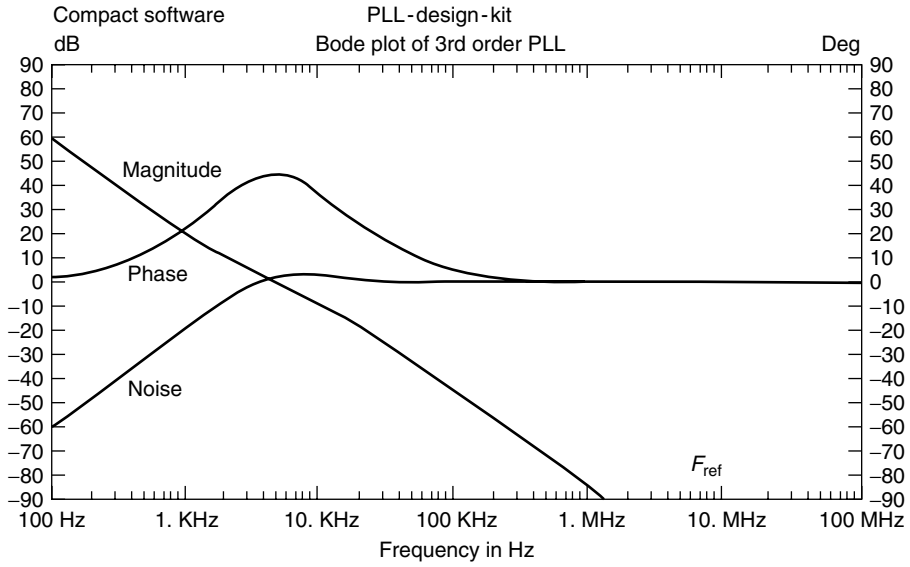


Figure 43. Open-loop Bode diagram for the type 2, third-order loop. It fulfills the requirement of 45° phase margin at the 0-dB crossover point, and corrects the slope down to −10 dB gain.

6. THE FRACTIONAL-*N* PRINCIPLE

The principle of the fractional-N PLL synthesizer was briefly mentioned in Section 2. The following is a numerical example for better understanding.

Example 1. Considering the problem of generating 899.8 MHz using a fractional-N loop with a 50-MHz reference frequency, we obtain

$$899.8\ \text{MHz} = 50\ \text{MHz}\left(N + \frac{K}{F}\right)$$

The integral part of the division N has to be set to 17 and the fractional part K/F needs to be $\frac{996}{1000}$; (the fractional part $\dfrac{K}{F}$ is not a integer) and the VCO output has to be divided by 996× every 1000 cycles. This can easily be implemented by adding the number 0.996 to the contents of an accumulator every cycle. Every time the accumulator overflows, the divider divides by 18 rather than by 17. Only the fractional value of the addition is retained in the phase accumulator. If we move to the lower band or try to generate 850.2 MHz, N remains 17, and K/F becomes $\frac{4}{1000}$. This method of using fractional division was first introduced by using analog implementation and noise cancellation, but today it is implemented totally as a digital approach. The necessary resolution is obtained from the dual-modulus prescaling, which allows for a well-established method for achieving a high-performance frequency synthesizer operating at UHF and higher frequencies. Dual-modulus prescaling avoids the loss of resolution in a system compared to a simple prescaler; it allows a VCO step

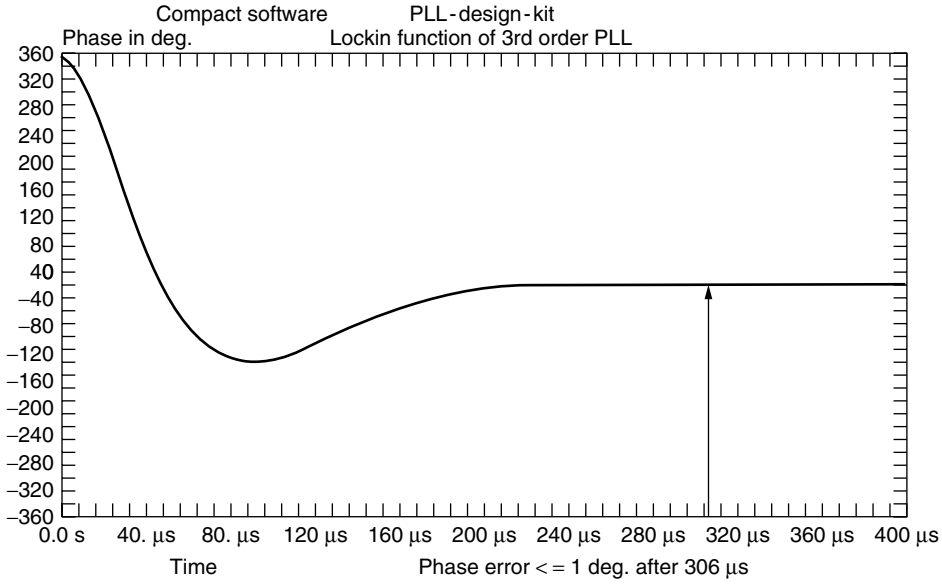


Figure 44. Lockin function of the type 2, third-order loop for an ideal 45° phase margin.

equal to the value of the reference frequency to be obtained. This method needs an additional counter and the dual modulus prescaler then divides one or two values depending on the state of its control. The only drawback of prescalers is the minimum division ratio of the prescaler for approximately N^2. The dual-modulus divider is the key to implementing the fractional-N synthesizer principle. Although the fractional-N technique appears to have a good potential of solving the resolution limitation, it is not free of having its own complications. Typically, an overflow from the phase accumulator, which is the adder with the output feedback to the input after being latched, is used to change the instantaneous division ratio. Each overflow produces a jitter at the output frequency, caused by the fractional division, and is limited to the fractional portion of the desired division ratio.

In our case, we had chosen a step size of 200 kHz, and yet the discrete sidebands vary from 200 kHz for $K/F = \frac{4}{1000}$ to 49.8 MHz for $K/F = \frac{996}{1000}$. It will become the task of the loop filter to remove those discrete spurious components. While in the past the removal of the discrete spurs has been accomplished by using analog techniques, various digital methods are now available. The microprocessor has to solve the following equation:

$$N* = \left(N + \frac{K}{F}\right) = [N(F - K) + (N + 1)K]$$

Example 2. For $F_O = 850.2$ MHz, we obtain

$$N* = \frac{850.2 \text{ MHz}}{50 \text{ MHz}} = 17.004$$

Following the formula above we have

$$N* = \left(N + \frac{K}{F}\right) = \frac{[17(1000 - 4) + (17 + 1) \times 4]}{1000}$$
$$= \frac{[16932 + 72]}{1000} = 17.004$$

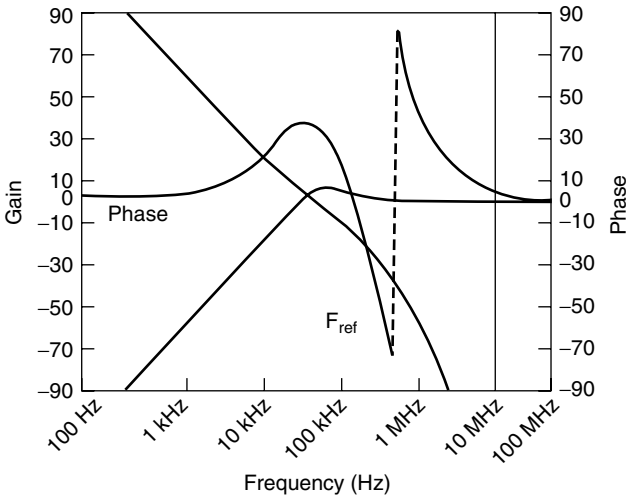


Figure 45. Bode plot of the fifth-order PLL system for a microwave synthesizer. The theoretical reference suppression is better than 90 dB.

$$F_{\text{out}} = 50 \text{ MHz} \times \frac{[16932 + 72]}{1000}$$
$$= 846.6 \text{ MHz} + 3.6 \text{ MHz}$$
$$= 850.2 \text{ MHz}$$

By increasing the number of accumulators, frequency resolution much below 1-Hz step size is possible with the same switching speed.

There is an interesting, generic problem associated with *all* fractional-N synthesizers. Assume for a moment that we use our 50-MHz reference and generate a 550-MHz output frequency. This means that our division factor is 11. Aside from reference frequency sidebands (±50 MHz) and harmonics, there will be no unwanted spurious frequencies. Of course, the reference sidebands

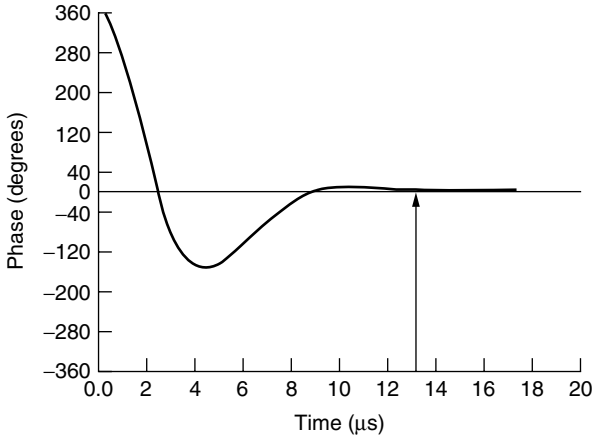


Figure 46. Lockin function of the fifth-order PLL. Note that the phase lock time is approximately 13.3 μs.

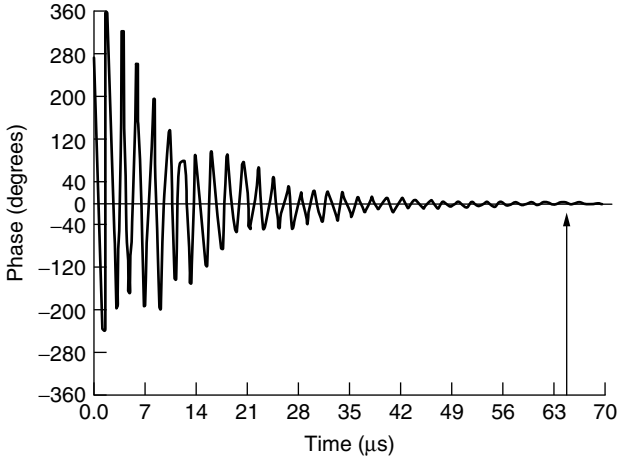


Figure 47. Lockin function of the fifth-order PLL. Note that the phase margin has been reduced from the ideal 45°. This results in a much longer settling time of 62 μs.

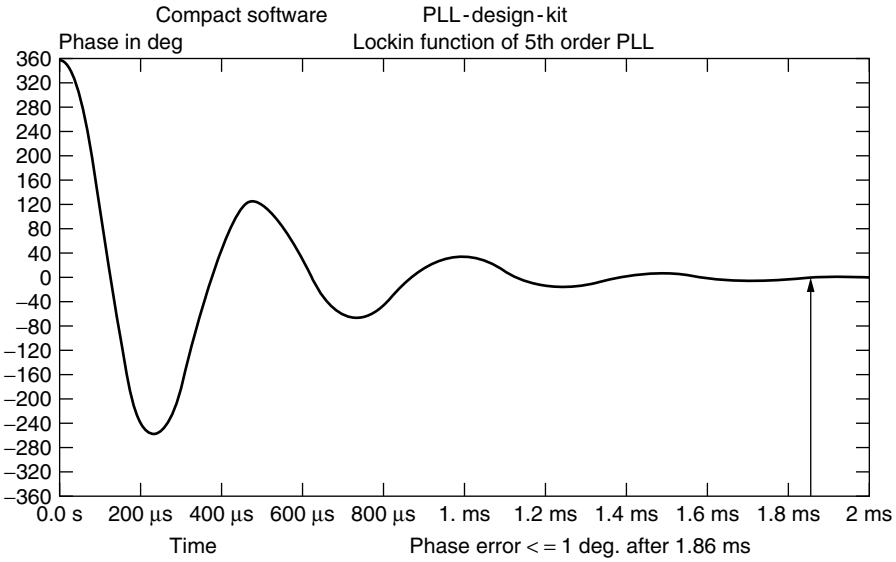


Figure 48. Lockin function of another type 2, fifth-order loop with a 25° phase margin. Noticeable ringing occurs, lengthening the lockin time to 1.86 ms.

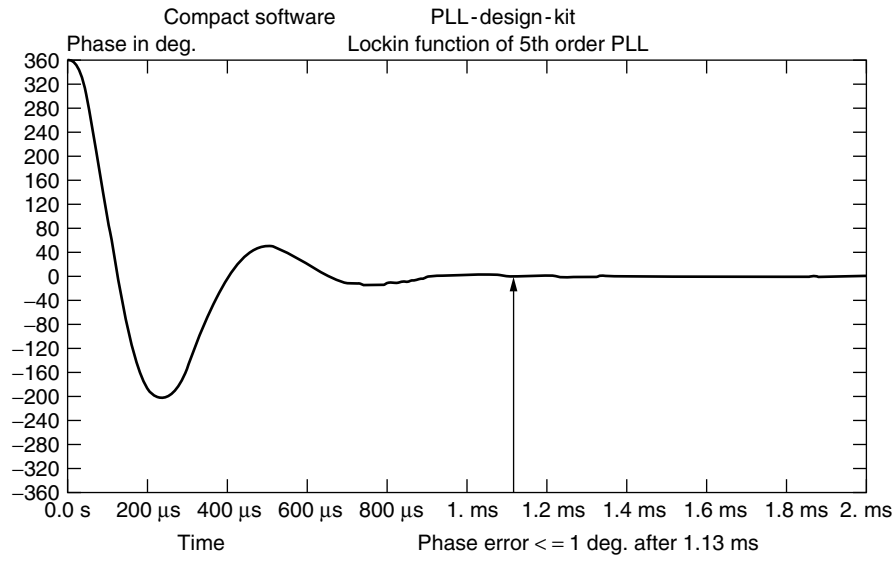


Figure 49. Lockin function of the type 2, fifth-order loop with a 35° phase margin. Ringing still occurs, but the lockin time has decreased to 1.13 ms.

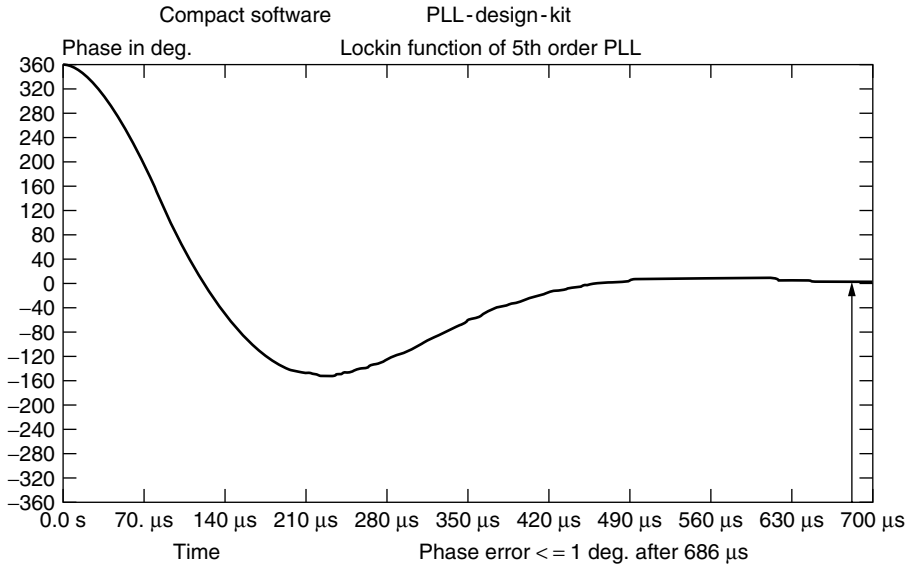


Figure 50. Lockin function of the type 2, third-order loop with an ideal 45° phase margin. The lockin time is 686 µs.

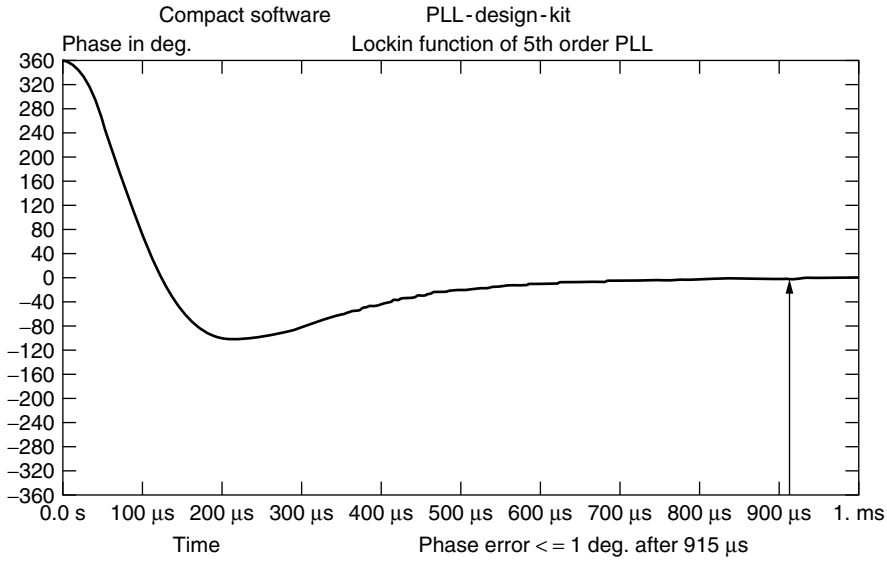


Figure 51. Lockin function of the type 2, fifth-order loop, for a 55° phase margin. The lockin time has increased to 915 µs.

will be suppressed by the loop filter by more than 90 dB. For reasons of phase noise and switching speed, a loop bandwidth of 100 kHz has been considered. Now, taking advantage of the fractional-*N* principle, say that we want to operate at an offset of 30 kHz (550.03 MHz). With this new output frequency, the inherent spurious signal reduction mechanism in the fractional-*N* chip limits the reduction to about 55 dB. Part of the reason why the spurious signal suppression is less in this case is that the phase frequency detector acts as a mixer, collecting both the 50-MHz reference (and its harmonics) and 550.03 MHz. Mixing the 11th reference harmonic (550 MHz) and the output frequency (550.03 MHz) results in output at 30 kHz; since the loop bandwidth is 100 kHz, it adds nothing to the suppression of this signal. To solve this, we could consider narrowing the loop bandwidth to 10% of the offset. A 30-kHz offset would equate to a loop bandwidth of 3 kHz, at which the loop speed might still be acceptable, but for a 1-kHz offset, the necessary loop bandwidth of 100 Hz would make the loop too slow. A better way is to use a different reference frequency — one that would place the resulting spurious product considerably outside the 100-kHz loop filter window. If, for instance, we used a 49-MHz reference, multiplication by 11 would result in 539 MHz. Mixing this with 550.03 MHz would result in spurious signals at ±11.03 MHz, a frequency so far outside the loop bandwidth that it would essentially disappear. Starting with a VHF, low-phase-noise crystal oscillator, such as 130 MHz, one can implement an intelligent reference frequency selection to avoid these discrete spurious signals. An additional method of reducing the spurious contents is maintaining a division ratio greater than 12 in all cases. Actual tests have shown that these reference-based spurious frequencies can be repeatedly suppressed by 80 to 90 dB.

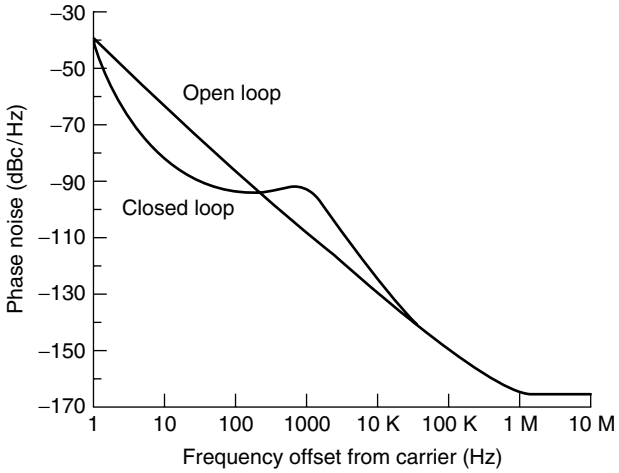


Figure 52. Comparison between open- and closed-loop noise prediction. Note the overshoot of around 1 kHz off the carrier.

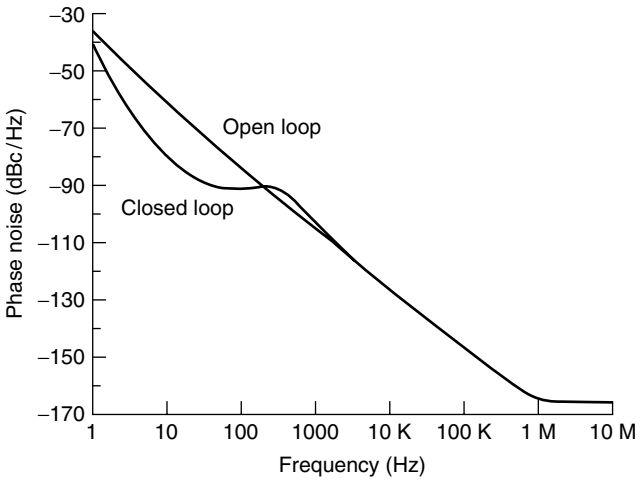


Figure 53. Comparison between open- and closed-loop noise prediction. Note the overshoot at around 300 Hz off the carrier.

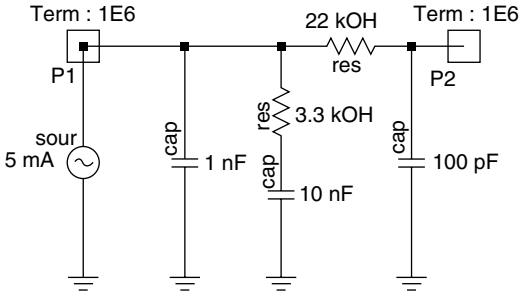


Figure 54. Type 1 high-order loop filter used for passive filter evaluation. The 1-nF capacitor is used for spike suppression as explained in the text. The filter consists of a lag portion and an additional lowpass section.

6.1. Spur Suppression Techniques

While several methods have been proposed in the literature, the method of reducing the noise by using a sigma–delta modulator has shown to be most promising. The concept is to get rid of the low-frequency phase error by rapidly switching the division ratio to eliminate the gradual phase error at the discriminator input. By changing the division ratio rapidly between different values, the phase errors occur in both polarities, positive as well as negative, and at an accelerated rate that explains the phenomenon of high-frequency noise pushup. This noise, which is converted to a voltage by the phase/frequency discriminator and loop filter, is filtered out by the lowpass filter. The main problem associated with this noise shaping technique is that the noise power rises rapidly with frequency. Figure 58 shows noise contributions with such a sigma–delta modulator in place.

On the other hand, we can now, for the first time, build a single-loop synthesizer with switching times as fast as 6 μs and very little phase noise deterioration inside the loop bandwidth, as seen in Fig. 58. Since this system maintains the good phase noise of the ceramic-resonator-based oscillator, the resulting performance is significantly better than the phase noise expected from high-end signal

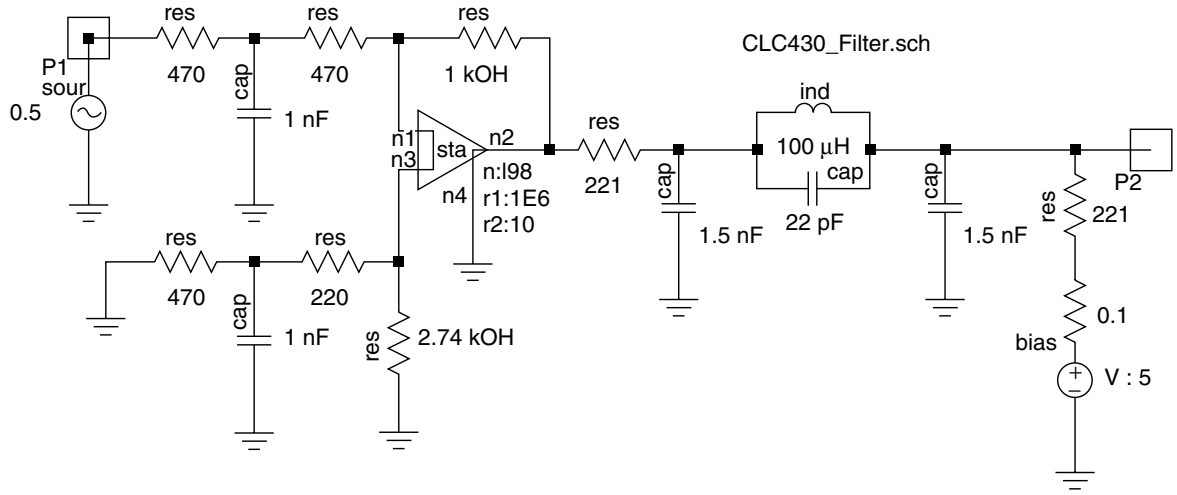


Figure 55. A type 2 high-order filter with a notch to suppress the discrete reference spurs.

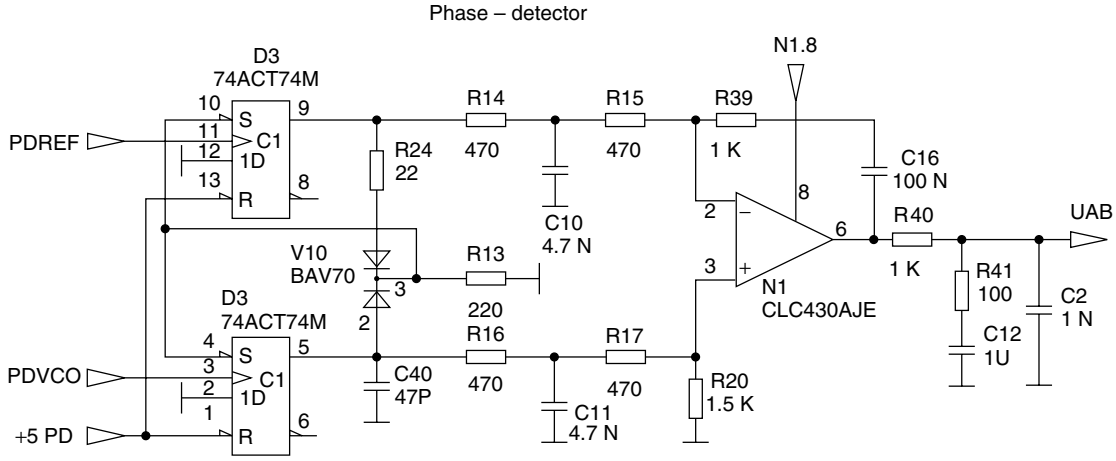


Figure 56. Phase/frequency discriminator including an active loop filter capable of operating up to 100 MHz.

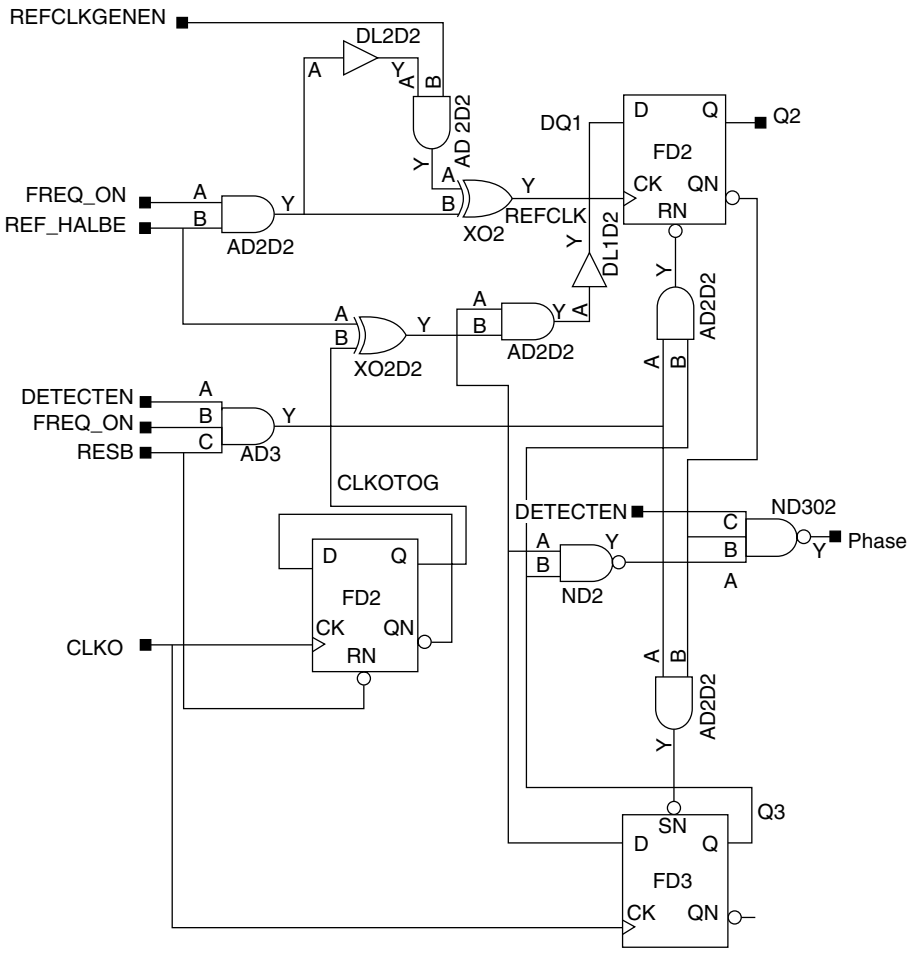


Figure 57. Custom-built phase detector with a noise floor of better than −168 dBc/Hz. This phase detector shows extremely low phase jitter.

generators. However, this method does not allow us to increase the loop bandwidth beyond the 100-kHz limit, where the noise contribution of the sigma–delta modulator takes over.

Table 2 shows some of the modern spur suppression methods. These three-stage sigma–delta methods with larger accumulators have the most potential.

The power spectral response of the phase noise for the three-stage sigma–delta modulator is calculated from

$$L(f) = \frac{(2\pi)^2}{12 \times f_{\text{ref}}} \times \left[2 \sin\left(\frac{\pi f}{f_{\text{ref}}}\right)\right]^{2(n-1)} \text{ rad}^2/\text{Hz} \qquad (78)$$

where n is the number of the stage of the cascaded sigma–delta modulator. Equation (78) shows that the

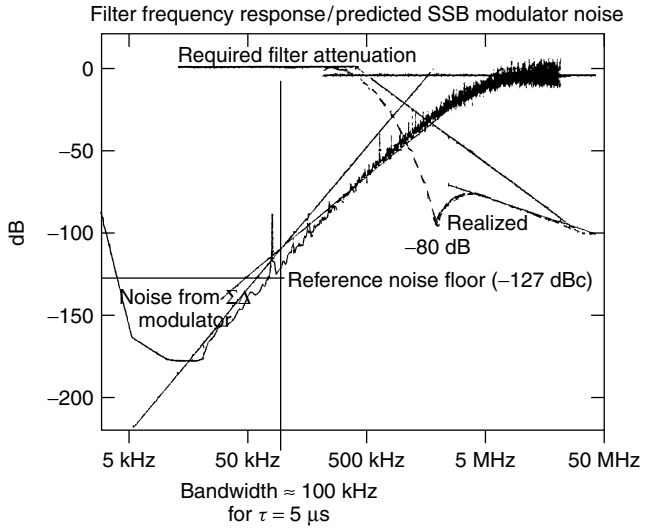


Figure 58. This filter frequency response/phase noise analysis graph shows the required attenuation for the reference frequency of 50 MHz and the noise generated by the sigma–delta converter (three steps) as a function of the offset frequency. It becomes apparent that the sigma–delta converter noise dominates above 80 kHz unless attenuated.

phase noise resulting from the fractional controller is attenuated to negligible levels close to the center frequency, and further from the center frequency, the phase noise is increased rapidly and must be filtered out prior to the tuning input of the VCO to prevent unacceptable degradation of spectral purity. A loop filter must be used to filter the noise in the PLL loop. Figure 58 shows the plot of the phase noise versus the offset frequency from the center frequency. A fractional-N synthesizer with a three-stage sigma–delta modulator as shown in Fig. 59 has been built. The synthesizer consists of a phase/frequency detector, an active lowpass filter (LPF), a voltage-controlled oscillator (VCO), a dual-modulus prescaler, a three-stage sigma–delta modulator, and a buffer. Figure 60 shows the inner workings of the chip in greater detail.

Table 2. Spur Suppression Methods

Technique	Feature	Problem
DAC phase estimation	Cancel spur by DAC	Analog mismatch
Pulse generation	Insert pulses	Interpolation jitter
Phase interpolation	Inherent fractional divider	Interpolation jitter
Random jittering	Randomize divider	Frequency jitter
Sigma–delta modulation	Modulate division ratio	Quantization noise

After designing, building, and predicting the phase noise performance of this synthesizer, it becomes clear that measuring the phase noise of such a system becomes tricky. Standard measurement techniques that use a reference synthesizer will not provide enough resolution because there are no synthesized signal generators on the market sufficiently good to measure such low values of

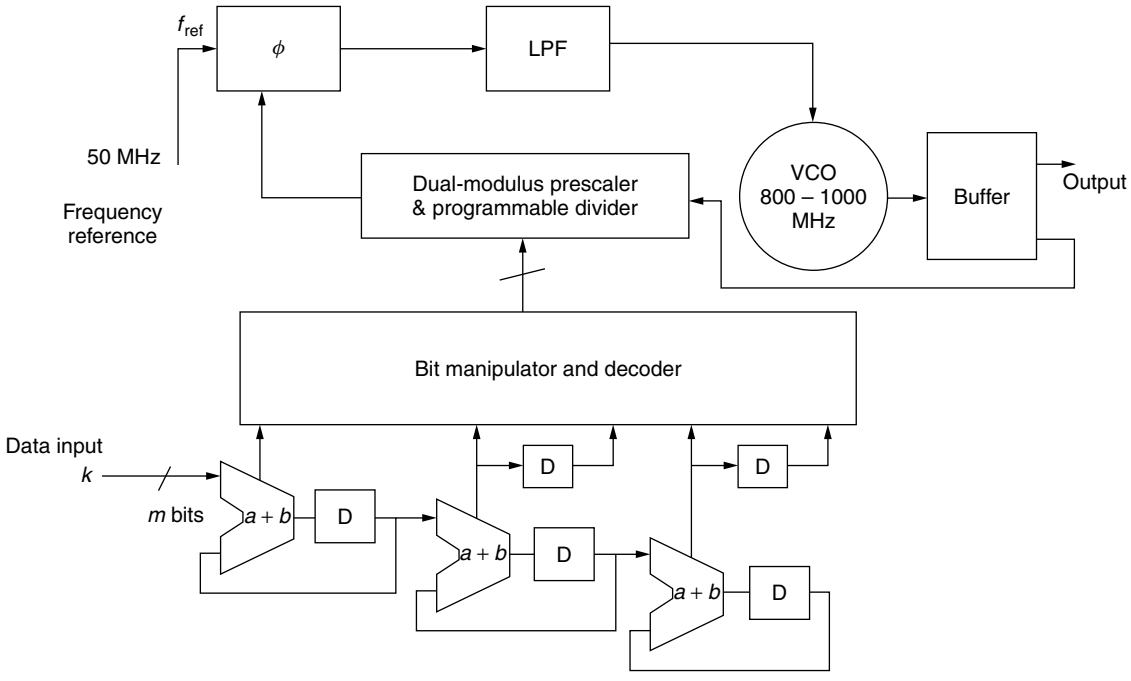


Figure 59. Block diagram of the fractional-N synthesizer built using a custom IC capable of operation at reference frequencies up to 150 MHz. The frequency is extensible up to 3 GHz using binary (÷2, ÷4, ÷8, etc.) and fixed-division counters.

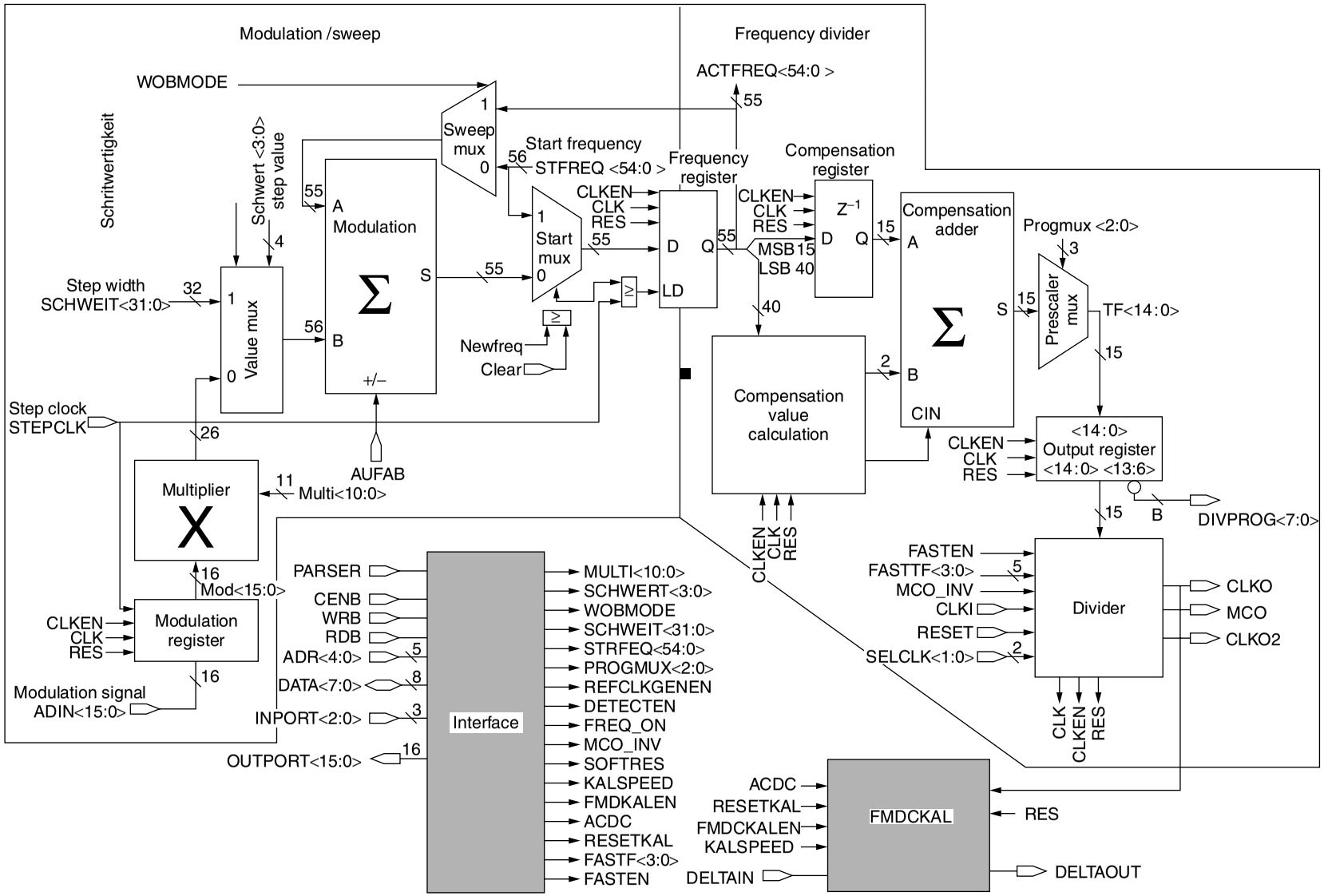


Figure 60. Detailed block diagram of the inner workings of the fractional-*N*-division synthesizer chip.

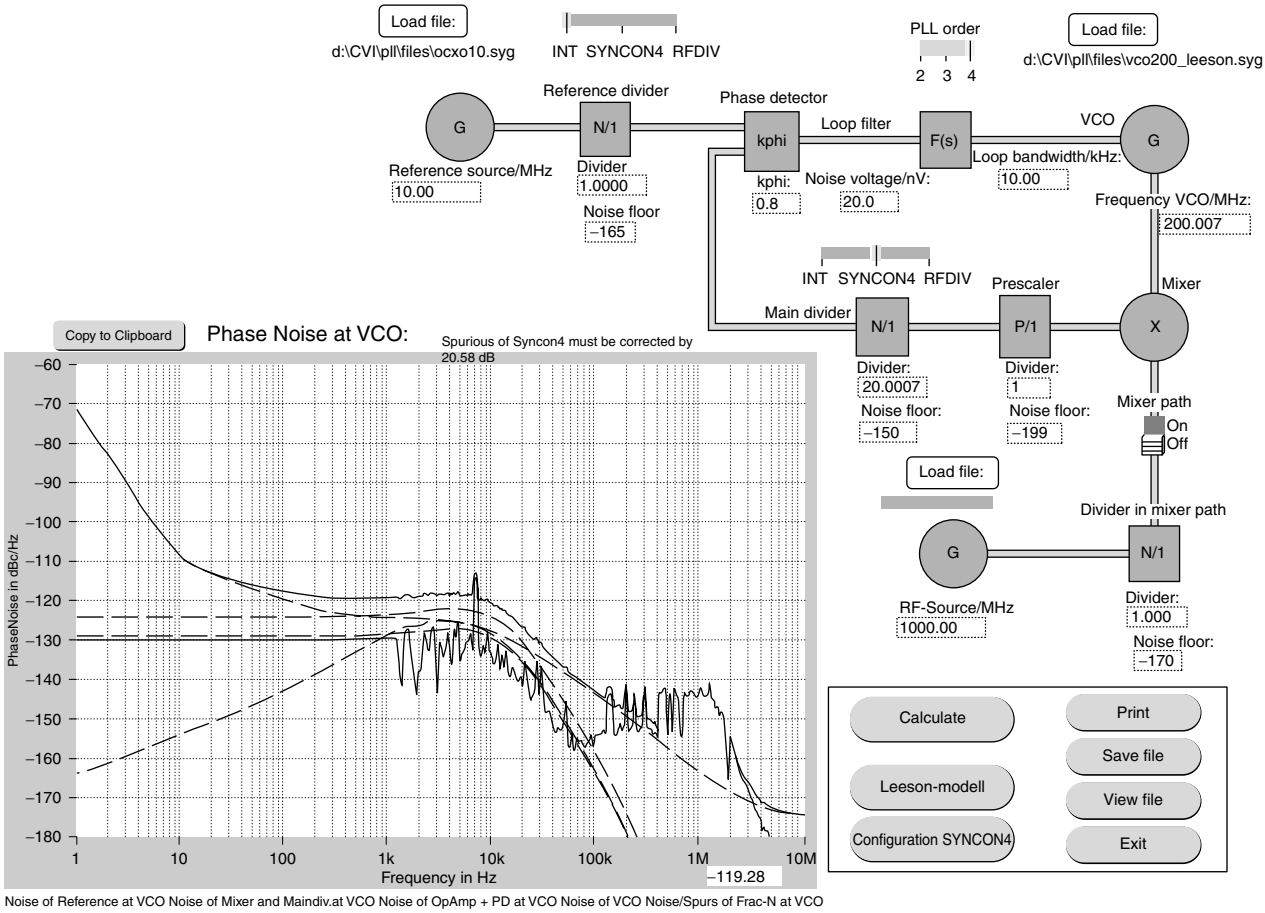


Figure 61. Composite phase noise of the fractional-N synthesizer system, including all noise and spurious signals generated within the system. The discrete spurious of 7 kHz is due to the nonlinearity of the phase detector. Its value needs to be corrected by 20.58 dB to a lesser value because of the bandwidth of the FFT analyzer.

phase noise. Therefore, a comb generator is needed that would take the output of the oscillator and multiply this up 10–20 times.

Figure 61 shows a simulated phase noise and termination of spurious outputs for the fraction-N-division synthesizer with $\Sigma\Delta$ converter. At the moment, it is unclear if the PLL with DDS or the fractional-N-division synthesizer principle with $\Sigma\Delta$ converter is the winning approach. The DDS typically requires two or three loops and is much more expensive, while the fractional-N approach requires only one loop and is a very intelligent spurious removal circuit, and high-end solutions are typically custom-built [96–111].

BIBLIOGRAPHY

1. U. L. Rohde and J. Whitaker, *Communications Receivers*, 3rd ed., McGraw-Hill, New York, Dec. 2000.
2. L. E. Larson, *RF and Microwave Circuit Design for Wireless Communications*, Artech House, Norwood, MA, 1996.
3. U. L. Rohde, *Microwave and Wireless Synthesizers: Theory and Design*, Wiley, New York, Aug. 1997.
4. W. F. Egan, *Frequency Synthesis by Phase Lock*, Wiley-Interscience, New York, 1981.
5. F. Gardner, *Phaselock Techniques*, 2nd ed., Wiley-Interscience, New York, 1979.
6. J. Gorski-Popiel, *Frequency Synthesis: Techniques and Applications*, IEEE, New York, 1975.
7. V. F. Kroupa, *Frequency Synthesis: Theory, Design, and Applications*, Wiley, New York, 1973.
8. V. Manassewitsch, *Frequency Synthesizers: Theory and Design*, Wiley, New York, 1976.
9. R. E. Best, *Phase-Locked Loops: Theory, Design, and Applications*, McGraw-Hill, New York, 1989.
10. P. Danzer, ed., *The ARRL Handbook for Radio Amateurs*, 75th ed., ARRL, Newington, 1997, Chap. 14, AC/RF sources (Oscillators and Synthesizers).
11. C. R. Chang and U. L. Rohde, The accurate simulation of oscillator and PLL phase noise in RF sources, *Wireless '94 Symp.*, Santa Clara, CA Feb. 15–18, 1994.
12. M. M. Driscoll, Low noise oscillator design using acoustic and other high Q resonators, *44th Annual Symp. Frequency Control*, Baltimore, MD, May 1990.
13. U. L. Rohde, Low-noise frequency synthesizers fractional N phase-locked loops, *Proc. SOUTHCON/81*, Jan. 1981.
14. W. C. Lindsey and C. M. Chie, eds., *Phase-Locked Loops*, IEEE Press, New York, 1986.

15. U.S. Patent 4,492,936 (Jan. 8, 1985), A. Albarello, A. Roullet, and A. Pimentel, Fractional-division frequency synthesizer for digital angle-modulation (Thomson-CSF).

16. U.S. Patent 4,686,488 (Aug. 11, 1987), C. Attenborough, Fractional-*N* frequency synthesizer with modulation compensation, (Plessey Overseas Ltd., Ilford, UK).

17. U.S. Patent 3,913,928 (Oct., 1975), R. J. Bosselaers, PLL including an arithmetic unit.

18. U.S. Patent 2,976,945, R. G. Cox, Frequency synthesizer (Hewlett-Packard).

19. U.S. Patent 4,586,005 (April 29, 1986), J. A. Crawford, Enhanced analog phase interpolation for fractional-*N* frequency synthesis (Hughes Aircraft Co., Los Angeles).

20. U.S. Patent 4,468,632 (Aug. 28, 1984), A. T. Crowley, PLL frequency synthesizer including fractional digital frequency divider (RCA Corp., New York).

21. U.S. Patent 4,763,083 (Aug. 9, 1988), A. P. Edwards, Low phase noise RF synthesizer (Hewlett-Packard), Palo Alto, CA.

22. U.S. Patent 4,752,902 (June 21, 1988), B. G. Goldberg, Digital frequency synthesizer (Sciteq Electronics, Inc., San Diego, CA).

23. U.S. Patent 4,958,310 (Sept. 18, 1990), B. G. Goldberg, Digital frequency synthesizer having multiple processing paths.

24. U.S. Patent 5,224,132 (June 29, 1993), B. G. Goldberg, Programmable fractional-*N* frequency synthesizer (Sciteq Electronics, Inc., San Diego, CA).

25. B. G. Goldberg, Analog and digital fractional-*n* PLL frequency synthesis: A survey and update, *Appl. Microwave Wireless* 32–42 (June 1999).

26. U.S. Patent 3,882,403, W. G. Greken, Digital frequency synthesizer (General Dynamics).

27. U.S. Patent 5,093,632 (March 3, 1992), A. W. Hietala and D. C. Rabe, Latched accumulator fractional *N* synthesis with residual error reduction (Motorola, Inc., schaumburg, IL).

28. U.S. Patent 3,734,269 (May, 1973), L. Jackson, Digital frequency synthesizer.

29. U.S. Patent 4,758,802 (July 19, 1988), T. Jackson, Fractional-*N* synthesizer (Plessey Overseas Ltd., Ilford, UK).

30. U.S. Patent 4,800,342 (Jan. 24, 1989), T. Jackson, Frequency synthesizer of the fractional type (Plessey Overseas Ltd., Ilford, UK).

31. Eur. Patent 0214217B1 (June 6, 1996), T. Jackson, Improvement in or relating to synthesizers (Plessey Overseas Ltd., Ilford, Essex, UK).

32. Eur. Patent WO86/05046 (Aug. 28, 1996), T. Jackson, Improvement in or relating to synthesizers (Plessey Overseas Ltd., Ilford, Essex, UK).

33. U.S. Patent 4,204,174 (May 20, 1980), N. J. R. King, Phase locked loop variable frequency generator (Racal Communications Equipment Ltd., England).

34. U.S. Patent 4,179,670 (Dec. 18, 1979), N. G. Kingsbury, Frequency synthesizer with fractional division ratio and jitter compensation (Marconi Co. Ltd., Chelmsford, UK).

35. U.S. Patent 3,928,813 (Dec. 23, 1975), C. A. Kingsford-Smith, Device for synthesizing frequencies which are rational multiples of a fundamental frequency (Hewlett-Packard, Palo Alto, CA).

36. U.S. Patent 4,918,403 (April 17, 1990), F. L. Martin, Frequency synthesizer with spur compensation (Motorola, Inc., Schaumburg, IL).

37. U.S. Patent 4,816,774 (March 28, 1989), F. L. Martin, Frequency synthesizer with spur compensation (Motorola).

38. U.S. Patent 4,599,579 (July 8, 1986), K. D. McCann, Frequency synthesizer having jitter compensation (U.S. Philips Corp., New York).

39. U.S. Patent 5,038,117 (Aug. 6, 1991), B. M. Miller, Multiple-modulator fractional-*N* divider (Hewlett-Packard).

40. U.S. Patent 4,206,425 (June 3, 1980), E. J. Nossen, Digitized frequency synthesizer (RCA Corp., New York).

41. O. Peña, SPICE tools provide behavioral modeling of PLLs, *Microwaves RF Mag.* 71–80 (Nov. 1997).

42. U.S. Patent 4,815,018 (March 21, 1989), V. S. Reinhardt and I. Shahriary, Spurless fractional divider direct digital frequency synthesizer and method (Hughes Aircraft Co., Los Angeles).

43. U.S. Patent 4,458,329 (July 3, 1984), J. Remy, Frequency synthesizer including a fractional multiplier, (Adret Electronique, Paris).

44. U.S. Patent 4,965,531 (Oct. 23, 1990), T. A. D. Riley, Frequency synthesizers having dividing ratio controlled sigma-delta modulator (Carleton Univ., Ottawa, Canada).

45. U.S. Patent 5,021,754 (June 4, 1991), W. P. Shepherd, D. E. Davis, and W. F. Tay, Fractional-*N* synthesizer having modulation spur compensation (Motorola, Inc., Schaumburg, IL).

46. U.S. Patent 3,959,737 (May 25, 1976), W. J. Tanis, Frequency synthesizer having fractional-*N* frequency divider in PLL (Engelman Microwave).

47. Eur. Patent 0125790B2 (July 5, 1995), J. N. Wells, Frequency synthesizers (Marconi Instruments, St. Albans, Hertfordshire, UK).

48. U.S. Patent 4,609,881 (Sept. 2, 1986), J. N. Wells, Frequency synthesizers (Marconi Instruments Ltd., St. Albans, UK).

49. U.S. Patent 4,410,954 (Oct. 18, 1983), C. E. Wheatley, III, Digital frequency synthesizer with random jittering for reducing discrete spectral spurs (Rockwell International Corp., El Segundo, CA).

50. U.S. Patent 5,038,120 (Aug. 6, 1991), M. A. Wheatley, L. A. Lepper, and N. K. Webb, Frequency modulated phase locked loop with fractional divider and jitter compensation (Racal-Dana Instruments Ltd., Berkshire, UK).

51. U.S. Patent 4,573,176 (Feb. 25, 1986), R. O. Yaeger, Fractional frequency divider, (RCA Corp., Princeton, NJ).

52. U. L. Rohde, A high performance synthesizer for base stations based on the fractional-*N* synthesis principle, *Microwaves RF Mag.* (April 1998).

53. U. L. Rohde and G. Klage, Analyze VCOs and fractional-*N* synthesizers, *Microwaves RF Mag.* (Aug. 2000).

54. R. Hassun and A. Kovalic, An arbitrary waveform synthesizer for DC to 50 MHz, *Hewlett-Packard J.* (Palo Alto, CA) (April 1988).

55. L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975, Chap. 2.

56. H. T. Nicholas and H. Samueli, An analysis of the output spectrum of direct digital frequency synthesizers in the

presence of phase-accumulator truncation, *41st Annual Frequency Control Symp.*, IEEE Press, New York, 1987.

57. L. B. Jackson, Roundoff noise for fixed point digital filters realized in cascade or parallel form, *IEEE Trans. Audio Electroacous.* **AU-18**: 107–122 (June 1970).
58. Technical Staff of Bell Laboratories, *Transmission Systems for Communication*, Bell Labs, Inc., 1970, Chap. 25.
59. U.S. Patent 4,482,974, A. Kovalick, Apparatus and method of phase to amplitude conversion in a SIN function generator.
60. C. J. Paull and W. A. Evans, Waveform shaping techniques for the design of signal sources, *Radio Electron. Eng.* **44**(10): (Oct. 1974).
61. L. Barnes, Linear-segment approximations to a sinewave, *Electron. Eng.* **40**: (Sept. 1968).
62. U.S. Patent 4,454,486, R. Hassun and A. Kovalic, Waveform synthesis using multiplexed parallel synthesizers.
63. D. K. Kikuchi, R. F. Miranda, and P. A. Thysel, A waveform generation language for arbitrary waveform synthesis, *Hewlett-Packard J.* (Palo Alto, CA) (April 1988).
64. H. M. Stark, *An Introduction to Number Theory*, MIT Press, Cambridge, MA, 1978, Chap. 7.
65. W. Sagun, Generate complex waveforms at very high frequencies, *Electron. Design* (Jan. 26 1989).
66. G. Lowitz and R. Armitano, Predistortion improves digital synthesizer accuracy, *Electron. Design* (March 31 1988).
67. A. Kovalic, Digital synthesizer aids in testing of complex waveforms, *EDN Mag.* (Sept. 1 1988).
68. G. Lowitz and C. Pederson, RF testing with complex waveforms, *RF Design* (Nov. 1988).
69. C. M. Merigold, in Kamilo Fehrer, ed., *Telecommunications Measurement Analysis and Instrumentation*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
70. Synergy Microwave Corporation Designer Handbook, *Specifications of Synthesizers*, 2001.
71. G. Vendelin, A. M. Pavio, and U. L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, Wiley, New York, 1990.
72. J. A. Crawford, *Frequency Synthesizer Design Handbook*, Artech House, Norwood, MA, 1994.
73. D. B. Leeson, Short-term stable microwave sources, *Microwave J.* 59–69 (June 1970).
74. J. Smith, *Modern Communication Circuits*, McGraw-Hill, New York, 1986, pp. 252–261.
75. J. S. Yuan, Modeling the bipolar oscillator phase noise, *Solid State Electron.* **37**(10): 1765–1767 (Oct. 1994).
76. D. Scherer, Design principles and test methods for low phase noise RF and microwave sources, *RF & Microwave Measurement Symp. Exhibition*, Hewlett-Packard.
77. S. Alechno, Analysis method characterizes microwave oscillators, *Microwaves RF Mag.* 82–86 (Nov. 1997).
78. W. Anzill, F. X. Kärtner, and P. Russer, Simulation of the single-sideband phase noise of oscillators, *2nd Int. Workshop of Integrated Nonlinear Microwave and Millimeterwave Circuits*, 1992.
79. N. Boutin, RF oscillator analysis and design by the loop gain method, *Appl. Microwave Wireless* 32–48 (Aug. 1999).
80. P. Braun, B. Roth, and A. Beyer, A measurement setup for the analysis of integrated microwave oscillators, *2nd Int. Workshop of Integrated Nonlinear Microwave and Millimeterwave Circuits*, 1992.
81. C. R. Chang et al., Computer-aided analysis of free-running microwave oscillators, *IEEE Trans. Microwave Theory Tech.* **39**: 1735–1745 (Oct. 1991).
82. P. Davis et al., Silicon-on-silicon integration of a GSM transceiver with VCO resonator, *1998 IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers*, pp. 248–249.
83. P. J. Garner, M. H. Howes, and C. M. Snowden, *Ka*-band and MMIC pHEMT-basd VCO's with low phase-noise properties, *IEEE Trans. Microwave Theory Tech.* **46**: 1531–1536 (Oct. 1998).
84. A. V. Grebennikov and V. V. Nikiforov, An analytic method of microwave transistor oscillator design, *Int. J. Electron.* **83**: 849–858 (Dec. 1997).
85. A. Hajimiri and T. H. Lee, A general theory of phase noise in electrical oscillators, *IEEE J. Solid-State Circuits* **33**: 179–194 (Feb. 1998).
86. Q. Huang, On the exact design of RF oscillators, *Proc. IEEE 1998 Custom Integrated Circuits Conf.*, pp. 41–44.
87. Fairchild Data Sheet, *Phase/Frequency Detector, 11C44*, Fairchild Semiconductor, Mountain View, CA.
88. Fairchild Preliminary Data Sheet, *SH8096 Programmable Divider-Fairchild Integrated Microsystems*, April 1970.
89. U. L. Rohde, *Digital PLL Frequency Synthesizers — Theory and Design*, Prentice-Hall, Englewood Cliffs, NJ, Jan. 1983.
90. W. Egan and E. Clark, Test your charge-pump phase detectors, *Electron. Design* **26**(12): 134–137 (June 7 1978).
91. S. Krishnan, Diode phase detectors, *Electron. Radio Eng.* 45–50 (Feb. 1959).
92. Motorola Data Sheet, *MC12012*, Motorola Semiconductor Products, Inc. Phoenix, AZ, 1973.
93. Motorola Data Sheet, *Phase-Frequency Detector, MC4344, MC4044*.
94. U. L. Rohde and D. P. Newkirk, *RF/Microwave Circuit Design for Wireless Applications*, Wiley, 2000.
95. W. C. Lindsey and M. K. Simon, eds., *Phase-Locked Loops & Their Application*, IEEE Press, New York, 1978.
96. W. Z. Chen and J. T. Wu, A 2 V 1.6 GHz BJT phase-locked loop, *Proc. IEEE 1998 Custom Integrated Circuits Conf.*, pp. 563–566.
97. J. Craninckx and M. Steyaert, A fully integrated CMOS DCS-1800 frequency synthesizer, 1998 *IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers*, pp. 372–373.
98. B. De Smedt and G. Gielen, Nonlinear behavioral modeling and phase noise evaluation in phase locked loops, *Proc. IEEE 1998 Custom Integrated Circuits Conf.*, pp. 53–56.
99. N. M. Filiol et al., An agile ISM band frequency synthesizer with built-in GMSK data modulation, *IEEE J. Solid-State Circuits* **33**(7): 998–1007 (July 1998).
100. Fujitsu Microelectronics, Inc., *Super PLL Application Guide*, 1998.
101. Hewlett-Packard Application Note 164-3, *New Technique for Analyzing Phase-Locked Loops*, June 1975.
102. V. F. Kroupa, ed., *Direct Digital Frequency Synthesizers*, IEEE Press, New York, 1999.

103. S. Lo, C. Olgaard, and D. Rose, A 1.8V/3.5 mA 1.1 GHz/300 MHz CMOS dual PLL frequency synthesizer IC for RF communications, *Proc. IEEE 1998 Custom Integrated Circuits Conf.*, pp. 571–574.

104. G. Palmisano et al., Noise in fully-integrated PLLs, *Proc. 6th AACD 97*, Como, Italy, 1997.

105. B. H. Park and P. E. Allen, A 1 GHz, low-phase-noise CMOS frequency synthesizer with integrated LC VCO for wireless communications, *Proc. IEEE 1998 Custom Integrated Circuits Conf.*, pp. 567–570.

106. B. Sam, Hybrid frequency synthesizer combines octave tuning range and millihertz steps, *Appl. Microwave Wireless* 76–84 (May 1999).

107. M. Smith, *Phase Noise Measurement Using the Phase Lock Technique*, Wireless Subscriber Systems Group (WSSG) AN1639, Motorola.

108. V. von Kaenel et al., A 600 MHz CMOS PLL microprocessor clock generator with a 1.2 GHz VCO, *1998 IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers*, pp. 396–397.

109. Motorola MECL Data Book, Chapter 6, *Phase-Locked Loops*, 1993.

110. U. L. Rohde, Low noise microwave synthesizers, WFFDS: Advances in microwave and millimeter-wave synthesizer technology, *IEEE-MTT-Symp.*, Orlando, FL, May 19, 1995.

111. U. L. Rohde, Oscillator design for lowest phase noise, *Microwave Eng. Eur.* 35–40 (May 1994).

G

GENERAL PACKET RADIO SERVICE (GPRS)

CHRISTIAN BETTSTETTER
CHRISTIAN HARTMANN
Technische Universität München
Institute of Communication Networks
Munich, Germany

1. INTRODUCTION

The *General Packet Radio Service* (GPRS) is a data bearer service for GSM and IS136 cellular networks. Its packet-oriented radio transmission technology enables efficient and simplified wireless access to Internet protocol-based networks and services. With GPRS-enabled mobile devices, users benefit from higher data rates [up to ~50 kbps (kilobits per second)], shorter access times, an "always on" wireless connectivity, and volume-based billing.

In conventional GSM networks without GPRS, access to external data networks from GSM mobile devices has already been standardized in GSM phase 2; however, on the air interface, such access occupied a complete circuit-switched traffic channel for the entire call period. In case of bursty traffic (e.g., Internet traffic), it is obvious that packet-switched bearer services result in a much better utilization of the traffic channels. A packet channel will be allocated only when needed and will be released after the transmission of the packets. With this principle, multiple users can share one physical channel (statistical multiplexing). Moreover, in conventional GSM exactly one channel is assigned for both uplink and downlink. This entails two disadvantages: (1) only symmetric connections are supported and (2) the maximum data rate is restricted since the use of multiple channels in parallel is not possible. In order to address the inefficiencies of GSM phase 2 data services, the General Packet Radio Service has been developed in GSM phase 2+ by the *European Telecommunications Standards Institute* (ETSI). It offers a genuine packet-switched transmission technology also over the air interface, and provides access to networks based on the *Internet Protocol* (IP) (e.g., the global Internet or private/corporate intranets) and X.25. GPRS enables asymmetric data rates as well as simultaneous transmission on several channels (multislot operation). Initial work on the GPRS standardization began in 1994, and the main set of specifications was approved in 1997 and completed in 1999. Market introduction took place in the year 2000. GPRS was also standardized for IS136 [1]; however, in this description we focus on GPRS for GSM.

Data transmission in conventional circuit switched GSM is restricted to 14.4 kbps, and the connection setup takes several seconds. GPRS offers almost ISDN-like data rates up to ~50 kbps and session establishment times below one second. Furthermore, GPRS supports a more user-friendly billing than that offered by circuit-switched data services. In circuit-switched services, billing is based on the duration of the connection. This is unsuitable for applications with bursty traffic, since the user must pay for the entire airtime even for idle periods when no packets are sent (e.g., when the user reads a Webpage). In contrast to this, packet-switched services allow charging based on the amount of transmitted data (e.g., in kilobytes) and the *quality of service* (QoS) rather than connection time. The advantage for the user is that he/she can be online over a long period of time ("always online") but will be billed only when data are actually transmitted. The network operators can utilize their radio resources in a more efficient way and simplify the access to external data networks.

Typical scenarios for GPRS are the wireless access to the *World Wide Web* (WWW) and corporate local-area networks. Here, GPRS provides an end-to-end IP connectivity; that is, users can access the Internet without first requiring to dial in to an Internet service provider. Examples for applications that can be offered over GPRS are mobile e-commerce services, location-based tourist guides, and applications in the telemetry field. GPRS can also be used as a bearer for the *Wireless Application Protocol* (WAP) and the *Short Message Service* (SMS).

Considering the network evolution of second generation cellular networks to third generation, GPRS enables a smooth transition path from GSM/TDMA networks toward the *Universal Mobile Telecommunication System* (UMTS). Especially, the IP backbone network of GPRS forms the basis for the UMTS core network. With the introduction of *Enhanced Data Rates for GSM Evolution* (EDGE), which will use 8-PSK modulation, GPRS will offer approximately a 3 times higher data rate and a higher spectral efficiency. This mode will be called *Enhanced GPRS* (EGPRS).

2. SYSTEM ARCHITECTURE

To incorporate GPRS into existing GSM networks, several modifications and enhancements have been made in the GSM network infrastructure as well as in the mobile stations. On the network side, a new class of nodes has been introduced, namely, the *GPRS support nodes* (GSNs). They are responsible for routing and delivery of data packets between the mobile stations and external packet data networks. Figure 1 shows the resulting GSM/GPRS system architecture [2,3]. A mobile user carries a *mobile station* (MS), which can communicate over a wireless link with a *base transceiver station* (BTS). Several BTSs are controlled by a *base station controller* (BSC). The BTSs and BSC together form a *base station subsystem* (BSS).

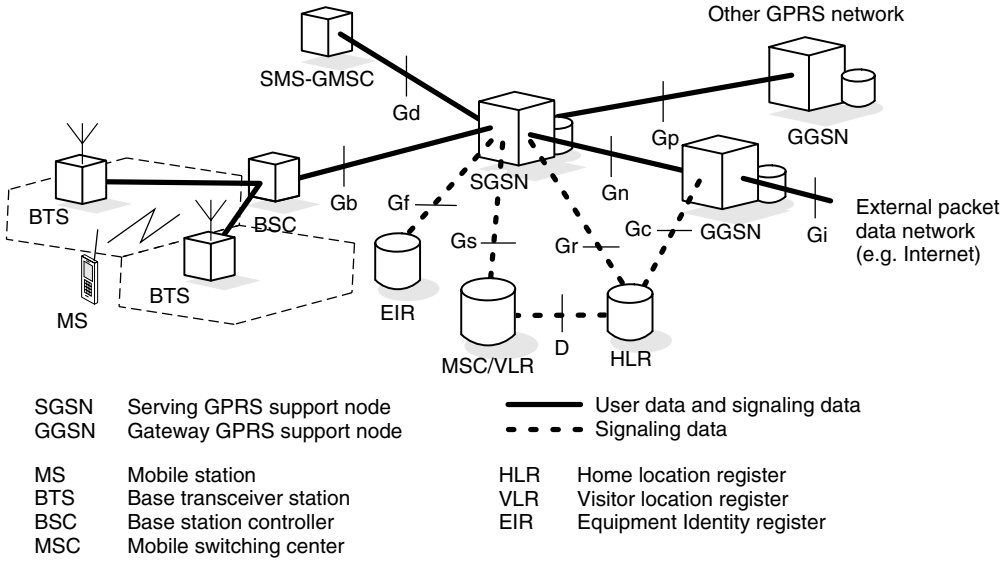


Figure 1. GPRS system architecture and interfaces.

For an detailed explanation of the GSM elements and the basic addresses, see the → GSM entry.

A *serving GPRS support node* (SGSN) delivers packets from and to the MSs within its service area. Its tasks include packet routing and transfer, functions for attach/detach of MSs and their authentication, mobility management, radio resource management, and logical link management. The location register of the SGSN stores location information of all GPRS users registered with this SGSN [e.g., current cell, current *visitor location register* (VLR)] and their user profiles [e.g., *international mobile subscriber identity* (IMSI), address used in the packet data network].

A *gateway GPRS support node* (GGSN) acts as an interface to external packet data networks (e.g., to the Internet). It converts GPRS packets coming from the SGSN into the appropriate *packet data protocol* (PDP) format (i.e., IP) and sends them out on the corresponding external network. In the other direction, the PDP address of incoming data packets (e.g., the IP destination address) is mapped to the GSM address of the destination user. The readdressed packets are sent to the responsible SGSN. For this purpose, the GGSN stores the current SGSN addresses and profiles of registered users in its location register.

The functionality of the SGSN and GGSN can be implemented in a single physical unit or in separate units. All GSNs are connected via an IP-based backbone network. Within this GPRS backbone, the GSNs encapsulate the external packets and transmit (tunnel) them using the so-called *GPRS tunneling protocol* (GTP). If there is a roaming agreement between two operators, they may install an inter-operator backbone between their networks (see Fig. 2). The *border gateways* (BGs) perform security functions in order to protect the private GPRS networks against attacks and unauthorized users.

Via the Gn and the Gp interfaces (see Figs. 1 and 2), user payload and signaling data are transmitted between the GSNs. The Gn interface will be used, if SGSN and GGSN are located in the same network, whereas the Gp interface will be used, if they are in different networks. These interfaces are also defined between two SGSNs. This allows the SGSNs to exchange user profiles when a mobile station moves from one SGSN area to another. The Gi interface connects the GGSN with external networks. From an external network's point of view, the GGSN looks like a usual IP router, and the GPRS network looks like any other IP subnetwork.

Figure 1 also shows the signaling interfaces between the GSNs and the conventional GSM network entities [2]. Across the Gf interface, the SGSN may query and verify the *international mobile equipment identity* (IMEI) of a mobile station trying to attach to the network. GPRS also adds some entries to the GSM registers. For mobility management, a user's entry in the *home location register* (HLR) is extended with a link to its current SGSN. Moreover, his/her GPRS-specific profile and current PDP address(es) are stored. The Gr interface is used to exchange this information between HLR and SGSN. For example, the SGSN informs the HLR about the current location of the mobile station, and when a mobile station registers with a new SGSN, the HLR will send the user profile to the new SGSN. In a similar manner, the signaling interface between GGSN and HLR (Gc interface) may be used by the GGSN to query the location and profile of a user who is unknown to the GGSN. In addition, the MSC/VLR may be extended with functions and register entries that allow efficient coordination between packet-switched and conventional circuit-switched GSM services. Examples for this optional feature are combined GPRS and GSM location updates and combined attachment procedures. Moreover, paging requests of circuit-switched GSM calls can be performed via the SGSN. For this purpose, the Gs interface connects the registers of SGSN and MSC/VLR.

In order to exchange messages of the Short Message Service via GPRS, the Gd interface interconnects the *SMS gateway MSC* (SMS-GMSC) with the SGSN.

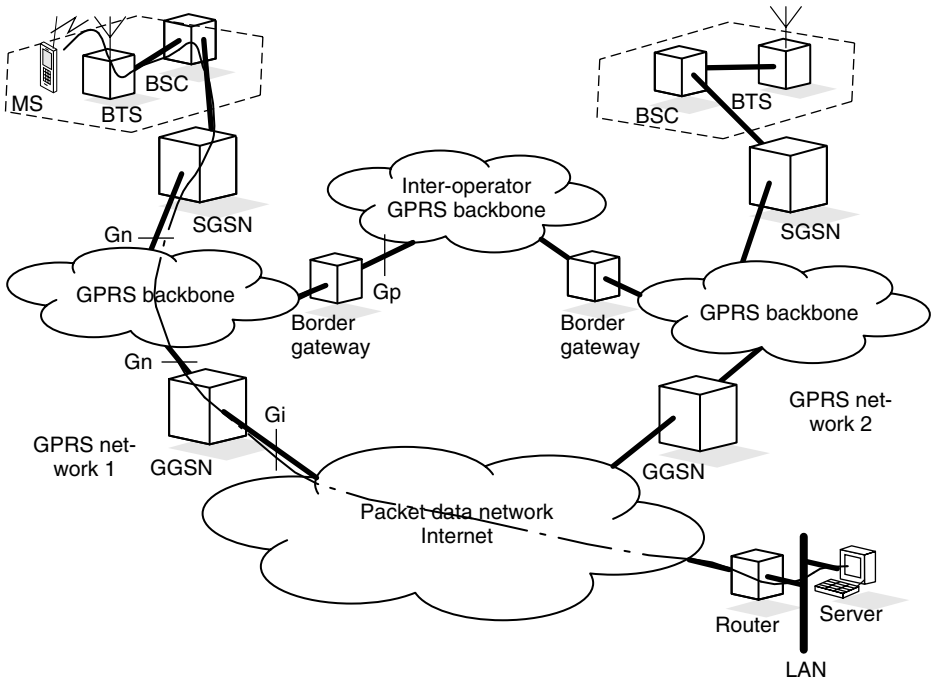


Figure 2. GPRS system architecture and routing example.

3. SERVICES

The bearer services of GPRS offer end-to-end packet-switched data transfer to mobile subscribers. A *point-to-point* (PTP) service is specified, which comes in two variants [4]: a connectionless mode for IP and a connection-oriented mode for X.25. Furthermore, SMS messages can be sent and received over GPRS.

It is planned to implement a *point-to-multipoint* (PTM) service, which offers transfer of data packets to several mobile stations. For example, IP multicast routing protocols can be employed over GPRS for this purpose [5]. Packets addressed to an IP multicast group will then be routed to all group members. Also supplementary services can be implemented, such as reverse charging and barring.

Based on these standardized bearer and supplementary services, a huge variety of nonstandardized services can be offered over GPRS. The most important application scenario is the wireless access to the Internet and to corporate intranets (for e-mail communication, database access, Web browsing). Also WAP services can be accessed in a more efficient manner than with circuit-switched GSM. Especially mobility-related services that require an "always on" connectivity but only infrequent data transmissions (e.g., interactive city guides) benefit from the packet-oriented transmission and billing method of GPRS.

3.1. Quality of Service (QoS)

Support of different QoS classes is an important feature in order to support a broad variety of applications but still preserve radio and network resources in an efficient way. Moreover, QoS classes enable providers to offer different billing options. The billing can be based on the amount of transmitted data, the service type itself, and the QoS profile. The GPRS standard [4] defines four QoS parameters: service precedence, reliability, delay, and throughput. Using these parameters, QoS profiles can be negotiated between the mobile user and the network for each session, depending on the QoS demand and the currently available resources.

The service precedence is the priority of a service (in relation to other services). There are three levels of priority defined in GPRS. In case of heavy traffic load, for example, packets of low priority will be discarded first.

The reliability indicates the transmission characteristics required by an application. Three reliability classes are defined, which guarantee certain maximum values for the probability of packet loss, packet duplication, missequencing, and packet corruption (i.e., undetected errors in a packet); see Table 1.

As shown in Table 2, the delay parameters define maximum values for the mean delay and the 95-percentile delay. The latter is the maximum delay guaranteed in 95% of all transfers. Here, delay is defined as the end-to-end transfer time between two communicating mobile stations or between a mobile station and the Gi interface to an external network, respectively. This includes all delays within the GPRS network, such as the delay for request and assignment of radio resources, transmission over the air interface, and the transit delay in the GPRS backbone network. Delays outside the GPRS network, for example in external transit networks, are not taken into account.

Table 1. QoS Reliability Classes

Class	Loss	Duplication	Missequencing	Corruption
1	10^{-9}	10^{-9}	10^{-9}	10^{-9}
2	10^{-4}	10^{-5}	10^{-5}	10^{-6}
3	10^{-2}	10^{-5}	10^{-5}	10^{-2}

Table 2. QoS Delay Classes (in seconds)

	128 Byte		1024 Byte	
Class	Mean	95%	Mean	95%
1	<0.5	<1.5	<2	<7
2	<5	<25	<15	<75
3	<50	<250	<75	<375
4	Best effort			

Finally, the throughput parameter specifies the maximum and mean bit rate.

3.2. Simultaneous Usage of Packet-Switched and Circuit-Switched Services

GPRS services can be used in parallel to circuit-switched GSM services. The GPRS standard defines three classes of mobile stations [4]. Mobile stations of class A fully support simultaneous operation of GPRS and conventional GSM services. Class B mobile stations are able to register with the network for both GPRS and conventional GSM services simultaneously and listen to both types of signaling messages, but they can use only one of the service types at a given time. Finally, class C mobile stations can attach for either GPRS or conventional GSM services at a given time. Simultaneous registration (and usage) is not possible, except for SMS messages, which can be received and sent at any time.

4. SESSION MANAGEMENT, MOBILITY MANAGEMENT, AND ROUTING

In this section we describe how a mobile station registers with the GPRS network and becomes known to an external packet data network. We show how packets are routed to or from mobile stations, and how the network keeps track of the user's current location [2].

4.1. Attachment and Detachment Procedure

Before a mobile station can use GPRS services, it must attach to the network (similar to the IMSI attach used for circuit-switched GSM services). The mobile station's ATTACH REQUEST message is sent to the SGSN. The network then checks if the user is authorized, copies the user profile from the HLR to the SGSN, and assigns a *packet temporary mobile subscriber identity* (P-TMSI) to the user. This procedure is called *GPRS attach*. It establishes a logical link between the mobile station and the SGSN, such that the SGSN can perform paging of the mobile station and deliver SMS messages. For mobile stations using circuit- and packet-switched services, it is possible to implement combined GPRS/IMSI attach procedures. The disconnection from the GPRS network is called *GPRS detach*. It can be initiated by the mobile station or by the network.

4.2. Session Management and PDP Context

To exchange data packets with external packet data networks after a successful GPRS attach, a mobile station must apply for an address to be used in the external network. In general, this address is called *PDP address* (packet data protocol address). In case the external network is an IP network, the PDP address is an IP address.

For each session, a so-called PDP context is created [2], which describes the characteristics of the session. It includes the PDP type (e.g., IPv6), the PDP address assigned to the mobile station (e.g., an IP address), the requested QoS class, and the address of a GGSN that serves as the access point to the external network. This context is stored in the MS, the SGSN, and the GGSN. Once a mobile station has an active PDP context, it is visible for the external network and can send and receive data packets. The mapping between the two addresses (PDP $\leftrightarrow$ GSM address) makes the transfer of packets between MS and GGSN possible. In the following we assume that access to an IP-based network is intended.

The allocation of an IP address can be static or dynamic. In the first case, the mobile station permanently owns an IP address. In the second case, using a dynamic addressing concept, an IP address is assigned on activation of a PDP context. In other words, the network provider has reserved a certain number of IP addresses, and each time a mobile station attaches to the GPRS network, it will obtain an IP address. After its GPRS detach, this IP address will be available to other users again. The IP address can be assigned either by the user's home network operator or by the operator of the visited network. The GGSN is responsible for the allocation and deactivation of addresses. Thus, the GGSN should also include DHCP (*Dynamic Host Configuration Protocol* [6]) functionality, which automatically manages the available IP address space.

A basic PDP context activation procedure initialized by an MS is as follows [2]: Using the message ACTIVATE PDP CONTEXT REQUEST, the MS informs the SGSN about the requested PDP context. Afterward, the usual GSM security functions are performed. If access is granted, the SGSN will send a CREATE PDP CONTEXT REQUEST to the affected GGSN. The GGSN creates a new entry in its PDP context table, which enables the GGSN to route data packets between the SGSN and the external network. It confirms this to the SGSN and transmits the dynamic PDP address (if needed). Finally, the SGSN updates its PDP context table and confirms the activation of the new PDP context to the MS.

In case the GGSN receives packets from the external network that are addressed to a known static PDP address of an MS, it can perform a network-initiated PDP context activation procedure.

4.3. Routing

In Fig. 2 we give an example of how IP packets are routed to an external IP-based data network. A GPRS mobile station located in the GPRS network 1 addresses IP packets to a Web server connected to the Internet. The SGSN to which the mobile station is attached encapsulates the IP packets coming from the mobile station, examines the PDP context, and routes them through the GPRS backbone to the appropriate GGSN.

The GGSN decapsulates the IP packets and sends them out on the IP network, where IP routing mechanisms transfer the packets to the access router of the destination network. The latter delivers the IP packets to the Web server.

In the other direction, the Web server addresses its IP packets to the mobile station. They are routed to the GGSN from which the mobile station has its IP address (e.g., its home GGSN). The GGSN queries the HLR and obtains information about the current location of the user. In the following, it encapsulates the incoming IP packets and tunnels them to the appropriate SGSN in the current network of the user. The SGSN decapsulates the packets and delivers them to the mobile station.

4.4. Location Management

As in circuit-switched GSM, the main task of location management is to keep track of the user's current location, so that incoming packets can be routed to his/her MS. For this purpose, the MS frequently sends location update messages to its SGSN.

In order to use the radio resources occupied for mobility-related signaling traffic in an efficient way, a state model for GPRS mobile stations with three states has been defined [2]. In IDLE state the MS is not reachable. Performing a GPRS attach, it turns into READY state. With a GPRS detach it may deregister from the network and fall back to IDLE state, and all PDP contexts will be deleted. The STANDBY state will be reached when an MS in READY state does not send any packets for a long period of time, and therefore the READY timer (which was started at GPRS attach and is reset for each incoming and outgoing transmission) expires.

The location update frequency depends on the state in which the MS currently is. In IDLE state, no location updating is performed; that is, the current location of the MS is unknown. If an MS is in READY state, it will inform its SGSN of every movement to a new cell. For the location management of an MS in STANDBY state, a GSM location area (→ see GSM entry) is divided into so-called *routing areas*. In general, a routing area consists of several cells. The SGSN will be informed when an MS moves to a new routing area; cell changes will not be indicated. In addition to these event-triggered routing area updates, periodic routing area updating is also standardized. To determine the current cell of an MS that is in STANDBY state, paging of the MS within a certain routing area must be performed. For MSs in READY state, no paging is necessary.

Whenever a mobile user moves to a new routing area, it sends a ROUTING AREA UPDATE REQUEST to its assigned SGSN [2]. The message contains the *routing area identity* (RAI) of its old routing area. The BSS adds the *cell identifier* (CI) of the new cell to the request, from which the SGSN can derive the new RAI. Two different scenarios are possible: intra-SGSN routing area updates and inter-SGSN routing area updates.

In the first case, the mobile user has moved to a routing area that is assigned to the same SGSN as the old routing area. The SGSN has already stored the necessary user profile and can immediately assign a new P-TMSI. Since the routing context does not change, there is no need to inform other network elements, such as GGSN or HLR.

In the second case, the new routing area is administered by a SGSN different from the old routing area. The new SGSN realizes that the MS has entered its area. It requests the PDP context(s) of the user from the old SGSN and informs the involved GGSNs about the user's new routing context. In addition, the HLR and (if needed) the MSC/VLR are informed about the user's new SGSN number.

Besides pure routing updates, there also exist combined routing/location area updates. They are performed whenever an MS using GPRS as well as conventional GSM services moves to a new location area.

To sum up, we can say that GPRS mobility management consists of two levels: (1) micromobility management tracks the current routing area or cell of the user and (2) macromobility management keeps track of the user's current SGSN and stores it in the HLR, VLR, and GGSN.

5. PROTOCOL ARCHITECTURE

The protocol architecture of GPRS comprises transmission and signaling protocols. This includes standard GSM protocols (with slight modifications), standard protocols of the Internet Protocol suite, and protocols that have specifically been developed for GPRS. Figure 3 illustrates the protocol architecture of the transmission plane [2]. The architecture of the signaling plane includes functions for the execution of GPRS attach and detach, mobility management, PDP context activation, and the allocation of network resources.

5.1. GPRS Backbone: SGSN-GGSN

As mentioned earlier, the GPRS tunneling protocol (GTP) carries the user's IP or X.25 packets in an encapsulated manner within the GPRS backbone (see Fig. 3). GTP is defined both between GSNs within the same network (Gn interface) and between GSNs of different networks (Gp interface).

The signaling part of GTP specifies a tunneling control and management protocol. The signaling is used to create, modify, and delete tunnels. A *tunnel identifier* (TID), which is composed of the IMSI of the user and a *network-layer service access point identifier* (NSAPI), uniquely indicates a PDP context. Below GTP, one of the standard Internet protocols of the transport layer, *Transmission Control Protocol* (TCP) or *User Datagram Protocol* (UDP), are employed to transport the GTP packets within the backbone network. TCP is used for X.25 (since X.25 expects a reliable end-to-end connection), and UDP is used for access to IP-based networks (which do not expect reliability in the network layer or below). In the network layer, IP is employed to route the packets through the backbone. Ethernet, ISDN, or ATM-based protocols may be used below IP. To summarize, in the GPRS backbone we have an IP/X.25-over-GTP-over-UDP/TCP-over-IP protocol architecture.

For signaling between SGSN and the registers HLR, VLR, and EIR, protocols known from conventional GSM are employed, which have been partly extended with

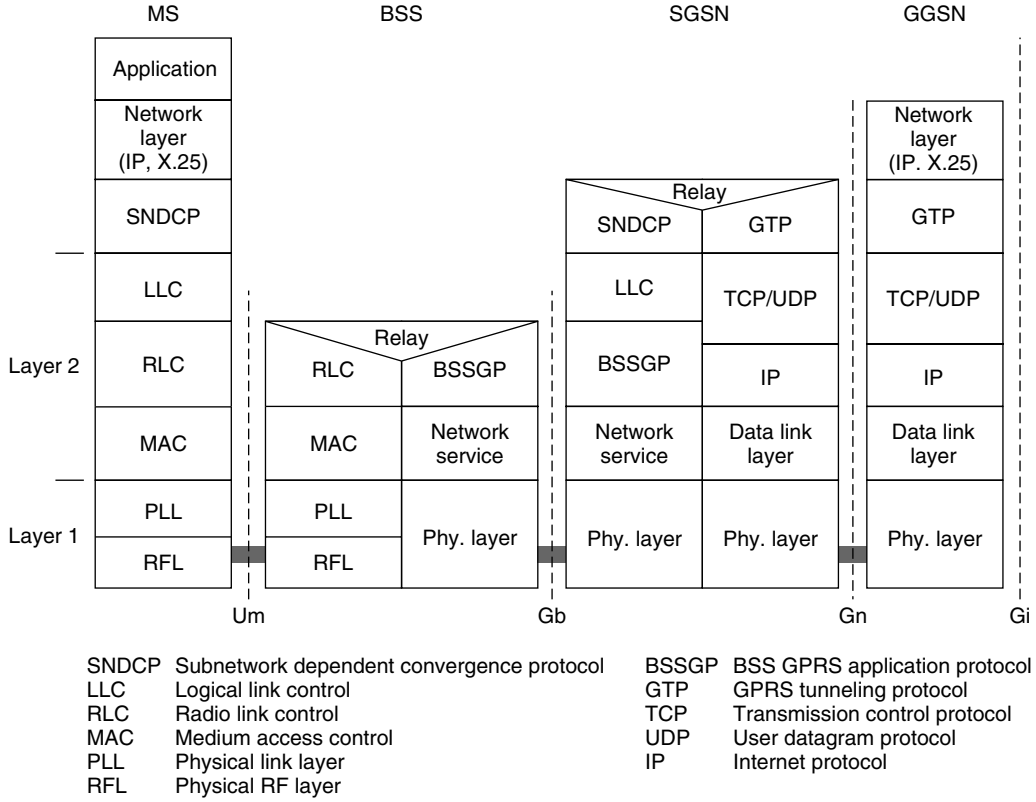


Figure 3. Protocol architecture: transmission plane.

GPRS-specific functionality. Between SGSN and HLR as well as between SGSN and EIR, an enhanced *mobile application part* (MAP) is used. The exchange of MAP messages is accomplished over the *transaction capabilities application part* (TCAP), the *signaling connection control part* (SCCP), and the *message transfer part* (MTP). The *BSS application part* (BSSAP+), which is based on GSM's BSSAP, is applied to transfer signaling information between SGSN and VLR (Gs interface). This includes, in particular, signaling of the mobility management when coordination of GPRS and conventional GSM functions is necessary (e.g., for combined GPRS and non-GPRS location updates, combined GPRS/IMSI attach, or paging of a user via GPRS for an incoming GSM call).

5.2. Air Interface

In the following we consider the transport and signaling protocols at the air interface (Um) between the mobile station and the BSS or SGSN, respectively.

5.2.1. Subnetwork-Dependent Convergence Protocol. An application running in the GPRS mobile station (e.g., a Web browser) uses IP or X.25, respectively, in the network layer. The *subnetwork-dependent convergence protocol* (SNDCP) is used to transfer these packets between the MSs and their SGSN. Its functionality includes multiplexing of several PDP contexts of the network layer onto one virtual logical connection of the underlying *logical link control* (LLC) layer and segmentation of network-layer packets onto LLC frames and reassembly on the receiver side. Moreover, SNDCP offers compression and decompression of user data and redundant header information (e.g., TCP/IP header compression).

5.2.2. GPRS Mobility Management and Session Management. For signaling between an MS and the SGSN, the *GPRS mobility management and session management* (GMM/SM) protocol is employed above the LLC layer. It includes functions for GPRS attach/detach, PDP context activation, routing area updates, and security procedures.

5.2.3. Data-Link Layer. The data-link layer is divided into two sublayers [7]:

- Logical link control (LLC) layer (between MS and SGSN)
- *Radio-link control/medium access control* (RLC/MAC) layer (between MS and BSS)

The LLC layer provides a reliable logical link between an MS and its assigned SGSN. Its functionality is based on the *link access procedure D mobile* (LAPDm) protocol, which is a protocol similar to *high-level data-link control* (HDLC). LLC includes in-order delivery, flow control, error detection, and retransmission of packets [*automatic repeat request* (ARQ)], and ciphering functions. It supports variable frame lengths and different QoS classes, and besides point-to-point, point-to-multipoint transfer is also possible. A logical link is uniquely addressed with a *temporary logical link identifier* (TLLI). The mapping

between TLLI and IMSI is unique within a routing area. However, the user's identity remains confidential, since the TLLI is derived from the P-TMSI of the user.

The RLC/MAC layer has two functions. The purpose of the radio-link control (RLC) layer is to establish a reliable link between the MS and the BSS. This includes the segmentation and reassembly of LLC frames into RLC data blocks and ARQ of uncorrectable blocks. The MAC layer employs algorithms for contention resolution of random access attempts of mobile stations on the radio channel (slotted ALOHA), statistical multiplexing of channels, and a scheduling and prioritizing scheme, which takes into account the negotiated QoS. On one hand, the MAC protocol allows for a single MS to simultaneously use several physical channels (several time slots of the same TDMA frame). On the other hand, it also controls the statistical multiplexing; that is, it controls how several MSs can access the same physical channel (the same time slot of successive TDMA frames). This will be explained in more detail in Section 6.

5.2.4. Physical Layer. The physical layer between MS and BSS can be divided into the two sublayers: *physical link layer* (PLL) and *physical radiofrequency layer* (RFL). The PLL provides a physical channel between the MS and the BSS. Its tasks include channel coding (i.e., detection of transmission errors, forward error correction, and indication of uncorrectable codewords), interleaving, and detection of physical link congestion. The RFL operates below the PLL and includes modulation and demodulation.

5.2.5. Data Flow. To conclude this section, Fig. 4 illustrates the data flow between the protocol layers in the mobile station. Packets of the network layer (e.g., IP packets) are passed down to the SNDCP layer, where they are segmented to LLC frames. After adding header information and a *framecheck sequence* (FCS) for error protection, these frames are segmented into one or several RLC data blocks. Those are then passed down to the MAC layer. One RLC/MAC block contains a MAC and RLC header, the RLC payload (information bits), and a *block-check sequence* (BCS) at the end. The channel coding of RLC/MAC blocks and the mapping to a burst in the physical layer will be explained in Section 6.3.

5.3. BSS-SGSN Interface

At the Gb interface, the *BSS GPRS application protocol* (BSSGP) is defined on layer 3. It delivers routing and QoS-related information between BSS and SGSN. The underlying *network service* (NS) protocol is based on the Frame Relay protocol.

5.4. Routing and Conversion of Addresses

We now explain the routing of incoming IP packets in more detail. Figure 5 illustrates how a packet arrives at the GGSN and is then routed through the GPRS backbone to the responsible SGSN and finally to the MS. Using the PDP context, the GGSN determines from the IP destination address a TID and the IP address of the relevant SGSN. Between GGSN and the SGSN, the GPRS tunneling protocol is employed. The SGSN derives the TLLI from the TID and finally transfers the IP packet to the MS. The NSAPI, which is part of the TID, maps a given IP address to the corresponding PDP context. An NSAPI/TLLI pair is unique within one routing area.

6. AIR INTERFACE

The packet-oriented air interface [7] is one of the key aspects of GPRS. Mobile stations with multislot capability can transmit on several time slots of a TDMA frame, uplink and downlink are allocated separately, and physical channels are assigned only for the duration of the transmission, which leads to a statistical multiplexing gain. This flexibility in the channel allocation results in a more efficient utilization of the radio resources. On top of the physical channels, a number of logical packet channels have been standardized. A special packet traffic channel is used for payload transmission. The GPRS signaling channels are used, such as for broadcast of system information, multiple access control, and paging. GPRS channel coding defines four different coding schemes, which allow one to adjust the tradeoff between the level of error protection and data rate.

6.1. Logical Channels

Several GPRS logical channels are defined in addition to the logical channels of GSM. As with logical channels in conventional GSM, they can be grouped into two categories: traffic channels and signaling (control) channels. The signaling channels can be further divided into packet broadcast control, packet common control, and packet dedicated control channels.

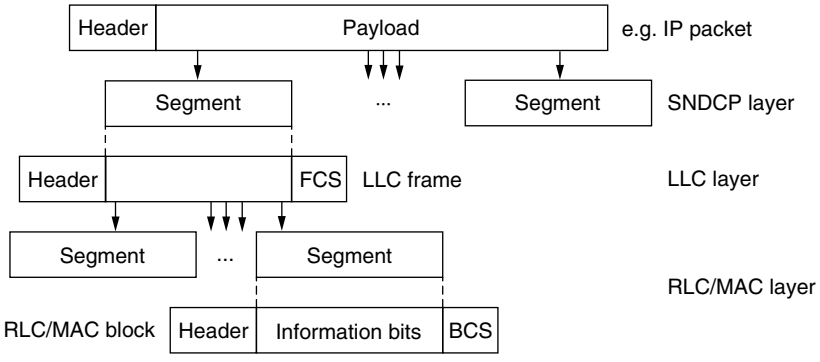


Figure 4. Data flow and segmentation between the protocol layers in the MS.

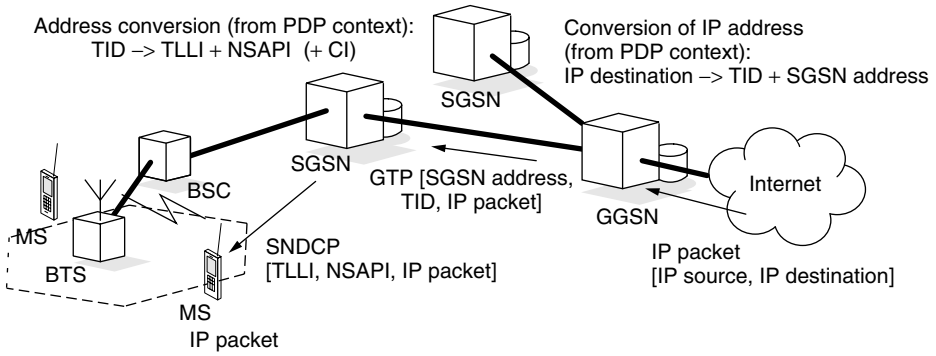


Figure 5. Routing and address conversion: incoming IP packet.

The *packet data traffic channel* (PDTCH) is employed for the transfer of user data. It is assigned to one mobile station (or in case of PTM to multiple mobile stations). One mobile station can use several PDTCHs simultaneously.

The *packet broadcast control channel* (PBCCH) is a unidirectional point-to-multipoint signaling channel from the BSS to the mobile stations. It is used to broadcast information about the organization of the GPRS radio network to all GPRS mobile stations of a cell. Besides system information about GPRS, the PBCCH should also broadcast important system information about circuit-switched services, so that a GSM/GPRS mobile station does not need to listen to the GSM *broadcast control channel* (BCCH).

The *packet common control channels* (PCCCHs) transport signaling information for functions of the network access management, specifically, for allocation of radio channels, medium access control, and paging. The group of PCCCHs comprises the following channels:

- The *packet random access channel* (PRACH) is used by the mobile stations to request one or more PDTCH.
- The *packet access grant channel* (PAGCH) is used to allocate one or more PDTCH to a mobile station.
- The *packet paging channel* (PPCH) is used by the BSS to determine the location of a mobile station (paging) prior to downlink packet transmission.
- The *packet notification channel* (PNCH) is used to inform mobile stations of incoming PTM messages.

The packet dedicated control channels are bidirectional point-to-point signaling channels. This group consists of the following two channels:

- The *packet-associated control channel* (PACCH) is always allocated in combination with one or more PDTCH. It transports signaling information related to one specific mobile station (e.g., power control information).
- The *packet timing-advance control channel* (PTCCH) is used for adaptive frame synchronization. The MS sends over the uplink part of the PTCCH, the PTCCH/U, access bursts to the BTS. From the delay of these bursts, the correct value for the timing advance can be derived. This value is then transmitted in the downlink part, the PTCCH/D, to inform the MS.

Coordination between GPRS and GSM logical channels is also possible here to save radio resources. If the PCCCH is not available in a cell, a GPRS mobile station can use the *common control channel* (CCCH) of circuit-switched GSM to initiate the packet transfer. Moreover, if the PBCCH is not available, it can obtain the necessary system information via the BCCH.

Four different coding schemes (CS1–CS4) are defined for data transmission on the PDTCH. Depending on the used coding scheme, the net data throughput on the PDTCH can be 9.05, 13.4, 15.6, or 21.4 kbps. The respective coding schemes are described in Section 6.3. Theoretically, a mobile station can be assigned up to eight PDTCHs, each of which can be either unidirectional or bidirectional.

6.2. Multiple Access and Radio Resource Management

On the physical layer, GPRS uses the GSM combination of FDMA and TDMA with eight time slots per TDMA frame. However, several new methods are used for channel allocation and multiple access [7]. They have significant impact on the performance of GPRS. In circuit-switched GSM, a physical channel (i.e., one time slot of successive TDMA frames) is permanently allocated for a particular MS during the entire call period (regardless of whether data are transmitted). Moreover, a GSM connection is always symmetric; that is, exactly one time slot is assigned to uplink and downlink.

GPRS enables a far more flexible resource allocation scheme for packet transmission. A GPRS mobile station can transmit on several of the eight time slots within the same TDMA frame (multislot operation). The number of time slots that an MS is able to use is called *multislot class*. In addition, uplink and downlink are allocated separately,

which saves radio resources for asymmetric traffic (e.g., Web browsing).

The radio resources of a cell are shared by all GSM and GPRS users located in this cell. A cell supporting GPRS must allocate physical channels for GPRS traffic. A physical channel that has been allocated for GPRS transmission is denoted as *packet data channel* (PDCH). The number of PDCHs can be adjusted according to the current traffic demand ("capacity on demand" principle). For example, physical channels not currently in use for GSM calls can be allocated as PDCHs for GPRS to increase the quality of service for GPRS. When there is a resource demand for GSM calls, PDCHs may be de-allocated.

As already mentioned, physical channels for packet-switched transmission (PDCHs) are allocated only for a particular MS when this MS sends or receives data packets, and they are released after the transmission. With this dynamic channel allocation principle, multiple MSs can share one physical channel. For bursty traffic this results in a much more efficient use of the radio resources.

The channel allocation is controlled by the BSC. To prevent collisions, the network indicates in the downlink which channels are currently available. An *uplink state flag* (USF) in the header of downlink packets shows which MS is allowed to use this channel in the uplink. The allocation of PDCHs to an MS also depends on its multislot class and the QoS of its current session.

In the following we describe the procedure of uplink channel allocation (mobile originated packet transfer). A mobile station requests a channel by sending a PACKET CHANNEL REQUEST message on the PRACH or RACH [the GSM equivalent of the PRACH (→ GSM entry)]. The BSS answers on the PAGCH or AGCH [the GSM equivalent of the PAGCH (→ GSM entry)], respectively. Once the PACKET CHANNEL REQUEST is successful, a *temporary block flow* (TBF) is established. With that, resources (e.g., PDTCH and buffers) are allocated for the mobile station, and data transmission can start. During transfer, the USF in the header of downlink blocks indicates to other MSs that this uplink PDTCH is already in use. On the receiver side, a *temporary flow identifier* (TFI) helps to reassemble the packets. Once all data have been transmitted, the TBF and the resources are released again.

The downlink channel allocation (mobile terminated packet transfer) is performed in a similar fashion. Here, the BSS sends a PACKET PAGING REQUEST message on the PPCH or PCH [the GSM equivalent of the PPCH (→ GSM entry)] to the mobile station. The mobile station replies with a PACKET CHANNEL REQUEST message on the PRACH or RACH, and the further channel allocation procedure is similar to the uplink case.

6.3. Channel Coding

Channel coding is used to protect the transmitted data packets against errors and perform forward error correction. The channel coding technique in GPRS is quite similar to the one employed in conventional GSM (→ see GSM entry). An outer block coding, an inner convolutional coding, and an interleaving scheme is used. Figure 6 shows how a block of the RLC/MAC layer is encoded and mapped onto four bursts.

As shown in Table 3, four coding schemes (CS1, CS2, CS3, and CS4) with different code rates are defined [8]. For each scheme, a block of 456 bits results after encoding; however, different data rates are obtained depending on the used coding scheme, due to the different code rates of the convolutional encoder and to different numbers of parity bits. Figure 6 illustrates the encoding process, which will be briefly explained in the following.

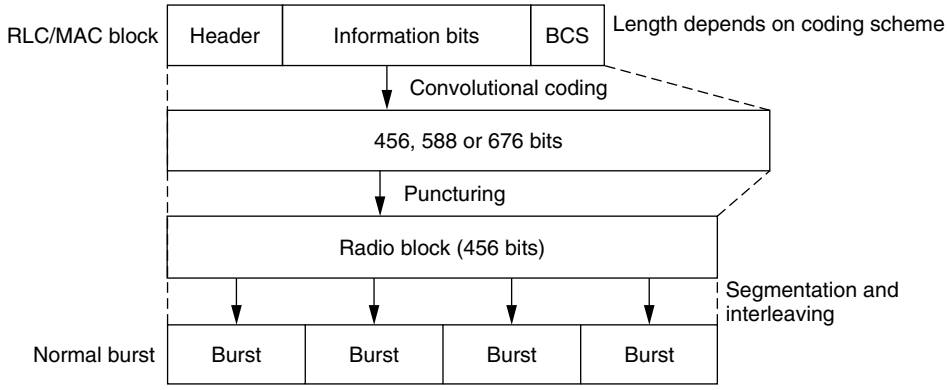


Figure 6. Physical layer at the air interface: channel coding, interleaving, and formation of bursts.

Table 3. Channel Coding Schemes for the Traffic Channels

CS	Preencoded USF	Info Bits without USF	Parity BCS	Tail Bits	Output Convolutional Encoder	Punctured Bits	Code Rate	Data Rate (kbps)
1	3	181	40	4	456	0	$\frac{1}{2}$	9.05
2	6	268	16	4	588	132	$\approx\frac{2}{3}$	13.4
3	6	312	16	4	676	220	$\approx\frac{3}{4}$	15.6
4	12	428	16	—	456	—	1	21.4

As an example we choose coding scheme CS2. First of all, the 271 information bits of an RLC/MAC block (268 bits plus 3 bits USF) are mapped to 287 bits using a systematic block encoder; thus 16 parity bits are added. These parity bits are denoted as *block-check sequence* (BCS). The USF pre-encoding maps the first 3 bits of the block (i.e., the USF) to 6 bits in a systematic way. Afterward, 4 zero bits (tail bits) are added at the end of the entire block. The tail bits are needed for termination of the subsequent convolutional coding. For the convolutional coding, a nonsystematic rate-$\frac{1}{2}$ encoder with memory 4 is used. This is the same encoder as used in conventional GSM for full-rate speech coding (→ see GSM entry). At the output of the convolutional encoder a codeword of length 588 bits results. Afterward, 132 bits are punctured (deleted), resulting in a radio block of length 456 bits. Thus, we obtain a code rate of the convolutional encoder (taking puncturing into account) of $\frac{2}{3}$. After encoding, the codewords are finally fed into a block interleaver of depth 4.

For the encoding of the traffic channels (PDTCH), one of the four coding schemes is chosen, depending on the quality of the signal. The two stealing flags in a normal burst (→ GSM entry) are used to indicate which coding scheme is applied. The signaling channels are encoded using CS1 (an exception is the PRACH).

For CS1 a systematic "fire" code is used for block coding and there is no precoding of the USF bits. The convolutional coding is done with the known rate-$\frac{1}{2}$ encoder, however, this time the output sequence is not punctured. Using CS4, the 3 USF bits are mapped to 12 bits, and no convolutional coding is applied. We achieve a data rate of 21.4 kbps per time slot, and thus obtain a theoretical maximum data rate of 171.2 kbps per TDMA frame. In practice, multiple users share the time slots, and, thus, a much lower bit rate is available to the individual user. Moreover, the quality of the radio channel will not always allow the use of CS4. The data rate available to the user depends (among other things) on the current total traffic load in the cell (i.e., the number of users and their traffic characteristics), the coding scheme used, and the multislot class of the MS.

7. SECURITY ASPECTS

The security principles of GSM (→ GSM entry) have been extended for GPRS [9]. As in GSM, they protect against unauthorized use of services (by authentication and service request validation), provide data confidentiality (using ciphering), and keep the subscriber identity confidential. The standard GSM algorithms are employed to generate security data. Moreover, the two keys known from GSM, the *subscriber authentication key* (Ki) and the *cipher key* (Kc), are used. The main difference is that not the MSC but the SGSN handles authentication. In addition, a special GPRS ciphering algorithm has been defined, which is optimized for encryption of packet data.

7.1. User Authentication

In order to authenticate a user, the SGSN offers a random number to the MS. Using the key Ki, the MS calculates a *signature response* (SRES) and transmits it back to the SGSN. If the mobile station's SRES is equal to the SRES calculated (or maintained) by the SGSN, the user is authenticated and is allowed to use GPRS services. If the SGSN does not have authentication sets for a user (i.e., Kc, random number, SRES), it requests them from the HLR.

7.2. Ciphering

The ciphering functionality is performed in the LLC layer between MS and SGSN (see Fig. 3). Thus, the ciphering scope reaches from the MS all the way to the SGSN (and vice versa), whereas in conventional GSM the scope is only between MS and BTS/BSC. The standard GSM algorithm generates the key Kc from the key Ki and a random number. Kc is then used by the GPRS encryption algorithm for encryption of user data and signaling. The key Kc that is handled by the SGSN is independent of the key Kc handled by the MSC for conventional GSM services. An MS may thus have more than one Kc key.

7.3. Confidentiality of User Identity

As in GSM, the identity of the subscriber (i.e., his/her IMSI) is held confidential. This is done by using temporary identities on the radio channel. In particular, a packet temporary mobile subscriber identity (P-TMSI) is assigned to each user by the SGSN. This address is valid and unique only in the service area of this SGSN. From the P-TMSI, a temporary logical link identity (TLLI) can be derived. The mapping between these temporary identities and the IMSI is stored only in the MS and in the SGSN.

BIOGRAPHIES

Christian Bettstetter is a research and teaching staff member at the Institute of Communication Networks at Technische Universität München TUM, Germany. He graduated from TUM in electrical engineering and information technology Dipl.-Ing. in 1998 and then joined the Institute of Communication Networks, where is he working toward his Ph.D. degree. Christian's interests are in the area of mobile communication networks, where his current main research area is wireless ad-hoc networking. His interests also include 2G and 3G cellular systems and protocols for a mobile Internet. He is coauthor of the book *GSM–Switching, Services and Protocols* (Wiley/Teubner) and a number of articles in journals, books, and conferences.

Christian Hartmann studied electrical engineering at the University of Karlsruhe (TH), Germany, where he received the Dipl.-Ing. degree in 1996. Since 1997, he has been with the Institute of Communication Networks at the Technische Universität München, Germany, as a member of the research and teaching staff, pursuing a doctoral degree. His main research interests are in the area of mobile and wireless networks including capacity and performance evaluation, radio resource management,

modeling, and simulation. Christian Hartmann is a student member of the IEEE.

BIBLIOGRAPHY

1. S. Faccin et al., GPRS and IS-136 integration for flexible network and services evolution, *IEEE Pers. Commun.* **6**: (June 1999).
2. ETSI/3GPP, *GSM 03.60: GPRS Service Description; Stage 2*, technical specification, Mar. 2001.
3. C. Bettstetter, H.-J. Vögel, and J. Eberspächer, GSM phase 2+ General Packet Radio Service GPRS: Architecture, protocols, and air interface, *IEEE Commun. Surv.* **2**(3): (1999).
4. ETSI/3GPP, *GSM 02.60: GPRS Service Description; Stage 1*, technical specification, July 2000.
5. 3GPP, *3G 22.060: GPRS: Service Description; Stage 1*, technical specification, Oct. 2000.
6. R. Droms, Automated configuration of TCP/IP with DHCP, *IEEE Internet Comput.* **3**: (July 1999).
7. ETSI/3GPP, *GSM 03.64: GPRS Overall Description of the Air Interface; Stage 2*, technical specification, Feb. 2001.
8. ETSI, *GSM 05.03: GSM Phase* 2+*: Channel Coding*, technical specification, April 1999.
9. ETSI, *GSM 03.10: GSM phase* 2+*: Security Related Network Functions*, technical specification, July 2001.
10. J. Eberspächer, H.-J. Vögel, and C. Bettstetter, *GSM — Switching, Services, and Protocols*, 2nd ed., Wiley, 2001.
11. Y.-B. Lin, H. C.-H. Rao, and I. Chlamtac, General packet radio service (GPRS): Architecture, interfaces, and deployment, *Wiley Wireless Commun. Mobile Comput.* **1**: (Jan. 2001).
12. B. Walke, *Mobile Radio Networks*, 2nd ed., Wiley, 2002.
13. G. Brasche and B. Walke, Concepts, services, and protocols of the new GSM phase 2+ General Packet Radio Service, *IEEE Commun.* (Aug. 1997).
14. H. Granbohm and J. Wiklund, GPRS: General Packet Radio Service, *Ericsson Rev.* (2): (1999).
15. R. Kalden, I. Meirick, and M. Meyer, Wireless Internet access based on GPRS, *IEEE Pers. Commun.* (April 2000).
16. D. Staehle, K. Leibnitz, and K. Tsipotis, QoS of Internet access with GPRS, *Proc. 4th ACM Int. Workshop on Modeling, Analysis, and Simulation of Wireless and Mobile Systems* (*MSWiM'01*), Rome, Italy, July 2001.
17. J. Korhonen, O. Aalto, A. Gurtov, and H. Laamanen, Measured performance of GSM HSCSD and GPRS, *Proc. IEEE Int. Conf. Commun.* (*ICC*), Helsinki, Finland, June 2001.

FURTHER READING

This article is based on the author's survey paper [3] (© 1999 IEEE) and the GPRS chapter of the Wiley book [10]. Both contain more detailed descriptions of the GPRS architecture, protocols, and air interface. An extensive description of the signaling protocols can be found in Ref. 11. The book [12] also contains a GPRS chapter. The paper [13] gives an overview of GPRS as of 1997 and proposes and analyzes a MAC protocol. The authors of Ref. 14 give an equipment manufacturer's view on the GPRS architecture. A simulative study on GPRS performance was done, for example, in Refs. 15 and 16. Measured performance analyzes of GPRS can be found in Ref. 17.

GEOSYNCHRONOUS SATELLITE COMMUNICATIONS

LIN-NAN LEE
KHALID KARIMULLAH
Hughes Network Systems
Germantown, Maryland

1. THE DEVELOPMENT OF GEOSYNCHRONOUS SATELLITE COMMUNICATIONS

The origin of the geosynchronous satellite concept can be traced back to Arthur C. Clarke. In his 1948 article, Clarke demonstrated the feasibility of providing worldwide communication by placing three radio repeaters as artificial satellites in space, each orbiting the earth with a period of 24 h. The geosynchronous orbit is about 22,300 m above the equator. Satellites on this orbit rotate around the earth at the same rate as the earth spins. They therefore appear as a fixed point in the sky to a fixed point on earth. Clarke's concept is visionary in its use of artificial satellites to bounce the radiowave for worldwide communication. It is also powerful, because a directional antenna at practically any fixed point on earth can be assured to maintain communications with at least one of these three satellites once accurate pointing is accomplished the first time.

In 1962, the United States created the Communications Satellite Corporation (COMSAT) to realize Clarke's vision under a Congressional Act. The first commercial application of geosynchronous satellites was transoceanic communications. In 1964, the International Satellite consortium, or INTELSAT, was formed by COMSAT and the Post Telephone and Telegraph entities (PTTs) around the world to operate a fleet of geosynchronous satellites and provide service to themselves as the carriers' carrier. In 1965, the *Early Bird*, also known as *INTELSAT I*, was successfully launched into orbit to demonstrate the feasibility with limited operational capability. Generations of INTELSAT satellites that followed have changed international telecommunications ever since.

In 1976, the launch of MARISAT satellites by COMSAT marked the beginning of a new era of maritime, mobile communications for ships at sea. The International Maritime Satellite Communications consortium, or INMARSAT, was formed by the PTTs in the following year to provide international maritime communications service. Generations of INMARSAT satellites have since played a very important role in maritime and aeronautical communications worldwide.

Following the success of the INTELSAT system, governments or private organizations in many countries since the 1970s have also launched many national and regional satellite systems. As the purpose of these national and regional systems is to facilitate radio communications within a country or region, a single satellite is typically capable of covering the entire service area, even though multiple satellites may be

employed to provide additional traffic-carrying capability and redundancy. The birth of these national and regional systems has resulted in many forms of new applications, including network television distribution, private data networking, and rural telephony services.

In the late 1980s and early 1990s, satellite direct broadcast of television to homes emerged as cost for receiving equipment came down. These direct broadcast satellites provide tens of millions of rural as well as urban and suburban consumers with access to a large number of channels of television programming using small, inexpensive satellite receivers within their coverage areas. New geosynchronous satellite systems capable of public switched telephone network (PSTN) access from consumer handheld mobile terminals have also been developed. At this writing, two audio broadcast satellite systems are in the process of launching their services in the United States to provide digital compact-disk (CD) quality audio programming to automobiles. A new generation of onboard processing satellites is being developed for broadband Internet access. It is clear that new geosynchronous satellite systems and applications continue to evolve well into the twenty-first century as the required technologies continue to become more available.

2. FREQUENCY BANDS AND ORBITAL SLOTS

Like all radio systems, bandwidth is a resource being shared by all communications. Coordination is required to prevent systems from interfering one another. Frequency bands are allocated by the International Telecommunications Union (ITU), based on the intended use. Satellite bands are typically designated as Fixed Satellite Services (FSS), Mobile Satellite Services (MSS), and Broadcast Satellite Services (BSS). Actual assignment of frequency bands may vary, however, from country to country because of nationalistic considerations as well as requirements to protect legacy systems. As device technology for the lower frequency bands are more available than the higher frequencies, lower-frequency bands are often more desirable and being used up first. Equipment for higher-frequency bands, however, can be constructed with much smaller physical dimension, due to their shorter wavelength. Higher-frequency bands are also less congested. The common frequency bands for geosynchronous satellite communications are C band and Ku band for FSS, L band for MSS, Ku band for BSS. After decades of intensive research and development, Ka-band technology is just becoming commercially viable at the turn of this century, and it has become the new frontier for intensive activities.

For FSS and BSS, highly directional antennas are used to transmit and receive signals at the ground terminal. It is therefore possible to reuse the same frequency band by other satellites a few degrees away. Orbital slots are allocated by the ITU along the geosynchronous ring around the world. The spacing between satellites is 2° for FSS. It is extended to 9° for BSS to allow the use of smaller receive antennas. To accommodate the use of low-gain handheld antennas, however, the MSS satellites need much wider orbit spacing.

To maximize reliability, early geosynchronous satellites were designed as simple repeaters. The signals are transmitted from the ground and received by the satellite in one frequency band. They are frequency translated to another frequency band, then amplified and transmitted back to the ground terminals. This approach generally fits well with the peer-to-peer network architecture common for the FSS. The communications link that ground terminals transmit and the satellite receives is referred to as *uplink*, and that satellite transmits and ground terminals receive is referred to as *downlink*.

For MSS, however, the mobile terminals generally need to communicate with a PSTN through a gateway station, which is fixed and much more capable. Also, the MSS frequency bands are generally quite limited. Therefore, signal from the mobile terminals in the MSS band is translated to a feeder link frequency, then sent down to the gateway stations. Similarly, signals from the gateway station are transmitted on a feeder link frequency to the satellite, and then amplified and send to the MSS downlink. In this manner, the more valuable MSS spectrum is more efficiently used.

3. SATELLITE ANTENNA BEAM PATTERNS

As the spectrum and orbital slots are fixed resources that cannot expand with traffic increase, frequency reuse on the same satellite is essential to increase the capacity of each satellite. One way to accomplish it is to use spatial separation between multiple antenna beams.

As a primary application for the INTELSAT series of satellites is transoceanic trunking, it becomes obvious that the same frequency band can be reused at both continents across the ocean with two separate, focused beams. The uplink from the one beam is connected to the downlink to the other beam and vice versa. This leads to the use of so called “hemibeam” and “spot beams”. Most of the modern INTELSAT satellites include a mix of “global beam,” hemibeams, and spot beams. Hemibeams and spot beams often use different polarization from the global beam, allowing more frequency reuse.

A narrower, more focused beam also provides higher gain, which helps reduce the size of the ground terminals. Therefore, most national and regional satellite systems employ shaped antenna beams matching their intended service area. For example, the a CONUS beam covers the lower 48 of the continental United States. Spot beams are often added to provide coverage of Hawaii and parts of Alaska in a number of domestic satellite systems.

As more spot beams are placed on a satellite, interconnection between these antenna beams becomes a significant problem. Beam interconnection evolves from a fixed cross-connect to an onboard RF switch, and eventually to an onboard baseband packet switch. (See Sections 10 and 11.)

Multibeam satellite antennas are typically based either on offset multiple feeds or phased-array technology. Higher antenna gain, however, implies larger physical dimensions. Given the limited launching vehicle alternatives, they are relatively easier to implement for higher-frequency bands. At lower frequencies, designs based on

foldable reflectors that are deployed after launch are often used to get around the problem.

4. THE SATELLITE TRANSPONDER

Most of the geosynchronous satellite systems assume a simple, wideband repeater architecture. The frequency band is partitioned into sections of a reasonable bandwidth, typically in the tens of megahertz (MHz). At the receive antenna output, the individual sections are filtered, frequency-converted and then amplified. The amplified signal is then filtered, multiplexed, and transmitted to the ground via the transmit antenna. The repeater chain of such a section is often called a *transponder*. A typical satellite may carry tens of transponders. The amplitude-to-amplitude (AM/AM) and amplitude-to-phase (AM/PM) response of its power amplifier, and the overall amplitude and group delay responses of the filter and multiplexer chain usually characterize the behavior of a satellite transponder. Solid-state power amplifiers are gradually replacing traveling-wave tube (TWT) amplifiers for low and medium power satellites. For further discussion, see Section 11.

Maximum power is delivered by a satellite transponder when its power amplifier is operated at saturation. The amplitude and phase response of the transponder is highly nonlinear when the power amplifier is saturated. Carriers of different frequencies may generate significant intermodulation as a result. Small signals may be suppressed by higher power signals when the transponder is operated in the nonlinear region. However, when a satellite transponder is used to transmit a single wideband carrier, such as a full-transponder analog television (TV) or a high-speed digital bitstream, operating the transponder at or close to saturation minimizes the size of receiving antennas on the ground. When the transponder is used to send many signals with different center frequencies, however, the common practice is to operate the transponder with sufficient input backoff so that the amplifier is operating at its linear region to minimize the impairments due to amplifier nonlinearity.

5. TRANSMISSION TECHNIQUES FOR GEOSYNCHRONOUS SATELLITES

As communications technology has evolved from applications to applications since the mid-1960s, the transmission techniques for geosynchronous satellites also changed. The most significant changes are certainly caused by the conversion from analog to digital in all forms of communications. For example, early transoceanic links carried groups of frequency-division multiplexed (FDM) analog telephony. The power amplifiers at the transmitting earth stations as well as the satellite transponder were required to operate with substantial backoff to ensure their linearity. This technique has very much been replaced by digital trunking, for which a single high-speed digital carrier containing hundreds of telephone circuits may be transmitted by the earth terminal. In this way the earth station power amplifier can be operated much more efficiently.

Because of the relatively higher bandwidth of television signals, most of the analog television signals have been transmitted with "full-transponder TV." The baseband video signal is frequency modulated (FM) to occupy a significant portion of the satellite transponder bandwidth, typically in the range of 24–30 MHz. The satellite transponder is often operated at saturation, because the FM/TV is the only signal in the transponder. In the INTELSAT system, a "half-transponder TV" scheme have also been used. In such case, there are two video signals, each FM modulated (with overdeviation) to ~17.5–20 MHz, depending on the width of the transponder. The satellite transponder is operated with moderate output backoff, typically 2 dB. The half-transponder TV is a tradeoff between transponder bandwidth efficiency and signal quality. Since network television distribution has been a large application segment for national and regional satellite systems, full-transponder FM/TV are still widely used throughout the world.

Advances in solid-state power amplifier technology and use of digital transmission techniques have brought dramatic cost reduction to satellite communications in terms of both space and ground segments. The single-channel per carrier (SCPC) mode of transmission coupled with power efficient forward error correction (FEC) coding allows small ground terminals to transmit a single digital circuit at speeds from a few kilo bits per second (kbps) up to hundreds of kbps with very small antenna and power amplifier. Digital transmission with power efficient FEC coding also allows direct broadcasting satellites to transmit tens of megabits per second of digitally encoded and multiplexed television signals to very small receive antennas with the satellite transponder operating at very close to saturation. Technology advancements in digital modulation, FEC coding, satellite power amplifiers, and antennas have enabled millions of consumers to watch direct satellite broadcasts of hundreds of television programming and to Web-surf via the geosynchronous satellites at speeds comparable to or higher than cable modems or digital subscriber lines (DSLs).

6. MULTIPLE-ACCESS TECHNIQUES FOR GEOSYNCHRONOUS SATELLITES

With the exception of television broadcast and distribution, most of the user applications do not require a single ground terminal to use the entire capacity of a satellite transponder. To use the transponder resource effectively, individual ground terminals may be assigned a different section of the frequency band within the transponder so that they can share the satellite transponder without interfering with one another. Similarly, terminals may use the same frequency, but each is assigned a different time slot periodically, so that only one terminal uses this frequency at any given time instant. Sharing of the same satellite transponder by frequency separation is known as *frequency-division multiple access* (FDMA), sharing by time separation is known as *time-division multiple access* (TDMA). Generally, FDMA requires less transmit power for each ground terminal, but the satellite transponder

must be operated at the linear region, resulting in less efficiency in the downlink. For TDMA, each ground station must send information in a small time slot at higher speed and not transmitting during the rest of the time. Therefore, higher power is needed at the ground transmitter, but the satellite transponder can be operated at saturation. When sending packetized data, TDMA also has the advantage of inherently high burst rate. Many practical systems use a combination of FDMA and TDMA to obtain the best engineering tradeoff.

The ground stations can also modulate their digital transmission using different code sequences with very low, or no cross-correlation, and transmit the resulting signal at the same frequency. The signals can then be separated at the receiver by correlating the received signal with the same sequence. This technique is known as *code-division multiple access* (CDMA). It is possible to combine CDMA with TDMA, FDMA, or TDMA and FDMA.

To achieve higher efficiency, a ground terminal is assigned a transmission resource, such as a frequency, a time slot, or a code sequence, only when they need it. This is accomplished by setting aside a separate communication channel for the ground terminals to request the resources. The transmission resource is then allocated based on the requests received from all the terminals. This general approach is known as *demand assignment multiple access* (DAMA). Most DAMA systems use centralized control, and the resource assignments are sent to the terminals via another separate channel. In some case, a distributed control approach has also been implemented in the INTELSAT system, since all national gateway stations in the INTELSAT network are considered to be equals. For distributed control, requests are received and used by all ground terminals. They then execute the same demand assignment algorithm to reach identical assignments. No separate assignment messages need to be sent. DAMA has been adopted by terrestrial cellular radio later, and is now commonly known as part of the media access control (MAC) protocol.

7. ROUND-TRIP PROPAGATION DELAY AND SERVICE QUALITY

Depending on the location of the ground terminals with respect to the geosynchronous satellites, it takes about 240–270 ms for the transmitted signal to propagate to a geosynchronous satellite and back down to the receive ground terminal. This round-trip propagation delay presents service quality issues for two-way communications. PSTN handsets are connected to the subscriber lines via 4-wire–2-wire 2/4-wire hybrid. As a result of limited isolation at the receiving end hybrid, an attenuated echo of one's own speech can be heard two round trips later. Because of the long round-trip delay of the geosynchronous satellites, the quality of conversational speech can be degraded significantly by the echo. This problem is overcome by the introduction of echo cancelers. An echo canceler stores a digital copy of the speech at the transmit end. It estimates the amount of the round-trip delay and the strength of the echo at the receive end. The properly delayed and attenuated replica of the stored signal is reproduced and subtracted from the received signal, thus canceling the echo. Although long propagation delay also introduces response delay in conversational speech, subjective tests have demonstrated that $\leq$400-ms round-trip delay is tolerable when echo cancelers are employed. The quality of conversational speech degrades considerably even with echo cancellation, however, if two hops of geosynchronous satellite links are in between.

The round-trip propagation delay must also be considered in designing data links using automatic repeat request (ARQ) for error recovery. For high-speed links, a very sizable amount of data can be transmitted during the two round-trip times it takes for an acknowledgment (ack) for the initial data packet to arrive at the transmit terminal. Well-known protocols such as Go-back-N can perform poorly in this situation. The round-trip delay also adversely affects the Internet end-to-end Transmit Control Protocol (TCP). The TCP transmission window is opened gradually on successful transmission of successive packets. The window is cut back quickly when a packet needs to be retransmitted. The longer propagation delay can cause the TCP to operate at a very small transmission window most of the time. Since the end user devices generally are unaware of the presence of a satellite link in between, such issues are best resolved by *performance enhancement proxies* (PEPs). PEP is a software proxy installed at the ground terminal. The proxy at the transmitting terminal emulates the destination device for the transmitting end user, whereas the proxy at the receiving terminal emulates the source device for the receiving end user. The proxy also implements a protocol optimized for the satellite link between the transmit and receive ground terminals. By breaking the communications into three segments, PEP is able to maximize the performance for the satellite link while maintaining the compatibility to the standard data communications protocols with the end user devices at both ends.

8. VERY SMALL APERTURE TERMINALS (VSAT) FOR GEOSYNCHRONOUS SATELLITES

Geosynchronous satellites were initially conceived as "repeaters in the sky." Every ground terminal is capable of communicating to any other ground terminals within a satellite's coverage area on a peer-to-peer basis. This fully connected mesh-shaped network is ideal for transoceanic trunking in the INTELSAT system. When maritime communications via satellite emerged, it became clear that most of the communications are between individual ship-terminals and their home country through their "home" coastal earth station. The INMARSAT system is essentially made up of many star-shaped networks with coastal earth stations as their hubs. Ship-to-ship communications must be relayed through the hub. By trading off the direct peer-to-peer communication capability, total ground segment cost is drastically reduced. Since there are far more remote terminals than hub terminals, minimizing the remote terminal cost tends to minimize the overall system cost. Very small aperture terminals (VSATs) exploit this property to create inexpensive private networks with geosynchronous FSS satellites.

In private networks, the remote VSAT terminals typically need to send data at speeds up to low hundreds of kbps (kilobits per second) to the hub. The remote-to-hub communication is often referred to as "inroute". At these rates, reliable data communications can typically be established via Ku-band FSS satellites using low-cost antennas with diameter less than a meter. Such antennas have reasonable beamwidth, requiring no constant tracking of the satellite position. Typically, VSATs use near-constant envelope digital modulation along with powerful FEC codes, and a small power amplifier operated at near saturation. With SCPC transmission, a large number of remote terminals share a part or a whole satellite transponder on a demand assigned TDMA/FDMA basis. Demand-assigned TDMA/FDMA takes advantage of the low-duty factor, bursty nature of the data traffic, allowing a number of terminals to share a single frequency. To minimize the size of the remote terminal antenna and power amplifier, the hub terminal typically uses a much larger antenna so that the downlink contributes very little to the overall link noise. Signal from each individual remote terminal is demodulated and decoded separately, and then routed to the appropriate terrestrial network interfaces.

For the "outroute," the hub station typically time multiplexes all the traffic toward the remote terminals into a single high-speed time-division multiplex (TDM) carrier, and transmits it to the entire network of terminals via either a part of the transponder at a different frequency, or through a separate transponder. Similarly, because of the much larger transmit antenna at the hub, the overall outroute link noise is dominated by the downlink to the remote terminal. The speed of the TDM data carrier is typically in the range from submegabits per second to tens of megabits to second, scaled according to network size.

In the rare cases for which data must be exchanged between two remote terminals, they are routed through the hub. The extra delay does not cause problem for most private network applications. But double hopping through the synchronous satellite creates too much delay for conversational telephony. Alternative full-mesh VSATs have also been developed for voice applications. In such private networks, a hub is responsible for interface to the PSTN and demand assign control. The remote terminals typically utilize low-bit-rate voice coders such as the ITU G.729 8-kbps voice coder. When a remote-to-remote call is initiated, the hub assigns a pair of frequencies for the two remote terminals to transmit on. They in turn tune their transmitters and receivers to the respective frequencies accordingly, and the circuit is established. The remote-to-remote communications is possible because the SCPC signals transmitted by these remotes are at lower bit rate than normally used for data communications.

9. DIGITAL VOICE AND TELEVISION FOR GEOSYNCHRONOUS SATELLITES

As just demonstrated by the example of one-hop voice communications between VSATs, low-bit-rate voice coders are instrumental for voice communications via geosynchronous satellites using small terminals. At the beginning of the digital conversion, 64-kbps pulse-coded modulation (PCM) and 56-kbps PCM were used by the INTELSAT system. In the late 1980s, 32-kbps adaptive differential PCM (ADPCM) combined with digital speech interpolation (DSI) was used to increase INTELSAT trunking capacity by a factor of four. DSI detects the silence periods of conversational speech and transmits only the active periods of the speech. By statistically aggregating a large number of circuits, DSI provides twofold capacity after taking into account the control overhead required.

Most of the early national and regional FSS networks use 64- or 56-kbps PCM. Newer networks have adopted the ITU G-728 16-kbps voice coder and the G.729 8-kbps voice coder. The lower bit rates not only increase the number of circuits carried by a satellite transponder, but also minimize the size of the remote terminal by reducing the transmit power needed to support a fixed number of voice circuits.

INMARSAT used low-bit-rate voice coders such as 16 and 9.6 kbps in their early digital terminals. It selected a 4-kbps voice coder for their smallest terminals in the early 1990s. A similar 4-kbps voice coder is also used by a geosynchronous mobile satellite system that supports handheld cell-phone-like terminals. In fact, with its low-gain antennas, handheld, high-quality, voice communications via satellite can only be made possible with the success of these very low bit rate voice coders.

Digital television signals based on ADPCM were demonstrated via satellites at 45 Mbps in the late 1970s. 1.5-Mbps and 384-kbps coders have been used for teleconference applications in the late 1980s. Not until digital direct satellite broadcast services were launched in the early 1990s, has digital television coding been widely used. Based on enhanced Motion Pictures Experts Group (MPEG) or MPEG-2 standards, these video coders are capable of compressing broadcast quality television signals to about 3–5 Mbps, depending on motion contents. The direct broadcast satellites typically time-multiplex up to 10 such signals into a single satellite transponder. Digital video compression is truly the underlining technology that makes satellite direct broadcast a commercial success.

10. ONBOARD PROCESSING FOR GEOSYNCHRONOUS SATELLITES

Advanced geosynchronous satellites use multibeam antennas to divide their coverage areas into cells like a cellular network to increase their traffic carrying capacity. Similar to cellular networks, the frequency band is reused by cell clusters. The adjacent cells within a cluster use different frequency bands or different polarization to provide the needed isolation in addition to the spatial discrimination provided by the antenna pattern. As the number of antenna beams increases, cross-connection between uplink beams and downlink beams becomes increasingly complicated. When the number of beams are more than a few tens on both the uplink and downlink, the task can no longer be handled with simple RF switches. Telephone switch technology has evolved since the 1970s from cross-bar to digital, and again from circuit-based "hard" switch

to packet-based "soft" switch. With each step of the evolution, the capacity of the switch has increased and the cost reduced by orders of magnitude. Beam cross-connection on the latest high-capacity geosynchronous satellites requires a digital switch solution.

Onboard processing satellites typically demodulate the uplink in each antenna beam into digital signals. Data packets are then FEC decoded, and put into an input buffer. Based on the header information in each packet, an onboard switch routes the packets to the output queue of each individual downlink beam. Header information may also provide traffic classification and quality of service (QoS) requirements. The onboard packet switch can also assign priority and exercise flow control in case of traffic congestion. For each downlink beam, data in the output queue are FEC-encoded, modulated, and upconverted to the RF frequency. They are then amplified and transmitted.

Onboard processing, in addition, offers the opportunity to optimize the uplink and downlink independently. Satellite transponders are most efficiently used when operated in TDM or TDMA mode at saturation. The size of ground terminals is minimized, however, if the satellite transponder is accessed via FDMA or CDMA. With onboard processing, both links can be operated with their most efficient access approach. When a large number of downlink beams are needed to cover an area with uneven traffic load, a common practice is to switch a much smaller number of active downlink transmitters to all the beams by phase-array technology. The dwell time on each individual beam can be directly proportional to the downlink traffic to its corresponding cell, thus optimizing the utilization of the satellite resources.

Also, the bit error rate (BER) or the packet error rate (PER) of the overall link equals to the sum of those incurred in the two links in a processing satellite, whereas the thermal noise and interference for the overall link is the sum of those incurred in the two links with a classic "bent pipe" satellite. Since the slope of BER or PER versus noise and interference is very steep when the link is protected by the FEC, the overall link can be much better optimized for an onboard processing satellite.

As a satellite system generally needs to accommodate different sizes of terminals, one of the challenges for cost-effective onboard processing is to implement the capability of dynamically assigning demodulator and FEC decoder resources for different-sized uplinks. This is accomplished with digital signal processing techniques that are scalable within limits.

11. SATELLITE COMMUNICATIONS SYSTEM EQUIPMENT

11.1. Gateway Architecture

There are several configurations of the satellite communications payload. It would be difficult to develop a generic architecture for the system since satellite payloads and Gateways are designed based on the traffic requirements it is meant to serve. In Fig. 1, we illustrate a simplified gateway architecture that supports digitized voice and data for TDMA application.

The gateway will transmit, via a satellite transponder, to multiple subscriber terminals (STs) on a single wideband TDM signal occupying one transponder bandwidth (e.g., 30 MHz.). This outbound signal will be in the 6- or 14-GHz bands for C-band or K band systems, respectively. Similarly, the gateway will receive FDM/TDMA bursts from multiple STs via another satellite transponder. As an example, the composite 30-MHz inbound signal could be 100 FDM channels, each 300 kHz wide, carrying 200 ksymbol/sec QPSK modulated carrier time-shared by multiple users in a TDMA protocol.

The gateway transmit traffic is converted to data blocks (slots), CRC and FEC encoded, and multiplexed into a TDM frame structure. The wideband TDM baseband signal is applied to a QPSK modulator, which converts it to a fixed IF. The IF is up converted (U/C) to C or K band and transmitted to the satellite as the outbound signal, via the RF Subsystem, which includes the power amplifier (HPA) and the antenna.

The gateway inbound FDM/TDMA signal at the C or K band, which is composed of transmission from multiple STs, is received by the same antenna, amplified by the LNA and downconverted (D/C) to IF. The IF is demodulated by the TDMA burst demodulators, each operating on a separate 300-kHz channel in this example. Each demodulator decodes its traffic slot by slot which is subsequently routed to appropriate destination by the TDMA Rx controller.

The gateway transmit power requirements are high since it has to transmit a wideband TDM signal. The high effective isotropic radiated power (EIRP) requirement is adjusted by suitable choice of HPA power rating and the antenna gain. Typically, gateway antenna beamwidths are narrow because of the HPA output power limitations, and thus a tracking antenna is required at most Gateways stations to track the satellite motion.

11.2. Satellite Communications Payload

Satellite communications payload is also very specific to applications. Conventional designs evolved from a simple bent pipe, global beam approach that acted as a simple repeater in the sky, to spot beam, regenerative repeaters with traffic switched between spots at baseband. There are far more advanced designs expected to operate in the near future. The fundamental objective is to support as much simultaneously active traffic as possible within the allocated bandwidth and the satellite power limitation.

In general, noninterfering spot beams and polarization are used to achieve frequency reuse. Spot beams with high gain, also provide higher EIRP that may illuminate areas with dense traffic, thus providing bandwidth reuse and power efficiency. Further, hopping spot beams, at the expense of complex control, provide further improvement in power efficiency. In this approach many spots are arranged over the coverage area, but the available transmit power is dynamically assigned only to a small sub-set of spots, at any given time. A mix of configurations is also possible.

In Fig. 2 we show a simple fixed spot beam payload system with two spot beams where traffic can be routed between the East and West spots by a RF switch matrix.

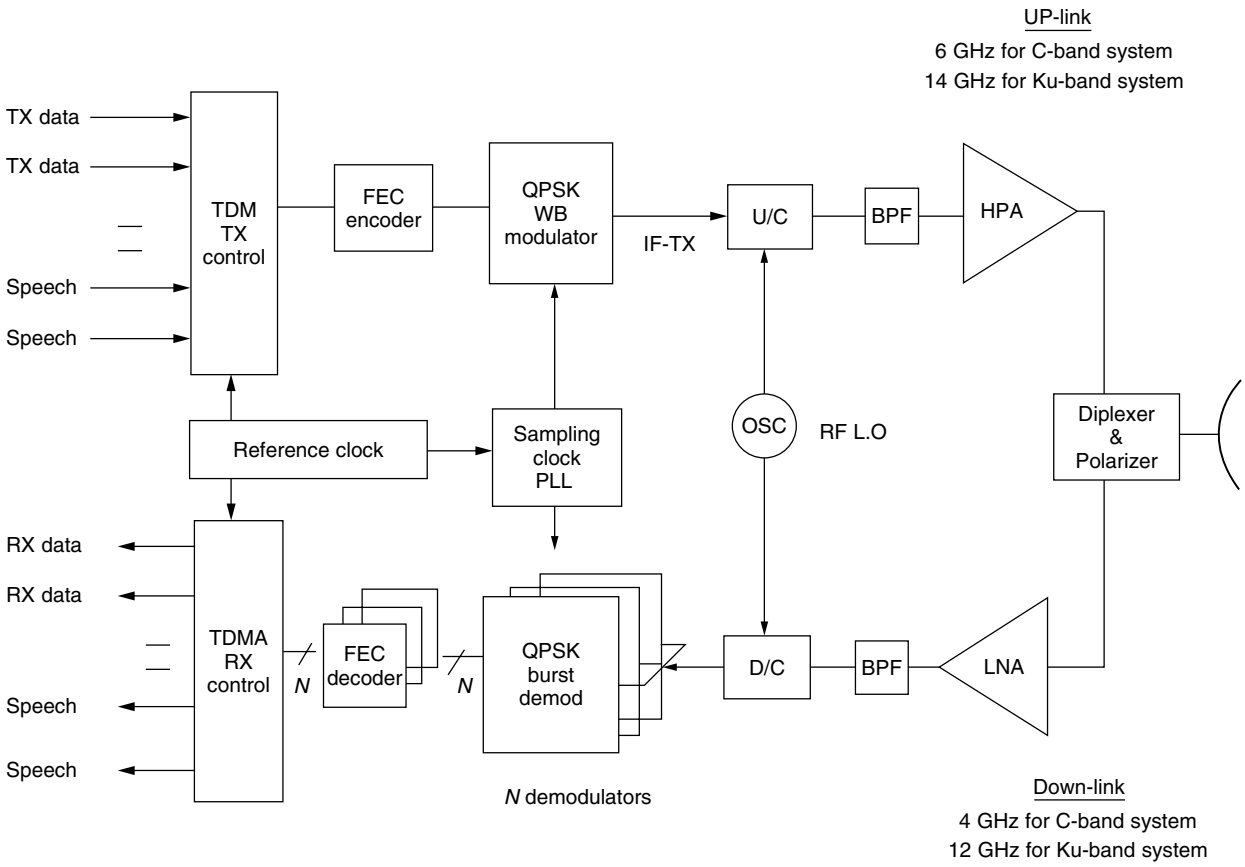


Figure 1. Simplified block diagram of a TDMA gateway.

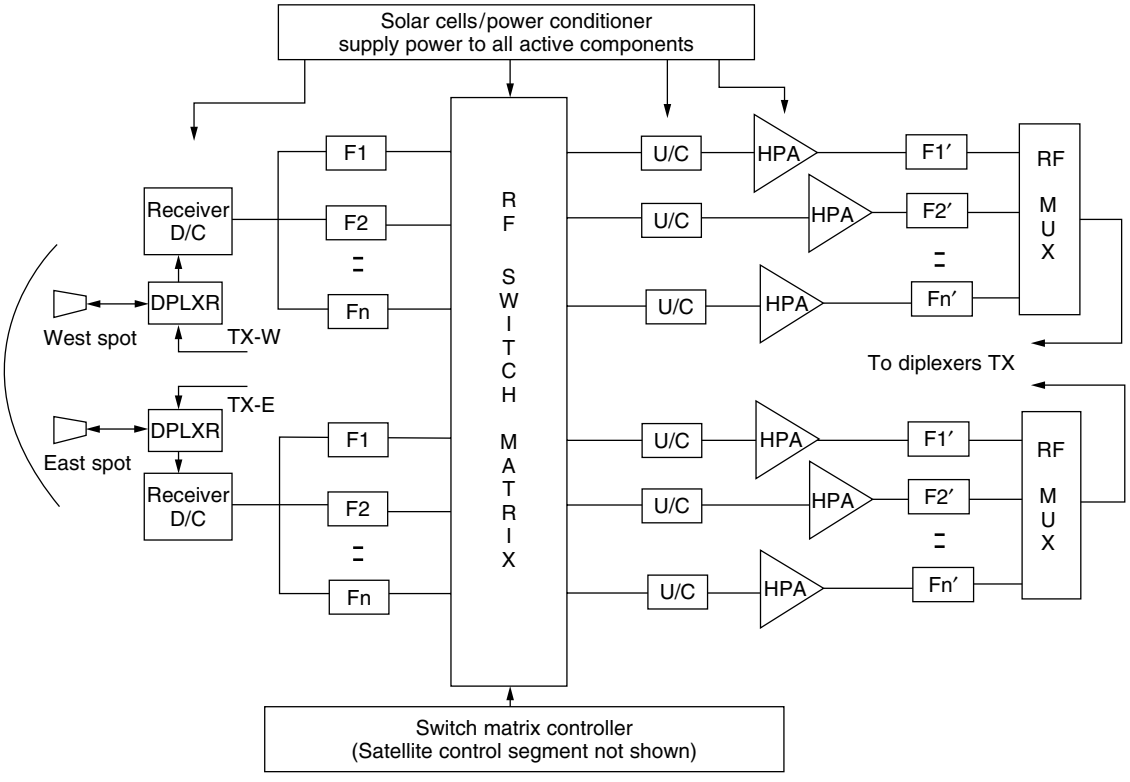


Figure 2. Simplified block diagram of communications payload.

The switch configuration is programmable by the control equipment. The satellite payload is controlled from ground by a dedicated command and control link with a control station.

Signals received from the ground are transmitted from either gateways or STs. The received signals are amplified by the LNA and demultiplexed by bandpass filters F. The bandwidth of these filters sets the transponder bandwidth. In our example F1 could be assigned to outbound transmission from a gateway, while F2 could be assigned to Inbound transmissions from a group of STs. The HPA for outbound may operate near saturation since there is only one (wideband TDMA) signal present. The HPA for Inbound traffic has to be backed off (typically output backoff of 4 dB) to support FDM/ TDMA carriers. The RF switch provides a simple return path to the same region or to a different region, based on the switch configuration. Typically this configuration would support traffic between hotspots on two ends of the coverage area for example New York and Los Angeles. Actual payload designs are more complicated than the example here. There could be a mix of CONUS beams and many spot beams, and a mix of C- and K-band transponders all interconnected via the RF switch matrix or matrices. It all depends on the system requirements and payload cost.

12. GEO SATELLITE SYSTEM LINK BUDGETS

As in any radio communications system design, link budgets are needed to size the system components to achieve performance objectives. For geosynchronous satellites, the uplink and downlink microwave path losses should be overcome by the antenna gain and transmit power to overcome the inherent noise and interference in the receivers. This analysis provides the basis of designing the hardware for the gateway and the ST, constrained by the cost and data rate capability tradeoffs. The following terminology is used for the various parameters of a link budget calculation.

Transmit Power. Actual power delivered to the antenna subsystem. If there are no losses between the HPA and the antenna, this will be the power output of the HPA.

Antenna Gain. The actual gain of the Tx (transmitting) antenna, i.e., directivity minus losses from illumination efficiency and other reflector and/or radome losses. The gain of a parabolic reflector fed by a feed-horn is given by $G = 4\pi \mathrm{A}\eta/\lambda^2$, where A is the planar area of the aperture, λ is the wavelength, and η is the aperture efficiency of the antenna (typically 65%).

EIRP. Effective isotropic radiated power is the product of the antenna gain and transmit power (or the sum if these are in decibels).

Edge-of-Beam Loss. The satellite antenna boresight points to one location on earth. The coverage area is defined by constant power contours decreasing in power as one moves away from the boresight. If the edge of a satellite antenna beam is defined by −3 dB contour, then this loss will be 3 dB.

Antenna Pointing Loss. The ground equipment antenna boresight may not point exactly at the satellite. To account for this pointing error, typically 0.5 dB pointing loss is assumed.

Free-Space Path Loss. The square-law propagation loss factor normalized with the wavelength. This definition simplifies calculation of received carrier power based on transmit EIRP propagation loss and receive antenna gain. Loss $= (4\pi d/\lambda)^2$, where d is the distance to the satellite, a function of elevation angle. A value of 40,000 km is typically used.

Atmospheric Loss. The atmospheric loss is due to the gaseous content of the atmosphere and higher layers, encountered during wave propagation even in clear-sky conditions. Typical values range from 0.5 (C band) to 0.8 dB (Ka band).

Receive Flux Density. The receive flux density at the satellite is a measure of received power per unit area assuming isotropic radiation from the transmit antenna. It can be obtained from the EIRP and the distance d (after accounting for other fixed losses such as pointing error, edge-of-beam loss, and atmospheric loss but excluding free-space path loss).

$$\text{Rx flux density}\left[\frac{(\text{dBW})}{\text{m}^2}\right] = 10\log\frac{\text{EIRP}}{4\pi d^2} - \text{losses (dB)}$$

Typically the satellite repeater function is specified in terms of the Rx flux density and transmit EIRP at saturation. Actual EIRP is computed by first computing the incident flux density, hence, the backoff from the saturation point, then finding the output backoff from saturation EIRP, based on the power amplifier nonlinear gain characteristics.

Receiver Noise Figure. The receiver noise figure (NF) is a measure of the thermal noise created in the receive chain referenced to the antenna port (LNA input). In satellite applications this can be assumed to come entirely from the LNA. Gateway LNA noise figures are very low (<1 dB), while satellite LNA NF could be >3 dB.

Equivalent Noise Temperature. The equivalent noise temperature (T_e), a more practical parameter, is directly related to the NF by the relationship $T_e = (\text{NF} - 1)T_0$, where T_0 is ambient temperature (290 K is assumed for room temp).

Antenna Noise Temperature. The antenna noise temperature (T_a) is the noise picked up by the antenna environment, which includes galactic, atmospheric and spillover noise. For a gateway looking at the satellite, this parameter is only 40–60 K. For the satellite looking at the earth, this noise temperature could be >500 K.

System Noise Temperature (T_s). This is the total noise temperature at antenna port: $T_s = T_e + T_a$.

Typically, system noise temperature of satellites is around 1000 K.

Receive G/T. This ratio establishes the figure of merit for the receiver subsystem. It is the ratio of Rx antenna gain to the system noise temperature.

Received C/N_0. The carrier-to-noise density ratio is computed directly from the G/T and EIRP after accounting for all the losses. It is independent of the bit rate. (In fact, it gives an easy way to determine what bit rates can be supported by the system given the threshold E_b/N_0).

$$C/N_0(\text{dB}-\text{Hz}) = \text{EIRP (dBW)} + G/T\ (\text{dB/K}) - \text{losses (dB)} + 228.6\ \text{dB/K/Hz}$$

(Boltzmann)

In a bent-pipe repeater model the uplink and downlink system noise temperatures should be accounted for using the relationship

$$\left(\frac{C}{N_0}\right)^{-1} = \left(\frac{C}{N_{\text{ou}}}\right)^{-1} + \left(\frac{C}{N_{\text{od}}}\right)^{-1}$$

Received C/I_0. The carrier-to-interference density is computed by finding the ratio of the received carrier power to the interference density. The interference density is obtained by the interference power in the channel divided by the channel bandwidth. In a bent-pipe repeater model the uplink and downlink interference should be accounted for by a similar relationship as given above:

$$\left(\frac{C}{I_0}\right)^{-1} = \left(\frac{C}{I_{\text{ou}}}\right)^{-1} + \left(\frac{C}{I_{\text{od}}}\right)^{-1}$$

Received $E_{\text{bi}}/(N_0+I_0)$. The $E_{\text{bi}}/(N_0+I_0)$ is the net information bit energy-to-noise plus interference density received by the demodulator. This includes the effects of thermal noise and interference, and for bent-pipe satellites, includes effects of the uplink and the downlink. This quantity should be compared with the threshold E_{bi}/N_0 to determine the link margin:

$$\frac{E_{\text{bi}}}{N_0+I_0}\ \text{dB} = \frac{C}{N_0+I_0}\ \text{dB} - 10\log \times (\text{information bit rate})$$

where the quantity $(C/(N_0+I_0))^{-1} = (C/N_0)^{-1} + (C/I_0)^{-1}$, to account for system noise temperature and interference.

E_b/N_0 Threshold. The modem is guaranteed to deliver a target information bit error rate at this threshold. Usually this threshold is derived by system simulation E_b/N_0 and ideal modem assumption, based on a chosen FEC approach and then a modem implementation margin is added.

Available Fade Margin. Typically the link budget is set up for clear-sky conditions. The difference between the received $E_{\text{bi}}/(N_0+I_0)$ and the threshold E_{bi}/N_0 gives the fade margin. The fade margin reflects the availability of the system during rain fades. System availability as a function of rain statistics is beyond the scope of this paper. For the C band, a 1.5 dB margin is adequate to achieve reasonable availability for typical locations.

12.1. Example of a Link Budget

A hypothetical system is chosen as an example to be used with an INTELSAT standard B, C-band (6/4-GHz) gateway (GW) earth station. The outbound is a 20-Mbps TDM, transmitted with antenna gain of 58 dBi and HPA power 200 W. The G/T of the GW is 33.5 dB/K.

The satellite is a bent-pipe model with saturation flux density of -83 dBW/m^2 corresponding to an EIRP of 37 dBW, giving a gain of 120 dB/m^2 at saturation or 125 dB/m^2 in linear region, based on a standard TWTA nonlinear transfer characteristics. The system noise temperature is 1000 K. The G/T is -6.0 dB, based on a gain of 24 dBi, which gives an 8° half-power CONUS beam. The inbound traffic is FDM/TDMA with 250 kbps information per channel, 120 channels spaced 300 khz apart using one transponder. The subscriber terminal (ST) contains a 2-W power amplifier and 1.8 m dish antenna. Its receiver noise figure is 1.5 dB.

Threshold E_b/N_0 of 3.0 dB (including 1 dB modem implementation loss) and fade margin of 1.5 dB are assumed in both directions. Other parameters are in the link budgets shown in Tables 1 and 2 for outbound and inbound time-division multiplexing, respectively.

BIOGRAPHY

Lin-Nan Lee is vice president of engineering at Hughes Network Systems (HNS), Germantown, Maryland, responsible for advanced technology development. Dr. Lee received his B.S. degree from National Taiwan University and his M.S. and Ph.D. from University of Notre Dame, Indiana, all in electrical engineering. He started his career at Linkabit Corporation, where he was a senior scientist working on packet communications over satellites at the dawn of the Internet age. He then worked at Communication Satellite Corporation (COMSAT) in various research and development capacities with emphasis on source and channel coding technology development for satellite transmission and eventually assumed the position of chief scientist, COMSAT systems division. After joining HNS in late 1992, he has contributed to HNS' effort in wireless and satellite communications areas. Dr. Lee is a fellow of IEEE. He was the corecipient of the COMSAT Exceptional Invention Award, and the 1985 and 1988 COMSAT Research Award. He has authored or coauthored more than 20 U.S. patents.

Khalid Karimullah received his Ph.D in electrical engineering from Michigan State University, East Lansing, Michigan in 1980. He started his professional career at

Table 1. Outbound Link Budget for 20-Mbps TDM

	Units	Clear Sky
6-GHz Uplink		
Outbound bit rate	Mbps	20.0
Gateway antenna gain	dB	58.0
HPA Tx power	dBW	23.0
Transmit EIRP	dBW	81.0
Uplink frequency	GHz	6.0
Wavelength	m	0.05
Distance to satellite (max)	km	40,000
Free-space path loss at 6 GHz	dB	200
Atmospheric loss	dB	0.5
Edge-of-beam loss	dB	3.0
Gateway antenna pointing loss	dB	0.5
Satellite G/T ($G = 24$ dB; $T_s = 1000$ K)	dB/K	−6.0
Carrier noise density $(C/N_0)_u$	dB−Hz	99.6
Rx flux density (saturation = −83)	(dBW)/m^2	−86
4-GHz Downlink		
Satellite EIRP (0.5 dB backoff)	dBW	36.5
Downlink Frequency	GHz	4.0
Wavelength	m	0.075
Free-space path loss at 4 GHz	dB	196.5
Atmospheric loss	dB	0.5
Edge-of-beam loss	dB	3.0
ST antenna diameter	m	1.8
ST antenna Rx gain	dB	36.0
ST antenna pointing loss	dB	0.5
ST noise figure	dB	1.5
Equivalent noise temp (NF = 1.5 dB)	K	119.6
Antenna noise temperature	K	52.0
System noise temperature	dBK	22.3
ST G/T	dB/K	13.7
Carrier noise density $(C/N_0)_d$	dB−Hz	78.2
Outbound Overall		
(C/N_0) received	dB−Hz	78.2
(C/I_0) assumed	dB−Hz	86.0
$C/(N_0 + I_0)$	dB−Hz	77.5
$E_b/(N_0 + I_0)$	dB	4.5
Threshold E_b/N_0	dB	3.0
Fade margin	dB	1.5

COMSAT Laboratories, Maryland, where he worked on regenerative satellite transponder modem designs. He later joined MA-Com Linkabit, San Diego, California, in April 1987, where he developed his expertise in the areas of communications and signal processing. He joined Hughes Network Systems, Germantown, Maryland, in 1989 and has since worked on CDMA technology. Currently, he works at HNS as a senior Director, engineering, involved in R.F. and CDMA technology related activities. He has been active in the TIA/EIA TR45.5 cdma2000 standards development, participating in the physical layer and enhanced access procedures development. He chaired the TR45.5 cdma2000 enhanced access ad-hoc group. He has authored/ coauthored several patents in the CDMA physical layer and enhanced access procedures. In 1998, he received Hughes Electronics Patent Award and was the corecipient of the 1998 CDMA Technical Achievement Award.

Table 2. Inbound Link Budget for 250-kbps/s FDMA/TDMA (120 Channels)

	Units	Clear Sky
6 GHz Uplink		
Inbound bit rate	Mbps	0.25
ST Tx antenna gain	dB	39.5
ST SSPA Tx power	dBW	3.0
Transmit EIRP	dBW	42.5
Uplink frequency	GHz	6.0
Wavelength	m	0.05
Distance to satellite	km	40,000
Free-space path loss at 6 GHz	dB	200
Atmospheric loss	dB	0.5
Edge-of-beam loss	dB	3.0
ST antenna pointing loss	dB	0.5
Satellite G/T ($G = 24$ dB; $T_s = 1000$ K)	dB/K	−6.0
Carrier noise density $(C/N_0)_u$	dB−Hz	61.1
Rx flux density (saturation = −83)	(dBW)/m^2	−124.5
4 GHz Downlink		
Downlink EIRP/ ST (linear)	dBW	0.5
Downlink frequency	GHz	4.0
Wavelength	m	0.075
Free-space path loss at 4 GHz	dB	196.5
Atmospheric loss	dB	0.5
Edge-of-beam loss	dB	3.0
Gateway antenna diameter	m	15.2
Gateway antenna Rx gain	dB	54.5
GW antenna pointing loss	dB	0.5
GW noise figure	dB	1.0
Equivalent noise temperature (NF = 1.0 dB)	K	75.1
Antenna noise temperature	K	52.0
System noise temperature	dBK	21.0
GW G/T	dB/K	33.5
Carrier noise Density $(C/N_0)_d$	dB−Hz	62.0
Inbound Overall		
(C/N_0) Received	dB−Hz	58.5
(C/I_0) Assumed	dB−Hz	86.0
$C/(N_0 + I_0)$	dB−Hz	58.5
$E_b/(N_0 + I_0)$	dB	4.5
Threshold E_b/N_0	dB	3.0
Fade margin	dB	1.5

GOLAY CODES

MARCUS GREFERATH
San Diego State University
San Diego, California

1. INTRODUCTION

Among the various codes and code families that have been enjoying the attention of coding theorists, two sporadic examples of (extended) cyclic codes have continuously attracted the interest of many scholars. These two codes, named after their discoverer M. Golay, have been known

since the very first days of algebraic coding theory in the late 1940s. Their enduring role in contemporary coding theory results from their simplicity, structural beauty, depth of mathematical background, and connection to other fields of discrete mathematics, such as finite geometry and the theory of lattices.

This presentation is devoted to a description of the basics about these codes. The reader should be aware that the issues that we have chosen for a more detailed discussion form only a narrow selection out of the material that is available about these codes. So the article at hand does not claim in any way to be a comprehensive or complete treatment of the issue. For the many aspects that we are not discussing here, the reader is referred to the Bibliography, and in particular to the treatise by Conway and Sloane [6]. This article is organized as follows. In Section 1 we will introduce the Golay codes as (extended) quadratic residue codes. We will discuss their parameters and immediate properties. In Section 2 we will address alternative descriptions of the Golay codes. Besides a collection of nice generator matrices, we will briefly mention the miracle octad generator (MOG) construction and also Pasquier's description of the binary Golay code. Section 3 is devoted to the decoding of these codes. Among the various algorithms from the literature we have picked a particular one due to M. Elia, which is an algebraic decoder.

We will then briefly come to the relationship of these codes to finite geometry (design theory) and the theory of lattices. In particular, we will mention an important lattice called the *Leech lattice*, which closely connected to the binary Golay code, more precisely, with a $\mathbb{Z}_4$-linear version of this code. This gives us a natural bridge to the final discussion of the article, the role of the Golay codes in the investigation of ring-linear codes.

1.1. Basic Notions

We define the Hamming distance d_H on a finite set $\mathbb{F}$ (usually a field or a ring) defined by

$$d_H\colon \mathbb{F}\times\mathbb{F}\to\mathbb{N},\quad (x,y)\mapsto \begin{cases} 0, & x=y \\ 1, & \text{else} \end{cases}$$

This function is usually extended additively to the space $\mathbb{F}^n$ for every $n\in\mathbb{N}$. A block code is a subset C of $\mathbb{F}^n$, and its Hamming minimum distance is

$$d_{\min}(C) := \{d_H(x,y) \mid x,y\in C, x\neq y\}$$

A block code of length n and minimum distance d is referred to as (n,M,d) code, where M is the number of its words.

If $\mathbb{F}$ is a (finite) field, then we call a code *C-linear* if it is a subspace of F^n. A linear code of dimension k is usually denoted as an $[n,k]$ code. If it has minimum distance d, we also speak of an $[n,k,d]$ code. Note that for a linear code C, the minimum Hamming weight, namely, $\min\{d_H(c,0) \mid c\in C-\{0\}\}$, coincides with its minimum distance.

A linear code $C\leq\mathbb{F}^n$ is called *cyclic* if is invariant under a cyclic shift of its coordinates. Algebraically, cyclic linear codes are exactly the ideals of the (residual) polynomial ring $\mathbb{F}[x]/(x^n-1)$, and for a cyclic code C there exists a unique monic divisor g of x^n-1 in $\mathbb{F}[x]$ such that $C=\mathbb{F}[x]g/(x^n-1)$.

Given a linear code C, we define the dual code

$$C^\perp := \{x\in\mathbb{F}^n \mid c\cdot x = 0 \text{ for all } c\in C\}$$

and we call C *self-dual*, if $C=C^\perp$.

2. THE GOLAY CODES AS (EXTENDED) QUADRATIC RESIDUE CODES

M. J. E. Golay (1902–1989) was a Swiss physicist known for his work in infrared spectroscopy, among other things. He was one of the founding fathers of coding theory, discovering the two binary Golay codes in 1949 [11] and the ternary Golay codes in 1954. These codes, which have been called *Golay codes* since then, have been important not only because of theoretical but also practical reasons — the extended binary Golay code, for example, has frequently been applied in the U.S. space program, most notably with the *Voyager I and II* spacecraft that transmitted clear color pictures of Jupiter and Saturn.

Let p and q be prime numbers such that p is odd and q is a quadratic residue modulo p, and let ω be a primitive pth root in a suitable extension field of $\mathbb{F}_q$. Let $\mathrm{QR} := \{i^2 \mid i\in\mathbb{F}_p^\times\}$ denote the set of quadratic residues modulo p and set $\mathrm{NQR} := \mathbb{F}_p^\times \setminus \mathrm{QR}$. Both of these sets have size $(p-1)/2$, the polynomials

$$f_{\mathrm{QR}} = \prod_{i\in\mathrm{QR}}(x-\omega^i) \text{ and } f_{\mathrm{NQR}} = \prod_{i\in\mathrm{NQR}}(x-\omega^i)$$

have coefficients in $\mathbb{F}_q$, and $x^p-1 = f_{\mathrm{QR}}\cdot f_{\mathrm{NQR}}\cdot(x-1)$.

Example 1

(a) For $p=23$ and $q=2$ and choosing a suitable primitive 23rd root in $\mathbb{F}_{2^{11}}$, we obtain

$$f_{\mathrm{QR}} = x^{11}+x^9+x^7+x^6+x^5+x+1 \text{ and}$$
$$f_{\mathrm{NQR}} = x^{11}+x^{10}+x^6+x^5+x^4+x^2+1.$$

(b) For $p=11$ and $q=3$ and a suitable primitive 11th root in $\mathbb{F}_{3^5}$, we obtain

$$f_{\mathrm{QR}} = x^5+x^4-x^3+x^2-1 \text{ and}$$
$$f_{\mathrm{NQR}} = x^5-x^3+x^2-x-1$$

Definition 1

(a) The cyclic binary code of length 23 generated by the polynomial $x^{11}+x^9+x^7+x^6+x^5+x+1\in\mathbb{F}_2[x]$ is called the *binary Golay code*. This code is a [23,12,7] code; its extension by a parity check will be denoted by $\overline{G}_2$ and is a [24,12,8] code.

(b) The cyclic ternary code G_3 of length 11 generated by the polynomial $x^5+x^4-x^3+x^2-1\in\mathbb{F}_3[x]$ is called the ternary Golay code. This code is a [11,6,5] code;

its extension by a parity check will be denoted by $\overline{G}_3$ and is a [12,6,6] code.

The sphere packing bound says that if $|F| = q$ and $d_{\min}(C) = 2t + 1$, then

$$|C| \sum_{i=0}^{t} \binom{n}{i} (q-1)^i \leq q^n$$

Codes meeting this bound with equality are called *perfect codes*. Given their minimum distance, it is indeed easily verified that G_2 as well as G_3 are perfect codes, since

$$2^{12}\left(\binom{23}{0} + \binom{23}{1} + \binom{23}{2} + \binom{23}{3}\right) = 2^{23} \text{ and}$$

$$3^6\left(\binom{11}{0} + \binom{11}{1}2 + \binom{11}{2}4\right) = 3^{11}$$

The codes $\overline{G}_2$ and $\overline{G}_3$ are, of course, not perfect, but they are still what is called *quasiperfect*: the spheres of radius $t = 3$ (or $t = 2$, respectively) are still disjoint, whereas the spheres of radius $t + 1$ already cover the ambient space.

2.1. Weight Enumerators

Let C be a linear $[n, k, d]$ code. Defining the Hamming weight enumerator of C as the integer polynomial

$$W_C(t) := \sum_{c \in C} t^{w_H(c)} = \sum_{i=0}^{n} A_i t^i$$

where $A_i := |\{c \in C \mid w_H(c) = i\}|$, we obtain the Hamming weight enumerators for the four Golay codes as listed in Table 1.

Call a linear code C *self-dual*, if it coincides with its dual code

$$C^\perp := \{x \in \mathbb{F}^n \mid c \cdot x = 0 \text{ for all } c \in C\}$$

It can be shown that the codes $\overline{G}_2$ and $\overline{G}_3$ are self-dual codes. Moreover $\overline{G}_2$ is called a doubly even code because all of its weights are divisible by 4. Note that all weights of $\overline{G}_3$ are divisible by 3.

2.2. Uniqueness of the Golay codes

Two block codes C and D of length n over the alphabet $\mathbb{F}$ are called *equivalent* if there is a coordinate permutation π and a set $\sigma_1, \ldots, \sigma_n$ of permutation on $\mathbb{F}$ such that C is mapped to D under the bijection

$$\mathbb{F}^n \to \mathbb{F}^n, (c_1, \ldots, c_n) \mapsto (\sigma_1(c_\pi(1)), \ldots, \sigma_n(c_\pi(n)))$$

It has been proved by Pless, Goethals, and Delsarte that the Golay codes are uniquely determined up to equivalence by their parameters. This is the content of the following theorem. Proofs can be found in the literature [8,21].

Table 1. Weight Enumerators for the Golay Codes

Code	Weight Enumerator
G_2	$t^{23} + 253\,t^{16} + 506\,t^{15} + 1288\,t^{12} + 1288\,t^{11} + 506\,t^8 + 253\,t^7 + 1$
$\overline{G}_2$	$t^{24} + 759\,t^{16} + 2576\,t^{12} + 759\,t^8 + 1$
G_3	$t^{11} + 132\,t^6 + 132\,t^5 + 330\,t^3 + 110\,t^2 + 24$
$\overline{G}_3$	$t^{12} + 264\,t^6 + 440\,t^3 + 24$

Theorem 1

(a) Every binary $(23, 2^{12}, 7)$ code is equivalent to the Golay code G_2, and every binary $(24, 2^{12}, 8)$ code is equivalent to $\overline{G}_2$.
(b) Every ternary $(11, 3^6, 5)$ code is equivalent to the Golay code G_3, and every ternary $(12, 3^6, 6)$ code is equivalent to $\overline{G}_3$.

3. ALTERNATIVE CONSTRUCTIONS

In this section we will discuss alternative constructions of the Golay codes. First, we will find generator matrices of rather beautiful form for equivalent versions of these.

In the binary case consider the matrix

$$[I \mid A] := \left[\begin{array}{cccccccccccc|cccccccccccc}
1&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&1\\
0&1&0&0&0&0&0&0&0&0&0&0&1&1&1&0&1&1&1&0&0&0&1&0\\
0&0&1&0&0&0&0&0&0&0&0&0&1&1&0&1&1&1&0&0&0&1&0&1\\
0&0&0&1&0&0&0&0&0&0&0&0&1&0&1&1&1&0&0&0&1&0&1&1\\
0&0&0&0&1&0&0&0&0&0&0&0&1&1&1&1&0&0&0&1&0&1&1&0\\
0&0&0&0&0&1&0&0&0&0&0&0&1&1&1&0&0&0&1&0&1&1&0&1\\
0&0&0&0&0&0&1&0&0&0&0&0&1&1&0&0&0&1&0&1&1&0&1&1\\
0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&1&0&1&1&0&1&1&1\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&1&0&1&1&0&1&1&1&0\\
0&0&0&0&0&0&0&0&0&1&0&0&1&0&1&0&1&1&0&1&1&1&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&1&1&0&1&1&0&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&1&0&1&1&0&1&1&1&0&0&0&1
\end{array}\right]$$

Then consider the ternary matrix

$$[I \mid B] = \left[\begin{array}{cccccc|cccccc}
1&0&0&0&0&0&0&1&1&1&1&1\\
0&1&0&0&0&0&1&0&1&2&2&1\\
0&0&1&0&0&0&1&1&0&1&2&2\\
0&0&0&1&0&0&1&2&1&0&1&2\\
0&0&0&0&1&0&1&2&2&1&0&1\\
0&0&0&0&0&1&1&1&2&2&1&0
\end{array}\right]$$

These matrices have the following properties.

1. Both A and B are symmetric and satisfy $A^2 = I$ and $B^2 = -I$.
2. Each row of A has exactly 7 or exactly 11 ones. Each row of B has exactly 5 nonzero entries.
3. Each pair of rows of A differs in exactly 6 places. For B the sum and the difference of each pair of rows have at least 4 nonzero entries.

It can be seen that the code generated by $[I \mid A]$ has the parameters of the extended binary Golay code, and that the code generated by $[I \mid B]$ has those of the ternary Golay code. By the uniqueness theorem (Theorem 1, above) we

therefore see that they are (up to equivalence) the binary and ternary Golay codes.

3.1. The MOG Construction

We now discuss a construction of the binary Golay code using the $\mathbb{F}_4$-linear hexacode [6, Chap. 11].

The *hexacode* is the [6,3,4] code H generated by the matrix

$$\begin{bmatrix} 1 & 0 & 0 & \omega^2 & 1 & \omega \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & \omega & 1 & \omega^2 \end{bmatrix}$$

Its Hamming weight enumerator is given by $1 + 45t^4 + 18t^6$, and it is clear that any word of the hexacode has 0, 2, or 6 zeros.

Let $\mathbb{F}_2^{4\times 6}$ denote the space of all 4×6 matrices with binary entries. As an $\mathbb{F}_2$-vector space it can clearly be identified with $\mathbb{F}_2^{24}$. We are going to isolate a subset of $\mathbb{F}_2^{4\times 6}$ that will in a natural way turn out to be equivalent to the Golay code.

For this we first define a mapping $\varphi\colon \mathbb{F}_2^{4\times 6} \to \mathbb{F}_4^6, G \mapsto (0, 1, \omega, \omega^2)G$. Now define C to be the subset of all matrices $G \in \mathbb{F}_2^{4\times 6}$ having the following two properties:

1. For every column $j \in \{1, \ldots, 6\}$ there holds $\sum_{i=1}^{4} g_{ij} = \sum_{j=1}^{6} g_{1j}$; thus, the parity of each column is that of the first row of G.
2. $\varphi(G) \in H$.

As both of these conditions are preserved under addition of matrices that satisfy these conditions, we have C to be a subspace of $\mathbb{F}_2^{4\times 6}$. The first condition imposes 6 restrictions on the matrix, and the second (at most) another 6. For this reason $\dim(C) \geq 24 - (6 + 6) = 12$. To find that this code is equivalent to the Golay code, we only have to check if its minimum weight is given by at least 8, because this forces its dimension down to 12 and hence makes it a [24,12,8] code, and we are finished by Theorem 1.

Let G be a nonzero element of C. First, assume that $\sum_{j=1}^{6} g_{1j} = 0$, which by the preceding description means that $\sum_{i=1}^{4} g_{ij} = 0$ for all $j \in \{1, \ldots, 6\}$. If $\varphi(G) \neq 0$, then $\varphi(G)$ has at least 4 nonzero entries, and hence G must have at least 4 nonzero columns. As each of these columns has an even number of ones, the weight of G is at least 8. If $\varphi(G) = 0$ then each column of G must be the all-zero or the all-one vector because $0, 1, \omega$ and ω^2 can be combined to zero under even parity only in these two ways. Hence, again the weight of G is at least 8 because we have an even number of nonzero columns.

Now assume that the parity of the first row of G is 1. Then each column of G has either 1 or 3 nonzero entries. Its weight will therefore be at least 8 unless each column has exactly 1 nonzero entry. Exactly the zero-entries of $\varphi(G)$ correspond to those columns of G having their nonzero entry in the first row. As $\varphi(G) \in H$ we conclude that there is an even number of zero entries in $\varphi(G)$ contradicting the fact that the parity of the first row of G is 1. Hence, this case cannot happen, and we know that the weight of G must be at least 8.

3.2. Pasquier's Construction

Pasquier observed [20] that the extended binary Golay code can be obtained from a Reed-Solomon code over $\mathbb{F}_8$. In order to see this let

$$\mathrm{tr}\colon \mathbb{F}_8 \to \mathbb{F}_2, x \mapsto x + x^2 + x^4$$

be the trace map and let $\alpha \in \mathbb{F}_8$ be an element satisfying the equation $\alpha^3 + \alpha^2 + 1 = 0$. It is easy to see that $\mathrm{tr}(\alpha) = 1$, and that $B := \{\alpha, \alpha^2, \alpha^4\}$ forms a trace-orthogonal basis of $\mathbb{F}_8$ over $\mathbb{F}_2$. This means that

$$\mathrm{tr}(xy) = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases}$$

for all $x, y \in B$.

Now consider the [7,4,4] Reed–Solomon code generated by the polynomial $\Pi_{i=1}^{3}(x + \alpha^i)$. Its extension by a parity check yields the self-dual [8,4,5] extended Reed–Solomon code generated by the matrix

$$R := \begin{bmatrix} 1 & \alpha^5 & 1 & \alpha^6 & 0 & 0 & 0 & \alpha^3 \\ 0 & 1 & \alpha^5 & 1 & \alpha^6 & 0 & 0 & \alpha^3 \\ 0 & 0 & 1 & \alpha^5 & 1 & \alpha^6 & 0 & \alpha^3 \\ 0 & 0 & 0 & 1 & \alpha^5 & 1 & \alpha^6 & \alpha^3 \end{bmatrix}$$

We now consider the $\mathbb{F}_2$-linear mapping

$$\mathbb{F}_8 \to \mathbb{F}_2^3, \quad a_0\alpha + a_2\alpha^2 + a_1\alpha^4 \mapsto (a_0, a_1, a_2)$$

and extend it componentwise to an $\mathbb{F}_2$-linear mapping $\varphi\colon \mathbb{F}_8^8 \to \mathbb{F}_2^{24}$. The image of the extended Reed–Solomon code described above is clearly a binary [24,12] code, because the vectors $\{\alpha v, \alpha^2 v, \alpha^4 v\}$ are independent over $\mathbb{F}_2$ for each of the rows v of the preceding matrix R. Hence this binary image is a [24,12] code.

Observing $\mathrm{tr}(xy) = \varphi(x)\varphi(x)$ for all vectors $x, y \in \mathbb{F}_8^8$, we see that our [24,12] code is self-dual because the Reed–Solomon code that we started with is so. Even more is true: this code is "doubly even," which means that the Hamming weight of its vectors is always a multiple of 4. We can easily check this by finding a basis consisting of doubly even words and keeping in mind that under self-duality this property inherits to linear combinations of vectors that have this property. Hence the minimum weight of the above code is a multiple of 4. However the underlying Reed–Solomon code has already minimum weight 5, and so our code must have minimum weight ≥ 8. Finally we can apply Theorem 1 and find that it is the binary Golay code.

Remark 1. It is worth noting that Goldberg [12] has constructed the ternary Golay code as an image of an $\mathbb{F}_9$-linear [6,3,4] code in a similar manner. Again an obvious mapping between $\mathbb{F}_9^6$ and $\mathbb{F}_3^{12}$ is used to map the [6,3,4] code

into a [12,6] code, which finally turns out to be equivalent to the ternary Golay code.

4. DECODING THE GOLAY CODES

There are various decoders for the (cyclic) binary and ternary Golay codes; the most common are probably that by Kasami and the systematic search decoder (both have been nicely described in Ref. 16). Focusing on the binary Golay code G_2, we prefer to discuss an algebraic decoder that has been developed by M. Elia [9]. Even though this decoder is not the fastest known, it is of interest because it makes use of the algebraic structure of the code in question. Furthermore in a modified form it has been used in order to decode ring-linear versions of the Golay code [14].

4.1. Decoding the Binary [23,12,7] Code

Let $\alpha \in \mathbb{F}_{2^{11}}$ be a root of the generator polynomial $g = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$ of G_2. It clearly satisfies $\alpha^{23} = 1$, and its associated cyclotomic coset is given by $B = \{1, 2, 4, 8, 16, 9, 18, 13, 3, 6, 12\}$. This shows that g has roots α, α^3 and α^9.

As we are dealing with a cyclic code, we will represent its words by polynomials in the sequel. Assume the word $r = fg + e$ has been received, where the Hamming weight of e is at most 3. We compute the syndromes

$$s_1 := r(\alpha) = e(\alpha), \quad s_3 := r(\alpha^3) = e(\alpha^3), \text{ and}$$
$$s_9 := r(\alpha^9) = e(\alpha^9)$$

and our plan is to recover e and hence fg and f from these. This can be done in a quite simple way. As we are in a binary situation, we are interested only in what is called the *error locator polynomial*

$$L(z) := \prod\{(z + \alpha^i) \mid i \in \{0, \ldots, 22\}, e_i \neq 0\}$$

Once we can express this in terms of the syndromes described above, we only have to find its roots, and we know the error locations. Let us distinguish the following four cases:

1. There is no error, which means that $e = 0$. Then $L(z) = 1$.
2. There is one error in the position i, which means that $L(z) := z + \sigma_1$ where we have set $\sigma_1 = \alpha^i$.
3. There are two errors in position i and j. Then $L(z) = z^2 + \sigma_1 z + \sigma_2$ where
$$\sigma_1 = \alpha^i + \alpha^j \text{ and } \sigma_2 = \alpha^i \alpha^j$$
4. There are three errors in position i, j and k. Then $L(z) = z^3 + \sigma_1 z^2 + \sigma_2 z + \sigma_3$, where
$$\sigma_1 = \alpha^i + \alpha^j + \alpha^k$$
$$\sigma_2 = \alpha^i \alpha^j + \alpha_j \alpha^k + \alpha^i \alpha^k$$
$$\sigma_3 = \alpha^i \alpha^j \alpha^k$$

Case 1 is easily recognized by the fact that here $s_1 = s_3 = s_9 = 0$. Case 2 occurs exactly if $s_1^3 = s_3$ and $s_3^3 = s_9$, which is easily verified considering the above mentioned definitions. For the remaining cases we observe first that $s_1^3 \neq s_3$. Furthermore we still have $\sigma_1 = s_1$ by definition. Setting

$$D = (s_1^3 + s_3)^2 + \frac{s_1^9 + s_9}{s_1^3 + s_3}$$

and verifying that $D = (\sigma_2 + s_1^2)^3$, we easily get

$$\sigma_2 = s_1^2 + \sqrt[3]{D} \text{ and } \sigma_3 = s_3 + s_1 \sqrt[3]{D}$$

where in the finite field $\mathbb{F}_{2^{11}}$ the (unique) third root of D can is also computed as D^{1365}.

All in all, the knowledge of s_1, s_3, and s_9 can be used to compute the error locator polynomial $L(z)$ and by finding its roots, we are able to solve for e and f.

Remark 1. It is worth noting that Elia and Viterbo have developed an algebraic decoder also for G_3 [10]. This decoder works in a similar fashion and makes use of two syndromes. Even though it has to consider error values in addition to error locations, it treats the entire problem just by determining one polynomial.

5. GOLAY CODES AND FINITE GEOMETRY — AUTOMORPHISM GROUPS

Given a finite set S of v elements, we recall that a subset $B \subseteq \binom{S}{k}$ is called a $t-(v,k,\lambda)$ *block design*, provided that every t-element subset of S is contained in exactly λ elements of B. The elements of B are called *blocks*, and B is often referred to simply as a t design.

Now let C be a binary code. For a word $c \in C$, we call the set $\{i \mid c_i \neq 0\}$ the *support* of c. Furthermore, we say that the word C covers the word c' if $\mathrm{Supp}(C) \supseteq \mathrm{Supp}(c')$. Let C_d be the set of codewords that have weight d. We say that C_d holds a $t-(n,d,\lambda)$ design, if the supports of the words in C_d form the blocks of such a design.

Theorem 2. If C is a perfect binary (n, M, d) code (containing the all-zero word), then the set C_d of all codewords of minimum weight d hold a $t-(n,d,1)$ design, where $t = (d+1)/2$.

Proof The spheres of radius t are disjoint and cover $\mathbb{F}_2^n$. Hence for every binary word x of weight t there exists exactly one codeword $c \in C$ such that x is contained in the sphere of radius $t-1$ centered in c. By $d_H(c,x) \leq t-1$ we immediately get

$$w_H(c) \leq d_H(c,x) + w_H(x) \leq 2t - 1 = d$$

and so $d_H(c,x) = t-1$ and $c \in C_d$. All in all, we now have

$$2|\mathrm{Supp}(x) \cap \mathrm{Supp}(c)| = w_H(x) + w_H(c)$$
$$- d_H(x,c) \geq t + 2t - 1 - t = 2t$$

and therefore C covers x, which finishes our proof.

Corollary 1. The words of weight 7 of the binary Golay code hold a $4-(23,7,1)$ design.

For the extended binary Golay code, there is another interesting result.

Theorem 3. The codewords of weight 8 in the extended binary Golay code hold a $5-(24,8,1)$ design.

Proof If a word of weight 5 in $\mathbb{F}_2^{24}$ were covered by two codewords of minimum weight, then the distance between these words would be at most 6, which contradicts the minimum distance of binary Golay code. Hence every word of weight 5 is covered by at most one minimum weight codeword of G_2. We now count the number of words of weight 5 in $\mathbb{F}_2^{24}$. On one hand, this is clearly given by $\binom{24}{5}$; on the other hand, it is by the foregoing arguments clear that it is at least $|C_8|\binom{8}{5}$. We have seen however that $|C_8| = 759$, and hence by $759\binom{8}{5} = \binom{24}{5}$ every word of weight 5 is covered by at least one word in C_8. This completes the proof.

Remark 2. Using a modified technique of proof it can be shown that the supports of minimum weight words of G_3 hold a $4-(11,5,1)$ design, and that the supports of words of weight 6 of the extended ternary Golay code $\overline{G}_3$ form the blocks of a $5-(12,6,1)$ design.

The *automorphism group* of a design is the permutation group acting on its points that maps the set of blocks into itself. The automorphism group of a binary linear code is the set of all coordinate permutations that map the code into itself. In the ternary case we need to consider coordinate permutations and sign flips (monomial transformations) that map the code in question into itself. Without proof, we state the following basic facts.

Theorem 4

(a) The automorphism group of the $5-(24,8,1)$ design held by the words of weight 8 in the extended binary Golay code is given by the (simple) Mathieu group M_{24} of order $24\cdot 23\cdot 22\cdot 21\cdot 20\cdot 48$. This group is also the automorphism group of the extended binary Golay code $\overline{G}_2$.

(b) The automorphism group of the $5-(12,6,1)$ design held by the words of weight 6 in the extended ternary Golay code is given by the (simple) Mathieu group M_{12} of order $12\cdot 11\cdot 10\cdot 9\cdot 8$. There is a normal subgroup N of the automorphism group $\mathrm{Aut}(\overline{G}_3)$. This group has order 2, and $\mathrm{Aut}(\overline{G}_3)/N$ is isomorphic to M_{12}.

6. GOLAY CODES AND RING-LINEAR CODES

One very important observation in algebraic coding theory in the early 1990s was the discovery of the $\mathbb{Z}_4$-linearity of the Preparata codes, the Kerdock codes, and related families [15]. These had previously been known as notoriously nonlinear binary codes that had more codewords than any known linear code of the same length and minimum distance.

Defining the *Lee weight* on $\mathbb{Z}_4$ as $w_{\mathrm{Lee}}\colon \mathbb{Z}_4 \to \mathbb{N}, r \mapsto \min\{|r|, |4-r|\}$, we obtain what is called the *Gray isometry* of $\mathbb{Z}_4$ into $\mathbb{F}_2^2$ as

$$\begin{aligned}\gamma\colon (\mathbb{Z}_4, w_{\mathrm{Lee}}) &\to (\mathbb{Z}_2^2, w_H)\\ r &\mapsto r_0(0,1) + r_1(1,1)\end{aligned}$$

where $r_i \in \mathbb{Z}_2$ are the coefficients of the binary representation of $r \in \mathbb{Z}_4$, i.e., $r = r_0 + 2r_1$. The image of γ is the full space $\mathbb{F}_2^2$, and by abuse of notation we denote by γ also its componentwise extension to $\mathbb{Z}_4^n \to \mathbb{Z}_2^{2n}$. A (linear or nonlinear) binary code C of length $2n$ is said to have a $\mathbb{Z}_4$-linear representation if there is a $\mathbb{Z}_4$-linear code D of length n such that C is equivalent to $\gamma(D)$.

The results in Ref. 15 show that the Nordstrom–Robinson code has a $\mathbb{Z}_4$-linear representation by the [8,4] octacode, which is a lift of the Reed–Muller code RM(2,3). Long before there had been known another interesting way of constructing the Nordstrom–Robinson code. We present this code in the following paragraphs.

6.1. The Extended Binary Golay Code and the Nordstrom–Robinson Code

As the extended binary Golay code has minimum weight 8, we can assume (after column permutations) that it contains the word $(1^8 0^{16})$, where x^n is an abbreviation for n (consecutive) occurrences of the element x. If A is a generator matrix for this version of the Golay code, then it is clear that the first 7 columns of A must be linearly independent since A also serves as a check matrix for the code because of self-duality. On the other hand, the first 8 columns of A are linearly dependent, and hence we have the 8th column as the sum of the foregoing 7 columns. By elementary row operations we can finally achieve that the last 5 rows of this matrix have zeros in their first 8 positions. Hence, we end up with a generator matrix of the form

$$\left[\begin{array}{ccccccc|c|ccc}
1&0&0&0&0&0&0&1&*&\cdots&*\\
0&1&0&0&0&0&0&1&*&\cdots&*\\
0&0&1&0&0&0&0&1&*&\cdots&*\\
0&0&0&1&0&0&0&1&*&\cdots&*\\
0&0&0&0&1&0&0&1&*&\cdots&*\\
0&0&0&0&0&1&0&1&*&\cdots&*\\
0&0&0&0&0&0&1&1&*&\cdots&*\\
\hline
0&0&0&0&0&0&0&0&*&\cdots&*\\
\hline
0&0&0&0&0&0&0&0&*&\cdots&*\\
0&0&0&0&0&0&0&0&*&\cdots&*\\
0&0&0&0&0&0&0&0&*&\cdots&*\\
0&0&0&0&0&0&0&0&*&\cdots&*
\end{array}\right]$$

where the asterisks represent some binary entries.

From this matrix it can be seen that there are $32 = 2^{12-7}$ codewords that have zeros in their first 8 positions. Furthermore for every $i \in \{1,\ldots,7\}$ there are 32 codewords that have a 1 in the ith and 8th positions. We define N to be the (nonlinear) binary code to consist of the union of all these $8\cdot 32$ words, where we have cut off the first 8

entries. This is certainly a code of length 16. To determine its minimum distance we observe that any word in this code results from truncation of a word of the Golay code starting with either $(0, \ldots, 1 \ldots, 0, 1)$ or $(0, \ldots, 0)$. Any two words of these words therefore differ in at most two entries in the first 8 positions, and since the minimum distance of the Golay code is 8, the initially given words must differ in at least 6 positions. Applying the sphere packing bound we see that the minimum distance is at most 6, and overall this proves it to be given by 6.

Remark 3. In light of the statements at the beginning of this section we have questioned whether the binary Golay code itself might enjoy a $\mathbb{Z}_4$-linear representation. This has been answered to the negative in Ref. 15.

6.2. The $\mathbb{Z}_4$-Linear Golay Code and the Leech Lattice

Bonnecaze et al. investigated $\mathbb{Z}_4$-linear lifts of binary codes in connection with lattices. To explain how this works for the binary Golay code, we first Hensel-lift its generator polynomial $f = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1 \in \mathbb{F}_2[x]$ to the divisor

$$F = x^{11} + 2x^{10} - x^9 - x^7 - x^6 - x^5 + 2x^4 + x - 1$$

of $x^{23} - 1$ in $\mathbb{Z}_4[x]$. (Recall that according to Hensel's lemma [18, Sect. XIII.4] the polynomial F is the unique polynomial that reduces to f modulo 2 and divides $x^{23} - 1$ in $\mathbb{Z}_4[x]$.)

Extending the cyclic [23, 12] code G_4 that is generated by F, we obtain a [24,12] code $\overline{G}_4$ of minimal *Lee weight* 12. Defining the *Euclidean weight* on $\mathbb{Z}_4$ as the squared Lee weight and extending it additively, we obtain the minimum Euclidean weight as 16.

The so-called construction A provides a means to construct a lattice Λ from G_4, by setting

$$\Lambda := \nu^{-1}(\overline{G}_4)$$

where ν denotes the natural map $\mathbb{Z}^{24} \to \mathbb{Z}_4^{24}$. Surprisingly, this lattice is one of the best studied lattices so far. It has maximal density that is achievable in 24 dimensions. Referring to the proof in Ref. 2 we state the following theorem.

Theorem 5. Construction A of the quaternary Golay code $\overline{G}_4$ yields up to lattice equivalence the Leech lattice.

We should mention that there are several different constructions for the Leech lattice involving the binary Golay code, but not purely via construction A. This construction, however, is one of the most natural ways to construct a lattice from a code, and hence inspired us to mention it here.

6.3. Higher Lifts of Golay Codes

It is possible to define a weight on rings (e.g., $\mathbb{Z}_8$ or $\mathbb{Z}_9$) for which a generalized version of the above Gray map exists. The only difference to keep in mind is that these maps are not necessarily subjective anymore.

Specifically, the normalized homogeneous weight (defined in [4]) on $\mathbb{Z}_8$ as

$$w_{\text{hom}}: \mathbb{Z}_8 \to \mathbb{N},\ r \mapsto \begin{cases} 0, & \text{r = 0} \\ 2, & \text{r = 4} \\ 1, & \text{else} \end{cases}$$

is such a weight. The according Gray map of this ring into $\mathbb{Z}_2^4$ is given by

$$\begin{aligned} \gamma: (\mathbb{Z}_8, 2w_{\text{hom}}) &\to (\mathbb{Z}_2^4, w_H) \\ r &\mapsto r_0(0,0,1,1) + r_1(0,1,0,1) + r_2(1,1,1,1) \end{aligned}$$

where $r_i \in \mathbb{Z}_2$ are the coefficients of the binary representation of $r \in \mathbb{Z}_8$, namely, $r = r_0 + 2r_1 + 4r_2$. Again we also denote by γ its componentwise extension to $\mathbb{Z}_8^n \to \mathbb{Z}_2^{4n}$.

Following the presentation in Ref. 7 Hensel lifting the generator polynomial of G_2 to $\mathbb{Z}_8[x]$ results in the polynomial $x^{11} + 2x^{10} - x^9 + 4x^8 + 3x^7 + 3x^6 - x^5 + 2x^4 + 4x^3 + 4x^2 + x - 1$, which generates a free [23,12] code G_8 over $\mathbb{Z}_8$. Extending the latter code by a parity check, we obtain a free $\mathbb{Z}_8$-linear self-dual [24,12] code $\overline{G}_8$.

Looking into Brouwer's and Litsyn's tables [3,17], it is remarkable that this code has more codewords than does any presently known binary code of length 96 and minimum distance 24. Hence we have an outperforming example of a nonlinear code that is constructed using the binary Golay code.

Remark 4. Using a similar technique it has been shown that a non-linear ternary $(36, 3^{12}, 15)$ code can be constructed as the image of a $\mathbb{Z}_9$-linear lift of the ternary Golay code. This code is not outperforming, but so far no ternary $(36, 3^{12})$ codes with a better minimum distance is known. The details can be found in Ref. 13.

BIOGRAPHY

Marcus Greferath received his Diploma and Ph.D. degrees in mathematics in 1990 and 1993, respectively, from Mainz University (Germany). He joined the Department of Mathematics of Duisburg University 1992 as Research Assistant. In Duisburg he obtained the position of an Assistant Professor in 1994 and finished his Habilitation in 2000. Dr. Greferath has published more than 20 papers in the areas of ring geometry, coding theory, and cryptography. In 1997 he started a 2-year research stay at AT&T Shannon Laboratory in New Jersey. In 1999 he held a one-year visiting professorship at Ohio University in Athens (Ohio) and was appointed as Assistant Professor at the Department of Mathematics of San Diego State University in 2001. His areas of interest are finite geometry, coding theory, and cryptography with a focus on the role of rings and modules in these disciplines.

BIBLIOGRAPHY

1. A. Barg, At the dawn of the theory of codes, *Math. Intelligencer* **15**(1): 20–26 (1993).

2. A. Bonnecaze, P. Sole and A. R. Calderbank, Quaternary quadratic residue codes and unimodular lattices, *IEEE Trans. Inform. Theory* **41**(2): 366–377 (1995).
3. A. E. Brouwer, Bounds on the minimum distance of linear codes, *http://www.win.tue.nl/math/dw/voorlincod.html*, 2002.
4. I. Constantinescu and W. Heise, A metric for codes over residue class rings of integers, *Problemy Peredachi Informatsii* **33**(3): 22–28 (1997).
5. J. H. Conway et al., M12, *Atlas of Finite Groups*, Clarendon Press, Oxford, UK, 1985.
6. J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed., Grundlehren der Mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences), Springer-Verlag, New York, 1999.
7. I. M. Duursma, M. Greferath, S. N. Litsyn, and S. E. Schmidt, A $\mathbb{Z}_8$-linear lift of the binary Golay code and a nonlinear binary (96, 2^{37}, 24)-code, *IEEE Trans. Inform. Theory* **47**(4): 1596–1598 (2001).
8. P. Delsarte and J.-M. Goethals, Unrestricted codes with the Golay parameters are unique, *Discrete Math.* **12**: 211–224 (1975)
9. M. Elia, Algebraic decoding of the (23, 12, 7) Golay code, *IEEE Trans. Inform. Theory* **33**(1): 150–151 (1987).
10. M. Elia and E. Viterbo, Algebraic decoding of the ternary (11, 6, 5) Golay code, *Electron. Lett.* **28**(21): 2021–2022 (1992).
11. M. J. E. Golay, Notes on digital coding, *Proc. IRE* **37**: 657 (1949).
12. D. Y. Goldberg, Reconstructing the ternary Golay code, *J. Combin. Theory S A* **42**(2): 296–299 (1986).
13. M. Greferath and S. E. Schmidt, Gray isometries for finite chain rings and a nonlinear ternary (36, 3^{12}, 15) code, *IEEE Trans. Inform. Theory* **45**(7): 2522–2524 (1999).
14. M. Greferath and E. Viterbo, On $\mathbb{Z}_4$- and $\mathbb{Z}_9$-linear lifts of the Golay codes, *IEEE Trans. Inform. Theory* **45**(7): 2524–2527 (1999).
15. A. R. Hammons et al., The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes, *IEEE Trans. Inform. Theory* **40**(2): 301–319 (1994).
16. S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, Inc. 1983.
17. S. Litsyn, An updated table of the best binary codes known, in *Handbook of Coding Theory*, North-Holland, Amsterdam, 1998, Vols. I and II, pp. 463–498.
18. B. R. McDonald, *Finite Rings with Identity*, Marcel Dekker, New York, 1974.
19. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, N, 1977.
20. G. Pasquier, The binary Golay code obtained from an extended cyclic code over $\mathbb{F}_8$, *Eur. J. Combin. Theory* **1**(4): 369–370 (1980).
21. V. Pless, On the uniqueness of the Golay codes, *J. Combin. Theory* **5**: 215–228 (1968).
22. V. S. Pless, W. C. Huffman, and R. A. Brualdi, *Handbook of Coding Theory*, North-Holland, Amsterdam, 1998, Vols. I and II.
23. S. Roman, *Coding and Information Theory*, Springer, New York, 1992.

GOLAY COMPLEMENTARY SEQUENCES

MATTHEW G. PARKER
University of Bergen
Bergen, Norway

KENNETH G. PATERSON
University of London
Egham, Surrey, United Kingdom

CHINTHA TELLAMBURA
University of Alberta
Edmonton, Alberta, Canada

1. INTRODUCTION

Complementary sequences were introduced by Marcel Golay [1] in the context of infrared spectrometry. A *complementary pair* of sequences (CS pair) satisfies the useful property that their out-of-phase *aperiodic autocorrelation* coefficients sum to zero [1,2]. Let $\mathbf{a} = (a_0, a_1, \ldots, a_{N-1})$ be a sequence of length N such that $a_i \in \{+1, -1\}$ (we say that $\mathbf{a}$ is bipolar). Define the aperiodic autocorrelation function (AACF) of $\mathbf{a}$ by

$$\rho_{\mathbf{a}}(k) = \sum_{i=0}^{N-k-1} a_i a_{i+k}, \quad 0 \le k \le N-1 \tag{1}$$

Let $\mathbf{b}$ be defined similarly to $\mathbf{a}$. The pair $(\mathbf{a}, \mathbf{b})$ is called a Golay complementary pair (GCP) if

$$\rho_{\mathbf{a}}(k) + \rho_{\mathbf{b}}(k) = 0, \quad k \ne 0 \tag{2}$$

Each member of a GCP is called a Golay complementary sequence (GCS, or simply Golay sequence). Note that this definition (2) can be generalized to nonbinary sequences. For example, a_i and b_i can be selected from the set $\{\zeta^0, \zeta^1, \ldots, \zeta^{2^h-1}\}$ where ζ is a primitive q-th root of unity, which yields so-called polyphase Golay sequences. In this survey, however, we emphasize binary GCPs.

It is helpful to view (2) in polynomial form. A sequence $\mathbf{a}$ can be associated with the polynomial $a(z) = a_{N-1}z^{N-1} + a_{N-2}z^{N-2} + \cdots + a_1 z + a_0$ in indeterminate z with coefficients ± 1. The pair $(\mathbf{a}, \mathbf{b})$ is then a GCP if the associated polynomials $(a(z), b(z))$ satisfy

$$a(z)a(z^{-1}) + b(z)b(z^{-1}) = 2N \tag{3}$$

Equations (2) and (3) are equivalent expressions because $a(z)a(z^{-1}) = \rho_{\mathbf{a}}(0) + \sum_{1}^{N-1} \rho_{\mathbf{a}}(k)(z^k + z^{-k})$. A further condition can be obtained by restricting z to lie on the unit circle in the complex plane, i.e., $z \in \{e^{2\pi jt} \mid j^2 = -1, 0 \le t < 1\}$. Then $|a(z)|^2 = a(z)a(z^{-1})$ and we have

$$|a(z)|^2 + |b(z)|^2 = 2N, \quad |z| = 1 \tag{4}$$

This means that the absolute value of each polynomial on the unit circle is bounded by $\sqrt{2N}$.

Golay complementary pairs and sequences have found application in physics (Ising spin systems), combinatorics (*orthogonal designs* and *Hadamard matrices*) and telecommunications (e.g., to surface-acoustic

wave design, the Loran C precision navigation system, channel-measurement, optical time-domain reflectometry [3], synchronization, spread-spectrum communications, and, recently, *orthogonal frequency division multiplexing* (OFDM) systems [4–7]. Initially, the properties of the *pair* were primarily exploited [1] in a two-channel setting, and periodic GCPs have lately been proposed for two-sided *channel-estimation*, where the two sequences in the pair form a preamble and postamble training sequence, respectively [8]. In recent years, the spectral spread properties of each individual sequence in the pair have also been used. As an example of this, we briefly describe the application of Golay sequences in OFDM. Here, given a data sequence, $\mathbf{a} = (a_0, a_1, \ldots, a_{N-1})$, the transmitted signal $s_{\mathbf{a}}(t)$ as a function of time t is essentially the real part of a discrete Fourier transform (DFT) of $\mathbf{a}$:

$$s_{\mathbf{a}}(t) = \sum_{i=0}^{N-1} a_i e^{2\pi j(i\Delta f + f_0)t} \tag{5}$$

where Δf the frequency separation between adjacent subcarrier pairs and f_0 is the base frequency. Notice that $|s_{\mathbf{a}}(t)| = |a(e^{2\pi j i \Delta f t})|$ where $a(z)$ is the polynomial corresponding to $\mathbf{a}$. Thus, the power characteristics of the OFDM signal can be studied by examining the behavior of an associated polynomial on $|z| = 1$. In particular, if $\mathbf{a}$ is a GCS, then we have that $|s_{\mathbf{a}}(t)|^2 \le 2N$ so that the *peak-to-mean envelope power ratio* (PMEPR) of the signal is at most 2.0. Having such tightly bounded OFDM signals eases amplifier specification at the OFDM transmitter.

Let $\mathbf{A} = (A_0, A_1, \ldots, A_{N'-1})$ be the N'-point oversampled DFT of $\mathbf{a}$, where $N' \ge N$, i.e.,

$$A_k = \sum_{i=0}^{N-1} a_i \omega^{ik} = a(\omega^k) \quad 0 \le k < N',$$

where $\omega = e^{2\pi j/N'}$ is a complex N'-th root of unity. For N' large, the values of the N'-point oversampled DFT of $\mathbf{a}$ can be used to approximate the values $a(z)$, $|z| = 1$, and thus the complex OFDM signal in (5).

Example 1. Let $\mathbf{a} = - + + - + - + + + -$, $\mathbf{b} = - + + + + + + - - +$, where '+' and '−' mean 1 and −1, respectively. The AACFs of $\mathbf{a}$ and $\mathbf{b}$ are:

$$\rho_{\mathbf{a}}(k) = (10, -3, 0, -1, 0, 1, 2, -1, -2, 1),$$
$$\rho_{\mathbf{b}}(k) = (10, 3, 0, 1, 0, -1, -2, 1, 2, -1).$$

It is evident that the AACFs of $\mathbf{a}$ and $\mathbf{b}$ sum to a δ-function, as required by (2) and (3), so $(\mathbf{a}, \mathbf{b})$ is a GCP. The absolute squared values of the 20-point oversampled DFT of $\mathbf{a}$ and $\mathbf{b}$ are:

$$\mathbf{A} = 10 \cdot (0.40, 0.44, 0.15, 0.73, 1.85, 0.20, 1.05, 1.67, 0.95, 1.96, 1.60, 1.96, 0.95, 1.67, 1.05, 0.20, 1.85, 0.73, 0.15, 0.44)$$
$$\mathbf{B} = 10 \cdot (1.60, 1.56, 1.85, 1.27, 0.15, 1.80, 0.95, 0.33, 1.05, 0.04, 0.40, 0.04, 1.05, 0.33, 0.95, 1.80, 0.15, 1.27, 1.85, 1.56)$$

At every point these two power spectra add to 20, as required by (4).

It should be stressed that the bound of $\sqrt{2N}$ on the amplitude of $a(z)$ on $|z| = 1$ is extremely low for any bipolar sequence of length greater or equal to about 16. One would not find such sequences by chance, and the complementary sequence/aperiodic correlation approach is currently the only construction method known that tightly upper bounds these values for bipolar sequences. There is also the *Rudin-Shapiro* (RuS) construction [9], which appeared soon after Golay's initial work, but the RuS construction can be viewed as a basic recursive Golay construction technique. This construction is described in Section 2.2. Indeed, research on the uniformity of polynomials on the unit circle has continued largely independently in the mathematical community for many years [10–13], and this work indicates that sequences with good AACFs and flat DFT spectra, or equivalently, polynomials that are approximately uniform on $|z| = 1$, are rather difficult to construct. For example, the celebrated conjecture of Littlewood on *flat polynomials* on the unit circle is still open:

Conjecture 1 [12]. There exist a pair of constants C_0, C_1 and a series of degree $N-1$ polynomials $a(z)$ with ± 1 coefficients such that, as $N \to \infty$,

$$C_0\sqrt{N} \le |a(z)| \le C_1\sqrt{N}, \quad |z| = 1 \tag{6}$$

There is no known construction that produces polynomials satisfying the lower bound of Conjecture 1, and the complementary sequence approach is the only one known that gives polynomials satisfying the upper bound.

2. EXISTENCE AND CONSTRUCTION

2.1. Necessary Conditions

As we shall see in the next section, GCPs are known to exist for all lengths $N = 2^{\alpha}10^{\beta}26^{\gamma}$, $\alpha, \beta, \gamma \ge 0$, [14]. GCPs are not known for any other lengths. Golay showed that the length N of a Golay sequence must be the sum of two squares (where one square may be 0) [2]. More recently, it has been shown that GCPs of length N do not exist if there is a prime p with $p = 3 \pmod 4$ such that $p \mid N$, [16]. This generalized earlier, weaker nonexistence results. Therefore, the admissible lengths <100 are

$$1, 2, 4, 8, 10, 16, 20, 26, 32, 34^*, 40, 50^*, 52, 58^*, 64, 68^*, 74^*, 80, 82^*$$

Moreover, various computer searches have eliminated lengths marked by "$*$" in the preceding list. Therefore, the lengths, $N < 100$, for which GCPs exist are:

$$1, 2, 4, 8, 10, 16, 20, 26, 32, 40, 52, 64, 80 \tag{7}$$

In the next sections, we provide constructions covering all these lengths.

2.2. Recursive Constructions

Using Eq. (3), many recursive constructions for GCPs can be obtained via simple algebraic manipulation. For example, if $a(z)$ and $b(z)$ are a Golay pair of length N, then simple algebraic manipulation shows that $a(z)+z^N b(z)$ and $a(z)-z^N b(z)$ also satisfy Eq. (3) with $2N$ being replaced by $4N$. This is in fact the well-known Golay–Rudin–Shapiro recursion, generating a length $2N$ GCP from a length N GCP [13]. We may write this more simply in terms of sequences as

$$(\mathbf{a}, \mathbf{b}) \rightarrow (\mathbf{a} \mid \mathbf{b}, \mathbf{a} \mid \overline{\mathbf{b}}) \tag{8}$$

where '|' means concatenation.

The following are a few other recursive constructions:

- The construction of Turyn [14] can be stated as follows. Let $(\mathbf{a}, \mathbf{b})$ and $(\mathbf{c}, \mathbf{d})$ be GCPs of length M and N, respectively. Then

$$\begin{aligned} &a(z^N)(c(z)+d(z))/2 + z^{N(M-1)} b(z^{-N})(c(z)-d(z))/2, \\ &b(z^N)(c(z)+d(z))/2 - z^{N(M-1)} a(z^{-N})(c(z)-d(z))/2 \end{aligned} \tag{9}$$

  is a GCP of length MN.
- The constructions of Golay in [2] are obtained as follows. Let $(\mathbf{a}, \tilde{\mathbf{b}})$ and $(\mathbf{c}, \mathbf{d})$ be GCPs of lengths M and N, respectively, where $\tilde{\mathbf{b}}$ means reversal of $\mathbf{b}$.

  Golay's *concatenation* construction can be stated as

$$a(z^N)c(z) + b(z^N)d(z)z^{MN}, \quad \tilde{b}(z^N)c(z) - \tilde{a}(z^N)d(z)z^{MN} \tag{10}$$

  is a GCP of length $2MN$.

  Golay's *interleaving* construction can be stated as

$$\begin{aligned} &a(z^{2N})c(z^2) + b(z^{2N})d(z^2)z, \\ &\tilde{b}(z^{2N})c(z^2) - \tilde{a}(z^{2N})d(z^2)z \end{aligned} \tag{11}$$

  is a GCP of length $2MN$.

Repeated application of Turyn's construction, beginning with pairs of lengths 2, 10, and 26 given in Section 2.4, can be used to construct GCPs for all lengths $N = 2^\alpha 10^\beta 26^\gamma$, $\alpha, \beta, \gamma \geq 0$.

2.3. Direct Constructions

In Ref. 2, Golay gave a direct construction for GCPs of length $N = 2^m$. Reference [4] gave a particularly compact description of this construction by using algebraic normal forms (ANFs). With a Boolean function $a(\mathbf{x}) = a(\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{m-1}})$ in m variables, we associate a length 2^m sequence $\mathbf{a} = (a_0, a_1, \ldots, a_{2^m-1})$, where

$$a_i = (-1)^{a(i_0, i_1, \ldots, i_{m-1})}, \quad i = \sum_{k=0}^{m-1} i_k 2^k$$

Thus the i-th term of the sequence $\mathbf{a}$ is obtained by evaluating the function a at the 2-adic decomposition of i. Then Ref. 4 showed that, for any permutation π of $\{0, 1, \ldots, m-1\}$, and any choice of constants $c_j, c, c' \in \mathbb{Z}_2$, the pair of functions

$$\begin{aligned} a(\mathbf{x}) &= \sum_{i=0}^{m-2} x_{\pi(i)} x_{\pi(i+1)} + \left(\sum_{j=0}^{m-1} c_j x_j \right) + c \\ b(\mathbf{x}) &= a(\mathbf{x}) + x_{\pi(0)} + c' \end{aligned} \tag{12}$$

yields a length 2^m GCP $(\mathbf{a}, \mathbf{b})$.

It is simple, given this representation, to show that this construction gives a set of $m!2^m$ distinct Golay sequences of length 2^m, each of which occurs in at least 4 GCPs. Perhaps more important, by expressing this set in the form Eq. (12), Ref. 4 identified a large set of GCS occurring as a subset of the binary Reed–Muller code RM$(2, m)$. Consequently, each sequence in the set has PMEPR at most 2, and the Hamming distance between any two sequences in the set is at least 2^{m-2}. This set therefore has a very attractive combination of PMEPR and error-correcting properties making it applicable in OFDM applications. For further details, see Ref. 4.

It was shown in Ref. 7 that the direct construction of Golay described above and Golay's recursive constructions Ref. 2 described in Section 2.2 in fact result in the same set of Golay sequences.

2.4. Symmetry and Primitivity

We consider simple symmetry operations that leave complementary properties and sequence length of a GCP invariant.

A length N GCP $(\mathbf{a}, \mathbf{b})$ remains a GCP under the operations [2]:

- Swapping the sequences in the pair.
- Negation (element multiplication by -1) of either or both sequences.
- Reversal of either or both sequences.
- The "linear offset" transformation $a_i \rightarrow (-1)^i a_i$, $b_i \rightarrow (-1)^i b_i$.

The action of these symmetries on a GCP generates a set of GCPs of size at most 64, which we call the *conjugates* of the original GCP. These symmetries can aid in computer searches for new GCPs.

We define a *primitive GCP* to be one that cannot be constructed from any shorter GCP by means of the recursive constructions described in Section 2.2. Primitive GCPs are only known to exist for lengths 2, 10, and 26. There is one primitive pair for lengths 2 and 26, and two primitive pairs for length 10, up to equivalence via the above symmetry operations. These primitive pairs are as follows:

$$(++,\ +-);$$

$$(-++-+-+++-,\ -++++++--+);$$

$$(+-+-++++--,\ ++++-++--+);$$

$$(+-++--+----+-+----++---+-+,$$
$$\quad -+--++-+++++-------++---+-+).$$

Golay points out that the two GCPs of length 10 are equivalent under decimation [2]. Specifically, the second pair of length 10 above is obtained from the first pair by taking successive 3rd sequence elements, cyclically. There is no proof that more primitive pairs cannot exist for $N > 100$, but none have been discovered for 40 years and it is conjectured that all GCPs arise from the four primitive pairs of lengths 2, 10, and 26, as given above.[1]

2.5. Enumeration

Two main types of enumeration are possible. One can enumerate the number of GCPs of a given length. Second, one can enumerate the number of Golay *sequences* of a given length. Since a Golay sequence is present in more than one GCP, the number of the former is greater than the number of the latter. As we have already seen in our brief discussion of OFDM, the enumeration of Golay sequences is of some practical importance. Table 1 provides a complete enumeration of GCPs for all possible lengths up to 100.

From Table 1 it is evident that the largest sets occur for lengths that have a large power of 2 as a factor. Here is a useful enumeration theorem:

Theorem 1 [2]. There are exactly $2^{m+2}m!$ GCPs of length 2^m that can be derived from the primitive pair $\{++, +-\}$ by repeated application of the symmetry operations and Golay's recursive constructions.

Next we consider the enumeration of Golay sequences. We have:

Theorem 2 [4,18]. Golay's direct construction produces exactly $m!2^m$ Golay sequences of length 2^m.

It was shown in Ref. 7 that the set of sequences in this theorem is identical to that which can be obtained from the primitive pair $(++, +-)$ by repeated application of Golay's recursive constructions. Theorem 2 accounts for all Golay sequences of lengths 2^m when $1 \le m \le 6$. It is not known if every Golay sequence of length 2^m must arise from Golay's direct construction when $m \ge 7$.

3. THE MERIT FACTOR OF COMPLEMENTARY SEQUENCES

The *merit factor* is a useful measure of the quality of sequences in certain applications where aperiodic correlations are important. It was introduced by Golay in Ref. 18. Let $\mathbf{a}$ be any length N sequence. Then the merit factor of $\mathbf{a}$ is defined to be

$$F(\mathbf{a}) = \frac{N^2}{2\sum_{k=1}^{N-1} |\rho_{\mathbf{a}}(k)|^2} \tag{13}$$

where $\rho_{\mathbf{a}}(k)$ is the AACF of $\mathbf{a}$.

The merit factor is, in fact, a spectral measure; it measures the mean-square deviation from the flat Fourier spectrum. Specifically,

$$1/F(\mathbf{a}) = \frac{1}{N^2}\int_0^1 (|a(e^{j2\pi t})|^2 - N)^2\, dt \tag{14}$$

where $a(z)$ is a polynomial whose values on $|z| = 1$ gives the Fourier transform of $\mathbf{a}$.

It is desirable to find sequences with high merit factor. A random sequence has merit factor around 1. It has been established that the asymptotic merit factor of a length 2^m Golay–Rudin–Shapiro (RuS) sequence is 3.0, which is high [19]. This is not the best possible; for example, shifted-Legendre sequences attain an asymptotic merit factor of 6.0 [20], and computer searches up to length 200 have revealed ± 1-sequences of merit factor around 8.5. There is also the celebrated Barker sequence of length 13, which has merit factor 14.08. However the length $N = 2^m$ RuS sequences $\mathbf{a}_m$ are notable as the quantities $\sigma_m = \sum_{k=0}^{2^m-1} |\rho_{\mathbf{a}_m}(k)|^2$ obey a simple *generalized Fibonacci recursion*, namely,

$$\sigma_m = 2\sigma_{m-1} + 8\sigma_{m-2} \tag{15}$$

with initial conditions $\sigma_1 = 1$, $\sigma_2 = 2$ [19].

This recursion immediately gives an asymptotic merit factor for the RuS sequences of 3.0 and is significant because it demonstrates the existence of a sequence family with large merit factor for which the merit factors do not need to be computed explicitly. This asymptotic value of 3.0 also holds for any Golay sequence obtained by applying symmetry operations to the RuS sequences [19]. Taking any other non-Golay pair as a starting seed always gives an asymptotic merit factor of K, for some constant, $K < 3.0$ [21].

4. LOW-COMPLEXITY CORRELATION

The pairwise property of GCPs has been exploited for channel measurement [1], but until 1990 or thereabouts the properties of individual sequences of a GCP were not wholly exploited [22], although Shapiro had stated Eq. (4) in his master's thesis of 1951 [9]. Budisin argues that Golay sequences are as good as, if not better than,

[1] However, a very recent paper by Borwein and Ferguson [17] also regards a length 20 pair as primitive, specifically: {+ + + + − + − − − + + − − + + − + − − +, + + + + − + + + + + − − − + − + − + + −}.

Table 1. The Number of Golay Complementary Pairs for All Lengths, $N < 100$

N	1	2	4	8	10	16	20	26	32	40	52	64	80
#GCPs [17]	4	8	32	192	128	1536	1088	64	15360	9728	512	184320	102912

m-sequences for application as pseudo-noise sequences due to their superior aperiodic spectral properties [22]. Also, there are more Golay sequences than m-sequences. Figure 1 lends some support for this view, where the Fourier spectra (from left to right) of a length 127 m-sequence and a length 127 shifted-Legendre sequence are compared with that of a length 128 RuS sequence.

Budisin [22] proposed a highly efficient method to perform correlation of an incoming data stream with a Golay sequence of length N, which achieves a complexity of $2\log_2(N)$ operations per sample, as opposed to N operations per sample for direct correlation. To do this, he interpreted the Golay construction for length $N = 2^m$ sequences using delay to implement concatenation, i.e., $\mathbf{a} \mid \mathbf{b}$ can be implemented as $\mathbf{a}[k] + \mathbf{b}[k + D]$, where $[k]$ indicates the starting time of $\mathbf{a}$ and D is the length (duration) of $\mathbf{a}$. Implementing the recursion of (8) is then achieved by serially combining delay elements, D_i, of duration 2^i, as shown in Fig. 2. Now we commence our Rudin–Shapiro recursion with $\mathbf{a} = \mathbf{b} = 1$. In other words, we input the δ function to the left-hand side of Fig. 2, and output our GCP, $(\mathbf{a}', \mathbf{b}')$, on the right. So the pair of Golay sequences realized by Fig. 2 are two impulse responses. We can therefore reinterpret Fig. 2 as a filter that correlates a received sequence, input from the left, with the reversals of $\mathbf{a}$ and $\mathbf{b}$. By choosing the ω_i in Fig. 2 from $\{1, -1\}$, we can choose to correlate with different length 2^m GCPs.

5. COMPLEMENTARY SETS AND ORTHOGONAL MATRICES

5.1. Complementarity with Respect to a Set of Sequences

Golay complementary pairs can be generalized to sets containing two or more sequences [23]. Analogously to Eq. (2), we say that a set of T bipolar sequences of length N $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_T)$ form a complementary sequence set of size T (a T-CSS) if

$$\sum_{i=1}^{T} \rho_{\mathbf{a}_i}(k) = 0 \quad k \neq 0 \tag{16}$$

In terms of polynomials, this is equivalent to

$$\sum_{i=1}^{T} |a_i(z)|^2 = TN, \quad |z| = 1 \tag{17}$$

Thus, all the DFT components of a sequence $\mathbf{a}$ that lies in a T-CSS are of size at most $\sqrt{TN}$. Of course, a 2-CSS is just a Golay complementary pair.

To date, little work has been done to formally establish primitivity conditions for T-CSS, $T > 2$, although Golay already found 4-CSS in Ref. 1. It can be shown that CSS only occur for T even, and Turyn showed that 4-CSS are only admissible at lengths N if N is a sum of at most *three* squares [14]. Dokovic later showed that 4-CSS exist for all even $N < 66$ [24]. Tseng and Liu [23] showed that for a

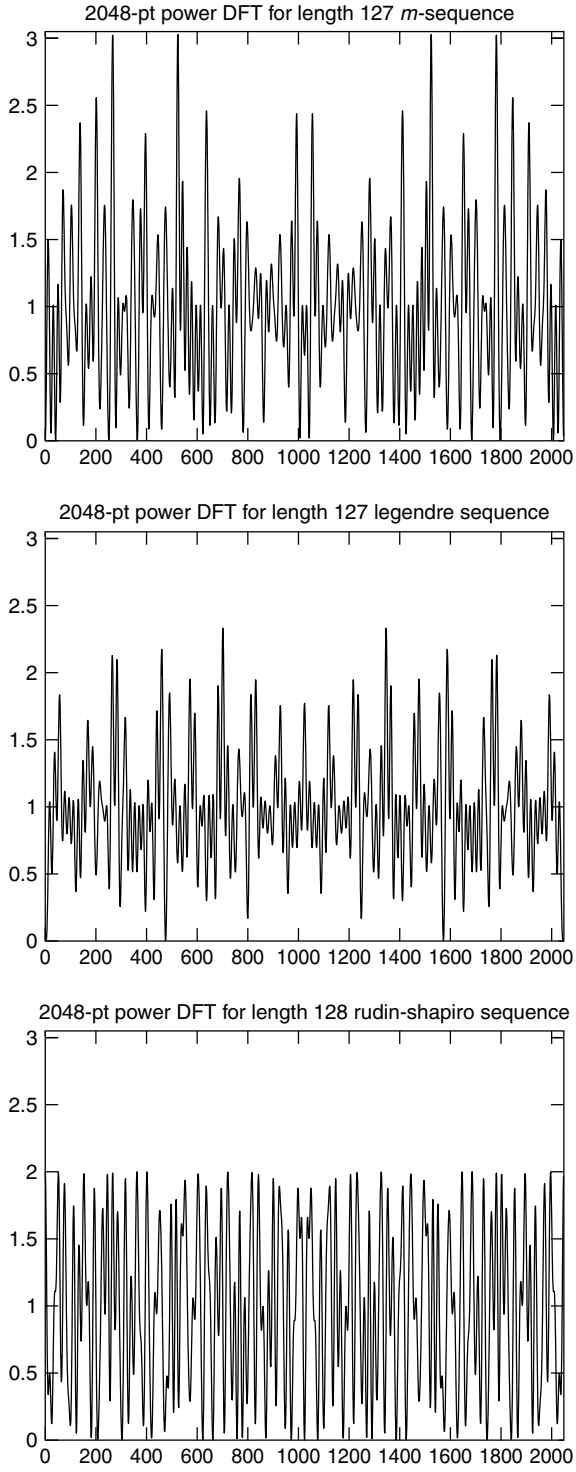


Figure 1. Power spectra for length 127 m-sequence, length 127 shifted-Legendre, and length 128 Rudin-Shapiro sequences, (power on y-axis, spectral index on x-axis).

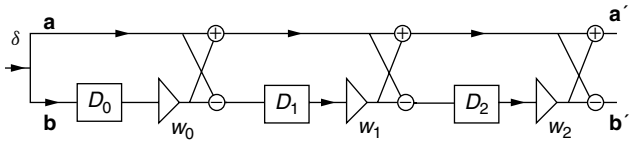


Figure 2. Fast Golay correlator.

Table 2. Number of Possibly Primitive Quadriphase Golay Pairs

Length	2	3	4	5	6	7	8	9	10	11	12	13
#Inequivalent Pairs [29]	—	1	—	1	2	—	4	—	14	1	32	1

CSS of odd length N, T must be a multiple of 4. By way of example, here are 4-CSS of lengths 3, 5, 7:

$$\{+++,\ -++,\ +-+,\ ++-\}$$

$$\{+----,\ -++-+,\ +---+,\ ---+-\}$$

$$\{+++-+++,\ +-+++--,\ +--+-++,\ ++-+---\}$$

Thus, 4-CSS can exist at lengths N where 2-CSS cannot. Turyn presented constructions for 4-CSS for all odd lengths $N \leq 33$, and $N = 59$ [14].

An orthogonal matrix is defined as a matrix whose columns are pairwise orthogonal. The following theorem is straightforward.

Theorem 3 [23]. Let $\mathbf{P}$ be a $T \times N$ orthogonal matrix, $N \leq T$. Then the rows of P form a T-CSS of length N.

The primitive GCP $(++, +-)$, is an example of Theorem 3. When the elements of $\mathbf{P}$ are ± 1 and $N = T$, then $\mathbf{P}$ is a Hadamard matrix, so a subset of T-CSS is provided by the set of Hadamard matrices.

5.2. Symmetries and Constructions

The symmetries and constructions for GCPs given in Sections 2.2 and 2.3 generalize to give constructions for T-CSS. One can also construct T'-CSS by combining T-CSS, $T' > T$. As one example, we have:

- Let $(\mathbf{u}, \mathbf{v})$ and $(\mathbf{x}, \mathbf{y})$ be GCPs of length N_0 and N_1, respectively. Then, $(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ is a 4-CSS of length $N_0 + N_1$, where

$$\mathbf{a} = (\mathbf{u} \mid \mathbf{x}),\ \mathbf{b} = (\mathbf{u} \mid -\mathbf{x}),\ \mathbf{c} = (\mathbf{v} \mid \mathbf{y}),\ \mathbf{d} = (\mathbf{v} \mid -\mathbf{y})$$

A fundamental recursive construction generalizing Golay's recursive constructions and relating CSS to orthogonal matrices is given in Ref. 23.

Theorem 4 [23]. Let $(\mathbf{a_0}, \mathbf{a_1}, \ldots, \mathbf{a_{T-1}})$ be a T-CSS of length N, represented by a $T \times N$ matrix, $\{\mathbf{F}\}$, with rows $\mathbf{a_j}$. Let $\mathbf{O} = (o_{ik})$ be an $S \times T$ orthogonal matrix (so $S \geq T$). Define

$$\mathbf{F}' = \mathbf{F} \odot \mathbf{O} = \begin{pmatrix} o_{00}\mathbf{a_0} & o_{01}\mathbf{a_1} & \ldots & o_{0(T-1)}\mathbf{a_{T-1}} \\ o_{10}\mathbf{a_0} & o_{11}\mathbf{a_1} & \ldots & o_{1(T-1)}\mathbf{a}_{T-1} \\ \ldots & \ldots & \ldots & \ldots \\ o_{(S-1)0}\mathbf{a_0} & o_{(S-1)1}\mathbf{a_1} & \ldots & o_{(S-1)(T-1)}\mathbf{a_{T-1}} \end{pmatrix} \tag{18}$$

Then $\mathbf{F}'$ is an $S \times TN$ matrix whose rows form an S-CCS of length TN.

Taking $(\mathbf{a}_0, \mathbf{a}_1)$ to be a GCP and $O = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, we recover Golay's concatenation construction. The basic symmetry operations can be interpreted as row/column permutations of $\mathbf{O}$ and as point-multiplication of rows of $\mathbf{O}$ by a constant vector (these operations maintain the orthogonality of $\mathbf{O}$).

The following theorem combines T-CSS of different lengths to build T'-CSS, where $T' > T$.

Theorem 5 [25]. Suppose there exist T_0-CSS, T_1-CSS, ..., T_{t-1}-CSS, of lengths $N_0, N_1, \ldots, N_{t-1}$, respectively. Let $T = \mathrm{lcm}(T_0, T_1, \ldots, T_{t-1})$. Suppose there also exists an $S \times T$ orthogonal matrix with ± 1 entries. Then there exists an ST-CSS of length $N' = N_0 + N_1 + \ldots + N_{t-1}$.

As with GCPs, CSS also have applications to OFDM: a primary drawback with the proposal to use the set of length 2^m GCPs as a codeset for OFDM is that the code rate of the set rapidly decreases as m increases. To obtain a larger codeset, one can consider sequences that lie in T-CSS for some $T > 2$. The resulting codeset will have PMEPR at most T.

5.3. Complementarity with Respect to a Larger Set of Transforms

Although CS pairs and, more generally, CSS, are usually defined to be complementary with respect to their AACFs (with a corresponding property on power spectra under the one-dimensional DFT), one can more generally define and discover sets that are complementary with respect to *any* specified transform. It can be shown, Ref. 26, that Golay CSS of length 2^m, as constructed using Theorem 5, have a very strong property:

Theorem 6. Let $\mathbf{U}$ be a 2×2 complex-valued matrix such that $\mathbf{U}\mathbf{U}^\dagger = 2\mathbf{I}$, where $\mathbf{I}$ is the 2×2 identity matrix, '†' means transpose-conjugate, and the elements of $\mathbf{U}$, u_{ij}, satisfy $|u_{00}| = |u_{01}| = |u_{10}| = |u_{11}|$. Let $\{\mathbf{U}_k, 0 \leq k < m\}$ be a set of any m of these matrices. Define $\mathbf{M} = \mathbf{U}_0 \otimes \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_{m-1}$. Let $N = 2^m$ and let $(\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1})$ be any T-CSS of length N constructed by Theorem 5. Finally, let $\mathbf{A}_i^M = \mathbf{M}\mathbf{a_i}$ be the N-point spectrum of $\mathbf{a}$ with respect to $\mathbf{M}\mathbf{a}$ with elements, $A_{k,i}^M$, $0 \leq k < N$. Then:

$$\sum_{i=0}^{T-1} |A_{k,i}^M|^2 = TN \tag{19}$$

Theorem 6 implies that

$$|A_{k,i}^M|^2 \leq TN \quad \forall k, i, T, \mathbf{M} \tag{20}$$

The combined set of all rows of all possible transform matrices, $\mathbf{M}$, includes the one-dimensional DFT and the *Walsh–Hadamard Transform* (WHT), along with infinitely many other transforms.

As an example of the application of this result, recall that cryptographers typically perform linear cryptanalysis of cipher components by looking for peaks in the WHT

spectrum, (see, e.g., Ref. 27). It is known that for m even, length 2^m GCPs are bent, that is, have a completely flat WHT spectrum [28]. Reference [26] shows that Golay constructions generate a large set of GCPs and CSS that also have a relatively flat spectrum with respect to the WHT, among other transforms. So Golay CSS may have applications to cryptography.

5.4. CSS Mates

Let $\mathbf{A} = (\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1})$ and $\mathbf{B} = (\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_{T-1})$ be two T-CSS. Then $\mathbf{A}$ and $\mathbf{B}$ are called "mates" if $\mathbf{a}_i$ and $\mathbf{b}_i$ are orthogonal as vectors for each i. We say that $\mathbf{A}$ and $\mathbf{B}$ are "mutually orthogonal CSS" (although, in general, $\mathbf{a}_i$ is not orthogonal to $\mathbf{b}_j$, $i \neq j$). Sets $(\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_{U-1})$ of pairwise mutually orthogonal CSS can be recursively constructed in a similar way to CSS [23].

6. COMPLEMENTARY SEQUENCES OVER LARGER ALPHABETS

Virtually all the symmetries and constructions mentioned so far for bipolar sequences can be generalized to sequences over other alphabets, but note that autocorrelation is now modified to include conjugacy, i.e., Eq. (1) is modified to,

$$\rho_{\mathbf{a}}(k) = \sum_{i=0}^{N-k-1} a_i a_{i+k}^*, \quad 0 \leq k \leq N-1 \tag{21}$$

where * means "complex conjugate."

For quadriphase CSS, the symmetry operations generate an equivalence class of up to 1,024 sequences [29]. For polyphase pairs, unlike the GCP case, there is no restriction that the length must be the sum of two squares. Sivaswamy and Frank investigated and discovered many polyphase CSS, including those of odd length [30,31]. The simplest polyphase T-CSS of length T is formed from the rows of the T-point DFT matrix. In fact, from Theorem 3, the rows of any $T \times T$ orthogonal polyphase matrix form a T-CSS of length T. Sivaswamy [30] identified a length 3 quadriphase primitive pair, (002, 010) (where $0, 1, 2, 3$ mean i^0, i^1, i^2, i^3, respectively), derived quadriphase versions of GCPs and synthesized sequence pairs of lengths 3.2^k. Frank [31] further presented the following primitive quadriphase Golay pairs, of lengths 5 and 13: (01321, 00013), and (0001200302031, 0122212003203). Note that the lengths here, 5 and 13, are half the length of the lengths 10 and 26 primitive bipolar GCPs, but no transform is known between the sets.

Davis and Jedwab [4] and Paterson [7] constructed many CSS with phase alphabet 2^h and any even phase alphabet, respectively. To do so, they worked with nonbinary generalizations of the Reed–Muller codes. The resulting sequences have application to OFDM with nonbinary modulation.

Many polyphase 3-CSS exist. For example, Frank [31] presented the triphase 3-CSS (01110,11210,00201) and provided a (possibly nonexhaustive) list of lengths, N, for which a CSS exists, N up to 100: Polyphase 3-CSS: 1–22,24–27,30,32,33,36,37,39–42,45,48,49,51–54,57,58, 60,61,63–66,72,73,75,78,80,81,90,96,97,100 Polyphase 4-CSS: All lengths except 71,89 $T > 4$: All lengths.

A recent exhaustive search [29] found *all* quadriphase Golay pairs up to length $N = 13$. These are summarized below where only those for which no construction is known have been counted. These are the *possible* primitive pairs. The figures also omit the GCPs that are a subset of quadriphase pairs:

Golay pairs over the alphabet $\{0, 1, -1\}$ have been found for all lengths, N. For such a set, the weight of the set becomes an important extra parameter. The weight W is the sum of the in-phase AACF coefficients, i.e., for a set $(\mathbf{a}_j)$, we have $W = \sum_j \rho_0(\mathbf{a}_j)$. For example, here is a 4-CSS of weight 7 and length 7 over the alphabet $\{0, 1, -1\}$:

$$\{+000+00,\ 0+0+0-0,\ 00+0000,\ 000000+\}$$

The larger W, the closer is the CSS to one over a bipolar alphabet.

Yet more CSS can be found by considering multilevel and QAM alphabets.

7. HADAMARD MATRICES FROM COMPLEMENTARY SEQUENCES

In Section 5, we showed that Hadamard matrices can be used to construct CSS. The converse is also true: CSS can be used to construct Hadamard matrices. Here, we present the two best-known constructions, where Theorems 7 and 8 use the *periodic* complementary property of a complementary set (see Section 8).

Theorem 7 [32]. Let $(\mathbf{a}, \mathbf{b})$ be a GCP of length N. Let $\mathbf{A}$ and $\mathbf{B}$ be $N \times N$ circulant matrices with first rows $\mathbf{a}$ and $\mathbf{b}$, respectively. Then,

$$\begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B}^{\mathbf{T}} & \mathbf{A}^{\mathbf{T}} \end{pmatrix} \text{ is a } 2N \times 2N \text{ Hadamard matrix}$$

Theorem 7 can be generalized to quadriphase Hadamard matrices by making $(\mathbf{a}, \mathbf{b})$ a quadriphase Golay pair and by substituting conjugation for transpose.

Theorem 8 [14,15]. Let $(\mathbf{u}, \mathbf{v})$ and $(\mathbf{x}, \mathbf{y})$ be GCPs of lengths N_0 and N_1, respectively. Then, $\mathbf{a} = \mathbf{u} \mid \mathbf{x}$, $\mathbf{b} = \mathbf{u} \mid -\mathbf{x}$, $\mathbf{c} = \mathbf{v} \mid \mathbf{y}$, and $\mathbf{d} = \mathbf{v} \mid -\mathbf{y}$ form a length $N = N_0 + N_1$ 4-CSS. Let $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ be $N \times N$ circulant matrices with first rows $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$, respectively. Let $\mathbf{R}$ be a back-circulant $N \times N$ permutation matrix. Then,

$$\begin{pmatrix} \mathbf{A} & -\mathbf{BR} & -\mathbf{CR} & -\mathbf{DR} \\ \mathbf{BR} & \mathbf{A} & -\mathbf{D}^{\mathbf{T}}\mathbf{R} & \mathbf{C}^{\mathbf{T}}\mathbf{R} \\ \mathbf{CR} & \mathbf{D}^{\mathbf{T}}\mathbf{R} & \mathbf{A} & -\mathbf{B}^{\mathbf{T}}\mathbf{R} \\ \mathbf{DR} & -\mathbf{C}^{\mathbf{T}}\mathbf{R} & \mathbf{B}^{\mathbf{T}}\mathbf{R} & \mathbf{A} \end{pmatrix} \text{ is a } 4N \times 4N \text{ Goethals–Seidel (Hadamard) matrix.}$$

8. PERIODIC AND ODD-PERIODIC (NEGAPERIODIC) COMPLEMENTARY SEQUENCES

Researchers have recently become interested in constructing *periodic* and/or *odd-periodic* CSS. Periodic CSS were considered in Ref. 33.

Using polynomial form, periodic autocorrelation (PACF) of the length N sequence $\mathbf{a}$ can be expressed as

$$\mathrm{PACF}(a(x)) = \langle a(x)a(x^{-1})\rangle_{x^N-1}$$

where '$\langle * \rangle_M$' reduces $*$ mod M.

Similarly, negaperiodic (odd-periodic) autocorrelation (NACF) can be expressed as

$$\mathrm{NACF}(a(x)) = \langle a(x)a(x^{-1})\rangle_{x^N+1}$$

There is a simple relationship relating aperiodic, periodic, and odd-periodic AACF, as follows:

Let $a(x)$ represent a length N sequence. Then, the AACF of $\mathbf{a}$ can be computed, via the Chinese remainder theorem, as

$$a(x)a(x^{-1}) = \frac{x^N+1}{2}\langle a(x)a(x^{-1})\rangle_{x^N-1} - \frac{x^N-1}{2}\langle a(x)a(x^{-1})\rangle_{x^N+1} \quad (22)$$

Equation (22) expresses AACF in terms of PACF and NACF. It follows that

- A T-CSS is also a periodic and a negaperiodic T-CSS.
- A set of length T sequences is only a T-CSS if it is both a periodic and negaperiodic T-CSS.

Periodic and negaperiodic CSS can be used instead of, say, m-sequences, for their desirable correlation and spectral properties. They are much easier to find than aperiodic CSS due to their algebraic structure via embedding in a finite polynomial ring. Moreover, in a search for aperiodic CSS, an initial sieve can be undertaken by first identifying periodic and negaperiodic CSS and then looking for the intersection of the two sets.

It is known that periodic GCP do not exist for lengths 36 and 18, respectively.

We have the following theorem.

Theorem 9 [34]. If a length N periodic GCP exists, such that $N = p^{2l}u$, $p \neq u$, p prime, $p = 3 (\mathrm{mod}\, 4)$, then $u \geq 2p^l$.

Dokovic [24] discovered a periodic GCP of length 34. This is significant because no aperiodic GCP exists at that length. The next unresolved case for periodic GCPs is at length 50. References 33 and 24 also present many T-CSS for $T > 2$.

Lüke [35] has found many odd-periodic GCPs for even lengths N where an aperiodic GCP cannot exist, e.g., $N = \frac{p^{u+1}}{2}$, p an odd prime, and also for all even N, $N < 50$. He did not find any odd-length pairs.

BIOGRAPHIES

Matthew G. Parker received a B.Sc. in electrical and electronic engineering in 1982 from University of Manchester Institute of Science and Technology, U.K. and, in 1995, a Ph.D. in residue and polynomial residue number systems from University of Huddersfield, U.K. From 1995 to 1998 he was a postdoctoral researcher in the Telecommunications Research Group at the University of Bradford, U.K., researching into coding for peak factor reduction in OFDM systems. Since 1998 he has been working as a postdoctoral researcher with the Coding and Cryptography Group at the University of Bergen, Norway. He has published on residue number systems, number-theoretic transforms, complementary sequences, sequence design, quantum entanglement, coding for peak power reduction, factor graphs, linear cryptanalysis, and VLSI implementation of Modular arithmetic.

Kenneth G. Paterson received a B.Sc. (Hons) degree in mathematics in 1990 from the University of Glasgow and a Ph.D. degree, also in mathematics, from the University of London in 1993. He was a Royal Society Fellow at Institute for Signal and Information Processing at the Swiss Federal Institute of Technology, Zurich, from 1993 to 1994, investigating algebraic properties of block chipers. He was then a Lloyd's of London Tercentenary Foundation Fellow at Royal Holloway, University of London from 1994 to 1996, working on digital signatures. He joined the mathematics group at Hewlett-Packard Laboratories Bristol in 1996, becoming project manager of the Mathematics, Cryptography and Security Group in 1999. While at Hewlett-Packard, he worked on a wide variety of pure and applied problems in cryptography and communications theory. In 2001, he joined the Information Security Group at Royal Holloway, University of London, becoming a Reader in 2002. His areas of research are sequences, the mathematics of communications, and cryptography and cryptographic protocols.

C. Tellambura received his B.Sc. degree with honors from the University of Moratuwa, Sri Lanka, in 1986; his M.Sc. in electronics from the King's College, U.K., in 1988; and his Ph.D. in electrical engineering from the University of Victoria, Canada, in 1993. He was a postdoctoral research fellow with the University of Victoria and the University of Bradford. Currently, he is a senior lecturer at Monash University, Australia. He is an editor for the *IEEE Transactions on Communications* and the *IEEE Journal on Selected Areas in Communications* (Wireless Communications Series). His research interests include coding, communications theory, modulation, equalization, and wireless communications.

BIBLIOGRAPHY

1. M. J. E. Golay, Multislit spectroscopy, *J. Opt. Soc. Amer.* **39**: 437–444 (1949).
2. M. J. E. Golay, Complementary series, *IRE Trans. Inform. Theory* **IT-7**: 82–87 (1961).
3. M. Nazarathy et al., Jr., Real-time long range complementary correlation optical time domain reflectometer, *IEEE J. Lightwave Technol.* **7**: 24–38 (1989).
4. J. A. Davis and J. Jedwab, Peak-to-mean power control in OFDM, Golay complementary sequences and Reed-Muller codes, *IEEE Trans. Inform. Theory* **IT-45**: 2397–2417 (1999).

5. R. D. J. van Nee, OFDM codes for peak-to-average power reduction and error correction, in *IEEE Globecomm 1996* (London, U.K., Nov. 1996), pp. 740–744.
6. H. Ochiai and H. Imai, Block coding scheme based on complementary sequences for multicarrier signals, *IEICE Trans. Fundamentals* **E80-A**: 2136–2143, (1997).
7. K. G. Paterson, Generalized Reed-Muller codes and power control in OFDM modulation, *IEEE Trans. Inform. Theory* **IT-46**: 104–120 (2000).
8. P. Spasojevic and C. N. Georghiades, Complementary sequences for ISI channel estimation, *IEEE Trans. Inform. Theory* **IT-47**: 1145–1152 (2001).
9. H. S. Shapiro, Extremal Problems for Polynomials, M.S. Thesis, M.I.T., 1951.
10. J. Beck, Flat polynomials on the unit circle — note on a problem of Littlewood, *Bull. London Math. Soc.* **23**: 269–277 (1991).
11. J. Kahane, Sur les polynomes à coefficients unimodulaires, *Bull. London Math. Soc.* **12**: 321–342 (1980).
12. J. E. Littlewood, On polynomials $\sum \pm z^m$, $\sum \exp(\alpha_m) z^m$, $z = e^{i\theta}$, *J. London Math. Soc.* **41**: 367–376 (1966).
13. W. Rudin, Some theorems on Fourier coefficients, *Proc. Amer. Math. Soc.* **10**: 855–859 (1959).
14. R. Turyn, Hadamard matrices, Baumert-Hall units, four-symbol sequences, pulse compression, and surface wave encodings, *J. Comb. Theory Ser. A* **16**: 313–333 (1974).
15. C. H. Yang, Hadamard matrices, finite sequences, and polynomials defined on the unit circle, *Math. Comp.* **33**: 688–693 (1979).
16. S. Eliahou, M. Kervaire, and B. Saffari, A new restriction on the lengths of Golay complementary sequences, *J. Comb. Theory Ser. A* **55**: 49–59 (1990).
17. P. Borwein and R. A. Ferguson, A complete description of Golay pairs for lengths up to 100, *in preparation*, preprint available [Online]. Simon Fraser University, *http://www.cecm.sfu.ca/~pborwein/* [May, 2002].
18. M. J. E. Golay, Sieves for low autocorrelation binary sequences, *IEEE Trans. Inform. Theory* **IT-23**: 43–51 (1977).
19. T. Høholdt, H. E. Jensen, and J. Justesen, Aperiodic correlations and merit factor of a class of binary sequences, *IEEE Trans. Inform. Theory* **IT-31**: 549–552 (1985).
20. T. Høholdt and H. E. Jensen, Determination of the merit factor of Legendre sequences, *IEEE Trans. Inform. Theory* **IT-34**: 161–164 (1988).
21. P. Borwein and M. Mossinghoff, Rudin-Shapiro like polynomials in L_4, *Math. Comp.* **69**: 1157–1166 (2000).
22. S. Z. Budisin, Efficient pulse compressor for Golay complementary sequences, *Elec. Lett.* **27**: 219–220 (1991).
23. C.-C. Tseng and C. L. Liu, Complementary sets of sequences, *IEEE Trans. Inform. Theory* **IT-18**: 644–651 (1972).
24. D. Z. Dokovic, Note on periodic complementary sets of binary sequences, *Designs, Codes and Cryptography* **13**: 251–256 (1998).
25. K. Feng, P. J.-S. Shiue, and Q. Xiang, On aperiodic and periodic complementary binary sequences, *IEEE Trans. Inf. Theory* **45**: 296–303 (1999).
26. M. G. Parker and C. Tellambura, A construction for binary sequence sets with low peak-to-average power ratio, Int. Symp. Inform. Theory, Lausanne, Switzerland, June 30–July 5, 2002.
27. M. Matsui, Linear cryptanalysis method for DES, in *Advances in Cryptology — Eurocrypt93*, Lecture Notes in Computer Science, Vol. 765, pp. 386–397, Springer, 1993.
28. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, Amsterdam, North-Holland, 1977.
29. W. H. Holzmann and H. Kharaghani, A computer search for complex Golay sequences, *Aust. J. Comb.* **10**: 251–258 (1994).
30. R. Sivaswamy, Multiphase complementary codes, *IEEE Trans. Inform. Theory* **IT-24**: 546–552 (1978).
31. R. L. Frank, Polyphase complementary codes, *IEEE Trans. Inform. Theory* **IT-26**: 641–647 (1980).
32. C. H. Yang, On Hadamard matrices constructible by circulant submatrices, *Math. Comp.* **25**: 181–186 (1971).
33. L. Bomer and M. Antweiler, Periodic complementary binary sequences, *IEEE Trans. Inform. Theory* **IT-36**: 1487–1494 (1990).
34. K. T. Arasu and Q. Xiang, On the existence of periodic complementary binary sequences, *Designs, Codes and Cryptography* **2**: 257–262 (1992).
35. H. D. Lüke, Binary odd-periodic complementary sequences, *IEEE Trans. Inform. Theory* **IT-43**: 365–367 (1997).

GOLD SEQUENCES

HABONG CHUNG
Hongik University
Seoul, Korea

1. INTRODUCTION

In such applications as CDMA communications, ranging, and synchronization, sequences having good correlation properties have played an important role in signal designs. When a single sequence is used for the purpose of synchronization, it must possess a good autocorrelation property so that it can be easily distinguished from its time-delayed versions. Similarly, when a set of sequences are used for CDMA communications systems or multitarget ranging systems, the sequences should exhibit good cross-correlation properties so that each sequence is easy to distinguish from every other sequence in the set. Many individual sequences as well as families of sequences with desirable correlation properties have been found and reported. The Gold sequence family [1] is one of the oldest and best-known families of binary sequences with optimal correlation properties. This article focuses on the binary Gold sequences and their relatives. After briefly reviewing basic definitions, some of the well-known bounds on the correlation magnitude, and cross-correlation properties of m sequences in Sections 2 and 3, we discuss the Gold sequences and Gold-like sequences in Section 4.

A longer and more detailed overview on the subject of the well-correlated sequences in general can be found in Chapter 21 of the book by Pless and Huffman [11]. The reader can also find the article by Sarwate and Pursley [3] to be an excellent survey.

2. PRELIMINARIES

Given two complex-valued sequences $a(t)$ and $b(t)$, $t = 0, 1, \ldots, N-1$, of length N, the (periodic) cross-correlation function $R_{a,b}(\tau)$ of the sequences $a(t)$ and $b(t)$ is defined as follows

$$R_{a,b}(\tau) = \sum_{t=0}^{N-1} a(t+\tau)b^*(t) \tag{1}$$

where the asterisk $(*)$ denotes complex conjugation and the sum $t+\tau$ is computed modulo n. When $a(t) = b(t)$, $R_{a,a}(\tau)$ is called the *autocorrelation function* of the sequence $a(t)$, and will be denoted by $R_a(\tau)$.

Here, for the sake of simplicity and for practical reasons, we will restrict our discussion mainly on sequences whose symbols are qth roots of unity for some integer q. In this situation, it is manifest that $R_a(0) = N$ for a sequence $a(t)$ of length N. In applications such as ranging systems, radar systems, and CDMA systems, one may need a set of sequences such that the magnitude of the cross-correlation function between any two sequences in the set as well as that of the out-of-phase autocorrelation function of each sequence in the set must be small compared to N. More precisely, let $\mathcal{S}$ be a set of M cyclically distinct sequences of length N whose symbols are qth roots of unity given by

$$\mathcal{S} = \{s_0(t), s_1(t), \ldots, s_{M-1}(t)\}$$

Then for the given set $\mathcal{S}$, we can define the peak out-of-phase autocorrelation magnitude θ_a and the peak cross-correlation magnitude θ_c as follows

$$\theta_a = \max_i \max_{1 \le l \le N-1} |R_{s_i}(l)|$$

and

$$\theta_c = \max_{i \ne j} \max_{0 \le l \le N-1} |R_{s_i,s_j}(l)|$$

The maximum of θ_a and θ_c is called the *maximum correlation parameter* $\theta_{\max}$ of the set $\mathcal{S}$:

$$\theta_{\max} = \max\{\theta_a, \theta_c\}$$

Conventionally, by the term "set of sequences with optimal correlation property," we imply the set $\mathcal{S}$, $\theta_{\max}$ of which is the smallest possible for a given length N and the set size M. Certainly, $\theta_{\max}$ of a given set must be the function of sequence length N and the set size M. For example, Sarwate [2] shows that

$$\left(\frac{\theta_c^2}{N}\right) + \frac{N-1}{N(M-1)}\left(\frac{\theta_a^2}{N}\right) \ge 1 \tag{2}$$

From Eq. (2), one can obtain a simple lower bound on $\theta_{\max}$ as

$$\theta_{\max} \ge N\sqrt{\frac{M-1}{NM-1}} \tag{3}$$

Other than the bound in (3), various lower bounds on $\theta_{\max}$ in terms of N and M are known. The following bound is due to Welch [8].

Theorem 1: Welch Bound. Given a set of M complex-valued sequences of length N whose in-phase autocorrelation magnitude is N, and for an integer $k(\ge 1)$, we obtain

$$(\theta_{\max})^{2k} \ge \frac{1}{(MN-1)}\left\{\frac{MN^{2k+1}}{\binom{k+N-1}{N-1}} - N^{2k}\right\} \tag{4}$$

Especially when the sequence symbols are complex qth roots of unity, the Sidelnikov [9] bound given as the following theorem is known.

Theorem 2: Sidelnikov Bound. In the case $q = 2$, then

$$\theta_{\max}^2 > (2k+1)(N-k) + \frac{k(k+1)}{2} - \frac{2^k N^{2k+1}}{M(2k)!\binom{N}{k}},$$

$$0 \le k < \frac{2}{5}N \tag{5}$$

In the case $q > 2$, then

$$\theta_{\max}^2 > \left(\frac{k+1}{2}\right)(2N-k) - \frac{2^k N^{2k+1}}{M(k!)^2\binom{2N}{k}}, \quad k \ge 0 \tag{6}$$

3. CROSS-CORRELATION OF *m* SEQUENCES

A maximal-length linear feedback shift register sequence (m sequence) may be the best-known sequence with an ideal autocorrelation property. Like other finite field sequences, m sequences are best described in terms of the trace function over a finite field.

Let F_{q^n} be the finite field with q^n elements. Then the trace function from F_{q^n} to F_{q^m} is defined as

$$\mathrm{Tr}_m^n(x) = \sum_{i=0}^{n/m-1} x^{q^{m \cdot i}}$$

where $x \in F_{q^n}$ and $m \mid n$. The trace function satisfies the following:

1. $\mathrm{Tr}_m^n(ax+by) = a\mathrm{Tr}_m^n(x) + b\mathrm{Tr}_m^n(y)$, for all $a, b \in F_{q^m}$, $x, y \in F_{q^n}$.
2. $\mathrm{Tr}_m^n(x^{q^m}) = \mathrm{Tr}_m^n(x)$, for all $x \in F_{q^n}$.
3. Let k be an integer such that $m|k|n$. Then

$$\mathrm{Tr}_m^n(x) = \mathrm{Tr}_m^k(\mathrm{Tr}_k^n(x)) \qquad \text{forall} \quad x \in F_{q^n}$$

A q-ary m sequence $s(t)$ of length $q^n - 1$ can be expressed as

$$s(t) = \mathrm{Tr}_1^n(a\alpha^t) \tag{7}$$

where a is some nonzero element in F_{q^n} and α is a primitive element of F_{q^n}. Note that the m sequence in (7) is not complex-valued. Its symbols are the elements of the finite field F_q. When q is a prime, the natural way of converting this finite-field sequence $s(t)$ into a complex-valued sequence is taking $\omega^{s(t)}$, where ω is the primitive

qth root of unity, $e^{j2\pi/q}$. For example, an m sequence $s(t)$ of length 7 is given by

$$s(t) = \mathrm{Tr}_1^3(\alpha^t) = 1001011$$

when α is the primitive element of F_8 having minimal polynomial $x^3 + x + 1$. This m sequence $s(t)$ is easily converted to its complex-valued counterpart:

$$(-1)^{s(t)} = - + + - + - -$$

An m sequence possesses many desirable properties such as balance property, run property, shift-and-add property, and ideal autocorrelation property [15]. Given a finite field F_{q^n}, there are $\phi(q^n - 1)/n$ cyclically distinct m sequences whose symbols are drawn from F_q. Each of them corresponds to different primitive element α values with different minimal polynomials. Thus, in other words, each of the cyclically distinct m sequences of given length can be viewed as the decimation $s(dt)$ of a given m sequence $s(t)$ by some d relatively prime to $q^n - 1$. The cross-correlation properties between an m sequence and its decimation are very important, since many sequence families, including the Gold sequence family, having optimal cross-correlation properties, are constructed from pairs of m sequences.

Now, consider the case of $q = 2$, that is, of binary m sequences. Without loss of generality, we can assume that an m sequence $s(t)$ is given by

$$s(t) = \mathrm{Tr}_1^n(\alpha^t)$$

Let $\theta_{1,d}(\tau)$ be the cross-correlation function between $s(t)$ and its d-decimation $s(dt)$, where d is some integer relatively prime to $2^n - 1$:

$$\theta_{1,d}(\tau) = \sum_{t=0}^{N-1} (-1)^{s(t+\tau)+s(dt)} \tag{8}$$

From previous research, the values $\theta_{1,d}(\tau)$ have been known for various d [11], although the complete evaluation of $\theta_{1,d}(\tau)$ for each possible d is still ongoing. One well-known result on $\theta_{1,d}(\tau)$ is that $\theta_{1,d}(\tau)$ takes on at least three distinct values as τ varies from 0 to $2^n - 2$, as long as the decimation $s(dt)$ is cyclically distinct to $s(t)$. One can obtain two examples of such decimation d from the following theorem which is in part due to Gold [1], Kasami [4], and Welch [10].

Theorem 3. Let $e = \gcd(n, k)$ and $\frac{n}{e}$ be odd. Let $d = 2^k + 1$ or $d = 2^{2k} - 2^k + 1$. Then the cross-correlation $\theta_{1,d}(\tau)$ of m sequence $\mathrm{Tr}_1^n(\alpha^t)$ and its decimated sequence $\mathrm{Tr}_1^n(\alpha^{dt})$ by d takes on the following three values:

$$\begin{cases} -1 + 2^{(n+e)/2}, & 2^{n-e-1} + 2^{(n-e-2)/2} \text{ times} \\ -1, & 2^n - 2^{n-e} - 1 \text{ times} \\ -1 - 2^{(n+e)/2}, & 2^{n-e-1} - 2^{(n-e-2)/2} \text{ times} \end{cases} \tag{9}$$

When $\theta_{1,d}(\tau)$ takes on the following three values

$$-1, -1 + 2^{\lfloor (n+2)/2 \rfloor}, -1 - 2^{\lfloor (n+2)/2 \rfloor}$$

the pair of m sequences $s(t)$ and $s(dt)$ is called a preferred pair. Note that either when $n = 2m$ or $n = 2m + 1$, the above three values become $-1, -1 + 2^{m+1}$, and $-1 - 2^{m+1}$. Theorem 3 can be applied to obtain a preferred pair as long as $n \not\equiv 0 \pmod 4$. In the case when n is odd, selecting k relatively prime to n yields a preferred pair, and in the case when $n \equiv 2 \pmod 4$, making $e = 2$ also yields a preferred pair. When $n \equiv 0 \pmod 4$, Calderbank and McGuire [12] proved the nonexistence of a preferred pair.

4. GOLD SEQUENCES AND GOLD-LIKE SEQUENCES

Consider the set $\mathcal{G}$ of $(N + 2)$ sequences constructed from two binary sequences $u(t)$ and $v(t)$ of length N given as follows:

$$\mathcal{G} = \{u(t), v(t), u(t) + v(t + i) \mid i = 0, 1, \ldots, N - 1\} \tag{10}$$

Especially when both $u(t)$ and $v(t)$ are m sequences, the cross-correlation function between any two members in the set becomes either the cross-correlation function between $u(t)$ and $v(t)$, or simply the autocorrelation function of m sequence $u(t)$ or $v(t)$, due to the shift-and-add property of m sequences. In the late 1960s, Gold used this method to construct the set called *Gold sequences family*. *Gold sequences family* is defined as the set $\mathcal{G}$ when $u(t)$ and $v(t)$ are preferred pair of m sequences of length $2^n - 1$. The cross-correlation values of the Gold sequence family can be directly computed from Theorem 3. Applying the set construction method above with $u(t) = \mathrm{Tr}_1^n(\alpha^{dt})$ and $v(t) = \mathrm{Tr}_1^n(\alpha^t)$, the pair of m sequences in Theorem 3, one can easily construct the set $\mathcal{W}$ (referred to here as the *sequences family*) of size $2^n + 1$ as follows

$$\mathcal{W} = \{w_i(t) \mid 0 \le i \le 2^n\}$$

where

$$w_i(t) = \begin{cases} \mathrm{Tr}_1^n(\alpha^{(t+i)}) + \mathrm{Tr}_1^n(\alpha^{dt}), & \text{for} \quad 0 \le i \le 2^n - 2 \\ \mathrm{Tr}_1^n(\alpha^{dt}), & \text{for} \quad i = 2^n - 1 \\ \mathrm{Tr}_1^n(\alpha^t), & \text{for} \quad i = 2^n \end{cases}$$

As mentioned in the previous section, when n is odd and $e = 1$, the two m sequences $\mathrm{Tr}_1^n(\alpha^t)$ and $\mathrm{Tr}_1^n(\alpha^{dt})$ in $\mathcal{W}$ are preferred pair, and in this case, the family $\mathcal{W}$ becomes the Gold sequence family. Then $\theta_{\max}$ is given by

$$\theta_{\max} = 2^{(n+1)/2} + 1$$

When we apply the Sidelnikov bound in Eq. (5) with $k = 1$, $N = 2^n - 1$, and $M = 2^n + 1$, we have $\theta_{\max}^2 > 2^{n+1} - 2$, specifically, $\theta_{\max} \ge 2^{(n+1)/2}$. But, since N is odd, $\theta_{\max}$ must be odd. Therefore, the Sidelnikov bound tells us that

$$\theta_{\max} \ge 2^{(n+1)/2} + 1 \tag{11}$$

which, in turn, implies that the Gold sequence family is optimal with respect to the Sidelnikov bound when n is odd. On the other hand, the Gold sequence family in the case when $n \equiv 2 \pmod 4$ is not optimal, since the actual $\theta_{\max}$ in this case is $2^{(n/2)+1}$, which is roughly $\sqrt{2}$ times

the bound in (11). Finally, when $n \equiv 0 \pmod 4$, no Gold sequence family exists, since no preferred pair exists.

The following example shows the Gold sequence family of length 31.

Example 1: Gold Sequence of Length 31. There are 32 sequences of length 31 in the set. By taking α as the primitive element in F_{2^5} having minimal polynomial $x^5 + x^2 + 1$, we have

$$w_{32}(t) = \mathrm{Tr}_1^5(\alpha^t) = 1001011001111100011011101010000$$

Setting $k = 1$ and $d = 2^k + 1 = 3$, we have

$$w_{31}(t) = \mathrm{Tr}_1^5(\alpha^{3t}) = 1111101111000101011010000110010 0$$

and $w_i(t) = w_{32}(t + i) + w_{31}(t), 0 \le i \le 30$.

In the literature, the term *Gold-like* has been used in at least two different contexts. Sarwate and Pursley [3] used this term to introduce the set $\mathcal{H}$ in the following example.

Example 2: Gold-Like Sequences in Ref. 3. Let $n = 2m \equiv 0 \pmod 4$, $N = 2^n - 1$, α be a primitive element of F_{2^n}, and $d = 1 + 2^{m+1}$. Let $s(t) = \mathrm{Tr}_1^n(\alpha^t)$ and $s_j(t) = s(dt + j)$ for $j = 0, 1, 2$. Then, the set $\mathcal{H}$ is given as follows:

$$\mathcal{H} = \left\{ s(t), s(t) + s_j(t + i) \mid j = 0, 1, 2, i = 0, 1, \ldots, \frac{N}{3} - 1 \right\}$$

Certainly, there are 2^n sequences in the set $\mathcal{H}$, and all except one are the sums of the shifted m sequence and the decimated sequence just like the members in $\mathcal{W}$. The major distinction of this set $\mathcal{H}$ with the Gold family $\mathcal{W}$ is that the decimation of the m sequence $s(t)$ by d results in three distinct subsequences $s_j(t), j = 0, 1, 2$ of period $N/3$ according to the initial decimation position j, since $\gcd(N, d) = \gcd(2^{2m} - 1, 2^{m+1} + 1) = 3$ (where gcd = greatest common divisor). Kasami [4] showed that the cross-correlation function between any two sequences in $\mathcal{H}$ takes on values in the set

$$\{-1, -1 - 2^m, -1 + 2^m, -1 - 2^{m+1}, -1 + 2^{m+1}\}$$

The set $\mathcal{H}$ in the preceding example has parameters very similar to those of the Gold sequence family for the case when $n \equiv 2 \pmod 4$. The size of the set $\mathcal{H}$ is $N + 1$ while that of Gold sequence family is $N + 2$, and $\theta_{\max}$ for both family is the same in terms of N. If the term *Gold-like* was used in this context, as it seems, then there are at least two better candidates that are entitled by this term when $n \equiv 0 \pmod 4$. Niho [13] found the following family of binary sequences.

Theorem 4. Let $n \equiv 0 \pmod 4$ and $n = 2m$. Let $d = 2^{m+1} - 1$ and $s(t) = \mathrm{Tr}_1^n(\alpha^t)$. Let the set

$$\mathcal{N} = \{s(t), s(dt), s(t + i) + s(dt) \mid i = 0, 1, \ldots, 2^n - 2\}$$

be a family of $2^n + 1$ binary sequences of length $N = 2^n - 1$. Then the cross-correlation function of the sequences in $\mathcal{N}$ takes on the following four values:

$$\{-1 + 2^{m+1}, -1 + 2^m, -1, -1 - 2^m\}$$

Note that compared to the Gold-like sequences in Example 2, the Niho family has one more sequence in the set and slightly smaller $\theta_{\max}$.

Udaya [14] introduced the family of binary sequences for even n with five-valued cross-correlation property as in the following definition and theorem.

Definition 1. For even $n = 2m$, a family $\mathcal{G}_e$ of $2^n + 1$ sequences is defined as

$$\mathcal{G}_e = \{g_e(t) \mid 0 \le i \le 2^n, 0 \le t \le 2^n - 2\}$$

where

$$g_e(t) = \begin{cases} \mathrm{Tr}_1^n(\alpha^{(t+i)}) + \sum_{k=1}^{m-1} \mathrm{Tr}_1^n(\alpha^{(2^k+1)t}) + \mathrm{Tr}_1^m(\alpha^{(2^m+1)t}), \\ \qquad\qquad \text{for} \quad 0 \le i \le 2^n - 2 \\ \sum_{k=1}^{m-1} \mathrm{Tr}_1^n(\alpha^{(2^k+1)t}) + \mathrm{Tr}_1^m(\alpha^{(2^m+1)t}), \\ \qquad\qquad \text{for} \quad i = 2^n - 1 \\ \mathrm{Tr}_1^n(\alpha^t), \ \text{for} \quad i = 2^n \end{cases}$$

Theorem 5. This theorem was proposed by Udaya [14]. For the family of sequences in (12), the cross-correlation function takes on the following values:

$$\{-1, -1 + 2^m, -1 - 2^m, -1 + 2^{m+1}, -1 - 2^{m+1}\}$$

The term *Gold-like* appeared much later, in 1994, when Boztas and Kumar [7] introduced the family of binary sequences with the three-valued cross-correlation property. They called this family *Gold-like sequences* since they are identical to Gold sequences in terms of family size, correlation parameter $\theta_{\max}$, and even the range of symbol imbalance. The following give their definition and cross-correlation values.

Definition 2. For odd $n = 2m + 1$, a family $\mathcal{G}_o$ of Gold-like sequences of period $2^n - 1$ is defined as

$$\mathcal{G}_o = \{g_i(t) \mid 0 \le i \le 2^n, 0 \le t \le 2^n - 2\}$$

where

$$g_i(t) = \begin{cases} \mathrm{Tr}_1^n(\alpha^{(t+i)}) + \sum_{k=1}^{m} \mathrm{Tr}_1^n(\alpha^{(2^k+1)t}), & \text{for} \quad 0 \le i \le 2^n - 2 \\ \sum_{k=1}^{m} \mathrm{Tr}_1^n(\alpha^{(2^k+1)t}), & \text{for} \quad i = 2^n - 1 \\ \mathrm{Tr}_1^n(\alpha^t), & \text{for} \quad i = 2^n \end{cases}$$

Theorem 6. This theorem was proposed by Boztas and Kumar [7]. The cross-correlation function of the sequences in the family $\mathcal{G}_o$ of Gold-like sequences defined in (13) takes on the following three values:

$$-1, -1 + 2^{m+1}, -1 - 2^{m+1}$$

The set construction method for the families $\mathcal{G}_e$ and $\mathcal{G}_o$ is identical to that of the Gold sequence family. In other words, the sets $\mathcal{G}_e$ and $\mathcal{G}_o$ are of the same type as $\mathcal{G}$ in Eq. (10). The difference is that in $\mathcal{G}_e$ and $\mathcal{G}_o$, the sequence $u(t)$ is the sum of many m sequences, whereas it is a single m sequence in the Gold sequence family. For this reason, the linear span of the sequences in the families $\mathcal{G}_e$ and $\mathcal{G}_o$ is much larger than that of the Gold sequences.

The Gold sequences and Gold-like sequences we have reviewed are not optimal when n is even. Here, we end this article by introducing two examples of optimal families in the case when n is even.

Example 3: Small Set of Kasami Sequences. Let $n = 2m$, $m \geq 2$, α be a primitive element of F_{2^n}, $d = 2^m + 1$ and the set

$$\mathcal{K} = \{s_b(t) = \mathrm{Tr}_1^n(\alpha^t) + \mathrm{Tr}_1^m(b\alpha^{dt}) \mid b \in F_{2^m}\}$$

be a family of 2^m binary sequences of length $N = 2^n - 1$. The cross-correlation function of the sequences in the family $\mathcal{K}$ takes on the following three values:

$$-1, -1 + 2^m, -1 - 2^m$$

The set $\mathcal{K}$ is called the *small set of Kasami sequences*. The Welch bound in (4) with $k = 1$ gives us

$$\theta_{\max} > 2^m - 1$$

when $N = 2^{2m} - 1$ and $M = 2^m$. But, since $\theta_{\max}$ in this case must be an odd integer, we have

$$\theta_{\max} \geq 2^m + 1$$

which implies that the small set of Kasami sequences is optimal with respect to the Welch bound.

"No sequence family" [6] is another example of an optimal family when n is even. It has the same size and correlation distribution as the small set of Kasami sequences.

Example 4: No Sequences. Let $n = 2m$, $m \geq 2$, α be a primitive element of F_{2^n} and $d = 2^m + 1$. Let the integer $r, 1 \leq r \leq 2^m - 1, r \neq 2^i, 1 \leq i \leq m$, satisfy $\gcd(r, 2^m - 1) = 1$. Then the set

$$\mathcal{F} = \{\mathrm{Tr}_1^m[\mathrm{Tr}_m^n(\alpha^t + b\alpha^{dt})]^r \mid b \in F_{2^m}\}$$

is called a "no sequence family." The cross-correlation function of the sequences in the family $\mathcal{F}$ takes on the following three values:

$$-1, -1 + 2^m, -1 - 2^m$$

BIOGRAPHY

Habong Chung was born in Seoul, Korea. He received the B.S. degree in 1981 in electronics from Seoul National University, Seoul, Korea, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California in 1985 and 1988, respectively. From 1988 to 1991, he was an Assistant Professor in the Department of Electrical and Computer Engineering, the State University of New York at Buffalo. Since 1991, he has been with the School of Electronic and Electrical Engineering, Hongik University, Seoul, Korea, where he is a professor. His research interests include coding theory, combinatorics, and sequence design.

BIBLIOGRAPHY

1. R. Gold, Maximal recursive sequences with 3-valued recursive cross-correlation functions, *IEEE Trans. Inform. Theory* **14**: 154–156 (Jan. 1968).
2. D. V. Sarwate, Bounds on cross-correlation and autocorrelation of sequences, *IEEE Trans. Inform. Theory* **IT-25**: 720–724 (1979).
3. D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudorandom and related sequences, *Proc. IEEE* **68**: 593–619 (1980).
4. T. Kasami, *Weight Distribution Formula for Some Class of Cyclic Codes*, Technical Report R-285 (AD 632574), Coordinated Science Laboratory, Univ. Illinois, Urbana, April 1966.
5. T. Kasami, Weight distribution of Bose-Chaudhuri-Hocquenghem codes, in *Combinatorial Mathematics and Its Applications*, Univ. North Carolina Press, Chapel Hill, NC, 1969.
6. J. S. No and P. V. Kumar, A new family of binary pseudorandom sequences having optimal correlation properties and large linear span, *IEEE Trans. Inform. Theory* **35**: 371–379 (March 1989).
7. S. Boztas and P. V. Kumar, Binary sequences with Gold-like correlation but larger linear span, *IEEE Trans. Inform. Theory* **40**: 532–537 (March 1994).
8. L. R. Welch, Lower bounds on the maximum cross correlation of signals, *IEEE Trans. Inform. Theory* **IT-20**: 397–399 (1974).
9. V. M. Sidelnikov, On mutual correlation of sequences, *Soviet Math. Dokl.* **12**: 480–483 (1979).
10. H. M. Trachtenberg, *On the Crosscorrelation Functions of Maximal Linear Recurring Sequences*, Ph.D. thesis, Univ. Southern California, 1970.
11. V. S. Pless and W. C. Huffman, *Handbook of Coding Theory*, North-Holland, 1998.
12. G. McGuire and A. R. Calderbank, Proof of a conjecture of Sarwate and Pursley regarding pairs of binary m-sequences, *IEEE Trans. Inform. Theory* **IT-41**: 1153–1155 (1995).
13. Y. Niho, *Multi-valued Cross-correlation Functions between Two Maximal Linear Recursive Sequences*, Ph.D. dissertation, Univ. of Southern California, 1972.
14. P. Udaya, *Polyphase and Frequency Hopping Sequences Obtained from Finite Rings*, Ph.D. dissertation, Dept. Electrical Engineering, Indian Institute of Technology (IIT), Kanpur, India, 1992.
15. M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread-Spectrum Communications Handbook*, McGraw-Hill, New York, 1994.
16. F. J. Macwilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, New York, 1977.

17. J. S. No, K. Yang, H. Chung, and H. Y. Song, New construction for families of binary sequences with optimal correlation properties, *IEEE Trans. Inform. Theory* **43**: 1596–1602 (Sept. 1997).

GSM DIGITAL CELLULAR COMMUNICATION SYSTEM

CHRISTIAN BETTSTETTER
CHRISTIAN HARTMANN
Technische Universität München
Institute of Communication Networks
Munich, Germany

1. INTRODUCTION AND OVERVIEW

The *Global System for Mobile Communication* (GSM) is an international standard for wireless, cellular, digital telecommunication networks. GSM subscribers can use their mobile phones almost worldwide for high-quality voice telephony and low-rate data applications. International roaming and automatic handover functions make GSM a system that supports seamless connectivity and mobility.

Work on GSM was started by the Groupe Spécial Mobile of the European CEPT (Conférence Européenne des Administrations des Postes et des Télécommunications) in 1982. The aim of this working group was to develop and standardize a new pan-European mobile digital communication system to replace the multitude of incompatible analog cellular systems existing at that time. The acronym GSM was derived from the name of this group; later it was changed to Global System for Mobile Communication.

In 1987 the prospective network operators and the national administrations signed a common memorandum of understanding, which confirmed their commitment to introducing the new system based on a comprehensive set of GSM guidelines. This was an important step for international operation of the new system. In 1989 the GSM group became a technical committee of the newly founded European Telecommunications Standards Institute (ETSI). The first set of GSM technical specifications was published in 1990, and in 1991 the first GSM networks started operation. After 2 years, more than one million users made phone calls in GSM networks. The GSM standard soon received recognition also outside Europe: At the end of 1993, networks were installed for example in Australia, Hong Kong, and New Zealand. In the following years the number of subscribers increased rapidly, and GSM was deployed in many countries on all continents. Figure 1 shows the development of the GSM subscribers worldwide and the number of networks and countries on the air.

The aim of this article is to give an overview of the technical aspects of a GSM network. We first explain the functionality of the GSM components and their interworking. Next, in Section 3, we describe the services that GSM offers to its subscribers. Section 4 explains how data is transmitted over the radio interface (frequencies, modulation, channels, multiple access, coding). Section 5 covers networking-related topics, such as mobility management and handover. Section 6 discusses security-related aspects.

2. SYSTEM ARCHITECTURE

A GSM network consists of several components, whose tasks, functions, and interfaces are defined in the standard [1]. Figure 2 shows the fundamental components of a typical GSM network. A mobile user carries a *mobile station* (MS) that can communicate over the radio interface with a *base transceiver station* (BTS). The BTS contains transmitter and receiver equipment as well as a few components for signal and protocol processing. The radio range of a BTS in Fig. 2 forms one cell. In practice, a BTS with sectorized antennas can supply several cells (typically three). Also, transcoding and rate adaption of speech, error protection coding, and link control are performed in the BTS. The essential control and protocol functions reside in the *base station controller* (BSC). It handles the allocation of radio channels, channel setup, frequency hopping, and management of handovers. Typically, one BSC controls several BTSs. The BTSs and BSC together form a *base station subsystem* (BSS).

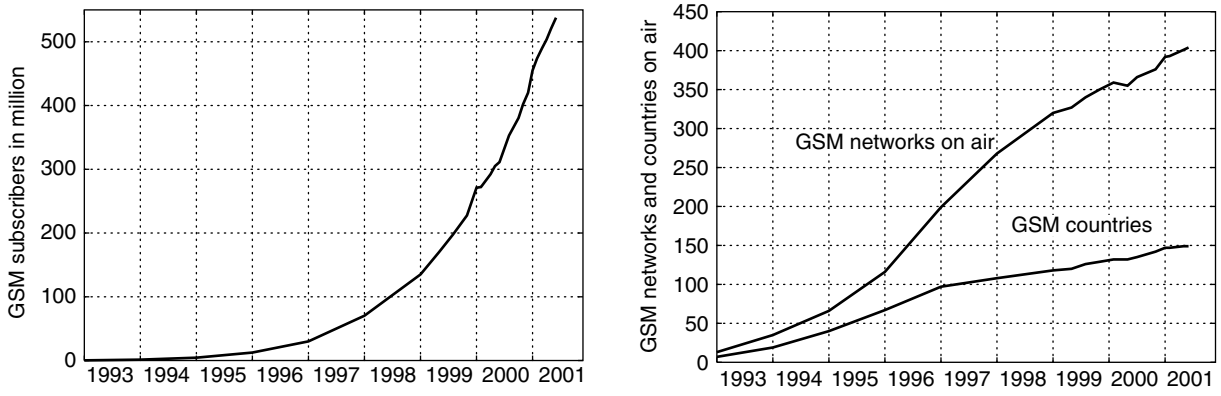


Figure 1. GSM subscriber and network statistics. (*Source*: GSM Association.)

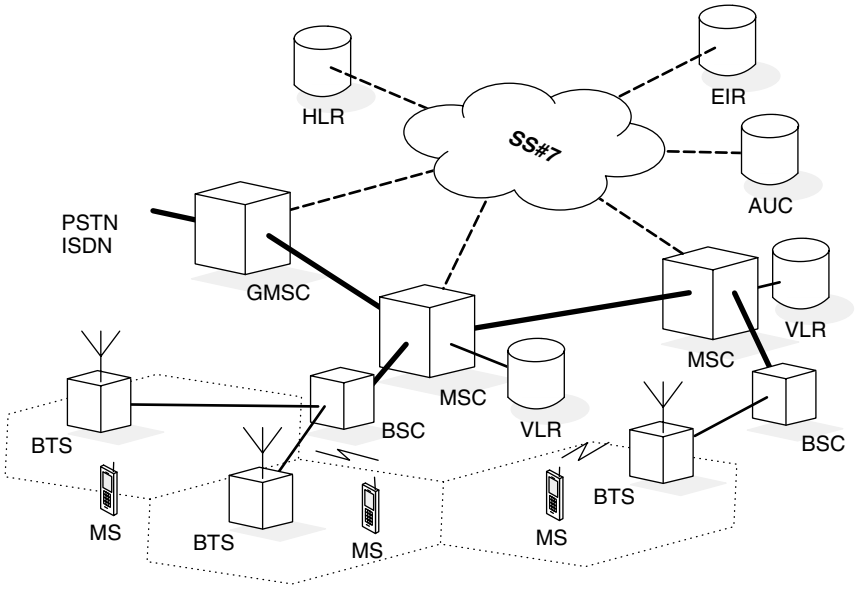


Figure 2. GSM system architecture.

The *mobile switching center* (MSC) performs the switching functions needed to route a phone call toward its destination user. Usually one MSC is allocated to several BSCs. In addition to the functionality known from usual switches in fixed ISDN networks, it must also handle the mobility of users. Such functions include the authentication of users, location updating, handover, and call routing to roaming users. Traffic between the GSM network and the fixed network [e.g., PSTN (public switched telephone network) and ISDN] is handled by a dedicated *gateway MSC* (GMSC). GSM was designed to be compatible with ISDN systems using standardized interfaces.

Two types of databases, namely, the *home location register* (HLR) and the *visitor location registers* (VLRs), are responsible for storage of profiles and location information of mobile users. There is one central HLR per network operator and typically one VLR for each MSC. The specific configuration is left to the network operator.

The HLR has a record for all subscribers registered with a network operator. It stores, for example, each user's telephone number, subscription profile, and authentication data. Besides this permanent administrative data, it also contains temporary data, such as the current location of a user. In case of incoming traffic to a mobile user, the HLR is queried in order to determine the user's current location. This allows for routing the traffic to the appropriate MSC. The mobile station must periodically inform the network about its current location. To assist this process, several cells are combined to a so-called *location area*. Whenever a mobile station changes its location area, it sends a location update to the network, indicating its current location.

A VLR is responsible for a group of location areas and stores data of all users that are currently located in this area. The data include part of the permanent subscriber data, which have been copied from the HLR to the VLR for faster access. In addition to this, the VLR may also assign and store local data, such as temporary identifiers. A user may be registered either with a VLR of his/her home network or with a VLR of a "foreign" network. On a location update, the MSC forwards the identity of the user and his/her current location area to the VLR, which subsequently updates its database. If the user has not been registered with this VLR before, the HLR is informed about the current VLR of the user.

Each user is identified by a so-called *international mobile subscriber identity* (IMSI). Together with all other personal information about the subscriber, the IMSI is stored on a chip card. This card is denoted as the *subscriber identity module* (SIM) in GSM and must be inserted into the mobile terminal in order to access the network and use the services. The IMSI is also stored in the HLR. In addition to this worldwide unique address, a mobile user receives a temporary identifier, denoted as the *temporary mobile subscriber identity* (TMSI). It is assigned by the VLR currently responsible for the user and has only local validity. The TMSI is used instead of the IMSI for transmissions over the air interface. This way, nobody can determine the identity of the subscriber.

The actual "telephone number" of a user is denoted as the *mobile subscriber ISDN number* (MSISDN). It is stored in the SIM card and in the HLR. In general, one user can have several MSISDNs.

Two further databases are responsible for various aspects of security (verification of equipment and subscriber identities, ciphering). The *authentication center* (AUC) generates and stores keys employed for user authentication and encryption over the radio channel. The *equipment identity register* (EIR) contains a list of all serial numbers of the mobile terminals, denoted as *international mobile equipment identities* (IMEI). This register allows the network to identify stolen or faulty terminals and deny network access.

As shown in Fig. 2, signaling between the GSM components in the mobile switching network is based on the Signaling System Number 7 (SS#7). For mobility-specific signaling, the MSC, HLR, and VLR hold extensions of SS#7, the so-called *mobile application part* (MAP).

Operation and maintenance of a GSM network are organized from a central *operation and maintenance center* (OMC), which is not shown in Fig. 2. Its functions include network configuration, operation, and performance management, as well as administration of subscribers and terminals.

To summarize, the entire GSM network can be divided into three major subsystems: the radio network (BSS), the switching network (including MSCs, databases, and wired core network), and the *operation and maintenance subsystem* (OSS).

3. SERVICES AND EVOLUTION

The first GSM networks mainly offered basic telecommunication services — in the first place, mobile voice telephony — and a few supplementary services. This step in the GSM evolution is called phase 1. The supplementary services of phase 1 include features for call forwarding (e.g., call forwarding on mobile subscriber busy, call forwarding on mobile subscriber not reachable) and call restriction (e.g., barring of all outgoing/incoming calls, barring of all outgoing international calls, and barring of incoming international calls when roaming outside the home network). All these services had to be implemented as mandatory features by all network operators.

Besides mobile voice telephony, GSM also offers services for data transmission, such as fax and circuit-switched access to data networks (e.g., X.25 networks) with data rates up to 9.6 kbps (Kilobits per second).

Of particular importance is the *short message service* (SMS). It allows users to exchange short text messages in a store-and-forward fashion. The network operator establishes a service center that accepts and stores text messages. Later, some value-added services, such as SMS cell broadcast and conversion of SMS messages from/to email and to speech, have been implemented.

The standardization of phase 2 basically added further supplementary services, such as call waiting, call holding, conference calling, call transfer, and calling line identification. Many parts of the GSM technical specifications had to be reworked, but all networks and terminals retained compatibility to phase 1. Phase 2 was completed in 1995, and market introduction followed in 1996.

In the following years, a broad number of additional services and improvements have been developed in independent standardization units (GSM phase 2+) [2]. These topics affect almost all aspects of GSM, and enable a smooth transition from GSM to the *Universal Mobile Telecommunication System* (UMTS); see Fig. 3. For example, the GSM voice codecs have been improved to achieve a much better speech quality (see Section 4.2). Furthermore, a set of group call and push-to-talk speech services with fast connection setup has been

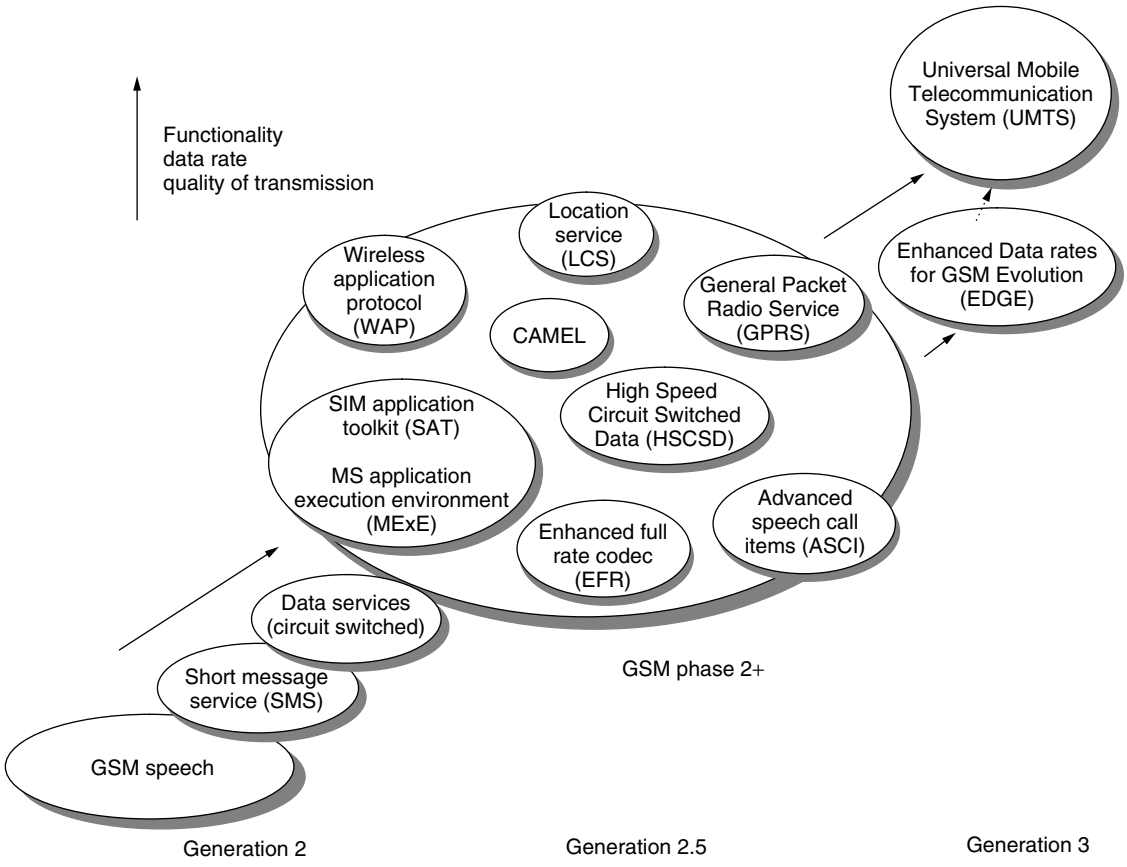


Figure 3. GSM evolution.

standardized under the name *advanced speech call items* (ASCIs). These services are especially important for closed user groups, such as police and railroad operators.

Another important aspect is the definition of service platforms. Instead of standardizing services, only features to create services and mechanisms that enable their introduction are specified. This allows the network providers to introduce new operator-specific services in a faster way. *Customized application for mobile network enhanced logic* (CAMEL) represents the integration of intelligent network (IN) principles into GSM. It enables use of operator-specific services also in foreign networks. For example, a subscriber roaming in a different country can easily check his/her voicebox with the usual short number. Other features include roaming services for prepaid subscriptions and speed dial for closed user groups. Also on the mobile station side, service platforms have been developed: the SIM application toolkit and the *MS application execution environment* (MExE). The SIM application toolkit allows the operator to run specific applications on the SIM card. With the toolkit, the SIM card is able to display new operator-specific items and logos and to play sounds. For example, users can download new ringing tones to their SIM card. The new applications can be transmitted, for example, via SMS to the mobile station. The most important components of MExE are a virtual machine for execution of Java code and the *Wireless Application Protocol* (WAP). With a virtual machine running on the mobile station, applications can be downloaded and executed. The WAP defines a system architecture, a protocol family, and an application environment for transmission and display of Web-like pages for mobile devices. WAP has been developed by the WAP forum; services and terminals have been available since 1999. Using a WAP-enabled mobile GSM phone, subscribers can download information pages, such as news, weather forecasts, stock reports, and local city information. Furthermore, mobile e-commerce services (e.g., ticket reservation, mobile banking) are offered.

If information about the current physical location of a user is provided by the GSM network or by GPS (Global Positioning System), location-aware applications are possible. A typical example is a location-aware city guide that informs mobile users about nearby sightseeing attractions, restaurants, hotels, and public transportation.

Development also continued with improved bearer services for data transmission. The *High-Speed Circuit-Switched Data* (HSCSD) service achieves higher data rates by transmitting in parallel on several traffic channels (multislot operation). The *General Packet Radio Service* (GPRS) offers a packet-switched transmission at the air interface. It improves and simplifies wireless access to the Internet. Users of GPRS benefit from shorter access times, higher data rates, volume-based billing, and an "always on" wireless connectivity. (For further details, see the GPRS entry of this encyclopedia.) The *Enhanced Data Rates for GSM Evolution* (EDGE) service achieves even higher data rates and a better spectral efficiency. It replaces the original GSM modulation by an 8-PSK (8-phase shift keying) modulation scheme.

4. AIR INTERFACE: PHYSICAL LAYER

Figure 4 gives a schematic overview of the basic elements of the GSM transmission chain. The stream of sampled speech is fed into a source encoder, which compresses the data. The resulting bit sequence is passed to the channel encoder. Its purpose is to add, in a controlled manner, some redundancy to the bit sequence. This redundancy serves to protect the data against the negative effects of noise and interference encountered in the transmission over the radio channel. On the receiver side, the introduced redundancy allows the channel decoder to detect and correct transmission errors. Without channel coding, the achieved bit error rate would be insufficient, not only for speech but also for data communication. Reasonable bit error rates are on the order of 10^{-5} to 10^{-6}. To achieve these rates, GSM uses a combination of block and convolutional coding. Moreover, an interleaving scheme is used to deal with burst errors that occur over multipath and fading channels. After coding and interleaving, the data are encrypted to guarantee secure and confident data transmission. The encryption technique as well as the methods for subscriber authentication and secrecy of the subscriber identity are explained in Section 6. The encrypted data are subsequently mapped to bursts, which are then multiplexed. Finally, the stream of bits is differentially coded, modulated, and transmitted on the respective carrier frequency over the mobile radio channel.

After transmission, the demodulator processes the signal, which was corrupted by the noisy channel. It

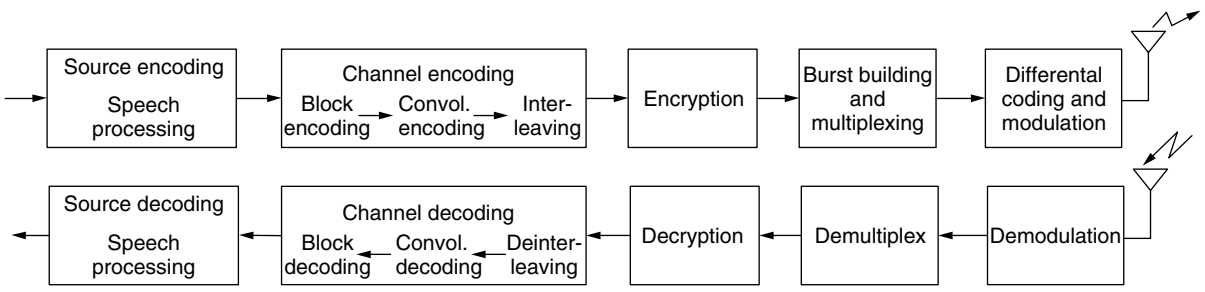


Figure 4. Basic function of the GSM transmission chain on the physical layer at the air interface.

attempts to recover the actual signal from the received signal. The next steps are demultiplexing and decryption. The channel decoder attempts to reconstruct the original bit sequence, and, as a final step, the source decoder tries to rebuild the original source signal.

4.1. Logical Channels

GSM defines a set of logical channels [3], which are divided into two categories: traffic channels and signaling channels. Some logical channels are assigned in a dedicated manner to a specific user; others are common channels that are of interest for all users in a cell.

The traffic channels (TCHs) are used for the transmission of user data (e.g., speech, fax). A TCH may either be fully used (full-rate TCH; TCH/F) or split into two half-rate channels that can be allocated to different subscribers (half-rate TCH; TCH/H). A TCH/F carries either 13 kbps of coded speech or datastreams at 14.5, 12, 6, or 3.6 kbps. A TCH/H transmits 5.6 kbps of half-rate coded speech or datastreams at 6 or 3.6 kbps.

In addition to the traffic channels, GSM specifies various signaling channels. They are grouped into broadcast channels (BCHs), common control channels (CCCHs), and dedicated control channels (DCCHs).

The BCHs are unidirectional and used to continually distribute information to all MSs in a cell. The following three broadcast channels are defined:

- The broadcast control channel (BCCH) broadcasts configuration information, including the network and BTS identity, the frequency allocations, and availability of optional features such as voice activity detection (Section 4.2) and frequency hopping (Section 4.5).
- The frequency correction channel (FCCH) is used to distribute information about correction of the transmission frequency.
- The synchronization channel (SCH) broadcasts data for frame synchronization of an MS.

The group of CCCHs consists of four unidirectional channels that are used for radio access control:

- The random-access channel (RACH) is an uplink channel that is used by the MSs in a slotted ALOHA fashion for the purpose of requesting network access.
- The access grant channel (AGCH) is a downlink channel used to assign a TCH or a DCCH to an MS.
- The paging channel (PCH), which is also a downlink channel, is employed to locate an MS in order to inform it about an incoming call.
- The notification channel (NCH) serves to inform MSs about incoming group and broadcast calls.

Finally, GSM uses three different DCCHs:

- The stand-alone dedicated control channel (SDCCH), which is applied for signaling between BSS and MS when there is no active TCH connection. This is necessary to update location information, for instance.
- The slow associated control channel (SACCH), which carries information for synchronization, power control, and channel measurements. It is always assigned in conjunction with a TCH or an SDCCH.
- The fast associated control channel (FACCH), which is used for short time signaling. It can be made available by stealing bursts from a TCH.

GSM also defines a set of logical channels for the General Packet Radio Service (GPRS). They are treated in the GPRS entry of this encyclopedia.

4.2. Speech Processing Functions and Codecs

Transmission of voice is one of the most important services in GSM. The user's analog speech signal is sampled at the transmitter at a rate of 8000 samples/s, and these samples are quantized with a resolution of 13 bits. At the input of the speech encoder, a speech frame containing 160 samples, each 13 bits long, arrives every 20 ms. This corresponds to a bit rate of 104 kbps for the speech signal. The compression of this speech signal is performed in the speech encoder. The functions of encoder and decoder are typically combined in a single building block, called a *codec*.

An optional speech processing function is *discontinuous transmission* (DTX). It allows the radio transmitter to be switched off during speech pauses. This saves battery power of the MS and reduces the overall interference level at the air interface. Voice pauses are recognized by the *voice activity detection* (VAD). During pauses, the missing speech frames are replaced by a synthetic background noise generated by the comfort noise synthesizer.

Another function on the receiver side is the replacement of bad frames. If a transmitted speech frame cannot be corrected by the channel coding mechanism, it is discarded and replaced by a frame that is predictively calculated from the preceding frame.

In the following paragraphs, the speech codecs used in GSM are briefly characterized. The channel coding is described in Section 4.3.

4.2.1. Full-Rate Speech Codec. The first set of GSM standards defined a full-rate speech codec for transmission via the TCH/F. It is an RPE-LTP (regular pulse excitation–long-term prediction) codec [4], which is based on linear predictive coding (LPC). The RPE-LTP codec has a compression rate of 1/8 and thus produces a data rate of 13 kbps at its output.

With the further development of GSM the speech codecs have also been improved. Two competing objectives have been considered: (1) the improvement of speech quality toward the quality offered by fixed ISDN networks and (2) better utilization of the frequency bands assigned to GSM, in order to increase the network capacity.

4.2.2. Half-Rate Speech Codec. The half-rate speech codec has been developed to improve bandwidth utilization. It produces a bit stream of 5.6 kbps and is used for

speech transmission over the TCH/H. Instead of using the RPE-LTP coding scheme, the algorithm is based on code-excited linear prediction (CELP), in which the excitation signal is an entry in a very large stochastically populated codebook. The codec has a higher complexity and higher latency. Under normal channel conditions, it achieves — in spite of half the bit rate — almost the same speech quality as the full-rate codec. However, quality loss occurs for mobile-to-mobile communication, since in this case (due to the ISDN architecture) one has to go twice through the GSM speech coding/decoding process. A method to avoid multiple transcoding has been passed under the name tandem free operation in GSM Release 98.

4.2.3. EFR Speech Codec. The *enhanced full-rate* (EFR) speech codec [5] was standardized by ETSI in 1996 and has been implemented in GSM since 1998. It improves the speech quality compared to the full- and half-rate speech codecs without using more system capacity than the full-rate codec. The EFR codec produces a bitstream of 12.2 kbps (244 code bits for each 20-ms frame) and is based on the algebraic code excitation linear prediction (ACELP) technique [6].

A detailed explanation of the full-rate, half-rate, and EFR codecs can be found in Ref. 4.

4.2.4. AMR Codec. The speech codecs mentioned before all use a fixed source bit rate, which has been optimized for typical radio channel conditions. The problem with this approach is its inflexibility; whenever the channel conditions are much worse than usual, very poor speech quality will result, since the channel capacity assigned to the mobile station is too small for error-free transmission. On the other hand, radio resources will be wasted for unneeded error protection if the radio conditions are better than usual. To overcome these problems, a much more flexible codec has been developed and standardized: the *adaptive multirate* (AMR) codec. It can improve speech quality by adaptively switching between different speech coding schemes (with different levels of error protection) according to the current channel quality.

To be more precise, AMR has two principles of adaptability [7]: channel mode adaptation and codec mode adaptation. Channel mode adaptation dynamically selects the type of traffic channel that a connection should be assigned to: either a full-rate (TCH/F) or a half-rate traffic channel (TCH/H). The basic idea here is to adapt a user's gross bit rate in order to optimize the usage of radio resources. The task of codec mode adaptation is to adapt the coding rate (i.e., the tradeoff between the level of error protection versus the source bit rate) according to the current channel conditions. When the radio channel is bad, the encoder operates at low source bit rates at its input and uses more bits for forward error protection. When the quality of the channel is good, less error protection is employed.

The AMR codec consists of eight different modes with source bit rates ranging from 12.2 to 4.75 kbps. All modes are scaled versions of a common ACELP basis codec; the 12.2-kbps mode is equivalent to the EFR codec.

4.3. Channel Coding

GSM uses a combination of block coding (as external error protection) and convolutional coding (as internal error protection). Additionally, interleaving of the encoded bits is performed in order to spread the channel symbol errors. The channel encoding and decoding chain is depicted in Fig. 4. The bits coming from the source encoder are first block-encoded; that is, parity and tail bits are appended to the blocks of bits. The resulting stream of coded bits is then fed into the convolutional encoder, where further redundancy is added for error correction. Finally, the blocks are interleaved. This is done because the mobile radio channel frequently causes burst errors (a sequence of erroneous bits), due to long and deep fading events. Spreading the channel bit errors by means of interleaving diminishes this statistical dependence and transforms the mobile radio channel into a memoryless binary channel. On the receiver side, the sequence of received channel symbols — after demodulation, demultiplexing, and deciphering — is deinterleaved before it is fed into the convolutional decoder for error correction. Finally, a parity check is performed based on the respective block encoding.

It depends on the channel type which specific channel coding scheme is employed. Different codes are used, for example, for speech traffic channels, data traffic channels, and signaling channels. The block coding used for external error protection is either a *cyclic redundancy check* (CRC) code or a fire code. In some cases, just some tail bits are added. For the convolutional coding, the memory of the used codes is either 4 or 6. The basic code rates are $r = \frac{1}{2}$ and $r = \frac{1}{3}$; however, other code rates can be obtained by means of puncturing (deleting some bits after encoding). In the following, the basic coding process is explained for some channels. The detailed channel coding and interleaving procedures of all channels are described in Ref. 8.

4.3.1. Full-Rate Speech Codec. Let us explain how the speech of the full-rate codec is protected. One speech block coming from the source encoder consists of 260 bits. These bits are graded into different classes, which have different impact on speech quality. The 182 bits of class 1 have more impact on speech quality and thus must be better protected. The block coding stage calculates three parity bits for the most important 50 bits of class 1 (known as class 1a bits). Next, four tail bits are added, and the resulting 189 bits are fed into a rate-$\frac{1}{2}$ convolutional encoder of memory 4 (see Fig. 5). The 78 bits of class 2 are not encoded. Finally, the resulting 456 bits are interleaved.

A frame in which the bits of class 1 have been recognized as erroneous in the receiver is reported to the speech codec using the *bad-frame indication* (BFI). In order to maintain a good speech quality, these frames are discarded, and the last frame received correctly is repeated instead, or an extrapolation of received speech data is performed.

4.3.2. EFR Speech Codec. Using the EFR codec, a special preliminary channel coding is employed for the most significant bits; 8 parity bits (generated by a CRC

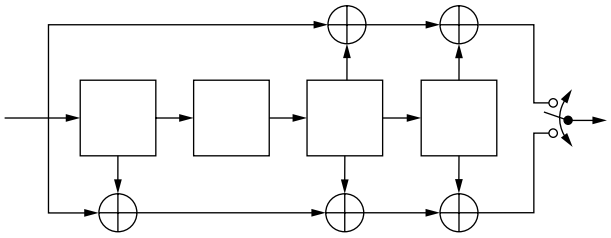


Figure 5. Convolutional encoder (used for full-rate speech, EFR speech, some data channels, and signaling channels).

encoder) and 8 repetition bits are added to the 244 bits at the output of the source encoder for additional error detection. The coding process for the resulting 260 bits is the same as for the full-rate codec.

4.3.3. AMR Speech Codec. The error protection using the AMR codec is more sophisticated. From the results of link quality measures, an adaptation unit selects the most appropriate codec mode. Channel coding is performed using a punctured recursive systematic convolutional code. Since not all bits of the voice data are equally important for audibility, AMR also employs an unequal error protection structure. The most important bits are additionally protected by a CRC code. Also, the degree of puncturing depends on the importance of the bits.

4.3.4. Data Traffic. For data traffic channels, no actual block coding is performed. Instead, a tail of all-zero bits is appended to the data blocks in order to obtain block sizes that are suitable for the convolutional encoder. In the convolutional coding stage, the channels TCH/F14.4, TCH/F9.6, TCH/H4.8 use a punctured version of the rate-$\frac{1}{2}$ encoder depicted in Fig. 5. For the TCH/F9.6 and TCH/H4.8, the 244 bit blocks at the input of the encoder are mapped to 488 bits. These blocks are reduced to 456 bits by puncturing 32 bits. Using TCH/F14.4, the 294 bits are mapped to 588 bits, followed by a puncturing of 132 bits. The channels TCH/F4.8, TCH/F2.4, and TCH/H2.4 use a rate-$\frac{1}{3}$ channel encoder.

4.3.5. Signaling Traffic. The majority of the signaling channels (SACCH, FACCH, SDCCH, BCCH, PCH, AGCH) use a fire code for error detection. It is a powerful cyclic block code that appends 40 redundancy bits to a 184-bit data block. A different approach is taken for error detection on the RACH. The very short random access burst of the RACH allows a data block length of only 8 bits, which is supplemented by six redundancy bits using a cyclic code. The SCH, as an important synchronization channel, uses a somewhat more elaborate error protection than the RACH channel. The SCH data blocks have a length of 25 bits and receive another 10 bits of redundancy for error detection through a cyclic code. The convolutional coding is performed using the encoder in Fig. 5 for all signaling channels.

4.4. Multiple Access

As illustrated in Fig. 6, GSM uses the frequency bands 890–915 MHz for transmission from the MS to the BTS (uplink) and 935–960 MHz for transmission from the BTS to the MS (downlink). Hence, the duplexing method used in GSM is *frequency-division duplex* (FDD). In addition, GSM systems developed later have been assigned frequency ranges around 1800 and 1900 MHz.

The multiple-access method uses a combination of *frequency-division multiple access* (FDMA) and *time-division multiple access* (TDMA). The entire 25-MHz frequency range is divided into 124 carrier frequencies of 200 kHz bandwidth. Each of the resulting 200 kHz radiofrequency channels is further divided into eight time slots, namely, eight TDMA channels (see Fig. 6). One timeslot carries one data burst and lasts 576.9 μs (156.25 bits with a data rate of 270.833 kbps). It can be

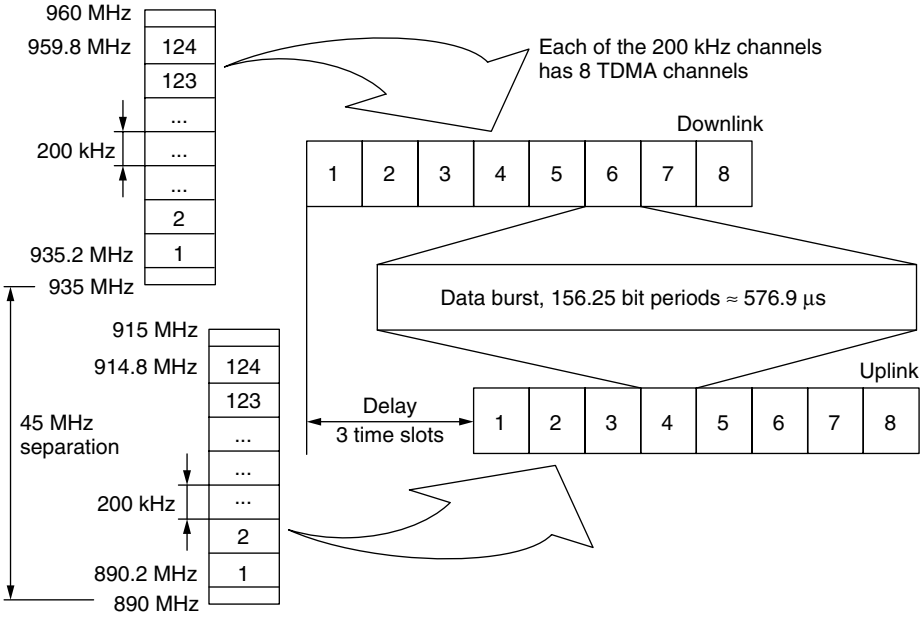


Figure 6. Carrier frequencies, duplexing, and TDMA frames.

considered the basic unit of a TDMA frame. A TDMA frame contains eight time slots and thus lasts 4.615 ms. Consequently, each TDMA channel is able to transmit at a data rate of 33.9 kbps.

Five different data bursts are distinguished in GSM [3]. A normal burst consists of two sequences of 57 data bits that are separated by a 26-bit midamble containing a training sequence. The training sequence is used for channel estimation and equalization. Furthermore, a normal burst contains two signaling bits, called "stealing flags," which indicate whether the burst contains traffic data or signaling data. Three tail bits, which are always set to logical 0, mark the beginning as well as the end of each burst. A guard period lasting the equivalent of 8.25 bits separates consecutive bursts. The other burst types are the frequency correction burst (used for frequency synchronization of an MS), the synchronization burst (used for time synchronization of the MS with the BTS), the dummy burst (used to avoid empty bursts on the BCCH in order to ensure continuous transmission on the BCCH carrier), and the access burst (used by MSs for random access on the RACH).

In order to efficiently use the bandwidth assigned to a GSM network, frequency channels have to be reused in a certain spatial distance (cellular principle). Determining a good allocation of frequencies is a complicated optimization problem, called *frequency planning*. The goal is to provide each cell with as many channels as possible, while securing that frequencies are only reused in cells that are sufficiently far apart to avoid severe cochannel interference. Assuming idealistic regular hexagonal cell patterns, frequency planning leads to dividing the system area into clusters that contain a certain number of cells (see Fig. 7). Each cluster can use all available channels of the network. Each channel can be allocated to exactly one cell of the cluster (i.e., channel reuse within a cluster is not possible). The set of frequency channels assigned to a cell is denoted as cell allocation. If each cluster of the network has the same assignment pattern, that is, if cells that have the same relative position within their respective clusters receive the same cell allocation throughout the network, a minimal cochannel reuse distance is secured in the whole network if the cluster size is reasonably chosen. Figure 7 gives an example in which 14 frequencies ($f_1 \cdots f_{14}$) have been allocated to clusters with seven cells each.

In real systems, the cluster size should be chosen high enough to avoid severe cochannel interference but not too high to obtain as many frequency channels per cell as possible. Given realistic propagation conditions and BTS positions, a typical cluster size in real GSM networks is on the order of 12. The actual frequency planning process becomes even more complicated since traffic conditions are seldom uniform and some cells will need a higher number of channels than others.

One channel of the cell allocation is used for broadcasting the configuration data on the BCCH as well as synchronization data (FCCH and SCH). This channel is called the BCCH carrier.

4.5. Frequency Hopping

As a result of multipath propagation, mobile radio channels experience frequency selective fading; that is, the instantaneous fading values that an MS experiences at a certain position are frequency-dependent. Therefore, GSM provides an optional frequency hopping procedure [9] that changes the transmission frequency periodically in order to average the interference over the frequencies in one cell. The frequency is changed with each burst, which is considered slow frequency hopping with a resulting hopping rate of about 217 changes per second. The use of frequency hopping is an option left to the network operator, which can be decided on an individual cell basis.

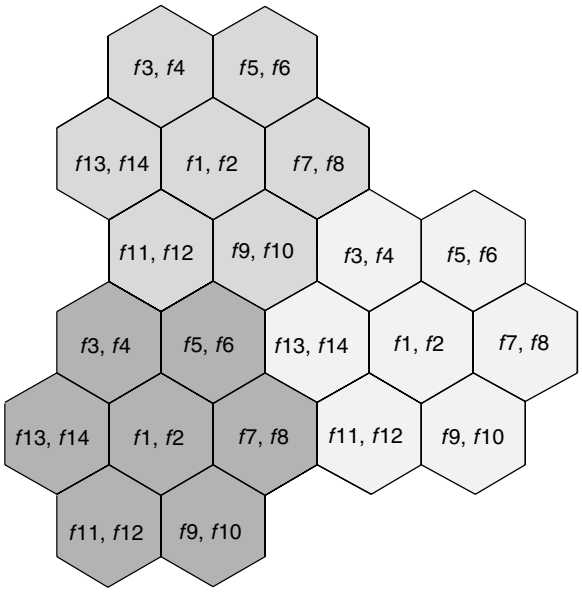


Figure 7. Cellular principle (clusters with seven cells).

4.6. Synchronization

For successful operation of a mobile radio system, synchronization between MSs and BTSs is necessary [10]. Two kinds of synchronization are distinguished: frequency synchronization and time synchronization.

Frequency synchronization is necessary so that the transmitter and receiver operate on the same frequency (in both up- and downlink). By periodically monitoring the FCCH, the MS can synchronize its oscillators with the BTS.

Time synchronization is important to adjust propagation time differences of signals from different MSs, in order to achieve synchronous reception of time slots at the BTS. This way, adjacent time slots do not overlap and interfere with each other. Furthermore, synchrony is needed for the frame structure, since there is a higher-level frame structure superimposed on the TDMA frames for multiplexing logical signaling channels onto one physical channel. To keep track of the frame structure, the MSs monitor the synchronization bursts on the SCH, which contain information on the frame number. To synchronize the BTS's reception of data bursts from all MSs within the cell, a parameter called *timing advance* is used. The mobile station receives the timing advance value that it must use from the BTS on the SACCH downlink, and it reports the currently used value on the SACCH uplink.

There are 64 steps for the timing advance, where one step corresponds to 1-bit duration. The value tells the MS when it must transmit relative to the downlink. Thus, the required adjustment always corresponds to the round-trip delay between the respective mobile station and the base station. Therefore, the maximum timing advance value defined in GSM also determines the maximum cell size (radius 35 km).

4.7. Modulation

The modulation technique used in GSM is *Gaussian minimum shift keying* (GMSK). GMSK belongs to the family of continuous-phase modulation, which have the advantages of a narrow transmitter power spectrum with low adjacent channel interference and a constant amplitude envelope. This allows for use of simple and inexpensive amplifiers in the transmitters without stringent linearity requirements. In order to facilitate demodulation, each burst is encoded differentially before modulation is performed.

4.8. Power Control

The main purpose of power control in a mobile communication system is to minimize the overall transmitted power in order to keep interference levels low, while providing sufficient signal quality for all ongoing communications.

In GSM, the transmit power of the BTS and each MS can be controlled adaptively. As part of the radio subsystem link control [11], the transmit power is controlled in steps of 2 dBm. For the uplink power control, 16 control steps are defined from step 0 (43 dBm = 20 W) to step 15 (13 dBm) with a gap of 2 dBm between neighboring values. Similarly, the downlink can be controlled in steps of 2 dBm. However, the number of downlink power control steps depends on the power class of the BTS, which defines the maximum transmission power of a BTS (up to 320 W). It should be noted that downlink power control is not applied to the BCCH carrier, which must maintain constant power to allow comparative measurements of neighboring BCCH carriers by the mobile stations.

GSM uses two parameters to describe the quality of a connection: the *received signal level* (RXLEV, measured in dBm) and the *received signal quality* (RXQUAL, measured as bit error ratio before error correction). Power control as well as handover decisions are based on these parameters. The received signal power is measured continuously by mobile and base stations in each received burst within a range of −110 to −48 dBm. The respective RXLEV values are obtained by averaging.

For power control, upper and lower threshold values for uplink and downlink are defined for the parameters RXLEV and RXQUAL. If a certain number of consecutive values of RXLEV or RXQUAL are above or below the respective threshold value, the BSS can adjust the transmitter power. For example, if the upper threshold value for RXLEV on the uplink is exceeded, the transmission power of the MS will be reduced. In the other case, if RXLEV on the uplink falls below the respective lower threshold, the MS will be ordered to increase its transmission power. If the criteria for the downlink are exceeded, the transmitter power of the BTS will be adjusted. Equivalent procedures will be performed if the values for RXQUAL violate the respective range.

5. NETWORKING ASPECTS AND PROTOCOLS

The GSM standard defines a complex protocol architecture that includes protocols for transport of user data as well as for signaling. The fact that users can roam within a network from cell to cell and also internationally to other networks, requires the GSM system to handle signaling functions for registration, user authentication, location updating, routing, handover, allocation of channels, and so on. Some signaling protocols for these tasks are explained in the following; a more comprehensive explanation can be found in Ref. 2.

Figure 8 illustrates the signaling protocol architecture between MS and MSC. Layer 1 at the air interface

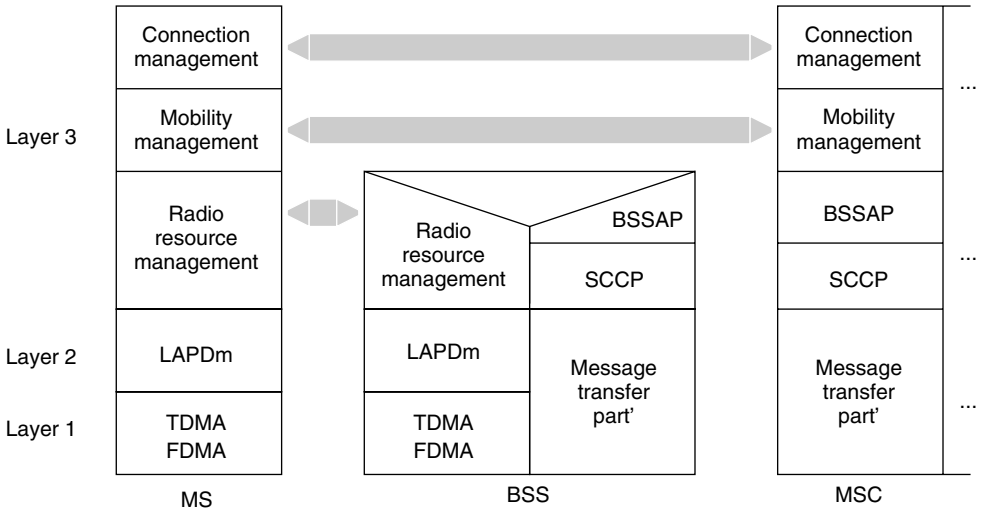


Figure 8. Signaling protocols.

(between MS and BTS) is the physical layer as described in Section 4. It implements the logical signaling channels. The data-link layer (layer 2) at the air interface employs a modified version of the LAPD (*link access procedure D*) protocol used in ISDN. It is denoted as LAPDm (LAPD mobile). The signaling protocols in Layer 3 between mobile stations and the network are divided into three sublayers: (1) radio resource management, (2) mobility management, and (3) connection management [12]. They are explained in the following subsections. In the fixed network part, a slightly modified version of the *message transfer part* (MTP) of SS#7 is used to transport signaling messages between BSC and MSC. Above this protocol, the *signaling connection control part* (SCCP) and the *BSS application part* (BSSAP) have been defined. The latter consists of the *direct transfer application part* (DTAP) and the *BSS mobile application part* (BSSMAP).

5.1. Radio Resource Management

The objective of radio resource management is to set up, maintain, and release traffic and signaling channels between a mobile station and the network. This also includes cell selection and handover procedures.

A mobile station is continuously monitoring the BCCH and CCCH on the downlink. It measures the signaling strength of the BCCHs broadcasted by the nearby BTSs in order to select an appropriate cell. At the same time, the MS periodically monitors the PCH for incoming calls.

A radio resource management connection establishment can be initiated either by the network or by the MS. In either case, the MS sends a channel request on the RACH in order to get a channel assigned on the AGCH. In case of a network-initiated connection, this procedure is preceded by a paging call to be answered by the MS.

Once a radio resource management connection has been set up, the MS has either an SDCCH or a TCH with associated SACCH/FACCH available for exclusive bidirectional use. On the SACCH the MS continuously sends channel measurements if no other messages need to be sent. These measurements include the values RXLEV and RXQUAL (see Section 4.8) of the serving cell and RXLEV of up to six neighboring cells. The system information sent by the BSS on the SACCH downlink contains information about the current and neighboring cells and their BCCH carriers.

In order to change the configuration of the physical channel in use, a channel change within the cell can be performed. The channel change can be requested by the radio resource management sublayer or by higher protocol layers. However, it is always initiated by the network and reported to the MS by means of an assignment command. A second signaling procedure to change the physical channel configuration of an established radio resource management connection is a handover, which is described in Section 5.4.

The release of radio resource management connections is always initiated by the network. Reasons for the channel release could be the end of the signaling transaction, insufficient signal quality, removal of the channel in favor of a higher priority call (e.g., emergency call), or the end of a call. After receiving the channel release command, the MS changes back to idle state.

Another important procedure of GSM radio resource management is the activation of ciphering, which is initiated by the BSS by means of the cipher mode command.

5.2. Mobility Management and Call Routing

Mobility management includes tasks that are related to the mobility of a user. It keeps track of a user's current location (location management) and performs attach and detach procedures, including user authentication.

Before a subscriber can make a phone call or use other GSM services, his/her mobile station must attach to the network and register its current location. Usually, the subscriber will attach to its home network, that is, the network with which he/she has a contract. Attachment to other networks is possible if there is a roaming agreement between the providers. To perform an attach, the MS sends the IMSI of the user to the current network. The MSC/VLR queries the HLR to check whether the user is allowed to access the network. If authentication is successful, the subscriber will be assigned a TMSI and an MSRN (*mobile station roaming number*) for further use. The TMSI is only valid within a location area. The MSRN contains routing information, so that incoming traffic can be routed to the appropriate MSC of the user.

After a successful attach, GSM's location management functions are responsible for keeping track of a user's current location, so that incoming calls can be routed to the user. Whenever a powered-on MS crosses the boundary of a location area, it sends a location update request message to its MSC. The VLR issues a new MSRN and informs the HLR about the new location. As opposed to the location registration procedure, the MS has already got a valid TMSI in this case. In addition to this event-triggered location update procedure, GSM also supports periodic location updating.

The telephone number dialed to reach a user (MSISDN) gives no information about the current location of this user. Clearly, we always dial the same number no matter whether our communication partner is attached to its home network or he/she is currently located in another country. To establish a connection, the GSM network must determine the cell in which the user resides.

Figure 9 gives an example for an incoming call from the fixed ISDN network to a mobile user. An ISDN switch realizes that the dialed number corresponds to a subscriber in a mobile GSM network. From the country and network code of the MSISDN it can determine the home network of the called user and forward the call to the appropriate GMSC. On arrival of the call, the GMSC queries the HLR in order to obtain the current MSRN of the user. With this number, the call can be forwarded to the responsible MSC. Subsequently, the MSC obtains the user's TMSI from its VLR and sends out a paging message for the user in the cells of the relevant location area. The MS responds, and the call can be switched through.

It is important to note that a subscriber can attach to a GSM network irrespective of a specific mobile terminal. By inserting the SIM card into another terminal, he/she

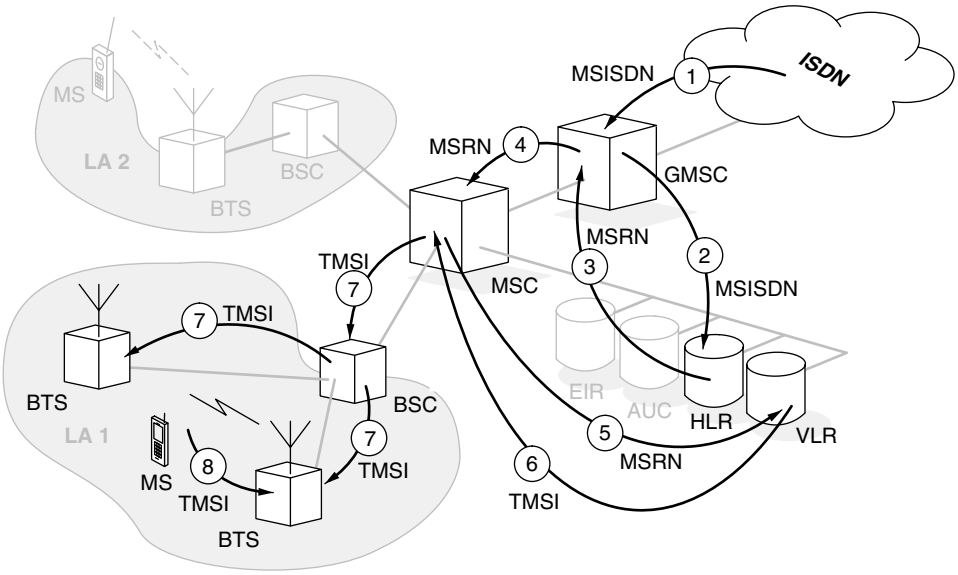


Figure 9. Routing calls to mobile users.

can make or receive calls at that terminal and use other subscribed services. This is possible because the IMEI (identifying the terminal) and the IMSI (identifying the user, stored in the SIM) are independent of each other. Consequently, GSM offers personal mobility in addition to terminal mobility.

5.3. Connection Management

The GSM connection management consists of three entities: call control, supplementary services, and short message service (SMS). The call control handles the tasks related to setting up, maintaining, and taking down calls. The services of call control include the establishment of normal calls (MS-originating and -terminating) and emergency calls (MS-originating only), the termination of calls, dual-tone multifrequency (DTMF) signaling, and incall modifications. The latter allows for changing the bearer service during a connection, for example, from speech to data transmission. Essentially, call control signaling in GSM corresponds to the call setup according to Q.931 in ISDN. Some additional features are incorporated to account for the limited resources in a wireless system. For example, it is possible to queue call requests if no traffic channel is immediately available. Furthermore, the network operator has the option to choose between early and late assignment procedures. With early assignment, the TCH is assigned immediately after acknowledging the call request. In case of late assignment, the call is first processed completely, and the TCH assignment occurs only after the destination subscriber has been called. This variant avoids unnecessary allocation of radio resources if the called subscriber is not available. Thus, the call blocking probability can be reduced.

5.4. Roaming and Handover

GSM supports roaming of mobile users within a network as well as between different GSM networks (as long as a roaming agreement exists between the respective network providers). While roaming within a single network supports the continuation of ongoing connections, roaming from and to other networks terminates the connection and requires a new attachment.

When a mobile user moves within a network and crosses cell borders, the ongoing connection is switched to a different channel through the neighboring BTS (see Fig. 10). This procedure is called *intercell handover* [13]. Handovers may even take place within the same cell, when the signal quality on the current channel becomes insufficient and an alternative channel of the same BTS can offer improved reception. This event is called *intracell handover*.

In case of an intercell handover, GSM distinguishes between intra-MSC handover (if both old and new BTS are connected to the same MSC) and inter-MSC handover (in case the new BTS is connected to a different MSC than the old BTS). In the latter case, the connection is rerouted from the old MSC (MSC-A) and through the new MSC (MSC-B). If during the same connection another inter-MSC handover takes place, either back to MSC-A or to a third one (MSC-C), the connection between MSC-A and MSC-B is taken down and the connection is newly routed from MSC-A to MSC-C (unless MSC-C equals MSC-A). This repeated inter-MSC handover event is called a *subsequent handover* (see Fig. 10).

Since handovers not only induce additional signaling traffic load but also temporarily reduce the speech quality, the importance of a well-dimensioned handover decision algorithm, which should even be locally optimized, is obvious. This is one reason why GSM does not have a standardized uniform algorithm for the determination of the moment of a handover. Network operators can develop and deploy their own algorithms that are optimally tuned for their networks. This is made possible through standardizing only the signaling interface that defines the processing of the handover and through transferring the handover decision to the BSS. The GSM handover is thus a network-originated handover as opposed to a mobile-originated handover, where the handover decision is made by the mobile station. An advantage of the GSM handover

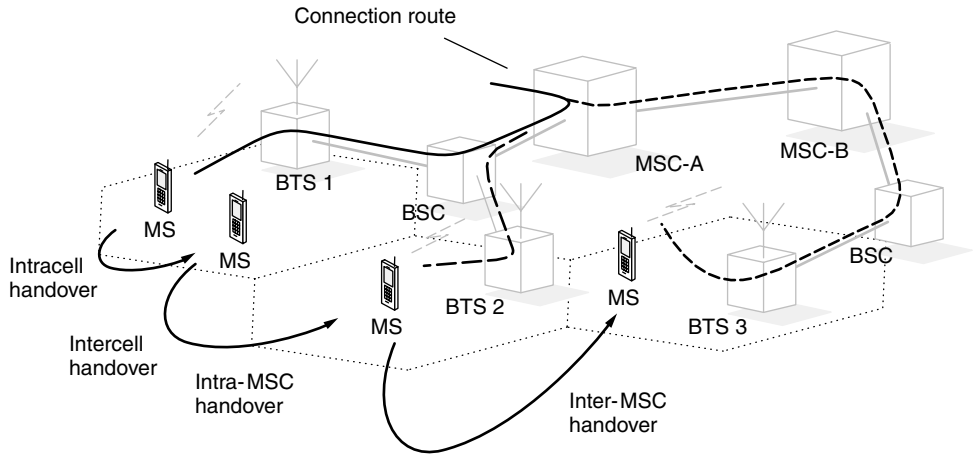


Figure 10. Handover types.

approach is that the software of the MS does not need to be changed when the handover strategy or the handover decision algorithm is changed in the network. Even though the GSM standard does not prescribe a mandatory decision algorithm, a simple algorithm is proposed.

The GSM handover decision is based on measurement data transmitted on the SACCH, most importantly the values RXLEV and RXQUAL. At least 32 values of RXLEV and RXQUAL must be averaged and the resulting mean values are continuously compared with threshold values. Additional handover decision criteria are the distance to the current BTS, an indication of unsuccessful power control, and the pathloss values to neighboring BTSs. The latter are obtained by monitoring the BCCH carriers of nearby cells. Thus, either of the following handover causes can be distinguished:

- Received signal level too low (uplink or downlink)
- Bit error ratio too high (uplink or downlink)
- Power control range exceeded
- Distance between MS and BTS too high
- Pathloss to neighboring BTS is lower than pathloss to the current BTS

A special handover situation exists if the bit error ratio (uplink or downlink) is too high but at the same time the uplink and downlink received signal levels are sufficiently high. This strongly hints at severe cochannel interference. This problem can be solved with an intracell handover, which the BSS can perform on its own without support from the MSC.

Once the decision on an intercell handover is made, a target cell must be chosen. Therefore it is determined which neighboring cell's BCCH is received with sufficient signal level. All potential target channels with lower path loss than the current channel are then reported to the MSC with the message HANDOVER REQUIRED. Once the target cell has been determined, the BSS sends the signaling message HANDOVER COMMAND to the MS on the FACCH. The HANDOVER COMMAND contains the new channel configuration and also information about the new cell (e.g., BTS identity and BCCH frequency) and a handover reference number. On reception of the HANDOVER COMMAND, the mobile station interrupts the current connection, deactivates the old physical channel, and switches over to the channel newly assigned in the HANDOVER COMMAND. On the FACCH, the mobile sends the message HANDOVER ACCESS in an access burst to the new base station. In case both BTSs have synchronized their TDMA transmission, the access burst is sent in exactly four subsequent time slots of the FACCH. Thereafter, the mobile station activates the new physical channel in both directions, activates encryption, and sends the message HANDOVER COMPLETE to the BSS. In case the BTSs are not synchronized, the mobile station repeats the access burst until either a timer expires (handover failure) or the BTS answers with a message PHYSICAL INFORMATION that contains the currently needed timing advance to enable the mobile station to activate the new physical channel.

To avoid a series of repeated handover events at the cell boundary where varying channel conditions can lead to frequent changes of the pathloss to both base stations, a hysteresis is used in GSM. This is done by defining a handover margin for each cell. A handover will be performed only if the path-loss difference of both base stations is higher than the handover margin of the potential new cell.

6. SECURITY ASPECTS

The wireless transmission over the air interface leads to the danger that unauthorized people may eavesdrop on the communication of a subscriber or that they use radio resources at the cost of the network provider. The security functions of GSM protect against unauthorized use of services (by authentication), provide data confidentiality (by encryption), and ensure the confidentiality of the subscriber identity [14].

Authentication is needed to verify that users are who they claim to be. Each subscriber has a secret key (*subscriber authentication key* Ki), which is stored in the SIM as well as in the AUC. The SIM is protected by a *personal identity number* (PIN) that is known only by the

subscriber. To authenticate a user, the network sends a random number to the mobile station. Using the key Ki and a specified algorithm, the mobile station calculates a *signature response* (SRES) from the random number and sends it back to the network. If the SRES value received from the MS matches the value calculated at the network side, the user is authenticated. In addition to user authentication, the mobile terminal must also be authenticated. On the basis of the IMEI, the network can restrict access for terminals that are listed in the EIR as stolen or corrupted.

The encryption of speech and data is a feature of GSM that is not supported in analog cellular and fixed ISDN networks. It is used to protect the transmission over the air interface against eavesdropping. As illustrated in the transmission chain of Fig. 4, data coming from the channel encoder is encrypted before it is passed to the multiplexer. The initial random number and the key Ki are used on both sides, the network and the MS, to calculate a *ciphering key* Kc. This key is then employed by the encryption algorithm for the symmetric ciphering of user data and signaling information. On the receiver side, the key Kc can be used to decipher the data.

Another security feature in GSM is that a user's identity remains undisclosed to listeners on the radio channel. The IMSI uniquely identifies a subscriber worldwide and should thus not be transmitted over the air interface. Thus, in order to protect his/her identity on network attach, the subscriber obtains a temporary identifier, the TMSI, from the network. The TMSI is used instead of the IMSI for transmission over the air interface. The association between the two values is stored in the VLR. The TMSI changes frequently and has only local validity within a location area. Thus, an attacker listening on the radio channel cannot deduce the subscriber's identity from the TMSI.

7. BIBLIOGRAPHIC NOTES

Article 15 is an early survey article mainly describing the system architecture and signaling protocols. The GSM book [2] covers both network aspects (system architecture, protocols, services, roaming) and transmission aspects. It also includes a detailed description of the GSM phase 2+ services. The book [16] also contains a detailed GSM part. The GSM chapter of Ref. 4 covers the physical layer at the air interface in detail.

BIOGRAPHIES

Christian Bettstetter is a research and teaching staff member at the Institute of Communication Networks at Technische Universität München TUM, Germany. He graduated from TUM in electrical engineering and information technology Dipl.-Ing. in 1998 and then joined the Institute of Communication Networks, where is he working toward his Ph.D. degree. Christian's interests are in the area of mobile communication networks, where his current main research area is wireless ad-hoc networking. His interests also include 2G and 3G cellular systems and protocols for a mobile Internet. He is coauthor of the book *GSM–Switching, Services and Protocols* (Wiley/Teubner) and a number of articles in journals, books, and conferences.

Christian Hartmann studied electrical engineering at the University of Karlsruhe (TH), Germany, where he received the Dipl.-Ing. degree in 1996. Since 1997, he has been with the Institute of Communication Networks at the Technische Universität München, Germany, as a member of the research and teaching staff, pursuing a doctoral degree. His main research interests are in the area of mobile and wireless networks including capacity and performance evaluation, radio resource management, modeling, and simulation. Christian Hartmann is a student member of the IEEE.

BIBLIOGRAPHY

1. ETSI, *GSM 03.02: Network Architecture*, Technical Specification, 2000.
2. J. Eberspächer, H.-J. Vögel, and C. Bettstetter, *GSM — Switching, Services, and Protocols*, 2nd ed., Wiley, Chichester, March 2001.
3. ETSI, *GSM 05.02: Multiplexing and Multiple Access on the Radio Path*, Technical Specification, 2000.
4. R. Steele and L. Hanzo, eds., *Mobile Radio Communications*, 2nd ed., Wiley, 1999.
5. R. Salami et al., Description of the GSM enhanced full rate speech codec, in *Proc. IEEE Int. Conf. Communications* (*ICC'97*), Montreal, Canada, June 1997, 725–729.
6. J. Adoul, P. Mabilleau, M. Delprat, and S. Morisette, Fast CELP coding based on algebraic codes, *Proc. ICASSP*, April 1987, pp. 1957–1960.
7. D. Bruhn, E. Ekudden, and K. Hellwig, Adaptive multi-rate: a new speech service for GSM and beyond, *Proc. 3rd ITG Conf. Source and Channel Coding*, Munich, Germany, Jan. 2000.
8. ETSI, *GSM 05.03: Channel Coding*, Technical Specification, 1999.
9. ETSI, *GSM 05.01: Physical Layer on the Radio Path; General Description*, Technical Specification, 2000.
10. ETSI, *GSM 05.10: Radio Subsystem Synchronization*, Technical Specification, 2001.
11. ETSI, *GSM 05.08: Radio Subsystem Link Control*, Technical Specification, 2000.
12. ETSI, *GSM 04.08: Mobile Radio Interface Layer 3 Specification*, Technical Specification, 2000.
13. ETSI, *GSM 03.09: Handover Procedures*, Technical Specification, 1999.
14. ETSI, *GSM 03.20: Security Related Network Functions*, Technical Specification, 2001.
15. M. Rahnema, Overview of the GSM system and protocol architecture, *IEEE Commun. Mag.* 92–100 (April 1993).
16. B. Walke, *Mobile Radio Networks*, 2nd ed., Wiley, Chichester, New York, 2002.

H.324: VIDEOTELEPHONY AND MULTIMEDIA FOR CIRCUIT-SWITCHED AND WIRELESS NETWORKS*

DAVE LINDBERGH
Polycom, Inc.
Andover, Massachusetts

BERNHARD WIMMER
Siemens AG
Munich, Germany

1. INTRODUCTION

ITU-T Recommendation H.324 [1] is the international standard for videotelephony and real-time multimedia communication systems on low-bit-rate circuit-switched networks. The basic H.324 protocol can be used over almost any circuit-switched network, including modems on the PSTN (public switched telephone network, often known as POTS — "plain old telephone service"), on ISDN networks, and on wireless digital cellular networks.

H.324 is most commonly used for dialup videotelephony service over modems, and as the basis for videotelephony service in the Universal Mobile Telecommunications System (UMTS) of the Third Generation Partnership Project (3GPP). The standard enables interoperability among a diverse variety of terminal devices, including PC-based multimedia videoconferencing systems, inexpensive voice/data modems, encrypted telephones, and remote security cameras, as well as standalone videophones.

H.324 is a "toolkit" standard that gives implementers flexibility to decide which media types (audio, data, video, etc.) and features are needed in a given product, but ensures interoperability by specifying a common baseline mode of operation for all systems that support a given feature. In addition to the baseline modes, H.324 allows other optional modes, standard or nonstandard, which may be better in various ways, to be used automatically if both ends have the capability to do so.

Above all, H.324 is designed to provide the best performance possible (video and audio quality, delay, etc.) on low-bit-rate networks. This is achieved primarily by reducing protocol overhead to the minimum extent possible, and by designing the multiplexer and channel aggregation protocols to allow different media channels to interleave frequently and flexibly, minimizing latency. These protocol optimizations mean that H.324 is not suitable for use on packet-switched networks, including IP networks, because H.324 depends on reasonably constant end-to-end latency in the connection, which can't be guaranteed in packet-routed networks.

The design of the H.324 standard benefits from industry's earlier experience with ITU-T H.320, the widespread international standard for ISDN videoconferencing, approved in 1990. H.324 shares H.320's basic architecture, consisting of a multiplexer that mixes the various media types into a single bitstream (H.223), audio and video compression algorithms (G.723.1 and H.263), and a control protocol that performs automatic capability negotiation and logical channel control (H.245). Other parts of the H.324 set, such as H.261 video compression, H.233/234 encryption, and H.224/281 far-end camera control, come directly from H.320. One of the considerations in the development of H.324 was practical interworking, through a gateway, with the installed base of H.320 ISDN systems, as well as with the ITU-T standards for multimedia on LANs and ATM networks, H.323 and H.310 respectively, and the ITU-T T.120 series of data conferencing standards. Table 1 compares H.324 with H.320 and the other major ITU-T multimedia conference standards.

As a second-generation standard, the design of H.324 was able to avoid the limitations of H.320. As a result, H.324 has features missing from (or retrofitted to) H.320 such as receiver-controlled mode preferences, the ability to support multiple channels of each media type, and dynamic assignment of bandwidth to different channels.

1.1. Variations

The original version of the H.324 standard was approved in 1995, exclusively for use on PSTN connections over

* This article is based on the article entitled The H.324 Multimedia Communication Standard, by Dave Lindbergh, which appeared in *IEEE Communications Magazine*, Dec. 1996, Vol. 24, No. 12, pp. 46–51.

Table 1. Major ITU-T Multimedia Conferencing Standards (Basic Modes)

Standard	Initial Approval	Networks Supported	Baseline Video	Baseline Audio	Multiplex	Control
H.324	1995	PSTN, ISDN, wireless	H.263	G.723.1	H.223	H.245
H.320	1990	ISDN	H.261	G.711	H.221	H.242
H.323	1996	Internet, LANs, Intranets	H.261	G.711	H.225.0	H.245
H.310	1996	ATM/B-ISDN	H.262	MPEG-1	H.222	H.245

dialup V.34 modems, at rates of up to 28,800 bps. Since then, modems have increased in speed and H.324 has been extended to support not only PSTN connections but also ISDN connections and wireless mobile radio links.

The original PSTN version of H.324 is called "H.324." The ISDN version is called "H.324/I," and the mobile version "H.324/M." The specific variant of H.324/M used in UMTS is called "3G-324" (see Table 2).

The requirements, capabilities, and protocols during a call are essentially the same for all variants of H.324. The main differences are the network interfaces and call setup procedures, and in the case of H.324/M, the use of a variant of the H.223 multiplex that is more resilient to the very high bit error rates encountered on wireless networks.

We begin by describing the base PSTN version of H.324, and describe the differences in the other variations afterwards.

2. SYSTEM OVERVIEW

Figure 1 illustrates the major elements of the H.324 system. The mandatory components are the V.34 modem (for PSTN use), H.223 multiplexer, and H.245 control protocol. The data, video, and audio streams are optional, and several of each kind may be present, set up dynamically during a call. H.324 terminals negotiate a common set of capabilities with the far end, independently for each direction of transmission (terminals with asymmetric capabilities are allowed). Multipoint operation in H.324, in which three or more sites can join in the same call, is possible through the use of a multipoint control unit (MCU, also known as a "bridge").

Table 2. Variants of H.324

Name	Description	Main Differences
H.324	For PSTN over modems	—
H.324/I	For ISDN	Network interface, call setup
H.324/M	For mobile wireless networks	Network interface, error-robust multiplex
3G-324M	Variant of H.324/M for 3GPP UMTS wireless video telephony	Network interface, error-robust multiplex, AMR audio codec instead of G.723.1

3. MODEM

When operating on PSTN circuits, H.324 requires support for the full-duplex V.34 modem [2], which runs at rates of up to 33,600 bps. The modem can operate at lower rates, in steps of 2400 bps, as line conditions require. H.324 is intended to work at rates down to 9600 bps. The mandatory V.8 or optional V.8bis protocol is used at call startup to identify the modem type and operation mode. In most implementations, the preferred V.8bis protocol allows a normal voice telephone call to switch into a multimedia call at any time. ITU-T V.25ter (the "AT" command set) is used to control local modem functions such as dialing. The V.34 modulation/demodulation is quite complex, and typically adds 30–40 ms of end-to-end delay.

An important V.34 characteristic for H.324 system design is its error burst behavior. Because of the V.34 trellis coder and other factors, bit errors rarely occur singly, but instead in bursts. The rate of these error bursts is also highly dependent on the bit rate of the modem. A single "step" of 2400 bps changes the error rate by more than an order of magnitude. This influences the proper "tuning" of the modem for use with H.324: a modem used for simple V.42 data transfer is often more efficient at the "next higher" rate, since retransmissions will consume less bandwidth than will the loss from stepping down. But a H.324 system should be run at the "next lower" rate,

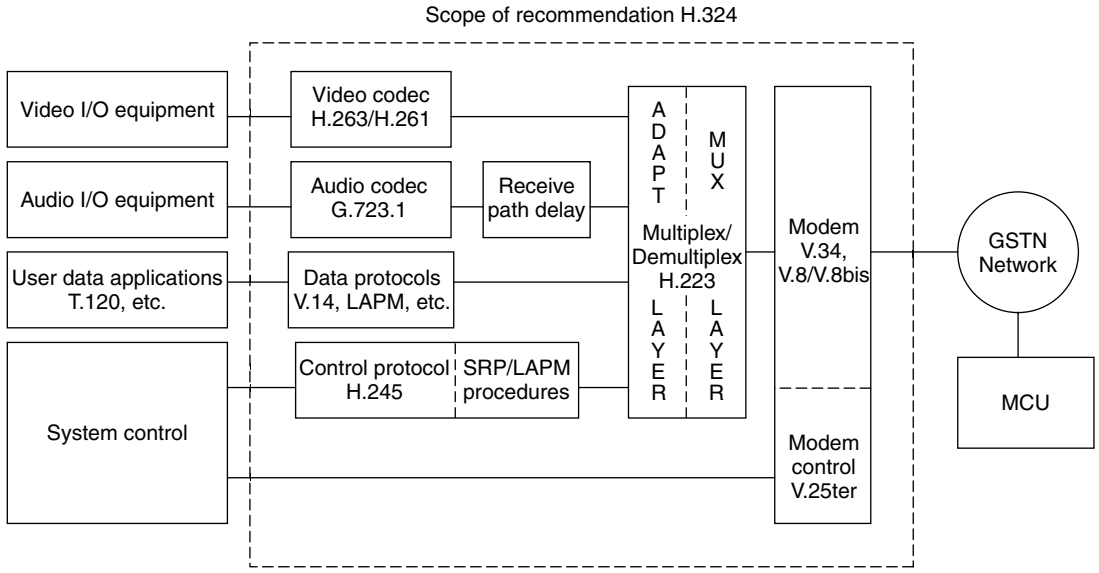


Figure 1. H.324 block diagram.

since the real-time requirements make retransmission inappropriate in many cases, requiring a lower error rate.

H.324 uses the V.34 modem directly as a synchronous data pump, to send and receive the bitstream generated by the H.223 multiplexer. Data compression such as V.42bis, and retransmission protocols such as V.42 link-access procedure for modems (LAPM) or the Microcom network protocol (MNP) are not used at the modem level, although these same protocols can be used at a higher layer for some types of data streams.

Like the V.42 error correction protocol supported in most modems, H.324 requires a synchronous interface to the V.34 data pump. Since most PCs have only asynchronous RS-232 modem interfaces, PC-based H.324 terminals require either synchronous interface hardware or some other method to pass the synchronous bitstream to the modem over the PC's asynchronous interface. Many modems support ITU-T V.80, which provides a standard sync-over-async "tunneling" protocol between the PC and the modem for this purpose.

Faster modems such as V.90 can also be used, but in most cases these modems offer asymmetric bit rates, where the modem is able to receive data faster than it can transmit. When two such modems interconnect, both sides transmit at the lower rate, yielding little or no improvement over V.34 speeds.

When operating on circuit-switched networks other than the PSTN, the modem is replaced by an equivalent interface that allows the output from the H.223 multiplex to be carried over the network in use. H.324 Annex C specifies operation of H.324 on wireless (digital cellular and satellite) networks, and H.324 Annex D specifies H.324 operation on ISDN.

4. CALL SETUP AND TEARDOWN

The setup and teardown of a H.324 call proceeds in seven phases, phase A–phase G.

In *phase A*, an ordinary telephone call is setup using the normal procedures for dialing, ringing, and answering. This can be completely manual, with the user dialing and answering by hand; or automatic, with the V.25ter protocol (the "AT" command set) used to control the modem's dialing and answering functions.

Once the call is connected, the H.324 terminal can immediately start the multimedia call, or, if both terminals support the optional V.8bis protocol, the two users can choose to have an ordinary voice conversation first. This is called *phase B*, which can continue indefinitely, until one of the users decides to switch the call into H.324 multimedia mode. V.8bis allows this "late start" mode, in which an ordinary phone call can be switched into multimedia mode at any time. On the ISDN network, the V.140 protocol substitutes for V.8bis.

In *phase C*, digital communication between the terminals is set up. For ISDN and digital wireless networks, this is inherent in the network connection. For PSTN calls, either terminal starts the modem negotiation by sending V.8 or V.8bis messages, using V.23 300 bps FSK modulation (which doesn't require any "training" time). The modems exchange capabilities and select V.34 modulation and H.324 protocol. A PC or other external DTE can control this V.8 or V.8bis negotiation using the procedures of V.25ter Annex A. If V.8bis is used, the terminal's gross ability to support H.324 video, audio, data, or encryption is also communicated, to quickly determine if the desired mode is available. The V.34 startup procedure then takes about 10 seconds, after which time end-to-end digital communication is established at the full V.34 data rate allowed by the quality of the telephone connection, up to 33,600 bps.

Phase D then starts, in which the H.324 terminals first communicate with each other on the H.245 control channel. The terminals are initialized, and detailed terminal capabilities are exchanged using H.245. This happens very quickly (less than 1 second), as the full connection bit rate is available for transfer of this control information. Logical channels can then be opened for the various media types, and table entries can be downloaded to support those channels in the H.223 multiplexer.

At this point the terminal enters *phase E*, normal multimedia communication mode. The number and type of logical channels can be changed during the call, and each terminal can change its capabilities dynamically.

Phase F is entered when either user wants to end the call. All logical channels are closed, and an H.245 message tells the far-end terminal to terminate the H.324 portion of the call, and specifies the new mode (disconnect, back to voice mode, or other digital mode) to use.

Finally, in *phase G*, the terminals disconnect, return to voice telephone mode, or go into whatever new mode (fax, V.34 data, etc.) was specified.

5. MULTIPLEX

H.324 uses a new multiplexer standard, H.223, to mix the various streams of video, audio, data, and the control channel, together into a single bitstream for transmission. The H.324 application required the development of a new multiplexing method, as the goal was to combine low multiplexer delay with high efficiency and the ability to handle bursty data traffic from a variable number of logical channels.

Time-division multiplexers (TDMs), such as H.221, were considered unsuitable because they can't easily adapt to dynamically changing modem and payload data rates, and are difficult to implement in software because of complex frame synchronization and bit-oriented channel allocation.

Packet multiplexers, such as V.42 (LAPM) and Q.922 (LAPF), avoid these problems but suffer from "blocking delay," where transmission of urgent data, such as audio, must wait for the completion of a large packet already started. This problem occurs when the underlying channel bit rate is low enough to make the transmission time of a single packet significant. In a packet multiplexer, this delay can be reduced only by limiting the maximum packet size or aborting large packets, both of which reduce efficiency.

The H.223 multiplexer combines the best features of TDM and packet multiplexers, along with some new ideas. It incurs less delay than TDM and packet multiplexers,

has low overhead, and is extensible to multiple channels of each data type. H.223 is byte-oriented for ease of implementation, able to byte-fill with flags to match differing data rates, and uses a unique synchronization character that cannot occur in valid data. In H.223, each HDLC (*h*igh-level *d*ata *l*ink *c*ontrol) framed [3] multiplex–protocol data unit (MUX-PDU) can carry a mix of different data streams in different proportions, allowing fully dynamic allocation of bandwidth to the different channels.

H.223 consists of a lower multiplex layer, which actually mixes the different media streams, and a set of adaptation layers that perform logical framing, sequence numbering, error detection, and error correction by retransmission, as appropriate to each media type.

5.1. Multiplex Layer

The multiplex layer uses normal HDLC zero insertion, which inserts a 0 bit after every sequence of five 1 bits, making the HDLC flag (01111110) unique. The multiplex consists of one or more flags followed by a one-byte header, followed by a variable number of bytes in the information field. This sequence repeats. Each sequence of header byte and information field is called a MUX-PDU, as shown in Fig. 2.

The header byte includes a multiplex code that specifies, by reference to a multiplex table, the mapping of the information field bytes to various logical channels. Each MUX-PDU may contain a different multiplex code, and therefore a different mix of logical channels.

The selected multiplex table entry specifies a pattern of bytes and corresponding logical channels contained in the information field, which repeats until the closing flag. Each MUX-PDU may contain bytes from several different logical channels, and may select a different multiplex table entry, so bandwidth allocation may change with each MUX-PDU. This allows many logical channels, low-overhead switching of bandwidth allocation, and many different types of channel interleave and priority. All of this is under control of the transmitter, which may choose any appropriate multiplex for the application, and change multiplex table entries as needed. Many syntactically compliant multiplexing algorithms, optimized for different applications, are possible [4].

Figure 3 illustrates only four logical channels (audio, video, data, and control), but there may be any number of channels, as specified by the multiplex table. This allows multiple audio, video, and data channels for different data protocols, multiple languages, continuous presence, or other uses.

A slightly different, but equally valid, way of thinking about H.223 is as a continuous stream of bit-stuffed bytes carrying information in a given pattern of logical channels. This pattern (again refer to Fig. 3) is occasionally interrupted by an "escape sequence" consisting of an HDLC flag followed by a single "header" byte. The header byte contains a multiplex code that indicates a change to a new pattern of logical channel bytes.

Generally, an HDLC flag may be inserted on any byte boundary, terminating the previous MUX-PDU and beginning a new one. Unlike traditional packet multiplexers, H.223 can interrupt a adaptation-layer variable-length packet at any byte boundary to reallocate bandwidth, such as when urgent real-time data must be sent. The reallocation can occur within 16–24 bit times (16 bits for the flag and header, plus 0 to 8 bits for the byte already being transmitted), less than 1 ms at 28,800 bps. This compares very favorably with both TDM muxes and packet multiplexes which suffer from "blocking delay."

H.223 maintains a 16-entry multiplex table at all times, selected by 4 bits of the MUX-PDU header byte. Entries in the multiplex table may be changed during the call, to meet requirements as various logical channels are opened or closed.

Of the remaining 4 bits in the header, 3 are used as a CRC (cyclic redundancy code) on the multiplex code. The final bit is the packet marker (PM) bit, which is used for marking the end of some types of higher-level packets.

5.2. Adaptation Layer

Different media types (audio, video, data) require different levels of error protection. For example, T.120 application data are relatively delay-insensitive, but require full error correction. Real-time audio is extremely delay sensitive, but may be able to accept occasional errors with only minor degradation of performance. Video falls between these two extremes. Ideally, each media type should use an error-handling scheme appropriate to the requirements of the application.

The H.223 adaptation layers provide this function, working above the multiplex layer on the unmultiplexed sequence of logical channel bytes. H.223 defines three adaptation layers, AL1, AL2, and AL3. AL1 is intended primarily for variable-rate framed information, such as HDLC protocols and the H.245 control channel. AL2 is intended primarily for digital audio such as G.723.1, and includes an 8-bit CRC and optional sequence numbers. AL3 is intended primarily for digital video such as H.261 and H.263, and includes provision for retransmission using sequence numbers and a 16-bit CRC.

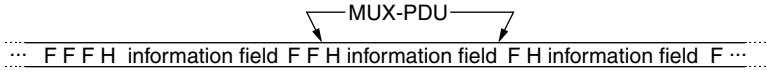


Figure 2. MUX-PDU in an H.223 multiplex stream (F = HDLC flag, H = header byte).

...VVVCVVVCV **FH** AAAAVVDCAAAAVVDCAAAAVV **FH** VDVDVDVDVDVDV **FH** CCCCCC **FH** AAAAVVDCAAA...

(Each letter represents one byte, F = Flag, H = Header, A = Audio, V = Video, D = Data, C = Control)

Figure 3. Example of an H.223 multiplex bitstream.

5.3. Encryption

Encryption is an option in H.324, and makes use of ITU-T H.233 and H.234, both of which were originally developed for use with H.320. H.233 covers encryption procedures and algorithm selection, in which FEAL, B-CRYPT, and DES may be used as well as other standardized and nonstandardized algorithms. H.234 covers key exchange and authentication using the ISO 8732, Diffie-Hellman, or RSA algorithms.

In H.324, encryption is applied at the H.223 multiplexer, where the encryptor produces a pseudorandom bitstream (cipher stream), which, prior to flag insertion and HDLC zero insertion, is exclusive-ORed with the output of the H.223 multiplexer. The exclusive-OR procedure is not applied to the H.223 MUX-PDU header and the H.245 control channel. The receiver reverses this process.

6. CONTROL

H.324 uses the H.245 multimedia system control protocol, which is also used by H.310 (for ATM networks) and H.323 (for packet-switched LANs and corporate intranets). Many of the extensions added to H.245 to support H.323 on IP networks are not used by H.324.

The control model of H.245 is based on "logical channels," independent unidirectional bitstreams with defined content, identified by unique numbers arbitrarily chosen by the transmitter. There may be up to 65,535 logical channels.

The H.245 control channel carries end-to-end control messages governing operation of the H.324 system, including capabilities exchange, opening and closing of logical channels, mode preference requests, multiplex table entry transmission, flow control messages, and general commands and indications. The H.245 structure allows future expansion to additional capabilities, as well as manufacturer-defined nonstandard extensions to support additional features.

H.245 messages are defined using ASN.1 syntax, coded according to the packed encoding rules (PERs) [5] of ITU-T X.691, which provide both clarity of definition and flexibility of specification. In H.324, the H.245 control channel runs over logical channel (LC) 0, a separate channel out of band from the various media streams. LC 0 is considered to be already open when the H.324 call starts up.

Within LC 0, H.245 is carried on a numbered simplified retransmission protocol (NSRP) based on V.42 exchange identification (XID) procedures. Optionally, the full V.42 protocol may be used instead. H.245 itself assumes that this link layer guarantees correct, in-order delivery of messages.

6.1. Capabilities Exchange

The large set of optional features in H.324 necessitates a method for the exchange of capabilities, so that terminals can become aware of the common subset of capabilities supported by both ends.

H.245 capabilities exchange provides for separate receive and transmit capabilities, as well as a system by which the terminal may describe its ability (or inability) to operate in various combinations of modes simultaneously, as some implementations are limited in processing cycles or memory availability.

Receive capabilities describe the terminal's ability to receive and process incoming information streams. Transmitters are required to limit the content of their transmitted information to that which the receiver has indicated it is capable of receiving. The absence of a receive capability indicates that the terminal cannot receive (is a transmitter only).

Transmit capabilities describe the terminal's ability to transmit information streams. They serve to offer receivers a choice of possible modes of operation, so the receiver can request the mode it prefers to receive using the H.245 RequestMode message. This is an important feature, as local terminals directly control only what they transmit, but users care about controlling what they receive. The absence of a transmit capability indicates that the terminal is not offering a choice of preferred modes to the receiver (but it may still transmit anything within the capability of the receiver).

Terminals may dynamically add or remove capabilities during a call. Since many H.324 implementations are on general-purpose PCs, other application activity on the machine can result in changing resource levels available to H.324. H.245 is flexible enough to handle such a scenario.

Nonstandard capabilities and control messages may be issued using the NonStandardParameter structure defined in H.245. This allows nonstandardized features to be supported automatically, in the same way as standardized features.

6.2. Logical Channel Signaling

H.324 terminals transmit media information from transmitter to receiver over unidirectional streams called *logical channels* (LCs). Each LC carries exactly one channel of one media type, and is identified by a logical channel number (LCN) arbitrarily chosen by the transmitter. Since transmitters completely control allocation of LCNs, there is no need for end-to-end negotiation of LCNs.

When a transmitter opens a logical channel, the H.245 OpenLogicalChannel message fully describes the content of the logical channel, including media type, codec in use, H.223 adaptation layer and any options, and all other information needed for the receiver to interpret the content of the logical channel. Logical channels may be closed when no longer needed. Open logical channels may be inactive, if the information source has nothing to send.

Logical channels in H.324 are unidirectional, so asymmetrical operation, in which the number and type of information streams is different in each direction of transmission, is allowed. However, if a terminal is capable only of certain symmetrical modes of operation, it may send a capability set that reflects its limitations.

6.3. Bidirectional Logical Channels

Certain media types, including data protocols such as T.120 and LAPM, and video carried over AL3, inherently require a bidirectional channel for their operation. In such cases a pair of unidirectional logical channels, one in each

direction, may be opened and associated together to form a bidirectional channel. Such pairs of associated channels need not share the same logical channel number, since logical channel numbers are independent in each direction of transmission. To avoid race conditions that could cause duplicate sets of bidirectional channels to be opened, a symmetry-breaking procedure is used, in which master and slave terminals are chosen on the basis of random numbers. There is no advantage to being master or slave.

7. AUDIO CHANNELS

The baseline audio mode for H.324 is the G.723.1 speech coder/decoder (codec), which runs at 5.3 or 6.4 kbps. The 3G-324M variant of H.324 for 3G wireless networks uses the adaptive multirate (AMR) codec as the baseline.

G.723.1 provides near-toll-quality speech, using a 30 ms frame size and 7.5 ms lookahead. A G.723.1 implementation is estimated to require 18–20 fixed-point (MIPS) (million instructions per second) in a general-purpose DSP (digital signal processor). Transmitters may use either of the two rates, and can change rates for each transmitted frame, as the coder rate is sent as part of the syntax of each frame. The average audio bit rate can be lowered further by using silence suppression, in which silence frames are not transmitted, or are replaced with smaller frames carrying background noise information. In typical conversations both ends rarely speak at the same time, so this can save significant bandwidth for use by video or data channels.

Receivers can use a H.245 message to signal their preference for low- or high-rate audio. The audio channel uses H.223 adaptation layer AL2, which includes an 8-bit cyclic redundancy code (CRC) on each audio frame or group of frames.

The G.723.1 codec imposes about 97.5 ms of end-to-end audio delay, which, together with modem, jitter buffer, transmission time, multiplexer, and other system delays, results in about 150 ms of total end-to-end audio delay (exclusive of propagation delay) [6]. On ISDN connections, modem latency is eliminated, leading to about 115 ms end-to-end delay. These audio delays are generally less than that of the video codec, so additional delay needs to be added in the audio path if lip synchronization is desired. This is achieved by adding audio delay in the receiver only, as shown in Fig. 1. H.245 is used to send a message containing the time skew between the transmitted video and audio signals. Since the receiver knows its decoding delay for each stream, the time skew message allows the receiver to insert the correct audio delay, or alternatively, to bypass lip synchronization, and present the audio with minimal delay. While multiple channels of audio can be transmitted, H.324 does not currently provide for the exact sample-level channel synchronization needed for stereo audio.

Many H.324 applications do not require lip synchronization, or do not require video at all. For these applications, optional H.324 audio codecs such as G.729, an 8-kbps speech codec which can reduce the total end-to-end audio delay to about 85 ms for PSTN use, or 50 ms on ISDN, can be important.

Other optional audio modes can also be used, such as the wideband ITU-T G.722.1 codec, which offers 7-kHz audio bandwidth (approximately double that of conventional telephone lines) at rates of 24 or 32 kbps. Nonstandard audio modes can be also used in the same manner as standardized codecs.

8. VIDEO CHANNELS

H.324 can send color motion video over any desired fraction of the available modem bandwidth. H.324 supports both H.263 and H.261 for video coding. H.263 is the preferred method, with H.261 available to allow interworking with older ISDN H.320 videoconferencing systems without the need to convert video formats, which would add an unacceptable delay.

H.263 is based on the same video compression techniques as H.261, but includes many enhancements, including much improved motion compensation, which result in H.324 video quality estimated equivalent to H.261 at a 50–100% higher bitrate. This dramatic improvement is most apparent at the low bit rates used by H.324; the difference is less when H.263 is used at higher bitrates. H.263 also includes a broader range of picture formats, as shown in Table 3.

H.324 video can range from 5 to 30 frames/second, depending on bitrate and picture format, H.263 options in use, and the amount of complexity and movement in the scene. A H.245 control message, the videoTemporalSpatialTradeOff, feature allows the receiver to specify a preference for the tradeoff between frame rate and picture resolution.

Video channels use H.223 adaptation layer AL3, which includes a 16-bit CRC and sequence numbering, and provision for retransmission of errored video data, at the option of the receiver.

Since H.324 systems can support multiple channels of video (although bandwidth constraints may make this impractical for many applications), continuous-presence multipoint operation, in which separate images of each transmitting site are presented at the receiver "Hollywood Squares" style, can be easily implemented via multiple logical channels of video. It is up to the receiver to locally arrange an appropriate set of the channels for display.

9. DATA CHANNELS

The H.324 multimedia system can carry data channels as well as video and audio channels. These can be used for any

Table 3. Video Picture Formats

Picture Format	Luminance Pixels	H.324 Video Decoder Requirements	
—	—	H.261	H.263
SQCIF	128 × 96 for H.263[a]	Optional[a]	Required
QCIF	176 × 144	Required	Required
CIF	352 × 288	Optional	Optional
4CIF	704 × 576	Not defined	Optional
16CIF	1408 × 1152	Not defined	Optional
Custom formats	≤2048 × 1152	Not defined	Optional

[a]H.261 SQCIF is any active size less than QCIF, filled out by a black border, coded in QCIF format.

data protocol or application in the same way as an ordinary modem. Standardized protocols include T.120 data [7] for conferencing applications such as electronic whiteboards and computer application sharing, user data via V.14 or V.42 (with retransmission), T.84 still image transfer, T.434 binary file transfers, H.224/H.281 for remote control of far-end cameras with pan, tilt, and zoom functions, and ISO/IEC TR9577 network-layer protocols [8] such as IP (Internet protocol) and IETF PPP (point-to-point protocol) [9], which can be used to provide Internet access over H.324. As with other media types, H.324 provides for extension to non-standard protocols, which can be negotiated automatically and used just like standardized protocols.

The same capability exchange and logical channel signaling procedures defined for video and audio channels are also used for data channels, so automatic negotiation of data capabilities can be performed. This represents a major step forward from current practice on data-only PSTN modems, where data protocols and applications to be used must generally be arranged manually by users before a call.

As with all other media types, data channels are carried as distinct logical channels over the H.223 multiplexer, which can accommodate bursty data traffic by dynamically altering the allocation of bandwidth among the different channels in use. For instance, a video channel can be reduced in rate (or stopped altogether) when a data channel requires additional bandwidth, such as during a file or still image transfer.

10. MULTIPOINT OPERATION AND H.320 ISDN INTEROPERABILITY

H.324 terminals can directly participate in multipoint calls, in which three or more participants join the same call, through a central bridge called a multipoint control unit (MCU). Since the connections on each link in a multipoint call may be operating at different rates, MCUs can send H.245 FlowControlCommand messages to limit the overall bit rate of one or more logical channels, or the entire multiplex, to a supportable common mode.

A similar situation arises when a H.324 terminal interoperates with an H.320 terminal on the ISDN. For interworking with H.320 terminals, an H.324/H.320 gateway can transcode the H.223 and H.221 multiplexes, and the content of control, audio, and data logical channels between the H.324 and H.320 protocols. If the H.320 terminal doesn't support H.263, then H.261 (QCIF) Quarter CIF video can be used to avoid the delay of video transcoding. In this case the gateway, like the MCU in the multipoint case, can send the H.245 FlowControlCommand to force the transmitted H.324 video bit rate to match the H.320 video bit rate in use by the H.221 multiplexer.

One way for a dual-mode (H.320 ISDN and H.324 PSTN) terminal on the ISDN to work directly with H.324 terminals on the PSTN is by using a "virtual modem" on the ISDN, which generates a G.711 audio bitstream representing the V.34 analog modem signal.

11. CHANNEL AGGREGATION FOR MULTILINK OPERATION

H.324 Annex F specifies the use of the H.226 channel aggregation protocol, which allows H.324 PSTN and ISDN calls to aggregate the capacity of multiple separate connections into one faster channel. This can provide improved video quality, compared to what can be delivered by a single V.34 or V.90 connection (at no more than 56 kbps) or by a single ISDN B-channel (at 64 kbps).

Although H.226 is a general-purpose channel aggregation protocol, it was designed specifically to meet the requirements of H.324. Modems like V.34 vary their bit rate as telephone-line noise conditions change. When multiple modems are used on separate lines, each modem changes its bit rate independently, so the bitrate available on each channel varies over time without channel-to-channel coordination.

This characteristic rules out the use of conventional synchronous channel aggregation protocols like H.221 and ISO/IEC CD 13871 "BONDING" [10]. These synchronous protocols work by distributing units of data to each channel in a fixed pattern, "spreading out" a higher-rate data stream over a number of lower-rate channels. The fixed pattern avoids the need to send overhead data to tell the receiver to how to reconstruct the original data. With no overhead, the units of distributed data can be arbitrarily small, resulting in very little latency added by the channel aggregation protocol, as well as high efficiency. However, because the distribution pattern is fixed, these synchronous protocols cannot make use of channels with arbitrary or varying bitrates, such as those provided by V.34 modems.

Packet-oriented channel aggregation protocols avoid this problem, but cannot simultaneously provide both very low latency and very low overhead. In a packet-oriented channel aggregation protocol, such as the "PPP Multilink Protocol" of RFC 1990 [11], data are repeatedly divided into packets, and each packet is transmitted on the channel whose transmit queue is least full. This proportionally distributes packets on all channels regardless of their rate. Because the receiver doesn't know which packets were sent on which channels, the transmitter must include a header to delineate packet boundaries and identify the original order of the packets. The overhead of such a protocol is inversely proportional the size of the packets — on small packets, the header overhead forms a larger proportion of the total rate. The latency of transmission, however, is proportional to the size of the packets — larger packets mean greater buffering, and thus more latency. This results in an unavoidable tradeoff between latency and overhead in a packet-based channel aggregation protocol. H.226 provides the ability to operate on channels with arbitrary or varying bitrates, while still achieving low latency and efficiency close to that of synchronous channel aggregation protocols.

11.1. H.226 Operation

The H.226 channel aggregation protocol operates between the H.223 multiplexer and multiple physical network connections, as shown in Fig. 4.

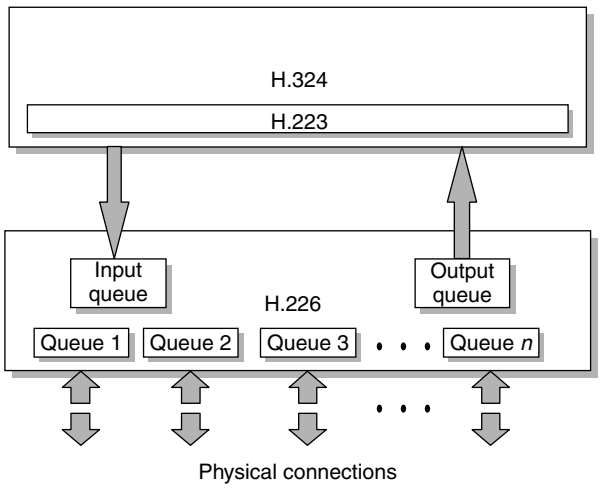


Figure 4. H.226/H.324 protocol stack.

H.226 operates by maintaining a separate transmit (and receive) queue for each individual transmission channel. The full-rate bitstream is divided into 8-bit samples, and each sample is placed on a transmit queue corresponding to one of the channels. The channel used to transmit each sample is determined not by actual queue fullness, but by the output of a finite-state machine called the *channel reference model* (CRM). The CRM simulates queue fullness for each channel using a simple algorithm that is also used identically and in synchrony at the receiver, so that the transmitter and receiver choose the same channel to transmit and receive each sample.

The CRM state machine in H.226 distributes the samples among the different channels in proportion to the bitrate of each channel, so that the full capacity of all the channels is used. For example, if there are two channels, and the first operates at double the bitrate of the other, the first channel will send two samples for each one that the other channel sends.

For each available transmission channel, the CRM maintains a state variable which represents the modeled (not actual) "queue emptiness" of that channel. It also maintains a set of "channel proportion" values, which represent the ratio of the bitrate of that channel to the total combined bitrate of all available channels. As each sample is processed, the state variable of each channel is incremented by the amount of the "channel proportion" value, so that the state variables reflect each channel's capacity to transmit. The CRM selects the channel having the largest state variable to transmit the sample. The state variable for that channel is then decreased by the sum of the "channel proportion" values for all channels, which compensates for the transmission of the sample.

Periodically, and when the transmitter detects an excessive divergence between the actual transmit queues and their expected fullness (possibly as a result of changing channel bitrates), the transmitter sends an updated set of "channel proportion" values to the receiver, and the CRM on each side is reset. This allows H.226 to track changing channel rates and correct for any other divergences between the Channel Reference Model and the actual transmit queues.

11.2. Call Setup for Multilink

Use of H.226 is negotiated during *phase C* of the H.324 call setup, using the procedures of V.8bis (for PSTN) or V.140 (for ISDN). If both terminals indicate in phase C that they wish to use the multilink mode, H.226 is used as a layer between the H.223 multiplex and the physical network connection on the channel that started the call.

The setup of additional channels is coordinated by H.245 message exchange as specified in H.324 Annex F. To avoid the need for the end user to dial additional telephone numbers for each extra channel, a special procedure has the answering terminal automatically provide for the answering terminal to automatically provide telephone numbers for additional connections. This information is sent in a H.245 DialingInformation message that contains only the rightmost digits of the telephone number that differ from the original number dialed by the end user.

For example, if the initial connection was established by dialing "0019786234349," and the DialingInformation message contains "51," the number to be dialed for the additional connection is "0019786234351." This differential digit method is used instead of the full E.164 [12] international number because the first few digits of the number to be dialed can vary according to the geographic location of the two terminals, such as whether they are located in the same building, city, or country.

12. H.324/I FOR ISDN

Operation of H.324 on ISDN is covered by Annex D of H.324, and is essentially identical to the PSTN use of H.324, except for call setup. For ISDN, the V.34 modem, which is used on PSTN lines to provide a digital bitstream, is replaced by an I.400 series ISDN user–network interface [13], which provides a direct digital connection to the ISDN network.

In *phase A* of the call setup, the end-to-end connection is made using normal ISDN procedures that make use of D-channel signaling. When this is complete, the V.140 ISDN call setup protocol is started.

12.1. V.140 ISDN Call Setup

V.140 is used to characterize the end-to-end digital connection before use, and has an automatic mode selection system that allows the terminals to automatically determine if a H.324/I, H.320, or ordinary G.711 voice telephone call is desired when the call first connects. V.140 proceeds in three phases: *phases 1–3*.

In *phase 1*, a V.140 "signature" pattern is sent in the low-order bits of each byte. The signature indicates that the terminal supports V.140, and is capable of proceeding to *phases 2* and *3*. The signature is designed to avoid conflict with a similar pattern called the *frame alignment*

sequence (FAS) used to start H.320. By transmitting in the low-order bits, these patterns create minimal disruption to G.711 coded audio signals, creating only low-level background noise. Therefore G.711 audio, which can contain speech or tones from PSTN modems, the H.320 FAS signal, and the V.140 signature, can all be sent simultaneously. This allows the terminal to negotiate and interoperate correctly if the far-end terminal supports any of these protocols, even if it doesn't support H.324/I or V.140. Once both terminals detect the V.140 signature, *phase 2* is entered.

Phase 2 of V.140 probes the network connection to determine which of a variety of national ISDN and ISDN-like digital network types is use. Most ISDN networks supply an 8.0-kHz clock that times each 8-bit byte transmitted over the network. The clock indicates the boundaries between bytes. Some networks transmit only 7 bits out of each byte end-to-end (providing 56 kbps), while others carry all 8 bits (for 64 kbps). Other network types don't provide any byte timing at all. Even between two terminals with full 64-kbps connections with byte timing, sometimes there can be an intervening network that drops 1 out of each 8 bits. Phase 2 of V.140 sorts this out by exchanging a series of test patterns across the connections. Each terminal can then determine which bits are usable for transmission, and in what order. This process can make H.324/I calls possible on connections that would be mismatched otherwise.

Finally, in *phase 3*, the two terminals exchange a list of modes that they are capable of using, and choose one to be used for the call. This capability exchange process can also be used as a substitute for the V.8bis "late start" procedure in *phase B* of the PSTN H.324 call startup.

After V.140 is complete, call setup continues from phase D. Operation of H.324/I after that is identical to H.324 on the PSTN.

13. H.324/M FOR MOBILE (WIRELESS) NETWORKS

The success of digital cellular telephone systems led to a strong industry desire for a real-time videophone service for $2\frac{1}{2}$ and 3G cellular systems offering higher data rates, such as GPRS (general packet radio service), EDGE (enhanced data rates for GSM evolution), cmda2000, and UMTS.

The H.324 variant called H.324/M supports this application, and is described in H.324 Annex C. The main difference from H.324 is the addition of robust error-detection and error-correction features in the H.223 multiplexer, and the use of G.723.1 Annex C scalable error-protection coding for audio. These allow H.324/M to cope with the high bit error rates (as poor as 5×10^{-2}) of these wireless networks.

The 3rd Generation Partnership Project (3GPP) videophone standard for UMTS, 3G-324M [14], uses the structure of H.324/M with a slightly different set of codecs and requirements.

13.1. Level Structure of H.324/M

H.324/M supports five different error-detection/correction schemes, called "levels," organized in a hierarchical structure. Level 0 is the same as the PSTN and ISDN variants of H.324. Levels 1 and 2 (in H.223 Annexes A and B) add robustness to the H.223 multiplex layer, while levels 3 and 3a add more error-robust adaptation layers in H.223 Annexes C and D. Each higher-numbered level is more error-robust than the previous level, and also requires more overhead and processing cycles. Each level supports the features of all lower levels. Table 4 summarizes these levels.

13.1.1. Level 0. Level 0 is identical to the base H.324 and H.223.

13.1.2. Level 1. Defined in H.223 Annex A, Level 1 replaces the 8-bit HDLC synchronization flag "01111110" with a 16-bit pseudo-noise (PN) flag "1011001010000111."

Longer flags allow more robust detection of MUX-PDU boundaries in the case of high bit error rates, because implementations can use correlation thresholds for flag detection. For example, if 15 out of 16 bits are correct, a flag may be detected. In addition, implementations should also consider the correctness of the header byte to verify the correct detection of the flag. In an optional extension, two consecutive PN flags may be used (double flag mode), permitting even more robust flag detection.

At level 1, there is no equivalent to the HDLC zero-insertion process, so emulation of the PN flag in the data stream is not prevented. If possible, transmitters should prevent the appearance of a PN flag pattern in the information field by terminating the MUX-PDU and starting a new one when a PN flag emulation occurs.

Avoiding HDLC zero-insertion ensures that bit errors cannot unsynchronize the zero-removal process, which could lead to error propagation through the entire MUX-PDU [15].

13.1.3. Level 2. Level 2, defined in H.223 Annex B, uses the 16-bit PN flag from level 1 and additionally

Table 4. Level Structure of H.324/M

Level	Multiplex Summary	Ref.
3a	16-bit PN flag	H.324, Annex C
	24-bit multiplex header error-robust AL1M′, AL2M′, AL3M′	H.223, Annexes A, B, C, D
3	16-bit PN flag	H.324, Annex C
	24-bit multiplex header error-robust AL1M, AL2M, AL3M	H.223, Annexes A, B, C
2	16-bit PN flag	H.324, Annex C
	24-bit multiplex header AL1, AL2, AL3	H.223, Annexes A, B
1	16-bit PN flag	H.324, Annex C
	8-bit multiplex header AL1, AL2, AL3	H.223, Annex A
0	8-bit HDLC flag	H.324
	8-bit multiplex header AL1, AL2, AL3 (same as base PSTN H.324)	H.223

replaces the 8-bit H.223 header with an extended 24-bit level 2 header.

The level 2 header includes a 4-bit multiplex code (the same as in base H.324), an 8-bit multiplex payload length (MPL), and a 12-bit extended Golay code.

The MPL field indicates the count of bytes in the information field. The combination of PN flag synchronization and the MPL makes it possible to overcome critical bit error or packet-loss situations that are likely in the mobile environment.

The extended Golay code is used for error detection. It can also be used for a combination of error detection and error correction, where error-detection capability decreases as error-correction increases.

The packet marker (PM) bit is not included in the MUX-PDU header, as in levels 0 and 1. Instead, the PM is signaled by inverting the following PN flag.

An optional extension provides even more redundancy by re-sending the previous MUX-PDU header, in the 8-bit level 0 header format, immediately after each level 2 header.

When the transmitter has no information to send, the "stuffing sequence," sent repeatedly to fill the channel, is the 2-byte PN flag followed by the 3-byte level 2 header, containing a multiplex code of "0000."

13.1.4. Level 3. H.223 Annex C defines level 3, which, in addition to the level 2 multiplex enhancements, replaces H.223 adaptation layers AL1, AL2, and AL3 with more error-robust "mobile" versions: AL1M, AL2M, and AL3M.

The level 3 adaptation layers can protect data with forward error correction (FEC) using a rate-compatible punctured convolutional code (RCPC) [16], and with automatic repeat request (ARQ) schemes, to allow incorrectly received data to be automatically retransmitted. A single higher-level packet can also be split into several smaller parts, to reduce the amount of data that can be damaged by uncorrected bit errors.

An optional interleaving scheme, which scrambles the order of the payload data, can be used to make the FEC more effective by spreading out the effect of a short burst of errors over a longer period. The headers of the higher-level packets are also protected with FEC, using either a systematic extended BCH code, or an extended Golay code.

The level 3 stuffing sequence is the 2-byte PN flag followed by the 3-byte level 2 header, containing a multiplex code of "1111."

13.1.5. Level 3a. Level 3a, defined in H.323 Annex D, is identical to level 3, except that the RCPC FEC codes are replaced by shortened Reed–Solomon (RS) codes. Depending on the error characteristics of the transmission channel, either RS or RCPC codes can produce a lower residual error rate.

The choice between levels 3 and 3a is selected by the H.245 protocol.

13.2. Call Setup

The H.324/M call setup is essentially the same as for the PSTN and ISDN variants of H.324. Similar to H.324/I, the V.34 modem is replaced by a suitable wireless interface, and a digital end-to-end connection is established.

The main difference is the negotiation used to select which of the different H.223 levels, just described, will be used. The "stuffing sequence," sent when the transmitter is has no information to send, is unique for each level and is therefore used to signal the sending terminal's highest supported level. Table 5 shows the stuffing sequences sent by each level.

Each terminal starts by sending consecutive stuffing sequences for the highest level it supports. Each receiver tries to detect stuffing sequences that correspond to one of the levels supported by the receiver. If either terminal receives stuffing corresponding to a level lower than the one it is sending, it changes its transmitted level to match the far-end terminal. Once both terminals detect the same stuffing sequence, the terminals operate at the same level and proceed with phase D of the H.324 call setup. (The choice between levels 3 and 3a is signaled using H.245.) This procedure guarantees that the terminals initially start at the highest commonly supported level.

H.324/M also defines an optional procedure to change levels dynamically during the call (in *phase E*). This procedure is signaled via H.245 and applies separately to each direction of transmission. For example, it is possible to operate at level 1 in one direction and at level 2 in the opposite direction. This procedure is useful if the transmission characteristics (primarily bit error rate) change, and if this doesn't happen too frequently.

13.3. Control Channel Segmentation and Reassembly Layer (CCSRL)

At levels 1 and higher a *control channel segmentation and reassembly layer* (CCSRL) is added between the NSRP and H.245 layers for logical channel 0. H.324/M has to guarantee error-free delivery for H.245 messages. In general, longer NSRP frames are more likely to be affected by errors than are shorter ones. Therefore the CCSRL layer segments long H.245 messages into smaller NSRP frames. This reduces the amount of NSRP data that needs to be retransmitted due to errors.

13.4. Videotelephony in UMTS: 3G-324M

3G-324M [14] is the videotelephony standard for the 3GPP's UMTS. H.324/M was taken as the baseline, but with a few changes, including

- Support for H.324/M level 2 is mandatory.
- The Adaptive Multirate Codec (AMR) [17], the mandatory speech codec in UMTS, is required; support for G.723.1 is also encouraged.

Table 5. Stuffing Sequences for Each Level

Level	Stuffing Sequence (Repeated)
0	HDLC flag
1	PN flag
2	PN flag + level 2 header, multiplex code 0000
3 and 3a	PN flag + level 2 header, multiplex code 1111

Figure 5. Prototype H.324/M mobile videotelephone.

- The H.263 video codec is mandatory, but support for the ISO/IEC 14496-1 (MPEG-4) [18] codec is strongly encouraged; it is expected that this will be widely supported. H.324 Annex G defines the use of MPEG-4 in H.324.

The North American and Asian-based 3GPP2 project also plans to use a variant of H.324/M for video telephony in cdma2000 [19].

14. CONCLUSION

The H.324 standard makes possible multimedia terminal devices capable of handling a variety of media types, including audio, video, and data, over switched circuit networks of all types. (Useful reference material for implementers of H.324 is available at *http://www.packetizer.com*.)

BIOGRAPHIES

Dave Lindbergh is an engineer with Polycom Inc. Since 1993, he has been active in U.S. and international standards organizations, including Committee T1, TIA, IETF, and ITU, where he was a principal contributor to ITU-T Recommendations, H.223, H.224, H.281, and V.140; served as editor for Recs. H.226 and H.324; and was chairman of the ITU-T H.324 Systems Experts Group. He currently is chairman of the IMTC Requirements Work Group and cochairman of the IMTC Media Processing Activity Group. In 2002 he received an IMTC leadership award for his leadership role in the standards community. Dave was a coauthor of *Digital Compression for Multimedia: Principles and Standards* (Morgan Kaufmann, 1998) and a contributor to *Multimedia Communications* (Academic Press, 2001). In 1990 Dave cofounded CD Atlas Company, a multimedia mapping start-up, and as a consulting engineer, he designed modem protocols and software and developed the APT (Asynchronous Performance Tester) data communications measurement tool used in the modem industry. In 1981 he founded Lindbergh Systems, maker of OMBITERM data communications software. He is credited with two U.S. patents in the field of data compression and data communications.

Bernhard G. Wimmer received the diploma degree in electrical engineering from Technical University Munich, Germany, in 1995. His diploma thesis was honored by the Siemens Information Technology User Group and by IEEE (Student Paper Award) in 1996. He joined the Corporate Research and Development of Siemens AG in 1996 as research engineer for multimedia technologies. In 1999 he changed to the division of Siemens Mobile Phones. Since 2002 he has headed the Multimedia Technology Group for 3G terminals. Since 1996 he has been active in standardization, including ETSI, IETF, ITU-T, and 3GPP, where he contributed to H.223, H.324M, H.263, 3G-324M, 3GPP PSS, and TS26.140 (MMS codecs). From 1997to 1999 he served as chairman of the ITU-T H.324 Systems Experts Group and currently as coeditor for codecs for MMS in 3GPP SA4 group. He presently is chairman of the IMTC H.324 Activity Group. His areas of interest are video/image coding algorithms, multimedia protocols, and applications for mobile devices. Wimmer is holding several patents in the area of multimedia technologies and applications.

BIBLIOGRAPHY

1. ITU-T Study Group 16, Recommendations of the H Series: H.324, H.223, H.245, G.723.1, H.263, H.226, ITU, Geneva, 1995–2001. *http://www.itu.int*.
2. ITU-T Study Group 16, Recommendations of the V Series: V.8, V.8bis, V.14, V.23, V.25, V.25ter, V.34, V.42, V.42bis, V.80, V.90, V.140, ITU, Geneva, 1990–2001, *http://www.itu.int*.
3. ISO/IEC 3309, *Information Technology — Telecommunications and Information Exchange between Systems — High-Level Data Link Control (HDLC) Procedures — Frame Structure*, 1991, *http://www.iso.ch*.
4. D. Lindbergh and H. Malvar, Multimedia teleconferencing with H.324, in K. R. Rao, ed., *Standards and Common Interfaces for Video Information Systems*, SPIE, Philadelphia, Oct. 1995, pp. 206–232.
5. ITU-T Recommendation X.691, *Information Technology; ASN.1 Encoding Rules — Specification of Packed Encoding Rules (PER)*, ITU, Geneva, 1995, *http://www.itu.int*.
6. J. Gibson et al., *Digital Compression for Multimedia*, Morgan Kaufmann, San Francisco, 1998, pp. 356–361.
7. ITU-T Study Group 8, Recommendations of the T Series: T.84, T.434, T.120, T.122–T.128, ITU, Geneva, 1992–1998. *http://www.itu.int*.
8. ISO/IEC TR9577, *Information Technology — Telecommunications Information Exchange between Systems — Protocol Identification in the Network Layer*, ISO, Geneva, 1990 *http://www.iso.ch*.
9. W. Simpson, ed., *The Point to Point Protocol*, IETF RFC 1661, July 1994, *http://www.ietf.org/rfc*.
10. ISO/IEC 13871: 1995, *Information Technology — Telecommunications and Information Exchange between*

Systems — Private Telecommunications Networks — Digital Channel Aggregation, ISO, Geneva, 1995, *http://www.iso.ch*.

11. K. Sklower et al., *The PPP Multilink Protocol (MP)*, IETF RFC 1990, Aug. 1996. *http://www.ietf.org/rfc*.
12. ITU-T Study Group 2, *Recommendation E.164*, ITU, Geneva, 1997, *http://www.itu.int*.
13. ITU-T Study Group 15, *Recommendations of the I Series*, ITU, Geneva, 1988–2000, *http://www.itu.int*.
14. 3rd Generation Partnership Project (3GPP), TSG-SA Codec Working Group, 3G TS 26.110, *Codec(s) for Circuit Switched Multimedia Telephony Service: General Description*, V.4.0.0, *http://www.3gpp.org*.
15. J. Hagenauer, E. Hundt, T. Stockhammer, and B. Wimmer, Error robust multiplexing for multimedia applications, *Signal Process. Image Commun.* **14**: 585–597 (1999).
16. J. Hagenauer, Rate-compatible punctured convolutional codes (RCPC codes) and their applications, *IEEE Trans. Commun.* **36**(4): 389–400 (April 1988).
17. 3GPP TS 26.071: *Mandatory Speech Codec; General Description*, *http://www.3gpp.org*.
18. ISO/IEC 14496-1: 1999, *Information Technology — Coding of Audio-visual Objects, Part 1: Systems*, ISO, Geneva, 1999, *http://www.iso.ch*.
19. 3GPP2, *Video Conferencing Service in cdma2000*, S.R0022, V.1.0, July 2000, *http://www.3gpp2.org*.

HADAMARD MATRICES AND CODES

MARTIN BOSSERT
University of Ulm
Ulm, Germany

1. INTRODUCTION

Hadamard matrices were born as a theoretical topic in combinatorics and have found various applications in different fields. They consist of n rows and n columns, where n is called the *order* and the matrix elements are either $+1$ or -1. The rows are pairwise orthogonal.

Their main property, the orthogonality, is among others exploited in communication systems for the separation of users. If different users simultaneously send orthogonal signals, each user can be extracted from the signal sum at the receiver. Examples therefore are the European standard UMTS (Universal Mobile Telecommunications System), the U.S. American standards IS95 (Interim Standard), GPS (Global Positioning System), and many more. Furthermore, the rows of specific Hadamard matrices possess a good periodic autocorrelation property that can be used for the resolution of multipath propagation in mobile communication.

Hadamard codes is a general name for binary codes that can be constructed on the basis of Hadamard matrices. Historically, most such codes were constructed independently and, only later, relationships to Hadamard matrices were discovered. Famous representatives of such code constructions are the class of simplex codes and the class of first- order Reed–Muller codes. The unique Golay code can also be derived based on Hadamard matrices.

In signal processing the concept of orthogonal transforms have a wide range of applications. Hadamard matrices can be used as an orthogonal transform and, as in the case of the Fourier transform, a fast Hadamard transform can be derived.

First we will give the definition and two possible constructions of Hadamard matrices, namely, the Sylvester and the Payley constructions. For $n \geq 4$ only orders n divisible by 4 may exist. By these two constructions, all but six possible Hadamard matrices of order up to 256 will be constructed. Namely, the orders 156, 172, 184, 188, 232, and 236 cannot be constructed; however, there exist constructions for these orders, for example, in Refs. 5 and 3. Of course, orders larger than 256 may be obtained by the two constructions presented.

Afterward we will describe the connections to coding theory and then derive the fast Hadamard transform for signal processing. Finally, we comment on several applications.

2. DEFINITIONS AND CONSTRUCTIONS OF HADAMARD MATRICES

Definition 1. *A Hadamard matrix, $\boldsymbol{H} = \boldsymbol{H}(n)$, of order n, is an $n \times n$ matrix with entries $+1$ and -1, such that $\mathbf{H}\mathbf{H}^T = n\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\boldsymbol{H}^T$ is the transposed matrix of $\boldsymbol{H}$.*

Example 1. As examples we give possible Hadamard matrices of orders 1, 2, and 4:

$$\boldsymbol{H}(1) = 1; \quad \boldsymbol{H}(2) = \begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix};$$

$$\boldsymbol{H}(4) = \begin{pmatrix} +1 & -1 & +1 & +1 \\ +1 & -1 & -1 & -1 \\ -1 & -1 & +1 & -1 \\ +1 & +1 & +1 & -1 \end{pmatrix}.$$

Properties. The rows of $\boldsymbol{H}(n)$ are pairwise orthogonal; that is, the scalar product of any two distinct rows is zero.

Permutations of rows or columns or inversion (multiplication by -1) of some rows and columns do not change this property. A matrix that can be derived by row or column operations from another is called *equivalent*. For any given Hadamard matrix there exists an equivalent one for which the first row and the first column consist entirely of $+1$s. Such a Hadamard matrix is called *normalized*.

Except for the cases $n = 1$ and $n = 2$, a Hadamard matrix may exist if $n = 4s$ for some integer $s > 0$. It is conjectured that a Hadamard matrix exists for all such integers n. At present, the smallest n for which no Hadamard matrix is known is 428.

2.1. The Sylvester Construction

The *Kronecker product* of matrices $\mathbf{A} = (a_{ij}), i, j \in \{1, 2, \ldots, n\}$ and $\mathbf{B}$ is defined as a blockwise matrix

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B})$$

Let $\boldsymbol{H}(n)$ and $\boldsymbol{H}(m)$ be two Hadamard matrices. Then $\boldsymbol{H}(nm) := \boldsymbol{H}(n) \otimes \mathbf{H}(m)$ is a Hadamard matrix of order nm.

This construction was given by Sylvester [1]. Note that often only constructions with $n = 2$ are called Sylvester type, however his original definition was more general.

Example 2. Let $\boldsymbol{H}(2) = \begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix}$ and $\boldsymbol{H}(n)$ be Hadamard matrices. Then the Hadamard matrix of order $2n$, using Sylvester's construction, is as follows:

$$\boldsymbol{H}(2n) = \begin{pmatrix} \boldsymbol{H}(n) & \boldsymbol{H}(n) \\ \boldsymbol{H}(n) & -\boldsymbol{H}(n) \end{pmatrix}$$

Walsh–Hadamard. In particular, the mth Kronecker powers of the normalized matrix $\boldsymbol{S}_1 := \boldsymbol{H}(2)$ gives a sequence of Hadamard matrices of the Sylvester type

$$\boldsymbol{S}_m := \boldsymbol{H}(2^m) = \boldsymbol{S}_1 \otimes \boldsymbol{S}_{m-1}, m = 2, 3, \ldots$$

which are of special interest in communications. The rows are often called Walsh or Walsh–Hadamard sequences.

2.2. The Paley Construction

A *conference matrix* $\boldsymbol{C}(n)$ of order n is an $n \times n$ matrix with diagonal entries 0 and other entries $+1$ or -1, which satisfies $\mathbf{C}\mathbf{C}^T = (n-1)\boldsymbol{I}$.

Example 3. As an example, we give a conference matrix of order 4:

$$\boldsymbol{C}(4) = \begin{pmatrix} 0 & -1 & -1 & -1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \end{pmatrix}.$$

The property $\mathbf{C}\mathbf{C}^T = (n-1)\boldsymbol{I}$ remains unchanged for permutations of rows and columns that do not change the diagonal elements or negations of rows and columns. Again, matrices constructed by such operations are called *equivalent*.

Conference matrices $\boldsymbol{C}(n)$ exist only for even n. Two special cases are distinguished: (1) if $\boldsymbol{C}(n) = -\boldsymbol{C}^T(n)$ this matrix is called a *skew-symmetric conference matrix*, which is possible only for $n \equiv 0 \pmod 4$; and (2) if $\boldsymbol{C}(n) = \boldsymbol{C}^T(n)$, this is called a *symmetric conference matrix* and $n \equiv 2 \pmod 4$.

We denote by $\boldsymbol{I}$ an identity matrix of order n. With both, symmetric and skew-symmetric conference matrices, Hadamard matrices can be constructed in the following way.

Payley Construction P1 (*Skew-Symmetric Conference Matrices*). If $n \equiv 0 \pmod 4$ and there exists a skew-symmetric conference matrix $\boldsymbol{C}(n)$, then

$$\boldsymbol{H}(n) = \boldsymbol{I} + \boldsymbol{C}(n) \tag{1}$$

is a Hadamard matrix.

Payley Construction P2 (*Symmetric Conference Matrices*). If $n \equiv 2 \pmod 4$ and there exists a symmetric conference matrix $\boldsymbol{C}(n)$, then

$$\boldsymbol{H}(2n) = \begin{pmatrix} \boldsymbol{I} + \boldsymbol{C}(n) & -\boldsymbol{I} + \boldsymbol{C}(n) \\ -\boldsymbol{I} + \boldsymbol{C}(n) & -\boldsymbol{I} - \boldsymbol{C}(n) \end{pmatrix} \tag{2}$$

is a Hadamard matrix.

Construction of Conference Matrices. Now we need to construct conference matrices $\boldsymbol{C}(n)$, and for this we will describe Paley's constructions based on quadratic residues in finite fields [2]. Note that in any field, addition and multiplication are defined and, in the following, we assume that the operations with field elements are done accordingly. Let $\mathbb{F}_q$ be a finite field, q odd. A non-zero-element $x \in \mathbb{F}_q$ is called a *quadratic residue* if an element $y \in \mathbb{F}_q$ exists such that $x = y^2$. Among the $q-1$ nonzero elements of $\mathbb{F}_q$ there exist exactly $(q-1)/2$ quadratic residues and $(q-1)/2$ quadratic nonresidues. The *Legendre* function $\chi : \mathbb{F}_q \Rightarrow \{0, \pm 1\}$ is defined by.

$$\chi(x) = \begin{cases} 0 & \text{if } x = 0 \\ +1 & \text{if } x \text{ is a quadratic residue} \\ -1 & \text{if } x \text{ is a quadratic nonresidue} \end{cases}$$

Clearly the Legendre symbol is an indicator for quadratic residues and nonresidues. A $q \times q$ matrix $\boldsymbol{Q} = (Q_{x,y})$ can be defined where the indices x,y of the entries $Q_{x,y}$ are the elements of the field, $x, y \in \mathbb{F}_q$ and their value is $Q_{x,y} := \chi(x-y)$. We denote by $\mathbf{1}$ the all one vector.

Skew-Symmetric Payley Conference Matrix. If $q \equiv 3 \pmod 4$, then

$$\boldsymbol{C}(q+1) = \begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{1}^T & \boldsymbol{Q} \end{pmatrix} \tag{3}$$

is a skew-symmetric Paley conference matrix of order $q+1$.

Symmetric Payley Conference Matrix. If $q \equiv 1 \pmod 4$, then

$$\boldsymbol{C}(q+1) = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1}^T & \boldsymbol{Q} \end{pmatrix} \tag{4}$$

is a symmetric Paley conference matrix of order $q+1$.

Clearly Eq. (1) gives Paley construction P1 for a Hadamard matrix of order $q+1$ if $q \equiv 3 \pmod 4$, and Eq. (2) gives Paley construction P2 for a Hadamard matrix of order $2(q+1)$ if $q \equiv 1 \pmod 4$. For both cases we will give an example.

Example 4. Let $q = 7$, $\mathbb{F}_q = \mathbb{F}_7 = \{0, 1, 2, 3, 4, 5, 6\}$. The elements $\{1, 2, 4\}$ are quadratic residues while $\{3, 5, 6\}$ are quadratic nonresidues. We have $q \equiv 3 \bmod 4$, and thus we will get a Hadamard matrix of order $q + 1 = 8$ using construction P1 in Eq. (1). First we construct the skew-symmetric conference matrix with the help of matrix $\boldsymbol{Q}$ defined by the Legendre symbols. The rows and

columns of $\boldsymbol{Q}$ are labeled by the elements of the field $\mathbb{F}_7 = \{0, 1, 2, 3, 4, 5, 6\}$, respectively.

$$\boldsymbol{Q} = \begin{pmatrix} 0 & 1 & 1 & -1 & 1 & -1 & -1 \\ -1 & 0 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & 0 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 0 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 0 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 0 \end{pmatrix},$$

$$\boldsymbol{C}(8) = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 0 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 0 & 1 & 1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 0 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 & -1 & 0 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 & 0 & 1 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 0 \end{pmatrix}$$

With $\boldsymbol{C}(8)$ we get the Hadamard matrix as

$$\boldsymbol{H}(8) = \boldsymbol{C}(8) + \boldsymbol{I}$$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{pmatrix}$$

Example 5. Let $q = 3^2 = 9$, $\mathbb{F}_q = \mathbb{F}_9 = \{0, 1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5, \alpha^6, \alpha^7\}$, where α is a root of the polynomial $x^2 - x - 1$. Elements $\{1, \alpha^2, \alpha^4, \alpha^6\}$ are quadratic residues, while $\{\alpha, \alpha^3, \alpha^5, \alpha^7\}$ are quadratic nonresidues. With this, we again can construct the matrix $\boldsymbol{Q}$ as

$$\boldsymbol{Q} = \begin{pmatrix} 0 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 0 & -1 & -1 & -1 & 1 & 1 & -1 & 1 \\ -1 & -1 & 0 & 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 0 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & -1 & 0 & 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 & 1 & 0 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 0 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & 1 & 0 & -1 \\ -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & 0 \end{pmatrix}$$

The symmetric conference matrix is

$$\boldsymbol{C}(10)$$

$$= \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 0 & -1 & -1 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 0 & 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 0 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 0 & 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 & 0 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 0 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 0 & -1 \\ 1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & 0 \end{pmatrix}$$

and the Hadamard matrix of order 20 is given by:

$$\boldsymbol{H}(20) = \begin{pmatrix} \mathbf{I} + \mathbf{C}(10) & -\mathbf{I} + \mathbf{C}(10) \\ -\mathbf{I} + \mathbf{C}(10) & -\mathbf{I} - \mathbf{C}(10) \end{pmatrix}$$

There exist many other constructions of Hadamard matrices (see, e.g., constructions and tables in Refs. 3–5). However, the two constructions we have presented give almost all Hadamard matrices of order ≤ 256 as shown in Table 1. The matrices are constructed by combining the Sylvester and Paley constructions. $\boldsymbol{S}_m$ denotes the Hadamard matrix $\boldsymbol{H}(2^m)$ of order 2^m constructed by recursion with the Hadamard matrix $\boldsymbol{H}(2)$. P1(q) and P2(q) denote Hadamard matrices of order $q + 1$ and $2(q + 1)$, respectively, given by the Paley constructions P1 and the P2.

3. HADAMARD CODES

A binary code C of length n, of cardinality (number of code words) M, and of minimal Hamming distance d is called an [n, M, d] code.

3.1. Binary Codes from a Hadamard Matrix

By replacing $+1$s by 0s and -1s by 1s in a normalized Hadamard matrix $\mathbf{H}(n)$ of order n, a *base* binary code

Table 1. List of Hadamard Matrices

n	Construction	n	Construction
4	$\mathbf{S}_2$	132	P1(131)
8	$\mathbf{S}_3$	136	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(68)$
12	P1(11)	140	P1(139)
16	$\mathbf{S}_4$, **P2**(7)	144	$\boldsymbol{S}_2 \otimes \boldsymbol{H}(36)$
20	P1(19), P2(3^2)	148	P2(73)
24	P1(23), $\boldsymbol{S}_1 \otimes \boldsymbol{H}(12)$	152	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(76)$
28	P2(13)	156	—
32	$\mathbf{S}_5$	160	$\boldsymbol{S}_2 \otimes \boldsymbol{H}(40)$
36	P2(17)	164	P1(163), P2(3^4)
40	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(20)$	168	P1(167)
44	P1(23)	172	—
48	P1(23), $\boldsymbol{S}_2 \otimes \boldsymbol{H}(12)$	176	$\boldsymbol{S}_2 \otimes \boldsymbol{H}(44)$
52	P2(5^2)	180	P1(179), P2(89)
56	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(28)$	184	—
60	P1(59)	188	—
64	$\boldsymbol{S}_6$	192	$\boldsymbol{S}_4 \otimes \boldsymbol{H}(12)$
68	P1(67)	196	P2(97)
72	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(36)$	200	P1(199)
76	P2(37)	204	P2(101)
80	$P1(67)$, $\boldsymbol{S}_2 \otimes \boldsymbol{H}(20)$	208	$\boldsymbol{S}_2 \otimes \boldsymbol{H}(52)$
84	P1(83)	212	P1(211)
88	P1(87), $\boldsymbol{S}_1 \otimes \boldsymbol{H}(44)$	216	—
92	—	220	P2(109)
96	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(48)$, $\boldsymbol{S}_2 \otimes \boldsymbol{H}(24)$	224	$\boldsymbol{S}_3 \otimes \boldsymbol{H}(28)$
100	P2(7^2)	228	P2(113)
104	P1(103)	232	—
108	P2(53)	236	—
112	$\boldsymbol{S}_2 \otimes \boldsymbol{H}(28)$	240	$\boldsymbol{S}_2 \otimes \boldsymbol{H}(60)$
116	—	244	P1(3^5)
120	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(60)$	248	$\boldsymbol{S}_1 \otimes \boldsymbol{H}(124)$
124	P2(61)	252	P1(251), P2(5^3)
128	$\boldsymbol{S}_7$	256	$\boldsymbol{S}_8$

$C(=[n,n,d])$ of length n and cardinality $M=n$ is obtained. Since the rows of $\mathbf{H}(n)$ are orthogonal, any two codewords of C differ exactly in $n/2$ coordinates, hence the minimal (Hamming) distance of $\mathcal{C}$ is $d=n/2$. This code is called the *Walsh–Hadamard* code. With small manipulations we can get three types of code from the base code $\mathcal{C}$.

1. $\mathcal{C}1[n-1, M=n, d=n/2]$. Note that the first column of $\mathbf{H}(n)$ consists of $+1$s, since we assumed a normalized Hadamard matrix. Correspondingly, the first coordinate of any codeword of $\mathcal{C}$ is equal to 0. So it can be deleted without changing the cardinality and minimal distance. Thus one gets an $[n-1, M=n, d=n/2]$ code $\mathcal{C}1$.
2. $\mathcal{C}2[n, M=2n, d=n/2]$. Let $\mathcal{C}2$ be a code consisting of all codewords of the base code $\mathcal{C}$ and all their complements. Then $\mathcal{C}2$ is an $[n, M=2n, d=n/2]$ code.
3. $\mathcal{C}3[n-1, M=2n, d=n/2-1]$. Let $\mathcal{C}3$ be a code obtained from the code $\mathcal{C}2$ by deleting first coordinates. Then $\mathcal{C}3$ is an $[n-1, M=2n, d=n/2-1]$ code.

3.2. The Plotkin Bound and Hadamard Matrices

The Plotkin bound states that for an [*n, M, d*] binary code

$$\begin{aligned} &M \le 2\left\lfloor \frac{d}{[2d-n]} \right\rfloor && \text{if } d \text{ is even}, 2d > n \\ &M \le 4d && \text{if } d \text{ is even}, n = 2d \\ &M \le 2\left\lfloor \frac{d+1}{[2d+1-n]} \right\rfloor && \text{if } d \text{ is odd}, 2d+1 > n \\ &M \le 4d+4 && \text{if } d \text{ is odd}, n = 2d+1. \end{aligned} \tag{5}$$

There exists a conjecture that for any integer $s>0$ there exists a Hadamard matrix $\boldsymbol{H}(4s)$. Levenstein proved that if this conjecture is true then codes exist meeting the Plotkin bound (5) (for details, see Ref. 6).

3.3. Simplex Codes

The class of codes $\mathcal{C}1$ is known as simplex codes, because the Hamming distance of every pair of codewords is the same. Codewords can be geometrically interpreted as vertices of a unit cube in $n-1$ dimensions and form a regular simplex. If $n=2^m$ and $\mathcal{C}1$ is obtained from the Hadamard matrix $\boldsymbol{S}_m = \boldsymbol{S}_1 \otimes \boldsymbol{S}_{m-1}$, then $\mathcal{C}1$ is a *linear* $(2^m-1, m, 2^{m-1})$ code[1] known also as a *maximal-length feedback shift register code*. Any nonzero codeword is known also as an *m-sequence*, or as a *pseudonoise* (PN) *sequence*. If $\mathbf{c}=(c_0, c_1, \ldots, c_2 m_{-2})$ is a nonzero codeword, then

$$c_s = c_{s-1}h_{m-1} + c_{s-2}h_{m-2} + \cdots + c_{s-m+1}h_1 + c_{s-m},$$
$$s = m, m+1, \ldots,$$

where subscripts are calculated (mod 2^m-1), and

$$1 + h_1 x + \cdots + h_{m-1}x^{m-1} + x^m$$

is a primitive irreducible polynomial.

Simplex codes have many applications, such as range-finding, synchronizing, modulation, scrambling, or pseudonoise sequences. The dual code of a simplex code is the Hamming code, which is a perfect code with minimal distance 3. In the application section below, we will describe the use of m sequences. For further information, see also Ref. 7.

3.4. Reed–Muller Codes of First Order

Codes $\mathcal{C}2$ obtained from the Hadamard matrix $\boldsymbol{S}_m = \boldsymbol{S}_1 \otimes \boldsymbol{S}_{m-1}$ are *linear* $(2^m, m+1, 2^{m-1})$ codes. They are known also as first-order Reed–Muller codes and are denoted by $\mathcal{R}(1,m)$. The all-zero vector $\mathbf{0}$ and the all-one vector $\mathbf{1}$ are valid codewords. All the other codewords have Hamming weight 2^{m-1}.

The encoding is the mapping between an information vector into a codeword. This mapping can be a multiplication of the information vector by a so-called generator matrix. One possible encoding of Reed–Muller codes is based on a special representation of the generator matrix $\mathbf{G}_m$ consisting of $m+1$ rows and 2^m columns. The first row $\mathbf{v}_0$ of $\mathbf{G}$ is the all-one vector $\mathbf{1}$. Consider this vector as a run of 1s of length 2^m. The second vector $\mathbf{v}_1$ is constructed by replacing this run by two runs of identical length 2^{m-1} where the first run consists of 0s and the second of 1s. In the third stage ($\mathbf{v}_2$), each run of length 2^{m-1} is replaced by two runs of 0s and 1s with lengths 2^{m-2}. The procedure is continued in a similar manner until one gets the last row $\mathbf{v}_m$ of the form $0101\cdots0101$.

Example 6. *Let $m=4, n=2^m=16$. Then the generator matrix of $\mathcal{R}(1,m)$ is as follows*:

$$\mathbf{G}_m = \begin{pmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \end{pmatrix} = \begin{pmatrix} 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1 \\ 0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1 \\ 0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1 \\ 0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1 \end{pmatrix}$$

If the information bits are $u_0, u_1, \ldots, u_m$, then the corresponding codeword is

$$\mathbf{c} = u_0\mathbf{v}_0 + u_1\mathbf{v}_1 + \cdots + u_m\mathbf{v}_m$$

For the code $\mathcal{R}(1,m)$, a systematic encoding is also possible as an extended cyclic code [7].

The decoding can be performed by the fast Hadamard transform described below. However, there exist several decoding methods; so-called soft decoding, which uses reliability information on the code symbols (if available) is also possible [7]. In the following we describe the multistep

[1] In case of linear codes one can give the dimension k instead of the number of codewords M. It is $M=2^k$ for binary linear codes.

majority-logic decoding algorithm (the Reed algorithm), using Example 6. Let u_0, u_1, u_2, u_3, u_4 be information bits and $c_0, c_1, \ldots, c_{15}$ be the corresponding code symbols. Note that

$$u_1 = \begin{cases} c_0 + c_8 \text{ or} \\ c_1 + c_9 \text{ or} \\ c_2 + c_{10} \text{ or} \\ c_3 + c_{11} \text{ or} \\ c_4 + c_{12} \text{ or} \\ c_5 + c_{13} \text{ or} \\ c_6 + c_{14} \text{ or} \\ c_7 + c_{15} \end{cases}, \quad u_2 = \begin{cases} c_0 + c_4 \text{ or} \\ c_1 + c_5 \text{ or} \\ c_2 + c_6 \text{ or} \\ c_3 + c_7 \text{ or} \\ c_8 + c_{12} \text{ or} \\ c_9 + c_{13} \text{ or} \\ c_{10} + c_{14} \text{ or} \\ c_{11} + c_{15} \end{cases},$$

$$u_3 = \begin{cases} c_0 + c_2 \text{ or} \\ c_1 + c_3 \text{ or} \\ c_4 + c_6 \text{ or} \\ c_5 + c_7 \text{ or} \\ c_8 + c_{10} \text{ or} \\ c_9 + c_{11} \text{ or} \\ c_{12} + c_{14} \text{ or} \\ c_{13} + c_{15} \end{cases}, \quad u_4 = \begin{cases} c_0 + c_1 \text{ or} \\ c_2 + c_3 \text{ or} \\ c_4 + c_5 \text{ or} \\ c_6 + c_7 \text{ or} \\ c_8 + c_9 \text{ or} \\ c_{10} + c_{11} \text{ or} \\ c_{12} + c_{13} \text{ or} \\ c_{14} + c_{15} \end{cases}.$$

These equations give 8 votes for each information bit u_1, u_2, u_3, u_4 and the majority of the votes determines the value of each information bit. If 3 or less errors occurred during transmission then all those bits will be decoded correctly. It remains to decode u_0. Since $\mathbf{c} - (u_1\mathbf{v}_1 + u_2\mathbf{v}_2 + u_3\mathbf{v}_3 + u_4\mathbf{v}) = u_0\mathbf{v}_0$, we have an equation to determine u_0 by the majority rule.

One of the first applications of coding was the first order Reed–Muller code of length 32, which was used in the beginning of the 1970s for data transmission in the *Mariner* 9 mission.

3.5. Quadratic Residue Codes

Binary codes obtained from Hadamard matrices of the Paley type are nonlinear for $n > 8$. The linear span of these codes results in *linear quadratic residue codes*. Most known codes are linear spans of codes $\mathcal{C}1$ and $\mathcal{C}3$ obtained from Paley constructions $P1$ for prime q. Such linear codes have length q, dimension $k = \frac{1}{2}(q \pm 1)$, and minimum distance satisfying the inequality $d^2 - d + 1 \geq q$. The linear quadratic residue code obtained from $\boldsymbol{H}(24)$ is the famous perfect Golay (23,12,7) code.

4. THE HADAMARD TRANSFORM

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a real-valued vector. If a Hadamard matrix $\boldsymbol{H}(n)$ exists, then the vector $\mathbf{y} = \mathbf{x}\mathbf{H}(n)$ is said to be its *Hadamard* transform (or *Walsh–Hadamard* transform). The Hadamard transform is used in communications and physics, mostly for $n = 2^m$ and $\boldsymbol{H}(n) = \boldsymbol{S}_m$. In general, the Hadamard transform with $\boldsymbol{S}_m$ requires $2^m \times 2^m$ additions and subtractions. A computationally more efficient implementation is the *fast Hadamard transform* based on the representation of $\boldsymbol{S}_m$ as a product of sparse matrices. Specifically, define for $1 \leq i \leq m$ matrices by

$$\mathbf{M}_n^{(i)} := \boldsymbol{I}_{2^{m-i}} \otimes \boldsymbol{S}_1 \otimes \boldsymbol{I}_{2^{i-1}}$$

where $\boldsymbol{I}_s$ denotes an identity matrix of order s. Such a matrix has exactly two nonzero entries ($+1$ or -1) in each row and in each column.

Example 7. Let $m = 3, n = 8$. Then

$$\mathrm{M}_8^{(1)} = \begin{pmatrix} +1 & +1 & & & & & & \\ +1 & -1 & & & & & & \\ & & +1 & +1 & & & & \\ & & +1 & -1 & & & & \\ & & & & +1 & +1 & & \\ & & & & +1 & -1 & & \\ & & & & & & +1 & +1 \\ & & & & & & +1 & -1 \end{pmatrix},$$

$$\mathrm{M}_8^{(2)} = \begin{pmatrix} +1 & & +1 & & & & & \\ & +1 & & +1 & & & & \\ +1 & & -1 & & & & & \\ & +1 & & -1 & & & & \\ & & & & +1 & & +1 & \\ & & & & & +1 & & +1 \\ & & & & +1 & & -1 & \\ & & & & & +1 & & -1 \end{pmatrix},$$

$$\mathrm{M}_8^{(3)} = \begin{pmatrix} +1 & & & & +1 & & & \\ & +1 & & & & +1 & & \\ & & +1 & & & & +1 & \\ & & & +1 & & & & +1 \\ +1 & & & & -1 & & & \\ & +1 & & & & -1 & & \\ & & +1 & & & & -1 & \\ & & & +1 & & & & -1 \end{pmatrix}.$$

The matrix $\boldsymbol{S}_m$ can be written as

$$\boldsymbol{S}_m = \mathbf{M}_n^{(1)}\mathbf{M}_n^{(2)} \cdots \mathbf{M}_n^{(m)}$$

Thus, to evaluate the Hadamard transform $\mathbf{y} = \mathbf{x}\mathbf{S}_\mathbf{m} = \mathbf{x}\mathbf{M}_\mathbf{n}^{(1)}\mathbf{M}_\mathbf{n}^{(2)} \cdots \mathbf{M}_\mathbf{n}^{(\mathbf{m})}$, one calculates in the first stage $\mathbf{y}_1 = \mathbf{x}\mathbf{M}_\mathbf{n}^{(1)}$. This requires 2^m additions and subtractions. Next, $\mathbf{y}_2 = \mathbf{y}_1\mathbf{M}_\mathbf{n}^{(2)}$ is obtained with the same number of calculations and so on. In the mth stage, one obtains $\mathbf{y}$ with a total number of $m2^m$ calculations.

5. COMMUNICATION APPLICATIONS

In this section we describe two main principles as examples of the various applications of Hadamard matrices and related codes and sequences. Example 8 is for user separation in communication systems; Example 9 is for correlation in distance measurement or to measure the channel impulse response.

5.1. Code-Division Multiple Access (CDMA)

In CDMA communication systems, such as the standards IS95 and UMTS, the user separation is done with Walsh–Hadamard sequences. Hereby, the link from the mobiles to the base station uses other means of user separation but the so-called downlink (base station to mobiles) uses Walsh–Hadamard sequences. First we consider a small example.

Example 8. Each user is assigned one row of the Hadamard matrix. In the case of $\boldsymbol{H}(4)$ we can have four users:

$$\textit{Alice}: +1, +1, +1, +1 \quad \textit{Bob}: +1, -1, +1, -1$$
$$\textit{Charlie}: +1, +1, -1, -1 \quad \textit{Dan}: +1, -1, -1, +1$$

Now for each user, information bits can be sent by using her/his sequence or the inverted (negative) sequence, for example, in the case of Charlie $+1, +1, -1, -1$ for bit zero and $-1, -1, +1, +1$ for bit one. We assume that there is a synchronous transmission of information for all users. For Alice, Bob, and Dan, bit 0 is send and for Charlie a one. Each of the four user's receiver gets the sum of all signals

$$(+2, -2, +2, +2) = (+1, +1, +1, +1) + (+1, -1, +1, -1) + (-1, -1, +1, +1) + (+1, -1, -1, +1)$$

Now the receiver can detect which bit was send by so called correlation of the user's sequence with the received sequence as follows. For example, Bob computes

$$(+1) \cdot (+2) + (-1) \cdot (-2) + (+1) \cdot (+2) + (-1) \cdot (+2) = +4$$

From the result $+4$, Bob concludes that a zero (the sequence itself) was transmitted.

For Charlie we get

$$(+1) \cdot (+2) + (+1) \cdot (-2) + (-1) \cdot (+2) + (-1) \cdot (+2) = -4$$

From the result -4, Charlie concludes that the transmission was a one (the inverted sequence).

For the description of the CDMA principle, we restrict ourselves to real-valued sequences. Then the crosscorrelation $\Phi_{\mathbf{x},\mathbf{y}}$ of two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n is defined by

$$\Phi_{x,y} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

Clearly, with this definition, the crosscorrelation $\Phi_{\boldsymbol{x},\boldsymbol{y}} = 0$, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal and $\Phi_{\boldsymbol{x},\boldsymbol{y}} = 1$ if $\boldsymbol{x} = \boldsymbol{y}$. Let $\mathbf{r} = u_1\boldsymbol{x}_1 + u_2\boldsymbol{x}_2 + \cdots + u_n\boldsymbol{x}_n$ be a received signal, where $u_1, u_2, \ldots, u_n$ are the information bits $\{+1, -1\}$ and the $\boldsymbol{x}_i$ are the rows of a Hadamard matrix. Then the CDMA principle gives

$$\Phi_{\mathbf{r},\boldsymbol{x}_i} = u_i, \quad i = 1, 2, \ldots, n$$

because of the orthogonality of the sequences.

This shows that the users do not mutually interfere. This is a direct consequence of the use of orthogonal sequences. However, in practice there is noise and multipath propagation, which introduces difficulties within this concept. The IS95 standard in the downlink uses Walsh–Hadamard sequences of order 64, while in the uplink the same set of sequences are used as for a Reed–Muller code. The UMTS standard uses Walsh–Hadamard sequences with variable orders, namely, between 4 and 256 in order to adapt to variable data rates and transmission conditions.

5.2. Measurement of the Channel Impulse Response and Scrambling

For user separation, the receiver evaluates the crosscorrelation function between the received sequence and each user sequence. Measurement of distances or the channel impulse response exploits the autocorrelation function. Again we restrict ourself to real-valued sequences. Let $\boldsymbol{x}^{(k)}$ be the cyclic shift of $\boldsymbol{x}$ by k positions. Then, the autocorrelation is defined as

$$\Phi_{\boldsymbol{x}}(k) = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^{(k)}$$

The autocorrelation can take values between -1 and $+1$. In the case of PN sequences the autocorrelation has either the value 1 for $k = 0$ and $k = n$ or

$$\Phi_{\boldsymbol{x}}(k) = \begin{cases} -\dfrac{1}{n} & 0 < k < n, n \text{ odd}, \\ -\dfrac{1}{n-1} & 0 < k < n, n \text{ even}, \end{cases}$$

The mobile communication channel suffers from multipath propagation when the receiver gets the sum of delayed and attenuated versions of the transmitted signal. Usually it is assumed that the channel is time-invariant for a small period and can be described by the impulse response of a linear time-invariant system. We will explain the measurement of the channel impulse response using an example.

Example 9. We assume that the sender transmits periodically the PN sequence $\boldsymbol{pn} = +1, +1, -1, -1, -1, +1, -1$ of length 7. Therefore the transmitted signal is (";" is used to mark the period):

$$\boldsymbol{x} = \ldots; +1, +1, -1, -1, -1, +1, -1; +1, +1, -1, -1, -1, +1, -1; \ldots$$

Suppose the receiver gets

$$y_i = x_i + \tfrac{1}{2}x_{i-2} = \ldots; \tfrac{3}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{3}{2}, \tfrac{1}{2}, -\tfrac{3}{2}; \tfrac{3}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}, \ldots$$

If we correlate the received sequence with the PN sequence, we get the following results:

$(\ldots; \frac{3}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}, \frac{1}{2}, -\frac{3}{2}; \ldots)$ *correlated with* $\boldsymbol{pn}$ *is* $\frac{13}{14}$

$(\ldots, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}, \frac{1}{2}, -\frac{3}{2}; \frac{3}{2}, \ldots)$ *correlated with* $\boldsymbol{pn}$ *is* $-\frac{3}{14}$

$(\ldots, -\frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}, \frac{1}{2}, -\frac{3}{2}; \frac{3}{2}, \frac{1}{2}, \ldots)$ *correlated with* $\boldsymbol{pn}$ *is* $\frac{5}{14}$

$(\ldots, -\frac{1}{2}, -\frac{3}{2}, \frac{1}{2}, -\frac{3}{2}; \frac{3}{2}, \frac{1}{2}, -\frac{1}{2}, \ldots)$ *correlated with* $\boldsymbol{pn}$ *is* $-\frac{3}{14}$

The other three values of the period will also be $-\frac{3}{14}$. Thus the influence of the channel can be measured.

In many CDMA systems, base stations transmit pilot tones consisting of PN sequences. Using the measurement method presented above, the mobile terminals can determine the channel impulse response and use it to adjust a so-called RAKE receiver. Because of the good autocorrelation and good cross-correlation properties of PN sequences, they are also used to scramble the transmitted signals. Clearly the autocorrelation property reduces the influence of the paths on each other and the crosscorrelation property guarantees the user separation. Thus, if we know the channel paths at the receiver, we can use a correlator for each path and add their results. This is the concept of a RAKE receiver. In IS95, a PN sequence of length $2^{42} - 1$ is used for this purpose, and in UMTS, scrambling codes are used on the basis of the PN sequences of the polynomials $x^{41} + x^3 + 1$ and $x^{41} + x^{20} + 1$.

The GPS system exploits both the good autocorrelation and crosscorrelation properties of PN sequences. Several satellites synchronously send their PN sequences. A GPS receiver exploits the crosscorrelation to separate the signals from different satellites and the autocorrelation to determine the timing offsets between different satellites caused by signal propagation. Combining these timing offsets with information on the actual position of the satellites allows the precise calculations of the receiver's position. The public PN sequences of length 1023 used in GPS are generated by the polynomials $x^{10} + x^3 + 1$ or $x^{10} + x^9 + x^8 + x^6 = x^3 + x^2 + 1$.

Acknowledgment

The author would like to acknowledge the valuable hints and help from Professor Ernst Gabidulin from the Institute of Physics and Technology, Moscow, Russia.

BIOGRAPHY

Martin Bossert received the Dipl.-Ing. degree in electrical engineering from the Technical University of Karlsruhe, Germany, in 1981, and a Ph.D. degree form the Technical University of Darmstadt, Germany, in 1987. After a DFG scholarship at Lingkoeping University, Sweden, he joined AEG Mobile Communication, Ulm, Germany, where he was, among others, involved in the specification and developement of the GSM System. Since 1993, he has been a professor with the University of Ulm, Germany, where he is currently head of the Departement of Telecommunications and Applied Information Theory. His main research areas concern secure and reliable data transmission, especially generalized concatenation/coded modulation, code constructions for block and convolutional codes, and soft-decision decoding.

BIBLIOGRAPHY

1. J. J. Sylvester, Thoughts on orthogonal matrices, simultaneous sign-successions, and tessellated pavements in two or more colours, with applications to Newton's rule, ornamental tile-work, and the theory of numbers, *Phil. Mag.* **34**: 461–475 (1867).
2. R. E. A. C. Paley, On orthogonal matrices, *J. Math. Phys.* **12**: 311–320 (1933).
3. W. D. Wallis, A. P. Street, and J. Seberry Wallis, Combinatorics: Room squares, sum-free sets, Hadamard matrices, in *Lecture Notes in Mathematics*, Springer, Berlin, 1972, p. 292.
4. J. Seberry and M. Yamada, Hadamard matrices, sequences, and block designs, in J. H. Dinitz and D. R. Stinson, eds., *Contemporary Design Theory: A Collection of Surveys*, Wiley, New York, 1992, pp. 431–560.
5. R. Craigen, Hadamard matrices and designs, in C. J. Colbourn and J. H. Dinitz, eds., *The CRC Handbook of Combinatorial Designs*, CRC Press, New York, 1996, pp. 370–377.
6. F. J. MacWilliams and N. J. A. Sloan, *The Theory of Error-Correcting Codes*, (3rd printing) North-Holland, 1993.
7. M. Bossert, *Channel Coding for Telecommunications*, Wiley, New York, 1999.

HELICAL AND SPIRAL ANTENNAS

HISAMATSU NAKANO
Hosei University
Koganei, Tokyo, Japan

1. INTRODUCTION

This article discusses the radiation characteristics of various helical and spiral antennas. For this, numerical techniques applied to these antennas are summarized in Section 2, where fundamental formulas to evaluate the radiation characteristics are presented. Section 3 on helical antennas, is composed of three subsections, which present the radiation characteristics of normal-mode, axial-mode, and conical-mode helical antennas, respectively. Section 4, on spiral antennas, is composed of five subsections; Sections 4.2 and 4.3 *qualitatively* describe the radiation mechanism of spiral antennas, and Sections 4.4 and 4.5 *quantitatively* refer to the radiation characteristics of the spirals. Finally, Section 5 presents additional information on helical and spiral antennas: a backfire-mode helix and techniques for changing the beam direction of a spiral antenna.

2. NUMERICAL ANALYSIS TECHNIQUES

This section summarizes numerical analysis techniques for helical and spiral antennas. The analysis is based on an electric field integral equation [1,2]. Using the current distribution obtained by solving the integral equation, the radiation characteristics, including the radiation field, axial ratio, input impedance, and gain, are formulated.

2.1. Current on a Wire

Figure 1 shows an arbitrarily shaped wire with a length of L_{arm} (from S_0 to S_E) in free space. It is assumed that the wire is thin relative to an operating wavelength and the current flows only in the wire axis direction. It is also assumed that the wire is perfectly conducting and hence the tangential component of the electric field on the wire

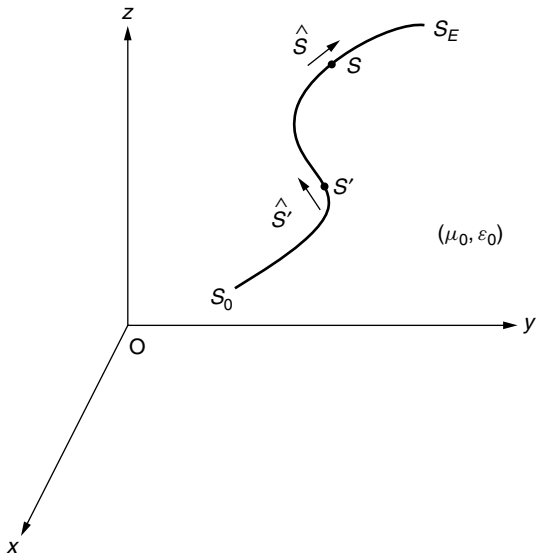


Figure 1. Arbitrarily shaped wire.

surface is zero. This boundary condition of the electric field leads to an *integral equation*.

$$\frac{1}{j\omega\varepsilon_0}\int_{S_0}^{S_E} I(s')\left[-\frac{\partial^2 G(s,\,s')}{\partial s\,\partial s'}+\beta^2 G(s,s')\hat{s}\bullet\hat{s}'\right]ds' = -E_s^i(s) \tag{1}$$

where $j^2=-1$; ω is the angular frequency ($=2\pi f$, $f=$ frequency); ε_0 is the permittivity of free space; s and s' are the distances measured along the wire from the origin (that can arbitrarily be chosen on the wire) to an observation point and a source point, respectively; $I(s')$ is the current at the source point; β is the phase constant ($=2\pi/\lambda$, $\lambda=$ wavelength); $\hat{s}$ and $\hat{s}'$ are unit vectors, parallel to the wire axis, at the observation and source points, respectively; $E_s^i(s)$ is the tangential component of an incident electric field on the wire; and $G(s,s')$ is Green's function, which is defined as

$$G(s,s') = \frac{1}{4\pi}\frac{e^{-j\beta r_{o,s}(s,s')}}{r_{o,s}(s,s')} \tag{2}$$

where $r_{o,s}(s,s')$ is the distance between the observation and source points.

The *method of moments* (MoM) [3] is adopted to obtain the current $I(s')$ in Eq. (1). For this, the current is expanded as

$$I(s') = \sum_{n=1}^{N} I_n J_n(s') \tag{3}$$

where $J_n(s')$ and I_n $(n=1,2,\ldots,N)$ are called the "expansion functions" and "unknown coefficients of the expansion functions," respectively. Note that one can arbitrarily choose $J_n(s')$. Therefore, $J_n(s')$ are known functions in Eq. (3).

Substituting Eq. (3) into Eq. (1), one obtains

$$\sum_{n=1}^{N} I_n e_n(s) = -E_s^i(s) \tag{4}$$

where

$$e_n(s) = \frac{1}{j\omega\varepsilon_0}\int_{S_0}^{S_E} J_n(s')\left[-\frac{\partial^2 G(s,s')}{\partial s\,\partial s'}+\beta^2 G(s,s')\hat{s}\bullet\hat{s}'\right]ds' \tag{5}$$

Multiplying both sides of Eq. (1) by functions $W_m(s)(m=1,2,\ldots,N)$ and integrating the multiplied results over the wire length from S_0 to S_E, one obtains

$$\sum_{n=1}^{N} I_n Z_{mn} = V_m, \qquad m=1,2,\ldots,N \tag{6}$$

where

$$Z_{mn} = \int_{S_0}^{S_E} e_n(s) W_m(s) ds \tag{7}$$

$$V_m = -\int_{S_0}^{S_E} E_s^i(s) W_m(s) ds \tag{8}$$

Note that one can arbitrarily choose $W_m(s)$, which are called "weighting functions." When the $W_m(s)$ have the same form as the expansion functions $J_m(s')$, the MoM is called the "Galerkin method."

Equation (6) is written in matrix form:

$$[Z_{mn}][I_n] = [V_m] \tag{9}$$

where $[Z_{mn}]$, $[I_n]$, and $[V_m]$ are called the "impedance, current, and voltage matrices," respectively. The unknown coefficients are obtained as

$$[I_n] = [Z_{mn}]^{-1}[V_m] \tag{10}$$

Substituting the obtained $I_n(n=1,2,\ldots,N)$ into Eq. (3), one can determine the current distributed along the wire.

2.2. Radiation Field, Axial Ratio, Input Impedance, and Gain

The electric field $\mathbf{E}$ at a far-field point, radiated from the current $I(s')$ and called the "radiation field," is calculated to be

$$\mathbf{E}(r,\theta,\phi) = E_\theta(r,\theta,\phi)\hat{\theta} + E_\phi(r,\theta,\phi)\hat{\phi} \tag{11}$$

where

$$E_\theta(r,\theta,\phi) = -\frac{j\omega\mu_0}{4\pi r}e^{-j\beta r}\hat{\theta}\bullet\int_{S_0}^{S_E}\hat{s}' I(s') e^{j\beta\hat{\mathbf{r}}\bullet\mathbf{r}'}ds' \tag{12}$$

$$E_\phi(r,\theta,\phi) = -\frac{j\omega\mu_0}{4\pi r}e^{-j\beta r}\hat{\phi}\bullet\int_{S_0}^{S_E}\hat{s}' I(s') e^{j\beta\hat{\mathbf{r}}\bullet\mathbf{r}'}ds' \tag{13}$$

in which (r,θ,ϕ) and $(\hat{r},\hat{\theta},\hat{\phi})$ are the spherical coordinates and their unit vectors, respectively; μ_0 is the permeability of free space; and the vector $\mathbf{r}'$ is the position vector where current $I(s')$ exists. Other notations are defined in Section 2.1.

The radiation field of Eq. (11) is decomposed into two circularly polarized (CP) wave components:

$$\mathbf{E}(r,\theta,\phi) = E_R(r,\theta,\phi)(\hat{\theta}-j\hat{\phi}) + E_L(r,\theta,\phi)(\hat{\theta}+j\hat{\phi}) \tag{14}$$

where the first term represents *a right-hand CP wave component* and the second represents *a left-hand CP wave*

component. Using these two components, the axial ratio (AR) is defined as $\mathrm{AR} = \{|E_R| + |E_L|\}/\{\|E_R| - |E_L\|\}$. The AR is an indicator of the uniformity of a CP wave. Note that $\mathrm{AR} = 1$ (0 dB) when the radiation is *perfectly* circularly polarized.

The input impedance Z_{in} is defined as $Z_{\mathrm{in}} = R_{\mathrm{in}} + jX_{\mathrm{in}} = V_{\mathrm{in}}/I_{\mathrm{in}}$, where V_{in} and I_{in} are the voltage and current at antenna feed terminals, respectively. The power input to the antenna is given as $P_{\mathrm{in}} = R_{\mathrm{in}}|I_{\mathrm{in}}/\sqrt{2}|^2$. The gain G is defined as

$$G(\theta, \phi) = \frac{|\mathbf{E}(r, \theta, \phi)/\sqrt{2}|^2/Z_0}{P_{\mathrm{in}}/4\pi r^2}$$

$$G(\theta, \phi) = \frac{|\mathbf{D}(\theta, \phi)|^2}{60P_{\mathrm{in}}} \tag{15}$$

where Z_0 is the intrinsic impedance of free space ($Z_0 = 120\ \Omega$) and $\mathbf{D}(\theta, \phi)$ is defined as

$$\mathbf{D}(\theta, \phi) = \left(\frac{r}{e^{-j\beta r}}\right)\mathbf{E}(r, \theta, \phi) \tag{16}$$

$\mathbf{D}(\theta, \phi)$ is decomposed into two components as $\mathbf{E}(r, \theta, \phi)$ in Eq. (14):

$$\mathbf{D}(\theta, \phi) = D_R(\theta, \phi)(\hat{\theta} - j\hat{\phi}) + D_L(\theta, \phi)(\hat{\theta} + j\hat{\phi}) \tag{17}$$

Therefore, the gains for right- and left-hand CP waves are calculated by $G_R(\theta, \phi) = |D_R(\theta, \phi)|^2/30P_{\mathrm{in}}$ and $G_{\mathrm{L}}(\theta, \phi) = |D_L(\theta, \phi)|^2/30P_{\mathrm{in}}$, respectively.

3. HELICAL ANTENNAS

Figure 2 shows a helical arm, which is specified by the pitch angle α, the number of helical turns n, and the circumference of the helix C. Helical antennas are classified into three groups in terms of the circumference C relative to a given wavelength λ [4]: a normal-mode helical antenna ($C \ll \lambda$), an axial-mode helical antenna ($C \approx \lambda$), and a conical-mode helical antenna ($C \approx 2\lambda$). The normal-mode helical antenna radiates a linearly polarized wave. The axial-mode and conical-mode helical antennas radiate CP waves. The beam direction for each mode is illustrated using arrows in Fig. 3.

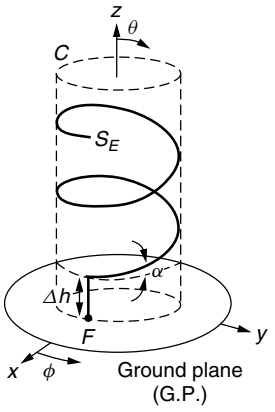


Figure 2. Helical arm.

The ground plane in Fig. 3, which backs each helical arm, is very large and assumed to be of infinite extent in the following analysis. This assumption allows the use of image theory, where the ground plane is removed. The removal of the ground plane enables one to use the techniques described in Section 2.

3.1. Normal-Mode Helical Antenna

The radiation when the circumference C is very small relative to a given wavelength is investigated in this subsection. For this, a circumference of $C = 0.0422\lambda_{0.5}$ (25.3 mm) is chosen for the antenna structure shown in Fig. 3a, together with a pitch angle of $\alpha = 40^\circ$ and number of helical turns $n = 4.5$, where $\lambda_{0.5}$ is the wavelength at a test frequency of 0.5 GHz = 500 MHz. The total arm length L_{arm}, including an initial wire length of $\Delta h =$ 1.5 mm, is $\frac{1}{4}\lambda_{0.5}$. Therefore, the antenna characteristics are expected to be similar to those of a monopole antenna of one-quarter wavelength.

Figure 4 shows the current $I(= I_r + jI_i)$ distributed along the helical arm at 0.5 GHz. The helical arm is chosen to be thin: wire radius $\rho = 0.001\lambda_{0.5}$. It is found that the current distribution is a standing wave, as seen from the phase distribution. Note that the phase is calculated from $\tan^{-1}(I_i/I_r)$.

The radiation field from the current distribution at 0.5 GHz is shown in Fig. 5, where parts (a) and (b) are the radiation patterns in the x–z and y–z planes, respectively, and part (c) is the azimuth radiation pattern in the horizontal plane ($\theta = 90^\circ, 0^\circ \le \phi \le 360^\circ$). It is clearly seen that the helical antenna radiates a linearly polarized wave: $E_\theta \neq 0$ and $E_\phi = 0$. The maximum radiation is in the horizontal direction ($\theta = 90^\circ$), where the polarization is in the antenna axis (z-axis) direction. The radiation field component E_θ in the horizontal plane is omnidirectional. The gain in the x direction is calculated to be approximately 4.9 dB.

Additionally, Fig. 6 shows the radiation patterns at 0.5 GHz as a function of the number of helical turns n, where the pitch angle and circumference remain unchanged: $\alpha = 40^\circ$ and $C = 0.0422\lambda_{0.5}$. It is found that, as n increases, the radiation beam becomes sharper.

3.2. Axial-Mode Helical Antenna

When the frequency is chosen to be 11.85 GHz, the physical circumference of the helix, $C = 25.3$ mm used in Section 3.1, corresponds to a length of one wavelength (1λ). An antenna having this circumference is analyzed in this subsection, using helical configuration parameters of $\alpha = 12.5^\circ$ and $n = 15$. The total arm length, including an initial wire length of $\Delta h = 1$ mm, is $L_{\mathrm{arm}} = 15.4\lambda_{11.85}$, where $\lambda_{11.85}$ is the wavelength at a test frequency of 11.85 GHz.

Figure 7 shows the current $I(= I_r + jI_i)$ at 11.85 GHz, where the helical arm is chosen to be thin: wire radius $\rho = 0.001\lambda_{11.85}$. It is found that the current distribution has three distinct regions: a region from the feed point F to point P_1, a region from point P_1 to point P_2, and a region from point P_2 to the arm end S_E. The amplitude of the current in each region shows a different form.

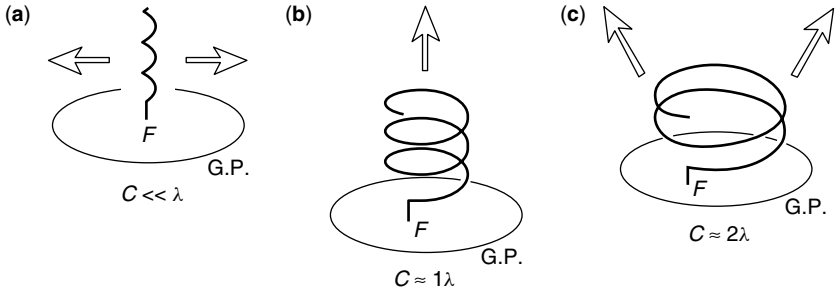


Figure 3. Helical antennas: (**a**) a normal-mode helical antenna ($C \ll \lambda$); (**b**) an axial-mode helical antenna ($C \approx \lambda$); and (**c**) a conical-mode helical antenna ($C \approx 2\lambda$).

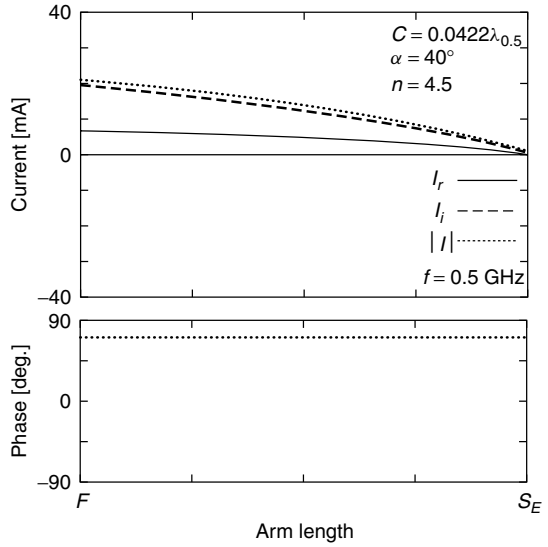


Figure 4. Current distribution of a normal-mode helical antenna.

The amplitude $|I|$ in the first region F–P_1 shows a rapid decay. Point P_1 is the position where the current first becomes minimal. The current in this region F–P_1 is a traveling wave, as seen from the phase progression. It is noted that the first region acts as an *exciter* for the remaining helical turns. This is proved by the fact that the current distribution remains almost the same even when the helical wire is cut at point P_1 [5].

The second region P_1–P_2 is the region called the "director," to which part of the power in the first region F–P_1 is guided. The amplitude of the current, $|I|$, in the second region is relatively constant. This is obviously different from the amplitude of the current in the first region. Detailed calculations reveal that the phase velocity of the current in the second region is such that the field in the z direction from each turn adds nearly in phase over a wide frequency bandwidth [6].

The outgoing current flowing along the director (the forward current) reaches the arm end S_E and is reflected. As a result, the current forms a standing wave near the arm end. The third region P_2–S_E reflects this fact. The reflected current from the arm end S_E is not desirable for a CP antenna, because it radiates a CP wave whose rotational sense is opposite that of the forward current, resulting in degradation of the axial ratio.

Figure 8 shows radiation patterns at 11.85 GHz, where the electric field at a far-field point is decomposed into right- and left-hand CP wave components. The forward current traveling toward the arm end generates the copolarization component. The copolarization component for this helical antenna is a right-hand CP wave component E_R. The component E_L for this helical antenna is called the "cross-polarization component." The cross-polarization component is generated by the undesired reflected current. The axial ratio and gain in the z direction are calculated to be approximately 0.8 and 11.5 dB, respectively.

Additionally, Fig. 9 shows the gains for a right-hand CP wave at 11.85 GHz as a function of the number of helical turns, n, for various pitch angles α, where the

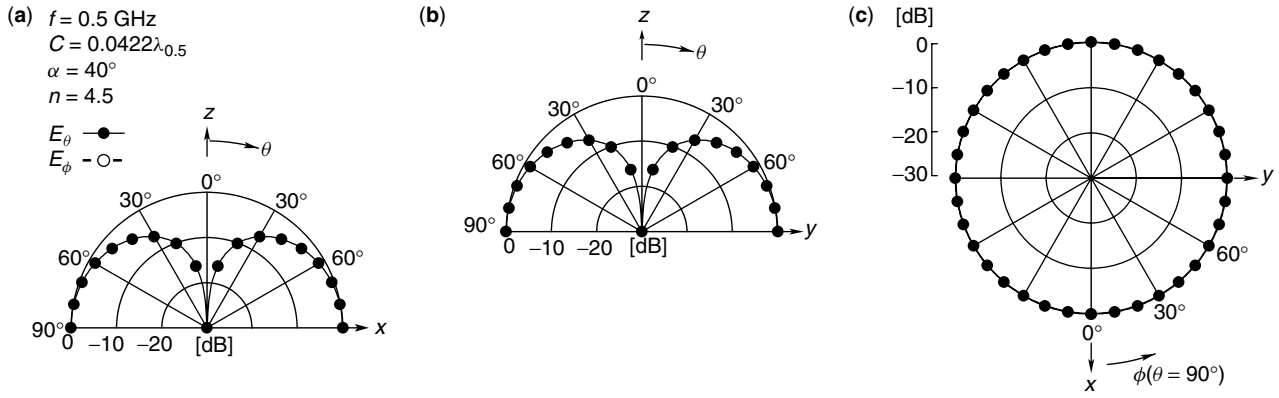


Figure 5. Radiation patterns of a normal-mode helical antenna: (**a**) in the x–z plane; (**b**) in the y–z plane; (**c**) in the horizontal plane ($\theta = 90°$, $0° \le \phi \le 360°$).

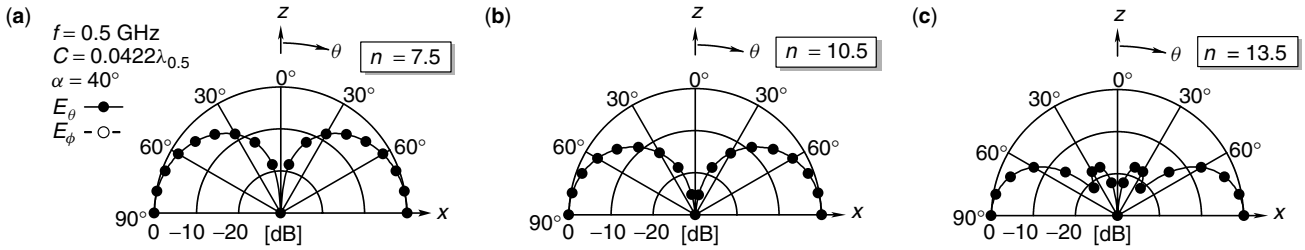


Figure 6. Radiation pattern of a normal-mode helical antenna as a function of number of helical turns: (**a**) $n = 7.5$; (**b**) $n = 10.5$; (**c**) $n = 13.5$.

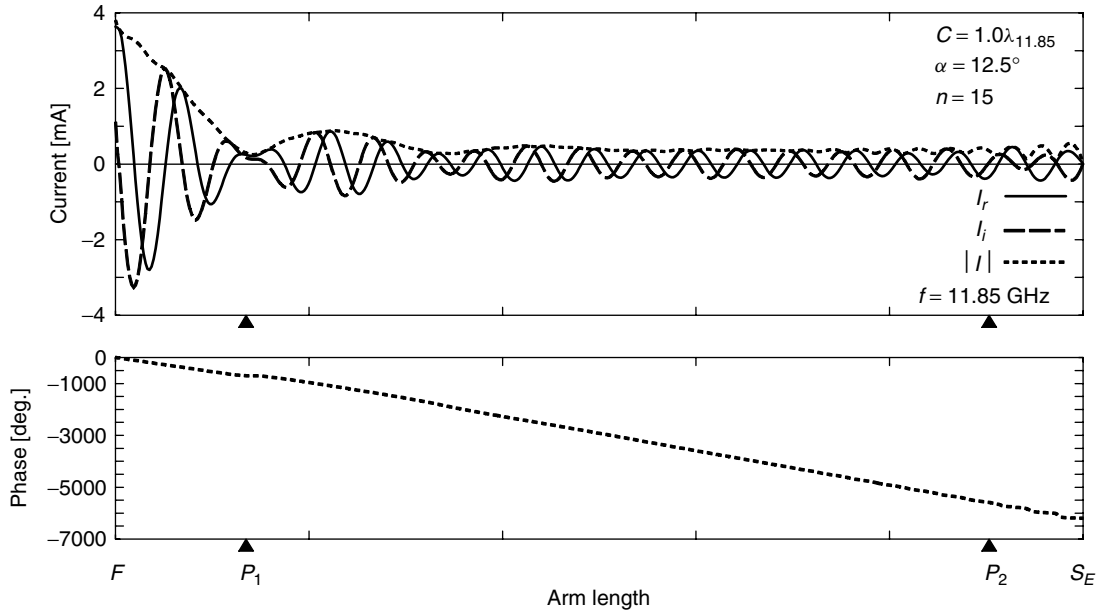


Figure 7. Current distribution of an axial-mode helical antenna.

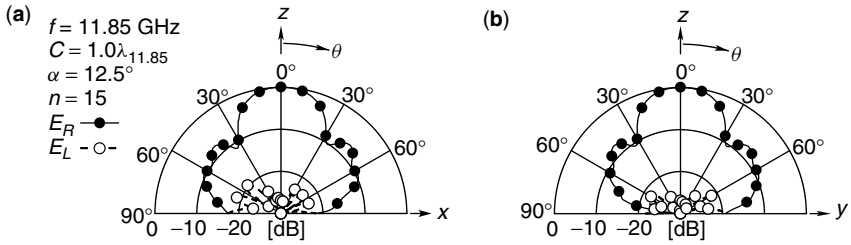


Figure 8. Radiation patterns of an axial-mode helical antenna: (**a**) in the x–z plane; (**b**) in the y–z plane.

circumference is kept constant: $C = 1.0\lambda_{11.85}$. Note that each gain increases as n increases. However, the gain increase is bounded; that is, there is a maximum gain value for a given pitch angle α.

3.3. Conical-Mode Helical Antenna

When the helical arm has a circumference of approximately two wavelengths (2λ), the radiation from the helix forms a conical beam [7,8]. To reflect this fact, a frequency of 23.7 GHz is used for an antenna with a circumference of $C = 25.3$ mm $= 2.0\lambda_{23.7}$, where $\lambda_{23.7}$ is the wavelength at 23.7 GHz. Other configuration parameters are arbitrarily chosen as follows: pitch angle $\alpha = 4^\circ$ and number of helical turns $n = 2$. The total arm length, including the initial wire length $\Delta h = 1$ mm, is $L_{\text{arm}} = 4.09\lambda_{23.7}$.

Figure 10 shows the current distribution along the helical arm, whose wire radius is $\rho = 0.001\lambda_{23.7}$. As observed in the first region F–P_1 of the axial-mode helical antenna (see Fig. 7), the current is a decaying traveling wave, which radiates a CP wave.

It is obvious that local radiation from the helical arm forms the total radiation. If each turn of the helix is a local radiation element and approximated by a loop whose circumference is two wavelengths, the currents along n local loops produce a zero field in the z direction and a maximum radiation off the z axis; that is, the total

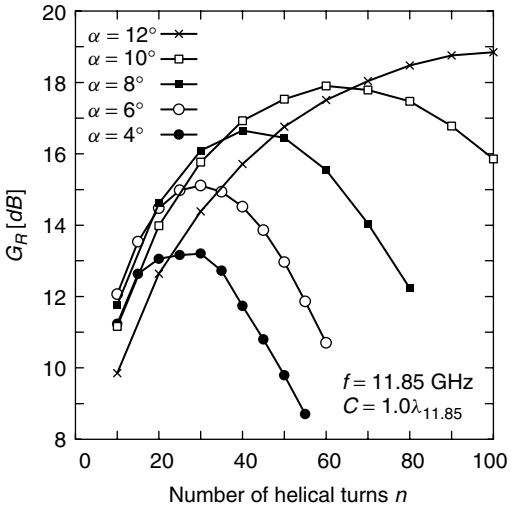


Figure 9. Gains as a function of the helical turns for various pitch angles.

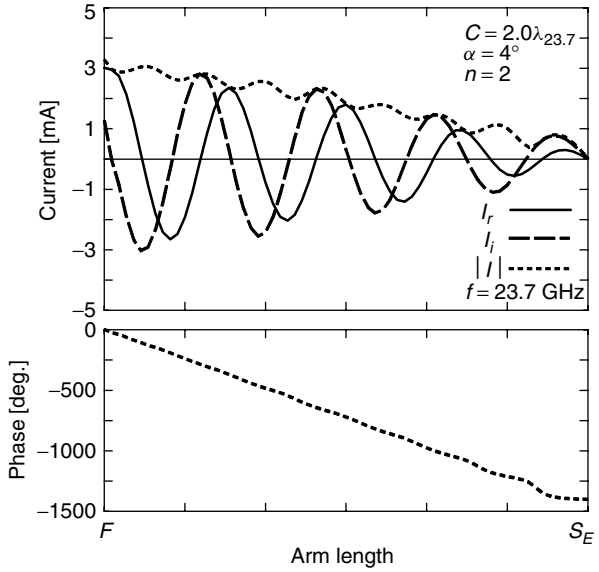


Figure 10. Current distribution of a conical-mode helical antenna.

radiation forms a conical beam. Figure 11 reflects this fact, where the radiation in the beam direction ($\theta = 34^\circ$) is circularly polarized. Note that the azimuth pattern of Fig. 11c is calculated in a plane of ($\theta = 34^\circ, 0^\circ \leq \phi \leq 360^\circ$), where the copolarization component E_R shows an omnidirectional pattern.

4. SPIRAL ANTENNA

A spiral antenna is an element that radiates a CP wave over a wide frequency bandwidth [9–12]. The mechanism of the CP wave radiation is investigated for two cases of excitation: (1) antiphase and (2) in-phase. First, the CP radiation is qualitatively explained in terms of *current band theory* [9], which is based on transmission line theory; second, the numerical results (quantitative results) obtained by the MoM are presented and discussed.

4.1. Configuration

Figure 12 shows the configuration of a two-arm spiral antenna. The two arms, A and B, are symmetrically wound with respect to the centerpoint o. The radial distance from the centerpoint o to a point on arm A is defined as $r_A = a_{\text{sp}}\phi_{\text{wn}}$, where a_{sp} and ϕ_{wn} are the spiral constant and winding angle, respectively. The winding angle starts at ϕ_{st} and ends at ϕ_{end}. Similarly, the radial distance from the centerpoint o to a point on arm B is defined as $r_B = a_{\text{sp}}(\phi_{\text{wn}} - \pi)$, with a starting angle of $\phi_{\text{st}} + \pi$ and an ending angle of $\phi_{\text{end}} + \pi$. It is noted that the spiral specified by r_A and r_B is an *Archimedean spiral antenna*.

The spiral antenna is fed from terminals T_A and T_B in the center region. The mechanism of CP wave radiation is qualitatively explained by current band theory [9], in which it is assumed that the two arms A and B are tightly wound. It is also assumed that the currents along the two arms gradually decrease, radiating electromagnetic power into free space.

To apply current band theory, four points P_A, P'_A, P_B, and P'_B are defined as follows. P_A and P'_A are points on arm A, and P_B and P'_B are points on arm B. Points P_A and P_B are symmetric with respect to the centerpoint o. P'_A is a point on arm A and a neighboring point of P_B. P'_B is a point on arm B and a neighboring point of P_A.

4.2. Antiphase Excitation

Discussion is devoted to the radiation from the spiral when terminals T_A and T_B are fed with the same amplitude and a phase difference of 180°. The excitation is called "antiphase excitation," which is realized by inserting a voltage source between terminals T_A and T_B.

4.2.1. First-Mode Radiation. The current along arm A travels through point P_A and reaches arm end E_A. Similarly, the current along arm B travels through point P_B and reaches arm end E_B. The phase of the current at point P_B always differs from that at point P_A by 180°, because of the antiphase excitation at terminals T_A and T_B. This is illustrated using arrows at P_A and P_B in Fig. 13. Note that the direction of the arrow at P_B is opposite that of the arm growth, and the two arrows at P_A and P_B are in the same direction.

An interesting phenomenon is found when points P_A and P_B are located on a ring region, with a center circumference of one wavelength (λ), in the spiral plane. Points P_A and P_B in this case are separated by approximately one-half wavelength (0.5λ) along the circumference. The current at P_A and the current at its neighboring point P'_B on arm B are approximately in phase, because the current traveling from point P_B to point P'_B along arm B (traveling by approximately one-half wavelength, since the spiral arms are tightly wounded) experiences a phase change of approximately 180°. Similarly, the current at P_B and the current at its neighboring point P'_A on arm A are approximately in phase. Figure 13 illustrates these four currents at points P_A, P'_B,

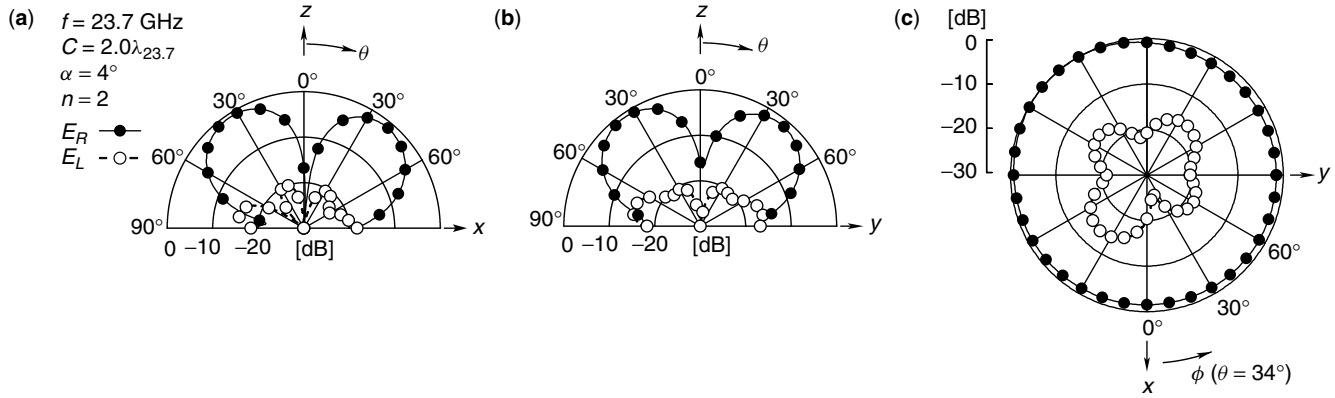


Figure 11. Radiation patterns of a conical-mode helical antenna: (**a**) in the x–z plane; (**b**) in the y–z plane; (**c**) in a plane of ($\theta = 34^\circ, 0^\circ \le \phi \le 360^\circ$).

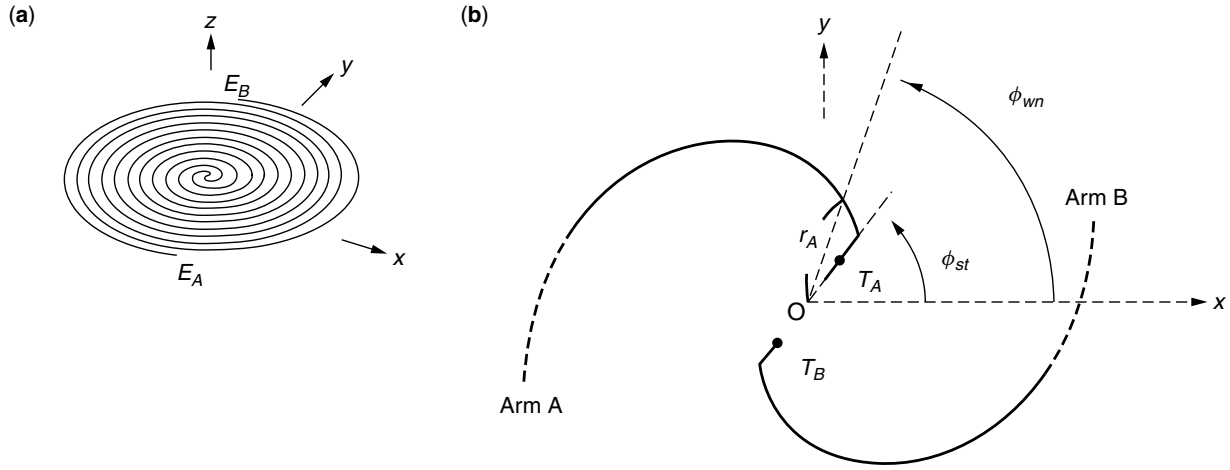


Figure 12. A two-wire spiral antenna: (**a**) perspective view; (**b**) top view.

P_B, and P'_A, where each pair of currents forms a band of current.

The two current bands in Fig. 13 rotate around the centerpoint o with time. This means that the electric field radiated from each current band also rotates. In other words, the radiation field is circularly polarized. The two circularly polarized waves radiated from the two current bands are in phase on the z axis, resulting in maximum radiation on the z axis. This radiation is called "first-mode radiation."

4.2.2. Third-Mode Radiation. When points P_A and P_B are located on a ring region of three-wavelength (3λ) circumference in the spiral plane (see Fig. 14), points P_A and P_B are separated by 1.5 wavelengths along this 3λ circumference. Therefore, the current along arm B experiences a phase change of approximately $360^\circ + 180^\circ$ from point P_B to point P'_B. As a result, the currents at points P_A and P'_B are approximately in phase; that is, a current band is formed. Similarly, the currents at points P_B and P'_A are approximately in phase, and a current band is formed.

The currents on arms A and B become in-phase with a period of one-half wavelength along the 3λ circumference.

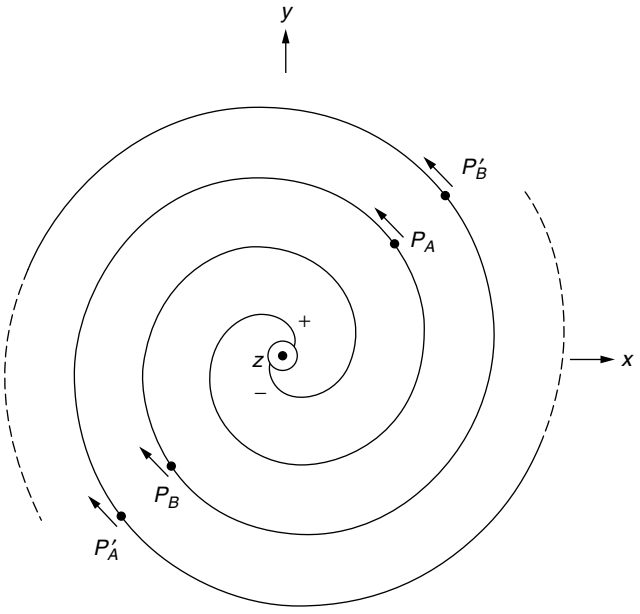


Figure 13. Current bands for first-mode radiation.

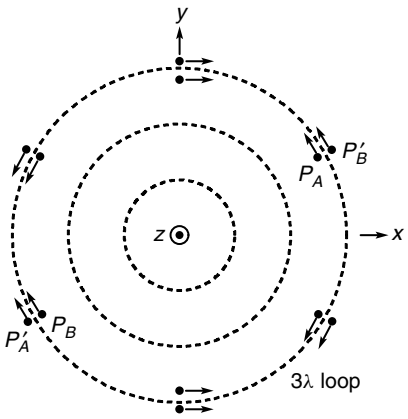


Figure 14. Current bands for third-mode radiation.

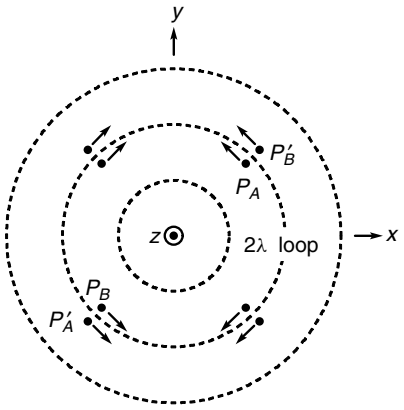


Figure 15. Current bands for second-mode radiation.

As a result, four more current bands are formed between points P_A and P_B, as shown in Fig. 14. The directions of all current bands along the 3λ circumference cause the electric fields, rotating with time, to cancel on the z axis and add off the z axis. This radiation is called "third-mode radiation."

4.2.3. Odd-Mode Radiation. So far, first-mode radiation and third-mode radiation have been discussed under the condition that the excitation at terminals T_A and T_B is antiphase. The first- and third-mode radiation components result from the current bands formed over two regions of 1λ and 3λ circumferences in the spiral plane, respectively. The mechanism described in the previous section leads to the formation of higher odd mth-mode radiation ($m = 5, 7, \ldots$) as long as the currents exist over regions of circumference of $m\lambda$ in the spiral plane. Note that each higher odd mth-mode radiation component becomes maximal off the z axis.

As the currents leave the regions of circumference of $m\lambda (m = 1, 3, 5, \ldots)$ in the spiral plane, the in-phase condition of the neighboring currents on arms A and B becomes destructive. The destructive currents contribute little to the radiation. The radiation, therefore, is characterized by a sum of odd-mode radiation components that the spiral supports.

4.3. In-Phase Excitation

When terminals T_A and T_B are excited in antiphase, the spiral has odd-mode radiation components, as discussed in the previous subsection. Now the radiation when terminals T_A and T_B are excited in phase (terminals T_A and T_B are excited with the same amplitude and the same phase) is considered. Realization of in-phase excitation is found in Section 4.5.

4.3.1. Second-Mode Radiation. A situation where points P_A and P_B are located on a ring region of 2λ circumference in the spiral plane is investigated. The phases of the currents at points P_A and P_B are the same due to two facts; terminals T_A and T_B are excited in phase, and the distance from terminal T_A to point P_A along arm A is equal to that from terminal T_B to point P_B along arm B. Arrows P_A and P_B in Fig. 15 illustrate the phases of the currents at points P_A and P_B. The directions of the arrows at points P_A and P_B are in the arm-growth direction and opposite each other.

The current traveling from points P_B to P'_B experiences a phase change of approximately 360°, because the distance from points P_B to P'_B along arm B is approximately one wavelength. It follows that the direction of the arrow for the current at point P'_B is in the arm-growth direction, as shown in Fig. 15. The arrows at points P_A and P'_B have the same direction, and indicate the formation of a current band.

Similarly, the current traveling from points P_A to P'_A experiences a phase change of approximately 360°, due to an approximately one-wavelength difference between these points. An arrow at point P'_A in Fig. 15 shows this fact. The arrows at points P'_A and P_B have the same direction, and indicate the formation of a current band.

The two middle points between points P_A and P_B along the 2λ circumference are separated from P_A (and P_B) by one-half wavelength. A current band is formed at each middle point. It follows that four current bands are formed over a ring region of 2λ circumference in the spiral plane. As seen from the direction of the arrows in Fig. 15, the radiation from these four current bands is zero on the z axis and maximal off the z axis. This radiation is called "second-mode radiation."

4.3.2. Even-Mode Radiation. One can conclude from the previous observation that the spiral with in-phase excitation does not have current bands in the ring region whose circumference is 3λ in the spiral plane. The phase relationship of the currents over the 3λ ring region is destructive (out of phase), thereby not forming current bands. However, as the currents on arms A and B further travel toward their arm ends, the phase relationship of the currents gradually becomes constructive. When the currents reach a ring region whose circumference is four wavelengths in the spiral plane, the phases of the neighboring currents become in-phase. Again current bands are formed. This radiation is called "fourth-mode radiation." Similarly, higher even-mode radiation occurs until the currents die out. It is noted that even mth-mode

radiation ($m = 2, 4, \ldots$) have zero intensity on the z axis and maximal off the z axis.

4.4. Numerical Results of a Spiral Antenna with Antiphase Excitation

The radiation mechanisms of a spiral antenna have been *qualitatively* discussed in Sections 4.2 and 4.3. In this subsection, the radiation characteristics of a spiral antenna with antiphase excitation are *quantitatively* obtained on the basis of the numerical techniques presented in Section 2.

The configuration parameters of the spiral are chosen as follows: spiral constant $a_{\text{sp}} = 0.0764$ cm/rad and winding angle ϕ_{wn} ranging from $\phi_{\text{st}} = 2.60$ rad to $\phi_{\text{end}} = 36.46$ rad. These configuration parameters lead to an antenna diameter of $2\pi a_{\text{sp}}\phi_{\text{end}} = 17.5 \text{ cm} = 3.5\lambda_6$, where λ_6 is the wavelength at a frequency of 6 GHz. In other words, the spiral at a frequency of 6 GHz includes a ring region of three-wavelength circumference in the spiral plane. Note that the wire radius of the spiral arm is $\rho = 0.012\lambda_6$.

Figure 16 shows the current $I(= I_r + jI_i)$ distributed along arm A at a frequency of 6 GHz. Since the currents on arms A and B are symmetric with respect to the centerpoint o (note that the distance between terminals T_A and T_B is assumed to be infinitesimal), this figure shows only the current along arm A. It is found that the current decreases, traveling toward the arm end. The phase progression of the current close to that in free space ($= -2\pi s/\lambda_6$, where s is the distance measured along the spiral arm from the centerpoint o to an observation point). This means that the wavelength of the current along the arm, called the "guided wavelength λ_g," is close to the wavelength propagating in free space, λ_6.

Figure 17 shows the radiation patterns in the x–z plane and y–z plane. The spiral equally radiates waves in the $\pm z$ hemispheres. The radiation has a maximum value in the $\pm z$ directions and is circularly polarized. The axial ratio in the $\pm z$ directions is approximately 0.1 dB, and the gain is approximately 5.5 dB.

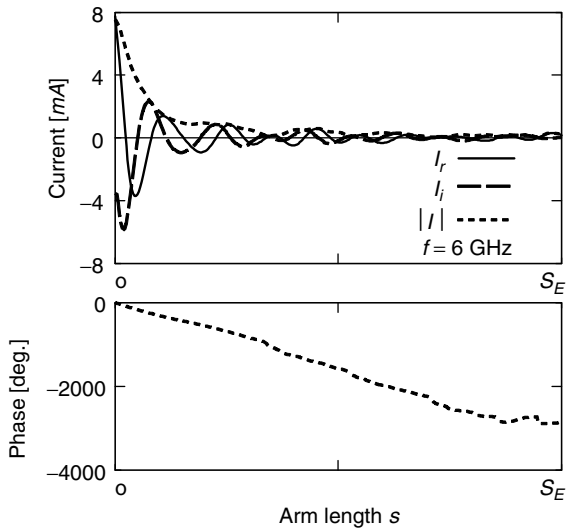


Figure 16. Current distribution of a spiral antenna with antiphase excitation.

So far, the antenna characteristics at a frequency of 6 GHz have been discussed. Now, the frequency responses of the radiation characteristics are investigated. Figure 18 shows the axial ratio (AR) in the positive z direction as a function of frequency, together with the gain for a right-hand CP wave, G_R, in the positive z direction. It is observed that, as the frequency decreases, the axial ratio deteriorates. This is due to the fact that, as the frequency decreases, the ring region of one-wavelength circumference in the spiral plane moves toward the periphery of the spiral and finally disappears. With the movement of the one-wavelength circumference ring, the polarization of the radiation becomes elliptical, that is, the axial ratio increases.

Figure 19 shows the input impedance $Z_{\text{in}}(= R_{\text{in}} + jX_{\text{in}})$ as a function of frequency. The input impedance is relatively constant over a wide frequency bandwidth. Note that the input impedance is always $Z_{\text{in}} = 60\pi\,\Omega$ when the following two conditions are satisfied: (1) arms A and B, each made of a *strip* conductor, are infinitely wound; and (2) the spacing between the two arms equals the width of the strip conductor. The antenna satisfying these conditions is called the "self-complementary antenna" [13].

4.5. Numerical Results of a Spiral Antenna with In-Phase Excitation

Figure 20 shows a spiral antenna with in-phase excitation, where a round conducting disk, approximated by wires for analysis (see Fig. 20c), is used for exciting the spiral. A voltage source is inserted between the spiral and the conducting disk. The spiral is a backed by a conducting plane of infinite extent.

The configuration parameters are as follows: spiral constant $a_{\text{sp}} = 0.04817$ cm/rad, winding angle ϕ_{wn} ranging from $\phi_{st} = 8\pi$ rad to $\phi_{\text{end}} = 37.5$ rad, wire radius $\rho = 0.00314\lambda_6$, disk diameter $D_{\text{disk}} = 0.49\lambda_6$, spacing between spiral and disk $H = 0.046\lambda_6$, and spacing between spiral and conducting plane $H_r = \frac{1}{4}\lambda_6$. The spiral at 6 GHz includes a ring region of two-wavelength circumference in the spiral plane. Note that a ring region of one-wavelength circumference does not contribute to the radiation, and hence the arms inside the one–wavelength circumference are deleted, as shown in Fig. 20a.

Figure 21 shows the radiation at a frequency of 6 GHz. The radiation occurs in the positive z hemisphere because the conducting plane is of infinite extent. As expected from current band theory, the maximum value of the radiation is off the z axis, as shown in Fig. 21a. Figure 21b shows the azimuth radiation pattern at $\theta = 40^\circ$ (beam direction angle from the z axis). The variation in the azimuth radiation component E_R is very small. The axial ratio in a plane of ($\theta = 40^\circ$, $0^\circ \le \phi \le 360^\circ$) is presented in Fig. 21c. It is concluded that the spiral forms a circularly polarized conical beam.

5. ADDITIONAL INFORMATION

The axial-mode helix in Section 3.2 has been analyzed under the condition that its ground plane is of infinite extent. An interesting phenomenon is observed when

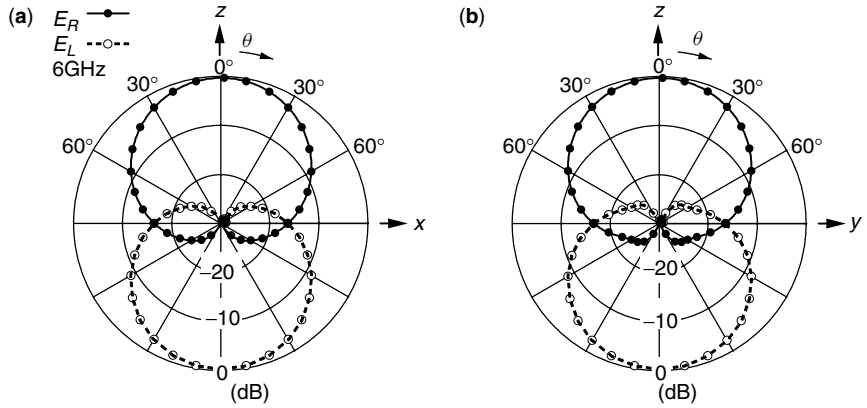


Figure 17. Radiation patterns of a spiral antenna with antiphase excitation: (**a**) in the x–z plane; (**b**) in the y–z plane.

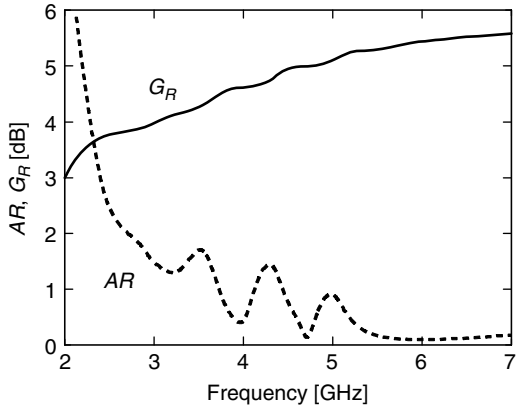


Figure 18. Axial ratio and gain as a function of frequency.

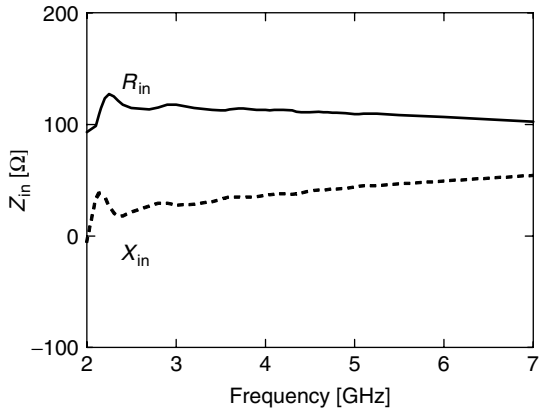


Figure 19. Input impedance as a function of frequency.

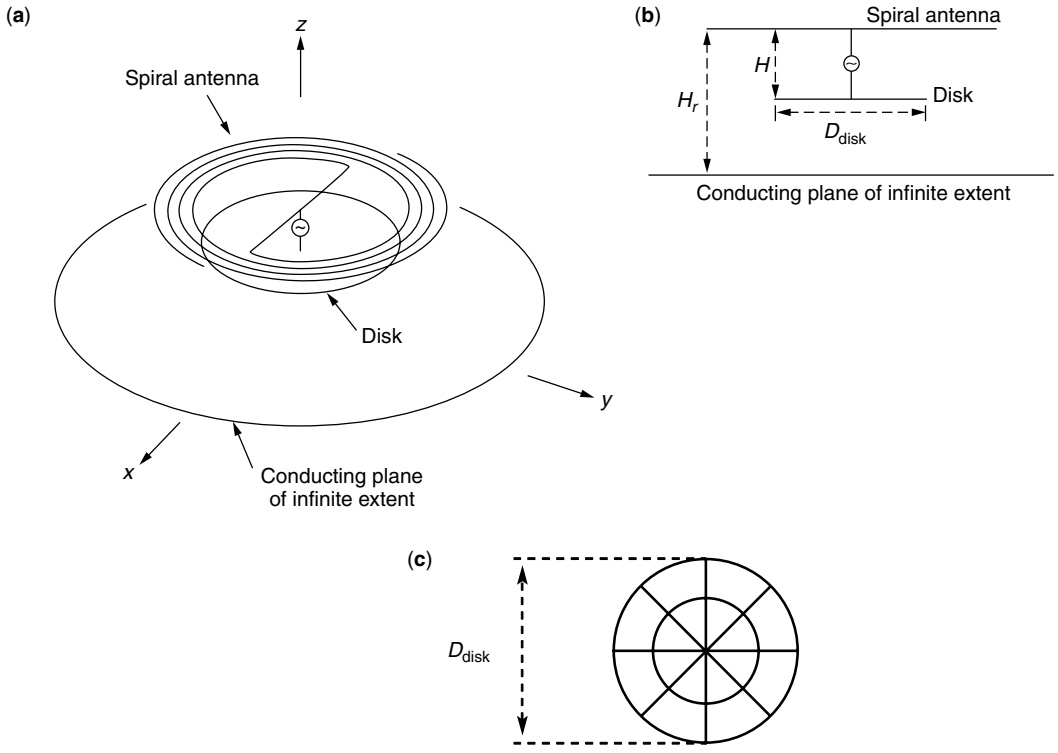


Figure 20. A spiral antenna with in-phase excitation: (**a**) perspective view; (**b**) side view; (**c**) disk approximated by wires.

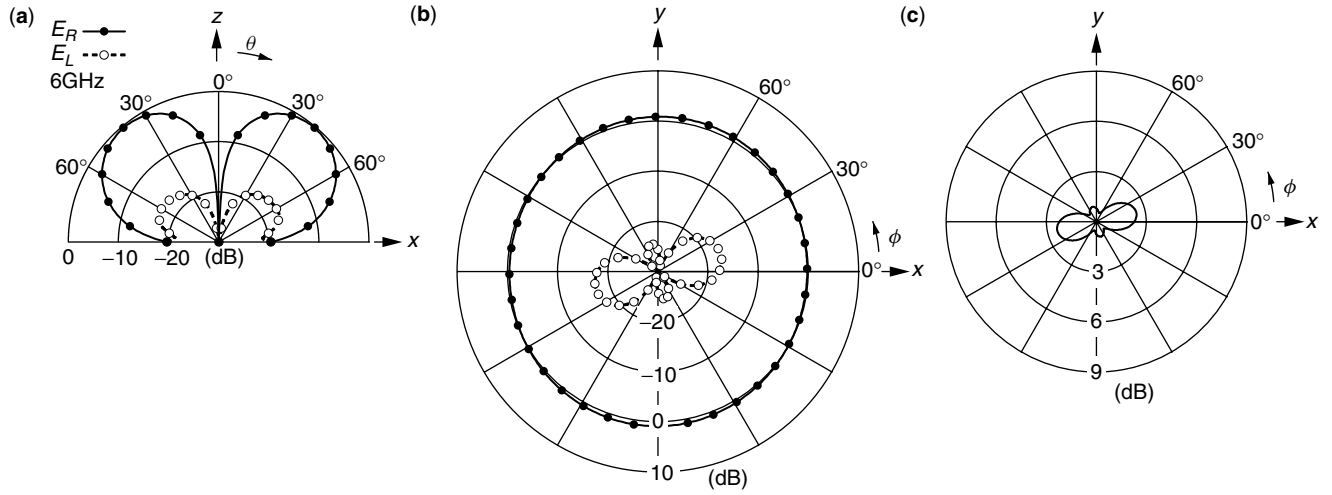


Figure 21. Radiation from a spiral antenna with in-phase excitation: (**a**) radiation pattern in the x–z plane; (**b**) radiation pattern in a plane of ($\theta = 40^\circ$, $0^\circ \leq \phi \leq 360^\circ$); and (**c**) axial ratio pattern in a plane of ($\theta = 40^\circ$, $0 \leq \phi \leq 360^\circ$).

the size of the ground plane is made comparable to the circumference of the helix — the helix radiates a CP wave in the *backward* direction (in the negative z direction). This is called the "backfiremode." An application of the backfiremode is found in Ref. 14, where the helix is used as a CP primary feed for a parabolic reflector. This reflector antenna has a high gain of more than 30 dB with an aperture efficiency of more than 70%. It is widely used for *direct broadcasting satellite* (DBS) signal reception. For more recent research on helical antennas, including the analysis of a helix wound on a dielectric rod and four helices with a cavity, readers are directed to Refs. 15 and 16.

The spiral antenna discussed in Section 4.4 has a, "bidirectional beam"; that is, the radiation occurs in the $\pm z$-hemispheres, as shown in Fig. 17. It is possible to change the bi-directional beam to a "unidirectional beam" (radiation in the $+z$ hemisphere) using two techniques. One is to use a conducting plane, as shown in Fig. 22a, and the other is to use a cavity, as shown in Fig. 22b.

The conducting plane in Fig. 22a is put behind the spiral usually with a spacing of one-quarter wavelength (0.25λ) at the operating center frequency. The gain is increased because the rear radiation is reflected by the conducting plane and added to the front radiation. However, the conductor plane affects the wideband characteristics of the spiral. To keep the antenna characteristics over a wide frequency bandwidth, it is recommended that the spacing between the spiral and the conducting plane be small and absorbing material be inserted between the outermost arms and the conducting plane [12]. A theoretical analysis for this case using a finite-difference time-domain method [17] is found in Ref. 18.

The inside of the cavity in Fig. 22b is filled with absorbing material. The rear radiation is absorbed and does not contribute to the front radiation. Only one-half of the power input to the spiral is used for the radiation. Therefore, the gain does not increase, unlike the gain for the spiral with a conducting plane. However, the antenna characteristics are stable over a wide frequency band, for example, ranging from 1 to 10 GHz.

BIOGRAPHY

Hisamatsu Nakano received his B.E., M.E., and Dr. E. degrees in electrical engineering from Hosei University, Tokyo, Japan, in 1968, 1970, and 1974, respectively. Since 1973 he has been a member of the faculty of Hosei University, where he is now a professor of electronic informatics. His research topics include numerical methods for antennas, electromagnetic wave scattering problems, and light wave problems. He has published more than 170 refereed journal papers and 140 international symposium papers on antenna and relevant problems. He is the author of *Helical and Spiral Antennas* (Research Studies Press, England, Wiley, 1987). He published the chapter "Antenna analysis using integral equations," in *Analysis Methods of Electromagnetic Wave Problems*, vol. 2 (Norwood, MA: Artech House, 1996).

He was a visiting associate professor at Syracuse University, New York, during March–September 1981, a visiting professor at University of Manitoba, Canada, during March–September 1986, and a visiting professor at the University of California, Los Angeles, during September 1986–March 1987.

Dr. Nakano received the Best Paper Award from the IEEE 5th International Conference on antennas and propagation in 1987. In 1994, he received the IEEE AP-S Best Application Paper Award (H. A. Wheeler Award).

BIBLIOGRAPHY

1. K. K. Mei, On the integral equations of thin wire antennas, *IEEE Trans. Antennas Propag.* **13**(3): 374–378 (May 1965).
2. E. Yamashita, ed., *Analysis Methods for Electromagnetic Wave Problems*, Artech House, Boston, 1996, Chap. 3.
3. R. F. Harrington, *Field Computation by Moment Methods*, Macmillan, New York, 1968.

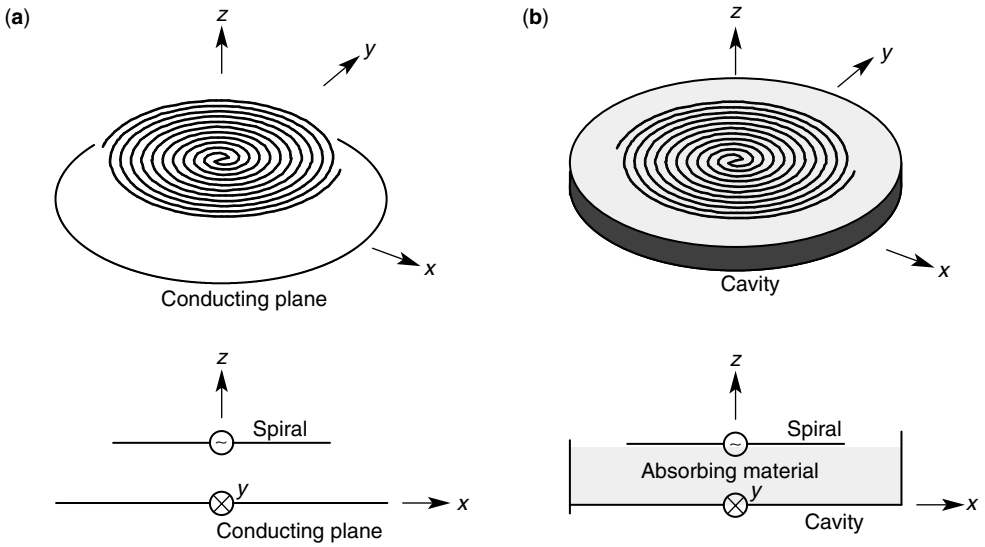


Figure 22. Techniques for unidirectional beam: (**a**) a conducting plane; (**b**) a cavity filled with absorbing material.

4. J. D. Kraus, *Antennas*, 2nd ed., McGraw-Hill, New York, 1988, Chap. 7.
5. H. Nakano and J. Yamauchi, Radiation characteristics of helix antenna with parasitic elements, *Electron. Lett.* **16**(18): 687–688 (Aug. 1980).
6. H. Nakano and J. Yamauchi, The balanced helices radiating in the axial mode, *IEEE AP-S Int. Symp. Digest*, Seattle, WA, 1979, pp. 404–407.
7. R. C. Johnson, *Antenna Engineering Handbook*, 3rd ed., McGraw-Hill, New York, 1993, Chap. 13.
8. H. Mimaki and H. Nakano, A small pitch helical antenna radiating a circularly polarized conical beam, *Proc. 2000 Communications Society Conf. IECE*, Nagoya, Japan, Sept. 2000, p. B-1-82.
9. J. A. Kaiser, The Archimedean two-wire spiral antenna, *IRE Trans. Antennas Propag.* **AP-8**(3): 312–323 (May 1960).
10. H. Nakano and J. Yamauchi, Characteristics of modified spiral and helical antennas, *IEE Proc.* **129**(5)(Pt. H): 232–237 (Oct. 1982).
11. H. Nakano et al., A spiral antenna backed by a conducting plane reflector, *IEEE Trans. Antennas Propag.* **AP-34**(6): 791–796 (June 1986).
12. J. J. H. Wang and V. K. Tripp, Design of multioctave spiral-mode microstrip antennas, *IEEE Trans. Antennas Propag.* **39**(3): 332–335 (March 1991).
13. Y. Mushiake, Self-complementary antennas, *IEEE Antennas Propag. Mag.* **34**(6): 23–29 (Dec. 1992).
14. H. Nakano, J. Yamauchi, and H. Mimaki, Backfire radiation from a monofilar helix with a small ground plane, *IEEE Trans. Antennas Propag.* **36**(10): 1359–1364 (Oct. 1988).
15. H. Nakano, M. Sakai, and J. Yamauchi, A quadrifilar helical antenna printed on a dielectric prism, Proc. *2000 Asia-Pacific Microwave Conf.*, Sydney, Australia, (Dec. 2000), Vol. 1, pp. 1432–1435.
16. M. Ikeda, J. Yamauchi, and H. Nakano, A quadrifilar helical antenna with a conducting wall, *Proc. 2001 Communications Society Conf.* IEICE, Choufu, Japan, (Sept. 2001), p. B-1-70.
17. A. Taflove, *Computational Electrodynamics: The Finite-Difference Time Domain Method*, Artech House, Norwood, MA, 1995.
18. Y. Okabe, J. Yamauchi, and H. Nakano, A strip spiral antenna with absorbing layers inside a cavity, *Proc. 2001 Communications Society Conf. IEICE*, Choufu, Japan, Sept. 2001, p. B-1-107.

HF COMMUNICATIONS

JULIAN J. BUSSGANG
STEEN A. PARL
Signatron Technology Corporation
Concord, Massachusetts

1. INTRODUCTION TO HF COMMUNICATION

High-frequency (HF) *communications* is defined by the International Telecommunication Union (ITU) as radio transmission in the frequency range 3–30 MHz. However, it is common to consider the lower end of the HF band as extending to 2 MHz.

HF radio communications originally became popular because of its long-range capabilities and low cost. For high-rate digital communications, the HF channel is, however, a rather difficult communications medium. It subjects the transmitted signal to large variations in signal level and multiple propagation paths with delay-time differences large enough to cause signal overlap, and hence self-interference. A relatively high rate of fading can result and hinder the ability of receivers to adapt to the channel. High-data-rate communications over HF have also been hampered by having to contend with narrowband channel allocations, a large number of legacy users, and restrictive regulatory requirements.

Communications in the HF band rely on either ground-wave propagation or sky-wave propagation, which

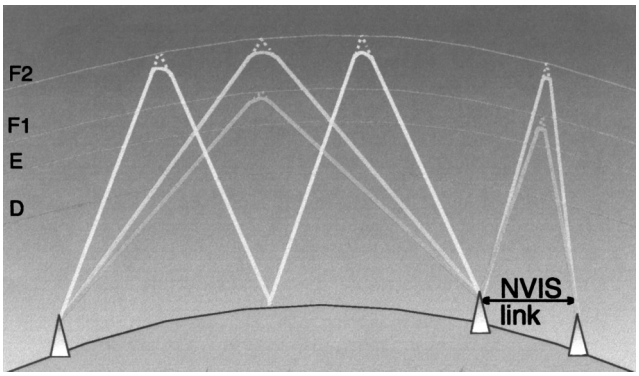


Figure 1. HF sky-wave propagation examples, showing multipath from different layer reflections and multihop propagation.

involves reflections from the ionosphere (Fig. 1). Ground-wave propagation is composed of a direct ray, a ground reflected ray, and a surface wave due to ground currents. The lower the frequency is, the farther ground-wave signals propagate. At the upper end of HF frequencies, the ground-wave attenuates rather rapidly with distance and plays less of a role for long-distance communications. The other mode of HF propagation, *sky-wave*, relies on the ionosphere, which is a series of ionized layers above the earth. Ionization is induced by solar radiation on the day side of the earth and by cosmic rays and meteors on the night side. This causes rays to be reflected, or rather *refracted*, from the ionosphere.

Refraction in the ionosphere is stronger at lower frequencies. As a result, frequencies below HF are refracted from lower layers. At frequencies above HF rays do not refract enough, even from the higher layers. Thus there is a highest frequency supported by HF sky-wave propagation. D-layer absorption determines the lowest frequency in the HF band supporting a sky-wave mode.

Well below HF, in the very low frequency (VLF) band ground-wave and sky-wave propagation combine to form the earth–ionosphere waveguide, which can become nearly lossless, allowing worldwide communications. HF is normally the highest frequency band that can propagate over long distances within this waveguide.

The importance of HF derives from the fact that HF radiowaves are capable of long-distance transmission with relatively little power, which was crucial before satellites and long-range cables became available. The ability to communicate using radio transmission at long distances was discovered by Guglielmo Marconi when he demonstrated the first transatlantic wireless telegraphy in December 1901. He used radio transmission to send Morse code signals from England to Newfoundland. The existence of the ionosphere was not known at that time. A year later Oliver Heaviside and independently Arthur Edwin Kennelly suggested the existence of an ionized layer to explain long-range propagation. It was not until 1924 that Edward V. Appleton confirmed this.

More recently the role of HF has to some extent been taken over by communications satellites, which hover over the earth in geostationary positions and permit over-the-horizon communication independent of day or night. However, satellite communications is more costly than HF, is less convenient for stations on a moving platform, and is not available without prior arrangements and usage fees. Thus, military services, commercial airplanes and ships, and emergency services continue to use HF radio. The lower cost of HF communications and advances in technology are the major driving forces toward continued use and further enhancements.

Today HF is increasingly used for data transmission. With advances in electronic signal processing, the interest in reviving HF is growing. The popularity of the Internet has created a demand for higher data rates. However, significant moves in that direction are stifled by the fact that the existing HF frequency allocations are very narrowband.

2. HF RADIO APPLICATIONS

HF radio was initially used for Morse code and voice transmission. Accordingly, under international radio conventions, much of the spectrum was organized into a series of narrowband channels, with voice channel allotted a nominal 3-kHz band for single sideband (SSB) transmission. Either the upper (USB) or the lower (LSB) sideband could be selected. For voice communications, HF radio has sometimes been combined with vocoders (speech bandwidth compression devices), particularly when double-sideband modulation of speech had to be accommodated within the narrow bandwidth of a typical HF channel.

Nowadays even voice is transmitted digitally, so that the principal interest has shifted to various digital modulation schemes and modems tailored to be effective over HF channels.

As digital transmission was introduced, frequency shift keying (FSK) and phase shift keying (PSK) became increasingly applied as modulation techniques. In addition, frequency hopping has been used for military digital transmission, when it is desirable to avoid detection of the transmissions so as to prevent a jammer from locating and concentrating on a particular transmission frequency.

A typical application of HF radio has been marine radio from coastal stations to ships and back, and between naval vessels. The frequency is selected according to the time of day, season of the year, and distance to be covered. Some of the marine communications have switched to satellite links, but HF is still commonly used because of its lower cost. HF voice links continue to be very much in use for emergency or distress calls.

A new digital service has been introduced by ARINC, augmenting existing HF voice radios to provide data communications between commercial aircraft and ground. This service is intended primarily for commercial aircraft flying over the oceans, out of reach of other radio communications. The service, HF Data Link (HFDL) [1], and 11 HF shore stations around the world under control of two long-distance operational control (LDOC) centers, one in New York and the other in San Francisco. HF links operated by ARINC are used by the Federal Aviation Agency (FAA) for air traffic control (ATC) communications. Above 80° north, HFDL is the only communications

medium available to commercial aviation. Elsewhere, the HFDL service competes with satellite services primarily on cost. HF is still used in Australia for the Royal Flying Doctor Service (RFDS), but its role has shifted to function more as an emergency backup to the telephone network and to satellite systems.

As HF data links become more sophisticated and better able to cope with link variability, they find more applications in digital networking, data transmission, and email. Indeed, email is a growing marine application. The shipboard HF radio stays in contact with one of several shore-based HF stations, which act as mail servers. Several commercial HF email networks have been established as an outgrowth of marine amateur radio operations [2]. The Red Cross has used the CLOVER protocol. Another network protocol is called PACTOR, which is an FSK based scheme developed in Germany in the late 1980s by a group of ham operators. A newer proprietary protocol is PACTOR-II, which uses a two-tone differential phase shift keying (DPSK) [3]. Both use a parallel-tone modulation scheme. The data rates for PACTOR-II range from 100 to 800 baud depending on conditions. The throughput is up to 140 characters per second.

HF radios serve as important links for data communications for all military services. The U.S. Air Force operates Scope Command, a network of HF stations. The U.S. Army uses HF for long-range handheld radio communications. The U.S. Navy uses HF for ship-to-ship and shore-to-ship communications, mostly at short range using ground waves. The U.S. Army and the U.S. Marine Corps often use HF in near-vertical incidence sky-wave (NVIS) mode to communicate short range over mountains and other obstacles (Fig. 1). The National Communications System (NCS) links a large network of HF stations (SHARES) and a large number of allocated frequencies for national emergency communications. The NATO Standardization Agreement (STANAG) 5066 data-link protocol is used in HF email gateways by the NCS and by U.S. and European military services, and has been adopted by NATO for the Broadcast and Ship-Shore (BRASS) system.

Other important HF radio applications include encrypted communications with diplomatic posts in various countries and private networks linking remote outposts around the world, such as International Red Cross field sites and oil exploration stations.

The allocation of HF frequencies is strictly regulated by licensing in individual countries, and is coordinated worldwide by the ITU and in the United States by the Federal Communications Commission (FCC). Some of the frequencies are allocated to citizens band (CB), some to amateur fixed public use, and some to marine and aeronautical communications. Government groups, including the military, have their own allocations. The full FCC table of HF allocations is available on the Internet [4].

Wideband HF has been an active topic of research, but it is still considered impractical by many because of the legacy of narrowband frequency allocations.

3. HF CHANNEL PROPAGATION

HF frequencies propagate by ground-wave signals along the conducting earth or by skywave signals refracted from the layers of the ionosphere, as summarized in Table 1. Many additional details about HF propagation and communication methods may be found in the literature [5–7].

Ground-wave propagation is quite predictable and steady. Signal strength drops off inversely with distance, but depends strongly on the polarization, the ground conductivity, the dielectric constant, and the transmission frequency. Saltwater provides excellent conductivity, and therefore transmission over seawater is subject to the least attenuation. At frequencies below HF, ground-wave communications reach several thousand kilometers, whereas at HF communications can range from tens of kilometers over dry ground to hundreds of kilometers over seawater.

Skywave signals depend primarily on the ionospheric layer from which they reflect. The layers can often support several rays, or paths, between two terminals. The transmissions at lower frequencies tend to reflect from lower layers, while the transmissions at higher frequency penetrate deeper and reflect from higher layers. Transmissions at frequencies above the HF band tend to

Table 1. Earth–Ionosphere Propagation at HF and Lower Frequencies

	Propagation Medium			
Feature	Ground Wave	D Layer	E Layer	F Layer
Height above earth	0	50–90 km	90–140 km	140–250 km (day) 250–400 km (night)
Variability	Varies with surface characteristics and frequency	Daytime only	Reduced or disappeared at night	Split into F1 and F2 during day Highly dependent on solar activity
Approximate communications range	Depends on frequency, ground conductivity, and noise	2100 km (single-hop)	2800 km (single-hop)	3500 km (day) 4500 km (night)

go right through the ionosphere, as they are not reflected back to earth.

We use the term *reflection* to indicate that a ray returns to earth. Actually as a radiowave hits the ionosphere, it is not reflected as if from a smooth surface boundary; rather, it is *refracted*, or somewhat bent, so that it appears as if it had been reflected from a mirrorlike surface at a greater height. This apparent height of reflection is called the *virtual height*.

The lowest ionospheric layer is the D layer, which thins out at night. The E layer above it also thins out at night. During the day it can refract frequencies in the lower HF band. The highest layer, F, is split during the day into the F1 layer and the higher F2 layer. By refracting from a higher layer, the radiowaves can reach out further. The F2 layer is the most important layer for daytime propagation of long-range HF rays; it permits communications in the higher end of the HF band.

The ionization in the F2 layer, so important for long-range HF, is highly variable because of its dependence on the sun. These variations have cycles of 1 day (due to the rotation of the earth), about 27 days (due to the rotation of the sun), seasonal (due to the movement of the earth around the sun), and about 11 years (due to the observed period of sunspot activity).

The reflecting property of the ionosphere is often characterized by the *critical frequency*, which is the highest frequency that can be reflected from the ionosphere at vertical incidence. The critical frequency f_c is determined by the electron density N (the number of electrons per cm^3) and it varies with time and location, but has a typical value around 12 MHz. The critical frequency can be determined by vertical sounding or predicted from

$$f_c = 9 \cdot 10^{-3}\sqrt{N}$$

The *maximum usable frequency* (MUF) is the highest frequency that connects transmitter and receiver on a given link:

$$\text{MUF} = \frac{kf_c}{\sin\varphi} = \frac{kf_c}{\cos(\theta/2)}$$

where φ is the angle of incidence, θ is the angle by which the ray is refracted, and $k \cong 1$ is a correction factor accounting for ray bending in lower layers. This equation can also be used to determine the *critical angle*, the maximum angle of incidence at which reflection can occur at a particular frequency.

When the critical frequency is high enough, it is possible to establish HF sky-wave communications at short ranges by tilting the HF antennas so that enough energy is radiated straight up. This unique mode of HF propagation, called near-vertical incidence sky-wave (NVIS) communication, is used for short-range valley-to-valley-type communication links. The NVIS transmission mode is possible only at the lower HF frequencies.

At the low end of the HF band D-layer absorption increases, making daytime communications difficult. *The lowest usable frequency* (LUF) is the frequency below which the signal becomes too weak. Whereas the MUF is defined entirely by propagation effects, the LUF depends on system parameters (transmitter power, antenna, modulation, and noise) as well.

The E layer sometimes contains patches of denser ionization that generate an additional reflection called "sporadic E." Sporadic E is strong and usually nonfading. It can degrade long-range HF communications by preventing signals from reaching the F2 layer. On the other hand, it can make possible medium-range communications at frequencies well above the HF band.

A ray reflected from the ionosphere can exhibit fading and time-delay dispersion due to absorption and small variations of the propagation medium along the path. The fading of a single ray is generally uniform across a wide bandwidth and is called *flat fading*.

Multiple rays can be created by reflections from different layers (see Fig. 1). The different rays arrive at the receiver with different delays and combine to cause a very-frequency-dependent fading called *frequency-selective fading*.

A sky-wave signal can be rereflected by the ground (especially seawater) to create a ray with multiple ionospheric reflections, usually referred to as *hops*, as also illustrated in Fig. 1. When this happens, delay spreads can be especially large.

The decomposition of a transmitted ray into multiple rays can also be caused by the influence of the earth's magnetic field. This can be explained by noting that a transmitted vertically polarized electromagnetic wave can be considered as a superposition of two waves of opposite circular polarizations. Because of the interaction of the magnetic field with electrons in the ionosphere, each of these component waves is subjected to different refraction levels and therefore follows a slightly different ray path between transmitter and receiver. These two rays, which are termed the *ordinary ray* (O-mode) and the *extraordinary ray* (X-mode), can arrive with different delays, also generating multipath.

Sky-wave signals exhibit a considerable amount of variability. The ionosphere is not static and has variations in the electron density distribution. Solar flares and sunspots cause ionospheric turbulence and atmospheric phenomena such as sudden ionospheric disturbances (SIDs), and polar cap absorption (PCA) can disrupt and disturb transmission.

The initial frequency can be selected from ionospheric predictions based on historical data, vertical sounding, or oblique ionospheric sounding probes. It can be updated by monitoring the transmitted signal directly and switching frequencies as necessary.

External noise is often a limiting factor on HF links. Its three components, *atmospheric noise*, *galactic noise*, and *man-made (human-generated) noise*, are important sources of noise in the HF band. We note that below HF only atmospheric noise and synthetic noise are significant. Atmospheric noise is due mainly to lightning and is therefore highly variable. Atmospheric noise levels tend to be more severe in equatorial regions and are 30–40 dB weaker near the poles of the earth. Atmospheric noise can be dominant at nighttime at frequencies below 16 MHz. Like the other external noise sources, it decreases with frequency. Galactic noise is fairly predictable and can

be dominant in radio-quiet areas at frequencies above 4 MHz. Synthetic noise is highly variable and depends on transmission activity in adjoining frequency bands.

In summary, the properties of HF sky-wave transmission depend on the height and density of the ionospheric layers and on ambient noise sources. Because the ionosphere has several layers and sublayers, several refracted ray paths are possible. Thus multipath and fading effects are common. Typical time spread within each path is 20–40 μs, with individual paths separated by 2–8 ms. Some of the fades last just a fraction of a second and some a few minutes. Fading and the limited bandwidth impose critical constraints on the achievable data rates, which can be quantified using Shannon's channel capacity [8].

4. KEY ELEMENTS OF THE HF DIGITAL TERMINAL

A typical HF digital terminal consists of an antenna, a radio (transmitter, receiver), control equipment (frequency scanning, frequency selection, link establishment, synchronization, selective calling, and link quality control), a modem, and computer or network interfaces as illustrated in Fig. 2. These functions are elements of the ISO layered protocol model illustrated in Fig. 3.[1] The terminals can range in complexity from fairly simple "ham" (amateur) radios to sophisticated automated radio network terminals.

HF voice terminals traditionally operate in a one-way (simplex) mode using manual or automated "push to talk" (PTT). Modern data modems operate in a full-duplex mode, requiring separate frequency allocations for each direction of communications. Typically each terminal is allocated several frequencies on which it can receive, so as to increase the likelihood that one of them will permit a good connection.

More recent advances in electronic hardware, radiofrequency circuits, programmable chips, and software technology have permitted the building of HF radio terminals that are miniaturized and relatively inexpensive. These

[1] The International Standards Organization (ISO) is a worldwide federation of national standards bodies from some 140 countries that publishes international standards.

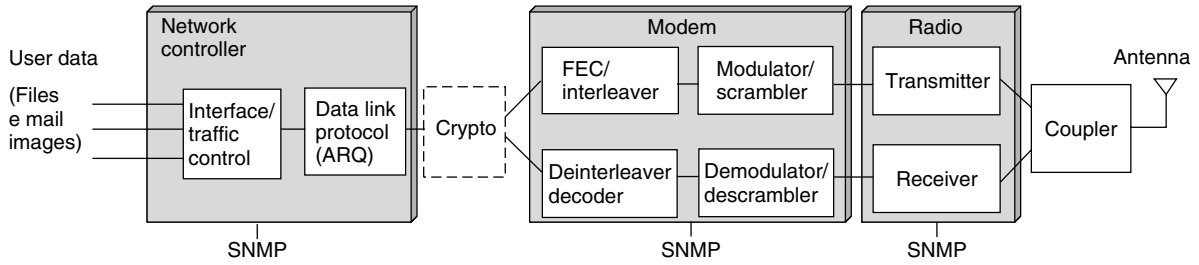


Figure 2. Block diagram of digital HF terminal.

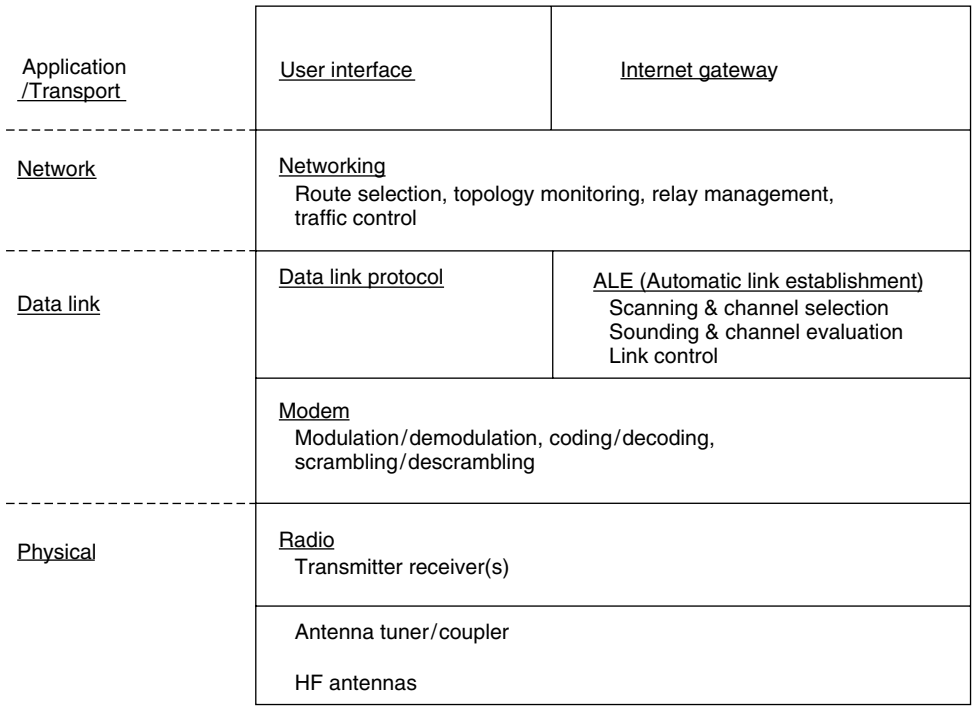


Figure 3. HF terminal functions.

radios can be versatile. They operate in different modes, use a range of data rates, and interface with different networks.

With this background, we describe the components shown in Fig. 2, starting with antennas (the bottom layer in Fig. 3).

4.1. Antennas and Couplers

Antennas are characterized by directivity, gain, bandwidth, radiation efficiency, physical size, and polarization. HF antennas [5] on the ground often require a ground screen to improve ground conductivity. Directivity is measured by assessing the proportion of the total radiated power propagating in a given direction. Gain expresses the combined directivity and radiation efficiency and thus measures the portion of the total transmitted power radiated in a given direction. For ground-wave propagation, the transmitter and receiver must have the same polarization. For sky-wave propagation, the received polarization is generally elliptic and can vary greatly from the transmitted polarization due to different propagation of the ordinary and extraordinary waves. Therefore, some receiver terminals use diversity, whereby the received signals from two orthogonally polarized antennas are combined to get the best signal.

Vertically polarized antennas, such as the vertical whip and high-power towers, tend to be narrowband. These antennas are best suited for ground-wave propagation or for long-range sky-wave propagation. Their antenna patterns are omnidirectional in azimuth, making such antennas ideal for point-to-multipoint communications. A vertical whip can be tilted or bent for intermediate- or short-range NVIS skywave. Horizontal polarization requires that the antenna be raised above the ground. Another common HF antenna type is the half-wave horizontal dipole on a mast, which can be used at low frequencies and medium range. The horizontal Yagi antenna is also used at HF, but becomes large at low frequencies (wavelength at 3 MHz is 50 m or 150 ft) and is very narrowband. An omnidirectional pattern can be achieved with a crossed-dipole antenna. Rhombic antennas can achieve significantly better antenna gain, but when designed for the low end of the HF band (2–10 MHz) can be very large, requiring several acres.

Narrowband antennas require an antenna *tuner*, which is a circuit of capacitors and inductors that allows adjustment of the antenna impedance to maximize radiation efficiency at the desired frequency. An antenna tuner automated under microprocessor control is called a *coupler*.

Wideband antennas that do not require a tuner or coupler are usually a variation of the logperiodic antenna, which can be either vertically or horizontally polarized. The logperiodic antenna is a combination of several narrowband elements, each tuned to a different frequency. Properly designed, they can cover the entire 2–30 MHz HF band.

Small wideband antennas are used primarily for receive-only functions, as they have low radiation efficiency due to the difficulty of matching the impedance over a wide bandwidth. Low-efficiency antennas do not degrade the signal-to-noise ratio when external noise dominates, as is the case in the HF band.

4.2. HF Radio

The radio transmitter generates the radiated power. The required amount of power (hence, the size of the power supply) depends, of course, on the intended transmission distance. Power can be emitted in bursts or continuously, according to the selected type of modulation. Transmitted power can range from tens of watts for handheld units to several kilowatts for point-to-multipoint (broadcast) ground stations. Automatic level control (ALC) or transmit level control (TLC) is used to moderate the increase in transmitted RF (radiofrequency) power when audio input increases.

A single HF radio channel accommodates transmission of voice in the range of 300–3000 Hz. HF radios generally use single-sideband (SSB) modulation, transmitting only one of the two sidebands generated by modulating the voiceband signal onto a carrier frequency. An SSB HF radio is typically designed for voice and supports only low-data-rate transmission.

To introduce higher data throughput, some modern radios operate over two channels (6-kHz band) or four channels (12-kHz band) at once. Two adjoining channels use both sidebands in an independent sideband (ISB) modulation. In a four-channel mode two sidebands on either side of the carrier are used [9].

At the receiving end, it is possible to improve performance by using polarization and space diversity. This works best if the antennas can be spaced several wavelengths apart, which may not always be practical at HF frequencies. However, some diversity improvement can still be achieved with smaller antenna spacing [10,11].

4.3. HF Modem

While the earliest radios used analog amplitude modulation (AM) or frequency modulation (FM), modern digital HF systems achieve significant information throughput by using a modem (modulator/demodulator) at each end of the link. The modem modulates the data onto a transmitted carrier, and inversely demodulates the received carrier waveform to get the transmitted data.

The modem must be designed to effectively cope with the severe multipath and fading found on the HF channel. A well-designed modem can take advantage of the redundancies generated by encoding and by propagation through the different paths spaced in time or space. It can reassemble the signal that may have been spread out, taking advantage of all propagation paths.

The modem is the heart of the digital HF terminal and is discussed in more detail in a later section.

4.4. Link Establishment and Maintenance

The station desiring to transmit has to select from a number of preassigned frequencies. While in the past link establishment was a difficult manual task for the operator, more and more modern HF radios use automatic link establishment (ALE) [12]. At the receiving station the ALE receiver scans the allocated frequencies to detect the

transmission and to make the best frequency selection for establishing the link. The selection may be based on what is best for one particular link or for connecting a transmitting station to a number of users at the same time.

An ALE-equipped radio can passively or actively evaluate the channel for the transmission frequency best suited for the intended receiver. Active evaluation of what is the proper frequency might be the result of a broadband sounder identifying the ionospheric layers. It can involve feedback from an ALE receiver to report when a carrier frequency change is needed. Passive evaluation, such as link prediction based on stored or broadcast ionospheric data, can be reasonably effective at estimating gross ionospheric effects.

The ALE module coordinates communications for automatic HF link establishment (handshake) and status information between the two nodes, controlling both half- and full-duplex modes. Another function of ALE is *selective calling*. This feature allows a radio to mute the received signal until the transmission intended for that particular radio is received. Selective calling requires that the receiving radio have a unique address and that this address be included as part of the transmission.

ALE functions international rely on standards to make radios from different manufacturers as compatible as possible, such as FED-STD-1045 [13] and MIL-STD-188-141B [9]. Combined with availability of adaptive equalization in the modem, ALE enhances the likelihood of link establishment and good communications.

Reliable HF communication also requires monitoring of the connectivity between the transmitter and the receiver to enable automatic switching to another frequency when the current frequency has faded out or has become less effective for the desired range of transmission.

4.5. Network Interface and Traffic Management

A network controller couples the modem to the network. It controls network traffic using the link and manages changing conditions over the HF link. Interfaces may be able to adapt the modem to more than one local-area network (LAN) standard (e.g., ATM or Ethernet). The interface performs appropriate handshake operations with the network equipment and may also implement a traffic management function, such as congestion control.

As an example, consider a marine email application. The modem at the user's end of the link is connected through a network interface to the computer running the local email software, while the modem at the shore station is connected to the Internet.

4.6. Data-Link Protocol

Digital transmission over the HF link must be controlled by a well-defined protocol. For instance, military radios use the data-link protocol in MIL-STD-188-110B, Appendix E [14]. The link data frame usually includes cyclic redundancy code (CRC) check bits, which allow detection of errors in a frame. When errors are detected, the link protocol may use a feedback channel to initiate retransmissions. Since such a feature is usually automatic, it is known as automatic repeat request (ARQ).

ARQ can be considered an alternative, or supplement, to interleaving and coding, which are discussed below. Essentially error-free communications can be achieved this way at the cost of significant delay variations. Combining interleaving and coding with an ARQ method can get the benefits of both approaches. ARQ is commonly used by higher layer protocols, such as TCP/IP, to provide end-to-end reliability. It can also be very useful at the link level to minimize end-to-end retransmissions.

The most common ARQ scheme is "go-Back-N" [15], where retransmissions start over with the frame that contains errors. The most advanced ARQ method is selective ARQ, which allows the terminals to select individual data frames for retransmission. One advantage of selective ARQ over coding and interleaving is that the delay is determined by the actual channel characteristics, while an interleaver generally has to be designed for the worst-case scenario (long fades).

5. HF MODEMS

HF modems perform several functions other than modulation and demodulation. These include interleaving and deinterleaving, coding and decoding, and security. Modems incorporate circuits for initial acquisition, adaptive equalizers for automatic compensation of multipath dispersion, and circuits for performance monitoring and for controlling the switching of carrier frequencies. This section describes some applicable standards and key functions of the modem.

5.1. Standards

The proliferation of different modulations, link protocols, and data rates has led to the establishment of a series of standards promoting interoperability.

The ITU, the Department of Defense (DoD), NATO, Telecommunications Industries Association (TIA), and others publish carefully crafted standards. The DoD, in particular, has issued a comprehensive series of mandatory and recommended standards for military communications, many of which have been adopted for commercial use. Military HF modems follow the MIL-STD-188-110B [14], which specifies a number of modulation and coding waveforms for various data rates and for both single- and multichannel modems. The HF terminals follow the MIL-STD-188-141B [9], which covers transmitters, receivers, link establishment procedures, data-link protocols, security, antijam technology, and network maintenance (using SNMP). This standard is coordinated with FED-STD-1045 [13]. Military standards are available on the Internet [16]. International standards, and many others, are available from the American National Standards Institute (ANSI) [17].

NATO has established its own Standardization Agreements (STANAGs). The U.S., NATO, and ITU standards are highly coordinated. For instance, MIL-188-STD-110B specifies STANAG 5066 as an optional data-link protocol.

Table 2 lists representative standardized HF modulations, several of which are discussed further below.

Table 2. Characteristics of Selected HF Modem Standards

	16-Tone DPSK Modem, MIL-STD-188-110B, Appendix A	39-Tone DPSK Modem, MIL-STD-188-110B, Appendix B	ALE Single Modem, MIL-STD-188-141B, Appendix A	PSK Single-Tone Modem, MIL-STD-188-110B, Fixed Frequency	PSK Single-Tone Modem, MIL-STD-188-110B, Appendix C
Data rates (bps)	75–2400	75–2400	61.2	75–4800	3200–12,800
Symbol rate/period	75 baud/13.3 ms	44.44 baud/22.5 ms	125 baud/8 ms	2400 baud/416.6 μs	2400 baud/416.6 μs
Known (training) transmissions	Tone for Doppler correction	Tone for Doppler correction; 1 framing bit per 31.5 data bits	1 stuff bit per 49 bits transmitted	2400 or 4800 bps: 16 training symbols following 32 data symbols	31 training symbols every 256 data symbols
				150–1200 bps: 20 training symbols following 20 data symbols	72 preamble symbols reinserted after every 72nd training sequence
				75 bps no training symbols	
Modulation	DQPSK (1200, 2400 bps) DBPSK (75–600 bps) Parallel tones, spaced 110 Hz	DQPSK Parallel tones, spaced 56.25 Hz	8-FSK	8-PSK (2400, 4800 bps) QPSK (1200 bps) BPSK (150–600 bps) QPSK spread by 32 (75 bps)	QPSK (3200 bps) 8-PSK (4800 bps) 16-QAM (6400 bps) 32-QAM (8000 bps) 64-QAM (9600, 12,800 bps)
Coding and interleaving	Bits repeated at rates below 2400 bps	Shortened Reed-Solomon (15,11):(14,10) at 2400 bps, (7,3) otherwise; repeat coded bits at rates below 2400	Rate-$\frac{1}{2}$ Golay code, triple redundancy	4800 bps: no coding 75, 600–2400 bps: rate $\frac{1}{2}$ 150–300 bps: rate $\frac{1}{2}$ plus repetitions	Rate $\frac{3}{4}$ convolutional code (with tail-biting to match interleaver length)
Interleaving	None	Selectable	48-bit interleaving	0, 0.6, or 4.8 s	0.12-8.64 s (6 steps)
Preamble	2 tones for 66.6 ms followed by all tones for 13.3 ms	4 tones over 14 or 27 periods followed by 3 tones over 8 or 27 periods followed by all tones over 1 or 12 periods (second number for optional extended preamble)	No fixed preamble; transmission is an asynchronous sequence of 24-bit words; each word has three 7-bit characters and a 3-bit preamble identifying word type	0.6 or 4.8 s of a fixed BPSK sequence	0-7 blocks of 184 8PSK symbols followed by 184 sync symbols followed by 103 symbols with information on data rate and interleaver settings

5.2. Signal Acquisition

HF modems must be able to achieve signal acquisition in the presence of several channel disturbances: severe fading, multipath, non-Gaussian noise, and interference from other transmitters. Transmissions normally include a known preamble designed to permit the modem to adjust as necessary to achieve signal acquisition.

After the initial acquisition, transmissions may incorporate a known signal to permit continuous monitoring of channel conditions. Such a signal may be in the form of a pilot tone in a parallel-tone system or a periodically inserted training sequence in a single-tone system. Equalization without a training sequence is known as "blind" equalization [18] and has been studied for HF [19].

5.3. Types of Modulation

The easiest way to communicate over a channel with multipath is to modulate tones with symbols of duration much longer than the multipath spread. By ignoring the first part of each received tone, intersymbol interference can be avoided. This principle is used in older frequency shift keying (FSK) modems, which transmit one or more of several possible tones at a time, and in the parallel-tone phase shift keying (PSK) modems commonly employed by military and commercial users. Table 2 shows an example of two parallel-tone PSK modems (16-tone modem in column 1 and 39-tone modem in column 2) and a robust FSK modem used for ALE (column 3). Parallel tone modems are also used with marine single-sideband (SSB) radios for commercial HF email and HF Web applications [3].

Modems with a narrow HF bandwidth allocation can benefit from using multiphase PSK (e.g., 4, 8, or 16 phases), to achieve higher data rates. Differential PSK (DPSK), which compares the phase received in each time slot to the phase received in the preceding slot, offers a simple implementation that does not require accurate phase tracking. However, since the phase in the previous slot is also noisy, DPSK incurs a performance loss relative to coherent PSK.

Parallel-tone modems with narrow frequency bands for each tone and low keying rates have several advantages: (1) they are easy to implement, (2) they can be quite bandwidth-efficient, and (3) the multipath spread is usually a fraction of the tone duration. The main disadvantage is the high peak-to-average power ratio resulting from the superposition of several parallel tones. This leads to the need for a linear power amplifier with a dynamic range much larger than the average transmitted power. Another disadvantage is the fact that Doppler spread can cause interference between the parallel tones. Coding is used to provide an in-band diversity gain, compensating for the selective fading caused by multipath.

In serial transmission, with single-tone modems, the modulation can have virtually constant amplitude, so the peak-to-average power ratio is close to unity. A higher average power is then achieved with a given peak power. This means that a more efficient class C amplifier can be used, and the transmitter can be made more compact. A single-tone modem requires faster keying to match the data rate. This means that intersymbol interference cannot be ignored, except perhaps at the lowest data rates. Single-tone modems use adaptive equalization to undo the intersymbol interference. By using the entire received signal, a serial modem can get better performance than a modem that ignores the part of the signal containing intersymbol interference.

Thanks to more recent advances in signal processing and linear amplification technologies, single-tone modems can use amplitude modulation to achieve even higher data rates. For example, the MIL-STD-188-110B includes 16-, 32-, and 64-quadrature amplitude modulation (QAM) modems (see the last column in Table 2). These new modulations can have rates up to 12,800 bps within a 3-kHz band. Although QAM spaces signal points as far apart as possible, it requires significantly higher signal-to-noise ratio (SNR), due the smaller spacing between signal constellation points. In general, one result of this "compressed" signal constellation is the *constellation loss*, meaning that the required power grows faster than the increase in data rate. Higher-data-rate modes can be used in short-range ground-wave applications where the signal is sufficiently strong. With sufficient transmitter power, they may also be appropriate for short- or medium-range sky-wave communications.

5.4. Coding

Coding can be used to correct errors by introducing redundancy in the transmitted data [20]. It can be used with both parallel- and serial-tone modulations. The selection of a proper code is one of the key design issues. Forward error correction (FEC) block codes are simple to use because they are inherently compatible with the data frame structure. However, short-constraint-length convolutional codes are often applied and can offer performance advantages, particularly as they can easily be decoded using the actual received samples (soft decision), as opposed to only using demodulated symbols (hard decision). Table 2 lists several types of codes used in standard HF systems. All these coding techniques add redundancy, thus increasing the required channel data rate. With conventional modulations it is then necessary to expand the bandwidth.

Trellis modulation coding [21] is a way of combining coding and modulation that can offer coding gain without expanding bandwidth. It has seen few applications in HF in spite of the band-limited nature of HF channels. A newer form of coding, Turbo coding [22], promises performance even closer to Shannon's capacity limit by interleaving the coded data over relatively long time periods.

5.5. Interleaving

Fades cause errors to appear in clusters. The objective of interleaving is to randomize error occurrences by forming the encoded block, not out of consecutive bits, but out of bits that are selected in a pattern spread out over several blocks. When the inverse process is implemented at the receiver, the received clusters of error are converted into randomly distributed errors. Coding then permits the random errors to be corrected at the receiver.

To be effective, the interleaving period needs to be longer than the duration of a fade, which can be several seconds on the HF channel. The resulting delay can be undesirable, and many modems using interleaving include means for selecting the interleaving period.

5.6. Link Protection

A number of methods may be used to protect the data transmitted over the link. These include data scrambling and encryption. Similarly, ALE transmissions may be protected, for instance, as described in Appendix B of the military standard for automatic link establishment [9].

5.7. Adaptive Equalization

Adaptive equalization is a modem feature that reassembles a signal that has been dispersed in time by the channel propagation effects. It was first introduced in telephony [23] and later applied to fading channels [24]. Adaptive equalization is an effective technique for HF links to overcome multipath.

When the delay spread is small compared to the symbol duration, a simpler technique called adaptive matched filtering, or a RAKE receiver, may be used [20]. A RAKE receiver estimates individual channel gains and uses these estimates to adjust the gains of a tapped delay line to best match the channel. This approach is most often used with spread-spectrum systems, where bandwidth is large enough to resolve the individual paths, and symbols are long enough to permit neglecting intersymbol interference.

When intersymbol interference cannot be neglected, adaptive equalization is needed. Many forms of adaptive equalization may be used to undo the channel multipath [20]. The main concept is to combine different delayed versions of the received signal in order to maximize the signal-to-interference ratio in each symbol being demodulated. The interference meant here includes that from adjoining symbols. The equalizer applies proper amplitude and phase to each delayed tap and combines the resulting signals in such a way as to best reconstruct the transmitted symbol.

Two common adaptive equalizer techniques, the decision-feedback equalizer (DFE) and the maximum-likelihood sequence estimation (MLSE) equalizer, have not been widely used with HF transmission. The MLSE equalizer is an optimal receiver when the channel is known, but it has rarely been used because of its complexity and sensitivity to channel perturbations. It has been used mostly at HF in experimental modems but could become more practical as the processing power of new digital signal processing (DSP) chips increases. Because of its simplicity, the DFE is popular for modems with data rates that are high relative to the fade rate, but it has not been used much at HF because of the low data rates involved. As the trend to increase HF modem data rates continues, however, the DFE may become more common.

Since the channel is continuously changing as a result of fading, a training sequence is periodically inserted to effectively measure the individual path gains and phase shifts. The MIL-STD-188-110B serial-mode waveform uses 33–50% of the transmissions for training sequences (see Table 2, column 4). The extra reference permits using other equalization methods offering performance slightly better than a DFE, such as the data decision estimation method [25].

6. LINK PREDICTION AND SIMULATION

Performance prediction, based on HF channel models, is used to plan and operate HF communications links. Channel simulators are used to test modem performance using synthesized model channels.

Multipath characterization of the HF communications channel is traditionally based on Watterson's narrowband HF channel model [26]. The Watterson model represents the channel as an ideal tapped-delay line with tap spacings selected to match the relative propagation delays of the resolvable multipath components. Each tap has a complex gain multiplier, and summation elements combine the delayed and weighted signal components (see Fig. 4). The transmitted signal feeds the tapped-delay-line channel model. The tap gains of each path are modeled as mutually independent complex-valued Gaussian random processes, producing Rayleigh fading. Each tap gain fades in accordance with a Gaussian spectrum.

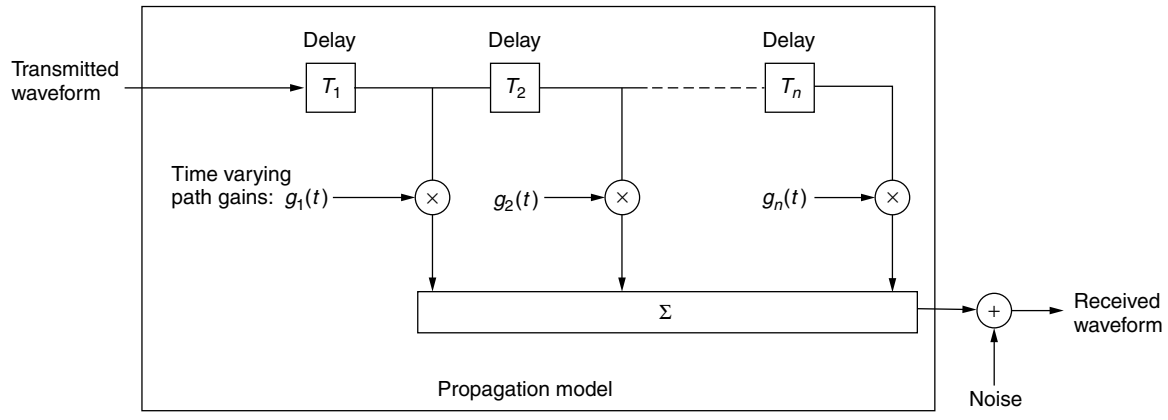


Figure 4. HF channel modeled as a topped delay line.

The Watterson model emulates the propagation channel with the addition of random noise. Atmospheric noise, galactic noise, and synthetic noise are important sources of noise in the HF band, and a proper noise model should combine all three.

Galactic noise is generally Gaussian, whereas atmospheric noise is due to lightning and therefore impulsive in nature. Atmospheric noise values have been measured on a worldwide basis, and typical ranges can be predicted based on geographic location, season, and time of day using charts calculated by the International Telecommunications Union (ITU) [27].

A number of models of synthetic noise have been developed; however, because they are often based on data that are decades old, they should be used with caution. The ITU man-made noise model recommends a distinction between several classes of areas [27]. The most important of these are in business, residential, and rural areas, where the median noise falls off by 27.7 dB per decade in frequency. Synthetic noise can dominate for transmissions at frequencies as low as the low-frequency (LF) band and as high as 4 MHz in the HF band.

The high variability of HF propagation means that the best frequency to transmit depends strongly on ionospheric parameters. Path loss prediction is an important component of link planning. Aided by tabulated data or ionosonde data, it can help select the best frequency. Several propagation models are available [28], most of them derived from the Ionospheric Communications Analysis and Prediction (IONCAP) model developed by the National Telecommunications and Information Administration — Institute for Telecommunications Sciences (NTIA/ITS). One such model is the ICEPAC model by the NTIA/ITS, which adds a newer ionospheric electron density model that provides better worldwide coverage. Another model is the Voice of America Communications Analysis and Prediction (VOACAP) program [28], which extends the IONCAP program with several features and adds a graphical user interface. The ITU skywave model [29] is available as both a table-based model and a computer program, as is the ITU ground-wave model [30]. NTIA/ITS offers a computer program (HFWIN32) combining all three models [31].

For digital communications it is important to model the delay spreads and Doppler spreads, especially when different modems are to be tested and compared. One such prediction model [32] generates the parameters of the Watterson model directly for a hardware simulator.

Hardware channel simulators are essential laboratory instruments for testing the performance of actual modems and network equipment under controlled repeatable HF channel conditions [33]. Such simulators are most valuable in that they model not just the channel but also additive Gaussian noise, impulse noise, and typical types of interference. Hardware channel simulators are available with interfaces at baseband, intermediate frequency (IF), a fixed radiofrequency (RF), and at a frequency-hopping RF. Comparisons of link performance and of simulator test data have validated that simulators are particularly useful for quantitative and comparative performance evaluation.

7. FUTURE TRENDS

Future HF radios and modems are likely to develop in several directions: higher data rates, different modulations, software-controlled multifunction radios, more intelligent signal processing, and improved networking and traffic management.

Apart from the propagation effects on the HF channel, the greatest impediment to achieving higher data rates is the current narrowband channel allocations, which allow only about 3 kHz of spectrum per channel. The simplest design to increase the data rate is to modulate separately each of several parallel channels. Military standards have been developed for combining two or four channels (see [14], Appendix F). However, if the channels could be lumped together and guard bands eliminated, then data rates could be increased and a wider choice of modulations would be available.

Radio users are assigned a group of frequencies to select from, depending on channel conditions. Combining the current 3-kHz channels into wider bands would mean a paradigm shift in HF frequency allocations and radio design. A wideband HF radio would have to use fewer frequencies with wider allocations, but would achieve higher data rates without necessarily increasing the total allocated bandwidth. Experimental HF modems have been studied by DARPA, aiming at data rates as high as 64 or 96 kbps.

Another approach to achieving higher data rates is to combine noncontiguous narrowband frequency channels, so as to be compatible with the current frequency allocations. However, in that case the use of the spectrum would be less efficient, because of the multiple guard bands required on the edges of each transmission band. Amplifier linearity would be a critical problem that might have to be overcome by introducing multiple amplifiers to service each individual frequency band. Another problem would be antennas, which would need to be sufficiently wideband. Having separate antennas for individual bands is probably impractical.

Wideband HF has also been proposed for secure military applications [34]. In that situation the idea is not necessarily to increase the data rate, but to use code-division multiple access (CDMA) with a spread-spectrum technique to reduce power density across the band. The signal is spread so much that interference to narrowband users in the same band is negligible and the transmissions are not readily detected.

Bandwidth-on-demand protocols are being used increasingly in wireless and satellite communications systems. Demand assignment multiple access (DAMA) is standardized for satellite systems [35]. It is possible that the principles of DAMA will be applied to improve the bandwidth utilization of HF in the future.

The so-called third-generation (3G) radios will exhibit enhanced automation and capability to respond to circumstances. Modems of the future will be increasingly software-controlled, and consequently may be reprogrammable depending on the link situation and the network into which they are being interfaced.

BIOGRAPHIES

Steen A. Parl received the degree of Civ. Ing. from the Technical University of Denmark in 1970 and the doctor of philosophy degree from the Massachusetts Institute of Technology (MIT) in 1974. He joined Signatron, Inc., Concord, Massachusetts, in 1972, conducting research in detection and estimation, microwave propagation prediction, underwater acoustic communications, adaptive modems, wideband interference cancellation, adaptive arrays for radio communications, coding, and statistical methods of channel simulation. Since 1993, he has been president and chief scientist of Signatron Technology Corporation, where he has worked on Over-the-Horizon (OTH) systems for applications such as providing Internet access for schools in rural areas, and on nonGPS position-location systems designed to find or track tagged objects. Dr. Parl holds seven patents in the areas of radio communications, interference suppression, channel simulation, and position location. He is a fellow of the IEEE. His current technical interests include wireless telecommunications, efficient data link protocols, digital HF and troposcatter communication systems, antenna nulling techniques, and radiolocation techniques.

Julian J. Bussgang received his B.Sc. (Engineering) degree in 1949 from University of London, England, M.S.E.E. degree from M.I.T. in 1951, and Ph.D. degree in applied physics from Harvard University, Cambridge, Massachusetts, in 1955. From 1951 to 1955 he was a staff member and then a consultant at the M.I.T. Lincoln Laboratory. He joined RCA in 1955 and became manager in Radar Development, and later manager in Applied Research at the Aerospace Division in Burlington, Massachusetts. In 1962 he founded Signatron, Inc. and served as President till 1987. He now consults to Signatron Inc., Concord, Massachusetts.

He is a patentee in the field and has published and presented over forty technical papers, some reprinted in the collections of significant publications in the field. He was visiting lecturer at Harvard University in the Division of Engineering and Applied Science one year and taught a graduate communications course at Northeastern University, Boston, Massachusetts.

Dr. Bussgang is fellow of the IEEE, served on the board of the Information Theory Group, and as chair of the Boston section of the IEEE.

His areas of interest have been correlation functions, sequential detection, radar system design, radio communications over multipath channels, and coding/decoding algorithms.

BIBLIOGRAPHY

1. *HF Data Link Protocols*, ARINC Specification 635, 2000.
2. Technical descriptions CLOVER, CLOVER-2000, G-TOR, PACTOR, PACTOR II, & PSK 31, American Radio Relay League (ARRL), #6982, 2000.
3. J. Corenman, PACTOR primer (Jan. 16, 1998) (online): *http://www.airmail2000.com/pprimer.htm* (Oct. 2001).
4. FCC Office of Engineering and Technology (Sep. 28, 2001), FCC Radio Spectrum Homepage (online): *http://www.fcc.gov/oet/spectrum* (Oct. 2001).
5. J. M. Goodman, *HF Communications*, Van Nostrand-Reinhold, New York, 1992.
6. K. Davies, *Ionospheric Radio*, Peregrinus, London (IEE), 1990.
7. L. Wiesner, *Telegraph and Data Transmission over Shortwave Radio Links: Fundamental Principles and Networks*, Wiley, New York, 1984.
8. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: Information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* **IT-44**(6): 2619–2692 (1998).
9. *Interoperability and Performance Standards for Medium and High Frequency Radio Systems*, U.S. Department of Defense MIL-STD-188-141B, March 1, 1999.
10. W. C.-Y. Lee and Y. S. Yeh, Polarization diversity system for mobile radio, *IEEE Trans. Commun.* **COM-20**: 912–913 (Oct. 1972).
11. W. C.-Y. Lee, *Mobile Communication, Design Fundamentals*, Wiley, New York, 1993.
12. E. E. Johnson et al., *Advanced High-Frequency Radio Communications*, Artech House, Boston, 1997.
13. *High Frequency (HF) Radio Automatic Link Establishment*, National Communications System, Office of Technology and Standards, FED-STD-1045A, 1993.
14. *Interoperability and Performance Standards for Data Modems*, U.S. Dept. Defense MIL-STD-188-110B, April 27, 2000.
15. W. Stallings, *Data and Computer Communications*, 6th ed., Prentice-Hall, Englewood Cliffs, NJ, 1999.
16. Defense Technical Information Center (Oct. 9, 2001), *Scientific and Technical Information Network* (online): *http://stinet.dtic.mil/* (Oct. 2001).
17. American National Standards Institute (no date), *Electronic Standards Store* (online): *http://webstore.ansi.org* (Oct. 2001).
18. S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
19. J. Q. Bao and L. Tong, Protocol-aided channel equalization for HF ATM networks, *IEEE J. Select. Areas Commun.* **18**(3): 418–435 (2000).
20. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill Higher Education, New York, 2000.
21. C. Schlegel and L. Perez, *Trellis Coding*, IEEE Press, Piscataway, NJ, 1997.
22. B. Vucetic and J. Yuan, *Turbo Codes: Principles and Applications*, Kluwer, Boston, 2000.
23. R. W. Lucky, Automatic equalization for digital communications, *Bell Syst. Tech. J.* (April 1965).
24. U.S. Patent 3,879,664 (April, 1975), P. Monsen (Signatron, Inc.), High speed digital communications receiver.
25. F. M. Hsu, Data directed estimation techniques for single tone HF modems, *Proc. IEEE MILCOM85*, 1985, pp. 12.4.1–12.4.10.

26. C. C. Watterson, J. R. Juroshek, and W. D. Bensema, Experimental confirmation of an HF channel model, *IEEE Trans. Commun. Technol.* **COM-18**(6): 792–803 (1970).

27. *Radio Noise*, International Telecommunications Union Recommendation ITU-R P.372-7, 2001.

28. J. Coleman (no date), *Propagation Theory and Software, Part II* (online): *http://www.n2hos.com/digital/prop2.html* (Oct. 2001).

29. *HF Propagation Prediction Method*, International Telecommunications Union Recommendation ITU-R, 2001, pp. 533–537.

30. *Ground-Wave Propagation Curves for Frequencies between 10 kHz and 30 MHz*, International Telecommunications Union Recommendation ITU-R, 2000, pp. 368–377.

31. G. R. Hand (no date), *NTIA/ITS High Frequency Propagation Models* (online): *http://elbert.its.bldrdoc.gov/hf.html* (Oct. 2001).

32. A. Malaga, A characterization and prediction of wideband HF skywave propagation, *Proc IEEE MILCOM85*: 1985, pp. 281–288.

33. L. Ehrman, L. Bates, J. Eschle, and J. Kates, Simulation of the HF channel, *IEEE Trans. Commun.* **COM-30**(8): 1809–1817 (1982).

34. B. D. Perry, A new wideband HF technique for MHz-bandwidth spread spectrum radio communications, *IEEE Commun. Mag.* **21**(6): 28–36 (1983).

35. *DAMA Demand Assignment Multiple Access*, National Communications System, Office of Technology and Standards, FED-STD-1037C, 2000.

HIDDEN MARKOV MODELS

BIING-HWANG JUANG
Bell Laboratories
Lucent Technologies
Holmdel, New Jersey

1. INTRODUCTION

Many real-world processes or systems change their characteristics over time. For example, the traffic condition at the Lincoln Tunnel connecting New York City and New Jersey displays drastically different volume and congestion situations several times a day — morning rush hours, midday flow, evening rush hours, night shifts, and perhaps occasional congestions due to construction. Telephony traffic bears a similar resemblance. As another example, the speech sound carries the so-called linguistic codes for a language in an acoustic wave with varying characteristics in terms of its energy distribution across time and frequency. In order to properly characterize these processes or systems, one has to employ models of measurement beyond the simple long-term average. The average number of daily phone calls going in and out of a city does not allow efficient, dynamic resource management to meet the telephony–traffic needs during peak hours. And the average power spectrum of a spoken utterance does not convey the linguistic content of the speech. The need to have a model that permits characterization of the average behavior as well as the nature of behavioral changes of the system gives rise to the mathematical formalism called the *hidden Markov model* (HMM).

A hidden Markov model is a doubly stochastic process with an underlying stochastic process that is not readily observable but can only be observed through another set of stochastic processes that produce the sequence of observations [1–3]. By combining two levels of stochastic processes in a hierarchy, one is able to address the short(er) term or instantaneous randomness in the observed event via one set of probabilistic measures while coping with the long(er) term, characteristic variation with another stochastic process, namely, the Markov chain. This formalism underwent rigorous theoretical developments in the 1960s and 1970s and reached some important milestones in the 1980s [1–7]. Today, it has found widespread applications in stock market prediction [6], ecological modeling [7], cryptoanalysis [8], computer vision [9], and most notably, automatic speech recognition [10]. Most, if not all, of the modern speech recognition systems are based on the HMM methodology.

2. DEFINITION OF HIDDEN MARKOV MODEL

Let $\boldsymbol{X}$ be a sequence of observations, $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T})$, where $\mathbf{x}_t$ denotes an observation or measurement, possibly vector-valued. Further consider a first-order N-state Markov chain governed by a *state transition probability* matrix $A = [a_{ij}]$, where a_{ij} is the probability of the Markov system making a transition from state i to state j; that is

$$a_{ij} = P(q_t = j | q_{t-1} = i) \qquad (1 \le i, j \le N) \tag{1}$$

where q_t denotes the system state at time t. Note that

$$a_{ij} \ge 0 \qquad \forall \quad i \text{ and } j \tag{2a}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \forall \quad i \tag{2b}$$

Assume that at $t = 0$ the state of the system q_0 is specified by the *initial-state probability* vector $\pi = [\pi_i]_{i=1}^{N}$, where

$$\begin{aligned} \pi_i &= P(q_0 = i) \\ \sum_{i=1}^{N} \pi_i &= 1 \end{aligned} \tag{3}$$

Then, for any state sequence $\boldsymbol{q} = (q_0, q_1, \ldots, q_T)$, the probability of $\boldsymbol{q}$ being generated by the Markov chain is

$$P(\boldsymbol{q} | A, \pi) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T} \tag{4}$$

Suppose that the system, when in state q_t, puts out an observation $\boldsymbol{x}_t$ according to a probability density function $b_{q_t}(\boldsymbol{x}_t) = P(\boldsymbol{x}_t | q_t), q_t \in \{1, 2, \ldots, N\}$. The hidden Markov

model thus defines a density function for the observation sequence $\boldsymbol{X}$ as follows:

$$\begin{aligned} P(\boldsymbol{X}|\pi, A, \{b_j\}_{j=1}^N) &= P(\boldsymbol{X}|\Lambda) \\ &= \sum_{\boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{q}|\Lambda) \\ &= \sum_{\boldsymbol{q}} P(\boldsymbol{X}|\boldsymbol{q}, \Lambda) P(\boldsymbol{q}|\Lambda) \\ &= \sum_{\boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{x}_t) \end{aligned} \tag{5}$$

where $\Lambda = (\pi, A, \{b_j\}_{j=1}^N)$ is the parameter set for the model.

In this model, $\{b_{q_t}\}$ defines the distribution for short-time observations $\boldsymbol{x}_t$ and A characterizes the behavior and interrelationship between different states of the system. In other words, the structure of a hidden Markov model provides a reasonable means for defining the distribution of a signal in which characteristic changes take place from one state to another in a stochastic manner. Normally N, the total number of states in the system, is much smaller than T, the time duration of the observation sequence. The state sequence $\boldsymbol{q}$ displays a certain degree of stability among adjacent q_ts if the rate of change of state is slow compared to the rate of change in observations. The use of HMM has been shown to be practically effective for many real-world processes such as a speech signal.

2.1. An Example of HMM: Drawing Colored Balls in Urns

To fix the idea of an HMM model, let us try to analyze an observation sequence consisting of a series of colors, say, {G(reen), G, B(lue), R(ed), R, G, Y(ellow), B, ...} (see [3]). The scenario may be as depicted in Fig. 1, in which N urns each with a large quantity of colored balls are shown. We assume there are M distinct colors of the balls. We choose an urn, according to some random procedure, and then pick out a ball from the chosen urn again randomly. The color of the ball is recorded as the observation. The ball is replaced back in the urn from which it was selected and the procedure repeats itself — a new urn is chosen followed by a random selection of a colored ball. The entire process generates a sequence of colors (i.e., observations) that can be a candidate for hidden Markov modeling.

It should be obvious that the number of states, N, defines the "resolution" or "complexity" of the model, which is intended to explain the generation of the color sequence as accurately as possible. One could attempt to solve the modeling problem using a single-state machine, namely, $N = 1$, with no possibility of addressing the potential distinction among various urns in their composition of colored balls. Alternatively, one can construct a model with a large N, such that detailed distinctions in the collection of colored balls among urns can be analyzed and the sequence of urns that led to the color observations can be hypothesized. This is the essence of the hidden Markov model.

2.2. Elements of HMM

A hidden Markov model is parametrically defined by the triplet $\{\pi, A, B = \{b_j\}_{j=1}^N\}$. The significance of each of these elements is as follows:

1. N, the number of states. The number of states defines the resolution and complexity of the model. Although the states are hidden, for many applications there is often some physical significance attached to the states or to sets of states of the model. In the urn-and-ball model, the states correspond to the urns. If in the application it is important to recover at any time the state the system is in, some prior knowledge on the meaning or the utility of the states has to be assumed. In other applications, this prior assumption may not be necessary or desirable. In any event, the choice of N implies our assumed knowledge of the source in terms of the number of distinct states in the observation.
2. $A = [a_{ij}]$, the state-transition probability matrix. The states of the Markov model are generally interconnected according to a certain *topology*. An ergodic model [3] is one in which each state can be reached from any other state (in a nonperiodic fashion). If the process is progressive in nature, say, from state $1, 2, \ldots$ to state N as time goes on and observations are made, then a so-called left-to-right topology may be appropriate. Figure 2 shows examples of an ergodic model and two left-to-right models. The topological interconnections of the

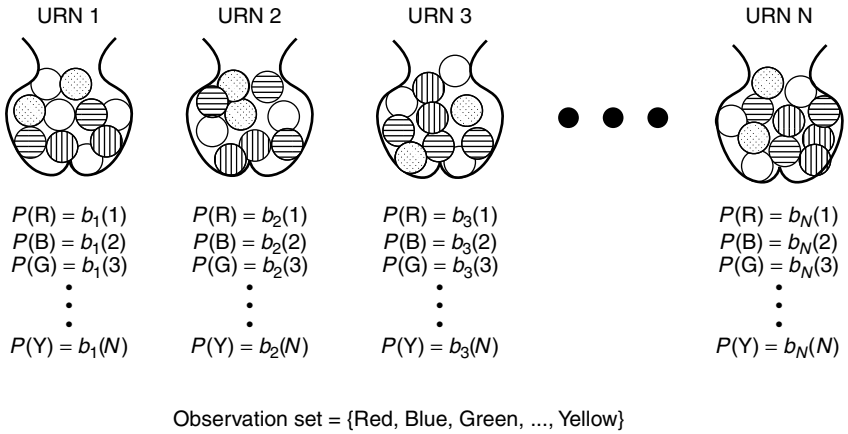


Figure 1. An N-state urn-and-ball model illustrating a case of discrete HMM.

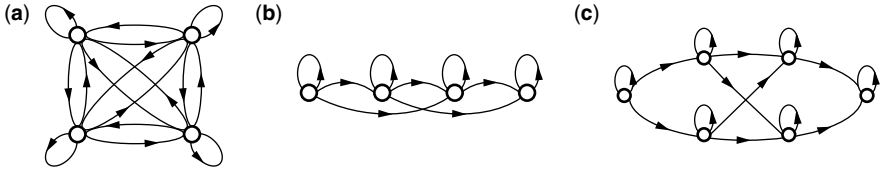


Figure 2. Three types of HMM differentiated according to the topology of the Markov chain: (**a**) a 4-state ergodic model; (**b**) a 4-state left–right model; (**c**) a 6-state parallel left–right model.

Markov chain can be expressed in a state-transition probability matrix. The state-transition probability is defined by Eq. (1) and satisfies the probability constraints of Eq. (2) For an ergodic model, one can calculate the "stationary" or "equilibrium" state probability, the probability that the system is in a particular state at an arbitrary time, by finding the eigenvalues of the state-transition probability matrix [11].

3. $B = \{b_j\}_{j=1}^{N}$, the set of observation probability distributions. Each function in the set defines the distribution of the observation in the corresponding state. These functions can take the form of a discrete density when the observation assumes one of a finite set of values, or a probability density function for observations that are real-valued over a continuous range. The former is often referred to as a *discrete* HMM; the latter, a *continuous-density* HMM.
4. $\pi = \{\pi_i\}_{i=1}^{N}$, the initial-state distribution. The matrix is defined according to Eq. (3).

3. THREE BASIC PROBLEMS OF HMM

Three basic problems associated with the HMM must be solved for the model to be useful in real-world applications. The first, the evaluation problem, relates to the computation of the probability of an observed sequence evaluated on a given model. This is important because a given model represents our knowledge of an information source, and the evaluated probability indicates how likely it is that the observed sequence came from the information source. The second, the decoding problem, is to uncover the sequence of states that is most likely to have led to the generation of the observed sequence in some optimal sense. We have mentioned that in many real-world problems a state often carries a certain physical meaning, such as realization of a phoneme in a speech utterance or a letter in handwritten script. Decoding aims at making the associated meaning explicit. The third, the estimation problem, is about obtaining the values of the model parameters, given an observation sequence or a set of observation sequences. A model can be viewed as a characterization of the regularity in the "random" event, which forms the basis of our knowledge of the source. The regularity is encapsulated by the model parameters, which have to be estimated from a set of given observation sequences, known to have come from the source.

3.1. The Evaluation Problem

We wish to calculate the probability of the observation sequence $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T})$, given the model Λ, specifically, to compute $P(\boldsymbol{X}|\Lambda)$. Obviously, this can be accomplished by enumerating every possible state sequence of length T together with the observation sequence as defined in Eqs. (4) and (5):

$$\begin{aligned} P(\boldsymbol{X}|\Lambda) &= \sum_{\boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{q}|\Lambda) \\ &= \sum_{q_0, q_1, \ldots, q_T} \pi_{q_0} a_{q_0 q_1} b_{q_1}(\boldsymbol{x}_1) a_{q_1 q_2} b_{q_2}(\boldsymbol{x}_2) \\ &\quad \times \cdots a_{q_{T-1} q_T} b_{q_T}(\boldsymbol{x}_T) \end{aligned} \tag{6}$$

This direct calculation, however, is not particularly efficient as its computational complexity is exponential in time. The total number of state sequences is N^T and approximately $2T$ essential calculations are required for each state sequence. It thus involves on the order of $2TN^T$ calculations. This is computationally infeasible even for small values of N and T. For example, for $N = 5$ and $T = 100$, the total number of calculations would be on the order of 10^{72}!

An efficient procedure, called the *forward algorithm*, exists that evaluates the HMM probability in linear time. Define the forward probability $\alpha_t(i)$ as

$$\alpha_t(i) = P(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_t}, q_t = i|\Lambda) \tag{7}$$

that is, the probability of the partial observation sequence $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t)$ and state i at time t. Calculation of the forward probabilities can be realized inductively as follows:

1. *Initialization:*

$$\alpha_0(i) = \pi_i \tag{8}$$

2. *Induction:*

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\boldsymbol{x_{t+1}}) \qquad (0 \le t \le T-1, \quad 1 \le j \le N) \tag{9}$$

3. *Termination:*

$$P(\boldsymbol{X}|\Lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{10}$$

This procedure requires on the order of N^2T calculations. Using the same example of $N = 5$ and $T = 100$, we see that this procedure entails about 3000 calculations instead of 10^{72} as required in the direct procedure.

In a similar manner, one can define the backward probability $\beta_t(i)$ as the probability of the partial observation

sequence from $t+1$ to the end, given the system state i at time t:

$$\beta_t(i) = P(\boldsymbol{x_{t+1}}, \boldsymbol{x_{t+2}}, \ldots, \boldsymbol{x_T} | q_t = i, \Lambda) \tag{11}$$

$$\beta_T(i) = 1 \quad (1 \le i \le N) \tag{12}$$

An induction procedure similar to Eq. (9) can be employed to compute the backward probability for all i and t. The forward and the backward probabilities are very useful in solving other fundamental problems in hidden Markov modeling.

3.2. The Decoding Problem

When the state of the system carries information of interest, it may be necessary to "uncover" the state sequence that led to the observation event $\boldsymbol{X}$. There are several ways of solving this problem, depending on the definition of the "optimal" state sequence and the associated criterion that one may choose to optimize. For example, one possible optimality criterion is to choose the states q_t^* that are individually most likely at each time t:

$$q_t^* = \arg \max_{1 \le i \le N} P(q_t = i | \boldsymbol{X}, \Lambda) \qquad (1 \le t \le T) \tag{13}$$

This optimality criterion maximizes the expected number of correct individual states. Other criteria are obviously possible, such as one that solves for the state sequence that maximizes the expected number of correct pairs of states (q_t, q_{t+1}) or triples of states (q_t, q_{t+1}, q_{t+2}).

The most widely used criterion is to find the *single* best state sequence to maximize $P(\boldsymbol{q}|\boldsymbol{X}, \Lambda)$, which is equivalent to maximizing $P(\boldsymbol{q}, \boldsymbol{X}|\Lambda)$, the joint state-observation probability. A dynamic programming based algorithm, called the Viterbi algorithm [12], can be used to efficiently find the best single-state sequence.

3.2.1. The Viterbi Algorithm. The Viterbi algorithm can be used to find the single best state sequence $\boldsymbol{q}^*$ defined as

$$\boldsymbol{q}^* = \arg \max_{\boldsymbol{q}} P(\boldsymbol{q}, \boldsymbol{X}|\Lambda) \tag{14}$$

for a given observation sequence $\boldsymbol{X}$. To accomplish the goal of maximization, define the following best partial score

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1, q_2, \ldots, q_{t-1}, q_t = i, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t | \Lambda) \tag{15}$$

that is, the maximum probability at time t along a single path that accounts for the first t observations and ends in state i. The best partial score can be computed by induction as follows:

1. *Initialization:*

$$\delta_0(i) = \pi_i \tag{16}$$

$$\psi_0(i) = 0 \tag{17}$$

2. *Recursion:*

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(\boldsymbol{x}_t) \qquad (1 \le t \le T, \quad 1 \le j \le N) \tag{18}$$

$$\psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] \qquad (1 \le t \le T, \quad 1 \le j \le N) \tag{19}$$

3. *Termination:*

$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \tag{20}$$

$$q_T^* = \arg \max_{1 \le i \le N} [\delta_T(i)] \tag{21}$$

4. *Backtracking:*

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \qquad (t = T-1, \quad T-2, \ldots, 0) \tag{22}$$

The array $\psi_t(j), t = T, \ T-1, \ldots, 1$ and $j = 1, 2, \ldots, N$ records the partial optimal state sequence, and is necessary in producing the single best state sequence $\boldsymbol{q}^*$. The Viterbi algorithm has been widely used in many applications such as speech recognition and data communication.

3.3. The Estimation Problem

A signal-modeling task involves estimating the parameter values from a given set of observations, say, $\Omega_X = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L\}$. The HMM parameters to be estimated is the set $\Lambda = (\pi, A, B = \{b_j\}_{j=1}^N)$. (In the following, we shall speak of the model and the model parameter set interchangeably without ambiguity.) Parameter estimation is normally carried out according to some well-known optimization criterion; "maximum likelihood" (ML) is one of the most prevalent.

3.3.1. Maximum Likelihood. The ML estimate of the model is obtained as

$$\Lambda_{\mathrm{ML}} = \arg \max_{\Lambda} P(\Omega_X | \Lambda) \tag{23}$$

For HMM, unfortunately, there is no known way to analytically solve for the model parameter optimization problem in closed form. Instead, a general hill-climbing algorithm is used to iteratively improve the model parameter set until the procedure reaches a fixed-point solution, which is at least locally optimal. The algorithm is called the *Baum–Welch algorithm* [13] [or the expectation–maximization (EM) algorithm [14] in other statistical contexts].

The Baum–Welch algorithm accomplishes likelihood maximization in a two-step procedure, known as "reestimation." On the basis of an existing model parameter set Λ, the first step of the algorithm is to transform the objective function $P(\Omega_X|\Lambda)$ into an auxiliary function $\boldsymbol{Q}(\Lambda, \Lambda')$, which measures a divergence between the model Λ and

another model Λ', a variable to be optimized. The auxiliary function is defined, for the simplest case with a single observation sequence $\boldsymbol{X}$, as

$$Q(\Lambda, \Lambda') = \sum_{\boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{q}|\Lambda) \log P(\boldsymbol{X}, \boldsymbol{q}|\Lambda') \tag{24}$$

where $P(\boldsymbol{X}, \boldsymbol{q}|\Lambda)$ can be found in Eq. (6). It can be shown that $Q(\Lambda, \Lambda') \geq Q(\Lambda, \Lambda)$ for a certain Λ' implies $P(\boldsymbol{X}, \boldsymbol{q}|\Lambda') \geq P(\boldsymbol{X}, \boldsymbol{q}|\Lambda)$. Therefore, the second step of the algorithm involves maximizing $Q(\Lambda, \Lambda')$ as a function of Λ' to obtain a higher, improved likelihood. These two steps iterate interleavingly until the likelihood reaches a fixed point. Detailed derivation of the reestimation formulas can be found in three papers [1–3].

3.3.2. Other Optimization Criteria. The need to consider other optimization criteria comes primarily from the potential inconsistency between the form of the chosen model (i.e., an HMM) and that of the true distribution of the data sequence. If inconsistency or model mismatch exists, the optimality achieved in ML (maximum likelihood) estimation may not represent its real significance, either in the sense of data fitting or in indirect applications such as statistical pattern recognition. In statistical pattern recognition, one needs to evaluate and compare the a posteriori probabilities of all the classes, on receipt of an unknown observation, in order to achieve the theoretically optimal performance of the so-called Bayes risk [15]. A model mismatch prevents one from achieving the goal of Bayes risk. In these situations, application of optimization criteria other than maximum likelihood may be advisable.

One proposal to deal with the potential problem of a model mismatch is the method of minimum discrimination information (MDI) [16]. Let the observation sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ be associated with a constraint $\boldsymbol{R}$, for example, $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_T)$, in which each $\boldsymbol{r}_t$ is an autocorrelation vector corresponding to a short-time observation $\boldsymbol{x}_t$. Note that $\boldsymbol{X}$ is just a realization of a random event, which might be governed by a set of possibly uncountably many distributions that satisfy the constraint $\boldsymbol{R}$. Let's denote this set of distributions by $\Theta(\boldsymbol{R})$. The minimum discrimination information is a measure of closeness between two probability measures under the given constraint $\boldsymbol{R}$ and is defined by

$$v(\boldsymbol{R}, P(\boldsymbol{X}|\Lambda)) \equiv \inf_{G \in \Theta(R)} I(G : P(\boldsymbol{X}|\Lambda)) \tag{25}$$

where

$$I(G : P(\boldsymbol{X}|\Lambda)) \equiv \int g(\boldsymbol{X}) \log \frac{g(\boldsymbol{X})}{p(\boldsymbol{X}|\Lambda)} d\boldsymbol{X} \tag{26}$$

where $g(\cdot)$ and $p(\cdot|\Lambda)$ denote the probability density functions corresponding to G and $P(\boldsymbol{X}|\Lambda)$, respectively. The MDI criterion tries to choose a model parameter set such that $v(R, P(\boldsymbol{X}|\Lambda))$ is minimized. An interpretation of MDI is that it attempts to find not just the HMM parameter set to fit the data $\boldsymbol{X}$ but also an HMM that is as close as it can be to a member of the distribution set $\Theta(\boldsymbol{R})$, of which $\boldsymbol{X}$ could have been a true realization. If $\boldsymbol{X}$ is indeed from a hidden Markov source (of the right order), the attainable minimum discrimination information is zero. While the MDI criterion does not fundamentally change the problem in model selection, it provides a way to deemphasize the potentially acute mismatch between the data and the chosen model.

Another concern about the choice of the HMM optimization criterion arises in pattern (e.g., speech) recognition problems in which one needs to estimate a number of distributions, each corresponding to a class of events to be recognized. Let V be the number of classes of events, each of which is characterized by an HMM Λ_v, $1 \leq v \leq V$. Also let $P(v)$ be the prior distribution of the classes of events. The set of HMMs and the prior distribution thus define a probability measure for an arbitrary observation sequence $\boldsymbol{X}$:

$$P(\boldsymbol{X}) = \sum_{v=1}^{V} P(\boldsymbol{X}|\Lambda_v) P(v) \tag{27}$$

A measure of mutual information between class v and a realized class v observation $\boldsymbol{X}^v$ can be defined, in the context of the composite probability distribution $\{P(\boldsymbol{X}^v|\Lambda_v)\}_{v=1}^V$, as

$$\begin{aligned} I(\boldsymbol{X}^v, v; \{\Lambda_v\}_{v=1}^V) &= \log \frac{P(\boldsymbol{X}^v, v|\{\Lambda_v\}_{v=1}^V)}{P(\boldsymbol{X})P(v)} \\ &= \log P(\boldsymbol{X}^v|\Lambda_v) - \log \sum_{w=1}^{V} P(\boldsymbol{X}^v|\Lambda_w) P(w) \end{aligned} \tag{28}$$

The maximum mutual information (MMI) criterion [17] is to find the entire model parameter set such that the mutual information is maximized:

$$(\{\Lambda_v\}_{v=1}^V)_{\text{MMI}} = \arg \max_{\{\Lambda_v\}_{v=1}^V} \left\{ \sum_{v=1}^{V} I(\boldsymbol{X}^v, v; \{\Lambda_v\}_{v=1}^V) \right\} \tag{29}$$

The key difference between an MMI and an ML estimate is that optimization of model parameters is carried out on all the models in the set for any given observation sequence. Equation (28) also indicates that the mutual information is a measure of the class likelihood evaluated in contrast to an overall "probability background" (the second term in the equation). Since all the distributions are involved in the optimization process for the purpose of comparison rather than individual class data fitting, it in some way mitigates the model mismatch problem. For direct minimization of recognition errors in pattern recognition problems, the minimum classification error (MCE) criterion provides another alternative [18].

Optimization of these alternative criteria is usually more difficult than the ML hill-climbing method and requires general optimization procedures such as the descent algorithm to obtain the parameter values.

4. TYPES OF HIDDEN MARKOV MODELS

Hidden Markov models can be classified according to the Markov chain topology and the form of the in-state observation distribution functions. Examples of the

Markov chain topology have been shown in Fig. 2. The in-state observation distribution warrants further discussion as it significantly affects the re-estimation procedure.

In many applications, the observation is discrete, assuming a value from a finite set or alphabet, with cardinality, say, M, and thus warrants the use of a discrete HMM. Associated with a discrete HMM is the set of observation symbols $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M\}$, where M is the number of distinct observation symbols in each state. The observation probability distribution is defined, for state j, as

$$b_j(\boldsymbol{x}_t = \boldsymbol{v}_k) = b_j(k) = P(\boldsymbol{x}_t = \boldsymbol{v}_k | q_t = j) = b_{jk} \tag{30}$$

forming a matrix $B = [b_{jk}]$, $1 \le j \le N$, $1 \le k \le M$. To obtain an ML estimate, the Baum–Welch algorithm involves maximization of the auxiliary function $Q(\Lambda, \Lambda')$ defined in Eq. (24). Note that

$$\log P(\boldsymbol{X}, \boldsymbol{q} | \Lambda') = \log \pi'_{q_0} + \sum_{t=1}^{T} \log a'_{q_{t-1} q_t} + \sum_{t=1}^{T} \log b'_{q_t}(\boldsymbol{x}_t) \tag{31}$$

which allows us to write the auxiliary function in the following form

$$Q(\Lambda, \Lambda') = Q_\pi(\Lambda, \pi') + \sum_{i=1}^{N} Q_a(\Lambda, \boldsymbol{a}'_i) + \sum_{i=1}^{N} Q_b(\Lambda, \boldsymbol{b}'_i) \tag{32}$$

where $\pi' = [\pi'_1, \pi'_2, \ldots, \pi'_N]$, $\boldsymbol{a}'_i = [a'_{i1}, a'_{i2}, \ldots, a'_{iN}]$, $\boldsymbol{b}'_i = [b'_{i1}, b'_{i2}, \ldots, b'_{iM}]$, and

$$Q_\pi(\Lambda, \pi') = \sum_{i=1}^{N} P(\boldsymbol{X}, q_0 = i | \Lambda) \log \pi'_i \tag{33}$$

$$Q_a(\Lambda, \boldsymbol{a}'_i) = \sum_{j=1}^{N} \sum_{t=1}^{T} P(\boldsymbol{X}, q_{t-1} = i, q_t = j | \Lambda) \log a'_{ij} \tag{34}$$

$$Q_b(\Lambda, \boldsymbol{b}'_i) = \sum_{t=1}^{T} P(\boldsymbol{X}, q_t = i | \Lambda) \log b'_i(\boldsymbol{x}_t) \tag{35a}$$

$$= \sum_{m=1}^{M} \sum_{t=1}^{T} P(\boldsymbol{X}, q_t = i | \Lambda) \delta(\boldsymbol{x}_t, \boldsymbol{v}_m) \log b'_{im} \tag{35b}$$

Note that in Eq. 35b, which is for a discrete HMM specifically, we define

$$\delta(\boldsymbol{x}_t, \boldsymbol{v}_m) = 1 \quad \text{if} \quad \boldsymbol{x}_t = \boldsymbol{v}_m; \quad = 0, \qquad \text{otherwise.}$$

These individual terms can be maximized separately over π', $\{\boldsymbol{a}'_i\}_{i=1}^N$, $\{\boldsymbol{b}'_i\}_{i=1}^N$. They share an identical form of an optimization problem. Find y_i, $i = 1, 2, \ldots, N$, (or M in 35b) to maximize $\sum_{i=1}^{N} w_i \log y_i$ subject to $\sum_{i=1}^{N} y_i = 1$ and $y_i \ge 0$. This problem attains a global maximum at the single point

$$y_i = \frac{w_i}{\sum_{k=1}^{N} w_k}, \; i = 1, 2, \ldots, N \tag{36}$$

This leads to the following reestimation transformation for the HMM parameters

$$\overline{\pi}_i = \frac{P(\boldsymbol{X}, q_0 = i | \Lambda)}{\sum_{j=1}^{N} P(\boldsymbol{X}, q_0 = j | \Lambda)} = \frac{P(\boldsymbol{X}, q_0 = i | \Lambda)}{P(\boldsymbol{X} | \Lambda)} \tag{37}$$

$$\begin{aligned} \overline{a}_{ij} &= \frac{\sum_{t=1}^{T} P(\boldsymbol{X}, q_{t-1} = i, q_t = j | \Lambda)}{\sum_{k=1}^{N} \sum_{t=1}^{T} P(\boldsymbol{X}, q_{t-1} = i, q_t = k | \Lambda)} \\ &= \frac{\sum_{t=1}^{T} P(\boldsymbol{X}, q_{t-1} = i, q_t = j | \Lambda)}{\sum_{t=1}^{T} P(\boldsymbol{X}, q_{t-1} = i | \Lambda)} \end{aligned} \tag{38}$$

$$\begin{aligned} \overline{b}_{im} &= \frac{\sum_{t=1}^{T} P(\boldsymbol{X}, q_t = i | \Lambda) \delta(\boldsymbol{x}_t, \boldsymbol{v}_m)}{\sum_{k=1}^{M} \sum_{t=1}^{T} P(\boldsymbol{X}, q_t = i | \Lambda) \delta(\boldsymbol{x}_t, \boldsymbol{v}_k)} \\ &= \frac{\sum_{t=1}^{T} P(\boldsymbol{X}, q_t = i | \Lambda) \delta(\boldsymbol{x}_t, \boldsymbol{v}_m)}{\sum_{t=1}^{T} P(\boldsymbol{X}, q_t = i | \Lambda)} \end{aligned} \tag{39}$$

where $\{\overline{\pi}\}$, $\{\overline{a}\}$, and $\{\overline{b}\}$ achieve $\max Q(\Lambda, \Lambda')$ as a function of Λ'.

For continuous-density HMMs, the same iterative procedure applies, although one needs to construct the right form of the auxiliary function in order to be able to carry out the maximization step. During the course of development of the HMM theory, Baum et al. [6] first showed an algorithm to accomplish ML estimation for an HMM with continuous in-state observation densities that are log-concave. The Cauchy distribution, which is not log-concave, was cited as one that the algorithm would have difficulty with. Later, Liporace [4], by invoking a representation theorem, relaxed this restriction and extended the algorithm to enable it to cope with the family of elliptically symmetric density functions. The algorithm was further enhanced at Bell Laboratories in the early 1980s [5] and extended to the case of general mixture densities taking the form, for state i

$$b_i(\boldsymbol{x}) = \sum_{k=1}^{M} c_{ik} f(\boldsymbol{x}; \boldsymbol{b}_{ik}) \tag{40}$$

where M is the number of mixture components, c_{ik} is the weight for mixture component k, and $f(\boldsymbol{x}; \boldsymbol{b}_{ik})$ is a continuous probability density function, log-concave or elliptically symmetric, parameterized by $\boldsymbol{b}_{ik}$. The significance of this development is that the modeling technique now has the capacity to deal with virtually any distribution function since the mixture density can be

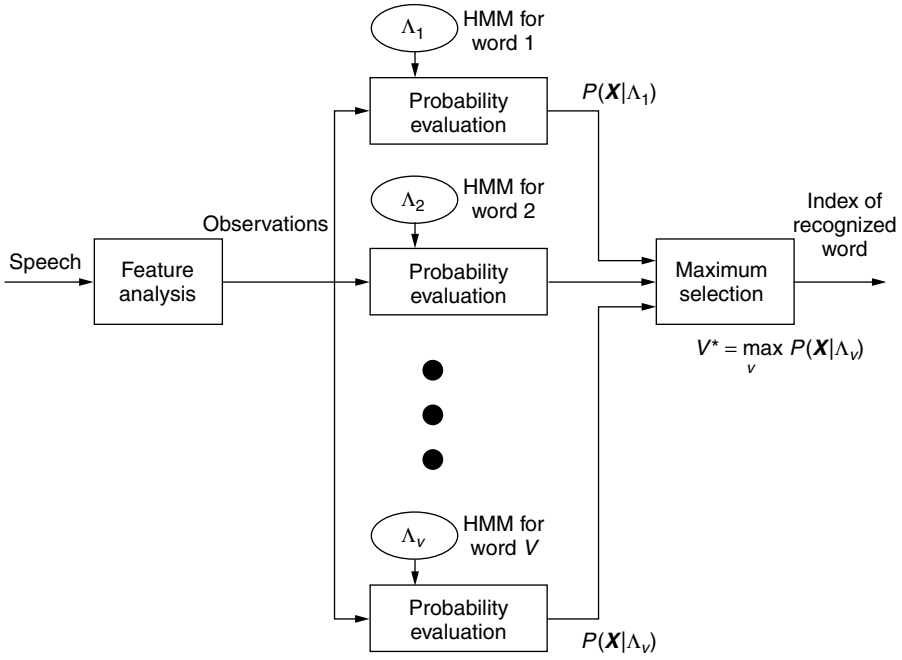


Figure 3. Block diagram of an isolated word HMM recognizer (after Rabiner [3]).

used to approximate a distribution with arbitrary accuracy (by increasing M). This is important because many real-world processes such as speech parameters do not follow the usual symmetric or single modal distribution form. For detailed derivations of the reestimation formulas for various kinds of density functions embedded in a hidden Markov model, consult Rabiner [3] and Rabiner and Juang [10].

5. APPLICATIONS OF HMM

Since its inception in the late 1960s and through its development in the 1970s and 1980s, the HMM has found widespread use in many problem areas, such as stockmarket analysis, cryptoanalysis, telecommunication modeling and management, and most notably automatic speech recognition, which we highlight here.

Automatic speech recognition (ASR) was traditionally treated as a speech analysis problem because of the complex nature of the signal. The main impact of the HMM methodology was to formalize the statistical approach as a feasible paradigm in dealing with the vast variability in the speech signal. Major sources of variability in a speech signal include the inherent uncertainty in speaking rate variation, physiological variation such as speaker-specific articulatory characteristics, changes in speaking conditions (e.g., the Lombard effect [19]), uncertainty and coarticulation in speech production (e.g., underarticulated phonemes), regional accents, and use of language. The HMM has an intrinsic ability to cope with the speaking rate variation due to the Markov chain, but the broad nature of variability calls for the use of a mixture-density HMM, particularly in the design of a system that is supposed to perform equally well for a large population of users (i.e., a speaker-independent speech recognition system [10]). Today, most high-performance large-vocabulary ASR systems are based on mixture-density HMMs.

Figure 3 illustrates the use of the HMM in an isolated word recognition task. A recognizer requires the knowledge of the a posteriori probabilities $P(v|\boldsymbol{X})$ for all the words/classes, v, $1 \leq v \leq V$, evaluated on an observed pattern or sequence, $\boldsymbol{X}$. For simplicity, we assume that the prior probabilities for all the classes are equal and the recognition decision is thus based on the likelihood functions $P(\boldsymbol{X}|\Lambda_v)$, $1 \leq v \leq V$, with each model Λ_v corresponding to a class v. The models $P(\boldsymbol{X}|\Lambda_v)$, $1 \leq v \leq V$, need to be "trained" on the basis of a set of known samples during the design phase. "Training" is the task of model parameter estimation using one or more observation sequences or patterns. Section 3.3 outlines the procedure for parameter estimation. Once these models are trained, they are stored in the system for likelihood evaluation on receipt of an unknown observation sequence. The solution based on the forward probability calculation as discussed in Section 3.1 is an efficient way to obtain the likelihood for each word in the vocabulary. The word model that achieves the highest likelihood determines the recognized word. This simple design principle has become the foundation of many ASR systems.

For more sophisticated speech recognition tasks, such as large-vocabulary, continuous speech recognition, the design principle described above has to be modified to cope with the increased complexity. In continuous speech recognition, event classes (whether they are based on "lexical words" or "phonemes") are observed consecutively without clear demarcation (or pause), making it necessary to consider composite HMMs, constructed from elementary HMMs. For example, an elementary HMM may be chosen to represent a phoneme, and a sequence of words would thus be modeled by a concatenation of the corresponding phoneme models. The concatenated model is then used

Table 1. Performance Benchmark of HMM-Based Automatic Speech Recognition Systems for Various Tasks or Applications

Task/Application	Vocabulary Size	Perplexity[a]	Word Accuracy (%)
Isolated word/digits	10	N/A	~100
Connected digit strings	10	10	~99
Isolated spell letters and digits	37	N/A	95
Navy resource management	991	<60	97
Air travel information system	1,800	<25	97
Wall St. Journal transcription	64,000	<140	94
Broadcast news transcription	64,000	<140	86

[a]Perplexity in this context is defined as the average number of words that can follow another word and is a rough measure of the difficulty or complexity of the task.

in the likelihood calculation for making the recognition decision. The procedure is rather complex, and dynamic programming techniques together with heuristics and combinatorics (to form search algorithms) are usually employed to obtain the result. Also, because of the much-increased variability in the realization of a continuous speech signal, an elementary model is often qualified by its context (e.g., an "e-**l**-i" model for the phoneme /l/ to be used in the word "element") in order to achieve the needed accuracy in modeling. These context-dependent models greatly increase the complexity of the system design.

Today, HMM-based ASR systems are deployed in telecommunication networks for automatic call routing and information services, and in personal computers for dictation and word-processing. Table 1 provides a gist of the performance, in terms of word accuracy of various systems and tasks. Voice-enabled portals that bring information services to the user over the Internet are also emerging and are expected to gain popular acceptance in the near future. The HMM is the underpinning technology of these modern-day communication services.

BIOGRAPHY

Dr. Biing-Hwang Juang received his Ph.D. in electrical and computer engineering from the University of California, Santa Barbara in 1981 and is currently Director of Multimedia Technologies Research at AVAYA Labs Research, a spin-off of Bell Laboratories. Before joining AVAYA Labs Research, he was Director of Acoustics and Speech Research at Bell Laboratories. He is engaged in and responsible for a wide range of research activities, from speech coding, speech recognition, and intelligent systems to multimedia and broadband communications. He has published extensively and holds a number of patents in the area of pattern recognition, speech communication and telecommunication services. He is co-authr of the book *Fundamentals of Speech Recognition* published by Prentice-Hall. He received the Technical Achievement Award from the IEEE Signal Processing Society in 1998 and the IEEE Third Millennium Medal in 2000. He is an IEEE Fellow and a Bell Labs Fellow.

BIBLIOGRAPHY

1. L. E. Baum, An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes, *Inequalities* **3**: 1–8 (1972).
2. L. R. Rabiner and B. H. Juang, An introduction to hidden Markov models, *IEEE ASSP Mag.* **3**(1): 4–16 (Jan. 1986).
3. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE* **77**(2): 257–286 (Feb. 1989).
4. L. R. Liporace, Maximum likelihood estimation for multivariate observations of Markov sources, *IEEE Trans. Inform. Theory* **IT-28**: 729–734 (Sept. 1982).
5. B. H. Juang, Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains, *AT&T Tech. J.* **64**: 1235–1249 (1985).
6. L. E. Baum, T. Petri, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.* **41**: 164–171 (1970).
7. L. E. Baum and J. A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, *Bull. Am. Math. Soc.* **73**: 360–363 (1967).
8. D. Andelman and J. Reeds, On the cryptoanalysis of rotor machines and substitution-permutation networks, *IEEE Trans. Inform. Theory* **IT-28**(4): 578–584 (July 1982).
9. H. Bunke and T. Caelli, *Hidden Markov Model in Vision,* a special issue of the *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, Singapore, Vol. 45, June 2001.
10. L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
11. A. T. Bharucha-Reid, *Elements of the Theory of Markov Processes and Their Applications*, McGraw-Hill, New York, 1960.
12. G. D. Forney, The Viterbi algorithm, *Proc. IEEE* **61**: 268–278 (March 1973).
13. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, *Bell Syst. Tech. J.* **62**(4): 1035–1074 (April 1983).
14. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc.* **39**(1): 1–38 (1977).
15. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
16. Y. Ephraim, A. Dembo, and L. R. Rabiner, A minimum discrimination information approach for hidden Markov modeling, *IEEE. Trans. Inform. Theory* **IT-35**(5): 1001–1003 (Sept. 1989).

17. A. Nadas, D. Nahamoo, and M. A. Picheny, On a model-robust training method for speech recognition, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-36**(9): 1432–1436 (Sept. 1988).

18. B. H. Juang, Wu Chou, and C. H. Lee, Minimum classification error rate methods for speech recognition, *IEEE Trans. Speech Audio Process.* **SA-5**(3): 257–265 (May 1997).

19. R. P. Lippmann, E. A. Martin, and D. B. Paul, Multi-style training for robust isolated-word speech recognition, *IEEE ICASSP-87 Proc.*, April 1987, pp. 705–708.

HIGH-DEFINITION TELEVISION*

JAE S. LIM
Massachusetts Institute of Technology
Cambridge, Massachusetts

A high-definition television (HDTV) system is a television system whose performance is significantly better than a conventional television system. An HDTV system delivers spectacular video and multichannel CD (compact disk) quality sound. The system also has many features that are absent in conventional systems, such as auxiliary data channels and easy interoperability with computers and telecommunications networks.

Conventional television systems were developed during the 1940s and 1950s. Examples of conventional systems are NTSC (National Television Systems Committee), SECAM (Sequential Couleur a Memoire), and PAL (phase-alternating line). These systems have comparable performance in video quality, audio quality, and transmission robustness. The NTSC system, used in North America, will be used as a reference for conventional television systems when discussing HDTV in this article.

For many decades, conventional television systems have been quite successful. However, they were developed on the basis of the technology that was available during the 1940s and 1950s. Advances in technologies such as communications, signal processing, and very-large-scale integration (VLSI) has enabled a major redesign of a television system with substantial improvements over conventional television systems. An HDTV system is one result of this technological revolution.

1. CHARACTERISTICS OF HDTV

Many characteristics of an HDTV system markedly differ from a conventional television system. These characteristics are described in this section.

1.1. High Resolution

An HDTV system can deliver video with spatial resolution much higher than a conventional television system. Typically, video with a spatial resolution of at least four times that of a conventional television system is called *high-resolution video*. Resolution represents the amount of detail contained within the video, which can also be called "definition." This is the basis for high-definition television. An NTSC system delivers video at a resolution of approximately 480 lines in an interlaced format at an approximate rate of 60 fields/s (it is actually 59.94 Hz, but we will not make a distinction between 59.94 and 60). Each line of resolution contains approximately 420 pixels or picture elements. The number of lines represents the vertical spatial resolution in the picture, and the number of pixels per line represents the horizontal spatial resolution. *Interlaced scanning* refers to the scanning format. All conventional television systems use this format. Television systems deliver pictures that are snapshots of a scene recorded at a certain number of times per second. In interlaced scanning, a single snapshot consists of only odd lines, the next snapshot consists of only even lines, and this sequence repeats. A snapshot in interlaced scanning is called a *field*. In the NTSC system, 60 fields are used per second. Although only snapshots of a scene are shown, the human visual system perceives it as continuous motion, as long as the snapshots are shown at a sufficiently high rate. In this way, the video provides accurate motion rendition.

More lines and more pixels per line in a field provide more spatial details than the field can retain. An HDTV system may have 1080 lines and 1920-pixel/line resolution in an interlaced format of 60 fields/sec. In this case, the spatial resolution of an HDTV system would be almost 10 times that of an NTSC system. This high spatial resolution is capable of showing details in the picture much more clearly, and the resultant video appears much sharper. It is particularly useful for sports events, graphic material, written letters, and movies.

The high spatial resolution in an HDTV system enables a large-screen display and increased realism. For an NTSC system, the spatial resolution is not high. To avoid the visibility of a line structure in an NTSC system, the recommended viewing distance is approximately 7 times the picture height. For a 2-ft-high display screen the recommended viewing distance from the screen is 14 ft, 7 times the picture height. This makes it difficult to have a large-screen television receiver in many homes. Because of the long viewing distance, the viewing angle is approximately 10°, which limits realism. For an HDTV system with more than twice the number of lines, the recommended viewing distance is typically 3 times the picture height. For a 2-ft-high display, the recommended viewing distance would be 6 ft. This can accommodate a large-screen display in many environments. Because of a short viewing distance and wider aspect (width-to-height) ratio, the viewing angle for an HDTV system is approximately 30°, which significantly increases realism.

An HDTV system can also deliver higher temporal resolution by using progressive scanning. Unlike interlaced scanning, where a snapshot (field) consists of only even lines or only odd lines, all the lines in progressive scanning are scanned for each snapshot. The snapshot in progressive scanning is called a *frame*. Both progressive scanning and interlaced scanning have their own merits, and the choice between the two generated much discussion during the digital television standardization process in the

* This article is a modified version of High Definition Television, published in the *Wiley Encyclopedia of Electrical and Electronics Engineering*; 1999, Vol. 8, pp. 725–739.

United States. An HDTV system can have only interlaced scanning, or only progressive scanning, or a combination of the two.

An HDTV system delivers video with substantially higher spatial and temporal resolution than does a conventional television system. In addition to its superior resolution, an HDTV system typically has other important features, discussed below.

1.2. Wide Aspect Ratio

An NTSC television receiver has a display area with an aspect ratio of 4:3. The aspect ratio is a width-to-height ratio. The 4:3 aspect ratio was chosen because movies were made with a 4:3 aspect ratio when the NTSC system was first developed. Since then, movies have been made with a wider aspect ratio. To reflect this change, an HDTV system typically has a wider aspect ratio, such as 16:9. The difference in spatial resolution and aspect ratio between an NTSC system and an HDTV system is illustrated in Fig. 1. Figure 1a is a frame with an aspect ratio of 4:3, and Fig. 1b is a frame with an aspect ratio of 16:9. The difference in spatial detail between the two pictures is approximately the difference in spatial resolution between a conventional television and HDTV.

1.3. Digital Representation and Transmission

In a conventional television system, the video signal is represented in an analog format, and the analog representation is transmitted. However, the analog representation is highly susceptible to channel transmission degradations such as multipath effects or random noise. In a conventional television system, video received through the air (terrestrial broadcasting) often has visible degradations such as ghosts and snowlike noise.

In an HDTV system, the video signal is represented digitally and transmitted. An HDTV system can be developed using an analog transmission system. However, transmission of the digital representation of the video signal is much more efficient in the bandwidth utilization. Digital transmission can utilize such modern technologies as digital video compression and digital communications. The effects of channel degradation manifest themselves differently in a digital transmission system. In an HDTV system that is broadcast digitally over the air, the video received is essentially perfect within a certain coverage area (within a certain level of channel degradation). Outside that area, the video is not viewable. Unlike an analog NTSC system, where the video degrades more as the channel degradation increases, a digital HDTV system will deliver either an essentially perfect picture or no picture at all. This is referred to as the "cliff effect" or "digitally clean" video.

1.4. Multichannel Digital Audio

An HDTV system has the capability to deliver multichannel sound. The number of audio channels that can accompany a video program may be as many as one desires. Multiple audio channels can be used to produce the effect of surround sound, which is often used in movie theaters. They can also be used for transmitting different languages in the same video program. In addition to multichannel sound, the reproduced sound achieves the quality of an audio CD (compact disk).

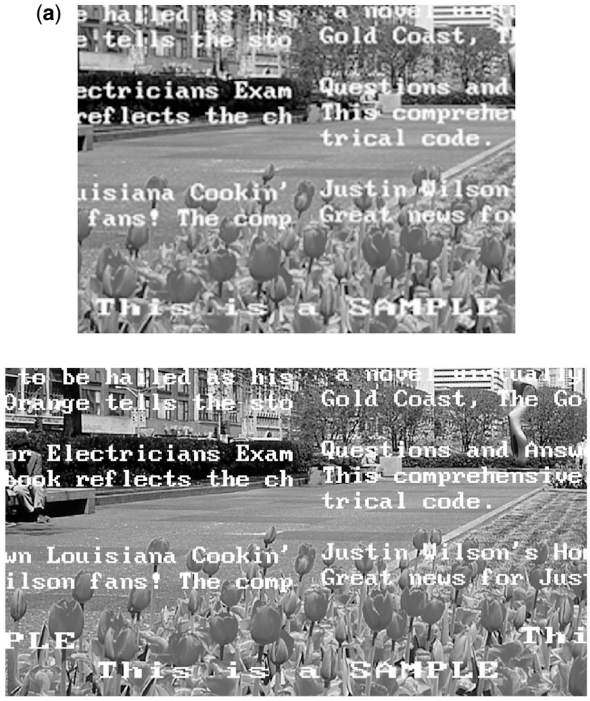


Figure 1. Resolution and aspect ratio of a conventional and a high-definition television system: (**a**) a segment of a conventional television video frame — 4:3 aspect ratio; (**b**) the corresponding segment of a high-definition television video frame — 16:9 aspect ratio with higher spatial resolution than in part (a).

A television system is often considered as primarily a video service. However, the audio service is particularly important for HDTV applications. Generally, people will not watch video with poor-quality audio, even when the video quality may be comparable to HDTV. In addition, high-quality audio enhances our visual experience. The same video, when accompanied by higher-quality audio, gives the impression of higher-quality video than when it is accompanied by low-quality audio. An HDTV system delivers multichannel audio with CD-quality sound. In addition to a superb listening experience, it enhances our visual experience beyond what is possible with high-resolution video alone.

1.5. Data Channel

A conventional television system is a standalone system whose primary objective is entertainment. A digital HDTV system utilizes a data transmission channel. Its data can represent not only high-resolution video and audio but also any digital data like computer data, newspapers, telephone books, and stockmarket quotes. The digital HDTV system can be integrated easily to operate with computers and telecommunication networks.

2. HISTORY OF HDTV IN THE UNITED STATES

The NTSC system was developed for the terrestrial transmission of television signals. Since the NTSC system

requires 6 MHz of bandwidth, the available VHF (very-high-frequency) and UHF (ultra-high-frequency) bands, which are suitable for the terrestrial broadcasting of television signals, were divided into 6-MHz channels. Initially, there was plenty of spectrum. The NTSC system, however, utilizes its given 6 MHz of spectrum quite inefficiently. This inefficiency generates interference among the different NTSC signals. As the number of NTSC signals that were broadcast terrestrially increased, the interference problem became serious. The solution was to avoid using certain channels. These unused channels are known as "taboo channels." In a typical highly populated geographic location in the United States, only one of two VHF channels is used and only one of six UHF channels is used. In addition, in the 1980s, other services such as mobile radio requested the use of the UHF band spectrum. As a result, an HDTV system that requires a large amount of bandwidth, such as Japan's MUSE system, was not an acceptable solution for terrestrial broadcasting in the United States.

At the request of the broadcast organizations, the United States Federal Communications Commission (FCC) created the Advisory Committee on Advanced Television Service (ACATS) in September 1987. ACATS was chartered to advise the FCC on matters related to the standardization of the advanced television service in the United States, including establishment of a technical standard. At the request of ACATS, industries, universities, and research laboratories submitted proposals for the ATV (advanced television) technical standard in 1988.

While the ACATS screened the proposals and prepared testing laboratories for their formal technical evaluation, the FCC made a key decision. In March 1990, the FCC selected the simulcast approach for advanced television service rather than the receiver-compatible approach, in which existing NTSC television receivers can receive an HDTV signal and generate a viewable picture. This was the approach taken when the NTSC introduced color. A black-and-white television receiver can receive a color television signal and display it as a viewable black-and-white picture. In this way, the then-existing black-and-white television receivers would not become obsolete. It was possible to use the receiver-compatible approach in the case of color introduction. This is because color information did not require a large amount of bandwidth and a small portion of the 6-MHz channel used for a black-and-white picture could be used to insert the color information without seriously affecting the black-and-white picture.

In the case of HDTV, the additional information needed was much more than the original NTSC signal and the receiver-compatibility requirement would require additional spectrum to carry the HDTV signal. Among the proposals received, the receiver-compatible approaches typically required a 6-MHz augmentation channel that carried the enhancement information, which was the difference between the HDTV signal and the NTSC signal. Even though the augmentation approach solves the receiver-compatibility problem, it has several major problems. The approach requires an NTSC channel as a basis to transmit an HDTV signal. This means that the highly spectrum-inefficient NTSC system cannot be converted to a more efficient technical system. In addition, the introduction of HDTV would permanently require a new channel for each existing NTSC channel. The FCC rejected this spectrum-inefficient augmentation channel approach.

Although the FCC's decision did not require receiver compatibility, it did require transmission of an entire HDTV signal within a single 6-MHz channel. In the simulcast approach adopted by the FCC, an HDTV signal that can be transmitted in a single 6-MHz channel can be designed independently of the NTSC signal. An NTSC television receiver cannot receive an HDTV signal. In order to receive an HDTV signal, a new television receiver would be needed. To ensure that existing television receivers do not become obsolete when HDTV service is introduced, the FCC would give one new channel for HDTV service to each NTSC station that requested it. During the transition period, both NTSC and HDTV services coexist. After sufficient penetration of HDTV service, NTSC service will be discontinued. The spectrum previously occupied by NTSC services will be used for additional HDTV channels or for other services. Initially, the FCC envisioned that the new HDTV channel and the existing NTSC channel would carry the same programs, so as not to disadvantage NTSC receivers during the transition period. This is the basis for the term "simulcasting." Later, this requirement was removed. The simulcast approach is illustrated in Fig. 2.

The simulcast approach provides several major advantages. It presents the possibility of designing a new spectrum-efficient HDTV signal that requires significantly less power and that does not interfere with other signals, including the NTSC signal. This allows the use of the taboo channels, which could not be used for additional NTSC service because of the strong interference characteristics of the NTSC signals. Without the taboo channels, it would not have been possible to give an additional channel to each existing NTSC broadcaster for HDTV service. In addition, it eliminates the spectrum-inefficient NTSC channels following the transition period. The elimination

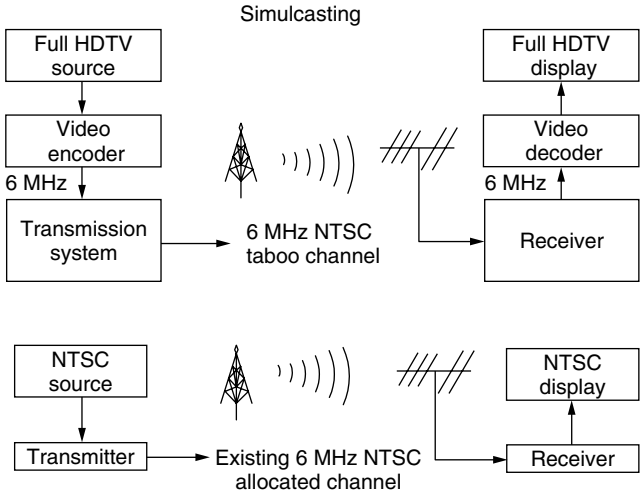


Figure 2. Flowchart of simulcasting approach for transition from an NTSC system to a digital high-definition television system.

of NTSC broadcasting vacates the spectrum that it occupied. Furthermore, by removing the NTSC signals that have strong interference characteristics, other channels could be used more efficiently. The 1990 FCC ruling was a key decision in the process to standardize the HDTV system in the United States.

The 1990 decision also created several technical challenges. The HDTV signal had to be transmitted in a single 6-MHz channel. In addition, the signal was required to produce minimal interference with NTSC signals and other HDTV signals. At the time of the FCC's decision in 1990, it was not clear whether such a system could be developed within a reasonable period of time. Later events proved that developing such a system at a reasonable cost to broadcasters and consumers was possible using modern communications, signal processing, and VLSI technologies.

Before the formal technical evaluation of the initial HDTV proposals began, some were eliminated, others were substantially modified, and still others combined their efforts. Five HDTV system proposals were ultimately approved for formal evaluation. One proposed an analog system while four others proposed all-digital systems. The five systems were evaluated in laboratory tests at the Advanced Television Testing Center (ATTC) in Alexandria, Virginia. Subjective evaluation of picture quality was performed at the Advanced Television Evaluation Laboratory (ATEL) in Ottawa, Canada. In February 1993, a special panel of experts reviewed the test results of the five HDTV system proposals and made a recommendation to the ACATS.

The panel concluded that the four digital systems performed substantially better than the analog system. The panel also concluded that each of the four digital systems excelled in different aspects. Therefore, the panel could not recommend one particular system. The panel recommended that each digital system be retested after improvements were made by the proponents. The four digital proponents had stated earlier that substantial improvements could be made to their respective system. The ACATS accepted the panel's recommendation and decided to retest the four systems after improvements were made. As an alternative to the retest, the ACATS encouraged the four proponents to combine the best elements of the different systems and submit one single system for evaluation.

The four digital system proponents evaluated their options and decided to submit a single system. In May 1993, they formed a consortium called the Grand Alliance to design and construct an HDTV prototype system. The Grand Alliance consisted of seven organizations who were members at the inception of the Grand Alliance. Later, some member organizations changed their names: (1) *General Instrument* first proposed digital transmission of an HDTV signal and submitted one of the four initial systems; *Massachusetts Institute of Technology* submitted a system together with General Instrument; *AT&T and Zenith* submitted one system together; and *Philips*, the *David Sarnoff Research Center*, and *Thomson Consumer Electronics* submitted one system together. Between the years 1993 and 1995, the Grand Alliance chose the best technical elements from the four systems and made further improvements on them. The Grand Alliance HDTV system was submitted to the ATTC and ATEL for performance verification. Test results verified that the Grand Alliance system performed better than the previous four digital systems. A technical standard based on the Grand Alliance HDTV prototype system was documented by the Advanced Television System Committee (ATSC), an industry consortium.

The HDTV prototype proposed by the Grand Alliance was a flexible system that carried approximately 20 million bits per second (20 Mbps). Even though it used the available bit capacity to transmit one HDTV program, the bit capacity could also be used to transmit several programs of standard definition television (SDTV) or other digital data such as stock quotes. SDTV resolution is comparable to that of the NTSC, but it is substantially less than the HDTV. The documented technical standard (known as the ATSC standard) allowed the transmission of SDTV programs as well as HDTV programs.

In November 1995, the ACATS recommended the ATSC standard as the United States advanced television standard to the FCC. The ATSC standard had allowed a set of only 18 video resolution formats for HDTV and SDTV programs. The FCC eased this restriction in December 1996, and decided that the ATSC standard with a relaxation of the requirements for video resolution format would be the United States digital television standard. In early 1997, the FCC made additional rulings to support the new technical standard, such as channel allocation for digital television service.

3. AN HDTV SYSTEM

A block diagram of a typical HDTV system is shown in Fig. 3. The information transmitted includes video, audio, and other auxiliary data such as stock quotes. The input video source may have a format (spatial resolution, temporal resolution, or scanning format) different from the formats used or preferred by the video encoder. In this case, the input video format is converted to a format used or preferred by the video encoder. The video is then compressed by a video encoder. Compression is needed because the bit rate supported by the modulation system is typically much less than the bit rate needed for digital video without compression. The audio, which may be multichannel for one video program, is also compressed. Since the bit rate required for audio is much less than that for video, the need to compress the audio is not as crucial. Any bit-rate savings, however, can be used for additional bits for video or other auxiliary data. The data may represent any digital data, including additional information for video and audio. The compressed video data, compressed audio data, and any other data are multiplexed by a transport system. The resulting bitstream is modulated. The modulated signal is then transmitted over a communication channel.

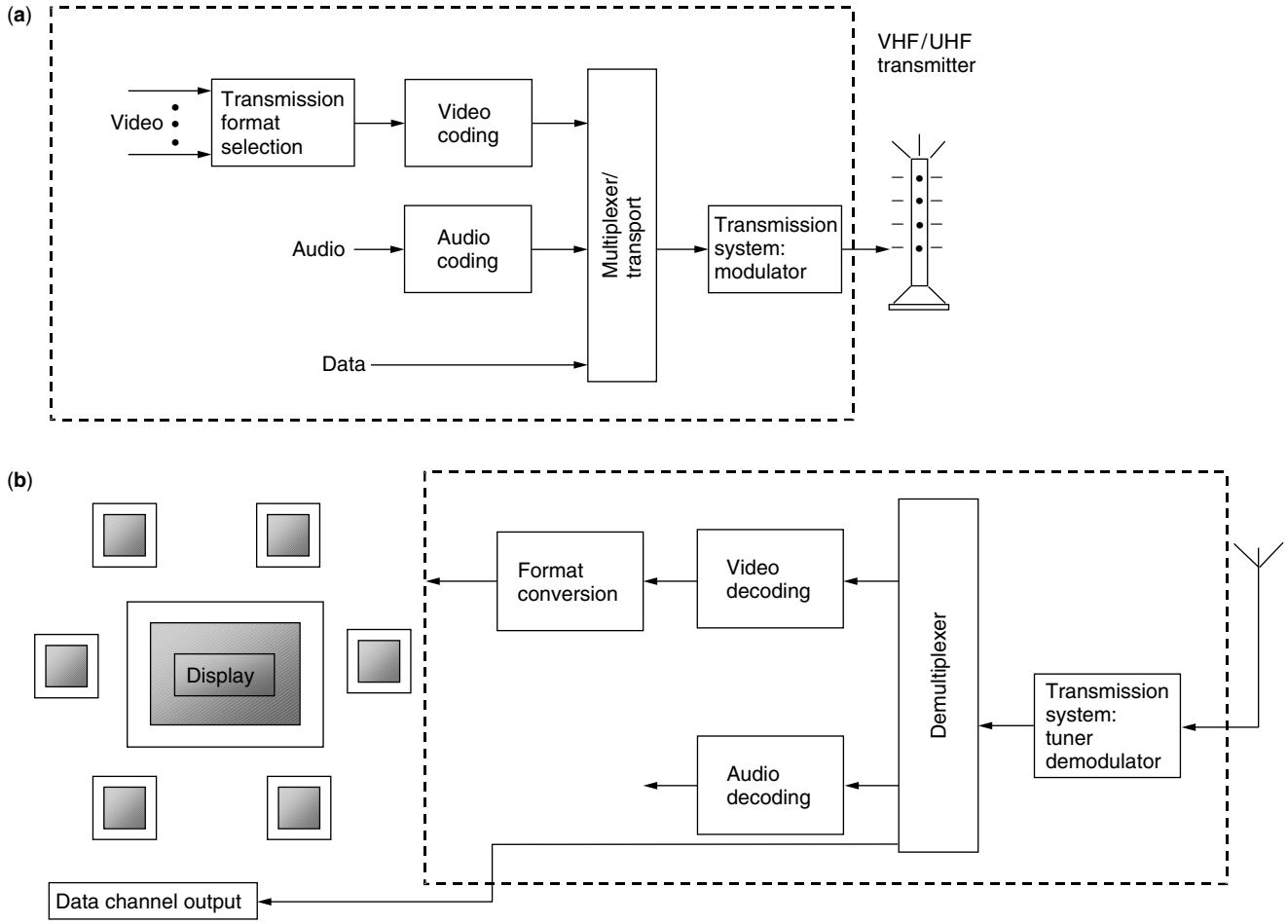


Figure 3. A block diagram of a typical HDTV system: (**a**) transmitter; (**b**) receiver.

At the receiver, the received signal is demodulated to generate a bitstream, which is demultiplexed to produce compressed video, compressed audio, and other data. The compressed video is then decompressed. The video format received may not be the same as the format used in the display. In this case, the received video format is converted to the proper display format. The compressed audio, which may be multichannel, is decompressed and distributed to different speakers. The use of the data received depends on the type of information the data contains. The communication channel in Fig. 3 may represent a storage device such as digital video disk. If the available bit rate can support more than one video program, multiple video programs can be transmitted.

There are many different possibilities for the design of an HDTV system. For example, various methods can be used for video compression, audio compression, and modulation. Some modulation methods may be more suitable for terrestrial transmission, while others may be more suitable for satellite transmission. Among the many possibilities, this article will focus on the Grand Alliance HDTV system. This system was designed over many years of industry competition and cooperation. The system's performance was carefully evaluated by laboratory and field tests, and was judged to be acceptable for its intended application. The system was the basis for the United States digital television standard. Even though this article focuses on one system, many issues and design considerations encountered in the Grand Alliance HDTV system could be applied to any HDTV system.

The overall Grand Alliance HDTV system consists of five elements: transmission format selection, video coding, audio coding, multiplexing, and modulation. These are described in the following sections.

3.1. Transmission Format Selection

A television system accommodates many video input sources such as videocameras, film, magnetic and optical media, and synthetic imagery. Even though these different input sources have different video formats, a conventional television system such as the NTSC uses only one single transmission format. This means the various input sources are converted to one format and then transmitted. Using one format simplifies the receiver design because a receiver can eliminate format conversion by designating the display format to be the same as the transmission format. This is shown in Fig. 4. When the NTSC system was standardized in the 1940s and 1950s, format conversion would have been costly.

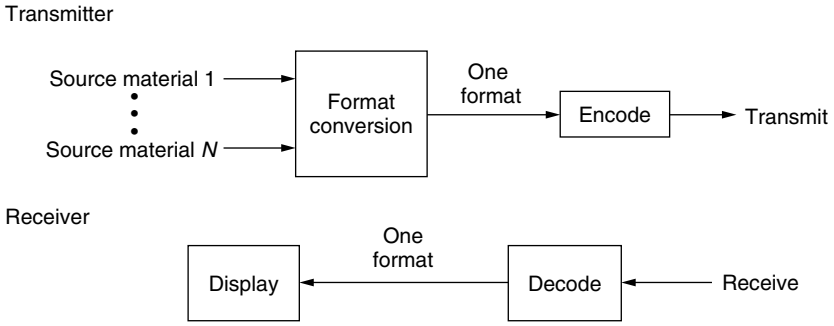


Figure 4. A television system with one single video transmission format.

The disadvantage of using one transmission format is the inefficient use of the available spectrum, since all the video input sources must be converted to one format and then transmitted in that format. For example, in the NTSC system, film (whose native format is 24 frames/s with progressive scanning) is converted to 60 fields/sec with interlaced scanning. It is then transmitted in the NTSC format. Transmission of video in a format other than its native format is an inefficient use of the spectrum.

The Grand Alliance HDTV system utilizes multiple transmission formats. This allows the use of a video transmission format that is identical to or approximates the native video source format. In addition, the system allows the use of different formats for various applications. From the viewpoint of spectrum efficiency, allowing all possible video formats would be ideal. Since a display (such as a CRT) has typically one display format, the different formats received must be converted to one display format, as shown in Fig. 5. Allowing for too many formats can complicate the format conversion operation. In addition, most of the benefits derived from multiple formats can be obtained by carefully selecting a small set of formats. For HDTV applications, the Grand Alliance system utilizes six transmission formats as shown in Table 1. In the table, the spatial resolution of $C \times D$ means C lines of vertical resolution with D pixels of horizontal resolution. The scanning format is either a progressive scan or an interlaced scan format. The "frame/field rate" refers to the number of frames/s for progressive scan and the number of fields/s for interlaced scan.

Table 1. HDTV Transmission Formats Used in the Grand Alliance System

Spatial Resolution	Scanning Format	Frame/Field Rate (Frames/s)
720×1280	Progressive scanning	60
720×1280	Progressive scanning	30
720×1280	Progressive scanning	24
1080×1920	Progressive scanning	30
1080×1920	Progressive scanning	24
1080×1920	Interlaced scanning	60

The Grand Alliance system utilizes both 720 lines and 1080 lines. The number of pixels per line was chosen so that the aspect ratio (width-to-height ratio) is 16×9 with square pixels. When the spatial vertical dimension that corresponds to one line equals the spatial horizontal dimension that corresponds to one pixel, it is called "square pixel." For 720 lines, the scanning format is progressive. The highest frame rate is 60 frames/s. The pixel rate is approximately 55 Mpixels/s (million pixels per second). For the video compression and modulation technologies used, a substantial increase in the pixel rate above 60–70 Mpixels/s may result in a noticeable degradation in video quality. At 60 frames/s with progressive scanning, the temporal resolution is very high, and smooth motion rendition results. This format is useful for sports events and commercials. The 720-line format also allows the temporal resolution of 30 frames/s and 24 frames/s. These frame rates were

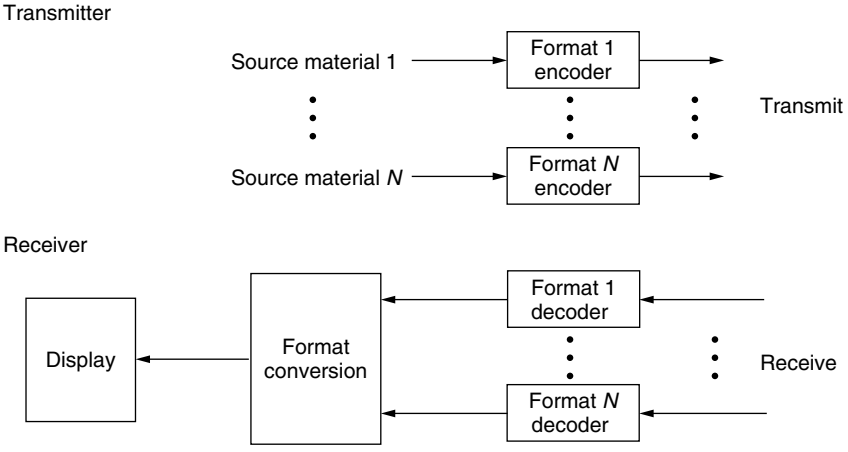


Figure 5. A television system with multiple video transmission formats.

chosen to accommodate film and graphics. For film, whose native format is 24 frames/s, conversion to 60 frames/s and then compressing it will result in a substantial inefficiency in spectrum utilization. For 720 lines at 24 frames/s, it is possible to simultaneously transmit two high-resolution video programs within a 6-MHz channel because of the lower pixel rate (approximately 22 Mpixels/s each).

In the 1080-line format, two temporal rates in progressive scan are 30 and 24 frames/s. These temporal rates were chosen for film and graphics with the highest spatial resolution. Another temporal rate used is the 1080-line interlaced scan at 60 fields/s. This is the only interlaced scan HDTV format used in the Grand Alliance system. It is useful for scenes obtained with a 1080-line interlaced-scan camera. The pixel rate for 1080-line progressive scan at 60 frames/s would be more than 120 Mpixels/s. The video encoded for such a high pixel rate can result in a substantial degradation in video quality for the compression and modulation technologies used in the Grand Alliance system. Therefore, there is no 1080-line, 60-frame/s progressive scan format in the system.

All conventional television systems, such as the NTSC, utilize only interlaced scanning. In such systems, a display format is matched to the single-transmission format. The display requires at least a 50–60 Hz rate with a reasonable amount of spatial resolution (approximately 480 active lines for the NTSC system). An alternative strategy in the NTSC system would be to preserve 480 lines with progressive scan, but at 30 frames/s. To avoid display flicker, each frame can be repeated twice at the display, making the display rate 60 Hz. Repetition of a frame at the receiver would require frame memory, which was not possible with the technologies available when the NTSC system was standardized. Because of the exclusive use of interlaced scanning in conventional television systems, early HDTV video equipment, such as videocameras, was developed for interlaced scanning.

An interlaced display has video artifacts such as interline flicker. Consider a sharp horizontal line that is in the odd field, but not in the even field. Even though the overall large area flicker rate is 60 Hz, the flicker rate for the sharp horizontal line is only 30 Hz. As a result, the line flickers in a phenomenon called *interline flicker*. Interline flickers are particularly troublesome for computer graphics or written material that contains many sharp lines. Partly for this reason, almost all computer monitors use progressive scanning.

When a television system is used as a standalone entertainment device, its interoperability with computers is not a serious issue. For a digital HDTV system, however, it is no longer a standalone entertainment device, and its interoperability with computers and telecommunications networks is useful. When a display device uses progressive scan and an interlaced transmission format is used, a conversion process called *deinterlacing* must convert the interlaced scan format to a progressive scan format before it is displayed. A high-performance deinterlacer requires complex signal processing. Even when a high-performance deinterlacer is used, a progressive transmission format yields better performance than an interlaced transmission format for graphics, animation, and written material. For this and other reasons like simple processing, the computer industry preferred only the progressive transmission format for television. Other industries, such as television manufacturers and broadcasters, preferred interlaced scanning. This is because they were accustomed to interlaced scanning. Interlaced scanning worked well for entertainment material. Early video equipment, such as videocameras, was developed only for interlaced scanning. The Grand Alliance HDTV system used five progressive scan formats and one interlaced scan format. The FCC decision in December 1996 removed most of the restrictions on transmission formats and allowed both progressive and interlaced formats. This decision left the choice of transmission format to free market forces.

A multiple transmission format system utilizes the available spectrum more efficiently than does a single transmission format system by better accommodating video source materials with different native formats. In addition, multiple transmission formats can be used for various applications. Table 2 shows possible applications for the six Grand Alliance HDTV formats. For a multiple transmission format system, one of the allowed transmission formats is chosen for a given video program prior to video encoding. The specific choice depends on the native format of the input video material and its intended application. The same video program within a given format may be assigned to a different transmission format, depending on the time of broadcast. If the transmission format chosen is different from the native format of the video material, format conversion occurs.

3.2. Video Coding

The Grand Alliance HDTV system transmits at least one HDTV program in a single 6-MHz channel. For the modulation technology used in the Grand Alliance system the maximum bit rate available for video is approximately 19 Mbps. For a typical HDTV video input, the incompressed bit rate is on the order of 1 Gbps. This means the input video must be compressed by a factor of >50.

For example, consider an HDTV video input of 720×1280 with progressive scan at 60 frames/s. The pixel rate is 55.296 Mpixels/s. A color picture consists of three monochrome images: red, green, and blue. The red, green, and blue colors are three primary colors of an

Table 2. Applications of HDTV Transmission Formats Used in the Grand Alliance System

Format	Applications
720×1280, PS, 60 frames/s	Sports, concerts, animation, graphics, upconverted NTSC, commercials
720×1280, PS, 24 frames/s or 30 frames/s	Complex film scenes, graphics, animation
1080×1920, PS, 24 frames/s or 30 frames/s	Films with highest spatial resolution
1080×1920, IS, 60 fields/s	Scenes shot with an interlaced scan camera

additive color system. By mixing the appropriate amounts of red, green, and blue lights, many different color lights can be generated. By mixing a red light and a green light, for example, a yellow light can be generated. A pixel of a color picture consists of the red, green, and blue components. Each component is typically represented by 8 bits (256 levels) of quantization. For many video applications, such as television, 8 bits of quantization are considered sufficient to avoid video quality degradation by quantization. Each pixel is then represented by 24 bits. The bit rate for the video input of 720×1280 with progressive scan at 60 frames/s is approximately 1.3 Gbps. In this example, reducing the data rate to 19 Mbps requires video compression by a factor of 70.

Video compression is achieved by exploiting the redundancy in the video data and the limitations of the human visual system. For typical video, there is a considerable amount of redundancy. For example, much of the change between two consecutive frames is due to the motion of an object or the camera. Therefore, a considerable amount of similarity exists between two consecutive frames. Even within the same frame, the pixels in a neighborhood region typically do not vary randomly. By removing the redundancy, the same (redundant) information is not transmitted.

For television applications, the video is displayed for human viewers. Even though the human visual system has enormous capabilities, it has many limitations. For example, the human visual system does not perceive well the spatial details of fast-changing regions. The high spatial resolution in such cases does not need to be preserved. By removing the redundancy in the data and exploiting the limitations of the human visual system, many methods of digital video compression were developed. A digital video encoder usually consists of the three basic elements shown in Fig. 6. The first element is representation. This element maps the input video to a domain more suitable for subsequent quantization and codeword assignment. The quantization element assigns reconstruction (quantization) levels to the output of the representation element. The codeword assignment selects specific codewords (a string of zeros and ones) to the reconstruction levels. The three elements work together to reduce the required bit rate by removing the redundancy in the data and exploiting the limitations of the human visual system.

Many different methods exist for each of the three elements in the image coder. The Grand Alliance system utilizes a combination of video compression techniques that conform to the specifications of the MPEG-2 (Moving Pictures Expert Group) video compression standard. This is one of many possible approaches to video compression.

MPEG-2 Standard. The International Standard Organization (ISO) established the Moving Pictures Expert Group (MPEG) in 1988. Its mission was to develop video coding standards for moving pictures and associated audio. In 1991, the group developed the ISO standard 11172, called *coding of moving pictures and associated audio*. This standard, known as MPEG-1, is used for digital storage media at up to $\sim$1.5 Mbps.

In 1996, the group developed the ISO standard 13818 called *Generic Coding of Moving Pictures and Associated Audio*. This standard, known as MPEG-2, is an extension of MPEG-1 that allows flexibility in the input format and bit rates. The MPEG-2 standard specifies only the syntax of the coded bitstream and the decoding process. This means that there is some flexibility in the encoder. As long as the encoder generates a bitstream that is consistent with the MPEG-2 bitstream syntax and the MPEG-2 decoding process, it is considered a "valid" encoder. Since many methods of generating the coded bitstream are consistent with the syntax and the decoding process, some optimizations and improvements can be made without changing the standard. The Grand Alliance HDTV system uses some compression methods included in MPEG-2 to generate a bitstream that conforms to the MPEG-2 syntax. An MPEG-2 decoder can decode a video bitstream generated by the Grand Alliance video coder.

3.3. Audio Processing. The Grand Alliance HDTV system compresses the audio signal to efficiently use the available spectrum. To reconstruct CD-quality audio after compression, the compression factor for audio is substantially less than that of the video. High-quality audio is important for HDTV viewing. The data rate for audio is inherently much lower than that for video, and the additional bit-rate efficiency that can be achieved at the expense of audio quality is not worthwhile for HDTV applications.

Consider one channel of audio. The human auditory system is not sensitive to frequencies above 20 kHz. The audio signal sampled at a 48-kHz rate is sufficient to ensure that the audio information up to 20 kHz is preserved. Each audio sample is typically quantized at 16 bits/sample. The total bit rate for one channel of audio input is 0.768 Mbps. Exploiting the limitations of the human auditory system, the bit-rate requirement is reduced to 0.128 Mbps, with the reproduced audio quality almost indistinguishable from that of the input audio. The compression factor achieved is 6, which is substantially less than the video's compression factor of >50. In the case of video, it is necessary to obtain a very high compression factor, even at the expense of some noticeable quality degradation for difficult scenes because of the very high-input video bit rate (>1 Gbps). In the case of audio, additional bit-rate savings from 0.128 Mbps at the expense of possible audio quality degradation is not considered worthwhile for HDTV applications. Similar to video compression, audio compression utilizes reduction of redundancy in the audio data and the limitations of the human auditory system.

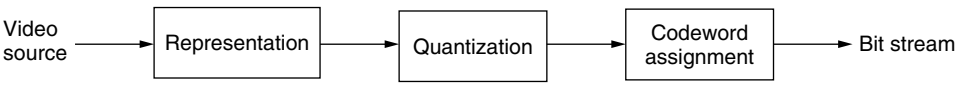


Figure 6. Three basic elements of a video encoder.

The Grand Alliance HDTV system uses a modular approach to the overall system design. Various technologies needed for a complete HDTV system can be chosen independently from each other. The audio compression method, for example, can be chosen independent from the video compression method. Even though the MPEG-2 standard used for video compression in the Grand Alliance system includes an audio compression method, the Grand Alliance selected the Audio Coder 3 (AC-3) standard on the basis of several factors including performance, bit-rate requirement, and cost.

The Grand Alliance system can encode a maximum of six audio channels per audio program. The channelization follows the ITU-R recommendation BS-775: *Multichannel Stereophonic Sound System with and without Accompanying Picture*. The six audio channels are left, center, right, left surround, right surround, and low-frequency enhancement. The bandwidth of the low-frequency enhancement channel extends to 120 Hz, while the other five channels extend to 20 kHz. The six audio channels are also referred to as "5.1 channels." Since the six channels are not completely independent for a given audio program, this dependence can be exploited to reduce the bit rate. The Grand Alliance system encodes 5.1 channel audio at a bit rate of 384 kbps with audio quality essentially the same as that of the original.

3.4. Transport System

The bitstreams generated by the video and audio encoders and the data channel must be multiplexed in an organized manner so that the receiver can demultiplex them efficiently. This is the main function of the transport system. The Grand Alliance system uses a transport format that conforms to the MPEG-2 system standard, but it does not utilize all of its capabilities. This means that the Grand Alliance decoder cannot decode an arbitrary MPEG-2 systems bitstream, but all MPEG-2 decoders can decode the Grand Alliance bitstream.

The bitstream that results from a particular application such as video, audio, or data is called an *elementary bitstream*. The elementary bitstreams transmitted in a 6-MHz channel are multiplexed to form the program transport bitstream. Each elementary bitstream has a unique program identification (PID) number, and all the elementary bitstreams within a program transport bitstream have a common timebase.

An example of the multiplex function used to form a program transport stream is shown in Fig. 7. The first two elementary streams are from one television program. The next two elementary streams are from another television program. As long as the available bit rate for the channel can accommodate more than one television program, the transport system will accommodate them. The fifth elementary stream is only an audio program without the corresponding video. The next two elementary streams are from two different datastreams. The last elementary stream contains the control information, which includes a program table that lists all the elementary bit streams, their PIDs, and the applications such as video, audio, or data. All eight elementary streams that form the program transport bitstream have the same timebase.

The Grand Alliance transport system uses the fixed-length packet structure shown in Fig. 8. Each packet consists of 188 bytes, which is divided into a header field and payload. The header field contains overhead information, and the payload contains the actual data that must be transmitted. The size of the packet is chosen to ensure that the actual payload-to-overhead ratio is sufficiently high and that a packet lost during

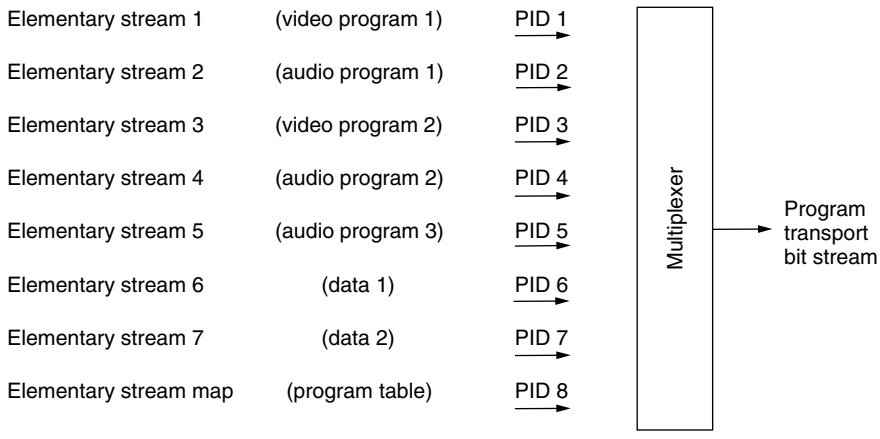


Figure 7. An example of the multiplex function used to form a program transport bitstream.

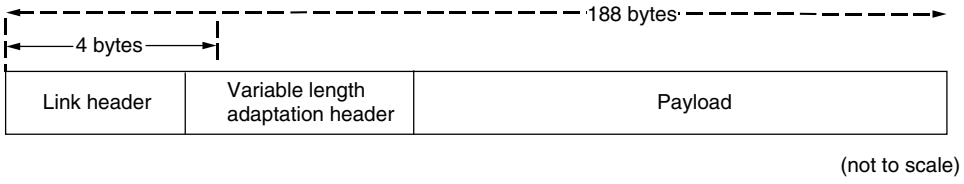


Figure 8. Fixed-length packet structure used in the Grand Alliance transport system.

transmission will not have serious consequences on the received video, audio, and data. The payload of each packet contains bits from only one particular elementary bitstream. For example, a packet cannot have bits from both a video elementary bitstream and an audio elementary bitstream.

Information that is not actual data but is important or useful for decoding the received bitstream is contained in the header. The header of each packet includes a 4-byte link header and may also include a variable-length adaptation header when needed. The link header includes 1-byte synchronization to indicate the beginning of each packet, a 13-bit PID to identify which elementary stream is contained in the packet, and information as to whether the adaptation header is included in the packet. The adaptation header contains timing information to synchronize decoding and a presentation of applications such as video and audio. This information can be inserted into a selected set of packets. The adaptation header also contains information that facilitates random entry into application bitstreams in order to support functions such as program acquisition and change.

On the transmitter side, the bits from each elementary stream are divided into packets that contain the same PID. The packets are multiplexed to form the program transport bitstream. An example is shown in Fig. 9. At the receiver, the synchronization byte, which is the first byte in each packet, is used to identify the beginning of each packet. From the program table contained in the control packet, information can be obtained on which elementary streams are in the received program transport bitstream. This information, together with the PID in each packet, is used to separate the packets into different elementary bitstreams. The information contained in the adaptation header in a selected set of packets is used for the timing and synchronization of the decoding and for the presentation of different applications (video, audio, data, etc.).

The transport system used in the Grand Alliance system has many advantages. The system is very flexible in dynamically allocating the available channel capacity to video, audio, and data. The system can devote all available bits to video, audio, or data, or any combination thereof. The system also can allocate available bits to more than one television program. If video resolution is not high, several standard definition television programs (comparable to the NTSC resolution) can be transmitted. This is in sharp contrast with the NTSC system, where a fixed bandwidth is allocated to one video program and a fixed bandwidth is allocated to audio. The capability to dynamically allocate bits as the need arises is a major feature of the transport system.

The transport system is also scalable. If a higher-bit-rate channel is available, the same transport system can be used by simply adding elementary bitstreams. The system is also extensible. If future services become available, such as 3D television, they can be added as an elementary stream with a new PID. Existing receivers that do not recognize the new PID will ignore the new elementary stream. New receivers will recognize the new PID.

The transport system is also robust in terms of transmission errors, and is amenable to cost-effective implementation. The detection and correction of transmission errors can be synchronized easily because of the fixed-length packet structure; this structure also facilitates simple demultiplex designs for low-cost, high-speed implementation.

3.5. Transmission System

The bitstream generated by the transport system must be processed in preparation for modulation, and then modulated for transmission. Choosing from the many modulation methods depends on several factors, including the transmission medium and the specific application. The best method for terrestrial broadcasting may not be the best for satellite broadcasting. Even for terrestrial broadcasting, the use of taboo channels means that interference with existing NTSC channels must be considered to determine the specific modulation technology. Considering several factors, such as coverage area, available bit rate, and complexity, the Grand Alliance system uses an 8-VSB system for terrestrial broadcasting. A block diagram of the 8-VSB system is shown in Fig. 10.

3.5.1. Data Processing. The data processing part of the 8-VSB system consists of a data randomizer, a Reed–Solomon encoder, a data interleaver, and trellis encoder. Data packets of 188 bytes per packet are received from the transport system and randomized. Portions of the bitstream from the transport system

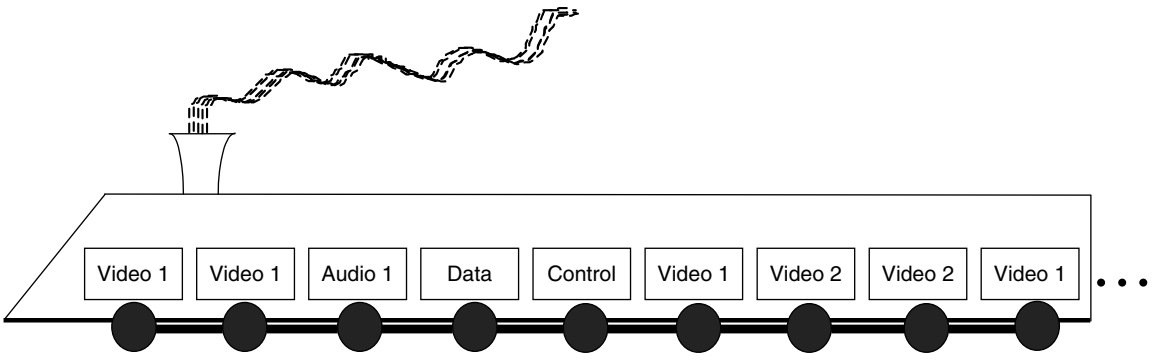


Figure 9. Example of packet multiplexing to form the program transport bitstream.

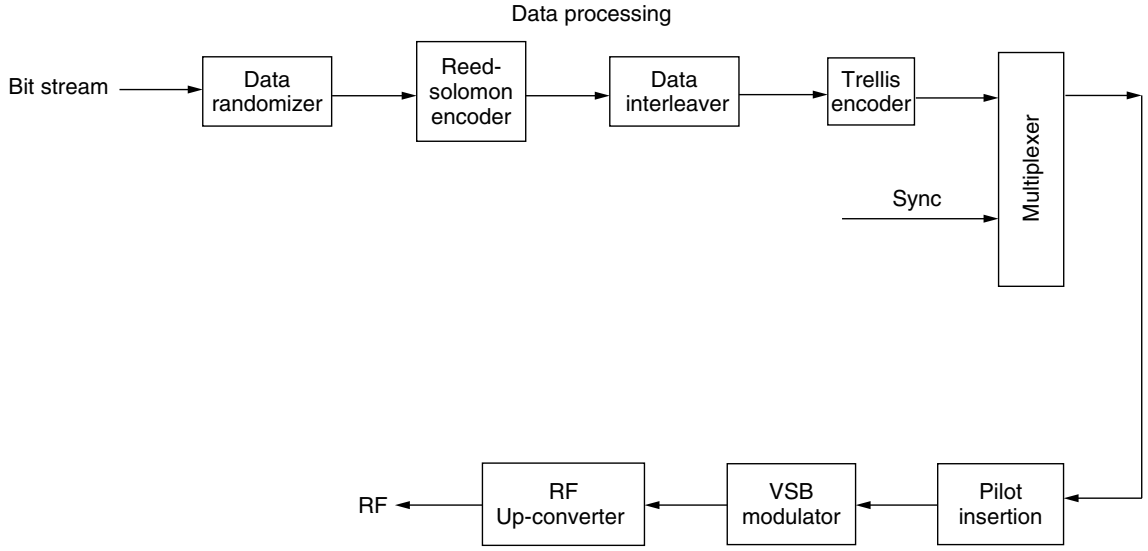


Figure 10. A block diagram of an 8-VSB system used in the Grand Alliance HDTV system.

may have some pattern and may not be completely random. Randomization ensures that the spectrum of the transmitted signal is flat and is used efficiently. In addition, the receiver exhibits optimal performance when the spectrum is flat.

The transmission channel introduces various types of noise such as random noise and multipath. They manifest themselves as random and bursty bit errors. To handle these bit errors, two forward error correction (FEC) schemes are used. These FEC schemes add redundancy to the bitstream. Errors are then detected and corrected by exploiting this added redundancy. Even though error correction bits use some of the available bit rate, they increase overall performance of the system by detecting and correcting errors.

The first error correction method is the Reed–Solomon code, known for its burst noise correction capability and efficiency in overhead bits. The burst noise correction capability is particularly useful when the Reed–Solomon code is used with the trellis code. The trellis code is effective in combating random and short impulsive noise, but it tends to generate burst errors in the presence of strong interference and burst noise. The Reed–Solomon code used in the 8-VSB system adds approximately 10% of overhead to each packet of data. The result is a data segment that consists of a packet from the transport system and the Reed–Solomon code bits. The resulting bytes are convolutionally interleaved over many data segments. The convolutional interleaving that is useful in combating the effects of burst noise and interference is part of the trellis encoding. The trellis encoder, which is a powerful technique for correcting random and short-burst bit errors, adds additional redundancy. In the 8-VSB system, the trellis coder creates a 3-bit symbol (eight levels) from a 2-bit data symbol.

At the transmitter, the Reed–Solomon encoder precedes the trellis encoder. At the receiver, the trellis decoder precedes the Reed–Solomon decoder. The trellis decoder is effective in combating the random and short-burst bit errors, but it can create long-burst bit errors in the presence of strong interference and bursts. The Reed–Solomon code is effective in combating long-burst bit errors.

The 8-VSB system transmits approximately 10.76 million symbols per second (SPS), with each symbol representing 3 bits (eight levels). When accounting for the overhead associated with the trellis coder, the Reed–Solomon coder, and additional synchronization bytes, the bit rate available to the transport system's decoder is approximately 19.4 Mbps. The bit rate is not only for applications such as video, audio, and data, but also for overhead information (link header, etc.) in the 188-byte packets. The actual bit rate available for the applications is less than 19.4 Mbps. The VSB system can deliver a higher bit rate, for example, by reducing the redundancy in the error correction codes and by increasing the number of levels per symbol. However, the result is loss of performance in other aspects such as the coverage area. The 8-VSB system delivers the 19.4-Mbps bit rate to ensure that an HDTV program can be delivered within a 6-MHz channel with a coverage area at least comparable to an NTSC system, and with an average power level below the NTSC power level in order to reduce interference with the NTSC signals.

The trellis encoding results in data segments. At the beginning of each set of 312 data segments, a data segment is inserted that contains the synchronization information for the set of 312 data segments. This data segment also contains a training sequence that can be used for channel equalization at the receiver. Linear distortion in the channel can be accounted for by an equalizer at the receiver. At the beginning of each data segment, a 4-symbol synchronization signal is inserted. The data segment sync and the 312-segment set sync are not affected by the trellis encoder and can provide synchronization independent of the data.

3.5.2. Pilot Insertion. Prior to modulation, a small pilot is inserted in the lower band within the 6-MHz band. The location of the pilot is on the Nyquist slope of NTSC receivers. This ensures that the pilot does not seriously impair existing NTSC service. The channel assigned for the HDTV service may be a taboo channel that is currently unused because of cochannel interference with an existing NTSC service located some distance away. The HDTV signal must be designed to ensure that its effect on existing service is minimal.

The NTSC system is rugged and reliable. The main reason for this is the use of additional signals for synchronization that do not depend on the video signals. The NTSC receiver reliably synchronizes at noise levels well below the loss of pictures. In the 8-VSB system, a similar approach is taken. The pilot signal that does not depend on the data is used for carrier acquisition. In addition, the data segment sync is used to synchronize the data clock for both frequency and phase. The 312-segment-set sync is used to synchronize the 312 segment-set and equalizer training. Reliance on additional signals for carrier acquisition and clock recovery is very useful. Even when occasional noise in the field causes a temporary loss of data, a quick recovery is possible as long as the carrier acquisition and clock recovery remain locked during the data loss. The 8-VSB system ensures that carrier acquisition and clock recovery remain intact well below the threshold level of data loss by relying on additional signals for such functions.

3.5.3. Modulation. For transmission of the prepared bitstream (message) over the air, the bitstream must be mapped to a bandpass signal that occupies a 6-MHz channel allocated to a station's HDTV service. A modulator modulates a carrier wave according to the prepared bitstream. The result is a bandpass signal that occupies a given 6-MHz channel. The 8-VSB system modulates the signal onto an IF (intermediate-frequency) carrier, which is the same frequency for all channels. It is followed by an upconversion to the desired HDTV channel.

3.5.4. Receiver. At the receiver, the signal is processed in order to obtain the data segments. One feature of the receiver is the NTSC rejection filter, which is useful because of the way HDTV service is being introduced in the United States. In some locations, an HDTV channel occupies the same frequency band as an existing NTSC channel located some distance away. The interference of the HDTV signal with the NTSC channel is minimized by a very-low-power level of the HDTV signal. The low-power level was made possible for HDTV service because of its efficient use of the spectrum. To reduce interference between the NTSC signal and the HDTV channel, the 8-VSB receiver contains an NTSC rejection filter. This is a simple comb filter whose rejection null frequencies are close to the video carrier, chroma carrier, and audio carrier frequencies of the NTSC signal. The comb filter, which reduces the overall performance of the system, can be activated only when a strong cochannel NTSC signal interferes with the HDTV channel.

3.5.5. Cliff Effect. Although additive white Gaussian noise does not represent typical channel noise, it is often used to characterize the robustness of a digital communication system. The segment error probability for the 8-VSB system in the presence of additive white Gaussian noise is shown in Fig. 11. At the signal-to-noise ratio (S/N) of 14.9 dB, the segment error probability is 1.93×10^{-4} or 2.5 segment errors/sec. At this threshold the segment errors become visible. Thus, up to an S/N of 14.9 dB, the system is perfect. At an S/N of 14 dB, which is just 0.9 dB less than the threshold of visibility (ToV), practically all segments are in error. This means a system that operates perfectly above the threshold of visibility becomes unusable when the signal level decreases by 1 dB or when the noise level increases by 1 dB. This is known as the "cliff effect" in a digital communication system. In the case of the NTSC system, the picture quality decreases gradually as the S/N decreases. To avoid operating near the cliff region, a digital system is designed to operate well above the threshold region within the intended coverage area. Both laboratory and field tests have indicated that the coverage area of the HDTV channel is comparable to or greater than the NTSC

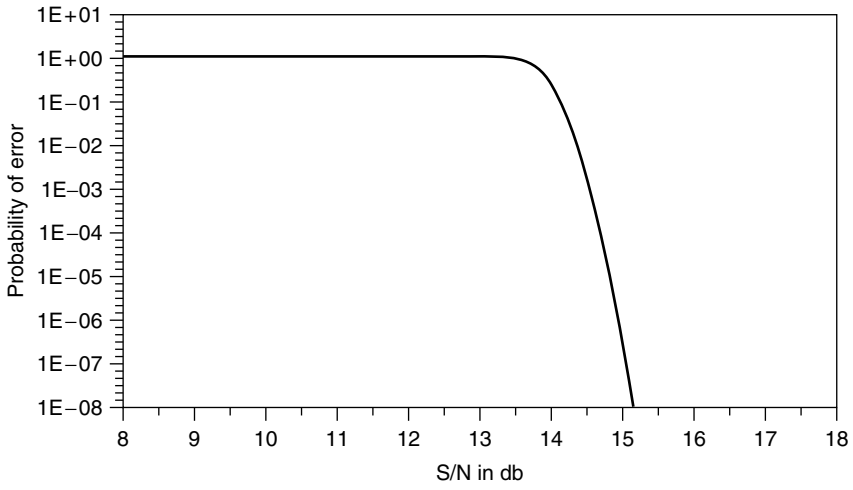


Figure 11. Segment error probability for the 8-VSB system in the presence of additive white Gaussian noise.

channel, despite the substantially lower power level used for the HDTV signal.

3.5.6. Cable Mode. A cable environment introduces substantially less channel impairment than does terrestrial broadcasting for a variety of reasons. In the case of cable, for example, cochannel interference from an existing NTSC station is not an issue. This can be exploited to deliver a higher bit rate for applications in a 6-MHz cable channel. Although the Grand Alliance HDTV system was designed for terrestrial broadcasting, the system includes a cable mode that doubles the available bit rate using a small modification. Doubling the bit rate means the ability to transmit two HDTV programs within a single 6-MHz cable channel.

In order to double the bit rate for cable mode, the Grand Alliance System uses 16-VSB rather than 8-VSB. All other aspects of the system such as video compression, audio compression, and transport remain the same. In the 8-VSB system, a symbol is represented by eight levels or 3 bits. One of the 3 bits is due to the redundancy created by trellis encoding. For a cable environment, error correction from the powerful trellis coding is no longer needed. In addition, the higher available S/N for cable means that a symbol can be represented by 16 levels or 4 bits. Since the symbol rate remains the same between the 8-VSB and 16-VSB systems, the available bit rate for the 16-VSB system doubles in comparison with the 8-VSB system. For the 16-VSB system without trellis coding, the S/N ratio for the segment error rate that corresponds to the threshold of visibility in the environment of additive white Gaussian noise is approximately 28 dB. This is 13 dB higher than the 8-VSB system with the trellis coding. This increase in the S/N is acceptable in a cable environment that has substantially less channel impairments than a typical terrestrial environment.

4. HDTV AND INTEROPERABILITY

The Grand Alliance HDTV system has served as the basis for the digital television standard in the United States. The standard itself defines a significant number of technical elements. The technologies involved, however, will continue to develop without the need to modify the standard. For example, the video compression system that was adopted defines syntax for only the decoder. There is much room for improvement and advances in the encoder. Technical elements can also be added in a backward-compatible manner. The transmission of a very high-definition (VHD) television format, which was not provided in the initial standard, can be accomplished in a backward-compatible manner. This can be achieved by standardizing a method to transmit enhancement data. This, in turn can be combined with an allowed video transmission format to deliver the VHD format.

The Grand Alliance system was designed for HDTV delivery in terrestrial environments in the United States. For other delivery environments such as satellite, cable, and environments in other parts of the world, other standards will emerge. Depending on the degree of common elements, interoperability among different standards will become an important issue. In terms of program exchange for different delivery media and throughout the world, technologies that convert one standard to another will continue to be developed. Efforts to facilitate the conversion process by adopting common elements among the different standards also will continue.

Interoperability will be an issue not only among the different HDTV standards, but among telecommunication services and computers as well. A traditional television system has been a standalone device whose primary purpose was entertainment. Although an HDTV system is used for entertainment, it can be an integral part of a home center for entertainment, telecommunications, and information. The HDTV display can be used as a videophone, a newspaper service, or a computer display. Interoperability between an HDTV system and other services that will be integrated in the future is an important consideration.

BIBLIOGRAPHY

1. ATSC, *Digital Audio Compression* (*AC-3*), Dec. 20, 1995.
2. ATSC, *Digital Television Standard*, Sept. 16, 1995.
3. ATSC, *Guide to the Use of the ATSC Digital Television Standard*, Oct. 4, 1995.
4. V. Bhaskaran and K. Konstaninides, *Image and Video Compression Standards and Architectures*, Kluwer Academic Publishers, 1995.
5. W. Bretl, G. Sgrignoli, and P. Snopko, *VSB Modem Subsystem Design for Grand Alliance Digital Television Receivers*, ICCE, 1995.
6. A. B. Carlson, *Communication Systems*, 3rd ed., McGraw-Hill, 1986.
7. The Grand Alliance Members, The U.S. HDTV standard. The Grand Alliance, *IEEE Spectrum* 36–45 (April 1995).
8. ISO/IEC JTC1 CD 11172, *Coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbits/s*, International Organization for Standardization (ISO), 1992.
9. ISO/IEC JTC1 CD 13818, *Generic Coding of Moving Pictures and Associated Audio*, International Organization for Standardization (ISO), 1994.
10. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
11. J. S. Lim, *Two-Dimensional Signal and Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1990.
12. A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression*, Plenum Press, New York, 1988.
13. C. C. Todd et al., *AC-3: Flexible Perceptual Coding for Audio Transmission and Storage*, Audio Engineering Society Convention, Amsterdam, Feb. 28, 1994.
14. J. S. Lim, Digital television: Here at last, *Sci. Am.* 78–83 (May 1998).
15. B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to MPEG-2*, Digital Multimedia Standard Series, Chapman and Hill, New York, 1997.

16. J. L. Mitchell, W. B. Pennebaker, C. E. Fogg, and D. J. LeGall, *MPEG Video Compression Standard*, Digital Multimedia Standard Series, Chapman and Hill, New York, 1997.

17. T. Sikora, MPEG digital video-coding standards, *IEEE Signal Process. Mag.* **14**(5): 82–100 (Sept. 1997).

HIGH-RATE PUNCTURED CONVOLUTIONAL CODES

DAVID HACCOUN
Ecole Polytechnique de Montréal
Montréal, Quebec, Canada

1. INTRODUCTION

For discrete memoryless channels where the noise is essentially white (such as the space and satellite channels), error control coding systems using convolutional encoding and probabilistic decoding are among the most attractive means of approaching the reliability of communication predicted by the Shannon theory; these systems provide substantial coding gains while being readily implementable [1–3].

By far, error control techniques using convolutional codes have been dominated by low-rate $R = 1/v$ codes [1–3,11,16,19]. Optimal low-rate codes providing large coding gains are available in the literature [1,3,11,16,19], and practical implementations of powerful decoders using Viterbi, sequential, iterative, or Turbo decoding schemes exist for high data rates in tens of Mbits/s. However, as the trend for ever-increasing data transmission rates and high error performance continues while conserving bandwidth, the needs arise for good high-rate $R = b/v$ convolutional codes as well as practical encoding and decoding techniques for these codes. Unfortunately, a straightforward application of the usual decoding techniques for rates $R = 1/v$ codes to high-rate $R = b/v$ codes becomes very rapidly impractical as the coding rate increases. Furthermore, a conspicuous lack of good nonsystematic long-memory ($M > 9$) convolutional codes with rates R larger than 2/3 prevails in the literature.

With the advent of high-rate punctured convolutional codes [6], the inherent difficulties of coding and decoding of high-rate codes can be almost entirely circumvented. Decoding of rate $R = b/v$ punctured convolutional codes is hardly more complex than for rate $R = 1/v$ codes; furthermore, puncturing facilitates the implementation of rate-adaptive, variable-rate, and rate-compatible coding-decoding [3,6–9].

In this article, we present high-rate punctured convolutional codes especially suitable for Viterbi and sequential decoding. We provide the weight spectra and upper bounds on the error probability of the best-known punctured codes having memory $2 \le M \le 8$ and coding rates $2/3 \le R \le 7/8$ together with rate-2/3 and 3/4 long-memory punctured convolutional codes having $9 \le M \le 23$. All these codes are derived from the best-known rate-1/2 codes of the same memory lengths. The short memory codes are useful for Viterbi decoding, whereas the long memory codes are provided for archival purposes in addition to being suitable for sequential decoding.

We assume that the reader is familiar with the basic notions of convolutional encoding and the tree, trellis, and state diagram representations of convolutional codes. Without loss of generality, we consider binary convolutional codes of coding rate $R = b/v$, b and v integers with $1 \le b \le v$. A rate $R = b/v$ convolutional code produced by a b-input/ v-output encoder may also be denoted as a (v, b) convolutional code. The encoder is specified by a generating matrix $G(D)$ of dimension $b.v$ whose elements are the generator polynomials

$$g_{ij}(D) = \sum_{k=0}^{m_i} g_{ij}^k D^k = g_{ij}^0 + g_{ij}^1 D + \ldots + g_{ij}^{m_i} D^{m_i} \qquad (1)$$

where $i = 1, \ldots, b; j = 1, \ldots, b$.

The total memory of the encoder is $M = \sum_i m_i$. Hence, for low rate $R = 1/v$ codes, there are two branches emerging from each node in all the representations of the code, with two branches or paths remerging at each node or state of the encoder. For usual high rate $R = b/v$ codes, 2^b branches enter and leave each encoder state. As a consequence, compared to $R = 1/v$ codes, for $R = b/v$ codes the encoding complexity is multiplied by b, whereas the Viterbi decoding complexity is multiplied by 2^b. By using the notion of puncturing, these difficulties can be entirely circumvented, since regardless of the coding rate $R = (v - 1)/v$, the encoding or decoding procedure is hardly more complex than for the rate $R = 1/v$ codes.

The article is structured as follows. Section 2 introduces the basic concepts of encoding for punctured codes and their perforation patterns. Section 3 presents Viterbi decoding for punctured codes and their error performances. The search for good punctured codes is the objective of Section 4. This section presents extensive lists of the best-known punctured codes of coding rates varying from $R = 2/3$ to $R = 7/8$ together with their weight spectra and bit error probability bounds. Finally, the problem of generating punctured codes equivalent to the best-known usual nonpunctured high-rate codes is presented in Section 5. Again, extensive lists of long memory punctured equivalent to the best usual codes of the same rate are provided.

2. BASIC CONCEPTS OF PUNCTURED CONVOLUTIONAL CODES

2.1. Encoding of Punctured Codes

A punctured convolutional code is a high-rate code obtained by the periodic elimination (i.e., puncturing) of specific code symbols from the output of a low-rate

encoder. The resulting high-rate code depends on both the low-rate code, called the *original code* or *mother code*, and the number and specific positions of the punctured symbols. The pattern of punctured symbols is called the perforation pattern of the punctured code, and it is conveniently described in a matrix form called the *perforation matrix*.

Consider constructing a high-rate $R = b/v$ punctured convolutional code from a given original code of any low-rate $R = 1/v_o$. From every $v_o b$ code symbols corresponding to the encoding of b information bits by the original low-rate encoder, a number $S = (v_o b - v)$ symbols are deleted according to some specific perforation pattern. The resulting rate is then $R = b/(v_o b - S)$, which is equal to the desired target rate $R = b/v$. By a judicious choice of the original low-rate code and perforation pattern, any rate code may thus be obtained [6–9,21–24].

For example, Fig. 1 shows the trellis diagram of a rate-1/2, memory $M = 2$ code where every fourth symbol is punctured (indicated by X on every second branch on the diagram). Reading this trellis two branches (i.e., two information bits) at a time and redrawing it as in Fig. 2, we see that it corresponds to a rate-2/3, memory $M = 2$ code. A punctured rate-2/3 code has therefore been obtained from an original rate-1/2 encoder. The procedure can be generalized as described below.

2.2. Perforation Patterns

As shown in Fig. 3, an encoder for rate $R = b/v$ punctured codes may be visualized as consisting of an original low-rate $R = 1/v_o$ convolutional encoder followed by a symbol selector or sampler that deletes specific code symbols according to a given perforation pattern. The perforation pattern may be expressed as a perforation matrix $\mathbf{P}$ having v_o rows and b columns, with only binary elements 0s and 1s, corresponding to the deleting or keeping of the corresponding code symbol delivered by the original encoder, that is, for $i \in \{1, \ldots, v_o\}$, $j \in \{1, 2, \ldots, b\}$ the elements of $\mathbf{P}$ are

$$p_{ij} = \begin{cases} 0 & \text{if symbol } i \text{ of every } j\text{th branch is punctured} \\ 1 & \text{if symbol } i \text{ of every } j\text{th branch is not punctured} \end{cases} \tag{2}$$

Clearly, both the punctured code and its rate can be varied by suitably modifying the elements of the perforation matrix. For example, starting from an original

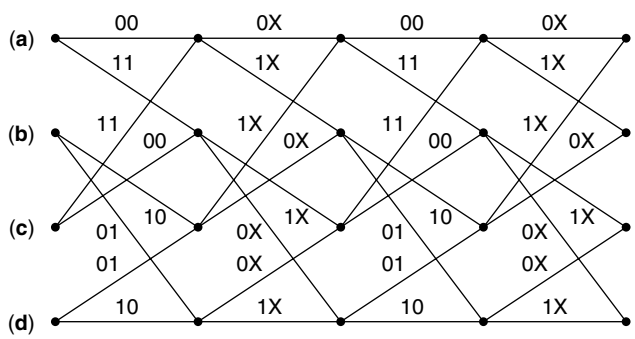


Figure 1. Trellis for $M = 2$, $R = 1/2$ convolutional code.

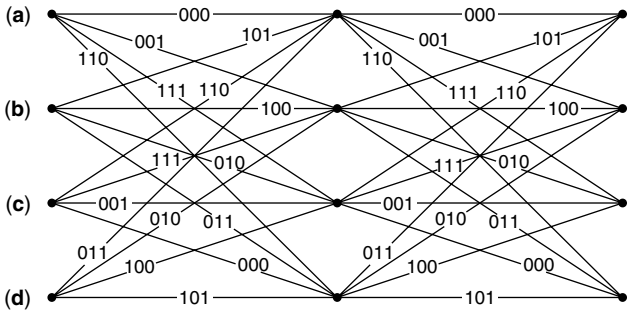


Figure 2. Trellis for punctured $M = 2$, $R = 2/3$ code.

rate-1/2 code, the perforation matrix of the rate-2/3 punctured code of Fig. 1 is given by:

$$\mathbf{P_1} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

whereas a rate-4/5 code could be obtained using the perforation matrix

$$\mathbf{P_2} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

Likewise, using an original $R = 1/3$ code, perforation matrix $\mathbf{P_3}$ also yields a punctured $R = 2/3$ code

$$\mathbf{P_3} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Variable-rate coding may be readily obtained if all punctured rates of interest are obtained from the same low-rate encoder where only the perforation matrices are modified accordingly, as illustrated by $\mathbf{P_1}$ and $\mathbf{P_2}$.

Variable-rate coding may be further specialized by adding the restriction that all the code symbols of the higher rate punctured codes are required by the lower rate punctured codes. This restriction implies minimal modifications of the perforation matrix as the coding rates vary. Punctured codes satisfying this restriction are said to be *rate-compatible*.

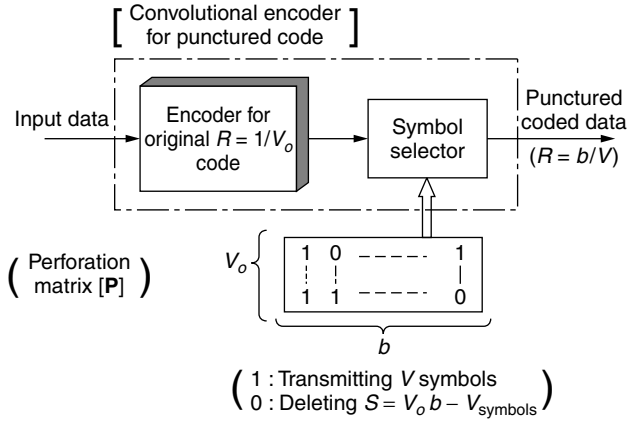


Figure 3. Schematic of an encoder for high rate punctured codes.

For example, using a mother code of rate $R = 1/2$, the sequence of perforation matrices $\mathbf{P_4}$, $\mathbf{P_5}$, and $\mathbf{P_6}$, yields the rate-compatible punctured codes with coding rates $R = 4/5$, 4/6, and 4/7, respectively,

$$\mathbf{P_4} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{P_5} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

$$\mathbf{P_6} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

Rate-compatible punctured codes are especially useful in some rate-adaptive ARQ/FEC applications since only the incremental redundancy must be transmitted as the coding rate is decreased. Families of good noncatastrophic short-memory rate-compatible punctured codes with rates varying from 8/9 to 1/4 have been found by Hagenauer [9].

Finally, another class of perforation patterns called *orthogonal perforation patterns* plays an important part in the generation of specific punctured codes [21,22]. An orthogonal perforation pattern is a pattern in which any code symbol that is not punctured on one of the b branches is punctured on all the other $(b-1)$ branches of the resulting rate-b/v punctured code. With an orthogonal perforation pattern, the perforation matrix has $v_o = v$ rows and b columns, with each row containing only one element 1.

For example the perforation patterns $\mathbf{P_7}$ is orthogonal

$$\mathbf{P_7} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and generates a $R = 3/4$ punctured code from a $R = 1/4$ original code.

Orthogonal perforation patterns ensure that all the different generators of the original low-rate $1/v_o$ code are used in deriving a desired punctured rate-b/v code. In particular, it can be shown that any punctured code can be obtained by means of an orthogonal perforation pattern. Using this concept, punctured codes strictly identical to the best-known usual rate-2/3 and 3/4 codes have been obtained [22]. These notions of puncturing may be generalized for obtaining any coding rate $R = b/v$ or $R = 1/v$, $v < v_o$.

For example, starting from a low rate code, $R = 1/v_o$, one could obtain a series of low-rate codes $R = 1/v$, $v < v_o$ using degenerate perforation matrices (or perforation vectors) having 1 column and v_o rows. Clearly, then, for a rate $R = 1/v$ code, the perforation vector will have v 1s and $(v_o - v)$ 0s. For example, starting from a mother code of rate $R = 1/6$, the perforation vectors $\mathbf{P_8}$ to $\mathbf{P_{11}}$ yield punctured codes of rates $R = 1/5$, 1/4, 1/3, and 1/2, respectively,

$$\mathbf{P_8} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{P_9} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{P_{10}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \mathbf{P_{11}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \tag{3}$$

Clearly, all the above codes are rate-compatible, and in principle one could further extend the procedure to the usual perforation patterns and obtain punctured high-rate codes $R = b/v$, which would all be issued from the same very low-rate original $R = 1/v_o$ code.

Finally, instead of puncturing some of the code symbols at the output of the original encoder, using a *repetition matrix* one could perform instead a repetition of some code symbols leading to a decrease of the coding rate. Starting with a coding rate $R = \dfrac{b}{bv_o}$, a repetition matrix will have v_o rows and b columns, and repetition of $(n - bv_o)$ code symbols will yield a coding rate $R = b/n$, $n > bv_o$, thus allowing one to obtain practically any coding rate. In a repetition matrix, the code symbol being repeated is identified by an integer equal to the number of repetition. For example, starting from a coding rate $R = 5/10$, the repetition matrices $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ yield the coding rates $R = 5/11$, 5/13, and 5/16, respectively,

$$\mathbf{Q_1} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\mathbf{Q_2} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 2 & 1 & 2 \end{pmatrix}$$

$$\mathbf{Q_3} = \begin{pmatrix} 3 & 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 & 1 \end{pmatrix}$$

The basic notions of encoding punctured codes having been established, the problem of their decoding by either the Viterbi algorithm or sequential decoding is examined in Section 3.

3. DECODING OF PUNCTURES CODES

3.1. Viterbi Decoding

Given the received sequence from the channel, Viterbi decoding consists essentially of computing for every distinct encoder state the likelihood (also called the metric) that a particular sequence has been transmitted. For rate $R = b/v$ codes, there are 2^b paths merging at every trellis state, and only the path with the largest metric is selected at each state. The process is repeated at each of the 2^M encoder states and for each trellis depth so that, clearly, for a given M as b increases, the complexity of decoding also increases very rapidly.

For punctured high-rate b/v codes, Viterbi decoding is hardly more complex than for the original low-rate $1/v_o$ code from which the punctured codes are derived. The decoding is performed on the trellis of the original low-rate code, where the only modification consists of discarding the metric increments corresponding to the punctured code

symbols. Given the perforation pattern of the punctured code, this can be readily performed by inserting dummy data into the positions corresponding to the deleted code symbols. In the decoding process, the dummy data are discarded by assigning them the same metric value (usually zero) regardless of the code symbol, 0 or 1. For either hard- or soft-quantized channels, this procedure in effect inhibits the conventional metric calculation for the punctured symbols. In addition to that metric inhibition, the only coding rate-dependent modification in a variable-rate codec is the truncation path length, which must be increased with the coding rate. All other operations of the decoder remain essentially unchanged [6–9,21].

Therefore, Viterbi codecs for high-rate punctured codes involve none of the complexity required for the straightforward decoding of high-rate b/v codes. They can be implemented by adding relatively simple circuitry to the codecs of the original low-rate $1/v_o$ code. Furthermore, since a given low-rate $1/v_o$ code can give rise to a large number of high-rate punctured codes, the punctured approach leads to very attractive realizations of variable-rate Viterbi decoding. In fact, virtually all hardware implementations of convolutional codecs for high-rate codes use the punctured approach.

The punctured approach to high-rate codes can just as easily be applied to sequential decoding, where the decoding of the received message is performed one tree-branch at a time, without searching the entire tree. Here again, decoding is performed on the tree of the original low-rate code rather than on the tree on the high-rate code where, like for Viterbi decoding, the metric of the punctured symbols is inhibited. Therefore, either the Fano [10] or the Ziganzirov–Jelinek stack decoding algorithm could be used for the decoding of high-rate punctured codes with minimal modifications [13–15,19].

Finally, the same approach as described above can also be applied for iterative Turbo decoding using the MAP algorithm or its variants [4,5].

3.2. Error Performance

For discrete memoryless channels, upper bounds on both the sequence and bit error probabilities of a convolutional code can be obtained. The derivation of the bounds is based on a union bound argument on the transfer function $T(D, B)$ of the code that describes the weight distribution, or weight spectrum, of the incorrect codewords and the number of bit errors on these codewords or paths [1,16]. Except for short memory codes, the entire transfer function of the code is rarely known in closed form, but upper bounds on the error performances can still be calculated using only the first few terms of two series expansions related to the transfer function $T(D, B)$, that is

$$T(D,B)\ |_{B=1} = \sum_{j=d_{\text{free}}}^{\infty} a_j D^j \tag{4}$$

and

$$\frac{dT(D,B)}{dB}\ |_{B=1} = \sum_{j=d_{\text{free}}}^{\infty} c_j D^j \tag{5}$$

In these expressions, d_{free} is the free distance of the code and a_j is the number of incorrect paths or adversaries of Hamming weight $j, j \geq d_{\text{free}}$, that diverge from the correct path and remerge with it sometime later. As for c_j, it is simply the total number of bit errors in all the adversaries having a given Hamming weight j.

Using the weight spectrum, upper bounds on the sequence error probability P_E of a code of rate $R = b/v$ is given by

$$P_E \leq \sum_{j=d_{\text{free}}}^{\infty} a_j P_j \tag{6}$$

where P_j is the pair-wise error probability between two codewords having Hamming distance j. As for the upper bound on the bit error probability P_B, it is easily obtained by weighing each erroneous path by the number of information bits 1 contained in that path.

Since the coding rate is $R = b/v$, P_B is bounded by

$$P_B \leq \frac{1}{b} \sum_{j=d_{\text{free}}}^{\infty} c_j P_j \tag{7}$$

Evaluation of this bound depends on the actual expression of the pair-wise error probability P_j, which in turn depends on the type of modulation and channel parameters used [1,16]. For coherent PSK modulation and unquantized additive white Gaussian noise channels, the error probability between two codewords that differ over j symbols is given by

$$P_j = Q(\sqrt{2jRE_b/No}) \tag{8}$$

where

$$Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz \tag{9}$$

and where E_b/No is the energy per bit-to-noise density ratio.

For binary symmetric memoryless channels having transition probability p, $p < 0.5$, the pair-wise error probabilities P_j may be bounded by

$$P_j \leq 2^j (p(1-p))^{j/2} \tag{10}$$

However, using the exact expressions for P_j yields tighter bounds for P_B and P_E that closely match performance simulation results, that is

$$P_j = \begin{cases} \displaystyle\sum_{i=(j+1)/(2)}^{j} C_i^j p^i (1-p)^{j-i}, & j \text{ odd} \\ \displaystyle\sum_{i=(j)/(2)+1}^{j} C_i^j p^i (1-p)^{j-i} + \frac{1}{2} C_{j/2}^j (p(1-p))^{j/2}, & j \text{ even} \end{cases}$$

where $C_i^j = \dfrac{j!}{i!(j-i)!}, j > i$.

A good evaluation of the bounds on either P_E or P_B requires knowledge of the transfer functions (4) or (5). However, for the vast majority of codes, only the first few terms of either functions are known, and sometimes

only the leading coefficients $a_{d_{\text{free}}}$ and $c_{d_{\text{free}}}$ are available. However, for channels with large E_b/No values such as those usually used with high-rate codes, the bounds on P_E and P_B are dominated by the very first terms $a_{d_{\text{free}}}$ and $c_{d_{\text{free}}}$.

Naturally, bounds (6) and (7) are also applicable for punctured codes. Therefore, in deriving the error performances of punctured codes, the free distances and at least the first few terms of the weight spectra of these codes must be known. Partial weight spectra are provided here for both the best known short-memory and for long-memory codes.

4. SEARCH FOR GOOD PUNCTURED CODES

Since punctured coding was originally devised for Viterbi decoding, the criterion of goodness for these codes was the free distance, and hence the best free distance punctured codes that first appeared in the literature were all short-memory codes [6–9]. For sequential decoding, good long-memory punctured codes should have, in addition to a large free distance, a good distance profile.

In searching for good punctured codes of rate b/v and memory M, one is confronted with the problem of determining both an original code of memory M and rate $R = 1/v_o$, and its accompanying perforation pattern. Not unlike the search for usual convolutional codes, the search for punctured codes is often based on intuition and trial and error rather than on a strict mathematical construction [21–23]. An approach that yielded good results is based on the notion that "good codes generate good codes." Consequently, one could choose as the mother code a known good (even the best) code of memory M and rate $R = 1/v_o$, (e.g., R = 1/2, 1/3, 1/4, ...) and exhaustively try out all possible perforation patterns to generate the best possible punctured codes of rates $R = b/v$ and same memory M.

For each perforation pattern, the error probability must be evaluated, which in turn implies determining the corresponding weight spectrum. For a punctured code of rate $R = b/v$ obtained from a mother code of rate $R = 1/v_o$, the number of possible perforation matrices to consider is equal to $C_v^{bv_o}$. For example, using a mother code of rate $R = 1/2$, there are $C_8^{14} = 3003$ different perforation matrices to consider for determining the best punctured codes of rate $R = 7/8$. This number may be substantially reduced if the code must satisfy some specific properties. For example, if the code is systematic, that is the information sequence appears unaltered in the output code sequence, then clearly the first row of the perforation matrix is all composed of 1s, and the puncturing may take place only in the other rows. For example, for $R = 7/8$ systematic punctured codes, the puncturing matrices are of the following form

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

There are only seven possible alternatives, and in general the mother codes for $R = b/v$ systematic punctured codes must also be systematic.

Naturally, if families of variable-rate punctured codes are desired, then all the required perforation patterns must be applied to the same low-rate original code. Furthermore, if the codes are to be rate-compatible, then the perforation patterns must be selected accordingly [9].

Obviously, puncturing a code reduces its free distance, and hence a punctured code cannot achieve the free distance of its original code. Although the free distance of a code increases as its rate decreases, using original codes with rates $1/v_o$ lower than 1/2 does not always warrant punctured codes with larger free distances since, for a given b and v, the proportion of deleted symbols, $S/v_o b = 1 - (v/v_o b)$, also increases with v_o. Consequently, good results and ease of implementation tend to favor the use of rate-1/2 original codes for generating good punctured codes with coding rates of the form $R = (v - 1)/v$. Results on both short-($M \leq 8$) and long-memory punctured codes ($M \geq 9$) that are derived from the best-known $R = 1/2$ codes are provided in this article [21–23].

Although one could select the punctured code on the basis of its free distance only, a finer method consists of determining the weight spectrum of the punctured code according to Eqs. (4) and (5) and then calculating the error probability bounds Eqs. (6) and (7). The code yielding the best error performance may thus be selected as the best punctured code, provided that it is not catastrophic.

Clearly, then, starting from a known optimal low-rate code, a successful search for good punctured codes hinges on the ability for determining the weight spectrum corresponding to each possible perforation pattern. Although seemingly simple, finding the weight spectrum of a punctured code is a difficult task. This is due to the fact that even if the spectrum of the low-rate original code were available, the spectrum of the punctured code cannot easily be derived from it. One has to go back exploring the tree or trellis of the low-rate original code and apply the perforation pattern to each path of interest. For the well-known short-memory codes, the procedure is at best a rediscovery of their weight spectra, whereas for long-memory codes where often only the free distance is known, it is a novel determination of their spectra. The problem is further compounded by the fact that since puncturing a path reduces its Hamming weight, in obtaining spectral terms up to some target Hamming distance, a larger number of paths must be explored over much longer lengths for a punctured than for a usual non-punctured low-rate code.

For each perforation pattern, the corresponding weight spectrum consists essentially in the triplet $\{D^l, a_l, c_l\}$ from which the bounds on the error probabilities P_E and P_B can be calculated. Of course, the best punctured code will correspond to the perforation pattern yielding the best error probability bounds. In the search for the best punctured codes, the difficulty is compounded by the fact that in determining the weight spectrum of a code of a memory length M up to some Hamming weight value L, the computational effort is exponential in both M and L.

The best punctured codes presented here have been obtained using very efficient trellis-search algorithms for determining the weight spectra of convolutional codes [17,18]. Each algorithm has been designed to be

especially efficient within a given range of memory lengths, making it possible to obtain substantial extensions of the known spectra of the best convolutional codes. The spectral terms obtained by these exploration methods have been shown to match exactly the corresponding terms derived from the transfer function of the code [21], thus validating the search procedure.

4.1. Short-Memory Punctured Codes

A number of short-memory lengths ($M \leq 8$) punctured codes of rates $R = (v-1)/v$ varying from 2/3 to 16/17 have been determined first by Cain et al. [6], Yasuda et al. [7] and [8], and then by Hagenauer [9] for rate-compatible codes. In particular, in Ref. [7] all the memory-6 punctured codes of rates varying from 2/3 to 16/17 have been derived from the same original memory-6, rate-1/2, optimal convolutional code due to Odenwalder [16]. A more complete list of rate 2/3 to 13/14 punctured codes has been derived from the optimal rate-1/2 codes with memory M varying from 2 to 8 and compiled by Yasuda et al. [8]. In this list, the perforation matrix for each code is provided, but the weight spectrum is limited to the first term only, that is, the term $c_{d_{\text{free}}}$ corresponding to d_{free}.

Using the given perforation patterns, Haccoun et al. [21–23] have extended Yasuda's results by determining the first 10 spectral terms a_n and c_n for all the codes having memory lengths $2 \leq M \leq 8$ and coding rates 2/3, 3/4, 4/5, 5/6, 6/7, and 7/8. These results are given in Tables 1 to 6, respectively. Each table lists the generators of the original low-rate code, the perforation matrix, the free distance of the resulting punctured code, and the coefficients a_n and c_n of the series expansions of the corresponding weight spectra up to the 10 spectral terms. Beyond 10 spectral coefficients the required computer time becomes rather prohibitive, and hardly any further precision of the bounds is gained by considering more than 10 spectral terms.

To verify the accuracy and validity of these results, the entire transfer functions $T(D, B)$ and $\dfrac{D(T, B)}{dB}$ have been analytically derived for both rate-2/3 and rate-4/5 memory-2 punctured codes. The transfer functions of the codes have been determined by solving the linear equations describing the transitions between the different states of the encoder. As expected, all the spectral terms obtained by the search algorithms have been shown to match exactly the corresponding terms of the transfer functions [22].

The bit error probability upper bound P_B over both unquantized and hard quantized additive white Gaussian noise channels has been evaluated for all the codes listed in Tables 1 to 6 [21–23], using all the listed weight spectra terms. For all these codes, the error performance improves as the coding rate decreases, indicating well-chosen perforation patterns. A notable exception concerns the memory $M = 3$ code, where the bit error performance turned out to be slightly better at rate 4/5 than at rate 3/4. This anomaly may be explained by an examination of the spectra of these two codes. As shown in Tables 2 and 3, although the free distances of the rates 3/4 and 4/5 codes are $d_f = 4$ and $d_f = 3$, respectively, the number of bit errors c_n on the various spectral terms are far larger for the rate 3/4 code than for the rate 4/5 code. Clearly, it is the coefficients c_n that adversely affect the error performance of the rate 3/4 code. This anomaly illustrates the fact that selecting a code according to only the free distance is good in general but may sometimes be insufficient. Knowledge of further terms of the spectrum will always provide more insight on the code performance.

The theoretical bound on P_B for the original $M = 6$, $R = 1/2$ optimal-free-distance code is plotted in Fig. 4, together with those of all the best punctured codes with rates $R = 2/3$, 3/4, 4/5, 5/6, 6/7, and 7/8 derived from it. For comparison purposes, Fig. 4 also includes the performance curve for the uncoded coherent PSK modulation. It shows that the performance degradation from the original rate-1/2 code is rather gentle as the coding rate increases from 1/2 to 7/8. For example, at $P_B = 10^{-5}$, the coding gains for the punctured rate-2/3 and 3/4 codes are 5.1 dB and 4.6 dB, respectively. These results indicate that these codes are

Table 1. Weight Spectra of Yasuda et al. Punctured Codes with $R = 2/3$, $2 \leq M \leq 8$, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
2	5	7	5	1 0 1 1	3	(1, 4, 14, 40, 116, 339, 991, 2897, 8468, 24752) [1, 10, 54, 226, 856, 3072, 10647, 35998, 119478, 390904]
3	15	17	6	1 1 1 0	4	(3, 11, 35, 114, 381, 1276, 4257, 14208, 47413, 158245) [10, 43, 200, 826, 3336, 13032, 49836, 187480, 696290, 2559521]
4	23	35	7	1 1 1 0	4	(1, 0, 27, 0, 345, 0, 4528, 0, 59435, 0) [1, 0, 124, 0, 2721, 0, 50738, 0, 862127, 0]
5	53	75	8	1 0 1 1	6	(19, 0, 220, 0, 3089, 0, 42790, 0, 588022, 0) [96, 0, 1904, 0, 35936, 0, 638393, 0, 10657411]
6	133	171	10	1 1 1 0	6	(1, 16, 48, 158, 642, 2435, 9174, 34705, 131585, 499608) [3, 70, 285, 1276, 6160, 27128, 117019, 498860, 2103891, 8784123]
7	247	371	10	1 0 1 1	7	(9, 35, 104, 372, 1552, 5905, 22148, 85189, 323823, 1232139) [47, 237, 835, 3637, 17770, 76162, 322120, 1374323, 5730015, 23763275]
8	561	753	12	1 1 1 0	7	(3, 9, 50, 190, 641, 2507, 9745, 37121, 142226, 545002) [11, 46, 324, 1594, 6425, 29069, 127923, 544616, 2313272, 9721227]

Table 2. Weight Spectra of Yasuda et al. Punctured Codes with $R = 3/4$, $2 \le M \le 8$, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
2	5	7	5	1 0 1 1 1 0	3	(6, 23, 80, 290, 1050, 3804, 13782, 49930, 180890, 655342) [15, 104, 540, 2557, 11441, 49340, 207335, 854699, 3471621, 13936381]
3	15	17	6	1 1 0 1 0 1	4	(29, 0, 532, 0, 10059, 0, 190112, 0, 3593147, 0) [124, 0, 4504, 0, 126049, 0, 3156062, 0, 74273624, 0]
4	23	35	7	1 0 1 1 1 0	3	(1, 2, 23, 124, 576, 2852, 14192, 70301, 348427, 1726620) [1, 7, 125, 936, 5915, 36608, 216972, 1250139, 7064198, 39308779]
5	53	75	8	1 0 0 1 1 1	4	(1, 15, 65, 321, 1661, 8396, 42626, 216131, 1095495, 5557252) [3, 85, 490, 3198, 20557, 123384, 725389, 4184444, 23776067, 133597207]
6	133	171	10	1 1 0 1 0 1	5	(8, 31, 160, 892, 4512, 23307, 121077, 625059, 3234886, 16753077) [42, 201, 1492, 10469, 62935, 379644, 2253373, 13073811, 75152755, 428005675]
7	247	371	10	1 1 0 1 0 1	6	(36, 0, 990, 0, 26668, 0, 726863, 0, 19778653, 0) [239, 0, 11165, 0, 422030, 0, 14812557, 0, 493081189, 0]
8	561	753	12	1 1 1 1 0 1	6	(10, 77, 303, 1599, 8565, 44820, 236294, 1236990, 6488527, 34056195) [52, 659, 3265, 21442, 133697, 805582, 4812492, 28107867, 162840763, 935232173]

Table 3. Weight Spectra of Yasuda et al. Punctured Codes with $R = 4/5$, $2 \le M \le 8$, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
2	5	7	5	1 0 1 1 1 1 0 0	2	(1, 12, 53, 238, 1091, 4947, 22459, 102030, 463451) [1, 36, 309, 2060, 12320, 69343, 375784, 1983150, 10262827]
3	15	17	6	1 0 1 1 1 1 0 0	3	(5, 36, 200, 1070, 5919, 32721, 180476, 995885, 5495386, 30323667) [14, 194, 1579, 11313, 77947, 514705, 3305113, 20808587, 129003699, 790098445]
4	23	35	7	1 0 1 0 1 1 0 1	3	(3, 16, 103, 675, 3969, 24328, 147313, 897523, 5447618, 33133398) [11, 78, 753, 6901, 51737, 386465, 2746036, 19259760, 132078031, 896198879]
5	53	75	8	1 0 0 0 1 1 1 1	4	(7, 54, 307, 2005, 12970, 83276, 534556, 3431703, 22040110) [40, 381, 3251, 27123, 213451, 1621873, 12011339, 87380826, 627189942]
6	133	171	10	1 1 1 1 1 0 0 0	4	(3, 24, 172, 1158, 7409, 48729, 319861, 2097971, 13765538, 90315667) [12, 188, 1732, 15256, 121372, 945645, 7171532, 53399130, 392137968, 2846810288]
7	247	371	10	1 0 1 0 1 1 0 1	5	(20, 115, 694, 4816, 32027, 210909, 1392866, 9223171, 61013236) [168, 1232, 9120, 78715, 626483, 4758850, 35623239, 263865149, 1930228800]
8	561	753	12	1 1 0 1 1 0 1 0	5	(7, 49, 351, 2259, 14749, 99602, 663936, 4431049, 29536078, 197041141) [31, 469, 4205, 34011, 268650, 2113955, 16118309, 121208809, 898282660, 2301585211]

very good indeed, even though their free distances, which are equal to 6 and 5, respectively, are slightly smaller than the free distances of the best-known usual rate-2/3 and 3/4 codes, which are equal to 7 and 6, respectively [11].

The error performance of these codes has been verified using an actual punctured Viterbi codec [7,8], and independently, using computer simulation [21–23]. Both evaluations have been performed using 8-level soft decision Viterbi decoding with truncation path lengths equal to 50, 56, 96, and 240 bits for the coding rates 2/3, 3/4, 7/8, and 15/16, respectively. Both hardware and software evaluations have yielded identical error performances that closely match the theoretical upper bounds.

Even for the rate 15/16, the coding gain of the $M = 6$ code has been shown to reach a substantial 3 dB at $P_B = 10^{-6}$ [3]. The fact that such a coding gain can be achieved with only a 7% redundancy and a Viterbi decoder that is hardly more complex than for a rate-1/2 code makes the punctured coding technique very attractive indeed for short-memory codes. For longer codes and larger coding gains, Viterbi decoding becomes impractical and other decoding techniques such as sequential decoding should be considered instead. Long-memory length punctured codes and their error performance are presented next.

4.2. Long-Memory Punctured Codes

Following essentially the same approach as for the short-memory codes, one could choose a known optimal long-memory code of rate 1/2 and exhaustively try out all possible perforation patterns to generate all punctured codes of rate $R = b/v$. The selection of the best punctured code is again based on its bit error performance, which is calculated from the series expansion of its transfer function. Here, one of the difficulties is that for the original

Table 4. Weight Spectra of Yasuda et al. Punctured Codes with $R = 5/6$, $2 \le M \le 8$, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
2	5	7	5	1 0 1 1 1 1 1 0 0 0	2	(2, 26, 129, 633, 3316, 17194, 88800, 459295, 2375897, 12288610) [2, 111, 974, 6857, 45555, 288020, 1758617, 10487425, 61445892, 355061333]
3	15	17	6	1 0 1 0 0 1 1 0 1 1	3	(15, 96, 601, 3918, 25391, 164481, 1065835, 6906182, 44749517) [63, 697, 6367, 53574, 426471, 3277878, 24573195, 180823448, 1311630186]
4	23	35	7	1 0 1 1 1 1 1 0 0 0	3	(5, 37, 309, 2282, 16614, 122308, 900991, 6634698, 48853474) [20, 265, 3248, 32328, 297825, 2638257, 22710170, 191432589, 1587788458]
5	53	75	8	1 0 0 0 0 1 1 1 1 1	4	(19, 171, 1251, 9573, 75167, 585675, 4558463, 35513472) [100, 1592, 17441, 166331, 1591841, 14627480, 131090525, 1155743839]
6	133	171	10	1 1 0 1 0 1 0 1 0 1	4	(14, 69, 654, 4996, 39699, 315371, 2507890, 19921920, 158275483) [92, 528, 8694, 79453, 792114, 7375573, 67884974, 610875423, 1132308080]
7	247	371	10	1 1 1 0 0 1 0 0 1 1	4	(2, 51, 415, 3044, 25530, 200878, 1628427, 12995292, 104837990) [7, 426, 5244, 49920, 514857, 4779338, 44929071, 406470311, 3672580016]
8	561	753	12	1 0 1 1 0 1 1 0 0 1	5	(19, 187, 1499, 11809, 95407, 775775, 6281882, 50851245) [168, 2469, 25174, 242850, 2320429, 21768364, 199755735, 1807353406]

Table 5. Weight Spectra of Yasuda et al. Punctured Codes with $R = 6/7$, $2 \le M \le 8$, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
2	5	7	5	1 0 1 1 1 1 1 1 0 0 0 0	2	(4, 39, 221, 1330, 8190, 49754, 302405, 1840129, 11194714, 68101647) [5, 186, 1942, 16642, 131415, 981578, 7076932, 49784878, 343825123, 2340813323]
3	15	17	6	1 0 0 0 1 1 1 1 1 1 0 0	2	(1, 25, 188, 1416, 10757, 81593, 619023, 4697330, 35643844) [2, 134, 1696, 18284, 179989, 1676667, 15082912, 132368246, 1140378555]
4	23	35	7	1 0 1 0 1 0 1 1 0 1 0 1	3	(14, 100, 828, 7198, 60847, 513573, 4344769, 36751720) [69, 779, 9770, 113537, 1203746, 12217198, 120704682, 1167799637]
5	53	75	8	1 1 0 1 1 0 1 0 1 0 0 1	3	(5, 55, 517, 4523, 40476, 362074, 3232848, 28872572) [25, 475, 6302, 73704, 823440, 8816634, 91722717, 935227325]
6	133	171	10	1 1 1 0 1 0 1 0 0 1 0 1	3	(1, 20, 223, 1961, 18093, 169175, 1576108, 14656816, 136394365) [5, 169, 2725, 32233, 370861, 4169788, 45417406, 483171499, 768072194]
7	247	371	10	1 0 1 0 0 1 1 1 0 1 1 0	4	(11, 155, 1399, 13018, 122560, 1154067, 10875198, 102494819, 965649475) [85, 1979, 24038, 282998, 3224456, 35514447, 383469825, 4075982541, 4092715598]
8	561	753	12	1 1 0 1 1 0 1 0 1 0 0 1	4	(2, 48, 427, 4153, 39645, 377500, 3600650, 34334182) [9, 447, 5954, 76660, 912140, 10399543, 115459173, 1256388707]

low-rate and long-memory codes of interest, only very partial knowledge of their weight spectra is available. In fact, beyond memory length $M = 15$, very often only the free distances of these codes are available in the literature [12].

Results of computer search for the best rate-2/3 and 3/4 punctured codes with memory lengths ranging from 9 to 23 that are derived from the best-known rate-1/2 are provided in Refs. 21–23, where for each code the first few terms of the weight spectrum have been obtained for each possible distinct perforation pattern. The final selection of the best punctured codes was based on the evaluation of the upper bound on the bit error probability. Naturally, the codes obtained with this approach are suitable for variable-rate codecs using an appropriate decoding technique such as sequential decoding. Only these two rates have been considered because a definite comparison of the resulting punctured codes with the best-known nonsystematic high-rate codes is limited to the rate-2/3 and 3/4 codes, since with very few exceptions, optimal long memory codes suitable for sequential decoding are known only for rates 2/3 and 3/4.

Tables 7 and 8 list the characteristics of the best punctured codes of rate 2/3 and 3/4, respectively, with memory lengths M varying from 9 to 23, that are derived from the best nonsystematic rate-1/2 codes [21]. From $M = 9$ to $M = 13$ the original codes are the maximal free distance codes discovered by Larsen [19], whereas for $14 \le M \le 23$ the original codes are those of Johannesson and Paaske [12]. In both of these tables, for each memory length the generators of the original code and its perforation matrix are given, together with the free distances of both the original and resulting punctured code. Just as with short-memory codes, the first few terms

Table 6. Weight Spectra of Yasuda et al. Punctured Codes with $R = 7/8$, $2 \le M \le 8$, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
2	5	7	5	1 0 1 1 1 1 1 1 1 0 0 0 0 0	2	(6, 66, 408, 2636, 17844, 119144, 793483, 5293846, 35318216) [8, 393, 4248, 38142, 325739, 2647528, 20794494, 159653495, 1204812440]
3	15	17	6	1 0 0 0 0 1 0 1 1 1 1 1 0 1	2	(2, 38, 346, 2772, 23958, 201842, 1717289, 14547758, 123478503) [4, 219, 3456, 38973, 437072, 4492304, 45303102, 442940668, 4265246076]
4	23	35	7	1 0 1 0 0 1 1 1 1 0 1 1 0 0	3	(13, 145, 1471, 14473, 143110, 1416407, 14019214, 138760394) [49, 1414, 21358, 284324, 3544716, 42278392, 489726840, 1257797047]
5	53	75	8	1 0 1 1 1 0 1 1 1 0 0 0 1 0	3	(9, 122, 1195, 12139, 123889, 1259682, 12834712, 130730637, 1331513258) [60, 1360, 18971, 252751, 3165885, 38226720, 450898174, 923001734, 3683554219]
6	133	171	10	1 1 1 1 0 1 0 1 0 0 0 1 0 1	3	(2, 46, 499, 5291, 56179, 599557, 6387194, 68117821) [9, 500, 7437, 105707, 1402743, 17909268, 222292299, 2706822556]
7	247	371	10	1 0 1 0 1 0 0 1 1 0 1 0 1 1	4	(26, 264, 2732, 30389, 328927, 3571607, 38799203) [258, 3652, 52824, 746564, 9825110, 125472545, 1567656165]
8	561	753	12	1 1 0 1 0 1 1 1 0 1 0 1 0 0	4	(6, 132, 1289, 13986, 154839, 1694634, 18532566) [70, 1842, 24096, 337514, 4548454, 58634237, 738611595]

a_n and c_n, $n = d_{\text{free}}, d_{\text{free}} + 1, d_{\text{free}} + 2, \ldots$ of the weight spectra are also given for each punctured codes. In deriving these spectral coefficients, the tree exploration of the original code had to be performed over a considerable length [21].

In the search for the best punctured codes, the perforation patterns were chosen as to yield both a maximal free distance and a good distance profile. Although all perforation patterns were exhaustively examined, the search was somewhat reduced by exploiting equivalences of the perforation patterns under cyclical shifts of their rows [21–23]. Among all the codes that were found, Tables 7 and 8 list only those having the smallest number of bit errors $c_{d_{\text{free}}}$ at the free distance d_{free}, and obviously all the codes listed are noncatastrophic.

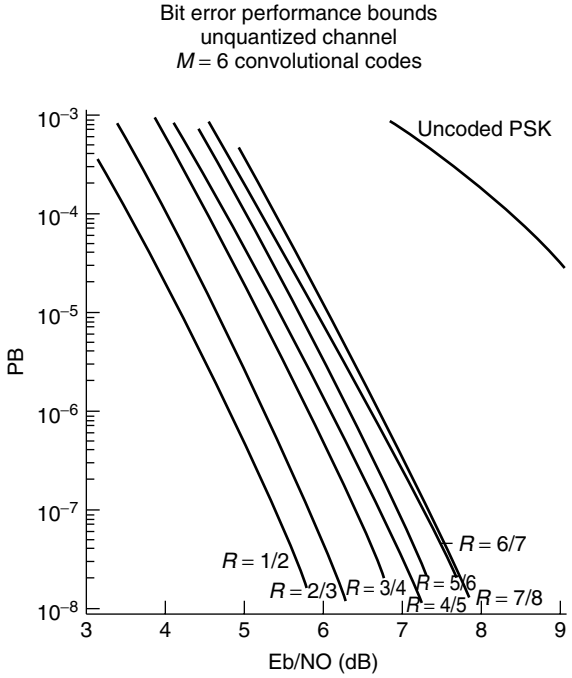


Figure 4. Bit error performance bounds for the $M = 6$, $R = 1/2$ original and punctured rates 2/3, 3/4, 4/5, 5/6, 6/7, and 7/8 codes, [21].

Table 7 lists two $M = 19$, $R = 2/3$ punctured codes derived from two distinct good $M = 19$, $R = 1/2$ original codes. The free distances of these two punctured codes are equal to 12 and 13, respectively, but the coefficients c_n are larger for the $d_{\text{free}} = 13$ code than they are for the $d_{\text{free}} = 12$ code. Consequently, as confirmed by the calculation of the bit error bound P_B, the code with $d_{\text{free}} = 13$ turned out to be slightly worse by approximately 0.35 dB than the code with $d_{\text{free}} = 12$. This anomaly again confirms the need to determine at least the first few terms of the weight spectra when searching for the best punctured codes.

Figure 5 plots the free distances of the original rate 1/2 codes and the punctured rate-2/3 and 3/4 codes as a function of the memory length M, $2 \le M \le 22$. Except for the two anomalies with the $M = 19$ mentioned above, the behavior of the free distances is as expected: the free distance of the punctured codes of a given rate is generally nondecreasing with the memory length, and for a given memory length the free distance decreases with increasing coding rates.

When the punctured codes of rate b/v are determined from the best original low-rate $1/v$ code, an upper bound on the free distance of the punctured code can be derived [21,22]. This derivation, which is based on an analysis of the effect of the different perforation patterns on the spectrum of the original code, yields the bound

$$d_{\text{free}(p)} \le (1/b)\, d_{\text{free}(0)} \tag{12}$$

where $d_{\text{free}(p)}$ and $d_{\text{free}(o)}$ are the free distances of the punctured and best codes of rate $1/v$, respectively. This bound is also plotted on Fig. 5 for the rate-2/3 and 3/4 codes with memory $M \le 13$.

Table 7. Best Rate-2/3, $9 \le M \le 23$ Punctured Codes with Their Weight Spectra, Perforation Matrix, and Original $R = 1/2$ Codes, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
9	1167	1545	12	1 1	7	(1, 10, 29, 94, 415, 1589, 5956)
				1 0		[3, 70, 207, 836, 4411, 19580, 82154]
10	2335	3661	14	1 0	8	(1, 21, 65, 226, 907, 3397, 13223)
				1 1		[8, 165, 560, 2321, 10932, 46921, 204372]
11	4335	5723	15	1 1	9	(10, 38, 137, 518, 1990, 7495, 28907)
				1 0		[86, 326, 1379, 6350, 27194, 114590, 492275]
12	10533	17661	16	1 1	9	(4, 8, 45, 193, 604, 2383, 9412)
				1 0		[25, 65, 413, 1991, 6925, 31304, 139555]
13	21675	27123	16	1 1	10	(5, 30, 104, 380, 1486)
				1 0		[46, 268, 1066, 4344, 19992]
14	55367	63121	18	1 1	10	(2, 6, 37, 153, 582)
				1 0		[13, 62, 334, 1606, 7321]
15	111653	145665	19	1 1	10	(3, 0, 46, 0, 683, 0)
				1 0		[28, 0, 397, 0, 7735, 0]
16	347241	246277	20	1 1	12	(8, 45, 145, 567, 2182)
				1 0		[68, 495, 1569, 7112, 31556]
17	506477	673711	20	1 0	12	(2, 24, 79, 320, 1251)
				1 1		[11, 253, 889, 3978, 18056]
18	1352755	1771563	21	1 1	12	(2, 11, 27, 137)
				1 0		[18, 105, 276, 1679]
19	2451321	3546713	22	1 1	12	(1, 3, 14, 71)
				1 0		[9, 21, 139, 715]
19	2142513	3276177	22	1 1	13	(10, 34, 101, 417, 1539)
				1 0		[99, 425, 1425, 6158, 25037]
20	6567413	5322305	22	1 1	12	(1, 0, 18, 0, 333, 0)
				1 0		[8, 0, 210, 0, 4290, 0]
21	15724153	12076311	24	1 1	13	(1, 1, 20, 62)
				1 0		[11, 5, 231, 736]
22	33455341	24247063	24	1 0	14	(1, 12, 67)
				1 1		[17, 163, 927]
23	55076157	75501351	25	1 0	15	(2, 14, 71)
				1 1		[39, 170, 852]

The upper bounds on the bit error probability over unquantized white Gaussian channels have been evaluated for all the punctured codes of rates $R = 2/3$ and 3/4 and are shown for even values of memory lengths in Figs. 6 and 7. These bounds indicate a normal behavior for all the punctured codes listed in Tables 7 and 8. All the bit error performances improve as the coding rate decreases and/or as the memory increases, with approximately 0.4 dB improvement for each unit increase of the memory length. At $P_B = 10^{-5}$, the $M = 22, R = 2/3$, and $R = 3/4$ punctured codes can yield substantial coding gains of 8.3 dB and 7.7 dB, respectively.

The selection of the best punctured codes listed in Tables 7 and 8 has been based on both the maximal free distance and the calculated bit error probability bound. However, the choice of the best punctured code is not always clear-cut as different puncturing patterns may yield only marginally different error performances. In some cases, the performance curves may even be undistinguishable, leading to several "best" punctured codes having the same coding rate and memory length. However, since these long codes are typically for sequential decoding applications, the final selection of the code should also be based on the distance profile and computational performance. Short of analyzing the computational behavior, when in doubt, the codes finally selected and listed in the tables had the fastest-growing column distance function.

In the search for punctured codes, the above approach will produce good but not necessarily optimal codes since the original low-rate code is imposed at the outset. A measure of the discrepancy between optimal and punctured codes of the same rate and memory length is provided in Fig. 8, which shows the bit error performance bound for both the best punctured and maximal free distance codes of memory length 9 and rates 2/3 and 3/4. These bounds have been computed using only the term at d_{free} for the maximum free distance (MFD) codes and using both the terms at d_{free} and $d_{\text{free}} + 1$ for the punctured codes. Based on these terms only, the two MFD codes appear to be only slightly better than the punctured codes. Therefore, it may be concluded that, although not optimal, the error performances of the rate-2/3 and 3/4 punctured codes of memory 9 closely match those of the MFD codes of the same rates and memory lengths. The same general

Table 8. Best Rate-3/4, $9 \le M \le 23$ Punctured Codes with Their Weight Spectra, Perforation Matrix, and Original $R = 1/2$ Codes, [21]

Original Code				Punctured Code		
M	G_1	G_2	d_f	$[P]$	d_f	$(a_n, n = d_f, d_{f+1}, d_{f+2}, \ldots)$ $[c_n, n = d_f, d_{f+1}, d_{f+2}, \ldots]$
9	1167	1545	12	1 0 0	6	(4, 31, 151, 774, 3967, 21140)
				1 1 1		[38, 270, 1640, 10554, 63601, 387227]
10	2335	3661	14	1 0 1	6	(2, 7, 59, 338, 1646)
				1 1 0		[9, 40, 517, 3731, 22869]
11	4335	5723	15	1 0 0	7	(12, 55, 236, 1271, 6853)
				1 1 1		[107, 628, 3365, 20655, 126960]
12	10533	17661	16	1 1 0	7	(4, 18, 90, 476, 2466)
				1 0 1		[34, 182, 965, 6294, 38461]
13	21675	27123	16	1 1 0	7	(1, 11, 41, 202, 1334)
				1 0 1		[12, 109, 387, 2711, 20403]
14	55367	63121	18	1 0 1	8	(3, 19, 95, 529)
				1 1 0		[28, 159, 1186, 7461]
15	111653	145665	19	1 0 0	8	(1, 14, 47, 259)
				1 1 1		[9, 143, 512, 3571]
16	347241	246277	20	1 1 0	8	(1, 5, 28, 167)
				1 0 1		[5, 49, 311, 2266]
17	506477	673711	20	1 0 0	9	(1, 13, 101, 427)
				1 1 1		[5, 142, 1375, 6842]
18	1352755	1771563	21	1 1 1	10	(6, 51, 217, 1014)
				1 0 0		[104, 735, 3368, 18736]
19	2451321	3546713	22	1 1 0	10	(4, 18, 81, 429)
				1 0 1		[40, 240, 1219, 6934]
19	2142513	3276177	22	1 1 0	10	(4, 18, 89, 461, 2529)
				1 0 1		[48, 202, 1248, 7445, 46981]
20	6567413	5322305	22	1 1 1	10	(4, 19, 82, 436, 2443)
				1 0 0		[40, 249, 1510, 8120, 53164]
21	15724153	12076311	24	1 1 1	11	(8, 19, 120)
				1 0 0		[143, 266, 2038]
22	33455341	24247063	24	1 0 0	12	(7, 68, 298)
				1 1 1		[79, 1275, 5279]
23	55076157	75501351	25	1 0 0	13	(21, 141, 707)
				1 1 1		[292, 2340, 13196]

conclusions may be made about the slight suboptimality of the other punctured codes with different memory lengths [21–23]. However, as mentioned earlier, the small error performance degradation of the punctured codes is compensated for by a far simpler practical implementation.

Given an optimal usual high-rate code of rate $R = b/v$ and memory length M, one could attempt to determine the low-rate $1/v$ code that, after puncturing, will yield a punctured code that is equivalent to that optimal code. This approach, which is the converse of the usual code searching method, can be used to find the punctured code equivalent to any known usual high-rate code. Based on this approach and using the notion of orthogonal perforation patterns, a systematic construction technique has been developed by Begin and Haccoun [22]. Using this technique, punctured codes strictly equivalent to the best-known nonsystematic rate-2/3 codes with memory lengths up to $M = 24$ have been found. Likewise, punctured codes equivalent to the best-known rate-3/4 codes with memory lengths up to $M = 9$ have also been determined [22]. Therefore, optimal high-rate codes may be obtained as punctured codes, but it should be pointed out that the punctured codes generated by this latter approach are not suitable for variable-rate applications since each punctured code has its own distinct low-rate original code and orthogonal perforation pattern.

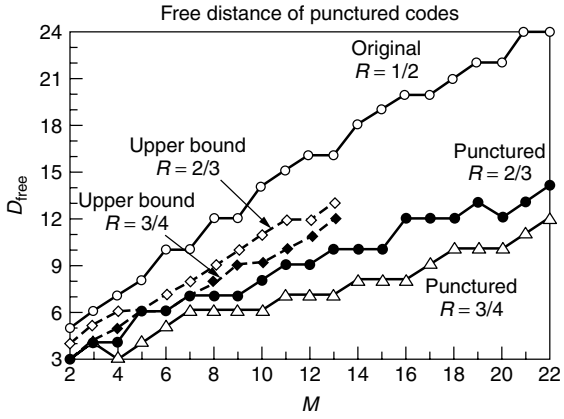


Figure 5. Free distances of original rate-1/2 codes and punctured $R = 2/3$ and 3/4 codes derived from them as a function of M, $2 \le M \le 22$, and upper bounds, [21].

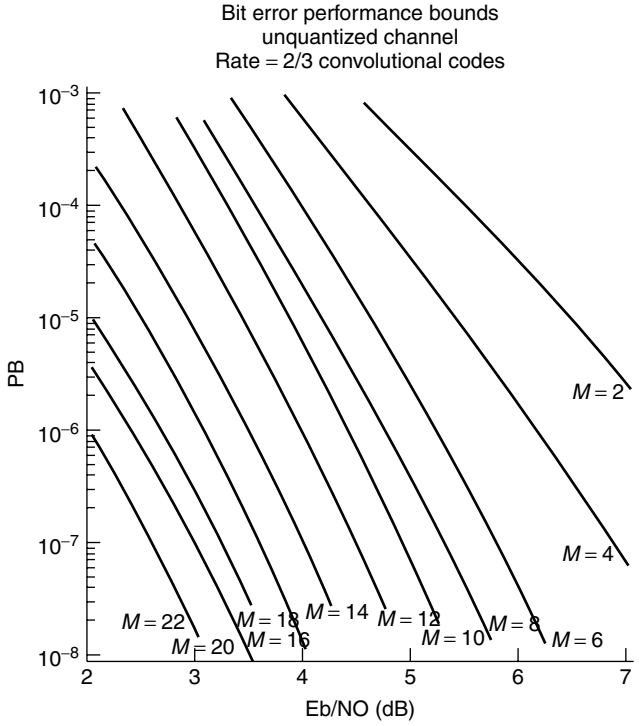


Figure 6. Upper bounds on the bit error probability over unquantized white Gaussian noise channels for $R = 2/3$ punctured codes with $2 \leq M \leq 22$, M even, [21].

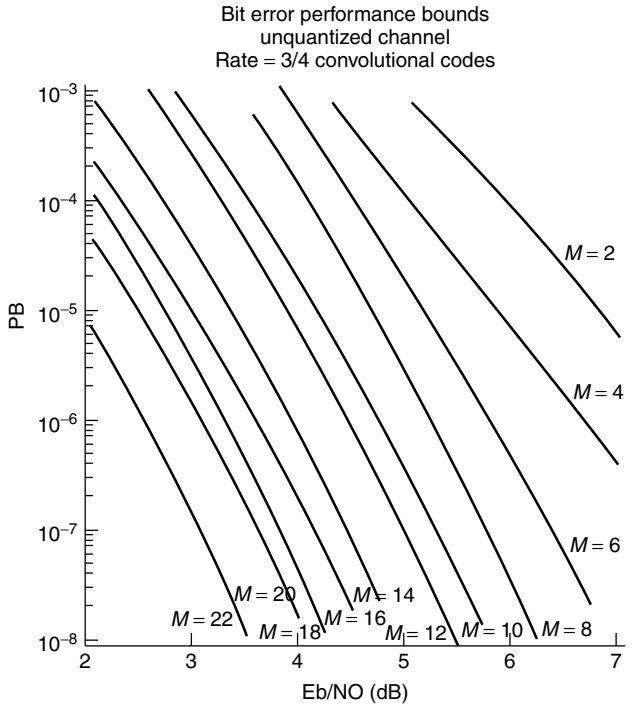


Figure 7. Upper bounds on the bit error probability over unquantized white Gaussian noise channels for $R = 3/4$ punctured codes with $2 \leq M \leq 22$, M even, [21].

5. PUNCTURED CODES EQUIVALENT TO THE BEST USUAL HIGH-RATE CODES

5.1. Nonsystematic Codes

Given a best-known high-rate usual nonsystematic code, Haccoun and Begin have devised a construction method for generating a low-rate original code that, when punctured by an orthogonal perforation pattern, yields a punctured code having both the same rate and same weight spectrum as the best-known usual code [22]. These codes are listed in Tables 9 to 12. Tables 9 to 11 use the same orthogonal perforation matrix $\mathbf{P_{01}}$ whereas Table 12 uses the orthogonal perforation matrix $\mathbf{P_{02}}$:

$$\mathbf{P_{01}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \mathbf{P_{02}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{13}$$

The $R = 2/3$ codes are obtained by puncturing $R = 1/3$ original codes, and the $R = 3/4$ punctured codes are obtained from $R = 1/4$ original codes. Tables 9 to 12 provide the memory lengths and generators of the original low-rate codes and the resulting punctured codes equivalent to their respective best usual codes.

We can observe from the tables that, for most of the cases, the memory of the required original code is larger than that of the resultant equivalent punctured code. The

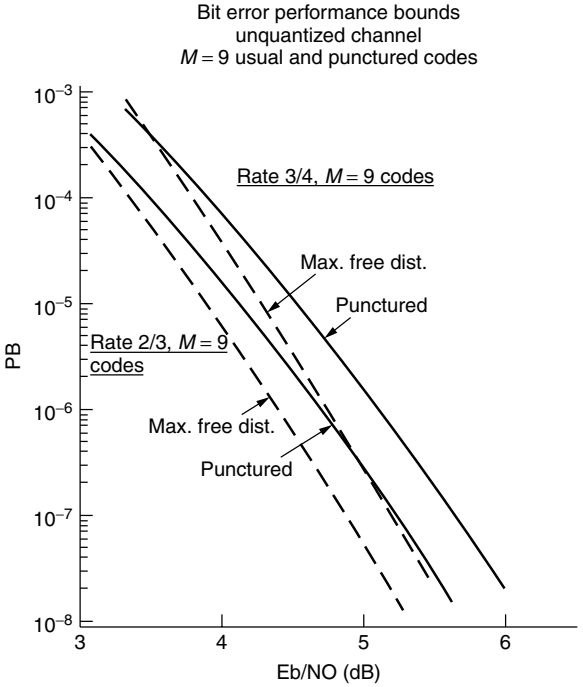


Figure 8. Bit error performance bounds for maximal free distance (MFD) codes and punctured codes of rates 2/3 and 3/4 with memory $M = 9$, [21].

difference in memory length is quite small, usually one or two units, and this difference appears to be independent of the actual memory length of the code. Therefore, its relative importance decreases as the memory length of the code increases. For large memory codes, this memory

increment is of no consequence whatsoever since the codes certainly would be decoded by sequential decoding methods.

As mentioned earlier, the punctured codes listed in Tables 9 to 12 are not suitable for variable-rate or rate-compatible decoding applications. However, they do provide a practical method for coding and decoding the best-known long memory rate 2/3 and 3/4 codes, especially using sequential decoding. Decoding these codes by the normal (nonpunctured) approach is cumbersome because of the large number (2^b) of nodes stemming from each node in the tree of the high-rate $R = b/v$ code. With the punctured approach, the decoding proceeds on the low-rate tree structure, so that the number of nodes stemming from a single node is always 2, regardless of the actual coding rate.

5.2. Systematic Codes

Using a similar procedure, all high-rate systematic codes, whether known or yet to be discovered, may be obtained by puncturing some low-rate original (systematic) code. Furthermore, since for any $R = b/(b+1)$ target code the branches of the high-rate code consist of b information digits and a single parity symbol, the original code need only be a $R = 1/2$ systematic code: the b information digits are directly issued from the information digits of the

Table 9. Original Codes that Yield the $R = 2/3$ Codes of Johannesson and Paaske, Perforation Pattern P_{01}, [22]

Original $R = 1/3$				Punctured $R = 2/3$			
M_0	G_1	G_2	G_3	M	G_{11} G_{12}	G_{21} G_{22}	G_{31} G_{32}
4	26	22	35	3	6 1	2 4	4 7
5	54	47	67	4	6 1	3 5	7 5
6	172	137	152	5	14 07	06 17	16 10
7	314	271	317	6	12 05	05 16	13 13
8	424	455	747	7	26 00	14 23	32 33
9	1634	1233	1431	8	32 13	05 33	25 22
10	3162	2553	3612	9	54 25	16 71	66 60
11	6732	4617	7153	10	53 36	23 53	51 67
12	17444	11051	17457	11	162 064	054 101	156 163

Table 10. Original Codes that Yield the $R = 2/3$ Codes of Johannesson and Paaske, Perforation Pattern P_{01}, [22]

Original $R = 1/3$				Punctured $R = 2/3$			
M_0	G_1	G_2	G_3	M	G_{11} G_{12}	G_{21} G_{22}	G_{31} G_{32}
16	377052	221320	314321	12	740 367	260 414	520 515
16	274100	233221	331745	13	710 140	260 545	670 533
16	163376	101657	153006	14	337 127	023 237	342 221
16	370414	203175	321523	15	722 302	054 457	642 435
18	1277142	1144571	1526370	16	1750 0165	0514 1235	1734 1054
18	1066424	1373146	1471265	17	1266 0140	0652 1752	1270 1307
19	2667576	2153625	3502436	18	1567 0337	0367 1230	1066 1603
20	4600614	4773271	6275153	19	2422 0412	1674 2745	2356 2711
20	12400344	13365473	15646505	20	3414 0005	1625 3367	3673 2440
22	24613606	22226172	35045621	21	6562 0431	2316 4454	4160 7225
24	117356622	126100341	151373474	22	13764 03251	02430 16011	14654 11766
24	106172264	130463065	141102467	23	12346 01314	05250 14247	10412 11067

Table 11. Original Codes that Yield the $R = 2/3$ Codes of Johannesson and Paaske, Perforation Pattern P_{01}, [22]

Original $R = 1/3$				Punctured $R = 2/3$			
					G_{11}	G_{21}	G_{31}
M_0	G_1	G_2	G_3	M	G_{12}	G_{22}	G_{32}
3	16	11	15	2	3	1	3
					1	2	2
4	22	22	37	3	4	2	6
					1	4	7
6	72	43	72	4	7	1	4
					2	5	7
7	132	112	177	5	14	06	16
					03	10	17
8	362	266	373	6	15	06	15
					06	15	17
9	552	457	736	7	30	16	26
					07	23	36
10	2146	2512	3355	9	52	06	74
					05	70	53
11	7432	5163	7026	10	63	15	46
					32	65	61

Table 12. Original Codes that Yield the $R = 3/4$ Codes of Johannesson and Paaske, Perforation Pattern P_{02}, [22]

Original $R = 1/4$					Punctured $R = 3/4$				
						G_{11}	G_{21}	G_{31}	G_{41}
						G_{12}	G_{22}	G_{32}	G_{42}
M_0	G_1	G_2	G_3	G_4	M	G_{13}	G_{23}	G_{33}	G_{43}
6	100	170	125	161	3	4	4	4	4
						0	6	2	4
						0	2	5	5
7	224	270	206	357	5	6	2	2	6
						1	6	0	7
						0	2	5	5
8	750	512	446	731	6	6	1	0	7
						3	4	1	6
						2	3	7	4
10	2274	2170	3262	3411	8	16	06	04	10
						03	12	00	13
						01	02	17	10
11	6230	4426	4711	7724	9	10	03	07	14
						01	15	04	16
						07	00	14	15

original code and the single parity symbol is obtained by puncturing all but one of the b remaining code symbols at the output of the original encoder. Equivalently, the perforation matrix has two rows, whereby the first one is filled with 1s and the second one is filled with 0s except at one position.

Using this procedure, Haccoun and Begin [22] have obtained the original codes and associated perforation matrices allowing to duplicate all the systematic codes of Hagenauer [24]. These codes are listed in Table 13. The possibility of generating these codes by perforation

Table 13. Systematic Punctured Codes that Duplicate the Codes of Hagenauer. G2 is Given in Octal [22]

R	P	G_2
2/3	1 1	33275606556377737
	0 1	
3/4	1 1 1	756730246717030774725
	0 0 1	
4/5	1 1 1 1	747546446652113345672547 5223
	0 0 0 1	
5/6	1 1 1 1 1	171751131171227722336701 06777
	0 0 0 0 1	
7/8	1 1 1 1 1 1 1	177363445377401454137543755 3121
	0 0 0 0 0 0 1	

allows once again their easy and practical decoding by any sequential decoding algorithm. However, just like the nonsystematic codes, the systematic punctured codes found here are optimal but do not readily lend themselves to variable-rate or rate-compatible decoding.

6. CONCLUSION

In this article, we have presented the basic notions, properties and error performances of high-rate punctured convolutional codes. These high-rate codes, which are derived from well-known optimal low-rate codes and a judicious choice of the perforation patterns, are no more complex to encode and decode than low-rate codes, yielding easy implementations of high coding rate codecs as well as variable-rate and rate-compatible coding and decoding. Extensive lists of both short- and long-memory length punctured codes have been provided together with up to the first 10 terms of their weight spectra and their bit error probability bounds. The substantial advantages of using high-rate punctured codes over the usual high-rate codes open the way for powerful, versatile, and yet practical implementations of variable-rate codecs extending from very low to very high coding rates.

BIOGRAPHY

David Haccoun received the Engineer and B.Sc.Ap. degrees (Magna Cum Laude) in engineering physics from École Polytechnique de Montréal, Canada, in 1965; the S.M. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, Massachusetts, in 1966; and the Ph.D. degree in electrical engineering from McGill University, Montréal, Canada, in 1974.

Since 1966 he has been with the Department of Electrical Engineering, École Polytechnique de Montréal, where he has been Professor of Electrical Engineering since 1980 and was the (founding) Head of the Communication and Computer Section from 1980 to 1996. He was a Research Visiting Professor at several universities in Canada and in France. Dr. Haccoun is involved in teaching at both undergraduate and graduate levels, conducting research, and

performing consulting work for both government agencies and industries.

His current research interests include the theory and applications of error-control coding, mobile and personal communications, and digital communications systems by satellite. He is the author or coauthor of a large number of journal and conference papers in these areas. He holds a patent on an error control technique and is a coauthor of the books *The communications Handbook* (CRC press and IEEE Press, 1997 and 2001), and *Digital Communications by Satellite: Modulation, Multiple-Access and Coding* (New York : Wiley, 1981). A Japanese translation of that book was published in 1984.

BIBLIOGRAPHY

1. A. J. Viterbi, Convolutional codes and their performance in communications systems, *IEEE Trans. Commun. Technol.* **COM-19** (1971).
2. I. M. Jacobs, Practical applications of coding, *IEEE Trans. Inform. Theory* **IT-20**: 305–310 (1974).
3. W. W. Wu, D. Haccoun, R. Peile, and Y. Hirata, Coding for satellite communication, *IEEE J. Select. Areas. Commun.* **SAC-5**: 724–748 (1987).
4. C. Berrou, A. Glavieux, and P. Thitimasjhima, Near Shannon Limit Error Correcting Coding and Decoding: Turbo Codes, Proceedings of ICC'93, pp. 1064–1070 (1993).
5. D. Divsalar and F. Pollara, Turbo codes for deep-space communications, *TDA Progress Report* **42–120**: 29–39 (1995).
6. J. B. Cain, G. C. Clark, and J. Geist, Punctured convolutional codes of rate $(n-1)/n$ and simplified maximum likelihood decoding, *IEEE Trans. Inform. Theory* **IT-25**: 97–100 (1979).
7. Y. Yasuda, Y. Hirata, K. Nakamura, and S. Otani, Development of a variable-rate Viterbi decoder and its performance characteristics, 6th Int. Conf. Digital Satellite Commun., Phoenix, Arizona, Sept. 1983.
8. Y. Yasuda, K. Kashiki, and Y. Hirata, High-rate punctured convolutional codes for soft decision Viterbi decoding, *IEEE Trans. Commun.* **COM-32**: 315–319 (1984).
9. J. Hagenauer, Rate compatible punctured convolutional codes and their applications, *IEEE Trans. Commun.* **36**: 389–400 (1988).
10. R. M. Fano, A heuristic discussion of probabilistic decoding, *IEEE Trans. Inform. Theory* **IT-9** (1962).
11. E. Paaske, Short binary convolutional codes with maximal free distance for rates 2/3 and 3/4, *IEEE Trans. Inform. Theory* **IT-20**: 683–686 (1974).
12. R. Johannesson and E. Paaske, Further results on binary convolutional codes with an optimum distance profile, *IEEE Trans. Inform. Theory* **IT-24**: 264–268 (1978).
13. F. Jelinek, A fast sequential decoding algorithm using a stack, *IBM J. Res. Develop.* **13**: 675–685 (1969).
14. K. Zigangirov, Some sequential decoding procedures, *Problemii Peradachi Informatsii* **2**: 13–15 (1966).
15. D. Haccoun and M. J. Ferguson, Generalized stack algorithms for decoding convolutional codes, *IEEE Trans. Inform. Theory* **IT-21**: 638–651 (1975).
16. J. P. Odenwalder, Optimal decoding of convolutional codes, Ph.D. dissertation, Dept. Elect. Eng., U.C.L.A., Los Angeles, 1970.
17. P. Montreuil, Algorithmes de determination de spectres des codes convolutionnels, M.Sc.A. thesis, Dep. Elect. Eng., Ecole Polytechnique de Montreal, 1987.
18. D. Haccoun and P. Montreuil, Weight spectrum determination of convolutional codes, to be submitted to *IEEE Trans. Commun.*, Book of Abstracts, 1988 Int. Symp. Inform. Theory, Kobe, Japan, June 1988, 49–50.
19. K. Larsen, Short convolutional codes with maximal free distance for rates 1/2, 1/3, and 1/4, *IEEE Trans. Inform. Theory* **IT-19**: 371–372 (1973).
20. G. Begin and D. Haccoun, Performance of sequential decoding of high rate punctured convolutional codes, *IEEE Trans. Commun.* **42**: 996–978 (1994).
21. D. Haccoun and G. Begin, High-rate punctured convolutional codes for Viterbi and sequential decoding, *IEEE Trans. Commun.* **37**: 1113–1125 (1989).
22. G. Begin and D. Haccoun, High-rate punctured convolutional codes: structure properties and construction technique, *IEEE Trans. Commun.* **37**: 1381–1385 (1989).
23. G. Begin, D. Haccoun, and C. Paquin, Further results on high-rate punctures convolutional codes for Viterbi and sequential decoding, *IEEE Trans. Commun.* **38**: 1922–1928 (1990).
24. J. Hagenauer, High rate convolutional codes with good profiles, *IEEE Trans. Inform. Theory* **IT-23**: 615–618 (1977).

HIGH-SPEED PHOTODETECTORS FOR OPTICAL COMMUNICATIONS

M. Selim Ünlü
Olufemi Dosunmu
Matthew Emsley
Boston University
Boston, Massachusetts

1. INTRODUCTION

The capability to detect, quantify, and analyze an optical signal is the first requirement of any optical system. In optical communication systems, the detector is a crucial element whose function is to convert the optical signal at the receiver end into an electrical signal, which is then amplified and processed to extract the information content. The performance characteristics of the photodetector, therefore, determines the requirements for the received optical power and dictates the overall system performance along with other system parameters such as the allowed optical attenuation, and thus the length of the transmission channel.

Progress in optical communications and processing requires simultaneous development in light sources, interconnects, and photodetectors. While optical fibers have been developed for nearly ideal links for optical signal transmission, semiconductors have become the material of choice for optical sources and detectors, primarily because of their well-established technology, fast electrical response, and optical generation and absorption properties. Many years of research on semiconductor devices have led to the development of high-performance photodetectors

in all of the relevant wavelengths throughout the visible and near-infrared spectra. The choice of wavelengths in optical communications is driven by limitations relating to the availability of suitable transmission media and, to a lesser extent, light-emitting devices (LEDs). In general, photodetectors are not the limiting factors. As a result, discussion of photodetectors is usually limited to one or two chapters in books on optoelectronic devices [1–4] or optical communications [5,6] with only a few dedicated books on photodetectors [7–9].

1.1. Optical Communication Wavelengths

The early development work on optical fiber waveguides focused on the 0.8–0.9-μm wavelength range because the first semiconductor optical sources, based in GaAs/AlAs alloys, operated in this region. As silica fibers were further refined, however, it was discovered that transmission at longer wavelengths (1.3–1.6 μm) would result in lower losses [6]. The rapid development of long wavelength fibers has led to attenuation values as low as 0.2 dB/km, which is very close to the theoretical limit for silicate glass fiber [10]. In addition to the intrinsic absorption due to electronic transitions at short wavelengths (high photon energies) and interaction of photons with molecular vibrations within the glass at long (infrared) wavelengths, absorption due to impurities in the silica host, most notably water incorporated into the glass as hydroxyl or OH ions, limits the transmission properties of the optical fiber. Furthermore, Rayleigh scattering resulting from the unavoidable inhomogeneities in glass density, manifesting themselves as subwavelength refractive-index fluctuations, represents the dominant loss mechanism at short wavelengths. The overall typical silica fiber loss has the well-known form given in Fig. 1. Two of the important communication wavelength regions are clearly identifiable as 1.3 and 1.55 μm as a direct result of the fiber attenuation characteristics. The third wavelength region we will consider is at 0.85 μm, due to the advent of GaAs-based light emitters and detectors.

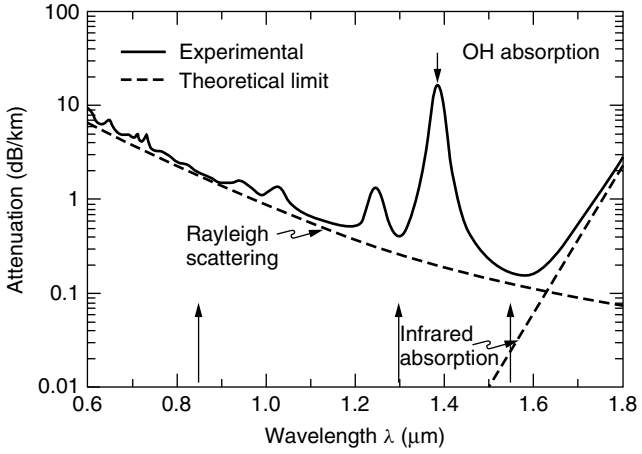


Figure 1. Measured attenuation and theoretical limits in silica fibers (see Ref. 1, for example). The three communication wavelengths we consider are indicated by arrows.

1.2. Basic Performance Requirements

Before we continue with specific photodetector structures, we must first discuss some of the performance requirements. Since the relevant wavelengths for optical communications are in the near-infrared region of the spectrum, both external and internal photoemission of electrons can be utilized to convert the optical signal to electrical signal. External photoemission devices such as vacuum tubes not only are very bulky but also require high voltages, and therefore cannot be used for optical communication applications. Internal photoemission devices, especially semiconductor photodiodes, provide high performance in compact and relatively inexpensive structures. Semiconductor photodetectors are made in a variety of materials such as silicon, germanium and alloys of III–V compounds, and satisfy many of the important performance and compatibility requirements:

1. Compact size for efficient coupling with fibers and easy packaging
2. High responsivity (efficiency) to produce a maximum electrical signal for given optical power
3. Wide spectral coverage through the use of different materials to allow for photodetectors in all communication wavelengths
4. Short response time to operate at high bit rates (wide bandwidth)
5. Low operating (bias) voltage to be compatible with electronic circuits
6. Low noise operation to minimize the received power requirements
7. Low cost to reduce the overall cost of the communication link
8. Reliability (long mean time to failure) to prevent failure of the system
9. Stability of performance characteristics working within a variety of ambient conditions, especially over a wide range of temperatures
10. Good uniformity of performance parameters to allow for batch production

2. FUNDAMENTAL PROPERTIES AND DEFINITIONS

In this section, we describe various definitions relating to the performance of photodetectors, including detection efficiency and noise, and identify suitable materials for a given operation wavelength. We will consider the high-speed properties in the following section. A typical photodetection process in semiconductors can be summarized in three steps:

1. Photons are absorbed in the material, resulting in the generation of mobile charge carriers (electron–hole pairs).
2. The charge carriers drift under an internal electric field.
3. The carriers are collected at the contacts and detection is completed with an electrical response in the external circuit.

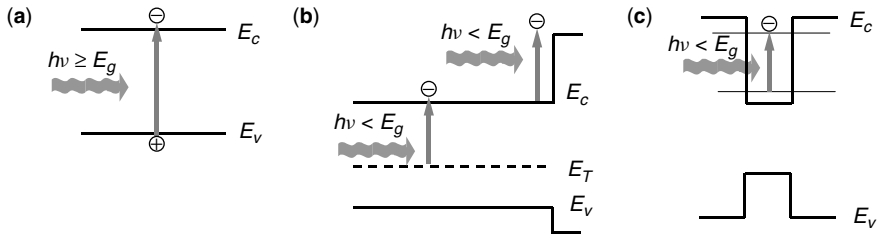


Figure 2. Various absorption mechanisms for photodetection: (**a**) intrinsic (band-to -band); (**b**) extrinsic utilizing an impurity level or free-carrier absorption; (**c**) intersubband transition in a quantum well.

2.1. Optical Absorption and Quantum Efficiency

The first requirement of photodetection — absorption of light — implies that photons have sufficient energy to excite a charge carrier inside the semiconductor. The most common absorption event results in the generation of an electron–hole pair in typical intrinsic photodetectors. In contrast, an *extrinsic* photodetector responds to photons of energy less than the bandgap energy. In these photodetectors, the absorption of photons result in transitions to or from impurity levels, or between the subband energies in quantum wells as is depicted in Fig. 2.

As will be shown below, thin semiconductor devices for high-speed operation will be required and, therefore, it is crucial to have materials that absorb the incident light very quickly. The absorption of photons at a particular wavelength is dependent on the absorption coefficient α, which is the measure of how fast the photon flux Φ decays in the material:

$$\frac{d\Phi}{dx} = -\alpha\Phi \Rightarrow \Phi(x) = \Phi(0)\exp(-\alpha x) \tag{1}$$

Therefore, the amount of photon flux that is absorbed in a material of thickness (or length) L and absorption coefficient α, can be expressed as [11]

$$\Delta\Phi = \Phi(0)\cdot(1-R)\cdot[1-\exp(-\alpha L)] \tag{2}$$

where R is the power reflectivity at the incidence surface.

The *quantum efficiency* $(0 \leq \eta \leq 1)$ of a photodetector is the probability that an incident photon will create or excite a charge carrier that contributes to the detected photocurrent. When many photons are present, we consider the ratio of the detected flux to the incident flux of photons. Assuming that all photoexcited carriers contribute to the photocurrent, we obtain

$$\eta = \frac{\Delta\Phi}{\Phi(0)} \tag{3}$$

$$\eta = (1-R)\cdot[1-\exp(-\alpha L)] \tag{4}$$

2.2. Material Selection

As can be deduced from the equations above, it is necessary to have a large α to realize high-efficiency photodetectors. To capture most of the photons for single-pass absorption, a semiconductor with a thickness of several absorption depth lengths ($L_{\text{abs}} = 1/\alpha$) is needed. Figure 3 shows the wavelength dependence of

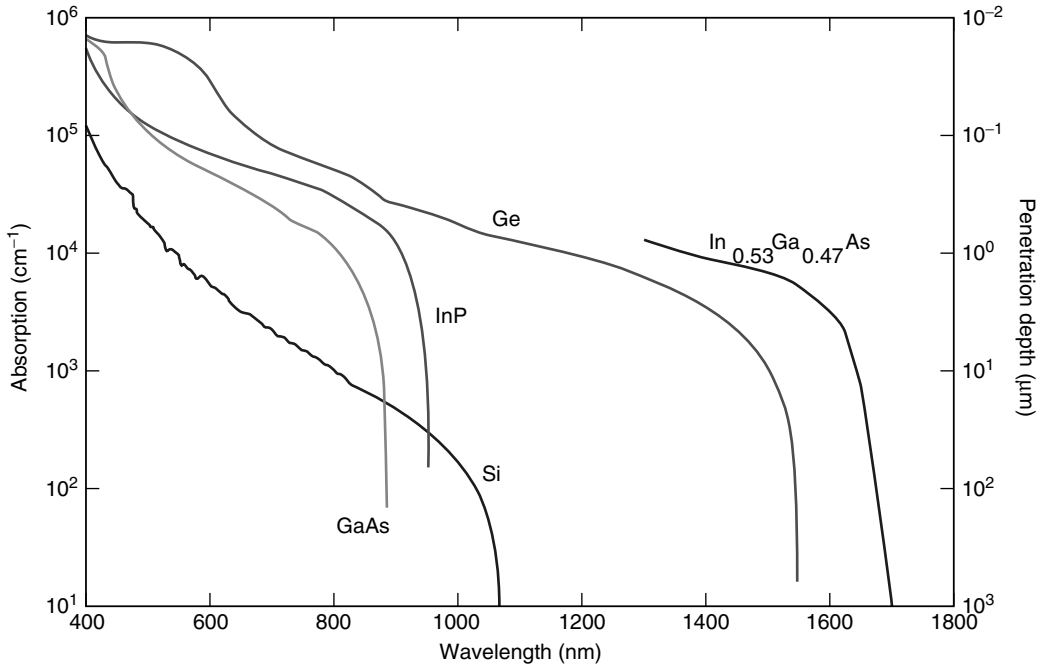


Figure 3. Absorption coefficient for various semiconductors.

α and absorption (or penetration) depth for a variety of semiconductor materials. For very small α, L_{abs} is very large; this results not only in impractical detector structures but also in detectors with a necessarily slow response. At the other extreme, L_{abs} is very small; that is, all the radiation is absorbed in the immediate vicinity of the surface. For typical high-speed semiconductor photodetectors, it is desirable to have an absorption region between a fraction of a micron and a few micrometers in length; specifically, α should be about 10^4 cm^{-1}. Therefore, the high-speed application of the most commonly used detector material, silicon, is limited to the visible wavelengths ($0.4\ \mu\text{m} < \lambda < 0.7\ \mu\text{m}$). While the absorption spectrum of an indirect bandgap semiconductor like germanium covers all the optical communication wavelengths, direct bandgap semiconductors such as III–V compounds are more commonly used for high-speed applications. For example, GaAs has a cutoff wavelength of $\lambda_c = 0.87\ \mu$m and is ideal for optical communications at 0.85 μm. At longer wavelengths, ternary compounds such as InGa(Al)As and quaternary materials such as InGa(Al)AsP and a variety of their combinations are used. An important consideration when various semiconductor materials are used together to form heterostructures is the lattice matching of the different constituents and the availability of a suitable substrate. Figure 4 shows the lattice constants and bandgap energies of various compound semiconductors. For most of the photodetector structures designed to operate at 1.3 and 1.55 μm wavelengths, compound semiconductors lattice-matched to InP are used.

2.3. Responsivity

In an analog photodetection system η is recast into a new variable, namely, *responsivity*, or the ratio of the photocurrent (in amperes) to the incident light (in watts):

$$\Re = \frac{\text{photocurrent}}{\text{incident optical power}} = \frac{I_p}{P_0} \quad \text{(A/W)} \tag{5}$$

From here, a simple relationship between quantum efficiency and responsivity can be developed:

$$\{P_0 = \Phi(0)h\nu \text{ and } I_p = q \cdot \Delta\Phi\} \Rightarrow \Re = \frac{q\eta}{h\nu} \tag{6}$$

Responsivity ($\Re$) can also be expressed in terms of wavelength (λ):

$$\nu = \frac{c}{\lambda} \Rightarrow \Re = \frac{q\lambda\eta}{hc} \simeq \frac{\eta \cdot \lambda(\mu\text{ m})}{1.24} \tag{7}$$

It should be noted that the responsivity is directly proportional to η at a given wavelength. The ideal responsivity versus λ is illustrated in Fig. 5 together with the response of a generic InGaAs photodiode showing long- and short-wavelength cutoff and deviation from the ideal behavior.

2.4. Noise Performance and Detectivity

The responsivity equation $I_p = \Re P_0$ suggests that the transformation from optical power to photocurrent is deterministic or noise-free. In reality, even in an ideal photoreceiver, two fundamental noise mechanisms [12] — *shot noise* [13] and *thermal noise* [14,15] — lead to fluctuations in the current even with constant optical input. The overall sensitivity of the photodetector is determined by these random fluctuations of current that occur in the presence and absence of an incident optical signal. Various figures of merit have been developed to assess the noise performance for photodetectors. Although these are not always

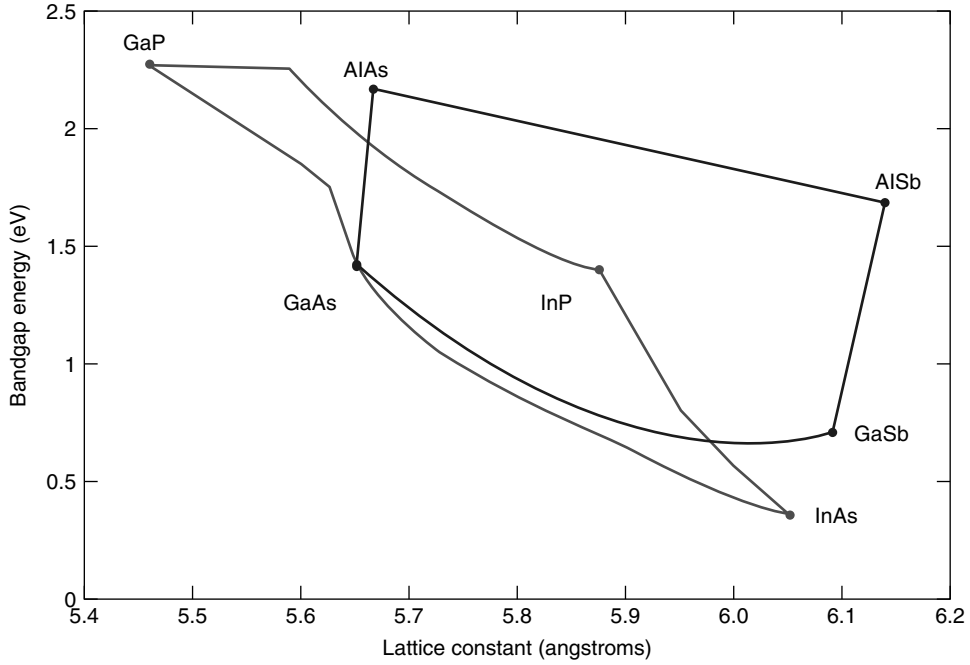


Figure 4. Correlation between lattice constant and bandgap energy for various semiconductor materials.

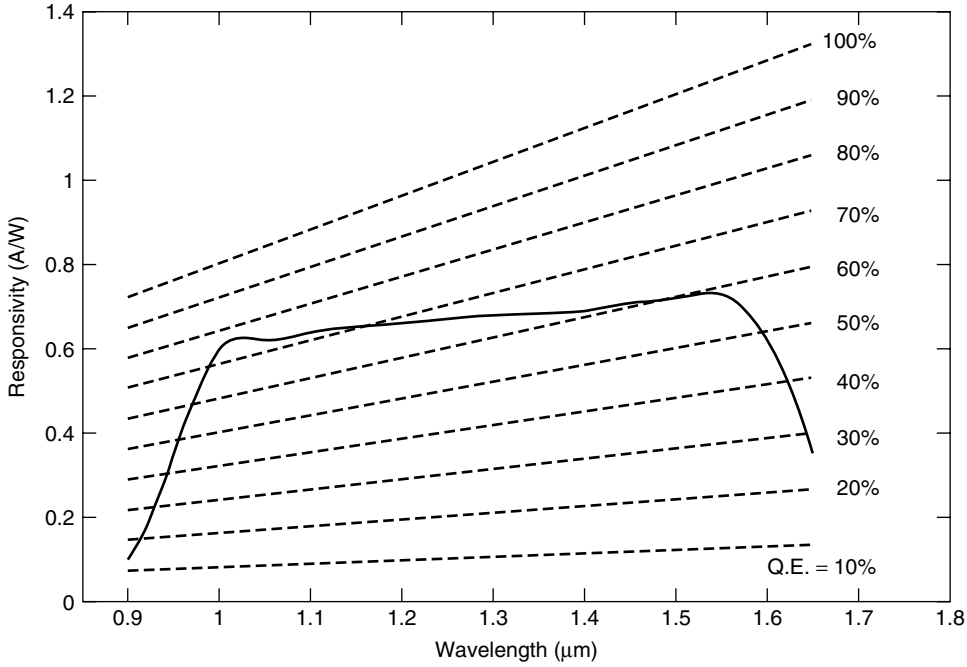


Figure 5. Responsivity versus wavelength for ideal case (dashed) and a generic InGaAs photodiode (solid).

appropriate for high-speed photodetectors for optical communications, it is instructive to discuss the most commonly used definitions.

The *noise equivalent power* (NEP) is the input optical power that results in unity signal-to-noise ratio (SNR = 1). The *detectivity* D is defined as the inverse of NEP ($D = 1/\text{NEP}$), while *specific detectivity* D^* (D-star) incorporates the area of the photodetector A and the bandwidth Δf of the signal:

$$D^* = \frac{(A \times \Delta f)^{1/2}}{\text{NEP}} \tag{8}$$

For thermal-noise-limited photodetectors (most semiconductor detectors without internal gain), the noise power (rms) is given by

$$\sigma_T^2 = \langle i_T^2(t) \rangle = \left(\frac{4k_B T}{R_L} \right) \Delta f \tag{9}$$

where R_L is the load resistance, T is temperature in degrees Kelvin and k_B is the Boltzmann constant. Therefore, the NEP has the units of W/Hz$^{1/2}$ as shown below:

$$\text{SNR} = \frac{I_p^2}{\sigma^2} = 1 = \frac{R_L \Re^2}{4k_B T F_n \Delta f} P_{\text{in}}^2 \tag{10}$$

$$\text{NEP} = \frac{P_{\text{in}}}{\sqrt{\Delta f}} = \frac{1}{\Re} \sqrt{\frac{4k_B T F_n}{R_L}} = \frac{h\nu}{\eta q} \sqrt{\frac{4k_B T F_n}{R_L}} \tag{11}$$

where P_{in} is the input optical power and F_n represents the noise figure. The NEP or detectivity can be used to estimate the optical power needed to obtain a specific value of SNR if the bandwidth Δf is known. In the case of shot noise limit, the noise power scales with the total current ($I_p + I_d$, photocurrent + dark current) and bandwidth Δf:

$$\sigma_s^2 = \langle i_s^2(t) \rangle = 2q(I_p + I_d)\Delta f \tag{12}$$

where q is the electronic charge. For high-gain detectors such as avalanche photodetectors, or at high incident power limit, shot noise is much greater than thermal noise:

$$\text{SNR} = \frac{\Re P_{\text{in}}}{2q\Delta f} = \frac{\eta P_{\text{in}}}{2h\nu \Delta f} \tag{13}$$

This further neglects the dark current:

$$\text{SNR} = \frac{I_p^2}{\sigma^2} = \frac{(\Re P_{\text{in}})^2}{2q(\Re P_{\text{in}} + I_d)\Delta f} \tag{14}$$

In this case, the SNR scales linearly with the input optical power and the NEP would not have the traditional units of W/Hz$^{1/2}$ and is not traditionally used. Instead, one can calculate and refer to the power (or number of photons) in a "one" bit at given bit rate and SNR values.

3. HIGH-SPEED PERFORMANCE AND LIMITATIONS

The most common high-speed photodetector response limitations include drift, diffusion, capacitance, and charge-trapping limitation. The type of detector used, or even changes in the detector's geometry, can result in one or more of these limitations, severely degrading the overall bandwidth of the photodetector. In addition, as will be discussed later in this section, the direct relationship between the detector responsivity and bandwidth often

serves as an obstacle in designing detectors that are both fast and efficient.

3.1. Transit-Time Limitation

Drift, or transit-time, limitation is directly related to the time needed for photogenerated carriers to traverse the depletion region. To fully understand the mechanism behind this limitation, one must first examine the transient response of photogenerated carriers within a typical PN junction photodiode. A cross-sectional view of a basic PN junction photodiode under reverse bias is illustrated in Fig. 6. In this basic design, photogeneration occurs throughout the photodiode structure. In particular, those carriers photogenerated within the depletion region are swept across at or close to their saturation velocities due to the high electric field within the depletion region.

Assuming that no absorption takes place within the neutral P and N regions of the photodiode, an incident optical pulse generates carriers only within the depletion region. The length of the depletion region, L, can be determined from the following expression:

$$L = \sqrt{\frac{2\varepsilon_s(V_0 - V_a)}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)} \tag{15}$$

Here, ε_s and V_0 represent the semiconductor permittivity and contact potential at the junction, while N_a and N_d represent the doping concentration in the P and N regions, respectively. The external bias, V_a, is negative when reverse-biased, and positive when forward-biased. As is illustrated in Fig. 6, these carriers are then swept to the opposite ends of the depletion region both by the built-in potential induced by the ionized dopants as well as any externally applied potential. At the same time, current at the photodiode contacts is "induced" by the movement of charge within the depletion region. The induced current can be seen as displacement current, where any delay between the movement of charge through the depletion region and the induced current at the contacts is set by the electromagnetic propagation time through the device, which will always be less than 10 fs [16].

Figure 7 shows that as an incremental sheet of photogenerated carriers moves from one particular point

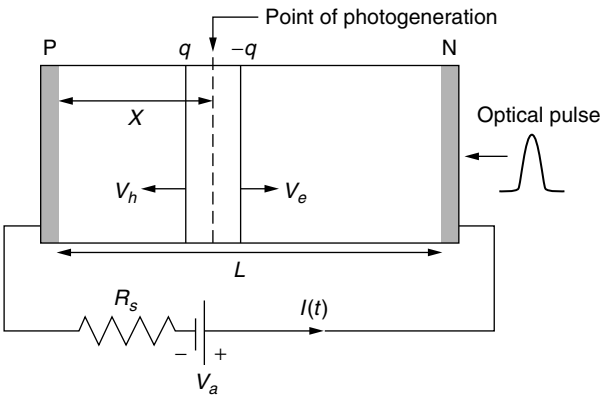


Figure 6. Movement of an incremental sheet of photogenerated carriers within a depletion region.

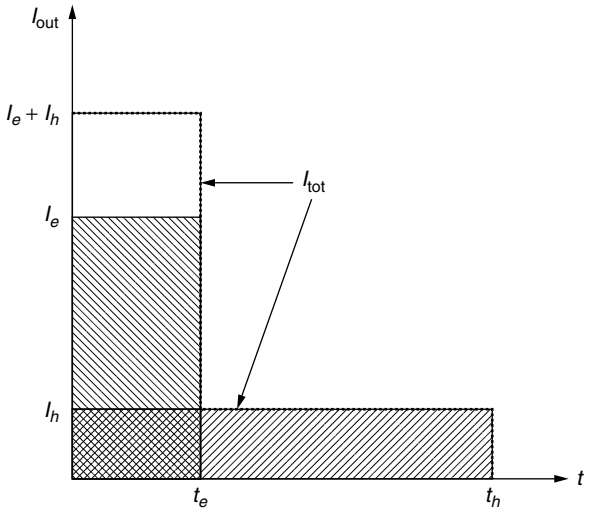


Figure 7. Current induced by incremental sheet of electrons and holes.

within the depletion region to its edge, the ideal current waveform would be a step function whose magnitude is determined by the electron and hole velocities, as well as the width of the depletion region:

$$I_{tot} = I_e(t) + I_h(t) \tag{16}$$

$$I_e(t) = \frac{q}{L}v_e, \text{ for } 0 < t < t_e \tag{17}$$

$$I_h(t) = \frac{q}{L}v_h, \text{ for } 0 < t < t_h \tag{18}$$

where q represents the electron charge and $v_e(v_h)$ and $t_e(t_h)$ represent the electron (hole) terminal velocities and transport times within the depletion region, respectively. In this figure, it is assumed that both carriers immediately reach their terminal velocities, and the electron velocity is higher than that of the holes. Now, taking into account carriers generated throughout the depletion region, the resulting current waveforms induced by all the photogenerated electrons and holes would be as shown in Fig. 8. Here, $t_e(t_h)$ represents the time required for the electron (hole) generated furthest from its respective depletion edge to reach that edge.

The response speed of any photodiode is limited, in part, by the speed of the slower carrier; in this case, holes. The temporal response of a typical high-speed photodiode, illustrated in Fig. 9, represents the combination of responses by both electrons and holes. Because current at the detector contacts is induced from the time the carriers are photogenerated to the time they reach the depletion region boundaries, the response current does not end until both carriers reach their respective depletion edges. The 3-dB bandwidth of a transit-time-limited photodetector can be expressed as [17]

$$f_{tr} = 0.4\frac{v}{L} \tag{19}$$

where v represents the speed of the slower carrier.

One way to reduce the effects of this transit-time limitation is to make only part of the depletion region

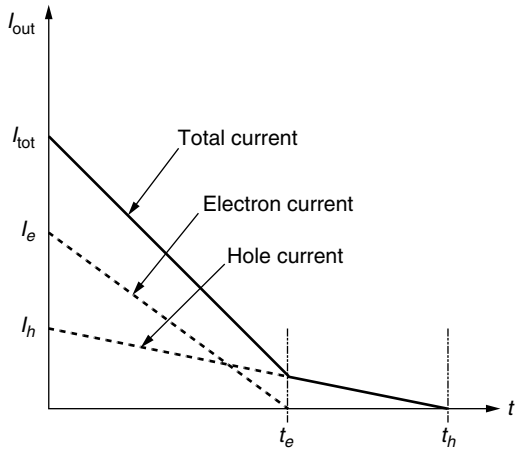


Figure 8. Current induced by all photogenerated electrons and holes.

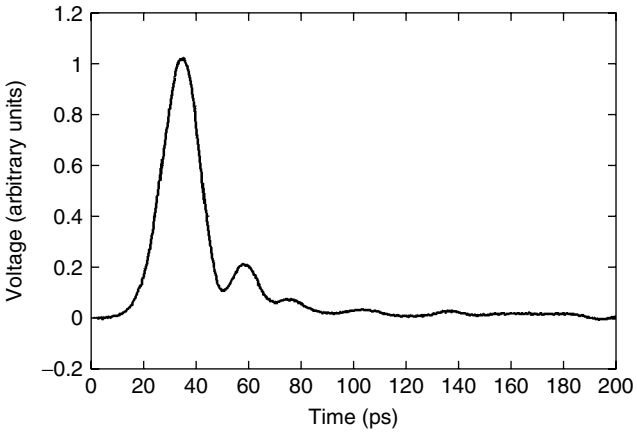


Figure 9. Typical temporal response of a high speed photodetector.

absorptive over the wavelength range of interest. As a result, photogeneration does not occur at the extremes of the depletion region, thus reducing the duration of induced current at the contacts. In addition, the absorption layer should be optimally positioned such that the average electron and hole transit times are equal.

3.2. Capacitance Limitation

Another possible way to reduce the effects of transit-time limitation would be to reduce the length of the depletion region itself. However, a reduction of the depletion length will also result in an increase of the junction capacitance. The bandwidth of a junction-capacitance-limited photodetector can be formulated by the following expression:

$$f_{RC} = \frac{L}{2\pi R_L \varepsilon A} \tag{20}$$

where R_L is the load resistance, ε is the semiconductor permittivity, and A represents the detector's cross-sectional area. To minimize this limitation, the photodetector should be designed such that the depletion length is not so small that capacitance limitations are dominant. At the same time, however, the length should not be increased to the point where transit-time limitations become an overwhelming limiting factor. Basically, there exists an optimum point where the combined effect of the transit-time and capacitance limitations is at a minimum. In addition to junction capacitance, other capacitances relating to the detector contact pad and other external elements can also serve to degrade the overall detector bandwidth. Called *parasitic capacitance*, they can be minimized by using an airbridge directly from the detector to the contact pad [18]. A top-view image of such a photodetector can be seen in Fig. 10.

3.3. Diffusion Limitation

If a photodetector is designed such that all the semiconductor regions, including the nondepleted or neutral regions, are absorptive over the wavelength range of interest, then limitations due to carrier diffusion will dominate any bandwidth limitations imposed on the detector. Unlike the depletion region, there is no potential drop across the neutral regions to drive the photogenerated minority carriers in any particular direction. Instead, the movement of carriers within these regions is dictated by the diffusion process, where a net current induced by the photogenerated minority carriers in the neutral regions does not occur until some of these carriers reach the depletion edge and are swept across to the other end by the high electric field. Here, the average amount of time needed for the photogenerated carriers within the nondepleted regions to reach the depletion edge is dictated in part by the thickness of the region, which can lead to a time period on the order of 10 ns or more [5]. As a result, the temporal response of a diffusion-limited photodetector will exhibit a "tail" at the end of the pulse as long as the diffusion time period. Limitations caused by carrier diffusion can be minimized, even totally eliminated, simply by making the neutral regions nonabsorptive.

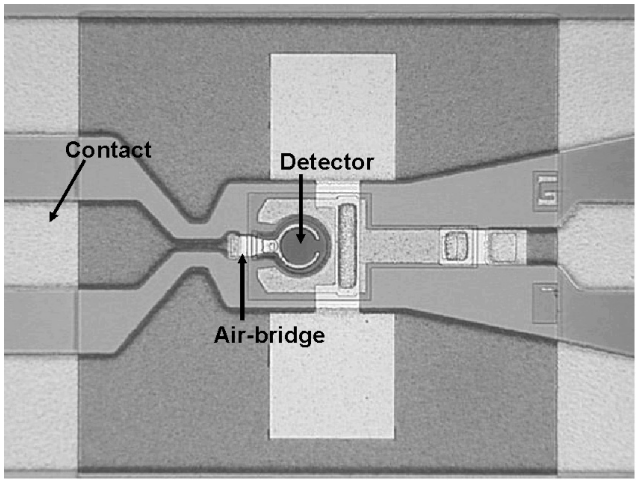


Figure 10. Top view of high-speed detector with airbridge and contacts for high-speed probe.

3.4. Charge-Trapping Limitation

Making only certain regions within a photodetector absorptive over a given wavelength range involves forming heterojunctions, or junctions between two different semiconductor materials, where one possesses a much higher absorption coefficient than the other over the wavelength range of interest. When two different semiconductors are brought abruptly into contact, however, the differences in their respective valence and conduction bands will result in the temporary storage of charge at the interface. This effect, referred to as *charge trapping*, ultimately impedes the transport of charge through the depletion region, reducing the device bandwidth as a result. One effective way to reduce the amount of charge trapping in heterojunction-based devices is to compositionally grade the region around the heterojunction itself, thereby smoothing out any abruptness in the energy bands between the two semiconductor materials.

3.5. Bandwidth–Efficiency Product

Designing photodetectors for high-speed optical communications requires the optimization of both the bandwidth, or speed, as well as the quantum efficiency of the detection system. Unfortunately, there is often a trade-off between bandwidth and efficiency, especially when carries are collected along the same direction as the photons being absorbed. The bandwidth–efficiency (BWE) product serves as an important figure of merit for high-speed detectors.

Assuming that all other speed limitations are minimized, the BWE product for a transit-time-limited photodetector can be expressed as

$$\text{BWE} = f_{tr} * \eta = 0.4\frac{v}{L}(1-R)(1-e^{-\alpha L}) \tag{21}$$

For a thin photodetector, the transit-time-limited BWE product can be approximated as

$$\text{BWE} = f_{tr} * \eta = 0.4v\alpha(1-R) \tag{22}$$

which is independent of the depletion region length and depends only on material properties, presenting a fundamental limitation.

4. SEMICONDUCTOR PHOTODETECTORS

4.1. Classification and Structures

Semiconductor photodetectors can be classified into two major groups:

1. *Photovoltaics*. These detectors produce a voltage drop across its terminals when illuminated by an external optical source. A subcategory of photovoltaics is the photodiode, the most commonly used detector in high-speed optical communications. Photodiodes detect light through the photogeneration of carriers within the depletion region of a diode structure, typically under reverse bias. Below, we will discuss the most common types of photodiodes, including PIN, Schottky, and avalanche photodetectors (APDs).
2. *Photoconductors*. These detectors respond to an external optical stimulus through a change in their conductivity. The optical signal is detected by applying an external bias voltage and observing the change in current due to a change in resistivity. Unlike other types of photodetectors, photoconductors can exhibit gain, since the photoexcited carriers may contribute to the external current multiple times when the recombination lifetimes of the carriers are greater than their transit times [19].

It is also possible to categorize semiconductor photodetectors according to a variety of other properties. Below we emphasize different photodetector structures based on illumination geometry and collection of photoexcited carriers. It is important to distinguish between structures where photon absorption and carrier collection occur along the same direction, and those that do not. In the first case, there is a direct trade-off between the efficiency and speed of the photodetector, as discussed in the previous section.

Most common photodetectors are vertically illuminated and carriers are collected along the same direction as schematically shown in Fig. 11a. Alternatively, carrier collection can be in the lateral direction (Fig. 11b). In a conventional (one-pass) structure, the efficiency is limited by the thickness and absorption coefficient of the detector material. To increase the efficiency without requiring a thick absorption region in vertically illuminated photodetectors, a resonant cavity enhanced (RCE) structure can be utilized, effectively resulting in multiple passes through the detector (Fig. 11c). Photodetectors can be formed as optical waveguides in an edge-illuminated configuration as shown in Fig. 11d. In waveguide photodetectors (WGPD), light is absorbed along the waveguide and the carriers are collected in the transverse direction, permitting nearly independent

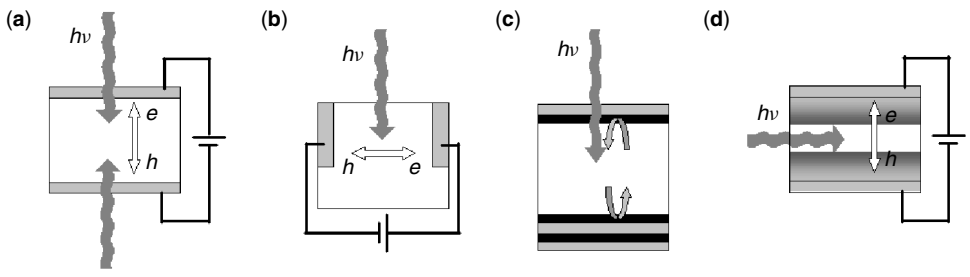


Figure 11. Photodetector structures: (**a**) vertically illuminated (top or bottom); (**b**) vertically illuminated with lateral carrier collections; (**c**) resonant cavity enhanced (RCE); (**d**) waveguide photodetector (edge-illuminated).

optimization of both the bandwidth and efficiency. A further refinement of the WGPD is the traveling-wave photodetector (TWPD), in which the optical signal and resulting electrical signal travel at the same speed, overcoming the capacitance limitations [20].

4.2. PIN Detectors

PN junctions are formed simply by bringing into contact n- (donors) and p-doped (acceptors) semiconductor regions [21]. When a junction between these two regions is formed, diffusion of the electrons from the n region and holes from the p region will cause a charge imbalance to form as a result of the uncompensated ions that are left behind. This charge imbalance will create an opposing electric field that will prevent further diffusion of electrons and holes, resulting in a region depleted of carriers at the junction. The length of the depletion was stated earlier in Eq. (15).

As was described earlier, photons absorbed in the depletion region will create electron–hole pairs that are separated by the built-in or externally applied field and drift across the depletion region, inducing current at the contacts until reaching the depletion edge. Conversely, photons that are absorbed in the neutral regions can contribute to the photocurrent if the minority carrier is sufficiently close to the depletion region, or if the diffusion length is long enough that recombination does not occur before it reaches the depletion region, where it will be swept across while inducing current at the contacts. The minority carrier moves by diffusion through the neutral region, which is an inherently slow process. Therefore, great care must be taken to inhibit this process by making the highly doped P and N regions extremely thin.

In general, the depletion region is designed to be as large as possible to increase the absorption of photons. For PN junctions, applying a reverse bias increases the depletion region. An alternative approach is a one-sided abrupt junction, formed between a highly doped $p(n)$ region and lightly doped $n(p)$ region. The depletion length is given by [22]

$$L = \sqrt{\frac{2\varepsilon_s(V_{bi} - V)}{qN_B}} \tag{23}$$

where N_B is the doping concentration on the lightly doped side. However, one-sided PN junctions suffer from high series resistance, which is detrimental to high-speed photodiodes. Increasing the background doping, N_B, in a PN junction decreases the transit time τ_d for the carriers by increasing the maximum field strength in the depletion region. Since the field is triangular, however, the mean electric field will be half the maximum E field:

$$\tau_d \approx \frac{L^2}{2V\mu} \tag{24}$$

where μ represents the carrier mobility.

PIN photodiodes, where "I" denotes intrinsic, improve on the conventional PN photodiode by having a large intrinsic region sandwiched between two heavily doped regions. Since the intrinsic region is depleted and is much larger than the depletion length in the highly doped p and n regions, it is sufficient to say that the field is at a maximum across the entire depletion length. Therefore, the mean electric field is twice that of a conventional PN junction and, correspondingly, the transit time is cut in half [7], assuming the carriers have not reached their saturation velocities:

$$\tau_d \approx \frac{L^2}{4V\mu} \tag{25}$$

An added benefit of the PIN photodiode is that one does not need to apply an external bias to deplete the intrinsic region, as it is usually fully depleted by the built-in potential. Also, having L much greater in length than the neutral regions results in a greatly reduced diffusion contribution to the photocurrent. PIN photodiodes can also have very highly doped p and n regions, thereby reducing access resistance as compared to the one-sided PN junction photodiode.

An alternative approach to making the neutral regions thin is the use of heterojunctions, or junctions between two different semiconductors. As was discussed earlier, the neutral regions can be made of materials that do not absorb at the operating wavelength. Therefore, carriers will be photogenerated only within the depletion region, thereby eliminating any diffusion current. For example, InGaAs-InP heterojunction PIN photodiodes operating at speeds in excess of 100 GHz have been reported [23,24].

4.3. MSM Photodetectors

Unlike most photodetectors used in high-speed optical communications, metal–semiconductor–metal (MSM) photodetectors are photoconductors. Photoconductors operate by illuminating a biased semiconductor material layer, which results in the creation of electron–hole pairs that raise the carrier concentration and, in turn, increase the conductivity as given by

$$\sigma = q(\mu_n n + \mu_p p) \tag{26}$$

where μ_n and μ_p are the electron and hole mobility, and n and p represent the electron and hole carrier concentrations, respectively. This increase in carrier concentration and subsequent increase in conductivity results in an increase of the photocurrent given by [4]

$$I_{ph} = \frac{q\eta GP}{h\nu} \tag{27}$$

where q is the electron charge, G is the photoconductor gain, and P represents the optical input power. The photoconductor gain G, which results from the difference between the transit time for the majority carrier and the recombination lifetime of the minority carrier, is given by

$$G = \frac{\tau}{\tau_{tr}} \tag{28}$$

where τ is the minority carrier lifetime and τ_{tr} is the majority carrier transit time. It is usually advantageous to have a high defect density in the photoconductor

absorption area so that the carrier lifetime is reduced, resulting in a reduction of the impulse response [9].

MSMs are important devices because they can be easily monolithically integrated into FET-based circuit technologies. MSMs, for example, have been demonstrated on InGaAs with bandwidths in excess of 40 GHz operating at 1.3 μm [25], and it has been shown that monolithic integration with InP-based devices is possible [26].

4.4. Schottky Photodetectors

A Schottky detector consists of a conductor in contact with a semiconductor, which results in a rectifying contact. As a result, a depletion region forms at the conductor/semiconductor junction, entirely on the semiconductor side. The conductor can be metal or silicide, while the semiconductor region is usually moderately doped ($\sim 10^{17}$ cm^{-3}) to prevent tunneling through a thin barrier. A generic energy band diagram of a Schottky junction is illustrated in Fig. 12.

In a basic Schottky detector, the depletion length W and semiconductor work function Φ_s are determined by the semiconductor doping, while the metal-to-semiconductor barrier height Φ_b is set by the high density of surface states formed at the metal/semiconductor interface, where the Fermi level is "pinned" at a certain position, regardless of doping. The semiconductor-to-metal barrier height Φ_v, on the other hand, can be controlled through doping, as is described through the following equation [21]:

$$\Phi_v = \Phi_m - \Phi_s \tag{29}$$

where Φ_m represents the metal work function.

In the case where the photon energy is greater than the semiconductor bandgap, the Schottky barrier behaves much like a one-sided PN junction; for photons incident on the metal side of the Schottky barrier, however, the metal must be thin such that it is effectively transparent. For energies less than the semiconductor bandgap, but greater than the barrier formed at the metal–semiconductor junction ($\Phi_v < h\nu < E_g$), photons incident on the semiconductor side will cross the region without any absorption and will excite carriers within the metal over the metal–semiconductor barrier.

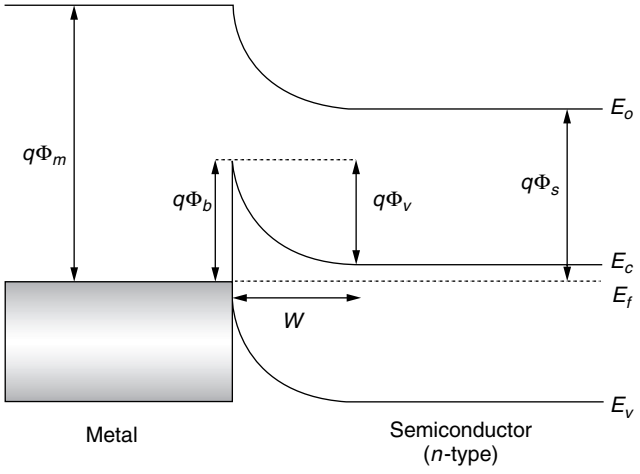


Figure 12. Energy band diagram of Schottky junction.

One advantage that Schottky detectors have over their PIN counterparts is their reduced contact resistance, resulting in a faster response. For example, Schottky detectors with measured 3-dB bandwidths well above 100 GHz have been reported [9,27]. An additional advantage is their simple material structure and relative ease in fabrication [18]; however, detector illumination is a common design issue. Because metals and silicides are very absorptive because of the effects of free-carrier absorption, one cannot design a traditional Schottky detector to be illuminated from the metal side, unless the metal layer is very thin such that reflection is minimized. Ignoring this issue would result in a photodiode with poor quantum efficiency.

4.5. Avalanche Photodetectors

Avalanche photodetectors are PN junctions with a large field applied across the depletion region so that electron–hole pairs created in this region will have enough energy to cause impact ionization, and in turn avalanche multiplication, while they are swept across the depletion length. Under this condition, a carrier that is generated by photon absorption is accelerated across the depletion region, lifting it to a kinetic energy state that is large enough to ionize the neighboring crystal lattice, resulting in new electron–hole pair generation. Likewise, these new pairs are accelerated and multiply with the production of more electron–hole pairs, until finally these carriers reach the contacts. This avalanche effect results in a multiplication of the photogenerated carriers and, in turn, an increase in the collected photocurrent. The multiplication factor M is used to refer to the total number of pairs produced by the initial photoelectron. This factor represents the internal gain of the photodiode; as a result, APDs can have quantum efficiencies greater than unity

$$M = \frac{1}{1 - (V/V_{\mathrm{BR}})^r} \tag{30}$$

where V_{BR} represents the breakdown voltage and r is a material dependent coefficient. Since APDs exhibit gain, they are highly sensitive to incident light and therefore are limited by shot noise, as a single incident photon has the potential to create many carrier pairs. Also, there is a finite time associated with the avalanche process, resulting in a slower impulse response. Regardless, APDs for optical communications have been demonstrated exhibiting internal gains in excess of 200 at 9 V reverse bias [28], as well as devices with BWE product of 90 GHz [29].

4.6. RCE Photodetectors

As discussed earlier, the important figure of merit for high speed photodetectors is the BWE product. That is, the transit time of the photogenerated carriers must be kept to a minimum while the absorption length must be sufficiently long so that a reasonable number of photons are absorbed and, in turn, carriers generated. Because these two quantities are inversely related, an increase in efficiency will result in the reduction of

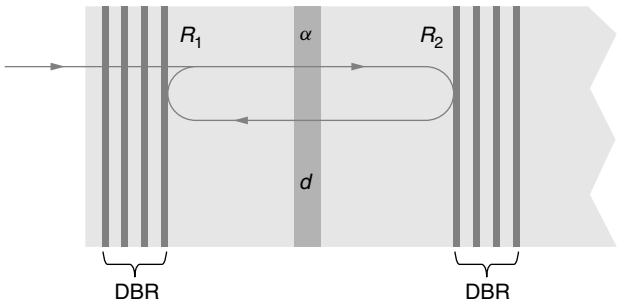


Figure 13. Simplified cross-sectional view of a typical RCE detector.

bandwidth. One way to circumvent this problem involves fabricating the photodetector within a Fabry–Perot cavity. Fabry–Perot cavities are formed by sandwiching a region between two reflecting surfaces, where light incident perpendicular to the cavity results in spectrally dependent constructive and destructive interference of the electric field inside the cavity. A cross-sectional illustration of a typical RCE photodetector can be seen in Fig. 13. The constructive interference results in an increase of the electric field amplitude at specific points within the cavity over a limited spectral range. Since high-speed optical communications typically rely on narrow-linewidth laser sources, fabricating photodetectors within a Fabry–Perot cavity is an ideal solution. For an RCE photodetector, a conventional PIN diode is placed within a Fabry–Perot cavity so that the electric field maximum occurs over the absorption region, thereby increasing the number of absorbed photons and, in turn, increasing the quantum efficiency. For a given absorption length, the quantum efficiency of the detector over a specific wavelength range is increased, while the transit time and bandwidth remain constant as compared to a conventional photodiode. This results in an overall increase in the bandwidth–efficiency product for the RCE photodiode. The peak quantum efficiency for an RCE photodetector is given by

$$\eta_{\max} = \left\{ \frac{(1 + R_2 e^{-\alpha d})}{(1 - \sqrt{R_1 R_2} e^{-\alpha d})^2} \right\} \times (1 - R_1) \cdot (1 - e^{-\alpha d}) \quad (31)$$

RCE structures can be fabricated in semiconductors using distributed Bragg reflectors (DBRs), which are alternating layers of different refractive index materials. DBRs can yield reflectivities in excess of 90% when many periods are used. Typical resonant cavity photodetectors are fabricated from GaAs-based materials by molecular beam epitaxy (MBE). AlAs/GaAs distributed Bragg reflectors are lattice matched to GaAs and achieve high reflectivity using greater than 10 periods. Figure 14 illustrates the cross section of a GaAs-based RCE photodiode optimized for 900 nm.

Figure 15 shows the quantum efficiency of a GaAs PIN photodetector [30]. The RCE photodiode exhibits greatly improved efficiency at 850 nm over a conventional single-pass photodiode. Additionally, RCE Schottky GaAs-based photodetectors have been demonstrated with bandwidths in excess of 50 GHz and peak quantum

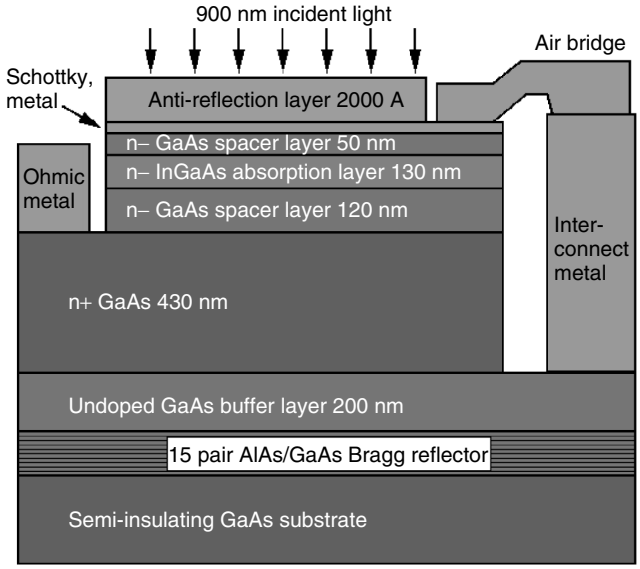


Figure 14. Cross-sectional view of a GaAs-based RCE photodiode optimized for 900 nm.

efficiency of 75% [31,32]. Making use of silicon on insulator technology (SOI), silicon-based RCE photodiodes have also been developed [33]. Conventional silicon photodiodes provide cost-efficient alternatives to GaAs- or InP-based semiconductors, due mainly to the ubiquitous silicon processing industry, but typically suffer from poor response at the communication wavelengths compared to its more expensive counterparts. The use of an RCE structure greatly improves the efficiency of silicon photodiodes, making them perform on par with GaAs- and InP-based photodiodes operating at 850 nm. Figure 16 shows the cross section of a silicon RCE photodiode.

4.7. Waveguide Photodetectors

One characteristic common to all the photodetectors described so far is the fact that the propagation direction

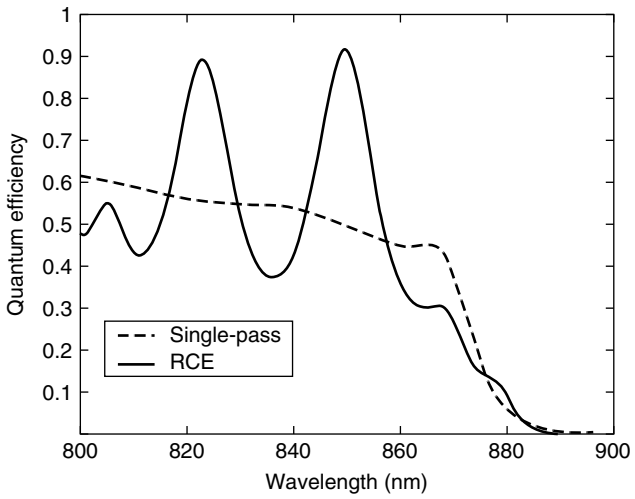


Figure 15. Quantum efficiency of a single-pass (dashed line) and a RCE (solid line) PIN photodetector.

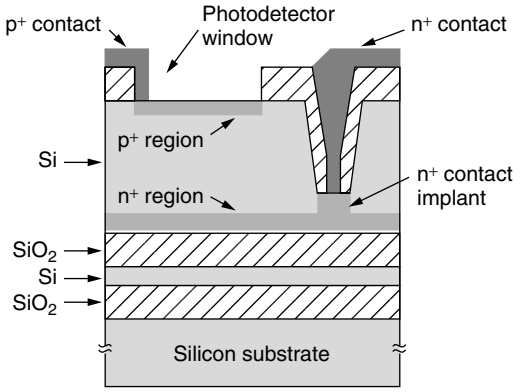


Figure 16. Cross-section of a silicon RCE PIN photodetector.

of both the incident light and the photogenerated carriers are parallel to one another. This characteristic is at the very heart of the bandwidth/efficiency "tug of war." Another interesting solution to this detector design issue involves making the propagation of the incident light and photogenerated carriers perpendicular to one another. This particular characteristic can be seen in waveguide detectors where, instead of top or bottom illumination, the photodetector is *side*-illuminated through a waveguide-like structure [34]. The guiding region doubles as the absorbing region and, under biasing, the photogenerated carriers within the guiding region are swept across the depletion region in a direction perpendicular to the propagation of the guided light. As a result, the photodetector quantum efficiency and bandwidth are not coupled to one another; because the quantum efficiency here depends on the waveguide length, absorption coefficient and confinement factor, while the bandwidth is related to the waveguide thickness, one can design a waveguide photodetector that has both high quantum efficiency and wide bandwidth. For example, a multimode waveguide PIN photodetector with a 3-dB bandwidth of 110 GHz and 50% quantum efficiency was demonstrated at 1.55 μm [35,36].

One design issue encountered with waveguide photodetectors involves saturation caused by photocarrier screening. This particular effect occurs when the absorption coefficient of the guiding region is high enough that a high carrier density is photogenerated near the entrance of the waveguide. As a result, the density of empty energy states for other photogenerated carriers are dramatically reduced, thereby preventing or *screening* further carrier photogeneration. This problem is often solved by reducing the absorption coefficient, in effect diluting and distributing the photogeneration of carriers along the guiding layer. However, an even more severe limitation of the waveguide photodetector involves its low coupling efficiency with respect to the incident light source.

5. CONCLUSIONS

In this section, we discussed the basic operation principles and properties of photodetectors, focusing on requirements relevant to high-speed optical communications. High-speed photodetectors are available for all communication wavelengths, utilizing a variety of semiconductor materials. Silicon photodetectors dominate applications for short wavelengths (visible to 0.85 μm) where low-cost, high-volume production is the most important consideration. Compound semiconductors, benefiting from their direct bandgap and availability of heterostructures, provide higher performance and coverage over the longer communication wavelengths (1.3–1.6 μm). The performance of conventional photodetectors has been refined to the point of fundamental material limitations. To meet the increasing demand for higher bit rates, more recent photodetector development has focused on innovative designs such as avalanche photodiodes (APDs), resonant cavity enhanced (RCE), and waveguide photodetectors, to name only a few.

Optical communications offering higher bandwidths have replaced their electrical counterparts, starting with long-distance applications. In these applications, the cost of a photodetector is not significant in the context of the overall system, and thus the choice for a receiver is performance-driven. Current trends indicate that optical communications will dominate medium- to short-distance applications in the near future. While the research and development efforts for faster and more efficient photodetectors will continue, the development of photodetectors with low-cost manufacturing potential will be increasingly important. We expect that there will be a greater effort behind the integration of high-speed photodetectors with electronic circuits and photonic components, leading not only to improved performance and functionality but also ultimately to the reduction of cost.

BIOGRAPHIES

Professor M. Selim Ünlü was born in Sinop, Turkey, in 1964. He received the B.S. degree in electrical engineering from Middle East Technical University, Ankara, Turkey, in 1986, and the M.S.E.E. and Ph.D. in electrical engineering from the University of Illinois, Urbana-Champaign, in 1988 and 1992, respectively. His dissertation topic dealt with resonant cavity enhanced (RCE) photodetectors and optoelectronic switches. In 1992, he joined the Department of Electrical and Computer Engineering, Boston University, as an Assistant Professor, and he has been an Associate Professor since 1998. From January to July 2000, he worked as a visiting professor at University of Ulm, Germany.

Dr. Ünlü's career interest is in *research and development of photonic materials, devices and systems* focusing on the design, processing, characterization, and modeling of semiconductor optoelectronic devices, especially photodetectors.

During 1994/95, Dr. Ünlü served as the Chair of IEEE Laser and Electro-Optics Society, Boston Chapter, winning the LEOS Chapter-of-the-Year Award. He served as the vice president of SPIE New England Chapter in 1998/99. He was awarded National Science Foundation Research Initiation Award in 1993, United Nations TOKTEN award in 1995 and 1996, and both the National Science Foundation CAREER and Office of Naval Research

Young Investigator Awards in 1996. Dr. Ünlü has authored and co-authored more than 150 technical articles and several book chapters and magazine articles; edited one book; holds one U.S. patent; and has several patents pending. During 1999–2001, he served as the chair of the IEEE/LEOS technical subcommittee on photodetectors and imaging, and he is currently an associate editor for *IEEE Journal of Quantum Electronics*.

Olufemi Dosunmu was born in Bronx, New York, in 1977. He graduated as Salutatorian from Lakewood High School in Lakewood, New Jersey in 1995, and received both his B.S. and M.S. degrees in Electrical Engineering with his thesis entitled *Modeling and Simulation of Intrinsic and Measured Response of High Speed Photodiodes*, along with a minor in Physics from Boston University in May 1999. While at Boston University, he was awarded the 4-year Trustee Scholarship in 1995, the Golden Key National Honor Society award in 1998, as well as the National Defense Science & Engineering Graduate (NDSEG) Fellowship in 2001. In 1997, he was a summer intern at AT&T Labs in Red Bank, New Jersey and, in 1998, at Princeton Plasma Physics Laboratories in Princeton, New Jersey. Between 1999 and 2000, he was employed at Lucent Technologies in the area of ASIC design for high-speed telecommunication systems. Mr. Dosunmu is expected to complete his PhD in January 2004.

Matthew K. Emsley was born in 1975, in the northern suburbs of Wilmington, Delaware. He graduated from Brandywine High School in 1993 and received his B.S. degree in Electrical Engineering from The Pennsylvania State University in December 1996, and a M.S. degree in Electrical Engineering from Boston University in May 2000 for his thesis entitled *Reflecting Silicon-on-Insulator (SOI) Substrates for Optoelectronic Applications*. While at Boston University Mr. Emsley was awarded the Electrical and Computer Engineering Chair Fellowship in 1997 and the Outstanding Graduate Teaching Fellow award for 1997/98. In 2001 Matthew was awarded a LEOS Travel Award as well as the H.J. Berman "Future of Light" Prize in Photonics for his poster entitled *Silicon Resonant-Cavity-Enhanced Photodetectors Using Reflecting Silicon on Insulator Substrates* at the annual Boston University Science Day. Mr. Emsley is expected to complete his Ph.D. in May 2003.

BIBLIOGRAPHY

1. D. Wood, *Optoelectronic Semiconductor Devices*, Prentice-Hall, New York, 1994.
2. J. Singh, *Semiconductor Optoelectronics*, McGraw Hill, New York, 1995.
3. K. J. Ebeling, *Integrated Opto-electronics*, Springer-Verlag, New York, 1992.
4. P. Bhattacharya, *Semiconductor Optoelectronic Devices*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1997.
5. G. P. Agrawal, *Fiber-Optic Communication Systems*, Wiley, New York, 1992.
6. J. M. Senior, *Optical Fiber Communications*, Prentice-Hall, New York, 1992.
7. S. Donati, *Photodetectors: Devices, Circuits, and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 2000.
8. E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*, Wiley, New York, 1996.
9. H. S. Nalwa, ed., *Photodetectors and Fiber-Optics*, Academic Press, New York, 2001.
10. T. Miya, Y. Terunuma, T. Hosaka, and T. Miyashita, Ultimate low-loss single-mode fibre at 1.55 μm, *Electron. Lett.* **15**(4): 106–108 (1979).
11. T. P. Lee and T. Li, *Photodetectors*, in S. E. Miller and A. G. Chynoweth, eds., *Optical Fiber Telecommunications*, Academic Press, New York, 1979, pp. 593–626.
12. D. K. C. MacDonald, *Noise and Fluctuations in Electronic Devices and Circuits*, Oxford Univ. Press, Oxford, 1962.
13. W. Schottky, *Ann. Phys.* **57**: 541 (1918).
14. J. B. Johnson, *Phys. Rev.* **32**: 97 (1928).
15. H. Nyquist, *Phys. Rev.* **32**: 110 (1928).
16. D. G. Parker, The theory, fabrication and assessment of ultra high speed photodiodes, *Gec. J. Res.* **6**(2): 106–117 (1988).
17. S. M. Sze, *Semiconductor Device Physics and Technology*, Wiley, New York, 1985.
18. M. Gökkavas et al., Design and optimization of high-speed resonant cavity enhanced schottky photodiodes, *IEEE J. Quant. Electron.* **35**(2): 208–215 (1999).
19. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York, 1991.
20. K. S. Giboney, M. J. W. Rodwell, and J. E. Bowers, Traveling-wave photodetectors, *IEEE Photon. Technol. Lett.* **4**(12): 1363–1365 (1992).
21. B. G. Streetman, *Solid State Electronic Devices*, 4th ed., Prentice-Hall, Englewood Cliffs, NJ, 1995.
22. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981.
23. D. L. Crawford et al., High speed InGaAs-InP p-i-n photodiodes fabricated on a semi-insulating substrate, *IEEE Photon. Technol. Lett.* **2**(9): 647–649 (1990).
24. Y. Wey et al., 108-GHz GaInAs/InP p-i-n photodiodes with integrated bias tees and matched resistors, *IEEE Photon. Tech. Lett.* **5**(11): 1310–1312 (1993).
25. E. H. Böttcher et al., Ultra-wide-band (>40 GHz) submicron InGaAs metal-semiconductor-metal photodetectors, *IEEE Photon. Technol. Lett.* **8**(9): 1226–1228 (1996).
26. J. B. D. Soole and H. Schumacher, InGaAs metal-semiconductor-metal photodetectors for long wavelength optical communications, *IEEE J. Quantum Electron.* **27**(3): (1991).
27. E. Özbay, K. D. Li, and D. M. Bloom, 2.0 ps, 150 GHz GaAs monolithic photodiode and all-electronic sampler, *IEEE Photon. Technol. Lett.* **3**(6): 570–572 (1991).
28. R. Kuchibhotla et al., Low-voltage high-gain resonant-cavity avalanche photodiode, *IEEE Photon. Technol. Lett.* **3**(4): 354–356 (1991).
29. H. Kuwatsuka et al., An $Al_xGa_{1-x}Sb$ avalanche photodiode with a gain bandwidth product of 90 GHz, *IEEE Photon. Technol. Lett.* **2**(1): 54–55 (1990).
30. M. Gökkavas et al., High-speed high-efficiency large-area resonant cavity enhanced p-i-n photodiodes for multimode

fiber communications, *IEEE Photon. Technol. Lett.* **13**(12): 1349–1351 (2001).

31. M. S. Ünlü et al., High bandwidth-efficiency resonant cavity enhanced schottky photodiodes for 800–850 nm wavelength operation, *Appl. Phys. Lett.* **72**(21): 2727–2729 (1998).
32. N. Biyikli et al., 45 GHz bandwidth-efficiency resonant cavity enhanced ITO-schottky photodiodes, *IEEE Photon. Technol. Lett.* **13**(7): 705–707 (2001).
33. M. K. Emsley, O. I. Dosunmu, and M. S. Ünlü, High-speed resonant-cavity-enhanced silicon photodetectors on reflecting silicon-on-insulator substrates, *IEEE Photon. Technol. Lett.* **14**(4): 519–521 (2002).
34. D. Dragoman and M. Dragoman, *Advanced Optoelectronic Devices*, Springer, 1999.
35. K. Kato et al., 110 GHz, 50% efficiency mushroom-mesa waveguide p-i-n photodiode for a 1.55 μm wavelength, *IEEE Photon. Technol. Lett.* **6**(6): 719–721 (1994).
36. K. Kato, Ultrawide-band/high-frequency photodetectors, *IEEE Trans. Microwave Theory Tech.* **47**(7): 1265–1281 (1999).

HORN ANTENNAS

EDWARD V. JULL
University of British Columbia
Vancouver, British Columbia,
Canada

A "horn" antenna is a length of conducting tube, flared at one end, and used for the transmission and reception of electromagnetic waves. For an efficient transition between guided and radiated waves, the horn dimensions must be comparable to the wavelength. Consequently horns are used mostly at centimeter and millimeter wavelengths. At lower or higher frequencies they are inconveniently large or small, respectively. They are most popular at microwave frequencies (3–30 GHz), as antennas of moderate directivity or as feeds for reflectors or elements of arrays.

Since acoustic horns have been in use since prehistoric times, the design of horns as musical instruments was a highly developed art well before the appearance of the first electromagnetic horns. This occurred shortly after Hertz in 1888 first demonstrated the existence of electromagnetic waves. Experimenters placed their spark-gap sources in hollow copper tubes (Figs. 1a,5a). These tubes acted as highpass filters for microwave and millimeter wave radiation from the open end. In London in 1897 Chunder Bose used rectangular conducting tubes with "collecting funnels," or pyramidal horns (Fig. 1d) in his demonstrations at 5 and 25 mm wavelengths [1]. Thus the electromagnetic horn antenna was introduced but this early beginning of microwave invention closed with Marconi's demonstration that longer wavelengths could be received at greater distances. Horns were too large to be practical at those wavelengths, and it was almost 40 years before microwave horns reappeared with the need for directive antennas for communications and radar. Horns alone were often not sufficiently directive but combined in an array or with a lens (Fig. 4a), or more often a parabolic reflector (Fig. 4b,c) highly directive antenna beams are obtained.

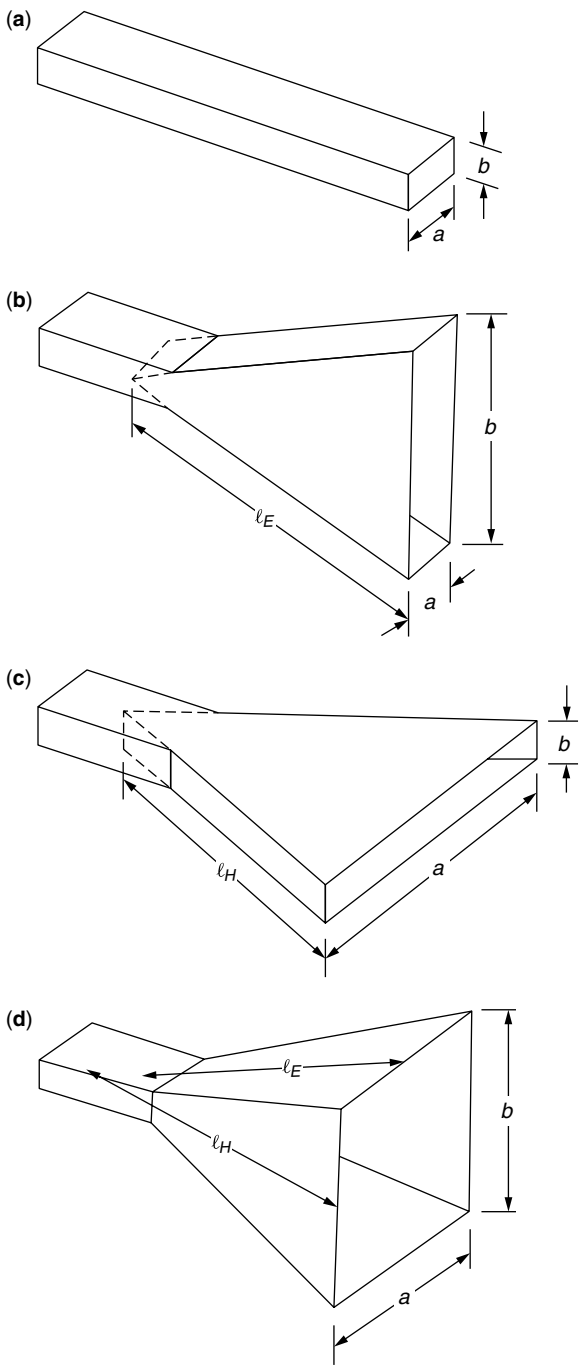


Figure 1. (**a**) Open-ended rectangular waveguide; (**b**) *E*-plane sectoral horn; (**c**) *H*-plane sectoral horn; (**d**) Pyramidal horn.

1. RADIATING WAVEGUIDES AND HORNS

Horns are normally fed by waveguides supporting only the dominant waveguide mode. For a rectangular waveguide (Fig. 1a) with TE_{01} mode propagation only, these dimensions in wavelengths λ are $\lambda/2 < a < \lambda$ and $b \approx a/2$. Open-ended waveguides have broad radiation

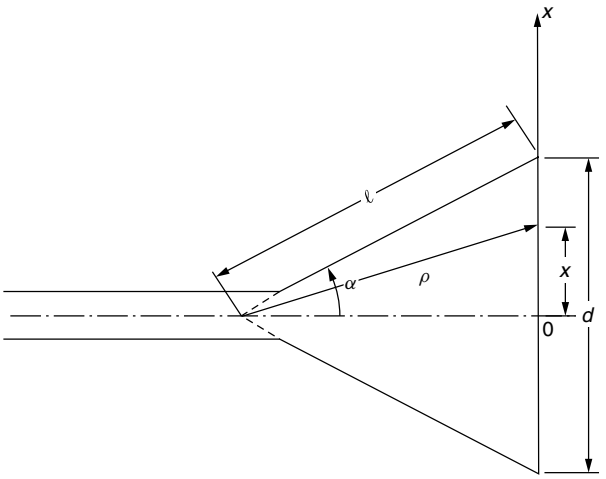


Figure 2. Effect of horn flare on the aperture field phase of a horn.

patterns, so when used as a feed for a reflector, there is substantial spillover, or radiation missing the reflector and radiation directly backward from the feed. To increase the directivity of a radiating waveguide and its efficiency as a reflector feed, for example, its aperture dimensions must be enlarged, for the beamwidth of an aperture of width $a \gg \lambda$ is proportional to λ/a radians.

This waveguide enlargement by a flare characterizes horns. The aperture fields of a horn are spherical waves originating at the horn apex (Fig. 2). The path from the horn apex to the aperture plane at a distance x from the aperture center of a horn of slant length ℓ is

$$\rho = ((\ell \cos\alpha)^2 + x^2)^{1/2} \approx \ell \cos\alpha + \frac{x^2}{2\ell \cos\alpha} \tag{1}$$

when $x \ll \ell \cos\alpha$. Thus the phase variation in radians across the aperture for small flare angles α is approximately $kx^2/(2\ell)$, where $k = 2\pi/\lambda$ is the propagation constant. This quadratic phase variation increases with increasing flare angle, thus reducing directivity increase due to the enlarged aperture dimension. It is convenient to quantify aperture phase variation by the parameter

$$s = \frac{\ell(1 - \cos\alpha)}{\lambda} \approx \frac{d^2}{8\lambda\ell}, \; d \ll \ell \tag{2}$$

which is the approximate difference in wavelengths between the distance from the apex to the edge ($x = d/2$) and the center ($x = 0$) of the aperture. The radiation patterns of Fig. 3a,b [2] show the effect of increasing s on the E- and H-plane radiation patterns of sectoral and pyramidal horns. The main beam is broadened, the pattern nulls are filled, and the sidelobe levels are raised over those for an in-phase aperture field ($s = 0$). With large flare angles radiation from the extremities of the aperture can be so out of phase with that from the center that the horn directivity decreases with increasing aperture width.

The adverse effects of the flare can be compensated by a lens in the aperture (Fig. 4a), but because that adds to the weight and cost and because bandwidth limitations

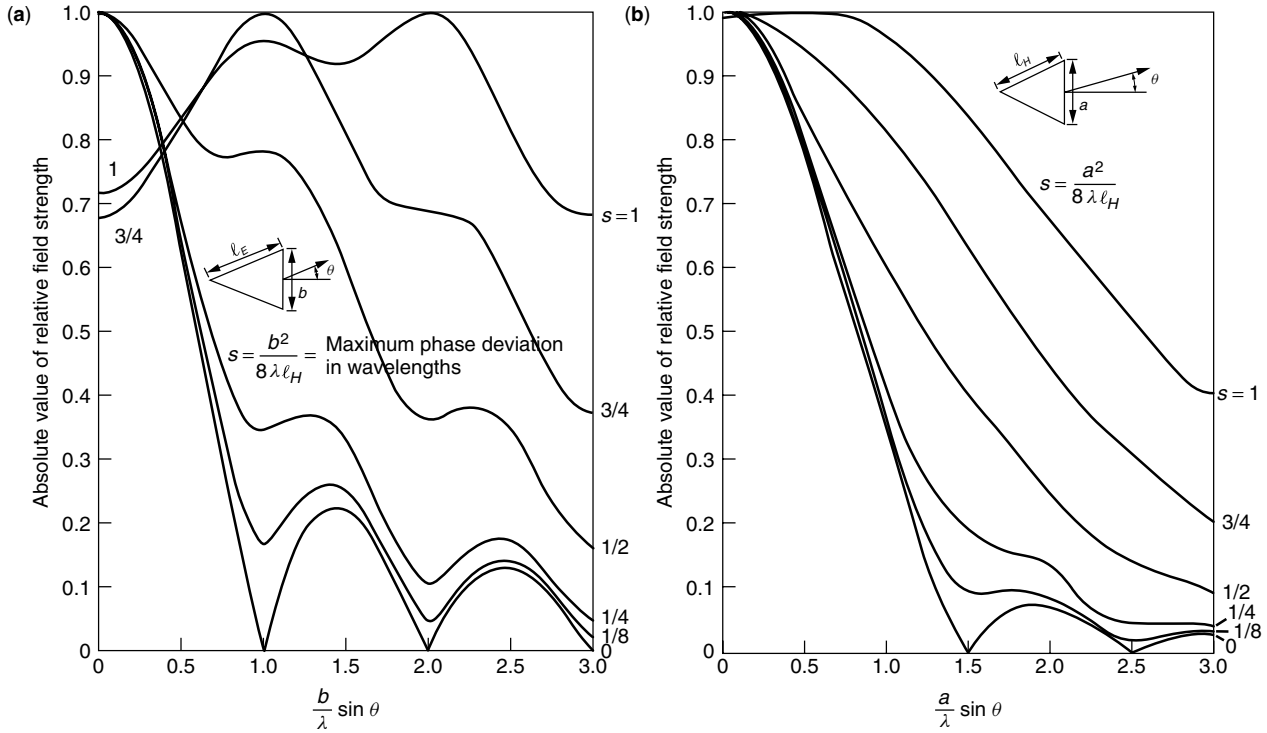


Figure 3. Universal radiation patterns of sectoral and pyramidal horns flared in the (**a**) E and (**b**) H-planes. The parameter $s = b^2/8\lambda\ell_E$ in (a) and $a^2/8\lambda\ell_H$ in (b); $2\pi s/\lambda$ is the maximum phase difference between the fields at the center and the edge of the aperture. (Copyright 1984, McGraw-Hill, Inc. from Love [2].)

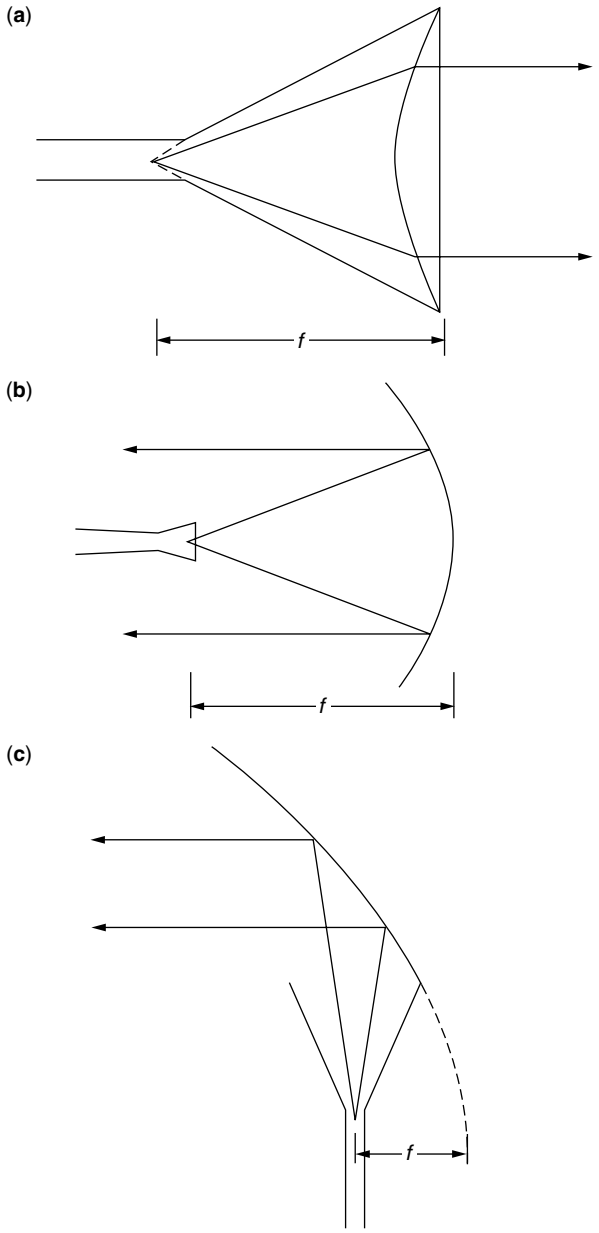


Figure 4. (**a**) Horn aperture field phase correction by a lens; (**b**) parabolic reflector fed by a horn; (**c**) horn reflector antenna (f = focal length of the lens or reflector).

are introduced by matching the lens surfaces to reduce reflections, it is seldom done. Instead, a combination of aperture width and flare length in wavelengths is chosen that provides maximum axial directivity or minimum beamwidth. This is an "optimum" horn design. To achieve higher directivity or narrower beamwidth for a given aperture width, a longer horn is required.

Sectoral horns (Fig. 1b,c) are rectangular waveguides flared in one dimension only. The incident waveguide mode becomes a radial cylindrical mode in the flared region of the horn. Since radiation pattern beamwidths are inversely proportional to aperture dimensions in wavelengths, sectoral horns have beams that are narrow in the plane containing the broad dimension. Such fan shaped beams may be useful for illuminating elongated parabolic reflectors or parabolic cylinder reflectors.

A pyramidal horn (Fig. 1d) is flared in both waveguide dimensions and so is more adaptable both as a reflector feed and on its own. The forward radiation pattern may be calculated quite accurately from Kirchhoff diffraction theory for all but small horns. The TE_{01} rectangular waveguide mode yields an aperture field uniform in one dimension (in the E plane) and cosinusoidal in the other (the H plane). A comparison of parts (a) and (b) of Fig. 3 shows that this results in a higher sidelobes in the E-plane and, for a square aperture, a narrower beam. Pyramidal horns are relatively easily constructed, and for all except small horns their axial gain can be predicted accurately. Consequently, they are used as gain standards at microwave frequencies; that is, they are used to experimentally establish the gain of other microwave antennas by comparing their response to the same illuminating field.

Most of the preceding remarks on open-ended rectangular waveguides and pyramidal horns also apply to open-ended circular waveguides and conical horns (Fig. 5a,b). For propagation of the lowest-order mode (TE_{11}) only in a circular waveguide, the interior diameter must be $0.59\lambda < a < 0.77\lambda$. This mode has a uniform aperture field in the E plane and a cosinusoidal distribution in the orthogonal H plane. This appears, modified by a quadratic phase variation introduced by the flare, in the aperture field of the horn. Consequently the E-plane radiation pattern of the horn is narrower, but with higher sidelobes than the H-plane pattern and the radiated beam is elliptical in cross-section. In addition, cross-polarized fields appear in pattern lobes outside the principal planes.

2. HORN FEEDS FOR REFLECTORS

Many refinements to horns arise from their use as efficient feeds for parabolic reflectors, particularly in satellite and space communications and radio astronomy. The phase center, where a horn's far radiation field appears to originate, must be placed at the focus of the reflector (Fig. 4b). This phase center is within the horn on the horn axis and depends on the flare angle and aperture distribution. For both rectangular and conical horns the position of the phase center is not the same in the E and H planes, or planes containing the electric and magnetic field vectors, respectively. A phase center can be calculated from the average of the positions of the E- and H-plane phase centers or determined from the position of the feed that maximizes the gain of the reflector antenna.

For efficient aperture illumination the feed horn radiation pattern should approximately match the shape of the aperture, and illuminate it essentially uniformly and with minimal spillover, or radiation missing the reflector. Pyramidal horns may seem suitable for rectangular apertures because their beams are rectangular in cross section, and conical horns may seem a natural choice for a circular aperture, but efficient aperture illumination is not obtained in either case, because their principal plane patterns differ. Both horns have high E-plane pattern

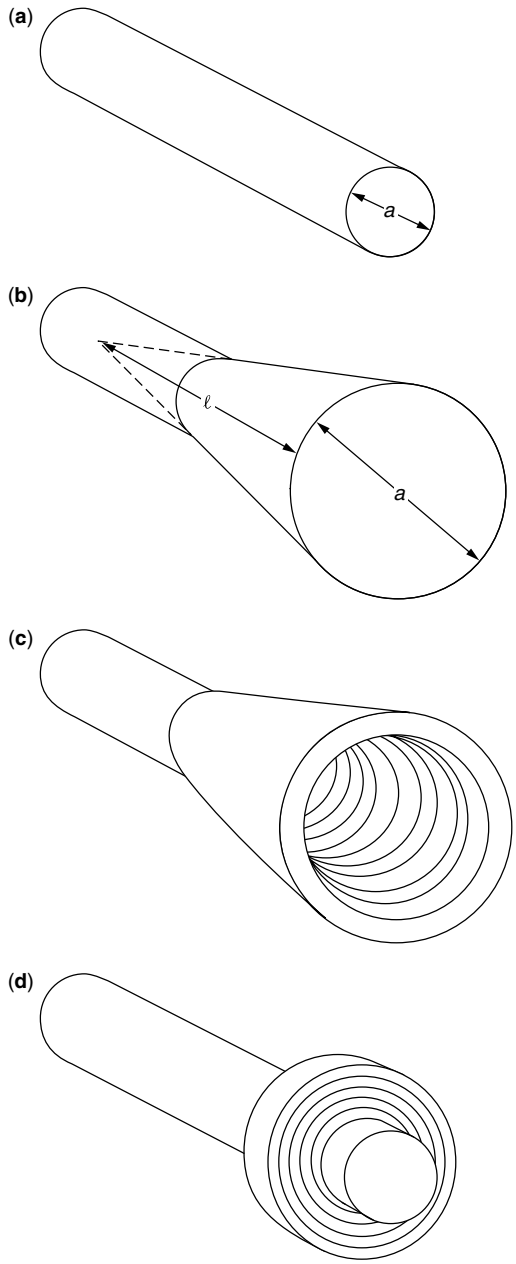


Figure 5. (**a**) Open-ended circular waveguide; (**b**) conical horn; (**c**) corrugated horn; (**d**) circular waveguide with corrugated flange.

sidelobes and low H-plane sidelobes. A dual (TE_{11}/TM_{11})-mode conical horn provides equal E- and H-plane beamwidths and equally low sidelobes, and is an efficient feed for a circular aperture over a narrow frequency band (see Love [3], p. 195; Ref. 3 also contains reprints of most earlier significant papers on horn antennas). A broadband solution achieves an axisymmetric beam with annular corrugations on the interior surfaces of a conical horn (Fig. 5c). These produce a horn aperture field distribution that is approximately cosinusoidal across the conical horn aperture in all directions and hence an axisymmetric radiation pattern with low sidelobes. Such corrugations in the E-plane interior walls only of a pyramidal horn will produce a cosinusoidal E-plane aperture distribution, and consequently similar E-plane and H-plane radiation patterns for a square horn aperture.

A feed for a small circular reflector that is more easily constructed than a corrugated conical horn but with a less axisymmetric radiation pattern, is an open-ended circular waveguide ringed by a recessed disk of approximately quarter-wavelength-deep corrugations (Fig. 5d). These corrugations suppress back radiation from the feed and so improve the aperture illumination over that of a simple open circular waveguide (Ref. 3, pp. 181, 226). Combined with dual-mode excitation, this arrangement provides a simple and efficient feed for a front-fed paraboloidal reflector.

3. RADIATION FROM APERTURES

The far-field radiation pattern of an aperture can be calculated exactly from the Fourier transform of the tangential fields in the entire aperture plane. Either electric or magnetic aperture fields may be used but for apertures in space, a combination of the two gives the best results from the usual assumption that aperture plane fields are confined to the aperture and negligible outside it. This aperture field is assumed to be the undisturbed incident field from the waveguide. For apertures with dimensions larger than several wavelengths, a further simplifying assumption usually made is that the aperture electric and magnetic fields are related as in free space.

3.1. Rectangular Apertures

With the above assumptions, at a distance much greater than the aperture dimensions, the radiated electric field intensity of a linearly polarized aperture field $E_x(x, y, 0)$ in the coordinates of Fig. 6a is

$$\overline{E}(r, \theta, \phi) = \overline{A}(r, \theta, \phi) \int_{-(b/2)}^{b/2} \int_{-(a/2)}^{a/2} E_x(x, y, 0) e^{j(k_1 x + k_2 y)}\, dx\, dy \tag{3}$$

Here

$$\left.\begin{aligned} k_1 &= k \sin\theta \cos\phi \\ k_2 &= k \sin\theta \sin\phi \end{aligned}\right\} \tag{4}$$

and

$$\overline{A}(r, \theta, \phi) = j \frac{e^{-jkr}}{2\lambda r} (1 + \cos\theta)(\hat{\theta} \cos\phi - \hat{\phi} \sin\phi) \tag{5}$$

is a vector defining the angular behaviour of the radiation polarization for an aperture in space. For an aperture in a conducting plane, it is more accurate to use

$$\overline{A}(r, \theta, \phi) = j \frac{e^{-jkr}}{\lambda r} (\hat{\theta} \cos\phi - \hat{\phi} \sin\phi \cos\theta) \tag{6}$$

which, since it is based on the aperture plane electric fields only, fully satisfies the assumption of a vanishing tangential field in the aperture plane outside the aperture. Consequently radiation fields of open-ended waveguides and small horns can be calculated accurately from (3) with (6) if they are mounted in a conducting plane.

Clearly (5) and (6) differ significantly only for large angles θ off the beam axis.

If the aperture field is separable in the aperture coordinates — that is, if in (3), $E_x(x, y, 0) = E_0 E_1(x) E_2(y)$ where $E_1(x)$ and $E_2(y)$ are field distributions normalized to E_0, the double integral is the product of two single integrals.

$$E(r, \theta, \phi) = \overline{A}(r, \theta, \phi) E_0 F_1(k_1) F_2(k_2) \tag{7}$$

where

$$F_1(k_1) = \int_{-(a/2)}^{a/2} E_1(x) e^{jk_1 y}\, dx \tag{8}$$

$$F_2(k_2) = \int_{-(b/2)}^{b/2} E_2(y) e^{jk_2 y}\, dy \tag{9}$$

define the radiation field.

4. OPEN-ENDED WAVEGUIDES

4.1. Rectangular Waveguides

With the TE_{10} waveguide mode the aperture field

$$E_x(x, y, 0) = E_0 \cos \frac{\pi y}{a} \tag{10}$$

in (7) yields the following equations for (8) and (9):

$$F_1(k_1) = b \frac{\sin\left(\dfrac{k_1 b}{2}\right)}{\dfrac{k_1 b}{2}} \tag{11}$$

$$F_2(k_2) = a \left(\frac{\cos\left(\dfrac{k_2 a}{2}\right)}{\pi^2 - (k_2 a)^2} \right) \tag{12}$$

This defines the radiation pattern in the forward hemisphere $-\pi/2 < \theta < \pi/2, 0 < \phi < 2\pi$. If the aperture is in space, then (5) is used for $\overline{A}(r, \theta, \phi)$, but this is not an accurate solution since the aperture dimensions are not large. Rectangular waveguides mounted in conducting planes use (6) for $\overline{A}(r, \theta, \phi)$ in (7), which then accurately provides the far field. The pattern has a single broad lobe with no sidelobes. For large apertures plots of the normalized E-plane ($\phi = 0$) and H-plane ($\phi = \pi/2$) patterns of (7) appear in Fig. 3a,b for those of a horn with no flare ($s = 0$), but without the factor $(1 + \cos\theta)/2$ from (5) or $\cos\theta$ from (6).

4.2. Circular Waveguides

The dominant TE_{11} mode field in circular waveguide produces an aperture field distribution, which in the aperture coordinates ρ',ϕ' of Fig. 6b is

$$\overline{E}(\rho', \phi') = E_0 \left[\hat{\rho}' \frac{J_1(k_c \rho')}{k_c \rho'} \cos\phi' + \hat{\phi}' J_1'(k_c \rho') \sin\phi' \right] \tag{13}$$

where J_1 is the Bessel function of the first kind and order, J_1' is its derivative with respect to its argument $k_c \rho'$, and $k_c a/2 = 1.841$ is its first root; E_0 is the electric field at the aperture center ($\rho' = 0$). Since (13) is not linearly polarized, its use in (3) provides only part of the total radiated far field. The total field

$$\overline{E}(r, \theta, \phi) = jkaE_0 J_1(1.841) \frac{e^{-jkr}}{r} \left\{ \hat{\theta} \cos\phi \frac{J_1\left(\dfrac{k'a}{2}\right)}{\dfrac{k'a}{2}} + \hat{\phi} \sin\phi \cos\theta \frac{J_1\left(\dfrac{k'a}{2}\right)}{1 - \left(\dfrac{k'a}{3.682}\right)^2} \right\} \tag{14}$$

in which $k' = k \sin\theta$.

In the E and H planes ($\phi = 0$ and $\pi/2$) the cross-polarized fields cancel and the patterns shown in Fig. 14a are similar to those of (11) and (12), respectively, but with slightly broader beams and lower sidelobes for

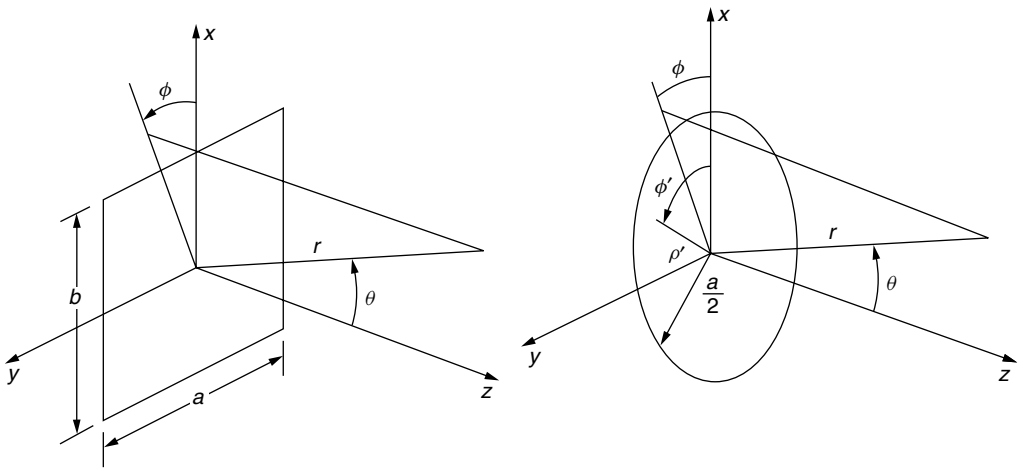


Figure 6. Coordinates for radiation from (**a**) rectangular and (**b**) circular apertures.

the same aperture dimensions. As with rectangular waveguides, open-ended circular waveguide apertures are insufficiently large for (14) to represent all the radiated fields accurately. In the principal planes ($\phi = 0, \pi/2$), it can give a reasonable approximation for the copolarized fields but fails to accurately represent the cross-polarized field patterns in $\phi = \pi/4$. This is evident from a comparison of numerical results from approximate and exact solutions (see Collin [4], p. 233).

5. PYRAMIDAL AND SECTORAL HORNS

5.1. Radiation Patterns

A pyramidal horn fed by a rectangular waveguide supporting the TE_{10} mode has an incident electric field in the aperture of Fig. 6a that is approximately the mode distribution modified by a quadratic phase variation in the two aperture dimensions:

$$E_x(x, y, 0) = E_0 \cos\left(\frac{\pi y}{a}\right) \exp\left(-jk\left(\frac{x^2}{2\ell_E} + \frac{y^2}{2\ell_H}\right)\right) \quad (15)$$

With (15), Eq. (3) becomes

$$\overline{E}(r, \theta, \phi) = \overline{A}(r, \theta, \phi) E_0 I_1(k_1) I_2(k_2) \quad (16)$$

where (5) is used for $\overline{A}(r, \theta, \phi)$ and

$$I_1(k_1) = \int_{-(b/2)}^{b/2} \exp\left(-j\left(\frac{\pi x^2}{\lambda \ell_E} - k_1 x\right)\right) dx \quad (17)$$

$$I_2(k_2) = \int_{-(a/2)}^{a/2} \cos\left(\frac{\pi y}{a}\right) \exp\left(-j\left(\frac{\pi y^2}{\lambda \ell_H} - k_2 y\right)\right) dy \quad (18)$$

The E-plane ($\phi = 0$) and H-plane ($\phi = \pi/2$) radiation patterns are, respectively

$$\frac{E_\theta(r, \theta)}{E_\theta(r, 0)} = \frac{1 + \cos\theta}{2} \frac{I_1(k \sin\theta)}{I_1(0)} \quad (19)$$

$$\frac{E_\theta(r, \theta)}{E_\theta(r, 0)} = \frac{1 + \cos\theta}{2} \frac{I_2(k \sin\theta)}{I_2(0)} \quad (20)$$

These integrals can be reduced to the Fresnel integrals

$$C(u) - jS(u) = \int_0^u e^{-j(\pi/2)t^2}\, dt \quad (21)$$

which are tabulated and for which computer subroutines are available. For example,

$$\frac{I_1(k \sin\theta)}{I_1(0)} = \frac{e^{j(\pi \ell_E/\lambda)\sin^2\theta}}{2} \frac{C(u_2) - C(u_1) - j[S(u_2) - S(u_1)]}{C(u) - jS(u)} \quad (22)$$

with

$$u = \frac{b}{\sqrt{2\lambda \ell_E}} \quad (23)$$

$$u_{\frac{2}{1}} = \pm u - \sqrt{\frac{2\ell_E}{\lambda}} \sin\theta \quad (24)$$

Figure 3a shows plots of the magnitude of (22) for various values of the E-plane flare parameter $s = b^2/8\lambda \ell_E$, while Fig. 3b shows corresponding plots of $|I_2(k \sin\theta)/I_2(0)|$ for the H-plane flare parameter $s = a^2/8\lambda \ell_H$. For no flare ($s = 0$) the patterns are those of a large open-ended rectangular waveguide supporting only the TE_{10} mode. The effect of the flare is to broaden the main beam, raise the sidelobes, and fill the pattern nulls. For larger values of s, there is enhanced pattern beam broadening and eventually a splitting of the main beam on its axis.

These curves also represent the radiation patterns of the E/H-plane sectoral horns of Fig. 1b,c. For an E-plane sectoral horn ($\ell_H \to \infty$), the E-plane pattern is given by (19) and the H-plane pattern approximately by (12). For an H-plane sectoral horn ($\ell_E \to \infty$), the E-plane pattern is given approximately by (11) and the H-plane pattern by (20).

In comparing parts (a) and (b) of Fig. 3 it is evident that E-plane beamwidths of a square aperture are narrower than H-plane beamwidths. For horns of moderate flare angle and optimum horns the E-plane half-power beamwidth is $0.89\lambda/b$ radians and the H-plane half-power beamwidth is $1.22\lambda/a$ radians. E-plane patterns have minimum sidelobes of -13.3 dB below peak power, while H-plane pattern minimum sidelobes levels are -23.1 dB.

The universal patterns of Fig. 3a,b can also be used to predict the approximate near-field radiation patterns of horns by including the quadratic phase error; which is a first-order effect of finite range r. This is done by including

$$\exp\left(-j\frac{\pi}{r\lambda}(x^2 + y^2)\right) \quad (25)$$

in (15). Then the near field principal plane patterns of a pyramidal horn are given by (17) and (18) with ℓ_E, ℓ_H replaced by

$$\ell'_H = \frac{r\ell_H}{r + \ell_H} \quad (26)$$

and

$$\ell'_E = \frac{r\ell_E}{r + \ell_E} \quad (27)$$

These near-field effects are analogous to decreasing the flare length of a horn with a fixed aperture width. The main beam broadens, nulls are filled in, and sidelobes rise.

5.2. Limitations and Extensions

Results from (16) do not apply to small horns and are limited to the forward direction ($\theta < 90^\circ$). They are most accurate on and around the beam axis ($\theta = 0$), becoming progressively less accurate as θ increases. The simplest method for extending the analysis is by the uniform geometric theory of diffraction (see, e.g., Ref. 3, p. 66), which provides the edge-diffracted fields in the lateral and rear directions, which receive no direct illumination from the aperture. Only the edges normal to the plane of the pattern contribute significantly to the E-plane pattern, but the rear H-plane pattern requires contributions from all four aperture edges and so is difficult to calculate this way.

While the geometry of the pyramidal horn defies rigorous analysis, numerical methods have been used with some success for open waveguides and small horns. For larger horns this approach becomes computationally

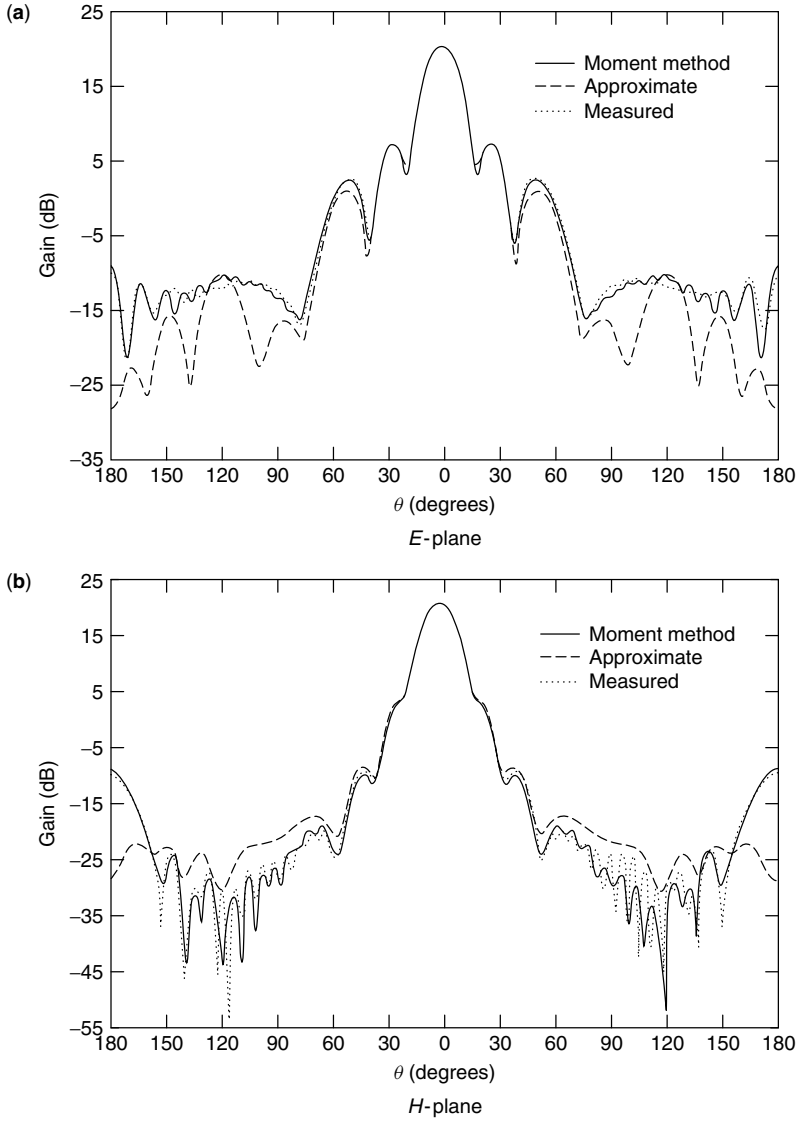


Figure 7. Calculated and measured (**a**) E-plane and (**b**) H-plane radiation patterns of a pyramidal horn of dimensions $a = 4.12\lambda$, $b = 3.06\lambda$, $\ell_E = 10.52\lambda$ $\ell_E = 9.70\lambda$. (Copyright 1993, IEEE, from Liu et al. [5].)

intensive, but some results from Liu et al. (5) are shown in Fig. 7 and compared with measurements and approximate computations. Their numerical computations and measurements by Nye and Liang (6) of the aperture fields show that higher-order modes need to be added to the dominant mode field of (15) and that the parabolic phase approximation of (1) improves as the aperture size increases.

5.3. Gain

Pyramidal horns are used as gain standards at microwave frequencies because they can be accurately constructed and their axial directive gain reliably predicted from a relatively simple formula. The ratio of axial far-field power density to the average radiated power density from (16) yields

$$G = G_0 R_E(u) R_H(v, w) \tag{28}$$

where $G_0 = 32ab/(\pi\lambda^2)$ is the gain of an in-phase uniform and cosinusoidal aperture distribution. The reduction of this gain due to the phase variation introduced by the E-plane flare of the horn is

$$R_E(u) = \frac{C^2(u) + S^2(u)}{u^2} \tag{29}$$

where the Fresnel integrals and their argument are defined by (21) and (23). Similarly the gain reduction factor due to the H-plane flare of the horn is

$$R_H(v, w) = \frac{\pi^2}{4} \frac{[C(v) - C(w)]^2 + [S(v) - S(w)]^2}{(v - w)^2} \tag{30}$$

where

$$\begin{matrix} v \\ w \end{matrix} = \pm \frac{a}{\sqrt{2\lambda\ell_H}} + \frac{1}{a}\sqrt{\frac{\lambda\ell_H}{2}} \tag{31}$$

A plot of R_E and R_H in decibels as a function of the parameter $2d^2/\lambda\ell$, where d is the appropriate aperture dimension b or a and ℓ the slant length ℓ_E or ℓ_H, respectively, is shown in Fig. 8. Calculation of the gain

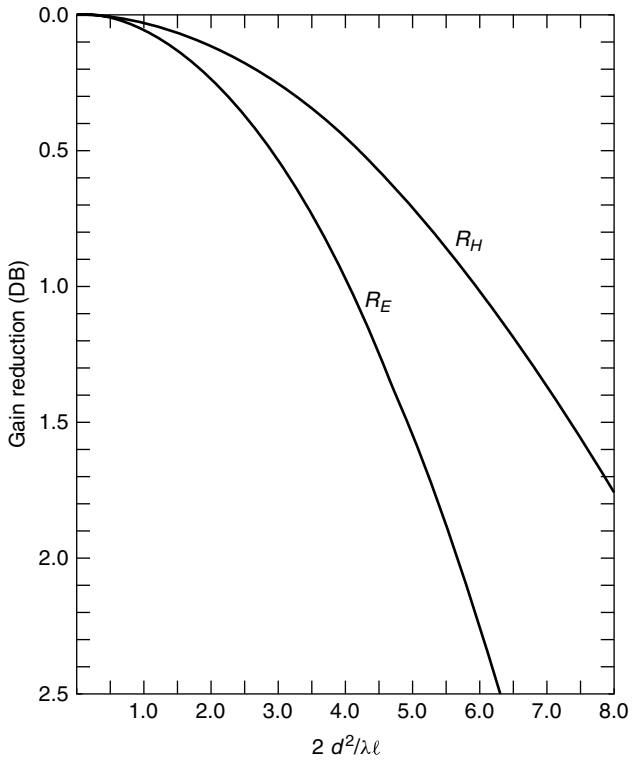


Figure 8. E- and H-plane flare and near-field gain reduction factors R_E and R_H of pyramidal and sectoral horns in decibels. (Copyright 1981, IEE, from Jull [11].)

from (28) is accurate to about ± 0.1 dB for 22 dB standard gain pyramidal horns: optimum horns with dimensions of at least 5λ. For 18-dB-gain horns, the accuracy is about ± 0.2 dB, and for 15 dB horns, ± 0.5 dB. Since optimum gain pyramidal horns have an aperture efficiency of approximately 50%, the gain is approximately

$$G = 0.5\frac{4\pi}{\lambda^2}ab \tag{32}$$

For an E-plane sectoral horn $\ell_H \to \infty$ and $R_H(v, w) \to 1$ the axial gain is then $G_E = G_0 R_E(u)$, an inaccurate formula because aperture dimension a is less than a wavelength. A result that includes the fact that aperture electric and magnetic fields are not related by free-space conditions and that interaction occurs across the narrow aperture of the horn is

$$G_E = \frac{16ab}{\lambda^2(1+\lambda g/\lambda)}R_E(u')\exp\left[\frac{\pi a}{\lambda}\left(1-\frac{\lambda}{\lambda_g}\right)\right] \tag{33}$$

where

$$u' = \frac{b}{\sqrt{2\lambda_g \ell_E}} \tag{34}$$

and

$$\lambda_g = \frac{\lambda}{\sqrt{1-(\lambda/2a)^2}} \tag{35}$$

is the guide wavelength. The accuracy of (33) is comparable to that of (28) for the horns of similar b dimension.

The gain of an H-plane sectoral horn, obtained by letting $\ell_E \to \infty$ so that $R_E(u) \to 1$, is $G_H = G_0 R_H(v, w)$. It probably is reasonably accurate, but there appears to be no experimental evidence available to verify it.

The near-field gain of pyramidal and sectoral horns can be calculated from the preceding expressions by replacing ℓ_E and ℓ_H by (26) and (27), respectively.

6. CONICAL HORNS

The aperture field of a conical horn fed by a circular waveguide supporting the TE_{11} mode is approximately

$$E(\rho', \phi')\exp\left(\frac{-jk\rho'^2}{2\ell}\right) \tag{36}$$

where $\overline{E}(\rho', \phi')$ is given by (13) and ℓ is the slant length of the horn. Numerical calculation of the radiation patterns is necessary. In the example of Fig. 9 [7] with a flare angle $\alpha = 5^\circ$ and aperture width $a = 4\lambda$, the E-plane ($\phi = 0$) pattern is narrower than the H-plane ($\phi = \pi/2$) pattern as in square rectangular horns. The cross-polar ($\phi = \pi/4$) radiation pattern peak level is -18.7 dB relative to the copolar pattern peak levels, a level typical of conical horn apertures larger than about 2λ. Smaller conical horns can have more axisymmetric patterns. E- and H-plane patterns have equal beamwidths for an aperture diameter $a = 0.96\lambda$, and cross-polarized fields cancel for $a = 1.15\lambda$. This makes small conical horns efficient as reflector feeds and as array elements with high polarization purity.

Larger conical horns are similar to rectangular horns in their lack of axial pattern symmetry. Optimum gain conical horns have an aperture efficiency of about 54% and half-power beamwidths in the E and H planes of $1.05\lambda/a$ and $1.22\lambda/a$ radians, respectively, for aperture diameters of more than a few wavelengths.

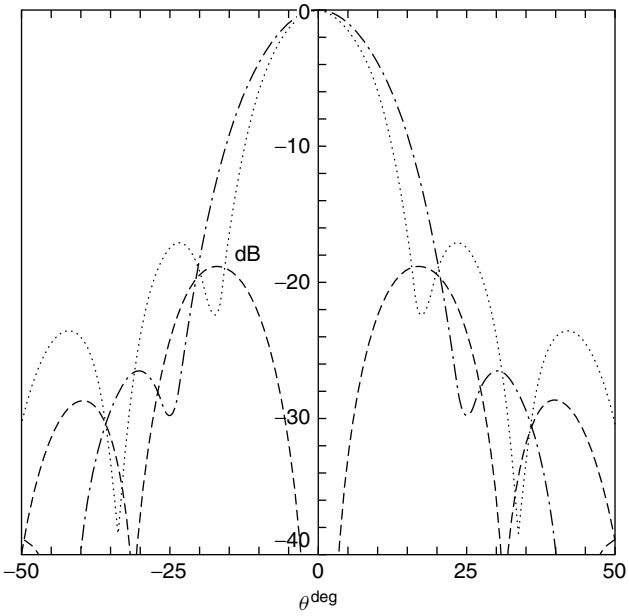


Figure 9. Copolar and crosspolar radiation patterns for a conical horn with dimensions $a = 4\lambda$, $\ell = 23\lambda$ - - - - E-plane, $-\cdot-\cdot-$ H-plane $-\ -\ -\ -$ crosspolarization. (Copyright 1994, IEE, from Olver et al. [7].)

7. MULTIMODE AND CORRUGATED HORNS

Lack of axisymmetric radiation patterns make rectangular and conical horns inefficient reflector feeds. Conical horns also have unacceptably high cross-polarization levels if used as reflector feeds in a system with dual polarization. Multimode and corrugated horns were developed largely to overcome these deficiencies. In a dual-mode horn in (Ref. 3, p. 195), this is done by exciting the TM_{11} mode, which propagates for circular waveguide diameters $a > 1.22\lambda$, in addition to the TE_{11} mode, which propagates for $a > 0.59\lambda$. The electric field configuration of these modes in a waveguide cross section is shown in Fig. 10a,b. Added in phase and in the right proportion, cross-polarized and aperture perimeter fields cancel, while the copolar fields around the aperture center add, yielding the aperture field configuration of Fig. 10c. These mixed mode fields are linearly polarized and taper approximately cosinusoidally radially across the aperture. This yields the essentially linearly polarized and axisymmetric radiation patterns desired.

Partial conversion of TE_{11} to TM_{11} fields can be effected by a step discontinuity in the circular waveguide feed, as in Fig. 10d, or by a circular iris or dielectric ring in the horn. The TM_{11}/TE_{11} amplitude ratio depends on the ratio of waveguide diameters, and the relative phase of the modes depends on the length of larger-diameter circular waveguide and the horn. This dependence limits the frequency bandwidth of the horn to about 5%. A multimode square pyramidal horn has similarly low sidelobe levels in its E- and H-plane radiation patterns because of an essentially cosinusoidal aperture distribution in both E and H planes [2]. This is achieved by excitation of a hybrid TE_{21}/TM_{21} mode either by an E-plane step discontinuity or by changes in the E-plane flare. With their bandwidth very limited, dual-mode horns have largely been replaced by corrugated horns in dual-polarization systems, except where a lack of space may give an advantage to a thin-walled horn.

Corrugated horns have aperture fields similar to those of Fig. 10c and consequently similar radiation patterns,

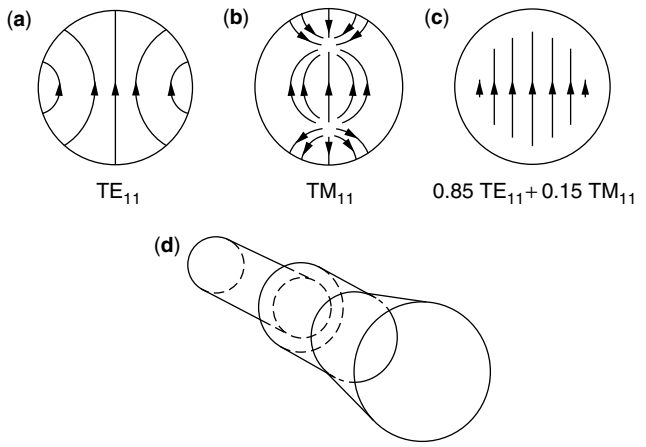


Figure 10. Excitation of axisymmetric linearly polarized aperture fields in a stepped conical horn. (Copyright 1984, McGraw-Hill, Inc. from Love [2].)

but without the frequency bandwidth limitations of the multimode horn. This is achieved by introducing annular corrugations to the interior walls of a conical horn. There must be sufficient corrugations per wavelength (at least 3) that the annular electric field E_ϕ is essentially zero on the interior walls. The corrugations make the annular magnetic field H_ϕ also vanish. This requires corrugation depths such that short circuits at the bottom of the grooves appear as open circuits at the top, suppressing axial current flow on the interior walls of the horn. This groove depth is $\lambda/4$ on a plane corrugated surface or a curved surface of large radius. For a curved surface of smaller radius, such as near the throat of the horn, the slot depths need to be increased; For example, for a surface radius of 2λ, the depth required is 0.3λ. Usually slots are normal to the conical surface in wide-flare horns but are often perpendicular to the horn axis with small flares. To provide a gradual transition from the TE_{11} mode in the wave guide to a hybrid HE_{11} mode in the aperture, the depth of the first corrugation in the throat should be about 0.5λ so that the surface there resembles that of a conducting cone interior. Propagation in corrugated conical horns can be accurately calculated numerically by mode matching techniques. The aperture field is approximately

$$E_x(\rho') = AJ_0(k_c\rho') \exp\left(\frac{-jk\rho'^2}{2\ell}\right) \tag{37}$$

where $k_c a/2$ is 2.405, the first zero of the zero order Bessel function J_0; ℓ is the slant length of the horn; and A is a constant. This aperture field is similar to that of Fig. 10c, and the resulting E and H patterns are similarly equal down to about -25 dB. Some universal patterns are shown in Fig. 11. Cross-polarization fields are also about -30 dB from the axial values, but now over a bandwidth of 2–1 or more.

Broadband axisymmetric patterns with low cross-polarization make corrugated horns particularly attractive as feeds for reflectors. Low cross-polarization allows the use of dual polarization to double the capacity of the system. Another notable feature for this application is that the position of the E- and H-plane pattern phase centers coincide. Figure 12 shows the distance of the phase center from the horn apex, divided by the slant length, of small flare angle conical [8] and corrugated [9] horns for values of the phase parameter s given by (2). For a conical horn the E-plane phase center is significantly farther from the aperture than the H-plane phase center. Thus, if a conical horn is used to feed a parabolic reflector, the best location for the feed is approximately midway between the E- and H-plane phase centers. With a corrugated horn such a compromise is not required, so it is inherently more efficient.

Corrugated horns may have wide flare angles, and their aperture size for optimum gain decreases correspondingly. For example, with a semiflare angle of 20°, the optimum aperture diameter is about 8λ, whereas for a semiflare angle of 70° it is 2λ. Wide-flare corrugated horns are sometimes called "scalar horns" because of their low cross-polarization levels.

For radio astronomy telescope feeds and other space-science applications, efficient corrugated horns have been

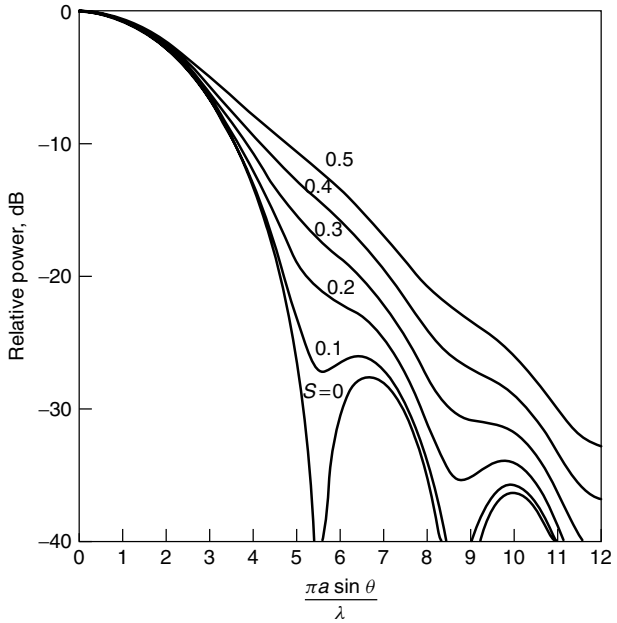


Figure 11. Universal patterns of small-flare-angle corrugated horns as a function of the parameter $s = a^2/8\lambda\ell$. (Copyright 1984, McGraw-Hill, Inc. from Love [2].)

made by electroforming techniques for frequencies up to 640 GHz. Their axisymmetric radiation patterns with very low sidelobe levels resemble Gaussian beams, which is often essential at submillimeter wavelengths.

8. PROFILE HORNS

Most corrugated horns are conical with a constant flare angle. Figure 13 shows a profile conical horn in which the flare angle varies as on a sine-squared or similar

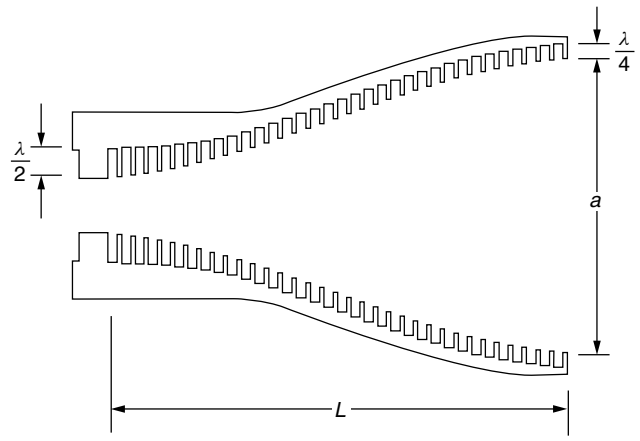


Figure 13. A profile corrugated horn. (Copyright 1994, IEE, from Olver et al. [7].)

curve along its length. This arrangement provides a horn shorter than a conical corrugated horn of similar beamwidth, with a better impedance match due to the curved profile at the throat and an essentially in-phase aperture field distribution due to the profile at the aperture. Consequently the aperture efficiency is higher than that of conical corrugated horns. The phase center of the horn is near the aperture center and remains nearly fixed over a wide frequency band. Radiation patterns of a short profile horn similar to that of Fig. 13, but with hyperbolic profile curves, are shown in Fig. 14 [10]. A Gaussian profile curve has also been used. All produce patterns similar to those of a Gaussian beam, such as is radiated from the end of an optical fiber supporting the HE_{11} mode. The performance of this small horn as a feed seems close to ideal, but larger-profile horns may exhibit higher sidelobe levels due to excitation of the HE_{12} mode at the aperture.

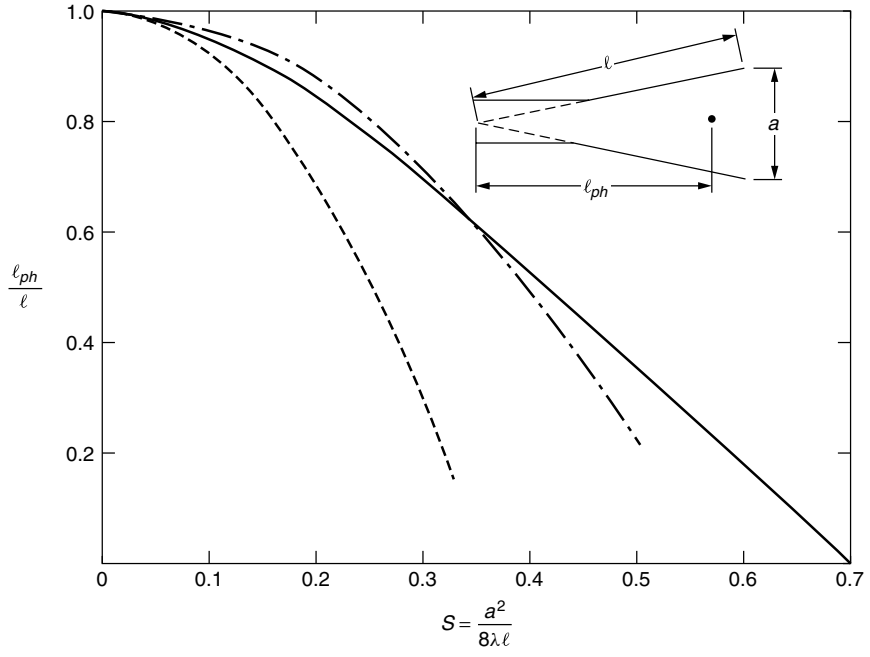


Figure 12. Normalized distance of the phase center from the apex of conical (- - - - *E*-plane, — · — · — *H*-plane and corrugated (_____) horns). (Data from Milligan [8] and Thomas [9].)

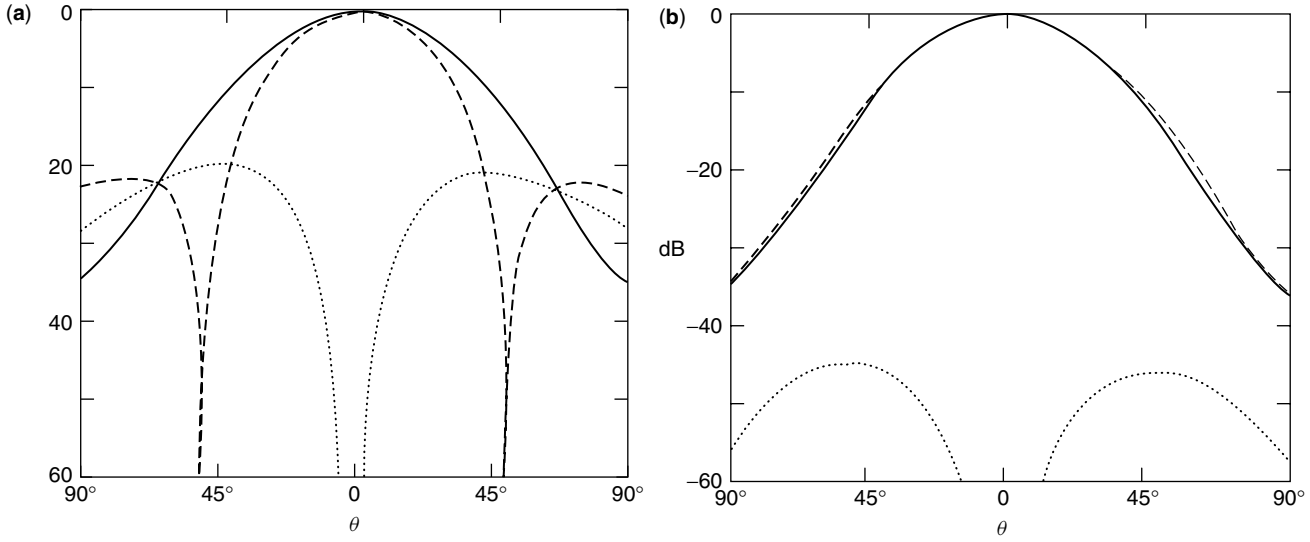


Figure 14. (**a**) Far field radiation patterns of TE_{11} mode and (**b**) radiation patterns of a profile corrugated horn of aperture $a = 15.8$ mm and length $L = 26.7$ mm at 30 GHz (- - - - - E-plane, _____ H-plane . . . . . crosspolarization). (Copyright 1997, IEEE, from Gonzalo et al. [10].)

9. HORN IMPEDANCE

Antennas must be well matched to their transmission lines to ensure a low level of reflected power. In microwave communications systems levels below −30 dB are commonly required.

The impedance behavior of a horn depends on the mismatch at the waveguide/horn junction and at its aperture. For an E-plane sectoral horn, reflections from these discontinuities are comparable in magnitude, and since they interfere, the total reflection coefficient oscillates with frequency and the input voltage standing-wave ratio (VSWR) may vary from 1.05 at high frequencies to 1.5 at the lowest frequency. With E-plane sectoral horns aperture reflection is much stronger than junction reflection, so their VSWRs increase almost monotonically with decreasing frequency. An inductive iris in the waveguide near the E-plane horn junction can match its discontinuity. A capacitive iris may be similarly used for an H-plane sectoral horn. Aperture reflections in these horns may be matched with dielectric covers.

Pyramidal horns of sufficient size and optimum design tend to be inherently well matched to their waveguide feeds because the E/H-plane aperture and flare discontinuities partially cancel. For example, a 22-dB-gain horn has a VSWR of about 1.04 and an 18 dB horn a VSWR of less than 1.1.

Conical horns fed by circular waveguides supporting the dominant TE_{11} mode have an impedance behavior similar to that of pyramidal horns of comparable size fed by rectangular waveguides. The waveguide/horn discontinuities of both horns may be matched by an iris placed in the waveguide near the junction. A broader bandwidth match is provided by a curved transition between the interior walls of the waveguide and the horn. Broadband reduction of aperture reflection may be similarly reduced by a curved surface of a few wavelengths' radius. Such "aperture-matched" horns also have lower sidelobe levels and less back radiation in their E-plane patterns than do conventional pyramidal and conical horns. Their H-plane flare patterns are affected little by such aperture matching because the electric field vanishes at the relevant edges.

For dual-mode and corrugated horns there are also negligible fields at the aperture edges and hence little diffraction there. Corrugated horns with initial groove depths near the throat of about a half-wavelength and which gradually decrease to a quarter-wavelength near the aperture, as in Fig. 13, are well matched at both throat and aperture. For most well-designed corrugated horns a VSWR of less than 1.25 is possible over a frequency range of about 1.5−1. Dual-mode horns using a step discontinuity as in Fig. 10d may have a VSWR of 1.2−1.4. If an iris is required for a match, the frequency bandwidth will, of course, be limited. Conical and pyramidal horns using flare angle changes to generate the higher-order modes can have VSWRs less than 1.03 and require no matching devices.

BIOGRAPHY

Edward V. Jull received his B.Sc.degree in engineering physics from Queen's University, Kingston, Ontario, Canada in 1956, a Ph.D in electrical engineering in 1960 and a D.Sc.(Eng.) in 1979, both from the University of London, United Kingdom. He was with the Division of Radio and Electrical Engineering of the National Research Council, Ottawa, Canada, from 1956 to 1957 in the microwave section and from 1961 to 1972 in the antenna engineering section. From 1963 to 1965 he was a guest researcher in the Laboratory of Electromagnetic Theory of the Technical University of Denmark in Lyngby, and the Microwave Department of the Royal Institute of Technology in Stockholm, Sweden. In 1972, he joined the Department of Electrical Engineering at the University

of British Columbia in Vancouver, British Columbia, Canada. He became professor emeritus in 2000, but remains involved in teaching and research on aperture antennas and diffraction theory and is the author of a book so titled. He was president of the International Union of Radio Science (URSI) from 1990 to 1993.

BIBLIOGRAPHY

1. J. F. Ramsay, Microwave antenna and waveguide techniques before 1900, *Proc. IRE* **46**: 405–415 (1958).
2. A. W. Love, Horn antennas, in R. C. Johnson and H. Jasik, eds., *Antenna Engineering Handbook*, 2nd ed., McGraw-Hill, New York, 1984, Chap. 15.
3. A. W. Love, ed., *Electromagnetic Horn Antennas*, IEEE Press, Piscataway, NJ, 1976.
4. R. E. Collin, *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985.
5. K. Liu, C. A. Balanis, C. R. Birtcher, and G. C. Barber, Analysis of pyramidal horn antennas using moment methods, *IEEE Trans. Antennas Propag.* **41**: 1379–1389 (1993).
6. J. F. Nye and W. Liang, Theory and measurement of the field of a pyramidal horn, *IEEE Trans. Antennas Propag.* **44**: 1488–1498 (1996).
7. A. D. Olver, P. J. B. Clarricoats, A. A. Kishk, and L. Shafai, *Microwave Horns and Feeds*, IEE Electromagnetic Waves Series, Vol. 39, IEE, London, 1994.
8. T. Milligan, *Modern Antenna Design*, McGraw-Hill, New York, 1985, Chap. 7.
9. B. MacA. Thomas, Design of corrugated horns, *IEEE Trans. Antennas Propag.* **26**: 367–372 (1978).
10. R. Gonzalo, J. Teniente, and C. del Rio, Very short and efficient feeder design from monomode waveguide, *IEEE Antennas and Propagation Soc. Int. Symp. Digest*, Montreal, 1997, pp. 468–470.
11. E. V. Jull, *Aperture Antennas and Diffraction Theory*, IEE Electromagnetic Waves Series, Vol. 10, IEE, London, 1981.

HUFFMAN CODING

EN-HUI YANG
DA-KE HE
University of Waterloo
Waterloo, Ontario, Canada

1. INTRODUCTION

Consider a data sequence $x = x_1x_2 \cdots x_n$, where each x_i is a letter from an alphabet $\mathcal{A}$. For example, the sequence x may be a text file, an image file, or a video file. To achieve efficient communication or storage, one seldom transmits or stores the raw-data sequence x directly; rather, one usually transmits or stores an efficient binary representation of x, from which the raw data x can be reconstructed. The process of converting the raw data sequence into its efficient binary representation is called *data compression*.

A simple data compression scheme assigns each letter $a \in \mathcal{A}$ a binary sequence $C(a)$ called the *codeword* of the letter a and then replaces each letter x_i in x by its codeword $C(x_i)$, yielding a binary sequence $C(x_1)C(x_2) \cdots C(x_n)$. The mapping C, which maps each letter $a \in \mathcal{A}$ into its codeword $C(a)$, is called a *code*. For example, the American Standard Code for Information Interchange (ASCII) assigns a 7-bit codeword to each letter in the alphabet of size 128 consisting of numbers, English letters, punctuation marks, and some special characters.

The ASCII code is a *fixed-length code* because all codewords have the same length. Fixed-length codes are efficient only when all letters are equally likely. In practice, however, letters appear with different frequencies in the raw-data sequence. In this case, one may want to assign short codewords to letters with high frequencies and long codewords to letters with low frequencies, thus making the whole binary sequence $C(x_1)C(x_2) \cdots C(x_n)$ shorter. In general, *variable-length codes*, in which codewords may be of different lengths, are more efficient than fixed-length codes.

For a variable-length code, the *decoding* process — the process of recovering x from $C(x_1)C(x_2) \cdots C(x_n)$ — is a bit more complicated than in the case of fixed-length codes. A variable-length code (or simply code) is said to be *uniquely decodable* if one can recover x from $C(x_1)C(x_2) \cdots C(x_n)$ without ambiguity. As the uniquely decodable concept suggests, not all one-to-one mappings from letters to codewords are uniquely decodable codes [1, Table 5.1]. Also, for some uniquely decodable code, the decoding process is quite involved; one may have to look at the entire binary sequence $C(x_1)C(x_2) \cdots C(x_n)$ to even determine the first letter x_1. A uniquely decodable code with *instantaneous* decoding capability — once a codeword is received, it can be decoded immediately without reference to the future codewords — is called a *prefix code*. A simple way to check whether a code is a prefix code is to verify the *prefix-free* condition: a code C is a prefix code if and only if no codeword in C is a prefix of any other codeword. A prefix code can also be represented by a binary tree in which terminal nodes are assigned letters from $\mathcal{A}$ and the codeword of a letter is the sequence of labels read from the root to the terminal node corresponding to the letter.

Example 1. Consider a code C that maps letters in $\mathcal{A} = \{a_1, a_2, \ldots, a_5\}$ to codewords in $\mathcal{B} = \{0, 100, 101, 110, 111\}$, where $C(a_i)$ is the ith binary string in $\mathcal{B}$ in the indicated order. The code C is a prefix code. If a string 011001010110 is received, one can easily parse it into 0, 110, 0, 101, 0, 110 and decode it into $a_1a_4a_1a_3a_1a_4$. Note that the first 0 is a codeword. After it is received, it can be decoded immediately into a_1 without looking at the next digit. The binary tree corresponding to C is shown in Fig. 1.

A uniquely decodable code C satisfies the following *Kraft–McMillan inequality* [1, Chap. 5]:

$$\sum_{a \in \mathcal{A}} 2^{-l(a)} \leq 1$$

where $l(a)$ is the length of $C(a)$. Conversely, it can be shown that given any set of lengths $l(a)$ satisfying the Kraft–McMillan inequality, one can find a prefix code C such that $C(a)$ has length $l(a)$ for any $a \in \mathcal{A}$.

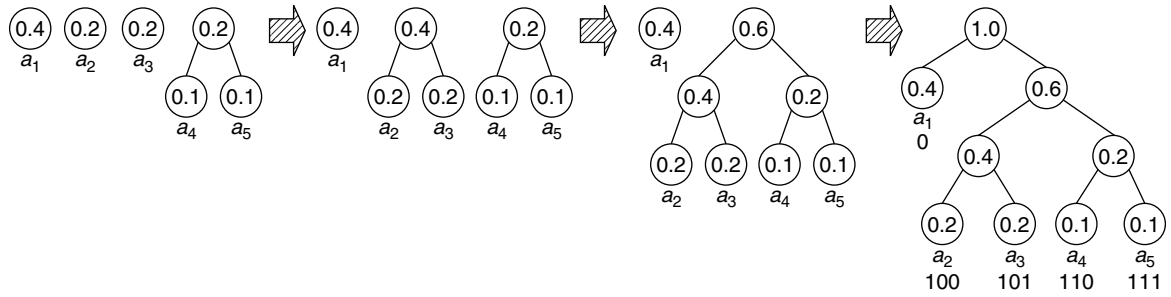


Figure 1. The Huffman code C generated by the Huffman coding algorithm.

Therefore, in view of the above result, it suffices to find an optimal prefix code C such that the total length $nr(C)$ of $C(x_1)C(x_2)\cdots C(x_n)$ is a minimum, where

$$r(C) = \sum_{a\in\mathcal{A}} p(a)l(a)$$

denotes the *average codeword length* of C in bits per letter, and $p(a)$ is the normalized frequency of a in x if x is deterministic and the probability of a if x is random.

Given a probability distribution or a set of normalized frequencies $\{p(a) : a \in \mathcal{A}\}$, an interesting problem is how to find such an optimal prefix code C. This problem is solved when the size of A is finite. Instead of performing an exhaustive search among all prefix codes, Huffman [2] proposed an elegant algorithm in 1952, which is now known as the *Huffman coding algorithm*, to generate optimal prefix codes based on a set of probabilities or frequencies. The resulting optimal codes are called *Huffman codes*.

2. HUFFMAN CODING ALGORITHM

Any elegant algorithm has its recursive procedure. There is no exception in the case of the Huffman coding algorithm. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ with $2 \le J < \infty$. For each $1 \le j \le J$, rewrite $p(a_j)$ and $l(a_j)$ as p_j and l_j, respectively. Let C be an optimal prefix code such that

$$r(C) = \sum_{j=1}^{J} p_j l_j \tag{1}$$

is minimized among all prefix codes. The following properties of the optimal code C are helpful to uncover the recursive procedure of the Huffman coding algorithm:

P1. C is a complete prefix code, that is, the binary tree corresponding to C is a complete binary tree in which every node other than the root has its sibling.

P2. If $p_i > p_j$, then $l_i \le l_j$.

P3. The two longest codewords of C, which correspond to the two least likely letters, have the same length.

P4. If in the binary tree corresponding to C, terminal nodes with the same depth are rearranged properly, then the two least likely letters are sibling terminal nodes.

Property P1 is straightforward. If a node other than the root has no sibling, then this node can be merged with its parent. As a result, all codewords of the terminal nodes in the subtree rooted at this node are shortened by 1, and the resulting prefix code will be better than C, which contradicts to the optimality of C. Property P2 is equivalent to the principle of assigning short codewords to highly likely letters and long codewords to less likely letters. Properties P3 and P4 are implied by properties P1 and P2.

Property P4 implies a recursive procedure to design an optimal prefix code C given the probabilities $p_1, p_2, \ldots, p_J$. Rewrite C as C_J. Suppose that p_{J-1} and p_J are the two least probabilities. From property P4, we know that a_{J-1} and a_J are sibling terminal nodes in the binary tree corresponding to C_J. Merge a_{J-1} and a_J with their common parent, which then becomes a terminal node corresponding to a merged letter with probability $p_{J-1} + p_J$. The new binary tree has $J-1$ terminal nodes and gives rise to a new prefix code C_{J-1} with $J-1$ codewords of lengths $l_1, l_2, \ldots, l_{J-2}, l_{J-1} - 1$. The average codeword length of C_J is related to that of C_{J-1} by

$$r(C_J) = r(C_{J-1}) + p_{J-1} + p_J$$

Since the quantity $p_{J-1} + p_J$ is independent of C_{J-1}, minimizing $r(C_J)$ is equivalent to minimizing $r(C_{J-1})$. Consequently, in order to design an optimal code C_J for J letters, we can first design an optimal code C_{J-1} for $J-1$ letters, and then extend C_{J-1} to C_J while maintaining optimality. Recursively reduce the alphabet size by merging the two least likely letters of the corresponding alphabet. We finally reduce the problem to design an optimal code C_2 for two letters, for which the solution is obvious; that is, we assign 0 to one letter and 1 to the other. Since $r(C_j)$ is minimized for each $2 \le j \le J$, we find that $C_2, C_3, \ldots, C_J$ are all optimal prefix codes with respect to their corresponding probability distributions. This is the essence of the recursive procedure of the Huffman coding algorithm.

To summarize, given a set of probabilities, the Huffman coding algorithm generates optimal prefix codes for $\mathcal{A}$ according to the following procedure:

Step 1. Start with $m = J$ trees, each of which consists of exactly one node corresponding to a letter in $\mathcal{A}$. Set the weight of each node as the probability of the corresponding letter.

Step 2. For $m \geq 2$, find the two trees $T_1^{(m)}$ and $T_2^{(m)}$ with the least weights at their roots. Combine $T_1^{(m)}$ and $T_2^{(m)}$ into a new tree so that in the new tree, the roots of $T_1^{(m)}$ and $T_2^{(m)}$ are siblings with the root of the new tree as their parent. The root of the new tree carries a weight equal to the sum of the weights of its two children. The total number of trees m is reduced by 1.

Step 3. Repeat step 2 for $m = J, J-1, \ldots, 2$ until only one tree is left. The final tree gives rise to an optimal prefix code.

The following example further illustrates the procedure.

Example 2. Consider an alphabet $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5\}$ with the probability distribution $(0.4, 0.2, 0.2, 0.1, 0.1)$. The Huffman coding algorithm generates the prefix code C in Example 1 recursively as shown in Fig. 1. By convention, the left and right branches emanating from an internal node in a binary tree are labeled 0 and 1, respectively.

It is worthwhile to point out that the Huffman codes generated by the Huffman coding procedure are not unique. Whenever there are two or more pairs of trees having the least weights at their roots in step 2, one can combine any such pair of trees into a new tree, resulting in a possibly different final tree. In Example 2, after merging a_4 and a_5 into one letter, say, a_4', with probability 0.2, we can choose to merge a_3 and a_4' instead of merging a_2 and a_3 as in Fig. 1. Continuing the algorithm, we can get another prefix code C' with the binary codeword set $\mathcal{B}' = \{0, 10, 110, 1110, 1111\}$ in which the jth string represents the codeword for a_j in $\mathcal{A}$, $1 \leq j \leq 5$. It is easy to verify that these both C' and C have the same average codeword length of 2.1 bits. This non-uniqueness will allow us later to develop *adaptive Huffman coding algorithms*.

3. ENTROPY AND PERFORMANCE

The entropy of a probability distribution $(p_1, p_2, \ldots, p_J)$ is defined as

$$H(p_1, p_2, \ldots, p_J) \triangleq -\sum_{j=1}^{J} p_j \log p_j$$

where log stands for the logarithm with base 2 and the entropy is measured in *bits*. The entropy represents the ultimate compression rate in bits per letter one can possibly achieve with all possible data compression schemes.

A *figure of merit* of a prefix code is its performance against the entropy. It can be shown [1] that if C is a Huffman code with respect to the distribution $(p_1, p_2, \ldots, p_J)$, then its average codeword length $r(C)$ in bits per letter is within one bit of the entropy.

Theorem 1. The average codeword length $r(C)$ of a Huffman code with respect to $(p_1, p_2, \ldots, p_J)$ satisfies

$$H(p_1, p_2, \ldots, p_J) \leq r(C) < H(p_1, p_2, \ldots, p_J) + 1 \qquad (2)$$

The upper bound in (2) is uniform and applies to every distribution. For some distributions, however, the difference between $r(C)$ and $H(p_1, p_2, \ldots, p_J)$ may be well below 1. For detailed improvements on the upper and lower bounds, the reader is referred to Gallager [3], Capocelli et al. [4], and Capocelli and De Santis [5].

Another method to improve the upper bound is to design a Huffman code C' for an extended alphabet $\mathcal{A}^N$, where $\mathcal{A}^N$ consists of all length N strings from $\mathcal{A}$. In this case, one assigns a codeword to each block of N letters rather than a single letter. Accordingly, $r(C')$ is the average codeword length in bits per block and is within one bit of the block entropy. Thus, $r(C')/N$, the average codeword length in bits per letter, is within $1/N$ bits of the entropy per letter. As N increases, $r(C')/N$ can be arbitrarily close to the entropy per letter. Since the complexity of the Huffman coding algorithm grows exponentially with respect to N, this method works only when both N and the size of $\mathcal{A}$ are small.

4. HUFFMAN CODING FOR AN INFINITE ALPHABET

Since the Huffman coding algorithm constructs a binary tree using the bottom–up approach by successively merging the two least probable letters in the alphabet, it cannot be applied directly to infinite alphabets. Let $\mathcal{A} = \{0, 1, 2, \ldots\}$. Indeed, given an arbitrary distribution $(p_0, p_1, p_2, \ldots)$, the problem of constructing an optimal prefix code C with respect to $(p_0, p_1, p_2, \ldots)$ is still open at the writing of this article, even though it can be shown [6] that such an optimal prefix code exists.

However, when the distribution is a geometric probability distribution

$$p_i = (1-\theta)\theta^i, \qquad i \geq 0 \qquad (3)$$

where $0 < \theta < 1$, a simple procedure does exist for construction of the optimal prefix code C. In this case, given each $i \in \mathcal{A}$, one can even compute the codeword $C(i)$ without involving any tree manipulation; this property is desirable in the case of infinite alphabets since there is no way to store an infinite binary tree. The procedure was first observed by Golomb [7] in the case of $\theta^k = \frac{1}{2}$ for some integer k, and later extended by Gallager and Voorhis [8] to general geometric distributions in (3). The corresponding optimal prefix code is now called the *Golomb code* in the case of $\theta^k = \frac{1}{2}$ for some integer k, and the *Gallager–Voorhis code* in the general case.

The procedure used to construct the Gallager–Voorhis code for a geometrical probability distribution with the parameter θ is as follows:

Step 1. Find the unique positive integer k such that

$$\theta^k(1+\theta) \leq 1 < \theta^{k-1}(1+\theta) \qquad (4)$$

This unique k exists because $0 < \theta < 1$.

Step 2. If $k = 1$, let $C_k(0)$ denote the empty binary string. Otherwise, let $k = 2^n + m$, where n and m are positive integers satisfying $0 \leq m < 2^n$. For any integer $0 \leq j < 2^n - m$, let $C_k(j)$ be the binary representation of j padded with possible zeros to the left to ensure that the length of $C_k(j)$ is n. For $2^n - m \leq j < k$, let $C_k(j)$ be the binary

representation of $j + 2^n - m$ padded with possible zeros to the left to ensure that the length of $C_k(j)$ is $n + 1$. The constructed code C_k is a prefix code for $\{0, 1, \ldots, k-1\}$.

Step 3. To encode an integer $i \geq 0$, we find a pair of nonnegative integers (s, j) such that $i = sk + j$ and $0 \leq j < k$. The Gallager–Voorhis code encodes i into a codeword $G_k(i)$ consisting of s zeros followed by a single one and then by $C_k(i)$.

Example 3. Suppose that $\theta^3 = \frac{1}{2}$. In this case, $k = 3$, and

$$C_3(0) = 0, C_3(1) = 10, C_3(2) = 11$$

Table 1 illustrates Golomb (Gallager-Voorhis) codewords $G_3(i)$ for integers $i = 0, 1, 2, \ldots, 11$.

5. ADAPTIVE HUFFMAN CODING

In previous sections, we have assumed that the probability distribution is available and known to both the encoder and decoder. In many practical applications, however, the distribution is unknown. Since Huffman coding needs a distribution to begin with, to encode a sequence $x = x_1 x_2 \cdots x_n$, one way to apply Huffman coding is to use the following *two-pass* approach:

Pass 1. Read the sequence x to collect the frequency of each letter in the sequence. Use the frequencies of all letters to estimate the probabilities of these letters, and design a Huffman code based on the estimated probability distribution. Send the estimated probability distribution or the designed Huffman code to the user;

Pass 2. Use the designed Huffman code to encode the sequence.

This two-pass coding scheme is not desirable in applications such as streaming procedures, where timely encoding of current letters is required. It would be nice to have a *one-pass* coding scheme in which we estimate the probabilities of letters on the fly on the basis of the previously encoded letters in x and adaptively choose a prefix code based on the estimated probability distribution to encode the current letter. Faller [9] and Gallager [3] independently developed a one-pass coding algorithm called the *adaptive Huffman coding algorithm*, also known as the *dynamic Huffman coding algorithm*. It was later improved by Knuth [10] and Vitter [11].

Table 1. Golomb Codewords $G_3(i)$ for Integers $i = 0, 1, 2, \ldots, 11$

Integer i	$G_3(i)$	Integer i	$G_3(i)$
0	10	6	0010
1	110	7	00110
2	111	8	00111
3	010	9	00010
4	0110	10	000110
5	0111	11	000111

Suppose that the sequence $x = x_1 x_2 \cdots x_n$ is drawn from the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ with $2 \leq J < \infty$. Before encoding x_1, it is reasonable to assume that all letters $a \in \mathcal{A}$ are equally likely since we have no knowledge about x other than the alphabet $\mathcal{A}$. Maintain a counter $c(a_j)$ for each letter a_j, $1 \leq j \leq J$. All counters are initially set to 1. The initial probability distribution is

$$\left(\frac{c(a_1)}{\sum_{j=1}^{J} c(a_j)}, \frac{c(a_2)}{\sum_{j=1}^{J} c(a_j)}, \ldots, \frac{c(a_J)}{\sum_{j=1}^{J} c(a_j)} \right) = \left(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J} \right)$$

Pick an initial Huffman code C_1 for $\mathcal{A}$ based on the initial probability distribution. To encode $x = x_1 x_2 \cdots x_n$, the adaptive Huffman coding algorithm works as follows:

Step 1. Use the Huffman code C_i to encode x_i.

Step 2. Increase $c(x_i)$ by 1.

Step 3. Update C_i into a new prefix code C_{i+1}, so that C_{i+1} is a Huffman code for the new probability distribution

$$\frac{c(a_1)}{\sum_{j=1}^{J} c(a_j)}, \frac{c(a_2)}{\sum_{j=1}^{J} c(a_j)}, \ldots, \frac{c(a_J)}{\sum_{j=1}^{J} c(a_j)}$$

Step 4. Repeat steps 1–3 for $i = 1, 2, \ldots, n$ until all letters in x are encoded.

It is clear that step 3 is the major step. To understand it properly, let us first discuss the sibling property of a complete binary tree.

5.1. Sibling Property

Assign a positive weight w_j to each letter a_j, $1 \leq j \leq J$. Let C be a complete prefix code for $\mathcal{A}$. Let T be the binary tree corresponding to C. Label the nodes of T by an index from $\{1, 2, \ldots, 2J-1\}$. Assign a weight to each node in T recursively; each terminal node carries the weight of the corresponding letter in $\mathcal{A}$, and each internal node carries a weight equal to the sum of the weights of its two children. A complete prefix code, C, is said to satisfy the *sibling property* with respect to weights $w_1, w_2, \ldots, w_J$ if the nodes of T can be arranged in a sequence $i_1, i_2, \ldots, i_{2J-1}$ such that (continuing the "properties" list from Section 2, above)

P5. $w(i_1) \leq w(i_2) \leq \cdots \leq w(i_{2J-1})$, where $w(i_j)$ denotes the weight of the node i_j.

P6. nodes i_{2j-1} and i_{2j} are siblings for any $1 \leq j \leq J$; and the parent of nodes i_{2j-1} and i_{2j} does not precede nodes i_{2j-1} and i_{2j} in the sequence.

The sibling property is related to the Huffman coding algorithm.

Example 4. We now revisit the Huffman code in Example 2. It is not hard to see that the Huffman code shown in Fig. 1 satisfies the sibling property with respect to weights $w_1 = 4$, $w_2 = 2$, $w_3 = 2$, $w_4 = 1$, and $w_5 = 1$.

Example 5. Let us now increase the weight w_2 in Example 4 by 1. Accordingly, the weight of each node along the path from the terminal node corresponding to a_2 to the root increases by 1, as shown in Fig. 2. It is easy to see that the prefix code in Fig. 2 is not a Huffman code for the probability distribution

$$\left(\frac{w_1}{\sum_{j=1}^{5} w_j}, \frac{w_2}{\sum_{j=1}^{5} w_j}, \ldots, \frac{w_5}{\sum_{j=1}^{5} w_j}\right) = \left(\frac{4}{11}, \frac{3}{11}, \frac{2}{11}, \frac{1}{11}, \frac{1}{11}\right)$$

Moreover, this prefix code does not satisfy the sibling property with respect to weights 4, 3, 2, 1, and 1, either.

In general, the following theorem is implied by the Huffman coding procedure.

Theorem 2. A complete prefix code C is a Huffman code for the probability distribution

$$\frac{w_1}{\sum_{j=1}^{J} w_j}, \frac{w_2}{\sum_{j=1}^{J} w_j}, \ldots, \frac{w_J}{\sum_{j=1}^{J} w_j}$$

if and only if C satisfies the sibling property with respect to weights $w_1, w_2, \ldots, w_J$.

Update C_i into C_{i+1}: Note that C_i is a Huffman code for the current probability distribution

$$\frac{c(a_1)}{\sum_{j=1}^{J} c(a_j)}, \frac{c(a_2)}{\sum_{j=1}^{J} c(a_j)}, \ldots, \frac{c(a_J)}{\sum_{j=1}^{J} c(a_j)}$$

Think of $c(a_j)$ as the weight of a_j, that is, $w_j = c(a_j)$, $1 \le j \le J$. Let the nodes of the binary tree corresponding to C_i be arranged in a sequence $i_1, i_2, \ldots, i_{2J-1}$ such that properties P5 and P6 are satisfied. Let $j_0, j_1, \ldots, j_l$ be a sequence such that $i_{j_0}, i_{j_1}, \ldots, i_{j_l}$ is the sequence of nodes leading from the terminal node corresponding to x_i to the root.

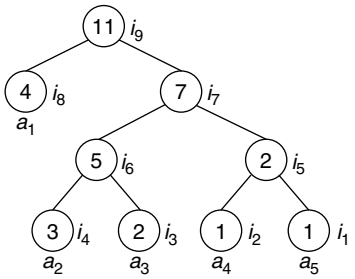


Figure 2. The prefix code in Example 4 for weights 4, 3, 2, 1, and 1.

Case 1. If for any $0 \le k < l$

$$w(i_{j_k}) < w(i_{j_k+1}) \tag{5}$$

then after the weights of nodes i_{j_k}, $0 \le k < l$, are updated, C_i still satisfies the sibling property. The update of the weights of nodes is due to the increment of $c(x_i)$ by 1 after x_i is encoded by C_i. In this case, C_i is also a Huffman code for the new probability distribution after $c(x_i)$ increases by 1. Hence $C_{i+1} = C_i$.

Case 2. If the inequality (5) is not valid for some k, then we can obtain C_{i+1} from C_i by exchanging some subtrees rooted at nodes of equal weight. Let j'_0 be the largest integer such that $w(i_{j'_0}) = w(i_{j_0})$. Having j'_k defined, we let j'_{k+1} be the largest integer such that $w(i_{j'_{k+1}})$ is equal to the weight of the parent of the node $i_{j'_k}$. If $j'_k = 2J - 1$, that is, if the node $i_{j'_k}$ is the root of the binary tree, the procedure terminates. Denote the maximum k by l'. It is easy to see that $l' < l$. In this case, the following must be done:

Step 1. Exchange the subtree rooted at the node i_{j_0} with the subtree rooted at the node $i_{j'_0}$.

Step 2. For $k = 0, 1, \ldots, l' - 1$, exchange (in the new binary tree resulting from the last exchange operation) the subtree rooted at the parent of the node $i_{j'_k}$ with the subtree rooted at the node $i_{j'_{k+1}}$. (The two roots of the two subtrees are not exchanged since they have the same weight.)

Step 3. Update the weight of each node along the path from node $i_{j'_0}$ to the root.

The final binary tree satisfies properties P5 and P6, and gives rise to the Huffman code C_{i+1}.

Example 6. Suppose that a_2 is the current letter to be encoded by the prefix code in Fig. 2. Denote the code by C_i. In Example 5, we know that C_i does not satisfy the sibling property with respect to weights 4, 3, 2, 1, and 1. Exchange the subtree rooted at node i_4 with the subtree rooted at node i_5, and update relevant weights accordingly. We get the two binary trees shown in Fig. 3. The binary tree on the right-hand side of Fig. 3 gives rise to a Huffman code for the probability distribution $\left(\frac{4}{11}, \frac{3}{11}, \frac{2}{11}, \frac{1}{11}, \frac{1}{11}\right)$ after the current letter a_2 is encoded by C_i.

The following example illustrates the complete process of the adaptive Huffman coding algorithm.

Example 7. Let $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5, a_6\}$. Apply the adaptive Huffman coding algorithm to encode $x = a_1a_1a_2a_3$. The initial counters are $c(a_1) = c(a_2) = c(a_3) = c(a_4) = c(a_5) = c(a_6) = 1$. Pick C_1 in Fig. 4 to be the Huffman code for $\mathcal{A}$, based on the probability distribution $\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$. Encode the first letter a_1 by C_1, and then increase the counter $c(a_1)$ by 1. In this case, $l = 3, j_0 = 1, j_1 = 7, j_2 = 10, j_3 = 11, j'_0 = 6, j'_1 = 9, j'_2 = 11$, and $l' = 2$.

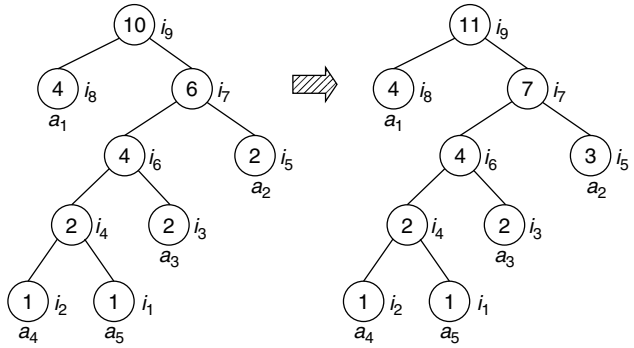


Figure 3. Exchange of subtrees and update of relevant weights.

By exchanging subtrees and updating all relevant weights, we get the Huffman code C_2 in Fig. 4. Encode the second letter a_1 by C_2, and then increase the counter $c(a_1)$ by 1. C_2 is updated into C_3 in Fig. 4 by exchanging subtrees rooted at nodes i_6 and i_8, and increasing all relevant weights by 1. Encode the third letter a_2 by C_3, and then increase the counter $c(a_2)$ by 1. C_3 is updated into C_4 in Fig. 4 by exchanging subtrees rooted at nodes i_2 and i_5, and updating relevant weights. Finally, encode the fourth letter a_3 by C_4. The codeword sequence is

$$C_1(a_1)C_2(a_1)C_3(a_2)C_4(a_3) = 00011001110$$

A very interesting fact about the adaptive Huffman coding algorithm is that as i is large enough, the adaptive prefix code C_i converges and is indeed a Huffman code for the true distribution. This is expressed in the following theorem.

Theorem 3. Apply the adaptive Huffman coding algorithm to encode a stationary ergodic source $X_1X_2 \cdots X_n \cdots$ taking values from the finite alphabet $\mathcal{A}$ with a common distribution $(p_1, \ldots, p_J)$. Then with probability one, C_i converges and for sufficiently large i, C_i itself is a Huffman code for the true distribution $(p_1, \ldots, p_J)$.

6. APPLICATIONS

The years since 1977 have witnessed widespread applications of Huffman coding in data compression, in sharp contrast with the first 25 years since the appearance of Huffman's groundbreaking paper [2]. This phenomenon can be explained by the increasing affordability of computing power and the growing demand of data compression to save transmission time and/or storage space. Nowadays, Huffman coding competes with state-of-the-art coding schemes such as arithmetic coding and Lempel–Ziv coding in applications requiring data compression. Since each letter in a data sequence to be compressed must be encoded into an integer number of bits, the compression performance of Huffman coding is often worse than that of arithmetic coding or Lempel–Ziv coding. However, in practical applications, the selection of a data compression scheme is based not only on its compression performance but also on its computational speed and memory requirement. Since a fixed Huffman code can be easily implemented as a lookup table in practice, Huffman coding has the advantages of real-time computational speed and the need of only a fixed amount of memory. Thus, Huffman coding remains a popular choice in time-critical applications or applications in which computing resources such as memory or computing power are limited. Some typical applications of Huffman coding in telecommunications are described in the following paragraphs.

6.1. Facsimile Compression

One of the earliest applications of Huffman coding in telecommunications is in *facsimile compression*. In 1980, the CCITT Group 3 digital facsimile standard was put into force by the Consultative Committee on International

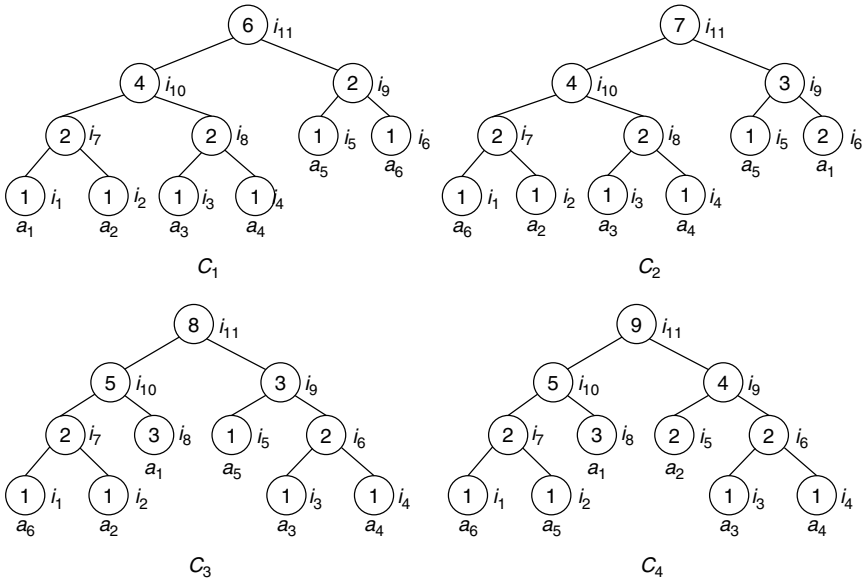


Figure 4. The Huffman codes C_1, C_2, C_3, and C_4 in Example 6.

Telephony and Telegraphy (CCITT), now the International Telecommunications Union (ITU). In the CCITT Group 3 standard, a predefined Huffman code was designed based on 11 typical fax documents recommended by the CCITT. For real-time processing of fax images, this Huffman code is implemented as a lookup table. It is estimated that a Group 3 fax system saves more than 85% of transmission time by compressing a typical business document of letter size.

6.2. Modem Data Compression

Early Modem data compression is another application of Huffman coding in telecommunications. The Microcom Networking Protocol (MNP) is a de facto standard for the modem industry. In MNP, MNP 5 is a modem data compression method that uses a much simplified variant of adaptive Huffman coding. In MNP 5, a set of 256 predefined prefix-free codewords are maintained in a table. When a letter from an alphabet of size 256 is to be encoded, MNP 5 selects a codeword according to the letter's recorded frequency. Thus, the mapping between codewords and letters are adaptively changing according to the data sequence. The compression performance of MNP 5 varies for different data. For standard text data, a modem generally can double its transmission speed by applying MNP 5. However, for some data like compressed images, MNP 5 will result in actual expansions of data, and thus slow down the modem's transmission.

6.3. Image Compression

The JPEG image compression standard developed by the Joint Photographic Experts Group uses Huffman coding as a residual coder after discrete-cosine transform (DCT), quantization, and run-length encoding [7]. Since the encoding of an image is not necessarily real-time, the JPEG standard allows two-pass Huffman coding in addition to the use of a predefined Huffman code, which is generated based on a group of typical images. The more recent JPEG-LS standard defines a new image compression algorithm that allows any image to be encoded, and then decoded, without any loss of information. In the JPEG-LS standard, variants of Golomb codes, which are called *Golomb–Rice codes* [12], are used to efficiently encode integers into easily computed codewords.

6.4. Audio/Video Compression

International standards for full-motion video compression include MPEG 1 and 2, which are both developed by the Moving Pictures Experts Group. MPEG layer 3, also known as MP3, is now a very popular audio coding standard. Similar to JPEG, MPEG 1 and 2 use Huffman coding as a residual coder after DCT, quantization, and run-length encoding. MP3 combines Huffman coding with modified DCT and quantization. A critical requirement for MP3, MPEG 1, and MPEG 2 is that the decoding of compressed audio and video data must be real-time. Therefore, in these standards Huffman codes are all predefined and implemented as lookup tables.

BIOGRAPHIES

En-hui Yang received the B.S. degree in applied mathematics from HuaQiao University, Qianzhou, China, and Ph.D. degree in mathematics from Nankai University, Tianjin, China, in 1986 and 1991, respectively. He joined the faculty of Nankai University in June 1991 and was promoted to Associate Professor in 1992. From January 1993 to May 1997, he held positions of Research Associate and Visiting Scientist at the University of Minnesota, Minneapolis–St. Paul (USA), the University of Bielefeld, Bielefeld, Germany, and the University of Southern California, Los Angeles (USA). Since June 1997, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada, where he is now a Professor and leading the Leitch — University of Waterloo multimedia communications lab. He also holds the position of Canada Research Chair in information theory and multimedia compression, and is a founding partner of SlipStream Data Inc, a high-tech company specializing solutions to reduce communication bandwidth requirements and speed up Internet connection. His current research interests are multimedia compression, digital communications, information theory, Kolmogorov complexity theory, source and channel coding, digital watermarking, quantum information theory, and applied probability theory and statistics. Dr. Yang is a recipient of several research awards, including the 2000 Ontario Premier's Research Excellence Award, Canada, and the 2002 Ontario Distinguished Researcher Award, Canada.

Da-ke He received the B.S. and M.S. degrees in electrical engineering from Huazhong University of Science and Technology, China, in 1993 and 1996, respectively. He joined Apple Technology, Zhuhai, China, in 1996 as a software engineer. Since 1999, he has been a Ph.D. candidate in the Electrical and Computer Engineering Department at the University of Waterloo, Canada, where he has been working on grammar-based data compression algorithms and multimedia data compression. His areas of interest are source coding theory, multimedia data compression, and digital communications.

BIBLIOGRAPHY

1. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
2. D. A. Huffman, A method for the construction of minimum redundancy codes, *Proc. IRE* **40**: 1098–1101 (1952).
3. R. G. Gallager, Variations on a theme by Huffman, *IEEE Trans. Inform. Theory* **IT-24**: 668–674 (1978).
4. R. M. Capocelli, R. Giancarlo, and I. J. Taneja, Bounds on the redundancy of Huffman codes, *IEEE Trans. Inform. Theory* **IT-32**: 854–857 (1986).
5. R. M. Capocelli and A. De Santis, New bounds on the redundancy of Huffman codes, *IEEE Trans. Inform. Theory* **IT-37**: 1095–1104 (1991).
6. T. Linder, V. Tarokh, and K. Zeger, Existence of optimal prefix codes for infinite source alphabets, *IEEE Trans. Inform. Theory* **IT-43**: 2026–2028 (1997).

7. S. W. Golomb, Run-length encodings, *IEEE Trans. Inform. Theory* **IT-12**: 399–401 (1966).

8. R. G. Gallager and D. C. Van Voorhis, Optimal source codes for geometrically distributed integer alphabets, *IEEE Trans. Inform. Theory* **IT-21**: 228–230 (1975).

9. N. Faller, An adaptive system for data compression, *Record 7th Asilomar Conf. Circuits, Systems and Computers*, 1973, pp. 593–597.

10. D. E. Knuth, Dynamic Huffman coding, *J. Algorithms* **6**: 163–180 (1985).

11. J. S. Vitter, Design and analysis of dynamic Huffman codes, *J. Assoc. Comput. Mach.* **34**: 825–845 (1987).

12. M. Weinberger, G. Seroussi, and G. Sapiro, The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS, *IEEE Trans. Image Process.* **9**: 1309–1324 (2000).

IMAGE AND VIDEO CODING

SHIPENG LI
Microsoft Research Asia
Beijing, P. R. China

WEIPING LI
WebCast Technologies, Inc.
Sunnyvale, California

1. INTRODUCTION

Image and video coding has been a very active research area for a long time. While transmission bandwidth and storage capacity have been growing dramatically, the demand for better image and video coding technology has also been growing. The reason is the ever-increasing demand for higher quality of images and video, which requires ever-increasing quantities of data to be transmitted and/or stored.

There are many different types of image and video coding techniques available. The term coding used to solely refer to compression of image and video signals. However, in recent years it is generalized more toward representation of image and video data that provides not only compression but also other functionalities. In this article, we still focus on discussions in the traditional coding sense (i.e., compression). We briefly touch on the topic of coding with different functionalities at the end.

To better understand the details of different image and video coding techniques and standards, many of which seem to be more art than science, a good understanding of the general problem of space of image and video coding is extremely important. The basic problem is to reduce the data rate required for representing images and video as much as possible. In compressing image and video data, often some distortion is introduced so that the received image and video signals may not be exactly the same as the original. Therefore, the second objective is to have as little distortion as possible. In theoretic source coding, data rate and distortion are the only two dimensions to be considered, and the objective is to have both rate and distortion as small as possible. Image and video coding is a type of source coding that also requires other practical considerations. One of the practical concerns is the complexity of a coding technique, which is further divided into encoding complexity and decoding complexity. Therefore, in image and video coding, complexity is the third dimension to be considered, and the additional objective is to have complexity lower than a given threshold. Yet another practical concern is the delay (or latency) of a coding technique, which is the time between when an image or a video frame is available at the input of the encoder and when the reconstructed image or video frame is available at the output of the decoder, excluding the transmission time from the output of the encoder to the input of the decoder. Delay is a critical parameter for two-way communications. For such applications, delay is the fourth dimension in image and video coding, and the objective is to have the delay lower than a given threshold. Therefore, in general, the problem space has four dimensions, namely, rate, distortion, complexity, and delay. The overall problem is to minimize both rate and distortion under a constraint of complexity and possibly another constraint of delay.

This article is organized as follows. In Section 2, some basic concepts of image and video signals are presented. Its goal is to establish a good basis for the characteristics of the source in image and video coding. Section 3 reviews some basic principles and techniques of image and video coding. This is to present some components in an image or video coding system. Section 4 is devoted to the existing and emerging image and video coding standards that exemplify the usage of many basic principles and techniques. Because image and video coding is a part of a communication system, standards are extremely important for practical applications. Section 5 concludes with some thoughts on the possible future research in image and video coding.

2. BASIC CONCEPTS OF IMAGE AND VIDEO SIGNALS

Before discussing image and video coding, some basic concepts about the characteristics of image and video signals are briefly presented in this section. Images and video are multidimensional signals. A grayscale image can be considered as a function of two variables $f(x, y)$ where x and y are the horizontal and vertical coordinates, respectively, of a two-dimensional plane, and $f(x, y)$ is the intensity of brightness at the point with coordinates (x, y). In most practical cases, the coordinates x and y are sampled into discrete values so that a grayscale image is represented by a two-dimensional array of intensity values. Each of the array elements is often called a picture element or pixel. By adding a new dimension in time, an image sequence is usually represented by a time sequence of two-dimensional spatial intensity arrays (images) or, equivalently, a three-dimensional spacetime array. Figure 1 illustrates the concepts of image and image sequence. A video signal is a special type of image sequence. It is different from a film, which is another type of image sequence. In addition to such a simple description, more aspects of image and video are presented in the following subsections.

2.1. Imaging

The term *imaging* usually refers to the process of generating an image by a certain physical means. Images and video may come from a rich variety of sources, from natural photographs to all sorts of medical images, from microscopy to meteorology, not necessarily directly

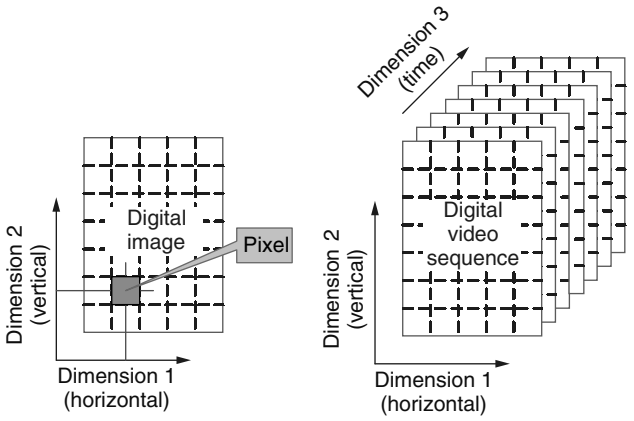


Figure 1. Image and image sequence.

perceptible by human eyes. They can be informally classified into three types according to their radiation sources: reflection sources, such as natural photographs and video; emission sources, such as MRI images; and absorption sources, such as X-ray images. Moreover, with the rapid deployment of multimedia computers, more and more artistic images and video are generated or processed synthetically with or without natural images or video as a basis.

2.2. Color Space

The visual experience of human eyes is much enriched with color information. Corresponding to the human perception system, a color image is most often represented with three primary color components: red (R), green (G), and blue (B). A number of other color coordinate systems can also be used in image processing, printing, and display systems. One particularly interesting color space is YIQ (luminance, in-phase chromatic, quadratic chromatic, also referred to as YUV, or YCbCr) commonly used in television or video systems. Luminance represents the brightness of the image and chrominance represents the color of the image. Conversion from one color space to another is usually defined by a color-conversion matrix. For example, in ITU-R Recommendation BT.709, the following color conversion is defined:

$$\begin{cases} Y = 0.7152G + 0.0722B + 0.2126R \\ Cb = -0.386G + 0.500B - 0.115R \\ Cr = -0.454G - 0.046B + 0.500R \end{cases} \quad (1)$$

2.3. Color Subsampling

Because chrominance is usually associated with slower amplitude variations than luminance and the human eyes are more sensitive to luminance than to chrominance, image and video coding algorithms can exploit this feature to increase coding efficiency by representing the chromatic components with a reduced spatial resolution or allocating fewer bits for them. Figure 2 gives some examples of different sampling schemes in the YUV color space. Three sampling schemes are commonly used in a video system. They are: $4:4:4$ (Y, U, and V with the same resolutions, Fig. 2 (a)); $4:2:2$ (U and V with half the horizontal resolution of Y, but the same vertical resolution, Fig. 2 (b)); and $4:2:0$ (U and V with both half the horizontal and vertical resolutions of Y, Fig. 2 (c)).

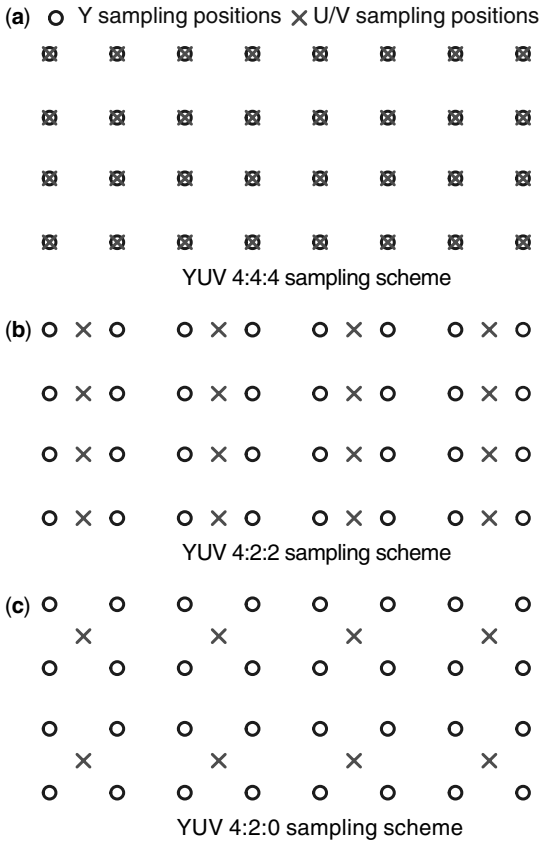


Figure 2. Examples of different sampling schemes in YUV color space.

2.4. Pixel Quantization

The image and video signals that exist abundantly in the environment are naturally analog. An analog image or video signal is a continuous function in a space/time domain and takes values that come from a continuum of possibilities. Before an analog image or video signal can be processed digitally, it must be converted to a digital format (or digitized) so that it becomes enumerable in pixel values in addition to space and time dimensions. Such an analog-to-digital (A/D) conversion process for a pixel value is called quantization. A commonly used quantization scheme when digitizing an analog image is uniform quantization, which maps the continuous-valued intensity to a finite set of nonnegative integers $\{0, \ldots, N-1\}$ through rounding operations, where N is a power of 2: $N = 2^n$. N is usually referred to as the number of gray levels and n is the number of bits allocated to each pixel. The most commonly used bit-depth is $n = 8$ for natural gray level images, whereas $n = 1$ is used for binary images, $n = 10$ is often used for studio quality video, and $n = 12$ is often used for medical images and infrared images. For color images, although different color components can be quantized jointly, most often they are quantized individually. For example, an RGB color image

is frequently represented with 24 bits per pixel, with 8 bits for each color component, which is commonly called a true color image. For an image with a YUV 4:2:2 color space and 8 bits per color component, there are $4*8 + 2*8 + 2*8 = 64$ bits for 4 pixels and equivalently 16 bits per pixel, which is commonly called a 16-bit color image. Similarly, an image with a YUV 4:2:0 color space and 8 bits per color component is commonly called a 12-bit color image.

2.5. Video Scanning and Frame Rate

Besides these commonalities with image signals, video signals have some special features. Unlike an image sequence obtained from film that is a time sequence of two-dimensional arrays, a video signal is actually a one-dimensional function of time. A video signal is not only sampled along the time axis, but also sampled along one space dimension (vertical). Such a sampling process is called scanning. The result is a series of time samples, or frames, each of which is composed of space samples, or scan lines. Therefore, video is a one-dimensional analog signal over time. There are two types of video scanning: progressive scanning and interlaced scanning. A progressive scan traces a complete frame, line by line from top to bottom, at a high refresh rate (>50 frames per second to avoid flickering). For example, video displayed on most computer monitors uses progressive scanning. It is well known that the frame rate of a film is 24 frames per second because the human brain interprets a film at 24 or more frames per second as "*continuous*" without a "*gap*" between any two frames. A major difference between a (scanned) video signal and a film is that all pixels in a film frame are illuminated at the same time for the same period of time, and the pixels at the upper left corner and the lower right corner of a frame of the (scanned) video signal are illuminated at different times due to the time period for scanning from the upper left corner to the lower right corner. This is why the frame rate of a video signal must be more than 50 frames per second to avoid flickering, while 24 frames per second is a sufficient rate for film.

2.6. Interlaced Scanning

A TV signal is a good example of interlaced (scan) video. Historically, interlaced video format was invented to achieve a good balance between signal bandwidth, flickering, and vertical resolution. As discussed in the previous subsection, the refresh rate for a video frame must be more than 50 times per second. The number of scan lines per frame determines the vertical resolution of a video frame. For a given refresh rate, the number of scan lines determines the bandwidth of the video signal, because, within the same time period of scanning one frame, more scan lines result in a faster change of intensity (i.e., higher bandwidth). To reduce the video signal bandwidth while maintaining the same refresh rate to avoid flickering, one must reduce the number of scan lines in one refresh period. The advantage of using interlaced scanning is that the vertical resolution is not *noticeably* reduced while the number of scan lines is reduced. This is possible because interlaced scanning refreshes every other line at each frame refresh and the full vertical resolution is covered in two refresh periods. However, the interlaced scan of more than two lines would result in noticeable reduction of the vertical resolution. The subframes formed by all the even or odd scan lines are called fields. Correspondingly, fields can be classified into even fields and odd fields, or top fields and bottom fields according to their relative vertical positions. The top and bottom fields are sent alternately to an interlaced monitor at a field rate equal to the refresh rate. Figure 3 (a) and (b) depict the progressive and interlaced video scanning processes, respectively. A video signal of either scan type can be digitized naturally by sampling horizontally along the scan line, which results in a single rectangular frame for progressive scan and two interlaced fields (in one frame) for interlaced scan. Figure 3 (c) and (d) illustrate such digitized progressive and interlaced video frames, respectively.

2.7. Uncompressed Digital Video

With the image and video signals in digital format, we can process, store, and transmit them digitally using computers and computer networks. However, the high volume nature of image and video data is prohibitive to many applications without efficient compression. For example, considering a 2-hour video with spatial resolution of 720×480 pixels per frame, YUV 4:2:2 format, 30 frames per second frame rate, and each color component quantized to 8 bits (1 byte), the total uncompressed data size of such a video sequence is $2 \times 60 \times 60 \times 30 \times 720 \times 480 \times 2 = 149$ Gbytes with a bit rate of $30 \times 720 \times 480 \times 2 \times 8 = 165$ Mbps, and it is still nowhere near the cinema quality. This is a huge amount of data for computers and networks with today's technology.

2.8. Redundancy and Irrelevancy

Fortunately, image and video data contain a great deal of redundancy and irrelevancy that can be reduced or removed. Redundancy refers to the redundant or duplicated information in an image or video signal and can generally be classified as spatial redundancy (correlation between neighboring pixel values), spectral redundancy (correlation between different color planes or spectral bands), temporal redundancy (correlation between adjacent frames in a sequence of images), statistical redundancy (nonuniform distribution of image and video pixel value), and so on. Irrelevancy refers to the part of the signal that is not noticeable by the receiver (e.g., the human visual system, or HVS). Image and video compression research aims at reducing the number of bits needed to represent an image or video signal by removing the redundancy and irrelevancy as much as possible, with or without noticeable difference to the human eye. Of course, with the rapid developments in high-capacity computers and high-speed computer networks, the tractable data size and bit rates will be increased significantly. However,

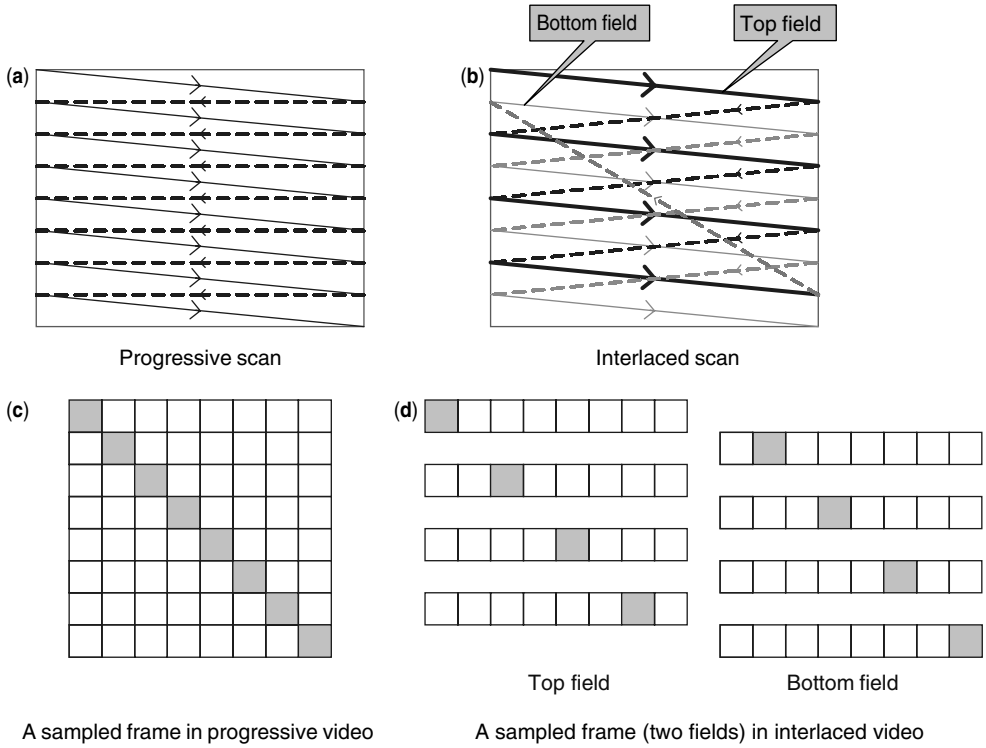


Figure 3. The comparison of progressive video and interlaced video.

the research and development of advanced algorithms for digital image and video compression continue to be necessary in the context of providing a novel and rich multimedia experience with much improved quality more efficiently, more flexibly, more robustly, and more ubiquitously.

2.9. Evaluation of Image and Video Coding Schemes

In practice, we face many choices of various image and video coding schemes. Even for the same coding scheme, we have the choice of different sets of parameters. It is important to first establish measures for quality and performance before any attempt to select the ones that best fit our needs. Some basic measurements commonly used in image and video compression are explained as follows.

The effectiveness of an image or video compression session is normally measured by the bit rate of the compressed bit stream generated, which is the average number of bits representing the compressed image or video signal, with average number of bits per pixel (bpp) for images and average number of bits per second (bps) for video. It can also be measured by the compression ratio defined as follows:

$$\begin{aligned}&\text{Compression Ratio}\\&= \frac{\text{total size in bits of the original image or video}}{\text{total size in bits of the compressed bitstream}}\\&= \frac{\text{bit rate of the original image or video}}{\text{bit rate of the compressed bitstream}}\end{aligned} \tag{2}$$

Ultimately, the quality of a compressed image or video bit stream should be judged subjectively by human eyes. However, it is normally very costly and time-consuming to perform a formal subjective test. Although subjective quality evaluation is still a necessary step for formal tests in various image and video coding standardization processes, many objective measurements can also be used as rough quality indicators. For example, it can be measured by the distortion of the decoded image with reference to the original image under different criteria, and a particular criterion among them is the mean squared error defined as follows:

$$D = \frac{\sum_{i,j}[(r(i,j) - o(i,j)]^2}{N} \tag{3}$$

where $r(i,j)$ and $o(i,j)$ are the reconstructed and original image intensities of pixel position (i, j), respectively, and N is the number of pixels in the image, and summation is carried out for all pixels in an image. Another commonly used measurement is PSNR (peak signal to noise ratio) defined as follows:

$$PSNR = 10 \log_{10} \frac{P^2}{D} \tag{4}$$

where D is the mean squared error calculated in Eq. (3) and P is the maximum possible pixel value of the original image; for example, if the intensities of the original images are represented by 8-bit integers, then the peak value would be $P = 2^8 - 1 = 255$.

The performance or coding efficiency of an image or video coding scheme is best illustrated in a rate-distortion or bit rate versus PSNR curve, where the encoder would encode the same image or video signal at a few bit rates and the decoded quality is measured using the above-defined criteria. This makes it a very intuitive tool for evaluating various coding schemes.

3. Basic Principles and Techniques for Image and Video Coding

Image and video coding techniques can be classified into two categories: lossless compression and lossy compression. In lossless compression, video data can be identically recovered (decompressed) both quantitatively (numerically) and qualitatively (visually) from a compressed bit stream. Lossless compression tries to represent the video data with the smallest possible number of bits without loss of *any* information. Lossless compression works by removing the *redundancy* present in video data information that, if removed, can be recreated from the remaining data. Although lossless compression preserves exactly the accuracy of image or video representation, it typically offers a relatively small compression ratio, normally a factor of 2 or 3. Moreover, the compression ratio is very dependent on the input data, and there is no guarantee that a given output bit rate can always be achieved.

By allowing a certain amount of distortion or information loss, a much higher compression ratio can be achieved. In lossy compression, once compressed, the original data cannot be identically reconstructed. The reconstructed data are similar to the original but not identical. Lossy compression attempts to achieve the best possible fidelity given an available bit-rate capacity, or to minimize the number of bits representing the image or video signal subject to some allowable loss of information. Lossy compression may take advantage of the human visual system that is insensitive to certain distortion in image and video data and enables rate control in the compressed data stream. As a special case of lossy compression, perceptually lossless or near lossless coding methods attempt to remove redundant, as well as perceptually irrelevant, information so that the original and the decoded images may be visually but not numerically identical. Unfortunately, the measure of perceptive quality of images or video is a rather complex one, especially for video. Many efforts have been devoted to derive an objective measurement by modeling the human visual system, and an effective one is yet to be found. Many practical coding systems still need close supervision by so-called compressionists.

Lossy compression provides a significant reduction in bit rate that enables a multitude of real-time applications involving processing, storing, and transmission of audio-visual information, such as digital camera, multimedia web, video conferencing, digital TV, and so on. Most image and video coding standards, such as JPEG, JPEG2000, MPEG-1, MPGE-2, MPEG-4, H.26x, and the like, are all examples of lossy compression.

Although a higher compression ratio can be achieved with lossy compression, there exist several applications that require lossless coding, such as digital medical imagery and facsimile. A couple of lossless coding standards have been developed for lossless compression, such as lossless JPEG, JPEG-LS, ITU Tele-Fax standards, and the JBIG. Furthermore, lossless compression components generally can be used in lossy compression to further reduce the redundancy of the signal being compressed.

3.1. A Generic Model for Image and Video Coding

To help better understand the basic ideas behind different image and video coding schemes. Figure 4 illustrates the components and their relationship in a typical image and video coding system. Depending on the target applications, not all these components may appear in every coding scheme. In Figure 4, the encoder takes as input an image or video sequence and generates as output a compressed bit stream. The decoder, conversely, takes as input the compressed bit stream and reconstructs an image or video sequence that may or may not be the same as the original input. The components in Fig. 4 are explained in this subsection.

An image or video coding scheme normally starts with certain preprocessing of the input image or video signal, such as denoising, color space conversion, or spatial resolution conversion to better serve the target applications or to improve the coding efficiency for the subsequent compression stages. Normally, preprocessing is a lossy process.

The preprocessed visual signal is passed to a reversible (one-to-one) transformation stage, where the visual data are transformed into a form that can be more efficiently compressed. After the transformation, the correlation (interdependency, redundancy) among transformed coefficients is reduced, the statistical distribution of the coefficients can be shaped to make subsequent entropy

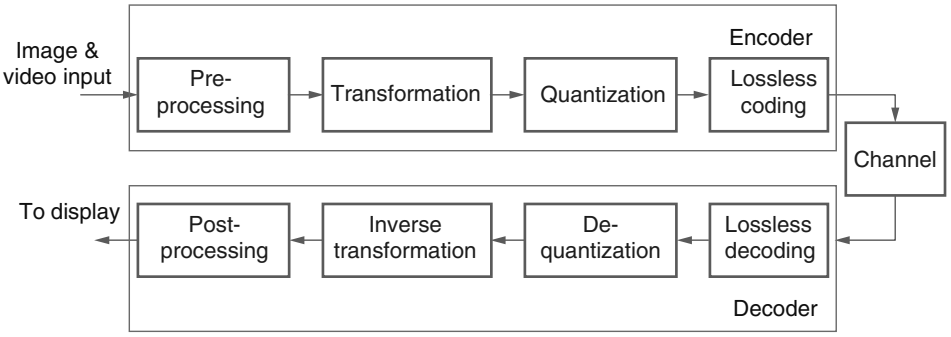


Figure 4. A generic image and video codec model.

coding more efficient, and/or most energy is packed into only a few coefficients or subband regions so that the majority of the transformed coefficients are zeros. Depending on the transform type and the arithmetic precision, even if this step is not lossless, it is close to it. Typical transformations include *differential* or *predictive* mapping (spatially or temporally); unitary transforms such as the discrete cosine transform (DCT); subband decomposition such as wavelet transform; and adaptive transforms such as adaptive DCT, wavelet packets, fractals, and so on. Modern compression systems normally use a combination of these techniques and allow different modes in the transformation stage to decompose the image or video signal adaptively. For example, there are intramodes and intermodes in MPEG video coding standards.

The transformed signal is fed into a quantization module where the coefficients are mapped into a smaller set of discrete values, thus requiring fewer bits to represent these coefficients. Because quantization causes loss of information, it is a lossy process. Quantization is actually where a significant amount of data reduction can be attained. Rate control in lossy compression can be easily achieved by adjusting the quantization step size. There are two main types of quantization: scalar quantization (SQ), where each coefficient is quantized independently, and vector quantization (VQ), where several coefficients are grouped together and quantized jointly.

In some cases, especially in predictive coding, the differential or predictive operation and the quantization may work iteratively in a feedback loop to prevent error propagation, because otherwise the original reference could not be reconstructed perfectly in the decoder after quantization.

Although not explicitly shown in Fig. 4, in practical lossy compression systems, there is always an optimization process that is closely related to rate control. The goal of this process is to minimize the bit rate given a certain quality constraint, or to maximize the decoded quality given a certain bit rate budget, by adjust encoding parameters such as transformation or prediction modes, quantization parameters, and bit allocation among different portions of the image or video signal. As a matter of fact, many compression standards specify only the decoding process and leave the encoding process, especially the optimization part, open to the implementers as long as they produce a bit stream that can be decoded by a compliant decoder. This allows the encoding algorithms to be improved over time, and yet compliant decoders will continue to be able to decode them. The compression process rarely just stops after quantization. The redundancy of the quantized coefficients can be further removed by some generic or specific lossless data compression schemes. Lossless data compression normally involves two parts, symbol formation and entropy coding.

The symbol formation part tries to convert data or coefficients into symbols that can be more efficiently encoded by entropy coding. Such a mapping can be achieved through, for example, coefficients partitioning, or run length coding (RLC). Image or video coefficients can be partitioned and grouped into data blocks based on their potential correlation. Such data blocks can be either mapped into a single symbol or formed as a context to address different probability distribution tables for one or more of the coefficients to be entropy coded. In either case, the correlation between the coefficients can be exploited to result in higher compression ratios by the entropy coder that follows. The entropy coding part tries to generate the shortest binary bit stream by assigning binary codes to the symbols according to their frequency of occurrence. Entropy coding can usually be achieved by either statistical schemes or dictionary-based schemes. In a statistical scheme, fixed-length symbols are coded using variable-length codewords (VLC), where the shorter codewords are assigned to the symbols that occur more frequently (e.g., Huffman coders and arithmetic coders). Alternatively, in a dictionary-based scheme, variable-length strings of symbols are coded using fixed-length binary codewords, where *a priori* knowledge is not required (e.g., Lempel Ziv coder).

After the lossless data compression stage, the image or video data are encoded into a bit stream that can be either directly sent to, or wrapped with some system layer information such as synchronization information or encryption, and then sent to a channel that can be either a storage device, a dedicated link, an IP network, or a processing unit. The decoder receives the bit stream from the channel with or without possible corruptions and starts to decode the received bit stream. The decoding is just the reverse process of the encoding except for the postprocessing part.

Applications of image and video coding may be classified into symmetric or asymmetric. For example, video conferencing applications are symmetric because both ends of the communication have the same requirements. In video streaming applications, where the same video content can be preencoded and stored in a server and accessed by many users, encoding may be much more complicated than decoding, because encoding is only required to be performed once but decoding is required by many users for many times. With more allowable encoding complexity, asymmetric algorithms usually lead to much better coding efficiency than the symmetric ones.

For predictive video coding, since previous decoded frames are normally needed as prediction references, the encoder would include the major parts of a decoder in it. In this sense, modules designed for decoder are also parts of the encoder, and to prevent drifting errors, the decoder must also follow the same procedures as in the decoding loop of the encoder.

There are many reasons for postprocessing. For example, if the decoded video format does not match the display, a scaling of spatial resolution or temporal frame rate, or a de-interlacing operation will be used for the conversion. If the bit-rate is so low that there are many artifacts in the decoded image or video, some de-blocking or de-ringing operations will be used to reduce these artifacts. If there are channel errors or packet losses, there would be some distortions in the decoded image or video, then an error concealment operation is required to repair the image or video to the best possible extent. If there are multiple image or video objects that need to be presented on the same display, a composition operation

will be used to generate a meaningful scene to present to the user. Moreover, it has been shown that in-loop filtering of the decoded frames can improve the efficiency of predictive video coding significantly at low bit rates. Although the in-loop filtering results may or may not be output directly to the display, we regard such filtering as a kind of postprocessing as well. Postprocessing is a stage that prepares the decoded image or video data into a more favorable form for improving visual quality or subsequent coding efficiency.

As mentioned before, not all stages would appear in every image or video coder. For example, a lossless coder normally does not contain lossy stages, such as, preprocessing and quantization.

So far, we only provided a generic overview of what components are involved in an image and video compression system. In the following subsections, we present common principles and basic techniques of image and video coding in some details.

Although color components in images or video can be coded jointly as a vector based signal to achieve improved compression efficiency, practical image and video coding systems normally choose to compress them independently due to its simplicity. From now on in this article, unless especially noted, the algorithms and schemes described are for a single color component.

3.2. Entropy Coding

Entropy is a measure of disorder, or uncertainty. In information systems, the degree of unpredictability of a message can be used as a measure of the information carried by the message. In 1948, Shannon defined the information conveyed by an event $I(E)$, measured in bits, in terms of the probability of the event $P(E)$,

$$I(E) = -\log_2(P(E)). \tag{5}$$

The physical meaning of the above definition is not hard to understand. The higher the probability of an event (i.e., the more predictable event), the less information is conveyed by that event when it happens. Moreover, information conveyed by a particular sequence of independent events is the sum of the information conveyed by each event in the sequence, whereas the probability of the sequence is the product of the individual probabilities of the events in the sequence, which is exactly what the log function reflects. A discrete memoryless source (DMS) generates symbols from a known set of alphabet symbols one at a time. It is memoryless since the probability of any symbol being generated is independent of the past history. Assume a DMS U_0 with alphabet $\{a_0, a_1, \ldots, a_{K-1}\}$ and probabilities $\{P(a_0), P(a_1), \ldots, P(a_{K-1})\}$, the entropy of such an information source is defined as the average amount of information conveyed by each symbol output by the source,

$$H(U_0) = -\sum_{k=0}^{K-1} P(a_i)\log_2(P(a_i)). \tag{6}$$

Generally, a source with nonuniform distribution can be represented or compressed using a variable-length code where shorter code words are assigned to frequently occurring symbols, and vice versa. According to Shannon's noiseless source encoding theorem, the entropy $H(U_0)$ is the lower bound for the average word length of a uniquely decodable variable-length code for the symbols. Conversely, the average word length can approach $H(U_0)$ if sufficiently large blocks of symbols are encoded jointly.

3.2.1. Huffman Coding. One particular set of uniquely decodable codes are called prefix codes. In such a code, one code word cannot be the prefix of another one. In 1952 Huffman proposed an algorithm for constructing optimal variable-length prefix codes with minimum redundancy for memoryless sources. This method remains the most commonly used today, for example, in the JPEG and MPEG compression standards.

The construction of a Huffman code is as follows:

1. Pick the two symbols in the alphabet with lowest probabilities and merge them into a new combined symbol. This generates a new alphabet with one less symbol. Assign "0" and "1" to the two branches linking the two original symbols to the new combined one, respectively.
2. Calculate the probability of the combined symbol by adding up the probabilities of the two original symbols.
3. If the new alphabet contains more than one symbol, repeat steps 1 and 2 for the new alphabet. Otherwise, the last combined symbol becomes the Huffman tree root.
4. For each original symbol in the alphabet, traverse all the branches from the root and append the assigned "0" or "1" for each branch along the way to generate a Huffman codeword for the original symbol.

We now have a Huffman code for each member of the alphabet. Huffman codes are prefix codes and each of them is uniquely decodable. A Huffman code is not unique. For each symbol set, there exist several possible Huffman codes with equal efficiency. It can be shown that it is not possible to generate a code that is both uniquely decodable and more efficient than a Huffman code [16]. However, if the probability distribution somehow changes, such preconstructed codes would be less efficient and sometimes would even bring expansion. Moreover, Huffman codes are most efficient for data sources with nonuniform probability distributions. Sometimes, we may have to manipulate the data so as to achieve such a distribution.

In many cases, the data source contains a large alphabet but with only a few frequent symbols. Huffman codes constructed for such a source would require a very large code table and it would be very difficult to adapt to any probability distribution variations. An alternative method used in many image and video coding systems is to group all infrequent symbols as one composite symbol and construct a Huffman table for the reduced alphabet. The composite symbol is assigned a special escape code used to signal that it is followed by a fixed-length index of one of the infrequent symbols in the composite group. Only a very small code table is used and the statistics of the vast

majority of infrequent symbols in the alphabet is shielded by the composite symbol. This greatly simplifies the Huffman code while maintaining good coding efficiency.

3.2.2. Arithmetic Coding. Huffman codes and derivatives can provide efficient coding for many sources. However, the Huffman coding schemes cannot optimally adapt to given symbol probabilities because they encode each input symbol separately with an integer number of bits. It is optimal only for a "quantized" version of the original probability distribution so the average code length is always close to but seldom reaches the entropy of the source. Moreover, no code is shorter than 1 bit, so it is not efficient for an alphabet with highly skewed probability distribution. Furthermore, there are no easy methods to make Huffman coding adapt to changing statistics.

On the other hand, an arithmetic encoder computes a code representing the entire sequence of symbols (called a string) rather than encodes each symbol separately. Coding is performed by representing the string by a subinterval through a sequence of divisions of an initial interval according to the probability of each symbol to be encoded.

Assume that a string of N symbols, $S = \{s_0, s_1, \ldots, s_t, \ldots, s_{N-1}\}$ are from an alphabet with K symbols $\{a_0, a_1, \ldots, a_i, \ldots, a_{K-1}\}$ with a probability distribution that is varied with time t as $P(a_i, t)$. Then the arithmetic encoding process of such a string can be described as follows:

1. Set the initial interval $[b, e)$ to the unit interval [0,1) and $t = 0$.
2. Divide the interval $[b, e)$ into K subintervals proportional to the probability distribution $P(a_i, t)$ for each symbol a_i at time t, that is,

$$b_i = b + (e - b) \sum_{j=0}^{i-1} P(a_j, t) \quad \text{and}$$

$$e_i = b + (e - b) \sum_{j=0}^{i} P(a_j, t).$$

3. Pick up the subinterval corresponding to symbol s_t, say $s_t = a_i$, update $[b, e)$ with $[b_i, e_i)$ and $t = t + 1$.
4. Repeat step 2 and 3 until $t = N$. Then output a binary arithmetic code that can uniquely identify the final interval selected.

Disregarding the numerical precision issue, the width of the final subinterval is equal to the probability of the string $P(S) = \prod_{t=0}^{N-1} P(s_t, t)$. It can be shown that the final subinterval of width $P(S)$ is guaranteed to contain one number that can be represented by B binary digits, with

$$-\log_2(P(S)) + 1 \le B < -\log_2(P(S)) + 2, \tag{7}$$

which means that the subinterval can be represented by a number which needs 1 to 2 bits more than the ideal code word length. Any number within that interval can now be used as the code for the string (usually the one with the smallest number of digits is chosen). In the encoding process, there is no assumption that the probability distribution would stay the same at all time. Therefore, by nature the arithmetic encoding can well adapt to the changing statistics of the input and this is a significant advantage over Huffman coding. From an information theory point of view, arithmetic coding is better than Huffman coding; it can generate fractional bits for a symbol, and the total length of the encoded data stream is minimal. There are also many implementation issues associated with arithmetic coding, for example, limited numerical precision, a marker for the end of a string, multiplication free algorithm, and so on. For a detailed discussion of arithmetic coding, please see [17]. In practice, arithmetic and Huffman coding often offer similar average compression rates while arithmetic coding is a little better (ranging from 0 to 10% less bits). Arithmetic coding is an option for many image and video coding standards, such as JPEG, MPEG, H.26x, and so forth.

3.2.3. Lempel–Ziv Coding. Huffman coding and arithmetic coding require *a priori* knowledge of the probabilities or an accurate statistical model of the source which in some cases is difficult to obtain, especially with mixed data types. Conversely, Lempel–Ziv (LZ) coding developed by Ziv and Lempel [18] does not need an explicit model of the source statistics. It is a dictionary-based universal coding that can dynamically adapt to any sources.

In LZ coding, the code table (dictionary) of variable-length symbol strings is constructed dynamically. Fixed-length binary codewords are assigned to the variable-length input symbol strings by indexing into the code table. The basic idea is always to encode a symbol string that the encoder has encountered before as a whole. The longest symbol string the encoder has not seen so far is added as a new entry in the dictionary, and will in turn be used to encode all future occurrences of the same string. At any time, the dictionary contains all the substrings (prefixes) the encoder has already seen. With the initial code table and the indices received, the decoder can also dynamically reconstruct the same dictionary without any overhead information.

A popular implementation of LZ coding is the Lempel–Ziv–Welch (LZW) algorithm developed by Welch [20]. Let A be the source alphabet consisting K symbols $\{a_k, k = 0, \ldots, K - 1\}$. The LZW algorithm can be described as follows,

1. Initialize the first K entries of the dictionary with each symbol a_k from A and set the scan string w to empty.
2. Input the next symbol S and concatenate it with w to form a new string wS.
3. If wS has a matching entry in the dictionary, update the scan string w with wS, and go to 2. Otherwise, add wS as a new entry in the dictionary, output the index of the entry matching w update the scan string w with S and go to 2.

4. When the end of the input sequence is reached, process the scan string w from left to right, output the indices of entries in the dictionary that match with the longest possible substrings of w.

If the maximum dictionary size is M entries, the length of the codewords would be $\log_2(M)$ rounded to the next smallest integer. The larger the dictionary is, the better the compression. In practice the size of the dictionary is a trade-off between speed and compression ratio. It can be shown that LZ coding asymptotically approaches the source entropy rate for very long sequences [19]. For short sequences, however, LZ codes are not very efficient because of their adaptive nature. LZ coding is used in the UNIX compress utility and in many other modern file compression programs.

3.3. Markov Sources

Though some sources are indeed memoryless, many others, for example image and video data where the probability distribution of values for one symbol can be very dependent on one or more previous values, are sources with memory. A source with memory can be modeled as a Markov source. If a symbol from a source is dependent on N previous value(s), the source is known as an Nth-order Markov source. Natural or computer rendered images and video data are examples of Markov sources.

Conversely, joint sources generate N symbols simultaneously. A coding gain can be achieved by encoding those symbols jointly. The lower bound for the average code word length is the joint entropy,

$$H(U_1, U_2, \ldots, U_N) = -\sum_{u_1}\sum_{u_2}\cdots\sum_{u_N} P(u_1, u_2, \ldots, u_N) \times \log_2(P(u_1, u_2, \ldots, u_N)). \quad (8)$$

It generally holds that

$$H(U_1, U_2, \ldots, U_N) \le H(U_1) + H(U_2) + \cdots + H(U_N) \quad (9)$$

with equality, if $U_1, U_2, \ldots, U_N$ are statistically independent. This states that for sources with memory, they can be best coded jointly with a smaller lower bound (the joint entropy) for the average word length than otherwise coded independently. Moreover, the word length of jointly coding the memoryless sources has the same lower bound as independently coding them. However, coding each memoryless source independently is much easier to implement than coding jointly in real applications.

For an image frame or a video sequence, since each pixel in it is correlated with neighboring pixels, to obtain the best coding efficiency, it is ideal to encode all the pixels in the whole image frame or video sequence jointly. However, practical implementation complexity prohibits us to do so. Fortunately, with the Markov model for image and video data, the problem can be much simplified.

For an Nth-order Markov source, assume the first symbol starts at time T_0 the conditional probabilities of the source symbols are,

$$P(u_T, Z_T) = P(u_T \mid u_{T-1}, u_{T-2}, \ldots, u_{T-N}) = P(u_T \mid u_{T-1}, u_{T-2}, \ldots, u_{T-N}, u_{T-N-1}, \ldots u_{T_0}) \quad (10)$$

where Z_T represents the state of the Markov source at time T The conditional entropy of such an Nth-order of Markov source is given by,

$$\begin{aligned} H(U_T, Z_T) &= H(U_T \mid U_{T-1}, U_{T-2}, \ldots, U_{T-N}) \\ &= E(-\log_2(P(u_T \mid u_{T-1}, u_{T-2}, \ldots u_{T-N}))) \\ &= -\sum_{u_T}\cdots\sum_{u_{T-N}} P(u_T, u_{T-1}, u_{T-2}, \ldots u_{T-N}) \\ &\quad \times \log_2(P(u_T \mid u_{T-1}, u_{T-2}, \ldots u_{T-N})) \\ &= \sum_{u_{T-1}}\cdots\sum_{u_{T-N}} P(u_{T-1}, u_{T-2}, \ldots u_{T-N}) \\ &\quad \times H(U_T \mid u_{T-1}, u_{T-2}, \ldots u_{T-N}). \end{aligned} \quad (11)$$

Moreover, it can be shown that,

$$H(U_T, U_{T-1}, \ldots, U_{T_0}) = \sum_{t=T_0}^{T} H(U_t, Z_t). \quad (12)$$

From the above equation, we can clearly see that for a Markov source, the complicated joint entropy of a whole symbol sequence can be simplified to the sum of the conditional entropy of each symbol in the sequence, which means that the entropy coding of such joint Markov sources can be simplified to the conditional entropy coding of each symbol in the sequence given N previous symbols. In addition, the last equation in (11) suggests a simple conditional entropy coding method: any entropy coding method for a memoryless source can be used to encode the symbol at time T as long as the probability distribution used is switched according to the contexts (or *states*) of N previous symbols.

Image and video data are normally highly correlated. The value of a pixel has dependency with a few neighboring pixels. Even for the simplest first-order Markov separable model, a pixel in a 2-D image would have dependency with at least 3 neighboring pixels and a pixel in a 3-D video sequence would have dependency with at least 7 neighboring pixels. If each pixel is represented by 8 bits, in order to most efficiently compress the pixel value with the above derived simplified context-based entropy coding method, it would require 256^3 probability tables for image coding and 256^7 probability tables for video coding. Apparently, it is impractical for the encoder or decoder to maintain such a huge number of probability distribution tables. For an efficient, independent coding of symbols, statistical dependencies should be reduced.

3.4. Predictive Coding

Predictive coding is a way to reduce correlation between data from Markov sources. It is much simpler than a conditional entropy coder described in the previous section. Instead of coding an original symbol value, the difference or error between the original value and a predicted value based on values of one or more past symbols is encoded. The decoder will perform the same prediction and use

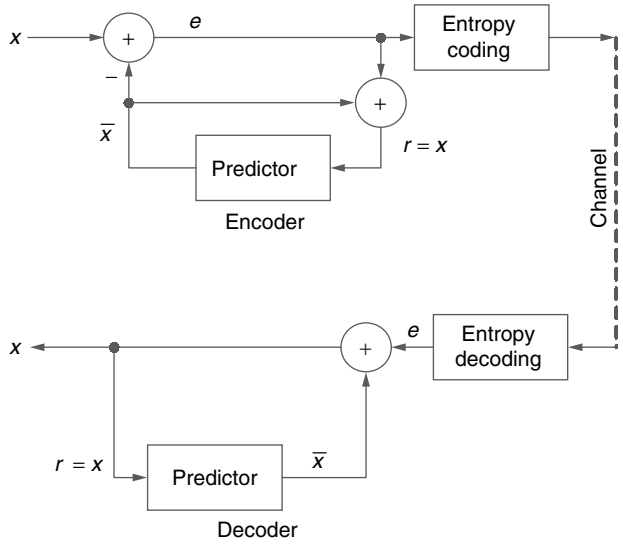


Figure 5. Diagram of a predictive coder.

the encoded error to reconstruct the original value of that symbol. Figure 5 illustrates the predictive coding process.

The linear predictor is the most used one in predictive coding. It creates a linear weighting of the last N symbols (Nth-order Markov) to predict the next symbol. That is,

$$\hat{S}_0 = \alpha_{-1}S_{-1} + \alpha_{-2}S_{-2} + \cdots + \alpha_{-N}S_{-N}, \tag{13}$$

where $\{S_{-1}, S_{-2}, \ldots, S_{-N}\}$ are the last N symbols and α_{-i} are the weights in the linear predictor. Without loss of generality, assume the input symbols have zero mean, i.e., $E\{S\} = 0$. The variance of the prediction error $e_0 = S_0 - \hat{S}_0$ is given by,

$$\sigma_{e_0}^2 = \sum_{i=0}^{N}\sum_{j=0}^{N} \alpha_{-i}\alpha_{-j}R_{ij}, \tag{14}$$

where R_{ij} is the covariance of symbols S_{-i} and S_{-j}, and $\alpha_0 = -1$.

Minimization of the prediction error variance leads to the orthogonality principle,

$$E\{e_0 S_{-i}\} = 0, \text{ for all } i = 1, 2, \ldots, N. \tag{15}$$

The optimum coefficients $\{\alpha_{-1}, \alpha_{-2}, \ldots, \alpha_{-N}\}$ can be obtained by solving the above equation. Moreover, the orthogonality principle implies de-correlation of errors,

$$E\{e_0 e_{-i}\} = E\{e_0\}E\{e_{-i}\} = 0, \text{ for all } i = 1, 2, \ldots, N. \tag{16}$$

For Gaussian random processes, de-correlation means statistical independence. Clearly, after optimum linear prediction, the prediction errors are uncorrelated and the much simpler and independent entropy coding can be used to efficiently encode these errors. Intuitively, after the process we have transformed a source of highly-correlated values with any possible distribution into a source of much less correlated values but with consistently high probability of being small. Such transformations that make the source more suitable for subsequent compression are very common in an image or video compression system.

For nonstationary sources, adaptive prediction can be used for more accurate prediction where the system switches between several predefined predictors according to the characteristics of the data being compressed. The choices of the predictors can be either explicitly coded with a small overhead or implicitly derived from the reconstructed values to avoid the overhead.

Predictive coding is important in a number of ways. In its lossless form, predictive coding is used in many sophisticated compression schemes to compress critical data, such as DC coefficients and motion vectors. If some degree of loss is acceptable, the technique can be used more extensively, (e.g., interframe motion compensated prediction in video coding).

3.5. Signal Models for Images and Video

In general, there is no good model to describe exactly the nature of real world images and video. A first-order Gaussian-Markov source is often used as a first order approximation due to its tractability. Though crude, this model provides many insights in understanding image and video coding principles.

A one-dimensional (1-D) Gaussian-Markov (AR(1) source) can be defined as,

$$x(n) = \alpha x(n-1) + \varepsilon(n), \text{ for all } n > n_0, \tag{17}$$

where $|\alpha| < 1$ is the regression coefficient, and $\{\varepsilon(n)\}$ is the i.i.d (independent identically distributed) zero mean normal random process with variance σ_ε^2, and $x(n_0)$ is a zero mean finite variance random variable. Such a source is known to be asymptotically stationary [1]. For a two-dimensional (2-D) image signal, the simplest source model is a two-dimensional separable correlation AR(1) model,

$$\begin{aligned} x(m,n) = {} & \alpha_h x(m-1,n) + \alpha_v(m,n-1) \\ & - \alpha_h\alpha_v x(m-1,n-1) + \varepsilon(m,n), \end{aligned} \tag{18}$$

where $\varepsilon(m,n)$ is an i.i.d zero mean Gaussian noise source with variance σ_N, and α_h, α_v denote the first order horizontal and vertical correlation coefficients, respectively. Its autocorrelation function is separable and can be expressed as a product of two 1-D autocorrelations. The separable correlation model of a 2-D image enables us to use 2-D separable transforms or other signal processing techniques to process the 2-D sources using separate 1-D processing in both horizontal and vertical directions, respectively. Therefore, in most cases, 1-D results can be generalized to 2-D cases in accordance with the 2-D separable correlation model.

For video signals, a 3-D separable signal model could still apply. However, the special feature of moving pictures in natural video is the relative motion of video objects in adjacent frames. A more precise signal model should incorporate the motion information into the 3-D signal model, for example, forming signal threads along the temporal direction and then applying the 3-D separable signal model. Fortunately, most modern video compression

technologies have already explicitly used such motion information in the encoding process to take advantages of the temporal redundancy.

3.6. Quantization

The coding theory and techniques we discussed so far are mostly focused on lossless coding. However, the state-of-the-art lossless image and coding schemes exploiting the statistical redundancy of image and video data can only achieve an average compression factor of 2 or 3. From Shannon's rate distortion theory, we know that the coding rate of a data source can be significantly reduced by introducing some numerical distortion [4]. Fortunately, because the human visual system can tolerate some distortion under certain circumstances, such a distortion may or may not be perceptible. Lossy compression algorithms reduce both the redundant and irrelevant information to achieve higher compression ratio. Quantization is usually the only lossy operation that removes perceptual irrelevancy. It is a many-to-one mapping that reduces the number of possible signal values at the cost of introducing some numerical errors in the reconstructed signal. Quantization can be performed either on individual values (called scalar quantization) or on a group of values (a coding block, called vector quantization). The rate-distortion theory also indicates that, as the size of the coding block increases, the distortion asymptotically approaches Shannon lower bound; in other words, if a source is coded as an infinitely large block, then it is possible to find a block-coding scheme with rate $R(D)$ that can achieve distortion D where $R(D)$ is the minimum possible rate necessary to achieve an average distortion D [4]. This states that vector quantization is always better than scalar quantization. However, due to the complexity issue, many practical image and video coders still prefer to use scalar quantization. The basics on scalar and vector quantization techniques are discussed as follows.

3.6.1. Scalar Quantization. An N-point scalar quantization is a mapping from a real one-dimensional space to a finite set C of discrete points in the real space, $C = \{c_0, c_1, \ldots, c_{N-1}\}$. Normally, the mapping is to find the closest match in C for the input signal according to a certain distortion criterion, for example, mean squared error (MSE). The values of c_i are referred to as reproduction values. The output is the index of the best matched reproduction value. The transmission rate $r = \log_2 N$ is defined to indicate the number of bits per sample. The uniform quantizer with c_i distributed uniformly on the real axis is the optimal solution when quantizing a uniformly distributed source. For a random nonuniformly distributed source (such as image luminance levels) and even for a source with an unknown distribution, the Lloyd–Max quantizer design algorithm [2] provides an essential approach to the (locally) optimal scalar quantizer design. As a special case of the generalized Lloyd algorithm to be discussed for vector quantization, the Lloyd-Max quantizer tries to minimize the distortion for a given number of levels without the need for a subsequent entropy coder. Though optimal, it involves a complicated iterative training process. On the other hand, the study of entropy-constrained quantization (ECQ) shows that, if an efficient entropy coder is applied after quantization, the optimal coding gain can be always achieved by a uniform quantizer [3]. In other words, for most applications in image and video compression, the simplest possible quantizer, followed by variable-length coding, produces the best results. In order to take advantage of the subsequent entropy coder, many practical scalar quantizers have included a dead-zone, where a relatively larger partition is allocated for zero. Note that if the probability distribution of the data to be quantized is highly skewed, the inverse quantizer might achieve smaller distortion if choosing a *biased* reconstruction point towards the higher probability end rather than the usual mid-point for uniform distribution.

As shown above, the uniform quantizer combined with entropy coding is simple and well suited in image and video coding. However, new applications such as delivery of image or video over unstable or low bandwidth networks require progressive transmission and/or exact rate control of the bit stream. To equip the image or video coding with these new functionalities, progressive quantization strategies have to be adopted where a coefficient is quantized in a multi-pass fashion and the quantization error can be reduced by successive enhancement information obtained from each pass. Successive-Approximation Quantization (SAQ) of the coefficients is one such approach and is widely used in image and video coding systems [37,38,44,47]. As a special case of SAQ, bit-plane coding encodes each bit in a binary representation of the coefficient from the most significant bit (MSB) to the least significant bit (LSB) at each quantization scan.

3.6.2. Vector Quantization. In contrast to a scalar quantizer that operates upon a single, one-dimensional, variable, a vector quantizer acts on a multidimensional vector. Vector quantization is a mapping of k-dimensional Euclidean space R^k into a finite subset $\mathbf{C}$ of R^k, where $\mathbf{C} = \{\mathbf{c}_i : i = 1, 2, \ldots, N\}$ is the set of N reproduction vectors or codebook, and the element $\mathbf{c}_i$ is referred to as a codeword or a code-vector. An encoder $Q(\mathbf{x})$ takes an input vector $\mathbf{x}$ and generates the index i of the best matched vector $\mathbf{c}_i$ to $\mathbf{x}$ in the set $\mathbf{C}$ according to certain distortion criteria, for example, MSE. A decoder $Q^{-1}(i)$ regenerates the codeword $\mathbf{c}_i$ using the input index i Fig. 6 illustrates the VQ encoding and decoding processes.

Vector quantization always outperforms the scalar quantization in terms of error measurement under the same bit rate by Shannon's rate-distortion theory [3–6]. Vector quantization takes advantage of the joint probability distribution of a set of random variables while scalar quantization only uses the marginal probability distribution of a one-dimensional random variable [3]. In [7], Lookabaugh and Gray summarized the vector quantization advantages over scalar quantization in three categories: memory advantage (correlation between vector components, Fig. 7), shape advantage (probability distribution, Fig. 8) and space filling advantage (higher dimensionality, Fig. 9). Table 1 lists the high-rate approximation of coding gain brought by these VQ advantages over a scalar quantizer for different VQ dimensionalities. The results

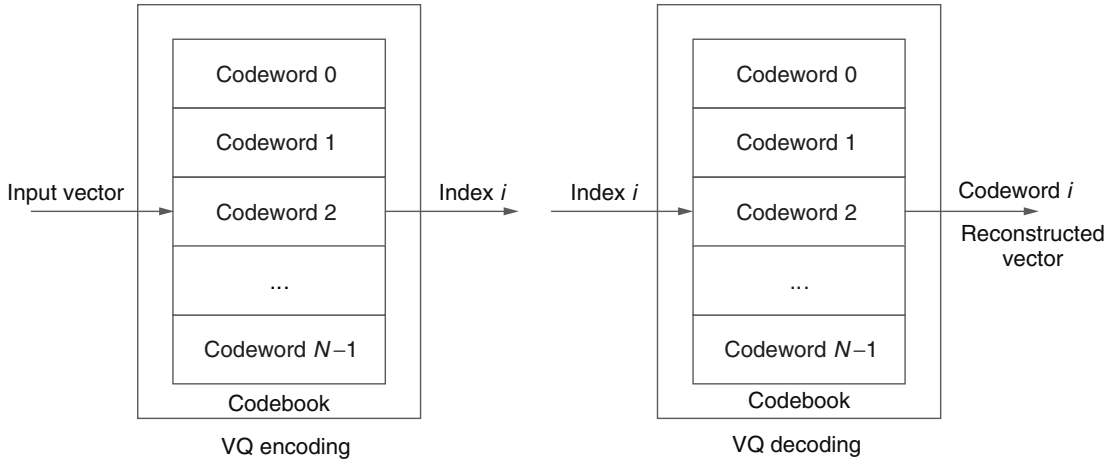


Figure 6. Vector quantization encoding and decoding processes.

are based on a first-order Gaussian-Markov source with regression coefficient $\rho = 0.95$, which is a typical value for a natural image or video source. Table 1 provides us some very useful insights on vector quantization. Firstly, as the VQ dimensionality increases, the coding gain also increases. The higher the VQ dimensionality is, the better the VQ performance is. Secondly, the increase in coding gain slows down beyond a certain VQ dimensionality. There is a delicate tradeoff between the increased complexity and the extra coding gain. The practical rule of thumb in image and video coding is that the coding gain of vector dimensionality beyond 16 may not be worth the added complexity. Thirdly, most of the VQ coding gain comes from the memory advantage. If we can completely decorrelate the samples within a VQ vector, and apply a scalar quantizer, we can still achieve most of the coding gain.

A good codebook design is crucial for the performance of a vector quantizer. The ideal codebook should minimize the average distortion for a given number of codewords. The most commonly used algorithms for codebook generation is the LBG algorithm (Linde, Buzo, and Gray) [11], which is also referred to as Generalized Lloyd Algorithm (GLA). The LBG algorithm is an iterative process based on the two necessary conditions for an optimal codebook: the nearest neighbor condition, where the optimal partition cell for a codeword should cover any vector that has the shortest distance to it; the centroid condition, where the optimal codeword should be the centroid of the partition cell.

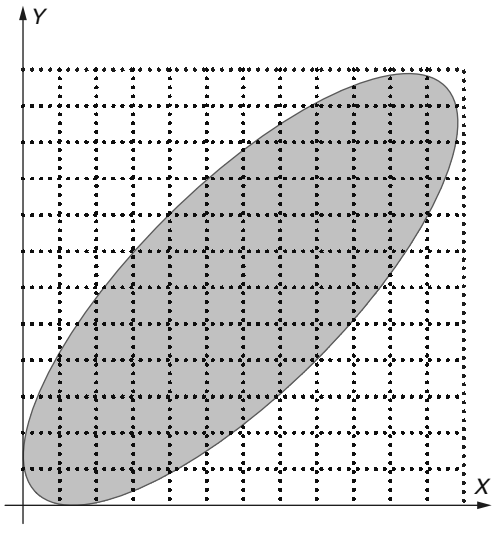


Figure 7. Memory advantage of vector quantization.

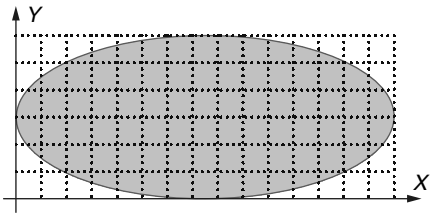


Figure 8. Shape advantage of vector quantization.

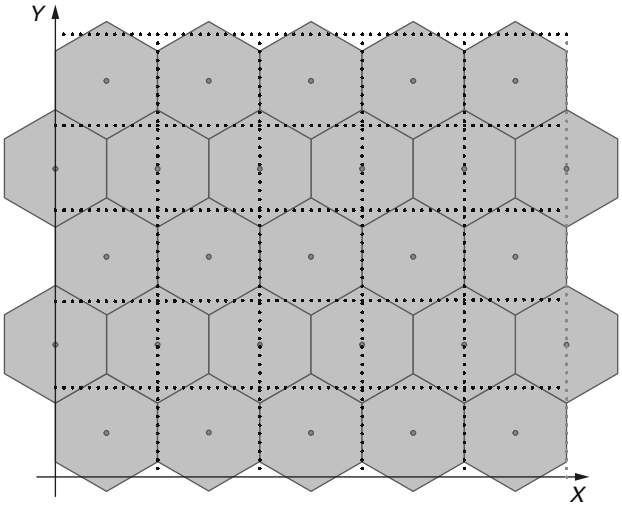


Figure 9. Space-filling advantage of vector quantization.

Table 1. The Coding Gain (dB) of Vector Quantization with Different Dimensionalities Over Scalar Quantization

VQ Dimension	Space Filling Advantage (dB)	Shape Advantage (dB)	Memory Advantage (dB)	Total Coding Gain (dB)
1	0	0	0	0
2	0.17	1.14	5.05	6.36
3	0.29	1.61	6.74	8.64
4	0.39	1.87	7.58	9.84
5	0.47	2.04	8.09	10.6
6	0.54	2.16	8.42	11.12
7	0.60	2.25	8.67	11.52
8	0.66	2.31	8.85	11.82
9	0.70	2.36	8.99	12.05
10	0.74	2.41	9.10	12.25
12	0.81	2.47	9.27	12.55
16	0.91	2.55	9.48	12.94
24	1.04	2.64	9.67	13.35
100	1.35	2.77	10.01	14.13
∞	1.53	2.81	10.11	14.45

Although vector quantizers offer unparallel quantization performance, they can be very complex in both codebook design and encoding (searching for the best match). Normally, they would be applied in very low bit rate coding case where only a small codebook size is required. There are many continuing investigations on how to reduce the complexity of codebook training and codebook searching [3,12–15,43]. Some efforts have also been put on the analogy of a uniform scalar quantizer in multidimensions — lattice VQ (LVQ) [8–10,45], where a codebook is not necessary. However, it seems that LVQ just puts off the burden of designing and searching for a large codebook to the design of a complex entropy coder.

3.7. Predictive Coding with Quantization

From the discussion on VQ, we know that the optimal (lossy) compression of an image is to take the image as a whole vector and perform vector quantization (VQ). However, the complexity of such a vector quantizer always prohibits us to do so in practice. We are constrained to very small dimensional VQ, or in the extreme case, scalar quantization. However, the performance of such a small dimensional VQ would degrade too much if there is no proper signal processing to decorrelate successive vectors. The ideal signal processing scheme should totally decorrelate the quantization unit (scalars or vectors) so that there is not much performance loss when quantizing independently and encoding with a DMS entropy coder or an entropy coder with low-order models.

Let's revisit the predictive coding discussed before but now combined with quantization. We have seen that predictive coding as a powerful lossless coding technique could decorrelate source data by prediction. The resultant error values could be regarded as a DMS with quite low entropy. However, we can save more bits if we can tolerate some small errors in the reconstructed signal. Predictive coding with linear prediction is also referred to as Differential Pulse Code Modulation (DPCM).

The difference between lossy and lossless DPCM lies in the handling of the prediction error. In order to lower the bit rate, the error in lossy DPCM is quantized prior to encoding. A block diagram for a basic DPCM encoder and decoder system is shown in Fig. 10, where e^* represents the quantized prediction error.

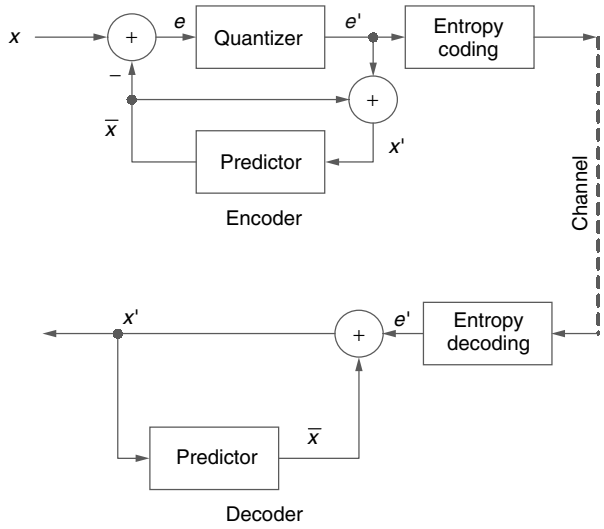


Figure 10. Diagram of a predictive coder combined with quantizer.

It is important to realize that in forming a prediction, the decoder only has access to the reconstructed values. Because the quantization of the prediction error introduces distortion, the reconstructed values typically differ from the original ones. To assure that identical predictions are formed at both the decoder and the encoder, the encoder also bases its prediction on the reconstructed values. This is accomplished by including the quantizer within the prediction loop as shown in Fig. 10. Essentially, each DPCM encoder includes the decoder within its structure. For the case of a successive-approximation quantizer, the encoder should fix a quantization level to be included in the prediction loop and the decoder should make certain

the same quantization level can be transmitted to avoid mismatch errors.

The design of a DPCM system should consist of optimizing both the predictor and the quantizer jointly. However, it has been shown that under the mean-squared error optimization criterion, independent optimizations of the predictor and the quantizer discussed in previous sections are good approximations to the jointly optimal solution. Because of the reconstruction dependency of a predictive coder, any channel errors could be propagated throughout the remainder of the reconstructed values. Usually, the sum of the coefficients is made slightly less than one (called leaky prediction) to reduce the effects of channel errors.

3.8. Linear Transformations

Predictive coding offers excellent de-correlation capability for sources with linear dependence. However, it has several drawbacks. First, its IIR filtering nature decides that the correct reconstruction of future values is always dependent on the previously correctly reconstructed values. Thus, channel errors are not only propagated but also accumulated to future reconstructed values. This makes predictive coding an unstable system under channel errors. Secondly, for lossy coding, because of the iterative prediction and quantization processes, it is hard to establish a direct relation between average distortion and rate of the quantizer, which in turn makes it hard for optimal rate control. Thirdly, predictive coding is a model-based approach and is less robust to source statistics. When source statistics changes, adaptive predictive coding normally has to choose different predictors to match the source. Moreover, the prediction coding is a waveform compression technique. Since the human visual system model is best described in the frequency domain, it is difficult to apply visual masking to the prediction error.

Alternatively, transform coding techniques can be used to reduce the correlation in source data. Transform coders take an M input source samples and perform a reversible linear transform or decomposition to obtain M transform domain coefficients that are decorrelated and more energy compacted for better compression. They decorrelate coefficients to make them amendable to efficient entropy coding with low-order models, and distribute energy to only a few coefficients and thus make it easy to remove redundancy and irrelevancy. Theoretically, the asymptotic MSE performance is the same for both predictive coding and transform coding [21]. However, transform coding is more robust to channel errors and source statistics. There is an explicit relationship between distortion and data rate after transformation, and optimal bit allocation or rate control can be easily implemented. The subjective quality is better at low bit rates since transform coding is normally a frequency domain approach and the HVS model can be easily incorporated in the encoding process.

Figure 11 illustrates the advantage of transformation over a highly correlated two-dimensional vector source with scalar quantization. We can see that the transformed signal would require much fewer bits to code than the original signal with the same distortion although they both use a uniform scalar quantizer in each vector dimension. Furthermore, the complexity of entropy coding following the quantization step is reduced in the transformed domain since the number of codewords is reduced.

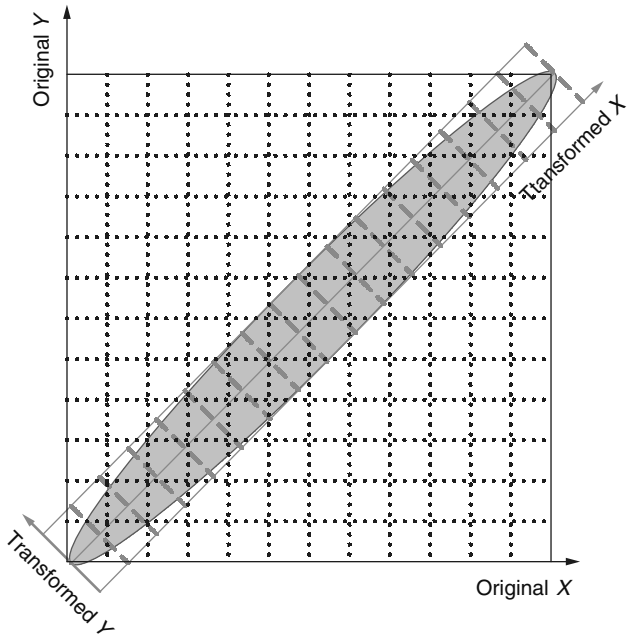


Figure 11. Advantage of transformation in signal compression.

The efficiency of a transform coding system will depend on the type of linear transform and the nature of bit allocation for quantizing transform coefficients. Most practical systems are based on suboptimal approaches for transform operation as well as bit allocation. There are mainly two forms of linear transformations that are commonly used in transform coding: block transforms and subband decompositions.

3.8.1. Block Transforms. Block transform coding, also called block quantization, is a widely used technique in image and video compression. A block of data is transformed so that a large portion of its energy is packed in relatively few transform coefficients, which are then quantized independently.

In general, a 1-D transformation scheme can normally be represented by a kernel matrix $T = \{t(k, n)\}_{k,n=0,1,\ldots,N-1}$ and the transformed results $Y = \{y_0, y_1, \ldots, y_{N-1}\}^T$ can be represented as the multiplications of the transform matrix T and the signal vector $X = \{x_0, x_1, \ldots, x_{N-1}\}^T$. That is,

$$Y = T \bullet X. \tag{19}$$

The original signal can be recovered by multiplying the inverse matrix of T with the transformed signal without any distortion if we ignore the rounding errors caused by limited arithmetic precision. From a signal analysis point of view, the original signal is represented as the weighted sum of basis vectors, where the weights are just the transform coefficients.

Block transforms are normally orthonormal (unitary) transforms, which means that,

$$T \bullet T^T = T^T \bullet T = I_{N\times N}, \tag{20}$$

where $I_{N\times N}$ is the identity matrix. Orthogonality is clearly a necessary condition for basis vectors to decompose an input into uncorrelated components in an N-dimensional space. Orthonormality of basis vectors is a stronger condition that leads to the signal energy preservation property in both the signal domain and the transform domain. Moreover, the mean squared error caused by quantization in the transform domain is the same as that in the signal domain and is independent of the transform. This greatly eases the encoding process where otherwise an inverse transform is needed to find the distortion in the original domain.

Image signals are two-dimensional signals. By nature the correlation between pixels is not separable, so a non-separable transform should be applied to decorrelate the pixels. However, since a 2-D separable correlation model provides practical simplicity and sufficiently good performance, 2-D separable transforms are widely used in practical image coding systems. A 2-D separable transforms can be easily implemented with two steps of 1-D transforms for both all rows and all columns subsequently. Similarly, for 3-D video signals, 3-D separable transforms can be implemented with 1-D transforms in each direction: horizontal, vertical and temporal, separately.

In practice, the block transforms are not applied to a whole image itself. Normally, the image is divided into subimages or blocks and each block is then transformed and coded independently. The transform coding based on a small block size does not necessarily degrade too much the coding efficiency, because: (1) from the VQ theory, we learned that beyond a certain vector size, the coding gain increase tends to saturate, so in this case, larger block sizes don't bring us significant additional coding gain anyway; (2) natural images and video are generally nonstationary signals, dividing them into small blocks is particularly efficient in cases where correlations are localized to neighboring pixels, and where structural details tend to cluster; and (3) the inter block correlation can still be exploited by predictive coding techniques for certain transform domain coefficients, for example, the DC component. From a theoretical analysis and simulation results, it is shown that for natural images the block size is optimal around 8 to 16. In most image and video coding standards such as JPEG, MPEG, a value of 8 has been chosen. Recently, there is also a trend to use adaptive block transforms where transform blocks may adapt to signal local statistics. The block processing has a significant drawback, however, since it introduces a distortion termed blocking artifact, which becomes visible at high compression ratios, especially in image regions with low local variance. Lapped Orthogonal Transforms (LOT) [30] attempt to reduce the blocking artifacts by using smoothly overlapping blocks. However, the increased computational complexity of such algorithms does not seem to justify wide replacement of block transforms by LOT.

The optimal block transform is the Karhunen-Loeve Transform (KLT) that yields decorrelated transform coefficients and optimum energy concentration. The basis vectors of KLT are eigenvectors of the covariance matrix of the input signal. However, the KLT depends on the second-order signal statistics as well as the size of the block, and the basis vectors are not known analytically. Even when a transform matrix is available, it still involves quite a large amount of transformation operations because the KLT is not separable for image blocks and the transform matrix cannot be factored into sparse matrices for fast calculation. Therefore, the KLT is not appropriate for image coding applications. Fortunately, there exists a unitary transform that performs nearly as well as the KLT on natural images but without the disadvantages of KLT. This leads us to the Discrete Cosine Transforms (DCT).

The Discrete Cosine Transform kernel matrix is defined as follows,

$$t_{kn} = u(k)\cos\left(\frac{\pi(2n+1)k}{2N}\right), \tag{21}$$

where

$$u(k) = \begin{cases} \dfrac{1}{\sqrt{N}}, & \text{if} \quad k = 0; \\ \sqrt{\dfrac{2}{N}}, & \text{if} \quad k \neq 0. \end{cases} \tag{22}$$

The DCT has some very interesting properties. First, it is verified that the DCT is very close — in terms of energy compaction and decorrelation — to the optimal KLT for a highly correlated first-order stationary Markov sequence [22]. Secondly, its transform kernel is a real function, so only the real part of the transform domain coefficients of a natural image or video must be coded. Moreover, there exist fast algorithms for computing the DCT in one or two dimensions [25–28]. All these have made DCT a popular transform in various image and video coding schemes. For a natural image block, after the DCT, the DC coefficient is typically uniformly distributed, whereas the distribution for the other coefficients resembles a Laplacian one.

There are some other transforms that also could be used for image and video coding, such as Haar Transform, or a Walsh-Hadamard Transform. The reason to use them is not because of their performance but because of their simplicity.

3.8.2. Subband Decomposition. Another form of linear transformation that brings energy compaction and decorrelation is subband decomposition. In subband decomposition (called analysis process), the source to be compressed is passed through a bank of analysis filters (filter bank) followed by critical subsampling to generate signal subbands. Each subband represents a particular portion of the frequency spectrum of the image. At the decoder, the subband signals are decoded, upsampled and passed through a bank of synthesis filters and properly summed up to yield the reconstructed signal. This process is called the synthesis process. The fundamental concept behind subband coding is to split up the frequency band of a signal

and then to code each subband using a coder and bit rate accurately matched to the statistics of the band.

Compared with block transforms, subband decomposition is normally applied to the entire signal and thus it can decorrelate a signal across a larger scale than block transforms, which translates into more potential coding gain. At high compression ratios, block transform coding suffers severe blocking artifacts at block boundaries whereas subband coding does not. The capability to encode each subband separately in accordance with its visual importance leads to visually pleasing image reconstruction. As a subset of subband decomposition, wavelet decomposition provides an intrinsic multi-resolution representation of signals that is important for many attractive image and video coding functionalities, such as adaptive coding, scalable coding, progressive transmission, optimal bit allocation and rate control, error robustness, and so forth.

It has been shown that the analysis and synthesis filters play an important role in the performance of the decomposition for compression purposes. One of the original challenges in subband coding was to design subband filters that cover well the desired frequency band but without aliasing upon the reconstruction step caused by the intermediate subsampling. The key advance was the development of quadrature mirror filters (QMF) [29]. Although aliasing is allowed in the subsampling step at the encoder, the QMF filters cancel the aliasing during the reconstruction at the receiver. These ideas continue to be generalized and extended.

A two-band filter bank is illustrated in Fig. 12. The filters $F_0(\omega)$ and $F_1(\omega)$ are the analysis lowpass and highpass filters, respectively, while $G_0(\omega)$ and $G_1(\omega)$ are the synthesis filters. In this system, the input/output relationship is given by

$$\begin{aligned} X'(\omega) = {} & \frac{1}{2}[F_0(\omega)G_0(w) + F_1(\omega)G_1(\omega)]X(\omega) \\ & + \frac{1}{2}[F_0(\omega+\pi)G_0(w) \\ & + F_1(\omega+\pi)G_1(\omega)]X(\omega+\pi), \end{aligned} \tag{23}$$

where the underlined term is where the aliasing occurs. Perfect reconstruction can be achieved by removing the aliasing distortion. One such condition could be, $G_0(\omega) = F_1(w+\pi)$ and $-G_1(\omega) = F_0(w+\pi)$. Furthermore QMF filters achieve aliasing cancellation by choosing, $F_0(\omega) = F_1(\omega+\pi) = -G_0(\omega) = G_1(\omega+\pi)$, where the highpass band is the mirror image of the lowpass band in the frequency domain.

As a special case to subband decomposition, wavelet decomposition allows nonuniform tiling of the time-frequency plane. All wavelet basis functions (baby wavelets) are derived from a single prototype (mother wavelet) by dilations (scaling) and translations (shifts). Besides the perfect reconstruction property, wavelet filters used for image and video compression face additional and often conflicting requirements, such as, compact support (short impulse response) of the analysis filters to preserve the localization of image features; compact support of the synthesis filters to prevent spreading of ringing artifacts; linear phase to avoid unpleasant waveform distortions around edges; and orthogonality to provide preservation of energy. Among them, orthogonality is mutually exclusive with linear phase in two-band FIR systems, so it is often sacrificed for linear phase. More information on wavelets and filter banks can be found in [146–148]. Wavelet decomposition of discrete sources is also often referred to as Discrete Wavelet Transforms (DWT).

Two-band systems are the basic component of most subband decomposition schemes. Recursive application of a two-band filter bank to the subbands of the previous stage yields subbands with various tree structures. Examples are uniform decomposition, octave-band (pyramid) decomposition, and adaptive or wavelet-packet decomposition. Among them, pyramid decomposition is the most widely used in image and video coding where a multi-resolution representation of image and video is generated. Based on the fact that most of the frequency of an image is concentrated in low-frequency regions, pyramid decomposition further splits the lower frequency part while keeping the high-pass part at each level untouched. Figure 13 illustrates such 2-D separable pyramid decomposition intuitively. Pyramid decomposition provides inherent spatial scalability in the encoded bit stream. On the other hand, an adaptive wavelet transform or wavelet-packet decomposition chooses the band splitting method according to the local characteristics of the signal source in each subband to achieve better compression.

Although applied to a whole image frame, the discrete wavelet transform is not as complicated as it seems. Lifting [48] is a fast algorithm that can be used to efficiently compute DWT transforms while providing some new insights on wavelet transforms.

As we mentioned before, one advantage of subband coding over block transform coding is the absence of the blocking effect. However, it introduces another major artifact — a ringing effect which occurs around high-contrast edges due to the Gibbs phenomenon of linear filters. This artifact can be reduced or even removed by an

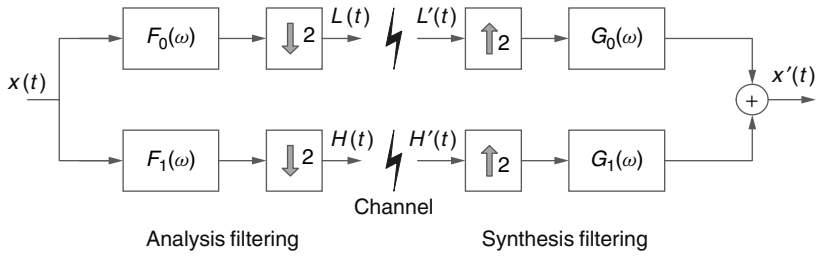


Figure 12. Two-band wavelet analysis/synthesis system. $F_0(\varpi)$ and $F_1(\varpi)$ are the lowpass and highpass analysis filters, respectively; $G_0(\varpi)$ and $G_1(\varpi)$ are the lowpass and highpass synthesis filters, respectively.

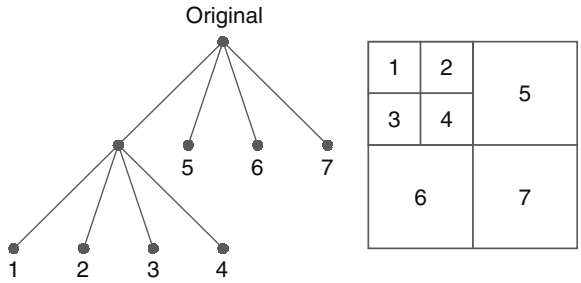


Figure 13. Illustration of a wavelet decomposition of depth two.

appropriate design of the filter bank or alleviated by some de-ringing post-filtering.

It is worth mentioning that to avoid the ringing artifacts caused by linear filtering, some nonlinear morphological filter banks are also being investigated [31–33].

It can be shown that block transforms are a special case of subband decompositions where the synthesis filters' impulse responses are the transform basis functions, the analysis filters' impulse responses are the time-reversed basis functions, and the decimation factor in each subband is the transform block length.

Note that the transformations (block transforms or subband decompositions) themselves do not compress data. The compression occurs when quantization and/or entropy coding techniques are applied to the transform coefficients. In fact, with sufficient arithmetic precision, the transformations should losslessly recover the original signal. There is also a class of transformations (reversible transforms) that can perfectly recover the original signal with limited precision. They are especially useful for lossless compression.

3.8.3. Vector Transformations. It is known from both information theory and practice that vector quantization always outperforms scalar quantization. However, when we replace the scalar quantization with vector quantization in the traditional predictive coding and transform coding schemes, the improvement is not significant. The main reason is because various transformations have a good decorrelation capability and have already decorrelated the components in the vectors to be vector quantized. As we learned from the vector quantization discussion in subsection 3.6.2 the most significant gain of VQ is from memory advantage. Is there a way to jointly optimize the transformation stage and the VQ stage so that the overall coding gain is maximized? The answer to this question is vector based signal processing [39], including vector transform coding [41,42], and vector wavelet/subband coding [40,46]. There are various kinds of vector transformation schemes, but the principle is the same, that is, (1) reduction of intervector correlation: the signal processing operations reduce correlation between the vectors as much as possible (2) preservation of intravector correlation: the signal processing operations preserve correlation between the components of each vector as much as possible. Image and video coding experiments show that vector-based transformation can achieve additional gain compared with scalar transformation followed by vector quantization with the cost of more complexity.

3.8.4. Shape-Adaptive Transformations. New applications in multimedia communications result in the need for object-based functionalities. Object-based functionalities require a prior segmentation of a scene into objects of interests. The objects are normally of arbitrary shapes. A new challenge that arises is how to efficiently compress the texture information within an arbitrarily-shaped object because the signal processing techniques we discussed so far are all based on rectangular regions. Of course, various padding schemes can be applied to expand the arbitrarily-shaped region into a rectangular one. However, such schemes are not very efficient because they need to code more coefficients than pixels within that region.

The objective of transformation over an arbitrarily-shaped region is still to decorrelate the data and achieve a high energy compaction. To be consistent with the techniques for rectangular regions, various approaches have been proposed to extend the block transforms or subband decomposition techniques to support arbitrarily-shaped regions. Among them, most notably are POCS-based block transforms (PBT) [22], shape-adaptive DCT (SA-DCT) [34,35] and shape-adaptive DWT (SA-DWT) [36,136,149,150].

The PBT, where POCS stands for projection onto convex sets, is based on two iterative steps to determine the best transform domain coefficient values. The first step transforms the pixels within an arbitrarily-shaped region through a normal rectangular transform and resets a selected set of coefficients to zeros, so that the number of nonzero coefficients equals to the number of pixels within the region. The second step performs an inverse transform of the coefficients obtained and resets the pixels within the arbitrarily-shaped region back to their original values. These two steps are iterated until the result converges to within a certain criterion. Although the convergence of the algorithm is guaranteed, this transform cannot achieve perfect reconstruction. However, as most of the energy of a natural image is concentrated in the low-frequency coefficients, the loss is minimized.

The SA-DCT is based on rectangular DCT transforms. A 2-D SA-DCT operates first vertically then horizontally. The SA-DCT begins by flushing all the pixels of the region up to the upper border of a rectangular block so that there are no holes within each column. Then an n-point DCT transform is applied to each column and where n is the number of pixels in that column. The column-transformed coefficients are normalized for the subsequent row processing. The same procedure repeats horizontally for each row to obtain the final transformed coefficients. The advantages of the SA-DCT algorithm include: computational simplicity, reversibility as long as shape information is available, effectiveness and the same number of transform coefficients as that of the pixels within the original shape. The main drawbacks of this approach lie on decorrelation of non-neighboring pixels and potential blocking artifacts.

On the other hand, the SA-DWT is an extension of the normal DWT transforms. Each level of SA-DWT is

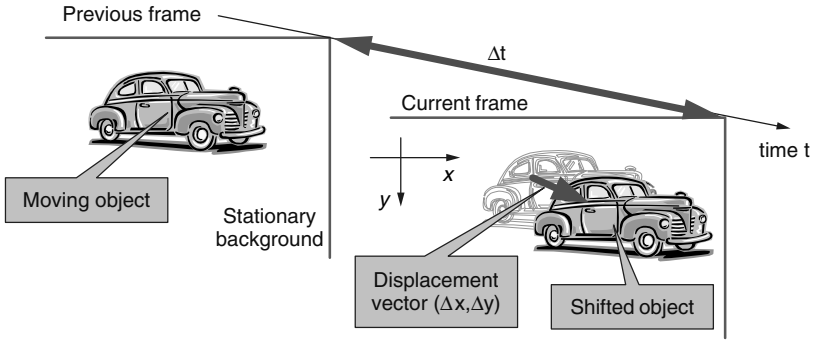


Figure 14. Illustration of motion compensation and estimation concept.

applied to each row of an image and then on each column. The key concept of SA-DWT is to keep the phase of each transformed coefficient aligned so that locality and spatial relation between neighboring pixels are well preserved. Based on the available shape information, SA-DWT starts by searching disjoined pixel segments in each row of an arbitrarily-shaped region and applies the normal DWT to each of the segments separately but the subsampling positions of each segment are aligned no matter where each segment starts. The subsampled coefficients are put in their corresponding spatial positions so that relative positions are preserved for efficient decorrelation in later steps. The same operation is then applied for each column of the row-transformed coefficients. SA-DWT also keeps the number of transform coefficients equal to the number of pixels in an original shape but with much improved decorrelation properties across an arbitrarily-shaped object texture. It has been shown that SA-DWT combined with advanced quantization and entropy coding techniques can achieve much improved efficiency over SA-DCT for still image object coding. Moreover, it will not suffer from the blocking artifacts existing in SA-DCT.

3.9. Motion Estimation and Compensation

The prediction or transformation techniques discussed so far all aim at de-correlating highly correlated *neighboring* signal samples. They are especially effective for spatial neighbors within an image or video frames. Video is a set of temporal samples that capture a moving scene across time. In a typical scene there is a great deal of similarity or correlation between neighboring images of the same sequence. However, unless the video sequence has slow motion, directly extending the above de-correlating techniques to the temporal direction would not always be very effective. The direct temporal neighboring samples that are spatially collocated are not necessarily highly correlated because of the motion in the video sequence. Rather, the pixels in neighboring frames that are more correlated are along the motion trajectory or optical flow and there usually exists a spatial displacement between them. Therefore, an efficient signal decorrelation scheme in the temporal direction should always operate on pixels along the same motion trajectory. Such a spatial displacement of visual objects (pixels, blocks or objects) is called a motion vector and the process of determining how objects move from one frame to another or finding the motion vector is called motion estimation. The operation of de-correlating samples in different frames using motion information is called motion compensation. The concept of motion estimation and compensation is illustrated in Fig. 14. Motion compensation is probably the most important factor in video coding that dominates most of the significant video coding advancements in recent years. It seems that further improvements of motion compensation are a continuing need.

The analogies to the normal prediction techniques and transformations in the temporal direction considering motion information are motion compensated predictions or motion predictions and motion compensated transformations or spatiotemporal transformations, respectively.

Motion estimation is the first step of these techniques, which is essentially a search process that normally involves heavy computation. Fortunately, there are many fast yet efficient motion estimation algorithms available today.

Ideally, each pixel in a video frame should be assigned a motion vector that refers to a pixel most correlated with it. However, it requires not only increases in the already heavy computational load for motion vector search, but also a significant amount of bits to encode the motion vector overhead. Therefore, in practice, pixels are usually grouped into a block with fixed or variable size for the motion estimation. Variable block size can adapt to the characteristics and object distribution in video frames and achieve better estimation accuracy and motion vector coding efficiency.

There are many motion estimation algorithms available, for example, block matching, where the motion vector is obtained by searching through all the possible positions according to a matching criterion; hierarchical block matching, where the estimated motion vectors are successively refined from larger size blocks to smaller size ones; gradient matching, where the spatial and temporal gradients are measured to calculate the motion vectors, and phase correlation, where the phases of the spectral components in two frames are used to derive the motion direction and speed. Among them, gradient matching and phase correlation techniques actually calculate the motion vectors of moving objects rather than estimating, extrapolating or searching for them. Block matching is the simplest and best studied one, and it is being widely used in practical video coding systems. To reduce the complexity of a brute-force full search while keeping the motion vector accuracy, many fast search schemes have been developed for block matching algorithms, for example,

3-step search (TSS) [49], Diamond search (DS) [50,51], Zonal-based search [52], and so on. Advanced search algorithms [53–55] even take advantage of the interframe motion field prediction to further speed up the motion estimation process.

Moreover, because the motion of objects in a video scene is not just limited to 2D translation, global motion estimation algorithms [56] try to capture the true object motion and further improve video coding efficiency by using an extended set of motion parameters for translation, rotation, and zooming.

3.9.1. Motion Compensated Prediction. The most widely used technique for de-correlating the samples along a temporal motion trajectory is motion compensated prediction. Motion compensated prediction forms a prediction image based on one or more previously *encoded* frames (not necessarily past frames in display order) in a video sequence and subtracts it from the frame to be encoded to generate a *residue* or *error* image to be encoded by various coding schemes. In fact, motion compensated prediction can be viewed as an adaptive predictor where the goal is to find minimum difference between the original image and the predicted image. Different techniques developed for improving motion prediction are just variations of the adaptive predictor. For example, integer-pel motion compensation is the basic form of the adaptive predictor where the choices of different predictions are signaled by the motion vectors; fractional-pel (half-pel, quarter-pel, or 1/8-pel) motion prediction is a refinement of integer-pel motion prediction where the choices of predictions are increased to provide more accurate predictions and they are signaled by the additional fractional-pel precision [57]; P-type prediction is formed by predictors using only a past frame as references; B-type prediction uses both the past frame and future frame; long-term prediction maintains a reference that is used frequently by other frames; advanced video coding schemes also use multiple frames as the input to the adaptive predictor to maximize the chance of further reducing the prediction errors with the expense of increased encoding complexity and more overhead for predictor parameters; overlapped block motion compensation (OBMC) [57] and de-blocking or loop filters in the prediction loop are ways to exploit more spatially correlated pixels at the input of the adaptive predictor without the need to transmit the overhead for more predictor parameters, while improving the visual quality of the decoded image; adaptive block size in motion prediction forces the motion predictor to adapt to the local characteristics of an image; global motion based prediction is not a linear prediction anymore and it forms more accurate prediction with complex warping operations considering the fact that the motion of video objects is not limited to 2-D translations. There are continuing research efforts on choosing better adaptive predictors, and there is always a delicate trade-off between complexity, overhead bits for predictor parameters, and prediction errors.

Ideally, if the motion compensated prediction simply removes the temporal correlation without affecting the spatial correlation between pixels, any efficient image coding schemes can be used to efficiently encode the prediction residue images with a little modification on the parts that heavily depend on the statistics of pixel values, for example, the entropy coding tables. Indeed, many existing video coding systems extend the still image coding methods to encode the residue image and achieve great results, for example, MPEG and H.26x coding standards using block-based motion compensation and block transformation for the residue coding.

Unfortunately, block-based systems have trouble coding sequences at very low-bit rates and the resultant coded images have noticeable blocking artifacts. The high-frequency components caused by the blocking artifacts in the residue image partly change the high-correlation nature of the residue. One solution to this problem is a lowpass filter in the motion compensation loop (either on the reference image or on the predicted image) to remove the high-frequency components. Another solution is to apply new signal decomposition techniques that match the statistics of the residue. Matching Pursuit [61–63] is one of these techniques that can be used to efficiently encode the prediction residue image. Instead of expanding the motion residual signal on a complete basis such as the DCT, an expansion on an overcomplete set of separable Gabor functions, which do not contain artificial block edges, is used. Meanwhile, it removes grid positioning restrictions, allowing elements of the basis set to exist at any pixel resolution location within the image. All these allow the system to avoid the artifacts most often produced by low-bit rate block-based systems.

Motion compensated prediction can work efficiently with dependency only on one immediate previous frame. This makes it very suitable for real-time video applications, such as video phone and video conferencing, where low delay is a critical prerequisite. As with any prediction coding schemes, the major drawback of predictive coding is the dependency on previously coded frames. This is especially more important for video. Features like random access and error robustness are limited in motion compensated predictive coding and transmission errors could be propagated and accumulated (called error drifting). Moreover, a predictive scalable coder must design ways to compensate the loss caused by the unavailability of part of the reference bit stream.

3.9.2. Motion Compensated Transformations. As we mentioned, direct extension of 2-D spatial signal decomposition to 3-D spatiotemporal decomposition without motion compensation [64] is not efficient especially for moderate to high motion sequences, though they offer computational simplicity and freedom from motion artifacts. Without motion compensation, the decoded video may normally present severe blurring and even ghost image artifacts.

With added complexity, motion compensated 3-D transformations align video samples along the motion trajectory to better decorrelate them for better coding efficiency. In addition, motion compensated transformations provide enhanced functionalities like scalability, easier rate-distortion optimization, easier error concealment and limited error propagation in error-prone channels, and so on. A particularly interesting example is the spatiotemporal subband/wavelet coding of video, where 3-D

subband/wavelet transforms are applied to the motion aligned video data [65–68]. One of the key issues is to align the pixels in successive frames with motion information. This is still an active research topic and some good results have been reported. Another issue is related to the boundary effect that may be present in the temporal direction due to limited-length wavelet transforms. A solution is proposed in [69] to use a lifting algorithm.

One of the major drawbacks of 3-D transformations is that normally a few frames are involved and there exist considerable encoding and decoding delays. Therefore, it is not suitable for real-time communications.

3.10. Fractal Compression

Quite different from all the other schemes we have presented so far, fractal compression exploits the piecewise self-transformability (self-similarity) property existing in natural images that each segment of an image can be properly expressed as a simple transformation (rotation, scaling, translation) of another part having a higher resolution. Fractal compression is based on the iterated functions systems (IFS) theory pioneered by Barnsley [70] and Jacquin [71] and followed by numerous contributions [72]. The fundamental idea of fractal coding is to represent an image as the attractor of a contractive function system through a piecewise matching algorithm. There is no need to encode any pixel level in fractal compression, and the encoded image can be retrieved simply by iterating the IFS starting from any initial arbitrary image.

An advantage of fractal coding is that image at different levels of resolution can be computed through the IFS, without using interpolation or the duplication of pixel values. However, since it involves a search process, it requires quite intensive computation. Also, self-similarity is not self-identity, and fractal coding is always lossy.

Fractal compression is in fact related to vector quantization, but in contrast to classical vector quantization it uses a vector codebook drawn from the image itself rather than a fixed codebook. Fractal-based image compression techniques have been shown to achieve very good performance at high compression ratios (about 70–80) [73,74].

3.11. Bit Allocation and Rate Control

The ultimate goal of lossy image and video compression is to squeeze image or video data into as few bits as possible under certain quality criteria, or to get as much as possible quality under certain bit budget constraint. Besides choosing the best combinations of transformations, quantizers, and entropy coders, we have to optimally distribute the available bits across different components so as to achieve the best overall performance. This brings us to bit allocation and rate control which are indispensable for practical image and video coding systems. Rate control usually regulates the bit rate of a compression unit according to conditions not just target bit rate, but also encoder and decoder buffer models, and constant quality criteria, and so forth. On the other hand, bit allocation tries to make the quality of a picture as good as possible given a bit budget that may be assigned by the rate control.

In fact, if the target is just concerned about bit rate versus quality, rate control and bit allocation are closely related and sometimes it is hard to distinguish the difference between them. In this case, a generic rate control (bit allocation) problem can be formulated as the problem to minimize the overall distortion (quantization error),

$$D = \sum_i D_i, \tag{24}$$

under the constraint of a given total bit rate

$$R = \sum_i R_i, \tag{25}$$

by assigning to each compression unit the appropriate quantizer having a distortion D_i and a rate R_i. It has to be emphasized here that the only assumption made is that the overall distortion can be written as a sum of the individual distortions. No assumption about the nature of the distortion measure and the quantizers is made. Each compression unit can have its own distortion measure and its own admissible quantizer. Normally, the MSE is used as a measure of the distortion, although it is not a good measure for the quality of natural images or video to be evaluated by the human visual system.

There are two scenarios when dealing with bit allocation in practice: independent coding and dependent coding. Independent coding refers to the cases where the compression units (image pixels, blocks, frames, or subbands) are quantized independently. However, many popular schemes involve *dependent* coding frameworks, that is, where the R-D performance for some compression units depends on the particular choice of R-D operating points for other units. Typical examples of dependent coding are various predictive coding schemes such as DPCM and motion compensated predictive video coding. For the simple independent coding case, the optimization problem leads to a necessary *Pareto* condition,

$$\frac{\partial D_i}{\partial R_i} = \frac{\partial D_j}{\partial R_j}, \text{ for all } i \text{ and } j, \tag{26}$$

which states that when the optimal bit allocation is achieved, the slopes of the rate-distortion curves at the optimal points should be the same.

The rate-distortion (R-D) theory is a powerful tool for bit allocation. Under the R-D framework, there are two approaches to the bit allocation problem: an analytical model-based approach and an operational R-D based approach. The model-based approach assumes various input distribution and quantizer characteristics [75–78]. Under this approach, closed-form solutions can be obtained based on the assumed models using continuous optimization theory. Conversely, the operational R-D based approach [79–81] considers practical coding environments where only a finite set of quantizers is admissible. Under the operational R-D based approach, the admissible quantizers are used by the bit allocation algorithm to determine the optimal strategy to minimize the distortion under the

constraint of a given bit budget. Integer programming theory is normally used in this approach to find the optimal discrete solution. Because the operational R-D approach can achieve exactly the practical optimal performance for completely arbitrary inputs and choices of discrete quantizers, it is often preferred in a practical coding system. However, it requires that: (1) the number of admissible quantizers or quantizer combinations is tractable for practical coding systems; (2) the coding system is capable of providing all the operational R-D points easily or tolerates the possible long delays and high complexity caused by calculation of these R-D points. Therefore, practically the operational approach is normally applied to independent coding cases or scalable coding schemes where a set of operational points can be easily obtained from the embedded bit streams. The model-based approach is normally applied in scenarios where delay and complexity cannot be tolerated or dependent coding cases where the combinations of admissible quantizers are out of control because of the dependency. If the input statistics are known, model-based bit allocation can be used to derive a predetermined bit allocation strategy that best fits the input source. Otherwise, it will work in a *feed-forward* fashion based on the heuristics where the parameters of the model can be updated adaptively and hopefully converge to an optimal bit allocation plan in the long run. The performance of model-based bit allocation can be also improved through multipass coding to refine the bit allocation among compression units.

For operational R-D based bit allocation, the problem normally involves finding the *convex hull* where the optimal points lie on [82]. Figure 15 shows an example of the convex hull on an R-D plot. Finding a convex hull of a set of points has been extensively investigated and many fast algorithms have been designed [83]. In order to avoid the impractical exhaustive computation of all the combinations, an algorithm was designed to find the convex hull in a limited number of computations [84]. For a single optimization point, the *BFOS* algorithm can be used to quickly allocate the available bits [85].

For a Gaussian source and MSE distortion criterion, assuming that the distortion error of each component

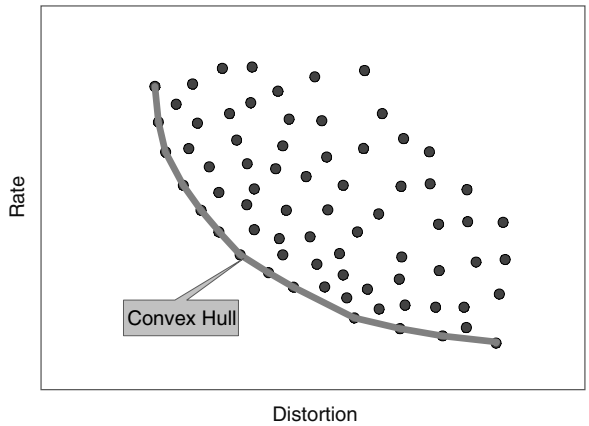


Figure 15. The convex hull of an R-D plot. The block dots represent all the admissible R-D points and the lower solid curve is the convex hull.

should be within a threshold θ, the optimal bit allocation can be derived as,

$$R_i = \begin{cases} \frac{1}{2}\log_2\left(\frac{\sigma_i^2}{\theta}\right), & \text{if } \sigma_i^2 > \theta; \\ 0, & \text{if } \sigma_i^2 \le \theta; \end{cases} \tag{27}$$

and the distortion for each component is given by,

$$D_i = \begin{cases} \theta, & \text{if } \sigma_i^2 > \theta; \\ \sigma_i^2, & \text{if } \sigma_i^2 \le \theta; \end{cases} \tag{28}$$

This result shows that all the components should have the same quantization error except for those with energy σ_i^2 lower than certain threshold θ. By adjusting the value θ, various bit rates can be achieved using the above equations.

Using such a fixed bit allocation, the number of bits used for each component is fixed for all compression units of the source, so the allocation information only needs to be sent out once and all other bits can be used to encode the coefficient values. Such an optimal bit allocation is totally dependent on the statistical characteristics of the source; specifically, the variances of transform components are needed and in general, the source has to be stationary. While producing a constant-rate bit stream, coders using fixed bit allocation cannot adapt to the spatial or temporal changes of the source, and thus coding distortion may vary from one compression unit to another due to changes of the source. Conversely, adaptive bit allocation schemes can be used to deal with the random changes of a source. However, overhead bits will be introduced to indicate the difference in bit allocation schemes. Combining with the possible entropy coding afterwards, a closed-form formula is generally not available to explicitly indicate the relationship between bits and distortion. A feed-forward heuristic or multipass approach can be applied to hopefully obtain an optimal bit allocation on average. A simple example is the threshold coding that is used widely in many image and video coding standards nowadays. Threshold coding fully takes advantage of the energy packing property of transformations by encoding only those significant coefficients. In such a coding scheme, only those coefficients whose energy is higher than the threshold are quantized and encoded; all others will be treated as zero and discarded. This significantly increases the number of zeros and reduces that of significant coefficients to be encoded. Besides, the threshold could be much larger than the quantization step to take advantage of efficient zero coding. This is the basis for many *dead-zone* quantizers. The threshold itself does not need to be encoded in the bit stream; therefore, in the extreme case it could be optimized differently from coefficient to coefficient. Because threshold coding depends only on local energy, it can easily adapt to changes of the source. The drawback is that the bit rate cannot be predicted exactly, and it depends on the threshold, quantization steps, and side information about the locations of zeros. It generates variable bit rate for each coding unit. However, with an output buffer and a proper rate control scheme, the output rate can be kept within a bit budget while maintaining optimal performance on the average.

When delivering image or video data over time-varying channels such as wireless channels and the best effort IP networks, the bit budget that best fits the channels could not be predetermined at the encoder time. Scalable image [86] and video coding [38,87] can dynamically adjust the bit rate on the fly to adapt to the channel conditions (bandwidth, throughput, or error rate, etc.). This essentially results in a new coding scheme where the rate control is put off from the encoder time to the delivery time. How to quickly and optimally allocate the bits on the fly is an active research topic [86,88].

3.12. Symbol Formation

As an important and sometimes crucial part of final step in image and video compression systems, symbol formation essentially organizes the final coefficients that may or may not be quantized in a form that is more suitable for efficient entropy coding subsequently. Common symbol formation techniques are run-length coding (RLC), zig-zag scanning, zerotree scanning, and context formation for conditional entropy coding, and so forth.

3.12.1. Run-Length Coding (RLC). *Run-length coding* (*RLC*) is a very simple form of data compression. It is based on a simple principle that every segment of the data stream formed of the same data values (called *run*), that is, sequence of repeated data values, is replaced with a pair of count number (length) and value (run). This intuitive principle works best on certain data streams that contain a large number of consecutive occurrences of the same data values. For example, in the image domain, the same values or prediction differences of neighboring pixels often appear consecutively; in the transform domain, if the highly compact coefficients are sorted according to their energy distribution, after quantization, they often contain a long run of zeros.

RLC is a lossless coding scheme and is often combined with the subsequent entropy coding. It is generally believed that for a Markov source, RLC combined with entropy coding would achieve the same efficiency as we encode each data item with a conditional entropy coding scheme. However, RLC can be easily implemented and quickly executed.

3.12.2. Zigzag Scanning. As we know, block transform can concentrate the energy of the image or video data to only a few transform domain coefficients. Conversely, threshold coding employs a dead-zone in the quantizer so that many coefficients with little energy distribution can be quantized to zeros. Also, threshold coding is an adaptive quantization scheme where the locations of nonzero coefficients have to be encoded. By sorting the transform coefficients according to their energy distribution, one can get a coefficient sequence where the quantized coefficients with lower probability to be zeros are put toward the beginning of the sequence and those with higher probability are put towards the end of the sequence. For image and video data, most of the energy would be concentrated in low-frequency coefficients. Therefore, such sorting would result a zigzag scan order starting from lower frequency coefficients to higher ones and with more consecutive zeros toward the end of the sequence; thus, it makes run-length coding more efficient. Zigzag scanning takes advantage of the inherent statistics in the image and video data and arranges it in an order that is more convenient for subsequent coding. Figure 16 gives an example of such a zigzag scan order for an 8×8 DCT domain coefficients, which is commonly used in JPEG and MPEG image and video coding standards.

3.12.3. Zerotree coding. For subband/wavelet based decompositions, symbol formation is a little bit different. In the tree structure of octave-band wavelet decomposition shown in Fig. 17, each coefficient in the high-pass subbands has four coefficients corresponding to its spatial position at a higher scale. Because of this very structure of the decomposition, there should be a good way to arrange its coefficients to achieve better compression efficiency. Based on the statistics of the decomposed coefficients in each subband, it has been observed that if a wavelet coefficient at a coarse scale is insignificant with respect

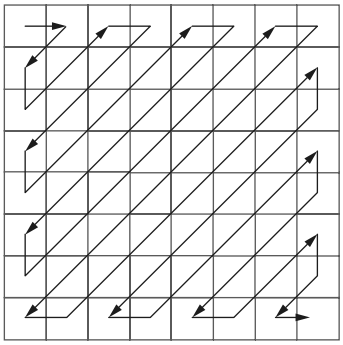


Figure 16. An exemplar zigzag scan order.

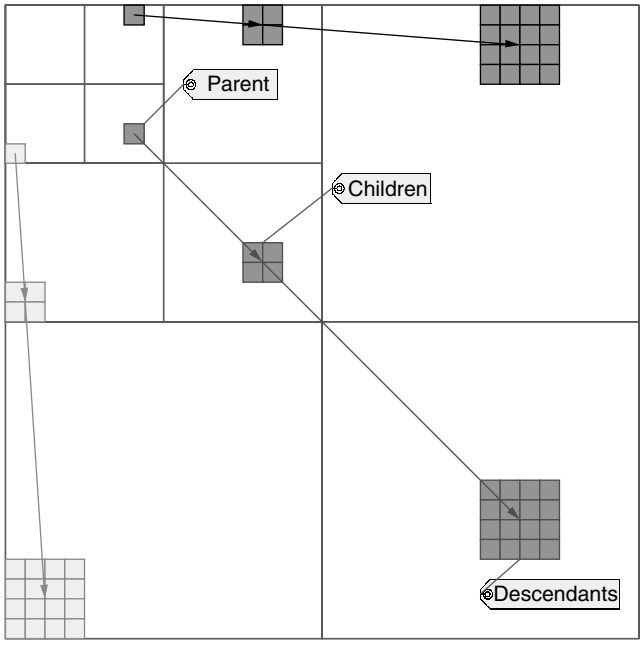


Figure 17. Illustration of parents and children relationship in a pyramid decomposition.

to a given threshold T, then all wavelet coefficients of the same orientation in the same spatial location at a finer scale are also likely to be insignificant with respect to T. EZW image compression is a wavelet based coder that was first proposed by Shapiro [89] in 1993. It uses the zerotree structure and takes advantage of the self-similarity between subbands at different scale. A zerotree [89] is formed with the root coefficient and all its children are insignificant. Figure 17 shows such a tree structure. An encoder can stop traversing the tree branch once it detects that the node is a zerotree root. Many insignificant coefficients at higher frequency subbands (finer resolutions) can be discarded. Zerotree coding can be considered as the counterpart of zigzag scan order in subband/wavelet coding. It also uses successive-approximation quantization to quantize the significant coefficients and context-based arithmetic coding to encode the generated bits. Bits are coded in order of importance, yielding a fully embedded code. The main advantage is that the encoder can terminate the encoding at any point, thereby allowing a target bit rate to be met exactly. Similarly, the decoder can also stop decoding at any point resulting in the best image at the rate of the truncated bit stream. The algorithm produces excellent results without any prestored tables or codebooks, training, or prior knowledge of the image source.

In general, ordered data are more easily compressed than un-ordered data. But when organizing data in an order that is not predefined, the order information has to be encoded in the bitstreams as overhead. Zigzag scan works well because it is a predetermined order and fits the source statistics well. On the other hand, for zerotree coding, partial order information such as a significance map has to be sent to the decoder. An efficient coding algorithm can be derived, if a good balance between overhead information for ordering and a good ordering scheme for subsequent entropy coding is achieved. Set Partitioning in Hierarchical Trees (SPIHT) algorithm [90], rearranges the transformed coefficients by partial ordering according to their magnitudes with a set partitioning sorting algorithm, transmits refinement bits in an order according to the ordered bitplane, and exploits the self-similarity of the wavelet transform across different scales. The algorithm enables progressive transmission of the coefficient values by ordering the coefficients and transmitting the most significant bits first. The ordering information also makes quantization more efficient by first allocating bits to coefficients with larger magnitudes. An efficient coding scheme of the ordering information is also included in the algorithm. The results [90] show that the SPIHT coding algorithm in most cases surpasses those schemes obtained from various zerotree algorithms [89].

3.12.4. Context Formation. It can be shown that when a data source is partitioned into different classes, the attainable entropy can be smaller than the unpartitioned one. Thus, it suggests that we can achieve more coding efficiency when a source is divided into different groups with significant different probability distributions. One way is to use a search algorithm to find which subgroup the encoded data belong to and use this information to drive a conditional entropy coding scheme. However, this would require overhead bits in explicitly coding the class of the subgroup. Fortunately, almost any data to be encoded in image and video compression are very dependent on its neighboring context. Such context information can be used to drive different entropy coders that are most suitable for the data to be coded. Context-based entropy coding is essentially a prediction coding scheme. The best part is that the prediction is *implicit* and the dependency or correlation between the data and the context does not have to be known explicitly.

As we have discussed in the entropy coding part, one challenge of context formation is to keep the contexts as small as possible while reflecting as much dependency as possible. Lower entropy can be achieved through higher order conditioning (larger contexts). However, larger context implies a larger *model cost* [92], which reflects the penalties of *context dilution* when count statistics must be spread over too many contexts, thus affecting the accuracy of the corresponding estimates, especially for adaptive context-based coding. Moreover, larger contexts means more memory is needed to store the probability tables. This observation suggests that the choice of context model should be guided by the use, whenever possible, of available prior knowledge on the data to be modeled, thus avoiding unnecessary learning costs. Often explicit prediction and context-based coding schemes can be combined to reduce the model cost even though they might be based on the same context [93].

3.13. Preprocessing and Postprocessing

In addition to the core techniques we have discussed, additional preprocessing and postprocessing stages are extensively used in image and video compression systems in order to render the input or output images in a more appropriate format for the purpose of coding or display. The possible operations include but are not limited to de-noising, format/color conversions, compression artifacts removal, error concealment, and so on.

As with any other signals captured from a natural source, image and video data normally contain noise in them. The problem becomes more severe as many low-end consumer-grade digital image and video devices are gaining in popularity. As we learned from information theory, truly random noise not only is very hard to compress but also degrades the image and video quality. It is very common to use a de-noising filter [94,95] prior to coding in order to enhance the quality of the final pictures and to remove the various noises that will affect the performance of compression algorithms.

The simplest compression techniques are based on interpolation and subsampling of the input image and video data to match with the resolution and format of the target display devices. For example, a mobile device is normally only equipped with limited display screen resolution, and a subsampled image that matches its screen size can just meet its needs while greatly reducing the bit rate. The challenge of subsampling is how to maintain the crispness of a picture without introducing aliasing, that is, how to select the lowpass filters, which is a classic image processing problem [96,97].

The human eyes are more sensitive to the difference in brightness than to the differences in color. To exploit this in image and video coding systems, the input images are often converted to YUV components (one luminance Y and two chrominance differences U and V) instead of using RGB components (red, green, blue). Furthermore, the components U and V often can be subsampled at a lower resolution.

Image format conversion is also a common postprocessing step for image and video decoding. Often the image or video signal is not encoded exactly as the same format supported by the display devices, for example, in terms of image size, frame rate, interlaced or progressive display, color space, and so on. Postprocessing must be able to convert these different formats of image video into the display format. There are active studies on this issue, especially on resizing of images [96,97], frame rate conversion [98,99], interlaced to progressive video (de-interlacing) [100–102].

It is normal to expect a certain degree of distortion of the decoded images for very low bit rate applications though a good coding scheme should introduce these distortions in areas less annoying for the users. Postprocessing can be used to further reduce these distortions. For block transform coding, solutions were proposed to reduce the blocking artifacts appearing at high-compression ratios [103–108]. Similar approaches have also been used to improve the quality of decoded signals in other coding schemes such as subband/wavelet coding, reducing different kinds of artifacts such as ringing, blurring, mosquito noise, and so on [109,110]. In addition to improving the visual quality, filtering in the prediction loop such as in motion compensation could also improve the coding efficiency [111,112].

When delivering encoded image and video bitstreams over error-prone or unreliable channels such as the best-effort Internet and wireless channels, it is quite possible to receive a partially corrupted bit stream due to packet losses or channel errors. Error concealment is a type of postprocessing technique [113–116] used to recover the corrupted areas in an image or video based on prior knowledge about the image and video characteristics, for example, spatial correlation for images or motion information for video. Other related preprocessing and postprocessing techniques are image and video object segmentation and composition for object-based video coding such as in MPEG-4 [117]; scene change detection for inserting key-frames in video coding [118]; variable frame-rate coding [119] according to the video contents and the related frame interpolation [120]; region of interests (ROI) coding of images [121]; and so forth.

Image and video coding is a very active research topic and progress on new compression technologies is being made rapidly. In a limited space, we can grasp only the fundamentals of these techniques. Hopefully, these basic principles can inspire and at least help readers to understand new image and video compression technologies.

4. IMAGE AND VIDEO CODING STANDARDS

We have introduced a number of well-known basic compression techniques commonly used in image and video coding. Practical coding systems all contain one or more of these basic algorithms. Now, we present several image and video coding standards. The importance of standards in image and video coding is due to the fact that an encoder and a decoder of a coding algorithm may be designed and manufactured by different parties. To ensure the interoperability, that is, one party's encoded bit stream can be decoded by the other party's decoder, a well-defined standard has to be established. It should be pointed out that a standard, such as MPEG-1, MPEG-2, or MPEG-4, only defines the bit stream syntax and the decoding process, leaving the encoding part open to different implementations. Therefore, strictly speaking, there is no such thing as video quality of a particular standard. The quality of a standard compliant bit stream at any given bit rate depends on the implementation of the encoder that generates the bit stream.

4.1. Image Coding Standards

4.1.1. ITU-T Group 3 and Group 4 Facsimile Compression Standards. One of the earliest lossless compression standards are the ITU-T (former CCITT) facsimile standards. The most common ones are *Group 3* [123] and *Group 4* [124] standards. Group 3 includes actually two distinct compression algorithms, known as Modified Huffman (MH) *coding* and Modified READ (Relative Element Address Designate) *coding* (MR) modes. The algorithm of Group 4 standard is commonly called Modified Modified READ (MMR) coding.

MH coding is a one-dimensional coding scheme that encodes each row of pixels in an image independently. It uses run-length coding with a static Huffman coder. It reduces the size of the Huffman code table by using a prefix markup code for runs over 63 and thus accounts for the term *modified* in the name. Each scan line ends with a unique EOL (end of line) code and it doubles as an error recovery code.

To exploit the fact that most transitions in bi-level facsimile images occur at one pixel to the right or left or directly below a transition on the line above, MR coding uses a 2-dimensional reference (Figure 18) with

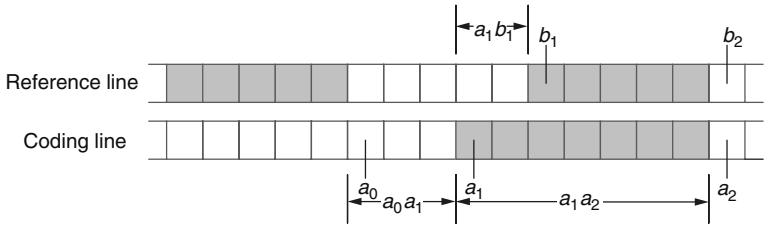


Figure 18. Reference point and lengths used during modified READ (MR) encoding.

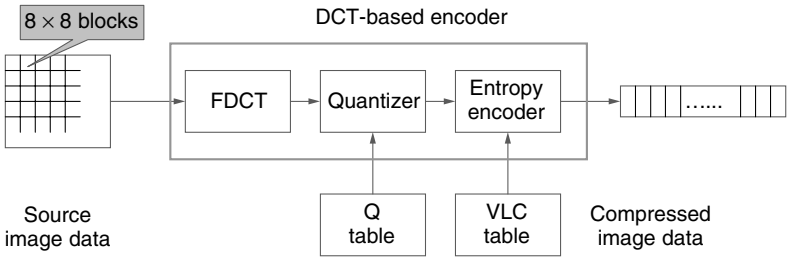


Figure 19. JPEG encoder block diagram.

run-length coding and a static Huffman coder. The name *Relative Element Address Designate* (*READ*) is based on the fact that only the relative positions of transitions are encoded. The most efficient MR coding mode is vertical mode (*interline prediction*) where only the small difference between the same transition positions in two adjacent lines is encoded. It also includes EOL codewords and periodically includes MH-coded lines (*intra-lines*) to minimize the effect of errors. The MR coding scheme, though simple, provides some very important concepts for modern image and video coding. We will see later in this section, it has much in common to the intra- and inter video coding schemes.

MMR coding in the Group 4 facsimile standard is based on MR coding in Group 3. It modifies the MR algorithm to maximize compression by removing the MR error prevention mechanisms. There is no EOL symbols and MH coding in Group 4. To enable the MR coding in Group 4, when coding the first line, a virtual white line is used as the reference. The transmission errors are corrected with lower level control procedures.

4.1.2. JBIG. Group 3 and Group 4 fax coding has proven adequate for text-based documents, but does not provide good compression or quality for documents with handwritten text or continuous tone images. As a consequence, a new set of fax standards, such as JBIG and the more recent JBIG2, has been created since the late 1980s.

JBIG stands for Joint Bi-level Image experts Group coding: The algorithm is defined in CCITT Recommendation T.82 [125,126] and it is a lossless bi-level image coding standard based on an adaptive arithmetic coding with adaptive 2D contexts.

Besides more efficient compression, JBIG provides progressive transmission of image through multi-resolution coding. As a new emerging standard for bi-level image coding, JBIG2 [127] allows both lossy and lossless bi-level image compression. It is the first international standard that provides lossy compression of bi-level images to achieve much higher compression ratio with almost no quality degradation. Besides higher compression performance, JBIG2 allows both quality-progressive coding from lower to higher (or lossless) quality, and content-progressive coding, successively adding different types of image data (e.g., first text, then halftones). The key technology in JBIG2 is pattern matching predictive coding schemes where a library of dynamically built templates is used to predict the repetitive character-based pixel blocks in a document.

4.1.3. JPEG. This is probably the best known ISO/ITU-T standard created in the late 1980s. The JPEG (Joint Photographic Experts Group) is a DCT-based standard that specifies three lossy encoding modes, namely, sequential, progressive, and hierarchical, and one lossless encoding mode. The baseline JPEG coder [122] is the sequential encoding in its simplest form. Baseline mode is the most popular one and it supports lossy coding only. Figure 19 shows the key processing steps in a baseline JPEG coder. It is based on the 8×8 block DCT, uniform scalar quantization with a perceptual weighting matrix, zigzag scanning of AC components, predictive coding of DC components, and Huffman coding (or arithmetic coding with more complexity but better performance).

The progressive and hierarchical modes of JPEG are both lossy and differ only in the way the DCT coefficients are coded or computed, when compared to the baseline mode. They allow a reconstruction of a lower quality or lower resolution version of the image, respectively, by partial decoding of the compressed bit stream. Progressive mode encodes the quantized coefficients by a mixture of spectral selection and successive approximation, while the hierarchical mode utilizes a pyramidal approach to computing the DCT coefficients in a multi-resolution way.

The lossless mode or lossless JPEG is a lossless coding scheme for continuous-tone image. It is an adaptive prediction coding scheme with a few predictors to choose from, which is similar to the DC component coding in the baseline mode but operated on the image domain pixel instead. The prediction difference is efficiently encoded with Huffman coding or arithmetic coding. It should be noted that lossless JPEG is not DCT-based as in the lossy modes. It is a pure pixel domain predictive coding.

4.1.4. JPEG-LS. Not to be confused with the lossless mode of JPEG (*Lossless JPEG*), JPEG-LS [131] is the latest and totally different ISO/ITU-T standard for lossless coding of still images and which also provides for *near-lossless* coding. The baseline system is based on the LOCO-I algorithm (LOw COmplexity LOssless COmpression for Images) [132] developed at Hewlett-Packard Laboratories.

LOCO-I combines the simplicity of Huffman coding with the compression potential of context models. The algorithm uses a nonlinear predictor with rudimentary edge detecting capability, and is based on a very simple context model, determined by quantized gradients. A small number of free statistical parameters are used to capture high-order dependencies, without the drawbacks of context dilution. The prediction residues are modeled by a *double-sided geometry distribution* with two parameters that

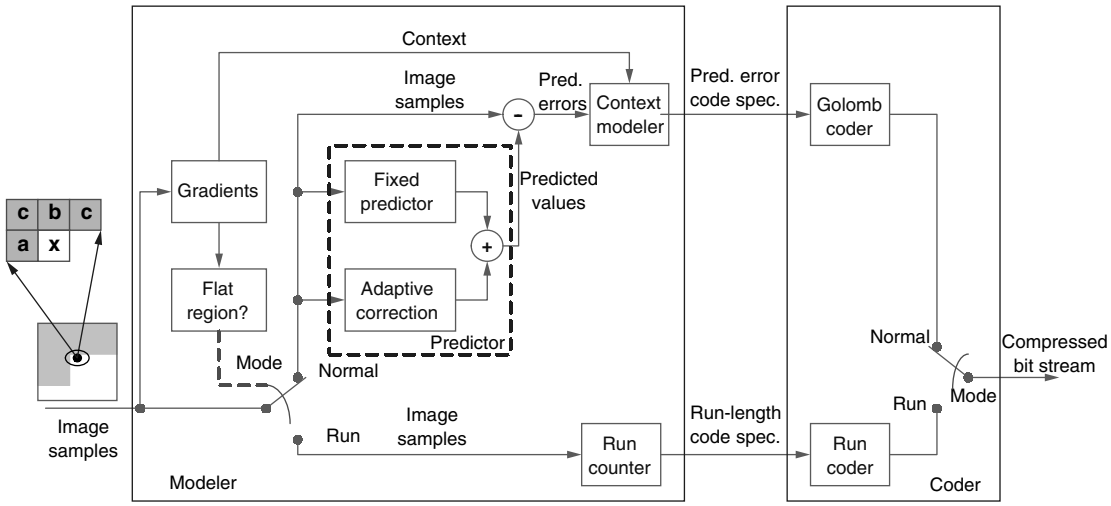


Figure 20. JPEG-LS encoder block diagram.

can be updated symbol by symbol based on the simple context, and in turn are efficiently encoded by a simple and adaptive Golomb code [133,134] that corresponds to Huffman coding for a geometric distribution. In addition, a run coding mode is defined for low-entropy flat area, where runs of identical symbols are encoded using extended Golomb coding with improved performance and adaptability. Figure 20 shows the block diagram of the JPEG-LS encoder. This algorithm was designed for low-complexity while providing high compression. However, it does not provide for scalability, error resilience, or other additional functionality.

4.1.5. Visual Texture Coding (VTC) in MPEG-4. Visual Texture Coding (VTC) [135] is the algorithm in the MPEG-4 standard (see Section 4.2.5) used to compress the texture information in photo realistic 3D models as well as still images. It is based on the discrete wavelet transform (DWT), scalar quantization, zerotree coding, and context-based arithmetic coding. Different quantization strategies are used to provide different SNR scalability: single quantization step (SQ) provides no SNR scalability; multiple quantization steps (MQ) provides discrete (coarse-grain) SNR scalability and bi-level quantization (BQ) supports fine grain SNR scalability at the bit level. In addition to the traditional tree-depth (TD) zerotree scanning similar to the EZW algorithm, band-by-band (BB) scanning is also used in MPEG-4 VTC to support resolution scalability. MPEG-4 VTC also supports coding of arbitrarily shaped objects, by the means of a shape adaptive DWT [36,150], but does not support lossless texture coding. Besides, a resolution scalable lossless shape coding algorithm [136] that matches the shape adaptive wavelet decomposition at different scales is also adopted. The scalable shape coding scheme uses fixed contexts and fixed probability tables that are suitable for the bi-level shape masks with large continuous regions of identical pixel values.

4.1.6. JPEG 2000. JPEG 2000 [86,137] is the latest emerging standard from the Joint Photographic Experts Group that is designed for different types of still images allowing different imaging models within a unified system. JPEG 2000 is intended to complement, not replace, the current JPEG standards and it has two coding modes: a *DCT-based coding* mode that uses currently baseline JPEG and a *wavelet-based coding* mode. JPEG 2000 normally refers to the wavelet-based coding mode and it is based on the discrete wavelet transform (DWT), scalar quantization, context modeling, arithmetic coding, and post-compression rate allocation techniques.

This core compression algorithm in JPEG 2000 is based on independent Embedded Block Coding with Optimized Truncation (EBCOT) of the embedded bit streams [91]. The EBCOT algorithm uses a wavelet transform to generate the subband coefficients, where the DWT can be performed with reversible filters for lossless coding, or nonreversible filters for lossy coding with higher compression efficiency. EBCOT partitions each subband into relatively small blocks of samples (called codeblocks) and encodes them independently. A multi-pass bitplane coding scheme based on context-based arithmetic coding is used to code the original or quantized coefficients in each codeblock into an embedded bit stream in the order of importance along with the rate-distortion pairs at each pass of the fractional bitplane (*truncation point*). It seems that failing to exploit the inter-subband redundancy would have a sizable adverse effect on coding efficiency. However, this is more than compensated by the finer scalability that results from the multi-pass coding.

JPEG 2000 also supports many new features, such as, compression of large images, single decompression architecture, compound documents, static and dynamic region-of-interest (ROI), multiple component images, content-based description, and protective image security.

4.2. Video Coding Standards

Most practical video coding systems including all international video coding standards are based on a hybrid coder that combines motion compensation in the temporal direction and DCT transform coding within each independent

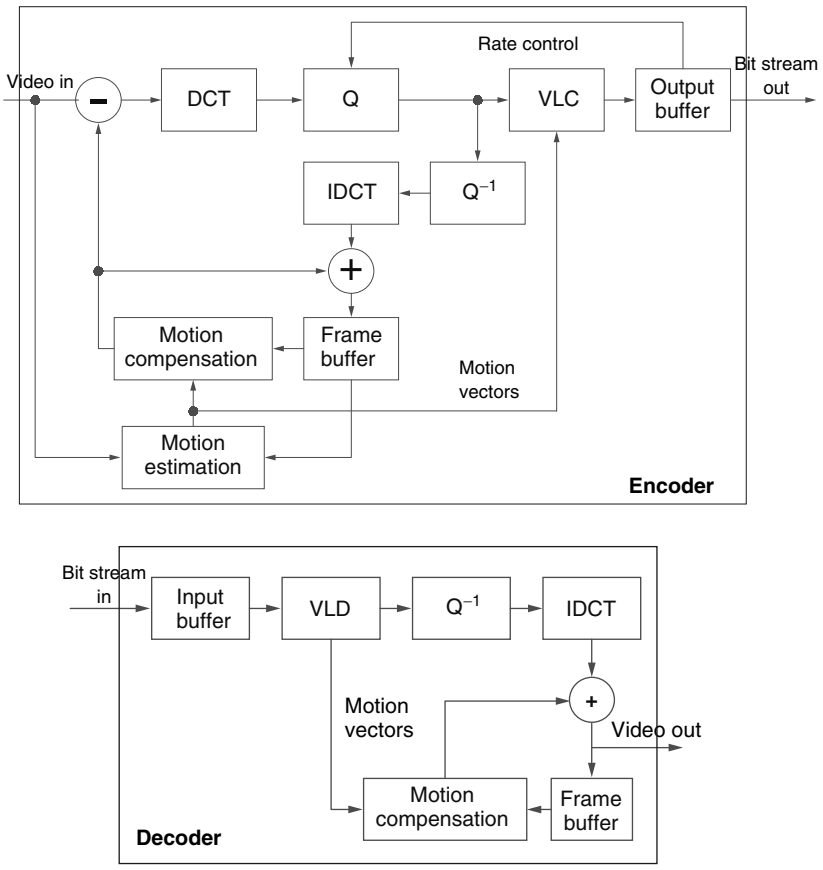


Figure 21. A generic DCT-based predictive video coder.

frame or prediction frame. A generic DCT-based predictive video coding system is illustrated in Fig. 21.

For independent frames, a DCT transform coder similar to the baseline JPEG is applied to compress the frame without using any information from other frames (referred to as intracoding). For prediction frames, a compensation module first generates a predicted image from one or more previously coded frames based on the motion vectors estimated and subtract it from the original frame to obtain the motion predicted residue image. Another DCT transform coder that fits the characteristics of the residue image is then applied to further exploit the spatial correlation to efficiently compress the residue image. In order to avoid the drifting errors caused by the mismatch between reference frames, the encoder normally embeds the same decoding structure to reconstruct exactly the same reference frames as in the decoder. Optionally, a loop filter to smooth out the blocking artifacts or generate a better reference image may be inserted in the reconstruction loop to generate a visually better image and enhance the prediction efficiency.

There have been several major initiatives in video coding that have led to a range of video standards: ITU video coding for video teleconferencing standards, H.261 for ISDN, H.262 (same as MPEG-2) for ATM/broadband, and H.263 for POTS (Plain Old telephone Service); ISO video coding standards MPEG-1 (Moving Picture Experts Group), MPEG-2 and MPEG-4; and the emerging standards such as H.264 | MPEG-4 Part 10 by JVT (Joint Video Team of ITU and ISO). These coding systems all belong to the DCT-based motion predictive coder category. In the following sections, we provide brief summaries of the video coders with emphasis on the differences of the coding algorithms.

4.2.1. H.261. The H.261 video codec [138] (1990), initially intended for ISDN teleconferencing, is the baseline video mode for most multimedia conferencing systems.

The basic coding unit in H.261 is *macroblock* that contains one 16×16 or four 8×8 luminance blocks and two 8×8 chrominance blocks. The H.261 codec encodes video frames using an 8×8 DCT. An initial frame (called an I or *intra* frame) is coded and transmitted as an independent frame. Subsequent frames are efficiently coded in the inter mode (P or *predictive* frame) using motion compensated predictive coding described above, where motion compensation is based on a 16×16 pixel block with integer pixel motion vectors and always referenced to the immediately previous frame. The DCT coefficients are quantized with a uniform quantizer and arranged in a zigzag scanning order. Run-length coding combined with variable length coding is used to compress the quantized coefficients into a video bit stream. An optional loop filter is introduced after motion compensation to improve the motion prediction efficiency.

H.261 is intended for head-and-shoulders type of scene in video conferencing applications where only small, controlled amounts of motion are present so the motion vector range can be limited. The supported video formats include both the CIF (Common Interchange Format with

a resolution of 352×288 and YCbCr $4:2:0$) and the QCIF (*quarter CIF*) format. All H.261 video is noninterlaced, using a simple progressive scanning pattern.

4.2.2. MPEG-1. The MPEG-1 standard [139] (1993) is a true multimedia standard that contains specifications for audio coding, video coding, and systems. MPEG-1 was intended for storage of multimedia content on a standard CD-ROM, with data rates of up to 1.5 Mbits/s and a storage capacity of about 600 Mbytes. Noninterlaced CIF video format (352×288 at 25 fps or 352×240 at 30 fps) is used to provide VHS-like video quality.

The video coding in MPEG-1 is very similar to the video coding of the H.261 described above with the difference that the uniform quantization is now based on perceptual weighting criteria. The temporal coding was based on both uni- and bi-directional motion-compensated prediction. A new *B* or bi-directionally predictive picture type is introduced in MPEG-1 which can be coded based on either the next and/or the previous I or P pictures. In contrast, an *I* or *intra* picture is encoded independently of all previous or future pictures and a *P* or *predictive* picture is coded based on only a previous I or P picture. MPEG-1 also allows half-pel motion vector precision to improve the prediction accuracy. There is no loop filter present in an MPEG-1 motion compensation loop.

4.2.3. MPEG-2. The MPEG-2 standard [140] (1995) was initially developed primarily for coding interlaced video at 4-9 Mbits/s for broadcast TV and high quality digital storage media (such as DVD video); it has now also been used in HDTV, cable/satellite TV, video services over broadband networks, and high quality video conferencing (same as H.262). The MPEG-2 standard includes video coding, audio coding, system format for program and transport streams, and other information related to practical implementations.

MPEG-2 video supports both interlaced and progressive video, multiple color format ($4:2:0$, $4:2:2$ and $4:4:4$), flexible picture size and frame rates, hierarchical or scalable video coding, and is backward compatible with MPEG-1. To best satisfy the needs of different applications, different profiles (subset of the entire admissible bit stream syntax) and levels (set of constraints imposed on the parameters of the bitstreams within a profile) are also defined in MPEG-2.

The most distinguishing feature of MPEG-2 from previous standards is the support for interlaced video coding. For interlaced video, it can be either encoded as a frame picture or a field picture, with adaptive frame or field DCT and frame or field motion compensation at macroblock-level. MPEG-2 also offer another new feature in providing the temporal, spatial, and SNR scalabilities. Temporal scalability is achieved with B-frame coding; spatial scalability is obtained by encoding the prediction error with a reference formed from both an upsampled low-resolution current frame and a high-resolution previous frame; SNR scalability is provided by finely quantizing the error residue from the coarsely quantized low-quality layer (there is a drifting problem).

Compared with MPEG-1, a number of improvements have been made to further improve the coding efficiency, including, a more flexible coding mode selection at the macroblock level; a nonlinear quantization table with increased accuracy for small values; an alternative zigzag scan for DCT coefficients especially for the interlaced video coding; much increased permissible motion vector range; new VLC tables for the increase bit rate range; customized perceptual quantization matrix support; and dual prime prediction for interlaced video encoded as P pictures, which mimics the B picture prediction especially for low-delay applications.

MPEG-2 also introduces some error resilience tools such as independent slice structure, data partitioning to separate data with different importance, concealment motion vectors in intra-pictures, and different scalabilities as we discussed.

MPEG-2 is probably the most widely used video coding standard so far that enables many applications, such as DVD, HDTV, and digital satellite and cable TV broadcast.

4.2.4. H.263, H.263+, and H.263++. The H.263 standard [141] (1995) was intended for use in POTS conferencing and the video codec is based on the same DCT and motion compensation techniques as used in H.261. H.263 now supports 5 picture formats including sub-QCIF (126×96) QCIF (176×144), CIF (352×288), 4CIF (704×576) and 16CIF (1408×1152). Several major differences exist between H.263 and H.261 including, more accurate half-pel motion compensation in H.263 but integer-pel in H.261; no loop filter in H.263 but optional in H.261 since the optional overlapped block motion compensation (OBMC) in H.263 could have the similar de-blocking filter effect; motion vector predictor in H.263 using the median value of three candidate vectors while only the preceding one being used in H.261; 3-dimensional run length coder (*last, run, level*) in H.263 versus (*run, level*) and an *eob* symbol in H.261.

As H.263 version 2, H.263+ [142] (1998) provides optional extensions to baseline H.263. These options broaden the range of useful applications and improve the compression performance. With these new options, H.263+ supports custom source format with different picture size, aspect ratio, and clock frequency. It also provides enhanced error-resilience capability with slice structured coding, prediction reference selection, motion vector coding with reversible VLC (RVLC), where RVLCs are codewords which can be decoded in a forward as well as a backward manner, and independent segment decoding. Backward-compatible supplemental enhancement information can be embedded into the video bit stream to assist operations in the receiver. In addition, temporal scalability can be provided by B picture, while SNR and spatial scalabilities are enabled by newly defined *EI* pictures (referencing to a temporally simultaneous picture) and *EP* pictures (referencing to both a temporally preceding picture and a temporally simultaneous one). Figure 22 gives an example of mixed scalabilities enabled by these new picture types.

H.263++ [143] (H.263 version 3, 2001) provides more additional options for H.263, including, *Enhanced Reference Picture Selection* to provide enhanced coding efficiency and enhanced error resilience (particularly

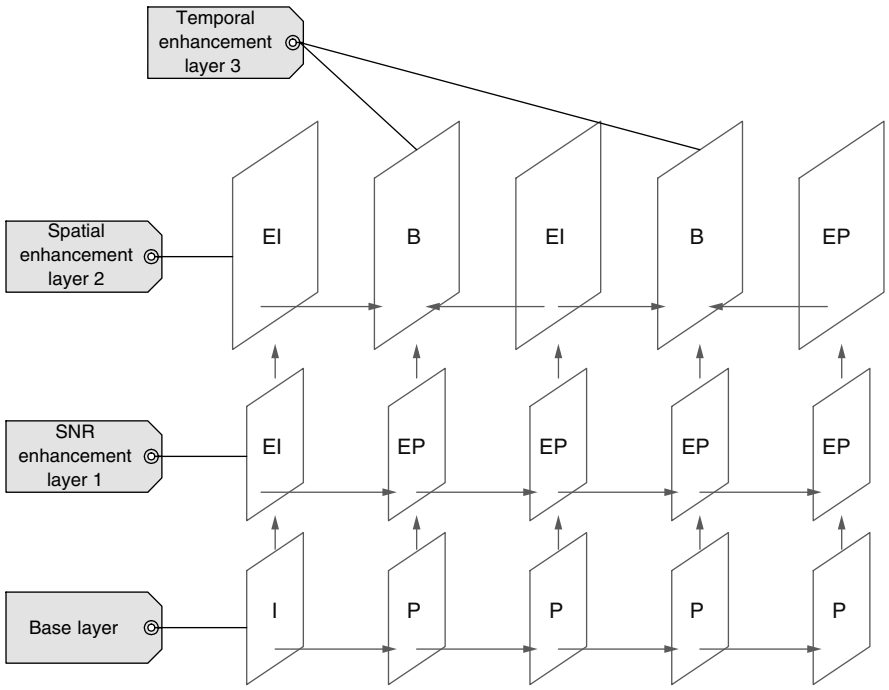


Figure 22. Example for mixed temporal, SNR and spatial scalabilities supported in H.263+.

against packet losses) through the use of multiple reference pictures; *Data Partitioned Slice* to enhance error resilience (particularly against localized corruption of bit stream) by separating header and motion vectors from DCT coefficients in the bit stream and protecting motion vectors using a reversibly decodable variable length coder (RVLC); and additional *Supplemental Enhancement Information* including support for interlaced video.

4.2.5. MPEG-4. Coding of separate audio–visual objects, both natural and synthetic, leads to the latest ISO MPEG-4 standard [144] (1998). The MPEG-4 visual standard is developed to provide users a new level of interaction with visual content. It provides technologies to view, access and manipulate objects rather than pixels, with great error robustness at a large range of bit rates. Application areas range from digital television, streaming video, to mobile multimedia, 2D/3D games, and visual communications.

The MPEG-4 visual standard consists of a set of tools supporting mainly three classes of functionalities: compression efficiency, content-based interactivity, and universal access. Among them, content-based interactivity is one of the most important novelties offered by MPEG-4. To support object-based functionality, the basic compression unit in MPEG-4 can be arbitrarily shaped video object plane (VOP) instead of always the rectangular frame. Each object can be encoded with different parameters, and at different qualities. Tools provided in the MPEG-4 standard include shape coding, motion estimation and compensation, texture coding, error resilience, sprite coding, and scalability. Conformance points, in the form of object types, profiles, and levels, provide the basis for interoperability. MPEG-4 provides support for both interlaced and progressive material. The chrominance format that is supported is $4:2:0$.

For reasons of efficiency and backward compatibility, video objects are coded in a hybrid coding scheme somewhat similar to previous MPEG standards. Figure 23 outlines the basic approach of the MPEG-4 video algorithms to encode rectangular as well as arbitrarily-shaped input image sequences. Compared with traditional rectangular video coding, there is an additional shape coding module in the encoder that compresses the shape information necessary for the decoding of a video object. Shape information is also used to control the behavior of the DCT transform, motion estimation and compensation, and residue coding.

Shape coding can be performed in binary mode, where the shape of each object is described by a binary mask, or in grayscale mode, where the shape is described in a form similar to an alpha channel, allowing transparency, and reducing aliasing. The binary shape can be losslessly or lossy coded using context-based arithmetic encoding (CAE) based on the context either from the current video object (intraCAE) or from the motion predicted video object (interCAE). By using binary shape coding for coding its support region, grayscale shape information is lossy encoded using a block based motion compensated DCT similar to that of texture coding.

Texture coding is based on an 8×8 DCT, with appropriate modifications for object boundary blocks for both intramacroblocks and intermacroblocks. Low-pass extrapolation (LPE) (also known mean-repetitive padding) is used for intra boundary blocks and zero padding is used for inter boundary blocks before DCT transforms. Furthermore, shape adaptive DCT (SA-DCT) can also be used to decompose the arbitrarily shaped boundary blocks to achieve more coding efficiency. The DCT coefficients are quantized with or without perceptual quantization matrices. MPEG-4 also allows for a nonlinear quantization of DC values. To further reduce the average energy of the

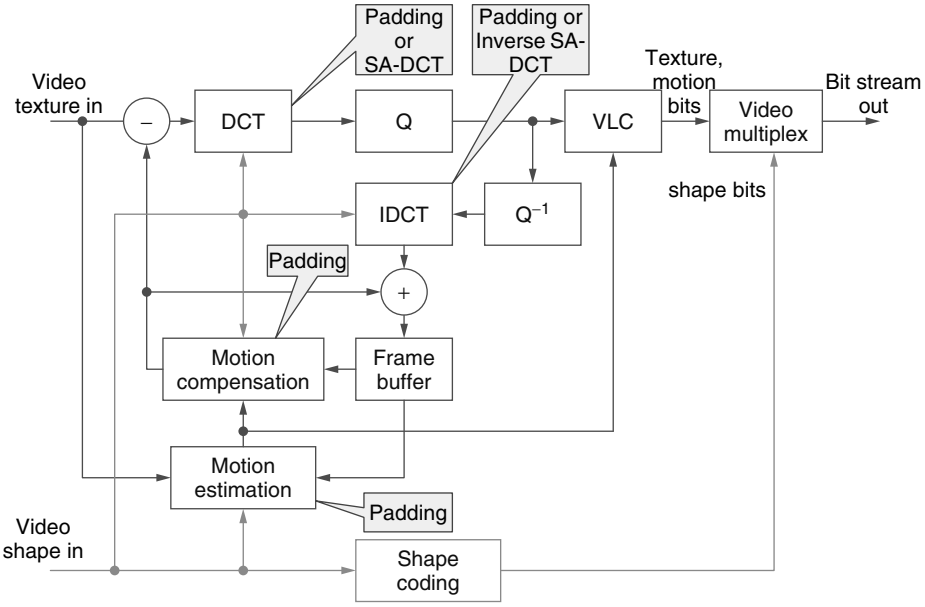


Figure 23. Block diagram of a MPEG-4 video coder.

quantized coefficients, the adaptive prediction of DC and AC coefficients from their neighboring blocks can be used to improve the coding efficiency. Three types of zigzag scans are allowed to reorder the quantized coefficients according to the DC prediction types. Two VLC tables switched by the quantization level can be used for the run length coding.

Still textures with possible arbitrary shapes can be encoded using wavelet transforms as described before with the difference that the shape-adaptive discrete wavelet transform (SA-DWT) and shape adaptive zerotree coding are used to extend the still texture coding to arbitrarily shaped image objects.

Motion compensation is block based, with appropriate modifications for object boundaries through repetitive padding and "polygon matching." Motion vectors are predictively encoded. The block size can be 16×16, or 8×8, with up to quarter pixel precision. MPEG-4 also provides modes for overlapped block motion compensation (OMBC) to reduce blocking artifacts and get better prediction quality at lower bit rates, and global motion compensation (GMC) through image warping. Moreover, static sprite can also be encoded efficiently with only 8 global motion parameters describing camera motion that represent the appropriate affine transform of the sprite.

When the video content to be coded is interlaced, additional coding efficiency can be achieved by adaptively switching between field and frame coding.

Error resilience functionality is important for universal access through error-prone environments, such as mobile communications. It offers means for resynchronization, error detection, data recovery, and error concealment. Error resilience in MPEG-4 is provided by resynchronization markers to resume decoding from errors; data partitioning to separate information with different importance; header extension codes to insert redundant important header information in the bit stream; and reversible variable length codes (RVLC) to decode portions of the corrupted bit stream in the reverse order.

MPEG-4 also includes tool for sprite coding where a sprite consists of those regions present in the scene throughout the video segment, for example, the background. Sprite-based coding is very suitable for synthetic objects as well as objects in natural scenes with rigid motion, where sprites can provide high compression efficiency by appropriate warping/cropping operations. Shape and texture information for a sprite is encoded as an intra VOP.

Scalability is provided for object, SNR, spatial and temporal resolution enhancement in MPEG-4. Object scalability is inherently supported by the object based coding scheme. SNR scalability is provided by fine granularity scalability (FGS) where the residue of the base layer data is further encoded as an out-of-loop enhancement layer. Spatial scalability is achieved by referencing both the temporally simultaneous base layer and the temporally preceding spatial enhancement layer. Temporal scalability is provided by using bi-directional coding. One of the most advantageous features of MPEG-4 temporal and spatial scalabilities is that they can be applied to both rectangular frames and arbitrarily-shaped video objects.

4.2.6. The Emerging MPEG-4 Part-10/ITU H.264 Standard. At the time of completing this manuscript, there is a very active new standardization effort in developing a video coding standard that can achieve substantially higher coding efficiency compared to what could be achieved using any of the existing video coding standards. This is the emerging MPEG-4 Part-10/ITU H.264 standard jointly developed by a Joint Video Team (JVT) [145]. In MPEG, this new standard is called the Advanced Video Coding (AVC) standard.

The new standard takes a detailed look at many fundamental issues in video coding. The underlying coding structure of the new standard is still a block-based motion compensated transform coder similar to that adopted in

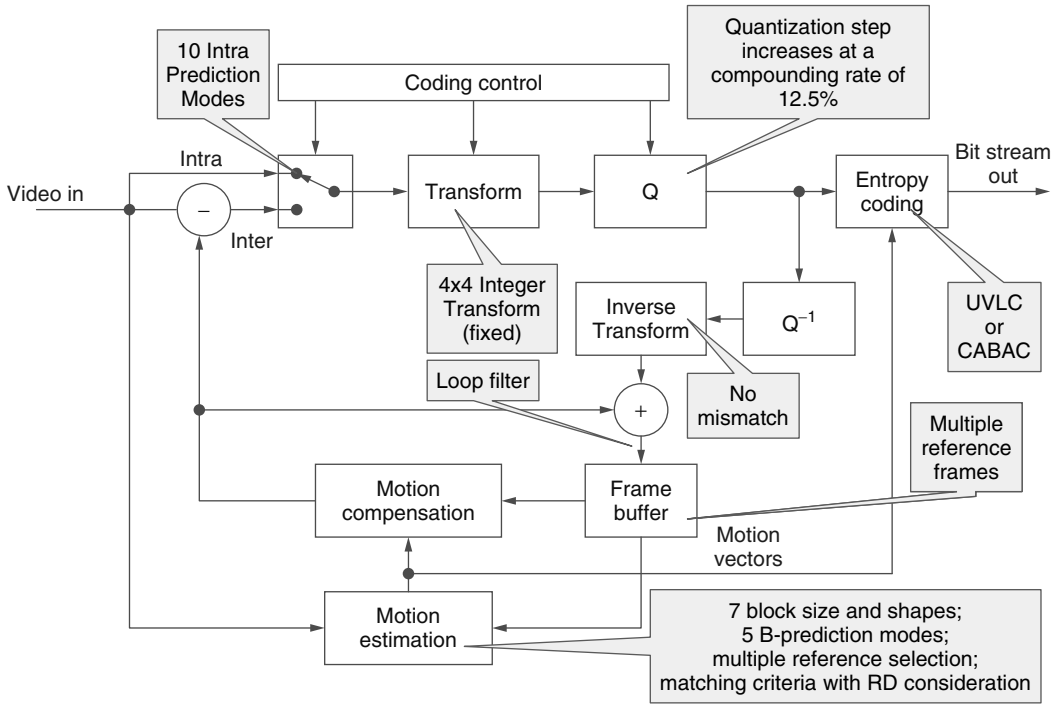


Figure 24. Block diagram of the AVC encoder.

previous standards. The block diagram of an AVC encoder is given in Fig. 24.

There are several significant advances in AVC that provides high coding efficiency. The key feature that enables such high coding efficiency is adaptivity. For intracoding, the flexible intra prediction is used to exploit spatial correlation between adjacent macroblocks and adapt to the local characteristics. AVC uses a simpler and integer-based 4×4 block transform that does not have a mismatch problem as in the 8×8 DCT due to different implementation precisions dealing with the real-valued DCT basis function. The smaller transform size also helps to reduce blocking and ringing artifacts.

Most of the coding gain achieved in AVC is obtained through advanced motion compensation. Motion compensation in AVC supports most of the key features found in earlier video standards, but its efficiency is improved through added flexibility and functionality. For motion estimation/compensation, AVC uses blocks of different sizes and shapes that provide capability to handle fine motion details, reduce high frequency components in the prediction residue and avoid large blocking artifacts. Using seven different block sizes and shapes can translate into bit savings of more than 15% as compared to using only a 16×16 block size. Higher precision sub-pel motion compensation is supported in AVC to improve the motion prediction efficiency. Using 1/4-pel spatial accuracy can yield more than 20% in bit savings as compared to using integer-pel spatial accuracy. AVC offers an option of multiple reference frame selection that can improve both the coding efficiency and error-resilience capability. Using five reference frames for prediction can yield 5-10% in bit savings as compared to using only one reference frame. The use of adaptive de-blocking filters in the prediction loop substantially reduces the blocking artifacts, with additional complexity. Because bit savings depend on video sequences and bit rates, the above numbers are only meant to give a rough idea about how much improvement one may expect from a coding tool.

In AVC, the quantization step sizes are increased at a compounding rate of approximately 12.5%, rather than increasing it by a constant increment, to provide finer quantization yet covering a greater bit rate range.

In AVC, entropy coding is performed using either variable length codes (VLC) or using context-based adaptive binary arithmetic coding (CABAC). The use of CABAC in AVC yields a consistent improvement in bit savings. AVC is still under development and new technologies continue to be adopted into the draft standard to further improve the coding efficiency. To provide a high level picture of the various video coding standards, Fig. 25 shows the approximate timeline and evolution of different standards.

5. FUTURE RESEARCH DIRECTIONS IN IMAGE AND VIDEO CODING

Further improving the coding efficiency of image and video compression algorithms continues to be the goal that researchers will be pursuing for a long time. From the success of the JPEG2000 and MPEG-4 AVC codecs, we learned that better schemes to enable rate-distortion optimization (normally generating a variable bit rate bit stream) and content-adaptivity are two areas worth further investigation. Adaptivity is the most important feature of AVC, block-sizes, intra prediction modes, motion

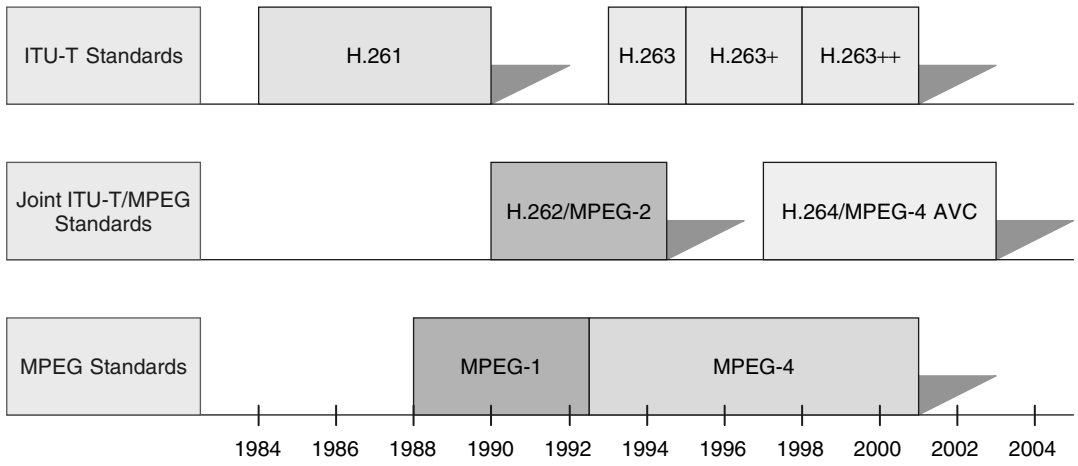


Figure 25. Progression of the ITU-T Recommendations and MPEG standards.

prediction modes, multiple reference frames, adaptive entropy coding, adaptive loop filtering, and so on, all reflect the significance of adaptivity. The main reason is that image and video signals by nature are not stationary signals, although traditional source coding theories all assume them as statistically stationary sources.

Besides coding efficiency, there are also increasing demands for more special functionalities, such as scalability, error-resilience, transcoding, object-based coding, model-based coding, and hybrid natural and synthetic coding, and so forth. These functionalities may require unconventional image and video coding techniques and may need to balance the trade-off between coding efficiency and functionalities.

Scalable coding refers to coding an image or a video sequence into a bit stream that is partially decodable to produce a reconstructed image or video sequence at possibly a lower spatial, temporal, or quantization resolution. Figure 26 shows the relationship between different scalabilities for a video bit stream. The principal concept in scalability is that an improvement in source reproduction can be achieved by sending only an incremental amount of bits over the current transmitted rate. Quantization, spatial, and temporal scalabilities are all important in applications. A key feature of scalable coding is that the encoding only needs to be done once and a single scalable bit stream will fit in all applications with different target bit rates, or spatial, or temporal resolutions. Scalable coding can bring many benefits in practical applications: potential high coding efficiency through rate-distortion optimization (e.g., JPEG-2000); easy adaptation to heterogeneous channels and user device capabilities and support for multicast; better received picture quality over dynamic changing networks through channel adaptation; robust delivery over error-prone, packet-loss-ridden through unequal error protection (UEP) and priority streaming because some enhancement bit stream can be lost or discarded without irreparable harm to the overall representation; reduced storage space required to store multiple bitstreams otherwise; no complicated transcoding needed for different bit rates and resolutions. There are two types of scalabilities, layered scalability and fine granularity scalability (FGS). Layered scalability only provides a limited number of ("quantized") layers of scalability (e.g., SNR and spatial scalabilities in MPEG-2 and temporal scalability in H.263) in which a layer is either completely useful or completely useless, while FGS provides a continuous enhancement layer that can be truncated at any bit location, and the reconstructed image or video quality is proportional to the number of decoded bits (e.g., JPEG-2000, MPEG-4 VTC, MPEG-4 FGS, 3-D wavelet video coding). Error-robust coding is necessary for error-prone channels to increase the capability of error recovery and error concealment. Error-robust features built in source coding can potentially provide better and more intelligent error resilience capability than "blind" error control channel coding. Resynchronization, data partitioning, multi-description coding are all source coding techniques aiming at error-robustness of video coding. Compared with scalable coding, error resilience coding

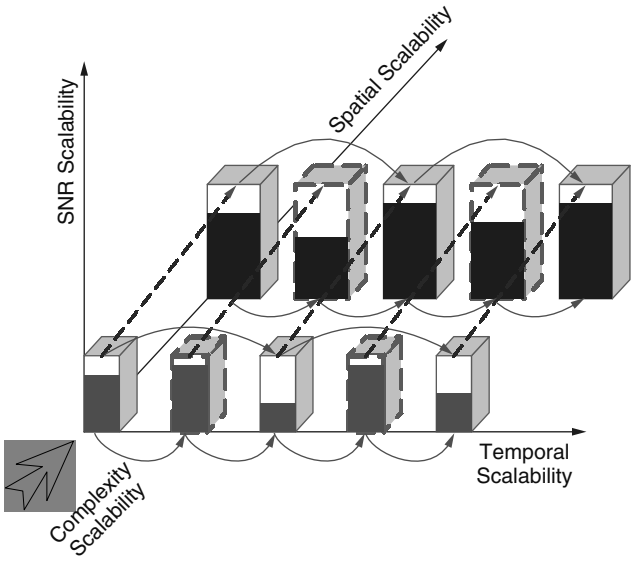


Figure 26. Relationship between different scalabilities.

may be considered as repairing damaged bitstreams while scalable coding is trying to prevent bitstreams from being damaged. For a given transmission bit rate, splitting bits between source coding and channel coding is a bit tricky. If bits are allocated to channel coding and the channel is ideal, there is a loss in performance compared to source coding alone. Similarly, if there are no bits allocated to channel coding and the channel is very noisy, there will be a loss in performance compared to using some error control coding. How to optimally allocate these bits between source coding and channel coding brings us to the problem of joint source and channel coding.

Transcoding is usually considered as converting one compressed bit stream into another compressed bit stream. There are several different reasons for using transcoding. One of them is to change the bit rate. Another reason is to convert the bit stream format from one standard to another. The challenge is to have the coding efficiency of the transcoded bit stream close to that of an encoded bit stream from the original image and video. Another challenge is to have the transcoding operation as simple as possible without a complete re-encoding process.

Object-based coding is required for object-based interactivity and manipulation. In MPEG-4 video coding standard, object-based coding is defined for many novel applications to provide content-based interactivity, content-based media search and indexing, and content scalable media delivery, and so on. Figure 27 illustrates the concept of an object-based scenario. One of the challenges in object-based coding is the preprocessing part that separates an image or video object from a normal scene. Good progresses have been reported in this direction [151–154].

For some visual objects, it is possible to establish a 2D or 3D structural model with which the movement of different parts of the object can be parameterized and transmitted. One of such successful visual models is the human face and body. In MPEG-4 visual coding standard, face and body animation is included. Using such a model-based coding, it is possible to transmit visual information at extremely low bit rates. The challenges are to establish good models for different visual objects and to accurately extract the model parameters.

Image and video coding techniques mainly deal with natural images and video. Computer graphics generate visual scenes as realistic as possible. Hybrid natural and synthetic coding is to combine these two areas. Usually, it is relatively easy for computer graphics to generate a structure of a visual object, but harder to generate realistic texture. With hybrid natural and synthetic coding, natural texture mapping can be used to make animated objects look more realistic. Latest developments in computer graphics/vision tend to use naturally captured image and video contents to render realistic environment (called image based rendering or IBR), such as lumigraph [155], light-fields [156], concentric mosaics [157], and so on. How to efficiently encode these multidimensional cross-correlated image and video contents is an active research topic [158].

In summary, image and video coding research has been a very active field with a significant impact to industry and society. It will continue to be active with many more interesting problems to be explored.

BIOGRAPHIES

Shipeng Li joined Microsoft Research Asia in May 1999. He is currently the research manager of Internet Media group. From October 1996 to May 1999, he was with Multimedia Technology Laboratory at Sarnoff Corporation in Princeton, NJ (formerly David Sarnoff Research Center, and RCA Laboratories) as a member of technical staff. He has been actively involved in research and development of digital television, MPEG, JPEG, image/video compression, next generation multimedia standards and applications,

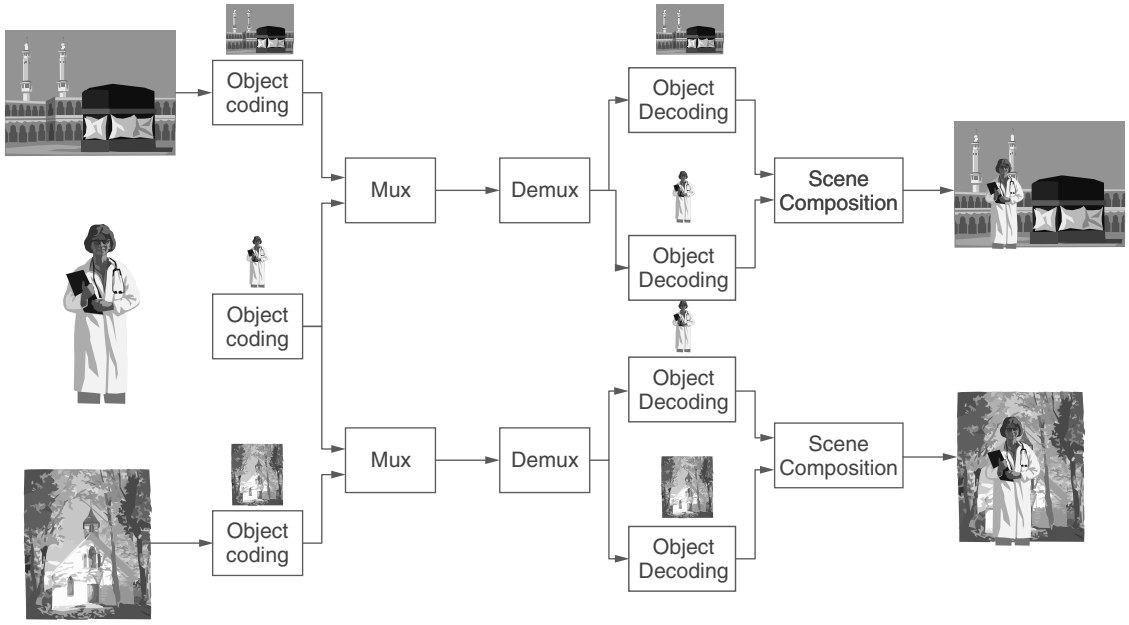


Figure 27. An object-based coding application scenario.

and consumer electronics. He made contributions in shape-adaptive wavelet transforms, scalable shape coding, and the error resilience tool in the FGS profile for the MPEG-4 standard. He has authored and coauthored more than 70 technical publications and more than 20 grants and pending U.S. patents in image/video compression and communications, digital television, multimedia, and wireless communication. He is the coauthor of a chapter in *Multimedia Systems and Standards* published by the Marcel Dekker, Inc. He is a member of Visual Signal Processing and Communications committee of IEEE Circuits and Systems Society. He received his B.S. and M.S. both in Electrical Engineering from the University of Science and Technology of China (USTC) in 1988 and 1991, respectively. He received his Ph.D. in Electrical Engineering from Lehigh University, Bethlehem, PA, in 1996. He was an assistant professor in Electrical Engineering department at University of Science and Technology of China in 1991–1992.

Weiping Li received his B.S. degree from University of Science and Technology of China (USTC) in 1982, and his M.S. and Ph.D. degrees from Stanford University in 1983 and 1988 respectively, all in electrical engineering. Since 1987, he has been a faculty member in the department of Electrical and Computer Engineering at Lehigh University. From 1998 to 1999, he was with Optivision, Inc., in Palo Alto, California, as the Director of R&D. He has been with WebCast Technologies, Inc., since 2000. Weiping Li has been elected a Fellow of IEEE for contributions to image and video coding algorithms, standards, and implementations. He served as the Editor-in-Chief of IEEE Transactions on Circuits and Systems for Video Technology from 1999 to 2001. He has served as an Editor for the Streaming Video Profile Amendment of MPEG-4 International Standard. He is a member of the Board of Directors for MPEG-4 Industry Forum. He served as one of the Guest Editors for a special issue of IEEE Proceedings on image and video compression (February 1995). Weiping Li was awarded Guest Professorships in University of Science and Technology of China and in Huazhong University of Science and Technology in 2001. He received the Spira Award for Excellence in Teaching in 1992 at Lehigh University and the Guo Mo-Ruo Prize for Outstanding Student in 1980 at University of Science and Technology of China.

BIBLIOGRAPHY

1. M. B. Priestley, *Spectral Analysis and Time Series*, Academic Press, NY, 1981.
2. S. P. Lloyd, Least squares quantization in PCM, *IEEE Trans. Inform. Theory* **28**: 127–135 (1982).
3. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.
4. C. E. Shannon, Coding theorems for a discrete source with a fidelity criteria, *IRE National Convention Record, Part 4*, 142–163, 1959.
5. R. M. Gray, *Source Coding Theory*, Kluwer Academic Press, 1992.
6. T. Berger, *Rate Distortion Theory*, Prentice-Hall Inc., Englewood Cliffs, 1971.
7. T. D. Lookabaugh and R. M. Gray, High-resolution quantization theory and the vector quantizer advantage, *IEEE Trans. Inform. Theory* **IT-35**: 1023–1033 (Sept. 1989).
8. T. R. Fischer, A pyramid vector quantizer, *IEEE Trans. Inform. Theory* **32**: 568–583 (July 1986).
9. M. Barlaud et al., Pyramid lattice vector quantization for multiscale image coding, *IEEE Trans. Image Process.* **3**: 367–381 (July 1994).
10. J. H. Conway and N. Sloane, A fast encoding method for lattice codes and quantizers, *IEEE Trans. Inform. Theory* **IT-29**: 820–824 (Nov. 1983).
11. A. Buzo, Y. Linde, and R. M. Gray, An algorithm for vector quantizer design, *IEEE Trans. Commun.* **20**: 84–95 (Jan. 1980).
12. B. Ramamurthi and A. Gersho, Classified vector quantization of image, *IEEE Trans. Commun.* **34**: 1105–1115 (Nov. 1986).
13. D.-Y. Cheng, A. Gersho, B. Ramamurthi, and Y. Shoham, Fast search algorithms for vector quantization and pattern matching, *Proc. of the International Conference on ASSP*, 911.1–911.4 (March 1984).
14. J. L. Bentley, Multidimensional binary search tree used for associative searching, *Comm. ACM*, 509–526 (Sept. 1975).
15. M. Soleymani and S. Morgera, An efficient nearest neighbor search method, *IEEE Trans. Commun.* **20**: 677–679 (1987).
16. Abramson, Norman, *Information Theory and Coding*, McGraw-Hill, New York, 1963.
17. P. G. Howard and J. S. Vitter, Practical implementations of arithmetic coding, in *Image and Text Compression*, J. A. Storer, ed., Kluwer Academic Publishers, Boston, MA, 1992, pp. 85–112.
18. J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans. Inform. Theory* **IT-23**: 337–343 (1977).
19. R. B. Wells, *Applied Coding and Information Theory for Engineers*, Prentice-Hall, Englewood Cliffs, NJ, 1999.
20. T. A. Welch, A technique for high-performance data compression, *IEEE Trans. Comput.* 17:8–19 (1997).
21. M. Rabani and P. Jones, *Digital Image Compression Techniques*, SPIE Optical Engineering Press, Bellingham, WA, 1991.
22. O. Egger, P. Fleury, T. Ebrahimi, and M. Kunt, High-performance compression of visual information — A tutorial review — Part I: Still pictures, *Proc. IEEE* **87**(6): 976–1011 (June 1999).
23. T. Berger and J. D. Gibson, Lossy source coding, *IEEE Trans. Inform. Theory* **44**(6): 2693–2723 (1998).
24. T. Ebrahimi and M. Kunt, Visual data compression for multimedia applications, *Proc. IEEE* **86**(6): 1109–1125 (Jun. 1998).
25. M. J. Narasimha and A. M. Peterson, On the computation of the discrete cosine transform, *IEEE Trans. Commun.* **COM-26**: 934–936 (Jun. 1978).
26. W. Li, A new algorithm to compute the DCT and its inverse, *IEEE Trans. Signal Proc.* **39**(6): 1305–1313 (June 1991).
27. D. Slawecki and W. Li, DCT/IDCT processor design for high data rate image coding, *IEEE Trans. Circuits Syst. Video Technol.* **2**(2): 135–146 (1992).

28. W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, 1993.
29. J. D. Johnson, A filter family designed for use in quadrature mirror filter bands, *Proc. Int. Conf. ASSP* (*ICASSP*) 291–294 (Apr. 1980).
30. H. S. Malavar, *Signal Processing with Lapped Transforms*, Artech House Norwood, MA, 1992.
31. Z. Zhou and A. N. Venetsanopoulos, Morphological methods in image coding, *Proc. Int. Conf. ASSP* (*ICASSP*) **3**: 481–484 (March 1992).
32. J. R. Casas, L. Torres, and M. Jare no, Efficient coding of residual images, *Proc. SPIE, Visual Communications Image Processing* **2094**: 694–705 (1993).
33. A. Toet, A morphological pyramid image decomposition, *Pattern Reconstruction Lett.* **9**(4): 255–261 (May 1989).
34. T. Sikora, Low complexity shape-adaptive DCT for coding of arbitrarily shaped image systems, *Signal Processing: Image Commun.* **7**: 381–395 (Nov. 1995).
35. T. Sikora and B. Makai, Shape-adaptive DCT for generic coding of video, *IEEE Trans. Circuits Syst. Video Technol.* **5**: 59–62 (Feb. 1995).
36. S. Li and W. Li, Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding, *IEEE Trans. Circuits Syst. Video Technol.* **10**(5) 725–743 (Aug. 2000).
37. J. M. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. Signal Process.* **41**: 3445–3462 (Dec. 1993).
38. W. Li, Overview of fine granularity scalability in MPEG-4 video standard, *IEEE Trans. Circuit Syst. Video Technol.* **11**(3): 301–317 (March 2001). Special issue on streaming video.
39. W. Li and Y.-Q. Zhang, Vector-based signal processing and quantization for image and video compression, *Proc. IEEE* **83**(2): 317–335 (Feb. 1995).
40. W. Li and Y.-Q. Zhang, Vector transform coding of subband-decomposed images, *IEEE Trans. Circuits Syst. Video Technol.* **4**(4): 383–391 (Aug. 1994).
41. W. Li, On vector transformation, *IEEE Trans Signal Proc.* **41**(11): 3114–3126 (Nov. 1993).
42. W. Li, Vector transform and image coding, *IEEE Trans. Circuits Syst. Video Technol.* **1**(4): 297–307 (Dec. 1991).
43. H. Q. Cao and W. Li, A fast search algorithm for vector quantization using a directed graph, *IEEE Trans. Circuits Syst. Video Technol.* **10**(4): 585–593 (June 2000).
44. F. Ling, W. Li, and H. Sun, Bitplane coding of DCT coefficients for image and video compression, *Proc. of SPIE VCIP'99* (Jan. 1999).
45. C. Wang, H. Q. Cao, W. Li, and K. K. Tzeng, Lattice labeling algorithms for vector quantization, *IEEE Trans. Circuits Syst. Video Technol.* **8**(2): 206–220 (Apr. 1998).
46. W. Li et al., A video coding algorithm using vector-based techniques, *IEEE Trans. Circuits Syst. Video Technol.* **7**(1): 146–157 (Feb. 1997).
47. F. Ling and W. Li, Dimensional adaptive arithmetic coding for image compression, *Proc. of IEEE International Symposium on Circuits and Systems* (May–June 1998).
48. I. Daubechies and W. Sweldens, Factoring wavelet transforms into lifting steps, *J. Fourier Anal. Appl.* **4**: 247–269 (1998).
49. T. Koga et al., Motion compensated interframe coding for video conferencing, *Proc. of the National Telecommunications Conference*, New Orleans, LA, pp. G5.3.1–G5.3.5 (Dec. 1981).
50. J. Y. Tham, S. Ranganath, M. Ranganath, and A. A. Kassim, A novel unrestricted center-biased diamond search algorithm for block motion estimation, *IEEE Trans. Circuits Syst. Video Technol.* **8**(4): 369–377 (Aug. 1998).
51. S. Zhu and K. K. Ma, A new Diamond search algorithm for fast block matching motion estimation, *Proc. 1st Int. Conf. Inform. Commun. Signal Process.* ICICS '97, **1**: 292–296 (Sept. 1997).
52. Z. L. He and M. L. Liou, A high performance fast search algorithm for block matching motion estimation, *IEEE Trans. Circuits Syst. Video Technol.* **7**(5): 826–828 (Oct. 1997).
53. S. Zafar, Y. Zhang, and J. S. Baras, Predictive block-matching motion estimation schemes for video compression — Part II. Inter-frame prediction of motion vectors, *IEEE Proc. SOUTHEASTCON 91* **2**: 1093–1095 (April 1991).
54. L. W. Lee, J. F. Wang, J. Y. Lee, and J. D. Shie, Dynamic search-window adjustment and interlaced search for block-matching algorithm, *IEEE Trans. Circuits Syst. Video Technol.* **3**(1): 85–87 (Feb. 1993).
55. B. Liu and A. Zaccarin, New fast algorithms for the estimation of block motion vectors, *IEEE Trans. Circuits Syst. Video Technol.* **3**(2): 148–157 (April 1993).
56. A. Smolic, T. Sikora, and J.-R. Ohm, Long-term global motion estimation and its application for sprite coding, content description and segmentation, *IEEE Trans. CSVT* **9**(8): 1227–1242 (Dec. 1999).
57. B. Girod, Motion-compensating prediction with fractional-pel accuracy, *IEEE Trans. Commun.* **41**: 604–612 (April 1993).
58. M. T. Orchard and G. J. Sullivan, Overlapped block motion compensation: an estimation-theoretic approach, *IEEE Trans. Image Proc.* **3**(5): 693–699 (Sept. 1994).
59. C. J. Hollier, J. F. Arnold, and M. C. Cavenor, The effect of a loop filter on circulating noise in interframe video coding, *IEEE Trans. Circuits Syst. Video Technol.* **4**(4): 442–446 (Aug. 1994).
60. M. Yuen and H. R. Wu, *Performance of loop filters in MC/DPCM/DCT video coding*, *Proceedings of ICSP'96*, pp. 1182–1186.
61. S. Mallat and Z. Zhang, Matching pursuit with time-frequency dictionaries, *IEEE Trans. Signal Proc.* **41**: 3397–3415 (Dec. 1993).
62. R. Neff and A. Zakhor, Very low bit-rate video coding based on matching pursuits, *IEEE Trans. Circuits Syst. Video Technol.* **7**(1): 158–171 (Feb. 1997).
63. O. K. Al-Shaykh et al., Video compression using matching pursuits, *IEEE Trans. Circuits Syst. Video Technol.* **9**(1): 123–143 (Feb. 1999).
64. W. E. Glenn, J. Marcinka, and R. Dhein, Simple scalable video compression using 3-D subband coding, *SMPTE J.* **106**: 140–143 (March 1996).
65. W. A. Pearlman, B.-J. Kim, and Z. Xiong, Embedded video coding with 3D SPIHT, P. N. Topiwala, ed., in *Wavelet Image and Video Compression*, Kluwer, Boston, MA, 1998.

66. S. J. Choi and J. W. Woods, Motion-compensated 3-D subband coding of video, *IEEE Trans. Image Proc.* **8**: 155–167 (Feb. 1999).

67. J. R. Ohm, Three-dimensional subband coding with motion compensation, *IEEE Trans. Image Process.* **3**(5): 559–571 (Sept. 1994).

68. J. Xu, S. Li, Z, Xiong, and Y.-Q. Zhang, 3-D embedded subband coding with optimized truncation (3-D ESCOT), *ACHA Special Issue on Wavelets* (May 2001).

69. J. Xu, S. Li, Z. Xiong, and Y.-Q. Zhang, On boundary effects in 3-D wavelet video coding, *Proc. Symposium on Optical Science and Technology*, (July 2000).

70. M. F. Barnsley, *Fractals Everywhere*, Academic Press, San Diego, CA, 1988.

71. A. E. Jacquin, Image coding based on a fractal theory of iterated contractive image transformations, *IEEE Trans. Image Process.* **1**: 18–30 (Jan. 1992).

72. Y. Fisher, A discussion of fractal image compression, in Saupe D. H. O. Peitgen, H. Jurgens, eds., *Chaos and Fractals*, Springer-Verlag, New York, 1992, pp. 903–919.

73. E. W. Jacobs, Y. Fisher, and R. D. Boss, Image compression: A study of iterated transform method, *Signal Process.* **29**: 251–263 (Dec. 1992).

74. K. Barthel, T. Voy'e, and P. Noll, Improved fractal image coding, in *Proc. Picture Coding Symp.* (*PCS*) **1.5**: (March 1993).

75. J. Y. Huang and O. M. Schultheiss, Block quantization of correlated Gaussian random variables, *IEEE Trans. Commun.* **11**: 289–296 (Sept. 1963).

76. N. S. Jayant and P. Noll, *Digital Coding of waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

77. A. Segall, Bit allocation and encoding for vector sources, *IEEE Trans. Inform. Theory* **IT-22**: 162–169 (March 1976).

78. T. Chiang and Y.-Q. Zhang, A new rate control scheme using quadratic rate distortion model, *IEEE Trans. Circuits Syst. Video Technol.* **7**(1): 246–250 (Feb. 1997).

79. Y. Shoham and A. Gersho, Efficient bit allocation for an arbitrary set of quantizers, *IEEE Trans. ASSP* **36**: 1445–1453 (Sept. 1988).

80. K. Ramchandran, A. Ortega, and M. Vetterli, Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders, *IEEE Trans. Image Process.* **3**(5): 533–545 (Sept. 1994).

81. T. Weigand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra, Rate-distortion optimized mode selection for very low bit-rate video coding and the emerging H.263 standard, *IEEE Trans. Circuits Syst. Video Technol.* **6**: 182–190 (April 1996).

82. H. Everett, Generalized lagrange multiplier method for solving problems of optimum allocation of resources, *Oper. Res.* **11**: 399–417 (1963).

83. G. T. Toussaint, Pattern recognition and geometrical complexity, in *Proc. 5th Int. Conf. Pattern Recognition* 1324–1347 (Dec. 1980).

84. P. H. Westerink, J. Biemond, and D. E. Boekee, An optimal bit allocation algorithm for subband coding, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) 757–760 (1988).

85. E. A. Riskin, Optimal bit allocation via the generalized BFOS algorithm, *IEEE Trans. Inform. Theory* **37**: 400–402 (March 1991).

86. M. Rabbani and R. Joshi, An overview of the JPEG 2000 still image compression standard, *Signal Process. Image Commun.* **17**(1): 3–48 (Jan. 2002). Special issue on JPEG 2000.

87. F. Wu, S. Li, and Y.-Q. Zhang, A framework for efficient progressive fine granular scalable video coding, *IEEE Trans. Circuits Syst. Video Technol.* **11**(3): 332–344 (March 2001). Special Issue for Streaming Video.

88. Q. Wang, Z. Xiong, F. Wu, and S. Li, Optimal rate allocation for progressive fine granularity scalable video coding, *IEEE Signal Process. Lett.* **9**: 33–39 (Feb. 2002).

89. J. M. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. SP* **41**(12): 3445–3462 (Dec. 1993).

90. A. Said and W. A. Pearlman, A new, fast, and efficient image codec based on set partitioning in hierarchical trees, *IEEE Trans. CSVT* **6**(3): 243–250 (June 1996).

91. D. Taubman, EBCOT (embedded block coding with optimized truncation): A complete reference, *ISO/IEC JTC1/SC29/WG1* N983, (Sept. 1998).

92. J. Rissanen, Universal coding, information, prediction, and estimation, *IEEE Trans. Inform. Theory* **IT-30**: 629–636 (July 1984).

93. M. J. Weinberger and G. Seroussi, Sequential prediction and ranking in universal context modeling and data compression, *IEEE Trans. Inform. Theory* **43**: 1697–1706 (Sept. 1997).

94. M. Mattavelli and A. Nicoulin, Pre and post processing for very low bit-rate video coding, in *Proc. Int. Workshop HDTV* (Oct. 1994).

95. M. I. Sezan, M. K. Ozkan, and S. V. Fogel, Temporally adaptive filtering of noisy image sequences using a robust motion estimation algorithm, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) **IV**: 2429–2432 (May 1991).

96. L. Chulhee, M. Eden, and M. Unser, High-quality image resizing using oblique projection operators, *IEEE Trans. Image Proc.* **7**(5): 679–692 (May 1998).

97. A. Munoz, T. Blu, and M. Unser, Least-squares image resizing using finite differences, *IEEE Trans. Image Proc.* **10**(9): 1365–1378 (Sept. 2001).

98. R. Castagno, P. Haavisto, and G. Ramponi, A method for motion adaptive frame rate up-conversion, *IEEE Trans. Circuits Systems Video Technol.* **6**(5): 436–466 (Oct. 1996).

99. Y.-K. Chen, A. Vetro, H. Sun, and S. Y. Kung, Frame-rate up-conversion using transmitted true motion vectors, in *Proc. IEEE Second Workshop on Multimedia Signal Processing*, 622–627 (Dec. 1998).

100. S.-K. Kwon, K.-S. Seo, J.-K. Kim, and Y.-G. Kim, A motion-adaptive de-interlacing method, *IEEE Trans. Consumer Electron.* **38**(3): 145–150 (Aug. 1992).

101. C. Sun, De-interlacing of video images using a shortest path technique, *IEEE Trans. Consumer Electron.* **47**(2): 225–230 (May 2001).

102. J. Schwendowius and G. R. Arce, Data-adaptive digital video format conversion algorithms, *IEEE Trans. Circuits Systs. Video Technol.* **7**(3): 511–526 (June 1997).

103. H. C. Reeve and J. S. Lim, Reduction of blocking effects in image coding, *Opt. Eng.* **23**(1): 34–37 (1984).

104. B. Ramamurthi and A. Gersho, Nonlinear space-variant postprocessing of block coded images, *IEEE Trans. Acoust. Speech Signal Process.* **34**: 1258–1268 (Oct. 1986).

105. R. L. Stevenson, Reduction of coding artifacts in transform image coding, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) **V**: 401–404 (April 1993).

106. L. Yan, A nonlinear algorithm for enhancing low bit-rate coded motion video sequence, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) **II**: 923–927 (Nov. 1994).

107. Y. Yang, N. P. Galatsanos, and A. K. Katsaggelos, Iterative projection algorithms for removing the blocking artifacts of block-DCT compressed images, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) **V**: 401–408 (April 1993).

108. B. Macq et al., Image visual quality restoration by cancellation of the unmasked noise, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) **V**: 53–56 (1994).

109. T. Chen, Elimination of subband-coding artifacts using the dithering technique, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) **II**: 874–877 (Nov. 1994).

110. W. Li, O. Egger, and M. Kunt, Efficient quantization noise reduction device for subband image coding schemes, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP*) **IV**: 2209–2212 (May 1995).

111. K. K. Pang and T. K. Tan, Optimum loop filter in hybrid coders, *IEEE Trans. Circuits Syst. Video Technol.* **4**(2): 158–167 (April 1994).

112. Tao Bo and M. T. Orchard, Removal of motion uncertainty and quantization noise in motion compensation, *IEEE Trans. Circuits Syst. Video Technol.* **11**(1): 80–90 (Jan. 2001).

113. S. Shirani, F. Kossentini, and R. Ward, A concealment method for video communications in an error-prone environment, *IEEE J. Select. Areas Commun.* **18**(6): 1122–1128 (June 2000).

114. S. Shirani, B. Erol, and F. Kossentini, Error concealment for MPEG-4 video communication in an error prone environment, in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (2000).

115. Y. Wang and Q.-F. Zhu, Error control and concealment for video communication: a review, *Proc. IEEE* **86**(5): 974–997 (May 1998).

116. H. R. Rabiee, H. Radha, and R. L. Kashyap, Error concealment of still image and video streams with multidirectional recursive nonlinear filters, in *Proceedings of International Conference on Image Processing* (1996).

117. D. Gatica-Perez, M.-T. Sun, and C. Gu, Semantic video object extraction using four-band watershed and partition lattice operators, *IEEE Trans. Circuits Syst. Video Technol.* **11**(5): 603–618 (May 2001).

118. K. Sethi and N. V. Patel, A statistical approach to scene change detection, in *IS&T SPIE Proc.: Storage and Retrieval for Image and Video Databases III* **2420**: 329–339 (Feb. 1995).

119. J.-J. Chen and H.-M. Hang, Source model for transform video coder and its application II. Variable frame rate coding, *IEEE Trans. Circuits Syst. Video Technol.* **7**(2): 299–311 (April 1997).

120. T.-Y. Kuo, J. W. Kim, and C.-J. Kuo, Motion-compensated frame interpolation scheme for H.263 codec, in *Proc. IEEE International Symposium on Circuits and Systems* (*ISCAS*) (1999).

121. C. Christopoulos, J. Askelof, and M. Larsson, Efficient methods for encoding regions of interest in the upcoming JPEG2000 still image coding standard, *IEEE Signal Process. Lett.* **7**(9): 247–249 (Sept. 2000).

122. K. R. Rao and P. Yip, *Discrete Cosine Transforms — Algorithms, Advantages, Applications*, Academic Press, (1990).

123. CCITT Recommendation T.4, *Standardization of Group 3 Facsimile Apparatus for Document Transmission*, (1980).

124. CCITT Recommendation T.6, *Facsimile Coding Schemes and Coding Control Functions for Group 4 Facsimile Apparatus*, (1984).

125. ITU-T Recommendation T.82, *Information technology — Coded representation of picture and audio information-Progressive bi-level image compression*, (1993).

126. ISO/IEC-11544, *Progressive Bi-level Image Compression. International Standard*, (1993).

127. ISO/IEC-14492 FCD, *Information Technology — Coded Representation of Picture and Audio Information — Lossy/Lossless Coding of Bi-Level Images*, Final Committee Draft. ISO/IEC JTC 1/SC 29/WG 1 N 1359, (July, 1999).

128. ISO/IEC JTC 1/SC 29/WG 1, ISO/IEC FCD 15444-1: *Information technology — JPEG 2000 image coding system: Core coding system* [WG1 N 1646], (March 2000).

129. W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, (1992).

130. ISO/IEC, ISO/IEC 14496-2:1999: *Information technology — Coding of audio-visual objects — Part 2*: Visual, (Dec. 1999).

131. ISO/IEC, ISO/IEC 14495-1:1999: *Information technology — Lossless and near-lossless compression of continuous-tone still images: Baseline*, (Dec. 1999).

132. M. J. Weinberger, G. Seroussi, and G. Sapiro, The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS, *IEEE Trans. Img. Process.* **9**(8): 1309–1324 (Aug. 2000).

133. S. W. Golomb, Run-length encodings, *IEEE Trans. Inform. Theory* **IT-12**: 399–401 (July 1966).

134. R. F. Rice, *Some practical universal noiseless coding techniques*, Tech. Rep. JPL-79-22, Jet Propulsion Laboratory, Pasadena, CA, (March 1979).

135. W. Li, Y.-Q. Zhang, I. Sodagar, J. Liang, and S. Li, MPEG-4 texture coding, A. Puri and C. Chen, eds., in *Multimedia Systems, Standards, and Networks*, Marcel Dekker, Inc., New York, 2000.

136. S. Li and I. Sodagar, Generic, scalable and efficient shape coding for visual texture objects in MPEG-4, In proc. *International Symposium on Circuits and Systems 2000* (May 2000).

137. ISO/IEC FCD15444-1:2000, Information Technology — Jpeg 2000 Image Coding System, V1.0, March 2000.

138. ITU-T Recommendation H.261, Video codec for audiovisual services at p*64 kbits/sec, 1990.

139. MPEG-1 Video Group, Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5 Mbit/s—: Part 2 — Video, ISO/IEC 11172-2, International Standard, 1993.

140. MPEG-2 Video Group, *Information Technology — Generic Coding of Moving Pictures and Associated Audio*: Part 2 — Video, ISO/IEC 13818-2, International Standard, 1995.

141. ITU-T Experts Group on Very Bit rate Visual Telephony, ITU-T Recommendation H.263: Video Coding for Low Bit rate Communication, Dec. 1995.

142. Video coding for low bit rate communication, ITU-T SG XVI, DRAFT 13, H.263+, Q15-A-60 rev. 0, 1997.

143. ITU-T Q.15/16, Draft for H.263 + + Annexes U, V, and W to Recommendation H.263, November, 2000.

144. MPEG-4 Video Group, Generic coding of audio-visual objects: Part 2 — Visual, ISO/IEC JTC1/SC29/WG11 N1902, FDIS of ISO/IEC 14496-2, Atlantic City, Nov. 1998.

145. Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG T, Working Draft Number 2, Revision 8 (WD-2 rev 8), JVT-B118r8, April, 2002.

146. Y. T. Chan, *Wavelet Basics*, Kluwer Academic Publishers, Norwell, MA, 1995.

147. G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996.

148. M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

149. S. Li and W. Li, Shape adaptive vector wavelet coding of arbitrarily shaped objects, *SPIE Proceeding: Visual Communications and Image Processing'97* San-Jose, (Feb. 1997).

150. S. Li, W. Li, H. Sun, and Z. Wu, Shape adaptive wavelet coding, *Proc. IEEE International Symposium on Circuits and Systems* ISCAS'98 5: 281–284 (May 1998).

151. C. Gu and M.-C. Lee, Semiautomatic segmentation and tracking of semantic video objects, *IEEE Trans. Circuits Syst. Video Technol.* **8**(5): (Sep. 1998).

152. J.-H. Pan, S. Li, and Y.-Q. Zhang, Automatic moving video object extraction using multiple features and multiple frames, *ISCAS 2000* (May 2000).

153. H. Zhong, W. Liu, and S. Li, A semi-automatic system for video object segmentation, *ICME 2001* (Aug. 2001).

154. N. Li, S. Li, and W. Liu, A novel framework for semi-automatic video object segmentation, *IEEE ISCAS 2002* (May 2002).

155. S. J. Gortler, R. Grzeszezuk, R. Szeliski, and M. F. Cohen, The lumigraph, *Computer Graphics* (*SIGGRAPH'96*) 43–54 (Aug. 1996).

156. M. Levoy and P. Hanrahan, Light field rendering, *Computer Graphics* (*SIGGRAPH'96*) (Aug. 1996).

157. H. Shum and L. He, Rendering with concentric mosaics, *Computer Graphics* (*SIGGRAPH'99*) 299–306 (Aug. 1999).

158. J. Li, H. Shum, and Y.-Q. Zhang, On the compression of image based rendering scene, in *Proc. International Conference on Image Processing*, **2**: 21–24 (Sept. 2000).

159. P. D. Symes, *Video Compression*, McGraw-Hill, New York, 1998.

160. J. Watkinson, *The MPEG Handbook*, Focal Press, 2001.

161. W. Effelsberg and R. Steinmetz, *Video Compression Techniques*, dpunkt-Verlag, 1998.

162. A. Bovik, ed., *Handbook of image and video processing*, Academic Press, 2000.

IMAGE COMPRESSION

ALFRED MERTINS
University of Wollongong
Wollongong, Australia

1. INTRODUCTION

The extensive use of digital imaging in recent years has led to an explosion of image data that need to be stored on disks or transmitted over networks. Application examples are facsimile, scanning, printing, digital photography, multimedia, Internet Websites, electronic commerce, digital libraries, medical imaging, and remote sensing. As a result of this ever-increasing volume of data, methods for the efficient compression of digital images have become more and more important. To give an example, a modern digital camera stores about 3.3 megapixels per image. With three color components and 8-bit resolution per color component, this makes an amount of approximately 10 MB (megabytes) of raw data per image. Using a 64-MB memory card, direct storage would allow only six images to be stored. With digital image compression, however, the same camera can store about 60 images or more on the memory card without noticeable differences in image quality. High compression factors without much degradation of image quality are possible because images usually contain a large amount of spatial correlation. In other words, images will typically have larger areas showing similar color, gray level, or texture, and these similarities can be exploited to obtain compression.

Image compression can generally be categorized into lossless and lossy compression. Obviously, lossless compression is most desirable as the reconstructed image is an exact copy of the original. The compression ratios that can be obtained with lossless compression, however, are fairly low, typically ranging from 3 and 5, depending on the image. Such low compression ratios are justified in applications where no loss of quality can be tolerated, as is often the case in the compression and storage of medical images. For most other applications such as internet browsing or storage for printing some loss is usually acceptable. Allowing for loss allows for much higher compression ratios.

Early image compression algorithms have mainly focused on achieving low distortion for a given (fixed) rate, or conversely, the lowest rate for a given maximum distortion. While these goals are still valid, modern multimedia applications have led to a series of further requirements such as spatial and signal-to-noise ratio (SNR) scalability and random access to parts of the image data. For example, in a typical Internet application the same image is to be accessed from various users with a wide range of devices (from a low-resolution palmtop

computer to a multimedia workstation) and via channels ranging from slow cable modems or wireless connections to high-speed wired local-area networks. To optimally utilize available bandwidth and device capabilities it is desirable to transcode image data within the network to the various resolutions that best serve the respective users. On the fly transcoding, however, requires the compressed image data to be organized in such a way that different content variations can be easily extracted from a given high-resolution codestream. The new JPEG2000 standard takes a step in this direction and combines abundant functionalities with very high compression ratio and random codestream access.

The aim of this article is to give an overview of some of the most important image compression tools and techniques. We start in Section 2 by looking at the theoretical background of data compression. Section 3 then reviews the discrete cosine and wavelet transforms, the two most important transforms in image compression. Section 4 looks at embedded state-of-the-art wavelet compression methods, and Section 5 gives an overview of standards for the compression of continuous-tone images. Finally, Section 6 gives a number of conclusions and outlook.

2. ELEMENTS OF SOURCE CODING THEORY

2.1. Rate Versus Distortion

The mathematical background that describes the tradeoff between compression ratio and fidelity was established by Shannon in his work on rate–distortion (RD) theory [1]. The aim of this work was to determine the minimum bit rate required to code the output of a stochastic source at a given maximum distortion. The theoretical results obtained by Shannon have clear implications on the performance of any practical coder, and in the following we want to have a brief look at what the RD tradeoff means in practice. For this we consider the graphs in Fig. 1, which show the distortion obtained with different coders versus the rate required to store the information. Such graphs obtained for a specific image and encoding algorithm are known as operational distortion–rate curves. The distortion is typically measured as the mean-squared error (MSE)

$$D = \frac{1}{NM} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |\hat{x}_{m,n} - x_{m,n}|^2 \qquad (1)$$

where $x_{m,n}$ is the original image and $\hat{x}_{m,n}$ is the reconstructed image. However, also of interest are other distortion measures that possibly better reflect the amount of distortion as perceived by a human observer. The rate is measured as the required average number of bits per pixel (bpp). As one might expect, all coders show the same maximum distortion at rate zero (no transmission). With increasing rate the distortion decreases and different coders reach the point of zero distortion at different rates. Coder 1 is an example of a coder that combines lossy and lossless compression in one bit stream and that is optimized for lossless compression. Coder 2, on the other hand, is a lossy coder that shows good performance for low rates, but reaches the lossless stage only at extremely high rates. The lowest curve shows the minimum rate required to code the information at a given distortion, assuming optimal compression for all rates. This is the desired curve, combining lossy and lossless compression in an optimal way.

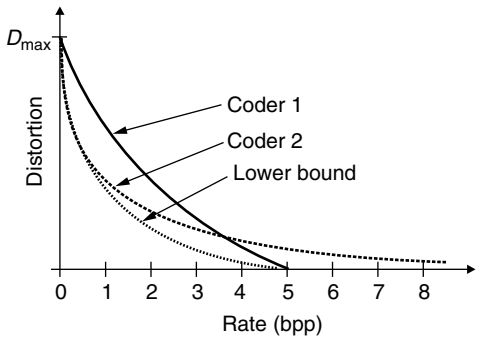


Figure 1. Distortion-rate curves for different coders.

2.2. Coding Gain Through Decorrelation

The simplest way to compress an image in a lossy manner would be to quantize all pixels separately and to provide a bit stream that represents the quantized values. This strategy is known as *pulse code modulation* (PCM). For example an 8-bit gray-scale image whose pixels $x_{m,n}$ are integers in the range from 0 to 255 could be compressed by a factor of four through neglecting the six least significant bits of the binary representations of $x_{m,n}$, resulting in an image with only four different levels of gray. Entropy coding (see Section 2.4) of the PCM values would generally allow increased compression, but such a strategy would still yield a poor tradeoff between the amount of distortion introduced and the compression ratio achieved. The reason for this is that the spatial relationships between pixels are not utilized by PCM.

Images usually contain a large amount of spatial correlation, which can be exploited to obtain compression schemes with a better RD tradeoff than that obtained with PCM. For this the data first needs to be decorrelated, and then quantization and entropy coding can take place. Figure 2 shows the basic structure of an image coder that follows such a strategy. The quantization step is to be seen as the assignment of a discrete symbol to a range of input values. In the simplest case this could be the assignment of symbol a through the operation $a = round(x/q)$, where q is the quantization step size. The inverse quantization then corresponds to the recovery of the actual numerical values that belong to the symbols (e.g., the operation $\hat{x} = q \cdot a$). The entropy coding stage subsequently endeavors to represent the generated discrete symbols with the minimum possible number of bits. This step is lossless. Errors occur only due to quantization. More details on entropy coding will be given in Section 2.4.

One of the simplest decorrelating transforms is a *prediction error filter*, which uses the knowledge of neighboring pixels to predict the value of the pixel of interest and then outputs the prediction error made.

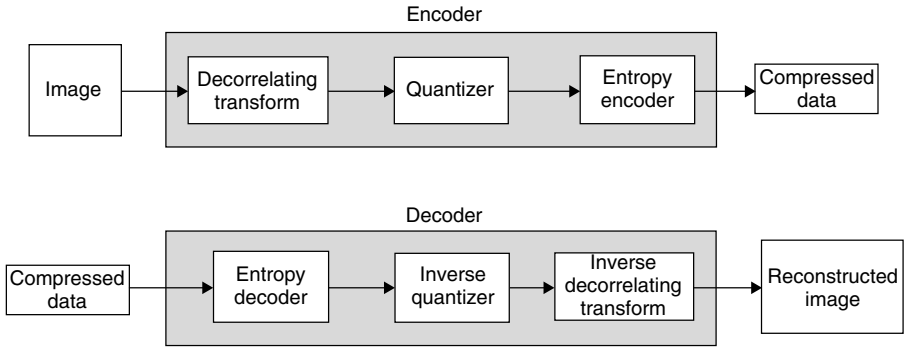


Figure 2. Typical image encoder and decoder structures.

In this case the coding paradigm changes from PCM to differential PCM, known as DPCM. To give a practical example, Fig. 12 shows the neighborhood relationship used in the JPEG-LS standard where the pixel values a, b, c, and d are used to predict an estimate $\hat{x}$ for the actual value x. Under ideal conditions (stationary source and optimal prediction), the output values $e = x - \hat{x}$ of a prediction error filter would be entirely mutually uncorrelated, but in practice there will still be some correlation left, for example due to spatially varying image properties. A coding gain of DPCM over PCM arises from the fact that the prediction error usually has much less power than the signal itself and can therefore be compressed more efficiently.

Modern state-of-the-art image compression algorithms use transforms like the discrete cosine transform (DCT) or the discrete wavelet transform (DWT) to carry out decorrelation [2–4]. These transforms are extremely efficient in decorrelating the pixels of an image and are employed in a number of compression standards [5,6]. To see how and why transform coding works, we consider the system in Fig. 3. The input is a zero-mean random vector $\boldsymbol{x} = [x_0, x_1, \ldots, x_{M-1}]^T$ with correlation matrix $\boldsymbol{R}_{xx} = E\{\boldsymbol{x}\boldsymbol{x}^T\}$, where $E\{\cdot\}$ denotes the expectation operation. The output $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{x}$ of the transform is then a zero-mean random process with correlation matrix $\boldsymbol{R}_{yy} = \boldsymbol{T}\boldsymbol{R}_{xx}\boldsymbol{T}^T$. Because the random variables $y_0, y_1, \ldots, y_{M-1}$ stored in $\boldsymbol{y}$ may have different variances, they are subsequently quantized with different quantizers $Q_0, Q_1, \ldots, Q_{M-1}$. In the synthesis stage the inverse transform is applied to reconstruct an approximation $\hat{\boldsymbol{x}}$ of the original vector $\boldsymbol{x}$ based on the quantized coefficients $\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{M-1}$. The aim is to design the system in such a way that the mean-squared error $E\{\|\boldsymbol{x} - \boldsymbol{x}\|^2\}$ becomes minimal for a given total bit budget. Questions arising are (1) what is the optimal transform $\boldsymbol{T}$ given the properties of the source, (2) how should an available bit budget B be distributed to the different quantizers, and (3) what are the optimal quantization levels? Answers to these questions have been derived for both unitary and biorthogonal transforms. We will briefly sketch the derivation for the unitary case. To simplify the expressions we assume that $x_0, x_1, \ldots, x_{M-1}$ and thus also $y_0, y_1, \ldots, y_{M-1}$ are zero-mean Gaussian random variables.

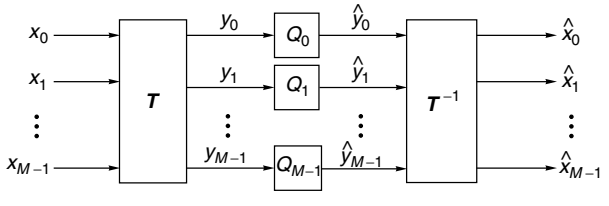


Figure 3. Transform coding.

The optimal scalar quantizers Q_k that minimize the individual error variances $\sigma_{q_k}^2 = E\{q_k^2\}$ with $q_k = y_k - \hat{y}_k$ for a given number of quantization steps are known as *Lloyd–Max quantizers* [7,8]. Important properties of these optimal quantizers are $E\{q_k\} = 0$ (zero-mean error) and $E\{q_k\hat{y}_k\} = 0$ (orthogonality between the quantized value and the error) [9].

The bit allocation can be derived under the assumption of mutually uncorrelated quantization errors and RD relationships of the form

$$\sigma_{q_k}^2 = \gamma_k \sigma_{y_k}^2 2^{-2B_k}, \quad k = 0, 1, \ldots, M-1 \tag{2}$$

for the individual quantizers. The term $\sigma_{q_k}^2$ in (2) is the variance of the quantization error produced by quantizer Q_k, $\sigma_{y_k}^2$ is the variance of y_k, and B_k is the number of bits spent for coding y_k. The values γ_k, $k = 0, 1, \ldots, M-1$ depend on the PDFs of the random variables y_k. Because of the Gaussian assumption made earlier we have equal γ_k for all k. The assumption of uncorrelated quantization errors stated above means that $E\{q_i q_k\} = 0$ for $i \neq k$, which is usually satisfied if the quantization is sufficiently fine, even if the random variables y_k are mutually correlated. Minimizing the average quantization error

$$\sigma_q^2 = \frac{1}{M}\sum_{k=0}^{M-1} \sigma_{q_k}^2 \tag{3}$$

under the constraint of a fixed average bit rate

$$\frac{1}{M}\sum_{k=0}^{M-1} B_k = B \tag{4}$$

using the Lagrange multiplier method yields the bit allocation [9]:

$$B_k = B + \frac{1}{2}\log_2 \frac{\sigma_{y_k}^2}{\left(\prod_{j=0}^{M-1} \sigma_{y_j}^2\right)^{1/M}}, \quad k = 0, 1, \ldots, M-1. \tag{5}$$

According to (5) the bits are to be allocated such that $2^{B_k} \sim \sigma_{y_k}^2$, which intuitively makes sense as more bits are assigned to random variables with a larger variance. Another interesting property of optimal bit allocation and quantizer selection can be seen when substituting B_k according to (5) into (2). One obtains

$$\sigma_{q_k}^2 = \gamma 2^{-2B} \left(\prod_{j=0}^{M-1} \sigma_{y_j}^2 \right)^{1/M}, \quad k = 0, 1, \ldots, M-1 \tag{6}$$

which means that in the optimum situation all quantizers contribute equally to the final output error.

In order to answer the question of which transform is optimal for a given coding task we consider the coding gain defined as

$$G_{\mathrm{TC}} = \frac{\sigma_{\mathrm{PCM}}^2}{\sigma_{\mathrm{TC}}^2} \tag{7}$$

where σ_{PCM}^2 is the error variance of simple PCM and σ_{TC}^2 is the error variance produced by transform coding, both at the same average bit rate B. Assuming that all random variables x_k have the same variance σ_x^2 the error of PCM amounts to $\sigma_{\mathrm{PCM}}^2 = \gamma 2^{-2B} \sigma_x^2$. Using (3) and (6) the error of transform coding can be written as $\sigma_{\mathrm{TC}}^2 = \gamma 2^{-2B} \left(\prod_{j=0}^{M-1} \sigma_{y_j}^2 \right)^{1/M}$, and the coding gain becomes

$$G_{\mathrm{TC}} = \frac{\sigma_x^2}{\left(\prod_{k=0}^{M-1} \sigma_{y_k}^2 \right)^{1/M}} \tag{8}$$

Bearing in mind that for a unitary transform

$$1/M \sum_{k=0}^{M-1} \sigma_{y_k}^2 = \sigma_x^2$$

the coding gain can be seen as the quotient of arithmetic and geometric mean of the coefficient variances. Among all unitary transforms $\boldsymbol{T}$, this quotient is maximized when $\boldsymbol{R}_{yy}$ is diagonal [10] and thus when the transform coefficients are uncorrelated. This decorrelation is accomplished by the Karhunen–Loève transform (KLT). The rows of the transform matrix $\boldsymbol{T}$ are then the transposed eigenvectors of the eigenvalue problem $\boldsymbol{R}_{xx}\boldsymbol{t}_k = \lambda_k \boldsymbol{t}_k$.

The Eq. (8) can be interpreted as follows:

1. A maximum coding gain is obtained if the coefficients y_k are mutually uncorrelated.
2. The more unequal the variances $\sigma_{y_k}^2$ are, the higher the coding gain is, because a high dissimilarity leads to a high ratio of arithmetic and geometric mean. Consequently, if the input values x_k are already mutually uncorrelated (white process), a transform cannot provide any further coding gain.
3. With increasing M one may expect an increasing coding gain that moves toward an upper limit as M goes to infinity. In fact, one can show that this is the same limit as the one obtained for DPCM with ideal prediction [9].

It is interesting to note that the expression (8) for the coding gain also holds for subband coding based on uniform, paraunitary filterbanks (i.e., filterbanks that carry out unitary transforms). The term σ_x^2 is then the variance of an ongoing stationary input process, and the values $\sigma_{y_k}^2$ are the subband variances. A more general expression for the coding gain has been derived by Katto and Yasuda [11]. Their formula also holds for biorthogonal transform and subband coding as well as other schemes such as DPCM.

2.3. Vector Quantization

Vector quantization (VQ) is a multidimensional extension of scalar quantization in that an entire vector of values is encoded as a unit. Let $\boldsymbol{x}$ be such an N-dimensional vector of values, and let $\boldsymbol{x}_i$, $i = 1, 2, \ldots, I$ be a set of N-dimensional codevectors stored in a codebook. Given $\boldsymbol{x}$ a vector quantizer finds the codevector $\boldsymbol{x}_\ell$ from the codebook that best matches $\boldsymbol{x}$ and transmits the corresponding index ℓ. Knowing ℓ the receiver reconstructs $\boldsymbol{x}$ as $\boldsymbol{x}_\ell$. An often used quality criterion to determine which vector from the codebook gives the best match for $\boldsymbol{x}$ is the Euclidean distance $d(\boldsymbol{x}, \boldsymbol{x}_i) = \|\boldsymbol{x} - \boldsymbol{x}_i\|^2$. Theoretically, if the vector length N tends to infinity, VQ becomes optimal and approaches the performance indicated by rate distortion theory. In practice, however, the cost associated with searching through a large codebook is the major obstacle. See Ref. 12 for discussions of computationally efficient forms of the VQ technique and details on codebook design.

2.4. Entropy Coding

Assigning the same code lengths to all symbols generated by a source is not optimal when the different symbols occur with different probabilities. In such a case it is better to assign short codewords to symbols that occur often and longer codewords to symbols that occur only occasionally. The latter strategy results in variable-length codes and is the basic principle of entropy coding.

A simple source model is the discrete memoryless source (DMS), which produces random variables X_i taken from an alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_L\}$. The symbols a_i may, for example, identify the various steps of a quantizer and are assumed to occur with probabilities $p(a_i)$. The entropy of the source is defined as

$$H = -\sum_{i=1}^{L} p(a_i) \log_2 p(a_i) \tag{9}$$

and describes the average information per symbol (in bits). According to this equation, the more skewed the probability distribution, the lower the entropy. For any given number of symbols L the maximum entropy is obtained if all symbols are equally likely. The entropy provides a lower bound for the average number of bits per symbol required to encode the symbols emitted by a DMS.

The most popular entropy coding methods are Huffman coding, arithmetic coding, and Lempel–Ziv coding; the

first two are frequently applied in image compression. Lempel–Ziv coding, a type of universal coding, is used more often for document compression, as it does not require a priori knowledge of the statistical properties of the source.

Huffman coding uses variable-length codewords and produces a uniquely decodable code. To construct a Huffman code, the symbol probabilities must be known a priori. As an example, consider the symbols a_1, a_2, a_3, a_4 with probabilities 0.5, 0.25, 0.125, 0.125, respectively. A fixed-length code would use two bits per symbol. A possible Huffman code is given by $a_1 : 0$, $a_2 : 10$, $a_3 : 110$, $a_4 : 111$. This code requires only 1.75 bits per symbol on average, which is the same as the entropy of the source. In fact, one can show that Huffman codes are optimal and reach the lower bound stated by the entropy when the symbol probabilities are powers of $\frac{1}{2}$. In order to decode a Huffman code, the decoder must know the code table that has been used by the encoder. In practice this means that either a specified standard code table must be employed or the code table must be transmitted to the decoder as side information.

In arithmetic coding there is no one-to-one correspondence between symbols and codewords, as it assigns variable-length codewords to variable-length blocks of symbols. The codeword representing a sequence of symbols is a binary number that points to a subinterval of the interval $[0, 1)$ that is associated with the given sequence. The length of the subinterval is equal to the probability of the sequence, and each possible sequence creates a different subinterval. The advantage of arithmetic over Huffman coding is that it usually results in a shorter average code length when the symbol probabilities are not powers of $\frac{1}{2}$, and arithmetic coders can be made adaptive to learn the symbol probabilities on the fly. No side information in form of a code table is required.

3. TRANSFORMS FOR IMAGE COMPRESSION

3.1. The Discrete Cosine Transform

The discrete cosine transform (DCT) is used in most current standards for image and video compression. Examples are JPEG, MPEG-1, MPEG-2, MPEG-4, H.261, and H.263. To be precise, there are four different DCTs defined in the literature [2], and in particular it is the DCT-II that is used for image compression. Because there is no ambiguity throughout this text we will simply call it the DCT. The DCT of a two-dimensional (2D) signal $x_{m,n}$ with $m, n = 0, 1, \ldots, M-1$ is defined as

$$y_{k,\ell} = \frac{2\gamma_k\gamma_\ell}{M} \sum_{m=0}^{M-1}\sum_{n=0}^{M-1} x_{m,n} \cos\frac{k(m+\frac{1}{2})\pi}{M} \cos\frac{\ell(n+\frac{1}{2})\pi}{M}, \quad k, \ell = 0, 1, \ldots, M-1 \tag{10}$$

with

$$\gamma_k = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } k = 0 \\ 1 & \text{otherwise} \end{cases}$$

The DCT is a unitary transform, so that the inverse transform (2D IDCT) uses the same basis sequences. It is given by

$$x_{m,n} = \sum_{k=0}^{M-1}\sum_{\ell=0}^{M-1} \frac{2\gamma_k\gamma_\ell}{M} y_{k,\ell} \cos\frac{k(m+\frac{1}{2})\pi}{M} \cos\frac{\ell(n+\frac{1}{2})\pi}{M}, \quad m, n = 0, 1, \ldots, M-1 \tag{11}$$

The popularity of the DCT comes from the fact that it almost reaches the coding gain obtained by the KLT for typical image data while having the advantage of fast implementation. In fact, for a 1D (one-dimensional) first-order autoregressive input process with autocorrelation sequence $r_{xx}(m) = \sigma_x^2 \rho^{|m|}$ and correlation coefficient $\rho \to 1$ it has been shown that the DCT asymptotically approaches the KLT [9]. Fast implementations can be obtained through the use of FFT (Fast Fourier Transform) algorithms or through direct factorization of the DCT formula [2]. The latter approach is especially interesting for the 2D case where 2D factorizations lead to the most efficient implementations [13,14].

In image compression the DCT is usually used on nonoverlapping 8×8 blocks of the image rather than on the entire image in one step. The following aspects have lead to this choice. First, from the theory outlined in Section 2.2 and the good decorrelation properties of the DCT for smooth signals, it is clear that in order to maximize the coding gain for typical images, the blocks should be as big as possible. On the other hand, with increasing block size the likelihood of capturing a nonstationary behavior within a block increases. This however decreases the usefulness of the DCT for decorrelation. Finally, quantization errors made for DCT coefficients will spread out over the entire block after reconstruction via the IDCT. At low rates this can lead to annoying artifacts when blocks consist of a combination of flat and highly textured regions, or if there are significant edges within a block. These effects are less visible if the block size is small. Altogether, the choice of 8×8 blocks has been found to be a good compromise between exploiting neighborhood relations in smooth regions and avoiding annoying artifacts due to inhomogeneous block content.

Figure 4 shows an example of the blockwise 2D DCT of an image. The original image of Fig. 4a has a size of 144×176 pixels (QCIF format) and the blocksize for the DCT is 8×8. Figure 4b shows the blockwise transformed image, and Fig. 4c shows the transformed image after rearranging the coefficients in such a way that all coefficients with the same physical meaning (i.e., coefficients $y_{k,\ell}$ from the different blocks) are gathered in a subimage. For example the subimage in the upper left corner of Fig. 4c contains the coefficients $y_{0,0}$ of all the blocks in Fig. 4b. These coefficients are often called *DC coefficients*, because they represent the average pixel value within a block. Correspondingly, the remaining 63 coefficients of a block are called *AC coefficients*. From Fig. 4c one can see that the DC coefficients contain the most important information on the entire image. Toward the lower right corner of Fig. 4c the amount of signal energy decreases significantly, represented by the average level of gray.

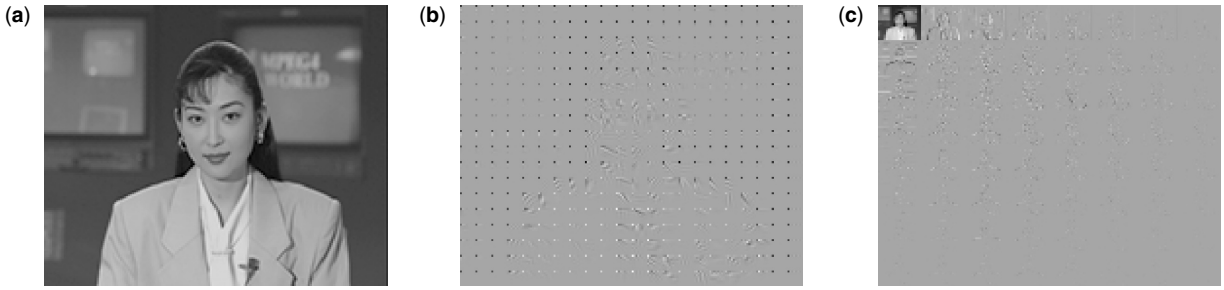


Figure 4. Example of a blockwise 2D DCT of an image: (**a**) original image (144 × 176 pixels); (**b**) transformed image (blocksize 8 × 8, 18 × 26 blocks); (**c**) transformed image after reordering of coefficients (8 × 8 subimages of size 18 × 26).

3.2. The Discrete Wavelet Transform

The discrete wavelet transform (DWT) is a tool to hierarchically decompose a signal into a multiresolution pyramid. It offers a series of advantages over the DCT. For example, contrary to the blockwise DCT the DWT has overlapping basis functions, resulting in less visible artifacts when coding at low bit rates. Moreover, the multiresolution signal representation offers functionalities such as spatial scalability in a simple and generic way. While most of the image compression standards are based on the DCT, the new still image compression standard JPEG2000 and parts of the MPEG4 multimedia standard rely on the DWT [6,15].

For discrete-time signals the DWT is essentially an octave-band signal decomposition, carried out through successive filtering operations and sampling rate changes. The basic building block of such an octave-band filterbank is the two-channel structure depicted in Fig. 5. $H_0(z)$ and $G_0(z)$ are lowpass, while $H_1(z)$ and $G_1(z)$ are highpass filters. The blocks with arrows pointing downward indicate downsampling by factor 2 (i.e., taking only every second sample), and the blocks with arrows pointing upward indicate upsampling by 2 (insertion of zeros between the samples). Downsampling serves to eliminate redundancies in the subband signals, while upsampling is used to recover the original sampling rate. Because of the filter characteristics (lowpass and highpass), most of the energy of a lowpass-type signal $x(n)$ will be stored in the subband samples $y_0(m)$. Because $y_0(m)$ occurs at half the sampling rate of $x(n)$ it appears that the filterbank structure concentrates the information in less samples, as required for efficient compression. More efficiency can be obtained by cascading two-channel filterbanks to obtain octave-band decompositions or other frequency resolutions. For the structure in Fig. 5 to allow perfect reconstruction (PR) of the input with a delay of n_0 samples [i.e., $\hat{x}(n) = x(n - n_0)$], the filters must satisfy

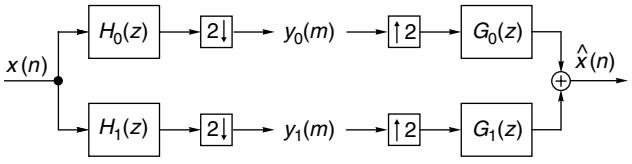


Figure 5. Two-channel filterbank.

$$H_0(-z)G_0(z) + H_1(-z)G_1(z) = 0 \tag{12}$$

and

$$H_0(z)G_0(z) + H_1(z)G_1(z) = 2z^{-n_0} \tag{13}$$

Equation (12) guarantees that the aliasing components that occur due to the subsampling operation will be compensated at the output, while (13) finally ensures perfect transmission of the signal through the system. In addition to the PR conditions (12) and (13) the filters should satisfy $H_0(1) = G_0(1) = \sqrt{2}$ and $H_1(1) = G_1(1) = 0$, which are essential requirements to make them valid wavelet filters. Moreover, they should satisfy some regularity constraints as outlined by Daubechies [3]. It should be noted that (12) and (13) are the PR constraints for biorthogonal two-channel filterbanks. Paraunitary filter banks and the corresponding orthonormal wavelets require the stronger condition

$$|H_0(e^{j\omega})|^2 + |H_0(e^{j(\omega+\pi)})|^2 = 2 \tag{14}$$

to hold. Apart from the special case where the filter length is 2, Eq. (14) can be satisfied only by filters with nonsymmetric impulse responses [16]. Symmetric filters are very desirable because they allow for simple boundary processing (see below). Therefore paraunitary two-channel filterbanks and the corresponding orthonormal wavelets are seldom used in image compression.

For the decomposition of images, the filtering and downsampling operations are usually carried out separately in the horizontal and vertical directions. Figure 6 shows an illustration of the principle that yields a 2D octave-band decomposition that corresponds to a DWT. In order to ensure that the analysis process results in the same number of subband samples as there are input pixels, special boundary processing steps are required. These will be explained below. An example of the decomposition of an image is depicted in Fig. 7. One can see that the DWT concentrates the essential information on the image in a few samples, resulting in a high coding gain.

When decomposing a finite-length signal (a row or column of an image) with a filterbank using linear convolution the total number of subband samples is generally higher than the number of input samples. Methods to

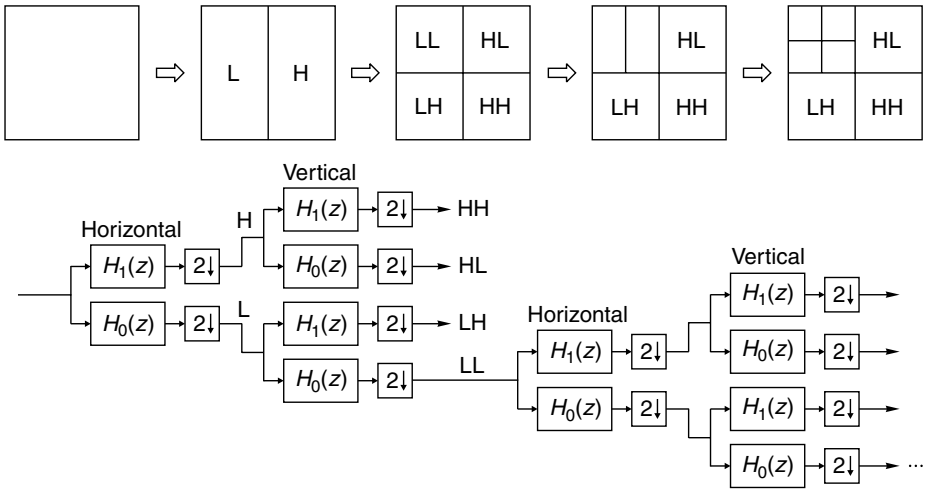


Figure 6. Separable 2D octave-band filterbank.

Figure 7. Example of a 2D octave-band decomposition.

resolve this problem are circular convolution [17], symmetric extension [18,19], and boundary filtering [20–22]. In the following we will describe the method of symmetric extension, which is the one most often used in practice. It requires the filters in the filterbank to have linear phase, which means that biorthogonal filters/wavelets have to be used. We will address the procedure for filters with odd and even length separately.

If the filter impulse responses have odd lengths, linear phase means that they obey the whole-sample symmetry (WSS) shown in Fig. 8a. For these filters the symmetric extension of the signal also has to be carried out with WSS, as depicted in Fig. 8b. A signal of length N is thus extended to a periodic signal $x_{\mathrm{WSS}}(n)$ with period $2N-2$ and symmetry within each period. When analyzing such an extended signal with the corresponding filterbank, the obtained subband signals will also be periodic and will show symmetry within each period. The type of subband symmetry depends on the filter and signal lengths. For the case where N is even, the obtained subband symmetries are depicted in Figs. 8c,d. One can easily see that only a total number of N distinct subband samples needs to be stored or transmitted, because from these N samples the periodic subband signals can be completely recovered. Feeding the periodic subband signals into the synthesis filterbank, and taking one period of the output finally yields perfect reconstruction.

Linear-phase filters with even length have the half-sample symmetry (HSS) in Fig. 9a for the lowpass and the half-sample antisymmetry (HSAS) in Fig. 9b for the highpass. In this case the HSS extension in Fig. 9c is required, resulting in an extended signal with period $2N$ and symmetry within each period. Again, the subband signals show symmetries that can be exploited to capture the entire information on a length-N input signal in a total number of N subband samples. The subband symmetries obtained for even N are depicted in Figs. 9d,e.

Note that the extension schemes outlined above can also be used for cases where N is odd, resulting in different subband symmetries. Moreover, with the introduction of two different subsampling phases, they can be used to define nonexpansive DWTs for arbitrarily shaped objects, see e.g., [23]. Such a scheme is included in the MPEG-4 standard [15].

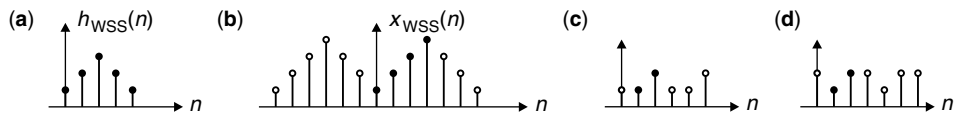


Figure 8. Symmetric extension for odd-length filters: (**a**) impulse response with whole-sample symmetry (WSS); (**b**) periodic signal extension with WSS (the original signal is marked with black dots); (**c**) HSS-WSS subband symmetry, obtained for filter lengths $L = 3 + 4k$, where k is an integer; (**d**) WSS-HSS subband symmetry, obtained for filter lengths $L = 5 + 4k$.

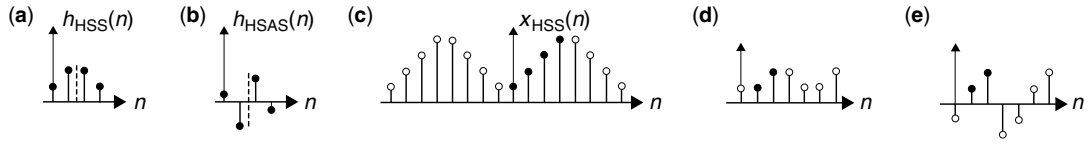


Figure 9. Symmetric extension for even-length filters: (**a**) impulse response with half-sample symmetry; (**b**) half-sample antisymmetry; (**c**) periodic signal extension with half-sample symmetry; (**d**) HSS-HSS, obtained with HSS filter; (**e**) HSAS-HSAS, obtained with HSAS filter.

4. EMBEDDED WAVELET CODING

Early wavelet image coders used to allocate bits to the various subbands according to the principles outlined in Section 2, followed by runlength and entropy coding for further compression of the quantized wavelet coefficients. Runlength coding was found particularly useful for coding of long stretches of zeros that frequently occur within the higher bands. Although such coders can perform reasonably well for a fixed bit rate, they do not offer much flexibility, and especially, they do not allow for progressive transmission of the wavelet coefficients in terms of accuracy. A new era of wavelet coders started with Shapiro's embedded zerotree wavelet (EZW) coder [24], which was the first coder that looked at simultaneous relationships between wavelet coefficients at different scales and produced an entirely embedded codestream that could be truncated at any point to achieve the best reconstruction for the number of symbols transmitted and/or received. The key idea of the EZW coder was the introduction of zerotrees, which are sets of coefficients gathered across scales that are all quantized to zero with regard to a given quantization step size and can be coded with a single symbol. All coefficients within a zerotree belong to the same image region. The formation of zerotrees and the parent–child relationships within a zerotree are shown in Fig. 10a. From looking at the wavelet transform in Fig. 7, it is clear that it is quite likely that all coefficients in such a tree may be quantized to zero in a smooth image region. The concept of EZW coding was refined by Said and Pearlman, who proposed a coding method known as *set partitioning in hierarchical trees* (SPIHT), a state-of-the-art coding method that offers high compression and fine granular SNR scalability [25]. Both the EZW and SPIHT coders follow the idea of transmitting the wavelet coefficients in a semiordered manner, bitplane by bitplane, together with the sorting information required to identify the positions of the transmitted coefficients. In the following we will take a closer look at the SPIHT coder, which is more efficient in transmitting the sorting information. In fact, the SPIHT algorithm is so efficient that additional arithmetic coding will result in only marginal improvements [25].

The SPIHT coder uses three lists to organize the sorting of coefficients and the creation of the bit stream. These are a list of insignificant pixels (LIP), a list of insignificant sets (LIS), and a list of significant pixels (LSP). During initialization the coordinates of the coefficients in the lowest band are stored in the LIP, and 3/4 of them are also stored in the LIS where they are seen as roots of insignificant sets. Figure 10b illustrates the structure of a

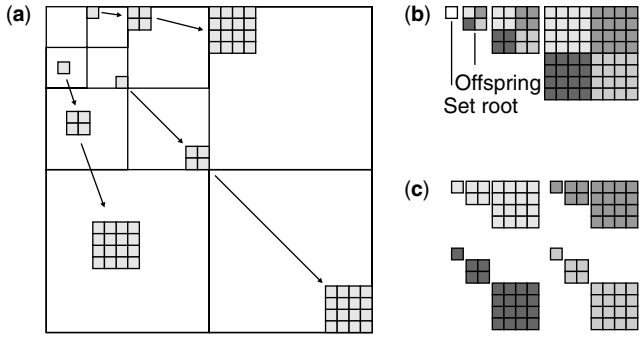


Figure 10. Formation of sets: (**a**) sets/zerotrees within the wavelet tree; (**b**) set; (**c**) set of (**b**) partitioned into four subsets.

set where each coefficient has four offspring. The LSP is left empty at this stage. A start threshold T is defined as $T = 2^n$, where $n = \lfloor \log_2(y_{max}) \rfloor$ and y_{max} is the magnitude of the largest subband coefficient. After initialization the algorithm goes through sorting and refinement stages with respect to T. During the sorting phase each coefficient in the LIP is compared with the threshold T and the result of the comparison (a symbol being 0 or 1) is sent to the channel. If a coefficient exceeds T, its sign is transmitted and its coordinate is moved to the LSP. In a second phase of the sorting pass, each set having its root in the LIS is compared with T, and if no coefficient exceeds T, a zero is sent to the channel. If at least one coefficient within a set is larger than T, then a one is sent and the set is subdivided into the four offspring and four smaller sets. The offspring are tested for significance and their coordinates are moved to the appropriate list (LIP or LSP). The offspring coordinates are also used as roots for the four smaller sets; see Fig. 10, Parts (b) and (c) for an illustration of set partitioning. The significance test and subdivision of sets are carried out until all significant coefficients with respect to T have been isolated. At the end of the procedure with threshold T, all coordinates of significant coefficients are stored in the LSP and their signs have been transmitted along with the sorting information required to identify the positions of the transmitted coefficients. In the next stage the threshold is halved and the accuracy of the coefficients in the LSP is refined by sending the information of whether a coefficient lies in the upper or lower halves of the uncertainty interval. Then the next sorting pass with the new threshold is carried out, and so on. The procedure is repeated until a bit budget is exhausted or the threshold falls below a given limit.

The SPIHT decoder looks at the same significance tests as the encoder and receives the answers to the tests from the bit stream. This allows the decoder to

(**a**)

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

(**b**)

17	18	24	47	99	99	99	99
18	21	26	66	99	99	99	99
24	26	56	99	99	99	99	99
47	66	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99

(**c**)

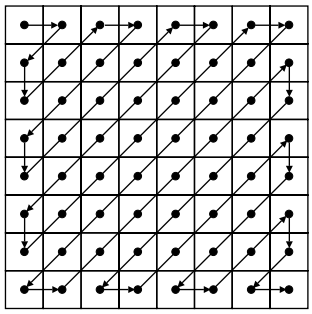


Figure 11. JPEG quantization matrices: (**a**) luminance; (**b**) chrominance; (**c**) zigzag scanning order.

reconstruct the subband coefficients bitplane by bitplane. The reconstruction levels are always in the middle of the uncertainty interval. This means that if the decoder knows for example that a coefficient is larger than or equal to 32 and smaller than 63, it reconstructs the coefficient as 48. The bit stream can be truncated at any point to meet a bit budget, and the decoder can reconstruct the best possible image for the number of bits received.

An interesting modification of the SPIHT algorithm has been proposed [26], called *virtual SPIHT*. This coder virtually extends the octave decomposition until only three set roots are required in the LIS at initialization. This formation of larger initial sets results in more efficient sorting during the first rounds where most coefficients are insignificant with respect to the threshold. Further modifications include 3-D extensions for coding of video [27].

5. COMPRESSION STANDARDS FOR CONTINUOUS-TONE IMAGES

This section presents the basic modes of the JPEG, JPEG-LS, and JPEG2000 compression standards for still images. There also exist standards for coding of bilevel images, such as JBIG and JBIG2, but these will not be discussed here.

5.1. JPEG

JPEG is an industry standard for digital image compression developed by the Joint Photographic Experts Group, which is a group of experts nominated by leading companies and national standards bodies. JPEG was approved by the principal members of ISO/IEC JTC1 as an international standard (IS 109181) in 1992 and by the CCITT as recommendation T.81, also in 1992. It includes the following modes of operation [5]: a sequential mode, a progressive mode, a hierarchical mode, and a lossless mode. In the following we will discuss the so-called baseline coder, which is a simple form of the sequential mode that encodes images block by block in scan order from left to right and top to bottom. Image data are allowed to have 8-bit or 12-bit precision. The baseline algorithm is designed for lossy compression with target bit rates in the range of 0.25–2 bits per pixel (bpp). The coder uses blockwise DCTs, followed by scalar quantization and entropy coding based on run-length and Huffman coding. The general structure is the same as in Fig. 2. The color space is not specified in the standard, but mostly the YUV space is used with each color component treated separately.

After the blockwise DCT (block size 8×8), scalar quantization is carried out with uniform quantizers. This is done by dividing the transform coefficients $y_{k,\ell}, k, \ell = 0, 1, \ldots, 7$ in a block by the corresponding entries $q_{k,\ell}$ in an 8×8 quantization matrix and rounding to the nearest integer: $a_{k,\ell} = \text{round}(y_{k,\ell}/q_{k,\ell})$. Later in the reconstruction stage an inverse quantization is carried out as $\hat{y}_{k,\ell} = q_{k,\ell} \cdot a_{k,\ell}$. The perceptually optimized quantization matrices in Fig. 11 have been included in the standard as a recommendation, but they are not a requirement and other quantizers may be used. To obtain more flexibility, the entire quantization matrices are often scaled such that step sizes $q'_{k,\ell} = D \cdot q_{k,\ell}$ are used instead of $q_{k,\ell}$. The factor D then gives control over the bit rate.

The quantized coefficients in each block are scanned in a zigzag manner, as shown in Fig. 11, and are then further entropy encoded. First the DC coefficient of a block is differentially encoded with respect to the DC coefficient of the previous block (DPCM), using a Huffman code. Then the remaining 63, quantized AC coefficients of a block are encoded. Because the occurrence of long stretches of zeros during the zigzag scan is quite likely, zero runs are encoded using a runlength code. If all remaining coefficients along the zigzag scan are zero, a special end-of-block (EoB) symbol is used. The actual coefficient values are Huffman-encoded.

5.2. JPEG-LS

JPEG-LS is a standard for lossless and near-lossless compression [28]. In the near-lossless mode the user can specify the maximum error ε that may occur during compression. Lossless coding means $\varepsilon = 0$. The performance of JPEG-LS for lossless coding is significantly better than the lossless mode in JPEG.

To achieve compression, context modeling, prediction, and entropy coding based on Golomb–Rice codes are used. Golomb–Rice codes are a special class of Huffman codes for geometric distributions. The context of a pixel x is determined from four reconstructed pixels a, b, c, d in the neighborhood of x, as shown in Fig. 12. Context is used to decide whether x will be encoded in the run or regular mode.

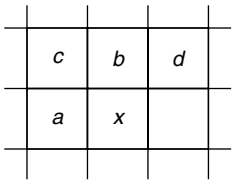


Figure 12. Context modeling in JPEG-LS.

The run mode (run-length coding) is chosen when the context determines that x is likely to be within the specified tolerance ε of the last encoded pixel. This may be the case in smooth image regions. The run mode is ended either at the end of the current line or when the reconstruction error for a pixel exceeds ε. To encode the runlength a modified Golomb–Rice code is applied.

The regular mode is used when, based on the context, it is unlikely that x lies within the tolerance ε of the previously encoded pixel. In this mode, a prediction for x is computed based on the three neighboring values a, b, and c according to

$$\hat{x} = \begin{cases} \min(a,b) & \text{if } c \geq \max(a,b) \\ \max(a,b) & \text{if } c \leq \min(a,b) \\ a+b-c & \text{otherwise} \end{cases} \tag{15}$$

This predictor adapts to local edges and is known as a median edge detection predictor. In a second step the prediction is corrected by a context-dependent term to remove systematic prediction biases. The difference between the bias-corrected prediction and the actual value is then encoded using a Golomb–Rice code.

5.3. JPEG2000

JPEG2000 is a new standard for still-image compression that provides a series of functionalities that were not addressed by the original JPEG standard [6,29]. It is meant to complement JPEG and not to replace it altogether. The main features of JPEG2000 are as follows:

- Any type of image (bilevel, continuous-tone, multicomponent) with virtually no restriction on image size
- A wide range of compression factors from 200 : 1 to lossless
- Progressive transmission by accuracy (signal-to-noise ratio) and spatial resolution
- Lossless and lossy compression within one bit stream
- Random codestream access
- Robustness to bit errors
- Region of interest with improved quality

To be backward-compatible with JPEG, the new JPEG2000 standard includes a DCT mode similar to JPEG, but the core of JPEG2000 is based on the wavelet transform. JPEG2000 is being developed in several stages, and in the following we refer to the baseline JPEG2000 coder as specified in Part I of the standard [6]. Part II includes optional techniques such as trellis-coded quantization [30], which are not required for all implementations. Further parts address "Motion JPEG2000," conformance, reference software, and compound image file formats for prepress and faxlike applications. The structure of the baseline JPEG2000 coder is essentially the one in Fig. 2 with the wavelet transform of Fig. 6 as the decorrelating transform. To allow for processing of extremely large images on hardware with limited memory resources, it is possible to divide images into tiles and carry out the wavelet transform and compression for each tile independently. The wavelet filters specified are the Daubechies 9-7 filters [4] for maximum performance in the lossy mode and the 5-3 integer-coefficient filters [31] for an integrated lossy/lossless mode. Boundary processing is carried out using the extension scheme for odd-length filters discussed in Section 3.2.

The quantization stage uses simple scalar quantizers with a dead zone around zero. The step sizes for the quantizers are determined from the dynamic ranges of the coefficients in the different subbands. In the lossless mode where all subband coefficients are integers, the step size is one. The quantized coefficients within each subband are then grouped into codeblocks that are encoded separately. The compression of a codeblock is carried out bitplane by bitplane using a context-dependent arithmetic coding technique. This results in independent embedded bit streams for each codeblock. Independent compression of subbands in codeblocks is the key to random codestream access and to simple spatial resolution scalability. Only the codeblocks referring to a certain region of interest at a desired spatial resolution level of the wavelet tree need to be transmitted or decoded. In order to facilitate for SNR scalability, a layer technique is used where each layer is composed of parts of the blockwise embedded bit streams. This is illustrated in Fig. 13, which shows the formation of layers from the individual bit streams. To ensure that the final codestream is optimally embedded (layer by layer) and that a target bit rate or distortion is met, the truncation points for the individual embedded bit streams can be determined via an operational postcompression rate–distortion optimization [32]. However, although the rate allocation proposed by Taubmann [32] is used in the JPEG2000 verification model [33], other methods may also be employed. Rate allocation is an encoder issue, and the standard specifies only the decoder and the structure of the codestream.

To demonstrate the performance of JPEG2000 and provide a comparison with the older JPEG standard, Fig. 14 shows some coding examples. One can see that JPEG produces severe blocking artifacts at low rates while JPEG2000 tends to produce slightly blurry images. At

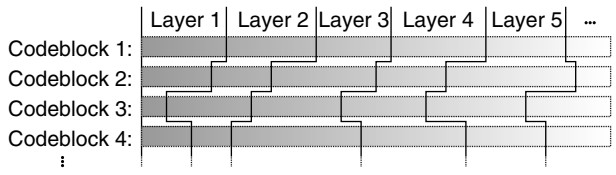


Figure 13. Formation of layered bit stream from embedded bit streams of individual codeblocks.

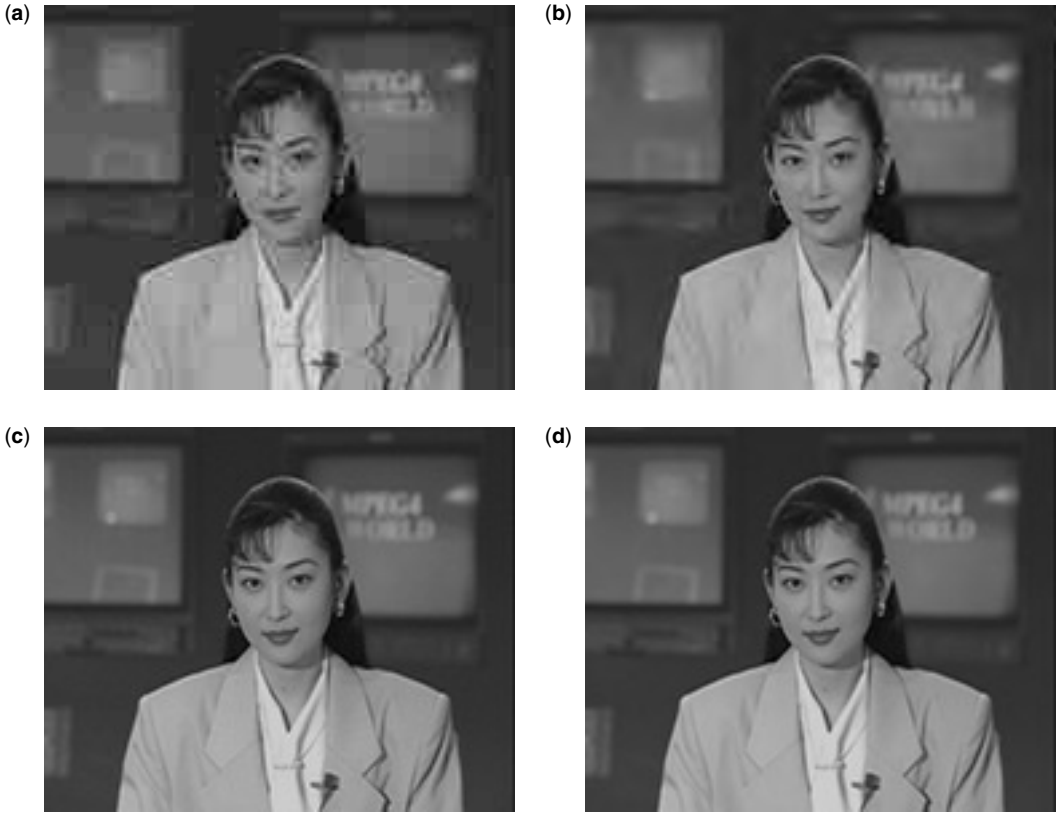


Figure 14. Coding examples (QCIF format): (**a**) JPEG at 0.5 bpp; (**b**) JPEG2000 at 0.5 bpp; (**c**) JPEG at 2 bpp; (**d**) JPEG2000 at 2 bpp.

higher rates both standards yield good-quality images with JPEG2000 still having the better signal-to-noise ratio.

6. CONCLUSIONS

This article has reviewed general concepts and standards for still-image compression. We started by looking at theoretical foundations of data compression and then discussed some of the most popular image compression techniques and standards. From todays point of view the diverse functionalities required by many multimedia applications are best provided by coders based on the wavelet transform. As demonstrated in the JPEG2000 standard, wavelet-based coders even allow the integration of lossy and lossless coding, which is a feature that is very desirable for applications such as medical imaging where highest quality is needed. The compression ratios obtained with lossless JPEG2000 are in the same order as the ones obtained with dedicated lossless methods. However, because lossless coding based on the wavelet transform is still in a very early stage, one may expect even better integrated lossy and lossless wavelet-based coders to be developed in the future.

BIOGRAPHY

Alfred Mertins received the Dipl.-Ing. degree in electrical engineering from University of Paderborn, Germany, and the Dr.-Ing. and Dr.-Ing. habil. degrees in electrical engineering from Hamburg University of Technology, Germany, in 1984, 1991, and 1994, respectively. From 1986 to 1991 he was with the Hamburg University of Technology, from 1991 to 1995 with the Microelectronics Applications Center Hamburg, Germany, from 1996 to 1997 with the University of Kiel, Germany, and from 1997 to 1998 with the University of Western Australia, Crawley. Since 1998, he has been a senior lecturer at the School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, New South Wales, Australia. His research interests include digital signal processing, wavelets and filter banks, image and video processing, and digital communications.

BIBLIOGRAPHY

1. C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE Nat. Conserv. Rec.* **4**: 142–163 (1959).
2. K. R. Rao and P. Yip, *Discrete Cosine Transform*, Academic Press, New York, 1990.
3. I. Daubechies, *Ten Lectures on Wavelets*, SIAM, 1992.
4. M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, Image coding using wavelet transform, *IEEE Trans. Image Process.* **1**(2): 205–220 (April 1992).
5. G. K. Wallace, The JPEG still picture compression standard, *IEEE Trans. Consumer Electron.* **38**(1): 18–34 (Feb. 1992).

6. Joint Photographic Experts Group, *JPEG2000 Final Draft International Standard, Part I*, ISO/IEC JTC1/SC29/WG1 FDIS15444-1, Aug. 2000.
7. S. P. Lloyd, Least squares quantization in PCM, *Institute of Mathematical Statistics Society Meeting*, Atlantic City, NJ, Sept. 1957, pp. 189–192.
8. J. Max, Quantizing for minimum distortion, *IRE Trans. Inform. Theory* 7–12 (March 1960).
9. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
10. R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.
11. J. Katto and Y. Yasuda, Performance evaluation of subband coding and optimization of its filter coefficients, *Proc. SPIE Visual Communication and Image Processing*, Nov. 1991, pp. 95–106.
12. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, Boston, 1991.
13. P. Duhamel and C. Guillemot, Polynomial transform computation of the 2-D DCT, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Albuquerqe, NM, April 1990, pp. 1515–1518.
14. W.-H. Fang, N.-C. Hu, and S.-K. Shih, Recursive fast computation of the two-dimensional discrete cosine transform, *IEE Proc. Visual Image Signal Process.* **146**(1): 25–33 (Feb. 1999).
15. *MPEG-4 Video Verification Model, Version 14. Generic Coding of Moving Pictures and Associated Audio*, ISO/IEC JTC1/SC 29/WG 11, 1999.
16. P. P. Vaidyanathan, On power-complementary FIR filters, *IEEE Trans. Circuits Syst.* **32**: 1308–1310 (Dec. 1985).
17. J. Woods and S. O'Neil, Subband coding of images, *IEEE Trans. Acoust. Speech Signal Process.* **34**(5): 1278–1288 (May 1986).
18. M. J. T. Smith and S. L. Eddins, Analysis/synthesis techniques for subband coding, *IEEE Trans. Acoust. Speech Signal Process.* 1446–1456 (Aug. 1990).
19. J. N. Bradley, C. M. Brislawn, and V. Faber, Reflected boundary conditions for multirate filter banks, *Proc. Int. Symp. Time-Frequency and Time-Scale Analysis*, Canada, 1992, pp. 307–310.
20. R. L. de Queiroz, Subband processing of finite length signals without border distortions, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, San Francisco, March 1992, Vol. IV, pp. 613–616.
21. C. Herley, Boundary filters for finite-length signals and time-varying filter banks, *IEEE Trans. Circuits Syst. II* **42**(2): 102–114 (Feb. 1995).
22. A. Mertins, Boundary filters for size-limited paraunitary filter banks with maximum coding gain and ideal DC behavior, *IEEE Trans. Circuits Syst. II* **48**(2): 183–188 (Feb. 2001).
23. A. Mertins, *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications*, Wiley, Chichester, UK, 1999.
24. J. M. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. Signal Process.* **41**(12): 3445–3462 (Dec. 1993).
25. A. Said and W. A. Pearlman, A new fast and efficient image codec based on set partitioning in hierarchical trees, *IEEE Trans. Circuits Syst. Video Technol.* **6**(3): 243–250 (June 1996).
26. E. Khan and M. Ghanbari, Very low bit rate video coding using virtual SPIHT, *Electron. Lett.* **37**(1): 40–42 (Jan. 2001).
27. B.-J. Kim, Z. Xiong, and W. A. Pearlman, Low bit-rate scalable video coding with 3-d set partitioning in hierarchical trees (3-D SPIHT), *IEEE Trans. Circuits Syst. Video Technol.* **10**(8): 1374–1387 (Dec. 2000).
28. Joint Photographic Experts Group, *JPEG-LS Final Committee Draft*, ISO/IEC JTC1/SC29/WG1 FCD14495-1, 1997.
29. C. Christopoulos, A. Skodras, and T. Ebrahimi, The JPEG-2000 still image coding system: An overview, *IEEE Trans. Consumer Electron.* **46**(4): 1103–1127 (Nov. 2000).
30. J. H. Kasner, M. W. Marcellin, and B. R. Hunt, Universal trellis coded quantization, *IEEE Trans. Image Process.* **8**(12): 1677–1687 (Dec. 1999).
31. D. LeGall and A. Tabatabai, Sub-band coding of digital images using short kernel filters and arithmetic coding techniques, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1988, pp. 761–764.
32. D. Taubmann, High performance scalable image compression with EBCOT, *IEEE Trans. Image Process.* **9**(7): 1158–1170 (July 2000).
33. Joint Photographic Experts Group, *JPEG2000 Verification Model 8*, ISO/IEC JTC1/SC29/WG1 N1822, July 2000.

IMAGE PROCESSING

MAJA BYSTROM
Drexel University
Philadelphia, Pennsylvania

Image processing for telecommunications is a broad field that can be divided into four general categories: acquisition, analysis, compression, and reconstruction. Images can likewise be classified by their color content into binary, monochrome, or color images, or by their method of acquisition or generation, such as natural, computer-generated, radar, or ultrasonic. Following acquisition, image data may be processed for efficient storage or representation. Image analysis may be employed to extract desired information for compression or further processing. Reconstruction or enhancement is posttransmission or postacquisition processing to recover lost or degraded data or to emphasize visually important image components. More recently significant attention has focused on the related field of digital watermarking or data hiding, in which marks or identification patterns are embedded in images for security purposes.

1. IMAGE ACQUISITION AND REPRESENTATION

Images are acquired through a variety of methods such as analog or digital cameras, radar, or sonar. Regardless of the method of image generation, in order to provide for digital transmission and storage, all input analog signals must be discretized and quantized. It is well known that for perfect reconstruction, images must be sampled above the Nyquist rate, specifically, twice the greatest frequency in a band-limited signal, and infinite-duration interpolation functions must be employed. In practice, however,

infinite-duration functions are infeasible and images are often represented with fewer than the optimum number of samples for conservation of storage space or transmission bandwidth. Subsampling of images reduces the number of picture elements, *pixels*, used to represent an image. However, interpolation of a subsampled image may result in visually apparent degradation, typically in the form of blockiness or blurring. Jain discusses image sampling and basic interpolation functions [1].

Following sampling, an image is represented by a two-dimensional signal denoted by $x_{i,j} = x(i,j)$; $i = 1 \cdots V$, $j = 1 \cdots H$, where V and H are the vertical and horizontal length in pixels, respectively. By convention, the upper left corner of the image is pixel $x_{1,1}$. The pixel values are continuous and must be quantized to further limit storage requirements. The most basic quantizer is the uniform quantizer in which the continuous range of values is subdivided into a finite number of equal-length intervals. All pixels with values falling within each interval are then assigned the same value. If the quantization is fine, that is, if a large number of quantization intervals is employed, then no subjective degradation will be apparent. In practice, sample values often have a nonuniform distribution such as Laplacian or Gaussian. In this case, better subjective results may be obtained with a nonuniform quantizer and quantization intervals are determined by the Lloyd–Max algorithm.

There is a significant trade-off between the storage requirement, which is a function of the number of quantization intervals, and the subjective quality of the image. The shades of gray in black-and-white (B/W) images are typically represented by 256 quantization levels. The storage requirement is then $\log_2 256 = 8$ bits per pixel. Thus, even a small image requires significant memory for storage. However, reducing the representation to 7 bits per pixel may result in loss of small objects in the image [2].

Color images can be represented in many color spaces. One of the most frequently used color spaces in image processing is the RGB color space, which indicates the proportion of the red, green, and blue components. The value of an image pixel, $x_{i,j}$, is then a vector in three dimensions. Each vector component assumes a value in the range $[0, max]$, where max is typically normalized to 2^n for n quantization levels. For full-color RGB each color is represented by 8 bits for a total of 24 bits or 2^{24} color combinations; this number of colors is significantly more than the human eye can recognize. Other, more visually intuitive, color spaces such as hue, saturation, and intensity (HSI) or hue, lightness, and saturation (HLS) can be employed as well.

A drawback of the RGB colorspace is that for natural images, there is significant correlation between the color components. Other spaces exploit this correlation and thus are more common for applications that require efficient color representation. In the YIQ space used in North American television, the YUV color space used in European television system, and the YCbCr used in image and video compression standards, there is a luminance or B/W component and two chrominance components.

To further reduce the storage space or transmission bandwidth required for each image, the chrominance components can be subsampled, typically by a factor of 2 in each direction, with little loss in the subjective quality.

2. FILTERING AND MORPHOLOGICAL OPERATORS

Sampling and interpolation are two examples of image filtering. Many other image processing applications, such as object identification, edge enhancement, or artifact removal, also rely on filtering. Given an original image, x, the filtering operation is

$$\hat{x}_{i,j} = \sum_{k,l \in \mathcal{N}} W_{k,l} x_{i+k,j+l}$$

where $\mathcal{N}$ is a set of pixels in the designated neighborhood of the (i,j)th pixel, $W_{k,l}$ is the kernel, or weighting function, and $\hat{x}$ is the filtered image. The neighborhood, $\mathcal{N}$, can be as small as one pixel or as large as the entire image. In the case of sampling, the kernel is a two-dimensional comb function; for interpolation, the kernel can be a two-dimensional rectangle or sinc function. Appropriate filter shapes for particular applications will be mentioned in the following sections.

For a class of images, the binary images in which pixels can be only black or white, morphological operators can be a valuable tool for many of the analysis and reconstruction processing operations. Morphological operators essentially perform filtering with different shape and size kernels to switch the binary value of the pixel, based on the pixel's neighbors. Through iterations of erosion, the shrinking of an object, or dilation, the expanding of an object, and other image operations, such as addition or subtraction, this class of operators can enhance lines, remove noise, and segment images. Morphological operators have been extended to grayscale images. Dougherty gives an introduction to this class of filters [3].

3. IMAGE TRANSFORMS

Often image processing applications, such as compression and reconstruction, are performed in a transform domain. The three most commonly used transforms are the discrete Fourier transform (DFT), the discrete cosine transform (DCT), and the discrete wavelet transform (DWT). The discrete Fourier transform over an $N \times N$ region of image pixels is given by

$$X_{k,l} = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} x_{m,n} e^{-j2\pi(mk+nl)/N}$$

while the inverse DFT is given by

$$x_{m,n} = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} X_{k,l} e^{j2\pi(mk+nl)/N}.$$

The DCT is more commonly used in image and video compression standards than the DFT, since it has excellent energy compaction properties and has performance close

to the optimal Karhunen–Loeve transform. Given a pixel block of size $N \times N$, the DCT is given by

$$X_{k,l} = \frac{2}{N} c(k)c(l) \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} x_{m,n} \cos\left(\frac{k\pi(2m+1)}{2N}\right) \times \cos\left(\frac{l\pi(2n+1)}{2N}\right)$$

while the inverse transform is given by

$$x_{m,n} = \frac{2}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} c(k)c(l)X_{k,l} \cos\left(\frac{k\pi(2m+1)}{2N}\right) \times \cos\left(\frac{l\pi(2n+1)}{2N}\right)$$

where $c(i) = 1/\sqrt{2}$ for $i = 0$ and $c(i) = 1$ otherwise. The DCT basis functions are shown in Fig. 1.

A final transform currently under investigation for many image processing applications and that will form the basis of the JPEG-2000 image compression standard is the DWT [4]. The DWT utilizes a combination of lowpass and highpass filters with downsampling and interpolation to separate an image into frequency bands that may be processed independently. The wavelet decomposition or analysis process is performed by filtering first along the rows and then the columns of an image and downsampling by a factor of 2. A two-level decomposition is given in Fig. 2. Note that each subband contains different frequency information. The lowest subband, in the upper left of the figure, contains the lowest frequencies, and appears as a smoothed version of the original image. Since this is a two-level transform, the lowest subband has been downsampled twice and thus is one-sixteenth the size of the original image.

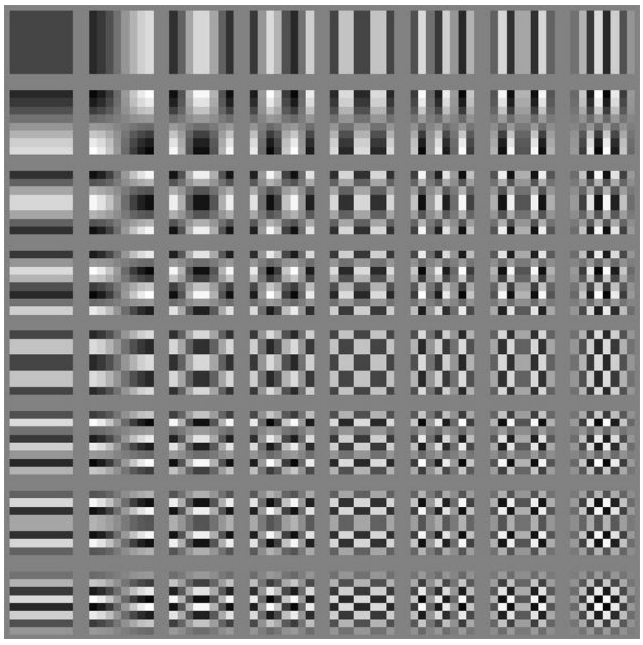


Figure 1. The DCT basis functions for an 8×8 transformation.

Synthesis, or reconstruction, is performed in the opposite manner. The synthesis filters are quadrature mirror filters designed to cancel the aliasing effects of the interpolation process. With appropriate filter design, that is, choice of wavelet basis, applications such as compression can produce perceptually better results than can DCT-based compression at the same rate, since the transform can efficiently act on large regions of the image.

4. IMAGE ANALYSIS

Image analysis is used to provide information about the image under consideration. This information could range from identification of a target in an image, segmentation of an object from a scene, or motion tracking of an object between video frames. Applications range from military target tracking, face recognition for security purposes, segmentation for compression, or object recognition for digital library indexing.

Object matching for target tracking and general object recognition is a challenging field because of the possibility of object movement, rotation, shape change, occlusion by other objects, and noise and clutter in the image. Typically for these applications, landmarks on the object are determined, or a model — based on pixel intensity, statistics, or other image features — is developed. The object is matched to known references taking into account the possible size and shape variations in the target as well as the other possible image degradation. Discussions of object recognition and target tracking are available in the literature [5–7].

Images may be segmented for applications such as efficient encoding or target location. For encoding, the foreground or more important objects are located and compressed less to maintain better quality. In target identification, targets are recognized from an often noisy and cluttered image and separated from the background. A number of segmentation methods ranging from segmentation on the basis of texture or color discrimination, motion between video frames, region growing from a starting point in a readily identifiable area, to boundary, feature, or edge identification are currently utilized. Often these methods are combined for better performance. Segmentation either is performed automatically by an algorithm or may be supervised by a user who provides input such as a starting point in a region or the number of objects to be located. An ongoing research focus is the development of unsupervised segmentation algorithms that extract perceptually important objects.

5. IMAGE COMPRESSION

Compression of an image involves reducing redundancy in the image through either lossless or lossy means. In lossy compression information is discarded through quantization or many-to-one mappings of component values. Lossy compression is typically used in applications such as broadcast images or video, since significant compression gains can be achieved at the expense of image quality. On the other hand, lossless compression is an

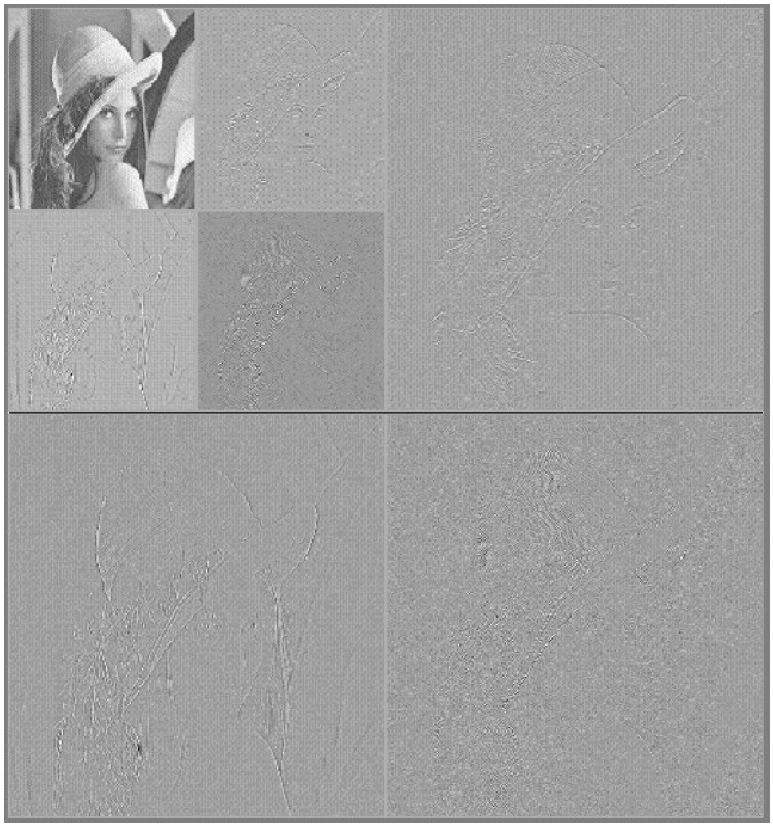


Figure 2. A two-level wavelet decomposition with the Haar filter.

invertible transform that preserves all image information and is employed on images in which all information is valuable, such as security, medical, or technical images.

Lossless compression typically involves some form of entropy coding such as Huffman or arithmetic coding. Image symbols are assigned codewords on the basis of their probabilities; more likely symbols are assigned shorter codewords. Often a form of prediction is employed as well. For example, pixel values can be predicted on the basis of their neighbors' values. The difference between the prediction and the actual pixel value is then entropy-coded. Naturally, these methods only work well if the symbol probability estimation and the prediction are good.

Lossy standards such as JPEG, JPEG-2000, MPEG-2, and MPEG-4 (JPEG — Joint Photographic Experts Group; MPEG — Moving Picture Experts Group) rely on a combination of either the DCT or DWT, quantization, and entropy coding in order to compress images or video. In block-based techniques, images are divided into blocks of pixels. These blocks are DCT transformed and the resulting coefficients are quantized and either Huffman or arithmetic-coded. Thus, spatial redundancy is taken advantage of and high frequencies are removed from the image. For video coding, temporal redundancy is utilized. Blocks of pixels in neighboring images tend to be similar, so for many blocks, a motion vector can be determined. This motion vector indicates the displacement of a block to a similar block in a neighboring frame. The difference between the two blocks is then transformed and coded.

The more recently developed lossy standards are wavelet-based. Because the lowest subband is a smoothed version of the original image, the higher subbands can be discarded or coarsely quantized with little effect on the reconstructed image quality. The wavelet coefficients are quantized and coded independently for each subband. The quantization of coefficients can be controlled by an algorithm such as the embedded zero-tree wavelet algorithm, which iteratively determines which coefficients are most significant and increases the precision of these coefficients. These wavelet compression techniques are flexible in that they permit coding of nonsquare regions of interest. Thus, arbitrarily shaped objects in an image may be segmented by an analysis algorithm and more important objects compressed less than others for better subjective quality. Future compression gains will likely be made through improved segmentation algorithms and region-of-interest compression.

6. IMAGE ENHANCEMENT AND RECONSTRUCTION

Because the human visual system is more sensitive to certain image components than others, the more perceptually important components of an image can be enhanced, often at the expense of the others, in order to provide a subjectively higher-quality image. Either point operations or histogram equalization can be used to modify the contrast and brightness of an image in order to enhance or emphasize objects that are washed out or hidden in a dark region. Contrast enhancement, either contrast stretching or brightness enhancement, is performed by manipulating pixel grayscale levels, that is, by nonlinear or linear mappings of grayscale values.

If the slope of a linear mapping is negative, then the pixel values are reversed and the mapping will result in a negative image. A similar process, histogram equalization, which stretches an image histogram so that it spans the color value range, will also result in the increase of image contrast.

Since the human visual system is sensitive to image edges, these high-frequency image components can be enhanced, typically by taking a local directional derivative, in order to make the image appear sharper. The Laplacian operating on the image luminance is an approximation to the second derivative of the brightness. It acts as a highpass filter, by increasing the contrast at distinct lines in images and setting pixels in homogeneous areas to zero. For edges that are step functions or plateaus in brightness, the Roberts and Sobel operators result in better performance than does the Laplacian operator. These filters approximate the first derivative and are applied oriented in multiple directions to capture diagonal edges.

Following compression, transmission over a noisy or fading channel, or capture by imperfect methods, there may be degradation in the received image. This resulting degradation may result from noise or speckling, coding artifacts, or data loss. Figure 3 illustrates three types of degradation. Error reconstruction or concealment measures must be taken to restore these degraded images.

There are various methods for noise removal; perhaps the simplest is lowpass filtering or smoothing by neighborhood averaging in the spatial domain. Common kernels employed in the filtering operation are uniform, Gaussian, and Savitsky–Golay. More complex iterative or adaptive noise and blur removal methods can be employed as well. Both stochastic and deterministic filtering and estimation have been performed with much success [8,9].

Low-bit-rate compression by standard methods often results in either block artifacts or ringing due to quantization of transform coefficients. An example of blockiness resulting from JPEG compression is shown in Fig. 3c. Techniques for compensating for this type of degradation are local adaptive filtering in the spatial or frequency domains. Molina et al. provide a survey of deblocking techniques [10].

When block-based coding schemes are employed, channel errors may affect the compressed bitstream and cause the decoder to lose synchronization and thus lose an entire block or multiple blocks. A typical example is shown in the left side of Fig. 3d. For intraimage recovery, typically some form of replacement or smoothing is employed. Blocks of pixels can be replaced by surrounding blocks, or interpolated from surrounding pixels, taking

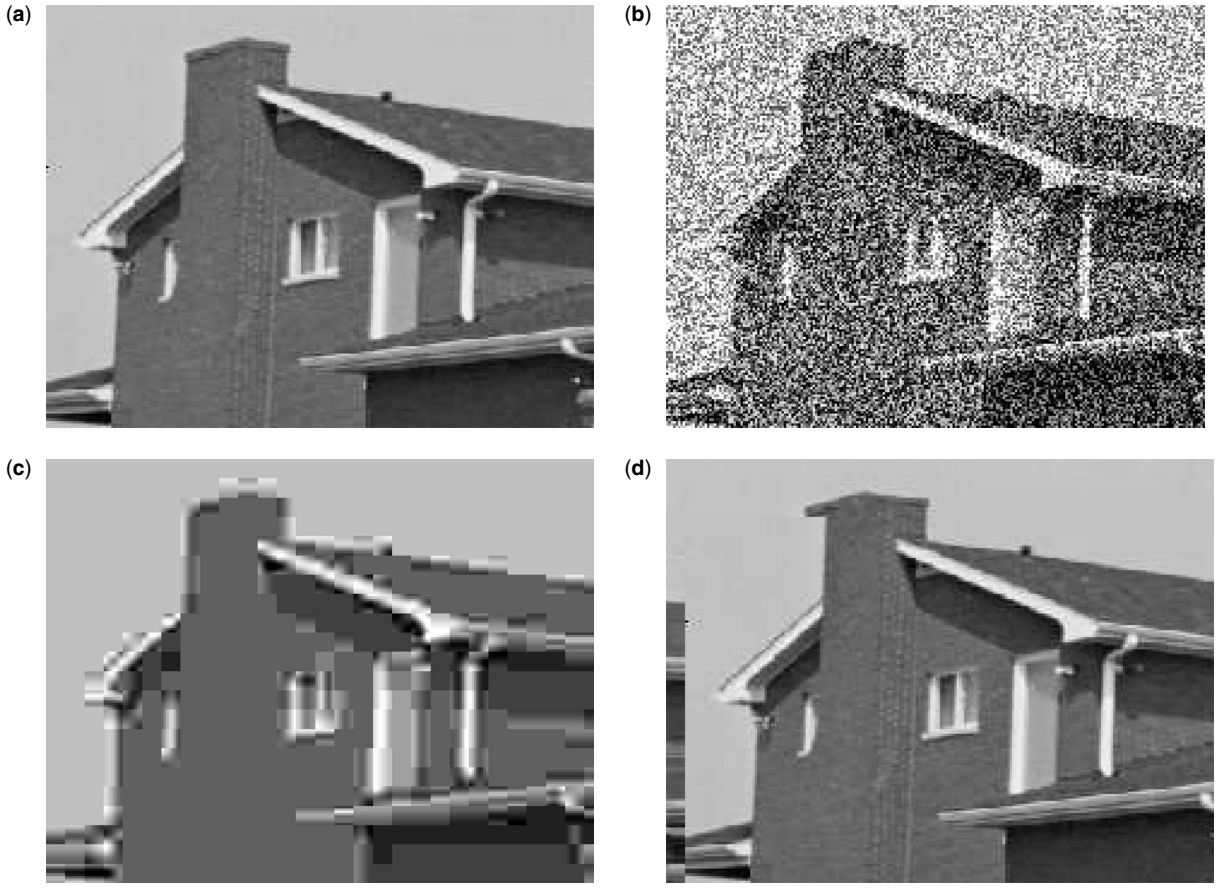


Figure 3. Selected examples of image degradation. (**a**) original image; (**b**) image with noise; (**c**) image with block artifacts; (**d**) image with block loss.

into consideration image features such as lines. If the errors arise in coded video, a combination of temporal and spatial error concealment is often employed.

7. QUALITY EVALUATION

A measure of image quality is vital for either evaluation of the appropriateness of an image for a particular application, or for evaluation of the effects of processing on an image. These measures differ broadly depending on the method of image acquisition and the image processing application. The simplest metrics are objective measures. For radar images, measures include the clutter : noise ratio, the resolution, and metrics of additive and multiplicative noise [11]. If, following image processing, the original image, x, is available for comparison, the quality of a reconstructed image denoted by $\hat{x}$ can be measured by one of four common distortion metrics:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{i=1}^{V}\sum_{j=1}^{H}x_{i,j}^2}{\sum_{i=1}^{V}\sum_{j=1}^{H}(x_{i,j}-\hat{x}_{i,j})^2}$$

$$\mathrm{PSNR} = 10\log_{10}\frac{(max x_{i,j} x_{i,j})^2}{\sum_{i=1}^{V}\sum_{j=1}^{H}(x_{i,j}-\hat{x}_{i,j})^2}$$

$$\mathrm{MSE} = \frac{1}{HV}\sum_{i=1}^{V}\sum_{j=1}^{H}(x_{i,j}-\hat{x}_{i,j})^2$$

$$\mathrm{MAD} = \frac{1}{HV}\sum_{i=1}^{V}\sum_{j=1}^{H}|x_{i,j}-\hat{x}_{i,j}|$$

where SNR = signal-to-noise ratio, PSNR = peak signal-to-noise ratio, MSE = mean-squared error, and MAD = mean absolute difference.

A more intuitive measure of quality would be a subjective measure. The (U.S.) National Image Interpretability Rating Scale (NIIRS) is a 10-level scale used to rate images for their usefulness in terms of resolution and clarity [12]. This scale is typically used to evaluate images acquired from airborne or space-based systems. For compression and enhancement applications a quality measure such as the five-point ITU-R BT.500-10 scale can be employed. Subjective quality is determined by taking a large sample of evaluations; that is, a large number of viewers rate the image on a selected scale, and the result is averaged. However, care must be taken in selecting evaluators, since trained viewers tend to notice different effects than do novices.

8. DIGITAL WATERMARKING AND DATA HIDING

With the growth in image transmission, reproduction, and storage capabilities, digital image information hiding and watermarking have become necessary tools for maintaining security and proving ownership of images. The primary goal in image watermarking and data hiding is to embed into an image an identifying mark that may or may not be readily identifiable to a viewer, but is easily detectable either when compared with the original image or with knowledge of a key. Typically, the watermark is spread throughout the image with the use of a pseudonoise key. Keys may be private or public; however, public keys permit the deletion of watermarks since the user can readily identify the location of the watermark and the method of watermarking. Because images may be stored, transmitted, copied, or printed, the watermark must be robust in the face of scanning, faxing, compression/decompression, transmission over a noisy channel, and the further addition of watermarks [13].

To effectively and imperceptibly embed data within an image, image components must be modified only slightly through the use of a key known to the watermark generator. Since many image operations, such as compression or filtering, tend to remove image components that are perceptually insignificant, the watermark must be embedded into perceptually important components. The image data are divided into perceptual components, such as the frequency or color components, and the watermark is embedded into one or more components depending on the components' robustness to distortion. Watermarking or data hiding may be performed in the spatial or transform domains. In the spatial domain one common watermarking technique is to amplitude-modulate a regular pattern of blocks of pixels by a small amount another is to increment or decrement the means of blocks. In the transform domain, the DFT, DCT, and DWT techniques are commonly used and information is embedded in the transform phase or by imposing relationships between transform coefficients.

BIOGRAPHY

Maja Bystrom received a B.S. in computer science and a B.S. in communications from Rensselaer Polytechnic Institute, Troy, New York, in 1991. She joined NASA-Goddard Space Flight Center where she worked as a computer engineer until August 1992. She then returned to Rensselaer where she received M.S. degrees in electrical engineering and mathematics, and a Ph.D. in electrical engineering. In 1997 she joined Drexel University, Philadelphia, Pennsylvania, as an assistant professor. She has received NSF CAREER and Fulbright awards. Her research interests are image processing for communications, and joint source-channel coding/decoding.

BIBLIOGRAPHY

1. A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
2. M. A. Sid-Ahmed, *Image Processing: Theory, Algorithms, & Architectures*, McGraw-Hill, New York, 1995.
3. E. R. Dougherty, *An Introduction to Morphological Image Processing*, SPIE Optical Engineering Press, Bellingham, WA, 1992.
4. G. Strang and T. Nguyen, *Wavelets and Filter Banks*, 2nd ed., Wellesley-Cambridge Press, Wellesley, MA, 1997.

5. R. Nitzberg, *Radar Signal Processing and Adaptive Systems*, Artech House, Boston, 1999.

6. H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, eds., *Face Recognition: From Theory to Applications*, Springer-Verlag, Berlin, 1998.

7. J. Weng, T. S. Huang, and N. Ahuja, *Motion and Structure from Image Sequences*, Springer-Verlag, Berlin, 1993.

8. A. K. Katsaggelos, ed., *Digital Image Restoration*, Springer-Verlag, Berlin, 1991.

9. G. Demoment, Image reconstruction and restoration: Overview of common estimation structures and problems, *IEEE Trans. Acoust. Speech Signal Process.* **37**: 2024–2036 (Dec. 1989).

10. R. Molina, A. K. Katsaggelos, and J. Mateos, Removal of blocking artifacts using a hierarchical bayesian approach, in A. Katsaggelos and N. Galatsanos, eds., *Signal Recovery Techniques for Image and Video Compression*, Kluwer, Boston, 1998.

11. W. G. Carrara, R. S. Goodman, and R. M. Majewski, *Spotlight Synthetic Aperture Radar: Signal Processing Algorithms*, Artech House, Boston, 1995.

12. J. Pike, National image interpretability rating scales (Jan. 10, 1998), Federation of American Scientists, *http://www.fas.org/irp/imint/niirs.htm* (posted Aug. 10, 2000); accessed 8/10/2000, updated 1/16/98.

13. S. Katzenbeisser and F. A. P. Petitcolas, eds., *Information Hiding Techniques for Stegnography and Digital Watermarking*, Artech House, Boston, 2000.

FURTHER READING

Aign S., Error concealment for MPEG-2 video, in A. Katsaggelos and N. Galatsanos, eds., *Signal Recovery Techniques for Image and Video Compression*, Kluwer, Boston, 1998.

Bhaskaran V. and K. Konstantinides, *Image and Video Compression Standards: Algorithms and Architectures*, Kluwer, Boston, 1997.

Bowyer K. and N. Ahuja, eds., *Advances in Image Understanding*, IEEE Computer Society Press, Los Alamitos, CA, 1996.

Giorgianni E. and T. Madden, *Digital Color Management: Encoding Solutions*, Addison-Wesley, Reading, MA, 1998.

Home site of the JPEG and JBIG committees, *http://www.jpeg.org* (Aug. 10, 2000).

Netravali A. and B. Haskell, *Digital Pictures: Representation, Compression and Standards*, 2nd ed., Plenum Press, New York, 1995.

Lindley C. A., *Practical Image Processing in C*, Wiley, New York, 1991.

Pennebaker W. B. and J. L. Mitchell, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, 1993.

Rihaczek A. and S. Hershkowitz, *Radar Resolution and Complex-Image Analysis*, Artech House, Boston, 1996.

Russ J. C., *The Image Processing Handbook*, 2nd ed., CRC Press, Boca Raton, FL, 1995.

Sangwine S. J. and R. E. N. Horne, eds., *The Colour Image Processing Handbook*, Chapman & Hall, London, 1998.

Sezan M. I. and A. M. Tekalp, Survey of recent developments in digital image restoration, *Opt. Eng.* **29**: (May 1990).

Vetterli M. and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Upper Saddle River, NJ, 1995.

IMAGE SAMPLING AND RECONSTRUCTION

H. J. TRUSSELL
North Carolina State University
Raleigh, North Carolina

1. INTRODUCTION

Images are the result of a spatial distribution of radiant energy. We see, record, and create images. The most common images are two-dimensional color images seen on television. Other everyday images include photographs, magazine and newspaper pictures, computer monitors, and motion pictures. Most of these images represent realistic or abstract versions of the real world. Medical and satellite images form classes of images where there is no equivalent scene in the physical world. Computer animation produces images that exit only in the mind of the graphic artist.

In the case of continuous variables of space, time, and wavelength, an image is described by a function

$$f(x, y, \lambda, t) \tag{1}$$

where x, y are spatial coordinates (angular coordinates can also be used), λ indicates the wavelength of the radiation, and t represents time. It is noted that images are inherently two-dimensional (2D) spatial distributions. Higher-dimensional functions can be represented by a straightforward extension. Such applications include medical CT and MRI, as well as seismic surveys. For this article, we will concentrate on the spatial and wavelength variables associated with still images. The temporal coordinate will be left for another chapter.

In order to process images on computers, the images must be sampled to create digital images. This represents a transformation from the analog domain to the discrete domain. In order to view or display the processed images, the discrete image must be transformed back into the analog domain. This article concentrates entirely on the very basic steps of sampling an image in preparation for processing and reconstructing or displaying an image. This may seem to be a very limited topic but let us consider what will not be covered in this limited space.

We introduced images as distributions of radiant energy. The exact representation of this energy and its measurement is the subject of radiometry. For this article, we will ignore the physical representation of the radiant source. We will treat the image as if everyone knows what the value of $f(x, y)$ means and how to interpret the two-dimensional gray-level distributions that will be used in this chapter to demonstrate various principles.

If we include the frequency or wavelength distribution of the energy, we can discuss spectrometry. Images for most consumer and commercial uses are the color images that we see everyday. These images are transformations of continuously varying spectral, temporal and spatial distributions. In order to fully understand the effects of sampling and reconstruction of color images, more understanding of the human visual system is required than can be presented here. Satellite images are now being recorded

in multispectral and hyperspectral bands. In this terminology, a hyperspectral image has more than 20 bands. We will only touch on the basics of color sampling.

All images exist in time and change with time. We're all familiar with the stroboscopic effects that we see in the movies that make car wheels and airplane propellers appear to move backward.[1] The same sampling principles can be used to explain these phenomena as will be used to explain the spatial sampling that is presented here. The description of object motion in time and its effect on images is another rich topic that will be omitted here.

Before presenting the fundamentals of image presentation, it necessary to define our notation and to review the prerequisite knowledge that is required to understand the following material. A review of rules for the display of images and functions is presented in Section 2, followed by a review of mathematical preliminaries and sampling effects in Section 3. Section 4 discusses simpling on a nonrectangular lattice. The practical case of using a finite aperture in the sampling process is presented in Section 5. Section 6 reviews color vision and describes multidimensional sampling with concentration on sampling color spectral signals. We will discuss the fundamental differences between sampling the wavelength and spatial dimensions of the multidimensional signal. Finally, Section 7 contains a mathematical description of the display of multidimensional data. This area is often neglected by many texts. The section will emphasize the requirements for displaying data in a fashion that is both accurate and effective.

2. PRELIMINARY NOTES ON DISPLAY OF IMAGES

One difference between 1D and 2D functions is the way they are displayed. One-dimensional functions are easily displayed in a graph where the scaling is obvious. The observer need examine only the numbers that label the axes to determine the scale of the graph and get a mental picture of the function. With two-dimensional scalar-valued functions, the display becomes more complicated. The accurate display of vector-valued two-dimensional functions, including color images, will be discussed after covering the necessary material on sampling and colorimetry.

Two-dimensional functions can be displayed as an isometric plot, a contour plot, or a grayscale plot. Since we are dealing with images, we will use the grayscale plot for images and the isometric plot for functions. All three types are supported by MATLAB [1]. The user should choose the right display for the information to be conveyed. For the images used in this article, we should review some basic rules for display.

Consider a monochrome image that has been digitized by some device, such as a scanner or camera. Without knowing the physical process that created the image, it is impossible to determine the best way to display the image. The proper display of images requires calibration of both the input and output devices [15,16]. This is another topic that must be omitted for lack of space. For now, it is reasonable to give some general rules about the display of monochrome images:

1. For the comparison of a sequences of images, it is *imperative* that all images be displayed using the same scaling.
2. Display a step-wedge, a strip of sequential gray levels from minimum to maximum values, with the image to show how the image gray levels are mapped to brightness or density. This allows some idea of the quantitative values associated with the pixels.
3. Use a graytone mapping which allows a wide range of gray levels to be visually distinguished. In software such as MATLAB, the user can control the mapping between the continuous values of the image and the values sent to the display device. It is recommended that adjustments be made so that a user is able to distinguish all levels of a step-wedge of about 32 levels.

3. SPATIAL SAMPLING

In most cases, the multidimensional process can be represented as a straightforward extension of one-dimensional processes. Thus, it is reasonable to mention the one-dimensional operations that are prerequisite to understanding this article and will form the basis of the mutidimensional processes.

3.1. Ideal Sampling in One Dimension

Mathematically, ideal sampling is usually represented with the use of a *generalized function*, the Dirac delta function, $\delta(t)$ [2]. The function is defined as zero for $t \neq 0$ and having an area of unity. The most useful property of the delta function is that of sifting, for instance, extracting single values of a continuous function. This is defined by the integral

$$s(t_0) = \int_{-\infty}^{\infty} s(t)\delta(t - t_0)\, dt = \int_{-\infty}^{\infty} s(t_0)\delta(t - t_0)\, dt \qquad (2)$$

This shows the production of a single sample. We would represent the sampled signal as a signal that is zero everywhere except at the sampling time, $s_{t_0}(t) = s(t_0)\delta(t - t_0)$. The sampled signal can be represented graphically by using the arrow, as shown in Fig. 1.

The entire sampled sequence can be represented using the *comb* function

$$comb(t) = \sum_{n=-\infty}^{\infty} \delta(t - n) \qquad (3)$$

where the sampling interval is unity. The sampled signal is obtained by multiplication

$$s_d(t) = s(t)comb(t) = s(t) \sum_{n=-\infty}^{\infty} \delta(t - n) = \sum_{n=-\infty}^{\infty} s(t)\delta(t - n) \qquad (4)$$

[1] We used to use the example of stagecoach wheels moving backward, but, alas, there are few Western movies anymore. Time marches on.

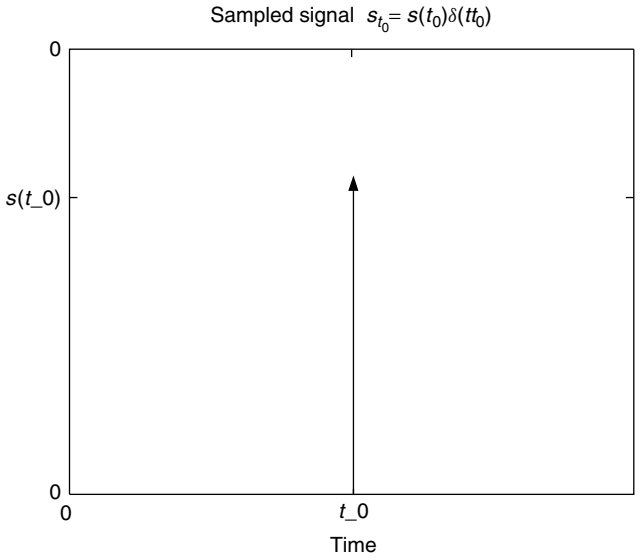


Figure 1. Sampled signal at $t = t_0$.

The sampling is represented graphically in Fig. 2. It is common to use the notation of $\{s(n)\}$ or $s(n)$ to represent the collection of samples in discrete space. The arguments n and t will serve to distinguish the discrete or continuous spaces, respectively.

The 1D effects in the frequency domain are shown in most undergraduate signals and systems texts. Briefly, we will review this graphically by considering the frequency domain representation of the signals in Fig. 2. The spectrum of the analog signal, $s(t)$ is denoted $S(\omega)$; the Fourier transform of the *comb(t)* is $2\pi comb(\omega)$.[2] The frequency-domain representation of the 1D sampling process is shown in Fig. 3. The spectra in this figure correspond to the time-domain signals in Fig. 2. The most important feature of the spectrum of the sampled signals is the replication of the analog spectrum. Mathematically, if the spectrum of $s(t)$ is denoted $S(\omega)$, then the spectrum of the sampled signal, $s_d(t)$, is given by

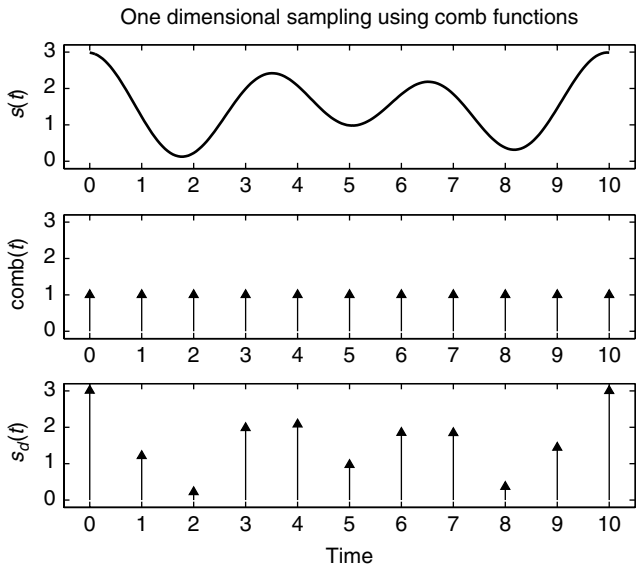


Figure 2. One-dimensional sampling.

[2] The proof of this is also available in the undergraduate signals and systems texts [e.g., 2].

$$S_d(\omega) = \sum_{k=-\infty}^{\infty} S(\omega - k2\pi F_s)$$

where F_s is the sampling rate. Note that reconstruction is possible only if there is no overlap of the replicated spectra. This, of course, corresponds to having a sampling rate that is greater than twice the highest frequency in the analog signal, $F_{\max}$, namely, $F_s > 2F_{\max}$.

From the frequency domain figures, it is easy to see that reconstruction of the original signal requires that the fundamental spectrum, the one centered at zero, be retained, while the replicated spectra by eliminated. In the time domain, this can be accomplished by passing the sampled signal through a lowpass filter. While ideal lowpass filters are not possible, it is possible to realize sufficiently good approximations that the reconstruction is close enough to ideal for practical applications. This is a major difference with two-dimensional image reproduction. There is no equivalent analog low pass filter that can be used with optical images. This will be addressed in a later section.

If the sampling rate is not adequate, then the original signal cannot be reconstructed from the sample values. This is seen by considering the samples of a sinusoid of frequency, F, which are given by

$$s_F(n) = \cos\frac{2\pi Fn}{F_s} + \theta\cos\left(\frac{2\pi Fn}{F_s} + \theta\right)$$

where F_s is the sampling rate and θ is the phase of the sinusoid. The samples are taken at $t_n = n/F_s$. We see that the samples are the same for all frequencies $F = F_m$ that are related to the sampling frequency by $F_m = F_0 + mF_s$. The samples of these sinusoids are all identical to those of the sinusoid of frequency F_0. We will refer to F_0 as an alias of the frequencies F_m under the sampling rate of F_s.

3.2. Ideal Sampling in Two Dimensions

The two-dimensional Dirac delta function can be defined as the separable product of one-dimensional delta functions, $\delta(x, y) = \delta(x)\delta(y)$. The extension of the comb function to two dimensions should probably be called a "brush," but we will continue to use the term comb and define it by

$$comb(x, y) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \delta(x - m, y - n)$$

The equation for 2D sampling is

$$s(m, n) = s_d(x, y) = s(x, y)comb(x, y) \tag{5}$$

where a normalized sampling interval of unity is assumed. We have the same constraints on the sampling rate in two

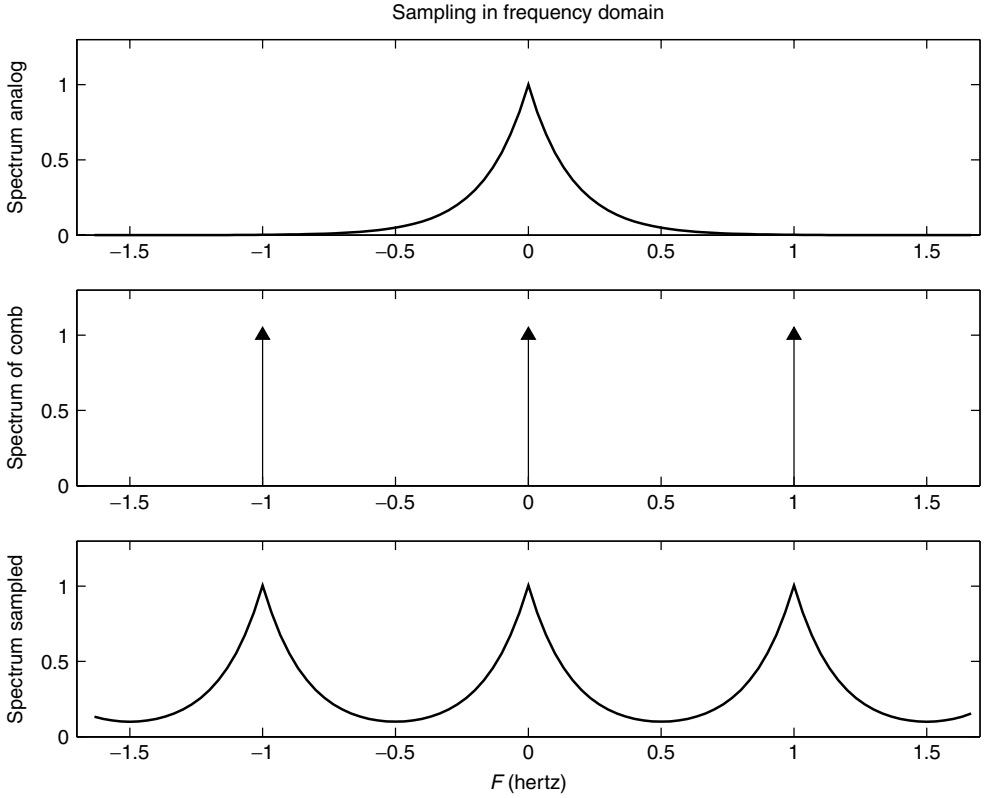


Figure 3. One-dimensional sampling: frequency domain.

dimensions as in one. Of course, the frequency is measured not in hertz, but in cycles per millimeter or inch.[3] Spatial sampling is illustrated in Figs. 4–6. Undersampling in the spatial domain signal results in spatial aliasing. This is easy to demonstrate using simple sinusoidal images. First, let us consider the mathematics.

Taking Fourier transforms of Eq. (5) yields

$$S_d(u,v) = S(u,v) * comb(u,v) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} S(u-k, v-l) \tag{6}$$

where the asterisk $(*)$ denotes convolution. Note that the sampled spectrum is periodic, as in the one-dimensional case.

Consider the effect of changing the sampling interval

$$s(m,n) = s(x,y) comb\left(\frac{x}{\Delta x}, \frac{y}{\Delta y}\right) \tag{7}$$

which yields

$$S_d(u,v) = S(u,v) * comb(u,v) = \frac{1}{|\Delta x \Delta y|} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} S(u\Delta x - k, v\Delta y - l) \tag{8}$$

Figure 7 shows the spectrum of a continuous analog image. If the image is sampled with intervals of δx in each direction, the sampled image has a spectrum that shows periodic replications at $1/\delta x$. The central portion of the periodic spectrum is shown in Fig. 8. For Fig. 8, we have used $\delta x = \frac{1}{30}$ mm.

Note that if the analog image, $s(x,y)$, is bandlimited to some 2D region, it is possible to recover the original signal from the samples by using an ideal lowpass filter. The proper sampling intervals are determined from the requirement that the region of support in the frequency domain (band limit) is contained in the rectangle defined by $|u| \leq (1/2\Delta x)$ and $|v| \leq (1/2\Delta\, y)$.

The effect of sampling can be demonstrated in the following examples. In these examples, aliasing will be demonstrated by subsampling a high-resolution digital image to produce a low resolution image. First let us consider a pure sinusoid. The function

$$s(x,y) = \cos\left[2\pi\left(\frac{36x}{128} + \frac{24y}{128}\right)\right]$$

where x is measured in mm, is sampled at 1 mm spacing in each direction. This produces no aliasing. The function and its spectrum are shown in Figs. 9 and 10, respectively.[4] Note that the frequency of the spectrum is in normalized

[3] Image processors use linear distance most often, but occasionally use angular measurement, which yields cycles per degree. This is done when considering the resolution of the eye or an optical system.

[4] The spectra of the sinusoids appear as crosses instead of points because of the truncation of the image to a finite region. The full explanation is beyond the scope of this article.

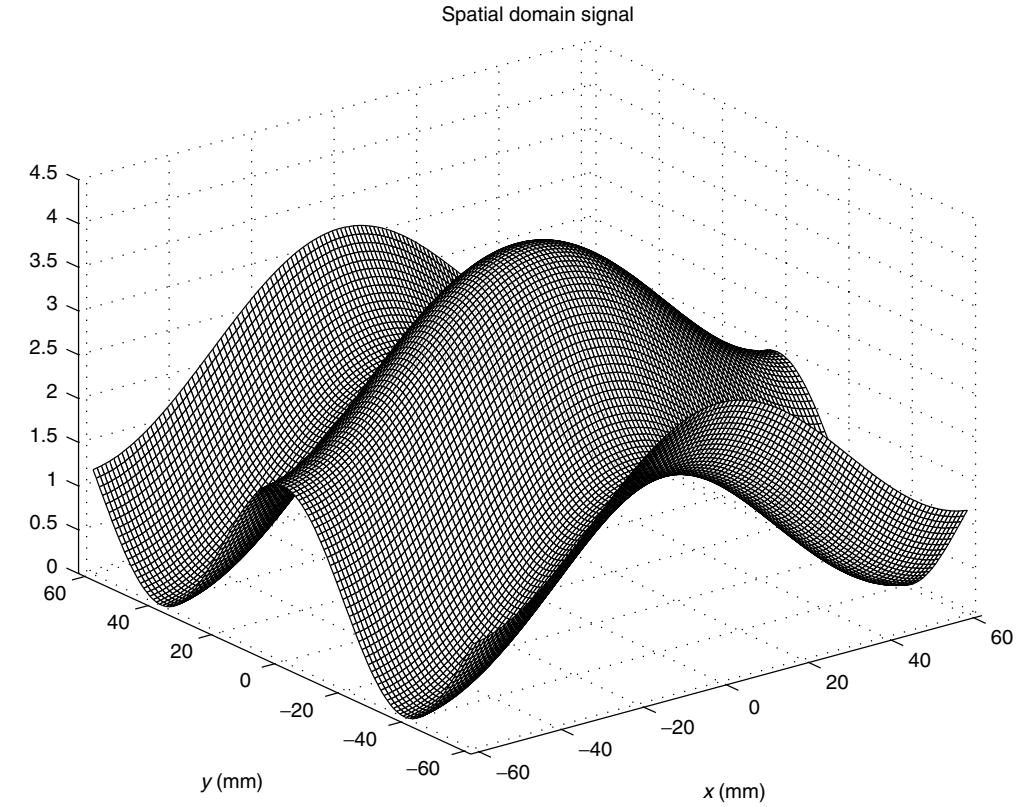


Figure 4. Two-dimensional analog signal.

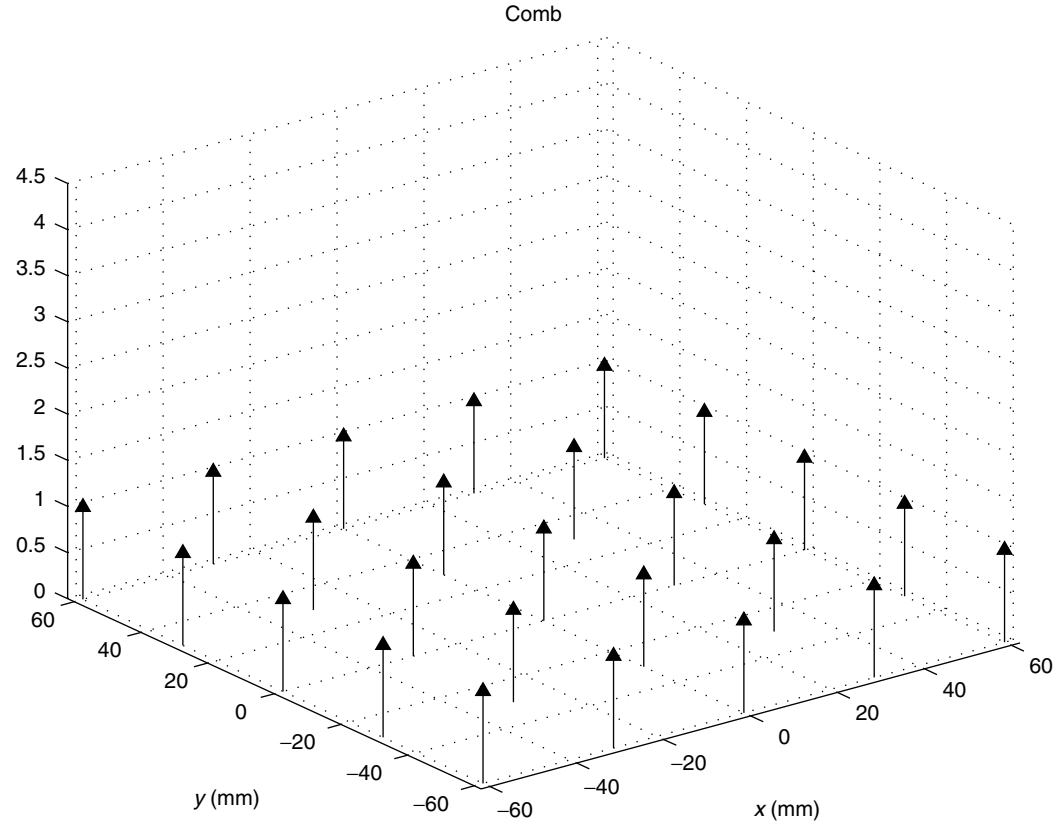


Figure 5. Two-dimensional comb.

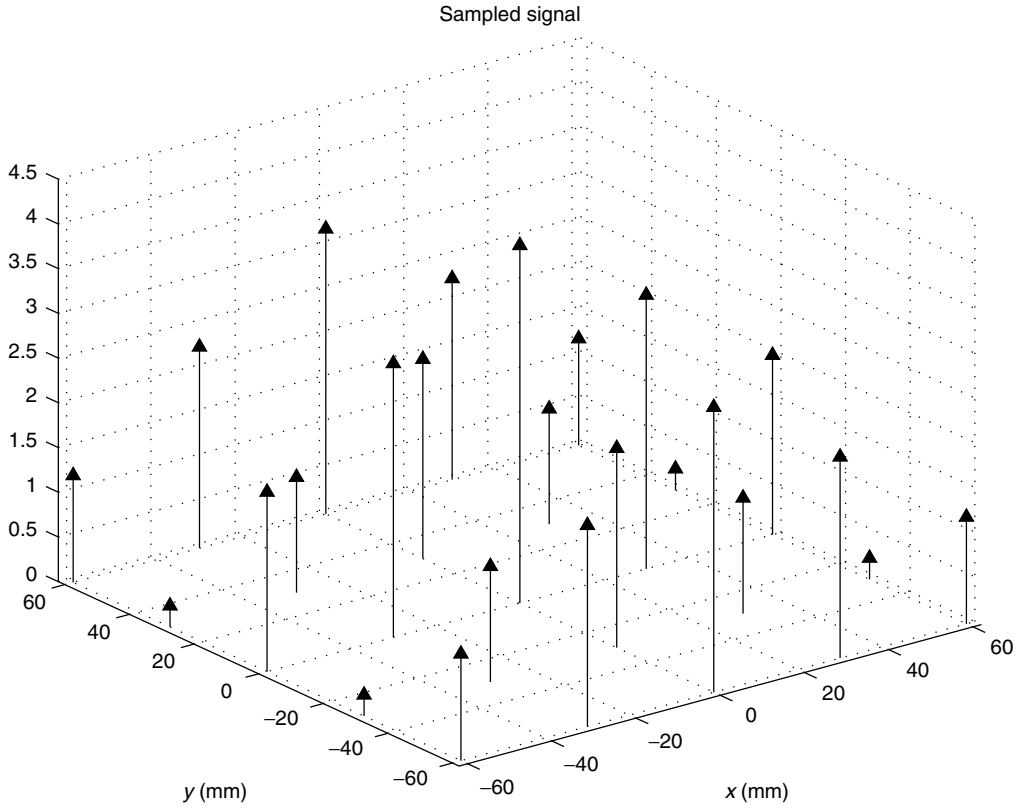


Figure 6. Two-dimensional sampled signal.

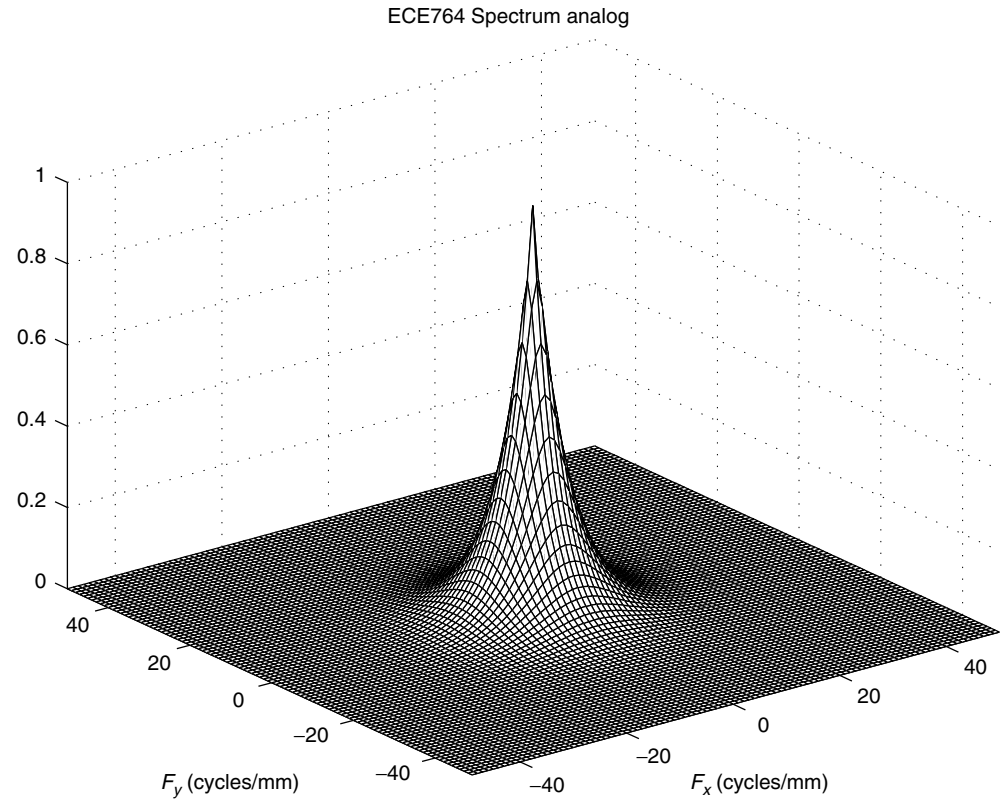


Figure 7. Analog spectrum.

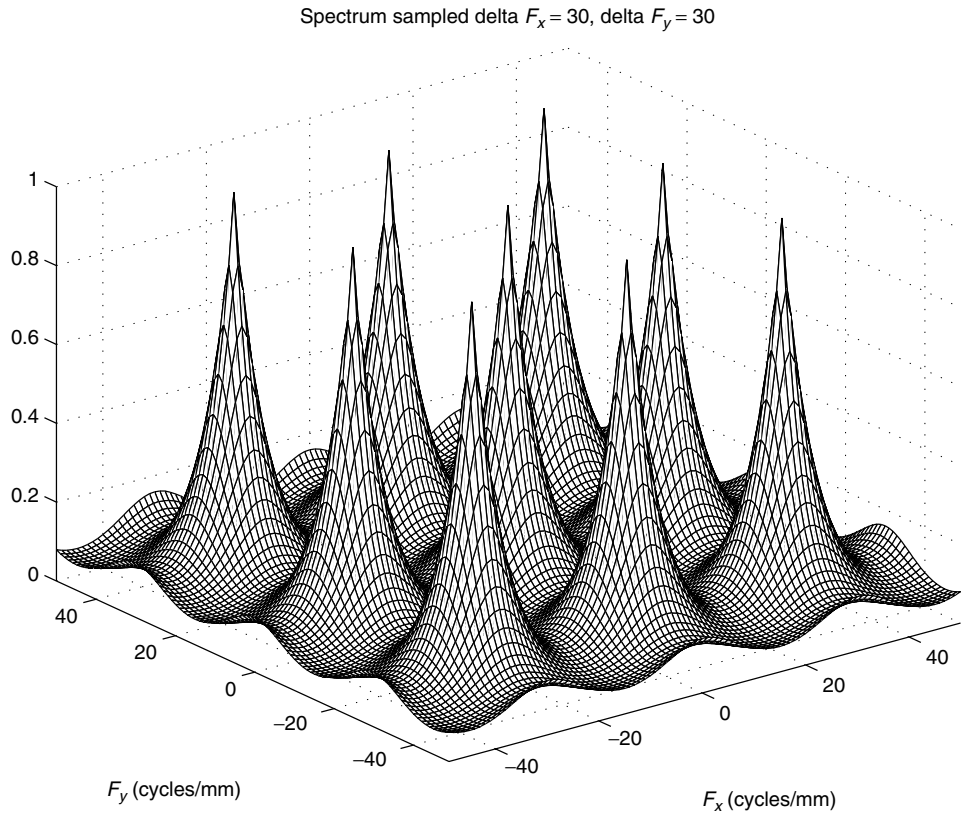


Figure 8. Digital spectrum with aliased images of analog spectrum.

digital frequency.[5] For this case, the analog frequency, F and the digital frequency, f, are the same. This yields $F_x = f_x = \frac{36}{128} = 0.28125$ and $F_y = f_4 = \frac{24}{128} = 0.1875$. The spectrum shows peaks at this 2D frequency. The image is subsampled by a factor of 4 and shown in Fig. 11. This is equivalent to a sampling of the analog signal with an interval of 4mm in each direction. The aliased 2D frequency can be found by finding k and l so that both $|F_x - kF_s| < 0.5F_s$ and $|F_y - lF_s| < 0.5F_s$ hold. For this case, $k = l = 1$ and the aliased analog 2D frequency is $(F'_x, F'_y) = (0.03125, -0.0625)$. This means that the function

$$s'(x, y) = \cos[2\pi(0.03125x - 0.0625y)]$$

will yield the same samples as $s(x, y)$ above when sampled at 4mm intervals in each direction. The spectrum of the sampled signal is shown in Fig. 12. The digital frequencies can be found by normalizing the aliased analog frequency by the sampling rate. For this case, $(f'_x, f'_y) = (0.03125/0.25, -0.0625/0.25) = (0.125, -0.25)$.

An example of sampling a pictorial image is shown in Figs. 13–16, where Fig. 13 is the original; Fig. 14 is its spectrum; Fig. 15 is a 2:1 subsampling of the original; Fig. 16 is the spectrum of the subsampled image. For this case, we can see that the lower frequencies have been preserved but the higher frequencies have been aliased.

[5] Normalized digital frequency is denoted by F and has the constraint $|F| \leq \frac{1}{2}$[2].

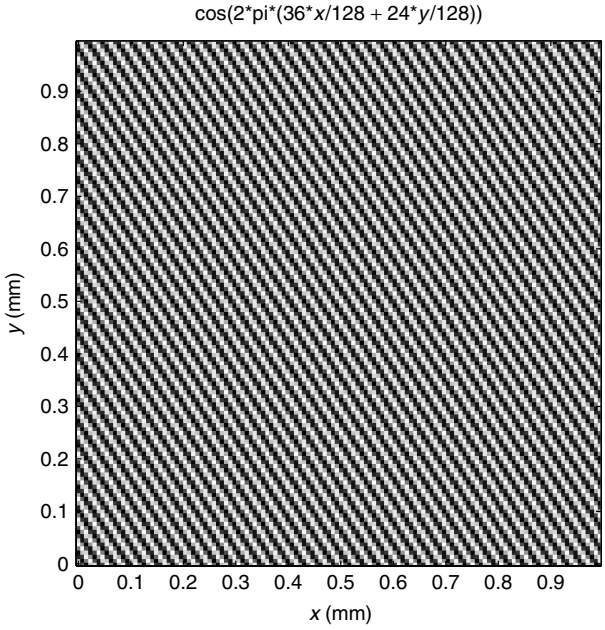


Figure 9. $\cos[2\pi(36x/128 + 24y/128)]$.

4. SAMPLING ON NONRECTANGULAR LATTICES

Because images may have oddly shaped regions of support in the frequency domain, it is often more efficient to sample with a nonrectangular lattice. A thorough discussion of this concept is found in the book by Dudgeon and Mersereau [3]. To develop this concept, it is convenient to

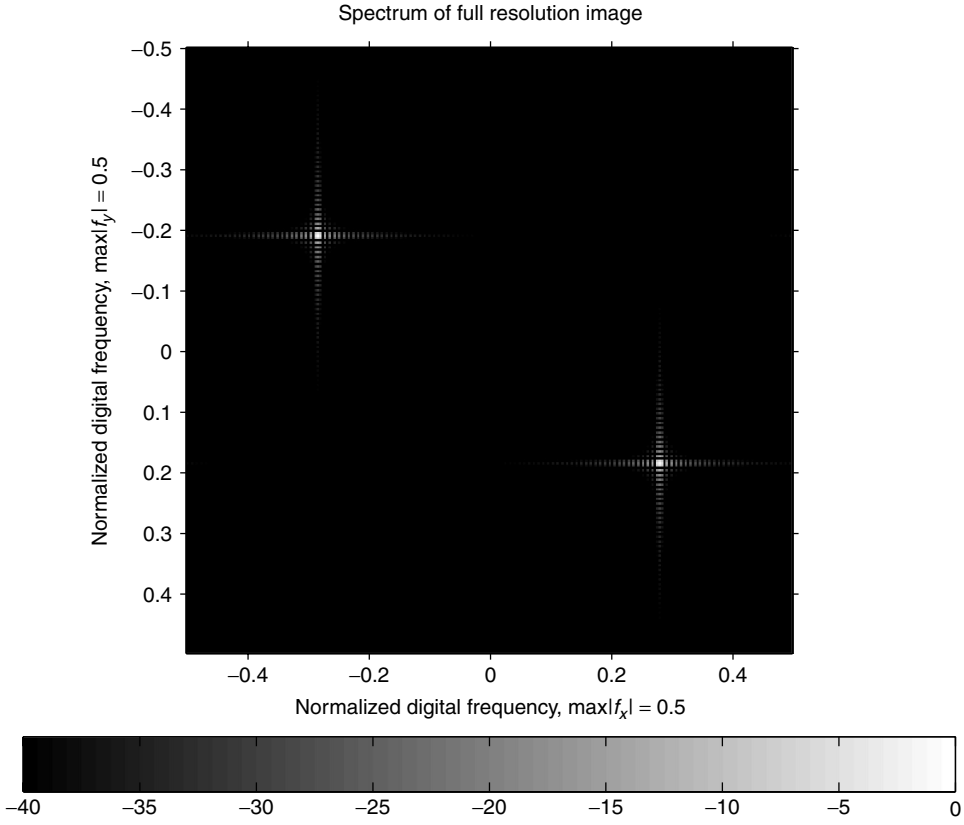


Figure 10. Spectrum of $\cos[2\pi(36x/128 + 24y/128)]$.

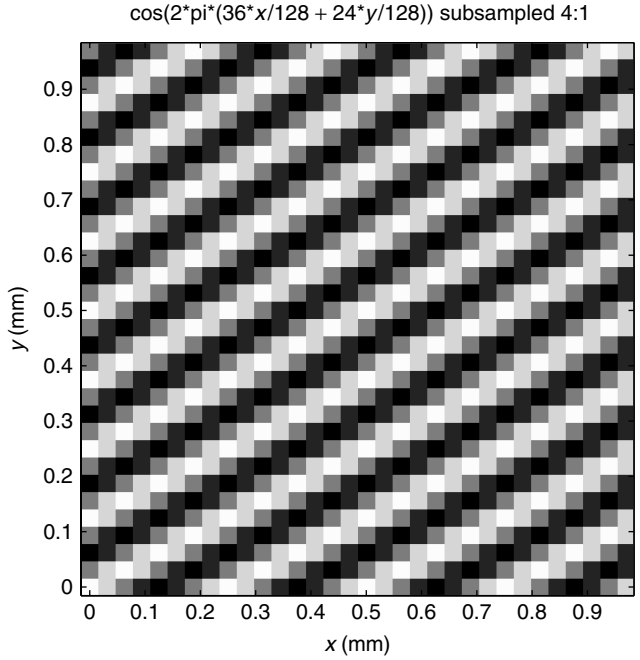


Figure 11. Sampled $\cos[2\pi(36x/128 + 24y/128)]$.

write the sampling process in vector form. Let $\mathbf{x} = [x, y]$ and the basis vectors for sampling in the space domain be given by $\mathbf{x}_1$ and $\mathbf{x}_2$. The sampling function or comb can be written

$$comb(\mathbf{r}) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(\mathbf{r} - m\mathbf{x}_1 - n\mathbf{x}_2) \tag{9}$$

This yields the functional form

$$s(m, n) = s(m\mathbf{x}_1 + n\mathbf{x}_2) \tag{10}$$

writing this in matrix form

$$s(\mathbf{n}) = s(\mathbf{X}\mathbf{n}) \tag{11}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]$ and $\mathbf{n} = [m, n]$.

The basis vectors in the frequency domain, $\mathbf{w}_1$ and $\mathbf{w}_2$, are defined by the relation

$$\mathbf{x}_k \mathbf{w}_l^T = \delta(k - l) \tag{12}$$

or using matrix notation

$$\mathbf{X}^T\mathbf{W} = \mathbf{I} \tag{13}$$

The Fourier transform in matrix notation is written

$$S(\mathbf{w}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(\mathbf{x}) \exp(-j2\pi \mathbf{W}^T\mathbf{x})\, d\mathbf{x} \tag{14}$$

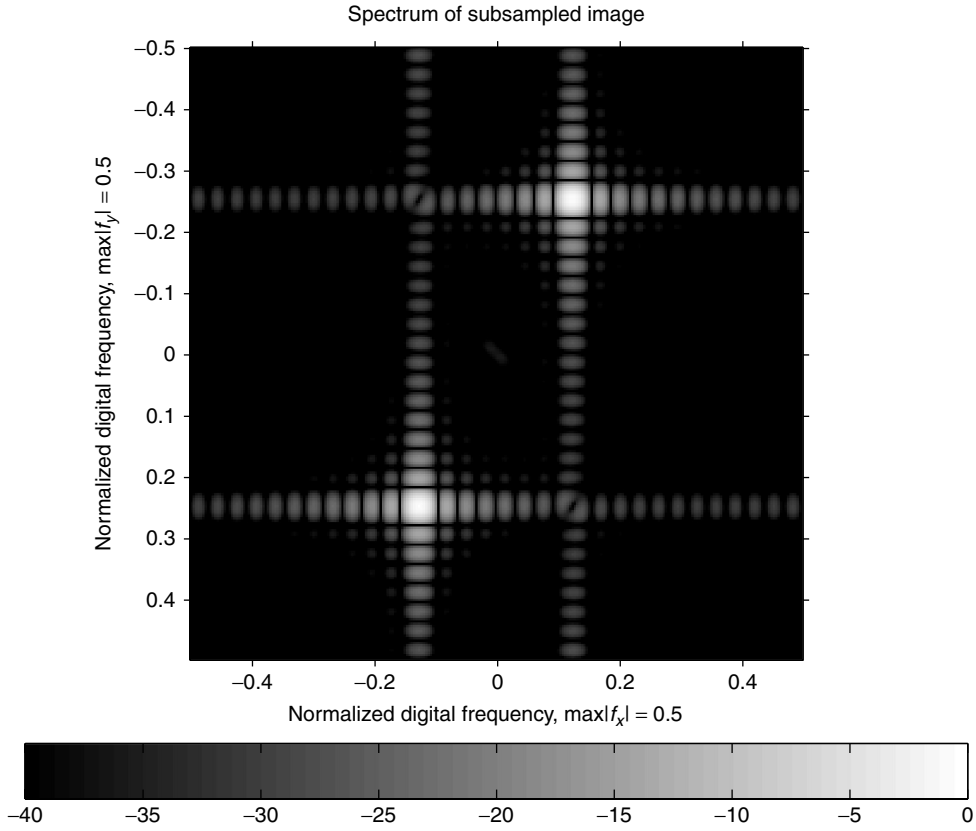


Figure 12. Spectrum of Subsampled $\cos[2\pi(36x/128 + 24y/128)]$.

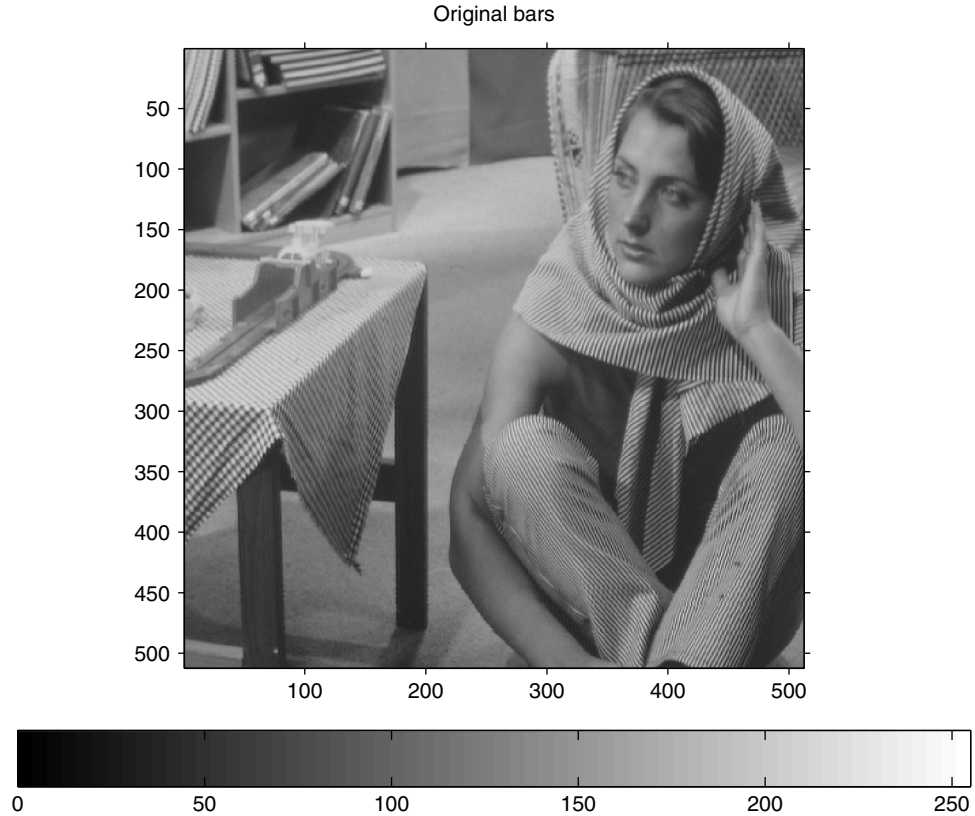


Figure 13. Original bars image.

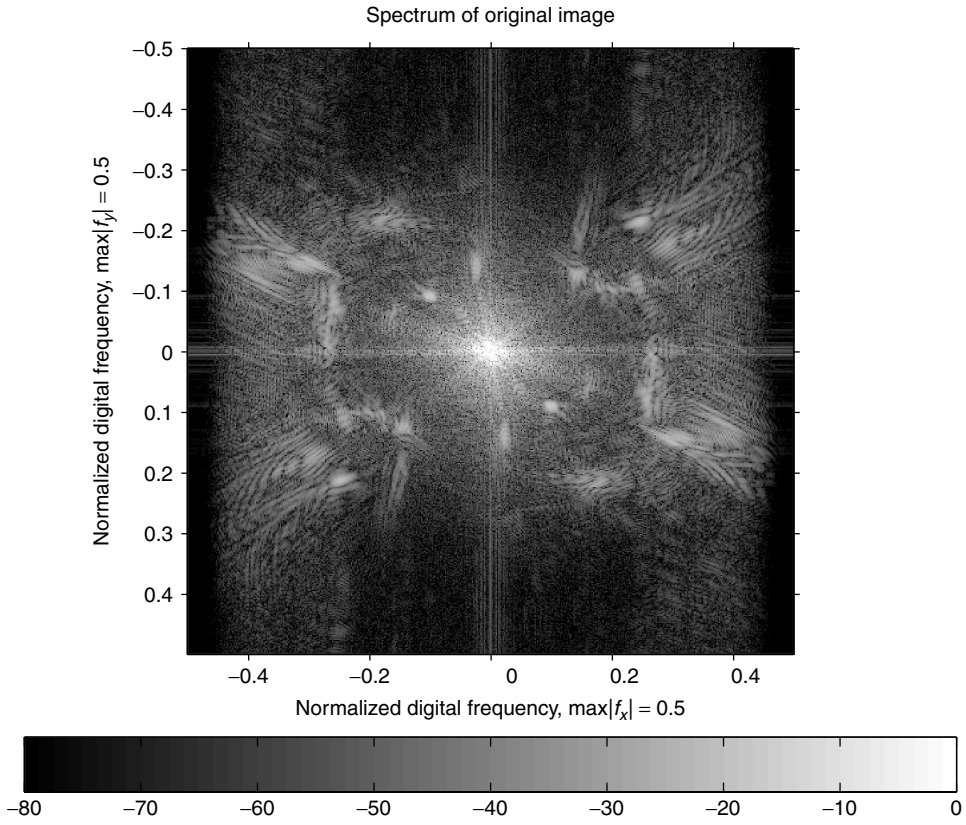


Figure 14. Spectrum of original bars image.

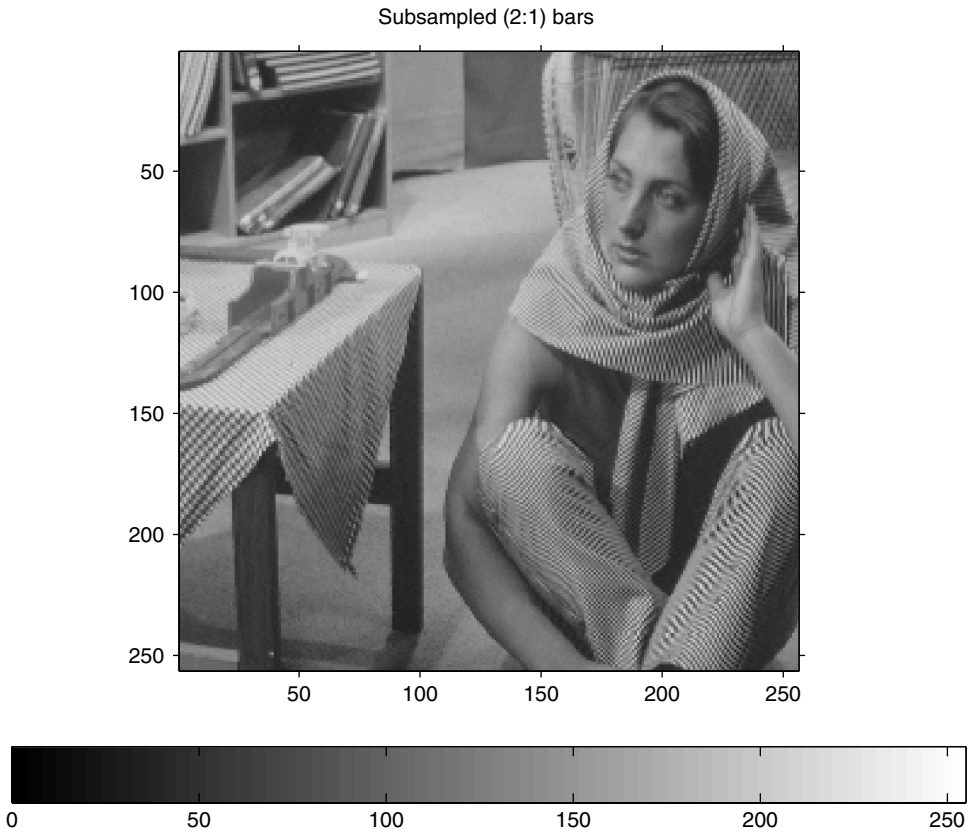


Figure 15. Sampled bars image 2 : 1.

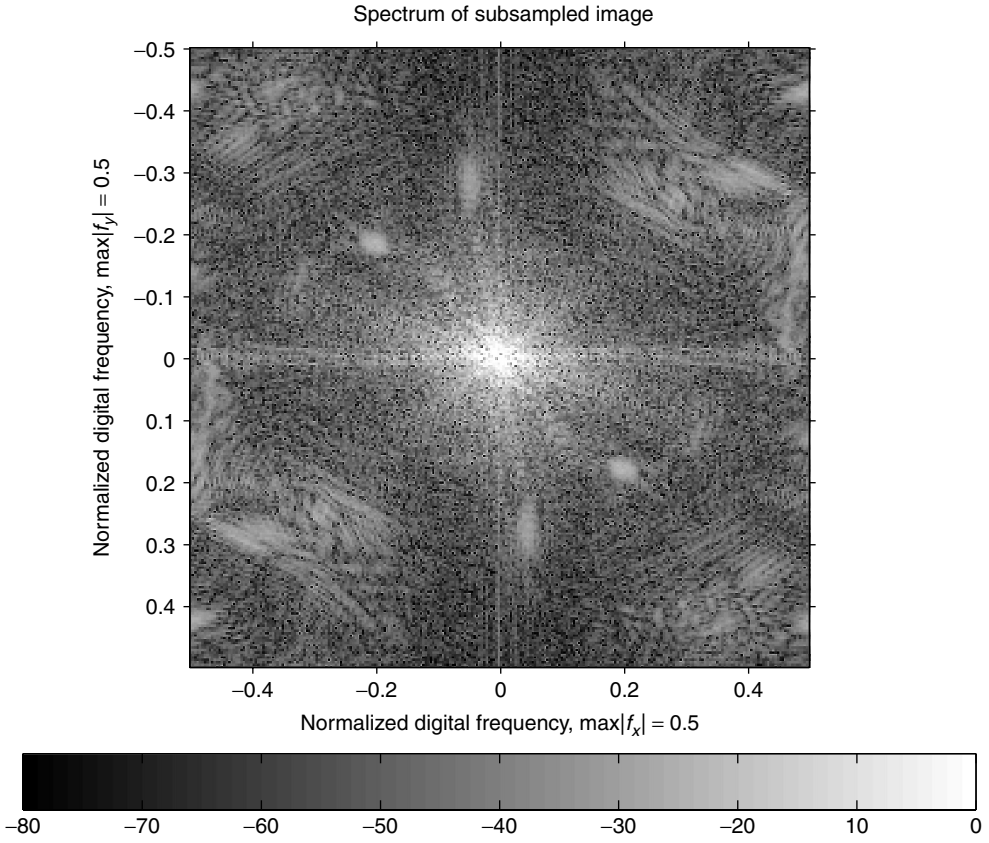


Figure 16. Spectrum of subsampled bars image.

The sampled spectrum can be written

$$S(\mathbf{w}) = S(\mathbf{w}) * comb(\mathbf{w}) = \frac{1}{|\mathbf{X}|} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} S(\mathbf{w} - k\mathbf{w}_1 - l\mathbf{w}_2) \tag{15}$$

(see the books by Dudgeon and Mersereau [3] and Jain [8]).

5. SAMPLING USING FINITE APERTURES

Practical imaging devices, such as videocameras, CCD (charge-coupled device) arrays, and scanners, must use a finite aperture for sampling. The *comb* function cannot be realized by actual devices. The finite aperture is required to obtain a finite amount of energy from the scene. The engineering tradeoff is one of signal-to-noise ratio (SNR) versus spatial resolution. Large apertures, receive more light, and thus, will have higher SNR's than smaller apertures, while smaller apertures permit higher spatial resolution than will larger ones. This is true for apertures larger than the order of the wavelength of light. For smaller apertures, diffraction limits the resolution.

The aperture may cause the light intensity to vary over the finite region of integration. For a single sample of a one-dimensional signal at time nT, the sample value can be obtained by

$$s(n) = \int_{(n-1)T}^{nT} s(t)a(t - NT)\,dt \tag{16}$$

where $a(t)$ represents the impulse response (or light variation) of the aperture. This is simple correlation and assumes that the same aperture is used for every sample. The mathematical representation can be written as convolution if the aperture is symmetric or, we replace the function $a(t)$ with $a(-t)$. The sampling of the signal can be represented by

$$s(n) = [s(t) * a(t)]comb\frac{t}{T} \tag{17}$$

where $*$ represents convolution. This model is reasonably accurate for spatial sampling of most cameras and scanning systems.

The sampling model can be generalized to include the case where each sample is obtained with a different aperture. For this case, the samples which need not be equally spaced, are given by

$$s(n) = \int_{l_n}^{u_n} s(t)a_n(t)\,dt \tag{18}$$

where the limits of integration correspond to the region of support for each aperture. A common application of this representation in two dimensions is the finite area of a CCD element of an imaging chip. The aperture function $a(t)$ may also take into account the leakage of charge from one cell to another. Equation (18) is also important in representing sampling the wavelength dimension of the image signals.

The generalized signal reconstruction equation has the form

$$s(t) = \sum_{n=-\infty}^{\infty} s(n) g_n(t) \tag{19}$$

where the collection of functions, $\{g_n(t)\}$, provide the interpolation from discrete to continuous space. The exact form of $\{g_n(t)\}$ depends on the form of $\{a_n(t)\}$. For sampling using the ideal *comb* function with a sample interval of T, $g_n(t)$ is a shift of the sinc function that represents the ideal band-limited filter

$$g_n(t) = \frac{\sin(2\pi(t - nT)/T)}{2\pi(t - nT)/T} \tag{20}$$

The two-dimensional aperture in the mathematical model of sampling can be written

$$s(m, n) = [s(x, y) * a(x, y)] comb(x, y) \tag{21}$$

where $*$ represents convolution and $a(x, y)$ represents the aperture. Note that using the finite aperture model can be written

$$s(m, n) = \iint_A s(x - m, y - n) a(x, y)\, dx\, dy \tag{22}$$

where the 2D integral is taken over the region of support of the aperture denoted by A. This equation is actually a correlation. The model can be written as a convolution with a space reversed aperture, $a_r(x, y) = a(-x, -y)$:

$$s(m, n) = \iint_A s(m - x, n - y) a_r(x, y)\, dx\, dy \tag{23}$$

For a symmetric aperture, which is most often the case in optical systems, $a_r(x, y) = a(x, y)$. Commonly used apertures include circular disks, rectangles, and Gaussians. Note that these have some symmetry that permits the substitution of convolution for correlation.

The Fourier representation of the sampled image is now given by

$$\begin{aligned} S_d(u, v) &= S(u, v) A(u, v) * comb(u, v) \\ &= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} S_a(u - k, v - l) A(u - k, v - l) \end{aligned} \tag{24}$$

With a finite aperture, the band-limited function to be sampled is the convolution $s(x, y) * a(x, y)$. The common aperture functions are generally lowpass in character; thus, the sampled function is more nearly band-limited. The aperture is effectively a filter. While aliasing caused by undersampling is diminished, the resultant spectrum is still a distorted version of the original.

The recovery of the original image must not only eliminate the periodic replications of the product $S(u, v)A(u, v)$ but also compensate for the effect of the aperture. The straightforward approach is to filter the sampled image with a kernel of the form

$$H(u, v) = \frac{1}{A(u, v)} \tag{25}$$

The problem with this approach is that the spectrum $A(u, v)$ often has values that are very small or zero which makes the inverse filter ill-conditioned; that is, $H(u, v)$ will have very large values that will amplify noise. Since most apertures are lowpass, the small or zero values usually occur at higher frequencies. A common modification of the above correction is to include a term to make the filter well-conditioned. Such a form is given by

$$H(u, v) = \frac{H_{lp}(u, v)}{A(u, v)} \tag{26}$$

where $H_{lp}(u, v)$ is a lowpass filter.

6. COLOR SAMPLING

There is a fundamental difference of philosophy about sampling in the wavelength domain from that of sampling in the spatial domain. To understand this difference, it is necessary to describe some of the fundamentals of color vision and color measurement. A more complete description of the human color visual system can be found in the books by Wandell [13] and Wyszecki and Stiles [14].

The retina contains two types of light sensors, rods, and cones. The rods are used for monochrome vision at low light levels; the cones are used for color vision at higher light levels. There are three types of cones. Each type is maximally sensitive to a different part of the spectrum. They are often referred to as long, medium, and short wavelength regions. A common description refers to them as red, green, and blue cones, although their maximal sensitivity is in the yellow, green, and blue regions of the spectrum. The visible spectrum extends from about 400 nm (blue) to about 700 nm (red).

Grassmann formulated a set of laws for additive color mixture in 1853 [5,6,15]. In addition, Grassmann conjectured that any additive color mixture could be matched by the proper amounts of three primary stimuli. Considering what was known about the physiology of the eye at that time, these laws represent considerable insight.[6] There have been several papers which have taken a linear systems approach to describing Grassmann's laws and color spaces as defined by a standard human observer, [4,7,10,12]. For the purposes of this work, it is sufficient to note that the spectral responses of the three types of sensors are sufficiently different so as to define a three-dimensional vector space. This three-dimensional representation is the basic principle of color displays in television, motion pictures, and computer monitors.

The mathematical model for the color sensor of a camera or the human eye can be represented by

$$v_k = \int_{-\infty}^{\infty} r(\lambda) m_k(\lambda) d\lambda, \quad k = 1, 2, 3 \tag{27}$$

where $r(\lambda)$ is the radiant distribution of light as a function of wavelength and $m_k(\lambda)$ is the sensitivity of the kth color

[6] The laws are not exact and there is considerable debate among color scientists today about their most accurate form.

sensor. The sensitivity functions of the eye are shown in any of the references [9–14].

Sampling of the radiant power signal associated with a color image can be viewed in at least two ways. If the goal of the sampling is to reproduce the spectral distribution, then the same criteria for sampling the usual electronic signals can be directly applied. However, the goal of color sampling is not often to reproduce the spectral distribution but to allow reproduction of the color sensation. The goal is to sample the continuous color spectrum in such a way that the color sensation of the spectrum can be reproduced by the monitor. To keep this discussion as simple as possible, we will treat the color sampling problem as a subsampling of a high-resolution discrete space, that is, the N samples are sufficient to reconstruct the original spectrum using the uniform sampling of Section 3.

Let us assume that the visual wavelength spectrum is sampled finely enough to allow the accurate use of numerical approximation of integration. A common sample spacing is 10 nanometers over the range 400–700 nm. Finer sampling is required for some illuminants with line emitters. Sampling of color signals is discussed in detail in Ref. 9. With the assumption of proper sampling, the space of all possible visible spectra lies in an N-dimensional vector space, where $N = 31$. The spectral response of each of the eye's sensors can be sampled as well, giving three linearly independent N vectors that define the visual subspace.

Under the assumption of proper sampling, the integral of Eq. (27) can be well approximated by a summation

$$v_k = \sum_{n=L}^{U} r(n\Delta\lambda) m_k(n\Delta\lambda) \tag{28}$$

where $\Delta\lambda$ represents the sampling interval and the summation limits are determined by the region of support of the sensitivity of the eye. The sensor $m_k(\cdot)$ can represent the eye as well as a photonic device.

The response of the sensors can be represented by a matrix, $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3]$, where the N vectors, $\mathbf{m}_i$, represent the response of the ith-type sensor (or cone). Any visible spectrum can be represented by an N vector, $\mathbf{r}$. The response of the sensors to the input spectrum is a 3 vector, v, obtained by

$$v = \mathbf{M}^T\mathbf{r} \tag{29}$$

Since we are interested in sampling to represent human color sensitivity, let the matrix $\mathbf{S} = [s_1, s_2, s_3]$, represent the sensitivity of the eye. The result of sensing with the eye, $\mathbf{t}$

$$\mathbf{t} = \mathbf{S}^T\mathbf{r} \tag{30}$$

is given the special name of *tristimulus* vector or values.

Two visible spectra are said to have the same color if they appear the same to the human observer. In our linear model, this means that if $\mathbf{r}_1$ and $\mathbf{r}_2$ are two N vectors representing different spectral distributions, they are equivalent colors if

$$\mathbf{S}^T\mathbf{r}_1 = \mathbf{S}^T\mathbf{r}_2 \tag{31}$$

It is clear that there may be many different spectra that appear to be the same color to the observer. Two spectra that appear the same are called *metamers*. Metamerism is one of the greatest and most fascinating problems in color science. It is basically color "aliasing" and can be described by the generalized sampling described earlier.

The N-dimensional spectral space can be decomposed into a 3D subspace known as the *human visual subspace* (HVSS) and an $N-3$D subspace known as the *black space*. All metamers of a particular visible spectrum, $\mathbf{r}$, are given by

$$\mathbf{x} = \mathbf{P}_v\mathbf{r} + \mathbf{P}_b\mathbf{g} \tag{32}$$

where $\mathbf{P}_v = \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T$ is the orthogonal projection operator to the visual space, $\mathbf{P}_b = [\mathbf{I} - \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T]$ is the orthogonal projection operator to the black space, and $\mathbf{g}$ is any vector in N space.

It should be noted that humans cannot see (or detect) all possible spectra in the visual space. Since it is a vector space, there exist elements with negative values. These elements are not realizable and thus cannot be seen. All vectors in the black space have negative elements. While the vectors in the black space are not realizable and cannot be seen, they can be combined with vectors in the visible space to produce a realizable spectrum.

If sampling by an optical device, with sensor $\mathbf{M}$, is done correctly, the tristimulus values can be computed from the optical measurements, $\mathbf{v}$, that is, $\mathbf{B}$ can be chosen so that

$$\mathbf{t} = (\mathbf{S}^T\mathbf{r}) = \mathbf{B}\mathbf{M}^T\mathbf{r} = \mathbf{B}\mathbf{v} \tag{33}$$

From the vector space viewpoint, the sampling is correct if the three-dimensional vector space defined by the cone sensitivity functions is the same as the space defined by the device sensitivity functions. Using matrix terminology, the range spaces of the $\mathbf{S}$ and $\mathbf{M}$ are the same.

When we consider the sampling of reflective spectra, we note that a reflective object must be illuminated to be seen. The resulting radiant spectra is the product of the illuminant and the reflection of the object

$$\mathbf{r} = \mathbf{L}\mathbf{r}_0 \tag{34}$$

where $\mathbf{L}$ is diagonal matrix containing the sampled radiant spectrum of the illuminant and the elements of the reflectance of the object are constrained, $0 \le \mathbf{r}_0(k) \le 1$. The measurement of the appearance of the reflective object can be computed in the same way as the radiant object with the note that the sensor matrices now include the illuminant, where $\mathbf{LS}$ and $\mathbf{LM}$ must have the same range space.

It is noted here that most physical models of the eye include some type of nonlinearity in the sensing process. This nonlinearity is often modelled as a logarithm; in any case, it is always assumed to be monotonic within the intensity range of interest. The nonlinear function, $\mathbf{v} = V(\mathbf{c})$, transforms the 3-vector in an element independent manner:

$$[v_1, v_2, v_3]^T = [V(c_1), V(c_2), V(c_3)]^T \tag{35}$$

Since equality is required for a color match by Eq. (31), the function $V(\cdot)$ does not affect our definition of equivalent colors. Mathematically

$$V(\mathbf{S}^T\mathbf{r}_1) = V(\mathbf{S}^T\mathbf{r}_2) \tag{36}$$

is true if, and only if, $\mathbf{S}^T\mathbf{r}_1 = \mathbf{S}^T\mathbf{r}_2$. This nonlinearity does have a definite effect on the relative sensitivity in the color matching process and is one of the causes of much searching for the "uniform color space."

7. PRACTICAL RECONSTRUCTION OF IMAGES

The theory of sampling states that a band-limited signal can be reconstructed if it is sampled properly. The reconstruction requires the infinite summation of a weighted sum of sinc functions. From the practical viewpoint, it is impossible to sum an infinite number of terms and the sinc function cannot be realized with incoherent illumination. Let us consider the two problems separately.

The finite sum can be modelled as the truncation of the infinite sum

$$\hat{s}(x,y) = \sum_{m=-M}^{M}\sum_{n=-N}^{N} s(m,n)\frac{\sin[\pi(x-m)]}{\pi(x-m)}\frac{\sin[\pi(y-n)]}{\pi(y-n)} \tag{37}$$

This is equivalent to truncating the number of samples by the use of the $rect(\cdot)$ function:

$$\hat{s}(x,y) = \left[s(x,y)comb(x,y)rect\left(\frac{x}{M},\frac{y}{N}\right)\right] * \operatorname{sinc}(x,y) \tag{38}$$

Clearly, as the number of terms approaches infinity, the estimate of the function improves. Furthermore, the $\operatorname{sinc}(x,y)$ is the optimal interpolation function for the mean-square error measure. Unfortunately, truncation by the ideal lowpass filter produces ringing at high-contrast edges caused by Gibbs phenomenon.

The inclusion of a practical reconstruction interpolation function can be modeled by replacing the $sinc(x,y)$ by the general function $h(x,y)$. The model is now given by

$$\hat{f}_a(x,y) = \left[f_a(x,y)comb(x,y)rect\left(\frac{x}{M},\frac{y}{N}\right)\right] * h(x,y) \tag{39}$$

The common forms of the interpolation function are the same as the sampling aperture when considering actual hardware, such as circular, rectangular, and Gaussian. These are used when considering output devices, such as CRT (cathode ray tube) or flat-panel monitors, photographic film, and laser printers. The optimal design of these apertures is primarily a hardware or optical problem. Software simulation usually plays a significant role in developing the hardware. Halftone devices, such as inkjet printers, offset and gravure printing, use more complex models that are a combination of linear and nonlinear processes [18]. There is another application that can be considered here that uses a wider variety of functions.

Image interpolation is often done when a sampled image is enlarged many times its original size. For example, an image may be very small, say 64×64. If the image is displayed as one screen pixel for each image pixel, the display device would show a reproduction that is too small, say, 1 in. $\times$ 1 in. The viewer could not see this well at normal viewing distances. To use the entire area of the display requires producing an image with more pixels. For this example, $8 \times$ enlargement would produce a 512×512 image that would be 8×8 in. This type of interpolation reconstruction is very common when using variable sized windows on a monitor.

The simple method of pixel replication is equivalent to using a rectangular interpolating function, $h(x,y)$. This is shown in Figs. 9 and 11. The square aperture is readily apparent. The figure of the pictorial image, Fig. 13, has more pixels and uses a proportionally smaller reproduction aperture. The aperture is not apparent to the eye.[5] In the case of the sinusoidal images of Figs. 9 and 11, the purpose of the image was to demonstrate sampling effects. Thus, the obvious image of the aperture helps to make the sampling rate apparent. If we desire to camouflage the sampling and produce a smoother image for viewing, other methods are more appropriate.

Bilinear interpolation uses a separable function composed of triangle functions in each coordinate direction. This is an extension of linear interpolation in one dimension. The image of Fig. 11 is displayed using bilinear interpolation in Fig. 17. The separability of the interpolation is noticed in the rectilinear artifacts in the image.

A more computationally expensive interpolation is the cubic spline. This method is designed to produce continuous derivatives, in addition to producing a continuous function. The result of this method is shown in Fig. 18. For the smooth sinusoidal image, this method works extremely well. One can imagine that images exist where the increased smoothness of the spline interpolation would produce a result that appears more blurred than the bilinear method. There is no interpolation method that is guaranteed to be optimal for a particular image. There are reasons to use the spline method for a wide variety of images [19]. There are many interpolating functions that have been investigated for many different applications [20].

8. SUMMARY

The article has given an introduction to the basics of sampling and reconstruction of images. There are clearly several areas that the interested reader should expand on by additional reading. In-depth treatment of the frequency domain can be obtained from many of the common image processing texts, such as that by Jain, [8]. The processing of video and temporal imaging is covered well in Tekalp's text [21]. Color imaging is treated in a special issue of the *IEEE Transactions on Image Processing* [17]. A survey paper in that issue is a good starting point on the current state of the art. Sampling and reconstruction of medical images is treated in the treatise by Macovski [22].

[5] This is true even when the image is viewed without the effect of halftone reproduction, which is used here.

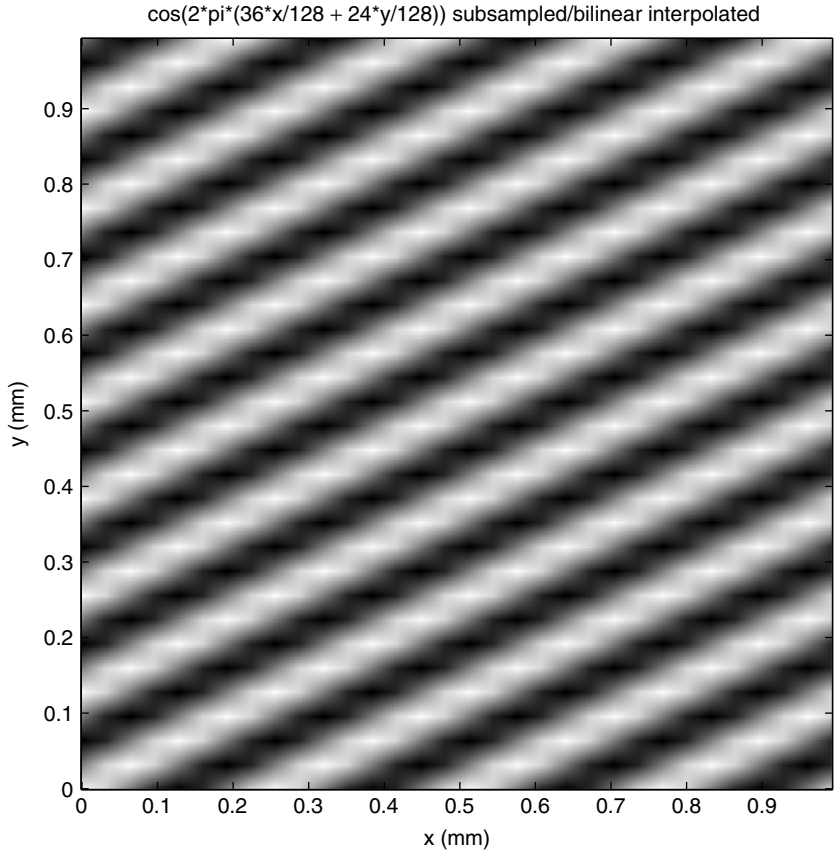


Figure 17. Linear interpolation of subsampled $\cos[2\pi(36x/128 + 24y/128)]$.

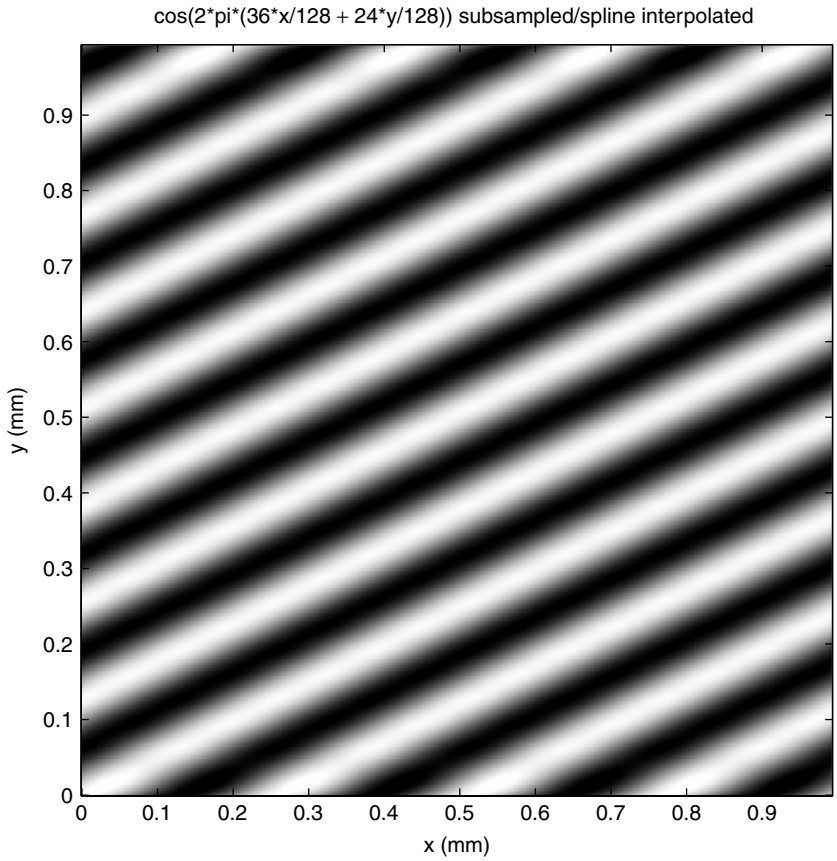


Figure 18. Spline interpolation of subsampled $\cos[2\pi(36x/128 + 24y/128)]$.

BIOGRAPHY

Joel Trussell received degrees from Georgia Tech (1967), Florida State (1968), and the University of New Mexico (1976). In 1969 he joined the Los Alamos Scientific Laboratory, Los Alamos, New Mexico, where he began working in the image and signal processing in 1971. During 1978/79, he was a Visiting Professor at Heriot-Watt University, Edinburgh, Scotland, where he worked with both the university and with industry on image processing problems. In 1980, he joined the Electrical and Computer Engineering Department at North Carolina State University, in Raleigh, where is now a Professor. During 1988/89, he was a Visiting Scientist at the Eastman Kodak Company in Rochester, New York, and in 1997/98 was a Visiting Scientist at Color Savvy Systems in Springboro, Ohio. He is a past Associate Editor for the journals *Transactions on ASSP* and *Signal Processing Letters*. He is a past Chairman of the Image and Multidimensional Digital Signal Processing Committee of the Signal Processing Society of the IEEE. He founded and edited the electronic newsletter published by this committee. He is Fellow of the IEEE and has shared the IEEE-ASSP Society Senior Paper Award (1986, with M. R. Civanlar) and the IEEE-SP Society Paper Award (1993, with P. L. Combettes).

BIBLIOGRAPHY

1. MATLAB, *High Performance Numeric Computation and Visualization Software*, The Mathworks Inc., 24 Prime Park Way, Natick, MA 01760.
2. A. V. Oppenheim and A. S. Willsky, *Signals and Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1997.
3. D. E. Dudgeon and R. M. Mersereau, *Multidimensional Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
4. J. B. Cohen and W. E. Kappauf, Metameric color stimuli, fundamental metamers, and Wyszecki's metameric blacks, *Am. J. Psychol.* **95**(4): 537–564 (1982).
5. H. Grassmann, Zur Therorie der Farbenmischung, *Annalen der Physik und Chemie* **89**: 69–84 (1853).
6. H. Grassmann, On the theory of compound colours, *Philos. Mag.* **7**(4): 254–264 (1854).
7. B. K. P. Horn, Exact reproduction of colored images, *Comput. Vision Graph. Image Proc.* **26**: 135–167 (1984).
8. A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
9. H. J. Trussell and M. S. Kulkarni, Sampling and processing of color signals, *IEEE Trans. Image Proc.* **5**(4): 677–681 (April 1996).
10. H. J. Trussell, Application of set theoretic methods to color systems, *Color Res. Appl.* **16**(1): 31–41 (Feb. 1991).
11. P. L. Vora and H. J. Trussell, Measure of goodness of a set of colour scanning filters, *J. Opt. Soc. Am.* **10**(7): 1499–1508 (July 1993).
12. B. A Wandell, The Synthesis and Analysis of Color Images, *IEEE Trans. Patt. Anal. Mach. Intel.* **PAMI-9**(1): 2–13 (Jan. 1987).
13. B. A Wandell, *Foundations of Vision*, Sinauer Assoc. Inc, Sunderland, MA, 1995.
14. G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd ed., Wiley, New York, 1982.
15. W. B. Cowan, An inexpensive scheme for calibration of a color monitor in terms of standard CIE coordinates, *Comput. Graph.* **17**: 315–321 (July 1983).
16. M. J. Vrhel and H. J. Trussell, Color device calibration: A mathematical formulation, *IEEE Trans. Image Process.* **8**(12): 1796–1806 (Dec. 1999).
17. *IEEE Trans. Image Proc.* **6**(7): (July 1997).
18. R. Ulichney, *Digital Halftoning*, MIT Press, Cambridge, MA, 1987.
19. M. Unser, Splines — a perfect fit for signal and image processing, *IEEE Signal Process. Mag.* **16**(6): 22–38 (Nov. 1999).
20. R. N. Bracewell, *Two-Dimensional Imaging*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
21. A. M. Tekalp, *Digital Video Imaging*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
22. A. Macovski, *Medical Imaging Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

IMT-2000 3G MOBILE SYSTEMS

ANNE CERBONI
JEAN-PIERRE CHARLES
France Télécom R&D
Issy Moulineaux, France

PIERRE GHANDOUR
JOSEP SOLE I TRESSERRES
France Télécom R&D
South San Franciso, California

1. INTRODUCTION

The ability to communicate anywhere, at any time, with anyone in the world, using moving pictures, graphics, sound, and data has been a longstanding challenge to telecommunications operators. With the advent of IMT-2000 third-generation mobile systems, people on all continents will be able to take advantage of most of these capabilities. Standardization, ensuring that customers' terminals are compatible with IMT-2000 networks throughout the world, has had to accommodate regional differences. While ongoing work within different standards forums is striving to achieve a high degree of universality, IMT-2000 in fact covers a family of standards. Relevant standards issues and their implications in different regions are discussed in Section 3. The emergence of third-generation (3G) mobile systems is, of course, rooted in the development and increasingly widespread use of previous generations. The migration paths toward 3G systems are reviewed in Section 4. Enhanced data rates offered by IMT-2000 are expected to provide a broad range of mobile multimedia services on a multiple choice of terminals. These new facilities and applications

are examined in Section 5. Section 6 describes IMT-2000 radio access and core network architecture. In Section 7, the authors address the most salient economic implications of migration toward 3G, with emphasis on license costs and new business models. Section 8 discusses evolution of mobile systems beyond 3G, which involves not only an all-IP core network but also optimization of scarce spectral resources through cooperation among heterogeneous networks.

2. DRIVING FORCES

Two major trends are reshaping the world of telecommunications. On one hand, mobile services have made great strides throughout the world, with penetration rates exceeding 50% in many countries. On the other hand, the explosion of Internet traffic testifies to the rapid development of the multimedia market. The prospect for convergence of these two trends, giving rise to mobile multimedia services, and the resulting need for greater spectral resources, has driven equipment suppliers, operators, standards bodies, and regulators throughout the world to develop a new generation of mobile systems. The stakes are considerable: around 2010, mobile traffic should be equal to that of fixed telephony [1]. The convergence of the mobile and Internet worlds, the strong dynamics of innovation, and anticipated cost reductions in these areas have opened new opportunities for 3G services as of 2001 in Japan and possibly in the United States, and 2002 in Europe.

3. STANDARDIZATION

3.1. IMT-2000 Frequency Spectrum

The International Telecommunications Union (ITU) initiated 3G mobile standardization with the ambition of defining a global standard replacing the broad variety of second-generation (2G) mobile systems, which implied common spectrum throughout the world. Hence, the first efforts on 3G systems truly started once the World Administrative Radio Conference (WARC) 92 had identified a total of 230 MHz for IMT-2000, as illustrated in Fig. 1.

IMT-2000 standardisation activities for the radio interface are conducted in ITU-R/WP 8F. At WARC 2000, additional frequency bands totalling 400 MHz were allocated for IMT-2000.

3.2. IMT-2000 Radio Interface

The early impetus for standardizing an IMT-2000 radio interface, specifically, the interface between the mobile terminal and the base station, can be attributed to an observation of the contrasting situation of 2G systems in Europe and in the United States. In Europe, the Global System for Mobile Communications (GSM) standard was developed and universally adopted before 2G systems were first deployed in 1991, ensuring that customers' terminals are compatible with mobile networks throughout the continent. In the United States, the lack of country or continentwide harmonization, and the relatively greater success of first-generation analog systems such as advanced mobile phone service (AMPS), led to the parallel development of three different 2G digital standards:

- Time-division multiple access (TDMA) including digital AMPS (DAMPS) and IS136
- Code-division multiple access (CDMA), known as IS95
- Global System for Mobile Communications (GSM)

Table 1 provides a brief overview of the market share of these standards in the United States.

Despite the initial goal of a single worldwide 3G air interface, standardization was strongly influenced by

Table 1. Market Share of 2G Cellular Standards in the United States

Technology	1999	2002
GSM	4%	11–15%
AMPS	55%	20%
TDMA	25%	36%
CDMA	16%	28%
Total subscribers (millions)	85	135

Source: France Télécom North America.

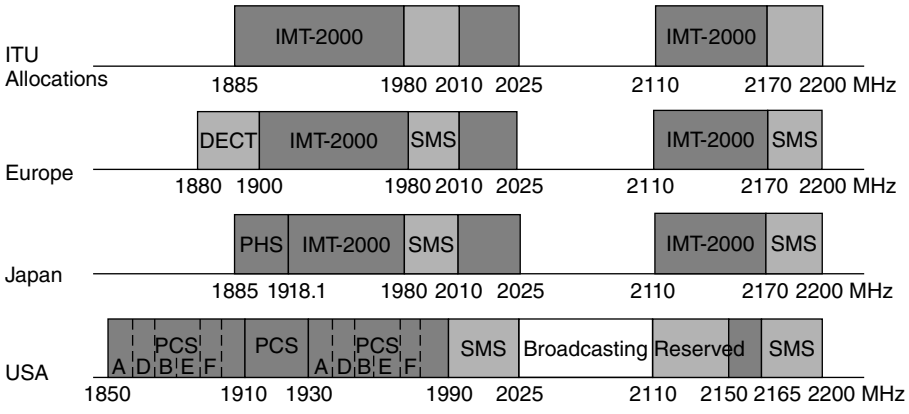


Figure 1. IMT-2000 spectrum.

mobile operators' need to leverage their 2G investment. Specifically, this meant that 3G systems must ensure backward compatibility with existing systems, providing seamless handover to and from 2G systems, in addition to sharing tower infrastructure and transmission resources. As a result, in November 1999, the International Telecommunications Union — Radiocommunications (ITU-R) could not establish a consensus on any one 3G air interface and thus adopted five different solutions:

- Universal Mobile Telecommunications System/Wideband CDMA (UMTS/WCDMA) one of the two modes supported by NTT DoCoMo, Nokia, and Ericsson, and developed by the Third-Generation Partnership Project (3GPP).
- cdma2000, an evolution of the American CDMA IS95A solution originally developed by Qualcomm and currently standardized by the Third-Generation Partnership Project 2 (3GPP2).
- UMTS/TD-CDMA: UMTS mode combining time-division (TD) and code-division multiple access, supported by Siemens. This solution also includes a specific option developed for China called TDSCDMA (the S stands for synchronous). This mode is also developed by 3GPP.
- Enhanced Data for GSM Evolution (EDGE) or UWC-136. This solution represents an evolution of both TDMA and GSM.
- Digital enhanced cordless telephony DECT developed by ETSI.

Among these solutions, cdma2000, WCDMA, and EDGE are discussed in greater detail in the following paragraphs.

3.3. Regional Standardization

3.3.1. Europe. In Europe, impetus for the development of IMT-2000 was largely provided by European Commission research programs in the late 1980s and early 1990s (RACE I and II,[1] ACTS/FRAMES[2] project). In 1991, with the deployment of the first 2G systems, the European Telecommunications Standards Institute (ETSI) created a subcommittee to develop an IMT-2000 system called *UMTS*. Efforts first focused on defining the technical requirements for the radio interface, and various solutions were presented at the December 1996 ETSI conference, including three proposals by ACTS/FRAMES. Following a vote in January 1998, a compromise was found based on two harmonized modes: WCDMA and TDCDMA. WCDMA was adopted for the frequency-domain duplex (FDD) mode, namely, one frequency per transmission direction, and TDCDMA for the time-domain duplex (TDD) mode, specifically, time-division multiplexing of the two directions on the same frequency. This combined solution offers the advantage of enabling full use of the IMT-2000 frequency bands; the FDD mode is used in priority in the paired bands and the TDD mode, in unpaired bands. This compromise was then submitted to ITU-R as the European proposal for the IMT-2000 radio interface.

Standardization in Europe was strongly influenced by lobbying within forums such as the GSM Association and the UMTS Forum, which strove to federate the stances of GSM operators and manufacturers regarding the development of 3G systems. National regulation authorities also played a fundamental role in defining the use of the spectrum identified by the WARC 92, and for the attribution of UMTS licenses.

3.3.2. Japan, Korea, and China. In Japan, most 3G developments were financed by mobile operator NTT DoCoMo. Japanese industry supported this R&D effort to develop a new standard and take the lead in this very competitive market. European manufacturers Nokia and Ericsson took part in this effort, which led to the establishment of a common solution between them and Japan, based on WCDMA. This compromise was reached just as ETSI was seeking candidates for its 3G mobile system, leading to a convergent view between Japan and Europe in favor of WCDMA for the air interface. Although other carriers like Japan Telecom followed this direction, KDDI, the second largest Japanese carrier, is strongly involved in cdma2000.

In Korea, the Telecommunications Technology Association (TTA) kept two paths open for the evolution of the country's CDMA mobile networks. Both WCDMA and cdma2000 are officially supported by the Ministry of Information and Communications. In fact, the three mobile carriers (SK Telecom, KT Freetel and LG Telecom) have already started to offer cdma2000 1x services in overlay of their 2G networks. The first two operators obtained 3G licenses to deploy WCDMA, mainly for ease of roaming. The government plans to grant another 3G license based on cdma2000.

In China, work on a 3G standard started in 1992 within the First Research Institute of the Datang Group, now the Chinese Academy of Telecommunication Technology (CATT). After studying the European GSM standard in detail, the group developed its own standard, Synchronous code-division multiple access (SCDMA), which became the basis for subsequent 3G developments. With the support of Siemens, TDSCDMA was adopted as an official ITU 3G standard. The Chinese government is devoting significant resources and energy to building a national industry around this standard, instead of relying on imported network equipment and terminals. The country which, in 2001, represents the world's second largest mobiles market, can afford to develop its own standard, touted as offering greater spectral efficiency than rivals WCDMA and cdma2000 and being more cost-effective and more easily integrated in the GSM environment. Carriers such as state-owned China Mobile and China Unicom may potentially deploy TDSCDMA.

3.3.3. United States. In the United States, tough and unrestricted competition driven by the various

[1] Research on Advanced Communication Technologies in Europe.
[2] Advanced Communication Technologies and Services/Future Radio Wideband Multiple Access System). The main partners were France Télécom, Nokia, Siemens, Ericsson, and CSEM/Pro Telecom.

manufacturers led to the creation of several standards committees. Initially, the Telecommunication Industry Association (TIA) was in charge of standardizing TDMA IS136 (TIA/TR43.3) and CDMA IS95A (TIA/TR45.5) while T1P1, the GSM Alliance, and the GSM Association were involved in GSM standardization. When 3GPP2 was created in 1999, the TIA/TR45.5 working group became a major player in cdma2000 standardization. TIA/TR45.34, on the other hand, decided that 3GPP2 was not the suitable group for TDMA standardization and joined standardization efforts within the Universal Wireless Communications Consortium (UWCC).

Establishing digital cellular coverage in the United States is costly and carriers have had to invest a substantial amount of money. It is thus not surprising to note that the primary recommendations of American proposals for IMT-2000 often correspond to evolutions of existing second-generation systems maintaining backward compatibility in order to capitalize on the initial investment. In particular, cdma2000 1x was designed for smooth evolution from the second-generation IS95A standard.

The United States is facing a spectrum shortage for 3G systems. A large part of the frequency band allocated by WARC 92 (see Fig. 1) is currently used by second-generation personal communication systems. Although the United States has allotted only 210 MHz of spectrum for mobile wireless use, compared to an average of 355 MHz per country in Europe, companies are currently pressing forward with their plans for third-generation (3G) networks, while industry efforts at obtaining more spectrum, led by the Cellular Telecommunications and Internet Association (CTIA), continue. Sprint PCS and Cingular, for example, have announced plans to squeeze more capacity out of existing networks by upgrading technology, although spectrum in the United States is already more crowded than in other markets. The United States has nearly 530,000 mobile customers per megahertz of spectrum while the United Kingdom has just more than 80,000 users, and Finland, the world's leader in wireless penetration, has only 15,000 users per megahertz.

In October 2000, a memorandum was issued to define the need for a radiofrequency spectrum for future mobile voice, high-speed data, and Internet-accessible wireless services. The memorandum directed the Secretary of Commerce to work cooperatively with the Federal Communications Commission (FCC).

Various frequency bands had been identified for possible 3G use. The FCC and the National Telecommunications and Information Administration (NTIA) undertook studies of the 2500–2690 MHz and the 1755–1850 MHz frequency bands in order to provide a full understanding of all the spectrum options available. Both bodies stated possible sharing and segmentation options, but a review of the reports showed that, for every option, there were several caveats. Therefore, 3G carriers will mainly count on existing spectrum for 3G services. Sprint PCS and Cingular are heading this way when releasing 3G services toward the end of 2001.

3.3.4. 3GPP and 3GPP2. In this international context, standardization activities led within the ITU and regional entities[3] developed with increasingly close contacts. In 1998, ETSI, the Association of Radio Industries and Businesses (ARIB, Japan), TTC, as well as the Telecommunications Technology Association (TTA) of Korea and T1P1 of the United States founded the Third Generation Partnership Project (3GPP) as a forum to develop a common WCDMA standard, assuming GSM as a basis. The following year, the American National Standards Institute (ANSI) International Committee initiated 3GPP2, geared toward developing standards for the cdma2000 standard, in continuity with the CDMA IS95A standard. Member organizations include ARIB, China Wireless Telecommunication Standards Group (CWTS), Telecommunications Industry Association (TIA) from North America, Telecommunications Technology Association (TTA) from Korea and Telecommunications Technology Committee (TTC) from Japan.

Harmonization efforts between 3GPP and 3GPP2 resulted in a number of common features in the competing technologies, and enabled work to be divided between them such that 3GPP handles direct-sequence WCDMA and 3GPP2 focuses on the multicarrier (MC) mode in cdma2000. 3GPP successfully produced a common set of standards for the WCDMA air interface, known as "Release 99." Meanwhile, 3GPP2 issued Release A of cdma2000 1x and is currently (at the time of writing) working on Release B. In parallel, 3GPP2 also released an evolution of cdma2000 1x called "cdma2000 1xEV" (1xEVOLUTION) Phase 1 [also called "HDR (High Data Rate standard) Data Only"] and is currently working on Phase 2 (Data and Voice). Section 4 provides more information on these releases. Table 2 presents an overview of the different CDMA technologies envisioned for 3G.

An Operators Harmonization Group (OHG) was also created in 1999 for promoting and facilitating convergence of 3G networks One goal is to provide seamless global roaming among the different CDMA 3G modes (cdma2000, WCDMA, and TDD modes). The OHG was involved in specifying what mode should be used for the 3G radio access network. While cdma2000 was specifically oriented toward the multicarrier mode, WCDMA was to be only direct-spread. The objective is to achieve a flexible connection between the radio transmission technologies (RTTs) and the core networks (either evolved GSM MAP (mobile application part) or evolved ANSI-41).

4. CONTEXT AND EVOLUTION PATHS

4.1. Development Context

Why are IMT-2000 systems called "3G"? First-generation mobile systems were based on analog standards including AMPS in the United States, TACS (Total Access Communication System) in the United Kingdom, CT-2

[3] ETSI for Europe, the Telecommunications Technology Committee (TTC) and the Association for Radio Industries and Businesses (ARIB) for Japan, the Telecommunication Industry Association (TIA) and American National Standards Institute (ANSI) for the United States.

Table 2. Comparison of Different CDMA Technologies

	CDMA Technology Comparisons			
CDMA technology	Peak Data Rate	Average Data Throughput	Approved Standard?	Company
cdma2000-1x Phase 1	153.6 kbps	150 kbps	Yes	—
cdma2000-1x RTT A	614.4 kbps	415 kbps	Yes	—
1X Plus Phase 1	1.38 Mbps	560 kbps	—	Motorola
WCDMA (5 MHz)	2.048 Mbps	1.126 Mbps	Yes	—
HDR	2.4 Mbps	621 kbps	—	Qualcomm
cdma2000-3x multicarrier	2.072 Mbps	1.117 Mbps	—	—
1x Plus (Phase 2)	5.184 Mbps	1.200 Mbps	—	Motorola

Source: Soundview Technology Group and Motorola (published in *Global Wireless Magazine*).

in Europe, and that of NTT (Nippon Telephone and Telegraph) in Japan. The term "second generation" refers to digital mobile systems such as TDMA and CDMA in the United States, GSM (USA, Europe, China) and Personal Digital Cellular (PDC) in Japan. These systems provide mobile telephony, short-message services (SMSs) and low rate data services relying on standards such as the Wireless Applications Protocol (WAP) and I-mode. WAP is a standard for delivering Internet content and applications to mobile telephones and other wireless devices such as personal digital assistants (PDAs). It can be used independently of the type of terminal and network. WAP content is written in wireless markup language (WML), a version of HTML designed specifically for display on wireless terminal screens. To provide WAP content, the operator must implement a server between the wireless network and the Internet, which translates the protocol and optimizes data transfer to and from the wireless device. However, slow data rates, poor connections, and a limited number of services have significantly limited the use of mobiles for data applications. I-mode, on the other hand, attracted millions of users within its first year of existence, starting in 1999, in Japan. Despite the 9.6-kbps (kilobits per second) data, this proprietary packet-data standard developed by NTT DoCoMo met with widespread success for several reasons: no dial-up, volume-based billing, services adapted to the low bit rate, and an extensive choice of content providers.

4.2. Migration Paths

The technological options taken by different cellular operators to deploy 3G networks depend on the 2G technology employed. In Europe, the starting point is GSM. In the United States, in addition to GSM, operators are focusing on two other paths, TDMA, and IS95A (CDMA). Each of these migration paths is described hereafter.

4.2.1. GSM Migration Path. The first step was high-speed circuit-switched data (HSCSD), introduced commercially in Finland in 1999. HSCSD supports data rates of up to 57.6 kbps by grouping four GSM time slots. Access to HSCSD, however, involves a new terminal for the customer. This service has not been developed extensively, and operators are putting more energy into developing packet-based data technologies such as General Packet Radio Service (GPRS or GSM phase 2+).

GPRS theoretically supports up to 115.2 kbps packet-switched mobile data alongside circuit-switched telephony. The principle is to utilize GSM time slots to carry data. The amount of data per time slot varies from about 9 to 21 kbps, depending on the coding scheme used. GPRS is well adapted to asymmetrical traffic, with a greater number of time slots dedicated to the downlink. It also enables per volume billing, which, according to the I-mode example, encourages use of the service. GPRS requires new equipment in the GSM base station subsystem (BSS) and network subsystem (NSS) in order to handle the packet data. The packet control unit (PCU) located in the BSS handles the lower levels of the radio interface: the radio-link control (RLC) and medium access control (MAC) protocols. In the NSS, there are two important elements, which are also used in 3G networks. The first is the Serving GPRS Support Node (SGSN), basically an IP (Internet Protocol) router with specific functional features. For the subscribers in a given area of the mobile network, it handles authentication and security mechanisms, mobility management, session management, billing functions, in addition to the transmission of data packets. The second element is the Gateway GPRS Support Node (GGSN), another IP router which acts as gateway for data transmission between the GPRS network and other packet data networks. Other elements in the GPRS network include a domain name server (DNS) and a legal interception gateway. The home location register (HLR) of the mobile network is modified to take into account the data capabilities of GPRS customers. Mobile terminals (MTs) must, of course, be GPRS-compatible. There are several MT classes; the most basic is a GPRS radio modem for a laptop or handheld device, the most sophisticated handles both voice and data flows simultaneously.

EDGE or Enhanced GPRS (EGPRS), one of the five standard IMT-2000 radio interfaces, represents an upgrade not only from GSM but also from TDMA. The EDGE approach is similar to that of GPRS, in that it makes use of existing time slots for packet data transmission. However, EDGE uses a more elaborate coding scheme providing up to 48 kbps per time slot, yielding an overall rate of up to 384 kbps. A potential drawback of EDGE is that, as data rates increase, range decreases. While services are not interrupted, this fact may require the

operator to build out a denser network. Furthermore, as in the case of HSCSD and GPRS, a specific terminal is required. EDGE is described in greater detail in Section 6.3.

In its Release 99 (R99), UMTS, taken to be synonymous with WCDMA, offers symmetrical links at 384 kbps for customers at a speed of about 120 km/h, and 128 kbps for customers at a speed of up to 300 km/h. Technically, this technology allows a data rate of up to 2 Mbps for a quasistationary user. In order to ensure compatibility with existing GSM networks, dual band, dual-mode UMTS/GSM terminals are required for access to nationwide voice and data (GPRS or EDGE) services, at least until UMTS coverage attains a significant percentage of the country, and GSM networks are progressively phased out.

While intermediate steps HSCSD, GPRS, and EDGE are overlaid onto the GSM radio network, UMTS requires an entirely different radio access network. The network architecture is described in Section 6.2.

4.2.2. TDMA Migration Path. The first step in the evolution of TDMA is IS136+, which advances the data speed to 64 kbps through packet switching over a 200-kHz channel. Since IS136+ require a 200-kHz channel, instead of the 30-kHz bandwidth previously used by TDMA, the base stations must be upgraded with new hardware, which is expensive. The next phase is IS136 HS (high-speed), which uses EDGE technology, allowing the network to reach a theoretical data rate of 384 kbps, and also requires a hardware upgrade to the network's base stations. From there, the network requires another expensive hardware (base station) upgrade to support WCDMA. As a result, even if a carrier skipped IS136+, the TDMA migration involves two expensive hardware upgrades and therefore represents the most costly evolution path.

4.2.3. CDMA Migration Path. Given that the two stages after CDMA IS95A — CDMA IS95B and cdma2000 1x RTT (also known as IS2000 or IS95C) — operate on the same 1.25-MHz channel bandwidth as CDMA IS95A, the migration of this network architecture is the easiest to implement. Indeed, both IS95B and cdma2000 1x[4] require only a relatively cheap software upgrade.

However, those stages are completely independent; it is not necessary for a CDMA IS95A carrier to move to IS95B before moving to cdma2000. As a matter of fact, IS95B standardization work took such a long time that, when it was released, U.S. carriers chose to bypass IS95B, judging that IS95A was sufficient and that they could wait for cdma2000. Indeed, while IS95B offers a speed of up to 64 kbps, cdma2000 1x more than doubles the data speed to 144 kbps, and doubles the network's voice capacity as well. Finally, CDMA 3x RTT, which supplies a peak data rate of up to 2 Mbps, operates on a 3.75 MHz (3×1.25 MHz) frequency channel. Given the larger channel requirements, hardware upgrades of base stations are necessary for this transition.

[4] 1x means one times 1.25 MHz, the bandwidth used for CDMA IS-95.

Another potential alternative called "cdma2000 1x EV" (1x Evolution) also uses a 1.25-MHz channel. This transition includes two phases. For Phase 1, the High Data Rate (HDR) standard was released in August 2000. This technology, also named 1xEV-DO (meaning 1x Evolution Data Only), is supported by Qualcomm, Ericsson and Lucent. It is expected to increase peak data rates of up to 2.4 Mbps. Phase 2 (1xEV-DV meaning 1xEvolution Data and Voice), under standardization, will enable both voice and data channels (up to around 5 Mbps), while enhancing capacity and coverage.

Figure 2 illustrates the main alternatives for operators of 2G cellular networks.

5. IMT-2000 SERVICES AND TERMINALS

5.1. Services

A key feature of IMT-2000 is the wide range of services offered. In addition to voice, videotelephony, and videoconferencing applications, it covers asymmetrical data flows (Web browsing, video or audio streaming), as well as low-data-rate machine-to-machine communications (metering, e-wallet applications). Unlike 2G networks and fixed networks, IMT-2000 does not preassign a bit rate and service quality level to each type of service, but rather provides a framework within which communications can be characterized in terms of their requirements. In a mobile context, in particular, services should remain available in a flexible manner (at a lower rate or with less error protection, e.g., in case of degraded radio conditions) to ensure optimal use of the allocated frequency bands while guaranteeing an acceptable service level from the user's viewpoint. These considerations led to the definition of four quality-of-service (QoS) classes:

- Conversational
- Streaming
- Interactive
- Background

The main characteristics of each class are indicated in Table 3 [2], along with examples:

The characterization in Table 3 enables the mapping of applications onto the UMTS and radio access bearer services. There are no specifications set out by 3GPP for this mapping, even for a service as simple as voice. Instead, the framework for mapping (QoS, transport formats and channels, channel coding, physical channel, etc.) is described and the actual choice of forward error correction code and physical bearer service is left up to the manufacturer and/or operator; the idea is to leave as much latitude as possible for the implementation of existing and especially new services.

In the 3GPP2 Environment, the *quality of service* refers to a set of capabilities that a network may provide to a communications session. These capabilities can be specified so that particular applications (e.g., voice, video, streaming audio) fulfill human factors or other requirements with respect to fidelity and performance.

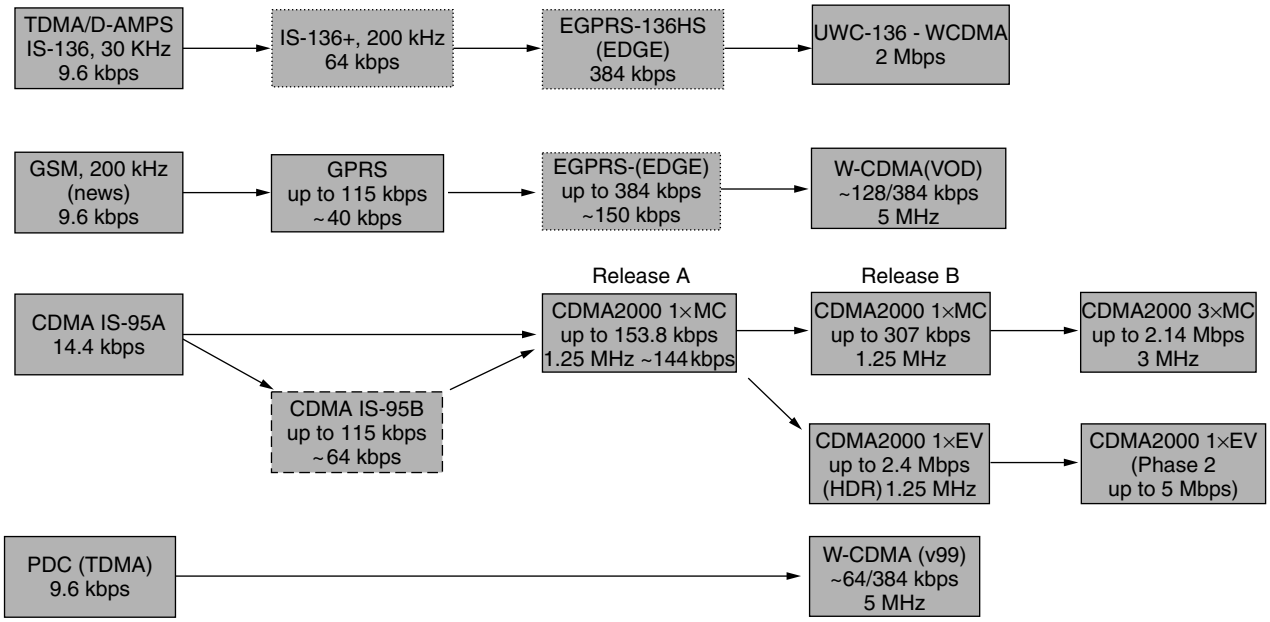


Figure 2. Possible migration paths from 2G to 3G.

Table 3. UMTS Quality-of-Service Classes

Traffic Class	Conversational Class Real Time	Streaming Class Real Time	Interactive Class Best Effort	Background Best Effort
Fundamental characteristics	Preserve time relation (variation) between information entities of the stream Conversational pattern: stringent and low delay	Preserve time relation (variation) between information entities of the stream One-way flow	Request response pattern Preserve payload content	Destination is not expecting the data within a certain time Preserve payload content
Example of the application	Voice, video telephony	Video, audio streaming	Web browsing	Background download of emails

Quality of service in a packet network consists of at least the following components:

Bandwidth — data rate
Delay — end-to-end or round-trip latency
Jitter — interpacket latency variation
Loss — rate at which packets are dropped

Additionally, this QoS may be

Unidirectional or bidirectional
Guaranteed or statistical
End-to-end or limited to a particular domain or domains
Applied to all traffic or just to a particular session or sets of sessions.

5.1.1. Open Service Architecture (OSA). Second-generation mobile systems offered fully standardized services such as voice, fax, short messages, and supplementary services (call hold, call forward, call conference, etc.). However, it was difficult for operators to propose innovative services to attract the customer. To provide greater flexibility in service creation, the second phase of GSM standardization included the introduction of "toolkits": CAMEL (*c*ustomized *a*pplications of *m*obile network *e*nhanced *l*ogic[5]), an intelligent network concept for GSM, SIM toolkit (STK), and MExE (mobile execution environment), which includes WAP. These toolkits were used in GSM to introduce prepaid services (CAMEL) and mobile internet portals (WAP). In 3G, these principles are still valid, but efforts are focused on integrating the various toolkits in a single one called Open Service Architecture (OSA), which is, in fact, an application programming interface (API) based on PARLAY, a forum developing a

[5] CAMEL is based on an intelligent network architecture that separates service logic and database from the basic switching functions, and implements the Intelligent Network Application Protocol (INAP).

common API for the different networks. This new concept is still under development in 3GPP and will be introduced in post-R99 UMTS releases.

An important feature of open service architecture is that service design and provision can be ensured by companies other than the network operator.

5.1.2. Virtual Home Environment. The virtual home environment (VHE) concept, based on CAMEL, will provide customers with the same set of services whatever the location.

When a subscriber is roaming, his or her service profile, or even the service logic registered in the home location register (HLR), is transferred to the visited network to offer the required services with the same ergonomics.

5.2. Terminals

Mobile customers will be able to use one or several mobile terminals (Fig. 3), including regular mobile phones, pocket videophones, and PDAs to manage agendas, addresses, transportation, and email, and to send and receive multiple types of information. Typical handsets are smaller and lighter than 2G handsets: 100 g and 100 cm^3 [3] versus about 130 g on average for 2G. IMT-2000 enabled laptops will provide traveling professionals, executives, and employees with direct access to their corporate intranets, offering videoconferencing, cooperative worktools, and shared programs and network resources facilitating work outside the office. Specific applications will use IMT-2000 capacity to provide data, sound, and fixed or moving images. Among the most widely cited services are location-specific information services, remote control and monitoring, remote health maintenance services, and driving assistance (navigation aids, traffic and weather information, engine maintenance information, etc.). In these instances, IMT-2000 terminals can be standard equipment in vehicles, or coupled with the application devices used (e.g., health monitors).

With large-scale 3G network deployment and mass production of 3G terminals, significant cost reductions are expected to open up a mass market for personal multimedia tools. Young people, who have been nurtured with today's mobile phones and game consoles, will undoubtedly drive the development of this market through their needs for entertainment, education, and information.

A major change in 3G terminals with respect to second-generation mobile terminals lies in the replacement of the 2G subscriber identity module (SIM) card with a more general-purpose card called a universal integrated circuit card (UICC) of the same size. The UICC contains one or more user services identity modules (USIMs) as well as other applications. Communications-related advantages include enhanced security, the ability to use the same handset for business and private use and to roam from UMTS to GSM networks as needed. The UICC can also contain payment mechanisms (micro-payment, credit card), access badge functions, and user profile information.

6. IMT-2000 NETWORK ARCHITECTURE

This section describes the network architecture of the three most prevalent IMT-2000 standards: cdma2000, W-CDMA, and EDGE.

Network architecture is divided into the radio access network (RAN) and the core network (CN). To an increasing extent, efforts are focused on developing standard interfaces between functional domains to enable interworking between network elements manufactured by different suppliers.

6.1. CDMA 2000

6.1.1. Radio Access Network. The cdma2000 radio access network architecture, illustrated in Fig. 4, is similar to that of CDMA IS95A. In cdma2000 1x is mainly a software upgrade.

The key difference between the two architectures is the Ater reference point (A3 and A7), which allows the source BSC (base station controller) to manage soft handover of communications between two base transceiver stations (BTSs) belonging to different BSCs. In Fig. 4, the source BTS is the initial BTS managing the communication. The target BTS is the BTS asked to enter in communication with the mobile. As shown, the source BTS is still managing the communication and remains the primary link with the mobile switching center (MSC) whereas, in IS95A, the MSC always manages the soft handover between two different BSCs (traffic and signaling go from the MSC to the two BSCs).

Figure 3. Examples of mobile multimedia terminals for UMTS.

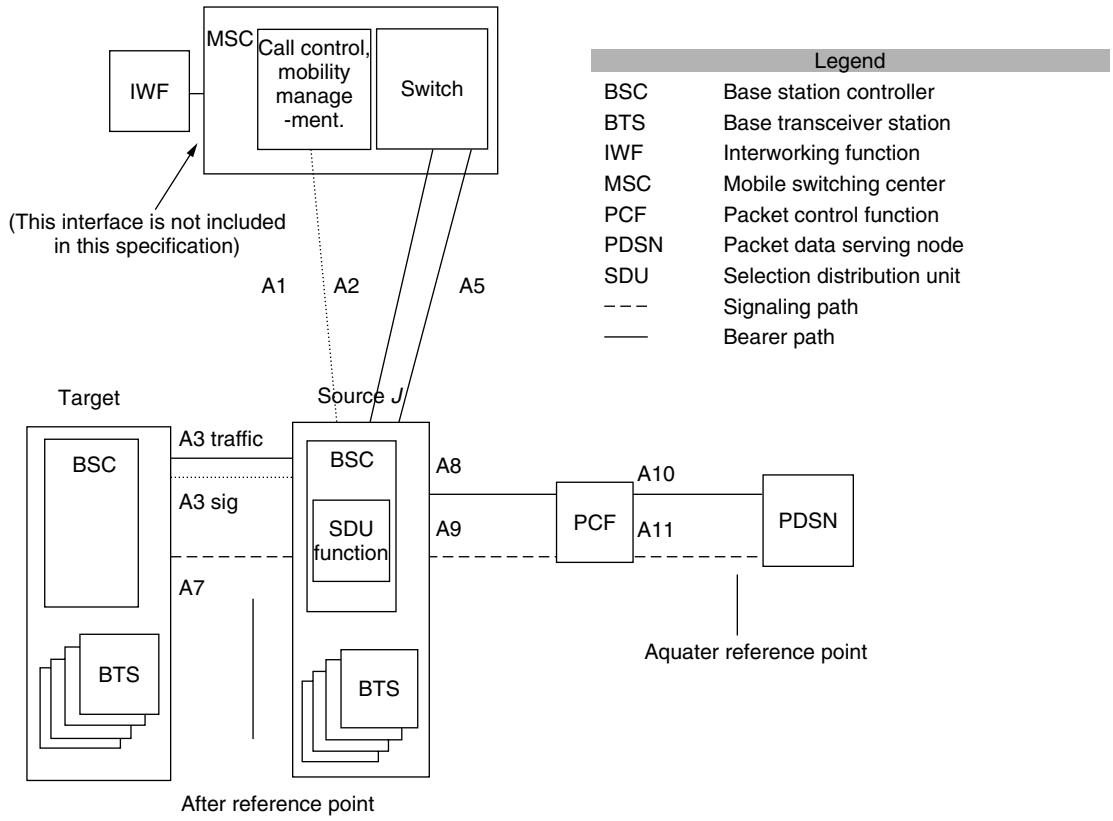


Figure 4. cdma2000 radio access network. (*Source*: 3GPP2 Website.)

Figure 4 depicts the logical architecture of the radio access network. It describes the overall system functions, including services and features required for interfacing a BTS with other BTSs, with the MSC for the circuit transmission mode, and with the packet control function (PCF) and the packet data serving node (PDSN) in the packet transmission mode.

The interfaces defined in this standard are described below.

A1 Carries signaling information between the call control (CC) and mobility management (MM) functions of the MSC and the call control component of the BTS (BSC).

A2 Carries 64/56-kbps pulse-code modulation (PCM) information (voice/data) or 64-kbps unrestricted digital information (UDI, for ISDN) between the MSC switch component and one of the following:
- The channel element component of the BTS (in the case of an analog air interface)
- The selection/distribution unit (SDU) function (in the case of a voice call over a digital air interface)

A3 Carries coded user information (voice/data) and signaling information between the SDU function and the channel element component of the BTS. The A3 interface is composed of two parts: signaling and user traffic. The signaling information is carried across a separate logical channel from the user traffic channel, and controls the allocation and use of channels for transporting user traffic.

A5 Carries a full duplex stream of bytes between the interworking function (IWF) and the SDU function.

A7 Carries signaling information between a source BTS and a target BTS.

A8 Carries user traffic between the BTS and the PCF.

A9 Carries signaling information between the BTS and the PCF.

A10 Carries user traffic between the PCF and the PDSN.

A11 Carries signaling information between the PCF and the PDSN.

A8, A9, A10, and A11 are all based on the use of Internet Protocol (IP). IP can operate across various physical layer media and link layer protocols. Conversely, A3 and A7 are based on ATM Transport; and A1, on SS7 Signaling.

In 2000, the Abis interface (between the BTS and the BSC) and Tandem Free Operation (TFO) for cdma2000 systems were also standardized.

6.1.2. Core Network. The 3GPP2 architecture is based on the same wireless intelligent network (WIN) concept as that developed by 2G mobile networks, but this structure is partially modified to introduce packet data technologies such as IP. The objective of these circuit-switched networks was to bring intelligent network (IN) capabilities, based on ANSI-41, to wireless networks in a seamless manner without making the network infrastructure obsolete. ANSI-41 was the standard backed by wireless providers because it facilitated roaming. It has capabilities for switching and connecting different systems

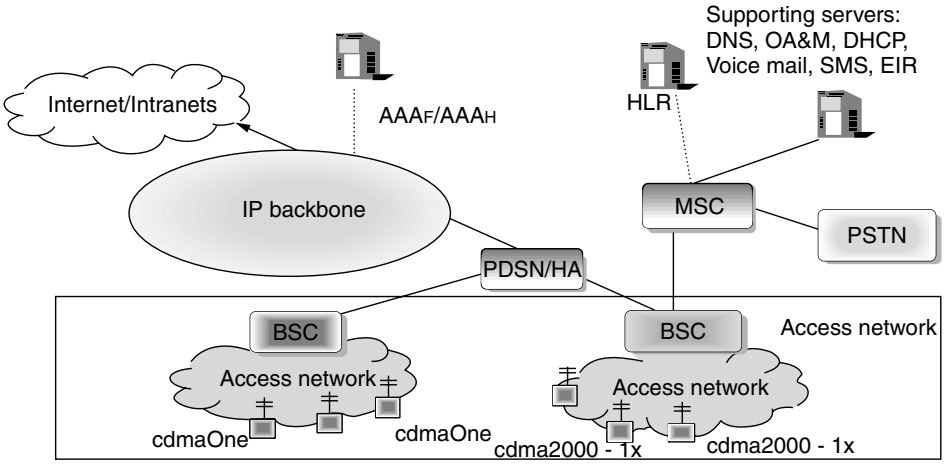


Figure 5. 3GPP2 architecture.

and the ability for performing direct connections between different elements of the WIN based on SS7 Signaling such as GSM/MAP.

Wireless data packet networking is based on the IS835 standard, whose architecture is illustrated in Fig. 5. This architecture provides access, via the cdma2000 air interface, to public networks (Internet) and private networks (Intranets) considering the required quality of service (QoS) and the pertinent accounting support. 3GPP2 strives to reuse IETF open standards whenever possible in order to keep a high level of interoperability with other cellular standards, and to increase marketability. These standards include mobile IP for interPDSN (packet data serving node) mobility management, radius for authentication, authorization and accounting, and differentiated services for the QoS.

Figure 6 introduces the general model for the 3GPP2 packet domain. This domain includes two modes, "simple IP" and "mobile IP." Only the mobile IP mode requires the use of home agent (HA) and PDSN/foreign agent (PDSN/FA) entities. It is clear that these two modes are not equivalent from a service point of view. Mobile IP supports mobility toward different PDSNs during a session. In this sense, the mobile IP solution is similar to the General Packet Radio System (GPRS — see Section 4.2). Simple IP only offers a connection to the packet domain without any real-time mobility between PDSNs. Simple IP supports mobility within a given PDSN.

It is important to note that this standard uses circuit-switched MSC resources (ANSI-41 Domain and SS7 Signaling) to handle both the voice and packet radio control resources (access registration, QoS profile based verification, paging, etc.).

This model will certainly be modified with the emergence of the all-IP network, which will support data capacities largely exceeding those offered in the

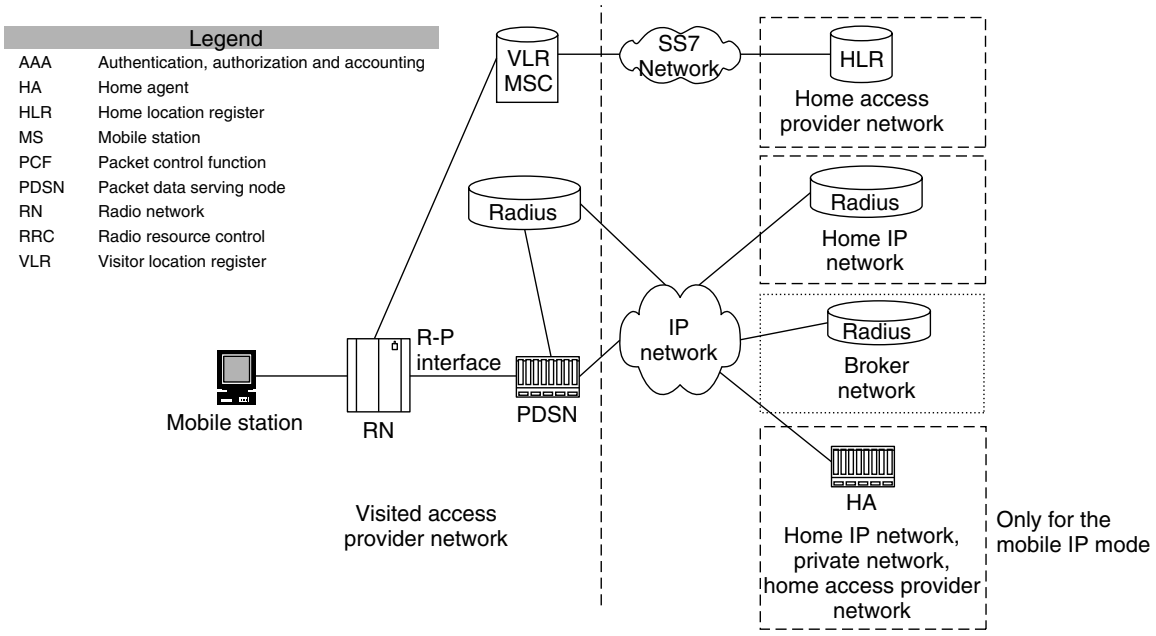


Figure 6. Simple IP and mobile IP according to 3GPP2. (*Source*: 3GPP2 Website.)

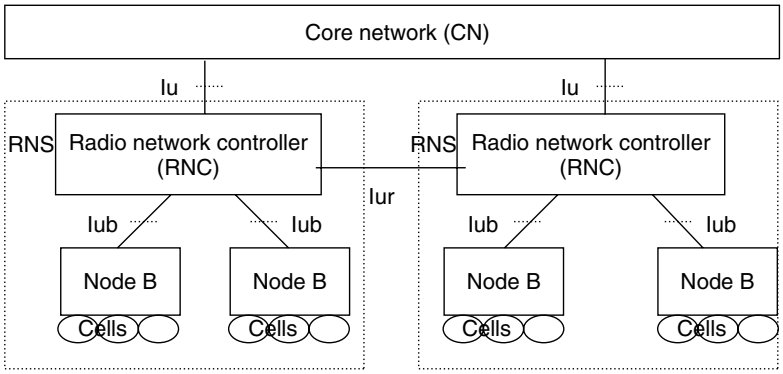


Figure 7. UMTS (W-CDMA) radio access network architecture

Table 4. Main Characteristics of TDD and FDD Modes [4]

Mode	FDD	TDD
Multiple-access method	DS-CDMA (direct-sequence CDMA)	TD/CDMA
Carrier chip rate	3.84 Mchips/s	
Channel spacing	5 MHz (nominal)	
Frame length	10 ms	
Frame structure	15 slots/frame	
Modulation	QPSK	
Spreading factors	4–512	1–16
Forward error correction (FEC) codes	Convolutional coding ($R = \frac{1}{2}$ or $\frac{1}{3}$, constraint length $K = 9$)	
	Turbo coding for BER $<10^{-3}$ (8-state PCCC $R = \frac{1}{3}$)	
	Service-specific coding	
Frequency bands	1920–1980 MHz — uplink (mobile to BTS) 2110–2170 MHz — downlink (BTS to mobile)	1900–1920 MHz 2010–2025 MHz
	Duplex separation — 190 MHz	

mixed circuit/packet data scheme. This new architecture is expected to be ready by the end of 2002. The future network will enable services such as voice over IP, multimedia calls, and video streaming, based on full use of IP Transport in both the radio access and core networks. This step will mark the end of the classical circuit-switched core network.

6.2. WCDMA

6.2.1. Radio Access Network

6.2.1.1. Deployment, Duplex Mode. WCDMA deployment involves a multilayer cellular network, with macrocells (0.5–10 km in range) for large-scale coverage, microcells (50–500 m) for hotspots, and picocells (5–50 m) for indoor coverage. Handover is ensured both among WCDMA cells and between WCDMA and GSM cells, without any perceptible cut or degradation of quality.

As indicated in Section 4, the air interface adopted by ETSI in January 1998 is based on two harmonized modes: FDD/WCDMA for the paired bands and TDD/TDCDMA for the unpaired bands. UMTS is to be deployed using at least two duplex 5-MHz bands, and must ensure interworking with GSM and dual-mode FDD/TDD operation.

FDD mode is appropriate for all types of cells, including large cells, but is not well adapted to asymmetric traffic. TDD is, by definition, more flexible to support traffic asymmetry, but it requires synchronization of the base stations, and is not appropriate for large cells because of the limited guard times between time slots. Table 4 lists the main characteristics of the two modes.

The WCDMA FDD mode is based on CDMA principles with a bandwidth of 5 MHz. One major difference with the IS95 standard is that no synchronization is required among base stations, thus allowing easier deployment for operators. One of the key advantages of WCDMA is its high spectral efficiency, or capacity per unit of spectrum [typically expressed in kilobits per second per megahertz or (kbps/MHz)]. Depending on the services offered, WCDMA offers 2–3 times the capacity of GSM with the same amount of spectrum.

The TDD mode is based on a mix between TDMA and CDMA. The TDD frame has 15 time slots and each time slot supports several simultaneous CDMA communications.

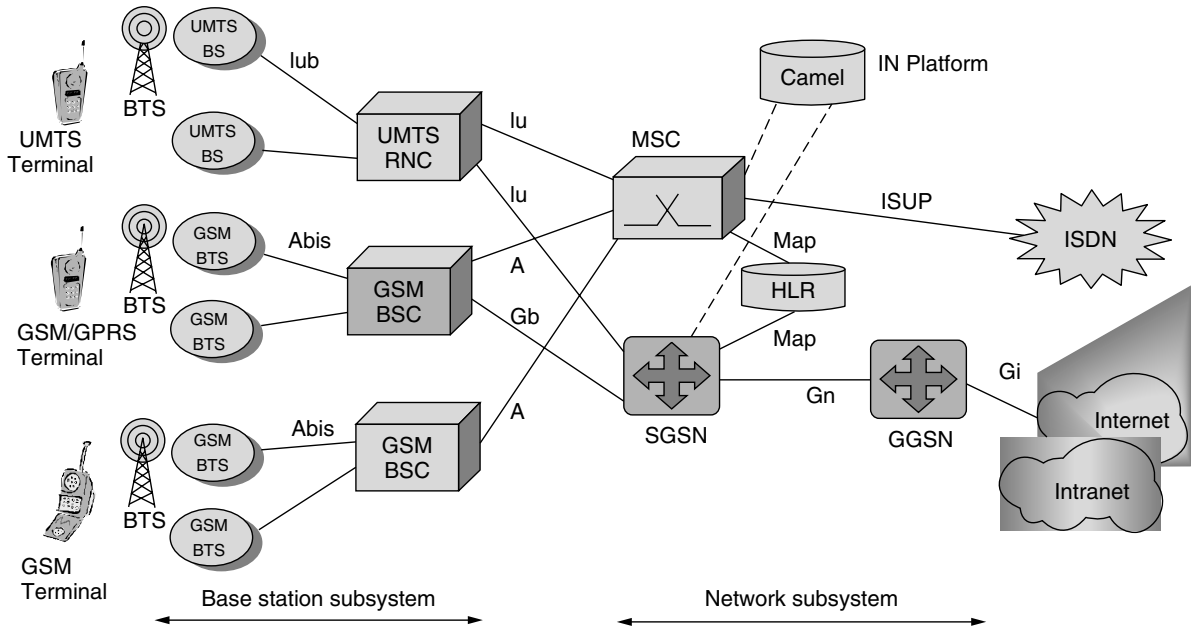


Figure 8. General architecture of the UMTS network (release 99).

6.2.1.2. Radio Access Network Architecture. Figure 8 represents the logical architecture of the UMTS radio access network. The radio network subsystem includes the radio base stations (node B) and the radio network controller (RNC).

This architecture is similar to that of the GSM radio access network. Iu represents the interface between the RNC and the core network. Iub represents the interface between the node B and the RNC. One key difference with GSM is the existence of the Iur interface between RNCs. This interface enables the management of soft handover between two node Bs belonging to two separate RNCs, independently from the core network. "Soft handover" means that the mobile terminal moving from one cell to another has links with both base stations during handover.

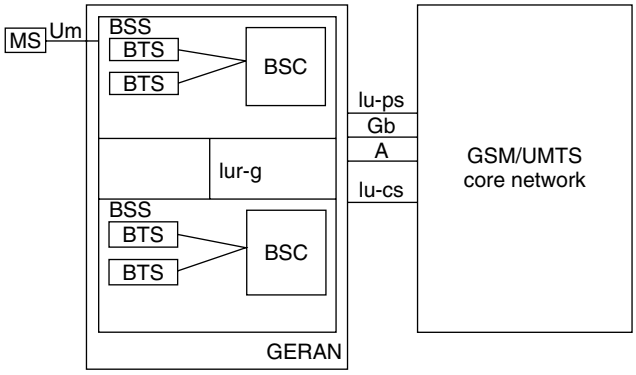


Figure 9. GSM/EDGE radio access network (GERAN) architecture.

6.2.1.3. WCDMA Radio Dimensioning. Dimensioning of the radio access network takes into account both coverage and capacity. To ensure physical coverage, the cell radius is based on link budgets calculated according to the different service types, propagation environment, indoor penetration assumptions, and loading factor. It is recalled that, in CDMA networks, each user contributes to the interference level, which causes the cell to "shrink," a phenomenon called "cell breathing." Generally, a loading factor of 50–70% is taken. This factor is based on a theoretical "100%" load at which interference tends to infinity.

On the basis of geomarketing data, the number of users per square kilometer and the average data rate per user during busy hour yields a load in terms of data rate (Mbps or kbps) or erlangs per square kilometer. The corresponding cell size is computed to handle this load. This cell area is matched against that computed from the link budgets, and when the offered load exceeds the capacity made available based on the coverage criterion, it is necessary to add carriers and/or sites.

6.2.2. WCDMA Core Network. The WCDMA network architecture is illustrated in Fig. 9. As mentioned, the radio access network is comprised of specific WCDMA base stations and RNCs. The network subsystem requires an SGSN (described in Section 4.2), which may be specific to the UMTS system or shared with the GPRS service, and a specific operations and maintenance center (OMC). Other NSS components, such as the DNS and GGSN, can also be shared with the existing GPRS system.

The UMTS Release 99 core network comprises two distinct domains: circuit-switched (CS) and packet-switched (PS), as in GSM/GPRS networks. The core network elements are the same as in these networks: MSC (mobile switching center) for CS services, and SGSN and GGSN for PS services.

In Release 99, ATM was chosen for transport in the access network. This choice makes it possible to support all the types of services offered: voice, circuit-switched data, and packet-switched data. Different ATM adaptation layers (AALs) are used: AAL2 for the voice or data on the Iu-cs circuit-switched domain interfaces, Iur and Iubis. AAL5 is used for signaling and for user data on the Iu-ps packed-switched domain interface.

6.3. EDGE

EDGE, as mentioned in Section 3.2., is a convergent solution for TDMA (IS136), and GSM.

Since July 2000, the 3GPP has been responsible for standardization of the GSM/EDGE Radio Access Network (GERAN). In order to harmonize this access technology with others developed for IMT-2000, notably WCDMA and cdma2000, the 3GPP chose to align the QoS requirements for EDGE with those of the other technologies (see Table 1). However, because the data rate varies with distance from the base station in EDGE, only interactive and background type services are expected to be offered initially, even though the EDGE standard encompasses enhanced circuit-switched data (ECSD) services.

The core network does not differ from that of a GPRS core network; however, because of the increased data rate available, capacity of the data links between base stations and BSCs, and between the BSCs and MSC and SGSN must be dimensioned appropriately. The radio access network is based on existing 2G infrastructure, as illustrated in Fig. 9.

The introduction of the Iu interfaces enables fast handover in the packet-switched domain, real-time services in the packet-switched domain, and enhanced multiplexing capabilities in both the packet- and circuit-switched domains. The A and Gb interfaces are maintained for compatibility with legacy systems.

The GERAN has been standardized for the following GSM frequency bands: 900, 1800, 1900, 400, and 700 MHz. Note that these bands do not include those allocated to IMT-2000 systems; hence this technology can be used by operators without IMT-2000 spectrum. However, it is recalled that the highest data rate possible is 384 kbps versus the theoretical maximum of 2 and 2.4 Mbps, for WCDMA and cdma2000, respectively.

7. LICENSE ASSIGNMENT AND ECONOMIC IMPLICATIONS OF IMT-2000 (4/2001)

As mentioned in Section 3.3, national regulators played a major role in shaping the development of IMT-2000 3G networks and services. In Europe, they focused on three important factors: 3G license cost and allocation mechanism, number of 3G operators per country, and rollout/coverage obligations. With an initial goal of encouraging competition to enhance service offerings and ensure reasonable prices for the end customer, regulatory agencies chose to allocate spectrum to at least one or two new entrants in each country, requiring, in most cases, incumbent GSM operators to sign roaming agreements with these new entrants, enabling them to offer their customers national coverage for voice service. The choice of assignment method — auctions or comparative hearings — led to highly divergent license costs in Europe. While auctions generated overall state revenues ranging from 85 to 630 Euros per capita (with Germany and the United Kingdom gaining the highest amounts), comparative hearings yielded revenues ranging from approximately 0 to 45 Euros per capita; a notable exception was France, where the license fee was set, shortly after the English auctions, at 335 Euros per capita. However, only two candidates submitted applications for the four licenses available; other candidates cited the high cost as the main deterrent. Rollout requirements, set primarily in conjunction with comparative hearings, generally call for initial deployment as of 2002, with coverage of the main cities and extension to 50–80% of the population within the following 6–8 years. In Japan, license costs were minimal, while in the United States, 3G spectrum auctions led to per capita costs similar to those observed in the United Kingdom and Germany.

The combined obligations of paying license fees and building out entirely new radio access networks, which represent some 80% of the initial IMT-2000 investment within a set time frame, have placed a considerable burden on both operators and equipment manufacturers. In Sweden, for example, operators have formed joint ventures to build the radio infrastructure in areas outside of the major metropolitan centers. In such cases, each operator exploits its own spectrum, but shares the cost of buildout in order to focus on service development [6]. Opinions diverge as to the economic prospects for IMT-2000. The outlook is optimistic in Japan, where customer awareness of mobile data services is high because of I-mode, and where willingness to pay for such services is among the highest in the world. Although Europe is well advanced in terms of second-generation mobiles penetration, mobile data services have gotten off to a slow start, with the disappointingly limited scope and speed of WAP. As better data rates and volume-based billing become available with GPRS, customers may more willingly adopt new data services offered by a wide panel of providers [6]. This “education” stage is considered crucial to the rapid adoption of IMT-2000.

In the United States, mobile data have generally been restricted to very low-data-rate exchanges, while fixed lines remain the preference for Internet access. Moreover, customers are used to free Internet content. This means that innovative location-based, customized services will have to be developed for operators to recoup their expenses.

In all cases, with respect to 2G, new players will be joining the value chain: content suppliers, service brokers, virtual mobile network operators, and network resource brokers. While some of these new players may simply be affiliates of today’s major mobile operators, some will be entirely new entities such as retailers, banks, insurance companies, and entertainment companies. This means that the corresponding 3G revenues, whether

from the end user, a third party, or advertising, will be spread more thinly than in the 2G world, leading in all likelihood to continentwide consolidation [7] of operators.

8. BEYOND 3G

With many uncertainties remaining as to the economic viability of IMT-2000 in the near term, research and development efforts are already turning toward the next phase, coined "beyond 3G." This phase, expected to emerge as of 2005, is based on seamless roaming among heterogeneous wireless environments, including 3G mobile networks and indoor wireless facilities such as radio LANs and Bluetooth networks. Another topic currently being explored is the cooperation of 3G networks and broadcast standards such as digital video or audio broadcasting (DVB, DAB).

The goal is to optimize the service offered to the end customer by taking advantage of the spectral resources available. For example, a train passenger can connect with the company intranet via the UMTS network with the available bit rate and quality inherent in this support service; then, when this passenger enters a train station or airport lounge equipped with a wireless LAN, the terminal detects the new network and vice versa, and switches over to this new broadband resource. If any background tasks are being performed, they are uninterrupted when the terminal goes from one environment to the other.

As a complement to IMT-2000, DVB-T can provide fast one-to-many services such as weather, sports scores, or stockmarket values. It can thus significantly ease the burden on IMT-2000 frequency resources which are then used for bi-directional wideband links.

9. CONCLUSION AND DISCUSSION

IMT-2000, a family of third-generation mobile standards, is on the verge of deployment. The prospect of offering a wide range of mobile multimedia services prompted numerous incumbent operators and new entrants to spend sometimes exorbitant amounts on 3G licenses. Undoubtedly, the first years will be characterized by turbulence in infrastructure rollout; technical teams will have to be trained, the manufacturers will have to ensure adequate levels of production, and — most importantly — new sites will have to be negotiated. Operators' finances will be strained as they seek to attract customers and generate adequate revenues early enough to offset the investments made. Smaller operators and new entrants with no infrastructure may find the expenditure and task too daunting and, willingly or unwillingly, merge with a larger competitor.

As networks and services reach cruising speed, operators will focus on enhancing services by finding new spectral resources in other frequency bands (e.g., GSM bands) and by encouraging cooperation among different wireless networks, both fixed and mobile. Underlying the success of this evolution are the continued efforts of standards organizations such as 3GPP and 3GPP2, the policies developed by national and international regulatory bodies that will have learned the lessons of 3G, the ability of operators and service suppliers to anticipate customers' needs and desires and, most importantly, the customers themselves.

Acknowledgments

The authors would like to thank their colleagues at France Télécom R&D as well as the partners of the ACTS 364/TERA project for their useful support and discussions.

BIBLIOGRAPHY

1. UMTS Forum, Report 5, 1998.
2. *3rd Generation Partnership Project Technical Specification* TS 23.107 v3.3.0, June 2000.
3. *Global Wireless Magazine*, Crain Communications Inc., March–April 2001, p. 18.
4. H. Holma and A. Toskala, eds., *WCDMA for UMTS — Radio Access for Third Generation Mobile Communications*, Wiley, Chichester, UK, 2000, p. 2285.
5. *Pyramid Research Perspective*, Europe, Feb. 2, 2001.
6. *Strategy Analytics SA Insight*, Dec. 8, 2000.
7. L. Godell et al., *Forrester Report: Europe's UMTS Meltdown*, Dec. 2000.

FURTHER READING

Useful Websites

Standards

ITU: *http://www.itu.int/imt/*
3GPP: *http://www.3gpp.org*
3GPP2: *http://www.3gpp2.org*
(member organizations can be accessed from these sites)

Manufacturers (Network Equipment and/or Terminals)

Nokia: *http://www.nokia.com*
Ericsson: *http://www.ericsson.com*
Motorola: *http://www.motorola.com*
Lucent Technologies: *http://www.lucent.com*
Qualcomm: *http://www.qualcomm.com*
Alcatel: *http://www.alcatel.com*
NEC: *http://www.nec.com*
Nortel Networks: *http://www.nortel.com*
Siemens: *http://www.siemens.com*
Toshiba: *http://www.toshiba.co.jp/worldwide/*
Sharp: *http://www.sharp-usa.com*
Sanyo: *http://www.sanyo.com*
Sony: *http://sony.com*
Panasonic, Matsushita Electric Industrial Co., Ltd.: *http://www.panasonic.co.jp/global/*

Kyocera Wireless Corp.: *http://www.kyocera-wireless.com*

Philips: *http://www.philips.com*

Journals

Global Wireless Magazine: *http://www.globalwireless-news.com*

Books

T. Ojanperä and R. Prasad, eds., *Wideband CDMA for Third Generation Mobile Communications*, Artech House, Boston, 1998.

ACRONYMS

AMPS	American Mobile Phone System
ANSI	American National Standards Institute
API	Application programming interface
ARIB	Association of Radio Industries and Businesses (Japan)
BSC	Base station controller
BSS	Base station subsystem
BTS	Base transceiver station
CAMEL	Customized Applications of Mobile Network Enhanced Logic
CATT	Chinese Academy of Telecommunication Technology
CDMA	Code-division multiple access
CN	Core network
CS	Circuit switched
CWTS	China Wireless Telecommunication Standards Group
DAB	Digital audio broadcasting
D-AMPS	Digital AMPS
DECT	Digital European Cordless Telephony
DHCP	Dynamic Host Configuration Protocol
DNS	Domain name server
DVB	Digital video broadcasting
DVB-T	Digital video broadcasting-terrestrial
ECSD	Enhanced circuit-swtiched data
EDGE	Enhanced Data for GSM Evolution
FCC	Federal Communications Commission (USA)
FDD	Frequency-division duplex
GERAN	GSM EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HA	Home agent
HDR	High Data Rate standard
HLR	Home location register
HS	High-speed
HSCSD	High-speed circuit-switched data
HTML	Hypertext markup language
IETF	Internet Engineering Task Force
IMT	International Mobile Telecommunications
IN	Intelligent network
INAP	Intelligent network application protocol
IP	Internet Protocol
ITU	International Telecommunications Union
IWF	Inter-working function
LAN	Local area network
MAC	Medium access control
MAP	Mobile application part
MC	Multicarrier
MExE	Mobile execution environment
MSC	Mobile switching center
MS	Mobile station
MT	Mobile terminal
NSS	Network subsystem
NTIA	National Telecommunications and Information Administration (USA)
OA&M	Operations, administration and maintenance
OHG	Operators Harmonization Group
OMC	Operations and maintenance center
PCF	Packet control function
PCM	Pulse code modulation
PCS	Personal communications system
PCU	Packet control unit
PDA	Personal digital assistant
PDSN	Packet data serving node
PS	Packet-switched
QOS	Quality of service
RAN	Radio access network
RLC	Radio link control
RN	Radio network
RNC	Radio network controller
RRC	Radio resource control
RTT	Radio transmission technology
SDU	Selection distribution unit
SGSN	Serving GPRS Support Node
SIM	Subscriber identity module
SMS	Satellite-based mobile service; Short message service
STK	SIM tool kit
TACS	Total Access Communication System (UK)
TDD	Time-division duplex
TDMA	Time Division Multiple Access
TD-SCDMA	Time-division synchronous CDMA
TFO	Tandem Free Operation
TIA	Telecommunications Industry Association (North America)
TTA	Telecommunications Technology Association (Korea)
TTC	Telecommunications Technology Committee (Japan)
UDI	Unrestricted digital information
UICC	Universal integrated circuit card
UMTS	Universal Mobile Telecommunication System
USIM	User services identity module
UWCC	Universal Wireless Communications Consortium
VHE	Virtual home environment
WAP	Wireless Applications Protocol
WARC	World Administrative Radio Conference
W-CDMA	Wideband code-division multiple access
WIN	Wireless intelligent network
WML	Wireless markup language

INFORMATION THEORY

RANDALL BERRY
Northwestern University
Evanston, Illinois

1. INTRODUCTION

The key concepts for the field of information theory were introduced by Claude E. Shannon in the landmark two-part paper "A mathematical theory of communication" [1,2] published in 1948. The emphasis of Shannon's work was on understanding the fundamental limits of communication systems. Subsequently, this theory has expanded and is now widely recognized as providing the theoretical framework for modern digital communication systems. The following surveys some of the basic ideas of information theory. In addition to communication theory, information theory has proved useful in a variety of other fields, including probability, statistics, physics, linguistics, and economics. Our emphasis here is on those aspects related to communication systems, specifically, the areas closest to Shannon's original work; such topics are also referred to as Shannon theory.

The basic problems to be discussed concern a generic communication system as in Fig. 1. The object is to convey a message, generated by the information source, to the destination. Information sources are modeled probabilistically, namely, the source generates a message from a set of possible messages according to a given probability distribution. The transmitter takes this message and maps it into a signal that is then transmitted over the communication channel. The received signal may differ from the transmitted signal as a result of additive noise and other channel impairments; these are also modeled in a probabilistic framework. The receiver attempts to extract the original message from the received signal. The above framework is general enough to accommodate a wide range of systems. For example, the message could be written text, an analog speech signal or digital data; possibilities for the channel include an optical fiber, the atmosphere or a storage medium.

As shown in Fig. 1, the transmitter and receiver are often divided into two stages. At the transmitter, the source coder first represents the incoming message as a binary sequence; the channel coder maps this binary sequence into the transmitted signal. The corresponding inverse operations are performed at the receiver. This two-stage implementation has many practical advantages. For example, this allows the same channel coder/decoder to be used for several different information sources. Moreover, a key result of information theory, the *joint source channel coding theorem*, states that for a wide range of situations, the source coder and channel coder can be designed separately without any loss of optimality.

Two of the main questions addressed by this theory concern the achievable performance of the source coder/decoder and channel coder/decoder:

1. What is the minimum number of bits needed to represent messages from a given source?
2. What is the maximum rate that the information can be transmitted over the channel with arbitrarily small probability of error?

The first question addresses data compression; the second, reliable communication. In Shannon's original work, both of these questions where posed and largely answered for basic models of the source and channel. The answer to these questions is given in terms of two basic information measures, *entropy* and *mutual information*.

In the following, we first discuss the basic information measures used in this theory. We then discuss the role of these quantities in the problems of source coding and channel coding.

2. INFORMATION MEASURES

In information theory, information sources and communication channels are modeled probabilistically. The information measures used are defined in terms of these stochastic models. First, we give the definitions of these measures for discrete random variables, For example, a random variable X taking values in a finite set χ. The set χ is often called the *alphabet* and its elements are called *letters*. In Section 2.2, continuous random variables are addressed.

2.1. Discrete Models

The first information measure we consider is the *entropy* $H(X)$ of a random variable X with alphabet χ. This

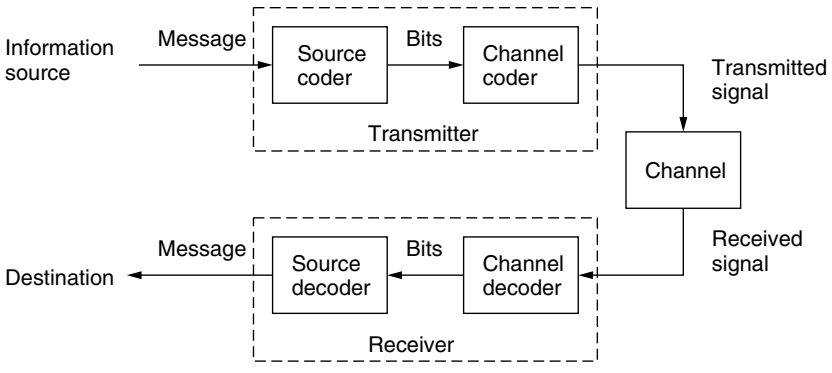


Figure 1. Model of a generic communication system.

quantity is defined by

$$H(X) = -\sum_{x\in\chi} p(x)\log p(x)$$

where $p(x) = \Pr(X = x)$ is the probability mass function of X. The base of the logarithm in this definition determines the units in which entropy is measured. Common bases are 2 and e with corresponding units called *bits* and *nats*, respectively. Entropy can be interpreted as a measure of the uncertainty in a random variable. For example, entropy satisfies

$$0 \le H(X) \le \log|\chi|$$

where $|\chi|$ denotes the size of the alphabet. The upper bound is achieved if and only if the random variable is uniformly distributed on this set; this corresponds to maximum uncertainty. The lower bound is achieved if and only if the random variable is deterministic, which corresponds to minimum uncertainty.

The significance of entropy arises from considering long sequences of random variables. Using entropy, the set of all sample sequences of a given length N can be divided into a set of *typical* sequences, which occur with high probability, and a set of atypical sequences, which occur with a probability that vanishes as N becomes large. Specifically, let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent, identically distributed (i.i.d.) random variables, where each X_i has entropy $H(X)$. We denote by $p(x^N)$ the probability of a sequence x^N of length N. For a given $\delta > 0$, the set of all sequences x^N whose probability satisfies

$$\left|\frac{-\log p(x^N)}{N} - H(X)\right| < \delta$$

is called the *typical set*.[1] For any $\delta > 0$, as N increases, the probability of the typical set approaches one. This result, proved by Shannon, is known as the "asymptotic equipartition property" (AEP). A consequence of the AEP is that the size of the typical set is approximately $2^{NH(X)}$; thus, entropy characterizes the rate at which the typical set grows with N. As will be discussed in the next section, this result has a natural source coding interpretation.

The entropy of two or more random variables is defined analogously; For instance, for two random variables X and Y, we obtain

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y)$$

where $p(x,y)$ is the joint probability mass function. For a discrete stochastic process, $\{X_i\}_{i=1}^{\infty}$, the *entropy rate* is defined to be the asymptotic per letter entropy:

$$H_\infty(X) = \lim_{n\to\infty}\frac{1}{n}H(X_1,\ldots,X_n)$$

Assuming that the process is stationary will ensure that this limit exists. Using the entropy rate, the AEP can be generalized to every stationary ergodic process with a finite alphabet; this result is known as the *Shannon–McMillan theorem* [4].

The *conditional entropy* of Y given X is defined by

$$H(Y\mid X) = -\sum_{x,y} p(x,y)\log p(y\mid x)$$

where $p(y\mid x) = \Pr(Y = y\mid X = x)$. Conditional entropy represents the average uncertainty in Y when X is given. From these definitions, various relations can be derived such as

$$H(X,Y) = H(X) + H(Y\mid X)$$

The next information measure we discuss is the *mutual information* between two random variables X and Y. This is defined by

$$I(X;Y) = H(X) - H(X\mid Y)$$

or equivalently

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

Mutual information is a measure of the reduction of uncertainty in X when Y is given. This quantity is symmetric: $I(X;Y) = I(Y;X)$. It is also nonnegative and equal to zero only when X and Y are independent.

Mutual information is a special case of another quantity called *relative entropy*.[2] Relative entropy is a measure of the difference between two probability mass functions, $p(x)$ and $q(x)$. This quantity is defined as

$$D(p\|q) = \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right).$$

Relative entropy is nonnegative and only equal to zero when $p = q$. The mutual information between X and Y is given by the relative entropy between the joint distribution $p(x,y)$ and the product of the marginal distributions $p(x)p(y)$.

2.2. Continuous Models

Corresponding information measures can be defined for continuous valued random variables. In this case, instead of entropy, the appropriate measure is given by the *differential entropy*.

Differential entropy is defined by

$$h(x) = -\int f(x)\log f(x)\,dx$$

where $f(x)$ denotes the probability density function of the random variable X. Differential entropy has many similar properties to entropy for a discrete random variable,

[1] These sequences are also called *weakly typical*. A stronger version of typicality can be defined using the method of types [3].

[2] Relative entropy is also referred to by a variety of other names including *Kullback–Leibler distance* and *information divergence*.

however it does not share all of the properties. One important difference is that differential entropy depends on the choice of coordinates, while entropy is invariant to coordinate transformations.

The mutual information between two continuous random variables is given by

$$I(X;Y) = h(X) - h(X \mid Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy$$

Unlike differential entropy, mutual information is invariant to coordinate transformations, and is a natural generalization of its discrete counterpart.

3. SOURCE CODING

Source coding or *data compression* refers to the problem of efficiently representing messages from a given information source as binary sequences.[3] Here efficiency refers to the length of the resulting representation.

Information sources are modeled as random processes, depending on the situation several different models may be appropriate. One class of model consists of discrete sources; in this case, the source is modeled as a sequence of discrete random variables. When these random variables are i.i.d. (independent, identically distributed), this model is called a *discrete memoryless source* (DMS). Another class of sources are analog source models, where the output of the source is a continuous-time, continuous-valued waveform. For discrete sources, compression may be *lossless*; that is, the original source output can be exactly recovered from the encoded bit sequence. On the other hand, for analog sources, compression must be *lossy*, since an infinite number of bits are required to represent any real number. In this case, source coding consists of sampling the source and quantizing the samples. We first discuss lossless source coding of discrete sources, here the information theoretic limitations are provided by the source's entropy rate. In Section 3.5 we consider analog sources, in this case the fundamental limits are characterized by the *rate distortion function*, which is defined using mutual information.

3.1. Classification of Source Codes

For a discrete source, a source code is a mapping or encoding of sequences generated by the source into a corresponding binary strings. An example of a source code for a source with an alphabet $\{a, b, c, d\}$ is

$$a \leftrightarrow 00 \qquad b \leftrightarrow 01 \qquad c \leftrightarrow 10 \qquad d \leftrightarrow 11$$

Notice this code is lossless, because each source letter is assigned a unique binary string. This is a fixed-length code, since each codeword has the same size. Source codes may also be variable-length, as in the Morse code used for English text. The performance of a variable-length source code is given by the expected code length, averaged over the source statistics. This will clearly be smaller if shorter codewords are used for more likely symbols. Instead of encoding each source letter individually as in the above example, a source code may encode blocks of source letters at once.

For variable-length codes, it is important to be able to tell where one codeword ends and the next begins. For example, consider a code which assigns 0 to a and 00 to b; in this case the string 00 could represent aa or b. A code with the property that any sequence of concatenated codewords can be correctly decoded is called *uniquely decodable*. A special class of uniquely decodable codes are codes where no codeword is the prefix of another; these are called *prefix-free codes* or *instantaneous codes*. The class of uniquely decodable codes is larger than the class of prefix-free codes. However, it can be shown that the performance attained by any uniquely decodable code can also be achieved by a prefix-free code, that is, it is sufficient to only consider prefix-free codes. This results follows from the *Kraft inequality*, which states that a prefix-free code can be found with codewords of lengths $l_1, l_2, \ldots, l_k$ if and only if

$$\sum_{i=1}^{k} 2^{-l_i} \leq 1$$

The set of codeword lengths for any uniquely decodable code must also satisfy this inequality.

3.2. Variable-Length Source Codes

In this section we discuss the performance of variable-length, uniquely decodable codes. Suppose that the code encodes a single letter from a DMS at a time. From the Kraft inequality, it can be shown that the minimum average length of any such code must be greater than or equal to the source's entropy. For a given distribution of source letters, a variable-length source code with minimal average length can be found using a simple iterative algorithm discovered by Huffman [5]. The average length of a Huffman code can be shown to be within one bit of source's entropy

$$H(X) \leq L_H \leq H(X) + 1$$

where $H(X)$ is the source's entropy and L_H is the average codeword length of the Huffman code. The above can be extended by considering a code that encodes blocks of N source letters at a time. A Huffman code can then be designed by treating each block as a single "supersymbol." In this case, if L_H^N is the average codeword length, then the compression ratio (the average number of encoded bits per source symbol) satisfies

$$H(X) \leq \frac{L_H^N}{N} \leq H(X) + \frac{1}{N}$$

Hence, as N becomes large, one can achieve a compression ratio that is arbitrarily close to the source's entropy. At times, this result is referred to as the *variable-length source coding theorem*. The converse to this theorem

[3] This easily generalizes to the case where messages are to be represented as strings in an arbitrary, finite-sized *code alphabet*; we focus on the binary case here.

states that no lossless, variable-rate code can achieve a compression ratio smaller than the sources entropy. This can be generalized to discrete sources with memory; in this case, the entropy rate of the source is used.

3.3. The AEP and Shannon's Source Coding Theorem

The AEP, discussed in Section 2.1, provides additional insight into the attainable performance of source codes. Consider a DMS with an alphabet of size K. There are K^L possible sequences of length L that the source could generate. Encoding these with a fixed-length, lossless source code requires approximately $L\log_2(K)$ bits, or $\log_2(K)$ bits per source letter. However, from the AEP, approximately $2^{LH(X)}$ of these sequences are typical. Therefore a lossy source code can be designed that assigns a fixed-length codeword to each typical sequence and simply disregards all nontypical sequences. This requires approximately $H(X)$ bits per source letter. The probability of not being able to recover a sequence is equal to the probability that the sequence is atypical, which becomes negligible as L increases. Making this more precise, it can be shown that for any $\varepsilon > 0$ a source code using no more than $H(X) + \varepsilon$ bits per source letter can be found with arbitrary small probability of decoding error. This is the direct part of *Shannon's source coding theorem*. The converse to this theorem states that any code that uses fewer than $H(X)$ bits per source letter will have a probability of decoding failure that increases to one as the block length increases.[4]

3.4. Universal Source Codes

Approaches such as Huffman codes require knowledge of the source statistics. In practice it is often desirable to have a data compression algorithm that does not need a priori knowledge of the source statistics. Among the more widely used approaches of this type are those based on two algorithms developed in 1977 and 1978 by Lempel and Ziv [6,7]. For example, the 1978 version is used in the UNIX compress utility. The 1977 algorithm is based on matching strings of uncoded data with strings of data already encoded. Pointers to the matched strings are then used to encode the data. The 1978 algorithm uses a dictionary that changes dynamically based on previous sequences that have been encoded. Asymptotically, the Lempel–Ziv algorithms can be shown to compress any stationary, ergodic source to its entropy rate [8]; such an approach is said to be *universal*.

3.5. Lossy Compression: Rate Distortion Theory

Next we consider encoding analog sources. We restrict our attention to the case where the source generates real-valued continuous-time waveforms.[5] The basic approach for compressing such a source is to first sample the waveform, which yields a sequence of continuous-valued random variables. If the source is a band-limited, stationary random process, then it can be fully recovered when sampled at twice its bandwidth (the Nyquist rate). After sampling, the random variables are then quantized so that they can be represented with a finite number of bits. Quantization inherently introduces some loss or distortion between the original signal and the reconstructed signal at the decoder. Distortion can be reduced by using a finer quantizer, but at the expense of needing more bits to represent each quantized value. The branch of information theory called rate-distortion theory investigates this tradeoff.

A scalar quantizer can be regarded as a function that assigns a quantized value or reconstruction point $\hat{x}$ to each possible sample value x. A rate R quantizer has 2^R quantization points. The loss incurred by quantization is quantified via a distortion measure, $d(x,\hat{x})$, defined for each sample value and reconstruction point. A common distortion measure is the squared error distortion given by

$$d(x,\hat{x}) - (x-\hat{x})^2$$

Given a probability distribution for the sample values and a fixed rate, the set of quantization points that minimize the expected distortion can be found using an iterative approach call the Lloyd–Max algorithm [9].

Instead of quantizing one sample at a time, performance can be improved by using a vector quantizer, that is, by quantizing a block of N samples at once. A rate R vector quantizer consists of 2^{NR} reconstruction points, where each point is an N dimensional vector. In this case, the distortion between a sequence of sample values and the corresponding reconstruction points is defined to be the average per sample distortion. A distortion D is said to be achievable at rate R if there exists a sequence of rate R quantizers with increasing block size N for which the limiting distortion is no greater than D. For each distortion D, the infimum of the rates R for which D is achievable is given by the *operational rate distortion function*, $R_{\text{op}}(D)$.

For an i.i.d. source $\{X_n\}$, the (*Shannon*) *rate distortion function*, $R(D)$, is defined to be

$$R(D) = \min I(X;\hat{X})$$

where the minimization is over all conditional distributions of $\hat{X}$ given X which yield a joint distribution such that the expected value of $d(X,\hat{X})$ is no greater than D. Here X is a scalar random variable with the same distribution as X_n. For a large class of sources and distortion measures, rate distortion theorems have been proved showing that $R_{\text{op}}(D) = R(D)$. Therefore, the fundamental limits of lossy compression are characterized by the information rate distortion function. As an example, consider an i.i.d. Gaussian source with variance σ^2. In this case, for squared error distortion, the rate distortion function is given by

$$R(D) = \begin{cases} \dfrac{1}{2}\log\dfrac{\sigma^2}{D}, & 0 \le D \le \sigma^2 \\ 0, & D > \sigma^2 \end{cases}$$

[4] This is sometimes referred to as the *strong converse* to the source coding theorem. The *weak converse* simply states that if fewer than $H(X)$ bits per source letter are used then the probability of error can not be zero.

[5] More generally one can consider cases where the source is modeled as a vector-valued function of time or a vector field.

In addition to i.i.d. sources, rate distortion theory has been generalized to a large variety of other sources [e.g., 10].

3.6. Distributed Source Coding

An interesting extension of the preceding models are various distributed source coding problems, that is situations where there are two or more correlated information sources being encoded at different locations. One information theoretic result for this type of situation is known as the *Slepian–Wolf theorem* [11]; this theorem implies the surprising result that the best (lossless) compression achievable by the two sources without cooperating is as good as what can be achieved with complete cooperation. For distributed lossy compression, characterizing the rate distortion region remains an open problem.

4. CHANNEL CODING

We now turn to the problem of communicating reliably over a noisy channel. Specifically given a bit stream arriving at the channel coder, the object is to reproduce this bit stream at the channel decoder with a small probability of error. In addition, it is desirable to transmit bits over the channel with as high a rate as possible. A key result of information theory is that for a large class of channels, it is possible to achieve arbitrarily small error probabilities provided that the data rate is below a certain threshold, called the *channel capacity*. As with the rate distortion function, capacity is defined in terms of mutual information. Conversely, at rates above capacity, the error probability is bounded away from zero. Results of this type are known as *channel coding theorems*.

Mathematically a channel is modeled as a set of possible input signals, a set of possible output signals and a set of conditional distributions giving the probability for each output signal conditioned on a given input signal. As with sources, several different types of channel models are studied. One class consists of discrete channels where the input and output signals are sequences from a discrete alphabet. Another class of channels consists of waveform channels where the input and output are viewed as continuous functions of time; the best-known example is the band-limited additive white Gaussian noise channel.

5. DISCRETE MEMORYLESS CHANNELS

A *discrete memoryless channel* (DMC) is a discrete channel, where the channel input is a sequence of letters $\{x_n\}$ chosen from a finite alphabet. Likewise, the channel output is also a sequence $\{y_n\}$ from another finite alphabet. Each output letter y_n depends statistically only on the corresponding input letter x_n. This dependence is specified by a set of transition probabilities $P(y \mid x)$ for each x and y. Two examples of DMC's are shown in Fig. 2. On the left is a binary symmetric channel, where both the input and output alphabet are $\{0, 1\}$. Each letter in the input sequence is reproduced exactly at the output with probability $1 - \varepsilon$ and is converted into the opposite letter with probability ε. On the right of Fig. 2 is a binary erasure

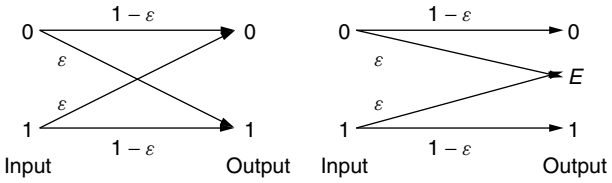


Figure 2. A binary symmetric channel (left) and a binary erasure channel (right).

channel; in this case the input alphabet is $\{0, 1\}$, but the output alphabet is $\{0, 1, E\}$. Each input letter is either received correctly, with probability $1 - \varepsilon$, or is received as an erasure E, with probability ε.

Uncoded transmission over a binary symmetric channel consists of simply mapping each bit arriving at the channel coder into the corresponding input symbol. The decoder would map the received signal back into the corresponding bit. In this case, one bit is sent in each channel use and the probability of error is simply ε. To reduce the probability of error, an error-correcting code may be used. For example, suppose each bit is repeated 3 times when transmitted over the channel. In this case, if at most one bit out of three is in error, the decoder can correct the error. This reduces the error probability, but also reduces the transmission rate by a factor of 3.

The preceding example is called a *repetition code*. More generally, a (M, n) block code for a discrete channel consists of a set of M codewords, where each codeword is a sequence of n symbols in the channel input alphabet. The encoder is a function that maps each sequence of $\log_2 M$ data bits into one of these codewords and the decoder maps each possible received sequence of length n back into one of the original sequences of data bits. In the previous example $M = 2$ and $n = 3$. The *rate* of the code is

$$R = \frac{\log_2 M}{n} \quad \text{bits per channel use}$$

In a (M, n) code, the probability of error for each codeword, $e_i, i = 1, \ldots, M$, can be calculated from the channel model and the specification of the decoder. The *maximal error probability* for a code is defined to be $\max_i e_i$, and the *average error probability* of the code is defined to be $(1/M) \sum_i e_i$.

For a DMC, if the input letters are chosen with the probability distribution $p(x)$, then the joint probability mass function of the channel input and output is given by $p(x, y) = p(x)p(y \mid x)$. Using this probability mass function, the mutual information between the channel input and output can be calculated for any input distribution. The channel capacity of a DMC is defined to be the maximum mutual information

$$C = \max I(X; Y)$$

where the maximization is over all possible probability distributions on the input alphabet. In many special cases, the solution to this optimization can be found in closed form; for a general DMC, various numerical approaches may be used, such as the Arimoto–Blahut algorithm [12].

The definition of capacity is given operational significance by the channel coding theorem, proved by Shannon for DMCs. The direct part states that for any $R < C$, there exists rate R codes with arbitrarily small probability of error. This applies for either the maximal or the average probability of error. The converse to the channel coding theorem[6] states that for any rate $R > C$, the probability of error is bounded away from zero. The proof of the converse relies on Fano's inequality, which relates the probability of error to the conditional entropy of the channel output given the channel input.

Shannon's proof of the direct part of the coding theorem relied on a *random coding argument*. In this approach, the expected error probability for a set randomly chosen codes is studied. By showing that this expectation can be made small, it follows that there must be at least one code in this set with an error probability that is also small. In fact, it can be shown that most codes in this set have low error probabilities.

At rates below capacity, the probability of error for a good code can be shown to go to zero exponentially fast as the block length of the code increases. For a each rate R, the fastest rate at which the probability of error can go to zero is given by the channel's *reliability function*, $E(R)$. Various upper and lower bounds on this function have been studied [e.g., 13].

In addition to DMC, coding theorems have been proven for a variety of other discrete channels, we briefly mention several examples. One class of channels are called *finite-state channels*. In a finite state channel each input symbol is transmitted over one a set of possible DMCs. The specific channel is determined by the channel state that is modeled as a Markov chain. The capacity of a finite state channel will depend on if the transmitter or receiver has any *side information* available about the channel state [14]. A related model is a *compound channel*. In this case, one channel from a set is chosen when transmission begins; the channel then remains fixed for the duration of the transmission. *Universal coding* for compound channels has also been studied [15], which parallels the universal source coding discussed in Section 3.4. Another example is a channel with feedback, where the transmitter receives information about what is received. For a DMC it has been shown that feedback does not increase capacity [16]; however, feedback may reduce the complexity needed.

5.1. Gaussian Channels

A widely used model for communication channels is the additive white Gaussian noise (AWGN) channel shown in Fig. 3. This is a waveform channel with output $Y(t)$ given by

$$Y(t) = X(t) + Z(t)$$

where $X(t)$ is the input signal and $Z(t)$ is a white Gaussian noise process. Assuming that the input is limited to a bandwidth of W, then using the sampling theorem this channel can be reduced to the equivalent discrete-time channel:

$$Y_n = X_n + Z_n$$

Here, $\{X_n\}$ and $\{Y_n\}$ are sequences representing the input and output samples respectively and $\{Z_n\}$ is a sequence of i.i.d. Gaussian random variables with zero mean and variance $N_0/2$, where N_0 is the one-sided power spectral density of $Z(t)$. The samples occur at a rate of $2W$ samples per second.

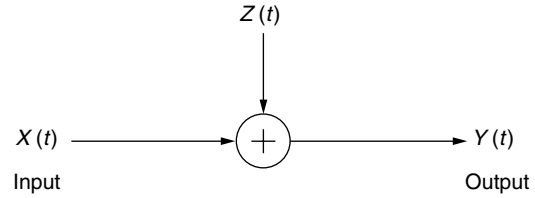


Figure 3. An additive white Gaussian noise channel.

The input to a channel is often required to satisfy certain constraints. A common example is an average power constraint, namely, a requirement that

$$\frac{1}{T}\int_T X^2(t)\,dt \le P$$

or, for the discrete time channel,

$$\frac{1}{N}\sum_{n=1}^{N} X_n^2 \le \frac{P}{2W}$$

The capacity of the AWGN channel with an average power constraint is given by the well-known formula

$$C = W\log\left(1 + \frac{P}{N_0 W}\right) \quad \text{bits per second (bps)}$$

where P/N_0W is the signal-to-noise ratio (SNR).

Many variations of the AWGN channel have also been studied, including the generalization of those discussed for DMCs above. One variation is channel with colored Gaussian noise, where the noise has power spectral density $N(f)$. In this case, achieving capacity requires that the transmission power be allocated over the frequency band. The optimal power allocation is given by

$$P(f) = \max\{\lambda - N(f), 0\}$$

where λ is chosen to satisfy the average power constraint. This allocation is often referred to as the "water-pouring" solution, because the power used at each frequency can be interpreted as the difference between $N(f)$ and the "water level" λ. Motivated in part by wireless applications, time-varying Gaussian channels and multiple-input/multiple-output Gaussian channels have also been widely studied.

6. NETWORK INFORMATION THEORY

Network information theory refers to problems where there are more than one transmitter and/or receiver.

[6] This is also known as the "weak converse to the coding theorem." The "strong converse" states that at rates above capacity the probability of error goes to one for codes with long enough lengths.

A variety of such problems have been considered. This includes the *multiaccess channel*, where several users are communicating to a single receiver. In this case, a *capacity region* is specified indicating the set of all rate pairs (for the two-user case). Coding theorems have been found that establish the multiple access capacity region for both the discrete memoryless case [17] and for the Gaussian case [18]. Another example is the *interference channel*, where two senders each transmit to a separate receivers. The complete capacity region for this channel remains open.

7. FURTHER STUDIES

For more in-depth reading there are a variety of textbooks on information theory including Cover [19] and Gallager [13]. Shannon's original papers [1,2] are quite readable; all of Shannon's work, including [1,2], can be found in the compilation by Sloane and Wyner [20]. A collection of good survey articles can also be found in the volume edited by [21].

BIOGRAPHY

Randall Berry received the B.S. degree in Electrical Engineering from the University of Missouri-Rolla in 1993 and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 1996 and 2000, respectively. Since 2000, he has been an Assistant Professor in the Department of Electrical and Computer Engineering at Northwestern University. In 1998 he was on the technical staff at MIT Lincoln Laboratory in the Advanced Networks Group. His primary research interests include wireless communication, data networks, and information theory.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communication (Part 1), *Bell Syst. Tech. J.* **27**: 379–423 (1948).
2. C. E. Shannon, A mathematical theory of communication (Part 2), *Bell Syst. Tech. J.* **27**: 623–656 (1948).
3. I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
4. B. McMillan, The basic theorems of information theory, *Ann. Math. Stat.* **24**: 196–219 (June 1953).
5. D. Huffman, A method for the construction of minimum redundancy codes, *Proc. IRE* **40**: 1098–1101 (Sept. 1952).
6. J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans. Inform. Theory* **24**: 337–343 (May 1977).
7. J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Trans. Inform. Theory* **24**: 530–536 (Sept. 1978).
8. A. Wyner and J. Ziv, The sliding window Lempel-Ziv algorithm is asymptotically optimal, *Proc. IEEE* **82**(6): 872–877 (June 1994).
9. S. Lloyd, *Least Squares Quantization in PCM*, Bell Laboratories Technical Note, 1957.
10. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
11. D. Slepian and J. Wolf, Noiseless coding of correlated information sources, *IEEE Trans. Inform. Theory* **19**: 471–480 (1973).
12. R. Blahut, Computation of channel capacity and rate distortion functions, *IEEE Trans. Inform. Theory* **18**: 460–473 (1972).
13. R. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.
14. J. Wolfowitz, *Coding Theorems of Information Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1978.
15. A. Lapidoth and P. Narayan, Reliable communication under channel uncertainty, *IEEE Trans. Inform. Theory* **44**: 2148–2175 (Oct. 1998).
16. C. Shannon, The zero error capacity of a noisy channel, *IRE Trans. Inform. Theory* **2**: 112–124 (Sept. 1956).
17. R. Ahlswede, Multi-way communication channels, *Proc. 2nd Int. Symp. Information Theory*, 1971, pp. 103–135.
18. T. Cover, Some advances in broadcast channels, *Adv. Commun. Syst.* **4**: 229–260 (1975).
19. T. Cover, *Elements of Information Theory*, Wiley, New York, 1991.
20. N. J. A. Sloane and A. D. Wyner, eds., *Claude Elwood Shannon: Collected Papers*, IEEE Press, New York, 1993.
21. S.Verdú, ed., *IEEE Trans. Inform. Theory* (*Special Commemorative Issue*) **44**: (Oct. 1998).

INTERFERENCE AVOIDANCE FOR WIRELESS SYSTEMS

DIMITRIE C. POPESCU
CHRISTOPHER ROSE
Rutgers WINLAB
Piscataway, New Jersey

1. INTRODUCTION

Interference from natural sources and from other users of the medium have always been factors in the design of reliable communication systems. While for wired or optical systems the amount of interference may be limited through hardware means that restrict access to the medium and/or reduce relative noise energy, the wireless medium is shared by all users. Usually, restrictions on use are legislative in nature and imposed by government regulating agencies such as the Federal Commission for Communications (FCC). Specifically, the oldest method of interference mitigation for wireless systems is spectrum licensing and the implied exclusive spectrum use. Unfortunately, licensing can indirectly staunch creative wireless applications owing to high licensing fees and the concomitant need for stable returns on large investments.

In an effort to promote creative wireless applications, the FCC has released 300 MHz of unlicensed spectrum

in the 5 GHz range [1] with few restrictions other than absolute power levels — a "Wild West" environment of sorts where the only restriction is the size of the weapon! Thus, no central control is assumed. Needless to say, such a scenario seems ripe for chaos. Specifically, self-interest by individual users often results in unstable behavior. Such behavior has been seen anecdotally in distributed channelized systems such as early cordless telephones in apartment buildings where the number of collocated phones exceeded the number of channels available. Phones incessantly changed channels in an attempt to find a usable one. Thus, some means of assuring efficient use in such distributed wireless environments is needed.

Reaction to mutual interference is the heart of the shared-spectrum problem, and traditional approaches to combating wireless interference start with channel measurement and/or prediction, followed by an appropriate selection of modulation methods and signal processing algorithms for reliable transmission — possibly coupled to exclusive use contracts (licensing). Interestingly, since the time of the first radio transmissions, the methods used to deal with interference can be loosely grouped into three categories:

- Build a fence (licensing)
- Use only what you need (efficient modulation, power control)
- Grin and bear it (signal processing at the receiver)

Examples of the first item are legion, while examples of the last two items include single-sideband amplitude modulation, frequency modulation with preemphasis at the transmitter and deemphasis at the receiver, power control in cellular wireless systems [2], and code-division multiple access (CDMA) coupled to sophisticated signal processing algorithms for interference suppression/cancellation at the receiver [3].

However, Moore's law advances in microelectronics have led to the emergence of new transceiver hardware that add a new weapon to the interference mitigation arsenal. Specifically, a class of radios that can be programmed to transmit almost arbitrary waveforms and can act as almost arbitrary receiver types is emerging — the so-called software radios [4–7]. So, as opposed to traditional radios, which owing to complex transceiver hardware are difficult to modify once a modulation method has been chosen, one can now imagine programming transceivers to use more effective modulation methods. Thus, wireless systems of the near future will be able to choose modulation methods that *avoid* ambient interference as opposed to precluding it via sole-use licenses, overpowering it with increased transmission power, or mitigating it with receiver signal processing.

Interference avoidance is the term used for adaptive modulation methods where individual users — simply put — employ their signal energy in "places" where interference is weak. Such methods have been shown to optimize shared use. More precisely, iterative interference avoidance algorithms yield optimal waveforms that maximize the signal-to-interference plus noise-ratio (SINR) for all users while maximizing the sum of rates at which all users can reliably transmit information (sum capacity). In other words, interference avoidance methods, through the self-interested action of each user, lead to a socially optimum[1] equilibrium (Pareto efficient [9,10]) in various mutual interference "games."

Interference avoidance was originally introduced in the context of "chip-based" DS-CDMA systems [11] and minimum mean-square error (MMSE) receivers, but was subsequently developed in a general signal space [12–14] framework [15,16] which makes them applicable to a wide variety of communication scenarios. Related methods for transmitter and receiver adaptation have also been used in the CDMA context for asynchronous systems [17] and systems affected by multipath [18].

The relationship between codeword assignment in a CDMA system and sum capacity has been studied in several papers [19,20]; the paper by Viswanath and Anantharam [20] provides an algorithm to obtain sum capacity optimal codeword ensembles in a finite number of steps — and perhaps more importantly, also shows that the optimal linear receiver for such ensembles is a *matched filter* for each codeword. Interference avoidance algorithms also yield optimal codeword ensembles but seem conceptually simpler and suitable for distributed adaptive implementation in multiuser systems.

It is worth expanding upon this last point. Interference avoidance is envisioned as a distributed method for unlicensed bands as opposed to a centralized procedure done by an omnescient receiver. Of course, we will see that the mathematics of the algorithm also lends itself to central application, so the distinction is really only important for practical application. However, throughout we will assume distributed application which implicitly suggests that each user knows its associated channel and in addition that each user has access to the whole system covariance through a side-channel beacon. The receiver can adaptively track codeword variation in a manner reminiscent of adaptive equalization. Since communication is two-way and physical channels are reciprocal, it is not unreasonable to assume that both the user and the system can know the channel. More important is the rate at which the channel varies. We will assume that channel variation is slow relative the frame rate [21,22], or if the channel variation rate is rapid that the average channel varies slowly enough for interference avoidance to be applied [23].

2. THE EIGEN-ALGORITHM FOR INTERFERENCE AVOIDANCE

We consider the uplink of a synchronous CDMA communication system with L users having signature waveforms $\{S_\ell(t)\}_{\ell=1}^{L}$ of finite duration T, with equal received power at the base station and ideal channels. The received signal is

$$R(t) = \sum_{\ell=1}^{L} b_\ell S_\ell(t) + n(t) \tag{1}$$

[1] Maximum sum capacity or *user capacity* [8] in a single-receiver multiuser system.

where b_ℓ represents the information symbol sent by user ℓ with signature $S_\ell(t)$, and $n(t)$ is an additive Gaussian noise process. We assume that all signals are representable in an arbitrary N-dimensional signal space. Hence, each user's signature waveform $S_\ell(t)$ is equivalent to an N-dimensional vector $\mathbf{s}_\ell$ and the noise process $n(t)$ is equivalent to a noise vector $\mathbf{n}$. The equivalent received signal vector $\mathbf{r}$ at the base station is

$$\mathbf{r} = \sum_{\ell=1}^{L} b_\ell \mathbf{s}_\ell + \mathbf{n} \tag{2}$$

By defining the $N \times L$ matrix S having as columns the user codewords $\mathbf{s}_\ell$

$$\mathbf{D} = \begin{bmatrix} | & | & & | \\ \mathbf{s}_1 & \mathbf{s}_2 & \dots & \mathbf{s}_L \\ | & | & & | \end{bmatrix} \tag{3}$$

the received signal can be rewritten in vector matrix form as

$$\mathbf{r} = \mathbf{S}\mathbf{b} + \mathbf{n} \tag{4}$$

where $\mathbf{b} = [b_1 \cdots b_L]^T$ is the vector containing the symbols sent by users.

Assuming simple matched filters at the receiver for all users and unit energy codewords $\mathbf{s}_k$, the SINR for user k is

$$\gamma_k = \frac{(\mathbf{s}_k^T \mathbf{s}_k)^2}{\sum_{j=1, j\neq k}^{L} (\mathbf{s}_k^T \mathbf{s}_j)^2 + E[(\mathbf{s}_k^T \mathbf{n})^2]} = \frac{1}{\mathbf{s}_k^T \mathbf{R}_k \mathbf{s}_k} \tag{5}$$

where $\mathbf{R}_k$ is the correlation matrix of the interference plus noise seen by user k having the expression

$$\mathbf{R}_k = \mathbf{S}\mathbf{S}^T - \mathbf{s}_k \mathbf{s}_k^T - \mathbf{W} \tag{6}$$

where $\mathbf{W} = E[\mathbf{n}\mathbf{n}^T]$ is the correlation matrix of the additive Gaussian noise.

Interference avoidance algorithms maximize the SINR through adaptation of user codewords. This is also equivalent to minimizing the inverse SINR defined as

$$\beta_k = \frac{1}{\gamma_k} = \mathbf{s}_k^T \mathbf{R}_k \mathbf{s}_k \tag{7}$$

Note that for unit power codewords, Eq. (7) represents the Rayleigh quotient for matrix $\mathbf{R}_k$, and recall from linear algebra [21, p. 348] that equation (7) is minimized by the eigenvector corresponding to the minimum eigenvalue of the given matrix $\mathbf{R}_k$. Therefore, the SINR for user k can be maximized by replacing codeword $\mathbf{s}_k$ with the minimum eigenvector of the correlation matrix $\mathbf{R}_k$. That is, user k *avoids interference* by seeking a place in the signal space where interference is least. Sequential application by all users of this greedy procedure defines the minimum eigenvector algorithm for interference avoidance, or the *eigen-algorithm* [15], formally stated below:

1. Start with a randomly chosen codeword ensemble specified by the codeword matrix $\mathbf{S}$.
2. For each user $\ell = 1 \cdots L$, replace user ℓ's codeword $\mathbf{s}_\ell$ with the minimum eigenvector of the correlation matrix $\mathbf{R}_k$ of the corresponding interference-plus-noise process.
3. Repeat step 2 until a fixed point is reached for which further modification of codewords will bring no improvement.

It has been shown [22] that in a colored noise background a variant of this algorithm, in which step 3 is augmented with a procedure to escape suboptimal fixed points, converges to the optimal fixed point where the resulting codeword ensemble "waterfills" over the background noise energy and maximizes sum capacity. If the background noise is white and the system is not overloaded (fewer users than signal space dimensions $L \leq N$), the algorithm yields a set of orthonormal codewords that corresponds to an ideal situation when users are orthogonal and therefore noninterfering. In the case of overloaded systems ($L > N$) in white noise, the resulting codeword ensembles form Welch bound equality (WBE) sets [19], which also minimize total squared correlation [15], a measure of the total interference in the system. In both underloaded and overloaded cases, the absolute minimum attainable total squared correlation is often used as a stopping criterion for the eigen-algorithm.

Finally, we note that signal space "waterfilling" and the implied maximization of sum capacity are *emergent* properties of interference avoidance algorithms. Thus, individual users do not attempt maximization of sum capacity via an individual or ensemble waterfilling scheme, but rather, they greedily maximize the SINR of their own codeword. In fact, individual waterfilling schemes over the whole signal space are impossible in this framework since each user's transmit covariance matrix $\mathbf{X}_\ell = \mathbf{s}_\ell \mathbf{s}_\ell^T$ is of rank one and cannot possibly span an N-dimensional signal space. So, emergent waterfilling and sum capacity maximization is a pleasantly surprising property of interference avoidance algorithms.

3. GENERALIZING THE EIGENALGORITHM

In order to extend application of the eigen-algorithm to more general scenarios, we consider the general multiaccess vector channel defined by [23]

$$\mathbf{r} = \sum_{\ell=1}^{L} \mathbf{H}_\ell \mathbf{x}_\ell + \mathbf{n} \tag{8}$$

where $\mathbf{x}_\ell$ of dimension N_ℓ is the input vector corresponding to user ℓ ($\ell = 1, \dots, L$), $\mathbf{r}$ of dimension N is the received vector at the common receiver corrupted by additive noise vector $\mathbf{n}$ of the same dimension, and $\mathbf{H}_\ell$ is the $N \times N_\ell$ channel matrix corresponding to user ℓ. It is assumed that $N \geq N_\ell, \forall \ell = 1, \dots, L$. This is a general approach to a multiuser communication system in which different users reside in different signal subspaces, with possibly different dimensions and potential overlap between them, but all of which are subspaces of the receiver signal space. We note that each user's signal space as well as the receiver signal space are of finite dimension — implied by a finite transmission frame $\mathcal{T}$, finite bandwidths W_ℓ for each

user ℓ, respectively, and by a finite receiver bandwidth W (which includes all W_ℓ values corresponding to all users) [24]. We also note that for memoryless channels the channel matrix $\mathbf{H}_\ell$ merely relates the bases of user ℓ's signal space and receiver signal space, but a similar model applies to channels with memory, in which case the channel matrix $\mathbf{H}_\ell$ also incorporates channel attenuation and multipath [16,18,23,25]. Figure 1 provides a graphical illustration of such a signal space configuration for two users residing in 2-dimensional subspaces with a three-dimensional receiver signal space.

In this signal space setting we assume that in a finite time interval of duration $\mathcal{T}$, each user ℓ sends a "frame" of data using a multicode CDMA approach wherein each symbol is transmitted using a distinct signature waveform that spans the frame. This scenario is depicted in Fig. 2. In other words, the sequence of information symbols $\mathbf{b}_\ell = [b_1^{(\ell)} \cdots b_{M_\ell}^{(\ell)}]^T$ is transmitted as a linear superposition of distinct, unit energy waveforms $s_m^{(\ell)}(t)$

$$x_\ell(t) = \sum_{m=1}^{M_\ell} b_m^{(\ell)} s_m^{(\ell)}(t) \tag{9}$$

as if each symbol in the frame corresponded to a distinct virtual user.

In the N_ℓ-dimensional signal space corresponding to user ℓ, each waveform can be represented as an N_ℓ-dimensional vector, thus the input vector $\mathbf{x}_\ell$ corresponding to user ℓ is equivalent to a linear superposition of unit norm codeword column vectors $\mathbf{s}_m^{(\ell)}$ scaled by the corresponding $b_m^{(\ell)}$. Therefore, each user uses an $N_\ell \times M_\ell$ codeword matrix $\mathbf{S}_\ell$

$$\mathbf{S}_\ell = \begin{bmatrix} | & | & & | \\ \mathbf{s}_1^{(\ell)} & \mathbf{s}_2^{(\ell)} & \cdots & \mathbf{s}_{M_\ell}^{(\ell)} \\ | & | & & | \end{bmatrix} \tag{10}$$

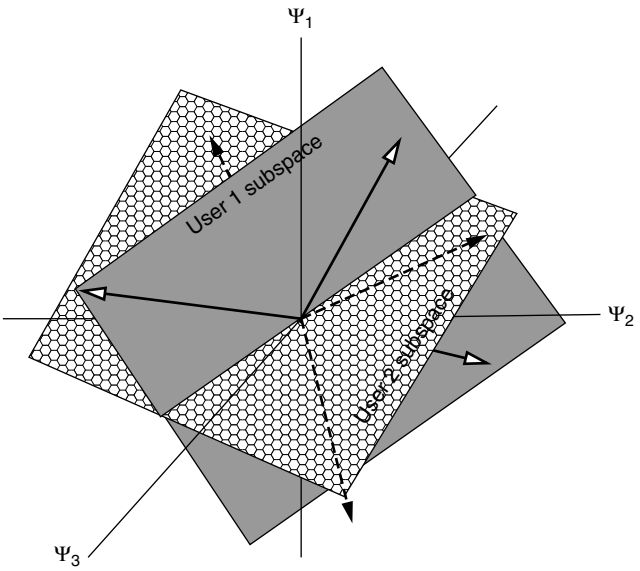


Figure 1. Three-dimensional receiver signal space with two users residing in two-dimensional subspaces. Vectors represent particular signals in user 1 (continuous line), respectively, user 2 (dashed line) signal spaces.

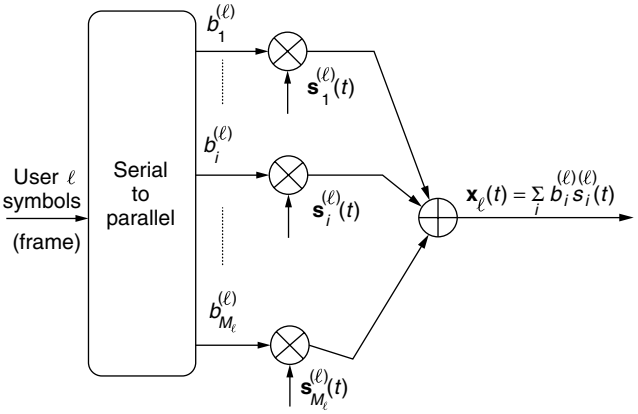


Figure 2. Multicode CDMA approach for sending frames of information. Each symbol in user ℓ's frame is assigned a distinct signature waveform, and the transmitted signal is a superposition of all signatures scaled by their corresponding information symbols.

so that

$$\mathbf{x}_\ell = \mathbf{S}_\ell \mathbf{b}_\ell \tag{11}$$

Therefore, the received signal can be rewritten as

$$\mathbf{r} = \sum_{\ell=1}^{L} \mathbf{H}_\ell \mathbf{S}_\ell \mathbf{b}_\ell + \mathbf{n} \tag{12}$$

Note that under the assumption that $M_\ell \geq N_\ell$, the $N_\ell \times N_\ell$ transmit covariance matrix of user ℓ, $\mathbf{X}_\ell = E[\mathbf{x}_\ell \mathbf{x}_\ell^T] = \mathbf{S}_\ell \mathbf{S}_\ell^T$, has full rank and spans user ℓ's signal space.

Extension of the eigen-algorithm to this general multiaccess vector channel setting is presented elsewhere in the literature [16,26]. The procedure starts by separating the interference-plus-noise seen by a given user k and rewriting the received signal in equation (12) from the perspective of user k as

$$\mathbf{r} = \mathbf{H}_k \mathbf{S}_k \mathbf{b}_k + \underbrace{\sum_{\ell=1, \ell \neq k}^{L} \mathbf{H}_\ell \mathbf{S}_\ell \mathbf{b}_\ell + \mathbf{n}}_{\mathbf{z}_k = \text{interference + noise}} = \mathbf{H}_k \mathbf{S}_k \mathbf{b}_k + \mathbf{z}_k \tag{13}$$

The covariance matrix of the interference-plus-noise seen by user k

$$\mathbf{Z}_k = E[\mathbf{z}_k \mathbf{z}_k^T] = \sum_{\ell=1, \ell \neq k}^{L} \mathbf{H}_\ell \mathbf{S}_\ell \mathbf{S}_\ell^T \mathbf{H}_\ell^T + \mathbf{W} \tag{14}$$

is then used to define a whitening transformation $\mathbf{T}_k$ of the interference-plus-noise seen by user k. The equivalent problem in which user k sees white interference-plus-noise is then projected onto the user k signal space using the singular value decomposition (SVD) [21, p. 442] of user k's transformed channel matrix. This reduces the problem to an equivalent one given by an equation identical in form to Eq. (4) and therefore allows straightforward application of the eigen-algorithm for optimization of user k's codewords.

One possible generalized eigen-algorithm is formally stated below:

1. Start with a randomly chosen codeword ensemble specified by user codeword matrices $\{\mathbf{S}_k\}_{k=1}^{L}$.
2. For each user $k = 1 \cdots L$
   a. Compute the transformation matrix $\mathbf{T}_k$ that whitens the interference-plus-noise seen by user k.
   b. Change coordinates and compute the transformed user k channel matrix $\tilde{\mathbf{H}}_k = \mathbf{T}_k \mathbf{H}_k$.
   c. Apply SVD for $\tilde{\mathbf{H}}_k$ and project the problem onto user k's signal space.
   d. Adjust user k's codewords sequentially using the greedy procedure of the basic eigen-algorithm; the codeword corresponding to symbol m of user k is replaced by the minimum eigenvector of the autocorrelation matrix of the corresponding interference-plus-noise process.
   e. Iterate the previous step until convergence (making use of escape methods [22] if the procedure stops in suboptimal points).
3. Repeat step 2 until a fixed point is reached.

We note that in steps 2d,e of the algorithm, user k waterfills its signal space by application of the basic eigen-algorithm [22] for interference avoidance by regarding all other user's signals as noise. Thus, the generalized eigen-algorithm *iteratively waterfills* each user's signal space. It has been shown [23] that an *iterative waterfilling algorithm* converges to a fixed point where the sum capacity of the vector multiaccess channel is maximized, which implies that the generalized eigen-algorithm will always yield a codeword ensemble that maximizes sum capacity.

4. THE GENERALIZED EIGEN-ALGORITHM: A VERSATILE TOOL FOR CODEWORD OPTIMIZATION

The generalized eigen-algorithm is a powerful tool that enables application of interference avoidance methods to various communication problems in which the underlying model is a multiaccess vector channel. Among these we mention codeword optimization in the uplink of a CDMA system with nonideal (dispersive) channels, multiuser systems with multiple inputs and outputs (MIMO), and asynchronous CDMA systems, for which an appropriate selection of signal space basis functions leads to particular cases of the general multiaccess vector channel model for which application of the generalized eigen-algorithm to codeword optimization becomes straightforward.

For the uplink of a CDMA system with dispersive channels considered in earlier studies [16,25], the spanning set of the signal space consists of a set of real sinusoids (sine and cosine functions) that are approximately eigenfunctions for all uplink channels corresponding to all users. Introduced in 1964 [27], channel eigenfunctions form an orthonormal spanning set with the property that their corresponding channel responses are also orthogonal, thus allowing convenient representation of channel outputs as scaled versions of the input vectors. In this case, the channel matrix of a given user is a diagonal matrix in which diagonal elements correspond to channel gain factors for the frequencies that define the signal space, and the modulation scheme turns out to be a form of multicarrier CDMA [28]. We note that even though interference avoidance is applied in this multicarrier modulation framework [25], the method is completely general and applicable to various scenarios with appropriate selection of signal space basis functions. For example using sinc functions the method is applicable to time-domain representations in which the vector channel model is obtained from Nyquist-sampled waveforms (see, e.g., Ref. 29). Using "time chips" as basis functions a vector model for DSCDMA systems with multipath is obtained [16,18].

Application of the generalized eigen-algorithm to multiuser MIMO systems is also possible [16,30]. The same multicarrier modulation framework with multiple antennas at the transmitter and receiver imply a MIMO channel matrix composed of block diagonal matrices corresponding to each transmit/receive antenna pair. We mention again that other MIMO channel models — for example, the spatiotemporal MIMO channel model [31], in which the MIMO channel matrix is composed of convolution matrices corresponding to each transmit/receive antenna pair — are perfectly valid and the generalized eigen-algorithm can be used for codeword optimization in conjunction with such models as well.

A similar signal space approach is used to apply the generalized eigen-algorithm for codeword optimization in very general asynchronous CDMA system models [16,32]. We note that for particular cases less general interference avoidance algorithms can be used [33].

5. CONCLUSION

Motivated by the emergence of software radios and the desire to foster creative uses of wireless spectrum in unlicensed bands, interference avoidance methods have been developed for wireless systems. The underlying idea of interference avoidance is for each user to greedily optimize spectrum use (SINR or capacity) through appropriate signal placement in response to interferers. Interference avoidance is applicable to a wide range of communications scenarios including dispersive channels, multiple antenna systems, and asynchronous systems.

The utility of interference avoidance methods in real wireless systems with multiple receivers — such as is found in a cellular environment — although currently under study, is still an open question in general. Specifically, theoretical results have been established for application of interference avoidance methods in a *collaborative* scenario wherein information from all receivers is pooled and used to decode all users in the system [34,35]. Since all information is available for use in decoding users, the collaborative scenario constitutes a best case of sorts. Unfortunately, real systems may not be collaborative or even cooperative.

Early experiments with geographically dispersed users assigned to different bases showed unstable behavior under direct application of the eigen-algorithm, mirroring the anecdotally reported behavior of early cordless phones (see Section 1, above). However, by allowing each user to send and adapt *multiple codewords* — a

multicode approach similar to that used for dispersive channels — the algorithm became stable [36] since each user could then waterfill their signal energy over the entire signal space when necessary as opposed to choosing exactly one channel. Of course, such convergence though welcome is a bit chimeric since the implied multiple receiver system model is an instance of the interference channel [37] for which very little is known in general. Regardless, the fact that interference avoidance can attain any equilibrium [9,10] in the implied "game" of mutually interfering wireless access is interesting, and can perhaps illuminate paths toward greater understanding of efficient and peaceful coexistence in unlicensed wireless systems.

BIOGRAPHIES

Dimitrie C. Popescu received the Engineering Diploma and M.S. degrees in 1991 from the Polytechnic Institute of Bucharest, Romania, and the Ph.D. degree from Rutgers University in 2002, all in Electrical Engineering. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, the University of Texas at San Antonio. His research interests are in the general areas of communication systems, control engineering, and signal processing. In the summer of 1997 he worked at AT&T Labs in Florham Park, New Jersey, on signal processing algorithms for speech enhancement, and in the summer of 2000 he worked at Telcordia Technologies in Red Bank, New Jersey, on wideband CDMA systems. His work on interference avoidance and dispersive channels was awarded second prize in the AT&T Student Research Symposium in 1999.

Dr. Christopher Rose received the B.S. (1979), M.S. (1981), and Ph.D. (1985) degrees all from the Massachusetts Institute of Technology in Cambridge, Massachusetts. Dr. Rose joined At&T Bell Laboratories in Holmdel, New Jersey as a member of the Network Systems Research Department in 1985 and in 1990 moved to Rutgers University, where he is currently an Associate Professor of Electrical and Computer Engineering and Associate Director of the Wireless Networks Laboratory. He is Editor for the *Wireless Networks* (ACM), *Computer Networks* (Elsevier) and Transactions on Vehicular Technology (IEEG) journals and has served on many conference technical program committees. Dr. Rose was technical program Co-Chair for MobiCom'97 and Co-Chair of the WINLAB Focus'98 on the U-NII, the WINLAB Berkeley Focus'99 on Radio Networks for Everything and the Berkeley WINLAB Focus 2000 on Picoradio Networks. Dr. Rose, a past member of the ACM SIGMobile Executive Committee, is currently a member of the ACM MobiCom Steering Committee and has also served as General Chair of ACM SIGMobile MobiCom 2001 (Rome, July 2001). In December 1999 he served on an international panel to evaluate engineering teaching and research in Portugal.

His current technical interests include mobility management, short-range high-speed wireless (Infostations), and interference avoidance methods for unlicensed band networks.

BIBLIOGRAPHY

1. Federal Communications Commission, *FCC Report and Order 97-5: Amendment of the Commission's Rules to Provide for Operation of Unlicensed NII Devices in the 5 GHz Frequency Range*, ET Docket 96–102, 1997.
2. R. Yates, A framework for uplink power control in cellular radio systems, *IEEE J. Select. Areas Commun.* **13**(7): 1341–1348 (Sept. 1995).
3. S. Verdu, *Multiuser Detection*, Cambridge Univ. Press, 1998.
4. I. Seskar and N. Mandayam, A software radio architecture for linear multiuser detection, *IEEE J. Select. Areas Commun.* **17**(5): 814–823 (May 1999).
5. I. Seskar and N. Mandayam, Software defined radio architectures for interference cancellation in DS-CDMA systems, *IEEE Pers. Commun. Mag.* **6**(4): 26–34 (Aug. 1999).
6. Special issue on software radio, *IEEE Pers. Commun. Mag.* **6**(4) (Aug. 1999) K.-C. Chen, R. Prasad, and H. V. Poor, eds.
7. J. Mitola, The software radio architecture, *IEEE Commun. Mag.* **33**(5): 26–38 (May 1995).
8. P. Viswanath, V. Anantharam, and D. Tse, Optimal sequences, power control and capacity of spread spectrum systems with multiuser linear receivers, *IEEE Trans. Inform. Theory* **45**(6): 1968–1983 (Sept. 1999).
9. E. M. Valsbord and V. I. Zhukovski, *Introduction to Multi-Player Differential Games and Their Applications*, Gordon and Breach Science Publishers, 1988.
10. R. B. Meyerson, *Game Theory: Analysis of Conflict*, Harvard Univ. Press, 1991.
11. S. Ulukus, *Power Control, Multiuser Detection and Interference Avoidance in CDMA Systems*, Ph.D. thesis, Rutgers Univ., Dept. Electrical and Computer Engineering, 1998 (thesis director: Prof. R. D. Yates).
12. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2000.
13. S. Haykin, *Communication Systems*, Wiley, New York, 2001.
14. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, Part I, Wiley, New York, 1968.
15. C. Rose, S. Ulukus, and R. Yates, Wireless systems and interference avoidance, *IEEE Trans. Wireless Commun.* **1**(3): (July 2002). preprint available at http:/steph.rutgers.edu/~crose/papers/avoid17.ps.
16. D. C. Popescu, *Interference Avoidance for Wireless Systems*, Ph.D. thesis, Rutgers Univ., Dept. Electrical and Computer Engineering (2002) (thesis director: Prof. C. Rose).
17. P. B. Rapajic and B. S. Vucetic, Linear adaptive transmitter-receiver structures for asynchronous CDMA systems, *Eur. Trans. Telecommun.* **6**(1): 21–27 (Jan.–Feb. 1995).
18. G. S. Rajappan and M. L. Honig, Signature sequence adaptation for DS-CDMA with multipath, *IEEE J. Select. Areas Commun.* **20**(2): 384–395 (Feb. 2002).
19. M. Rupf and J. L. Massey, Optimum sequence multisets for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **40**(4): 1226–1266 (July 1994).
20. P. Viswanath and V. Anantharam, Optimal sequences and sum capacity of synchronous CDMA systems, *IEEE Trans. Inform. Theory* **45**(6): 1984–1991 (Sept. 1999).
21. D. C. Popescu and C. Rose, Interference avoidance and multiaccess dispersive channels. In *Proc. 35th Annual*

Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 2001.

22. D. C. Popescu and C. Rose, CDMA codeword optimization for uplink dispersive channels through interference avoidance. *IEEE Transactions on Information Theory*. (Submitted 12/2002, revised 10/2001. Preprint available at *http://www.winlab.rutgers.edu/~cripop/papers*).
23. D. C. Popescu and C. Rose, Fading channels and interference avoidance. In *Proc. 39th Allerton Conference on Communication, Control, and Computing*, Monticello, IL, October 2001.
24. G. Strang, *Linear Algebra and Its Applications*, 3rd ed., Harcourt Brace Jovanovich College Publishers, 1988.
25. C. Rose, CDMA codeword optimization: Interference avoidance and convergence via class warfare, *IEEE Trans. Inform. Theory* **47**(4): 2368–2382 (Sept. 2001).
26. W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, Iterative water-Filling for Gaussian Vector multiple access channels, *2001 IEEE Int. Symp. Information Theory, ISIT'01*, Washington, DC, June 2001 (submitted for journal publication).
27. H. J. Landau and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty — III: The dimension of the space of essentially time- and band-limited signals, *Bell Syst. Tech. J.* **40**(1): 43–64 (Jan. 1961).
28. D. C. Popescu and C. Rose, Interference avoidance for multiaccess vector channels, *2002 IEEE Int. Symp. Information Theory, ISIT'02*, Lausanne, Switzerland, July 2002 (submitted for publication in *IEEE Trans. Inform. Theory*).
29. J. L. Holsinger, *Digital Communication over Fixed Time-Continuous Channels with Memory — with Special Application to Telephone Channels*, Technical Report 366, MIT — Lincoln Lab., 1964.
30. N. Yee and J. P. Linnartz, *Multi-Carrier CDMA in an Indoor Wireless Radio Channel*, Technical Memorandum UCB/ERL M94/6, Univ. California, Berkeley, 1994.
31. S. N. Diggavi, On achievable performance of spatial diversity fading channels, *IEEE Trans. Inform. Theory* **47**(1): 308–325 (Jan. 2001).
32. D. C. Popescu and C. Rose, Interference avoidance and multiuser MIMO systems, 2002. Preprint available at *http://www.winlab.rutgers.edu/~cripop/papers*).
33. G. G. Raleigh and J. M. Cioffi, Spatio-temporal coding for wireless communication, *IEEE Trans. Commun.* **46**(3): 357–366 (March 1998).
34. D. C. Popescu and C. Rose, Codeword optimization for asynchronous CDMA systems through interference avoidance, *Proc. 36th Conf. Information Sciences and Systems, CISS 2002*, Princeton, NJ, March 2002.
35. S. Ulukus and R. Yates, Optimum signature sequence sets for asynchronous CDMA systems, *Proc. 38th Allerton Conf. Communication, Control, and Computing*, Oct. 2000.
36. O. Popescu and C. Rose, Minimizing total squared correlation for multibase systems, *Proc. 39th Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2001.
37. O. Popescu and C. Rose, Interference avoidance and sum capacity for multibase systems, *Proc. 39th Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2001.
38. D. Tabora, *An Analysis of Covariance Estimation, Codeword Feedback, and Multiple Base Performance of Interference Avoidance*, Master's thesis, Rutgers Univ., Dept. Electrical and Computer Engineering, 2001 (thesis director: Prof. C. Rose).
39. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.

INTERFERENCE MODELING IN WIRELESS COMMUNICATIONS

Xueshi Yang
Athina P. Petropulu
Drexel University
Philadelphia, Pennsylvania

1. INTRODUCTION

In wireless communication networks, signal reception is often corrupted by interference from other sources that share the same propagation medium. Knowledge of the statistics of interference is important in achieving optimum signal detection and estimation. Construction of most receivers is based on the assumption that the interference is i.i.d. (independent, identically distributed) Gaussian. The Gaussianity assumption is based on the central-limit theory. However, in situations such as underwater acoustic noise, urban and synthetic (human-made) RF noise, low-frequency atmospheric noise, and radar clutter noise, the mathematically appealing Gaussian noise model is seldom appropriate [14]. For such cases, several mathematical models have been proposed, including the class A noise model [18], the Gaussian mixture model [28], and the α-stable model [20]. All these noise models share a common feature: the tails of their probability density function decay in a power-law fashion, as opposed to the exponentially decaying tails of the Gaussian model. This implies that the corresponding time series appear bursty, or impulsive, since the probability of attaining very large values can be significant.

Existing models for non-Gaussian impulsive noise are usually divided into two groups: empirical models and statistical–physical models. The former include the hyperbolic distribution and the Gaussian mixture models [28,29]. They fit a mathematical model to the measured data, without taking into account the physical mechanism that generated the data. Although empirical models offer mathematical simplicity in modeling, their parameters are not related to any physical quantities related to the data. On the other hand, statistical–physical models are grounded on the physical noise generation process, and their parameters are linked to physical parameters. The first physical models for noise can be traced back in the works of Furutsu and Ishida [9], Middleton [18], and later in the works of Sousa [26], Nikias [20], Ilow [13], and Yang [33] and colleagues. All of these models consider the following scenario. A receiver is surrounded by interfering sources that are randomly distributed in space according to a Poisson point process. The receiver picks up the superposition of the contributions of all the interferers. By assuming that the propagation path loss is inversely proportional

to the power of the distance between the receiver and the interferer, and that the pulses have a symmetrically distributed random amplitude, it has been shown that the resulting instantaneous noise is impulsive and non-Gaussian [18,20]. The power-law assumption for path loss is consistent with empirical measurement data [10,18].

The same model can be applicable for modeling interference in a spread-spectrum wireless system, where randomly located users access the transmission medium simultaneously. The user identity is determined by a signature, which is embedded in the transmitted signal. Ideally, signatures of different users are orthogonal; thus, by correlating a certain signature with the total received signal, the corresponding user signal can be recovered. However, in practice the received signals from other users are not perfectly orthogonal to the signature waveform of the user of interest, due to non-perfect orthogonal codes and channel distortion. Therefore, correlation at the receiver results in an interference term, usually referred to as *cochannel interference*. Cochannel interference is the determining factor as far as quality of service is concerned, and sets a limit to the capacity of a spread-spectrum communication system. Modeling, analysis, and mitigation of cochannel interference has been the subject of numerous studies.

The goal of this article is to provide a mathematical treatment of the interference that occurs in wireless communication systems. It is organized as follows. In Section 2, we provide a brief treatment of α-stable distributions and heavy-tail processes. We also outline some techniques for deciding whether a set of data follows an α-stable distribution, and for estimating the model parameters. In Section 3, we discuss two widely studied statistical–physical models for interference: the α-stable model and the class A noise model.

2. PROBABILISTIC BACKGROUND

2.1. α-Stable Distributions

α-Stable distributions are defined in terms of their characteristic function:

$$\Phi(\rho) = \exp\{\imath\mu\rho - \sigma^\alpha|\rho|^\alpha(1 + \imath\eta\,\mathrm{sign}(\rho)\varphi(\rho,\alpha))\} \tag{1a}$$

with

$$\varphi(\rho,\alpha) = \begin{cases} \tan\dfrac{\alpha\pi}{2} & \text{if} \quad \alpha \neq 1 \\ \dfrac{2}{\pi}\ln|\rho| & \text{if} \quad \alpha = 1 \end{cases} \tag{1b}$$

$$\mathrm{sign}(\rho) = \begin{cases} 1, & \text{if} \quad \rho > 0 \\ 0, & \text{if} \quad \rho = 0 \\ -1, & \text{if} \quad \rho < 0 \end{cases} \tag{1c}$$

where $\alpha \in (0, 2]$: characteristic exponent — a measure of rate of decay of the distribution tails; the smaller the α the heavier the tails of the distribution

$\eta \in [-1, 1]$: symmetry index

$\sigma > 0$: scale parameter; also referred to as *dispersion*. $2\sigma^2$ equals to the variance in the Gaussian distribution case

μ : location parameter

The distribution is called symmetric α-stable ($S\alpha S$) if $\eta = 0$.

Since (1) is characterized by four parameters, we denote stable distributions by $S_\alpha(\sigma, \eta, \mu)$, and indicate that the random variable X is α-stable distributed by

$$X \sim S_\alpha(\sigma, \eta, \mu) \tag{2}$$

The probability density functions of α-stable random variables are seldom given in closed form. Some exceptions are the following:

- The Gaussian distribution, which can be expressed as $S_2(\sigma, 0, \mu)$, and whose density is

$$f(x) = \frac{1}{2\sigma\sqrt{\pi}}e^{-\frac{(x-\mu)^2}{4\sigma^2}} = N(\mu, 2\sigma^2) \tag{3}$$

- The Cauchy distribution, $S_1(\sigma, 0, \mu)$, whose density is

$$f(x) = \frac{\sigma}{\pi((x-\mu)^2 + \sigma^2)} \tag{4}$$

Two important properties of stable distributions are the stability property and the generalized central-limit theorem.

The *stability property* is defined as follows. A random variable X has a stable distribution if and only if for arbitrary constants a_1 and a_2, there exist constants a and b such that $a_1X_1 + a_2X_2$ has the same distribution as $aX + b$, where X_1 and X_2 are independent copies of X. A consequence of the stability property is the generalized central-limit theorem, according to which, the stable distribution is the only possible limit distribution of sums of i.i.d. random variables. In particular, if $X_1, X_2, \ldots$ are i.i.d. random variables, and there exists sequences of positive numbers $\{a_n\}$ and real numbers $\{b_n\}$, such that

$$S_n = \frac{X_1 + \cdots + X_n}{a_n} - b_n \tag{5}$$

converges to X in distribution, then X is stable distributed.

If the X_i are i.i.d. and have finite variance, then the limit distribution is Gaussian and the generalized central-limit theorem reduces to the ordinary central-limit theorem.

α-Stable distributions are a special class of the so called heavy-tail distributions. A random variable X is said be heavy-tail distributed if

$$\Pr(|X| \geq x) \sim L(x)x^{-\alpha}, \quad 0 < \alpha < 2 \tag{6}$$

where $L(x)$ is a slowly varying function at infinity, that is, $\lim_{x\to\infty} L(cx)/L(x) = 1$ for all $c > 0$. Thus, α-stable distributions with $\alpha < 2$ have power-law tails, as opposed to the exponential tails of the Gaussian distribution. Figure 1 illustrates the survival function $P(X > x)$ of α-stable distributions for different α ($\alpha = 0.5, 1, 1.5, 2$). We

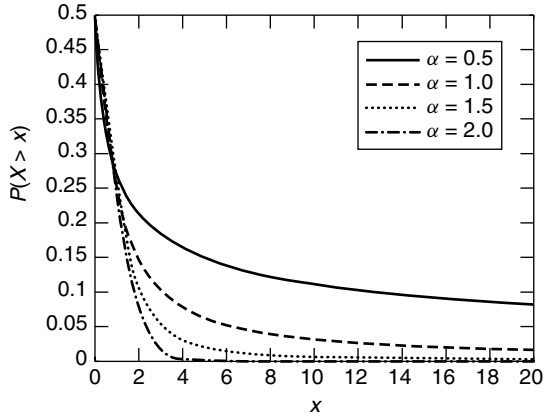


Figure 1. Survival function of α-stable distributions for different $\alpha = 0.5, 1.0, 1.5, 2.0$.

observe that the tail of the distribution decays slower as α becomes smaller.

A set of n real random variables $X_1, X_2, \ldots, X_n$ are jointly symmetric α-stable ($S\alpha S$) if and only if the linear combination $a_1X_1 + \cdots + a_nX_n$ is $S\alpha S$ for all real $a_1, \ldots, a_n$.

A random process $x(t)$ is said to be $S\alpha S$ if for any $n \geq 1$ and distinct instants $t_1, \ldots, t_n$ the random variables $x(t_1), \ldots, x(t_n)$ are jointly $S\alpha S$ with the same α.

α-Stable distributions are known for their lack of moments of order larger than α, in particular, for $\alpha < 2$ second-order statistics do not exist. In such a case, the role of covariance is played by the covariation or the codifference [24]; when $\alpha = 2$ both functions reduce to the covariance. While the covariation may not be defined for $\alpha < 1$, the codifference is defined for all $0 < \alpha \leq 2$.

The *codifference* of two jointly $S\alpha S$, $0 < \alpha \leq 2$, random variables X_1 and X_2 equals

$$R_{X_1,X_2} = \sigma_{X_1}^{\alpha} + \sigma_{X_2}^{\alpha} - \sigma_{X_1-X_2}^{\alpha} \tag{7}$$

where σ_X is the scale parameter of the $S\alpha S$ variable X. If X_1 and X_2 are independent, then their codifference equals zero. However, a zero codifference does not in general imply that the random variables are independent, except in the case $0 < \alpha < 1$.

Let $x(t)$ be a stationary $S\alpha S$ random process. The codifference, $R_{x(t+\tau),x(t)}$, provides a measure of structure of the process. For processes that are marginally only α-stable, or in general, for heavy-tail processes, the codifference is not defined. An alternative measure of structure for such processes is the so called generalized codifference [22,24]:

$$\begin{aligned} I(\rho_1, \rho_2; \tau) = &- \ln E\{e^{\imath(\rho_1 x(t+\tau) + \rho_2 x(t))}\} \\ &+ \ln E\{e^{\imath\rho_1 x(t+\tau)}\} + \ln E\{e^{\imath\rho_2 x(t)}\} \end{aligned} \tag{8}$$

The generalized codifference is closely related to the codifference; if the process is $S\alpha S$, then

$$R_{x(t+\tau),x(t)} = -I(1, -1; \tau) \tag{9}$$

2.2. Estimation of Parameters of α-Stable Distributions

As was already mentioned, α-stable random variables with $\alpha < 2$ do not have finite variance. However, when dealing with a finite-length data record, the estimated variance will always be finite. A natural question is how one assesses the validity of the α-stable model based on a finite set of samples of a random process, and if the model is valid, how one determines the characteristic exponent α.

A concept that appears in all the methods that will be described next is that of *order statistics*. The order statistics of a random sequence $X_1, \ldots, X_N$ are the samples placed in ascending order, usually denoted by $X_{(1)}, \ldots, X_{(N)}$. Thus, $X_{(1)} = \min\{X_1, \ldots, X_N\}$, and $X_{(N)} = \max\{X_1, \ldots, X_N\}$.

A popular method for determining whether a given data set follows a specific distribution is the quantile–quantile (QQ) plot (see Ref. 23 for a detailed coverage of the topic). The principle of the QQ plot is based on the following observation. Assume that the random sequence $U_1, U_2, \ldots, U_N$ are uniformly distributed on [0,1]. By symmetry, the following will hold for the order statistics of U_i:

$$E(U_{(i+1)} - U_{(i)}) = \frac{1}{N+1}$$

and hence

$$EU_{(i)} = \frac{i}{N+1}.$$

Since $U_{(i)}$ should be close to its mean $i/(N+1)$, the plot of $\{(i/(N+1), U_{(i)}), 1 \leq i \leq N\}$ should be roughly linear. Now if we suspect that $X_1, X_2, \ldots, X_N$ come from a distribution G, we can plot $i/N + 1$ against $G(X_{(i)})$ for $1 \leq i \leq N$. $X_{(i)}$ is the order statistics of X_i. If our suspicion is correct, the plot should be approximately linear, and should be the plot of $\{G^{-1}(i/N+1), X_{(i)}, 1 \leq i \leq N\}$. Here G^{-1} is the inverse function of the distribution function G. $G^{-1}(i/N+1)$ is the theoretical quantile, and $X_{(i)}$ is the empirical quantile, and hence the name QQ-plot.

In practice, one usually replaces the theoretical quantile, $G^{-1}(\cdot)$, with the empirical quantile of a sequence of data that are known (or simulated) to be distributed according to G. For example, to check whether a given data set follows the distribution $S_\alpha(\sigma, \eta, \mu)$, we can plot the QQ plot of the data against an ideally $S_\alpha(\sigma, \eta, \mu)$ distributed sample data set. We conclude that the data set follow the distribution of $S_\alpha(\sigma, \eta, \mu)$, if the QQ plot is close to a 45° line.

Additional tests are the probability plot [6], and the chi-square goodness-of-fit test [25]. The book by D'Agostino and Stephens [1] is an excellent reference on these tests.

Estimation of the tail index α can be obtained by several methods [20], two of which, namely the Hill plot and the QQ estimator, are briefly discussed next.

2.2.1. Hill Plot. Let $k < N$. The Hill estimator is defined as [12]:

$$\alpha_k^{-1} = k^{-1} \sum_{j=1}^{k} \log \frac{X_{(N-k+j)}}{X_{(N-k)}}$$

Note that k is the number of upper order statistics used in the estimation.

The choice of k is a difficult issue. A useful graphical tool to choose α is to use a so called *dynamic Hill plot*, which

consists in plotting α_k as a function of k. The sequence α_k^{-1} is a consistent sequence of estimators of α^{-1}, thus the Hill plot should stabilize after some k to a value roughly equal to α. In practice, finding that stable region of the Hill plot is not a straightforward task, since the estimator exhibits extreme volatility.

2.2.2. The QQ Estimator. The slope of regression $1 - \log(1 - j/(k+1))$ through the points $\log(X_{(N-k+j)}/X_{(N-k)})$ can be used as an estimator of the tail index. This estimator is referred to as the QQ estimator [15]. Computing the slope we find that the QQ estimator is given by

$$\hat{\alpha}_k^{-1} = \frac{\displaystyle\sum_{i=1}^{k} -\log\left(\frac{i}{k+1}\right) \times \left(n\log(X_{(N-i+1)}) - \sum_{j=1}^{k}\log(X_{(N-j+1)})\right)}{\displaystyle k\sum_{i=1}^{k}\left(-\log\left(\frac{i}{k+1}\right)\right)^2 - \left(\sum_{i=1}^{k} -\log\left(\frac{i}{N+1}\right)\right)^2} \tag{10}$$

The QQ estimator is consistent for i.i.d. data if $k^{-1} + k/N \to 0$. Its variance is larger than the asymptotic variance of the Hill estimator; however, the variability of the QQ estimator always seems to be less than that of the Hill estimator.

Again, the choice of k is a difficult problem. Two different plots can be constructed on the basis of the QQ estimator: (1) the dynamic QQ plot obtained from plotting $(k, \hat{\alpha}_k), l < k < N$ (similar to the Hill plot) and (2) the static QQ estimator, which is obtained by representing the data $\log(X_{(N-k+1)}/X_{(N-k)})$ as a function of $\log(1 - j/(k+1))$ together with the least-squares regression line. The slope of that line is used to compute the QQ estimator $\hat{\alpha}_{k,n}^{-1}$.

3. MODELING INTERFERENCE

3.1. A α-Stable Noise Model

Consider a packet radio network, where the users are distributed on a plane. The basic unit of time is the *slot*, which is equal to the packet transmission time T. Let us assume that all users are synchronized at the slot level.

In a typical situation, a user transmits a packet to some destination at a distance. The success of the reception depends partly on the amount of interference experienced by the receiving user. The interference at a certain network location consists of two terms: *self-interference*, or interference caused by other transmitting users; and *external interference*, such as thermal noise, coming from other systems. We are here concerned with the former type of interference.

Self-interference is shaped by the positions of the network users and the transmission characteristics of each user. Traditionally, the locations of interfering users have been assumed to be distributed on a plane according to a Poisson point process. Although this is a simplistic assumption, especially in the case of mobile users where their locations can change in a time-varying fashion, it facilitates analysis. The packets are assumed to arrive at the receiver according to a Poisson process, if continuous time is considered, or according to a Bernoulli process if discrete time is considered.

During each symbol interval T at the receiver there are a random number of transmitting users, which have emitted pulses that interfere with signal reception during the particular interval. The users are assumed to be Poisson-distributed in space with density λ:

$$P[k \text{ transmitting users in a region } \mathbf{R}] = \frac{e^{-\lambda R}(\lambda R)^k}{k!} \tag{11}$$

The signal transmitted from the ith interfering user, that is, $p_i(t)$ propagates through the transmission medium and the receiver filters, and as a result become attenuated and distorted. For simplicity, we will assume that distortion and attenuation can be separated. Let us first consider only the filtering effect. For short time intervals, the propagation channel and the receiver can be represented by a time-invariant filter with impulse response $h(t)$. As a result of filtering only, the contribution of the ith interfering source at the receiver is of the form $x_i(t) = p_i(t) * h(t)$, where the asterisk denotes convolution.

The attenuation of the propagation is determined by the transmission medium and environment. In wireless communications, the power loss increases logarithmically with the distance between the transmitter and the receiver [18]. If the distance between the transmitter and the receiver is r, the power loss function may be expressed in terms of signal amplitude loss function, $a(r)$:

$$a(r) = \frac{1}{r^{\gamma/2}} \tag{12}$$

where γ is the path loss exponent and γ is a function of the antenna height and the signal propagation environment. This may vary from slightly less than 2, for hallways within buildings, to >5, in dense urban environments and hard-partitioned office buildings [21].

Taking into account attenuation and filtering, the contribution of the ith interfering source at the receiver equals $a(r_i)x_i(t)$, where r_i, is the distance between the interferer and the receiver.

Assuming that the interferers within the region of consideration have the same isotropic radiation pattern, and that the receiver has an omnidirectional antenna, the total signal at the receiver is

$$x(t) = s(t) + \sum_{i \in \mathcal{N}} a(r_i)x_i(t) \tag{13}$$

where $s(t)$ is the signal of interest and $\mathcal{N}$ denotes the set of interferers at time t. The number of interfering users, as already discussed, is a Poisson-distributed random variable with parameter λ.

The receiver consists of a signal demodulator followed by the detector. A correlation demodulator decomposes the received signal into an N-dimensional vector. The signal is expanded into a series of orthonormal basis functions $\{f_n(s), 0 < s \le T, n = 1, \ldots, N\}$. Let $Z_n(m)$ be the projection of $x(t)$ onto $f_n(\cdot)$ at time slot m:

$$Z_n(m) = \int_0^T x(s + (m-1)T))f_n(s)\,ds \tag{14}$$

To compute the probability of symbol error, one needs to determine first the joint probability density function of the samples $Z_n(m)$. The samples $Z_n(m)$ may be expressed as

$$Z_n(m) = S_n(m) + \sum_{i \in \mathcal{N}} a(r_i) X_{i,n}(m) \tag{15}$$

$$= S_n(m) + Y_n(m) \tag{16}$$

where $X_{i,n}(m)$ and $S_n(m)$ are, respectively, the result of the correlations of $x_i(t)$ and $s(t)$ with the basis functions $f_n(\cdot)$, and $Y_n(m)$ represents interference.

Let us assume that the $X_{i,n}(m)$ are spatially independent [e.g., $X_{i,n}(m)$ independent $X_{j,n}(m)$ for $i \neq j$]. For simplicity, we assume tht $X_{i,k}(m)$, $X_{i,l}(m)$ are identically distributed. Therefore, we shall concentrate on one dimension only. For notational convenience, we shall drop the subscript n, thus denoting $Y_n(m)$ and $X_{i,n}(m)$ by respectively $Y(m)$ and $X_i(m)$. According to the i.i.d. assumption, we will later on denote $X_i(m)$ by $X(m)$.

For some time m, $Y(m)$ is the sum of a random number of i.i.d. random variables, which are contributions of interfering users. To compute the characteristic function of $Y(m)$, we first restrict the sum to contain interferers within a disk, D_b, centered at the receiver and has radius b. Later, we will let the disk radius $b \to \infty$. The characteristic function of interference received from D_b, to be denoted by $Y_b(m)$, is

$$\Phi_{Y_b}(\omega) = E\{e^{j\omega Y_b(m)}\}$$

$$= E\{e^{j\omega \sum_{i=0}^{N_b} a(r_i) X_i(m)}\} \tag{17}$$

where N_b is a random variable representing the number of interferers at times m in D_b. Since the interferers are Poisson-distributed in the space, given that there are k of them in a disk D_b, they are i.i.d. and uniformly distributed. Thus, the distance r_i between the ith interferer and the receiver has density function

$$f(r) = \begin{cases} \dfrac{2r}{b^2} & r \leq b \\ 0 & \text{elsewhere} \end{cases} \tag{18}$$

From (17), using the i.i.d. property of $X_i(m)$ and r_i, and also using the property of conditional expected values

$$E\{X\} = E\{E\{X \mid Y\}\}$$

we obtain

$$\Phi_{Y_b}(\omega) = E\{E\{e^{j\omega \sum_{i=0}^{N} a(r_i) X_i(m))} \mid N\}\}$$

$$= E\{[E\{e^{j\omega a(r) X(m)}\}]^N\} \tag{19}$$

where r_i and $X_i(m)$ are random variables, which, according to the i.i.d. assumption, are generically denoted by r and $X(m)$, respectively. The inner expectation in (19) equals

$$A \triangleq E\{e^{j\omega a(r) X(m)}\}$$

$$= E\{E\{e^{j\omega a(r) X(m)} \mid r\}\}$$

$$= E\{\Phi_X(a(r)\omega)\}$$

$$= \int_0^b f(r) \Phi_X(a(r)\omega)\, dr \tag{20}$$

where $\Phi_X(.)$ is the characteristic function of $X(\cdot)$.

By substituting (20) in (19), we obtain

$$\Phi_{Y_b}(\omega) = E\{A^N\}$$

$$= \sum_{k=0}^{\infty} P(N = k) A^k$$

$$= \sum_{k=0}^{\infty} \frac{(\lambda \pi b^2)^k e^{-\lambda \pi b^2}}{k!} A^k$$

$$= e^{\lambda \pi b^2 (A-1)}$$

By considering the logarithm of the characteristic function, and also using (20), we obtain

$$\Psi_{Y_b}(\omega) \triangleq \log \Phi_{Y_b}(\omega)$$

$$= \lambda \pi b^2 \left(\int_0^b f(r) \Phi_X(a(r)\omega)\, dr - 1 \right) \tag{21}$$

Taking into account (12), setting

$$\alpha = \frac{4}{\gamma} \tag{22}$$

and after a variable substitution, the equation above becomes

$$\Psi_{Y_b}(\omega) = \lambda \pi b^2 \left(\frac{\alpha \omega^\alpha}{b^2} \int_{\omega b^{-2/\alpha}}^{\infty} t^{-1-\alpha} \Phi_X(t)\, dt - 1 \right) \tag{23}$$

To obtain the first-order characteristic function of $Y(n)$, we need to evaluate this expression for $b \to \infty$. Using integration by parts, and noting that $\Phi_X(\omega b^{-2/\alpha}) - 1 \to 0$ as $b \to \infty$, we obtain

$$\Psi_Y(\omega) = \lim_{b \to \infty} \Psi_{Y_b}(\omega) = -\sigma |\omega|^\alpha, \quad \text{for } 0 < \alpha < 2 \tag{24}$$

where

$$\sigma = -\pi \lambda \int_0^{\infty} \frac{\partial \Phi_X(x)}{\partial x} x^{-\alpha}\, dx \tag{25}$$

Equation (25) corresponds to the log-characteristic function of a $S\alpha S$ distribution with exponent α; thus, it implies that the interference $Y(n)$, for a fixed n, is $S\alpha S$, or equivalently, marginally $S\alpha S$.

The basic propagation characteristic that led to the nice closed form result shown above is the power-law attenuation. However, the attenuation expression in (12) is valid for large values of r. As $r \to 0$, the signal amplitude appears to approach infinity. To avoid this problem, an alternative loss function was proposed [11]

$$a'(r) = \min(s, r^{-\gamma/2}) \tag{26}$$

for some $s > 0$. Under this form of attenuation, the log-characteristic function is as in (24), except that now $\sigma = \sigma(\omega)$. The difference between the two log-characteristic functions approaches zero as $\omega \to \infty$ while s is fixed. For small ω the difference tends to zeros as $s \to 0$.

The meaning of (24) is illustrated next via an example. Consider a direct-sequence spread-spectrum

(DSSS) system with large processing gain and chip synchronization. Assume that users' contributions, $X_i(m)$ values, are Gaussian, and that the attenuation law is $1/r^4$. Equation (24) suggests that the resulting interference at the receiver will be Cauchy-distributed.

The discussion above applies to both narrowband interference and wideband interference. In most communication systems, the receiver is narrowband. In those cases, we need to derive the statistics of the envelope and phase of the impulsive interference. The joint characteristic function of the in-phase and quadrature components of the interference can be found to be [20]

$$\log \Phi(\omega_1, \omega_2) = -c(\alpha)(\omega_1^2 + \omega_2^2)^{\alpha/2} \tag{27}$$

where $c(\alpha)$ is a constant that depends on α. This form of joint characteristics function is referred to as *isotropic α-stable*.

Let Y_c and Y_s denote respectively the in-phase and quadrature components of the interference. The envelope A and phase Ψ are then

$$A = \sqrt{Y_c^2 + Y_s^2}, \quad \Psi = \arctan \frac{Y_s}{Y_c} \tag{28}$$

It can be shown that the phase is uniformly distributed in $[0, 2\pi]$, and is independent of the envelope. The probability density function of the envelope cannot be obtained in closed form. However, it can be shown that

$$\lim_{x\to\infty} x^\alpha P(A > x) = \beta(\alpha, \gamma) \tag{29}$$

where β is some function that does not vary with x. According to this equation, the envelope is heavy-tailed.

To implement optimum signal detection and estimation, one needs to obtain joint statistics of the interference. In the literature, for simplicity reasons, interference is traditionally assumed to be i.i.d. However, as we will see next, such assumptions are oversimplified and it is inconsistent with practical communication systems. Temporary dependence arises when the interference sources are the cochannel users, whose activity is decided by the information type being exchanged and the user behavior.

Let us define the term *emerging interferers at symbol interval l*, which describes the interfering sources whose contribution arrive for the first time at the receiver in the beginning of the symbol interval l. It is assumed that the *emerging* interferers at some symbol interval are located according to a Poisson point process in space. In particular, at any given symbol interval, the expected number of emerging interferers in a unit area/volume is given by λ_e. Once a user initiates a transmission, it may last for a random duration of time, referred to as *session life L*. The distribution of L is assumed to be known a priori. It is not difficult to show that at any symbol interval m, the active transmitting users are Poisson-distributed in the space with density $\lambda = \lambda_e E\{L\}$.

Consider an interference source, namely, a cochannel user, whose transmission starts at some random time slot m. According to the previous analysis, it contributes to the interference observed at time m at the targeted receiver, $a(r_i)x_i(m)$ [see Eq. (15)]. Since this particular cochannel user continues to be active for some random time duration after its initiation of the session, it also contributes to the interference at time $m+l$ with probability $\overline{F}_L(l)$. Here $\overline{F}_L(\cdot)$ is the survival function of session life L. Therefore, the interference at time m and $m+l$ are correlated, in contrast to conventional i.i.d. assumption for the interference at different times.

The joint statistics of the interference can be analyzed through the joint characteristic function (JCF) of the interference at time m and n ($m < n$):

$$\begin{aligned}\Phi_{m,n}(\omega_1, \omega_2) &\triangleq E \exp\{j\omega_1 Y(m) + j\omega_2 Y(n)\} \\ &= E \exp\left\{ j\omega_1 \sum_{i\in\mathcal{N}_n} a(r_i) X_i(m) \right. \\ &\qquad \left. + j\omega_2 \sum_{i\in\mathcal{N}_n} a(r_i) X_i(m) \right\}\end{aligned} \tag{30}$$

where $\mathcal{N}_m$ and $\mathcal{N}_n$ denote the set of active interferers at m and n, respectively. It can be shown that although the discretized interference is marginally α-stable, in general, it is not jointly α-stable distributed [33].

We next consider a special case when the symbols $X_i(m)$ are symmetric binary, such as 1 or -1 with equal probability and independent from slot to slot. Then, it can be shown that [33]

$$\begin{aligned}\ln \Phi_{m,n}(\omega_1, \omega_2) &= -\sigma^\alpha \sum_{l=1}^{n-m} \overline{F}_L(l)(|\omega_1|^\alpha + |\omega_2|^\alpha) \\ &\quad - \frac{1}{2}\sigma^\alpha \sum_{l=n-m+1}^{\infty} \overline{F}_L(l)(|\omega_1 + \omega_2|^\alpha + |\omega_1 - \omega_2|^\alpha)\end{aligned} \tag{31}$$

with α as defined in (22), and

$$\sigma = \frac{1}{2}\left(\frac{\lambda\pi^{3/2}\Gamma(1-\alpha/2)}{\Gamma(1/2+\alpha/2)}\right)^{1/\alpha} \tag{32}$$

Equation (31) implies that the interference at m and n are jointly α-stable distributed [24, p. 69].

In high-speed data networks, where large variations of file sizes are exchanged, the holding time of a single transmission session exhibits high variability. It has been shown [30,34] that the holding times can be well modeled by heavy-tail-distributed random variables. As data service becomes increasingly popular in wireless communication networks, where more and more bandwidth is available for users, it should be expected that session life of users will behave similarly as in wired data networks, that is, will be heavy-tail-distributed. Indeed, preliminary verification of the latter has been presented in Ref. 16 through statistical analysis of wireless network traffic data.

By modeling the session life of the interferers as a heavy-tail-distributed random variable with tail index $1 < \alpha_L < 2$, and $X_i(m)$ as i.i.d. Bernoulli random variables taking possible values 1 and -1 with equal probability $\frac{1}{2}$,

it can be shown that the resulting interference is strongly correlated. In particular, assuming a Zipf distribution for the session life,[1] the following holds [33]:

$$\lim_{\tau\to\infty} \frac{-I(1,-1;\tau)}{\tau^{-(\alpha_L-1)}} = \frac{(2-2^{\alpha-1})\sigma^{\alpha}}{\alpha_L - 1} \tag{33}$$

where $I(\cdot)$ = codifference of the resulted interference
τ = time lag between symbol intervals
α_L = tail index of the session life distribution
σ = as defined in (32)

The form of (33) implies an important phenomenon, referred to as long-range dependence (LRD) [3,22]. LRD refers to strong dependence between samples of stochastic processes. For processes with finite autocorrelation, this implies that the correlation of well separated samples is not negligible, and decays in a power-law fashion with the distance between the samples. As a result, the autocorrelation of a long-range dependent process is not summable. When the noise exhibits LRD the performance of signal detection and estimation algorithms, which are optimized for i.i.d. noise, can degrade [2,32]. For processes with infinite autocorrelation, such as marginally α-stable processes, LRD is defined in terms of a power-law decaying generalized codifference [22], which is the case in (33).

We next present some simulation results based on the model described above. A wireless network is simulated, where a receiver is subjected to interference from users that are spatially Poisson distributed over a plane with density $\lambda = 2/\pi$. The path loss is power-law with $\gamma = 4$. Once the interferers start to emit pulses, they remain active for a random time duration, which in our simulations was taken to be Zipf-distributed with $k_0 = \sigma = 1$ and $\alpha_L = 1.8$. The $X_i(m)$ values were taken to be i.i.d. Bernoulli-distributed, taking values ± 1 with equal probabilities. Then, according to (32), $\sigma = \pi$, and the instantaneous interference is Cauchy distributed with scale parameter $\sigma\mu^{1/\alpha}$, where μ is the mean of the session life. Note that $\mu = \zeta(1.8) \simeq 1.88$.

One segment of the simulated interference process is shown in Fig. 2, where it can be seen that the simulated process is highly impulsive.

We use the static Hill estimator presented in Section 2.B to estimate the tail index. Figure 3 clearly shows that the interference is heavy-tail-distributed with tail index very close to the theoretical value 1. The QQ plot of the interference against ideally Cauchy-distributed random variable with scale parameter 1.88π is also illustrated in Fig. 4, which shows that the instantaneous interference can be well modeled by Cauchy distributed random variables.

[1] A random variable X has a Zipf distribution if

$$P\{X \geq k\} = \left[1 + \left(\frac{k-k_0}{\sigma}\right)\right]^{-\alpha}, \quad k = k_0, k_0+1, k_0+2\cdots$$

where k_0 is an integer denoting the location parameter, $\sigma > 0$ denotes the scale parameter and $\alpha > 0$ is the tail index.

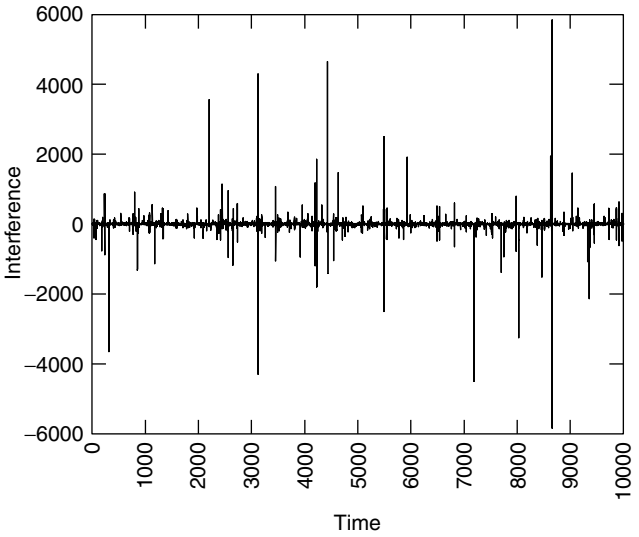


Figure 2. Interference at a receiver surrounded by Poisson distributed interferers.

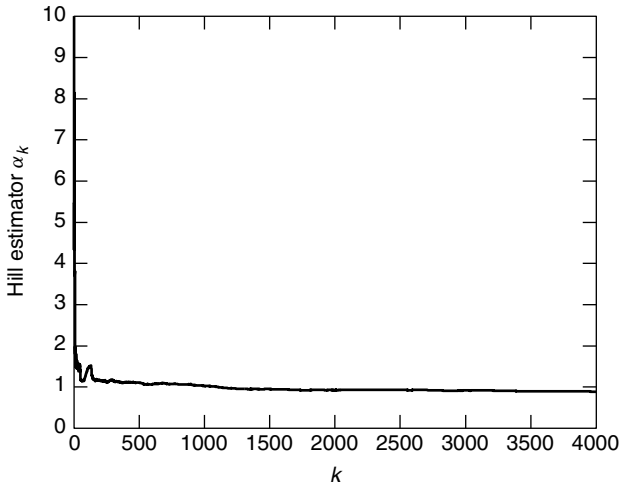


Figure 3. Dynamic Hill estimator α_k.

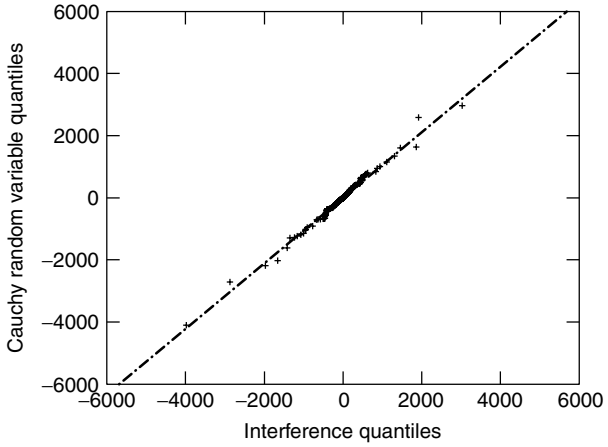


Figure 4. QQ plot of simulated interference and Cauchy random variables.

3.2. The Class A Noise Model

Another model, which has finite variance, but has been used to approximate impulsive interference, is the Middleton class A noise model [18].

The class A model assumes that a narrowband receiver is exposed to an infinite number of potential interfering sources, which emit basic waveforms with common form, scale, durations, frequencies, and so on. The parameters are randomly distributed. As assumed in the α-stable noise model, the locations of the sources are Poisson-distributed in space. The sources are statistically independent, and their emission times are Poisson-distributed in time. The propagation model employed in the class A model is power-law as defined in (12). Moreover, to take into account system noise and external interference, resultant of many independent sources none of which is exceptionally dominant with respect to the others, the class A noise includes a Gaussian component, denoted by its variance σ_G^2.

Under these assumptions, Middleton [18] derived the exceeding probability function of the envelop of the resulted noise:

$$\Pr(\varepsilon > \varepsilon_0) \cong e^{-A} \sum_{m=0}^{\infty} \frac{A^m}{m!} e^{-\varepsilon_0^2/2\sigma_m^2}, \quad 0 \le \varepsilon_0 < \infty \tag{34}$$

with $2\sigma_m^2 = (m/A + \Gamma)/(1 + \Gamma)$. Here ε, ε_0 are normalized envelopes

$$\varepsilon \equiv \frac{E}{\sqrt{2\Omega(1+\Gamma)}} \tag{35}$$

$$\varepsilon_0 \equiv \frac{E_0}{\sqrt{2\Omega(1+\Gamma)}} \tag{36}$$

where E_0 is some preselected threshold value of the envelope E. There are three parameters involved in the class A model, namely, (A, Γ, Ω). They all have physical significance as stated next.

1. $A =$ the *impulsive index*, which is defined as the average number of emissions multiplied by the mean duration of a typical interfering source emission. When A is large, the central-limit theory comes into play, and one approaches Gaussian statistics (for envelopes, it is Rayleigh).
2. $\Gamma \equiv \sigma_G^2/\Omega =$ the ratio of the independent Gaussian component σ_G^2 to the intensity of the non-Gaussian, impulsive component.
3. $\Omega =$ the intensity of the impulsive component.

The phase statistics of the resulting noise is uniformly distributed in $(0, 2\pi)$.

The class A noise model is a canonical model, in the sense that it is invariant of the particular noise source and associated quantifying parameters. The parameters of the class A noise model can be deduced from physical measurement, because its derivation is based on a general physical mechanism. The class A noise model fits a variety of measurements, and has been applied in various communication scenaria, where non-Gaussian noise dominates the interfering sources. However, it is often difficult to evaluate (34). A simplification has been proposed by Spaulding [27], where the infinite sum in (34) has been replaced by a finite sum. Note that the density function of class A noise can be written as

$$f(\varepsilon) = \sum_{m=0}^{\infty} a_m g(\varepsilon; 0, \sigma_k^2) \tag{37}$$

where

$$a_m = \frac{e^A A^m}{m!} \quad \text{and} \quad g(\varepsilon; \mu, \sigma_m^2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-(\varepsilon-\mu)^2/2\sigma_m^2} \tag{38}$$

Essentially, (37) indicates that class A noise is a Poisson weighted sum of zero mean Gaussian functions with monotonic increasing variance σ_m^2. Since A is small, the weights in (37) for large m can be truncated without much loss in the accuracy. Spaulding [27] suggested that (37) can be well approximated (2% error) by the first two terms when A and Γ are small. The truncated class A noise now reduces to a subset of the Gauss-Gauss mixture model [14]:

$$f_{\text{mix}}(x) = ag(x; 0, \sigma_0^2) + (1-a)g(x; 0, \sigma_1^2) \tag{39}$$

where $0 < a < 1$ and $\sigma_0^2 < \sigma_1^2$.

Efforts have been devoted to developing a multivariate class A noise model. The multivariate case arises in the scenario of communication systems with multiple antennas or antenna arrays reception, or in times where multiple time observations are available. In Ref. 7, multivariate class A models are developed based on mathematical extension of the univariate class A model. Three different extensions — independent, dependent, and uncorrelated cases — are discussed in Ref. 7. Issues of signal detection under these models are also investigated. A more physical-mechanism oriented approach is presented in Ref. 17. Under assumptions similar to those expressed in Ref. 18, the authors consider the noise presented in the intermediate-frequency (IF) stage in a communication system with two antennas. By assuming the emission times of the sources are uniformly distributed over a long time interval, Ref. 17 obtains the characteristic function of the resulting noise for cases where the antenna observations may be statistically dependent from antenna to antenna. The probability density function thus can be deduced by utilizing Fourier transform techniques.

4. CONCLUSION

In this chapter, we present two mathematical models for interference in wireless communications: the α-stable model and the class A noise model. Both models are based on the physical mechanism of the noise generation process, and lead to a non-Gaussian scenario, where noise may become impulsive, or heavy-tail-distributed. The noise process may not be i.i.d. Moreover, under heavy-tailed holding times, the noise becomes highly correlated.

Impulsiveness of noise can have severe degrading effects on system performance, particularly on most conventional systems designed for optimal or near-optimal

performance against white Gaussian noise. A significant amount of research has been devoted to signal detection and estimation when the noise is non-Gaussian and/or correlated [e.g., 14,19,20].

BIOGRAPHIES

Xueshi Yang received his B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1998, and a Ph.D. degree in electrical engineering from Drexel University, Philadelphia, Pennsylvania, in 2001. From September 1999 to April 2000 he was a visiting researcher with Laboratoire de Signaux et Systemes, CNRS-Universite de Paris-Sud, SUPELEC, Paris, France. He is currently a postdoc research associate in Electrical Engineering at Princeton University, New Jersey, and Drexel University. His research interests are in the area of non-Gaussian signal processing, self-similar processes, and wireless/wireline data networking. Dr. Yang received the George Hill Jr. Fellowship in 2001 and the Allen Rothwarf Outstanding Graduate Student Award in 2000, respectively, from Drexel University.

Athina P. Petropulu received the Diploma in electrical engineering from the National Technical University of Athens, Greece, in 1986, the M.Sc. degree in electrical and computer engineering in 1988, and her Ph.D. degree in electrical and computer Engineering in 1991, both from Northeastern University, Boston, Massachusetts.

In 1992, she joined the Department of Electrical and Computer Engineering at Drexel University, Philadelphia, Pennsylvania, where she is now a professor. During the academic year 1999–2000 she was an associate professor at LSS, CNRS-Université Paris Sud, École Supérieure d'Electrícité in France. Dr. Petropulu's research interests span the area of statistical signal processing, communications, higher-order statistics, fractional-order statistics and ultrasound imaging. She is the coauthor of the textbook entitled, *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*, (Englewood Cliffs, NJ: Prentice-Hall, Inc., 1993). She is the recipient of the 1995 Presidential Faculty Fellow Award.

BIBLIOGRAPHY

1. R. B. D'Agostino and M. A. Stephens, eds., *Goodness-of-fit Techniques*, Marcel Dekker, New York, 1986.
2. R. J. Barton and H. V. Poor, Signal detection in fractional Gaussian noise, *IEEE Trans. Inform. Theory* **34**: 943–959 (Sept. 1988).
3. J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, New York, 1994.
4. K. L. Blackard, T. S. Rappaport, and C. W. Bostian, Measurements and models of radio frequency impulsive noise for indoor wireless communications, *IEEE J. Select. Areas Commun.* **11**: 991–1001 (Sept. 1993).
5. O. Cappe et al., Long-range dependence and heavy-tail modeling for teletraffic data, *IEEE Signal Process. Mag.* (*Special Issue on Analysis and Modeling of High-Speed Network Traffic*) (in press).
6. J. M. Chambers, *Graphical Methods for Data Analysis*, PWS Publishing, 1983.
7. P. A. Delaney, Signal detection in multivariate Class-A interference, *IEEE Trans. Commun.* **43**: 365–373 (Feb.–April 1995).
8. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 3rd ed., Wiley, New York, 1971.
9. K. Furutsu and T. Ishida, On the theory of amplitude distribution of impulsive random noise, *J. Appl. Phys.* **32**(7): (1961).
10. A. Giordano and F. Haber, Modeling of atmosphere noise, *Radio Sci.* **7**(11): (Nov. 1972).
11. J. W. Gluck and E. Geraniotis, Throughput and packet error probability in cellular frequency-hopped spread-spectrum radio networks, *IEEE J. Select. Areas Commun.* **7**: 148–160 (Jan. 1989).
12. B. M. Hill, A simple general approach to inference about the tail of a distribution, *Ann. Stat.* **3**: 1163–1174 (1975).
13. J. Ilow, D. Hatzinakos, and A. N. Venetsanopoulos, Performance of FH SS radio networks with interference modeled as a mixture of Gaussian and Alpha-stable noise, *IEEE Trans. Commun.* **46**(4): (April 1998).
14. S. A. Kassam, *Signal Detection in Non-Gaussian Noise*, Springer-Verlag, New York, 1987.
15. M. F. Kratz and S. I. Resnick, The QQ estimator and heavy tails, *Stoch. Models* **12**(4): 699–724 (1996).
16. T. Kunz et al., WAP traffic: Description and comparison to WWW traffic, *Proc. 3rd ACM Int. Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Boston, Aug. 2000.
17. K. F. McDonald and R. S. Blum, A statistical and physical mechanisms-based interference and noise model for array observations, *IEEE Trans. Signal Process.* **48**(7): (July 2000).
18. D. Middelton, Statistical-physical models of electromagnetic interference, *IEEE Trans. Electromagn. Compat.* **EMC-19**(3): (Aug. 1977).
19. D. Middleton and A. D. Spaulding, Elements of weak signal detection in non-Gaussian noise environments, in H. V. Poor and J. B. Thomas, eds., *Advances in Statistical Signal Processing*, Vol. 2, *Signal Detection*, JAI Press, Greenwich, CT, 1993.
20. C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*, Wiley, New York, 1995.
21. J. D. Parsons, *The Mobile Radio Propagation Channel*, Wiley, New York, 1996.
22. A. P. Petropulu, J.-C. Pesquet, X. Yang, and J. Yin, Power-law shot noise and relationship to long-memory processes, *IEEE Trans. Signal Process.* **48**(7): (July 2000).
23. J. Rice, *Mathematical Statistics and Data Analysis*, Duxbury Press, Belmont, CA, 1995.
24. G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman & Hall, New York, 1994.
25. G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Iowa State Press, 1990.
26. E. S. Sousa, Performance of a spread spectrum packet radio network link in a Poisson field of interferers, *IEEE Trans. Inform. Theory* **38**(6): (Nov. 1992).

27. A. D. Spaulding, Locally optimum and suboptimum detector performance in a non-Gaussina interference environment, *IEEE Trans. Commun.* **COM-33**(6): 509–517 (1985).

28. G. V. Trunk, Non-Rayleigh sea clutter: Properties and detection of targets, in D. C. Shleher ed., *Automatic Detection and Radar Data Processing*, Artech House, Dedham, 1980.

29. E. J. Wegman, S. C. Schwartz, and J. B. Thomas, eds., *Topics in Non-Gaussian Signal Processing*, Springer, New York, 1989.

30. W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, Self-similarity through high-variability: Statistical analysis of Ethernets LAN traffic at the source level, *IEEE/ACM Trans. Network.* **5**(1): (Feb. 1997).

31. B. D. Woerner, J. H. Reed, and T. S. Rappaport, Simulation issues for future wireless modems, *IEEE Commun. Mag.* (July 1994).

32. G. W. Wornell, Wavelet-based representations for the 1/f family of fractal processes, *Proc. IEEE* **81**(10): 1428–1450 (Oct. 1993).

33. X. Yang and A. P. Petropulu, Joint statistics of impulsive noise resulted from a Poisson field of interferers, *IEEE Trans. on Signal Processing* (submitted).

34. X. Yang and A. P. Petropulu, The extended alternating fractal renewal process for modeling traffic in high-speed communication networks, *IEEE Trans. Signal Process.* **49**(7): (July 2001).

INTERFERENCE SUPPRESSION IN SPREAD-SPECTRUM COMMUNICATION SYSTEMS

Moeness G. Amin
Yimin Zhang
Villanova University
Villanova, Pennsylvania

1. INTRODUCTION

Suppression of correlated interference is an important aspect of modern broadband communication platforms. For wireless communications, in addition to the presence of benign interferers, relatively narrowband cellular systems, employing time-division multiple access (TDMA) or the Advanced Mobile Phone System (AMPS) may coexist within the same frequency band of the broadband code-division multiple access (CDMA) systems. Hostile jamming is certainly a significant issue in military communication systems. Global Positioning System (GPS) receivers potentially experience a mixture of both narrowband and wideband interference, both intentionally and unintentionally.

One of the fundamental applications of spread-spectrum (SS) communication systems is its inherent capability of interference suppression. SS systems are implicitly able to provide a certain degree of protection against intentional or unintentional interferers. However, in some cases, the interference might be much stronger than the SS signal, and the limitations on the spectrum bandwidth render the processing gain insufficient to decode the useful signal reliably. For this reason, signal processing techniques are frequently used in conjunction with the SS receiver to augment the processing gain, permitting greater interference protection without an increase in the bandwidth. Although much of the work in this area has been motivated by the applications of SS as an antijamming method in military communications, it is equally applicable in commercial communication systems where SS systems and narrowband communication systems may share the same frequency bands.

This article covers both the direct-sequence spread-spectrum (DSSS) and frequency-hopping (FH) communication systems, but the main focus is on the DSSS communication systems. For DSSS communication systems, two types of interference signals are considered, namely, narrowband interference (NBI) and nonstationary interference, such as instantaneously narrowband interference (INBI).

The early work on narrowband interference rejection techniques in DSSS communications has been reviewed comprehensively by Milstein in [1]. Milstein discusses in depth two classes of rejection schemes: (1) those based on least-mean-square (LMS) estimation techniques and (2) those based on transform domain processing structures. The improvement achieved by these techniques is subject to the constraint that the interference be relatively narrowband with respect to the DSSS signal waveform. Poor and Rusch [2] have given an overview of NBI suppression in SS with the focus on CDMA communications. They categorize CDMA interference suppression by linear techniques, nonlinear estimation techniques, and multiuser detection techniques (multiuser detection is outside the scope of this article). Laster and Reed [3] have provided a comprehensive review of interference rejection techniques in digital wireless communications, with the focus on advances not covered by the previous review articles.

Interference suppression techniques for nonstationary signals, such as INBI, have been summarized by Amin and Akansu [4]. The ideas behind NBI suppression techniques can be extended to account for the nonstationary nature of the interference. For time-domain processing, time-varying notch filters and subspace projection techniques can be used to mitigate interferers characterized by their instantaneous frequencies and instantaneous bandwidths. Interference suppression is achieved using linear and bilinear transforms, where the time–frequency domain and wavelet/Gabor domain are typically considered. Several methods are available to synthesize the nonstationary interference waveform from the time–frequency domain and subtract it from the received signal.

Interference rejection for FH is not as well developed as interference rejection for DS or for CDMA. In FH systems, the fast FH (FFH) is of most interest, and the modulation most commonly used in FH is frequency shift keying (FSK). Two types of interference waveforms can be categorized, namely, partial-band interference (PBI) and multitone interference (MTI). Typically, interference suppression techniques for FH communication systems often employ a whitening or clipping stage to reject interference, and then combined by diversity techniques.

2. SIGNAL MODEL

The received waveform consists of a binary phase shift keying (BPSK) DSSS signal $s(t)$, an interfering signal $u(t)$, and thermal noise $b(t)$. Without loss of generality, we consider the single-interferer case, and additive Gaussian white noise (AGWN) that is uncorrelated with both the DSSS and the interference signals. The input to the receiver, $x(t)$, is given by

$$x(t) = s(t) + u(t) + b(t) \tag{1}$$

The DSSS signal can be expressed as

$$s(t) = \sum_{l=-\infty}^{\infty} d(l)p(t - lT_c) \tag{2}$$

where T_c is the chip duration, $p(t)$ is the shaping waveform

$$d(l) = s(n)c(n, l) \tag{3}$$

is the chip sequence, $s(n) \in [-1, +1]$ is the nth symbol, and $c(n, l) \in [-1, +1]$ is a pseudonoise (PN) sequence used as the spreading code for the nth symbol. The PN sequence can be either periodic or aperiodic. Different types of interference signals are considered.

For discrete-time filter implementations, signals are sampled at the rate $1/T$. Typically, the sampling interval T is equal to the chip duration T_c. The input to the receiver, after sampling, becomes

$$x[n] = x(nT) \tag{4}$$

The samples of the DSSS signal, interference, and noise can be defined accordingly as $s[n]$, $u[n]$, and $b[n]$, respectively.

3. NARROWBAND INTERFERENCE SUPPRESSION

Interference suppression techniques for DSSS systems are numerous. In particular, much literature exists on the adaptive notch filtering as it relates to suppress NBI on a wideband DSSS signal. Synthesis/subtraction is another well-established technique for sinusoidal interference suppression. Other techniques include nonlinear adaptive filtering and multiuser detection techniques.

3.1. Adaptive Notch Filtering

The basic idea in employing an adaptive notch filter is to flatten the filter input spectrum. An SS signal tends to have a uniform wide spectrum and is affected little by the filtering process, whereas the NBI is characterized by spectral spikes and frequency regions of high concentrated power. The adaptive notch filter places notches at the frequency location of the NBI to bring the interference level to the level of the SS signal. At least two main approaches exist for constructing an adaptive notch filter: (1) estimation/subtraction-type filters and (2) transform-domain processing structures.

3.1.1. Estimation/Subtraction-Type Filters. If the interference is relatively narrowband compared with the bandwidth of the spread-spectrum waveform, then the technique of interference cancellation by the use of notch filters often results in a large improvement in system performance. This technique, described in many references [e.g., 5–8] uses a tapped-delay line to implement the prediction error filter (Wiener filter [9]). Since both the DS signal and the thermal noise are wideband processes, their future values cannot be readily predicted from their past values. On the other hand, the interference, which is a narrowband process, can indeed have its current and future values predicted from past values. Hence, the current value, once predicted, can be subtracted from the incoming signal, leaving an interference-free waveform comprised primarily of the DS signal and the thermal noise. A general diagram of this technique is depicted in Fig. 1. Both one-sided and two-sided transversal filters can be used for this purpose. When two-sided filters are used, the estimation of current interference value is based on both past and future values of the interference. Consider a single-sided filter as shown in Fig. 2. Define an N-dimensional vector $\mathbf{x}[n]$, denoted as

$$\mathbf{x}[n] = (x[n-1], \ldots, x[n-N])^T \tag{5}$$

where the superscript T denotes transpose of a vector or a matrix. The DSSS signal, interference, and noise vectors can be defined similarly as $\mathbf{s}[n]$, $\mathbf{u}[n]$, and $\mathbf{b}[n]$, respectively. We also define the corresponding weight vector $\mathbf{w}$ as

$$\mathbf{w} = [w_1, \ldots, w_N]^T \tag{6}$$

Hence, the output sample of the filter is

$$y[n] = x[n] - \mathbf{w}^T\mathbf{x}[n] \tag{7}$$

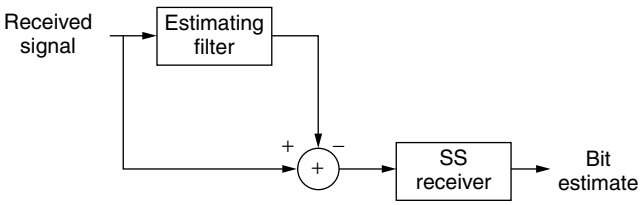


Figure 1. Estimator/subtracter-based interference suppression.

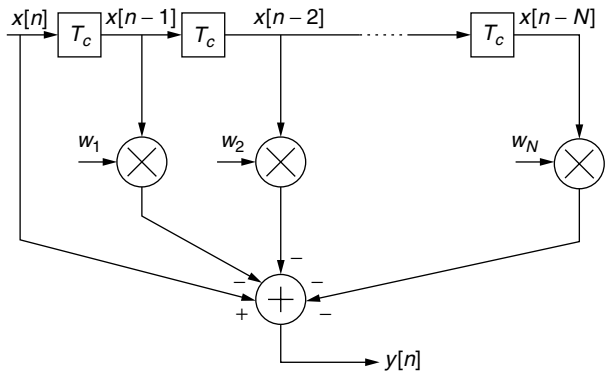


Figure 2. Single-sided transversal filter.

The mean-square value $E[y^2[n]]$, representing the output average power, is given by

$$E(y^2[n]) = E(x^2[n]) - 2\mathbf{w}^T E(x[n]\mathbf{x}[n]) + \mathbf{w}^T E(\mathbf{x}[n]\mathbf{x}^T[n])\mathbf{w} \triangleq E(x^2[n]) - 2\mathbf{w}^T\mathbf{p} + \mathbf{w}^T\mathbf{R}\mathbf{w} \quad (8)$$

where $\mathbf{p} = E(x[n]\mathbf{x}[n])$ is the correlation vector between $x[n]$ and $\mathbf{x}[n]$, and

$$\mathbf{R} = E(\mathbf{x}[n]\mathbf{x}^T[n]) \quad (9)$$

is the covariance matrix of $\mathbf{x}[n]$. It is noted that, when the PN sequence is sufficiently long, the PN signal samples at different taps are approximately uncorrelated. On the other hand, samples of the narrowband interference at different taps has high correlations. Since the DSSS signal, interference, and thermal noise are mutually uncorrelated, it follows that

$$\begin{aligned}\mathbf{p} &= E(x[n]\mathbf{x}[n]) \\ &= E\{(s[n] + u[n] + b[n])(\mathbf{s}[n] + \mathbf{u}[n] + \mathbf{b}[n])\} \\ &= E(u[n]\mathbf{u}[n]).\end{aligned} \quad (10)$$

Minimizing the output power $E[y^2[n]]$ yields the following well-known Wiener–Hopf solution for the optimum weight vector $\mathbf{w}_{\text{opt}}$:

$$\mathbf{w}_{\text{opt}} = \mathbf{R}^{-1}\mathbf{p} \quad (11)$$

The cost of notch filtering is the introduction of some distortion into the SS signal. Such distortion results in power loss of the desired DSSS signal as well as the introduction of self-noise. Both effects become negligible when the processing gain is sufficiently large.

Note that when precise statistical knowledge of the interference cannot be assumed, adaptive filtering can be used to update the tap weights. There are a variety of adaptive algorithms and receiver structures [6,9–11]. The optimum Wiener–Hopf filter can be implemented by using direct matrix inversion (DMI) or recursive adaptation methods. For the DMI method, the covariance matrix $\mathbf{R}$ and $\mathbf{p}$ are estimated at time n by using most recent N_t data samples:

$$\hat{\mathbf{R}}[n] = \frac{1}{N_t}\sum_{l=0}^{N_t-1}\mathbf{x}[n-l]\mathbf{x}^T[n-l] \quad (12)$$

$$\hat{\mathbf{p}}[n] = \frac{1}{N_t}\sum_{l=0}^{N_t-1}x[n-l]\mathbf{x}^T[n-l] \quad (13)$$

The least-mean-square (LMS) algorithm is a simple and stable method to implement an iterative solution to the Wiener–Hopf equation without making use of any a priori statistical information about the received signal. Using the instantaneous estimates of the covariance matrix and cross-correlation vector of Eqs. (9) and (10), the LMS algorithm can be expressed as

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu y[n]\mathbf{x}[n] \quad (14)$$

where $\mathbf{w}[n]$ is the filter weight vector of elements $w_l[n]$, $l = 1, 2, \ldots, N$, $\mathbf{x}[n]$ is the vector that includes the data within the filter, and $y[n]$ is the output, all at the nth adaptation, and μ is a parameter that determines the rate of convergence of the algorithm. It is noted that in most applications of the LMS algorithm, an external reference waveform is needed in order to correctly adjust the tap weights. However, in this particular application, the signal on the reference tap $x[n]$ serves as the external reference.

The drawback of the LMS algorithm is its slow convergence. To improve the convergence performance, techniques including self-orthogonalizing LMS, recursive least-squares (RLS), and lattice structure can be used [9,12].

3.1.2. SINR and BER Analysis. The output signal-to-interference-plus-noise ratio (SINR) and bit error rate (BER) are two important measures for communication quality and the performance enhancement using the signal processing techniques.

To derive the output SINR, we rewrite the filter output as[1]

$$y[n] = \sum_{l=0}^{N} w_l x[n-l] = \sum_{l=0}^{N} w_l(c[n-l] + u[n-l] + b[n-l]) \quad (15)$$

where $w_0 = 1$. The signal $\{y[n]\}$ is then fed to the PN correlator. Denote L as the number of chips per information bit. Then the output of the PN correlator, which is the decision variable for recovering the binary information, is expressed as

$$\begin{aligned} r &= \sum_{n=1}^{L} y[n]c[n] = \sum_{n=1}^{L} c[n] \\ &\quad \times \sum_{l=0}^{N} w_l(c[n-l] + u[n-l] + b[n-l]) \\ &= \sum_{n=1}^{L} c^2[n] + \sum_{n=1}^{L} c[n] \sum_{l=1}^{N} w_l c[n-l] \\ &\quad + \sum_{n=1}^{L} c[n] \times \sum_{l=0}^{N} w_l(u[n-l] + b[n-l]) \\ &= L + \sum_{n=1}^{L} c[n] \sum_{l=1}^{N} w_l c[n-l] + \sum_{n=1}^{L}\sum_{l=0}^{N} c[n]w_l u[n-l] \\ &\quad + \sum_{n=1}^{L}\sum_{l=0}^{N} c[n]w_l b[n-l] \end{aligned} \quad (16)$$

The first term on the right-hand side of (16) represents the desired signal component, the second term amounts to the self-noise caused by the dispersive characteristic of the filter, and the third term is the residual narrowband

[1] To keep the notation simple, we have used in (15) the same symbol w_l as in (6). The two sets of weights, however, differ in sign.

interference escaping the excision process and appearing at the output of the PN correlator. The last term in (16) is the additive noise component.

The mean value of r is

$$E(r) = L \tag{17}$$

and the variance is [6]

$$\text{var}(r) \triangleq \sigma^2 = L\sum_{l=1}^{N} w_l^2 + L\sum_{n=1}^{N}\sum_{l=0}^{N} w_n w_l \rho[n-l] + L\sigma_n^2 \sum_{l=0}^{N} w_l^2 \tag{18}$$

where $\sigma_n^2 = E(b^2[n])$ is the AGWN variance and

$$\rho[n-l] = E(u[k]u[k+n-l])$$

The three terms of the right-hand side of (18) represent the mean square values caused by the self-noise, residual narrowband interference, and noise, respectively.

The output SINR is defined as the ratio of the square of the mean to the variance. Thus

$$\text{SINR}_0 = \frac{E^2(r)}{\text{var}(r)} = \frac{L}{\sum_{l=1}^{N} w_l^2 + \sum_{n=1}^{N}\sum_{l=0}^{N} w_n w_l \rho[n-l] + \sigma_n^2 \sum_{l=0}^{N} w_l^2} \tag{19}$$

Note that if there is no interference suppression filter, $w_l = 1$ for $l = 0$ and zero otherwise. Therefore, the corresponding output SINR is

$$\text{SINR}_{\text{no}} = \frac{L}{\rho[0] + \sigma_n^2} \tag{20}$$

If we assume that the self-noise, residual interference, and noise components at the output of correlator is Gaussian, then the BER can be evaluated in the same manner as the conventional BPSK corrupted only by AWGN. Under such assumption, the BER is given by

$$P_b = P(r < 0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{-(r-L)^2/2\sigma^2}\, dr = Q(\sqrt{\text{SINR}_0}) \tag{21}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-v^2/2}\, dv \tag{22}$$

is the Q function [12].

3.1.3. Transform-Domain Processing Structures. An alternative to time-domain excision as described in the preceding section is to transform the received signal to the frequency domain and perform the excision in that domain. Clipping and gating methods can then be applied on those transform bins contaminated by the interference.

Surface acoustic wave (SAW) device technology can be used to produce the continuous-time Fourier transform of the received waveform [13,14]. The discrete Fourier transform (DFT), with FFT implementations, is commonly applied for time-sampled signals [15]. Adaptive subband transforms generalize transform-domain processing [16,17], and can yield uncorrelated transform coefficients.

The interference-suppressed signal based on a block transform can be written as

$$\mathbf{x}_s[n] = \mathbf{BEAx}[n] \tag{23}$$

where $\mathbf{x}[n]$ is the received input vector; $\mathbf{A}$ and $\mathbf{B}$ are the forward transform matrix and inverse transform matrix, respectively; and $\mathbf{E}$ is a diagonal matrix with each diagonal element acting as a weight multiplied to the input signal at each transform bin. The weights can be controlled by different schemes. Two commonly used methods are either to set the weights binary (i.e., a weight is either one or zero) or to adjust the weights adaptively. In applying the first method, powerful NBI is detected by observing the envelope of the spectral waveform. Substantial interference suppression can be achieved by multiplying the input signal with a weight that is set to zero when the output of the envelope detector at a transform bin exceeds a predetermined level. Figure 3 illustrates the concept of transform-domain notch filtering. Adaptive algorithms, such as LMS and RLS, can be used to determine the excision weights adaptively. The application of these algorithms, however, requires a reference signal that is correlated with the DSSS signal.

When the binary weights are used, the transform-domain processing technique may suffer from the interference sidelobes. With block transforms, the energy in the narrowband interference, which initially occupies a small region of the frequency spectrum, is dispersed in a relatively large spectral region. In this case, excision of a large frequency band may become necessary to effectively remove the interference power from most transform bins. The frequency dispersion of the interference can be nearly eliminated by weighting the received signal

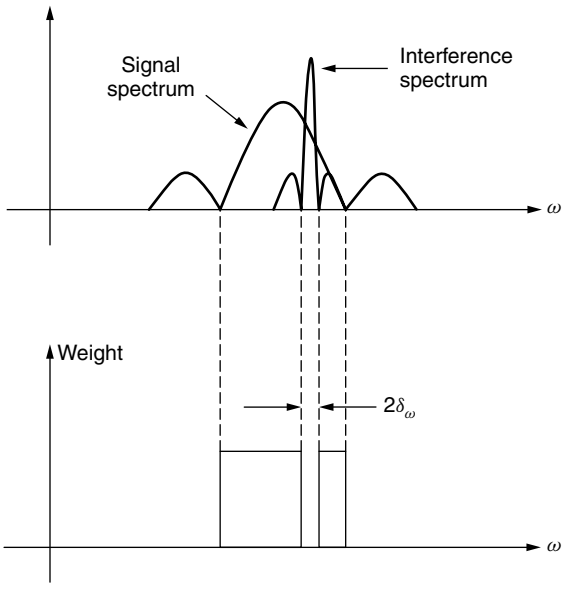


Figure 3. Transform-domain notch filtering.

in the time domain with a nonrectangular weighting function prior to evaluating the transform. In doing so, the levels of the sidelobes of the interference frequency spectrum are attenuated at the expense of broadening the mainlobe [18,14]. In this case, the conventional matched filter is no longer optimal. Using adapted demodulation accordingly can improve the receiver performance [19].

It is important to point out that for transform-domain processing, symbol detection can be performed in either the time or the transform domain. In the later case, filtering and correlation operations can be combined in one step.

The BER expression for transform-domain interference excision can be easily formulated using the Gaussian tail probability or the Q function. The residual filtered and despreaded interference is treated as an equivalent AWGN source. Typically, a uniform interference phase distribution is assumed. When transform-domain filtering is considered, the BER depends on both the excision coefficients and the error misadjustment.

3.2. Synthesis and Subtraction for Sinusoidal Interference

In this section, we view the interference signal as the one that is corrupted by the additive noise and the DSSS signal. In a typical situation, the power level of the DSSS signal is negligible relative to the power level of the interference and, in most cases, relative to the additive noise. For high interference-to-noise ratio (INR), the correlation matrix of the received signal vector consists of a limited number of large eigenvalues contributed mainly by the narrowband interference, and a large number of small and almost equal eigenvalues contributed by the DSSS signal and noise. The eigenanalysis interference canceller is designed with a weight vector orthogonal to the eigenvectors corresponding to the large eigenvalues [20]. The eigendecomposition of the correlation matrix, defined in (9), results in

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^H = [\mathbf{U}_r\mathbf{U}_n]\begin{bmatrix}\boldsymbol{\Sigma}_r & 0\\ 0 & \boldsymbol{\Sigma}_n\end{bmatrix}\begin{bmatrix}\mathbf{U}_r^H\\ \mathbf{U}_n^H\end{bmatrix} \tag{24}$$

where the columns of $\mathbf{U}_r$ span the interference subspace, whereas the columns of $\mathbf{U}_n$ span the signal with noise subspace, and $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_n$ are diagonal matrices whose elements are the eigenvalues of $\mathbf{R}$. For real sinusoidal interference, the number of dimensions of the interference subspace is twice the number of interfering tones.

The projection of the signal vector on the noise subspace results in interference suppressed data sequence

$$\hat{\mathbf{x}}[n] = \mathbf{U}_n\mathbf{U}_n^H\mathbf{x}[n] = (\mathbf{I} - \mathbf{U}_r\mathbf{U}_r^H)\mathbf{x}[n] \tag{25}$$

where $\mathbf{I}$ is the identity matrix.

The subspace projection approach can also be performed using the singular value decomposition (SVD) for the sample data matrix [21].

3.3. Nonlinear Estimation Techniques

The commonly applied predictor/subtracter technique for narrowband interference suppression previously discussed is optimum in the minimum mean-square error (MMSE) sense when trying to predict a Gaussian autoregressive process in the presence of AWGN. If the prediction is done in a non-Gaussian environment, as in the case of SS signals, linear prediction methods will no longer be optimum. In Ref. 2, depending on whether the statistics of the AR process is known or unknown, time-recursive and data-adaptive nonlinear filters with soft-decision feedback are used to estimate the SS signal. For known interference statistics, the interference suppression problem is cast in state space for use with Kalman–Bucy and approximate conditional mean (ACM) filters. A fixed-length LMS transversal filter, on the other hand, is used when there is no a priori statistical information is provided. With the same AR model, both schemes are shown to achieve similar performance, which is an improvement over the Gaussian assumed environment.

3.4. Multiuser Detection Techniques

A narrowband interference could be a digital communication signal with a data rate much lower than the spread-spectrum chip rate. This is typically the case when spread-spectrum signals are used in services overlaying on existing frequency band occupants. In this case, the narrowband interference is a strong communication signal that interferes with commercial DSSS communication systems. This type of interferer is poorly modeled as either a sinusoid or an autoregressive process. Because of the similarity of the spread spectrum signal and the digital interference, techniques from multiuser detection theory are applied to decode the SS user signal and simultaneously suppressing the interferer [22].

In order to apply methods from multiuser detection, the single narrowband interferer is treated as a collection of m spread-spectrum users, where m is a function of the relative data rates of the true SS signal and the interference. That is, m bits of the narrowband user occur for each bit of the SS user. As shown in Fig. 4, and using square waves for illustrations, each narrowband user's bit can be regarded as a signal arising from a virtual user having a signature sequence with only one nonzero entry. The virtual users are orthogonal, but correlated with the SS user.

The optimum receiver implementing the maximum-likelihood (ML) detector has a complexity that is exponential in the number of virtual users, m. To overcome such complexity, the optimal linear detector and

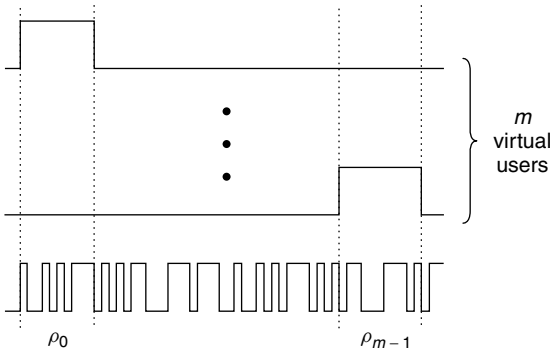


Figure 4. Virtual CDMA systems (synchronous case).

decorrelating detector are applied. While the first requires knowledge of relative energies of both the narrowband interferer and the SS user and maximizes the receiver asymptotic efficiency, the latter is independent of the receiver energies and achieves the near–far resistance of the ML detector. The asymptotic efficiency is the limit of the receiver efficiency as the AWGN goes to zero. It characterizes the detector performance when the dominant source of corruption is the narrowband interferer rather than the AWGN. The receiver efficiency, on the other hand, quantifies the SS user energy that would achieve the same probability of error in a system with the same AWGN and no other users. The input of both detectors is the output of the filter bank and consists of filters matched to the spreading codes of each active user, as depicted in Fig. 5.

The following expressions have been derived [22] for the probability of errors of four different detectors (in all four cases, it is assumed that the narrowband signal is synchronized with the SS signal):

1. *Conventional Detector* (CD), where the received signal is sent directly to a single filter matched to the spreading code. The output of the filter is then compared to a threshold to yield the spread-spectrum bit estimate. This detector is only optimum in the case of a single spread-spectrum user in AWGN. The BER is given by

$$P_{\text{cd}} = \frac{1}{2^m} \sum_{i=0}^{2^m-1} Q\left(\frac{\sqrt{w_2}(1-\alpha \mathbf{p}^T \mathbf{q}^i)}{\sigma_n}\right) \tag{26}$$

   where $\alpha = \sqrt{w_1/w_2}$, w_1 is the received energy of the narrowband interference, w_2 is the received energy of the SS user (including the process gain), $\mathbf{p}$ is the vector formed by the cross correlation between the narrowband interference waveform and the DSSS signal waveform, $\mathbf{q}$ is the narrowband interference data bits, and $\{\mathbf{q}^i\}$ is an ordering of the 2^m possible values of the vector of narrowband bits.

2. *Decorrelating Detector* (DD), where the last row of the inverse of the cross-correlation matrix of the $m+1$ users is used to multiply the output of the $m+1$ matched filters, followed by a threshold comparison for bit estimate. The BER is given by

$$P_{\text{dd}} = Q\left(\frac{\sqrt{w_2(1-\mathbf{p}^T\mathbf{p})}}{\sigma_n}\right) \tag{27}$$

3. *Optimum Linear Detector* (OLD), where the user energies are used to maximize the asymptotic efficiency. The BER is given by

$$P_{\text{old}} = \frac{1}{2^m} \sum_{i=0}^{2^m-1} Q\left(\frac{\sqrt{w_2}(1+\alpha \mathbf{v}^T\mathbf{p} - \alpha(\alpha \mathbf{v}^T + \mathbf{p}^T)\mathbf{q}^i)}{\sigma_n \sqrt{1+2\alpha \mathbf{v}^T\mathbf{p} + \alpha^2 \mathbf{v}^T\mathbf{v}}}\right) \tag{28}$$

   where the ith element of vector $\mathbf{v}$ is given by

$$v_i = \begin{cases} 1 & -\rho_i > \alpha \\ -1 & \rho_i > \alpha \\ -\dfrac{\rho_i}{\alpha} & \text{otherwise.} \end{cases} \tag{29}$$

4. *Ideal Predictor/Subtracter* (IPS), which is similar to the transversal filter excision techniques described in Section 3.1. Perfect knowledge of the narrowband signal is assumed. Further, it is assume that perfect prediction to the sample interior to the narrowband bit is achieved and the only error occurs when predicting at bit transitions. The expressions have been derived [22], where one detector assumes zero bit estimate of the narrowband bit at the transition and the other detector takes this estimate to be random. For the former detector

$$P_{\text{ips}} = \frac{1}{2^m} \sum_{i=0}^{2^m-1} Q\left(\frac{\sqrt{w_2}(1-\alpha \tilde{\mathbf{p}}^T \mathbf{q}^i)}{\sigma_n}\right) \tag{30}$$

   and for the other detector

$$P_{\text{ips}} = \frac{1}{2^m} \sum_{i=0}^{2^m-1} \frac{1}{2^m} \sum_{j=0}^{2^m-1} Q\left(\frac{\sqrt{w_2}(1-\alpha \tilde{\mathbf{p}}^T(\mathbf{q}^i - \hat{\mathbf{q}}^j))}{\sigma_n}\right) \tag{31}$$

   where $\hat{\mathbf{q}}^j$ is the estimate of $\mathbf{q}^j$, and $\tilde{\mathbf{p}}$, defined only over the chip interval encompassing a narrowband bit transition, is the vector formed by the cross-correlation between the narrowband interference and the DSSS signal.

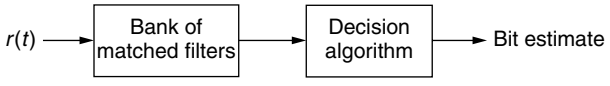


Figure 5. Multiuser detector structure.

The performance of the optimum linear and decorrelator detectors, representing the multiuser detection techniques, has been shown [22] to be similar for different interference power and bandwidth. Both techniques significantly outperform the conventional detector and the predictor/subtracter, when using a 7-tap LMS prediction filter. The improvement in BER is more pronounced for stronger and less narrowband interferers. The advantage of the decorrelator detector over the other conventional and adaptive prediction filters remains unchanged when considering asynchronous interference.

Figure 6 depicts the BER comparison between the conventional detector, decorrelating detector, optimum linear detector, and ideal predictor/subtracter with $m = 2$ and $L = 63$. This figure is in agreement with the performance figures [22] and conforms with the same observations stated above. It is clear from Fig. 6 that the matched filter performs well for weak interferers. The optimum linear detector offers slight improvement over the decorrelating detector. The ideal predictor/subtracter outperforms the decorrelating detector for moderate values of interference power. It is important to note,

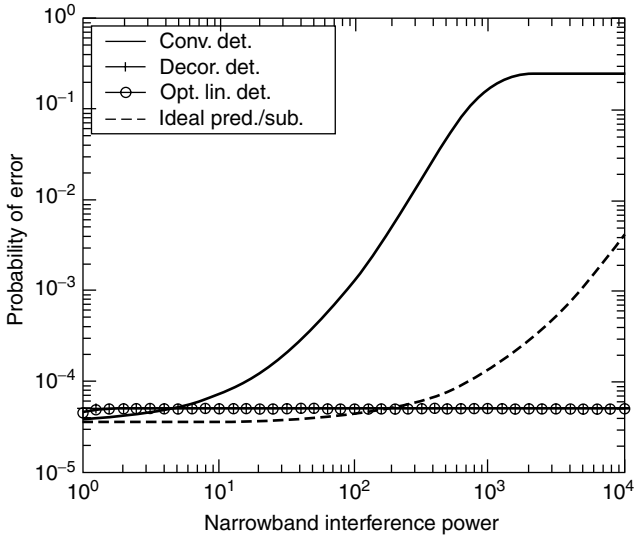


Figure 6. BER performance for different multiuser detection techniques.

however, that the actual predictor/subtracter performance will have much greater error probability [22].

3.5. Minimum Mean-Square Error Algorithm

The minimum mean-square error (MMSE) algorithm, originally proposed for suppressing multiple-access interference in CDMA multiuser detection problems, has been employed for narrowband interference mitigation [23]. Using the signal-to-interference ratio and its upper bounds as a performance measure, the MMSE has been compared with linear and nonlinear techniques in suppressing three types of interference: single-tone and multitone signals, autoregressive process, and digital communications signal with a data rate much lower than the spread-spectrum chip rate. The linear estimators include the conventional matched filter detector, the predictor/subtracter, and the interpolator/subtracter techniques. The latter is based on using a fixed number of past and future samples [23]. The nonlinear techniques include those based on prediction and interpolation. It is shown that the MMSE detector completely suppresses the digital interference, irrespective of its power, and provides performance similar to that using the nonlinear interpolator/subtracter method, when dealing with AR type of interference.

4. NONSTATIONARY INTERFERENCE SUPPRESSION

The interference excision techniques discussed in the previous sections deal with stationary or quasistationary environment. The interference frequency signature, or characteristics, is assumed fixed or slowly time-varying. None of these techniques is capable of effectively incorporating the suddenly changing or evolutionary rapidly time-varying nature of the frequency characteristics of the interference. These techniques all suffer from their lack of intelligence about interference behavior in the joint time–frequency (t-f) domain and therefore are limited in their results and their applicability. For the time-varying

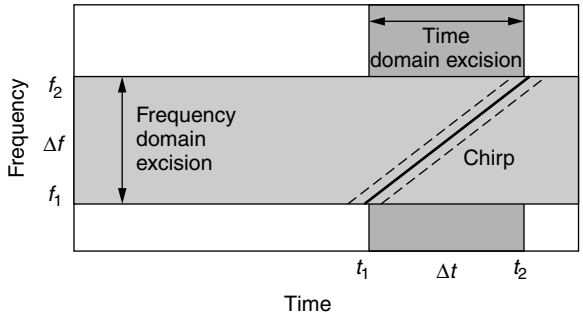


Figure 7. Excision methods for nonstationary interferers.

interference depicted in Fig. 7, frequency-domain methods remove the frequency band Δf and ignore the fact that only few frequency bins are contaminated by the interference at a given time. Dually, time domain excision techniques, through gating or clipping the interference over Δt, do not account for the cases where only few time samples are contaminated by the interference for a given frequency. Applying either method will indeed eliminate the interference but at the cost of unnecessarily reducing the desired signal energy. Adaptive excision methods might be able to track and remove the nonstationary interference, but would fail if the interference is highly nonlinear FM or linear FM, as in Fig. 7, with high sweep rates. Further, the adaptive filtering length or block transform length trades off the temporal and the spectral resolutions of the interference. Increasing the step-size parameter increases the filter output errors at convergence, and causes an unstable estimate of the interference waveform.

The preceding example clearly demonstrates that nonstationary interferers, which have model parameters that rapidly change with time, are particularly troublesome due to the inability of single-domain mitigation algorithms to adequately ameliorate their effects. In this challenging situation, and others like it, joint t-f techniques can provide significant performance gains, since the instantaneous frequency (IF), the instantaneous bandwidth, and the energy measurement, in addition to myriad other parameters, are available. The objective is then to estimate the t-f signature of the received data using t-f analysis, attenuating the received signal in those t-f regions that contain strong interference. This is depicted by the region in between the dashed lines in Fig. 7.

An FM interference in the form $u(n) = e^{j\phi(n)}$ is solely characterized by its IF, which can be estimated using a variety of IF estimators, including the time–frequency distributions (TFDs) [24,25].

The TFD of the data, $x(n)$, at time t and radian frequency ω, is given by

$$C_f(t,\omega,\phi) = \sum_{l=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \phi(m,l)x(n+m+l) \times x^*(n+m-l)e^{-j2\omega l} \tag{32}$$

where $\phi(m,l)$ is the time–frequency kernel, which is a function of the lag l and time lag m. Several requirements have been imposed on $\phi(m,l)$ to satisfy desirable

distribution properties, including power localization at the IF. As shown in Eq. (32), the TFD is the Fourier transform of a time-average estimate of the autocorrelation function.

A time–frequency notch filter can be designed, in which the position of the filter notch is synchronous with the interference IF estimate. Based on the IF, two constraints should exist to construct an interference excision filter with desirable characteristics. First, an FIR filter with short impulse response must be used. Long-extent filters are likely to span segments of changing frequency contents and, as such, allow some of the interference components to escape to the filter output. Second, at any given time, the filter frequency response must be close to an ideal notch filter to be able to null the interference with minimum possible distortion of the signal. This property, however, requires filters with infinite or relatively long impulse responses.

Amin [26] has shown that a linear-phase 5-coefficient filter is effective in FM interference excision. Assuming exact IF values, the corresponding receiver SINR is given by

$$\text{SINR} = \frac{L}{11/8 + 9\sigma_n^2/4} \tag{33}$$

which shows that full interference excision comes at the expense of a change in the noise variance in addition of a self-noise form, as compared with the noninterference case. The main objective of any excision process is to reduce both effects. The SINR in (33) assumes a random IF with uniform distribution over $[0, 2\pi]$. For an interference with fixed frequency ω_0, the receiver SINR becomes dependent on ω_0. The receiver performance sensitivity to the interference frequency is discussed in detail in Ref. 26.

Wang and Amin [27] considered the performance analysis of the IF-based excision system using a general class of multiple-zero FIR excision filters showing the dependence of the BER on the filter order and its group delay. The effect of inaccuracies in the interference IF on receiver performance was also considered as a function of the filter notch bandwidth. Closed-form approximations for SINR at the receiver are given for the various cases.

One of the drawbacks to the notch filter approach [26] is the infinite notch depth due to the placement of the filter zeros. The effect is a "self-noise" inflicted on the received signal by the action of the filter on the PN sequence underlying the spread information signal. This problem led to the design of an open-loop filter with adjustable notch depth based on the interference energy. The notch depth is determined by a variable embedded in the filter coefficients chosen as the solution to an optimization problem that maximizes receiver SINR. The TFD is necessary for this work, even for single component signals, because simple IF estimators do not provide energy information. Amin et al. accomplished this work [28], incorporating a "depth factor" into the analysis and redeveloping all the SINR calculations. The result was a significant improvement in SINR, especially at midrange interference-to-signal ratios (ISR's), typically around 0–20 dB.

Instead of using time-varying excision filters, Barbarossa and Scaglione [29] proposed a two-step procedure based on dechirping techniques commonly applied in radar algorithms. In the first step the time-varying interference is converted to a fixed-frequency sinusoid eliminated by time-invariant filters. The process is reversed. In the second step and the interference-free signal is multiplied by the interference t-f signature to restore the DSSS signal and noise characteristics that have been strongly impacted in the first phase.

Similar to the predictor/subtracter method discussed in Section 3, Lach et al. proposed synthesis/subtracter technique for FM interference using TFD [30]. A replica of the interference can be synthesized from the t-f domain and subtracted from the incoming signal to produce an essentially interference-free channel.

Another synthesis/subtracter method has been introduced [31] where the discrete evolutionary and the Hough transforms are used to estimate the IF. The interference amplitude is found by conventional methods such as linear filtering or singular value decomposition. This excision technique applies equally well to one or multicomponent chirp interferers with constant or time-varying amplitudes and with instantaneous frequencies not necessarily parametrically modeled.

To overcome the drawbacks of the potential amplitude and phase errors produced by the synthesis methods, Amin et al. [32] proposed a projection filter approach in which the FM interference subspace is constructed from its t-f signature. Since the signal space at the receiver is not specifically mandated, it can be rotated such that a single interferer becomes one of the basis functions. In this way, the interference subspace is one-dimensional and its orthogonal subspace is interference-free. A projection of the received signal onto the orthogonal subspace accomplishes interference excision with a minimal message degradation. The projection filtering methods compare favorably over the previous notch filtering systems.

Zhang et al. [33] proposed a method to suppress more general INBI signals. The interference subspace is constructed using t-f synthesis methods. In contrast to the work in Ref. 30, the interferer is removed by projection rather than subtraction. To estimate the interference waveform, a mask is constructed and applied such that the masked t-f region captures the interference energy, but leaves out most of the DSSS signals.

Seong and Loughlin have also extended the projection method developed by Amin et al. [32] for excising constant amplitude FM interferers from DSSS signals to the case of AM/FM interferers [34]. Theoretical performance results (correlator SNR and BER) for the AM/FM projector filter show that FM estimation errors generally cause greater performance degradation than the same level of error in estimating the AM. The lower-bound for the correlator SINR for the AM/FM projection filter for the case of both AM and FM errors is given by

$$\text{SINR} = \frac{L-1}{\dfrac{1}{L} + \sigma_n^2 + A^2\left[\dfrac{1}{1+\sigma_{\Delta a}^2}(1 - e^{-\sigma_{\Delta\phi}^2}) + \sigma_{\Delta a}^2\right]} \tag{34}$$

where L is the PN sequence length, A^2 is the interference power, σ_n^2 is the variance of AWGN, and $\sigma_{\Delta a}^2$ and $\sigma_{\Delta\phi}^2$ are

the variances of the estimation errors in the AM and FM, respectively.

Linear t-f signal analysis has also been shown effective to characterize a large class of nonstationary interferers. Roberts and Amin [35] proposed the use of the discrete Gabor transform (DGT) as a linear joint time–frequency representation. The DGT can attenuate a large class of nonstationary wideband interferers whose spectra are localized in the t-f domain. Compared to bilinear TFDs, the DGT does not suffer from the cross-term interference problems, and enjoys a low computational complexity. Wei et al. [36] devised a DGT-based, iterative time-varying excision filtering, in which a hypothesis testing approach was used to design a binary mask in the DGT domain. The time–frequency geometric shape of the mask is adapted to the time-varying spectrum of the interference. They show that such a statistical framework for the transform-domain mask design can be extended to any linear transform. Both the maximum-likelihood test and the local optimal test are presented to demonstrate performance versus complexity.

Application of the short-time Fourier transform (STFT) to nonstationary interference excision in DSSS communications has been considered [37,38]. In those studies [37,38], due to the inherent property of STFT to trade off temporal and spectral resolutions, several STFTs corresponding to different analysis windows were generated. Ouyang and Amin [37] used a multiple-pole data window to obtain a large class of recursive STFTs. Subsequently, they employed concentration measures to select the STFT that localizes the interference in the t-f domain. This procedure is followed by applying a binary excision mask to remove the high-power t-f region. The remainder is synthesized to yield a DSSS signal with improved signal-to-interference ratio (SIR).

Krongold et al. [38] proposed multiple overdetermined tiling techniques and utilized a collection of STFTs for the purpose of interference excision. Unlike the Ouyang–Amin procedure [37], Krongold et al. [38] removed the high-value coefficients in all generated STFTs, and used the combined results, via efficient least-square synthesis, to reconstruct an interference-reduced signal. Bultan and Akansu [39] proposed a chirplet-transform-based exciser to handle chirplike interference types in SS communications.

The block diagram in Fig. 8 depicts the various interference rejection techniques using the time–frequency methods cited above.

4.1. Example

At this point, in order to further illustrate these excision methods, the work by Amin et al. [32] will be detailed since it includes comparisons between the two most prominent techniques based on TFDs currently being studied: notch filtering and projection filtering. The signal model is, as expected, given by Eq. (1), and the major theme of the work is to annihilate interference via projection of the received signal onto an "interference-free" subspace generated from the estimated interference characteristics. This paper includes a figure, reprinted here as Fig. 9, which clearly illustrates the tradeoffs between projection and notch filtering based on the ISR. In the legend, the variable a represents the adaptation parameter for the notch filtering scheme and N represents the block size, in samples, for a 128-sample bit duration in the projection method. Thus, $N = 128$ means no block processing and $N = 2$ corresponds to 64 blocks per bit being processed for projection. Since the projection and nonadaptive notch filter techniques are assumed to completely annihilate the interference, their performance is decoupled from the interference power, and therefore correctly indicate constant SINR across the graph. The dashed line representing the notch filter with $a = 0$ really indicates no filtering at all, since the adaptation parameter controls the depth of the notch.

It is evident from Fig. 9 that without adaptation a crossover point occurs around 2 dB, where filtering with an infinitely deep notch is advantageous. Thus, when the interference power exceeds this point, presumably

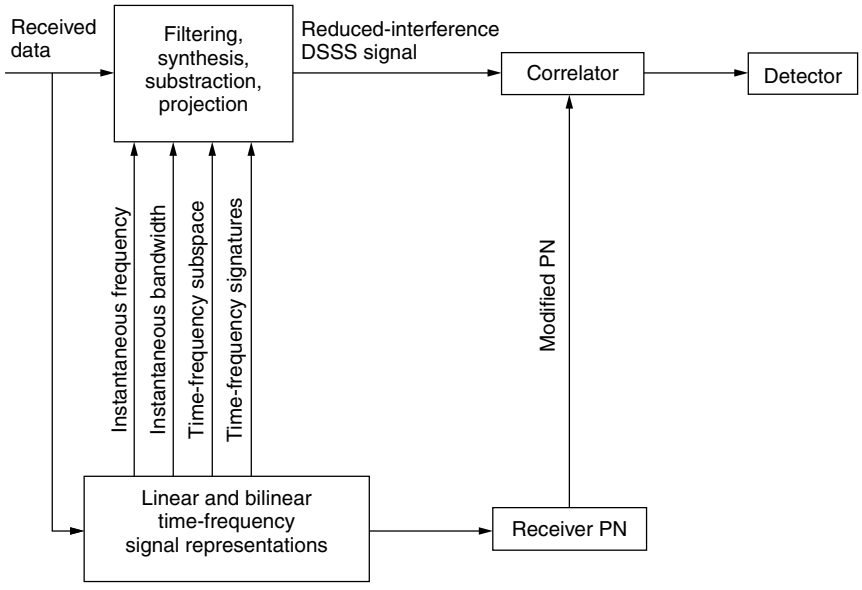


Figure 8. Interference rejection techniques.

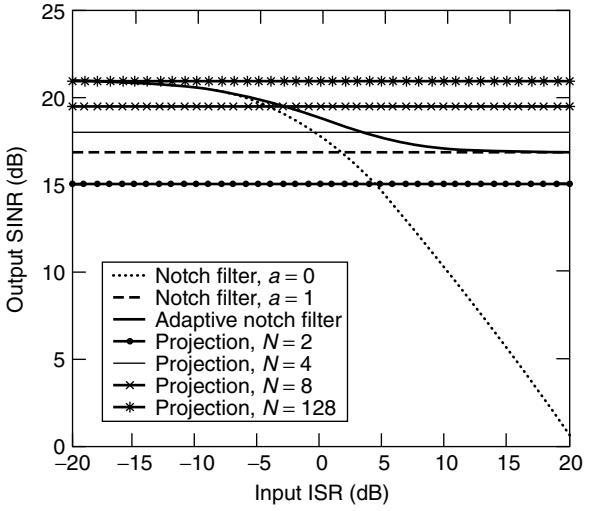


Figure 9. Comparison between projection and notch filtering excision methods.

a user would flip a switch to turn on the excision subsystem. However, with adaptation, this process happens automatically, while giving superior performance in the midrange. For the projection technique, the block size determines receiver performance conspicuously (ceteris paribus). Most important to note, however, is the superior performance of projection over all methods when the block size is equal to the bit duration, namely, no block processing. It is feasible that computational complexity may warrant a tradeoff between SINR and block size, in which case a hybrid implementation may be of benefit — one that automatically switches between adaptive notch filtering and projection depending on the desired SINR. In any case, this example illustrates the parameters involved in the design of modern excision filters for nonstationary interferers.

5. INTERFERENCE SUPPRESSION FOR FREQUENCY-HOPPING COMMUNICATIONS

Interference rejection for FH is not as well developed as interference rejection for DS or for CDMA. In FH systems, the fast FH (FFH) is of most interest, and the modulation most commonly used in FH is frequency shift keying (FSK). Two types of interference waveforms can be categorized, namely, partial-band interference (PBI) and multitone interference (MTI).

The effects of PBI and AWGN on several diversity-combining receivers in FFH/FSK SS communication systems have been investigated [40–42,44]. An alternative method using a fast Fourier transform (FFT) has been proposed [45]. An automatic gain-control (AGC) receiver using a diversity technique has also been presented [40]. In this method, each soft-decision square-law detected MARK and SPACE filter output is weighted by the inverse of the noise power in the slot prior to linear combining. This method is near-optimal (in terms of SNR) if the exact information of noise and interference power can be obtained. A similar clipped-linear combining receiver was also reported [42]. Because of the difficulty of such information, self-normalizing receivers [41] and the ratio-statistic combining technique [43] use the output values of the square-law detector in each hop to derive a weight or normalizing factor. The performance of these two methods is shown to be comparable to that of the square-law clipper receiver.

An FFH receiver that employs a prewhitening filter to reject NBI has been described [44]. For binary FSK modulations, it is shown that the FFH signal is statistically uncorrelated at lag values of $T_h/(4N_h)$, where T_h is the hop duration and $2N_h$ is the total number of frequency slots (i.e., there are N_h MARK and N_h SPACEs). Thus, as in the DS case, NBI can be predicted and suppressed independently of the desired FFH signal. Using the complex LMS algorithm to update the prewhitening filter coefficients, this technique is shown to compare favorably with the maximal-ratio combiner diversity technique. When the interferer is wide-sense stationary, the prewhitening-filter-based receiver provides performance approaching that of the AGC receiver and at least 2–3 dB superior to that of the self-normalizing receiver. However, when hostile interference is present, the adaptive prewhitening filter technique may not be able to track the interference rapidly enough. In this case, nonparametric techniques such as the self-normalizing receiver must be used to reject the jammed hops.

Reed and Agee [46] have extended and improved on the idea of whitening by using a time-dependent filter structure to estimate and remove interference, based on the interference spectral correlation properties. In this method, the detection of FHSS in the presence of spectrally correlated interference is nearly independent of the SIR. The process can be viewed as a time-dependent whitening process with suppression of signals that exhibit a particular spectral correlation. The technique is developed from the maximum-likelihood estimate of the spectral frequency of a frequency agile signal received in complex Gaussian interference with unknown spectral correlation. The resulting algorithm uses the correlation between spectrally separated interference components to reduce the interference content in each spectral bin prior to the whitening/detection operation.

An alternative approach to suppress PBI using the FFT has been proposed [45]. The major attraction of FFT-based implementation lies in the ability to achieve guaranteed accuracy and perfect reproducibility.

For suppression of MTI, basically the same processing methods applied for PBI can be employed. However, the performance analyses differ from those for PBI situations. The performance depends on the distribution of the MTI and, in turn, how many bands are contaminated by MTI. Performance analyses of FFH SS systems have been presented for linear combining diversity [47,48], for clipped diversity [49], for maximum likelihood and product-combining receivers [50,51].

BIOGRAPHIES

Dr. Moeness Amin received his B.Sc. degree in 1976 from Cairo University, Egypt, M.Sc. degree in 1980 from

University of Petroleum and Minerals, Dhahran, Saudi Arabia, and his Ph.D. degree in 1984 from University of Colorado at Boulder. All degrees are in electrical engineering. He has been on the faculty of the Department of Electrical and Computer Engineering at Villanova University, Villanova, Pennsylvania, since 1985, where is now a professor. Dr. Amin is a fellow of the IEEE and the recipient of the IEEE Third Millennium Medal. He is also a recipient of the 1997–Villanova University Outstanding Faculty Research Award as well as the recipient of the 1997–IEEE Philadelphia Section Service Award. He is a member of the Franklin Institute Committee of Science and Arts. Dr. Amin has 4 book chapters, 60 journal articles, 2 review articles, and over 150 conference publications. His research includes the areas of wireless communications, time-frequency analysis, smart antennas, anti-jamming gPS, interference cancellation in broadband communication platforms, digitized battlefield, high definition TV, target tracking and direction finding, channel equalization, signal coding and modulation, and radar systems. He has two U.S. patents and served as a consultant to Micronetics Wireless, ELCOM Technology Corporation, and VIZ Manufacturing Company.

Yimin Zhang received his M.S. and Ph.D. degrees from the University of Tsukuba, Japan, in 1985 and 1988, respectively. He joined the faculty of the Department of Radio Engineering, Southeast University, Nanjing, China, in 1988. He served as a technical manager at Communication Laboratory Japan, Kawasaki, Japan, in 1995–1997, and a visiting researcher at ATR Adaptive Communications Research Laboratories, Kyoto, Japan, in 1997–1998. Currently, he is a research fellow at the Department of Electrical and Computer Engineering, Villanova University, Villanova, Pennsylvania. His current research interests are in the areas of array signal processing, space-time adaptive processing, multiuser detection, blind signal processing, digital mobile communications, and time-frequency analysis.

BIBLIOGRAPHY

1. L. B. Milstein, Interference rejection techniques in spread spectrum communications, *Proc. IEEE* **76**(6): 657–671 (June 1988).
2. H. V. Poor and L. A. Rusch, Narrowband interference suppression in spread-spectrum CDMA, *IEEE Pers. Commun. Mag.* **1**(8): 14–27 (Aug. 1994).
3. J. D. Laster and J. H. Reed, Interference rejection in digital wireless communications, *IEEE Signal Process. Mag.* **14**(3): 37–62 (May 1997).
4. M. G. Amin and A. N. Akansu, Time-frequency for interference excision in spread-spectrum communications (section in Highlights of signal processing for communications), *IEEE Signal Process. Mag.* **15**(5): (Sept. 1998).
5. F. M. Hsu and A. A. Giordano, Digital whitening techniques for improving spread-spectrum communications performance in the presence of narrow-band jamming and interference, *IEEE Trans. Commun.* **COM-26**: 209–216 (Feb. 1978).
6. J. W. Ketchum and J. G. Proakis, Adaptive algorithms for estimating and suppressing narrow-band interference in PN spread-spectrum systems, *IEEE Trans. Commun.* **COM-30**: 913–924 (May 1982).
7. L. Li and L. B. Milstein, Rejection of narrow-band interference in PN spread-spectrum system using transversal filters, *IEEE Trans. Commun.* **COM-30**: 925–928 (May 1982).
8. R. A. Iltis and L. B. Milstein, Performance analysis of narrow-band interference rejection techniques in DS spread-spectrum systems, *IEEE Trans. Commun.* **COM-32**: 1169–1177 (Nov. 1984).
9. S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.
10. R. A. Iltis and L. B. Milstein, An approximate statistical analysis of the Widrow LMS algorithm with application to narrow-band interference rejection, *IEEE Trans. Commun.* **COM-33**: 121–130 (Feb. 1985).
11. F. Takawira and L. B. Milstein, Narrowband interference rejection in PN spread spectrum system using decision feedback filters, *Proc. MILCOM*, Oct. 1986, pp. 20.4.1–20.4.5.
12. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
13. L. B. Milstein and P. K. Das, Spread spectrum receiver using surface acoustic wave technology, *IEEE Trans. Commun.* **COM-25**: 841–847 (Aug. 1977).
14. S. Davidovici and E. G. Kanterakis, Narrowband interference rejection using real-time Fourier transform, *IEEE Trans. Commun.* **37**: 713–722 (July 1989).
15. R. C. Dipietro, An FFT based technique for suppressing narrow-band interference in PN spread spectrum communication systems, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1989, pp. 1360–1363.
16. M. V. Tazebay and A. N. Akansu, Adaptive subband transforms in time-frequency excisers for DSSS communication systems, *IEEE Trans. Signal Process.* **43**: 1776–1782 (Nov. 1995).
17. M. Medley, G. J. Saulnier, and P. Das, Adaptive subband filtering of narrowband interference, in H. Szu, ed., *SPIE Proc. — Wavelet Appls. III*, Vol. 2762, April 1996.
18. J. Gevargiz, M. Rosenmann, P. Das, and L. B. Milstein, A comparison of weighted and non-weighted transform domain processing systems for narrowband interference excision, *Proc. MILCOM*, 1984, pp. 32.3.1–32.3.4.
19. S. D. Sandberg, Adapted demodulation for spread-spectrum receivers which employ transform-domain interference excision, *IEEE Trans. Commun.* **43**: 2502–2510 (Sept. 1995).
20. A. Haimovich and A. Vadhri, Rejection of narrowband interferences in PN spread spectrum systems using an eigenanalysis approach, *Proc. IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, Quebec, Canada, June 1994, pp. 1002–1006.
21. B. K. Poh, T. S. Quek, C. M. S. See, and A. C. Kot, Suppression of strong narrowband interference using eigen-structure-based algorithm, *Proc. MILCOM*, July 1995, pp. 1205–1208.
22. L. A. Rusch and H. V. Poor, Multiuser detection techniques for narrow-band interference suppression in spread spectrum communications, *IEEE Trans. Commun.* **43**: 1725–1737 (Feb.–April 1995).
23. H. V. Poor and X. Wang, Code-aided interference suppression for DS/CDMA communications. I. Interference suppression capability, *IEEE Trans. Commun.* **45**: 1101–1111 (Sept. 1997).

24. B. Boashash, Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals, *Proc. IEEE* **80**: 520–538 (April 1992).

25. B. Boashash, Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications, *Proc. IEEE* **80**: 540–568 (April 1992).

26. M. G. Amin, Interference mitigation in spread-spectrum communication systems using time-frequency distributions, *IEEE Trans. Signal Process.* **45**(1): 90–102 (Jan. 1997).

27. C. Wang and M. G. Amin, Performance analysis of instantaneous frequency based interference excision techniques in spread spectrum communications, *IEEE Trans. Signal Process.* **46**(1): 70–83 (Jan. 1998).

28. M. G. Amin, C. Wang, and A. R. Lindsey, Optimum interference excision in spread-spectrum communications using open-loop adaptive filters, *IEEE Trans. Signal Process.* (July 1999).

29. S. Barbarossa and A. Scaglione, Adaptive time-varying cancellations of wideband interferences in spread-spectrum communications based on time-frequency distributions, *IEEE Trans. Signal Process.* **47**(4): 957–965 (April 1999).

30. S. Lach, M. G. Amin, and A. R. Lindsey, Broadband nonstationary interference excision in spread-spectrum communications using time-frequency synthesis techniques, *IEEE J. Select. Areas Commun.* **17**(4): 704–714 (April 1999).

31. H. A. Khan and L. F. Chaparro, Formulation and implementation of the non-stationary evolutionary Wiener filtering, *Signal Process.* **76**: 253–267 (1999).

32. M. G. Amin, R. S. Ramineni, and A. R. Lindsey, Interference excision in DSSS communication systems using projection techniques, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000.

33. Y. Zhang, M. G. Amin, and A. R. Lindsey, Combined synthesis and projection techniques for jammer suppression in DS/SS communications, *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Orlando, FL, May 2002.

34. S.-C. Jang and P. J. Loughlin, AM-FM interference excision in spread spectrum communications via projection filtering, *J. Appl. Signal Process.* **2001**(4): 239–248 (Dec. 2001).

35. S. Roberts and M. Amin, Linear vs. bilinear time-frequency methods for interference mitigation in direct-sequence spread-spectrum communication systems, *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1995.

36. D. Wei, D. S. Harding, and A. C. Bovik, Interference rejection in direct-sequence spread-spectrum communications using the discrete Gabor transform, *Proc. IEEE Digital Signal Processing Workshop*, Bryce Canyon, UT, Aug. 1998.

37. X. Ouyang and M. G. Amin, Short-time Fourier transform receiver for nonstationary interference excision in direct sequence spread spectrum communications, *IEEE Trans. Signal Process.* **49**(4): 851–863 (April 2001).

38. B. S. Krongold, M. L. Kramer, K. Ramchandran, and D. L. Jones, Spread-spectrum interference suppression using adaptive time-frequency tilings, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Munich, Germany, April 1997.

39. A. Bultan and A. N. Akansu, A novel time-frequency exciser in spread-spectrum communications for chirp-like interference, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Seattle, WA, May 1998.

40. J. S. Lee, L. E. Miller, and Y. K. Kim, Probability of error analysis of a BFSK frequency-hopping system with diversity — Part II, *IEEE Trans. Commun.* **COM-32**: 1243–1250 (Dec. 1984).

41. L. E. Miller, J. S. Lee, and A. P. Kadrichu, Probability of error analyses of a BPSK frequency-hopping system with diversity under partial-band jamming interference — Part III: Performance of a square-law self-normalizing soft decision receiver, *IEEE Trans. Commun.* **COM-34**: 669–675 (July 1986).

42. C. M. Keller and M. B. Pursley, Clipper diversity combining for channels with partial-band interference — Part I: Clipper linear combining, *IEEE Trans. Commun.* **COM-35**: 1320–1328 (Dec. 1987).

43. C. M. Keller and M. B. Pursley, Clipper diversity combining for channels with partial-band interference — Part II: Ratio-statistic combining, *IEEE Trans. Commun.* **COM-37**: 145–151 (Feb. 1989).

44. R. A. Iltis, J. A. Ritcey, and L. B. Milstein, Interference rejection in FFH systems using least squares estimation techniques, *IEEE Trans. Commun.* **38**: 2174–2183 (Dec. 1990).

45. K. C. Teh, A. C. Kot, and K. H. Li, Partial-band jammer suppression in FFH spread-spectrum system using FFT, *IEEE Trans. Vehic. Technol.* **48**: 478–486 (March 1999).

46. J. H. Reed and B. Agee, A technique for instantaneous tracking of frequency agile signals in the presence of spectrally correlated interference, *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1992.

47. B. K. Livitt, FH/MFSK performance in multitone jamming, *IEEE J. Select. Areas Commun.* **SAC-3**: 627–643 (Sept. 1985).

48. R. E. Ezers, E. B. Felstead, T. A. Gulliver, and J. S. Wight, An analytical method for linear combining with application to FFH NCFSK receivers, *IEEE J. Select. Areas Commun.* **11**: 454–464 (April 1993).

49. J. J. Chang and L. S. Lee, An exact performance analysis of the clipped diversity combining receiver for FH/MFSK systems against a band multitone jammer, *IEEE Trans. Commun.* **42**: 700–710 (Feb.–April 1994).

50. K. C. Teh, A. C. Kot, and K. H. Li, Performance study of a maximum-likelihood receiver for FFH/BFSK systems with multitone jamming, *IEEE Trans. Commun.* **47**: 766–772 (May 1999).

51. K. C. Teh, A. C. Kot, and K. H. Li, Performance analysis of an FFH/BFSK product-combining receiver under multitone jamming, *IEEE Trans. Vehic. Technol.* **48**: 1946–1953 (Nov. 1999).

INTERLEAVERS FOR SERIAL AND PARALLEL CONCATENATED (TURBO) CODES

TOLGA M. DUMAN
Arizona State University
Tempe, Arizona

1. INTRODUCTION

Interleavers are commonly used devices in digital communication systems. Basically, an interleaver is used to reorder a block or a sequence of binary digits [1].

Figure 1. Usage of an interleaver in a coded digital communication system.

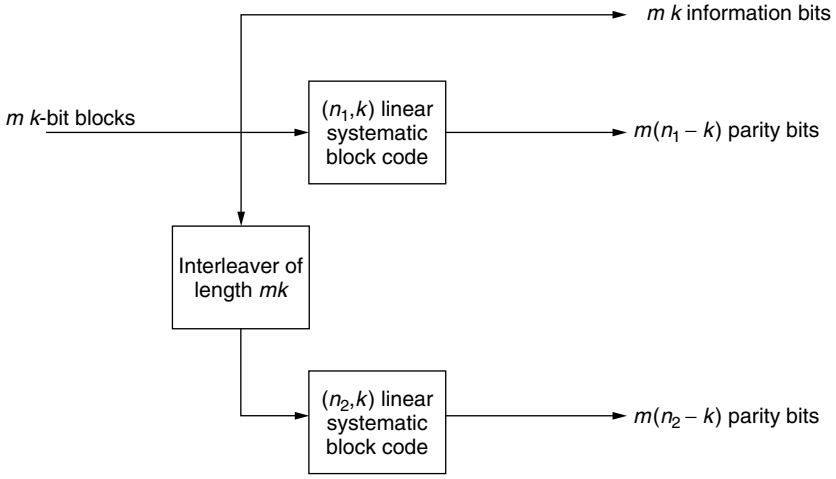


Figure 2. Parallel concatenated block code via an interleaver.

Traditionally interleavers have been employed in coded digital communications, as shown in Fig. 1, to reorder the coded bits in order to combat the burst errors that may be caused by signal fading or different types of interfering signals.

A simple way to reorder the coded bits for this purpose is to use a block interleaver where the sequence is written to a matrix rowwise, and read columnwise. For example, if the original sequence is $\{u_1, u_2, u_3, \ldots, u_{N_1N_2}\}$, the $N_1 \times N_2$ matrix

$$\begin{bmatrix} u_1 & u_2 & \cdots & u_{N_1} \\ u_{N_1+1} & u_{N_1+2} & \cdots & u_{2N_1} \\ u_{2N_1+1} & u_{2N_1+2} & \cdots & u_{3N_1} \\ \cdots & \cdots & \cdots & \cdots \\ u_{(N_2-1)N_1+1} & u_{(N_2-1)N_1+2} & \cdots & u_{N_1N_2} \end{bmatrix}$$

is constructed and read columnwise, resulting in the interleaved sequence

$$\{u_1, u_{N_1+1}, u_{2N_1+1}, \ldots, u_{(N_2-1)N_1+1}, u_2, u_{N_1+2}, u_{2N_1+2}, \ldots, u_{N_1N_2}\}$$

Clearly, the order in which the data are written or read may change, resulting in different variations of the block interleaver.

A more current use of interleavers is in the construction of channel codes having very large codelengths. In his classic paper [2], Shannon showed that the codes with large blocklengths chosen randomly achieve the channel capacity. However, such codes are in general very difficult to encode or decode. As a remedy, one can construct codes with large blocklengths by concatenating two simple (short-blocklength) codes either in parallel or in series using an interleaver.

An example of a parallel concatenated code is shown in Fig. 2 [3]. Two linear systematic block codes are used as component encoders: m k-bit blocks are input to the first component encoder, which produces $m(n_1 - k)$ parity bits, and the interleaved version is input to the second encoder, which produces $m(n_2 - k)$ parity bits. The information bits together with the two sets of parity bits constitute the codeword corresponding to the original mk information sequence. The overall code is an $(m(n_1 + n_2 - k), mk)$ systematic block code. Clearly, by proper selection of the parameter m, we can obtain a large-blocklength code. Different code rates can be obtained by selecting the component codes properly, and by puncturing some of the redundant bits produced by the component codes.

Similarly, simple block codes can be concatenated in series to construct codes with large blocklengths. An example is shown in Fig. 3. In this case, m k-bit blocks are input to an outer linear systematic block code that produces m p-bit coded blocks. These coded bits are interleaved and then input to an inner encoder that generates m n-bit blocks that constitute

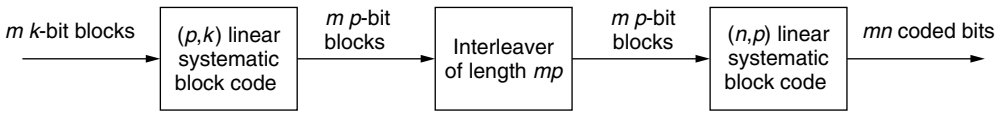


Figure 3. Serially concatenated block code via an interleaver.

the overall codeword corresponding to the original information sequence. The overall code is an (mn, mk) linear systematic block code.

As will be demonstrated later, a simple block interleaver is seldom a good choice for constructing concatenated codes because they fail to eliminate problematic error events because of their inherent structure. Therefore, we need to develop methods of generating good interleavers for code concatenation.

The article is organized as follows. In Section 2, we briefly describe Turbo codes together with the iterative decoding algorithms and explain the role of the interleaver. We summarize the results on the interleaver gain for serial and parallel concatenated convolutional codes in Section 3. We devote Section 4 to review of various promising interleaver design techniques and present several examples. We specifically focus on the use of Turbo codes over AWGN (additive white Gaussian noise) channels. Finally, we conclude in Section 5.

2. TURBO CODES

While concatenation of simple block codes with an interleaver is a viable solution for constructing large-blocklength codes to approach the Shannon limit, encoding and decoding of such codes is difficult since block codes rarely admit simple (soft-decision) maximum-likelihood decoding. Therefore, it is desirable to design concatenated codes using simple convolutional codes as the building blocks.

Parallel or serial concatenation of convolutional codes via an interleaver (i.e., Turbo codes [4,5]), coupled with a suboptimal iterative decoder, has proved to be one of the most important developments in the coding literature since the early 1990s. In particular, long "randomlike" block codes with rather simple decoding algorithms are constructed, and it is shown that their performance is very close to the Shannon limit on the AWGN channel. To be specific, at a bit error rate of 10^{-5}, performance within 1 dB of the channel capacity is common. Let us now describe these codes in more detail.

2.1. Parallel Concatenated Convolutional Codes

The idea in Turbo coding is to concatenate two recursive systematic convolutional codes in parallel via an interleaver as shown in Fig. 4. The information sequence is divided into blocks of a certain length. The input of the first encoder is the information block, and the input of the second encoder is an interleaved version of the information block. The encoded sequence (codeword) corresponding to that information sequence is then the information block itself, the first parity block, and the second parity block. The block diagram shown in Fig. 4 assumes a rate-$\frac{1}{3}$ Turbo code with $\frac{5}{7}$ (in octal notation) component convolutional codes.[1] As in the case of parallel concatenated block codes, higher rate codes can be obtained by puncturing some of the parity bits.

The main role of the interleaver in this construction is to help the code imitate a "random" code with a large blocklength.

2.2. Serially Concatenated Convolutional Codes

Convolutional codes can also be concatenated in series via an interleaver to construct powerful error-correcting

[1] The term p/q *convolutional code* is used to indicate the places of the feedforward and feedback connections in the convolutional code in octal notation.

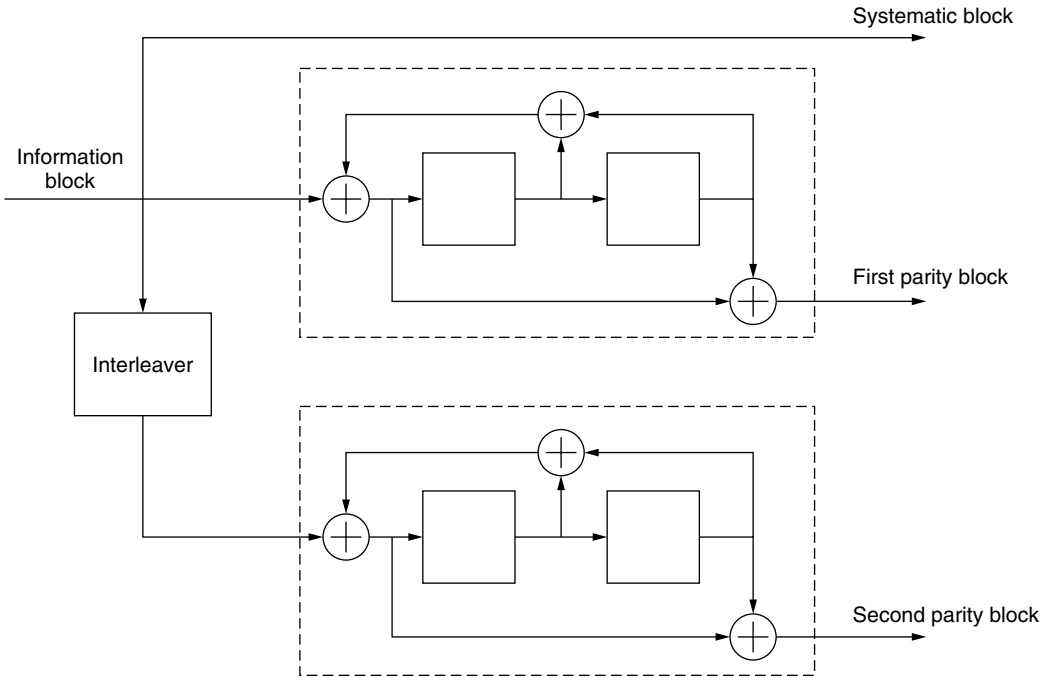


Figure 4. Rate-$\frac{1}{3}$ Turbo code with $\frac{5}{7}$ component codes.

Figure 5. Block diagram for a serially concatenated convolutional code.

(Turbo) codes. In this case, the information sequence is divided into blocks, and each block is encoded using an outer convolutional code. Then, the encoded block is scrambled using an interleaver and passed to the inner convolutional code. Both inner and outer codes are usually selected to be systematic. In order to obtain a good performance, the inner code needs to be recursive, whereas the outer code may be selected as a feedback–free convolutional code [5]. The block diagram of the encoder is shown in Fig. 5.

Similarly, the interleaver is employed to generate a "randomlike" long-blocklength code.

2.3. Iterative Decoding

Assume that the parallel or serial concatenated code is used for transmission over an AWGN channel. Because of the usage of the interleaver, performing maximum-likelihood decoding is very difficult. In general, one has to consider all the possible codewords (there are 2^N possibilities, where N is the length of the information block), compute the squared Euclidean distance corresponding to each one, and declare the one with the lowest distance as the transmitted codeword. Even for short interleavers, this is a tedious task and cannot be done in practice. Fortunately, there is a suboptimal iterative decoding algorithm that achieves near-optimal performance.

Consider the case of parallel concatenation. The iterative decoding algorithm is based on two component decoders, one for each convolutional code, that can produce soft information about the transmitted bits. At each iteration step, one of the decoders takes the systematic information (directly from the observation of the systematic part) and the extrinsic loglikelihood information produced by the other decoder in the previous iteration step, and computes its new extrinsic loglikelihood information. The newly produced extrinsic information is ideally independent of the systematic information and the extrinsic information of the other decoder. The block diagram of the iterative decoding algorithm is presented in Fig. 6, where the extrinsic information of the jth component decoder about the information bit d_k is denoted by $L_{je}(d_k)$ and its interleaved version by $L'_{je}(d_k)$, $j = 1, 2$. The updated extrinsic information is fed into the other decoder for the next iteration step. The extrinsic information of both decoders is initialized to zero before the iterations start. After a number of iterations, the algorithm converges and the decision on each transmitted bit is made based on its "total" likelihood (i.e., sum of two extrinsic information terms and the systematic loglikelihood).

For serial concatenated convolutional codes, a similar iterative (suboptimal) decoder is employed. As in the case of parallel concatenation, there are two component decoders for the inner and the outer code, which can be implemented using a soft-input/soft-output decoder. However, in this case the information exchanged between the component decoders include the information about the parity bits of the outer code along with the systematic

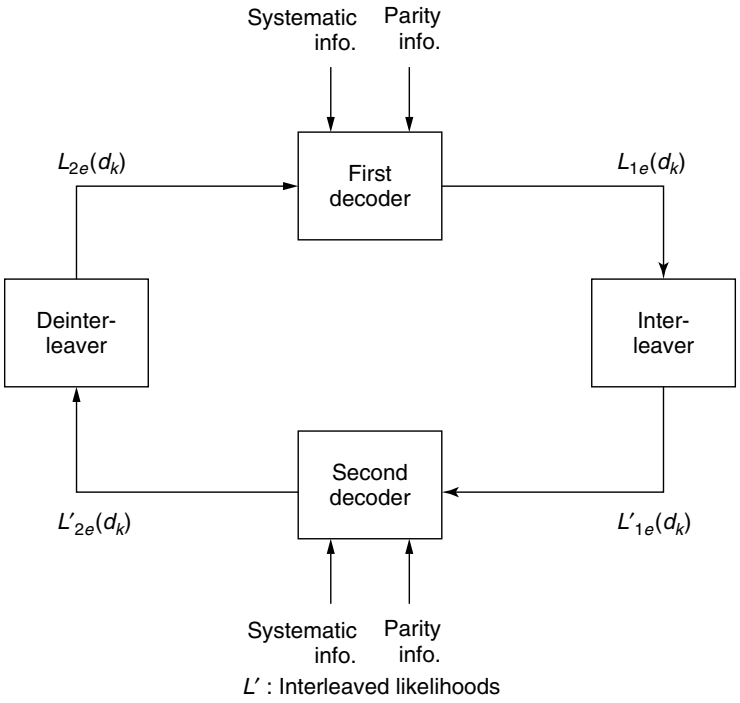


Figure 6. Block diagram of the iterative Turbo decoder for parallel concatenated convolutional codes.

bits. Therefore, the soft-in/soft-out algorithms suitable for component decoders of parallel concatenated codes should be modified accordingly.

3. UNIFORM INTERLEAVER AND INTERLEAVER GAIN

A uniform interleaver devised by Benedetto and Montorsi [6] is a probabilistic device that takes on any one of the possible $N!$ interleavers with equal probability. It is used primarily in analysis of Turbo codes, and can be employed to determine the interleaver gains provided by the parallel and serial concatenation of convolutional codes. The performance predicted by the uniform interleaver is an average over the ensemble of all possible interleavers, and therefore, an interleaver selected at random is expected to achieve this performance. Another interpretation is that there exists an interleaver whose performance is at least as good as the performance predicted by the uniform interleaver.

It is shown [7] that if two recursive convolutional codes are concatenated in parallel, the average number of lowest-weight error events for the overall code (under the assumption of uniform interleaving) is proportional to $1/N$, where N is the interleaver length. If we consider the union bound on the bit error probability, the most important terms of the bound decays with $1/N$ providing an interleaver gain. Hence, the performance of the code improves with the interleaver length (for maximum-likelihood decoding). It is also important to note that if nonrecursive component convolutional codes are employed, there is no interleaver gain. Therefore the use of the recursive codes is critical.

For serially concatenated convolutional codes [5], the outer code may be selected recursive or nonrecursive, whereas the inner code must be selected to be a recursive convolutional code to make sure that an interleaver gain is observed. Then, under the assumption of uniform interleaving, the interleaver gain is given by $N^{-\lceil d_f^o/2 \rceil}$, where d_f^o is the free distance of the outer convolutional code.

It is important to emphasize that concatenation of block codes in parallel or series does not result in any interleaver gain that increases with interleaver size [5,6]. Therefore, using convolutional codes as component codes is advantageous.

4. INTERLEAVER DESIGN FOR PARALLEL CONCATENATED CONVOLUTIONAL CODES

Mathematically, an interleaver of size N is a one-to-one mapping from the set of integers $\{1, 2, \ldots, N\}$ to the same set. Let $\pi(\cdot)$ denote this mapping; then $\pi(i) = j$ means that the ith bit of the original sequence is placed to the jth position after interleaving.

A very simple but effective way of obtaining interleavers for Turbo codes is to select them in a pseudorandom manner. To construct such interleavers, one can easily pick random integers from $\{1, 2, \ldots, N\}$ without replacement. If the ith number picked is j, then $\pi(i) = j$. Pseudorandom interleavers are shown to perform well for Turbo codes. For example, Fig. 7 shows the bit error rate versus signal-to-noise ratio (SNR) obtained by a rate-$\frac{1}{2}$ (obtained by

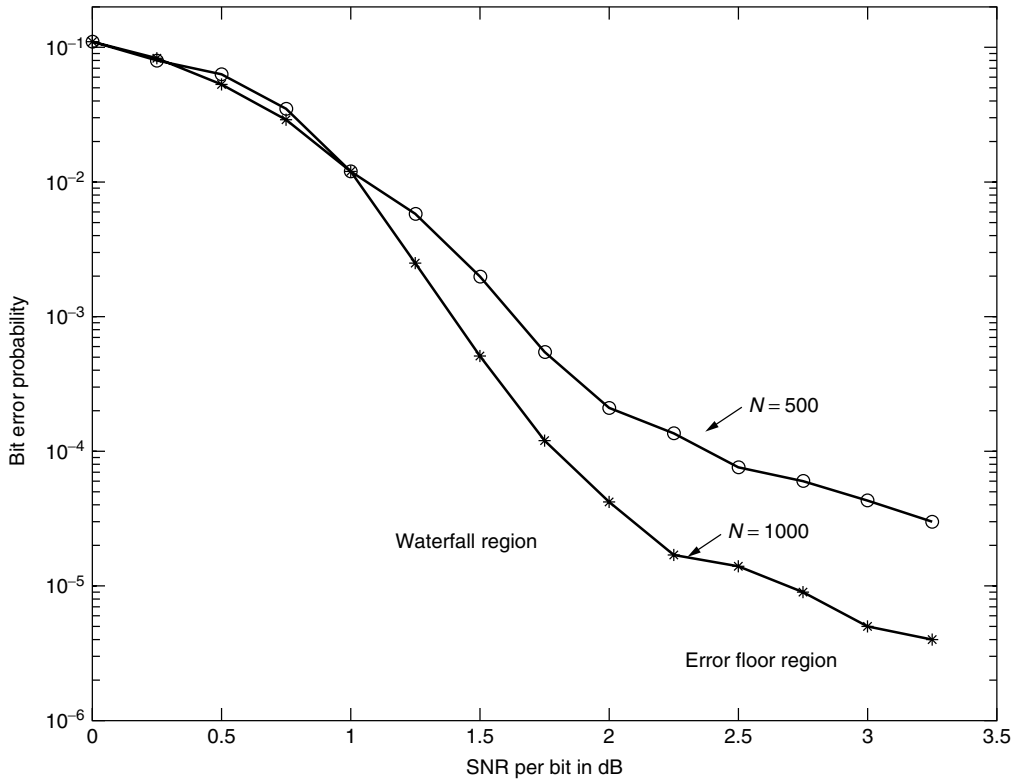


Figure 7. Probability of error for a rate-$\frac{1}{2}$ Turbo code with $\frac{5}{7}$ component codes.

puncturing half of the parity bits) parallel concatenated Turbo code with $\frac{5}{7}$ component codes. However, as we will see shortly, the Turbo code performance can be improved considerably by proper interleaver design techniques.

Before going into the details of interleaver design, it is worthwhile to mention a basic asymptotic result. Turbo codes are linear; therefore the weight distribution determines the distance spectrum of the code. Khandani has shown [8] that the weight of the systematic bitstream, and the two parity bitstreams have Gaussian distribution for large interleaver sizes, and that the correlation coefficients between these three bitstreams are nonnegative and go to zero as $N \to \infty$ for almost any random interleaver. For proper interleaver design, it is desired that these correlation coefficients be as small as possible, and asymptotically this is already satisfied by a random interleaver. Therefore, for large interleaver sizes a randomly selected interleaver is as good as it gets; that is, there are no interleavers that will perform significantly better than the average over the ensemble of all possible interleavers.

However, Khandani's result is only an asymptotic result, and for practical block sizes, it is important to design the interleavers properly. There are two basic approaches to accomplish this goal. One is to attack the weight distribution of the overall Turbo code assuming maximum-likelihood decoding, and the other is to design the interleavers by considering the suboptimal iterative decoding algorithms. We consider these two approaches separately.

4.1. Interleaver Design Based on Distance Spectrum

The recursive convolutional code used to construct the Turbo code is simply a division circuit in a Galois field GF(2). Let us denote its transfer function by $F(D)/G(D)$. Turbo codes are linear block codes; thus let us assume that the all-zero codeword is transmitted over an AWGN channel. The possible error sequences corresponding to this transmission are all the nonzero codewords of the Turbo code. Consider a codeword corresponding to a weight 1 information block. Since the component encoders are selected as recursive convolutional encoders, the parity sequences corresponding to this information sequence will not terminate until the end of the block is reached. With a good selection of the interleaver, if the single 1 occurs toward the end of the input sequence, it will occur toward the beginning of the block for the other component encoder. Therefore, the codewords with information weight 1 will have a large parity weight, hence a large total weight, provided the interleavers that map the bits toward the end of the original sequence too close to the end of the interleaved sequence are avoided.

For larger information weight sequences, the premise is that the interleaver "breaks down" the "bad" sequences. In other words, if the information block results in a lower-weight parity sequence corresponding to one of the encoders, it will have a larger weight parity sequence corresponding to the other one. Therefore, most of the codewords will have large Hamming weights, or the average distance spectrum will be "thinned" [9], and the code will perform well over an AWGN channel.

Among the larger input weight information sequences, the most problematic ones are some of the weight 2 information sequences. Clearly, if the input polynomial is divisible with $G(D)$, then the parity sequence produced by the component code terminates, resulting in a low-weight error event. For example, for the code described in Fig. 4, the feedback polynomial is $1 + D + D^2$, and any input polynomial of the form $D^j(1 + D^{3l})$ corresponds to a "bad" weight 2 error sequence. In general, if the degree of the feedback polynomial is m, then it divides the input polynomials of the form $D^j(1 + D^{kl})$ for some k with $k \leq 2^m - 1$. If the feedback polynomial is primitive, then the lowest-degree polynomial that is divisible by the feedback polynomial is $1 + D^{2^m - 1}$. Therefore, as a side note, we emphasize that in order to reduce the number of "bad" error events, the feedback polynomial is usually selected to be primitive.

The information weight 2 sequences are the most difficult to break down because the percentage of the self-terminating weight 2 sequences is much larger than that for the larger input weight sequences. For example, it is easy to see that with uniform interleaving the number of error events with input sequence weight 2 drops only with $1/N$, whereas the number of higher input weight error events with low overall Hamming weights reduce with $1/N^l$, where $l \geq 2$. Therefore, asymptotically, the performance of the turbo code is determined by the error events of information weight two.

Figure 7 illustrates the bit error rate versus SNR for a typical Turbo code. From the figure two regimes are apparent: the waterfall region and the error floor region. The waterfall region is due to the "thin" distance spectrum of the Turbo code, and the error floor is due to the fact that the minimum distance of the Turbo code is usually small caused by "bad" weight 2 information sequences.

Interleaver design based on the distance spectrum of the overall Turbo code is concerned mainly with lowering the error floor present by attacking the problematic lower information weight sequences, in particular sequences with information weight 2.

4.1.1. Block Interleaver. A simple block interleaver is not suitable for use in Turbo codes as there are certain "bad" information sequences that cannot be broken down. For example, consider the information sequence written rowwise

$$\begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & \cdots & 0 & \mathbf{1} & 0 & 0 & \mathbf{1} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \mathbf{1} & 0 & 0 & \mathbf{1} & 0 & \cdots & 0 \\ \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

and read columnwise. The pattern formed by four 1s shown in boldface cannot be broken down by the block interleaver since both the original sequence and the interleaved version contain two terminating weight 2 error patterns

(for the $\frac{5}{7}$ convolutional code) regardless of the size of the interleaver. Since we have to consider all possible binary N-tuples, there are many information sequences containing the same problematic pattern. Clearly, for any component code, there always exists similar problematic weight 4 information sequences.

Another important information sequence that cannot be broken down is the one with a single 1 at the end of the block.

Although the block interleavers are well suited for coded communications over bursty channels (e.g., fading channels), or breaking up bursts errors due to an inner code decoded using the Viterbi algorithm in a concatenated coding scheme, they are seldom appropriate for Turbo codes.

4.1.2. Reverse Block Interleavers. The problematic information sequence that contains a single 1 at the end can be accommodated easily by a "reverse" block interleaver as proposed by Herzberg [10]. In this case, the information bits are written rowwise and read columnwise; however, the last column is read first. Clearly, this approach does not address problematic weight 4 sequences; nevertheless it is shown to perform well for very short-blocklength codes, and for moderate- and long-blocklength codes at very low probability-of-error values.

Figure 8 shows the performance of the reverse block interleavers together with the block interleavers and the pseudorandom interleavers. The Turbo code in the example is a rate-$\frac{1}{3}$ Turbo code with $\frac{7}{5}$ component codes. This example clearly justifies the superiority of reverse block interleaver for short blocklengths and for larger blocklengths at low bit error probabilities.

4.1.3. *s*-Random Interleaver and Its Variations. As we have discussed in the previous section, the main problem for Turbo codes is the "bad" weight 2 information sequences. If the two 1s resulting in the terminating parity sequence are close to each other in both the original sequence and the interleaved version, both sequences result in low parity weights, and thus the overall codeword has a low Hamming weight. In order to increase the Hamming distance and thus reduce the error floor, an s-random interleaver [11] ensures that any pair of positions that are close to each other in the original sequence are separated by more than a preselected value s (called "spread") in the interleaved version. More precisely, we want

$$\max_{i,j}\{|i-j|, |\pi(i)-\pi(j)|\} > s$$

This condition will ensure that the two 1s do not occur in close proximity in both the original and the interleaved information sequences. Thus, at least, one of the parity weights will be relatively large because of the weight that the first 1 will accumulate until the second 1 is inserted.

An s-random interleaver may be designed as follows [11]. Assume that $\pi(1), \pi(2), \ldots, \pi(n-1)$ are selected. To select $\pi(n)$, we pick a random number $j \in \{1, 2, \ldots, N\} \setminus \{\pi(1), \pi(2), \ldots, \pi(n-1)\}$. If the number selected does not violate the spread constraint, then we let $\pi(n) = j$. If it violates the spread constraint, we reject this number, and select a new one at random. We continue this process until all the integers that describe the interleaver are selected. Obviously, if the desired spread is selected too large, the algorithm may not terminate. However, it is shown by experiments that if $s < \sqrt{N/2}$, then the algorithm converges in reasonable time.

The performance of an s-random interleaver is significantly better that of the pseudorandom interleaver. In particular, the error floor is dramatically reduced since

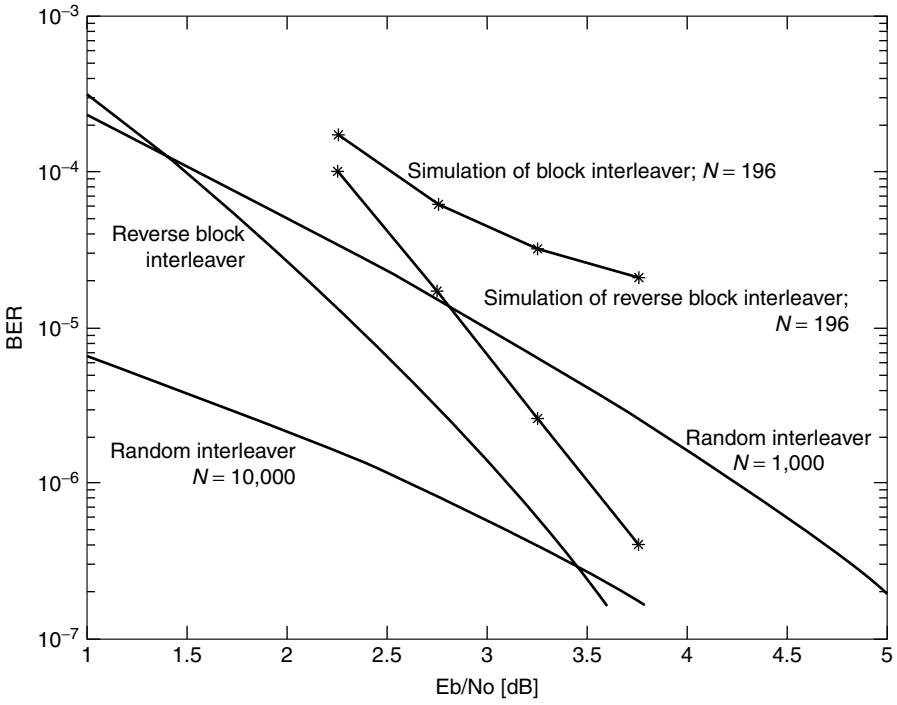


Figure 8. Performance of the reverse block interleaver, block interleaver, and pseudorandom interleaver. (From Herzberg [10], © 1998 IEEE.)

the most problematic weight 2 information sequences that cause the error floor are accommodated. We will give a specific example in the following section and compare it with other interleaver design methods.

We note that block interleavers can be used to achieve a good spread, up to $\sqrt{N}$ as opposed to $\sqrt{N/2}$, which is achieved with the s-random interleavers. However, with block interleaving, the higher-weight sequences become very problematic and the overall interleaver does not perform very well.

In addition to the weight 2 input sequences, we can consider larger input weight sequences to improve the interleaver structure. For instance, Fragouli and Wesel [12] consider the multiple error events and ensure that the interleavers that map the problematic information sequences with multiple error events to another problematic sequence are avoided. It is shown that such interleavers, when concatenated with a higher-order modulation scheme, perform well compared to the s-random interleaver. However, there is no significant improvement when they are used for binary communications over an AWGN channel.

Another approach [13] is to consider "bad" [i.e., divisible by $G(D)$] information sequences of higher weights, and make sure that the interleaver does not map such sequences to another similar sequence. Simulations show that such interleavers perform slightly better than do the s-random interleavers.

"Swap" interleavers are considered in another study [14], where the interleaver is initialized to a block interleaver, and two randomly selected positions are swapped. If the swapping results in violation of the spread constraint, the modification is rejected, and a new pair of positions are selected. After a number of iterations the interleaver is finalized. With swap interleavers, a slight performance improvement over the s-random interleaver is observed (in the order of 0.05 dB).

4.1.4. Iterative Interleaver Growth Algorithms. Daneshgaran and Mondin [15,16] have proposed and studied the idea of iterative interleaver growth algorithms in a systematic way. The main idea is to argue that we should be able to construct good interleavers of size $n+1$ from good interleavers of size n. They exploit the algebraic structure of interleavers and represent any interleaver with an equivalent "transposition vector" of size N that defines the permutation precisely. They prove what is called the "prefix symbol substitution property," which intuitively states that if a new transposition vector is obtained by appending an index to an existing transposition vector, the new vector defines a permutation of one larger size, and it is very closely related to the old one. For details, see Ref. 15.

They then define a cost function that is closely related to the bit error rate that is the ultimate performance measure. Problematic error sequences are identified using the component code structure, and elementary cost functions for each error pattern are defined. The elementary cost functions are directly related to the pairwise error probability that corresponds to the particular error pattern over an AWGN channel. The overall cost function is formed as the sum of these elementary cost functions. Finally, since for the parallel concatenation, both the actual information sequence and its interleaved version are inputs to convolutional codes, the cost function is defined for the inverse interleaver as well. The sum of the two are considered for the overall interleaver design. Extensions of these cost functions for designing interleavers for other channels are straightforward by considering the new appropriate pairwise probability of error expressions.

Now that a cost function is defined, and a method of increasing the interleaver size by one is described, the details of the interleaver growth algorithm can be presented. We start with a small-size interleaver optimized with respect to the cost function defined by an exhaustive search over all interleavers. Then, we extend this interleaver by appending the "best" prefix (with respect to the cost function defined) to the transposition vector that describes the original permutation. Since there are only a limited number of possible prefixes to consider, this step is simple to implement. We repeat this process until an interleaver of a desired size is obtained. We simply are looking for a "greedy" solution to find a good interleaver. Clearly, the process is not optimal; however, experiments show that it results in very good interleavers. We also note that the overall algorithm has only polynomial complexity.

In Fig. 9, we present the performance of the Turbo code with $\frac{5}{7}$ component codes with three different interleavers of size $N = 160$, that is, with the interleaver designed using the algorithm in Ref. 15, the reverse block interleaver and the s-random interleaver. We observe that the interleaver designed using the iterative interleaver growth algorithm performs significantly better than the others. In particular, it is $\sim$0.3 dB better than the reverse block interleaver and $\sim$0.5 dB better than the s-random interleaver at a bit error rate of 10^{-6}. Also shown in the figure is the ML (maximum-likelihood) decoding bound for a uniform interleaver (i.e., the average performance of all the possible interleavers). We observe that the interleavers designed outperform the average significantly, particularly, in the error floor region.

4.1.5. Other Deterministic Interleavers. Several other deterministic interleaver design algorithms are proposed in the literature with varying levels of success [e.g., 17–20].

4.2. Interleaver Design Based on Iterative Decoder Performance

The previous interleaver design algorithms described improve the distance spectrum of the turbo code by avoiding or reducing the number of problematic low-input-weight error sequences. They consider the maximum likelihood decoding performance and inherently assume that the performance of the suboptimal iterative decoding employed will be close to the optimal decoder. Another approach to interleaver design is to consider the performance of the iterative decoder as proposed by Hokfelt et al. [21], and design interleavers according to their suitability for iterative decoding.

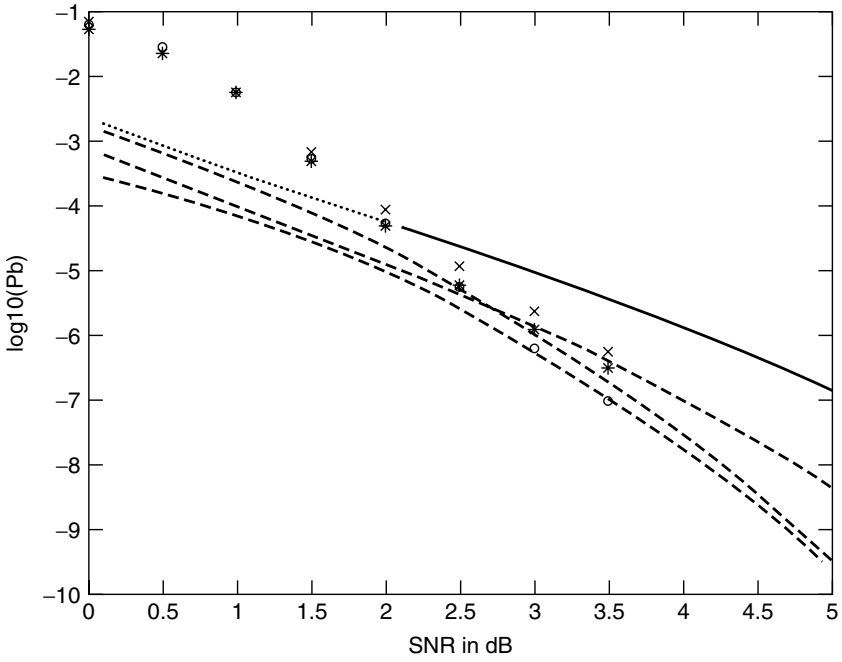


Figure 9. Performance of the Daneshgaran–Mondin (DM) interleaver (simulated points shown by "o"), the s-random (SR) interleaver (simulated points shown by "*"), and the reverse block (RB) interleaver with $N = 169$ (simulated points shown by "x"). The asymptotic BER curves are shown by dashed lines; at 5 dB, the uppermost curve is for SR, the middle one is for RB, and the lowermost one is for DM. The union bound for the uniform interleaver is shown by the solid line. (From Daneshgaran and Mondin [15], © 1999 IEEE.)

Consider the iterative decoding algorithm summarized in Section 2.3 (Fig. 6). Hokfelt et al. [21] showed that the extrinsic information produced at the output of the component decoders correlates with the systematic input. Let us denote the jth systematic information input to the lth component decoder with $x_{j,l}$, $j = 1, 2, \ldots, N$, $l = 1, 2$. The extrinsic information produced by the lth decoder about the ith bit is denoted by $L_{le,i}$.

Empirical results show that after the first decoding step (first half of the iteration), the correlation coefficient between the ith extrinsic information produced by the decoder and the jth systematic input can be approximated by [21]

$$\rho^{(1)}_{L_{e,i},x_j} = \begin{cases} ae^{-c|i-j|} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $i, j = 1, 2, \ldots, N$, and the subscript l denoting the component decoder index is suppressed. The constants a and c depend on the particular component code selected, and can be computed using simulations.

Similarly, after the second iteration step, these correlations can be approximated by

$$\rho^{(2)}_{L_{e,i},x_j} = \frac{1}{2}ae^{-c|\pi(j)-i|} + \frac{1}{2}\sum_{\substack{m=1 \\ m\neq i}}^{N} a^2 e^{-c(|\pi^{-1}(m)-j|+|i-m|)}$$

where π^{-1} denotes the inverse of the interleaver.

The nonzero correlation between the systematic input and the extrinsic information to the component decoders deteriorates the performance of the iterative decoder. With this interleaver design approach, the objective of the interleaver is to make the nearby extrinsic inputs as uncorrelated to each other as possible, or to ensure that the extrinsic information at the output of the second decoder is uniformly correlated with the systematic information. The interleaver is designed as follows [21]. Starting with $i = 1$, the ith entry of the interleaver is selected as

$$\pi(i) = \underset{j}{\operatorname{argmin}} \sum_{m} e^{-c(|\pi(m)-j|+|i-m|)}$$

where the summation is performed over predefined interleaver elements, and the minimization is over all permissible positions to choose from.

We note that this interleaver design technique tries to minimize the number of short cycle error events that deteriorate the iterative decoder performance due to the excessive correlation of the extrinsic information. We also note that the interleaver designed in this manner competes very well with the s-random interleaver. In particular, the iterative decoding algorithm converges faster, and BER performance improves in the order of 0.1 dB.

Sadjadpour et al. [22] combined the two basic interleaver design approaches to improve the interleaver structure; that is, both the distance spectrum of the code and the correlation of the extrinsic information are considered. In addition to the iterative decoding suitability criterion of Hokfelt et al. [21], Sadjadpour et al. [22] propose another related criterion that tries to ensure that the extrinsic information is uniformly correlated with the input, and that the "power" of the correlation coefficients (sum of squares) is minimized. Then, the interleaver is designed in two steps: (1) an s-random interleaver is selected and then (2) all the low-weight error sequences that result in a lower weight than a predetermined value — typically, problematic weight 2 information sequences — are considered. On the basis of these problematic sequences, the interleaver is updated by using the iterative decoding suitability criterion. This operation is continued until all the problematic sequences are considered. Finally, step 2 is repeated with the new

interleaver until there are no low-weight error patterns with a minimum distance less than the preselected value. Clearly, this procedure may not converge if the desired minimum distance of the overall code is selected to be too large.

Examples presented by Sadjadpour et al. [22] indicate that this two-step s-random interleaver design technique has a better performance than that of the s-random interleaver, typically in the error floor region.

5. INTERLEAVER DESIGN FOR SERIALLY CONCATENATED CONVOLUTIONAL CODES

Although there are many results available for the interleaver design for parallel concatenated convolutional codes, the literature on interleaver design for serially concatenated convolutional codes is very scarce. The main reason for this fact is that for serial concatenation the interleaver gain is much larger than that of parallel concatenation. To reiterate, for serial concatenation, under uniform interleaving, the probability of error decays with $N^{-\lceil d_f^o/2 \rceil}$ asymptotically, whereas for parallel concatenation this is only $1/N$. As a result, the error floor observed for parallel concatenation is much less of an issue for serial concatenation.

Clearly, there are certain interleavers that need to be avoided. For instance, the identity operation (no interleaving) does not result in any interleaver gain and is not useful. Also, we need to avoid interleavers that map the last few bits of the sequence to the end of the frame, since for input sequences that contain 1s only at the end of the block, the minimum distance of the overall concatenated code will be small. However, a pseudorandom interleaver selected will perform well; that is, the error floor will be much lower than the one observed in the parallel concatenated Turbo codes. If the interleaver is further selected to be an s-random interleaver, it should perform even better as the short error events of the outer convolutional code will be distributed across the entire sequence at the input of the inner encoder, and likely will result in a large Hamming weight at the output [23].

For parallel concatenation, weight 2 error patterns are the dominant events that warrant special attention regardless of the component convolutional codes used. However, for serial concatenation, the problematic error sequences are highly dependent on the component codes, and the interleaver design should be performed specifically for the component codes.

At the time of this writing, only two papers deal with interleaver design for serially concatenated convolutional codes [23,24]. Daneshgaran et al. [23] formulate the interleaver design algorithm as an optimization problem where the cost function to be minimized considers important error events of the outer component code, and depends on the specific interleaver and the inner component code. If one has a good interleaver of size n, then this interleaver can be grown to size $n+1$ in such a way that the cost function is minimized. This process is continued until an interleaver of desired size is obtained. The interleaver growth algorithm has polynomial complexity, and therefore is easy to implement. The authors present an example design and show that much larger minimum distances can be obtained compared to codes that employ random or s-random interleaving. Therefore, the error floor of the code designed in this manner is much smaller.

Another natural approach is to design the interleaver based on the suboptimal iterative decoding algorithm. In particular, as in the case of parallel concatenation, one can select the interleaver to make sure that the short cycles are avoided. However, at this point, this approach has not been explored.

6. CONCLUSIONS

Interleavers play a major role in the construction of both parallel and serially concatenated convolutional (Turbo) codes as they provide what is called the "interleaver gain." In this article, we have discussed the role of the interleavers in Turbo codes in detail, and reviewed some of the important results. We also have identified the design of interleavers as an important issue, and summarized several promising interleaver design algorithms developed in the literature.

BIOGRAPHY

Tolga M. Duman received the B.S. degree from Bilkent University, Turkey, in 1993, and M.S. and Ph.D. degrees from Northeastern University, Boston, in 1995 and 1998, respectively, all in electrical engineering. He joined the Electrical Engineering faculty of Arizona State University as an Assistant Professor in August 1998. Dr. Duman's current research interests are in digital communications, wireless and mobile communications, channel coding, Turbo codes, coding for recording channels, and coding for wireless communications.

Dr. Duman is the recipient of the National Science Foundation CAREER Award, IEEE Third Millennium medal, and IEEE Benelux Joint Chapter best-paper award (1999). He is a member of the IEEE Information Theory and Communication Societies.

BIBLIOGRAPHY

1. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice Hall, 1995.
2. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* 1–10 (Jan. 1948).
3. John G. Proakis, *Digital Communications*, McGraw-Hill, New York, 2001.
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. IEEE Int. Conf. Communications* (*ICC*), 1993, pp. 1064–1070.
5. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design and iterative decoding, *IEEE Trans. Inform. Theory* 909–929 (May 1998).

6. S. Benedetto and G. Montorsi, Unveiling Turbo codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* 409–428 (March 1996).
7. S. Benedetto and G. Montorsi, Design of parallel concatenated convolutional codes, *IEEE Trans. Commun.* 591–600 (May 1996).
8. A. K. Khandani, Optimization of the interleaver structure for Turbo codes, *Proc. Canadian Workshop on Information Theory*, June 1997, pp. 25–28.
9. L. C. Perez, J. Seghers, and D. J. Costello, Jr., A distance spectrum interpretation of Turbo codes, *IEEE Trans. Inform. Theory* 1698–1709 (Nov. 1996).
10. H. Herzberg, Multilevel Turbo coding with short interleavers, *IEEE J. Select. Areas Commun.* 303–309 (Feb. 1998).
11. S. Dolinar and D. Divsalar, *Weight Distributions for Turbo Codes Using Random and Nonrandom Permutations*, TDA Progress Report 42–122, JPL, Aug. 1995.
12. C. Fragouli and R. D. Wesel, Semi-random interleaver design criteria, *Proc. IEEE Global Communications Conf.* (*GLOBECOM*), 1999, pp. 2352–2356.
13. F. Said, A. H. Aghvami, and W. G. Chambers, Improving random interleaver for Turbo codes, *Electron. Lett.* **35**(25): 2194–2195 (Dec. 1999).
14. B. G. Lee, S. J. Bae, S. G. Kang, and E. K. Joo, Design of swap interleaver for Turbo codes, *Electron. Lett.* 1939–1940 (Oct. 1999).
15. F. Daneshgaran and M. Mondin, Design of interleavers for Turbo codes: Iterative interleaver growth algorithms of polynomial complexity, *IEEE Trans. Inform. Theory* **45**(6): 1845–1859 (Sept. 1999).
16. F. Daneshgaran and M. Mondin, Optimized Turbo codes for delay constrained applications, *IEEE Trans. Inform. Theory* **48**(1): 293–305 (Jan. 2002).
17. D. Wang and H. Kobayashi, On design of interleavers with practical size for Turbo codes, *Proc. IEEE Int. Conf. Communications* (*ICC*), Oct. 2000, pp. 618–622.
18. M. Z. Wang, A. Sheikh, and F. Qi, Interleaver design for short Turbo codes, *Proc. IEEE Global Communications Conf.* (*GLOBECOM*), Dec. 1999, pp. 894–898.
19. A. K. Khandani, Group structure of Turbo codes with applications to the interleaver design, *Proc. IEEE Int. Symp. Information Theory* (*ISIT*), Aug. 1998, p. 421.
20. S. Crozier, J. Lodge, P. Guinand, and A. Hunt, Performance of Turbo codes with relative prime and golden interleaving strategies, *Proc. 6th Int. Mobile Satellite Conf.* (*IMSC99*), June 1999, pp. 268–275.
21. J. Hokfelt, O. Edfors, and T. Maseng, A Turbo code interleaver design criterion based on the performance of iterative decoding, *IEEE Commun. Lett.* 52–54 (Feb. 2001).
22. H. R. Sadjadpour, N. J. A. Sloane, M. Salehi, and G. Nebe, Interleaver design for Turbo codes, *IEEE J. Select. Areas Commun.* **19**(5): 831–837 (May 2001).
23. F. Daneshgaran, M. Laddomada, and M. Mondin, Interleaver design for serially concatenated convolutional codes: Theory and application, preprint, 2002.
24. R. Jordan, S. Host, and R. Johannesson, On interleaver design for serially concatenated convolutional codes, *Proc. IEEE Int. Symp. Information Theory* (*ISIT*), 2001, p. 212.

INTERNET SECURITY

BÜLENT YENER*
Rensselaer Polytechnic University
Troy, New York

1. INTRODUCTION

As the Internet utilized as a new commercial infrastructure, meeting security requirements of diverse applications becomes imminent. Furthermore, the Web and browsers bring have brought the Internet to homes of average people, creating not only a surge in use of the Internet but also a risk to their privacy.

Internet security aims to ensure *confidentiality, authentication, integrity*, and *nonreputiation* of the "information" carried over a collection of interconnected, heterogeneous networks via messages. Confidentiality or privacy prevents unauthorized parties from accessing the message. Authentication requires that source of a message has correct and verifiable identity. Integrity protection of information ensures that unauthorized parties cannot modify the information. Nonreputiation of information requires that the sender and receiver of the information cannot deny the transmission of the message. In general, the security attacks can be grouped into several classes:

1. *Interception* attacks, which are directed at the confidentiality of information by unauthorized access. It is a passive attack where the adversary simply observes the communication channel without modifying the information. Eavesdropping on a communication channel is a typical example.
2. *Modification* attacks, which violate the integrity of information. It is an active attack in which adversary changes the content of information. Man-in-the-middle attacks are typical examples.
3. *Fabrication* attacks, in which the adversary generates and inserts malicious information to the system. This is also an active attack and it violates the authenticity of information.
4. *Interruption*, which is also an active attack that targets the availability of the system. An example is malicious jamming in a wireless network to generate intentional interference.

1.1. Cryptography

Cryptography provides the essential techniques and algorithms to keep information secure [1,2]. Confidentiality is done by encryption and decryption. Authentication is ensured with digital certificates while integrity is protected with hash functions. Nonreputiation of messages is ensured with digital signatures.

* Department of Computer Science at Rensselaer Polytechnic Institute. This article was written in part while the author was visiting Bogazici University Department of Computer Science, Istanbul, Turkey.

The roots of cryptographic research can be traced to William F. Friedman's report *Index of Coincidence and Its Applications* [3], and Edward H. Hebern's rotor machine [4] in 1918. In 1948 Claude Shannon presented his work on the communication theory of secrecy systems in the *Bell System Technical Journal* [5]. In early 1970s work by Horst Feistel from IBM Watson Laboratory led the first U.S. Data Encryption Standard (DES) [6]. In DES both parties must share the same secret key of 56 bits before communication begins. However, Michael Weiner showed that exhaustive search can be used to find any DES key [7]. More recently, the National Institute of Standards and Technology (NIST) has selected the Advanced Encryption Standard (AES), the successor to the venerable DES. AES was invented by Joan Daemen and Vincent Rijmen [8].

1.1.1. Confidentiality. Encryption is a function E that takes plaintext message M as input and produces the encrypted ciphertext C of M: $E(M) = C$. Decryption is the function for the reverse process: $D(C) = M$. Note that $D(E(M)) = M$. In modern cryptography, a *key* K is used for encryption and decryption so that $D_K(E_K(M)) = M$. Cryptographic algorithms, based on using a single key, (i.e., both encryption and decryption are done by the same key) are called *symmetric ciphers* and they have two drawbacks: (1) an arrangement must be made to ensure that two parties have the same key prior to communicate with each other and (2) the number of keys required for a complete communication mash for an n party network is $O(n^2)$. Although a trusted third party such as a *key distribution center* (KDC) can be used to circumvent these two problems, it requires that KDC must be available in real-time to initiate a communication.

Public key cryptography proposed by Whitfield Diffie and Martin Hellman in 1975 [9] is based on *asymmetric ciphers*. In such systems, the encryption key K_1 is different from the decryption key K_2 so that $D_{K_2}(E_{K_1}(M)) = M$. The encryption key K_1 is called the *public key*, and it is not secret. The second key K_2 is called the *private key*, and it is kept confidential. The best known public key crypto system, proposed by Ronald Rivest, Adi Shamir, and Leonard Adleman (RSA) [10]. Although the public key ciphers reduce the number of keys to $O(n)$, they also suffer two problems: (1) key size is much larger than in symmetric systems and (2) the encryption and decryption are much slower. These two issues become problematic in a bandwidth processing-constrained environments such as wireless networks. Thus, the main use of public key systems is limited to distribution of symmetric cipher keys.

1.1.2. Integrity. Cryptographic solutions for integrity protection are based on *one-way hash functions*. A hash function is a one-way function that is easy to compute but significantly harder to compute in reverse (e.g., $a^x \bmod n$). It takes a variable-length input and produces a fixed-length hash value (also known as "message digest"). In general it works as follows. The sender computes the hash value of the message, encrypts it using the receiver's public key, and appends it to the message. The receiver decrypts the message digest using his/her private key and then computes the hash value of the message. If the computed hash value is the same as the decrypted one, the message integrity is considered to be preserved during transmission. However, this is not an absolute guarantee since a hash collision is possible (i.e., a modified or fabricated message may have the same hash value as the original. Furthermore, the hash function is public so that the attacker can intercept a message, modify it, and compute a new hash value for the modified message. Thus, it would be a good idea to encrypt the message as well or use a message authentication code (MAC), which is a one-way function with a key.

1.1.3. Nonreputiation. In order to prove the source of a message, one-way functions called *digital signatures* can be used in conjunction with public key cryptography. To stamp a message with its digital signature, the sender encrypts the message digest with its private key. The receiver first decrypts the message using the sender's public key and then computes the message digest to compare it to the one that arrives with the message.

1.1.4. Authentication. To prevent an attacker from impersonating a legitimate party, *digital certificates* are used. At the beginning of a secure Internet session, sender transmits its digital certificate to have his/her identity to be verified. Digital certificates may be issued in a hierarchical way for distributed administration. A digital certificate may follow the ITU standard X.509 [11,12] and is issued by a *certificate authority* (CA) as a part of public key infrastructure (PKI). The certificate contains the sender's public key, the certificate serial number and validity period, and the sender's and the CA's domain names. CA must ensure integrity of the issued certificate; thus it may encrypt the hash value of it using its private key and append it to the certificate.

1.2. Cryptography and Security

Although cryptography provides the building blocks, there is much to consider for Internet security. First, the rapid growth of the Internet increases its heterogeneity and complexity. A communication channel between a pair of users may pass through different network elements running diverse protocols. Most of these protocols are designed with performance considerations and carry design and implementation holes from security point of view. For example, consider the *Anonymous File Transfer Protocol* (FTP) [13], which provides one of the most important services in the Internet for distribution of information. There are several problems with FTP and its variants. For example, the *ftpd* deamon runs with superuser privileges for password and login processing. Thus, leaving a sensitive file such as the password file in the anonymous FTP site will be a serious security gap. Another example is the *Transport Control Protocol* (TCP) [14], which provides a connection between a pair of users. In TCP each connection is identified with a 4-tuple: <*source* (*local*) *host IP address, local port number, destination* (*remote*) *host IP address ID, remote port number*>. Since the same 4-tuple can be reused,

the *sequence numbers* are used to detect the lingering packets from the previous uses of the tuple. There is a potential threat here since an attacker can "guess" the initial sequence number and convince the remote host that it is communicating with a trusted host. This is known as a *sequence number attack* [15]. A remedy for this attack would be to hide the target host behind a dedicated gateway called a *firewall* to prevent direct connections [16].

Also, the network software is hierarchical and layered. At each layer a different protocol is in charge and interacts with the protocols in the adjacent layers. In the following sections we will examine layer-specific security concerns.

2. LINK-LAYER SECURITY

Link-layer security issues in *local-area networks* (LANs) and *wide-area networks* (WANs) are fundamentally different. In a WAN link-layer security requires that end point of each link is secure and equipped with encryption devices. Although institutions such as the military have been using link-layer encryption, it is not feasible in the Internet.

In a LAN hosts share the same communication medium that has (in general) a broadcast nature (i.e., transmission from one node can be received by all others on the same LAN). Thus eavesdropping is easy and, to ensure confidentiality encryption, is required. For example, in a LAN it is better to use the SSH protocol [17] instead of Telnet to avoid compromising passwords.

2.1. Access Control

Typically, a LAN connects hosts who are in the same security or administrative domain. While allowing legitimate user accessing to a LAN from outside (via a dialup modem, or a DSL, or a cable modem), it is crucial to prevent unauthorized access. Firewalls are dedicated gateways used for access control and can be grouped into three classes [16]: packet filter gateways, circuit-level gateways, and application-layer gateways. Packet filters operate by selectively dropping packets based on source address, destination address, or port number. In a firewall the security policies can be specified in a table that contains the filtering rules to which the incoming or outgoing packets are subject. For example, all outgoing mail traffic can be permitted to pass through a firewall while Telnet requests from a list of hosts can be dropped. Filtering rules must be managed carefully to prevent loopholes in a firewall. In case of multiple firewalls managed by the same security domain, it is crucial to eliminate inconsistencies between the rules.

Application- and circuit-level gateways are firewalls that can secure the usage of a particular application by screening the commands. Logically it resides in the middle of a protocol exchange and ensures that only valid commands are sent. For example, it may monitor a FTP session to ensure that only a specific file is accessed with read-only permission.

3. NETWORK-LAYER SECURITY

The Internet is composed of many independent management domains called *autonomous systems* (ASes). Internet routing algorithms within an AS and among ASes are different. Most important intra-AS routing (interior routing) and inter-AS (exterior routing) protocols are the Open Shortest Paths First (OSPF), and Border Gateway Protocol (BGP), respectively. However, the common theme in these protocols is the exchange of routing information to converge in a stable routing state. However, because of the lack of scalable authenticity check, routing information exchanged between the peers is subject to attacks. The attacker can eavesdrop, modify, and reinject the exchanged messages. Most of these attacks can be addressed by deployment of public key infrastructures (PKI) and certificates for authentication and validation of messages. For example, Kent et al. [18] discuss how to secure BGP protocol using PKI with X.509 certificates, and IPsec protocol suite. The solution proposes to use a new BGP path attribute to ensure the authenticity and integrity of BGP messages and validate the source of UPDATE messages. However, if a legitimate router is compromised, then such cryptographic mechanisms cannot be sufficient and the security problem degenerates to the Byzantine agreement problem [19] in distributed computing.

Next we discuss the IPsec protocols for securing IP-based intranets and then review the security issues in ATM networks.

3.1. IP Security:IPsec

The suite of IPsec protocols are designed to provide security for Internet Protocol version 4 (IPv4) and version 6 (IPv6) [20]. IPsec offers access control, connectionless integrity, source authentication, and confidentiality. IPsec defines two headers that are placed after the IP header and before the header of layer 4 protocols (i.e., TCP or UDP). These headers are used in two traffic security protocols: the *authentication header* (AH) and the *encapsulating security payload* (ESP). AH is recommended when confidentiality is not required, while ESP provides optional encryption. They both ensure integrity and authentication using tools such as keyed hash functions. Both AH and ESP use a simplex connection called a *security association* (SA). An SA is uniquely identified by a triple that contains a security parameter index (SPI), a destination IP address, and an identifier for the security protocol (i.e., AH or ESP). The negotiation of security association between two entities and exchange of keys can be done by using the Internet Key Exchange (IKE) protocol [22]. Conceptually, an SA is a virtual tunnel based on encapsulation. Two types of SAs are defined in the standard: *transport mode*, and *tunnel mode*. The former is a security association between two hosts, while the latter is established between network elements. Thus, in the transport mode the security protocol header comes right after the IP header and encapsulates any higher-level protocols. In the tunnel mode there is an outer header and an inner IP header. The *outer* header specifies the next hop, while the *inner* header indicates the final destination. The security protocol header in the

tunnel mode is inserted after the outer IP header and before the inner one, thus protecting the inner header.

There are successful attacks on IPsec in spite of the secure ciphers used by the protocol. For example, consider the cut-and-paste attack by Bellovin [21]. In this attack, an encrypted ciphertext from a packet carrying sensitive (targeted) information is cut and pasted into the ciphertext of another packet. The objective is to trick the receiver to decrypt the modified ciphertext and reveal the information.

3.2. ATM Security

Asynchronous transfer mode (ATM) technology is based on establishing switched or permanent virtual circuits (SVCs, PVCs) to transmit fixed-size (53-byte) cells. There are no standards for ATM security, and work is in progress at the ATM Forum [23]. Next we review some of the security threats inherent in the architecture. All the cells carrying the same VPI/VCI (virtual path identifier, virtual connection identifier, respectively) are carried on the same virtual channel. Thus, eavesdropping and integrity violation attacks can be mounted to *all* the cells of a connection from a single point. In particular the cells carrying signaling information can be used to identify communicating parties. For example, capturing CONNECT or CALL PROCEEDING messages during signaling will reveal the VPI/VCIs assigned by the network to a particular connection. Flooding network with SET UP requests can be used to achieve denial-of-service attacks. Management cells can be abused to disrupt or disconnect legitimate connections. For example, by tampering with AIS/FERF cells, the attacker can cause a connection to be terminated.

3.2.1. IP over ATM. ATM networks has been deployed in high-speed backbone as the switching plane for IP traffic using IP over ATM protocols. The IP over ATM suite brings security concerns in ATM networks, many of which are similar to those in IP networks; however, their remedies are more difficult. For example, firewalls and packet filters used for access control in IP networks will require termination of ATM connection, inducing large delays and overhead. Authentication between ATMARP (ATM Address Resolution Protocol) server and hosts is a must for preventing various threads, including *spoofing, denial-of-service*, and *man-in-the-middle* attacks. For example, it is possible to send spoofed IP packets over an ATM connection, if the ATM address of an ATMARP server is known. The attacker can first establish a virtual connection to the server and then use the IP address of the victim to spoof the packets on this connection. Since the server will reply back to the attacker using the same connection the victim may not even know the attack. Similarly, the attacker can use the IP addresses of victims to register them with the ATMARP server. Since each IP address can be used only once, the victims will be denied service.

4. TRANSPORT-LAYER SECURITY

The Secure Sockets Layer (SSL) protocol [24] was developed to ensure secure communication between the Internet browsers and servers by Netscape Corporation. Protocols such as HTTP run over SSL to provide secure connections. The Transport Layer Security (TLS) protocol [25] is expected to become the standard for secure client/server applications over the Internet. TLS v.1.0 is based on SSL v.3.0 and considered to be SSL v.3.1. Both SSL and TLS provide encryption, authentication, and integrity protection over a public network. They are composed of two subprotocols (layers): the Record Protocol and the Handshake Protocol. The *Record Protocol* is at the lowest layer and resides above a reliable transport layer protocol such as TCP. It provides encryption using symmetric cryptography (e.g., DES), and message integrity check using a keyed MAC. The *Handshake Protocol* is used to agree on cryptographic algorithms, to establish a set of keys to be used by the ciphers, and to authenticate the client. The Handshake Protocol starts with a *ClientHello* message sent to a server by a client. This message contains a random number, version information, encryption algorithms that the client supports, and a session ID. The server sends back a *ServerHello* message that also contains random data, session ID, and indicates selected cipher. In addition, the server sends a *Certificate* message that contains the server's RSA public key in an X.509 [11,12] certificate. The client verifies the server's public key, generates a 48-byte random number called the *premaster key*, encrypts it using server's public key, and sends it to the server. The client also computes a *master secret* and uses the master key to derive at a symmetric *session key*. The server performs similar operations to compute a master key and a symmetric key. After the keys are installed to the record layer, the handshake is completed. Although SSL and TLS are similar there are also important differences between them. For example, in SSL each message is transmitted with a new socket while in TLS multiple messages can be transmitted over the same socket. Security of the SSL protocol is well examined and reported to be sound, although there are some easy-to-fix problems [26]. For example, unprotected data structures (e.g., server key exchange) can be exploited to perform cryptographic attacks (e.g., *ciphersuite rollback attack* [26]).

4.1. Multicast Security

Multicasting is a group communication with single or multiple sources and multiple receivers. It has considerably more challenging security concerns than does a single source–destination communication:

1. Message authentication and confidentiality in a multicast group requires efficient key management protocols. Establishing a different key between each pair of multicast members is not scalable. Thus, most of the solutions focus on a *shared* key [27–29]. However, a single-key approach is not sufficient to authenticate the identity of a sender since it is shared by all the members. Signing each message using a public key scheme is a costly solution; thus MAC-based solutions have been proposed [30].
2. The membership dynamics (i.e., joining and leaving a multicast group) requires efficient key revocation

algorithms. In particular deletion of a user from the multicast group must not reset all the keys. The solution is based on assigning multiple keys to each member and organizing the key allocation into a data structure that is easy to update [32–34]. For example, the Wallner et al. [33] key allocation scheme uses a (binary) tree structure. The group members are the leaves, and each intermediate node represents a distinct key. Each user will receive all the keys on the path to the root of the tree, and the root contains the shared group key. A group controller manages the data structure for delete and insert operations. Thus, in the key-based scheme each user gets $\log(n+1)$ keys and deletion of a user cost $2\log n - 1$ key encryptions.

5. APPLICATION-LEVEL SECURITY: KERBEROS

Kerberos is an authentication service that allows users and services to authenticate themselves to each other [31]. It is typically used when a user on a network requests a network service, and the server needs to ensure that the user is a legitimate one. For example, it enables users to log in to remote computers over the network without exposing their passwords to network packet-sniffing programs. User authentication is based on a "ticket" issued by the Kerberos key distribution center (KDC), which has two modules: authentication server (AS) and ticket-granting server (TGS). Both the user and the server are required to have keys registered with the AS. The user's key is derived from a password that is seen by only the local machine; the server key is selected randomly. The authentication between a user u and server S has the following steps:

1. The user u sends a message to the AS specifying the server S.
2. The AS produces two copies of a key called the *session key* to be used between u and S. AS encrypts one of the session keys and the identity S of the server using the user's key. Similarly, it encrypts the other session key and identity of the user with the server key. It sends both of the encrypted messages, called the "tickets" (say, m_1 and m_2, respectively) to u.
3. u can decrypt m_1 with its own key, extracting the session key and the identity of the server S. However, u cannot decrypt m_2 instead it timestamps a new message m_3 (called the *authenticator*), encrypts it with the session key and sends both m_2 and m_3 to S.
4. S decrypts m_2 with its own key to obtain the session key and the identity of user u. It then decrypts m_3 with the session key to extract the timestamp in order to authenticate the identity of the user u.

Following Step 4, all the communication between u and S will be done using the session key. However, in order to avoid performing all the steps above for each request, the TGS module in KDC issues a special ticket called the *ticket-granting ticket* (TGT). TGT behaves like a temporary password, with a lifetime of several hours only, and all other tickets are obtained using TGT.

6. WIRELESS SECURITY

Security in wireless networks is a challenging problem — the bandwidth and power limitations encourage the use of weaker cryptographic tools or keys with smaller sizes; also, the lack of point-to-point links makes it more difficult to protect the communication.

Elliptic curve crypto (ECC) systems [35] provide a remedy to some of these problems. ECC is based on discrete logarithm problem defined over the points on an elliptic curve. It is considered to be harder than the factorization problem and can provide works with much smaller key size than can other public key crypto systems [36]. Smaller key size reduces the processing overhead, and smaller digital signatures save on the bandwidth consumption.

6.1. Wireless LAN (WLAN)

WLANs use RF technology to receive and transmit data in a local-area network domain. In contrast with a wired LAN, a WLAN offers mobility and flexibility due to lack of any fixed topology. IEEE 802.11 is the most widely adapted standard for WLANs and it operates in the 2.4–2.48-GHz band.

There are several security vulnerabilities of a WLAN due to its nature: (1) any node within the transmission range of the source can eavesdrop easily, (2) unsuccessful attempts to access to a WLAN may be interpreted as a high *bit error rate* (BER) — this misinterpretation can be used to conceal an intruder's unauthorized access attack to a WLAN, and (3) the transmission medium is "shared" among the users. Thus, intentional interference (called "jamming") can be produced in a WLAN for denial of service attacks.

The *spread-spectrum* transmission technology helps countermeasure some of these problems in the WLANs. In spread spectrum, a signal is spread over the channel using two different techniques: (1) the frequency-hopping spread spectrum (FHSS), and (2) the direct-sequence spread spectrum (DSSS). An attacker must know the hopping pattern in FHSS or the codewords in DSSS to tune into the right frequency for eavesdropping. (Ironically these parameters are made public in the IEEE 802.11 standard.) Additional help comes from sophisticated network interface cards (NICs) of IEEE 802.11b devices. These cards can be equipped with a unique public and private key pair, in addition to their unique address, to prevent unauthorized access to a WLAN.

IEEE 802.11 standard provides a security capability called *wired equivalent privacy* (WEP). In WEP there is a secret 40-bit or a 128-bit key that is shared between a wireless node and an access point. Communication between a wireless station and its access point can be encrypted using the key and RSA's RC4 encryption algorithm. RC4 is a stream cipher with a variable key size and uses an *initialization vector* (IV). IV is used to produce different ciphertexts for identical plaintexts by initializing the shift registers with random bits. IV does not need to be secret but it should be unique for each transmission. However, IEEE 802.11 does not enforce the uniqueness of IV. Thus, one potential problem with the WEP is the reuse

of IV, which may be exploited for cryptanalyzing and for fabricating new messages [37].

6.2. Wireless Transport Layer (WTLS)

The Wireless Transport Layer Security (WTLS) [38] protocol provides authentication, privacy and integrity for the Wireless Application Protocol (WAP) [39]. The WTLS is based on TLS v.1.0 and takes into account the characteristics of wireless world (e.g., low bandwidth, limited processing and power capacity, and connectionless datagram service). WTLS supports a rich set of cryptographic algorithms. Confidentiality is provided by using block ciphers such as DES CBC, integrity is ensured by SHA-1 [41] and MD5 [40] MAC algorithms, and the authentication is checked by RSA and Diffie–Hellman-based key exchange algorithms. WTLS does not contain any serious security problems to force an architectural change. Nevertheless there are several weak points of the protocol: (1) the computation of initialization vector (IV) is not a secret, (2) some fields in the data structures used by the protocol are not protected (one example is the sequence numbering, which enables an attacker to generate replay attacks), and (3) the key size should be at least 56 bits since 40-bit keys are not sufficient.

7. CONCLUSIONS

Heterogeneity of the Internet requires a skillful integration of the cryptographic building blocks with protocols for ensuring end-to-end security. Thus, deployment of security in the Internet cannot be confined to a particular crypto algorithm or to a particular architecture. The limitations on the processing capability or bandwidth forces sacrifices on the security (e.g., smaller key sizes, IV reuse, CRC for integrity check). Some of these problems can be addressed by efficient cryptographic tools such as ECC, and some will disappear as the technology improves. Many attacks exploit the way protocols are designed and implemented, even if these protocols may use very secure ciphers. Examples include unprotected fields in the data structures (e.g., SSL 3.0 server key exchange message) and lack of authentication in ATMARP between server and client. Finally, the security problem in the Internet degenerates to the distributed consensus problem if network elements are compromised by the adversary. For example there is no easy way to check the "correctness" of a routing exchange message if it is signed by a once-legitimate-but-compromised router.

Thus the Internet will never be absolutely secure, and creating a high cost–benefit tradeoff for the attacker, to reduce the incentive, will always remain a practical security measure.

BIOGRAPHY

Bulent Yener is an Associate Professor at the Computer Science Department at Rensselaer Polytechnic Institute. Dr. Yener received B.S. and M.S. degrees in Industrial Engineering from the Technical University of Istanbul, Turkey, and M.S. and Ph.D. degrees in Computer Science, both from Columbia University, in 1987 and 1994, respectively. He was a Member of Technical Staff at the Bell Laboratories in Murray Hill, New Jersey during 1998–2001. Before joining to the Bell Laboratories in 1998, he served as an Assistant Professor at Lehigh University and NJIT. His current research interests include quality of service in the IP networks, wireless networks, and Internet security. He has served on the Technical Program Committee of leading IEEE conferences and workshops. Dr. Yener is a member of the IEEE and serves on the editorial boards of the *Computer Networks Journal* and the *IEEE Network Magazine*. He is a member of IEEE.

BIBLIOGRAPHY

1. B. Schneier, *Applied Cryptography*, 2nd ed., Wiley, New York, 1996.
2. D. R. Stinson, *Cryptography: Theory and Practice* (*Discrete Mathematics and Its Applications*), Chapman & Hall, 1995.
3. William F. Friedman, *Index of Coincidence and Its Applications in Cryptography*, Riverbank Publication 22, Riverbank Labs., 1920, reprinted by Agean Park Press, 1976.
4. E. H. Hebern, Electronic coding machine, U.S. Patent 1,510,441,30.
5. C. E. Shannon, in N. J. A. Sloane and A. D. Wyner, eds., *Collected Papers: Claude Elmwood Shannon*, IEEE Press, New York, 1993.
6. H. Feistel, Cryptography and computer privacy, *Sci. Am.* **228**(5): 15–23 (1973).
7. M. J. Weiner, Efficient DES key search, *Proc. CRYPTO'93*, 1993.
8. J. Daemen and V. Rijmen, *Rijndael Home Page* (online) *http://www.esat.kueven.ac.be/rijmen/rijndael* (March 26, 2002).
9. W. Diffie and M. E. Hellman, New directions in cryptography, *IEEE Trans. Inform. Theory* **IT-22**: 644–654 (1976).
10. R. Rivest, A. Shamir, and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Commun. ACM* **21**(2): 120–126 (1978).
11. ITU-T Recommendation X.509, *Information Technology — Open System Interconnection — The Directory: Authentication Framework*, 1997.
12. R. Housley, W. Ford, W. Polk, and D. Solo, *Internet X.509 Public Key Infrastructure Certificate and CRL Profile*, IETF RFC 2459, 1999.
13. J. Postel and J. Reynolds, File Transfer Protocol, *IETF RFC 959*, 1985.
14. J. Postel, *Transmission Control Protocol*, IETF RFC 791, 19.
15. S. M. Bellovin, Security Problems in the TCP/IP Protocol Suite, *ACM Comput. Commun. Rev.* **19**(2): 32–48 (1989).
16. S. M. Bellovin and W. R. Cheswick, *Firewalls and Internet Security*, Addison-Wesley, New York, 1994.
17. T. Ylonen et al., *SSH Protocol Architecture* (online) *draft-ietfsecsh-architecture-12.txt*, 2002.

18. S. Kent, C. Lynn, and K. Seo, Secure Border Gateway Protocol (S-BGP), *IEEE JSAC Network Security* **18**(34): 582–592 (2000).

19. M. Raynal, *Distributed Algorithms and Protocols*, Wiley, New York, 1988.

20. S. Kent and R. Atkinson, *Security Architecture for the Internet Protocol*. IETF RFC 2401, 1998.

21. S. Bellovin, Problem areas for the IP security protocols, *Proc. 6th USENIX Security Symp.* 1996, pp. 205–214.

22. D. Harkins and D. Carrel, *The Internet Key Exchange*, IETF RFC 2409, 1998.

23. ATM Forum. *http://www.atmforum.org*.

24. A. Frier, P. Karlton, and P. Koccher, *The SSL3.0 Protocol Version 3.0,* Netscape, 1996 (online) *http://home.netscape.com/eng/ssl3/* (March 26, 2002).

25. T. Dierks and C. Allen, *The TLS Protocol Version 1.0*, IETF RFC 2246, 1999.

26. D. Wagner and B. Schneier, Analysis of the SSL 3.0 protocol, *Proc. 2nd USENIX Workshop on Electronic Commerce*, USENIX Press, 1996, pp. 29–40 (online) *http://citeseer.nj.nec.com/article/wagner96analysi.html* (March 26, 2002).

27. H. Harney and C. Muckenhirn, *Group Key Management Protocol* (*GKMP*) *Specification*, IETF RFC 2093, 1997.

28. H. Harney and C. Muckenhirn, *Group Key Management Protocol* (*GKMP*) *Architecture*, IETF RFC 2094, 1997.

29. S. Mittra, Iolus: A framework for scalable secure multicasting, *Proc. ACM SIGCOMM'97*, 1997.

30. R. Canetti et al., Multicast security: A taxonomy and efficient constructions, *Proc. IEEE INFOCOM'99*, 1999.

31. J. Kohl and C. Neuman, *The Kerberos Network Authentication Service* (*V5*), IETF RFC 1510, 1993.

32. A. Fiat and M. Naor, Broadcast encryption, *Advances in Cryptography — Crypto'92*, 1995, Vol. 8, pp. 189–200.

33. D. M. Wallner, E. J. Harder, and R. C. Agee, *Key Management for Multicast: Issues and Architectures*, IETF RFC 2627, 1999.

34. C. K. Wong and S. Lam, Digital signature for flows and multicasts, *Proc. IEEE ICNP'98'*, 1998.

35. N. Koblitz, Elliptic curve cryptosystems, *Math. Comput.* **48**: 203–209 (1987).

36. Certicom White Paper, *Current Public-Key Cryptographic Systems*, 1997 (online) *http://www.certicom.com* (March 26, 2002).

37. N. Borisov, I. Goldberg, and D. Wagner, Intercepting mobile communications: The insecurity of 802.11, *Proc. Mobile Computing and Networking*, 2001.

38. WAP Forum, *Wireless Application Protocol — Wireless Transport Layer Security Specification version 1* (online) *http://www.wapforum.org* (March 26, 2002).

39. WAP Forum, *Wireless Application Protocol* (online) *http://www.wapforum.org* (March 26, 2002).

40. R. Rivest, *The MD5 Message-Digest Algorithm*, IETF RFC 1321, 1992.

41. Federal Information Processing Standard Publication 180-1, 1995 (online) *http://www.itl.nist.gov/fibspubs/fip180-1.htm* (March 26, 2002).

INTERSYMBOL INTERFERENCE IN DIGITAL COMMUNICATION SYSTEMS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Intersymbol interference arises in both wireline and wireless communication systems when data are transmitted at symbol rates approaching the Nyquist rate and the characteristics of the physical channels through which the data are transmitted are nonideal. Such interference can severely limit the performance that can be achieved in digital data transmission, where performance is measured in terms of the data transmission rate and the resulting probability of error in recovering the data from the received channel corrupted signal.

The degree to which one must be concerned with channel impairments generally depends on the transmission rate of data through the channel. If R is the transmission bit rate and W is the available channel bandwidth, intersymbol interference (ISI) caused by channel distortion generally arises when $R/W > 1$. Since bandwidth is usually a precious commodity in communication systems, it is desirable to utilize the channel to as near its capacity as possible. In such cases, the communication system designer must employ techniques that mitigate the effects of ISI caused by the channel.

This article provides a characterization of ISI resulting from channel distortion.

2. CHANNEL DISTORTION

Wireline channels may be characterized as linear time-invariant filters with specified frequency-response characteristics. If a channel is band-limited to W Hz, then its frequency response $C(f) = 0$ for $|f| > W$. Within the bandwidth of the channel, the frequency response $C(f)$ may be expressed as

$$C(f) = |C(f)|e^{j\theta(f)} \tag{1}$$

where $|C(f)|$ is the amplitude-response characteristic and $\theta(f)$ is the phase-response characteristic. Furthermore, the envelope delay characteristic is defined as

$$\tau(f) = -\frac{1}{2\pi}\frac{d\theta(f)}{df} \tag{2}$$

A channel is said to be *nondistorting* or *ideal* if the amplitude response $|C(f)|$ is constant for all $|f| \leq W$ and $\theta(f)$ is a linear function of frequency, that is, if $\tau(f)$ is a constant for all $|f| \leq W$. On the other hand, if $|C(f)|$ is not constant for all $|f| \leq W$, we say that the channel *distorts the transmitted signal in amplitude*, and, if $\tau(f)$ is not constant for all $|f| \leq W$, we say that the channel *distorts the signal in delay*.

As a result of the amplitude and delay distortion caused by the nonideal channel frequency-response

characteristic $C(f)$, a succession of pulses transmitted through the channel at rates comparable to the bandwidth W are smeared to the point that they are no longer distinguishable as well-defined pulses at the receiving terminal. Instead, they overlap, and, thus, we have intersymbol interference. As an example of the effect of delay distortion on a transmitted pulse, Fig. 1a illustrates a band-limited pulse having zeros periodically spaced in time at points labeled $\pm T, \pm 2T$, and so on. If information is conveyed by the pulse amplitude, as in pulse amplitude modulation (PAM), for example, then one can transmit a sequence of pulses each of which has a peak at the periodic zeros of the other pulses. However, transmission of the pulse through a channel modeled as having a linear envelope delay characteristic $\tau(f)$ [quadratic phase $\theta(f)$] results in the received pulse shown in Fig. 1b having zero crossing that are no longer periodically spaced. Consequently, a sequence of successive pulses would be smeared into one another and the peaks of the pulses would no longer be distinguishable. Thus, the channel delay distortion results in intersymbol interference.

Another view of channel distortion is obtained by considering the impulse response of a channel with nonideal frequency response characteristics. For example, Fig. 2 illustrates the average amplitude response $|C(f)|$ and the average envelope delay $\tau(f)$ for a medium-range (180–725-mi) telephone channel of the switched telecommunications network. It is observed that the usable band of the channel extends from ~300 to ~3000 Hz. The corresponding impulse response of this average channel is shown in Fig. 3. Its duration is about 10 ms. In comparison, the transmitted symbol rates on such a channel may be of the order of 2500 pulses or symbols per second. Hence, intersymbol interference might extend over 20–30 symbols.

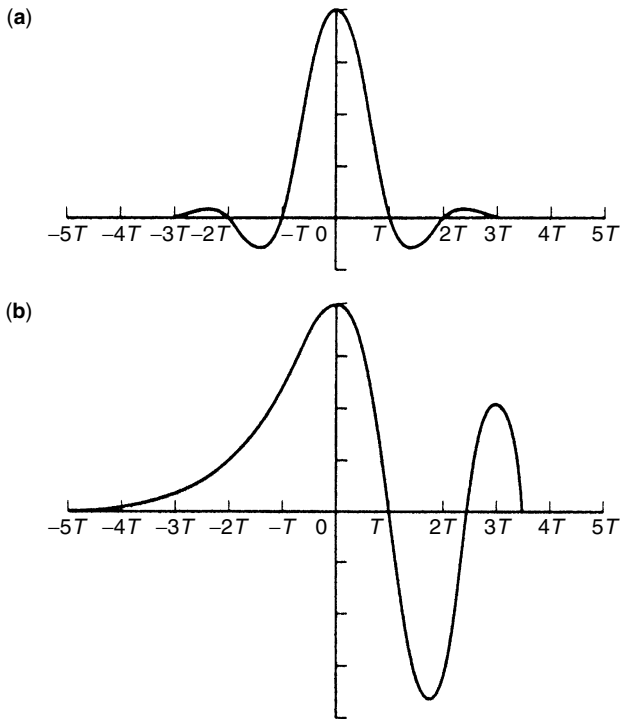


Figure 1. Effect of channel distortion (**a**) channel input (**b**) channel output.

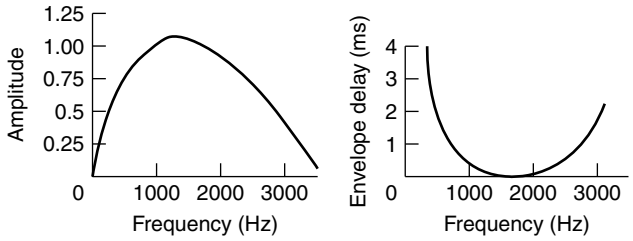


Figure 2. Average amplitude and delay characteristics of medium-range telephone channel.

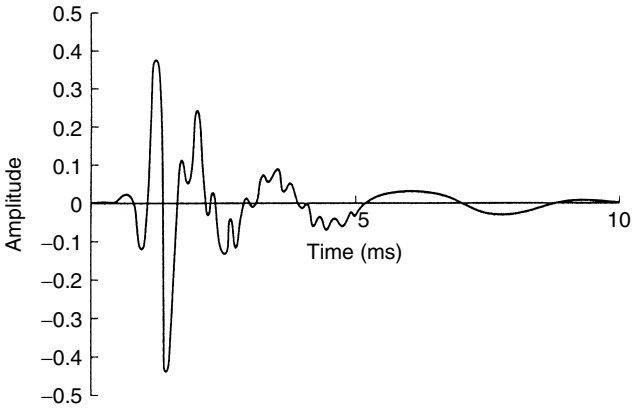


Figure 3. Impulse response of medium-range telephone channel with frequency response shown in Fig. 2.

Besides wireline channels, there are other physical channels that exhibit some form of time dispersion and, thus, introduce intersymbol interference. Radio channels such as shortwave ionospheric channels (HF), tropospheric scatter channels, and mobile cellular radio channels are examples of time-dispersive channels. In these channels, time dispersion and, hence, intersymbol interference are the result of multiple propagation paths with different path delays. The number of paths and the relative time delays among the paths vary with time, and, for this reason, these radio channels are usually called *time-variant multipath channels*. The time-variant multipath conditions give rise to a wide variety of frequency-response characteristics and the resulting phenomenon called *signal fading*.

In the following section, a mathematical model for the intersymbol interference (ISI) is developed and transmitted signal characteristics are described for avoiding ISI.

3. CHARACTERIZATION OF INTERSYMBOL INTERFERENCE

In conventional linear digital modulation such as pulse amplitude modulation (PAM), phase shift keying (PSK), and quadrature amplitude modulation (QAM), the transmitted signal is generally represented as an equivalent

lowpass signal (prior to frequency conversion[1] for transmission over the bandpass channel) of the form

$$v(t) = \sum_{n=0}^{\infty} I_n g(t - nT) \tag{3}$$

where $\{I_n\}$ represents the discrete information-bearing sequence of symbols and $g(t)$ is a modulation filter pulse that, for the purposes of this discussion, is assumed to have a band-limited frequency-response characteristic $G(f)$, specifically, $G(f) = 0$ for $|f| > W$. For PAM, the information-bearing sequence $\{I_n\}$ consists of symbols taken from the alphabet $\{\pm 1, \pm 3, \ldots, +(M-1)\}$ for M-level amplitude modulation. In the case of PSK, the information-bearing sequence $\{I_n\}$ consists of symbols taken from the alphabet $\left\{e^{j\theta_m}, \theta_m = \frac{2\pi}{M}m, \quad m = 0, 1, \ldots, M-1\right\}$. QAM may be considered as a combined form of digital amplitude and phase modulation, so that the sequence $\{I_n\}$ takes values of the form $\{A_m e^{j\theta_m}, \quad m = 0, 1, \ldots, M-1\}$.

The signal given by Eq. (3) is transmitted over a channel having a frequency response $C(f)$, also limited to $|f| < W$. Consequently, the received signal can be represented as

$$r(t) = \sum_{n=0}^{\infty} I_n h(t - nT) + z(t) \tag{4}$$

where

$$h(t) = \int_{-\infty}^{\infty} g(\tau) c(t - \tau)\, d\tau \tag{5}$$

and $z(t)$ represents the additive white Gaussian noise that originates at the front end of the receiver. The channel impulse response is denoted as $c(t)$.

Suppose that the received signal is passed first through a filter and then sampled at a rate $1/T$ samples. Since the additive noise is white Gaussian, the optimum filter at the receiver is the filter that is matched to the signal pulse $h(t)$; that is, the frequency response of the receiving filter is $H^*(f)$. The output of the receiving filter[2] may be expressed as

$$y(t) = \sum_{n=0}^{\infty} I_n x(t - nT) + v(t) \tag{6}$$

where $x(t)$ is the pulse representing the response of the receiving filter to the input pulse $h(t)$ and $v(t)$ is the response of the receiving filter to the noise $z(t)$.

[1] The frequency upconversion that is performed at the transmitter and the corresponding downconversion performed at the receiver may be considered transparent, so that transmitted and received signals are treated in terms of the equivalent lowpass characteristics.

[2] Often, the frequency-response characteristics of the channel are unknown to the receiver. In such a case, the receiver cannot implement the optimum matched filer to the signal pulse $h(t)$. Instead, the receiver may implement the filter matched to the transmitted pulse $g(t)$.

Now, if $y(t)$ is sampled at times $t = kT + \tau_0, k = 0, 1, \ldots$, we have

$$y(kT + \tau_0) = y_k = \sum_{n=0}^{\infty} I_n x(kT - nT + \tau_0) + v(kT + \tau_0) \tag{7}$$

or, equivalently

$$y_k = \sum_{n=0}^{\infty} I_n x_{k-n} + v_k, \quad k = 0, 1, \ldots \tag{8}$$

where τ_0 is the transmission delay through the channel. The sample values can be expressed as

$$y_k = x_0 \left(I_k + \frac{1}{x_0} \sum_{\substack{n=0 \\ n \neq k}}^{\infty} I_n x_{k-n} \right) + v_k, \quad k = 0, 1, \ldots \tag{9}$$

We regard x_0 as an arbitrary scale factor, which can be set equal to unity for convenience. Then

$$y_k = I_k + \sum_{\substack{n=0 \\ n \neq k}}^{\infty} I_n x_{k-n} + v_k \tag{10}$$

The term I_k represents the desired information symbol at the kth sampling instant, the term

$$\sum_{\substack{n=0 \\ n \neq k}}^{\infty} I_n x_{k-n} \tag{11}$$

represents the ISI, and v_k is the additive Gaussian noise variable at the kth sampling instant.

The amount of intersymbol interference and noise in a digital communication system can be viewed on an oscilloscope. For PAM signals, we can display the received signal $y(t)$ on the vertical input with the horizontal sweep rate set at $1/T$. The resulting oscilloscope display is called an "eye pattern" because of its resemblance to the human eye. For example, Fig. 4 illustrates the eye patterns for binary and quaternary PAM modulation. The effect of ISI is to cause the eye to close, thereby reducing the margin for additive noise to cause errors. Figure 5 graphically illustrates the effect of intersymbol interference in reducing the opening of a binary eye. Note that intersymbol interference distorts the position of the zero crossings and causes a reduction in the eye openings. Thus, it causes the system to be more sensitive to a synchronization error.

For PSK and QAM it is customary to display the "eye pattern" as a two-dimensional scatter diagram illustrating the sampled values $\{y_k\}$ that represent the decision variables at the sampling instants. Figure 6 illustrates such an eye pattern for an 8-PSK signal. In the absence of intersymbol interference and noise, the superimposed signals at the sampling instants would result in eight distinct points corresponding to the eight transmitted signal phases. Intersymbol interference and noise result in a deviation of the received samples $\{y_k\}$ from the desired 8-PSK signal. The larger the intersymbol interference

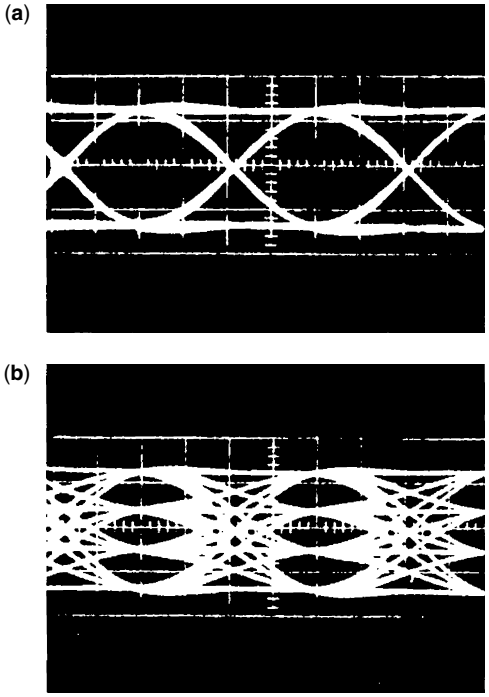


Figure 4. Examples of eye patterns for binary (**a**) and quaternary (**b**) PAM.

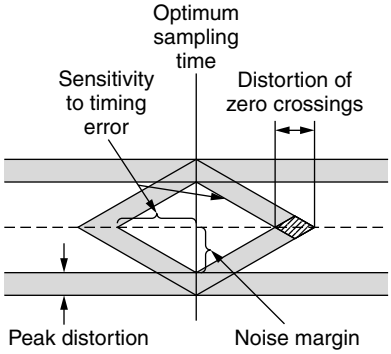


Figure 5. Effect of intersymbol interference on eye opening.

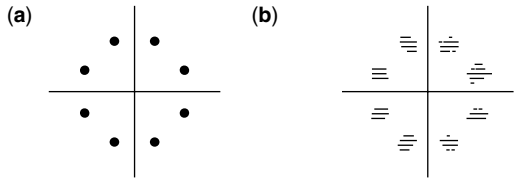


Figure 6. "Eye Patterns" for a two-dimensional signal constellation: (**a**) transmitted eight-phase signal; (**b**) received signal samples at the output of demodulator.

and noise, the larger the scattering of the received signal samples relative to the transmitted signal points.

The following section considers the design of transmitted pulses that result in no ISI in the transmission through a band-limited channel.

4. SIGNAL DESIGN FOR ZERO ISI — THE NYQUIST CRITERION

The problem treated in this section is the design of the transmitter and receiver filters in a modem so that the received signal has zero ISI, assuming the condition that the channel is an ideal channel. The subsequent section treats the problem of filter design when there is ISI due to channel distortion.

Under the assumption that the channel frequency response is ideal, i.e., $C(f) = 1$ for $|f| \leq W$, the pulse $x(t)$ has a spectral characteristic $X(f) = |G(f)|^2$, where

$$x(t) = \int_{-W}^{W} X(f) e^{j2\pi ft}\, df \tag{12}$$

We are interested in determining the spectral properties of the pulse $x(t)$ and, hence, the transmitted pulse $g(t)$, which results in no intersymbol interference. Since

$$y_k = I_k + \sum_{\substack{n=0 \\ n \neq k}}^{\infty} I_n x_{k-n} + v_k \tag{13}$$

the condition for no intersymbol interference is

$$x(t = kT) \equiv x_k = \begin{cases} 1 & (k = 0) \\ 0 & (k \neq 0) \end{cases} \tag{14}$$

Nyquist [1] formulated and solved this problem in the late 1920s. He showed that a necessary and sufficient condition for zero ISI, given by Eq. (14) is that the Fourier transform $X(f)$ of the signal pulse $x(t)$ satisfy the condition

$$\sum_{m=-\infty}^{\infty} X(f + m)T) = T \tag{15}$$

Nyquist also demonstrated that if the symbol transmission rate $1/T > 2W$, where $2W$ is called the Nyquist rate, it is impossible to design a signal $x(t)$ that has zero ISI. If the symbol rate $1/T = 2W$, the only possible signal pulse that yields zero ISI has the spectrum

$$X(f) = \begin{cases} T, & |f| \leq W \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

and the time response

$$x(t) = \frac{\sin(\pi t/T)}{t\pi/T} \equiv \text{sinc}\left(\frac{\pi t}{T}\right) \tag{17}$$

This means that the smallest value of T for which transmission with zero ISI is possible is $T = 1/2W$, and for this value, $x(t)$ has to be the sinc function. The difficulty with this choice of $x(t)$ is that it is concausal and, therefore, nonrealizable. To make it realizable, usually a delayed version of it, $\text{sinc}[\pi(t - t_0)/T]$, is used and t_0 is chosen such that for $t < 0$, we have $\text{sinc}[\pi(t - t_0)T] \approx 0$. Of course, with this choice of $x(t)$, the sampling time must also be shifted to $mT + t_0$. A second difficulty with this pulseshape is that its rate of convergence to zero is slow. The tails of $x(t)$ decay as $1/t$; consequently, a small mistiming error in sampling

the output of the receiver filter at the demodulator results in an infinite series of ISI components. Such a series is not absolutely summable because of the $1/t$ rate of decay of the pulse, and, hence, the sum of the resulting ISI does not converge. Consequently, the signal pulse given by Eq. (17) does not provide a practical solution to the signal design problem.

By reducing the symbol rate $1/T$ to be slower than the Nyquist rate, $1/T < 2W$, there exists numerous choices for $X(f)$ that satisfy Eq. (17). A particular pulse spectrum that has desirable spectral properties and has been widely used in practice is the raised-cosine spectrum. The raised-cosine frequency characteristic is given as

$$\mathrm{X_{rc}(f)} = \begin{cases} T & \left(0 \le |f| \le \dfrac{1-\beta}{2T}\right) \\ \dfrac{T}{2}\left\{1+\cos\left[\dfrac{\pi T}{\beta}\left(|f|-\dfrac{1-\beta}{2T}\right)\right]\right\} & \left(\dfrac{1-\beta}{2T} \le |f| \le \dfrac{1-\beta}{2T}\right) \\ 0 & \left(|f| > \dfrac{1+\beta}{2T}\right) \end{cases} \tag{18}$$

where β is called the *rolloff factor* and takes values in the range $0 \le \beta \le 1$. The bandwidth occupied by the signal beyond the Nyquist frequency $1/2T$ is called the *excess bandwidth* and is usually expressed as a percentage of the Nyquist frequency. For example, when $\beta = \frac{1}{2}$, the excess bandwidth is 50% and when $\beta = 1$, the excess bandwidth is 100%. The pulse $x(t)$, having the raised-cosine spectrum, is

$$\begin{aligned} x(t) &= \frac{\sin(\pi t/T)}{\pi t/T}\frac{\cos(\pi\beta t/T)}{1-4\beta^2t^2/T^2} \\ &= \mathrm{sinc}(\pi t/T)\frac{\cos(\pi\beta t/T)}{1-4\beta^2t^2/T^2} \end{aligned} \tag{19}$$

Note that $x(t)$ is normalized so that $x(0) = 1$. Figure 7 illustrates the raised-cosine spectral characteristics and the corresponding pulses for $\beta = 0, \frac{1}{2}$, and 1. Note that for $\beta = 0$, the pulse reduces to $x(t) = \mathrm{sinc}(\pi t/T)$, and the symbol rate $1/T = 2W$. When $\beta = 1$, the symbol rate is $1/T = W$. In general, the tails of $x(t)$ decay as $1/t^3$ for $\beta > 0$. Consequently, a mistiming error in sampling leads

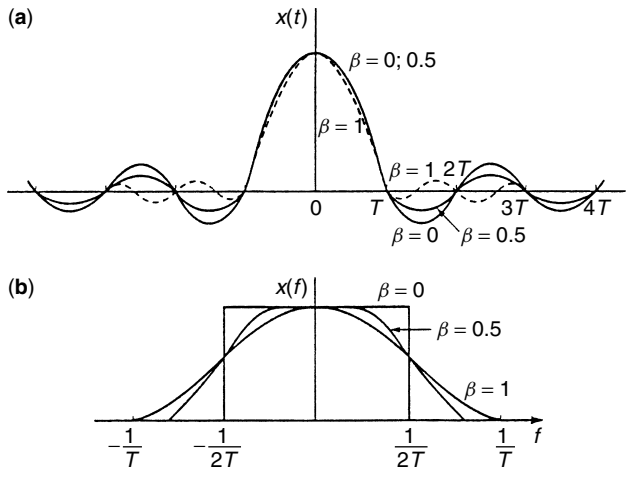


Figure 7. Signal Pulses with a Raised-Cosine Spectrum.

to a series of ISI components that converges to a finite value.

Because of the smooth characteristics of the raised-cosine spectrum, it is possible to design practical filters for the transmitter and the receiver that approximate the overall desired frequency response. In the special case where the channel is ideal, $C(f) = 1$, $|f| \le W$, we have

$$X_{rc}(f) = G_T(f)G_R(f) \tag{20}$$

where $G_T(f)$and $G_R(f)$ are the frequency responses of the filters at the transmitter and the receiver, respectively. In this case, if the receiver filter is matched to the transmitter filter, we have $X_{rc}(f) = G_T(f)G_R(f) = |G_T(f)|^2$. Ideally

$$G_T(f) = \sqrt{|X_{rc}(f)|}\, e^{-j2\pi f t_0} \tag{21}$$

and $G_R(f) = G_T^*(f)$, where t_0 is some nominal delay that is required to ensure physical realizability of the filter. Thus, the overall raised-cosine spectral characteristic is split evenly between the transmitting filter and the receiving filter. Note also that an additional delay is necessary to ensure the physical realizability of the receiving filter.

5. CHANNEL DISTORTION AND ADAPTIVE EQUALIZATION

Modems that are used for transmitting data either on wireline or wireless channels are designed to deal with channel distortion, which usually differs from channel to channel. For example, a modem that is designed for data transmission on the switched telephone network encounters a different channel response every time a telephone number is dialed, even if it is the same telephone number. This is due to the fact that the route (circuit) from the calling modem to the called modem will vary from one telephone call to another.

Since the channel distortion is variable and unknown a priori, a modem contains an additional component, called an *adaptive equalizer*, that follows the receiving filter $G_R(f)$, which further processes the received signal samples $\{y_k\}$ to reduce the ISI. The most commonly used equalizer is a linear, discrete-time, finite-impulse-response duration (FIR) filter with coefficients that are adjusted to reduce the ISI. With $\{y_k\}$ as its input, given by Eq. (10) and coefficients $\{b_k\}$, the equalizer output sequence is an estimate of the transmitted symbols $\{I_k\}$

$$\hat{I}_k = \sum_{m=-N}^{N} b_m y_{k-m} \tag{22}$$

where $2N + 1$ is the number of equalizer coefficients.

When a call is initiated, a sequence $\{I_k\}$ of training symbols (known to the receiver) are transmitted. The receiver compares the known training symbols with the sequence of estimates $\{\hat{I}_k\}$ from the equalizer and computes the sequence of error signals

$$\begin{aligned} e_k &= I_k - \hat{I}_k \\ &= I_k - \sum_{m=-N}^{N} b_m y_{k-m} \end{aligned} \tag{23}$$

The equalizer coefficients $\{b_k\}$ are the adjusted to minimize the mean (average) squared value of the error sequence. Thus, the equalizer adapts to the channel characteristics by adjusting its coefficients to reduce ISI.

The effect of the linear adaptive equalizer in compensating for the channel distortion can also be viewed in the frequency domain. The transmitter filter $G_T(f)$, the channel frequency response $C(f)$, the receiver filter $G_R(f)$, and the equalizer $B(f)$ are basically liner filters in cascade. Hence, their overall frequency response is $G_T(f)C(f)G_R(f)B(f)$. If the cascade $G_T(f)G_R(f)$ is designed for zero ISI as in Eq. (20), then by designing the equalizer frequency response $B(f)$ such that $B(f) = 1/C(f)$, the ISI is eliminated. In this case, the adaptive equalizer has a frequency response that is the inverse of the channel response. If the adaptive equalizer had an infinite number of coefficients (infinite-duration impulse response), its frequency response would be exactly equal to the inverse channel characteristic and, thus, the ISI would be completely eliminated. However, with a finite number of coefficients, the equalizer response can only approximately equal the inverse channel response. Consequently, some residual ISI will always exist at the output of the equalizer. For more detailed treatments of adaptive equalization, the interested reader may refer to Refs. 2–4.

6. CONCLUDING REMARKS

The treatment of ISI in this article was based on the premise that the signal transmitted through the channel was carried on a single carrier frequency. The effect of channel distortion and, hence, ISI can be significantly reduced if a channel with bandwidth W is subdivided into a large number of subchannels, say, N, where each subchannel has a bandwidth $B = W/N$. Modulated carrier signals are then transmitted in all the N subchannels. In this manner, each information symbol on each subchannel has a symbol duration of NT, where T is the symbol duration in a single carrier modulated signal. The modulation of all the carriers in the N subchannels is performed synchronously. This type of multicarrier modulation is called *orthogonal frequency-division multiplexing* (OFDM). If the bandwidth B of each subchannel is made sufficiently small, each subchannel will have (approximately) an ideal frequency-response characteristic. Hence, the ISI in each subchannel becomes insignificant. However, OFDM is particularly vulnerable to interference among the subcarriers (interchannel interference) due to frequency offsets or Doppler frequency spread effects, which are present when either the transmitter and/or the receiver terminals are mobile.

ISI in digital communications systems is treated numerous publications in the technical literature and various textbooks. The interested reader may consult Refs. 2–4 for more detailed treatments.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. H. Nyquist, Certain topics in telegraph transmission theory, *AIEE Trans* **47**: 617–644 (1928).
2. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
3. E. A. Lee and D. G. Messerschmitt, *Digital Communications*, 2nd ed., Kluwer, Boston, 1994.
4. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communications*, McGraw-Hall, New York, 1968.

INTRANETS AND EXTRANETS

ALGIRDAS PAKŠTAS
London Metropolitan University
London, England

1. INTRODUCTION

The intranet is a private network that extensively uses established Web technologies based on the Internet protocols (TCP/IP, HTTP, etc.) [1]. An intranet is accessible only by the organization's members, employees, or other users who have authorization. The Internet is a public access network. Therefore, an organization's Webpage on the Internet is its public face that helps to create its image. Such pages may be built with many graphics and special features. The intranet, however, is the organization's private face where employees get their information and then get off and go back to work. An intranet's Websites may look and act just like any other Websites, but most often their appearance is simpler and more casual. Both, the Internet and an intranet use the same types of hardware and software, but they are used for two very different purposes. An intranet is protected by the firewall surrounding it from the unauthorized access. Like the Internet itself, intranets are used to share information, but additionally the intranet is the communications platform for group work — "the 'electronic brain' employees tap into," as expressed by Steve McCormick, a consultant with Watson Wyatt Worldwide, Washington, D.C., USA. Secure intranets are now the fastest growing segment of the Internet because they are much less expensive to build and manage than private networks based on proprietary protocols. Since 1996, intranets have been embraced by corporate users of information services and have made substantial inroads in strategic vision documents and procurement practices.

An extranet is an intranet that is partially accessible to authorized outsiders, such as an organization's mobile workers or representatives of partner (sharing common goals) businesses such as suppliers, vendors, customers, and so on [2]. Thus, an extranet, or extended intranet, is a private business network of several cooperating organizations located outside the corporate firewall. Whereas an intranet resides behind a firewall and is accessible only to people who are members of the same company or organization, an extranet provides various levels of accessibility to outsiders. A user can access an extranet only if he or she has a valid username and password, and the user's identity determines which parts of the extranet can be viewed or accessed. Extranets are becoming a very popular means for business partners to exchange information. Extranets are using not only internal network resources but also the public telecommunication system. Therefore, an extranet requires security and privacy. These require firewall server management, the issuance and use of digital certificates or similar means of user authentication, encryption of messages, and the use of virtual private networks (VPN) that tunnel through the public network.

The extranets were introduced shortly after intranets. At the same time, some experts are arguing that extranets have been around since time when the first rudimentary LAN-to-LAN networks began connecting two different business entities together to form WANs and that in its basic form, an extranet is the interconnection of two previously separate LANs or WANs with origins from different business entities [3].

Table 1 summarizes the discussed features of Internet, intranet, and extranet.

1.1. Examples of Extranet Applications

The most obvious examples of extranet applications are the following:

- Exchange of large volumes of data using electronic data interchange (EDI)
- Shared product catalogs accessible only to wholesalers or those "in the trade"
- Collaboration with other companies on joint development efforts, e.g., establishing private newsgroups that cooperating companies use to share valuable experiences and ideas
- Groupware in which several companies collaborate in developing a new application program they can all use
- Project management and control for companies that are part of a common work project
- Provide or access services provided by one company to a group of other companies, such as an online banking application managed by one company on behalf of affiliated banks

Table 1. Summary of the Internet, Intranet, and Extranet Features

	Internet	Intranet	Extranet
Users	Everyone	Members of the specific firm	Group of closely related firms
Information	Fragmented	Proprietary	Shared in closely trusted held circles
Access	Public	Private	Semi-private
Security mechanism	None	Firewall, encryption	Intelligent firewall, encryption, various document security standards

- Training programs or other educational material that companies could develop and share

2. INTRANET'S HARDWARE AND SOFTWARE

The intranet's hardware consists of the client/server network. The clients are computers that are connected either to the LAN or to the WAN. The servers are high-performance computers with a large hard disk capacity and RAM.

The network operating system is the software required to run the network and share its resources (printers, files, applications, etc.) among the users. There are three basic software packages needed for a corporate intranet: Web software, firewall software, and browser software. Web software allows the server to support HTTP (Hyper Text Transfer Protocol) so it can exchange information with the clients. Firewall software provides the security needed to protect corporate information from the outside world. Browser software allows the user to read electronic documents published on the Internet or corporate intranet.

Other functions can be provided by adding software for Internet access and searching, authoring and publishing documents, collaboration and conferencing, database archive and retrieval, and document management access. A properly constructed intranet allows for the following to be done effectively [1]: centralized storage of data (real-time and archived), scalability and management of application services, centralized control over access to knowledge, decentralized management of knowledge, and universal access to the information that decision makers have deemed it appropriate to view.

These of established hardware and software technologies for intranet and extranet infrastructure makes intranets and extranets very economical in comparison with the creation and maintenance of proprietary networks. The following main approaches are used for building extranet software and inevitably are linked to the backing industrial groups or corporations:

- Crossware: Netscape, Oracle, and Sun Microsystems have formed an alliance to ensure that their extranet products can work together by standardizing on JavaScript and the Common Object Request Broker Architecture (CORBA).
- Microsoft supports the Point-to-Point Tunneling Protocol (PPTP) and is working with American Express and other companies on an Open Buying on the Internet (OBI) standard.
- The Lotus Corporation is promoting its groupware product, Notes, as well suited for extranet use.

The first approach (Crossware) is oriented to development of open application standards, while the other two are more focused on a "one company's product" type of philosophy.

Thus, intranets and extranets are rather classes of applications than *categories of networks*. Applications in the public Internet, intranet, and extranet will all run on the same type of network infrastructure, but their software and data content resources will be *administered* for different levels of accessibility and security. This can be achieved with the help of tools for network management.

2.1. Extranets and Intergroupware

Workgroups may often need special tools for *communications*, *collaboration*, and *coordination*, which are referred as *groupware* [4]. Emphasis is on computer-aided help to implement message-based human communications and information sharing as well as to support typical workgroup tasks such as scheduling and routing of the workflow tasks.

In the typical enterprise communications the inter- and intraorganizational media-based communications activities are forming two separate parallel planes [4]. The dimensions of each plane are the degree of *structure* and the degree of *mutuality* in the communications activities. *Structure* lies between informal (ad hoc) and formally structured, defined, and managed or edited processes. *Mutuality* ranges from unidirectional or sequential back-and-forth message passing, to true joint work or *collaborative transactions* in a shared space of information.

The resulting four regions characterize four different kinds of interaction, which place distinct demands on their media vehicles or tools. Separate tools have been developed in each region, usually as isolated solutions, but the need to apply them to the wide spectrum of electronic documents and in coordinated form is causing them to converge. Thus, we can describe groupware in terms of the following categories representing three of the mentioned four regions. These are [4]:

- Communications or messaging (notably E-mail)
- Collaboration or conferencing (notably forums or "bulletin board" systems that organize messages into topical "threads" of group discussion, maintained in a shared database)
- Coordination or workflow and transactions (applying predefined rules to automatically process and route messages)

Figure 1 shows the distinct forms of electronic media supported by the groupware, the kinds of interactions they support, and how they are converging [4]. Intergroupware is just groupware applied with the flexibility to support multiple interacting groups, which may be open or closed, and which may share communications selectively, as appropriate (as in an extranet). It should be noted that Lotus Notes is one of the approaches to implement intergroupware, and its success is based on the recognition of the fact that while these mentioned categories have distinct characteristics, they can only be served effectively by a unified platform that allows them to interact seamlessly.

2.2. Open Application Standards for Creating Extranets

Broad use of the Internet technology in general and development of extranets in particular is now supported by the existence of *the open application standards* that offer a range of features and functionality across all client and

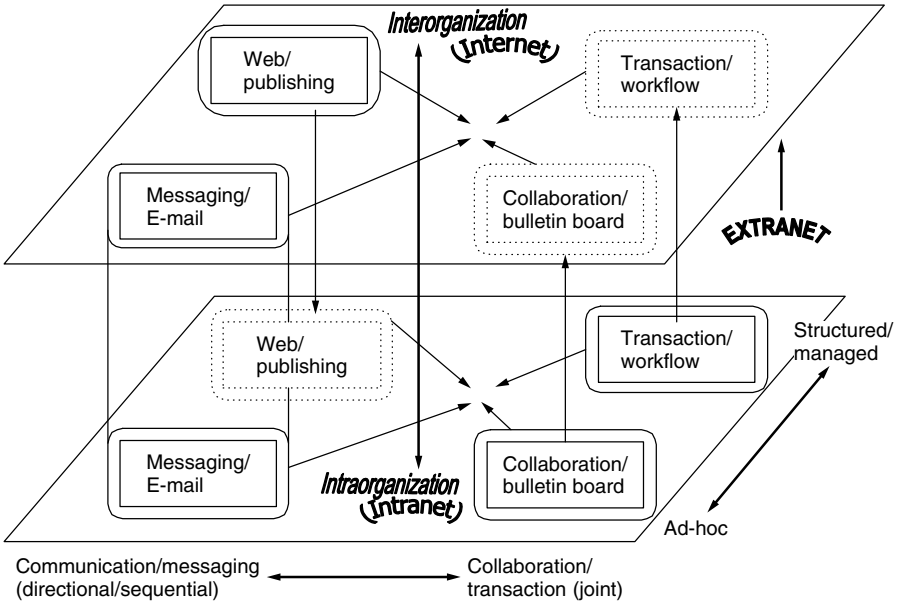


Figure 1. Relations between extranets and intergroupware. (*Source*: Teleshuttle Corporation, (4), 1997).

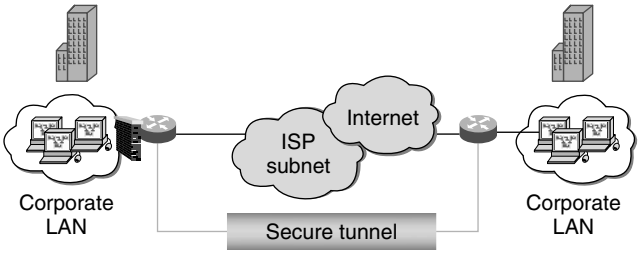


Figure 2. Site-to-site VPN. (*Source*: Core Competence, (14), 2001).

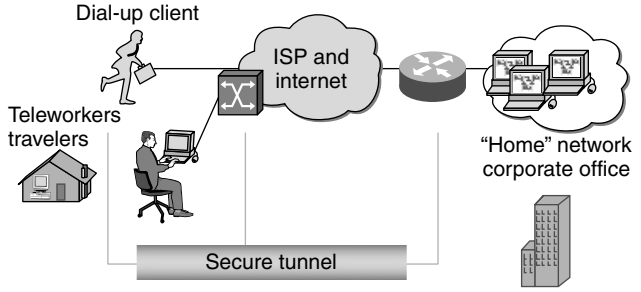


Figure 3. Remote access VPN. (*Source*: Core Competence, (14), 2001).

server platforms. Among these are the following groups of standards:

- Platform-independent content creation, publishing, and sharing of the information: HTML and HTTP
- Platform-independent software development as well as creation and deployment of distributed objects: Java, JavaScript, Common Object Request Broker Architecture (CORBA)
- Platform-independent messaging and collaboration (E-mail, discussion, and conferencing capabilities): Simple Mail Transfer Protocol (SMTP), Internet Message Access Protocol (IMAP), Multipurpose Internet Mail Extensions (MIME), Secure MIME (S/MIME), Network News Transport Protocol (NNTP), Real-Time Protocol (RTP)
- Directory and security services, network management capabilities: Lightweight Directory Access Protocol (LDAP), X.509, Simple Network Management Protocol (SNMP)

Netscape Communications has established a partnership with other companies who have agreed on a collection of standards and "best practices" for use in extranet deployment and the creation of *Crossware*. For enterprises, this offers two significant benefits:

1. An assurance of interoperability among products from multiple vendors.
2. A virtual roadmap for efficient implementation of an extranet.

Netscape's partners have committed to support the following Internet standards: LDAP, X.509 v3, S/MIME, vCards, JavaSoft, and EDI INT. Together, these standards create a comprehensive infrastructure that enables Crossware applications to interoperate across the Internet and the intranets of business partners, suppliers, and customers.

They also serve to provide a secure environment that supports much more than simple exchange of HTML pages between enterprises. In fact, the standards agreed upon by Netscape's partners represent by far the most secure and the best supported open standards technology. These standards are briefly described in the following sections.

LDAP intelligent directory services store and deliver contact information, registration data, certificates, configuration data, and server state information. These services provide support for single-user logon applications

and strong authentication capabilities throughout the extranet. Key benefits include:

- Users can search for contact information across enterprises, partners, and customers using the same interface and protocols as internal corporate directories.
- A standard format for storage and exchange of X.509 digital certificates allows single-user logon applications and secure exchange of documents and information via S/MIME.
- Replication over open LDAP protocol allows secure distribution of directory data between enterprises.
- Enables extranet applications that rely on fast and flexible queries of structure information.

X.509 v3 digital certificates provide a secure container of validated and digitally signed information. They offer strong authentication between parties, content, or devices on a network including secure servers, firewalls, email, and payment systems. They are a foundation for the security in S/MIME, object signing, and Electronic Document Interchange over the Internet (EDI INT). Digital certificates can be limited to operate within an intranet or they can operate between enterprises with public certificates coissued by the company and a certification authority such as VeriSign. Certificates surpass passwords in providing strong security by authenticating identity, verifying message and content integrity, ensuring privacy, authorizing access, authorizing transactions, and supporting nonrepudiation.

Key benefits include:

- Digital certificates eliminate cumbersome login and password dialog boxes when connecting to secure resources.
- Each party can be confident of the other's identity.
- Digital certificates ensure that only the intended recipient can read messages sent.
- Sophisticated access privileges and permissions can be built in, creating precise levels of authority for Internet transactions.

S/MIME message transmission uses certificate-based authentication and encryption to transmit messages between users and applications. S/MIME enables the exchange of confidential information without concerns about inappropriate access.

vCards provide a structured format for exchanging personal contact information with other users and applications, eliminating the need to retype personal information repeatedly.

JavaSoft's signed objects allow trusted distribution and execution of software applications and applets as part of an extranet. With signed objects, tasks can be automated and access to applications and services within the extended network can be granted based on capability. A digital certificate is used with a signed object to authenticate the identity of the publisher and grant appropriate access rights to the object.

EDI INT provides a set of recommendations and guidelines that combine existing EDI standards for transmission of transaction data with the Internet protocol suite. By using S/MIME and digital signatures, EDI transactions between enterprises can be exchanged in a secure and standard fashion.

Thus, open standards provide the most flexible, efficient, and effective foundation for enterprise networking.

3. NETWORK SECURITY ISSUES

Exposing of the organization's private data and networking infrastructure to the Internet crackers definitely increases concerns of the network administrators about the security of their networks [5]. It has been very well expressed that "security should not be a reason for avoiding cyberspace, but any corporation that remains *amateurish* about security is asking for trouble" [6].

Each user who needs external access has a unique set of computing and data requirements, and the solution will not and should not be the same for all [7]. Six basic extranet components can be identified as well as some specialized and hybrid solutions [7]: (1) access to the external resources; (2) Internet protocol (IP) address filtering; (3) authentication servers; (4) application layer management; (5) proxy servers; and (6) encryption services. Each is sufficient to initiate business communications, but each carries different performance, cost, and security. Namely, security (as recent statistics shows [8]) is the main issue preventing organizations from establishing extranets.

The term *security* in general refers to techniques for ensuring that data stored in a computer cannot be read or compromised. Obvious security measures involve *data encryption* and *passwords*. Data encryption by definition is the translation of data into a form that is unintelligible without a deciphering mechanism. A password, obviously, is a secret word or phrase that gives a user access to a particular program or system. Thus, in the rest of this section we focus on the following issues: risk assessment; development of the security policy; and establishment of the authentication, authorization, and encryption.

3.1. Risk Assessment

Risk assessment procedures should answer the following typical questions:

- What are the organization's most valuable intellectual and network assets?
- Where do these assets reside?
- What is the risk if they are subjected to unauthorized access?
- How much damage could be done — can it be estimated in terms of money?
- Which protocols are involved?

3.2. Security Policy

To provide the required level of protection, an organization needs a security policy to prevent unauthorized users from accessing resources on the private network and to protect

against the unauthorized export of private information. Even if an organization is not connected to the Internet, it may still want to establish an internal security policy to manage user access to portions of the network and protect sensitive or secret information. According to the FBI, 80% of all break-ins are internal.

Policy is the allocation, revocation, and management of permission as a network resource to define who gets access to what [9]. Rules and policy should be set by business managers, the chief information officer, and a security specialist — someone who understands policy writing and the impact of security decisions. Network managers can define policy for a given resource by creating an entry in access control lists, which are two-dimensional tables that map users to resources.

A firewall is an implementation of *access rules*, which are an articulation of the company's security policy. It is important to make sure that some particular firewall supports all the necessary protocols. If LANs are segmented along departmental lines, firewalls can be set up at the departmental level. However, multiple departments often share a LAN. In this case, the creation of virtual private network (VPN) for *each* person is highly advisable.

The following are recognized as basic steps for developing a security policy:

1. Assessment of the types of risks to the data will help to identify weak spots. After correction, regular assessments will help to determine the ongoing security of the environment.
2. Identification of the vulnerabilities in the system and possible responses, including operating system vulnerabilities, vulnerabilities via clients and modems, internal vulnerabilities, packet sniffing vulnerabilities, and means to test these vulnerabilities. Possible responses include encrypting data and authenticating users via passwords and biometrically.
3. Analysis of the needs of user communities:
   - Grouping data in categories and determining access needs. Access rights make the most sense on a project basis.
   - Determining the time of day, day of week, and duration of access per individual are the most typical procedures.

   Determination and assignment of the security levels can include the following, five levels:
   - Level one for top-secret data such as pre-released quarterly financials or a pharmaceutical firm's product formula database
   - Level two for highly sensitive data such as the inventory positions at a retailer
   - Level three for data covered by nondisclosure agreements such as six month product plans
   - Level four for key internal documents such as a letter from the CEO to the staff
   - Level five for public domain information

   To implement this security hierarchy, it is recommended that firewalls be placed at the personal desktop, workgroup, team, project, application, division, and enterprise level.
4. Writing the policy.
5. Development of a procedure for revisiting the policy as changes are made.
6. Writing an implementation plan.
7. Implementation of the policy.

3.3. Authentication, Authorization, Encryption

When it comes to the security aspects of teleworking and remote access, there is a tension between the goals of the participants in all aspects of information technology security. Users want access to information as quickly and as easily as possible, whereas information owners want to make sure that users can access only the information they are allowed to access. Security professionals often find it difficult to reduce this tension because of the demands of users in a rapidly changing business world. Smartcards may provide the solution to this problem [10].

Encryption requires introduction of the key management/updating procedure. Encryption can be implemented:

- *At the application*: Examples of this are Pretty Good Privacy (PGP) and Secure Multipurpose Internet Mail Extensions (S/MIME), which provide encryption for email.
- *At the client or host network layer*: The advantage of this approach is that it will provide extra protection for the hosts that will be in place even if there is no firewall or if it is compromised. The other advantage is that it allows distribution of the burden of processing the encryption among the individual hosts involved. This can be done at the client with products such as Netlock (see [11]), which provides encryption on multiple operating system platforms at the IP level. A system can be set up so that it will accept only encrypted communications with certain hosts. There are similar approaches from Netmanage and FTP Software.
- *At the firewall network layer*: The advantage to this approach is that there is centralized control of encryption, which can be set up based on IP address or port filter. It can cause a processing burden on the firewall, especially if many streams must be encrypted or decrypted. Many firewalls come with a feature called virtual private network (VPN). VPNs allows encryption to take place as data leave the firewall. It must be decrypted at a firewall on the other end before it is sent to the receiving host.
- *At the link level*: The hardware in this case is dedicated solely to the encryption process, thus offloading the burden from a firewall or router. The other advantage of this method is that the whole stream is encrypted, without even a clue as to the IP addresses of the devices communicating. This can be used only on a *point-to-point link*, as the IP header would not be intact, which would be necessary for routing.

Products such as those manufactured by Cylink [12] can encrypt data after they leave the firewall or router connected to a WAN link.

Extranet routers combine the functions of a VPN server, an encryption device, and a row address strobe [13]. The benefits of using the extranet routers include the network's ability to build secure VPNs and tunnel corporate Internet protocol traffic across public networks like the Internet. Management is much easier than it is in a multivendor, multidevice setup. Expenses are significantly lower because there is no need for leased lines.

3.4. Secure Virtual Private Networks

The goal of all VPN products is to enable deployment of logical networks, independent of physical topology [14]. That is the "virtual" part — allowing a geographically distributed group of hosts to interact and be managed as a single network, extending the user dynamics of LAN or workgroup without concern as to physical location.

Some interpret private simply as a "closed" user group — any virtual network with controlled access. This definition lets the term *VPN* fit a wide variety of carrier services, from traditional frame relay and ATM networks to emerging MPLS-based networks. Others interpret private as "secure" virtual networks that provide confidentiality, message integrity, and authentication among participating users and hosts. We focus on secure VPNs.

VPN topologies try to satisfy one of the three applications.

1. *Site-to-site connectivity*. Private workgroups can be provided with secure site-to-site connectivity, even when LANs that comprise that workgroup are physically distributed throughout a corporate network or campus. Intranet services can be offered to entire LANs or to a select set of authorized hosts on several LANs, for example, allowing accounting users to securely access a payroll server over network segments that are not secured. In site-to-site VPNs, dedicated site-to-site WAN links are replaced by shared network infrastructure — a service provider network or the public Internet.
2. *Remote access*. Low-cost, ubiquitous, and secure remote access can be offered by using the public Internet to connect teleworkers and mobile users to the corporate intranet, thus forming a Virtual Private Dial Network (VPDN). In remote access VPNs, privately operated dial access servers are again replaced by shared infrastructure — the dial access servers at any convenient ISP POP.
3. *Extranet or business-to-business services*. Site-to-site and remote access topologies can also be used to provide business partners and customers with economical access to extranet or business-to-business (B2B) services. In extranet and B2B VPNs, a shared infrastructure and the associated soft provisioning of sites and users makes it possible to respond more rapidly and cost-effectively to changing business relationships.

It is a current trend to outsource such VPN applications to a commercial service provider (e.g., regional ISP, top-tier network access provider, public carrier, or a service provider specializing in managed security services). In such outsourced VPNs, the service provider is responsible for VPN configuration (provisioning) and monitoring. Service providers may locate their VPN devices at customer premises or at the carrier's POP.

3.5. Use of Tunnels for Enabling Secure VPNs

Secure VPN applications are supported by secure, network-to-network, host-to-network, or host-to-host tunnels — virtual point-to-point connections. VPN tunnels may offer three important security services:

- Authentication to prove the identity of tunnel endpoints
- Encryption to prevent eavesdropping or copying of sensitive information transferred through the tunnel
- Integrity checks to ensure that data are not changed in transit

Tunnels can exist at several protocol layers.

Layer 2 tunnels carry point-to-point data link (PPP) connections between tunnel endpoints in remote access VPNs. In a compulsory mode, an ISP's network access server intercepts a corporate user's PPP connections and tunnels these to the corporate network. In a voluntary mode, VPN tunnels extend all the way across the public network, from dial-up client to corporate network. Two layer 2 tunneling protocols are commonly used:

- The point-to-point tunnel protocol (PPTP) provides authenticated, encrypted access from Windows desktops to Microsoft or third-party remote access servers.
- The IETF standard Layer 2 Tunneling Protocol (L2TP) also provides authenticated tunneling, in compulsory and voluntary modes. However, L2TP by itself does not provide message integrity or confidentiality. To do so, it must be combined with IPsec.

Layer 3 tunnels provide IP-based virtual connections. In this approach, normal IP packets are routed between tunnel endpoints that are separated by any intervening network topology. Tunneled packets are wrapped inside IETF-defined headers that provide message integrity and confidentiality. These IP security (IPsec) protocol extensions, together with the Internet Key Exchange (IKE), can be used with many authentication and encryption algorithms (e.g., MD5, SHA1, DES, 3DES). In site-to-site VPNs, a security gateway — an IPsec-enabled router, firewall, or appliance — tunnels IP from one LAN to another. In remote access VPNs, dial-up clients tunnel IP to security gateways, gaining access to the private network behind the gateway.

Companies with "email only" or "web only" security requirements may consider other alternatives, such as Secure Shell (SSH, or SecSH). SSH was originally developed as a secure replacement for UNIX "r" commands

(rsh, rcp, and rlogin) and is often used for remote system administration. But SSH can actually forward any application protocol over an authenticated, encrypted client–server connection. For example, SSH clients can securely forward POP and SMTP to a mail server that is running SSH. SSH can be an inexpensive method of providing trusted users with secure remote access to a single application, but it does require installing SSH client software.

A far more ubiquitous alternative is Netscape's Secure Sockets Layer (SSL). Because SSL is supported by every web browser today, it can be used to secure HTTP without adding client software. SSL evolved into IETF-standard Transport Layer Security (TLS), used to "add security" application protocols like POP, SMTP, IMAP, and Telnet. Both SSL and TLS provide digital certificate authentication and confidentiality. In most cases, SSL clients (e.g., browsers) authenticate SSL servers (e.g., e-Commerce sites). This is sometimes followed by server-to-client subauthentication (e.g., user login). SSL or TLS can be a simple, inexpensive alternative for secure remote access to a single application (e.g., secure extranet "portals").

Table 2 shows a comparison of these alternatives. Combining approaches is also possible and, to satisfy some security policies, absolutely necessary. L2TP does not provide message integrity or confidentiality. Standard IPsec does not provide user-level authentication. The Windows 2000 VPN client layers L2TP on top of IPsec to satisfy both of these secure remote access requirements. On the other hand, vanilla IPsec is more appropriate for site-to-site VPNs, and SSL is often the simplest secure extranet solution.

3.6. Summary of the Weak Points and Security Hazards

Table 2 provides a summary of the weak points in the system security, identifies and shows how these problems can be addressed, and suggests technical solutions for them. Additional information for various platforms can be obtained in Refs. 15–19.

4. COST OF RUNNING WEBSITES

Evolutionary scale of Websites suggested by the Positive Support Review Inc. of Santa Monica, California [20] includes the following:

1. *Promotional*: A site focused on a particular product, service, or company. Cost: $300,000 to $400,000 per year (17–20% on hardware and software, 5–10% on marketing, and the balance on content and servicing).
2. *Knowledge-based*: A site that publishes information that is updated constantly. Cost: $1 to $1.5 million annually (20–22% on hardware and software, 20–25% on marketing, and 55–60% on content and servicing).
3. *Transaction-based*: A site that allows surfers to shop, to receive customer services, or to process orders. Cost: $3 million per year (20–24% on hardware and software, 30–35% on marketing, and 45–50% on content and servicing).

A similar classification by Zona Research Inc. of Redwood City, California (cited in Ref. 21) divides Websites into:

1. *Static presence* (*"Screaming and Yelling"*). According to Zona Research, the page cost for such sites is less than $5,000. At present, the over whelming majority of Websites belong to this category.
2. *Interactive* (*"Business Processes and Data Support"*), with page costs ranging from $5,000 to $30,000. Perhaps 15 to 20% of all current Websites are in this category.
3. *Strategic* (*"Large-Scale Commerce"*), with dynamic pages that cost more than $30,000 each to produce

Table 2. VPN Protocol Comparison

Feature	L2TP	IPsec with IKE	SSL/TLS
System-Level Authentication	Control Session: Challenge/Response	Mutual Endpoint Authentication: Preshared Secret Raw Public Keys, Digital Certificates	Server Authentication: Digital Certificates
User-Level Authentication	PPP Authentication: PAP/CHAP/EAP	Vendor Extensions: XAUTH, Hybrid, CRACK, etc.	Client Sub Authentication: Optional
Message Integrity	None (use with IPsec)	IP Header & Payload: IPsec AH or ESP, Keyed Hash, HMAC-MD5, or SHA-1	Application Payload: Keyed Hash, MD5 or SHA-1
Tunnel Policy Granularity	Network Adapter: Tunnels all packets in PPP session, bidirectional	Security Associations: Unidirectional policies defined by IP address, port, user id, system name, data sensitivity, protocol	Application Specific
Data Confidentiality	None (use with IPsec)	IP Header & Payload: IPsec ESP, DES-CBC, 3DES, other symmetric ciphers	Application Stream: RC4, RC2, DES, 3DES, Fortezza
Compression	IPPCP	IPPCP	LZS

Source: Core Competence, (14), 2001.

and maintain. Currently, less than 0.5% of all Websites are in this category.

Thus, electronic commerce sites are not toys, and before entering these *WWWaters* organization must develop a clear idea about current and strategic investments to this part of business.

4.1. ISDN Cost Model

The mobile workers should look at ISDN as an important technology for building extranet infrastructure because of its ability to support flexible access. The ISDN cost model should consider the following:

- Service fees from local telecom for each location (installation + monthly + per minute)
- Long-distance charges, if applicable
- Cost of equipment (NT1 + TA, NT1 + bridges, NT1 + routers, etc.)
- Cost of Internet Service Provider (ISP) services, if applicable

Thus, planning a budget for a month, for example, will include \$30 a month + 3 cents per minute per channel. Therefore, it is \$3.60 per hour for two B connections. If the user needs to be connected three hours per day, 20 days per month, it will cost \$16 + \$30 = \$246 for a month of service, just for the local telecom portion of your ISDN connection. Long-distance fees and ISP charges naturally would be added on the top.

4.2. Mobile Connection Cost Model

Mobile workers can consider wireless communications as another option that can be highly cost-effective, but its costs generally are *higher* than wireline communications. It should be mentioned that with packet data the modem occupies the radio channel only for the time it takes to transmit that packet. In ordinary data networks, users are usually billed for the amount of data they send. In contrast, for data communication over a wireless link, there is established a *circuit connection* and *payment is based on the duration* of the call just as with an ordinary voice call. The per-minute charges are usually the same.

Wireless modems are complex electronic devices containing interface logic and circuitry, sophisticated radios, considerable central processing power, and digital signal processing. As such, they cost more than landline modems.

4.3. Cost of Downtime

Electronic commerce is difficult in case of unreliable infrastructure. In this section, we use an example from Ref. 22 to examine the cost of downtime for some consumer-oriented business, such as an airline or hotel reservation center. If customers have a choice, they will call a competitor and place their order there.

We will use an example where customer service center has a staff of 500 people, each of whom carries a burdened cost of \$25 an hour. They make an average of 60 transactions per hour and an average of three high-priced sales per hour. Hours of operation are 24 hours a day, seven days a week, 365 days a year.

In actuality, the line managers of the site, not the IT staff, should calculate the costs of downtime. This information often is not forthcoming, however. So, this example can be presented to give a general sense of the impact that downtime has on the company's finances as well as a guideline for estimation of the cost of outages.

The cost of outages in the hypothetical network with an availability rate of 99.9% is about \$500,000 a year (see Table 3 to Table 5). If hardware and software necessary to do the job have already been bought, this estimate can be considered as a guideline for the additional budget to be spent on providing redundancy. This is separate and apart from the funds required to provide a base level of network functionality. Thus, it really is not worth rushing headlong into designing a fault-tolerant network unless all parties agree on all the implications that downtime has on the operation.

Table 3. Weak Points and Security Hazards

Weak Point/Hazard	Technical Solution
Operating system/ applications on servers	Research vulnerabilities; Monitor CERT advisories; Work with vendors; Apply appropriate patches or remove services/applications; Limit access to services on host and firewall; Limit complexity
Viruses	Include rules for importing files on disks and from the Internet in security policy; Use virus scanning on client, servers, and at Internet firewall
Modems	Restrict use; Provide secured alternatives when possible (such as a dial-out only modem pool)
Clients	Unix: Same as server issues above; Windows for Workgroups, Win95, NT: Filter TCP/UDP ports 137,138,139 at firewall; Be careful with shared services, use Microsoft's Service Pack for Win95 to fix bugs
Network snooping	Use encryption; Isolate networks with switch or router
Network attacks	Internet firewall; Internal firewall or router; Simple router filters that do not have an impact on performance
Network spoofing	Filter out at router or firewall

Table 4. Cost of Outages

Downtime Percentage	0.9990
Number of hours/year	8.76
Number of employees	500
Average burdened cost	825
Idle sale	\$109,500
Impact to production	\$131,400
Opportunity lost	\$262,500
Total downtime impact	\$503,700

Table 5. Impact to Production

Profit Per Transaction	0.5
Transactions per hour per employee	60
Missed transactions per hour	30,000
Total missed transactions	262,800
Impact of missed transactions	$131,400

Table 6. Opportunity Lost

Profit Per Sale	20
Sales per hour per employee	3
Missed sales per hour	1,500
Total missed sales	13,140
Impact of missed sales	$262,600

5. INTRANET AND EXTRANET TRENDS

Mindbridge.com Inc. has been analyzing and describing trends typical for intranets [1]. We suggest that these finding are also applicable to extranets.

Trend 1: Shifting the focus around the customer. Enhancing and simplifying customer interactions with the supplier or dealer has become a major focus of many industries with the proliferation of the Internet. The primary focus has been to increase the success of the customer, which in turn creates a greater sense of loyalty and increases profits.

Trend 2: Automate routine tasks. No longer is it necessary to fill out and file through request forms for daily routine activities. Now employees can complete requests for products such as office supplies when they log onto the intranet or extranet site and click on what is needed. This saves not only time but also considerable money by allowing the purchasing department to order in bulk, thus acquiring a rebate on such orders.

Trend 3: Delivering information where it is needed. No longer is location an issue. Organizations can deliver information where it is needed by an upload to the network server. Communications have become streamlined and efficient as virtual teams are now connected and able to collaborate without concern to time or distance.

Trend 4: The acceptance of the intranet or extranet by top management. Now that the intranet or extranet has had a chance to prove itself, top management and many businesses are buying in. It has been identified as a critical part of the organization, and many have moved to improve and expand their existing intranet or extranet capabilities.

Trend 5: More interesting features. Web construction kits, tools that help customers create their own homepages, and tickers that track the status of major projects are just a few of the new and interesting features being added to the intranet or extranet. This should continue as the intranet or extranet becomes an increasingly vital tool for doing business.

Trend 6: Knowledge management is growing. Knowledge management is growing, which is helping company communications between parts and developing a set of tools to help the staff find out who is doing what. These tools include a project registry, a technology and a methodology library, employee profiles, and discussion areas for consultants. Building these communities creates groups who can help each other out so as not to repeat mistakes previously made or to prevent others from reinventing the wheel.

Trend 7: New business opportunities and revenues. As time continues and organizations continue to implement and enhance their intranet or extranet technologies, capturing and sharing information will become more efficient. New opportunities can arise from this effectiveness, such as learning that the organization had internal skills that were not obvious, leading it to extend its offerings.

Trend 8: Enhancing learning environments. The intranet or extranet has entered the educational arena as it has become feasible to do so with the vast numbers of adolescents online. This will empower educators to develop their own intranet or extranet pages and update them accordingly. This will increase student–teacher interaction and develop a stronger and more productive relationship. Distance learning, particularly in regard to overseas students, will also be affected, as the boundaries of the classroom walls will continue to fade.

CONCLUSION

Currently, the extranet is conceptualized as the key technology that can enable development of the third wave of electronic commerce sites. Although technical and cost advantages are of very high importance, the real significance of the extranet is that it is the first nonproprietary technical tool that can support rapid evolution of electronic commerce.

It is already clear that the Internet impacted retail sales, the use of credit cards, and various digital cash and payment settlement schemes. However, the experts predict that the real revolution will be in systems for global procurement of goods and services at the wholesale level and that the role of extranets is crucial for this. It is also expected that on a more fundamental level the extranets will stimulate the business evolution of conventional corporations into knowledge factories.

BIOGRAPHY

Algirdas Pakštas received the M.Sc. degree in radiophysics and electronics in 1980 from Irkutsk State University, Irkutsk, Russia; Ph.D. degree in system programming from the Institute of Control Sciences in 1987; and DrTech degree from the Lithuanian Science Council in 1993, respectively. He first joined the Siberian Institute of Solar-Terrestrial Physics (SibIZMIR) in 1980. At SibIZMIR he worked on the design and development of distributed data acquisition and visualization systems. Since 1987, he has been a senior research scientist and later head of the department at the Institute of Mathematics and Informatics, Lithuanian Academy of Sciences, where he has been working on software engineering for distributed computer systems. From 1991 to 1993, he worked as research fellow at the Norwegian University of Technology (NTH),

Trondheim, and from 1994 to 1998 as professor and later full professor at the Agder University College, Grimstad, Norway. Currently, Pakstas is with London Metropolitan University, London, England. He has published 2 research monographs, edited 1 book, and authored more than 130 publications. His areas of interest are communications software, multimedia communications, as well as enterprise networking and applications.

BIBLIOGRAPHY

1. Mindbridge, (2000). Intranet WhitePapers. [Online]. Mindbridge.com Inc. *http://www.mindbridge.com/whitepaperinfo.htm* [2001, August 2].
2. R. H. Baker, *Extranets: The Complete Sourcebook*, McGraw-Hill Book Company, New York, 1997.
3. P. Q. Maier, Implementing and supporting extranets, *Information-Systems-Security* **7**(4): 52–59 (1997).
4. R. R. Reisman, (1997, March 21). Extranets and intergroupware: a convergence for the next generation in electronic media-based activity. [Online]. Teleshuttle Corporation. *http://www.teleshuttle.com/media/InterGW.htm* [2001, August 2].
5. R. Herold and S. Warigon, Extranet audit and security, *Computer Security Journal* **14**(1): 35–40 (1998).
6. J. Martin, *Cybercorp: the New Business Revolution*, Amacom Book Division, New York, 1996.
7. S. Trolan, Extranet security: what's right for the business?, *Information-Systems-Security* **7**(1): 47–56 (1998).
8. T. Lister, Ten commandments for converting your intranet into a secure extranet, *UNIX Review's: Performance Computing* **16**(8): 37–39 (1998).
9. R. Thayer, Network security: locking in to policy, *Data Communications* **27**(4): 77–8 (1998).
10. D. Birch, Smart solutions for net security, *Telecommunications* **32**(4): 53–54, 56 (1998).
11. Netlock Version 1.4.1, (1997). [Online]. Interlink Computer Sciences, Inc. *http://www.interlink.com/NetLOCK/* [1998, September 21].
12. Cylink Corporation, (1998). Global Network Security Products. [Online]. *http://www.cylink.com/products* [2001, August 2].
13. E. Roberts, Extranet routers: the security and savings are out there, *Data Communications* **27**(12): 9 (1998).
14. L. Phifer, (2001, April 12). VPNs: Virtually Anything? [Online]. Core Competence. *http://www.corecom.com/html/vpn.html* [2002, May 17].
15. S. Castano, ed., *Database Security*, Addison Wesley Publishing Co. - ACM Press 1995.
16. R. Farrow, *Unix Systems Security*, Addison-Wesley Publishing Co 1991.
17. TruSecure Media Group (2000). TruSecure publications. [Online]. *http://www.trusecure.com/html/tspub/index.shtml* [2001, August 2].
18. ALC Press, (2001) Linux boot camp. [Online]. *http://alcpress.com/* [2001, August 2].
19. S. A. Sutton, *Windows NT Security Guide*, Addison Wesley Developers Press 1995.
20. V. Junalaitis, The true cost of the Web, *PC Week* **18**: 85 (Nov. 1996).
21. E. Shein, Natural selection, *PC Week* **14**: E2 (Oct. 1996).
22. B. Walsh, (1996). Fault-Tolerant networking. [Online]. CMP Net. *http://techweb.cmp.com/nc/netdesign/faultmgmt.html* [1997, September 21].

IP TELEPHONY

MATTHEW CHAPMAN CAESAR
University of California at Berkeley
Berkeley, California

DIPAK GHOSAL
University of California at Davis
Davis, California

1. INTRODUCTION

IP telephony systems are designed to transmit packetized voice using the *Internet Protocol* (IP). IP telephony networks can offer video, high-quality audio, and directory services, in addition to services traditionally offered by the public switched telephone network (PSTN). Enterprises implement IP telephony networks to avoid the cost of supporting both a circuit switched network and a data network. Individual users save money by using IP telephony software to bypass long distance charges in the PSTN. IP telephony makes use of statistical multiplexing and voice compression to support increased efficiency over circuit-switched networks. However, many IP networks do not support quality-of-service (QoS) guarantees. Care must be taken to design these networks to minimize factors such as packet delay, jitter, and loss that decrease the quality of the received audio [29].

IP telephony services can be implemented over campus networks and wide-area networks (WANs). In a campus, a converged network with IP telephony services can take the place of a private branch exchange (PBX). These networks usually consist of one or more Internet telephony gateways (ITGs) to interface with the PSTN and a set of intelligent telephony devices offering users connectivity to the ITG. Campus networks can be interconnected over a WAN to more efficiently utilize leased lines and to bypass tariffs imposed by regulatory agencies.

The functionality of IP allows for IP telephony systems to offer a wide variety of services in addition to those offered by the PSTN. From a telephony provider's perspective, IP telephony allows integration of data, fax, voice, and video onto a single network resulting in cost savings. Converged networks eliminate the need to purchase equipment for different types of traffic, and derive increased efficiency by allowing the bandwidth to be shared among the various traffic types. From a user's perspective, the increased intelligence of the network endpoints allows for next-generation services to be implemented and quickly deployed [4,5,8]. Examples of these services include caller ID, higher quality of sound, video telephony, unified messaging, Web interfaces to

call centers, virtual phone lines, and real-time billing. The increased capabilities of the endpoints allow for better user interfaces than can be supported in the PSTN. Cable providers could deploy an interactive set-top box to offer directory services and caller ID on the television screen [30]. Unlike the PSTN, new services can be deployed without making expensive, time-consuming modifications to the network. In addition, third parties may deploy services on an IP telephony service provider's network.

Several difficulties can arise in deploying IP telephony networks:

1. Delay, loss, and jitter can decrease audio quality in IP networks. Audio quality can be improved by increasing network bandwidth, implementing QoS enhancements inside the network, applying protocol enhancements such as forward error correction (FEC) to send redundant data, and using strategies to limit and load balance congestion in the network.
2. Some standards for IP telephony are complex, partially developed, or vague. Many companies are resorting to proprietary protocols that limit interoperability. Thirdly, in order to recover the cost of deployment and recover their huge sunken costs in PSTN equipment, providers need to implement competitive and economically feasible pricing strategies.
3. It is not clear how IP telephony will be regulated. IP telephony traffic may be unregulated, tariffed, or even outlawed by national governments [19].

This article gives an overview of IP telephony. The architecture and the related protocols are discussed in Section 2. Section 3 describes several common deployment scenarios. Delivering good audio quality over best effort IP networks is a challenging task, and Section 4 highlights the key issues. Section 5 discusses design issues of the various entities in the IP telephony architecture. Section 7 gives an overview of security considerations, while Section 8 lists regulatory agencies and standardization bodies guiding and accelerating the deployment of IP telephony solutions.

2. ARCHITECTURE AND PROTOCOLS

A typical IP telephony architecture is shown in Fig. 1. The key network elements include the *terminal*, the *IP telephony gateway* (ITG), *the multipoint control unit* (MCU), and the *gatekeeper*. Terminals may be a specialized device such as an IP phone, or a personal computer (PC) equipped with IP telephony software. Terminals form the endpoints of an IP telephony call. The ITG provides a bridge between the IP network and the PSTN, which uses the Signaling System 7 (SS7) protocol to set up and tear down dedicated circuits to service a call. A gatekeeper is used to support admission control functionality, while the MCU enables advanced services such as multiparty conference calls.

The two key standards for real-time voice and video conferencing are (1) *H.323*, which is developed by the International Telecommunications Union's Telecommunication Standardization Sector (ITU-T) [14]; and (2) *Session Initiation Protocol* (SIP), which is developed by the Internet Engineering Task Force (IETF) [15]. There is some overlap between these two standards and combinations of the two may be used. Many vendors are moving to support both these protocols to increase interoperability, and several application programming interfaces (APIs) have been developed to speed development of IP telephony systems [3,27,28].

2.1. Functional Entities

2.1.1. H.323. H.323 operations require interactions between several different components, including terminals, gateways, gatekeepers, and multipoint control units.

2.1.1.1. Terminals. IP telephony terminals are able to initiate and receive calls. They provide an interface for a user to transmit bidirectional voice, and optionally video or data: (1) the terminal must use a compressor/decompressor (codec) to encode and compress voice for

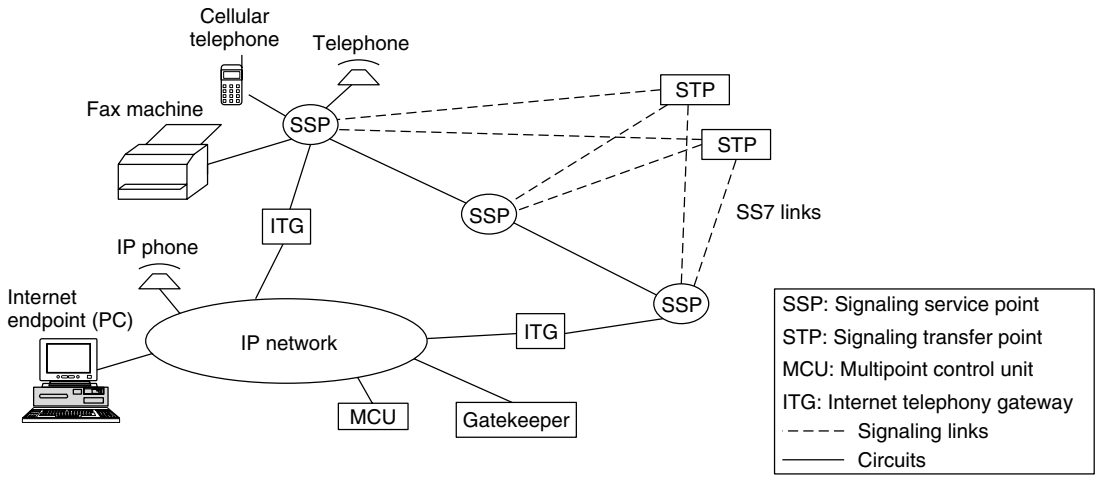


Figure 1. IP telephony architecture.

transport across the network — the receiving host must decode and decompress the voice for playback to the user; (2) terminals must perform signaling to negotiate codec type, data transfer speed, and perform call setup procedures; and (3) the terminal may provide an interface to choose the remote host and services desired. The interface may also display real-time feedback as to the connection quality and price charged.

H.323 supports several codecs for voice and video in the terminal, including G.711, G.723.1, and G.722. Codecs vary in bit rate, processing delay, frame length and size, memory and processor overhead, and resulting sound quality [21].

H.323 networks can support a wide variety of end systems. A PC may be configured to act as an IP telephony end system by installing appropriate client software. Many of these clients may be downloaded free of charge, but may require multimedia hardware such as a microphone, speakers, and possibly a videocamera. Specialized sound cards are available that allow the use of a standard telephone in place of a microphone. Performance may be improved and software cost reduced by installing a card to perform voice encoding in hardware. PC-less solutions are also available; specialized IP phones are manufactured that can make voice calls across a campus data network. Standard PSTN telephones may be equipped with an adapter to support similar functionality.

2.1.1.2. Gateway. A gateway connects IP networks to other types of networks, such as the PSTN: (1) it uses a signaling module to translate signaling information such as call setup requests from one network to another, (2) it uses a media gateway to perform translations between different data encoding types, and (3) it allows for real-time control of the data stream through the use of a media controller. The signaling module and the media controller are sometimes collectively referred to as the *signaling gateway*. Home users may purchase telephony cards to transform their PC into a gateway. More commonly, gateways are sold as specialized devices and deployed in service provider networks.

For example, a gateway might bridge the PSTN network to an H.323 network by translating PSTN Signaling System 7 (SS7) into H.323 style signaling, and encoding PSTN voice. Endpoints communicate with the gateway using H.245 and Q.931. The PSTN user dials a phone number to connect to the gateway. Address translation may be performed by requesting a name or extension from the user, and forming a connection to the appropriate IP address. An IP user may call a PSTN user by requesting the gateway to dial the appropriate phone number.

Telecommunications companies (telcos) are often hesitant to expose their SS7 network because of its critical nature. A preferred deployment method is to install the signaling gateway at the telco premises and allow it to control media gateways in data service provider networks [7,21].

2.1.1.3. Gatekeeper. A gatekeeper is responsible for managing and administrating the initiated calls in the H.323 network. It is often implemented as part of a gateway. Each gatekeeper has authority over a portion of the H.323 network called a *zone*. It performs admissions control of incoming calls originating in the zone based on H.323 guidelines and user defined policies. These policies may control admission based on bandwidth usage, network load, or other criteria. The gatekeeper is also responsible for address translation from alias addresses such as an H.323 identifier, URL (Uniform Resource Locator), or email address to an IP address. The gatekeeper maintains a table of these translations, which may be updated using the registration–admission–status (RAS) channel. The gateway also monitors the bandwidth and can restrict usage to provide enhanced QoS to other applications on the network.

Several enhancements may be made to gatekeepers to improve utility:

1. It may act as a proxy to process call control signals rather than passing them directly between endpoints.
2. Calls may be authorized and potentially rejected. Policies may be implemented that reject a call based on a user's access rights, or to terminate calls if higher priority ones come in.
3. The gatekeeper may send a billing system call details on call termination.
4. The gatekeeper may perform call routing to collect more detailed call information, or to route the call to other terminals if the called terminal is not available. The gatekeeper may perform load balancing strategies across multiple gatekeepers to avoid congestion.
5. The gatekeeper may limit the bandwidth of a call to less than what was requested to decrease network congestion.
6. The gatekeeper may keep a database of user profiles for directory services. Additional supplementary services such as call park/pickup may also be implemented with a gatekeeper.

2.1.1.4. Multipoint Control Unit (MCU). A *multipoint control unit* (MCU) is used to manage conference calls between 3 or more H.323 endpoints. It may be implemented as a standalone unit, as part of the gateway, or as part of the terminal. It performs mixing, transcoding, and redistribution operations on a set of media streams.

2.1.2. SIP. The SIP architecture [15] is composed of terminals, proxy servers, and redirect servers. Unlike H.323, terminals are the only required entities. Terminals have functionality similar to that of an H.323 terminal. Servers are used to implement directory services. Redirect servers inform the calling SIP terminal the location of the called party, and the terminal may then connect directly to that location. Proxy servers act as an intermediary for the connection, and forward the call setup request to a new location. SIP allows any codec registered with the Internet Assigned Numbers Authority (IANA) to be used, in addition to the ITU-T codecs supported in H.323.

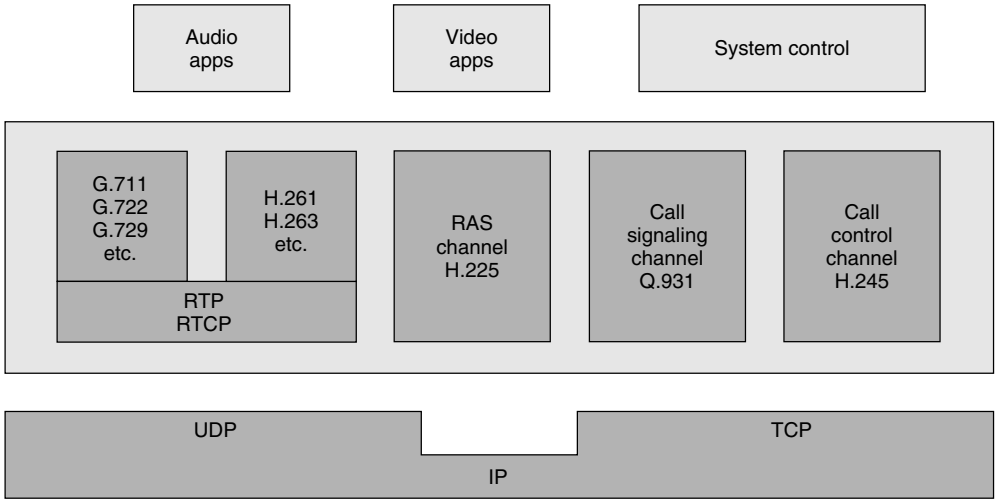


Figure 2. The H.323 protocol architecture.

2.2. Protocols

2.2.1. H.323. H.323 is an umbrella specification developed by the ITU-T which defines several protocols for IP telephony. The overall protocol architecture is shown in Fig. 2 [1]. *Q.931* is specified for call signaling, which is used to establish a connection between two terminals or between a terminal and a gateway; Q.931 implements traditional telephone functionality by exchanging signals for call setup and termination. *H.245* signaling is used for controlling media between endpoints. The control messages can be used to exchange capabilities, start and terminate media streams, and send flow control messages. *H.225* RAS is used for communication between a gatekeeper and an H.323 endpoint. RAS channels are established over the User Datagram Protocol (UDP) and are used to perform admission control, status queries, bandwidth control, and management of connections to gatekeepers. The *Real-Time Transport Protocol* (RTP) used to transport and encapsulate media streams between endpoints [13]. RTP provides payload type identification, sequence numbers, and timestamps. It is often used in conjunction with the Real-Time Transport Control Protocol (RTCP), a protocol that allows receivers to provide connection feedback information regarding the quality of the connection. *H.323* defines a set of codecs to be used to encode voice and video. These codecs vary by processor and memory utilization, bit rates, and resulting audio quality. At a minimum, an H.323 system must support the G.711 codec, RAS, Q.931, H.245, and RTP/RTCP. Finally, it specifies negotiation of codec type and supplementary services. An H.323 network must support reliable communication for control signaling and data, unreliable communication for the RAS channel, and voice and video data.

A sample H.323 call is shown in Fig. 3. Suppose that user A at terminal 1 (T1) wishes to contact user B at terminal 2 (T2). First, T1 sends an admission request (ARQ) over the RAS channel to the zone's gatekeeper asking for the IP address where user B can be reached. The gatekeeper chooses to accept the call by returning an admission confirm (ACF) with T2's

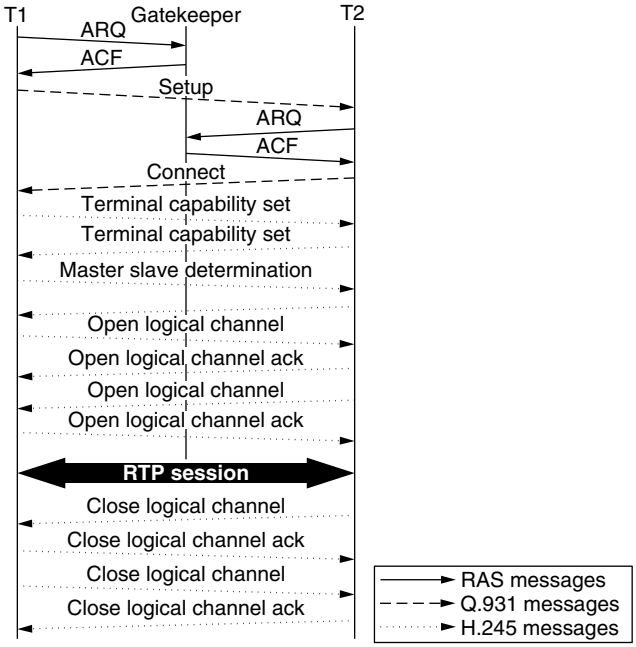


Figure 3. A sample H.323 call setup and teardown.

IP address. T1 opens a Q.931 channel to T2 over the Transmission Control Protocol (TCP) and sends a setup message. When T2 receives the message it sends an ARQ over RAS to the gatekeeper requesting permission to communicate with T2. The gatekeeper replies with an ACF and T2 returns a connect to T1 containing the TCP address it wishes to use for the H.245 channel. T1 opens an H.245 TCP channel to the specified port, and capabilities are exchanged. Either side may initiate a master/slave determination procedure or a capabilities exchange procedure. Both terminals must perform a capabilities exchange but only one master/slave determination is required. At the completion of these exchanges, logical channels may then be opened. T1 sends an *open logical channel* over the H.245 channel

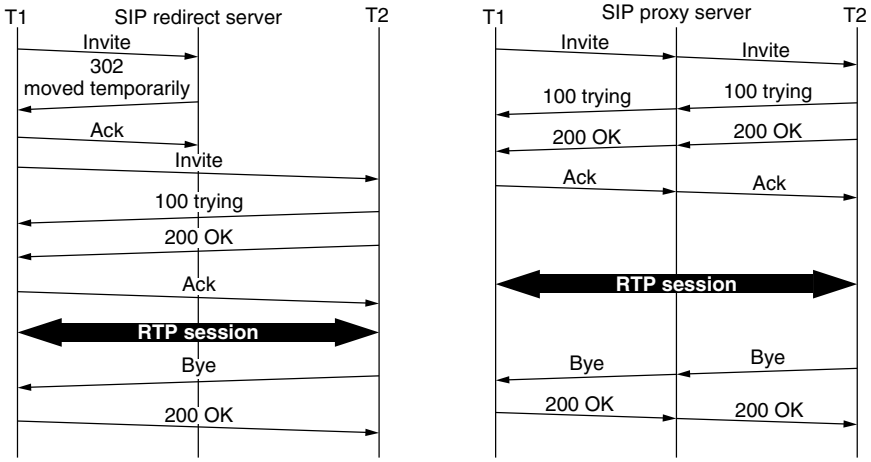


Figure 4. Sample SIP call establishment and teardown.

to request an audio channel. This message contains the UDP ports it wishes to use for the RTP audio stream and RTCP messages. T2 acknowledges this and attempts to set up a channel in the reverse direction to complete the bidirectional call, and RTP flows are established. Eventually, user B wishes to terminate the call, and causes T2 to send a *close logical channel*, which is acknowledged by T1. T1 sends a similar message to T2, and both sides send disengage requests to the gatekeeper.

2.2.2. SIP. Like H.323, SIP is a standard for signaling and control for IP telephony networks. SIP may be used in conjunction with H.323. SIP is based on HTTP, and provides call setup and termination, call control, and call data transfer. SIP is a client/server protocol, and can operate over reliable and unreliable transport protocols. SIP 2.0 contains client requests to invite a user to join a session, to acknowledge a new connection, to request the server's capability information, to register a user's location for directory services, to terminate a call, and to terminate directory searches. SIP uses the Session Description Protocol (SDP) for session announcement and invitation, and to choose the type of communication [43].

There are several differences between SIP and H.323. First, H.323 uses a binary format to exchange information. SIP is text-based, making parsers easier to build and the protocol easier to debug. Although H.323's format reduces some bandwidth and processing overhead, this overhead is minimal given that signaling and control happens infrequently during a call. SIP is considered to be more adaptable, as H.323 is required to be fully backward compatible. SIP exchanges fewer messages on connection setup, and hence can more quickly establish connections. H.323 may use only ITU-T developed codecs, whereas SIP may use any codec registered with IANA. Furthermore, SIP is considered to be more scalable, allows the servers to be stateless, provides loop detection, and has a smaller call setup delay. On the other hand, following the early publication of the standard, H.323 has been more widely supported. Furthermore, the H.323 specification provides improved fault tolerance as well as more thorough coverage of data sharing, conferencing, and video. Recent revisions of H.323 are addressing some of its shortcomings.

A sample SIP call through a redirect server is shown in Fig. 4. Suppose user A at T1 wants to contact user B at T2. T1 sends an "Invite" message to the SIP server running in redirect mode to indicate user B is being invited to participate in a session. This message usually contains a session description written in SDP including the media types T1 is willing to receive. The redirect server contacts a location server to lookup user B's current address, and returns a 302 "Moved temporarily" message containing the location of user B. T1 then sends a second invite message to T2. When T2 receives the message it sends back a 100 "trying" message and tries to alert the user. When user B picks up, T2 sends a "200 OK" message containing its capabilities. T1 responds with an acknowledgement (Ack) and RTP flows are established, completing the call. Either side may send a "Bye" message to terminate the call, which must be answered with a 200 OK.

A call handled by a SIP proxy server is also shown in Fig. 4. All messages pass through the server, and the calling terminal exchanges connection setup messages as if it is connecting directly to the callee's terminal. This scheme can cause the SIP server to become a bottleneck, limiting scalability of the IP telephony network. However, proxy servers allow for the accurate measurement of usage necessary for billing services. Furthermore, the use of a proxy server can enhance privacy by hiding each party's IP address and identification.

3. DEPLOYMENT SCENARIOS

In order to understand how an IP telephony system is constructed, it is useful to review several common scenarios for voice telephony. We will note how IP telephony can affect each of these service markets.

3.1. Data Service Provider

Data service providers can provide telephony services to increase revenue, and can provide voice with advanced supplementary services and a regulatory price advantage over the PSTN carriers. Traditionally this has been done by becoming a competitive local exchange carrier (CLEC) and installing PSTN services in rented space in a central

office (CO). Today, data service providers that wish to offer voice services can implement IP telephony solutions on top of their existing data networks, eliminating the need to purchase equipment for and train personnel to run a separate PSTN network. The unreliability of IP networks requires an investment in upgrading network element software and hardware to increase availability and service quality. To succeed, data service providers must convince the public that IP telephony is not a low-quality service, reengineer their networks to give good service quality under high load, and implement reliable account control and billing systems [6].

The major players in the data service provider space include companies that deploy data lines to customer premises, and those that provide long-haul data transport services. Many cable television (CATV) providers use their hybrid fiber coaxial (HFC) networks for bidirectional data transfer to subscribers. These networks consist of an optical node that converts optical signaling from the control center transmitter to electrical signaling on coaxial cable. Instead of deploying specialized HFC telephony equipment, CATV providers are considering using Voice over IP (VoIP) to decrease cost and attract customers with a set of supplementary services [30]. Problems that need to be solved include updates to the data over cable service interface specification (DOCSIS) to support voice requirements, and forming interconnection agreements to more effectively route the IP telephony calls. Also, IP telephony support will need to be deployed through DOCSIS-compliant products at the client and inside the network. Because of the large investment in circuit switched products and the strict QoS requirements of voice traffic, many CATV providers favor initially using IP telephony for local calls. The revenue generated may then be used to implement a nationwide IP telephony network. Finally, providers will need to decide between an all-IP-based telephony solution or a hybrid IP/circuit switched solution involving a network call signaling gateway (NCSG). Other providers that deploy data connections directly to the subscriber, like digital subscriber line (DSL) and wireless data service providers, face similar issues.

Internet service providers (ISPs) are also considering implementing IP telephony solutions in their networks. ISPs provide long-haul data transport. There are several types of ISPs, including backbone ISPs and access ISPs. A typical access ISP consists of *points of presence* (POPs) to which users may connect through modem, DSL, cable, Integrated Services Digital Network (ISDN), or other types of leased lines. Access ISPs can terminate IP telephony calls at gateways placed at one or more POPs. IP telephony is expected to significantly increase the operating costs of an ISP, so it is critical that the ISP implement good pricing strategies to recover these costs. Backbone ISPs provide a long-distance network to interconnect access ISPs. Backbone ISPs can expect to generate more revenue with the introduction of IP telephony services because of the resulting increase in data traffic carrying voice flowing between POPs. There are several pure Internet telephony service providers (ITSPs) that lease capacity from existing data network to provide IP telephony service. ITSPs have recently entered the market and are facing similar issues.

3.2. Local Exchange Carrier (LEC)

Local exchange carriers (LECs) provide local telephone service. Phone lines run from the customer premises to a CO. COs are connected in a *local access and transport area* (LATA) or to an interexchange carrier (IXC) for long-distance traffic. LECs already provide voice services to a large number of customers, giving them a significant advantage over data service providers in the IP telephony market. LECs also have well-developed operations for marketing, billing, and complying with federal regulations. Implementing an IP telephony network can improve a LEC's efficiency by eliminating congestion points in and decreasing costs associated with their PSTN network.

LECs face several hurdles in implementing an IP telephony solution. LECs have a large investment in circuit switched voice services, and implementing VoIP diverts revenue away from these services. Furthermore, in the United States, LECs are required by law to subsidize certain users, and hence have higher operating costs. Also the Regional Bell Operating Companies (RBOCs) are forbidden by law from offering long distance services. This may interfere with their ability to deploy large-scale IP telephony networks. Finally, LECs must build a reliable, overprovisioned data network to handle the stringent requirements of voice traffic.

3.3. Interexchange Carrier (IXC)

IXCs are long distance carriers. They provide transport for inter-LATA calls. Like the LECs, IXCs lack an economic incentive to implement IP telephony networks, due to their large investment in PSTN equipment. IXCs must also build a large-scale management architecture for billing, provisioning, and monitoring [5] and face similar problems in developing or partnering with the provider of a data network scalable enough to handle a large number of calls.

IP telephony can provide considerable benefits to an IXC:

(1) IXCs are required by law in several countries by law to pay a minimum fixed price per minute to the LECs terminating and originating the long-distance call — costs may be reduced by using a data network to bypass the LECs on each end of the call;
(2) IP telephony allows IXCs to implement a bundled solution for nationwide data, video, and voice transport with supplementary services;
(3) cost savings may be achieved by multiplexing PSTN-originated voice calls over a nationwide data network.

3.4. IP Host to IP Host

Two users may install IP telephony software on their computers to communicate. If the call does not need to pass through the PSTN, there is no need for any sort of

network intelligence, and only plain data services need to be purchased from the service provider. It is not clear how an ISP should treat calls that do not require transit on the PSTN network. Although these calls increase network load, the ISP may not be able to distinguish VoIP traffic, and hence may be unable to charge separately for IP telephony services. However, users are often willing to pay more for high-quality connections in IP telephony networks, and so ISPs may wish to charge more for increased QoS [32]. Furthermore, the IP hosts may wish for part of the call path to be routed over the PSTN, improving QoS.

4. ISSUES OF PACKET AUDIO OVER IP NETWORKS

IP networks do not provide the QoS guarantees found in the PSTN. Hence, they are inherently limited in meeting the requirements of voice transport. However, proper network and end system design can provide an infrastructure with good end-to-end voice quality.

4.1. Factors Affecting Audio Quality

Several factors affect the quality of the received audio. Delay and echo can interfere with normal conversations and make communication over the network unpleasant. Packet-switched networks suffer from packet jitter and loss, which further decrease audio quality.

4.1.1. Delay. Delay is the time for the voice signal to be encoded, sent through the network, and decoded at the receiver. One way delays greater than 150 ms cause the listener to hesitate before responding, making the mood of the conversation sound cold and unfriendly [25]. PSTN delays are typically 30 ms, whereas Internet delays tend to range from 50 to 400 ms. Delay in IP telephony networks is caused by IP network delays and signal processing delays.

There are five types of IP network delays: propagation delays, packet capture delays, serialization delays, switching delays, and queuing delays. *Propagation delay* is the time for the signal to propagate through a link, and is a function of transmission distance. *Packet capture delay* is the time required for the router to receive a complete IP packet before processing it and forwarding it towards the destination. *Serialization delay* is the time it takes to place a packet on the transmission link. *Switching delay* is the time for the router to choose the correct output port on the basis of information contained in the IP packet header. *Queuing delays* occur when packets are buffered at the router while other packets are being processed.

Signal processing delays occur in the codec algorithms implemented at the gateway or end-user systems. Codecs perform encoding and compression operations on the analog voice signal, resulting in digital signal processing (DSP) delays. These delays include the time to detect dual-tone multifrequency (DTMF) tones, detect silence, and cancel echo. DSP algorithms depend on processing an entire frame at a time, and the time to fill one of these frames before passing it to the DSP algorithm increases delay. The time for a transmitter to fill a packet with voice data is the packetization delay. Using shorter packet sizes decreases this delay, but increases network load, as more bandwidth must be expended on sending packet headers. Some coders examine the next frame to exploit any correlation with the current frame, causing lookahead delays.

4.1.2. Loss. IP network congestion can cause router buffers to overflow, resulting in packet loss. Unlike the PSTN, no end-to-end circuits are established, and IP packets from many links are queued for transmission over an outgoing link. Packets are dropped if there is no space in the queue. Packet loss interferes with the ability for the receiving host's codec to reconstruct the audio signal. Each packet contains 40–80 ms of speech, matching the duration of the phoneme, a critical phonetic unit used to convey meaning in speech. If many packets are lost, the number of lost phonemes increases, making it more difficult for the human brain to reconstruct the missing phonemes and understand the talker. Packet loss in the Internet is improving, and even long distance paths average less than 1% packet loss.

4.1.3. Jitter. Random variation in packet interarrival times at the receiver is referred to as *jitter*. Jitter occurs as a result of the variance in queueing delays at IP network routers. If this delay is too long, the packet will be considered lost by the receiver, decreasing audio quality. Packets that arrive late require complex processing in the receiver's codec. Jitter increases with both network load and distance. Jitter tends to range from near-zero values for overprovisioned local area networks (LANs) and 20 ms for intercity links up to 99 ms for international links [32].

4.1.4. Echo. Echo is the sound of the talker's voice being reflected back to the talker. The most common cause of echo in IP telephony systems is an impedance mismatch in the network, causing some of the speech energy to be reflected back to the talker. Echo may also be caused by acoustic coupling between the terminal's speaker and microphone. Echo is seldom noticed in the PSTN, because low delays allow it to return quickly to the talker's ear. The long delays of IP networks make echo more perceptible and annoying.

4.1.5. End-System Design. Several other factors can influence packet audio. Poor design of end-user equipment can accentuate background noise, decreasing intelligibility of speech. Noise can also originate from analog components or bit errors introduced by the end-user system. Finally, voice activity detectors (VADs) decrease bandwidth utilization by only sending packets when the user is speaking. Improper design of these devices can lead to inadvertently clipping parts of the audiostream.

4.2. Measurement Techniques

The quality of the received audio affects the ability of the listener to understand the talker. Hence, appropriate measurement algorithms are needed to evaluate audio quality in telephony networks. These measurement algorithms often operate by comparing an encoded

waveform to an uncoded reference, and often give outputs in terms of a mean opinion score (MOS).

The ITU-T defines the *perceptual speech quality measurement* (PSQM) method to provide a relative score of the quality of a voice signal, taking into account cognitive factors and the physiology of the human ear. Output and reference waveforms must be time synchronized before they are input to the algorithm. PSQM+ was later defined as an improvement to PSQM. PSQM+ more accurately reflects audio quality in the presence of packet drops and other quality impairments. The *perceptual analysis measurement system* (PAMS) is similar to PSQM, but uses a different signal-processing model. PAMS performs time alignment to eliminate delay effects, and improves on the accuracy of PSQM+. A second alternative to PSQM is *measuring normalized blocks* (MNB). PESQ combines the cognitive modeling of PSQM with the time alignment techniques of PAMS. PESQ has been found to give, on average, the most accurate measurements of these techniques [26].

5. SYSTEM DESIGN

IP telephony service providers must implement well-designed systems to meet the strict requirements of voice traffic. Data service providers must provide networks capable of handling a large number of voice calls. Equipment vendors must implement hardware and software that is highly available and routes voice traffic with low delay. Finally, protocols and standards must be developed to support scalable, robust, and secure wide-area telephony systems.

5.1. IP Network

High availability is a prerequisite for telecommunications. The PSTN has been engineered to provide virtually uninterrupted dial tone to over a billion users worldwide. Matching this impressive achievement with a fully IP-based network will be a challenging task. If IP telephony is to be accepted as a viable alternative to the PSTN, IP network hardware and software must be engineered to provide similar reliability, scalability, and ease of use. This can be achieved through redundancy, intelligent network design, and failover to PSTN networks in cases of severe congestion or failure.

The network must provide good QoS to the voice application. Designers of campus networks should over-provision to avoid call blocking, and should mark traffic to support routing QoS enhancements [5,8]. WANs may run at near-link capacity to reap the benefits of statistical multiplexing, and should use QoS enhancements such as traffic prioritization and QoS routing to improve service [2,38]. Gateways should be placed close to terminals to decrease loss, jitter, and delay. Network topologies that duplicate voice traffic over several links can improve performance at the cost of decreased efficiency. Packet classification may also be used to provide priority service to voice traffic. QoS improvements may be made in IP networks following the Integrated Services (IntServ) or Differentiated Services (DiffServ) models, or operating over ATM. Finally, increasing link speeds decreases serialization delay, further increasing service quality.

5.2. End System

IP telephony endpoints are much more intelligent than their PSTN counterparts, allowing us to implement sophisticated algorithms to achieve improved QoS and more effective use of bandwidth.

Proper design of codecs can greatly improve service quality. Using compression decreases network utilization, but can lower audio quality. Furthermore, it increases delay, sensitivity to dropped packets, and computational overhead. Silence suppression can decrease bandwidth utilization by up to 50% by only transmitting packets when the speaker is talking. However, quality of the signal may be reduced if the transmitter does not accurately detect when speech is present. Increasing the number of frames in an IP packet improves network utilization by removing packet overhead, but increases packetization delay and sensitivity to dropped packets [21]. For example, the ITU-T's G.711 codec gives the best received audio quality under poor network conditions, while the G.723.1 codec offers much lower network utilization at the expense of audio quality. Furthermore, codec DSP algorithms may be implemented in hardware, significantly decreasing the processing latency.

Packet loss in today's IP networks is fairly common. Voice transport in IP networks usually takes place over unreliable transport mechanisms, as the overhead for reliable transport significantly increases delay. However, packet loss may be acceptable because of the ability of the human brain to reconstruct speech from lost fragments and the advanced processing capabilities at IP telephony terminals. FEC can be used to reconstruct lost packets by transmitting redundant information along with the original information [22–24]. The redundant information is used to reconstruct some of the lost original data. For example, a low-quality copy of the previous packet may be transmitted with the current packet.

Packet jitter can be alleviated by using a playback buffer at the receiver. The sender transmits RTP packets with a sequence number and a timestamp, and the receiver queues each packet to be played at a certain time in the future. This playout delay must be long enough so that the majority of packets are received by their playout times. However, too much delay will be uncomfortable to the listener. Receivers may estimate the network delay and jitter, and calculate this playout delay accordingly.

To deal with unwanted echo, echo cancelers may be placed in the PSTN, the gateway, or the end-user terminal. These devices should be placed as close as possible to the source of the echo. Long-distance networks usually implement two echo cancelers per path, one for each direction. A good echo canceler delivers good audio quality without echo, distortion, or background noise. Echo cancelers work by obtaining a replica of the echo by a linear filter and subtracting it from the original signal, then using a nonlinear processor (NLP) to eliminate the remaining echo by attenuating the signal. PSTN echo cancelers are seldom constructed to recognize the large

delays from packet networks and do not attempt to perform echo cancellation in such circumstances. Echo cancellation must therefore also be implemented in the IP network or terminal.

6. RESOURCE MANAGEMENT

IP telephony networks consist of limited resources that must be allocated to a group of subscribers. These resources include IP network bandwidth, PSTN trunks, gateway functionality, and access to MCUs. IP telephony service providers must design and implement schemes to allocate resources to the users with greater demand. IntServ and DiffServ are two architectures useful for designing networks in which resources may be easily allocated. Pricing may then be used to control the way in which these resources are allocated to different users.

6.1. Resource Allocation

6.1.1. Integrated Services (IntServ). A simple way to partition resources among a group of users is to let the network decide whether a call will be admitted on the basis of factors such as user requirements and current network load. In the IntServ model, each router in the network is required to know what percentage of its link bandwidths are reserved [18]. An endpoint wishing to place a call must request enough bandwidth for the call, and the network may decide whether there are sufficient resources. If not, the call is blocked. The Resource Reservation Protocol (RSVP) is used in an IntServ framework to deliver QoS requests from a host to a router, and between intermediate routers [11]. QoS routing may be used to find the optimal path for the call to take through the IP network, as RSVP does not determine the links in which reservations are made.

Some networks, such as the Internet, are not constructed using the IntServ model. However, an administrator defined policy may be implemented at the H.323 gatekeeper or SIP server to provide admission control and resource reservation. For example, the gatekeeper may choose to block incoming calls when the total bandwidth in use by callers passing through that gateway exceeds a certain threshold.

6.1.2. Differentiated Services (DiffServ). The Differentiated Services (DiffServ) model provides different services for different classes of traffic by marking packets with the service class they belong to [17]. This class may be chosen on the basis of payload type, price charged to user, or price charged to network provider. The network allocates a set of resources for each service class, and each router provides services associated with the packet markings. Both H.323 and SIP may be used over networks architected in the IntServ or DiffServ models.

Although IntServ can provide QoS guarantees to network applications, it requires intermediate routers to store per-flow information, decreasing scalability. DiffServ provides enhanced scalability by allowing the routers to be stateless. DiffServ further improves on the IntServ model by providing more flexible service models and eliminating the need for QoS request signaling and channel setup delay. However, since DiffServ does not perform admission control, it cannot provide QoS guarantees to VoIP applications.

6.2. Pricing

It is expected that a moderate use of IP telephony will bring about an increase of 50% in an ISP's operating costs [37]. It is important that these costs be passed on to the users. The signaling architecture along with the greater intelligence of IP telephony endpoints allows service providers to formulate and implement complex pricing models that take into account both the dynamic congestion at the gateway along with the desired QoS of the call [33–35]. The price charged to the user should take into account the amount of resources consumed as well as the expected revenue lost to the provider because higher-paying users are blocked from using those resources [43].

Several pricing schemes have been proposed for IP telephony services. Although flat pricing schemes are well accepted by users and tend to dominate in most Internet services, light users subsidize heavy users, making them economically inefficient [41]. Pricing resources based on the current congestion in the network causes low paying users to back off, freeing resources for high-paying users. In congestion sensitive pricing, users may be charged per byte or per minute. In "smart market" (SM) pricing, the user puts a bid into each packet. During times of congestion, the router will drop the packet if it is below the current congestion price. Charging users a higher price for a higher-QoS connection leaves high-quality connections free for higher-paying users. In QoS-sensitive pricing, the user may be charged a higher price for closer gateways or gateways offering supplementary services. In Paris Metro Pricing [36], the network bandwidth is partitioned into several subnetworks, each priced differently. Users with low demand will tend to congregate in the lower priced partitions, increasing the service quality of more expensive partitions. The provider should choose a pricing scheme that is economically efficient and accepted by users. These pricing schemes may be implemented using the Gateway Location Protocol (GLP) to distribute price information in real time [10,12,16].

7. SECURITY ISSUES

IP telephony service providers must implement a secure system to ensure authentication, integrity, privacy, and nonrepudiation. This system must prevent eavesdropping, avoiding payment, and calls being made by nonsubscribers. Replay, spoofing, man-in-the-middle, and denial of service (DoS) attacks must also be constrained. The PSTN provides a vertical, intelligent, circuit-switched network robust against many types of attack. In the IP network, most of the intelligence is concentrated at the endpoints, making it more challenging to implement secure services. These services may be implemented in the IP telephony software, and may rely on lower-layer protocols such as IP Security (IPSec) [39] or Transport Layer Security (TLS) [40].

In the H.323 protocol architecture, security of the videostream, call setup, call control, and communications with the gatekeeper is achieved by H.235. This is achieved by having the endpoints authenticate themselves with gateways, gatekeepers, and MCUs. Encrypting data and control channels can prevent attackers from making unauthorized modifications to the information in transit, thereby providing data integrity and privacy. This can, for example, prevent attackers from using a LAN analyzer to listen to an unencrypted call on a broadcast network. Finally, nonrepudiation ensures that no endpoint can deny that it participated in a call, and is provided by the gatekeeper. Being sure of a caller's identity is critical for billing services.

SIP derives most of its security properties from HTTP [15]. In addition to these, SIP also supports public key encryption. SIP does not specify the security of user sessions, but SDP messages may be exchanged to allow the endpoints to decide on a common scheme. Data integrity and privacy can be implemented with encryption. Authentication of end users may be performed with HTTP authentication or PGP. Having callers register with a SIP server provides nonrepudiation.

8. REGULATORY AND STANDARDIZATION ISSUES

There are several standards organizations designing standards for IP telephony. The Internet Engineering Task Force (IETF) develops IP related standards for the Internet. The IETF publishes *requests for comments* (RFCs) on many areas, including IntServ, DiffServ, Gateway location, Telephony Routing over IP (TRIP), RTP, and SIP. The Asynchronous Transfer Mode (ATM) Forum deals with IP and ATM issues. Unlike IP, ATM was designed to support applications with different characteristics and service requirements, and networks may use IP over ATM networks to improve QoS. Telephony and Internet Protocol Harmonization Over Networks (TIPHON) develops specifications covering interoperability, architecture, and functionality for IP telephony networks. The Digital Audio Video Council (DAVIC) was created to address end-to-end interoperability of audiovisual information and multimedia communication. Its charter is to develop specifications for development based on Moving Picture Experts Group version 2 (MPEG-2) coding. The ITU-T is concerned with developing standards for all fields of telecommunications except radio aspects. The ITU-T began to study IP related projects in 1997, and established formal collaborations with the IETF, ATM Forum, and DAVIC.

National governments have a wide range of Internet and telephony regulatory policies. The technological advance that brought about IP telephony has moved much faster than most countries were able to change policy. Furthermore, it is not clear whether VoIP should be regulated and taxed as voice in the PSTN. Countries such as the United States, Sweden, Italy, and Russia do not restrict access to the Internet nor regulate IP telephony services. China disallows ISPs from operating international gateways or providing a nationwide infrastructure, although they may run their own switches. Canada allows ISPs to establish international gateways. However, they must pay tariffs to interface with the PSTN. Japan requires ISPs to obtain a license to establish and operate circuits and facilities. Regulatory issues are difficult to resolve and need to be reviewed in detail.

BIOGRAPHIES

Matthew Chapman Caesar completed his B.S. degree in computer science at the University of California at Davis in 1999. During 2000, he was a member of technical staff at iScale, Mountain View, California. Currently, he is enrolled in the Ph.D. program in computer science at the University of California at Berkeley. In 2001, he was awarded the National Science Foundation (NSF) Graduate Research Fellowship and the Department of Defense National Defense Science and Engineering Graduate (NDSEG) Fellowship. His research interests include Internet economics, congestion control, and real-time multimedia services.

Dipak Ghosal received his B.Tech degree in electrical engineering from Indian Institute of Technology, Kanpur, India, in 1983, his M.S. degree in computer science from Indian Institute of Science, Bangalore, India, in 1985, and his Ph.D. degree in computer science from University of Louisiana, Lafayette, in 1988. From 1988 to 1990 he was a research associate at the Institute for Advanced Computer Studies at University of Maryland (UMIACS) at College Park. From 1990 to 1996 he was a member of technical staff at Bell Communications Research (Bellcore) at Red Bank. Currently, he is an associate professor of the Computer Science Department at the University of California at Davis. His research interests are in the areas of IP telephony, wireless network, web caching, and performance evaluation of computer and communication systems.

BIBLIOGRAPHY

1. J. Kurose and K. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*, Addison-Wesley, 2000.
2. B. Li, M. Hamdi, D. Jiang, and X. Cao, QoS enabled voice support in the next generation internet: Issues, existing approaches and challenges, *IEEE Commun. Mag.* **38**(4): 54–61 (April 2000).
3. D. Bergmark and S. Keshav, Building blocks for IP telephony, *IEEE Commun. Mag.* **38**(4): 88–94 (April 2000).
4. F. Anjum et al., ChaiTime: A system for rapid creation of portable next-generation telephony services using third-party software components, *Proc. 2nd IEEE Conf. Open Architectures and Network Programming* (*OPENARCH*), New York, March 1999.
5. M. Hassan, A. Nayandoro, and M. Atiquzzaman, Internet telephony: Services, technical challenges, and products, *IEEE Commun. Mag.* **38**(4): 96–103 (April 2000).
6. A. Rayes and K. Sage, Integrated management architecture for IP-based networks, *IEEE Commun. Mag.* **38**(4): 48–53 (April 2000).

7. M. Hamdi et al., Voice service interworking for PSTN and IP networks, *IEEE Commun. Mag.* (May 1999).
8. Cisco Systems, *Architecture for Voice, Video and Integrated Data*, technical white paper, *http://www:cisco:com/warp/public/cc/so/neso/vvda/iptl/avvid_wp:htm*.
9. M. Korpi and V. Kumar, Supplementary services in the H.323 IP telephony network, *IEEE Commun. Mag.* **37**(7): 118–125 (July 1999).
10. J. Rosenberg and H. Schulzrinne, Internet telephony gateway location, *IEEE INFOCOM*, San Francisco, March–April 1998.
11. R. Braden et al., *Resource reservation protocol (RSVP)—Version 1 Functional Specification*, RFC 2205, September 1997.
12. J. Rosenberg and H. Schulzrinne, *A Framework for Telephony Routing over IP*, IETF, RFC 2871, June 2000.
13. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, Audio-Video Transport Working Group, *RTP: A Transport Protocol for Real-Time Applications*, IETF, RFC 1889, Jan. 1996.
14. Recommendation H.323, *Visual Telephone Systems and Equipment for Local Area Networks Which Provide a Non-guaranteed Quality of Service*, International Telecommunications Union, Telecommunications Standardization Sector (ITU-T), Geneva, Switzerland, Nov. 1996.
15. M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, *SIP: Session Initiation Protocol*, IETF, RFC 2543, March 1999.
16. C. Agapi et al., *Internet Telephony Gateway Location Service Protocol*, IETF, Internet Draft, November 1998.
17. S. Black et al., *An Architecture for Differentiated Service*, IETF, RFC 2475, December 1998.
18. R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, IETF, RFC 1633, June 1994.
19. Organization for Economic Co-operation and Development, *OECD Communications Outlook 1999*, OECD, 1999.
20. M. Arango et al., *Media Gateway Control Protocol (MGCP)*, IETF, RFC 2705, Oct. 1999.
21. M. Perkins, C. Dvorak, B. Lerich, and J. Zebarth, Speech transmission performance planning in hybrid IP/SCN networks, *IEEE Commun. Mag.* **37**(7): 126–131 (July 1999).
22. J. C. Bolot, S. Fosse-Parisis, and D. Towsley, Adaptive FEC-based error control for Internet telephony, *Proc. IEEE INFOCOM*, New York, March 1999.
23. M. Podolsky, C. Romer, and S. McCanne, Simulation of FEC-based control for packet audio on the Internet, *Proc. IEEE INFOCOM*, San Francisco, March–April 1998.
24. J. C. Bolot and H. Crepin, Analysis and control of audio packet loss over packet-switched networks, *Proc NOSSDAV*, Durham, NC, April 1995.
25. A. Percy, *Understanding Latency in IP telephony*, technical white paper, *http://www:brooktrout.com/whitepaper/iptel_latency.htm*.
26. J. Anderson, *Methods for Measuring Perceptual Speech Quality*, technical white paper, *http://onenetworks.comms.agilent.com/downloads/PerceptSpeech2.pdf*.
27. G. Herlein, The Linux telephony kernel API, *Linux J.* **82**: (Feb. 2001).
28. Microsoft Corp., *IP telephony with TAPI 3.0* (*Microsoft Corp), technical white paper*, *http://www.microsoft.com/windows-2000/techinfo/howitworks/communications/telephony/iptelephony.asp*.
29. Cisco, *IP telephony Solution Guide: Planning the IP telephony Network*, technical white paper, *http://www.cisco.com/warp/public/788/solution_guide/3_planni.htm*.
30. G. Cook, Jr., Taking the hybrid road to IP telephony, *Commun. Eng. Design Mag.* (Dec. 2000).
31. W. Matthews, L. Cottrell, and R. Nitzan, 1-800-CALL-HEP, presented at Computing in High Energy and Nuclear Physics 2000 (CHEP'00), Feb. 2000 (full text on website *http://www.slac.stanford.edu/pubs/slacpubs/8000/slac-pub-8384.html)*.
32. J. Altmann and K. Chu, A proposal for a flexible service plan that is attractive to users and Internet service providers, *IEEE INFOCOM*, Anchorage, April 2001.
33. X. Wang and H. Schulzrinne, Pricing network resources for adaptive applications in a differentiated services network, *IEEE INFOCOM*, Anchorage, April 2001.
34. M. Falkner, M. Devetsikiotis, and I. Lambadaris, An overview of pricing concepts for broadband IP networks, *IEEE Commun. Surv.* **3**(2): (April 2000).
35. L. DaSilva, Pricing for QoS-enabled networks: A survey, *IEEE Commun. Surv.* **3**(2): (April 2000).
36. A. Odlyzko, Paris Metro Pricing for the Internet, *Proc. ACM Conf. Electronic Commerce*, Denver, Nov. 1999, pp. 140–147.
37. L. McKnight, Internet telephony: Costs, pricing, and policy, *Proc. 25th Annual Telecommunications Policy Research Conf.*, 1997.
38. A. Dubrovsky, M. Gerla, S. Lee, and D. Cavendish, Internet QoS routing with IP telephony and TCP traffic, *Proc. ICC, New Orleans, June 2000*.
39. S. Frankel, *Demystifying the Ipsec Puzzle*, Artech House, Boston, 2001.
40. T. Dierks and C. Allen, *The TLS Protocol: Version 1.0*, IETF, RFC 2246, Jan. 1999.
41. G. Huston, *ISP Survival Guide: Strategies for Running a Competitive ISP*, Wiley, New York, 1999.
42. M. Handley, and V. Jacobson, *SDP: Session Description Protocol*, RFC 2327, April 1998.
43. M. Caesar, S. Balaraman, and D. Ghosal, "A Comparative Study of Pricing strategies for IP telephony", IEEE Globecom 2000, Global Internet Symposium, San Francisco, Nov. 2000.
44. M. Caesar and D. Ghosal, "IP Telephony Annotated Bibliography", *http://www.cs.berkeley.edu/~mccaesar/research/iptel_litsurv.html.*

ITERATIVE DETECTION ALGORITHMS IN COMMUNICATIONS

KEITH M. CHUGG
University of Southern California
Los Angeles, California

1. INTRODUCTION

Iterative algorithms are those that repeatedly refine a current solution to a computational problem until

an optimal or suitable solution is yielded. Iterative algorithms have a long history and are widely used in modern computing applications. The focus of this article is iterative algorithms applied to communication systems, particularly to receiver signal processing in digital communication systems. In this context, the basic problem is for the receiver to infer the digital information that was encoded at the transmitter using only a waveform observed after corruption by a noisy transmission channel. Thus, one can imagine that an iterative receiver algorithm would obtain an initial rough estimate of the transmitted block of data, and then refine this estimate. The refinement process can take into account various sources of corruption to the observed signal and various sources of structure that have been embedded into the transmitted waveform. Even within this fairly narrow context, there are a large number of distinct iterative algorithms that have been suggested in the literature.

There is one elegant and relatively simple iterative approach that has recently emerged as a powerful tool in modern communication system design. This approach, which is the focus of this article, is known by many names in the literature, including iterative detection and decoding [1], belief propagation [2–4], message-passing algorithms [5], and the "turbo principle."[1] Appreciation for the power and generality of this approach in the communications and coding literature resulted from the invention of "turbo codes" in 1993 [6,7] and the associated decoding algorithm, which is a direct application of the standard iterative algorithm addressed in this article [8]. Turbo codes are error-correction codes with large memory and strong structure that are constructed using multiple constituent codes, each with relatively low complexity, connected together by data permutation devices (interleavers). Turbo codes and similar turbo-like codes were found to perform very close to the theoretical limit, a feat that evaded coding theorist for years and was thought by many to be practically impossible.

Intrigued by the effectiveness of the iterative decoding algorithm, researchers in the field of data detection and decoding pursued several parallel directions shortly after the invention of turbo codes. One subject addressed was the general rules for the iterative algorithm, i.e., *Is there a standard view of the various iterative algorithms suggested to decode turbo-like codes?* A related issue is *In what sense and under what conditions is the algorithm optimal and, otherwise, how may it be viewed as a good approximation to optimal processing?* A third area of research was the application of the iterative decoding paradigm to other receiver processing tasks such as channel equalization, interference mitigation, and parameter tracking. By the late 1990s, all of these issues had been relatively well addressed in the literature [1]. A more recent accomplishment was the development of tools for predicting the convergence properties of the standard iterative algorithms [9,10].

Currently, there is a consensus in the literature regarding the standard iterative detection paradigm and an understanding of its practical characteristics and range of application. The transfer of this understanding in the research community to engineering practice is well under way. The use of turbo-like codes is now common is systems and standards designed after 1996. Practical implementations of other applications and a more aggressive exploitation of the potential benefits of iterative detection are currently emerging in industry. The motivation, basic ideas underlying the approach, and potential advantages are described in the following subsections.

[1] Unless emphasized otherwise, these terms are used interchangeably in this article.

1.1. Motivation for Iterative Detection

A common, generic, digital communication system block diagram is shown in Fig. 1 (a) and (b). In fact, the receiver block diagram in Fig. 1 (b) mirrors the processing performed in most practical receiver implementations. This *segregated* design paradigm allows each component of the receiver to be designed and "optimized" without much regard to the inner workings of the other blocks of the receiver. As long as each block does the job it is intended for, the overall receiver should perform the desired task: extracting the input bits.

Despite the ubiquity of the diagram in Fig. 1 (b) and the associated design paradigm, it clearly is not optimal from the standpoint of performance. More specifically, the probability of error for the data estimates is not minimized by this structure. This segregated processing is adapted for tractability — both conceptual tractability and tractability of hardware implementation. The optimal receiver for virtually any system is conceptually simple, yet typically prohibitively complex to implement. For example, consider the transmission of 1000 bits through a system of the form shown in Fig. 1 (a). These bits may undergo forward error-correction coding (FEC), interleaving, training insertion (pilots, synchronization fields, training sequences, etc.) before modulation and transmission. The channel may corrupt the modulated signal through random distortions (possibly time-varying and nonlinear), like-signal interference (co-channel, multiple access, crosstalk, etc.), and additive noise. The point is, regardless of the complexity of the transmitter and/or channel, the optimal receiver would compute 2^{1000} likelihoods and select the data sequence that most closely matches the assumed model. Intuitively, this is one of the primary advantages of digital modulation techniques — i.e., the receiver knows what to look for and there are but a finite number of possibilities. This is shown in Fig. 1 (c). Ignoring the obvious complexity problems, this requires a good model of the transmitter formatting and the channel effects. For example, the aforementioned likelihood computation may include averaging over the statistics of a fading channel model or the possible data values of like-signal interferers.

The iterative approaches described in this article are exciting because they enable receiver processing that can closely approximate the above optimal solution with feasible conceptual and implementation complexity. Specifically, data detection and parameter estimation are done using the entire global system structure. Unlike the direct approach in Fig. 1 (c), the iterative receiver in Fig. 1 (d) exploits this structure indirectly. The key

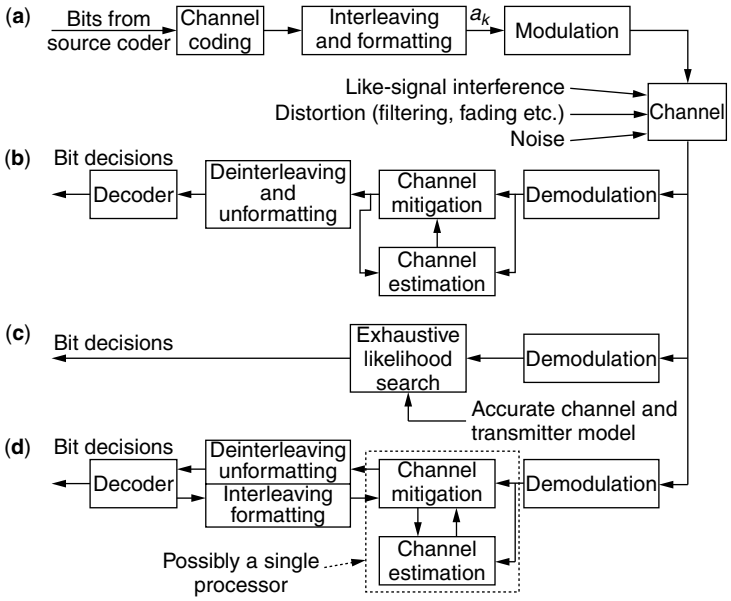


Figure 1. (**a**) A typical communication block diagram, (**b**) a traditional segmented receiver design, (**c**) the optimal receiver processing, and (**d**) a receiver based on iterative detection principles. The formatting includes insertion of framing and training information (overhead). Channel mitigation includes tasks such as equalization, diversity combining, array combining, and so on.

concept in this approach is the exchange and updating of "soft information" on digital quantities in the system (e.g., the coded modulation symbols). This concept is shown in Fig. 1 (d). The iterative detection receiver is similar to the conventional segregated design in that, for each subsystem block in the model, there is a corresponding processing block. In fact, each of these corresponding processing blocks in the receiver of Fig. 1 (c) exploits only *local* system structure (e.g., the FEC decoder does not use any explicit knowledge of the channel structure). As a consequence, the complexity of the receiver in Fig. 1 (d) is comparable to the traditional segregated design in Fig. 1 (b) [i.e., the increase in complexity usually is linear as opposed to the exponential increase in complexity associated with the optimal processing in Fig. 1 (c)].

1.2. Components of the Standard Iterative Detection Algorithm

There are several basic concepts that form the basis of the standard iterative detection paradigm. At the core of this paradigm is the exchange and update of soft information on digital variables. A digital variable is one that takes only a finite number of values, which are known to the receiver. Soft information on a digital variable is an array of numbers that gives a measure "belief" that a given conditional value is correct. For example, soft information on a binary variable $b \in \{0, 1\}$ can be represented by two numbers. If such soft information were represented as probabilities, $\mathrm{P}[b = 0] = 0.2$ and $\mathrm{P}[b = 1] = 0.8$ would represent the case where the current belief is that $b = 1$ is four times as likely as the zero hypothesis. Soft information can be represented in various forms and therefore is also referred to as beliefs, reliabilities, soft decisions, messages, and metrics. Soft information should be contrasted with hard decision information. For the example discussed previously, the associated hard decision information would simply be the best current guess for the value of b — i.e., $\hat{b} = 1$.

The standard iterative detection paradigm defines a way of refining soft information on the system variables, most importantly the digital data inputs to the system. The components of this approach are:

- **Modeling System Structure:** This involves specifying the structure of the system or computational problem of interest such that local dependencies are captured and the global system structure is accurately modeled. For example, in Fig. 1 (a) this is accomplished by modeling the system (global structure) as a concatenated network of subsystems (local structure). This may seem like a trivial step, but, in fact, this is possibly the most important and least well-understood aspect of the field. As will become apparent, once the model has been selected, the standard iterative detection algorithm is defined for the most part. The properties of the model selected determine in large part the complexity of the processing, the optimality or the effectiveness of a suboptimal algorithm, and many relevant implementation properties such as latency and area tradeoffs. This basic problem has provided the impetus for applying graph theory to this field. Specifically, graphical models are simply explicit index block diagrams that capture local dependencies between relevant system variables.
- **Directly and Fully Exploiting Local Structure:** For each defined local subsystem or node in the model, the associated processing performed is defined in terms of the optimal receiver for that subsystem. This leads to the notion of a *marginal soft inverse* (or, for brevity, the soft inverse) of a subsystem. The soft inverse takes in marginal soft information on each of the its digital inputs and outputs and updates this information by exploiting the local subsystem structure. The term *marginal soft information* means that the soft information is provided separately for

each variable indexed in a sequence. For example, if $\{b_0, b_1, \ldots b_{1023}\}$ is a sequence of binary inputs to the system, the marginal soft information is 1024 arrays of size two as opposed to one array of size 2^{1024} (i.e., the latter is joint soft information). The subsystem soft inverse is defined based only on the local subsystem structure. Specifically, it is based on the optimal data detection algorithm under the assumption of a sequence of independent inputs and a memoryless channel corruption of the outputs. Thus, the soft inverse takes in "soft-in" information on the system inputs and "soft-in" information on the system outputs, and it produces "soft-out" information on the system inputs and "soft-out" information on the system outputs. For this reason, the soft inverse processor is sometimes referred to as a soft-in/soft-out (SISO) processor.

- **Exploit Global Structure via Marginal Soft-Information Exchange:** Since the soft inverse does not take into account the overall global system structure, this structure should be accounted for somehow if the globally optimal receiver is to be achieved or well approximated. This occurs by the exchange of soft information between soft inverse processors that correspond to subsystems having direct dependencies on the same variables. For example, if the output of one subsystem is the input to another subsystem, they will exchange soft information on these common variables. The soft information can be viewed as a method to bias subsequent soft inverse processing. Specifically, soft-in information on the system inputs plays the role of a priori probabilities on inputs and soft-in information on the system outputs plays the role of channel likelihoods.

Consider the preceding general description in the context of the example in Fig. 1. As mentioned, the concatenated block diagram defines the model. For each block in the model, there is a soft inverse processor in the iterative receiver of Fig. 1 (d). The arrows leading into each soft inverse block correspond to the soft-in information and the arrows departing represent soft-out information. The task of these processing units is to *update* the beliefs on the input and output variables of the corresponding system sub-block in Fig. 1 (a). Each sub-block processing unit will be activated several times, each time biased by a different (updated) set of beliefs.

The iterative receiver offers significant performance advantages over the segregated design. For example, suppose that a system using convolutional coding and interleaving experiences severe like-signal interference and distortion over the channel. In this case, the channel mitigation block in Fig. 1 (b) will output hard decisions on the coded/interleaved bit sequence a_k. Suppose that, given the severity of the channel, the error probability associated with these coded-bit decisions will be approximately 0.4. Deinterleaving these decisions and performing *hard-in* (Hamming distance) decoding of the convolutional code will provide a very high bit error rate (BER) (i.e., nearly 0.5).

For the receiver in Fig. 1 (d), however, the channel mitigation block produces soft-decision information on the coded/interleaved bit sequence a_k. For example, this may be thought of as two numbers $\mathrm{P}[a_k = 1]$ and $\mathrm{P}[a_k = 0]$ that represent a measure of current probability or belief that the k-th coded bit a_k takes on the value 1 or 0, respectively. Clearly, soft decisions contain more information than the corresponding hard decisions. In this example, it is possible that even though the hard decisions on a_k associated with the receiver of Fig. 1 (b) are hopelessly inaccurate, the soft-decision information contains enough information to jump-start a decoding procedure. For example, two different possible sequences of soft-decision information are shown in Table 1. Note that each of these correspond to exactly the same hard-decision information (i.e., the hard decisions obtained by *thresholding* the soft information is the same). However, the soft information in case B is much worse than that in case A. Specifically, for case A, there is a high degree of confidence for correct decisions and very low confidence for incorrect decisions. For case B, there is little confidence in any of the decisions.

A receiver of the form in Fig. 1 (d) would pass the soft information through a deinterleaver to a soft-in/soft-out decoder for the convolutional code. Thus, after activation of this decoder, one could make a decision on the uncoded bits. Alternatively, the updated beliefs on the coded bits could be interleaved and used in the role of a priori probabilities to bias another activation of the channel mitigation processing unit in Fig. 1 (d). In fact, this process could be repeated with the channel mitigation and FEC decoder exchanging and updating beliefs on the coded bits through the interleaver/deinterleaver pair. After several iterations, final decisions can be made on the uncoded bits by thresholding the corresponding beliefs generated by the code processing unit. This is what is meant by iterative detection.

Note that in this example, even though the hard-decision information on the coded bits after activating the channel mitigation processing unit is very unreliable (e.g., an error rate of 0.4), the soft information may allow

Table 1. Example of Two Sequences of Soft Information Implying the Same Hard Decisions, But Containing Very Different Soft Information. The Soft Information is Given as ($P[a_k = 0]$, $P[a_k = 1]$)

k:	0	1	2	3	4...
true data:	0	0	1	0	1
case A:	(0.99, 0.01)	(0.97, 0.03)	(0.51, 0.49)	(0.48, 0.52)	(0.03, 0.97)
case B:	(0.51, 0.49)	(0.55, 0.45)	(0.51, 0.49)	(0.48, 0.52)	(0.48, 0.52)
decisions:	0	0	0	1	1

the FEC decoder to draw some reasonable inference. For example, if the soft information is that of case A in Table 1, then the FEC decoder may update the beliefs as to overturn the unreliable (incorrect) decisions and reenforce the reliable decisions (i.e., the correct decisions in this example). Note that this updating takes into account only these marginal beliefs on the coded bits and the local code structure.

In summary, the receiver processing in Fig. 1 (d) closely approximates the performance of the optimal processing in Fig. 1 (c), with complexity roughly comparable to that of the traditional segmented design in Fig. 1 (b). It does so by performing locally optimal processing that exploits the local system structure and by updating and exchanging marginal soft information on the subsystem inputs and outputs.

1.3. Summary of Article Contents

The remainder of this article expands on the preceding material and discusses some modifications that may be required in practical applications. Specifically, in Section 2 the soft inverse operation is defined more precisely and some important special cases are presented. The standard rules for iterative detection are described in Section 3, which also contains a discussion of graphical models and a sufficient condition for optimality of the processing. A brief summary of selected applications of iterative detection is presented in Section 4.

2. SYSTEM SOFT INVERSE

Consider a system comprising a concatenated network of subsystems as illustrated in Fig. 2. To illustrate that the soft inverse is the key concept underlying iterative detection, the standard iterative detector for the system of Fig. 2 is illustrated in Fig. 3. Note that the iterative detector is specified by replacing each subsystem in the system model by the corresponding soft inverse in the iterative detector. Subsystems connected in the model correspond to soft inverse processors that directly exchange soft information messages. Given the soft inverse concept, the only remaining issues are related to scheduling, specifically in what order the soft inverses are activated and when to stop iterating. These scheduling issues are discussed in Section 3, while the soft inverse is defined more precisely in this section.

A system with digital inputs a_m and outputs x_n is shown in Fig. 4 along with the corresponding soft inverse. Two conventions for block diagrams are used. The implicit

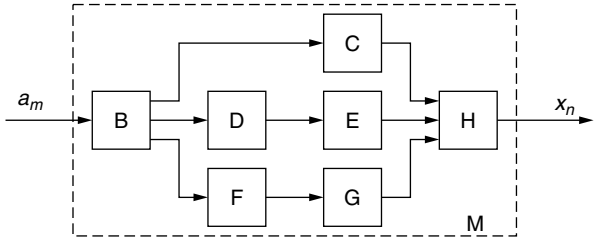


Figure 2. The block diagram of a generic concatenated network.

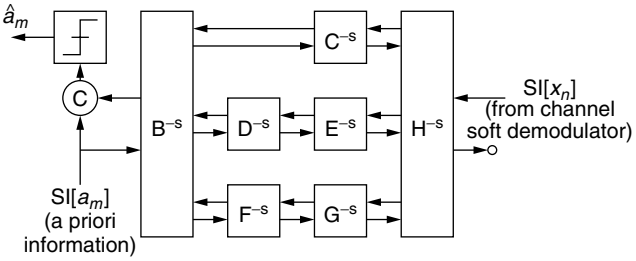


Figure 3. The iterative detector implied by the Fig. 2.

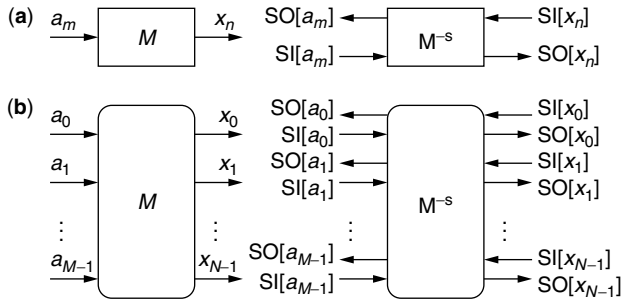


Figure 4. The marginal soft inverse of a system in the (**a**) implicit and (**b**) explicit index block diagram conventions.

(time) index convention is that most commonly used in signal processing and communications where, for example, the label a_m implies a sequence of variables. In the explicit (time) index convention, each variable in this sequence is shown explicitly. Each of these conventions has its own advantages, with the latter leading directly to the graphical modeling approaches popularized by the connection to work in computer science. The notation SI[·] and SO[·] denotes the soft-in and soft-out information, respectively. For example, if x_{10} takes on eight values, then SI[x_{10}] and SO[x_{10}] each corresponds to an array of eight numbers, much like a probability mass function.

There are many possible variations on how one represents soft information. The two most important conventions are representation in the probability domain and representation in the metric domain. In the probability domain, the soft information may be viewed as beliefs with larger numbers implying more confidence. For the special case of soft-in/soft-out information in the probability domain, the notation PI[·] and PO[·] is adopted and P[·] is used for generic probability domain soft information. Most practical software and hardware implementations are based on representation in the log-probability domain. The term *metric* is used herein to refer to the negative log of a probability domain quantity. For the special case of soft-in/soft-out information in the metric domain, the notation MI[·] and MO[·] is adopted and M[·] is used for generic metric domain soft information.

Soft inversion is a computational problem as opposed to a computational algorithm. A soft inverse computes the equivalent of a two-step process. The first step is combining of marginal soft-in information to obtain (joint) soft information on each possible system configuration. A

system configuration is defined by an allowable input-output sequence pair [e.g., $(\mathbf{a}, \mathbf{x}(\mathbf{a}))$ where boldface represents the vector of variables]. In the probability domain, combining is performed by multiplication, in the same way that the joint probability is the product of the marginal probabilities for independent quantities. Thus, for each configuration one could compute

$$\begin{aligned} \mathrm{P}[\mathbf{a}, \mathbf{x}(\mathbf{a})] &= \mathrm{P}[\mathbf{x}(\mathbf{a})] \times \mathrm{P}[\mathbf{a}] \\ &= \left(\prod_{n=0}^{N-1} \mathrm{PI}[x_n(\mathbf{a})]\right) \times \left(\prod_{m=0}^{M-1} \mathrm{PI}[a_m]\right) \end{aligned} \tag{1}$$

where each element in the second equality is consistent with the configuration $(\mathbf{a}, \mathbf{x}(\mathbf{a}))$.

The second step is marginalization of this joint soft information to produce updated marginal soft information. A natural way to marginalize is to sum over all $\mathrm{P}[\mathbf{a}, \mathbf{x}(\mathbf{a})]$ consistent with a specific conditional value of a specific input or output variable. For example, to get $\mathrm{PO}[x_{10} = 3]$, one could sum $\mathrm{P}[\mathbf{a}, \mathbf{x}(\mathbf{a})]$ over all configurations with $x_{10} = 3$. The notation $z : y$ is used to denote "all z consistent with y." Thus, marginalization to obtain $\mathrm{PO}[\cdot]$ for the input and output variables is

$$\mathrm{PO}[a_m] = \left(\sum_{\mathbf{a}:a_m} \mathrm{P}[\mathbf{a}, \mathbf{x}(\mathbf{a})]\right) \div \mathrm{PI}[a_m] \tag{2a}$$

$$\mathrm{PO}[x_n] = \left(\sum_{\mathbf{a}:x_n} \mathrm{P}[\mathbf{a}, \mathbf{x}(\mathbf{a})]\right) \div \mathrm{PI}[x_n] \tag{2b}$$

Thus, for example, $\mathrm{PO}[x_{10} = 3]$ is computed by summing $\mathrm{P}[\mathbf{a}, \mathbf{x}(\mathbf{a})]$ over all configurations with $x_{10} = 3$ and then dividing by $\mathrm{PI}[x_{10} = 3]$.[2] This division converts the soft-out information to the so-called *extrinsic form*, which as will be discussed, can be viewed as a form of likelihood. The convention is to refer to processing by the marginalization and combining operators; thus, the above is *sum-product* processing.

Another reasonable marginalization operator for the probability domain is the max operator. Without altering the combining, therefore, *max-product* soft inversion is obtained by replacing the summations in Eq. (2) by max operations.

Soft inversion using sum-product or max-product marginalization and combining is based on optimal decision theory. For example, assume that $\mathrm{PI}[a_m] = p(a_m)$ is the a priori probability for the independent input variables. Further, assume that $\mathrm{PI}[x_n] = p(z_n \mid x_n)$ is the channel likelihood for system output x_n based on channel observation z_n, where the channel in memoryless. Then, using sum-product processing, $\mathrm{PO}[a_m] \times \mathrm{PI}[a_m] \equiv p(a_m \mid \mathbf{z}) = \mathrm{APP}[a_m]$ or the a posteriori probability (APP) of input a_m given the entire observation sequence $\mathbf{z}$. Thresholding this soft information therefore yields the maximum a posteriori probability (MAP) decision on a_m. So, sum-product processing under the assumption of independent inputs and a memoryless channel yields MAP symbol detection (MAP-SyD). Under the same assumptions, max-product processing yields soft information that, when thresholded, yields decisions optimal under the MAP sequence decision (MAP-SqD) criterion. In summary, while the soft inversion problem was presented as a computational problem, it is based on Bayesian decision theory for the system in isolated, ideal conditions.

As mentioned previously, for numerical stability and hardware efficiency, the soft inversion processing is almost always implemented in the log domain. Both sum-product and max-product processing can be equivalently carried out in the metric domain. For the max-product case, this is straightforward since the negative-log and max operations commute. Furthermore, in the metric domain, the product combining becomes sum combining. Thus, in the metric domain, max-product processing corresponds to *min-sum* processing of metrics that are defined as the negative-log of the probability-domain quantities defined above [e.g., $\mathrm{MI}[a_m] = -\ln(\mathrm{PI}[a_m])$]. In particular, the min-sum soft inversion problem is

$$\mathrm{M}[\mathbf{a}, \mathbf{x}(\mathbf{a})] = \mathrm{M}[\mathbf{x}(\mathbf{a})] + \mathrm{M}[\mathbf{a}] \tag{3a}$$

$$= \left(\sum_{n=0}^{N-1} \mathrm{MI}[x_n(\mathbf{a})]\right) + \left(\sum_{m=0}^{M-1} \mathrm{MI}[a_m]\right) \tag{3b}$$

$$\mathrm{MO}[a_m] = \left(\min_{\mathbf{a}:a_m} \mathrm{M}[\mathbf{a}, \mathbf{x}(\mathbf{a})]\right) - \mathrm{MI}[a_m] \tag{3c}$$

$$\mathrm{MO}[x_n] = \left(\min_{\mathbf{a}:x_n} \mathrm{M}[\mathbf{a}, \mathbf{x}(\mathbf{a})]\right) - \mathrm{MI}[x_n] \tag{3d}$$

Conversion of the sum-product to the metric domain is more complicated because the negative-log operation does not commute with the summation operator. This is handled nicely by introducing the $\min^*(\cdot)$ operation [11] as

$$\min^*(x, y) \triangleq -\ln\left(e^{-x} + e^{-y}\right) \tag{4a}$$

$$= \min(x, y) - \ln(1 + e^{-|x-y|}) \tag{4b}$$

$$\min^*(x, y, z) \triangleq -\ln\left(e^{-x} + e^{-y} + e^{-z}\right) \tag{4c}$$

$$= \min^*(\min^*(x, y), z) \tag{4d}$$

Then, the metric domain version of the sum-product soft inversion problem is the $\min^*$-sum problem obtained by replacing the min operations in (3) by $\min^*$ operations. Notice that $\min^*(x, y)$ is neither x nor y in general. Also, when $|x - y|$ is large $\min^*(x, y) \approx \min(x, y)$, which implies that the two basic marginalization approaches should yield similar results at moderate to high SNR (i.e., MAP-SqD and MAP-SyD should yield similar decisions).

The soft inversion problem as described here is actually very general and arises in many problems of interest inside and outside the area of communications. This problem is sometimes referred to as a Marginalize a Product Function (MPF) problem [5], based on the sum-product version. In the min-sum version, the problem has a very intuitive form. For example, $\mathrm{MO}[a_m = 0] + \mathrm{MI}[a_m = 0]$ is the minimum total configuration metric over all system configurations consistent with $a_m = 0$. For this reason, the

[2] Note that this division can be avoided if the term in the denominator is excluded from the combining operation.

term Minimum Sequence Metric or Minimum Sum Metric (MSM) is used to denote the minimum configuration metric consistent with a particular conditional value of a variable. The notation MSM[u] is used to denote this quantity. Thresholding MSM[a_m] yields MAP-SqD. For this reason, the min-sum version of the soft inversion problem is referred to as a shortest path problem [12].

A very useful result is that under certain conditions on the marginalization and combining operators, an algorithm to perform the soft inversion under one convention can be directly converted to another convention. Specifically, the marginalization and combining operators together with the soft information representation should form a commutative semi-ring [5,12]. In this case, any algorithm that uses only the properties of the marginalization-combining semi-ring[3] can be converted to any other marginalization-combining convention that forms a semi-ring. For example, this allows one to change a proper MAP-SyD algorithm to a MAP-SqD algorithm by simply redefining operators. This allows one to work with the most convenient form to obtain a soft inverse algorithm, then this algorithm can be converted as necessary. For example, min-sum algorithms can often be derived using pictures and intuition, while sum-product algorithms are the most straightforward to prove analytically.

Consider a toy example of a system and its min-sum soft inverse described in Table 2. This system has two inputs, $a \in \{0, 1\}$ and $b \in \{0, 1, 2, 3\}$ and one output $c \in \{0, 1\}$, with soft-in metrics as defined in the caption. Note that the best of the eight configurations is $(a = 1, b = 1, c = 1)$. It follows that $\text{MSM}[a = 1] = \text{MSM}[b = 1] = \text{MSM}[c = 1] = -5$ and $\text{MO}[a = 1] = -7$, $\text{MO}[b = 1] = -2$, and $\text{MO}[c = 1] = -1$. Other values can be computed in a similar manner. For example, $\text{MSM}[b = 2] = -2$ (when $a = 0$ and $c = 1$) and $\text{MO}[b = 2] = -4$. Note that if one desires the best hard decision for a given variable, the soft-in and soft-out information should be combined and this information should be thresholded (this is the MSM information in the min-sum case and the APP information in the sum-product case). This information is sometimes referred to as *intrinsic* information, while the soft-in/soft-out information to be passed to other soft inverse modules is in extrinsic (likelihood) form.

Table 2. Toy Example of a System with Two Inputs a and b and One Output c. The Soft-In Information is $\text{MI}[a = 0] = 0$, $\text{MI}[a = 1] = 2$, $\text{MI}[b = 0] = 0$, $\text{MI}[b = 1] = -3$, $\text{MI}[b = 2] = 2$, $\text{MI}[b = 3] = 6$, $\text{MI}[c = 0] = 0$, $\text{MI}[c = 1] = -4$

a	b	c	Configuration Metric
0	0	0	$0 + 0 + 0 = 0$
0	1	0	$0 - 3 + 0 = -3$
0	2	1	$0 + 2 - 4 = -2$
0	3	0	$0 + 6 + 0 = 6$
1	0	1	$2 + 0 - 4 = -2$
1	1	1	$2 - 3 - 4 = -5$
1	2	0	$2 + 2 + 0 = 4$
1	3	0	$2 + 6 + 0 = 8$

[3] These are called semi-ring algorithms in Ref. [1].

In summary, the soft inverse of any system is defined based on the optimal receiver processing (MAP) for the system in isolation under the assumption of independent inputs and a memoryless channel. The exact form of the soft inversion problem depends on the optimality criterion (i.e., MAP-SyD or MAP-SqD) and the format used to represent the soft information (i.e., metric or probability domain). For the marginalization-combining operators discussed, the semi-ring properties hold. Thus, most soft inversion algorithms of interest can be converted by simply replacing the combining and marginalization operators.

2.1. Specific Example Sub-Subsystems

It is important to note that the soft inversion [e.g., as stated in Eq. (3)] is a computational problem rather than an algorithm. The problem statement does suggest a method of computing the soft inverse, but this brute-force approach will typically be prohibitively complex. For example, if the system in Fig. 4 has M binary inputs, it will have 2^M configurations. Computing soft information for each of these configurations and then performing the subsequent marginalization will be prohibitively complex even for moderate values of M. This brute-force method is referred to as *exhaustive combining and marginalization*. In many cases, it is possible to compute the soft inverse with dramatically less computational effort by exploiting the special structure of the system. Thus, while there is really one unique computational problem, it is useful to consider special cases of systems, find efficient algorithms for their soft inversion, and then use these as standard modules for iterative detection–based receivers.

For the implicit block diagram convention, Benedetto et al. [13] defined the marginal soft inverse for a variety of systems commonly encountered. These are shown in Fig. 5 with minor modification. Some of these are quite trivial. For example, the soft inverse of an interleaver is an interleaver/deinterleaver pair. Similarly for rate converters, no computation is required for the soft inversion.

The memoryless mapper is a mapper that maps small blocks of inputs onto outputs without memory between blocks. For example, if four bits are collected and mapped onto a 16-ary constellation, then the memoryless mapper is the proper subsystem model. The soft inverse of the memoryless mapper is computed using exhaustive combining and marginalization over the small block size (e.g., 16 configurations in the modulator example). Another example in which the memoryless mapper can be applicable is block error-correction codes. A common special case of the mapper is the broadcaster or repeater where all inputs and outputs are necessarily equal.

One of the most important subsystems is the finite state machine (FSM), which is discussed in detail in the next section. With the modules shown in Fig. 5, a large number

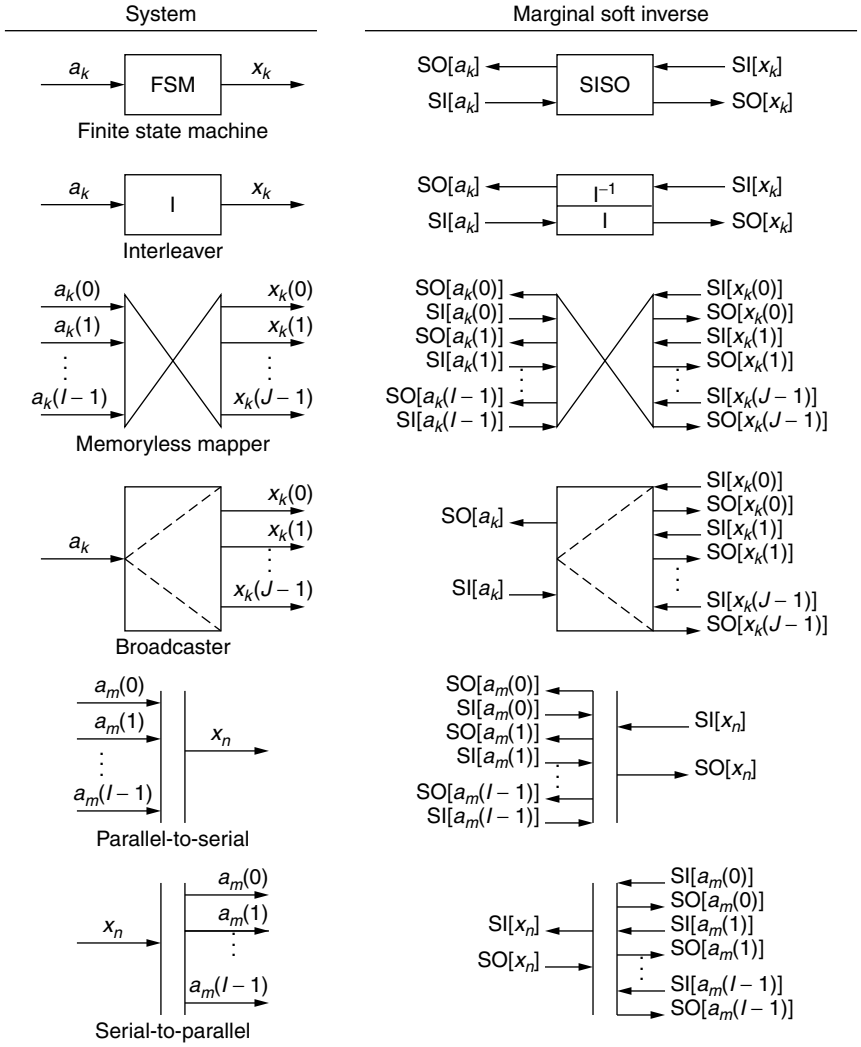


Figure 5. Several common systems and the associated marginal soft inverses in implicit index block diagrams.

of practical systems can be modeled and the corresponding standard iterative receivers defined.

2.1.1. Soft Inversions of an FSM via the Forward–Backward Algorithm. An FSM is a common model for subsystems of digital communications systems. An FSM is a system that at time k has a current state s_k that takes on a finite number of possible values. Application of the input at time k, a_k, results in the system producing output x_k and transitioning to the next state s_{k+1}. The transition at time k is denoted by t_k, which is defined by (s_k, a_k) or with some redundancy (s_k, a_k, s_{k+1}). It is common to represent an FSM by a diagram that is explicit in time and in value, the so-called trellis diagram.

Convolutional codes and related trellis-coded modulation schemes are naturally modeled as FSMs. Certain modulation formats and precoding methods such as differential encoding, line coding, and continuous phase modulation (CPM) are also routinely defined in terms of FSMs. Channel impairments are also often well modeled as FSMs. This includes, for example, the intersymbol interference (ISI) channel [14], the multiple access interference channel [15], and even channels with random fading effects [16]. So, an efficient method for soft inversion of an FSM is desirable.

Using the development of the soft inverse notion and a presumed familiarity with the Viterbi algorithm [17] or similar dynamic programming tools [12], an efficient algorithm can be derived pictorially. This is the *forward–backward algorithm* (*FBA*), which can be used to compute the MSM of the inputs and/or outputs of an FSM [18–20]. First, note that for a given transition, the FSM output x_k and input a_k are specified uniquely. Thus, each well-defined trellis transition can be assigned a transition metric of the form

$$\mathrm{M}_k[t_k] = \mathrm{MI}[a_k(t_k)] + \mathrm{MI}[x_k(t_k)]. \tag{5}$$

Second, the metric of the shortest path through a given transition t_k can be obtained if one has the metric of the shortest path to the states s_k from the left edge of the index range and the metric of the shortest path from the right edge to s_{k+1}. This concept is shown in Fig. 6, where $\mathrm{MSM}_i^j[\cdot]$ denotes the MSM using input metrics over the indices from i to j inclusive. Mathematically, the claim

is that

$$\mathrm{MSM}_0^{K-1}[t_k] = \mathrm{MSM}_0^{k-1}[s_k(t_k)] + \mathrm{M}_k[t_k] + \mathrm{MSM}_{k+1}^{K-1}[s_{k+1}(t_k)] \quad (6)$$

Third, the MSM of the input a_k or output x_k can be obtained by marginalizing (minimizing in this min-sum case) over all transitions t_k consistent with those conditional values. Finally, the quantities $F_{k-1}[s_k] = \mathrm{MSM}_0^{k-1}[s_k]$ and $B_{k+1}[s_{k+1}] = \mathrm{MSM}_{k+1}^{K-1}[s_{k+1}]$ can be updated by a forward recursion and a backward recursion, respectively. The forward recursion is identical to that of the Viterbi algorithm and the backward recursion is the same, only run in reverse.

In summary, the soft inverse of an FSM can be computed via the FBA using three steps: a forward state metric recursion, a backward state metric recursion, and a completion operation that performs the marginalization over transitions to obtain the desired soft outputs. Given the transition metrics defined in (5), the steps are

$$F_k[s_{k+1}] = \min_{t_k : s_{k+1}} \left(F_{k-1}[s_k] + \mathrm{M}_k[t_k]\right) \qquad k = 0, 1, 2, \ldots, K-1 \quad (7a)$$

$$B_k[s_k] = \min_{t_k : s_k} \left(\mathrm{M}_k[t_k] + B_{k+1}[s_{k+1}]\right) \qquad k = K-1, K-2, \ldots, 0 \quad (7b)$$

$$\mathrm{MO}[a_k] = \min_{t_k : a_k} \left(F_{k-1}[s_k] + \mathrm{M}_k[t_k] + B_{k+1}[s_{k+1}]\right) - \mathrm{MI}[a_k] \quad (7c)$$

$$\mathrm{MO}[x_k] = \min_{t_k : x_k} \left(F_{k-1}[s_k] + \mathrm{M}_k[t_k] + B_{k+1}[s_{k+1}]\right) - \mathrm{MI}[x_k] \quad (7d)$$

Initialization of the forward and backward metrics is performed according to available initial edge information [1]. This is a semi-ring algorithm, so it may be directly converted to other marginalization-combining forms by simply replacing the min-sum operations appropriately. One step through a four-state trellis for the FBA is shown in Fig. 7.

In hardware implementations, it often is useful to approximate the FBA using only a portion of the soft inputs to compute a given soft output. For example,

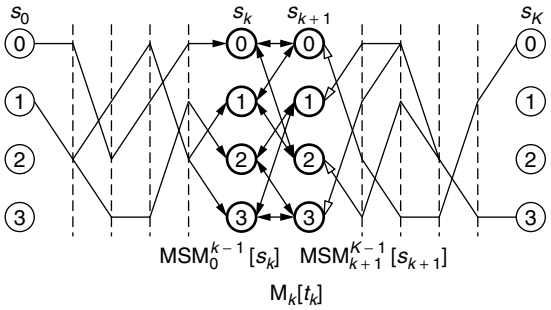


Figure 6. The MSM for a given transition may be computed by summing the transition metric and the forward and backward state metrics.

a fixed-lag algorithm uses soft-in information for times $\{0, 1, \ldots, k+D\}$ to compute the soft-out information for the input or output variables at time k. For sufficiently large lag D, no significant performance degradation will occur. A particularly attractive algorithm for hardware is the min-lag/max-lag algorithm suggested by Viterbi [21]. In this case, the lag varies with index k, but is bounded below by D and above by $2D$. This algorithm can be implemented in hardware with one forward state metric processor and two backward state metric processors in such a way that computation, memory, and speed are attractively balanced [21,22].

Finally, note that the FBA is just one algorithm for computing the soft inverse of an FSM. While the FBA solution has low complexity, it has a bottleneck in the state metric recursions (i.e., the "ACS bottleneck"). An alternative algorithm that computes the soft inverse based on a low-latency tree structure was suggested in Refs. 23 and 24.

3. STANDARD RULES FOR ITERATIVE DETECTION

The standard iterative detection technique can be summarized as follows:

- Given a system comprising a concatenated network of subsystems, construct the marginal soft inverse of each subsystem. The marginal soft inverse is found by considering the subsystem in isolation with independent inputs and a memoryless channel. Using these operators, specify an algorithm to compute the extrinsic soft outputs for the system inputs and outputs.
- Construct the block diagram of the iterative detector by replacing each subsystem by the corresponding marginal soft inverse and connecting these soft inverses accordingly. Specifically, each connection between subsystems in the system block diagram is replaced by a corresponding pair of connections in the iterative detector block diagram so that the soft-out port of each is connected to the soft-in port of the other.
- Specify an activation schedule that begins by activating the soft inverses corresponding to some subsystems providing the global outputs and ends with activation of some soft inverses corresponding to subsystems with global inputs.
- Specify a stopping criterion.
- Take the soft-inputs on global output symbols as the channel likelihoods (metrics) obtained by appropriate soft-demodulation. The soft-inputs for the global inputs are the a priori probabilities (metrics), which are typically uniform.
- At the activation of each subsystem soft inverse, take as the soft-in on the digital inputs/outputs the soft-outputs from connected subsystem soft inverses. If no soft information is available at the soft-in port, take this to be uniform soft-in information (i.e., this applies to the first activation of soft inverses that

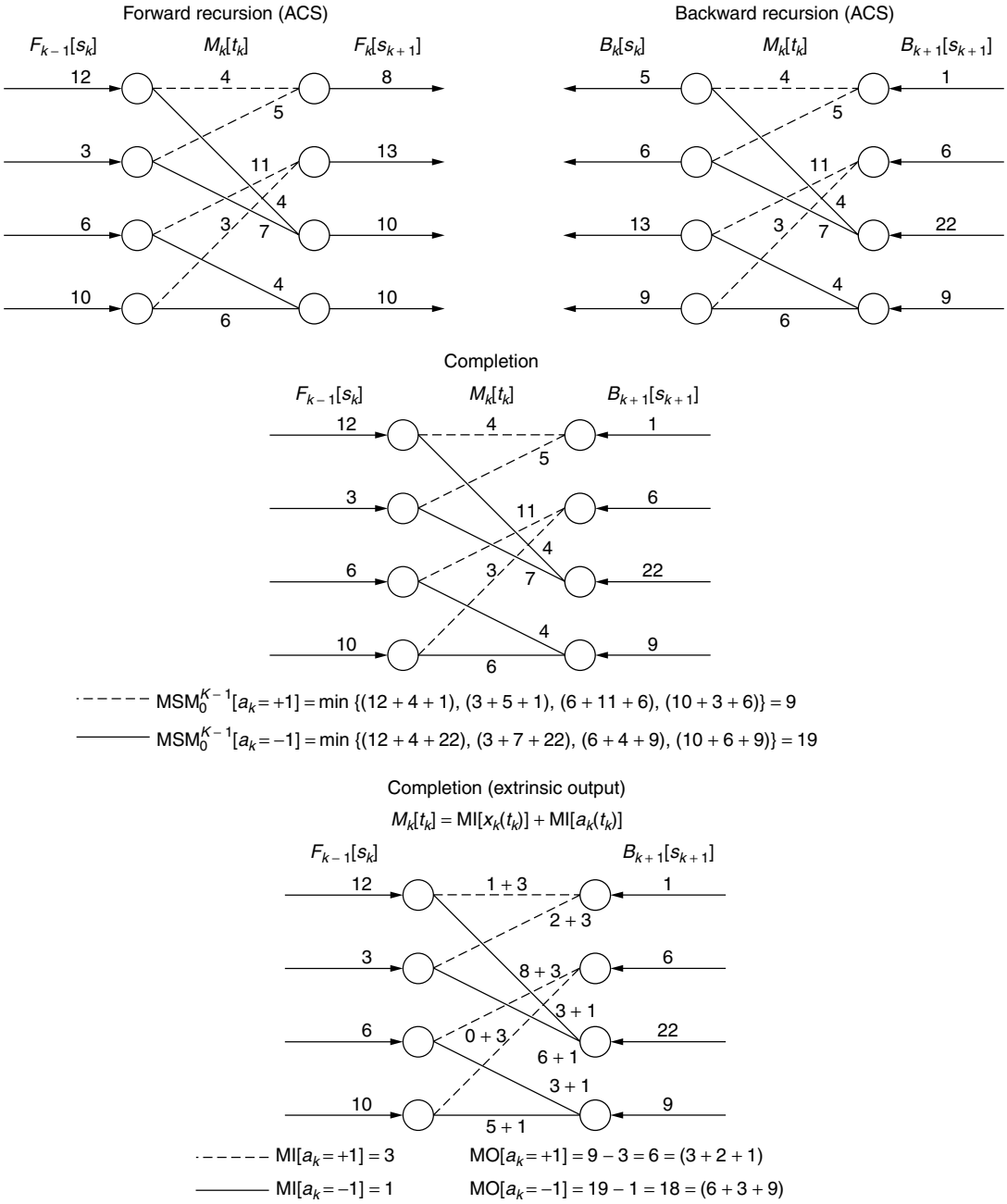


Figure 7. An example of the forward–backward algorithm processing steps.

have inputs or outputs that are internal or hidden variables).

A common stopping criterion is that a fixed number of iterations are to be performed, with this number determined by computer simulation. For example, while formal proofs of convergence for complicated iterative detectors are difficult, in most cases of practical interest, the performance improvement from iteration reaches a point of diminishing returns. Alternatively, performance may be sacrificed to reduce the number of iterations. It also is possible to define a stopping rule that results in variable number of iterations.

In most cases of practical interest, there either is a natural activation schedule or different activation schedules produce similar results. Thus, for the most part, the iterative detector is specified once the block diagram is given and the subsystem soft inverses are determined. For example, the iterative detector for the general concatenated system in Fig. 2 is shown in Fig. 3.

As a simple example of this paradigm, a simple turbo code, or parallel concatenated convolutional code (PCCC) is considered. The encoder is shown in Fig. 8 and the corresponding standard decoder is shown in Fig. 9. The blocks used in the encoder are two FSMs, a one-to-two broadcaster and a puncture/binary mapper. The

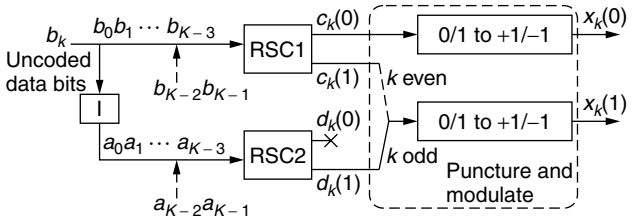


Figure 8. A rate one-half PCCC encoder with four state constituent recursive systematic convolutional codes.

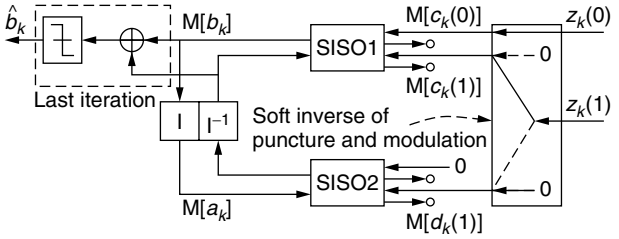


Figure 9. The min-sum iterative decoder for the PCCC in Fig. 8.

decoder exploits the special properties of a one-to-two broadcaster and efficient min-sum metric representation using normalization techniques. Even such a simple code is capable of approaching the theoretical limits within 1 dB of SNR.

3.1. Sufficient Condition for Optimality of the Standard Rules

Note that the iterative detector in Fig. 3 has the same input/output format as a soft inverse. Specifically, the detector takes in soft-in information on the global system inputs and outputs and performs an update to produce soft-out information on these same variables. Since the soft inverse is defined relative to well-established optimal receiver criteria, the natural question arises: *When does applying the standard iterative detection rules to a system model yield the global system soft inverse, and hence the optimal receiver for the global system?* Interestingly, there is a simple sufficient condition for this to occur. This is developed through the following examples.

Note that the general rules were not dependent on the implicit index convention and apply as well to explicit index diagrams. Consider, for example, the explicit index diagram for the general FSM as sown in Fig. 10 (a) in which the FSM has been decomposed or modeled by a series of small transition nodes or subsystems, each defining one transition in the trellis. Applying the standard definitions, it can be shown that the soft inverse of each of these transition nodes performs one forward state recursion step, one backward state recursion step, and a completion step for both the FSM input and output. This is illustrated in Fig. 11. As a result, running the standard iterative detection rules on the receiver shown in Fig. 10 (b), with a specific activation schedule, yields exactly the FBA-based soft inverse of the FSM. Even more remarkable, it can be shown that after some point, further activation of the soft inverse nodes does not change the soft information. In fact, the soft information will stabilize

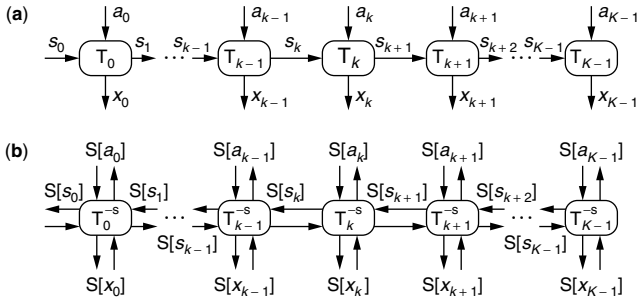


Figure 10. (**a**) An explicit index block diagram for an arbitrary FSM, and (**b**) the associated concatenated detector.

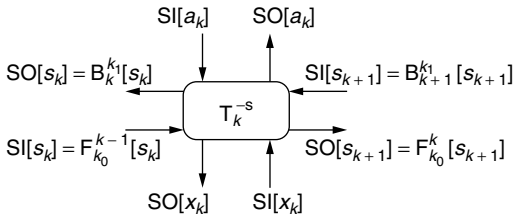


Figure 11. The soft inverse of the transition subsystem. Activation is equivalent to one update of the backward and forward state metric recursions, and completion for both the FSM input and output.

under any activation schedule as long as it satisfies some basic requirements. Intuitively, the soft-in information for each node must be passed to all other nodes before the values stabilize. So, in this example, the globally optimal solution is achieved by applying the standard rules to a concatenated system model.

This is not the case in general. For example, for the iterative decoder shown in Fig. 9, one will observe conditions where the soft information passed does not completely stabilize. Consider the explicit index diagram for the PCCC encoder of Fig. 8 as illustrated in Fig. 12. Note that this contains two FSM subgraphs of the form in Fig. 10 (a), but with connections due to the interleaver. Ignoring the directions of the arrows, one can trace a loop by following the connections in the diagram. Intuitively,

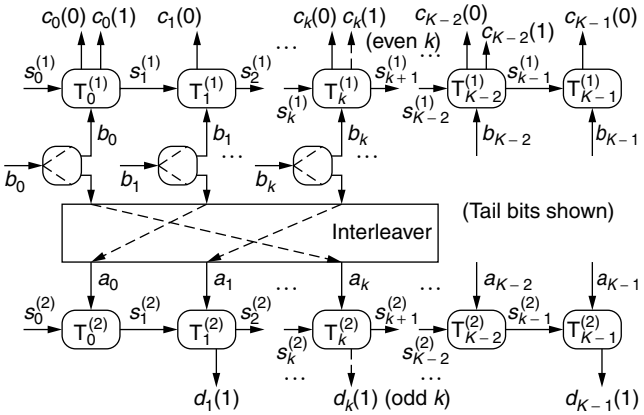


Figure 12. The explicit index block diagram of the PCCC shown in Fig. 8.

one may understand how this could compromise the optimality of the global detector. Specifically, the initial belief obtained from the channel for a particular variable may propagate through the processing and return to be "double counted" by the original processing node. This can lead to convergence to a false minimum.

In fact, it can be shown rigorously that if no cycles exist in the explicit index model (or equivalently the graphical model), then under some simple requirements on the activation schedule, the standard rules result in convergence to the globally optimal solution. This fact was known in computer science for some time [2,25]. Two remarkable facts did arise, however, from the research efforts in communications and coding. First, at a conceptual level, there really is only one algorithm. For example, the FBA and the turbo decoding algorithm are both just examples of running the standard rules on different system models [3,5,8,26]. Second, although this processing on graphical models with loops is suboptimal in general, it is highly effective in many practical scenarios. This can be motivated intuitively in a manner similar to the finite traceback approximation in the Viterbi algorithm or the fixed-lag approximation in the FBA. Specifically, if the local neighborhood of all nodes is cycle-free, then one can expect the processing to well-approximate the optimal solution.

In summary, there is one standard iterative detection algorithm that is optimal when applied to cycle-free graphs and generally suboptimal, yet very effective, when applied to graphical models with loops. One reasonable convention for terminology is to refer to the algorithm on cycle-free graphs as *message passing* and on loopy graphs as *iterative message passing*. Also note that selection of the system model is critical in determining the complexity and performance of the algorithm. For example, use of the approach on models with cycles is most effective when there exists some sparse structure in the system that can be exploited locally by proper reindexing of the variables.

3.2. Modified Rules

Despite the elegance of the single set of message-passing rules, in some applications modification of these rules is necessary or desirable. One example is when there is some uncertainty in the structure of one or more subsystems. This occurs, for example, when some channel parameters are not completely known. In this case, the input–output relation for subsystems in contact with the channel may not be completely defined. This requires consideration of the proper marginal soft inverse definition in the presence of parametric uncertainty. Typically, the theory will suggest that exhaustive combining and marginalization is required to perform soft inversion in the presence of unknown parameters. However, greedy approximations to this processing work well in practice. Such solutions typically are based on applying decision feedback or memory truncation techniques to recursive formulations of the soft inverse computation. This yields practical *adaptive SISO* modules that approximate the soft inverse. Incorporating the parameter estimation and tracking tasks into the iteration process is a powerful tool. Such *adaptive iterative detection* algorithms allow tracking of severe parameter dynamics at very low SNR. Work in this area can be found in [1,27–31].

A similar situation can arise when all parameters are known, but there is a subsystem that has a prohibitively complex soft inverse. In such cases, one can attempt to approximate the soft inverse with an algorithm of reasonable complexity. One important case is an FSM subsystem with a large number of states. One approach is to use decision feedback techniques similar to those used in reduced state sequence detectors [32]. There are a number of variations on this theme suggested for *reduced state SISO* algorithms primarily applied to ISI mitigation (equalization) [33,34]. Another approach is to use an approximate soft inverse based on a constrained receiver structure. For example, one may use a linear or decision feedback equalizer, modified to update soft information, in place of an FBA-based soft inverse [35].

Another case in which some modification of the rules can be helpful is when there are short cycles in the underlying graphical models. In such cases, it is often observed that convergence occurs quickly, but performance is poor. As is the case in many iterative algorithms, this effect can be alleviated somewhat by attempting to slow down the convergence of the algorithm. This can be accomplished in a number of ways. For example, filtering the messages over iterations so as to slow their evolution or simply scaling them to degrade the associated confidence can provide significant improvements in applications where the basic approximations break down [1].

4. APPLICATIONS AND IMPACT

Once a general view of the iterative detection algorithm is understood, applying the approach to various problems is relatively straightforward. A detailed description of these applications is beyond the scope of this article. Instead, a brief summary of the applications, the performance gains, complexity issues, and representative references is given in this section.

- **Turbo-Like Codes:** Following the invention of turbo codes, many variations on the theme were described in the literature. This includes serial concatenated convolutional codes (SCCCs) [36], low-density parity check (LDPC) codes [37,38], and product codes [39]. The most significant difference in the decoding algorithm from what has been presented herein is the use of indirect system models to perform the soft inversion. For example, it is possible to perform soft inversion using the parity check structure of a block code. This is possible because one needs only to be able to identify allowable configurations to combine and marginalize over. Decoding of LDPC codes and product codes based on high-rate block codes can be performed efficiently using this approach. Codes similar in performance and structure to LDPC codes, but with simpler encoding have also been suggested. These include the generalized repeat-accumulate codes [40,41] and parallel concatenated zig-zag codes [42], which are both based on a recursive single parity check codes and broadcasters.

Code design techniques have developed to the point where turbo-like codes are attractive alternatives to previous approaches at virtually all code rates and reasonable input block sizes (e.g., >64). For reasonable block sizes and rates, it is not unusual to achieve 3 to 8 dB of additional coding gain relative to more conventional FEC designs. Codes based on PCCCs and SCCCs are commonly adopted for standardized systems. In the coding application, code construction allows a designer to achieve *interleaver gain* as well as *iteration gain*. A system with interleaver gain will perform better as the size of the interleaver increases, while iteration gain simply means that iteration improves the performance.

- **Modulation and Coding:** Based on the knowledge of the design rules for turbo-like codes and the standard iterative detection paradigm, several common modulation and coding system designs have been demonstrated to benefit greatly from iterative processing. One example is the serial concatenation of a convolutional code, an interleaver, and a recursive inner modulator, which can be viewed as an effective, simple SCCC. This has been exploited in the case of the inner system being a differential encoder [43] and in the case of the inner modulation being CPM, which has a recursive representation [44,45]. Similarly, simple schemes such as bit interleaved coded modulation (BICM), where coded bits are interleaved and then mapped onto a nonbinary constellation, have been shown to benefit significantly from iterative decoding/demodulation [46]. The benefits in these various applications range from 1 to 6 dB of SNR improvement for typical scenarios.
- **Equalization and Decoding:** Many systems are designed with coding, interleaving, and a propagation channel that results in ISI. A number of researchers realized the applicability of the standard iterative approach to such a scenario, typically with convolutional coding and a relatively short ISI channel delay spread [47–49]. In this case, the ISI and code SISOs are both implemented using the FBA. For typical scenarios, soft-out equalization provides approximately 2.5 dB of gain in SNR over hard decision processing and iteration provides another 4 dB in SNR. If the channel is fading, then most of the iteration gain is lost when the average performance is considered. This is because the worst-case fading conditions, for which there is little iteration gain, dominate the performance [1]. In the case of severe channel dynamics, significant iteration gain will be achieved in fading if a properly designed adaptive iterative detection algorithm is used. Unless combined with a turbo-like code or equivalent modulation techniques, there is no interleaver gain in this application.
- **Interference Mitigation:** Similar to the previous application, one can use the standard iterative algorithm to perform joint decoding and like-signal interference mitigation. Specifically, considering the case where each user's data is coded and interleaved, they can be effectively separated by using an interference mitigation SISO module and a bank of code SISO modules. Utilizing the code structure is especially helpful if there is a high degree of correlation between the signals on the multiple access channel (i.e., the channel is heavily loaded). In fact, it has been demonstrated that users can be separated when the only unique feature is an interleaver pattern (see [1] and references therein).

In applications where the operating SNR is very low, there can be an advantage to using $\min^*$-sum processing over min-sum processing. This advantage typically is 0.2 to 1.0 dB for these applications, which are the turbo-like codes and similar coding-modulation constructions. For other applications, such as joint equalization and decoding or interference mitigation, there is little practical advantage to using $\min^*$-sum processing over min-sum processing.

The number of iterations required varies significantly with application as well. Generally, systems with weak local structure converge more slowly. This is the case with LDPC and similar codes that are based on single-parity-check codes. For a typical SCCC or PCCC 6 to 10 iterations will yield the majority of the gains for most practical scenarios. In many of the interference mitigation and equalization applications, most of the gains are achieved with 3 to 5 iterations.

In all cases, the complexity increase relative to a standard solution is moderate. In the case of turbo-like codes, it is not uncommon to achieve better performance with less complexity than conventional designs. In other applications, the soft inverse typically is 2 to 4 times as complex as a subsystem processor based on the segregated design paradigm. Substantially more memory is also required. While not insignificant, these increases in memory and computation requirements impact digital circuitry, which continues to experience a steady, rapid improvement in capabilities.

5. CONCLUSION

Iterating soft-decision information to improve the performance of practical digital communication systems can be motivated as an approach to approximate the optimal receiver and improve upon the performance of the segregated design. This intuitive notion can be formalized using the standard tools of communication receiver design, namely, Bayesian decision theory. This results in a standard approach that is based on the notion of the soft inverse of a system and the exchange and update of soft information. This approach, viewed generally, describes a standard algorithm for optimal or near-optimal data detection for complex systems. A sufficient condition for optimality (i.e., minimum error probability) is that the underlying graphical model describing dependencies between system variables be cycle-free. In the case where cycles exist, which is the case in many practical applications of interest, the performance of the standard approach is often extremely good. The approach has been demonstrated to improve performance substantially in a number of relevant applications. The general approach is becoming

better known to researchers in the field, and the results are finding adoption in engineering practice rather quickly.

Acknowledgments
The author thanks John Proakis for his encouragement and understanding during the preparation of this article; Kluwer Academic Publishers for permission to use material from [1]; and the co-authors of [1], Achilleas Anastasopoulos and Xiaopeng Chen.

BIBLIOGRAPHY

1. K. M. Chugg, A. Anastasopoulos, and X. Chen, *Iterative Detection: Adaptivity, Complexity Reduction, and Applications*, Kluwer Academic Publishers, 2001.
2. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, 1988.
3. R. J. McEliece, D. J. C. MacKay, and J. F. Cheng, Turbo decoding as an instance of Pearl's "belief propagation" algorithm, *IEEE J. Select. Areas Commun.* **16**: 140–152 (Feb. 1998).
4. F. Kschischang and B. Frey, Iterative decoding of compound codes by probability propagation in graphical models, *IEEE J. Select. Areas Commun.* **16**: 219–231 (Feb. 1998).
5. S. M. Aji and R. J. McEliece, The generalized distributive law, *IEEE Trans. Inform. Theory* **46**: 325–343 (March 2000).
6. C. Berrou, A. Glavieux, and P. Thitmajshima, Near shannon limit error-correcting coding and decoding: turbo-codes, in *Proc. International Conf. Communications*, (Geneva, Switzerland), pp. 1064–1070, May 1993.
7. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: turbo-codes, *IEEE Trans. Commun.* **44**: 1261–1271 (Oct. 1996).
8. N. Wiberg, *Codes and Decoding on General Graphs*. PhD thesis, Linköping University (Sweden), 1996.
9. T. Richardson and R. Urbanke, The capacity of low-density parity-check codes under message-passing decoding, *IEEE Trans. Inform. Theory* **47**: 599–618 (Feb. 2001).
10. H. E. Gamal and J. A. R. Hammons, Analyzing the turbo decoder using the Gaussian approximation, *IEEE Trans. Inform. Theory* **47**: 671–686 (Feb. 2001).
11. P. Robertson, E. Villebrum, and P. Hoeher, A comparison of optimal and suboptimal MAP decoding algorithms operating in the log domain, in *Proc. International Conf. Communications*, (Seattle, WA), pp. 1009–1013, 1995.
12. T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, Cambridge, Mass.: MIT Press, 1990.
13. S. Benedetto, G. Montorsi, D. Divsalar, and F. Pollara, Soft-input soft-output modules for the construction and distributed iterative decoding of code networks, *European Trans. Telecommun.* **9**: 155–172 (March/Apr. 1998).
14. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 284–287 (May 1972).
15. S. Verdú, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **32**: 85–96 (Jan. 1986).
16. J. Lodge and M. Moher, Maximum likelihood estimation of CPM signals transmitted over Rayleigh flat fading channels, *IEEE Trans. Commun.* **38**: 787–794 (June 1990).
17. G. D. Forney, Jr., The Viterbi algorithm, *Proc. IEEE* **61**: 268–278 (March 1973).
18. R. W. Chang and J. C. Hancock, On receiver structures for channels having memory, *IEEE Trans. Inform. Theory* **IT-12**: 463–468 (Oct. 1966).
19. P. L. McAdam, L. R. Welch, and C. L. Weber, M.A.P. bit decoding of convolutional codes, *Proc. IEEE Int. Symp. Info. Theory* (1972).
20. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).
21. A. J. Viterbi, Justification and implementation of the MAP decoder for convolutional codes, *IEEE J. Select. Areas Commun.* **16**: 260–264 (Feb. 1998).
22. G. Masera, G. Piccinini, M. R. Roch, and M. Zamboni, VLSI architectures for turbo codes, *IEEE Trans. VLSI* **7**: (Sept. 1999).
23. P. A. Beerel and K. M. Chugg, A low latency SISO with application to broadband turbo decoding, *IEEE J. Select. Areas Commun.* **19**: 860–870 (May 2001).
24. P. Thiennviboon and K. M. Chugg, A low-latency SISO via message passing on a binary tree, in *Proc. Allerton Conf. Commun., Control, Comp.* (Oct. 2000).
25. F. V. Jensen, *An Introduction to Bayesian Networks*, Springer-Verlag, 1996.
26. F. Kschischang, B. Frey, and H.-A. Loeliger, Factor graphs and the sum-product algorithm, *IEEE Trans. Inform. Theory* **47**: 498–519 (Feb. 2001).
27. A. Anastasopoulos and K. M. Chugg, Adaptive soft-input soft-output algorithms for iterative detection with parametric uncertainty, *IEEE Trans. Commun.* **48**: 1638–1649 (Oct. 2000).
28. A. Anastasopoulos and K. M. Chugg, Adaptive iterative detection for phase tracking in turbo-coded systems, *IEEE Trans. Commun.* **49**: 2135–2144 (Dec. 2001).
29. M. C. Valenti and B. D. Woerner, Refined channel estimation for coherent detection of turbo codes over flat-fading channels, *IEE Electron. Lett.* **34**: 1033–1039 (Aug. 1998).
30. J. Garcí a-Frí as and J. Villasenor, Joint turbo decoding and estimation of hidden Markov sources, *IEEE J. Select. Areas Commun.* 1671–1679 (Sept. 2001).
31. G. Colavolpe, G. Ferrari, and R. Raheli, Noncoherent iterative (turbo) detection, *IEEE Trans. Commun.* **48**: 1488–1498 (Sept. 2000).
32. M. V. Eyuboğlu and S. U. Qureshi, Reduced-state sequence estimation with set partitioning and decision feedback, *IEEE Trans. Commun.* **COM-38**: 13–20 (Jan. 1988).
33. X. Chen and K. M. Chugg, Reduced state soft-in/soft-out algorithms for complexity reduction in iterative and non-iterative data detection, in *Proc. International Conf. Communications*, (New Orleans, LA), 2000.
34. P. Thiennviboon, G. Ferrari, and K. Chugg, Generalized trellis-based reduced-state soft-input/soft-output algorithms, in *Proc. International Conf. Communications*, (New York), pp. 1667–1671, May 2002.

35. M. Tuchler, R. Koetter, and A. Singer, Turbo equalization: principles and new results, *IEEE Trans. Commun.* **50**: 754–767 (May 2002).

36. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* **44**: 909–926 (May 1998).

37. R. G. Gallager, Low density parity check codes, *IEEE Trans. Inform. Theory* **8**: 21–28 (Jan. 1962).

38. D. J. C. MacKay, Good error-correcting codes based on very sparse matrices, *IEE Electron. Lett.* **33**: 457–458 (March 1997).

39. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**: 429–445 (March 1996).

40. H. Jin, A. Khandekar, and R. McEliece, Irregular repeat accumulate codes, in *Turbo Code Conf.*, (Brest, France), 2000.

41. K. R. Narayanan, I. Altunbas, and R. Narayanaswami, On the design of LDPC codes for MSK, in *Proc. Globecom Conf.*, (San Antonio, TX), pp. 1011–1015, Nov. 2001.

42. L. Ping, X. Huang, and N. Phamdo, Zigzag codes and concatenated zigzag codes, *IEEE Trans. Inform. Theory* **47**: 800–807 (Feb. 2001).

43. P. Hoeher and J. Lodge, Turbo DPSK: iterative differential PSK demodulation and channel decoding, *IEEE Trans. Commun.* **47**: 837–843 (June 1999).

44. K. Narayanan and G. Stuber, Performance of trellis-coded CPM with iterative demodulation and decoding, *IEEE Trans. Commun.* **49**: 676–687 (Apr. 2001).

45. P. Moqvist and T. Aulin, Serially concatenated continuous phase modulation with iterative decoding, *IEEE Trans. Commun.* **49**: 1901–1915 (Nov. 2001).

46. X. Li and J. A. Ritcey, Trellis-coded modulation with bit interleaving and iterative decoding, *IEEE J. Select. Areas Commun.* **17**: 715–724 (Apr. 1999).

47. C. Douillard, Iterative correction of intersymbol interference: Turbo equalization, *European Trans. Telecommun.* **6**: 507–511 (Sept. 1995).

48. A. Anastasopoulos and K. M. Chugg, Iterative equalization/decoding of TCM for frequency-selective fading channels, in *Proc. Asilomar Conf. Signals, Systems, Comp.*, pp. 177–181, Nov. 1997.

49. A. Picart, P. Didier, and A. Glavieux, Turbo-detection: a new approach to combat channel frequency selectivity, in *Proc. International Conf. Communications*, (Montreal, Canada), 1997.

ITERATIVE DETECTION METHODS FOR MULTIUSER DIRECT-SEQUENCE CDMA SYSTEMS

LARS K. RASMUSSEN
University of South Australia
Mawson Lakes, Australia

1. INTRODUCTION

The continuing development of the Internet, wireless communication and wireless Internet is rapidly increasing the demands on enabling communication networks. The continuous development of ever-faster computers allows for the development of ever larger communication systems to meet the ever-increasing demand for capacity to support the never-ending supply of new communications services. The third-generation (3G) mobile network, the so-called IMT2000 (International Mobile Telecommunication 2000) system, is the next step to bringing wireless Internet to the general consumer [1]. The 3G cellular mobile network provides up to 2 megabits per second (Mbps) for indoor environments and 144 kilobits per second (kbps) for vehicular environments. The two dominating standards for 3G networks are UMTS (Universal Mobile Telephone System) and cdma2000, respectively [1]. They are both based on direct sequence, code-division multiple-access (DSCDMA) technology, which was considered to provide the best alternative within the standardization process. DSCDMA is a spread-spectrum transmission technology where each user in principle is assigned a unique signature waveform, creating a distinguishing feature separating multiple users and thus providing multiple access [2].

In popular terms, these principles may be likened to conversations at a cocktail party where each conversation is conducted in a unique language. Even though a particular conversation can be clearly heard by surrounding people engaged in other conversations, they do not become disturbed since they do not understand the language and thus consider it as background noise. In case two languages are closely related, two simultaneous conversations may interfere with each other. The same phenomenon occurs in CDMA if two users are assigned signature waveforms that are closely correlated. This is termed *multiple-access interference* (MAI) and is one of the performance-limiting factors when conventional single-user technologies are used in CDMA systems [3,4]. The basics of spread-spectrum and CDMA technologies are described in more details elsewhere in this book. For further information, see also Refs. 1,2,5 and 6.

Third-generation mobile cellular systems are designed for multimedia communications. The standards of person-to-person communications can be improved through better voice quality and the possibility of exchanging high-quality images and video. In addition, access to information and services on public and private networks will improve through higher data rates and variable data rate options, introducing a high level of flexibility. Through the standardization process, CDMA technologies came out as the overall winners. The UMTS system is based on so-called wideband CDMA (WCDMA) [7]. There are no conceptual differences between WCDMA and CDMA. The former merely uses a bandwidth that is significantly larger than second-generation CDMA systems, leading to additional spread-spectrum advantages such a robustness toward hostile mobile radio channels.

The flexible data rates, and especially the high data rates of 2 Mbps, represent significant challenges for equipment manufacturers. For the system load to be commercially viable, technologies providing considerable system capacity improvements are required. Three areas of technology have been identified as enabling techniques for 3G CDMA systems:

- Error control coding [8,9]
- Multiuser detection [5]
- Space–time processing [10]

In this article, we discuss the use of multiuser detection for CDMA systems in general, and the use of iterative multiuser detection strategies in particular.

1.1. Multiuser Detection

Historically, multiple-access systems have been designed to avoid the MAI problem. This is accomplished by dividing the available system resources into dedicated portions for the exclusive use by a designated communication connection. In a CDMA system, we break with these principles and allow users to utilize all resources simultaneously. Under certain system conditions, it is possible to avoid MAI. However, such conditions are in general impossible to achieve in a practical setting, and thus, a certain level of MAI is to be expected.

Initially, the MAI was considered as unavoidable interference and assumed to possess similar statistical characteristics as thermal background noise generated in electronic components. Based on such arguments, the optimal receiver structures developed for the case of thermal noise only, are also optimal in the case of MAI with similar statistical behavior [11]. The performance of such systems are determined by the signal power to noise power ratio. It is therefore tempting to increase the transmitted signal power to improve performance. However, if all active users do that, the power of the interfering MAI increases with the same amount, providing no performance gains at all. On the basis of these assumptions and techniques, it follows that the systems are interference limited. A strict upper limit on active users, corresponding to a certain signal to total noise level, decides the system capacity [5].

The problem with this approach is that each active user is making decisions in isolation, regarding the corresponding MAI as unstructured noise. As an alternative, a joint decision among all users simultaneously takes the known structure of the MAI into account, treating the MAI as additional information in the decision process. Assuming binary transmission (two possible signal waveform alternatives for transmission for each user), a decision made in isolation is based on a choice between which of the two signal waveform alternatives were transmitted, ignoring the known structure of the MAI. In contrast, a joint multiuser decision is based on evaluating a suitable cost function for all possible transmitted waveform combinations between the active users, selecting the combination that maximizes the cost function. For optimal detection, we shall attempt to maximize the probability of a certain combination of data symbols being transmitted, given the particular received signal. Assuming that all data symbols are equally likely, this optimization criterion is equivalent to maximizing the corresponding likelihood function [11]. For three active users, each using binary transmission, we need to evaluate the likelihood function for 2^3 combinations of data symbols, or in general for K active users, 2^K combinations of data symbols in order to find the combination that is maximum likely. Maximum-likelihood (ML) joint multiuser detection has been suggested by several authors in the literature [12,13], most notably by Verdú [14]. Verdú [14] suggested the use of the Viterbi search algorithm [16] for detection was as an efficient implementation of the exhaustive search.

The ML problem is known to be a so-called NP-hard problem [16] and can be solved only by an exhaustive search as described above, leading to a detection complexity that grows exponentially with the number of users. In some important cases, this complexity growth is beyond practical implementation and will remain so for the foreseeable future. To address this complexity problem, an abundance of receiver structures have been proposed [5,17]. Most of these structures are sub-optimal approximations to classic design criteria.

Among the first complexity reducing schemes, the decorrelating detector was suggested [18]. The detector decorrelates the data symbols, removing the MAI entirely through linear filtering. The filter is determined by the inverse of the channel matrix. The filter removes all MAI at the expense of noise enhancement. A slightly different approach is taken by the linear minimum mean-squared error (LMMSE) detector, which minimizes the mean squared error between detector output and the transmitted symbol [19]. This detector takes the thermal noise as well as the correlation between users into account and therefore generally performs better than the decorrelator in terms of bit error rate (BER). Both the decorrelator and the LMMSE detector require matrix inversion, which is generally also prohibitively complex for practical implementation of realistically sized systems. As an alternative, the LMMSE detector can be approximated by adaptive detectors such as described in the literature [20–22].

Since even linear detectors have complexity problems, multiuser detection for CDMA was initially considered to be prohibitively complex for practical implementation. With the massive research conducted in conjunction with IMT2000, however, this view has now changed. For practical implementation, interference cancellation schemes have been subject to most attention. These techniques rely on simple processing elements constructed around conventional receiver concepts. The main component in a conventional receiver is a filter matched to the user-specific signaling waveform. An estimate of the contribution of a specific user to the composite received signal of all users can be generated based on the corresponding matched-filter output. For a specific user, we can now subtract the contributions to the received signal made from all other users. If this estimate of the MAI experienced by that specific user is correct, we can effectively eliminate the disturbance, obtaining a clean signal with no MAI. In practise the MAI estimate is rarely completely correct, leaving some residual interference. Cancellation can then be done iteratively to improve the quality of the resulting signal for detection of a specific user [23,24]. A family of structures can be defined based on how MAI estimates are generated [23,25,26].

Let us assume that we have tentative decisions for the data symbols for all users. On the basis of these

decisions we can make an estimate of the MAI experienced by each user, and subtract the corresponding estimates from the received signal. With K parallel processing units, one iteration for each user can be done in parallel. This approach is naturally denoted parallel interference cancellation (PIC) [23]. As an alternative, we can process one user at a time. From a single cancellation operation, we get an updated tentative decision for a specific user that can now be used in the process of generating an updated MAI estimate for the following user. This way, the most updated information is always used in the cancellation process. In this approach, the users are processed successively, introducing a detection delay between users. The strategy is known as successive interference cancellation (SIC) [25,26].

In the discussion above, we have assumed we have tentative decisions available. Tentative decisions are obtained based on so-called decision statistics, which are passed through a corresponding tentative decision function. The collection of all matched filter outputs represents a *sufficient statistic*, including all information required for making an optimal ML decision for all users simultaneously. The cancellation process, however, is not in general an iterative *joint* detection process. The decision statistics resulting from cancellation are thus not necessarily representing sufficient statistics. In some special cases, such as linear cancellation [27], however, they are.

Assume that each user is transmitting binary data d_k, for example represented by $d_k = +1$ or $d_k = -1$. The corresponding received decision statistic then contains contributions from the desired transmitted data symbol, noise generated in the receiver and corresponding MAI. On the basis of this received decision statistic, a tentative decision is to be made. In the literature mainly four different tentative decision functions have been suggested for interference cancellation. These are shown in Fig. 1. The first one is a linear decision where in fact the decision statistic is left untouched [28,29]. Alternatively, a hard decision can be made where it is decided whether a $+1$ or a -1 was transmitted [23,25]. This is done on the basis of the polarity of the decision statistic. These two principles can be combined to generate the linear clip function, which is piecewise linear [30]. A similar shape can be obtained on the basis of nonlinear principles as shown in Fig. 1 for a hyperbolic tangent function [31,32]. It should be noted that once the cancellation process is completed, hard decisions are applied to the resulting decision statistics in case a final decision is required. In some cases, the soft cancellation output is used directly as input to an error control decoder [33], in which case no final decisions are made in the iterative detector.

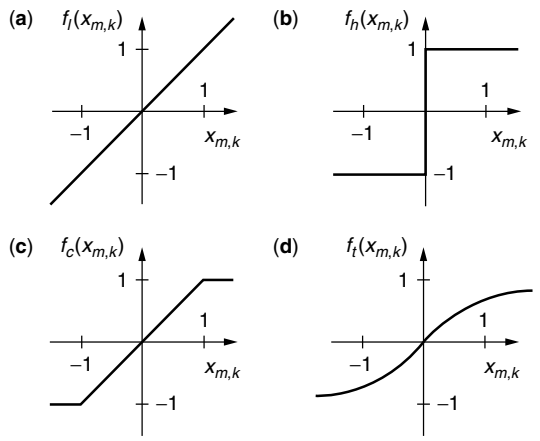


Figure 1. Tentative decision functions for interference cancellation: (**a**) linear; (**b**) hard; (**c**) clipped linear; (**d**) hyperbolic tangent.

The cancellation strategy and the tentative decision function define the character of an interference cancellation structure. The best performance in terms of bit error rate is obtained by SIC schemes as compared to PIC schemes. This advantage is, however, achieved at the expense of detection delays due to the successive processing. For the tentative decision functions, the linear clip and the hyperbolic tangent generally provide better performance than do linear and hard decisions. These issues are discussed at length in the remainder of the article.

The rest of the article is organized as follows. In Section 2, an algebraic model is derived for a simple synchronous CDMA system. The model is kept simple to more clearly illustrate the principles of iterative multiuser detection. In Section 3 the fundamental principles for interference cancellation are formalized and motivated. Corresponding modular structures are suggested, constructed around a simple interference cancellation unit. The hard tentative decision function is discussed in Section 4, while the concept of weighted cancellation is introduced in Section 5, as a powerful technique for improving convergence speed and BER performance. Linear cancellation based on the linear tentative decision function is presented in Section 6, where the connection to classic iterations for solving linear equation systems is explained. Here it is shown that the cancellation structure in fact leads to an iterative solution to the constrained ML problem. The advantages of the clipped linear tentative decision function are detailed in Section 7, and in Section 8, structures using the hyperbolic tentative decision function are discussed. Numerical examples are included in Section 9 to illustrate the characteristics of the schemes discussed, and in Section 10, concluding remarks are made.

2. SYSTEM MODEL

Let us consider a CDMA channel that is simultaneously shared by K users. Each user is assigned a signature waveform $p_k(t)$ of duration T where either $T = T_s$, or $T \gg T_s$. Here, T_s denotes the duration of one data symbol. In the first case, the signature waveform is the same for each symbol interval, while in the latter case, the signature waveform changes for each symbol interval. The former case is termed *short codes*, while the latter is termed *long codes*. For notational simplicity, the mathematical description is based on the former case. It is conceptually easy to extend the description to the long code case. A signature waveform may thus be expressed as

$$p_k(t) = \sum_{j=0}^{N-1} a_k(j)p(t - jT_c), \quad 0 \le t \le T_s \tag{1}$$

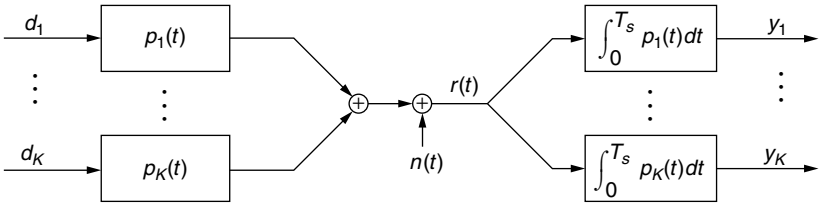


Figure 2. A simple continuous-time model for a synchronous CDMA system.

where $\{a_k(j): 0 \le j \le N-1\}$ is a spreading code sequence consisting of N chips that take on values $\{-1, +1\}$, and $p(t)$ is a chip pulse of duration T_c where T_c is the chip interval. Here, $p_k(t) = 0$ outside the symbol interval, $t < 0$, $t > T_s$ since a short-code system is considered. Thus, we have N chips per symbol and $T_s = NT_c$. Without loss of generality, we assume that all K signature waveforms have unit energy:

$$\int_0^{T_s} |p_k(t)|^2 dt = 1 \tag{2}$$

The information sequence of the kth user is denoted $\{d_k(l): 0 \le l \le L-1\}$, where the value of each information symbol may be chosen from the set $\mathcal{D}$, and L denotes the block length of the transmission. For binary transmission, we have $\mathcal{D} = \{-1, +1\}$. The corresponding equivalent lowpass transmitted waveform may be expressed as

$$s_k(t) = \sum_{l=0}^{L-1} e^{j\phi_k(l)} d_k(l) p_k(t - lT_s) \tag{3}$$

The composite signal for the K users may be expressed as follows, assuming for simplicity a single-path channel with unit magnitude:

$$s(t) = \sum_{k=1}^{K} s_k(t - \tau_k) = \sum_{k=1}^{K} \sum_{l=0}^{L-1} e^{j\phi_k(l)} d_k(l) p_k(t - lT_s - \tau_k) \tag{4}$$

where $\{\tau_k: 1 \le k \le K\}$ are the transmission delays, which satisfy the condition $0 \le \tau_k \le T_s$ for $1 \le k \le K$ and $\{\phi_k(l): 1 \le k \le K, 0 \le l \le L-1\}$ are the random-phase rotations, assumed constant over one symbol interval. This is the model for a multiuser signal in asynchronous mode, which is typical for an uplink scenario. In the special case of synchronous transmission, $\tau_k = 0$ for $1 \le k \le K$, which is a typical scenario for downlink transmission. This model can easily be extended to model transmission over multipath channels. The extension is discussion in more detail in, for example, Ref. 34.

To make the presentation in the following sections notationwise and conceptually simple, we will focus on a synchronous CDMA system without any random phase rotation, specifically, $\tau_k = 0$ and $\phi_k(l) = 0$ for $1 \le k \le K$, $0 \le l \le L-1$. A synchronous system is naturally associated with downlink transmission while multiuser detection strategies are intended primarily for uplink transmission. We still maintain a synchronous model in this presentation as it greatly simplifies notation and conception. The extensions to asynchronous systems are straightforward once the basic principles are understood.

We also assume that binary phase shift keying (BPSK) modulation formats are used: $\mathcal{D} = \{-1, +1\}$. All the presented concepts, however, generalize to more elaborate cases. For this simplified case, it is sufficient to only consider one symbol interval. The symbol interval index is therefore omitted in the following. The corresponding K user model is shown in Fig. 2, where $\phi_k = 0$ for $1 \le k \le K$.

The transmitted signal is assumed to be corrupted by additive white Gaussian noise (AWGN). Hence, the received signal may be expressed as

$$r(t) = s(t) + n(t) \tag{5}$$

where $n(t)$ is the noise with double-sided power spectral density $\sigma_n^2 = N_0/2$. The sampled output of a chip matched filter (CMF) for chip interval j, user k and an arbitrary symbol interval is [11]

$$r_j = \int_{jT_c}^{(j+1)T_c} r(t) p(t - jT_c) dt, \quad 0 \le j \le N-1 \tag{6}$$

Collecting all CMF outputs in a column vector

$$\mathbf{r} = (r_1, r_2, \ldots, r_N)^T \tag{7}$$

we have

$$\mathbf{r} = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} \mathbf{a}_k d_k + \mathbf{n} \tag{8}$$

where $\mathbf{a}_k$ is a length N vector representing the code sequence $\{a_k(j): 0 \le j \le N-1\}$ and $\mathbf{n}$ is a length N Gaussian noise vector with autocorrelation matrix

$$\mathrm{E}\{\mathbf{n}\mathbf{n}^T\} = \sigma_n^2 \mathbf{I} \tag{9}$$

Define

$$\mathbf{s}_k = \frac{1}{\sqrt{N}} \mathbf{a}_k \tag{10}$$

On the basis of Eq. (8), discrete-time code matched filtering for user k can be conveniently described as

$$\begin{aligned} y_k = \mathbf{s}_k^T \mathbf{r} &= \sum_{i=1}^{K} \mathbf{s}_k^T \mathbf{s}_i d_i + \mathbf{s}_k^T \mathbf{n} = \sum_{i=1}^{K} \mathbf{s}_k^T \mathbf{s}_i d_i + z_k \\ &= d_k + \sum_{i \ne k} \rho_{ki} d_i + z_k = d_k + w_k + z_k \end{aligned} \tag{11}$$

where $\rho_{ki} = \mathbf{s}_k^T \mathbf{s}_i$ is the cross-correlation between the spreading codes of users k and i, z_k is the Gaussian noise experienced by user k, and w_k is the MAI experienced by user k, respectively.

Considering the algebraic structure of the discrete-time received signal y_k described in (11), we can conclude the following. Interference arises since, in general, every output, y_k, has a contribution from every input, $\{d_k: 1 \le k \le K\}$. Under ideal conditions where all users are orthogonal to each other, all the cross-correlations are zero and there is no interference. In practice such a scenario is virtually impossible to achieve, and thus each output has contributions from all users. When the MAI is ignored in the detection process, the performance is strongly interference limited. However, when the structure of the MAI is considered in the detector, considerable gains are possible.

The models in (8) and (11) can conveniently be extended to include all users, applying linear algebra to provide a compact description. Equation (8) can be described as

$$\mathbf{r} = \mathbf{S}\mathbf{d} + \mathbf{n} \tag{12}$$

where $\mathbf{S}$ is a $N \times K$ matrix containing the spreading codes (10) of all users as columns

$$\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K) \tag{13}$$

and $\mathbf{d}$ is a length K vector of user symbols

$$\mathbf{d} = (d_1, d_2, \ldots, d_K)^T \tag{14}$$

Each decision statistic is described by (11). Collecting all decision statistics in a vector

$$\mathbf{y} = (y_1, y_2, \ldots, y_K)^T \tag{15}$$

we arrive at the following model

$$\mathbf{y} = \mathbf{S}^T\mathbf{S}\mathbf{d} + \mathbf{S}^T\mathbf{n} = \mathbf{R}\mathbf{d} + \mathbf{z} \tag{16}$$

where $\mathbf{R}$ is a symmetric, positive semi-definite correlation matrix of dimension $K \times K$

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ \vdots & & & \vdots \\ \rho_{K1} & \rho_{K2} & \cdots & 1 \end{bmatrix} \tag{17}$$

and $\mathbf{z}$ is a vector of length K, containing the Gaussian noise samples with autocorrelation function:

$$\mathrm{E}\{\mathbf{z}\mathbf{z}^T\} = \sigma_n^2\mathbf{R} \tag{18}$$

An equivalent discrete-time system model can be defined on the basis of these results and is depicted in Fig. 3.

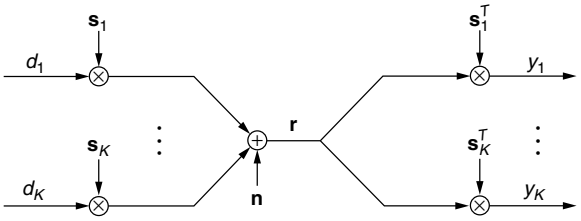


Figure 3. A simple discrete-time model for a synchronous CDMA system.

The correlation matrix can be partitioned as $\mathbf{R} = \mathbf{I} + \mathbf{M}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{M}$ is the corresponding off-diagonal matrix. We can then write (16) as follows

$$\mathbf{y} = (\mathbf{I} + \mathbf{M})\mathbf{d} + \mathbf{z} = \mathbf{d} + \mathbf{M}\mathbf{d} + \mathbf{z} \tag{19}$$

where $\mathbf{d}$ is the desired signal vector and $\mathbf{M}\mathbf{d}$ is MAI.

3. PRINCIPLE STRUCTURE

It was argued in Section 2 that the decision statistics for detection are polluted by MAI. Let us for a moment assume that, somehow, the MAI is perfectly known at the receiver. It is then possible to eliminate the interference simply by subtracting it from the received signal:

$$x_k = y_k - \sum_{i \ne k} \rho_{ki} d_i = d_k + z_k \tag{20}$$

By subtracting the known MAI, we obtain an interference-free received signal, and thus the performance is identical to the case where only one user is present, the so-called single-user (SU) case.

Unfortunately, the MAI is not perfectly known at the receiver. To have perfect knowledge of the MAI requires perfect knowledge of the transmitted symbols, in which case there would be no information contained in the transmission. Instead of perfect knowledge, we can use an estimate of the transmitted symbols. For example, let the initial estimate for user k, $u_{1,k}$ be determined by a hard decision based on the corresponding received matched-filter output, y_k, i.e., the polarity of y_k decides whether $u_{1,k} = 1$ or $u_{1,k} = -1$

$$u_{1,k} = \mathrm{Sgn}\ (y_k) \tag{21}$$

where $\mathrm{Sgn}(\cdot)$ denotes the polarity check function, which is the same as the hard decision in Fig. 1:

$$\mathrm{Sgn}\ (x) = \begin{cases} 1 & x \ge 0 \\ -1 & x < 0 \end{cases} \tag{22}$$

An updated decision statistic, $x_{2,k}$ after one step of MAI subtraction is then

$$x_{2,k} = y_k - \sum_{i \ne k} \rho_{ki} u_{1,i} = d_k + \sum_{i \ne k} \rho_{ki}(d_i - u_{1,i}) + z_k \tag{23}$$

and

$$u_{2,k} = \mathrm{Sgn}\ (x_{2,k}) \tag{24}$$

If all the tentative decisions were correct, we have successfully eliminated all MAI and we obtain single-user (SU) performance. Each wrong decision, however, doubles the particular MAI contribution rather than eliminating it. As long as we eliminate "more" MAI than we introduce, it seems to be intuitively a good approach. The resulting decision statistic, $u_{2,k}$ can now be used to generate a new, and hopefully better, estimate of the MAI:

$$x_{3,k} = y_k - \sum_{i \ne k} \rho_{ki} u_{2,i} = d_k + \sum_{i \ne k} \rho_{ki}(d_i - u_{2,i}) + z_k. \tag{25}$$

This strategy can be continued until no further improvements are obtained. In case the updates for all the users are done simultaneously, this iterative multiuser detection scheme is also known as hard-decision, multistage parallel interference cancellation. It is called *interference cancellation* (IC) for obvious reasons as we attempt to cancel MAI. It is termed "parallel" since updated decision statistics for all the users are determined in parallel, based on the same tentative estimates of the transmitted symbols. The update process is obviously recursive or iterative, a characteristic that initially was termed "multistage detection" [23]. Finally the scheme is based on hard decisions, namely, polarity check, on the resulting decision statistics.

This is, however, not necessarily the best tentative decision strategy. In Fig. 1, four alternatives are shown: linear decision, hard decision, clipped linear decision, and hyperbolic tangent decision. We will later examine these tentative decision functions in more detail and try to establish theoretical justification.

For a general tentative decision function, the above scheme can be described by

$$x_{m+1,k} = y_k - \sum_{i \neq k} \rho_{ki} u_{m,i} \tag{26}$$

$$u_{m+1,k} = f_x(x_{m+1,k}), \tag{27}$$

or in a more compact form

$$u_{m+1,k} = f_x\left(y_k - \sum_{i \neq k} \rho_{ki} u_{m,i}\right) \tag{28}$$

with $u_{0,k} = 0$.

Cancellation can also be based on the most current estimate of the MAI. In this case, the MAI estimate is updated for each new tentative decision. As a consequence, the users are processed successively, leading to SIC in contrast to the PIC described above. An iterative SIC scheme is described by

$$x_{m+1,k} = y_k - \sum_{i=1}^{k-1} \rho_{ki} u_{m+1,i} - \sum_{i=k+1}^{K} \rho_{ki} u_{m,i} \tag{29}$$

$$u_{m+1,k} = f_x(x_{m+1,k}). \tag{30}$$

To arrive at a description convenient for implementation, we first use the fact that $y_k = \mathbf{s}_k^T \mathbf{r}$ and $\rho_{ki} = \mathbf{s}_k^T \mathbf{s}_i$:

$$x_{m+1,k} = \mathbf{s}_k^T \left(\mathbf{r} - \sum_{i=1}^{k-1} \mathbf{s}_i u_{m+1,i} - \sum_{i=k+1}^{K} \mathbf{s}_i u_{m,i}\right) \tag{31}$$

Then we add and subtract the term $u_{m,k}$:

$$x_{m+1,k} = \mathbf{s}_k^T \left(\mathbf{r} - \sum_{i=1}^{k-1} \mathbf{s}_i u_{m+1,i} - \sum_{i=k}^{K} \mathbf{s}_i u_{m,i}\right) + u_{m,k} \tag{32}$$

Finally, we define the residual error vector for user k, $\mathbf{e}_{m+1,k}$ as the term within the parentheses and get

$$x_{m+1,k} = \mathbf{s}_k^T \mathbf{e}_{m+1,k} + u_{m,k} \tag{33}$$

The residual error vector can be updated recursively:

$$\mathbf{e}_{m+1,k+1} = \mathbf{r} - \sum_{i=1}^{k} \mathbf{s}_i u_{m+1,i} - \sum_{i=k+1}^{K} \mathbf{s}_i u_{m,i} \tag{34}$$

$$= \mathbf{r} - \sum_{i=1}^{k-1} \mathbf{s}_i u_{m+1,i} - \sum_{i=k}^{K} \mathbf{s}_i u_{m,i} - \mathbf{s}_k u_{m+1,k} + \mathbf{s}_k u_{m,k} \tag{35}$$

$$= \mathbf{e}_{m+1,k} - \mathbf{s}_k (u_{m+1,k} - u_{m,k}) = \mathbf{e}_{m+1,k} - \Delta \mathbf{e}_{m+1,k} \tag{36}$$

Here, $\mathbf{e}_{m+1,K+1} = \mathbf{e}_{m+2,1}$.

The PIC structure described previously can also be described in this manner. In this case

$$\mathbf{e}_{m+1,k} = \mathbf{r} - \sum_{i=1}^{K} \mathbf{s}_i u_{m,i} \tag{37}$$

and thus the residual error signal is the same for all users. We can therefore drop the user index. Rewriting (37) for the PIC case, we get

$$\mathbf{e}_{m+1} = \mathbf{r} - \sum_{i=1}^{K} \mathbf{s}_i u_{m,i} = \mathbf{r} - \sum_{i=1}^{K} \mathbf{s}_i u_{m,i} + \sum_{i=1}^{K} \mathbf{s}_i u_{m-1,i} - \sum_{i=1}^{K} \mathbf{s}_i u_{m-1,i}$$

$$= \mathbf{e}_m - \sum_{i=1}^{K} \mathbf{s}_i (u_{m+1,k} - u_{m,k}) = \mathbf{e}_m - \sum_{i=1}^{K} \Delta \mathbf{e}_{m+1,i} \tag{38}$$

where $\Delta \mathbf{e}_{m+1,k} = \mathbf{s}_k (u_{m+1,k} - u_{m,k})$ as before. Comparing the cases for SIC and PIC, we see that for user k, the required input to make an updated tentative decision is $u_{m,k}$ and $\mathbf{e}_{m+1,k}$ while the output is conveniently $u_{m+1,k}$ and $\Delta \mathbf{e}_{m+1,k}$. We can thus define a basic interference cancellation unit (ICU) as shown in Fig. 4, where $f_x(\cdot)$ is a predetermined tentative decision function, possibly selected among the four alternatives illustrated in Fig. 1.

The SIC and the PIC structures are then obtained by different interconnection strategies of ICUs. In Fig. 5, we have an SIC structure. The residual error vector is updated according to (36), as it should be. In contrast, we have a PIC structure in Fig. 6, where the same residual error vector is input to all ICUs at the same iteration and it is updated according to (37). This modular structure is quite attractive for practical implementation and thus IC structures have received most attention

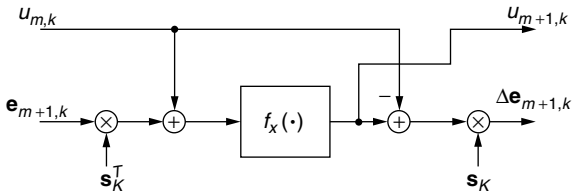


Figure 4. Basic structure of an ICU.

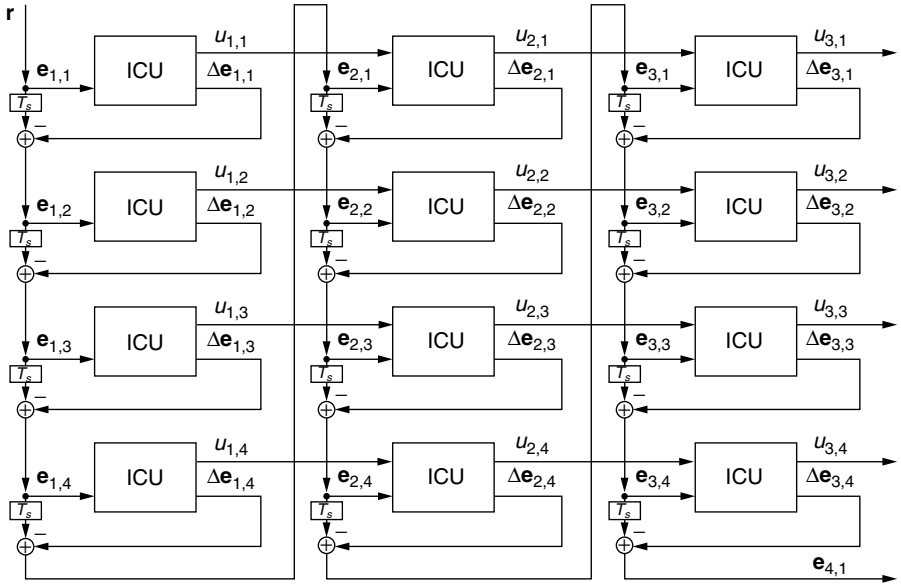


Figure 5. A modular SIC structure.

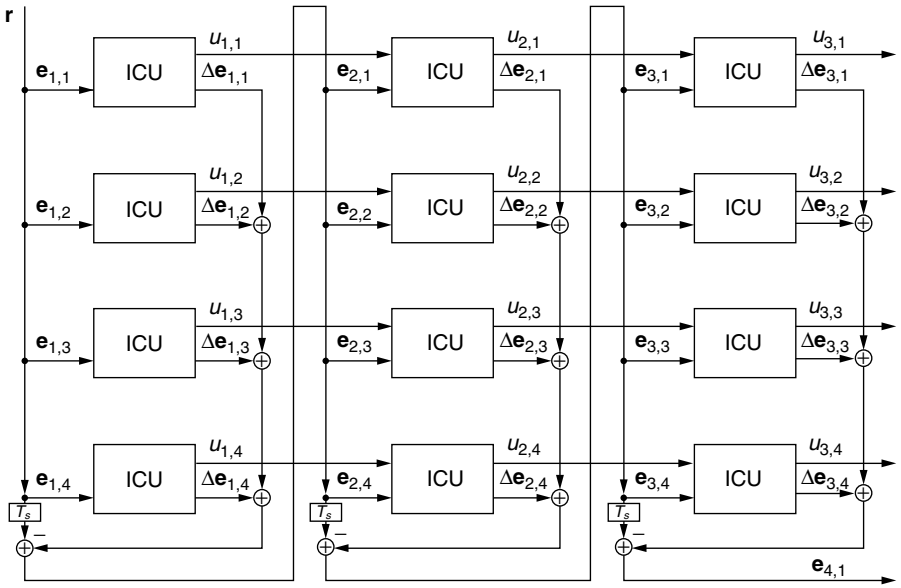


Figure 6. A modular PIC structure.

for potential commercial use. The regular structure also motivates hybrid, or group, cancellation strategies, combining successive and parallel techniques [35,36]. A groupwise cancellation structure is shown in Fig. 7.

As mentioned previously, it should be noted that once the cancellation process is completed, hard decisions are applied to the resulting decision statistics in case a final decision out of the detector is required.

As in Section 2, the decision statistics can be collected in vectors and the cancellation process can be described conveniently through matrix algebra. A PIC scheme can thus be described as

$$\mathbf{x}_{m+1} = \mathbf{y} - \mathbf{M}\mathbf{u}_m \tag{39}$$

$$\mathbf{u}_{m+1} = \mathbf{f}_x(\mathbf{x}_{m+1}) \tag{40}$$

or in a more compact form

$$\mathbf{u}_{m+1} = \mathbf{f}_x(\mathbf{y} - \mathbf{M}\mathbf{u}_m) \tag{41}$$

with $\mathbf{u}_0 = \mathbf{0}$. For a convenient algebraic description of SIC, the following partition of $\mathbf{M}$ is helpful

$$\mathbf{M} = \mathbf{L} + \mathbf{U} \tag{42}$$

where $\mathbf{L}$ is a strictly lower left triangular matrix and $\mathbf{U}$ is a strictly upper right triangular matrix, respectively. An iterative SIC scheme is then described by

$$\mathbf{x}_{m+1} = \mathbf{y} - \mathbf{L}\mathbf{u}_{m+1} - \mathbf{U}\mathbf{u}_m \tag{43}$$

$$\mathbf{u}_{m+1} = \mathbf{f}_x(\mathbf{x}_{m+1}) \tag{44}$$

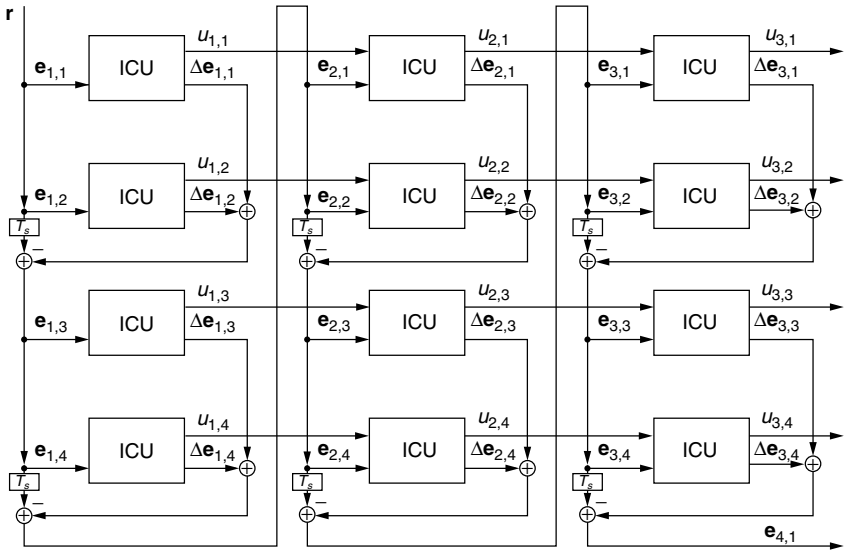


Figure 7. A modular groupwise structure.

4. HARD TENTATIVE DECISION FUNCTION

In the previous section, we defined SIC and PIC principles and established how they are related. In the continuing discussion, we focus on PIC structures. The extension to SIC schemes is usually straightforward, although more cumbersome notation is required. Known difficulties are explicitly pointed out. Also in the previous section, the principles of cancellation were presented in an intuitive manner. In the following sections, we show that certain iterative interference cancellation structures are in fact iterative realizations of known theoretically justified detector structures.

One of the first suggested IC structures was hard decision PIC [23]. This structure can be derived as an approximation to the optimal ML detector, given certain simplifying assumptions. Considering the signal model in (16), then given perfect knowledge of the correlation matrix, the decision statistics are jointly Gaussian distributed, leading to the following probability density function which is also the likelihood function:

$$p(\mathbf{y} \mid \mathbf{d}) = C \exp\left(\tfrac{1}{2}(\mathbf{y} - \mathbf{R}\mathbf{d})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{R}\mathbf{d})\right) \tag{45}$$

Here, C is a constant which is independent of the conditional data vector $\mathbf{d}$. The corresponding log-likelihood function is [37]

$$\Lambda(\mathbf{d}) = \mathbf{d}^T \mathbf{R}\mathbf{d} - 2\mathbf{y}^T \mathbf{d} \tag{46}$$

and thus the optimal ML decision is the argument $\mathbf{d} \in \{-1, 1\}^K$ that minimizes the loglikelihood function:

$$\mathbf{d}_{\mathrm{ML}} = \arg \min_{\mathbf{d} \in \{-1,1\}^K} [\mathbf{d}^T \mathbf{R}\mathbf{d} - 2\mathbf{y}^T \mathbf{d}] \tag{47}$$

Assume now that we know all the transmitted symbols, except the symbol for user k, d_k. We want a detector that maximizes the probability of d_k given the received signal and given knowledge of all other transmitted symbols. Let us first define

$$\mathbf{d}(k) = (d_1, d_2, \ldots, d_{k-1}, d_{k+1}, \ldots, d_K)^T \tag{48}$$

We now want to maximize

$$\begin{aligned} P(d_k \mid \mathbf{y}, \mathbf{d}(k)) &= P(d_k) \frac{p(\mathbf{y}, \mathbf{d}(k) \mid d_k)}{p(\mathbf{y}, \mathbf{d}(k))} \\ &= P(d_k) \frac{p(\mathbf{y} \mid \mathbf{d}(k), d_k) P(\mathbf{d}(k) \mid d_k)}{p(\mathbf{y}, \mathbf{d}(k))} \end{aligned} \tag{49}$$

with respect to d_k. Here, $p(\cdot)$ denotes a probability density function while $P(\cdot)$ denotes a probability. The equality follows from Baye's rule [11]. Maximizing (49) is equivalent to maximizing $p(\mathbf{y} \mid \mathbf{d}(k), d_k)$, a problem that can be described by the loglikelihood function,

$$\begin{aligned} \hat{d}_k &= \arg \min_{d_k \in \{-1,1\}} [\mathbf{d}^T \mathbf{R}\mathbf{d} - 2\mathbf{y}^T \mathbf{d}] \\ &= \arg \min_{d_k \in \{-1,1\}} \left[d_k^2 + 2 d_k \sum_{i \neq k} \rho_{ki} d_i - 2 y_k d_k \right] \end{aligned} \tag{50}$$

where we have thrown away all terms independent of d_k and thus do not influence the optimization problem. We can also write this as

$$\begin{aligned} \hat{d}_k &= \arg \max_{d_k \in \{-1,1\}} \left[d_k \left(y_k - \sum_{i \neq k} \rho_{ki} d_i \right) \right] \\ &= \mathrm{Sgn}\left(y_k - \sum_{i \neq k} \rho_{ki} d_i \right) \end{aligned} \tag{51}$$

which is in fact identical in form to (23) and (24), describing hard-decision PIC. Since $\mathbf{d}(k)$ is not known, we use the most recent estimate instead and arrive exactly at (23) and (24). In conclusion, given perfect knowledge of interfering symbols, interference cancellation is an

optimal structure. Using current estimates of these symbols of course leads to an approximating, suboptimal approach. Depending on the strategy of cancellation, PIC, SIC or hybrids of the two are obtained.

5. WEIGHTED CANCELLATION

Hard decision cancellation is prone to error propagation and can exhibit a significant error floor for high SNR, as it is in effect interference-limited. This is due to the coarse approximation that previous symbol estimates provide for the MAI

$$x_{m+1,k} = d_k + \sum_{i \neq k} \mathbf{s}_k^T \mathbf{s}_i (d_i - u_{m,i}) + \mathbf{s}_k^T \mathbf{n} = d_k + \hat{\mathbf{z}}_k \quad (52)$$

$$u_{m+1,k} = \mathrm{Sgn}\,(x_{m+1,k}) \quad (53)$$

where

$$\hat{z}_k = \sum_{i \neq k} \mathbf{s}_k^T \mathbf{s}_i (d_i - u_{m,i}) + \mathbf{s}_k^T \mathbf{n} \quad (54)$$

Inherent assumptions are that the cancellation error, $\sum_{i \neq k} \mathbf{s}_k^T \mathbf{s}_i (d_i - u_{m,i})$ is Gaussian and independent of the thermal noise $\mathbf{n}$. Neither of these assumptions is true. Obviously the second assumption cannot be true since each tentative decision depends on the thermal noise. This was taken into account in the improved cancellation scheme suggested by Divsalar et al. [38]. Here, it is assumed that the cancellation error and the thermal noise are correlated. Also, at iteration $(m+1)$, the detector is derived based on observing the received signal y_k as well as the previous decision statistic $x_{m,k}$. The joint likelihood function is still derived as conditioned on perfect knowledge of $\mathbf{d}(k)$, however, it now depends on the correlation between the Gaussian noise and the residual interference.

Following some manipulations, some simplifying assumptions and substituting the most current estimate $\mathbf{u}_m(k)$ in place of $\mathbf{d}(k)$, we arrive at the following revised decision statistic:

$$x_{m+1,k} = \mu_{m+1,k} \left(y_k - \sum_{i \neq k} \rho_{ki} u_{m,i} \right) + (1 - \mu_{m+1,k}) x_{m,k} \quad (55)$$

The updated decision statistic is now determined as a weighted sum of the previous decision statistic and the corresponding decision statistic determined by a traditional PIC cancellation. The weighting factor, $\mu_{m+1,k}$, is described by an involved combination of the correlation parameters between the Gaussian noise and the residual error term [38]. It may not be possible to accurately determine the weighting factor analytically, but trial-and-error selection has shown that the general structure is very powerful and provides significant performance gains over traditional hard-decision structures [38]. This technique was originally termed *partial cancellation*, and the principles are now used in most practical studies of IC techniques [39–42].

6. LINEAR TENTATIVE DECISION FUNCTION

Let us now focus on the linear tentative decision function. In this case, the corresponding iterative detectors are also linear. We therefore start by considering optimal linear detectors. Allowing the symbol estimate vector to be any real-valued vector, $\mathbf{u} \in \mathbb{R}^K$, the corresponding ML solution is easily found to be [18]

$$\mathbf{u} = \mathbf{R}^{-1} \mathbf{y} \quad (56)$$

Similarly, on the basis of the linear minimum mean-squared error criterion, the solution is [19]

$$\mathbf{u} = (\mathbf{R} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \quad (57)$$

Before data are delivered, the real-valued estimate must, of course, be mapped to a valid data symbol. Both optimal linear detectors rely on matrix inversion, which has a complexity of the order of $\mathcal{O}(K^3)$. The matrix inverse represents the solution to a set of linear equations [43]. There exist, however, efficient iterative techniques for solving a set of linear equations. As an example, let us focus on the implementation of the decorrelator, Eq. (56). The set of linear equations to be solved is described by

$$\mathbf{y} = \mathbf{R}\mathbf{u} = (\mathbf{I} + \mathbf{M})\mathbf{u} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{u} \quad (58)$$

where we have applied the partition of $\mathbf{R}$ described previously. A common iteration used for matrix inversion is the Jacobi iteration

$$\mathbf{u}_{m+1} = \mathbf{y} - \mathbf{M}\mathbf{u}_m \quad (59)$$

which is identical to (39) given a linear tentative decision as $\mathbf{u}_m = \mathbf{f}_x(\mathbf{x}_m) = \mathbf{x}_m$. Here, $\mathbf{f}_x(\cdot)$ denotes a vector function applying the decision function, $f_x(\cdot)$ to each element of the argument vector independently. Similarly, the well-known Gauss–Seidel (GS) iteration is described as

$$\mathbf{u}_{m+1} = \mathbf{y} - \mathbf{L}\mathbf{u}_{m+1} - \mathbf{U}\mathbf{u}_m \quad (60)$$

which, in turn, is equivalent to (43). It follows that these two classic iterations are linear PIC and linear SIC, respectively. This was first realized by Elders-Boll et al. [27]. For the linear case, IC therefore represents an iterative implementation of optimal linear detectors, given that the particular iteration converges. The GS iteration is guaranteed to converge while the Jacobi iteration is not. The linear PIC converges if the iteration matrix $\mathbf{M}^m = (\mathbf{R} - \mathbf{I})^m$ converges for increasing m. For $\mathbf{M}^m$ to converge, the maximum eigenvalue of $\mathbf{R}$ must be constrained by [44]

$$\lambda_{\max} \leq 2 \quad (61)$$

which is not true for all possible correlation matrices. To guarantee convergence for a linear PIC scheme, and in general also to increase convergence speed for linear IC, more advanced iterations can be used [44–46]. The concept of over-relaxation can be used for both PIC and

SIC structures. The Jacobi over-relaxation iteration is described by

$$\mathbf{u}_{m+1} = \mu(\mathbf{y} - \mathbf{R}\mathbf{u}_m) + \mathbf{u}_m \tag{62}$$

The corresponding iteration matrix is now $(\mu\mathbf{R} - \mathbf{I})^m$, and thus the relaxation parameter directly scales the eigenvalues of $\mathbf{R}$, allowing for tuned, guaranteed convergence. Considering the decision statistic for user k, we get

$$\begin{aligned} u_{m+1,k} &= \mu\left(y_k - \sum_{i=1}^{K} \rho_{ki} u_{m,i}\right) + u_{m,k} \\ &= \mu\left(y_k - \sum_{i \neq k} \rho_{ki} u_{m,i}\right) + (1-\mu)u_{m,k} \end{aligned} \tag{63}$$

which has the structure of (55), emphasizing that these advanced iterations correspond to weighted linear IC. The optimal weighting factor for the Jacobi overrelaxation is determined by the eigenvalue spread of the correlation matrix $\mathbf{R}$ [43]. Fastest convergence is assured when the positive and the negative mode of convergence for the iteration matrix are equal, which is obtained by

$$\mu = \frac{2}{\lambda_{\max} + \lambda_{\min}} \tag{64}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of $\mathbf{R}$, respectively. An asymptotic analysis for large systems [44] has shown that this optimal weighting factor is well approximated by

$$\mu = \frac{N}{N+K} \tag{65}$$

when N and K are large. The ICU of Fig. 4 should be modified as shown in Fig. 8 to accommodate the weighted cancellation of (62).

First-order and second-order iterations have also been suggested, such as the steepest-descent iteration and the conjugant gradient iteration [45–47]. For the steepest-descent iteration, expressions for optimal weighting factors for both short and long codes have been derived Guo et al. [29,46]. For a more thorough analysis and discussion of linear IC schemes, please consult Refs. 44–46.

7. CLIPPED LINEAR TENTATIVE DECISION FUNCTION

A hard decision is effective when the decision statistic is large, in which case the corresponding decision should be relatively reliable. This is in contrast to a hard decision based on a decision statistic only marginally different from zero, which is bound to be relatively unreliable, potentially introducing additional interference with high probability. A linear tentative decision function is effective when the decision statistic has a magnitude less than one. For very small decision statistics, only a very small interference estimate is subtracted, limiting the potential damage of a wrong decision. As the decision statistic increases, increasingly larger interference estimates are subtracted. For decision statistics with magnitude larger than one, a linear decision will therefore invariably introduce additional interference. The transmitted symbol is limited to unit magnitude and thus we should not attempt to cancel out more than that.

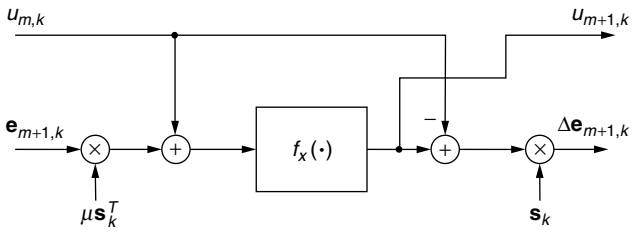


Figure 8. Basic structure of a weighted ICU.

The clipped linear decision function depicted in Fig. 1 seems to be an appropriate choice for combining the benefits of a hard decision and a linear decision, respectively, avoiding the inherent drawbacks. Therefore, assume now that we constrain the allowable solution for each user to $-1 \leq u_{m,k} \leq 1$ for all k and m. Enforcing these constraints for all users simultaneously describe a K-dimensional hypercube in Euclidean space. Such a constraint is denoted a box-constraint and is formally defined as

$$\mathbb{B}^K = \{\mathbf{d} \in \mathbb{R}^K : \mathbf{d} \in [-\mathbf{b}, \mathbf{b}]\} \tag{66}$$

where $\mathbf{b}$ is an all ones vector. The corresponding box-constrained optimization problem is described as

$$\mathbf{u} = \arg\min_{\mathbf{d} \in \mathbb{B}^K} [\mathbf{d}^T\mathbf{R}\mathbf{d} - 2\mathbf{y}^T\mathbf{d}] \tag{67}$$

Ahn [48] first suggested an iterative algorithm for solving the general problem of constraining the solution to any convex set. A hypercube, a so-called box, is a tight convex set. For any convex set we can define an orthogonal projection. For the box constraint, the orthogonal projection is merely the clipped linear decision function applied independently to all the users as demonstrated in Fig. 9. As shown in Refs. 30,49, and 50, the algorithm suggested by Ahn is in fact a generalization of first order iterations for solving linear equation systems:

$$\mathbf{x}_{m+1} = \mu(\mathbf{y} - \mathbf{Q}\mathbf{u}_{m+1} - (\mathbf{R} - \mathbf{Q})\mathbf{u}_m) + \mathbf{u}_m \tag{68}$$

$$\mathbf{u}_{m+1} = \mathbf{f}_x(\mathbf{x}_{m+1}) \tag{69}$$

Letting $\mathbf{Q} = \mathbf{0}$, we have a weighted PIC structure, while $\mathbf{Q} = \mathbf{L}$ leads to a weighted SIC structure. It has been shown that the same convergence conditions as for the linear case apply [48]. It follows that a traditional SIC scheme based on a clipped linear decision always converges to the solution of the box-constrained ML problem. The PIC also converges to this solution when an appropriate weighting factor is chosen. Unfortunately, no analytic results have yet been obtained for deriving optimal weighting factors. In general an SIC structure converges faster than a PIC structure. The clipped linear decision is quite attractive as it provides good performance, relatively fast convergence, and it is simple to implement in hardware. The corresponding ICU has the same structure

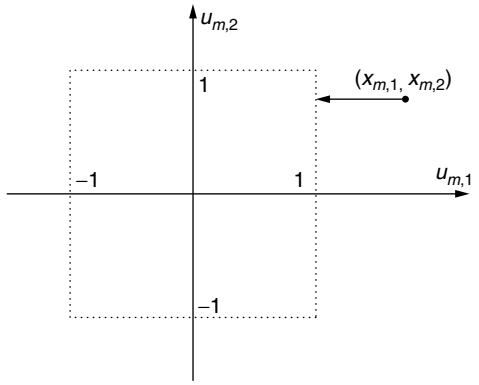


Figure 9. The orthogonal projection onto a hypercube.

as shown in Fig. 8. This approach has been suggested for practical implementation in several papers [e.g., 41].

8. HYPERBOLIC TANGENT TENTATIVE DECISION FUNCTION

The final tentative decision function we consider is based on an MMSE optimized estimate.

$$u_{m,k} = \mathrm{E}\{d_k \mid x_{m,k}\} \tag{70}$$

This was first suggested by Tarköy [51] and later developed further by other authors [31,32]. Considering the general description of PIC, we can write the decision statistic for user k and iteration m as done in (23)

$$x_{m+1,k} = d_k - \sum_{i \neq k} \rho_{ki}(d_i - u_{m,i}) + z_k = d_k - \sum_{i \neq k} \rho_{ki}\varepsilon_{m,i} + z_k \tag{71}$$

Assuming that the cancellation error is a zero-mean Gaussian random variable, independent of the thermal noise, $x_{m+1,k}$ is also a Gaussian random variable with mean d_k and variance $\sigma^2_{m,k}$:

$$x_{m+1,k} \sim N(d_k, \sigma^2_{m,k}) \tag{72}$$

$$\sigma^2_{m,k} = \sigma^2_{\varepsilon,m,k} + \sigma^2_n \tag{73}$$

The expectation in (70) is obviously determined by

$$u_{m+1,k} = P(d_k = 1 \mid x_{m+1,k}) - P(d_k = -1 \mid x_{m+1,k}) \tag{74}$$

Considering one term at a time

$$P(d_k = 1 \mid x_{m+1,k}) = \frac{P(d_k)p(x_{m+1,k} \mid d_k = 1)}{p(x_{m+1,k})} \tag{75}$$

Using the fact that

$$P(d_k = 1 \mid x_{m+1,k}) + P(d_k = -1 \mid x_{m+1,k}) = 1 \tag{76}$$

we arrive at

$$P(d_k = 1 \mid x_{m+1,k}) = \frac{p(x_{m+1,k} \mid d_k = 1)}{p(x_{m+1,k} \mid d_k = 1) + p(x_{m+1,k} \mid d_k = -1)} \tag{77}$$

From (74), we then have

$$u_{m+1,k} = \frac{p(x_{m+1,k} \mid d_k = 1) - p(x_{m+1,k} \mid d_k = 1)}{p(x_{m+1,k} \mid d_k = 1) + p(x_{m+1,k} \mid d_k = -1)} \tag{78}$$

Since $x_{m+1,k}$ is assumed Gaussian with known statistics, we obtain

$$u_{m+1,k} = \tanh\left(\frac{x_{m+1,k}}{\sigma^2_{m,k}}\right) \tag{79}$$

which is a hyperbolic tangent function as shown in Fig. 1. The variance can be determined as devised by Müller and Huber [31]:

$$\sigma^2_{m,k} = \sum_{i \neq k} \rho^2_{ki}(1 - u^2_{m,k}) + \sigma^2_n \tag{80}$$

In this case, the corresponding ICU has the form shown in Fig. 4.

9. NUMERICAL EXAMPLES

In this section numerical examples are presented, illustrating the characteristics of the different cancellation strategies and the different tentative decision functions, respectively. The impact of weighted cancellation is demonstrated, although no attempts have been made for optimizing the weighting factors. Only general trends are illustrated, leaving the interested reader to consult the vast literature on the topic for more details regarding weight optimization.

A symbol synchronous CDMA system with processing gain $N = 32$ and the simple channel model of Eq. (16) is considered. Long codes are assumed, so a new random spreading sequence is used for each user and each symbol interval. In Fig. 10, the BER performance of PIC and weighted PIC (WPIC) is shown as a function of the number of users in the system. Here, the four tentative decision functions depicted in Fig. 1 are used, respectively. For weighted cancellation, a factor of $\mu = 0.5$ have been used for all cases. Better performance can be obtained for more carefully selected weighting factors. The PIC can be considered as the case of $\mu = 1$. With caution it is therefore possible to roughly predict performance for factors $0.5 \leq \mu \leq 1$ as the optimal weights in most cases are within this interval. The performance is captured at a bit energy to noise ratio $E_b/N_0 = 5$ dB. The iterative detectors have been restricted to five iterations which is considered reasonable for potential practical applications. In the following paragraphs we will denote the use of the tentative functions in Fig. 1 as LIN, HARD, CLIP, and TANH, respectively.

Considering first PIC, significant performance losses are observed as the system load, K/N, increases. Especially a linear tentative decision function is sensitive to the load. As K increases, it becomes more likely that the iteration matrix is diverging, leading to detector collapse with very poor performance. For the other decision functions, the performance degradation is more graceful, but still severe as the load increases. A load of 10–15% for the linear case and of 25–30% for the others can be

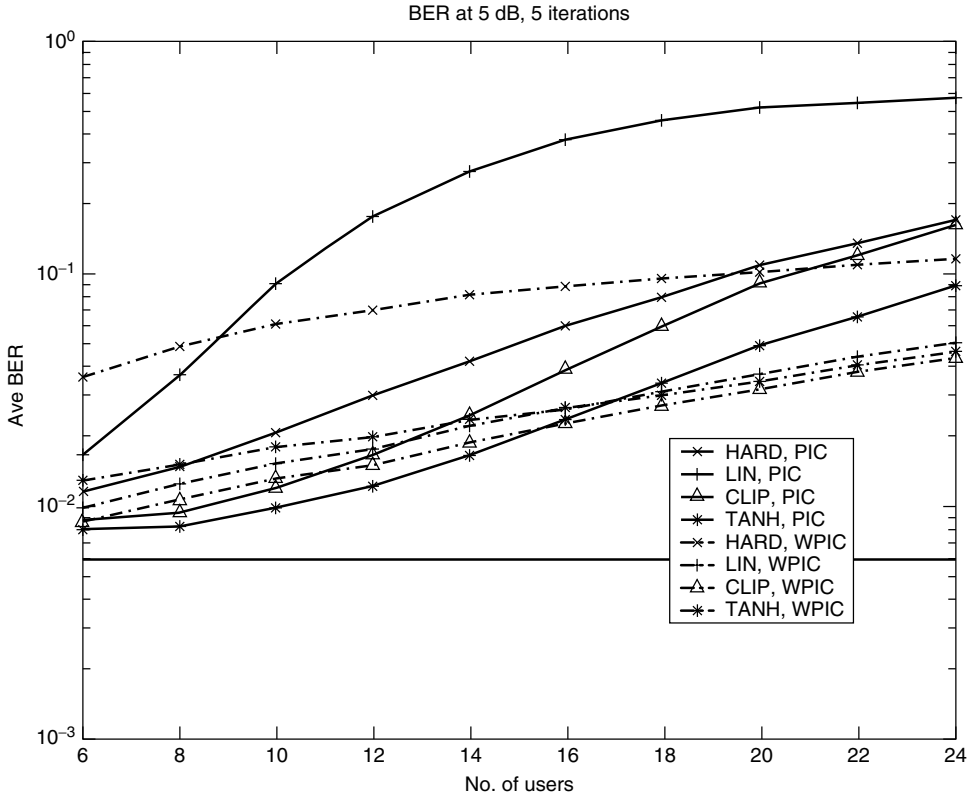


Figure 10. The BER performance for PIC and WPIC as a function of the number of users. HARD, LIN, CLIP, and TANH are used together with the cases of $\mu = 1$ and $\mu = 0.5$. The solid horizontal line represents single-user performance at 5 dB.

accommodated with reasonable losses at a processing gain of 32. The TANH shows the best performance, followed by the CLIP and HARD. For high load, CLIP and HARD provide similar performance.

Introducing a fixed weighting factor of $\mu = 0.5$ does not provide better performance for small loads. The performance degradation for WPIC as the load increases is noticeably more graceful, extending potential load with appropriately selected weights to about 50% for TANH, CLIP, and LIN. The performance of WPIC for LIN, CLIP, and TANH are quite similar to each other. This is mainly due to the limited number of iterations allowed. For a larger number of iterations, differentiating performance is obtained as illustrated in Fig. 11.

For HARD, a low weighting factor at small to moderate loads is not appropriate. For $\mu = 0.5$, the performance is considerably worse than $\mu = 1$ up to a load of 50%. This illustrates the difficulty of selecting weighting factors for HARD. Since a hard decision leads to cancellation of a "full" MAI contribution scaled by μ, regardless of the quality of the decision statistic, weighting may do more harm than good. In the first iteration decisions are based on the matched-filter output, which usually provides a BER of less than 0.5. With a weighting factor of $\mu = 0.5$, the additional MAI introduced due to wrong decisions are reduced, but at the same time only half of the MAI contributions corresponding to correct decisions are eliminated. The distinct nonlinearity of hard decisions makes the selection of weighting factors more complicated and is mainly left to a trial-and-error approach with little analytic justification. To avoid these drawbacks, the decision function should differentiate on decision statistic quality as done by the other three alternatives.

Comparing PIC ($\mu = 1$) and WPIC ($\mu = 0.5$), we can conclude that the weighting factor should decrease with load, starting at $\mu = 1$ for small loads. For appropriately varying weights, reasonable performance is to be expected at least up to a load of 50%.

It should be kept in mind that only 5 iterations are allowed in Fig. 10. In Fig. 11, the BER as a function of the number of iterations is shown for the case of $K = 24$ and $E_b/N_0 = 7$ dB. In this case, $\mu = 1$ does not provide reasonable performance for any decision function. In all cases, pingpong effects are observed where the BER oscillates with iterations [52]. For $\mu = 0.5$, HARD is still not useful as previously discussed, while LIN and TANH improve gradually up to 8 iterations, after which the performance converges to a level determined by the residual MAI. The CLIP continues to improve even beyond 15 iterations, representing the best alternative. The CLIP will, however, also converge to a level above the SU performance since this strategy provides a box-constrained ML solution and not necessarily a SU solution. At 7 dB, the SU performance is just below 10^{-3}. The benefits of a larger number of iterations at higher loads are nicely illustrated.

Successive cancellation is expected to provide better performance than PIC. This is illustrated in Fig. 12,

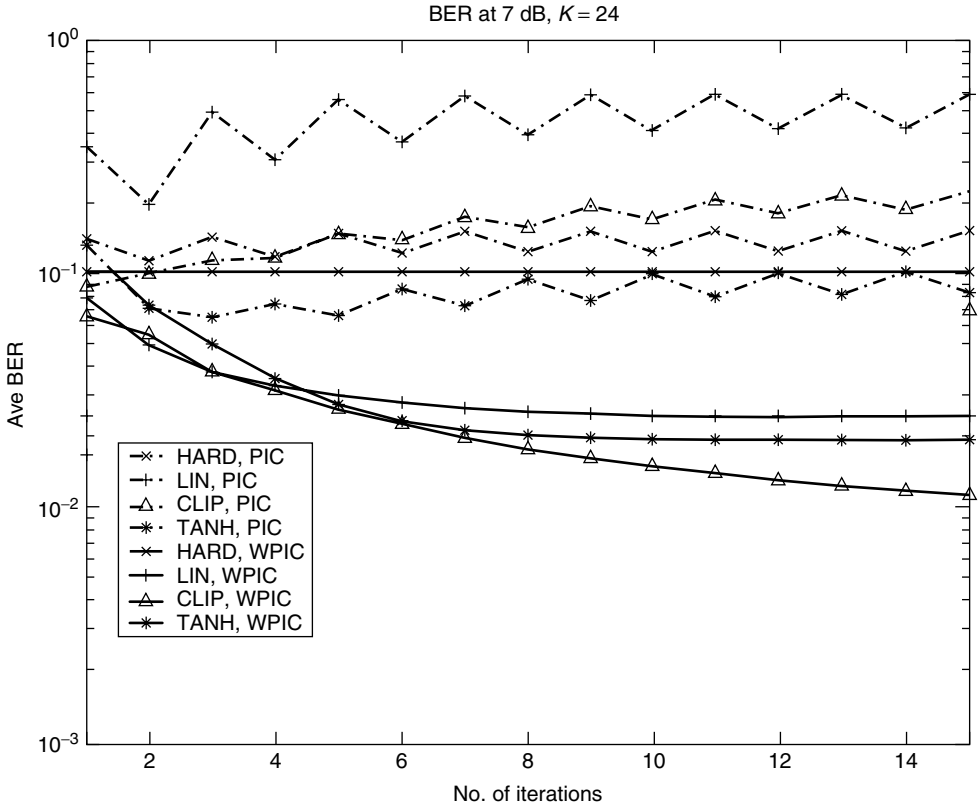


Figure 11. The BER performance for PIC and WPIC as a function of the number of iterations. HARD, LIN, CLIP, and TANH are used together with the cases of $\mu = 1$ and $\mu = 0.5$.

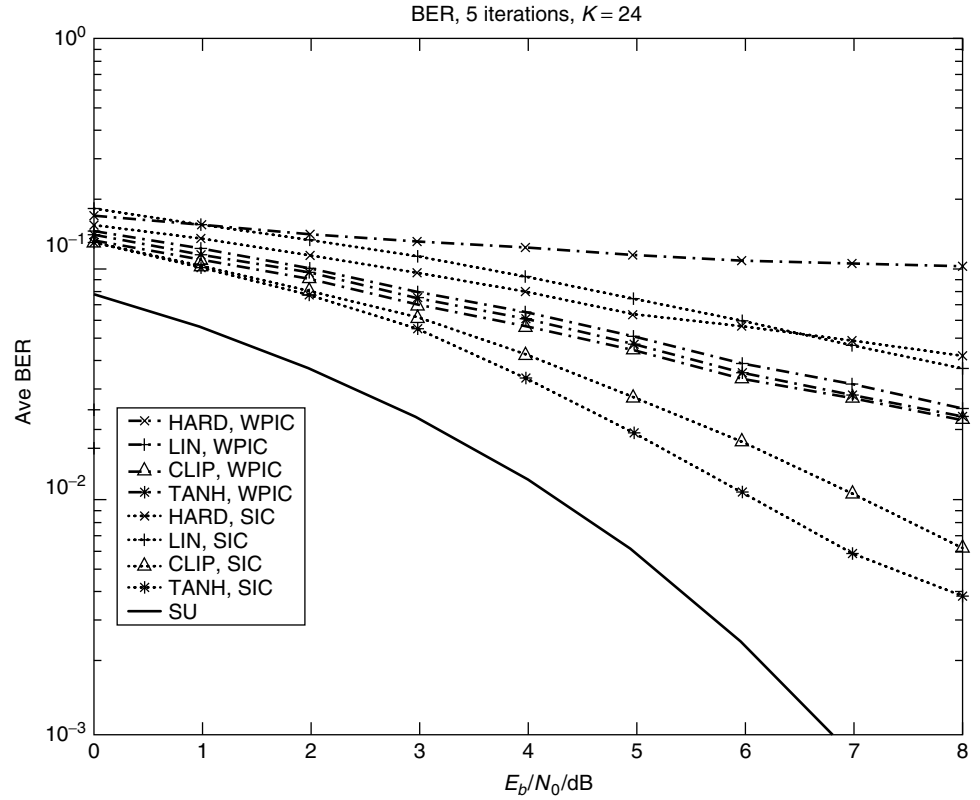


Figure 12. The BER performance for SIC and WPIC as a function of E_b/N_0, where HARD, LIN, CLIP, and TANH tentative decisions are used. For WPIC, $\mu = 0.5$.

where the BER is shown as a function of E_b/N_0. Here the performance of the SIC with the four tentative decision functions is contrasted with the performance of WPIC at a load of $K = 24$. Again, five iterations are allowed. The benefits of SIC are clear, especially for TANH and CLIP, where significant improvements in terms of E_b/N_0 are obtained.

For more detailed numerical examples and discussions on cancellation structures and tentative decision functions, the reader is referred to the open literature on the subject.

10. CONCLUDING REMARKS

In this article, we have presented the fundamental principles of iterative multiuser detection in CDMA, also known as interference cancellation strategies. The basic building block in a cancellation structure is an interference cancellation unit, taking as input the residual error signal and the tentative symbol decision from the previous iteration, giving as output an updated tentative symbol decision and an updated residual error signal. Different strategies, implementing serial and parallel cancellation structures or combinations thereof, can be constructed through different interconnections of ICUs. The ICU itself is characterized mainly by a tentative decision function. Here, we have presented four such functions, namely the linear, hard, clipped linear, and hyperbolic tangent tentative decision functions. The principles behind weighted cancellation are also presented and explained. Selected numerical examples are presented to illustrate the characteristics of each cancellation strategy and each decision function.

Acknowledgments
The author gratefully acknowledges Mr. Peng Hui Tan for providing the numerical examples and Mr. Fredrik Brännström for providing input to the presentation.

BIOGRAPHY

Lars K. Rasmussen was born on March 8, 1965 in Copenhagen, Denmark. He got his M.Eng. in 1989 from the Technical University of Denmark, and his Ph.D. degree from Georgia Institute of Technology (Atlanta, Georgia, USA) in 1993, both in electrical engineering.

From 1993 tp 1995, he was at the Mobile Communication Research Centre, University of South Australia as a Research Fellow. From 1995 to 1998 he was with the Centre for Wireless Communications at the National University of Singapore as a Senior Member of Technical Staff. He then spent 3 months at the University of Pretoria, South Africa as a Visiting Research Fellow, followed by three years at Chalmers University of Technology in Gothenburg, Sweden as an Associate Professor. He is now a professor of telecommunications at the Institute for Telecommunications Research, University of South Australia. He also maintains a part-time appointment at Chalmers.

BIBLIOGRAPHY

1. H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, rev. ed., Wiley, New York, 2001.
2. A. Viterbi, *CDMA, Principles of Spread Spectrum Communication*, Addison-Wesley, Reading, MA, 1995.
3. E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* **36**: 48–54 (Sept. 1998).
4. A. J. Viterbi, A. M. Viterbi, and E. Zehavi, Other-cell interference in cellular power-controlled CDMA, *IEEE Trans. Commun.* **42**: 1501–1504 (Feb.–April 1994).
5. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, 1998.
6. V. K. Garg, K. Smolik, and J. E. Wilkes, *Applications of CDMA in Wireless/Personal Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1997.
7. F. Adachi and M. Sawahashi, Wideband wireless access based on DS-CDMA, *IEICE Trans. Commun.* **E81-B**: 1305–1316 (July 1998).
8. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
9. C. Heegard, S. B. Wicker, and C. Heegaard, *Turbo Coding*, Kluwer, 1999.
10. P. van Rooyen, M. Lötter, and D. van Wyk, *Space-Time Processing for CDMA Mobile Communications*, Kluwer, 2000.
11. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw–Hill, 1995.
12. K. Schneider, Optimum detection of code division multiplexed signals, *IEEE Trans. Aerospace Electron. Syst.* **15**: 181–185 (Jan. 1979).
13. R. Kohno, H. Imai, and M. Hatori, Cancellation techniques of co-channel interference and application of Viterbi algorithm in asynchronous spread spectrum muliple access systems, *Prod. Symp. Inform. Theory Appl.* 659–666 (Oct. 1982).
14. S. Verdú, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **32**: 85–96 (Jan. 1986).
15. A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **13**: 260–269 (1967).
16. S. Verdú, Computational complexity of optimum multiuser detection, *Algorithmica* **4**: 303–312 (1989).
17. S. Moshavi, Multi-user detection for DS-CDMA communications, *IEEE Pers. Commun.* **34**: 132–136 (Oct. 1996).
18. R. Lupas and S. Verdú, Linear multiuser detectors for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **35**: 123–136 (Jan. 1989).
19. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detector for coherent multiuser communications, *IEEE J. Select. Areas Commun.* **8**: 683–690 (May 1990).
20. U. Madhow and M. L. Honig, MMSE interference suppression for direct-sequence spread spectrum CDMA, *IEEE Trans. Commun.* **42**: 3178–3188 (Dec. 1994).
21. T. J. Lim and Y. Ma, The Kalman filter as the optimal linear minimum mean-squared error multiuser CDMA detector, *IEEE Trans. Inform. Theory* **46**: 2561–2566 (Nov. 2000).

22. M. Honig, U. Madhow, and S. Verdú, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **41**: 944–960 (July 1995).

23. M. K. Varanasi and B. Aazhang, Multistage detection in asynchronous code–division multiple–access communications, *IEEE Trans. Commun.* **38**: 509–519 (April 1990).

24. K. Jamal and E. Dahlman, Multi-stage serial interference cancellation for DS-CDMA, *Proc. IEEE VTC '96*, Atlanta, April 1996, pp. 671–675.

25. P. Dent, B. Gudmundson, and M. Ewerbring, CDMA-IC: A novel code divisions multiple access scheme based on interference cancellation, *Proc. 3rd IEEE Int. Symp. PIMRC '92*, Boston, Oct. 1992, pp. 98–102.

26. P. Patel and J. Holtzman, Analysis of simple successive interference cancellation scheme in a DS/CDMA, *IEEE J. Select. Areas Commun.* **12**: 796–807 (June 1994).

27. H. Elders-Boll, H. D. Schotten, and A. Busboom, Efficient implementation of linear multiuser detectors for asynchronous CDMA systems by linear interference cancellation, *Eur. Trans. Telecommun.* **9**(4): 427–437 (Sept.–Nov. 1998).

28. L. K. Rasmussen, T. J. Lim, and A.-L. Johansson, A matrix-algebraic approach to successive interference cancellation in CDMA, *IEEE Trans. Commun.* **48**: 145–151 (Jan. 2000).

29. D. Guo, L. K. Rasmussen, S. Sun, and T. J. Lim, A matrix-algebraic approach to linear parallel interference cancellation in CDMA, *IEEE Trans. Commun.* **48**: 152–161 (Jan. 2000).

30. P. H. Tan, L. K. Rasmussen, and T. J. Lim, Constrained maximum-likelihood detection in CDMA, *IEEE Trans. Commun.* **49**: 142–153 (Jan. 2001).

31. R. R. Müller and J. B. Huber, Iterative soft-decision interference cancellation for CDMA, in Louise and Pupolin, eds., *Digital Wireless Communications*, Springer Verlag, 1998, pp. 110–115.

32. S. Gollamudi, S. Nagaraj, Y. -F. Huang, and R. M. Buehrer, Optimal multistage interference cancellation for CDMA systems using nonlinear MMSE criterion, *Proc. Asilomar Conf. Signals, Systems, Computers 98*, Oct. 1998, Vol. 5, pp. 665–669.

33. P. D. Alexander, A. J. Grant, and M. C. Reed, Iterative detection in code-division multiple-access with error control coding, *Eur. Trans. Telecommun.* **9**: (July–Aug. 1998).

34. L. K. Rasmussen, P. D. Alexander, and T. J. Lim, A linear model for CDMA signals received with multiple antennas over multipath fading channels, in F. Swarts, P. van Rooyen, I. Oppermann, and M. Lötter, eds., *CDMA Techniques for 3rd Generation Mobile Systems*, Kluwer, Sept. 1998, Chap. 2.

35. S. Sumei, L. K. Rasmussen, T. J. Lim, and H. Sugimoto, A hybrid interference canceller in CDMA, *Proc. IEEE Int. Symp. Spread Spectrum Techniques and Applications*, Sun City, South Africa, Sept. 1998, pp. 150–154.

36. S. Sumei, L. K. Rasmussen, T. J. Lim, and H. Sugimoto, A matrix-algebraic approach to linear hybrid interference canceller in CDMA, *Proc. Int. Conf. Univ. Personal Communication*, Florence, Italy, Oct. 1998, pp. 1319–1323.

37. L. K. Rasmussen, T. J. Lim, and T. M. Aulin, Breadth-first maximum-likelihood detection in multiuser CDMA, *IEEE Trans. Commun.* **45**: 1176–1178 (Oct. 1997).

38. D. Divsalar, M. Simon, and D. Raphaeli, Improved parallel interference cancellation for CDMA, *IEEE Trans. Commun.* **46**: 258–268 (Feb. 1998).

39. M. Sawahashi, Y. Miki, H. Andoh, and K. Higuchi, *Serial Canceler Using Channel Estimation by Pilot Symbols for DS-CDMA*, IEICE Technical Report RCS95-50, July 1995, Vol. 12, pp. 43–48.

40. T. Ojaperä et al., Design of a 3rd generation multirate CDMA system with multiuser detection, MUD-CDMA. *Proc. IEEE Int. Symp. Spread Spectrum Techniques and Applications* (*ISSSTA*), Mainz, Germany, Sept. 1996, pp. 334–338.

41. H. Seki, T. Toda, and Y. Tanaka, Low delay multistage parallel interference canceller for asynchronous DS/CDMA systems and its performance with closed-loop TPC, *Proc. 3rd Asia-Pacific Conf. Communications*, Sydney, Australia, Dec. 1997, pp. 832–836.

42. M. Sawahashi, H. Andoh, and K. Higuchi, Interference rejection weight control for pilot symbol-assisted coherent multistage interference canceller using recursive channel estimation in DS-CDMA mobile radio, *IEICE Trans. Fund.* **E81-A**: 957–970 (May 1998).

43. O. Axelsson, *Iterative Solution Methods*, Cambridge Univ. Press, 1994.

44. A. Grant and C. Schlegel, Convergence of linear interference cancellation multiuser receivers, *IEEE Trans. Commun.* **49**: 1824–1834 (Oct. 2001).

45. R. M. Buehrer, S. P. Nicoloso, and S. Gollamudi, Linear versus nonlinear interference cancellation, *J. Commun. Networks* **1**: 118–133 (June 1999).

46. D. Guo, L. K. Rasmussen, and T. J. Lim, Linear parallel interference cancellation in random-code CDMA, *IEEE J. Select. Areas Commun.* **17**: 2074–2081 (Dec. 1999).

47. P. H. Tan and L. K. Rasmussen, Linear interference cancellation in CDMA based on iterative solution techniques for linear equation systems, *IEEE Trans. Commun.* **48**: 2099–2108 (Dec. 2000).

48. B. H. Ahn, Iterative methods for linear complementary problems with upper bounds on primary variables, *Math. Prog.* **26**(3): 295–315 (1983).

49. A. Yener, R. D. Yates, and S. Ulukus, A nonlinear programming approach to CDMA multiuser detection, *Proc. Asilomar Conf. Signals, Systems, Computers 99*, Pacific Grove, CA, Oct. 1999, pp. 1579–1583.

50. P. H. Tan, L. K. Rasmussen, and T. J. Lim, Iterative interference cancellation as maximum-likelihood detection in CDMA, in *Proc. Int. Conf. Information, Communication, Signal Processing 99*, Singapore, Dec. 1999.

51. F. Tarköy, MMSE-optimal feedback and its applications, *Proc. IEEE Int. Symp. Information Theory*, Whistler, Canada, Sept. 1995, p. 334.

52. L. K. Rasmussen and I. J. Oppermann, Ping-pong effects in linear parallel interference cancellation for CDMA, to *IEEE Trans. Wireless Commun.* (in press).

JPEG2000 IMAGE CODING STANDARD

B. E. USEVITCH
University of Texas at El Paso
El Paso, Texas

1. INTRODUCTION

JPEG2000 is a new international standard for the coding (or compression) of still images. The standard was developed in order to address some of the shortcomings of the original JPEG standard, and to implement improved compression methods discovered since the original JPEG standard first appeared. The JPEG2000 standard offers a number of new features, with one of the most significant being the flexibility of the compressed bit stream. As a result of this flexibility, many image processing operations such as rotation, cropping, random spatial access, panning, and zooming, can be performed in JPEG2000 either directly on the compressed data, or by only decompressing a relevant subset of the compressed data. The flexible bit stream allows the compressed data to be reordered such that decompression will result in images of progressively larger size, higher quality, or more colors. All the bit streams, original and reordered, maintain a strict embedded property in which the most important bits come first in the bit stream. Consequently, truncating these embedded bit streams at any point gives an optimal compressed representation for the given bitlength. Other desirable features of JPEG2000 include

- Lossy and lossless compression using the same algorithm flow (truncated lossless bit streams give lossy image representations)
- Efficient algorithm implementation on both small memory and parallel processing devices
- Region of interest coding
- Improved error resilience.

Amazingly, all the features of JPEG2000 are realized by coding the image data only once. This contrasts sharply with previously used image compression methods where many of the image properties, such as image size or quality, are fixed at compression time. Thus, to get compressed data representing different image sizes or quality, multiple codings of the original data, and the subsequent storage of multiple compressed representations were required. Given all the advantages to JPEG2000, are there any disadvantages? One main disadvantage, which should become clear after reading this description, is that JPEG2000 is considerably more complex than its predecessor JPEG. Thus JPEG may still be the preferred method for coding images at medium compression rates where it is only slightly worse than JPEG2000 in terms of distortion performance.

This article gives a brief overview of the JPEG2000 coding algorithm, covering only Part 1 or the baseline of the standard. Section 2 first describes some basic concepts, namely, image progressions and embedded representations, that are needed to understand the standard. Section 3, the longest and most detailed section, describes the JPEG2000 coding algorithm from the perspective of encoding an image. Section 4 gives some performance results of JPEG2000 relative to the SPIHT algorithm and JPEG, and give conclusions.

2. BACKGROUND

2.1. Image Scaling

The JPEG2000 standard is capable of progressively decoding images in several different ways corresponding to different ways of scaling image data. The current standard uses four methods of image scaling: resolution, quality, position, and color. *Resolution scaling* corresponds to changing the image size, where increasingly larger resolution gives larger image sizes (see Fig. 1). Resolution

Figure 1. An example of image progression by resolution (or size).

scaling is useful in applications where an image is decoded to different display sizes, such as a palmtop display or a 21-in. monitor, and in Web serving applications where a small thumbnail image is typically downloaded prior to downloading a full-sized image.

Quality scaling, also called *signal-to-noise* (SNR) *scaling*, refers to altering the numerical precision used to represent the pixels in an image. Increasing the quality corresponds to higher fidelity images as shown in Fig. 2. Quality scaling is useful when decompressing pictures to displays having different capabilities, such as displays having only black-and-white pixels or those having 256 levels of grayscale. Position or spatial scaling refers to altering the order in which smaller portions of an image are used to progressively build up the entire image. For example, most printers print images from top to bottom in a raster scan order, corresponding to increasing the spatial scale of the image. *Spatial scaling* is useful in printers which may have to print large images with limited memory and in applications that require random access to locations of an image. *Color scaling* refers to changing the number of colors planes [such as RGB (red-green-blue)] used to represent an image and is useful when a color image is printed or displayed on a black-and-white printer or monitor.

2.2. Embedded Bit Streams

A powerful tool used in the JPEG2000 coding algorithm is that of embedded coding. A binary bit stream is said to be embedded if any truncation of the bit stream to length L results in an optimal compressed representation. By *optimal* we mean that no other compressed representation of length L, embedded or not, will have better resulting distortion than the truncated embedded bit stream. EZW [1] and SPIHT [2] are good examples of algorithms that give embedded representations that are quality scalable. Truncating the bit streams resulting from these algorithms thus gives lower SNR representations, where each representation is optimal for the truncated length. A significant advantage of embedded bit streams is that the compressed representations of an image at many different rates can be achieved by coding the image only once. The final optimal compressed representation is achieved by simply truncating the bit stream from the single coding to the desired bitlength.

Embedded bit streams can be constructed such that when sequentially decompressed they give the image progressions described above. Since the image progressions are different, it would appear that a separate coding run would be required to create the embedded bit stream corresponding to each progression. A major breakthrough achieved by the JPEG2000 coding algorithm is that the embedded bit streams corresponding to all the basic image progressions can be achieved by doing only one coding. The way JPEG2000 is able to do this is by dividing up a transformed image into a number of independent codeblocks. Each codeblock is independently compressed to form its own quality scalable bit stream called an *elementary embedded bit stream*. The set of all elementary bit streams from all codeblocks are annotated and collected together into a database called a *generalized embedded representation* [3]. The creation of this database is the initial step of the coding process. The second and final step consists of extracting, annotating, and ordering the bits from this generalized representation to give the final coded image representation. This final coded representation gives the progression desired as well as being an embedded representation, and thus optimal for any given truncation point. JPEG2000 is able to create these final coded representations by only selecting and rearranging the bits resulting from the single initial coding.

3. JPEG2000 ENCODING

3.1. Preprocessing (Tiling and Color Transform)

The first step in the JPEG2000 coding algorithm is to divide the image into rectangular regions called "tiles." These tiles are all of the same size and completely cover the image in a regular array. The tile size can be chosen to be as large as the image itself, in which case the image consists of only one tile. For purposes of coding, each

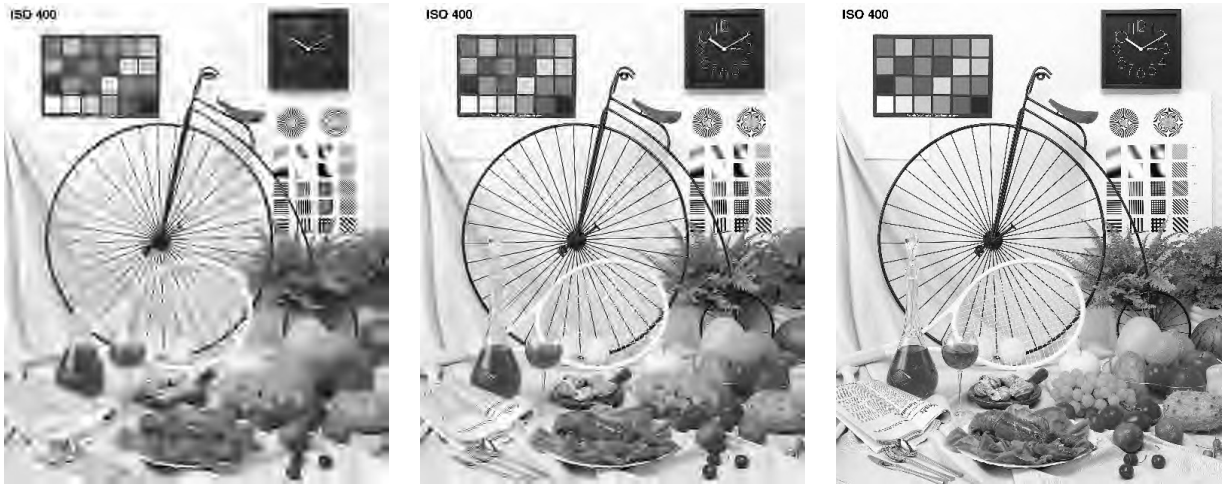


Figure 2. An example of image progression by quality (or SNR).

tile can be treated as an independent image having its own set of coding parameters. Thus tiling can be useful in coding compound documents, which are documents having separate subregions of texts, graphics, and picture data. Tiles from compound images that contain only text data can be compressed with parameters that are very different from tiles that contain only picture data. Tiling can also be used to process very large images using systems with small amounts of memory, and to give random spatial access to compressed data (although this spatial access can also be done in a single tile). The main disadvantage of tiling is that it leads to blocking artifacts and reduced compression performance similar to that found in the original JPEG standard.

After tiling, color images, or images consisting of more than one color component, can be transformed to reduce the redundancy amongst the color components. Two transforms are defined in JPEG2000 for use on standard *RGB* color data. The first is a linear transform which converts the three *RGB* components into a luminance (or black-and-white) component, and two chrominance (or color) components denoted YC_bC_r. This transform, called the *irreversible color transform* (ICT), cannot exactly recover the original data from the transformed data and therefore is used for lossy coding applications. The second transform, a nonlinear transform called the *reversible color transform* (RCT), converts RGB data into three components $Y'D_bD_r$. The RCT is an integer approximation to the ICT that maps integers to integers in a reversible manner. Because it is reversible, the RCT is the only transform used for lossless compression. The separate color components resulting from either the ICT or RCT are then coded independently. Since tiles and color components are coded independently and in the same manner, this article assumes for simplicity that the image being coded has only one tile and one color component.

The final preprocessing step is to remove any average (or DC) value in the image coefficients by adding a constant value to the image. Eliminating the average value reduces the probability of overflow and allows the JPEG2000 algorithm to take full advantage of fixed precision arithmetic hardware.

3.2. Wavelet Transform

After preprocessing the image coefficients are transformed using a standard dyadic (power of 2) discrete wavelet transform (DWT). An example three-level DWT with corresponding notation is shown in Fig. 3. Note that the number of levels of wavelet decomposition M gives rise to $M+1$ well defined image resolutions (sizes). These resolutions are numbered from the smallest (r_0) to the full image size before transformation (r_M) as shown in Fig. 3. JPEG2000 uses a special form of the DWT called the *symmetric wavelet transform* (SWT), which is able to handle border effects and has the property that the transformed image has the same number of coefficients as the original image [4]. The wavelet transform can be implemented using either a filterbank or lifting methods, and special methods can be employed by the transform to reduce memory requirements [5].

Part 1 of the JPEG2000 standard specifies only two sets of filter coefficients to be used in the transform: the 9/7 and 5/3 filters (where the numbers correspond to filter lengths). Wavelet transforms using the 9/7 filters and finite precision arithmetic cannot exactly recover the original image from the wavelet transformed coefficients. Thus the 9/7 filters are only used in lossy coding applications. Wavelet transforms using the 5/3 filters and lifting map integers to integers such that the exact original image can be reconstructed from the wavelet-transformed coefficients. Thus the 5/3 filters are the only ones used for lossless compression. The 5/3 filters can also be used

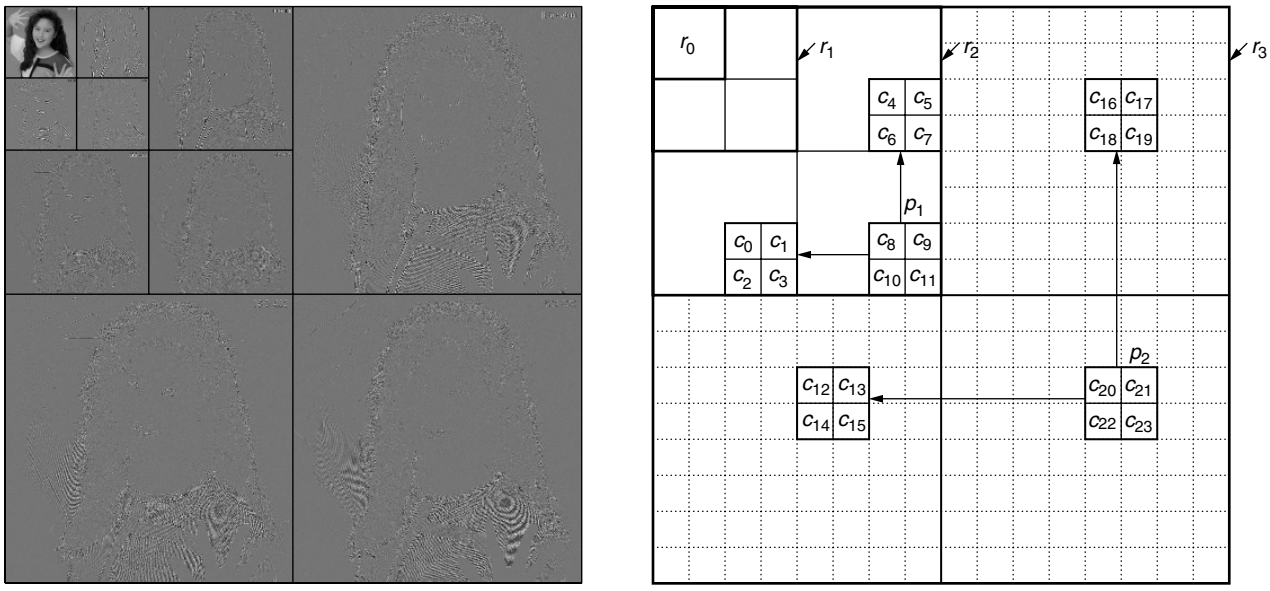


Figure 3. An example three-level wavelet decomposition and notation depicting resolutions, r_i, example codeblocks, c_i, and example precincts, p_i. Codeblocks $c_0 - c_{11}$ correspond to precinct p_1 and codeblocks $c_{12} - c_{23}$ correspond to precinct p_2.

in lossy compression but are not preferred since the 9/7 filters give better lossy compression performance.

3.3. Embedded Quantization

The wavelet coefficients are quantized using a uniform "deadzone" quantizer, where the deadzone is twice as large as the quantization step size and symmetric about zero (see Fig. 4). A deadzone quantizer is used since the deadzone improves coding performance when the quantizer is used with entropy coding. The deadzone quantizer can be represented mathematically as

$$q = \text{sign}(w) \left\lfloor \frac{|w|}{\Delta_b} \right\rfloor \tag{1}$$

where w is the unquantized wavelet coefficient in subband b, and Δ_b is the quantization step size for subband b. The quantized wavelet coefficients are represented in binary signed magnitude form. Specifically, one bit of the binary number represents the sign, zero for positive, and the remainder of the bits represent the quantized magnitude (see Fig. 4). The quantization step includes some loss of information since all wavelet coefficients in a quantization interval of size Δ_b are mapped into the same quantized value. For lossless coding applications, the wavelet coefficients are not quantized, which corresponds to setting $\Delta_b = 1$ in Eq. (1).

The quantized wavelet coefficients are coded using what is called bit plane coding [6]. In bit plane coding the most significant magnitude bits of all the coefficients are coded prior to coding the next most significant bits of all the coefficient magnitudes, and so forth. The combination of quantization and bit plane coding can be viewed as quantizing data with a set of successively finer quantizers, and this process is called embedded quantization. A set of embedded scalar quantizers is shown in Fig. 4. These quantizers have the property that finer quantizers are formed by subdividing the quantization intervals of a more coarse quantizer. The result of using embedded quantization in JPEG2000 is that the quantization using interval Δ_b includes all the quantizations having coarser quantization intervals of $2^k\Delta_b$ where $0 \le k \le k_{\max}$. Note also from Fig. 4 that the quantization interval index is formed by appending a bit to the index of the next coarser interval to which it belongs.

Embedded quantization is the method in which JPEG2000 is able to construct quality scalable compressed data. Because of embedded quantization the choice of finest quantization interval Δ_b is not critical. The interval Δ_b is typically chosen to be rather narrow (or fine), and the resulting quantization interval of the compressed image ($2^k\Delta_b$) is determined by truncation of the embedded bit stream.

3.4. Bit Plane Encoding

Prior to coding the quantized wavelet coefficients, the wavelet subbands are divided into a regular array of relatively small rectangles called *codeblocks* (see Fig. 3). Each of these codeblocks is then coded independently to form its own elementary embedded bit stream (see Section 2.2). Coding small blocks rather than entire subbands or the whole image offers several advantages. Since the blocks are small, they can be coded in hardware having limited memory, and since the blocks are independently coded, several blocks can be coded in parallel. Independently coding blocks also gives better error resilience since the errors in one block will not propagate into other blocks (error propagation can be a significant problem in EZW and SPIHT). Having a large number of blocks makes the resulting coded bit stream more flexible. The coded blocks can be arranged such that different progressive decodings are possible. Also, by using a large number of codeblocks the rate distortion performance of the compressed image can be optimized without further coding. This is accomplished by selecting only the best bits from each compressed codeblock in a process called *postcompression rate distortion optimization* (see book by Taubman and Marcellin [3] and Section 3.6).

Each codeblock is coded to give a quality embedded bit stream. The bit stream has the property that those bits that reduce the output distortion the most come first in the bit stream. This is accomplished by bit plane encoding the wavelet coefficients starting with the most significant bit plane and ending with the least significant one. To see why the most significant bits must be coded first to get a quality embedded bit stream, consider the case of an encoder wanting to send a coefficient to a decoder. If the binary value of the coefficient is $10{,}010_2$ and only one bit could be sent, which would be the best bit? The answer from a squared error perspective is to send the most significant bit since this results in a squared reconstruction error of only $(10{,}010_2 - 10{,}000_2)^2 = (10_2)^2 = 4$ while sending the lower significant bit results in a larger squared error of $(10{,}010_2 - 10_2)^2 = (10{,}000_2)^2 = 256$.

JPEG2000 uses a quantized coefficient's significance to determine how it will be coded in a bit plane. A coefficient is defined to be significant with respect to a bit plane if the coefficient's magnitude has a nonzero bit in the current or higher bit plane. Defining the significance function $\sigma_k(q)$

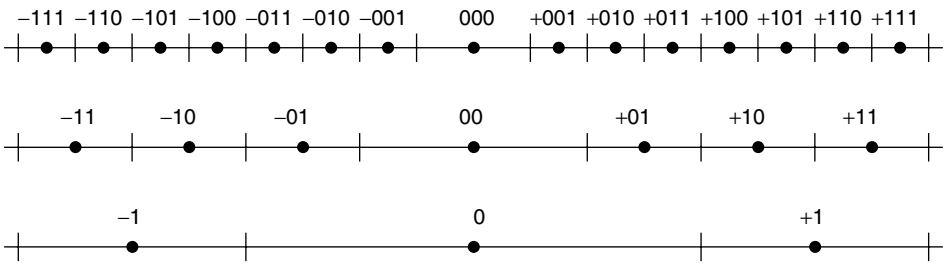


Figure 4. A set of embedded deadzone quantizers. Quantization intervals are indexed with sign magnitude binary format. Increasing quantization resolution is implemented by subdividing intervals and appending bits to interval indices.

of a quantized wavelet coefficient q at bit plane k as

$$\sigma_k(q) = \left\lfloor \frac{|q|}{2^k} \right\rfloor$$

then a wavelet coefficient is significant in bit plane k if $\sigma_k(q) > 0$. Coding of coefficients in a codeblock begins in bit plane $k_{\max}$, where $k_{\max}$ is the highest bit plane such that at least one of the wavelet coefficients in the block is significant ($k_{\max} = \lfloor \log_2(\max|q|) \rfloor$, where the maximization is over all quantized coefficients q in the codeblock).

JPEG2000 codes each bit plane using three separate passes, where each pass corresponds to what is called a *fractional bit plane*. This is different from earlier coders such as EZW and SPIHT, which used only one pass per bit plane. These previous methods argued that all bits in a bit plane reduced the output distortion by the same amount. Although this is true, further research showed that some bits in a bit plane have a higher distortion rate slope than do others [7,8]. The intuition behind this is that all bits in a bit plane (input bits) reduce the image distortion by the same amount ΔD. However, after entropy coding (discussed in the next section), some input bits require fewer output bits ΔL to encode. The result is that the distortion rate slopes $\Delta D/\Delta L$ are different for different input bits. Coding bits with higher slopes first leads to a coding advantage as illustrated in Fig. 5. The bottom curve in this figure shows the optimal distortion rate curve resulting from entropy-coded quantization and continuously varying the quantization step size. The two dark dots give the distortion rate results at the end of coding the k and $k-1$ bit planes, corresponding to quantizing with step sizes $2^k\Delta_b$ and $2^{k-1}\Delta_b$ respectively. Without fractional bit planes truncation of the bit stream results in distortion decreasing linearly with increased code length, since on the average keeping a fraction $\alpha\Delta L$ of the bits reduces the distortion by the amount $\alpha\Delta D$. By coding the higher slope bits first, truncated bit streams give distortion results that are much closer to the optimum rate distortion curve.

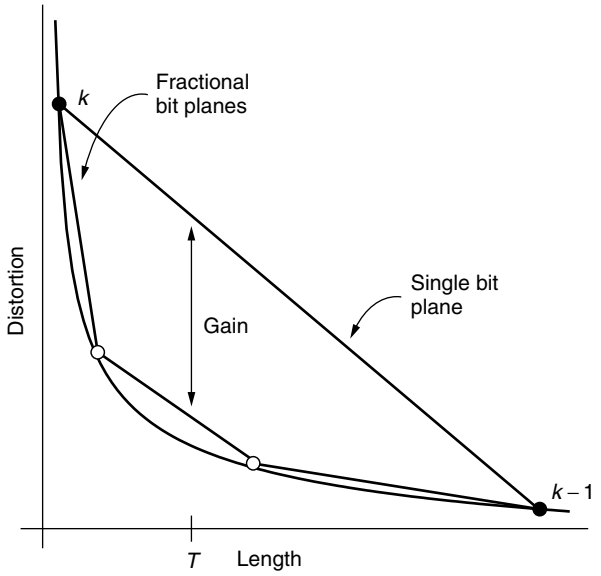


Figure 5. An illustration showing the benefit of using fractional bit plane coding when truncating to an arbitrary length T. Fractional bit planes result in reduced distortion when the code stream is not truncated at bit plane boundaries.

The 3 passes corresponding to the three fractional bit planes scan the codeblock using the scanning pattern shown in Fig. 6. The scan pattern is basically a raster scan where each scan line consists of several height four columns of coefficients. The first pass, called the *significance propagation* (SP) pass, codes all the insignificant bits that are immediate neighbors of coefficients that have been previously found significant. By "previously significant coefficients," we mean coefficients found significant in a previous bit plane or in the current bit plane earlier on in the scanning pattern of this pass. Each coefficient in this pass is coded with what is called *standard coding*, which is either a 0 to indicate that the coefficient remains insignificant, or a 1 and a sign bit to indicate that the coefficient has become significant. The second pass is called the *magnitude refinement* (MR) pass. It codes the current bit in the bit plane corresponding to coefficients found significant in a prior bit plane. That is it codes the next most significant bit in the binary representation of the these coefficients. Coefficients that became significant in the current bit plane are not coded in this pass since their values were already coded in the SP pass.

The final pass is the "cleanup" (CL) pass, which codes the significance of all coefficients in the bit plane not coded in the previous two passes. The bits from this pass are coded either using standard coding or with a special run mode designed for coding sequences of zeros. The special run mode is advantageous since at medium to high compression ratios most of the output bits from this pass will be zero [3]. Run mode is entered if a height 4 column from the scanning pattern satisfies the following: (1) the four coefficients in the column are currently insignificant (i.e., all the coefficients are to be coded in this pass) and (2) all the neighbors of the four coefficients are currently insignificant (i.e., the neighbors are either outside the codeblock and thus considered insignificant or are all to be coded in this pass and if already coded, have not become significant). Note that the conditions for run mode can be deduced at both the encoder and decoder so that no extra

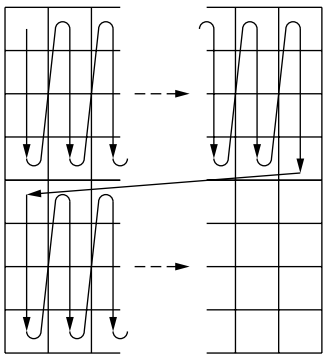


Figure 6. The scanning pattern for the three fractional bit plane passes.

side information needs to be sent. When in run mode, a 0 indicates that all four coefficients in the column remain insignificant while a 1 or "run interrupt" indicates that one or more of the coefficients have become significant. After a run interrupt 2 bits are sent to indicate the length of the run of zeros in the column (0 to 3), followed by the sign of the first significant bit. Standard coding resumes on the next coefficient.

The ordering of the passes is SP, MR, and CL for each bit plane except the first. Because there are no significant bits from previous bit planes, the first bit plane starts with the CL pass.

3.4.1. Bit Plane Coding Example. Table 1 shows the results of bit plane coding some example quantized wavelet coefficients. For clarity signs are shown as +/− instead of 0/1 and the following comments refer to the table:

1. This column satisfies the conditions for run mode, and since all the coefficients remain insignificant in this pass, they can be coded with a single 0.
2. This column satisfies the conditions for run mode but has a coefficient that becomes significant, namely, 25. The first 1 ends the run mode, and the 01 indicates a run of one insignificant coefficient. The next coefficient, 25, is known to be significant so only its sign is coded.
3. Since −11 became significant in this pass and 3 is its neighbor later in the scanning pattern, it is coded in the SP pass.

Table 1. The First Three Passes Resulting from Bit Plane Coding the Quantized Wavelet Coefficients[a]

3	5	−1
−1	25	2
7	−2	−11
1	−6	3

	Pass						
Coef.	cl4	sp3	mr3	cl3	sp2	mr2	cl2
3	0(1)	0			0		
−1		0			0		
7		0			1+		
1				0	0		
5	101(2)	0			1+		
25	+		1			0	
−2	0	0			0		
−6	0			0	1−		
−1	0	0			0		
2	0	0			0		
−11	0	1−	(4)			0	
3	0	0(3)			0		

[a]Superscript numbers in parentheses refer to listed items in Section 3.4.1.

4. Since −11 was coded in the current bit plane in the SP pass, it is not coded in the MR pass.

3.5. Entropy Coding

The sequence of ones and zeros resulting from bit plane coding is further compressed through entropy coding using a binary arithmetic encoder. Information theory results show that the minimum average number of bits required to code a sequence of independent input symbols is given by the average entropy:

$$E_{\text{av}} = -P_1 \log_2 P_1 - (1 - P_1) \log_2 (1 - P_1).$$

The input symbols are the zeros and ones from bit plane coding, and P_1 is the probability that the symbol is a 1. If $P_1 = 0.5$, the formula above shows that it requires on average 1 bit per symbol to code a sequence; thus, coding doesn't help. However, if P_1 differs or is skewed from 0.5, then it will require less than 1 bit per symbol on average to encode the binary input sequence. Arithmetic encoders are very useful since they are able to code sequences of symbols at rates approaching the average entropy. A binary arithmetic encoder requires two pieces of information in order to do its encoding: (1) the symbol (1 or 0) to be coded and (2) a probability model (the value of P_1). A great utility of the arithmetic encoder is that it can dynamically estimate P_1 by keeping a running count of the input symbols it receives. The more symbols the arithmetic encoder processes, the closer these counts will be to approximating the true probability P_1.

The specific arithmetic encoder used in JPEG2000 is called the *MQ coder*, and is the same coder used by the JBIG standard. In order to get probabilities skewed from 0.5, JPEG2000 codes bits according to their context, which is determined by the value of the eight nearest-neighbor coefficients to the coefficient being coded. A symbol probability P_1 is computed for each context, and each symbol is coded with the probability corresponding to its own context. By using contexts, the final symbol probabilities are more skewed from 0.5 than they would be without using contexts. Thus the use of contexts results in a coding gain.

In JPEG2000 contexts are formed by labeling neighbors as significant or insignificant, making a total of 2^8 contexts possible. By careful experimentation and consideration of symmetries the JPEG2000 algorithm was able to reduce this number down to only 18 contexts for arithmetic encoding: 9 for standard significance coding, 5 for sign coding, and 3 for magnitude refinement coding. A full description of the contexts and how they were derived can be found in Taubman and Marcellin's book [3]. Having a small number of contexts allows the probability models formed by counting symbols to quickly adapt to the true probability models. This is important since JPEG2000 uses small codeblocks and the probability models are reinitialized for each codeblock.

3.6. Final Embedded Bit Stream Formation

The result of bit plane coding is that each codeblock is represented by an elementary embedded bit stream. The

next step is to process and arrange these elementary bit streams into a generalized embedded representation. The structure of this generalized representation is designed to make it easy to extract final coded image representations having the scaling and embedded properties discussed in Section 2. It is easy to construct resolution and scalable compressed images from the elementary bit streams. For example, resolution scalable data can be formed by concatenating elementary bit streams, starting with the lowest-resolution codeblocks and following with the higher-resolution codeblocks (extra data need to be included to indicate codeblock location and lengths). Sequentially decoding this bit stream gives resolution scaling since lower-resolution subband data appears before higher resolution subband data. Spatial scaling is implemented from this same bit stream by decoding only those blocks in the datastream associated with a particular region of interest. The only scaling not possible from this simple bit stream is quality scaling.

Quality scaling is introduced in JPEG2000 by dividing up the bits from the elementary embedded bit streams into a collection of quality layers. A quality layer is defined as the set of bits from all subbands necessary to increase the quality of the full sized image by one quality increment. These layers can be implemented by adding information to the elementary bit streams to indicate those bits belonging to each quality layer. Since each elementary bit stream is embedded, this amounts to selecting a set of increasing truncation points in each elementary embedded bit stream. These truncation points are found by selecting a set of increasing final codelengths $L_1 < L_2 < \cdots < L_N$, where L_N is the sum total of all the bits in all the elementary bit streams. The optimal set of L_1 bits from all the codeblocks that minimizes the distortion is then selected. These bits are indicated in each elementary bit stream by a truncation point and constitute the bits in the first quality layer. Next, the optimal set of L_2 bits from all the codeblocks that minimizes the distortion is selected. Since the elementary bit streams are embedded, this set consists of the L_1 bits in the first quality layer with an additional $L_2 - L_1$ bits. These additional bits are indicated in each elementary bit stream by another truncation point (one per elementary bit stream) and constitute the second quality layer. The process, called *postcompression rate distortion optimization*, is repeated for the remaining lengths L_n to form all the quality layers.

The data structure JPEG2000 uses in the generalized representation to track resolution, spatial location, and quality information is called a "packet." Each packet represents one quality increment for one resolution level at one spatial location. Since spatial locations are spread across three subbands for a particular resolution level, JPEG2000 uses precincts to refer to spatial locations. A "precinct" is defined as a set of codeblocks at one resolution level that correspond to the same spatial location as shown in Fig. 3. A full-quality layer can then be defined as one packet, from each precinct, at each resolution level.

Packets are the fundamental data structure that JPEG2000 uses to achieve highly scalable embedded bit streams since packets contain incremental resolution, quality, and spatial data. Simply rearranging the packet ordering results in final coded bit streams that are resolution, quality, or spatially scalable. The only complexity involved in forming these final coded representations is that of reordering data.

4. PERFORMANCE AND CONCLUSIONS

Table 2 shows lossy image coding results (in terms of peak SNR [6]) of JPEG2000 relative to SPIHT and the original JPEG standard. The results were generated using publicly available computer programs for JPEG [9] and SPIHT [10], and programs distributed in [3] for JPEG2000. The test images used are the monochrome, 8 bit grayscale woman and bike images shown in Figs. 1 and 2. To give an idea of visual quality versus bit rate, the bike images in Fig. 2 were coded at rates of 0.0625, 0.25, and 1 bit per pixel. The results show that the performance of SPIHT is very close to JPEG2000. However, remember that JPEG2000 is more flexible (SPIHT is not resolution or color scalable) and thus incurs some overhead due to this flexibility.

The JPEG2000 standard represents a culmination of over a decade of wavelet based compression research. It represents a fundamental shift from Fourier based techniques to wavelet based techniques that was enabled by the discovery of wavelet analysis. Research in image compression for the foreseeable future will focus on improving and extending wavelet-based techniques such as JPEG2000 until the discovery of new enabling technologies and theory.

BIOGRAPHY

Bryan E. Usevitch received the B.S. degree in electrical engineering (magna cum laude) from Brigham Young University, Provo, Utah, in 1986, and the M.S. and Ph.D. degrees from the University of Illinois at Urbana–Champaign in 1989 and 1993, respectively. From 1986 to 1987 and 1993 to 1995 he was a staff engineer at

Table 2. Peak SNR Results from Lossy Coding the Images of Figs. 1 and 2[a]

Bit Rate (bpp)	0.1	0.2	0.5	1.0
	Woman			
JPEG	—	24.88	29.68	33.55
SPIHT (lossless)	26.34	28.56	33.04	37.58
SPIHT (lossy)	26.72	28.93	33.56	38.33
JPEG2000 (lossless)	26.16	28.31	32.84	37.54
JPEG2000 (lossy)	26.76	29.02	33.62	38.48
	Bike			
JPEG	—	24.67	30.14	34.37
SPIHT (lossless)	24.67	27.64	32.64	36.98
SPIHT (lossy)	24.92	28.04	33.01	37.70
JPEG2000 (lossless)	24.98	28.00	32.95	37.34
JPEG2000 (lossy)	35.51	28.49	33.52	38.09

[a]Recall that truncating lossless encoded data gives lossy compression. JPEG values are only approximate due to the difficulty in coding at exact bit rates.

TRW designing satellite communication systems. In 1995 he joined the department of electrical engineering at the University of Texas at El Paso where he is currently an associate professor. Dr. Usevitch's research interests are in signal processing, focusing on wavelet based image compression and multicarrier modulation.

BIBLIOGRAPHY

1. J. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. Signal Process.* **41**: 3445–3462 (Dec. 1993).
2. A. Said and W. Pearlman, A new, fast and efficient image codec based on set partitioning, *IEEE Trans. Circuits Syst. Video Technol.* **6**: 243–250 (June 1996).
3. D. Taubman and M. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*, Kluwer, Boston, 2002.
4. B. Usevitch, A tutorial on modern lossy wavelet compression: Foundations of JPEG2000, *IEEE Signal Process. Mag.* **18**: 22–35 (Sept. 2001).
5. C. Chrysafis and A. Ortega, Line based, reduced memory, wavelet image compression, *IEEE Trans. Image Process.* **9**: 378–389 (March 2000).
6. A. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
7. J. Li and S. Lei, Rate-distortion optimized embedding, *Proc. Picture Coding Symp.*, Berlin, Sept. 1997, pp. 201–206.
8. E. Ordentlich, M. Weinberger, and G. Seroussi, A low-complexity modeling technique for embedded coding of wavelet coefficients, *Proc. IEEE Data Compression Conf.*, Snowbird, UT, March 1998, pp. 408–417.
9. I. J. Group, *Home Page*, (online) (no date), *http://www.ijg.org*.
10. A. Said and W. Pearlman, *SPIHT Image Compression*, (online) (no date), *http://www.cipr.rpi.edu/research/SPIHT/spiht3.html*.

KASAMI SEQUENCES

TADAO KASAMI
RYUJI KOHNO
Hiroshima City University
Hiroshima, Japan

1. INTRODUCTION

Code-division multiple-access, (CDMA) based on spread-spectrum communication systems requires many pseudo-noise sequences with sharp autocorrelation and small cross-correlation. On one hand, sharp autocorrelation results in not only reliable and quick acquisition but also tracking of sequence synchronization. On the other hand, small cross-correlation reduces interference between multiple accessing users' signals, and thus results in increasing the number of multiple accessing users in CDMA [1–5].

Kasami sequences can be generated with a binary linear feedback shift register in the same manner as maximum-length sequences abbreviated as m sequences, Gold, Gold-like, and dual-BCH sequences. Kasami sequences are classified into two sets: (1) the *small set of Kasami sequences* and (2) the *large set of Kasami sequences*, which have period $2^n - 1$, n even. For a small set of Kasami sequences, the peak correlation magnitude, the maximum absolute value of cross-correlations, is optimal and approximately half of that achieved by the Gold [6,7] and Gold-like sequences of the same period. However, its size — the number of sequences in the set — is approximately the square root of that of the Gold or Gold-like sequences. A large set of Kasami sequences contains a set of Gold or Gold-like sequences and the small set of Kasami sequences as subsets. Compared with a set of Gold or Gold-like sequences of the same period, its peak correlation magnitude is the same, and its size is approximately $2^{n/2}$ times larger.

In this article, we consider only binary sequences. Sarwate suggested extension to nonbinary or polyphase sequences in a manner similar to that for m sequences (for the term *nonbinary Kasami sequences*, see Ref. 2).

2. DEFINITIONS AND BASIC CONCEPTS

(For further details, see Refs. 1, 2, 8, and 9.) By a *sequence*, we mean a binary infinitely long sequence with a finite period. A sequence $\mathbf{u} = \ldots, u_{-2}, u_{-1}, u_0, u_1, u_2, \ldots$ is abbreviated as $\{u_j\}$. Let T denote the left-shift operator by one bit. For an integer i, T^i denotes the i-times applications of T; that is, $T^i\{u_j\} = \{u_{j+i}\}$. The period N of $\mathbf{u}$ is the least positive integer such that $u_i = u_{i+N}$ for all i. $T^i\mathbf{u}$ is called a phase shift of $\mathbf{u}$.

Define $X(0) = +1$ and $X(1) = -1$, where X represents binary shift keying modulation, that is, $X(t) = \exp\{i\pi t\} = (-1)^t$ for $t \in \{0, 1\}$, $i = \sqrt{-1}$. For a sequence $\mathbf{u}$, $wt(\mathbf{u})$ denotes the number of ones per period in $\mathbf{u}$. For sequences $\mathbf{u} = \{u_i\}$ and $\mathbf{v} = \{v_j\}$ with the same period N, the *periodic cross-correlation function* $\theta_{u,v}(\cdot)$ is defined by

$$\theta_{u,v}(\tau) \triangleq \sum_{j=0}^{N-1} X(u_j)X(v_{j+\tau}) \qquad \text{for} \qquad 0 \le \tau < N$$
$$= N - 2wt(\mathbf{u} \oplus T^\tau \mathbf{v}) \tag{1}$$

For a special case of $\mathbf{u} = \mathbf{v}$, the expression $\theta_{u,v}(\cdot)$ is called the *periodic autocorrelation function*. For a set S of sequences with period N, the *peak cross-correlation magnitude* $\theta_{\max}$ of S is defined by

$$\theta_{\max} \triangleq \max\{\theta_{u,v}(\tau) : \mathbf{u}, \mathbf{v} \in S \text{ and } 0 \le \tau < N\} \tag{2}$$

For sequences $\mathbf{u} = \{u_j\}$ and $\mathbf{v} = \{v_j\}$, $\mathbf{u} \oplus \mathbf{v}$ denotes $\{u_j \oplus v_j\}$, that is, the sequence whose jth element is $u_j \oplus v_j$, where $\oplus$ denotes addition modulo 2, that is, the exclusive OR operation. For a positive integer f, consider the sequence $\mathbf{v} = \{v_j\}$ formed by taking every fth bit of sequence $\mathbf{u} = \{u_j\}$, that is, $v_j = u_{fj}$ for every integer j. This sequence, denoted $\mathbf{u}[f]$, is said to be a decimation by f of $\mathbf{u}$.

Let $h(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ denote a binary polynomial of degree n where $a_0 = a_n = 1$ and other coefficients are value 0 or 1. A sequence $\mathbf{u} = \{u_j\}$ is said to be a *sequence generated by* $h(x)$ if for all integers j

$$a_0u_j \oplus a_1u_{j-1} \oplus a_2u_{j-2} \oplus \cdots \oplus a_nu_{j-n} = 0 \tag{3}$$

Subsequence $u_0, u_1, u_2, \ldots$ can be generated by an n-stage binary linear feedback shift register that has a feedback tap connected to the ith cell if $h_i = 1$ for $0 < i \le n$. If $\mathbf{u}$ is a sequence generated by $h(x)$, then its phase shift is also generated by $h(x)$. The period of a nonzero sequence $\mathbf{u}$ generated by the polynomial $h(x)$ of degree n cannot exceed $2^n - 1$. If $\mathbf{u}$ has this maximal period $N = 2^n - 1$, it is called a *maximum-length sequence* or *m sequence*, and $h(x)$ is called a *primitive* binary polynomial of degree n.

Let $gcd(i, j)$ denote the greatest common divisor of the integers i and j. Let $\mathbf{u}$ be an m sequence. For a positive integer f, if $\mathbf{u}[f]$ is not identically zero, $\mathbf{u}[f]$ has period $N/gcd(N, f)$, and is generated by the polynomial $h_f(x)$ whose roots are the fth powers of the roots of $h(x)$. The degree of $h_f(x)$ is the smallest positive integer m such that $2^m - 1$ is divisible by $N/gcd(N, f)$. For positive integers $f_1, f_2, \ldots, f_t$ such that $h_{f_1}(x), h_{f_2}(x), \ldots, h_{f_t}(x)$ are all different, the set of sequences generated by $h_{f_1}(x)h_{f_2}(x)\ldots h_{f_t}(x)$ is the set of linear (with respect to $\oplus$) sums of some phase shifts of $\mathbf{u}[f_1], \mathbf{u}[f_2], \ldots, \mathbf{u}[f_t]$.

The *linear span* of a sequence is defined as the smallest degree of polynomials that generate the sequence, and it is also called the *linear complexity*. The *imbalance* between zeros and ones per period in a sequence $\mathbf{u}$ is defined as

the absolute value of the difference between the numbers of zeros and ones per period in $\mathbf{u}$, which is equal to $|N - 2wt(\mathbf{u})|$, where N is the period of $\mathbf{u}$.

3. SMALL SETS OF KASAMI SEQUENCES

See Refs. 1 and 2.

3.1. Definition

Let n be even and let $\mathbf{u} = \{u_j\}$ denote an m sequence of period $N = 2^n - 1$ generated by a primitive polynomial $h_1(x)$ of degree n. Define $s(n) \triangleq 2^{n/2} + 1$. Consider the sequence $\mathbf{w} = \mathbf{u}[s(n)]$ derived by decimating or sampling the sequence $\{u_j\}$ with every $s(n)$ bits. Then, $\mathbf{w}$ is a sequence of period $N' = N/gcd(N, s(n)) = (2^n - 1)/(2^{n/2} + 1) = 2^{n/2} - 1$ which is generated by the polynomial $h_{s(n)}(x)$ whose roots are the $s(n)$th powers of the roots of $h_1(x)$. Since $h_{s(n)}$ has degree $n/2$ and is primitive, $\mathbf{w}$ is an m sequence of period $2^{n/2} - 1$.

Now consider the nonzero sequences generated by the polynomial $h_S(x) \triangleq h_1(x)h_{s(n)}(x)$ of degree $3n/2$. As stated in Section 2, any such sequence must be one of the forms $T^i\mathbf{u}, T^j\mathbf{w}, T^i\mathbf{u} \oplus T^j\mathbf{w}, 0 \le i < 2^n - 1, 0 \le j < 2^{n/2} - 1$. Thus any sequence of period N generated by $h_S(x)$ is some phase shift of a sequence in the following set $K_S(\mathbf{u}, \mathbf{w})$ defined by

$$K_S(\mathbf{u}, \mathbf{w}) = \{\mathbf{u}, \mathbf{u} \oplus \mathbf{w}, \mathbf{u} \oplus T\mathbf{w}, \mathbf{u} \oplus T^2\mathbf{w}, \ldots, \mathbf{u} \oplus T^{2^{n/2}-2}\mathbf{w}\} \tag{4}$$

This set of sequences is called the *small set of Kasami sequences*.

3.2. Correlation Properties

It has been proved [10,11] that the periodic correlation functions $\theta_{x,y}(\tau)$ of any sequences $\mathbf{x}$ and $\mathbf{y}$ belonging to the small sets of Kasami sequences $K_S(\mathbf{u}, \mathbf{w})$ take only three values:

$$\theta_{x,y}(\tau) = -1 \quad \text{or} \quad -2^{n/2} - 1 \quad \text{or} \quad 2^{n/2} - 1 \tag{5}$$

It is obvious that the peak correlation magnitude of the periodic correlation function $\theta_{\max} = 2^{n/2} + 1 = s(n)$ for the small set of Kasami sequences is approximately one half of the values of $\theta_{\max} = 2^{(n+2)/2} + 1$ achieved by the Gold and Gold-like sequences.

The Welch bound [12] applied to a set of $M = 2^{n/2}$ sequences of period $N = 2^n - 1$ provides a lower bound of $\theta_{\max}$ for all binary sequences:

$$\theta_{\max} \ge N\left(\frac{M-1}{NM-1}\right)^{1/2} > 2^{n/2} \tag{6}$$

Since N is odd, it follows from Eq. (1) and Eq. (2) that $\theta_{\max}$ is also odd. This implies that $\theta_{\max} \ge 2^{n/2} + 1$. Comparing this Welch bound with $\theta_{\max} = 2^{n/2} + 1$ of the small set of Kasami sequences, it is noted that the small set of Kasami sequences is an optimal collection of binary sequences with respect to the bound.

Small sets of Kasami sequences contain only $M = 2^{n/2} = (N + 1)^{1/2}$ sequences, while Gold and Gold-like sequences contain $N + 2$ and $N + 1$ sequences, respectively.

Regarding the linear span, since $h_S(x)$ has degree $3n/2$, the maximum linear span of sequences in the set is $3n/2$. Considering the *imbalance* between the numbers of zeros and ones per period, the maximum is $2^{n/2} + 1 : 1$.

4. LARGE SETS OF KASAMI SEQUENCES

See Refs. 1 and 2.

4.1. Definition

Let n be even. Define $t(n) \triangleq 2^{(n+2)/2} + 1$. Let $\mathbf{u}$ and $\mathbf{w}$ be defined as the nonzero sequences generated by the polynomials $h_1(x)$ of degree n and $h_{s(n)}(x)$ of degree $n/2$, respectively, as mentioned above for the small sets of Kasami sequences, and let $\mathbf{v} = \mathbf{u}[t(n)]$ be a nonzero sequence generated by $h_{t(n)}(x)$ of degree n [derived by decimating or sampling the sequence $\mathbf{u}$ with every $t(n)$ bits].

The period of $\mathbf{v}$ is given by

$$\frac{N}{gcd(N, t(n))} = \begin{cases} N/3 & \text{for } n \equiv 0 \bmod 4 \\ N & \text{for } n \equiv 2 \bmod 4 \end{cases} \tag{7}$$

Then, the set of sequences generated by $h_L(x) \triangleq h_1(x)h_{s(n)}(x)h_{t(n)}(x)$ of degree $5n/2$ has period $N = 2^n - 1$, and any sequence with the period N in the set is some phase shift of a sequence in the following set $K_L(\mathbf{u}, \mathbf{v}, \mathbf{w})$, called the *large set of Kasami sequences*. There are two cases.

Case 1. If $n \equiv 2 \bmod 4$, then

$$K_L(\mathbf{u}, \mathbf{v}, \mathbf{w}) \triangleq \{G(\mathbf{u}, \mathbf{v}), G(\mathbf{u}, \mathbf{v}) \oplus \mathbf{w}, G(\mathbf{u}, \mathbf{v}) \oplus T\mathbf{w}, \ldots, G(\mathbf{u}, \mathbf{v}) \oplus T^{2^{n/2}-2}\mathbf{w}\} \tag{8}$$

where $G(\mathbf{u}, \mathbf{v})$ is the set of Gold sequences [6,7] defined by

$$G(\mathbf{u}, \mathbf{v}) = \{\mathbf{u}, \mathbf{v}, \mathbf{u} \oplus \mathbf{v}, \mathbf{u} \oplus T\mathbf{v}, \mathbf{u} \oplus T^2\mathbf{v}, \ldots, \mathbf{u} \oplus T^{2^n-2}\mathbf{v}\} \tag{9}$$

and $G(\mathbf{u}, \mathbf{v}) \oplus T^i\mathbf{w}$ denotes the set $\{\mathbf{x} \oplus T^i\mathbf{w} : \mathbf{x} \in G(\mathbf{u}, \mathbf{v})\}$.

Regarding the size of the sequences, $K_L(\mathbf{u}, \mathbf{v}, \mathbf{w})$ contains $2^{n/2}(2^n + 1)$ sequences.

Case 2. If $n \equiv 0 \bmod 4$, then

$$\begin{aligned} &K_L(\mathbf{u}, \mathbf{v}, \mathbf{w}) \\ &\quad \triangleq \{H(\mathbf{u}, \mathbf{v}), H(\mathbf{u}, \mathbf{v}) \oplus \mathbf{w}, H(\mathbf{u}, \mathbf{v}) \oplus T\mathbf{w}, \\ &\quad \ldots, H(\mathbf{u}, \mathbf{v}) \oplus T^{2^{n/2}-2}\mathbf{w} \\ &\quad \mathbf{v}^{(0)} \oplus \mathbf{w}, \mathbf{v}^{(0)} \oplus T\mathbf{w}, \ldots, \mathbf{v}^{(0)} \oplus T^{(2^{n/2}-1)/3-1}\mathbf{w} \\ &\quad \mathbf{v}^{(1)} \oplus \mathbf{w}, \mathbf{v}^{(1)} \oplus T\mathbf{w}, \ldots, \mathbf{v}^{(1)} \oplus T^{(2^{n/2}-1)/3-1}\mathbf{w} \\ &\quad \mathbf{v}^{(2)} \oplus \mathbf{w}, \mathbf{v}^{(2)} \oplus T\mathbf{w}, \ldots, \mathbf{v}^{(2)} \oplus T^{(2^{n/2}-1)/3-1}\mathbf{w}\} \end{aligned} \tag{10}$$

where $H(\mathbf{u}, \mathbf{v})$ is the set of Gold-like sequences [1–2] defined by

$$\begin{aligned} H(\mathbf{u}, \mathbf{v}) = \{&\mathbf{u}, \mathbf{v}^{(0)} \oplus \mathbf{u}, \mathbf{v}^{(0)} \oplus T\mathbf{u}, \ldots, \mathbf{v}^{(0)} \oplus T^{(2^n-1)/3-1}\mathbf{u} \\ &\mathbf{v}^{(1)} \oplus \mathbf{u}, \mathbf{v}^{(1)} \oplus T\mathbf{u}, \ldots, \mathbf{v}^{(1)} \oplus T^{(2^n-1)/3-1}\mathbf{u} \\ &\mathbf{v}^{(2)} \oplus \mathbf{u}, \mathbf{v}^{(2)} \oplus T\mathbf{u}, \ldots, \mathbf{v}^{(2)} \oplus T^{(2^n-1)/3-1}\mathbf{u}\} \end{aligned} \tag{11}$$

and $H(\mathbf{u}, \mathbf{v}) \oplus T^i\mathbf{w}$ denotes the set $\{\mathbf{x} \oplus T^i\mathbf{w} : \mathbf{x} \in H(\mathbf{u}, \mathbf{v})\}$ and $\mathbf{v}^{(i)} = (T^i\mathbf{u})[t(n)]$ is the result of decimating $T^i\mathbf{u}$ by every $t(n)$ bits.

Regarding the size of the sequences, $K_L(\mathbf{u}, \mathbf{v}, \mathbf{w})$ contains $2^{n/2}(2^n + 1) - 1$ sequences.

4.2. Correlation Properties

In either case 1 or 2, the correlation functions for any sequences $\mathbf{x}, \mathbf{y} \in K_L(\mathbf{u}, \mathbf{v}, \mathbf{w})$ take only the following five values [11]; for $0 \le \tau < N = 2^n - 1$, we obtain

$$\begin{aligned} \theta_{x,y}(\tau) = -1 \quad \text{or} \quad &-2^{(n+2)/2} - 1 \quad \text{or} \quad -2^{n/2} - 1 \\ &\text{or } 2^{(n+2)/2} - 1 \quad \text{or} \quad 2^{n/2} - 1 \end{aligned} \tag{12}$$

Thus, although the large set of Kasami sequences involves the small set of Kasami sequences and a set of Gold or Gold-like sequences as subsets, the correlation bound equals that of Gold or Gold-like sequences, that is, $\theta_{\max} = 2^{(n+2)/2} + 1$.

The maximum linear complexity of the large set of Kasami sequences is $5n/2$. The range of imbalance between the numbers of zeros and ones per period is $2^{(n+2)/2} + 1 : 1$.

Table 1 shows several measures of Kasami sequences compared with those of Gold and Gold-like sequences [2].

5. RELATION TO BINARY CYCLIC CODES

Let $h(x)$ be a binary polynomial of degree k, and let S_0 denote the set of sequences generated by $h(x)$. The greatest period N of sequences in S_0 is the smallest positive integer such that $x^N - 1$ is divisible by $h(x)$. For a sequence $\mathbf{u} = \{u_j\}$ in S_0, define $\mathbf{u}_c \triangleq (u_0, u_1, u_2, \ldots, u_{N-1})$, and let C denote $\{\mathbf{u}_c : \mathbf{u} \in S_0\}$. There is a one-to-one correspondence between S_0 and C; C is a binary cyclic code of length N whose parity-check polynomial is $h(x)$. Sequence $\mathbf{u}_c$ is called a *codeword* of C. Corresponding to the left-shift operator T, the cyclic left-shift operator T_c is defined by $T_c(u_0, u_1, u_2, \ldots, u_{N-1}) \triangleq (u_1, u_2, \ldots, u_{N-1}, u_0)$. For $\mathbf{u_c} \in C$, $T^i\mathbf{u_c} \in C$ and $T^i\mathbf{u}_c$ is called a *cyclic shift* of $\mathbf{u}_c$. The period of $\mathbf{u}$ is the least positive integer N' such that $T_c^{N'}\mathbf{u}_c = \mathbf{u}_c$, which is the same as the period of $\mathbf{u}$. The code C can be partitioned into blocks in such a way that codewords $\mathbf{u}_c$ and $\mathbf{v}_c$ belong to a block if and only if they are cyclic shifts of the other. The period of a codeword in a block is equal to the size of the block.

Let S be a minimal subset of S_0 such that any nonzero sequence with period N in S_0 is some phase shift of a sequence in S. Define $C_s \triangleq \{\mathbf{u}_c : \mathbf{u} \in S\}$. Then, C_s consists of codewords chosen as a unique representative from each block of size N. Thus the size of S is equal to the number of blocks of size N.

For a codeword $\mathbf{u}_c$, let $wt(\mathbf{u}_c)$ denote the weight of $\mathbf{u}_c$, that is, the number of ones in $\mathbf{u}_c$. The set $W \triangleq \{wt(\mathbf{u}_c) : \mathbf{u}_c \in C\}$ is called the *weight profile* of C. From Eq.(1), we obtain

$$\theta_{u,v}(\tau) = N - 2wt(\mathbf{u}_c \oplus T_c^{\tau}\mathbf{v}_c). \tag{13}$$

Since a cyclic shift of a codeword has the same weight as the codeword, the set of those values on which the correlation functions for the sequences in S take can be readily found if the weight profile of C_s is known. The profile can be easily derived from weight enumerators that give the number of codewords with any weight for C and its certain subcodes. For a class R_2 of subcodes of the second order (punctured) Reed–Muller codes, weight enumerators have been derived [10,11,13]. The class R_2 contains the codes corresponding to Gold sequences generated by $h_1(x)h_f(x)$ whose f is of form $2^e + 1$, Gold-like, dual-BCH, and Kasami sequences.

As a historical remark, Sarwate and Pursley [1,14] chose two subclasses of R_2 from a point of view of sequence design for communications applications. They translated the results on weight spectra into results on correlation spectra, and named the small and large sets of Kasami sequences. They later discovered that some of the results were already known to Massey and Uhram [15].

Table 1. Parameters of Gold, Gold-like and Kasami Sequences (2)

Sequence Set		Order n	Period N	Size M	Linear Span	Peak Cross Correlation $\theta_{\max}$	Decimation Sampler f
Gold		1 (mod 2)	$2^n - 1$	$2^n + 1$	2^n	$2^{(n+1)/2} + 1$	$2^{(n+1)/2} + 1$
		2 (mod 4)				$2^{(n+2)/2} + 1$	$2^{(n+2)/2} + 1$
Gold-like		0 (mod 4)		2^n			
Kasami	small	0 (mod 2)		$2^{n/2}$	$3n/2$	$2^{n/2} + 1$	$2^{n/2} + 1$
	large	2 (mod 4)		$2^{n/2}(2^n + 1)$	$5n/2$	$2^{(n+2)/2} + 1$	$2^{(n+2)/2} + 1$, $2^{n/2} + 1$
		0 (mod 4)		$2^{n/2}(2^n + 1) - 1$	$5n/2$		

BIOGRAPHIES

Tadao Kasami received the B.E., M.E., and D.E. degrees in communication engineering from Osaka University, Osaka, Japan, in 1958, 1960, and 1963, respectively. He joined the faculty of Osaka University in 1963, and he was a professor of engineering science from 1966 to 1994. From 1992 to 1998, he was a professor at the Graduate School of Information Science of Nara Institute of Science and Technology, Nara, Japan. He is an emeritus professor of Osaka University and Nara Institute of Science and Technology. Since 1998, he has been a professor of information science at Hiroshima City University, Hiroshima, Japan. His research and teaching interests have been in coding theory and algorithms. He is a life fellow of IEEE and a recipient of the 1999 Claude E. Shannon Award from the IEEE Information Theory Socity; a fellow of the Institute of Electronics, Information and Communication Engineers in Japan; and a recipient of the 1987 Achievement Award and 2001 Distinguished Services Award from the Institute.

Ryuji Kohno received his Ph.D. degree in electrical engineering from the University of Tokyo in 1984. Since 1998, he has been a professor in the Division of Physics, Electrical and Computer Engineering, Graduate School of Engineering, Yokohama National University. Dr. Kohno was elected a member of the Board of Governors of the IEEE IT Society in 2000. He was an associate editor of the *IEEE Transactions on Information Theory* from 1995 to 1998 and an editor of the *IEICE* (Institute of Electronics, Information, Communications Engineers) *Transactions on Communications* from 1990 to 1993. He was chairman of the IEICE Professional Group on Spread Spectrum Technology from 1995 to 1998. From 1998 to 2000, he was chairman of the IEICE Technical Group on Intelligent Transport System (ITS), and currently he is chairman of the IEICE Technical Group on Software Radio. Dr. Kohno also is an associate editor of both the *IEEE Transactions on Communications* and the *IEEE Transactions of Intelligent Transport Systems* (ITS).

BIBLIOGRAPHY

1. D. V. Sarwate and M. B. Pursley, Cross-correlation properties of pseudorandom and related sequences, *Proc. IEEE* **68**(5): 593–619 (1980).
2. P. Fan and M. Darnell, *Sequence Design for Communication Applications*, Wiley, New York, 1996.
3. E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* **9**: 48–54 (1998).
4. M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications*, Vol. 1, Computer Science Press, Rockville, MD, 1985.
5. R. C. Dixon, *Spread Spectrum Systems with Commercial Applications*, 3rd ed., Wiley, New York, 1994.
6. R. Gold, Optimal binary sequences for spread spectrum multiplexing, *IEEE Trans. Inform. Theory* **IT-13**: 619–621 (1967).
7. R. Gold, Maximal recursive sequences with 3 values recursive cross-correlation functions, *IEEE Trans. Inform. Theory* **IT-14**: 154–156 (1968).
8. S. W. Golomb, *Shift Register Sequences*, Holden-Day, San Francisco, 1967.
9. W. W. Peterson and E. J. Weldon, Jr., *Error-Correcting Codes*, 2nd ed., MIT Press, Cambridge, MA, 1972.
10. T. Kasami, *Weight Distribution Formula for Some Class of Cyclic Codes*, Coordinated Sci. Lab., Univ. Illinois, Urbana, Tech. Rep. R-285, 1966.
11. T. Kasami, *Weight Distribution of Bose–Chauduri–Hocquenghem Codes*, Coordinated Sci. Lab., Univ. Illinois, Urbana, Tech. Rep. R-317, 1966 (also in *Combinatorial Mathematics and Its Applications*, Univ. North Carolina Press, Chapel Hill, NC, 1969; reprinted in E. R. Berlekamp, ed., *Key Papers in the Development of Coding Theory*, IEEE Press, New York, 1974.
12. L. R. Welch, Lower bounds on the maximum cross-correlation of signals, *IEEE Trans. Inform. Theory* **IT-20**: 397–399 (1974).
13. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
14. D. V. Sarwate and M. B. Pursley, Applications of coding theory to spread-spectrum multiple-access satellite communications, *Proc. 1976 IEEE Canadian Commun. Power Conf.*, 1976, pp. 72–75.
15. J. L. Massey and J. J. Uhran, *Final Report for Multipath Study*, Contract NAS5-10786, Univ. Notre Dame, IN, 1969.

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 3

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 3

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at
http://www.mrw.interscience.wiley.com/eot

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor.
p. cm.
includes index.
ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G.
TK5102 .W55 2002
621.382′03 — dc21

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

L

LAND-MOBILE SATELLITE COMMUNICATIONS*

JEFFREY B. SCHODORF
MIT Lincoln Laboratory
Lexington, Massachusetts

1. INTRODUCTION

Since the early 1990s there has been significant progress in the development of land-mobile satellite communications (LMSC) systems and technology. In general, LMSC service providers have struggled to compete with their terrestrial mobile wireless counterparts. However, few doubt that LMSC systems have a meaningful role to play in the quest for global wireless access in the twenty-first century. The key advantage enjoyed by satellite communications systems is their ability to cover broad geographic areas, substantially decreasing the terrestrial infrastructure required and potentially simplifying issues relating to the coordination of this infrastructure for tasks such as channel assignment and handover. Moreover, a sufficient amount of spectrum has been allocated to LMSC systems such that they represent a good choice for the delivery of broadband services such as multimedia. Of course, LMSC systems are not perfect. While fewer satellites may be required to cover an area, satellites are very expensive to build and deploy relative to terrestrial base stations. Moreover, depending on operating frequency, significant channel impairments must be overcome in LMSC systems. Nonetheless, the potential of LMSC systems ensures they will remain an area of intense research and development for the foreseeable future.

The purpose of this article is to describe in moderate detail technical issues surrounding LMSC systems. Where appropriate, references are cited so that the interested reader can pursue these topics further. In Section 2 a brief description of several existing and planned LMSC systems is given. These systems are categorized loosely by their orbital type and are further subdivided according to the services they provide. Section 3 discusses propagation issues, including path loss, signal fading, shadowing, and the effects of directional antenna mispointing. Strategies for dealing with channel impairments are discussed in Section 4. These approaches fall into one of two main categories: error control techniques such as forward error correction (FEC) coding and automatic repeat request (ARQ) protocols, and diversity combining methods. In LMSC systems, satellite resources are typically a limiting factor. Hence, efficient use of these resources is critical. Section 5 addresses this issue with a discussion of multiple access schemes. Finally, network aspects of LMSC systems are discussed in Section 6. The primary emphasis of this section is the issue of internetworking LMSC and terrestrial data networks.

* This work was sponsored by the Department of the Army under A/F Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States government.

2. LAND-MOBILE SATELLITE SYSTEMS

Land-mobile satellite systems come in a variety of orbital configurations, including geostationary or geosynchronous earth orbit (GEO), medium earth orbit (MEO), and low earth orbit (LEO). GEO systems operate at an altitude of 35,786 km, and have an orbital period of 24 hs. MEO systems have altitudes ranging from 5000 to 10,000 km and have orbital periods of 4–6 hs. LEO satellites orbit at altitudes from 500 to 1500 km with periods of approximately 2 hs. Orbital mechanics will not be addressed here, but thorough treatments of this topic may be found in Refs. 1 and 2. Services provided by land mobile satellite systems include navigation, fleet management, broadcast, and duplex voice and data communications, the primary service of interest in this article.

LMSC systems are characterized by a forward path and a reverse path between two user terminals. Each path comprises an uplink between the transmitting terminal and the satellite, and a downlink between the satellite and the receiving terminal, as depicted in Fig. 1. Traditionally, LMSC systems satellites have been transponders (sometimes referred to as "bent pipes"), where the received uplink signal is simply amplified and translated to a downlink frequency. More recent systems have begun to employ processing satellites, where in addition to amplification and frequency translation, additional processing, such as demodulation, decoding, and remodulation is performed.

Table 1 summarizes the nomenclature used to categorize the various operating frequency bands of LMSC and other wireless communications systems. Original spectrum allocations for LMSC systems were in the L and

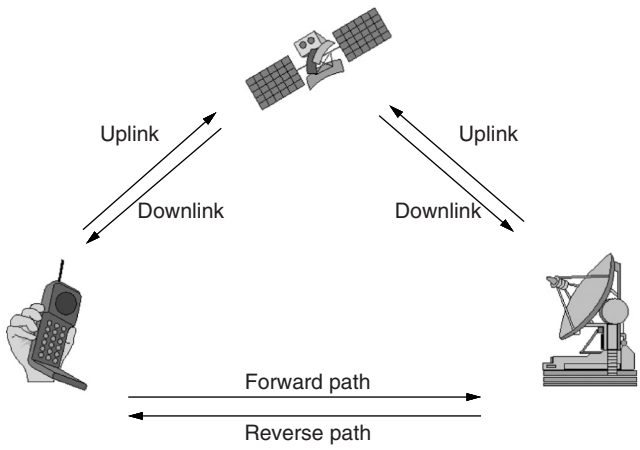


Figure 1. LMSC link nomenclature.

Table 1. Frequency Band Designations

Band	Frequency Range (MHz)
P	225–390
L	390–1550
S	1550–3900
C	3900–8500
X	8500–10,900
Ku	10,900–17,250
Ka	17,250–36,000
Q	36,000–46,000
V	46,000–56,000
W	56,000–100,000

S bands, where most systems continue to operate today. However, demand for bandwidth has resulted in additional allocations at higher frequencies. In many cases, these allocations are for fixed systems. However, at the higher frequencies terminals are typically small, and thus portable. Moreover, the distinction between the services provided by fixed and mobile systems is becoming increasingly vague. For example, consider the situation today where commercial vendors supply antenna and positioning systems that allow land mobile platforms to acquire and track DirecTV, a fixed service system. At present there is at least one operational or planned satellite system in each of the bands in Table 1, with the exception of W band.

Because of their 24-h orbital period, GEO satellites appear stationary above the equator to an observer on earth. These systems are well suited to broadcast services because a constellation size of only three or four satellites provides total earth coverage. Examples of GEO broadcast systems include DirecTV and the relatively new XM satellite radio service. For voice and data service, GEO systems have the attractive feature that no handover between satellites is necessary. On the other hand, because of their high altitude, GEO systems have long propagation delays (i.e., ~240 ms, one-way) relative to their MEO and LEO counterparts. Numerous GEO systems have been deployed for the delivery of duplex voice and data services. For example, the INMARSAT-M system provides 4.8 kbps (kilobits per second) voice capability, 2.4 kbps fax service, and 1.2–2.4 kbps data services. In addition to land-mobile terminals, INMARSAT-M also supports maritime users. Higher-data-rate systems such as the INMARSAT-4, which will support mobile communications services at rates of 144–432 kbps, are planned for the near future.

The lower orbital altitude of MEO satellites implies that they move across the sky relative to a fixed point on earth. This movement necessitates handovers between satellites. Generally, connection to the terrestrial infrastructure is achieved with a reasonable number of earth stations, or gateways. Hence, intersatellite links (ISLs) are not typically employed in MEO systems. The Intermediate Circular Orbits (ICO) system is an example of a planned MEO LMSC system. Scheduled to launch in 2004, ICO will deliver 4.8 kbps voice service, 2.4–9.6 kbps data service to handheld terminals, and 8–38.4 kbps data service to land mobile terminals. The popular Global Positioning System (GPS) is another example of a MEO satellite system, although GPS is a navigation system, as opposed to a duplex communications system.

LEO systems represent the most recent architectural system concept in LMSC systems, as well as the most complex. LEO satellites have the shortest orbital period of the three configurations discussed here. The fact that they pass rapidly over a fixed point on earth (e.g., a typical LEO satellite is in a user's field of view for 8–12 mins) implies that sophisticated handover procedures are necessary. Moreover, to connect to the terrestrial infrastructure, either a significant number of gateways are required, or else ISLs must be used. On the other hand, LEO systems offer superior delay performance and suffer less propagation loss relative to MEO and GEO systems. LEO systems are generally categorized as either "little LEO" systems or satellite personal communication networks (S-PCNs), sometimes called "big LEO" systems. Little LEO systems provide nonvoice, low-bit-rate mobile data and messaging services. Orbcomm is an example of a little LEO system currently in operation. Orbcomm offers 2.4–4.8 kbps data service to fixed and mobile users. Iridium and Globalstar are examples of S-PCNs. More information on these and other LMSC systems can be found in the literature [2,3].

3. CHANNEL CHARACTERISTICS

Fundamental to the design of any communications system is an accurate understanding of the channel over which the communications signals will be propagating. While the propagation modeling field is relatively mature for terrestrial wireless communications systems [4], it continues to be an area of active research in LMSC systems, especially at higher frequencies. In this section, numerous LMSC channel characteristics will be discussed, including random noise, path loss, weather and atmospheric effects, signal fading, shadowing, and fluctuations due to spatial tracking errors in LMSC systems that use directional antennas.

3.1. Random Noise

The dominant noise source in a communications system is usually thermal noise generated at the input to the receiver. A common practice is to lump all other noise sources together with the thermal noise and represent them as a single source added directly to the received signal. This collection of random noise is typically assumed to follow a Gaussian distribution. In digital communications systems, the effects of noise are to introduce bit errors in the received signal. The probability of error, or bit error rate (BER) in a digital communications system, is generally parameterized by the signal-to-noise ratio (SNR) per bit, or bit energy : noise ratio, E_b/N_0, where $N_0/2$ is the noise power spectral density. The relationship between BER and E_b/N_0 depends on the modulation scheme employed. Most LMSC systems use some form of constant-envelope modulation so that transmit amplifiers can be operated at saturation without introducing

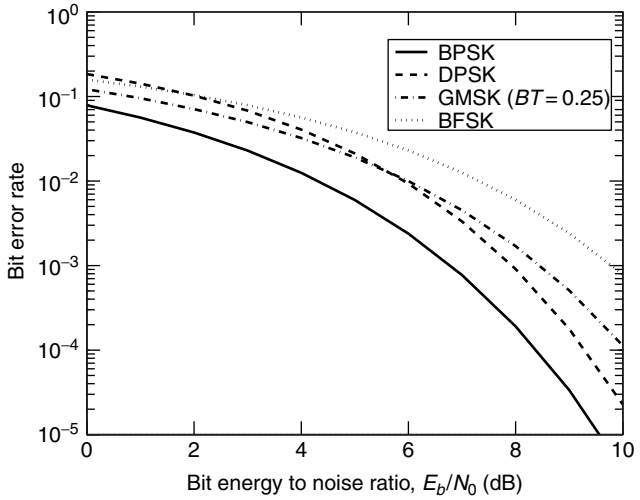


Figure 2. Performance of several modulation schemes for the Gaussian channel.

distortion into the modulated signal. Common modulation schemes include phase shift keying (PSK), frequency shift keying (FSK), and continuous-phase modulation (CPM). These techniques are described in Refs. 5 and 6, where BER expressions are derived as a function of E_b/N_0. For reference, Fig. 2 summarizes the BER performance of several modulation schemes, including binary PSK (BPSK), differential PSK (DPSK), binary FSK (BFSK), and a CPM scheme known as *Gaussian minimum shift keying* (GMSK). The corresponding equations for BER are given below:

$$\mathrm{BER}_{\mathrm{BPSK}} = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{1}$$

$$\mathrm{BER}_{\mathrm{DPSK}} = \frac{1}{2}e^{-E_b/N_0} \tag{2}$$

$$\mathrm{BER}_{\mathrm{BFSK}} = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{3}$$

$$\mathrm{BER}_{\mathrm{GMSK}} \approx Q\left(\sqrt{\frac{2\alpha E_b}{N_0}}\right) \tag{4}$$

where $Q(x) = \int_x^\infty 1/\sqrt{2\pi}e^{-u^2}\,du$ is the Gaussian Q function. The term α in the approximation for BER in GMSK systems is a scalar that depends on the time–bandwidth product, BT. For $BT = 0.25$, $\alpha = 0.68$ [6].

In order to asses the quality of a satellite link between two terminals, both uplink and downlink must be considered. In transponded satellite systems, the following relationship holds between the bit energy to noise of the total link, $(E_b/N_0)_{\mathrm{tot}}$, and the bit energy to noise ratio of the uplink and downlink, assuming that the transponder and receiving terminal bandwidths are the same:

$$\left(\frac{E_b}{N_0}\right)_{\mathrm{tot}} = \frac{(E_b/N_0)_{\mathrm{ul}}(E_b/N_0)_{\mathrm{dl}}}{(E_b/N_0)_{\mathrm{ul}} + (E_b/N_0)_{\mathrm{dl}}} \tag{5}$$

where $(E_b/N_0)_{\mathrm{ul}}$ and $(E_b/N_0)_{\mathrm{dl}}$ represent the bit energy : noise ratio of the uplink and downlink, respectively. The BER of the total link is then based on $(E_b/N_0)_{\mathrm{tot}}$ and the modulation scheme used. In processing satellites, the uplink and downlink can be analyzed separately, and the following approximation holds:

$$(\mathrm{BER})_{\mathrm{tot}} \approx (\mathrm{BER})_{\mathrm{ul}} + (\mathrm{BER})_{\mathrm{dl}} \tag{6}$$

where $(\mathrm{BER})_{\mathrm{tot}}$ is the BER of the total link, $(\mathrm{BER})_{\mathrm{ul}}$ is the uplink BER, and $(\mathrm{BER})_{\mathrm{dl}}$ is the downlink BER.

3.2. Path Loss

In LMSC systems, path loss arises from several sources. Free-space path loss arises in wireless communications systems due to the spatial dispersion of the radiated power, and is quantified as follows:

$$L_0 = 10\log\left(\frac{4\pi d}{\lambda}\right)^2 \tag{7}$$

where L_0 is the free-space path loss in dB, d is the distance between the transmitter and receiver, and λ is the signal wavelength. Figure 3 illustrates free-space loss for several frequencies as a function of satellite altitude, d_a. Note, however, that in LMSC systems the distance between a user terminal and a satellite is not the same as the satellite altitude. The user's location on the earth must be considered as well. A conceptually simple way to consider the problem is to treat the total distance between a satellite and user terminal as the sum of two terms:

$$d_{sr} = d_a + d_e \tag{8}$$

where d_e is an elevation dependent term that, when added to the satellite altitude d_a, yields the total distance between the satellite and user terminal, referred to as the *slant range*. The additional free space loss, in decibels, due to the slant range may be expressed as

$$\Delta L_{sr} = 20\log\frac{d_{sr}}{d_a} \tag{9}$$

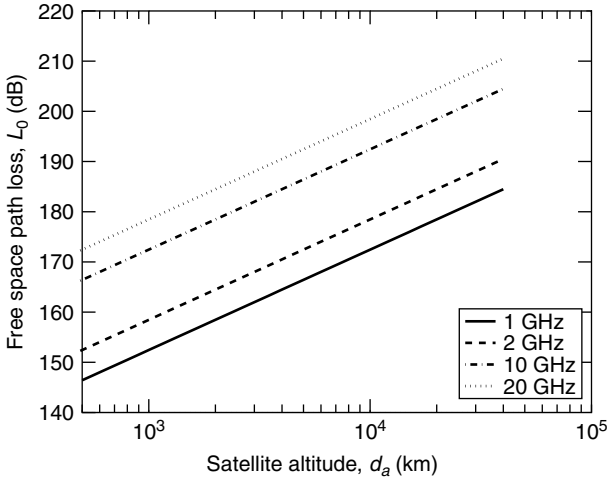


Figure 3. Free-space path loss as a function of satellite altitude d_a for several different operating frequencies.

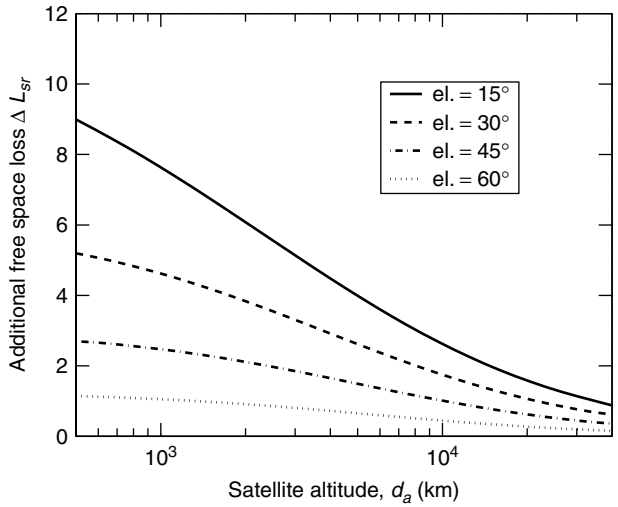


Figure 4. Additional free-space path loss due to slant range.

Figure 4 gives ΔL_{sr} as a function of satellite altitude for several different elevation angles.

Weather and atmospheric effects are another factor that contribute to path loss in LMSC systems. Above X band, signal scattering and absorption by water droplets are the dominating factors. In general, attenuation due to rain is a function of the amount of rainfall, drop size, temperature, operating frequency, and pathlength through the rain. Results from measurement campaigns at millimeter wavelengths [7,8] suggest that "typical" rain rates of 20 mm/h yield losses on the order of 2 dB/km. Obviously, rain rate statistics will vary from region to region, thus affecting average losses.

3.3. Multipath Fading

The term *multipath fading* is used to describe the phenomenon whereby multiple, reflected versions of a transmitted signal (i.e., multipath components) combine at a receiver in either a constructive or destructive fashion depending on their relative amplitudes and phases. When either the transmitter or receiver is in motion, the dynamic, random combining of multipath components leads to received signals that can vary by several tens of decibels with relatively small changes in spatial location. Statistically, multipath fading may be treated as a random process. In the event that no single dominant propagation path exists, the fluctuations in received signal power S are described by the central chi-square distribution [5]:

$$p(S \mid S_0) = \frac{1}{S_0} \exp\left(-\frac{S}{S_0}\right) \tag{10}$$

where S_0 is the mean received signal power, due entirely to multipath. This class of channel is often referred to as a *Rayleigh channel* because the received signal envelope follows the Rayleigh distribution. This statistical model holds well in practice for terrestrial microwave cellular systems, where typically no line-of-sight (LoS) path between transmitter and receiver exists [9]. In LMSC systems, a LoS path often exists in addition to the multipath. In this case, the received signal power, S, is described by the noncentral chi-square distribution [5]

$$p(S \mid A, \sigma_d^2) = \frac{1}{\sigma_d^2} \exp\left\{-\frac{A^2 + 2S}{2\sigma_d^2}\right\} I_0\left(\sqrt{2S}\frac{A}{\sigma_d^2}\right) \tag{11}$$

where A is the amplitude of the LoS component, σ_d^2 is the diffuse signal power, and I_0 is the modified zeroth-order Bessel function of the first kind. This class of channel is often referred to as a Ricean channel because the received signal envelope follows a Ricean distribution. Ricean channels are frequently parameterized by the *Rice factor:*

$$c = \frac{A^2}{2\sigma_d^2} \tag{12}$$

The Rice factor is simply the ratio of the power in the direct and multipath components. When no LoS component exists (i.e., $A = 0$), $c = 0$ and (11) reduces to (10) with the mean received signal power given by $S_0 = \sigma_d^2$. When $c = \infty$, the channel does not exhibit fading. Figure 5 compares

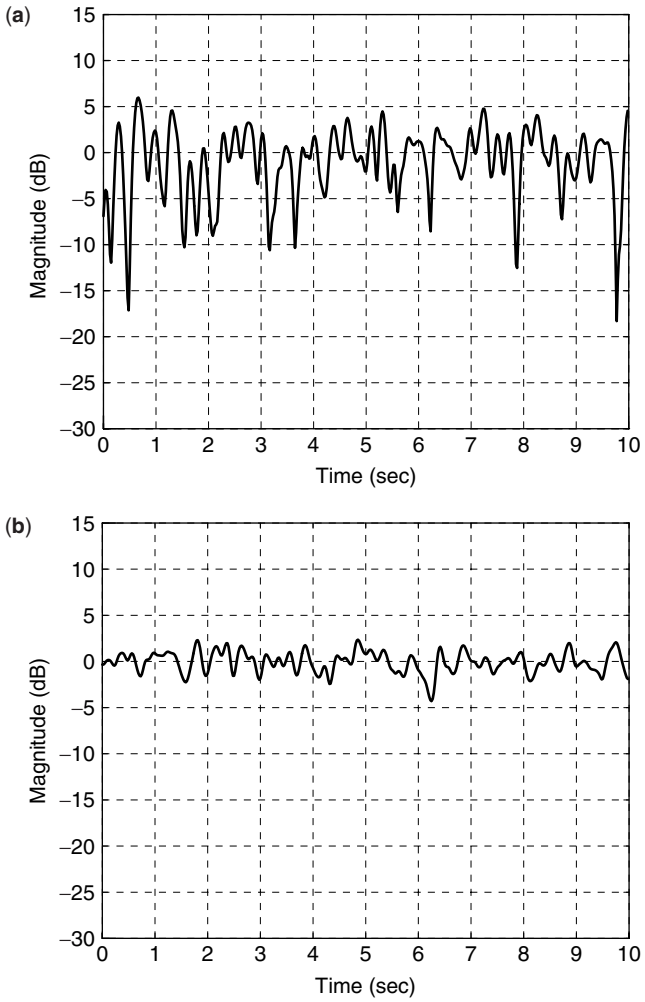


Figure 5. Ricean fading envelopes for (**a**) $c = 10$ dB and (**b**) $c = 20$ dB.

fading signal envelopes for two values of the Rice factor: $c = 10$ dB, and $c = 20$ dB.

In LMSC systems, the Rice factor depends on a number of parameters, including operating frequency, elevation angle, and antenna type. In general, systems that operate at higher frequencies will experience less multipath due to their use of directive antennas and the tendency of shorter wavelengths to scatter off objects in the propagation path. Hence, these systems are typically characterized by larger Rice factors. For example, whereas Rice factors reported for the L-band system studied by Lutz et al. [10] average approximately 10 dB, the Rice factors reported from NASA's Advanced Communications Technology Satellite (ACTS) propagation experiments [11–13], conducted at 20 GHz, average more than 20 dB. Vogel and Goldhirsh [14] examined the multipath fading phenomenon in detail at low elevation angles in unshadowed LOS environments for the INMARSAT LMSC system, which operates at L band. Experimental results show that at elevation angles from 7° to 14°, fades exceeding 7 dB occur for approximately 1% of the driving distance. Moreover, the authors note that the fading is typically dominated by a single multipath reflection from a nearby terrain feature.

The effects of multipath fading on the BER of an LMSC system are quite severe. Because of the variations in received signal strength, the average bit energy : noise ratio $\overline{E_b}/N_0$ must be used in characterizing the average BER performance. Proakis has derived [5] expressions for average BER as a function of $\overline{E_b}/N_0$ for BPSK and DPSK modulations in a Rayleigh fading environment:

$$\overline{\mathrm{BER}_{\mathrm{BPSK}}} = \frac{1}{2}\left(1 - \sqrt{\frac{\overline{E_b}/N_0}{1+\overline{E_b}/N_0}}\right) \tag{13}$$

$$\overline{\mathrm{BER}_{\mathrm{DPSK}}} = \frac{1}{2\left(1+\overline{E_b}/N_0\right)} \tag{14}$$

where $\overline{\mathrm{BER}}$ denotes average BER. Average BER performance in Ricean fading environments was examined in a 1995 article [15]. Figure 6 summarizes these results with BPSK and a couple of different Rice factors. The average BER of BPSK for the Gaussian and Rayleigh channels are also included for reference. Note the significant difference in required $\overline{E_b}/N_0$ necessary to achieve a given BER in the presence of multipath fading. For example, more than 10 dB separates the Gaussian channel and the Ricean channel with $c = 7$ dB at $\overline{\mathrm{BER}} = 1e-4$.

3.4. Shadowing

Signal shadowing is caused when relatively large-scale objects, such as buildings or terrain features, either partially or completely intersect the propagation path. Although difficult to model mathematically, variations in the received signal power S_0, caused by shadowing have been observed to follow a lognormal distribution (i.e., a distribution whose values, when plotted on a log scale, appear Gaussian)

$$p(S_0) = \frac{10}{\sqrt{2\pi}\sigma \ln 10}\frac{1}{S_0}\exp\left[-\frac{(10\log S_0 - \mu)^2}{2\sigma^2}\right] \tag{15}$$

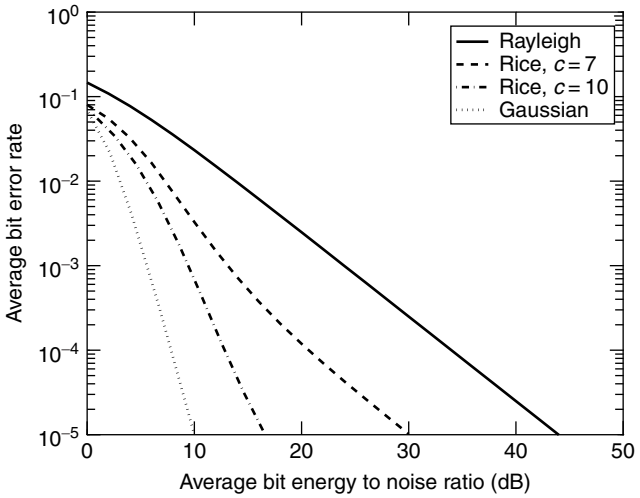


Figure 6. Performance of BPSK in the presence of Ricean and Rayleigh multipath fading.

with a mean μ and standard deviation σ that depend on the carrier frequency and environment [4]. Above X band, the effects of shadowing tend to be more severe, due to the absence of received multipath energy. In these systems the LoS path is critical, and obstruction of this path results in a nearly total loss of received signal power (i.e., signal blockage). Published results from the ACTS mobile propagation experiments [11–13] at 20 GHz report typical means from −15 to −20 dB and standard deviations in the range of 5–10 dB. It is also important to note that path losses due to foliage are significant enough to classify trees as objects that give rise to shadowing in systems that operate above X band. According to certain foliage path loss models [4,16], 5–10 ms of foliage is sufficient to yield losses on the order of 10–15 dB at 20 GHz.

Lutz et al. [10] proposed a total shadowing model (TSM) that effectively combines the densities given by (10), (11), and (15) to describe a LMSC propagation channel at L band. A timeshare parameter $0 \le X \le 1$ was introduced such that a fraction X of the time the received signal power is unaffected by shadowing and described by (11), while the remaining fraction, $(1-X)$, of the time the LoS component is totally blocked (i.e., $A = 0$) and the received signal power follows (10) with the average power S_0, given by the lognormal density in (15). Expressed mathematically. The TSM is given by

$$\begin{aligned} p(S) &= X\,p(S \mid A, \sigma_d^2) + (1-X)\int_0^\infty p(S \mid S_0)\,p(S_0)\,dS_0 \\ &= X\frac{1}{\sigma_d^2}\exp\left\{-\frac{A^2+2S}{2\sigma_d^2}\right\} I_0\left(\sqrt{2S}\frac{A}{\sigma_d^2}\right) \\ &\quad + (1-X)\int_0^\infty \frac{1}{S_0}\exp\left(-\frac{S}{S_0}\right) \\ &\quad \times \frac{10}{\sqrt{2\pi}\sigma \ln 10}\frac{1}{S_0}\exp\left[-\frac{(10\log S_0 - \mu)^2}{2\sigma^2}\right] dS_0 \end{aligned} \tag{16}$$

According to this equation, the fading behavior of the channel consists of two dominant modes or states. In

the unshadowed state (i.e., the "good" channel state) the channel is characterized by the presence of a LoS component, which implies high received power and Ricean fading, while in the shadowed state (i.e., the "bad" channel state) the channel is characterized by the absence of a LoS component, which implies low received power and Rayleigh fading. The timeshare parameter X is a long-term average that describes the fractional amount of time spent in each state. The short-term characteristics of the switching process are accurately described by a two-state Markov model [10]. The situation is depicted in Fig. 7. When the channel is in the good state G, there is a probability p_{GG} associated with remaining in that state and a crossover probability p_{GB} associated with the transition to the bad state B such that $p_{GG} + p_{GB} = 1$. Likewise, there is a probability p_{BB} associated with remaining in the bad state and a probability p_{BG} associated with switching from the bad state to the good state such that $p_{BB} + p_{BG} = 1$. According to the model, the mean duration, in bits, of a good or bad channel state is given by

$$G_b = \frac{1}{p_{GB}}$$
$$B_b = \frac{1}{p_{BG}} \tag{17}$$

and the probability that a good or bad channel state lasts longer than n bits is given by

$$p_G(> n) = p_{GG}^n$$
$$p_B(> n) = p_{BB}^n \tag{18}$$

In addition, the timeshare parameter X can be expressed in terms of the Markov model parameters:

$$X = \frac{G_b}{G_b + B_b} = \frac{p_{BG}}{p_{BG} + p_{GB}} \tag{19}$$

In [10] the Markov model parameters were estimated by fitting the statistics to actual recorded data. The validity of the model described by (16)–(19) for the Ka-band LMSC channel was verified over the course of the ACTS mobile propagation experiments and values for the various model parameters, including X, c, μ, σ, p_{GG}, p_{BB}, G_b, and B_b were reported by Rice [11].

In addition to the TSM, numerous other statistical models have been proposed to describe the LMSC channel. Loo [17] presented a statistical model for L-band LMSC systems. Expressions for level crossing rate and average fade duration are derived and compared to measured data, where reasonably good agreement is observed. Another L-band statistical model, proposed for nongeostationary (i.e., LEO and MEO) LMSC systems has been presented [18]. The model is tunable over a range of environments. Moreover, comparisons to real data are used to derive empirical formulas for the model parameters for several different elevation angles. Finally, a comprehensive statistical model has been proposed [19]. This model is intended to cover a broad range of operating environments and frequencies. In addition, the model can be used to generate time series for LMSC signal features that include amplitude, phase, instantaneous power delay profiles, and Doppler spectra.

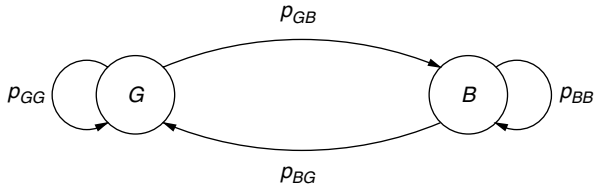


Figure 7. Two-state Markov model that describes the switching process in a LMSC channel.

3.5. Fading Due to Antenna Mispointing

In many LMSC systems directional antennas are typically employed for their high gain. One challenge associated with this practice is maintaining accurate pointing of the receive terminal's directive antenna despite the vehicle dynamics. Regardless of the pointing system used, there will always be residual mispointing error. Characterizing the fluctuations in received signal strength due to antenna mispointing is difficult because of the wide variety of factors that contribute to pointing errors. These factors include terrain, vehicle type and speed, antenna beamwidth, and the antenna controller.

In the ACTS system, mobile propagation experiments were conducted at 20 GHz using the ACTS mobile terminal (AMT). The AMT uses a mechanically steered elliptically shaped reflector antenna with dimensions of approximately 6×2.5 ins. More details on the AMT antenna and tracking system can be found papers by Densmore and others [20,21]. Rice et al. [22] experimentally characterized mispointing error for the AMT. Specifically, measurements of the vehicle pitch, roll, and heading were taken at 0.1-mi intervals along a specific route traveled by the terminal. The error associated with these measurements was used to upper-bound the azimuth and elevation angle mispointing errors at 3.9° and 3.3°, respectively. Finally, through logarithmic interpolation of the antenna gain pattern data, the loss in received signal power due to antenna mispointing was calculated to be on the order of 1.5 dB. These results were then used to rationalize the 1–2-dB variations in received signal power observed during previous AMT runs. Rice and Humphreys [12,13] used data from a wider range of experiments to develop a somewhat ad hoc statistical characterization of antenna mispointing with the AMT. Specifically, a bimodal density function was proposed since this model was observed to fit the experimental data.

A probabilistic analysis of antenna mispointing was presented in an earlier work in which, Gaussian distributed mispointing errors in the azimuth and elevation directions were assumed [23]. This assumption reflects heuristic observations made with pointing systems designed to support Ka-band LMSC systems over rugged terrain. The analysis is of sufficient generality to apply to a range of pointing systems and antenna types since the case where unequal azimuth and elevation mispointing variance are as well as the equal-variance

case examined. Probability density functions (PDFs) for the total mispointing error (i.e., the vector sum of the azimuth and elevation mispointing errors) are given. In addition, the antenna mispointing PDF is used to generate the PDF for received signal loss, from which the average BER is easily computed.

4. ERROR MITIGATION

Consider the following, somewhat simplified, link budget equation between a ground terminal and satellite:

$$\frac{E_b}{N_0} = \text{EIRP} - L + \frac{G}{T} - k_b - R - M \text{ dB} \qquad (20)$$

where $\text{EIRP} = P_t + G_t$ is the effective isotropic radiated power of the transmitter (i.e., the sum, in decibels, of the transmit power and antenna gain) and L is a loss term that accounts for free space and other path losses. The term G/T, where G is the receive antenna gain and T is the effective noise temperature, including the noise temperature of the antenna and the effective noise temperature of the receiver low-noise amplifier (LNA), is often referred to as the *figure of merit* of the receiver. The constant k_b is Boltzmann's constant, $k_b = -228.6$ dBW K^{-1} Hz^{-1} (i.e., decibel watts per degree Kelvin per hertz). The transmission rate, in bits per second (bps), that can be supported by the channel, is given by R. Finally, the link margin, M, is a contingency included to overcome implementation losses and channel effects such as weather, multipath fading, shadowing, and antenna mispointing.

Table 2 illustrates link budget parameters for a typical processing LEO satellite uplink, such as Iridium. In this example, $E_b/N_0 = 8$ dB. Assuming BPSK modulation, Fig. 6 shows that a BER of 2×10^{-4} is achieved for this scenario with full margin. However, in order to achieve the same BER in the presence of Ricean fading where $c = 7$ dB, an additional 10 dB of signal power is required, leaving only 2 dB of link margin available to address other losses. Clearly, any scheme that is capable of reducing the required channel bit energy : noise ratio, $(E_b/N_0)_{\text{req}}$ for a given performance (i.e., BER) level, is of interest. After all, if $(E_b/N_0)_{\text{req}}$ can be reduced, the savings can be applied directly to reduce the transmitter power, antenna gain, or other parameters, or to increase the channel data rate. In this section, two approaches to reducing $(E_b/N_0)_{\text{req}}$ are discussed: error control in the form of FEC coding and ARQ protocols, and diversity combining.

Table 2. LEO Satellite Uplink Budget

Terminal transmit power	$P_t = 1$ W $\equiv 0$ dBW
Terminal antenna gain	$G_t = 2$ dBi
Terminal EIRP	2 dBW
Free-space path loss (2 GHz operating frequency, 800 km orbit altitude)	$L_0 = 166.1$ dB
Additional loss due to 30° satellite elevation (i.e., slant range)	$\Delta L_{sr} = 5.1$ dB
Total budgeted path loss	$L = 171.2$ dB
Gain of satellite receive antenna	26 dBi (edge of coverage)
Antenna noise temperature	290 K
LNA noise temperature	75 K
Effective noise temperature	$T = 365$ K $\equiv 25.6$ dBK
Satellite G/T	$G/T = 0.4$ dB
Boltzmann's constant	$k_b = -228.6$ dBW/K/Hz
Channel rate	9.6 Kbps $\equiv 39.8$ dBHz
Link margin	12 dB

4.1. FEC Coding and ARQ Protocols

The main idea behind power-efficient FEC coding is that the introduction of redundancy into the transmitted bit stream can be exploited by a receiver to correct bit errors caused by channel impairments. The redundancy comes in the form of additional (i.e., parity) bits that are carefully selected by some encoding algorithm and inserted into the transmitted bit stream. With knowledge of the encoding algorithm, the receiver is able to reverse the encoding process, or decode the received sequence. Depending on the sophistication of the encoding/decoding procedures, errors caused by channel noise, fading, or other anomalies can be detected and/or corrected. Note also that as the name implies, FEC coding algorithms operate on the forward link only. No feedback or return link is necessary.

Numerous FEC coding strategies have been developed over the years, including block coding techniques and convolutional codes. Block coding strategies operate on fixed-size blocks of information bits. For each k-bit input block of information bits, the encoding algorithm produces a unique n-symbol output block (i.e., codeword). Note that for a given channel rate, in bps, the *information rate*, that is, the channel capacity devoted to carrying the information bits, is decreased by a factor $R_c = k/n$, called the *code rate*. As with block codes, convolutional codes can be used to generate n-symbol outputs for k-bit inputs yielding a rate $R_c = k/n$ code. However, the primary characteristic that distinguishes convolutional codes from block codes is the fact that convolutional codes have memory. In other words, an n-symbol output block depends not only on the corresponding k-bit input block, but also the m previous input blocks, where m is the memory order of the encoder. Detailed treatments of these approaches are available from a variety of sources [e.g., 5,24]. In recent years, a new and extremely powerful approach to FEC coding, referred to as *Turbo coding* [25], has been introduced. At the heart of Turbo coding schemes are the concepts of code concatenation, soft-decision decoding, and iterative decoding [26]. In general, Turbo coding schemes are more computationally demanding than either block or convolutional codes. Within the context of LMSC systems, convolutional codes remain a popular choice because of their good performance and relatively simple implementation. However, the superior performance advantages of Turbo codes, coupled with advances in decreased-complexity implementations and ever-increasing microprocessor speeds, suggest that Turbo

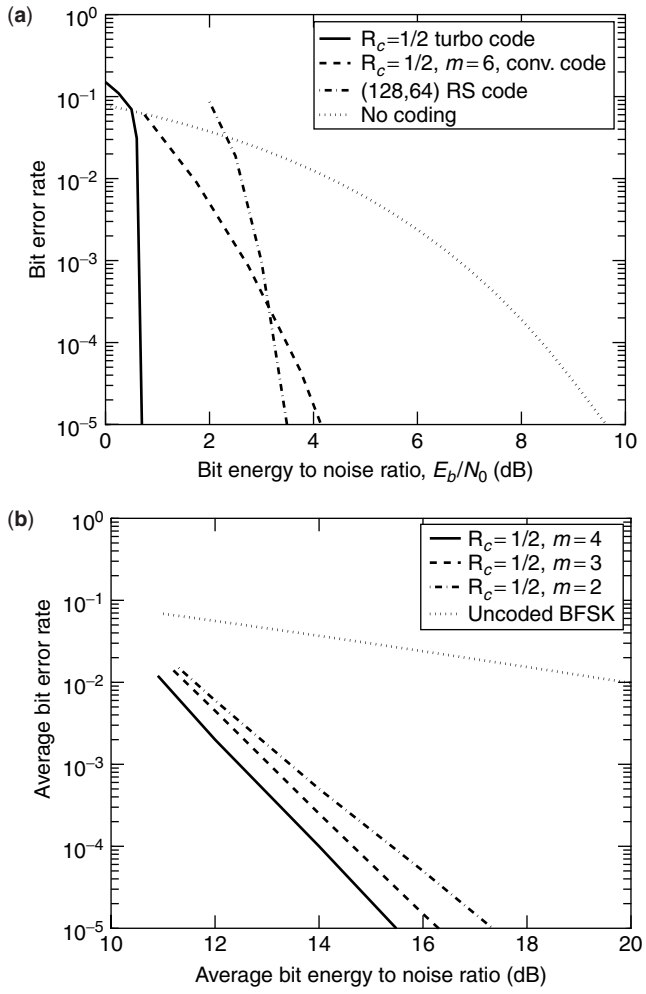


Figure 8. FEC coding performance: (**a**) a $R_c = \frac{1}{2}$ Turbo code, a $R_c = \frac{1}{2}$, $m = 6$ convolutional code (with soft-decision decoding), and a $n = 128$, $k = 64$, (i.e., $R_c = \frac{1}{2}$) Reed–Solomon (RS) block code with BPSK modulation for the Gaussian channel; (**b**) FEC convolutional coding with BFSK modulation for the Rayleigh fading channel; $R_c = \frac{1}{2}$ codes with $m = 2, 3, 4$ are compared.

codes are a logical choice for future systems. Figure 8a illustrates the performance of several different $R_c = \frac{1}{2}$ codes, including the parallel concatenated Turbo code in [25], with BPSK modulation and a Gaussian channel. Proakis [5], examined the performance of convolutional codes with noncoherent FSK modulation for Rayleigh fading channels and derived upper bounds. Figure 8b summarizes these results.

As opposed to FEC coding, ARQ schemes operate by requesting a retransmission of the codeword rather than attempting to correct it at the receiver. Of course, the existence of a feedback path (i.e., a return link) is required with such an approach. Typically, coding is still used in ARQ strategies, but only to alert the receiver to the presence of errors, not to correct them. Since the probability of an undetected error is usually much smaller than that of a decoding error, ARQ schemes are an inherently more reliable form of error control than FEC coding. Hence, these schemes are most often associated with data communications where very low error rates are required. As established previously, FEC coding is appropriate for addressing bit errors introduced by random noise and multipath fading in LMSC systems. On the other hand, ARQ protocols are well suited to situations where long deep fades, such as those caused by shadowing and signal blockage in high-frequency systems, are expected [27]. A thorough description of the main forms of ARQ, including stop and wait, go-back N, and selective repeat is available [24]. It is also possible to combine FEC coding and ARQ schemes into hybrid ARQ (HARQ) protocols that offer performance advantages as well as other desirable attributes such as rate adaptation.

4.2. Diversity Techniques

The basic idea of diversity signaling is that the effects of signal fading can be reduced by supplying the receiver with replicas of the same transmitted signal information over independently faded channels. Through proper selection or combining of these replicas, the likelihood that the receiver experiences a deep fade is reduced considerably. For example, if the probability that any one signal fades below some threshold is p_f, then the probability that D independently faded replicas of the same signal will simultaneously fade below this threshold is given by p_f^D. The optimum strategy for diversity reception is that the D signal replicas be combined coherently in proportion to their received SNR. This type of combining scheme is often referred to as *maximal ratio combining*. A very simple approximation to the average BER of a BPSK signal in a Rayleigh fading environment with maximal ratio combining of D independently fading diversity branches is given by

$$\overline{\mathrm{BER}_{\mathrm{BPSK}}^{mr}} = \left(\frac{1}{4(\overline{E_b}/N_0)_d}\right)^D \binom{2D-1}{D} \tag{21}$$

where $\overline{\mathrm{BER}^{mr}}$ is the average BER for maximal ratio combining and $(\overline{E_b}/N_0)_d$ is the average bit energy : noise ratio received from the D diversity branches:

$$\left(\frac{\overline{E_b}}{N_0}\right)_d = \sum_{i=1}^{D} \left(\frac{\overline{E_b}}{N_0}\right)_i \tag{22}$$

and $\binom{a}{b}$ represents the number of possible combinations of a objects taken b at a time. Figure 9 summarizes the results for several values of D.

There are many ways in which diversity can be introduced into a communications system, most of which have been explored thoroughly within the context of terrestrial cellular systems [9]. These techniques include time diversity, frequency diversity, polarization diversity, and spatial diversity. With spatial diversity, multiple receive antennas are spaced far enough apart so as to yield sufficiently low correlation between their outputs. In environments where significant multipath energy is received from numerous different directions, spacings on the order of $\lambda/4$ are appropriate [9]. In situations where

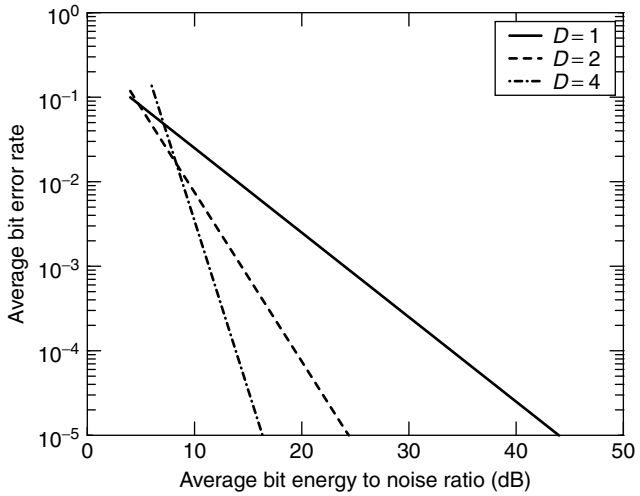


Figure 9. Performance of D-branch maximal ratio diversity combining with BPSK signaling in a Rayleigh multipath fading channel.

received power is confined to a relatively narrow sector in space, such as those where a strong LoS component exists, wider spacings are required. Spatial diversity is especially popular in cellular radio because no modifications to the waveform or transmitter are required. Instead, only $D - 1$ additional antennas, including RF chains, and combining logic are required at the receiver.

An idea similar to spatial diversity has received attention in the LMSC community. The concept is referred to as *satellite diversity*, and is applicable in situations where multiple satellites are potentially within the field of view of a terminal (e.g., LEO and MEO systems). With satellite diversity, terminals communicate with the satellite that provides the best link quality [28], thus improving link availability in the presence of signal shadowing and blockage. In some cases, multiple satellites may transmit the same signal to a user terminal, where these downlinks are combined for improved performance. Obviously, the effectiveness of satellite diversity depends on a variety of factors, including the satellite constellation and the propagation environment. Results from a measurement campaign in Japan [29] suggest that fade margins can be reduced by approximately 10 dB in urban environments using twofold satellite diversity in the Globalstar (i.e., LEO) satellite system. Satellite diversity is an integral part of both the Globalstar and ICO system concepts. The XM satellite radiobroadcast system uses a GEO constellation but achieves the same effect as satellite diversity through the use of terrestrial repeater stations located in heavily shadowed environments such as dense urban areas.

5. MULTIPLE ACCESS

Multiple access is concerned with ways in which a group of users share a common communications resource in an efficient manner. With respect to LMSC systems, the common communications resource is the satellite. Because the satellite resource is limited, and because upgrading this resource is difficult once the satellite is in orbit, efficient multiple-access schemes represent a critical component to LMSC systems. Satellite multiple-access schemes fall into one of four categories [30]: fixed-assignment techniques, random-access methods, demand assignment protocols, and adaptive assignment protocols.

Fixed assignment schemes are most appropriate in situations where users' needs are such that resources should be dedicated for a relatively long period of time (at least long enough to justify the overhead associated with allocating and deallocating the resources), such as voice circuits in a satellite telephony system. Frequency division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access (CDMA) are the basic forms of fixed assignment. With these schemes, the satellite resource is partitioned into orthogonal, or quasiorthogonal segments, referred to as *channels*. In FDMA, the channels are fixed slices of bandwidth, to which users are granted exclusive use for the duration of their call, or session. In TDMA, channels are created through the use of a framing structure that divides the resource into time slots. Time slots are then assigned to users whereby they are allowed access to the entire bandwidth for the duration of their slot. Time slots are typically quite short but repeat on a regular basis so that from the users' perspective a constant data rate is achieved. Finally, with CDMA, channels are associated with special periodic spreading codes, which are used to modulate the users' bit streams, resulting in bandwidth expansion. Hence, with CDMA, users overlap one another in both frequency and time. However, provided the codes are orthogonal, they are distinguished at the receiver by correlating with the desired user's code. Figure 10 illustrates the three concepts. FDMA and TDMA have been used in LMSC systems for quite some time. However, CDMA is gaining popularity [31] and is currently used in the Globalstar system.

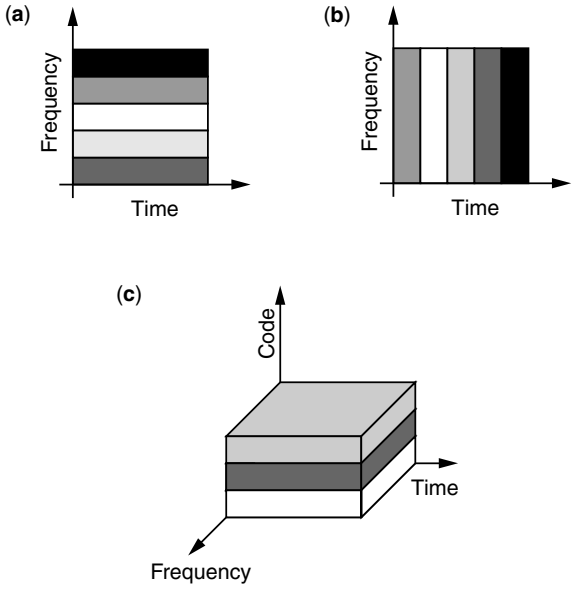


Figure 10. Fixed assignment multiple access schemes: (**a**) FDMA; (**b**) TDMA; (**c**) CDMA.

Random access schemes are appropriate in situations where user traffic is packetized and bursty. With random access, users contend for the common communications resource. A user who has a packet to send simply sends it. The user then monitors the satellite downlink to determine whether the packet was correctly received, and hence forwarded, by the satellite. Or, an acknowledgment scheme may be used whereby the receiving terminal notifies the sender via an acknowledgment message (ACK). If the packet is not heard on the downlink, or the ACK is not received, a collision with another user's packet is assumed and the packet is resent after a random delay. ALOHA and slotted ALOHA are examples of random-access protocols that operate in this manner [32]. The throughput, defined as the expected number of successful transmissions per unit time of ALOHA, is approximately 18.4%. The slotted time structure imposed by slotted ALOHA reduces the likelihood of a collision and effectively doubles the throughput to 36.8%. The obvious drawback to random-access schemes is that as traffic loading increases, so does the probability of collision, which negatively impacts performance.

Demand assignment protocols typically consist of two phases: a *reservation phase*, where resources are requested by users according to their needs; and a *communications phase*, where these resources are actually used. Reservations are typically carried out on a separate channel via random access, with the logic that reservation packets are typically short and less frequent than data messages. Packet reservation multiple access [33] is an example of a demand assignment protocol. These schemes are most appropriate when users have, on average, moderate communications requirements (e.g., occasional large file transfers). Of course, the penalty associated with demand assignment schemes is the overhead and latency (an extra round-trip delay through the LMSC system) associated with the reservation request.

Adaptive assignment protocols use a variety of means to dynamically adjust to the traffic type and load. In most cases, these schemes represent a hybrid, or superset, of other multiple access schemes. For example, the approach in [34], referred to as the "urn" scheme because of the analogy of drawing balls from an urn used in its development, adapts smoothly from the slotted ALOHA random access scheme in lightly loaded conditions to the TDMA fixed assignment scheme in heavily loaded conditions. In priority oriented demand assignment (PODA) [35], a framing structure is imposed whereby frames are subdivided into reservation and information subframes. Reservation subframes are used to make explicit reservations as in a typical demand assignment scheme. However, the protocol also supports implicit reservations whereby reservations for subsequent packets may be piggybacked onto transmitted information packets, thus improving performance for streaming traffic. In general, adaptive assignment approaches will work best when users' communications requirements are mixed, since the schemes will adapt as necessary. However, adaptive assignment protocols are also more complex relative to other multiple-access schemes.

6. NETWORK ASPECTS

Networking in LMSC systems is a relatively broad topic that has at least two major components: issues relating to LMSC cellular voice networks and those related to the internetworking of LMSC and terrestrial data networks. Only a few points will be made in regard to LMSC voice networks since there is a great deal of similarity between these systems and terrestrial cellular voice networks. The discussion on internetworking focuses exclusively on the problem of improving the performance of the standard TCP/IP protocol suite in data networks that contain both terrestrial and LMSC links.

In most respects, network procedures in LMSC cellular voice systems are quite similar to those in terrestrial cellular networks. Of course, these similarities are largely by design to promote interoperability among terrestrial and LMSC networks. In cases where distinctions occur, they are due mostly to differences in the topology of LMSC systems compared to terrestrial cellular systems, and/or the fact that satellites in LMSC systems might move relative to a fixed point on earth. For example, in terrestrial cellular systems, mobile users communicate directly with a base station that is responsible for serving a particular cell. This affiliation facilitates a number of network procedures, including those associated with mobility management (e.g., location registration and handover of live calls from one base station to another), and those associated with resource management (e.g., call setup and teardown, channel allocation, and paging). Because in terrestrial systems there is generally a one to one correspondence between a base station and a cell, the affiliation process is relatively straightforward. As users roam from cell to cell, they associate with the appropriate base station as determined by the quality of a received broadcast transmission that emanates from the various base stations in a service region. In LMSC systems, fixed earth stations, referred to as "gateways," act as the interface to the terrestrial wired infrastructure, and are similar to base stations in this regard. However, one important distinction between base stations and gateways is that whereas in terrestrial cellular systems users communicate with the base station directly, in LMSC systems users reach a gateway by communicating through a satellite link (see Fig. 1). This difference in network topology complicates the relationship between a gateway and the users it serves in a particular area, especially in LEO and MEO systems, where the satellites move relative to the gateways. For example, consider a LEO or MEO system that does not employ ISLs. In this case, the service area of a gateway, defined as the area in which both the mobile user and gateway have a simultaneous view of the same satellite, actually changes over time as the satellites move relative to the earth. In systems where ISLs are used, the service area of a gateway is completely arbitrary. In fact, it is theoretically possible that a single gateway could be used in these systems to serve all users [2].

Just as interoperability among terrestrial and LMSC voice networks is desirable, the same is true with data networks. Internetworking of disparate terrestrial

data networks is often achieved in part through the use of the TCP/IP protocol suite [32], where IP is responsible primarily for message routing and TCP includes mechanisms for flow control and error recovery. Unfortunately, the use of TCP in wireless and satellite networks generally results in suboptimal performance. The primary reason for poor TCP performance in these situations is that TCP flow control and error recovery algorithms were originally designed for use in wired systems where BERs are relatively low, latencies are short and forward and return paths are generally symmetric with respect to data rate. On the other hand, networks that include satellite links will exhibit higher BERs, longer latencies, and possibly asymmetric data rates on the forward and return paths. These differences in link conditions adversely affect the performance of TCP and result in suboptimal performance [36]. For example, because it was designed for use in terrestrial wired networks where the BERs are usually low, the TCP flow control algorithm attributes packet loss to congestion and reduces link utilization to accommodate the situation. In LMSC systems, where packets are often lost because of bit errors as opposed to congestion, the resulting reduction in utilization is unwarranted and represents an inefficiency. The relatively long round-trip time in LMSC systems adversely impacts both TCP flow control and error recovery, especially when high data rates are used (i.e., the system has a large bandwidth–delay product). With respect to flow control, TCP uses an algorithm known as "slow start," whereby the packet transmission rate is increased gradually on the basis of received acknowledgments. In large bandwidth–delay product environments, slow start will take a relatively long time to achieve full link utilization. For error recovery, TCP uses the go-back-*N* ARQ strategy, also known to perform poorly in large bandwidth–delay product environments. Finally, asymmetric links may result in poor performance if the low-rate return path becomes congested with acknowledgments before the high-rate forward path is fully utilized.

There are several ways in which the problem of poor TCP performance in LMSC and other wireless systems may be addressed. One solution is to simply use an alternate protocol, one optimized for the satellite environment. Several such protocols have been proposed, including the Satellite Transport Protocol (STP) [37] and the Wireless Transmission Control Protocol (WTCP) [38]. Another approach is to modify or extend TCP to improve its performance over satellite links. Efforts in this area include the extensions proposed by Jacobson et al. [39], the selective acknowledgments (SACK) extension proposed by Mathis et al. [40], and the space communications protocol standards (SCPS) [41]. One serious drawback to TCP replacement and extensions is that in order to achieve the performance gains, all participating hosts must comply with the modification, thus negating the value of TCP's high penetration as a standard transport-layer protocol. Another approach is to employ TCP splitting. The idea behind TCP splitting is that the end-to-end TCP connection between two users is split into three segments. The first segment consists of a (presumably wired) TCP connection between the first user and a satellite terminal, the second segment consists of the satellite link, over which a protocol optimized for this environment is used, and the third link (also assumed to be wired) is a TCP connection between the receiving terminal and the second user. The splitting is done in such a way that it is transparent to the end users. In other words, from their perspective, an end-to-end TCP connection exists between them. However, because TCP is not actually used over the satellite link, the inefficiencies associated with this practice are not experienced. Of course, additional complexity must be introduced to perform the splitting, but this complexity is confined to the satellite terminals, and no modifications to end-user equipment (i.e., host computers) are required. Stadler et al. [36] and Mineweaser et al. [42] proposed and evaluated a TCP splitting mechanism, referred to as the *wireless IP suite enhancer* (WISE). In Fig. 11 the performance of WISE as a function of round-trip time is contrasted with that of TCP and TCP with the SACK extensions. In the figure, the transfer time of a 256-kilobyte (KB) file as a function of end-to-end round trip time is presented as an average over 20 simulation runs. The channel rate was taken to be 128 kbps and the average BER was 10^{-5}. In the WISE simulations, the protocol used over the satellite portion of the link is the Lincoln Laboratory Link Layer (LLLL) protocol [43]. Because this protocol resides at the link layer, it is also possible to use it directly with TCP (i.e., no splitting) to achieve some performance gain. With this approach, LLLL conditions the underlying satellite link according to the requirements of TCP. The performance of this scheme is also characterized in the figure. Note that WISE delivers nearly uniform performance regardless of round-trip time, while the other approaches are more sensitive to latency. Similar results can be achieved as a function of BER.

7. CONCLUSIONS

A brief description of several LMSC systems, their characteristics, and technical issues surrounding their

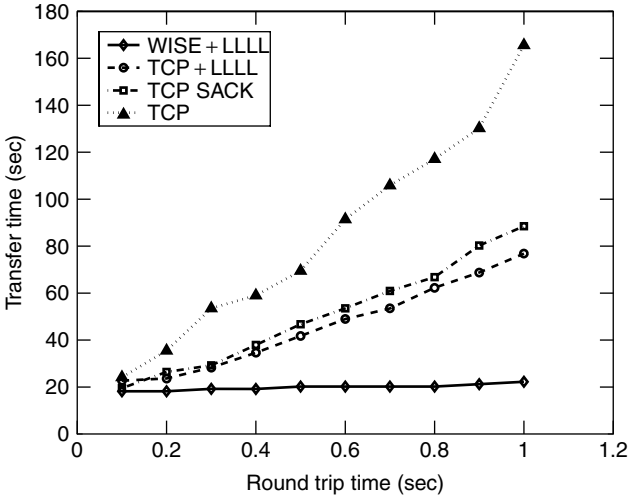


Figure 11. Average file transfer time as a function of round-trip time for several protocols.

implementation and operation has been presented. Major topics of discussion include the following. The LMSC propagation channel, including the effects of random noise, path loss, multipath fading, shadowing, attenuation due to antenna mispointing, was investigated. Also, error mitigation strategies such as FEC coding, ARQ protocols, and diversity techniques were examined. Multiple access schemes, including fixed assignment, random access, demand assignment, and adaptive assignment protocols, were also discussed. Finally, network aspects were covered, including a brief discussion of the similarities and differences in LMSC and terrestrial cellular voice networks, and an examination of the performance of TCP in networks that include LMSC links. Where appropriate, references were cited so that interested readers can pursue these topics in further detail.

BIOGRAPHY

Jeff Schodorf received his B.S.E.E., M.S.E.E., and Ph.D. in 1991, 1994, and 1996, respectively, all from the Georgia Institute of Technology. Since 1996 he has been a member of the Technical Staff in the Tactical Communication Systems Group at MIT Lincoln Laboratory. His research at the Laboratory has covered a variety of topics, including reduced complexity demodulation and decoding, satellite multiple access, and channel models and error control protocols for the land mobile satellite channel.

BIBLIOGRAPHY

1. T. T. Ha, *Digital Satellite Communications*, McGraw Hill, New York, 1990.
2. E. Lutz, M. Werner, and A. Jahn, *Satellite Systems for Personal and Broadband Communications*, Springer, 2000.
3. R. E. Sheriff and Y. F. Hu, *Mobile Satellite Communication Networks*, Wiley, New York, 2001.
4. D. Parsons, *The Mobile Radio Propagation Channel*, Halsted Press, New York, 1992.
5. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1989.
6. F. Xiong, *Digital Modulation Techniqes*, Artech House, Boston, 2000.
7. P. K. Karmakar et al., Radiometric measurements of rain attenuation at 22.2 and 31.4 GHz over Calcutta, *Int. J. Infrared and Millimeters Waves* 493–501 (1998).
8. S. Poonam and T. K. Bandopadhyaya, Rain rate statistics and fade distribution of millimeter waves in Indian continents, *Int. J. Infrared and Millimeters Waves* 503–509 (1998).
9. W. C. Jakes, *Microwave Mobile Communications*, IEEE Press, Piscataway, NJ, 1993.
10. E. Lutz et al., The land mobile satellite communications channel — recording, statistics and channel model, *IEEE Trans. Vehic. Technol.* 375–386 (May 1991).
11. M. Rice et al., K-band land-mobile satellite characterization using ACTS, *Int. J. Sat. Commun.* 283–296 (Jan. 1996).
12. M. Rice and B. Humphreys, Statistical models for the ACTS K-band land mobile satellite channel, *Proc. IEEE Vehicular Technology Conf.*, 1997.
13. M. Rice and B. Humphreys, A new model for the ACTS land mobile satellite channel, *Proc. Int. Mobile Satellite Conf.*, 1997.
14. W. J. Vogel and J. Goldhirsh, Multipath fading at L band for low elevation angle, land mobile satellite scenarios, *IEEE J. Select. Areas Commun.* 197–204 (Feb. 1995).
15. M. G. Shayesteh and A. Aghamohammadi, On the error probability of linearly modulated signals on frequency-flat Ricean, Rayleigh, and AWGN channels, *IEEE Trans. Commun.* 1454–1466 (Feb. 1995).
16. A. J. Simmons, *EHF Propagation through Foliage*, MIT Lincoln Laboratory Technical Report TR-594 1981.
17. C. Loo, A statistical model for a land mobile satellite link, *IEEE Trans. Vehic. Technol.* 122–127 (Aug. 1985).
18. G. E. Corazzo and F. Vatalaro, A statistical model for land mobile satellite channels and its application to nongeostationary orbit systems, *IEEE Trans. Vehic. Technol.* 738–742 (Aug. 1994).
19. F. P. Fontan et al., Statistical modeling of the LMS channel, *IEEE Trans. Vehic. Technol.* 1549–1567 (Nov. 2001).
20. A. C. Densmore and V. Jamnejad, A satellite tracking K- and Ka-band mobile vehicle antenna system, *IEEE Trans. Vehic. Technol.* 502–513 (Nov. 1993).
21. A. Densmore et al., K- and Ka- band land mobile satellite-tracking reflector antenna system for the NASA ACTS mobile terminal, *Proc. Int. Mobile Satellite Conf.*, 1993.
22. M. Rice, B. J. Mott, and K. D. Wise, A pointing error analysis of the ACTS mobile terminal, *Proc. Int. Mobile Satellite Conf.*, 1997.
23. J. B. Schodorf, A probabilistic mispointing analysis for land mobile satellite communications systems with directive antennas, *Proc. IEEE Vehicular Technology Conf.*, 2001.
24. S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
25. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error correcting coding and decoding: Turbo codes, *Proc. IEEE Int. Conf. Communication*, 1993.
26. B. Sklar, A primer on turbo code concepts, *IEEE Commun. Mag.* 94–102 (Dec. 1997).
27. J. B. Schodorf, Error control for Ka-band land mobile satellite communications systems, *Proc. IEEE Vehicular Technology Conf.*, 2000.
28. J. Schindall, Concept and implementation of the Globalstar mobile satellite system, *Proc. Intl. Mobile Satellite Conf.*, 1995.
29. R. Akturan and W. J. Vogel, Path diversity for LEO satellite-PCS in the urban environment, *IEEE Trans. Antennas Propag.* 1107–1116 (July 1997).
30. T. Nguyen and T. Suda, Survey and evaluation of multiple access protocols in multimedia satellite networks, *Proc. Southeastcon*, 1990, 408–412.
31. A. J. Viterbi, A perspective on the evolution of multiple access satellite communication, *IEEE J. Select. Areas Commun.* 980–983 (Aug. 1992).
32. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1992.
33. L. G. Roberts, Dynamic allocation of satellite capacity through packet reservation, *Proc. Nat. Computer Conf., AFIPS Conf.*, 1973, pp. 711–716.

34. L. Kleinrock and Y. Yemini, An optimal adaptive scheme for multiple access broadcast communication, *Proc. IEEE Int. Conf. Communication*, 1978, pp. 7.2.1–7.2.5.

35. I. M. Jacobs, R. Binder, and E. V. Hoversten, General purpose packet satellite networks, *Proc. IEEE* 1448–1467 (Nov. 1978).

36. J. S. Stadler, J. Gelman, and J. Howard, Performance enhancements for TCP/IP on wireless links, *Proc. Virginia Tech/MPRG Symp. Wireless Personal Communications*, 1999.

37. T. R. Henderson and R. H. Katz, Transport protocol for internet-compatible satellite networks, *IEEE J. Select. Areas Commun.* 326–344 (Feb. 1999).

38. P. Sinha et al., WTCP: A reliable transport protocol for wireless wide-area networks, *Wireless Networks* 301–316 (2002).

39. V. Jacobson, R. Braden, and D. Borman, *TCP Extensions for High Performance*, IETF, RFC 1323, May 1992.

40. M. Mathis et al., TCP selective acknowledgment options, IETF, RFC 2018, Oct. 1996.

41. R. C. Durst, G. J. Miller, and E. J. Travis, TCP extensions for space communications, *Wireless Networks* 389–403 (1997).

42. J. L. Mineweaser et al., Improving TCP/IP performance for the land mobile satellite channel, *Proc. IEEE Military Communication Conf.*, 2001.

43. J. S. Stadler, A link layer protocol for efficient transmission of TCP/IP via satellite, *Proc. IEEE Military Communication Conf.*, 1997.

LEAKY-WAVE ANTENNAS

Fabrizio Frezza
Alessandro Galli
Paolo Lampariello
"La Sapienza" University of Rome
Roma, Italy

1. INTRODUCTION

1.1. Definition

Leaky-wave antennas (LWAs) constitute a type of radiators whose behavior can be described by an electromagnetic wave ("leaky wave") that propagates in guiding structures that do not completely confine the field, thus allowing a continuous loss of power towards the external environment ("leakage").

According to the IEEE Standard 145-1993, a leaky-wave antenna is "an antenna that couples power in small increments per unit length either continuously or discretely, from a traveling wave structure to free space."

1.2. General Properties and Applications

LWAs [1] belong to the class of traveling-wave line antennas, for which the illumination is produced by a wave that propagates along a guiding structure [2]. If compared with the wavelength, a LWA is "long" in the propagation direction z, while its cross section is usually

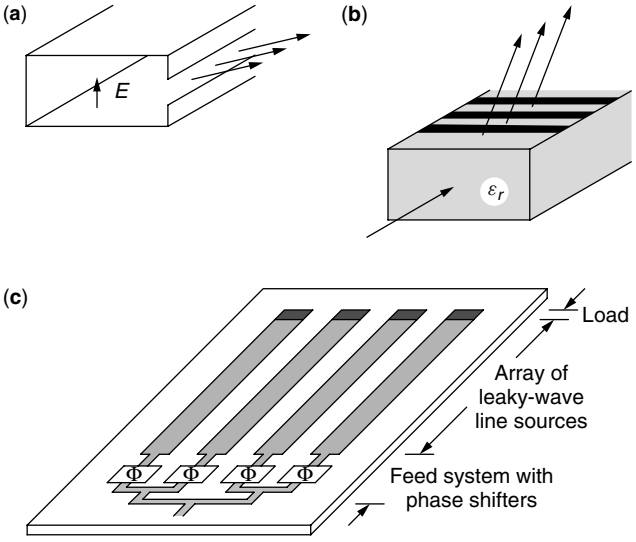


Figure 1. Basic structures of leaky-wave antennas (LWAs): (**a**) uniform LWAs: — geometry derivable by a partially open metallic waveguide; (**b**) periodic LWAs: — geometry derivable by a strip-loaded dielectric-rod waveguide; (**c**) topology of LWA arrays.

of the order of the wavelength (see the reference examples of Fig. 1a,b).

LWAs radiate along their lengths and in general are excited from one input of the open guiding structure with a traveling wave that propagates mainly in one longitudinal direction (e.g., $+z$) and is attenuated as a result of the power leakage toward the exterior region, thus leaving a negligible field at the end termination of the guide. In a harmonic regime (with an $\exp(j\omega t)$ time dependence), this wave is characterized by a complex propagation constant of the type $k_z = \beta_z - j\alpha_z$ [3,4], where β_z is the "phase constant" and α_z is the "attenuation constant" of the leaky wave (when only power loss due to radiation is taken into account, α_z is also said "leakage constant").

Usually the radiation pattern of a single LWA has a typical "fan" shape; in the elevation (or zenith) plane a narrow beam is achievable with the pointing direction that varies by frequency, while in the cross (or azimuth) plane the beam is usually wider in connection with the characteristics of a more reduced transverse aperture. Depending on the desired application, a suitable longitudinal variation of the aperture distribution, usually reached by modulating geometric parameters ("tapering"), allows a good control of the radiation pattern (sidelobe behavior, etc.). In some cases, in order to obtain a beam shaping or a physical matching with mounting curved surfaces, LWAs can be designed with certain amounts of curvature along their lengths [5].

The scanning properties in the elevation plane (pointing angle variable with the frequency) are related to the type of waveguide employed, which can be of either "uniform" (Fig. 1a) or "periodic" type (Fig. 1b) [1,2]. LWAs derived by waveguides that are longitudinally uniform (i.e., the structure maintains continuously the same transverse geometry) typically allow the angular scanning in one quadrant, from around broadside toward one endfire

(the "forward" one, which is concordant with the wave propagation direction). LWAs derived by waveguides that are longitudinally periodic (i.e., where the structure is periodically loaded with proper discontinuities, at intervals that are usually short with respect to the wavelength) allow a wider angular scanning in both the forward and backward quadrants. However, due to different causes, limitations in such scanning ranges generally exist for both the types of structures. A scan range in both quadrants may also be accomplished by using anisotropic media.

When a "pencil beam" is aimed with a possible two-dimensional (2D) scanning in both elevation and cross-planes (zenith and azimuth), a phased array of juxtaposed LWAs may be employed, thus enlarging the equivalent aperture also transversely [6,7] (Fig. 1c). LWA arrays are therefore constituted by a linear configuration of sources (i.e., one-dimensional elements), instead of the planar ones of standard arrays (i.e., two-dimensional elements). For LWA arrays a pointed-beam scanning is achievable by varying both the frequency for the elevation plane and the phase shift for the cross-plane.

Since LWAs are derived by partially open waveguides, they present a number of distinctive features as radiators: handling of high-power amounts, particularly for structures derivable by closed metallic waveguides; reduction of bulk problems, due to the usually small profiles in the cross sections; capability of designing a wide variety of aperture distributions and consequent flexibility for the beamshaping; possible use as wideband radiators, allowing large angular scanning by varying frequency (instead of using mechanical or other electronic means); achievement of very narrow beams with good polarization purity; and simplicity of feeding and economy for 2D scannable pencil-beam arrays (reduced number of phase shifters).

LWAs are used mainly in the microwave and millimeter wave regions. The first studies on LWAs were presented during the 1940s, basically for aerospace applications (radar, etc.); since then, a very wide number of different solutions for LWAs has been proposed in connection with changing requirements and constraints. Also the applicability of this type of antennas has been widened, involving various problems of traffic control, remote sensing, wireless communications, and so forth [8].

2. PRINCIPLES OF OPERATION

2.1. Leaky Waves in Open Structures

A leaky wave [3,4] has a complex longitudinal wavenumber k_z that can be derived by solving, as a function of the physical parameters (frequency and geometry of an open waveguiding structure), the characteristic equation (or dispersion relation), which is of the general type:

$$D(k_z, k_0) = 0 \tag{1}$$

where $k_0 = \omega(\mu_0\varepsilon_0)^{1/2}$ is the vacuum wavenumber.

As is well known, for lossless closed waveguides the dispersion relation (1) generally presents an infinite discrete set of eigensolutions giving the "guided modes," which individually satisfy all the relevant boundary conditions. Any field excited by a source in a closed guide can be expanded in terms of the complete set of the infinite discrete eigensolutions derived by Eq. (1). In conventional guides, the longitudinal wavenumbers k_z are either real [propagating waves above their cutoff, with $k_z = \beta_z < k = k_0(\varepsilon_r)^{1/2}$] or imaginary (attenuating waves below their cutoff, with $k_z = -j\alpha_z$).

In lossless open waveguides (e.g., dielectric guides), instead, only a finite number of propagating modes can exist as eigensolutions of Eq. (1) satisfying all the boundary conditions (particularly the radiation condition); these are the so-called bound "surface waves" (each one exists only above its cutoff, with $k_z = \beta_z > k_0$). In addition to this, for a complete representation of the field that is no longer confined to a closed section, a "continuous spectrum" of modes must be introduced to describe the radiated field as an integral contribution in terms of a set of plane waves having a continuous range of wavenumbers (e.g., such that $0 < k_z = \beta_z < k_0$ and $-j\infty < k_z = -j\alpha_z < 0$). Any field excited by a source in an open guide can therefore be expanded by means of a "spectral representation," that is, in terms of a finite set of guided modes and an integral contribution of the continuous spectrum.

On the other side, it is seen that the characteristic equation (1) for open guides presents additional discrete solutions that are "nonspectral," since they correspond to fields that violate the radiation condition (they attenuate along the propagation direction but exponentially increase in a transverse direction away from the structure) and are not included in the spectral representation of the field. In an open lossless structure the leaky-wave solutions that are of the type $k_z = \beta_z - j\alpha_z$ describe power flowing away from the structure.

In many practical circumstances, for describing the radiative effects of the open structures in the presence of a source, the evaluation of the field through the "spectral representation" (i.e., including the integral contribution of the continuous spectrum) can be very difficult and cumbersome to quantify. It is seen that the radiation field can be evaluated accurately in much a simpler fashion by considering just the contribution due to the presence of one complex mode, that is, a leaky wave, which can therefore be viewed as a simple rephrasing of the continuous spectrum. In fact, it is seen that in practical cases the remaining part of the continuous spectrum (viz., the "space wave" or "residual wave") is able to give negligible contributions to the description of the LWA's radiation.

It can be seen that, when properly excited by a source at a finite section, a leaky wave, even though nonspectral, assumes its physical validity within an angular sector close to the equivalent aperture of the open guiding structure, and the relevant field distribution is able to furnish a fundamental contribution to evaluation of the near field. Since the relevant far field is achieved by a simple Fourier transform of the field on the aperture, a leaky wave can definitively furnish a highly convergent and efficient quantification of the radiation of LWAs, as an extremely advantageous alternative to a continuous spectrum evaluation.

2.2. Characterization of Leaky-Wave Antennas

LWAs present the advantage of a rather simple characterization of their basic properties, with consequent straightforward approaches for the analysis and synthesis. The basic knowledge is reduced to the evaluation of a dominant complex eigensolution $k_z = \beta_z - j\alpha_z$ that can be supported and strongly excited in a specific open structure.

The characteristic behaviors of the real and imaginary parts of the longitudinal wavenumber of a leaky wave are presented in Fig. 2; specifically the dispersion behaviors of the normalized parameters β_z/k_0 and α_z/k_0 versus frequency f. The radiation region of LW structures lies approximately inside the frequency range where the wave becomes fast ($\beta_z/k_0 < 1$) and power can therefore leak out from the guiding structure toward the outside air region in the typical form of a TEM-like mode; in fact, $\beta_z/k_0 < 1$ is in general the so-called condition for leakage of a complex wave that can radiate in an external air region.

The valid frequency range for LWA applications is actually where, as the frequency decreases, β_z/k_0 diminishes monotonically from unity toward rather low values; in this region, to have an efficient directive beam, α_z/k_0 should assume rather limited values (e.g., typically α_z/k_0 varies from about 10^{-1} to 10^{-3}). As frequency decreases further, a sudden rise of α_z/k_0 is generally found, which describes the predominance of reactive phenomena instead of radiative ones, while β_z/k_0 can present a flat zone with approximately constant low values before showing a further steep rise as frequency goes to zero: in these ranges, radiative effects can no longer be represented by the leaky wave and the structures usually cannot work well as radiators [3,4,7].

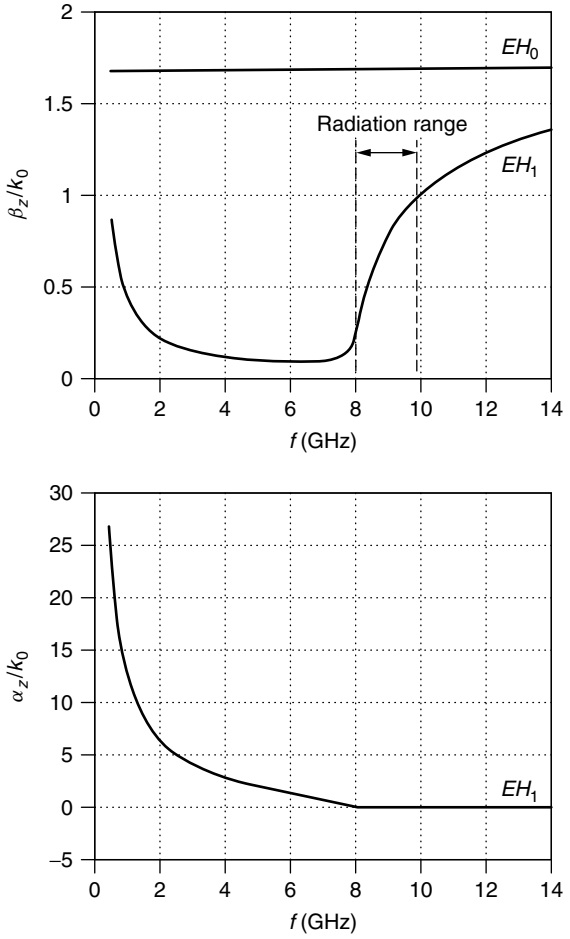


Figure 2. Typical dispersion behavior of the leaky-mode complex wavenumber (normalized phase β_z/k_0 and leakage α_z/k_0 constants vs. frequency f) for an open planar waveguide (microstrip).

It is worth noting here that in open planar structures a different type of leakage can occur as well, which is associable to "surface waves" (that are TE- or TM-like modes) propagating in the substrates [7], instead of the standard "space wave" (TEM-like mode) that carries out power in the outside air region; while the latter is able to account for useful contributions to far-field radiation in LWA's applications, the former usually describes power that leaks out transversely in a layered structure and accounts for loss and interference effects in the planar circuits.

2.3. Evaluation of the Leaky-Wave Phase and Leakage Constants

The evaluation of the complex eigensolutions for nonclosed waveguides depends on the physical parameters involved (frequency and geometry) and is generally achievable with numerical methods. Among them, the *transverse resonance technique* (TRT) [9,10] is one of the most efficient approaches for either rigorous or approximate (according to the antenna topology) evaluations. It first requires the introduction of a suitable equivalent transmission-line network, which describes the transverse geometry of the structure. Then, a numerically solvable transcendental equation in terms of transverse eigenvalues k_t and of physical parameters is usually achievable by imposing a resonance condition for the equivalent circuit. The complex eigenvalue k_z is derived by the additional link to the longitudinal problem given by a separation condition for the eigenvalues (e.g., in air: $k_0^2 = \omega^2\mu_0\varepsilon_0 = k_t^2 + k_z^2$). Where the separation condition holds rigorously also for the variables in the transverse plane (e.g., $k_t^2 = k_x^2 + k_y^2$), TRT in general gives exactly the characteristic equation of the geometry. An example is given in the next paragraph.

When separation of variables does not strictly hold, other numerical methods can nevertheless be employed to accurately determine the complex eigensolutions of the involved open waveguides. The most appropriate choice depends on several factors related to the computational features of the methods, the geometry of the open-type structures, and so on [9,10]. Among the various possible approaches, integral equation techniques can work particularly well. As is known, in particular spectral domain approaches appear well suited for the derivation of the eigensolutions in structures of printed type [9].

2.4. Interpretation of the Behavior of a Leaky-Wave Antenna

As stated above, LWAs are described by a fast wave that propagates on an equivalent aperture losing power

toward free space with a leakage amount that is usually rather limited to allow a sufficiently directive beam. The simplest LWA geometry for this purpose is derivable by a closed metallic waveguide in which a suitable "small" aperture is introduced longitudinally in order to get a continuous power loss along its length, as shown in Fig. 3a for a rectangular guide with a slit cut on a sidewall. This structure, besides having a historical importance as the first proposed LWA in 1940 [1,2], can be taken as a reference structure for explaining the basic behavior of LWA's in terms of a waveguide description.

For such a structure, a leaky wave can be considered as excited by a standard incident mode for the closed rectangular waveguide, that is the dominant TE_{10}, which travels in the $+z$ direction with a known phase constant β_{0z} for a fixed choice of the physical parameters (geometry and frequency). For a sufficiently small geometry perturbation due to the slit, the phase constant is changed just slightly to a value represented by β_z, and a "low" leakage rate

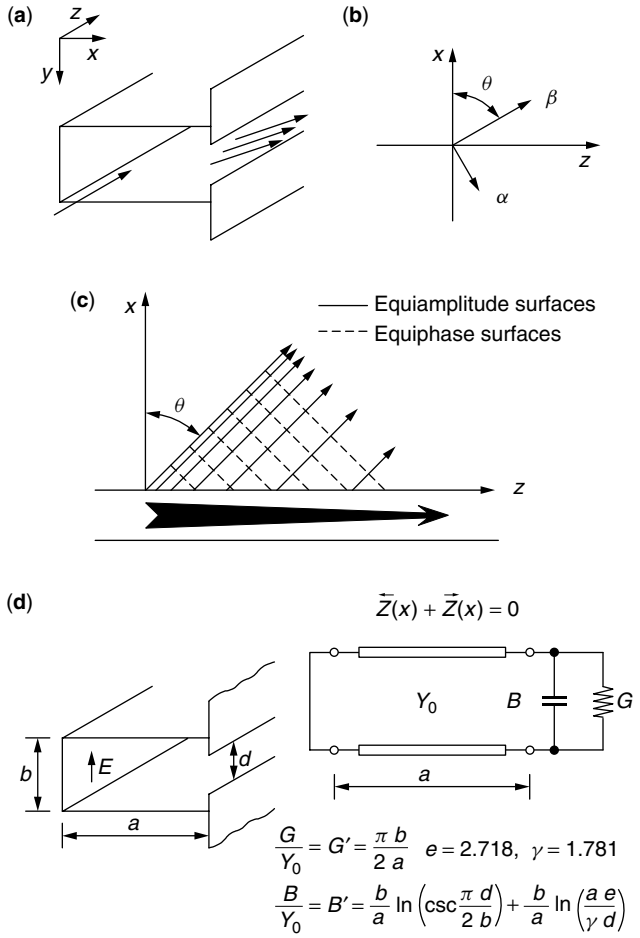


Figure 3. LWA derived by a sidewall slit rectangular waveguide: (**a**) geometry of the structure; (**b**) nature of the propagation vector of the inhomogeneous plane leaky wave (phase and attenuation vectors); (**c**) equiphase and equiamplitude planes of the leaky wave with the relevant leakage phenomenon along the guide; (**d**) equivalent transverse resonance network, resonance conditions, and network parameters for the numerical evaluation of the leaky-wave complex wavenumbers as a function of the physical parameters involved.

α_z originates, too, which as mentioned accounts for the longitudinal attenuation due to the field that is no longer confined and flows also in the outside region; the propagating field inside the waveguide and in the proximity of its aperture is therefore described by the complex longitudinal wavenumber $k_z = \beta_z - j\alpha_z$, whose quantification depends on the physical parameters.

In this case the leakage phenomenon is assumed along $+z$ ($\beta_z > 0$ and $\alpha_z > 0$), and by supposing that the vertical field variations are almost negligible ($k_y \cong 0$), it is easily seen that, from the general separation condition for waveguides ($k_0^2 = \omega^2\mu_0\varepsilon_0 = k_t^2 + k_z^2 \cong k_x^2 + k_z^2$), the horizontal wavenumber is also complex:

$$k_x = \beta_x - j\alpha_x, \tag{2}$$

with $\beta_x > 0$ and $\alpha_x < 0$ since it results $\beta_x\alpha_x = -\beta_z\alpha_z$. Therefore a plane wave of inhomogeneous type exists, having a complex propagation vector $\mathbf{k}$ of the type.

$$\begin{aligned} \mathbf{k} &= \boldsymbol{\beta} - j\boldsymbol{\alpha} \\ \boldsymbol{\beta} &= \beta_x\mathbf{x}_0 + \beta_z\mathbf{z}_0 \\ \boldsymbol{\alpha} &= \alpha_x\mathbf{x}_0 + \alpha_z\mathbf{z}_0 \end{aligned} \tag{3}$$

with the phase vector $\boldsymbol{\beta}$ directed at an angle that describes the outgoing of power from the guide to the external, and the attenuation (leakage) vector $\boldsymbol{\alpha}$ that is perpendicular to $\boldsymbol{\beta}$, and represents attenuation along z and amplification along x. Consequently, the field has a spatial dependence of the type

$$\exp[-j(\beta_x x + \beta_z z)]\exp[|\alpha_x|x - \alpha_z z] \tag{4}$$

Therefore, this plane wave travels at an angle $\theta = \sin^{-1}(\beta_z/|\boldsymbol{\beta}|)$ with respect to broadside carrying out power, and its amplitude is transversely increasing as expected in a leaky wave. It should be noted that the direction angle θ of the leaky wave is usually expressed under the approximate form: $\theta \cong \sin^{-1}(\beta_z/k_0)$, since in general the leakage constant is numerically negligible with respect to the phase constant. The nature of the propagation vector is sketched in Fig. 3b, while the distribution of equiphase and equiamplitude surfaces with the decreasing power flow along the guide is represented in Fig. 3c. It should be recalled that, even though the leaky wave has a nonspectral nature, the field generated from a source located at a finite distance along z still satisfies the radiation condition, since the field increases transversely only in a limited sector given by angles less than the θ value describing the direction of power leakage [3,4].

A quantitative description of this LWA is easily achieved with a simple analysis of the complex eigenvalue derivable as a modification of the dominant mode by employing a TRT [1,2,11]. To this aim, it is required a characterization of the slit aperture in the side wall as a circuit element in the equivalent transmission line. For the quantification of such discontinuities, a great deal of work was developed since the 1950s, basically through variational methods [2,7,9–12]. The description of the radiative and reactive effects of the slit in the side wall of the rectangular guide can be represented by a lumped

element (e.g., an admittance $Y_R = G_R + jB_R$) as a function of geometry and frequency. The transverse network is reported in Fig. 3d. The solution of the relevant resonance equation in the complex plane for the perturbed dominant mode describes the leaky-wave behavior.

3. DESIGN PROCEDURES

3.1. Basic Radiation Features

The basic design principles of LWAs are generally derivable from the knowledge of the desired beam width and of the pointing direction. In LWAs these quantities can be linked in a straightforward way to the complex longitudinal wavenumber.

In fact, the beam maximum direction θ_M is, as seen before, related mainly to the normalized phase constant, according to the simple relationship

$$\sin\theta_M \cong \frac{\beta_z}{k_0} \tag{5}$$

Since β_z has a dispersive behavior as is typical of waveguiding structures, a change in the frequency yields a scanning of the beam: typically, as the frequency is increased from the cutoff, the pointing angle varies its direction from around the broadside ($\theta_M = 0^\circ$), toward the forward endfire ($\theta_M = 90^\circ$).

In regard to the beamwidth, we recall that the leakage constant α_z quantifies the rate of power loss along the line due to the radiation, thus influencing primarily the effective dimension of the equivalent aperture for the line source: in fact, the more α_z increases, the more the actual illumination length reduces (and the less the beamwidth is focused).

A basic link between the leakage constant and the antenna length L derives from the specification of the radiation efficiency η, expressible in LWAs as $\eta = [P(0) - P(L)]/P(0)$, where $P(0)$ is the input power delivered to the structure and $P(L)$ is the output power left at the end termination. The link between efficiency, leakage rate, and length is generally dependent on the desired radiation pattern and therefore on the aperture distribution: referring to a uniform-section LWA, where α_z is independent of z, this results in $\eta = 1 - \exp(-2\alpha_z L)$. It should also be noted that, for narrowbeam applications, a very high increase in efficiency should require an extreme prolongation of the line source; actually, in LWAs it is typical to radiate around 90% or at most 95% of the input power, where the remaining power at the end termination is absorbed by a matched load to avoid a backlobe of radiation due to the reflected wave.

Once the efficiency is chosen, a fixed link therefore exists between the relative length in terms of wavelengths L/λ_0 and the normalized leakage constant α_z/k_0. For a uniform-section LWA, an inverse proportionality relationship between L and α_z is found of the type

$$\frac{L}{\lambda_0} \cong \frac{c}{\alpha_z/k_0}$$
$$c = \left(\frac{1}{4\pi}\right)\ln\left(\frac{1}{1-\eta}\right) \tag{6}$$

where c is related to the value of the desired efficiency (e.g., for 90% of efficiency it is $c = 0.185$). For a nonuniform section, since α_z depends on z, the link between efficiency, length, and leakage rate is related to the chosen illumination and is more complicated.

In order to achieve narrow beams in the elevation angle, the effective longitudinal aperture has to be sufficiently wide (usually several wavelengths), and this implies a rather low leakage rate. The half-power (−3-dB) beamwidth $\Delta\theta$ is directly linkable to the normalized antenna length L/λ_0 through an approximate relationship, which takes into account also the contribution of the scan angle [1]:

$$\Delta\theta \cong a \Big/ \left(\frac{L}{\lambda_0}\cos\theta_M\right) \quad \text{(rad)} \tag{7}$$

where the proportionality factor a is dependent on the aperture distribution; it has the most reduced value for a constant aperture distribution ($a \cong 0.88$) and increases for tapered distributions (typically, more than unity) [1]. From the previous expression, it is seen that, since $\cos\theta_M \cong k_t/k_0$, the beamwidth is also expressible as $\Delta\theta \cong 2\pi/(k_t L)$. This means that the beam width is, as a first approximation, practically constant when the beam is scanned away from broadside by varying the frequency for air-filled LWAs (where k_t is independent of frequency), while it changes for dielectric-filled LWAs (where k_t depends on frequency).

The effective aperture is anyway reduced for a fixed antenna length as the beam approaches endfire (where the previous expression becomes not accurate), and $\Delta\theta$ anyway tends in practice to enlarge. It can be seen that for an ideal semiinfinite uniform structure, that is an antenna aperture from $z = 0$ to $z = L \to \infty$, the beam width is determined by the leakage rate only, since in this case it can be found that $\Delta\theta \cong 2\alpha_z/k_t$. Moreover, in this situation the radiation pattern depends only on β_z and α_z and does not present sidelobes:

$$R(\theta) \approx \frac{\cos^2\theta}{(\alpha/k_0)^2 + (\beta/k_0 - \sin\theta)^2} \tag{8}$$

For finite antenna lengths, sidelobes are produced and the expression for $R(\theta)$ is more involved. In general the specifications on the sidelobe level are related to the choice of the aperture distribution, whose Fourier transform allows the derivation of the radiation pattern.

3.2. Scanning Properties

It is seen that the beams for LWAs derived by partially open air-filled metallic waveguides scan in theory an angular region from around the broadside ($\beta_z/k_0 \cong 0$) towards one endfire ($\beta_z/k_0 \cong 1$).

In practice, around broadside the structure works near the cutoff region of the closed waveguide where reactive effects are increasingly important. The leaky-wave values for β_z/k_0 cannot anyway be extremely low and at the same time α_z/k_0 tends to increase too much, adversely affecting the possibility of focusing radiation at broadside.

Concerning the behavior at endfire it is seen that, since β_z/k_0 tends to unity asymptotically as the frequency

increases, in the unimodal range (where these structures are usually employed) the beam cannot reach so closely the endfire radiation in an air-filled LWA. A way of improving the angular scanning is to fill these structures with dielectric materials. Thus, since in this case the normalized phase constant approaches the square root of the relative permittivity as the frequency is increased ($\beta_z/k_0 \to \varepsilon_r^{1/2}$), the $\beta_z/k_0 = 1$ value can actually be approached in a much more restricted frequency range. It should anyway be noted that for such dielectric-filled structures the beam width may change strongly as a function of frequency and therefore as the pointing angle varies [see comments on Eq. (7)].

Moreover, it should be noted that in many leaky structures (such as the dielectric and printed ones), as the frequency is increased, the leaky-mode solution changes into a guided-mode solution through a complicated "transition region" [13,14]; in this frequency range, also called "spectral gap," the contribution of the leaky wave to the field tends progressively to decrease, and generally the structure does not work well as a LWA.

As stated above, while the uniform LWAs usually radiate only in the forward quadrant, with the limits specified above, the LWAs derived from periodically modulated slow-wave guides can start to radiate from the backward endfire in the lower frequency range.

The design principles of periodic LWAs are in most part similar to those of uniform LWAs [1,2]. The main difference lies in the characterization of the fast wave that is now associated to a Floquet's space harmonic of the periodic guide [1,2,14,15]. One can see that if a uniform guide is considered whose operating mode is slow ($\beta_z/k_0 > 1$, e.g., a dielectric waveguide), and a longitudinally-periodic discontinuity is properly added (e.g., an array of metal strips or notches, placed at suitable distances p), such periodicity furnishes a field expressible in an infinite number of space harmonics ($\beta_{zn}p = \beta_{z0}p + 2n\pi$), where β_{z0} is the phase constant of the fundamental harmonic, which is slightly varied with respect to the original value β_z of the unperturbed guide. With proper choices of the physical parameters, it is in general possible to make only one harmonic fast (typically, the $n = -1$), so that it can radiate as a leaky wave (presence of an additional attenuation constant α_z).

In this case, the phase constant of this fast harmonic can assume both positive and negative values ($-1 < \beta_z/k_0 < 1$), as a function of the parameters involved; in particular, as frequency is increased, the beam starts to radiate from backward endfire toward the broadside. In general, also periodic LWA's have difficulties in working well in the broadside region, since usually for periodic structures there exists an "open stopband" [14], where the attenuation constant rapidly increases, resulting in a widening beamwidth.

As the frequency is further increased after broadside, the beam is then scanned also in the forward quadrant. In periodic LWAs, depending on the choice of the design parameters, additional limitations in the forward scanning behavior could exist when also a second harmonic starts to radiate before the first harmonic reaches its endfire, thus limiting the single-beam scanning range [1,14].

3.3. Leaky-Wave Arrays for Pencil-Beam Radiation

If an increase of directivity in the cross-plane is desired, a simple improvement for LWAs based on long radiating slots can be achieved by a physical enlargement of the transverse aperture (e.g., with a flared transition to enlarge the effective cross-aperture). As said before, a more efficient way to increase directivity in the cross-plane is to use a number of radiators placed side by side at suitable lateral distances, thus constituting a linear array; it is then possible to achieve radiation with a focused pencil beam. In addition, if properly phased, these arrays of LWAs allow a 2D scanning of the beam: in the elevation plane, as is typical for LWAs, the scanning is achievable by varying the frequency, while in the cross-plane the scanning is achievable with phase shifters that vary the phase difference among the single line sources. As noted, in LWAs only a unidimensional number of phase shifters is therefore necessary, with particular structural simplicity and economic advantage if compared to all the usual radiators requiring a 2D number of shifters for the scanning. Additional desirable features of such arrays are in general the absence of grating lobes and of blind spots, and good polarization properties.

For analysis of such LW arrays, an efficient method is that one based on the "unit cell" approach [6,7]. In this way, it is possible to derive the behavior of the global structure by referring to a single radiator taking into account the mutual effects due to the presence of all the others. In the equivalent network this is achievable by changing only the description of the radiation termination for a periodic-type array environment (infinite number of linear elements): in particular, an "active admittance" can be quantified, which describes the external radiating region as a function of the geometry and of the scan angle. More sophisticated techniques also allow accurate analyses of arrays by taking into account the mutual couplings for a finite number of elements [6].

3.4. Radiation Pattern Shaping

In the basic requirements of the radiation pattern, in addition to the specification for the maximum of the beam direction and for its half-power width, also the sidelobe behavior has a primary importance. In a general sense, it is desired to derive the properties of the source in connection with a desired radiation pattern. Since LWAs can be viewed as aperture antennas with a current distribution having a certain illumination $A(z)$, it is possible to obtain the far field through a standard relationship:

$$E(\theta) = G(\theta) \int_0^L |A(z')| e^{j\mathrm{Arg}[A(z')]} e^{jkz' \sin\theta} dz' \qquad (9)$$

The radiation pattern for E is expressed in terms of a Fourier transform of the line-source complex current distribution on the aperture multiplied by the pattern of the element current G (e.g., a magnetic dipole).

It is easily seen that if the LWA geometry is kept longitudinally constant, the amplitude distribution has always an exponential decay of the type: $\exp(-\alpha_z z)$. As is known, this behavior furnishes a quite poor radiation

pattern for the sidelobes that are rather high (around −13 dB). It therefore derives that, in conjunction with the choice of a fixed illumination function $A(z)$ giving a desired sidelobe behavior (e.g., cosine, square cosine, triangular, Taylor), the leakage rate has to be modulated along the main direction z of the line source: in practice this is achievable by properly modifying the cross section of the LWA structure along z, with the procedure usually known as "tapering." Considering that, for a smoothly tapered antenna, the power radiated per unit length from the antenna aperture is simply related to the aperture distribution [viz. $-dP(z)/dz = 2\alpha_z(z)P(z) = c|A(z)|^2$], a useful analytic expression for $\alpha_z(z)$ as a function of the amplitude $A(z)$, of the line-source length L, and of the efficiency η is obtainable [1,2,16]:

$$\alpha_z(z) = \frac{1}{2}\frac{|A(z)|^2}{\frac{1}{\eta}\int_0^L |A(z')|^2 dz' - \int_0^z |A(z')|^2 dz'} \tag{10}$$

From this equation it is also seen that the more the efficiency is desired high (around unity), the more α_z has to increase toward extremely high values around the terminal section (as mentioned, efficiency in common practice does not exceed 90–95%).

In general, in the tapering procedure the longitudinal modification of the geometry should be made in an appropriate way in order to affect only the leakage constant, taking into account that the phase constant should conversely be maintained the same (in pencil-beam applications, β_z should not depend on z in order to have the correct pointing angle for each elementary current contribution on the aperture).

The pattern shaping procedure requires therefore the knowledge of the phase and leakage constants as a function of the geometric and physical parameters of the chosen structure, and this is achievable, as stated, by finding the suitable complex eigensolution with numerical methods. Since the pattern shaping requires a proper α_z distribution with β_z constant, the procedure is strongly simplified if it is possible to find geometric parameters through which the leakage and phase constants are varied as independently as possible. This property is related to the topology characteristics of the waveguiding structure.

An example of tapering is sketched in Fig. 4 for a leaky structure, the so-called "stepped" LWA (Fig. 4a), proposed for high-performance applications with well-controlled radiation patterns [17] and additional general desirable features (increased geometrical flexibility, compactness, low profiles for aerospace applications, etc.).

In Fig. 4b the detailed behavior of the modulation in the height of the lateral steps is shown as a function of z for a desired illumination (cosine type). A first action only on the steps' unbalance, with their mean value kept constant (dashed profile), modifies appropriately the longitudinal distribution of the leakage constant, maintaining almost constant the phase constant. A second action is advisable to compensate the phase nonlinearity, which can give rather disturbing effects on the radiation patterns: in this topology it is possible to slightly vary the steps' mean value, with the previously fixed unbalance, to obtain the final valid profile (solid line). The relevant radiation patterns are then illustrated in Fig. 4c,d, for the single-shot and the double-shot tapering procedures, respectively; Fig. 4c is a rather "distorted" pattern related to the nonoptimized tapering (dashed profile), while Fig. 4d is a "correct" cosine-type pattern related to the optimized tapering (solid profile). The tapering procedure can be performed numerically in an easy way from a TRT network representation of the structure. The typical scanning behavior of these kinds of antennas is finally illustrated in Fig. 4e for the pointed beam variable by frequency.

4. FURTHER EXAMPLES OF SPECIFIC STRUCTURES

4.1. Partially Open Metallic Waveguides

One of the main drawbacks of the antenna shown in Fig. 1a is related to the leakage constant, which in general cannot be reduced below a certain limit. Reduced leakage amounts are achievable by slitting the top wall of the rectangular guide, decreasing the current modification due to the cut (Fig. 5a). By shifting the cut with respect to the central vertical plane, it is possible to modulate the leakage rate: investigations were also performed with tapered meander profiles for sidelobe control [18].

A way of improving the polarization purity in the basic geometry of a top-wall slitted rectangular guide is to use an aperture parallel-plate stub, able to reduce the contribution of the higher modes on the aperture, which are below cutoff, while the dominant leaky wave travels nonattenuated as a TEM-like mode at an angle [19] (Fig. 5b). Metal wide flanges, simulating an open half-space on the upper aperture, can increase the directivity of this type of LWA.

4.2. Printed Lines: Microstrip LWAs

The possibility of using LWAs also in printed circuitry has received interest that is probably destined to increase in the near future due to the wide use of planar technology for light, compact, and low-cost microwave integrated circuits (MICs). Among the various printed waveguides that can act as leaky-wave radiators (coplanar guides, slot and strip lines, etc.) [7,20], we can refer to structures derivable from lengths of microstrip. Many different configurations can be employed with microstrips acting as traveling-wave radiators. A first class is based on modulating the dominant mode of the structure with periodic loadings, such as resonant patches or slots (Fig. 6a), and also by varying the lineshape periodically with different meander contours (Fig. 6b) [21]. Even though different solutions have been tested, theory on this topic seems to deserve further studies.

A different way of operation concerns the use of uniform structures acting on higher-order modes that can become leaky for certain values of the parameters involved (Fig. 6c). Analysis of the complex propagation characteristics of the microstrip line shows in fact that, in addition to the dominant quasi-TEM mode, the higher-order modes generally become leaky in suitable frequency ranges [7,20] (see Fig. 2). In particular, it is seen that the

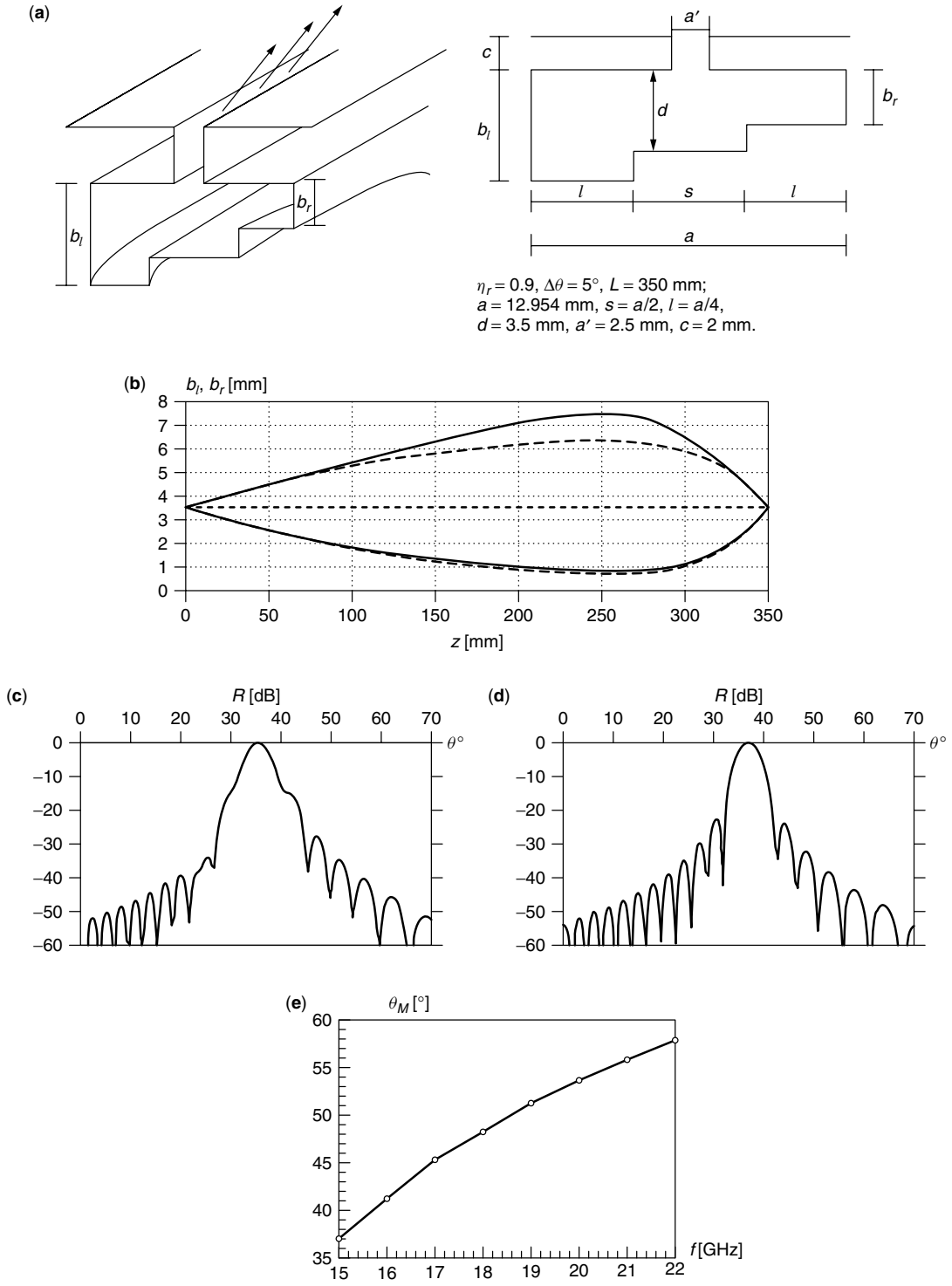


Figure 4. Example of LWAs tapering procedure to achieve a required aperture distribution for pattern shaping: (**a**) reference structure of a stepped rectangular guide LWA with relevant parameters. (**b**) longitudinal modulation of the lateral steps (b_l, b_r vs. z) related to a cosine-type illumination function for a microwave application. The dashed line of b_l, b_r vs. z profile is obtained with a single-shot tapering procedure, that is only an action on the imbalance $\Delta b = (b_l - b_r)/(b_l + b_r)$ taking a constant value of mean height $b_m = (b_l + b_r)/2$ (thus, variations on the phase constant are anyway introduced). The solid-line profile is due to a double-shot tapering procedure, where phase errors are compensated by suitably varying b_m. (**c**) "Distorted" normalized radiation pattern R (dB) according to the dashed-line profile. (**d**) "Correct" radiation pattern according to the solid-line profile for the cosine illumination of the stepped LWA. (**e**) Typical scanning properties for the pointed beam as a function of the frequency (stepped LWA under investigation).

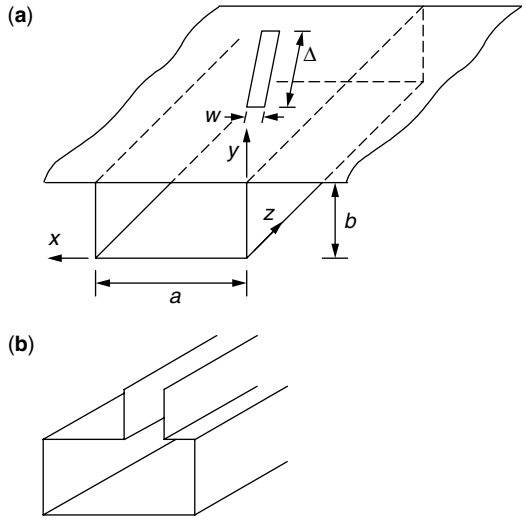


Figure 5. (**a**) Top-wall slitted rectangular guide LWA; (**b**) stub-loaded rectangular guide LWA.

first higher-mode EH_1 can be excited with a proper odd-type source (the mid-plane of symmetry is a perfect electric conductor) and, as frequency is raised, starts to leak power. In general, for the planar structures, leakage can occur in two forms: the surface wave leakage (power that is carried away through the TE and/or TM surface modes of the layered structure), and the "space-wave" leakage (power that is carried away through the TEM mode of the free space) [22]. It is found that, for suitable choices of the parameters with an appropriate excitation, the EH_1 mode can represent rather efficiently the radiation of the microstrip in a certain frequency range (see, e.g., Fig. 2). The coupling phenomenon between the feeding and the radiating line is an aspect to be accurately evaluated, and simplified equivalent networks can be convenient to this aim (23). Radiation performance of printed-circuit LWAs (concerning power handling, polarization, efficiency, pattern shaping, etc.) can be less versatile and satisfactory if compared with LWAs derived from metal guides. From a practical point of view, difficulties can be found particularly in acting independently on the phase and leakage constants through the physical parameters. Uniform-type microstrip LWAs have also been investigated in array configurations for 2D pencil-beam scanning [7,24,25].

4.3. Nonradiative Dielectric (NRD) Guide LWAs

Nonradiative-dielectric (NRD) waveguide, proposed for millimeter-wave applications [26] (Fig. 7a), is a hybrid metal/dielectric guide; it consists of a dielectric rod inserted between metal plates placed at a distance apart that is less than the free-space wavelength. In this way, each discontinuity that preserves the central horizontal-plane symmetry gives only reactive contributions, reducing interference and radiation effects in integrated circuits. A number of passive and active components has been realized with such topology, and also integrated antennas and arrays have been proposed [27,28]. Usually NRD LWAs employ some asymmetry in the basic geometry in order to make leaky the operating mode. A first possible choice [27] (Fig. 7b) is to shorten the length of the plates so that the bound operating mode (LSM_{01}) [26] presents a nonnegligible amplitude contribution on the equivalent aperture, and can give rise to an outgoing leaky wave in the fast-wave range. Another possible choice [7] (Fig. 7c) is to insert some geometrical asymmetry with respect to the central plane (typically an airgap between dielectric and metal), so that a field having a net electric component perpendicular to the plates can be excited, and power can leak out in the form of a TEM-like mode traveling at an angle in the parallel-plate region toward the external environment. Various

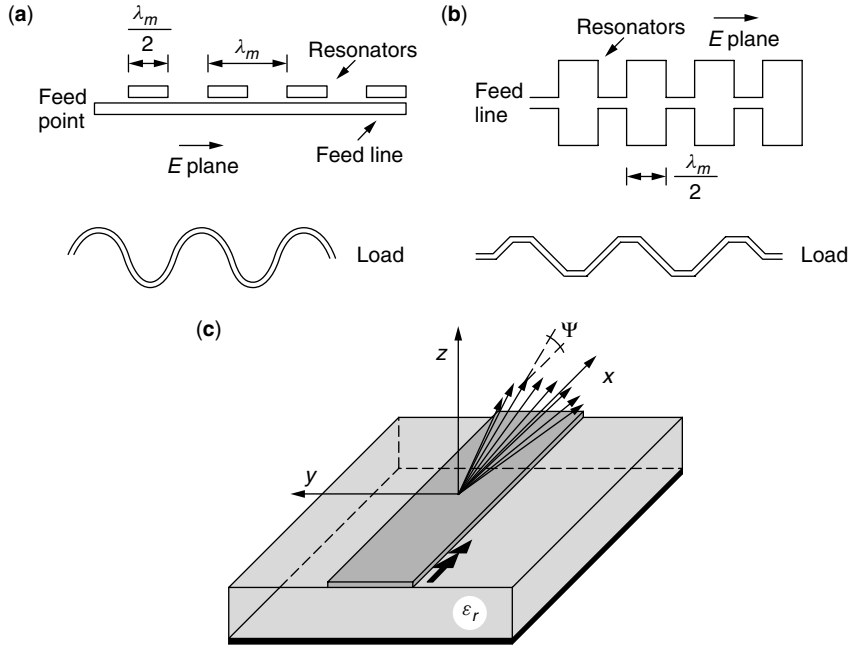


Figure 6. (**a**) Periodically loaded microstrip LWA; (**b**) periodical meander microstrip LWA; (**c**) uniform higher-mode microstrip LWA—space-wave radiation can be associated, for example, with the strip current distribution of the EH_1 mode, which is leaky in a suitable frequency range (see Fig. 2).

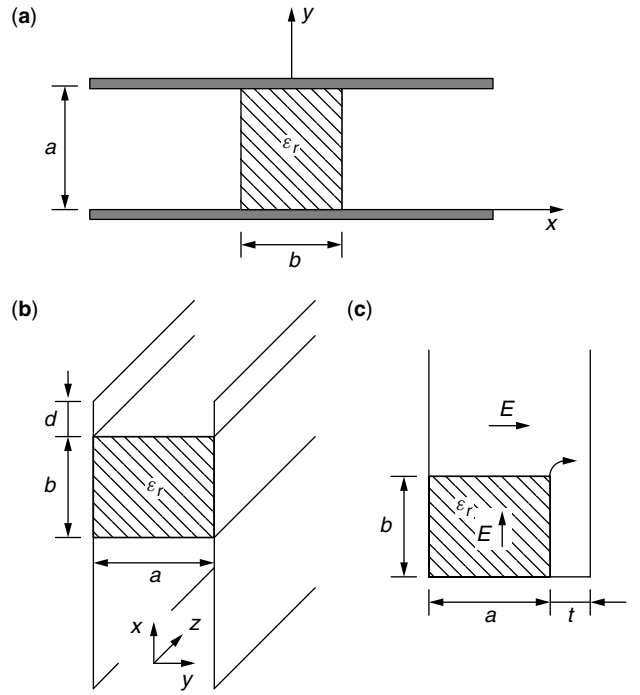


Figure 7. (**a**) Nonradiative dielectric (NRD) waveguide; (**b**) foreshortened NRD LWA; (**c**) asymmetrical NRD LWA.

analyses and design procedures have been developed for these configurations in conjunction with measurements on prototypes.

4.4. Dielectric LWAs

As said, in basic dielectric guides a periodic loading is required in order to isolate a suitable fast-wave space harmonic from the intrinsically slow-wave structure. The most usual periodic perturbation is represented by grating of grooves [29] or metal strips [30,31], usually placed in the top surface of the guide (Fig. 8a); also lateral metal patches can be used in hybrid forms (dielectric/microstrip) (Fig. 8c) [32]. When a sidelobe control is required, the taper is realized on the periodic perturbation (e.g., with grooves or strips slightly changing their dimensions longitudinally). Various studies have been developed to characterize the theoretical performances for these radiators [33]; also, practical aspects have been analyzed, such as the proper feeding elements in order to avoid spurious radiation, and the reduction of the beamwidth in the cross-plane with flared horns [34] (Fig. 8d). All these topologies are good candidates particularly for high-frequency applications (millimeter and submillimeter waves), where the use of dielectric instead of metal for the guidance can reduce the loss effects.

4.5. Layered Dielectric-Guide LWAs

It has been observed that LWAs based on single dielectric layers, also with a ground plane on one side, usually present quite high leakage values, with consequent weak capability in focusing radiation. A significant improvement is achievable by using additional dielectric layers (Fig. 9a); in particular, interesting analyses were performed on substrate/superstrate layered structures [35–37]. By properly dimensioning the heights and the dielectric constants (usually the substrate has lower permittivity than the superstrate), it is possible to excite with a simple element (dipole or slot) a leaky wave giving a conical (due to the symmetries of the topology) highly directive beam. More recently, this basic substrate/superstrate topology has been arranged to allow for a very focused pencil beam with a limited number of radiating elements in form of widely spaced array, exploiting an interaction between leaky and Floquet's modes (Fig. 9b) [38]. Through very simple design procedures, such configurations have the advantages of good radiative performance (high directivity, absence of grating lobes, etc.) with an array

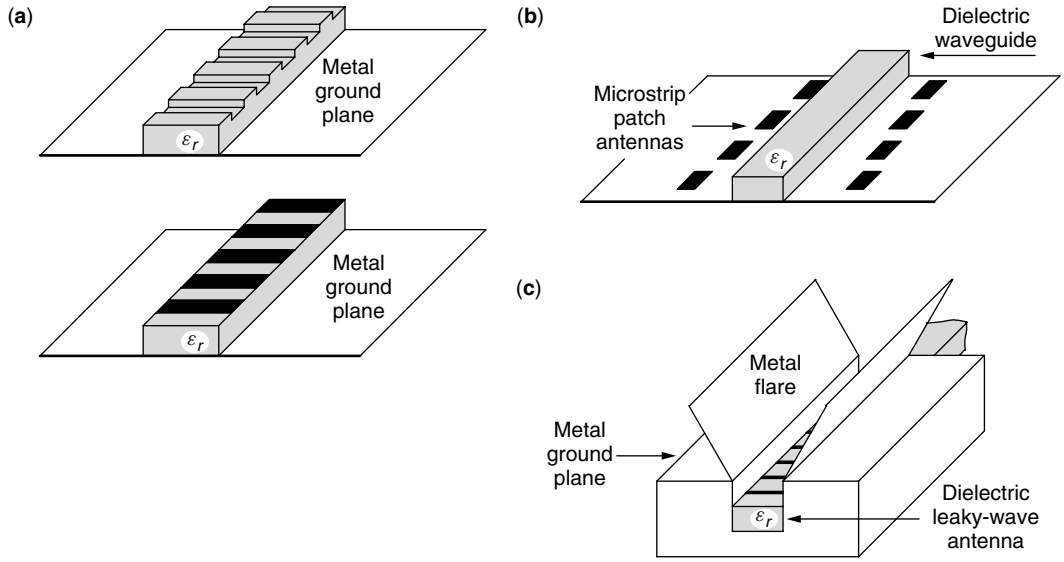


Figure 8. (**a**) Periodically-loaded dielectric LWAs; (**b**) hybrid dielectric/microstrip (insular guide with patches) LWA; (**c**) dielectric LWA with a flared horn to reduce the cross-plane beam width.

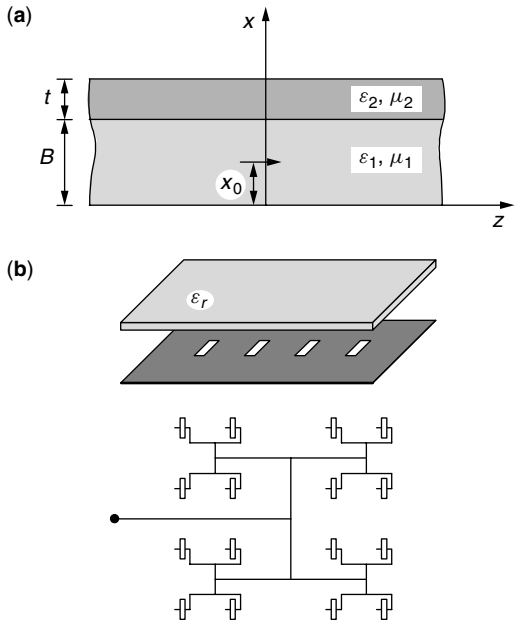


Figure 9. (**a**) Layered dielectric LWA based on a substrate/superstrate structure with a dipole excitation; (**b**) high-gain LW arrays of widely spaced elements in a substrate/superstrate structure: linear and planar configurations (for the latter case, a top view is shown for a microstrip feeding network of widely spaced slot elements on the ground plane of the substrate/superstrate structure).

of few spaced 1D or 2D elements, reducing the cost of the beam forming network and exploiting the greater interspace available at high frequencies (dual-polarization applications, etc.).

5. PRACTICAL CONSIDERATIONS AND MEASUREMENTS

5.1. Feed, Losses, and Manufacture

Feeding LWAs is usually quite simple. In particular, for LWAs derived by metal guides, the feed is represented by a continuous transition from the closed structure acting on a suitable guided mode to the related open one acting on the perturbed (leaky) mode [1–4]. Tapered transitions from the closed to the open structures can be realized to reduce the discontinuity effects and the possible excitation of spurious modes that could arise from abrupt transitions. At the output termination, the introduction of a matched load drastically decreases the remaining power that, if reflected, should give rise to a backlobe, in a direction symmetrical to the main beam with respect to the broadside. The use of dielectric structures can present more difficulties in feeding, in particular in planar configurations. For efficiency and radiation performance, attention has to be paid in avoiding the excitation of additional guided and leaky modes, and also in obtaining a good excitation of the desired leaky wave. For planar guides, such as microstrip or layered dielectrics, local coupling elements (such as slot or dipoles) are usually employed to excite the leaky mode from an input line toward the radiating line.

Ohmic losses do not usually affect much the radiative performance (efficiency, etc.) of LWAs, since the attenuation due to the leakage of radiated power is generally more influent than the attenuation due to dissipated power in the nonideal guiding structure [16]. However, as frequency increases, power loss can be excessive, particularly for LWAs based on closed metal guides. Therefore, for millimeter wave applications the choice of open guides with dielectrics and limited use of metal is often advisable.

The general simplicity of LWA structures makes their manufacture usually easy to perform, even though different construction problems can arise depending on the chosen topology and the frequency range. Simple structures are particularly desirable at millimeter waves, due to the reduced dimensions. On the other hand, too simplified shapes seldom can allow a good control of the radiation performance. In particular, a delicate aspect concerns the usually small longitudinal modifications of the geometry related to tapering for sidelobe control. In this case, an accurate determination of the fabrication imprecisions and tolerances has particular importance to avoid overwhelming the required geometric variations for tapering, thus degrading the improvements of the pattern shaping. Finally, the effects of radomes, used for environmental protection, have also been analyzed [39].

5.2. Measurement Techniques

The radiation properties of LWAs can be tested experimentally through different types of measurement, most of them applicable to aperture antennas [40]. Some basic parameters, such as efficiency and mismatching effects, can be measured directly through the transmission and/or reflection scattering parameters with a network analyzer. Radiation patterns and directivity properties as a function of the observation angles (θ and ϕ in the zenith and azimuth planes, respectively) can be measured for various frequency values with different techniques, on the aperture, in the radiating near field (Fresnel region), and in the far field (Fraunhofer region) [17].

Measurements on the aperture are quite easy to perform, in particular for LWAs derived from partially open metal guides. As already said, the basic parameters to be determined in LWAs, from which a complete knowledge of the radiative characteristics is achieved, are the phase and the leakage constants. A measurement of the field in the close proximity of the aperture can be achieved with a small pickup element (e.g., an electric dipole probe placed parallel to the aperture electric field). Amplitude and phase of the signal received by the probe are thus measurable through a network analyzer, with possible compensations related to the mutual coupling between the current distribution on the aperture and the current probe element.

BIOGRAPHIES

Fabrizio Frezza received the Laurea (degree) cum laude in electronic engineering from "La Sapienza" University of Rome, Italy, in 1986. In 1991 he obtained a doctorate in applied electromagnetics from the same university.

In 1986, he joined the Electronic Engineering Department of the same university, where he has been a researcher from 1990 to 1998, a temporary professor of electromagnetics from 1994 to 1998, and an associate professor since 1998. His main research activity concerns guiding structures, antennas and resonators for microwaves and millimeter waves, numerical methods, scattering, optical propagation, plasma heating, and anisotropic media.

Dr. Frezza is a senior member of IEEE, a member of Sigma Xi, of AEI (Electrical and Electronic Italian Association), of SIOF (Italian Society of Optics and Photonics), of SIMAI (Italian Society for Industrial and Applied Mathematics), and of AIDAA (Italian Society of Aeronautics and Astronautics).

Alessandro Galli received the Laurea degree in electronic engineering in 1990 and a Ph.D. in applied electromagnetics in 1994, both from "La Sapienza" University of Rome, Italy. In 1990, he joined the Electronic Engineering Department of "La Sapienza" University of Rome for his research activity. In 2000, he became temporary professor of electromagnetic fields for telecommunications engineering at "La Sapienza" University of Rome, and in 2002 he became associate professor of electromagnetics at the same university.

His scientific interests mainly involve electromagnetic theory and applications, particularly regarding analysis and design of passive devices and antennas (dielectric and anisotropic waveguides and resonators, leaky-wave antennas, etc.) for microwaves and millimetre waves. He is also active in bioelectromagnetics (modeling of interaction mechanisms with living matter, health safety problems for low-frequency applications, and mobile communications, etc.). In 2000, he was selected as a member of the Technical Committee of the Advisor chosen by the Italian Government for the licenses of the third-generation cellular phones (UMTS).

Dr. Galli is a member of IEEE (the Institute of Electrical and Electronics Engineers). In 1994, he received the Barzilai Prize for the best scientific work of under-35 researchers at the 10th National Meeting of Electromagnetism. In 1994 and 1995, he was the recipient of the Quality Presentation Recognition Award presented by the IEEE Microwave Theory and Techniques Society (MTT-S).

Paolo Lampariello obtained the Laurea degree (cum laude) in electronic engineering at the University of Rome, Italy in 1971.

In 1971, he joined the Institute of Electronics, University of Rome. Since 1976, he has been engaged in educational activities involving electromagnetic field theory. He was made professor of electromagnetic fields in 1986. From November 1988 to October 1994 he served as head of the Department of Electronic Engineering of the "La Sapienza" University of Rome. Since November 1993, he has been the president of the Electronic Engineering Curriculum and since September 1995 he has been the president of the Center Interdepartmental for Scientific Computing. From September 1980 to August 1981 he was a NATO postdoctoral research fellow at the Polytechnic Institute of New York, Brooklyn.

Professor Lampariello has been engaged in research in a wide variety of topics in the microwave field, including electromagnetic and elastic wave propagation in anisotropic media, thermal effects of electromagnetic waves, network representations of microwave structures, guided-wave theory with stress on surface waves and leaky waves, traveling-wave antennas, phased arrays, and, more recently, guiding and radiating structures for the millimeter and near-millimeter wave ranges.

Professor Lampariello is a fellow of the Institute of Electrical and Electronics Engineers, and a member of the Associazione Elettrotecnica ed Elettronica Italiana.

BIBLIOGRAPHY

1. A. A. Oliner, Leaky-wave antennas, in R. C. Johnson, ed., *Antenna Engineering Handbook*, 3rd ed., McGraw-Hill, New York, 1993, Chap. 10.
2. C. H. Walter, *Traveling Wave Antennas*, McGraw-Hill, New York, 1965; Peninsula Publishing, Los Altos, CA, reprint, 1990.
3. T. Tamir and A. A. Oliner, Guided complex waves, Parts I and II, *Proc. IEE* **110**: 310–334 (1963).
4. T. Tamir, Inhomogeneous wave types at planar interfaces: III — Leaky waves, *Optik* **38**: 269–297 (1973).
5. I. Ohtera, Diverging/focusing of electromagnetic waves by utilizing the curved leakywave structure: Application to broad-beam antenna for radiating within specified wide-angle, *IEEE Trans. Antennas Propag.* **AP-47**: 1470–1475 (1999).
6. R. C. Hansen, *Phased Array Antennas*, Wiley, New York, 1998.
7. A. A. Oliner (principal investigator), *Scannable Millimeter Wave Arrays*, Final Report on RADC Contract F19628-84-K-0025, Polytechnic Univ., New York, 1988.
8. T. Itoh, Millimeter-wave leaky-wave antennas, *Proc. Int. Workshop Millimeter Waves*, Italy, 1996, pp. 58–78.
9. T. Itoh, ed., *Numerical Techniques for Microwave and Millimeter-Wave Passive Structures*, Chap. 3 (J. R. Mosig), Chap. 5 (T. Uwano and T. Itoh), and Chap. 11 (R. Sorrentino), Wiley, New York, 1989.
10. R. Sorrentino, ed., *Numerical Methods for Passive Microwave and Millimeter Wave Structures*, IEEE Press, New York, 1989.
11. L. O. Goldstone and A. A. Oliner, Leaky-wave antennas — Part I: Rectangular waveguides, *IRE Trans. Antennas Propag.* **AP-7**: 307–319 (1959).
12. N. Marcuvitz, *Waveguide Handbook*, McGraw-Hill, New York, 1951.
13. P. Lampariello, F. Frezza, and A. A. Oliner, The transition region between bound-wave and leaky-wave ranges for a partially dielectric-loaded open guiding structure, *IEEE Trans. Microwave Theory Tech.* **MTT-38**: 1831–1836 (1990).
14. S. Majumder, D. R. Jackson, A. A. Oliner, and M. Guglielmi, The nature of the spectral gap for leaky waves on a periodic strip-grating structure, *IEEE Trans. Microwave Theory Tech.* **MTT-45**: 2296–2307 (1997).

15. R. E. Collin, *Field Theory of Guided Waves*, 2nd ed., IEEE Press, New York, 1991.

16. C. Di Nallo, F. Frezza, A. Galli, and P. Lampariello, Rigorous evaluation of ohmic-loss effects for accurate design of traveling-wave antennas, *J. Electromagn. Wave Appl.* **12**: 39–58 (1998).

17. C. Di Nallo et al., Stepped leaky-wave antennas for microwave and millimeter-wave applications, *Ann. Télécommun.* **52**: 202–208 (1997).

18. F. L. Whetten and C. A. Balanis, Meandering long slot leaky-wave waveguide antennas, *IEEE Trans. Antennas Propag.* **AP-39**: 1553–1560 (1991).

19. P. Lampariello et al., A versatile leaky-wave antenna based on stub-loaded rectangular waveguide: Parts I–III, *IEEE Trans. Antennas Propag.* **AP-46**: 1032–1055 (1998).

20. H. Shigesawa, M. Tsuji, and A. A. Oliner, New improper real and complex solutions for printed-circuit transmission lines and their influence on physical effects, *Radio Sci.* **31**: 1639–1649 (1996).

21. J. R. James and P. S. Hall, *Handbook of Microstrip Antennas*, Peter Peregrinus, London, 1989.

22. F. Mesa, C. Di Nallo, and D. R. Jackson, The theory of surface-wave and space-wave leaky-mode excitation on microstrip lines, *IEEE Trans. Microwave Theory Tech.* **MTT-47**: 207–215 (1999).

23. P. Burghignoli et al., An unconventional circuit model for an efficient description of impedance and radiation features in printed-circuit leaky-wave structures, *IEEE Trans. Microwave Theory Tech.* **MTT-48**: 1661–1672 (2000).

24. C. N. Hu and C. K. C. Tzuang, Microstrip leaky-mode antenna array, *IEEE Trans. Antennas Propag.* **AP-45**: 1698–1699 (1997).

25. P. Baccarelli et al., Full-wave analysis of printed leaky-wave phased arrays, *Int. J. RF Microwave Comput. Aid. Eng.* (in press).

26. T. Yoneyama, Nonradiative dielectric waveguide, in K. J. Button, ed., *Infrared and Millimeter-Waves*, Academic Press, New York, 1984, Vol. 11, pp. 61–98.

27. A. Sanchez and A. A. Oliner, A new leaky waveguide for millimeter waves using nonradiative dielectric (NRD) waveguide — Parts I and II, *IEEE Trans. Microwave Theory Tech.* **MTT-35**: 737–752 (1987).

28. J. A. G. Malherbe, An array of coupled nonradiative dielectric waveguide radiators, *IEEE Trans. Antennas Propag.* **AP-46**: 1121–1125 (1998).

29. F. Schwering and S. T. Peng, Design of dielectric grating antennas for millimeter-wave applications, *IEEE Trans. Microwave Theory Tech.* **MTT-31**: 199–209 (1983).

30. M. Ghomi, B. Lejay, J. L. Amalric, and H. Baudrand, Radiation characteristics of uniform and nonuniform dielectric leaky-wave antennas, *IEEE Trans. Antennas Propag.* **AP-41**: 1177–1186 (1998).

31. S. Kobayashi, R. Lampe, R. Mittra, and S. Ray, Dielectric-rod leaky-wave antennas for millimeter-wave applications, *IEEE Trans. Antennas Propag.* **AP-29**: 822–824 (1981).

32. A. Henderson, A. E. England, and J. R. James, New low-loss millimeter-wave hybrid microstrip antenna array, *Proc. 11th Eur. Microwave Conf.*, 1981, pp. 825–830.

33. M. Guglielmi and A. A. Oliner, Multimode network description of a planar periodic metal-strip grating at a dielectric interface — Parts I and II, *IEEE Trans. Microwave Theory Tech.* **MTT-37**: 534–552 (1989).

34. T. N. Trinh, R. Mittra, and R. J. Paleta, Horn image-guide leaky-wave antenna, *IEEE Trans. Microwave Theory Tech.* **MTT-29**: 1310–1314 (1981).

35. D. R. Jackson and N. G. Alexopoulos, Gain enhancement methods for printed circuit antennas, *IEEE Trans. Antennas Propag.* **AP-33**: 976–987 (1985).

36. D. R. Jackson and A. A. Oliner, A leaky-wave analysis of the high-gain printed antenna configuration, *IEEE Trans. Antennas Propag.* **AP-36**: 905–910 (1988).

37. H. Ostner, J. Detlefsen, and D. R. Jackson, Radiation from one-dimensional dielectric leaky-wave antennas, *IEEE Trans. Antennas Propag.* **AP-43**: 331–339 (1995).

38. L. Borselli, C. Di Nallo, A. Galli, and S. Maci, Arrays with widely-spaced high-gain planar elements, *1998 IEEE AP-S Int. Symp. Dig.*, 1998, pp. 1446–1449.

39. C. Di Nallo, F. Frezza, A. Galli, and P. Lampariello, Analysis of the propagation and leakage effects for various classes of traveling-wave sources in the presence of covering dielectric layers, *1997 IEEE MTT-S Int. Microwave Symp. Dig.*, 1997, pp. 605–608.

40. C. A. Balanis, *Antenna Theory: Analysis and Design*, Wiley, New York, 1997, Chap. 16.

LEO SATELLITE NETWORKS

THOMAS R. HENDERSON
Boeing Phantom Works
Seattle, Washington

1. INTRODUCTION

Since the mid-1960s, most communications satellites have been deployed in a geostationary orbit, so named because the satellite appears to an earth-bound observer to remain nearly fixed in the sky. The geostationary orbit is a circular equatorial orbit at an altitude of 35,786 km, in which the angular velocity and direction of the satellite matches the angular rate of the rotation of the earth's surface. Satellites in this orbit provide telecommunications trunking services, VSAT (very-small-aperture terminal) data networks, direct-to-home television broadcasts, and even mobile services.

Although the very first satellites were launched into low orbits (since lower orbits were cheaper and less risky to attain), the convenience of geostationary orbits soon became the dominant factor in orbit selection. However, the latter half of the 1990s witnessed a renewal of interest in deploying communications satellites in orbits much closer to the earth [hence the term *low-earth-orbit* (LEO)], driven by the desire to extend voice, low-speed data, and Internet access services to mobile or remote users. Satellites at lower orbits have the drawback that they do not appear fixed in the sky. To provide continuous coverage within a given service region, more than one satellite (i.e., a network or *constellation* of satellites) is needed. Several such commercial systems have already been deployed (most notably the Iridium [1] and Globalstar [2] systems),

and even more ambitious systems have been proposed. Satellite constellations at lower orbits offer the following primary advantages:

- The end-to-end propagation delays can be significantly lower, thereby improving the quality of service provided to voice-based and data applications.
- Advances in cellular telephony electronics for handheld devices have enabled truly handheld satellite terminals equipped with small low-gain antennas, reachable by satellites at these lower orbits.
- As the orbital altitude increases, it is necessary to use larger antennas onboard to support the small spot beam (i.e., cell) sizes required for large system capacities.

This article summarizes the key issues regarding satellite networks employing LEO or other nongeostationary orbits. We first describe the various orbital geometries available to the satellite system designer. Next, we highlight several differences between satellite systems designed using geostationary (GSO or GEO) and nongeostationary orbits (non-GSO). Finally, we describe networking issues that arise from the need to use many satellites over time to serve terminals on the ground.

Our focus is on satellite networks that provide *continuous communications services* to a given service region. Therefore, we will not be explicitly focusing on store-and-forward satellite communications networks, or on satellites used for remote sensing, position determination, or military purposes, although many of the same principles apply.

2. SATELLITE ORBITS

2.1. Basic Orbital Geometry

To first order, a satellite's orbit can be described by an ellipse lying in a fixed orbital plane, with the earth's center positioned at one of the foci of this ellipse. As the satellite proceeds around this orbit, the earth rotates underneath it. The combination of the earth's rotation and the satellite's movement within the orbital plane contribute to its apparent motion in the sky as viewed from earth. The shape of the orbit is defined by its eccentricity (e) and its semimajor axis (a). The point at which the satellite is furthest from the earth is known as the *apogee* and, conversely, the closest point is the *perigee*. Additionally, there are three parameters that describe the orientation of this ellipse with respect to the earth. The right ascension of the ascending node (Ω) is a positive angle measured in the equatorial plane between two directions — a reference direction in the coordinate system, and the direction of the ascending node. The reference direction is given by intersection of the equatorial plane and the plane of the ecliptic, and is known as the direction of vernal equinox.[1]

[1] This reference direction maintains a fixed orientation in space with time and is so named because passes through both the earth and the sun on the vernal (spring) equinox.

The *ascending node* is the point of intersection between the orbital plane and the plane of the equator, the satellite crossing this plane from south to north. The inclination (i) is the positive angle between the normal to the direction of ascending node (pointed toward the east) in the equatorial plane and the normal to the line of nodes (in the direction of the velocity) in the orbital plane. The inclination can range from 0° to 180°; orbits with inclination greater than 90° are called *retrograde* orbits. The argument of perigee (ω) defines how the elliptical orbit is oriented in the plane. It is defined as the positive angle in the orbital plane between the direction of ascending node and direction of perigee (w ranges from 0° to 360°). A sixth parameter, the time of perigee passage (τ), defines the position of the satellite within this orbit (i.e., it specifies an initial condition). The period of the orbit (T) is given by the following relationship

$$T = 2\pi \left(\frac{a^3}{\mu}\right)^{1/2} \quad \text{(s)} \tag{1}$$

where $\mu = 3.9866 \times 10^{14}$ m^3/s^2 is the gravitational parameter for the earth, and a is the semimajor axis. Figure 1 illustrates these orbital parameters.

There are a variety of orbital perturbations (asymmetry of terrestrial gravitational potential due to the earth's oblateness, solar radiation pressure, solar and lunar gravitational influences, and atmospheric drag) that cause the actual orbit to deviate from this idealized model. To counteract such perturbations, satellites periodically apply controlled motor thrusts in a process known as *station keeping*. Station keeping requires the storage onboard of excess fuel reserves (such as pressurized nitrogen), the quantity of which may determine the operating lifetime of the satellite since they are not replenishable. To ensure that the satellite constellation geometry can remain fixed in the face of such

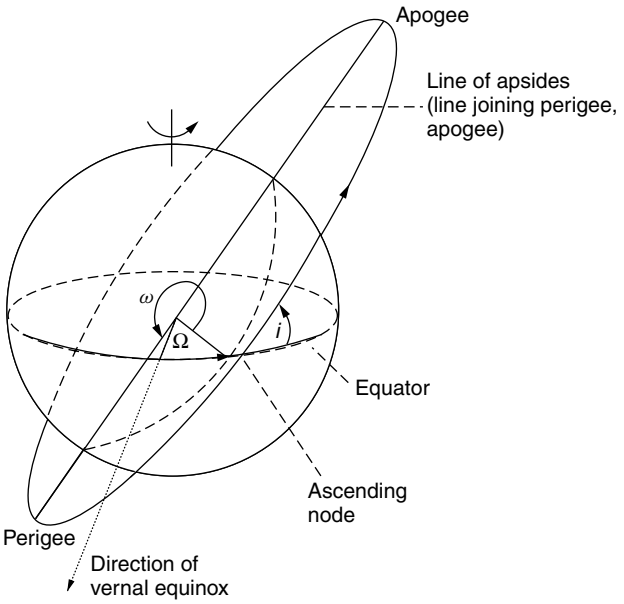


Figure 1. Illustration of Keplerian orbital parameters that define a satellite orbit (from Pattan [3], used with permission).

perturbations, all satellites in a given constellation should assume the same inclination and altitude [3]. However, even under these preferred geometries, the network operator may choose not to expend the fuel required to maintain the relative distances between satellites in the same orbital plane or in different orbits. Constellations that maintain these relationships are known as *phased* constellations, and provide more optimal solutions to the network design, at the expense of requiring larger satellites and a control station network [4]. The book by Maral and Bousquet [5] provides an extensive treatment of orbital perturbations.

2.2. Types of Orbit

In principle, satellites may be deployed into any orbit, but there are practical and economic considerations that favor certain orbit types over others. The choice of orbital configuration and number of satellites is the result of a system optimization combining a large number of factors that we will summarize shortly. As far as the orbits themselves, the International Telecommunications Union (ITU) has defined three broad classifications for nongeostationary orbits:

- *Low-Earth Orbit* (*LEO*). LEO satellite orbits lie between roughly 700 and 1500 km. The lower altitude bound is governed by limits on atmospheric drag, while the upper bound is the approximate beginning of the inner Van Allen radiation belt.[2] LEO orbits are typically circular. For coverage of the entire earth's surface, a near-polar inclination can be selected — this, however, causes a concentration of coverage near the poles. If the inclination angle is relaxed, a higher degree of system capacity can be concentrated at midlatitudes, at the expense of polar coverage.
- *Medium-Earth Orbit* (*MEO*). Systems that lie between the two Van Allen radiation belts (between 5000 and 13,000 km), or above the outer belt (greater than 20,000 km) are typically classified as MEO satellite systems. The term *intermediate circular orbit* (ICO) is also sometimes used when the orbit is circular. Because of their use of a higher altitude, MEO systems require fewer satellites than do LEO systems to provide similar coverage. For example, the Iridium (LEO) system uses 66 active satellites for global coverage, while the commercially proposed ICO constellation [6] requires only 10.
- *Highly Elliptical Orbit* (*HEO*). A third option has been to use elliptical, inclined orbits. The key property of an elliptical orbit is that the velocity of the orbit is not constant but instead is slowest at the orbit apogee.[3] Therefore, satellites in such orbits can remain visible for longer stretches of time if the apogee is situated over the desired region of coverage. Furthermore, unlike GSO satellites, HEO satellites can serve latitudes higher than 75°. One drawback to elliptical orbits is that the oblateness of the earth, and the resulting anomalies in the gravitational field, causes the apogee to rotate slowly around the orbit (a phenomenon known as *apsidal rotation*). There are, however, two orbital inclinations (63.4° and 116.6°) for which no apsidal rotation occurs. One such orbit, known as the *Molnya* (lightning) orbit, uses an inclination of 63.4°. Molnya orbits, pioneered by the former Soviet Union, have an apogee at roughly 40,000 km, a perigee of about 1000 km, an argument of perigee of about 270°, and a period of 12 h. By using multiple satellites in such orbits and ground stations that can track the slowly moving satellites, communications at high latitudes can be enabled with a high elevation angle over the horizon. Note that these orbits must pass through the Van Allen radiation belts. Typically, this requires more radiation shielding of the electronics and results in a shorter satellite lifetime; as a result, variations to this orbit that do not require crossing the radiation belts have been studied.

2.3. Coverage

The maximal satellite *footprint*, or coverage area, is governed by the altitude above the earth's surface and the minimum elevation angle supported. Details on the geometry of this relationship are covered by Maral and Bousquet [5]. The actual coverage area may be smaller if the antenna pattern is more focused on a smaller surface area. Furthermore, the coverage area is usually segmented into a collection of smaller *spot beams*. This is done primarily for two reasons: (1) as in cellular networks, the overall system capacity can be increased through frequency reuse — for example, in the Iridium system, the satellite footprint is divided into 48 smaller spot beams, with a frequency reuse factor of 12 [1]; and (2) the communications link performance is inversely related to the spot size illuminated, because smaller spot beams result in more focused RF carrier power. The costs of supporting smaller spot beams include larger aperture antennas on board the satellite, more frequent link handoffs for terminals, and a more sophisticated payload to route traffic to the correct spot beam if onboard switching is performed.

2.4. Constellation Design

Satellite constellations are typically designed based on a requirement of having one or more satellites continuously in view of earth stations (above some minimum elevation angle) throughout a given service area. One of the main objectives is to minimize the number of satellites needed

[2] The Van Allen radiation belts consist of two toroidally shaped regions around the earth's magnetic equator where highly charged particles are trapped by the magnetic field. The inner belt lies between approximately 1500 and 5000 km, and the outer belt between 13,000 and 20,000 km. It is preferable to avoid prolonged exposure to such regions because of damaging effects on solid-state electronics.

[3] This is a consequence of Kepler's second law of planetary motion, which states that the radius vector of the orbit sweeps out equal areas in equal times (the "law of areas").

to meet this requirement. Walker originally explored different types of constellations using circular orbits [7], which are generally classified into two categories. The first category, constellations with orbits using near-polar inclination (sometimes called *Walker star* or *polar* constellations), have the property that the ascending nodes of the orbits are regularly distributed over a hemisphere (180°). As a result, there are two hemispheres in which all the satellite orbits are corotating, in either a north–south or south–north direction. The division between these hemispheres of coverage, across which the satellite orbits are counterrotating, is commonly called a *seam*. Although this type of constellation has a concentration of coverage at the poles, it efficiently covers the lower latitudes, and has the desirable property that satellites in corotating planes move slowly with respect to one another, allowing for easier establishment of intersatellite communications links between them. The Iridium constellation uses a design of this type, as illustrated in Fig. 2.

The second category of circular-orbiting constellations, known as *Walker delta* or *rosette* constellations, have the ascending nodes distributed uniformly across 360° of longitude, with the orbits all at the same inclination. The result of this design is that any area of the earth's surface has both ascending and descending satellites. This type of constellation design is most commonly applied when the inclination angle is relaxed below 90°, so that coverage can be concentrated at the populated midlatitudes. When the satellites are connected via intersatellite links, this constellation design also offers networking path diversity not achievable with polar constellations [9]. Globalstar uses a rosette constellation design, as illustrated in Fig. 3. The book by Pattan provides further details on polar and rosette constellation designs [3].

As emphasized above, nongeostationary satellites are not limited to circular orbits. Draim has derived optimal constellation geometries based on elliptical orbits, the principles of which have been incorporated into the proposed Ellipso constellation [10]. We have already introduced the Molnya orbit as an example of a highly elliptical orbit. Another elliptical orbit known as the *Tundra* orbit shares the same orbital inclination of 63.4° but with a period of 24 h. Since the visibility of each Tundra satellite beneath the apogee is greater than 12 h, two satellites in such orbits are sufficient for continuous coverage [5].

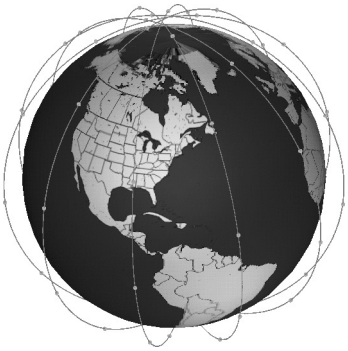


Figure 2. Orbital geometry for the Iridium constellation, an example of a polar-orbiting constellation (from Thurman and Worfolk [8], used with permission).

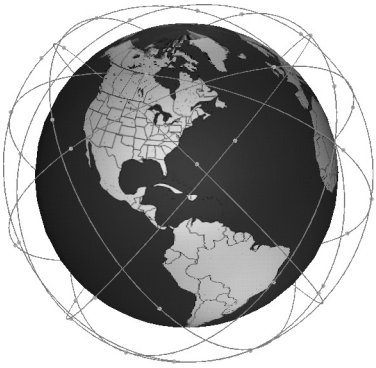


Figure 3. Orbital geometry for the Globalstar constellation, an example of a rosette constellation (from Thurman and Worfolk [8], used with permission).

3. GEOSTATIONARY VERSUS NONGEOSTATIONARY SATELLITES

In this section, we highlight some of the distinguishing properties of nongeostationary satellites by contrasting them with geostationary satellites.

3.1. Propagation Delay

One of the reasons most frequently cited for moving to lower-earth orbits is the resultant reduction in propagation delay. For GSO satellites, the one-way propagation delay from the earth to the satellite is between 120 and 140 ms. This makes the round-trip delay based on propagation delay alone somewhere between 480 and 560 ms, and additional delays for coding, queueing, and multiple access typically push this number above 600 ms. For voice traffic, this amount of delay tends to disrupt the rhythms of conversation and can lead to annoying echos when analog phones are involved. Furthermore, delays of this magnitude can seriously compromise the throughput performance of Internet transport protocols, as well as affect interactive data applications. However, for other applications such as broadcasting, the delay is not important. Similar delays are experienced when using satellites in the Molnya orbit.

By moving satellites to lower orbits, the resultant round-trip delays can be as low as 10–20 ms — similar to what is experienced over the wide-area Internet. However, it must be emphasized that this is only a lower bound; the actual delay is a function of distance between terminals and the number of satellites traversed and can vary over time as the relative positions change. Furthermore, at low link rates, the delays to access the channel (multiple access) can be a significant component. Analysis of the Iridium system has shown that the average end-to-end delay encountered is roughly 100–200 ms, with a large portion of this delay due to multiple access [1]. Moreover, while the delay is constantly varying due to orbital motions, it also is subject to step changes as link handoffs cause a reconfiguration of the communications

path. Nevertheless, as long as end-to-end delays can be kept in the region of 100 ms or less, the qualitative performance improvement for both voice and data traffic can be substantial.

3.2. Link Performance Issues

The error performance of a radio link is a function of the carrier power transferred (by the antenna) to the receiver input, interference from undesired signals also captured by the antenna, and noise power within the receiver. Many books on satellite communications provide an extensive treatment of link budget issues [e.g., 5]. Many of the same principles apply to nongeostationary channels, but there are some significant differences. Geostationary satellite links providing service to fixed terminals can generally be modeled as additive white Gaussian noise (AWGN) channels, with a fixed free-space path loss. In contrast, LEO channels are subject to Rician fading (particularly if low-gain antennas are used by the terminal), high Doppler shifts, variable path loss (unless compensated for by the spacecraft), and irregular terrestrial interference [11,12]. A popular model for the LEO channel, combining Rice and lognormal statistics, is due to Corazza [13]. Because of the severe fading, researchers have shown the benefit of satellite diversity in improving overall system availability and capacity [14].

3.3. Interference

Geostationary satellites using the same frequency bands and carrier polarizations are typically spaced about two to four degrees apart in the orbital arc. For fixed satellite service, this dictates a minimum size of the terrestrial antenna such that the desired pattern directivity can be maintained. Satellite systems can therefore share frequency bands by exploiting these fixed geometries and directional antennas. The situation becomes considerably more complex when satellites appear to move in the sky, and when mobile handsets (with low directivity antennas) are being used. Satellites and terminals from LEO and GSO systems using the same frequencies are likely to interfere with one another as their relative geometries change. While this can be alleviated by placing LEO and GSO systems at different frequency bands, LEO systems attempting to share the same frequency bands are still likely to interfere with one another. Two possible solutions to this problem are to employ spread-spectrum modulation techniques using code sets with low cross-correlation, or to simply divide the available spectrum among users and have each operate an independent system. For systems with user terminals employing low-gain antennas, the current consensus seems to be that spectrum separation is required, while the jury is still out for broadband systems using terminals with highly directive antennas. In summary, the international regulatory procedures required to operate a non-GSO system are considerable.

3.4. Frequencies

Because satellite orbits and frequencies do not belong to any nationality exclusively, their use is coordinated by the ITU. It should be emphasized that the particular allocations change over time and the details are complicated, so we will simply provide an overview of the main frequency bands herein. Briefly, the ITU has established that geostationary satellite links use frequencies found mainly in the L (roughly 1.5 GHz downlink, 1.6 GHz uplink), C (4/6 GHz), Ku (12/14 GHz), and Ka (20/30 GHz) bands. The ITU further classifies satellite systems as providing either fixed satellite service (FSS) (to fixed terminals on the ground), broadcast satellite service (BSS), or mobile satellite service (MSS). Frequency allocations for nongeostationary FSS systems are found in the same general frequency bands used by geostationary satellites. For mobile satellite service, links carrying subscriber traffic can be categorized as either *feeder links* or *subscriber links*. Feeder links connect the satellites to a gateway earth station; these types of links also have been allocated frequencies in the C, Ku, and Ka bands. However, because of the low-gain antennas typical of mobile handsets, there is a strong incentive to use as low a frequency as possible for link performance issues. Specifically, roughly 4 MHz of spectrum in the VHF/UHF bands (around 150 and 400 MHz) and 32 MHz of spectrum in the L band (1.6/2.5 GHz) are allocated to nongeostationary MSS systems.

In the United States, the spectrum in the VHF/UHF bands has been set aside for low-data-rate data systems. Such systems have been coined "little LEOs"; an example is the Orbcomm system used for paging and short data messaging. "Big LEO" systems such as Iridium and Globalstar use the L-band frequencies and are permitted to offer both voice and data services. LEO systems offering broadband data rates will use frequencies in the Ku and Ka bands, or even higher frequencies. The book by Pattan [3] discusses the various frequency allocations in more detail.

3.5. Launch and Spacecraft

As the orbital altitude is increased, the cost of deploying a satellite into that orbit also increases. The geostationary orbit is expensive to attain, requiring a multistep approach. The first step involves placing the satellite into a circular low-earth orbit, then into an elliptical geostationary transfer orbit (where the apogee of this orbit corresponds to the altitude of the geostationary orbit), then into its final orbit. Since only a small fraction of the mass deployed at low-earth orbit is eventually deployed in the final orbit (the rest is fuel), the cost penalty to achieve geostationary orbit is substantial. In contrast, LEO satellites can be launched directly to their final orbital altitude, and for small satellites, multiple satellites can often be launched using the same vehicle. However, while the cost of launching an individual LEO satellite is cheaper than a GSO satellite, the cost of launching a whole constellation is seldom. At least one launch is typically required for each orbital plane of the constellation.

A detailed treatment in the difference between geostationary and nongeostationary spacecraft is beyond this article's scope; the interested reader is directed to the overview in Ref. 5. However, we note that since LEO satellites are closer to the earth, they more frequently undergo shadowing by the earth, and therefore

are subjected to frequent thermal stresses and require batteries to continue to operate while within the shadow. LEO satellites can also be smaller in some dimension (antenna size or transmit power) than GSO satellites while still providing an equivalent RF carrier flux density on the ground. However, the perception that LEO satellites are much smaller than GSO satellites is not necessarily valid in general, but rather is due to the initially deployed LEO systems using a small amount of spectrum. The size of a satellite is directly related to the power required, which is directly related to the throughput; consequently, the proposed broadband LEO satellites plan to use very large satellites.

3.6. Tracking and Link Handoff

Handoff (also known as *handover*) is defined as the procedure for changing the radio communications path to maintain an active communications session. The most significant challenge for a nongeostationary satellite system is the need to track satellites and to hand off active communications links from one satellite to another or between different beams of the same satellite. The rate at which a ground terminal must hand off a connection between satellites (*intersatellite* handoff) varies with the altitude, ranging from 12 h (HEO orbits of type Tundra) to 10 min (LEO). However, handoffs can be even more frequent if the coverage area of the satellite is further segmented into spot beams. For example, the Iridium satellites employ 48 spot beams within the coverage area of one satellite. In this case, handoff between beams on the same satellite can occur every minute or two.

Beam handoffs typically require a change in carrier frequency (unless spread-spectrum modulation is used) and acquisition of the new link. Intersatellite handoffs may require the additional step of repointing the terminal's antenna, which could cause an interruption of service. Such an interruption may be avoided in one of several ways. One brute-force method is to equip the terminal with two mechanically or electronically steered antennas, and engage the nonactive antenna in finding the next satellite. Depending on the service, if a single electronically steered beam is used, the switchover may occur rapidly enough, especially if the approximate position of the next satellite is known by the terminal.

Satellite antenna patterns are typically nadir-pointing, which means that the pattern drags across the surface of the earth with a constant velocity. As a result, handoffs are asynchronous in the system — there will always be some subset of user terminals in the process of handing off at any given time. A proposed alternative would be to electronically or mechanically steer the antenna on board the satellite to keep the coverage fixed on the earth's surface until some point in time at which all of the patterns synchronously switch. This proposed technique would reduce or eliminate intrasatellite handoffs and would cause all intersatellite handoffs to occur synchronously, thereby simplifying the algorithms that deal with handoff [15].

3.7. Intersatellite Links

At low orbital altitudes, the satellite footprint may be relatively small. In a communications session, if both ground terminals are not within the same footprint, some means of transmitting signals between the satellites is necessary. If there are gateway stations located in each satellite's footprint, then one solution is to route traffic from an earth station to a gateway in each footprint, and then to use landlines to interconnect the gateways. Such an approach, while greatly simplifying the satellite payload, has the drawback of requiring a large network of ground-based gateway stations interconnected by terrestrial links.[4]

An alternative solution is to use communications links to interconnect the satellites themselves. These links, known as *intersatellite links* (ISLs), create a mesh network in the sky, and obviate the need to have gateway stations in every coverage footprint (note that this advantage diminishes for MEO/ICO satellites, which have broader coverage footprints). Each Iridium satellite, for example, has ISLs to the two closest satellites within the same orbital plane (black lines illustrated in Fig. 4), and either one or two links to the nearest neighboring satellite in an adjacent plane (lighter lines in Fig. 4). The drawbacks to using ISLs are an increase in complexity of the satellite payload, the establishment and maintenance of such links, as well as the requirement to route traffic between satellites.

The frequencies allocated for ISLs correspond to strong absorption by the atmosphere (to protect against terrestrial interference). Selected radio links at frequencies between 23 and 58 GHz and optical wavelengths between 0.8 and 10.6 μm may be used. For high-capacity links, optical link hardware requires less mass and power consumption.

ISLs require steerable antennas for link pointing, acquisition, and tracking. ISLs between satellites in the same orbiting plane (known as *intraplane* ISLs) do not require tracking in a phased constellation, because the orbital relationship between such satellites is fixed. ISLs that connect satellites in different orbital planes (*interplane* ISLs) will require tracking. The pointing requirements depend strongly on the constellation design. As an example, the Iridium constellation requires a pointing range of roughly 10° in the vertical direction and 140° in the horizontal direction [16]. Furthermore, the pointing angles may become so severe that ISLs will need to be deactivated for a portion of an orbit, or handed off to another satellite. This condition holds in the high-latitude regions of polar-orbiting constellations. Finally, ISLs may be handed off from one satellite to another if the relative locations of the satellites change with respect to one another; for example, ISLs connecting satellites across the seam of a polar-orbiting constellation. As we discuss in the next section, such ISL link changes have implications on network routing.

4. NETWORKING CONSIDERATIONS FOR SATELLITE CONSTELLATIONS

Satellite constellations are considerably more complicated than geostationary satellites from a networking

[4] This approach is used by the Globalstar system.

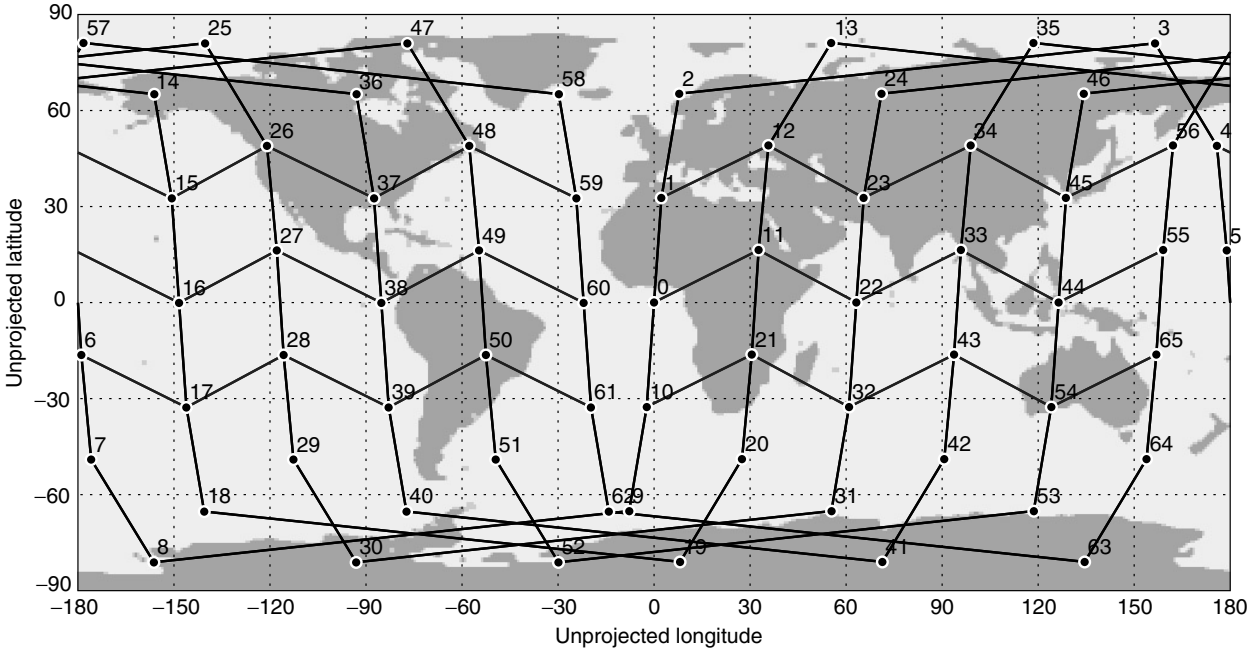


Figure 4. Snapshot of Iridium satellites and active ISLs on an unprojected map of the earth's surface. Lighter lines indicate interplane ISLs; black lines denote intraplane ISLs.

standpoint. Geostationary satellites were originally used as repeaters ("bent pipes") for fixed or semipermanent communications channels. With the desire to share bandwidth among many users, multiple-access protocols such as time-division multiple access (TDMA) and ALOHA[5] were developed. VSAT networks, for example, rely on the coordinated sharing of uplink capacity among hundreds or thousands of terminals. More recently, the construction of multibeam satellites has led to onboard switching between transponders. Only relatively recently have satellite payloads with demodulation, baseband processing, and signal regeneration begun to be deployed. Nevertheless, even with sophisticated onboard processing, geostationary satellites are still not much more than a (fixed location) switch in the sky.

In contrast, satellite constellations are an interesting variant of mobile networks — one in which the network nodes move and the terminals stay (relatively) fixed. There is a mixture of permanent (intraplane ISL), semipermanent (interplane ISL), and transient (ground-to-satellite) links. Unlike traditional mobile networks, however, the network topology is somewhat regular in structure, and many topological changes can be predicted in advance. The regularity and predictability of the network geometry can be exploited in the network architecture design.

In this section, we survey the different approaches that have been proposed for networking in satellite constellations. Since satellite networks are designed to extend services of terrestrial networks, it should not be surprising that satellite network architectures can generally be classified as either circuit switching or packet switching. Before examining both of these approaches in turn, we first note that some common architectural principles apply to both types of networks. The first principle is *flexibility*. Satellite networks are expected to last many years and are difficult or prohibitively expensive to upgrade in space. Therefore, system designers strive to implement general solutions for the space segment that will not become obsolete. For example, packet-switched satellite architectures will likely not implement pure IP (Internet Protocol) packet switching in space, but instead will strive for a generic, satellite-optimized packet switching infrastructure into which IP and other protocols can be mapped. A second related principle is *simplicity*, which argues for deploying functions, when possible, within the ground segment so as to relieve the onboard complexity of the electronics, and hence the mass and power requirements of the spacecraft.

4.1. Circuit-Switched Architectures

Many LEO and MEO satellite systems have been positioned to offer traditional PSTN services (voice, low-bit-rate data, facsimile, paging, etc.). The architectures for these networks typically resemble those of "second generation" mobile communications systems such as GSM. The functions required of the network include basic routing and switching of the call, mobility management, privacy, security, and radio resource management. The chief difference between satellite constellations and traditional mobile networks lies in the mobility management function. As mentioned above, the network nodes, as well as the terminals, in fact move. This has implications on location registration and handoff [17].

Location registration is the procedure by which mobile units are identified as being in a particular location so

[5] ALOHA was developed by N. Abramson at the University of Hawaii in 1970.

that calls can be correctly routed. In mobile satellite networks, this may be network-based or terminal-based. Terminal-based solutions typically rely on signals from the Global Positioning System (GPS); an accurate approach that has the downside of only working when several GPS satellite signals can be received. Network-based solutions rely on the estimation of location on the basis of timing, arrival angle, and signal strength from one or more base stations. This approach provides less accurate position information. Both of these approaches may fail in dense urban canyons. The need for precise terminal location information depends on the amount that this information is used by handoff and routing algorithms to predict the future evolution of the network topology. Of course, satellite position information can be known accurately due to the predictable movement of the satellites and telemetry, tracking, and control (TTC) communications links with ground stations.

4.1.1. Handoffs. Handoff management is a key determinant in the performance of LEO satellite networks. We have already discussed some of the issues regarding handoff. In general, handoffs should be optimized (which in many situations means minimized), since they are typically accompanied by signaling and call processing overhead, and may result in a degradation in the call's quality of service, due to either blocking or suboptimal routing.

There are two types of handoff in a LEO satellite network. ISL handoffs are generally regular and predictable events. Terminal-to-satellite link-handoffs are not predictable with as much accuracy. Terminals may initiate handoffs to new spot beams or satellites by monitoring signal strength and interference from different carriers. When the terminal enters an overlap area, it requests a handoff. Conversely, in some systems such as Iridium, a gateway station may control the handoff between two satellites by instructing the leading satellite to prepare to handoff and the trailing satellite to prepare to accept the call [18].

Links that must be transferred from one beam to another are subject to blocking if insufficient resources are available in the new beam. Two common techniques for minimizing the possibility that an active call is dropped are to implement guard channels or to queue handover requests. Guard channels are a pool of channels explicitly reserved for handoffs (i.e., no new call arrivals can be assigned a guard channel). Queueing techniques prescribe that when the terminal is in an overlap area that it hold onto its existing channel until a channel in the new beam becomes available. On release of a channel, queued handover requests are served before new calls are admitted. Both approaches lead to fewer handover failures at the expense of a higher initial call blocking probability.

More sophisticated handoff techniques attempt to exploit predictable aspects of the constellation evolution. For example, if it can be estimated when calls will need to be handed off, active channels in one beam can be reserved to handle channels that are predicted to need handoff from another beam in the future [19]. If terminal locations are precisely known, connection admission control can be optimized by predicting the future spotbeam handoff path of each new call arrival [20].

Finally, if ISLs are not used, the use of CDMA for multiple access can yield link performance gains during handoffs. In the Globalstar system, since the same gateway station is used before and after intersatellite handoff, the so-called soft handoff technique of CDMA can be implemented. In this technique, the mobile terminal signal is passed through both the old and new satellites, and the two signals are independently demodulated, selected, and constructively combined to yield processing gain at the edges of satellite coverage.

4.1.2. Multihop Satellite Routing. Consider the establishment of a satellite-based connection traversing ISLs. This requires that a candidate route be picked, each node be signaled about the new connection, the connection be maintained even in the face of changes to the topology, and the connection resources be released once the call is completed. Note that, as exemplified by Fig. 4, there may be many similar routes (topologically) between distant stations. The process whereby a multihop satellite route is established and maintained is a routing problem.

First, consider the establishment of the initial route. In a traditional network, a shortest-path algorithm, perhaps weighted by the amount of current congestion at the nodes, would be used. In the satellite case, the additional consideration of *link permanence* can be applied. In this scenario, routes can be avoided for which it is likely that one of the constituent links is known to be short-lived (such as an ISL about to be deactivated or handed off). In order to minimize handoffs, researchers have studied techniques that consider the time-varying topology during route selection, favoring routes requiring fewer handoffs [21–23]. Another consideration that may be included in routing decisions involves accounting for nonuniform traffic densities in different areas [24,25].

Next, consider a terminal to satellite handoff, which are frequent in LEO constellations. If an ISL exists between the previous satellite and the new satellite, then this link can be grafted onto the existing route without disrupting the other nodes along the path [26]. Note, however, that the new route may no longer be optimal, and over time, may become grossly distorted. Satellite constellations, therefore, may consider this *route augmentation* as a preferred option so long as the resultant route does not fall below some threshold of optimality.

Finally, consider a topological change in which the route from ingress satellite to egress satellite cannot be maintained, or in which the augmented route becomes too suboptimal. In these cases, it may be necessary to determine a new route altogether, and inform the affected nodes along the paths to synchronously switch over at some time instant.

Note that these topological changes can cause the overall circuit delay to drastically change, which may be a problem for some services. One way to compensate for this is to use buffers at the endpoints of the satellite connections to smooth out any delay variations, at the expense of consistently larger delays.

4.2. Packet-Switched Architectures

An alternative to circuit switching, especially suited for interworking with actual packet-switched networks like the Internet, is a packet-switched satellite network. This approach has the advantage of not requiring per connection state to be kept and maintained on board the spacecraft. Nevertheless, many of the same handoff challenges described above still persist, because for reasons of link efficiency, channel reservations between terminals and satellites are still desirable.

There are several different techniques available for implementing packet routing in satellite networks, differing chiefly in their implementation and processing complexity onboard the satellites. A general discussion of several IP networking issues, including address translation, multicast, interfacing with exterior routing protocols, tunneling, and quality of service can be found in Ref. 27. In this section, we focus on the basic packet routing problem in a satellite constellation.

Consider the problem of routing a packet from one terminal to another through one or more satellite nodes. The simplest approach to route the packet from the standpoint of satellite complexity would be to flood the packet (i.e., transmit the packet out of all the active link interfaces except the one on which the packet arrived) and limit the number of hops for which the packet can be forwarded (such as by decrementing a counter). Because of the densely interconnected mesh, such an approach would be grossly suboptimal, leading to extremely congested networks. Another simple approach from the satellite standpoint would be to determine the entire route of the packet a priori at the ingress terminal, and affix this route to the packet before sending it to the first satellite node. Each satellite would then forward the packet by simply following the next-hop instructions attached to the packet, and an onboard routing table would not need to be maintained. This would require, however, that the terminals affixing the route to the packet determine the optimal route; that is, they must have access to full instantaneous routing state of the network. The burden placed on terminals on the network may be considerable in this case. An alternative could be to have route servers distributed throughout the network that could be queried by terminals whenever a new route was needed. However, this approach would incur extra latency in the initiation of the communications. A more serious impediment to this approach would occur if route topological changes were not predicted (such as a terminal-initiated handoff). In this case, packets could be lost to a deadend until new route information is made available to endpoints, and because of the latency in the system, it would take some time for the routing information to stabilize on any unanticipated topological change.

To offload the responsibility of routing from terminals to the satellite network, again, different approaches may be used. For example, centralized routing servers could periodically upload routing tables to satellites, incrementally updating them as topology changes become known. Again, there may be latency issues with this approach upon unexpected topological changes. This approach, while requiring the satellite to maintain memory for and lookup routes from a routing table, the task of actually computing the routing tables is left to the ground segment.

Latency issues in the propagation of state information can be minimized if the satellite nodes implement fully distributed routing, such as used in the Internet. The drawback to this approach is that it requires satellites to not only build and maintain routing tables but also incur the processing and signaling overhead of a distributed routing protocol. Indeed, terrestrial Internet routing protocols such as traditional distance vector or link-state protocols, applied to this type of a dynamic network topology, would either be slow to converge or would overwhelm nodes with update messages. General flooding of routing update messages would also be problematic, even if there were sufficient ISL capacity to handle the messages, due to the sheer volume of routing updates that would need to be processed. Nevertheless, distributed routing techniques have been the focus of most recent research, as researchers have studied ways to simplify the problem by exploiting the regularity of the network topology and the predictability of ISL topological changes.

One general technique for simplifying distributed routing is to try to hide the mobility of satellite nodes from the terrestrial nodes. The semiregular structure of most satellite constellations facilitates this. For example, if one overlays a cellular structure over the earth's surface, with the cell size roughly corresponding to the coverage area of a satellite footprint, then it may be possible to overlay a logical network structure of "virtual nodes," in which different satellites over time embody each virtual node [28]. Another possible approach is to assume that the satellite network evolves through a finite series of topologies, and have the satellite network store the appropriate routing table for each state and iterate through these tables [29]. Although such approaches appear to have some promise when considering idealized constellation geometries, they have yet to be demonstrated as a robust approach when applied to practical constellations [30].

5. FUTURE DIRECTIONS

The design of a satellite constellation is a complex optimization problem with the cost a function of various link parameters as well as terminal and satellite complexity. In this article we have provided an overview of nongeostationary satellite fundamentals and surveyed many of the design features that differentiate these networks from systems based on geostationary satellites. Unlike GSO systems, LEO and MEO satellite networks are still in their infancy, and several of the initial attempts to deploy large-scale commercial constellations have been a financial failure. Nevertheless, the promises of global ubiquitous coverage, accompanied by significantly lower propagation delays, will continue to spur development of nongeostationary satellite network architectures. Technically, many issues will be the subject of ongoing research and development, including interference mitigation, link issues (such as error control coding and handoff algorithms), electronically steerable antennas,

routing algorithms, onboard switching architectures, electronics based on more radiation-resistant substrates (such as gallium arsenide), and regulatory issues.

BIOGRAPHY

Thomas R. Henderson received his B.S. and M.S. degrees from Stanford University, Stanford, California, and a Ph.D. from the University of California, Berkeley, California, all in electrical engineering. He is currently a researcher at Boeing Phantom Works, the research and development division of The Boeing Company. He is also presently a part-time lecturer in the Electrical Engineering department at the University of Washington, Seattle, Washington. Previously, he was director of digital television research and standards at Geocast Network Systems in Menlo Park, California. Prior to attending Berkeley, he worked at COMSAT Laboratories in Maryland, where his responsibilities included performance analysis, protocol development, and standards activities in the areas of ATM and ISDN over satellite networks. He has been a rapporteur of ITU-T Study Group 13, a vice chair of ANSI T1S1.5, and editor of several national and international standards. His current research interests are focused on network-layer mobility and routing for wireless IP networks.

BIBLIOGRAPHY

1. S. R. Pratt, R. E. Raines, C. E. Fossa Jr., and M. A. Temple, An operational and performance overview of the IRIDIUM low Earth orbit satellite system, *IEEE Commun. Surv.* Second Quarter: **2**(2): 2–8 (1999).
2. E. Hirshfield, The Globalstar system: Breakthroughs in efficiency in microwave and signal processing technology, *Space Commun.* **14**: 69–82 (1996).
3. B. Pattan, *Satellite-Based Cellular Communications*, McGraw-Hill, New York, 1998.
4. G. Maral, J.-J. De Ridder, B. G. Evans, and M. Richharia, Low Earth orbit satellite systems for communications, *Int. J. Satellite Commun.* **9**: 209–225 (1991).
5. G. Maral and M. Bousquet, *Satellite Communications Systems*, Wiley, Chichester, UK, 2000.
6. L. Ghedia, K. Smith, and G. Titzer, Satellite PCN — the ICO system, *Int. J. Satellite Commun.* **17**: 273–289 (1999).
7. J. Walker, Some circular orbit patterns providing continuous whole earth coverage, *J. Br. Interplan. Soc.* **24**: 369–381 (1971).
8. R. Thurman and P. Worfolk (No date), SaVi (online), *http://www.geom.umn.edu/worfolk/SaVi/*, April 4, 2001.
9. L. Wood, *Internetworking with Satellite Constellations*, Ph.D. thesis, Univ. Surrey, 2001.
10. J. E. Draim, Design philosophy for the ELLIPSO™ satellite system, *Proc. 17th AIAA Int. Comm. Satellite Conf.*, Yokohama, Japan, 1998.
11. I. Ali, N. Al-Dhahir, and J. E. Hershey, Doppler characterization for LEO satellites, *IEEE Trans. Commun.* **46**: 309–313 (1998).
12. F. Vatalaro and G. E. Corazza, Probability of error and outage in a Rice-lognormal channel for terrestrial and satellite personal communications, *IEEE Trans. Commun.* **44**: 921–924 (1996).
13. G. E. Corazza and F. Vatalaro, A statistical model for land mobile satellite channels and its application to nongeostationary orbit systems, *IEEE Trans. Vehic. Technol.* **43**: 738–742 (1994).
14. G. E. Corazza and C. Caini, Satellite diversity exploitation in mobile satellite CDMA systems, *Proc. Wireless Comm. Networking Conf.* (*WCNC*), 1999, pp. 1203–1207.
15. J. Restrepo and G. Maral, Cellular geometry for world-wide coverage by non-GEO satellites using "Earth-fixed cell" technique, *Space Commun.* **14**: 179–189 (1996).
16. M. Werner, A. Jahn, E. Lutz, and A. Bottcher, Analysis of system parameters for LEO/ICO-satellite communication networks, *IEEE J. Select. Areas Commun.* **13**: 371–381 (Feb. 1995).
17. F. Ananasso and M. Carosi, Architecture and networking issues in satellite systems for personal communications, *Int. J. Satellite Commun.* **12**: 33–44 (1994).
18. Y. C. Hubbel, A comparison of the IRIDIUM and AMPS Systems, *IEEE Network Mag.* **11**: 52–59 (1997).
19. P. Wan, V. Nguyen, and H. Bai, Advance handovers arrangement and channel allocation in LEO satellite systems, *Proc. IEEE Globecom*, 1999, pp. 286–290.
20. S. R. Cho, I. F. Akyildiz, M. D. Bender, and H. Uzunalioglu, A new spotbeam handover management technique for LEO satellite networks, *Proc. IEEE Globecom*, 2000, pp. 1156–1160.
21. M. Werner et al., ATM-based routing in LEO/MEO satellite networks with intersatellite links, *IEEE J. Select. Areas Commun.* **15**: 69–82 (1997).
22. A. Jukan, H. N. Nguyen, and G. Franzl, QoS-based routing methods for multihop LEO satellite networks, *Proc. IEEE Int. Conf. Networks* (*ICON*), 2000, pp. 399–405.
23. H. Uzunalioglu, Probabilistic routing protocol for low earth orbit satellite networks, *Proc. IEEE Int. Conf. Commun.* (*ICC*), 1998, pp. 89–93.
24. A. Jamalipour, *Low Earth Orbital Satellites for Personal Communication Networks*, Artech, Norwood, MA, 1998.
25. Y. Kim and W. Park, Adaptive routing in LEO satellite networks, *IEEE Vehicular Tech. Conf.*, 2000, pp. 1983–1987.
26. H. Uzunalioglu and W. Yen, Managing connection handover in satellite networks, *Proc. ACM Mobicom*, 1997, pp. 204–214.
27. L. Wood et al., IP routing issues in satellite constellation networks, *Int. J. Satellite Commun.* **19**: 69–92 (2001).
28. R. Mauger and C. Rosenberg, QoS guarantees for multimedia services on a TDMA-based satellite network, *IEEE Commun. Mag.* **35**: 56–65 (1997).
29. H. S. Chang et al., Topological design and routing for LEO satellite networks, *Proc. IEEE Globecom*, 1995, pp. 529–535.
30. T. R. Henderson and R. H. Katz, On distributed, geographic-based packet routing for LEO satellite networks, *Proc. IEEE Globecom*, 2000, pp. 1119–1123.
31. L. Wood (n.d.), Lloyd's satellite constellations (Online), *http://www.ee.surrey.ac.uk/Personal/L.Wood/*, April 4, 2001.

LINEAR ANTENNAS

SHELDON S. SANDLER
Lexington, Massachusetts

An antenna constructed of a few straight-line segments, made of conducting, partially conducting, or nonconducting material, is known as a *linear antenna*. Because of their simplicity, linear antennas are probably the most common type of radiator for communication between distant points. They exist in many varieties delineated by (1) geometry (e.g., straight dipole, V-shaped dipole, L-shaped antenna), (2) electrical characteristics (e.g., resonant, antiresonant, wideband), and (3) radiation properties (e.g., isotropic, directive). With a view toward application, linear antennas and antennas in general are evaluated with respect to the spatial and frequency characteristics of the radiation and their circuit or electrical properties. For example, if a designer wants to send narrowband signals to all parts of the world without any preference in direction, the ideal radiator would have an isotropic distribution of energy in space. Furthermore, as a circuit element the antenna must be matched to a low-impedance source through a transmission line. Here a resonant antenna is the right choice.

1. LINEAR DIPOLES

To better understand the linear antenna, it is best to start out with the simplest example, namely, a linear dipole of half-length h driven in the center by a sinusoidal voltage. The dipole is constructed of thin wire or rods, with each rod being connected to one end of the transmission line, as shown in Fig. 1.

The dipole of Fig. 1 has radiation characteristics that are dependent on the current on the wires. A good analogy in visualizing the current distribution is to consider a string fixed at both ends and plucked at various points along the length. After plucking, nodes and antinodes appear in a configuration called a *standing wave*. For our dipole, the first "mode" corresponding to a resonant length resembles a cosine with a maximum at the center (antinode) and zero at the ends (nodes) as shown in Fig. 2. This would be a half-wavelength dipole, $h/\lambda = 0.25$. If the dipole is electrically smaller, say, $h/\lambda = 0.05$, the current would still be zero at the ends and the maximum would still be at the center. As the electrical length of the antenna increases, by increasing the frequency of the source, more complicated current distributions arise. In fact, with our simple model, there will be times when the driving point is at a node, and the driving point impedance is infinite. This scenario is shown in Fig. 2 for $h/\lambda = 0.5$.

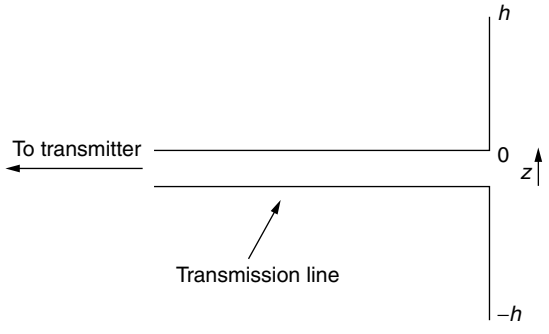


Figure 1. A dipole antenna.

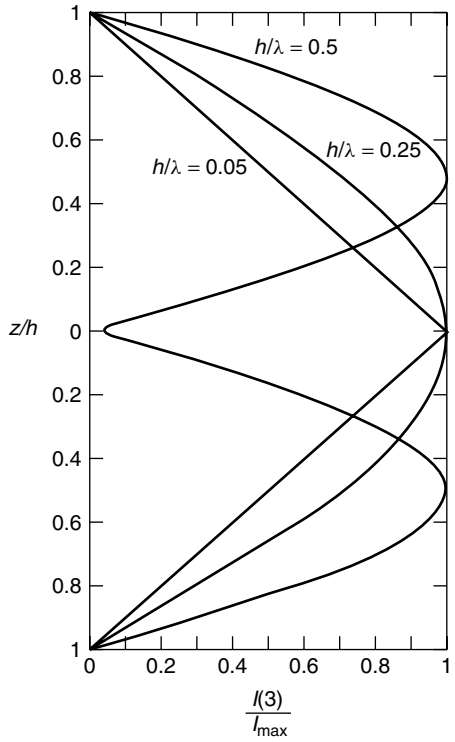


Figure 2. Idealized antenna currents.

In practice the current is never a pure sinusoid, so that the driving-point impedance will never be infinite. The spatial radiation characteristics, called the *radiation pattern*, after calculation from an assumed current, have their maximum perpendicular to the dipole when the half-length is less than about a wavelength and a half. This maximum is in the plane of the antenna (i.e., E plane) in the broadside direction and in the plane perpendicular to the antenna (i.e., H plane). The radiation is uniform or isotropic. Figure 3 shows the E plane radiation pattern for a linear antenna that has values of $h/\lambda = 0.05$, 0.25, and 0.5. For increasing lengths, the maximum radiation can be in oblique directions to the dipole axis. It is instructive to quantitatively examine the radiation pattern of a linear antenna based on the geometry shown in Fig. 4. The far-zone electric field E_θ^R, also called the *radiation field*, is in the direction of the θ arrow and is tangent to a sphere whose radius is R. Together, R and θ form the E plane. Note that one must be roughly in the range $k_0R \gg 1$ to meet the conditions for the far zone. A simplified relation for E_θ^R is

$$E_\theta^R = CF_0(\theta, k_0h) \tag{1}$$

where C is a factor that contains the R dependence of the field, and F_0 is the field pattern normalized with respect to the value of the current at $z = 0$.

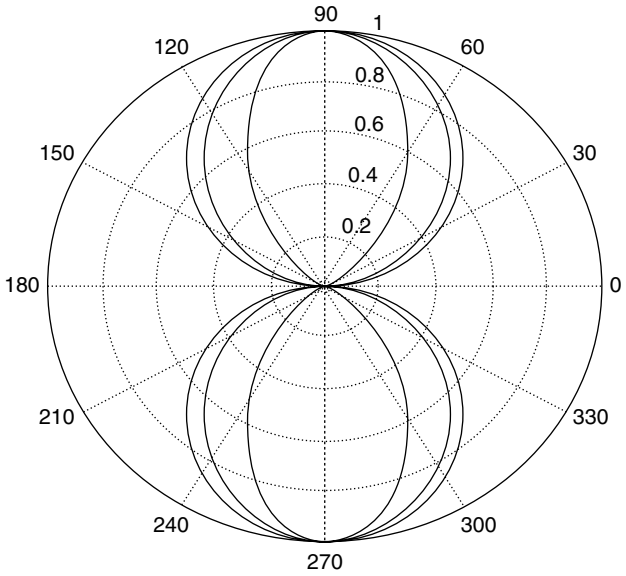


Figure 3. Radiation pattern for linear antenna with different electrical lengths.

The field pattern is given by [1]

$$F_0(\theta, k_0 h) = \frac{k_0 \sin\theta}{2I(0)} \int_{-h}^{h} I(z') e^{jk_0 z' \cos\theta} dz' \qquad (2)$$

When the current at the base of the antenna is zero (see Fig. 2 for $h/\lambda = 0.5$), the field pattern can be normalized to the maximum value of the current. This produces a radiation pattern $F_m(\theta, k_0 h)$. Patterns shown in Fig. 3 were computed from the F_0 or F_m relation and are valid for the E plane. The far-zone magnetic field E_ϕ^R is orthogonal to the E_θ^r field and is located in the H plane formed by the ϕ and r arrows in Fig. 4.

The antenna current, which in the frequency domain is a complex quantity, completely determines the driving-point impedance of the antenna. For an electrically short antenna, the driving-point resistance is small and the reactance is capacitive and large. At the first resonance, the complex driving-point impedance is approximately

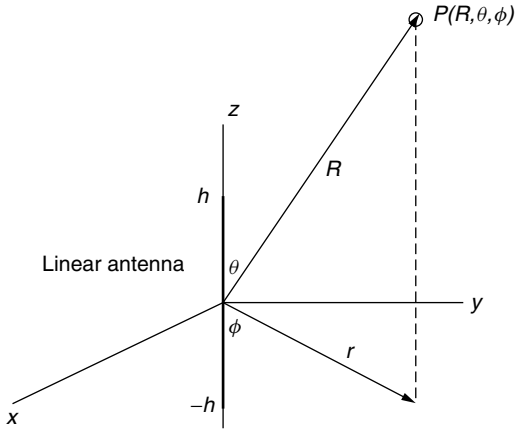


Figure 4. Gain and directivity.

$73 + j42$ ohms (Ω) (half-wavelength antenna). This explains why the half-wavelength linear antenna is so popular, since the driving-point resistance is easy to match with available transmitters and transmission lines. Many sources giving the impedance characteristics of linear antennas are available; perhaps the best is in *Tables of Antenna Characteristics* by King [2]. The design of a linear antenna system for a single frequency is not very complicated. Choose a half-wavelength antenna, design a matching network to cancel out the driving-point reactance, and find a compatible transmitter and coaxial line to match the antenna. The design of a linear antenna system where the bandwidth is important requires that the response of the antenna at different frequencies not cause excessive degradation in the amplitude of the transmitted signal.

One important antenna parameter has to do with the concentration of radiation in a specific direction. The gain of an antenna is proportional to the power radiated in a given direction divided by the average power. Another parameter, called the *directivity*, is equal to the maximum value of the gain and is expressed as a numeric or in decibels (dB). The directivity of a half-wavelength dipole is 1.64 or 2.1 dB. (i.e., $10 \log 1.64$). An example to illustrate these concepts is to design a more directive linear antenna that carries a sinusoidal current. In analogy to the radiation (light pattern) from a thin optical slit, it is known that a more directive light pattern is obtained by increasing the length of the slit. However, when the same idea is tried with a linear antenna, an increase in directivity is not present when the antenna is lengthened. This is because successive half-wavelengths of current are of opposite sign and serve to reduce the radiated field. To overcome this difficulty, phase-reversing stubs can be placed every half-wavelength along the antenna. The current along the antenna is now unidirectional and a closer approximation to the uniform light in a slit.

Sometimes the designer is limited in the physical length available for the antenna. For example, linear antennas that are short in electrical length can have reduced resistance and increased capacitance when compared with a half-wavelength dipole. The driving-point impedance of a quarter-wavelength dipole is about $14 - j195$ Ω, while the impedance of a half wavelength dipole is about $73 + j42$ Ω. To make increase the apparent electrical length of the quarter-wavelength dipole, a series inductance can be placed near the base of the antenna, say, with a coil of wire. Top loading and series loading at any point is also possible to change the current distribution on the antenna.

2. TRAVELING-WAVE ANTENNAS

So far, the discussion has been concerned mainly with the standing-wave linear antenna, since the current must be zero at the ends of the antenna (i.e., $z = \pm h$). A different type of antenna current distribution is concerned with traveling waves instead of standing waves. It is well known that a standing wave can be decomposed into a forward/backward-traveling wave. For example, in a half

wavelength dipole the ideal current is given by $I_z(z) = I_0 \cos k_0 z$, $|z| \le h$. Using the exponential representation for the current, we obtain

$$I_z = I_0 \cos k_0 z = I_0 \left(\frac{e^{jk_0 z} - e^{-jk_0 z}}{2} \right) \tag{3}$$

Using $e^{j\omega t}$ time dependence,

$$I_z = I_0 e^{j(\omega t + k_0 z)} + I_0 e^{j(\omega t - k_0 z)} \tag{4}$$

where $k_0 = (2\pi/\lambda_0)$ = free-space propagation constant
$\omega = 2\pi f$
f = frequency in Hz
t = time
z = distance along the antenna (5)

The first term on the right represents a traveling wave moving inward to the base, and the second term represents a wave moving outward toward the end of the antenna at $z = h$.

From transmission-line theory it is also known that using a termination equal to the characteristic impedance can produce a reflectionless line. A traveling-wave antenna, called a "beverage antenna," is constructed by placing a conductor parallel to the earth and terminating it with terminal impedance, producing minimum reflections at the end.

For a monopole structure the radiation pattern in the E plane roughly resembles a set of rabbit ears, where each ear is at an oblique angle to the antenna axis. As the monopole elongates electrically, the ears move closer to the antenna axis. The radiation pattern of a traveling-wave dipole antenna consists of two rabbit ears roughly in the shape of the letter X. If a unidirectional pattern is desired, a V-shaped antenna may be used. It is constructed by bending the arms of a dipole about the center. The apex angle is chosen such that the inside radiation lobes are completely superimposed on one another. A diagram of the radiation patterns for two representative traveling wave antennas is shown in Fig. 5.

The reflectionless antenna described by Wu and King [3] has a prescribed resistive coating that produces a traveling wave along a finite-length antenna. It has an important application as an antenna for pulses that have a very wideband frequency spectrum. One major drawback for this antenna is that it is about 50% efficient since half of the power is dissipated in the resistance. Increased attention has been given to antennas energized by short temporal pulses for use in GPR (ground-penetrating radar) systems. The design of antennas for use in such systems involves somewhat different criteria and physical viewpoint than antennas used for CW (continuous-wave) systems. The following simple example will illustrate the physics involved in driving a linear antenna with a carrierless temporal pulse. A dipole of finite length is energized with a short temporal pulse, say, with a half-width of an nanosecond (1 ns). In order to get the whole pulse to exist on the antenna, the antenna length must be greater than a foot. Roughly, the pulse travels at a velocity of 1 ns/ft along the antenna. This is the velocity of an electromagnetic wave in free space.

2.1. Example of a Traveling-Wave Antenna

To gain some insight into the radiation properties of a pulsed antenna, the traveling-wave antenna with a Gaussian pulse excitation will be considered. A more detailed analysis can be found in the book by Smith [4]. The radiated field for a linear antenna in the time domain has a form different from that in the frequency-domain integral given earlier. It is possible to find the time-domain radiation expression from the frequency-domain representation by using a Fourier transform. Qualitatively the time-domain radiated field is proportional to the integral along the antenna of the time derivative of the current, evaluated in retarded time. Expressed in quantitative terms, this radiated field for a monopole is given by

$$E^r = \frac{\mu_0}{4\pi} \sin\theta \int_0^h \left[\frac{\partial}{\partial t'} I_z \left(t' - \frac{z'}{c} \right) \right] dz' \tag{6}$$

where

μ_0 = magnetic constant
$= 4\pi \times 10^{-7}$ H/m (henries per meter) (7)
= retarded time

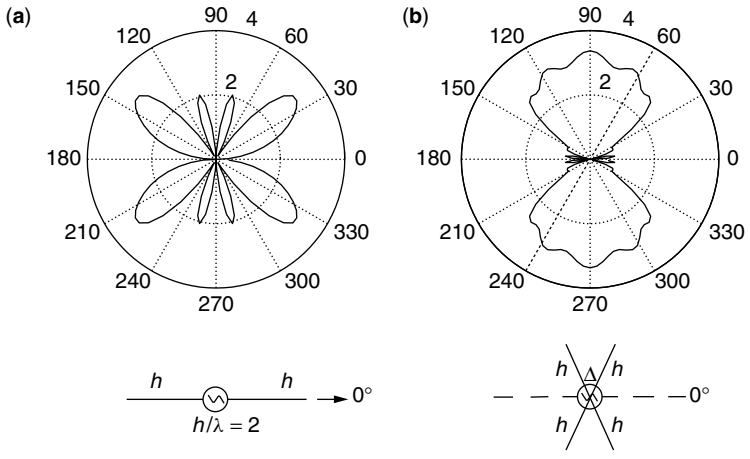


Figure 5. Two representative traveling wave antennas: (**a**) linear antenna with an assumed traveling wave current ($h/\lambda = 2.0$); (**b**) X-shaped antenna with traveling-wave currents ($\Delta = 32°$, $\Delta/\lambda = 2$).

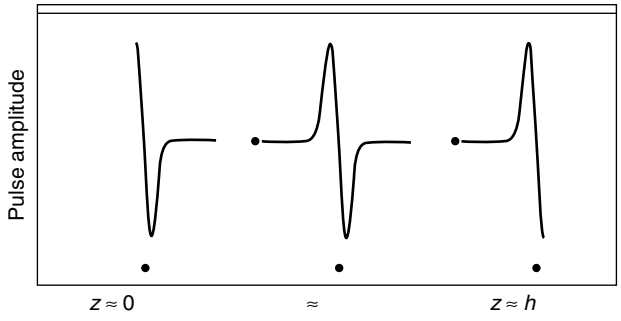


Figure 6. Pulse example.

In expression (6) the current is traveling along the positive z direction with a velocity c. Figure 6 shows the integrand of the radiation integral of (6). The first derivative of a Gaussian pulse has two equal sections, one positive and one negative, and at the driving point there are times (i.e., near $z = 0$) that areas under the positive and negative sections do not cancel. Here radiation exists. When the entire pulse is present in the antenna, say, near $z = h/2$, the positive and negative areas do cancel and there is no radiation. When the pulse is near the end of the antenna, $z = h$, the incident and reflected pulse areas may not cancel at certain times, giving rise to a second radiated pulse. As time progresses, pulses continue to be radiated at $z = 0$ and $z = h$.

2.2. Receiving Antenna

Linear antennas are also used as receptors for electromagnetic signals. Important quantities are the voltage at the terminals of the antenna, V_0, and the current through the load impedance, Z_L. A sketch of a linear antenna with length $2h$ used for reception is shown in Fig.7, along with its Thévinen equivalent circuit. The equivalent circuit has a series arrangement of the driving-point impedances Z_0 and Z_L driven by the open-circuit voltage V_0. From this arrangement the current in the circuit is given by

$$I_z(z = 0) = I_z(0) = \frac{V_0}{Z_0 + Z_L} \tag{8}$$

An important parameter for the receiving antenna is the complex effective length $h_e(k_0 h)$, which relates V_0 to the antenna. Thus

$$V_0 = 2h_e(k_0 h)E_z^{\text{inc}} \text{ volts } (V) \tag{9}$$

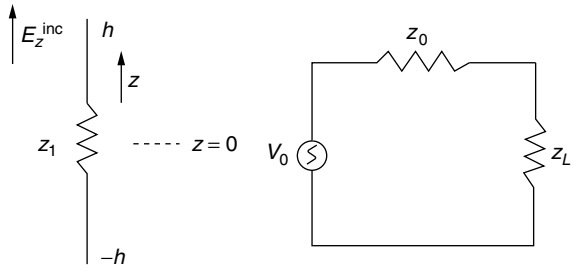


Figure 7. Receiving antenna and its Thévenin equivalent circuit.

A good reference on the receiving antenna is the book by King [5]. When the antenna is short, $k_0 h \leq 0.5$, the effective length is approximately equal to half the physical length. Stated in another way, the total length of the antenna is about equal to twice the effective length. As the antenna becomes longer than $k_0 h = 0.5$, the simple approximation breaks down. For example, a resonant antenna with $k_0 h = \pi/2$ has an effective length of about $h_e \approx 1.21$. So far the discussion has been about the circuit properties of transmitting and receiving antennas. When attention is given to the currents on transmitting and receiving antennas, the situation is more complicated. These currents differ because on a receiving antenna, both the incident electric field and the load impedance are involved. With a zero value of Z_L, the current distribution is of the receiving type. For antennas of moderate length the current has a shifted cosine distribution given by $\cos k_0 z - \cos k_0 h$. When the load impedance is increased, a transmitting current is added to the receiving current. It has the form $\sin k_0(h - |z|)$.

BIOGRAPHY

Sheldon S. Sandler received his Ph.D. and M.A. from Harvard University, Massachusetts, his M.Eng.E.E. from Yale University, and his B.S.E.E. from Case Institute of Technology. He is a professor emeritus in the Department of Electrical and Computer Engineering at Northeastern University, Boston, Massachusetts. He is also a research fellow in the Department of Archaeology at Boston University, Massachusetts, working on remote sensing for site evaluation. At Geo-Centers, Inc., Massachusetts, he is a senior engineer and a technical advisor to the CEO. There he has developed new GPR systems and time domain antennas as well as designing algorithms to detect targets in noisy data. Dr. Sandler has been a guest professor at both the E.T.H and the University of Zurich in Switzerland and at the Robotics Center at the University of Rhode Island. At the MRC Laboratory in Cambridge, England, he was a visiting scholar. Dr. Sandler is the author of *Picture Processing and Reconstruction* and the coauthor of *Arrays of Cylindrical Dipoles* with R.W.P. King and R. Mack. He is a member of the International Radio Union (URSI), IEEE, Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.

BIBLIOGRAPHY

1. R. W. P. King, R. B. Mack, and S. S. Sandler, *Arrays of Cylindrical Dipoles*, Cambridge Univ. Press, Cambridge, UK, 1968.
2. R. W. P. King, *Tables of Antenna Characteristics*, IFI/Plenum, New York, 1971.
3. T. T. Wu and R. W. P. King, The cylindrical antenna with nonreflecting resistive loading, *IEEE Trans. Antennas Propag.* **AP-13**: (1975).
4. G. S. Smith, *An Introduction to Classical Electromagnetic Radiation*, Cambridge Univ. Press, Cambridge, UK, 1997, Chap. 8.
5. R. W. P. King, *The Theory of Linear Antennas*, Harvard Univ. Press, Cambridge, MA, 1956, Chap. 4.

LINEAR PREDICTIVE CODING

AMRO EL-JAROUDI
University of Pittsburgh
Pittsburgh, Pennsylvania

1. INTRODUCTION

Most real-world signals carry redundant information from one sample (or snapshot) to the next. For example, a television video signal is made of a sequence of frames (about 30 frames per second, depending on the video standard) where often very little changes in the picture from one frame to the next. Even within a frame, neighboring pixels are likely to be related in terms of intensity and color. It is not unusual for an office document to contain long strings of consecutive white pixels or long strings of consecutive black pixels.

From an efficient communication standpoint, it is extremely wasteful to spend valuable bits (or bandwidth) on encoding the redundant information from one sample to the next. Instead, it is more efficient to use the bits to encode only the novel information. Consequently, a preprocessing procedure to remove the intersample redundancy becomes necessary before encoding a signal for storage or transmission. This procedure has two steps. In the first step, an estimate of the current sample is *predicted* (guessed scientifically) based on its neighbor(s). This predicted value is the redundant portion of the current sample since it is based solely on neighboring samples. The second step is simply to subtract the predicted value (the redundancy) from the current sample, thereby, leaving only the novel information to be encoded. Although, in general, one may use any parametric function to predict a sample from its neighbors, the discussion below focuses on *linear prediction* (LP), where the sample is predicted as a linear combination of other samples. While the focus of this article is on the use of LP in redundancy removal or data compression for coding applications [also known as *linear predictive coding* (LPC)], it is important to note that LP is used in a variety of other applications, including forecasting, control, system modeling and identification, and spectral estimation, to name only a few [1,2].

The remainder of the article is organized as follows. In Section 2, LP is formulated and the optimal prediction parameters are derived. In Section 3, the computational aspects and algorithms for the implementation of LP are explored. In Section 4, examples and applications of LPC are presented. Finally, in Section 5, variations on LPC used for coding speech signals are discussed.

2. FORMULATION OF LINEAR PREDICTION

Given a discrete-time signal,[1] x_n, defined over a finite interval $n = 0, 1, 2, \ldots, N-1$, define $\hat{x}_n$ to be the predicted value of x_n based on the p previous values of x_n. In other words

$$\hat{x}_n = \sum_{k=1}^{p} a_k x_{n-k} \tag{1}$$

where $a_1, a_2, \ldots, a_p$ are the prediction parameters (or coefficients). The prediction error e_n is then defined as the difference between x_n and $\hat{x}_n$:

$$e_n = x_n - \hat{x}_n = x_n - \sum_{k=1}^{p} a_k x_{n-k} \tag{2}$$

The error signal is often referred to as the *residual signal* since it describes the residual information after the redundancy removal.

The prediction parameters are chosen to minimize the prediction error subject to an optimality criterion. Different prediction parameters can be obtained depending on the criterion selected. Below, we examine the method of *least squares* (LS), which is one of the more popular criteria. For an example of other methods, please see Refs. 3 and 4.

2.1. LS Minimization

The least-squares criterion takes on different forms depending on the assumptions made regarding the signal, x_n. If we treat the signal as a random signal, we minimize the expected value of the squared error

$$\mathbf{E} = E\{e_n^2\} = E\left\{\left[x_n - \sum_{k=1}^{p} a_k x_{n-k}\right]^2\right\} \tag{3}$$

where $E\{\cdot\}$ stands for the expectation operator. Assuming the signal to be a sample of a stationary process, substituting for e_n in Eq. (3), taking the derivative with respect to a_i for $1 \le i \le p$, and setting the derivative to zero yields the following set of equations

$$\sum_{k=1}^{p} a_k R_{i-k} = R_i \tag{4}$$

where R_i is the autocorrelation of the random process and is defined as

$$R_i = E\{x_n x_{n-i}\} \tag{5}$$

Here, R_i measures how a sample is related to another sample which is i lags away, with a value of zero indicating no correlation between the samples. In other words, the autocorrelation function is a measure of the average redundancy between samples. It is then no surprise that the autocorrelation information is used to determine the optimal prediction parameters.

Instead of treating x_n as a random signal, we may treat it as a deterministic signal. Then, we minimize the sum of the squared errors

$$\mathbf{E} = \frac{1}{N}\sum_n e_n^2 = \frac{1}{N}\sum_n \left[x_n - \sum_{k=1}^{p} a_k x_{n-k}\right]^2 \tag{6}$$

[1] A discrete-time signal is usually obtained by sampling an analog signal using an analog-to-digital converter.

If we assume that the range of minimization is infinite, then taking the derivative and setting it to zero yields

$$\sum_{k=1}^{p} a_k \hat{R}_{i-k} = \hat{R}_i \tag{7}$$

where $\hat{R}_i$ is the (time-average) autocorrelation function of the signal x_n and is given by

$$\hat{R}_i = \frac{1}{N} \sum_{n=-\infty}^{\infty} x_n x_{n-i} \tag{8}$$

Note that $\hat{R}_i$ here is calculated over all the values of x_n. Unfortunately, in practice, the signal x_n is not given over an infinite interval. To reduce the range of summation in Eq. (8), the given signal is often multiplied by a window function creating a new signal $\tilde{x}_n$, which is zero outside the interval $0 \le n \le N-1$. The autocorrelation function is then calculated for $\tilde{x}_n$,

$$\hat{R}_i = \frac{1}{N} \sum_{n=0}^{N-1} \tilde{x}_n \tilde{x}_{n-i} \tag{9}$$

where, as described above, $\tilde{x}_n$ is given by

$$\tilde{x}_n = \begin{cases} x_n w_n, & 0 \le n \le N-1 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Clearly the window function, w_n, will influence the value of the autocorrelation function and the resulting predictor coefficients. Consequently, great care should be used in selecting it. Two of the more popular window functions are the rectangular window given by $w_n = 1$ for $0 \le n \le N-1$, and the Hamming window given by

$$w_n = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \qquad \text{for} \qquad 0 \le n \le N-1 \tag{11}$$

It is important to note the similarity between Eqs. (4) and (7). They differ only in the definition of the autocorrelation function. Fortunately, in practice, (9) is often used to estimate the autocorrelation of the stationary process. In this case, the predictor coefficients produced under the random stationary process assumption would be the same as those produced under the deterministic infinite interval assumptions. Since the methods rely on the autocorrelation information, they are often referred to as the *autocorrelation method of LP*.

If we assume that the error in Eq. (6) is defined over a the finite interval $0 \le n \le N-1$ and minimize it only over this interval, we obtain the following set of equations

$$\sum_{k=1}^{p} a_k \varphi_{i,k} = \varphi_{0,i} \tag{12}$$

where $\varphi_{i,k}$ is called the *covariance of the signal* x_n and is given by

$$\varphi_{i,k} = \frac{1}{N} \sum_{n=0}^{N-1} x_{n-i} x_{n-k} \tag{13}$$

In Eq. (13), it is required that the values of the signal x_n be known over the range $-p \le n \le N-1$ for all the terms in the summation to be calculated. If the values for $-p \le n \le -1$ are not known, then the summation limits in (13) must be changed to $p \le n \le N-1$. This method is often referred to as the *covariance method of LP*. For the details and properties of this method, please see the article by Makhoul [1].

2.2. The Minimum Error

In order to gauge the quality of the obtained predictor, one can examine the final prediction error which is the minimum value of the error criterion used. This value is, of course, dependent on the method of LP used. For the autocorrelation method, the final prediction error is obtained by substituting Eq. (7) in (3) or (6), thereby producing

$$\mathbf{E}_{\min} = R_0 - \sum_{k=1}^{p} a_k R_k \tag{14}$$

This error is a measure of the variance (or power) in the residual. Consequently, this value is used to calculate a factor, G, to normalize the residual signal and produce a unit-variance excitation signal, u_n. In other words

$$u_n = \frac{e_n}{G} \tag{15}$$

where

$$G = \sqrt{\mathbf{E}_{\min}} \tag{16}$$

The advantage of this normalization lies in that, independent of the power of the original signal, the resulting excitation signal will be consistently unit-variance making it easier to encode. Figure 1 is a block diagram of the relation between the given signal x_n, the error signal e_n, and the excitation signal u_n. It is easy to see that the relation is that of a discrete-time moving-average (MA) linear time-invariant filter with input x_n, output u_n, and parameters $\{1, -a_1, -a_2, \ldots, -a_p\}$ and G. The MA (also known as the *analysis*) filter has a transfer function $A(z)/G$, where

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{17}$$

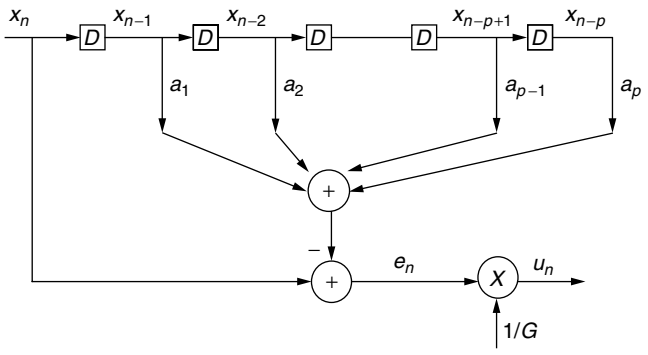


Figure 1. LPC analysis.

As discussed above, the excitation is then encoded for storage or transmission through a communications channel. The prediction parameters and gain are also encoded for storage or transmission. The decoder then uses the residual and the prediction parameters and gain to synthesize the original signal. The synthesis is performed by passing the excitation signal (hence the name) through a filter $H(z)$ that is the inverse of the MA filter used by the encoder. In other words

$$H(z) = \frac{G}{A(z)} \tag{18}$$

and

$$s_n = Gu_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{19}$$

Figure 2 is a block diagram of the synthesis filter that is a discrete-time autoregressive (AR) LTI system. It is important to note that, in Eq. (18), we assumed that the analysis filter $A(z)$ is invertible which is true only if all its roots lie inside the unit circle. This issue will be addressed in a later section. If the encoding of u_n, the prediction coefficients and gain introduced no errors, the synthesized signal s_n would be identical to the original signal x_n. This latter assumption however is not realistic since encoding invariably introduces quantization errors. The effect of quantization on the excitation and prediction parameters will be discussed later.

3. COMPUTATION OF PREDICTION PARAMETERS

In the case of the autocorrelation method, the minimization equations in Eq. (4) form a set of p linear equations in p unknowns, which can be written in matrix form as

$$\mathbf{Ra} = \mathbf{r} \tag{20}$$

where the autocorrelation matrix $\mathbf{R}$ is given by

$$\mathbf{R} = \begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_1 & R_0 & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \tag{21}$$

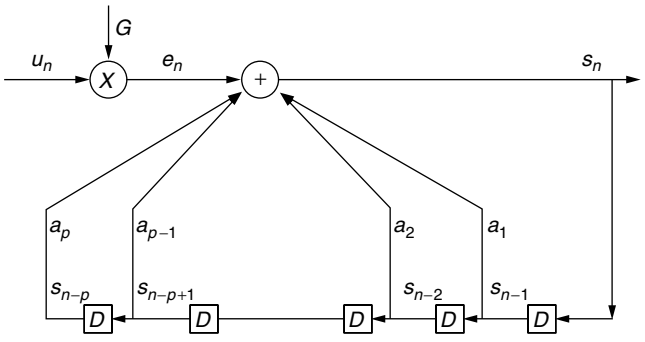


Figure 2. LPC synthesis.

and the vectors $\mathbf{a}$ and $\mathbf{r}$ are given by

$$\mathbf{a} = [a_1, a_2, \ldots, a_p]^T \tag{22}$$

$$\mathbf{r} = [r_1, r_2, \ldots, r_p]^T \tag{23}$$

It is easy to see that $\mathbf{R}$ has a special structure; namely, the elements along each diagonal are equal and the matrix is symmetric. A matrix of this form is called *Toeplitz symmetric*. The vector of prediction coefficients can be obtained by solving Eq. (20) using standard methods such as Gaussian elimination. These methods usually require on the order of p^3 operations. These methods however do not take advantage of the special structure of $\mathbf{R}$. The special method of Levinson–Durbin takes into account the Toeplitz nature of $\mathbf{R}$ and the fact that the $\mathbf{r}$ vector is composed of the same elements in $\mathbf{R}$. This method requires on the order of p^2 operations and is described as follows:

1. The initialization step

$$E_0 = R_0 \tag{24}$$

2. The recursion steps repeated for $i = 1, 2, \ldots, p$

$$K_i = \frac{R_i - \sum_{k=1}^{i-1} a_k^{(i-1)} R_{i-k}}{E_{i-1}} \tag{25}$$

$$a_k^{(i)} = a_k^{(i-1)} - K_i a_{i-k}^{(i-1)} \qquad \text{for} \qquad 1 \le k \le i-1 \tag{26}$$

$$a_i^{(i)} = K_i \tag{27}$$

$$E_i = \left(1 - K_i^2\right) E_{i-1} \tag{28}$$

The solution for the order p prediction coefficients is then given by

$$a_k = a_k^{(p)} \quad \text{for} \quad 1 \le k \le p \tag{29}$$

In addition to its great computational advantage, the Levinson–Durbin method provides procedural advantages as well. It is important to note that during the recursion steps, the prediction coefficients for predictors of order less than p are calculated, namely, $a_k^{(i)}$ in Eqs. (25)–(27) refers to the kth coefficient of the optimal predictor of order i. Moreover, the prediction error for the lower-order predictors is also produced, $\mathbf{E}_i$ in Eq. (28). This information may be used to select the most appropriate prediction order, p. In contrast, using the standard methods, one would have to solve the equations multiple times to compare the performance of the various order predictors and select the appropriate p. The Levinson–Durbin method also produces an alternate set of coefficients $K_i, 1 \le i \le p$. It is often these coefficients (or a function of them) that are encoded and transmitted. The decoder then uses Eqs. (26) and (27) to reconstruct the prediction coefficients and synthesize the original signal.

One is always faced with the question of which method to choose for estimating the LP coefficients. While there is no rule of thumb, an understanding of the advantages and disadvantages of each may help the reader choose the method most appropriate for the application at hand. For

the autocorrelation method, the advantages are twofold: (1) it utilizes a fast computational algorithm with useful intermediate information and (2) the resulting $A(z)$ is guaranteed to have its roots inside the unit circle, which makes it invertible. This is of great importance since the synthesis filter $H(z)$ becomes unstable if $A(z)$ has roots outside the unit circle. The main disadvantage of the autocorrelation method is the effect of the windowing function on the estimated LP coefficients. The inaccuracies introduced due to windowing lead to suboptimal predictors.

4. EXAMPLE OF LPC

To demonstrate the effectiveness of LPC in signal compression and redundancy removal, consider the speech signal shown in Fig. 3 (this is a portion of the vowel /ee/ as in beet). The signal is sampled at 11,025 samples per second and is 350 samples long. When encoded using a 2-bit per sample pulse code modulation (PCM) encoder, the resulting signal at the receiver is shown in Fig. 4. The coding error, which is the difference between the original and the reconstructed signal, is shown in Fig. 5. The signal-to-noise ratio (SNR) defined as 10 times the base_{10} logarithm of the ratio of the power in the original signal over the power in the coding error is 7.4 dB. If we perform LPC on the original signal using a 10^{th} order predictor, we can then quantize the excitation signal using 2 bits per sample of PCM and use it to synthesize the speech at the decoder (the quantization is performed using adaptive predictive LPC to maximize the SNR [2]). The quantized excitation signal is shown in Fig. 6, while the reconstructed speech signal is shown in Fig. 7. In this case the coding error is shown in Fig. 8. Note that the coding error using LPC is much smaller than the error using PCM. In fact, the SNR for LPC is 19.7 dB, representing a gain of 12 dB at the expense of encoding and transmitting the prediction coefficients and gain (usually on the order of 50 bits). If we use a single bit per sample to encode the excitation signal, thereby cutting the LPC bit rate in half, the resulting signal has a SNR of 13.6 dB, which is still an improvement over 2-bit PCM. In this case, LPC reduced the bit rate and improved the quality of the resulting signal compared to PCM.

5. APPLICATIONS OF LPC

A common model for speech generation (or synthesis) is shown in Fig. 9, where the output speech is produced by passing an excitation signal through an all-pole filter [1,2]. The system shown in Fig. 9 is similar to the synthesis (decoder) system in Fig. 2; the important difference is that the excitation signal is generated locally at the decoder. The nature of the excitation depends on the desired sound to be produced. For voiced speech (e.g., vowel sounds), the excitation consists of periodic pulses. The period of these pulses (also known as the *pitch period*) corresponds to the desired fundamental frequency (or pitch frequency) of the produced sound which varies from speaker to speaker. For unvoiced speech (e.g., fricative sounds such as /s/ and /sh/), the excitation consists of the output of a random-noise generator. Note that, under this synthesis model, the decoder does not need the actual excitation signal. Instead, it needs the type of excitation, the pitch period in case of voiced speech, along with the gain and prediction parameters. Consequently the analysis (encoder) system will differ from the one shown in Fig. 1 and is shown in Fig. 10.

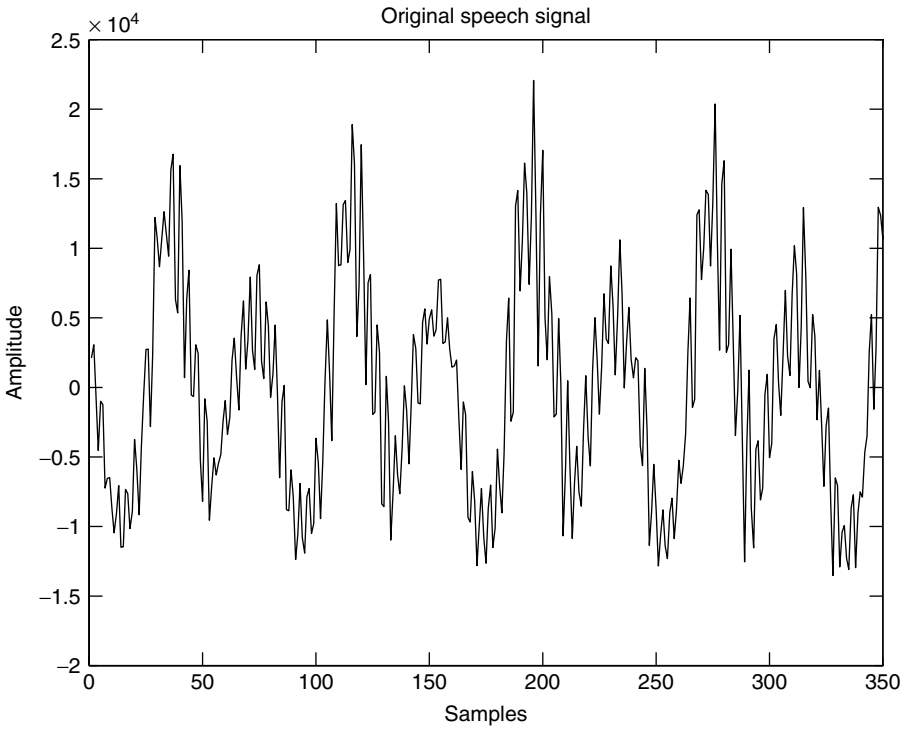


Figure 3. Original speech signal.

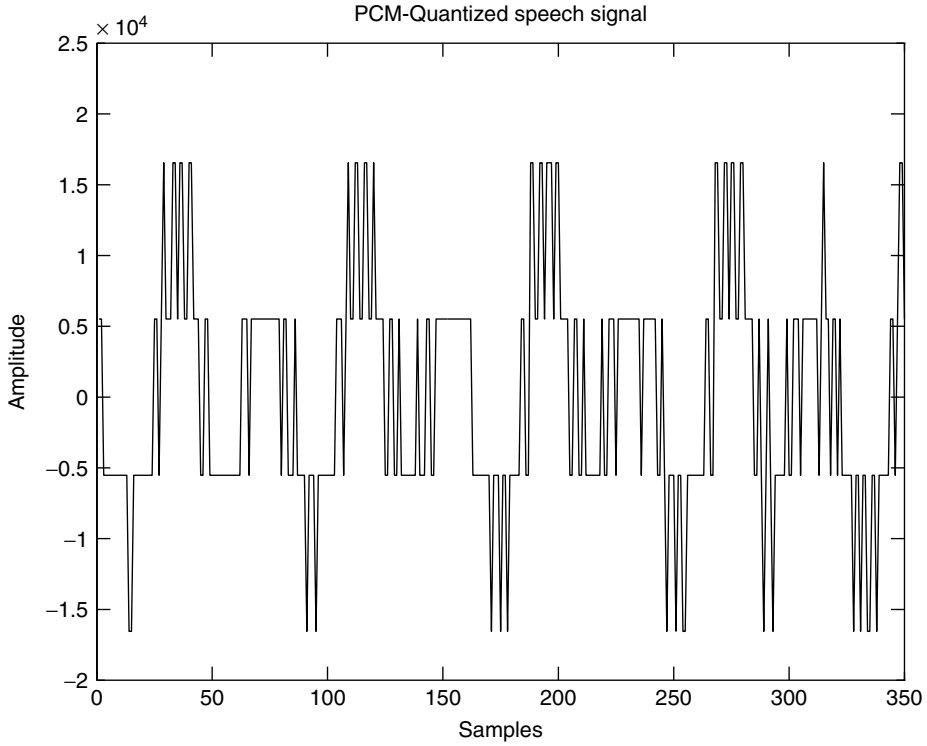


Figure 4. PCM-quantized speech signal using 2 bits per sample.

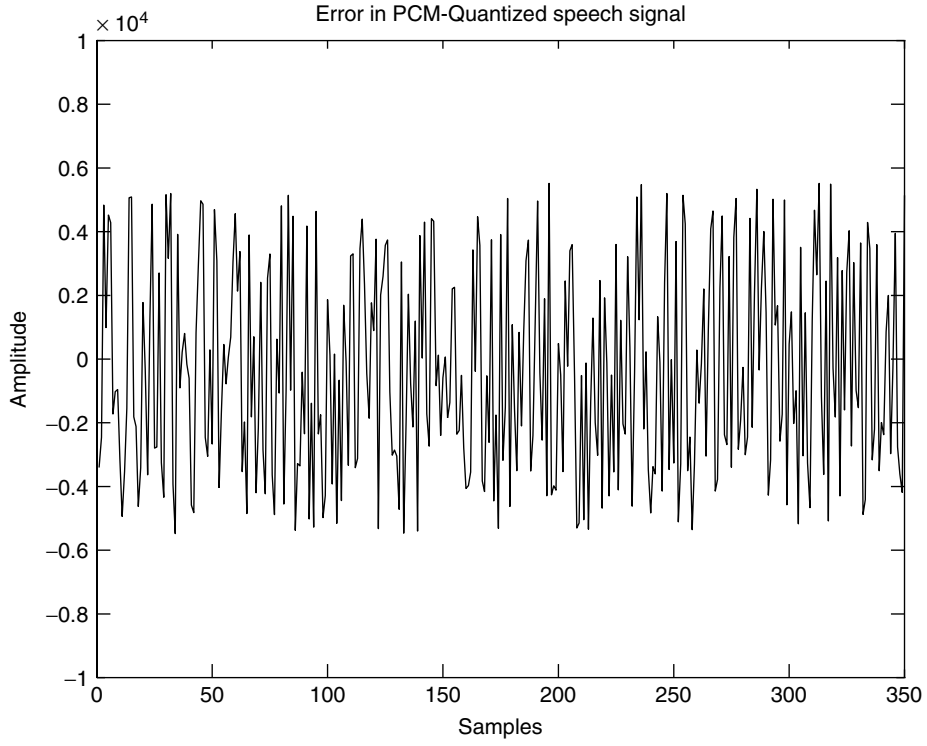


Figure 5. Coding error in PCM-quantized speech signal.

The voiced/unvoiced and pitch information is typically transmitted every 10 ms to track changes in speech. The prediction parameters and gain may be transmitted every 20–30 ms. For speech sampled at 8000 samples per second, it is typical to use a predictor of order 10. Typical bit assignment for the various parameters which would lead to a bit rate on the order of 2400 bits per second (bps) (i.e., approximately 0.25 bits per sample of the original signal). Additional compression may be obtained by applying vector quantization to the parameters. If we had used PCM to encode the original speech, the required bit rate would be on the order of 64,000 bps.

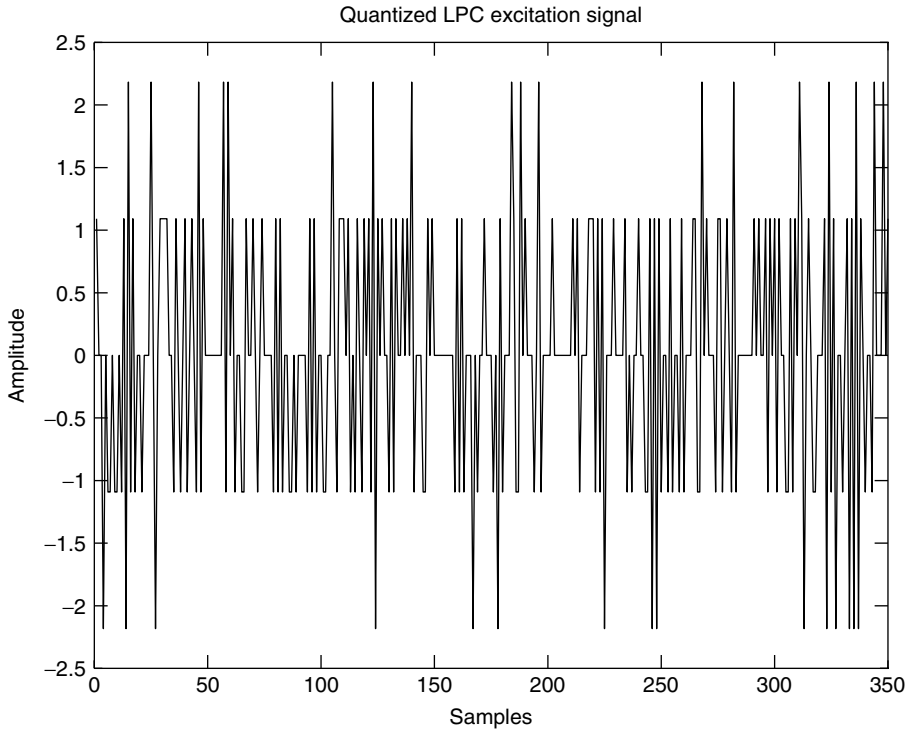


Figure 6. Quantized LPC excitation signal.

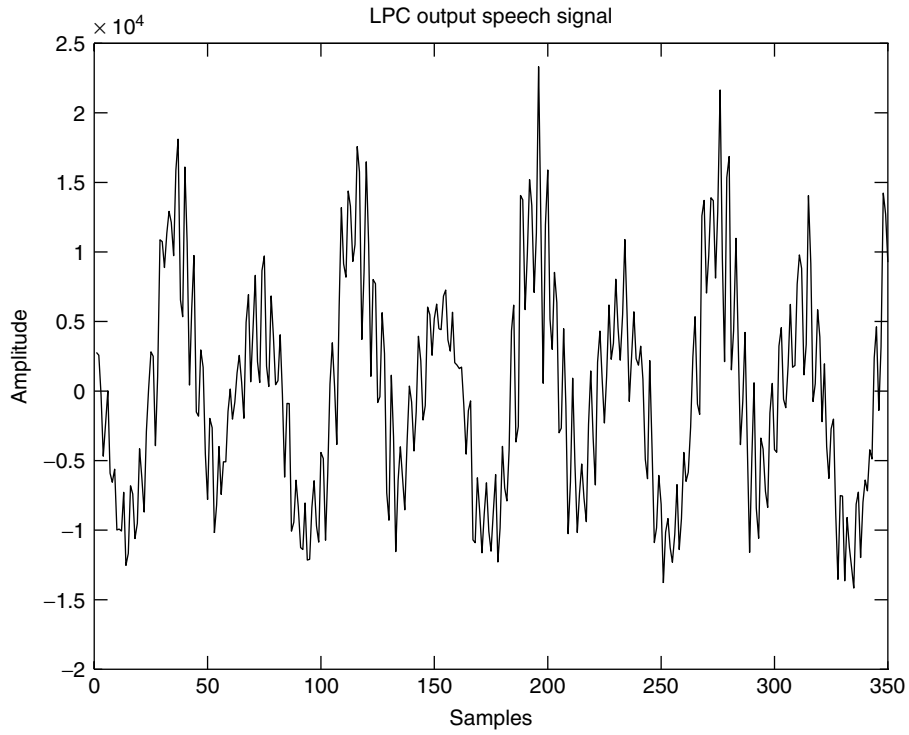


Figure 7. Reconstructed speech signal using LPC.

Clearly, LPC provided great reduction in bit rate in this case. Some of this reduction, however, comes at the expense of the quality of the output speech. Because of the synthesis model used with the hard switching between sources of excitation, the resulting speech lacks naturalness and is often described as choppy. Over the years, more sophisticated variations on LPC have been devised. We describe two of the more recent methods below: code-excited LP (CELP) [5] and mixed-excitation LP (MELP) [6].

In CELP, the excitation signal is selected from a library (referred to as a *codebook*) of possible excitation signals. The encoder conducts an exhaustive search of all the excitations in the codebook to determine the one that

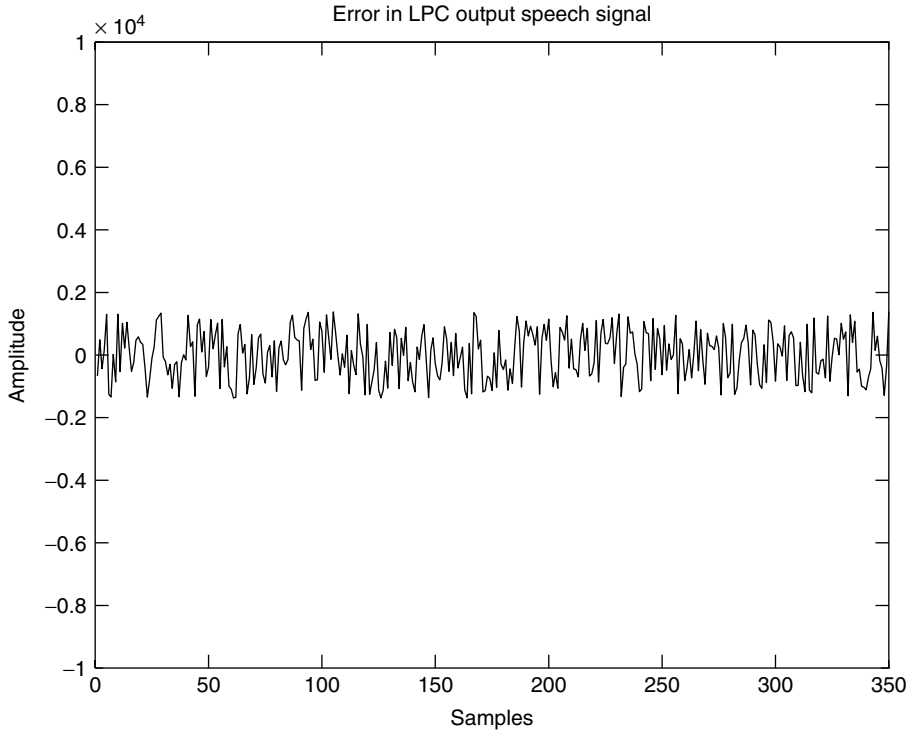


Figure 8. Coding error in LPC generated speech signal (same scale as in Fig. 5).

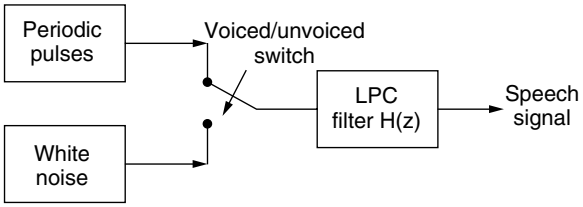


Figure 9. Typical speech synthesis system using LPC.

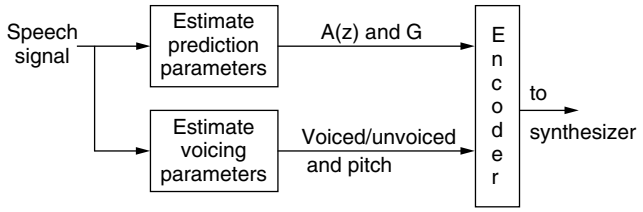


Figure 10. Speech analysis system for use with the synthesizer in Fig. 9.

produces the best output speech, then transmits the index of this best excitation to the decoder along with the LP parameters and gain. If the codebook is composed of 1000 possible excitations, for example, 10 bits would be needed to encode the index information. This information is transmitted every 5 ms; consequently 2000 bps are required for the excitation information. Often in CELP, two predictors are employed. The first predictor is used to remove the short-term redundancy as described earlier, while the second predictor is used to remove the long-term redundancy associated with the periodicity of voiced speech. In the latter case, the predicted sample is no longer based on neighboring samples, but on ones a pitch period away. In essence the long-term predictor carries the pitch information. The set of parameters associated with CELP are therefore the excitation index in the codebook, the LP coefficients, the long-term predictor coefficients, and the delay of the long-term predictor. CELP offers improved output speech quality when compared to the system in Figs. 9 and 10 at the expense of an increase in bit rate to around 4800 bps.

In MELP, the excitation signal is chosen as a combination of the excitation functions in Fig. 9. Each excitation function (i.e., periodic pulses and white noise) is passed through a multiband filter before they are combined. As a result, the excitation signal is considered to be voiced in some frequency bands and unvoiced in others. This model reflects more closely the true nature of speech and, as a result, produces more natural output speech. The encoder in MELP transmits the voiced/unvoiced information associated with each band, as well as pitch information, gain and LP coefficients. Using efficient encoding and quantization, MELP produces good quality speech, rivaling that of CELP at about half the bit rate, namely, 2400 bps. Methods aimed at reducing the bit rate are currently under investigation with goals of LPC-based coders operating in 600–1200 bps range while producing high quality natural-sounding speech.

BIOGRAPHY

Amro El-Jaroudi was born in Cairo, Egypt, in 1963. He received his B.S. and M.S. degrees in 1984, and his Ph.D. in 1988 from Northeastern University in electrical engineering. In 1988, he joined the Department of Electrical Engineering at the University of Pittsburgh,

Pennsylvania, where he is now associate professor. His research interests include digital processing of speech signals with applications to speech coding and recognition, spectral estimation of nonstationary signals, and pattern classification.

BIBLIOGRAPHY

1. J. Makhoul, Linear prediction: A tutorial review, *Proc. IEEE* **63**(4): 561–580 (April 1975).
2. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
3. C. Lee, Robust linear prediction for speech analysis, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, April 1987, pp. 289–292.
4. A. El-Jaroudi and J. Makhoul, Discrete all-pole modeling, *IEEE Trans. Signal Process.* **39**(2): 411–423 (Feb. 1991).
5. M. R. Schroeder and B. S. Atal, Code-excited linear prediction (CELP): High quality speech at very low bit rates, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, March 1985, pp. 937–940.
6. A. V. McCree and T. P. Barnwell III, Mixed excitation LPC vocoder model for low bit rate speech coding, *IEEE Trans. Speech Audio Process.* **3**: 242–250 (July 1995).

LOCAL MULTIPOINT DISTRIBUTION SERVICES (LMDS)

Peter Papazian
Roger Dalke
Institute for Telecommunication Sciences
Boulder, Colorado

1. INTRODUCTION

LMDS is the acronym for Local Multipoint Distribution Service, a broadband wireless access (BWA) service being developed in the United States at millimeter wave(length) frequencies. Similar BWA services, sometimes with different names but also at millimeter wave frequencies, are concurrently being developed and deployed in Canada [Local Multipoint Communication Service or (LMCS)], Europe, Asia, and Central and South America. As the name implies, LMDS is a short-range (local), point-to-multipoint broadcast service. The service will allow two-way communication and has been allocated more than 1 GHz of radio spectrum in the United States. This large bandwidth enables high-speed (high-bit-rate) wireless communication. LMDS is envisioned as a wireless link to a metropolitan-area network (MAN) capable of providing simultaneous interactive video, digital telephony, data, and Internet services. These services are allowed two modes of operation: point-to-point and broadcast. The point-to-point mode operation is similar to fixed microwave links. However, larger link budgets must be allocated for signal fading due to rain and for attenuation due to atmospheric adsorption. Point-to-point radio links can serve medium to large size business customers and have also been used to provide service to niche markets, small areas not served by cable or urban buildings where cable or fiber would be too expensive to install. The broadcast service was initially envisioned as providing internet, video, and telephony services to consumers on a large scale. This market has been slow to develop due to the costs of infrastructure development and the technical difficulties of obtaining adequate signal coverage. Both of these factors have made these systems economically unfeasible to deploy in the United States so far.

The advantages and disadvantages of LMDS are related to the use of the extremely high/superhigh frequency (EHF/SHF) or millimeter wave portion of the radio spectrum. The millimeter wave spectrum allows some equipment miniaturization and has large available bandwidths necessary for high-speed digital communication. But the high radiofrequencies also cause problems due to radiowave propagation impairments, the higher cost of electronic components and unavailability of high power solid-state linear amplifiers. For a summary of the most recent advances in amplifier technology, see Ref. 1.

The remainder of this article is organized in the following manner. First an overview of the LMDS band (spectrum) allocation and some technical rules related to the use of this band are given. This section also discusses work done by standards groups to help speed development and deployment of LMDS. Then millimeter wave radiowave propagation impairments are presented analytically. Finally, radiowave propagation measurements for an LMDS broadcast system are summarized.

2. REGULATORY AND STANDARDS OVERVIEW

Figure 1 is the band allocation chart for LMDS in the United States as specified by the Federal Communications Commission (FCC) [2]. Under this band plan, two blocks of frequencies (A and B) near 30 GHz are allocated in 493 basic trading areas (BTAs). These areas are defined in the *Rand McNally Commercial Atlas* and details about BTAs can also be obtained from the FCC website (*www.fcc.gov*). The LMDS spectrum was then licensed by block and BTA to successful bidders at the FCC LMDS spectrum auction. Block A has 1150 MHz of radio spectrum, and block B has a 150-MHz allocation. A critical issue for LMDS operation is the task of avoiding interference at BTA boundaries and in shared bands. This task is called *frequency coordination and interference control*. Frequency coordination and interference rules are more difficult to define for systems that can operate in the broadcast mode such as LMDS than for point-to-point operation more typical in the microwave bands. Usually the procedure requires a knowledge of the following transmitter and receiver parameters: (1) *effective isotropic radiated power* (EIRP) or *power flux density* (PFD), (2) channelization and frequency plan, (3) modulation type and channel bandwidth, (4) frequency stability, (5) receiver parameters (noise figure, bandwidth, and

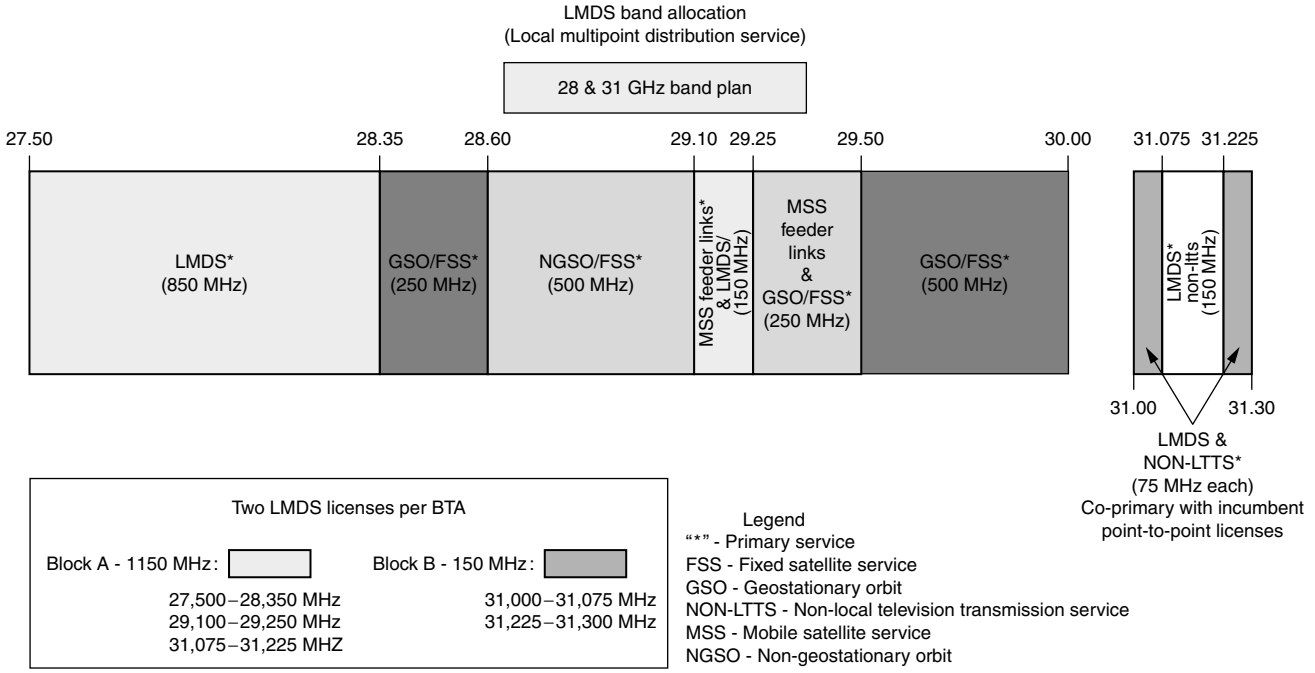


Figure 1. LMDS band allocation chart for the United States, source FCC.

thresholds), (6) antenna characteristics, and (7) system geometry. The FCC declined to set these standards, citing the technical difficulties of calculating a reasonable limit and a lack of support by industry for establishment of such a limit. Instead, frequency coordination between adjacent BTAs was left to a cooperative effort between license holders as specified in Section 101.103(d) of the *Code of Federal Regulations* (CFR) [3]. These coordination rules are applied to LMDS stations within 20 km of the BTA boundary. Within each BTA and frequency block, operators are left with the task of establishing their own frequency coordination rules to avoid interfering with adjacent hubs in their own cellular-type broadcast or point-to-point system. The FCC did set maximum allowable EIRP for any system by frequency band; these limits are listed in Table 1.

Referring to Fig. 1, we see that between 29.1 and 29.25 GHz, LMDS coexists with mobile satellite feeder links. These coexistence rules are specified by the FCC to protect existing satellite links. Since LMDS stations can transmit in the point-to-point mode as well as in the broadcast mode, two types of EIRP coexistence rules were specified. For point-to-point narrowband operation, the EIRP per carrier is limited to −23 to −26 dBW/MHz, depending on climate zone. To prevent LMDS broadcasting base stations from interfering with mobile satellite stations, the EIRP aggregate power spectral density per unit area for all LMDS hub transmitters in a BTA is limited to between −23 and −26 dBW/MHz · km^2 [3,4].

Table 1. FCC EIRP Limitations by Band for LMDS Systems

	Maximum Allowable EIRP	
Frequency Band (GHz)	Fixed (dBW/MHz)	Mobile (dBW/MHz)
27.50–28.35	30[a]	—
29.10–29.25	−23 to −26[b]	—
31.00–31.075	30	30
31.225–31.30	30	30

[a] 42 dBW/MHz for subscriber terminals.
[b] See text.

It was envisioned by the FCC that each operator would install a sufficient number of base stations in the BTA to meet subscriber demand, develop interference and coexistence rules, and manage frequency reuse in their own frequency block. The IEEE Wireless LAN/MAN Standards Committee group, 802.16, has been convened to develop these coexistence and channelization rules. This will be accomplished by defining the Physical and media access control (MAC) layer standards for proposed LMDS systems. At present only a draft standard is available. The work of the 802.16 group can be retrieved online from the IEEE web site (*www.ieee.org*). Results from 802.16 that are available indicate some of the channelization and the capacity or spectral efficiency of the proposed LMDS and BWA systems both in the United States and abroad. See Table 2 for a summary of these proposed frequencies and aggregate transmission rates. It should be noted that 802.16 regards its work as encompassing both LMDS and other BWA systems in the millimeter wave and microwave frequency range. This can cause some difficulty when trying to focus on LMDS standards.

The FCC has also required operators to provide a substantial level of service in their BTA. For an LMDS license that provides point-to-multipoint service, coverage to 20 percent of the population in the service area at the 10-year mark would constitute substantial service. For a license holder choosing to deploy point-to-point service,

Table 2. Summary by Country of Frequency Allocations and Estimated Aggregate Data Rates for LMDS and BWA Services Operating in the Millimeter-Wave Frequency Range

Country	Frequency (GHz)	Bandwidth (MHz)	Proposed Rates[a] (Mbps)
USA	LMDS block A 28,29,31	1150	862
	LMDS block B 31	150	115
	38 (point to point)	$N \times 50$	$N \times 75$
	40	Future	
	60	Future	
Canada	LMCS 25–28	3000	
Japan	23–28	Various	
Europe	26	Various	
	40	3000	
	28	LMDS equivalent	
Korea	25–27		
Asia	26–31,38	Various	

[a]Total data rate for the entire frequency block.

four permanent links per million people in the service area at the 10-year mark would constitute substantial service. More details on these rulings can be found in Refs. 2–4.

3. MILLIMETER WAVE PROPAGATION

3.1. Clear Air Absorption

At frequencies above 10 GHz, radiowaves propagating through the atmosphere are subject to molecular absorption. Although typical LMDS frequencies near 28 GHz are in a "window" — comfortably between the water vapor absorption line at 22 GHz and the band of oxygen lines near 60 GHz — there will nevertheless be some residual effects from the tails of these and other lines. Such effects can be evaluated using the millimeterwave propagation model of Liebe [5,6] (see also Rec. ITU-R P.676-4 [7]). For example, Fig. 2 shows clear air absorption as a function of frequency and humidity for a standard atmosphere (15 °C, 1013.25 mbar) and a frequency range of 10–100 GHz. Figure 3 shows the absorption as a function of relative humidity and temperature for 28 GHz and 1013.25 mbar. Note that on a hot, muggy day, a 6-km path could suffer perhaps 5 dB clear air attenuation.

3.2. Effects of Rain

Absorption and scattering of radiowave energy, due to the presence of raindrops, can severely degrade the reliability and performance of communication links. Attenuation resulting from propagation through raindrops is perhaps the most significant threat to line-of-sight (LoS) radio links operating in the millimeter waveband. For such systems, reliability predictions based on rain attenuation alone are often sufficient, since the error due to the exclusion of the other atmospheric propagation effects is much less than

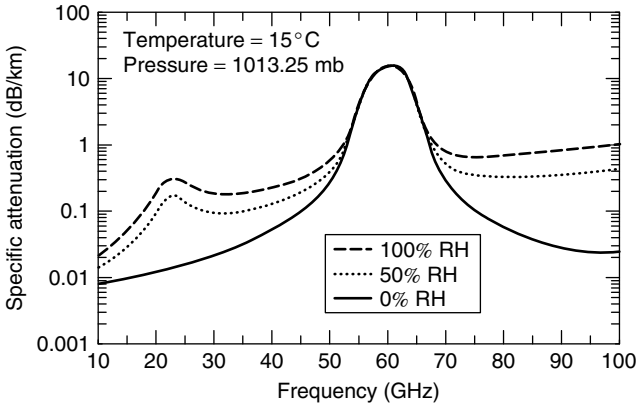


Figure 2. Clear air absorption as a function of frequency and humidity at 15 °C and 1013.25 mbar.

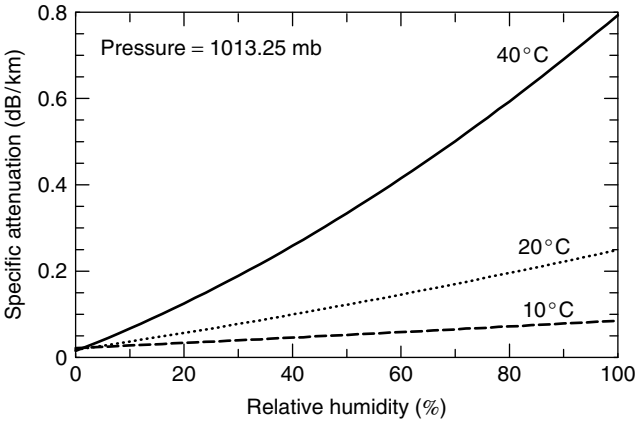


Figure 3. Clear air absorption as a function of relative humidity and at 28 GHz and 1013.25 mbar.

the normal year-to-year variation in rain attenuation. In general, rain-induced dispersion (frequency selectivity) is not considered significant for bandwidths of less than 1 GHz [8]. Note that proposed channelization schemes within the LMDS bands are all much smaller than 1 GHz, hence this effect can be ignored.

Rain attenuation is a function of drop shape, drop size, rain rate, and wavelength. Since the drops are randomly distributed in the atmosphere, the net scattering is an incoherent superposition of contributions from individual drops. The power scattered and absorbed per drop for a unit incident energy flux is called the *absorption cross section* σ, which for spherical drops is a function of the wavelength, drop radius, and the refractive index.

In traversing an incremental distance ds through spherical raindrops of radius r, the fractional loss of flux is $n_r \sigma ds$, where n_r is the number of drops per unit volume with radius r. The beam intensity decays exponentially, i.e., $I(x) = I_0 e^{-ax}$, where $a = n_r \sigma$ is the attenuation or extinction coefficient. In a rain storm, the actual drop sizes vary with rain rate and type of storm activity, and hence, the total attenuation is obtained by summing the

contribution from all drop sizes or

$$a = \int \sigma(r, \lambda, m) n(r)\, dr$$

where $n(r)$ is the drop size distribution, λ is the wavelength, and m is the complex refractive index. The specific attenuation over a path of length L in decibels per unit length is

$$\alpha = \frac{10 \log_{10}\{I_0/I(L)\}}{L} = 4.343\, a$$

The drop size distribution is a function of rain rate and type of storm activity and is well represented by an exponential of the form [9]

$$n(r) = N_0 e^{-cR^{-d}r}$$

where R is the rain rate, in mm/hr, r is the drop radius in millimeters, and c and d are empirical constants. The absorption cross section can be calculated using the classic scattering theory of Mie for a plane wave incident on an absorbing sphere [9]. For frequencies of $\leq$40 GHz, where the wavelength is much greater than the drop size, the *Rayleigh approximation* can be used. The Rayleigh scattering cross section is given by

$$\sigma = \frac{8\pi^2}{\lambda} r^3 \mathrm{Im}\left[\frac{m^2 - 1}{m^2 + 2}\right]$$

where Im refers to the imaginary part of the argument.

Integrating over all possible drop sizes and assuming Rayleigh scattering gives a relatively simple relationship between the specific attenuation and the rain rate: $\alpha = aR^b$ (dB/km). The coefficients a and b depend on the drop size distribution, refractive index, and frequency. By convention, the coefficients are given for rain rates in mm/h. This result is consistent with direct measurements of attenuation and is in agreement with Mie scattering [9] over a wide frequency range.

Several investigators have studied the distribution of raindrop sizes as a function of rain rate and type of storm activity. Olsen et al. [10] give tables of coefficients for several spherical drop size distributions as a function of temperature and frequency. The most commonly used distributions are those of Law and Parsons (LP), Marshall and Palmer (MP), and Joss and Waldvogel (JW). Law and Parsons propose two distributions, the LP(L) distribution for widespread rain (with rates less than 25 mm/hr), and the LP(H) distribution for convective rain with higher rates. In general the LP distributions seem to be favored for design purposes because they have been widely tested and compared to measurements. The LP(L) distribution gives approximately the same specific attenuation as the JW thunderstorm distribution, and the specific attenuation of the MP and LP(H) are approximately the same. Allen [11] points out that for millimeter waves there is a range of more than a factor of 2 in specific attenuation for different drop size distributions used by Olsen et al. and a range of a factor of 4 for the different climate regions used by Dutton et al. [12]. The resulting uncertainty is a critical limitation to predicting link reliability.

In general, drops are not spherical, in which case the coefficients depend on polarization. Coefficients for vertically and horizontally polarized electromagnetic waves have been calculated for oblate spheroidal drops using methods similar to those described above. Coefficients for nonspherical drops and methods for calculating coefficients for arbitrary polarizations are given in ITU-R P.838-1 [13]. Figure 4 compares the specific attenuation at 30 GHz as a function of rain rate for L-PL coefficients (spherical drops at 20 °C) and ITU coefficients for horizontal and vertical polarization. Note that vertical polarization provides a significant advantage for lengthy paths in moderate to severe rain.

In principle, the total attenuation is obtained by integrating the specific attenuation over a particular path. Accurately modeling the total attenuation is difficult since rain rate is a nonstationary random process with short and long term, as well as global and local variations. Local variations occur because the vertical distribution of precipitation varies with temperature as a function of height. Also, intense rain tends to be localized and the rain rate can vary significantly over terrestrial paths.

Two important models that are commonly used to predict terrestrial path attenuation are the global model of Crane [14] and the ITU model (ITU-R P.530-8 [15]). Both models provide empirical formulas for calculating an *effective pathlength* L_{eff} that is a function of the rain rate. The path attenuation is then the product of the specific attenuation based on the locality or *point* rain rate and the effective pathlength

$$A(\mathrm{dB}) = aR^b L_{\mathrm{eff}}$$

If measured rain rate statistics for the desired location are not available, both the Crane global model and the ITU model give methodologies for estimating rain rate statistics for an *average year*. The ITU model provides global data for calculating rainfall statistics with grid points spaced at 1.5° intervals in both latitude and longitude. The Crane global model partitions the world into 12 rain climate zones based on the assumption that the location-to-location variability within a zone

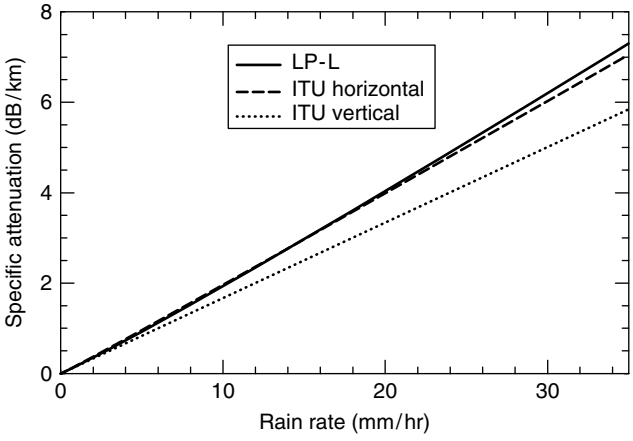


Figure 4. Specific attenuation as a function of rain rate for LP(L) (spherical drops at 20 °C) and ITU drop distributions.

is comparable to the year-to-year variation at a point. Location-to-location variability and year-to-year variation in the rain rate distribution are reported to be lognormal with a standard deviation of 50% for all climate regions [16]. The Crane model is widely used and has been shown by Dutton [17] to be one of the better models.

The Rice–Holmberg [18] global surface rain rate model can be used to calculate local rain rate statistics using historical meteorological data. This model is based on extensive long term rain rate statistics from 150 locations throughout the world. The Rice–Holmberg model gives the rain rate distribution in terms of commonly recorded climatologic parameters: the average annual rainfall accumulation and the average annual accumulation of thunderstorm rain. According to this model, the cumulative distribution of 1-minute average rainfall rates for an *average year* is given by

$$P\{R > \rho\} = \frac{M}{8766}\{0.03\beta e^{-0.03\rho} + 0.2(1-\beta) \times [e^{-0.258\rho} + 1.86e^{-1.63\rho}]\}$$

where M is the average annual rainfall accumulation in mm and β is the average annual ratio of thunderstorm rain to total rain. The required climatological data can be obtained from a variety of sources. Perhaps the best source for the rainfall data is the National Climatic Data Center (NOAA/National Weather Service, Asheville, North Carolina) which is the world's largest active archive of weather data. Extensions to the Rice–Holmberg model that include year-to-year and location-to-location variability are given by Dutton [19].

As an example, consider radio links of less than 20 km in the vicinity of San Francisco, California. Using the Rice–Holmberg model and historical data from a local weather station [20], the 1-min rain rate exceeded less than .01% of an average year is found to be 18.6 mm/h. Considering the LP(L) and the ITU coefficients for vertically and horizontally polarized waves, the specific attenuation for 18.6 mm/h is obtained from Fig. 4. The effective path lengths calculated using both Crane global model and ITU procedures are shown in Fig. 5. The total path attenuation is the product of the specific attenuation and the effective pathlength that corresponds to the actual pathlength. For a 10-km link assuming an LP(L) drop distribution ($\alpha = 3.73$ dB/km), the total path attenuation is 37.3 dB according to the Crane global model and 27.1 dB according to the ITU model. These results give the total attenuation exceeded less than 0.01% of an average year.

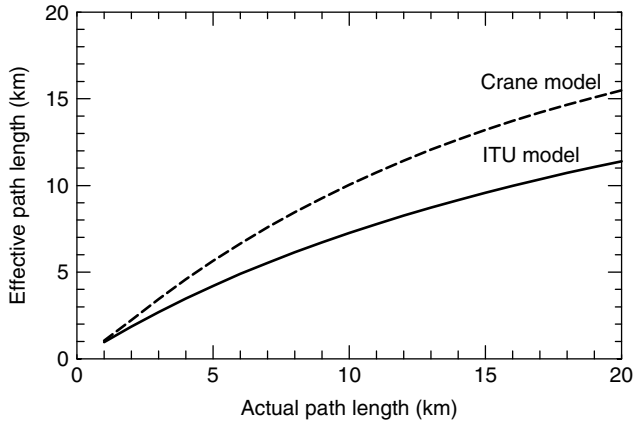


Figure 5. Effective pathlength based on the Crane global and ITU models assuming a rain rate of 18.6 mm/h.

3.3. Rain-Induced Depolarization

Rain-induced depolarization is due to differential attenuation and phase shifts caused by nonspherical raindrops. The classic model for a falling raindrop is an oblate spheroid with its major axis canted to the horizontal and with major and minor axes related to the radius of a sphere of equal volume. For practical applications a semiempirical relationship between rain attenuation and depolarization is provided by Ippolito [9] (see also Ref. 15):

$$\mathrm{XPD} = 30\log_{10} f_{\mathrm{GHz}} - 10\log_{10}(0.5 - 0.4697\cos 4\tau) - 40\log_{10}(\cos\theta) - 23\log_{10} A$$

where XPD is the "cross-polarization discrimination," that is, the ratio (in decibels) of the copolarized and cross-polarized field strengths, where τ is the tilt angle of the polarization with respect to horizontal, θ is the elevation angle of the path, and A is the rain attenuation in decibels. For 30-GHz terrestrial links ($\theta \approx 0$) with attenuation of less than 15 dB, and horizontal ($\tau = 0$) or vertical ($\tau = \pi/2$) polarization, the effects of rain-induced depolarization are quite small (XPD > 30 dB).

3.4. Attenuation Due to Fog

Fog results from the condensation of atmospheric water vapor into water droplets that remain suspended in air. There are two main types of fog. Advection fog is coastal fog that forms when warm, moist air moves over colder water. Liquid water content of advection fog does not normally exceed 0.4 g/m^3. Radiation fog forms inland at night, usually in valleys and low marshes, and along rivers. Radiation fog can have a liquid content of up to 1 g/m^3.

Specific attenuation for fog can be calculated using a model developed by Liebe [6]. Using this model and assuming dense fog conditions with 1 g/m^3 water result gives a specific attenuation of 0.5 dB/km. For a homogeneous fog path of 6 km, the total attenuation is 3 dB.

4. LMDS RADIO CHANNEL

Signal impairments due to atmospheric gases, rain, and rain depolarization are important radiowave propagation factors that can be computed using methods outlined in the previous sections. However, millimeter wave signal dispersion due to multipath and attenuation, and depolarization due to random distributions of vegetation are system and site dependent and must be measured. For example, multipath measurements are highly dependent on the beamwidth of the transmitting and receiving

antennas. Since narrow beam antennas will filter out multipath signals, a multipath metric such as delay spread (S), will be smaller for LMDS point-to-point systems using a narrow beamwidth antenna than an LMDS broadcast system using a wider beamwidth antenna. The site dependence of these parameters for broadcast systems is also critical. For instance, the percentage of LoS paths to potential subscribers (and hence signal attenuation) will vary depending on transmitter location and environment. An urban high-rise or hilly suburban environment will suffer more blocked paths than a flat rural environment with little vegetation. When dealing with attenuation due to vegetation, further environmental distinctions must be made to account for vegetation type, density, and distribution.

Fortunately some aspects of the millimeter wave radio channel can be modeled using advanced computer methods. Standard radio propagation models can calculate diffraction and signal blockage due to terrain by incorporation of digital terrain data for an area. More advanced models have been developed for millimeter wave and LMDS applications that incorporate higher resolution terrain and building elevation data from aerial photographs. These programs can then determine LoS

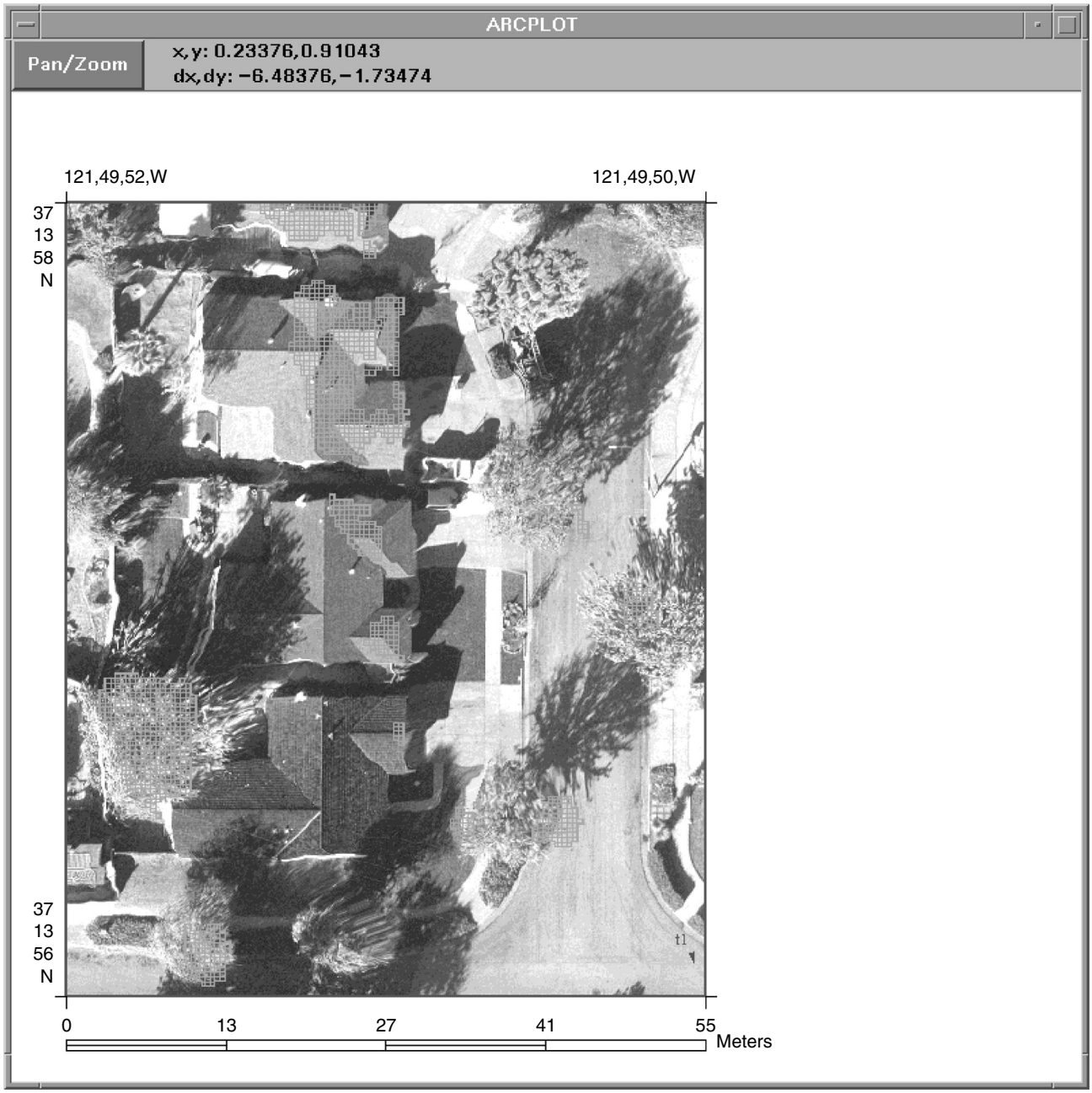


Figure 6. LMDS propagation modeling program output overlaid on an aerial photograph of San Jose, CA. Purple areas indicate line-of-sight coverage from a 12 m high transmitter located to the southeast.

Table 3. Measurement Equipment Parameters

	Antenna Beam Width (degrees)		EIRP (dBm)	Sensitivity (dBm)	
	Vertical	Azimuthal		Narrowband	Wideband
Transmitter	20	90	51	N/A	N/A
Receiver	7.5	7.5	N/A	−130	−102

paths at millimeter wave frequencies. Figure 6 is an example of such a computer simulation.[1] This figure is an aerial photograph of several houses in San Jose, CA. The purple areas indicate LoS coverage from a 12-m transmitter located to the southeast. Computer processing of the photograph was used to develop the surface contour used in conjunction with latitude, longitude and elevation of the transmitter location to determine LoS paths. However, these efforts still lack the ability to incorporate multipath and attenuation, diffraction, and depolarization due to vegetation. To incorporate these effects, measurement data are required.

4.1. LMDS Broadcast System Area Coverage and Radio Channel Measurements

LMDS is ideal for providing last-mile connectivity to a fixed, broadband network. To achieve this it is important to know the area coverage and the radio channel characteristics for specific sites and proposed systems, including vegetation. A typical last mile solution in suburban neighborhoods will utilize broadcasting base stations arranged in a cellular pattern with low antenna heights and spaced on a 1–2-km grid. Consumers could then install small directional antennas aimed at the base station. Typically the forward link from the base station would be high power and high bit rate while the reverse link would be low power and low bit rate. To quantify the percentage of households that can be reached (coverage achieved), as well as the radio channel characteristics, a set of 30-GHz radiowave propagation measurements is described below.

These measurements include area coverage, signal attenuation, signal depolarization, and the delay spread (S) for an LMDS radio channel. The measurement system transmits a 28.8-GHz narrowband continuous-wave (CW) signal and a 30.3-GHz wideband signal through a common traveling wavetube amplifier. The wideband signal, used to measure the radio channel impulse response, was created by modulating the carrier with a 500-Mb/s pseudo-random-noise code. The transmitter used a single vertically polarized horn with 14 dB gain. The antenna had a 90° azimuthal 3-dB beamwidth, and a 20° vertical beamwidth. The EIRP for the transmitter was 51 dBm (−6 dBW/MHz for 500 MHz BW).

The receiver antenna system consisted of two 7.5° dishes with linearly polarized feeds. One dish was aligned for vertical polarization, and the second was aligned for horizontal polarization. The received signals were split and processed in separate narrowband and wideband receivers. The wideband receiver provided cophase and quadrature-phase impulse response data with 2-ns resolution. The receiver had a sensitivity of −102 dBm and a dynamic range of 50 dB. The narrowband receiver was used to measure received signal power. It had a sensitivity of −130 dBm and a dynamic range of 70 dB. Some relevant equipment parameters are listed in Table 3.

4.2. Environment

The measurement area consisted of one- and two-story single-family residences in Northglenn, Colorado and San Jose, California. Both areas have small yearly rainfall totals and slow tree growth. Some relevant geographic statistics for each site are listed in Table 4.

Factors that can affect coverage at these sites include rainfall, terrain, shadowing by buildings, and attenuation by vegetation. Since the terrain at both survey sites is flat, this is not an issue when comparing results between the two sites. The distributions of roof heights for each site were estimated from measured data and are also similar. The most important difference between the sites is the vegetation, in particular the tree canopy. The tree population in Northglenn is dominated by elms, maples, cottonwoods, and ponderosa pines. Mature trees of these species are 9–15 m tall. In contrast, many trees in San Jose have tropical origins and are only 6–9 m tall.

4.3. Measurement Procedures

Both narrowband and wideband data were collected. The narrowband data includes a time series record of the signal power, which was used to study area coverage, short-term variations of the signal and depolarization. These data were recorded at 1000 samples/s for 50 s. Wideband data, used to measure multipath, consisted of 100 complex impulse responses at each site. Each impulse lasted for 254 ns and was sampled 1000 times. The repetition rate of

Table 4. House Density, Normal Temperature, and Rainfall Averages for Northglenn, CO, and San Jose, CA

Geographic Statistics	Northglenn, Colorado	San Jose, California
Number of houses/km^2	780	900
Temperature (°F)[a]	50.3	59.7
Precipitation (in.)[b]	15.31	13.86

[a] Monthly average.
[b] Yearly normal between 1951 and 1980.

[1] The user's guide to CSPT (communications system planning tool) is available from ITS by request. The software is available free of charge to users.

the impulses from the sliding correlator was 10 Hz. Both data sets were collected using vertical (copolarized) and horizontal (cross-polarized) receive antennas.

The receiver address was determined by randomly selecting houses, using aerial photographs of the survey area. Because it was assumed that the probability of acceptable coverage would decrease with distance, each broadcast cell was first subdivided into bands of increasing radii from the transmitter. Stations (houses) were then selected randomly from equal area subdivisions of each band. Figure 7 shows a typical 0.5-km square cell quadrant with its three sampling bands. The number of stations needed for an acceptable error was determined by assuming that the area coverage estimate could be modeled using a binomial distribution (see the next section for a description of this model).

At each receiver station the curbside location of the measurement van was selected using both aerial photographs and onsite inspection to avoid obvious obstructions between the roof of the house and the transmit antenna. The receive antenna height was determined using a mast-mounted videocamera to locate the height of the roof peak above street level and then by raising the mast an additional meter. The optimum receiver antenna azimuth and elevation angle were then determined using narrowband, vertically polarized, azimuth, and elevation scans to find the direction of maximum received power.

It was desired to estimate coverage for cells that could be separated into four symmetric quadrants. To save time, only one quadrant of each cell was sampled. It was assumed that the other quadrants would be sufficiently uniform and would produce similar results for the entire cell. In Northglenn, two 0.5-km (Fig. 7) square cell quadrants were surveyed using different 12-m-high transmitter locations, and the area coverage results were compared and found to yield similar results. In San Jose, one 0.5-km square cell quadrant and a 1-km circular cell quadrant were surveyed using a 12-m-high transmitter site. To study the area coverage dependence on transmitter height, a 24-m-high transmitter site was added. The 0.5-km quadrant and 1-km quadrant were re-surveyed using this transmit antenna. Then a 2-km circular cell quadrant was surveyed, also using the 24-m transmit antenna to study the coverage dependence on transmitter height.

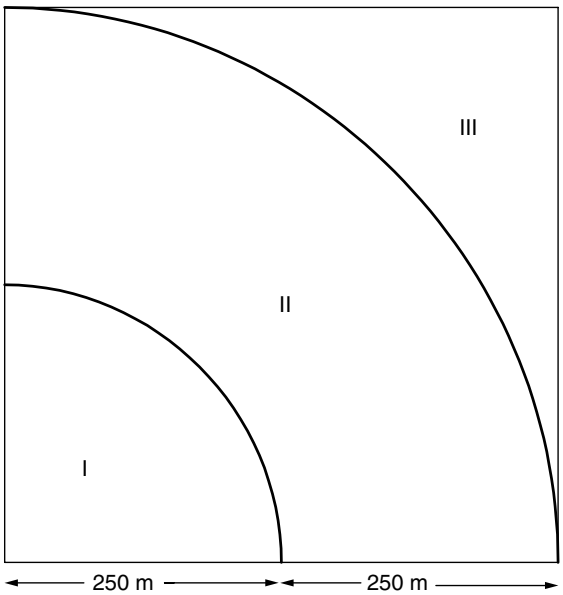


Figure 7. 0.5-km cell square cell quadrant with transmitter located in the lower left corner. Areas I, II, III indicate sampling zones for the calculation of area coverage statistics.

4.4. Area Coverage Model

The area coverage estimates are based on copolarized (vertical) received power data. The coverage in each cell band can be estimated as the fraction of houses for which an adequate signal is available for a given percentage of the time. If p_i is the area coverage probability for the band, n_i is the number of houses sampled, and n_{i1} is the number of houses in the ith band that meet the signal level requirements for coverage, the area coverage estimate is

$$p_i = \frac{n_{i1}}{n_i}$$

Assuming that the number of houses with coverage is binomially distributed and the area is sampled without replacement, the standard error σ_i in each cell band can be approximated as [21]

$$\sigma_i = \sqrt{p_i(1-p_i)\left(\frac{1}{n_i} - \frac{1}{N_i}\right)}$$

where N_i is the number of houses in the ith band. The area coverage p_c and error estimates σ_c for each cell are calculated by weighting the results from each band using their relative area a_i and summing the results from each band as follows:

$$p_c = \sum_{i=1} a_i p_i$$

$$\sigma_c = \sqrt{\sum_i a_i^2 \sigma_i^2}$$

4.5. Area Coverage Metric

The metric used to determine area coverage is *basic transmission loss* (L_b). L_b is the signal loss expected between ideal, loss-free, isotropic transmitting and receiving antennas [22]. This loss is a function of the frequency, pathlength, and attenuation on the path. The major source of attenuation in our survey area was obstruction of the radio path by buildings and vegetation.

Coverage is the percent of locations for which L_b does not exceed the allowable loss ($L_b^{\max}$) for a given system at the desired availability level. If one knows the operating parameters for a radio system, then an $L_b^{\max}$ can be determined based on the available transmitter power and the necessary SNR at the receiver to achieve the required bit error rate (BER). A station then has coverage if $L_b \le L_b^{\max}$.

Availability is based on the time variability of the received signal measured at each station. The cumulative

distribution function (CDF) of the received power is used to calculate the time statistics (i.e., availability) of L_b. For instance, the median signal power measured at a receiver station gives L_b for 50% availability, while the lower decile gives L_b for 90% availability. Using L_b calculated at specific availability levels and the statistical development of the previous section, area coverage and standard error estimates are made for a range of $L_b^{\max}$. As one would expect, coverage decreases at increased availability levels. Because a high level of availability is desirable, we have summarized area coverage results versus $L_b^{\max}$ assuming 99% availability. For coverage estimates at higher availability levels, more independent measurements would be required.

Because coverage estimates for both Northglenn and San Jose are similar, a sample of results from both sites are used to illustrate the general trends. Area coverage for a 0.5-km cell quadrant and a 1.0-km cell quadrant, both using a 12-m-high transmitter site in San Jose, are shown in Fig. 8. From the figure we can see that systems capable of sustaining an $L_b^{\max}$ of 150–155 dB can achieve 80% coverage at 99% availability in 0.5-km cell quadrants (1-km transmitter spacing). For the 1-km quadrant (2-km transmitter spacing), the coverage for L_b between 150 and 155 dB decreases to 75% versus 80% measured in the 0.5-km quadrant. In Fig. 9, we see that the area coverage for San Jose is improved significantly for the 0.5- and 1-km quadrants by using a 24-m-high transmitter site. Now, 80% coverage for the 0.5-km quadrant can be achieved at an L_b of 140 dB, 10–15 dB less signal loss than the 12-m transmitter results. When using a 24-m-high transmitter in the 1-km quadrant, 80% coverage can be reached at an L_b between 145 and 150 dB. For the 2-km quadrant we see that 80% coverage is not achieved for an L_b up to 155 dB. More details can be found in Ref. 23.

4.6. Attenuation

Attenuation is the additional power loss above the free space loss (spreading loss) between the transmit and receive antennas. It is convenient to separate L_b into its two components, attenuation A and basic free-space loss L_{fs}:

$$L_b(\text{dB}) = L_{fs}(\text{dB}) + A(\text{dB}).$$

Using this relationship, A is calculated by subtracting L_{fs} from L_b where L_{fs} is

$$L_{fs}(\text{dB}) = 32.4 + 20\log f(\text{MHz}) \cdot d(\text{km}).$$

An attenuation versus distance graph for San Jose using the 12-ft transmitter site is shown in Fig. 10. The data is highly scattered due to the random nature of the obstructions. However, a general trend can be seen by overlaying a linear least squares fit curve on the data. Similar linear fits were made using the other Northglenn and San Jose data. The slopes and intercepts of these curves are summarized in Table 5.

The slope of the attenuation data has an expected inverse correlation with the area coverage results. Northglenn, which had a smaller coverage estimate

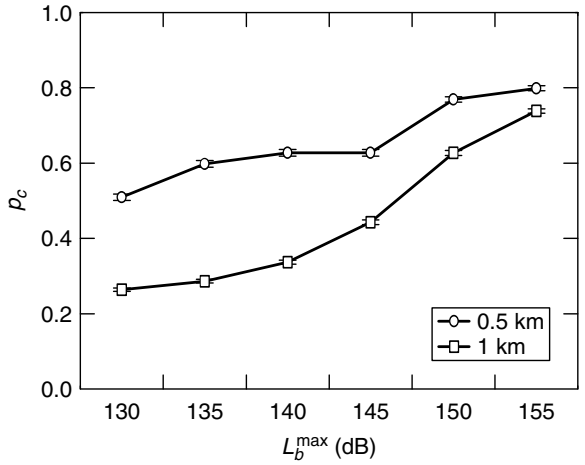


Figure 8. Area coverage estimate p_c versus $L_b^{\max}$ at 99% availability for 0.5 km and 1.0 km cells using a 40-ft transmitter, San Jose, California.

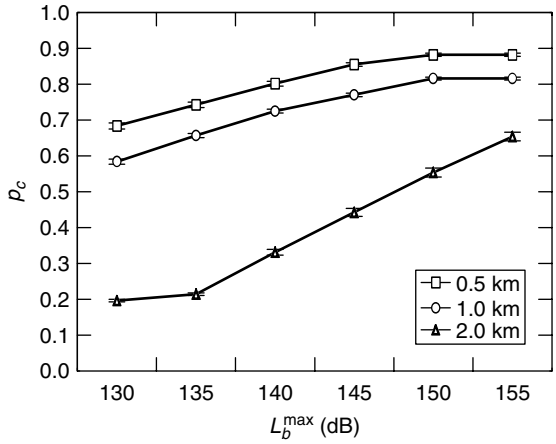


Figure 9. Area coverage probability estimate p_c versus $L_b^{\max}$ at 99% availability for 0.5-, 1.0-, and 2.0-km cells using the 80-ft transmitter, San Jose, California.

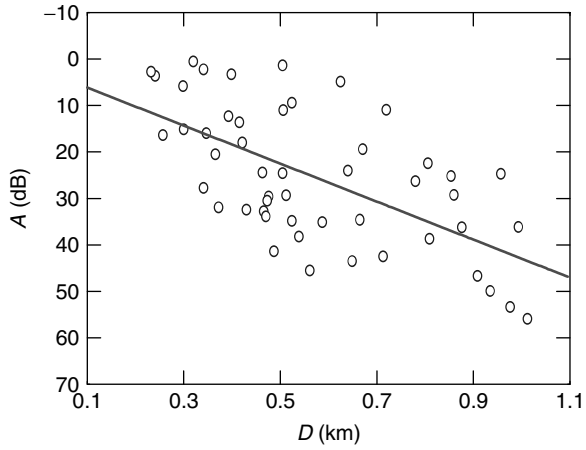


Figure 10. Attenuation versus distance using the 40-ft transmitter, San Jose, California.

Table 5. Attenuation versus Distance Slope and Intercept Data for Northglenn, CO and San Jose, CA

Site	Slope (dB/km)	Intercept (dB)
Northglenn (40-ft transmitter)	42.4	7.3
San Jose (40-ft transmitter)	40.9	2.1
San Jose (80-ft transmitter)	6.7	9.4

than San Jose, has the larger attenuation slope. The attenuation slope decreased significantly when the 24-ft transmitter site was used in San Jose, indicating that the radio path was able to clear many more obstructions. When a tree is blocking the radio path, signal propagation will be dependent on scattering and diffraction. In many cases when a tree was obstructing the radio path it was located within 10–20 m of the receiver site. For the 500-m cell, using an average pathlength of 250 m and assuming an obstruction (e.g., tree) at 235 m, the diameter of the first Fresnel zone for a 30 GHz signal is about 1 m. Usually LoS radio links require 60% of the first Fresnel zone to be free of obstructions to limit diffraction losses [22]. Hence, at least a 77-cm opening through the tree canopy is required for an unobstructed radio path.

In addition to arguments using Fresnel diffraction zones, large signal attenuation by trees is consistent with previous experiments to characterize millimeter wave propagation in vegetation [24–27]. Measurements in regularly planted orchards have found attenuation values between 12 and 20 dB per tree for one to three deciduous trees and up to 40 dB for one to three coniferous trees. The measured attenuation can be accounted for by a combination of one to four coniferous or deciduous trees on the radio path.

4.7. Cross-Polarization Discrimination

Cross-polarization discrimination measurements were made to test the practicality of frequency reuse schemes that employ signals of orthogonal polarization. A vertical linearly polarized signal was transmitted and both vertically and horizontally polarized signals were received. The larger XPD is, the more effective orthogonal frequency reuse will be. At millimeter-wave frequencies, rain-induced depolarization is produced by differential attenuation caused by nonspherical raindrops. As discussed previously, the effects of rain-induced depolarization for a short 30-GHz terrestrial link is expected to be small. Of more concern is the depolarization caused by scattering from vegetation. Experiments [24–27] have characterized millimeter wave depolarization in both coniferous and deciduous orchards. The most serious impairments are seen consistently in conifer tree stands where the average XPD at 28.8 GHz was 12 dB for foliage depths of 20 m and decreased to about 9 dB after 60 m. However, it is difficult to apply these results to cells proposed for LMDS applications because the foliage depth and tree species for any particular subscriber are random and unknown. Measured XPD results for Northglenn are presented as a function of attenuation in Fig. 11. The data are highly scattered but

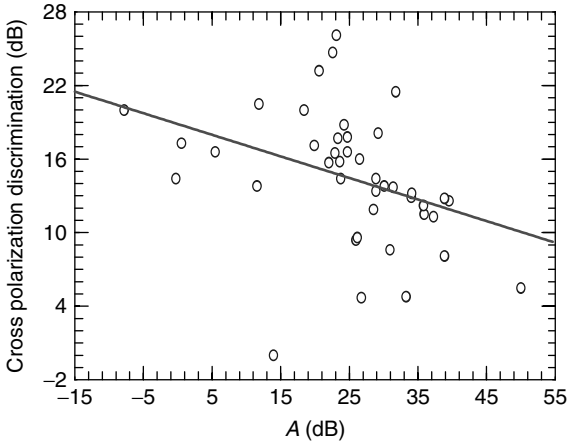


Figure 11. Cross-polarization discrimination (XPD) versus attenuation for 0.5-km cells using 40-ft transmitters in Northglenn, Colorado.

a linear fit predicts an XPD of 14 dB at an attenuation of 30 dB, which is 10 dB greater than predicted due to rain.

4.8. Characterization of Multipath Using the Tapped Delay Line Channel Model

The tapped delay line channel model is

$$h(t) = \sum_{n=1}^{N} \beta_n \delta(t - \tau_n) e^{-j\omega_c \tau_n}$$

where $h(t)$ is the complex channel impulse response, N is the maximum number of taps, n is the tap index, β is the tap gain, τ is the tap delay, and ω_c is the carrier frequency.

We selected three stations located at successively greater distances from the transmitters along the same cell radial to represent good, moderate, and bad wideband channels. Table 6 summarizes the channel model at these stations. The small delay spreads confirm that there are few specular reflections due to the filtering effect of the narrow beam receiver antennas. We note that delay spread is calculated using a 20-dB threshold. Table 7 lists the distance (D) between transmitter and receiver, attenuation (A), delay spread (S) and L_b for these paths. From Table 4 we see that links that exhibit multipath also have larger values of L_b and attenuation. Delay spreads are also plotted versus attenuation in Fig. 12. This plot also indicates that multipath is associated with larger signal attenuations.

5. SUMMARY

An LMDS band allocation was established by the FCC to provide broadband wireless access services to MANs and LANs. The spectrum allocation straddles the EHF and SHF bands near 30 GHz. Radiowaves in this part of the spectrum are commonly called millimeter waves. Although the radio bands were allocated and some technical specifications were made by the FCC, interference and coexistence rules were not defined for the broadcast

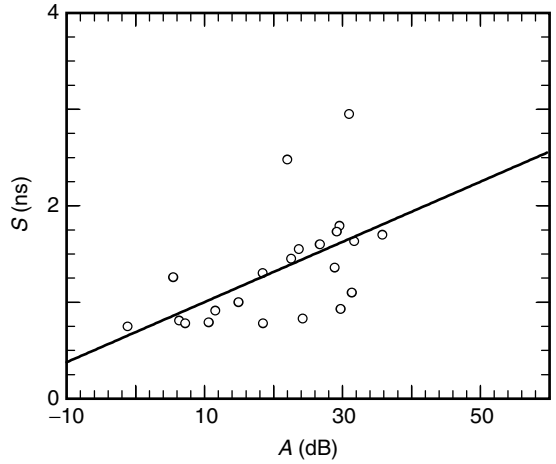


Figure 12. Delay spread (S) versus attenuation, 40-ft transmitters, Northglenn, Colorado.

Table 6. Summary of Tapped Delay Line Models for Good, Moderate, and Bad Channels from Northglenn, CO

Quality	Tap #	β_n (dBm)	τ_n (ns)
Good	1	0	0
Moderate	1	0	0
Moderate	2	−13.7	5.3
Bad	1	0	0
Bad	2	−2.8	3.6
Bad	3	−16.2	15.3

Table 7. Summary of Distance D, Attenuation A, Delay Spread S, and Basic Transmission Loss L_b at 99% Exceedance for Three Wideband Channels in Northglenn, CO

Quality	D (m)	A (dB)	S (ns)	L_b (dB)
Good	122	6.2	1.26	111.7
Moderate	309	32.2	1.60	145.9
Bad	419	32.6	2.95	159.4

mode of operation. To define these standards the IEEE LAN/MAN standards group 802.16 was established. This group is expected to publish physical layer (PHY) and media access control (MAC) Layer standards established by a consortium from private industry and government.

The advantages and disadvantages of LMDS are related to the use of the millimeter wave portion of the radio spectrum. The millimeter wave spectrum allows some equipment miniaturization and has large available bandwidths necessary for high-speed digital communication. But there are also significant problems associated with millimeter wave systems such as radiowave propagation impairments, the higher cost of electronic components, and unavailability of high-power solid-state linear amplifiers.

Radiowave propagation considerations for point-to-point links include attenuation caused by rain and atmospheric adsorption and depolarization due to nonspherical raindrops. These effects can be estimated using models and empirical formulas. Advanced computer models that incorporate high resolution areal photography and digital terrain data can be used to determine LoS paths excluding blockage due to vegetation. For broadcast systems, measurement data for specific environments must be used to determine coverage, availability levels, and radio channel characteristics such as multipath and signal depolarization. To date only limited measurement data are available at millimeter wave frequencies.

BIOGRAPHIES

Peter B. Papazian (M'91) received his B.S. in physics from the State University of New York at Stonybrook in 1973, and his M.S. in geophysics from the Colorado School of Mines in 1979. In 1990 he joined the radio research and standards group at the Institute for Telecommunication Sciences in Boulder, Colorado. At ITS, Peter has conducted research in the fields of millimeter-wave propagation, man-made radio noise, and impulse response measurements and systems. Currently, Mr. Papazian has developed an advanced antenna test-bed to study the capacity of data and mobile communication systems.

Roger A. Dalke received his bachelors degree in physics from the University of Colorado in 1971. He received his M.S. in geophysics in 1983 and his Ph.D. in 1986 from the Colorado School of Mines. As a research engineer, he has developed numerical techniques for a variety of electromagnetic scattering problems as well as signal processing and imaging methods used in exploration geophysics. More recently, he has been involved in the development of computer simulation models for digital radio systems, noise and interference measurements and analysis, and radio propagation in urban environments.

BIBLIOGRAPHY

1. R. H. Abrams, B. Levush, A. A. Mondelli, and R. K. Parker, Vacuum electronics for the 21st century, *IEEE Microwave Mag.* 61–72 (Sept. 2001).
2. FCC 96-311, *First Report and Order and Fourth Notice of Proposed Rule Making*, Docket 92-297, adopted July 17, 1996.
3. FCC 97-82, *Second Report and Order, Order on Reconsideration, and Fifth Notice of Proposed Rule Making*, Docket 92-297, adopted March 11, 1997.
4. *Code of Federal Regulations*, Title 47: Telecommunication, Section 101.103 to 101.113, Office of the Federal Register, National Archives and Records Administration, Oct. 1, 2000.
5. H. J. Liebe, MPM — an atmospheric millimeter-wave propagation model, *Int. J. Infrared Millimeter Waves* **10**: 631–650 (1989).
6. H. J. Liebe, G. A. Hufford, and M. G. Cotton, Propagation modeling of moist air and suspended water/ice particles at frequencies below 1000 GHz, *Proc. AGARD Conf. Atmospheric Propagation Effects through Natural and Man-Made Obscurants for Visible to MM-Wave Radiation*, 1993, pp. 3-1–3-11.

7. ITU-R (International Telecommunication Union, Radiocommunications Assembly), *Attenuation by Atmospheric Gases*, Rec. ITU-R P.676-4, Geneva, Switzerland, 1999.

8. R. H. Espeland, E. J. Violette, and K. C. Allen, *Atmospheric Channel Performance Measurements at 10 to 100 GHz*, NTIA Report 84-149, Apr. 1984 (NTIS Order PB 84-211325).

9. L. J. Ippolito, *Radiowave Propagation in Satellite Communications*, Van Nostrand Reinhold, New York, 1989.

10. R. L. Olsen, D. V. Rogers, and D. B. Hodge, The aR^b relation in the calculation of rain attenuation, *IEEE Trans. Antennas. Propag.* **AP-28**: 318–329 (March 1978).

11. K. C. Allen, *EHF Telecommunication System Engineering Model*, NTIA Report 86-192, April 1986 (NTIS Order PB 86-214814/AS).

12. E. J. Dutton, C. E. Lewis, and F. K. Steele, *Climatological Coefficients for Rain Attenuation at Millimeter Wavelengths*, NTIA Report 83-129, Aug. 1983 (NTIS Order PB 84-104272).

13. ITU-R (International Telecommunication Union, Radiocommunications Assembly), *Specific Attenuation Model for Rain for Use in Prediction Methods*, Rec. ITU-R P.838-1, Geneva, Switzerland, 1999.

14. R. K. Crane, Prediction of attenuation by rain, *IEEE Trans. Commun.* **COM-28**: 1717–1733 (Sept. 1980).

15. ITU-R (International Telecommunication Union, Radiocommunications Assembly), *Propagation Data and Prediction Methods Required for the Design of Terrestrial Line-of-Sight Systems*, Rec. ITU-R P.530-8, Geneva, Switzerland, 1999.

16. R. K. Crane, Comparative evaluation of several rain attenuation prediction models, *Radio Sci.* **20**(4): 843–863 (July–Aug. 1985).

17. E. J. Dutton and F. K. Steele, *Some Further Aspects of the Influence of Raindrop-Size Distributions on Millimeter-Wave Propagation*, NTIA Report 84-169, Dec. 1984 (NTIS Order PB 85-168334).

18. P. L. Rice and N. R. Holmberg, Cumulative time statistics of surface-point rainfall rates, *IEEE Trans. Commun.* **COM-21**: 1131–1136 (Oct. 1973).

19. E. J. Dutton and H. T. Dougherty, Year-to-year variability of rainfall for microwave applications in the U.S.A., *IEEE Trans. Commun.* **COM-27**(5): (May 1979).

20. National Oceanic and Atmospheric Administration, *Climates of the States*, 1978, Vol. 1, p. 136.

21. M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Macmillan, New York, 1977.

22. M. P. M. Hall, Effects of the troposphere on radiocommunication, in *IEE Electromagnetic Wave Series 8*, Peter Peregrinus Ltd., Stevenage, UK and New York, 1979, pp. 10–13, 81.

23. P. B. Papazian and G. A. Hufford, Study of the local multipoint distribution service radio channel, *IEEE Trans. Broadcast.* **43**(2): (June 1997).

24. D. Jones, R. Espeland, and E. Violette, *Vegetation Loss Measurements at 9.6, 28.8, 57.6, and 96.1 GHz through a Conifer Orchard in Washington State*, NTIA Report 89-251, Oct. 1989 (NTIS Order PB 90-168717).

25. E. Violette, R. Espeland, and F. Schwering, Vegetation loss measurements at 9.6, 28.8, and 57.6 GHz through a pecan orchard in Texas, in *Multiple Scattering of Waves in Random Media and Random Rough Surfaces*, Pennsylvania State Univ., State College, PA, 1985, pp. 457–472.

26. P. B. Papazian, D. Jones, and R. Espeland, *Millimeter-Wave Propagation at 30.3 GHz through a Pecan Orchard in Texas*, NTIA Report 92-287, Sept. 1992.

27. E. Violette, R. Espeland, and K. C. Allen, *Millimeter-Wave Propagation Characteristics and Channel Performance for Urban-Suburban Environments*, NTIA Report 88-239, Dec. 1988 (NTIS Order No. PB 89-180251/AS).

28. R. A. Dalke, G. A. Hufford, and R. L. Ketchum, *Radio Propagation Considerations for Local Multipoint Distribution Systems*, NTIA Report 96-331, Aug. 1996.

LOCAL AREA NETWORKS

JOHN H. CARSON
George Washington University
Washington, District of Columbia

1. INTRODUCTION

Local area networks (LANs) are private, high-speed networks that are limited in distance and typically serve as a distribution system for both Internet and local information services.

2. HISTORY

Although a number of research and development activities can be associated with the origin of the LAN, the best known early publication in this field appeared in 1976 [1] by Robert Metcalfe and David Boggs, who developed the Ethernet Local Area Network system while working at Xerox PARC. This coaxial cable-based system transmitted data at 2.94 Mbps. Following development of the Ethernet LAN, other local network technologies appeared and disappeared, most notable of which was the token ring system developed by IBM.

In 1980, Xerox, Digital Equipment Corporation, and Intel developed the "Ethernet Blue Book" or "DIX standard." The second version of this standard was completed in November 1982. Also in 1980, the IEEE formed the 802 Committee to standardize LAN/MAN (metropolitan area network) technology. This committee continues to develop and extend LAN standards.

2.1. Xerox PARC

As mentioned above, Robert Metcalfe and David Boggs published the details of Ethernet, a project developed at the Xerox Palo Alto Research Center in 1976 (Fig. 1). Although developed in 1976 the concepts presented in this paper are still the foundation for the contention based LANs of today.

It should be noted that Metcalfe later founded 3COM Corporation, which was instrumental in transitioning Ethernet from the laboratory to the commercial marketplace. Thus he developed the concept in a research environment and then guided its transition to the commercial world.

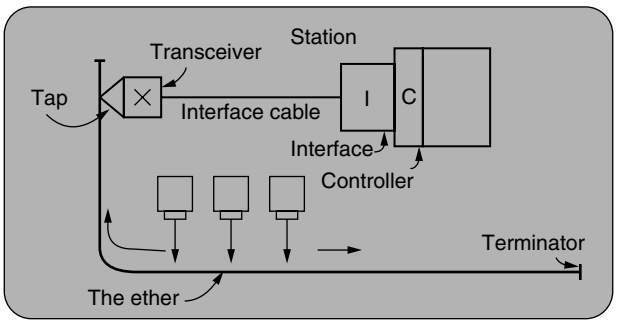


Figure 1. Diagram presented at the 1976 National Computer Conference by Robert M. Metcalfe.

2.2. 802 Activities

In 1980, noting the increasing popularity of Ethernet and the need to standardize existing and future LAN protocols, the IEEE formed the 802 Committee, whose duties were to oversee these standards. Later, the Committee's responsibilities were increased to address MAN standards. As can be seen by the sample of IEEE 802 working group activities listed below, this effort has been quite comprehensive (see Table 1).

As the standards were developed, many transitioned to ANSI and eventually ISO standards. Notable among these standards are 802.1, 802.2, 802.3, 802.4, 802.5, and 802.11. IEEE 802.1 defines the overall architecture of the set of standards; 802.2 defines the logical link control (LLC), which provides a standard set of network services to higher-level protocols; and 802.3, a contention-based system, addresses the copper-based Ethernet standards, which are still employed. 802.3 has expanded from the original 10-Mbps coaxial cable standard to one employing twisted pairs and fiber optics operating at 10, 100, 1000, and 10,000 Mbps.

While the 802.3/Ethernet standard describes a contention-based system, the 802 Committee has developed several contention-free, token passing systems, the most significant of which is the 802.5 — a token ring system popularized by IBM that has practically become extinct. 802.4, a token bus system designed primarily for/by the automotive industry, never achieved popularity and is also extinct.

Table 1. Sample of IEEE 802 Activities

P802.1, *High Level Interface* (HILI)	P802.2, *Logical Link Control*[a]
P802.3, *CSMA/CD*	P802.4, *Token Bus*[a]
P802.5, *Token Ring*[a]	P802.6, *Metropolitan Area Network* (MAN)[a]
P802.7, *Broadband TAG*[a]	P802.8, *Fiber Optic TAG*[b]
P802.9, *Integrated Services LAN* (ISLAN)[a]	P802.10, *Standard for Interoperable LAN Security* (SILS)[a]
P802.11, *Wireless Local Area Network* (WLAN)	P802.12, *Demand Priority*[a]
P802.14, *Cable-TV Based Broadband Communication Network*[a]	P802.15, *Wireless Personal Area Network* (WPAN)
P802.16, *Broadband Wireless Access*	P802.17, *Resilient Packet Ring*

[a]Inactive.
[b]Disbanded.

The token passing systems avoid contention by having the stations organized in a ring (physical for 802.5 and logical for 802.4). A token is passed from station to station. If it has nothing to send, the station receiving the token will directly pass the token to the next station or if it does have packets queued for delivery, it will send one or more of those packets before passing the token to the next station. This deterministic behavior offers capabilities not available in the contention-based 802.3 systems, such as supporting priority schemes and providing worst-case response times. However, these features were not adequate to overcome the overwhelming popularity of the Ethernet systems, and they became extinct as hub-based 802.3 systems eliminated the capability gap between the contention and token-based systems.

As the 802.3 hubs increased in sophistication, they began to use alternative technologies internally. This move changed the role of the IEEE standards from describing the overall activity and behavior of the network to that of describing the *network interface* to the hub. 802.11 covers a set of wireless Ethernet standards, including 802.11b, which is currently commercially popular and discussed in detail later in this article.

3. ETHERNET FUNDAMENTALS

Contention based LAN systems employ a broadcast approach where every station potentially hears every transmission. This means that overlapping transmissions (from different stations) will *collide* and interfere with each other. In order to avoid collisions or quickly recover from those collisions not avoided, the CSMA/CD concept is employed. Originally known as *listen before talk*, and now known as *carrier sense multiple access* (CSMA), the first part of this approach requires each station to listen to the medium and detect the presence of any transmission. If no transmission is detected, then the station may go ahead and transmit. If the medium is busy, the station defers until the medium becomes free. In order to allow transmission detection, the modulation scheme employs a carrier that is quickly distinguishable from a quiescent state. For the original Ethernet and 802.3 standards differential Manchester encoding was employed. This modulation scheme requires a minimum of one line state transition per bit, making it easy to distinguish from a quiescent line.

However, CSMA, by itself, is not sufficient to avoid collisions. Multiple stations could simultaneously sense an empty medium and decide to transmit at roughly the same time, thereby creating a collision. In order to operate efficiently, LANs must detect and quickly recover from collisions since (1) the time involved in a collision is wasteful; and (2) since the messages are obliterated, they become lost frames and thus require some action at higher levels (TCP in the Internet), which noticeably

degrades performance. Originally known as *listen while talk*, *collision detection* (CD) is handled a number of ways. In coaxial cable–based networking, the transmitting station listens to the network while transmitting. If the message observed is not identical to that being transmitted, a collision has occurred. At this point, the detecting station continues to transmit for a short jam-time period (or alternatively sends a *jamming* signal for that same interval) in order to allow the collision to be noticed by all involved parties.

Twisted-pair technologies use hubs that employ separate pairs for transmitting and receiving data. When receiving a transmission, the hubs relay it to all connected stations *except* the originating station. Thus, if a transmitting station hears an incoming transmission, a collision is taking place because the transmission has originated from a different station.

When a collision is detected by whatever means employed, the participating stations transition into a backoff state. Essential to this technique is the concept of a *slot time*. The value of a slot time varies with the implementation standards, but it must satisfy the following criteria [2]:

- *It must define an upper bound on the acquisition time for the medium.*
- *It must define an upper bound on the length of a frame fragment generated by a collision.*
- *It is used for scheduling retransmissions.*

The first two criteria dictate that the slot time will be at least equal to the longest round-trip propagation delay between two stations on the same LAN plus the jam time. This propagation delay involves the signal propagation through the medium plus any electronic delays induced by hubs, repeaters, and level 2 switching. More specifically, it is the longest time period for which a transmitting station must transmit before being assured that a collision will not take place.

The backoff technique employed by IEEE 802.3 and Ethernet is *truncated binary exponential backoff*. Here, when encountering a collision, each station waits until the medium is clear (the collision has ended); it then waits an integer number of slot times chosen from a randomly distributed set of integers in a specified range. If a collision again occurs, the integer range is increased. Eventually, the station either transmits successfully or gives up after a backoff attempt limit and declares a system failure.

The specific integer range for the nth transmission retry is $0 \leq r \leq 2^k$, where $k = \min(n, 10)$.

The set of stations that may enter into a collision or see a collision fragment is known as a *collision domain* while the set of stations that can receive a single broadcast is known as a *broadcast domain*. These are often the same two sets, but the collision domain may be a proper subset of the broadcast domain through the use of switching hubs and other devices that pass broadcast messages but block collisions.

4. IEEE STANDARDS

4.1. IEEE Model

As mentioned earlier, the IEEE 802 Committee has developed both an architecture and an associated set of protocols that are quite thorough. Unfortunately, a complete explanation would take longer than space here allows. Therefore, only an overview is provided here. The IEEE architecture is shown in Fig. 2.

The IEEE model addresses two major ISO model layers: the physical and the data link layers. The ISO physical layer corresponds closely to the IEEE physical layer, while the ISO data link layer contains two sublayers: the IEEE media access control (MAC) and logical link control (LLC) layers.

Multiple link service access points (LSAPs) provide LAN services to higher level layers. Unacknowledged connectionless (type 1), connection-oriented (type 2), and acknowledged connectionless (type 3) are defined in the standard. These LSAPs also hide the MAC/PHY level differences between the various options. IEEE 802.3 (the Ethernet style) employs the type 1 (unacknowledged connectionless) service.

The MAC sublayer supports the LLC sublayer by providing the necessary functions for the LLC to perform. Specifically, it provides for the transmission and reception of frames, which involves framing, addressing and frame check sequence generation and checking.

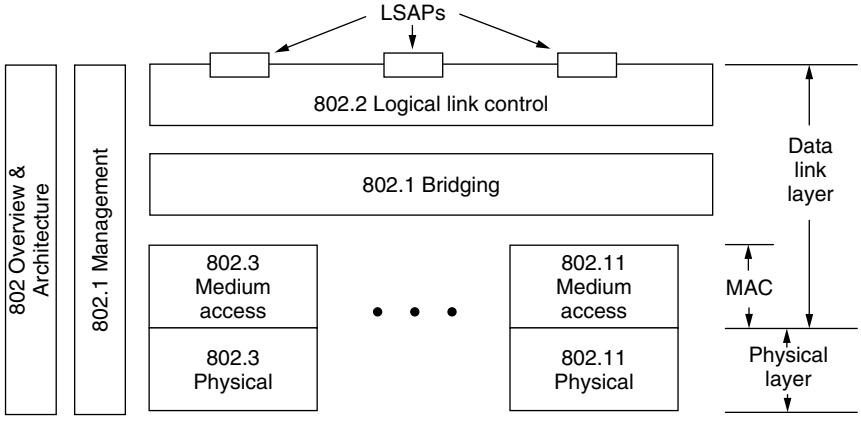


Figure 2. IEEE 802.11 standards.

The physical layer deals with the actual transmission and reception of signals, modulation and timing issues, and so on.

4.2. IEEE Address Management

The developers of Ethernet recognized the value of ensuring that each LAN device, regardless of the PHY layer, has a unique and standard address. In reality, this is only critical for devices on the same LAN since traffic across different LANs is handled by Internet protocols. Originally, Xerox administered LAN addresses, but now IEEE holds this responsibility.

Known as *MAC addresses*, LAN addresses are 48 bits in length (6 octets). The first 2 bits of the address define its nature. The first bit indicates whether the address is a unique (set to 0) or multicast address (set to 1). The second bit, if set to 0, indicates the address is administered by the IEEE. If set to 1, the address is locally administered and not subject to any of the following specifications. Obviously, virtually all LAN addresses are globally administered by the IEEE.

The IEEE addresses are further split into 24-bit portions. The first 24 bits define the *organizationally unique identifier* (OUI), which is administered by the IEEE and allocated uniquely to requesting organizations. Thus it is impossible for LAN addresses specified by one vendor (employing that vendor's OUI) to duplicate addresses from another vendor. Each vendor is responsible for avoiding duplication within its OUI address space. Each distinct OUI allows the vendor to develop approximately 4 million group and unique addresses. If an organization does exhaust its address space, it simply applies for another OUI. Duplication of IEEE addresses has been observed but attributed to manufacturing defects.

4.3. IEEE Media Access Control Frame Format

Originally, the IEEE 802.3 and Ethernet standards defined slightly different MPDU (MAC) frame formats as shown in Fig. 3.

The original differences between the IEEE 802.3 and Ethernet standards were minimal. The most significant was that the Ethernet standard employed a protocol type field to identify the protocol that either requested transmission of the frame or should receive the frame. The 802.3 standard employs the 802.2 layer between the 802.3 and IP layers. The 802.2 SNAP format presents an alternative means for providing the same identification. The 2-octet length/type field is used to distinguish between the approaches. Values less than or equal to 1500 indicate the frame is an IEEE 802.3 frame and the field is a length field containing the length of the remainder of the frame while values above or equal to 1536 represents a protocol id (e.g., 2048 indicates Internet IPv4) and an Ethernet frame. In 1997, revisions to 802.3 merged the Ethernet format into 802.3 as an option.

As shown in Fig. 4 the updated frame format merges the two fields into a single length/type field.

When virtual LANs (VLANs) appeared, the Ethernet frame format was again modified as discussed later. The preamble is a sequence of 56 bits that begins with 1, alternating zeros and ones, and ends with 0. This pattern allows the receiver to synchronize with the incoming data. The start of the important frame contents is indicated by a *start frame delimiter* containing 10101011. Following the start frame delimiter are the six-octet destination and source MAC addresses, and following the MAC addresses is the length/type field previously mentioned. The pad field is used to ensure that the MAC frame size meets the minimum requirements of the 802.3 MPDU. This minimum size (the number of octets beginning with the source address and including everything through the 32 octet FCS) varies with the particular 802.3 implementation — for example, 64 octets for the 10BASE options.

4.4. IEEE PHY Level

The Physical (PHY) Level provides the capability of transmitting and receiving bits between Physical Layer Entities [3] through the defined modulation and encoding schemes specified in the standard.

802.3:

Preamble (7 octets)	Starting delimiter (1 octet)	Destination address (6 octets)	Source address (6 octets)	Length (2 octets)	802.2 Frame (0 – n octets)	Pad	FCS (4 octets)

Ethernet:

Preamble (8 octets)	Destination address (6 octets)	Source address (6 octets)	Type field (2 octets)	Data (46 – 1600 octets)	FCS (4 octets)

Figure 3. Original 802.3/Ethernet MPDU organization.

Preamble (7 octets)	Starting delimiter (1 octet)	Destination address (6 octets)	Source address (6 octets)	Length/ Type field (2 octets)	MAC client data	Pad	FCS (4 octets)

Figure 4. Final 802.3 MPDU frame format.

4.5. IEEE 802.3: Current Copper LAN Implementations

Through the years, the IEEE has developed a large number of LAN standards, originally designated as follows:

<data rate in Mb/s> <medium type> <maximum segment length (× 100 m)>

For example, 10BASE5 would signify a 10-Mbps baseband system with a maximum cable length of 500 meters.

Later options dropped the maximum segment length portion for a medium designation such as "T." Thus, 10BASE-T would signify a 10-Mbps baseband twisted-pair system. The standards began with coaxial cable-based systems and then moved to twisted-pair systems also increasing the data rate. Table 2 lists the most common options.

4.5.1. 10BASE5. 10BASE5, the oldest member of the IEEE 802.3 effort, is now mostly extinct. A 10-Mbps descendant of Metcalfe's Xerox PARC system, 10BASE5 used a thick 50-Ω coaxial cable (polyvinyl chloride with a 0.40-in. diameter or fluoropolymer with a 0.37-in. diameter). 10BASE5, which allowed cable spans of up to 500 m, is the only IEEE 802.3 standard that employed an external medium attachment unit (MAU) known as a *transceiver*. These transceivers connected to the Ethernet coax (coaxial cable) via vampire taps, which attach to the inner conductor through a hole in the outer layers of the coax. An AUI (transceiver) cable connected the transceiver to the station. A failure in the cable would take the entire coax-based system down, thereby making coaxial cable problematic. If the cable were bent, crushed or a terminator removed from either end, the signal quality would be degraded enough, due to reflections, for the system to fail.

4.5.2. 10BASE2. The 10BASE2 standard, also known as *Thin Ethernet* or *Cheapernet*, employs a 50-Ω, RG-58 A/U cable, which is smaller in diameter than the 10base5 cable. With this standard cable sections threaded their way through each of the stations by employing a T-tap at each station that brought the signal into the station. The maximum cable length for this standard was only 185 m as compared to 500 m for the 10BASE5. This did not pose a significant problem, however, since computers had become cheaper and more plentiful and the cable could reach enough computers to make it practical. One weakness was that each cable segment required two BNC connectors which were a common point of failure.

4.5.3. 10BASE-T. Although not the first twisted-pair LAN since 1BASE-T and other non-IEEE systems existed previously, the 10BASE-T quickly became the standard LAN technology employed. This was true in part because the installation and maintenance of twisted-pair cable plants was far easier than that of the coaxial cable systems. Additionally, RJ-45 connectors, which connect up to four twisted pairs, are used to connect the cables to hubs. The use of hubs allowed fault tolerance and, when managed with SNMP or a similar protocol, provided useful management and administrative capabilities.

Table 2. Popular 802.3 Options

10BASE5	Thick coaxial cable	10 Mbps
10BASE2	Thin coaxial cable	10 Mbps
10BASE-T	Two twisted pairs	10 Mbps
100BASE-TX	Two twisted pairs	100 Mbps
1000BASE-T	Four twisted pairs	1 Gbps

The 10BASE-T systems center around a hub (see Fig. 5), known originally in the IEEE standard as a multiport repeater. Unlike the two-way transmission possible with coaxial cable, the twisted-pair systems employ separate pairs for transmitting and receiving. Originally, the system was designed to operate successfully with Category 3 voice-grade unshielded twisted pair (Cat 3 UTP) cable. However, the improved performance of Cat 5 cable coupled with the emergence of 100BASE-T systems has eliminated Cat 3 cable from consideration for today's installations. The hub operates by reflecting any signals received on the uplink from a station to the downlinks for all but the transmitting station. Hubs may be spliced together by using either a special gender switching cable or a special repeater port on the hub. (Otherwise, the downlink from one hub would connect to the downlink of the other hub.)

The two-pair wire plant required a few changes to the CSMA/CD implementation. While transmitting, the 10BASE-T listens to the downlink rather than to the signal on the cable. If a signal appears on the downlink, it must have come from a station other than the listening station thus indicating a collision and thereby negating the need to compare the received message with the transmitted message. When a collision is detected, the detecting station continues to transmit and sends a *jamming* signal for the jam interval in order to inform all involved parties of the collision.

When the 10BASE-T system was introduced, vendors often had a mixed systems employing it and the two coaxial cable technologies. PC NICs (network interface cards) that provided all three interfaces (10BASE5, 10BASE2, and 10BASE-T) became common. Virtually all LAN standards that have appeared after 10BASE-T have employed either twisted pair cable or fiber optic cabling.

4.5.4. 100BASE-T. As technology advanced, network speed became increasingly important. The IEEE 802.3u

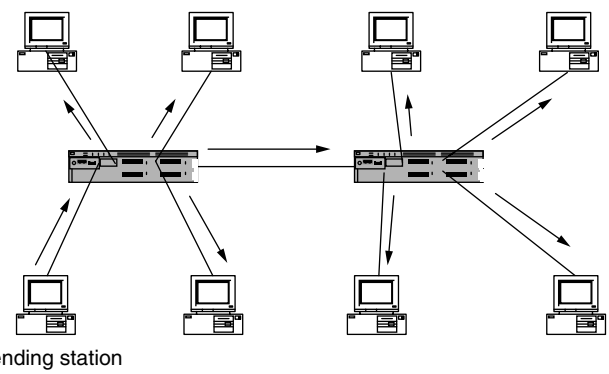


Figure 5. 10BASE-T hub system showing transmission path.

Table 3. 100BASE Options

TX — Two Cat 5 pairs
FX — Two optical fibers
T4 — Four Cat 3 pairs
T2 — Two Cat-4 pairs

Task Group was chartered with developing a 100 Mbps twisted-pair standard; a number of options were developed by the committee, with the 100BASE-TX the dominant option (see Table 3).

100BASE-TX does not differ significantly from 10BASE-T in that both systems can use Cat 5 cable plants. The physical modulation scheme does differ from 10BASE-T and follows that employed by *fiber distributed data interface* (FDDI). (The higher data rate makes differential Manchester encoding less practical as two transitions are often required for each bit.) Many components can automatically accommodate either 10BASE-T or 100BASE-TX through *autonegotiation*. This provides an effective upgrade path from 10 to 100 Mbps using Cat 5 cable plants.

4.5.5. Full-Duplex Ethernet. In 1997, IEEE 802.3x issued a standard for a full-duplex version of Ethernet that allows two stations to communicate over *point-to-point* links. It does not support hubs of other connections. Full-duplex transmission allows a maximum bandwidth of twice the conventional LAN since both transmission and reception can occur simultaneously. Additionally, since only two stations are involved, collisions do not occur, and therefore point-to-point links between the two stations are longer than that allowed in a true contention situation. PAUSE frames are added to the provide flow control necessary to support the higher bandwidth available in a contention free environment. The PAUSE frame specifies the amount of time that a receiving station must refrain from transmitting anything other than MAC control frames. Full-duplex Ethernet also allows *link aggregation* where multiple links (running at the same data rate) may exist between two stations. The bandwidths of the links may be combined to provide high bandwidth capability — for instance, two 100-Mbps full-duplex connections may be combined to provide a 200-Mbps connection.

4.5.6. 1000BASE. The IEEE Gigabit Ethernet standard can be logically divided into two groups. The first developed by the 802.3z taskforce, known as 1000BASE-X, contains three PHY options: 1000BASE-SX (short wavelength fiber) 1000BASE-LX (long wavelength fiber) and 1000BASE-CX (short-run copper). The second group, developed by the 802.3ab taskforce, 1000BASE-T, is designed as an extension to 10BASE-T and 100BASE-T.

The 1000BASE-X family uses the physical layer standards based on those employed by Fibre Channel technology. On the other hand, the 1000BASE-T standard uses four pairs of Cat 5 cable. (Note that 100BASE-T and 10BASE-T only use two pairs.) Each of the four pairs is modulated at the same clock rate (125 MHz) as 100BASE-T but employs a coding scheme that contains 2-bits/symbol and thus achieves 250 Mbps per twisted pair. Thus, four pairs transfer 1000 Mbps. As with the 100BASE-T standard, the maximum cable segment length is specified as 100 meters.

In order to meet the slot-time requirements of gigabit transmission, an extension field is added on to the end (after the FCS) of the Ethernet PDU. This is only employed in half-duplex operation as collisions do not occur in full-duplex mode. The extension bits are *non-data* symbols, which distinguish them from data bits.

For data rates above 100 Mbps, transmitting stations may employ a *burst mode* where a series of frames may be transmitted without relinquishing control of the transmission medium. *Burst mode* permits higher efficiency than would be possible with the conventional Ethernet protocol. The first frame may, if necessary, employ extension bits, but subsequent frames in the burst need not. Each frame is sent with the proper interframe gap, but the gap is filled with non-data symbols, which prevent the medium from appearing idle.

5. VIRTUAL LANs

Virtual LANS (VLANs) allow a set of stations to share the same broadcast domain while not necessarily in the same physical domain. A set of stations connected to a set of switches can be partitioned into several VLANs.

VLANs can be implemented by employing port-based VLAN hubs that partition stations based on the connection port. Connection ports are configured into VLANs through some form of station management, and then all stations plugged into the designed ports are in the same broadcast domain. Another approach is to partition based on MAC addresses. Both this and the previous approach do not require any special configuration of the stations.

VLANs implemented by using IEEE 802.1Q employ a tagged frame that contains a 4-octet tag inserted just after the source MAC address. The tag begins with 10000001 and contains a priority as well as a 12-bit VLAN identifier (the VID). This tagged frame is defined in the IEEE 302.3, 2000 edition standard (see Fig. 6).

6. WIRELESS LANs: 802.11

Wireless LANs, standardized by the IEEE 802.11 Working Group [4,5], extend the Ethernet concept to an RF connection to the desktop. As shown earlier in Fig. 2, the 802.11 medium-access and physical standards are

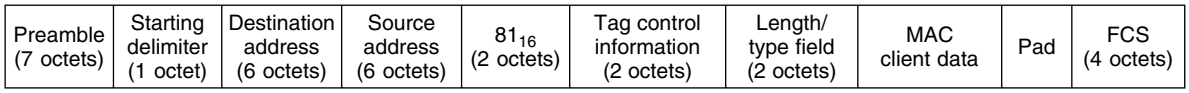


Figure 6. VLAN tagged frame format.

alternatives to the 802.3 or other 802 medium access and physical layer standards within the overall 802 architecture.

Designed to support mobile computing and improve flexibility of the "to the desktop" connection, wireless Ethernet has proved to be a very difficult technological chore. Issues such as multipath distortion, licensing, security, bandwidth, and possibly health hazards hindered the development of wireless systems. In 1997, the IEEE 802.11 Working Group published standards for both frequency-hopping spread spectrum (FHSS) and direct sequence spread spectrum (DSSS) 2.4 GHz and infrared technologies supporting data rates of 1 and 2 Mbps. Although some products appeared, they did not become commonplace until the 802.11b Task Group developed a 2.4 GHz direct sequence spread spectrum 11 Mbps system. The IEEE 802.11a Task Group also developed a 5 GHz orthogonal frequency division multiplexing system with data rates up to 54 Mbps. This technology is currently under development with the fabrication of integrated circuits that operate at this higher data rate.

In order to provide compatibility with existing wired LANs, the 802.11 Working Group designed 802.11 to provide the same interface as 802.3. Thus, it uses the 802.2 LLC sublayer and appears identical to 802.3 from any layer above 802.2.

Other wireless technologies that can be used for LANs, such as BlueTooth and HomeRF, are briefly addressed later in this section and covered in detail in separate entries in the encyclopedia.

6.1. 802.11 PHY

The 802.11b systems operate in the 2.4 GHz ISM (industrial, scientific, and medical) band, which does not require licensing. This band is split into 14 overlapping 22 MHz channels, but not all channels are available in all countries; for example, in the United States, only channels 1–11 are available. The power employed also varies by area, but most devices available in the United States limit the output power to 100 mW even though 1000 mW is the formal limitation. This puts the power output to less than that employed by the typical digital cellular phone (125 mW). Most systems also employ a speed backoff dropping from 11 Mbps to 5.5, 2, and then, finally, 1 Mbps.

The actual distances achieved through 802.11b are very sensitive to antenna placement, multipath distortion, walls, floors, and other barriers. Typical distances between components vary between 50 and 200 ft.

6.2. Network Topology

Due to the increased configuration flexibility and complexities associated with RF transmission, the overall architecture and protocols forming the 802.11 standard are more complex than their 802.3 counterparts. There are six major components of wireless LANs: the wireless stations themselves (STAs in IEEE terminology); access points (APs); the wireless medium; basic service sets (BSS); the distribution system (DS); and the extended service set (ESS) (see Fig. 7).

A set of stations (STAs), commonly called wireless NICs (network interface card), that talk to each other is called a *basic service set* (BSS). If all the STAs are wireless and there are no other components, the BSS is called an ad hoc or *independent BSS* (IBSS).

Access points (APs) serve both as relays between STAs and as bridges between STAs and wired LANs. When a BSS contains a single AP, it is called an *infrastructure BSS*, but the term "infrastructure" is typically omitted. In this mode, wireless stations communicate with each other with the AP acting as a relay. This slows STA-to-STA performance since two messages are needed for each STA-to-STA frame rather than the single STA-to-STA message employed in the ad hoc mode (see Fig. 8).

When a system contains more than one AP and associated BSS, it is called an *extended service set* (ESS). The BSSs are connected through an abstraction known as a *distribution system* (DS), which is typically a conventional copper-based LAN. The additional access points provide for increased geographic coverage in a manner similar to multiple cells for a cellular phone system.

6.3. 802.11 MAC Operation

The 802.11 MAC provides peer-to-peer best-effort, connectionless communication between LLC entities. Three types of MAC frames exist: data, control, and management (with a fourth type reserved for future use). MAC service data units (MSDUs) convey the peer-to-peer data supplied by high-level protocols, while MAC management PDUs

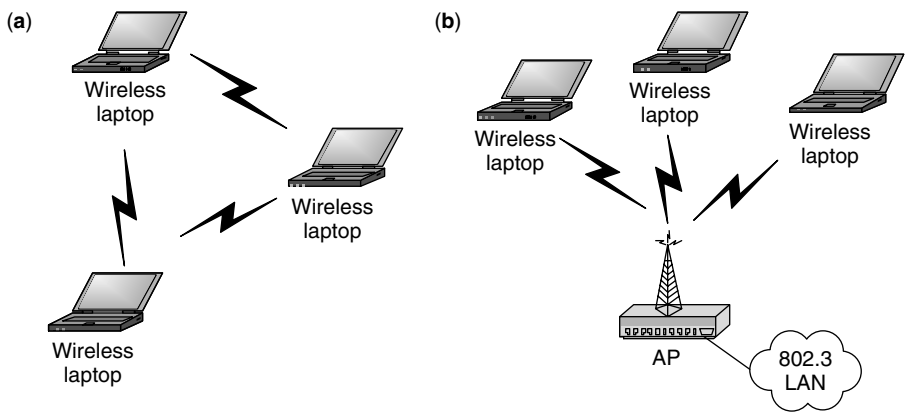


Figure 7. 802.11 architectures: (**a**) IBSS; (**b**) Infrastructure BSS.

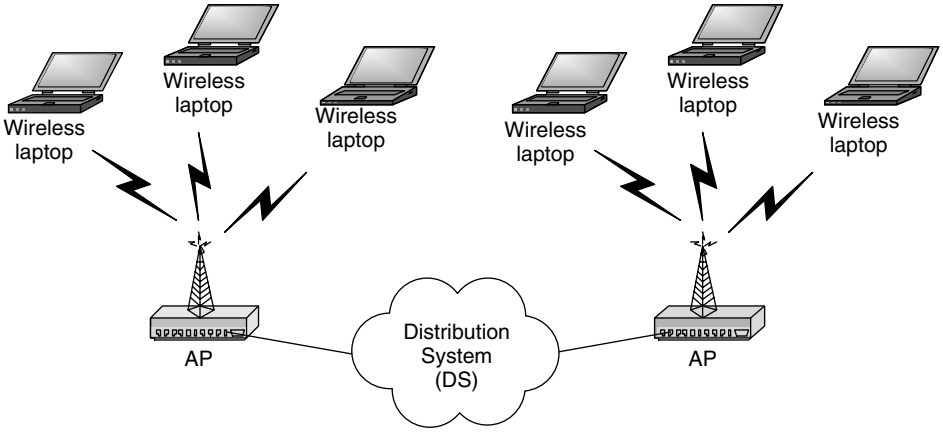


Figure 8. 802.11 Extended service set.

(MMPDUs) convey management messages supporting the communication. Additionally, control frames are employed to aid management and data transfer, and as with 802.3, multicast and broadcast services are available.

6.3.1. 802.11 MAC Data Frame Format. As shown in Fig. 9, the MAC data frame (MPDU) format differs from the 802.3 MAC data frame. MSDUs contain data supplied by a higher-level protocol and are one of the MPDU payload types. MMPDUs contain management information and are another MPDU payload.

MSDUs are prepended with a MAC header containing four 6-octet addresses, a frame control field, a duration/ID field, and a sequence control field. A 4-octet FCS (CCITT CRC-32) is appended to the MSDU.

The control field contains the 802.11 protocol version and a number of subfields to support fragmentation, power management, retries, WEP (*wired equivalent privacy* — discussed later), utilization, and other operational parameters.

Four 6-octet IEEE addresses are contained in the MPDU, although only some message types will use all four addresses; others will use between 1 and 3 addresses. In addition to the source and destination addresses found in 802.3, transmitter and receiver addresses as well as a BSSID address may be contained in the other two address fields. The transmitter address (TA) identifies the wireless source of the transmitted message (not always the originating source), while the receiver address (RA) specifies the wireless receiver (not always the ultimate destination.)

The duration/ID field contains information to update system NAVs or an association identificer (AID) used to obtained frames buffered in APs during STA power saving activities.

6.3.2. CSMA/CA. While conventional (wired) LANs employ CSMA/CD, wireless LANs cannot implement the collision detection (listen while talk), so instead they employ CSMA/CA (a collision avoidance scheme). Each STA employs the listen before talk aspect of the CSMA from 802.3. However, when the STA detects a busy medium, it defers for a time period determined by a binary exponential backoff algorithm. The range of values generated from this algorithm doubles each time the station consecutively defers due to a busy channel.

6.3.2.1. Hidden-Node Problem. A problem specific to wireless LANs is the *hidden node problem* (Fig. 10). Because of distance limitations of the RF system, two STAs may possibly communicate effectively with an AP

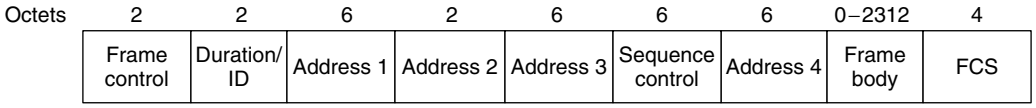


Figure 9. MPDU (MAC frame) format.

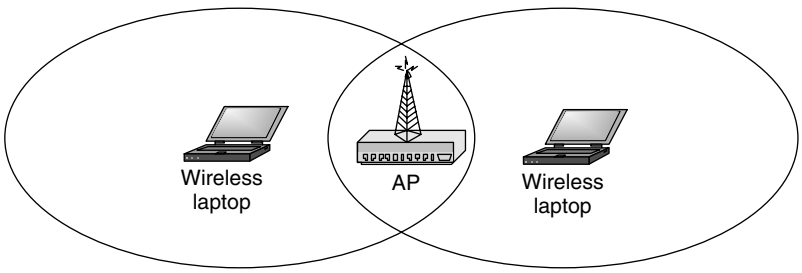


Figure 10. Hidden node problem.

but cannot hear each other. Thus, the CSMA algorithm cannot prevent collisions at the AP in all cases.

To work around this problem, STAs may send an *optional* RTS (request to send) message to the AP. Whether this RTS message is sent or not depends on the value of the dot11RTSThreshold attribute. If the frame is longer than this value, then a RTS message contains an appropriate reservation time period, which the access point echoes back to all stations in a CTS (clear to send) message. Then, the requesting station is clear, for the reserved time period, to send its frame without the possibility of interference from a hidden STA. A zero setting for dot11RTSThreshold requires all frames to be sent with RTS/CTS exchange, while a value of dot11RTSThreshold greater than the largest MSDU deactivates the RTS/CTS facility.

6.3.2.2. The Network Allocation Vector (NAV). Additional help for the collision avoidance system is provided through the NAV (network allocation vector), which indicates to a station the time period before the network medium becomes free again. NAVs are updated from data contained in each transmitted frame.

6.3.3. Roaming. When a STA enters the area of one or more APs, it chooses an AP (joining a BSS) based on signal strength and error rates. Joining is called *association*. When a STA shuts down, it *disassociates* with the system. *Reassociation* occurs when a station requests to switch associations between APs. Similar to the association service, this request also includes the AP previously involved in an association. The protocols to support reassociation coordination between APs are not standardized at this time, so APs from different vendors may not perform handoffs successfully.

6.3.4. Fragmentation. In order to increase system reliability, IEEE 802.11 allows transmitters to fragment unicast MSDUs and MMPDUs into smaller MPDUs. Defragmentation occurs when the frame arrives at the immediate receiving station All the fragments, each of which is acknowledged separately, are the same size except for the last, which may be smaller.

The value in the Sequence Number field is the same for all fragments of a single MSDU or MMPDU. The sequence order within the set of fragments is designated by the value in the Sequence Control field.

6.3.5. Power Management. To support power conservation in battery-powered equipment, STAs may employ a power management scheme. For an infrastructure BSS, a STA will be in one of two modes of operation. When a station is *awake*, it is fully powered and operational. When a STA is in the *doze* state, it is not able to transmit or receive. STAs notify their associated APs when they are changing state. If a STA enters the *doze* state, the AP must buffer any transmissions to it until the STA changes to the *awake* mode.

For an independent BSS, the operation is more complicated since there are no APs to buffer frames for dozing STAs. A dozing STA must periodically awake to allow any STA with a queued message for it to send the message. After draining the queued messages, the STA may again doze for a time period.

6.4. Security

The 802.11 standard identifies two security subsystems: the *wired equivalent privacy* (WEP) system and the *authentication services*. The goal of WEP is to provide security equivalent to that of a wire-based system. Thus, rather than being a true security system, WEP simply alleviates weaknesses introduced by the wireless nature of the network.

6.4.1. Wired Equivalent Privacy (WEP). WEP is an encryption *option* in the 802.11 standard required for WiFi™ certification (see Fig. 11).

WEP relies on the RC4 stream cipher [6] and a secret shared key. The secured payload consists of the data field and a 32-bit CRC (cyclic redundancy check), the integrity check value (ICV), of the data field, all of which are RC4 encrypted. The encryption seed consists of 40-bit shared secret key prepended with a 24-bit initialization vector (IV). The purpose of the IV is to randomize part of the key so that repeated encryptions of the same data are not

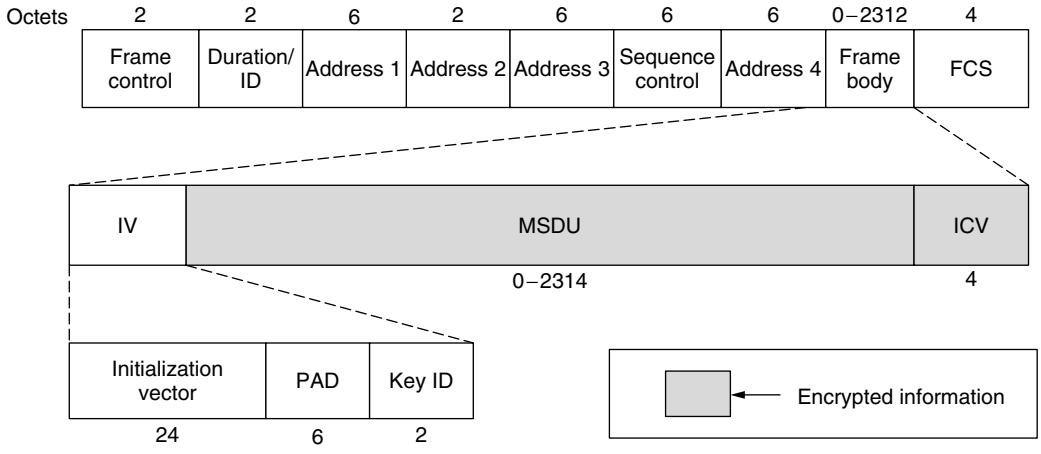


Figure 11. WAP frame body encryption.

identical, as this would provide a useful clue in breaking the encryption. The IV is transmitted in plaintext with the MAC frame. The rate at which the IV is changed is not specified in the standard but left to the discretion of the vendor.

6.4.2. Authentication Services. Two types of authentication services are specified within the 802.11 standard: *open system* and *shared key*. *Open-system* authentication automatically authenticates any station if the recipient station is operating in open system mode. The requesting station issues a message that contains its identity and a request for authentication. The recipient station responds with *successful* in the frame status code.

The *shared key* system differentiates stations that know a shared secret key from those that do not. The secret key is transmitted to the stations using an out-of-band mechanism and is stored as a write-only MIB (management information base) attribute. (Write-only attributes cannot be accessed *via* the MAC management system.)

Four frames are used in the authentication process. The first is the request for authentication by the station. The responding station sends the second frame containing a randomly generated 128-octet plaintext field to the requesting station. The requesting station encrypts the plaintext field and returns it to the responding station in the third frame. The responding station decrypts the message and compares the received plaintext to the original plaintext and, if they are identical, assumes that the requesting station knows the secret key and, therefore, should be authenticated. The responding station then sends the requesting station a successful authentication frame. If the decrypted text is not identical to the original plaintext message, then an unsuccessful authentication is returned.

6.4.3. Weaknesses of 802.11 Security. Almost concurrent with the development of the 802.11 standard were concerns for the strength of its security provisions. Researchers at the University of California at Berkeley [7] and others [8] have identified serious weaknesses in the WEP protocol. The concerns are centered on the way the RC4 cipher is employed rather than any fundamental weaknesses of the encryption technology itself. Four attacks have been identified.

The first attack, a *passive attack to decrypt traffic*, is based on the high potential volume of traffic coupled with the relatively short (24-bit) initialization vector. A system sending 1500-byte packets at 11 Mbps that is 10% utilized would be guaranteed to reuse the key in not more than 50 h. If the packets were smaller in size and the workload on the wireless system above 10%, a much shorter reuse period would occur. Further, these calculations assume that vendors change the IV for every transmission and use a random pattern with an externally generated seed. If multiple stations employ the same algorithm and start from the same seed, then reuse will occur much more quickly since all systems reinitialized will start from the same seed. Once multiple messages using the same IV have been recovered, statistical methods may be employed to identify the original contents of a message; all messages with that IV will be "open" for examination.

The second attack, an *active attack to inject traffic*, requires the plaintext of a message to be determined by the passive attack technique outlined above. Then using the relationship that RC4(X) xor X xor Y = RC4(Y), new messages can be encrypted successfully.

The third attack is an *active attack from both ends*. Here an attacker simply adjusts the bits of the header of a message changing the IP address to a host available to the attacker. This is possible since flipping a bit in the ciphertext flips the corresponding bit in the decrypted plaintext. It is also possible to calculate which bits of the CRC-32 must be flipped to compensate for the changes in the IP. All that is necessary is to capture a packet where the IP address is known and replay it on the system with the appropriate alternations. A plaintext IP packet will be sent to the controlled location on the Internet and examined to reveal the plaintext of the message. Then, any messages *using that same IV* can be decrypted.

The final attack is a *table-based attack* where a table of IVs and corresponding key streams are stored. Such a table would be large but eventually would allow a station to decrypt every packet sent through the system.

While these techniques require sophisticated monitoring and are not as simple as exploiting a hole in an operating system or application, they do imply that additional security work is required by the 802.11, Task Group i. WEP2 will employ 128-bit encryption and 128-bit IVs thus increasing the computation cost to break the system. The IEEE Task Group i has approved a draft to establish an authentication and key management system, tentatively called ESN *enhanced security network* (ESN), which will employ the draft Federal Information Processing Standard AES (advanced encryption standard).

6.5. WECA

The Wireless Ethernet Compatibility Alliance (WECA) [9] was formed in 1999 for the purpose of guaranteeing interoperability. Addressing 802.11, 2.4 GHz, DSSS, high data rate standards, this nonprofit organization's mission is to "certify interoperability of Wi-Fi (IEEE 802.11) products and to promote Wi-Fi as the global wireless LAN standard across all market segments." WECA has developed a certification test suite. Those products passing the compatibility tests are labeled Wi-Fi products.

6.6. Other Wireless Technologies

6.6.1. IEEE 802.11a. Standard 802.11a defines a PHY layer standard operating in the unlicensed 5 GHz band, known as U-NII (unlicensed national information infrastructure), and provides for data rates of 6, 9, 12, 18, 24, 36, 48, or 54 Mbps; 6, 12, and 24 Mbps are mandatory. It shares the same medium access controller (MAC) protocol as 802.11b. It achieves the higher data rate by using a higher carrier frequency along with orthogonal frequency-division multiplexing (OFDM) modulation.

OFDM survives multipath and intersymbol interference at these higher data rates by simultaneously transmitting multiple subcarriers on orthogonal frequency

channels where each subcarrier modulated at a low symbol rate.

A concern with 802.11a is that the lower power limit may restrict distances to less than 50 ft. Chipsets and products employing 802.11a are now (at the time of writing) beginning to appear but are considerably more expensive than those for 802.11b.

6.6.2. IEEE 802.11g. The IEEE 802.11 Task Group g is working toward a higher-speed version of 802.11b but has not yet approved a standard. To some, 802.11g appears to be in direct competition with 802.11a. For this and other reasons, 802.11g is politically charged, and its future is unclear at this time.

6.6.3. BlueTooth. BlueTooth [10] also operates in the 2.4 GHz ISM band using FHSS technology. Intended for personal area networking using low-cost interfaces for systems such as PDAs, mobile phones, and personal computers, it is limited to a 10-m distance. Bluetooth devices use the IEEE standard 48-bit addressing scheme. First generation devices operate up to 1 Mbps, while second-generation devices are expected to provide up to a 2 Mbps data rate.

More details on BlueTooth may be found elsewhere in the encyclopedia.

6.6.4. HomeRF. HomeRF [11] is a wireless network standard, specified in the shared wireless access protocol (SWAP) for home use where the distance requirements are limited. HomeRF 1.0 provides up to a 1.6 Mbps data rate at distances up to 150 ft. It uses FHSS transmission in the same 2.4 GHz band employed by 802.11b.

HomeRF incorporates the DECT (digital enhanced cordless telephony) standard to support mixed data and voice. The HomeRF 2.0 specification will support data rates in the 10 Mbps range. HomeRF also claims increased security over 802.11b. For example, it uses a 128-bit encryption key and a 32-bit IV that increases the IV reuse period significantly. Further, the way in which IVs are selected is specified in the protocol. Finally, the FHSS technology employed provides protection against denial of service attacks.

More details on HomeRF may be found elsewhere in the encyclopedia

7. THE FUTURE

Finally, the quest for increased speed continues. The 10 Gig (gigabit) Ethernet standard is in the final stages and expected to be completed and approved early in 2002, and 40 Gig Ethernet is being explored. The enormous popularity of Ethernet is not limited to *local* networks. Work is underway to employ variations of the Ethernet standard for metropolitan area networks currently employing PoS (packet over SONET) technology. Ethernet and its variations may eventually become the most popular standard for medium and long haul transmission. For example, the Metro Ethernet Forum [12] was created in June 2001 "to accelerate the adoption of optical Ethernet technology in metro networks around the globe."[1] The 10 Gig Ethernet contains a WAN PHY (physical) that describes a 40 km SMF (single-mode fiber) behavior suitable for competing with SONET. Additionally, the IEEE has created an 802.3ah *Ethernet in the First Mile* Task Force to explore the use of Ethernet fiber to the home/curb.

Although starting as a local network standard, Ethernet is rapidly expanding into the metropolitan and long haul environments and appears destined to become the prevalling standards for all classes of networking.

BIOGRAPHY

John H. Carson is a professor of management science at the George Washington University in Washington, D.C. Dr. Carson earned a B.S. in electrical engineering in 1969 and an M.S. and Ph.D. in information science in 1970 and 1976, respectively, from Lehigh University, Bethlehem, Pennsylvania. Dr. Carson served as manager of system engineering for the Software Productivity Consortium and as a principle scientist in the MITRE Corporation's networking center. He joined the George Washington University's Management Science Department in 1980, where he directed the M.S. program in information systems technology for 19 years. His areas of interest are software design, networking and communications technology, and information system design. Dr. Carson has authored and coauthored four books on computing and communications technology; he is a member of Eta Kapp Nu, Sigma Xi, and Beta Gamma Sigma.

BIBLIOGRAPHY

1. R. M. Metcalfe and D. R. Boggs, Ethernet: Distributed packet switching for local computer networks, *Commun. ACM* **19**(7): 395–404 (1976).
2. IEEE Std 802.3, 2000 ed., Part 3: *Carrier Sense Multiple Access with Collision Detection* (*CSMA/CD*) *Access Method and Physical Layer Specifications*, IEEE, 2000.
3. IEEE Standards for Local and Metropolitan Area Networks, *Overview and Architecture*, 1990.
4. IEEE 802.11 standards are available at *http://standards.ieee.org/getieee802/*.
5. B. O'Hara and A. Petrick, *802.11 Handbook, a Designer's Companion, IEEE Standards Information Network*, IEEE Press, 1999.
6. See papers on RC4 encryption from RSA Security at *http://www.rsa.com/*.
7. N. Borisov, I. Goldberg, and D. Wagner, Security of the WEP algorithm, *http://www.isaac.cs.berkeley.edu/isaac/wep-faq.html*, Univ. California at Berkeley, Feb. 2001.
8. P. C. Mehta, Wired equivalent privacy vulnerability, *http://www.sans.org/infosecFAQ/wireless/equiv.htm*, SANS Institute, April 4, 2001.
9. *http://www.weca.net*.
10. *http://www.bluetooth.com*.
11. *http://www.homerf.org*.
12. *www.metroethernetforum.org/*.

LOOP ANTENNAS

KAZIMIERZ (KAI) SIWIAK
Time Domain Corporation
Huntsville, Alabama

1. INTRODUCTION

The *IEEE Standard Definitions of Terms for Antennas* [1] defines the loop as "an antenna whose configuration is that of a loop," further noting that "if the current in the loop, or in the multiple parallel turns of the loop, is essentially uniform and the loop circumference is small compared with the wavelength, the radiation pattern approximates that of a magnetic dipole." That definition and the further note imply the two basic realms of loop antennas: electrically small, and electrically large structures.

There are hundreds of millions of loop antennas currently in use [2] by subscribers of personal communications devices, primarily pagers. Furthermore, loops have appeared as transmitting arrays, like the massive multielement loop array at shortwave station HCJB in Quito, Ecuador, and as fractional wavelength-size tunable HF transmitting antennas. The loop is indeed an important and pervasive communications antenna.

The following analysis of loop antennas reveals that the loop, when small compared with a wavelength, exhibits a radiation resistance proportional to the square of the enclosed area. Extremely low values of radiation resistance are encountered for such loops, and extreme care must be taken to effect efficient antenna designs. Furthermore, when the small loop is implemented as a transmitting resonant circuit, surprisingly high voltages can exist across the resonating capacitor even for modest applied transmitter power levels. The wave impedance in the immediate vicinity of the loop is low, but at close distances (0.1–2 wavelengths) exceeds the intrinsic free space impedance before approaching that value.

A loop analysis is summarized that applies to loops of arbitrary circular diameter and of arbitrary wire thickness. The analysis leads to some detail regarding the current density in the cross section of the wire. Loops of shapes other than circular are less easily analyzed, and are best handled by numerical methods such as moment method described by Burke and Poggio [3].

Loops are the antennas of choice in pager receivers, and appear as both ferrite loaded loops and as single-turn rectangular shaped structures within the radio housing. Body worn loops benefit from a field enhancement because of the resonant behavior of human body with respect to vertically polarized waves. In the high frequency bands, the loop is used as a series resonant circuit fed by a secondary loop. The structure can be tuned over a very large frequency band while maintaining a relatively constant feed point impedance. Large loop arrays comprised of one wavelength perimeter square loops have been successfully implemented as high-gain transmitting structures at high power shortwave stations.

2. ANALYSIS OF LOOP ANTENNAS

Loop antennas, particularly circular loops, were among the first radiating structures analyzed beginning as early as 1897 with Pocklington's analysis [4] of the thin wire loop excited by a plane wave. Later, Hallén [5] and Storer [6] studied driven loops. All these authors used a Fourier expansion of the loop current, and the latter two authors discovered numerical difficulties with the approach. The difficulties could be avoided, as pointed out by Wu [7], by integrating the Green function over the toroidal surface of the surface of the wire. The present author coauthored an improved theory [8,9] that specifically takes into account the finite dimension of the loop wire and extends the validity of the solution to fatter wires than previously considered. Additionally, the work revealed some detail of the loop current around the loop cross section. Arbitrarily shaped loops, such as triangular loops and square loops, as well as loop arrays can be conveniently analyzed using numerical methods.

2.1. The Infinitesimal Loop Antenna

The infinitesimal current loop consists of a circulating current I enclosing an infinitesimal surface area S, and is solved by analogy to the infinitesimal dipole. The fields of an elementary loop element of radius b can be written in terms of the loop enclosed area $S = \pi b^2$ and a constant excitation current, I (when I is RMS, then the fields are also RMS quantities). The fields are "near" in the sense that the distance parameter r is far smaller than the wavelength, but far larger than the loop dimension $2b$. Hence, this is *not* the *close* near field region. The term kIS is often called the *loop moment* and is analogous to the similar term Ih associated with the *dipole moment*. The infinitesimally small loop is pictured in Fig. 1a next to its elementary dipole analog (Fig. 1b). The dipole uniform current I flowing over an elemental length h is the dual of a "magnetic current" $M_z S = Ih$ and the surface area is $S = h/k$. The fields due to the infinitesimal loop are then found from the vector and scalar potentials.

2.1.1. Vector and Scalar Potentials. The wave equation, in the form of the inhomogeneous Helmholtz equation, is used here with most of the underlying vector arithmetic omitted; see Refs. 10–12 for more details. For a magnetic current element source, the electric displacement **D** is always solenoidal (the field lines do not originate or

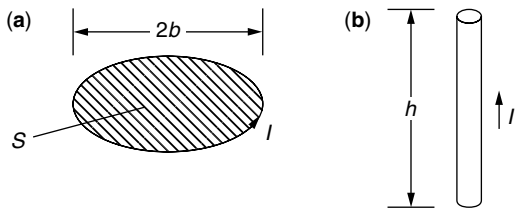


Figure 1. Small-antenna geometry showing (**a**) the parameters of the infinitesimal loop moment, and (**b**) its elementary dipole dual. (*Source*: Siwiak [2].)

terminate on sources); that is, in the absence of source charges the divergence is zero

$$\nabla \cdot \mathbf{D} = 0 \tag{1}$$

and the electric displacement field can be represented by the curl of an arbitrary vector $\mathbf{F}$

$$\mathbf{D} = \varepsilon_0 \mathbf{E} = \nabla \times \mathbf{F} \tag{2}$$

where $\mathbf{F}$ is the vector potential and obeys the vector identity $\nabla \cdot \nabla \times \mathbf{F} = 0$. Using Ampere's law in the absence of electric sources, we obtain

$$\nabla \times \mathbf{H} = j\omega\varepsilon_0 \mathbf{E} \tag{3}$$

and with the vector identity $\nabla \times (-\nabla\Phi) = 0$, where Φ represents an arbitrary scalar function of position, it follows that

$$\mathbf{H} = -\nabla\Phi - j\omega\mathbf{F} \tag{4}$$

and for a homogeneous medium, after some manipulation, we get

$$\nabla^2\mathbf{F} + k^2\mathbf{F} = -\varepsilon_0\mathbf{M} + \nabla(\nabla \cdot \mathbf{F} + j\omega\mu_0\varepsilon_0\Phi) \tag{5}$$

where k is the wavenumber and $k^2 = \omega^2\mu_0\varepsilon_0$. Although equation (2) defines the curl of $\mathbf{F}$, the divergence of $\mathbf{F}$ can be independently defined and the *Lorentz condition* is chosen

$$j\omega\mu_0\varepsilon_0\Phi = -\nabla \cdot \mathbf{F} \tag{6}$$

where ∇^2 is the Laplacian operator given by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{7}$$

Substituting the simplification of Eq. (6) into (5) leads to the inhomogeneous Helmholtz equation

$$\nabla^2\mathbf{F} + k^2\mathbf{F} = -\varepsilon_0\mathbf{M} \tag{8}$$

Similarly, by using Eqs. (6) and (4) it is seen that

$$\nabla^2\Phi + k^2\Phi = 0 \tag{9}$$

Using Eq. (4) and the Lorentz condition of Eq. (6), we can find the electric field solely in terms of the vector potential $\mathbf{F}$. The utility of that definition becomes apparent when we consider a magnetic current source aligned along a single vector direction, for example, $\mathbf{M} = \mathbf{z}M_z$, for which the vector potential is $\mathbf{F} = \mathbf{z}F_z$, where $\mathbf{z}$ is the unit vector aligned with the z axis, and Eq. (8) becomes a scalar equation.

2.1.2. Radiation from a Magnetic Current Element. The solution to the wave equation (8) presented here, with the details suppressed, is a spherical wave. The results are used to derive the radiation properties of the infinitesimal current loop as the dual of the infinitesimal current element. The infinitesimal magnetic current element $\mathbf{M} = \mathbf{z}M_z$ located at the origin satisfies a one-dimensional, and hence scalar form of Eq. (8). At points excluding the origin where the infinitesimal current element is located, Eq. (8) is source-free and is written as a function of radial distance r

$$\nabla^2 F_z(r) + k^2 F_z(r) = \frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2\frac{\partial F_z(r)}{\partial r}\right] + k^2 F_z(r) = 0 \tag{10}$$

which can be reduced to

$$\frac{d^2F_z(r)}{dr^2} + \frac{2}{r}\frac{dF_z(r)}{dr} + k^2F_z(r) = 0 \tag{11}$$

Since F_z is a function of only the radial coordinate, the partial derivative in Eq. (10) was replaced with the ordinary derivative. Eq. (11) has a solution

$$F_z = C_1\frac{e^{-jkr}}{r} \tag{12}$$

There is a second solution where the exponent of the phasor quantity is positive; however, we are interested here in outward traveling waves, so we discard that solution. In the static case the phasor quantity is unity. The constant C_1 is related to the strength of the source current, and is found by integrating Eq. (8) over the volume, including the source giving

$$C_1 = \frac{\varepsilon_0}{4\pi}kIS \tag{13}$$

and the solution for the vector potential is in the $\mathbf{z}$ unit vector direction

$$\mathbf{F} = \frac{\varepsilon_0}{4\pi}kIS\frac{e^{-jkr}}{r}\mathbf{z} \tag{14}$$

which is an outward propagating spherical wave with increasing phase delay (increasingly negative phase) and with amplitude decreasing as the inverse of distance. We may now solve for the magnetic fields of an infinitesimal current element by inserting Eq. (14) into (4) with Eq. (6) and then for the electric field by using Eq. (2). The fields, after sufficient manipulation, and for $r \gg kS$, are

$$H_r = \frac{kIS}{2\pi}e^{-jkr}k^2\left[\frac{j}{(kr)^2} + \frac{1}{(kr)^3}\right]\cos(\theta) \tag{15}$$

$$H_\theta = \frac{kIS}{4\pi}e^{-jkr}k^2\left[-\frac{1}{kr} + \frac{j}{(kr)^2} + \frac{1}{(kr)^3}\right]\sin(\theta) \tag{16}$$

$$E_\phi = \eta_0\frac{kIS}{4\pi}e^{-jkr}k^2\left[\frac{1}{kr} - \frac{j}{(kr)^2}\right]\sin(\theta) \tag{17}$$

where $\eta_0 = c\mu_0 = 376.730313$ is the intrinsic free-space impedance, c is the velocity of propagation (see Ref. 13 for definitions of constants), and I is the loop current.

Equations (15) and (16) for the *magnetic* fields H_r and H_θ (1.30) of the infinitesimal loop have exactly the same form as the *electric* fields E_r and E_θ for the infinitesimal dipole, while Eq. (17) for the *electric* field of the loop E_ϕ has exactly the same form as the *magnetic* field H_ϕ of the dipole when the term kIS of the loop expressions is replaced with Ih for the infinitesimal ideal (uniform current element) dipole. In the case where the loop moment

kIS is superimposed on, and equals the dipole moment Ih, the fields in all space will be circularly polarized.

Equations (15)–(17) describe a particularly complex field behavior for what is a very idealized selection of sources: a simple linear magnetic current M representing a current loop I encompassing an infinitesimal surface $S = \pi b^2$. Expressions (15)–(17) are valid only in the region sufficiently far ($r \gg kS$) from the region of the magnetic current source M.

2.1.3. The Wave Impedance of Loop Radiation. The wave impedance can be defined as the ratio of the total electric field magnitude divided by the total magnetic field magnitude. We can study the wave impedance of the loop fields by using Eqs. (15)–(17) for the infinitesimal loop fields, along with their dual quantities for the ideal electric dipole. Figure 2 shows the loop field wave impedance as a function of distance kr from the loop along the direction of maximum far field radiation. The wave impedance for the elementary dipole is shown for comparison. At distances near $kr = 1$ the wave impedance of loop radiation exceeds $\eta_0 = 376.73\ \Omega$, the intrinsic free-space impedance, while that of the infinitesimal loop is below 376.73 Ω. In this region, the electric fields of the loop dominate.

2.1.4. The Radiation Regions of Loops. Inspection of Eqs. (15)–(17) for the loop reveal a very complex field structure. There are components of the fields that vary as the inverse third power of distance r, inverse square of r, and the inverse of r. In the near field or induction region of the idealized infinitesimal loop, that is, for $kr \ll 1$ (however, $r \gg kS$ for the loop and $r \gg h$ for the dipole), the magnetic fields vary as the inverse third power of distance.

The region where kr is nearly unity is part of the radiating near field of the Fresnel zone. The inner boundary of that zone is taken by Jordan [12] to be

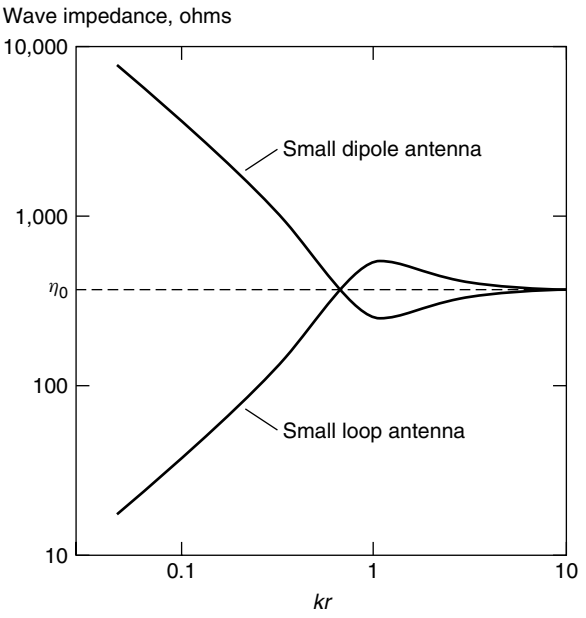


Figure 2. Small loop antenna and dipole antenna wave impedances compared. (*Source*: Siwiak [2].)

$r^2 > 0.38D^3/\lambda$, and the outer boundary is $r < 2D^2/\lambda$, where D is the largest dimension of the antenna, here equal to $2b$. The outer boundary criterion is based on a maximum phase error of $\pi/8$. There is a significant radial component of the field in the Fresnel zone.

The far-field or Fraunhofer zone is a region of the field for which the angular radiation pattern is essentially independent of distance. That region is usually defined as extending from $r < 2D^2/\lambda$ to infinity, and the field amplitudes there are essentially proportional to the inverse of distance from the source. Far-zone behavior is identified with the basic free space propagation law.

2.1.5. The Induction Zone of Loops. We can study the "induction zone" in comparison to the "far field" by considering "induction zone" coupling that was investigated by Hazeltine [14] and that was applied to low frequency radio receiver designs of his time. Today the problem might be applied to the design of a miniature radio module where inductors must be oriented for minimum coupling. The problem Hazeltine solved was one of finding the geometric orientation for which two loops in parallel planes have minimum coupling in the induction zone of their near fields and serves to illustrate that the "near field" behavior differs fundamentally and significantly from "far field" behavior. To study the problem we invoke the principle of reciprocity, which states

$$\int_V \left[\mathbf{E}_b \cdot \mathbf{J}_a - \mathbf{H}_b \cdot \mathbf{M}_a\right] dV \equiv \int_V \left[\mathbf{E}_a \cdot \mathbf{J}_b - \mathbf{H}_a \cdot \mathbf{M}_b\right] dV \tag{18}$$

That is, the reaction on antenna (a) of sources (b) equals the reaction on antenna (b) of sources (a). For two loops with loop moments parallel to the z axis, we want to find the angle θ for which the coupling between the loops vanishes; that is, both sides of equation (18) are zero. The reference geometry is shown in Fig. 3. In the case of the loop, there are no electric sources in Eq. (18), so $\mathbf{J}_a = \mathbf{J}_b = 0$, and both $\mathbf{M}_a$ and $\mathbf{M}_b$ are aligned with $\mathbf{z}$, the unit vector parallel to the z axis. Retaining only the inductive field components and clearing common constants in Eqs. (15) and (17) are placed into (18). We require that $(H_r\mathbf{r} + H_\theta\theta)\mathbf{z} = 0$. Since $\mathbf{r} \cdot \mathbf{z} = -\sin(\theta)$ and $\mathbf{r} \cdot \mathbf{z} = \cos(\theta)$, we are left with $2\cos^2(\theta) - \sin^2(\theta) = 0$, for which $\theta = 54.736°$. When oriented as shown in Fig. 3, two loops parallel to the x–y plane whose centers are displaced by an angle of 54.736° with respect to the z axis will not couple in their near fields. To be sure, the angle determined above is "exactly" correct for infinitesimally small loops; however, that angle will be nominally the same for larger loops. Hazeltine [14] used this principle, placing the axes of the inductors in a common plane each at an angle of 54.7° with respect to the normal form the radio chassis, to minimize the coupling between the inductors.

The same principle can be exploited in the design of a metal detector, as depicted in Fig. 4. The loop a is driven with an audiofrequency oscillator. Loop b, in a parallel plane and displaced so that nominally $\theta = 54.7°$, is connected to a detector that might contain an audio amplifier that feeds a set of headphones. Any conductive object near loop a will disrupt the balance of the system

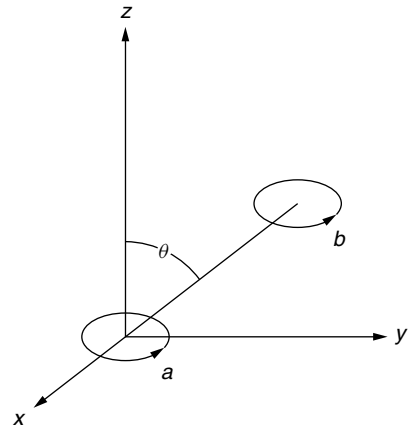


Figure 3. Two small loops in parallel planes and with $\theta = 54.736^\circ$ will not couple in their near fields. (*Source*: Siwiak [2].)

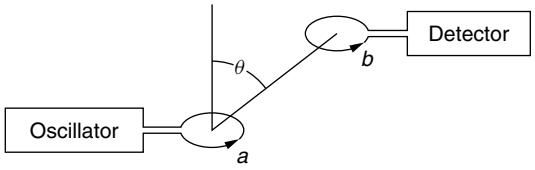


Figure 4. A metal detector employs two loops initially oriented to minimize coupling in their near fields.

and result in an increased coupling between the two loops, thus indicating the presence of a conducting object near a.

2.1.6. The Intermediate- and Far-Field Zones of Loops. The loop coupling problem provides us with a way to investigate the intermediate and far-field coupling by applying Eq. (18) with Eqs. (15) and (16) for various loop separations kr. In the far-field region only the H_θ term of the magnetic field survives, and by inspection of Eq. (16), the minimum coupling occurs for $\theta = 0$ or 180°. Figure 5 compares the coupling (normalized to their peak values) for loops in parallel planes whose fields are given by Eq. (15)–(17). Figure 5 shows the coupling as a function of angle θ for an intermediate region ($kr = 2$) and for the far-field case ($kr = 1000$) in comparison with the induction-zone case ($kr = 0.001$). The patterns are fundamentally and significantly different. The coupling null at $\theta = 54.7^\circ$ is clearly evident for the induction-zone case $kr = 0.001$ and for which the $(1/kr)^3$ terms dominate. Equally evident is the far-field coupling null for parallel loops on a common axis when the $1/kr$ terms dominate. The intermediate zone coupling shows a transitional behavior where all the terms in kr are comparable.

2.1.7. The Directivity and Impedance of Small Loops. The *directive gain* of the small loop can be found from the far-field radially directed Poynting vector in ratio to the average Poynting vector over the radian sphere:

$$D(\theta,\phi) = \frac{|(\mathbf{E}\times\mathbf{H})\cdot\mathbf{r}|}{\dfrac{1}{4\pi}\displaystyle\int_0^{2\pi}\int_0^{\pi}|(\mathbf{E}\times\mathbf{H})\cdot\mathbf{r}|\sin(\theta)\,d\theta\,d\phi} \tag{19}$$

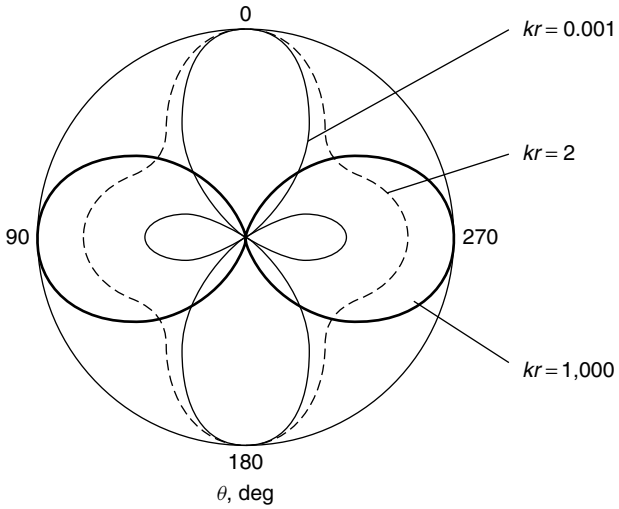


Figure 5. Normalized-induction-zone, intermediate-zone, and far-zone coupling between loops in parallel planes. (*Source*: Siwiak [2].)

Only the θ component of H and the ϕ component of E survive into the far field. Using Eq. (16) for H_θ and Eq. (17) for E_ϕ and retaining only the $1/kr$ terms, we see that Eq. (19) yields $D = 1.5\sin^2(\theta)$ by noting that the functional form of the product of E and H is simply $\sin^2(\theta)$ and by carrying out the simple integration in the denominator of Eq. (19).

Taking into account the directive gain, the far-field power density P_d in the peak of the pattern is

$$P_\mathrm{density} = \frac{1.5\,I^2R_\mathrm{radiation}}{4\pi r^2} = H_\theta^2\eta_0 = \left[\frac{kS}{4\pi}\frac{k}{r}I\right]^2\eta_0 \tag{20}$$

For radiated power $I^2R_\mathrm{radiation}$, hence, we can solve for the radiation resistance

$$R_\mathrm{radiation} = \frac{(kS)^2}{6\pi}\eta_0 = \eta_0\frac{\pi}{6}(kb)^4 \tag{21}$$

for the infinitesimal loop of loop radius b.

When fed by a gap, there is a dipole moment that adds terms not only to the impedance of the loop but also to the close near fields. For the geometry shown in Fig. 6, and using the analysis of King [15], the electrically small loop, having a diameter $2b$ and wire diameter $2a$, exhibits a feed point impedance given by

$$Z_\mathrm{loop} = \eta_0\frac{\pi}{6}(kb)^4\left[1+8(kb)^2\right]\left[1-\frac{a^2}{b^2}\right]+\cdots$$
$$+\,j\eta_0 kb\left[\ln\left[\frac{8b}{a}\right]-2+\frac{2}{3}(kb)^2\right][1+2(kb)^2] \tag{22}$$

including dipole-mode terms valid for $kb \ll 0.1$. The leading term of Eq. (22) is the same as derived in Eq. (21) for the infinitesimal loop. Expression (22) adds the detail of terms considering the dipole moment of the gap-fed loop as well as refinements for loop wire radius a. The small-loop antenna is characterized by a radiation resistance that is proportional to the Fourth power of the loop radius b.

The reactance is inductive; hence, it is proportional to the antenna radius. It follows that the Q is inversely proportional to the third power of the loop radius, a result that is consistent with the fundamental limit behavior for small antennas.

Using Eq. (22), and ignoring the dipole-mode terms and second order terms in a/b, the unloaded Q of the loop antenna, is

$$Q_{\text{loop}} = \frac{\frac{6}{\pi}\left[\ln\left[\frac{8b}{a}\right] - 2\right]}{(kb)^3} \tag{23}$$

which for $b/a = 6$ becomes

$$Q_{\text{loop}} = \frac{3.6}{(kb)^3} \tag{24}$$

which has the proper limiting behavior for small-loop radius. The Q of the small loop given by Eq. (23) is indeed larger than the minimum possible $Q_{\min} = (kb)^{-3}$ predicted in Siwiak [2] for a structure of its size. It must be emphasized that the actual Q of such an antenna will be smaller than given by Eq. (24) because of unavoidable dissipative losses not represented in Eq. (22)–(24). We can approach the minimum Q but never be smaller, except by introducing dissipative losses.

2.2. The Gap-Fed Loop

The analysis of arbitrarily fat wire loops follows the method in Ref. 8, shown in simplified form in Ref. 9 and summarized here. The toroid geometry of the loop is expressed in cylindrical coordinates ρ, ϕ, and z with the toroid located symmetrically in the $z = 0$ plane. The relevant geometry is shown in Fig. 6.

2.2.1. Loop Surface Current Density. The current density on the surface of the toroidal surface of the loop is given by

$$J_\phi = \sum_{n=-\infty}^{\infty} \sum_{p=-\infty}^{\infty} A_{n,p} e^{jn\phi} F_p \tag{25}$$

where the functions F_p are symmetric about the z axis and are simple functions of $\cos(n\psi)$, where ψ is in the cross section of the wire as shown in Fig. 6 and is related to the cylindrical coordinate by $z = a\sin(\psi)$. These function

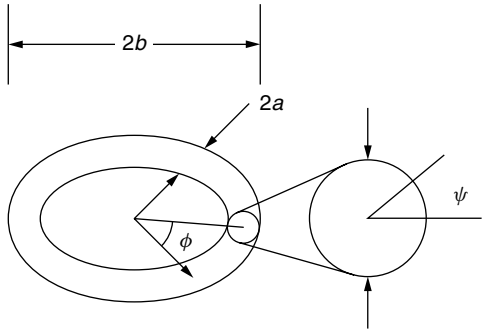


Figure 6. Parameters of the fat wire loop. (*Source*: Siwiak [2].)

are orthonormalized over the conductor surface using the Gram–Schmidt method described in (16), yielding

$$F_0 = \frac{1}{2\pi\sqrt{ab}} \tag{26}$$

and

$$F_1 = F_0\sqrt{\frac{2}{1-(a/2b)^2}}\left[\cos(\psi) - \frac{a}{2b}\right] \tag{27}$$

The higher-order functions are lengthy but simple functions of $\sin(p\psi)$ and $\cos(p\psi)$.

2.2.2. Scalar and Vector Potentials. The electric field is obtained form the vector and scalar potentials

$$\mathbf{E} = -\nabla\Phi - j\omega\mathbf{A} \tag{28}$$

The boundary conditions require that E_ϕ, E_ψ, and E_ρ are zero on the surface of the loop everywhere except at the feed gap $|\phi| \le \varepsilon$. Because this analysis will be limited to wire diameters significantly smaller than a wavelength, the boundary conditions on E_ψ and E_ρ will not be enforced. In the gap $E_\phi = V_0/2\varepsilon\rho$, where V_0 is the gap excitation voltage.

The components of the vector potential are simply

$$A_\phi = \frac{1}{4\pi}\int_S\int J_\phi \cos(\phi - \phi')\, dS \tag{29}$$

and

$$A_\rho = \frac{1}{4\pi}\int_S\int J_\phi \sin(\phi - \phi')\, dS \tag{30}$$

and the vector potential is

$$\Phi = \frac{j\eta_0}{4\pi k}\int_S\int \frac{1}{\rho}\frac{\partial J_\phi}{\partial\phi} G\, dS \tag{31}$$

where the value of $dS = [b + a\sin(\psi)]a\, d\psi$. *Green's function* G is expressed in terms of cylindrical waves to match the rotational symmetry of the loop

$$G = \frac{1}{2j}\sum_{m=-\infty}^{\infty} e^{-jm(\phi-\phi')}\int_{-\infty}^{\infty} J_m(\rho_1 - \nu)H_m^{(2)}(\rho_2 - \nu)e^{-j\zeta(z-z')}\, d\zeta \tag{32}$$

where $\nu = \sqrt{k^2 + \zeta^2}$
$\rho_1 = \rho - a\cos(\psi)$
$\rho_2 = \rho + a\cos(\psi)$

and where $J_m(\nu\rho)$ and $H_m^{(2)}(\nu\rho)$ are the Bessel and Hankel functions, respectively.

2.2.3. Matching the Boundary Conditions. Expression (8) is now inserted into Eqs. (5)–(8), and the electric field is then found from Eq. (2) and the boundary condition is enforced. For constant ρ on the wire

$$\int_{-\pi}^{\pi} E_\phi e^{jm\phi}\, d\phi = -\frac{V_0}{\rho}\frac{\sin(m\varepsilon)}{m\varepsilon} \tag{33}$$

This condition is enforced on the wire as many times as there are harmonics in ψ. Truncating the index p as

described in Ref. 9 to a small finite number P, we force $E_\phi = 0$ except in the feeding gap along the lines of constant ρ on the surface of the toroid. If we truncate to P, the number of harmonics F_p in ψ and to M the number of harmonics in ϕ, we find the radiation current by solving M systems of $P \times P$ algebraic equations in $A_{m,p}$. In Ref. 9, $P = 2$ and M in the several hundreds was found to be a reasonable computational task that led to useful solutions.

2.2.4. Loop Fields and Impedance. With the harmonic amplitudes $A_{m,p}$ known, the current density is found from Eq. (1). The electric field is found next from Eq. (2) and the magnetic field is given by

$$H_\rho = -\frac{\partial A_\phi}{\partial z} \tag{34}$$

$$H_\phi = -\frac{\partial A_\rho}{\partial z} \tag{35}$$

$$H_z = \frac{\partial A_\phi}{\partial \rho} + \frac{A_\phi}{\rho} - \frac{1}{\rho}\frac{\partial A_\rho}{\partial \rho} \tag{36}$$

The loop current across a section of the wire is found by integrating the function J_ϕ in Eq. (25) around the wire cross section. The loop radiation impedance is then the applied voltage V_0 in the gap divided by the current in the gap. Figure 7 shows the loop feed radiation resistance, and Fig. 8 shows the corresponding loop reactance, as a function of loop radius kr for a thin wire, $\Omega = 15$, and a fat wire, $\Omega = 10$, loop where $\Omega = 2\ln(2\pi b/a)$. The thin-wire loop has very sharp resonant behavior compared with the fat-wire loop, especially for a half-wavelength-diameter ($kb = 0.5$) structure. The higher resonances are less pronounced for both loops. Fat-wire loops exhibit an interesting behavior in that at a diameter of about a half-wavelength, the reactance is essentially always capacitive and the total impedance remains well behaved.

2.2.5. Small Gap-Fed Loops. The detailed analysis of the fat, gap-fed wire loop, as shown in Refs. 8 and 9, reveals

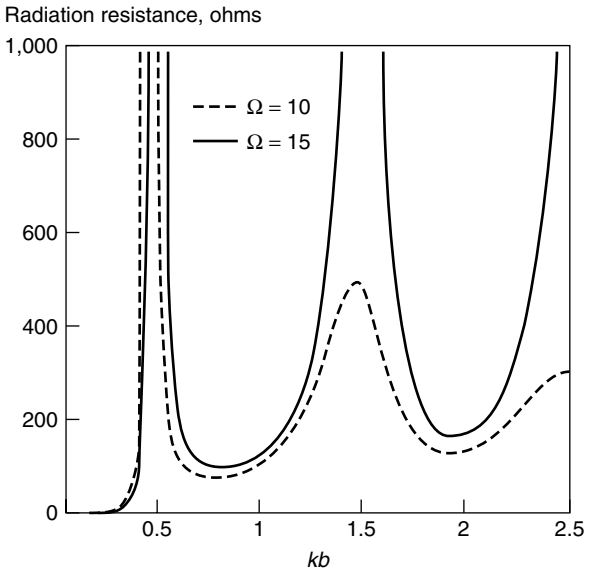


Figure 7. Loop radiation resistance.

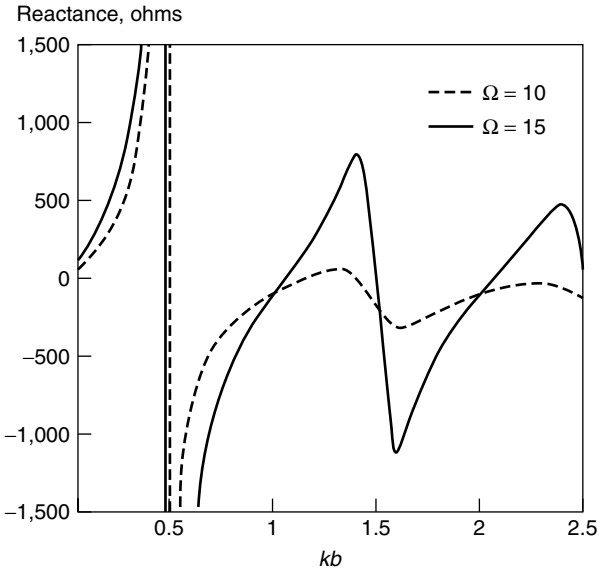


Figure 8. Loop reactance.

that the current density around the circumference of the wire, angle ψ in Fig. 6, is not constant. An approximation to the current density along the wire circumference for a small-diameter loop is

$$J_\phi = \frac{I_\phi}{2\pi a}[1 - 2\cos(\phi)(kb)^2][1 + Y\cos(\psi)] \tag{37}$$

where I_ϕ is the loop current, which has cosine variation along the *loop circumference*; and where the variation around the *wire circumference* is shown as a function of the angle ψ. Y is the ratio of the first to the zero-order mode in ϕ, and is not a simple function of loop dimensions a and b, but can be found numerically in Siwiak [2] and from the analysis of the previous section. For the small loop Y is negative and of order a/b, so Eq. (37) predicts that there is current bunching along the inner contour ($\psi = 180^\circ$) of the wire loop. Table 1 gives representative values for Y as a function of a/b.

This increased current density results in a corresponding increase in dissipative losses in the small loop. We can infer that the cross-sectional shape of the conductor formed into a loop antenna will impact the loss performance in a small loop.

The small loop fed with a voltage gap has a charge accumulation at the gap and will exhibit a close near electric field. For a small loop of radius b and centered in

Table 1. Parameter Y for Various Loop Thickness and $b = 0.01$ Wavelengths

Ω	a/λ	Y
19.899	0.000003	−0.0039
17.491	0.00001	−0.0090
15.294	0.00003	−0.020
12.886	0.0001	−0.048
10.689	0.0003	−0.098
8.2809	0.001	−0.179

the x–y plane, the fields at $(x, y) = (0, 0)$ are derived in Ref. 9 and given here as

$$E_\phi = -j\frac{\eta_0 kI}{2} \tag{38}$$

where I is the loop current and

$$H_z = \frac{I}{2b} \tag{39}$$

Expression (39) is recognized as the classic expression for the static magnetic field within a single-turn solenoid. Note that the electric field given by Eq. (38) does not depend on any loop dimensions, but was derived for an electrically small loop. The wave impedance, Z_w, at the origin, is the ratio of E_ϕ to H_z and from Eqs. (38) and (39) is

$$Z_w = -j\eta_0 kb \tag{40}$$

In addition to providing insight into the behavior of loop probes, Eqs. (38)–(40) are useful in testing the results of numerical codes like the numerical electromagnetic code (NEC) described in Ref. 3, and often used in the numerical analysis of wire antenna structures.

When the small loop is used as an untuned and unshielded field probe, the current induced in the loop will have a component due to the magnetic field normal to the loop plane as well as a component due to the electric field in the plane of the loop. A measure of E field to H field sensitivity is apparent from expression (40). The electric field to magnetic field sensitivity of a simple small-loop probe is proportional to the loop diameter. The small gap-fed loop, then, has a dipole moment that complicates its use as a purely magnetic field probe.

3. LOOP APPLICATIONS

Loop antennas appear in pager receivers as both ferrite-loaded loops and as single-turn rectangular structure within the radio housing. When worn on the belt, the loop benefits from coupling to the vertically resonant human body. In the high-frequency bands, the loop has been implemented as a series resonant circuit fed by a secondary loop. The structure can be tuned over a very large frequency band while maintaining a relatively constant feed point impedance. One wavelength perimeter square loops have been successfully implemented as high-gain transmitting structures.

3.1. The Ferrite-Loaded Loop Antenna: A Magnetic Dipole

Let us examine a small ferrite-loaded loop antenna with dimensions, $2h = 2.4$ cm, $2a = 0.4$ cm, and at a wavelength of about $\lambda = 8.6$ m as depicted in Fig. 9. When the permeability of the ferrite is sufficiently high, this antenna behaves like a magnetic dipole. The magnetic fields are strongly confined to the magnetic medium, especially near the midsection of the ferrite rod, and behave as the dual of the electric dipole excited by a triangular current distribution. We can therefore analyze its behavior using a small dipole analysis shown in Siwiak [2].

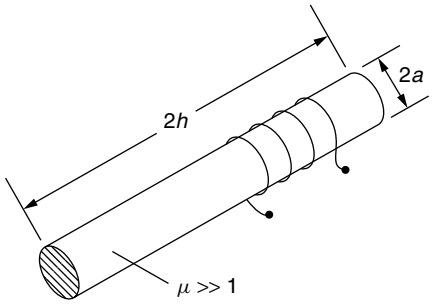


Figure 9. A ferrite loaded loop antenna. (*Source*: Siwiak [2].)

The impedance at the midpoint of a short dipole having a current uniformly decreasing from the feed point across its length $2h$ is

$$Z_{\text{dipole}} = \frac{\eta_0}{6\pi}(kh)^2 - j\frac{\dfrac{\eta_0}{2\pi}\left[\ln\left[\dfrac{2h}{a}\right] - 1\right]}{kh} \tag{41}$$

The corresponding unloaded Q of the dipole antenna is

$$Q_{\text{dipole}} = \frac{3\left[\ln\left[\dfrac{2h}{a}\right] - 1\right]}{(kh)^3} \tag{42}$$

Equation (42) has the expected inverse third power with size behavior for small antennas, and for $h/a = 6$

$$Q_{\text{dipole}} = \frac{4.5}{(kh)^3} \tag{43}$$

Comparing the Q for a small dipole given by Eq. (43) with the Q of a small loop of Eq. (24), we see that the loop Q is small even though the same ratio of antenna dimension to wire radius was used. We conclude that the small loop utilizes the smallest sphere that encloses it more efficiently than does the small dipole. Indeed, the thin dipole is essentially a one-dimensional structure, while the small loop is essentially a two-dimensional structure.

We can use Eqs. (41) and (42) for the elementary dipole to examine the ferrite load loop antenna since it resembles a magnetic dipole. The minimum ideal Q of this antenna is given by Eq. (42), 1.0×10^6. The corresponding bandwidth of such an antenna having no dissipative losses would be $2 \times 35f/Q = 70$ MHz$/1.3 \times 10^6 = 69$ Hz. A practical ferrite antenna at this frequency has an actual unloaded Q_A of nearer to 100, as can be inferred from the performance of belt-mounted radios shown in Table 2. Hence, an estimate of the actual antenna efficiency is

$$10\log\frac{Q_A}{Q} = -40\,\text{dB} \tag{44}$$

and the actual resultant 3 dB bandwidth is about 700 kHz. Such an antenna is typical of the type that would be used in a body-mounted paging receiver application. As detailed in Siwiak [2], the body exhibits an average magnetic field

Table 2. Paging Receiver Performance Using Loops

Frequency Band (MHz)	Paging Receiver, at Belt Average Gain (dBi)	Field Strength Sensitivity (dB·μV/m)
30–50	−32 to −37	12–17
85	−26	13
160	−19 to −23	10–14
280–300	−16	10
460	−12	12
800–960	−9	18–28

Source: After Siwiak [2].

enhancement of about 6 dB at this frequency, so the average belt-mounted antenna gain is −34 dBi. This is typical of a front-position body-mounted paging or personal communication receiver performance in this frequency range.

3.2. Body Enhancement in Body-Worn Loops

Loops are often implemented as internal antennas in pager receiver applications spanning the frequency bands from 30 to 960 MHz. Pagers are often worn at belt level, and benefit from the "body enhancement" effect. The standing adult human body resembles a lossy wire antenna that resonates in the range of 40–80 MHz. The frequency response, as seen in Fig. 10, is broad, and for belt-mounted loop antennas polarized in the body axis direction, enhances the loop antenna azimuth-averaged gain at frequencies below about 500 MHz.

The far-field radiation pattern of a body-worn receiver is nearly omnidirectional at very low frequency. As frequency is increased, the pattern behind the body develops a shadow that is manifest as a deepening null with increasing frequency. In the high-frequency limit, there is only a forward lobe with the back half-space essentially completely blocked by the body. For horizontal incident polarization, there is no longitudinal body resonance and there is only slight enhancement above 100 MHz.

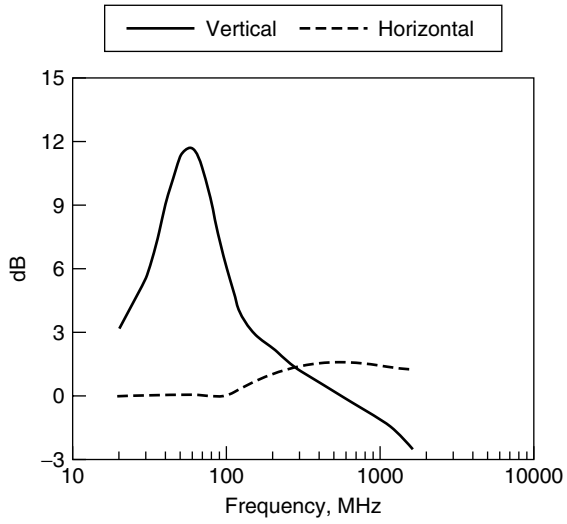


Figure 10. Gain-averaged body-enhanced loop response. (*Source*: Siwiak [2].)

3.3. The Small Resonated High-Frequency Loop

The simple loop may be resonated with a series capacitor having a magnitude of reactance equal to the loop reactance, and indeed, is effectively implemented that way for use in the HF bands as discovered by Dunlavy [17]. When fed by a second untuned loop, this antenna will exhibit a nearly constant feed point impedance over a three or four to one bandwidth by simply adjusting the capacitor to the desired resonant frequency. The reactive part of the loop impedance is inductive, where the inductance is given by $\{Z_L\} = \omega L$, so ignoring the higher-order terms

$$L = \frac{\eta_0 kb\left[\ln\left[\frac{8b}{a}\right] - 2\right]}{\omega} \tag{45}$$

which with the substitution $\eta_0 k/\omega = \mu_0$ becomes

$$L = \mu_0 b\left[\ln\left[\frac{8b}{a}\right] - 2\right] \tag{46}$$

The capacitance required to resonate this small loop at frequency f is

$$C = \frac{1}{(2\pi f)^2 L} \tag{47}$$

The loop may be coupled to a radio circuit in many different ways, including methods given in Refs. 17 and 18. When used in transmitter applications, the small-loop antenna is capable of impressing a substantial voltage across the resonating capacitor. For a power P delivered to a small loop with unloaded Q of Eq. (23) and with resonating the reactance X_C given by the reactive part of Eq. (22), it is easy to show that the peak voltage across the resonating capacitor is

$$V_p = \sqrt{X_C Q P} \tag{48}$$

by recognizing that

$$V_p = \sqrt{2} I_{\text{RMS}} X_C \tag{49}$$

where I_{RMS} is the total RMS loop current

$$I_{\text{RMS}} = \sqrt{\frac{P}{\text{Re}\{Z_{\text{loop}}\}}} \tag{50}$$

along with Q at the resonant frequency in Eq. (23).

Transmitter power levels as low as one watt delivered to a moderately efficient small-diameter ($\lambda/100$) loop can result in peak values of several hundred volts across the resonating capacitor. This is not intuitively expected; the small loop is often viewed as a high current circuit that is often described as a short-circuited ring. However, because it is usually implemented as a *resonant circuit* with a resonating capacitor, it can also be an extremely high-voltage circuit as will be shown below. Care must be exercised in selecting the voltage rating of the resonating capacitor even for modest transmitting power levels, just as care must be taken to keep resistive losses low in the loop structure.

As an example, consider the Q and bandwidth of a small-loop antenna: $2b = 10$ cm in diameter, resonated by a series capacitor and operating at 30 MHz. The example loop is constructed of $2a = 1$ cm diameter copper tubing with conductivity $\sigma = 5.7 \times 10^7$ S/m. The resistance per unit length of round wire of diameter $2a$ with conductivity σ is

$$R_s = \frac{1}{2\pi a \delta_s \sigma} = \frac{1}{2\pi a}\sqrt{\frac{\omega\mu_0}{2\sigma}} \tag{51}$$

where δ_s is the skin depth for good conductors and ω is the radian frequency and $\mu_0 = 4\pi \times 10^{-7}$ H/m is the permeability of free space, so $R_s = 0.046\ \Omega$. From Eq. (22) the loop impedance is $Z = 0.00792 + j71.41$. Hence the loop efficiency can be found by comparing the loop radiation resistance with loss resistance. The loop efficiency is $R_s/(R_s + \mathrm{Re}\{Z\}) = 0.147$ or 14.7%. From Eqs. (46) and (47) we find the resonating capacitance $C = 74.3\ \mu$F. From Eqs. (48)–(50) we see that if one watt is supplied to the loop, the peak voltage across the resonating capacitor is 308 V, and that the loop current 4.3 A. The *resonated* loop is by no means the "low impedance" structure that we normally imagine it to be.

3.4. The Rectangular Loop

Pager and other miniature receiver antennas used in the 30–940 MHz frequency range are most often implemented as electrically small rectangular loops. For a rectangle dimensioned $b_1 \times b_2$ of comparable length, and constructed with $2a$-diameter round wire, the loop impedance is given in Ref. 19 as

$$Z_{\mathrm{rect}} = \frac{\eta_0}{6\pi}(k^2A)^2 + j\frac{\eta_0}{\pi}\left[b_1 \ln\left[\frac{2A}{a(b_1+b_c)}\right] + \left[b_2 \ln\left[\frac{2A}{a(b_2+b_c)}\right] + 2(a + b_c - b_1 - b_2)\right]\right] \tag{52}$$

where $A = b_1 b_2$ and $b_c = (b_1^2 + b_2^2)^{1/2}$. The loss resistance is found by multiplying R_s in Eq. (51) by perimeter length of the loop, $2(b_1 + b_2)$. For a given antenna size the lowest loss occurs for the circular loop.

3.5. The Quad Loop Antenna

The quad loop antenna, sometimes called the *cubical quad*, was developed by Clarence C. Moore in the 1940s as a replacement for a four element parasitic dipole array (Yagi–Uda array). The dipole array exhibited corona arcing at the element tips severe enough to damage the antenna when operated at high power levels (10 kW) in a high-altitude (10,000-ft) shortwave broadcasting application in the 25-m band. Moore sought an antenna design with "no tips" that would support extremely high electric field strengths, which caused the destructive arcing. His solution was a one-wavelength perimeter square loop, and later with a loop director element as shown in Fig. 11. The configuration exhibited no arcing tendencies, and a new short wave antenna configuration was born.

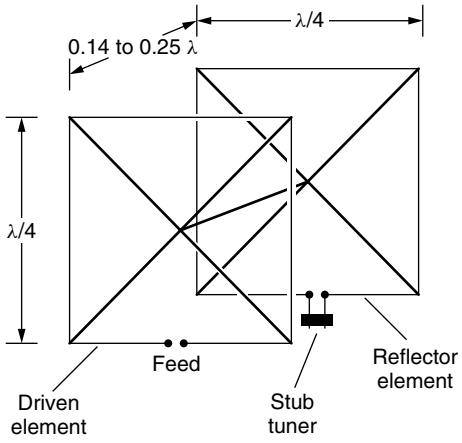


Figure 11. Two-element loop array.

As shown in Fig. 11, the driven element is approximately one quarter-wavelength on an edge. Actually, resonance occurs when the antenna perimeter is about 3% greater than a wavelength. The reflector element perimeter is approximately 6% larger than a wavelength, and may be implemented with a stub tuning arrangement. Typical element spacing is between 0.14 λ and 0.25 λ. The directivity of a quad loop is approximately 2 dB greater than that of a Yagi antenna with the same element spacing.

BIOGRAPHY

Kazimierz "Kai" Siwiak received his B.S.E.E. and M.S.E.E. degrees from the Polytechnic Institute of Brooklyn and his Ph.D. from Florida Atlantic University, Boca Raton, Florida. He designed radomes and phased array antennas at Raytheon before joining Motorola, where he received the Dan Noble Fellow Award for his research in antennas, propagation, and advanced communications systems. In 2000, he joined Time Domain Corporation to lead strategic technology development. He has lectured and published internationally; and holds more than 70 patents worldwide, including 31 issued in the United States. He was awarded Paper of the Year by IEEE–VTS and has authored, *Radiowave Propagation and Antennas for Personal Communications*, (Artech House), now in second edition, and contributed chapters to several other books and encyclopedias.

BIBLIOGRAPHY

1. *IEEE Standard Definitions of Terms for Antennas*, IEEE Std 145-1993, SH16279, March 18, 1993.
2. K. Siwiak, *Radiowave Propagation and Antennas for Personal Communications*, 2nd ed., Artech House, Norwood, MA, 1998.
3. G. J. Burke and A. J. Poggio, *Numerical Electromagnetics Code (NEC)—Method of Moments*, Lawrence Livermore Laboratory, NOSC Technical Document 116 (TD 116), Vols. 1 and 2, Jan. 1981.

4. H. C. Pocklington, Electrical oscillations in wires, *Proc. Cambridge Physical Society*, London, 1897, Vol. 9, pp. 324–333.
5. E. Hallén, Theoretical investigation into transmitting and receiving qualities of antennae, *Nova Acta Regiae Soc. Ser. Upps.* **II**(4): 1–44 (1938).
6. J. E. Storer, Impedance of thin-wire loop antennas, *Trans. AIEE* **75**(4): 609–619 (1965).
7. T. T. Wu, Theory of the thin circular antenna, *J. Math. Phys.* **3**: 1301–1304 (Nov.–Dec. 1962).
8. Q. Balzano and K. Siwiak, The near field of annular antennas, *IEEE Trans. Vehic. Technol.* **VT36**(4): 173–183 (Nov. 1987).
9. Q. Balzano and K. Siwiak, Radiation of annular antennas, *Correlations* (Motorola Eng. Bull., Motorola Inc., Schaumburg, IL, USA) **VI**(2): (1987).
10. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.
11. R. E. Collin, *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985.
12. E. C. Jordan and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1968.
13. R. Cohen and B. N. Taylor, The 1986 CODATA recommended values of the fundamental physical constants, *J. Res. Natl. Bureau Stand.* **92**(2): (March–April 1987).
14. U.S. Patent 1,577,421 (March 16, 1926), L. A. Hazeltine, Means for eliminating magnetic coupling between coils.
15. R. W. P. King and C. W. Harrison, Jr., *Antennas and Waves: A Modern Approach*, MIT Press, Cambridge, MA, 1969.
16. R. Courant and D. Hibert, *Methods of Mathematical Physics*, Interscience, New York, 1953.
17. U.S. Patent 3,588,905 (June 28, 1971), J. H. Dunlavy, Jr., Wide range tunable transmitting loop.
18. T. Hart, Small, high-efficiency loop antennas, *QST J. ARRL* 33–36 (June 1986).
19. K. Fujimoto, A. Henderson, K. Hirasawa, and J. R. James, *Small Antennas*, Wiley, New York, 1987.

LOW-BIT-RATE SPEECH CODING

Miguel Arjona Ramírez
Mario Minami
University of São Paulo
São Paulo, Brazil

1. INTRODUCTION

Speech coders were first used for encrypting the speech signal as they still are today for secure voice communications. But their most important use is bit rate saving to accommodate more users in a communications channel such as a mobile telephone cell or a packet network link. Alternatively, a high-resolution coder or a more elaborate coding method may be required to provide for a higher-fidelity playback.

Actually, the availability of ever broader-band connection and larger-capacity media has led some to consider speech coding as unnecessary but the increasing population of transmitters and the increasingly rich content have taken up the "bandwidth" made available by the introduction of broadband services.

Further, coding may be required to counter the noise present in the communication channel, such as a wireless connection, or the decay of the storage media, such as a magnetic or optical disk. In fact, such a coding, called *channel coding*, will increase the total bit rate, and this is usually on a par with encryption. In contrast, the coding mentioned before is called *source coding* and will be dealt with almost exclusively below.

The speech signal is an analog continuous waveform, and any digital representation of it incurs a distortion or lack of fidelity, which is irrelevant for high-fidelity rendering. High-fidelity representations are obtained by filtering the signal within a sufficiently wide frequency band, sampling it at regular intervals and then quantizing each amplitude so obtained with a large number of bits. This kind of direct digital coding is called *pulse-code modulation* (PCM). The sampling operation is reversible if properly done, and the large number of bits for quantizer codes makes it possible to have a large number of closely spaced coding levels, reducing quantization distortion.

Since human hearing has a finite sensitivity, a sufficiently fine digital representation may be considered "transparent" or essentially identical to the original signal. In the case of a general audio signal, a bit rate of 706 kbit/s per channel, compact-disk (CD) quality, is usually considered transparent, while for telephone speech 64 kbps (kilobits per second) is taken as toll quality (Table 1). Even though it is rather elusive to impose a range for low-bit-rate speech coding as it is a moving target, it seems that nowadays it is best bounded by 4 kbps from above, given the longlasting effort to settle for a toll quality speech coder at that rate at the ITU-T [1,2], and it is bounded by $\approx$1 kbps from below by considering mainly the expected range of leading coding techniques at the lower low-rate region and the upper very-low-rate region [3]. A very good and comprehensive reference to speech coding [4] located low rate between 2.4 kbps and 8 kbps just some years ago.

Table 1. Bit Rates of Typical Acoustic Signals

	Bandwidth (Hz–kHz)	Sampling Frequency	Bits per Sample	Bit Rate (kbps)
Narrowband speech	300–3.4	8.0	8	64
Wideband speech	50–7.0	16.0	14	224
Wideband audio (DAT format)	10–20.0	48.0	16	768
Wideband audio (CD format)	10–20.0	44.1	16	706

2. SPEECH MODELING FOR LOW-RATE SPEECH CODING

Speech is a time-varying signal that may be considered stationary during segments of some tens of milliseconds in general. For these segments, usually called *frames*, an overall characterization is often made by using a spectral model. Complementarily, the energy is imparted to a synthesis filter, which embodies the estimated spectral model, by an excitation signal also carrying more details of the fine structure of the signal spectrum, or else the spectral model may be sampled at selected frequencies or integrated over selected frequency bands in order to define a proper reconstructed signal. In addition, the incorporation into the excitation model of the requisite interpolation for the process of synthesis further extends it into the time–frequency domain.

2.1. Predictive Coders

During the first half of the twentieth century, filterbanks were used for synthesizing speech since the first voice coder or "vocoder" developed by Dudley. The major difficulty in vocoding was the separation of vocal source behavior from vocal-tract behavior in order to drive a source–filter model for synthesis. A didactic taxonomy of parametric coders is given by Deller et al. [5].

A manageable and accurate acoustical model of speech production was proposed by Fant in 1960, and a good approximation to it is provided by the linear prediction (LP) model. The LP model for speech analysis was originally proposed by Itakura and Saito in 1968 and Atal and Hanauer in 1971 [6], whose spectral models are short-term stationary and nonstationary, respectively. The stationary LP spectral model is the frequency response of

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

whose magnitude may be interpreted as a fit to the envelope of the short-term log spectrum of the signal as shown in Fig. 1. The order p, of the LP model has to be high enough to enable it to adjust to the overall shape of the spectrum, and the gain factor G allows an energy matching between the frequency response of the model and the spectrum of the signal. The LP model is particularly biased toward the peaks of the signal spectrum as opposed to the valleys and is particularly useful as a smooth peak-picking template for estimating the formants, sometimes not at likely places at first glance, like the second formant frequency in Fig. 1.

The excitation model proposed by Itakura and Saito combines two signal sources as shown in Fig. 2 whose relative intensities may be controlled by the two attenuation factors $U^{1/2}$ and $V^{1/2}$, which are interlocked by the relation

$$U + V = 1 \tag{2}$$

The pulse source, obtained for $V = 1$ and $U = 0$, is useful for generating voiced speech. In this mode, besides the gain factor G, the pulse repetition rate P has to be

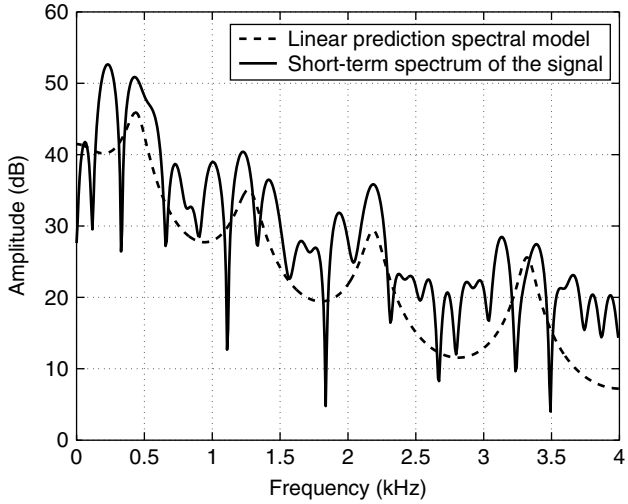


Figure 1. Linear prediction spectral fit to the envelope of the short-term log spectrum of the signal.

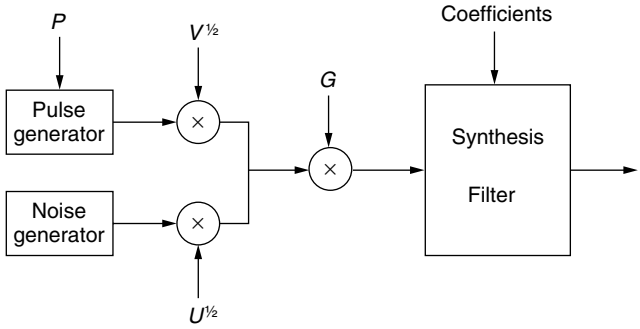


Figure 2. Mixed source–filter model for speech synthesis.

controlled. It is obtained in the coder as the pitch period of the speech signal through a pitch detection algorithm. The detected pitch period value may not be appropriate in many situations that may occur because of the quasiperiodic nature of voiced speech, the interaction of fundamental frequency (F_0) with the first formant or missing lower harmonics of F_0. On the other hand, for unvoiced speech the gain factor G is sufficient to match the power level of the pseudorandom source along with $U = 1$ and $V = 0$.

A better mixed excitation is produced by the mixed-excitation linear prediction (MELP) coder, which, besides combining pulse and noise excitations, is able to yield periodic and aperiodic pulses by position jitter [7]. Further, the composite mixed excitation undergoes adaptive spectral enhancement prior to going through the synthesis filter to produce the synthetic signal that is applied to the pulse dispersion filter.

2.2. Sinusoidal Coders

The voiced mode of speech production motivates the sine-wave representation of voiced speech segments by

$$s(n) = \sum_{k=1}^{K} A_{\cos}(\omega_k n + \phi_k) \tag{3}$$

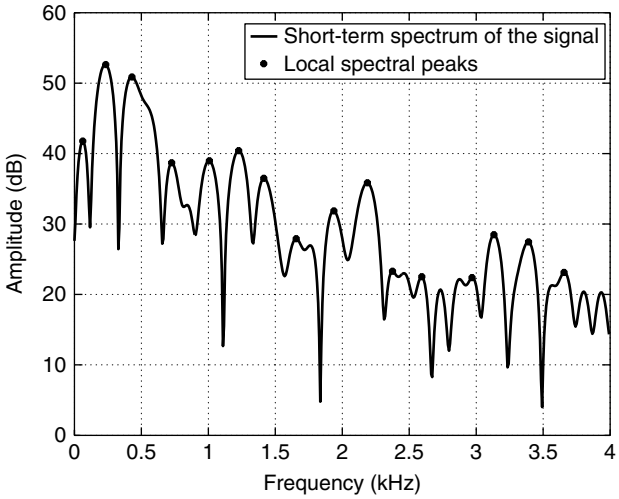


Figure 3. Short-term log spectrum of the signal with selected local peaks.

where A_k and ϕ_k are respectively the amplitude and phase of oscillator k, associated with the ω_k frequency track. This model quite makes sense in view of the spectrum of a voiced segment as can be seen in Fig. 3. As suggested in this figure, the peak frequencies $\{\omega_k,\ k = 1, 2, \ldots, K\}$ may be extracted and used as the oscillator frequencies in Eq. (3). For a strict periodic excitation model, $\omega_k = k\omega_0$, that is, the peak frequencies are equally interspaced and we have the so-called harmonic oscillator model. However, not all sinusoidal coders subscribe to this model because, by distinguishing small deviations from harmony, tonal artifacts may be guarded against. But the harmonic model is more amenable to low-rate implementation; thus other techniques have to be used to forestall the development of "buzzy" effects, which arise as a consequence of the forced additional periodicity.

The amplitudes may be constrained to lie on an envelope fit to the whole set of amplitudes, thereby enabling an efficient vector quantization of the amplitude spectrum. This amplitude model is compatible with the linear prediction filter described in Section 2.1, and the efficient quantization methods available for it may be borrowed, as is done for the sinusoidal transform coder (STC) [8].

Equation (3) may also be used for synthesizing unvoiced speech as long as the phases are random. In order to reduce the accuracy required of the voicing decision, a uniformly distributed random component is added to the phase of the oscillators with frequency above a voicing-dependent cutoff frequency in the STC as the lower harmonics of F_0 are responsible for the perception of pitch. In the multiband excitation (MBE) coder, the band around each frequency track is defined as either voiced or unvoiced, and Eq. (3) is not used for unvoiced synthesis; instead, filtered white noise is used. The bands are actually obtained after the signal has been windowed, and, as the windows have a finite bandwidth, this brings about a similarity of the sinusoidal coder with subband coders.

For low-rate coding, there is not enough rate for coding the phases, and phase models have to be used by the synthesizer such as the zero-phase model and the minimum-phase model. When there is a minimum-phase spectral model as in the latter case, the complex amplitude is obtained at no additional cost by sampling its frequency response as

$$H(e^{j\omega_k}) = A_k^{(r)} e^{j\phi_k^{(r)}} \tag{4}$$

where $A_k^{(r)}$ and $\phi_k^{(r)}$ are the reconstructed amplitude and phase of frequency track ω_k, respectively.

2.3. Waveform-Interpolation Coders

Waveform-interpolation coders usually apply linear prediction for estimating a filter whose excitation is made by interpolation of characteristic waveforms. Characteristic waveforms (CWs) are supposed to represent one cycle of excitation for voiced speech. The basic idea for the characteristic waveform stems from the Fourier series representation of a periodic signal, whose overtones are properly obtained by a Fourier series expansion. Therefore, the CW encapsulates the whole excitation spectrum, provided the signal is periodic. The rate of extraction of CWs may be as low as 40 Hz for voiced segments, as these waveforms are slowly varying in this case. On the other hand, for unvoiced segments the rate of extraction may have to be as high as 500 Hz but each segment may be represented with lower resolution [9].

The length of sampled characteristic waveforms varies as the pitch period. Therefore, their periods have to be normalized and aligned before coding for proper phase tracking. A continuous-time notation encapsulates a length normalization and the time-domain CW extraction process so that a two-dimensional surface may be built. The normalization of CW length is achieved by stretching or shrinking the waveforms to fit them within a normalized period of 2π radians. This normalized time within a period is referred to as the phase (ϕ). Assuming that linear prediction analysis has been performed and that the prediction residual has been determined for CW extraction and Fourier series representation, above and below the time–phase plane undulates the characteristic surface

$$u(t, \phi) = \sum_{k=1}^{K} \alpha_k(t) \cos(k\phi) + \beta_k(t) \sin(k\phi) \tag{5}$$

For the sake of coding efficiency, it is convenient to decompose the characteristic surface into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW). The SEW may be obtained by lowpass filtering $u(t, \phi)$ along the t axis as shown in Fig. 4 and represents the quasiperiodic component of speech excitation, whereas the REW may be obtained by highpass filtering $u(t, \phi)$ along the t axis, representing the random component of speech excitation. Both components must add up to the original surface:

$$u(t, \phi) = u_{\mathrm{SEW}}(t, \phi) + u_{\mathrm{REW}}(t, \phi) \tag{6}$$

Characteristic waveforms may be represented by means other than a Fourier series but in the latter case they may be compared to sinusoidal coders, having smaller

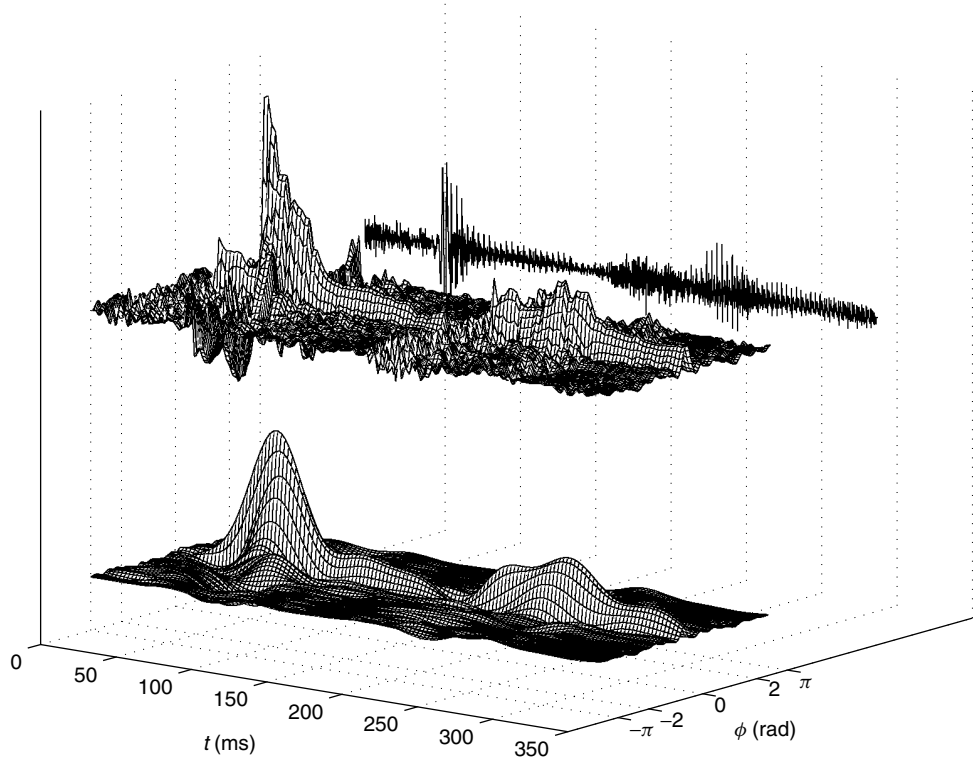


Figure 4. Characteristic surface for WI coding the residual signal given behind whose underlying CWs have been extracted at a 400 Hz rate. Its SEW component is also shown below, which has been obtained by lowpass filtering the characteristic surface along the time axis with a cutoff frequency of 20 Hz.

interpolation rates due to a more flexible time–frequency representation and to a higher resolution in time. For a common framework that encompasses both sinusoidal coding and waveform interpolation, please refer to Ref. 10, where the issue of perfect reconstruction in the absence of quantization errors is brought to bear.

3. PARAMETER ESTIMATION FROM SPEECH SEGMENTS

The linear prediction model was introduced in the last section along with the simplest excitation types for time-domain implementation, the frequency-domain parametric models of greater use for low-bit-rate coders and a harmonic excitation model, including waveform interpolation. In this section a more detailed description is provided of the structures used to constrain the excitation and the algorithms used for estimating its parameters. The segmentation of the speech signal for its analysis is complemented by its concatenation in the synthesis phase.

Although the initial goal was a medium bit rate range from 8 to 16 kbps, a different approach has come to be used for coding the excitation, called *code-excited linear prediction* (CELP) [11]. The two most important concepts in CELP coding are (1) an excitation quantization by sets of consecutive samples, which is a kind of vector quantization (VQ) of the excitation, and (2) a search criterion based on the reconstruction error instead of the prediction error or differential signal. Figure 5 has been drawn stressing these main distinguishing features.

A CELP coder is provided with a finite set of codevectors to be used for reconstructing each segment or subframe of the original signal. A collection of M codevectors is said to be a codebook of size M. Prior to searching the excitation, a filter is estimated through LP analysis (see Section 2.1) to have a frequency response matching the short-term spectral envelope of a block of the original signal called a "frame." Each frame typically consists of two to four excitation subframes, and the synthesis filter is determined for each subframe by interpolation from the LP filters of neighboring frames. As shown in Fig. 5, each codevector $\mathbf{c}_k$ in turn, for $k = 1, 2, \ldots, M$ is filtered by the synthesis filter

$$H(z) = \frac{1}{1 - P(z)} \tag{7}$$

generating all around the encoding loop a reconstruction error vector ε_k. This process of determining the signal to be synthesized within the coder is called the *analysis-by-synthesis method*. It allows the coder to anticipate the best strategy constrained to the situation that the synthesizer will face. Thus, the minimum square reconstruction error is identified as

$$i = \underset{k=1,2,\ldots,M}{\operatorname{argmin}} \{\|\varepsilon_k\|^2\} \tag{8}$$

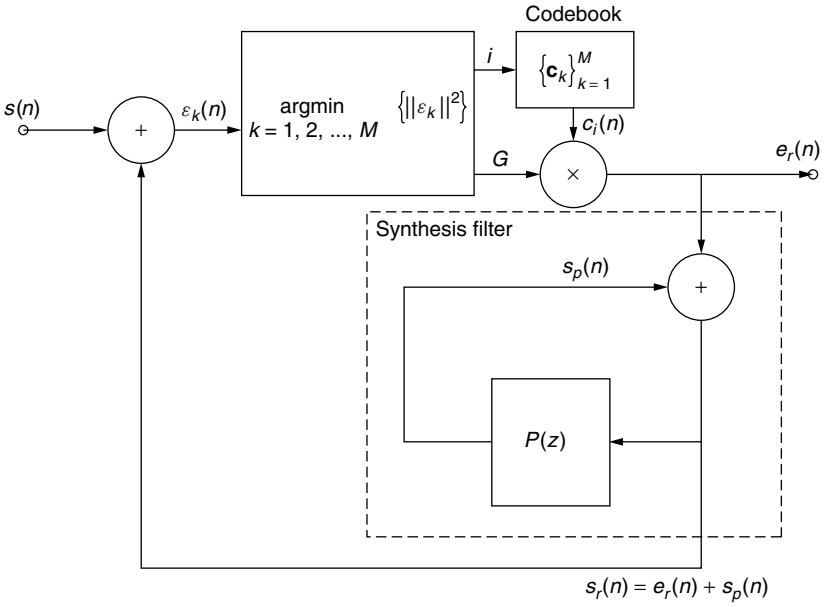


Figure 5. Conceptual block diagram for CELP coding.

after an exhaustive search all through the codebook and the actual excitation is delivered as the scaled version

$$\mathbf{e}_r = G\mathbf{c}_i \tag{9}$$

of codevector $\mathbf{c}_i$, where the scale factor $G = G_i$ has been calculated to minimize the square reconstruction error $\|\varepsilon_i\|^2$ for codevector $\mathbf{c}_i$.

Actually, a CELP coder applies a perceptual spectral weighting to the reconstruction error prior to the minimization by means of the weighting filter, defined by a function of the adaptive synthesis filter as

$$W(z) = \frac{H(z/\gamma_2)}{H(z/\gamma_1)} \tag{10}$$

where $0 < \gamma_2 < \gamma_1 \leq 1$ are bandwidth expansion factors. A very usual combination of values is $\gamma_2 = 0.8$ and $\gamma_1 = 1$. Overall, the weighting filter serves the dual purpose of deemphasizing the power spectral density of the reconstruction error around the formant frequencies where the power spectrum of the signal is higher and emphasizing the spectral density of the error in between the formant frequencies where hearing perception is more sensitive to an extraneous error. Both actions come about as consequences of the frequency response of $W(z)$ in Fig. 6. In much the same way, in order to achieve a reconstructed signal with a higher perceptual quality, an open-loop postfilter is usually applied to the reconstructed signal, which is defined as a function of the synthesis filter as well (see Fig. 7).

Additionally, toll quality reconstruction can be achieved only if there is a rather precise means of imposing the periodicity of voiced speech segments on the reconstructed signal. This goal can be achieved by using a second adaptive codebook in the CELP coder. This adaptive codebook is fed on a subframe basis the composite coded excitation

$$e(n) = G_a c_a(n) + G_f c_f(n) \tag{11}$$

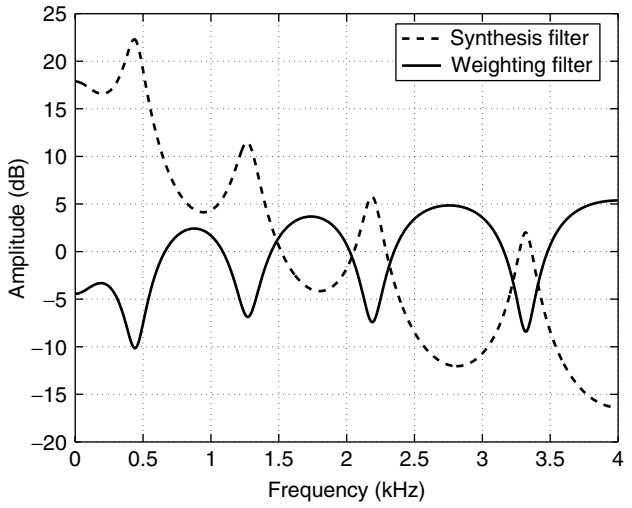


Figure 6. Frequency responses of synthesis filter and corresponding perceptual weighting filter.

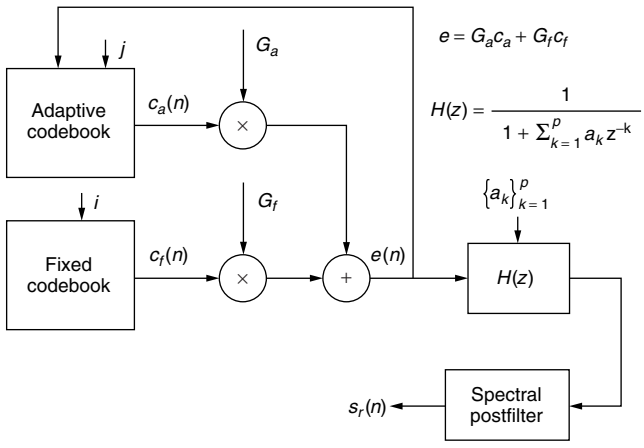


Figure 7. Two-codebook CELP synthesis model.

where $c_a(n)$ stands for the adaptive codevector with its gain factor G_a and $c_f(n)$ with its gain factor G_f represents the fixed excitation, depicted by the only codebook in Fig. 5. The enhanced synthesis model for this CELP coder is illustrated in Fig. 7.

Nonetheless, the fixed codebook structure and its search algorithms have been the target for developments leading to the widespread applicability of CELP coders. The fixed codebook in the original CELP coder was stochastically populated from samples of independent and identically Gaussian distributed vectors [11]. As the complexity of exhaustive searches through the codebook was overwhelming for the then-current signal processors, more efficient search methods were derived (discussed in Section 4), which required more structured codebooks such as the center-clipped and overlapped stochastic codebooks. Their searches have lower operational complexity due to the sparse amplitude distribution and the overlapped nature of their codevectors. The latter allows for the use of efficient search techniques originally developed for the adaptive codebook. Even more surprising, they enhance the speech quality as well [12] to a level considered good enough for secure voice and cellular applications at low to medium rates.

Meanwhile, predictive waveform coders borrow the idea of impulse excitation from parametric LP coders (see Section 2.1) in order to decrease the bit rate but with a twist to deliver higher quality, which involves the increase in the number of pulses per pitch period. A subframe of multipulse excitation is given by

$$e(n) = G \sum_{k=0}^{M-1} \alpha_k \delta(n - m_k), \quad n = 0, 1, \ldots, L-1 \qquad (12)$$

where M is the number of pulses per excitation subframe, L is the length of the subframe, α_k and m_k respectively represent individual pulse amplitude and position, and G is a common excitation vector gain. This new approach was called "multipulse excitation" and is very complex in its most general formulation [13]. Moreover, a constrained version of it, known by "regular pulse excitation with long-term predictor" (RPE-LTP), was adopted for the Global System for Mobile Communications (GSM) full-rate standard coder for digital telephony and is notable for its low complexity [14].

This kind of excitation was further structured and inserted into a CELP coder. Pulse positions were constrained to lie in different tracks, which cover in principle all the positions in the excitation subframe, whereas pulse amplitudes α_k were restricted to either plus or minus one. The latter feature and its conceptual connection to error-correction codes has established the name "algebraic CELP" for this kind of excitation. These deterministic sparse codebooks made their entrance into standard speech coding with the G.729 conjugate structure, algebraic CELP (CS-ACELP) coder [15]. A general ACELP position grid is given in Table 2 for an M-pulse codebook over an L-sample subframe.

As the bit rate is decreased, further modeling and classification of the signal has to be done at the encoder in order to keep speech quality about the same. For instance, the pitch synchronous innovation CELP (PSI-CELP) coder adapts the fixed random codevectors in voiced frames to have periodicity [16].

Table 2. ACELP Position Grid for *M*-Pulse Tracks over an *L-Sample* Subframe

Track	Positions				
0	0	M	$2M$	$\cdots$	$L-M$
1	1	$M+1$	$2M+1$	$\cdots$	$L-M+1$
2	2	$M+1$	$2M+2$	$\cdots$	$L-M+2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$M-1$	$M-1$	$2M-1$	$3M-1$	$\cdots$	$L-1$

Surprisingly, the analysis-by-synthesis operation of CELP is proving capable of delivering toll-quality speech at lower rates when generalized to allow for a mixture of open-loop and closed-loop procedures [2] where parameters and excitation are determined in an open-loop fashion for clearly recognizable subframe types such as stationary periodic or voiced segments and closed-loop algorithms are used for unvoiced or transient segments. Because of the scarcity of bits for representing the excitation, it makes sense to predistort the target vector for closed-loop searches when it is clearly voiced since it becomes easier to match a codevector to it. The predistortion has to be perceptually transparent such as the time warping described in Ref. 17.

In a different trend, the development of text-to-speech (TTS) systems has been moving away from the rule-based, expert system approach to the new framework of concatenative synthesis, based on model fitting with statistical signal processing [18]. In rule-based systems subword speech units are designed as well as rules for concatenating them that take into account the coarticulation between neighboring units as well as their exchange for allophonic variations. On the other hand, concatenative synthesis systems are based on the acquisition of a large database of connected speech from an individual speaker containing instances of coarticulation between all possible units. For the latter systems, the synthesis consists of selecting the largest possible string of original database subunits, thereby borrowing their natural concatenation. The final postprocessing stage of the TTS adjusts the prosody of the synthetic signal, mostly by pitch and timescale modifications. For segment selection, a concatenative synthesizer uses both an acoustic cost within each segment and a concatenation cost between consecutive segments [3]. If the input feature vector sequence $\mathbf{F} = \mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$ is to be synthesized by the unit sequence $\mathbf{U} = \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$, the acoustic cost may be defined by

$$J_A(\mathbf{f}_m, \mathbf{u}_m) = \sum_{k=1}^{K} (f_{m,k} - u_{m,k})^2 \qquad (13)$$

for segment m, where k indices through the K features are selected for comparison, normally the spectral representation of the subunits, and the concatenation cost

may be calculated by

$$J_C(\mathbf{u}_{m-1}, \mathbf{u}_m) = \sum_{k=1}^{K} (u_{m-1,k} - u_{m,k})^2 \tag{14}$$

The best subunit sequence is selected by minimization of the total cost $J(\mathbf{F}, \mathbf{U})$ whose simplest definition is

$$J(\mathbf{F}, \mathbf{U}) = \sum_{m=1}^{N} J_A(\mathbf{f}_m, \mathbf{u}_m) + \sum_{m=2}^{N} J_C(\mathbf{u}_{m-1}, \mathbf{u}_m) \tag{15}$$

With the use of these kinds of cost measures in their analysis, concatenative synthesizers are becoming more similar to speech coders.

4. LOW-RATE CODING APPROACHES

Speech coding allows more users to share a communications channel such as a mobile telephone cell or a packet network link and is concerned with the economical representation of a speech signal with a given distortion for a specified implementation complexity level. Traditionally, a fixed bit rate and an acceptable maximum distortion are specified. More generally, the required maximum bit rate or the acceptable maximum distortion level may be specified. Actually, for modern cellular or packet communications, sometimes the bit rate may be dictated by channel traffic constraints, requiring variable-bit-rate coders.

Objective fidelity measures such as the segmental signal-to-noise ratio (SNRSEG) are very practical for coder development, while more perceptual methods such as objective distortion measures, including the perceptual speech quality measure (PSQM) [19], which use to advantage the limitations of the human ear, may be used instead. But subjective the opinion of human listeners is still the best gauge of fidelity and may be assessed by the *mean opinion score* (MOS), obtained in formal listening tests where each listener classifies the speech stimulus on the 5-point scale shown in Table 3.

Coder complexity constrains the possibilities of rate distortion tradeoff. Its major component is operational complexity, liable to be measured in million instructions per second (MIPS) [20]. An artistic conception of the fidelity versus rate behavior of low-rate coders for two levels of complexity is presented in Fig. 8, anchored by some real coder test points, listed in Table 4. It should be mentioned that these fidelity curves pass through a kind of "knee" around the 4 kbps rate, where they evolve at a lower slope, eventually reaching a virtual plateau at high rates [21].

Low-bit-rate implementations of models tested at higher rates need compensation for the loss of resolution or reduction of parameters, whereas very-low-bit-rate

Table 3. Quality Scale for Subjective Listening Rating

Quality	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

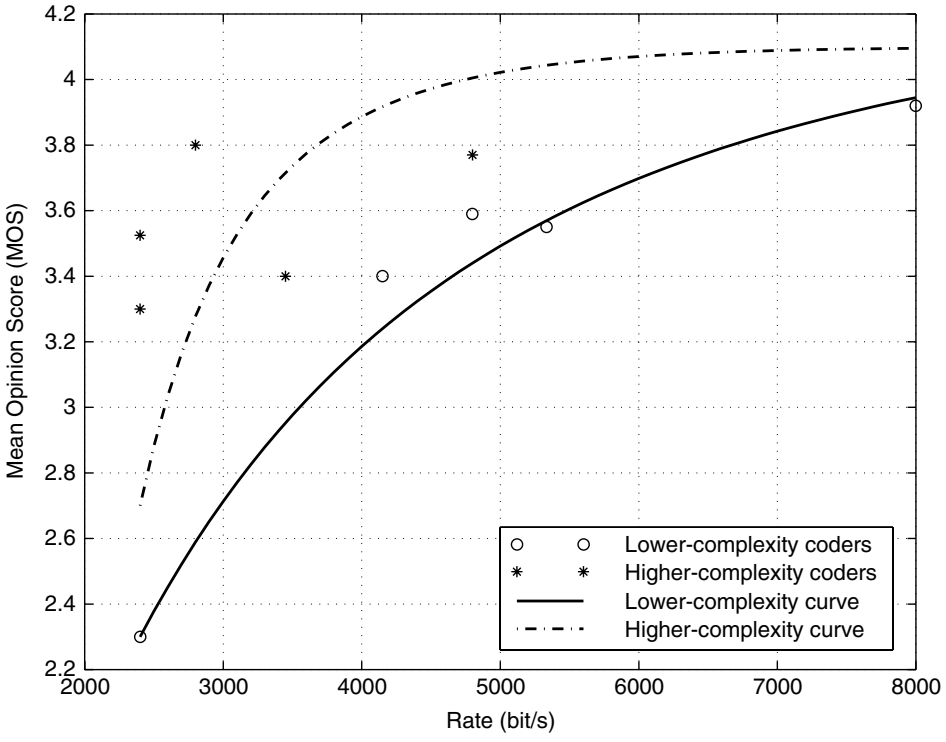


Figure 8. Conception of the fidelity versus rate behavior of low-rate speech coders for two levels of complexity, anchored by some real coder test points, listed in Table 4.

Table 4. Speech Quality and Operational Complexity of Some Selected Coders[a]

Coder	Bit rate (kbps)	Quality (MOS)	Complexity (Mips)[b]	Ref.
LPC-10e, FS-1015	2.40	2.30	8.7	37
MELP, FS-1017	2.40	3.30	20.4	37
EWI	2.80	~3.80	~30.0	33,35,38
PSI-CELP, RCR PDC half-rate	3.45	~3.40	23.0	14,16,38,39
IMBE, INMARSAT-M System	4.15	3.40	7.0	4,14
CELP, FS-1016	4.80	3.59	17.0	37,40
STC	4.80	3.53	~25.0	8
WI	4.80	3.77	~25.0	40
ACELP, G.723.1	5.33	3.55	16.0	33,41
CS-ACELP, G.729	8.00	3.92	20.0	38,41

[a] *Caution*: These performance and complexity figures were obtained under different test and implementation conditions and should be used only as a first guess in comparisons. Tilde (~) indicates estimate.
[b] Million instructions per second.

implementations admit refinements when upgraded to the low-rate range. In general, low-rate implementations require higher complexity algorithms and incur longer algorithmic delay. But a reduction in complexity may render the original algorithm useful for a number of applications. This is one reason why a number of efficient search algorithms have been proposed since the inception of the CELP coder such as that due to Hernández-Gómez et al. [22], who proposed a residual-based preselection of codevectors and the efficient transform-domain search algorithms elaborated by Trancoso and Atal [23]. Another preselection of codevectors was proposed [24] on the basis of the correlation between the backward-filtered target vector and segments of codevectors. The latter efficient search was called "focused search" and was adopted for the reference ITU-T 8-kbps CS-ACELP coder [15] with an open-loop signal-selected pulse amplitude approach. This coder is used for transmitting voice over packet networks among other applications.

In fact, the acceptance of this family of coders is so wide that most of the second-generation digital cellular coders use it, including the Telecommunications Industry Association (TIA) IS641 enhanced full-rate (EFR) coder [25] and the IS127 enhanced variable-rate coder (EVRC) [26] as well as the GSM EFR coder [27]. In addition, a general-purpose efficient search algorithm for ACELP fixed excitation codebook has been proposed, the joint position and amplitude search (JPAS) [28], which includes a closed-loop sequential pulse amplitude determination, and a more efficient search for the EVRC [29] has been advanced as well. Also, a generalization of "algebraic pulses" by "algebraic subvectors" is the basis for the algebraic vector quantized CELP (AVQ-CELP) search, which enhances the IS127 coder and uses open-loop subvector preselection in order to make it more efficient [30].

As the bit rate is decreased below 6 kbps, ACELP coder quality degrades because of the uniform pulse density in the pulse position grid [31] and the high level of sparsity in the resulting excitation waveform. In an effort to push down the bit rate for ACELP applications, pulse dispersion techniques have been proposed [32,33]. The former closed-loop technique is incorporated in a partially qualified candidate for the ITU-T 4-kbps coder [2]. Furthermore, parametric coders such as MELP also implement pulse dispersion but as an open-loop enhancement in the decoder as mentioned in Section 2.1. Along with pulse dispersion, the pulse position in the grid should be changed adaptively since it will not be able to cover all the positions [31,34].

Another technique that holds promise for lower-bit-rate coding is target vector predistortion. Time-warping predistortions have already been proposed as mentioned in Section 3 and even used in the IS127 EVRC.

The segments coded open loop may use enhanced vocoderlike techniques such as those used in the MELP or sinusoidal coders or, alternatively, WI techniques with a partial use of analysis-by-synthesis methods [35].

The judicious application of these enhancement techniques requires classification of the signal into voice or silence. In the former case, the speech signal is classified into voiced and unvoiced stationary segments at least. Even the identification of transients may be required as a next step. Branching out further, speech classification might get down to subunits such as triphones, diphones, and phones. In these cases the segmentation is event-driven, similar to the method used for very-low-rate coding [36]. Nevertheless, one should bear in mind that irregular segmentation requires timescale modification as a postprocessing stage, which may introduce annoying artifacts into the reconstructed signal. So sometimes it may be wise to maintain regular frame-based segmentation even at very low rates in order to ensure a certain uniform quality level [3].

In conclusion, the CELP framework with some relaxed waveform matching constraints, allowing for perceptual quality preserving signal predistortion and more segments of simple parametric coding, is very likely to be able to achieve toll quality at 4 kbps. It is anticipated as well that

coders based on codebooks of sequences of speech subunits with properly defined distortion measures will also play an important role in advancing the toll quality frontier into the low-bit-rate range.

BIOGRAPHIES

Miguel Arjona Ramírez received the E.E. degree from Instituto Tecnológico de Aeronáutica (ITA), Brazil, in 1980 and the M.S.E.E. and Dr.E.E. degrees from University of São Paulo, Brazil, in 1992 and 1997, respectively.

In 1981, while studying at Philips International Institute, he worked with coding algorithms for a formant speech synthesizer at Philips Electronic Components and Materials Laboratories, The Netherlands.

He joined Itautec Informática S.A., São Paulo, Brazil, as a Full Development Engineer in 1982, eventually becoming an Engineering Development Group Leader for Interactive Voice Response (IVR) Systems in 1988.

He is currently Assistant Professor at Escola Politécnica, University of São Paulo, where he has conducted research on predictive speech coders since 1991. Dr. Arjona Ramírez became Senior Member of the IEEE in 2000. His research interests include signal compression, speech coding and recognition, and audio coding with applications to circuit and packet telephony.

Mario Minami received the B.S. degree in physics in 1989 from Physics Institute of University of Sao Paulo, Sao Paulo, Brazil, and the M.S. and Ph.D. degrees in electrical engineering from the Politechnic School of University of Sao Paulo, Sao Paulo, Brazil in 1993 and 1998, respectively. He joined the OSUC — Obras Sociais, Universitarias e Culturais in 1989-91 as a Research Engineer and professor. At OSUC he worked on the design and development of didactic DSP projects. Since 1993, he has been a Research Scientist at LPS-EPUSP — Signal Processing Laboratory of Politechnic School of University of Sao Paulo. His areas of interest are speech and speaker recognition, speech coding, database for speech applications and channel equalization algorithms.

BIBLIOGRAPHY

1. S. Dimolitsas, C. Ravishankar, and G. Schröder, Current objectives in 4-kb/s wireline-quality speech coding standardization, *IEEE Signal Process. Lett.* **1**(11): 157–159 (Nov. 1994).
2. J. Thyssen et al., A candidate for the ITU-T 4 kbit/s speech coding standard, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, 2001, Vol. 2, pp. 681–684.
3. K.-S. Lee and R. V. Cox, A very low bit rate speech coder based on a recognition/synthesis paradigm, *IEEE Trans. Speech Audio Process.* **9**(5): 482–491 (July 2001).
4. A. S. Spanias, Speech coding: A tutorial review, *Proc. IEEE* **82**(10): 1541–1582 (Oct. 1994).
5. J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, 1993, Chap. 7, pp. 459–487.
6. J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer, Berlin, 1976.
7. A. McCree et al., A 2.4 kbit/s MELP coder candidate for the new U. S. Federal Standard, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Atlanta, 1996, Vol. 1, pp. 200–203.
8. R. J. McAulay and J. F. Quatieri, Sinusoidal coding, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 121–173.
9. W. B. Kleijn and K. K. Paliwal, An introduction to speech coding, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 1–47.
10. W. B. Kleijn, A frame interpretation of sinusoidal coding and waveform interpolation, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Istanbul, 2000, Vol. 3, pp. 1475–1478.
11. M. R. Schroeder and B. S. Atal, Code-excited linear prediction (CELP): High quality speech at very low bit rates, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tampa, 1985, Vol. 2, pp. 437–440.
12. W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, Fast methods for the CELP speech coding algorithm, *IEEE Trans. Acoust. Speech, Signal Process.* **38**(8): 1330–1342 (Aug. 1990).
13. B. S. Atal and J. R. Remde, A new model of LPC excitation for producing natural-sounding speech at low bit rates, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Paris, 1982, Vol. 1, pp. 614–617.
14. R. V. Cox, Speech coding standards, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 49–78.
15. R. Salami et al., Design and description of CS-ACELP, a toll quality 8 kb/s speech coder, *IEEE Trans. Speech Audio Process.* **6**(2): 116–130 (March 1998).
16. K. Mano et al., Design of a pitch synchronous innovation CELP coder for mobile communications, *IEEE J. Select. Areas Commun.* **13**(1): 31–40 (Jan. 1995).
17. W. B. Kleijn, R. P. Ramachandran, and P. Kroon, Generalized analysis-by-synthesis coding and its application to pitch prediction, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, San Francisco, 1992, Vol. 1, pp. 23–26.
18. Y. Sagisaka and N. Iwahashi, Objective optimization in algorithms for text-to-speech synthesis, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 685–706.
19. *Objective Quality Measurement of Telephone-Band (300–3400 Hz) Speech Codecs*, ITU-T Recommendation P.861, Aug. 1996.
20. P. Kroon, Evaluation of speech coders, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 467–494.
21. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
22. L. A. Hernández-Gómez, F. J. Casajús-Quirós, A. R. Figueiras-Vidal, and R. García-Gómez, On the behaviour of reduced complexity code-excited linear prediction (CELP), in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tokyo, 1986, Vol. 1, pp. 469–472.
23. I. M. Trancoso and B. S. Atal, Efficient procedures for finding the optimum innovation in stochastic coders, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tokyo, 1986, Vol. 4, pp. 2375–2378.

24. C. Laflamme et al., 16 kbps wideband speech coding technique based on algebraic CELP, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Toronto, 1991, Vol. 1, pp. 13–16.

25. T. Honkanen, J. Vainio, K. Järvinen, and P. Haavisto, Enhanced full rate codec for IS-136 digital cellular system, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 731–734.

26. *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, TIA/EIA/IS-127, July 1996.

27. K. Järvinen et al., GSM enhanced full rate speech codec, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 771–774.

28. M. A. Ramírez and M. Gerken, Joint position and amplitude search of algebraic multipulses, *IEEE Trans. Speech Audio Process.* **8**(5): 633–637 (Sept. 2000).

29. H. Park, Efficient codebook search method of EVRC speech codec, *IEEE Signal Process. Lett.* **7**(1): 1–2 (Jan. 2000).

30. F. Liu and R. Heidari, Improving EVRC half rate by the algebraic VQ-CELP, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, 1999, Vol. 4, pp. 2299–2302.

31. V. Cuperman et al., A novel approach to excitation coding in low-bit-rate high-quality CELP coders, *Proc. IEEE Workshop on Speech Coding*, Delavan, Wisconsin, 2000, pp. 14–16.

32. K. Yasunaga, H. Ehara, K. Yoshida, and T. Morii, Dispersed-pulse codebook and its application to a 4 kb/s speech coder, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Istanbul, 2000, Vol. 3, pp. 1503–1506.

33. M. A. Ramírez, Sparsity compensation for speech coders, *Proc. IEEE GLOBECOM*, San Antonio, 2001, Vol. 4, pp. 2475–2478.

34. T. Amada, K. Miseki, and M. Akamine, CELP speech coding based on an adaptive pulse position codebook, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, 1999, Vol. 1, pp. 13–16.

35. O. Gottesman and A. Gersho, Enhanced waveform interpolative coding at low bit-rate, *IEEE Trans. Speech Audio Process.* **9**(8): 786–798 (Nov. 2001).

36. C. S. Xydeas and T. M. Chapman, Segmental prototype interpolation coding, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, 1999, Vol. 4, pp. 2311–2314.

37. M. A. Kohler, A comparison of the new 2.4 kbps MELP federal standard with other standard coders, in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 1587–1590.

38. M. E. Perkins, K. Evans, D. Pascal, and L. A. Thorpe, Characterizing the subjective performance of the ITU-T 8 kb/s speech coding algorithm — ITU-T G.729, *IEEE Commun. Mag.* **35**(9): 74–81 (Sept. 1997).

39. K. Mano, Design of a toll-quality 4-kbit/s speech coder based on phase-adaptive PSI-CELP, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Munich, 1997, Vol. 2, pp. 755–758.

40. W. B. Kleijn and J. Haagen, Waveform interpolation for coding and synthesis, in W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, 1995, pp. 175–207.

41. R. V. Cox and P. Kroon, Low bit-rate speech coders for multimedia communication, *IEEE Commun. Mag.* **34**(12): 34–41 (Dec. 1996).

LOW-DENSITY PARITY-CHECK CODES: DESIGN AND DECODING

SARAH J. JOHNSON*
STEVEN R. WELLER†
University of Newcastle
Callaghan, Australia

1. INTRODUCTION

The publication of Claude Shannon's 1948 paper, "A mathematical theory of communication" [1], marked the beginning of coding theory. In his paper, Shannon established that every communication channel has associated with it a number called the channel *capacity*. He proved that arbitrarily reliable communication is possible even through channels that corrupt the data sent over them, but only if information is transmitted at a rate less than the channel capacity.

In the simplest case, transmitted messages consist of strings of 0s and 1s, and errors introduced by the channel consist of bit inversions: $0 \rightarrow 1$ and $1 \rightarrow 0$. The essential idea of forward error control coding is to augment messages to produce codewords containing deliberately introduced redundancy, or *check* bits. With care, these check bits can be added in such a way that codewords are sufficiently distinct from one another so that the transmitted message can be correctly inferred at the receiver, even when some bits in the codeword are corrupted during transmission over the channel.

While Shannon's noisy channel coding theorem establishes the existence of capacity-approaching codes, it provides no explicit guidance as to how the codes should be chosen, nor how messages can be recovered from the noise-corrupted channel output. The challenges to communicating reliably at rates close to the Shannon limit are therefore twofold: (1) to design sets of suitably distinct codewords and (2) to devise methods for extracting estimates of transmitted messages from the output of a noise-contaminated channel, and to do so without excessive decoder complexity.

In this article, we consider code design and decoding for a family of error correction codes known as *low-density parity-check* (LDPC) *block codes*. In the simplest form of a parity-check code, a single parity-check equation provides for the detection, but not correction, of a single bit inversion in a received codeword. To permit correction of errors induced by channel noise, additional parity checks can be added at the expense of a decrease in the rate of transmission. Low-density parity-check codes are a special case of such codes. Here "low density" refers to the sparsity of the parity-check matrix characterizing the

* Work supported by a CSIRO Telecommunications & Industrial Physics postgraduate scholarship and the Centre for Integrated Dynamics and Control (CIDAC).
† Work supported in part by Bell Laboratories Australia, Lucent Technologies, as well as the Australian Research Council under Linkage Project Grant LP0211210, and the Centre for Integrated Dynamics and Control (CIDAC).

code. Each parity-check equation checks few message bits, and each message bit is involved in only a few parity-check equations. A delicate balance exists in the construction of appropriate parity-check matrices, since excessive sparsity leads to uselessly weak codes.

First presented by Gallager in his 1962 thesis [2,3], low-density parity-check codes are capable of performance extraordinarily close to the Shannon limit when appropriately decoded. Codes that approach the Shannon limit to within 0.04 of a decibel have been constructed. Figure 1 shows a comparison of the performance of LDPC codes with the performance of some well-known error correction codes. The key to extracting maximal benefit from LDPC codes is *soft-decision* decoding, which starts with a more subtle model for channel-induced errors than simple bit inversions. Rather than requiring that the receiver initially make *hard decisions* at the channel output, and so insisting that each received bit be assessed as either 0 or 1, whatever is the more likely, soft-decision decoders use knowledge of the channel noise statistics to feed probabilistic (or "soft") information on received bits into the decorder.

The final ingredient in implementing soft-decision decoders with acceptable decoder complexity are *iterative* schemes that handle the soft information in an efficient manner. Soft iterative decoders for LDPC codes make essential use of *graphs* to represent codes, passing probabilistic messages along the edges of the graph. The use of graphs for iterative decoding can be traced to Gallager, although for over 30 years barely a handful of researchers pursued the consequences of Gallager's work. This situation changed dramatically with the independent rediscovery of LDPC codes by several researchers in the mid-1990s, and graph-based representations of codes are now an integral feature in the development of both the theoretical understanding and implementation of iterative decoders.

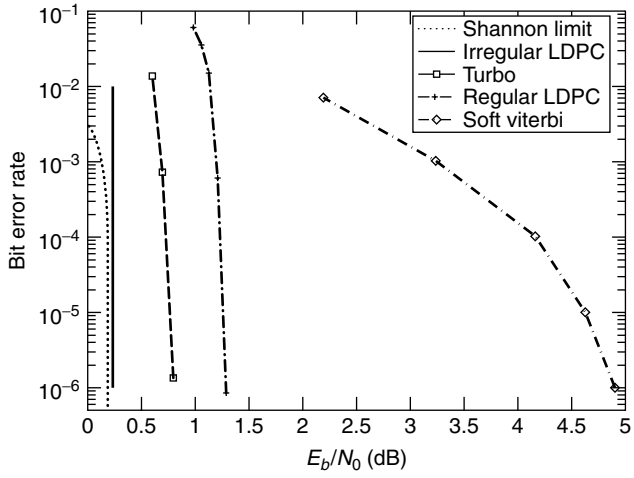


Figure 1. Bit error rate performance of rate-$\frac{1}{2}$ error correction codes on an additive white Gaussian noise channel. From right to left, soft Viterbi decoding of a constraint length 7 convolutional code; sum–product decoding of a regular Gallager code with blocklength 65,389 [8]; a Turbo code with $2 + 32$ states, 16,384-bit interleaver, and 18 iterations *http://www331.jp1.nasa.gov/public/TurboPerf.html*; sum–product decoding of a blocklength 10^7 optimized irregular code [16]; and the Shannon limit at rate $\frac{1}{2}$.

In this article, we begin by introducing parity checks and codes defined by their parity-check matrices. To introduce iterative decoding we present in Section 3 a hard-decision iterative algorithm that is not very powerful, but suggestive of how graph-based iterative decoding algorithms work. The soft-decision iterative decoding algorithm for LDPC codes known as *sum–product decoding* is presented in Section 4. Section 5 focuses on the relationship between the codes and the decoding algorithm, as expressed in the graphical representation of LDPC codes, and Section 6 considers the design of LDPC codes. The article concludes with a discussion of the connections of this work to other topics and future directions in the area.

2. LOW-DENSITY PARITY-CHECK CODES

2.1. Parity-Check Codes

The simplest possible error detection scheme is the single parity check, which involves the addition of a single extra bit to a binary message. Whether this parity bit should be a 0 or a 1 depends on whether even or odd parity is being used. In even parity, the additional bit added to each message ensures an even number of 1s in each transmitted codeword. For example, since the 7-bit ASCII code for the letter S is 1010011, a parity bit is added as the eighth bit. If even parity is being used, the value of the parity bit is 0 to form the codeword 10100110.

More formally, for the 7-bit ASCII plus even parity code, we define a codeword c to have the following structure:

$$c = c_1\, c_2\, c_3\, c_4\, c_5\, c_6\, c_7\, c_8$$

where each c_i is either 0 or 1, and every codeword satisfies the constraint

$$c_1 \oplus c_2 \oplus c_3 \oplus c_4 \oplus c_5 \oplus c_6 \oplus c_7 \oplus c_8 = 0 \qquad (1)$$

Here the symbol $\oplus$ represents modulo-2 addition, which is equal to 1 if the ordinary sum is odd and 0 if the ordinary sum is even. Whereas the inversion of a single bit due to channel noise can be easily detected with a single parity check [as (1) is no longer satisfied by the noise-corrupted codeword], this code is not sufficiently powerful to indicate which bit (or bits) was (were) inverted. Moreover, since any even number of bit inversions produces a word satisfying the constraint (1), any even numbers of errors go undetected by this simple code.

One measure of the ability of a code to detect errors is the *minimum distance* of the code. The *Hamming distance* between two codewords is defined as the number of bit positions in which they differ. For example, the codewords 10100110 and 10000111 differ in positions 3 and 8, so the Hamming distance between them is 2. The minimum distance of a code, $d_{\min}$, is defined as the smallest Hamming distance between any pair of codewords in the code. For the even parity code $d_{\min} = 2$, so the corruption of 2 bits in a codeword can result in another valid codeword and will consequently not be detected.

Detecting more than a single bit error calls for increased redundancy in the form of additional parity checks. To illustrate, suppose that we define a codeword c to have the following structure:

$$c = c_1\ c_2\ c_3\ c_4\ c_5\ c_6$$

where each c_i is either 0 or 1, and c is constrained by three parity-check equations:

$$\begin{array}{l} c_1 \oplus c_2 \oplus c_4 = 0 \\ c_2 \oplus c_3 \oplus c_5 = 0 \\ c_1 \oplus c_2 \oplus c_3 \oplus c_6 = 0 \end{array} \Leftrightarrow \underbrace{\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}}_{H} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (2)$$

In matrix form we have that $c = [c_1\ c_2\ c_3\ c_4\ c_5\ c_6]$ is a codeword if and only if it satisfies the constraint

$$Hc^T = 0 \quad (3)$$

where the *parity-check matrix*, H, contains the set of parity-check equations that define the code. To generate the codeword for a given message, the code constraints can be rewritten in the form

$$\begin{array}{l} c_4 = c_1 \oplus c_2 \\ c_5 = c_2 \oplus c_3 \\ c_6 = c_1 \oplus c_2 \oplus c_3 \end{array} \Leftrightarrow [c_1\ c_2\ c_3\ c_4\ c_5\ c_6] = [c_1\ c_2\ c_3] \underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}}_{G} \quad (4)$$

where bits c_1, c_2, and c_3 contain the 3-bit message, and parity-check bits c_4, c_5, and c_6 are calculated from the message. Thus, for example, the message 110 produces parity-check bits $c_4 = 1 \oplus 1 = 0$, $c_5 = 1 \oplus 0 = 1$, and $c_6 = 1 \oplus 1 \oplus 0 = 0$, and hence the codeword 110010. The matrix G is the *generator matrix* of the code. Substituting each of the $2^3 = 8$ distinct messages $c_1\ c_2\ c_3 = 000, 001, \ldots, 111$ into Eq. (4) yields the following set of codewords:

$$\begin{array}{cccc} 000000 & 001011 & 010111 & 011100 \\ 100101 & 101110 & 110010 & 111001 \end{array} \quad (5)$$

The reception of a word that is not in this set of codewords can be detected using the parity-check constraint equation (3). Suppose, for example, that the word $r = 101011$ is received from the channel. Substitution into Eq. (3) gives

$$Hr^T = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad (6)$$

which is nonzero, and so the word 101011 is not a codeword of our code.

To go further and correct the error requires that the decoder determine the codeword most likely to have been sent. Since it is reasonable to assume that the number of errors will more likely be small rather than large, the required codeword is the one closest in *Hamming distance* to the received word. By comparison of the received word $r = 101011$ with each codeword in (5), the closest codeword is $c = 001011$, which is at Hamming distance 1 from r. The minimum distance of this code is 3, so a single bit error always results in a word closer to the codeword that was sent than any other codeword, and hence can always be corrected. In general, for a code with minimum distance $d_{\min}$, e bit errors can always be corrected by choosing the closest codeword whenever

$$e \leq \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor \quad (7)$$

where $\lfloor x \rfloor$ is the largest integer that is at most x.

Error correction by direct search is feasible only when the number of distinct codewords is small. For codes with thousands of bits in a codeword, it becomes far too computationally expensive to directly compare the received word with every codeword in the code, and numerous ingenious solutions have been proposed, including choosing codes that are cyclic or, as presented in this article, devising iterative methods to decode the received word.

2.2. Low-Density Codes

LDPC codes are parity-check codes with the requirement that H is low-density, so that the vast majority of entries are zero. A parity-check matrix is *regular* if each code bit is contained in a fixed number, w_c, of parity checks and each parity-check equation contains a fixed number, w_r, of code bits. If an LDPC code is described by a regular parity-check matrix it is called a (w_c, w_r)-*regular LDPC* code; otherwise it is an *irregular LDPC* code.

Importantly, an error correction code can be described by more than one *parity-check matrix*, where H is a valid parity-check matrix for a code, provided (3) holds for all codewords in the code. Two parity-check matrices for the same code need not even have the same number of rows; what is required is that the rank over GF(2) of both be the same, since the number of message bits, k, in a binary code is

$$k = n - \text{rank}_2(H) \quad (8)$$

where $\text{rank}_2(H)$ is the number of rows in H that are linearly dependent over GF(2). To illustrate, we give a regular parity-check matrix for the code of (2) with $w_c = 2$, $w_r = 3$, and $\text{rank}_2(H) = 3$, which satisfies (3)

$$H = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \quad (9)$$

A *Tanner graph* is a graphical representation of H that facilitates iterative decoding of the code. The Tanner

graph consists of two sets of vertices: *n bit vertices* (or *bit nodes*) and *m parity-check vertices* (or *check nodes*), where there is a parity-check vertex for every parity-check equation in H and a bit vertex for every codeword bit. Each parity-check vertex is connected by an edge to the bit vertices corresponding to the code bits included in that parity-check equation. The Tanner graph of the parity-check matrix (9) is shown in Fig. 2. As the number of edges leaving the bit vertices must equal the number of edges leaving the parity-check vertices it follows that for a regular code:

$$m \cdot w_r = n \cdot w_c \tag{10}$$

A *cycle* in a Tanner graph is a sequence of connected vertices that start and end at the same vertex in the graph, and that contain other vertices no more than once. The length of a cycle is the number of edges it contains, and the *girth* of a graph is the size of its smallest cycle. A cycle of size 6 is shown in bold in Fig. 2.

Traditionally, the parity-check matrices of LDPC codes have been defined pseudorandomly subject to the requirement that H be sparse, and code construction of binary LDPC codes involves randomly assigning a small number of the values in an all-zero matrix to be 1. The lack of any obvious algebraic structure in randomly constructed LDPC codes sets them apart from traditional parity-check codes. The properties and performance of LDPC codes are often considered in terms of the ensemble performance of all possible codes with a specified structure (e.g., a certain node degree distribution), reminiscent of the methods used by Shannon in proving his noisy channel coding theorem. More recent research has considered the design of LDPC codes with specific properties, such as large girth, and we describe in Sections 5 and 6 methods to design LDPC codes. For sum–product decoding, however, no additional structure beyond a sparse parity-check matrix is required, and in the following two sections we present decoding algorithms requiring only the existence of a sparse H.

3. ITERATIVE DECODING

To illustrate the process of iterative decoding, a *bit-flipping* algorithm is presented, based on an initial hard decision (0 or 1) assessment of each received bit. An essential part of iterative decoding is the passing of messages between the nodes of the Tanner graph of the code. For the bit-flipping algorithm, the messages are simple; a bit node sends a message to each of the check nodes to which it is connected, declaring whether it is a 1 or a 0, and each check node sends a message to each of the bit nodes to which it is connected, declaring whether the parity check is satisfied. The sum–product algorithm for LDPC codes operates similarly but with more complicated messages.

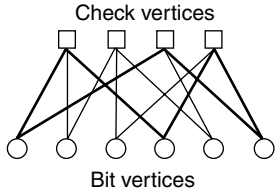


Figure 2. Tanner graph representation of the parity-check matrix in (9). A 6-cycle is shown in bold.

The bit-flipping decoding algorithm is as follows:

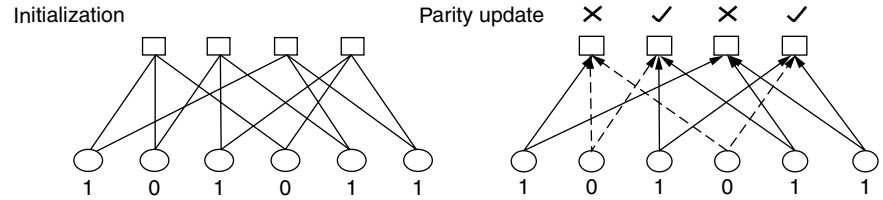
Step 1. Initialization. Each bit node is assigned the bit value received from the channel, and sends messages to the check nodes to which it is connected indicating this value.

Step 2. Parity update. Using the messages from the bit nodes, each check node calculates whether its parity-check equation is satisfied. If all parity-check equations are satisfied, the algorithm terminates; otherwise each check node sends messages to the bit nodes to which it is connected indicating whether the parity-check equation is satisfied.

Step 3. Bit update. If the majority of the messages received by each bit node are "not satisfied," the bit node flips its current value; otherwise the value is retained. If the maximum number of allowed iterations is reached, the algorithm terminates and a failure to converge is reported; otherwise each bit node sends new messages to the check nodes to which it is connected, indicating its value, and the algorithm returns to step 2.

To illustrate the operation of the bit-flipping decoder, we take the code of (9) and again assume that the codeword $c = 001011$ is sent, and the word $r = 101011$ is received from the channel. The steps required to decode this received word are shown in Fig. 3. In step 1 the bit values are initialized to be 1, 0, 1, 0, 1, and 1, respectively, and messages are sent to the check nodes indicating these values. In step 2 each parity-check equation is satisfied only if an even number of the bits included in the parity-check equation are 1. For the first and third check nodes this is not the case, and so they send "not satisfied" messages to the bits to which they are connected. In step 3 the first bit has the majority of its messages indicating "not satisfied" and so flips its value from 1 to 0. Step 2 is repeated, and since now all four parity-check equations are satisfied, the algorithm halts and returns $c = 001011$ as the decoded codeword. The received word has therefore been correctly decoded without requiring an explicit search over all possible codewords.

The existence of cycles in the Tanner graph of a code reduces the effectiveness of the iterative decoding process. To illustrate the detrimental effect of a 4-cycle, we adjust the code of the previous example to obtain the new code shown in Fig. 4. A valid codeword for this code is 001001, but again we assume that the first bit is corrupted, so that $r = 101001$ is received from the channel. The steps of the bit-flipping algorithm for this received word are shown in Fig. 4. In step 1 the initial bit values are 1, 0, 1, 0, 0, and 1, respectively, and messages are sent to the check nodes indicating these values. Step 2 reveals that the first and second parity-check equations are not satisfied. In step 3 both the first and second bits have the majority of their messages indicating "not satisfied," and so both flip their bit values. When step 2 is repeated, we see that the first and second parity-check equations are again not satisfied.



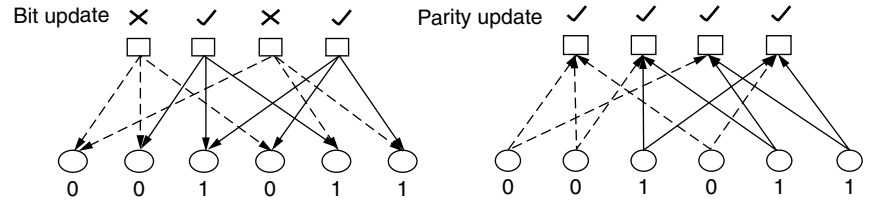
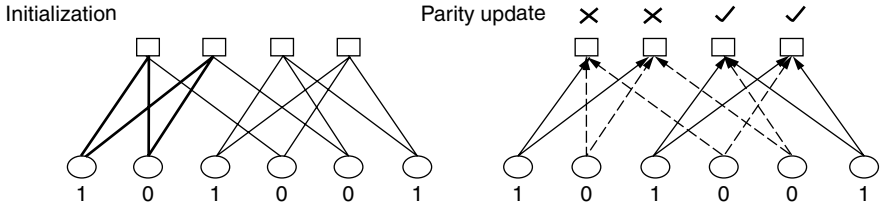


Figure 3. Bit-flipping decoding of the received word $r = 101011$. Each diagram indicates the decision made at each step of the decoding algorithm based on the messages from the previous step. A cross ($\times$) represents that the parity check is not satisfied, while a tick ($\surd$) indicates that it is satisfied. For the messages, a dashed arrow corresponds to the messages "bit = 0" or "check not satisfied," while a solid arrow corresponds to "bit = 1" or "check satisfied."



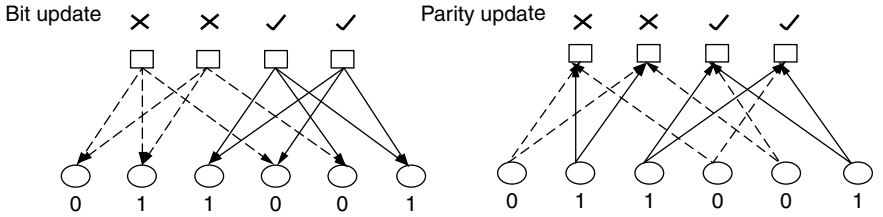


Figure 4. Bit-flipping decoding of the received word $r = 101001$. A 4-cycle is shown in bold in the first diagram.

Further iterations at this point simply cause the first 2 bits to flip their values in such a way that one of them is always incorrect; the algorithm fails to converge. As a result of the 4-cycle, each of the first two codeword bits is involved in the same two parity-check equations, and so when neither of the parity-check equations is satisfied, it is not possible to determine which bit is causing the error.

4. SUM–PRODUCT DECODING

The sum–product decoding algorithm, also called *belief propagation* decoding, was first introduced by Gallager in his 1962 thesis, where he applied it to the decoding of pseudorandomly constructed LDPC codes. For block lengths of 10^7, highly optimized irregular LDPC codes decoded with the sum-product algorithm are now known to be capable of approaching the Shannon limit to within hundredths of a decibel on the binary input additive white Gaussian noise (AWGN) channel. In the early 1960s, however, limited computing resources prevented Gallager from demonstrating the capabilities of iteratively decoded LDPC codes for blocklengths longer than $\sim$500, and for over 30 years his work was ignored by only a handful of researchers. It was only rediscovered by several researchers in the wake of Turbo decoding [4], which has subsequently been recognized as an instance of the sum–product algorithm.

The sum–product algorithm can be regarded as being similar to the bit-flipping algorithm described in the previous section, but with the messages representing each decision (check met, or bit value equal to 1) now probabilistic values represented by loglikelihood ratios. Whereas with bit-flipping decoding an initial hard decision is made on the signal from the channel, what is actually received is a string of real values where the sign of the received value represents a binary decision — 0 if positive and 1 if negative — and the magnitude of the received value is a measure of the confidence in that decision. A shortcoming of using only hard decisions when decoding is that the information relating to the confidence of the signal, the soft information, is discarded. Soft-decision decoders, such as the sum–product decoder, make use of the soft received information, together with knowledge of the channel properties, to obtain probabilistic expressions for the transmitted signal.

For a binary signal, if p is the probability of a 1, then $1 - p$ is the probability of a 0 that is represented as a

loglikelihood ratio (LLR) by

$$\mathrm{LLR}(p) = \log_e\left(\frac{1-p}{p}\right) \tag{11}$$

The sign of LLR(p) is the hard decision, and the magnitude |LLR(p)| is the reliability of this decision. One benefit of the logarithmic representation of probabilities is that whereas probabilities need to be multiplied, loglikelihood ratios need only be added, reducing implementation complexity.

The aim of sum–product decoding is to compute the *a posteriori probability* (APP) for each codeword bit, $P_i = P\{c_i = 1 \mid N\}$, which is the probability that the ith codeword bit is a 1 conditional on the event N that all parity-check constraints are satisfied. The *intrinsic* or *a priori probability*, P_i^{int}, is the original bit probability independent of knowledge of the code constraints, and the *extrinsic* probability P_i^{ext} represents what has been learnt from the event N.

The sum–product algorithm iteratively computes an approximation of the APP value for each code bit. The approximations are exact if the code is cycle-free. Extrinsic information gained from the parity-check constraints in one iteration is used as a priori information for the subsequent iteration. The extrinsic bit information obtained from a parity-check constraint is independent of the a priori value for that bit at the start of the iteration. The extrinsic information provided in subsequent iterations remains independent of the original a priori probability until that information is returned via a cycle.

To compute the extrinsic probability of a codeword bit i from the jth parity-check equation, we determine the probability that the parity-check equation is satisfied if bit i is assumed to be a 1, which is the probability that an odd number of the other codeword bits are a 1:

$$P_{i,j} = \frac{1}{2} + \frac{1}{2}\prod_{i' \in B_j, i' \neq i}(1 - 2P_{i'}^{\mathrm{int}}) \tag{12}$$

The notation B_j represents the set of column locations of the bits in the jth parity-check equation of the code considered. Similarly, A_i is the set of row locations of the parity-check equations which check on the ith bit of the code. To put (12) into loglikelihood notation we note that

$$\tanh\left(\frac{1}{2}\log_e\left(\frac{1-p}{p}\right)\right) = 1 - 2p$$

to give

$$\mathrm{LLR}(P_{i,j}^{\mathrm{ext}}) = \log_e\left(\frac{1 + \prod_{i' \in B_j, i' \neq i}\tanh(\mathrm{LLR}(P_{i'}^{\mathrm{int}})/2)}{1 - \prod_{i' \in B_j, i' \neq i}\tanh(\mathrm{LLR}(P_{i'}^{\mathrm{int}})/2)}\right)$$

The LLR of the estimated APP of the ith bit at each iteration is then simply

$$\mathrm{LLR}(P_i) = \mathrm{LLR}(P_i^{\mathrm{int}}) + \sum_{j \in A_i}\mathrm{LLR}(P_{i,j}^{\mathrm{ext}})$$

The sum–product algorithm is as follows:

Step 1. Initialization. The initial message sent from bit node i to the check node j is the LLR of the (soft) received signal y_i given knowledge of the channel properties. For an AWGN channel with signal-to-noise ratio E_b/N_0, this is

$$L_{i,j} = R_i = 4y_i\frac{E_b}{N_0} \tag{13}$$

Step 2. Check to bit. The extrinsic message from check node j to bit node i is the probability that parity check j is satisfied if bit i is assumed to be a 1 expressed as an LLR:

$$E_{i,j} = \log_e\left(\frac{1 + \prod_{i' \in B_j, i' \neq i}\tanh(L_{i'j}/2)}{1 - \prod_{i' \in B_j, i' \neq i}\tanh(L_{i'j}/2)}\right) \tag{14}$$

Step 3. Codeword test. The combined LLR is the sum of the extrinsic LLRs and the original LLR calculated in step 1:

$$L_i = \sum_{j \in A_i} E_{i,j} + R_i \tag{15}$$

For each bit a hard decision is made:

$$z_i = \begin{cases} 1, & L_i \leq 0 \\ 0, & L_i > 0 \end{cases}$$

If $z = [z_1, \ldots, z_n]$ is a valid codeword ($Hz^T = 0$), or if the maximum number of allowed iterations have been completed, the algorithm terminates.

Step 4. Bit to check. The message sent by each bit node to the check nodes to which it is connected is similar to (15), except that bit i sends to check node j a LLR calculated without using the information from check node j:

$$L_{i,j} = \sum_{j' \in A_i, j' \neq j} E_{i,j'} + R_i \tag{16}$$

Return to step 2.

The application of Eqs. (14) and (16) to the code in (9) is demonstrated in Fig. 5. The extrinsic information passed from a check node to a bit node is independent of the probability value for that bit. The extrinsic information from the check nodes is then used as a priori information for the bit nodes in the subsequent iteration.

To illustrate the power of sum–product decoding, we revisit the example of Fig. 3, where the codeword sent is 0 0 1 0 1 1. Suppose that the channel is AWGN with $E_b/N_0 = 1.25$ and the received signal is $y = -0.1\ \ 0.5\ \ -0.8\ \ 1.0\ \ -0.7\ \ 0.5$. There are now two bits in error if the hard decision of the signal is considered: bits 1 and 6. Figure 6 illustrates the operation of the sum–product decoding algorithm, as described in Eqs. (13)–(16), to decode this received signal which terminates in three iterations. The existence of an exact termination rule for the sum–product algorithm has two important benefits: (1) a failure to converge is always

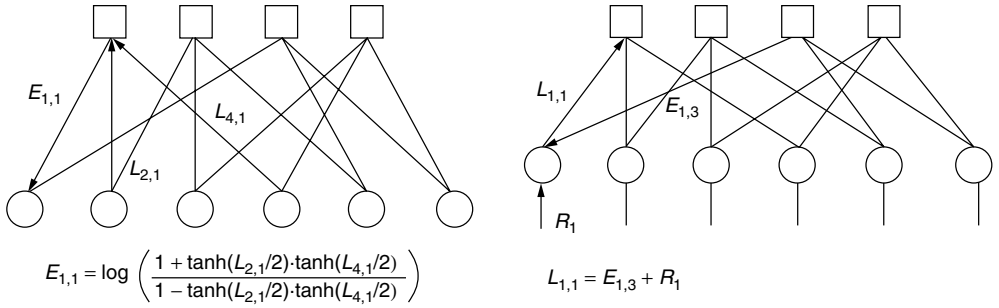


Figure 5. An example of the messages for sum–product decoding. Calculation of the extrinsic message sent to bit 1 depends on messages from bits 2 and 4 but not from bit 1. Similarly, the message sent to check 1 is independent of the message just received from it.

Iteration 1

$R = [\,-0.5000 \quad 2.5000 \quad -4.0000 \quad 5.0000 \quad -3.5000 \quad 2.5000\,]$

$[\,1\ 0\ 1\ 0\ 1\ 0\,]$ as a hard decision

$$E = \begin{bmatrix} 2.4217 & -0.4930 & . & -0.4217 & . & . \\ . & 3.0265 & -2.1892 & . & -2.3001 & . \\ -2.1892 & . & . & . & -0.4217 & 0.4696 \\ . & . & 2.4217 & -2.3001 & . & -3.6869 \end{bmatrix}$$

$L = [\,-0.2676 \quad 5.0334 \quad -3.7676 \quad 2.2783 \quad -6.2217 \quad -0.7173\,]$

$z = [\,1\ 0\ 1\ 0\ 1\ 1\,]$

$Hz^T = [\,1\ 0\ 1\ 0\,]^T \Rightarrow$ Continue

$$L = \begin{bmatrix} -2.6892 & 5.5265 & . & 2.6999 & . & . \\ . & 2.0070 & -1.5783 & . & -3.9217 & . \\ 1.9217 & . & . & . & -5.8001 & -1.1869 \\ . & . & -6.1892 & 4.5783 & . & 2.9696 \end{bmatrix}$$

Iteration 2

$$E = \begin{bmatrix} 2.6426 & -2.0060 & . & -2.6326 & . & . \\ . & 1.4907 & -1.8721 & . & -1.1041 & . \\ 1.1779 & . & . & . & -0.8388 & -1.9016 \\ . & . & 2.7877 & -2.9305 & . & -4.3963 \end{bmatrix}$$

$L = [\quad 3.3206 \quad 1.9848 \quad -3.0845 \quad -0.5630 \quad -5.4429 \quad -3.7979\,]$

$z = [\,0\ 0\ 1\ 1\ 1\ 1\,]$

$Hz^T = [\,1\ 0\ 0\ 1\,]^T \Rightarrow$ Continue

$$L = \begin{bmatrix} 0.6779 & 3.9907 & . & 2.0695 & . & . \\ . & 0.4940 & -1.2123 & . & -4.3388 & . \\ 2.1426 & . & . & . & -4.6041 & -1.8963 \\ . & . & -5.8721 & 2.3674 & . & 0.5984 \end{bmatrix}$$

Iteration 3

$$E = \begin{bmatrix} 1.9352 & 0.5180 & . & 0.6515 & . & . \\ . & 1.1733 & -0.4808 & . & -0.2637 & . \\ 1.8332 & . & . & . & -1.3362 & -2.0620 \\ . & . & 0.4912 & -0.5948 & . & -2.3381 \end{bmatrix}$$

$L = [\quad 3.2684 \quad 4.1912 \quad -3.9896 \quad 5.0567 \quad -5.0999 \quad -1.9001\,]$

$z = [\,0\ 0\ 1\ 0\ 1\ 1\,]$

$Hz^T = [\,0\ 0\ 0\ 0\,]^T \Rightarrow$ Terminate

Figure 6. Operation of sum–product decoding with the code from (9) when the codeword [001011] is sent through an AWGN channel with $E_b/N_0 = 1.25$ and the vector $[-0.1\ 0.5\ -0.8\ 1.0\ -0.7\ 0.5]$ is received. The sum–product decoder converges to the correct codeword after three iterations.

detected, and (2) additional iterations are avoided once a solution has been found.

There are variations to the sum–product algorithm presented here. The *min-sum* algorithm, for example, simplifies the calculation of (14) by recognizing that the term corresponding to the smallest $L_{i'j}$ dominates the product term, and so the product can be approximated by a minimum; the resulting algorithm thus requires calculation of only minimums and additions. An alternative approach, designed to bridge the gap between the error performance of sum-product decoding and that of maximum-likelihood (ML) decoding, finishes each iteration of sum–product decoding with ordered statistic decoding, with the algorithm terminating when a specified number of iterations have returned the same codeword [5].

5. CODES, GRAPHS, AND CYCLES

The relationship between LDPC codes and their decoding is closely associated with the graph-based representations of the codes. The most obvious example of this is the link between the existence of cycles in the Tanner graph of the code to both the analysis and performance of sum–product decoding of the code. In his work, Gallager used a graphical representation of the bit and parity-check sets of regular LDPC codes, to describe the application of iterative APP decoding. The systematic study of codes on graphs, however, is due largely to Tanner, who, in 1981, extended the single parity-check constraints of Gallager's LDPC codes to arbitrary linear code constraints, foresaw the advantages for very large-scale integration (VLSI) implementations of iterative decoders, and formalized the use of bipartite graphs for describing families of codes [6]. In so doing, Tanner also founded the topic of algebraic methods for constructing graphs suitable for sum–product decoding.

By proving the convergence of the sum–product algorithm for codes whose graphs are free of cycles, Tanner was also the first to formally recognize the importance of cycle-free graphs in the context of iterative decoding. The effect of cycles on the practical performance of LDPC codes was demonstrated by simulation experiments when LDPC codes were rediscovered by MacKay and Neal [7] (among others) in the mid-1990s, and the beneficial effects of using graphs free of short cycles were shown [8]. Given the detrimental effects of cycles on the convergence of iterative decoders, it is natural to seek strong codes whose Tanner graphs are free of cycles. An important negative result in this direction was established by Etzion et al. [9], who showed that for linear codes of rate $k/n \geq 0.5$, which can be represented by a Tanner graph without cycles, the minimum distance is at most 2.

As the existence of cycles in a graph makes analysis of the decoding algorithm difficult, most analyses consider the asymptotic performance of iterative decoding on graphs with asymptotically unbounded girth. This analysis provides thresholds to the performance of LDPC codes with sum–product decoding. As we will see in the following section, this process can be used to select LDPC code properties that improve the threshold values, a process that works well even though the resulting codes contain cycles.

To date very little analysis has been presented regarding the convergence of iterative decoding methods on graphs with cycles, and the majority of the work in this area can be found in Ref. 10. Gallager suggested that the dependencies introduced by cycles have a relatively minor effect and tend to cancel each other out somewhat. This "seems to work" philosophy has underlined the performance of sum–product decoding on graphs with cycles for much of the (short) history of the topic. It is only relatively recently that exact analysis on the expected performance of codes with cycles has emerged. Di et al. [11] use finite-length analysis to give the exact average bit and block error probabilities for any regular ensemble of LDPC codes over the binary erasure channel when decoded iteratively; however, there is as yet no such analysis for irregular codes or more general channel models.

Besides cycles, Sipser and Spielman [12] showed that the expansion of the graph is a significant factor in the application of iterative decoding. Using only a simple hard-decision decoding algorithm, they proved that a fixed fraction of errors in an LDPC code can be corrected in linear time provided that the Tanner graph of the code is a sufficiently good expander. That is, any subset S of bit vertices of size m or less is connected to at least $\in |S|$ constraint vertices, for some defined m and $\in$.

6. DESIGNING LDPC CODES

For the most part, LDPC codes are designed by first choosing the required blocklength and node degree distributions, then pseudorandomly constructing a parity-check matrix, or graph, with these properties. A generator matrix for the code can then be found using Gaussian elimination [8]. Gallager, for example, considered the ensemble of all (w_r, w_c)-regular matrices with rows divided into w_c submatrices, where the first contain w_r copies of the identity matrix and subsequent submatrices are random column permutations of the first. Using ensembles of matrices defined in this way, Gallager was able to find the maximum crossover probability of the binary symmetric channel (BSC) for which LDPC codes could be used to transmit information reliably using a simple hard-decision decoding algorithm.

Luby et al.extended the class of LDPC ensembles to those with irregular node degrees and showed that irregular codes are capable of outperforming regular codes [13]. In extending Gallager's analysis to irregular ensembles, Luby et al. introduced tools based on linear programming for designing irregular code ensembles for which the maximum allowed crossover probability of the binary symmetric channel is optimized [14]. Resulting from this work are the "tornado codes," a family of codes that approach the capacity of the erasure channel and can be encoded and decoded in linear time.

Richardson and Urbanke extended the work of Luby et al. to any binary input memoryless channel and to soft-decision message-passing decoding [15]. They determined the capacity of message-passing decoders applied to LDPC code ensembles by a method called *density evolution*.

For sum–product decoding density evolution makes it possible to determine the corresponding capacity to any degree of accuracy and hence determine the ensemble with node degree distribution that gives the best capacity. Once a code ensemble has been chosen a code from that ensemble is realized pseudorandomly. By carefully choosing a code from an optimized ensemble, Chung et al. have demonstrated the best performance to date of LDPC codes in terms of approaching the Shannon limit [16].

A more recent development in the design of LDPC codes is the introduction of algebraic LDPC codes, the most promising of which are the finite-geometry codes proposed by Lucas et al. [17], which are cyclic and described by sparse 4-cycle free graphs. An important outcome of this work with finite-geometry codes was the demonstration that highly redundant parity-check matrices can lead to very good iterative decoding performance without the need for very long blocklengths. Although the probability of a random graph having a highly redundant parity-check matrix is vanishingly small, the field of *combinatorial designs* offers a rich source of algebraic constructions for matrices that are both sparse and redundant. In particular, there has been much interest in balanced incomplete block designs (BIBDs) to produce sparse matrices for LDPC codes that are 4-cycle-free. For codes with greater girth, generalized quadrangle designs give the maximum possible girth for a graph with given diameter [18]. Both generalized quadrangles and BIBDs are subsets of the more general class of combinatorial structures called *partial geometries*, a possible source of further good algebraic LDPC codes [19].

In comparison with more traditional forms of error-correcting codes, the minimum distance of LDPC codes plays a substantially reduced role. There are two reasons for this: (1) the lack of any obvious algebraic structure in pseudorandomly constructed LDPC codes makes the calculation of minimum distance infeasible for long codes, and most analyses focus on the average distance function for an ensemble of LDPC codes; and (2) the absence of conspicuous flattening of the bit-error-rate (BER) curve at moderate to high signal-to-noise ratios (the "error floor") strongly suggests that minimum distance properties are simply not as important for LDPC codes as for traditional codes. Indeed, it has been established that to achieve capacity on the binary erasure channel when using irregular LDPC codes, the codes cannot have large minimum distances [20].

7. CONNECTIONS AND FUTURE DIRECTIONS

Following the rediscovery of Gallager's iterative LDPC decoding algorithm in the mid-1990s, the notion of an iterative algorithm operating on a graph has been generalized and is now capable of unifying a wide range of apparently different algorithms from the domains of digital communications, signal processing, and even artificial intelligence. An important generalization of Tanner graphs was presented by Wiberg in his 1996 Ph.D. thesis [21]. Wiberg introduced *state variables* into the graphical framework, thereby establishing a connection between codes on graphs and the trellis complexity of codes, and was the first to observe that on cycle-free graphs, the sum–product (respectively, min-sum) algorithm performs APP (respectively, ML) decoding.

In an even more general setting, the role of the Tanner graph is taken by a *factor graph* [22]. Central to the unification of message-passing algorithms via factor graphs is the recognition that many computationally efficient signal processing algorithms exploit the manner in which a global cost function acting on many variables can be factorized into the product of simpler local functions, each of which operates on a subset of the variables. In this setting a (bipartite) factor graph encodes the factorization of the global cost function, with each local function node connected by edges only to those variable nodes associated with its arguments.

The sum–product algorithm operating on a factor graph uses message passing to solve the *marginalize product-of-functions* (MPF) problem which lies at the heart of many signal processing problems. In addition to the iterative decoding of LDPC codes, specific instances of the sum–product algorithm operating on suitably defined factor graphs include the forward/backward algorithm (also known as the *BCJR* (Bahl–Cocke–Jelinek–Raviv) *algorithm* [23] or *APP decoding algorithm*), the Viterbi algorithm, the Kalman filter, Pearl's belief propagation algorithm for Bayesian networks, and the iterative decoding of "*Turbo codes*," or parallel concatenated convolutional codes.

For high-performance applications, LDPC codes are naturally seen as competitors to Turbo codes. LDPC codes are capable of outperforming Turbo codes for blocklengths greater than $\sim 10^5$, and the error floors of LDPC codes at BERs below $\sim 10^{-5}$ are typically much less pronounced than those of Turbo codes. Moreover, the inherent parallelism of the sum–product decoding algorithm is more readily exploited with LDPC codes than their Turbo counterparts, where block interleavers pose formidable challenges to achieving high throughput [24]. Despite these impressive advantages, LDPC codes lag behind Turbo codes in real-world applications. The exceptional simulation performance of the original Turbo codes [4,25] generated intense interest in these codes, and variants of them were subsequently incorporated into proposals for third-generation (3G) wireless systems such as the Third Generation Partnership Project (3GPP), a global consortium of standards-setting organizations [26]. Whatever performance advantages of very long LDPC codes over Turbo codes there may be, the invention of Turbo codes some 3 years prior to the (re)discovery of LDPC codes has given them a distinct advantage in wireless communications, where blocklengths of at most several thousand are typical, and where compliance with global standards is paramount.

One serious shortcoming of LDPC codes is their potentially high *encoding* complexity, which is in general

quadratic in the blocklength, and compares poorly with the linear time encoding of Turbo codes. Finding computationally efficient encoders is therefore critical for LDPC codes to be considered as serious contenders for replacing Turbo codes in future generations of forward error correction devices. Several approaches have been suggested, including the manipulation of the parity-check matrix to establish that while the complexity is, strictly speaking, quadratic, the actual number of encoding operations grows essentially linearly with blocklength. For some irregular LDPC codes whose degree distributions have been optimized to allow transmission near to capacity, the encoding complexity can be shown to be truly linear in blocklength [27]. A very different approach to the encoding complexity problem is to employ cyclic, or quasicyclic, codes as LDPC codes, as encoding can be achieved in linear time using simple feedback shift registers [28].

While addressing encoding complexity is driven by applications, two issues seem likely to dominate future theoretical investigations of LDPC codes. The first of these is to characterize the performance of LDPC codes with ML decoding and thus to assess how much loss in performance is due to the structure of the codes, and how much is due to the suboptimum iterative decoding algorithm. The second, and related, issue is to rigorously deal with the decoding of codes on graphs with cycles. Most analyses to date have assumed that the graphs are effectively cycle-free. What is not yet fully understood is just why the sum–product decoder performs as well as it does with LDPC codes having cycles.

BIOGRAPHIES

Sarah J. Johnson was born in 1977, and received the B.E. degree in electrical engineering in 2000 (Hons I and University Medal) from the University of Newcastle, Australia. She is presently a candidate for the Ph.D. degree in electrical engineering at the University of Newcastle, Australia, where her research interests include low-density parity-check codes, and iterative decoding algorithms.

Steven R. Weller was born in Sydney, Australia, in 1965. He received the B.E. (Hons I) degree in computer engineering in 1988, the M.E. degree in electrical engineering in 1992, and the Ph.D. degree in electrical engineering in 1994, all from the University of Newcastle, Australia. From April 1994 to July 1997 he was a lecturer in the Department of Electrical and Electronic Engineering at the University of Melbourne, Australia, and was a member of the Centre for Sensor Signal and Information Processing (CSSIP). Since July 1997 he has been at the University of Newcastle, Australia, where he is currently a Senior Lecturer in the School of Electrical Engineering and Computer Science, and a member of the Centre for Integrated Dynamics and Control (CIDAC). His research interests include low-density parity-check codes, iterative decoding algorithms, space time-coded communications, and combinatorics.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (July-Oct. 1948).
2. R. G. Gallager, Low-density parity-check codes, *IRE Trans. Inform. Theory* **IT-8**(1): 21–28 (Jan. 1962).
3. R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. IEEE Int. Conf. Communications* (*ICC'93*), Geneva, Switzerland, May 1993, pp. 1064–1070.
5. M. P. C. Fossorier, Iterative reliability-based decoding of low-density parity check codes, *IEEE J. Select. Areas Commun.* **19**(5): 908–917 (May 2001).
6. R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **IT-27**(5): 533–547 (Sept. 1981).
7. D. J. C. MacKay and R. M. Neal, Near Shannon limit performance of low density parity check codes, *Electron. Lett.* **32**(18): 1645–1646 (March 1996); reprinted in *Electron. Lett.* **33**(6): 457–458 (March 1997).
8. D. J. C. MacKay, Good error-correcting codes based on very sparse matrices, *IEEE Trans. Inform. Theory* **45**(2): 399–431 (March 1999).
9. T. Etzion, A. Trachtenberg, and A. Vardy, Which codes have cycle-free Tanner graphs? *IEEE Trans. Inform. Theory* **45**(6): 2173–2181 (Sept. 1999).
10. *IEEE Trans. Inform. Theory* (Special issue on Codes on Graphs and Iterative Algorithms) **47**(2) (Feb. 2001).
11. C. Di etal, Finite-length analysis of low-density parity-check codes on the binary erasure channel, *IEEE Trans. Inform. Theory* **48**(6): 1570–1579 (June 2002).
12. M. Sipser and D. A. Spielman, Expander codes, *IEEE Trans. Inform. Theory* **42**(6): 1710–1722 (Nov. 1996).
13. M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, Efficient erasure correcting codes, *IEEE Trans. Inform. Theory* **47**(2): 569–584 (Feb. 2001).
14. M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, Improved low-density parity-check codes using irregular graphs, *IEEE Trans. Inform. Theory* **47**(2): 585–598 (Feb. 2001).
15. T. J. Richardson and R. L. Urbanke, The capacity of low-density parity-check codes under message-passing decoding, *IEEE Trans. Inform. Theory* **47**(2): 599–618 (Feb. 2001).
16. S.-Y. Chung, G. D. Forney, Jr., T. J. Richardson, and R. Urbanke, On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit, *IEEE Commun. Lett.* **5**(2): 58–60 (Feb. 2001).
17. R. Lucas, M. P. C. Fossorier, Y. Kou, and S. Lin, Iterative decoding of one-step majority logic decodable codes based on belief propagation, *IEEE Trans. Commun.* **48**(6): 931–937 (June 2000).
18. P. O. Vontobel and R. M. Tanner, Construction of codes based on finite generalized quadrangles for iterative decoding, *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 24–29, 2001, p. 223.
19. S. J. Johnson and S. R. Weller, Codes for iterative decoding from partial geometries, *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, June 30–July 5, 2002, p. 310.

20. C. Di, T. J. Richardson, and R. L. Urbanke, Weight distributions: How deviant can you be? *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, June 24–29, 2001, p. 50.

21. N. Wiberg, *Codes and Decoding on General Graphs*, Ph.D. thesis, Dept. Electrical Engineering, Linköping Univ., Sweden, 1996.

22. F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, Factor graphs and the sum-product algorithm, *IEEE Trans. Inform. Theory* **47**(2): 498–519 (Feb. 2001).

23. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**(2): 284–287 (March 1974).

24. A. J. Blanksby and C. J. Howland, A 690-mW 1-Gb/s 1024-b, rate-1/2 low-density parity-check code decoder, *IEEE J. Solid-State Circuits* **37**(3): 404–412 (March 2002).

25. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo codes, *IEEE Trans. Commun.* **44**(10): 1261–1271 (Oct. 1996).

26. 3rd Generation Partnership Project (3GPP), *Technical Specification Group Radio Access Network; Multiplexing and Channel Coding* (*FDD*), 3GPP TS 25.212 V4.0.0 (2000-12) (online), *http://www.3gpp.org*.

27. T. J. Richardson and R. L. Urbanke, Efficient encoding of low-density parity-check codes, *IEEE Trans. Inform. Theory* **47**(2): 638–656 (Feb. 2001).

28. Y. Kou, S. Lin, and M. P. C. Fossorier, Low-density parity-check codes based on finite geometries: A rediscovery and new results, *IEEE Trans. Inform. Theory* **47**(7): 2711–2736 (Nov. 2001).

M

MAGNETIC STORAGE SYSTEMS

HEMANT K. THAPAR
LSI Logic Corporation
San Jose, California

1. INTRODUCTION

Data storage is an essential function within today's information delivery systems. While the communication of information focuses on the delivery "from here to there," the storage function is focused on the delivery of information "from now to then." The "now to then" may entail near real-time transactions, as in server-client systems, or a period of days or months, as in data archiving. Whatever the application, the demand for storage is exploding in computing, communication, consumer, and entertainment systems. Over 200 million magnetic hard disk drives and over 150 million optical drives, combining various forms of compact disk (CD) and digital video disk (DVD) drives, will be shipped worldwide in 2002. The market for magnetic hard disk drives alone is forecasted to grow to over 350 million drives by 2006. Using a conservative estimate of 100 Gbytes of average storage capacity per drive in that time frame, digital magnetic storage alone will support 35,000 petabytes (35×10^{18} bytes) of storage demand worldwide.

Storage devices can be broadly classified into two categories: solid-state and mechanical. Solid-state memories are based on semiconductor process technology, and they can be used as standalone components within a system, or integrated with other functions in monolithic form. Mechanical storage devices rely on the relative motion between a magnetic, optical, or hybrid (magneto-optic) transducer and an associated storage medium to store temporal signals as spatial patterns on the medium. They are complex standalone subsystems that are used to store large quantities of data in nonvolatile form; that is, the device power can be turned off while preserving the stored data. Solid-state memories consume less power during storage and retrieval of data and offer better mechanical reliability, but they are considerably more expensive than mechanical storage devices.

The trend in storage devices is similar to that for communication systems: digital storage is emerging as the technology of choice. While data storage is inherently digital, storage of ubiquitous analog sources of information, namely, audio and video, is being accomplished increasingly using digital techniques. Digital source coding methods, such as MPEG-x and its associated audio standard MP3, JPEG, and pulse code modulation (PCM), are used to convert the analog signals to digital bit sequences for the purposes of delivery and storage. The advantages of performance, cost, and flexibility are driving this trend. Digital storage offers the ability to maintain a low probability of error during repeated retrievals of the stored data. This advantage is similar to "regeneration" in digital communications. Sources with wide dynamic range signals, such as classical music, can be reproduced with low noise and distortion compared to analog storage. Digital audio tape (DAT), music CD, DVD, and PVR (personal video recorder) as a replacement for VCR are examples of the emerging trend to use digital storage. The cost of digital storage also continues to decline at a rapid rate because of the steady improvements in component technologies, as discussed later in the article, and the economies of scale associated with the mass personal computer market.

In order to meet the wide and varying demands of different applications, storage devices have become segmented on the basis of two factors: performance and cost per megabyte. Performance is measured in terms of the access time, which is defined as the average time spent to access the selected data. Figure 1 shows the segmentation of commonly used storage devices in various applications. Highly cost-sensitive applications, involving software distribution, consumer audio and video playback, use compact disk read only memory (CD-ROM) and DVD devices. These devices are based on the use of optical recording technology and are designed to support varying modes of storage, including read-only, write-once read-many (WORM), and erasable/rewriteable. Their access times typically are on the order of tens to hundreds of milliseconds, but the associated cost of the drive and media is very low. Data backup and archival storage rely largely on the lower cost, lower performing magnetic tape systems. Such systems have access times on the order of tens to hundreds of seconds, since they can only access the selected data sequentially. They, however, support very large volumetric densities (Mbytes/cu. ft.) at very low cost per megabyte. Near real-time transactional systems use the higher cost, higher performance hard disk drive (HDD) systems permitting direct access to the selected data in any arbitrary order. Such devices are based on the use of magnetic recording technology. Solid-state memory devices, with access times on the order of nanoseconds but much higher cost per megabyte, are used for real-time storage applications, such as caching, real-time data memory, and program control.

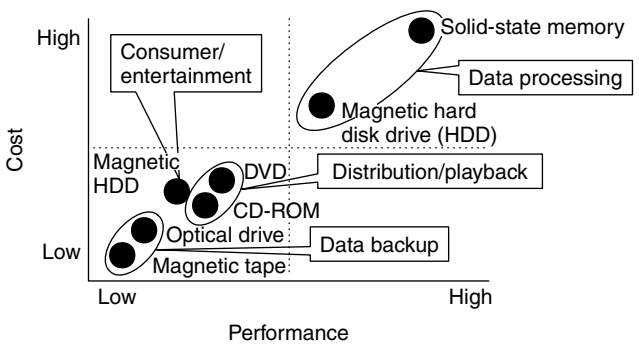


Figure 1. Storage devices segmentation.

Among mechanical storage devices, the magnetic HDD system has occupied a unique position in data storage since the introduction of IBM's RAMAC system in 1957 (see [15]). In order to keep this position unchallenged, the underlying magnetic component, signal processing, mechanical, and interface technologies have evolved at a very rapid pace to meet the growing demands of high-performance computing and peripheral devices. As a consequence of the rapid progress, digital magnetic storage technology is well poised to support the lower cost, lower performance segment previously serviced by optical or magnetic tape devices.[1] Magnetic HDD has already found its way in today's set-top boxes, allowing users to store in excess of 100 Gbytes of their favorite TV programs unattended. More applications in this area, dubbed "personal video recording," are likely to emerge as the much-vaunted convergence of communications, computing, entertainment, and mobility picks up pace. Similarly, IBM's one-inch diameter HDD is capable of storing 1 Gbyte of data in nonvolatile form at low cost. It is finding its way in digital cameras and other mobile applications. Except for removeability, digital magnetic recording offers everything that optical storage does, but with system attributes that include smaller, cheaper, denser, and faster.

This article provides an overview of digital magnetic storage based on the HDD system. Section 2 provides an overview of digital magnetic storage in HDD systems and the associated technology trends. Section 3 describes the digital magnetic recording channel, including the magnetic recording processes that underlie the generation of signals, noise, distortion, and interference during data retrieval. Section 4 describes the signal processing and coding techniques used in commercial HDD systems. Many of those techniques have their origins in data transmission and can be applied to other digital magnetic and optical storage channels with suitable modifications. Section 5 is devoted to concluding remarks.

2. HDD SYSTEM AND TRENDS

HDD systems are designed to deliver digital information "from now to then." Two processes are involved in such delivery: recording and retrieval.[2] The recording function, often referred to as "write," takes blocks of data (typically 4 kbits), appends control and synchronization information, and records it in the form of data sectors on the medium. The retrieval function, referred to as "read," processes the readback signal from the medium to deliver the recovered data. The two processes are similar to the transmit and the receive functions in data communication systems. While there are many similarities between data storage and communication, there are key differences that pertain to the error rate and synchronization requirements, which, in turn, have a bearing on the overall system design philosophy. Unlike data communications, HDD systems do not rely on "automatic request for retransmission" to recover from data errors; the retrieval process is designed to guarantee a prescribed worst-case bit error rate.[3] Similarly, since data recording involves mapping temporal data into spatial patterns, clock and data synchronization during retrieval must be fast and reliable to conserve the "real estate" on the medium. A great deal of effort is focused in HDD systems to minimize spatial overhead.

The HDD is a highly sophisticated electromechanical system. The mechanical assembly involves a slider mechanism holding a read/write head that flies about 10–20 nanometers (nm) over a rotating disk with speeds ranging from 3,600 rpm in the 1″ form factor (which refers to the disk diameter) HDD to 15,000 rpm in the 3.5″ form factor HDD. This head/media spacing places very stringent requirements on the disk surface in terms of uniformity, planarity, and defects. At the time of this writing, a typical drive will have 1 to 7 disks and 1 to 14 heads.

As shown in Fig. 2, data is stored on the disk in the form of spatial magnetic patterns along a *track*. The tracks are laid out as annuli of width W_t, which comprises the physical magnetic track width and a guard band between neighboring tracks. During data recording or retrieval, the head is positioned at a new track by moving the slider using servo control. To achieve accurate positioning, prewritten servo data patterns are interspersed along the tracks in the form of wedges, as shown in Fig. 2. These patterns are sensed during the head positioning process, which takes place in two major steps. First, the head seeks the track

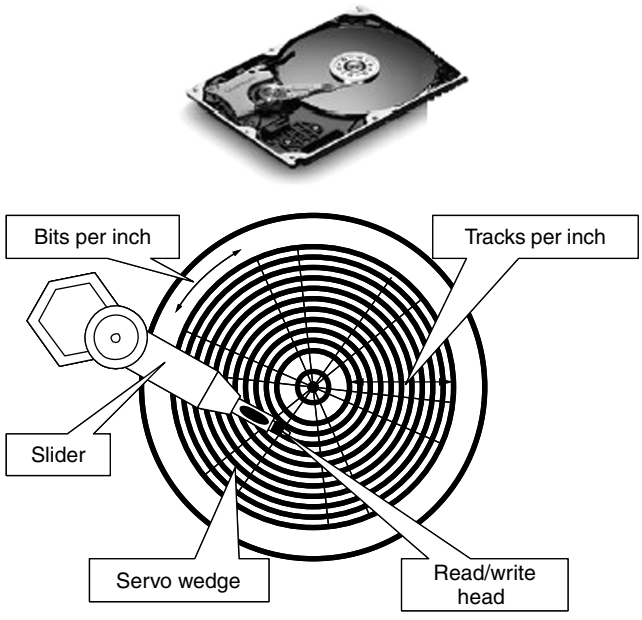


Figure 2. Data storage on hard disk drive.

[1] Even though Fig. 1 shows the cost for magnetic tape storage to be lower than HDD, the cost per megabyte for many HDD devices is comparable to that for tape systems.

[2] Accurate head positioning during recording and retrieval may be regarded as the third process for information delivery. It is very key to the operation of the HDD.

[3] Procedures that rely on the reread of the recorded data are also built into the HDD to recover from a rare error event, but their probability of use is minimized by design to maintain a high data throughput rate.

where the target data is to be located. The average amount of time required to seek the targeted track is called the *seek time*. The second step is to locate the data on the targeted track. The average time spent to locate the data is referred to as *latency*, which equals half the revolution period of the disk, since the target data location, on average, is halfway along the track from the initially positioned head. The sum of the seek time and latency defines the *access time*, which denotes the average time spent in going from one randomly selected location to another. Access time is a key measure of performance in data storage applications. Improvements in servo control algorithms, actuator design using lighter materials, and higher rotational speeds (e.g., 15000 rpm) are progressively reducing the access time in HDD. Today's products have access times ranging from 15 ms in IBM's 1″ Microdrive to below 6 ms in high performance server drives.

The number of bits stored per unit length of the track is referred to as *linear density*, measured in bits per inch (bpi). If the rotating speed of the disk is M revolutions per second, the data rate is R bits per second, and the track location is at radius r inches, then the linear density L is given by

$$L = \frac{R}{2\pi rM} \text{ bits/inch} \tag{1}$$

Note that the linear density grows towards infinity as r approaches 0. In practice, a nonzero inner radius, r_i, is selected to achieve a prescribed maximum linear density. Likewise, a prescribed outer radius, r_o, is used and the data storage is confined to the region between r_o and r_i. Based on capacity considerations (see Eq. (3) below), r_o is approximately equal to $2r_i$.

The number of tracks per unit length along the radial direction is referred to as the *track density*, measured in tracks per inch (tpi), and is given by

$$N_t = \frac{r_o - r_i}{W_t} \text{ tracks/inch} \tag{2}$$

The linear density is typically 8 to 15 times higher than the track density in commercially available products. The product of linear and track density defines the *areal density*, measured in bits per sq. inch. The storage capacity of a disk surface is the product of the areal density and the total surface area available for recording. Based on simple physical arguments, the capacity per surface, C, can be bounded as:

$$\frac{2\pi r_i(r_o - r_i)}{lW_t} \le C \le \frac{\pi({r_o}^2 - {r_i}^2)}{lW_t} \tag{3}$$

where W_t is the track width and l is the smallest bit cell length along the track. The upper bound assumes that bit cells of area lW_t are recorded over the entire disk surface.[4] Such a bound would be achieved if the linear velocity of the disk could be kept constant across all tracks. Since the linear velocity is radius-dependent (note $v = r\omega$, where ω is the angular velocity and r is the radius), it is impractical to keep it constant while supporting random access with low access time.[5] The lower bound is based on the use of constant data rate and rotational speed, wherein the number of bit cells at radius r_i, given by $2\pi r_i/l$, is kept constant across all tracks.

In practice, the capacity per surface lies between the two bounds. Instead of varying the rotational speed, the data rate is varied across the radii to effect better utilization of the disk surface. The disk surface is delineated into annular zones and the data rate is increased across the zones from the inner radius to the outer radius. This allows the zones along the outer radii to store more data, and hence yield higher storage capacity. The actual increase in capacity achieved from this so-called *zone-bit recording* scheme depends upon the number of zones and the linear density in each zone.

Areal density growth is key to increasing the capacity per disk surface. Indeed, the areal density has grown by a factor of 17 million since the introduction of the RAMAC drive in 1957 (see [13,14]). Figure 3 shows the areal density trends for commercial products and prototype demonstrations. At least two inflection points have occurred in the past decade. Since 1991, the rate of increase in areal density accelerated to 60% per year, and since 1997 this rate has further increased to 100% per year. Today's commercially available products store in excess of 50 Gbits/sq. inch, combining 80 ktpi in track density and 670 kbpi in linear density. Experimental prototype demonstrations exceeding 100 Gbits/sq. inch have been reported in the industry. The acceleration in 1991 of the annual growth rate was caused by the introduction of two key component technologies: magnetoresistive (MR) sensor for read heads and partial response maximum-likelihood (*PRML*) for read channels. The application of coding and PRML are discussed in more detail later in the chapter. The inflection point in 1997 was caused by the introduction of Giant MR (GMR) heads, which provide improved transducer sensitivity over their predecessors. Continual improvements in magnetic medium and signal processing technologies are also supporting this unprecedented growth rate in areal density.

The incredible growth in areal density has wrought similar trends in other figures-of-merit of interest in storage applications. Most importantly, the cost per megabyte decreases since the number of heads and disks required to achieve a prescribed capacity point decreases. Indeed, the cost per megabyte is declining at a rate of 40–50% annually, a rate currently higher than that for DRAM. Volumetric density also grows since smaller form factor disk drives,[6] with reduced head/disk count,

[4] In practice, the entire disk surface is not available for data storage. Some area is used to store servo data patterns for controlling the head positioning over the track as well as for storing overhead information related to defect skipping, calibration, etc.

[5] The linear velocity is kept constant across all tracks in compact audio players because the information is accessed sequentially from the inner radius to the outer radius and enough time is available to vary the rotational speed continually across the tracks.

[6] The 5.25 inches and larger form factors are all but obsolete now.

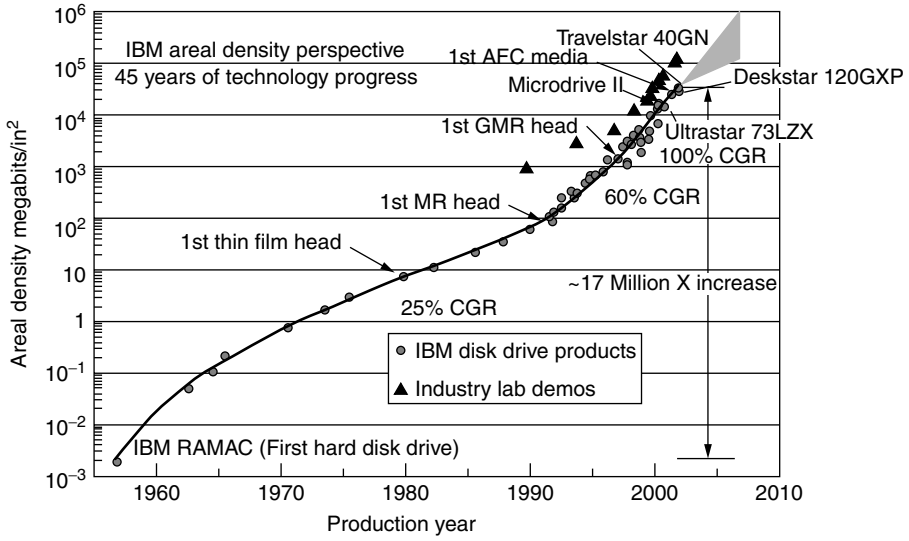


Figure 3. Area density trend in digital magnetic HDD.

achieve prescribed capacity points. The 3.5 inches and 2.5 inches form factors, which refer to the disk diameters, are mainstream today; the 1 inch is the emerging form factor. At the time of this writing, the 3.5 inches drive (with 1″ height), serving the high performance server market, stores approximately 80 Gbytes; the same form factor drive for desktop applications has a capacity exceeding 100 Gbytes. The 2.5 inches drive (with 1/2″ height) serving the mobile and notebook computers stores approximately 50 Gbytes. The 1.0 inch form factor (1/4″ height) stores 1 Gbyte. These same form factors are likely to double their storage capacity within a year, or support the same capacity with reduced number of heads and disks, and thus lowered cost per megabyte.

With smaller form factor and fewer disks, higher rotational speed and improved mechanical assembly can be achieved, thereby providing the means to reducing the access time. As the linear density grows with the areal density, the data transfer rate also increases (see Eq. (1)). The internal data transfer rates in today's disk drives is in the range of 400 Mbits/sec to over 1 Gbits/sec. It is growing at 30–40% annually. Together, the increasing transfer rates and decreasing access time are rendering HDD systems faster than before.

The above trends point to the following observation: magnetic disk drives are unequivocally becoming smaller, denser, faster, and cheaper. With these attributes, magnetic HDD is likely to become a viable storage device for such consumer applications as digital cameras, mobile communication devices, handheld computers, personal video recorders, and set-top boxes.

However, as the capacity per surface grows exponentially due to the increasing areal density, issues of reliability of the storage device become more acute. This trend has led to the development and proliferation redundant array of independent disks of (RAID) systems, which use redundancy within an array of disk drives to improve reliability and performance of the storage system. Tens of terabytes are aggregated in a RAID device with prescribed measures of data availability and reliability. Just as in error correction coding schemes, RAID devices are designed to reconstruct the stored data in the midst of a prescribed number of drive failures. Interconnected through data networks, these devices are used today to service the storage demands of the Internet and other information delivery systems with uncompromised availability.

The reader is referred to [13–15,28,30] for more detailed information on the trends for HDD systems.

3. DIGITAL MAGNETIC RECORDING CHANNEL

Digital magnetic recording is based on the elementary principles of electromagnetics wherein the temporal data signal to be recorded is converted into spatial patterns of magnets on a magnetic medium. It relies on the well-known *M-H* curve, shown in Fig. 4, which defines the switching behavior of the applied magnetic field, H, and the resulting remanent magnetization, M, of a magnetic material. Figure 4 shows that when the applied magnetic field exceeds the coercivity of the medium, H_c, the medium has remanent magnetization M_r. Likewise, when the field is reversed, the state of the magnetization is also reversed. Thus, the medium can be saturated into two

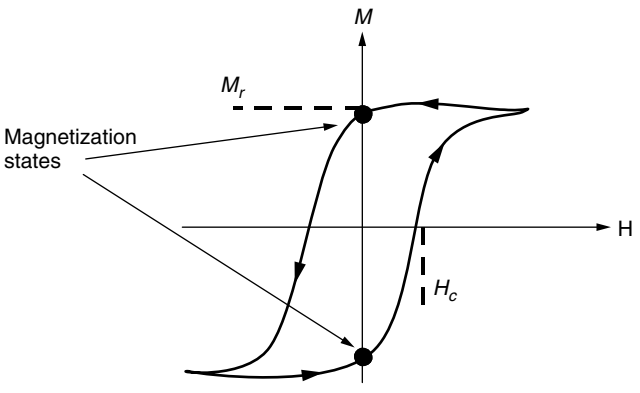


Figure 4. The *M-H* curve.

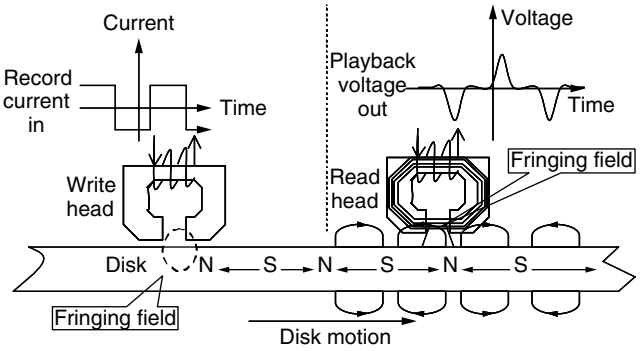


Figure 5. The digital magnetic recording and retrieval processes.

states, corresponding to binary 1s and 0s, by controlling the magnitude and direction of the applied field.

The recording and retrieval processes are illustrated in more detail in Fig. 5. As shown on the left side, the data sequence to be recorded is represented as a binary current waveform, which is applied to the write head with a gap. The flow of current in the windings of the write head generates a magnetic field within the head. The presence of the gap causes the field to fringe into a "bubble" and penetrate the magnetic recording medium. The field emanating from the gap can be decomposed into the longitudinal component (along the medium) and a perpendicular component. When the longitudinal component exceeds the coercivity of the medium, magnetic domains are created in accordance with field changes due to the current waveform. These domains are delineated in the figure by the symbols N and S, representing, respectively, the north and south poles of a magnet. Thus, temporal changes in the write current are mapped into spatial changes in the magnetic medium. Such a recording process is referred to as *longitudinal* saturation *recording*. With a different head-medium construction, one can use the perpendicular component of the head field to record the digital data. Such an approach, referred to as perpendicular recording, is regarded as a potential successor to longitudinal recording beyond 100 Gbits/sq. inch (see [28]).

During retrieval, the spatial magnetic patterns on the moving medium, represented by the alternating magnetic poles N and S, create fringing magnetic fields that are sensed by the read head, producing an analog voltage waveform that varies in accordance with the recorded patterns. Readback magnetic transducers are either inductive or magnetoresistive (MR). The inductive sensor produces a readback voltage that is proportional to the time derivative of the flux from the fringing field. The signal amplitude depends upon the rotational speed of the medium and the number of turns, N, in the winding of the inductive head [$V = -N(d\phi/dt)$]. MR heads use the MR stripe (or MR element), placed between two shields, to sense the flux from the external field. The change in flux causes a change in the resistance of the current-biased MR stripe, resulting in an output voltage that varies in accordance with the recorded magnetic patterns. The readback signal amplitude, unlike inductive heads, is independent of the rotational speed of the medium. Read heads based on MR-based sensors are ubiquitous in today's hard disk drive systems and remain major contributors to the accelerated areal density growth.

The preceding description is intended to capture the essence of the recording process. At the detailed level, the recording/retrieval processes are quite complex, especially as the bit cells continue to shrink. For example, the write current has finite rise times; the head fields do not switch instantaneously; magnetic transducers are frequency selective and often nonlinear; magnetic fields interact as transitions get closer due to the growth in linear density; and so on. All such factors cause nonidealities in the recording and replay processes, requiring sophisticated analysis, modeling, and experimental work to understand the signals, noise, interference, and distortion mechanisms. The reader is referred to Refs. 1, 4, 5, 30 for a more detailed treatment.

Based on the above description of the recording process, the recording channel characteristics are discussed below. The "channel" refers to the combined head and medium block that is responsible for generating the signals, noise, distortion, and interference mechanisms. Unlike communications channels, where the bandwidth and noise characteristics typically remain fixed after the spectral allocations are made, the digital magnetic recording channels continue to evolve and change. Signal and noise bandwidths get larger with the scaling of head/medium dimensions, and new noise phenomena arise as the bit cell dimensions shrink and new magnetic materials and transducers are introduced. The dynamic nature of the channel makes every new generation of HDD development more interesting, particularly from the viewpoint of deploying modern modulation and coding techniques.

The digital magnetic recording channel is inherently nonlinear because of the *M-H* hysteresis loop. That is, scaling the input current does not proportionately scale the output voltage since the remnant magnetization does not change appreciably for a large increase in the applied field. However, for a fixed write current that is sufficiently large to saturate the magnetic medium, the output signal can be constructed as a linear combination of the response due to the individual inputs symbols. Thus, in a limited regime of operation, the write process is nonlinear, but the readback process is linear. The output (readback) voltage waveform can be written as a pulse amplitude modulated (PAM) signal:

$$r(t) = \sum_k a_k h(t - kT) + \nu(t) \tag{4}$$

where $h(t)$ is the unit pulse response of the head/medium, $v(t)$ is the noise, and a_k is the sequence of input symbols forming the write current. Note that $a_k \in \{1, -1\}$. Using linearity once again, the unit pulse response can be written in terms of the more elementary response $s(t)$, called the transition response, as:

$$h(t) = s(t) - s(t - T) \tag{5}$$

where $1/T$ is the clock rate of the input sequence into the head/medium. The transition response corresponds to the head/medium response to a unit step change in

the write current polarity. Since the digital magnetic recording channel is peak amplitude limited, the peak of the transition response represents the maximum value of the output signal. The average power of the channel output for random data depends upon the operating density.

The readback signal can also be written in terms of the transition response as follows:

$$r(t) = \sum_k b_k s(t - kT) + \nu(t) \tag{6}$$

where $b_k = (a_k - a_{k-1})$ is the sequence of data transitions. Since a_k is binary, b_k is ternary ($b_k \in \{2, 0, -2\}$), where $b_k = -2$ denotes a change in write current polarity from positive to negative, $b_k = 0$ denotes no change, and $b_k = 2$ denotes a change from negative to positive. Note that successive nonzero data transitions alternate in polarity. Based on analytic results, the step response in digital magnetic recording channels is commonly modeled by a Lorentzian pulse given by

$$s_L(t) = \frac{A_L T}{\pi t_{50}} \frac{1}{1 + \left(\frac{2t}{t_{50}}\right)^2} = \frac{A_L}{\pi \delta} \frac{1}{1 + \left(\frac{2t}{\delta T}\right)^2} \tag{7}$$

where t_{50} is the width of the step response at half its maximum amplitude, A_L is the peak amplitude scaling factor, and δ is the *normalized linear density*, defined as:

$$\delta = \frac{t_{50}}{T} \tag{8}$$

The parameter t_{50} measures the temporal dispersion of the step response.[7] The model of Eq. (7) assumes that the head gap and the head/medium spacing are zero. More elaborate models are also available (see [5]), but the single-parameter Lorentzian pulse model is adequate for investigating the relative performance of different signal processing and coding schemes. The normalized linear density measures the number of bit cells that are packed per t_{50}, the half-amplitude-pulse-width, and is used as the parameter for comparing different detection methods. Today's HDD products have values of δ ranging from 2.3 to 3.0. Even with the application of zoned recording, the normalized linear density varies from the inner radius to the outer radius.

Figure 6 shows the Lorentzian transition response and Fig. 7 shows the corresponding pulse responses for varying values of δ. Both responses are symmetrical, and thus they have linear phase characteristics. The intersymbol interference (ISI) causes the amplitude and the energy of the pulse response to decrease as the linear density is increased.

In the frequency-domain, the amplitude spectra of the Lorentzian pulse and transition responses are given by:

$$H_L(\Omega) = j2TA_L \sin(\pi\Omega) \exp(-\pi\delta|\Omega|) \tag{9}$$

[7] When measured spatially, the spatial dispersion $pw_{50} = vt_{50}$ where v is the linear velocity of the disk. Many reported publications define the normalized linear density as pw_{50}/T.

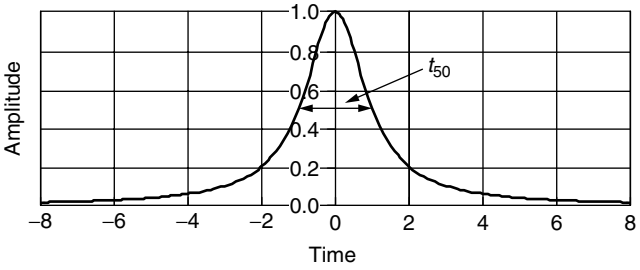


Figure 6. Lorentzian transition response.

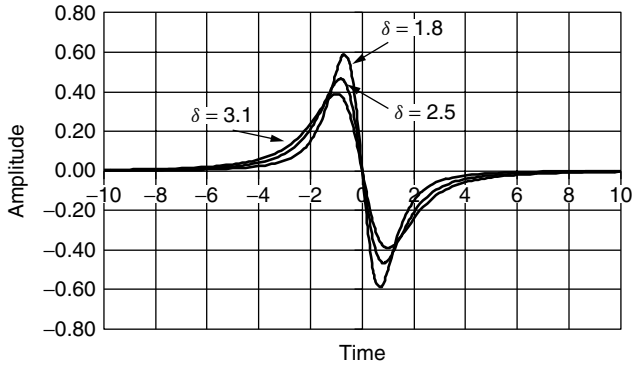


Figure 7. Lorentzian pulse response for varying linear density.

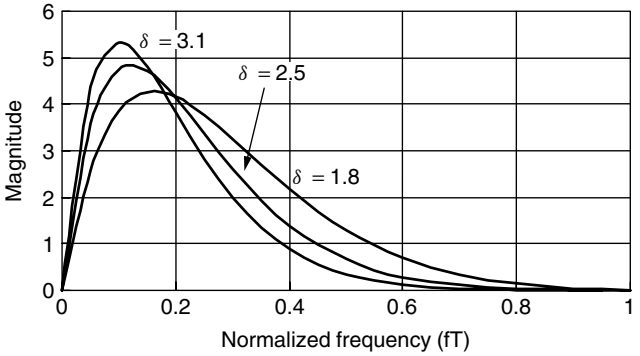


Figure 8. Amplitude spectra of lorentzian pulse.

and

$$S_L(\Omega) = A_L \exp(-\pi\delta|\Omega|) \tag{10}$$

where $\Omega = fT$ is the normalized frequency. Figure 8 shows the amplitude spectra of the Lorentzian pulse for different values of δ. The high-frequency content of the signal spectrum becomes increasingly attenuated because of the increased ISI at higher linear densities. Note that the amplitude spectrum extends beyond the Nyquist frequency ($f = 1/2T$), thereby requiring special consideration of sampling phase selection in symbol-spaced finite impulse response (FIR) equalizers. The phase spectrum of the pulse response is linear.

In summary, from the signal perspective, the digital magnetic recording channel is band-limited with a peak amplitude constraint, instead of the average power constraint generally associated with most communications

channels. Severe ISI occurs as the linear density is increased for a given recording channel, resulting in loss of pulse energy. The readback signal is corrupted by channel impairments, some of the major ones of which are outlined below. For a more detailed and exhaustive treatment, please refer to Refs. 1, 5.

In the digital magnetic recording channel, the *noise* sources include the following:

Media noise depends on the type of media. In particulate media, the noise is due to statistical distribution of the magnetic particles. It is modeled as additive, Gaussian, stationary, and with power spectrum similar to that of the signal. In thin film media, which are ubiquitous in today's disk drives, the noise is due to the randomness in the width of the recorded transitions. Figure 9 illustrates the source of this noise. As shown, the recorded transitions are far from being a straight line. Instead, they exhibit a zig-zag microstructure with a shape that varies randomly with each transition. The nominal transition response and its location then depend on the average width, w, and the average center of the recorded transitions. The statistical variation from these nominal values constitutes media noise. It is, in general, data-dependent and neither stationary nor additive. However, under some simplifying yet realistic assumptions, it can be modeled as additive and stationary (see Appendix 2C in [4]). *Head Noise* also depends on the head type. In MR heads, it is due primarily to the thermal resistances within the MR element and its contacts, and is, therefore, modeled as additive, white, and Gaussian. *Preamplifier noise* is added to the readback signal during its amplification. It is due to the electronic circuits in the signal path, and is also modeled as additive, stationary, Gaussian, and largely white.[8]

The above noise sources are mutually uncorrelated and their relative mix depends on the data rate and the magnetic and mechanical parameters of the head/media interface. For present-day systems, the media noise power is 1–4 dB higher than that of the electronics noise, which increases from the inner radii to the outer radii as the data rate is increased in zone-bit recording.

Inter-track interference (crosstalk) is caused by the pick up of signals from adjacent tracks as the head moves offtrack during the retrieval process. These interference signals are mitigated with accurate servo positioning, which attempts to keep the read head in the center of the track with very high probability. Further mitigation of interference signals is achieved by making the inductive write head wider than the MR read head. This so-called write-wide, read-narrow concept effectively introduces an additional guard band between tracks. The adjacent track interfering signal is, of course, a filtered version of the recorded data, and hence neither stationary nor Gaussian.

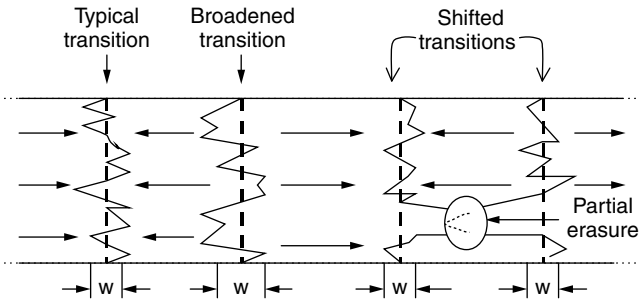


Figure 9. Recorded transitions in thin film media.

In addition to noise and interference, the readback signal may be corrupted by *distortion*, which may be linear or nonlinear, and generally grows as the recording density is increased. Some of the sources of nonlinear distortion are outlined below:

Nonlinear distortion in the form of *transition shift* occurs when the recorded transition is shifted from its intended location. Such shifts may occur when adjacent transitions get too close and bandwidth limitations in the write path, resulting in inadequate rise times of the write current or of the flux in the head gap, cause the transitions to move. Similarly, since successive transitions alternate in polarity, the magnetic field from the preceding transition[9] can interact with that of the new transition to aid its recording, thereby shifting the new transition earlier than intended. The amount of shift depends heavily on the operating conditions of the head/medium, and decreases rapidly as the minimum transition spacing increases. Such transition shift phenomenon is typically limited to transitions which are one symbol duration apart, but it can extend to two or more symbol durations if the linear density is very high. In practice, this nonlinear shift is virtually removed by: (1) ensuring adequate rise times in the write path, and (2) using precompensation during data recording wherein selective transitions in the write current are "delayed" by a prescribed amount to offset the subsequent "shift-early" effect.

The above transition shift is due to field interactions involving the new data sequence being recorded; similar shifts can occur because of residual fields from previous recordings. Because there is typically no dedicated erasure cycle in magnetic recording, the field from previously recorded data, especially those associated with low-frequency patterns, can "impede" or "aid" the recording of the new transition, causing it to shift. This effect is mitigated through careful design of the head/media parameters to achieve a prescribed "overwrite" signal-to-noise ratio, which guarantees a prescribed power ratio between the new readback signal and that from a previously recorded pattern.

Nonlinear distortion can occur with MR heads during readback. In single stripe MR head configurations,[10] which are widely deployed today, the stripe is biased to achieve a linear transfer characteristic between the change in flux and the associated change in MR resistance. Because of tolerances in MR stripe and the bias point, perfect linearity

[8] Some roll-off in the noise spectrum may occur at the upper and lower edges of the readback signal spectrum depending upon the amplifier design.

[9] This field is often referred to as demagnetizing field, which essentially has the effect of lowering the coercivity.

[10] Single stripe MR heads produce single-ended readback voltage signal; with dual stripe MR head, differential combining may be done to circumvent this effect.

is not achieved and some amount of memoryless, quadratic nonlinearity is introduced. The resulting signal has pulse asymmetry wherein the negative and the positive pulses may have different heights or widths, or both. This effect may be compensated by adaptively canceling the quadratic term before detection.

Nonlinear intersymbol interference can also occur in thin film media as recorded transitions get too close to each other. As noted earlier, the microstructure of the recorded transition in thin film media has a zig-zag signature (see Fig. 9). At very high linear density, portions of successive zig-zags may merge, causing the transitions to "weaken" and the readback signal amplitude to become smaller than that predicted by the linear model. This effect is referred to as *partial erasure*. It can be mitigated using precompensation during data recording wherein write current transitions separated by one symbol duration are moved away from each other by some prescribed amount.

In addition to the above nonlinear distortions, transient disturbances due to imperfections of the media may distort the readback signal. Media defects can cause "dropouts" in the readback signal amplitude. Such defects are screened during surface analysis of the media at the time of manufacturing of the disk drive. The defective sectors are precluded from storing information by the drive controller. Another form of distortion, referred to as *thermal asperity*, occurs when the single-stripe MR head bumps against a high spot on the medium. A large voltage transient is created because of the heating of the MR element, causing the small readback signal to modulate the transient signal. The transient decays exponentially as the MR element returns to its ambient temperature. This effect is mitigated during data retrieval by detecting the onset of the transient at the very earliest stage of the signal processing chain to avoid saturation of the subsequent blocks and loss of synchronization. Since the energy of the transient signal is located near the lower band-edge of the signal spectrum, the lower corner frequency of the receive filter is temporarily increased to filter out the ensuing transient. Such an approach is effective in limiting the span of the data errors to the correction capability of the error correcting code.

Channel identification based on the use of pseudo-random binary sequences has been developed to isolate the various nonlinear distortion effects outlined above. This method can be used in near real-time to define the precompensation parameters to linearize the channel (see [22]). As discussed in the next section, the signal processing methods deployed in magnetic recording assume the channel to be linear.

Linear distortion in the form of intersymbol interference (ISI) is by far the major contributor of distortion at high linear density, resulting in reduced energy and attenuated high-frequency content of the pulse response. The application of classical communication techniques to high density digital magnetic storage has been investigated over the past two decades, culminating in the development of many new coding and signal processing techniques that address the unique requirements of data storage (see [32,33]). Some of those techniques are discussed in the next section.

4. SIGNAL PROCESSING AND CODING METHODS

Signal processing and coding methods have played a vital role in digital magnetic recording systems, especially during the past decade as the bit cell dimensions have shrunk at an accelerated pace, requiring detection methods that are bandwidth and SNR efficient. Some of the methods used commercially are described in this section.

Figure 10 shows a block diagram of the data channel for digital magnetic recording. On the recording ("transmit") side, the data to be recorded are first encoded using an error-correction code (ECC), which is typically based on Reed-Solomon codes. The encoded output is applied to a modulation code and an associated precoder to achieve prescribed properties in the readback signal during data retrieval. The modulation code adds redundancy to improve signal detectability, but the precoder performs a one-to-one mapping on the encoded sequence. Different modulation codes have been used over the years along with different precoders, as discussed below. The encoded output is used to generate the write current waveform, which is then applied to the pre-compensation circuit to suitably time-shift the transitions associated with prescribed data patterns. As noted previously, symbol-spaced transitions are typically preshifted to linearize the recording channel.

On the retrieval side, the readback signal is amplified and then applied to the receiver, which is often referred to as "Read Channel" within the data storage community. The Read Channel is similar to a typical digital baseband communication receiver, comprising blocks that perform the functions of gain control, timing control, synchronization, receive filtering, equalization, detection, and decoding. In addition, compensation techniques may be incorporated to address other impairments, such as nonlinearity in MR heads and thermal asperity processing. This section will cover the evolution of the detection methods but not discuss important receiver functions like timing recovery, gain control, and synchronization. The reader is referred to Ref. 8 for a detailed treatment of those functions.

Figure 11 shows the evolution of the coding and detection methods in digital magnetic HDD. The upper legend in the box denotes the modulation code, and the lower legend denotes the detection method. The dashed box, denoting "Turbo Coded EPRML/GPRML," is not deployed commercially at the time of this writing. But, as

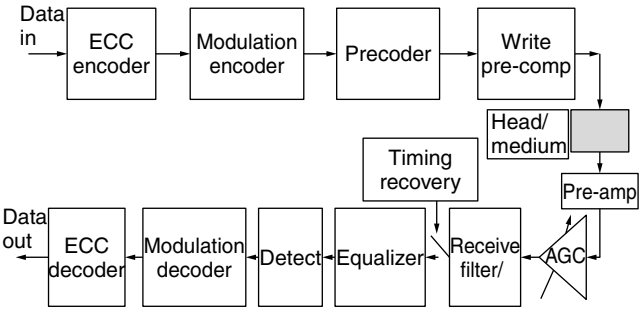


Figure 10. Magnetic data storage channel.

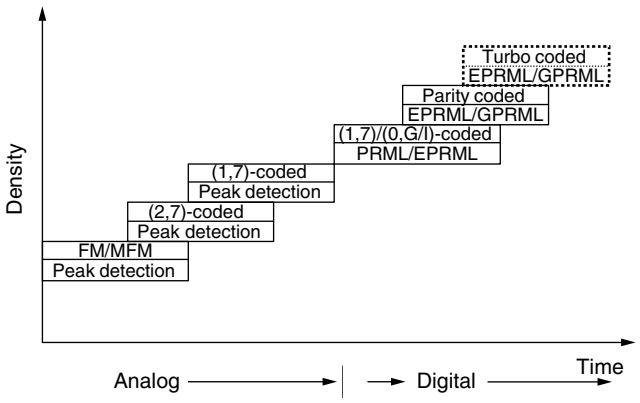


Figure 11. Evolution of coding and signal processing techniques.

with other communication channels, deployment of Turbo Coding in digital magnetic storage is of intense interest to researchers and practitioners alike, and simulation results (see [21]) to date show it to be highly effective in providing SNR benefits over other channel coding methods.

The Peak Detection method, relying on analog signal processing, was used ubiquitously in disk drives for over three decades to recover the recorded data from the analog readback signal corrupted by various impairments outlined in the previous section. It is based upon the simple observation that, in the absence of any inter symbol interference (ISI), the maximum (minimum) value of the transition response coincides with the location of the recorded transition, which, in turn, corresponds to a change in the polarity of the write current. Using this observation and the NRZI format[11] to represent the write current, the Peak Detection method detects the presence or absence of the signal peak within each symbol interval. A peak is considered to be present if the applied signal exceeds a prescribed threshold *and* its derivative has a zero crossing. Otherwise, the peak is considered to be absent. Thus, the presence of a peak denotes a "1" and the absence denotes a "0."

The performance of Peak Detection degrades rapidly in the presence of ISI because: (1) the signal peaks shift away from the location of the transitions, and (2) the peak amplitude associated with closest transitions decreases (see Fig. 7). The effect of ISI is mitigated with a class of modulation (line) codes called run-length limited (RLL) *codes*. These codes prescribe run-length constraints on recorded sequences to extend the applicability of Peak Detection. The run-length constraints are typically designated as (d, k), where d and k are nonnegative integers (with $d < k$) that denote, respectively, the minimum and the maximum number of "0s" between "1s" at the encoder output. For example, the (1,7) RLL code produces sequences that contain at least one "0" and at most seven "0s" between any pair of "1s." When combined with the NRZI format, wherein "1" produces a polarity change in the write current, these parameters determine the minimum $(= (d + 1))$ and the maximum $(= (k + 1))$ symbol intervals between recorded transitions. Since transitions produce signals with nonzero amplitudes, the k constraint acts to ensure that corrective updates are available at some minimal rate for the timing recovery and the automatic gain control loops. Similarly, the d constraint controls the separation between closest transitions, and thus the resulting ISI, if any. Together, the (d, k) pair defines the highest code rate, called the code capacity, which can be achieved for the prescribed constraints. The d constraint can be removed by setting $d = 0$; likewise, the k constraint can be removed by setting $k = \infty$. For a detailed description of RLL codes and their construction, refer to Ref. 19.

Early HDD systems were based on rate 1/2 codes, with the run-length constraints evolving from (0,1) for frequency modulation (FM) to (1,3) for modified-FM (MFM), to (2,7). By progressively increasing d, these codes achieved higher linear densities with peak detection without incurring a code rate penalty or performance degradation. The approach was, however, not extendible to $d = 3$, since such a constraint could not be achieved with a rate 1/2 code.[12] Instead, the (1,7) code with rate 2/3 was adopted. The ISI increased because of the $d = 1$ constraint, but the symbol duration also increased because of the higher code rate, resulting in more available energy for distinguishing signals most likely to be confused. With suitable equalization to mitigate the ISI, the (1,7) code provided a net performance gain over its predecessor (2,7) code. The equalization was based on the simple approach of boosting the high frequencies in the readback signal to slim the transition response.

As the magnetic bit cell continued to shrink, the combination of equalization, RLL coding, and peak detection was no longer adequate to achieve acceptable performance; new detection methods were required to cope with decreasing SNR and severe ISI. Classical transmission techniques, including partial response signaling [4,8,9,25,26,29], decision feedback equalization and its variations [2,3], along with powerful modulation coding [16] were investigated. An exhaustive treatment of the many results (see Refs. 32, 33) from these investigations is beyond the scope of this chapter. Today, partial response signaling is deployed in HDD systems ubiquitously. The remainder of this section is devoted to the theory and practice of partial response signaling in digital magnetic recording channels.

The partial response signaling concept as applied in the digital magnetic recording channel is illustrated in Fig. 12. The readback signal is suitably equalized to a target partial response signal and sampled before detection. Since the input to the write block is a sequence of data symbols, the entire signal path, comprising the write/read/pre-amplify/equalizer/sampling blocks, can be represented by a discrete-time transfer function based on the choice of the target partial response signal. The

[11] The non-return-to-zero-invert (NRZI) format inverts the write current polarity with every occurrence of "1," thereby causing a transition to be recorded on the medium. It is a form of precoding commonly referred to as differential encoding in data communications.

[12] The code capacity for $d = 3$ and unconstrained-k code (that is, the $(3, \infty)$ code) is 0.4650. It is even lower for a code with a finite k constraint.

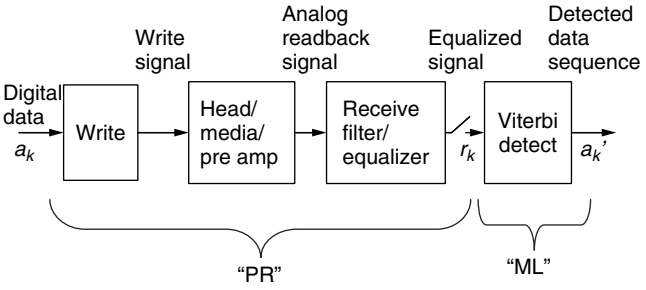


Figure 12. The PRML concept.

detector is based on the Viterbi algorithm performing maximum-likelihood sequence estimation. The overall approach is referred to as partial response maximum-likelihood (PRML) in the HDD industry, where the "PR" and "ML" parts are identified in the figure. The relevant partial response targets for the digital magnetic recording channel are discussed below.

Partial response signaling was developed for transmission of data over band-limited channels (see Ref. 18). Using the sampling theorem, any partial response signal u(t) can be expressed as:

$$u(t) = \sum_k u_k \frac{\sin[\pi(t - kT)/T]}{[\pi(t - kT)/T]} = \sum_k u_k \text{sinc}[(t - kT)/T] \tag{11}$$

where u_k represents the sample value u(kT) and the function sinc(y) is defined as the ratio $\sin(\pi y)/\pi y$. In the frequency domain, the spectrum of the partial response signal is given by:

$$U(f) = \sum_k u_k \exp(-j2\pi fkT), \quad |f| \le 1/2T \tag{12}$$

Partial response signals are designed to support a symbol rate of $1/T$ over a bandwidth of $1/2T$ Hz. With the binary input constraint for the digital magnetic recording channel, this represents a spectral efficiency of 2 bits/sec/Hz. By defining the transform of the unit delay operation, $D = \exp(-j2\pi fT)$, Eq. (12) can be written as a polynomial in D, given by[13]

$$U(D) = \sum_k u_k D^k \tag{13}$$

Infinitely many pulse shapes can be created by choosing different values of u_k in Eq. (11). In general, the number of nonzero u_k's is minimized to achieve the desired performance objectives at least cost.

To understand which partial response signals are well suited for digital magnetic recording, consider the model of saturation recording again. The differencing operation inherent in the recording process (see Eq. (5)) suggests that $(1 - D)$ must be a factor in the polynomial defining the target pulse response for the channel. The $(1 - D)$ partial response system has a high-pass amplitude spectrum with a null at dc. The low-pass filtering effect during readback, due to the gap between the head and the medium, can be modeled by the $(1 + D)^n$ partial response signals, where n is a nonnegative integer. The combined polynomials, representing a set of bandpass responses, are given by

$$P_n(D) = (1 - D)(1 + D)^n \tag{14}$$

These polynomials represent a class of partial response targets that are well suited for the digital magnetic recording channel. This class is called extended partial response (EPR) systems in the magnetic recording literature, where $n = 1$ is referred to as PRML, $n = 2$ as EPRML (for Extended-PRML), $n = 3$ as E^2PRML, and so on. The polynomial $(1 - D^2)$ corresponding to $n = 1$ is the well-known Class IV or Modified Duobinary partial response system [18]. Its application to digital magnetic recording was first noted in [17].

Note that, while $P_n(D)$ defines the target pulse response, the polynomial $(1 + D)^n$ can be interpreted to represent the target transition response since the $(1 - D)$ factor models the differencing operation in Eq. (5). The sample values of the target transition response are given by the binomial coefficients:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{15}$$

since

$$(1 + D)^n = 1 + \binom{n}{1} D + \binom{n}{2} D^2 + \cdots + \binom{n}{n} D^n \tag{16}$$

Note that for large n, the transition response is approximately Gaussian. Indeed, MR heads typically exhibit a transition response between a Lorentzian pulse and a Gaussian pulse. The sample values of the pulse response, p_k, can be derived from the binomial coefficients. Figure 13 shows the target EPR pulse shapes for $n = 1, 2, 3$, and 4. Qualitatively, the EPR signals with increasing n are similar to the Lorentzian signals with increasing δ (see Fig. 7).

Referring back to Fig. 12, $P_n(D)$ defines the "PR" part of the system; that is, it defines the input-output relationship of the sampled data sequences. Thus, if $\{a_k\}$ represents the input data sequence, the noise-free output sequence $\{y_k\}$ is given by:

$$Y(D) = P_n(D)A(D) \tag{17}$$

where

$$Y(D) = \sum_k y_k D^k \tag{18}$$

$$A(D) = \sum_k a_k D^k \tag{19}$$

and

$$P_n(D) = \sum_{k=0}^{n+1} p_k D^k \tag{20}$$

[13] The D is replaced by z^{-1} in digital signal processing literature. The two are equivalent transform representations of a sample sequence.

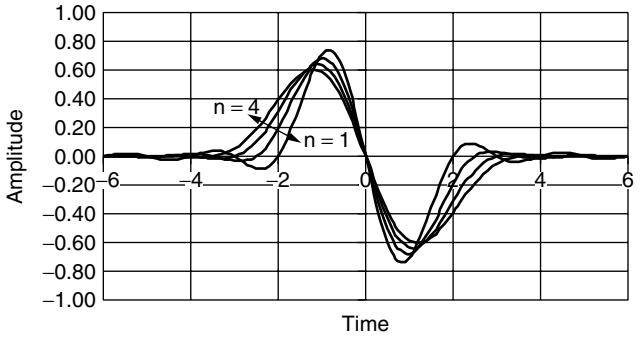


Figure 13. Target EPR pulse shapes for $n = 1, 2, 3$, and 4.

In the time-domain, the noise-free target sampled output can be written as:

$$y_k = \sum_{i=0}^{n+1} p_i a_{k-i} = p_0 a_k + p_1 a_{k-1} + \cdots + p_{n+1} a_{k-n-1} \qquad (21)$$

where the set $\{p_i\}$ is derived from the binomial coefficients. The above formulation stipulates a finite impulse response model for the equalized magnetic recording channel where controlled ISI is allowed between the responses due to the current and the $(n + 1)$ previous inputs. The controlled ISI is prescribed by the choice of $\{p_i\}$, the sample values of the target pulse response. Because of the controlled ISI, the number of output levels is greater than the number of input levels, and depends on the target partial response polynomial. The sampled input to the detector is the noisy sample, given by:

$$r_k = y_k + v_k = \sum_{i=0}^{n+1} p_i a_{k-i} + v_k \qquad (22)$$

where v_k is the sampled noise. The signal spectrum, $S_n(f)$, for each $P_n(D)$ is obtained by setting $D = \exp(-j2\pi fT)$ in Eq. (14), yielding

$$S_n(f) = jT2^n \cos^{n-1}(\pi fT) \sin(2\pi fT), \quad |f| \le 1/2T \qquad (23)$$

Figure 14 shows a plot of the amplitude spectrum $|S_n(f)|$ for different n. Since the factor $(1 + D)$ has the effect of introducing a null at the Nyquist frequency, higher values of n introduce higher order nulls, thereby attenuating the high-frequency content in the target response. This behavior is similar to that exhibited by the Lorentzian model shown in Fig. 8 as δ is increased. As discussed below, for a given linear density, the order of the EPR polynomial can be chosen to maximize the available SNR at the detector.

The equalization of the readback signal to the EPR target can be implemented using analog or digital filters, or combinations thereof. Even with zone bit recording, the linear density changes radially, thus requiring some level of adaptation, either real-time or during zone switching, of the equalizer response. Commercially, both analog and digital implementations have been deployed successfully. The optimization of the filter parameters is generally based on the minimum mean-squared error (MMSE) criterion (see Ref. 24).

Analog equalization typically is implemented using a continuous-time filter with linear phase response. The filter comprises a cascade of two real-axis zeros and a low-pass filter, typically the 7th order Bessel filter. The location of the zeros and the cutoff frequency of the low-pass filter are jointly optimized for each zone and preset by the HDD controller during zone switching. This approach was deployed commercially with EPR and E^2PR target signals (see Refs. 7, 23).

Discrete-time equalization is implemented with a programmable or adaptive finite impulse response (FIR) filter. The choice of the sampling phase of the unequalized readback signal is important to the error rate performance when using symbol-spaced FIR filters. As noted in Fig. 8, the readback signal spectrum typically extends beyond the Nyquist frequency. When sampled at the symbol rate before equalization, foldover of the readback signal spectrum occurs about the Nyquist frequency, resulting in the well-known aliasing effect. The folded spectrum determines the noise enhancement penalty depending on whether the aliasing is additive or subtractive, which, in turn, depends upon the sampling phase. Figure 15 illustrates this effect for the Lorentzian channel with $\delta = 2$. Phase (1) corresponds to sampling at the peak of the transition response, resulting in additive aliasing and no null at Nyquist frequency. Phase (2) corresponds to sampling at $\pm T/2$ seconds from Phase (1), resulting in a null at Nyquist frequency. Even though the target EPR signals require a null at Nyquist frequency, the folded spectrum without the null yields better error rate

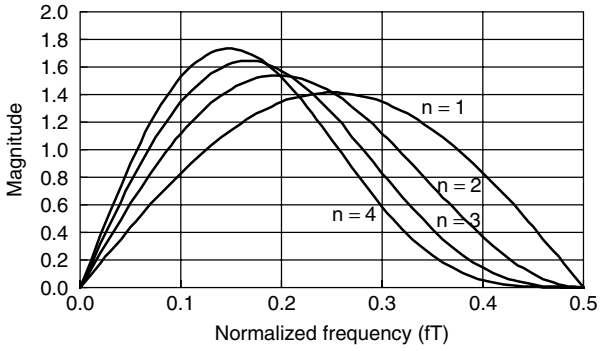


Figure 14. Amplitude spectra of the EPR pulse response signals.

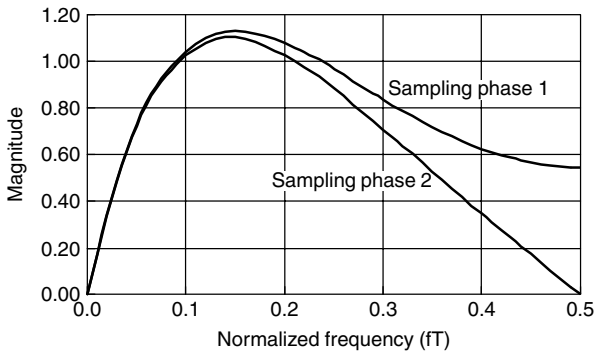


Figure 15. Folded Lorentzian spectrum for two different sampling phases.

performance because of less noise enhancement in the equalizer. Theoretical and experimental results show that the choice of sampling phase can, depending on the operating linear density, affect the SNR by 1–2.5 dB at the detector for PRML and EPRML (see Ref. 27). For higher order systems, the effect is less since the pulse energy of the target response is concentrated at the lower frequencies. The optimum sampling phase depends on the phase response of the channel, including the receive filter. With a linear phase receive filter and symmetric transition response, the optimal sampling phase typically corresponds to sampling at the peak of the transition response.

The controlled ISI inherent in partial response signaling introduces structure in the equalized waveform that can be used to unravel the ISI with little or no loss in performance relative to ISI-free signaling. The required optimum detector is based on maximum-likelihood (ML) sequence estimation instead of symbol-by-symbol detection, as is the case with peak detection. The ML estimation relies on the Viterbi algorithm, which takes the noisy samples r_k in Eq. (22), output from the equalizer, and determines the most likely recorded sequence $\{a_k'\}$. The estimation is based on minimizing recursively the squared-error between the r_k sequence and all *allowed* y_k sequences from the beginning to the end of the data sector. The recursive procedure is applied to a trellis diagram, which graphically depicts all possible states of the channel, defined by the $(n+1)$ most recent inputs $\{a_{k-1}, \ldots, a_{k-n-1}\}$, and the allowed transitions between channel states from time k to $(k+1)$. The complexity of the Viterbi algorithm is proportional to the number of states, which, for the EPR signals, equals 2^{n+1}. For details of the Viterbi algorithm, refer to Ref. 24.

Modulation (line) coding with run-length constraints is also needed with partial response signaling. However, unlike peak detection, the d constraint can be zero since partial response signaling is fundamentally based on allowing controlled ISI at the detector input. The purpose of modulation coding then becomes: (1) to provide frequent updates to the timing recovery and the automatic gain control loops, and (2) to facilitate survivor path merges within a prescribed length of the path memory in the Viterbi detector. The first requirement is similar to that encountered in all digital communication receivers. The second requirement stems from the observation that the EPR polynomials have nulls at both dc and Nyquist frequency. Thus, data sequences with contiguous 1s or -1s (dc), or alternating 1s and -1s (Nyquist frequency) produce zero output levels, and are hence indistinguishable. Such sequences traverse paths in the Viterbi trellis that do not merge, nor accumulate the minimum Euclidean distance, which determines the performance of the ML sequence detector (see Eq. (24)). To avoid performance degradation due to unmerged survivor sequences, the maximum run-length of like symbols (1s or -1s) in both global (contiguous) and interleaved strings of recorded data are constrained to at most G and I symbols, respectively γ. The resulting modulation code is designated (0, G/I), where the 0 represents the d constraint. The G and I constraints may be achieved by interleaving two $(0, k)$ RLL codes. However, a tighter G constraint is achieved for a given code rate by designing the (0, G/I) code as a single code. The (0, G/I) modulation codes typically use the interleaved NRZI (I-NRZI) format to represent the write current. The I-NRZI precoder is based upon applying NRZI precoding to the even and odd bits of the encoded data sequence independently. The first PRML Read Channel deployed commercially in HDD was based upon a rate 8/9, (0, 4/4) code. Subsequently, rate 16/17 codes with looser G and I constraints were used with PRML and EPRML systems.

To ascertain the relative performance of the EPR systems, consider first a channel that already has the prescribed EPR response without any equalization. With additive white Gaussian noise as the only impairment, the probability of error at moderate to high SNR with random data input is given by (see Ref. 12):

$$P_e = K_1 Q\left(\frac{d_{\min}}{2\sigma}\right) \tag{24}$$

where $d_{\min}$ is the minimum Euclidean distance for an error event in the Viterbi detector, K_1 is the average number of such error events, and Q is the area under the tail of the Gaussian density function, given by:

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty \exp(-q^2/2)\, dq \tag{25}$$

If the given EPR channel is used to transmit just one pulse, then no ISI is present. The probability of error for such single use of the channel is given by

$$P_{MF} = K_2 Q\left(\frac{|p|}{2\sigma}\right) \tag{26}$$

where

$$|p|^2 = \sum_{k=0}^{n+1} {p_k}^2 \tag{27}$$

is the energy of the pulse response and K_2 is the number of ways the pulse is incorrectly detected. P_{MF} denotes the probability of error in detecting the isolated pulse (without ISI) with the optimum linear receiver, namely, the matched filter receiver. Comparing Eqs. (24) and (26), the ratio ${d_{\min}}^2/|p|^2$ represents the loss due to the controlled ISI. It is the fraction of the pulse energy that the Viterbi detector is able to use toward discriminating the sequences most likely to be confused. The ISI loss and some key parameters for EPR systems are listed in Table 1. Note that the $n = 1$ and $n = 2$ systems do not incur any ISI loss; thus, they perform as well as an ISI-free signaling scheme. The ISI loss column also represents the relative SNR loss in the detector in choosing a higher order EPR polynomial. From this perspective, $n = 1$ and 2 require the same SNR at the detector for a desired error rate, whereas $n = 3$ requires 2.2 dB more, and so on. The number of output levels for binary inputs and the t_{50}/T of the target transition response, representing the best linear density match in white Gaussian noise, are also given in Table 1.

In deploying EPR systems, complexity and performance considerations alone dictate that the degree of the polynomial, determined by the choice of n, be as small as

Table 1. Key Parameters of EPR Systems

n	No. of Output Levels	$\|p\|^2$	$d^2_{\min}$	$d^2_{\min}/\|p\|^2$	ISI Loss (dB)	t_{50}/T
1	3	2	2	1	0	1.6
2	5	4	4	1	0	2.0
3	7	10	6	0.6	2.2	2.3
4	13	28	12	0.4	3.7	2.6

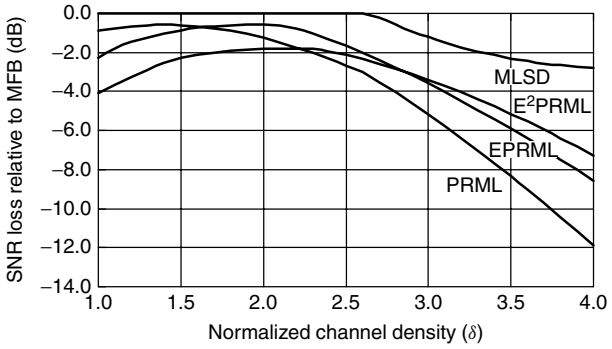


Figure 16. Performance of EPR systems on Lorentzian channel.

possible. Indeed, higher order partial response polynomials have not been of much interest in communication channels. But the digital magnetic recording channel is different in that the transition response is fixed for a given head/medium combination. As the linear density is increased with a given EPR target, the noise enhancement penalty from the equalizer increases because of the growing mismatch between the readback signal and target signal spectra. Indeed, when the ISI loss and the noise enhancement penalty are taken into account, a given polynomial will provide acceptable performance over a range of recording density; beyond that range, a higher degree polynomial is needed to reduce the noise enhancement penalty, and achieve acceptable performance. Figure 16 illustrates this point using the Lorentzian pulse generator followed by additive white Gaussian noise as the model for the recording channel. Using the MMSE criterion for the equalizer design and taking into account the effect of noise correlation in the Viterbi detector, the asymptotic SNR loss relative to the matched filter bound ($\mathrm{SNR}_{\mathrm{MF}} = |p|^2/\sigma^2$) is plotted as a function of normalized density δ. Note that 0 dB loss represents matched filter performance. As shown, PRML ($n = 1$) provides better performance for δ in the range of 1.0 to 1.6, EPRML ($n = 2$) provides better performance for δ in the range of 1.6 to 2.8; and so on. To provide a benchmark, Fig. 16 also shows the performance of the maximum likelihood sequence detector (MLSD), which is the optimum detector for ISI channels (see Ref. 12). As shown, the MLSD detector achieves the matched filter bound up to the normalized density of approximately 2.6; that is, its performance is the same as that of an ISI-free channel. Beyond that range, however, the performance of MLSD degrades relative to the matched filter bound because of the increasing ISI. Note that the performance of the EPR systems, featuring moderate complexity, is within 1–2 dB of MLSD over the range of linear densities of current interest.

Read Channel products based on PRML, EPRML, and *E²PRML* have been deployed commercially in hundreds of millions of magnetic disk drives during the past decade. More recently, there has been increased interest in the development of *generalized partial response* (GPRML) polynomials to bridge the performance gap between MLSD and EPR polynomials. Such polynomials are derived using search procedures that minimize the probability of error on an empirical model of the recording channel. Since the optimum target response is one which whitens the noise at the detector input, the GPRML schemes, in a sense, perform spectral matching and noise whitening jointly with a polynomial of prescribed degree and spectral null constraints.[14] Unlike the EPR targets, which are symmetrical like the ideal channel, the resulting GPR targets are asymmetrical and closer to a minimum phase representation of the channel.[15] The GPRML approach is shown to be quite promising at high linear density ($\delta > 2.7$) with polynomials of degree 4 and above. About 1 dB SNR advantage is achieved over an equivalent EPR polynomial with modest increase in the equalizer complexity (see Ref. 10).

As recording densities continue to increase, *channel coding* techniques have been developed for partial response signaling to deal with the reduced SNR. Unlike communication channels, however, channel coding is difficult to apply in the digital magnetic recording channel because of the binary constraint on the input waveform. Also, for a given head/medium combination, the pulse-energy-to-noise ratio, representing the matched filter bound, decreases rapidly because of the increased ISI and noise bandwidth associated with adding redundancy inherent in channel coding. The rate of this decrease is 6–9 dB per doubling of the symbol rate (or linear density). Thus, in order to achieve a net gain with a rate 1/2 code, the coding gain must be larger than 6–9 dB — a nontrivial task! Channel coding, therefore, must rely on the use of high code rates to eke out a net performance gain for a given recording channel.

Two types of channel coding methods have been developed and deployed commercially in HDD systems: *trellis coding* and *parity coding*. Both methods rely on suitably dealing with the most likely minimum distance error events for the target partial response signal, since it is these events that limit the performance of the detector (see Eq. (23)).

The underlying approach in the trellis coding method is similar to that used in data communications: increase the minimum Euclidean distance between all allowed sequences of output symbols, y_k. This objective is achieved by eliminating input data patterns that support the prescribed minimum distance error events for a given partial response system. The constraints on the input sequence together with those from the target partial response signal are suitably combined to create a new trellis diagram which has larger minimum Euclidean distance than the uncoded system. Interestingly, the

[14] First order nulls at dc and at Nyquist frequency are typically retained, although some reported polynomials have dropped the null at the Nyquist frequency.

Table 2. Some Input Error Sequences for E^2PR

d^2	Input Error Sequence ($\pm e_k^a$)
6	1-11
8	1-11001-11 1-11-11-1 1-11-11-11-1... 1-11-11-11 1-11-11-11-11
10	1 1-110-11-1 1-11001-11001-11 ...

resulting trellis may, at times, be simpler than that for the uncoded system.

For example, Table 2 lists a few of the minimum distance input error sequences for the E^2PR polynomial. The *input error sequence* $\{e_k{}^a\}$ is the difference between any two possible input data sequences, and the *error event* $\{e_k{}^y\}$ is the difference between the corresponding noise-free channel output sequences (see Eq. (21)). The two error sequences are related as follows:

$$e_k{}^y = y_k{}^1 - y_k{}^2 = \sum_{i=0}^{n+1} p_i a_{k-i}{}^1 - \sum_{i=0}^{n+1} p_i a_{k-i}{}^2$$
$$= \sum_{i=0}^{n+1} p_i (a_{k-i}{}^1 - a_{k-i}{}^2) = \sum_{i=0}^{n+1} p_i e_{k-i}{}^a \qquad (28)$$

where $y_k{}^1$ and $y_k{}^2$ are two noise-free output sequences due to input data sequences $a_k{}^1$ and $a_k{}^2$, respectively. Since $a_k{}^i$ is binary, the associated error sequence alphabet $e_k{}^a$ is ternary, as given in Table 2. By avoiding the input NRZ sequences of the form 1-11 and −11-1, both the squared-distance 6 and 8 error events can be eliminated, along with some of the squared-distance 10 events. This constraint on the input sequence is easily introduced through a suitable combination of the $d = 1$ RLL code and precoding, and eliminating the associated states in the Viterbi trellis. The resulting trellis diagram has 12 states, instead of 16 states in the uncoded $E^2PRML(n = 3)$ system, along with minimum squared-Euclidean distance for an error event of 10. The resulting coding gain is $10\log(10/6) = 2.2$ dB. This combined coding and equalization scheme was deployed commercially to replace the (1,7)-Coded Peak Detection scheme.

Even higher code rates can be used on the E^2PR channel to achieve the coding gain of 2.2 dB. These codes rely on the use of a 16-state time-varying trellis to represent the constraints on the input sequences, and are referred to as TMTR (Time-Varying Maximum Transition Run-Length) codes (see Refs. 6, 20, 23). A more general and systematic trellis code construction method for partial response channels relies on suitably matching the nulls of the code spectrum with those of the partial response channel to increase the minimum Euclidean distance. The theory and implementation of this method can be found in Refs. 11, 16.

The parity coding method aims to detect and selectively correct the dominant error events at the output of the Viterbi detector to improve the error rate performance. Parity bits are suitably inserted within codewords to detect the occurrence of parity violations at the output of the detector. Soft decision correction, in the maximum likelihood sense, is applied to correct the parity violations. The applicability of this approach relies on the unique patterns that characterize the most likely error events for EPR and GPR targets. Unlike trellis coding, parity coding does not add any constraints on the input sequences beyond those imposed by the modulation code. The resulting code rate can be, therefore, higher than that for trellis codes. Indeed, in commercial deployments, the parity-coded schemes achieved the same code rate of 16/17 as in modulation-code-only systems by combining the parity and the modulation constraints in longer block codes,[16] such as rate 32/34 or rate 96/102.

To illustrate the parity coding approach, consider the use of EPRML on the Lorentzian channel with additive white Gaussian noise. Table 3 lists the ordered sets of error sequences as a function of the normalized linear density. The ordering is based on the likelihood of occurrence of each event, defined in terms of the SNR level above the most likely error event (normalized to 0 dB). The ordering of error sequences changes as a function of the linear density because of the changing noise correlation introduced by the equalizer. Note that, except for 1-11-1, all other error sequences involve an odd number of input bits. Thus, by appending a bit to create codewords with even parity, the occurrence of most likely error events within a codeword can be detected. The same approach can be applied to measured signal and noise from the recording channel and to other EPR or GPR targets. Indeed, the approach can be extended to include multiple bits of parity, wherein the data to be recorded is organized into multidimensional arrays and parity bits are suitably appended in each dimension to achieve a prescribed detection and correction capability.

Table 3. Input Error Sequences for EprmL System

Normalized Linear Density	Input Error Event Sequence	SNR Level (dB)
2.25	1 1-11-11 1-11-11-11 ...	0 0.19 0.33
2.5	1-11 1-11-11 1-11-11-11 ...	0 0.14 0.14
2.75	1-11 1-11-11 1-11-1 ...	0 0.45 0.58

[16] The resulting G and I constraints are looser relative to the modulation code with rate 16/17.

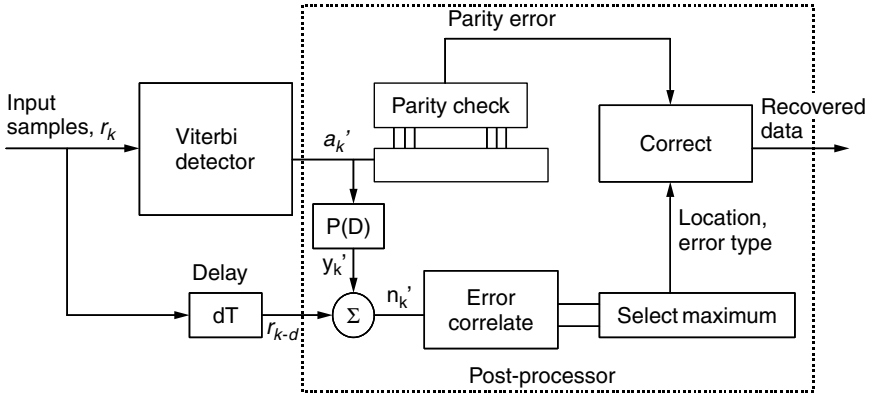


Figure 17. Detector structure for parity coding using postprocessor.

The detection and correction of parity violations may be performed within the Viterbi detector, or with a combination of the Viterbi detector and a postprocessor. With the former approach, the combined constraints of the parity code and the target response are incorporated into a time-varying trellis with twice the number of states of the original target response. The Viterbi algorithm is applied to the new trellis to recover the corrected data sequence. For a target response with 16 states, which represents the state of the art at the time of this writing, the increased complexity with this approach is quite significant. Instead, the combination of the Viterbi detector and postprocessor is commonly used with parity coding (see Refs. 10, 12, 31).

Figure 17 shows the postprocessor-based detector structure. The detection of parity violations is performed at the end of each codeword by the parity check block. The most likely location of the parity violation, if it occurs, is determined as follows: The Viterbi detector for the target partial response signal produces the estimated data sequence $\{a_k'\}$, which is used to reconstruct the noise-free partial response output symbols $\{y_k'\}$. Together with suitably delayed input samples, the sequence $\{y_k'\}$ is used to estimate the noise sequence $\{n_k'\}$, where $n_k' = r_k - y_k'$. These noise estimates are correlated at each bit time using a bank of matched filters, where each filter is matched to the most likely error event. The noise correlation outputs are only considered valid if the estimated sequence $\{a_k'\}$ supports the prescribed error events. The bit interval with the maximum valid noise correlation output is assumed to be the location of the error sequence. This location and the associated error event are then used to correct a parity violation, if necessary.

The parity coding scheme is effective in providing coding gains in excess of 1 dB without incurring any code rate penalty relative to systems with only modulation coding.

Unencumbered by signaling standards, the application of partial response signaling and channel coding techniques described above has been fast paced during the past decade. The peak detection method, which enjoyed widespread use for almost 30 years, was replaced by the more powerful PRML method in the early 1990s. PRML was completely displaced by EPRML and E^2PRML during the mid- to late 1990s. Parity and trellis coding techniques were introduced in late 1990s along with GPRML. Much research activity today is focused on combining Turbo coding techniques with partial response equalization (see Ref. 21). But issues related to the decoding delay inherent in Turbo decoding need to be resolved first. These issues pertain to delay requirements that exist within the datapath of the HDD as well as between the host CPU and the disk drive. Given the continuing thrust to double the areal density annually, it is just a matter of time before these issues are resolved. The trend in the semiconductor industry to integrate more and more functionality on a single chip is likely to also facilitate the development and deployment of Turbo coding methods.

5. CONCLUDING REMARKS

Digital magnetic storage has played a pivotal role in the evolution of storage systems over the past 45 years. Thousands of terabytes of storage are consumed annually worldwide, and the demand is exploding as new applications emerge in computing, communications, and entertainment systems. The underlying technologies in digital magnetic recording continue to progress at an exponential rate. And while some fundamental limitations due to the decreasing bit cell are anticipated above 100 Gbits/sq. inch, researchers are busy exploring means to overcoming those limitations. As noted in Ref. 28, no alternative storage technologies exist on the horizon which show promise for replacing the magnetic hard disk drive in the next ten years.

Over the past decade, equalization and coding methods have played a vital role in the growth of storage capacity per disk surface. New and powerful methods of combining equalization and coding are being developed and deployed commercially. This trend will continue as the digital magnetic channel itself continues to evolve based on new head, media, and recording technologies.

Acknowledgments

The author would like to convey special thanks to some colleagues from former DataPath Systems who, for six years, helped shape the trends in Read Channel technology and products beyond PRML: S. Altekar, J. Chern, C. Conroy, R. Contreras, L. Fang, Y. Hsieh, E. MacDonald, T. Pan, S. Shih, and A. Yeung; and to

former colleagues with whom he had the privilege of collaborating on modern signal processing and coding methods for digital magnetic storage: J. Cioffi, T. Howell, R. Karabed, M. Melas, A. Patel, P. Siegel, J. Wolf, and R. Wood. Thanks are also due to Ed Growchoski of IBM Corporation for providing the chart on the areal density trends.

BIOGRAPHY

Hemant K. Thapar received the M.S and Ph.D. degrees in Electrical Engineering from Purdue University in 1977 and 1979, respectively. He worked at Bell Telephone Laboratories, Holmdel (1979–84), at IBM Corporation, San Jose (1984–94), and at DataPath Systems (1994–2000), which he cofounded. He is presently a senior vice-president at LSI Logic Corporation, Milpitas, California, and an adjunct lecturer at Santa Clara University, California. He was corecipient of: the Best Paper Award at Interface 1984 for his work on high-speed, full-duplex data transmission; the Best Technical Report citation from IBM Almaden Research Center in 1989 for his work on PRML technology; and the 1991 IEEE Communications Magazine Prize Paper award for his paper on future technology directions in signal processing for data storage. Dr. Thapar is a Fellow of the IEEE and holds many patents and publications in the areas of digital communications, data storage, and networking. His technical interests are in the areas of data transmission and storage, networking, and VLSI architectures and design.

BIBLIOGRAPHY

1. T. C. Arnoldussen and L. L. Nunnelley, *Noise in Digital Magnetic Recording*, World Scientific Publishing Co. Pte. Ltd., Singapore, 1992.
2. P. S. Bednarz et al., Performance evaluation of an adaptive RAM-DFE read channel, *IEEE Trans. Magn.* **MAG-31**(2): 1121–1127 (March 1995).
3. J. W. M. Bergmans, Decisions feedback equalization for run-length limited modulation codes with $d = 1$, *IEEE Trans. Magn.* (1996).
4. J. W. M. Bergmans, *Digital Baseband Transmission and Recording*, Kluwer Academic, 1996.
5. H. N. Bertram, *Theory of Magnetic Recording*, Cambridge University Press, United Kingdom, 1994.
6. W. Bliss, An 8/9 rate time-varying trellis code for high density magnetic recording, *IEEE Trans. Magn.* **33**: 2746–2748 (Sept. 1997).
7. J. Chem et al., An EPRML digital read/write channel IC, *ISSCC Digest of Tech. Papers*, Paper 19.4, 320–322 (Feb. 1997).
8. R. D. Cideciyan, F. Dolivo, R. Hermann, W. Hirt, and W. Schott, A PRML system for digital magnetic recording, *IEEE J. Select. Areas Commun.* **SAC-10**(1): 38–56 (Jan. 1992).
9. J. M. Cioffi, W. L. Abbott, H. K. Thapar, C. M. Melas, and K. D. Fisher, Adaptive equalization in magnetic-disk storage channels, *IEEE Comm. Mag.* 14–29 (Feb. 1990).
10. T. Conway, A new target response with parity coding for high density magnetic recording channels, *IEEE Trans. Magn.* **34**(4): 2382–2386 (July 1998).
11. L. Fredrickson et al., Improved trellis coding for partial response channels, *IEEE Trans. Magn.* **31**: 1141–1148 (March 1995).
12. G. D. Forney, Jr., Maximum likelihood estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**(3): 363–378 (May 1972).
13. E. Grochowski and R. Hoyt, Future trends in hard disk drives, *IEEE Trans. Magn.* **32**(3): 1850–1854 (May 1996).
14. E. Grochowski, website: *http://www.storage.ibm.com/*
15. J. M. Harker, D. W. Bede, R. E. Pattison, G. R. Santana, and L. G. Taft, A quarter century of disk file innovation, *IBM J. Res. Devel.* **25**(5): 677–689 (Sept. 1981).
16. R. Karabed and P. Siegel, Matched spectral-null codes for partial response channels, *IEEE Trans. Inform. Theory* **37**(3): 818–855 (May 1991).
17. H. Kobayashi and D. T. Tang, Application of partial response channel coding to magnetic recording systems, *IBM J. Res. Devel.* **15**: (July 1970).
18. A. Lender, Correlative digital communication techniques, *IEEE Trans. Comm. Tech.* **COM-12**: (Dec. 1964).
19. B. Marcus, P. Siegel, and J. Wolf, Finite-state modulation codes for data storage, *IEEE J. Select. Areas Commun.* **SAC-10**(1): 5–37 (Jan. 1992).
20. J. Moon and B. Brickner, Maximum transition run codes for data storage systems, *IEEE Trans. Magn.* **32**: 3992–3994 (Sept. 1996).
21. M. Oberg and P. Siegel, Performance analysis of turbo-equalized partial response channels, *IEEE Trans. Commun.* **49**(3): 436–444 (March 2001).
22. D. Palmer, J. Hong, D. Stanek, and R. Wood, Characterization of the read/write process for magnetic recording, *IEEE Trans. Magn.* **MAG-31**(2): 1071–1076 (March 1995).
23. T. Pan et al., A trellis-coded E^2PRML digital read/write channel IC, *ISSCC Digest of Tech. Papers*, Paper MP 2.2, 36–37 (Feb. 1999).
24. J. Proakis, *Digital Communications*, 2nd Edition, McGraw-Hill, New York, 1989.
25. P. H. Siegel and J. K. Wolf, Modulation and coding for information storage, *IEEE Commun. Mag.* **29**(12): 68–86 (Dec. 1991).
26. H. K. Thapar and A. M. Patel, A class of partial response systems for increasing storage density in Magnetic Recording, *IEEE Trans. Magn.* **MAG-23**(5): 3666–3668 (Sept. 1987).
27. H. Thapar, P. Ziperovich, and R. Wood, On the performance of symbol- and fractionally-spaced equalization in digital magnetic recording, *Proc. of IEEE Audio, Video, and Data Recording*, (May 1990).
28. D. A. Thompson and J. S. Best, The future of magnetic data storage technology, *IBM J. Res. Devel.* **44**(3): 311–322 (May 2000).
29. R. W. Wood and D. A. Peterson, Viterbi detection of class IV partial response on a magnetic recording channel, *IEEE Trans. Commun.* **COM-34**(5): 454–461 (May 1986).
30. R. W. Wood, Magnetic recording systems, *Proc. IEEE* **74**(11): 1557–1569 (Nov. 1986).
31. R. Wood, Turbo-PRML: A compromise EPRML detector, *IEEE Trans. Magn.* **29**: 4018–4020 (Nov. 1993).
32. *IEEE J. Select. Areas Commun.* **SAC-10**(1): (Jan. 1992).
33. *IEEE J. Select. Areas Commun.* **SAC-19**(4): (April 2001).

MATCHED FILTERS IN SIGNAL DEMODULATION

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

In digital communication systems, the modulator maps a sequence of information bits into signal waveforms that are transmitted through the communication channel. The simplest form of digital modulation is binary modulation, in which the information bit 0 is mapped into a signal waveform $s_0(t)$ and the information bit 1 is mapped by the modulator into the signal waveform $s_1(t)$. Thus, if the binary data rate into the modulator is R bits per second, each waveform may be confined to occupy a time duration $T_b = 1/R$ seconds, where T_b is called the *bit interval*. Then, the mapping performed by the modulator for binary signalling may be expressed as

$$\begin{aligned} 0 &\rightarrow s_0(t), \quad 0 \le t \le T_b \\ 1 &\rightarrow s_1(t), \quad 0 \le t \le T_b \end{aligned}$$

Higher-level modulation can be performed by mapping multiple data bits into corresponding signal waveforms. Specifically, the modulator may employ $M = 2^k$ different signal waveforms to map groups of k bits at a time for transmission through the communication channel. A group of k bits is called a *symbol*, and, for a data rate of R bits per second, the corresponding signal waveforms generally may be of duration $T_s = k/R = kT_b$ seconds. T_s is called the *symbol duration*, and the modulator for $M > 2$ is generally said to perform M-ary signal modulation.

For example, $M = 4$ signal waveforms are used to transmit pairs of data bits. The mapping performed by the modulator for $M = 4$ may be expressed as (with $T_s = 2T_b$)

$$\begin{aligned} 00 &\rightarrow s_0(t), \quad 0 \le t \le T_s \\ 01 &\rightarrow s_1(t), \quad 0 \le t \le T_s \\ 10 &\rightarrow s_2(t), \quad 0 \le t \le T_s \\ 11 &\rightarrow s_3(t), \quad 0 \le t \le T_s \end{aligned}$$

The set of signal waveforms $\{s_m(t), m = 0, 1, \ldots, M-1\}$ for $M = 2^k, k = 1, 2, \ldots$, which convey the information bits may differ either in amplitude, as in pulse amplitude modulation (PAM), or in phase, as in phase shift keying (PSK), or in both amplitude and phase, as in quadrature amplitude modulation (QAM); or, more generally, they may be multidimensional signal waveforms constructed from different frequencies, as in M-ary frequency shift keying (MFSK), or from pulses transmitted in different time slots as in M-ary time shift keying (MTSK).

In the transmission of the signal through the channel, the signal is corrupted by additive noise. This noise originates at the front end of the receiver and is well modeled statistically as a Gaussian random process. Hence, if the transmitted signal is $s_m(t), 0 \le t \le T_s$, the received signal $r(t)$ may be expressed as

$$r(t) = s_m(t) + n(t), \ 0 \le t \le T_s$$

where $m = 0, 1, \ldots, M-1$.

2. SIGNAL DEMODULATION

The basic function of the demodulation process is to recover the transmitted data by processing the received signal $r(t)$. Since the noise in the received signal in any signaling interval of duration T_s is a sample function of a random process, the demodulation of $r(t)$ should be designed to minimize the probability of a symbol error or, equivalently, to maximize the probability of a correct decision. The probability of error is the probability of selecting a signal waveform $s_j(t)$ when, in fact, waveform $s_i(t)$ was transmitted where $i \neq j$. On the basis of this criterion, the optimum demodulator, having observed $r(t)$ over the time interval $0 \le t \le T_s$, computes the M (posterior) probabilities

$$P_m \equiv P[s_m(t) \text{ was transmitted} \mid r(t), \ 0 \le t \le T_s]$$

and selects the signal waveform corresponding to the largest probability. It is shown in basic textbooks on digital communications [1–3] that the optimum demodulator obtained by applying this maximum aposterior probability (MAP) design criterion, when the transmitted signal is corrupted by additive white Gaussian noise (AWGN), consists of a parallel bank of M matched filters, where the filter impulse responses are matched to the M possible transmitted signal waveforms $s_m(t), m = 0, 1, \ldots, M-1$. The outputs of these M linear filters are sampled at the end of the signaling interval T_s, and the samples are passed to the detector which selects the largest of the M samples and performs the inverse mapping from the corresponding waveform to the $k = \log_2 M$ data bits. A block diagram of the optimum demodulator is illustrated in Fig. 1.

3. THE MATCHED FILTER

Consider a time-limited signal waveform as shown in Fig. 2a. A filter is said to be matched to the signal waveform $s(t)$ if its impulse response $h(t)$ is given as

$$h(t) = s(T - t), \ 0 \le t \le T$$

For the signal waveform shown in Fig. 2, the impulse response of the matched filter is shown in Fig. 2b.

Now, suppose that the signal $s(t)$ is the input to the filter whose impulse response is matched to $s(t)$. The output of the matched filter is given by the convolution integral

$$y(t) = \int_0^t s(\tau) h(t - \tau)\, d\tau \tag{1}$$

where $h(t) = s(T - t)$. By substituting for $h(t)$ in Eq. (1), one obtains the result

$$y(t) = \int_0^t s(\tau) s(T - t + \tau)\, d\tau \tag{2}$$

Figure 3 illustrates the filter output waveform $y(t)$.

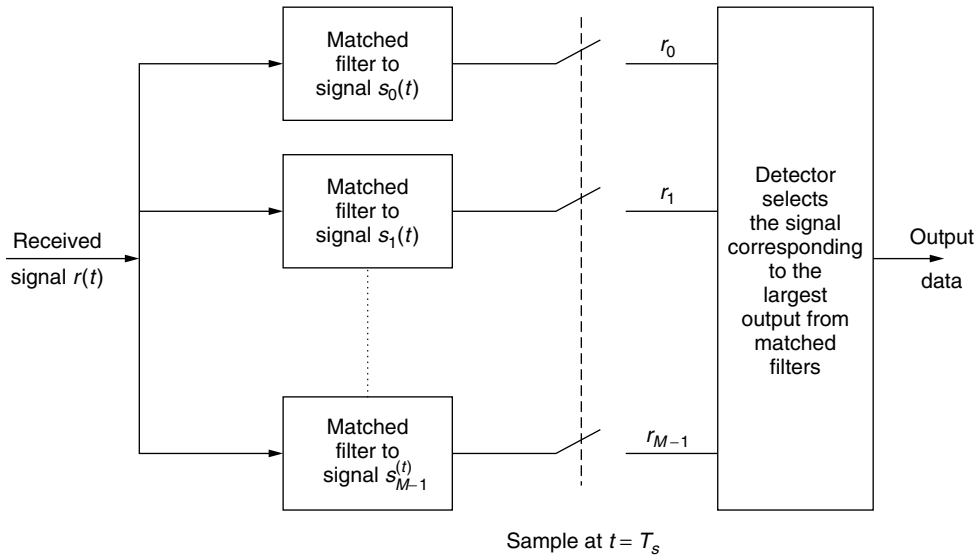


Figure 1. Signal demodulation using matched filters.

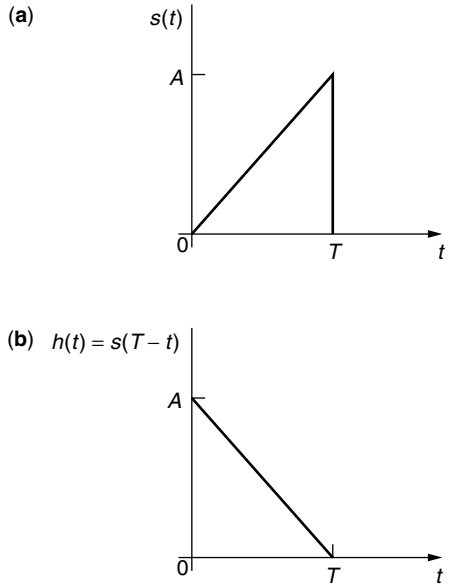


Figure 2. Signal $s(t)$ and filter matched to $s(t)$: (**a**) signal $s(t)$; (**b**) impluse response of filter matched to $s(t)$.

It is observed that $y(t)$ has a peak at $t = T$, whose value is

$$y(T) = \int_0^T s^2(\tau)\, d\tau = \mathcal{E} \tag{3}$$

which is the signal energy $\mathcal{E}$ in the signal waveform $s(t)$. Furthermore, $y(t)$ is symmetric with respect to the point $t = T$. In fact, the form of Eq. (2) is simply the time autocorrelation function of the signal $s(t)$, which is symmetric for any arbitrary signal waveform. Consequently, any signal waveform $s(t), 0 \le t \le T$, when passed through a filter matched to it, will result in an output that is the time-autocorrelation of $s(t)$, and the value of the output $y(t)$ at $t = T$ will be the energy in the signal $s(t)$, as given by equation (3).

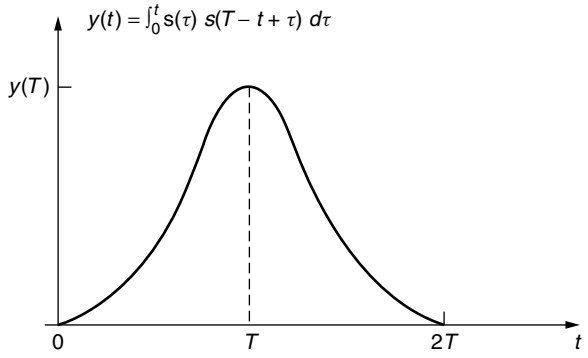


Figure 3. Output response of matched filter for the signal in Fig. 2.

4. PROPERTIES OF THE MATCHED FILTER

A matched filter has some interesting properties. We consider the most important property, which may be stated as follows: If a signal $s(t)$ is corrupted by AWGN, the filter with impulse response matched to $s(t)$ maximizes the output SNR.

To prove this property, let us assume that the received signal $r(t)$ consists of the signal $s(t)$ and AWGN $n(t)$ which has zero-mean and power-spectral density $\Phi_n(f) = N_0/2$ W/Hz. Suppose the signal $r(t)$ is passed through a filter with impulse response $h(t), 0 \le t \le T$, and its output is sampled at time $t = T$. The filter response to the signal and noise components is

$$\begin{aligned} y(t) &= \int_0^t r(\tau)h(t-\tau)\, d\tau \\ &= \int_0^t s(\tau)h(t-\tau)\, d\tau + \int_0^t n(\tau)h(t-\tau)\, d\tau \end{aligned} \tag{4}$$

At the sampling instant $t = T$, the signal and noise components are

$$\begin{aligned} y(T) &= \int_0^T s(\tau)h(T-\tau)\,d\tau + \int_0^T n(\tau)h(T-\tau)\,d\tau \\ &= y_s(T) + y_n(T) \end{aligned} \tag{5}$$

where $y_s(T)$ represents the signal component and $y_n(T)$ represents the noise component. The problem is to select the filter impulse response that maximizes the output SNR defined as

$$\left(\frac{S}{N}\right)_0 = \frac{y_s^2(T)}{E[y_n^2(T)]} \tag{6}$$

The denominator in Eq. (6) is simply the variance of the noise term at the output of the filter. Let us evaluate $E[y_n^2(T)]$. We have

$$\begin{aligned} E[y_n^2(T)] &= \int_0^T \int_0^T E[n(\tau)n(t)]h(T-\tau)h(T-t)\,dt\,d\tau \\ &= \int_0^T \int_0^T \frac{N_0}{2}\delta(t-\tau)h(T-\tau)h(T-t)\,dt\,d\tau \\ &= \frac{N_0}{2}\int_0^T h^2(T-t)\,dt \end{aligned} \tag{7}$$

Note that the variance depends on the power spectral density of the noise and the energy in the impulse response $h(t)$.

By substituting for $y_s(T)$ and $E\left[y_n^2(T)\right]$ into Eq. (6), we obtain the expression for the output SNR as

$$\left(\frac{S}{N}\right)_0 = \frac{\left[\int_0^T s(\tau)h(T-\tau)\,d\tau\right]^2}{\frac{N_0}{2}\int_0^T h^2(T-\tau)\,dt} = \frac{\left[\int_0^T h(\tau)s(T-\tau)\,d\tau\right]^2}{\frac{N_0}{2}\int_0^T h^2(T-\tau)\,dt} \tag{8}$$

Since the denominator of the SNR depends on the energy in $h(t)$, the maximum output SNR over $h(t)$ is obtained by maximizing the numerator of $(S/N)_0$ subject to the constraint that the denominator is held constant. The maximization of the numerator is most easily performed by use of the Cauchy–Schwarz inequality, which states, in general, that if $g_1(t)$ and $g_2(t)$ are finite-energy signals, then

$$\left[\int_{-\infty}^{\infty} g_1(t)g_2(t)\,dt\right]^2 \le \int_{-\infty}^{\infty} g_1^2(t)\,dt \int_{-\infty}^{\infty} g_2^2(t)\,dt$$

where equality holds when $g_1(t) = Cg_2(t)$ for any arbitrary constant C. If we set $g_1(t) = h(t)$ and $g_2(t) = s(T-t)$, it is clear that the $(S/N)_0$ is maximized when $h(t) = Cs(T-t)$; thus, $h(t)$ is matched to the signal $s(t)$. The scale factor C^2 drops out of the expression for $(S/N)_0$ since it appears in both the numerator and the denominator.

The output (maximum) SNR obtained with the matched filter is

$$\begin{aligned} \left(\frac{S}{N}\right)_0 &= \frac{2}{N_0}\int_0^T s^2(t)\,dt \\ &= \frac{2\mathcal{E}}{N_0} \end{aligned} \tag{9}$$

4.1. Frequency-Domain Interpretation of the Matched Filter

The matched filter has an interesting frequency-domain interpretation. Since $h(t) = s(T-t)$, the Fourier transform of this relationship is

$$\begin{aligned} H(f) &= \int_0^T s(T-t)e^{-j2\pi ft}\,dt \\ &= \left[\int_0^T s(\tau)e^{j2\pi f\tau}d\tau\right]e^{-j2\pi fT} \\ &= S^*(f)e^{-j2\pi fT} \end{aligned} \tag{10}$$

We observe that the matched filter has a frequency response that is the complex conjugate of the transmitted signal spectrum multiplied by the phase factor $e^{-j2\pi fT}$, which represents the sampling delay of T. In other words, $|H(f)| = |S(f)|$, so that the magnitude response of the matched filter is identical to the transmitted signal spectrum. On the other hand, the phase of $H(f)$ is the negative of the phase of $S(f)$.

Now, if the signal $s(t)$, with spectrum $S(f)$, is passed through the matched filter, the filter output has a spectrum $Y(f) = |S(f)|^2 e^{-j2\pi fT}$. Hence, the output waveform is

$$\begin{aligned} y_s(t) &= \int_{-\infty}^{\infty} Y(f)e^{j2\pi ft}\,df \\ &= \int_{-\infty}^{\infty} |S(f)|^2 e^{-j2\pi fT}e^{j2\pi ft}\,df \end{aligned} \tag{11}$$

By sampling the output of the matched filter at $t = T$, we obtain

$$y_s(T) = \int_{-\infty}^{\infty} |S(f)|^2\,df = \int_0^T s^2(t)\,dt = \mathcal{E} \tag{12}$$

where the last step follows from Parseval's relation.

The noise of the output of the matched filter has a power spectral density

$$\Phi_0(f) = |H(f)|^2\frac{N_0}{2} \tag{13}$$

Hence, the total noise power at the output of the matched filter is

$$\begin{aligned} P_n &= \int_{-\infty}^{\infty} \Phi_0(f)\,df \\ &= \int_{-\infty}^{\infty} \frac{N_0}{2}|H(f)|^2\,df = \frac{N_0}{2}\int_{-\infty}^{\infty} |S(f)|^2\,df = \frac{\mathcal{E}N_0}{2} \end{aligned} \tag{14}$$

The output SNR is simply the ratio of the signal power P_s, given by

$$P_s = y_s^2(T)$$

to the noise power P_n. Hence

$$\left(\frac{S}{N}\right)_0 = \frac{P_s}{P_n} = \frac{\mathcal{E}^2}{\mathcal{E}N_0/2} = \frac{2\mathcal{E}}{N_0} \tag{15}$$

which agrees with the result given by Eq. (9).

5. CONCLUDING REMARKS

Matched filters are widely used for signal demodulation in digital communication systems and in radar signal receivers. In the latter, the transmitted signal usually consists of a series of signal pulses. When the signal pulses are reflected from an object, such as an airplane, the received signal over the observation interval has the form $r(t) = s(t - t_0) + n(t)$, where $n(t)$ represents the additive noise and t_0 represents the round-trip time delay corresponding to the signal reflected from the object. By passing the received signal $r(t)$ through the filter matched to $s(t)$ and determining when the matched-filter output reaches a peak value that exceeds a predetermined threshold, an estimate of the time delay is obtained. From this measurement of t_0, the distance (range) of the object from the radar position is determined. If the threshold is not exceeded during an observation interval, a decision is made that no target or object is present at that corresponding range.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
2. S. Haykin, *Communication Systems*, 4th ed., Wiley, New York, 2000.
3. H. Stark, F. B. Tuteur, and J. B. Anderson, *Modern Electrical Communication Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1988.

MAXIMUM-LIKELIHOOD ESTIMATION

SIMON HAYKIN
McMaster University
Hamilton, Ontario, Canada

1. INTRODUCTION

Estimation theory is a branch of probability and statistics that deals with the problem of deriving information about properties of random variables and stochastic processes, given a set of observed data. This problem arises frequently in the study of communication and control systems. *Maximum likelihood* is a powerful method of parameter estimation, which was pioneered by Fisher [1]. In principle, the method of maximum likelihood may be applied to any estimation problem, with the proviso that we formulate the joint probability density function of the available set of observed data. The method then yields almost all the well-known estimates as special cases.

2. LIKELIHOOD FUNCTION

The method of maximum likelihood is based on a relatively simple idea:

> Different populations tend to generate different data samples, where the given data sample is more *likely* to have come from some population than from other populations.

Let $f_{\mathbf{U}}(\mathbf{u} \mid \theta)$ denote the *conditional joint probability density function* of the *random vector* $\mathbf{U}$ represented by the observed sample vector $\mathbf{u}$ with elements $u_1, u_2, \ldots, u_M$, where θ is a parameter vector with elements $\theta_1, \theta_2, \ldots, \theta_K$. The method of maximum likelihood is based on the principle that we should estimate the parameter vector θ by its most *plausible value*, given the observed sample vector $\mathbf{u}$. In other words, the maximum-likelihood

estimates of $\theta_1, \theta_2, \ldots, \theta_K$ are those values of the parameter vector for which the conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \theta)$ is a maximum.

The term *likelihood function*, denoted by $l(\theta)$, is given to the conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \theta)$, viewed as a function of the parameter vector θ. We thus write

$$l(\theta) = f_{\mathbf{U}}(\mathbf{u} \mid \theta) \tag{1}$$

Although the conditional joint probability density function and the likelihood function have exactly the same formula, it is vital that we appreciate the physical distinction between them. In the case of the conditional joint probability density function, the parameter vector θ is fixed and the observation vector $\mathbf{u}$ is variable. In the case of the likelihood function, we have the opposite situation in that the parameter vector θ is variable and the observation vector $\mathbf{u}$ is fixed.

In many cases, it turns out to be more convenient to work with the natural logarithm of the likelihood function rather than with the likelihood itself. Thus, using $L(\theta)$ to denote the *loglikelihood function*, we write

$$\begin{aligned} L(\theta) &= \ln[l(\theta)] \\ &= \ln[f_{\mathbf{U}}(\mathbf{u} \mid \theta)] \end{aligned} \tag{2}$$

The logarithmic function $L(\theta)$ is a *monotonic transformation* of $l(\theta)$. This means that whenever $l(\theta)$ decreases, its logarithm $L(\theta)$ also decreases. Where $l(\theta)$ is a formula for a conditional joint probability density function, it follows that it never becomes negative; hence there is no problem in evaluating the logarithmic function $L(\theta)$. We conclude, therefore, that the parameter vector for which the likelihood function $l(\theta)$ is a maximum is exactly the same as the parameter vector for which the loglikelihood function $L(\theta)$ is a maximum.

To obtain the ith element of the maximum-likelihood estimate of the parameter vector θ, we differentiate the loglikelihood function with respect to θ_i and set the result equal to zero. We thus get a set of first-order conditions:

$$\frac{\partial L}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, K \tag{3}$$

The first derivative of the loglikelihood function with respect to the parameter θ_i is called the *score* for that parameter. The vector of such parameters is known as the *scores vector* (i.e., the gradient vector). The scores vector is identically zero at the maximum-likelihood estimates of the parameters [i.e., at the values of parameter vector θ that result from the solutions of Eq. (3)].

To find how effective the method of maximum likelihood is, we need to compute the *bias* and *variance* for the estimate of each parameter. However, this is frequently difficult to do. Thus, rather than approach the computation directly, we may derive a *lower bound* on the variance of any *unbiased* estimate. We say an estimate is unbiased if the average value of the estimate equals the parameter we are trying to estimate. Later, we show how the variance of the maximum-likelihood estimate compares with this lower bound.

3. CRAMÉR–RAO INEQUALITY

Let $\mathbf{U}$ be a random vector with conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \boldsymbol{\theta})$, where $\mathbf{u}$ is the observed sample vector with elements $u_1, u_2, \ldots, u_M$ and θ is the parameter vector with elements $\theta_1, \theta_2, \ldots, \theta_K$. Using the definition of Eq. (2) for the loglikelihood function $L(\theta)$ in terms of the conditional joint probability density function $f_{\mathbf{U}}(\mathbf{u} \mid \boldsymbol{\theta})$, we form the $K \times K$ matrix

$$\mathbf{J} = -\begin{bmatrix} E\left[\dfrac{\partial^2 L}{\partial \theta_1^2}\right] & \left[\dfrac{\partial^2 L}{\partial \theta_1 \partial \theta_2}\right] & \cdots & E\left[\dfrac{\partial^2 L}{\partial \theta_1 \partial \theta_K}\right] \\ E\left[\dfrac{\partial^2 L}{\partial \theta_2 \partial \theta_1}\right] & E\left[\dfrac{\partial^2 L}{\partial \theta_2^2}\right] & \cdots & E\left[\dfrac{\partial^2 L}{\partial \theta_2 \partial \theta_K}\right] \\ \vdots & \vdots & & \vdots \\ E\left[\dfrac{\partial^2 L}{\partial \theta_K \partial \theta_1}\right] & E\left[\dfrac{\partial^2 L}{\partial \theta_K \partial \theta_2}\right] & \cdots & E\left[\dfrac{\partial^2 L}{\partial \theta_K^2}\right] \end{bmatrix} \tag{4}$$

The matrix $\mathbf{J}$ is called *Fisher's information matrix*.

Let $\mathbf{I}$ denote the inverse of Fisher's information matrix $\mathbf{J}$. Let I_{ii} denote the ith diagonal element (i.e., the element in the ith row and ith column) of the inverse matrix $\mathbf{I}$. Let $\hat{\theta}_i$ be *any* unbiased estimate of the parameter θ_i, based on the observed sample vector $\mathbf{u}$. We may then write [2]

$$\operatorname{var}[\hat{\theta}_i] \geq I_{ii}, \quad i = 1, 2, \ldots, K \tag{5}$$

This equation is called the *Cramér–Rao inequality*. It enables us to construct a lower limit (greater than zero) for the variance of any unbiased estimator, provided, of course, that we know the functional form of the loglikelihood function. The lower limit is called the *Cramér–Rao lower bound*.

If we can find an unbiased estimator whose variance equals the Cramér–Rao lower bound, then according to Eq. (5), there is no other unbiased estimator with a smaller variance. Such an estimator is said to be *efficient*.

4. PROPERTIES OF MAXIMUM-LIKELIHOOD ESTIMATORS

Not only is the method of maximum likelihood based on an intuitively appealing idea (that of choosing those parameters from which the actually observed sample vector is most likely to have come), but the resulting estimates also have some desirable properties. Indeed, under quite general conditions, the following *asymptotic* properties may be proved [2]:

1. Maximum-likelihood estimators are *consistent*; that is, the value of θ_i for which the score $\partial L/\partial \theta_i$ is identically zero *converges in probability* to the true value of the parameter θ_i, $i = 1, 2, \ldots, K$, as the *sample size* M approaches infinity.
2. Maximum-likelihood estimators are *asymptotically efficient*:

$$\lim_{M \to \infty} \left\{ \frac{\operatorname{var}[\theta_{i,\mathrm{ml}} - \theta_i]}{I_{ii}} \right\} = 1, \quad i = 1, 2, \ldots, K \tag{6}$$

where $\theta_{i,ml}$ is the maximum-likelihood estimate of parameter θ_i and I_{ii} is the ith diagonal element of the inverse of Fisher's information matrix.

3. Maximum-likelihood estimators are *asymptotically Gaussian*.

In practice, we find that the large-sample (i.e., asymptotic) properties of maximum-likelihood estimators hold rather well for sample size $M \geq 50$.

5. CONDITIONAL MEAN ESTIMATOR

Another classic problem in estimation theory is the *Bayes estimation of a random parameter*. There are different answers to this problem, depending on how the Bayes estimation is formulated [2]. A particular type of the Bayes estimator of interest is the *conditional mean estimator*. We now wish to do two things: (1) derive the formula for the conditional mean estimator and (2) show that such an estimator is the same as a minimum mean-square-error estimator.

Toward those ends, consider a *random parameter* x. We are given an observation y that depends on x, and the requirement is to estimate x. Let $\hat{x}(y)$ denote an *estimate* of the parameter x; the symbol $\hat{x}(y)$ emphasizes the fact that the estimate is a function of the observation y. Let $C(x, \hat{x}(y))$ denote a *cost function* that depends on both x and $\hat{x}(y)$. Let E denote the statistical expectation operator. Then according to Bayes' estimation theory, we may write the expression

$$\begin{aligned}\mathcal{R} &= E[C(x, \hat{x}(y))] \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} C(x, \hat{x}(y)) f_{X,Y}(x, y)\, dy \end{aligned} \tag{7}$$

for the risk [2]. Here, $f_{X,Y}(x, y)$ is the joint probability density function of x and y. For a specified cost function $C(x, \hat{x}(y))$, the *Bayes estimate* is defined as the estimate $\hat{x}(y)$ that *minimizes* the risk $\mathcal{R}$.

A cost function of particular interest is the mean-square error, specified as the square of the estimation error, which is itself defined as the difference between the actual parameter value x and the estimate $\hat{x}(y)$:

$$\varepsilon = x - \hat{x}(y) \tag{8}$$

Correspondingly, the cost function is defined by

$$C(x, \hat{x}(y)) = C(x - \hat{x}(y))$$

or simply

$$C(\varepsilon) = \varepsilon^2 \tag{9}$$

Thus the cost function $C(\varepsilon)$ varies with the estimation error ε in the manner indicated in Fig. 1. It is assumed here that x and y are both real. Accordingly, we may rewrite Eq. (7) for mean-square error as

$$\mathcal{R}_{\text{ms}} = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} [x - \hat{x}(y)]^2 f_{X,Y}(x, y)\, dy \tag{10}$$

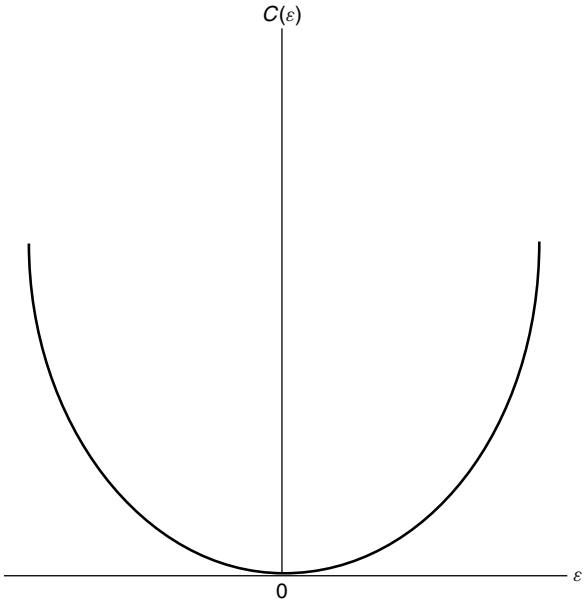


Figure 1. Mean-square error as a quadratic cost function.

where the subscripts ms in the risk $\mathcal{R}_{\text{ms}}$ indicate the use of mean-square error estimation. From probability theory, we have

$$f_{X,Y}(x, y) = f_X(x \mid y) f_Y(y) \tag{11}$$

where $f_X(x \mid y)$ is the conditional probability density function of x given y, and $f_Y(y)$ is the (marginal) probability density function of y. Hence, using Eq. (11) in Eq. (10), we may write

$$\mathcal{R}_{\text{ms}} = \int_{-\infty}^{\infty} dy f_Y(y) \int_{-\infty}^{\infty} [x - \hat{x}(y)]^2 f_X(x, y)\, dx \tag{12}$$

We now recognize that the inner integrand and the probability density function $f_Y(y)$ in Eq. (12) are both nonnegative. We may therefore simplify matters by minimizing the inner integral. Let the estimate so obtained be denoted by $\hat{x}_{\text{ms}}(y)$. We find $\hat{x}_{\text{ms}}(y)$ by differentiating the inner integral with respect to $\hat{x}(y)$ and then setting the result equal to zero. To simplify the minimization procedure, let I_{inner} denote the inner integral in Eq. (12). Then differentiating I_{inner} with respect to the estimate $\hat{x}(y)$ yields

$$\frac{dI_{\text{inner}}}{d\hat{x}} = -2 \int_{-\infty}^{\infty} x f_X(x \mid y)\, dx + 2\hat{x}(y) \int_{-\infty}^{\infty} f_X(x \mid y)\, dx \tag{13}$$

The second integral on the right-hand side of Eq. (13) represents the total area under a probability density function, which, by definition, equals unity. Hence, setting the derivative $dI_{\text{inner}}/d\hat{x}$ equal to zero and solving for the minimum mean-square error estimate, denoted by, we obtain

$$\hat{x}_{\text{ms}}(y) = \int_{-\infty}^{\infty} x f_X(x \mid y)\, dx \tag{14}$$

The optimum solution defined by Eq. (14) is unique by virtue of the assumed form of the cost function $C(\varepsilon)$. For another interpretation of the estimator $\hat{x}_{\text{ms}}(y)$, we

recognize that the integral on the right-hand side of the equation is just the *conditional mean* of the parameter x, given the observation y. On the basis of this result, we therefore conclude that *the minimum mean-square-error estimate and the conditional mean estimator are indeed one and the same*. In other words, we have

$$\hat{x}_{\text{ms}}(y) = E[x \mid y] \tag{15}$$

Substituting Eq. (15) for the estimate $\hat{x}(y)$ into Eq. (12), we find that the inner integral is just the conditional variance of the parameter x, given y. Accordingly, the minimum value of the risk $\mathcal{R}_{\text{ms}}$ is just the average of this conditional variance over all observations y.

6. EXPECTATION–MAXIMIZATION (EM) ALGORITHM

The *expectation–maximization algorithm*, popularly known as the *EM algorithm*, is an iterative algorithm for computing maximum-likelihood estimates when dealing with data that have a latent structure and/or are incomplete. Moreover, computation of the maximum-likelihood estimate is often greatly facilitated by formulating it as an incomplete data problem, which is invoked because the EM algorithm is able to exploit the reduced complexity of the maximum-likelihood estimate, given the complete data. Applications of the EM algorithm include hidden Markov models for speech recognition and hierarchical mixture of experts model for the design of neural networks.

The EM algorithm derives its name from the fact that, at each iteration of the algorithm, there are two basic steps [3,4]:

- *Expectation step* or *E-step*, which uses the given data set of an incomplete data problem and the current value of the parameter vector to manufacture data so as to postulate an augmented or so-called complete data set.
- *Maximization step* or *M-step*, which consists of deriving a new estimate of the parameter vector by maximizing the loglikelihood function of the complete data manufactured in the E-step.

The E-step, operating in the forward direction, and the M-step, operating in the backward direction, form a closed loop. Thus, starting from a suitable value for the parameter vector, the E-step and M-step are repeated on an alternating basis until convergence occurs.

Let the vector $\mathbf{z}$ denote the missing or hidden data. Let $\mathbf{r}$ denote the complete data vector, made up of some observable data d and the missing data vector $\mathbf{z}$. There are therefore two data spaces $\mathcal{R}$ and $\mathcal{D}$ to be considered, and the mapping from $\mathcal{R}$ to $\mathcal{D}$ is many-to-one. However, instead of observing the complete data vector $\mathbf{r}$, we are actually able to observe only the complete data $d = d(\mathbf{r})$ in $\mathcal{D}$. Let $f_c(\mathbf{r} \mid \theta)$ denote the conditional probability density function (pdf) or $\mathbf{r}$, given a parameter vector θ. It follows therefore that the conditional PDF of random variable D, given θ, is defined by

$$f_D(d \mid \theta) = \int_{\mathcal{R}(d)} f_c(\mathbf{r} \mid \theta)\, d\mathbf{r} \tag{16}$$

where $\mathcal{R}(d)$ is the subspace of $\mathcal{R}$ that is determined by $d = d(\mathbf{r})$. The EM algorithm is directed at finding a value of θ that maximizes the *incomplete data loglikelihood function*

$$L(\theta) = \log f_D(d \mid \theta) \tag{17}$$

This problem, however, is solved indirectly by working iteratively with the *complete data loglikelihood function*

$$L_c(\theta) = \log f_c(\mathbf{r} \mid \theta) \tag{18}$$

which is a random variable, because the missing data vector $\mathbf{z}$ is unknown.

To be more specific, let $\hat{\theta}(n)$ denote the value of the parameter vector θ on iteration n of the EM algorithm. In the E-step of this iteration, we calculate the expectation

$$Q(\theta, \hat{\theta}(n)) = E[L_c(\theta)] \tag{19}$$

where the expectation is performed with respect to $\hat{\theta}(n)$. In the M-step of this same iteration, we maximize $Q(\theta, \hat{\theta}(n))$ with respect to θ over the parameter (weight) space $\mathcal{W}$, and so find the updated parameter estimate $\hat{\theta}(n+1)$, as shown by

$$\hat{\theta}(n+1) = \arg\max_{\theta} Q(\theta, \hat{\theta}(n)) \tag{20}$$

The algorithm is started with some initial value $\hat{\theta}(0)$ of the parameter vector θ. The E-step and M-step are then alternately repeated in accordance with Eqs. (19) and (20), respectively, until the difference between $L(\hat{\theta}(n+1))$ and $L(\hat{\theta}(n))$ drops to some arbitrary small value; at that point the computation is terminated. Note that after an iteration of the EM algorithm, the incomplete data loglikelihood function is *not* decreased, as shown by

$$L(\hat{\theta}(n+1) \geq L\hat{\theta}(n)) \quad \text{for } n = 0, 1, 2, \ldots \tag{21}$$

Equality usually means that we are at a stationary point of the loglikelihood function.

Under fairly general conditions, the loglikelihood function computed by the EM algorithm converges to stationary values. However, a cautionary note is in order. The EM algorithm will not always lead to a local or global maximum of the loglikelihood function. This point is demonstrated in Chapter 3 of the book by McLachlan and Krishnam [4]; in one of two examples presented therein, the EM algorithm converges to a saddle point, and in the other example the algorithm converges to a local minimum of the loglikelihood function.

7. DISCUSSION

In this article, we presented a description of maximum-likelihood (ML) estimation, which, in mathematical terms, corresponds to the limiting case of maximum a posteriori probability (MAP) estimation when the a priori knowledge pertaining to the problem at hand approaches zero. ML estimates have some nice asymptotic properties as the size of the data set approaches infinity. Indeed, it is these properties that motivate the use of ML estimates even when there is no efficient estimate.

We also briefly described a forward–backward computation procedure known as the EM algorithm, which is

remarkable in part because of the simplified and generality of the underlying theory, and in part because of the wide range of applications that fall under its umbrella. The EM algorithm applies to incomplete data problems. Problems of this kind encompass situations where naturally there are hidden variables, and other situations where the incompleteness of data is not at all evident or natural to the problem of interest.

BIOGRAPHY

Simon Haykin received the degrees of B.Sc. (First Class Honours), Ph.D., and D.Sc., all in electrical engineering from the University of Birmingham, England. On the completion of his Ph.D. studies, he spent several years from 1956 to 1965 in industry and academe in England. In January 1966, he joined McMaster University, Hamilton, Ontario, Canada, as Full Professor of Electrical Engineering; he has stayed there since. In 1996, the Senate of McMaster University established the new title of University Professor; in April of that year, he was appointed the first University Professor from the Faculty of Engineering.

Professor Haykin is a Fellow of the IEEE and a Fellow of the Royal Society of Canada. In 1999 he was awarded the honorary degree of Doctor of Technical Sciences by ETH, Zurich, Switzerland.

Professor Haykin's research interests have focused on adaptive signal processing, for which he is recognized worldwide.

BIBLIOGRAPHY

1. R. A. Fisher, Theory of statistical estimation, *Proc. Cambridge Phil. Soc.* **22**: 700–725 (1925).
2. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, Part I, Wiley, 1968.
3. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc., B* **39**: 1–38 (1977).
4. G. J. McLachlan and T. Krishnam, *The EM Algorithm and Extensions*, Wiley, 1997.

MEDIUM ACCESS CONTROL (MAC) PROTOCOLS

ANDRÁS FARAGÓ
VIOLET R. SYROTIUK
University of Texas at Dallas
Richardson, Texas

1. INTRODUCTION

A number of communication networks use a *broadcast* (or *multiaccess*) *transmission medium*, where the signal transmitted by a node (station) is received by every other node that is in the listening area of the transmitting node. The most frequently occurring examples of such networks are wired or wireless local-area networks (LANs), and radio networks that include, as examples, satellite networks, wireless cellular networks, and mobile ad hoc radio networks (also called *packet radio networks*).

Because of the nature of the broadcast medium and the technological constraints of network nodes, the transmissions have to be controlled to ensure successful communication. Examples of typical technological constraints are (1) a node cannot both transmit and receive at the same time; and (2) a node can receive only one transmission at a time — in case of more than one simultaneous transmission, a *collision* occurs and nothing is received successfully. The constraints vary by technology and also by network type. The general task of the *medium access control* (MAC) *protocol* is to organize the transmissions such that under the given constraints successful communication takes place over the broadcast transmission medium.

The MAC protocol is fundamental to the ability to communicate in, and the performance of, networks based on broadcast channels. It is thus a vast and well-studied topic because of the wide range of broadcast media and the importance of the task. A large number of MAC protocols have been proposed in the literature, and quite a few of them are standardized and widely deployed in different commercial networks. In the next section we classify MAC protocols according to various criteria. Then, in subsequent sections, we provide a more detailed description of the most important protocols, categorized according to the way they implement the multiaccess communication.

[*Remark*: In point-to-point (rather than broadcast) communication-based networks, similar protocols are used for *multiplexing* several signals onto the same link. In this article, however, we discuss only MAC protocols and do not address multiplexing protocols.]

2. CLASSIFICATION OF MAC PROTOCOLS

MAC protocols can be classified by a number of different criteria, which we overview below. Details of the most important protocols are then given in the following sections.

2.1. Classification by Physical Domain of Sharing

2.1.1. Time Domain. Here the transmissions are organized in time, using the same frequency band. The time may be either *slotted* or *unslotted*. In slotted time, time is discrete and packet transmission can begin only at slot boundaries; it assumes that the nodes are synchronized, which is itself a challenging problem in a distributed setting. In unslotted or continuous time, there is no restriction on when packet transmission can begin. The key problem is how to organize the transmissions, such that collisions are either avoided or resolved with repeated transmissions.

2.1.2. Frequency Domain. If different stations can use different frequencies, then it is possible to separate the transmissions in the frequency domain by filtering or by tuning to a given frequency band. A traditional example is radio broadcasting.

2.1.3. Hybrid Domains. In several networks the physical domain of sharing is a combination of time and frequency. The most important group of hybrid techniques is *spread-spectrum* communication, commonly referred to as *code-division multiple access* (CDMA). Here we can symbolically say that the "code domain" is shared. Another example is the MAC protocol in the European digital cellular radio *Global System for Mobile Communication* (GSM) standard that combines a complex slotted timeframing hierarchy within 124 frequency channels [1].

2.2. Classification by Method of Sharing

2.2.1. Allocation-Based Protocols. In this category the transmission rights are allocated in advance to the nodes to avoid collisions. A typical example is *time-division multiple access* (TDMA) with a fixed assignment of time slots to users that they can use for transmission. If there is a change in the network topology or traffic pattern, reallocation of the slots may take place, but the standard operation is based on pre- or reallocated transmission rights.

2.2.2. Contention-Based Protocols. In these protocols the operation is fully distributed and there are no preallocated transmission rights. The stations *contend* for transmission, attempting it whenever needed. As a consequence, *collisions* may occur that make retransmission necessary. The key part of the protocol is how *collision resolution* is done to ensure that with appropriately organized retransmissions eventually all users have acceptable throughput and delay. A popular example is the IEEE 802.11 Wireless LAN protocol.

2.2.3. Hybrid Schemes. Many protocols exhibit both allocation- and contention-based features. For example, in *reservation*-based protocols there is a contention phase for reserving the channel. Once the reservations are made, the data are then transmitted as if the access were allocation-based.

2.3. Classification by Mode of Operation

2.3.1. Centralized Protocols. These protocols require a central entity that controls the channel access of all nodes in a given area. For example, in a cellular network the base station in a cell plays this role; in a satellite network the satellite provides control in its coverage area.

2.3.2. Distributed Protocols. If no central entity is available or desired, then the operation of the protocol must be distributed. This is necessary, for example, in mobile ad hoc networks and in most LANs. Contention based protocols typically operate in a distributed fashion.

2.4. Classification by Network Characteristics

The MAC protocols that are practically applied are tailored to the type of network in which they are used. Some typical factors that depend on the network characteristics are listed below.

2.4.1. Connectivity. In most wired LANs the network is logically fully connected; that is, every transmission is received by all stations. On the other hand, mobile wireless ad hoc networks have lower connectivity, which introduces new problems such as the *hidden-* and *exposed-terminal* problems (see Fig. 1). The hidden-terminal problem occurs when the destination B of a transmitting node A suffers a collision because of an interfering transmission from another node C that is not in the range of A. In this case, C is hidden from A. In the exposed-terminal problem, if node B is transmitting to A, C can transmit concurrently with B as long as its destination is not in the overlapping transmission range of B and C (e.g., node D). However, without additional information, node C cannot make this determination.

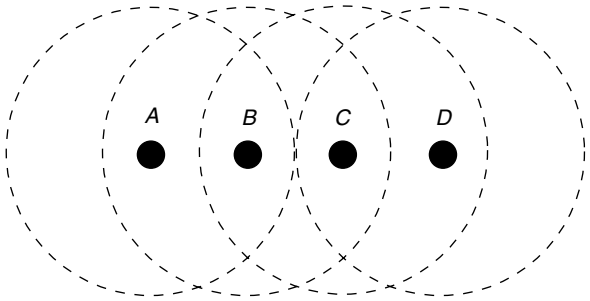


Figure 1. Each large circle indicates the transmission range of the node at its center.

2.4.2. Propagation Time. The performance of the MAC protocol is seriously influenced by the relationship of the propagation delay D and the packet transmission time T. This is often expressed by the ratio D/T. In most LANs, mobile ad hoc networks and cellular networks the ratio is small, around 10^{-2}. This implies that a collision can be quickly detected. On the other hand, in satellite networks the ratio can be as large as 100, so many packets may be sent before a collision is detected.

2.4.3. Physical Link Characteristics. Wireless links behave substantially differently from wired links. The signal-to-noise ratio is lower, while the bit error rate and signal attenuation are typically higher in wireless links. This also has an influence on optimizing the MAC protocol for a given network. An example is the *near–far* problem when the stronger signal of a node nearby the destination suppresses the weaker signal of a remote station. Instead of a collision being detected at the receiver, the closer node "captures" the receiver. As a result, the remote node may not be helped out by collision resolution.

2.5. Summary

As seen above, there are a number of ways to categorize the large number of existing MAC protocols. In the subsequent sections we review some of the most important MAC protocols, according to their method of sharing the medium (allocation- or contention-based, or a hybrid of these). We choose this categorization for our presentation because the most characteristic aspect of a MAC protocol is how it actually implements the sharing of the multiaccess medium.

The traditional performance measures of a MAC protocol are its throughput and delay characteristics at

various traffic load conditions. In this brief overview of MAC protocols, there is no opportunity to introduce the models and develop the mathematical foundations for the performance analysis of each protocol. The interested reader should consult the book by Bertsekas and Gallager [2] for a good introduction to the analysis of MAC protocols.

3. ALLOCATION-BASED MAC PROTOCOLS

In allocation-based MAC protocols parts of the communication resources (such as time or frequency) are assigned in advance to the stations in a way that excludes collision, so there is no contention for transmission. Below we discuss some of the most important allocation based MAC protocols.

3.1. Time-Division Multiple Access (TDMA)

In TDMA protocols time is slotted and each slot can incorporate the transmission of one packet. The key issue is how to allocate the slots to stations, such that no collision occurs. Typically, two constraints have to be satisfied: (1) a node cannot both transmit and receive in the same slot (primary conflict); and (2) no successful reception is possible if more than one transmission reaches the node in a slot (secondary conflict).

If the network is logically fully connected, that is, if each transmission is received by all stations, then the allocation is conceptually very simple — the time slots are grouped into frames and each node has its own unique slot in each frame. The frames are periodically repeated, so if there are N nodes, then each one has the opportunity to transmit once in each frame of N slots.

The above mentioned simple frame structure assumes that we want to give all nodes an equal chance to transmit. If, however, different nodes generate different amounts of traffic, then they may be assigned a different number of slots in a longer frame. The assignment is called *fair* if each node is assigned a number of slots proportional to its traffic rate, that is, the average number of packets per unit time that the node wants to transmit. On the other hand, if packets are generated randomly, some slots may remain unused at certain nodes, while at others the buffer may overflow. Thus, one may also want to assign the slots so that the *throughput*, that is, the average number of successful packet transmissions per slot, is maximized. It is interesting that one can prove under general modeling assumptions the maximum throughput is achieved precisely when the assignment is fair [3].

The slot assignment problem is more complicated in mobile ad hoc networks, where the network topology is not fully connected and furthermore, can change over time. In these networks it is possible that two nodes that are more than two hops away in the network topology can be assigned the same time slot without the danger of any conflict. (Here, a "hop" refers to nodes within direct transmission range of a node — since all nodes are not one hop away from each other in a mobile ad hoc network it is often called a *mobile multihop network*.) This makes it possible to achieve *spatial reuse* of the available spectrum. An example of such a conflict-free slot assignment for the multihop network is shown in Fig. 2. Notice that the shorter frame lengths must inter-operate with longer frame lengths.

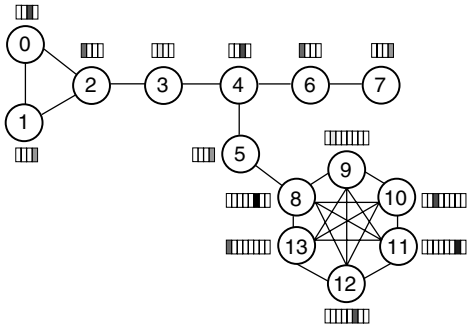


Figure 2. A conflict-free slot assignment for a multihop network.

A number of algorithms were proposed to find good slot assignments for spatial reuse TDMA [4,5]. The task can be mathematically modeled by a *graph coloring* problem where nodes of the same color can transmit concurrently. This is an algorithmically difficult (NP-complete) problem even for the restricted graphs that can occur as mobile ad hoc network topologies [6]. Nevertheless, acceptable solutions can be found with simple heuristics, such as given by a greedy algorithm.

The advantage of TDMA protocols is that they guarantee a certain throughput via the allocated transmission rights. On the other hand, for low traffic loads, TDMA introduces unnecessary delays, since a station has to wait for its turn even if others are not transmitting.

3.2. Frequency-Division Multiple Access (FDMA)

FDMA assigns a different frequency to each station. Since this makes it possible to separate the transmissions in the frequency domain, they do not have to be separated in time. In its pure form FDMA is best suited for analog systems. It has been used for a long time in radio and TV broadcasting. Note that broadcasting involves spatial reuse, since beyond the coverage area of a radio station its frequency can be reused. The way FDMA is used in broadcasting can only serve relatively few transmitters, as the radio spectrum is a scarce resource. FDMA is a component in a number of MAC protocols in cellular telephony, such as in the GSM system, which combines TDMA with FDMA [1].

3.3. Code-Division Multiple Access (CDMA)

One basic form of CDMA is *frequency hopping* (FH/CDMA). In this system there are a number of channels on different frequencies and the transmitter quickly hops between the different channels in a pseudo-random manner, as illustrated in Fig. 3. In this way the signal energy is spread over a larger frequency band, hence this technique is also called *spread-spectrum* communication. The spreading takes place according to a *spreading code* that specifies the hopping schedule. If the receiver "knows" the spreading code, then it can successfully receive the signal by following the same hopping schedule among the channels. Without the correct code, however, only noise is received, since the

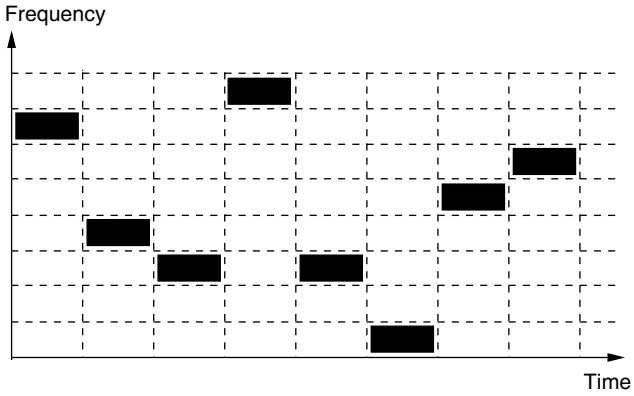


Figure 3. Frequency-hopping code-division multiple access (FH/CDMA).

signal is hidden in a broad frequency band in which the overall noise may be stronger than the useful signal that always falls into a narrow band.

Another basic form of CDMA is *direct-sequence* CDMA (DS/CDMA). Here each bit is replaced by a pseudo-random sequence of bits, which now represents the code (sometimes called a *chip sequence*). The receiver correlates the received sequence with the known code. If it is the same code, then the correlation is high; otherwise it is small (in case of orthogonal codes, it is zero). For example, assume that nodes A, B, and C are assigned the following codes, respectively: $(+1, +1, -1, -1)$; $(+1, +1, +1, +1)$; $(+1, -1, +1, -1)$. To transmit a binary one (1), a node transmits its code c; otherwise it transmits the negation of its code $\overline{c}$. When two or more nodes transmit simultaneously, their codes add linearly. To recover the transmission of a specific node, a receiver computes the normalized inner product of that node's code with the incoming signal. Since all of code pairs are orthogonal, all signals except that from the specific node are eliminated. For example, suppose that node D wants to recover the transmission from node C, and that node A transmits a zero and nodes B and C transmit ones at the same time. Node D then computes

$$C \cdot (\overline{A} + B + C) = C \cdot \overline{A} + C \cdot B + C \cdot C = 0 + 0 + 1$$

and thus node D recovers that node C transmitted a 1.

In both CDMA systems the appropriate choice of codes makes it possible for a receiver to lock onto a transmission even if other transmissions are present. This is due to the fact that, in case of orthogonal codes, the correlation with the interfering transmissions is zero, so they are effectively filtered out from the aggregated received signal. The number of orthogonal codes, however, is limited by the available spectrum. The number of codes can be significantly increased if we do not insist on full orthogonality. In this case the correlation with the interfering transmission will not be zero, but can be kept small as long as there are not too many interfering transmissions. In any case, since the codes are allocated in advance, CDMA is an allocation-based protocol, at least in its basic forms.

CDMA has several advantages. It is resistant to jamming and interception, which makes it desirable for tactical applications.[1] With appropriately designed codes, it can work without global synchronization. A characteristic feature is that CDMA has no hard limit on capacity, since the increasing number of interfering packets results in degrading signal-to-noise ratio, but not in sudden breakdown. In this way the network shows *graceful degradation* in case of increasing traffic load. In CDMA cellular networks frequency planning is easy, since each cell uses the same frequency band. Handoff is also easier and graceful degradation is an attractive feature. On the other hand, CDMA has some disadvantages, too. Its implementation technology is complex in addition to requiring power control and a large contiguous frequency band.

IS-95 CDMA is a digital cellular radio system that is used in over 35 countries worldwide. Some CDMA systems operate in the personal communications systems (PCS) frequency band.

3.4. Centralized and Distributed Polling, Token Ring

A simple allocation-based MAC solution is *polling*, when a master station polls each node in a round-robin manner to see whether it has a packet to transmit. The node that is being polled can send the packet to the master or directly to another node. This is similar to the sharing philosophy of TDMA in a fully connected network. The essential difference is, however, that if a station has nothing to send, then there is no need to wait; the master can immediately poll the next station. The simple round-robin polling scheme can be improved at the price of added complexity. There are more sophisticated polling strategies that perform a logarithmic search for a station with a packet to send (see Ref. 2, Section 4.5.6).

Polling is a centralized protocol that is also allocation-based; since the master station allocates the transmission rights, no contention is involved. To create a distributed version of polling, one can observe that it is not the master station that is essential in the protocol. What is really needed is the rotating transmission right that goes from station to station in a cyclic manner, and if a node has nothing to send, then the transmission right is quickly passed to the next station. This is the core idea of the *token-ring* protocol. In a token-ring network the nodes are arranged in a ring. Conceptually, the operation can be described such that there is a rotating token in the network that is passed from node to node. Only the station that currently has the token is allowed to transmit a packet, after which point it passes the token to the next node. If the station has no packet to transmit, then it passes the token immediately. This operation achieves, conceptually, the same effect as round-robin polling, but in a distributed way.

[1] Let us cite an interesting historical remark from Ref. 7, p. 168: "Spread spectrum (using frequency hopping) was invented, believe it or not, by Hollywood screen siren Hedy Lamarr in 1940 at the age of 26. She and a partner who later joined her effort were granted a patent in 1942 [U.S. Patent 2,292,387; Aug. 11, 1942]. Lamarr considered this her contribution to the war effort and never profited from her invention."

Of course, an implementation of the token ring has to pay attention to a number of practical issues, such as what information is put in the token (a type of control packet), how to recover from a token loss or failure, and how to recover from a node failure. The details of the protocol are standardized in the IEEE 802.4 (token bus) and IEEE 802.5 (token ring) standards [8]. In the token bus protocol, the nodes are logically arranged in a ring using a distributed algorithm to establish and maintain the ring. The token ring, on the other hand, is a physical ring, although it is common to wire the nodes to a wire center to permit the ring to remain functional in the presence of node failures or maintenance.

A higher-performance fiberoptic token ring is *Fiber Distributed Data Interface* (FDDI). The primary difference between FDDI and IEEE 802.5 is that the token is put on the ring immediately after the transmission of a packet rather than after the source drains its own packet from the ring. Thus, in FDDI, it is possible to have multiple packets on the ring at the same time, resulting in higher throughput in the presence of higher transmission rates and larger distances. FDDI is used primarily in backbone networks.

IEEE 802.4, IEEE 802.5, and FDDI are all capable of handling several priority classes of traffic with guaranteed throughput and delay [2], something that is not possible in contention based protocols.

4. CONTENTION-BASED MAC PROTOCOLS

While allocation-based MAC protocols have the advantage that they can guarantee a certain throughput and, therefore, can prevent complete breakdown of the network due to congestion, they are not very efficient for light-traffic-load situations. If the traffic load is light, a node, rather than waiting for its turn, can attempt transmission immediately when it has a packet to transmit. This is how contention-based protocols operate. Even though this savings in delay can cause collision, if the traffic load is light, the probability of collision is low. In case a collision still occurs, the protocol resolves it by resending the packet later, possibly trying multiple times. The heart of a contention based MAC protocol is in how it implements collision resolution.

Another important advantage of contention-based protocols is that they can automatically adapt to changing network topology in ad hoc networks, as opposed to TDMA that may need networkwide frame and slot reassignment whenever there is a change in the topology. On the other hand, contention protocols do not guarantee a deterministically bounded delay. Below we review some of the fundamental contention based MAC protocols.

4.1. ALOHA and Its Variants

ALOHA is the historically first contention-based MAC protocol [9], an influential milestone in MAC protocol history. It has slotted and unslotted versions, depending on whether packet transmission can be started only at time-slot boundaries or at any time, respectively. Let us explain the operating principle through the slotted version (the unslotted version is conceptually similar, only that a packet is vulnerable to collision for longer time periods).

When a node has a packet to transmit, it simply transmits it in the first available time slot. If the transmission is successful (which the node may know, e.g., from an acknowledgment sent on a separate channel or piggybacked onto a response), then there is nothing else to do with respect to that packet. If, however, the transmission is not successful, such as when a collision or transmission error occurs, then the node retransmits the packet with a *random delay*, that is, after waiting a random number of slots. In other words, the node backs off and then tries the transmission again after a random waiting time. Nothing excludes that the packet collides on successive transmission attempts, but this has lower probability, given that the network is not overloaded. Ultimately, after a number of retransmissions the packet has a very good chance to get through.

A important issue is to decide how to draw the random delay, that is, what should be the *backoff scheme*. A simple variant is when transmission occurs in each slot with a given probability p. This is called p-persistent ALOHA, and it corresponds to a geometrically distributed random delay. Another popular scheme is *binary exponential backoff*, where the next transmission slot is drawn uniformly at random from an interval and after each unsuccessful trial the length of the interval is doubled. In real networks, such as Ethernet, if the interval reaches a maximum length (1024), it remains fixed at that length. If the transmission is still unsuccessful at the maximum length, after some number (16) of collisions, failure is reported to and handled by the higher-layer protocol.

The analysis of various backoff schemes has been the subject of intense research, and it is not easy to determine which is the best one. For example, despite its wide usage, it is known that binary exponential backoff results in an unstable protocol (infinitely growing queues) under certain modeling assumptions, such as infinite user population [10]. The existence of stable protocols in this setting also depends on the type of feedback available from the channel and on how the user population is modeled. For acknowledgment-based protocols, it is known [11] that a large class of backoff, schemes, including polynomial backoff, is unstable in the infinite user population model (in polynomial backoff the backoff interval grows according to a polynomial function rather than an exponential function). In contrast, for a finite user population, any superlinear polynomial backoff protocol has been proved stable, while binary exponential backoff still remains unstable above a certain arrival rate [12].

4.2. CSMA and Its Variants, CSMA/CD and CSMA/CA

A natural improvement of ALOHA is possible if the stations can sense before transmission whether the channel is idle. After *carrier sensing*, a station starts transmitting only if the channel is idle, since otherwise the packet would surely collide with the ongoing transmission. This is the basis for the *Carrier Sense Multiple Access* (CSMA) protocol, which otherwise operates similarly to ALOHA. If the network had zero propagation delay, then collision could occur only if two nodes start transmission of a packet at precisely the same time. This has zero probability in continuous time. With finite propagation

delay, however, a node may sense that the channel is idle even if another node has already started transmitting, but due to the propagation delay, the signal has not reached yet the first node.

Further improvement to CSMA is possible via *collision detection* (CD); if a transmitting node detects collision, then it stops transmitting, since the rest of the packet transmission time is just wasted. CSMA/CD with binary exponential backoff is the basis for the MAC protocol in the IEEE 802.3 LAN standard [8]. This standard, popularly known as *Ethernet*, is by far the most widely used protocol in LANs today. Because of its widespread success, Fast Ethernet (IEEE 802.3u) did not change the protocol; it only made it run faster, with the faster technology. Another variation of CSMA uses *collision avoidance* rather than collision detection. CSMA/CA is commonly used in wireless networks since collision detection is not commonly available in wireless nodes. (Collision detection requires that a node can both transmit and receive simultaneously.) We discuss CSMA/CA in Section 5.1.

4.3. Splitting Algorithms

A more sophisticated collision resolution technique with higher theoretical performance is implemented in another class of protocols, called *splitting algorithms* (see Ref. 2, Section 4.3). The basic principle is described as follows. If in slot k a collision occurs, then the stations that are not involved in the collision go into a waiting mode. The involved nodes split into two roughly equally sized groups, for example, by drawing a random bit ("flipping a coin"). The first group attempts retransmission in slot $k+1$, while the second group in slot $k+2$. If there is no new collision (because only half of the nodes are involved), then the collision has been resolved. If there is a new collision in slot $k+1$, then all involved nodes hear it and then the first group starts resolving it using the same protocol recursively for the halved group. During this time the second group waits, since it can sense the collisions. After the first group's collisions have been resolved, the second group runs the same algorithm for itself, again recursively for a group that is half that of the original. In the worst case, the splitting continues until the group contains only one member, in which case no collision is guaranteed in the next slot. This special example of splitting protocols is called a *tree algorithm*.

5. HYBRID SCHEMES

Quite a few MAC protocols try to combine the advantages of allocation and contention. In Sections 5.1 and 5.2 we review a few interesting solutions.

5.1. Reservation and Collision Avoidance

If the network has a master station that can coordinate the operation of the other stations, such as in satellite and cellular networks, then a good way of getting rid of the wasted bandwidth caused by collisions is to make *reservations* with the master station. A typical solution is to dedicate a certain time period for making the reservation requests, and then the stations to which the master grants reservation can send their data without collision in an allocated period of time. There are different ways to make the reservations. For example, in the *Packet Demand Assignment Multiple Access* (PDAMA) scheme the stations contend for *reservation minislots* using slotted ALOHA. Then the master computes a transmission schedule and announces it to all stations [13]. This is clearly a combination of contention and allocation, where the contention phase is restricted to the reservation minislots that take only a small percentage of time. Since reservation messages are short, if the number of stations is not too large, contention can be fully eliminated, as in the *Fixed Priority Oriented Demand Assignment* (FPODA) protocol, where each station is assigned its own reservation minislot [13]. The reservation concept goes back to the beginnings of MAC protocol development, as *Reservation ALOHA* (*R-ALOHA*) was already proposed in the early 1970's [14]. Since then, many variants have been used in satellite and cellular systems [15]. Notice that even spatial reuse TDMA protocols, described in Section 3.1 use a contention period to recompute TDMA schedules in the presence of mobility.

If the network has no central master station or if it is not fully connected, reservation is not feasible. Then the *collision avoidance* (CA) technique can be applied to reduce the chance for collision. In the CSMA/CA protocol each station is assigned a time called an *interframe spacing* (IFS), which is related to the propagation delay. Additionally, lower-priority nodes are assigned a longer IFS. Before a station transmits, it waits for an IFS time. If a higher-priority station wanted to transmit simultaneously, then the lower priority station, because of its longer IFS, can already sense that the channel is busy and refrains from transmission, so collision in this case can be avoided. After waiting an IFS time the station sets a contention timer to a random value and starts counting down. When the timer expires, transmission is attempted. If during the countdown the node senses that another node transmits a packet, then the timer is frozen until the packet is completed and then the countdown continues.

In mobile ad hoc networks additional difficulties arise, due to the irregular topology, such as the *hidden-terminal problem* mentioned in Section 2. A possible solution is provided by the *Busy-Tone Multiple Access* (BTMA) protocol [16], where a node while receiving a packet transmits another signal ("busy tone") on a separate channel, thus informing all nodes in its range that they should refrain from transmission to prevent collision at the receiver. This solution also found application in cellular networks, in the *Cellular Digital Packet Data* (CDPD) standard, which provides an overlay packet mode access over circuit mode cellular networks [13].

A problem with BTMA and its variants is that a separate frequency band is needed for the busy tone. Radio propagation characteristics, however, depend on frequency, so the range for data and for the busy tone may not coincide, causing problems with protocols using multiple channels. This problem is solved by the *Multiple Access Collision Avoidance* (MACA) protocol [17] with a handshake of control packets. In MACA, the sender node A first sends a *request-to-send* (RTS) control packet

addressed to the intended receiver B, which replies with a *clear-to-send* (CTS) packet. On receiving the CTS, node A sends the data packet. If another node, C, also wanted to send data to B, then C, hearing the CTS of B, will know that B is busy, so it can refrain from transmission for the duration of the packet, which is included in the RTS control packet and copied into the CTS. If there is a collision of RTS packets at a destination, both senders use binary exponential backoff.

After a number of protocols applied the RTS/CTS concept, this line of development culminated in the IEEE 802.11 Wireless LAN standard [8]. In particular, the *distributed coordination function* (DCF) of IEEE 802.11 is a CSMA/CA access method utilizing several different IFSs. The standard also incorporates an optional access method called a *point coordination function* (PCF), which is essentially a centralized polling scheme. The DCF and PCF can coexist by having the two methods alternate, with a contention-free period followed by a contention period. HIPERLAN (*hi*gh-*per*formance *r*adio *l*ocal-*a*rea *n*etwork) is a set of wireless LAN standards used primarily in Europe [18]. There are two specifications (HIPERLAN1 and HIPERLAN2), which provide the features and capabilities similar to those of IEEE 802.11.

5.2. Combinations of Allocation and Contention

There are a number of other creative ways of combining the advantages of allocation and contention. The *Time-Spread Multiple Access* (TSMA) protocol [19] uses a fixed slot assignment, as in TDMA, but each node is assigned several slots in a frame. These slots are chosen by means of an algebraic method (finite, or Galois, fields), such that even if some of the transmissions may collide, eventually each node can successfully send a message in each frame; that is, collisions are resolved in a deterministic way, even though the frame length is much shorter than in TDMA. In an ad hoc network TSMA can provide deterministically bounded delay, which does not require rescheduling even if the network topology changes. It assumes, however, that the maximum nodal degree (number of neighbors of a node) remains bounded. This constraint is relaxed with the further idea of *protocol threading* [20] that interleaves several transmission schedules with different parameters.

Another combination is the *ADAPT* protocol [21], which utilizes the fact that in a TDMA schedule not all assigned slots are actually used for sending a packet, due to the random (bursty) nature of traffic. If a node leaves its assigned slot unused, then other nodes can sense this and can contend for the slot using an RTS/CTS based contention protocol. The CATA protocol [22] also incorporates contention within a slot, and provides explicit support for unicast, multicast, and broadcast packet transmissions. However, it is subject to instability at high traffic load due to the lack of a fixed frame length.

A very different type of combination is implemented in the *Meta-MAC* protocol [23]. Here a master protocol combines the "advice" of any set of MAC protocols that run independently. The combination is based on a weighted-majority decision with randomized rounding, using continuously updated weights depending on the feedback obtained from the channel. The Meta-MAC protocol can automatically and adaptively select the best protocol for the unknown or unpredictable conditions from a given set of protocols. While this set may contain different MAC protocols, it may instead consist of a single MAC protocol with different parameter settings. In this way Meta-MAC can be used for adaptively optimizing the parameters of a given MAC protocol for the current network conditions.

6. OUTLOOK

One might be tempted to think that the intensive research and development of MAC protocols, going on for several decades, has already produced all essential ideas in this field, and, consequently, that substantial new development is unlikely. This is, however, a wrong perception of this area. The design of MAC protocols critically depends on the technological constraints (examples of which are mentioned in Section 1). Thus, the emergence of new technologies and novel systems constantly poses new challenges to MAC protocol design. Let us mention a few examples of such emerging challenges.

Directional antennas can be used for radios, redefining the meaning of conflicts in packet radio networks [24] with the potential to improve spatial reuse. In *sensor networks* the sensor battery is a critical resource, as its replacement is seldom feasible. Even in more powerful wireless nodes, in addition to energy-efficient advances in hardware, corresponding improvements in software (i.e., protocols) to conserve energy at all layers of the protocol stack are required. In particular, the design of *energy-efficient* MAC protocols seek to reduce the energy wasted by nodes overhearing transmissions not intended for them [25]. Another interesting opportunity is to utilize the technological feasibility of more sophisticated radio hardware that can relax some of the traditional technological constraints. For example, it may be possible for a radio to implement a *multiple reception capability*, where more than one packet can be received successfully in the same time slot, utilizing the fact that some of the base technologies, such as CDMA, make it feasible [26]. In addition, new multimedia applications and services present many new challenges and opportunities for medium access control.

BIOGRAPHIES

Dr. Andras Farago received a B.S. in 1976, an M.S. in 1979, and a Ph.D. in 1981, all in electrical engineering from the Technical University of Budapest, Hungary. After graduation he joined the Department of Mathematics at the Technical University of Budapest. In 1982, he moved to the Department of Telecommunications and Telematics of the same university. He was also cofounder and research director of the High Speed Networks Laboratory, the first research center in high speed networking in Hungary. In 1997, he became Szechenyi Professor of Telecommunications at the Technical University of Budapest. In 1998, he joined the University of Texas at Dallas as a professor of Computer Science. His main

research area is in algorithms, protocols, and modeling of telecommunication networks. Dr. Farago authored over 100 research papers and in 1996 received the distinguished recognition Doctor of the Hungarian Academy of Sciences.

Violet R. Syrotiuk received her B.Sc. in 1983 from the University of Alberta, Canada, her M.Sc. in 1984 from the University of British Columbia, Canada, and her Ph.D. in computer science in 1992 from the University of Waterloo, Ontario, Canada. Dr. Syrotiuk is currently an assistant professor in the Department of Computer Science in the Erik Jonsson School of Engineering and Computer Science at the University of Texas at Dallas, where she is the codirector of the Scalable Network Engineering Techniques Laboratory (NET Lab). Dr. Syrotiuk's research has been funded by the Defense Advanced Research Projects Agency (DARPA), and is currently supported by grants from the National Science Foundation (NSF) and Raytheon Company. Her current research interests include medium access control (MAC) protocols with special emphasis on intelligent protocol adaptation to unknown or changing network conditions, and network layer protocols with an emphasis on scalable design.

BIBLIOGRAPHY

1. M. Ranhema, Overview of the GSM system and protocol architecture, *IEEE Commun. Mag.* 92–100 (April 1993).
2. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.
3. I. Chlamtac, A. Faragó, and H. Zhang, A fundamental relationship between fairness and optimum throughput in TDMA protocols, *IEEE Int. Conf. Universal Personal Communications* (*ICUPC'96*), Cambridge, MA, Sept. 1996, pp. 671–675.
4. I. Chlamtac and S. Pinter, Distributed node organization algorithm for channel access in a multi-hop packet radio network, *IEEE Trans. Comput.* **36**(6): (1987).
5. C. Zhu and S. Corson, A five-phase reservation protocol (FPRP) for mobile ad hoc networks, *Proc. IEEE INFOCOM'98*, 1998.
6. A. Sen and M. L. Huson, A new model for scheduling packet radio networks, *IEEE INFOCOM'96*, 1996, pp. 1116–1124.
7. W. Stallings, *Wireless Communications and Networks*, Prentice-Hall, 2002.
8. Local and Metropolitan Area Networks Drafts (LAN/MAN 802), IEEE Standards Association Home Page, *http://standards.ieee.org*.
9. N. Abramson, The ALOHA system — another alternative for computer communications, *Proc. Fall Joint Computer Conf.*, 1970.
10. D. Aldous, Ultimate stability of exponential backoff protocol for acknowledgement based transmission control of random access communication channels, *IEEE Trans. Inform. Theory* **33**(2): 219–223 (1987).
11. F. P. Kelly, Stochastic models of computer communication systems, *J. Roy. Stat. Soc. B* **47**: 379–395 (1985).
12. J. Håstad, F. T. Leighton, and B. Rogoff, Analysis of backoff protocols for multiple access channels, *ACM Symp. Theory of Computing* (*STOC'87*), New York, May 1987, pp. 241–253.
13. S. Keshav, *An Engineering Approach to Computer Networking*, Addison-Wesley, 1997.
14. W. Crowther et al., A system for broadcast communication: Reservation-ALOHA, *Proc. 6th Hawaii Int. System Science Conf.*, Jan. 1973, pp. 596–603.
15. W. Stallings, *Data and Computer Communications and Networks*, 2nd ed., Macmillan, 1988.
16. A. Tobagi and L. Kleinrock, Packet switching in radio channels, Part II: The hidden terminal problem in carrier sense multiple access and the busy-tone solution, *IEEE Trans. Commun.* **23**: 1517–1453 (1975).
17. P. Karn, MACA — a new channel access protocol for packet radio, *ARRL/CRRL Amateur Radio 9th Computer Networking Conf.*, 1990, pp. 134–140.
18. European Telecommunications Standards Institute (ETSI), *http://www.etsi.org*.
19. I. Chlamtac and A. Faragó, Making transmission schedules immune to topology changes in multi-hop packet radio networks, *IEEE/ACM Trans. Network.* **2**(1): 23–29 (1994).
20. I. Chlamtac, A. Faragó, and H. Zhang, Time spread multiple access (TSMA) protocols for multihop mobile radio networks, *IEEE/ACM Trans. Network.* **5**(6): 804–812 (1997).
21. I. Chlamtac et al., ADAPT to mobility, *IEEE GLOBECOM*, Rio de Janeiro, Brazil, Dec. 1999.
22. Z. Tang and J. J. Garcia-Luna-Aceves, A protocol for topology-dependent transmission scheduling in wireless networks, *Proc. IEEE WCNC'99*, New Orleans, Sept. 21–24, 1999.
23. A. Faragó, A. D. Myers, V. R. Syrotiuk, and G. Záruba, Meta-MAC protocols: Automatic combination of MAC protocols to optimize performance for unknown conditions, *IEEE J. Select. Areas Commun.* **18**(9): 1670–1681 (2000).
24. Y.-B. Ko, V. Shankarkumar, and N. Vaidya, Medium access control protocols using directional antennas in ad hoc networks, *IEEE INFOCOM 2000*.
25. J. P. Monks, V. Bharghavan, and W.-M. W. Hwu, A power controlled multiple access protocol for wireless packet networks, *IEEE INFOCOM 2001*.
26. I. Chlamtac and A. Faragó, An optimal channel access protocol with multiple reception capacity, *IEEE Trans. Comput.* **43**(4): 480–484 (1994).

MEMS FOR RF/WIRELESS APPLICATIONS

HÉCTOR J. DE LOS SANTOS
Coventor, Inc.
Irvine, California

1. INTRODUCTION

The number of radiofrequency (RF) wireless applications and appliances is expected to explode in the first decade of the twenty-first century because of the unabated consumer demand for ubiquitous access to information [1]. The diversity of these consumers, which include both individuals and businesses, and the nature of the information they demand, which encompasses not only voice communications but also video, broadband data, messaging, navigation, direct broadcast satellite links, and the Internet, in the context of global connectivity, imposes, in turn,

such extreme levels of functionality and sophistication on these appliances, that doubts have been cast on the ability of conventional integrated circuit (IC) technology and fabrication techniques to deliver the high-performance RF functions required [1–3]. The seriousness of the matter may be gauged from an examination of the evolution in wireless standards (Table 1). In particular, while the first-generation (1G) appliances provided only single-band analog cellular connectivity capabilities, those of the second (2G) had to provide dual-mode dual-band digital voice plus data, and now those of the third (3G) and fourth (4G) generations will have to provide multimode (i.e., analog/digital), multiband (i.e., various frequencies), and multistandard [i.e., various standards — Global System for Mobile Communications (GSM) — a leading digital cellular system, which allows eight simultaneous calls on the same radiofrequency, digital European Cordless Telecommunications (DECT) — a system for the transmission of integrated voice and data in the range of 1.8–1.9 GHz, cellular digital packet data (CDPD) — a data transmission technology that uses unused cellular channels to transmit data in packets in the range of 800–900 MHz, General Packet Radio Service (GPRS) — a standard for wireless communications that runs at 150 kilobits per second (kbps), and code-division multiple access (CDMA) — a North American standard for wireless communications that uses spread-spectrum technology to encode each channel with a pseudorandom digital sequence] performance capabilities. Furthermore, the desire to maintain seamless connectivity, on a global basis, as the user moves through independently operated Internet Protocol (IP) networks [4–5], such as among various countries, dictates that these appliances be equipped for operation over a wide variety of access and network technologies and standards, with their accompanying processing overhead associated with function management. This latter requirement, in view of limited battery power, makes power consumption minimization a prime factor in the successful implementation of these systems.

Beyond technical performance considerations, however, commercial success is largely dependent on achieving a cost-effective solution. Thus, as the need for interfacing with off-chip components reflects adversely on the manufacturer's ability to meet cost constraints, its avoidance provides a strong motivation for the exploration of alternatives. Potential paths toward these alternatives become readily manifest when the limitations of conventional IC technologies for implementing RF functions are examined. In particular, attempts produce high-quality on-chip passive RF components, such as inductors, capacitors, varactors, switches, resonators, and transmission lines [2] — the core of wireless functions — reveal that this is a virtually impossible task because of the poor properties of silicon substrates. Against this bleak picture, microelectromechanical systems (MEMS) technology is emerging as the disruptive technology whose versatile fabrication techniques might well provide the solution to these limitations, insofar as it is poised to render virtually parasitic-free passive RF devices, side by side with the electronics, while simultaneously reaping the low cost property that characterizes batch fabrication processes.

In this article we review the fundamentals of MEMS fabrication techniques, the performance of typical RF devices exploiting them, and the RF MEMS circuits and systems it enables.

2. FUNDAMENTALS OF MEMS FABRICATION TECHNIQUES

MEMS fabrication techniques enable the construction of three-dimensional mechanical structures in the context of the conventional process utilized in the production of planar integrated circuits. As such, the structures created encompass feature sizes between a few and several hundred micrometers. To introduce the various approaches to MEMS fabrication, we begin with a brief review of the conventional IC fabrication process [2].

Table 1. Wireless Standards: The Evolution Blueprint

1G	2G	3G	4G
Analog cellular (single band)	Digital (dual-mode, dual band)	Multimode, multiband software-defined radio	Multistandard + multiband
Voice telecom only	Voice + data telecom	New services markets beyond traditional telecom: *higher-speed data, improved voice, multimedia mobility*	
Macrocell only	Macro/micro/picocell	Data networks, Internet, VPN, *WINternet*	
Outdoor coverage	Seamless indoor/outdoor coverage		
Distinct from PSTN	Complementary to fixed PSTN		
Business customer focus	Business + consumer	Total communications subscriber: virtual personal networking	

Source: http://www .uwcc.org.

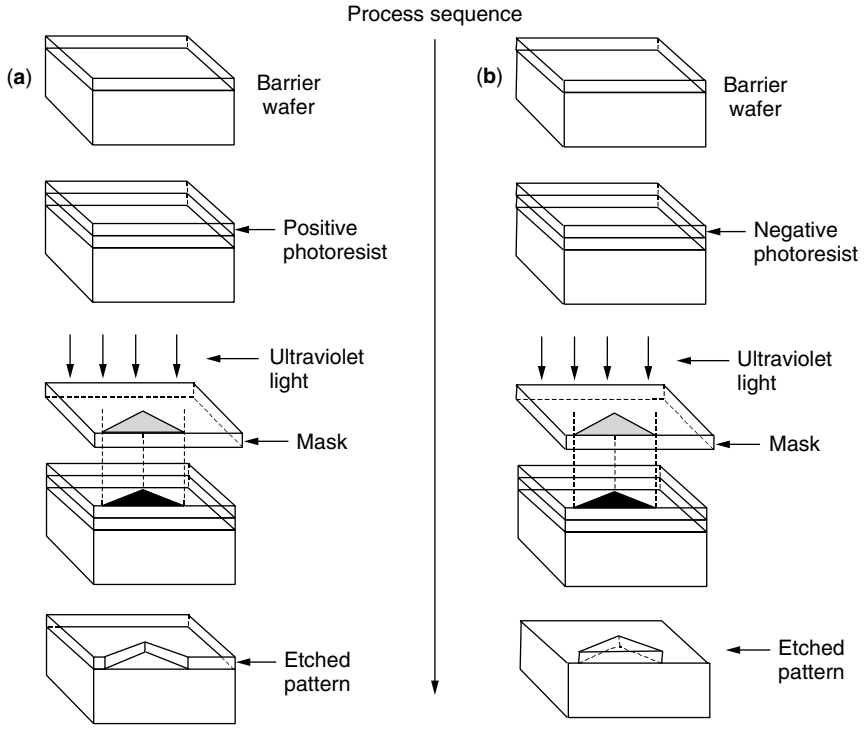


Figure 1. Conventional IC fabrication process using (**a**) positive and (**b**) negative photoresist.

2.1. Conventional IC Fabrication Process

The conventional IC fabrication process employs a sequence of photolithography and chemical etching steps to transfer a layout pattern onto the surface of a wafer. The process is illustrated by the sketch of Fig. 1. A semiconductor wafer, covered with a barrier material, is coated with a soft light-sensitive material called photoresist (PR). The PR may be positive or negative in the sense that, when exposed to ultraviolet (UV) light, it may harden or weaken, respectively. Thus, when positive PR is exposed through a mask containing transparent and opaque regions, representing the pattern to be transferred, a pattern identical to that on the mask is defined on the barrier upon subsequent chemical etching (Fig. 1a). On the other hand, if the PR is negative, the negative image of the pattern in the mask ends up being defined after etching (Fig. 1b).

2.2. RF MEMS Fabrication Approaches

Two main approaches, summarized below, dominate those employed to build three-dimensional mechanical structures for RF MEMS, namely, surface micromachining and bulk micromachining. Further information on these and other RF MEMS fabrication alternatives may be found in an earlier treatise [2].

2.2.1. Surface Micromachining. In surface micromachining, freestanding micromechanical structures are formed on the surface of a wafer by depositing and patterning a sequence of thin film material layers. Those layers that are deposited, and later removed, are called *sacrificial layers*; those layers that remain freestanding are called *structural layers*. Thus, the creation of every freestanding element involves the following main steps: (1) depositing a sacrificial layer; (2) opening a hole on the sacrificial layer that will permit access to the underlying wafer or structural layer; (3) deposition and patterning of the structural layer, such that it becomes anchored in the underlying substrate via a connection through the hole that was opened in the sacrificial layer; and (4) removal of the sacrificial layer to *release* the structure from it. These steps are sketched in Fig. 2. Examples of RF MEMS applications using this technique include switches, inductors, and varactors.

One fundamental limitation of surface micromachining [2] is the phenomenon of stiction. *Stiction* refers to the propensity of microscopic structures to stick together when they are close to each other or in a humid environment. An example of the former may occur when they experience van der Waals and electrostatic forces (due to random charging), whereas an example of the latter is when the structure is immersed in a wet etchant for dissolving the sacrificial layer, in which case the surface tension of the etchant may overcome the springback force that attempts to bring the structure to its equilibrium configuration. Popular approaches to overcome stiction during the release process involve the adoption of dry-etching chemistries, drying of the released wafer with supercritical CO_2, or freezing and then sublimating the release liquid.

2.2.2. Bulk Micromachining. In bulk micromachining, micromechanical structures are sculpted within the confines of a wafer. This is accomplished by the ingenious exploitation of highly directional (anisotropic) and nondirectional (isotropic) etchants, together with their etching rates, in relation to the various crystallographic planes of the wafer. Essentially, planes with a higher density of atoms will etch more slowly. Similarly, by defining heavily doped contour layers and pn (positive–negative) junctions,

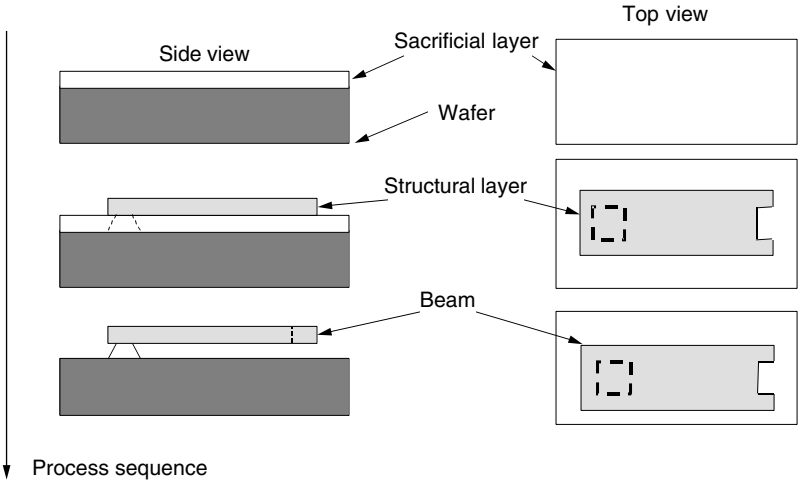


Figure 2. Sketch of surface micromachining process.

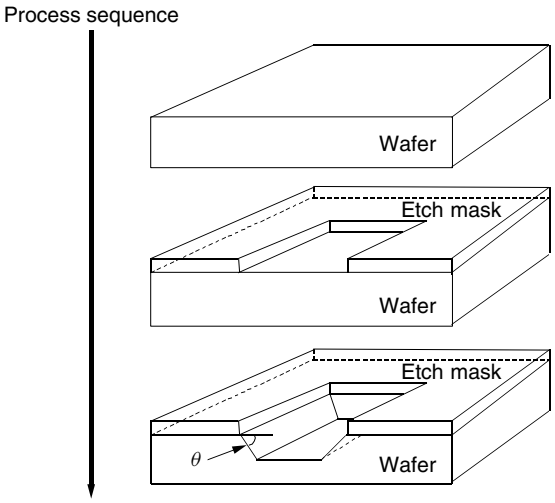


Figure 3. Sketch of bulk micromachining process.

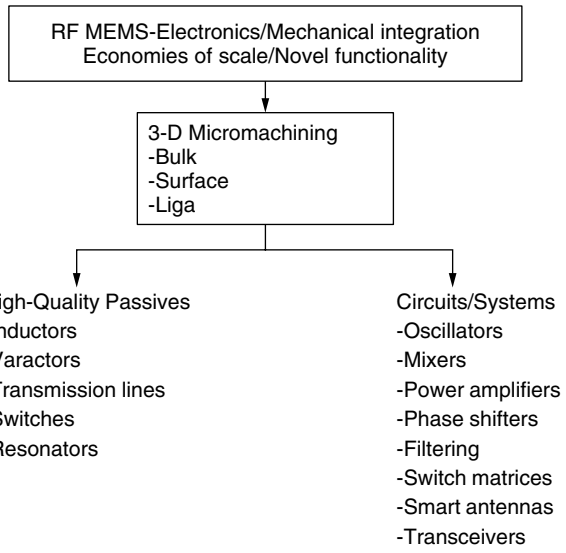


Figure 4. RF components enabled by MEMS technology (after Ref. 1).

for slowing down or totally stopping the etching process, respectively, the technique allows the creation of deep cavities. Figure 3 shows sketches of bulk micromachined structures. Examples of RF MEMS applications using this technique include transmission lines and inductors.

One fundamental limitation of bulk micromachining [2] is that the aspect ratio of the sculpted structures, such as the slope or verticality of the cavity walls, is a function of the angle between crystallographic planes. To overcome this limitation, a new technique called *deep reactive-ion etching* has been introduced.

3. RF MEMS DEVICES, CIRCUITS, AND SYSTEMS

The high level of interest in RF MEMS for wireless applications stems from the versatility of its fabrication approaches for producing virtually ideal (parasite-free) RF devices, in particular, in conjunction with integrated circuits, thus enabling new levels of circuits and systems functionality, together with the potential for achieving overall reductions in systems' weight/size and power, while exploiting the economies of scale germane to ICs. Figure 4 captures the RF MEMS arsenal and its potential areas of influence in wireless communications.

3.1. RF MEMS Passive Devices

Virtually all types of passive devices utilized in wireless applications (i.e., inductors, capacitors, varactors, transmission lines, switches, and resonators) have been demonstrated via MEMS fabrication techniques. In what follows, we describe representative examples of each of these.

3.1.1. Inductors. The performance of integrated inductors, in particular, their self-resonance frequency (the maximum frequency delimiting inductive behavior), and quality factor, is well known to be limited by the capacitance and resistance of the substrate on which they are disposed. Accordingly, a number of approaches aimed at separating the trace structure from the substrate have been advanced, including

1. Creating an airpit under the trace spiral via bulk micromachining [6], (Fig. 5a), which resulted

in a self-resonance frequency enhancement from 800 MHz to 3 GHz, on substrate removal, and a Q of 22 at 270 MHz on an 115-nH inductor;

2. Suspending the trace spiral a distance over the wafer surface via a combination of bulk and surface micromachining [8] (Fig. 5b), which resulted in a Q of 30 at 8 GHz on a 10.4-nH inductor with a self-resonance frequency of 10.1 GHz;
3. Implementing the inductor as a solenoid via surface micromachining [9] (Fig. 5c), which resulted in a Q of 16.7 at 2.4 GHz on 2.67-nH inductors;
4. Using self-assembly techniques to erect the plane of trace structure perpendicular to the substrate [10] (Fig. 5d), which resulted in improvements in the Q of 2nH meander inductors from 4 at 1 GHz, for the planar realization, to 20 at 3 GHz for the self-assembly implementation

3.1.2. Varactors. Varactors are indispensable in the operation of voltage-controlled oscillators (VCOs). High-quality varactors, however, are difficult to produce in the context on an IC because of processes are usually optimized for other devices, such as transistors [2]. Since MEMS devices may be integrated on chip without disrupting the process flow, specifically in a postprocessing step, several schemes using surface micromachining have been exploited to create potentially IC-compatible varactors. These are predicated upon varying one of the parameters defining capacitance, $C = \varepsilon A/d$, where ε is the dielectric constant, A is the area, and d is the plate separation. Schemes that vary the plate separation employ a square plate suspended by four beams and disposed over a bottom plate (electrode). At zero bias, there is a maximum distance between top and bottom plates. However, when a voltage is applied between the plates, the force of electrostatic attraction between them causes the gap to diminish [2], thus varying (increasing) the capacitance [11]. In another scheme, the capacitor structure consists of interdigitated plates. Thus by varying the degree of engagement, namely the effective device area, the overall capacitance is made to vary [12]. Finally, in a more recent scheme (Fig. 6), the effective dielectric constant of the structure is made to vary by sliding, in a lateral fashion, a dielectric between the parallel plates of the capacitor [13].

3.1.3. Transmission Lines. The performance of transmission lines, in particular, their attenuation, is a strong

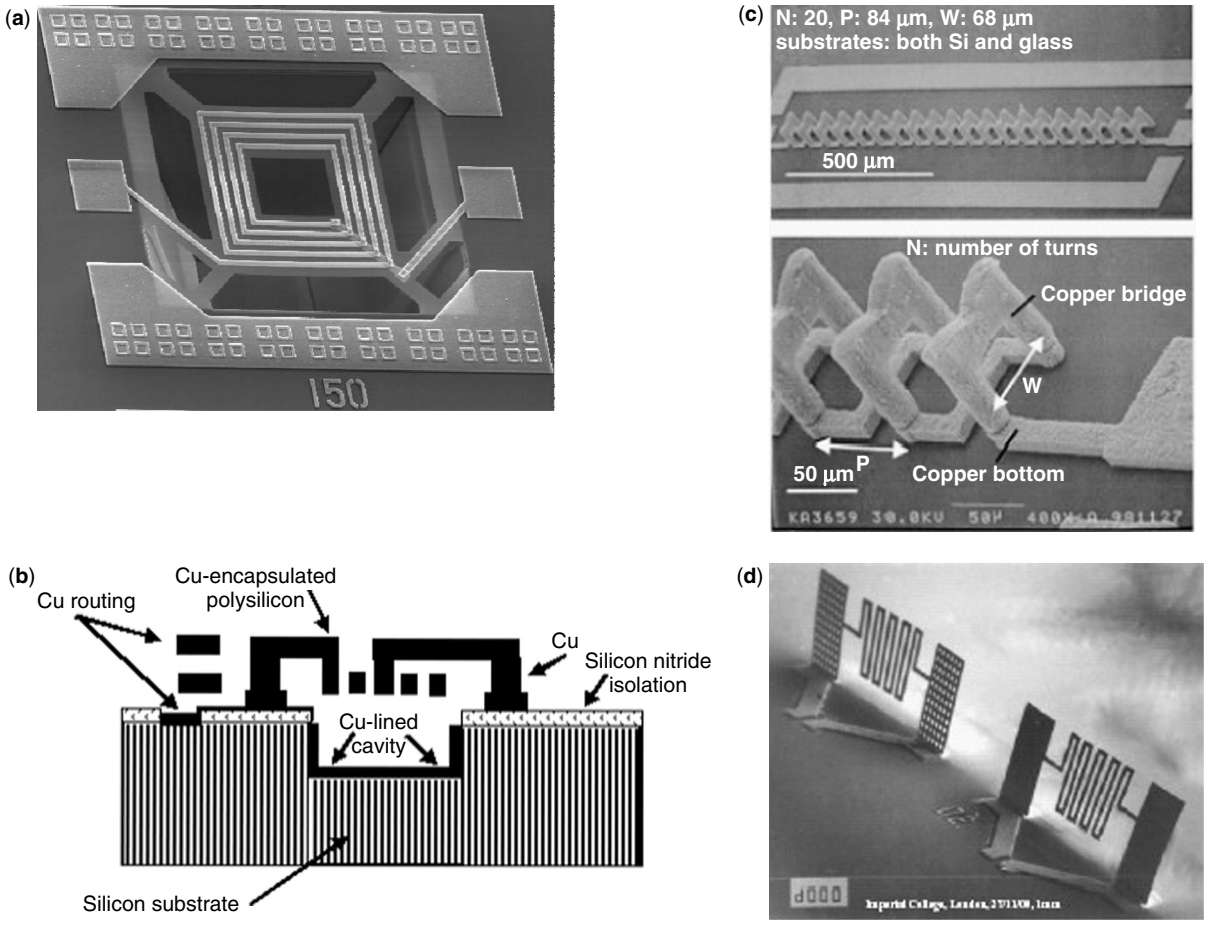


Figure 5. RF MEMS Inductors: (**a**) bulk-micromachined inductor [7]; (**b**) schematic of a copper-encapsulated polysilicon inductor suspended over a copper-lined cavity beneath [8]; (**c**) SEM photograph of 20-turn, on-Si, air-core, all-copper solenoid inductor (*upper* — overview; *lower* — magnified view) [9]; (**d**) $4\frac{1}{2}$-turn meander inductor after self-assembly [10].

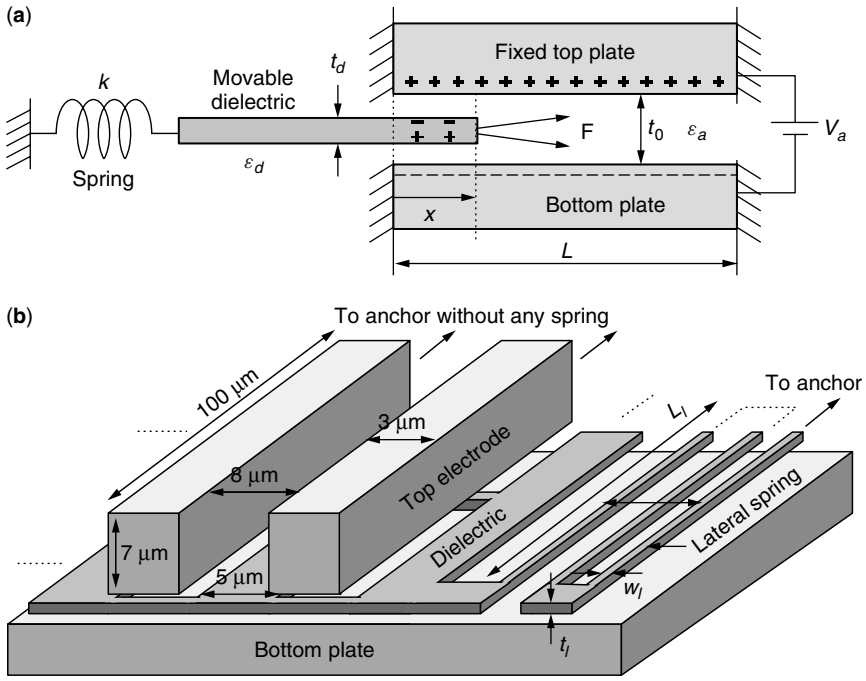


Figure 6. Micromachined varactor: (**a**) conceptual schematic; (**b**) actual implementation using a lateral spring. (from Ref. 13 © 1998 IEEE).

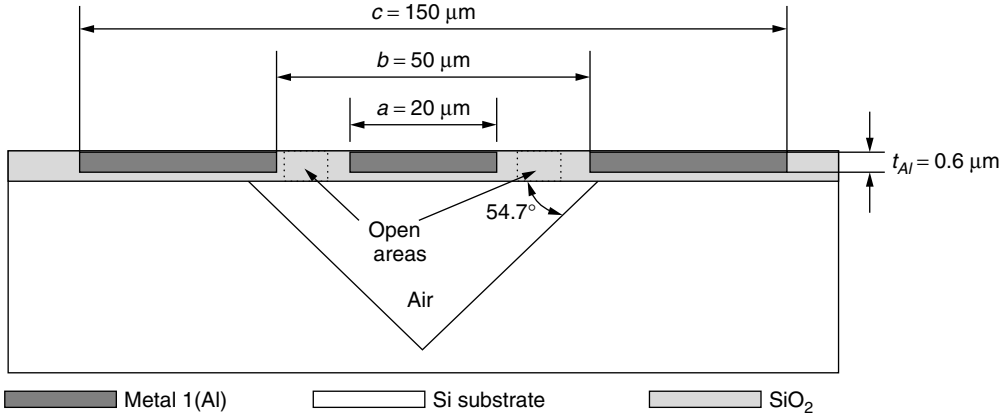


Figure 7. Cross-sectional view of bulk-etched transmission-line structure (after Ref. 14).

function of the substrate that mechanically supports them. Thus, silicon transmission lines tend to be lossy. To enable low-loss interconnects in the context of a silicon IC, Milanovic et al. [14] developed the structure shown in Fig. 7. This is a coplanar waveguide (CPW) transmission line, in which, by opening access windows in the top passivation, an airpit is formed underneath the center conductor to eliminate the substrate under it and, thus, minimize the structure's insertion loss (IL). Improvements in the IL of about 7 dB at 7 GHz, and 20 dB at 20 GHz, with respect to the nonetched reference, were obtained.

3.1.4. Switches. Switches may be considered one of the RF MEMS elements with the potential for greatest impact in wireless communications. With such capabilities as [3] series resistance $<1\ \Omega$, insertion loss at 1 GHz within 0.1 dB, Isolation at 1 GHz $>$ 40 dB, IP3 $>$ 66 dBm, 1 dB compression $>$33 dBm, size $<1\ mm^2$, switching speed of the order of 1 μs, control voltage between 3 and 30 V, and control current $<1\ \mu A$, with no standby power consumption, they have become the potential enabler for many systems, in particular, phased arrays and switch matrices [2]. Figure 8 [15] shows the structure and operation of a state-of-the-art RF MEMS switch. The switch consists of a metal bridge bar that is moved to make or break contact with an underlying signal line. Bridge motion is achieved by voltage biasing a mechanical actuator supporting the bridge. The device covers an area of approximately $250 \times 250\ \mu m$. The typical performance included and effective capacitance of 2fF in the OFF state, for an isolation of 30 dB at 40 GHz, an effective resistance of 1 Ω, giving an insertion loss of 0.2 dB, and a return loss of 25 dB. In addition, the actuation voltage was 85 V and the switching time about 10 μs.

3.1.5. Resonators. Resonators are essential elements for realizing filters and oscillators [2]. Their RF MEMS implementations take usually two main forms: a *cavity*,

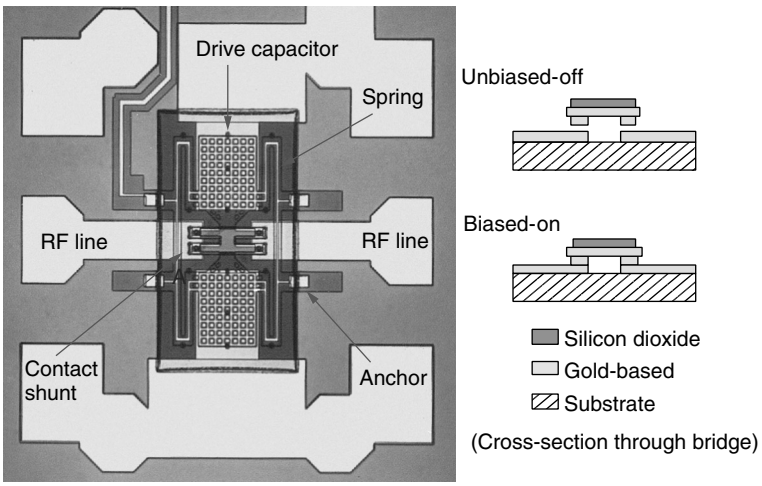


Figure 8. Structure of RF MEM switch (courtesy of Drs. R. E. Mihailovich and J. DeNatale, Rockwell Scientific) (from Ref. 15 © 2001 IEEE).

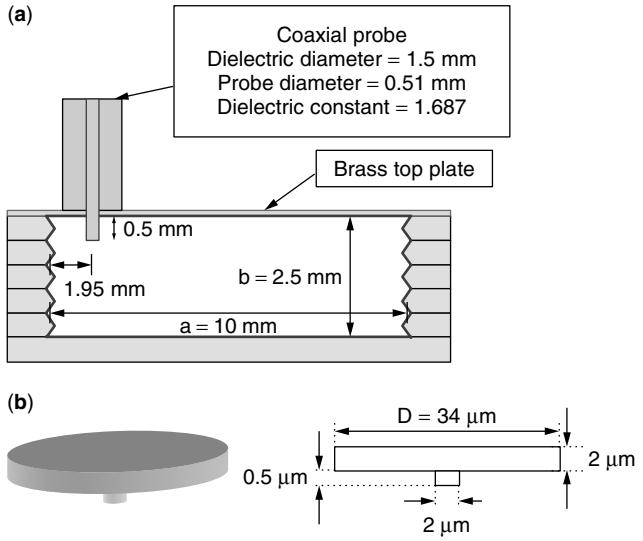


Figure 9. RF MEMS resonators: (**a**) schematic and photograph of bulk micromachined cavity resonator at 30 GHz (length $c = 5$ mm) (courtesy of Mr. M. Stickel and Prof. G. V. Eleftheriades, Univ. of Toronto); (**b**) schematic of contour-mode disk resonator (courtesy of Mr. Hideyuki Maekoba, Coventor, Inc.).

for applications beyond 10 GHz, and a *micromechanical resonator*, for applications below 1 GHz. Figure 9a shows an example of the former, which exhibited $Q > 2000$ at 30 GHz, and Fig. 9b, which exhibited $Q = 9200$ at 156 MHz, an example of the latter.

In addition to these resonators, there is the film bulk acoustic wave resonator (FBAR), which is based on the formation of an acoustic cavity out of a piezoelectric material, and which exhibits Q between 500 and >1000, at frequencies of several GHz [16].

3.2. Circuit Applications of RF MEMS Devices

While a number of RF MEMS-based circuits, notably, oscillators and filters, have been demonstrated [16–21], phase shifters exploiting MEM switches may be considered the major technology driver, as they are an enabling component for the realization of large phased arrays. An example of such a phase shifter is shown in Fig. 10 [23]. This is a line-switched true(real)-time delay (TTD) 4-bit phase shifter implemented with the RF MEMS switches described in Section 3.1.4. In the DC 40-GHz frequency band the circuit exhibited delay times in the 106.9–193.9 ps range. This was accomplished with a resolution of 5.8-ps-delay increments, which represents a phase shift of 22.5° at 10.8 GHz produced by using microstrip lines with a length of 600 μm. The total chip area was 6×5 mm^2.

4. SUMMARY

In this article we have presented a brief review of MEMS for RF/wireless applications. In particular, we have addressed the motivations propelling the high level of interest in this emerging technology, its fabrication fundamentals, and the sample realizations and performance of key devices that it enables, namely,

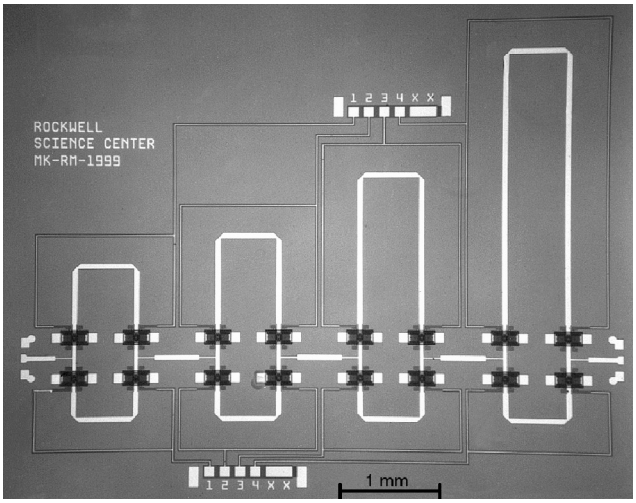


Figure 10. Photograph of 4-bit RF MEMS TTD phase shifter. The second longest bit (bit 3) was fabricated separately to analyze both the insertion loss and the impact of the matching section on TTD performance. (*Source*: Ref. 23 © 2001 IEEE. Courtesy of Drs. R. E. Mihailovich and J. DeNatale.)

inductors, varactors, transmission lines, switches, and resonators. We have also presented, perhaps, the major RF MEMS technology driver, namely, the phase shifter circuit, which is of great importance to large systems applications, in particular, phased arrays.

BIOGRAPHY

Héctor J. De Los Santos is Principal Scientist at Conventor, Inc., Irvine, California, where he leads Conventor's RF MEMS R&D. He received a Ph.D. from the School of Electrical Engineering, Purdue University, West Lafayette, Indiana, in 1989. From March 1989 to September 2000, he was employed at Hughes space and Communications Company, Los Angeles, where he served as Scientist and Principal Investigator and Director of the Future Enabling Technologies IR&D Program. Under this program he pursued research in the areas of RF MEMS, quantum functional devices and circuits, and photonic bandgap devices and circuits. Dr. De Los Santos holds a dozen patents and has over six patents pending. He is author of the bestseller textbook *Introduction to Microelectromechanical (MEM) Microwave Systems*, Artech House, Norwood, Massachusetts, 1999, and of the book *RF MEMS Circuit Design for Wireless Communications*, Artech House, June 2002. Dr. De Los Santos is a Senior Member of the IEEE, and member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi. He is an IEEE Distinguished Lecturer of the Microwave Theory and Techniques Society for the 2001–2003 term.

BIBLIOGRAPHY

1. H. J. De Los Santos, MEMS — a wireless vision, *Proc. 2001 International MEMS Workshop*, Singapore, July 4–6, 2001.
2. H. J. De Los Santos, *Introduction to Microelectromechanical (MEM) Microwave Systems*, Artech House, Norwood, MA, 1999.
3. R. J. Richards and H. J. De Los Santos, MEMS for RF/wireless applications: The next wave, *Microwave J.* (March 2001).
4. R. R. Parrish, Mobility and the Internet, *IEEE Potentials Mag.* 8–10 (April/May 1998).
5. A. Fasbender, F. Reichert, E. Geulen, and J. Hjelm, Any network, any terminal, anywhere, *IEEE Pers. Commun. Mag.* 22–30 (April 1999).
6. J.-Y. Chang, A. A. Abidi, and M. Gaitan, Large suspended inductors on silicon and their use in a 2um CMOS RF amplifier, *IEEE Electron Device Lett.* **14**: 246–248 (1993).
7. Y. Sun, H. van Zeijl, J. L. Tauritz, and R. G. F. Baets, Suspended membrane inductors and capacitors for application in silicon MMICs, *IEEE Microwave and Millimeter-wave Monolithic Circuits Symp. Digest of Papers*, 1996, pp. 99–102.
8. H. Jiang, Y. Wang, J.-L. A. Yeh, and N. C. Tien, Fabrication of high-performance on-chip suspended spiral inductors by micromachining and electroless copper plating, *2000 IEEE IMS Digest of Papers*, Boston, MA.
9. J.-B Yoon et al., Surface micromachined solenoid On-Si and On-Glass inductors for RF applications, *IEEE Electron Device Lett.* **20**: 487 (1999).
10. G. W. Dahlmann et al., MEMS high Q microwave inductors using solder surface tension self-assembly, *2001 IEEE IMS Digest of Papers*.
11. D. J. Young and B. E. Boser, A micromachined variable capacitor for monolithic low-noise VCOs, *Hilton Head '96*, pp. 86–89.
12. J. J. Yao, Topical review: RF MEMS from a device perspective, *J. Micromech. Microeng.* **10**: R9–R38 (2000).
13. J.-B. Yoon and C. T.-C. Nguyen, A high-Q tunable micromechanical capacitor with movable dielectric for RF applications, *1998 IEEE Int. Electron Devices Meeting Digest of Papers*, pp. 489–492 (Figs. 1, 4, 6).
14. V. Milanovic et al., Micromachined microwave transmission lines in CMOS technology, *IEEE Trans. Microwave Theory Tech.* **45**: 630–635 (1997).
15. R. E. Mihailovich et al., MEM relay for reconfigurable RF circuits, *IEEE Microwave Wireless Components Lett.* **11**: 53–55 (Feb. 2001).
16. H. J. De Los Santos, *RF MEMS Circuit Design for Wireless Communications*, Artech House, Norwood, MA, 2002.
17. P. Bradley, R. Ruby, and J. D. Larson III, A film bulk acoustic resonator (FBAR) duplexer for USPCS, *2001 IEEE Int. Microwave Symp.*, Phoenix, AZ.
18. A. R. Brown and G. M. Rebeiz, A high-performance integrated-band diplexer, *IEEE Trans. Microwave Theory Tech.* **47**: 1477–1481 (Aug. 1999).
19. H.-T. Kim, J.-H. Park, Y. Kim, and Y. Kwon, Millimeter-wave micromachined tunable filters, *1999 IEEE MTT-S Digest*, pp. 1235–1238.
20. F. D. Bannon III, J. R. Clark, and C. T.-C. Nguyen, High-Q HF microelectromechanical filters, *IEEE J. Solid-State Circuits* **35**: 512–526 (April 2000).
21. K. Wang and C. T.-C. Nguyen, High-order micromechanical electronic filters, *Proc. IEEE Micro Electro Mechanical Systems Workshop*, 1999, pp. 25–30.
22. C. T.-C. Nguyen and R. T. Howe, An integrated CMOS micromechanical resonator high-Q oscillator, *IEEE J. Solid-State Circuits* **34**: 440–445 (April 1999).
23. M. Kim, J. B. Hacker, R. E. Mihailovich, and J. F. DeNatale, A DC-40 GHz four-bit RF MEMS true-time delay network, *IEEE Microwave Wireless Components Lett.* **11**: 56–58 (Feb. 2001).

MICROSTRIP ANTENNAS

NAFTALI HERSCOVICI
Anteg, Inc.
Framingham, Massachusetts

1. INTRODUCTION

Microstrip antennas consist of a patch of metalization separated from a ground plane by a dielectric substrate (Fig. 1). The concept of the microstrip radiator was proposed in the early 1950s by Deschamps [1]. Only a couple of years after Deschamps' communication, a French patent on a similar geometry was awarded to Gutton and Baissinot [2]. However, no reports on this subject were published in the literature until the early

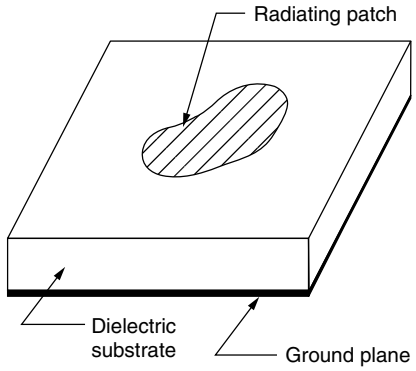


Figure 1. The microstrip antenna.

1970s, when Byron [3] proposed the "conductive strip radiator separated from the ground plane by a dielectric substrate." Since then, to the present, significant progress has been made in the development of dielectric substrates and computer technologies, which both contributed to the development of numerous variations of the basic concept proposed by Deschamps.

Since microstrip radiators are essentially planar in nature [two-and-one-half-dimensional ($2\frac{1}{2}$D) structures], they are narrowband antennas. This is true for the initial configurations, which consisted of a single patch. Since the late 1980s, considerable effort has been invested in developing broadband microstrip elements, which were succeeding in approaching 90% bandwidth for a VSWR < 2.

Microstrip antennas have a number of advantages in comparison with other types of antennas:

- Light weight, low volume, and, to a certain extent, flexibility, which allows integration on conformal surfaces
- Allow integration with active devices
- Easily arrayable, allowing a significant freedom of design and synthesis of various radiation patterns.
- Low fabrication cost
- All polarizations possible with relatively simple feeding mechanisms

The limitations of microstrip antennas and arrays are

- Narrow band.
- Large arrays can exhibit low efficiency due to the losses associated with the feeding network.
- Radiation coverage limited to one hemisphere.
- Very low sidelobe arrays are difficult to obtain because of the radiation of the feeding network.
- Losses associated with surface waves.
- Low power handling capability.

These limitations of basic microstrip antennas and arrays can be overcome with more sophisticated architectures, which might make the design expensive to mass production and sensitive to manufacturing tolerances.

In spite of their simple geometry, the design of microstrip antennas can be a complicated and iterative process. This is mostly because of their high-Q nature and the complexity of the analysis associated with the accurate modeling such structures. Many approximate models have been developed and, since the late 1980s, with the development of fast computers, numerical methods have been developed for accurate analysis.

The approximate methods treat the microstrip patch antenna as a transmission line (the *transmission-line model* and its derivatives) or a cavity (the *cavity model*) and provide a better physical insight, which is missing in the accurate CAD models. The formulas for the main characteristics of the microstrip antennas given below are based on approximate methods that are accurate enough for the initial design iteration.

2. ELECTRICAL CHARACTERISTICS OF A RECTANGULAR PATCH ANTENNA

2.1. Radiation Characteristics of the Rectangular Microstrip Antenna Element

Considering its Cartesian shape, the rectangular microstrip patch antenna is relatively easy to analyze, so the electrical characteristics of microstrip antennas presented below pertain to the rectangular patch. Furthermore, experience shows that the rectangular patch is much more used than any other type of patch. The typical rectangular microstrip radiator is shown in Fig. 2. For calculation of the radiation patterns, the rectangular patch can be seen as a line resonator, approximately a half-wavelength long [4]. The radiation occurs mostly from the fringing fields at the open-transmission-line ends (Figs. 3 and 4). Including the effect of the ground plane and substrate, the E-plane pattern is given by

$$E_\theta(\theta) = -jk_0V_0W\frac{e^{-jk_0r}}{2\pi r}\sin c\left(\frac{k_0h}{2}\sin\theta\right) \times \cos\left(\frac{k_0L}{2}\sin\theta\right)F_1(\theta) \tag{1}$$

$$E_\phi = 0 \tag{2}$$

and the H-plane pattern is given by

$$E_\phi(\theta) = jk_0V_0W\frac{e^{-jk_0r}}{2\pi r}\sin c\left(\frac{k_0W}{2}\sin\theta\right)\cos\theta F_2(\theta) \tag{3}$$

$$E_\theta(\theta) = 0 \tag{4}$$

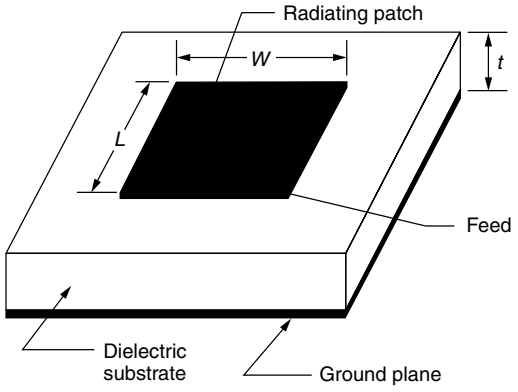


Figure 2. The rectangular, single-layer microstrip radiator.

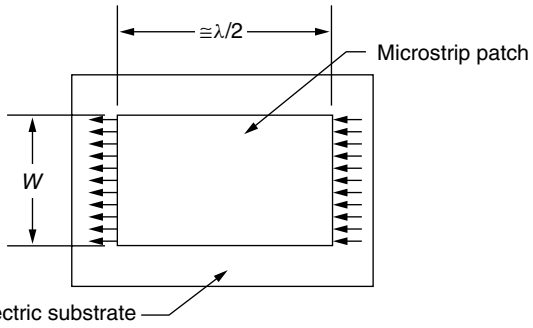


Figure 3. The rectangular microstrip patch with equivalent radiating slots.

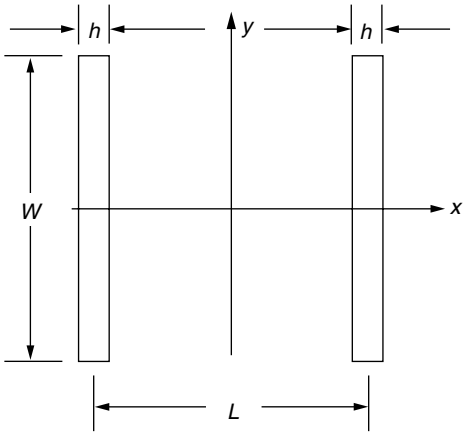


Figure 4. The rectangular microstrip antenna represented as two radiating slots.

where

$$F_1 = \frac{2\cos\theta\sqrt{\varepsilon_r - \sin^2\theta}}{\sqrt{\varepsilon_r - \sin^2\theta} - j\varepsilon_r\cos\theta\cot(k_0 h\sqrt{\varepsilon_r - \sin^2\theta})} \tag{5}$$

$$F_2 = \frac{2\cos\theta}{\cos\theta - j\sqrt{\varepsilon_r - \sin^2\theta}\cot(k_0 h\sqrt{\varepsilon_r - \sin^2\theta})} \tag{6}$$

and

W = the width of the patch (H-plane dimension)
L = the length of the patch (E-plane dimension also called the resonant dimension)
h = the thickness of the substrate
ε_r = the dielectric constant of the substrate
$k_0 = \dfrac{2\pi}{\lambda_0}$, where λ_0 is the wavelength

The half-power beamwidth can be approximated to give

$$\theta_{BH} = 2\arccos\sqrt{\frac{1}{2\left\{1 + \dfrac{k_0 W}{2}\right\}}} \tag{7}$$

$$\theta_{BE} = 2\arccos\sqrt{\frac{7.03}{3k_0^2 L^2 + k_0^2 h^2}} \tag{8}$$

2.2. Input Impedance of the Rectangular Microstrip Antenna Element

The transmission-line model [4] does not take the position of the feeding point along the length of the patch into consideration. Newman and Tulyathan proposed [5] an improved model, which solves this problem. They derived a simple formula for the input impedance, which also includes the reactance of the probe:

$$Z_{\text{in}} = Z_1 + jX_L \tag{9}$$

$$Z_1 = \frac{1}{Y_1} \tag{10}$$

$$Y_1 = Y_0\left[\frac{Z_0\cos\beta L_1 + jZ_w\sin\beta L_1}{Z_w\cos\beta L_1 + jZ_0\sin\beta L_1} + \frac{Z_0\cos\beta L_2 + jZ_w\sin\beta L_2}{Z_w\cos\beta L_2 + jZ_0\sin\beta L_2}\right] \tag{11}$$

where

$$X_L = \frac{377}{\sqrt{\varepsilon_r}}\tan\left(\frac{2\pi h}{\lambda_0}\right) \tag{12}$$

is the probe reactance, Z_0 is the characteristic impedance of the microstrip, and $Y_W = 1/Z_W$ is the wall admittance in the E plane as defined in the cavity model proposed by Bahl [6]. Then

$$Y_W = G_W + jB_W \tag{13}$$

$$G_W = \frac{0.00836W}{\lambda_0} \tag{14}$$

$$B_W = 0.01668\frac{\Delta I}{h}\frac{W}{\lambda_0}\varepsilon_e \tag{15}$$

where ε_e is the dielectric effective constant, given by

$$\varepsilon_e = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + \frac{12h}{W}\right)^{-1/2} \tag{16}$$

and Δl is a length correction factor given by

$$\Delta l = 0.412h\frac{(\varepsilon_e + 0.3)(W/h + 0.264)}{(\varepsilon_e - 0.258)(W/h + 0.8)} \tag{17}$$

Again, the quantities defined above are approximations required by the transmission-line model (and other approximate models) and for the same quantities, various expressions have been derived [6,7].

2.3. Design Procedure for Rectangular Microstrip Antennas

2.3.1. Choice of the Dielectric Substrate. The first step in the design of microstrip antenna is the choice of the dielectric substrate, which includes the following considerations: dielectric constant, thickness, and losses. As shown further, the Q factor of the antenna is strongly dependent on the dielectric constant and substrate thickness, as well as the efficiency associated with the surface wave excitation.

2.3.2. The Element Width. Once the dielectric substrate is chosen, the "effective dielectric constant" of the

substrate has to be calculated [Eq. (16)]. The dielectric constant ε_r has a definite impact on the resonance frequency which is determined not only by ε_r and h but also by the length of the patch, L. The width of the patch has an impact on the input impedance and a good approximation for W is

$$W = \frac{c}{2f_r}\sqrt{\frac{2}{\varepsilon_r + 1}}$$

where c is the velocity of light and f_r is the resonance frequency. The width of the patch, which also controls the radiation pattern [Eq. (7)], has to be chosen carefully to avoid the excitation of higher-order modes. The dependence of W on frequency for three different dielectric constants is shown in Fig. 5 [8].

2.3.3. The Element Length. The element length (often known as the *resonant dimension* of the rectangular patch) determines the resonant frequency. Here *resonant frequency* means "the frequency where the reactance of the antenna equals zero." In practice, the presence of additional components, such as feeding probes and stubs, alter the meaning of this definition, which is sometimes changed to "the frequency where the maximum of the real part of the input impedance occurs."

Knowing ε_e and Δ_l one can calculate L (the resonance dimension of the rectangular patch):

$$L = \frac{c}{2f_r\sqrt{\varepsilon_e}} - 2\Delta l \tag{18}$$

Figure 6 shows L versus the resonance frequency for a number of different substrates.

2.3.4. Radiation Patterns. Unlike other types of antennas where the characteristics of the radiation patterns are determined by parameters that do not significantly impact the input impedances and working bandwidth, in the case of microstrip elements, once the dielectric substrate, the patch length, and the patch width are determined, the characteristics of the radiation pattern are already set. To obtain narrower beamwidths (higher directivity) using microstrip radiators, arrays will have to be used.

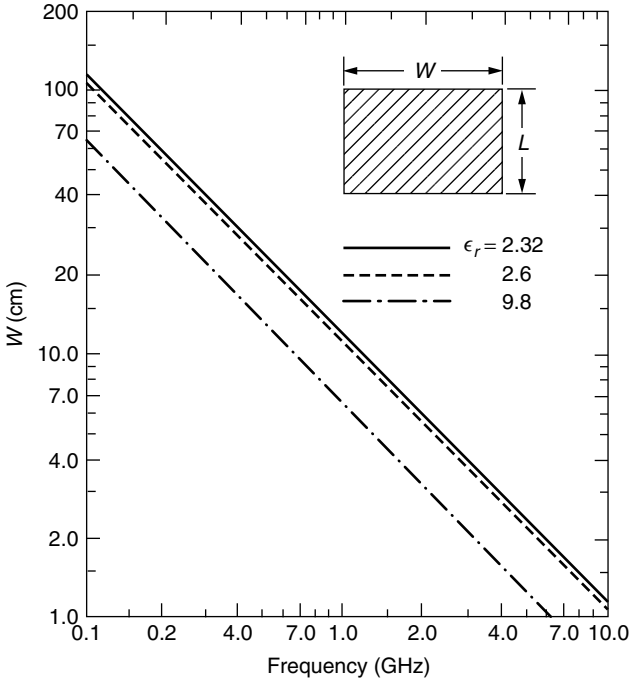


Figure 5. Element width versus frequency for different dielectric substrates [8].

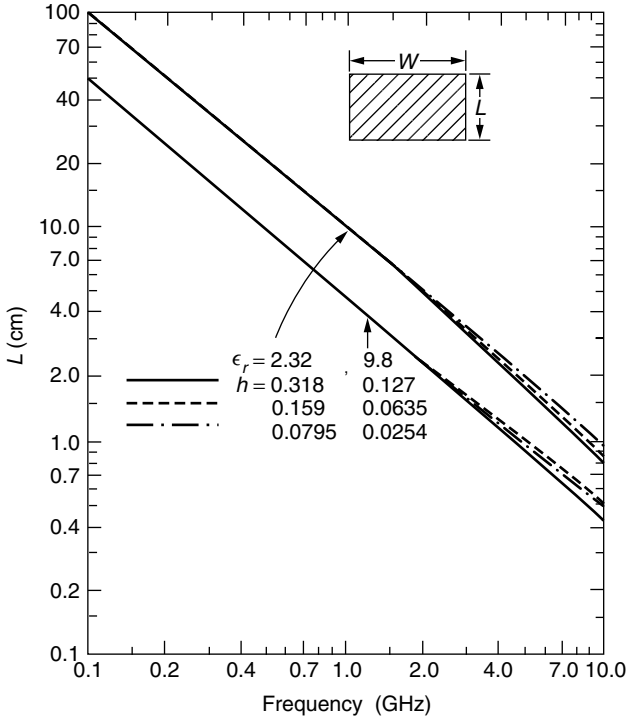


Figure 6. Element length versus frequency for different dielectric substrates [8].

2.3.5. Input Impedance. The only free parameter left is the location of the feeding point. Using this parameter, one can design a patch with almost any input impedance at resonance. Because of its fundamental current distribution, a patch fed in the center has a zero-ohm input impedance. To avoid the excitation of cross-polarization currents (for linear polarization), the feeding point has to be positioned symmetrically with respect of the width of the patch. Figure 7 shows the dependence of the patch input impedance on the location of the feeding point.

2.3.6. Q Factor and Losses. The quality factor of the patch is given by

$$Q = \frac{Q_r R_T}{R_r} \tag{19}$$

where Q_r is the quality factor associated with the radiation resistance [10]:

$$Q_r = \frac{c\sqrt{\varepsilon_e}}{4f_r h} \tag{20}$$

$$R_c = 0.00027\sqrt{f_r}\frac{L}{W}Q_r^2 \tag{21}$$

$$R_d = \frac{30\tan\delta}{\varepsilon_r}\frac{h\lambda_0}{LW}Q_r^2 \tag{22}$$

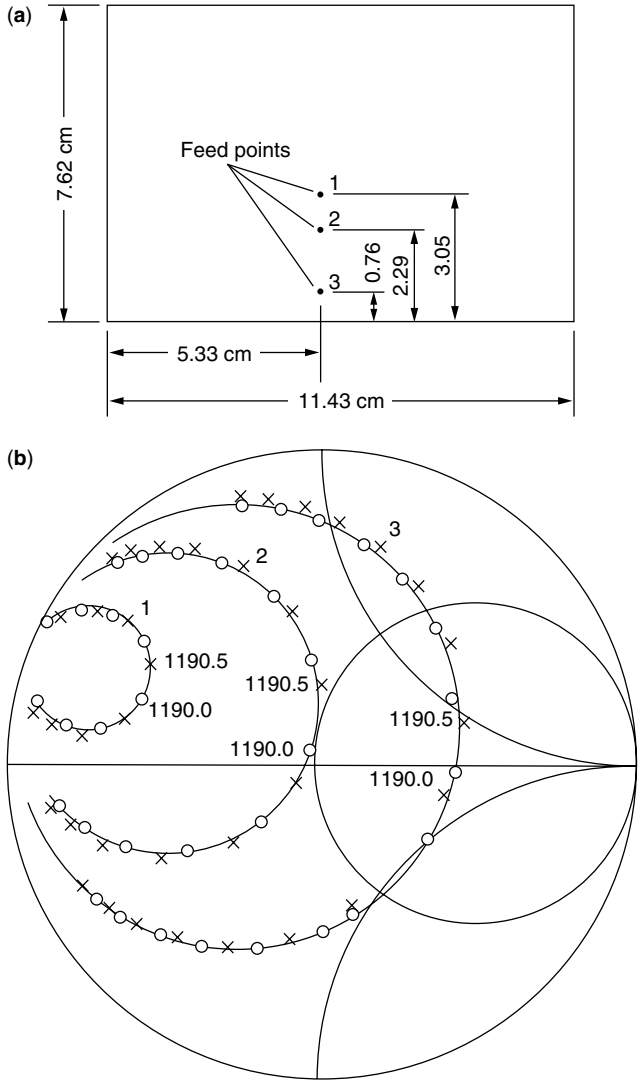


Figure 7. Experimental and theoretical loci for a microstrip rectangular patch antenna (xxx theoretical, ooo experimental) [9].

and

$$R_T = R_r + R_d + R_c \tag{23}$$

The radiation efficiency is thus

$$\eta\% = 100\frac{R_r}{R_T} \tag{24}$$

Figure 8 shows the radiation resistance R_T as a function of frequency for different dielectric substrates, and Fig. 9 shows the efficiency, $\eta(\%)$ as a function of frequency for different dielectric substrates.

2.3.7. Bandwidth. The bandwidth of the microstrip for VSWR < $\text{VSWR}_{\max}$ is given by

$$BW = \frac{\text{VSWR}_{\max} - 1}{Q_T\sqrt{\text{VSWR}_{\max}}} \tag{25}$$

This formula establishes two basic principles in the design of microstrip antennas; for a certain frequency,

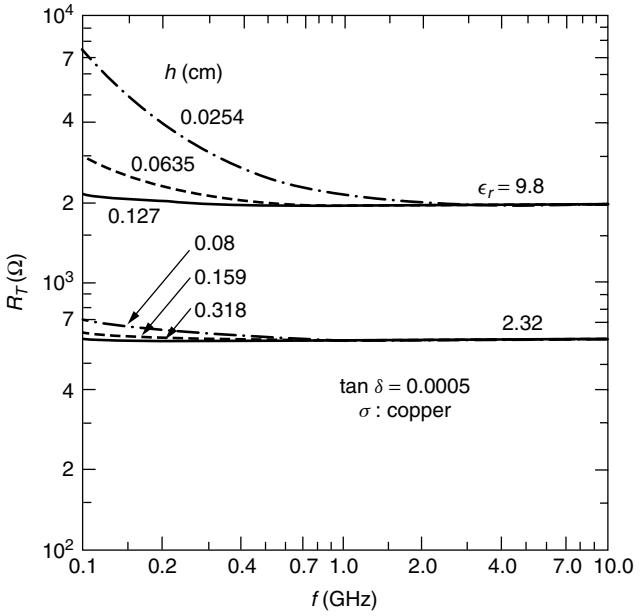


Figure 8. Radiation resistance as a function of frequency for different dielectric substrates [8].

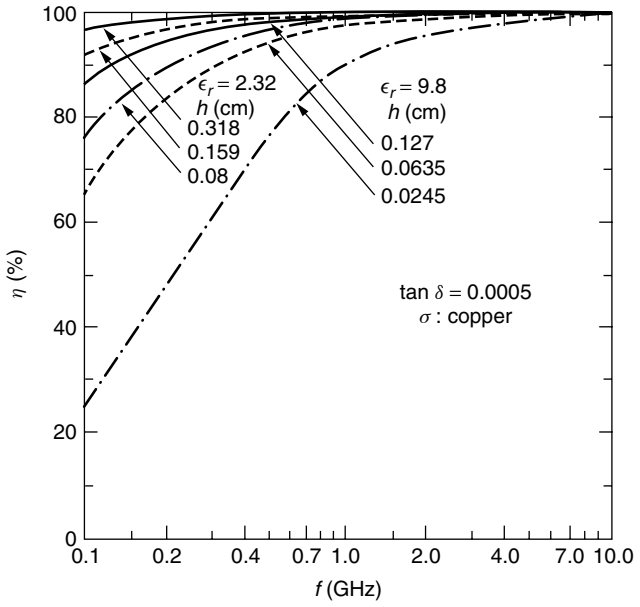


Figure 9. Efficiency as a function of frequency for different dielectric substrates [8].

the bandwidth of the antenna is directly proportional to the substrate thickness and inversely proportional to the dielectric constant of the substrate.

2.3.8. Directivity and Gain. Bahl and Bhartia [8] showed that a good approximation for the directivity is given by

$$D \cong 6.6 \quad W \ll \lambda_0 \tag{26}$$

$$D \cong \frac{8W}{\lambda_0} \quad W \gg \lambda_0 \tag{27}$$

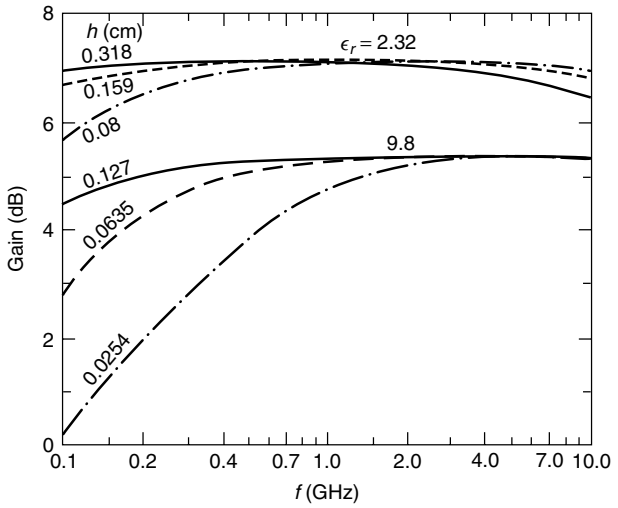


Figure 10. Gain as a function of frequency for various rectangular microstrip antennas [8].

For any antenna, the gain is defined as $G = \eta D$. Figure 10 shows the Gain as a function of frequency for various rectangular microstrip antennas.

2.4. The Impact of Manufacturing Tolerance on the Electrical Characteristics of the Rectangular Patch Antenna

A number of facts have to be considered in the fabrication of microstrip antennas:

1. Substrates with a low dielectric constant have variation in the dielectric constant of about $\pm 1\%$ and $\pm 5\%$ in thicknesses.
2. The tolerances for higher dielectric constant substrates are about $\pm 2\%$ (for dielectric constant) and $\pm 4\%$ for thickness.
3. A significant source for errors is in the etching process. A proper fabrication process has to include good control on the surface quality of the substrate, and adequate metalization thickness.

Bahl and Bhartia [8] present formulas for the sensitivity of some of the parameters discussed above. These formulas are reproduced below:

1. The change in the resonance frequency as a function of the variation of the length of the patch and the effective dielectric constant of the substrate

$$|\Delta f_r| = \sqrt{\left(\frac{\partial f_r}{\partial L}\Delta L\right)^2 + \left(\frac{\partial f_r}{\partial \varepsilon_e}\Delta \varepsilon_e\right)^2} \tag{28}$$

2. The change in effective dielectric constant of the substrate as a function of the variation of the width of the patch (W), the thickness of the substrate (h), the dielectric constant of the substrate (ε), and the thickness of the metalization (t) is

$$|\Delta \varepsilon_e| = \sqrt{\begin{array}{l}\left(\dfrac{\partial \varepsilon_e}{\partial W}\Delta W\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial h}\Delta h\right)^2 \\ + \left(\dfrac{\partial \varepsilon_e}{\partial \varepsilon_r}\Delta \varepsilon_r\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial t}\Delta t\right)^2\end{array}} \tag{29}$$

From the previous three equations we can derive the formula for the relative change in the resonance frequency:

$$\frac{|\Delta f_r|}{f_r} = \sqrt{\begin{array}{l}\left(\dfrac{\Delta L}{L}\right)^2 + \left(\dfrac{0.5}{\varepsilon_e}\right)^2 \left\{\left(\dfrac{\partial \varepsilon_e}{\partial W}\Delta W\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial h}\Delta h\right)^2 \right. \\ \left. + \left(\dfrac{\partial \varepsilon_e}{\partial \varepsilon_r}\Delta \varepsilon_r\right)^2 + \left(\dfrac{\partial \varepsilon_e}{\partial t}\Delta t\right)^2\right\}\end{array}} \tag{30}$$

Figure 11 shows the variation of change in the fractional resonant frequency of a rectangular microstrip antenna with frequency for $\varepsilon_r = 2.32$ and given tolerances. The complete design of a microstrip antenna has to factor also in polarization, frequency response (wideband, multiband), and feeding mechanism. The following sections address these issues in a broader context and present a large variety of geometries.

3. FEEDING METHODS FOR MICROSTRIP ANTENNAS

3.1. Microstrip Antenna Configurations

As they are essentially printed circuits, microstrip antennas can have different geometric shapes and dimensions. Since the early 1970s, numerous configurations have been proposed [11]. Some of these geometries are shown in Fig. 12. The electrical characteristics of these shapes are somewhat similar, all having a broadside beam generated by a fundamental mode. The slight difference in the physical area occupancy or multimode (or higher-mode) operation might make one geometry more appropriate for certain applications. All these geometries, however, can each be fed in similar ways.

3.2. Coaxial Feed

The coaxial feeding method is mostly appropriate for single elements. The location of the feeding point determines

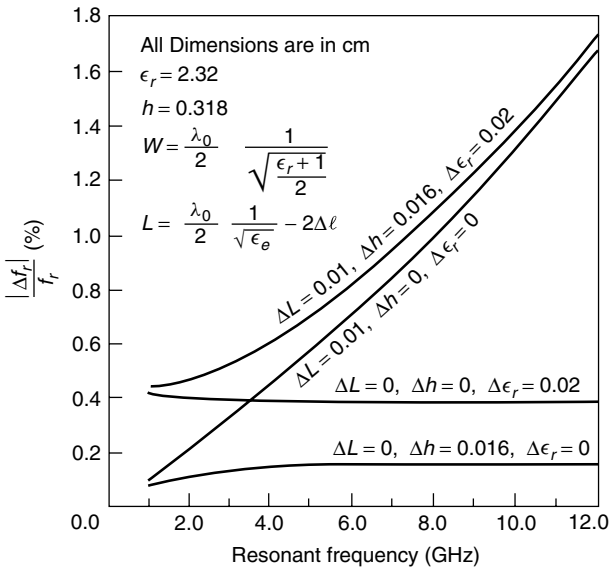


Figure 11. Variation of change on the fractional resonant frequency of a rectangular microstrip antenna with frequency for $\varepsilon_r = 2.32$ and given tolerances [8].

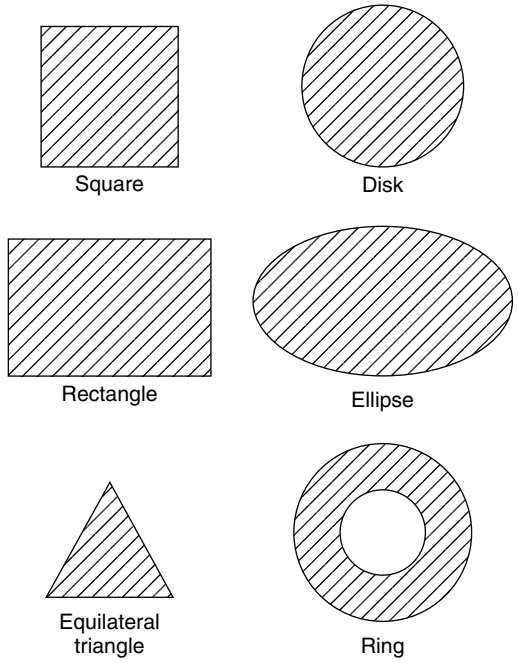


Figure 12. Basic microstrip patch antenna shapes commonly used in practice [8].

the input impedance and the polarization. The input impedance calculations of the single element will have to include besides the patch self-impedance a serial reactance component as shown in Eq. (12) The geometry of the coaxial fed microstrip patch is shown in Fig. 13. The central pin of the coaxial feed is connected to the patch at the "feeding point," and the shield of the coaxial feed is connected to the ground plane. A number of various coaxial-fed patch geometries are shown in Fig. 14.

3.3. Microstrip Line Feed

The coaxial-fed patch is not easy to array. In arrays, the most common way to feed the radiating element is using a microstrip line, which generally is an extension of the feed network.

The most used types of microstrip feeds are

1. The coplanar microstrip feed (Fig. 15)
2. The proximity (electromagnetic) coupled microstrip feed (Fig. 16)
3. The aperture coupled microstrip feed (Fig. 17)

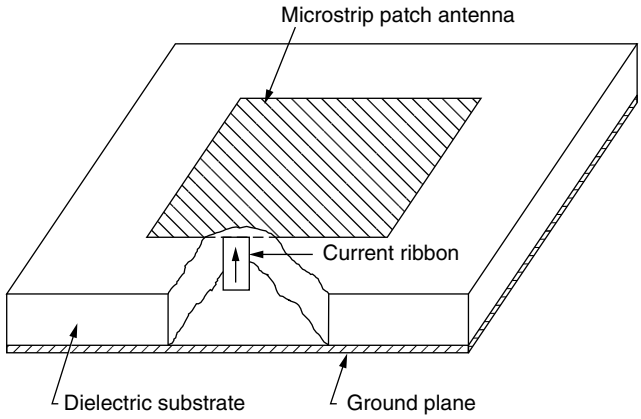


Figure 13. The coaxial-fed microstrip patch [8].

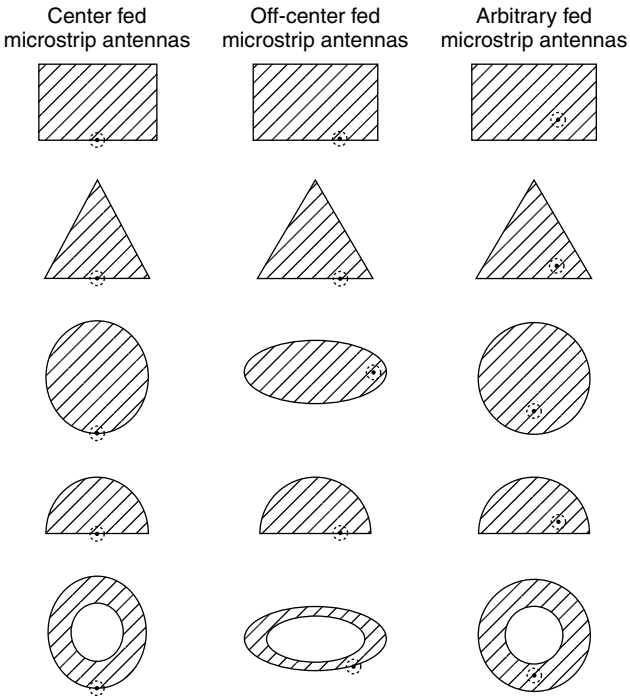


Figure 14. Coaxial fed microstrip antennas [8].

Figure 15 shows a number of variations of the coplanar microstrip feed: edge feed (a), gap feed (b), and inset feed (c).

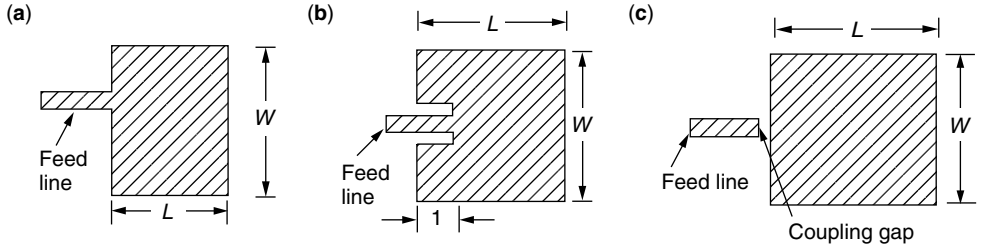


Figure 15. The coplanar microstrip feed.

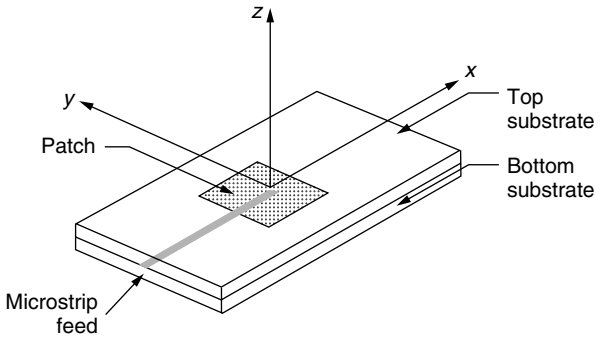


Figure 16. The proximity coupled microstrip patch.

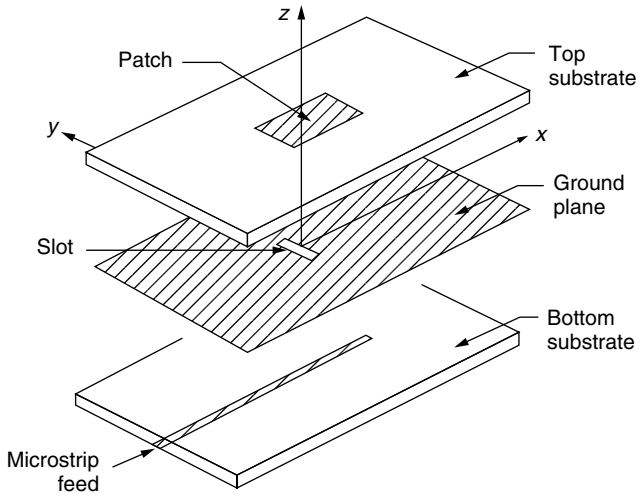


Figure 17. The aperture-coupled microstrip feed.

The patch impedance required to match the feed network determines the choice of any of the three. The patch self-impedance at the edge is typically high, and if 50 Ω is required, then the inset feed will have to reach inside the patch to a point where the patch self-impedance is 50 Ω. The coplanar microstrip feed has the advantage that it is printed on the same substrate as the radiating elements. In some cases, the design of such arrays might be difficult since the patches themselves might occupy most of the space on the substrate.

The mutual coupling between the radiating elements and the feeding network as well as the feeding network itself can create spurious radiation that might affect the overall performance of the array.

In order to avoid a crowded design, two substrates can be used, one for the radiating elements and one for the feeding network. In this case, the patches can be fed by a microstrip line sharing the same ground plane as the patch but located in between the patch and the ground plane (proximity feed). This feeding mechanism solves the "real estate" problem on the substrate; however, it does not address the spurious radiation problem. The total separation between the radiation of the radiating elements and the spurious radiation of the feeding network is achieved by using the aperture-coupled feeding method.

A comprehensive overview of the different type of feeding mechanisms is given by James and Hall [11] and Garg et al. [12].

4. POLARIZATION PROPERTIES OF MICROSTRIP ANTENNAS

4.1. Linear Polarization

In general, any rectangular or circular microstrip antenna fed in a symmetric way with respect to one axis will be linearly polarized. The difference between the different methods of feeding mentioned in the previous paragraphs is in the cross-polarization level. The probe feeding of a patch is symmetric in the H plane; however, it is not symmetric in the E plane, and this results in the excitation of cross-polarization currents. In addition, a thicker substrate implies a longer probe that radiates, and increases even more the cross-polarization level. The aperture-fed patch is symmetrically fed in both planes, and no cross-polarization currents are excited.

4.2. Circular Polarization

4.2.1. Singly Fed Circularly Polarized Microstrip Antennas. Traditionally, the singly fed circularly polarized microstrip antennas are very narrowband, both in terms of VSWR and axial ratio. A number of different geometries are shown in Fig. 18.

In general, circular polarization is obtained by superimposing two orthogonal current modes that are excited with equal amplitude and a phase differential of 90°. This can be achieved by introducing a perturbation segment, which excites a specific current distribution, consisting of

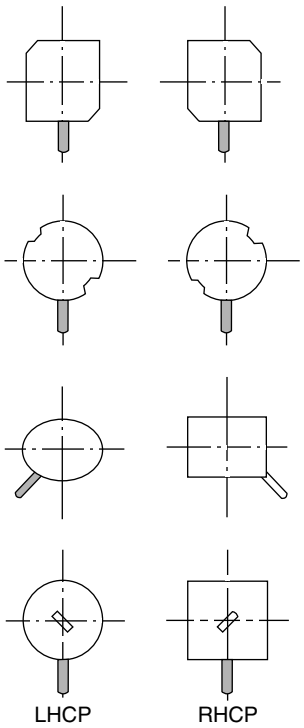


Figure 18. Singly fed circularly polarized patches [11].

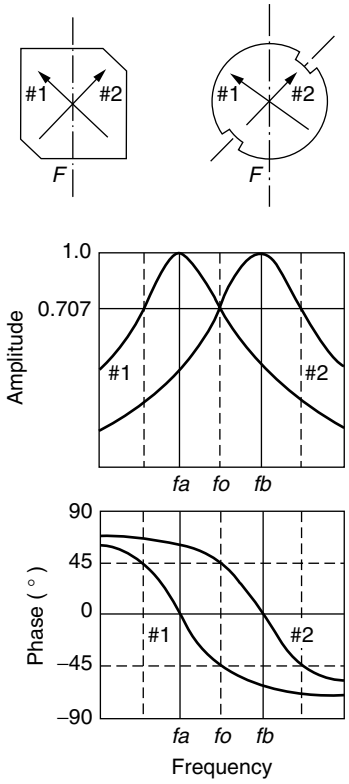


Figure 19. Amplitude and phase diagrams for singly fed circularly polarized microstrip antennas [11].

the two modes, resonant at slightly different frequencies. At the central frequency, the modes self-impedance fulfills the condition mentioned above (Fig. 19). This effect though, for single layered patches is very narrowband [6]. The advantage of the singly fed CP microstrip antennas is that they are easy to array, and in terms of array topology, they are similar to the linearly polarized version. Owing to their narrowband characteristics, they have few applications. For stacked patched however, a wider band can be achieved [13]. A detailed design procedure for the singly fed circularly polarized microstrip antennas is given in Chapter 4 of Ref. 11.

4.2.2. Dual-Fed Circularly Polarized Microstrip Antennas. When a wider band of operation is required (for VSWR and axial ratio), the dual-fed configuration is a better choice (Fig. 20). In this case, the excitation of the appropriate modes is done outside the radiating element. As shown in Fig. 20, the overall size of the element (which now includes the circuit generating the circular polarization) is considerably larger. In an array, a large element would force a large separation between elements, resulting in grating lobes.

An ingenious solution for the tradeoff between bandwidth and element size in arrays was presented by John Huang [14] (Fig. 21). The idea consists of creating circularly polarized subarrays from linearly polarized elements. This sequential feeding scheme allows for an excellent circular polarization over a relatively wide frequency bandwidth. Moreover, the array is capable of scanning in the principal planes to relatively wide angles from

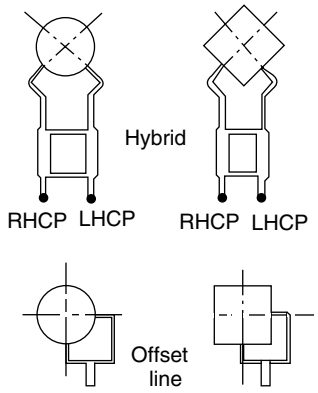


Figure 20. Dual-fed CP patches [11].

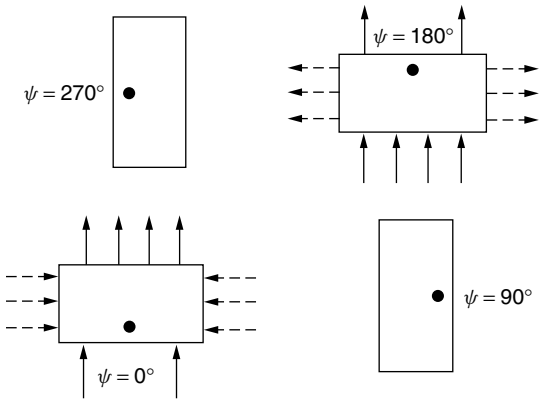


Figure 21. A 2×2 microstrip subarray that generates CP with LP elements [14].

its broadside direction without significant degradation to the axial ratio. This idea was developed further using singly fed circularly polarized instead of linearly polarized elements [15,16].

5. BANDWIDTH CHARACTERISTICS OF MICROSTRIP ANTENNAS

5.1. Introduction

The basic, single-layer microstrip antennas are $2\frac{1}{2}$D structures, and therefore, electrically very small. Since the early 1970s, many variations of the elementary patch were developed: multilayer microstrip antennas (or stacked patches), tall patches, cluster patches, and slotted patches. In essence, all these variations have as a goal the realization of radiating elements, which can be easily arrayed, with a radiation pattern similar to the elementary patch and that, finally, have a significantly larger bandwidth than the original microstrip patch antenna. All the techniques mentioned above effectively increase the electrical volume of the radiating elements, and generate radiating elements with a lower Q.

The difference between these techniques resides in the different tradeoffs they present, such as bandwidth versus physical volume, manufacturing cost, cross-polarization, and radiation pattern shape.

5.2. The Single Patch

The choice of the different parameters in the design of the single patch offers *some* latitude, even though quite limited, in the bandwidth characteristics:

1. The thicker the substrate, the wider the bandwidth.
2. The lower the dielectric constant of the substrate, the wider the bandwidth.

When the dielectric substrate is too thick, the efficiency and the crosspolarization of the antenna can be of concern; the surface wave dependency on the substrate thickness *is not* monotonic, and an excellent study of this phenomenon is given by Pozar [17]. However, when the substrate is excessively thick (allowing for the excitation of higher-order surface waves), the efficiency is considerably affected. Figures 22 and 23 show the surface wave efficiency as a function of the substrate thickness for two dielectrics, $\varepsilon_r = 2.55$ and $\varepsilon_r = 12.8$, respectively.

The losses due to the dielectric heating *are* monotonic and inversely proportional to the substrate thickness (Fig. 24). In addition, for a single-layer patch *resonating at a certain frequency*, by increasing the dielectric substrate, the directivity of the antenna is reduced. This is due to the fact that the antenna is smaller.

A special class of tall patches is the suspended patches. Rather than printing the patch on a grounded substrate,

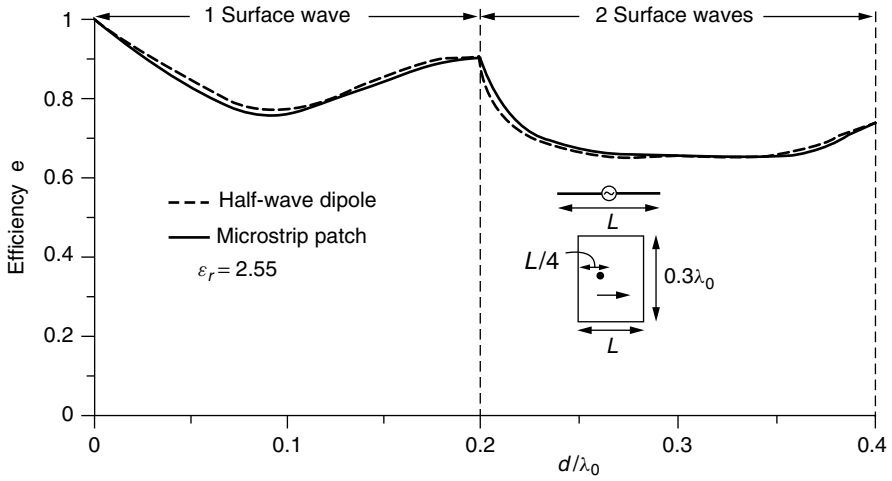


Figure 22. Loss due to surface wave for a half-wave printed dipole and a microstrip patch versus the substrate thickness for $\varepsilon_r = 2.55$, with patch width $0.3\lambda_0$ [17].

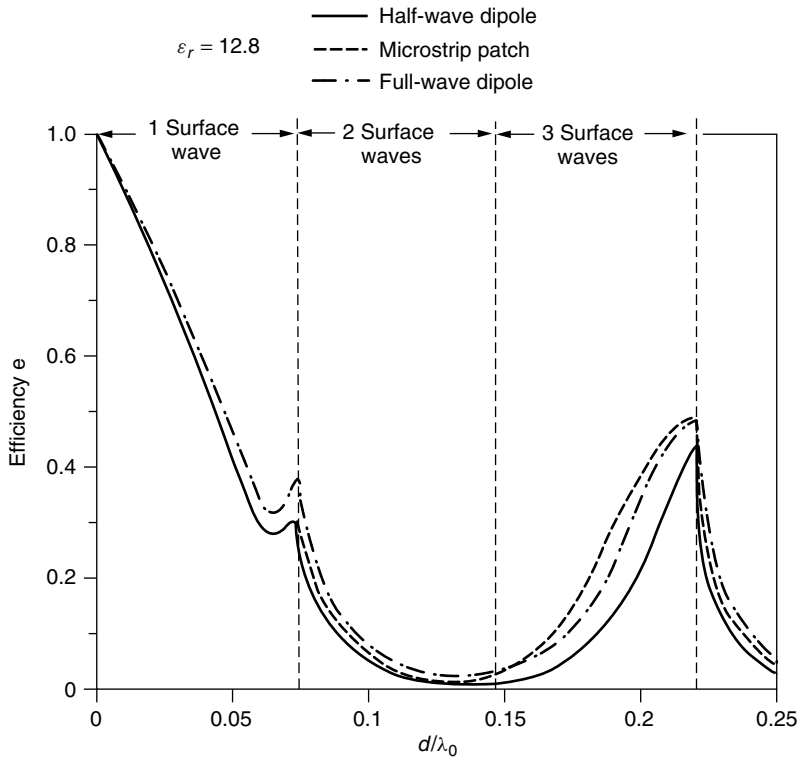


Figure 23. Loss due to surface wave for a half-wave printed dipole and a microstrip patch versus the substrate thickness for $\varepsilon_r = 12.8$, with patch width $0.15\lambda_0$ [17].

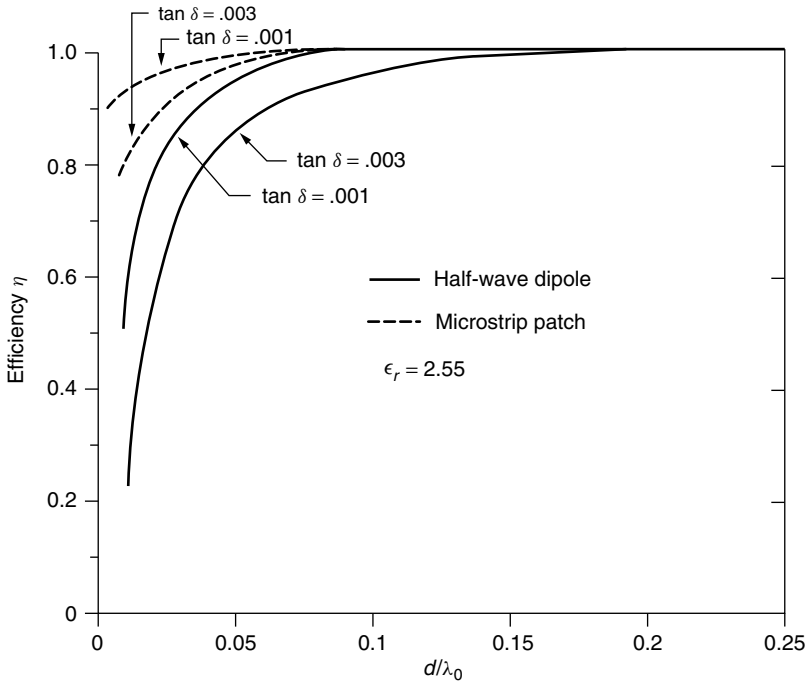


Figure 24. Loss due to dielectric for a half-wave printed dipole and a microstrip patch versus the substrate thickness for $\varepsilon_r = 2.55$, with patch width $0.3\lambda_0$ [17].

the patch is made of bare metal (or printed on a very thin dielectric substrate, typically 2–5 mils) and separated from the ground either by foam or by a supporting standoff. The widest bandwidth (in terms of VSWR) for a suspended patch reported in the literature is 95% [18].

The design reported in Ref.18 incorporates three principles of bandwidth enhancement.

1. Large separation between the patch and the ground plane
2. Low dielectric (air)
3. Multiresonant patch geometry

The geometry is shown in Fig. 25, and theoretical prediction and measurements are compared in Fig. 26.

The shape of the patch is such that different parts of the patch are resonant at different frequencies [19]. The distance between the radiating element and the ground plane is about $\lambda/4$ at midband, and using a $\lambda/4$ probe to feed the patch would allow the probe to radiate like a monopole. This is why a 3D transition was used to feed the patch. At least for 45% of the band, the cross-polarization of the element is better than 10 dB within the −3-dB beamwidth. On broadside, the cross-polarization is better than 30 dB. As shown in Fig. 26, the VSWR is less than 2 from 2.2 to 4.3 GHz. At frequencies higher than about 3.5 GHz, the 3D transition itself is radiating, and generates a high level of cross-polarization.

This example emphasizes the fact that the term *bandwidth* has to be carefully defined when referring to antenna performance; sometimes the *VSWR bandwidth* is different from the *cross-polarization bandwidth, directivity bandwidth, axial ratio*, and other parameters.

5.3. Nonresonant Methods for Bandwidth Enhancements

The simplest way to improve the frequency response of a microstrip antenna is to use a matching network. In Refs. 20 and 21, 10–12% bandwidths are reported for a relatively thin patch, using a lossless matching network. Lossless matching networks, though, have only a limited impact, and in some cases they might occupy too much

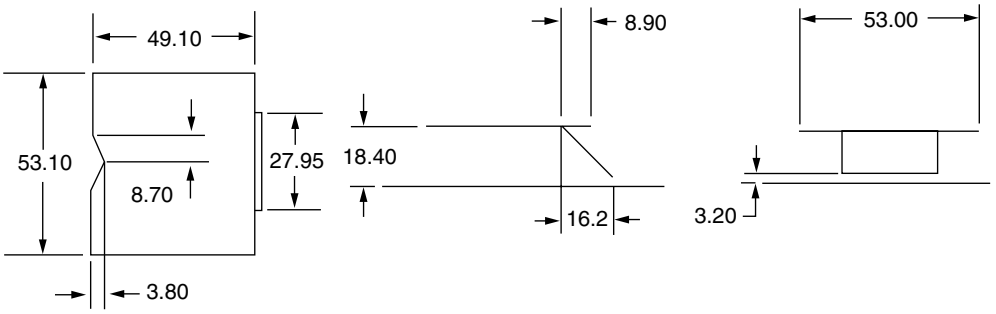


Figure 25. Dimensions (in millimeters) for the wideband microstrip single-layer patch antenna [18].

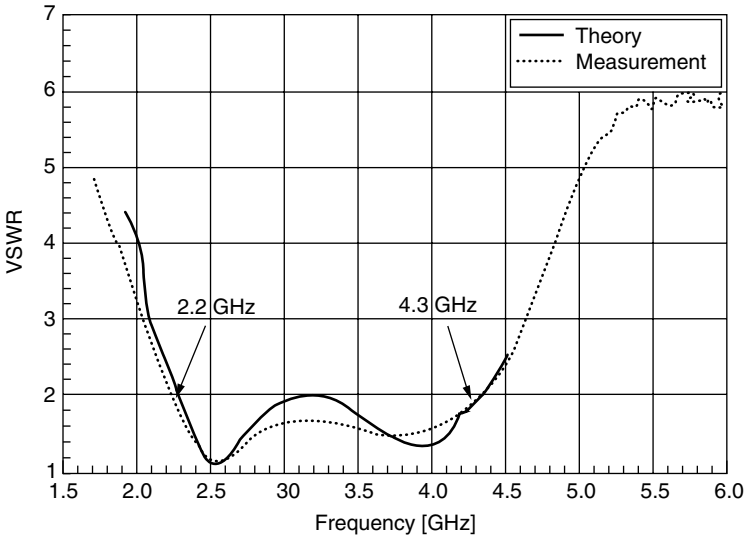


Figure 26. The VSWR of a wideband microstrip single-layer patch antenna [18].

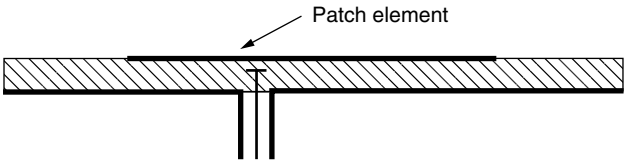


Figure 27. Capacitive feeding of a single patch.

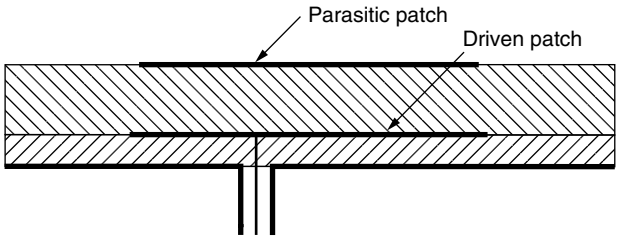


Figure 28. Bandwidth improvement using stacked patches.

space on the board. In addition, if the matching network were too complex, they would create spurious radiation.

A simple matching technique is capacitive feeding (Fig. 27). The matching is achieved by controlling the size of the tab and its distance from the patch.

5.4. Multiresonator Microstrip Antennas

5.4.1. The Stacked Patch. The single patch can be considered as a resonator. By adding an additional patch (Fig. 28), an additional resonator is created. By setting the resonance dimensions of the driven patch and the parasitic patch appropriately, the broadband or dual-band effect can obtained. The physical interpretation (or the equivalent circuit) of such a structure is extremely difficult to generate, mainly because of the mutual coupling between these two resonators. Therefore, only full-wave modeling can provide a good prediction of the electrical characteristics of this antenna [22,23]. Table 1 gives an idea of the bandwidths that can be achieved [24].

The case of the stacked patches is special in the sense that the radiating element has almost the same size (in the substrate plane) as the single patch itself, and it does not require additional space. The two patches have to be very close in size to obtain the broadband effect. The other methods for bandwidth enhancement, described below, involve larger elements and/or some price to pay in performance (front-to-back design, cross-polarization, complexity of fabrication, etc.).

Intuitively, the next step would be to add more parasitic elements. Since the excitation of the parasitic element is by coupling, the broadband effect is lost very quickly.

5.4.2. Coplanar Parasitic Elements. A different way to use parasitic elements is in the coplanar configuration. Figure 29 shows the geometry of a probe-fed patch

Table 1. Experimental Results for Stacked Two Layer Antennas [21]

Antenna Geometry	Frequency Band	Bandwidth (%)	Beamwidth H-Plane (Degrees)	Gain (dbi)	Sidelobe Levels H-Plane (dB)	Polarization
Circular disk	S	15	72	7.9	−22	Linear
Circular annular disk	S	11.5	78	6.6	−14	Linear
Rectangular	S	9	70	7.4	−25	Linear
Square	S	9	72	7	−22	Linear
Circular disk	S	10	72	7.5	−22	Circular
Circular disk	X	15	72	7.5	−25	Circular

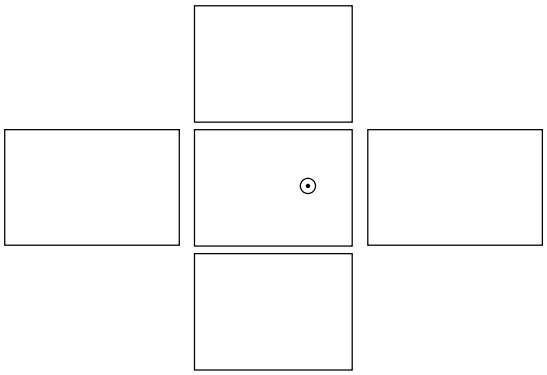


Figure 29. A probe-fed patch, with four edge-coupled parasitic elements [25].

feeding four coplanar parasitic patches. In order to obtain the appropriate level of coupling, the gap between the patches has to be very small. That requires very tight tolerances in fabrication, which might be difficult to achieve. Bandwidths up to 25% have been reported [25–26] (Fig. 30); however, control over the shape of the beams might be difficult (Fig. 31). Because of the tight tolerances required in the fabrication of edge-coupled elements, direct coupling was proposed [27]. Figure 32 shows the proposed geometry. As shown in Fig. 33, the experimental bandwidth is 810 MHz (24% at $f_o = 3.38$ GHz), which is about 7.4 times the bandwidth of the typical rectangular patch antenna printed on the same substrate. The radiation patterns (Fig. 34), however, vary quite significantly across the operating frequency band, which might be unacceptable in some applications.

5.4.3. Aperture-Coupled Microstrip Antennas. The basic configuration of the aperture-coupled microstrip antenna is shown in Fig. 17. Initially developed as a way

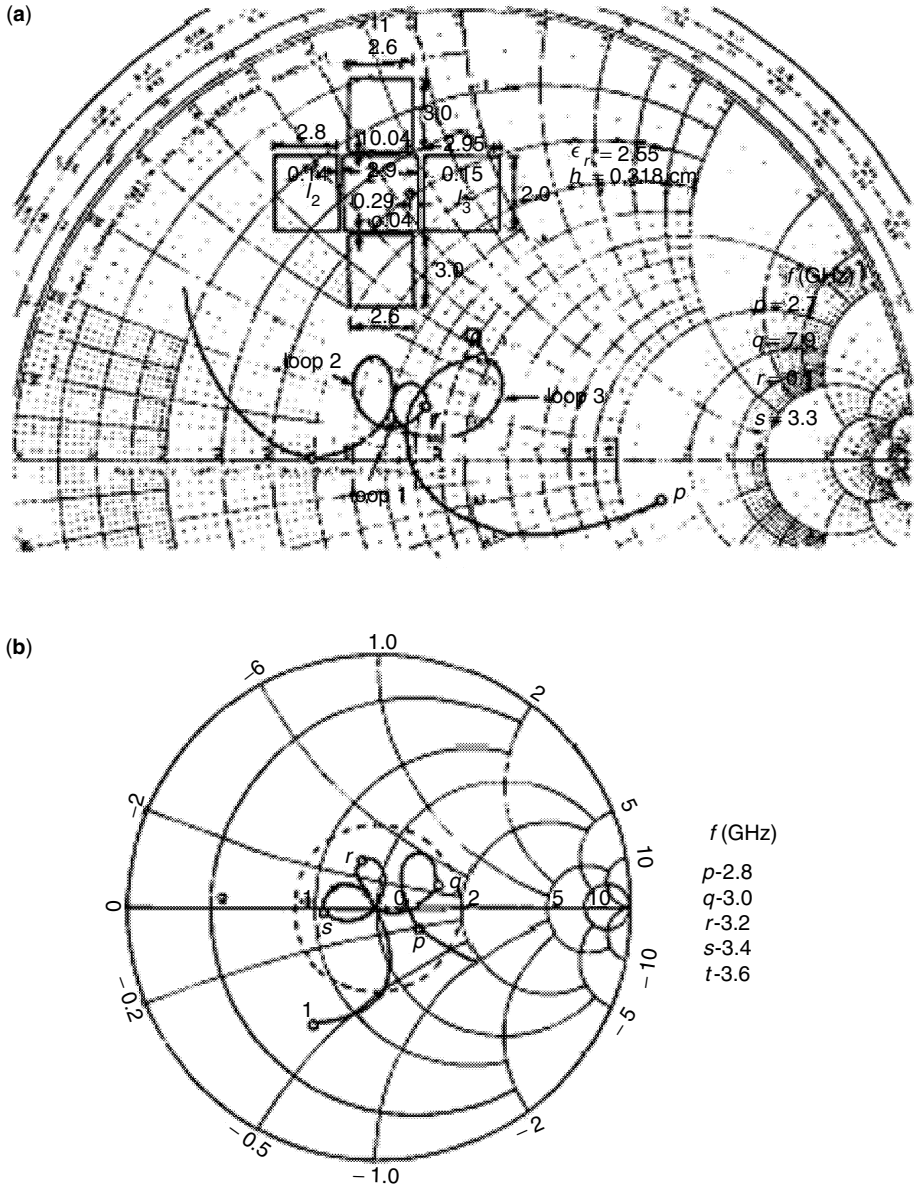


Figure 30. (**a**) Theoretical input impedance locus of FEGCOMA shown in inset and (**b**) experimental input impedance locus of FEGCOMA with modified dimensions [25].

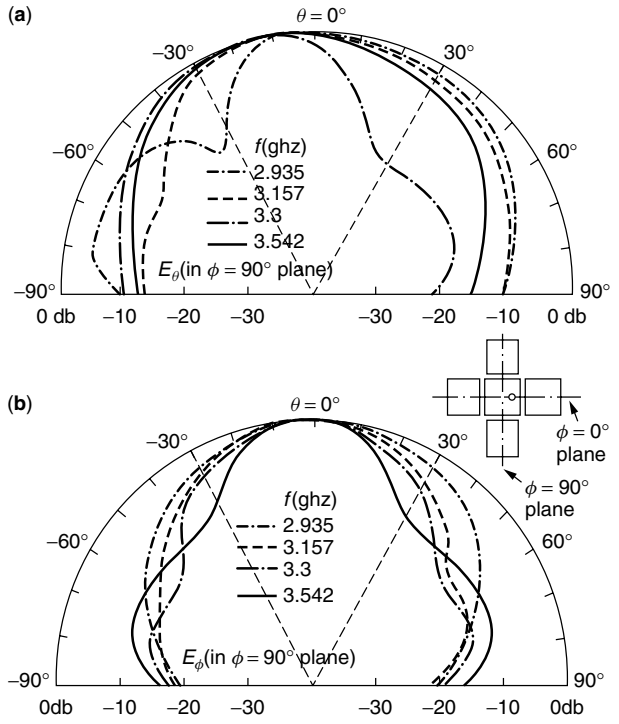


Figure 31. Experimental values of the radiation fields of FEGCOMA [25]: (**a**) E_θ in $\phi = 0^\circ$ plane and (**b**) E_ϕ in $\phi = 90^\circ$ plane.

to separate the feeding line from the radiating element, the aperture-coupled patch introduces an additional resonator: the coupling aperture. To avoid back radiation, the coupling aperture should not be resonant; however, its resonance can be *close* to the patch resonance, so that the antenna bandwidth is slightly increased. A number of different geometries based on the aperture-coupled microstrip antenna were developed:

1. The aperture-coupled coplanar dipole array [28] is shown in Figs. 35 and 36. Croq and Pozar [28] discuss only the multiband case; however, this geometry is conceivably appropriate for broadband applications.
2. The aperture-coupled stacked patch antenna, which, as reported [29,30], can achieve 50% bandwidth, is shown in Fig. 37. When using the aperture to feed the radiating elements, usually the tradeoff is between bandwidth and the amount of back radiation allowed. Some attempts were made to suppress the backradiation; a shielding plane was placed behind the antenna. While the back radiation is reduced, the shielding plane allows for the excitation of parallel-plate modes, which can seriously degrade the efficiency of the antenna. Furthermore, this bandwidth enhancement is done at the expense of much greater manufacturing complexity.

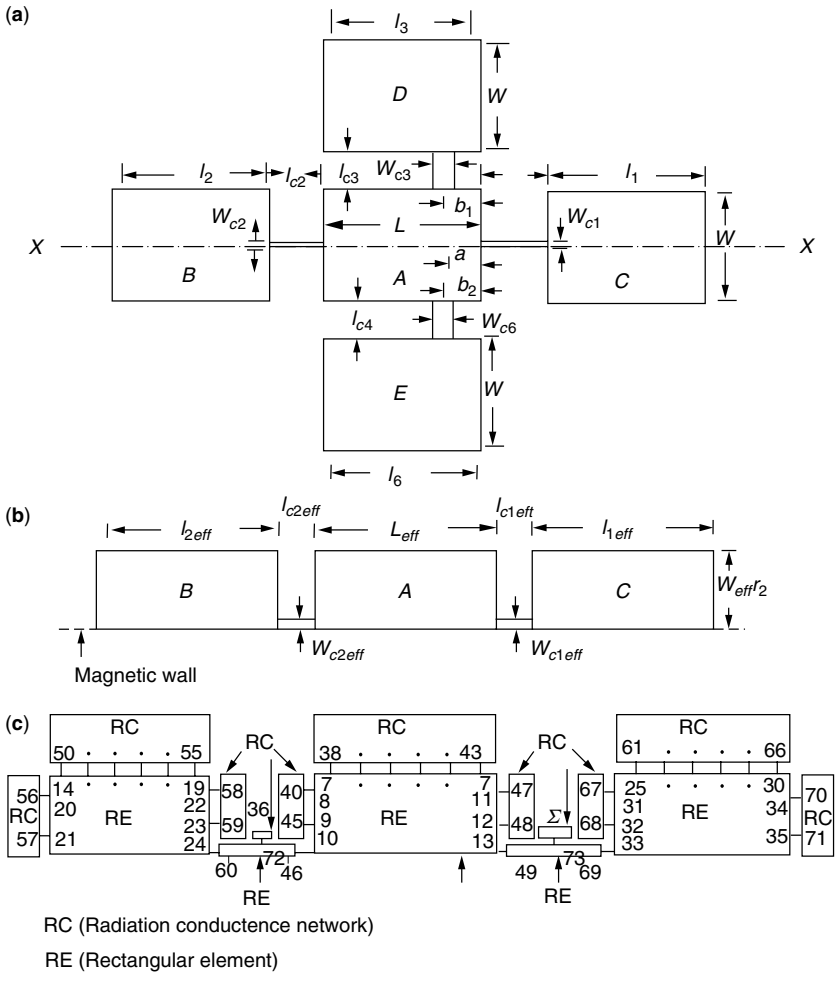


Figure 32. (**a**) Four edges directly coupled microstrip antenna (FEDCOMA) [26]; (**b**) even-mode half-section of REDCOMA; and (**c**) its segmented network.

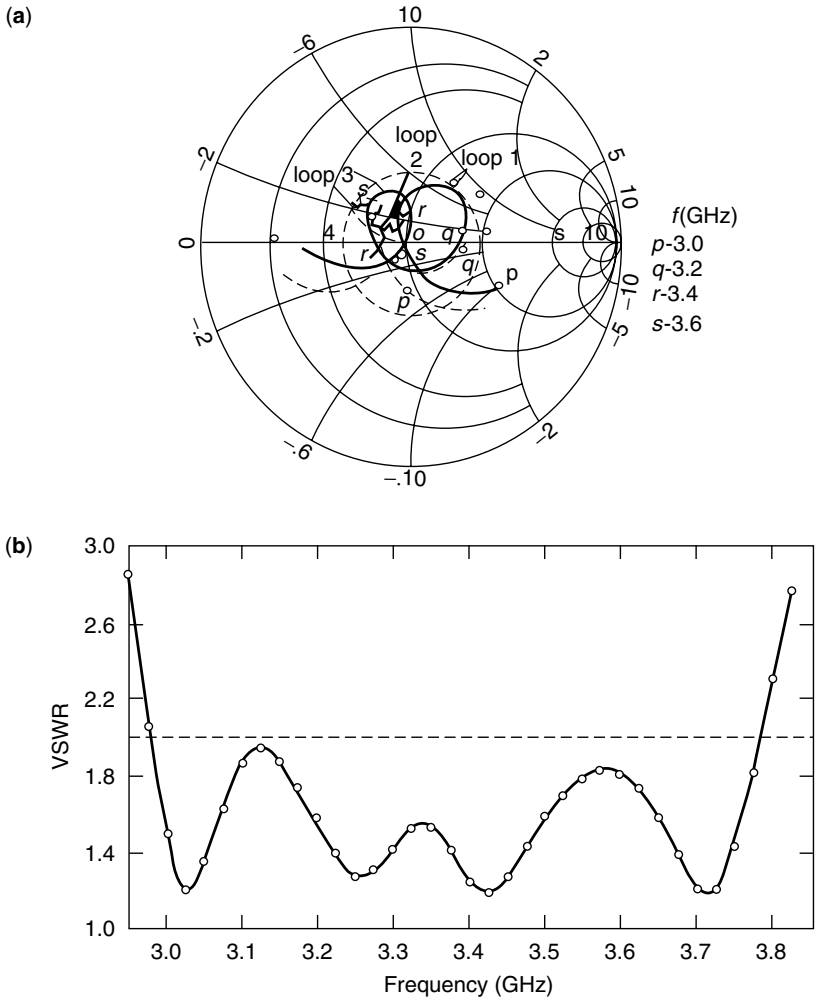


Figure 33. (**a**) Theoretical (---) and experimental (-o-) input impedance loci and (**b**) experimental VSWR variation with frequency of FEDCOMA [27].

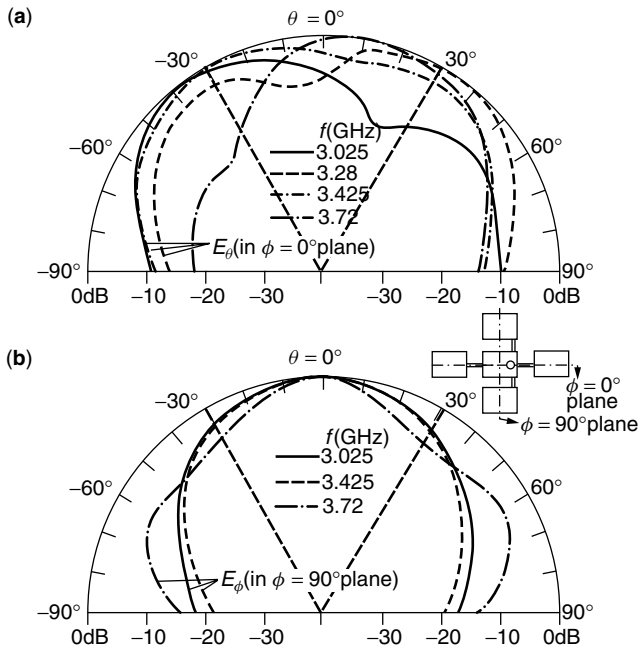


Figure 34. Experimental values of (**a**) E_θ in $\phi = 0^\circ$ plane and (**b**) E_ϕ in $\phi = 90^\circ$ plane of FEDCOMA [27].

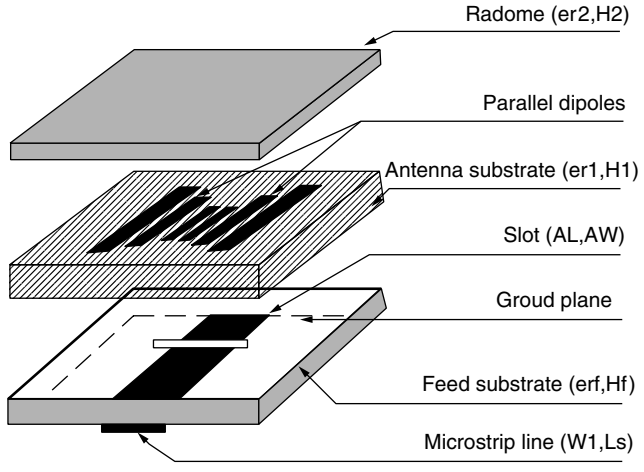


Figure 35. Multifrequency microstrip antenna composed of parallel dipoles aperture-coupled to a microstrip line [28].

6. MUTUAL COUPLING

In an array, the mutual coupling between elements can have a significant impact on the array radiation pattern as well as the input impedance. It can be either calculated

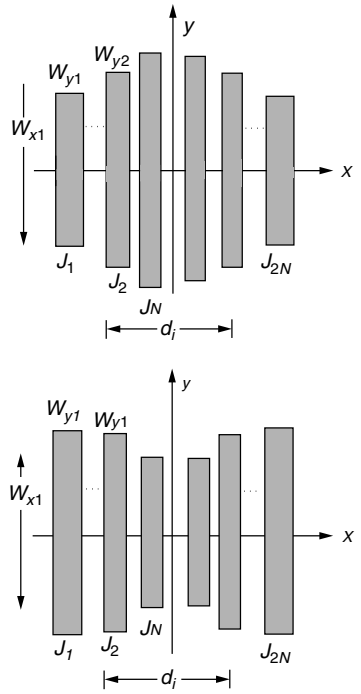


Figure 36. Two configurations of the multiple-resonator aperture-coupled antenna: (**a**) MFR1 and (**b**) MFRZ [28].

or measured, so that the feed network design can be modified to compensate (where possible) for its effect. In very large scanning arrays, the mutual coupling can create *blindness*, a situation in which the input reflection coefficient is very close to 1. The blindness effect will be discussed in Section 7.

The mutual coupling mechanism in microstrip antennas consists of two components: the radiation and the surface waves. A full-wave analysis of the mutual coupling between rectangular microstrip antennas is presented by Pozar [32] (Fig. 38). When the patches are very close (less than about $\lambda/10$), the *H*-plane coupling is slightly stronger than the *E*-plane coupling; however for larger distances, the *E*-plane coupling is significantly stronger, due to the excitation of surface waves. The mutual coupling depends on the substrate and the shape of the patch. A measurement study is presented by Jedlicka et al. [33]. Figure 39 shows the effect of the mutual coupling on its input impedance, and Figs. 40–42 show the mutual coupling for the principal planes and different geometries.

7. MICROSTRIP ARRAYS

7.1. Introduction

The electrical characteristics of microstrip antenna elements were described above. However, they are even

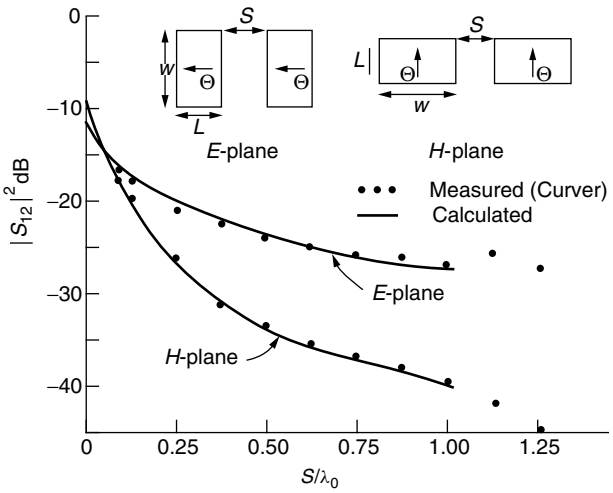


Figure 38. Measured and calculated mutual coupling between two coaxial-fed microstrip antennas for both *E*-plane and *H*-plane coupling ($W = 10.57$ cm, $L = 6.55$ cm, $d = 0.1588$ cm, $\varepsilon_r = 2.55$, $f_0 = 1410$ MHz) [32].

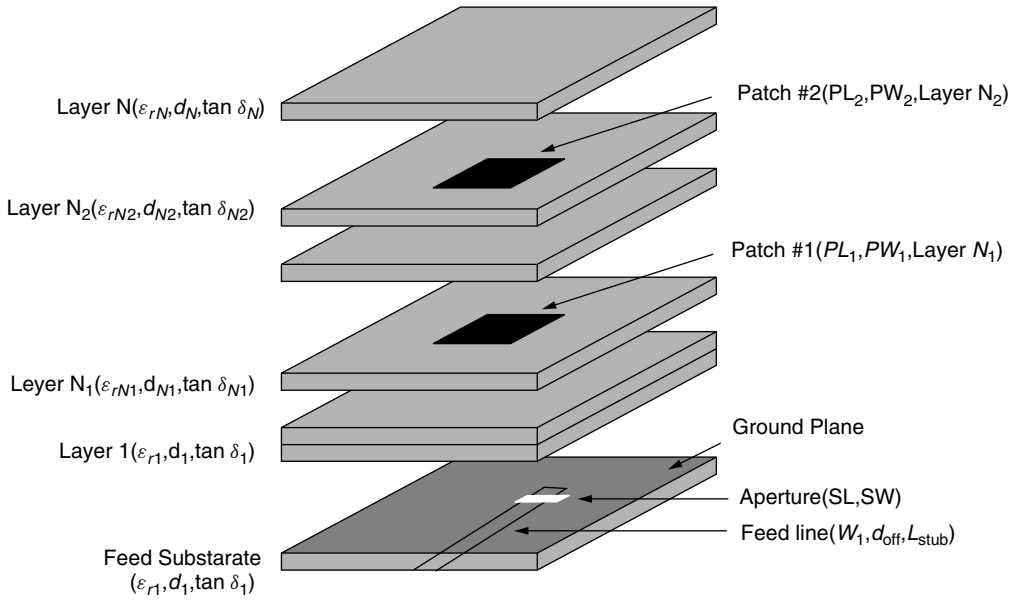


Figure 37. The wideband aperture-coupled stacked patch microstrip antenna [30].

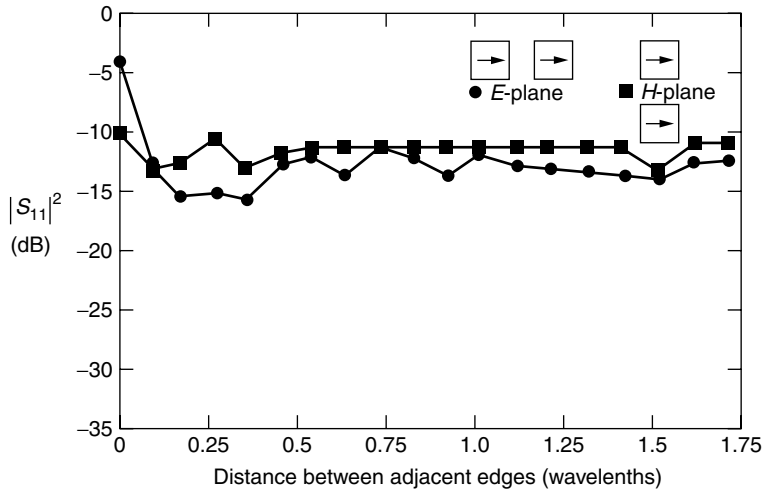


Figure 39. Measured $|S_{11}|^2$ values at 1410 MHz for 10.57 cm (radiating edge) × 6.55 cm rectangular patches with 0.1575 cm substrate thickness [33].

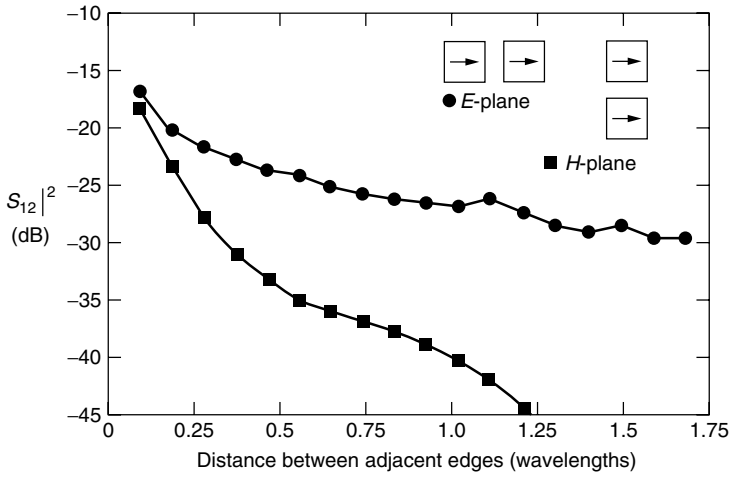


Figure 40. Measured $|S_{12}|^2$ values for the rectangular patch of Fig. 39 [38].

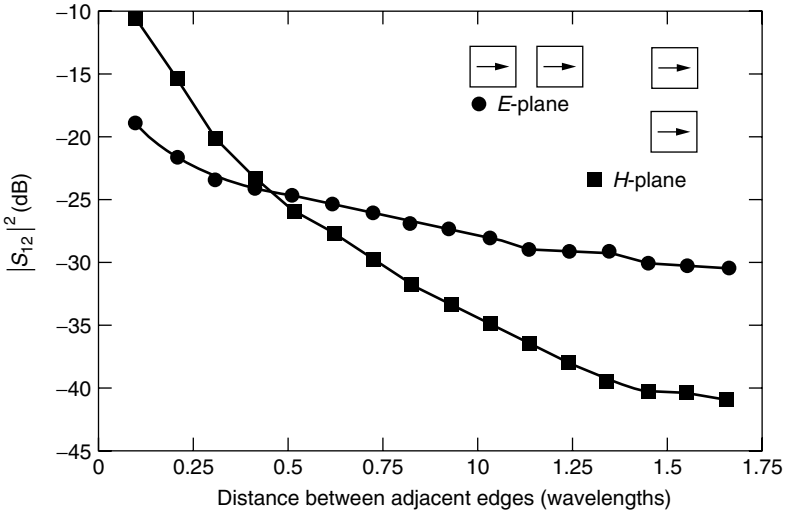


Figure 41. Measured $|S_{12}|^2$ values at 1560 MHz for 5.0 cm (radiating edge) × 6.0-cm nearly square patches with 0.305 cm substrate thickness [33].

more attractive in the array context. Compared to other types of arrays, microstrip arrays are relatively easy to manufacture, and are light and conformal. Since they essentially are printed circuits, they allow a significant freedom of design that results in a large variety of configurations: serial-fed arrays, parallel-fed arrays, and a significant number of different combinations between serial and parallel feeding techniques.

Array antennas in general and microstrip arrays in particular can be designed to have a fixed beam of a certain

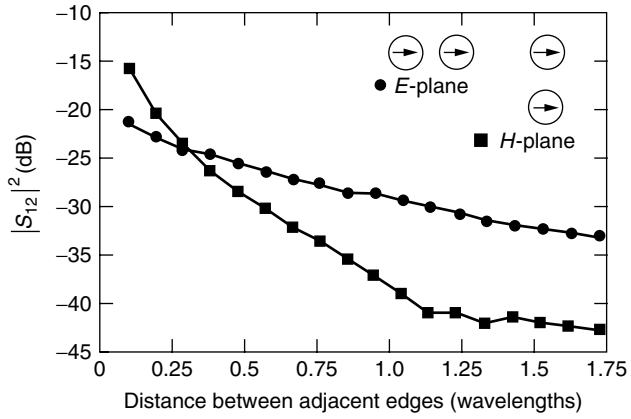


Figure 42. Measured $|S_{12}|^2$ values at 1440 MHz for circular patches with a 3.85-cm radius and a feed point location at 1.1 cm radius. The substrate thickness is 0.1575 cm [33].

shape or a beam that scans (using phase shifters or time-delay devices) or multiple beams (where the elements are fed by a special space feed).

7.2. Linear Arrays

By definition, linear arrays consist of radiating elements positioned at finite distances from each other along a straight line. In terms of the feed mechanism, linear arrays can be

1. Parallel-fed by a printed power divider
2. Serially fed with two-port radiating elements
3. Serially fed with one-port radiating elements

The parallel-fed array simply uses a power divider (usually printed on the same substrate as the radiating elements) to feed each radiating element (Fig. 43a). The junctions can be symmetric (for uniform amplitude) or asymmetric (e.g., for low-sidelobe design). In the case of low-sidelobe arrays, the bandwidth is determined not only by the bandwidth of the radiating elements but also by the bandwidth of the feed network. An important factor is played by the power dividers; the commonly used T junction does not have a very good isolation, and if the whole circuit is not very well matched, the amplitude and phase distribution generated will have errors. To alleviate this problem, Wilkinson power dividers could be used instead. When the electrical distances between the input port and all the other ports are identical, the phase distribution obtained is uniform and the beam generated is "squintless." The direction of the beam is independent of frequency and also of the spacing between elements.

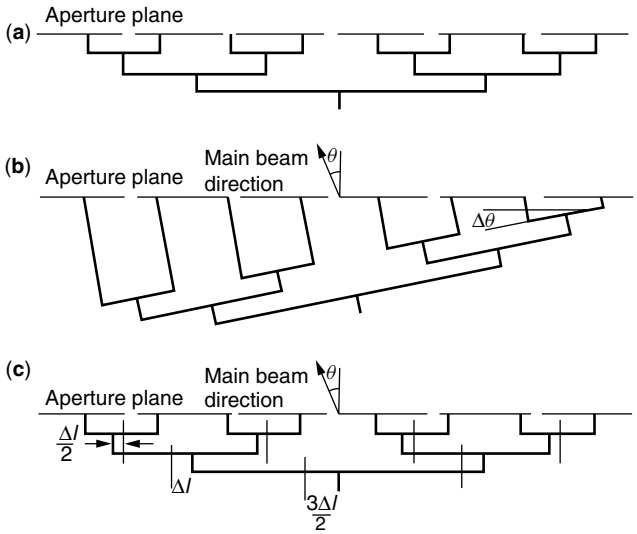


Figure 43. Parallel networks (**a**) without beam scan; (**b**) with beam scan, type 1; and (**c**) with beam scan, type 2.

A parallel feed network can be used to produce a scanning beam by simply using delay lines (Fig. 43b). The scanning angle is given by

$$\theta_0 = \arcsin\left(\frac{\delta}{2\pi}\frac{\lambda_0}{d}\right) \tag{31}$$

$$\delta = 2\pi\frac{\Delta l}{\lambda_t} \tag{32}$$

where δ = incremental phase difference between consecutive elements
d = distance between consecutive elements
λ_0 = wavelength in free space
λ_t = wavelength in transmission line
Δl = transmission line extension rate from one element to another

As shown in Eq. (31), the squint angle varies with frequency *and* the distance between elements. Another variation of this array is shown in Fig. 43c. As in the previous case, here, too, the phase gradient is realized using true time-delay lines, and the difference between the three layouts is in the implementation. Parallel feed networks (of the type shown in Fig. 44a) are typically used when a nonscanning (with frequency) beam is required. However, they are relatively complex (especially if low sidelobes are required) and therefore occupy significant space on the board. In addition, they radiate, and their spurious radiation can interfere with the radiation of the radiation elements, affecting the sidelobe level and/or the cross-polarization level of the whole array.

When the bandwidth of the array is very small (~1–2%), serial feeds can be used. Serial feeds have been used for decades in slotted waveguide arrays. In the waveguide case, they are of two types: traveling-wave arrays and resonant arrays. The resonant arrays end up in a short, and with the radiating elements separated by

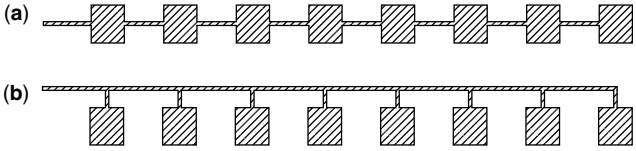


Figure 44. (**a**) two-port serial fed array; (**b**) One-port serial fed array.

a half-guide wavelength, all elements can be fed in phase so that a broadside beam can be obtained.

The traveling-wave arrays are fed at one end and are terminated into a matched load. Since most of the power is radiated through the slots, the matching load absorbs only a small fraction of the incident wave. The spacing between elements is not a half-guide wavelength (to avoid reflection in phase), and the direction of the beam is never broadside.

The main problem with waveguides is that the characteristic impedance of the waveguide cannot be (easily) changed. In the case of microstrip lines, this can be done by simply changing the width of the transmission line. This allows for a greater freedom of design and more types of serially fed microstrip arrays concepts to be introduced. Moreover, the serially fed microstrip array can have any polarization, unlike the slotted waveguide, where the polarization can be linear only.

The wide variety of serially fed microstrip arrays makes it somewhat difficult to divide them into specific groups. One classification would be in arrays where the microstrip element is used as both a two-port device and a one-port device (Fig. 44). In both cases resonant and traveling-wave arrays can be designed. The methods of feeding can vary: microstrip line or aperture-fed (Fig. 45). A full design procedure based on the transmission line model of these arrays is given in Chapter 14 of Ref. 11.

An excellent example of a shaped beam serially fed microstrip array is shown in Fig. 46 [34]. The design is based on the transmission-line model for the patches with measured values for the different components. The widths of the patches and their location are calculated so as to produce the desired radiation pattern.

Figure 47 shows the comparison between the calculated pattern and the measured one, while Fig. 48 shows the change of the radiation pattern with frequency. The one-port microstrip serially fed antennas allow for even greater freedom of design.

Three types of arrays have been described [35]:

1. The first array uses a standard standing-wave feed design. As in Ref. 34, the patch width is varied in order to obtain the desired amplitude taper (Fig. 49a). The transmission line connecting the patch to the main feeder is $\lambda_g/2$ long, so it transforms the input impedance of each patch directly to the main feeder. The characteristic impedance of the main feeder is constant for its entire length.
2. The second design also uses patches of varying widths, but the main feedline is matched at each patch tap point (Fig. 49b).
3. The third array uses a center-fed feed network with each half designed as a traveling-wave array with a main beam angle slightly off broadside. The combination of both halves will yield a broadside beam, which does not *scan* with frequency, but whose shape *will* slightly vary with frequency (Fig. 49c).

Three 16-element arrays with a 22-dB sidelobe level for each of the designs described above were designed and tested (see Table 2 [36]).

In addition to the two classes described above, other types of serially fed microstrip arrays can be mentioned:

1. Linear array with capacitively coupled microstrip patches (Fig. 50)
2. Comb-Line array with microstrip stubs (Fig. 51)

7.3. Planar Arrays

When a narrow pencil beam is required in both planes, planar arrays (rather than linear arrays) have to be used.

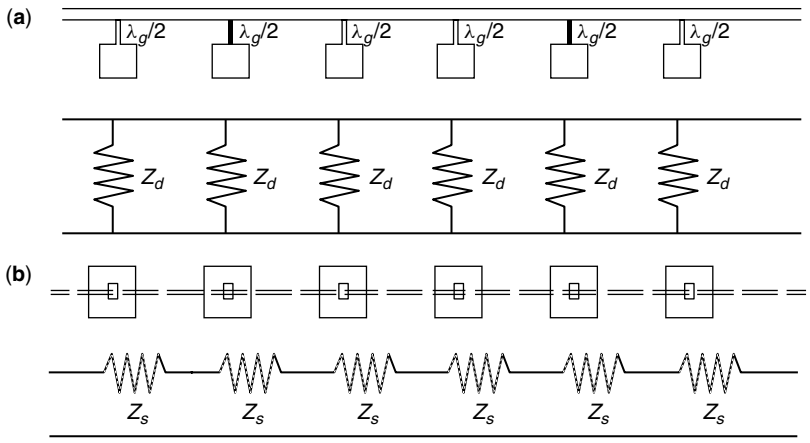


Figure 45. Two-port serial fed microstrip arrays: (**a**) microstrip fed and (**b**) aperture-coupled fed. (Courtesy of Prof. D. M. Pozar, Univ. of Massachusetts at Amherst).

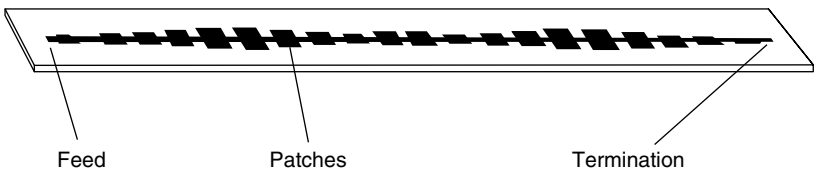


Figure 46. A serial fed microstrip array with a cosec^2 pattern [34].

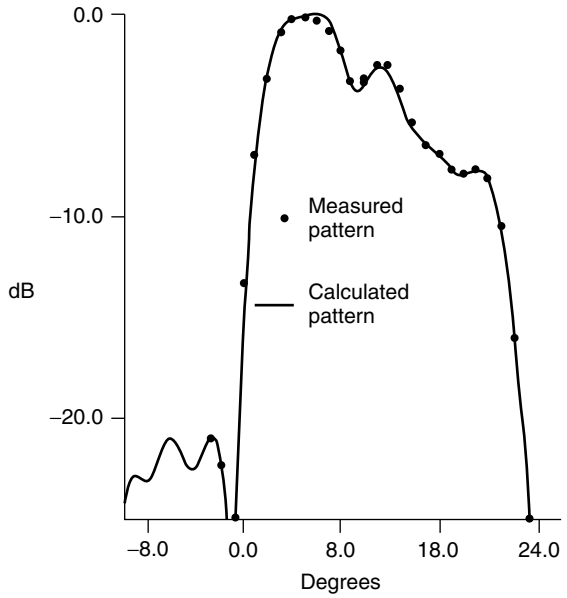


Figure 47. Comparison of measured and calculated amplitude patterns of the cosec^2 patch array [34].

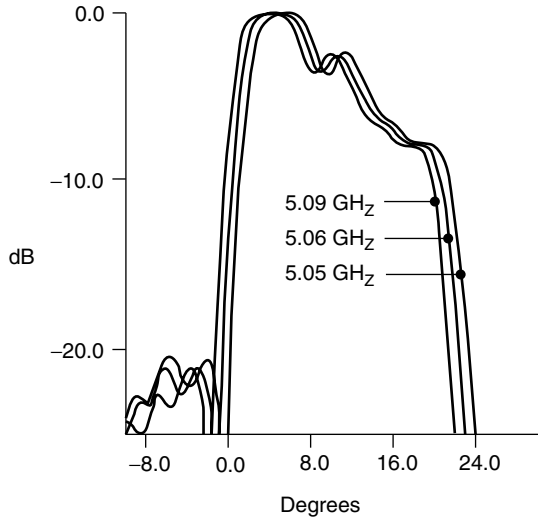


Figure 48. Effect on pattern of a 1% change in frequency [34].

Many configurations have been proposed for planar arrays, in which serial feed and/or parallel feed subarrays can be combined.

The main difficulty in planar arrays is the limited space within the unit cell. To avoid grating lobes, the unit cell has to be no larger than about 0.5–0.8 wavelength, while the typical patch printed on a low dielectric substrate is about 0.3–0.4 wavelength. Therefore, except for arrays with uniform amplitude and phase distribution (where the parallel feed is relatively simple to implement; see Fig. 52), the parallel feed network could be very complex. Such a feed network would couple to the radiating elements, resulting in significant degradation of array performance. In some cases, the degradation of the sidelobe level can be up to 10 dB [41]. To alleviate this problem, a combination between a corporate feed and a parallel feed can be used. Some examples are shown in Fig. 53.

In Fig. 53a square patches are used to yield the same resonance frequency for the two polarizations. For each polarization, four serially fed subarrays are combined by means of a parallel feeding network.

An interesting method to control the sidelobe level has been proposed [40] (Fig. 53b). The power tapering is achieved by connecting the equally wide patches diagonally. Changing the slope of the connecting feeding lines controls the beamwidth. Figure 53c shows a 9×9-element array. Here the serial feed is used in both planes. This is an example of the planar form comb array. The taper required for sidelobes is obtained by simply assigning the appropriate width to the microstrip stubs. An example of interlaced networks is shown Fig. 53d. The antenna consists of two arrays: one operating at 2.45 GHz and the other at 5.8 GHz. The 2.45-GHz radiating element is a rectangular stacked patch with a high aspect ratio. The input impedance of this element is 200 Ω, so a 3D linear transformer was required to reduce the impedance to 100 Ω. Note that the transformer does not change in *width* but in *height* above the ground plane. The length of the transformer was determined to match the bandwidth requirements. The 5.8-GHz elements are suspended square patches. In this design, the interlaced architecture is possible only because of the use of serially fed arrays.

7.4. Scanning Arrays

The array's scanning capability is frequently used in many military as well as commercial applications. This can be done electronically to achieve continuous coverage at very high scan rates. Unlike the situation in mechanical scanning, where the whole antenna is rotated, in electronic scanning, the radiating aperture is fed the appropriate phase distribution, which controls the direction of the main beam. The direction of the main beam is given by Eq. (31).

By scanning the beam of an array, besides the direction of the main beam axis, most of the array

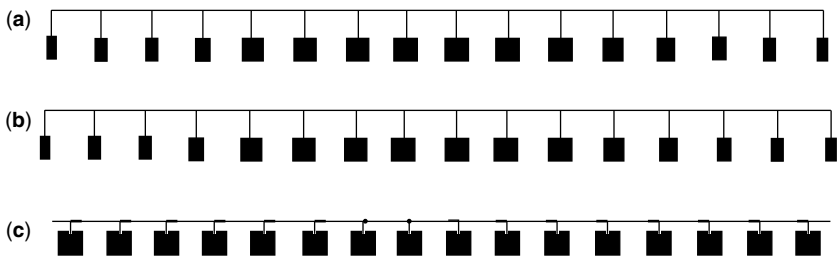


Figure 49. Series fed array designs: (**a**) standing-wave array with element spacing of λ_g; (**b**) traveling-wave array using matched feedlines and an element spacing of λ_g; and (**c**) traveling-wave array with element spacing less than λ_g [35].

Table 2. Summarized Performance of Three Arrays

Array Type/ Parameter	Standing-Wave Array	Matched Traveling-Wave Array	Phase-Compensated Traveling-Wave Array
Impedance BW (calculated)	1.8%	2.0%	4.0%
Impedance BW (measured)	1.7%	1.3%	4.2%
Directivity (calculated)	18.9 dB	18.9 dB	17.9 dB
Gain (calculated)	17.8 dB	17.8 dB	17.3 dB
Gain (measured)	17.4 dB	16.9 dB	16.5 dB
Efficiency (calculated)	77%	78%	88%
Sidelobe level (design)	22 dB	22 dB	20 dB
Sidelobe level (measured)	21 dB	22 dB	21 dB
Pattern BW[a] (measured)	2.3%	2.3%	12%

[a]For sidelobe level remaining below 13 dB.
Source: Courtesy of Prof. David Pozar, University of Massachusetts at Amherst.

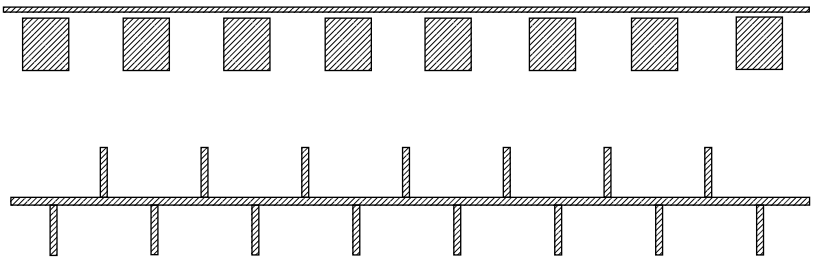


Figure 50. Linear array with capacitively coupled microstrip patches.

Figure 51. Comb-line array with microstrip stubs.

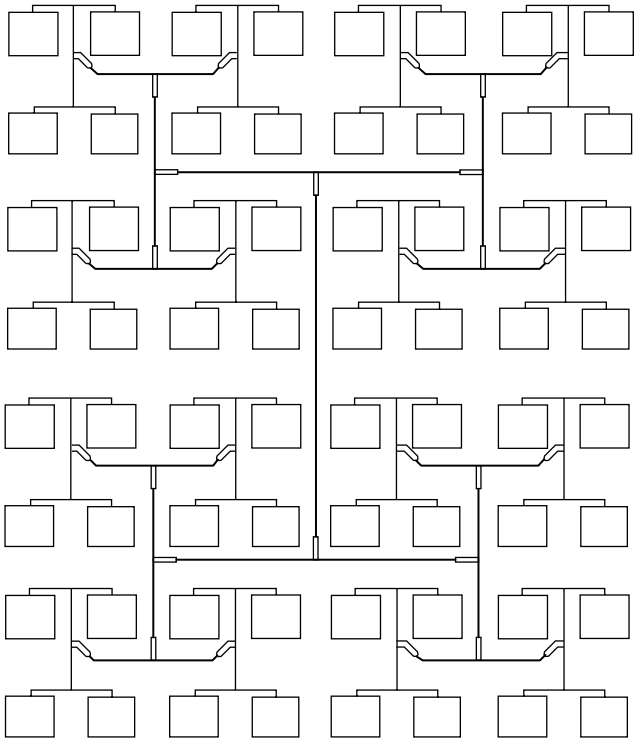


Figure 52. Typical parallel fed microstrip array [38].

characteristics change: beamwidth, radiation pattern, and input impedance. This is because the input impedance and the radiation pattern of each individual element in the array change, as well as the mutual coupling between elements. The *active-element pattern* of an element in an array is defined as the radiation pattern of the array when only that element is driven and all other elements are terminated in matched loads. In the absence of grating lobes, it can be shown that the active-element pattern is given by

$$F(\theta, \phi) = (1 - |R(\theta, \phi)|^2) \cos\theta \tag{33}$$

where

$$R(\theta, \phi) = \frac{Z_{\text{in}}(\theta, \phi) - Z_{\text{in}}(0, 0)}{Z_{\text{in}}(\theta, \phi) + Z_{\text{in}}(0, 0)} \tag{34}$$

and $|R(\theta, \phi)|$ is the active reflection coefficient. The term, $Z_{\text{in}}(0, 0)$ is the input impedance when the beam is at broadside and the array is assumed to be matched for maximum array gain.

Depending on the array geometry and its *physical implementation*, for some scanning angles, the active reflection coefficient might be close to unity. In this case, no power is radiated by the array. This phenomenon is generally known as *scan-blindness*. It was initially studied for the infinite array case; however, the same theory is applicable for finite arrays [43, 47]

The following example deals with a 9×9 array of rectangular patches printed on a $0.06\lambda_0$ thick substrate with a relative dielectric constant $\varepsilon_r = 12.8$ [45]. The calculations of the different infinite array parameters are compared to the infinite array case. Figure 54 shows the geometry of the array, and Fig. 55 summarizes the results.

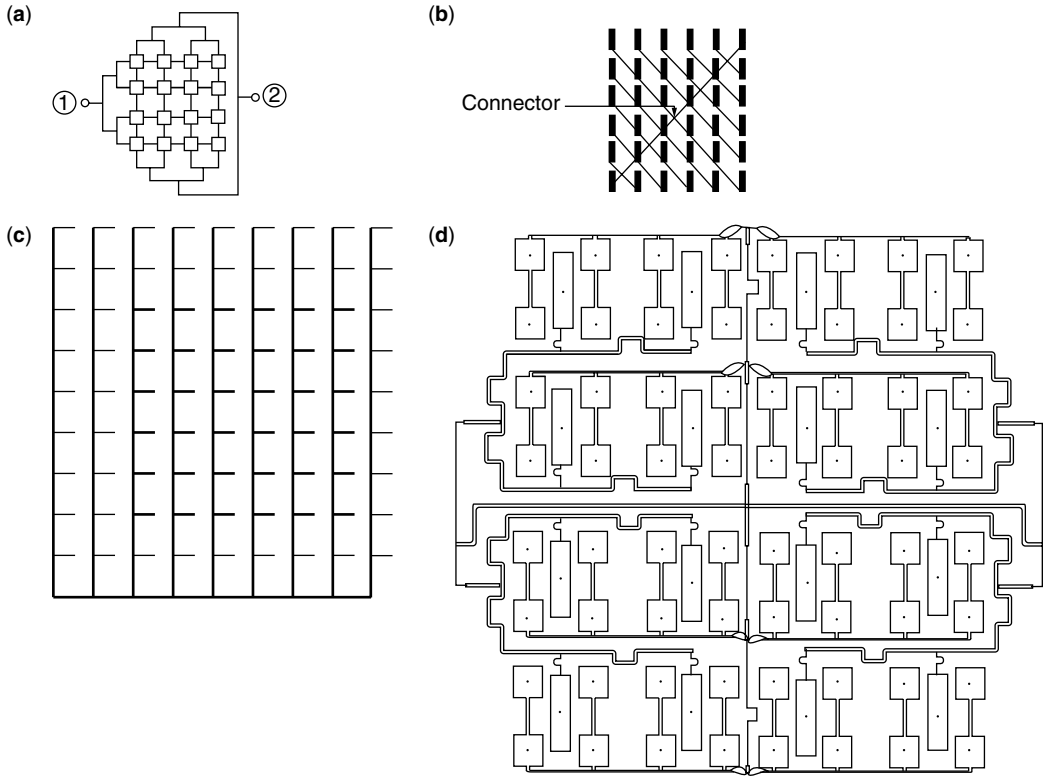


Figure 53. Examples of microstrip planar arrays: (**a**) a dual-polarized 4 × 4-element microstrip array (port 1 is for horizontal polarization; port 2 is for vertical polarization [39]); (**b**) a Cross-fed array [40]; (**c**) J-band planar array of nine linear arrays with nine cophase stubs [41]; and (**d**) a dual-band (2.45/5.7-GHz) planar array [42].

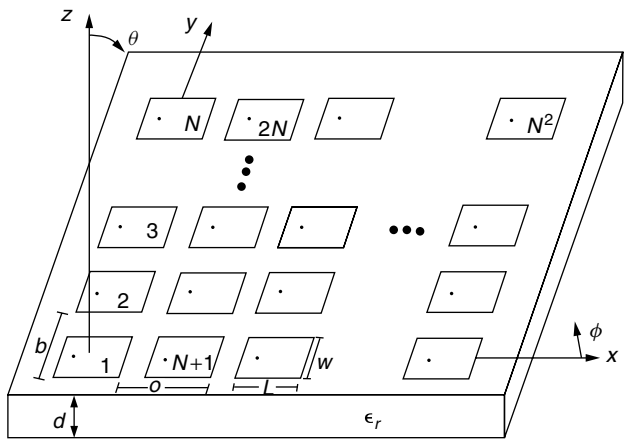


Figure 54. Geometry of the finite array of rectangular microstrip patches [45].

BIOGRAPHY

Naftali Herscovici was born in Bucuresti, Romania in 1954. He received his B.Sc. and M.Sc. from the Technion, Haifa, Israel and his Ph.D. from the University of Massachusetts, Amherst, in 1978, 1985, and 1992 respectively. Between 1982 and 1989 he was employed by Rafael, Haifa, Israel as an Antenna Research Engineer; there he was engaged in research and development of microwave antennas. He is currently the President and Founder of Anteg, Inc., Framingham, Massachusetts. His research interests include microstrip antennas and arrays, reflector antennas and feeds, pattern synthesis, and antenna modeling. Dr. Herscovici is the author of over 50 technical papers in various journal and conference publications.

BIBLIOGRAPHY

1. G. A. Deschamps, Microstrip microwave antennas, *Proc. 3rd USAF Symp. Antennas*, 1953.
2. H. Gutton, and G. Baissinot, Flat aerial for ultra high frequencies, French Patent 70313 (1955).
3. E. V. Byron, A new flush-mounted antenna element for phased array application, *Proc. Phased Array Antenna Symp.*, 1970, pp. 187–192.
4. R. Munson, Conformal microstrip antennas and microstrip phased arrays, *IEEE Trans. Antennas Propag.* **AP-22**: 74–78 (1974).
5. E. H. Newman, and P. Tulyathan, Microstrip analysis technique, *Proc. Workshop Printed Circuit Antennas*, New Mexico State Univ., Oct. 1979, pp. 9.1–9.8.
6. I. J. Bahl, Build microstrip antennas with paper-thin dimensions, *Microwaves* **18**: 50–63 (Oct. 1979).
7. K. R. Carver, Practical analytical techniques for the microstrip antenna, *Proc. Workshop Printed Circuit Antennas*, New Mexico State Univ., Oct. 1979, pp. 7.1–7.20.

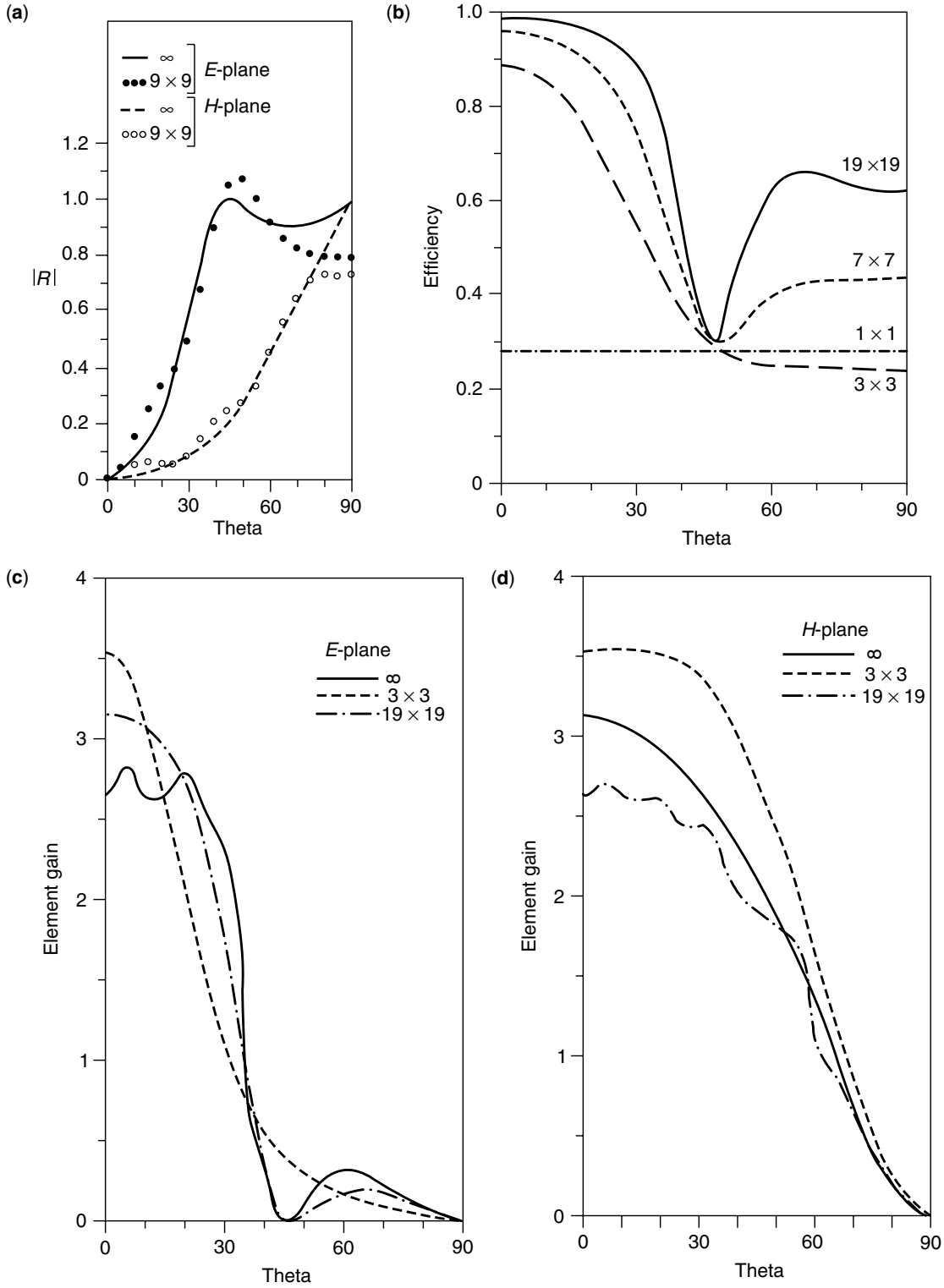


Figure 55. Calculated results for finite patch arrays on $0.06\lambda_0$ thick substrate with a relative dielectric constant $\varepsilon_r = 12.8$ ($a = b = 0.5\lambda_0$, $L = 0.1074\lambda_0$, $W = 0.15\lambda_0$, $X_p = -L/2$, $Y_p = 0$) [45]: (**a**) Reflection coefficient magnitude versus scan angle (E and H planes) for a finite (9×9, center element) patch array, compared with infinite array results; (**b**) efficiency of a finite patch array versus E-plane scan angle for various sizes; (**c**) E-plane active-center-element gains for patch arrays of various sizes; (**d**) H-plane active-center-element gains for patch arrays for various sizes, (**e**) E-plane active-element gain patterns for various elements of a 13×13 patch array; and (**f**) H-plane active-element gain patterns for various elements of a 13×13 patch array.

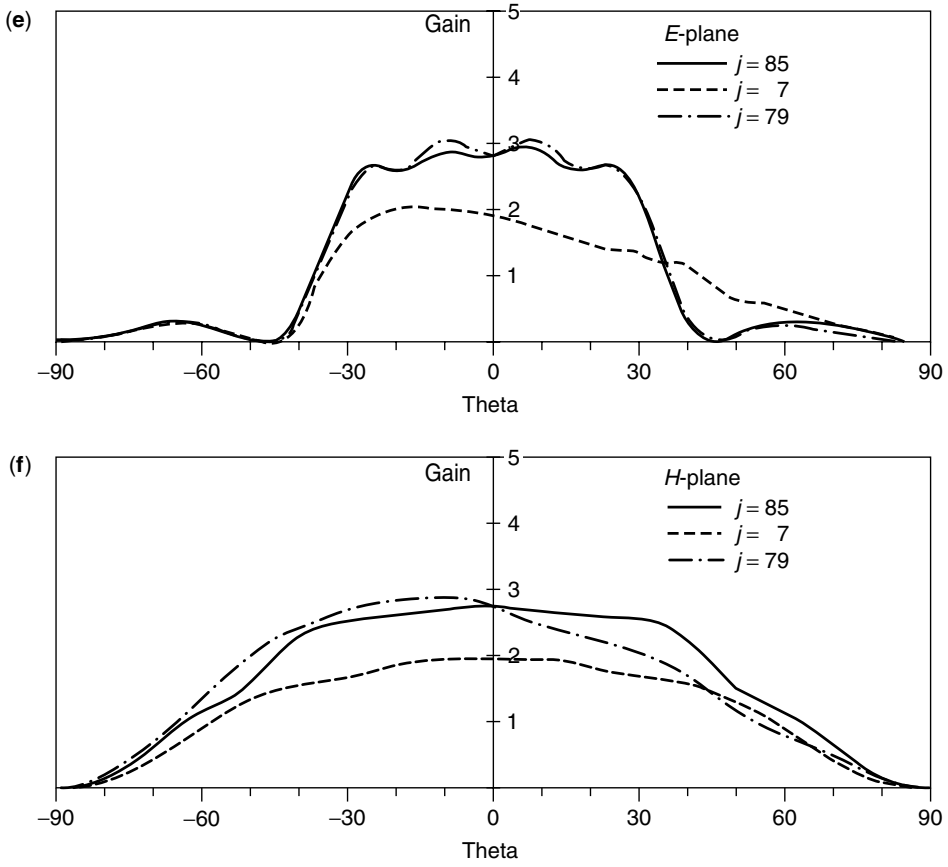


Figure 55. (*Continued*)

8. I. J. Bahl and P. Bhartia, *Microstrip Antennas*, Artech House, 1980.
9. W. F. Richard, Y. T. Lo, and D. D. Harrison, Theory and experiment on microstrip antennas, *Electron. Lett.* **15**: 42–44 (1979).
10. A. G. Derneryd and A. G. Lind, Extended analysis of rectangular microstrip antennas, *IEEE Trans. Antennas Propag.* **AP-27**: 846–849 (1979).
11. J. R. James and P. S. Hall, *Handbook of Microstrip Antennas*, Peter Peregrinus, London, 1989.
12. R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, *Microstrip Antenna Design Handbook*, Artech House, Norwood, MA, 2001.
13. N. Herscovici, Z. Sipus, and D. Bonefacic, Circularly polarized single-fed wide-band microstrip elements and arrays, *1999 IEEE Int. Antennas Propag. Symp. Dig.* **37**: 280–283 (June 1999).
14. J. Huang, A technique for an array to generate circular polarization with linearly polarized elements, *IEEE Trans. Antennas Propagat.* **34**: 1113–1124 (Sept. 1986).
15. M. Haneishi, S. Yoshida, and N. Goto, A broadband microstrip array composed of single-feed type circularly polarized microstrip antennas, *1982 IEEE Int. Antennas Propag. Symp. Dig.* **20**: 160–163 (May 1982).
16. T. Teshirogi, M. Tanaka, and W. Chujo, Wideband circularly polarized array antenna with sequential rotations and phase shifts of elements, *Proc. Int. Symp. Antennas and Propagation*, Japan, 1985, pp. 117–120.
17. D. M. Pozar, Considerations for millimeter wave printed antennas, *IEEE Trans. Antennas Propag.* **31**: 740–747 (Sept. 1983).
18. N. Herscovici, A wide-band single-layer patch antenna, *IEEE Trans. Antennas Propag.* **46**: 471–474 (April 1998).
19. R. Zetner, J. Bartolic, and E. Zetner, Electromagnetically coupled butterfly patch antenna, *J. Int. Nice Antennes* 588–591 (Nov. 1996).
20. H. F. Pues and A. R. Van de Capelle, An impedance-matching technique for increasing the bandwidth of microstrip antennas, *IEEE Trans. Antennas Propag.* **37**: 1345–1354 (Nov. 1989).
21. S. M. Duffy, An enhanced bandwidth design technique for electromagnetically coupled microstrip antennas, *IEEE Trans. Antennas Propag.* **48**: 161–164 (Feb. 2000).
22. R. Kastner, E. Heyman, and A. Sabban, Spectral domain iterative analysis of single- and double-layered microstrip antennas using the conjugate gradient algorithm, *IEEE Trans. Antennas Propag.* **36**: 1204–1212 (Sept. 1988).
23. F. Croq and D. M. Pozar, Millimeter-wave design of wide-band aperture-coupled stacked microstrip antennas, *IEEE Trans. Antennas Propag.* **39**: 1770–1776 (Dec. 1991).
24. A. Sabban, A new broadband stacked two-layer microstrip antenna, *1983 IEEE Int. Antennas Propag. Symp. Dig.* **21**: 63–66 (May 1983).
25. G. Kumar and K. C. Gupta, Broadband microstrip antennas using coupled resonators, *1983 IEEE Int. Antennas Propag. Symp. Dig.* **21**: 67–70 (May 1983).

26. G. Kumar and K. C. Gupta, Nonradiating edges and four edges gap-coupled multiple resonator broad-band microstrip antennas, *IEEE Trans. Antennas Propag.* **33**: 173–178 (Feb. 1985).

27. G. Kumar and K. C. Gupta, Directly coupled multiple resonator wide-band microstrip antennas, *IEEE Trans. Antennas Propag.* **33**: 588–593 (June 1985).

28. F. Croq and D. M. Pozar, Multifrequency operation of microstrip antennas using aperture-coupled parallel resonators, *IEEE Trans. Antennas Propag.* **40**: 1367–1374 (Nov. 1992).

29. S. D. Targonski, R. B. Waterhouse, and D. M. Pozar, An aperture coupled stacked patch antenna with 50% bandwidth, *IEEE Antennas Propagation Symp. Dig.*, Baltimore, MD, July 1996, 18–22.

30. S. D. Targonski, R. B. Waterhouse, and D. M. Pozar, Design of wide-band aperture-stacked patch microstrip antennas, *IEEE Trans. Antennas Propag.* **46**: 1245–1251 (Sept. 1998).

31. S. D. Targonski and R. B. Waterhouse, Reflector elements for aperture and aperture coupled microstrip antennas, *1997 IEEE Int. Antennas Propag. Symp. Dig.* **35**: 1840–1843 (June 1997).

32. D. M. Pozar, Input impedance and mutual coupling of rectangular microstrip antennas, *IEEE Trans. Antennas Propag.* **30**: 1191–1196 (Nov. 1982).

33. R. P. Jedlicka, M. T. Poe, and K. R. Carver, Measured mutual coupling between microstrip antennas, *IEEE Trans. Antennas Propag.* **29**: 147–149 (Jan. 1981).

34. B. B. Jones, F. Y. M. Chow, and A. W. Seeto, The synthesis of shaped patterns with series-fed microstrip patch arrays, *IEEE Trans. Antennas Propag.* **30**: 1206–1212 (Nov. 1982).

35. D. M. Pozar and D. H. Schaubert, Comparison of three series fed microstrip array geometries, *1993 IEEE Int. Antennas Propag. Symp. Dig.* **31**: 728–731 (June 1993).

36. D. M. Pozar, private communication.

37. P. S. Hall and C. M. Hall, Coplanar corporate feed effects in microstrip patch array design, *IEE Proc. Part H*, vol. 135, June 1988 pp. 180–186.

38. D. M. Pozar, *Workshop of Antennas for Wireless Communications*, Nov. 1998.

39. A. G. Derneryd, Microstrip array antenna, *Proc. 6th European Microwave Conf.*, 1976, 339–343.

40. J. C. Williams, Cross fed printed aerials, *Proc. 7th European Microwave Conf.*, 1977, 292–296.

41. J. R. James and P. S. Hall, Microstrip antennas and arrays, Part 2—New array design technique, *IEEE J. Microwaves Opt. Acoust.* **1**: 175–181 (1977).

42. N. Herscovici, New considerations in the design of microstrip antennas, *IEEE Trans. Antennas Propag.* **46**: 807–812 (June 1998).

43. R. C. Hansen, *Microwave Scanning Arrays,* Vol. 2, Academic Press, 1966.

44. R. J. Mailloux, *Phased Array Antenna Handbook*, Artech house, 1994.

45. D. M. Pozar, Finite phased arrays of rectangular microstrip patches, *IEEE Trans. Antennas Propag.* **34**: 658–665 (May 1986).

46. D. M. Pozar and D. H. Schaubert, Analysis of an infinite array of rectangular microstrip patches with idealized probe feeds, *IEEE Trans. Antennas Propag.* **32**: 1101–1107 (Oct. 1984).

47. D. M. Pozar and D. H. Schaubert, Scan blindness in infinite phased arrays of printed dipoles, *IEEE Trans. Antennas Propag.* **32**: 602–610 (June 1984).

MICROSTRIP PATCH ARRAYS

R. B. WATERHOUSE
K. GHORBANI
RMIT University
Melbourne, Australia

1. INTRODUCTION

Microstrip patch antennas have long been touted as one of the most versatile radiating structures. These printed radiating elements have several well-known advantages over conventionally styled antennas based on wires and metallic apertures including their low profile, low cost, robustness, and ease of integration with other components. Since 1980 or so this once considered problematic antenna has matured into one of the most commonly used interfaces for free-space/wireless communications. Most mobile communication base stations and handset terminals as well as spaceborne communication systems incorporate this form of radiator.

As a single radiating element, the microstrip patch antenna is generally classified as a low–moderate-gain antenna with gains in the order of 5–8 dBi in its conventional form. One critical advantage of the microstrip patch over its counterparts, which is related to some of the features mentioned before, is the relative ease in which these structures can be integrated or combined to form an array of antennas. By doing so greatly increases the flexibility in shaping the radiation pattern and other features of the antenna, which is consistent with arraying wire, metallic and other forms of radiators. However the distinct advantages of arraying microstrip elements are the ease in fabricating the entire structure, the simplicity of the array layout as well as the low cost of production. The fabrication of these antennas is based on printed circuit board (PCB) etching processes that has minimal labor costs.

In this article we review the development of arrays based on microstrip patch technology. Firstly we discuss the fundamental styles of linear arrays that can be developed using microstrip patches, namely, series feed, corporate feed and a combination feed technique. For each of these methods advantages, issues and design cases are given. The scanning performance (or radiation control) of a linear array is discussed and a design case is once again given. The concepts introduced for linear arrays are then expanded on to investigate planar arrays and methods on how these radiating structures can be developed are presented. Some printed antenna alternatives are summarized that can yield high-gain solutions, with minimal complexity. These printed antennas can overcome the feed loss problems associated with very large planar

arrays. Finally the scanning performance of large planar arrays of microstrip patch antennas is examined and once again the parameters affecting the control of the radiation distribution as well as the limiting performance are discussed.

2. LINEAR ARRAYS

Examining linear arrays of any antenna is probably the easiest means to see how the radiation performance can be controlled in a particular direction or dimension. Of course, the concepts developed or derived from a linear array can be readily expanded to a planar, or a two-dimensional solution. There are numerous books and articles on array theory and the reader should consult these to understand the fundamental properties of arrays [e.g., 1]. There are three types of microstrip patch linear arrays: the series-fed configuration, the corporate (or parallel)-fed geometry, and the combination technique. These methods are examined herein.

2.1. Series-Fed Arrays

One of the first realizations of a microstrip patch array was the series-fed array [e.g., 2]. Here each element of the array is connected in series via an arrangement of transmission lines. Figure 1 shows a schematic diagram of an 8-element series-fed array consisting of edge-fed patches. The array is fed from the left and is classified as a standing-wave array. Series-fed arrays have been developed in waveguide realizations for decades; however, microstrip forms have much more flexibility. This is due mainly to the fact that it is easy to change the impedance of the microstrip feedlines between the radiating elements to give the desired amplitude taper [3]. The patch width can also be varied to give this same effect.

The advantages of microstrip patch arrays utilizing a series feed configuration over other forms (to be discussed later) include it having a simple, more compact feed network, as evident from Fig. 1, and lower feedline loss. However, this form of microstrip patch array does suffer from several drawbacks. The most fundamental issue is the narrow radiation bandwidth of the array, which is typically much narrower than the inherent impedance bandwidth of the individual microstrip elements. There are only several reported cases of series-fed microstrip arrays in the literature, and these have bandwidths typically only fractions of a percent. As microstrip patches in their original form have a high Q value, placing them in series means that each will have a direct impact on the other, and therefore if there are any errors in fabrication or factors not taken into consideration with the design (such as mutual coupling), the overall array performance will be degraded. Because the power to be supplied to each element must by transferred from the previous element (see Fig. 1), the rapid impedance variation of the conventional microstrip patch inherently hinders the delivery of the power to the other elements. Although there have been several techniques over the years to increase the impedance bandwidth of individual microstrip patches, such as a proximity coupled or aperture-coupled patch, incorporating a series feed array solution of these radiators removes the open- or short-circuited tuning stub, reducing the number of degrees of freedom of these non-contact-excitation methods and hence their flexibility.

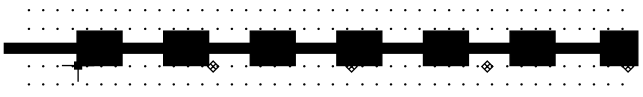


Figure 1. Schematic diagram of 8-element series-fed linear array of microstrip patches.

Figure 2 shows a 4-element edge-fed series microstrip array. The parameters for the four elements are shown in Fig. 2. The length of elements and distance between them were varied to achieve maximum gain near broadside using a full-wave simulator. The feedlines between the elements are 100-Ω transmission lines, so the disturbance in the field of radiators can be minimized. The overall antenna is matched to 50 Ω by a quarter-wave transformer. The widths of antenna elements were varied to reduce the sidelobe levels. The return loss response and the E-plane radiation performance of the array across the 10-dB return loss bandwidth of 1% is shown in Fig. 3a,b, respectively. As can be seen from Fig. 3b, the radiation pattern remains generally constant across the matched impedance bandwidth, unlike some of the early cases of series-fed microstrip patch arrays. There is a slight asymmetry in the radiation pattern that is due to the presence of the feed network. The H-plane pattern is similar to a conventional edge-fed microstrip patch [e.g., 1]. The cross-polarization level in both principal planes (E and H planes) was less than 40 dB below the copolar fields. The relatively constant radiation performance of the antenna can be attributed to using a full-wave simulator to synthesize the antenna, software tools that were not available in the early days of microstrip patch technology development or were simply too slow for

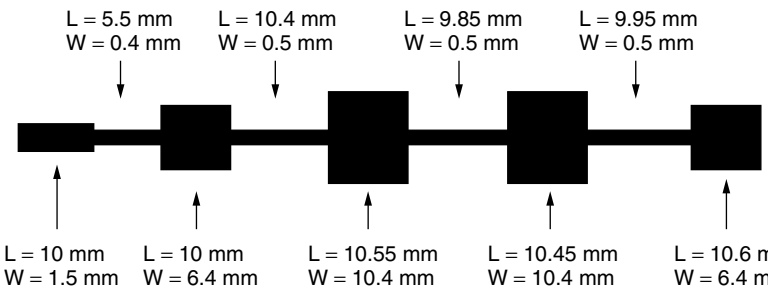


Figure 2. Schematic diagram of 4-element series-fed linear array of microstrip patches.

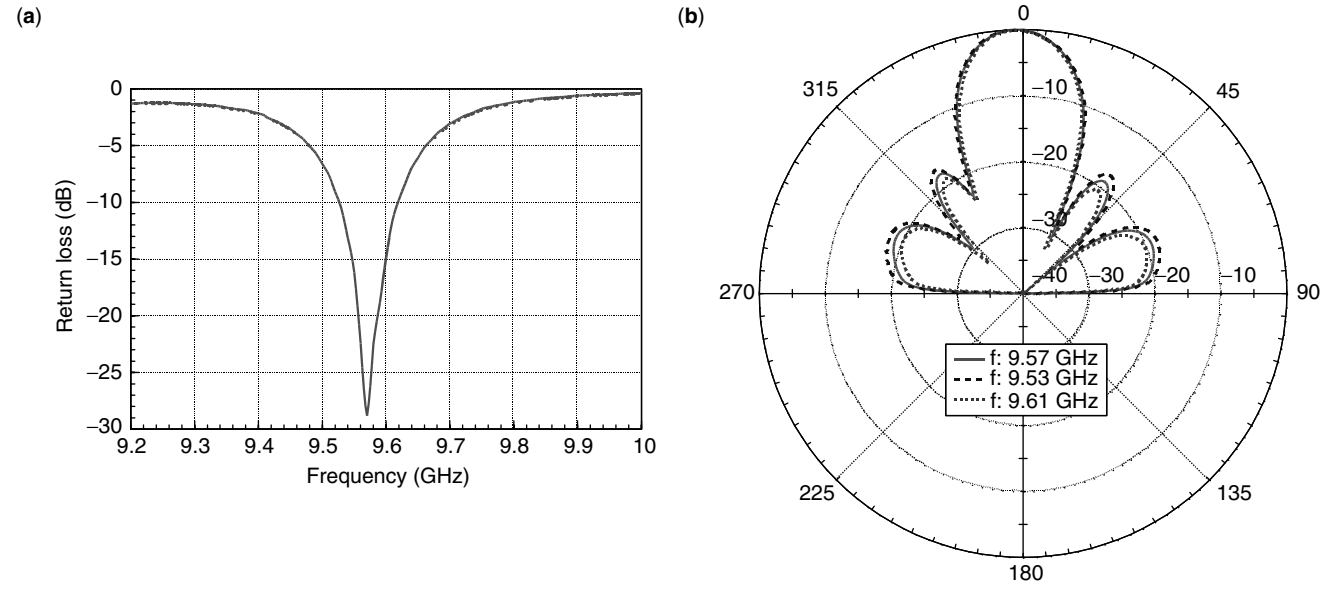


Figure 3. Characteristics of 4-element series fed array: (**a**) return loss; (**b**) radiation patterns.

design purposes. Utilizing such tools as well as enhanced bandwidth elements (such as stacked patches) should see an improvement in the performance of series-fed arrays, but probably not to the same degree as corporate (or parallel) fed microstrip arrays.

2.2. Parallel Fed Arrays

Parallel or corporate fed microstrip patch arrays are the most common type of array using microstrip patch technology. Here, unlike the series-fed array, each element has its own excitation transmission line, which can be made independent of the feedlines of the other elements as well as the other elements of the array. Figure 4 shows a schematic diagram of an 8-element corporate feed array of edge-fed microstrip patches. As can be seen from the figure, each element has its own excitation transmission line. Each of these transmission lines is then connected together via a series of two-way power combiners, although three-way dividers are commonly used if an odd number of elements are used in the array. The power combiners can either be reactive, such as shown in Fig. 4, or based on Wilkinson dividers. The Wilkinson divider gives broader band isolation between the elements at the expense of increased complexity and also loss. It should be noted that most microstrip patches have impedance bandwidths smaller than that of a reactive power divider.

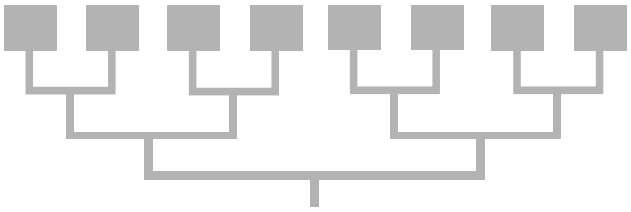


Figure 4. Schematic diagram of an 8-element corporate fed linear array of microstrip patches.

Of all the array formats, parallel configurations have the broadest bandwidths, in some cases even greater than that of the individual elements of the array. This effect can be attributed to the cancellation of unwanted reflections of power within the feed network. The good isolation between the individual feedlines allows the ready incorporation of phase shifters to allow scanning of the radiation beam of the array (referred to later in this section) as well as amplitude tapers to reduce the sidelobe level. An excellent paper outlining how to do this and the possible source of error is available [4]. The good isolation of the parallel feed allows the designer to separately address the issues related to the individual microstrip patch (the basis of the array) and then the feed network. Such an approach significantly reduces the computational power required to successfully design the array. Because of all these features, corporate fed microstrip patch arrays are utilized in many applications such as mobile base-station antennas.

Figure 5 shows a schematic diagram of an 8-element corporate array of aperture-coupled microstrip patches. For details on the design of aperture-coupled patches

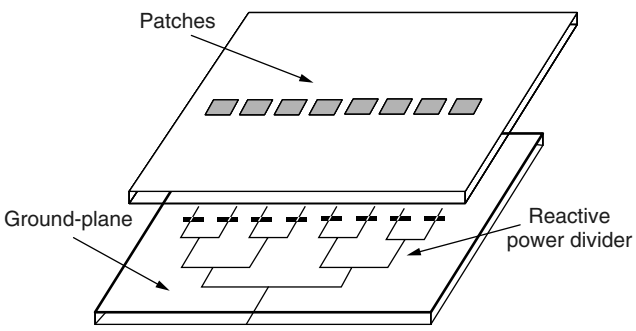


Figure 5. Schematic diagram of 8-element corporate fed linear array of aperture coupled microstrip patches.

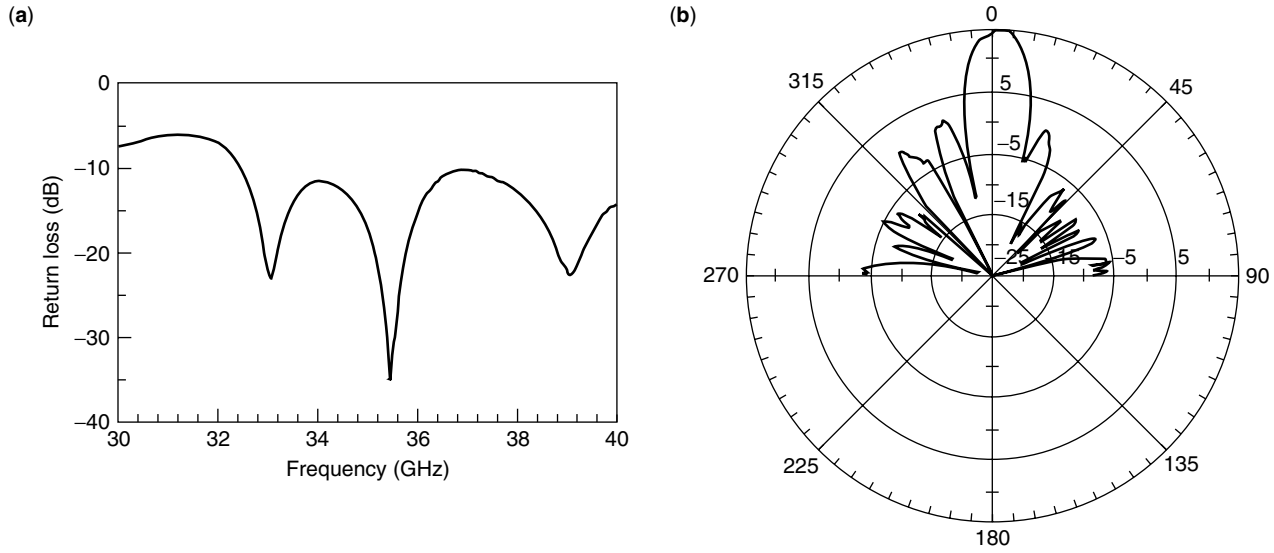


Figure 6. Characteristics of 8 element corporate fed array of aperture coupled patches: (**a**) measured return loss; (**b**) *H*-plane radiation pattern.

please refer to the article by Targonski and Pozar [5]. The array utilizes seven reactive power dividers to feed the elements. As mentioned before, reactive power dividers are more efficient than their counterparts. The array was designed for fiber radio applications at millimeter-wave frequencies [6]. The measured return loss of the array is given in Fig. 6a, and the 10-dB return loss bandwidth is approximately 30%. A sample of the *H*-plane radiation pattern is shown in Fig. 6b. The *E*-plane pattern, not shown here, is similar to a conventional microstrip patch element. The array has a gain of 15 dBi across the entire 10-dB return loss bandwidth, and the cross-polarization levels were less than 20 dB below the copolar fields in both principal planes.

2.3. Combination Feed Arrays

There is a third class of microstrip array, which is a combination of the series and parallel feed methods. Here a common feedline is used for the entire array and each element taps power off this feedline. To ensure that the bandwidth is greater than the conventional series-fed array, each tap and section of the common feedline is impedance matched. Thus the whole design of the array simplifies itself to uncomplicated impedance matching to ensure good impedance bandwidth as well as the appropriate distribution of power. Figure 7 shows a schematic diagram of how an 8-element version of this array can be realized. First, the array is split into two symmetrical parts, one of which is shown in Fig. 7. Here each element is designed for an input impedance at resonance of 200 Ω. Doing so allows for a relatively straightforward impedance-matching design to ensure that equal power is distributed to each microstrip patch. The feedlines connected to each element are also made to have a characteristic impedance of 200 Ω and a length of $\lambda_g/2$, where λ_g is the guided wavelength in microstrip. Doing this minimizes the effects of error in the design of the microstrip patch, which will further impact the performance of the array. It is also difficult to fabricate 200-Ω transmission lines on some material, so the $\lambda_g/2$ length transforms the input impedance of the patch to the central feedline (refer to Fig. 7).

An 8-element combination array was designed and developed centered at 9 GHz. A photograph of the array is shown in Fig. 8 (see also Fig. 9). The impedance bandwidth was measured as 5%. An example of the radiation patterns in the *H* and *E* planes of the antenna are shown in Fig. 9a,b, respectively. The *H*-plane pattern shows the expected focusing of the beam in the plane of the array (note that the slight off-broadside pattern is due to alignment errors in the measurement setup). A small

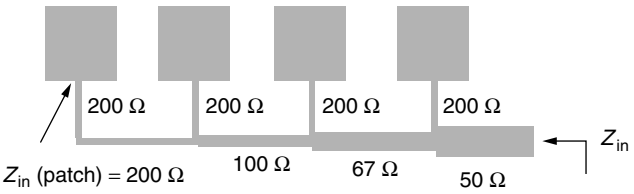


Figure 7. Schematic diagram outlining design of combination feed linear array.

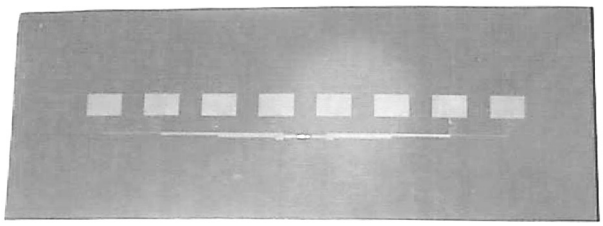


Figure 8. Photograph of 8-element combination feed linear array.

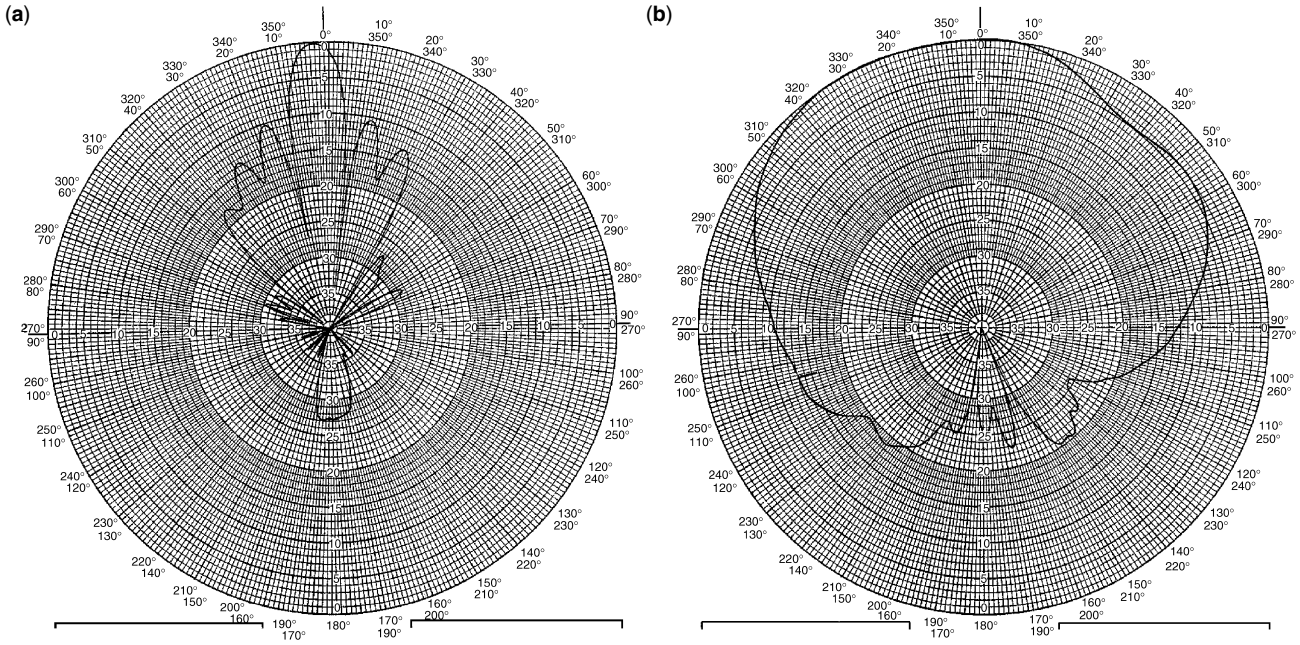


Figure 9. Radiation performance of 8-element combination feed linear array: (**a**) *H* plane; (**b**) *E* plane.

scalping of the pattern is evident in the *E* plane. This is due to the radiation from the feed network. The gain of the array was measured as 15 dBi across the impedance bandwidth. The overall performance of a combination feed array lies somewhere between that of the series-fed array and the parallel array. Its impedance and radiation bandwidth (3 dB gain) is greater than the series array but less than that of the corporate array. Its efficiency is greater than the parallel configuration, although less than the series method.

2.4. Scanned Linear Arrays

The linear array designs considered above are fixed-beam examples, with the main beam directed toward broadside. It is relatively straightforward to point this beam at a fixed angle off broadside by simply inserting a constant phase between the elements. There are many applications. For example, satellite communications, which require a beam that can be scanned or continually steered to ensure contact between a moving object (say, an aircraft) and a stationary or a moving object (say, a constellation of satellites) can be maintained at all times. Because microstrip patches can readily be formed into arrays and easily connected to phase shifting circuitry, these radiators are prime candidates for most phased-array applications. Microstrip elements have broad radiation patterns, which means that arrays of these elements should be able to be scanned to large angles, approaching endfire. This issue will be discussed in Section 3.

Of the three configurations presented, corporate feed arrays are the easiest to control the phasing to its elements. Figure 10 shows a photograph of a linear phased-array based on a corporate feed arrangement mounted on a test jig. The 6.4 GHz 8-element array is located at the top of the photograph, with the radome (a layer of Duroid 5880) that covers the top patches evident in the photograph (the white-grayish rectangular region). The size of the ground plane of the array is 22×10 cm. The array consists of edge-fed stacked patches fed by a 90° branchline coupler to produce circular polarization, a common requirement for satellite communications and

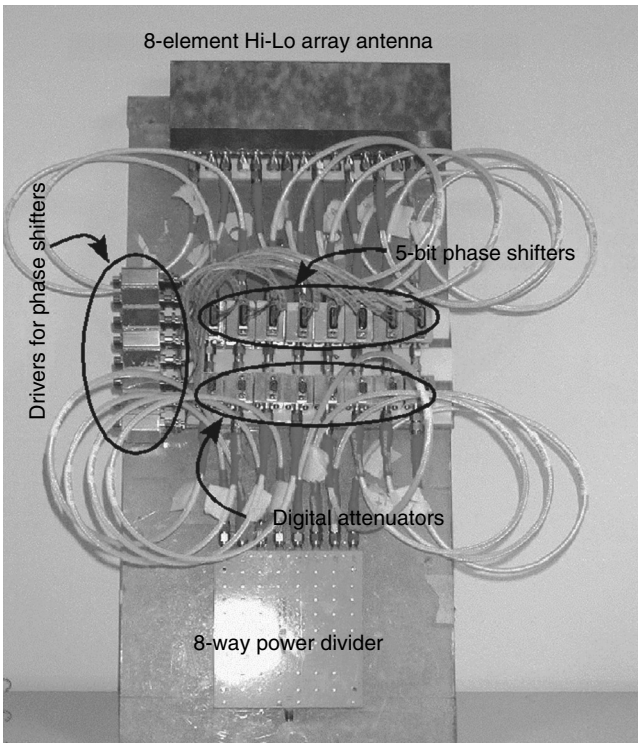


Figure 10. Photograph of linear 8-element phased array of microstrip patches.

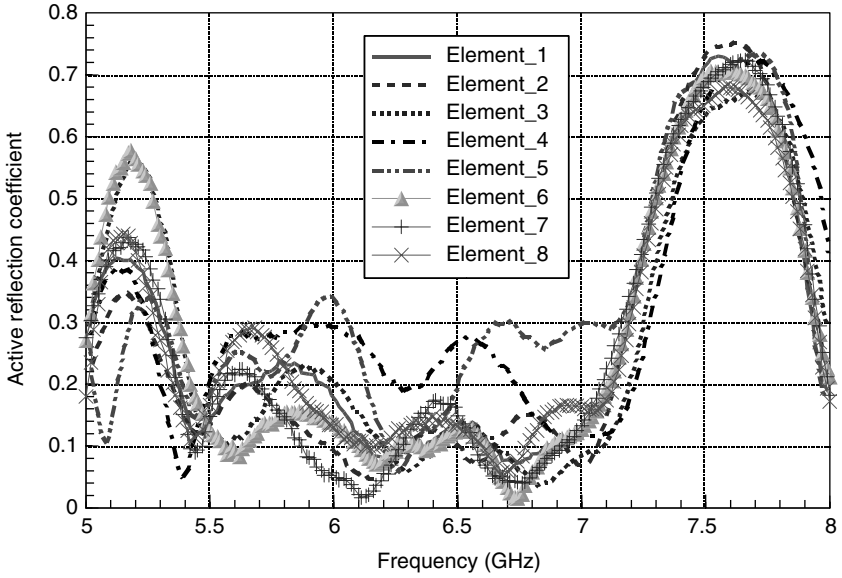


Figure 11. Active reflection coefficient of 8-element linear phased array.

easily achievable using microstrip technology. Feeding the left-hand-side port while terminating the right-hand-side port with matched loads results in right-hand circular polarization (RHCP) generation. Reversing the feed and terminated port results in left-hand circular polarization (LHCP) generation. For the experiments conducted here, only the RHCP was investigated.

Eight phase-matched cables are used to connect the stacked patch array to a bank of 5-bit switched-line phase shifters. As the losses of the phase shifters vary at different phase settings, a bank of 3-bit digital attenuators is included to achieve amplitude balance. In this experiment, the maximum phase error is $\pm 8^\circ$ and the maximum amplitude error is ± 0.7 dB.

Figure 11 shows the measured active reflection coefficient [7] of each element in the array at broadside with the other elements of the array terminated with matched loads. This is a common measurement procedure to ascertain the performance of the array. The worst-case measured 10 dB return loss for each element was 28.6%, centered at 6.3 GHz. The significantly increased impedance bandwidth, compared to the predicted individual stacked patch case of 20% can be attributed to the feed network. Such feed-networks typically cancel unwanted reflections as mentioned previously. It is interesting to note that most of the active reflection coefficients for the 8 elements have similar responses between 5.3 and 7.5 GHz, with the exception of elements 4 and 5. Although these elements still satisfy the 10-dB return loss criteria over the same bandwidth as the other elements, their active reflection coefficients are marginally higher. This may be due to a soldering problem where the microstrip lines are connected to the appropriate SMA-style connectors.

The radiation patterns and axial ratio of the scannable printed array were measured at a variety of frequencies and scan angles. A sample of the results is presented in Fig. 12. The array can be readily scanned to $\pm 45^\circ$ while maintaining an axial ratio of less than 3 dB. The gain across the scanned range of angles and frequencies is approximately 3 dB lower than the expected or predicted values due to the insertion loss of the phasing module.

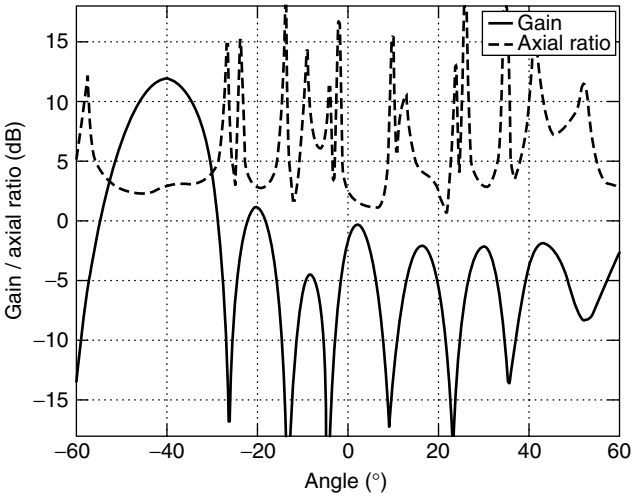


Figure 12. Sample of radiation performance of linear phased array.

3. PLANAR ARRAYS

Having outlined the fundamental structures for linear arrays in the previous section, it is relatively straightforward to extend these configurations into planar, or two-dimensional arrays. Thus you can effectively have a planar series-fed, corporate fed or a combination feed array. However, it is probably easier to categorize planar arrays in two styles: fixed-beam or scanned. Fixed-beam can consist of any of the three linear arrays introduced. Scanned arrays tend to always consist of corporate style feeding, simply because it is the easiest to achieve independent phase (and amplitude) control of the elements of the array. In this section we examine fixed and scanned beam planar arrays.

3.1. Fixed-Beam Planar Arrays

Looking through the literature, to the authors' knowledge there does not appear to be any case of planar series-fed arrays. This is intuitive, for as the length of the array increases, or in the planar case, the area, the elements at the end of the array are less likely to receive any power from the source. Thus these elements in a series feed array are redundant. It is perhaps possible to develop small (say, 8-element) two-dimensional arrays of series-fed microstrip patch elements; however, the shortcomings highlighted previously for the series-fed linear array would still hold for this case. There is reported in the literature a case of several linear series arrays connected in parallel, using the impedance-matching procedure summarized earlier [8].

It is possible to create a planar version of the combination feed array. Figure 13 shows a photograph of a 32-element array. The microstrip array consists of combining four of the 8-element linear arrays considered in the previous section. Once again, to combine these arrays requires the use of impedance matching and quarter-wave transformers. The radiation patterns of the array in both the E and H planes are shown in Fig. 14. The focusing of the radiation toward broadside is evident in this figure. The gain of this array was measured as 21 dBi and the impedance bandwidth as 5%. In Fig. 13 the $\lambda_g/2$ lines that feed the elements of each linear array have been folded on themselves to ensure that the array spacing in the E plane is not too large. The array spacing in each plane is $0.8\ \lambda_0$ to ensure maximum directivity [1].

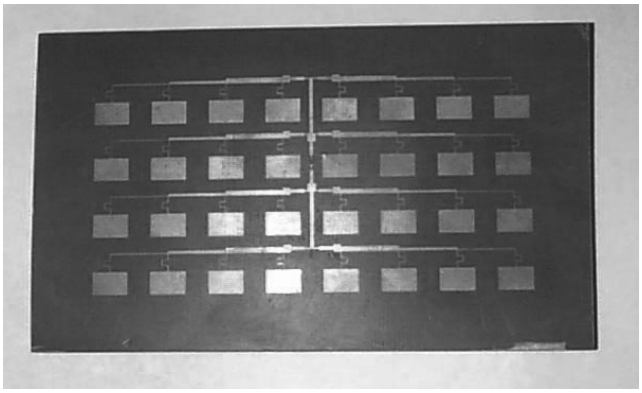


Figure 13. Photograph of 32-element planar combination array.

By far the most common type of fixed-beam microstrip patch array is based on the corporate feed [9]. These planar arrays are utilized in applications such as millimeter wave collision avoidance radar for vehicles, local multipoint distribution services and imaging. A schematic diagram of a 256-element corporate fed patch array is shown in Fig. 15 [10]. The design of these arrays can be somewhat complicated, not so much in terms of the antenna element design, but because of the feed network layout. A good rule of thumb to minimize spurious radiation from feedlines is to keep the structure as symmetric as possible, which tends to minimize cross-polarization levels and to use thin transmission lines. Levine et al. have contributed an excellent paper on the effect of the feed network on the overall performance of a corporate fed microstrip patch array [10]. In this paper, it was shown that as the array gets larger, the loss associated with the feed network gets more and more until it can be substantial. For a 32×32-element array, the loss was more than 7.5 dB. Table 1 shows a comparison of planar arrays of microstrip patches versus reflectors with efficiencies of 50% [10]. The table highlights the issues related to large patch arrays. We can see from this table that although the directivity increases as the number of elements increase

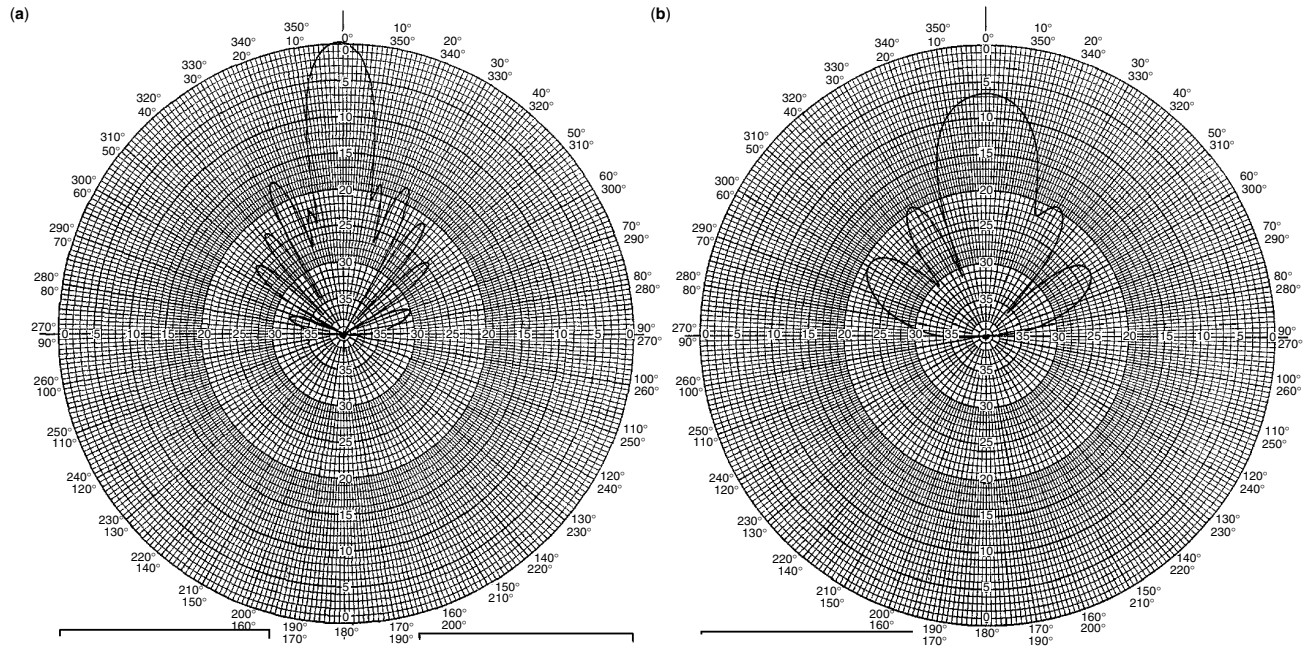


Figure 14. Radiation performance of 32-element combination array: (**a**) H plane; (**b**) E plane.

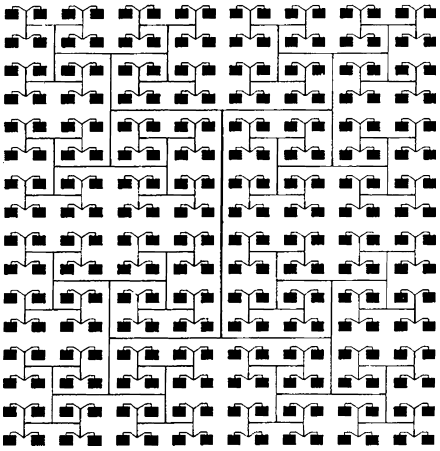


Figure 15. Schematic diagram of planar corporate fed array.

and microstrip technology can yield gains similar to those of a reflector for array sizes less than about 1000 elements, the feed-related losses (radiation, dielectric and ohmic) become significant.

There are printed alternatives to large arrays of patches to produce high-gain antennas for point-to-point applications. These include lens coupled printed antennas [e.g., 11] and reflectarrays [12]. Lens coupled microstrip patches remove the feed-associated losses as there is only one radiating element. These antennas can yield gains in excess of 30 dBi and importantly bandwidths (both radiation and impedance) as broad as the feed element [13]. Figure 16 shows a photograph of an aperture stacked patch lens coupled antenna, with a bandwidth that covers the entire Ka band (26–40 GHz). Printed reflectarrays are another promising alternative to large arrays of microstrip patches. These antennas can yield gains greater than 50 dBi, although the bandwidths are typically small, to date a couple of percent. Figure 17 shows a photograph of a millimeter wave reflectarray [14]. Of course, these printed antennas have many of the features of microstrip patch arrays; however, the conformal nature of the entire antenna is no longer a feature.

3.2. Scanning/Phased Planar Arrays

As mentioned previously, microstrip patch antennas can readily be integrated with active microwave devices. This is one of the key reasons as to why microstrip patch antennas are so advantageous when considering a scanning array, in particular a planar phased array. A planar phased array allows for pattern control in both dimensions and in doing so provides a very flexible or smart antenna.

Table 1. Planar Arrays of Microstrip Patches Versus Reflectors

Number of Elements	16	64	256	1024	4096
Directivity without network	20.9	27.0	33.0	39.2	45.1
Radiation loss	0.8	1.0	1.3	1.9	2.6
Surface wave loss	0.3	0.3	0.2	0.2	0.1
Dielectric loss	0.1	0.3	0.5	1.0	2.1
Ohmic loss	0.1	0.3	0.6	1.2	2.4
Calculated gain	19.5	25	30	34.5	37.5
Gain of reflector	18	24	30	36	42

Figure 16. Photograph of millimeter wave lens-coupled proximity-coupled microstrip patch antenna.

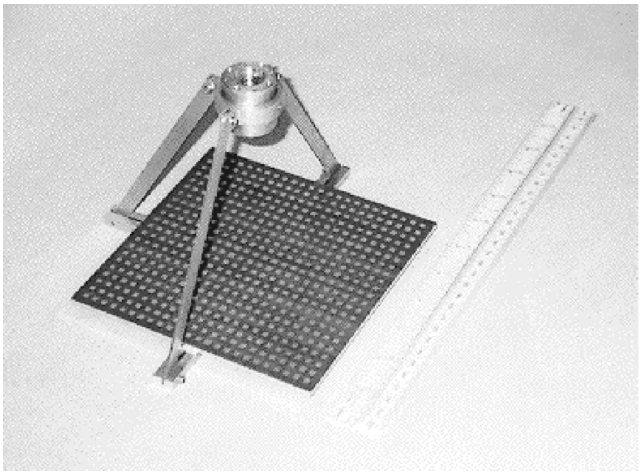


Figure 17. Photograph of millimeter wave reflectarray antenna.

There is one issue related to microstrip patch antenna technology that wasn't mentioned before in this article: surface wave excitation. Surface waves (sometime referred to as "leaky" waves) are "trapped" waves excited by the presence of the substrate or dielectric layers associated with the microstrip antenna. Because the energy is generally trapped within the material and not radiated,

surface waves are classified as a loss mechanism. The presence of a surface wave can cause increases in cross-polarization levels due to the trapped wave refracting off the finite edges of the ground plane of the antenna. Surface waves can also cause unwanted coupling between the antenna and any active devices.

For a single-layer microstrip patch antenna, the thicker the material used, the larger the power lost to the surface wave. Also the higher the dielectric constant, the less efficient the antenna becomes as a result of surface wave excitation. For large arrays of microstrip patches, the resonance of modes associated with these surface waves can severely limit the scan performance of the array by inducing a phenomenon known as a *scan blindness*. For a scan blindness, all (or at least most) of the power is coupled back to the source and subsequently is not radiated. A common means of examining the scanning potential of a large array of microstrip patches is to consider the theoretical active reflection coefficient of the array, which is defined as the reflection coefficient of an element in the array as a function of scan angle [15]. Figure 18 shows the active reflection coefficient of a large array of probe-fed patches when scanning in the E, H and D planes. As can be seen here, in the E plane, the active reflection becomes larger as the scan angle increases as a result of mutual coupling until a point where it levels off and then increases to unity at endfire. The scan angle where it is becomes large (approximately 75°) is the scan blindness. Note that the degree of blindness depends on what element is used. For example, if aperture-coupled patches were used in this array, the active reflection coefficient at the scan blindness would approach one. The magnitude at the scan blindness is dependent on the level of spurious radiation from the antenna.

The scan position of the blindness is very dependent on the element spacing of the array. Figure 19 shows the active reflection coefficient for an array of microstrip patches at three frequencies, at the lower edge of the 10-dB return loss bandwidth, the center frequency, and the upper frequency of the 10-dB return loss bandwidth. The element spacing of the array was 0.5 λ_0 at the center frequency. As can be seen from Fig. 19, the array has very limited scanning ability at the higher-frequency edge because of the impedance mismatch associated with the surface wave.

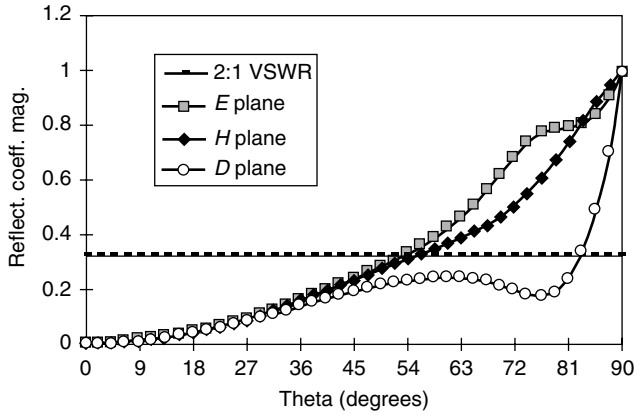


Figure 18. Scan active reflection coefficient of infinite array of probe-fed microstrip patches.

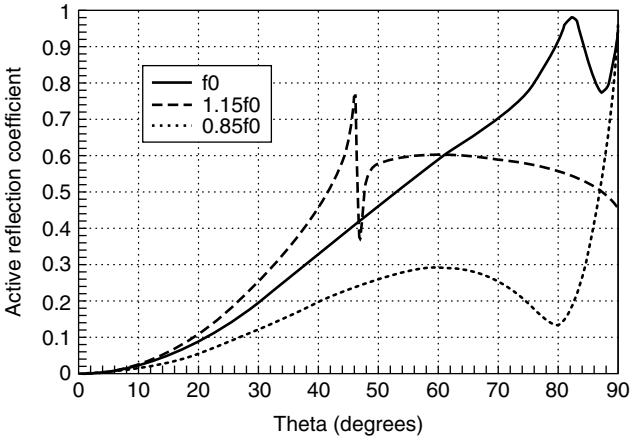


Figure 19. Frequency dependence of scan active reflection coefficient of infinite array of microstrip patches.

The sudden drop in reflection coefficient after the scan blindness is due to the presence of a grating lobe. Although the active reflection coefficient looks reasonable after this scan angle, the radiation efficiency of the array is low, due to power being dumped into the grating lobe [16,17]. Thus it would appear that microstrip patches would have very limited use for large scanning arrays because of the excitation of surface waves and the fact that to increase the bandwidth of a conventional patch the material thickness must be increased and therefore the surface wave content would also increase. However fortunately there are ways to alleviate this problem.

The previously mentioned scanning and/or material trends apply only to single-layer geometries and do not hold for more complicated patch configurations consisting of multiple layers. For example, an aperture stacked patch [18] can have a surface wave efficiency greater than 85% even when the overall thickness of the materials used is greater than $0.1\lambda_0$ (which is very thick for microstrip patches). Such a printed antenna can have an impedance bandwidth of over an octave. Also a stacked patch using high-dielectric-constant material for the lower layer can have an efficiency greater than 90%, even though a single-layer patch using the same high-dielectric-constant material has a surface wave efficiency of only <65% [19]. The 10-dB return loss bandwidth for this antenna can be greater than 30%. Both of these patch elements can also be used in large scanning arrays. It has since been was shown that a large array of aperture stacked patches can have a 10-dB return loss bandwidth in excess of an octave while being able to be scanned to angles greater than ±45° in the principal planes [20]. To design such arrays requires careful consideration of the impedance response of the array and its *spiders* (how the impedance changes as a function of scan angle [20]). It is imperative to try to minimize the impedance variation (as a function of frequency and scan angle) as much as possible for the array to ensure optimum performance.

Other techniques have been developed over the years that can improve the scanning performance of the conventional microstrip patch phased array, albeit at the expense of complexity. These include using cavity-backed

structures [21] and shorting pins [22]. These methods could be applied to the broadband solutions of Refs. 20 and 23 to give perhaps the ultimate microstrip patch phased arrays.

4. CONCLUSIONS

In this article, an overview of microstrip patch array technology has been presented. Various forms of linear arrays were discussed. Case studies were given and a comparison of the advantages and issues associated with each type of array were presented. Corporate fed arrays are probably the most versatile with the largest bandwidth, although these arrays suffer from higher feed loss than do series and combination arrays. Combination arrays can provide a relatively simple design procedure and also good radiation and bandwidth results. A linear phased array was also presented, and its scanning performance is summarized.

Planar fixed-beam and scanned arrays utilizing microstrip patches were also investigated. Once again, corporate feeding is probably the easiest means of forming a planar array, especially if scanning the beam is required. Surface waves associated with the dielectric materials can have detrimental effects on the scanning performance of a large array of patches, although several methods have been established to overcome this inherent problem. Also, the scan/materials trends for single-layer geometries do not necessarily hold for more complicated, broader bandwidth printed structures, which is very fortuitous. Finally, high-gain printed antenna alternatives to large arrays was briefly examined. These antennas are very suited to point-to-point applications.

From the arrays and trends presented, it should be apparent that microstrip patches will continue to be one of the preferred options when choosing an antenna for a communication system.

Acknowledgments

The authors would like to thank the following people for the valuable discussions and input into the design and realization of some of the arrays presented in this article: Dr. D. Chui, Dr. A. Hoorfar, Dr. A. Nirmalathas, Dr. D. Novak, Mr. W. Rowe, Dr. S. Targonski, and Mr. D. Welch.

BIOGRAPHIES

Rod Waterhouse (S'90–M'94–SM'01) received the degrees of BE (Hons), MEngSc (Research) and Ph.D. from the University of Queensland, Australia, in 1987, 1990, and 1994, respectively. In 1994 he joined the School of Electrical and Computer Engineering at the RMIT University, Melbourne, Australia. From mid–2000 to the beginning of 2001 he was a visiting professor at the Department of Electrical and Computer Engineering UCLA, California, for three months and then a visiting researcher in the Photonics Technology Branch at the Naval Research Laboratories, Washington D.C., for a another 3 months while on his sabbatical. In June 2001, he took a leave of absence from RMIT and joined Dorsal Networks, Columbia, Maryland. His research interests include printed antennas, optically distributed wireless systems, photonic devices and optical systems. He has published over 140 papers and has three patents in these areas. Dr. Waterhouse chaired the IEEE Victorian MTTS/APS Chapter from 1998–2001.

Kamran Ghorbani was born in Mashad, Iran, in 1966. He received his B.E. degree in communication and electronic engineering (first honor) from RMIT University, Melbourne, Australia, in 1995. He has completed his Ph.D. degree at RMIT in 2001. After working as a RF designer for AWA Defense Industries Adelaide, South Australia, he joined the RF and photonic research group at RMIT University in 1996. From 1999 to 2001 he worked as senior RF designer for a telecommunication company. He rejoined RF Photonic Group at RMIT in 2001, where he is currently a research fellow. His research interests include integrated optics, phased array antenna, and microwave system design.

BIBLIOGRAPHY

1. C. A. Balanis, *Antenna Theory: Analysis and Design*, 2nd ed., Wiley, New York, 1996.
2. J. R. James, P. S. Hall, and C. Wood, *Microstrip Antenna Theory and Design*, Peter Peregrinus, London, 1981.
3. D. M. Pozar and D. H. Schaubert, Comparison of three series fed microstrip array geometries, *IEEE Antennas Propagation Symp.*, Ann Arbor, MI, July 1993, pp. 728–731.
4. D. M. Pozar and B. Kaufman, Design considerations for low sidelobe microstrip arrays, *IEEE Trans. Antennas Propag.* **38**: 1176–1185 (Aug. 1990).
5. S. D. Targonski and D. M. Pozar, Design of wideband circularly polarized aperture coupled microstrip antennas, *IEEE Trans. Antennas Propag.* **41**: 214–220 (Feb. 1993).
6. A. Nirmalathas, C. Lim, D. Novak, and R. B. Waterhouse, Progress in millimeter-wave fiber-radio access networks (invited), *Ann. Telecommun.* **56**: 27–38 (Jan./Feb. 2001).
7. D. M. Pozar, The active element pattern, *IEEE Trans. Antennas Propag.* **29**: 1176–1178 (Aug. 1994).
8. J. Huang, A parallel-series-fed microstrip array with high efficiency and low cross-polarization, *Microwave Opt. Technol. Lett.* **5**: 230–233 (May 1992).
9. R. J. Mailloux, J. F. McIlvenna, and N. P. Kernweis, Microstrip array technology, *IEEE Trans. Antennas Propag.* **29**: 25–37 (Jan. 1981).
10. E. Levine, G. Malamud, S. Shtrikman and D. Treves, A study of microstrip array antennas with the feed network, *IEEE Trans. Antennas Propag.* **37**: 426–434 (April 1989).
11. L. Mall and R. B. Waterhouse, Millimeter-wave proximity-coupled microstrip antenna on an extended hemispherical dielectric lens, *IEEE Trans. Antennas Propag.* (in press).
12. D. M. Pozar, S. D. Targonski, and H. D. Syrigos, Design of millimeter-wave microstrip reflectarrays, *IEEE Trans. Antennas Propag.* **45**: 287–296 (Feb. 1997).
13. R. B. Waterhouse, D. Novak, A. Nirmalathas, and C. Lim, Broadband printed antennas for point-to-point and point-to-multipoint wireless millimetre-wave applications, *IEEE Antennas Propagation Symp.* Utah (USA), July 2000, pp. 1390–1393.

14. S. D. Targonski and R. B. Waterhouse, Microstrip reflectarray analysis and design techniques, *5th Australian Symp. Antennas*, Sydney, Australia, Feb. 1996, p. 20.

15. D. M. Pozar and D. H. Schaubert, Scan blindness in infinite arrays of printed dipoles, *IEEE Trans. Antennas Propag.* **32**: 602–610 (June 1984).

16. D. M. Pozar, Scanning characteristics of infinite arrays of printed antenna subarrays, *IEEE Trans. Antennas Propag.* **40**: 666–674 (June 1992).

17. D. Novak and R. B. Waterhouse, Impedance behaviour and scan performance of microstrip patch arrays configurations suitable for optical beamforming networks, *IEEE Trans. Antennas Propag.* **42**: 432–435 (March 1994).

18. S. D. Targonski, R. B. Waterhouse, and D. M. Pozar, Design of wideband aperture-stacked patch microstrip antennas, *IEEE Trans. Antennas Propag.* **46**: 1246–1251 (Sept. 1998).

19. R. B. Waterhouse, Stacked patches using high and low dielectric constant material combination, *IEEE Trans. Antennas Propag.* **47**: 1767–1771 (Dec. 1999).

20. R. B. Waterhouse, Design and performance of large arrays of aperture stacked patches, *IEEE Trans. Antennas Propag.* **49**: 292–297 (Feb. 2001).

21. F. Zavosh and J. T. Aberle, Infinite phased arrays of cavity-backed patches, *IEEE Trans. Antennas Propag.* **42**: 390–398 (March 1994).

22. R. B. Waterhouse, The use of shorting posts to improve the scanning range of probe-fed microstrip patch phased arrays, *IEEE Trans. Antennas Propag.* **44**: 302–309 (March 1996).

23. R. B. Waterhouse, Design and scan performance of large, probe-fed stacked microstrip patch arrays, *IEEE Trans. Antennas Propag.* (in press).

MICROWAVE WAVEGUIDES

M. HAJIAN
L. P. LIGTHART
Delft University of Technology
Delft, The Netherlands

1. INTRODUCTION

A waveguide is used to guide electromagnetic waves at microwave frequency regions, and an open-ended waveguide to radiate them; which one is used depends on the application. A waveguide usually consists of a hollow metal pipe whose cross section is rectangular or circular. In most applications the open-ended waveguide is used as a feed element for a large reflector antenna or as an antenna element in active or passive phased-array antennas. Waveguide antennas provide optimum RF performance, that is, high aperture efficiency, high polarization purity, and low VSWR. Apart from inherent wide-bandwidth characteristics, they have the unique feature of a highpass filter behavior. In general the waveguide antenna protects the receiver system from unwanted electromagnetic interference (EMI) at frequencies lower than the cutoff frequencies of the waveguide modes. The reasons for its popularity are:

- Robustness
- Wide bandwidth
- High polarization purity
- Easy to mount on surfaces of spacecraft or aircraft for space applications
- Standard for calibrating other antennas
- Element for protecting the fields of larger transmit antenna
- Easy to manufacture

Energy transport in a waveguide can be achieved through propagation of so-called electromagnetic wave modes. These modes are solutions of the Maxwell equations and satisfy the boundary conditions. Such a hollow waveguide has characteristic cutoff frequencies connected with the propagation modes. The propagation of these modes depends on the operational frequency. If the frequency of the signal entering the waveguide is higher than the cutoff frequency of a given mode, then the electromagnetic mode energy can be transported through the waveguide with minimum attenuation because of the conduction losses in the waveguide walls. If the frequency of the incoming signal is lower than the cutoff frequency of a given mode, then the electromagnetic mode field is attenuated to a very low value within a short distance. It is convenient to design the waveguide such that the electromagnetic energy can be guided through the mode with only the lowest cutoff frequency. This mode is called the *fundamental* or *dominant mode*.

2. WAVEGUIDE APPLICATIONS

The waveguide is a low-loss transmission line that can handle high power signals. Since losses increase with frequency, waveguide applications can be found in the microwave and millimeter-wave region. In telecommunication and radar systems where losses may give major problems waveguide components are attractive. In other applications such as those in satellite communications, ground stations, and radar, where high power is a necessary requirement, waveguide solutions become attractive because they satisfy high power-handling capabilities.

The shape of the waveguide and its inner structure can be reconfigured in order to realize passive microwave components such as filters, couplers, phase shifters or active components such as oscillators and amplifiers [1]. The design and measurement results of a multichannel waveguide power divider based on *H*-plane or *E*-plane geometry have been discussed [2]. A corrugated waveguide has been used [3] to realize a phase shifter as part of a high-power dual reflector array antenna. Yoneyama developed a transmit/receive system at 35 GHz based on the NRDW (nonradiative dielectric waveguide) [4]. In NRDW the millimeter-wave field is concentrated inside the dielectric and it propagates similar as in a metal waveguide, meaning that minimum "leakage" takes place.

The waveguide is extensively used in medical applications (e.g., cancer therapy) where electromagnetic energy is coupled into the human body [5]. For this purpose the waveguide is loaded with a lossless dielectric material

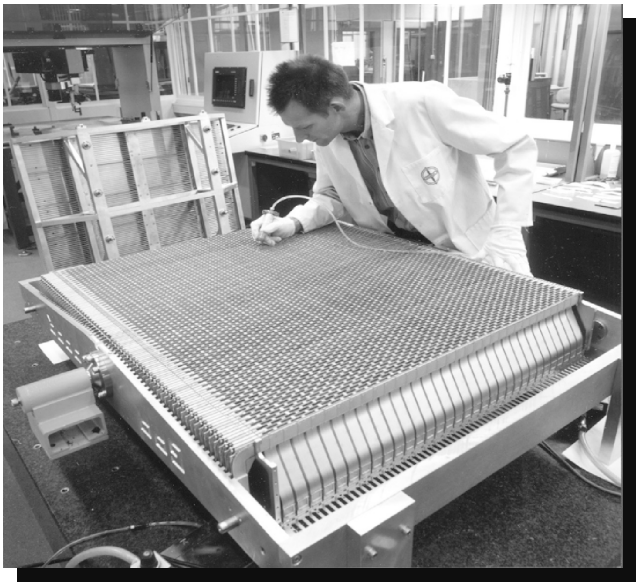


Figure 1. APAR open-ended waveguide array antenna under construction. (Courtesy of THALES.)

with a permittivity equal to that of the muscle tissue. This approach provides good impedance matching and results into a concentrated energy transfer.

The waveguide is often applied as an antenna element in large or phased-array antennas. For example, the active phased array antenna (APAR) from THALES uses open-ended waveguides as antenna elements [6,7]. APAR has four antenna-array panels; each panel consists of more than 4096 waveguide radiators (Fig. 1). Each waveguide radiator is connected to a T/R element, which comprises a sum channel and a combined transmit/delta elevation channel for monopulse tracking radar. The antennas are designed to have a wide angular scan range (up to 70° from the antenna broadside) for full 360° coverage, fast electronic beam steering to support search functions, and simultaneous tracking of hundreds of targets [8]. The APAR waveguide array uses a dielectric sheet for wide-angle impedance matching (WAIM sheet). Figure 2 shows an artist's impression of an antenna array panel integrated with the T/R modules and combiner networks.

It is possible to realize a linear array by using so-called slotted waveguides: narrow openings in the waveguide surface. A proper design of the slotted waveguide array may result in antennas with high efficiency, ultralow sidelobes and can sustain high peak power in the order of kilowatts. Figure 3 shows an example of a planar slotted waveguide array in X band.

The Earth Observation Satellites (ERS-1 and ERS-2) use a slotted waveguide array. The satellites were launched by the European Space Agency in 1991 and 1995, respectively. The ERS-1 antenna system under test can be seen in Fig. 4.

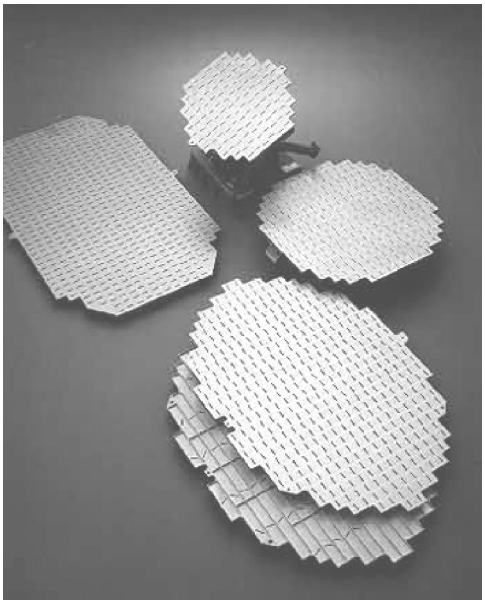


Figure 3. Slotted waveguide array. (Courtesy of ELTA Electronics Industries.)

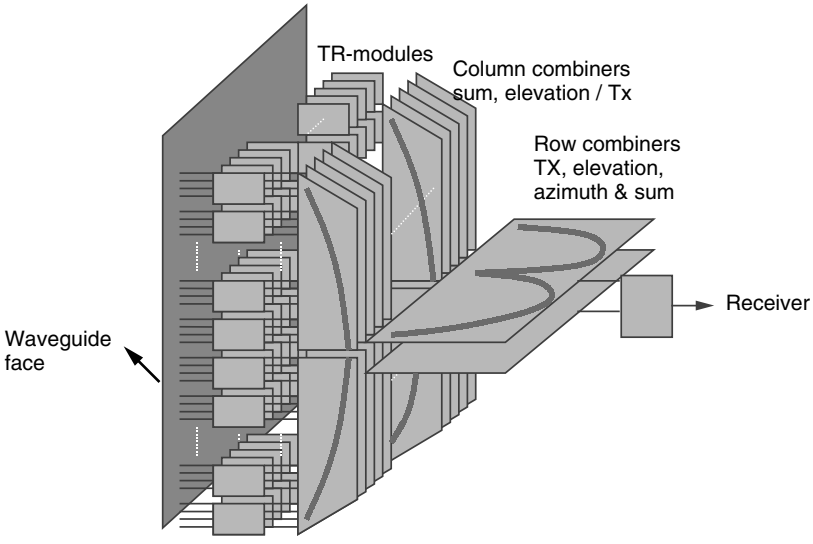


Figure 2. Artist impression of antenna RF network of APAR. (Courtesy of THALES.)

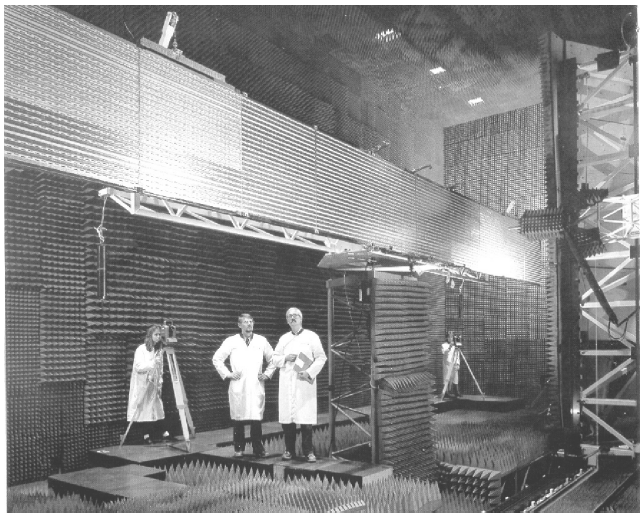


Figure 4. Slotted waveguide array antenna of the European remote-sensing satellite ERS-1 during planar near-field measurements. (Courtesy of Ericsson-ESA.)

Several applications of waveguide antennas in earth stations for satellite communication can be found in the literature. Bird et al. designed and measured [9] a compact high-power S-band dual frequency and dual polarized waveguide feed system with high power capability (handling 2 kW continuous RF) [9]. A second example is given by Bird and Sprey [10], who designed and measured a circularly polarized X-band feed system with high transmit/receive isolation for a dual-shaped Cassegrain reflector. A dual-mode waveguide filter and a 2-step waveguide *E*-plane filter has also been designed and measured [11,12].

The opening of the waveguide with different cross sections is tapered (flared) to a larger opening to form a so-called horn antenna. Such antennas are widely used as feed elements for large-sized radioastronomy, satellite and communication dishes. In the following an overview of different configurations with specific examples is given.

2.1. Single-Feed Systems

Figure 5 shows a selection of waveguide horn antennas for space applications.

2.1.1. TV-SAT Horn Antennas. The elliptical corrugated horn radiator is used in the TV-SAT feeding system in a reflector. The pattern has a high-gain elliptical beam (3-dB beamwidths $0.72^\circ \times 1.62^\circ$). The operational frequency is 11.7–12.1 GHz. It is circularly polarized and has low cross polarization (decoupling >35 dB).

2.1.2. IntelSAT 8 and Nahuel Horn Antennas. These conical corrugated horn antennas are part of the feed system of a dual-reflector Gregorian and a shaped reflector antenna in the frequency range of 10.95–14.5 GHz. They are linearly polarized and combine the transmit/receive function for both polarizations with low cross-polar coupling (decoupling >40 dB). The antennas can handle 12 high-power carriers up to 120 W per polarization.

2.2. Multifeed Systems

Figure 6 shows some multifeed antenna systems.

2.2.1. DFH-3 Feed. The seven-element diagonal-waveguide horn-antenna cluster combines a transmit (4 GHz) and receive (6 GHz) diplexer with a three-layer coaxial beamforming network (BFN).

2.2.2. AMOS 8 Feed. This three-element conical corrugated horn antenna cluster is part of an offset reflector antenna in the frequency range of 10.95–14.5 GHz.

2.2.3. IntelSAT 8 Feed Array. This 96-element conical corrugated horn antenna cluster as a feed system of a multibeam antenna provides eight beams with stringent

Figure 5. Waveguide horn antennas for space applications. (Courtesy of Astrium.)

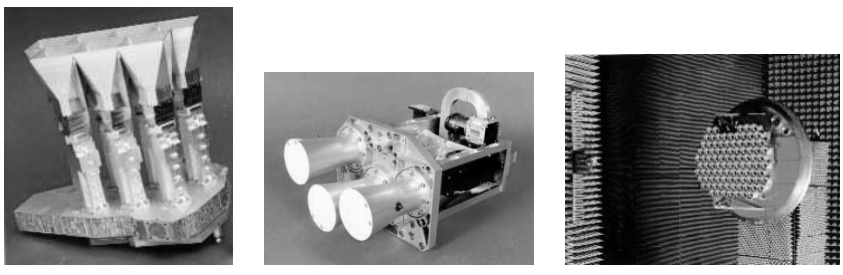


Figure 6. Multifeed waveguide systems for space applications. (Courtesy of Astrium.)

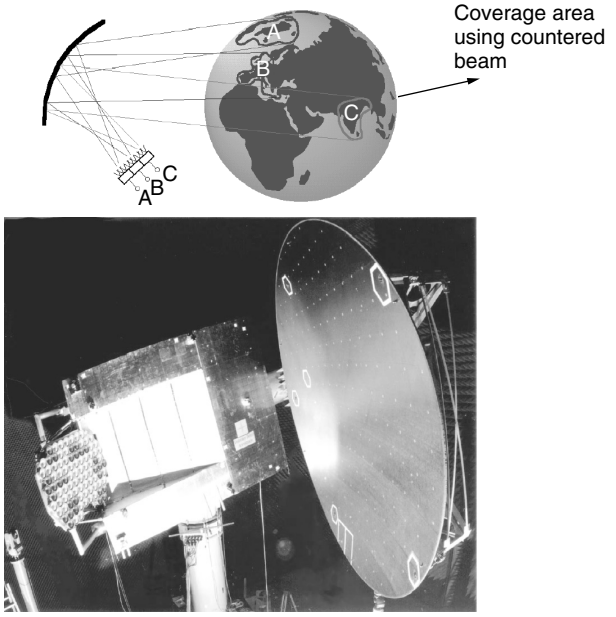


Figure 7. Reflector antenna with multiple contoured beams using the combined multifeed concept applied in Intelsat-8 (C band, transmitter and receiver) and in Intelsat-9 (C band, transmitter and receiver). (Courtesy of Astrium.)

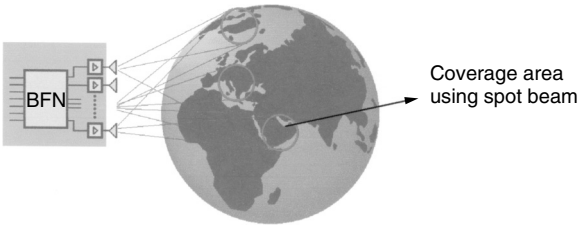
interbeam isolations (better than 27 dB) in order to support multifold frequency reuse. A three-layer coaxial beamforming network (BFN) generates eight individual beams: two left-hand circularly polarized hemispherical beams and six right-hand polarized "zone" beams. The power capability is 1.5 kW RF. Figure 7 illustrates the concept of generating multiple contoured beams. The same figure also shows the complete antenna system during the measurement phase.

Figure 8 shows a compact 8×8 dual-polarized waveguide array for application in a direct radiating antenna array. It was developed for space-based high-resolution polarimetric synthetic aperture radar (SAR) antennas in the X band. The height of the element including the dual polarized feed section is less than 0.3λ. A balun-type feed is used to excite the orthogonal fundamental modes with a polarization purity better than 40 dB over a bandwidth greater than 5%. This technique allows tight packaging in the array configuration and supports low-loss distribution/combiner networks.

In near-field antenna measurement techniques the waveguide antenna is used as a probe antenna to measure the radiation characteristics of the antenna under test (AUT). Since its characteristics are well measured and documented, correcting the probe to determine the far field accurately is more straightforward.

3. RECTANGULAR WAVEGUIDE

Figure 9 shows the cross section of a rectangular waveguide with a width and height of a and b respectively. It is assumed that the waveguide is filled with air and is infinite in length. There are a number of transverse electric- and magnetic modes (TE^x, TM^x, TE^y, TM^y, TE^z, TM^z) that satisfy the boundary conditions and are a solution to the Maxwell equations. The desired mode can be generated by the feed structure in the waveguide. This will be explained later in this article. However, without loss of generality in this part only TE^z is considered. Note that since the TEM mode does not satisfy the boundary conditions in the waveguide, it cannot be used to transport electromagnetic energy through this mode in the waveguide [11].



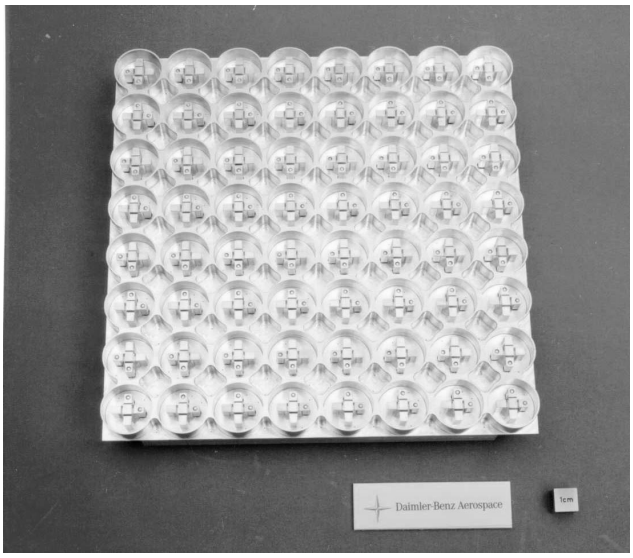


Figure 8. Open-waveguide planar array for space-based synthetic aperture radar (SAR) applications. (Courtesy of Astrium.)

3.1. Transverse Electric (TE^z)

Transverse electric modes are field configurations whose electric-field components lie in a plane that is transverse to the direction of the wave propagation. For example TE^z implies that $E_z = 0$. The other field components may or may not exist.

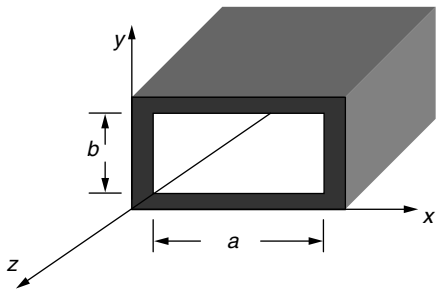


Figure 9. The rectangular waveguide with its dimensions and coordinates.

To derive the field expressions in a rectangular coordinate system that are TE to given direction, one needs only to let the magnetic vector potential $\mathbf{F}$ have only one component in that direction. Other components of electric vector potential $\mathbf{A}$, and $\mathbf{F}$ are set equal to zero. This corresponds to the following condition

$$\begin{aligned} \mathbf{A} &= 0 \\ \mathbf{F} &= \hat{a}_z F_z(x, y, z) \end{aligned} \tag{1}$$

where F_z is the scalar potential function and it represent the z component of vector potential $\mathbf{F}$. In the source-free region for the transverse electric modes, the components of the electric and magnetic fields satisfy the following equations [11]

$$\begin{aligned} E_x &= -\frac{1}{\varepsilon}\frac{\partial F_z}{\partial y} & H_x &= -j\frac{1}{\omega\mu\varepsilon}\frac{\partial^2 F_z}{\partial x\,\partial z} \\ E_y &= \frac{1}{\varepsilon}\frac{\partial F_z}{\partial x} & H_y &= -j\frac{1}{\omega\mu\varepsilon}\frac{\partial^2 F_z}{\partial y\,\partial z} \\ E_z &= 0 & H_z &= -j\frac{1}{\omega\mu\varepsilon}\left(\frac{\partial^2}{\partial z^2} + \beta^2\right)F_z \end{aligned} \tag{2}$$

where ε and μ are the permittivity in F/m and the permeability in H/m of the medium, respectively; ω is the frequency of the impressed signal; and β is the free-space wavenumber. The scalar potential F_z satisfies the scalar wave or Helmholtz equation

$$\nabla^2 F_z + \beta^2 F_z = 0 \tag{3}$$

In rectangular coordinates, this equation becomes

$$\frac{\partial^2 F_z}{\partial x^2} + \frac{\partial^2 F_z}{\partial y^2} + \frac{\partial^2 F_z}{\partial z^2} + \beta^2 F_z = 0 \tag{4}$$

The solution to Eqs. (3) and (4) is a well-known problem in the literature. The solution is based on the separation of variables and has the form of

$$F_z = \psi(x)\varphi(y)\zeta(z) \tag{5}$$

The variation in the z direction represents the propagating waves such that $\zeta(z)$ has the following form

$$\zeta(z) = A_1 e^{-j\beta_z z} + B_1 e^{+j\beta_z z} \tag{6}$$

where $\pm$ represents the waves traveling in the $+$ and $-z$ direction, respectively. It is assumed that the source in the waveguide is located such that only the waves in the $+z$ direction exist. In this case B_1 is zero.

The variation in the x and y directions represent the standing waves since the guide is bounded in these directions. The most appropriate solution is

$$\begin{aligned} \psi(x) &= A_2\cos(\beta_x x) + B_2\sin(\beta_x x) \\ \varphi(y) &= A_3\cos(\beta_y y) + B_3\sin(\beta_y y) \end{aligned} \tag{7}$$

where A_1, A_2, B_2, A_3, B_3 and $\beta_x, \beta_y, \beta_z$ are constants that need to be evaluated using the boundary conditions. $\beta_x, \beta_y, \beta_z$ are the wavenumbers in the x, y, and z directions, respectively, and they are related to the free-space wavenumber β in rad/m as follows:

$$\beta_x^2 + \beta_y^2 + \beta_z^2 = \beta^2 = \omega^2\mu\varepsilon \tag{8}$$

Substituting (6) and (7) in (5) with $B_1 = 0$ leads to

$$\begin{aligned} F_z(x, y, z) = {} & [A_2\cos(\beta_x x) + B_2\sin(\beta_x x)] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] * A_1 e^{-j\beta_z z} \end{aligned} \tag{9}$$

Equation (9) applies for $+z$ traveling waves. Note that here for simplicity the sign $+$ is omitted. Since the waveguide walls are good conductors, the tangential components of the electric field will vanish on the waveguide walls. For Fig. 9, the following boundary conditions exist for the left and right sidewalls:

$$\begin{aligned} E_y(x = 0, 0 \le y \le b, z) &= E_y(x = a, 0 \le y \le b, z) = 0 \\ E_z(x = 0, 0 \le y \le b, z) &= E_z(x = a, 0 \le y \le b, z) = 0 \end{aligned} \tag{10}$$

and for the top and bottom walls

$$\begin{aligned} E_x(0 \le x \le a, y = 0, z) &= E_x(0 \le x \le a, y = b, z) = 0 \\ E_z(0 \le x \le a, y = 0, z) &= E_z(0 \le x \le a, y = b, z) = 0 \end{aligned} \tag{11}$$

Equations (2), (10), and (11) are used to determine the constants in Eq. (9). Substituting (9) in (2), the y component of the electric field can be written as

$$\begin{aligned} E_y(x, y, z) = {} & \frac{\beta_x}{\varepsilon}[-A_2\sin(\beta_x x) + B_2\cos(\beta_x x)] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] * A_1 e^{-j\beta_z z} \end{aligned} \tag{12}$$

Applying the boundary condition given by Eq. (11) on the left wall for the E_y component to Eq. (12) gives

$$\begin{aligned} E_y(x = 0, 0 \le y \le b, z) = {} & \frac{\beta_x}{\varepsilon}[B_2] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] \\ & * A_1 e^{-j\beta_z z} = 0 \end{aligned} \tag{13}$$

Equation (13) can be satisfied if and only if B_2 is equal to zero. Applying the boundary condition of the right wall to Eq. (12) leads to

$$\begin{aligned} E_y(x = a, 0 \le y \le b, z) = {} & \frac{\beta_x}{\varepsilon}[-A_2\sin(\beta_x a)] \\ & * [A_3\cos(\beta_y y) + B_3\sin(\beta_y y)] \\ & * A_1 e^{-j\beta_z z} = 0 \end{aligned} \tag{14}$$

Equation (14) can be satisfied for a nontrivial solution if and only if

$$\sin(\beta_x a) = 0, \quad \beta_x a = m\pi \quad m = 0, 1, 2, \ldots \tag{15a}$$

$$\beta_x = \frac{m\pi}{a} \quad m = 0, 1, 2, \ldots \tag{15b}$$

Usually Eqs. (15a) and (15b) are called the *eigenfunction* and *eigenvalue*. It is straightforward to show that the following relations exists, using the same procedure

$$\begin{aligned} B_3 &= 0 \\ \beta_y &= \frac{n\pi}{b} \quad n = 0, 1, 2, \ldots \end{aligned} \tag{16}$$

Substituting Eqs. (16) and (15) in (9) and letting $A_1A_2A_3 = A$ leads to

$$F_z(x, y, z) = A \cos\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) e^{-j\beta_z z} \tag{17}$$

Substituting (17) in (2) leads to the complete solution of TE^z_{mn} modes

$$\begin{aligned} E_x &= A\frac{\beta_y}{\varepsilon} \cos(\beta_x x) \sin(\beta_y y) e^{-j\beta_z z} \\ E_y &= A\frac{\beta_x}{\varepsilon} \sin(\beta_x x) \cos(\beta_y y) e^{-j\beta_z z} \\ E_z &= 0 \\ H_x &= A\frac{\beta_x\beta_z}{\omega\mu\varepsilon} \sin(\beta_x x) \cos(\beta_y y) e^{-j\beta_z z} \\ H_y &= A\frac{\beta_z\beta_y}{\omega\mu\varepsilon} \cos(\beta_x x) \sin(\beta_y y) e^{-j\beta_z z} \\ H_z &= -jA\frac{\beta_x^2 + \beta_y^2}{\omega\mu\varepsilon} \cos(\beta_x x) \cos(\beta_y y) e^{-j\beta_z z} \end{aligned} \tag{18}$$

where β_x and β_y are the wavenumbers (eigenvalues) in the x and y directions, respectively. They are related to the wavelengths of the wave inside the waveguide in the x and y directions and the wave number in the z direction β_z and β as follows:

$$\begin{aligned} \beta_y &= \frac{n\pi}{b} = \frac{2\pi}{\lambda_y} \quad n = 0, 1, 2, \ldots \\ \beta_x &= \frac{m\pi}{a} = \frac{2\pi}{\lambda_x} \quad m = 0, 1, 2, \ldots \\ \beta_z^2 &= \beta^2 - (\beta_x^2 + \beta_y^2) = \beta^2 - \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right] \\ \beta &= \frac{2\pi}{\lambda} \end{aligned} \tag{19}$$

The values of β_z depend on the waveguide cutoff frequency and its value determines the propagating waves, standing waves, and evanescent waves. The cutoff frequency and cutoff wavenumber in turn are determined by letting $\beta_z = 0$ in Eq. (8) or (19)

$$\beta_c^2 = \beta^2 = \omega^2\mu\varepsilon = \omega_c^2\mu\varepsilon = (\beta_x^2 + \beta_y^2) = \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right] \tag{20}$$

which leads to

$$\begin{aligned} 2\pi f_c\sqrt{\mu\varepsilon} &= \sqrt{\left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right]} \\ &\times \begin{cases} (f_c)_{mn} = \dfrac{1}{2\pi\sqrt{\mu\varepsilon}}\sqrt{\left[\left(\dfrac{m\pi}{a}\right)^2 + \left(\dfrac{n\pi}{b}\right)^2\right]} \\ \qquad \times \left\{\begin{matrix} n = 0, 1, 2, \ldots \\ m = 0, 1, 2, \ldots \end{matrix}\right\} m = n \neq 0 \end{cases} \end{aligned} \tag{21}$$

Since it is assumed that the z direction is the propagation axis, the integers m and n denote the number of half-waves of electric or magnetic field intensity in the x and y directions. Depending on the value of the cutoff frequency, the propagation constant β_z can take different values. Three different cases are distinguished and they are given here as follows:

$$\beta_c^2 \triangleq \beta_x^2 + \beta_y^2 = \beta^2 - \beta_z^2$$

where

$$\beta_z = \begin{cases} \pm\sqrt{\beta^2 - \beta_c^2} = \pm\beta\sqrt{1 - \left(\dfrac{f_c}{f}\right)^2} \\ \qquad \text{for } f > f_c \text{ propagating waves} \\ 0 \qquad \text{for } f = f_c \text{ standing waves} \\ \pm j\sqrt{\beta_c^2 - \beta^2} = \pm j\beta\sqrt{\left(\dfrac{f_c}{f}\right)^2 - 1} \\ \qquad \text{for } f < f_c \text{ evanescent waves} \end{cases} \tag{22}$$

The evanescent waves are the fields that decay exponentially. Equation (22) also shows that the waveguide has highpass filter behavior. If the operational frequency is higher than the cutoff frequency, the fields propagate; if not, they attenuate. The ratio of suitable electric to magnetic field components has the same dimension as the impedance. Using (18) the following relation exists:

$$Z \triangleq \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \frac{\omega\mu}{\beta_z}\Omega \tag{23}$$

Inserting Eq. (22) in (23) leads to the following expression for the waveguide impedance

$$Z = \begin{cases} \dfrac{\eta}{\sqrt{1 - \left(\dfrac{f_c}{f}\right)^2}} & \text{for } f > f_c \text{ resistive} \\ \infty & \text{for } f = f_c \text{ open circuit} \\ j\dfrac{\eta}{\sqrt{\left(\dfrac{f_c}{f}\right)^2 - 1}} & \text{for } f < f_c \text{ inductive} \end{cases} \tag{24}$$

where η is the free-space impedance. The waveguide impedance behaves inductively for frequencies lower than the cutoff frequency, and resistively for frequencies higher than the cutoff frequency. Figure 10 shows the waveguide impedance as function of the normalized frequency.

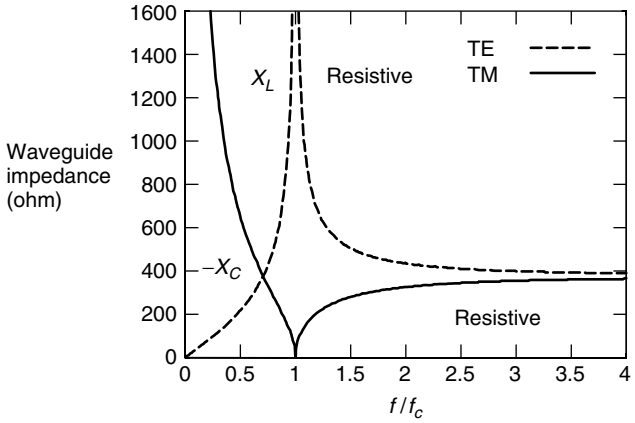


Figure 10. The wave impedance of a rectangular waveguide.

The expression for the wavenumber β_z along the z axis can be used to define the wavelength along the guide axis and is given as

$$\lambda_z = \begin{cases} \dfrac{\lambda}{\sqrt{1-\left(\dfrac{f_c}{f}\right)^2}} = \dfrac{\lambda}{\sqrt{1-\left(\dfrac{\lambda}{\lambda_c}\right)^2}} & \text{for } f > f_c \\ \infty & \text{for } f = f_c \\ j\dfrac{\lambda}{\sqrt{\left(\dfrac{f_c}{f}\right)^2-1}} = \dfrac{\lambda}{\sqrt{1-\left(\dfrac{\lambda}{\lambda_c}\right)^2}} & \text{for } f < f_c \end{cases} \tag{25}$$

Figure 11 shows the waveguide, wavelength, and wavenumber along the guide axis as function of the normalized frequency. Note that depending on the value of the signal and cutoff frequency, the waves are propagating or attenuating. The expression for the waveguide impedance for the TM mode is not given in this article. A procedure similar to TE can be used to derive the related parameters [11].

Example 1. The waveguide in Fig. 9 has inner dimensions of $a = 2$ cm, $b = 1$ cm, is filled with air and operates in the TE_{10} mode.

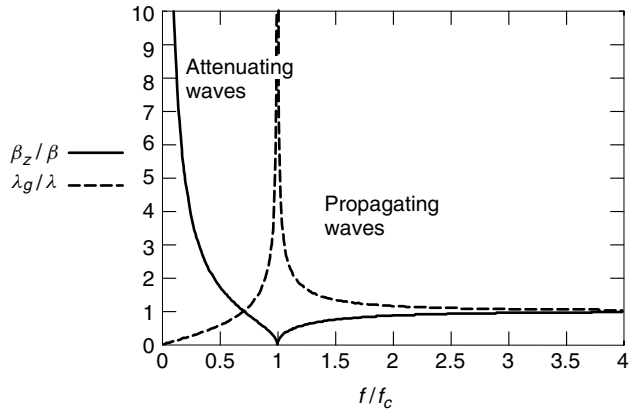


Figure 11. The normalized wavelength and propagation constant.

a. Determine the cutoff frequency.

b. Determine the propagation constant at 9.5 GHz.

c. Determine the waveguide impedance at 7.0 and 9.5 GHz.

Now the waveguide is filled with material. The dielectric constant of the material equals 5.

d. Find the cutoff frequency.

e. Determine the new waveguide dimensions to obtain the same cutoff frequency as derived in part a.

Solution

a. $(f_c)_{10} = \dfrac{c}{2a} = \dfrac{3\times10^8}{2\times2\times10^{-2}} = 7.5 \text{ GHz}$

b. $\beta_z = \beta\sqrt{1-\left(\dfrac{f_c}{f}\right)^2} = \dfrac{2\pi}{\lambda_0}\sqrt{1-\left(\dfrac{f_c}{f}\right)^2}$

$$= \frac{2\pi}{\dfrac{3\times10^8}{9.5\times10^9}}\sqrt{1-\left(\frac{7.5}{9.5}\right)^2} = 122.12 \text{ rad/m}$$

c. For 7.0 GHz:

$$Z = j\frac{\eta}{\sqrt{\left(\dfrac{f_c}{f}\right)^2-1}} = j\frac{120\pi}{\sqrt{\left(\dfrac{7.5}{7}\right)^2-1}}$$

$$= j980\ \Omega \text{ (inductive)}$$

For 9.5 GHz:

$$Z = \frac{\eta}{\sqrt{1-\left(\dfrac{f_c}{f}\right)^2}} = \frac{120\pi}{\sqrt{1-\left(\dfrac{7.5}{9.5}\right)^2}}$$

$$= 614\ \Omega \text{ (resistive)}$$

d. $(f_c)_{10} = \dfrac{c}{2a\sqrt{\varepsilon_r}} = \dfrac{3\times10^8}{2\times2\times10^{-2}\sqrt{5}} = 3354 \text{ GHz}$

e. $(f_c)_{10} = \dfrac{c}{2a\sqrt{\varepsilon_r}} = \dfrac{3\times10^8}{2\times a\sqrt{5}} = 7.5\times10^9, a = 8.94 \text{ mm}$

In the second part of Example 1 a miniaturization aspect of waveguides is introduced. In Section 4 the theory and practice of miniaturization and matching of a dielectric-filled waveguide will be discussed.

3.2. Power in Rectangular Waveguide

The power transport is associated with the fields propagating in the waveguide. The total power in the waveguide is the summation of the power of the TE_{mn} and TM_{mn} modes and is given as

$$P_{\text{total}} = \sum_m \sum_n P_{mn}^{\text{TE}} + \sum_m \sum_n P_{mn}^{\text{TM}} \tag{26}$$

In this section the expression for P_{mn}^{TE} is derived. A similar procedure can be used to derive P_{mn}^{TM}. The power is calculated by integrating the power density related to the

electromagnetic fields over the cross-sectional area of the waveguide and is given by

$$P_{mn} = \iint_{S_0} W_{mn} \cdot dS = \iint_{S_0} W_{mn} \cdot \hat{\boldsymbol{n}}\, ds$$
$$= \frac{1}{2} \iint_{S_0} \mathrm{Re}[\boldsymbol{E} \times \boldsymbol{H}^*]_{mn} \cdot \hat{\boldsymbol{n}}\, ds \tag{27}$$

where W_{mn} is the power density related to mn^{th} mode and $dS = dx\,dy$ is the infinitesimal area of the waveguide cross section and $\hat{\mathbf{n}} = \hat{n}_z$ is the unit vector normal to the waveguide cross section. The power density is given by

$$W_{mn} = \tfrac{1}{2}\mathrm{Re}[\boldsymbol{E} \times \boldsymbol{H}^*]_{mn} = \tfrac{1}{2}\mathrm{Re}[(\hat{n}_x E_x + \hat{n}_y E_y) \times (\hat{n}_x H_x + \hat{n}_y H_y)^*] = \tfrac{1}{2}\hat{n}_z \mathrm{Re}[E_x H_y^* - E_y H_x^*] \tag{28}$$

Inserting the given field components given by Eq. (18) in (28) and using (27) leads to the following expression for the power:

$$P_{mn} = \frac{1}{2}\int_0^a \int_0^b \hat{n}_z \mathrm{Re}[E_x H_y^* - E_y H_x^*]_{mn} \cdot \hat{n}_z\, dx\, dy$$
$$= \frac{1}{2}\int_0^a \int_0^b \mathrm{Re}[E_x H_y^* - E_y H_x^*]_{mn}\, dx\, dy$$
$$= \frac{1}{2}|A|^2 \frac{\beta_z}{\omega\mu\varepsilon^2} \int_0^a \int_0^b [\beta_y^2 \cos^2(\beta_x x)\sin^2(\beta_y y) + \beta_x^2 \cos^2(\beta_y y)\sin^2(\beta_x x)]\, dx\, dy \tag{29}$$

Performing the integration and inserting the expression for the β_z given by Eq. (22) in the propagating case leads to the final expression for the power transport by the TE mode as

$$P_{mn} = \frac{1}{2}|A|^2 \frac{\beta_z}{\omega\mu\varepsilon^2}(\beta_y^2 + \beta_x^2)\left(\frac{a}{\delta_m}\right)\left(\frac{b}{\delta_n}\right)$$
$$= \frac{1}{2}|A|^2 \frac{\beta(\beta_y^2 + \beta_x^2)}{\omega\mu\varepsilon^2}\left(\frac{a}{\delta_m}\right)\left(\frac{b}{\delta_n}\right)\sqrt{1 - \left(\frac{f_c}{f}\right)^2} \tag{30}$$

where

$$\delta_k = \begin{cases} 1 & k = 0 \\ 2 & k \neq 0 \end{cases} \tag{31}$$

Example 2. The waveguide in Fig. 9 has inner dimensions $a = 2$ cm, $b = 1$ cm, is filled with air and operates in the TE_{10} mode. The frequency is 9.5 GHz. The peak value of the electric field is 40 kV/m. Calculate the transport power in the waveguide.

Solution Since the waveguide operates in TE_{10} mode, the field components can be found using Eqs. (18) and (19) with $m = 1$ and $n = 0$. They are given by

$$E_x = 0$$
$$E_y = A\frac{\beta_x}{\varepsilon}\sin\left(\frac{\pi}{a}x\right)e^{-j\beta_z z}$$
$$E_z = 0$$
$$H_x = A\frac{\beta_x\beta_z}{\omega\mu\varepsilon}\sin\left(\frac{\pi}{a}x\right)e^{-j\beta_z z} \tag{32}$$
$$H_y = 0$$
$$H_z = -jA\frac{\beta_x^2 + \beta_y^2}{\omega\mu\varepsilon}\cos\left(\frac{\pi}{a}x\right)$$

The peak value of the electric field intensity can be obtained from the maximum electric field value by using equation (32) and is given by

$$|E_y|_{\max} = \frac{|A|}{\varepsilon_0}\beta_x = \frac{|E_{10y}|}{\varepsilon_0}\frac{\pi}{a} = 40 \text{ kV/m} \tag{33}$$

The cutoff frequency of the TE_{10} mode is

$$f_c = \frac{c}{2a} = \frac{3\times 10^8}{2\times 2\times 10^{-2}} = \frac{3\times 10^{10}}{4} = 7.5 \text{ GHz} \tag{34}$$

Inserting Eqs. (33) and (34) into Eq. (30) leads to the maximum power

$$P_{10} = \frac{|E_y|^2}{2\eta}\left(\frac{a}{\delta_{01}}\right)\left(\frac{b}{\delta_{00}}\right)\sqrt{1 - \left(\frac{f_c}{f}\right)^2}$$
$$P_{10} = \frac{|40\times 10^3|^2}{2\times 377}\frac{2\times 10^{-2}}{2}\frac{1\times 10^{-2}}{1}\sqrt{1 - \left(\frac{7.5}{9}\right)^2} \tag{35}$$
$$P_{10} \simeq 117.30 \text{ W}$$

3.3. Excitations of Modes in a Rectangular Waveguide

The analysis given in the previous sections focused on the wave propagation and power transport in the waveguide. However, the electric and magnetic fields first need to be generated in the waveguide. In general this can be done by an infinitesimal electric or magnetic dipole element (probe) [12], which in turn is connected to a generator. In order to achieve the optimal interface, it is necessary to match the impedance of the feed element to the waveguide impedance. This means that the return loss should be minimized. It is also desired that the dielectric losses, and losses caused by sharp bends are as low as possible.

The position, dimensions, and depth of the probe play a major role in coupling the energy from the feedline into the waveguide. The reflection caused by the waveguide walls and the generated field at the antenna probe need to be in phase in order to reinforce each other and to allow the waves to propagate in the aperture direction. In most cases the distance between the probe and the short-circuited end wall of the waveguide is in the order

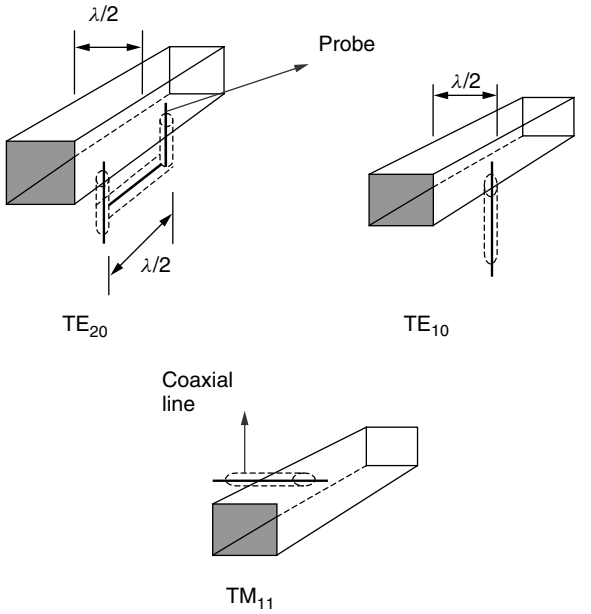


Figure 12. Methods used to excite various modes in a rectangular waveguide.

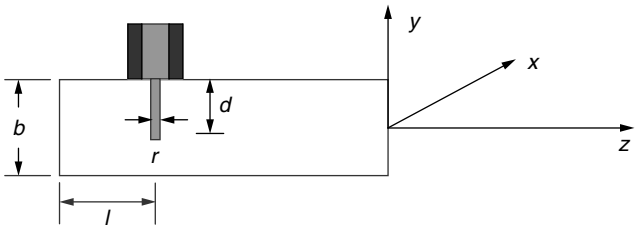


Figure 13. Configuration of coax-to-rectangular waveguide transition.

of half a wavelength. In this way the generated fields and reflected fields coming from the backside of the end wall are in phase. Figure 12 shows various methods to excite different modes in the waveguide. Figure 13 shows the geometry of a coaxial cable-to-waveguide transition. It consists of a coaxial line with its inner conductor extending over a distance d into the waveguide. To create a probe that radiates into one direction, a short is placed at a distance of l from the probe. By choosing l and d properly one can couple the power optimally from the coaxial line to the waveguide. r is the radius of the inner conductor of the probe.

The input impedance is inductive. Tian et al. [14] show that to obtain an optimal transition from coax to waveguide, one needs to introduce a capacitance between the coax end and the waveguide wall. When designing a coaxial feed, the major design problem is to find the optimal location and dimensions of the probe to achieve the best impedance matching. Equation (36) suggests that by properly choosing the probe length d and the short-circuit end position l, the radiation resistance R_{10} can be made equal to the characteristic impedance Z_0 of the coaxial line. In turn X can cancel the input reactance caused by the higher-order modes. The diameter of the coaxial feed is determined experimentally. The input impedance of the coaxial line is derived using the mode matching technique and is given by [13]

$$Z_{\text{in}} = R + jX$$

$$R_{10} = \frac{2Z_0}{ab\beta_{10}\beta}\sin^2(\beta_{10}l)\tan^2\left(\beta\frac{d}{2}\right)$$

$$X = \frac{Z_0}{ab\beta_{10}\beta}\tan^2\left(\beta\frac{d}{2}\right) \times \left\{\begin{array}{l} \ln\dfrac{2a}{\pi r} + \dfrac{0.0518\beta^2a^2}{\pi^2} \\ \quad + \dfrac{2\pi}{\beta_{10}a}\sin(2\beta_{10}l) \\ -2\left(1-\dfrac{2r}{a}\right) - 2\beta^2\displaystyle\sum_{m=1}^{\infty}\left[1-\frac{\sin^2\left(\dfrac{m\pi d}{2b}\right)}{\sin^2\left(\dfrac{\beta d}{2}\right)}\right] \\ \quad \times \dfrac{K_0(k_m r)}{k_m^2} \end{array}\right\}$$

$$k_m^2 = \left(\frac{m\pi}{b}\right)^2 - \beta^2$$

where a and b are the width and height of the waveguide, Z_0 is the free-space impedance, K_0 is the Bessel function of the second kind, and $\beta_{10} = \sqrt{\beta^2 - (\pi/a)^2}$ is the propagation constant of the fundamental mode. Figure 14 shows the values of l and d for tuning the input impedance of the probe for an X-band waveguide. The dimensions of the waveguide are a = 2.286 cm and b = 1.016 cm. The values make the inductive part of the input impedance equal to zero and force the resistive part to different characteristic impedances.

Figure 15 shows the optimum matching and measured input reflection of the coax transition of a dielectric filled waveguide in the L-band for $a = 83$ mm and $b = 10$ mm with a central frequency of 1.6 GHz. Since the height of the waveguide is very low, the waveguide behaves more

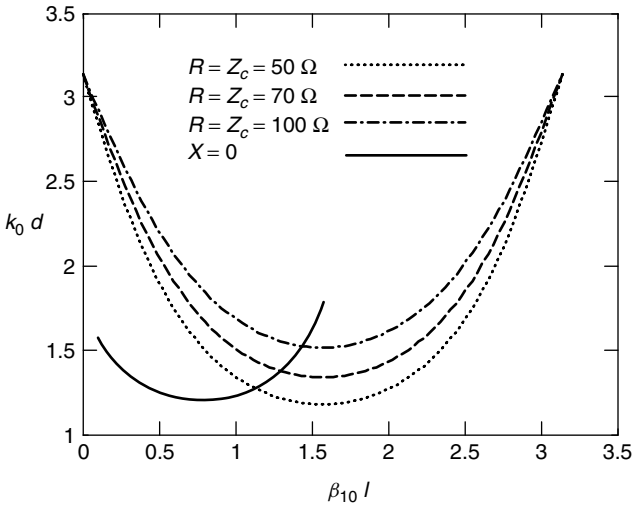


Figure 14. The design contour for matching the input impedance of the probe feed. (*Source*: R. E. Collin, *Field Theory of Guided Waves*.)

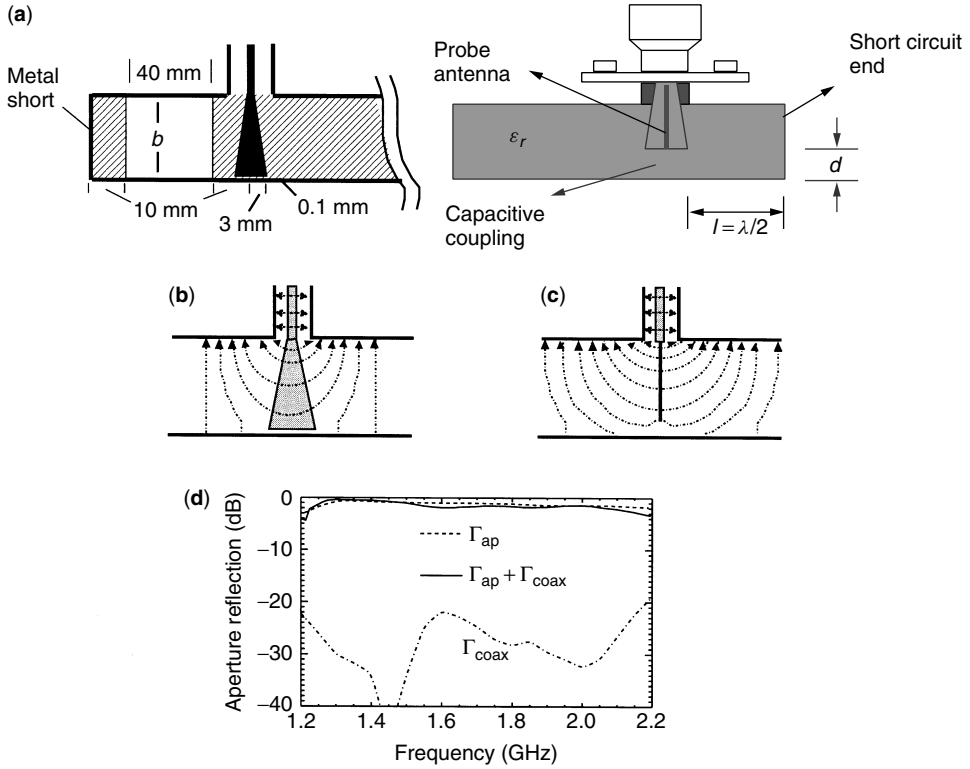


Figure 15. L-band coax-to-waveguide transition: (**a**) geometry of optimized capacitive coupling; (**b,c**) higher-order modes for a cone-shaped probe and monoprobe; (**d**) reflection matching of the coaxial waveguide transition, where Γ_{ap} and Γ_{coax} are the aperture and the input reflection at the coaxial interface.

or less as a cavity resonator. The probe feed needs to be tapered to a disk-cone form in order to increase the capacitive coupling and radiation resistance [14].

In many microwave and millimeter-wave planar circuit applications, such as active phased arrays or front ends in radar and radiocommunication systems, it is often necessary to use a microstrip line to excite the waveguide antenna [15]. Care is needed to couple the field generated at the source via the feed structure into the waveguide. The transition between the coaxial or microstrip feed is complex and needs to be analyzed, designed and experimentally verified.

There are several possible alternative transitions from microstrip to waveguide: microstrip *E*-plane probe (MEPP), finline transition, microstrip end launcher (MEL), ridged-waveguide transition, radiating-slot transition, and tapered-microstrip transition.

3.3.1. Microstrip *E*-Plane Probe (MEPP). Figure 16 shows the configuration of the MEPP. The probe behaves like a monopole antenna, which excites the waveguide. The reflector is used to reflect the excited waves in the desired direction.

A theoretical model has been developed [16] for calculating the input impedance. The model is based on an assumed current distribution in the probe, and a variational expression to calculate the input impedance.

Experimental results show that the MEPP has a −20 dB bandwidth of about 30% in the Ka band (26.5–40 GHz).

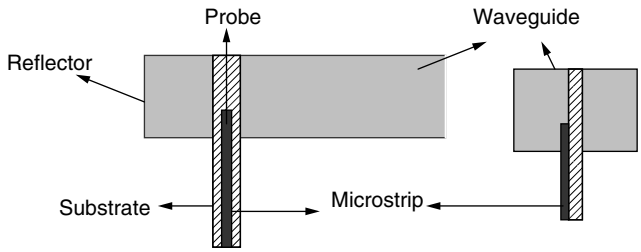


Figure 16. Microstrip *E*-plane probe (MEPP).

An advantage of MEPP is obviously its wide bandwidth; a disadvantage, however, is its non-planar construction. In the MEPP configuration, the microstrip T/R module would lie transversely to the waveguide, which is not very practical for miniature phased-array systems. Nevertheless, MEPP is one of the most widely used microstrip-to-waveguide transitions because of its simple configuration and its wide bandwidth.

3.3.2. Finline Transition. Figure 17 illustrates the finline transition where tapered antipodal fins are used to rotate the dominant TE_{10} mode of the waveguide into the TEM mode of the microstrip line. This transition does not require a reflector, since the bifurcation of the waveguide due to the microstrip ground plane serves as an imaginary reflector.

In Ref. 17 experimental results show a −20 dB bandwidth of 25% in the band 18–26 GHz. Unfortunately, this

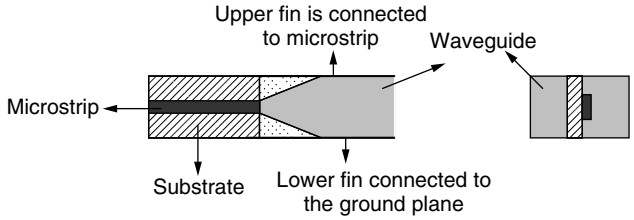


Figure 17. The geometry of finline transition.

paper does not present a theoretical model for analyzing such a transition. In Ref. 18 an empirical model based on a T matrix of a uniform finline section is given. The tapered finline is divided into a large number of uniform finlines. The total T matrix of the transition is calculated by taking the product of the T matrices of the individual uniform sections. In Ref. 19 an empirical expression for the resonance frequencies that may occur in this type of transition is given. By using this expression, one can place the resonance frequencies outside the operational frequency band in the design stage.

Advantages of this transition are the wide bandwidth and its configuration, which is suitable for miniature array systems. Disadvantages are its long complex structure and its empirical design.

3.3.3. Microstrip End Launcher (MEL). The MEL uses a loop antenna (launcher) to excite the waveguide (Fig. 18). A reflector is needed to reflect the excited waves into the desired direction.

In Ref. 20 a theoretical model has been derived to calculate the input impedance of the MEL. The model is based on an assumed current distribution in the launcher, and a variational expression for the input impedance. The experimental results of the MEL show a -20 dB bandwidth of 10% in the Ka band. The advantage of the MEL is its simple longitudinal configuration, which is suitable for miniature arrays. The disadvantage is that the model requires a rather narrow current strip (0.185 mm in the Ka band) of the launcher, and as a result only thin microstrip substrates can be used.

3.3.4. Ridged Waveguide Transition. Figure 19 shows the configuration of the ridged waveguide transition where a tapered or stepped ridge in a waveguide is used to convert the dominant TE_{10} mode of the waveguide into the TEM mode of the microstrip line.

In Ref. 21 the experimental result of such a transition was presented and it shows -20-dB difference over 25% bandwidth in the Ka band. Because of the complex structure of this transition, the final design was found

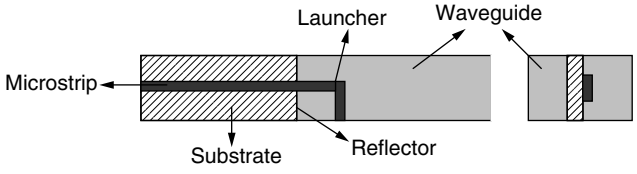


Figure 18. The microstrip end launcher.

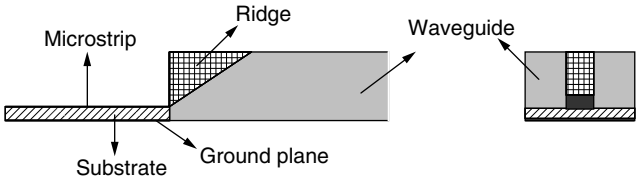


Figure 19. The ridged waveguide transition.

empirically. A possibility to analyze this transition is to use the T-matrix concept of a uniform ridged waveguide section. The advantages and disadvantages of the ridged waveguide transition are similar to those of the finline transition.

3.3.5. Radiating Slot Transition. The radiating slot transition is shown in Fig. 20. A slot in the ground plane of the microstrip is used to excite the waveguide.

In Ref. 22 a theoretical model of the input impedance of the radiating slot transition is given. The model is based on an assumed E-field distribution in the slot and charge distribution on the microstrip. The input impedance is calculated by using the complex power flow through the slot, and the modal voltage discontinuity in the microstrip. Unfortunately, the author does not give experimental results to verify the mathematical models. Nevertheless, simulation results show -20 dB bandwidth over 2% in the X band. Advantages of the transition are its simple structure and the use of the stub as a matching network. Disadvantages are its narrow bandwidth and perpendicular configuration, which is less suitable for miniature arrays.

3.3.6. Tapered Microstrip Transition. Figure 21 shows two views of the tapered microstrip transition, where the tapered-microstrip conductor and the ground plane are connected with the upper and lower waveguide walls, respectively. The waveguide is excited via a slot between the microstrip and the waveguide. Figure 22 shows the field patterns of a microstrip and a waveguide [23]. The microstrip has an E-field distribution that is almost

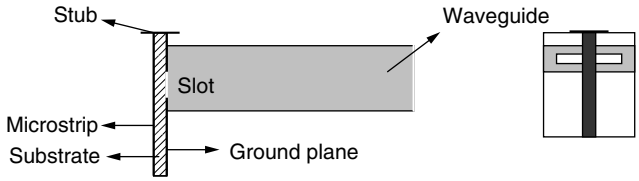


Figure 20. Radiating slot transition.

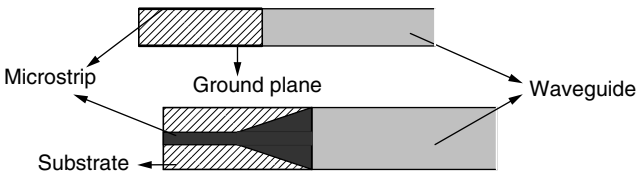


Figure 21. The tapered microstrip transition.

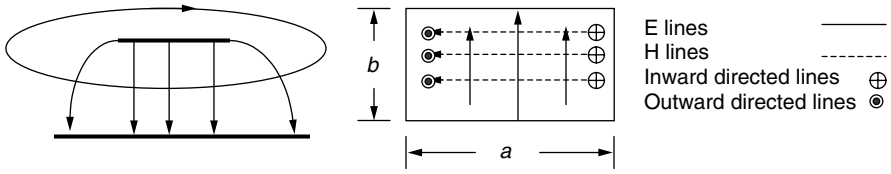


Figure 22. Field patterns of waveguide (at $t = T/4$) and microstrip [23].

uniform between the strip conductor and the ground plane, while the waveguide has a cosinusoidal E-field distribution. In addition, the H-field of the microstrip circle around the strip conductor, while the H-field lines of the waveguide circle around its E-field. It is obvious that a direct transition between these transmission lines will cause severe reflections. Another problem concerning this transition is that the waveguide height must be about the same as the microstrip height (usually less than 1 mm). An advantage of this transition is its longitudinal structure.

3.3.7. Analysis of the MEL. In this section the microstrip end launcher of Fig. 23 is analyzed. It shows a dielectric filled waveguide (DFW) transition where a printed circuit board is placed inside the waveguide. In order to avoid discontinuity effects such as LSM and LSE modes, the waveguide is filled with a dielectric material constant ε_r, which is the same as the dielectric constant of the DFW and the substrate of the microstrip line.

The current loop is divided into two different sections: the z-directed current section, which extends from the plane $z = 0$ to $z = z_1$; and the x-directed section, from $x = 0$ to $x = x_1$. The current is assumed to be continuous at the connecting point $x = x_1$ and $z = z_1$. The perfect ground planes, which are formed by the waveguide walls, are located at $x = 0$ and $x = a$, $y = 0$ and $y = b$, and $z = 0$ (the reflector). The current strip in the plane $y = y_1$ is assumed to be infinitely thin. The width $2w$ is sufficiently narrow so that the current distribution does not vary considerably in the transverse direction. In addition, for simplicity of the analysis, the effects due to the aperture in the reflector are neglected. The efficiency of the transition is characterized by the analysis of the input reflection coefficients.

3.3.8. Reflection Coefficient. The input reflection is given by

$$\Gamma_{\text{in}} = \frac{Z_{\text{in}} - Z_0}{Z_{\text{in}} + Z_0} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \tag{37}$$

where Z_0 and Z_{in} are the characteristic and input impedance of the microstrip line and the transition, respectively and Γ_L is the reflection coefficient of the load. When the load is not matched, the input reflection can be calculated by using the scattering coefficients $(S_{11}, S_{21}, S_{12}, S_{22})$ of the transition.

3.3.9. Input Impedance. The input impedance seen by the microstrip line satisfies the expression

$$Z_{\text{in}} = -\int_v \frac{E_z \cdot J_z}{I_{\text{in}}^2}\, dV - \int_v \frac{E_x \cdot J_x}{I_{\text{in}}^2}\, dV \tag{38}$$

where E_x and E_z are the electric fields inside the waveguide due to the current density components J_x and J_z, respectively. The current distribution is described as

$$\begin{aligned} J_z &= I_0 \cos[k(z_1 + x_1 - z)]\delta(y - y_1) \\ J_x &= I_0 \cos(kx_1)\delta(y - y_1) \end{aligned} \tag{39}$$

where I_0 is the amplitude of the input current. The current densities J_x and J_z are valid for the region $0 \le z \le z_1$ and $(x_1 - w) \le x \le (x_1 + w)$, and the region $0 \le x \le x_1$ and

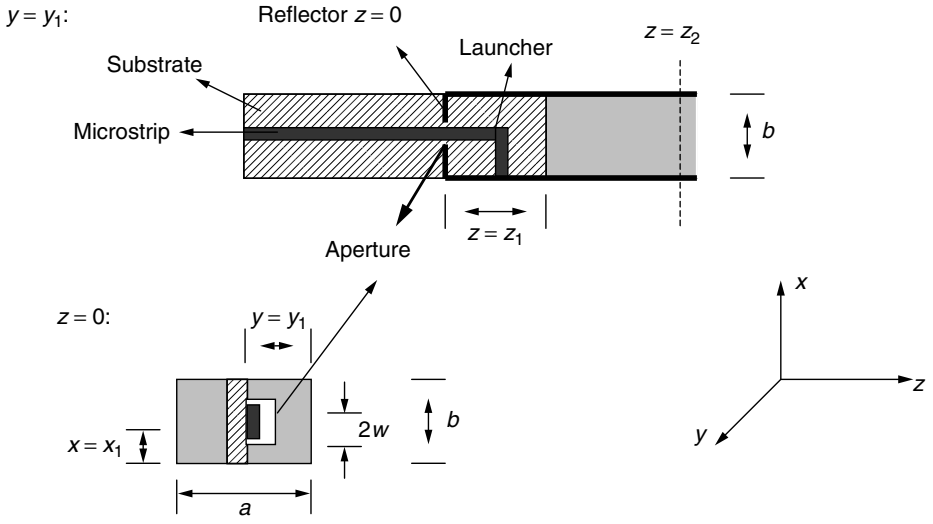


Figure 23. Microstrip line end launcher in a DFW.

$(z_1 - w) \le z \le (z_1 + w)$. The total input current I_{in} at the reference plane $z = 0$ becomes

$$I_{\text{in}} = 2wI_0 \cos[k(z_1 + x_1)] \tag{40}$$

where w is the strip width. Since there is a reflector at the plane $z = 0$, the excited field is caused by the current distribution given by Eq. (35) and by its image

$$\begin{aligned} J_z &= I_0 \cos[k(z_1 + x_1 + z)]\delta(y - y_1) \\ J_x &= -I_0 \cos(kx_1)\delta(y - y_1) \end{aligned} \tag{41}$$

which is valid for the regions $-z_1 \le z \le 0$ and $(x_1 - w) \le x \le (x_1 + w)$, and the region $0 \le x \le x_1$ and $-(z_{1+}w) \le z \le (-z_1 + w)$, respectively. Figure 24 shows the current distribution, its image and the integration domains V and V'.

Figure 25 shows the steps necessary to calculate the input impedance for further analysis. The electric fields are related to the magnetic potentials via [24]

$$\begin{aligned} E_z &= \frac{1}{j\omega\varepsilon}\left(\frac{\partial A_z}{\partial z^2} + k^2 A_z\right) \\ E_x &= \frac{1}{j\omega\varepsilon}\left(\frac{\partial A_x}{\partial x^2} + k^2 A_x\right) \end{aligned} \tag{42}$$

The magnetic vector potentials are defined as

$$\begin{aligned} A_x &= \int_{v'} G_{xx}\left(\frac{x,y,z}{x',y',z'}\right) \cdot J_x(x',y',z')\, dV' \\ A_z &= \int_{v'} G_{zz}\left(\frac{x,y,z}{x',y',z'}\right) \cdot J_z(x',y',z')\, dV' \end{aligned} \tag{43}$$

The primed coordinates x', y', z' represent the source point, while the unprimed coordinates represent the field points.

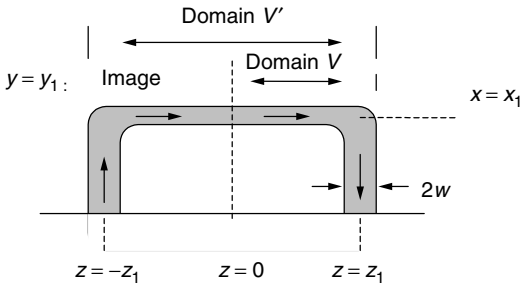


Figure 24. Current distribution and its image of the microstrip end launcher.

The Green function is given by [20]

$$\begin{aligned} G_{xx}\left(\frac{x,y,z}{x',y',z'}\right) &= \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \frac{\delta_n}{ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \cos\left(\frac{n\pi x'}{b}\right)\sin\left(\frac{n\pi y'}{a}\right) e^{-\gamma_{mn}|z-z'|} \\ G_{zz}\left(\frac{x,y,z}{x',y',z'}\right) &= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \frac{1}{2ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \cos\left(\frac{n\pi x'}{b}\right)\sin\left(\frac{n\pi y'}{a}\right) e^{-\gamma_{mn}|z-z'|} \end{aligned} \tag{44}$$

where

$$\gamma_{mn} = \begin{cases} \sqrt{k^2 - \left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right]} & \text{if } \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 \le k^2 \\ j\sqrt{\left[\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2\right] - k^2} & \text{otherwise} \end{cases} \tag{45}$$

Substituting (40) in (39), using (37) and performing the integration over V' leads to the following expressions:

$$\begin{aligned} A_x &= \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \frac{I_0\delta_n}{ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \left[\begin{array}{l} \displaystyle\int_0^a\int_0^b \cos\left(\frac{n\pi x'}{b}\right)\cos(kx')\sin\left(\frac{n\pi y'}{a}\right) \\ \qquad \times \delta(y' - y_1)\, dx'\, dy' \\ \displaystyle\left(-\int_{-z_1}^{0} e^{-\gamma_{mn}|z-z'|} + \int_0^{z_1} e^{-\gamma_{mn}|z-z'|}\right) dz' \end{array}\right] \\ &= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \frac{I_0}{2ab\gamma_{mn}} \cos\left(\frac{n\pi x}{b}\right)\sin\left(\frac{m\pi y}{a}\right) \\ &\quad \times \left[\begin{array}{l} \displaystyle\int_0^a\int_0^b \cos\left(\frac{n\pi x'}{b}\right)\cos(kx')\sin\left(\frac{n\pi y'}{a}\right) \\ \qquad \times \delta(y' - y_1)\, dx'\, dy' \\ \displaystyle\left(-\int_{-z_1}^{0} e^{-\gamma_{mn}|z-z'|}\cos[k(z_1 + x_1 + z')] \right. \\ \displaystyle\left. + \int_0^{z_1} e^{-\gamma_{mn}|z-z'|}\cos[k(z_1 + x_1 + z')]\right) dz' \end{array}\right] \end{aligned} \tag{46}$$

The integration is performed in the paper by Ho and Shin [20] and is not included here. Inserting the results in (38) leads to the expressions for E_x and E_z. Substituting

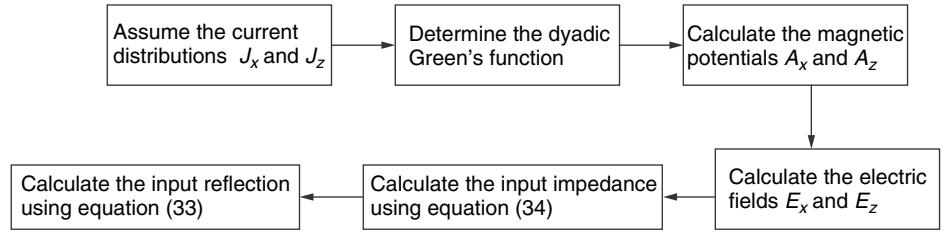


Figure 25. Steps for calculating the input reflection.

E_x and E_z in (34) gives the result for the Z_{in} [20]. It can be shown that

$$
\begin{aligned}
-\int_v \frac{E_z \cdot J_z}{I_{\text{in}}^2}\,dV &= j\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \frac{1}{\omega\varepsilon_0 ab\gamma_{mn}\cos^2[k(z_1+x_1)]}\frac{b}{n\pi w} \\
&\quad \times \sin\left(\frac{n\pi w}{b}\right)\sin^2\left(\frac{m\pi y_1}{a}\right)\sin^2\left(\frac{n\pi x_1}{b}\right) \\
&\quad \times \left\{ \begin{array}{l} \dfrac{k\sin(k(z_1+x_1))}{k^2+\gamma_{mn}^2}[\gamma_{mn}\cos^2(k(z_1+x_1)) \\ +k\sin(k(z_1+x_1)) - e^{-\gamma_{mn}z_1}(\gamma_{mn}\cos(kx_1) \\ +k\sin(kx_1))] \\ -\left(\begin{array}{c} \dfrac{e^{-\gamma_{mn}z_1}(\gamma_{mn}\cos(kx_1)+k\sin(kx_1))}{k^2+\gamma_{mn}^2} \\ \times k\sin(k(z_1+x_1)) \\ +\gamma_{mn}\cos(kx_1)\sin h(\gamma_{mn}z_1) \\ -k\sin(kx_1)\cos h(\gamma_{mn}z_1) \end{array} \right) \end{array} \right\} \\
-\int_v \frac{E_x \cdot J_x}{I_{\text{in}}^2}\,dV &= j\sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \frac{240\delta_n}{abk\gamma_{mn}}\frac{\sin h(\gamma_{mn}w)}{\gamma_{mn}w} \\
&\quad \times \sin^2\left(\frac{m\pi y_1}{a}\right)\sin h(\gamma_{mn}z_1) \\
&\quad \times e^{-\gamma_{mn}z_1}\frac{\left[\sin(kx_1)\cos\left(\dfrac{n\pi x_1}{b}\right) - \dfrac{n\pi}{ak}\cos(kx_1)\sin\left(\dfrac{n\pi x_1}{b}\right)\right]^2}{\cos^2(k(z_1+x_1))\left(1-\left(\dfrac{n\pi}{ak}\right)^2\right)}
\end{aligned}
\tag{47}
$$

The microstrip end launcher may be used to excite a DFW with two E-plane steps. This kind of waveguide will be discussed in Section 6. The waveguide with E-plane steps (Fig. 26) is filled with a dielectric material with a dielectric constant $\varepsilon_r = 2.53$. The airgap matching network is discussed in Section 5.

The effect of different parameters such as x_1, y_1, and z_1 (see Fig. 23) on the behavior of the input impedance in the X and Ka bands has been studied by Ho and Shih [20] and Lam et al. [25], respectively. An optimization routine has been developed in order to obtain the optimal values for the different parameters. Since the purpose is to realize a miniature antenna with a large bandwidth, a two $\lambda/4$ matching network is employed. Figure 27 illustrates this concept. The design parameters x_1, y_1, z_1, Z_{0k} and l are optimized. Several parameters were kept constant throughout the analysis, such as the physical dimensions of the width w and the height h of the microstrip, and the width of the DFW with steps. The values of w, a and b_3 were chosen to be 0.185, 14, and 5 mm, respectively. A substrate 0.25 mm thick with a dielectric constant of 2.53 is used as a dielectric slab. The end launcher is matched to a microstrip line with a characteristic impedance of 75 Ω.

Figure 28 shows the optimized calculated input reflection with and without a $\lambda/4$ section. It is observed that the $\lambda/4$ section has a remarkable effect on the bandwidth. The −20 dB bandwidth increases by almost 15%. The difference is due to the fact that the $\lambda/4$ section decreases the frequency sensitivity of the multiple reflections [23].

3.3.10. Design of the Microstrip End Launcher. Figure 29 shows the geometry of the end launcher with two $\lambda/4$ sections and its integration into the DFW. The thin substrate consists of the microstrip line with z_{01}. Two $\lambda/4$ sections, z_{0j} and z_{0k}, and the current loop launcher were inserted into the DFW.

In order to obtain a sufficiently narrow loop microstrip, a thin microstrip substrate with a dielectric constant of $\varepsilon_r = 2.53$ and $h = 0.25$ mm is chosen. This thickness was chosen because the dielectric constant of the substrate is not exactly the same as the dielectric constant of DFW. Furthermore, a thin substrate would cause fewer LSM and LSE mode effects.

The $\lambda/4$ section with z_{0j} is chosen to have the same width as the end launcher loop. The width of the $\lambda/4$ section with z_{0k} has been optimized, $w_1 = 0.32$ mm. The

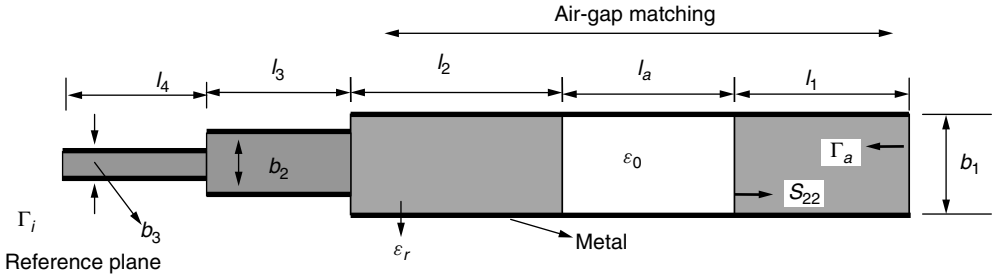


Figure 26. Two-plane step DFW.

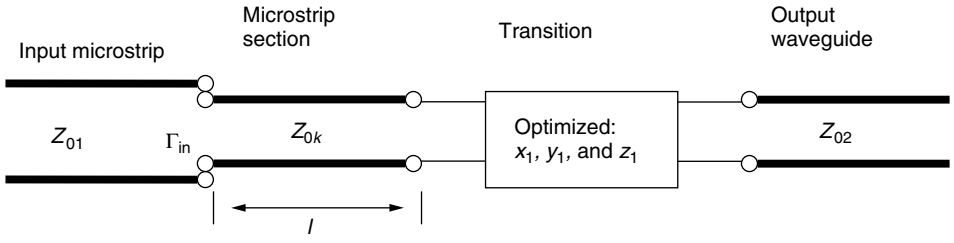


Figure 27. Optimization of matching network.

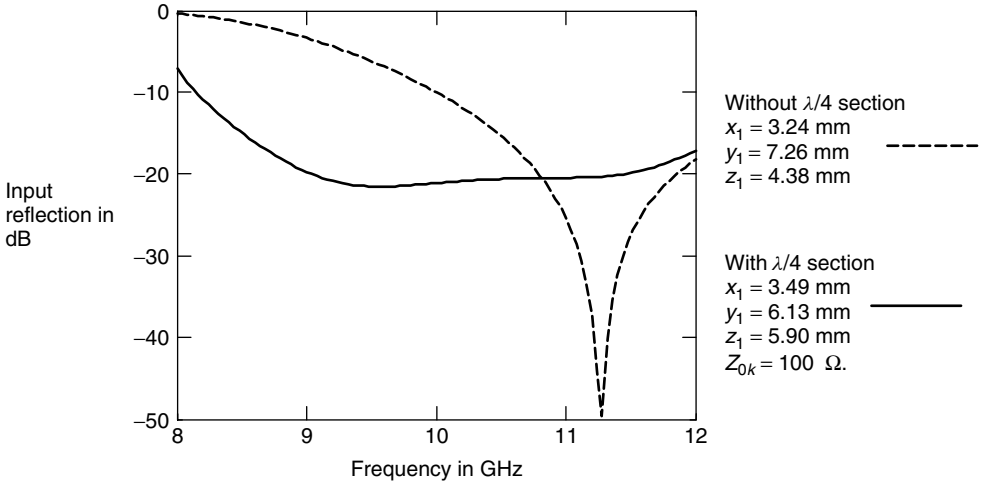


Figure 28. Calculated input reflection of the microstrip end launcher with and without $\lambda/4$ section.

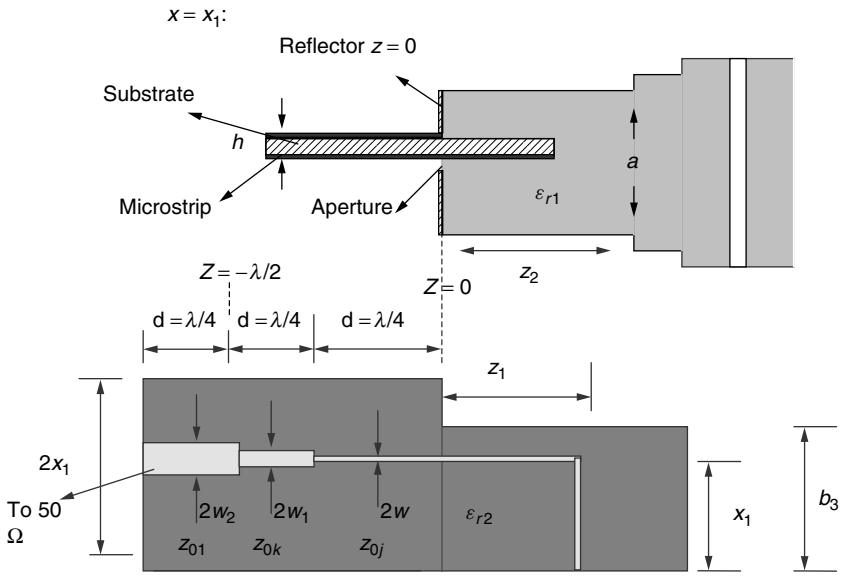


Figure 29. Microstrip end launcher transition at 10 GHz and its integration to DFW. (Courtesy of IRCTR.)

corresponding microstrip width is $w_2 = 0.37$ mm at the design frequency of 10 GHz. The characteristic impedance of the input microstrip line Z_{01} is chosen to be 75 Ω and has the same length as the $\lambda/4$ section. The current in the loop is assumed to be continuous. Therefore, a bend having a radius of $r = 4w$ is used to minimize the reflections [23] (not shown in Fig. 29).

Figure 30 shows the optimized design parameters corresponding to those in Fig. 29. Figure 31 shows the calculated input reflection as a function of frequency at different reference planes of the microstrip end launcher. The reference planes are $z = z_2$, $z = 0$, and $z = -\lambda/2$ (see Fig. 29). These planes indicate the input reflection of the DFW [26], the transition of the microstrip end launcher with DFW, and the total input reflection of the microstrip end launcher with the DFW and the two $\lambda/4$ sections.

The input reflection of the DFW has a −20 dB bandwidth over a 15% frequency band. There are two dips at 9.1 and 10.3 GHz. The input reflection of MEL with

a	14 mm		
b_3	5 mm		
$\varepsilon_{r1} = \varepsilon_{r1}$	2.33		
h	0.25 mm		
x_1	3.9 mm		
y_1	9.48 mm		
z_1	6.06 mm		
z_2	50 mm		
Z_{oj}	75 ohm	w	0.185 mm
Z_{0k}	55 ohm	w_1	0.32 mm
Z_{01}	50 ohm	w_2	0.37 mm

Figure 30. Optimal design parameters.

DFW has a −20 dB bandwidth over 5% at the frequency band. In addition, the dips at 9.1, 9.8, and 10.6 GHz can be distinguished.

A piece of the dielectric material inside the waveguide was removed in order to place the launcher section

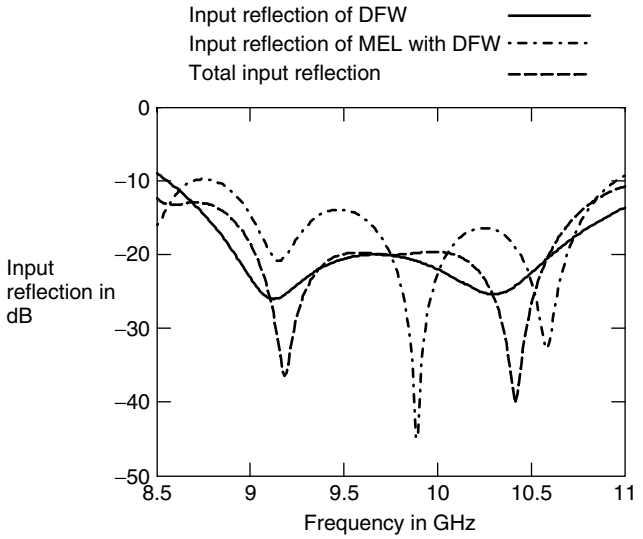


Figure 31. Input reflections as function of frequency at different reference planes of MEL.

of the microstrip circuit inside the waveguide. Then, an electrical connection between the launcher and the lower waveguide wall was made. A reflector was placed to close the waveguide, and practically all waves are reflected in the desired direction. To prevent a short circuit, a small aperture was made in the reflector wall. Since the theoretical model does not take the aperture effect into account, the dimensions of the aperture were chosen experimentally. A conducting post was used for the electrical connection between the end launcher and the waveguide wall [25].

The coax-to-microstrip transition has been realized empirically by tapering the pin of the SMA connector. The half-circle of the mounting plate with radius D is used to compensate the reactance of the coax-to-microstrip transition empirically [25,27]. Measurement results can be seen in Fig. 32.

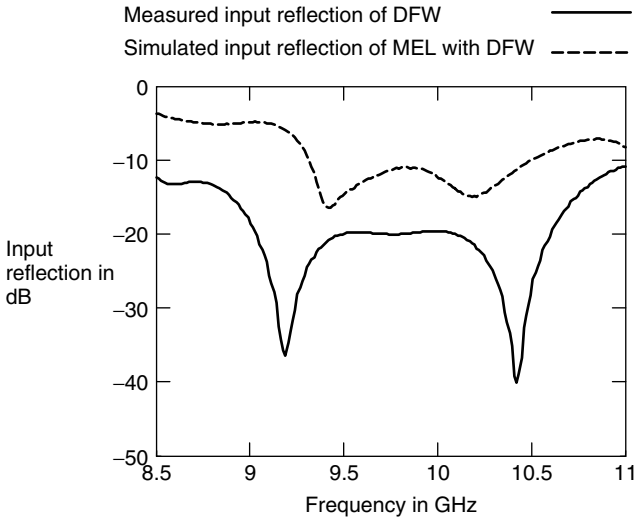


Figure 32. Measured and simulated results of input reflection of MEL as a function of frequency.

4. ATTENUATION

Although the waveguide has low losses, the electromagnetic waves still suffer from attenuation. Since the waveguide metal is not a perfect conductor, there are some ohmic (conduction) losses. If the waveguide is filled with dielectric material, an extra factor, called the *dielectric losses*, which contributes to the total losses, needs to be taken into the consideration. They are denoted with α_c and α_d, respectively, and are given by the following expression [11]

$$\alpha_c = \frac{2R_s}{\delta_m \delta_n b \eta \sqrt{1-\left(\frac{f_c}{f}\right)^2}} \left\{ \left(\delta_m + \delta_n \frac{b}{a}\right)\left(\frac{f_c}{f}\right)^2 + \frac{b}{a}\left[1-\left(\frac{f_c}{f}\right)^2\right] \frac{m^2 ab + (na)^2}{(ma)^2 + (na)^2} \right\} \tag{48}$$

where

$$R_s = \sqrt{\frac{\omega\mu}{2\sigma}} \quad \text{for } \sigma \gg \omega\varepsilon$$
$$\delta_m = \begin{cases} 2 & m = 0 \\ 1 & m \neq 0 \end{cases} \tag{49}$$

and

$$\alpha_d = 8.68 \left(\frac{\varepsilon_e''}{\varepsilon_e'}\right) \frac{\pi}{\lambda} \left(\frac{\lambda_g}{\lambda}\right) \quad \text{dB/m} \tag{50}$$

Figure 33 shows the TE_{10} conduction losses as a function of frequency for three different dielectric materials.

5. MINIATURIZATION TECHNIQUE

Large array antennas can benefit significantly from miniaturization. Since the propagation of the fundamental mode in the rectangular waveguide is independent of the height, the miniaturization can be achieved by lowering the height. Filling the waveguide with dielectric material can also contribute to miniaturization. When a dielectric

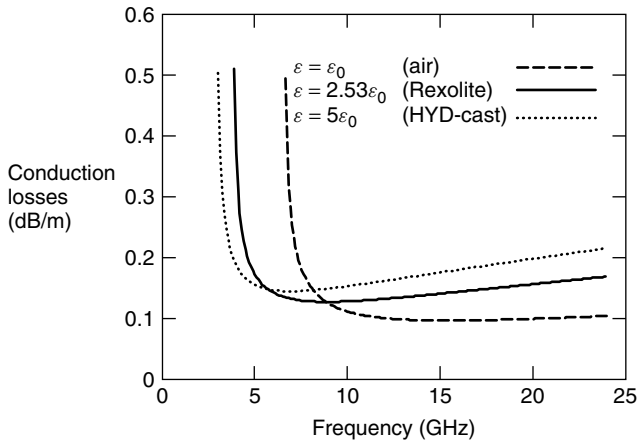


Figure 33. The conduction losses as function of frequency for different materials. (*Source*: C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, 1989.)

is used, the dimensions of the antenna will decrease by the square root of the dielectric constant.

This dielectric loading technique is also applied to protect the waveguide from environmental conditions and makes it possible to flesh mount the antenna on the surface of the spacecraft or aircraft. Figure 34 shows a number of dual-polarized DFW in different frequency bands.

Using this technique for miniaturization may in some cases lead to a high aperture reflection. This leads to a complexity in aperture matching. In this section the aperture matching technique is presented.

5.1. Aperture Characteristics

In order to derive an expression for the aperture admittance, the waveguide geometry in Fig. 35 is considered. The approach is based on power conservation across the aperture from region I to region II. It is assumed that the aperture is mounted in a perfectly conducting plane of infinite size. It is also assumed that the excitation is such that only symmetrical TE_{m0} modes are generated in region I (in Fig. 35).

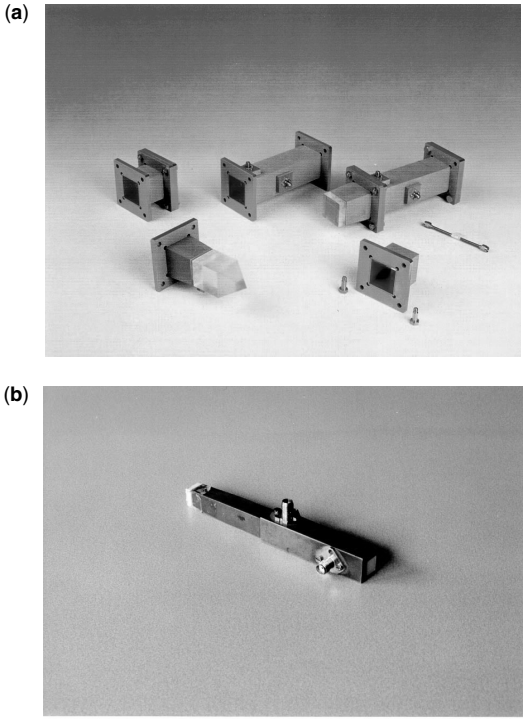


Figure 34. Miniaturized dual-polarized DFW: (**a**) S band, $\varepsilon_r = 2.53$; (**b**) X band, $\varepsilon_r = 5.0$. (Courtesy of IRCTR.)

5.1.1. Internal Field in Region I. When only the TE_{m0} mode (may) exist in the waveguide, the field components in region I (Fig. 35) can be written as [28]

$$\begin{aligned} E_y^{\mathrm{I}} &= E_0(e^{-j\beta_{z1}z} + \Gamma e^{j\beta_{z1}z})\cos\left(\frac{\pi x}{a}\right) \\ &\quad + \sum_{m=3,5,\ldots}^{\infty} A_m \cos\left(\frac{m\pi x}{a}\right) e^{j\beta_{zm}z} \\ E_x^{\mathrm{I}} &= H_y^{\mathrm{I}} = 0 \\ H_x^{\mathrm{I}} &= Y_{10}E_0(e^{-j\beta_{z1}z} - \Gamma e^{j\beta_{z1}z})\cos\left(\frac{\pi x}{a}\right) \\ &\quad - \sum_{m=3,5,\ldots}^{\infty} Y_{m0}A_m \cos\left(\frac{m\pi x}{a}\right) e^{j\beta_{zm}z} \end{aligned} \tag{51}$$

where Y_0 is the free-space admittance and Y_{m0} the wave impedance for the TE_{m0} modes. In this section the following changes in parameters are introduced:

$$\beta_{z1} = \beta\frac{Y_{10}}{Y_0}, \beta_{zm} = \beta\frac{Y_{m0}}{Y_0}, \beta = \beta\sqrt{\varepsilon_r} \tag{52}$$

Note that in this case the wavenumber is higher than the free-space wavenumber. At the boundary $z = 0$ and with $A_m = D_m E_0(1 + \Gamma)$, Eq. (47) becomes

$$\begin{aligned} E_y^{\mathrm{I}}(x, y, z = 0) &= E_0(1 + \Gamma)\left(\cos\left(\frac{\pi x}{a}\right)\right. \\ &\quad \left. + \sum_{m=3,5,\ldots}^{\infty} D_m \cos\left(\frac{m\pi x}{a}\right)\right) \\ E_x^{\mathrm{I}}(x, y, z = 0) &= H_y^{\mathrm{I}}(x, y, z = 0) = 0 \\ H_x^{\mathrm{I}}(x, y, z = 0) &= Y_{10}E_0(1 - \Gamma)\cos\left(\frac{\pi x}{a}\right) \\ &\quad - \sum_{m=3,5,\ldots}^{\infty} Y_{m0}D_m E_0(1 + \Gamma)\cos\left(\frac{m\pi x}{a}\right) \end{aligned} \tag{53}$$

The expression for the unknown coefficients D_m and the normalized aperture admittance will now be derived.

5.1.2. Aperture Admittance. The reaction integral I for the aperture fields are considered to compute the energy transfer through the aperture. The integral for region I becomes

$$I = \int_{-\frac{a}{2}}^{\frac{a}{2}} \int_{-\frac{b}{2}}^{\frac{b}{2}} E_y^{\mathrm{I}}(x, y, 0) H_x^{\mathrm{I}}(x, y, 0)\, dy\, dx \tag{54}$$

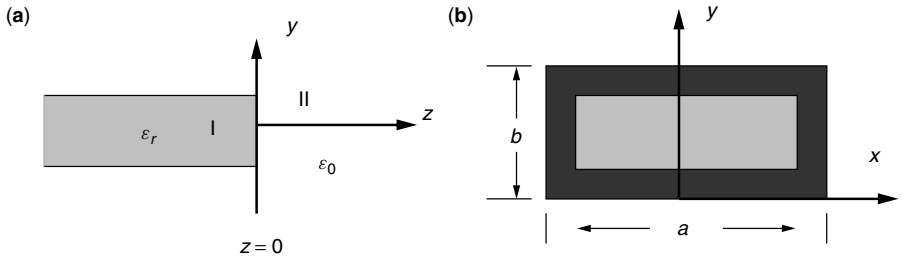


Figure 35. Configuration of the waveguide for analysis of aperture admittance: (**a**) side view; (**b**) front view [28].

Substituting Eq. (49) in (50) and performing the integration leads to

$$I = \frac{ab}{2} Y_{10} E_0^2 (1+\Gamma)^2 \left[\frac{1-\Gamma}{1+\Gamma} - \sum_{m=3,5,\ldots}^{\infty} \frac{Y_{m0}}{Y_{10}} D_m^2 \right] \tag{55}$$

Rearranging Eq. (51) gives the following expression for the normalized aperture admittance:

$$y_{\text{ap}} = \frac{1-\Gamma}{1+\Gamma} = \frac{2}{ab} \frac{I}{Y_{10} E_0^2 (1+\Gamma)^2} + \sum_{m=3,5,\ldots}^{\infty} \frac{Y_{m0}}{Y_{10}} D_m^2 \tag{56}$$

In order to calculate the integral I, it is assumed that the tangential fields are continuous across the aperture. The following relationship should exist:

$$\begin{aligned} I &= \int_{-(a/2)}^{a/2} \int_{-(b/2)}^{b/2} E_y^I(x,y,0) H_x^{\text{I}}(x,y,0)\, dy\, dx \\ &= \int_{-(a/2)}^{a/2} \int_{-(b/2)}^{b/2} E_y^{\text{II}}(x,y,0) H_x^{\text{II}}(x,y,0)\, dy\, dx \end{aligned} \tag{57}$$

In region II the aperture fields can be expressed in the spectral domain via

$$\hat{E}_y(k_x,k_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E_y(x,y,0) e^{j\{k_x x + k_y y\}}\, dy\, dx \tag{58}$$

where $\hat{E}_y(k_x,k_y)$ is the Fourier transformation of the aperture electric field and k_x, k_y are the spectral frequencies that extend over the entire frequency spectrum $-\infty \le k_x, k_y \le \infty$. Since the tangential fields are zero outside the aperture and are even functions with respect to x and y, Parseval's theorem can be used to obtain the following relationship [29]:

$$\begin{aligned} I &= \int_{-(a/2)}^{a/2} \int_{-(b/2)}^{b/2} E_y^{\text{II}}(x,y,0) H_x^{\text{II}}(x,y,0)\, dy\, dx \\ &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{E}_y^{\text{II}}(k_x,k_y,0) \hat{H}_y^{II}(k_x,k_y,0)\, dk_y\, dk_x \end{aligned} \tag{59}$$

Substituting (55) in (52) relates the normalized aperture admittance to the exterior fields in the spectral domain. The result becomes

$$\begin{aligned} y_{\text{ap}} = \frac{1-\Gamma}{1+\Gamma} &= \frac{2}{abY_{10}E_0^2(1+\Gamma)^2} \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{E}_y^{\text{II}}(k_x,k_y,0) \\ &\quad \times \hat{H}_y^{\text{II}}(k_x,k_y,0)\, dk_y\, dk_x + \sum_{m=3,5,\ldots}^{\infty} \frac{Y_{m0}}{Y_{10}} D_m^2 \end{aligned} \tag{60}$$

The next step is to solve the fields in the exterior region.

5.1.3. Exterior Field in Region II. The expressions for the exterior field are derived using the electric and magnetic potentials. In terms of the potentials, the electromagnetic fields are given as [28]

$$\begin{aligned} E &= -\nabla \times F - j\omega A + \frac{\nabla\nabla \cdot A}{j\omega\varepsilon} \\ H &= \nabla \times A - j\omega F + \frac{\nabla\nabla \cdot F}{j\omega\mu} \end{aligned} \tag{61}$$

where $\nabla \cdot A$ is the divergence of A and ∇ is the gradient operator [12]. The field components in region II can be expressed as [29]

$$\begin{aligned} E_y^{\text{II}} &= -\frac{\partial \psi}{\partial x} + \frac{1}{j\omega\varepsilon} \frac{\partial^2 \varphi}{\partial y\, \partial z} \\ E_x^{\text{II}} &= \frac{\partial \psi}{\partial y} + \frac{1}{j\omega\varepsilon} \frac{\partial^2 \varphi}{\partial x\, \partial z} \\ H_y^{\text{II}} &= \frac{1}{j\omega\mu} \frac{\partial^2 \psi}{\partial y\, \partial z} + \frac{\partial \varphi}{\partial x} \\ H_x^{\text{II}} &= \frac{1}{j\omega\mu} \frac{\partial^2 \psi}{\partial x\, \partial z} - \frac{\partial \varphi}{\partial y} \end{aligned} \tag{62}$$

where $F = \psi\hat{z}$, $A = \varphi\hat{z}$. A possible solution of Eq. (58) in the spectral domain is given by

$$\begin{aligned} \psi &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(k_x,k_y) e^{-j(k_x x + k_y y + k_z z)}\, dk_y\, dk_x \\ \varphi &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(k_x,k_y) e^{-j(k_x x + k_y y + k_z z)}\, dk_y\, dk_x \end{aligned} \tag{63}$$

The modal coefficients f and g are derived as follows. The Fourier transform of the fields in region II is given as

$$\begin{aligned} [E^{\text{II}}(x,y), H^{\text{II}}(x,y)] = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} &[\hat{E}^{\text{II}}(k_x,k_y), \\ &\times \hat{H}^{\text{II}}(k_x,k_y)] e^{-j\{k_x x + k_y y\}}\, dk_y\, dk_x \end{aligned} \tag{64}$$

Substituting Eqs. (59) and (60) into (58) leads to

$$\begin{aligned} \hat{E}_y^{\text{II}}(k_x,k_y) &= j\left(k_x f + \frac{k_x k_z}{\omega\varepsilon} g \right) e^{-jk_z z} \\ \hat{E}_x^{\text{II}}(k_x,k_y) &= -j\left(k_y f - \frac{k_x k_z}{\omega\varepsilon} g \right) e^{-jk_z z} \end{aligned} \tag{65}$$

The tangential electric field components are continuous across the aperture at $z = 0$:

$$\begin{aligned} \hat{E}_y^{\text{II}}(k_x,k_y,0) &= \hat{E}_y^{\text{I}}(k_x,k_y,0) \\ \hat{E}_x^{\text{II}}(k_x,k_y,0) &= \hat{E}_x^{\text{I}}(k_x,k_y,0) \end{aligned} \tag{66}$$

The modal coefficients f and g can be expressed in terms of the electric field components in region I:

$$\begin{aligned} f(k_x,k_y) &= j \frac{k_y \hat{E}_x^{\text{I}}(k_x,k_y,0) - k_x \hat{E}_y^{\text{I}}(k_x,k_y,0)}{k_y^2 + k_x^2} \\ g(k_x,k_y) &= -j \frac{(k_y \hat{E}_x^{\text{I}}(k_x,k_y,0) + k_x \hat{E}_y^{\text{I}}(k_x,k_y,0))\omega\varepsilon}{k_z(k_y^2 + k_x^2)} \end{aligned} \tag{67}$$

From Eqs. (58), (60), and (63) it can be deduced that

$$\hat{H}_x^{\text{II}}(k_x,k_y,0) = \frac{(k^2 - k_x^2)\hat{E}_x^{\text{I}}(k_x,k_y,0) + k_x k_y \hat{E}_y^{\text{I}}(k_x,k_y,0)}{\omega\mu k_z} \tag{68}$$

Inserting the Fourier transform of the electric field components from Eq. (49) yields

$$\begin{aligned}\hat{E}_x^{\mathrm{II}}(k_x, k_y, 0) &= 0\\ \hat{E}_y^{\mathrm{II}}(k_x, k_y, 0) &= \hat{E}_y^{\mathrm{I}}(k_x, k_y, 0)\\ &= E_0(1+\Gamma)\iint_{\mathrm{ap}}\left(\cos\left(\frac{\pi x}{a}\right)\right.\\ &\quad\left.+\sum_{m=3,5,\ldots}^{\infty} D_m\cos\left(\frac{m\pi x}{a}\right)\right)e^{+j\{k_x x+k_y y\}}\,dy\,dx\end{aligned} \tag{69}$$

Substituting $\hat{E}_y^{\mathrm{II}}(k_x, k_y, 0)\hat{H}_y^{\mathrm{II}}(k_x, k_y, 0) = \hat{E}_y^{\mathrm{I}}(k_x, k_y, 0)\hat{H}_y^{\mathrm{I}}(k_x, k_y, 0)$ in Eq. (57) gives the following scheme for the aperture admittance

$$\begin{aligned}y_{\mathrm{ap}} &= \frac{2}{abY_{10}\omega\mu}\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(k^2-k_x^2)}{k_z}C_0^2(k_x)\\ &\quad\times\left[C_1^2(k_x)+2C_1(k_x)\sum_{m=3,5,\ldots}^{\infty}D_mC_m(k_x)\right.\\ &\quad\left.+\left(\sum_{m=3,5,\ldots}^{\infty}D_mC_m(k_x)\right)^2\right]dk_y\,dk_x+\sum_{m=3,5,\ldots}^{\infty}\frac{Y_{m0}}{Y_{10}}D_m^2\end{aligned} \tag{70}$$

with

$$\begin{aligned}Y_{ij} = Y_{ji} &= \frac{2}{ab\omega\mu}\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(k^2-k_x^2)}{k_z}\\ &\quad\times C_0^2(k_y)C_j(k_x)C_j(k_x)\,dk_y\,dk_x\end{aligned} \tag{71}$$

for $i, j \neq 0$ the expression results in

$$y_{\mathrm{ap}} = \frac{Y_{11}}{Y_{10}} + 2\sum_{m=3,5,\ldots}D_m\frac{Y_{1m}}{Y_{10}} + \sum_{m=3,5,\ldots}D_m^2\left(\frac{Y_{mm}}{Y_{10}}+\frac{Y_{m0}}{Y_{10}}\right) \tag{72}$$

where

$$\begin{aligned}C_0(k_y) &= \frac{b\sin\left(k_y\frac{b}{2}\right)}{k_y\frac{b}{2}}\\ C_1(k_x) &= \frac{2\pi a\cos\left(k_x\frac{a}{2}\right)}{(\pi)^2-(k_xa)^2}\\ C_m(k_x) &= \frac{2m\pi aj^{m-1}\cos\left(k_x\frac{a}{2}\right)}{(m\pi)^2-(k_xa)^2}\end{aligned} \tag{73}$$

Figure 36 shows the normalized aperture impedance of a DFW as function of frequency in the L and X bands. The height of the waveguide is given as a parameter. The width

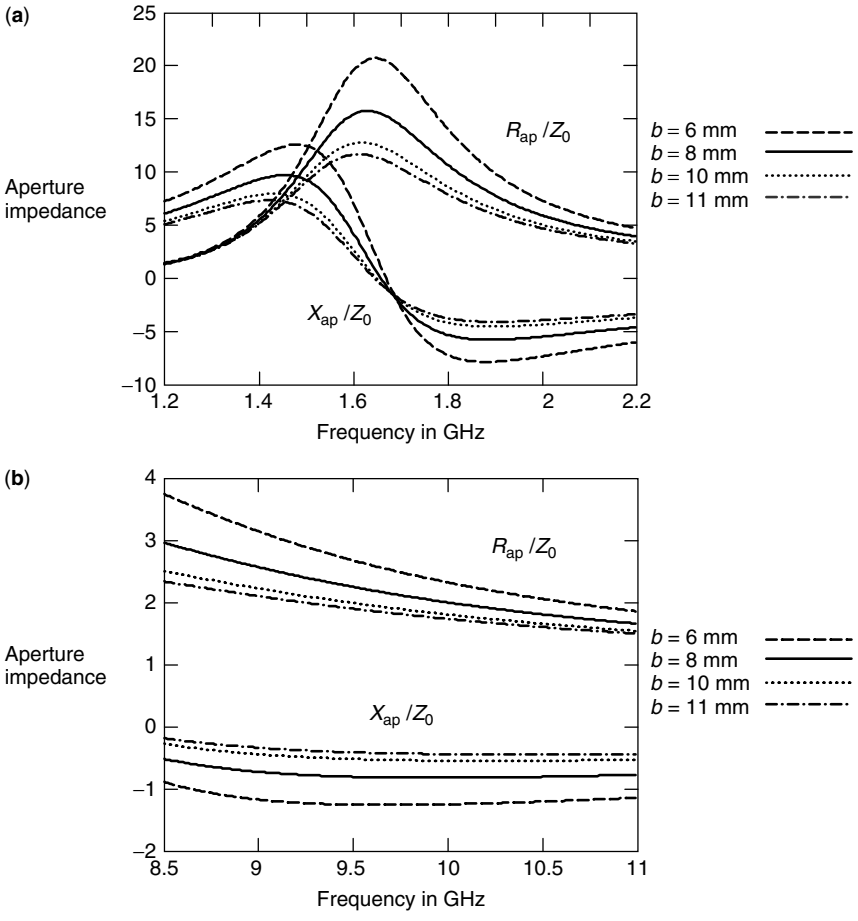


Figure 36. The normalized aperture impedance of DFW as a function of frequency: (**a**) L band; (**b**) X band, $\varepsilon_r = 2.53$.

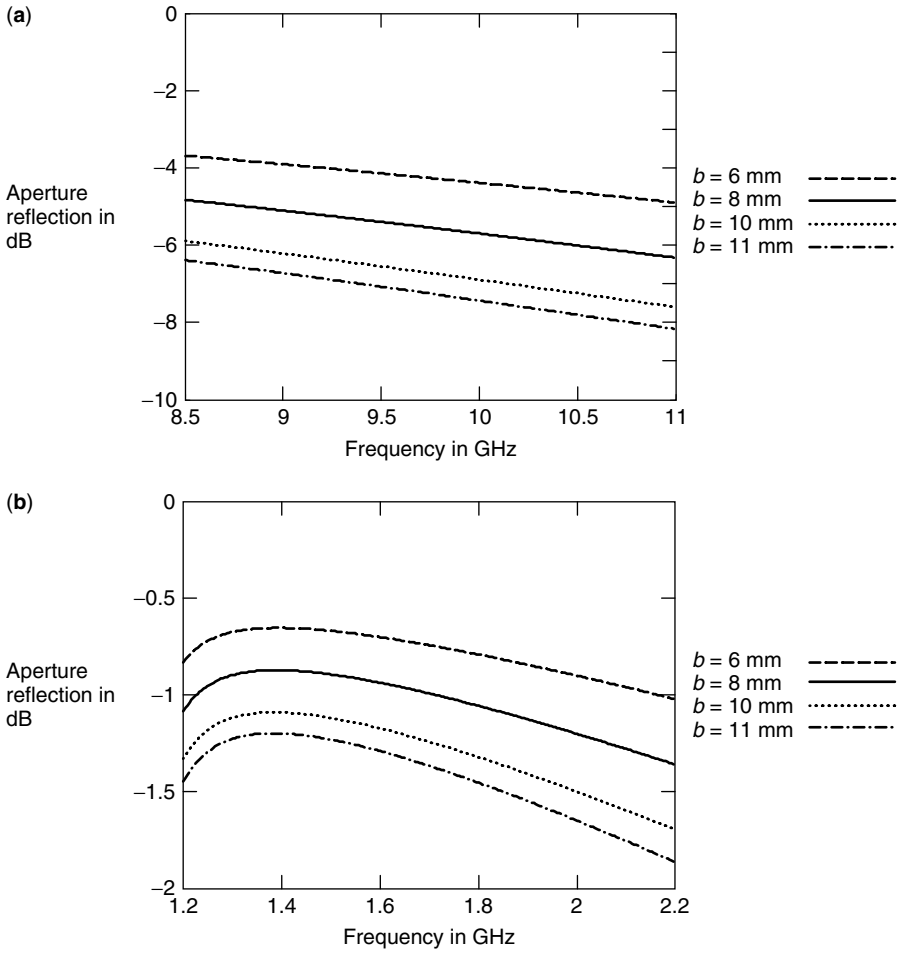


Figure 37. The aperture reflection as a function of frequency: (**a**) L band; (**b**) X band, $\varepsilon_r = 2.53$.

of the aperture is 83 and 17 mm in the L and X bands, respectively.

Figure 37 shows the aperture reflection for different heights of the aperture. The reflection losses increase with the decrease in the height of the aperture.

5.2. Matching Technique

The aperture reflection results given in the previous section indicate that the aperture reflection can be high. It is possible to match the aperture reflection using a microwave matching technique. In Ref. 30 a unique matching technique is proposed and analyzed. The aim of this section is to derive the mathematical expression for such a matching condition. For this aim the matching network is considered as a two-port device. This is shown in Fig. 38.

With the definition for the reflection coefficient

$$\Gamma_i \triangleq \frac{b_1}{a_1}, \Gamma_a \triangleq \frac{a_2}{b_2} \tag{74}$$

where an optimal matching network must satisfy the condition

$$\Gamma_i = 0 \tag{75}$$

The scattering matrix of the two-port network is given as

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \tag{76}$$

Substituting (70) in (72) leads to

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}\Gamma_a b_2 \\ b_2 &= S_{21}a_1 + S_{22}\Gamma_a b_2 \end{aligned} \tag{77}$$

Equation (73) leads to

$$\frac{b_1}{a_1} = \frac{S_{11} - (S_{11}S_{22} - S_{21}S_{12})\Gamma_a}{1 - S_{22}\Gamma_a} = \frac{S_{11} - (\det S)\Gamma_a}{1 - S_{22}\Gamma_a} \tag{78}$$

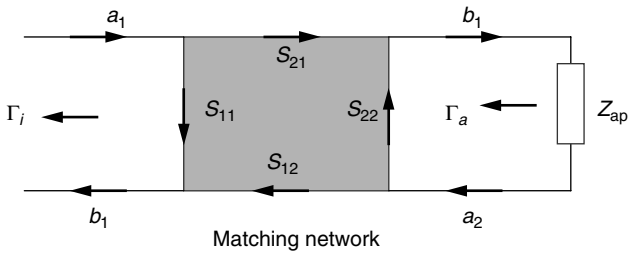


Figure 38. The two-port matching network and its scattering diagram [30].

Equation (74) can be rearranged into

$$\Gamma_i = \frac{b_1}{a_1} = \det S \frac{\dfrac{S_{11}}{\det S} - \Gamma_a}{1 - S_{22}\Gamma_a} \tag{79}$$

when considering

$$S_{22}^* \det S = S_{22}^*(S_{11}S_{22} - S_{21}S_{12}) \tag{80}$$

It has been shown [29] that a symmetric and lossless two-port device has the properties

$$\begin{aligned} S_{11}S_{22}^* &= -S_{12}S_{22}^* \\ |S_{11}| &= |S_{22}| \\ |S_{12}| &= \sqrt{1 - |S_{11}|^2} \end{aligned} \tag{81}$$

The S parameters at an arbitrary reference plane can be characterized by

$$\begin{aligned} &S_{11} = |S_{11}|e^{j\theta_1}, \quad S_{22} = |S_{22}|e^{j\theta_2} = |S_{11}|e^{j\theta_2} \\ &S_{12} = |S_{12}|e^{j\phi} = \sqrt{1 - |S_{11}|^2}e^{j\phi} = S_{21} \\ &\phi = \frac{\theta_1 + \theta_2}{2} + \frac{\pi}{2} \mp 2n\pi \end{aligned} \tag{82}$$

Equation (76) can now be written as

$$\begin{aligned} \det S S_{22}^* &= S_{11}|S_{11}|^2 - |S_{12}|^2 e^{2j\phi}|S_{11}|e^{-j\theta_2} \\ &= S_{11}|S_{11}|^2 - |S_{12}|^2 e^{2j\phi}|S_{11}|e^{-j\theta_1}e^{j\theta_1}e^{-j\theta_2} \\ &= S_{11}|S_{11}|^2 - |S_{12}|^2 e^{2j\phi}S_{11}e^{-j(\theta_1+\theta_2)} \\ &= S_{11}\{|S_{11}|^2 - (1 - |S_{11}|^2)e^{2j\phi}e^{-j(\theta_1+\theta_2)}\} \\ &= S_{11}\{|S_{11}|^2 + (1 - |S_{11}|^2)e^{2j\phi}e^{-j(\theta_1+\theta_2)}e^{j\pi}\} \end{aligned} \tag{83}$$

or

$$\det S S_{22}^* = S_{11}\{|S_{11}|^2 + (1 - |S_{11}|^2)\} = S_{11} \tag{84}$$

Substituting (80) into (75) gives

$$\Gamma_i = \frac{b_1}{a_1} = \det S \frac{S_{22}^* - \Gamma_a}{1 - S_{22}\Gamma_a} \tag{85}$$

The input reflection given by Eq. (71) can be minimized if S_{22}^* is adjusted to cancel the aperture reflection in amplitude and phase. The following necessary condition must exist for the two-port matching network.

$$\Gamma_a = S_{22}^* \tag{86}$$

5.2.1. Airgap Matching Network. On the basis of condition (82) in this section, it is shown that S_{22}^* can be tuned using an airgap matching network to minimize the input reflection. Consider therefore an air-filled homogeneous waveguide section with finite length l, which is bounded by two waveguide sections filled with dielectric material as illustrated in Fig. 39.

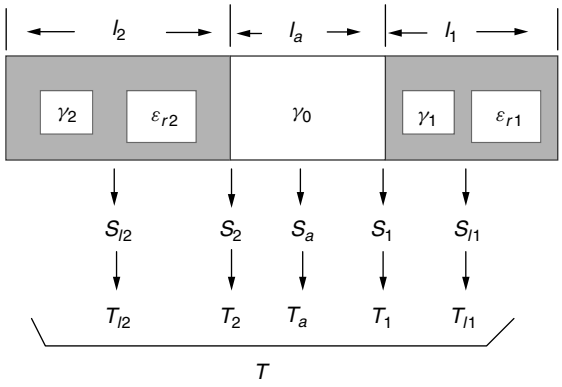


Figure 39. Configuration of S–T matrix with the airgap [30].

The overall T-matrix is formed by the multiplication of a series of successive T matrices

$$T = T_{l2}T_2T_{la}T_1T_{l1} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \tag{87}$$

The scattering matrix S^0 of the two-port networks is related to the transmission matrix T^0 and vice versa as

$$T^0 = \begin{bmatrix} T_{11}^0 & T_{12}^0 \\ T_{21}^0 & T_{22}^0 \end{bmatrix} = \begin{bmatrix} -\dfrac{\Delta S^0}{S_{21}^0} & \dfrac{S_{11}^0}{S_{21}^0} \\ -\dfrac{S_{22}^0}{S_{21}^0} & \dfrac{1}{S_{21}^0} \end{bmatrix} \tag{88}$$

where $\Delta S^0 \triangleq S_{11}^0 S_{22}^0 - S_{21}^0 S_{12}^0$:

$$S^0 = \begin{bmatrix} S_{11}^0 & S_{12}^0 \\ S_{21}^0 & S_{22}^0 \end{bmatrix} = \begin{bmatrix} \dfrac{T_{12}^0}{T_{22}^0} & \dfrac{\Delta T^0}{T_{22}^0} \\ \dfrac{1}{T_{22}^0} & \dfrac{T_{21}^0}{T_{22}^0} \end{bmatrix} \tag{89}$$

From (85) S_{22} is given by

$$S_{22}^0 = -\frac{T_{21}^0}{T_{22}^0} \tag{90}$$

The scattering matrix of the microwave network given in Fig. 39 can be written as [30]

$$\begin{aligned} S_{lm} &= \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} 0 & \exp(-\gamma_m l_m) \\ \exp(-\gamma_m l_m) & 0 \end{bmatrix} \\ S_1 &= \frac{1}{\gamma_a + \gamma_1} \begin{bmatrix} \gamma_0 - \gamma_1 & 2\sqrt{\gamma_a\gamma_1} \\ 2\sqrt{\gamma_a\gamma_1} & \gamma_0 - \gamma_1 \end{bmatrix} \\ S_2 &= \frac{1}{\gamma_2 + \gamma_a} \begin{bmatrix} \gamma_2 - \gamma_a & 2\sqrt{\gamma_a\gamma_2} \\ 2\sqrt{\gamma_a\gamma_2} & \gamma_2 - \gamma_a \end{bmatrix} \end{aligned} \tag{91}$$

where l_m is the length of each homogeneous section. γ_m is the propagation constant in different sections of the network and is given as

$$\gamma_m = \begin{cases} j\sqrt{\omega^2\mu_m\varepsilon_m - \left(\dfrac{\pi}{a}\right)^2} & \text{if } \omega^2\mu_m\varepsilon_m \geq \left(\dfrac{\pi}{a}\right)^2 \\ \sqrt{\left(\dfrac{\pi}{a}\right)^2 - \omega^2\mu_m\varepsilon_m} & \text{otherwise} \end{cases} \tag{92}$$

Using (84) the T matrix of the microwave network can be written as

$$T_{lm} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \exp(-\gamma_m l_m) & 0 \\ 0 & \exp(+\gamma_m l_m) \end{bmatrix}$$

$$T_1 = \frac{1}{2\sqrt{\gamma_a \gamma_1}} \begin{bmatrix} \gamma_a + \gamma_1 & \gamma_a - \gamma_1 \\ \gamma_a - \gamma_1 & \gamma_a + \gamma_1 \end{bmatrix} \tag{93}$$

$$T_2 = \frac{1}{2\sqrt{\gamma_a \gamma_2}} \begin{bmatrix} \gamma_2 + \gamma_a & \gamma_2 - \gamma_a \\ \gamma_2 - \gamma_a & \gamma_2 + \gamma_a \end{bmatrix}$$

The overall T matrix of the two-port network using (83) is given by

$$\begin{aligned} T_{11} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2+\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2-\gamma_a) \\ &\quad \times (\gamma_a-\gamma_1)\exp(+\gamma_a l_a)]\exp(-\gamma_1 l_1)\exp(\gamma_2 l_2) \\ T_{12} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2+\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2-\gamma_a) \\ &\quad \times (\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\exp(+\gamma_1 l_1)\exp(-\gamma_2 l_2) \\ T_{21} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2-\gamma_a)(\gamma_a+\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2+\gamma_a) \\ &\quad \times (\gamma_a-\gamma_1)\exp(+\gamma_a l_a)]\exp(-\gamma_1 l_1)\exp(-\gamma_2 l_2) \\ T_{22} &= \frac{1}{4\gamma_a\sqrt{\gamma_2\gamma_1}}[(\gamma_2-\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) + (\gamma_2+\gamma_a) \\ &\quad \times (\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\exp(+\gamma_1 l_1)\exp(\gamma_2 l_2) \end{aligned} \tag{94}$$

Substituting (90) in (86) leads to the expression for S_{22}

$$\begin{aligned} S_{22}^0 = -\frac{T_{21}^0}{T_{22}^0} &= \frac{\begin{matrix}[(\gamma_2+\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) \\ +(\gamma_2-\gamma_a)(\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\end{matrix}}{\begin{matrix}[(\gamma_2-\gamma_a)(\gamma_a-\gamma_1)\exp(-\gamma_a l_a) \\ +(\gamma_2+\gamma_a)(\gamma_a+\gamma_1)\exp(+\gamma_a l_a)]\end{matrix}} \exp(-2\gamma_1 l_1) \\ &= \Gamma_a^* \end{aligned} \tag{95}$$

The length of the airgap is used to tune the necessary condition given by (82). Figures 40 and 41 show the amplitude and phase of S_{22} as a function of the airgap length and with l_1 as parameter. The frequency is 1.8 GHz. Note that the length of l_1 does not have any effect on the amplitude of S_{22}. The dielectric material has a dielectric constant of 2.53. The flowchart in Fig. 42 gives the procedure for matching the aperture reflection. Using (81) and (84), the input reflection is related to the T matrix as follows:

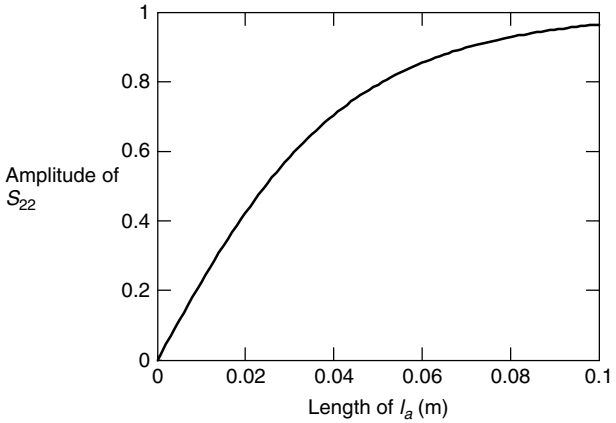


Figure 40. The amplitude of S_{22} as a function of l_a.

$$\Gamma_i = \frac{b_1}{a_1} = \frac{T_{11}^0 \Gamma_a + T_{12}^0}{T_{21}^0 \Gamma_a + T_{22}^0} \tag{96}$$

where the elements of the T matrix are as given by Eq. (90). Figure 43 shows the input reflection as a function of frequency in the L and X bands. The dielectric has a constant of 2.53. The length of the airgap is a parameter.

The length of l_1 is used to tune the minimum of the input reflection for a desired frequency. Figure 44 shows the input reflection as function of frequency for different values of l_1. From this figure the designer can choose the length, l_1, for tuning the resonance frequency.

6. *E*-PLANE STEPPED DFW

In many microwave applications it is necessary to have waveguides with larger aperture dimensions in order to

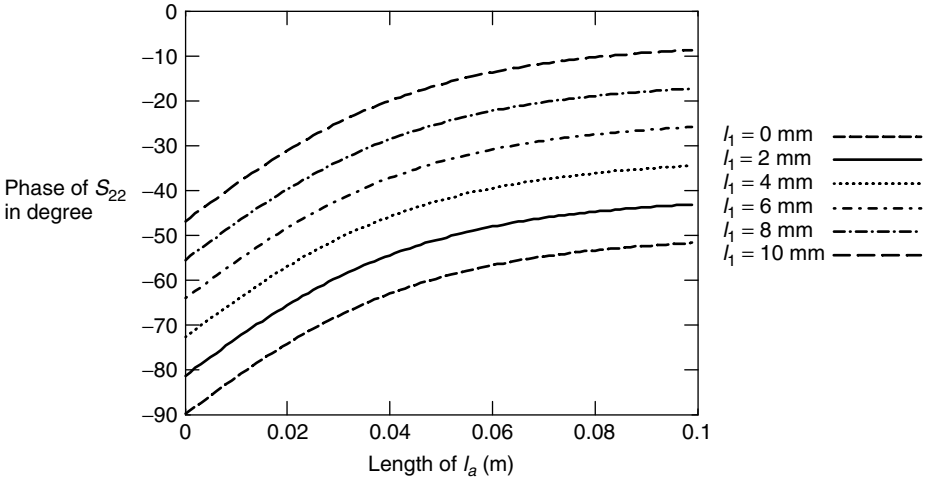


Figure 41. The phase of S_{22} as function of l_a with l_1 as parameter.

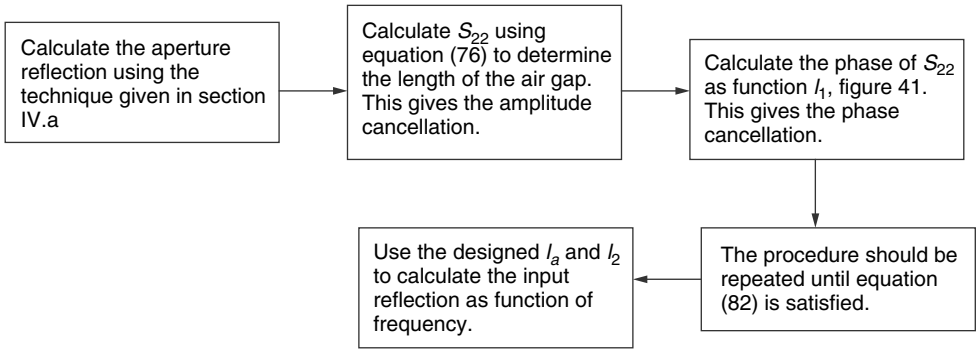


Figure 42. Flowchart for designing an airgap matching network.

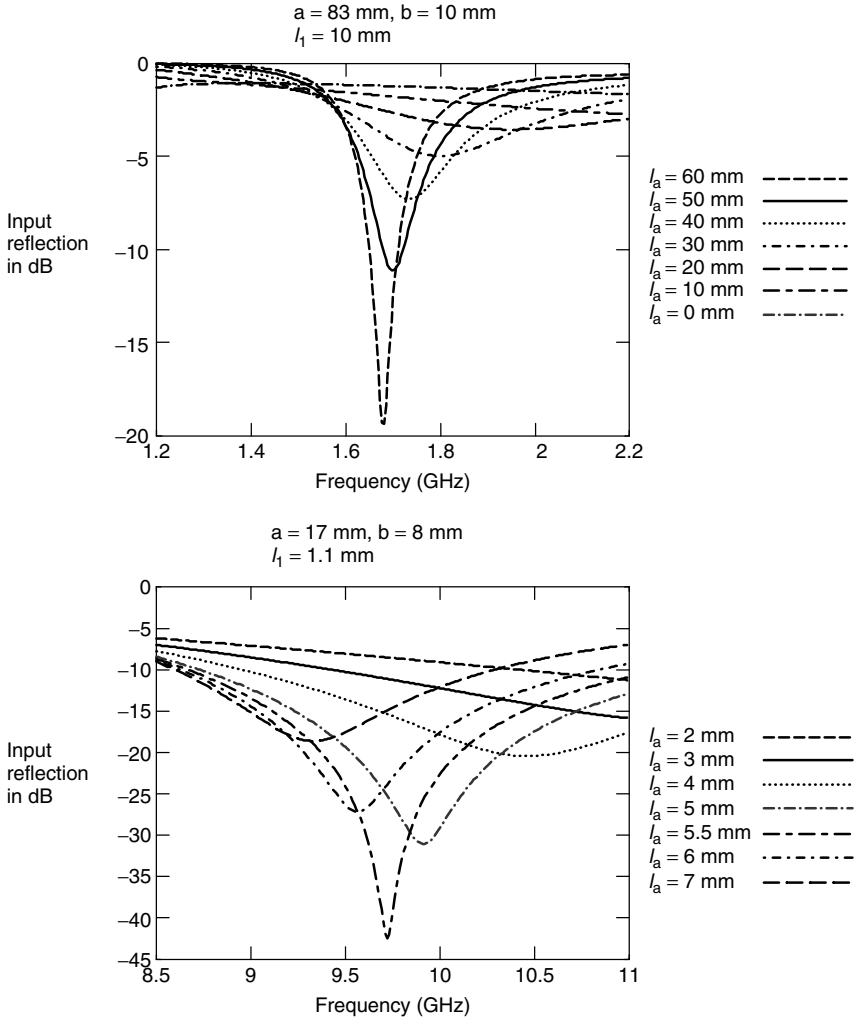


Figure 43. The input reflection as function of frequency in the L and X bands. The length of the airgap is the parameter.

have a better gain and to be able to integrate the antenna with a planar circuit. To achieve this goal, waveguide step discontinuities have been suggested [23]. The electromagnetic boundary conditions at a discontinuity usually require the presence of high-order modes. When the new dimensions of the waveguide are such that the higher-order modes are below cutoff, these modes are confined to a region very close to the discontinuity. A reactive network can then model these localized modes. Figure 45 illustrates the steps in height (*E*-plane stepped) and width (*H*-plane stepped).

The discontinuities in the waveguide can be used to realize matching networks, phase shifters, high- or lowpass filters, and resonators. In this section the *S-T* matrix approach and the airgap matching techniques discussed in Section 5 are used to characterize the input

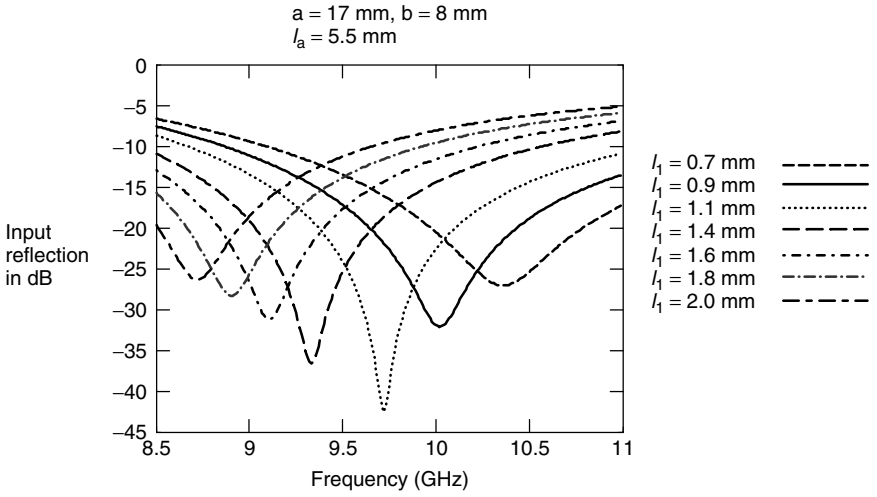


Figure 44. The input reflection as a function of frequency in the X band. The length l_1 is the parameter.

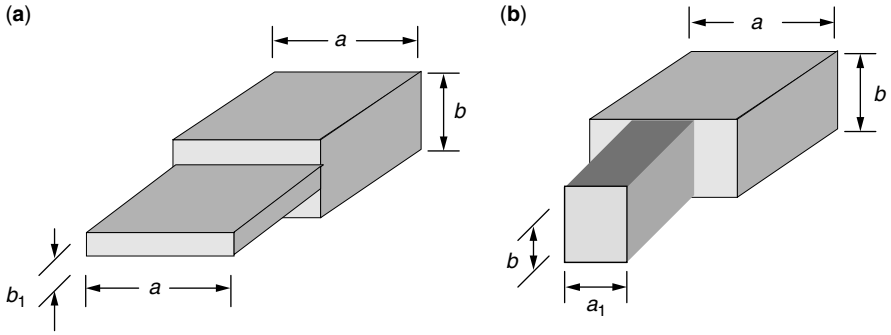


Figure 45. Symmetric step discontinuities in a waveguide: (**a**) E plane; (**b**) H plane.

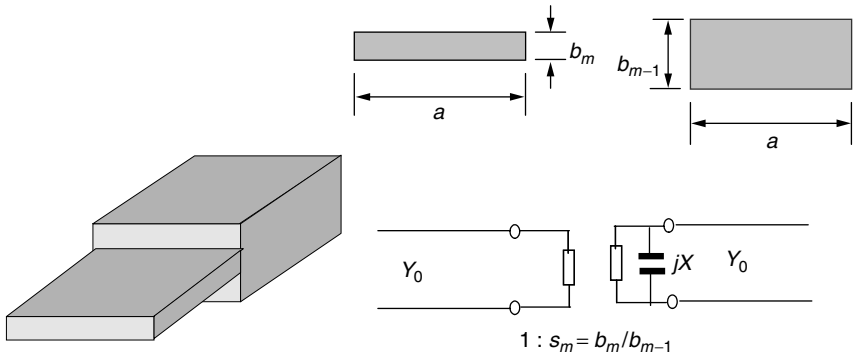


Figure 46. The stepped E plane and its lumped-circuit network representation.

reflection of a double-stepped E-plane DFW. The E-plane step configuration, its dimensions, and the lumped-circuit network are shown in Fig. 46.

The lumped-circuit network representation is a capacitance, and its susceptance value is given by [31]

$$X_m = Y_{10}\frac{2b_m}{\lambda_g}\left\{\ln\left(\frac{1-s_{m+1}^2}{4s_{m+1}}\right)\left(\frac{1+s_{m+1}}{1-s_{m+1}}\right)^{1/2(s_{m+1}+(1/s_{m+1}))} + \frac{2}{H_m}\right\}$$

$$\lambda_g = \frac{2\pi}{\gamma_m} = \frac{2\pi}{\sqrt{\omega^2\mu_m\varepsilon_m - \left(\frac{\pi}{a}\right)^2}}$$

$$s_m = \frac{b_m}{b_{m-1}}, H_m = \left(\frac{1+s_{m+1}}{1-s_{m+1}}\right)^{2s_{m+1}} \times \frac{1+\sqrt{1-\left(\frac{\gamma_m b_m}{2\pi}\right)^2}}{1-\sqrt{1-\left(\frac{\gamma_m b_m}{2\pi}\right)^2}} - \frac{1-3s_{m+1}^2}{1-s_{m+1}^2}$$

$$Y_{10} = \frac{\gamma_m}{\omega\mu} \quad m = 1, 2 \tag{97}$$

where Y_{10} is the waveguide admittance of the TE_{10} mode. b_{m-1} and b_m are the heights (steps) of the waveguides, λ_g is the wavelength of the dominant mode in the waveguide,

and S_m is the step ratio. It is shown that an accurate result can be achieved by considering only the fundamental mode. The transmission matrix elements of the E-plane step are

$$T_{11}^m = \frac{1 + S_{m+1} - jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}},$$

$$T_{12}^m = \frac{1 - S_{m+1} - jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}}$$

$$T_{21}^m = \frac{1 - S_{m+1} + jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}},$$

$$T_{22}^m = \frac{1 + S_{m+1} + jX_m \dfrac{\gamma_m b_m}{\pi} S_{m+1}}{2\sqrt{S_{m+1}}} \quad m = 1, 2 \tag{98}$$

where $m = 1,2$ for the first and second steps, respectively. The input reflection at the reference plane (see Fig. 47) is given by Eq. (92), where the total T matrix is given by

$$T = T_{l4}T_4T_{l3}T_3T_{l2}T_2T_{la}T_1T_{l1} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \tag{99}$$

Using Eqs. (95), (88), (89), and (94), the elements of the overall transmission matrix are

$$T_{11} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3C \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3E\}A_4G \\ \quad + \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3D \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3F\}B_4I \end{array} \right\} A_5$$

$$T_{12} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3C \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3E\}A_4H \\ \quad + \{[A_1T_{11}^1A_2T_{11}^2 \\ \quad + A_1T_{12}^1B_2T_{21}^2]A_3D \\ \quad + [A_1T_{11}^1A_2T_{12}^2 \\ \quad + A_1T_{12}^1B_2T_{22}^2]B_3F\}B_4J \end{array} \right\} B_5$$

$$T_{21} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3C \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3E\}A_4G \\ \quad + \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3D \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3F\}B_4I \end{array} \right\} A_5$$

$$T_{22} = \frac{1}{4\gamma_0\sqrt{\gamma_1\gamma_2}} \left\{ \begin{array}{l} \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3C \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3E\}A_4H \\ \quad + \{[B_1T_{21}^1A_2T_{11}^2 \\ \quad + B_1T_{22}^1B_2T_{21}^2]A_3D \\ \quad + [B_1T_{21}^1A_2T_{12}^2 \\ \quad + B_1T_{22}^1B_2T_{22}^2]B_3F\}B_4J \end{array} \right\} B_5 \tag{100}$$

where

$$T_{11}^1 = \frac{1 + S_2 - jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}}, \quad T_{12}^1 = \frac{1 - S_2 - jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}},$$

$$T_{21}^1 = \frac{1 - S_2 + jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}}, \quad T_{22}^1 = \frac{1 + S_2 + jB_1 \dfrac{\gamma_2 b_1}{\pi} S_2}{2\sqrt{S_2}}$$

$$T_{11}^2 = \frac{1 + S_3 - jB_2 \dfrac{\gamma_3 b_2}{\pi} S_3}{2\sqrt{S_3}}, \quad T_{12}^2 = \frac{1 - S_3 - jB_2 \dfrac{\gamma_3 b_2}{\pi} S_3}{2\sqrt{S_3}},$$

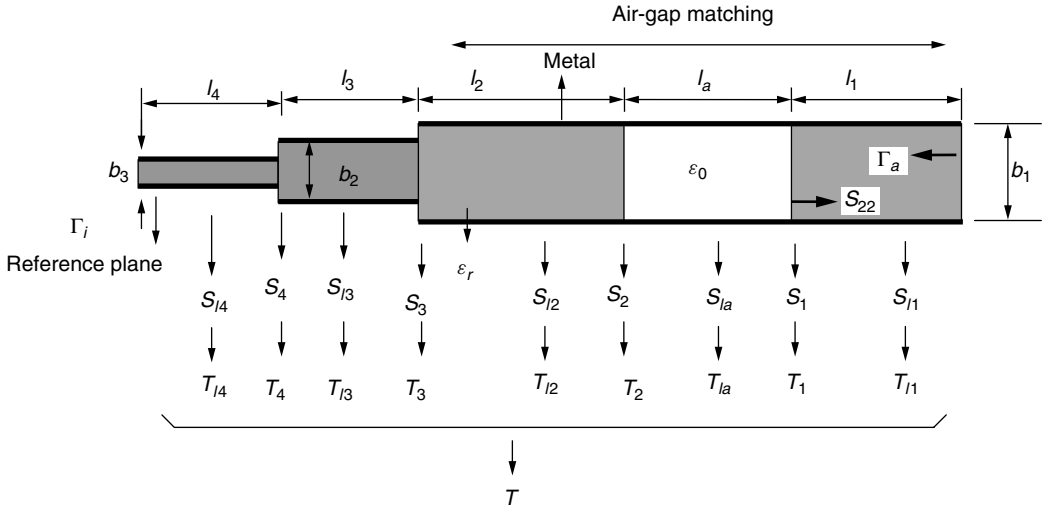


Figure 47. DFW with the airgap matching network and two E plane steps as a cascaded S–T network.

$$T_{12}^2 = \frac{1 - S_3 + jB_2\frac{\gamma_3 b_2}{\pi}S_3}{2\sqrt{S_3}}, \quad T_{22}^2 = \frac{1 + S_3 + jB_2\frac{\gamma_3 b_2}{\pi}S_3}{2\sqrt{S_3}}$$

$$A_1A_2 = \exp(-\gamma_4 l_4)\exp(-\gamma_3 l_3),$$

$$A_1B_2 = \exp(-\gamma_4 l_4)\exp(+\gamma_3 l_3)$$

$$A_3C = \exp(-\gamma_2 l_2)(\gamma_2 + \gamma_a), \quad B_3E = \exp(+\gamma_2 l_2)(\gamma_2 - \gamma_a),$$

$$A_4G = \exp(-\gamma_a l_a)(\gamma_a + \gamma_1) \quad A_3D = \exp(-\gamma_2 l_2)(\gamma_2 - \gamma_a),$$

$$B_3F = \exp(+\gamma_2 l_2)(\gamma_2 + \gamma_a), \quad B_4I = \exp(+\gamma_a l_a)(\gamma_a - \gamma_1)$$

$$A_4H = \exp(-\gamma_a l_a)(\gamma_a - \gamma_1), \quad B_4J = \exp(+\gamma_a l_a)(\gamma_a + \gamma_1)$$

$$B_1A_2 = \exp(+\gamma_4 l_4)\exp(-\gamma_3 l_3),$$

$$B_1B_2 = \exp(+\gamma_4 l_4)\exp(+\gamma_3 l_3),$$

$$A_5 = \exp(-\gamma_1 l_1), B_5 = \exp(+\gamma_1 l_1) \tag{101}$$

where $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$, since the dielectric constant of the filling material is the same for different sections. The input reflection can be calculated as a function of different parameters. These parameters are the frequency, the dielectric constant, the length of the airgap, and the length of the filled homogeneous section. Figure 48 shows the input reflection as a function of frequency for different parameters at reference plane 1 (see Fig. 49).

For this case $a_1 = a_2 = a_3 = 17$ mm, $b_1 = 11$ mm, $b_2 = 8$ mm, $b_3 = 5$ mm, and $\varepsilon_r = 2.53$. Note that the aperture reflection is calculated using the method given in Section 4. The length of l_4 does not affect the amplitude of the input reflection coefficient. These optimal values are calculated by the trial-and-error method. Note that the input reflection is very sensitive to the length of different sections.

6.1. Measurement Results

Figure 49 shows the realized E-plane stepped DFW in the X band, $a_1 = a_2 = a_3 = 17$ mm and $\varepsilon_r = 2.53$.

In order to carry out the input reflection measurement at the reference plane 1 (see Fig. 49), the antenna is calibrated using a modified waveguide calibration technique [32]. Three standard short circuits with different offset delays are designed for this purpose. Figure 50 shows the calibration set. The width of the three waveguide standards equals $a = 17$ mm.

Figure 51 compares the calculated and measured input reflections at the reference plane 1 as a function of frequency.

The measurement differs from the theoretical results since the two steps were not constructed in one piece. In order to ensure good galvanic contact between the waveguide steps, it was necessary to use screws. Such a technique can lead to an airgap between the steps. Also the inaccuracies in length of each section were not

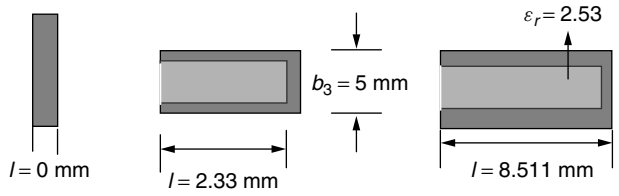


Figure 50. The standard short circuits for waveguide calibration.

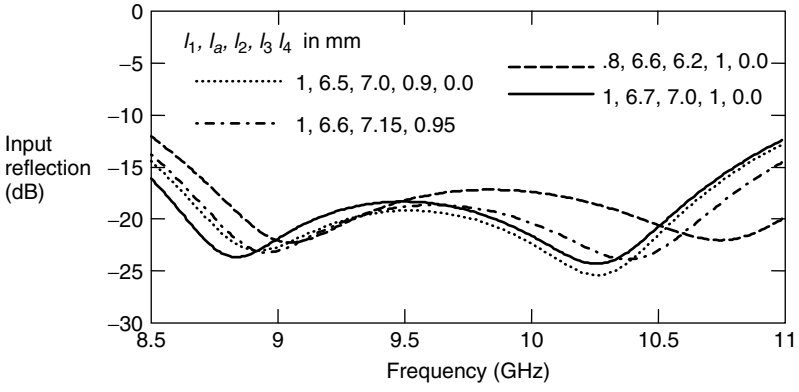


Figure 48. The input reflection as a function of frequency.

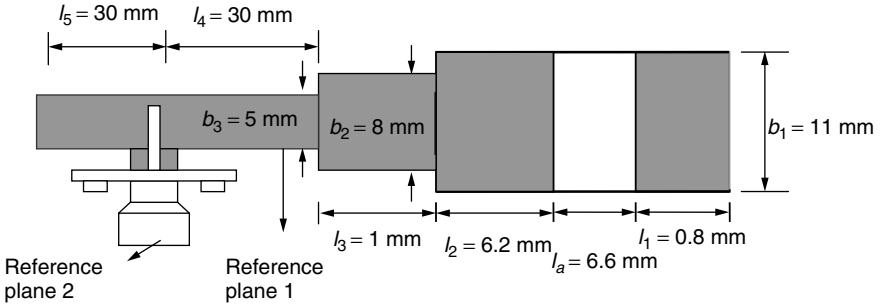


Figure 49. Realized DFW in X band.

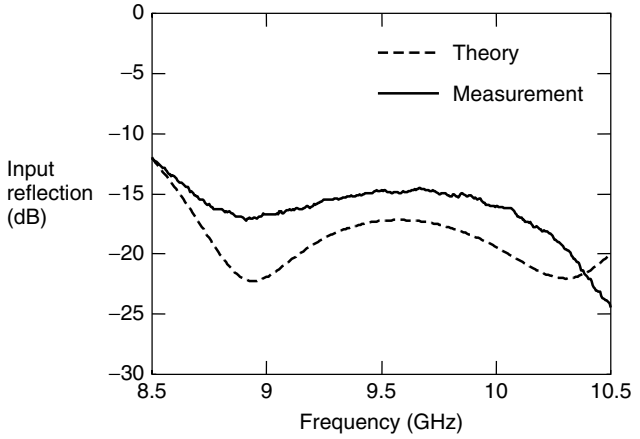


Figure 51. Comparison between the calculated and measured input reflection at reference plane 1.

taken into account. It can be seen from Fig. 48 that the input reflection is very sensitive to the values of the design parameters.

7. DUAL-POLARIZED WAVEGUIDE ANTENNA

In many applications polarization plays a major role. Since the late 1990s there has been an enormous expansion in wireless communication systems and in the number of people using their services. Limited bandwidth is the major concern that may prevent expansion of the capacity. In order to expand the capacity, the use of multiple antennas at the base station has been suggested [33]. Such antennas consist of a number of antenna elements that can transmit and receive the signals independently from each other. Using signal processing algorithms, multiple antenna systems can continuously distinguish between the desired signals and multipath and interfering signals by tracking the desired users with the main lobes and the interferers with the pattern minima. In this way it is possible to maximize the carrier-to-interference ratio (C/I) and the signal-to-noise-ratio (SNR). This is called *space-division multiplexing access* (SDMA). If dual-polarized antennas are used so that two orthogonal polarizations are received, this can enhance the array signal processing in two different ways. First, the number of available signals is increased, which can further improve C/I after signal processing [34,35]. It is also possible to use a postdetection maximum ratio combining technique based on signals coming from the two polarizations.

Radar polarimetry is a valuable technique for the extraction of geophysical parameters from synthetic aperture radar (SAR) images and terrain classification. In many radar applications SAR is used for target detection, classification, and identification. Cloude and Papathanassiou [36] describe the use of a dual-polarized antenna system in a spaceborne satellite in NASA's mission to the planet earth that is intended to provide measurements of the earth's environment (air, water, land). The measurements are used to determine land–surface soil moisture, ocean salinity, surface temperature, and vegetation water content in the L band (1.4 GHz). Liu et al. [37] use a high-resolution dual-polarization X-band radar at "low grazing" angle to obtain images from the ocean surface. Characteristics of low-grazing-angle backscatter marked differences in horizontally and vertically polarized Doppler properties.

7.1. Design

When designing dual polarized antennas, one needs to keep the input reflection of both polarizations at the coax reference plane and the aperture reflection as low as possible. At the same time the isolation between the two feeds must be as high as possible. The techniques described in the previous sections can be used to match the aperture reflection and the coax-to-waveguide transitions. For the isolation between the two polarizations, a polarization filter needs to be designed.

It is desired for the filter to be transparent for one polarization and to reflect the other one. Since the waveguide cutoff frequency depends on polarization, it is possible to design a polarization filter that satisfies this requirement. The polarization filter is a piece of thin metal, which is located vertically or horizontally inside the waveguide. In this way the piece of metal divides the waveguide section in two new ones, each with a different cut off frequency than the original one. The concept is illustrated in Fig. 54.

The polarization filter divides the original waveguide in two equal waveguides in which only the fundamental mode can be propagated with vertical polarization. The longer the filter length, the more the unwanted polarization will be attenuated. The thinner the filter, the less it would disturb the fundamental vertical mode and thus the more transparent it becomes for the desired polarization.

In order to design the length of the filter, one must to calculate the attenuation of the mode in the waveguide after the polarization filter. Using Eq. (21), for TE_{10}, the cutoff frequency of the waveguide given in Fig. 52 becomes

$$f_{c_{10}} = \frac{c}{2a} \tag{102}$$

Substituting (98) and $f = c/\lambda$ in Eq. (22), the propagation constant for the evanescent waves (horizontal polarization) becomes

$$\begin{aligned} \beta_z &= \beta\sqrt{\left(\frac{f_c}{f}\right)^2 - 1} = \frac{2\pi}{\lambda}\sqrt{\left(\frac{\lambda}{2b}\right)^2 - 1} \\ (\beta_z)^2 &= (2\pi)^2\left[\left(\frac{1}{2b}\right)^2 - \left(\frac{1}{\lambda}\right)^2\right] \end{aligned} \tag{103}$$

Consider

$$\delta = 20\log_{10}[e^{\beta_z \rho}] \tag{104}$$

where ρ is the length of the filter and δ is the attenuation of the wave in decibels. This becomes

$$\delta = \frac{20}{\ln(10)}\ln[e^{\beta_z \rho}] = \frac{20}{\ln(10)}\beta_z \rho \tag{105}$$

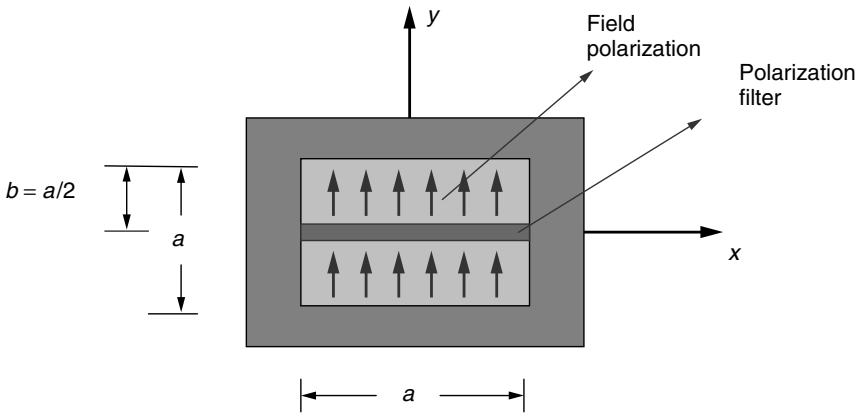


Figure 52. Front view of the polarization filter [38].

Substituting Eq. (99) into (101) leads to

$$\delta = \frac{20\rho}{\ln(10)}(2\pi)\sqrt{\left(\frac{1}{2b}\right)^2 - \left(\frac{1}{\lambda}\right)^2} \tag{106}$$

Rearranging (102) gives

$$\frac{\delta}{\rho} = \frac{40\pi}{\ln(10)}\frac{1}{\lambda}\sqrt{\left(\frac{\lambda}{2b}\right)^2 - 1} \tag{107}$$

An upper bound can be found and is given by the following equation:

$$\frac{\delta}{\rho} \le \frac{40\pi}{\ln(10)}\frac{1}{2b} \le 27.3\frac{1}{b} \tag{108}$$

Figure 53 shows the attenuation of the TE_{10} mode as a function of the length of the polarization filter for three different dielectric materials. It is shown that as the filter length increases, the field attenuates more.

Figure 54 shows the layout of a dual-polarized DFW with a polarization filter and an airgap matching network. In order to lower the diffraction effect from the edge of the waveguide, edge tapering is introduced.

Figure 55 shows the calibrated measurement results for input reflection and isolation after tuning. The *E*- and *H*-plane radiation patterns are shown in Fig. 56. The pattern computation is given in Section 8.

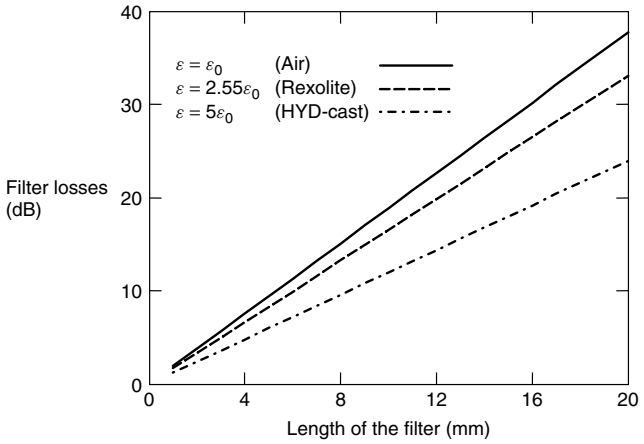


Figure 53. The attenuation of TE_{10} mode as function of polarization filter length at $f = 4$ GHz.

8. RADIATION PATTERN

The waveguide antenna is considered as an aperture antenna. The far-field radiation pattern of the waveguide can be calculated using the equivalent principle [40]. It states that based on the known electromagnetic modes propagating inside the waveguide, one can construct the electric and magnetic currents on the aperture plane. The

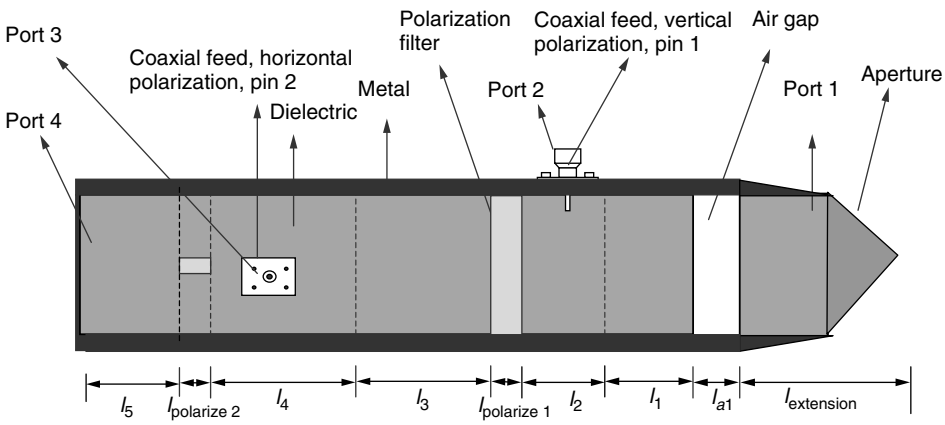


Figure 54. Dual-polarized dielectric-filled waveguide with polarization filter and edge tapering in S-band, side view. (Courtesy of IRCTR.)

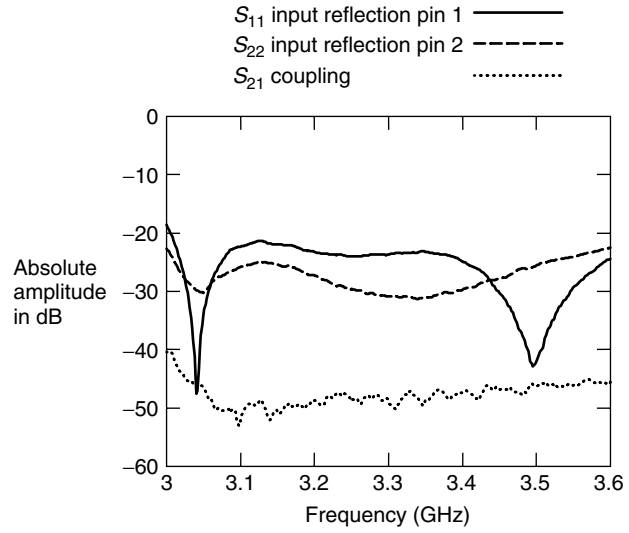


Figure 55. The measured input reflection and isolation of the dual-polarized dielectric-filled waveguide.

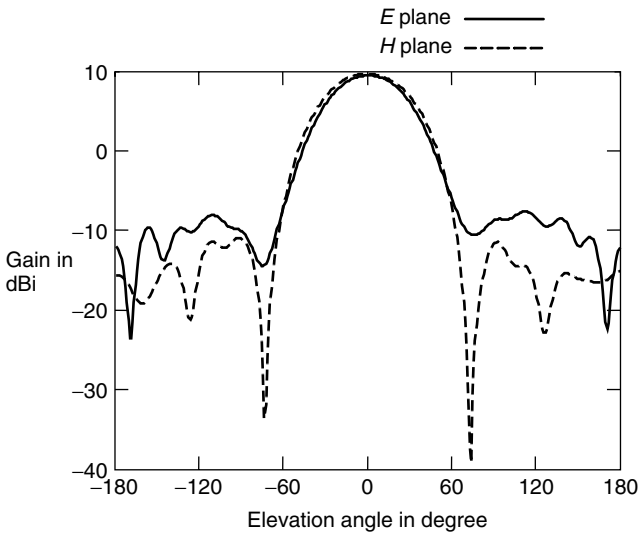


Figure 56. The measured E and H planes of the dielectric-filled waveguide at S band, $f = 3.3$ GHz.

electric and magnetic currents can then be used to set up the "potentials" integral equations to calculate the far-field radiation pattern.

The electromagnetic potentials are a solution to the vector wave equation and can be written as [40]

$$\mathbf{A} = \frac{\mu}{4\pi} \iint_s \mathbf{J} \frac{e^{-jkR}}{R}\, ds'$$
$$\mathbf{F} = \frac{\varepsilon}{4\pi} \iint_s \mathbf{M} \frac{e^{-jkR}}{R}\, ds' \tag{109}$$

where $\mathbf{J}$ and $\mathbf{M}$ are the electric and magnetic current sources; ds' is a differential area. The coordinate system to analyze the radiation pattern is shown in Fig. 57.

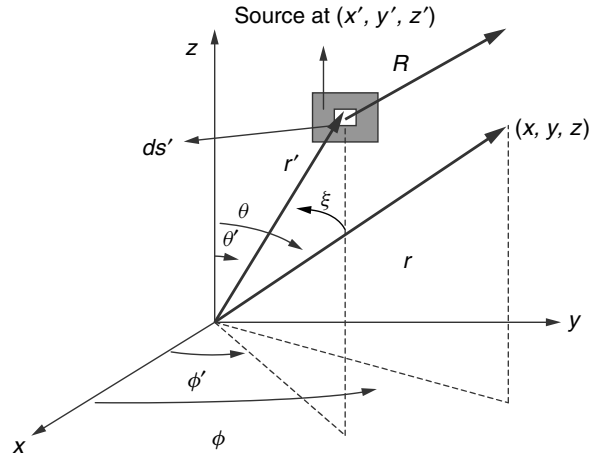


Figure 57. Coordinate system for aperture antenna analysis. (*Source:* C. Balanis.)

It is shown in [40] that in the far zone R can be approximated by

$$R \simeq r - r' \cos\xi \quad \text{for phase variations}$$
$$R \simeq r \quad \text{for amplitude variations} \tag{110}$$

where ξ is the angle between the vector r and r'. The complete electromagnetic field in the far zone is related to the electromagnetic potentials and can be written as

$$\begin{aligned} E_r &\simeq 0 \\ E_\theta &\simeq -\frac{je^{-jkr}}{4\pi r}(F_\phi + \eta A_\theta) \\ E_\phi &\simeq \frac{jke^{-jkr}}{4\pi r}(F_\theta - \eta A_\phi) \\ H_r &\simeq 0 \\ H_\theta &\simeq \frac{jke^{-jkr}}{4\pi r}\left(A_\phi - \frac{F_\theta}{\eta}\right) \\ H_\phi &\simeq -\frac{jke^{-jkr}}{4\pi r}\left(A_\theta + \frac{F_\phi}{\eta}\right) \end{aligned} \tag{111}$$

The potentials given in Eq. (107) can be derived from Eqs. (105) and (106) via

$$\begin{aligned} \mathbf{A} &= \frac{\mu e^{-jkr}}{4\pi r} \iint_s \mathbf{J} e^{jkr' \cos\xi}\, ds' \\ &= \frac{\mu e^{-jkr}}{4\pi r} \iint_s (\hat{\mathbf{a}}_x J_x + \hat{\mathbf{a}}_y J_y + \hat{\mathbf{a}}_z J_z) e^{jkr' \cos\xi}\, ds' \\ \mathbf{F} &= \frac{\varepsilon e^{-jkr}}{4\pi r} \iint_s \mathbf{M} e^{jkr' \cos\xi}\, ds' \\ &= \frac{\varepsilon e^{-jkr}}{4\pi r} \iint_s (\hat{\mathbf{a}}_x M_x + \hat{\mathbf{a}}_y M_y + \hat{\mathbf{a}}_z M_z) e^{jkr' \cos\xi}\, ds' \end{aligned} \tag{112}$$

Using the rectangular-to-spherical transformation, one can give the components of the potentials:

$$\begin{aligned}
A_\theta &= \frac{\mu e^{-jkr}}{4\pi r}\iint_s (J_x\cos\theta\cos\phi + J_y\cos\theta\sin\phi \\
&\quad - J_z\sin\theta)e^{jkr'\cos\xi}\,ds' \\
A_\phi &= \frac{\mu e^{-jkr}}{4\pi r}\iint_s (-J_x\sin\phi + J_y\cos\phi)e^{jkr'\cos\xi}\,ds' \\
F_\theta &= \frac{\varepsilon e^{-jkr}}{4\pi r}\iint_s (M_x\cos\theta\cos\phi + M_y\cos\theta\sin\phi \\
&\quad + M_z\sin\theta)e^{jkr'\cos\xi}\,ds' \\
F_\phi &= \frac{\varepsilon e^{-jkr}}{4\pi r}\iint_s (-M_x\sin\phi + M_y\cos\phi)e^{jkr'\cos\xi}\,ds'
\end{aligned} \tag{113}$$

The electric and magnetic current components in (108) and (109) in general can be found using the equivalent principle.

For a given field distribution, the analytic forms for the fields for an arrangement are not the same. However, the far-zone expression will be the same. For each geometry, the only difference in analysis is in the formulation of

1. The components of the equivalent current densities $(J_x, J_y, J_z, M_x, M_y, M_z)$
2. The difference in paths from the source to the observation point $r'\cos\xi$
3. The differential area ds'

8.1. Uniform Aperture Distribution

For simplicity, the first aperture field is considered to be constant and given by

$$E_a = \begin{cases} \hat{a}_y E_0 & -\dfrac{a}{2} \le x' \le \dfrac{a}{2}, -\dfrac{b}{2} \le y' \le \dfrac{b}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{114}$$

Figure 58 shows the geometry of the waveguide aperture antenna on an infinite electric ground plane. The equivalent principle is used to form the electromagnetic current densities on the opening of the waveguide. Once the current densities are calculated, Eqs. (107) and (109) are used to determine the components of the electromagnetic field in the far zone. The components of the current densities are given by

$$M_s = \begin{cases} -2\hat{n}\times E_a = -2\hat{a}_z\times\hat{a}_y E_0 \\ \qquad\quad = +2\hat{a}_x E_0 & -\dfrac{a}{2} \le x' \le \dfrac{a}{2}, \\ & -\dfrac{b}{2} \le y' \le \dfrac{b}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{115}$$

$$J_s = 0 \qquad \text{everywhere}$$

Substituting (111) in (109) leads to

$$\begin{aligned}
A_\theta &= A_\phi = 0 \\
F_\theta &= \frac{\varepsilon e^{-jkr}}{4\pi r}\cos\theta\cos\phi\left[\int_{-\frac{a}{2}}^{\frac{a}{2}}\int_{-\frac{b}{2}}^{\frac{b}{2}} M_x e^{jk(x'\sin\theta\cos\phi + y'\sin\theta\cos\phi)}\right. \\
&\quad \left.\times\, dx'\,dy'\right]
\end{aligned} \tag{116}$$

The integral within the brackets represents the *space factor* for a two-dimensional distribution. Using the following integral identity

$$\int_{-(c/2)}^{c/2} e^{j\alpha z}\,dz = c\left[\frac{\sin\left(\frac{\alpha}{2}c\right)}{\frac{\alpha}{2}c}\right] \tag{117}$$

Eq. (112) reduces to

$$F_\theta = \frac{\varepsilon e^{-jkr}}{4\pi r}2ab\left[\cos\theta\cos\phi\left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \tag{118}$$

where

$$\begin{aligned}
X &= \frac{ka}{2}\sin\theta\cos\phi \\
Y &= \frac{kb}{2}\sin\theta\sin\phi
\end{aligned} \tag{119}$$

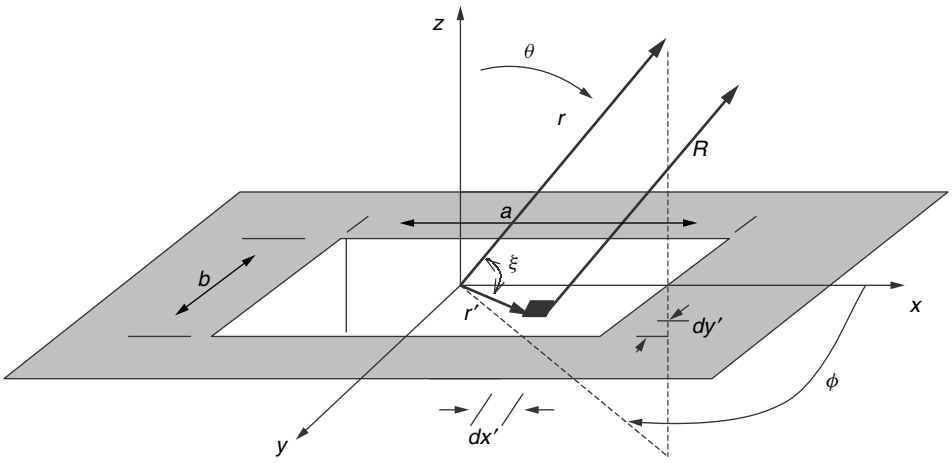


Figure 58. Rectangular waveguide aperture on an infinite electric ground plane.

Similarly it can be shown that

$$F_\phi = -\frac{\varepsilon e^{-jkr}}{4\pi r} 2ab \left[\sin\phi \left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \tag{120}$$

Inserting (116) and (114) in (107), the fields radiated by the waveguide aperture with uniform field distribution can be written as

$$\begin{aligned} E_r &= 0 \\ E_\theta &= -j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\sin\phi\left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \\ E_\phi &= j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\cos\theta\cos\phi\left(\frac{\sin X}{X}\right)\left(\frac{\sin Y}{Y}\right)\right] \\ H_r &= 0 \\ H_\theta &= -\frac{E_\phi}{\eta} \\ H_\phi &= \frac{E_\theta}{\eta} \end{aligned} \tag{121}$$

For the aperture given in Fig. 59, the E-plane pattern is on the y–z plane ($\phi = \pi/2$) and the H-plane pattern is on the x–z plane ($\phi = 0$). Thus

E plane($\phi = \pi/2$):

$$E_r = E_\phi = 0$$

$$E_\theta = j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\frac{\sin\left(\frac{kb}{2}\sin\theta\right)}{\frac{kb}{2}\sin\theta}\right] \tag{121}$$

H plane($\phi = 0$):

$$E_r = E_\theta = 0$$

$$E_\phi = j\frac{abkE_0 e^{-jkr}}{2\pi r}\left[\cos\theta\frac{\sin\left(\frac{ka}{2}\sin\theta\right)}{\frac{ka}{2}\sin\theta}\right] \tag{122}$$

Figures 59 and 60 show the three-dimensional patterns of a rectangular waveguide mounted on an infinite ground plane. Since the dimensions are greater than the wavelength, multiple lobes appear. The number of lobes is directly related to the dimension of the waveguide aperture. The pattern in the H plane is only a function of the dimension a, whereas that in the E plane is influenced only by b. In the E plane, the sidelobe formed on each side of the major lobe is a result of $\lambda < b \leq 2\lambda$. In the H plane, the first minor lobe on each side of the major lobe is formed when $\lambda < a \leq 2\lambda$ and the second sidelobe when $2\lambda < a \leq 3\lambda$. Additional lobes are formed when both aperture dimensions increase.

The patterns computed above assumed that the aperture was mounted on an infinite ground plane. In practice, infinite ground planes are not realizable. Edge effects on the patterns of apertures mounted on finite-size ground planes can be accounted for by the method of moment technique [39]. Figure 61 illustrates the electric and magnetic components of a DFW using the MoM [40].

The electric and magnetic currents are used to set up the far-field integral equations. Figure 62 shows the comparison between the simulations and measurements results for the E- and H-plane patterns.

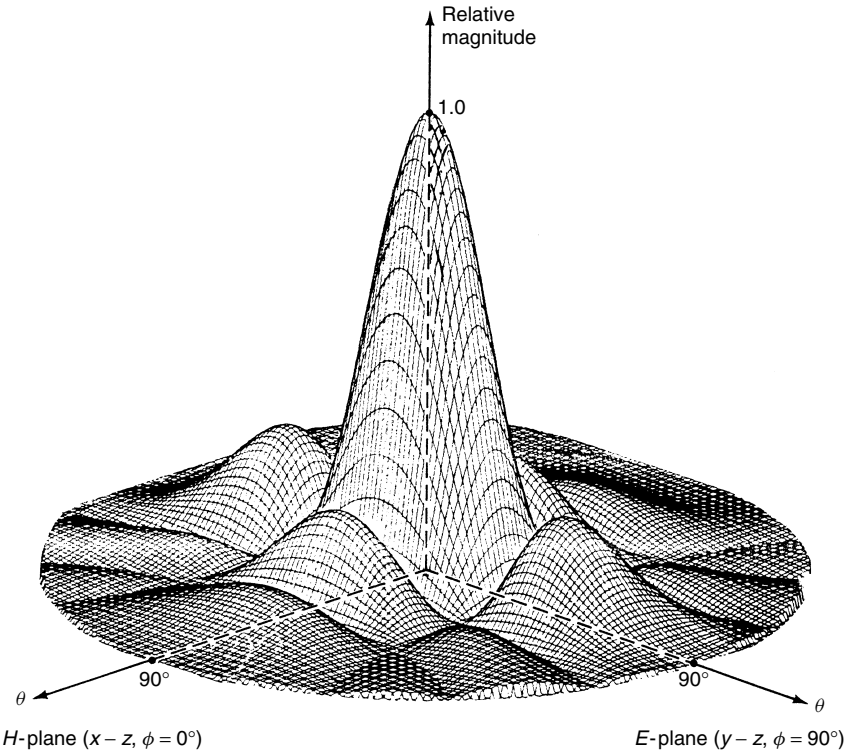


Figure 59. Three-dimensional field pattern of a constant field rectangular aperture mounted on an infinite ground plane ($a = 3\lambda$, $b = 2\lambda$).

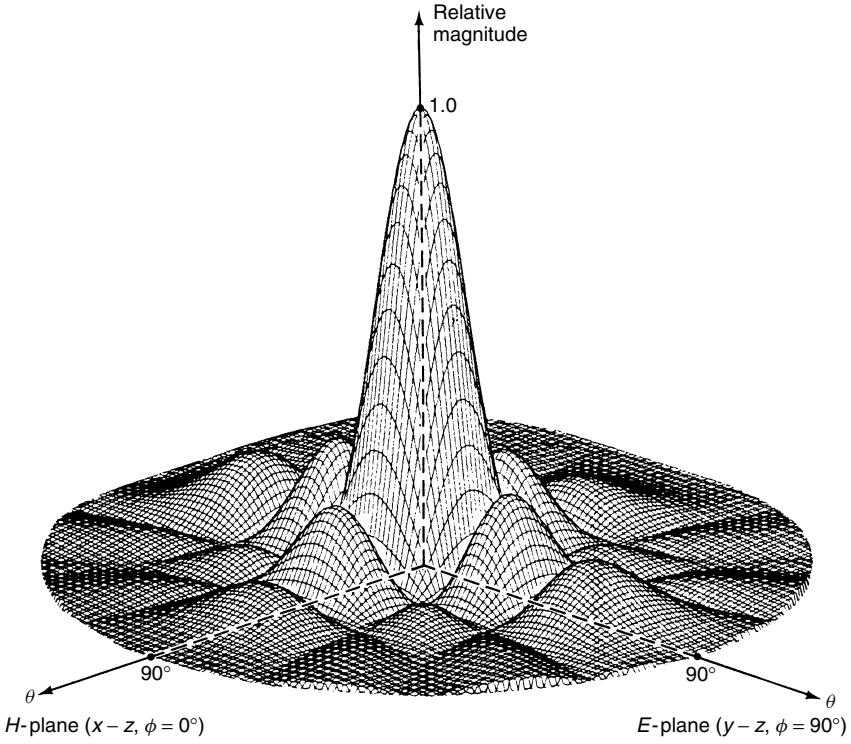


Figure 60. Three-dimensional field pattern of a constant field square aperture mounted on an infinite ground plane ($a = b = 3\lambda$).

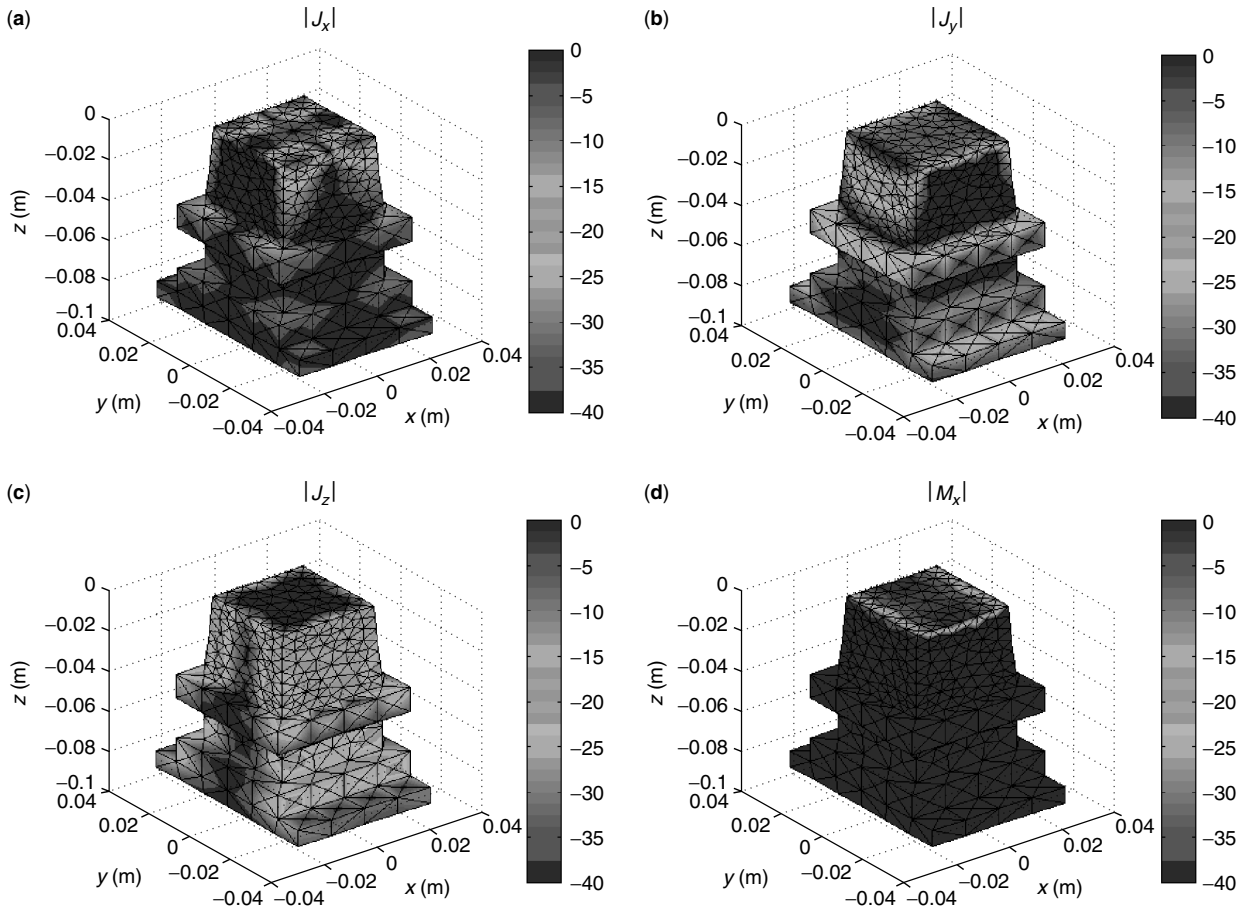


Figure 61. Relative amplitude of the different components of the induced surface currents for DFW. The electric (**a–c**) and magnetic (**d**) currents are given in decibels. Aperture dimensions: $0.38\lambda \times 0.38\lambda$ [40].

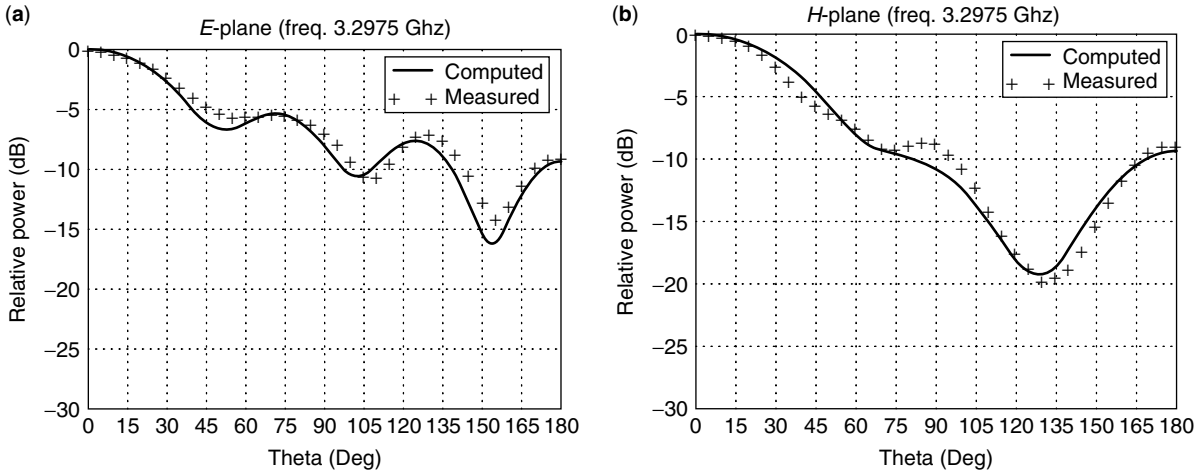


Figure 62. Predicted and measured *E*- and *H*-plane patterns of DFW.

BIOGRAPHIES

M. Hajian was born in Iran on April 21, 1957. He received his B.S. in Physics from the University of Oklahoma (USA) in 1981, and the M.S. degree in Electrical Engineering from Delft University of Technology (in the Netherlands) in 1990. Since 1990 he has been with the Microwave and Radar Laboratory of the Delft University of Technology. In 1995 he became a Senior Lecturer teaching a course on antennas. He is the Netherlands representative of EC/COST 260 on adaptive antennas. His major interests are antennas and propagation, smart antennnas, antenna near-field measurement techniques, and mobile communications systems.

L. P. Ligthart was born in Rotterdam, on September 15, 1946. He graduated with distinction in 1969 and received the M.S. degree in Electrical Engineering from Delft University of Technology. Since 1969 he has been with the Microwave Laboratory of the Delft University of Technology. In 1974 he became a Senior Lecturer teaching an undergraduate course on transmission line theory, antennnas, and propagation. From 1976 to 1977 he spent one year as a senior scientist at Chalmers University, Gothenburg in Sweden. In 1985 he received the Ph.D. degree in Technical Sciences based on his contributions in the design of miniaturized waveguide radiating elements.

Prof. Dr. Ligthart is Director of IRCTR, covering activities on antennas and propagation; radar, mobile, and satellite communication; remote sensing; and electromagnetic compatibility. His present interests include antennas and propagation, radar, and remote sensing.

He received the Vederprijs award in 1981, the IEE-Blumlein-Brown-Willams Premium Award in 1982, and the Doctor Honoris Cause from Moscow State Technical University of Civil Aviation in 1999. He is a fellow of IEE and IEEE, and the Netherlands representative of EC/COST 260 on adaptive antennas and EC/COST on advanced weather radar. He has published over 152 scientific papers.

BIBLIOGRAPHY

1. H.-C. Song et al., Four-branch single-mode waveguide power divider, *IEEE Photon. Technol. Lett.* **10**(12): 1760–1762 (Dec. 1998).
2. L. F. Libelo and C. M. Knop, A corrugated waveguide phase shifter and its use in HPM dual-reflector antenna arrays, *IEEE Trans. Microwave Theory Tech.* **43**(1): 31–35 (Jan. 1995).
3. T. Yoneyama, Millimeter-wave transmitter and receiver using the nonradiative dielectric waveguide, *IEEE MTT-S Digest* 1083–1086 (1989).
4. Y. T. Lo and S. W. Lee, *Antenna Handbook, Theory, Applications, and Design*, Van Nostrand Reinhold, New York, 1988, Chap. 24.
5. G. H. C. van Werkhoeven and A. K. Golshayan, Calibration aspects of the APAR antenna unit, *IEEE Trans. AP* **46**(6): 776–781 (June 1998).
6. A. B. Smolders, Design and construction of a broadband wide-scan angle phased array antenna with 4096 radiating elements, *IEEE-APS on Phased Array Systems and Technology*, Boston, 1996, pp. 87–92.
7. J. Bennett et al., Quadpack X-band T/R module for active phased array radar, *GAAS 98 Conf. Proc.*, Oct. 1998, pp. 63–67.
8. V. K. Lakshmeesha et al., A compact high-power S-band dual frequency, dual polarized feed, *Antennas and Propagation Society International Symposium*, Vol. 3 AP-S. Digest, 1991, pp. 1607–1610.
9. T. S. Bird, M. A. Sprey, K. J. Greene, and G. L. James, A circularly polarized X-band feed system with high transmit/receive port isolation, *Antennas and Propagation*, Vol. 1 Ninth International Conference on (Cof. Publ. No. 407), 1995 pp. 322–326.
10. P. Savi, D. Trinchero, R. Tascone, and R. Orta, A new approach to the design of dual-mode rectangular waveguide filters with distributed coupling, *IEEE Trans. Microwave Theory Tech.* **45**(2): 221–228 (Feb. 1997).
11. D. Crawford and M. Davidovitz, A 2-step waveguide E-plane filter design method using the semi-discrete finite element method, *IEEE Trans. Microwave Theory Tech.* **42**(7): 1407–1411 (July 1994).

12. C. A. Balanis, *Advanced Engineering Electromagnetic*, Wiley, 1989.

13. R. E. Collin, *Field Theory of Guided Waves*, IEEE Press, 1990.

14. M. Tian, P. D. Tran, M. Hajian, and L. P. Ligthart, Air-gap technique for matching the aperture of miniature waveguide antennas, *IEEE Instrumentation and Measurement Technology Conf.*, May 18–20, 1993, pp. 197–201.

15. M. Hajian, T. S. Lam, and L. P. Ligthart, Microstrip-to-waveguide transition for miniature dielectric-filled waveguide antenna, *Microwave Opt. Technol. Lett.* **12**(5): (Aug. 1996).

16. T. Q. Ho and Y.-C. Shih, Spectral-domain analysis of E-plane waveguide to microstrip transition, *IEEE-MTT* **37**(2): 388–392 (Feb. 1989).

17. Transition links waveguide and microstrip lines, *Microwaves RF* 119–120 (May 1994).

18. D. Li and R. Wang, Analysis of waveguide-to-microstrip transition, *Microwave Opt. Technol. Lett.* **5**(3): 128–130 (March 1992).

19. G. E. Ponchak and A. N. Downey, A new model for broadband waveguide-to-microstrip transition design, *Microwave J.* 333–343 (May 1988).

20. T. Q. Ho and Y.-C. Shih, Analysis of microstrip line to waveguide end launchers, *IEEE-MTT* **36**(3): 561–567 (March 1988).

21. P. M. Meaney, A novel transition from waveguide to microstrip, *Microwave J.* **33**(11): 145–148 (Nov. 1990).

22. B. N. Das and K. V. S. V. R. Prasad, Excitation of waveguide by stripline- and microstrip-line-fed slots, *IEEE-MTT* **34**(3): 321–327 (March 1986).

23. P. A. Rizzi, *Microwave Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

24. C. T. Tai, *Dyadic Green's Functions in Electromagnetic Theory*, Scranton Intext, Scranton, PA, 1971, Chap. 5, pp. 76–80.

25. T. Lam, M. Hajian, and L. Ligthart, *Excitation of MLA by Microstrip*, IRCTR internal thesis report, Aug. 1995.

26. M. Hajian, *Analysis and Design of Dielectric Filled Waveguide Antennas for Collision Avoidance Radar*, IRCTR internal report, Feb. 1995.

27. C. N. Capsalis, A rigorous analysis of a coaxial to shielded microstrip line transition, *IEEE-MTT* **37**: 1091–1098 (July 1989).

28. M. Tian, *Characterization of Miniature Dielectric Filled Open Ended Waveguide Antennas*, Ph.D. thesis, Delft Univ., Oct. 1995.

29. R. F. Harrington, *Time-Harmonic Electromagnetic Field*, McGraw-Hill, New York, 1961.

30. L. P. Ligthart, *Antenna Design and Characterization Based on the Elementary Antenna Concept*, Ph.D. thesis, Dutch Efficiency Bureau, 1985.

31. N. Marcuvitz, *Waveguide Handbook*, Dover, New York, 1965.

32. System Manual, HP 8510B, *HP Network Analyzer User's Guide* 1986.

33. J. C. Liberti, Jr. and T. S. Rappaport, *Smart Antennas for Wireless Communications: IS-95 and Third Generation CDMA Applications*, Prentice-Hall, 1999.

34. C. B. Dietrich, Jr., K. Dietze, J. R. Nealy, and W. L. Stutzman, Spatial, polarization, and pattern diversity for wireless handheld terminals, *IEEE Trans. on AP* **49**(9): 1271–1281 (Sept. 2001).

35. C. Passmann, G. Villino, and T. Wixforth, A polarization flexible phased array antenna for a mobile communication SDMA field trial, *Int. Microwave Symp. Digest*, Denver, June 8–13, 1997, Vol. 2, pp. 595–598.

36. S. R. Cloude and K. P. Papathanassiou, Polarimetric SAR interferometry, *IEEE Trans. Geosci. Remote Sens.* **36**(5): (Part 1) 1551–1565 (Sept. 1998).

37. Y. Liu, S. J. Frasier, and R. F. McIntosh, Measurement and classification of low-grazing-angle radar spikes, *IEEE Trans. Antennas Propag.* **46**(1): 27–40 (Jan. 1998).

38. R. F. M. Van den Brink, *Design a Miniaturized Feed at 4 GHz*, IRCTR internal thesis report, Aug. 1984.

39. C. A. Balanis, *Antenna Theory, Analysis and Design*, 2nd ed., Wiley, 1997.

40. A. R. Moumen, *Analysis and Synthesis of Compact Feeds for Large Multiple Beam Reflector Antennas*, Ph.D. thesis, Delft Univ., March 2001.

MILLIMETER-WAVE ANTENNAS

ZHIZHANG (DAVID) CHEN
Dalhousie University
Halifax, Nova Scotia, Canada

1. INTRODUCTION

1.1. Definition of Millimeter Waves — A Part of Electromagnetic Spectrum

As one of the key areas in information technology, telecommunications involve the transmission of electric signals that contain messages from one location to another location as well as processing of these signals. The carriers of the electrical signals are time-varying electromagnetic waves in the forms of electric current flow or radiowave propagation. Two parameters characterize a time-varying electromagnetic wave or signal. The first is called *frequency*, which is defined as the number of cycles of variations per second in unit of *hertz* (Hz). The second is called *wavelength* with units in meters. The wavelength describes the spatial period of repetition in space when a signal of a certain frequency travels in a medium. The relationship between a frequency f and a wavelength λ in free space is $f = c/\lambda$, where c is the speed of light. Theoretically, the frequency of an electromagnetic wave (or the electrical signal) can be from zero to infinity and the corresponding wavelength then goes from infinity to zero, leading to an electromagnetic spectrum of infinite extent.

Electromagnetic waves or electric signals of different frequencies have found many different applications due to their different characteristics. A zero-frequency signal is what we normally call DC (direct current). The most commonly used batteries produce DC energy. It is used in any battery-powered equipment such as CD players and flashlights. A 60-Hz signal is what is used in our electric power grid to distribute electric energy from power stations to our homes. Figure 1 illustrates different applications with

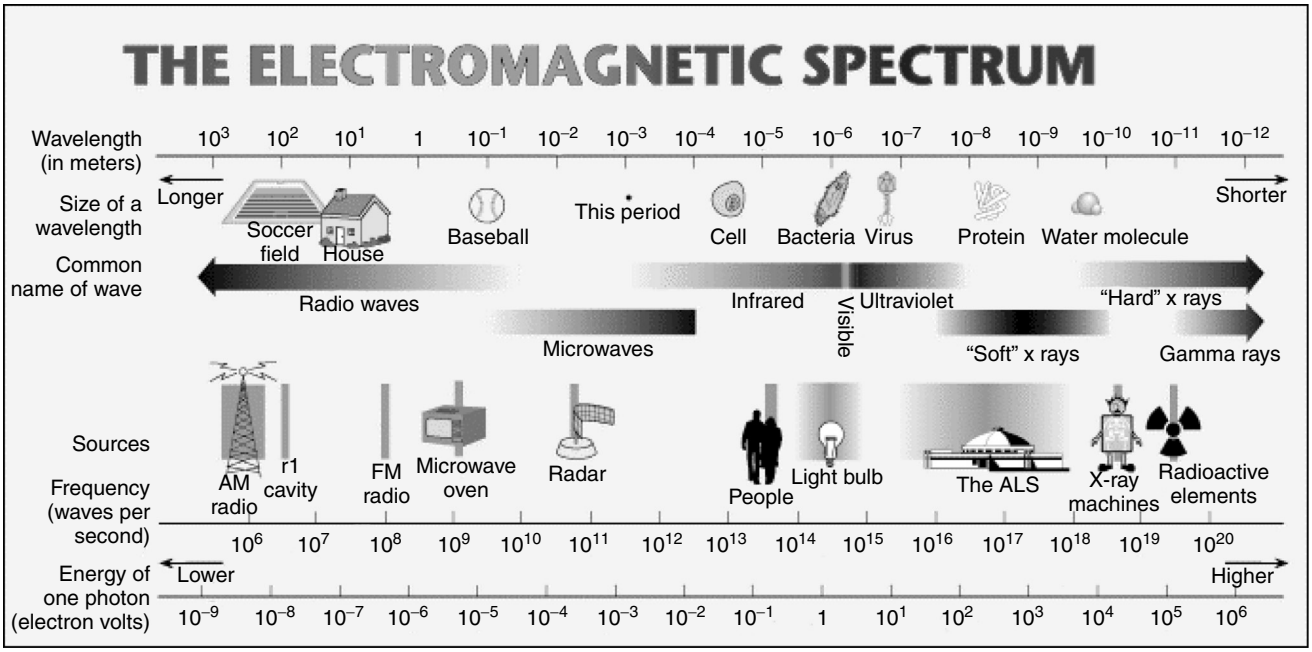


Figure 1. The electromagnetic spectrum (courtesy of the Advanced Light Source, Lawrence Berkeley National Laboratory).

Table 1. Band Classifications of the Radiofrequency Spectrum

Band Classification	Frequency f	Free-Space Wavelength λ	Applications
	<3 Hz	$>10^5$ (km)	Magnetotelluric sensing
Extremely low frequency (ELF)	3–30 Hz	10^5–10^4 km	Detection of buried metal objects
Superlow frequency (SLF)	30–300 Hz	10^4–10^3 km	Ionospheric sensing, electric power distribution, submarine communications
Ultralow frequency (ULF)	300 Hz–3 kHz	10^3–10^2 km	Audio signals on telephone
Very low frequency (VLF)	3–30 kHz	10–1 km	Navigation and position
Low frequency (LF)	30–300 kHz	1 km–100 m	Radio beacon, weather broadcast stations for air navigations
Medium frequency (MF)	300 kHz–3 MHz	100–10 m	AM broadcasting
High frequency (HF)	3–30 MHz	10–1 m	Shortwave broadcasting
Very high frequency (VHF)	30–300 MHz	1 m–10 cm	TV and FM broadcasting, mobile radio communication, air traffic control
Ultrahigh frequency (UHF)	300 MHz–3 GHz	10–1 cm	TV broadcasting, radar, radioastronomy, microwave ovens, cellular phones
Superhigh frequency (SHF)	3–30 GHz	1 cm–1 mm	Radar, satellite communication systems, aircraft navigation, radioastronomy, remote sensing
Extremely high frequency (EHF)	30–300 GHz	1–0.1 mm	Radar, advanced communication systems, remote sensing, radioastronomy

different frequency ranges of the electromagnetic spectrum. Note that contrary to what most people think, X-ray, infrared, or visible light are all parts of the electromagnetic wave spectrum with difference frequencies.

In theory, any part of the electromagnetic spectrum can be used for telecommunications. However, because of specific requirements for communications, most communication systems use the so-called radiofrequency spectrum where frequencies range from 3 Hz to 300 GHz (1 G = 10^9). For clarity, it is artificially divided into various bands, each being named in terms of its frequency and wavelength. Table 1 shows the classifications of the radio spectrum and their respective applications.

The millimeter-wave frequency band falls into the EHF band with frequencies ranging from 30 to 300 GHz (see Table 1). The term *millimeter waves* comes from the fact that the corresponding wavelength is in the range of millimeters in free space.

1.2. Applications of Millimeter Waves

As can be seen in Fig. 1, a millimeter wave is very close to the light in its spectrum position and possess very short wavelengths. Therefore, it has the properties similar to those of light, such as high resolution and large bandwidth. However, better than light, a millimeter wave experiences fewer environmental effects such as atmospheric absorption as it travels through the atmosphere. For this reason, millimeter waves, while possessing higher resolution and larger bandwidth than normal radio and microwave systems, can propagate through various transmission media. As the consequence of these properties, applications have been developed in areas of remote sensing/imaging, radioastronomy, plasma diagnosis, radar, and high-speed or broadband wireless and satellite communications. In particular, in telecommunications, in light of increasingly congested lower frequency bands and growing demands for high-speed data communications such as video transmissions, millimeter waves pose a very promising band of the radio spectrum to be utilized and have attracted growing attentions and research and development efforts.

1.3. Millimeter-Wave Antennas

One of the key components in any wireless telecommunication system is the antenna. It serves as the "eyes" and "ears" of a communication system, or technically, the interfacing system between the air and electronics. It radiates the radio signal it obtains from an electronic transmitter into space and sends the radio signals it detects in space to an electronic receiver. In general, to effect good radiation and detection, the size of an antenna must be proportional to the operating wavelength. The higher the operating frequency (or the shorter the wavelength), the smaller the antenna size. Roughly, the dimensions of an antenna are at least one-quarter to one-half of the wavelengths. The example is the antenna tower at an AM radiobroadcasting station. Because of the operating wavelengths of the AM signals are normally in the range of hundreds of meters, the antenna heights have to be tens or even hundreds of meters. In contrast, the millimeter-wave antennas tend to be very small and in the range of centimeters and millimeters, because millimeter waves have very short wavelengths. Such a feature, as well as the potential broad bandwidths with millimeter waves, has made millimeter-wave antennas very attractive for future short-range and high-speed communication systems. Figure 2 shows the millimeter-wave antennas made of traditional conical and pyramidal horn antenna structures. When designed to operate at 90–140 GHz, the dimensions for the conical horn are about 13 mm (in diameter) × 28 mm (in length) and for the pyramidal horn 20 mm (in width) × 15 mm (in height) × 41 mm (in length).

Like any radio antennas, the important specifications for the millimeter-wave antennas are gain, radiation pattern, return loss, bandwidth, and efficiency [1]. The *gain* of an antenna describes the degree of radio energy concentration by an antenna in a direction, in reference to the isotropic antenna that radiates the energy uniformly in all the directions. The *radiation pattern* of an antenna describes the radiation power intensity distribution along all the directions. The *return loss* of an antenna indicates the degree of the energy reflected by the antenna due to the impedance mismatch between the antenna and a transmitter or a receiver. The *bandwidth* is defined as the frequency range within which the performance of the antenna, with respect to some characteristics (e.g., return loss or gain), conforms to a specified standard. *Efficiency* takes into account the energy loss due to imperfections of conductors and substrates used.

Figure 2. Millimeter-wave horn antennas (courtesy of QuinStar Technology Inc).

In general, millimeter-wave (mm) antennas can be categorized into five groups in terms of their structures and configurations: (1) the traditional reflector and lens antennas, (2) waveguide-based antennas, (3) printed-circuit antennas, (4) active integrated antennas, and (5) optically controlled and integrated antennas. Figure 3 lists the categories and the various antenna types under each category.

It should be noted that Fig. 3 is simply a general presentation of the techniques used so far in developing millimeter-wave antennas. The groupings presented are not absolutely clearcut among various millimeter-wave antennas reported so far. An antenna may belong to two or three groups simultaneously. For instance, an active integrated antenna may also belong to the group of printed-circuit antennas as it may be fabricated on a planar structure.

2. REFLECTOR AND LENS ANTENNAS

The traditional reflector and lens antennas are still being used for millimeter-wave applications because of their simplicity in operational principles and constructions. A typical reflector, illustrated in Fig. 4, consists of a feed that illuminates the conducting reflector with radio millimeter waves. The reflector surface is shaped in such a way that the radio fields scattered by the reflector will illuminate the area in a desired pattern. For instance, a parabolic surface will focus the radio energy in one direction, while the others will focus the radio energy in a certain coverage area. A shaped reflector antenna for 60-GHz indoor wireless networks has been reported [2]. It achieved a very good circular coverage with edge illumination less

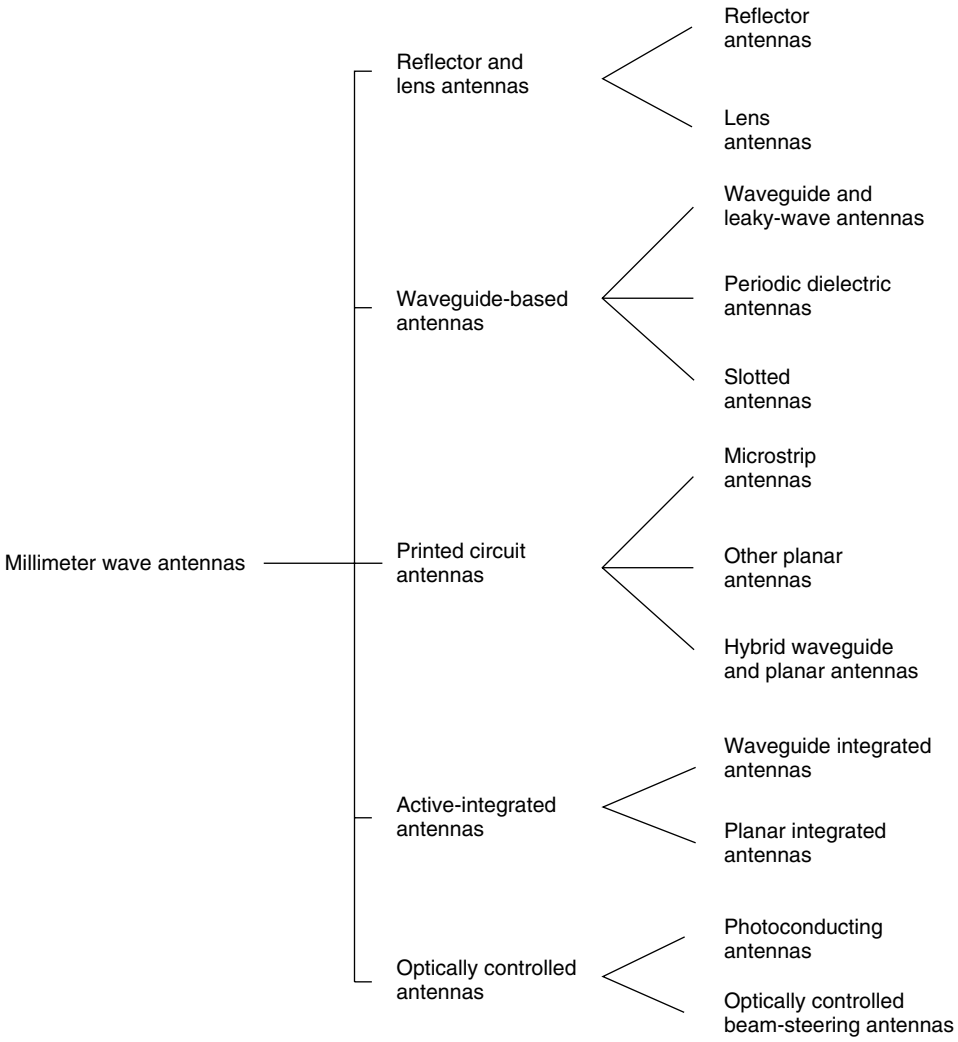


Figure 3. Division of millimeter-wave antennas in terms of structure.

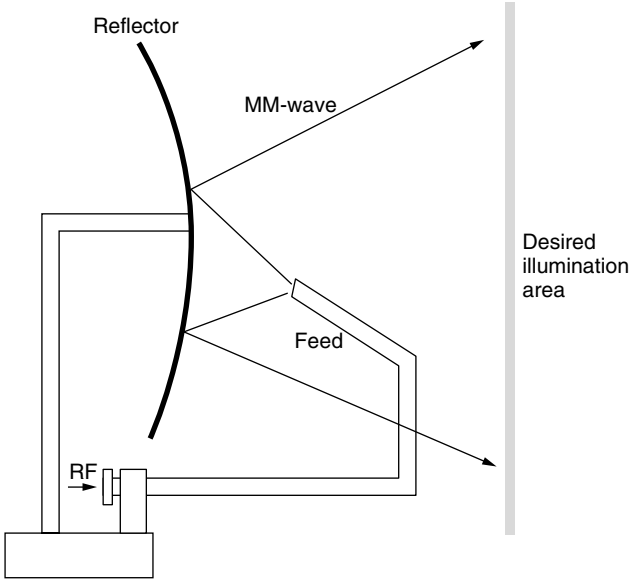


Figure 4. A shaped reflector antenna (the shape of the reflector is designed in such a way that the waves reflected by the reflector only illuminate the area desired).

than −10 dB of the desired boresight illumination. The size of the antenna is, however, rather bulky. The reflector has a diameter of 30 cm.

A millimeter-wave lens antenna is illustrated in Fig. 5. The lens is formed by a low-loss dielectric, and its surface is designed in such a way that the waves coming from the radiating patch will be diffracted at the lens–air interface into the air in the desired angle. When multiple radiating elements are placed at the different locations, a multibeam antenna can be achieved. It was reported that such an antenna achieved a directivity of 25.9 dB at 30 GHz with a beamwidth of 6.6° [3]. The diameter of the lens is ∼50 mm (∼5 cm), and the height of the whole structure is ∼80 mm (∼8 cm).

3. WAVEGUIDE-BASED ANTENNAS

As their name implies, waveguides are devices that guide and transmit microwave and millimeter (mm)-wave energy from one point to another. Traditional waveguides include coaxial, rectangular and circular waveguides, which are simply hollow metal tubes with cross-sections of two concentric circles, a rectangle and a circle (see

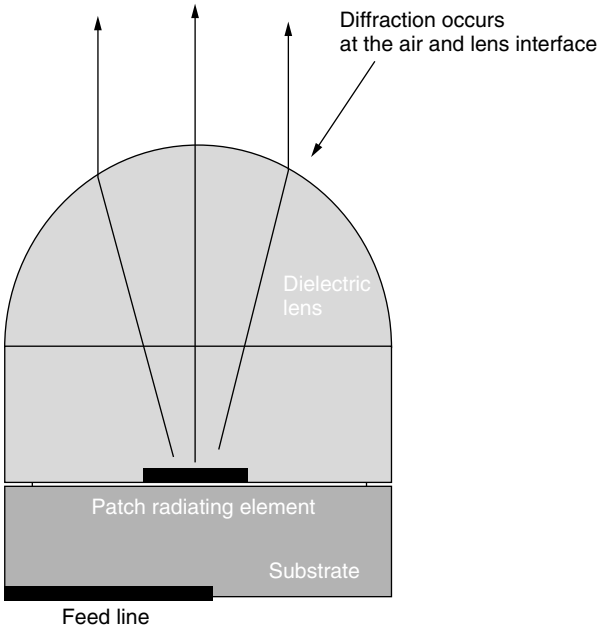


Figure 5. A lens millimeter-wave antenna (the shape of the lens is designed in such a way that the fields diffracted at the lens-air interface will radiate in the desired angles) (redrawn from Fig. 1 of Ref. 3, © 2001 IEEE).

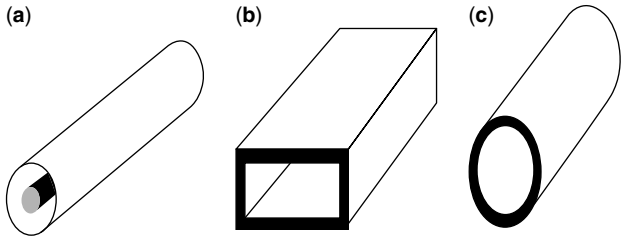


Figure 6. Traditional waveguides that are made of hollow metal tubes with different cross-sectional shapes sections: (**a**) coaxial line; (**b**) rectangular waveguide; (**c**) circular waveguide.

Fig. 6). The more modern waveguides, or guided-wave structures, are mostly planar structures that facilitate the integration with integrated circuits. They include striplines, microstrip lines, and coplanar lines as shown in Fig. 7. It has been proved theoretically and experimentally that when the shapes of the cross sections are uniform along the guides (i.e., do not vary along the longitudinal direction), microwave and mm-wave energy will propagate without incurring much energy leakage and radiation. However, when the uniformity of the cross sections is perturbed, energy leakage or radiation will occur. A simple example is to leave a waveguide abruptly cut open to air or free space. The microwave and mm-wave energy will then radiate into space. Consequently, by intentionally introducing perturbations in the wavequides along their longitudinal directions, radiation into space can be achieved in a controlled and desired manner. A specific advantage of these waveguide-based antennas is their compatibility with the waveguides from which they derived the energy, thus facilitating integrated designs with the waveguide structures. A few of these types of antennas are introduced below.

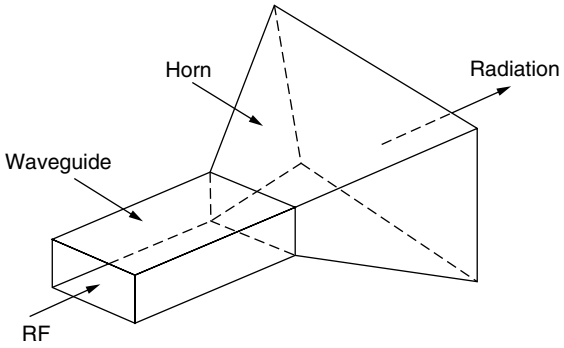


Figure 8. A pyramidal horn antenna.

3.1. Waveguide-Derived Antenna

The first type of antenna is the *waveguide-derived antennas*, where the waveguide structures are perturbed at their ends. They include horn antennas and leaky-wave antennas.

In a *horn antenna*, a longitudinally uniform waveguide is made with an open end. To ensure that most of energy in a waveguide radiates efficiently out into space, a transition from the waveguide to the open end is required. A typical transition is a horn-shape extension that transforms a small aperture of the waveguide to a large aperture (see Fig. 8). The radiation pattern of such an antenna is normally end-fire type. The gain of a horn antenna runs from 20 to 40 dB. The efficiency in light of the conducting loss and spillover energy can be as high as 85%. The length of the horn is less than 10 cm, and the width and height are less than 8 cm.

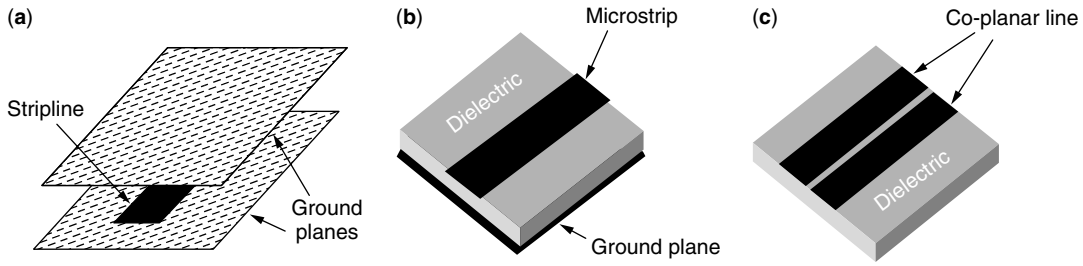


Figure 7. The planar guided wave structures with all the metal strips are planar: (**a**) striplines; (**b**) microstrip line; (**c**) coplanar lines.

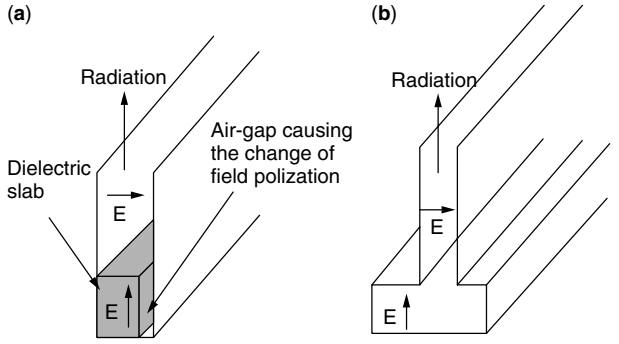


Figure 9. Two leaky-wave antennas (fields in the main guides are perturbed so that the fields in the upper arms propagate without cutoff and then radiate): (**a**) NRD guide antenna; (**b**) special groove guide antenna (redrawn from Fig. 2 of Ref. 18, © 1992 IEEE).

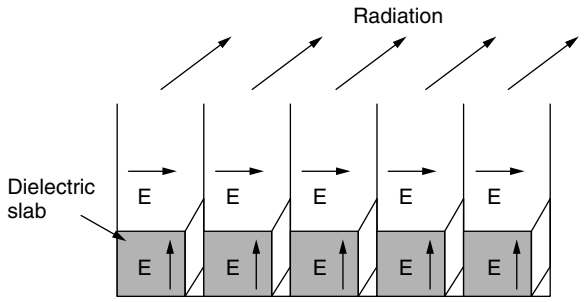


Figure 10. Array of leaky-wave NRD guide antennas (redrawn from Fig. 4 of Ref. 18, © 1992 IEEE).

Figure 9 shows two typical leaky-wave waveguide antennas, one with the use of a nonradiative dielectric (NRD) waveguide and the other with a special groove waveguide [4]. Both waveguides are bisected horizontally relative to their unperturbed waveguides with extended upper arms and closed at the bottom. Horizontally polarized fields will be produced in the foreshortening upper arms as a result of diffraction of the fundamental modes around the asymmetric air gaps or open aperture. The fields will then propagate without cutoff frequencies in a transverse electromagnetic parallel-plate mode to the end of the arms and radiate outward. Figure 10 shows an array of NRD waveguide antennas. It is typically 10–50 wavelengths long. By varying the phases to the feed systems for the antenna elements, the radiation beam can be steered in the longitudinal plane.

3.2. Periodic Dielectric Antenna

The second type of waveguide-based antenna is the *periodic dielectric antenna* [5–7], which normally consists of a uniform dielectric waveguide with a periodic surface perturbation that may take the form of a dielectric grating or a metal grating (see Fig. 11). The structure is designed in such a way that the fundamental mode is excited in the nonperturbed section of the waveguide and then transformed, as a result of the grating in the perturbed area, into a leaky wave that radiates into space. It has been shown theoretically and experimentally that as the frequency is increased, the main-beam direction scans from backfire, through broadside and into the forward quadrant. There exists, however, a possible stopband where an internal resonance occurs as a result of the periodic structures. Such a resonance will inhibit radiation. Fortunately, in most cases, such a stopband is narrow.

3.3. Slotted Waveguide Antenna

The third type of a waveguide-based antenna is a *slotted waveguide antenna*, where slots are opened on the sidewalls of the waveguide (see Fig. 12). As the waves propagate down the waveguide, fields radiate through the slots out into space. Since the number of slots can be large, the gain can be made as high as 35 dB with efficiency of 75% at 60 GHz [8].

4. PRINTED-CIRCUIT ANTENNAS

Microstrip and other printed-circuit antennas are planar types of antennas that are fabricated on multiplayer dielectric substrate structures. They are simple in structure, easy to fabricate by lithography, and convenient for circuit integration. Most of them are low-profile, lightweight, and low-cost devices potentially conformal to any planar surfaces. The most commonly seen are microstrip patch/dipole antennas, planar slotted antennas, and hybrid waveguide–planar antennas.

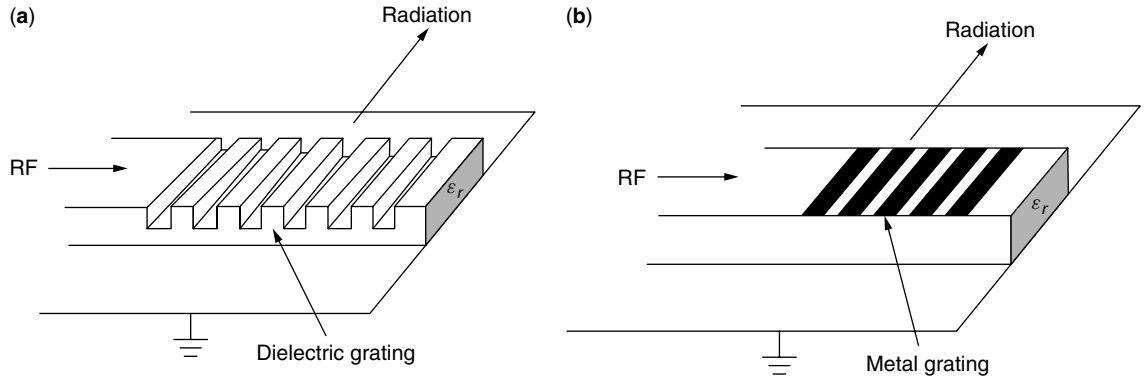


Figure 11. The periodic dielectric antennas (redrawn from Fig. 1 of Ref. 18, © 1992 IEEE).

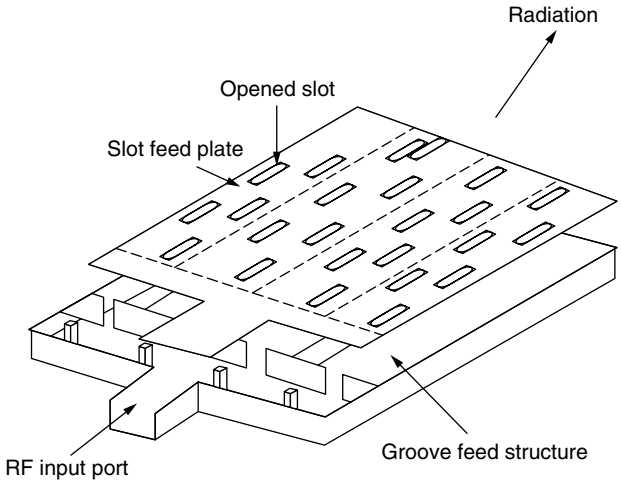


Figure 12. A single-layer waveguide slot antenna array (redrawn from Fig. 3 of Ref. 8, © 1997 IEEE).

The *microstrip patch antennas* are perhaps the simplest antennas, formed either by a patch or a planar dipole (see Fig. 13). The patch or the planar dipole serves as the radiating element. In general, the bandwidth of a microstrip antenna is quite narrow (e.g., <5%) because most of the fields are trapped between the patches and the ground planes. There are two other problems peculiar to mm-wave applications: fabrication tolerance and losses primarily associated with the feed systems.

Because of the small wavelengths, absolute fabrication tolerance for fabricating mm-wave lines is very small. For instance, for the width of a feedline, the order of a few tenths of a millimeter is required for operating frequencies of 30–100 GHz. In addition, conducting losses are relatively high and efficiency is potentially low because of the high frequencies of millimeter waves and the small cross section of a feedline. All these factors have limited the application of microstrip antennas to narrowband applications up to frequencies of 140 GHz [9,10].

To resolve the difficulties encountered in the microstrip antennas, other planar antennas have been proposed. One of them is the top-loaded coplanar waveguide fed aperture stacked patch antenna (Fig. 14), in which the co-planar transmission line is used as the feedline but is coupled to the radiation elements by means of aperture coupling. Two radiating patches resonating at two neighboring frequencies are stacked on top of each other with dielectric substrates in between. As a result, the bandwidth is increased from <1% to >15% [11].

A combination of waveguide structures with printed-circuit antennas was also been proposed. One of them is a two-dimensional printed dipoles with dual polarizations suspended in pyramidal horns (see Fig. 15). The horn is etched into a silicon wafer structure and the dipoles are printed by photolithographic techniques. The radiation characteristics of the antennas are determined by both the horn structures and the dipoles. Such a structure facilitates the integration with planar circuits while maintaining the efficiency and high gain of the horn antennas [12].

5. ACTIVE INTEGRATED ANTENNAS

The term *active integrated antenna* (AIA) refers to a class of radiating structures where the radiating elements are not only used for radiation but also serve as integral parts of active components such as resonators and filters. As a result, conventional 50-Ω feedline connections between radiating elements and active components and subsequent impedance matching elements are no longer necessary. In other words, both radiating elements and active components can be integrated and fabricated on the same substrate or multiplayer substrates. This leads to the obvious advantages of compactness, reliability, reproducibility, and potential low-cost.

A variety of AIA structures have been reported. It is difficult to comprehensively review all these antennas. Nevertheless, an attempt is made here to divide AIA structures into two groups in terms of the structures and configurations: waveguide-derived structures and planar structures.

Figure 16 shows a monolithic integrated waveguide single-slot mixer on a GaAs substrate mounted in a TE_{10} waveguide and a horn. RF signals received by the horn will be coupled to the GaAs active diode chip for signal

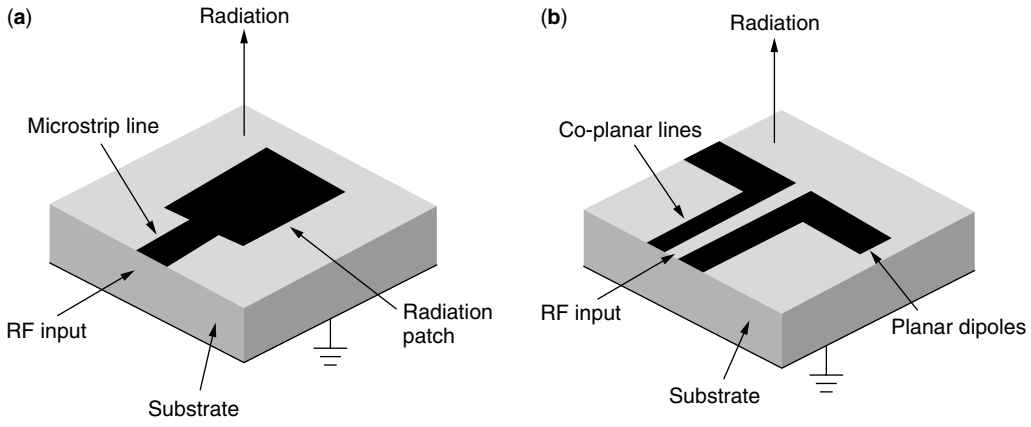


Figure 13. Microstrip patch (**a**) and planar dipole (**b**) antennas.

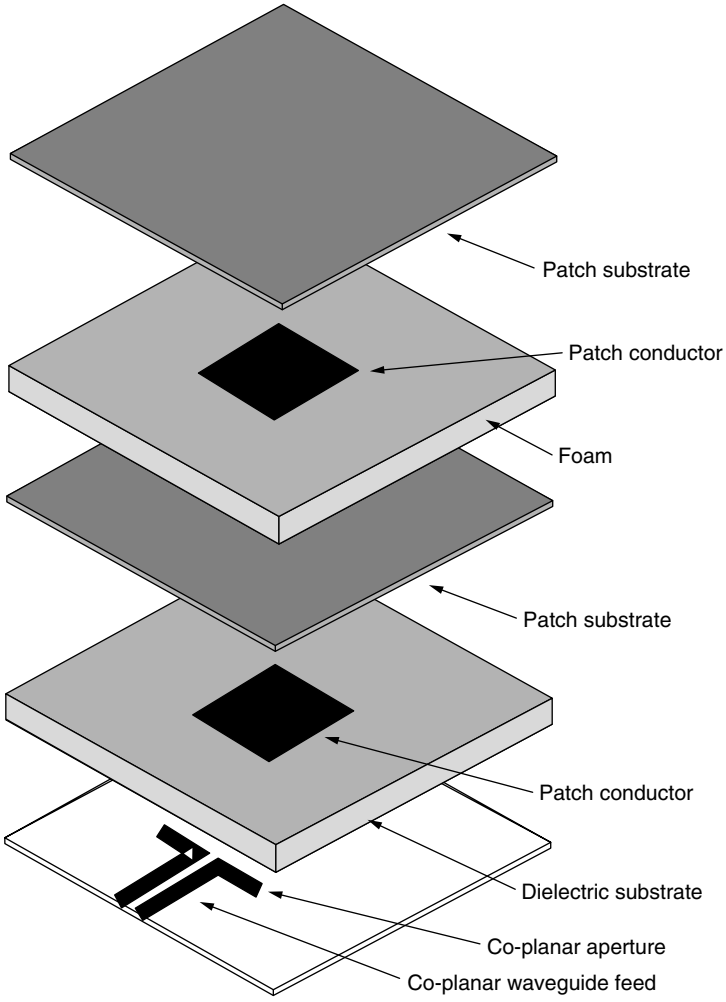


Figure 14. The stacked microstrip patch antenna (the radiating patch resonates at neighboring frequencies to increase the bandwidth) (redrawn from Fig. 2 of Ref. 11, © 2000 IEEE).

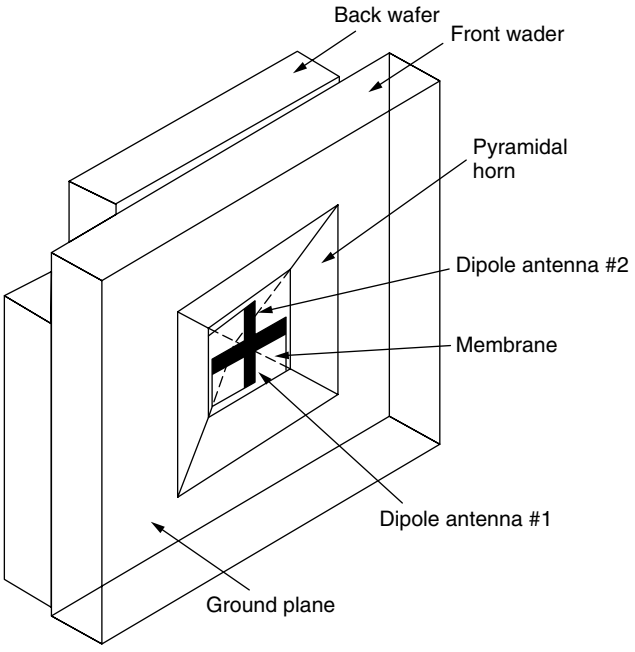


Figure 15. An integrated horn with dual polarizations for integrated balanced mixers (redrawn from Fig. 30 of Ref. 19, © 1992 IEEE).

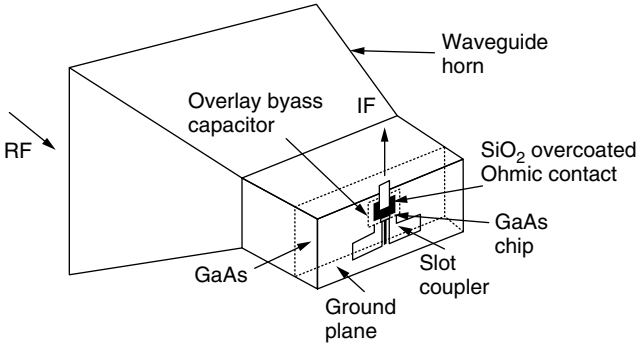


Figure 16. Monolithic integrated circuit single-slot mixer on a GaAs substrate in a TE_{10} waveguide (redrawn from Fig. 6 of Ref. 19, © 1992 IEEE).

mixings. The IF signal is directly output from the back of the waveguide horn [13].

Figure 17 illustrates a conceptually typical planar active integrated antenna. The radiating elements are a periodic, linear, series-fed microstrip patch array that also serves as the resonator to the oscillator. The field-effect transistor (FET) is used to improve the DC-to-RF conversion efficiency. The power output of the FET is

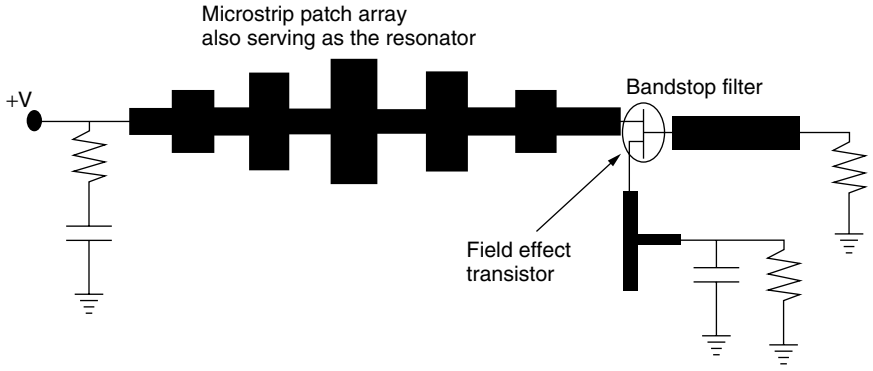


Figure 17. Periodic microstrip patch array with integrated FET (redrawn from Fig. 9 of Ref. 18, © 1992 IEEE).

delivered to the antenna that radiates in the broadside direction. The gate of the integrated FET is terminated in a bandstop filter that provides the correct reactance at the oscillating frequency [14]. A slight modification with additions of RF input circuitry could lead to a transceiver circuit, where the FET performs the dual functions, as the source for the transmitted signal and as self-oscillating mixer for downconversion of the received signal.

6. OPTICALLY CONTROLLED ANTENNAS

By integrating optical circuits and components with conventional millimeter-wave structures, optically controlled millimeter-wave antennas have been developed since the early 1990s. In these antenna structures, the properties of mm-wave radiations are controlled by lasers or optical signals. There are two groups of such antennas: (1) photoconducting antennas that result in generating millimeter waves and (2) optically controlled beam-steering antennas.

In the first case (see Fig. 18) the antenna consists of a planar dipole, a photoconductor deposited in between the feed gap of the dipole, coplanar strip transmission line, and contact pads for photoconductor biasing. The antenna is excited by illuminating the photoconductor with optical pulses, and the millimeter wave is generated as the result of the illumination and dipole resonance. By modulating the bias applied to the photoconductor, modulated millimeter waves can also be obtained [15].

In the second case, the antenna array is developed for beam scanning. The phases of RF signals fed to each array element are controlled with an optical means, either with photoconductors or with optical wavelength-dependent time-delay dispersive structures. Two examples are shown in Figs. 19 and 20. In Fig. 19, the semiconductor slab is illuminated with a special pattern formed by photomasks, creating a photoinduced plasma grating on the slab (that behaves similarly to a metal grating). The millimeter waves that couple to the slab through the dielectric waveguide will then interact with the plasma grating and radiate out of the slab in a specific direction. The direction is dependent on the grating pattern that is controlled by the photomasks [16]. In Fig. 20, millimeter-wave signals are first modulated onto an optical signal and split into four modulated optical signals. These four signals propagate through an optically controlled dispersive prism and thus have different time delays. They are demodulated with photodiodes and fed to antenna arrays, leading to beam steering that is dependent on the phase delays of the four signals. It was reported that such an antenna achieved squint-free steering across $\pm 60^{\circ}$ azimuthal span and over the entire Ka band (26.5–40 GHz) [17].

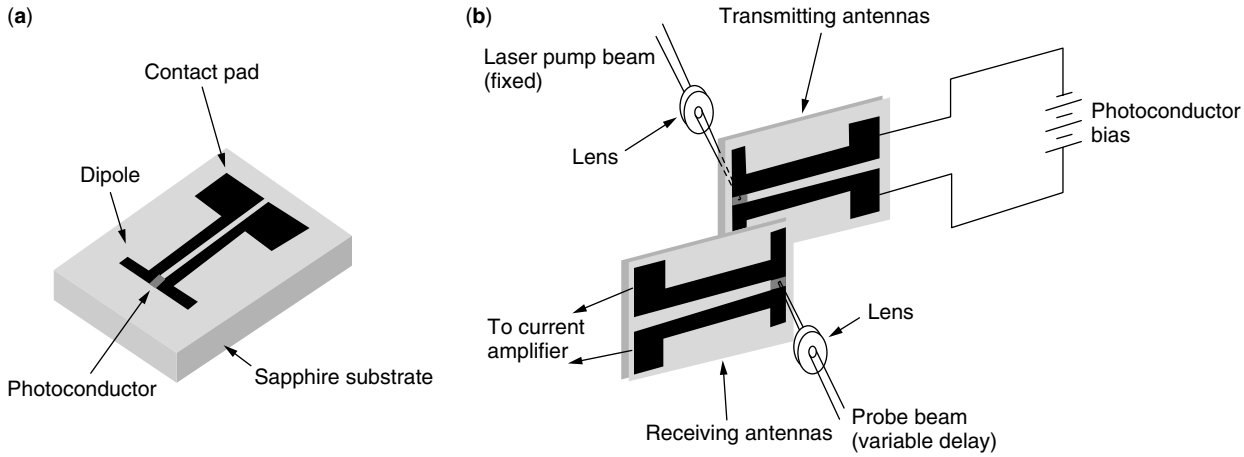


Figure 18. Photoconducting antenna structure (**a**) for generating electric short pulses that contain (**b**) millimeter-wave and submillimeter-wave components (redrawn from Figs. 1 and 2 of Ref. 15, © 1988 IEEE).

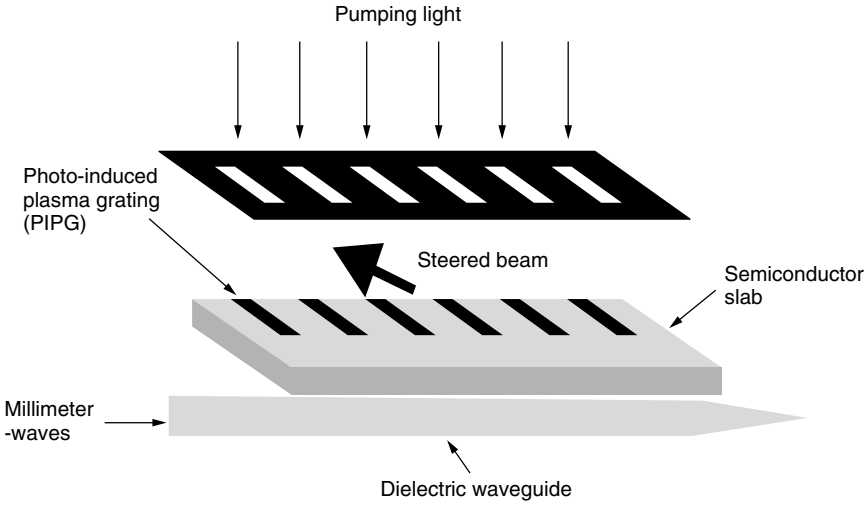


Figure 19. The beam-steering antenna with photoinducing plasma grating (redrawn from Fig. 1 of Ref. 16, © 1997 IEEE).

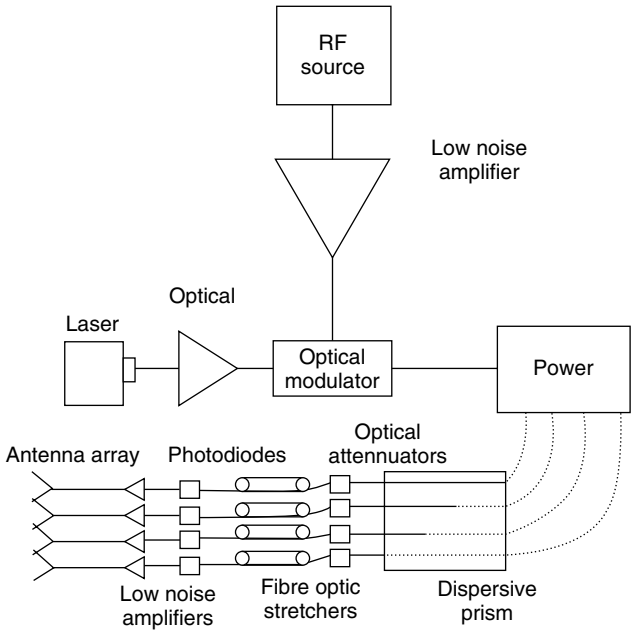


Figure 20. The fiberoptic beam-steering antenna (redrawn from Fig. 1 of Ref. 17, © 2001 IEEE).

7. FUTURE TRENDS OF MILLIMETER-WAVE ANTENNAS

Because of their high frequencies, millimeter waves are a promising medium for broadband and high-speed wireless communications. Their short wavelength permits circuits and systems to be small and compact. Their near-light properties allow for precise imaging and remote sensing in different environments such as clouds and damp air. As a result, millimeter-wave circuits and systems have been investigated and studied extensively since the early 1980s, including millimeter-wave antennas.

As described above, tremendous progress in the understanding and fabricating millimeter-wave antennas have been achieved [18,19]. Numerous antenna structures have been developed to meet different application requirements and to achieve better performance. In particular, active integrated antennas and optically controlled antennas have been attracting much attention.

Nevertheless, like the design of other millimeter-wave circuits and systems, the main challenges associated with the design of the millimeter-wave antennas are (1) low gains and low powers offered by active components, (2) relatively high conductor losses and substrate losses, and (3) the cost of components and systems. All these factors have the limiting effects in applications of millimeter-wave circuits and systems. It is expected that these three challenges will also continue to be the topics of interest of future research and development of millimeter-wave antennas.

BIOGRAPHY

Zhizhang (David) Chen received his B. Eng. degree in radio engineering in 1982 from Fuzhou University, P. R. China, his M. A. Sc. degree in radio engineering in 1986 from Southeast University, P. R. China, and his Ph.D. degree in electrical engineering in 1992 from the University of Ottawa, Ottawa, Ontario, Canada. He was a lecturer with the Department of Radio Engineering, Fuzhou University from 1985 to 1988, and has held a Natural Science and Engineering Research Council postdoctoral fellowship with the Department of Electrical and Computer Engineering, McGill University, Montreal, Québec, Canada, from January of 1993 to August of 1993. Since September of 1993, he has been with the Department of Engineering, Dalhousie University, (formerly Technical University of Nova Scotia), Halifax, Nova Scotia, Canada, where he is presently an associate professor. Dr. Chen has published over 100 refereed journal/conference papers and industrial reports in the areas of RF/microwave circuit and system design and computational electromagnetics. His general research areas are in RF/microwave engineering and applied electromagnetics for communications and microelectronics. His current teaching and research interests include RF/microwave CAD, RF interconnection & packaging, antenna design, numerical modeling and simulation, and wireless circuit and system design.

BIBLIOGRAPHY

1. C. A. Ballanis, *Antenna Theory*, 2nd ed., Wiley, 1997.
2. P. F. M. Smulder, S. Khusial, and M. H. A. J. Herben, A shaped reflector antenna for 60 GHz indoor wireless LAN access points, *IEEE Trans. Vehic. Technol.* **50**(2): 584–592 (March 2001).
3. X. Wu, G. V. Eleftheriades, and T. E. V. Deventer-Perkins, Design and characterization of single- and multiple-beam mm-wave circularly polarized substrate lens antennas for wireless communications, *IEEE Trans. Microwave Theory Tech.* **49**(3): 431–441 (March 2001).
4. F. K. Schwering and A. A. Oliner, Millimeter-wave antennas, in K. Chang, ed., *Handbook of Microwave and Optical Components*, Wiley, New York, 1988, Chap. 12.2.
5. T. Itoh and B. Adelseck, Trapped image guide leaky-wave antennas for millimeter-wave applications, *IEEE Trans. Antennas Propag.* **AP-30**(5): 505–509 (May 1982).
6. F. Schwering and S. T. Peng, Design of dielectric grating antennas for millimeter-wave applications, *IEEE Trans. Microwave Theory Tech.* **MTT-31**: 199–209 (Feb. 1983).
7. M. Guglielmi and A. A. Oliner, A practical theory for image guide leaky-wave antennas loaded by periodic metal strips, *Proc. 17th Eur. Microwave Conf.*, Rome, Italy, Sept. 11–17, 1987, pp. 549–554.
8. M. Ando and J. Hirokawa, Novel single high-gain and high-efficiency single-layer slotted waveguide arrays in 60 GHz band, *Proc. 10th Int. Conf. Antennas and Propagation*, Edinburg, UK, April 14–17, 1997, pp. 464–668.
9. F. K. Schwering and A. A. Oliner, Millimeter-wave antennas, in Y. T. Lo and S. W. Lee, eds., *Antenna Handbook*, Van Nostrand Reinhold, New York, 1988, Chap. 17.
10. M. A. Weiss, Microwave antennas for millimeter waves, *IEEE Trans. Antennas Propag.* **AP-29**: 171–174 (Jan. 1981).
11. W. S. T. Rowe and R. B. Waterhouse, Comparison of broad-band millimeter-wave antenna structures for MMIC and optical device integration, *Digest of 2000 IEEE Int. Antennas and Propagation Symp.*, Salt Lake City, UT, 2000, pp. 1390–1393.
12. W. Y. Ali-Ahmad and G. M. Rebeiz, 94 GHz integrated horn monopulse antennas, *IEEE Trans. Antennas Propag.* **39**(7): 820–825 (July 1991).
13. B. J. Clifton, G. D. Alley, R. A. Murphy, and I. H. Mroczkowski, High-performance quasioptical GaAs monolithic mixers at 110 GHz, *IEEE Trans. Electron Devices* **28**: 135–157 (Feb. 1981).
14. J. Birkeland and T. Itoh, FET-based planar circuits for quasi-optical sources and transceivers, *IEEE Trans. Microwave Theory Tech.* **37**(9): 1452–1459 (Sept. 1989).
15. P. R. Smith, D. H. Auston, and M. C. Nuss, Subpicosecond photoconducting dipole antennas, *IEEE J. Quant. Electron.* **24**(2): 255–260 (Feb. 1988).
16. V. A. Manasson, L. S. Sadovnik, V. A. Yepishin, and D. Marker, An optically controlled MMW beam-steering antenna based on a novel architecture, *IEEE Trans. Microwave Theory Tech.* **45**(8): 1497–1500 (Aug. 1997).
17. D. A. Tulchinsky and P. J. Matthews, Ultrawide-band fiber-optic control of a millimeter-wave transmit beamformer, *IEEE Trans. Microwave Theory Tech.* **49**(7): 1248–1253 (July 2001).
18. F. K. Schwering, Millimeter-wave antennas, *Proc. IEEE* **80**(1): 92–102 (Jan. 1992).
19. G. M. Rebeiz, Millimeter-wave and terahertz integrated circuit antennas, *Proc. IEEE* **80**(11): 1748–1770 (Nov. 1992).

MILLIMETER WAVE PROPAGATION

EDWARD E. ALTSHULER
Electromagnetics Technology Division
Hanscom AFB, Massachusetts

1. INTRODUCTION

The millimeter wave region of the electromagnetic spectrum generally covers wavelengths in the range from about 2 cm down to 1 mm (15–300 GHz). These limits are based on wavelength and on the nature and magnitude of the interaction between the wave and the atmosphere. The propagation characteristics of electromagnetic waves in this region are of particular interest because the waves have a strong interaction with lower atmospheric gases and particulates. Although the interaction with the atmosphere does not change abruptly at these limits, it does become weaker at longer wavelengths and stronger at shorter wavelengths. Thus the concepts presented here can generally be extended to either slightly longer or slightly shorter wavelengths. We shall often refer to the "window regions" of the millimeter wave spectrum. These are considered the low-attenuation regions between the gaseous absorption resonances. In particular, wavelengths between the 1.35-cm water vapor resonance and the 5-mm oxygen resonance, the 5- and 2.5-mm oxygen resonances, and the 2.5-mm oxygen resonance and 1.6-mm water vapor resonance compose the window regions. Low-attenuation regions also exist at wavelengths longer than 1.35 cm and shorter than 1.6 mm.

The physics of the interaction between millimeter waves and the atmosphere is extremely complex, so many facets are considered beyond the scope of this article. However, an effort is made to provide the reader with a general understanding of the mechanisms of this interaction; if in-depth details are required, they can be obtained from the references. Likewise, complicated mathematical expressions are used only to illustrate concepts that are considered important and cannot be satisfactorily explained otherwise. Many of the figures contain information that is "typical" or "average", that is, it is intended to provide the reader with an estimate of the magnitude of an interaction. More quantitative results are available from the cited references. In this article we first review the physics of the interaction of millimeter waves with the atmosphere and then describe the effects of the atmosphere on both terrestrial and earth–space communications.

1.1. Propagation Effects

Atmospheric gases and particulates often have a profound effect on millimeter waves and thus limit the performance

of many millimeter wave systems. Because the densities of these gases and particulates generally decrease with altitude, the effects of the atmosphere on the propagated wave are strongest very close to the earth and tend to diminish at higher altitudes. For this reason millimeter waves propagating above the tropopause — the altitude at which the temperature remains essentially constant with increasing height (~10–12 km) — are assumed to be unaffected. For the clear atmosphere there are essentially two types of interaction. The stronger interaction is the absorption–emission produced by oxygen and water vapor. The other interaction takes place with the refractivity structure of the atmosphere, which is often divided into two categories: gross structure and fine structure. For the gross structure it is assumed that the atmosphere is a horizontally stratified continuum characterized by a refractivity that normally decreases slowly with increasing altitude and is wavelength-independent; that is, it affects all wavelengths from microwaves through millimeter waves in the same way. For the refractivity fine structure, the atmosphere is viewed as an inhomogeneous medium consisting of small pockets of refractive index that vary both temporally and spatially. Because these pockets have different sizes, this interaction is wavelength-dependent. Although the terrain is not actually part of the atmosphere, it does interface with the atmosphere and can affect millimeter wave propagation. Thus, multipath propagation produced by the terrain and diffraction by prominent obstacles on the earth's surface are also reviewed. Atmospheric particulates range in size from micrometers to close to 1 cm in diameter. For particles very small compared to wavelength, the only significant propagation effects are those of absorption and emission. As the particles become larger with respect to wavelength, scattering effects become pronounced. In Section 2 we shall review the effects of the clear atmosphere and then of atmospheric particulates on millimeter wave propagation. We show that although the effects of the clear atmosphere are generally not as severe as those due to atmospheric particulates, they cannot be disregarded, even in the window regions and especially at short millimeter wavelengths.

1.2. Applications

Potential applications of millimeter waves have been considered for many years [1]. However, for most applications, atmospheric effects have always imposed limitations on system performance. The principal applications of millimeter waves have been in the areas of communications and radar. Probably the first application for which millimeter waves were considered was communications. The discovery of the circular electric waveguide mode TE_{01}, in the late 1930s prompted the use of these wavelengths for a waveguide communication system, because waveguide attenuation for that mode decreases with decreasing wavelength. It is believed that the most significant contribution resulting from this effort was not so much the development of the system itself but rather the research that was directed toward a whole new line of millimeter wave equipment and techniques required for this application namely, sources, amplifiers, detectors, and waveguide components. Through the years consideration was often given to utilizing millimeter waves for point-to-point communications. However, proposed systems never materialized, principally because they were not economically competitive with those at longer wavelengths; they also lacked reliability because of atmospheric effects and inferior components.

The need for new types of communication systems and the need to alleviate increasing spectrum congestion finally led to a reappraisal of millimeter waves. The availability of large bandwidths makes this region of the spectrum particularly attractive for high-data-rate earth–space communication channels. Furthermore, high-gain, high-resolution antennas of moderate size and lightweight compact system components are indeed applicable for space vehicle instrumentation. Millimeter waves provide an excellent means for obtaining secure communication channels. For satellite–satellite links, where all propagation is above the absorptive constituents of the lower atmosphere, narrow-beamwidth antennas may be operated at a wavelength where atmospheric attenuation is very high (i.e., $\lambda \sim 5$ mm); thus the signal is confined by the antenna to a narrow cone and then absorbed by the lower atmosphere before it reaches the earth. Another application for secure communications is ship-to-ship and short terrestrial links; in these cases attenuation is sufficiently high at millimeter waves to allow a detectable signal only over short distances. Finally, point-to-point radio-relay systems that were previously not considered feasible are now in operation. More recent studies have shown that attenuation in the lower atmosphere can be combatted using very short hops and diversity techniques; also, satisfactory system performance can now be obtained with solid-state components quite economically.

Because the beamwidth of an aperture is inversely proportional to wavelength, antennas having the same size aperture, have better resolution at millimeter wavelengths than at longer wavelengths. Furthermore, because range resolution is a function of bandwidth, improved range resolution is also possible at the shorter wavelengths.

2. ATMOSPHERIC EFFECTS ON PROPAGATED WAVES

Atmospheric gases and particulates may severely alter the properties of millimeter waves. For the clear atmosphere the most pronounced effect is absorption due to the gases, oxygen, and water vapor. However, refraction, scattering, diffraction, and depolarization effects are also reviewed, and it is shown that under special conditions their impact on the propagated wave can be significant. For an atmosphere containing particulates, the absorption and scattering by rain seriously limit propagation at millimeter wavelengths. However, the effects of smaller, water-based particulates are also significant, particularly at shorter millimeter wavelengths.

2.1. Clear Atmosphere

We shall consider an atmosphere "clear" if it is free of atmospheric particulates. Thus the only interaction that takes place is that between the propagated wave and

the atmospheric gases (and terrain). We shall see that for very low elevation angles the gross structure of the refractive index causes the propagated wave to be bent and delayed, whereas the fine structure of the refractive index scatters the propagated wave and may produce scintillations. The gases, oxygen, and water vapor absorb energy from the wave and reradiate this energy in the form of noise. Finally, if the wave is propagating close to the earth's surface, part of the energy may be reflected from the surface, or it may be diffracted by prominent obstacles on the surface and then interfere with the direct signal. If some form of reflection or scattering takes place, the propagated wave may also become depolarized. All of these effects are discussed in detail in Sections 2.1.1–2.1.5.

2.1.1. Refraction. The lower atmosphere is composed of about 78% nitrogen, 21% oxygen, and 1% water vapor, argon, carbon dioxide, and other rare gases. The densities of all these gases, except for water vapor, decrease gradually with height; the density of water vapor, on the other hand, is highly variable. The index of refraction of the atmosphere is a function of the temperature, pressure, and partial pressure of water vapor. Because the refractive index, n, is only about 1.0003 at the earth's surface, it is often expressed in terms of a refractivity N, where

$$N = (n - 1) \times 10^6 \tag{1}$$

The radio refractivity can be approximated by the following theoretical expression, which has empirically derived coefficients [2]

$$N = \frac{77.6P}{T} + \frac{3.73 \times 10^5 e}{T^2} \tag{2}$$

where P is the atmospheric pressure (in millibars), T the absolute temperature (in degrees kelvin), and e is the water vapor pressure (in millibars). This expression consists of two terms; the first is often referred to as the "dry" term, and the second as the "wet" term, because it is a function of water vapor. Equation (2) neglects dispersion effects and in principle does not hold near absorption lines; however, in practice it is assumed valid down to a wavelength of ~3 mm and is often considered acceptable down to wavelengths of even 1 mm, particularly in the window regions. The refractivity decreases approximately exponentially with height and is often expressed as

$$N(h) = N_s e^{-bh} \tag{3}$$

where h is the height above sea level (in kilometers); N_s, the surface refractivity; $b = 0.136\ \text{km}^{-1}$; $N(h)$ is the refractivity at height h (in kilometers).

As mentioned earlier, the gross behavior of atmospheric refractivity affects millimeter waves in much the same way as microwaves; because effects such as bending, time delay, and ducting are amply treated elsewhere [3], they will only be summarized here. The refractive index fine-scale structure is too complex to be treated by simple refraction theory. Because of the variability in the scale sizes of the refractive inhomogeneities, the effects are wavelength-dependent, and thus the interaction is different for millimeter waves than for microwaves.

2.1.1.1. Bending. The angular bending is due primarily to the change of the index of refraction of the atmosphere with height. Because the refractivity generally decreases with height, the wave passing obliquely through the atmosphere is bent downward. Thus the apparent elevation angle of a target tends to appear slightly higher than its true elevation angle, and this difference is called the *angle error*. For a horizontally stratified atmosphere, the angle error is zero at zenith and increases very slowly with decreasing elevation angle. This error becomes appreciable at low elevation angles, and for a standard atmosphere it approaches a value of about 0.7° near the horizon. For illustration we plot in Fig. 1 the angle errors of targets at altitudes of 90 km and infinity (radio source) for an atmosphere with a surface refractivity of $313N$ units that decreases exponentially with height. For a typical atmosphere the refractivity decreases at a rate of about $40N$ units/km near the surface and then more slowly at higher altitudes; this produces superrefraction. As the gradient of this decrease in refractivity increases, the wave is bent more toward the earth's surface, and when the gradient decreases at a rate of $157N$ units/km, the wave travels parallel to the surface; this condition is called *ducting*. For still steeper gradients the wave will actually be bent into the earth.

There are times when the decrease in refractivity is less than $40N$ units/km and the wave is bent less than normal toward the earth; this is called *subrefraction*. In actuality we have a curved ray passing over a curved earth. It is sometimes easier to visualize a straight ray over a curved earth (or a curved ray over a flat earth).

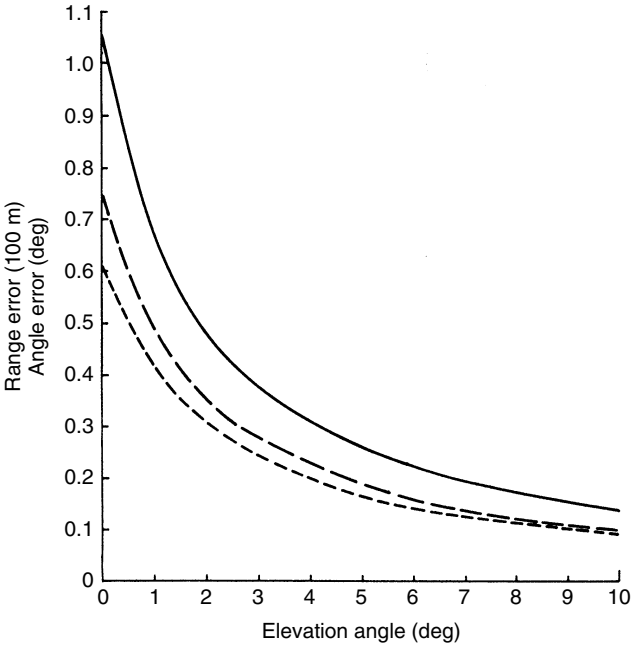


Figure 1. Typical tropospheric refraction angle and range errors:———, range error; — —, angle error ($h = \infty$);- - - - , angle error ($h = 90$ km).

For a vertical negative gradient of $40N$ units/km, it can be shown that an earth with an effective radius about four-thirds that of the true earth with the wave propagating in a straight line is equivalent to the true case — the curved ray path over the curved earth. The ratio of the effective earth radius to the true earth radius is often designated k. If the refractivity gradient is less than $40N$ units/km, then k decreases and reaches unity for the case of no gradient. As the gradient becomes more positive, k approaches zero. When the negative gradient is more negative than $40N$ units/km, k is larger than $\frac{4}{3}$ and approaches infinity for the special case of ducting, for which the gradient reaches $157N$ units/km. For still larger negative gradients, k becomes negative, because the flat earth for $k = \infty$ becomes curved in the opposite sense.

2.1.1.2. Time Delay. Time delay occurs primarily because the index of refraction of the atmosphere is greater than unity, thus slowing down the wave, and to a lesser extent because of the lengthening of the path by angular bending. For navigation systems, the range is determined from time-delay measurements; thus the additional time delay produced by the troposphere results in a corresponding range error. This error causes the target to appear farther away than its true distance. Navigation systems such as the Global Positioning System (GPS) must correct for this range error [4]. For a typical atmosphere the range error is slightly larger than 2 m in the zenith direction and increases very slowly with decreasing elevation angle. The range error becomes much larger at lower elevation angles, and for a standard atmosphere it approaches a value of about 100 m near the horizon. For illustration, we plot in Fig. 1 the range error for an atmosphere with a surface refractivity of $313N$ units that decreases exponentially with height.

2.1.1.3. Refraction Corrections. As seen in Fig. 1, the angle and range errors become appreciable for very low elevation angles. It has been shown that both errors are strongly correlated with the surface refractivity, and for many applications adequate corrections based on a linear regression on N are possible [3,4]. More accurate corrections can be obtained by actually measuring the vertical refractivity profile and then calculating the corrections. The principal limitation of this approach is that a horizontally stratified atmosphere is usually assumed, and this is not always valid for the long distances traversed at low elevation angles. It has been shown that the range error is correlated with the brightness temperature of the atmosphere, and techniques to take advantage of this dependence have been proposed [5–8]. One of the most effective methods for obtaining angle error corrections involves the use of "targets of opportunity." These may be either calibration satellites or radio sources, the angular positions of which are normally known to an accuracy of the order of microradians. In principle, the angular error of the calibration source is measured before the target is tracked. If the target is in the same general direction as the calibration source and the atmospheric refractivity does not change appreciably with time, then the correction can be determined directly. For range error corrections, differential GPS uses a known location on the surface as a reference.

2.1.1.4. Scintillations. So far we have discussed the effects of the gross refractivity structure of the atmosphere on the propagated wave. The atmosphere also has a fine-scale refractive index structure, which varies both temporally and spatially and thus causes the amplitude and phase of a wave to fluctuate; these fluctuations are often referred to as *scintillations* [9–11]. The refractive index structure is envisioned to consist of pockets of refractive inhomogeneities that are sometimes referred to as "turbulent eddies" and may be classified by size into three regions: the input range, the inertial subrange, and the dissipation range. The two boundaries that separate these regions are the outer and inner scales of turbulence, L_0 and l_0, respectively. These are the largest and smallest distances for which the fluctuations in the index of refraction are correlated. A meteorologic explanation of how these pockets are generated is beyond the scope of this article; however, in simple terms, large parcels of refractivity, possibly of the order of hundreds of meters in extent, continually break down into smaller-scale pockets. These pockets become smaller and smaller until they finally disappear. The very large pockets in the input range have a complex structure, and at the present time there is no acceptable formulation of the turbulence properties of this region. Pockets having scale sizes of less than $\sim$1 mm have essentially no turbulent activity, and for all practical purposes the spectrum of the covariance function of the refractive index fluctuations, $\phi_n(k)$, equals zero. The inertial subrange bounded by L_0 and l_0 has a spectrum

$$\varphi_n(k) = 0.033\, C_n^2\, k^{-11/3} \tag{4}$$

for $2\pi/L_0 < k < 2\pi/l_0$, where C_n is the structure constant and k the wavenumber (not to be confused with the ratio k of the effective earth radius to the true earth radius used earlier). The phase fluctuations arise from changes in the velocity of the wave as it passes through pockets of different refractive indices. As the wavelength becomes shorter, the changes in phase increase proportionally. The amplitude fluctuations arise from defocusing and focusing by the curvature of the pockets.

2.1.2. Absorption and Emission. Atmospheric gases can absorb energy from millimeter waves if the molecular structure of the gas is such that the individual molecules possess electric or magnetic dipole moments. It is known from quantum theory that, at specific wavelengths, energy from the wave is transferred to the molecule, causing it to rise to a higher energy level; if the gas is in thermal equilibrium, it will then reradiate this energy isotropically as a random process, thus falling back to its prior energy state. Because the incident wave has a preferred direction and the emitted energy is isotropic, the net result is a loss of energy from the beam. The emission characteristics of the atmosphere may be represented by those of a blackbody at a temperature that produces the same emission; therefore the atmospheric emission is often expressed as an apparent sky temperature.

Because absorption and emission are dependent on the same general laws of thermodynamics, both are expressed in terms of the absorption coefficient. Using Kirchhoff's law and the principle of conservation of energy, one can derive the radiative transfer equation, which describes the radiation field in the atmosphere that absorbs and emits energy. This emission is expressed as

$$T_a = \int_0^\infty T(s)\gamma(s) \exp\left(-\int_0^\infty \gamma(s')\,ds'\right) ds \qquad (5)$$

where T_a is the effective antenna temperature, $T(s)$ is the atmospheric temperature, $\gamma(s)$ is the absorption coefficient, and s is the distance from the antenna (ray path). In simpler terms

$$T_a = T_m(1 - e^{-\gamma s}) \qquad (6)$$

where T_m is the atmospheric mean absorption temperature within the antenna beam. Solving for the attenuation, we obtain

$$A = \gamma s = 10 \log\left(\frac{T_m}{T_m - T_a}\right) \qquad (7)$$

where A is in decibels. The only atmospheric gases with strong absorption lines at millimeter wavelengths are water vapor and oxygen. The absorption lines O_3, CO, N_2O, NO_2, and CH_2O are much too weak to affect propagation in this region.

2.1.2.1. Water Vapor. The water vapor molecule has an electric dipole moment with resonances at wavelengths of 13.49, 1.64, and 0.92 mm (22.24, 183.31, and 325.5 GHz) in the millimeter wave region. In general, the positions, intensities, and linewidths of these resonances agree well with experimental data. There are, however, serious discrepancies between theoretical and experimental absorption coefficients in the window regions between these strong lines; experimental attenuations are often a factor of 2–3 times larger than theoretical values.

Although the cause of the discrepancy is not known, indications are that either the lineshapes do not predict enough absorption in the wings of the resonances or there is an additional source of absorption that has not yet been identified. It should be mentioned that there are over 1800 water vapor lines in the millimeter wave/infrared spectrum, 28 of which are at wavelengths above 0.3 mm. Because the wings of these lines contribute to the absorption in the window regions, very small errors in the lineshapes could significantly affect the overall absorption. In an effort to overcome this problem, several workers have introduced an empirical correction term to account for the excess attenuation [12]. In addition to the uncertainty of the absorption coefficient of water vapor, there is also the problem of water vapor concentration. The amount of water vapor in the lower atmosphere is highly variable in time and altitude and has densities ranging from a fraction of a gram per cubic meter for very arid climates to 30 g/m^3 for hot and humid regions; for this reason it is very difficult to model. A plot of the water vapor absorption as a function of frequency is shown in Fig. 2 for a density of 7.5 g/m^3. Because the attenuation, α, is linearly proportional to the water vapor density, except for very high concentrations, attenuations for other water vapor densities are easily obtained.

2.1.2.2. Oxygen. The oxygen molecule has a magnetic dipole moment with a cluster of resonances near a wavelength of 5 mm (60 GHz) and a single resonance at 2.53 mm (118.75 GHz). Although the >30 lines near a wavelength of 5 mm are resolvable at low pressures (high altitudes), they appear as a single pressure-broadened line near sea level owing to a large number of molecular collisions. Even though the magnetic dipole moment of oxygen is approximately two orders of magnitude weaker than the electric dipole moment of water vapor, the net absorption due to oxygen is still very high, simply because it is so abundant. The fact that the distribution of oxygen throughout the atmosphere is very stable makes it easy to model. A plot of oxygen attenuation as a function of frequency is shown in Fig. 2 along with that of water vapor. Note the importance of water vapor attenuation at very short wavelengths.

2.1.3. Scattering. The principal effect of the pockets of refractive index on the propagated wave is to produce scintillations, as described in Section 2.1.1. When the pockets are of the order of a wavelength in size, they can also scatter the signal. At millimeter wavelengths and for line-of-sight paths, this scattered field is generally very weak compared to the direct signal and is not considered

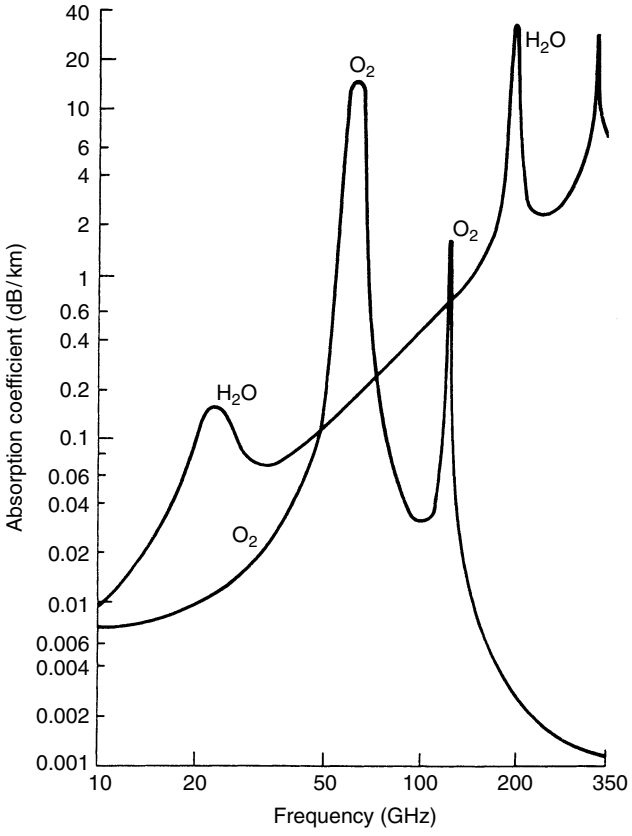


Figure 2. Absorption coefficients for water vapor and oxygen.

significant. Electromagnetic waves scattered from the earth's surface may interfere with the direct signal; this is called *multipath propagation*. The extent of multipath is dependent on the geometry of the transmitter and receiver with respect to the surface, their respective beamwidths and polarizations, and the dielectric constant and surface roughness of the terrain. Let us first consider a surface that is relatively smooth with respect to wavelength. The reflection coefficients of vertically and horizontally polarized waves for a nonmagnetic surface are

$$\Gamma_v = \frac{\varepsilon \sin\alpha - (\varepsilon - \cos^2\alpha)^{1/2}}{\varepsilon \sin\alpha + (\varepsilon - \cos^2\alpha)^{1/2}} \tag{8}$$

$$\Gamma_h = \frac{\sin\alpha - (\varepsilon - \cos^2\alpha)^{1/2}}{\sin\alpha + (\varepsilon - \cos^2\alpha)^{1/2}} \tag{9}$$

where α is the grazing angle and $\varepsilon = \varepsilon' - j\varepsilon''$ is the complex dielectric constant. For very small grazing angles the magnitudes of the reflection coefficients approach unity and the phases approach 180°. As the grazing angle increases, the magnitude and phase of the vertically polarized wave fall off sharply and those of the horizontally polarized wave decrease very slightly. The sharp falloff of the vertically polarized reflection coefficient can be explained as follows. For most surfaces, with the exception of very dry ground, $|\varepsilon| \gg 1$. With this approximation, Eq. (8) can be rewritten as

$$\Gamma_v = \frac{(\varepsilon)^{1/2}\sin\alpha - 1}{(\varepsilon)^{1/2}\sin\alpha + 1} \tag{10}$$

Note that the numerator is $((\varepsilon)^{1/2}\sin\alpha - 1)$. Thus at some angle the numerator approaches zero; this is the Brewster angle, and if the terrain were a perfect dielectric [$(\varepsilon)^{1/2}$ is real], then the reflection coefficient would actually go to zero. As the grazing angle approaches normal incidence the vertical reflection coefficient increases and finally equals the horizontally polarized reflection coefficient at normal incidence. When the surface is relatively smooth, the scattering is predominantly specular; that is, it can be considered coherent. As the surface becomes rougher, a diffuse, incoherent component appears, and for a very rough surface the scattered signal is predominantly diffuse. The criterion usually applied for characterizing surface roughness is that introduced by Rayleigh. It is based on the phase difference of adjacent rays reflected from a rough surface; when the path difference between these rays increases to about 90°, the surface is assumed to transform from smooth to rough. Obviously this transition is very gradual and should be interpreted as such.

Mathematically the surface can be considered smooth when $h\sin\alpha < \frac{1}{8}\lambda$, where h is the height of a surface irregularity. It must be emphasized that even at millimeter wavelengths, for which λ is very small, typical surfaces tend to look smooth at very low grazing angles. We have reviewed the general characteristics of multipath propagation. Now let us summarize multipath propagation in the context of millimeter waves. At longer wavelengths the reflection coefficient of a vertically polarized wave is significantly lower than that of a horizontally polarized wave, particularly in the vicinity of the Brewster angle, so microwave systems are often designed to operate with vertical polarization to minimize multipath interference. At millimeter wavelengths this polarization dependence is of less importance, because the reflection coefficients of vertically and horizontally polarized waves are comparable for most millimeter wave applications. First, multipath effects at these short wavelengths will generally occur only at very small grazing angles, because most surfaces based on the Rayleigh criterion appear rough for larger grazing angles. Furthermore, because the dielectric constants of most surfaces tend to remain constant or decrease with decreasing wavelength, the Brewster angle increases and the reflection coefficient of the vertically polarized wave does not drop off as rapidly with increasing grazing angle. Therefore at millimeter wavelengths multipath is confined to much lower grazing angles than at microwave wavelengths and is thus less sensitive to polarization. As the surface becomes rough with respect to wavelength, the grazing angle must become very small to have specular reflection.

2.1.4. Diffraction. Electromagnetic waves incident on an obstacle may be bent around that obstacle; this is known as *diffraction*. The extent of diffraction is dependent on the shape and composition of the obstacle, its position with respect to the direct path of the incident wave, and the wavelength. Tradition diffraction theory has been used to treat simple shapes such as individual knife edges, rounded edges, and in some instances sets of these edges. An underlying assumption is that the knife edge is very sharp or the rounded edge very smooth with respect to wavelength. It is also often assumed that the edge is a perfect conductor, although solutions have been obtained for edges having finite conductivity. Diffraction loss is often expressed as a function of the dimensionless Fresnel parameter v, which is, in turn, a function of the geometric parameters of the obstacle and path. For knife-edge diffraction the Fresnel parameter can be defined as

$$v = h\left[\frac{2}{\lambda}\left(\frac{1}{d_1} + \frac{1}{d_2}\right)\right]^{1/2} \tag{11}$$

where h is either the height of the obstacle above the direct path or the distance of the obstacle below the path and d_1 and d_2 are the respective distances of transmitter and receiver from the knife edge. For illustration, let us assume that the knife edge is midway between transmitter and receiver; then $d_1 = d_2 = \frac{1}{2}d$ and

$$|v|^2 = \frac{8h^2}{\lambda d} \tag{12}$$

where v is positive when the ray path is below the edge and negative when the ray path is above the edge. It is known that the diffraction loss is approximately zero for $v < -3$ and very high for $v > 3$, so $-3 < v < 3$ can be considered the region of interest.

Equation (12) can be expressed as

$$h = (\tfrac{1}{8}\lambda d v^2)^{1/2} = \tfrac{1}{2}v(\tfrac{1}{2}\lambda d)^{1/2} \tag{13}$$

Because d is generally on the order of kilometers and λ on the order of millimeters, h can be only on the order of meters. Thus, from a practical standpoint only isolated obstacles such as small hills or buildings would produce diffraction effects at millimeter wavelengths.

2.1.5. Depolarization. Depolarization of an electromagnetic wave can occur when the incident wave is scattered and a cross-polarized component is produced along with the copolarized component. It is defined as

$$|\text{ depolarization }| = 20 \log \left(\frac{|E_x|}{|E_y|} \right) \tag{14}$$

where E_x and E_y are the cross-polarized and copolarized components, respectively, and the depolarization is in decibels. Olsen [13] has summarized in detail both the mechanisms that can produce depolarization during clear air conditions and some experimental observations of this phenomenon. He divides these mechanisms into two groups: those that are independent of the cross-polarized pattern of the antenna (a perfect plane-polarized wave) and those that are dependent on the cross-polarized pattern. In principle, depolarization can arise from scattering by refractive inhomogeneities or from terrain. For the plane-polarized wave it appears that depolarization due to refractive multipath is insignificant but that depolarization due to terrain multipath can be much stronger [13]. For an antenna having a measurable cross-polarized pattern, it is believed that both atmospheric and terrain multipath mechanisms contribute to depolarization of the wave. Although most experimental results of depolarization by the clear atmosphere have been obtained at centimeter wavelengths, there is no reason to believe that the same effects will not occur at millimeter wavelengths. However, because both atmospheric and terrain multipath may normally be weaker at millimeter wavelengths than at microwave wavelengths, the depolarization may not be as severe.

2.2. Atmospheric Particulate Effects

In this section we discuss the degrees of absorption, scattering, and depolarization that may occur from atmospheric particulates. We shall see that rain is by far the most important of the particulates, for two reasons: (1) the interaction of rain with millimeter waves is very strong and (2) rain occurs more often than do other particulates. Thus the interaction between rain and millimeter waves is discussed in detail.

2.2.1. Absorption and Scattering. Millimeter waves incident on atmospheric particulates undergo absorption and scattering; the degree of each is dependent on the size, shape, and complex dielectric constant of the particle and the wavelength and polarization of the wave. The following Mie expression can be used for calculating the absorption and scattering from a dielectric sphere:

$$Q_t = \frac{\lambda^2}{2\pi} \text{ Re } \sum_{n=1}^{\infty} (2n+1)(a_n^s + b_n^s) \tag{15}$$

where Q_t represents losses due to both absorption and scattering, and a_n^s and b_n^s are complicated spherical Bessel functions that correspond to the magnetic and electric modes of the particle, respectively. Q_t has the dimension of area and is usually expressed in square centimeters. Physically, if a wave with a flux density of S (W/cm^2) is incident on the particle, then $S \times Q_t$ is the power absorbed or scattered. When the circumference of the particle is very small compared to wavelength (i.e., $\pi D \ll \lambda$), then the scattering and absorption losses can be represented by

$$Q_s = \left(\frac{\lambda^2}{2\pi} \right) \left(\frac{4}{3\rho^6} \right) \left| \frac{n^2 - 1}{n^2 + 2} \right|^2 \tag{16}$$

and

$$Q_a = \left(\frac{\lambda^2}{2\pi} \right) (2\rho^3) \text{ Im } \left| -\frac{(n^2 - 1)}{n^2 + 2} \right| \tag{17}$$

where $\rho = kD/2 = \pi D/\lambda \ll 1$ and n is the complex index of refraction. Because ρ is very small, the loss due to scattering, which is proportional to ρ^6, will be much smaller than that due to absorption, which is proportional to ρ^3. This condition is often referred to as the Rayleigh approximation, for which

$$Q_s \propto \frac{1}{\lambda^4} \quad \text{and} \quad Q_a \propto \frac{1}{\lambda}$$

Because the scattering loss is often assumed negligible, the total loss is proportional to the volume of the drop. Often the backscatter cross section (or radar cross section) is of interest:

$$\sigma = \left(\frac{\lambda^2 \rho^6}{\pi} \right) \left| \frac{n^2 - 1}{n^2 + 2} \right|^2 = \frac{3}{2} Q_s \tag{18}$$

The relationship with Q_s arises because Rayleigh scatterers are assumed to have the directional properties of a short dipole and the directivity of a dipole in the backscatter direction is 1.5 times that of an isotropic source. As the drop becomes large with respect to wavelength, the Rayleigh approximation becomes less valid and the Mie formulation in Eq. (16) must be used.

2.2.2. Depolarization. Atmospheric particulates having a nonspherical shape will depolarize a wave (produce a cross-polarized component) if the major or minor axis of the particulate is not aligned with the E field of the incident wave. The extent of the depolarization is a strong function of the size, shape, orientation, and dielectric constant of the scatterer. The depolarization defined in Eq. (14) arises because the orthogonal components of the scattered field undergo different attenuations and phase shifts. These differences are referred to as the differential attenuation and differential phase shift. An alternative definition related to depolarization is the cross-polarization discrimination, which is simply the reciprocal of the depolarization. In general, the depolarization increases as the particulate size and eccentricity increase. The depolarization also increases as the angle between the E field of the incident wave and the major axis of the particulate increases up

to approximately 45°, for which the depolarization passes through a maximum.

2.2.3. Types of Particulate. Rain is the most common particulate; drops range in size from a fraction of a millimeter to about 7 mm. Sleet and snow, which are considered quasisolid forms of water, are then treated. Because of their complexity of shape and composition, only limited theoretical work has been done on them, and because they are rare events in most locations, only limited experimental data have been obtained. Hail is frozen water, and does not occur very often. However, the losses due to hail can be calculated quite accurately, and are very small at millimeter waves because the complex dielectric constant of ice is small. Cloud, fog, and haze particulates are very similar in that they are all composed of very small water droplets suspended in air (clouds may also contain ice crystals) with diameters ranging from several microns (μ) up to about 100 μm. Therefore, through most of the millimeter wave region the Rayleigh approximation is valid for these particulates. Dust and sand particulates have size distributions comparable to that of clouds, but because their complex dielectric constants are low, their interaction with millimeter waves is very weak.

2.2.3.1. Rain. Rain is an extremely complex phenomenon, both meteorologically and electromagnetically. From a meteorologic standpoint it is generally nonuniform in shape, size, orientation, temperature, and distribution, thus making it very difficult to model. Electromagnetically, the absorption, scattering, and depolarization characteristics can be calculated only for very simple shapes and distributions. However, theoretical results do provide a qualitative understanding of the effects of rain on millimeter waves, and when they are combined with experimental data, empirical parameters can be derived and more quantitative results are possible. Let us first examine the absorption and scattering characteristics of a single spherical raindrop. From Eq. (15) we can calculate the total cross section Q_t, which is the sum of the absorption cross section Q_a and the scattering cross section Q_s. This cross section is a strong function of the drop diameter and its complex index of refraction. At millimeter wavelengths both the real and imaginary parts decrease with decreasing wavelength, and we shall see that this is one of the reasons why the cross section of a drop eventually starts to decrease at shorter wavelengths. In Fig. 3, the total cross section of a drop is plotted as a function of drop diameter for several wavelengths. When the drop is very small with respect to wavelength, the Rayleigh approximation is valid; it is seen from Eqs. (16) and (17) that the scattering and absorption cross sections Q_s, and Q_a are proportional to $(D/\lambda)^6$ and $(D/\lambda)^3$, respectively. Because the loss due to scattering is negligible compared to that due to absorption, the total cross section is proportional to the volume of the drop. As the drop increases in size, both the scattering and absorption cross sections continue to increase, with the scattering cross section increasing more rapidly. Finally, the total cross section begins to level off, and would eventually approach a value of twice the geometric cross section of the drop when it is very large with respect to wavelength [14]. Thus, as the drop becomes larger, the cross section, which is initially proportional to the drop volume, becomes proportional to the drop area. The dependence of the cross section on wavelength is more complicated than that of size, because both the relative drop size and the complex index of refraction are changing. In Fig. 4, the cross sections are plotted as a function of wavelength for a number of drop radii. The cross sections increase with decreasing wavelength, reach a peak, and then start to decrease very slightly for still smaller wavelengths. This behavior can be explained by considering that although the cross section increases as the drop becomes larger with respect to wavelength, the real and imaginary components of the index of refraction decrease as the wavelength becomes smaller, and this decrease eventually causes the total cross section to decrease.

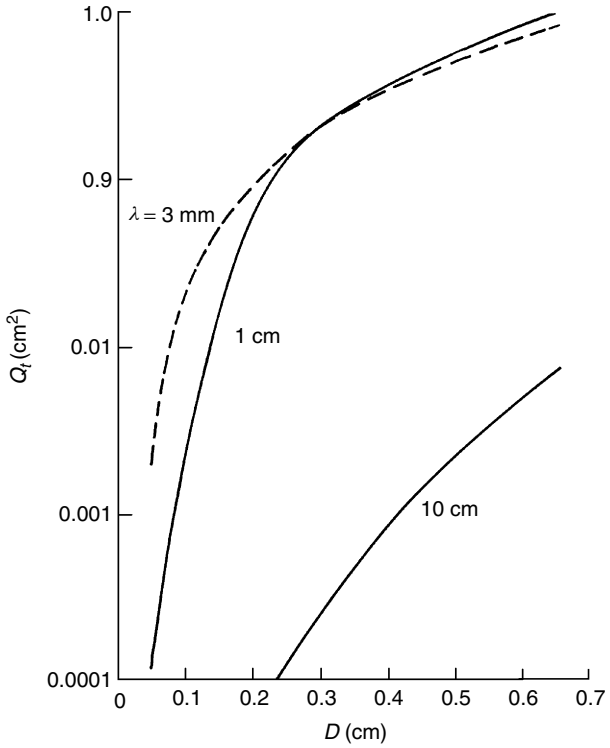


Figure 3. Total cross section Q_t of a raindrop as a function of drop diameter D.

We shall now consider the effect of the *shape* of raindrops on electromagnetic parameters. Whereas small drops tend to be spherical in shape, larger drops become oblate because of distortion due to air drag and are often modeled as oblate spheroids [15]. The cross section of an oblate drop is generally larger than that of a corresponding spherical drop having an equal volume of water [16]. The cross section is strongly dependent on the polarization of the wave, being larger when the polarization vector is aligned with the major axis of the spheroid and smaller when the polarization vector is aligned with the minor axis. We shall see that the most significant effect of a nonspherical drop is the depolarization it can produce.

We have reviewed the absorption and scattering characteristics of individual raindrops and found that they

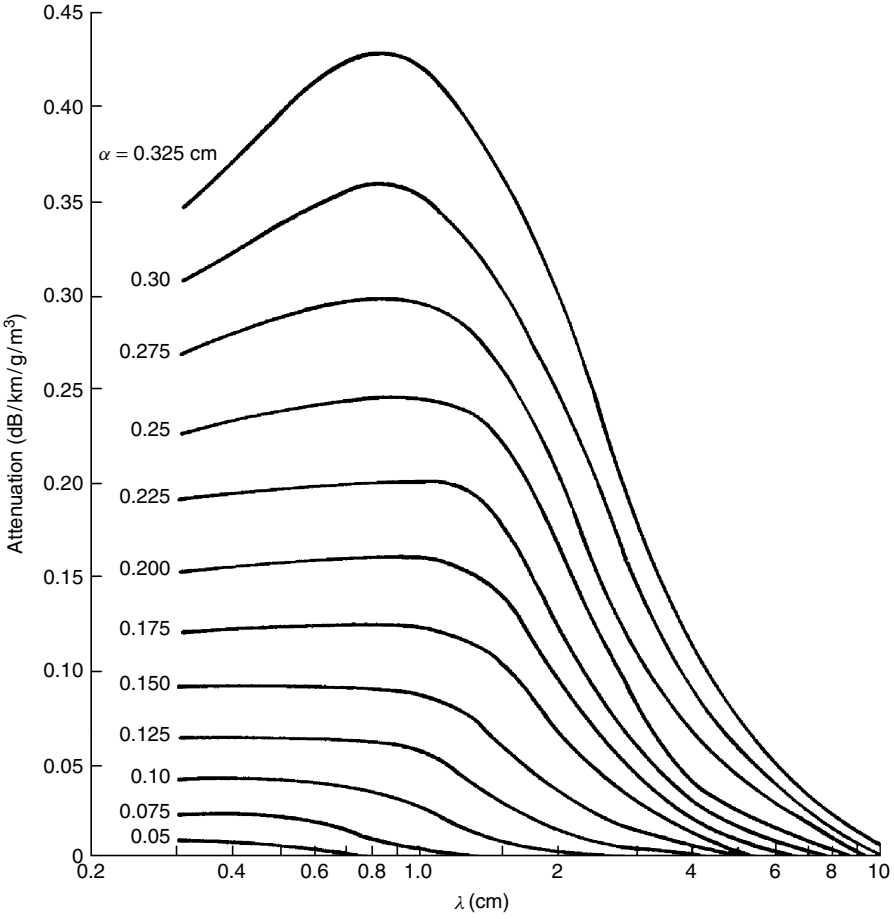


Figure 4. Theoretical values of attenuation by raindrops for various drop radii, expressed in decibels per kilometer per drop per cubic meter.

are a very complicated function of both the drop geometry and the index of refraction. Rain can be considered a collection of drops having diameters ranging from a fraction of a millimeter (mist) up to possibly 7 mm. To compute the attenuation of rain, the cross sections of the drops must be calculated and then summed. Because the characteristics of a precipitation system are controlled largely by the airflow, the net result is a collection of drops that is continually varying both spatially and temporally; it is thus very difficult to model. It has been found that the meteorologic parameter that is most easily measured and also most effectively characterizes rain is the rain rate. A number of investigators have shown that rain rate is correlated with drop size distribution; their results are summarized by Olsen et al. [17]. The attenuation can be expressed in the form

$$A = 0.4343 \int_0^\infty N(D) Q_t(D_t \lambda)\, dD \tag{19}$$

where $N(D)dD$ is the number of drops per cubic meter having diameters in the range dD, Q_t is the total cross section of each drop, and the attenuation is measured in decibels per kilometer. Rain is often assumed to have an exponential distribution of drop diameters, so that

$$N(D) = N_0\, e^{-\Lambda D}, \tag{20}$$

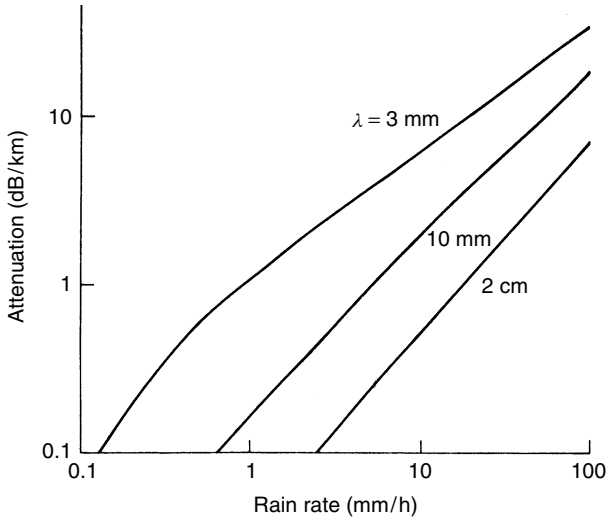


Figure 5. Rain attenuation as a function of rain rate.

where N_0 and Λ are empirical constants that are a function of the type of rain and more particularly the rain rate. Attenuations based on Medhurst's calculations [18] are plotted as a function of rain rate in Fig. 5. These curves can be approximated by

$$A = aR^b \tag{21}$$

where a and b are numerical constants that are a function of wavelength and type of rain, R is the rain rate in millimeters per hour, and the attenuation is measured in decibels per kilometer. Olsen et al. [17] calculated rain attenuation as a function of rain rate using the Mie formulation and then performed a logarithmic regression to obtain values for a and b. These values have been tabulated for frequencies from 1 to 1000 GHz, and although they are believed to provide a good approximation to the attenuation, it should be remembered that they are statistical and must be interpreted accordingly.

The depolarization characteristics of rain are very heavily dependent on the shape and orientation of the drops. Because light rain consists mostly of small drops and small drops tend to be spherical, depolarization effects are minimal. As the rain becomes heavier, the average drop size increases and the larger drops tend to become more oblate. The more oblate the drop, the larger the differential attenuation and phase between the orthogonal fields. However, it should be emphasized that these differentials do not in themselves produce depolarization; the incident field must also be tilted with respect to the axes of the drop. Because large oblates are easily canted by winds, their axes are seldom aligned with either horizontally or vertically polarized waves. Brussard [19] has shown that the canting angle of oblate raindrops is a function of the average drop diameter and vertical wind gradients. Typically the canting increases with drop size and levels off for drops on the order of a few millimeters in diameter. It naturally increases with wind speed and usually becomes smaller with increasing height. Because the differential attenuation is proportional to the total attenuation, it increases with rain rate. It also initially increases with shorter wavelengths, as does the total attenuation, but then reaches a peak and eventually starts to decrease at very short millimeter wavelengths, mostly because attenuation at very short wavelengths is produced primarily by the smaller drops and these tend to be spherical. The differential phase is affected mostly by the real part of the refractive index of water and decreases with shorter millimeter wavelengths. Thus, at millimeter wavelengths, differential attenuation is the dominant cause of depolarization

2.2.3.2. Sleet, Snow, and Hail. These very complex forms of precipitation have attenuation characteristics that vary markedly at millimeter wavelengths. We have seen that liquid water is a strong attenuator of millimeter waves; therefore sleet, which is a mixture of rain and snow, can also produce very high attenuations. In fact, these attenuations may exceed those of rain because nonspherical shapes have been shown to produce higher attenuations than equivalent spheres and sleet particulates are often very elongated [16]. The depolarization effects of sleet can be very strong if the flakes show a nonparallel preferential alignment with the E field. Wet snow has characteristics very similar to those of sleet. As the snow becomes drier, its composition approaches that of ice crystals, and because ice has a low imaginary index of refraction, the absorption is very small. Losses due to scattering are small at longer millimeter wavelengths but may become appreciable at shorter wavelengths if the flakes are large.

The effect of hail on millimeter waves is better understood than that of sleet or snow because there is less variability in its shape and composition. Because the imaginary part of the index of refraction of ice is about three orders of magnitude less than that of water, absorptive losses are negligible. The real part of the index of refraction is about one-fourth that of rain, so although scattering losses produced by hail are smaller than those of rain, they can be important, particularly at shorter millimeter wavelengths. If the hailstone is covered with even a very thin coat of water, its attenuation rises significantly and approaches that of a raindrop having an equivalent volume [20]. In summation, because it is extremely difficult to produce accurate models of sleet, snow, and hail, and because there are very few experimental attenuation data at millimeter wavelengths (or any other wavelengths), it is not possible to provide quantitative results. Although it is known that the absorption and scattering losses of sleet and wet snow are large and that the scattering losses of dry snow and hail are significant, there are presently no attenuation data for these particulates comparable to those for rain. However, from a practical standpoint, these particulates do not occur very often, and so their impact on millimeter wave systems is not considered critical.

2.2.3.3. Cloud, Fog, and Haze. Meteorologically, cloud, fog, and haze are very similar because they consist of small water droplets suspended in air. The droplets have diameters ranging from a fraction of a micron for fog and haze to over 100 μm for heavy fog and high-altitude clouds. Haze may be considered a light fog, and of course the principal difference between cloud and fog is that clouds exist at higher altitudes and often contain ice as well as water particulates.

Electromagnetically, cloud, fog, and haze can be treated identically. Because the droplets are very small with respect to wavelength, the Rayleigh approximation is valid, even at short millimeter wavelengths. Therefore, scattering losses are negligible and the attenuation, which is proportional to the volume of the drops, can be calculated from Eq. (17) and is seen to increase with decreasing wavelength. Because it is very difficult to measure fog or cloud water content, the attenuations produced by these particulates are not easily determined. Typical attenuations are plotted as a function of wavelength in Fig. 6 for several 'temperatures.

Fog is often characterized by visibility, which is much easier to measure than density; however, it should be emphasized that visibility is an optical parameter, a function of the scattering characteristics of the droplets, whereas the attenuation at millimeter wavelengths, as mentioned previously, is strictly a function of the fog density. Although there is a correlation between fog or cloud visibility and total liquid water content [21,22], this relationship must be used with caution because the correlation coefficient is strongly dependent on the type of fog. Fog is often divided into two types: *radiation fog*, which forms when the ground becomes cold at night and cools

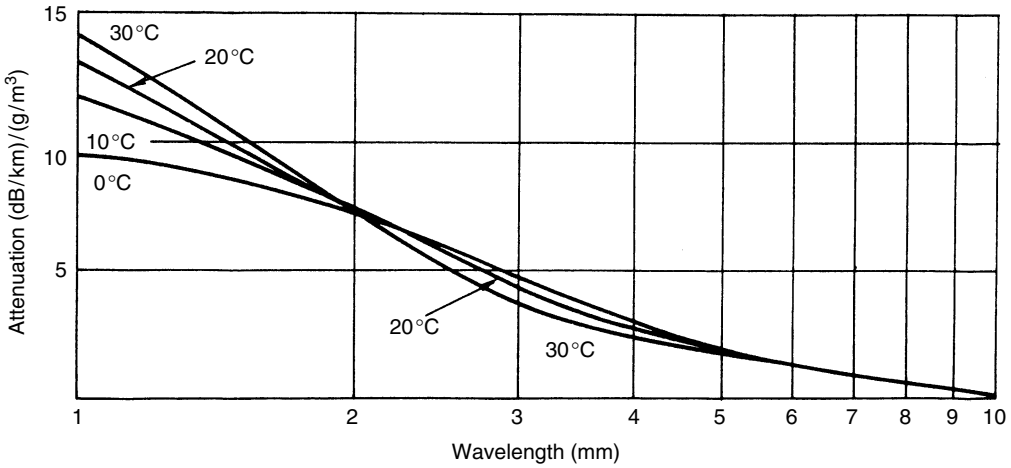


Figure 6. Attenuation of cloud and fog as a function of wavelength.

the adjacent air mass until it becomes supersaturated, and *advection fog*, which forms when warm, moist air moves across a cooler surface. The average drop size of an advection fog is usually larger than that of a radiation fog. Thus, if two fogs had the same liquid water densities but one consisted of relatively large drops, then that fog would have a much higher visibility. Another fog or cloud parameter not to be overlooked is temperature, because it has a strong influence on the complex index of refraction, particularly at longer millimeter wavelengths. Once again, as with snow and hail, the overall effects of cloud and fog are not as severe as those of rain.

2.2.3.4. Sand and Dust. Sand and dust are both fine-grained, quartz-type particles that have diameters of a fraction of a millimeter, densities on the order of 2600 kg/m^3, and a relative dielectric constant of about $2.5 - 0.025i$ [23]. The principal difference between sand and dust is that the average grain size of sand is larger. As a result, whereas larger windblown sand grains rise to a maximum height of only 2 m, heavy winds can carry fine dust particles to altitudes as high as 1 km. In general, the height to which the particles rise is proportional to the wind speed and inversely proportional to the particle size. Because sand and dust particulates are very small with respect to wavelength, even at millimeter wavelengths, scattering losses are negligible and the only losses are due to absorption. However, the imaginary part of the complex dielectric constant is very small; thus absorption losses can be considered minimal under naturally disturbed conditions. If a large amount of dirt or dust were to become suspended in the atmosphere as a result of an explosion, however, then significant attenuations might arise, particularly if the dust were moistened by the presence of water; these attenuations would last only for a duration of seconds.

3. TRANSMISSION PATHS

There are essentially two types of transmission paths: terrestrial and earth–space. We shall now take the results of Section 2 and apply them to systems that would use these propagation paths.

3.1. Terrestrial Line-of-Sight Paths

For terrestrial paths the atmospheric effects on propagation become more pronounced as the length of the path increases and the wavelength decreases. For short paths, attenuation is of principal concern; refraction and atmospheric multipath effects are unlikely, terrain multipath and diffraction problems arise only when transmitter and receiver are close to the surface, and depolarization and scintillations occur only under very extreme conditions of precipitation. For a clear atmosphere, gaseous absorption in the window regions is only a fraction of a decibel per kilometer at longer millimeter wavelengths but can become appreciable at shorter wavelengths. Sand and dust attenuations throughout the millimeter wave spectrum are well below 1 dB/km, except for conditions of a large cloud that may be produced by an explosion. Haze and fog attenuations increase with decreasing wavelength as shown in Fig. 6, and although they become appreciable, they are generally lower than water vapor absorption at very short wavelengths. Attenuation due to rain, sleet, and wet snow can be significant even for very short paths; this attenuation is lowest at longer wavelengths and gradually increases with decreasing wavelength, reaching a maximum at a wavelength of a few millimeters and then leveling off at still shorter wavelengths. As the path becomes longer, attenuation effects become more severe; in addition, all the other propagation effects mentioned in the previous paragraph are more likely to occur. The use of millimeter waves for applications requiring long terrestrial paths appears unlikely at this time because the attenuation would be prohibitive, except perhaps for a region having a relatively dry climate.

3.1.1. Attenuation. For many applications, particularly communications, it is important to be able to estimate the percentage of time that the path attenuation exceeds a certain value. To accomplish this one must first examine

the climate of the region of interest. If the absolute humidity is known, the gaseous absorption can be estimated from the expression

$$A = a + b\rho_0 - cT_0 \tag{22}$$

The coefficients are plotted in Fig. 7; ρ_0 is in grams per cubic meter and T_0 in degrees Celsius. This calculation is not very accurate at very short wavelengths, where there is a lack of agreement between the theoretical and experimental values of water vapor absorption. As mentioned in the previous paragraph, sand, dust, haze, and fog attenuations are low at longer wavelengths and generally small compared to water vapor absorption at very short wavelengths; only at wavelengths between about 2 and 3 mm is fog attenuation generally important.

Rain attenuation is by far the most serious propagation problem. Rain attenuation statistics are available only for some locations. A procedure for estimating rain attenuation for different climates has been outlined by Crane [24]. A global model representing typical rain climates throughout the world was developed, based on rain data provided by the World Meteorological Organization, and from this model the percentage of time that the point rain rate exceeds a particular value can be estimated. However, rain ordinarily consists of cells of different sizes and seldom is homogeneous in the horizontal plane. Thus, it is necessary to derive a path average rain rate from the point values. By pooling worldwide rain statistics it was possible to obtain an empirical relationship between point and path average rain rates; this has been done so far for distances up to 22.5 km. For low rain rates the rain is usually widespread; widespread rain, however, may contain convective cells having a high rain rate, so on an average the rain rate along the total path will be higher than that at a point. For high rain rates the rain tends to be localized, so the average rain rate for the total path would ordinarily be lower than that at a point. As expected, the correction factor is heavily dependent on the pathlength.

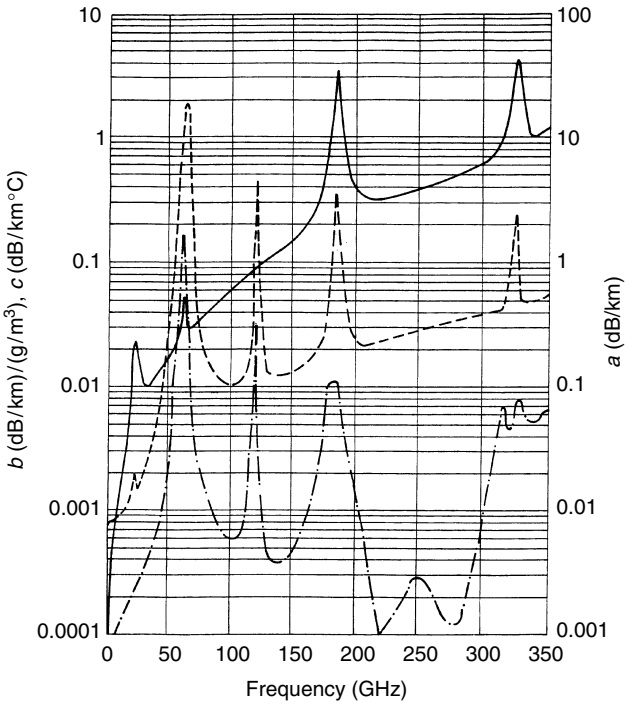


Figure 7. Coefficients for computing specific attenuation:- - - , a coefficient;———, b coefficient; —·—·, c coefficient.

In Section 2.2.3.1 the aR^b relationship for attenuation as a function of rain rate was introduced. By using the path average rain rate concept of the previous paragraph, it is possible to derive a correction term for the aR^b expression. From Crane [24], we have

$$A(R_p, D) = aR_p^b \left(\frac{e^{\beta d} - 1}{\mu\beta} - \frac{b^\beta e^{c\beta d}}{c\beta} + \frac{b^\beta e^{c\beta D}}{c\beta} \right) \qquad d \le D \le 22.5 \text{ km} \tag{23}$$

$$= aR_p^\beta \left(\frac{e^{u\beta D} - 1}{u\beta} \right) \qquad 0 < D \le d \tag{24}$$

where R_p is in millimeters per hour, D in kilometers, the specific attenuation $A(R_p, D)$ in decibels per kilometer, $\beta = 2\pi/\lambda$ and

$$u = \frac{\ln(be^{cd})}{d}, \qquad b = 2.3R_p^{-0.17}$$

$$c = 0.026 - 0.03 \ln R_p, \qquad d = 3.8 - 0.6 \ln R_p$$

For $D > 22.5$ km, the probability of occurrence P is replaced by a modified probability of occurrence

$$P' = \left(\frac{22.5}{D} \right) P \tag{25}$$

For example, suppose that for a particular climatic region the rain rate exceeds 28 mm/h 0.01% of the time and 41 mm/h 0.005% of the time ($R_{0.01} = 28$ mm/h, $R_{0.005} = 41$ mm/h). The attenuation along a 22.5-km path that would be exceeded 0.01% of the time can be calculated directly from Eq. (23) or (24). To determine the attenuation that would be exceeded 0.01% of the time over a 45-km path, a new percentage of time $P' = (22.5/45)P = 0.005\%$ would be used, with a corresponding rain rate of $R_{0.005} =$ 41 mm/h. Now the attenuation that would be exceeded 0.01% of the time for a 45-km path would be based on a rain rate of $R_{0.005} = 41$ mm/h for a 22.5-km path.

To improve the reliability of a terrestrial link, path diversity can be utilized. As mentioned previously, the heavy rain that has a severe impact on terrestrial link performance tends to be localized. By using redundant terminals, the probability of having a path free from heavy rain is increased. Ideally, the optimum separation of transmitter or receiver terminals (or both) is that for which the rain rates (and corresponding attenuations) for the pair of terminals are uncorrelated. This separation is a function of the climate and rain rate. Blomquist and Norbury [25] have studied diversity improvement for a number of paths with lengths of about 3–13 km and terminal separations of 4–12 km. On the basis of very limited data, they found that the diversity improvement increased as the terminal separation was increased from

4 to ~8 km; there was, however, no additional improvement as the separation was increased further. Because only limited statistical data on diversity advantage are presently available, it is not possible to determine optimal terminal spacing for most locations. However, there is sufficient evidence to indicate that a diversity mode of operation can significantly improve the performance of line-of-sight links.

3.1.2. Terrain Scatter and Diffraction. If the terminals of a line-of-sight path are close to the surface, then propagation losses due to multipath and diffraction are possible. These mechanisms were described in Sections 2.1.3.1 and 2.1.4, respectively, and it was seen that the multipath or diffracted signal can interfere with the direct signal; the net effect is a resultant signal that may vary in amplitude from zero intensity to twice the intensity of the direct signal. It should be emphasized that the interference is strongest when the multipath or diffracted signal is coherent. For multipath this occurs when the terrain is smooth with respect to wavelength; for diffraction it occurs from a prominent obstacle or a set of obstacles that happen to add constructively (or destructively). Regarding terrain multipath, even though most terrains have surface irregularities much larger than 1 mm, we see from Eqs. (8) and (9) that as the grazing angle becomes small, the surface becomes electromagnetically smooth (the reflection coefficient approaches unity) and large specular signals are possible. These signals interfere with the direct signal and can significantly degrade the performance of a line-of-sight system. Interference effects produced by a diffracted signal should not be as large as those produced by multipath, because the obstacles that would diffract the wave are not likely to produce a coherent signal. As mentioned previously, if the surface irregularities of a prominent obstacle are rough with respect to wavelength, there are many uncorrelated, diffracted rays and the resultant signal consists of a diffuse signal superimposed on a weak specular signal; if the surface is very rough, the specular component will disappear. Also, it is seen from Eq. (13) that the direct path must be within meters of the top of the obstacle for a diffracted signal to appear. However, diffracted signals are certainly possible at millimeter wavelengths under the "right" conditions.

3.1.3. Depolarization. As mentioned in Sections 2.1.5 and 2.2.2, depolarized signals may arise from either multipath or precipitation. A multipath ray obliquely incident on a paraboloidal receiving antenna can produce a cross-polarized component. Oblate raindrops canted with respect to the plane of the polarization vector of the incident wave also produce a cross-polarized component. The net effect is that the resultant signal is depolarized and the system performance compromised. Vander Vorst [26] has summarized the effects of depolarization on line-of-sight links. Often, depolarization produced by multipath may have a more severe effect on link performance than that produced by rain. The influence of rain caused only a very small degradation of the performance of a dual-polarized system with respect to a single-polarization system, whereas the same was not true for multipath.

3.1.4. Refraction and Atmospheric Multipath. Under normal atmospheric conditions an electromagnetic wave is bent toward the earth's surface. The amount of bending is proportional to the length of the path, so for long paths refractive bending corrections may be required, as discussed in Section 2.1.1.3. Under abnormal atmospheric conditions — for example, those producing a sharp negative gradient in the refractivity as a function of height — the wave may become trapped (ducting) or a multipath signal may be produced by the "layer" arising from the refractivity structure. Although refraction and multipath effects are possible in principle, they are not considered important as far as millimeter wave line-of-sight links are concerned, because, as mentioned previously, attenuation effects will normally prohibit the use of very long paths.

3.2. Earth–Space Paths

For earth–space paths, propagation effects become more severe with decreasing wavelength [27–29]. For elevation angles above about 6°, attenuation and emission from atmospheric gases and precipitation are of principal concern. In addition, backscatter and depolarization resulting from precipitation may also cause problems. For low elevation angles, all the problems associated with long terrestrial paths are present. The determination of propagation effects for slant paths is generally more difficult than for terrestrial paths. The modeling of a slant path under conditions of cloud and precipitation is particularly complicated, because the structure of the particulates is varying in both time and space. Experimentally, it is very expensive to place a millimeter wave beacon on a satellite or aircraft, so other means of obtaining attenuation information are sometimes used.

3.2.1. Attenuation and Emission. Atmospheric attenuation and emission are the most serious propagation problems for earth–space systems. Because the attenuation decreases the signal level and the emission (or effective noise temperature) sets a minimum noise level for the receiver, the only way to maintain the system signal-to-noise ratio is through increased transmitter power or a diversity mode of operation [30]. High powers are not readily available at millimeter wavelengths, and diversity systems are costly, so these options have their difficulties. For clear-sky conditions, the attenuation is a function of oxygen and water vapor density along the path. The vertical distributions of these gases are assumed to decrease exponentially with height and have scale heights of approximately 4 and 2 km, respectively. The zenith attenuation in decibels as a function of wavelength can be estimated from

$$A_{90^\circ} = \alpha + \beta\rho_0 - \xi T_0 \tag{26}$$

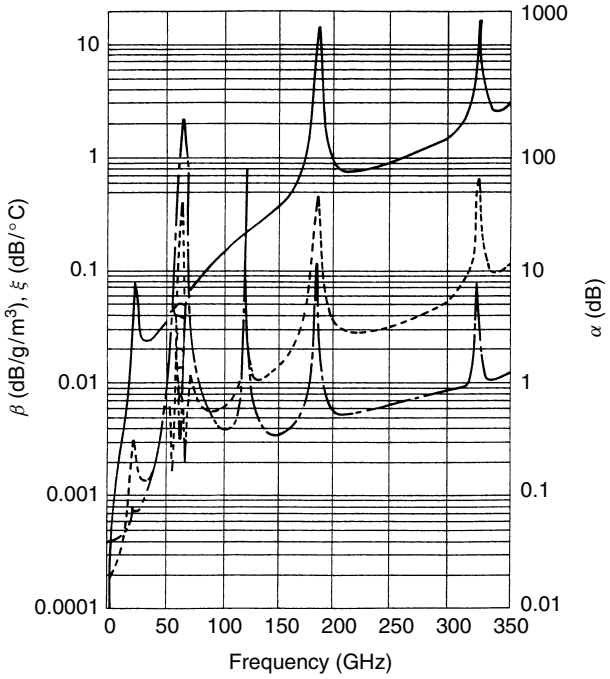


Figure 8. Coefficients for computing zenith attenuation: —·—·, α coefficient; ———, β coefficient; - - - , γ coefficient.

The coefficients α, β, and ξ are plotted in Fig. 8. The attenuation at elevation angles above ~6° can be calculated by multiplying the zenith attenuation by the cosecant of the elevation angle. For angles below 6° the attenuation is assumed to be proportional to the pathlength through the attenuating medium. This distance is given by Altshuler et al. [31] as

$$d(\theta) = [(a_e + h)^2 - a_e^2 \cos^2\theta]^{1/2} - a_e \sin\theta \qquad (27)$$

where θ is the elevation angle, a_e is $\frac{4}{3}$ the earth's radius ($a_e = 8500$ km), and h is the scale height of combined oxygen and water vapor gases (~3.2 km). Therefore

$$A(\theta) = \frac{A(90^\circ)d(\theta)}{h} \qquad (28)$$

The zenith attenuation is plotted as a function of frequency in Fig. 9 for a completely dry atmosphere and more typical atmospheres having surface absolute humidities of 3 and 10 g/m^3 [32]. For a dry atmosphere the total zenith attenuation in the window regions is only a fraction of a decibel. However, this attenuation increases very sharply as the atmosphere becomes moist, particularly below a wavelength of a few millimeters, at which losses on the order of tens of decibels are possible. A plot of the apparent sky temperature (emission) is shown in Fig. 10 as a function of frequency for a set of elevation angles from the horizon to zenith and for an atmosphere having a water vapor density of 7.5 g/m^3 [32]. The sky temperature is relatively low for higher-elevation angles and longer wavelengths and gradually increases for lower-elevation angles and

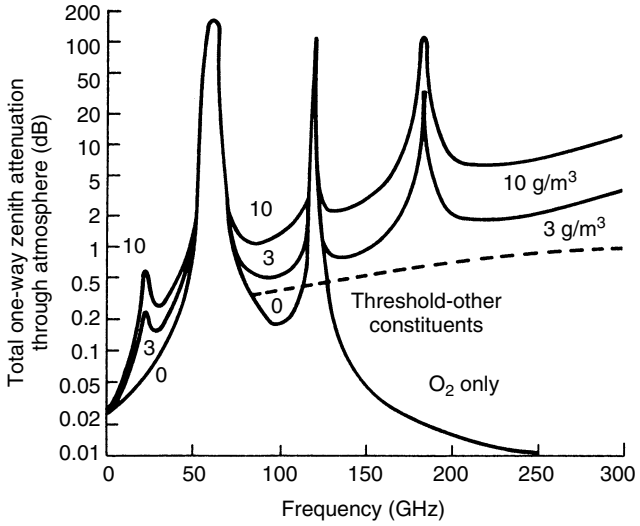


Figure 9. Total zenith attenuation through atmosphere as a function of frequency.

shorter wavelengths, approaching a terrain temperature of approximately 290 K.

For conditions of fog or cloud, the modeling of the atmosphere becomes increasingly difficult, particularly for slant paths close to the horizon. Fog and cloud models have been developed and the attenuation can be estimated using the information provided in Fig. 6. It must be emphasized that these attenuations are only approximate, because neither the true liquid water density of the cloud nor the extent of the cloud is accurately known. Because the cloud particulates are in the Rayleigh region and scattering losses are negligible, the corresponding brightness temperature (emission) can be calculated from Eq. (6). Several investigators have measured cloud attenuations at millimeter wavelengths. Altshuler et al. [31] have presented cloud attenuation statistics at frequencies of 15 and 35 GHz based on 440 sets of measured data. They characterized sky conditions as clear, mixed clouds, or heavy clouds. Average attenuations as a function of elevation angle are shown in Fig. 11. They also demonstrated a reasonable correlation between the slant path attenuation and surface absolute humidity. Typical slant path attenuations extrapolated to zenith were 0.1 and 0.36 dB at 15 and 35 GHz, respectively. Lo et al. [33] measured attenuations at wavelengths of 8.6 and 3.2 mm over a 6-month period, and obtained typical attenuations of 0.42 and 2.13 dB, respectively. Slobin [34] has estimated average-year statistics of cloud attenuation and emission down to a wavelength of 6 mm for various climatically distinct regions throughout the United States.

As was the case for terrestrial paths, rain, sleet, and wet snow present the most serious propagation limitations on earth–space millimeter wave systems. A number of investigators have derived models for predicting rain attenuation, and those results have been summarized by Ippolito [35] and Crane et al. [36–37]. All the techniques assume that the slant path attenuation can be estimated by modifying the attenuation for a terrestrial path by an

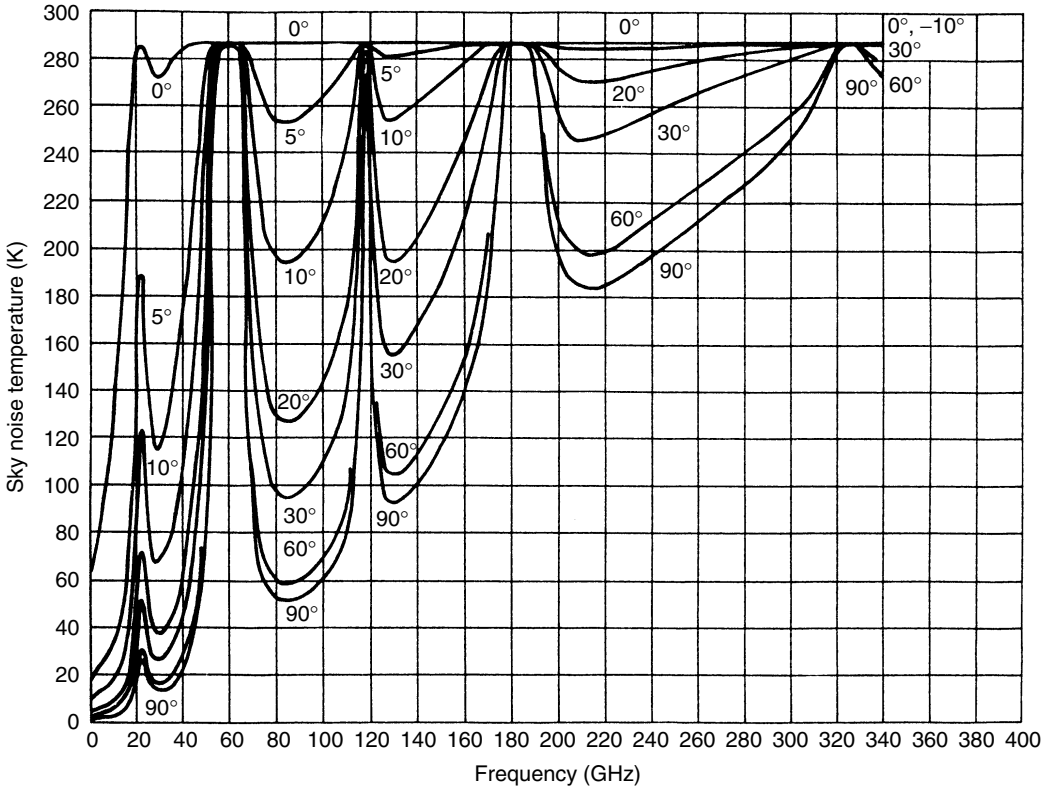


Figure 10. Sky noise temperature as a function of frequency for a water vapor density of 7.5 g/m^3.

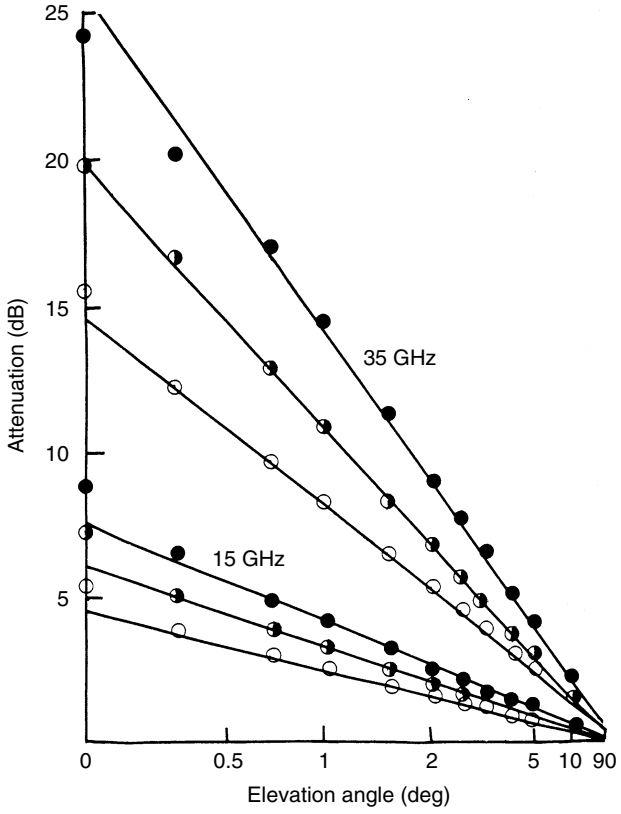


Figure 11. Typical cloud attenuations as a function of elevation angle: ☺ clear; ☻ cloudy; ◑ mixed.

effective pathlength parameter that is usually a function of the elevation angle and the type of rain. For example, the Crane model [24] for estimating rain attenuation for terrestrial paths, presented in Section 3.1.1, can be modified for slant path attenuations. It is assumed that the rain rate has a constant value between the station height h_0 and the height h of the 0 °C isotherm. Precipitation above h consists of ice particles, so the attenuation is considered negligible. Although the height of the 0 °C isotherm is variable, radar measurements have shown it to have a strong dependence on site latitude and rain rate, which in turn are linearly related to the logarithm of the probability of occurrence P. With h and h_0 known, the effective pathlength D through the rain can be calculated:

$$D = \frac{h - h_0}{\tan\theta} \qquad \theta \geq 10^\circ \tag{29}$$

$$= a_e \psi \qquad \theta < 10^\circ \tag{30}$$

where

$$\psi = \sin^{-1} \frac{\cos\theta}{h + a_e} \left\{ \begin{array}{l} [(h_0 + a_e)^2 \sin^2\theta + 2a_e(h - h_0) \\ +h^2 - h_0^2]^{1/2} - (h_0 + a_e)\sin\theta \end{array} \right\} \tag{31}$$

where a_e is the effective earth radius (8500 km) and θ is the elevation angle.

The surface-projected attenuation $A(R_p, D)$ is calculated from Eq. (23) or (24). Then, the value for the slant path A_s is estimated assuming a constant attenuation

below h by

$$A_s = \frac{LA(D)}{D} \tag{32}$$

$$L = \frac{D}{\cos\theta} \quad \theta \geq 10^\circ \tag{33}$$

$$= [(a_e + h_0)^2 + (a_e + h)^2 - 2(a_e + h_0)(a_e + h)\cos\psi]^{1/2}, \quad \theta < 10^\circ \tag{34}$$

Several methods can be used to measure slant path attenuations. The most straightforward, but also the most costly, is to place a millimeter wave beacon in space [38]. Measurements using satellite beacons have been made at wavelengths from approximately 30 to 10 mm and have been summarized by Ippolito [35] and Bauer [38]. A number of investigators have made attenuation measurements using the sun as a source [39–41]. When a radiometer is pointed at the sun, the noise power received consists of radiation from the sun and the atmosphere. The antenna temperature can be expressed as

$$T_a = T_a' e^{-\tau} + \int_0^\infty T(s)\gamma(s)\exp\left(-\int_0^\infty \gamma(s')\,ds'\right) ds \tag{35}$$

where T_a' is the effective antenna temperature of the sun with no intervening atmosphere (in degrees Kelvin), $T(s)$ is the atmospheric temperature, τ is the total attenuation (in nepers), $\gamma(s)$ is the absorption coefficient, and s is the distance from the antenna (ray path). In simpler terms

$$T_a = T_a' e^{-\gamma} + (1 - e^{-\gamma})T_m \tag{36}$$

where T_m is the atmospheric mean absorption temperature within the antenna beam. The attenuation γ appears in both terms on the right-hand side of Eq. (36). Because the second term is the emission, it can easily be canceled out by pointing the antenna beam toward and away from the sun. With the second term balanced off, Eq. (36) can be solved for γ, converted from nepers to decibels, and expressed as

$$A = 10\log\left(\frac{T_a'}{T_a}\right) \tag{37}$$

T_a' is determined from a set of antenna temperature measurements made under clear-sky conditions as a function of elevation angle; for these conditions the antenna temperature is proportional to the cosecant of the elevation angle, and it can be shown that T_a' is equal to the slope of the line $\log T_a$ versus $\csc\theta$ [41]. Because the attenuation also appears in the emission term in Eq. (36), it is possible to determine the attenuation from an emission measurement. In this method the antenna must be pointed away from the sun or the moon (millimeter wave radiation from all other natural sources is negligible), so the first term on the right-hand side of Eq. (36) can be considered zero and Eq. (36) reduces to Eq. (6). As before, the equation is then solved for the attenuation γ and expressed in decibels as

$$A = \frac{10\log T_m}{T_m - T_a} \tag{38}$$

Another technique for estimating rain attenuation at millimeter wavelengths is from a measurement of the reflectivity factor of the rain. Attenuation is derived from established relationships between these parameters [42,43]. McCormick [44] has compared attenuations derived from radar reflectivity with those obtained directly utilizing a beacon placed on an aircraft. Strickland [45] has measured slant path attenuations using radar, radiometers, and a satellite beacon simultaneously. When a single-wavelength radar is used, calibration errors and the uncertainty in the reflectivity–attenuation relationship, particularly for mixed-phase precipitation, limit the accuracy of this technique. Uncertainties in the reflectivity–attenuation relationship can, however, be reduced by using a dual-wavelength radar and measuring the differential attenuation.

In summation, there are four methods for measuring slant path attenuations. The most direct method is to place a source in space. This allows measurements to be made over a very wide dynamic range. An additional advantage is that the polarization and bandwidth limitations imposed by the atmosphere can also be measured. One disadvantage is cost; and, depending on the satellite orbit, measurements may be possible at only one elevation angle, which may or may not be a drawback. Total attenuation can be measured very accurately and economically using the sun as a source over dynamic ranges approaching 25–30 dB. A disadvantage is that measurements can be made only in the direction of the sun and during the day. Attenuation can easily be determined from an emission measurement on a continual basis and at any elevation angle. It must be emphasized that there are two major problems that limit the accuracy of this technique. The true value of T_m is not always known; if $T_m - T_a$ is large, this uncertainty is not serious, but if $T_m - T_a$ is small, a large error may arise. Also, the emission is related only to the absorption, whereas the attenuation includes losses due to scattering in addition to those due to absorption. Therefore, in cases for which the Rayleigh approximation is not valid and scattering losses are appreciable, errors will arise. Techniques for correcting for the additional losses due to scattering have been investigated by Zavody [46] and Ishimaru and Cheung [47]. For these reasons this method is not generally recommended for attenuations much above 10 dB. Attenuations determined from radar reflectivity measurements have the limitations in accuracy discussed previously, and in general this technique is not suitable for losses arising from very small particulates such as fog, cloud, or drizzle. This method does, however, have the advantage of measuring attenuation as a function of distance from the transmitter.

4. CONCLUSION

The interaction of millimeter waves with atmospheric gases and particulates has been examined. Precipitation in general and rain in particular limit the performance of longer millimeter wave systems. Systems operating at short millimeter wavelengths are significantly affected by

high water vapor absorption in addition to precipitation, so applications in this region of the spectrum will of necessity be limited to very short paths.

The use of millimeter waves has probably not progressed as rapidly as had been originally anticipated. For many years, the more optimistically inclined envisioned millimeter waves revolutionizing traditionally longer wavelength communications. When the discovery of the laser created a temporary lull in millimeter wave research, the more pessimistically inclined feared that millimeter waves had passed from infancy to obsolescence without having experienced a period of fruitfulness. Finally, there were realists who recognized that cost is a major consideration and that millimeter wave systems would reach the marketplace only when they could be shown to be competitive with systems that operate at longer or shorter wavelengths or to have unique properties such that needed applications could be realized only with millimeter waves. So far, history seems to be supporting the realists.

BIOGRAPHY

Edward E. Altshuler received a B.S. degree in physics from Northeastern University, Boston, Massachusetts, in 1953, an M.S. degree in physics from Tufts University, Medford, Massachusetts, in 1954, and the PhD. degree in applied physics from Harvard University, Cambridge, Massachusetts, in 1960. He joined Air Force Cambridge Research Labs (AFCRL), Hanscom Air Force Base (AFB), Massachusetts in 1960, but left in 1961 to become director of engineering at Gabriel Electronics, Millis, Massachusetts; he later returned to AFCRL in 1963 as chief of the propagation branch from 1963 to 1982. He was a lecturer in the Northeastern University Graduate School of Engineering from 1964 to 1991. He has served on the Air Force Scientific Advisory Board and was chairman of the NATO Research Study Group on millimeter wave propagation from 1974 to 1993. He was President of the Hanscom Chapter of Sigma Xi during 1989 through 1990. He received the IEEE Harry Diamond Memorial Award in 1997 and was awarded an IEEE Millennium Medal in 2000. He is a fellow of both the IEEE and AFRL. Dr. Altshuler has over 120 scientific publications, conference papers, and patents. He is currently conducting antenna research for the Air Force Research Laboratory at Hanscom AFB.

BIBLIOGRAPHY

1. E. E. Altshuler, New applications at millimeter wavelengths, *Microwave J.* **11**: 38–42 (1968).
2. E. K. Smith and S. Weintraub, The constants in the equation for atmospheric refractive index at radio frequencies, *Proc. IRE* **41**: 1035–1037 (1953).
3. B. R. Bean and E. J. Dutton, *Radio Meteorology*, Dover, New York, 1968.
4. E. E. Altshuler, Tropospheric range-error corrections for the global positioning system, *IEEE Trans. Antennas Propag.* **46**: 643–649 (1998).
5. G. K. Elgered, Tropospheric wet-path delay measurements, *IEEE Trans. Antennas Propag.* **30**: 502–505 (1982).
6. M. A. Gallop, Jr. and L. E. Telford, Use of atmospheric emission to estimate refractive errors in a non-horizontally stratified troposphere, *Radio Sci.* **11**: 935–945 (1975).
7. L. W. Schaper, Jr., D. H. Staelin, and J. W. Waters, The estimation of tropospheric electrical path length by microwave radiometry, *Proc. IEEE* **58**: 272–273 (1970).
8. S. C. Wu, Optimum frequencies of a passive microwave radiometer for tropospheric path-length correction, *IEEE Trans. Antennas Propag.* **27**: 233–239 (1979).
9. J. Goldhirsh, B. H. Musiani, and W J. Vogel, Cumulative fade distributions and frequency scaling techniques at 20 GHz from the advanced communications technology satellite and at 12 GHz from the digital satellite system, *Proc. IEEE* **85**: 910–916 (1997).
10. C. E. Mayer, B. E. Jaeger, R. K. Crane, and X. Wang, Ka-band scintillations: measurements and model predictions, *Proc. IEEE* **85**: 936–945 (1997).
11. F. S Marzano and C. Riva, Evidence of long-term correlation between clear-air attenuation and scintillation in microwave and millimeter-wave satellite links, *IEEE Trans. Antennas Propag.* **47**: 1749–1757 (1979).
12. J. W. Waters, *Methods of Experimental Physics*, 12B, Academic Press, New York, 1976, Chap. 23.
13. R. L. Olsen, Cross polarization during clear-air conditions on terrestrial links-a review, *Radio Sci.* **16**: 631–647 (1981).
14. H. C. Van de Hulst, *Light Scattering by Small Particles*, Wiley, New York, 1957.
15. H. R. Pruppacher and R. L. Pitter, A semi-empirical determination of the shape of cloud and rain drops, *J. Atmos. Sci.* **28**: 86–94 (1971).
16. D. Atlas, M. Kerker, and W. Hitschfeld, Scattering and attenuation by nonspherical atmospheric particles, *J. Atmos. Terr. Phys.* **3**: 108–119 (1953).
17. R. L. Olsen, D. V. Rogers, and D. E. Hodge, The aRb relation in the calculation of rain attenuation, *IEEE Trans. Antennas Propag.* **26**: 318–329 (1978).
18. R. G. Medhurst, Rainfall attenuation of centimeter waves: Comparison of theory and measurement, *IEEE Trans. Antennas Propag.* **13**: 550–563 (1965).
19. G. Brussaard, A meteorological model for rain-induced cross polarization, *IEEE Trans. Antennas Propag.* **24**: 5–11 (1976).
20. L. J. Battan, *Radar Observations of the Atmosphere*, Univ. Chicago Press, 1973.
21. R. G. Eldridge, Haze and fog aerosol distributions, *J. Atmos. Sci.* **23**: 605–613 (1966).
22. C. Platt, Transmission of submillimeter waves through water clouds and fogs, *J. Atmos. Sci.* **27**: 421–425 (1970).
23. T. S. Chu, Effects of sandstorms on microwave propagation, *Bell Syst. Tech. J.* **58**: 549–555 (1979).
24. R. K Crane, Prediction of attenuation by rain, *IEEE Trans. Commun.* **28**: 1717–1733 (1980).
25. A. Blomquist and J. R. Norbury, Attenuation due to rain or series, parallel and convergent terrestrial paths, *Alta Freq.* **66**: 185–190 (1979).
26. A. VanderVorst, Cross polarization on a terrestrial path, *Alta Freq.* **48**: 201–209 (1979).

27. D. V Rogers, L. J. Ippolito, Jr., and F. Davarian, System requirements for Ka-band earth-satellite propagation data, *Proc. IEEE* **85**: 810–820 (1997).

28. Y. Karasawa and Y. Maekawa, Ka-band earth-space propagation research in Japan, *Proc. IEEE* **85**: 821–842 (1997).

29. R. Arbesser-Rastburg and A. Paraboni, European research on Ka-band slant path propagation, *Proc. IEEE* **85**: 843–852 (1997).

30. H. Helmken et al., A three-site comparison of fade-duration measurements, *Proc. IEEE* **85**: 917–925 (1997).

31. E. E. Altshuler, M. A. Gallop, and L. E. Telford, Atmospheric attenuation statistics at 15 and 35 GHz for very low elevation angles, *Radio Sci.* **13**: 839–852 (1978).

32. E. K. Smith, Centimeter and millimeter wave attenuation and brightness temperature due to atmospheric oxygen and water vapor, *Radio Sci.* **17**: 1455–1464 (1982).

33. L. Lo, B. M. Fanning, and A. W. Straiton, Attenuation of 8.6 and 3.2 mm radio waves by clouds, *IEEE Trans. Antennas Propag.* **23**: 782–786 (1975).

34. S. D. Slobin, Microwave noise temperature and attenuation of clouds: Statistics of these effects at various sites in the United States, Alaska and Hawaii, *Radio Sci.* **17**: 1443–1454 (1982).

35. L. J. Ippolito, Radio propagation for space communications systems, *Proc. IEEE* **69**: 697–727 (1981).

36. R. K. Crane and A. W Dissanayake, ACTS propagation experiment: Attenuation distribution observations and prediction model comparisons, *Proc. IEEE* **85**: 879–892 (1997).

37. R. K. Crane and P. C. Robinson, ACTS propagation experiment: Rain-rate distribution observations and prediction model comparisons, *Proc. IEEE* **85**: 946–958 (1997).

38. R. Bauer, Ka-band propagation measurements: An opportunity with the advanced communications technology satellite (ACTS), *Proc. IEEE* **85**: 853–862 (1997).

39. E. E. Altshuler and L. E. Telford, Frequency dependence of slant path rain attenuations at 15 and 35 GHz, *Radio Sci.* **15**: 781–796 (1980).

40. R. W. Wilson, Suntracker measurements of attenuation by rain at 16 and 30 GHz, *Bell Syst. Tech. J.* **48**: 1383–1404 (1969).

41. K. N. Wulfsberg, Atmospheric attenuation at millimeter wavelengths, *Radio Sci.* **2**: 319–324 (1967).

42. J. Goldhirsh, A review on the application of non-attenuating frequency radars for estimating rain attenuation and space-diversity performance, *IEEE Trans. Geosci. Electron.* **17**: 218–239 (1979).

43. J. D. Beaver and V. N. Bringi, The application of S-band polarimetric radar measurements to Ka-band attenuation prediction, *Proc. IEEE* **85**: 893–909 (1997).

44. K. S. McCormick, A comparison of precipitation attenuation and radar backscatter along earth-space paths, *IEEE Trans. Antennas Propag.* **20**: 747–755 (1972).

45. J. I. Strickland, The measurement of slant path attenuation using radar, radiometers and a satellite beacon, *J. Rech. Atmos.* **VIII**: 347–358 (1974).

46. A. M. Zavody, Effect of scattering by rain or radiometer measurements at millimeter wavelengths, *Proc. IEE* **121**: 257–263 (1974).

47. A. Ishimaru and R. L. T. Cheung, Multiple-scattering effect on radiometric determination of rain attenuation at millimeter wavelengths, *Radio Sci.* **15**: 507–516 (1980).

MIMO COMMUNICATION SYSTEMS

ROHIT U. NABAR
AROGYASWAMI J. PAULRAJ
Stanford University
Stanford, California

1. INTRODUCTION

Successful deployment of wireless networks presents a number of challenges. These include limited availability of the radiofrequency spectrum and a complex time-varying wireless environment (fading and multipath). Meeting the increasing demand for higher data rates, better quality of service (QoS), fewer dropped calls, higher network capacity, and user coverage calls for innovative techniques that improve spectral efficiency and link reliability. The use of multiple antennas at both receiver and transmitter in a wireless system is an emerging technique that promises significant improvements in these measures. This technology, popularly known as multiple-input/multiple-output (MIMO) wireless technology, offers a variety of (often competing) leverages that, if exploited correctly, can significantly improve network performance. These leverages include *array gain, diversity gain, multiplexing gain*, and *interference reduction*.

Figure 1 shows a typical MIMO system with M_T transmit antennas and M_R receive antennas. The space–time modem at the transmitter (Tx) encodes and modulates the information bits to be conveyed to the receiver. Additionally, it maps the signals to be transmitted across space (M_T transmit antennas) and time. The space–time modem at the receiver (Rx) processes the signals received on each of the M_R receive antennas, and in accordance with the transmitter's signaling strategy, demodulates and decodes the received signal. Signaling strategies are designed on the basis of the transmitter's knowledge of the wireless channel (we assume channel knowledge at the receiver) and link requirements (data rate, error rate, etc.) and exploit one

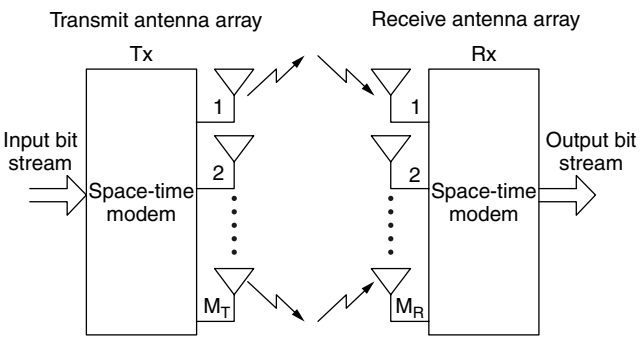


Figure 1. Schematic of a MIMO communication system.

or more of the four leverages of MIMO systems. In the following section, we describe these key leverages.

2. LEVERAGES OF MIMO TECHNOLOGY

To explore the leverages of MIMO technology we focus on various subsets of a MIMO link including MISO (multiple input/single output), SIMO (single input/multiple output) and SISO (single input/single output). Leverages such as array gain, diversity gain, and interference reduction that are available in MIMO systems are also offered by SIMO and MISO systems. Multiplexing gain, however, can only be exploited in MIMO systems.

2.1. Array Gain

Consider a SIMO system with one transmit antenna and two receive antennas as shown in Fig. 2. The two receive antennas see different versions, s_1 and s_2, of the same transmitted signal, s. The signals s_1 and s_2 have different amplitudes and phases as determined by the propagation conditions. If the channel is known to the receiver, appropriate signal processing techniques can be applied to combine the signals s_1 and s_2 coherently so that the resultant power of the signal at the receiver is enhanced, leading to an improvement in signal quality. More specifically, the signal-to-noise ratio [ratio of signal power to noise power (SNR)] at the output is equal to the sum of the SNR on the individual links. This result can be extended to systems with one transmit antenna and more than two receive antennas. The average increase in signal power at the receiver in such systems is defined as array gain and is proportional to the number of receive antennas. Array gain can also be exploited in systems with multiple antennas at the transmitter (MISO or MIMO systems). Extracting the maximum possible array gain in such systems requires channel knowledge at the transmitter, so that the signals may be optimally processed before transmission. Analogous to the SIMO case, the array gain in MISO systems is proportional to the number of transmit antennas. The array gain in MIMO systems depends on the number of transmit and receive antennas and is a function of the dominant singular value of the channel.

2.2. Diversity Gain

Signal power in a wireless channel fluctuates (or fades) with time–frequency–space. When the signal power drops dramatically, the channel is said to be in a fade. Diversity is used in wireless systems to combat fading. The basic principle behind diversity is to provide the receiver with several looks at the transmitted signal over independently fading links (or diversity branches). As the number of diversity branches increases, the probability that at any instant of time one or more branch is not in a fade increases. Thus diversity helps stabilize a wireless link.

Diversity is available in SISO links in the form of time or frequency diversity. The use of time or frequency diversity in SISO systems often incurs a penalty in data rate due to the utilization of time or bandwidth to introduce redundancy. The introduction of multiple antennas at the transmitter and/or receiver provides spatial diversity, the use of which does not incur a penalty in data rate while adding the array gain advantage discussed earlier. In this article we are concerned with this form of diversity. To utilize spatial diversity we must transmit and receive from antennas that are spaced by more than the coherence distance, which is the minimum spatial separation between antennas at the receiver (and/or transmitter) that ensures that the received signals (or their components) experience independent fading. In a rich scattering environment the coherence distance is approximately equal to half the wavelength ($\lambda/2$) [1] of the transmitted signal. There are two forms of spatial diversity: receive and transmit diversity.

Receive diversity applies to systems with multiple antennas only at the receiver (SIMO systems) [2]. Figure 3 illustrates a system with receive diversity. Signal s is transmitted from a single antenna at the transmitter. The two receive antennas see independently faded versions, s_1 and s_2, of the transmitted signal, s. The receiver combines these signals using appropriate signal processing techniques so that the resultant signal exhibits greatly reduced amplitude variability (fading) as compared to either s_1 or s_2. The amplitude variability can be further reduced by adding more antennas to the receiver. The diversity in a system is characterized by the number of independently fading diversity branches, also known as the diversity order. The diversity order of the system in Fig. 3 is two and in general is equal to the number of receive antennas, M_R, in a SIMO system.

Transmit diversity is applicable when multiple antennas are used at the transmitter and has become an active

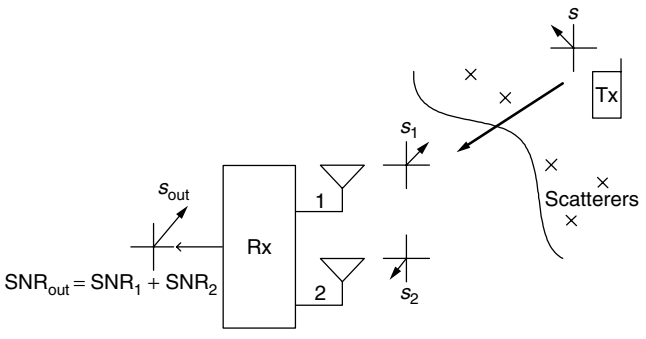


Figure 2. Array gain.

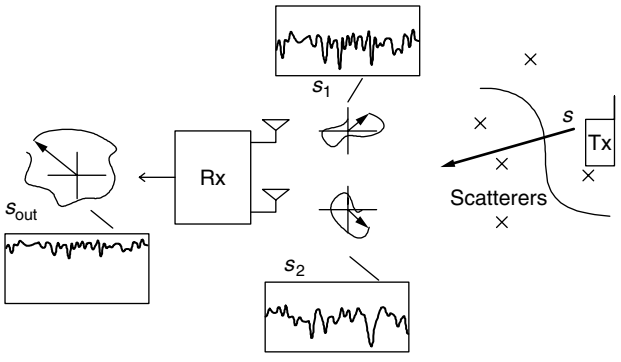


Figure 3. Receive diversity.

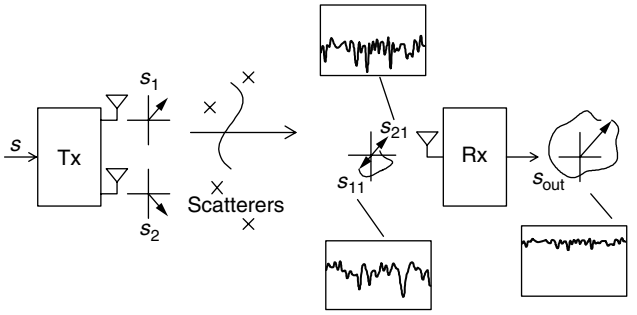


Figure 4. Transmit diversity.

area for research [3–5]. Extracting diversity in such systems does not necessarily require channel knowledge at the transmitter. However, suitable design of the transmitted signal is required to extract diversity. Space–time coding [6,7] is a powerful transmit diversity technique that relies on coding across space (transmit antennas) and time to extract diversity. Figure 4 shows a generic transmit diversity scheme for a system with two transmit antennas and one receive antenna. At the transmitter, signals s_1 and s_2 are derived from the original signal to be transmitted, s, such that the signal s can be recovered from either of the received signals s_{11} or s_{21}. The receiver combines the received signals in such a manner that the resultant output exhibits reduced fading when compared to s_{11} or s_{21}. The diversity order of this system is two and in general is equal to the number of transmit antennas, M_T, in a MISO system.

Utilization of diversity in MIMO systems requires a combination of receive and transmit diversity described above. A MIMO system can be decomposed into $M_T \times M_R$ SISO links. If the signals transmitted over each of these links experience independent fading, then the diversity order of the system is given by $M_T \times M_R$. Thus the diversity order in a MIMO system scales linearly with the product of the number of receive and transmit antennas.

2.3. Multiplexing Gain

MIMO systems offer a capacity (data rate) enhancing leverage not available in SIMO or MISO systems. We refer to this leverage as multiplexing gain which can be realized through a technique known as *spatial multiplexing* [8,9]. Figure 5 shows the basic principle of spatial multiplexing for a system with two transmit and two receive antennas. The symbol stream to be transmitted is split into two half-rate substreams and modulated to form the signals s_1 and s_2 that are transmitted simultaneously from separate antennas. Under favorable channel conditions, the spatial signatures of these signals [denoted by $[s_{11} \quad s_{12}]^T$ (the superscript ^{T} represents matrix transpose) and $[s_{21} \quad s_{22}]^T$] induced at the receive antennas are well separated (ideally orthogonal). The receiver can then extract the two substreams, s_1 and s_2, which it combines to give the original symbol stream, s.

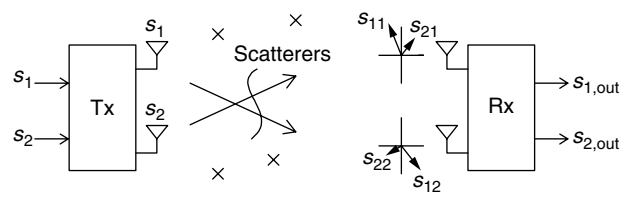


Figure 5. Spatial multiplexing.

2.4. Interference Reduction

Cochannel interference arises due to the reuse of frequency spectrum in wireless networks and adds to the overall noise in the system and deteriorates performance. Figure 6 illustrates the general principle of interference reduction for a receiver with two antennas. Typically, the desired signal (s) and the interference (i) arrive at the receiver with well separated spatial signatures — $[s_1 \quad s_2]^T$ and $[i_1 \quad i_2]^T$, respectively. The receiver can exploit the difference in signatures to reduce the interference, thereby enhancing the signal to interference ratio [ratio of signal power to interference power (SIR)]. Interference reduction requires knowledge of the desired signal's channel. Complete knowledge of the interfering signal's channel is not necessary. Interference reduction can also be implemented at the transmitter, where the goal is to enhance the signal power at the intended receiver and minimize the interference energy sent toward the cochannel users. Interference reduction allows the use of aggressive reuse factors and improves network capacity.

Having discussed the key advantages of MIMO technology we note that it may not be possible to exploit all the leverages simultaneously in a MIMO system. This is because some of the leverages and their methods of realization may be mutually conflicting. The optimal MIMO signaling strategy is a function of the wireless channel and network requirements. Exploiting the benefits of MIMO technology requires a good understanding of the MIMO channel. In the following section we introduce a simple MIMO channel model for an interference free environment.

3. MIMO CHANNEL MODEL

Consider a MIMO system with M_T transmit antennas and M_R receive antennas as shown in Fig. 7. For simplicity

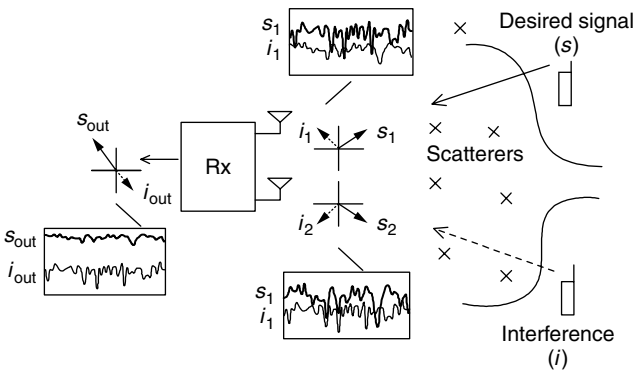


Figure 6. Interference reduction.

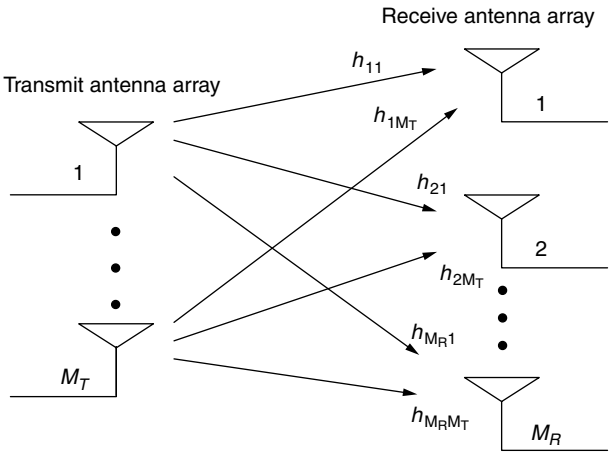


Figure 7. Flat fading MIMO channel model.

we consider only frequency flat fading; thus, the fading is not frequency-selective. When a continuous-wave (CW) probing signal, s, is launched from the jth transmit antenna, each of the M_R receive antennas see a complex weighted version of the transmitted signal. We denote the signal received at the ith receive antenna by $h_{ij}s$, where h_{ij} is the channel response between the jth transmit antenna and the ith receive antenna. The vector $[h_{1j}\, h_{2j} \cdots h_{M_Rj}]^T$ is the signature induced by the jth transmit antenna across the receive antenna array. It is convenient to denote the MIMO channel ($\mathbf{H}$) in matrix notation as shown below:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1M_T} \\ h_{21} & h_{22} & \cdots & h_{2M_T} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M_R1} & h_{M_R2} & \cdots & h_{M_RM_T} \end{bmatrix} \tag{1}$$

The channel matrix $\mathbf{H}$ defines the input–output relation of the MIMO system and is also known as the channel transfer function. If a signal vector $\mathbf{x} = [x_1\, x_2 \cdots x_{M_T}]^T$ is launched from the transmit antenna array (x_j is launched from the jth transmit antenna) then the signal received at the receive antenna array, $\mathbf{y} = [y_1 y_2 \cdots y_{M_R}]^T$ is given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{2}$$

where $\mathbf{n}$ is the $M_R \times 1$ noise vector consisting of independent complex Gaussian distributed elements with zero mean and variance σ_n^2 (white noise). Note that the discussion above pertains to a snapshot of the channel at a particular frequency and a specific instant of time. Channels with large delay spreads show greater variability of $\mathbf{H}$ with frequency. Likewise, channels with large Doppler spreads show greater variability of $\mathbf{H}$ with time. In a scattering environment with sufficient antenna separation at the transmitter and receiver, the elements of the channel matrix $\mathbf{H}$ can be assumed to be independent, zero-mean, complex Gaussian random variables (Rayleigh fading) with equal variances. This model is popularly referred to as the i.i.d. fading MIMO channel model.

The choice of MIMO signaling strategies that optimize performance depends on knowledge of the channel at the receiver and/or transmitter. Channel knowledge at the receiver is a common assumption. Knowledge of the channel can be maintained at the receiver via training and tracking. Maintaining channel knowledge at the transmitter requires the use of feedback from the receiver or through the reciprocity principle in a duplex system. If feedback is employed, the receiver must estimate the channel (using training symbols/tones in the transmit signal) and convey the channel state information to the transmitter via a return channel. Alternatively, in a full (frequency or time)-duplex system, the transmitter first learns the "return channel" (i.e., the reverse link) and estimates the channel for the forward link by invoking the reciprocity principle, which guarantees that the transmit and receive channels are identical if the frequency and time (and antennas) of operation on both links are identical, given, of course, that in a duplex system the frequency and time of operation on both links are not identical but close. Maintaining channel knowledge at the transmitter is difficult to implement in practice. For the remainder of this article we focus on the case when the channel is known perfectly to the receiver and is unknown to the transmitter.

4. MIMO CHANNEL CAPACITY

The spectral efficiency of a wireless link is defined as the data rate transmitted per unit bandwidth [bits per second per hertz (bps/Hz)]. The maximum error-free spectral efficiency that can be achieved over a communication link is upper-bounded by the Shannon capacity. In this section we briefly review the Shannon capacity of a MIMO channel for flat fading conditions and then extend the results to frequency selective fading.

The Shannon capacity of a SISO (scalar) channel[1] is given by

$$C = \log_2(1 + \rho\|\mathrm{H}\|^2) \quad \text{bps/Hz} \tag{3}$$

where ρ is the SNR and H is the scalar transfer function. We assume $\mathcal{E}[\mathrm{H}] = 0$ ($\mathcal{E}$ stands for the expectation operator) and $\mathcal{E}[\|\mathrm{H}\|^2] = 1$. As is well known, at high SNR an increase in capacity of 1 bps/Hz is achieved for every 3-dB increase in SNR.

The Shannon capacity of MIMO channels has been derived in [10,11] and is given by[2]

$$C = \log_2\left[\det\left(\mathbf{I}_{M_R} + \frac{\rho}{M_T}\mathbf{H}\mathbf{H}^{\dagger}\right)\right] \quad \text{bps/Hz} \tag{4}$$

where ρ is the SNR defined above for the SISO link and $\mathbf{H}$ is the $M_R \times M_T$ matrix transfer function in (1). We assume $\mathcal{E}[h_{ij}] = 0$ and $\mathcal{E}[\|h_{ij}\|^2] = 1$ (i.i.d. fading model).

Consider a system with an equal number of transmit and receive antennas, $M_T = M_R$. If the channel signatures

[1] A SISO channel can me modeled as a MIMO channel with a 1×1 transfer function.

[2] Here, det ($\mathbf{X}$) stands for the determinant of matrix $\mathbf{X}$. $\mathbf{I}_m$ is the $m \times m$ identity matrix. The superscript † stands for conjugate transpose.

are orthogonal such that $\mathbf{H}\mathbf{H}^{\dagger} = M_T \mathbf{I}_{M_R}$, then the capacity expression in Eq. (4) reduces to

$$C = M_R \log_2(1 + \rho) \text{ bps/Hz} \tag{5}$$

Hence M_R parallel channels are created within the same frequency bandwidth for no additional power expenditure and capacity scales linearly with number of antennas for increasing SNR; that is, the capacity increases by M_R bps/Hz for every 3 dB increase in SNR, leading to a significant capacity advantage. In general, it can be shown that an orthogonal channel of the form described above maximizes the Shannon capacity of a MIMO system. For the i.i.d. fading MIMO channel model discussed in the previous section, the channel realizations become approximately orthogonal when the number of antennas used is very large. When the number of transmit and receive antennas is not equal, $M_T \neq M_R$, the increase in capacity is limited by the minimum of M_T and M_R.

It is important to note that for a time-varying channel, the channel capacity, C, is a random variable whose distribution depends on the channel statistics. To study the capacities of time-varying channels, we can consider the time-averaged channel capacity:

$$\text{Average capacity} = \mathcal{E} \log_2 \left[\det \left(\mathbf{I}_{M_R} + \frac{\rho}{M_T} \mathbf{H}\mathbf{H}^{\dagger} \right) \right] \text{bps/Hz} \tag{6}$$

where $\mathcal{E}$ is the expected value over the distribution of the elements of $\mathbf{H}$. Figure 8 shows the average capacity as a function of the SNR for the i.i.d. fading channel model for different MIMO configurations. It is clear that the average capacity increases with the number of antennas in the system. At very low SNR, the gain in capacity due to multiple antennas is low, but it increases with increasing SNR becoming asymptotically constant.

Since the channel capacity fluctuates with time, it is also useful to define outage capacity for a fading channel.

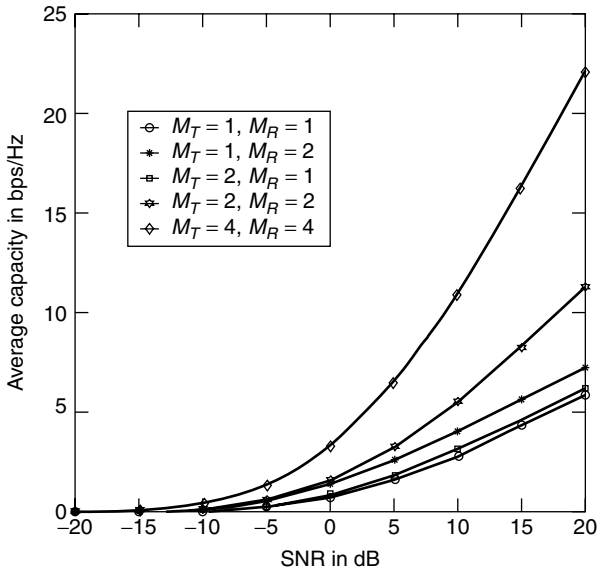


Figure 8. Average capacity for i.i.d. fading MIMO channel model.

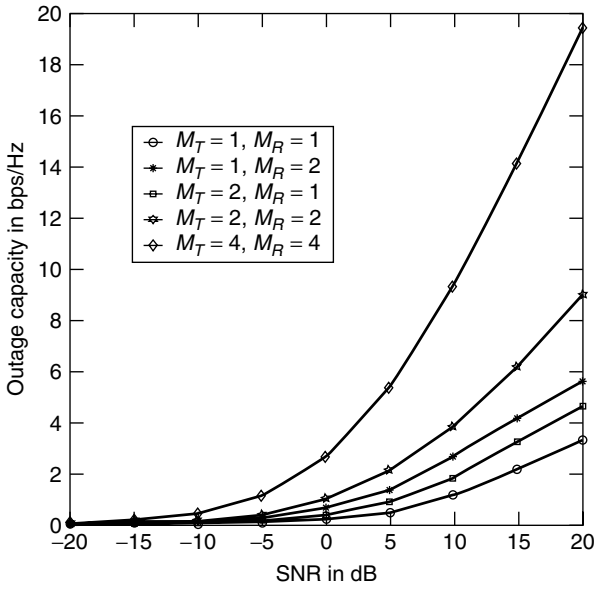


Figure 9. Plot showing 10% outage capacity for i.i.d. fading MIMO channel model.

This is the capacity that is guaranteed at some level of reliability. Thus the 10% outage capacity of a channel is the capacity that is guaranteed 90% of the time. Figure 9 shows the 10% outage capacity as a function of SNR for several MIMO configurations. It is clear that outage capacity also increases with an increasing number of antennas in the system, with better proportionality at larger antenna configurations. This can be attributed to the increased diversity gain at higher values of $M_T \times M_R$.

So far we have restricted our discussion on capacity to a flat fading MIMO channel. The capacity of a frequency-selective fading MIMO channel can be calculated by dividing the frequency band of interest, B, into N narrower flat fading subchannels (Fig. 10). The capacity of the system is then given by the sum of the individual

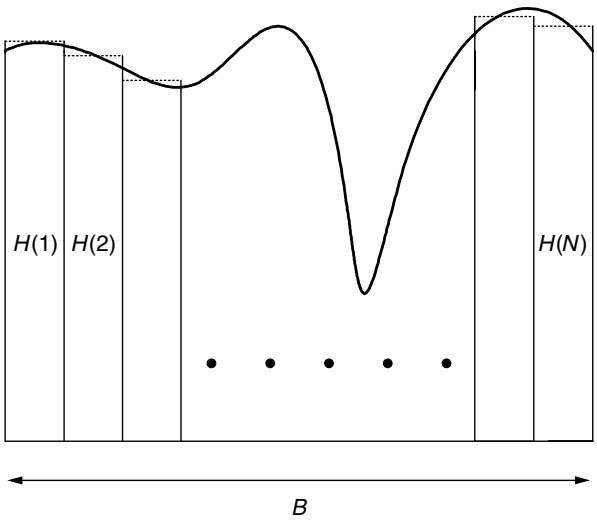


Figure 10. Flat fading approximation of frequency-selective channel.

subchannel capacities

$$C = \frac{1}{N}\sum_{i=1}^{N} \log_2 \det\left(\mathbf{I}_{M_R} + \frac{\rho}{M_T}\mathbf{H}(i)\mathbf{H}(i)^{\dagger}\right) \text{ bps/Hz} \qquad (7)$$

where $\mathbf{H}(i)$ is the channel transfer function corresponding to the ith subchannel. If all subchannels follow the i.i.d. fading MIMO channel model, then averaging over frequency and time reveals that the average capacity of a frequency-selective fading MIMO channel is the same as the average capacity of a frequency-flat fading MIMO channel. However, performance of a frequency-selective fading MIMO channel measured in terms of outage capacity will be better than a frequency-flat fading MIMO channel due to frequency diversity.

The capacity results discussed so far are for the case when the channel is known to the receiver and is unknown to the transmitter. Under this condition, equal power allocation across the transmit antenna array is optimal. If the channel is known to the transmitter, then the optimal transmission strategy from the point of view of maximizing capacity involves allocating possibly unequal amounts of power across the channel modes (corresponding to the singular values of the channel), a technique that is known as *water-filling* [12]. Water-filling may result in an unequal power distribution across the transmit antenna array.

5. MIMO SIGNALING

As described in the previous section, MIMO systems promise much higher spectral efficiency than SISO systems. MIMO systems can also be leveraged to improve the quality of transmission (reduce error rate). Most existing signaling schemes either maximize spectral efficiency (multiplexing mode) or minimize error rate (diversity mode) as will be described next.

5.1. Multiplexing Versus Diversity

We discuss the multiplexing and diversity modes of transmission in the context of a MIMO system with an equal number of transmit and receive antennas. In multiplexing mode, the objective is to maximize the data rate delivered to the receiver. Multiple symbol streams are transmitted to the receiver at the same time as described in Fig. 5. Ideally, the signatures induced at the receive antennas must be orthogonal for spatial multiplexing. The receiver can then perfectly separate the individual symbol streams. On the other hand, if the signatures are not orthogonal, more complex processing at the receiver is required to separate the individual symbol streams. In other words, a low condition number (ideally 1) of $\mathbf{H}$ is preferred for good multiplexing gain. If a channel is not able to support spatial multiplexing because of a high condition number, then data may be delivered to the receiver in diversity mode. Transmit diversity techniques such as space–time coding are employed at the transmitter to extract the spatial diversity in the system. In lieu of the orthogonality requirement of spatial multiplexing, the elements of $\mathbf{H}$ must undergo independent fading for maximum diversity gain. Diversity gain stabilizes the link between transmitter and receiver, improving link reliability.

From the discussion above it is clear that the choice of signaling mode depends on the structure of the MIMO channel. If appropriate channel knowledge is available to the transmitter, it can choose the signaling mode that optimizes performance. This is referred to as *link adaptation*. In general, if the channel is not known to the transmitter, the optimal signaling strategy is a mixture of spatial multiplexing and transmit diversity modes that optimize spectral efficiency and link reliability over the desired range of SNR. Designing efficient, low-complexity receivers for MIMO signaling techniques presents a number of challenges and is a promising area for future research and is described next.

5.2. Receiver Design

In multiplexing mode, each receive antenna observes a superposition of the transmitted signals. The receiver must be able to separate the constituent data streams based on channel knowledge. The separation step determines the computational complexity of the receiver. The problem is similar in nature to the multiuser detection problem in CDMA, and parallels can be drawn between the receiver architectures in these two areas. Maximum-likelihood (ML) detection is optimal but receiver complexity grows exponentially with the number of transmit antennas, making this scheme impractical. Lower-complexity suboptimal receivers include the zero-forcing (ZF) receiver or the minimum mean-square error (MMSE) receiver, the design principles of which are similar to equalization principles for SISO links with intersymbol interference (ISI). An attractive alternative to ZF and MMSE receivers is the V-BLAST algorithm described by Golden et al. [13], which is essentially a successive cancellation technique.

In diversity mode, receiver design is dependent on the diversity signaling technique applied; the most popular is space–time coding. There are two flavors of space–time coding-block codes and trellis codes. Both block codes as well as trellis codes can be designed to extract coding gain and diversity gain. ML receivers and suboptimal receivers similar to those for multiplexing mode have been studied in the context of block codes. The Alamouti scheme [14] is a popular space–time block code for systems with two transmit antennas that uses a simple receiver and extracts maximum diversity gain. Space–time trellis codes are decoded using traditional maximum-likelihood sequence estimation (MLSE) implemented via the Viterbi algorithm. Trellis codes offer better performance than do block codes at the cost of computational complexity. We now briefly describe coding and modulation for MIMO signaling.

5.3. Modulation and Coding for MIMO

MIMO technology is compatible with a wide variety of coding and modulation schemes. In general, the

best performance in achieved by generalizing standard (scalar) modulation and coding techniques to matrix channels. MIMO has been proposed for single-carrier (SC) modulation, direct-sequence code division multiple access (DSCDMA) and orthogonal frequency division multiplexing (OFDM) modulation techniques. MIMO has also been considered in conjunction with single or concatenated coding schemes. Turbo codes and low-density parity codes are currently being studied for MIMO use. As the need for high data rates increases, wireless communication is becoming broadband and there is an increasing trend [15] toward MIMO-OFDM techniques utilizing some version of space–frequency coding with concatenated Reed–Solomon codes.

6. CONCLUDING REMARKS

MIMO wireless communication systems provide significant gains in terms of spectral efficiency and link reliability. These benefits translate to wireless networks in the form of improved coverage and capacity. MIMO communication theory is an emerging area and full of challenging problems. Some promising research areas in the field of MIMO technology include channel estimation, new coding and modulation schemes, low complexity receivers, MIMO channel modeling and network design in the context of MIMO.

Acknowledgments
The authors would like to thank Helmut Bölcskei, Dhananjay Gore, Robert Heath, Sriram Mudulodu, Sumeet Sandhu, and Arak Sutivong for their valuable comments and suggestions. R. Nabar's work was supported by the Dr. T. J. Rodgers Stanford Graduate Fellowship.

BIOGRAPHIES

Arogyaswami J. Paulraj has been a professor at the Department of Electrical Engineering, Stanford University, California, since 1993, where he supervises the Smart Antennas Research Group. This group consists of approximately a dozen researchers working on applications of space-time signal processing for wireless communications networks. His research group has developed many key fundamentals of this new field and has helped shape a worldwide research and development focus on this technology.

Paulraj's research has spanned several disciplines, emphasizing estimation theory, sensor signal processing, parallel computer architectures/algorithms, and space-time wireless communications. His engineering experience includes development of sonar systems, massively parallel computers, and more recently, broadband wireless systems.

He is the author of over 250 research papers and holds 11 patents. Paulraj is a fellow of the Institute of Electrical and Electronics Engineers (IEEE) and a member of the Indian National Academy of Engineering.

Rohit U. Nabar received his B.S. degree in Electrical Engineering in 1998 from Cornell University, Ithaca, New York, and his M.S. degree in electrical engineering in 2000 from Stanford University, California. He is currently a doctoral student in the Smart Antennas Research Group at Stanford University and is the recipient of the Dr. T. J. Rodgers Stanford Graduate Fellowship. His research interests include signal processing and MIMO wireless.

BIBLIOGRAPHY

1. W. C. Y. Lee, *Mobile Communications Engineering*, McGraw-Hill, New York, 1982.
2. W. C. Jakes, *Microwave Mobile Communications*, Wiley, New York, 1974.
3. A. Wittneben, Base station modulation diversity for digital SIMULCAST, *Proc. IEEE VTC*, May 1991, pp. 848–853.
4. N. Seshadri and J. Winters, Two signaling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity, *Int. J. Wireless Inform. Networks* **1**(1): 49–60 (Jan. 1994).
5. J. Guey, M. Fitz, M. Bell, and W. Kuo, Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels, *Proc. IEEE VTC*, 1996, Vol. 1, pp. 136–140.
6. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**(2): 744–765 (March 1998).
7. V. Tarokh, H. Jafarkhani, and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* **45**(5): 1456–1467 (July 1999).
8. U.S. Patent 5,345,599 (1994), A. J. Paulraj and T. Kailath, Increasing capacity in wireless broadcast systems using distributed transmission/directional reception.
9. G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* **1**(2): 41–59 (1996).
10. I. E. Telatar, *Capacity of Multi-antenna Gaussian Channels*, Technical Report BL0112170950615-07TM, AT&T Bell Laboratories, 1995.
11. G. J. Foschini and M. J. Gans, On limits of wireless communications in a fading environment when using multiple antennas, *Wireless Pers. Commun.* **6**(3): 311–335 (March 1998).
12. C. Chuah, D. Tse, and J. M. Kahn, Capacity of multi-antenna array systems in indoor wireless environment, *Proc. IEEE GLOBECOM*, 1998, Vol. 4, pp. 1894–1899.
13. G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, Detection algorithm and initial laboratory results using the V-BLAST space-time communication architecture, *Electron. Lett.* **35**(1): 14–16 (Jan. 1999).
14. S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**(8): 1451–1458 (Oct. 1998).
15. H. Bölcskei et al., Fixed broadband wireless: State of the art, challenges and future directions, *IEEE Commun. Mag.* **39**(1): 100–108 (Jan. 2001).

MINIMUM-SHIFT-KEYING

MARVIN K. SIMON
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

1. INTRODUCTION

Minimum-shift-keying (MSK), originally invented by Doelz and Heald as disclosed in a 1961 U.S. Patent [1], is a constant envelope digital modulation that combines both power and bandwidth efficiencies. Signals with constant envelope are desirable when communicating over nonlinear channels, such as those whose transmitter contain a traveling wave tube (TWT) amplifier operated near power saturation, in order to eliminate the occurrence of extraneous spectral sidelobes brought about by amplitude fluctuations. In its native form (also see Hutchinson's 1973 U.S. Patent [2]), MSK is simply a form of binary frequency-shift-keying (BFSK) whose phase is kept continuous from data bit interval to data bit interval and whose modulation index (frequency deviation ratio: the ratio of peak-to-peak frequency deviation to data bit rate) is equal to 0.5. The term *minimum* in this context refers to the fact that, for a given information rate, the 0.5 modulation index corresponds to the minimum frequency shift (and thus the minimum bandwidth) that guarantees orthogonality of the two possible transmitted signals when coherent detection is employed at the receiver which in turn produces maximum power efficiency. Also implicit in the definition of MSK is the fact that the frequency that characterizes the modulation in each bit interval is, as in conventional (not necessarily phase continuous) BFSK, constant over this interval (equivalently, the frequency pulse is a rectangle of duration equal to the bit time).

While at first glance, it might appear that MSK and orthogonal BFSK should have similar performances and behaviors, the fact that the phase is kept continuous in the former introduces memory into the modulation that allows for important differences in the spectral behavior of the transmitted signal as well as the manner in which it can be coherently detected at the receiver. Specifically, the introduction of memory into the modulation produces spectral sidelobes that decay much more rapidly than do those of its conventional binary modulation counterparts. Furthermore, in contrast to a bit-by-bit (bitwise) detector, the deployment of a receiver that exploits the memory introduced at the transmitter offers a power performance more typical of binary *antipodal* signaling than that of binary *orthogonal* signaling.

Although at the time of its introduction MSK had significance in its own right, it gained increased popularity later on when viewed as a special case of a more generic modulation technique referred to as *continuous phase frequency modulation* (CPFM) or more simply *continuous phase modulation* (CPM) whose properties and performance characteristics are well documented in the textbook by Anderson et al. [3]. In particular, CPM allowed for modulation indices other than 0.5, frequency pulse shapes other than rectangular, and frequency pulse durations larger than a single bit time. In fact, it is the distinction between frequency pulses of a single bit duration and those that are longer that accounts for the classification of CPM into *full response* and *partial response* schemes, respectively. Clearly from our discussion above, MSK would fall into the full response CPM category. Furthermore, within the class of full response CPMs, the subclass of schemes having modulation index 0.5 but arbitrary frequency pulse shape resulted in a form of *generalized MSK* [4][1] and included as a special case Amoroso's *sinusoidal FSK* (SFSK) [7] possessing a sinusoidal (raised cosine) frequency pulse shape. Finally, the class of full response schemes with rectangular frequency pulse but arbitrary modulation index is referred to as *continuous phase frequency-shift-keying* (CPFSK) [8] and for all practical purposes served as the precursor to what later became known as CPM itself.

While the primary intent of this article is to focus specifically on the properties and performance of MSK in the form it is most commonly known, the reader should bear in mind that many of these very same characteristics, such as transmitter/receiver implementations, equivalent inphase-quadrature (I-Q) signal representations, and spectral and error probability analysis tools, apply equally well to generalized MSK. Whenever convenient, we shall draw attention to these analogies so as to alert the reader to the generality of our discussions.

In accordance with the above introduction, we begin the mathematical treatment by portraying MSK as a special case of the more general CPM signal whose characterization is given in the next section.

2. THE CONTINUOUS PHASE FREQUENCY MODULATION REPRESENTATION OF MSK

A binary single mode (one modulation index for all transmission intervals) continuous phase modulation (CPM) signal is a constant envelope waveform that has the generic form (see the implementation in Fig. 1).

$$s(t) = \sqrt{\frac{2E_b}{T_b}} \cos(2\pi f_c t + \phi(t, \boldsymbol{\alpha}) + \phi_0), \quad nT_b \le t \le (n+1)T_b \tag{1}$$

where E_b and T_b respectively denote the energy and duration of a bit ($P = E_b/T_b$ is the signal power), and f_c is the carrier frequency. In addition, $\phi(t, \boldsymbol{\alpha})$ is the phase modulation process, which is expressable in the form

$$\phi(t, \boldsymbol{\alpha}) = 2\pi \sum_{i \le n} \alpha_i h q(t - iT_b) \tag{2}$$

where $\boldsymbol{\alpha} = (\ldots, \alpha_{-2}, \alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \ldots)$ is an independent identically distributed (i.i.d.) binary data sequence with each element taking on equiprobable values ± 1, $h = 2\Delta f T_b$ is the *modulation index* (Δf is the peak frequency deviation

[1] Several other authors [5,6] coined the phrase "generalized MSK" to represent generalizations of MSK other than by pulse shaping.

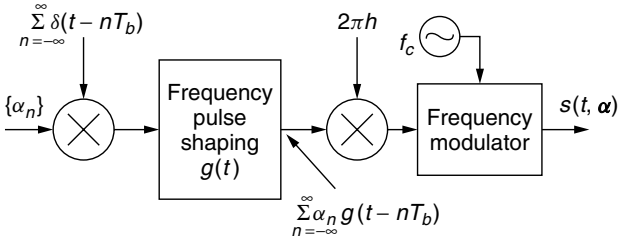


Figure 1. CPM transmitter.

of the carrier), and $q(t)$ is the *normalized phase smoothing response* that defines how the underlying phase $2\pi\alpha_i h$ evolves with time during the associated bit interval. Without loss of generality, the arbitrary phase constant ϕ_0 can be set to zero.

For our discussion here it is convenient to identify the derivative of $q(t)$, namely

$$g(t) = \frac{dq(t)}{dt} \tag{3}$$

which represents the *instantaneous frequency pulse* (relative to the nominal carrier frequency f_c) in the zeroth signaling interval. In view of Eq. (3), the phase smoothing response is given by

$$q(t) = \int_{-\infty}^{t} g(\tau)\, d\tau \tag{4}$$

which, in general, extends over infinite time. For full response CPM schemes, as will be the case of interest here, $q(t)$ satisfies the following:

$$q(t) = \begin{cases} 0, & t \le 0 \\ \frac{1}{2}, & t \ge T_b \end{cases} \tag{5}$$

and thus the frequency pulse $g(t)$ is nonzero only over the bit interval $0 \le t \le T_b$. In view of Eq. (5), we see that the ith data symbol α_i contributes a phase change of $\pi\alpha_i h$ radians to the total phase for all time after T_b seconds of its introduction and thus this fixed phase contribution extends over all future symbol intervals. Because of this overlap of the phase smoothing responses, the total phase in any signaling interval is a function of the present data symbol as well as all of the past symbols and accounts for the *memory* associated with this form of modulation. Thus, in general, optimum detection of CPM schemes must be performed by a *maximum-likelihood sequence estimator* (MLSE) form of receiver [9] as opposed to bit-by-bit detection, which is optimum for memoryless modulations such as conventional BFSK with discontinuous phase.

As previously mentioned, MSK is a full response CPM scheme with a modulation index $h = 0.5$ and a rectangular frequency pulse mathematically described by

$$g(t) = \begin{cases} \dfrac{1}{2T_b}, & 0 \le t \le T_b \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

For SFSK, one of the generalized MSK schemes mentioned in the introduction, $g(t)$ would be a raised cosine pulse given by

$$g(t) = \begin{cases} \dfrac{1}{2T_b}\left[1 - \cos\left(\dfrac{2\pi t}{T_b}\right)\right], & 0 \le t \le T_b \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The associated phase pulses defined by Eq. (4) are

$$q(t) = \begin{cases} \dfrac{t}{2T_b}, & 0 \le t \le T_b \\ \dfrac{1}{2}, & t \ge T_b \end{cases} \tag{8}$$

for MSK and

$$q(t) = \begin{cases} \dfrac{1}{2T_b}\left[t - \dfrac{\sin 2\pi t/T_b}{2\pi/T_b}\right], & 0 \le t \le T_b \\ \dfrac{1}{2}, & t \ge T_b \end{cases} \tag{9}$$

for SFSK.

Finally, substituting $h = 0.5$ and $g(t)$ of (6) in (1) combined with Eq. (2) gives the CPM representations of MSK and SFSK, respectively, as

$$s_{\text{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left(2\pi f_c t + \frac{\pi}{2T_b}\sum_{i \le n} \alpha_i (t - iT_b)\right), \quad nT_b \le t \le (n+1)T_b \tag{10}$$

and

$$s_{\text{SFSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left(2\pi f_c t + \frac{\pi}{2T_b} \times \sum_{i \le n} \alpha_i \left[t - iT_b - \frac{\sin 2\pi(t - iT_b)/T_b}{2\pi/T_b}\right]\right), \quad nT_b \le t \le (n+1)T_b \tag{11}$$

both of which are implemented as in Fig. 1 using $g(t)$ of Eqs. (6) or (7) as appropriate.

Associated with MSK (or SFSK) is a *phase trellis* that illustrates the evolution of the phase process with time corresponding to all possible transmitted sequences. For MSK, the phase variation with time is linear [see Eq. (8)] and thus paths in the phase trellis are straight lines with a slope of $\pm\pi/2T_b$. Figure 2 illustrates the MSK phase trellis where the branches are labeled with the data bits that produce the corresponding phase transition. Note that the change in phase over a single bit time is either $\pi/2$ or $-\pi/2$ depending on the polarity of the data bit α_i corresponding to that bit time. Also note that the trellis is *time-varying* in that the phase states (modulo 2π) alternate between 0 and π at even multiples of the bit time and $\pi/2$ and $3\pi/2$ at odd multiples of the bit time. For SFSK the phase trellis would appear as in Fig. 2 with, however, a sinusoidal variation in phase superimposed over the straight line paths. Here again the change in phase over a single bit time would be

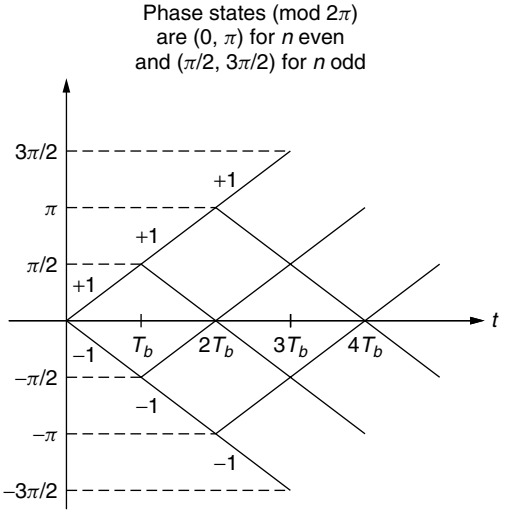


Figure 2. Phase trellis (time-varying) for conventional MSK.

either $\pi/2$ or $-\pi/2$ depending on the polarity of the data bit α_i corresponding to that bit time.

3. EQUIVALENT I-Q REPRESENTATION OF MSK

Although, as stated above, CPM schemes because of their inherent memory require a memory-type of detection, such as MLSE, full response modulations with $h = 0.5$ such as MSK and SFSK can in fact be detected using a memoryless I-Q form of receiver. The reason for this is that for these modulations the transmitter can be implemented in an I-Q form analogous to that of *offset quadrature-phase-shift-keying* (OQPSK). To see this mathematically, we first rewrite the excess phase in the nth transmission interval of the MSK signal in (10) as

$$\phi(t,\boldsymbol{\alpha}) = \frac{\pi}{2T_b}\sum_{i\le n}\alpha_i(t-iT_b) = \alpha_n\frac{\pi}{2T_b}(t-nT_b) + \frac{\pi}{2}\sum_{i\le n-1}\alpha_i = \alpha_n\frac{\pi}{2T_b}t + x_n, \quad nT_b \le t \le (n+1)T_b \tag{12}$$

where $(\pi/2)\sum_{i\le n-1}\alpha_i$ is the accumulated phase at the beginning of the nth transmission interval, which is equal to an odd integer (positive or negative) multiple of $\pi/2$ when n is odd and an even integer (positive or negative) multiple of $\pi/2$ when n is even, and x_n is a phase constant required to keep the phase continuous at the data transition points $t = nT_b$ and $t = (n+1)T_b$. Note also that x_n represents the y-intercept (when reduced modulo 2π) of the path in the phase trellis that represents $\phi(t,\boldsymbol{\alpha})$. In the previous transmission interval, the excess phase is given by

$$\phi(t,\boldsymbol{\alpha}) = \alpha_n\frac{\pi}{2T_b}(t-(n-1)T_b) + \frac{\pi}{2}\sum_{i\le n-2}\alpha_i = \alpha_{n-1}\frac{\pi}{2T_b}t + x_{n-1}, \quad (n-1)T_b \le t \le nT_b \tag{13}$$

For phase continuity at $t = nT_b$, we require that

$$\alpha_n\frac{\pi}{2T_b}(nT_b) + x_n = \alpha_{n-1}\frac{\pi}{2T_b}(nT_b) + x_{n-1} \tag{14}$$

or equivalently

$$x_n = x_{n-1} + \frac{\pi n}{2}(\alpha_{n-1} - \alpha_n) \tag{15}$$

Equation (15) is a recursive relation that allows x_n to be determined in any transmission interval given an initial condition, x_0.

We observe that $(\alpha_{n-1} - \alpha_n)/2$ is a ternary random variable (RV) taking on values $0, +1, -1$ with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$, respectively. Thus, from Eq. (15) when $\alpha_{n-1} = \alpha_n$, $x_n = x_{n-1}$ whereas when $\alpha_{n-1} \ne \alpha_n$, $x_n = x_{n-1} \pm \pi n$. If we arbitrary choose the initial condition $x_0 = 0$, then we see that x_n takes on values of 0 or π (when reduced modulo 2π). Using this fact in (12) and applying simple trigonometry to (10), we obtain

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}[\cos\phi(t,\boldsymbol{\alpha})\cos 2\pi f_c t - \sin\phi(t,\boldsymbol{\alpha})\sin 2\pi f_c t], \quad nT_b \le t \le (n+1)T_b \tag{16}$$

where

$$\begin{aligned}\cos\phi(t,\boldsymbol{\alpha}) &= \cos\left(\alpha_n\frac{\pi}{2T_b}t + x_n\right) \\ &= a_n\cos\frac{\pi}{2T_b}t,\ a_n = \cos x_n = \pm 1 \\ \sin\phi(t,\boldsymbol{\alpha}) &= \sin\left(\alpha_n\frac{\pi}{2T_b}t + x_n\right) \\ &= \alpha_n a_n\sin\frac{\pi}{2T_b}t = b_n\sin\frac{\pi}{2T_b}t, \\ b_n &= \alpha_n\cos x_n = \pm 1\end{aligned} \tag{17}$$

Finally, substituting (17) in (16) gives the I-Q representation of MSK as

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}[a_n C(t)\cos 2\pi f_c t - b_n S(t)\sin 2\pi f_c t], \quad nT_b \le t \le (n+1)T_b \tag{18}$$

where

$$C(t) = \cos\frac{\pi t}{2T_b}, \quad S(t) = \sin\frac{\pi t}{2T_b} \tag{19}$$

are the effective I and Q pulse shapes and $\{a_n\}, \{b_n\}$ as defined in (17) are the effective I and Q binary data sequences.

For SFSK, the representation of Eq. (18) would still be valid with a_n, b_n as defined in (17) but now the effective I and Q pulse shapes become

$$C(t) = \cos\left[\frac{\pi}{2T_b}\left(t - \frac{\sin 2\pi t/T_b}{2\pi/T_b}\right)\right], \quad S(t) = \sin\left[\frac{\pi}{2T_b}\left(t - \frac{\sin 2\pi t/T_b}{2\pi/T_b}\right)\right] \tag{20}$$

To tie the representation of (18) back to that of FSK, we observe that

$$C(t)\cos 2\pi f_c t = \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] + \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right]$$
$$S(t)\sin 2\pi f_c t = -\frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] + \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \quad (21)$$

Substituting (21) in (18) gives

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{a_n + b_n}{2}\right)\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] + \left(\frac{a_n - b_n}{2}\right)\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \quad (22)$$

Thus, when $a_n = b_n (\alpha_n = 1)$, we have

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] \quad (23)$$

whereas when $a_n \neq b_n (\alpha_n = -1)$, we have

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \quad (24)$$

which establishes the desired connection.

Note from (19) that since $C(t)$ and $S(t)$ are offset from each other by a time shift of T_b seconds, it might appear that $s_{\mathrm{MSK}}(t)$ of (18) is in the form of OQPSK with half sinusoidal pulse shaping.[2] To justify that this is indeed the case, we must examine more carefully the effective I and Q data sequences $\{a_n\}, \{b_n\}$ in so far as their relationship to the input data sequence $\{\alpha_i\}$ and the rate at which they can change. Since the input α_n data bit can change every bit time it might appear that the effective I and Q data bits a_n and b_n can also change every bit time. To the contrary, it can be shown that as a result of the phase continuity constraint of (15), $a_n = \cos x_n$ can change only at the zero crossings of $C(t)$, whereas $b_n = \alpha_n \cos x_n$ can change only at the zero crossings of $S(t)$. Since the zero crossings of $C(t)$ and $S(t)$ are each spaced $2T_b$ seconds apart, then a_n and b_n are constant over $2T_b$-second intervals (see Fig. 3 for an illustrative example). Further noting that the continuous waveforms $C(t)$ and $S(t)$ alternate in sign every $2T_b$ seconds, we can incorporate this sign change in the I and Q data sequences themselves and deal with a fixed *positive* time-limited pulse shape on each

[2] A similar statement can be made for SFSK where the pulse shaping is now described by Eq. (20).

k	α_k	$x_k \pmod{2\pi}$	a_k	b_k	Time interval
0	1	0	1	1	$0 \le t \le T_b$
1	−1	π	−1	1	$T_b \le t \le 2T_b$
2	−1	π	−1	1	$2T_b \le t \le 3T_b$
3	1	0	1	1	$3T_b \le t \le 4T_b$
4	1	0	1	1	$4T_b \le t \le 5T_b$
5	1	0	1	1	$5T_b \le t \le 6T_b$
6	−1	0	1	−1	$6T_b \le t \le 7T_b$
7	1	π	−1	−1	$7T_b \le t \le 8T_b$
8	−1	π	−1	1	$8T_b \le t \le 9T_b$

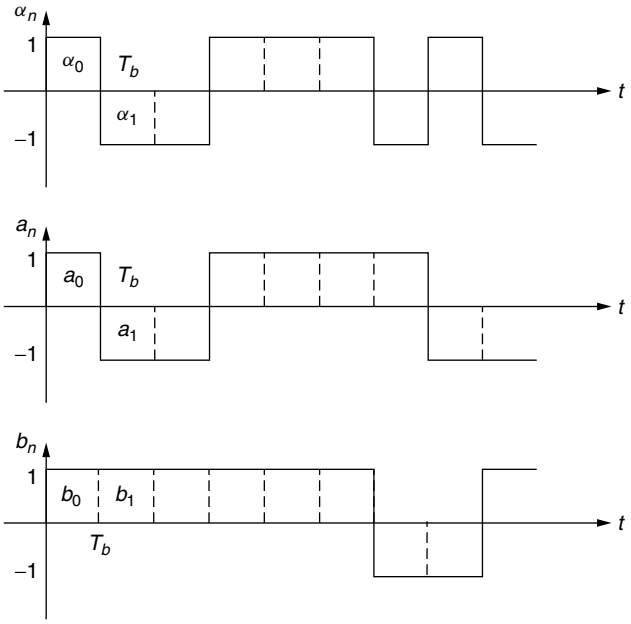


Figure 3. An example of the equivalent I and Q data sequences represented as rectangular pulse streams.

of the I and Q channels. Specifically, defining the pulse shape

$$p(t) = \begin{cases} \sin\dfrac{\pi t}{2T_b}, & 0 \le t \le 2T_b \\ 0, & \text{otherwise} \end{cases} \quad (25)$$

then the I-Q representation of MSK can be rewritten in the form

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}[d_c(t)\cos 2\pi f_c t - d_s(t)\sin 2\pi f_c t] \quad (26)$$

where

$$d_c(t) = \sum_n c_n p(t - (2n-1)T_b), \quad d_s(t) = \sum_n d_n p(t - 2nT_b) \quad (27)$$

with

$$c_n = (-1)^n a_{2n-1}, \quad d_n = (-1)^n b_{2n} \quad (28)$$

To complete the analogy between MSK and sinusoidally pulse shaped OQPSK, we must examine the manner in which the equivalent I and Q data sequences needed in (28) are obtained from the input data sequence $\{\alpha_n\}$. Without going into great mathematical detail, suffice it to say that it can be shown that the sequences

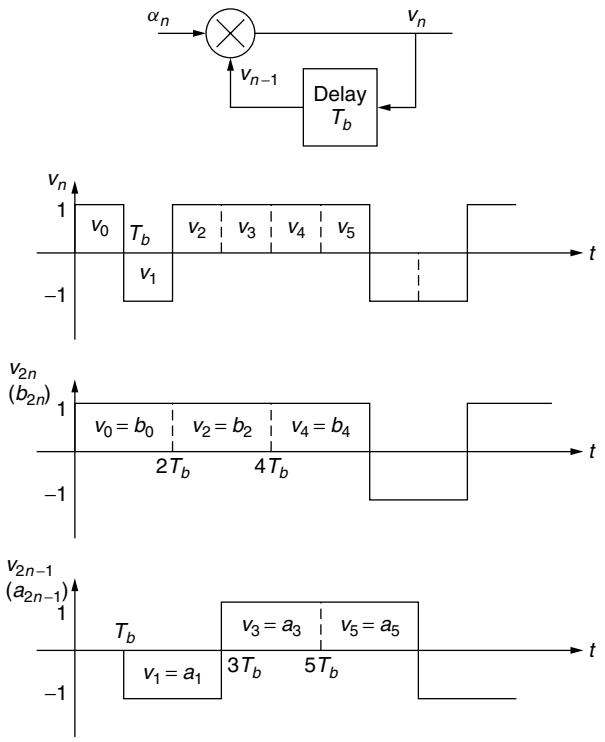


Figure 4. An example of the equivalence between differentially encoded input bits and effective I and Q bits.

$\{a_{2n-1}\}$ and $\{b_{2n}\}$ are the odd/even split of a sequence $\{v_n\}$ which is the *differentially encoded* version of $\{\alpha_n\}$, i.e., $v_n = \alpha_n v_{n-1}$ (see Fig. 4 for an illustrative example). Finally, the I-Q implementation of MSK as described by (27) is illustrated in Fig. 5. As anticipated, we observe that this figure resembles a transmitter for OQPSK except that here the pulse shaping is half-sinusoidal (of symbol duration $T_s = 2T_b$) rather than rectangular, and in addition a differential encoder is applied to the input data sequence prior to splitting it into even and odd sequences each at a rate $1/T_s$. The interpretation of MSK as a special case of OQPSK with sinusoidal pulse shaping along with tradeoffs and comparisons between the two modulations is discussed further in the literature [10,11].

Before concluding this section, we note that the alternative representation of MSK as in (22) can be also expressed in terms of the differentially encoded bits, v_n. In particular

For n odd

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{v_{n-1}+v_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] - \left(\frac{v_{n-1}-v_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \tag{29a}$$

For n even

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{v_{n-1}+v_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] + \left(\frac{v_{n-1}-v_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \tag{29b}$$

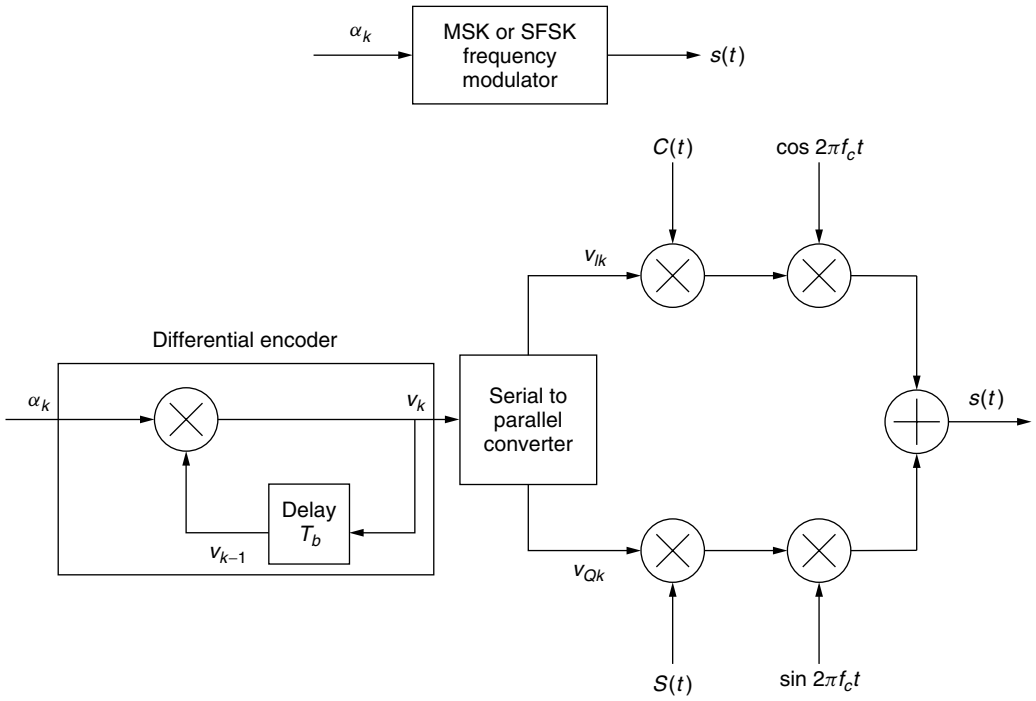


Figure 5. CPM and equivalent I-Q implementations of MSK or SFSK.

Combining these two results, we get

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{v_{n-1}+v_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] + (-1)^n\left(\frac{v_{n-1}-v_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\right], \quad nT_b \le t \le (n+1)T_b \tag{30}$$

4. PRECODED MSK

The differential encoder that precedes the I-Q portion of the transmitter in Fig. 5 requires a compensating differential decoder at the receiver following I-Q demodulation and detection (see Fig. 6). Such a combination of differential encoding at the transmitter and differential decoding at the receiver results in a loss in power performance relative to that obtained by conventional OQPSK (this is discussed in more detail later in the article). It is possible to modify MSK to avoid such a loss by first recognizing that the CPM form of modulator in Fig. 1 for implementing MSK can be preceded by the cascade of a differential encoder and a differential decoder without affecting its output (Fig. 7); that is, the cascade of a differential encoder and a differential decoder produces unity transmission, where input = output. Thus, comparing Fig. 7 with Fig. 5, we observe that precoding the CPM form of MSK modulator with a differential *decoder* resulting in what is referred to as *precoded MSK* [9, Chap. 10] will be equivalent to the I-Q implementation of the latter without the differential encoder at its input (see Fig. 8), and thus the receiver for precoded MSK is that of Fig. 6 without the differential decoder at its output. It goes without saying that a similar precoding applied to SFSK would also allow for dispensing with the differential decoder at the output of its I-Q receiver. Finally, we note that both MSK (or SFSK) and its precoded version have identical spectral characteristics and thus for all practical purposes the improvement

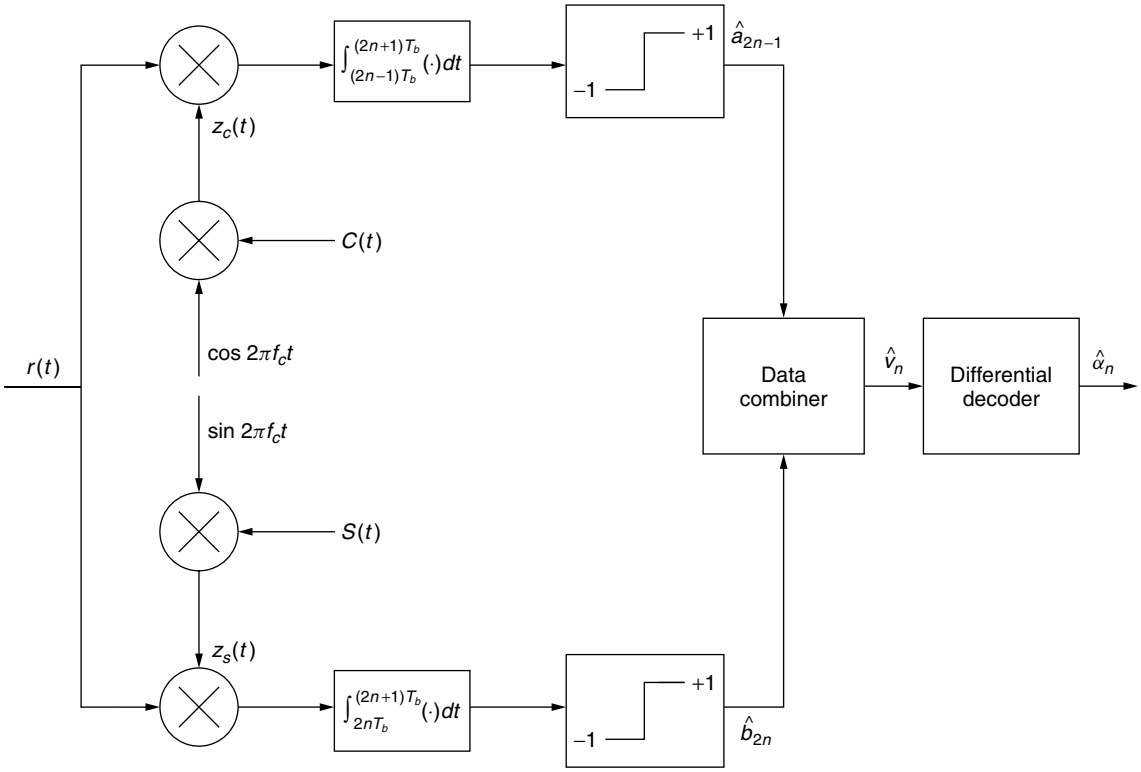


Figure 6. An I-Q receiver implementation of MSK.

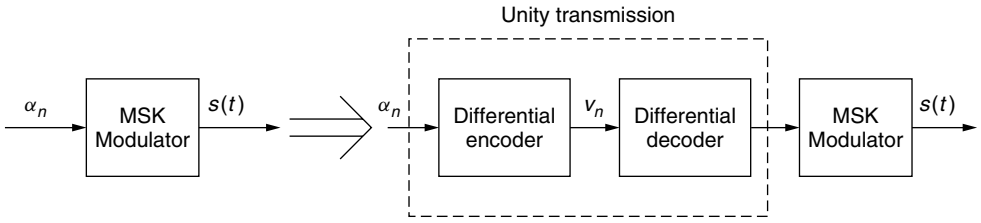


Figure 7. Two equivalent MSK transmitters.

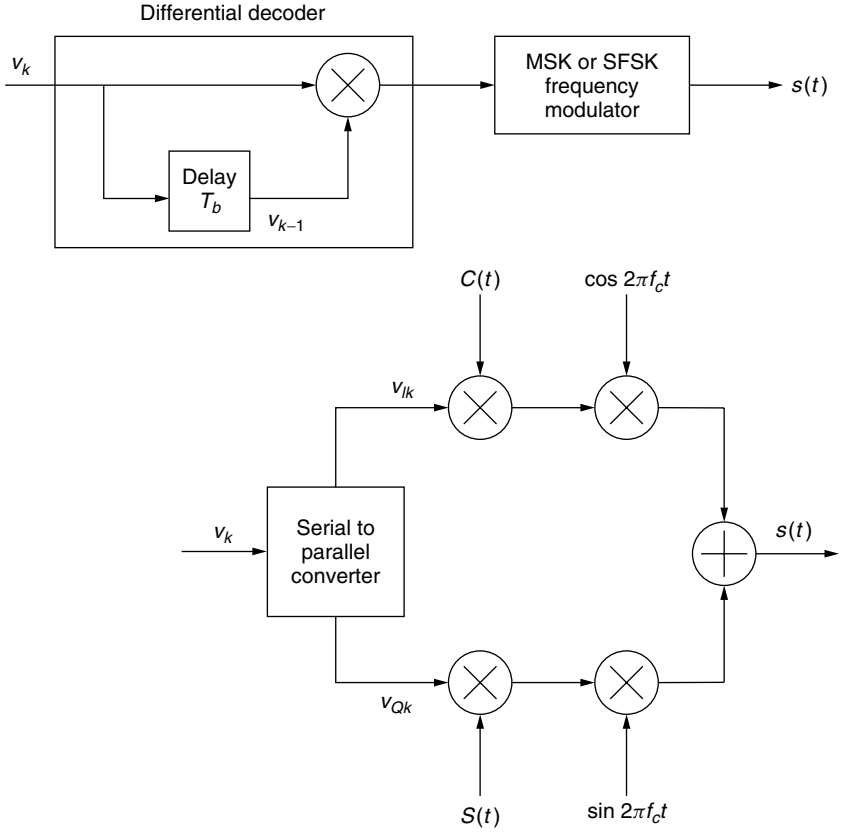


Figure 8. CPM and equivalent I-Q implementations of precoded MSK or SFSK.

in power performance provided by the latter comes at no expense.

5. SPECTRAL CHARACTERISTICS

The ability to express MSK in the offset I-Q form of Eq. (18) allows for simple evaluation of its power spectral density (PSD). In particular, for a generic offset I-Q modulation formed by impressing two lowpass modulations (random pulse trains of rate $1/2T_b$) of equal power and pulse shape on inphase and quadrature carriers:

$$\begin{aligned} s(t) &= Am_I(t)\cos 2\pi f_c t - Am_Q(t)\sin 2\pi f_c t \\ m_I(t) &= \sum_n a_n p(t-2nT_b),\ m_Q(t) = \sum_n b_n p(t-(2n-1)T_b) \end{aligned} \tag{31}$$

the PSD is given by [9, Chap. 2]

$$S_s(f) = \tfrac{1}{4}[G(f-f_c) + G(f+f_c)] \tag{32}$$

where $G(f)$ is the equivalent baseband PSD and is related to the PSD, $S_m(f)$, of $m_I(t)$ or $m_Q(t)$ by

$$G(f) = 2A^2 S_m(f); \quad S_m(f) = \frac{1}{2T_b}|P(f)|^2 \tag{33}$$

with $P(f)$ denoting the Fourier transform of the pulse shape $p(t)$. For MSK, we would have $A = \sqrt{2E_b/T_b}$ and $p(t)$ given by (25) with Fourier transform

$$P(f) = \frac{4T_b}{\pi} e^{-j2\pi f T_b} \frac{\cos 2\pi f T_b}{1 - 16f^2 T_b^2} \tag{34}$$

Substituting (34) in (33) gives the equivalent baseband PSD of MSK as

$$G(f) = \frac{32E_b}{\pi^2} \frac{\cos^2 2\pi f T_b}{(1 - 16f^2 T_b^2)^2} \tag{35}$$

and the corresponding bandpass PSD as [9, Chap. 2]

$$S_s(f) = \frac{8E_b}{\pi^2}\left[\frac{\cos^2 2\pi(f-f_c)T_b}{(1-16(f-f_c)^2T_b^2)^2} + \frac{\cos^2 2\pi(f+f_c)T_b}{(1-16(f+f_c)^2T_b^2)^2}\right] \tag{36}$$

We observe from (35) that the main lobe of the lowpass PSD has its first null at $f = 3/4T_b$. Also, asymptotically for large f, the spectral sidelobes roll off at a rate f^{-4}. By comparison, the equivalent PSD of OQPSK wherein $A = \sqrt{E_b/T_b}$ and $p(t)$ is a unit amplitude rectangular pulse of duration $2T_b$, is given by

$$G(f) = 4E_b \frac{\sin^2 2\pi f T_b}{(2\pi f T_b)^2} \tag{37}$$

whose main lobe has its first null at $f = \frac{1}{2}T_b$ and whose spectral sidelobes asymptotically roll off at a rate f^{-2}. Thus, we observe that while MSK (or precoded MSK) has

a wider main lobe than OQPSK (or QPSK) by a factor of $\frac{3}{2}$, its spectral sidelobes roll off at a rate two orders of magnitude faster. Figure 9 is an illustration of the normalized lowpass PSDs, $G(f)/2E_b$, of MSK and OQPSK obtained from (35) and (37), respectively, as well as that of SFSK, which is given by [9, Chap. 2]

$$G(f) = 2E_b\left[J_0\left(\frac{1}{4}\right)A_0(f) + 2\sum_{n=1}^{\infty} J_{2n}\left(\frac{1}{4}\right)B_{2n}(f) + 2\sum_{n=1}^{\infty} J_{2n-1}\left(\frac{1}{4}\right)B_{2n-1}(f)\right]^2$$

$$A(f) = 2\left(\frac{\sin 2\pi f T_b}{2\pi f T_b}\right), A_0(f) = \frac{1}{2}A\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A\left(f - \frac{1}{4T_b}\right) = \frac{4}{\pi}\frac{\cos 2\pi f T_b}{1 - 16f^2T_b^2}$$

$$A_{2n}(f) = \frac{1}{2}A\left(f + \frac{2n}{T_b}\right) + \frac{1}{2}A\left(f - \frac{2n}{T_b}\right), \quad (38)$$

$$A_{2n-1}(f) = \frac{1}{2}A\left(f + \frac{2n-1}{T_b}\right) - \frac{1}{2}A\left(f - \frac{2n-1}{T_b}\right)$$

$$B_{2n}(f) = \frac{1}{2}A_{2n}\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A_{2n}\left(f - \frac{1}{4T_b}\right),$$

$$B_{2n-1}(f) = -\frac{1}{2}A_{2n-1}\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A_{2n-1}\left(f - \frac{1}{4T_b}\right)$$

$J_n(x) = n$th order Bessel function of the first kind

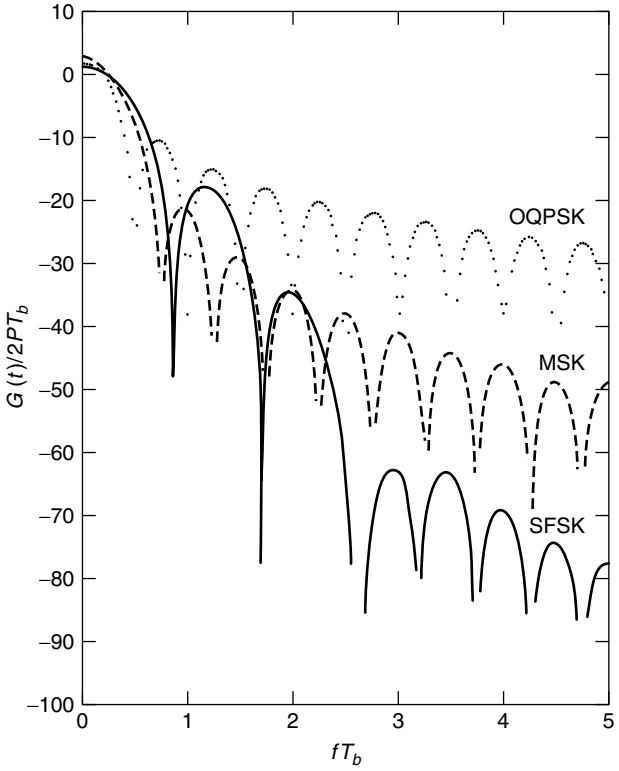


Figure 9. A comparison of the equivalent baseband PSDs of MSK, OQPSK, and SFSK.

whose main lobe is wider than that of MSK but whose spectral sidelobes asymptotically roll off four orders of magnitude faster: at a rate f^{-8}. In fact, for the class of generalized MSK schemes, we can conclude that the smoother we make the shape of the frequency pulse; specifically, the more derivatives that go to zero at the endpoints $t = 0$ and $t = 2T_b$, the wider will be the main lobe but the faster the sidelobes will roll off.

Another way of interpreting the improved bandwidth efficiency that accompanies the equivalent I and Q pulse shaping is in terms of the fractional out-of-band power defined as the fraction of the total power that lies outside a given bandwidth:

$$\eta = 1 - \frac{\int_{-B/2}^{B/2} G(f)\,df}{\int_{-\infty}^{\infty} G(f)\,df} \quad (39)$$

Figure 10 is a plot of the fractional out-of-band power (in decibels) versus BT_b for MSK, OQPSK, and SFSK using the appropriate expression for $G(f)$ as determined from Eqs. (35), (37), and (38), respectively.

6. SERIAL MSK

As an alternative to the parallel I-Q implementations previously discussed, Amoroso and Kivet [12] suggested a serial implementation[3] of MSK which, when the ratio of

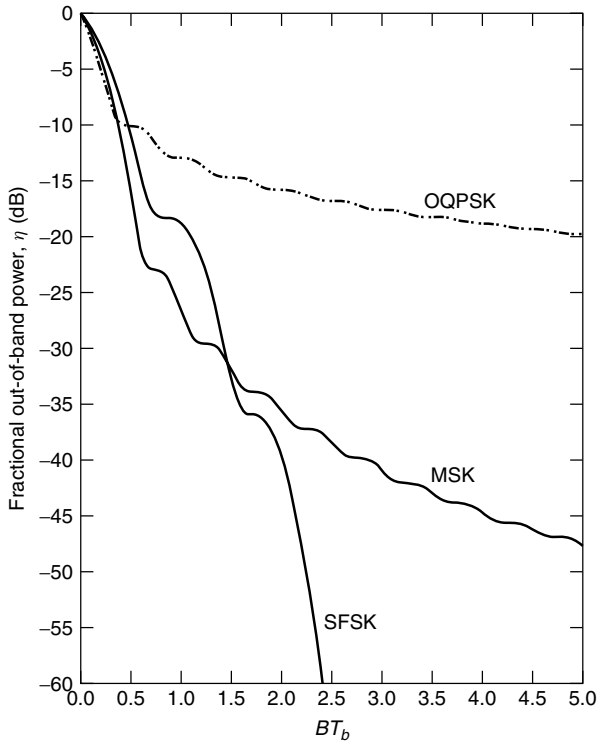


Figure 10. A comparison of the fractional out-of-band power performance of MSK, OQPSK, and SFSK.

[3] Other investigations of serial type implementations appear in [13,14].

carrier frequency to bit rate is high, avoids the requirement for precise relative phasing between the pair of transmitter oscillators needed in the former. In this implementation, MSK is synthesized using a simple biphase modulator that accepts the data in serial form (as opposed to splitting it into I and Q sequences) and as such biphase demodulation and detection are performed at the receiver. The operation of this synthesis method depends on compliance with the original frequency modulation concept of MSK proposed by Doelz and Heald [1] together with additional constraints imposed by Sullivan [15].

Figure 11 is a block diagram of the serial MSK transmitter and receiver. During any T_b-second interval, one of two frequencies f_1 or f_2 is transmitted where for some selected integer n

$$f_1 = \frac{n+1}{2T_b}, \quad f_2 = \frac{n}{2T_b} \tag{40}$$

The carrier frequency f_c is thought of as being midway between f_1 and f_2, namely, $f_c = (f_1 + f_2)/2 = (n + \frac{1}{2})/2T_b$ although it is actually never generated in this implementation. Note that, independent of n, the modulation index $h = (f_1 - f_2)/(1/T_b) = 0.5$ as required for MSK. The operation of the transmitter that synthesizes MSK is as follows. With reference to Fig. 11, a binary rectangular pulse train whose generating sequence is the data sequence $\{\alpha_n\}$ is PSK modulated onto a carrier $c(t) = \cos(2\pi f_2 t + \theta)$ producing the signal

$$x(t) = \sqrt{\frac{2E_b}{T_b}} \sum_{n=-\infty}^{\infty} \alpha_n p(t - nT_b) \cos(2\pi f_2 t + \theta) \tag{41}$$

where $p(t)$ is a unit amplitude pulse of duration T_b seconds and θ is a phase constant to be chosen. This signal is then passed through a lowpass filter with impulse response

$$h_T(t) = \begin{cases} \dfrac{\pi}{T_b} \sin 2\pi f_1 t, & 0 \le t \le T_b \\ 0, & \text{otherwise} \end{cases} \tag{42}$$

Convolving $x(t)$ with $h_T(t)$ gives the filter output in the kth transmission interval as (ignoring high frequency terms at $f_1 + f_2$):

$$\begin{aligned} s(t) = {} & \sqrt{\frac{2E_b}{T_b}} \frac{\pi}{2T_b} \alpha_{k-1} \int_{t-T_b}^{kT_b} \sin[2\pi f_1 t + \theta + 2\pi(f_2 - f_1)\tau]\, d\tau \\ & + \sqrt{\frac{2E_b}{T_b}} \frac{\pi}{2T_b} \alpha_k \int_{kT_b}^{t} \sin[2\pi f_1 t + \theta + 2\pi(f_2 - f_1)\tau]\, d\tau, \\ & kT_b \le t \le (k+1)T_b \end{aligned} \tag{43}$$

Note that $s(t)$ depends only on the 2 data bits α_{k-1} and α_k. Evaluating the integrals and using the fact that $(f_2 - f_1)T_b = 0.5$, we obtain after some simplification

$$\begin{aligned} s(t) = {} & \sqrt{\frac{2E_b}{T_b}} \left\{ -\left(\frac{\alpha_{k-1} + \alpha_k}{2}\right) \cos(2\pi f_2 t + \theta) - (-1)^k \right. \\ & \left. \times \left(\frac{\alpha_{k-1} - \alpha_k}{2}\right) \cos(2\pi f_1 t + \theta) \right\}, \\ & kT_b \le t \le (k+1)T_b \end{aligned} \tag{44}$$

Letting $\theta = \pi$, (44) becomes

$$\begin{aligned} s(t) = {} & \sqrt{\frac{2E_b}{T_b}} \left\{ \left(\frac{\alpha_{k-1} + \alpha_k}{2}\right) \cos 2\pi f_2 t + (-1)^k \left(\frac{\alpha_{k-1} - \alpha_k}{2}\right) \right. \\ & \left. \times \cos 2\pi f_1 t \right\}, kT_b \le t \le (k+1)T_b \end{aligned} \tag{45}$$

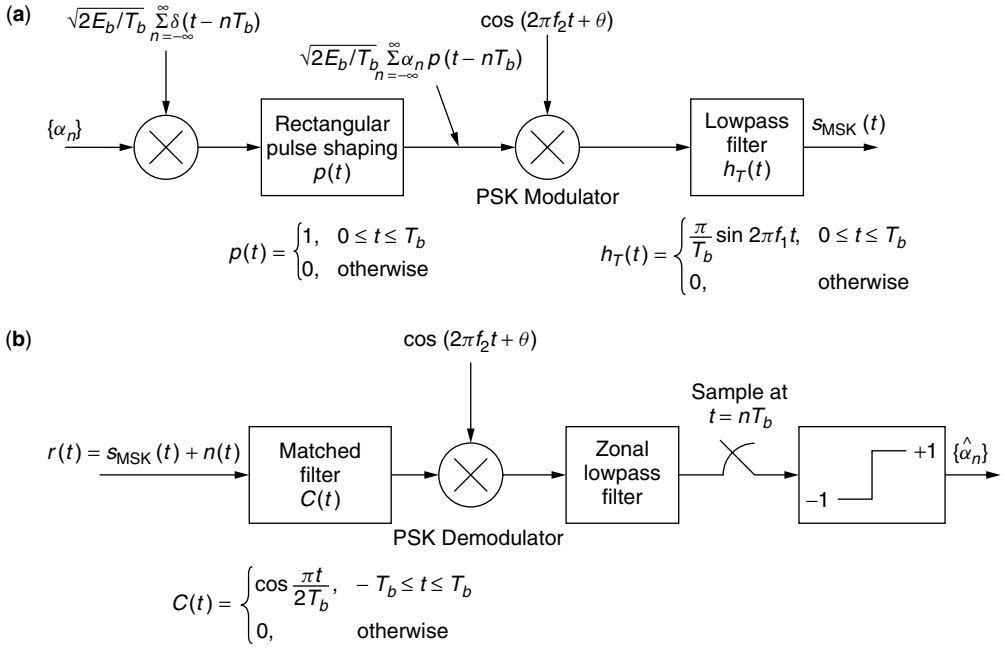


Figure 11. Serial MSK (**a**) transmitter and (**b**) receiver.

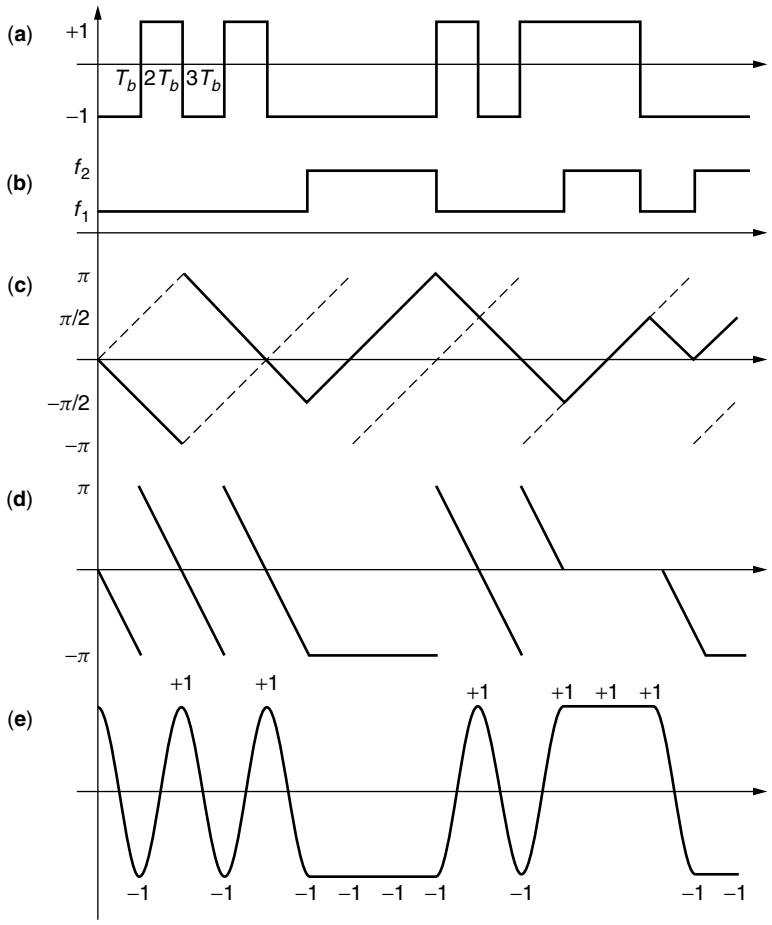


Figure 12. (**a**) Transmitted bit sequence; (**b**) transmitted frequency; (**c**) transmitted excess phase (mod 2π); (**d**) demodulated phase (mod 2π); (**e**) demodulator output (cosine of demodulated phase).

Alternatively, letting $\theta = 0$, we get

$$s(t) = -\sqrt{\frac{2E_b}{T_b}}\left\{\left(\frac{\alpha_{k-1}+\alpha_k}{2}\right)\cos 2\pi f_2 t + (-1)^k\left(\frac{\alpha_{k-1}-\alpha_k}{2}\right)\times\cos 2\pi f_1 t\right\}, kT_b \le t \le (k+1)T_b \quad (46)$$

Comparing Eq. (45) [or (46)] with (30) and noting that $f_1 = f_c - \frac{1}{4T_b}$ and $f_2 = f_c + 1/4T_b$, we see that the serial implementation in Fig. 11 produces a *precoded* MSK signal.[4]

The operation of the serial form of the receiver in Fig. 11 is best described in terms of the series of noise-free waveforms illustrated in Fig. 12 that ignore the presence of the matched filter. Figure 12a is a typical ± 1 data sequence for $\{\alpha_n\}$. Figure 12b shows the transmitted frequency corresponding to this typical data sequence in accordance with (45); that is, frequency f_1 is transmitted when the current bit is different from the previous one and frequency f_2 is transmitted when these 2 bits are the same. The solid line portion of Fig. 12c is the continuous excess (relative to $2\pi f_c t$) phase (reduced modulo 2π) corresponding to the frequency sequence of the second waveform. This is the excess phase of the signal component $s_{\mathrm{MSK}}(t)$ of the received waveform $r(t)$ in Fig. 11b. The excess phase of the PSK demodulation reference is given by $2\pi(f_2 - f_c)t$ and when reduced modulo 2π is illustrated as the dotted line portion of Fig. 12c. The phase of the PSK demodulator output after passing through the zonal lowpass filter (to remove the high frequency carrier term) is given by $\phi(t, \boldsymbol{\alpha}) - 2\pi(f_2 - f_c)t$ and when reduced modulo 2π is illustrated in Fig. 12d. Finally, the actual zonal lowpass filter output is $\cos(\phi(t, \boldsymbol{\alpha}) - 2\pi(f_2 - f_c)t)$, which is illustrated in Fig. 12e. Sampling this waveform at integer multiples of T_b produces a sequence identical to the original data sequence in Fig. 12a. Of course, in the presence of noise, these samples would be noisy, in which case a hard-limiting operation would be used to produce estimates of the data sequence. In the actual implementation of the receiver a matched (to the equivalent I and Q pulse shape) filter would be used which would produce an individual pulse contribution to the recovered bit stream that extends over an interval of $4T_b$ seconds (convolution of a $2T_b$ half-sinusoid with itself). However, it can be shown that the apparent intersymbol interference (ISI) introduced by this broadening of the pulse does not affect the sampled values of the demodulator output (after zonal filtering) and thus no loss

[4] In the Pelchat et al. paper [8], the role of f_1 and f_2 in the transmitter and receiver implementations are reversed with respect to their usage here whose purpose is to maintain consistency with our definition of precoded MSK.

in performance occurs; thus the serial MSK modulation and demodulation system illustrated in Fig. 11 has the same communication efficiency as its parallel counterpart.

Before concluding this section, we note that had we inverted the data sequence in Fig. 12a, the corresponding transmitted frequency sequence of Fig. 12b would remain identical, and likewise the detected data sequence obtained from Fig. 12e would also remain identical. This results in a detected data sequence that is opposite in polarity to the transmitted sequence which implies the presence of a 180° phase ambiguity in the receiver. This type of phase ambiguity is indemic to all binary phase coherent communication systems and as such a means must be provided in the receiver to resolve this ambiguity.

7. CROSS-COUPLED I-Q TRANSMITTER

A variation of the I-Q transmitter discussed in Section 4 is illustrated in Fig. 13 [16–18]. An modulated carrier at frequency f_c is multiplied by a lowpass sinusoidal signal at frequency $1/4T_b$ to produce a pair of unmodulated tones (carriers) at $f_2 = f_c + 1/4T_b$ and $f_1 = f_c - 1/4T_b$. These tones are separately extracted by narrow bandpass filters whose outputs, $s_1(t)$ and $s_2(t)$ are then summed and differenced to produce

$$\begin{aligned} z_c(t) = s_1(t) + s_2(t) &= \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \\ &\quad + \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] = \cos\left(\frac{\pi t}{2T_b}\right)\cos 2\pi f_c t \\ z_s(t) = s_1(t) - s_2(t) &= \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] \\ &\quad - \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] = \sin\left(\frac{\pi t}{2T_b}\right)\sin 2\pi f_c t \end{aligned} \tag{47}$$

The signals $z_c(t)$ and $z_s(t)$ are respectively multiplied by I and Q data sequences $\{\alpha_I\}$ and $\{\alpha_Q\}$ each at a rate of $\frac{1}{2}T_b$ (and offset from each other by T_b seconds) and then differenced to produce the MSK (actually precoded MSK) output. The advantage of the implementation of Fig. 13 is that the signal coherence and the frequency deviation ratio are largely unaffected by variations in the data rate [17].

8. RIMOLDI'S REPRESENTATION

As stated previously, the conventional CPM implementation of MSK produces a phase trellis that is symmetric about the horizontal axis but time-varying in that the possible phase states (reduced modulo 2π) alternate between $(0, \pi)$ and $(\pi/2, 3\pi/2)$ every T_b seconds. To remove this time-variation of the trellis, Rimoldi [19] demonstrated that CPM with a rational modulation index could be decomposed into the cascade of a memory encoder (finite state machine) and a memoryless demodulator (signal waveform mapper). For the specific case of MSK, Rimoldi's transmitter is illustrated in Fig. 14. Unbalanced (0s and 1s) binary 1 bits, $U_n = (1 - \alpha_n)/2$, are input to a memory one encoder. The current bit and the differentially encoded version of the previous bit (the encoder state) are used to define, via a binary coded decimal (BCD) mapping, a pair of baseband signals (each chosen from a set of four possible waveforms) to be modulated onto I and Q carriers for transmission over the channel. Because of the unbalance of the data, the phase trellis is *tilted* as shown in Fig. 15, but on the other hand it is now *time-invariant*; that is, the phase states (reduced modulo 2π) at all time instants (integer multiples of the bit time) are $(0, \pi)$. This transmitter implementation suggests the use of a simple two-state trellis decoder, which is discussed in the next section dealing with memory receiver structures.

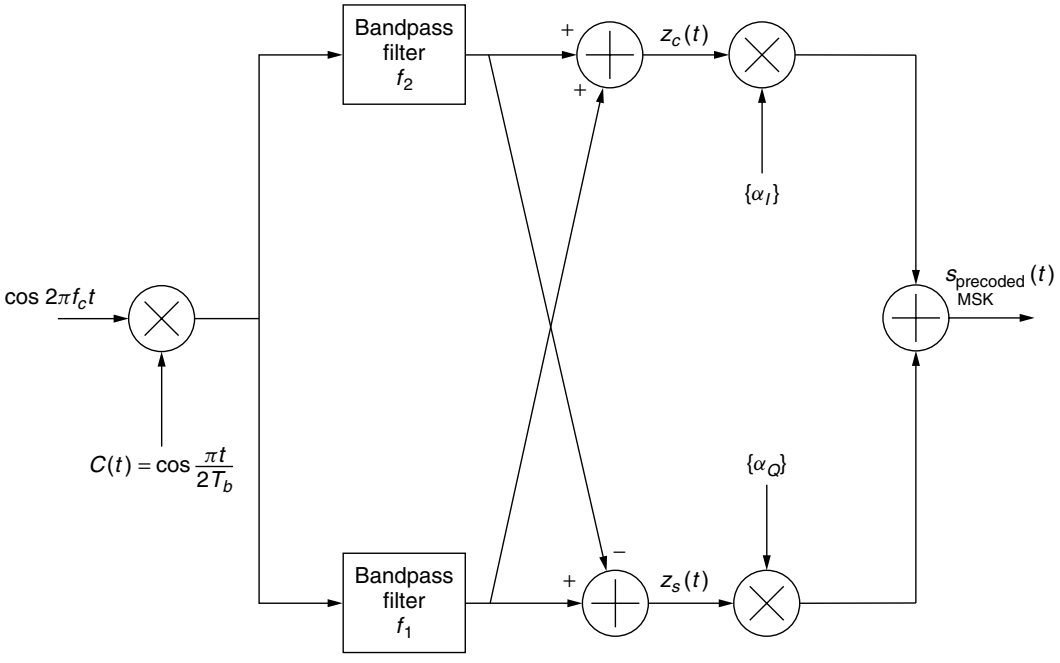


Figure 13. Cross-coupled implementation of precoded MSK.

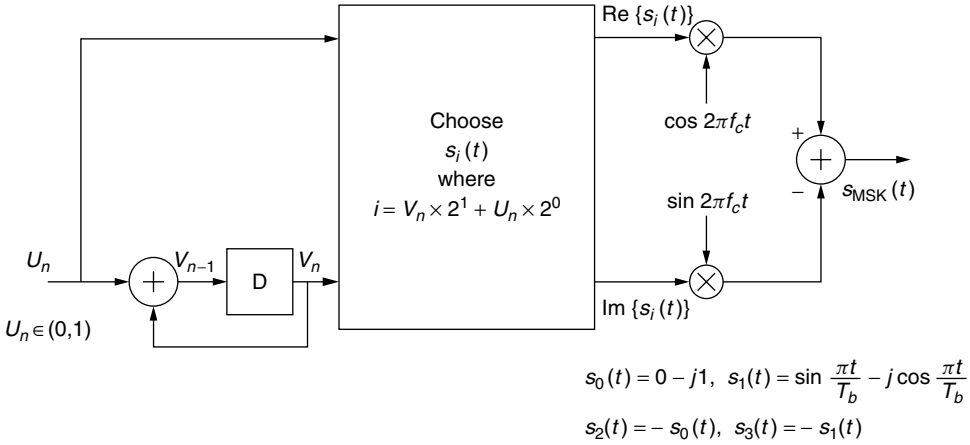


Figure 14. MSK transmitter based on Rimoldi decomposition of CPM.

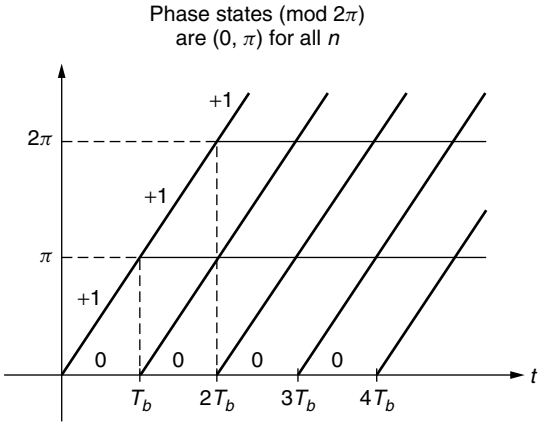


Figure 15. Tilted (time-invariant) phase trellis for Rimoldi's MSK representation.

Rimoldi's representation can also be used to implement precoded MSK. The appropriate transmitter is illustrated in Fig. 16.

9. COHERENT DETECTION

Depending on the particular form used to represent the MSK signal (e.g., CPM, serial or parallel I-Q), many different forms of receivers have been suggested in the literature for performing coherent detection. These various forms fall into two classes: structures based on a memoryless transmitter representation and structures based on a memory transmitter representation. As we shall see, all of these structures, however, are themselves memoryless.

9.1. Structures Based on a Memoryless Transmitter Representation

The two most popular structures for coherent reception of MSK that are based on a memoryless transmitter representation correspond to the parallel I-Q and serial representations and have already been illustrated in Figs. 6 and 11b, respectively. In the case of the former, the received signal plus noise is multiplied by the I and Q "carriers,"[5] $z_c(t)$ and $z_s(t)$, respectively, followed by integrate-and-dump (I&D) circuits of duration $2T_b$ seconds that are timed to match the zero crossings of the I and Q symbol waveforms. The multiplier–integrator combination constitutes a matched filter, which, in the case of additive white Gaussian noise (AWGN) and no ISI,

[5] The word "carrier" here is used to denote the combination (product) of the true carrier and the symbol waveform (clock).

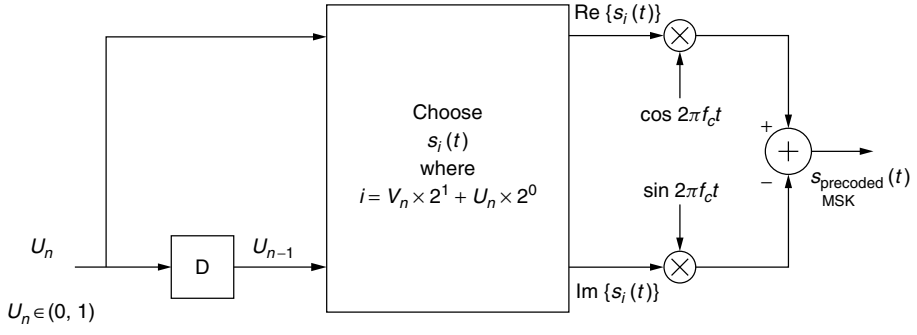


Figure 16. Precoded MSK transmitter based on Rimoldi decomposition of CPM.

results in optimum detection. Means for producing the I and Q demodulation signals $z_c(t)$ and $z_s(t)$ are discussed in the section on synchronization techniques.

9.2. Structures Based on a Memory Transmitter Representation

As noted in Section 7, MSK (or precoded MSK) can be viewed as a cascade of a memory one encoder and a memoryless modulator. As such, a receiver can be implemented on the basis of MLSE detection. For precoded MSK, the appropriate trellis diagram that represents the transitions between states is illustrated in Fig. 17. Each branch of the trellis is labeled with the input bit (0 or 1) that causes a transition and the corresponding waveform (complex) that is transmitted as a result of that transition. The decision metrics based on a two-symbol observation that result in the surviving paths illustrated in Fig. 17 are

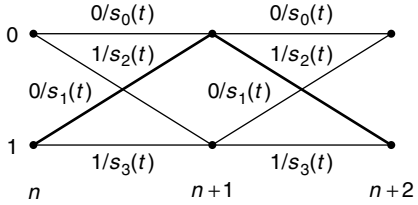
$$\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_0(t)\,dt > \int_{nT_b}^{(n+1)T_b} r(t)s_3(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_1(t)\,dt \quad (48a)$$

$$\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_2(t)\,dt > \int_{nT_b}^{(n+1)T_b} r(t)s_3(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_3(t)\,dt \quad (48b)$$

Noting from Fig. 16 that $s_3(t) = -s_0(t)$ and $s_2(t) = -s_1(t)$, (48a) and (48b) can be rewritten as

$$\int_{nT_b}^{(n+1)T_b} r(t)s_0(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_0(t)\,dt > -\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_1(t)\,dt \quad (49a)$$

$$\int_{nT_b}^{(n+1)T_b} r(t)s_0(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_0(t)\,dt > -\int_{nT_b}^{(n+1)T_b} r(t)s_1(t)\,dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)s_1(t)\,dt \quad (49b)$$



Surviving paths for decoding $\hat{U}_n = 0$ in the interval $(n, n+1)$ assuming state "1" at time n are indicated by heavy lines

Figure 17. Complex baseband trellis.

which are identical and suggest the memoryless receiver illustrated in Fig. 18 [19].[6] Thus we conclude that MSK (or precoded MSK) is a memory one type of trellis-coded modulation (TCM) which can be decoded with a finite (one bit) decoding delay, i.e., the decision on the nth bit can be made at the conclusion of observing the received signal for the $(n+1)$st transmission interval.

Massey [20] suggests an alternative representation of MSK (or precoded MSK) in the form of a single-input two-output sequential transducer followed by an RF selector switch (Fig. 19). For precoded MSK, the sequential transducer implements the ternary sequences $\alpha_k^+ = \frac{1}{2}(\alpha_{k-1} + \alpha_k)$ and $\alpha_k^- = (-1)^k \frac{1}{2}(\alpha_{k-1} - \alpha_k)$ in accordance with Eq. (45). Note as before that α_k^+ is nonzero only when α_k^- is zero and vice versa. The function of the RF selector switch is to select one of the carriers for the signal to be transmitted in each bit interval according to the rule

$$s(t) = \begin{cases} r_2(t) & \text{if } \alpha_k^+ = 1 \\ -r_2(t) & \text{if } \alpha_k^+ = -1 \\ r_1(t) & \text{if } \alpha_k^- = 1 \\ -r_1(t) & \text{if } \alpha_k^- = -1 \end{cases}, \qquad r_i(t) = \sqrt{\frac{2E_b}{T_b}} \cos 2\pi f_i t, \quad i = 1, 2 \quad (50)$$

which represents four mutually exclusive possibilities. This form of modulator has the practical advantage of not requiring addition of RF signals nor RF filtering since there is no actual mixing of the carriers with the modulating signals.

Massey shows that, analogous to Fig. 17, the output of the modulator can be represented by a trellis (Fig. 20) where again each branch is labeled with the input bit

[6] It can be shown that the surviving paths corresponding to being in state "0" at time n leads to the identical decision metric as that in (49a) or (49b).

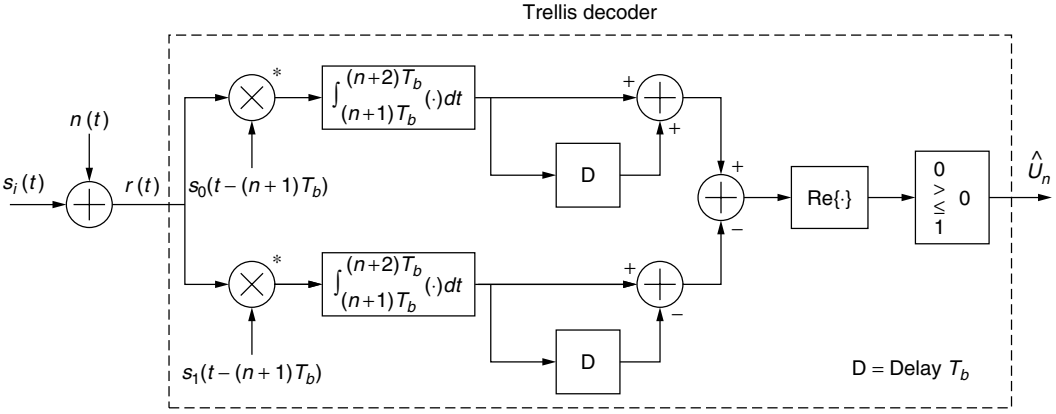


Figure 18. Complex MLSE receiver.

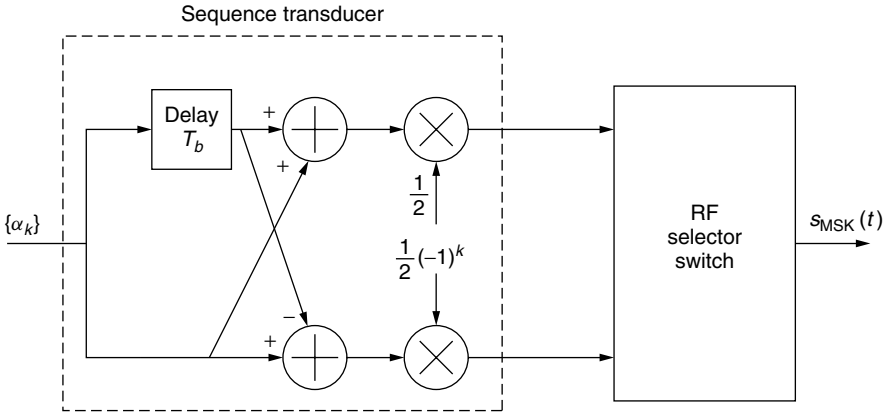


Figure 19. Massey's precoded MSK transmitter.

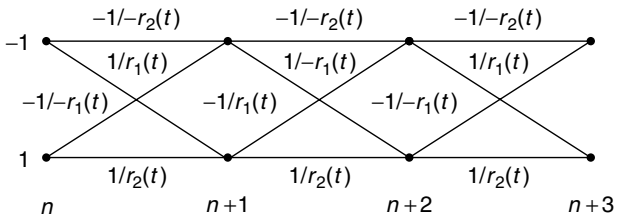


Figure 20. Transmitter output trellis diagram.

and the signal transmitted. Note that the trellis is time-varying (the branch labels alternate with a period of 2). In view of the trellis representation in Fig. 20 the optimum receiver is again an MLSE, which has the same structure as that in Fig. 18, where the complex demodulation signals $s_0(t-(n+1)T_b)$ and $s_1(t-(n+1)T_b)$ are replaced by the real carriers $r_1(t)$ and $r_2(t)$ of (50), the real part of the comparator (difference) output is omitted, and the decision device outputs balanced $+1, -1$ data rather than 0,1 data.

Regardless of the particular receiver implementation employed, the bit error probability (BEP) performance of ideal coherent detection[7] of MSK is given by

$$P_b(E) = \operatorname{erfc}\sqrt{\frac{E_b}{N_0}}\left(1-\frac{1}{2}\operatorname{erfc}\sqrt{\frac{E_b}{N_0}}\right) \tag{51}$$

whereas the equivalent performance of precoded MSK is

$$P_b(E) = \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E_b}{N_0}} \tag{52}$$

which is identical to that of ideal coherent detection of BPSK, QPSK, or OQPSK. Comparing (51) with (52), we observe that the former can be written in terms of the latter as

$$P_b(E)\Big|_{\text{MSK}} = 2P_b(E)\Big|_{\substack{\text{precoded}\\ \text{MSK}}}\left(1-P_b(E)\Big|_{\substack{\text{precoded}\\ \text{MSK}}}\right) \tag{53}$$

which reflects the penalty associated with the differential encoding/decoding operation inherent in MSK but not in precoded MSK as discussed previously. At a BEP of 10^{-5} this amounts to a penalty of approximately a factor of 2 in error probability or equivalently a loss of 0.75 dB in E_b/N_0.

10. DIFFERENTIALLY COHERENT DETECTION

In addition to coherent detection, MSK can be differentially detected [21] as illustrated in Fig. 21. The MSK signal plus noise is multiplied by itself delayed one bit and phase shifted 90°. The resulting product is passed through a lowpass zonal filter that simply removes second harmonics of the carrier frequency terms. Also assumed is that the carrier frequency and data rate are integer related, that is, $f_cT_b = k$ with k integer. Assuming that the MSK signal input to the receiver is in the form of (1) combined with (12):

$$\begin{aligned} s(t) &= \sqrt{\frac{2E_b}{T_b}}\cos\left(2\pi f_c t + \alpha_n\frac{\pi}{2T_b}t + x_n\right)\\ &= \sqrt{\frac{2E_b}{T_b}}\cos\Phi(t,\boldsymbol{\alpha}), \quad nT_b \le t \le (n+1)T_b \end{aligned} \tag{54}$$

then the differential phase $\Delta\Phi \triangleq \Phi(t,\boldsymbol{\alpha}) - \Phi(t-T_b,\boldsymbol{\alpha})$ is given by

$$\Delta\Phi \triangleq -(\alpha_{n-1}-\alpha_n)\frac{\pi}{2}\left(\frac{t}{T_b}-k\right)+\alpha_{n-1}\frac{\pi}{2} \tag{55}$$

where we have made use of the phase continuity relation in (15) in arriving at (55). The mean of the lowpass zonal filter output can be shown to be given by

$$\overline{y(t)} = s(t)s_{90}(t) = \frac{E_b/T_b}{2}\sin\Delta\Phi \tag{56}$$

[7] By "ideal coherent detection" we mean a scenario wherein the local supplied carrier reference is perfectly phase (and frequency) synchronous with the received signal carrier. In Section 10 we explore the practical implications of an imperfect carrier synchronization.

where the "90" subscript denotes a phase shift of 90° in the corresponding signal. Combining (55) and (56), the sampled mean of the lowpass zonal filter output at time $t = (n+1)T_b$ becomes

$$\overline{y((k+1)T_b)} = \frac{E_b/T_b}{2} \sin\left(\alpha_k \frac{\pi}{2}\right) = \alpha_k \frac{E_b/T_b}{2} \qquad (57)$$

which clearly indicates the appropriateness of a hard limiter detector in the presence of noise. Figure 22 illustrates the various waveforms present in the differentially coherent receiver of Fig. 21 for a typical input data sequence.

11. SYNCHRONIZATION TECHNIQUES

In our discussion of coherent reception in Section 6, we implicitly assumed that a means was provided in the receiver for synchronizing the phase of the local demodulation reference(s) with that of the received signal carrier

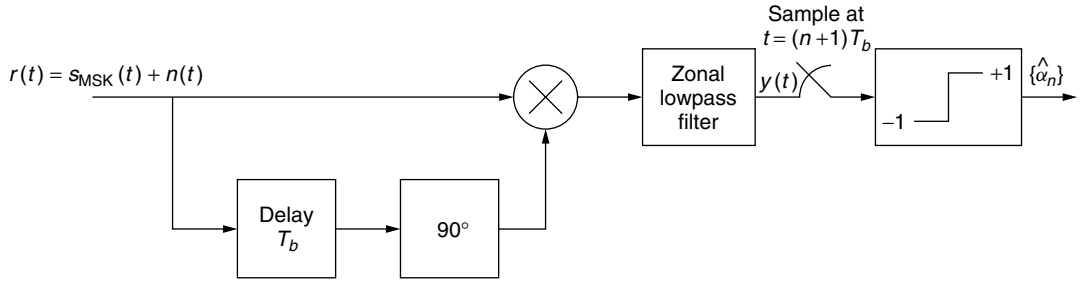


Figure 21. Differentially coherent MSK receiver.

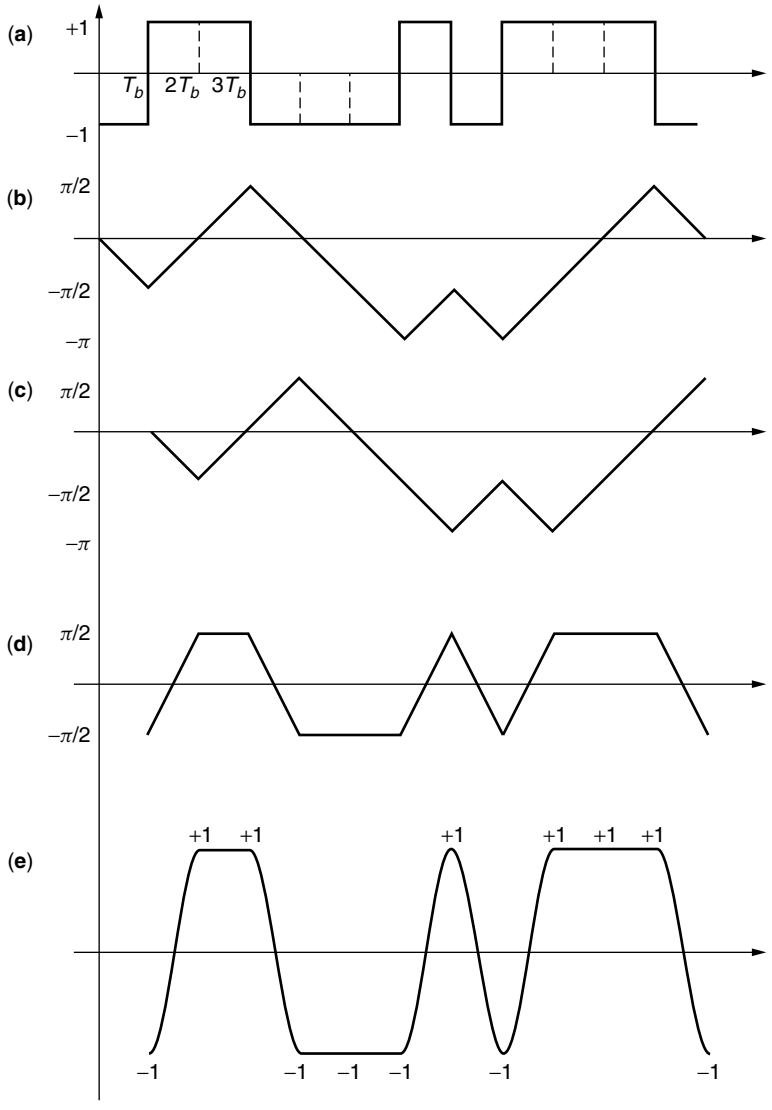


Figure 22. (**a**) Transmitted bit sequence; (**b**) transmitted phase; (**c**) transmitted phase delayed; (**d**) difference phase; (**e**) multiplier output (sine of difference phase).

and also for time synchronizing the I&D circuits. Here we discuss several options for implementing such means.

One form of combined carrier and clock recovery which is synergistic with the transmitter form in Fig. 13 was originally proposed by DeBuda [22,23].[8] With reference to Fig. 23, the received MSK signal is first squared to produce an FSK signal at twice the carrier frequency and with twice the modulation index, i.e., $h = 1$, which is known as *Sunde's FSK* [24]. Whereas the MSK signal has no discrete (line) spectral components, after being squared it has strong spectral components at $2f_1$ and $2f_2$ which can be used for synchronization. In fact, Sunde's FSK has 50% of its total power in these two line components (the other 50% of the total power is in a discrete line component at DC). To demonstrate this transformation from continuous to discrete spectrum, we square the MSK signal form in (30), which gives

$$\begin{aligned} s^2_{\mathrm{MSK}}(t) &= \frac{2E_b}{T_b}[(v_n^+)^2\cos^2 2\pi f_2 t + (v_n^-)^2\cos^2 2\pi f_1 t \\ &\quad + 2v_n^+ v_n^- \cos 2\pi f_2 t \cos 2\pi f_1] \\ &= \frac{2E_b}{T_b}\left[\frac{1}{2} + \frac{1}{2}(v_n^+)^2\cos 4\pi f_2 t + \frac{1}{2}(v_n^-)^2\cos 4\pi f_1 t\right], \\ &\qquad v_n^+ = \frac{v_{n-1}+v_n}{2},\ v_n^- = (-1)^n\left(\frac{v_{n-1}-v_n}{2}\right) \end{aligned} \tag{58}$$

where we have made use of the fact that since either v_n^+ or v_n^- is always equal to zero, then $v_n^+ v_n^- = 0$. Also, either $(v_n^+)^2 = 1$ and $(v_n^-)^2 = 0$ or vice versa, which establishes (58) as a signal with only discrete line components. The components at $2f_1$ and $2f_2$ are extracted by bandpass filters (in practice, phase-locked loops) and then frequency divided to produce $s_1(t) = \frac{1}{2}\cos 2\pi f_1 t$ and $s_2(t) = \frac{1}{2}\cos 2\pi f_2 t$. The sum and difference of these two signals produce the reference "carriers" $z_c(t) = C(t)\cos 2\pi f_c t$ and $z_s(t) = S(t)\sin 2\pi f_c t$, respectively, needed in Fig. 6. Finally, multiplying $s_1(t)$ and $s_2(t)$ and lowpass filtering the result produces $\frac{1}{8}\cos 2\pi t/2T_b$ (a signal at half the bit rate), which provides the desired timing information for the I&Ds in Fig. 6.

Another joint carrier and timing synchronization scheme for MSK was derived by Booth [25] in the form of a closed loop motivated by the maximum a posteriori (MAP) estimation of carrier phase and symbol timing (Fig. 24). The resulting structure (Fig. 24a) is an overlay of two MAP estimation I-Q closed loops — one typical of a carrier synchronization loop assuming known symbol timing (Fig. 24b) and one typical of a symbol timing loop assuming known carrier phase (Fig. 24c). In fact, the carrier synchronization component loop is identical to what would be obtained for sinusoidally pulse-shaped OQPSK.

Finally, many other synchronization structures have been developed for MSK and conventional (single modulation index) binary CPM, which, by definition, would also be suited to MSK. A sampling of these is given in the literature [26–32]. In the interest of brevity, however, we do not discuss these here. Instead the interested reader is referred to the cited references for the details.

12. FURTHER EXTENSIONS

As alluded to earlier, MSK is just one special case of a class of full response CPM modulations with modulation index 0.5, in particular, it is one that possesses a rectangular frequency pulse shape. We have also at times referred to another modulation in this class, namely,

[8] DeBuda also referred to MSK, in conjunction with his self-synchronizing circuit, as "fast FSK (FFSK)" which at the time was the more popular terminology in Canada.

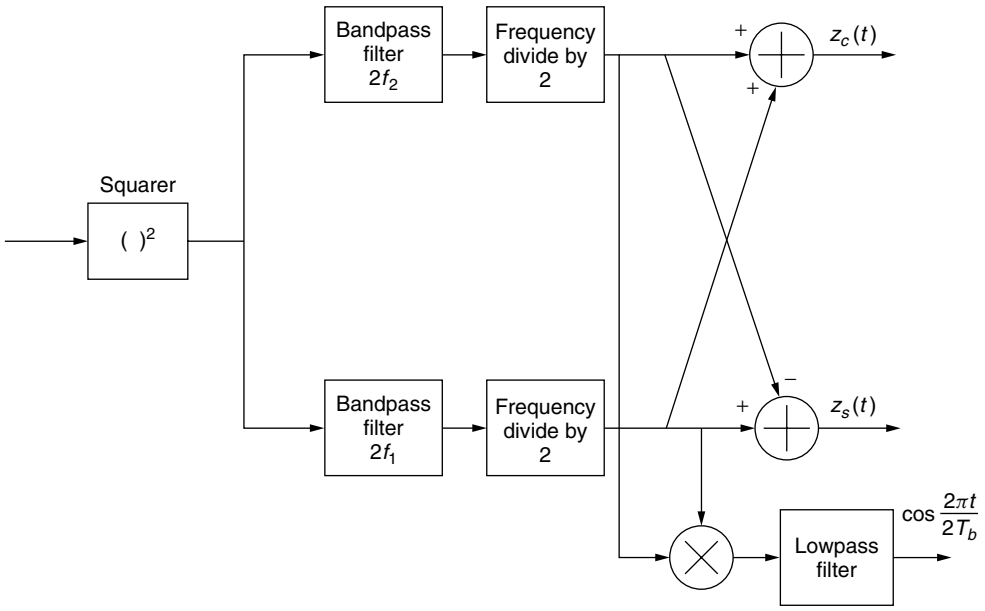


Figure 23. DeBuda's carrier and symbol synchronization scheme.

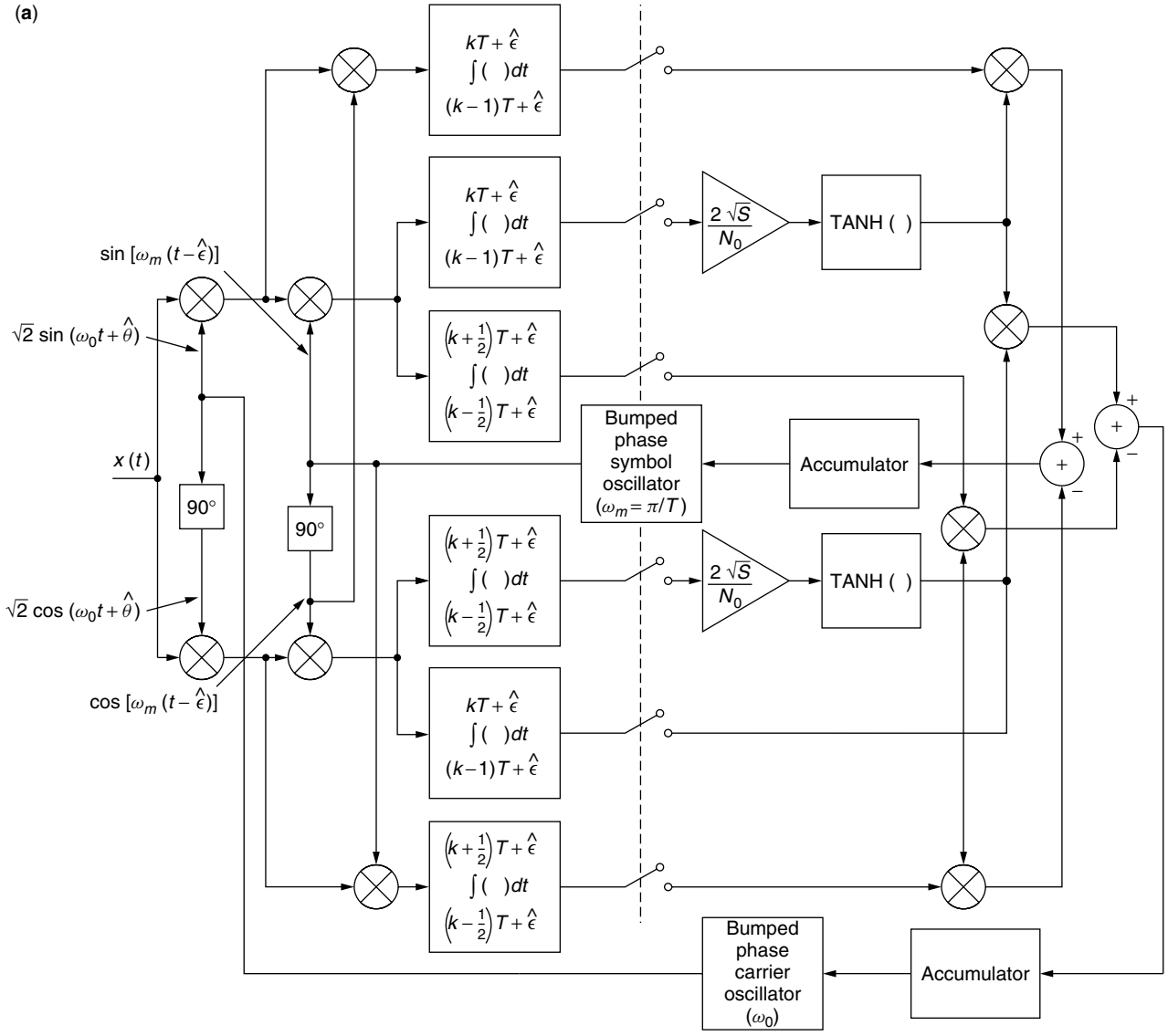


Figure 24. (**a**) Joint carrier and symbol MAP estimation loop for MSK modulation; (**b**) same (carrier synchronization component); (**c**) same (symbol synchronization component).

SFSK, which possesses a raised cosine pulse shape. Since the particular frequency pulse shape selected does not effect the power efficiency (error probability performance) of the system, provided an appropriate matched filter is used in the receiver, then the choice of a pulse shape is made primarily based on spectral efficiency considerations. In this regard, many other modulations in this class have been suggested in the literature. Here we briefly summarize these and provide the appropriate references for readers wishing to explore these in more detail.

Reiffen and White [33] proposed a generalization of MSK called *continuous shift keying* (CSK), in which the instantaneous frequency and perhaps higher derivatives of the phase, as well, are continuous. For CSK the effective inphase channel amplitude pulse shape (of duration $2T_b$) takes the form

$$C(t)=\frac{1}{\sqrt{T_b}}\cos\phi(t), \quad -T_b\le t\le T_b \tag{59}$$

where

$$\phi(t)=\begin{cases}\pm\dfrac{\pi}{2}, & t=-T_b\\ \text{arbitrary}, & -T_b<t<0\\ 0, & t=0\\ \pm\dfrac{\pi}{2}\pm\phi_0(t-T_b), & 0<t<T_b\\ \pm\dfrac{\pi}{2}, & t=T_b\end{cases} \tag{60}$$

It can be shown that the spectral efficiency of this class of schemes is directly related to the smoothness of the derivatives at the endpoints of the $2T_b$-second interval.

Rabzel and Pasupathy [34] proposed a pulse shape characterized by Eq. (59) but with

$$\phi(t)=\frac{\pi t}{2T_b}-\frac{1}{n}\sum_{i=1}^{M}K_i\left[\sin\frac{2\pi nt}{T_b}\right]^{2i-1},$$

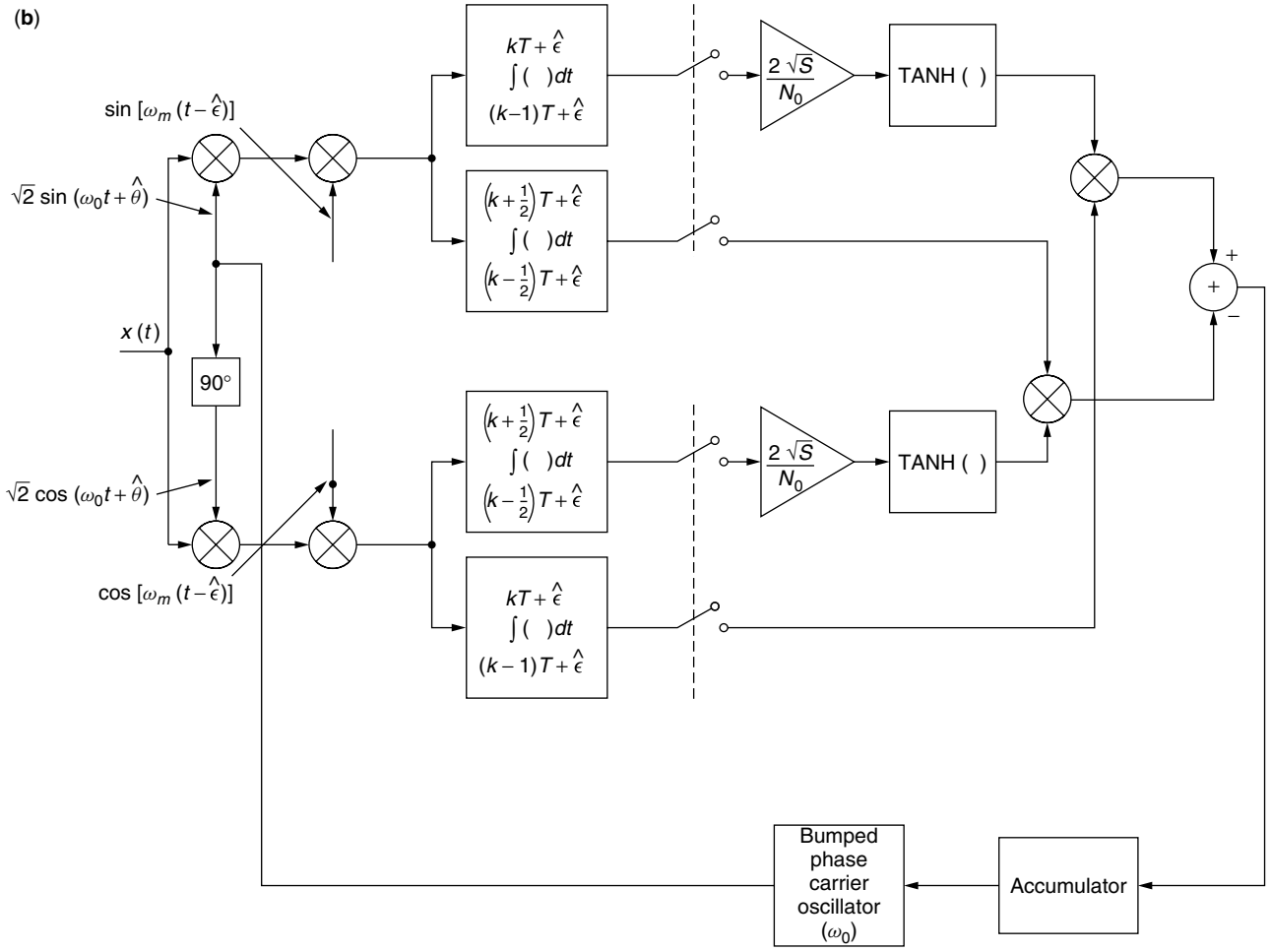


Figure 24. (*Continued*)

$$K_i = \frac{(2i-2)!}{2^{2i-2}[(i-1)!]^2(2i-1)} \tag{61}$$

where $n = 1, 2, 3, \ldots,$ and M is an integer parameter to be selected. The PSD of this class of generalized MSK modulations asymptotically rolls off as $f^{-(4M+4)}$. Special cases are $M = 0$ (MSK) and $M = 1, n = 1$ (SFSK).

Bazin [35] proposed a pulse shape also characterized by (59) but with

$$\phi(t) = \frac{\pi t}{2T_b} - \sum_{k=1}^{N'} A_k \sin \frac{2\pi k t}{T_b}, \quad N' \geq \frac{N}{2} \tag{62}$$

where N is an integer such that all the phase pulse function has all its derivatives up to the Nth order equal to zero at its endpoints (as such the PSD asymptotically rolls of as $f^{-(2N+4)}$) and the A_k coefficients are the solution of the linear system

$$\left.\frac{d^i s(t)}{dt^i}\right|_{t=\pm T_b} = 0, \quad i = 1, 2, \ldots, N \tag{63}$$

It can be shown that Rabzel's pulse format is a subclass of Bazin's. Aside from the well-known special cases of MSK and SFSK, a variation of the latter called DSFSK is the special case corresponding to $N = 4$, $N' = 2$ with coefficients $A_1 = \frac{1}{3}$, $A_2 = -\frac{1}{24}$ which, from Eq. (62), yields the phase pulse

$$\phi(t) = \frac{\pi t}{2T_b} - \frac{1}{3}\sin\frac{2\pi t}{T_b} + \frac{1}{24}\sin\frac{4\pi t}{T_b} \tag{64}$$

In accordance with the above, the PSD for this modulation scheme asymptotically rolls off as f^{-12}. Finally, the connection between CSK and MSK, SFSK, and the generalizations suggested in [34] and [35] was pointed out by Cruz [36].

As an alternative to the choice of a frequency pulse for shaping the transmitted PSD, one can accomplish spectral efficiency by introducing correlation into the transmitted data sequence via a precoder. In Section 3 we considered a simple differential decoder as a precoder in Fig. 8; however, such a precoder has no effect whatsoever on the PSD. The authors of Refs. 37–41 applied correlative encoding (duobinary modulation) [42] to MSK with the intention of obtaining spectral improvement with minimum sacrifice in power performance. The duobinary encoder is illustrated in Fig. 25 and at first glance resembles a differential decoder (see Fig. 8). However, one

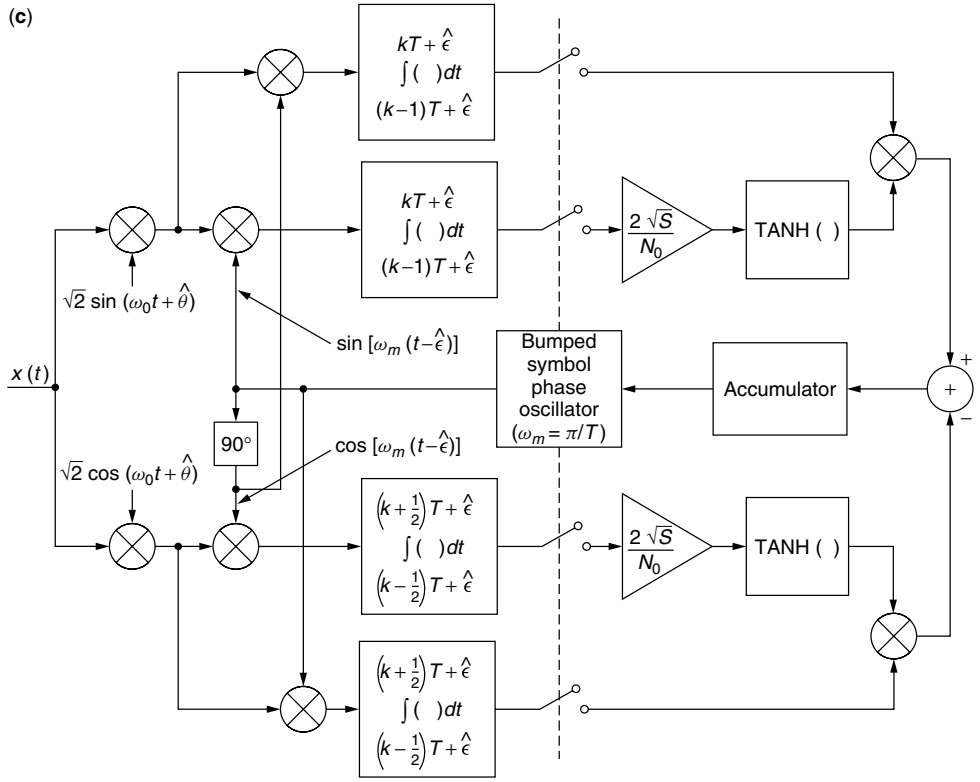


Figure 24. (*Continued*)

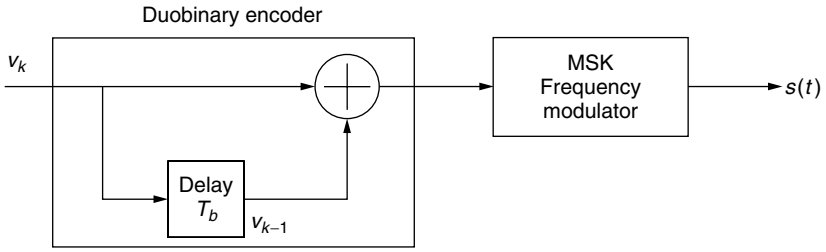


Figure 25. Duobinary encoded MSK.

should note that the multiplier in the latter is replaced by a summer in the former, which results in a ternary $(+1, 0, -1)$ rather than a binary $(+1, -1)$ output.

The last extension of MSK worth mentioning, which again attempts to trade improved bandwidth efficiency for a decrease in power efficiency, is the extension to multiple level pulses, known as *multiple amplitude MSK* (MAMSK) [43]. The simplest way of envisioning this modulation is to consider the I-Q representation of MSK where the input data $\{\alpha_n\}$ now takes on levels $\pm 1, \pm 3, \ldots, \pm M-1$. In this regard, MAMSK can be viewed as a form of offset QAM with sinusoidal pulse shaping [44]. Of course, since the modulation now contains multiple amplitudes, the modulation is no longer constant envelope but rather occupies a discrete number of envelope levels.

Before concluding this chapter we wish to point out to the reader a well-written and simple-to-read tutorial article on MSK by Pasupathy [45] that covers the basics of what is discussed here and provides a valuable set of additional references on the subject as of the time of that publication that include some of the early applications in commercial communication systems.

BIOGRAPHY

Dr. Marvin K. Simon is currently a principal scientist at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, where for the last 34 years he has performed research as applied to the design of NASA's deep-space and near-earth missions resulting in the issuance of nine patents and 23 NASA Tech Briefs. His research interests are modulation and demodulation, synchronization techniques for space, satellite and radio communications, trellis-coded modulation, spread spectrum and multiple access communications, and communication over fading channels. In the past, Dr. Simon also held a joint appointment with the Electrical Engineering Department at Caltech.

He has published over 160 papers and 10 textbooks on the above subjects. His work has also appeared as chapters in several other textbooks. He is the corecipient of

the 1988 Prize Paper Award in Communications of the IEEE Transactions on Vehicular Technology for his work on trellis-coded differential detection systems and also the 1999 Prize Paper of the IEEE Vehicular Technology Conference for his work on switched diversity. He is a fellow of the IEEE and a fellow of the IAE. Among his awards are the NASA Exceptional Service Medal, NASA Exceptional Engineering Achievement Medal, IEEE Edwin H. Armstrong Achievement Award, and most recently the IEEE Millennium Medal all in recognition of outstanding contributions to the field of digital communications and leadership in advancing this discipline.

BIBLIOGRAPHY

1. U.S. Patent 2,977,417 (March 28, 1961), M. L. Doelz and E. T. Heald, Minimum-shift data communication system.
2. U.S. Patent 3,731,233 (May 1, 1973), W. M. Hutchinson, Minimum shift keying modulating apparatus.
3. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
4. M. K. Simon, A generalization of MSK-Type signaling based upon input data symbol pulse shaping, *IEEE Trans. Commun.* **COM-24**(8): 845–856 (Aug. 1976).
5. P. Galko and S. Pasupathy, Generalized MSK, *Proc. IEEE Int. Electrical, Electronics, Conf. Exposition*, Toronto, Ontario, Canada, Oct. 5–7, 1981.
6. I. Korn, Generalized MSK, *IEEE Trans. Inform. Theory* **IT-26**(2): 234–238 (March 1980).
7. F. Amoroso, Pulse and spectrum manipulation in the minimum (frequency) shift keying (MSK) format, *IEEE Trans. Commun.* **COM-24**(3): 381–384 (March 1976).
8. M. G. Pelchat, R. C. Davis, and M. B. Luntz, Coherent demodulation of continuous phase binary FSK signals, *Proc. Int. Telemetry Conf.* Washington, DC, 1971.
9. M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Prentice-Hall, Upper Saddle River, NJ, 1995.
10. F. Amoroso and J. A. Kivett, Simplified MSK signaling technique, *IEEE Trans. Commun.* **25**(4): 433–441 (April 1977).
11. H. R. Mathwich, J. F. Balcewicz, and M. Hecht, The effect of tandem band and amplitude limiting on the E_b/N_0 performance of minimum (frequency) shift keying (MSK), *IEEE Trans. Commun.* **COM-22**(10): 1525–1540 (Oct. 1974).
12. S. A. Gronemeyer and A. L. McBride, MSK and offset QPSK modulation, *IEEE Trans. Commun.* **COM-24**(8): 809–820 (Aug. 1976).
13. R. E. Ziemer, C. R. Ryan, and J. R. Stilwell, Conversion and matched filter approximations for serial minimum-shift keyed modulation, *IEEE Trans. Commun.* **COM-30**(3): 495–509 (March 1982).
14. S. M. Ryu and C. K Un, A simple method for MSK modulation and demodulation, *Proc. IEEE* **73**(11): 1690–1691 (Nov. 1985).
15. W. A. Sullivan, High-capacity microwave system for digital data transmission, *IEEE Trans. Commun.* **COM-20**(P. I): 466–470 (June 1972).
16. D. M. Brady, A constant envelope digital modulation technique for millimeter-wave satellite system, *ICC'74 Conf. Record*, Minneapolis, MN, June 1974, p. 36C-1.
17. D. P. Taylor, A high speed digital modem for experimental work on the communications technology satellite, *Can. Elect. Eng. J.* **2**(1): 21–30 (1977).
18. R. M. Fielding, H. L. Berger, and D. L. Lochhead, Performance characterization of a high data rate MSK and QPSK channel, *ICC'77 Conf. Record*, Chicago, IL, June 1977, pp. 3.2.42–3.2.46.
19. B. E. Rimoldi, A decomposition approach to CPM, *IEEE Trans. Inform. Theory* **IT-34**: 260–270 (May 1988).
20. J. L. Massey, A generalized formulation of minimum shift keying modulation, *ICC'80 Conf. Record*, Seattle, WA, June 1980, pp. 26.5.1–26.5.5.
21. T. Masamura, S. Samejima, Y. Morihiro, and H. Fuketa, Differential detection of MSK with nonredundant error correction, *IEEE Trans. Commun.* **COM-27**(6): 912–918 (June 1979).
22. R. DeBuda, The Fast FSK modulation system, *ICC'71 Conf. Record*, Montreal, Canada, June 1971, pp. 41-25–45-27.
23. R. DeBuda, Coherent demodulation of frequency-shift-keying with low deviation ratio, *IEEE Trans. Commun.* **COM-20**(3): 429–435 (June 1972).
24. W. R. Bennett and S. O. Rice, Spectral density and autocorrelation functions associated with binary frequency shift keying, *Bell Syst. Tech. J.* **42**: 2355–2385 (Sept. 1963).
25. R. W. Booth, An illustration of the MAP estimation method for deriving closed-loop phase tracking topologies: the MSK signal structure, *IEEE Trans. Commun.* **COM-28**(8): 1137–1142 (Aug. 1980).
26. S. J. Simmons and P. J. McLane, Low-complexity carrier tracking decoders for continuous phase modulations, *IEEE Trans. Commun.* **COM-33**(12): 1285–1290 (Dec. 1985).
27. J. Huber and W. Liu, Data-aided synchronization of coherent CPM receivers, *IEEE Trans. Commun.* **40**(1): 178–189 (Jan. 1992).
28. M. Moeneclaey and I. Bruyland, The joint carrier and symbol synchonizability of continuous phase modulated waveforms, *ICC'86 Conf. Record*, Vol. 2, Toronto, Canada, June 1986, pp. 31.5.1–31.5.5.
29. A. N. D'Andrea, U. Mengali, and R. Reggiannini, A digital approach to clock recovery in generalized minimum shift keying, *IEEE Trans. Vehic. Technol.* **39**: 227–234 (Aug. 1990).
30. A. N. D'Andrea, U. Mengali, and M. Morelli, Multiple phase synchronization in continuous phase modulation, in *Digital Signal Processing 3*, Academic Press, New York, 1993, pp. 188–198.
31. U. Lambrette and H. Meyr, Two timing recovery algorithms for MSK, *ICC'94 Conf. Record*, New Orleans, LA, May 1994, pp. 918–992.
32. A. N. D'Andrea, U. Mengali, and M. Morelli, Symbol timing estimation with CPM modulation, *IEEE Trans. Commun.* **44**(10): 1362–1371 (Oct. 1996).
33. B. Reiffen and B. E. White, On low crosstalk data communication and its realization by continuous shift keyed modulation schemes, *IEEE Trans. Commun.* **COM-26**(1): 131–135 (Jan. 1978).

34. M. Rabzel and S. Pasupathy, Special shaping in minimum shift keying (MS)-type signals, *IEEE Trans. Commun.* **COM-26**(1): 189–195 (Jan. 1978).

35. B. Bazin, A class of MSK baseband pulse formats with sharp spectral roll-off, *IEEE Trans. Commun.* **COM-27**(5): 826–829 (May 1979).

36. J. R. Cruz, A note on spectral shaping of minimum-shift-keying-type signals, *Proc. IEEE* **68**(8): 1035–1036 (Aug. 1980).

37. G. J. Garrison, A power spectral density analysis for digital FM, *IEEE Trans. Commun.* **COM-23**(11): 1228–1243 (Nov. 1975).

38. F. De Jager and C. B. Dekker, Tamed frequency modulation: A novel method to achieve spectrum economy in digital transmission, *IEEE Trans. Commun.* **COM-26**(50): 534–542 (May 1978).

39. S. Gupta and S. Elnoubi, Error rate performance of coded MSK with discriminator detection in land mobile communication system, *Proc. Int. Communications and Computer Exposition*, Los Angeles, CA, Nov. 1980, pp. 120–124.

40. S. Gupta and S. Elnoubi, Error rate performance of duobinary coded MSK and TFM with differential detection in land mobile communication systems, *31st Vehicular Technology Conf. Record*, Washington, DC, April 1981.

41. S. Gupta and S. Elnoubi, Error rate performance of noncoherent detection of duobinary coded MSK and TFM in mobile radio communication systems (with S. Elnoubi), *IEEE Trans. Vehic. Technol.* **VT-30**: 62–76 (May 1981).

42. S. Pasupathy, Correlative coding: a bandwidth efficient signaling scheme, *IEEE Commun. Mag.* **17**(4): 4–11 (July 1977).

43. W. J. Weber, P. H. Stanton, and J. T. Sumida, A bandwidth compressive modulation system using multiamplitude minimum shift-keying (MAMSK), *IEEE Trans. Commun.* **COM-26**(5): 543–551 (May 1978).

44. M. K. Simon, An MSK approach to offset QASK, *IEEE Trans. Commun.* **COM-24**(8): 921–923 (Aug. 1976).

45. S. Pasupathy, Minimum shift keying: a spectrally efficient modulation, *IEEE Commun. Mag.* **17**(4): 14–22 (July 1979).

MOBILE RADIO COMMUNICATIONS

RODGER E. ZIEMER
University of Colorado
Colorado Springs, Colorado

WILLIAM H. TRANTER
R. MICHAEL BUEHRER
Virginia Tech
Blacksburg, Virginia

THEODORE S. RAPPAPORT
The University of Texas at Austin
Austin, Texas

1. THE EARLY HISTORY OF WIRELESS COMMUNICATIONS

Guglielmo Marconi's development and commercialization of wireless telegraphy in the 1890s marked the beginning of wireless communications. While nineteenth century researchers such as Volta, Hertz, and Tesla experimented with the electrostatic and inductive components of electromagnetic fields, Marconi accidentally discovered that a radiation field could be launched from an appropriately designed antenna, thereby allowing reliable propagation over great distances. In April 1901, Marconi successfully demonstrated wireless transmission across the Atlantic Ocean. The work of Marconi spawned a century of research and commercial activity that produced the AM radio, FM radio, television, land mobile radio, cellular radio, wireless data networks, and satellite communications.

Even during the early days of wireless communications, mobile communications was of interest. In 1902, Ernest Rutherford and his assistant, Howard Barnes, developed a wireless telegraphy system to communicate with moving trains. One of the pioneering practical uses of mobile wireless telegraphy was ship-to-ship and ship-to-shore communications in the early decades of the twentieth century. Wireless communications played an important role in rescue operations when the Titanic sank in 1912. Although many recognized the commercial potential of wireless communications, the dawn of the twentieth century witnessed a number of skeptics. For example, J. J. Thompson, who was to receive the Nobel Prize in physics in 1906, remarked that wireless communications was not likely to ever be of real commercial use [1]. In addition, Ernest Rutherford, who would receive the Nobel Prize in chemistry in 1908, remarked to his class at McGill University (Montreal, Canada) in 1898 that "... it is not safe or politic to invest much capital in a company for the transmission of signals by wireless [1]." The chief concerns of these early skeptics were limited range, reliability, and privacy. In reality, these concerns simply provided interesting challenges for future innovators.

Appleton and others throughout the first half of the twentieth century found through experimentation that when the carrier frequency of an electromagnetic wave was selected to match a particular channel, such as the ionosphere or the troposphere, surprisingly reliable worldwide communication was possible, depending on the particular time of day, season of the year, and sunspot activity [2]. Medium wave (100 kHz–3 MHz) and short wave (3 MHz–30 MHz) radiobands became the mainstay for the fledgling wireless broadcasting industry, as well as for ship-to-shore, telegraph, and military operations during the first several decades of the twentieth century, all relying on long-distance "skip" communications.

Amplitude modulation (AM) broadcasting, using medium wave frequencies in the 500 kHz–1700 kHz range, was launched throughout the world in the 1920s. Station KDKA, located in Pittsburgh, Pennsylvania, and owned by Westinghouse, was one of the first commercial AM broadcasting stations to go on the air in 1920. The initial broadcast provided listeners with the election results in which Warren G. Harding won the presidency over James Cox [3]. Inexpensive crystal radio sets were popular at the time and consisted of a small piece of germanium crystal that could be used with a conventional earpiece for local AM reception. An early (1922) factory

purchased crystals for 96 cents and from these developed radios that were sold for $2.25 [3]. To facilitate widespread adoption of AM reception by consumers, extremely low-cost receivers using envelope detectors were developed. Envelope detectors could be implemented easily with a simple diode and RC filter. The goal was to develop low-cost receivers in order to increase public access to this new form of communications.

Single-side band (SSB), which is a special form of AM, provides a spectrally efficient way of transmitting an AM waveform that also provides some security, because it cannot easily be detected by a standard AM envelope detector. In the early years of wireless communications, military personnel relied on SSB for wireless communications for both fixed and mobile applications throughout the world. Today, amateur radio operators and military operations (Military Amateur Radio Service — MARS) still use SSB because of its spectral efficiency.

In the 1920s and 1930s, Edwin Armstrong pioneered two fundamental inventions that still shape the wireless communications field [4]. As an engineer for the U.S. military, Armstrong invented the superheterodyne receiver. Using the concept of mixing, the superhetrodyne receiver (also called the superhet) allows a very high-frequency carrier wave to be translated down to a baseband frequency for detection. Until Armstrong's superhet design, receivers used direct conversion, where the receiver input signal was filtered directly from the antenna with a high-Q tunable bandpass filter, and then detected immediately. With the superhet receiver, it became possible to build much more sensitive receivers that could perform over a much wider frequency range, because it was no longer necessary to provide such a tight (and expensive) tunable filter at the incoming receiver frequency. Instead, a wider bandwidth fixed filter could be used at the antenna input, and a local oscillator could be tuned, thereby providing a mixed signal that could subsequently be filtered with better and less expensive filtering at a much lower intermediate (IF) frequency. The superheterodyne receiver allowed the received signal to be brought down to the baseband detector in stages. By standardizing on IF frequencies, component manufacturers were able develop devices such as filters, oscillators, and mixers that could be used over a wide range of wireless frequencies, thus allowing the wireless communications industry to begin its dramatic growth.

Armstrong's second invention, frequency modulation (FM), was patented in 1934 [3]. FM provided much greater fidelity than AM, as it was impervious to ignition and atmospheric noises that plagued AM transmissions. Since FM used constant envelope modulation, it also was much more power-efficient than AM, making it particularly well-suited for mobile radio telephone operation where battery preservation was key. The high-fidelity qualities of FM launched a new FM broadcasting industry, and frequency allocations for worldwide FM broadcasting were granted by the World Administrative Radio Conference (WARC) in the very high frequency (VHF) bands of 30 MHz–300 MHz, where wireless signals propagated reliably from a broadcast antenna to the visible horizon on earth. Unlike MF and HF waves, the shorter wavelengths of the VHF band are not generally propagated by skip mechanisms, thus providing much more predictable line-of-sight radio propagation behavior for both terrestrial and satellite use.[1] It is for this reason that virtually all modern wireless communication systems operate at or above the VHF frequency band.

Commercial television (TV) broadcasting evolved in the late 1940s and employed a type of AM modulation (vestigial sideband — VSB) for video transmission and FM for simultaneous aural transmission. Like FM broadcasting, TV broadcasting relied on the reliable "to-the-horizon" propagation offered by VHF radio waves as well as ultra-high-frequency (UHF) waves in the 300 MHz to 3 GHz bands.

Very early mobile telephone services also used FM for voice communications in the VHF and UHF bands, as mobile communications equipment became viable from a cost and reliability standpoint in the 1950s. Early taxicab and public service dispatch radio services soon realized that in order to handle a large population of mobile radio users with a finite set of channels, it would be necessary to employ trunking theory. Trunking theory determines how to allocate a finite set of channels to a large population of potential users, based on the calling patterns of an average user [6]. Early mobile radio systems used the concept of a control channel that is intermittently shared by all users of a radio service. The control channel is used by a central switch to broker instantaneous access to all of the available voice channels within the system. Trunking theory is used by a radio service to determine the appropriate number of channels to allocate for a large population of users, so that a particular average grade of service (or channel availability likelihood) can be provided to the users.

2. MOBILE CELLULAR TELEPHONY

The next major event in wireless communications was the development of cellular telephony. The development of cellular communications made mobile radio communications available to the general public at low cost. Cellular radio communications systems were developed in the United States by Bell Laboratories, Motorola, and other companies in the 1970s, and in Europe and Japan at about the same time. Test systems were installed in the United States in Washington, D.C., and Chicago in the late 1970s, and the first commercial cellular systems became operational in Japan in 1979, in Europe in 1981, and in the United States in 1983. Like other early cellular telephone systems, the first system in the United States used analog FM and operated in the UHF band. Designated AMPS (for advanced mobile phone system), it proved so successful that AMPS cellular telephones are still widely used today, especially in rural areas. The AMPS system is based on a channel spacing of 30 kHz.

[1] However, rare instances of skip from the ionosphere have been found to allow propagation distances of several thousand kilometers at frequencies well above 30 MHz [5].

In the early 1990s, the demand for cellular telephones exceeded available capacity, resulting in the development and adoption of so-called second-generation (2G) personal communications systems (PCS), with the first of these systems being fielded in the mid-1990s. All 2G PCS systems use the cellular concept, but employ digital transmission in place of analog FM. They have differing modulation and multiple access schemes as defined by their respective common air-interface standards. The European 2G standard, called global system for mobile communications (GSM), the Japanese system, and one U.S. standard [U.S. digital cellular (USDC) system] all employ time-division multiple access (TDMA), but with differing channel bandwidths and numbers of users per frame. A second U.S. 2G standard uses code division multiple access (CDMA). A goal of 2G system development in the United States was backward compatibility because of the large AMPS infrastructure that had been installed with the first generation. Europe, however, had several first-generation standards, depending on the country, and their goal with 2G was to have a common standard across all countries. As a result, GSM has been widely adopted, not only in Europe, but in much of the rest of the world. From the mid- to late 1990s, work began on third-generation (3G) standards, and these systems are beginning to be deployed after several widely publicized delays. A goal in the development of 3G systems is to have a common worldwide standard, but this proved to be too optimistic. Therefore, a family of standards was adopted, with one objective being to make migration from first-generation and second-generation systems possible.

Cellular systems were much more widely accepted by the public than was first expected when the first-generation systems were introduced. In many European and Pacific Rim countries, more than 50% of the population owns a cellular telephone, with the United States close behind. One might wonder why the United States is not leading the world in terms of cellular telephone usage. Perhaps there are three main reasons. The U.S. licensing process was not formalized until several years after the deployment of first-generation cellular systems in Europe and Japan. Also, the United States enjoyed the preexistence of a very good wireline system before cellular telephones were introduced. Finally, the United States had the practice of charging the cellular subscriber for both incoming and outgoing calls. It appears that the billing mechanism is changing with the further expansion of cellular telephone use in the United States, with cellular service providers not only trying to attract new customers but also trying to reduce the "churn," or changing of the customer from one provider to another.

2.1. Basic Principles of Cellular Radio[2]

Radio telephone systems had been in use long before the introduction of cellular radio, but their capacity was very limited because they were designed around the concept of a single base station servicing a large area — often the size of a large metropolitan area. Cellular telephone systems are based on the concept of dividing the geographic service area into a number of cells and servicing the area with low-power base stations placed within each cell, usually the geographic center. This allows the band of frequencies allocated for cellular radio use (currently there are two bands in the 900 and 1800 MHz regions of the radio spectrum) to be reused in physically separated geographic regions, depending on the accessing scheme involved. For example, with AMPS, the same radio channels are reused over a relatively small geographic area and are repeated once every seven cells. In today's CDMA (code division multiple access), the same channels are used in each cell. Another characteristic that the successful implementation of cellular radio depends on is the attenuation of transmitted power with frequency. For free space, power density decreases per the inverse square of the distance from the transmitter. However, because of the characteristics of terrestrial radio propagation, the decrease of power with distance is greater than an inverse square law, typically between the inverse third and fourth power of the distance. Were this not the case, the cellular concept would not work. Since the service area (the geographic area of interest to be covered with cellular service) is represented by tessallating cells overlaid on a map of the coverage region, it is necessary for the mobile user to be transferred from one base station to another as the mobile moves within the service area. This procedure is called *handoff* or *handover*. Also note that it is necessary to have some way of initializing a call to a given mobile and keeping track of it as it moves from one base station to another. This is the function of a *mobile switching center* (MSC). MSCs also interface with the public switched telephone network (PSTN).

Consider Fig. 1, which shows a typical cellular tessellation using hexagons to represent cells. It is emphasized that real cells are never hexagonal; indeed, some cells may have very irregular shapes because of

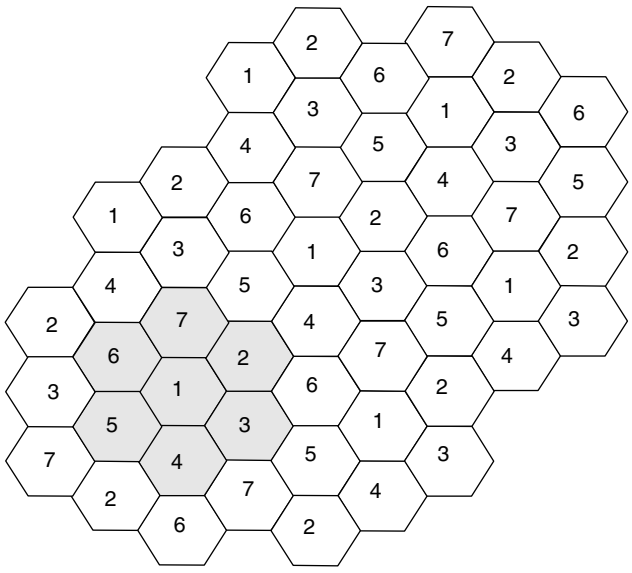


Figure 1. Hexagonal grid system representing cells in a cellular radio system; a reuse pattern of seven is illustrated.

[2] This section follows closely a previous publication by Ziemer and Tranter [7].

geographic features and illumination patterns of the transmit antenna. However, hexagons are typically used in theoretical discussions of cellular radio because a hexagon is one geometric shape that tessellates a plane and very closely approximates a circle, which is what we assume for the contours of equal transmit power in a relatively flat environment. Note that a seven-cell reuse pattern is indicated in Fig. 1 via the integers given in each cell. Obviously, there are only certain integers that work for reuse patterns, for example, 1, 3, 4, 7, 9, 12, ... A convenient way to describe the frequency reuse pattern of an ideal hexagonal tessellation is to use a nonorthogonal set of axes, U and V, intersecting at 60 degrees as shown in Fig. 2. The normalized grid spacing of one unit represents the distance between adjacent base stations, or hexagon centers. Thus, each hexagon center is at point (u, v) where u and v are integers. Using this normalized scale, each hexagon vertex is $R = 1/\sqrt{3}$ from the hexagon center. It can be shown that the number of cells in an allowed frequency reuse pattern is given by

$$N = i^2 + ij + j^2 \tag{1}$$

where i and j take on integer values. Letting $i = 1$ and $j = 2$ (or vice versa), it is seen that $N = 7$, as we already know from the pattern identified in Fig. 2. Considering other integers, the number of cells in various reuse patterns are as given in Table 1. Currently used reuse patterns are 1 (CDMA), 3 (GSM), and 7 (AMPS).

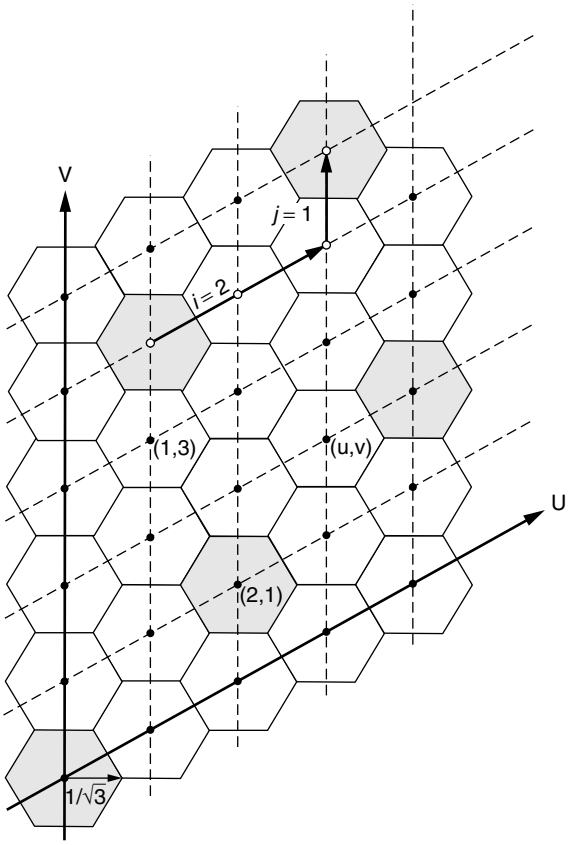


Figure 2. Hexagonal grid geometry showing coordinate directions; a reuse pattern of seven is illustrated.

Table 1. Possible Reuse Patterns in Cellular Systems

Reuse Coordinates		Number of cells in Reuse Pattern	Normalized Distance Between Repeat Cells
i	j	N	$\sqrt{N}$
1	**0**	1	1
2	**1**	3	1.732
1	**2**	7	2.646
2	**2**	12	3.464
1	**3**	13	3.606
2	**3**	19	4.359
1	**4**	21	4.583
2	**4**	28	5.292
1	**5**	31	5.568

Another useful relationship is the distance between like-cell centers, D_{co}, which can be shown to be $D_{co} = \sqrt{3N}R$, which is $D_{co} = \sqrt{N}$ since $R = 1/\sqrt{3}$. This is an important consideration in computing *cochannel interference*, that is, the interference from a second user in a nearby cell that is using the same frequency assignment as that of a user of interest. Clearly, if a reuse pattern has N cells in it, this interference could be a factor of N larger than that due to a single interfering user (not all cells at distance $\sqrt{N}$ from a user of interest may have an active call on that particular frequency). Note that there is a second ring of cells at $2\sqrt{N}$ that can interfere with a user of interest, but these usually are considered to be negligible compared with those within the first ring of interfering cells.

Assume a decrease in power with distance, R, of the form

$$P_r(R) = K\left(\frac{R_0}{R}\right)^{\alpha} \text{ watts} \tag{2}$$

where R_0 is a reference distance and the power is known to be K watts. As mentioned previously, the power law is typically in the range of 2.5 to 4 for terrestrial propagation, which can be analytically shown to be a direct consequence of the earth's surface acting as a partially conducting reflector (other factors such as scattering from buildings and other large objects also come into play, which accounts for the variation in α). In logarithmic terms, the received power is

$$P_{r,\text{dBW}}(R) = K_{\text{dB}} + 10\alpha \log_{10} R_0 - 10\alpha \log_{10} R \text{ dBW} \tag{3}$$

Now consider reception by a mobile from a base station of interest, A, at distance d_A, while at the same time being interfered with from a cochannel base station B, at distance D_{co} from A. We assume for simplicity that the mobile is on a line connecting A and B. Thus, the signal-to-interference ratio (SIR) in decibels is

$$\text{SIR}_{\text{dB}} = K_{\text{dB}} + 10\alpha \log_{10} R_0 - 10\alpha \log_{10} d_A - [K_{\text{dB}} + 10\alpha \log_{10} R_0 - 10\alpha \log_{10}(D_{co} - d_A)] \tag{4}$$

This gives

$$\mathrm{SIR_{dB}} = 10\alpha \log_{10}\left(\frac{D_{co}}{d_A} - 1\right) \mathrm{dB} \tag{5}$$

Clearly, as $d_A \to D_{co}/2$, the argument of the logarithm approaches 1 and the SIR_{dB} approaches 0. As $d_A \to D_{co}/2$, the mobile should ideally hand off from A and begin using B as its base station.

We can also compute a worst-case SIR for a mobile of interest. If the mobile is using base station A as its source, the interference from the six other cochannel base stations in the reuse pattern is no worse than that from B (the mobile is assumed to be on a line connecting A and B). Thus, $\mathrm{SIR_{dB}}$ is underbounded by

$$\begin{aligned}\mathrm{SIR_{dB,min}} &= 10\alpha \log_{10}\left(\frac{D_{co}}{d_A} - 1\right) - 10\log_{10}(6)\ \mathrm{dB} \\ &= 10\alpha \log_{10}\left(\frac{D_{co}}{d_A} - 1\right) - 7.7815\ \mathrm{dB}\end{aligned} \tag{6}$$

2.2. The Mobile Wireless Channel

A distinguishing factor in terrestrial mobile radio is the channel impairments experienced by the signal. In addition to the Gaussian noise present in every communication link due to the nonzero temperature of the receiver, and the cochannel interference, another important source of degradation is the mobile wireless channel. As the mobile moves, the signal strength varies drastically because of multiple transmission paths as well as objects that block the line-of-sight propagation path. The mobile wireless channel induces the attenuation and distortion seen at the receiver due to the environment and mobility. The attenuation and distortion caused by the mobile wireless channel can be broken into two main components: large-scale fading and small-scale fading.

2.2.1. Large-Scale Fading. Large-scale fading is related to the path loss discussed earlier. As mentioned previously, the terrestrial wireless channel (because it typically involves scattering and possibly non–line-of-sight conditions) experiences a loss in received power that increases with distance raised to the third to fourth power. This typically is called the path loss exponent. Additionally, for a given distance there is some statistical variation about the distance-dependent mean value represented by the path loss exponent. This variation normally is attributed to large objects in the environment and is termed *shadowing*. For a fixed propagation distance R, some radial paths from a transmitter experience more shadowing than others as a result of the spatial variations of objects over a particular geometry. The large-scale variation about a distance-dependent mean signal level typically follows a log-normal distribution with a standard deviation of 8–12 dB [6].

2.2.2. Small-Scale Fading. Small-scale fading refers to variations in the signal strength seen by a mobile receiver as it moves over very short (on the order of wavelengths) distances. This type of fading can be characterized in terms of a Doppler spectrum, which is determined by the motion of the mobile (and to a small degree the motion of the surroundings, such as a wind blowing trees or the motion of reflecting vehicles). Another characteristic of small-scale fading is time-varying delay spread due to the differing propagation distances of the received multipath components. As signaling rates increase, this becomes a more serious source of degradation due to intersymbol interference (ISI) of the transmitted signal. Equalization can be used to compensate for ISI.

Small-scale multipath can be divided into two types: *unresolvable*, where the resultant at the receiver can be approximated as a sum of phasors whose amplitudes and phases vary with motion of the transmitter, receiver, or environment; and *resolvable*, where propagation times are long compared with the inverse signal (receiver) bandwidth. The latter condition means that the channel is frequency selective (the signal bandwidth is wide with respect to variations in channel frequency response). It should be clear that multipath resolvability depends on receiver bandwidth relative to the time intervals between multipath components.

A common technique used to combat small-scale fading is diversity. In GSM and USDC, diversity takes the form of error-correction coding using interleaving. For CDMA, diversity can be added in the form of simultaneous reception from two different base stations near cell boundaries (soft handoff). Other combinations of simultaneous transmissions [9] and receptions in a rich multipath environment are being proposed for future-generation systems to significantly increase capacity (this is also called transmit diversity). Also used in CDMA is a tool called a RAKE receiver, which detects the separate resolvable multipath components and puts them back together in a constructive fashion. In general, the various diversity techniques can be divided into three general categories:

1. Space diversity — the use of more than one antenna to capture the propagating signal.
2. Frequency diversity — the use of several carrier frequencies or the use of a wideband signal with an equalizer or RAKE receiver
3. Time diversity — spreading out the effects of errors through interleaving and coding

All three of these techniques are commonly used.

2.3. Multiple Access Techniques

The implementation of a cellular radio system depends heavily on the use of a multiple access scheme. Multiple access is the technique that allows multiple users to access the system simultaneously. We have already touched on the idea of multiple access, but in this section we describe it in more detail. Historically, three common methods for multiple access are

1. Frequency division multiple access (FDMA)
2. Time division multiple access (TDMA)
3. Code division multiple access (CDMA)

Figure 3 schematically illustrates these three access schemes. Three dimensions are shown in each figure — time, frequency, and code. In Fig. 3 (a), this time-frequency-code resource is split into a number of frequency channels, each one of which may be assigned to a different user. In other words, we give potential users access to the communication resource using FDMA. In Fig. 3 (b), the time-frequency-code resource is split into a number of time slots, each of which may be assigned to a different user. In other words, we give potential users access to the communication resource using TDMA. Finally, in Fig. 3 (c), the time-frequency-code resource consists of a number of codes, each of which may be assigned to a different user. In other words, we give potential users access to the communication resource using CDMA.

In cellular radio systems, two of these typically are used together. For example, in the global system for mobile (GSM) communications system, TDMA and FDMA are used together in that the allocated frequency spectrum is divided into 200-kHz chunks for each TDMA frame, and each frame can then accommodate up to eight users using TDMA. As another example, the IS-95 standard was designed around the use of CDMA to accommodate up to 61 users (There are 64 potential channels. Three channels are set aside for synchronization, the pilot, and for paging. See Ref. 8). A Walsh code is assigned to each of the 64 channels, and each block of 61 users requires only 1.25 MHz of spectrum. Thus, the allocated frequency spectrum is divided into multiple 1.25-MHz chunks of frequency, each of which can be employed for up to 61 users that are distinguished from each other by their assigned CDMA codes.

3. CHARACTERISTICS OF SECOND-GENERATION CELLULAR SYSTEMS

As mentioned previously, the development of cellular telephony is commonly broken up into three distinct stages termed generations. First-generation cellular systems (including the AMPS standard in the United States) were analog systems that carried only voice traffic. As the capacity of first-generation systems filled, more efficient cellular systems were developed. These so-called second-generation systems used digital modulation and allowed roughly a three times capacity improvement over the analog systems. Also, because they were digital, these second-generation systems were capable of carrying rudimentary data services.

Space does not allow much more than a cursory glance at the technical characteristics of the most popular second-generation cellular radio systems — in particular, USDC, GSM, and CDMA (referred to as IS-95 in the past, where the "IS" stands for "interim standard"). For complete details, the standard for each may be consulted. Before doing so, however, the reader is warned that this amounts to thousands of pages in each case. Table 2 summarizes

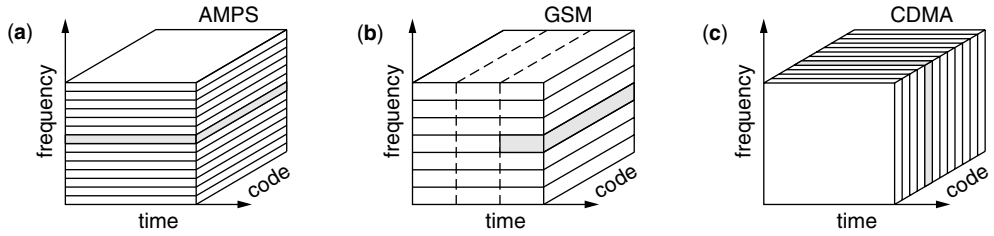


Figure 3. Illustrations of multiple access schemes: (**a**) FDMA; (**b**) TDMA; (**c**) CDMA.

Table 2. 3G AMPS, GSM, and CDMA Technologies Compared

	AMPS	GSM	CDMA
Carrier separation	30 kHz	200 kHz	1.25 MHz
No. channels/carrier	1	8	61
Accessing techniques	FDMA	TDMA-FDMA	CDMA-FDMA
Frame duration	NA	4.6 ms with 0.58 ms slots	20 ms
User modulation	FM	GMSK, $BT = 0.3$ Binary, diff. encoded	BPSK, FL 64-ary orthog. RL
Cell reuse pattern	7	3	1
Cochan. Inter. Protect.	$\leq$15 dB	$\leq$12 dB	NA
Error correction-coding	NA	Rate 1/2 convolutional Constraint length 5	Rate 1/2 convol., FL Rate 1/3 convol., RL Both constr. length 9
Diversity methods	NA	Freq. hop, 216.7 hops/s Equalization	Wideband signal Interleaving RAKE
Speech representation	Analog	Residual pulse excited, Linear prediction coder	Code-excited vocoder
Speech coder rate	NA	13 kbps	9.6 kbps max

some of the most pertinent features of these three systems. For further details, see Rappaport [6] and Ziemer and Peterson [8].

4. THIRD-GENERATION CELLULAR RADIO

Second-generation systems provides one of the most successful practical applications of many aspects of communication theory, including speech coding, modulation, channel coding, diversity techniques, equalization, and so on. The implementation of 3G cellular promises the same accommodation as 2G for voice in addition to much higher data rate capacity, including 64 kbps for high-speed vehicles over wide areas, 384 kbps for pedestrian speeds over smaller areas, and 2 Mbps for stationary (but movable) locations over building- or campus-size areas. With these capabilities, 3G cellular indeed promises ready access to information anytime and anywhere. Specifically, 3G will include the following options:

1. Flexible support of multiple services (data rates from kbps to Mbps; packet transmission)
2. Voice
3. Messaging — email, fax, etc.
4. Medium-rate multimedia — Internet access, educational
5. High-rate multimedia — file transfer, video
6. High-rate interactive multimedia — video teleconferencing, telemedicine, etc.
7. Mobility: quasi-stationary to high-speed platforms
8. Global roaming: ubiquitous, seamless coverage
9. Evolution from second-generation systems

Two CDMA-based radio transmission technologies (RTTs) have emerged as the leading answers for 3G radio. One, called WCDMA for wideband CDMA, originated out of several European and Japanese studies carried out in the mid-1990s. The other, called CDMA2000, originated from a coalition of U.S. companies working out a standard in the late 1990s that would provide an easy migration path for the CDMA 2G standard, IS-95. The characteristics of CDMA and CDMA2000 are summarized in Table 3.

Table 3. 3G Radio Transmission Technologies Compared

PARAMETER	WCDMA	CDMA2000
Carrier spacing	5 MHz	3.75 MHz
Chip rate	3.84 Mcps	3.684 Mcps
Data modulation	BPSK	DL — QPSK; UL — BPSK
Spreading	OQPSK	OQPSK
Power control frequency	1500 Hz	800 Hz
Variable data rate implementation	Variable SF; multicode	Repetition, puncturing, multicode
Frame duration	10 ms	20 ms
Coding	Turbo and convolution	Turbo and convolution
Base station synchronized?	Asynchronous	Synchronous
Base station acquisition/detect	3 step: slot, frame, code	Time shifted PN correl.
Forward link pilot	TDM dedicated pilot	CDM common pilot
Antenna beam forming	TDM dedicated pilot	Auxiliary pilot

Table 4. Rate Examples for the Uplink of WCDMA

Number of Data Channels	Bits per Slot	Spreading Factor	Channel Symbol Rate (kbps)	Data Rate (kbps)
1	10	256	15	7.5
1	20	128	30	15
1	40	64	60	30
1	80	32	120	60
1	160	16	240	120
1	320	8	480	240
1	640	4	960	480
6	640	4	5740	2370

5. CONCLUSIONS

Despite the fact that wireless communications is over 100 years old, it remains an active area of research, development, and commercialization. Cellular systems continue to grow in popularity, with penetration rates increasing all over the world. Wireless networks are becoming more popular in business and campus environments. Wireless mobile radio systems provide the ability to communicate and access information from any location. As the need or desire for these increases in both the business world and our daily lives, the need for seamless connectivity will grow. As a result, it is expected that mobile radio systems will continue to be an important technology for many years to come. The interested reader is encouraged to investigate this topic further. Refs. 6 and 8 provide a starting point for this endeavor.

BIBLIOGRAPHY

1. J. Campbell, *Rutherford: Scientist Supreme*, Christchurch, NZ, AAS Publications, 1999.
2. E. V. Appleton and W. J. G. Beynon, The application of ionospheric data to radio communication problems: Part I, *Proc. Phys. Soc.* **52**: 518–533 (1940).
3. T. Lewis, *Empire of the Air: The Men Who Made Radio*, New York, HarperCollins, 1991.
4. *A History of the Radio Club of America, Inc.: 1909–1984, Seventy-Fifth Diamond Jubilee Yearbook*, Radio Club of America, Inc., 1984. (Library of Congress Catalog Number 84–061879)
5. T. S. Rappaport, R. L. Campbell, and E. Pocok, A single-hop F2 propagation model for frequencies above 30 MHz and

path distances greater than 4000 km, *IEEE Trans. Antennas Propag.* **38**: 1967–1968 (1990).

6. T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, 2002.
7. R. E. Ziemer and W. H. Tranter, *Principles of Communications: Systems, Modulation and Noise*, 5th ed., Wiley, New York, 2002.
8. R. E. Ziemer and R. L. Peterson, *Introduction to Digital Communications*, 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, 2001.
9. S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (1998).

MODELING AND ANALYSIS OF DIGITAL OPTICAL COMMUNICATIONS SYSTEMS

MARK SHTAIF
Tel-Aviv University
Tel-Aviv, Israel

1. INTRODUCTION

Since the early 1990s the world of optical communications has experienced staggering growth driven by an unprecedented acceleration of demand. Technologically, what enabled this growth was major advancements in the area of fiberoptic components, the most significant of which were the invention of the erbium-doped fiber amplifier (EDFA) [1,2] and the implementation of devices that compensate for the chromatic dispersion of optical fibers [3]. These technologies revolutionized almost every aspect of fiberoptic transmission. They allowed systems to extend to multiple thousands of kilometers without electronic regeneration and made the concept of wavelength-division multiplexing (WDM) cost-effective for the first time, thereby increasing the aggregate capacity of optical systems by many orders of magnitude. Simultaneously with the enhancement in performance, the combination of optical amplifiers with effective dispersion compensation technology has also changed the way in which fiberoptic systems operate. Systems are no longer limited by the optical signal power impinging on the photoreceiver, and therefore effects such as shot noise and even thermal noise generated in the receivers became practically irrelevant. Instead, the dominant noise source is amplified spontaneous emission generated in the amplifiers themselves. Waveform distortions are no longer dominated by the chromatic dispersion of the link; instead it is the optical nonlinearity of the transmission fiber that has become the chief source of distortions and interference. Now, more than ever before, the performance of optical communications systems is dictated not by the imperfection of individual components but chiefly by fundamental physical principles pertaining to signal generation, transmission, and detection. The description of those principles is the main goal of this article.

The structure of a typical optical communication system is illustrated in Fig. 1. It consists of a stack of transmitters, a fiberoptic link, and a stack of optical receivers. The light generated by each transmitter has a specific central optical wavelength (or frequency), and the signals emitted by the various transmitters are optically multiplexed into a single fiber. On the receiver side the optical channels are demultiplexed such that each channel is fed into its own dedicated receiver. The link consists of multiple spans of fiber separated by optical amplifiers. The typical length of a single fiber span ranges from 40 to 120 km, and the length of the entire link varies between a few tens of kilometers (single span) in short-reach applications and 9000 km in transpacific systems. Each amplifier provides gain to compensate for the attenuation incurred in the preceding fiber span, and in most cases it also incorporates a dispersion-compensating module (DCM) that is included in its physical structure and whose role it is to balance the chromatic dispersion in the transmission fiber. As is always the case when modeling complicated systems, certain assumptions need to be made regarding its various components. In our case, since we wish to focus on fundamental limitations, we shall ignore the detailed description of components that do not impose fundamental constraints on system performance. Thus we shall assume that the optical transmitter is capable of generating any reasonable pulseshape that we desire, an assumption that is consistent with most waveforms that are considered favorable for fiberoptic transmission. We will also assume that the modulated electric field of each channel can be approximated as $\sum_k a_k g(t - kT) \exp(-i\omega_{jt})$, where a_k represents the value of the kth symbol, $g(t)$ is the slowly varying envelope of an individual pulse, and ω_j is the central optical frequency of the jth channel. The multiplexer and the demultiplexer used on the two sides of the system are assumed to be perfect in the sense that they combine and separate the individual channels, respectively, without affecting their waveforms. The photodetector is described as a module that generates an electric current that is proportional to the incident optical power. Finally, noise mechanisms such as shot noise, laser noise, and thermal noise at the receiver are ignored as they are negligible relative to the noise contributed by amplified spontaneous emission.

The dominant mechanisms that determine the performance of optical systems are the noise generated in the amplifiers on one hand and waveform distortions taking place in the optical fiber on the other hand. The distortions are due to a combination of fiber dispersion, nonlinearity, and polarization related effects. Although, as pointed out earlier, dispersion by itself can be perfectly compensated for, it affects the way in which fiber nonlinearities distort the optical waveforms, and therefore the interplay between the two is of outmost importance. Polarization-related distortions are caused by mechanical stress and geometric imperfections that break the cylindrical symmetry of optical fibers, thereby making its transmission properties polarization-dependent. They are particularly important in systems operating with exceptionally bad fiber [4], or when the data rates per channel are particularly high [$\geq$40 Gbps (gigabits per second)]. In the following section we discuss the modeling of optical amplifiers. Section 3 deals with the modeling of receivers and

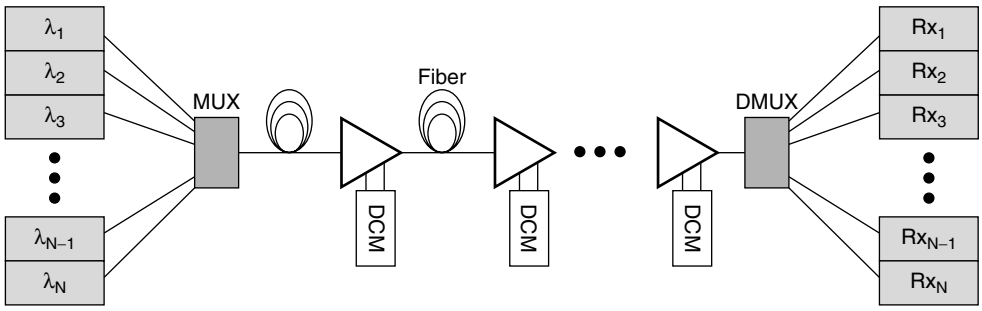


Figure 1. A schematic description of a WDM optical system. The transmitters are labeled by their central wavelengths λ_1 to λ_N. Triangles represent optical amplifiers, and DCM stands for dispersion compensation module.

the performance of systems limited by optical noise. The issues of fiber transmission are dealt with in Sections 4 and 5; Section 4 concentrates on the combination of chromatic dispersion and nonlinearities, and Section 5 reviews polarization-related effects.

2. AMPLIFICATION AND NOISE

One of the most fundamental properties of coherent optical amplification is that it is always accompanied by noise [5–7]. The principles that necessitate this noise and determine its properties can be fully understood only in a quantum-mechanical context, but in all relevant situations when the number of noise photons emitted by the amplifier is much greater than 1, the effect of an optical amplifier has a very accurate classical representation [8]

$$\vec{E}_{\text{out}} = \sqrt{G}\vec{E}_{\text{in}} + \vec{n}(t) \tag{1}$$

where $\vec{E}_{\text{in}}$ and $\vec{E}_{\text{out}}$ are the input and output electric field vectors, respectively, G is the power gain of the amplifier, and $\vec{n}(t)$ is a white Gaussian noise process whose power density spectrum measured along any given state of polarization is $\hbar\omega_0 n_{\text{sp}}(G-1)$. Here ω_0 is the central frequency of the amplified signal, $\hbar$ is Planck's constant, and the term n_{sp}, which is always greater than or equal to 1, accounts for the enhancement of noise in amplifiers that contain loss mechanisms. In amplifiers based on the inversion of carrier populations, losses may result from incomplete inversion, and therefore n_{sp} is commonly known as the inversion factor. Although physically it is the gain of the amplifier and its inversion factor that determine the noise power, amplifier manufacturers and designers often talk about optical amplifiers in terms of their noise figure (NF). The concept of the NF is borrowed from RF amplification, and it describes the deterioration in the signal-to-noise ratio (SNR) of a coherent (shot-noise-limited) signal as a result of optical amplification [9]. Unlike its RF equivalent, the optical noise figure definition is based on the SNR obtained in a measurement of optical energy, where the noise is not additive and therefore the interpretation of the NF is not obvious. A useful relation illustrating the significance of the noise figure can be obtained in the case of high-gain amplifiers, where it can be shown that $\text{NF} \simeq 2n_{\text{sp}}$.

As depicted in Fig. 1, optical systems usually consist of a large number of amplified spans such that each amplifier compensates for the loss of the fiber span that precedes it. In principle, the length of an individual span is a free parameter. One may choose to implement a system with a large number of short spans or a small number of long spans such that in both cases the signal power impinging on the receiver is identical. Yet when it comes to the accumulation of noise, the two scenarios are considerably different from each other. To observe the difference, we express the total noise power at the receiver within a signal bandwidth B as $P_{\text{ASE}} = \hbar\omega_0 n_{\text{sp}}(G-1)BN$, where N is the number of spans. If we denote the overall length of the system by L and the loss coefficient of the fiber by α, then the amplifier gain can be expressed as $G = \exp(\alpha L/N)$ such that it exactly compensates for the losses of the span, and therefore we obtain $P_{\text{ASE}} = \hbar\omega_0 n_{\text{sp}}\alpha LB(G-1)/\ln(G)$. Notice that the minimum noise power is equal to $P_{\text{ASE}} = \hbar\omega_0 n_{\text{sp}}\alpha LB$ and corresponds to the case where the number of spans approaches infinity or $G \to 1$. The penalty for introducing a finite number of spans is described by the function $(G-1)/\ln(G)$, which is plotted in Fig. 2. Notice that for typical gain values in optical systems (of the order of 20 dB) the dependence of the penalty on G is almost linear. The preceding calculation of the noise power assumed that the power that is launched into the system is independent of G, which implies that as we reduce G, the signal power averaged over the system length increases. This can be easily visualized in the limit in which the number of spans goes to infinity and G approaches 1, where the path-averaged power becomes equal to the power launched into the system. As we shall see in the next section, an increase in the optical power along the system enhances the nonlinear effects and increases waveform distortions. Therefore, a more meaningful quantity than the noise power calculated above is the ratio between the signal and noise powers, assuming that the launched signal power is adjusted as a function of G such that the path-averaged power is kept constant [10]. In this case the penalty for using a finite number of spans is given by $F(G) = [(G-1)/\ln(G)]^2/G$ and is shown by the dashed curve in Fig. 2. The obvious advantage in reducing the length of the span and increasing the number of amplified spans in an actual system is balanced by the higher cost of systems containing a larger number of amplifiers.

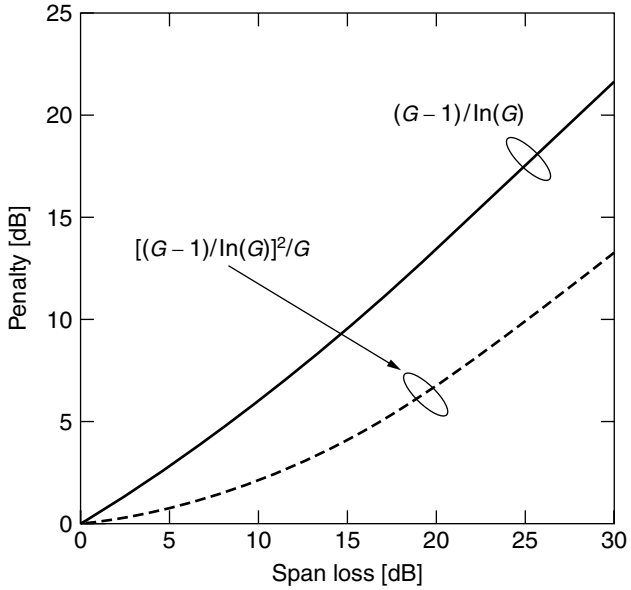


Figure 2. The optical SNR penalty caused by the use of lumped amplification as a function of the gain G (which is equal to the span losses). The solid curve corresponds to the case of fixed launched power, and the dashed curve represents the case of fixed path-averaged power. (After Ref. 10.)

3. MODELING OF OPTICAL RECEIVERS

The detection of light is based on the excitation of electrons by photons in photosensitive materials. This process generates an electric current whose average is proportional to the power of the incident light. The fluctuations around this average are caused by shot noise and as we have noted earlier, they are negligible in amplified systems where the dominant noise contribution comes from spontaneous emission. The vast majority of receivers that are used for optical communications are based on the direct detection of light and are therefore sensitive only to the incident optical power. Receivers that are also capable of detecting the optical phase are called *coherent receivers*, and their principle of operation relies on the coupling of the incoming optical field with a strong local oscillator prior to photodetection [11]. Coherent receivers offer two major advantages: (1) both the intensity and the phase of the optical field can be used for transmitting information and (2) they provide "free" amplification since the measured signal is proportional to the product of the incident optical field with an arbitrarily intense local oscillator. Historically, it was primarily the second advantage that drove the entire field to work on coherent transmission, and therefore with the invention of efficient optical amplifiers, interest in this topic became marginal. But the more fundamental reason for the loss of interest in coherent optical systems is provided by Gordon and Mollenauer [12] and is related to the fact that the optical phase is very easily corrupted by the combined effect of noise and fiber nonlinearities. Therefore in systems that are not limited by the available optical signal power, higher capacities can be achieved when only intensity modulation is used.

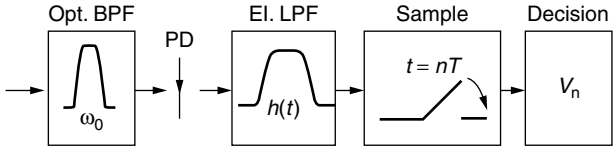


Figure 3. A schematic description of an intensity receiver (Opt. BPF — optical bandpass filter, PD — photodiode, El. LPF — electrical lowpass filter).

The structure of a generic intensity receiver is illustrated in Fig. 3. The optical signal is first filtered optically and photodetected. The electric current generated by the photodetector is filtered electrically and sampled. The samples are then fed into a decision circuit that determines the identity of the transmitted symbol. The sampling period and phase are matched to the received data by a separate clock recovery mechanism that is not shown in the figure. The modulation format that is almost exclusively used in intensity modulated systems is on/off keying, where logical ones and zeros are indicated by the existence or the absence of an optical pulse, respectively. The received signal in this case can be expressed by $\sum_k a_{kg}(t - kT)\exp(-i\omega_{jt})$ where the value of a_k is either 0 or 1. In spite of numerous attempts, modulation schemes using more than two intensity levels have not yet proved themselves usable in optical communications, as we shall explain later in this section. The details of the electric filter in the receiver and its model vary from system to system. Nevertheless, valuable insight into the problem can be obtained by assuming that the impulse response of this filter is square; that is, it is equal to 1 in the time interval between 0 and T and to zero otherwise. Then the effect of the filter is simply to integrate the optical energy of the received signal within the symbol duration so that the value of the kth sample is given by

$$V_k = \int_0^T dt|a_k|^2|g(t) + n(t)|^2 \tag{2}$$

where we have neglected the effect of intersymbol interference (ISI). This simplified version of the optical receiver is commonly referred to as "integrate and dump" and it is attractive as it allows convenient analytic handling. In particular, it can be shown [13] that the probability distributions of V_k corresponding to $a_k = 0$ and $a_k = 1$ are the central and the noncentral chi-square distributions, respectively, with mBT degrees of freedom [14]. The coefficient m is equal to 2 when a properly aligned polarizer is used to select the polarization of the received signal prior to photodetection, whereas in the absence of a polarizer $m = 4$. Figure 4a shows the two probability density functions corresponding to the case of $BT = 5$, $m = 4$, and where the ratio between the optical signal and noise powers (evaluated in a bandwidth B) is equal to 4. This would be the case for example in a typical 10-Gbps transmission system using a 50-GHz optical filter and characterized by an optical SNR (measured in a bandwidth of 0.1 nm) of approximately 12 dB. The same curves are also shown on a logarithmic scale in Fig. 4b. The

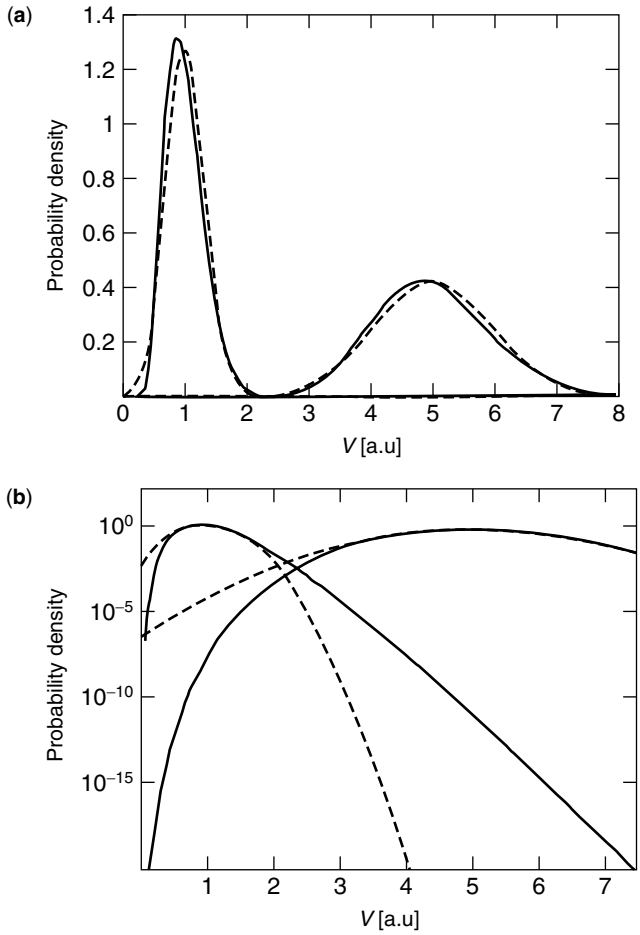


Figure 4. The probability density functions corresponding to the transmission of a logical 0 and 1: (**a**) linear scale; (**b**) logarithmic scale. The solid curves are the exact chi-square distributions, and the dashed curves represent a Gaussian fit to these distributions.

optimal decision threshold is equal to the value of V at the point where the two distributions intersect. Interestingly, the error probabilities given that the transmitted symbol is either 0 or 1 can be shown to be almost identical in a broad range of system parameters, so that the optical channel can be very accurately approximated as symmetric. Figure 4 also shows for comparison the Gaussian distribution functions (dashed curves) that are obtained based on the mean and the variance of the sampled signal. Although the Gaussian and the chi-square distributions are visibly different, it is common practice among optical system engineers to estimate system performance according to the assumption that the received signals are Gaussian distributed. Coincidently, in spite of the apparent difference between them, the two kinds of distributions give very similar average error rates [13]. This fact is illustrated in Fig. 5, which shows the ratio between the approximate Q factor resulting from the Gaussian approximation and the actual Q factor over a broad range of values [where the Q factor is related to the error probability in the usual way: $p_{\text{err}} = 1/\sqrt{2\pi}\int_Q^\infty \exp(-x^2/2)\,dx$]. While Fig. 5 justifies the

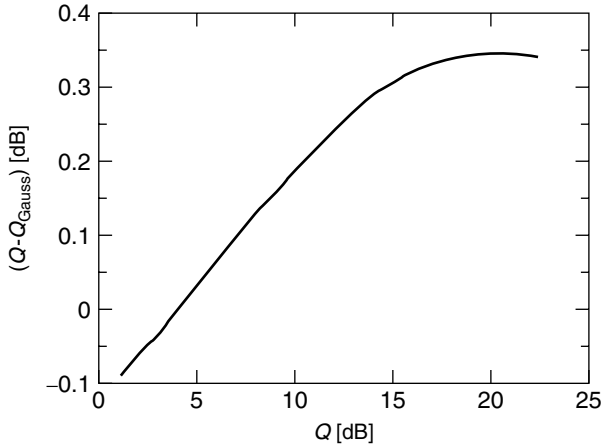


Figure 5. An illustration of the accuracy of the Gaussian approximation in estimations of the average error rate. The vertical axis (ordinate) shows the difference between the Q factor obtained from the Gaussian approximation and the actual Q factor obtained with the chi-square distributions. The horizontal axis corresponds to the actual Q.

use of the Gaussian approximation, it is important to emphasize that it is only the average error rate that can be accurately evaluated this way. The threshold level and the individual error rates conditioned on the transmission of either 0 or 1 cannot be obtained correctly from the Gaussian approximation.

To conclude the subject of linear transmission, we address the question of the ultimate limit to the information capacity of an optical channel. This question can be easily answered in the case of coherent optical systems where Shannon's capacity formula [15] corresponding to channels affected by additive Gaussian noise can be directly applied:

$$C_{\text{coh}} = 4 \times \frac{1}{2T}\log_2\left(1 + \frac{\mathcal{E}}{N_0}\right) \tag{3}$$

where $\mathcal{E}$ is the average energy per symbol duration and N_0 is the power density of the noise measured on both quadratures and both polarizations. The factor of 4 in front of the expression is due to the fact that the signal can be transmitted over 4 degrees of freedom (i.e., two states of polarization and two orthogonal quadratures). In the case of an intensity-modulated system the general calculation of the capacity is very difficult, but a simple and intuitive solution can be obtained in the limit where the optical SNR is much greater than 1, as is almost always the case in actual systems. In this limit it can be shown that the channel capacity (for both states of polarization) is given by [16]

$$C_{\text{Int}} \simeq \frac{1}{T}\log_2\left(\frac{\mathcal{E}}{2N_0}\right) \tag{4}$$

As usual, the capacity (4) represents the highest transmission rate in bits per second that can be reliably achieved in an intensity-modulated optical channel. It implies optimal coding and involves no constraints on

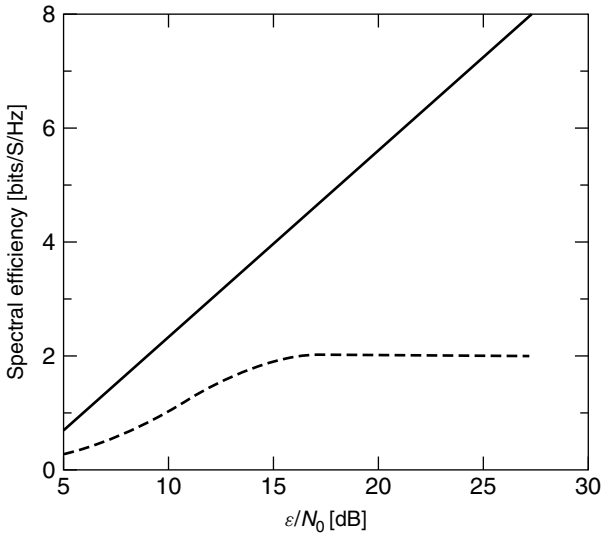


Figure 6. The spectral efficiency of a system using square-law detection. The solid curve corresponds to multilevel transmission constrained only by the average optical power. The dashed curve represents the case of on/off keying. (After Ref. 16.)

the number of transmitted intensity levels. The *spectral efficiency* of an intensity modulated fiberoptic system is defined as the channel capacity per unit of bandwidth. Since a symbol of a temporal duration T occupies a bandwidth equal to at least T^{-1}, the spectral efficiency is given by expression (4) multiplied by the symbol duration T. Figure 6 shows the spectral efficiency extracted from (4) together with the result for on/off keying transmission, as a function of the optical SNR $\mathcal{E}/N_0$. The difference between the two capacities shows the maximum advantage that can be extracted from the use of multilevel intensity modulation. As can be observed in the figure, for typical values of the optical signal to noise ratios (between 10 and 20 dB) multilevel signaling can increase the capacity of an on/off-modulated system by only very small factors. Such small improvement seldom justifies the enormous increase in the complexity and cost that are implied by the use of multilevel optical transmitters and receivers. Notice that the curves shown in Fig. 6 do not take into account the effect of optical nonlinearities, which would undoubtedly reduce the advantage of multilevel modulation even further [17].

4. FIBER TRANSMISSION

The electric field in single-mode optical fibers can be very accurately represented in the form [18]

$$\vec{E}(t,x,y,z) = E(t,z)F(x,y)\exp[-i(\omega_0 t - \beta_0 z)]\hat{u}(t,z) \qquad (5)$$

where z is the propagation axis and x, y are the lateral dimensions of the fiber. The parameter ω_0 is the central frequency of the signal, and β_0 denotes the wavevector at $\omega = \omega_0$. The lateral profile of the beam is $F(x,y)$ and is assumed to remain constant in the process of propagation. The vector $\hat{u}(t,z)$ is a unit polarization vector and lies almost entirely in the x,y plane. In most cases of relevance the transmitted information resides only in the complex envelope of the electric field, which is denoted by the term $E(t,z)$ and is normalized such that $|E(t,z)|^2$ is the optical power in watts. When the effect of polarization-related impairments on system performance is small, it can be shown that the evolution of $E(t,z)$ along the optical fiber is accurately described by the so-called nonlinear Schrödinger equation (NLSE) [18,19]

$$\frac{\partial E}{\partial z} = -i\frac{\beta_2}{2}\frac{\partial^2 E}{\partial t^2} + i\gamma|E|^2E - \frac{\alpha}{2}E \qquad (6)$$

where the first term on the right-hand side describes the effect of chromatic dispersion, the second corresponds to fiber nonlinearity, and the third term corresponds to the scattering losses of the fiber. The derivation of the NLSE relies on the fact that the index of refraction in the fiber can be approximated as $n(\omega,|E|^2) = n(\omega) + n_2|E|^2/A_{\text{eff}}$, where n_2 is the nonlinear refractive index and A_{eff} is the effective cross-sectional area of the beam [18]. Thus the dispersion coefficient is $\beta_2 = \partial^2/\partial\omega^2[\omega n(\omega)/c]$ and the nonlinearity coefficient is $\gamma = \omega_0 n_2/(cA_{\text{eff}})$, where c is the velocity of light in vacuum. The dispersion coefficient can also be expressed in terms of the group velocity v_g in the fiber $\beta_2 = \partial(v_g^{-1})/\partial\omega$ so that it reflects the dependence of the group velocity on the optical frequency. Finally, Eq. (6) is expressed in a delayed timeframe where the average group delay of the propagating pulse is factored out. This means that the time axis t should be interpreted as $t - z/v_g$ in the original representation.

Some insight into the evolution of optical signals in fibers can be obtained by considering cases in which either dispersion or nonlinearity is negligible. First, let us assume that the optical power is low enough to neglect the nonlinear term in the NLSE. Then Eq. (6) can be trivially solved in the Fourier domain yielding

$$\tilde{E}(\omega,z) = \tilde{E}(\omega,0)\exp\left(\frac{i\beta_2\omega^2 z}{2}\right)\exp\left(\frac{-\alpha z}{2}\right) \qquad (7)$$

where $\tilde{E}(\omega,z) = \int_{-\infty}^{\infty} dt\, E(t,z)\exp(i\omega t)$, and ω represents the deviation of the optical frequency from ω_0. Notice that chromatic dispersion does not change the spectral content of the propagating signal since it merely multiplies the original spectrum by a transfer function whose amplitude is equal to 1. Yet it describes a situation where the group velocity varies across the pulse spectrum such that the various frequency components of the pulse tend to separate in time. The most instructive example that illustrates this effect is the case in which the launched pulse is Gaussian: $E(t,0) = A\exp(-t^2/(2\tau_0^2))$. Then the dispersed pulse can be expressed analytically in the time domain $E(t,z) = A(z)\exp[-t^2(1+iC)/(2\tau^2)]$, where C is the chirp parameter $C = \text{sign}(\beta_2)z/L_d$, τ is the width of the dispersed pulse $\tau = \tau_0\sqrt{1+z^2/L_d^2}$, and $L_d = \tau^2/|\beta_2|$ is a parameter with units of length and it describes a characteristic length scale for the effect of dispersion. If we now consider the instantaneous frequency of the pulse, which is equal to minus the time derivative of

its phase $\Delta\omega = -\partial\varphi/\partial t = Ct/\tau^2$, we see that it changes linearly across the pulsewidth. When the fiber dispersion is negative, $\beta_2 < 0$, as is the case in most fibers used for transmission, the leading edge of the pulse consists primarily of the high-frequency content and the trailing edge consists primarily of the low-frequency part. The opposite occurs when $\beta_2 > 0$. Before we conclude the discussion of dispersion as a "standalone" process, we note that it is common practice among optical system engineers to define a different (although equivalent) dispersion coefficient that reflects the dependence of the group velocity on wavelength instead of frequency $D = \partial(v_g^{-1})/\partial\lambda$. It is easy to show that the two coefficients are related to each other by $D = -2\pi c\beta_2/\lambda^2$. It is also common to refer to "normal" and "anomalous" dispersion with regard to the cases $\beta_2 > 0$ ($D < 0$) and $\beta_2 < 0$ ($D > 0$), respectively. Most transmission fibers, as we have noted earlier, are used in the anomalous regime.

To illustrate the effect of fiber nonlinearities we consider the case in which dispersion is negligible. Then the evolution of the electric field assumes the form

$$E(t, z) = E(t, 0)\exp(i\gamma|E(t, 0)|^2 z_{\text{eff}})\exp(-\alpha z/2), \qquad (8)$$

where $z_{\text{eff}} = \int_0^z dz\exp(-\alpha z) = [1 - \exp(-\alpha z)]/\alpha$ is called the *effective length* of fiber, which is the length scale that characterizes the effect of scattering losses. In most relevant cases $\exp(-\alpha L) \ll 1$ so that $z_{\text{eff}} \simeq \alpha^{-1}$. Notice that the effect of the nonlinearity is to modulate the phase of the propagating signal while its intensity remains unperturbed. Therefore this phenomenon is frequently referred to as *self-phase modulation* (SPM). The characteristic length describing the effect of fiber nonlinearities is defined as the effective length of fiber in which the acquired maximum phase shift is equal to 1 radian. It is given by $L_{NL} = (\gamma P_{\text{peak}})^{-1}$, where P_{peak} is the peak power of the optical pulse. The relative significance of dispersion and nonlinearity can therefore be estimated by comparing the characteristic lengths L_{NL} and L_d. As in the case of dispersion, the propagated pulse is characterized by a frequency chirp as its phase becomes time-dependent. The effect on the instantaneous frequency is once again obtained from the derivative of the optical phase $\Delta\omega = -\partial\varphi/\partial t = -\gamma z_{\text{eff}}\partial|E(t, 0)|^2/\partial t$ and it is plotted in Fig. 7 for the case of a typical waveform. As illustrated in the figure, the optical frequency of the leading edge is downshifted and the frequency of the trailing edge is upshifted as a result of the nonlinearity. Consequently, the effect on initially un-chirped pulses will always be that of spectral broadening. Notice, however, that when the launched pulse contains chirp that is opposite to the one caused by nonlinearity (i.e., the leading edge consists of higher frequencies than the trailing edge), the effect of nonlinearity is to equalize the spectrum so that the bandwidth is compressed. This is exactly the situation that occurs when the pulse entering the nonlinear section of fiber is initially predispersed in a linear section of fiber that is characterized by anomalous dispersion.

The general analysis of fiber transmission in the presence of both dispersion and nonlinearity can be handled only numerically. The most efficient and widely used numerical technique for solving the NLSE is the split-step Fourier method [18]. This technique relies on the fact that when the fiber is divided into sufficiently short increments, the propagation of the electric field through each increment can be evaluated in two steps. In the first step the field is propagated through the fiber increment, assuming that it is purely dispersive. In the second step the dispersed signal is propagated through the same increment, assuming that it only contains nonlinearity. The order of the two steps is purely arbitrary and therefore can be reversed without affecting the accuracy of the computation. The representation of the split-step method offers some qualitative insight into the interaction between dispersion and nonlinearity. In particular, we may consider the case of fibers operated in the anomalous regime where the chirp induced by dispersion is opposite in sign to that induced by nonlinearity. Then we may expect that the dispersive step and the nonlinear step in the split-step method will tend to balance each other. A pulse that takes the most advantage of this balance is the optical soliton [19], whose waveform propagates unperturbed along the optical fiber. The electric field envelope of the soliton is given by $E(t, z) = A\,\text{sech}(t/\tau)\exp[-i\beta_2 z/(2\tau)]$, where the pulsewidth τ and the amplitude A are related to each other and to the fiber parameters through $A\tau = \sqrt{|\beta_2|/\gamma}$. It can be shown by direct substitution that the soliton is an exact solution of the NLSE in the absence of scattering losses. In the presence of scattering losses the soliton does not maintain its shape exactly because the strength of the nonlinear effect changes with propagation. Nevertheless, it has been demonstrated that as long as the dispersion length is much greater than the separation between adjacent amplifiers, soliton transmission can exist in a path-averaged sense [20]. Although optical solitons may appear to be a natural solution for fiberoptic transmission, their applicability to optical communications systems

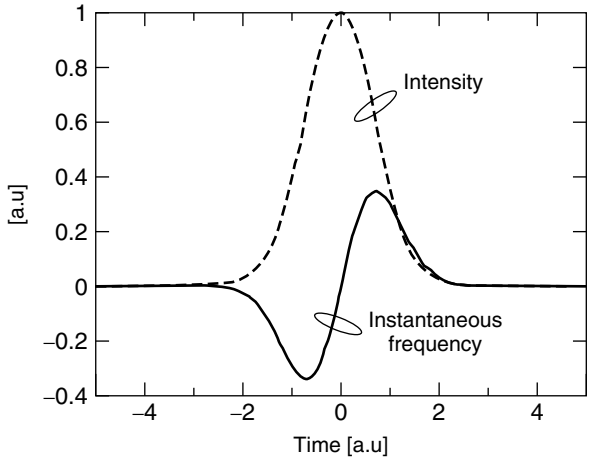


Figure 7. An illustration of SPM in optical fibers. The dashed curve represents the power profile of an optical pulse. The solid curve is the instantaneous frequency shift due to XPM. The shift is proportional to minus the time derivative of the intensity. The leading edge is therefore downshifted in frequency and the trailing edge is upshifted.

remains very limited. What limits their use is primarily the so-called Gordon–Haus effect [21] related to the periodic addition of amplified spontaneous emission noise to the propagating signals, and nonlinear interactions with pulses in adjacent WDM channels, which we review later in this section. Both these effects generate random perturbations of the central frequency of the soliton that are translated into group velocity perturbations due to chromatic dispersion and cause uncertainty in the arrival time of the pulse. Notice that such perturbations of the central frequency occur with all pulses, regardless of their shape. What makes solitons particularly sensitive to frequency perturbations is the fact that they are transmitted without any compensation for chromatic dispersion. Therefore small random variations of the central frequency are translated into enormous variations in the arrival time of the pulse. In spite of very clever ideas for preventing the accumulation of random frequency shifts [22,23], true soliton systems are not used in optical communications. Instead, systems that are designed for reliable long-haul communications are constructed such that most of the dispersion of the link is compensated. Pulses propagating in such systems necessarily change their properties in the process of propagation, and their evolution is strongly dependent on the initial pulseshape and on the amounts of dispersion compensation that are applied along the link. A particularly attractive solution for long-haul transmission is the so-called dispersion-managed soliton, which was originally described in 1995 and 1996 [24,25] and has been extensively studied since then. Dispersion-managed solitons are nearly Gaussian pulses whose evolution is periodic with the period of exactly one span. Specifically, the pulse parameters are chosen such that it undergoes spectral broadening in the first part of its propagation through the fiber span, and then its spectrum is recompressed as it continues to propagate. After applying a properly chosen amount of dispersion compensation at the end of the span, the pulse returns almost precisely to its original waveform. The evolution of dispersion managed solitons can be described quite accurately in a simplified model tracking only a few parameters. These are either the pulse duration and chirp [26], or its bandwidth and frequency dispersion [27]. The reduced models provide very simple numerical methods for extracting the system parameters that are required for dispersion managed soliton transmission. Actual systems parameters are frequently constrained by practical considerations that do not allow matching of the dispersion managed soliton condition exactly. In such cases, when the deviations are small, transmission is often characterized by a periodic "breathing" of the pulse duration and bandwidth with a period of several spans. When the deviation from the dispersion managed soliton requirements is large, the pulses may become completely corrupted, eventually preventing reliable transmission.

In high-channel-count WDM systems one of the most significant impairments to system performance is caused by nonlinear crosstalk between channels [28]. The physical mechanism is quite simple. Every WDM channel is affected by the nonlinear modulation of the refractive index that is caused by the combined intensity of all other channels in the fiber. This process manifests itself in two ways: (1) four-wave mixing (FWM) and the (2) cross-phase modulation (XPM). The *four wave mixing* effect is completely analogous to intermodulation distortions in electronic systems. Any two channels at optical frequencies ω_i and ω_j cause the total optical intensity to oscillate at the frequency difference $\Delta\omega_{i,j} = \omega_j - \omega_i$. These oscillations are imprinted on the refractive index of the fiber such that the propagation of all channels through the fiber is affected. Specifically, a channel at optical frequency ω_k is modulated by the refractive index oscillations and therefore generates new tones at frequencies $\omega_k \pm \Delta\omega_{i,j}$. This interaction occurs between all possible pairs ($\omega_k = \omega_i \neq \omega_j$) and triplets ($\omega_k \neq \omega_i \neq \omega_j$), and therefore, even in systems with a moderate number of channels, an enormous number of new components is created [28]. These components are added as noise to the transmitted channels and may cause significant penalties to system performance. A parameter that strongly affects the efficiency of the FWM process is the chromatic dispersion of the fiber. In the presence of chromatic dispersion, signals at different optical frequencies propagate with different velocities and their phases are acquired at different rates. This implies that the phase with which FWM products are created (which is defined by the phase of the signals that create them) does not match the phase that they acquire during propagation, and therefore their intensity is significantly reduced at the system output. For similar reasons the FWM efficiency also reduces monotonically when the frequency separation between adjacent WDM channels is increased.

The other form of interchannel interference is *cross-phase modulation* (XPM), which is caused by the modulation of the phase of each channel propagating through the fiber by the intensities of all other channels. It is instructive to consider this phenomenon first when the effect of chromatic dispersion is negligible. Then the evolution of the ith channel is expressed analytically as $E_i(t,z) = E_i(t,0)\exp\left[i\gamma z_{\text{eff}}\left(|E_i|^2 + 2\sum_{j\neq i}|E_j|^2\right)\right]$ [18], as can be obtained by direct substitution of the electric field corresponding to multiple channels into Eq. (8) and ignoring the added noise generated by the FWM products. The XPM effect is embodied in the second term in the exponent and similarly to SPM, which we reviewed earlier (represented here by the phase dependence on $|E_i|^2$), XPM affects the instantaneous frequency of the considered channel. Unlike the case of SPM, however, the frequency shift caused by XPM depends on the temporal overlap between the interacting pulses. To clarify this point let us consider the interference between two pulses belonging to two different WDM channels, as illustrated in Fig. 8. Recall that the shift of the optical frequency is given by minus the time derivative of the optical phase. The contribution of XPM to the instantaneous frequency of each pulse is therefore proportional to minus the time derivative of the intensity of the other pulse. As a result, the optical frequency of the leading pulse is reduced by the interaction, whereas the optical frequency of the trailing pulse is increased. In the absence of chromatic dispersion this interference

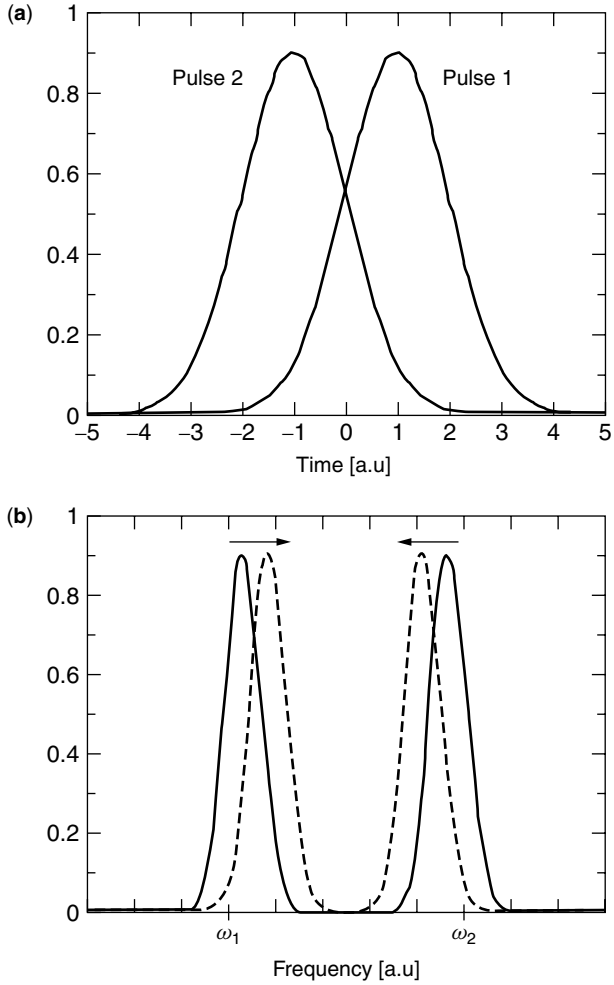


Figure 8. An illustration of a XPM interaction between two pulses transmitted over two different optical frequencies. The solid curves in (**b**) represent the original optical frequencies of the two pulses. The dashed curves represent the spectral shift caused by XPM. The frequency shift experienced by each pulse is proportional to minus the time derivative of the intensity of the other pulse.

turns into a problem only in coherent systems, where information is transmitted on the optical phase, or in intensity-modulated systems using very narrow optical filters that translate the frequency shifts caused by XPM into intensity modulation. Yet zero dispersion fibers are very rarely encountered with WDM transmission since their use would result in significant FWM impairments, as we have discussed earlier. Therefore the more relevant situation to consider is that of transmission in dispersive fibers. The presence of chromatic dispersion modifies the picture presented in Fig. 8 in a number of ways. The most significant modification is that the temporal overlap between the two interacting pulses changes as they propagate along the fiber. In fact, they can overlap only in a fiber section that is no longer than $2\tau_p/(\beta_2\Delta\omega_{i,j})$, where τ_p denotes their duration and $\Delta\omega_{i,j}$ is the difference between their central optical frequencies (so that $\beta_2\Delta\omega_{i,j}$ is the difference between their inverse group velocities). Since the frequency separation $\Delta\omega_{i,j}$ is typically much higher than the spectral width of the individual channels, the effect of dispersion on the individual pulseshapes is negligible within the section of fiber in which the pulses overlap. Finally, the shifts in the optical frequency of the pulses that occur as a result of their interaction are translated into shifts in their group velocities, which lead to uncertainty in their arrival times and cause timing jitter. Notice that chromatic dispersion has two contradicting effects on the significance of XPM-induced penalties. On one hand it is responsible for translating frequency shifts into timing jitter, as we explained earlier, but on the other hand the higher the dispersion, the shorter is the section of fiber in which the pulses overlap, and therefore the magnitude of the frequency shifts becomes smaller [29]. The key difference between these two contradicting mechanisms is that the first is caused by the accumulated dispersion between the section where the pulses interact and the receiver, whereas the latter is related to the local dispersion that characterizes the fiber section in which the nonlinear interaction takes place. Therefore, to reduce the significance of XPM in systems it is beneficial to use high dispersion fiber and compensate for most of the accumulated dispersion at the end of each span.

Other nonlinear phenomena that are not included in the NLSE are the Brillouin and Raman effects, which are both related to the interaction of light with vibrational modes of the medium, or phonons [18]. The Brillouin effect is caused by scattering of light from acoustic phonons and it manifests through the generation of a back scattered wave that carries most of the energy once the incident power exceeds a certain threshold. It is a narrowband phenomenon and can be very effectively avoided by artificially broadening the laser linewidth when necessary [30]. The Raman effect is related to scattering of light from optical phonons. Its most relevant manifestation in communications systems is observed when energy from high-frequency WDM channels is transferred to channels at low frequencies, creating a tilt in the transmitted optical spectrum. The efficiency of the energy transfer induced by Raman scattering peaks for channels separated by approximately 13 THz [31]. This implies that the interacting channels differ significantly in their group velocities as a result of chromatic dispersion, so that the interference between them is averaged with respect to the transmitted data. Consequently, the Raman effect does not constitute a significant mechanism for interchannel crosstalk [28]. The most important application of the Raman effect in fiberoptic systems is in Raman amplifiers, where an intense pump with an appropriately selected optical frequency is launched into the fiber such that it provides gain to the data-carrying signals [32]. The main advantage of this amplification technique is that the gain is distributed along the transmission fiber itself, thereby offering a SNR improvement relative to lumped amplification, consistent with the discussion in Section 2.

5. POLARIZATION-RELATED EFFECTS

Polarization effects in optical systems result from the dependence of the transmission properties of optical fibers and components on the polarization of the transmitted

light. This dependence stems from geometric distortions and mechanical stress that violate the cylindrical symmetry of optical fibers and are created in the process of fabrication and cabling [4]. The most significant consequence of these distortions is *optical birefringence*, which is characterized by the existence of two orthogonal polarization axes $\hat{e}_1$ and $\hat{e}_2$ having different indices of refraction. If we choose to express the electric field as a column vector in the representation of the base defined by the axes of birefringence, the relation between input and output fields is described by

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{out}} = \exp(i\beta_0 z) \begin{pmatrix} \exp\left(\frac{i\Delta\beta z}{2}\right) & 0 \\ 0 & \exp\left(\frac{-i\Delta\beta z}{2}\right) \end{pmatrix} \times \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{in}} \tag{9}$$

where z is the length of the birefringent section, β_0 is the average wavenumber, and $\Delta\beta$ is the difference between the wavenumbers corresponding to the two birefringent axes. It can be expressed approximately as $\Delta\beta = \Delta n\omega/c$, where Δn is the difference between the indices of refraction of the two axes. The electric field vectors expressed in this way are called *Jones vectors*, and the matrix relating them to each other is called a *Jones matrix*. Notice that the polarization state of a signal is defined jointly by the relative amplitudes of its components and the phase difference between them. Therefore the polarization state of signals that have nonzero components on both $\hat{e}_1$ and $\hat{e}_2$ changes as a result of birefringent propagation. Furthermore, the acquired phase difference and therefore the polarization of the output signal are frequency-dependent, describing a situation that is known as polarization mode dispersion (PMD). If we consider an optical pulse that is launched into a birefringent section of fiber, the output waveform consists of two orthogonally polarized replica of the pulse delayed by $\Delta\beta z$ relative to each other. This situation can be very harmful to the performance of optical communications systems, but it can be easily avoided in a number of ways. For example, one may consider transmitting the signal in a polarization state that coincides with either $\hat{e}_1$ or $\hat{e}_2$ such that no waveform distortions exist at the output. Yet, the more relevant situation to consider is that of a fiber in which the axes of birefringence change in the process of propagation. The most common way of modeling such fibers is by describing them as if they were made of many discrete birefringent sections with random and statistically independent axes of birefringence. This model is justified due to the fact that the correlation length of the birefringence in optical fibers is smaller by several orders of magnitude than the system length [33], so that the details of the local birefringence statistics become irrelevant. In this situation the relation between the input and output states of polarizations is given by

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{out}} = \mathbf{T}(i\omega) \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{in}} \tag{10}$$

where $\mathbf{T} = \mathbf{T}_N\mathbf{T}_{N-1}\cdots\mathbf{T}_1$, where $\mathbf{T}_i$ is the birefringence matrix of the ith fiber section. Notice that although each individual matrix can be expressed in a diagonal form as the matrix in Eq. (9), they are not diagonal in the same representation and therefore the combined matrix $\mathbf{T}(i\omega)$ is a generic 2×2 unitary matrix that may have an arbitrary dependence on the optical frequency. In this case there is no simple description of the waveform distortions generated by the transmission of a general pulse. An important result can be obtained for the case in which the bandwidth of the launched signal is small relative to the bandwidth that characterizes the frequency dependence of $\mathbf{T}$ [34]. Then Eq. (10) can be approximated in the vicinity of the central frequency of the launched signal ω_0, and it is possible to show that there are two orthogonal states of polarization $\hat{e}_1$ and $\hat{e}_2$ in whose representation equation (10) assumes the form [34]

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{out}} \simeq e^{i(\omega-\omega_0)\tau_0/2} \begin{pmatrix} e^{-i(\omega-\omega_0)\tau/2} & 0 \\ 0 & e^{i(\omega-\omega_0)\tau/2} \end{pmatrix} \times \mathbf{T}(i\omega_0) \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}_{\text{in}} \tag{11}$$

where $\mathbf{T}(i\omega_0)$ is the polarization rotation corresponding to the central frequency and τ_0 is the average, polarization-independent time delay. Notice that apart from the frequency-independent operator $\mathbf{T}(i\omega_0)$, Eq. (11) is identical to Eq. (9), which represents the case of birefringence with fixed axes. Therefore a pulse with nonzero components on $\hat{e}_1$ and $\hat{e}_2$ comes out of the fiber consisting of two orthogonally polarized replica of itself delayed by τ relative to each other. The time delay τ is known as the *differential group delay* (DGD), and the polarization states $\hat{e}_1$ and $\hat{e}_2$ are known as the *principal states of polarization* (PSPs). Equation (11) represents what is known as the first-order PMD approximation because it relies on the first-order expansion of the matrix $\mathbf{T}$ with respect to frequency. This description has been shown to be very useful in predicting the kind of penalties induced by PMD in fiberoptic systems in a reasonably broad range of parameters [4,35]. The further analysis of PMD requires tools whose description is beyond the scope of this article [36]. We will therefore limit ourselves to stating some of the most relevant results concerning this phenomenon. One obvious observation is that the DGD and the PSPs are stochastic processes with respect to both the central frequency and the position along the fiber. It can be shown that the differential group delay at any fixed frequency is a Maxwell distributed parameter and its mean value $\langle\tau\rangle$ is proportional to the square root of the fiber length [37]. The bandwidth that characterizes the frequency dependence of the transfer matrix $\mathbf{T}(i\omega)$ can be estimated by deriving frequency autocorrelation functions of the DGD and the principal states [38–40], which indicate that the bandwidth of all phenomena related to PMD is of the order of $\langle\tau\rangle^{-1}$. The nature of higher-order distortions that are caused by PMD is still one of the most active topics of research in optical communications. With the rapid increase in data rates transmitted over modern systems, the importance of this topic is growing accordingly. One important general property of PMD that is related to the rapidly increasing data rates is that its

effect can be rigorously scaled on the basis of the mean DGD of the fiber [40]. Thus for example a 10-Gbps channel transmitted over a system with 10 ps mean DGD experiences the same waveform distortions as a 40-Gbps channel transmitted over a system with a mean DGD of 2.5 ps.

In addition to the effect of birefringence, which limits the performance of optical systems by distorting the transmitted waveforms, optical systems can also be penalized by the existence of polarization-dependent loss (PDL) along the optical link. As is suggested by its name, PDL describes a situation in which there exist two orthogonal polarization components characterized by different attenuation. Whereas PMD is caused primarily by the birefringence of optical fibers, the main source of PDL is the imperfection of inline optical components such as isolators, dispersion compensating devices, and optical switches. In principle, PDL may contribute to the generation of waveform distortions through a nontrivial interaction with the fiber birefringence [41]. In practice, however, considering typical system parameters, the effect of PDL on the waveform is usually small. Instead, it effects the performance of optical communications systems by penalizing the optical signal-to-noise ratio of the received signals [42,43]. The deterioration of the signal-to-noise ratio occurs in two ways. The first is caused by the fact that components having PDL may attenuate the propagating signal by more than the average attenuation and therefore the SNR is reduced accordingly [42]. The second results from the fact that in the presence of PDL, noise that is originally emitted into a state of polarization that is orthogonal to the signal is coupled into the signal's polarization [43] and therefore mixes with the signal at photodetection. The significance of PDL is limited primarily to long-haul terrestrial systems using affordable components.

The preceding description of polarization-related phenomena assumed that they could be treated separately from nonlinear propagation. This assumption is characteristic of the vast majority of studies that have been reported in the literature so far. It is reasonable in systems operating over high PMD links [4], where PMD is the most significant limiting factor. A more general treatment of polarization related effects is based on the so-called coupled nonlinear Schrödinger equations [44–46], which can simultaneously take into account the effects of birefringence, PDL, dispersion, and nonlinear propagation. Such simultaneous treatment of all propagation phenomena is particularly important in long-haul systems where proper operation requires fine control of the dispersion map. In those cases even moderate levels of PMD can disturb the system and cause significant penalties to performance.

Acknowledgment

The author is pleased to acknowledge A. Mecozzi and J. P. Gordon for the multiple discussions that were conducted on the topics included in this article.

BIOGRAPHY

Mark Shtaif received his M.Sc. and Ph.D. degrees in electrical engineering at the Technion in 1993 and 1997, respectively. In 1997 he joined the Light-wave Networks Research Department at AT&T Labs Research as a Senior and then Principal Member of Technical Stuff. In AT&T his work was centered around the modeling and characterization of optical fiber communication systems, focusing on propagation effects in optical fibers, including fiber nonlinearities, polarization mode dispersion, special modulation formats, and interaction of signals and noise. During his employment in AT&T he served as a technical consultant to the AT&T business units, on the evaluation of fiberoptic technologies. In December 2001 he became a Principal Architect in Celion Networks, an optical networking company, where he worked on the analysis and design of long-haul optical transmission systems. In April 2002 Dr. Shtaif joined the faculty of the Engineering Department in Tel-Aviv University, where he conducts research and teaches courses in the area of optical communications.

BIBLIOGRAPHY

1. R. J. Mears, L. Reekie, I. M. Jauncey, and D. N. Payne, Low noise erbium doped fibre amplifier operating at 1.54 μm, *Electron. Lett.* **23**: 1026 (1987).
2. E. Desurvire, J. R. Simpson, and P. C. Becker, High-gain erbium doped fiber amplifier, *Opt. Lett.* **12**: 888 (1987).
3. A. H. Gnauck and R. M. Jopson, Dispersion compensation for optical fiber systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 7.
4. C. D Pool and J. A. Nagel, Polarization effects in lightwave systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 6.
5. K. Shimoda, H. Takahasi, and C. H. Townes, Fluctuations in the amplification of quanta with application to Maser amplifiers, *J. Phys. Soc. Jpn.* **12**: 686 (1957).
6. H. A. Haus and J. A. Mullen, Quantum noise in linear amplifiers, *Phys. Rev.* **128**: 2407 (1962).
7. H. Kogelnik and A. Yariv, Considerations of noise and schemes for its reduction in laser amplifiers, *Proc. IEEE* **52**: 165 (1964).
8. E. Desurvire, *Erbium Doped Fiber Amplifiers: Principles and Applications*, Wiley, New York, 1994, Chap. 2.
9. H. A. Haus, Noise figure definition valid from RF to optical frequencies, *IEEE J. Select. Top. Quant. Electron.* **6**: 240 (2000).
10. J. P. Gordon and L. F. Mollenauer, Effects of fiber nonlinearities and amplifier spacing on ultra-long distance transmission, *J. Lightwave Technol.* **9**: 170 (1991).
11. L. Kazovsky, S. Benedetto, and A. Willner, *Optical Fiber Communications Systems*, Artech House, Norwood, MA, 1996, Chap. 4.
12. G. P. Gordon and L. F. Mollenauer, Phase noise in photonic communications systems using optical amplifiers, *Opt. Lett.* **15**: 1351 (1990).
13. P. A. Humblet and M. Azizoglu, On the bit error rate of lightwave systems with optical amplifiers, *J. Lightwave Technol.* **9**: 1576 (1991).

14. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 2001.

15. C. E. Shannon, A mathematical theory of communication, *Bell. Syst. Tech. J.* **27**: 379 (July 1948); 623 (Oct. 1948).

16. A. Mecozzi and M. Shtaif, On the capacity of intensity modulated systems using optical amplifiers, *IEEE Photon. Technol. Lett.* **13**: 1029 (2001).

17. P. P. Mitra and J. B. Stark, Nonlinear limits to the information capacity of optical fibre communications, *Nature* **411**: 1027 (2001).

18. G. P. Agrawal, *Nonlinear Fiber Optics*, Academic Press, San Diego, CA, 1989.

19. L. F. Mollenauer, J. P. Gordon, and P. V. Mamyshev, Solitons in high bit-rate long-distance transmission, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 12.

20. L. F. Mollenauer, J. P. Gordon, and M. N Islam, Soliton propagation in long fibers with periodically compensated loss, *IEEE J. Quant. Electron.* **22**: 157 (1986).

21. J. P. Gordon and H. A. Haus, Random walk of coherently amplified solitons in optical fiber transmission, *Opt. Lett.* **11**: 665 (1986).

22. A. Mecozzi, J. D. Moores, H. A. Haus, and Y. Lai, Soliton transmission control, *Opt. Lett.* **16**: 1841 (1991).

23. L. F. Mollenauer, J. P. Gordon, and S. G. Evangelides, The sliding frequency guiding filter, an improved form of soliton jitter control, *Opt. Lett.* **17**: 1575 (1992).

24. M. Suzuki et al., Reduction of Gordon-Haus timing jitter by periodic dispersion compensation in soliton transmission, *Electron. Lett.* **31**: 2027 (1995).

25. N. J. Smith et al., *Electron.Lett.* **32**: 54 (1996).

26. J. N. Kutz, P. Holmes, S. G. Evangelides, and J. P. Gordon, Hamiltonian dynamics of dispersion managed breathers, *J. Opt. Soc. Am.* **15**: 87 (1998).

27. J. P. Gordon and L. F. Mollenauer, Scheme for the characterization of dispersion managed solitons, *Opt. Lett.* **24**: 223 (1999).

28. F. Forghieri, R. W. Tkach, and A. R. Chraplyvy, Fiber nonlinearities and their impact on transmission systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, CA, 1997, Chap. 8.

29. M. Shtaif, An analytical description of cross phase modulation in dispersive optical fibers, *Opt. Lett.* **23**: 1191 (1998).

30. D. A. Fishman and J. A. Nagel, Degradations due to stimulated Brillouin scattering in multigigabit intensity-modulated fiber-optic systems, *J. Lightwave Technol.* **11**: 1721 (1993).

31. R. H. Stolen and E. P. Ippen, Raman gain in glass optical waveguides, *Appl. Phys. Lett.* **22**: 294 (1973).

32. P. B. Hansen et al., Capacity upgrades of transmission systems by Raman amplification, *IEEE Photon. Technol. Lett.* **9**: 262 (1997).

33. A. Galtarossa, L. Palmieri, M. Schiano, and T. Tambosso, Measurement of birefringence correlation length in long, single-mode fibers, *Opt. Lett.* **26**: 962 (2001).

34. C. D. Poole and R. E. Wagner, Phenomenological approach to polarization dispersion in long single mode fibers, *Electron. Lett.* **22**: 1029 (1986).

35. H. Kogelnik, R. M. Jopson, and L. E. Nelson, Polarization mode dispersion, in I. P. Kaminow and T. Li, eds., *Optical Fiber Telecommunications Ivb: Systems and Impairments*, Academic Press, San Diego, 2002, Chap. 15.

36. J. P. Gordon and H. Kogelnik, PMD fundamentals, *Proc. Nat. Acad. Sci.* **97**: 4541 (2000).

37. G. J. Foschini and C. D. Poole, Statistical theory of polarization dispersion in single mode fibers, *J. Lightwave Technol.* **9**: 1439 (1991).

38. M. Karlsson and J. Brentel, Autocorrelation function of the polarization mode dispersion vector, *Opt. Lett.* **24**: 939 (1999).

39. M. Shtaif, A. Mecozzi, and J. A. Nagel, Mean square magnitude of all orders of polarization mode dispersion and the relation with the bandwidth of the principal states, *IEEE Photon. Technol. Lett.* **12**: 53 (2000).

40. M. Shtaif and A. Mecozzi, Study of the frequency autocorrelation of the differential group delay in fibers with polarization mode dispersion, *Opt. Lett.* **25**: 707 (2000).

41. B. Huttner, C. Geiser, and N. Gisin, Polarization-induced distortions in optical fiber networks with polarization mode dispersion, *IEEE J. Select. Top. Quant. Electron.* **6**: 317 (2000).

42. E. Lichtman, Limitations imposed by polarization dependent gain and loss on all-optical ultra-long communications systems, *J. Lightwave Technol.* **13**: 906–913 (1995).

43. M. Shtaif, A. Mecozzi, and R. W. Tkach, Noise enhancement caused by polarization dependent loss and the effect of gain equalizers, *Proc. Optical Fiber Communications Conf.*, Anaheim, CA, 2002, Paper TuL1.

44. C. R. Menyuk, Nonlinear pulse propagation in birefringent optical fibers, *IEEE J. Quant. Electron.* **23**: 174 (1987).

45. S. G. Evangelides, L. F. Mollenauer, J. P. Gordon, and N. S. Bergano, Polarization multiplexing with solitons, *J. Lightwave Technol.* **10**: 28 (1992).

46. P. K. Wai, W. L. Kath, C. R. Menyuk, and J. W. Zhang, Nonlinear polarization mode dispersion in optical fibers with randomly varying birefringence, *J. Opt. Soc. Am. B* **14**: 2967–2979 (1997).

MODEMS

RAVI BHAGAVATHULA
HYUCK KWON
Wichita State University
Wichita, Kansas

1. INTRODUCTION

Transmission of data between two devices requires the usage of a transmitter, a receiver, and a transmitting media that provides a path between the transmitter and the receiver. According to the manner in which data are transmitted, there are two fundamental modes of transmission: (1) parallel and (2) serial.

In a parallel mode of transmission, data are transmitted one byte (or character) at a time. This mode of transmission requires a minimum of 8 lines, with additional lines for control signaling, for transmitting the 8 bits (or one byte) of data from the transmitter to the receiver. This

transmission method yields a very high data rate at the expense of increased costs due to the presence of a large number of cables between the communicating devices. Hence, it is used in communication between computers and peripheral units where cable distances are relatively short and data transfers must occur rapidly (such as between a printer and a computer).

The parallel mode of transmission becomes increasingly expensive as the distance between the two communicating devices increases relative to the increase in the cost of the cables. An alternative to parallel transmission is the serial mode of transmission, wherein the data are transmitted in sequence over one line, that is, one bit at a time. Instead of requiring additional lines for control signals, a preset sequence of bits can be used for a similar purpose, allowing a two-wire circuit with one wire serving as an electrical ground to be used for data transmission. Since the public switched telephone network (PSTN) already provides a two-wire facility for voice transmission, it is quite logical for serial transmission to utilize the available infrastructure for a cost-effective transmission mechanism.

To communicate over serial lines, the terminal devices need to convert the parallel data into a serial datastream. A *universal asynchronous receiver/transmitter* (UART) on the terminal device usually handles this, and the resulting serial datastream is transmitted/received using a common serial interface. However, there is a basic incompatibility between the digital signals transmitted by a terminal device and the analog signals transmitted by a PSTN line since the PSTN was originally designed to carry only voice signals. Although digital signals can be transmitted over an analog telephone line, the digital pulse-distorting effects of resistance, inductance, and capacitance on the analog PSTN line limit their transmission distance. Further, the presence of analog amplifiers to boost analog voice signal levels in the PSTN pose additional problems with digital data since the analog amplifier would boost the digital signal along with the distortions and would, therefore, increase the distortion in the digital data transmission.

Because of the incompatibilities between the digital signals produced by terminal devices and the analog signals that telephone lines were designed to carry, a conversion device is required to enable digital signals to be carried on an analog transmission medium. This conversion device is a modem, a contraction of the term modulator–demodulator. The modulator portion of the device converts (or modulates) the digital signals into analog signals for transmission over the PSTN line, while the demodulator portion of the device converts (or demodulates) the analog signal into digital format. Therefore, the modulator portion of the modem can be considered to be the transmission component of a communication system, and the demodulator can be considered to be the receiver component of a communication system.

With the emergence of digital telephony, some portions of the PSTN are designed to carry voice signals in a digital format. Usage of services such as the Integrated Services Digital Network (ISDN) allows PSTN subscribers for end-to-end voice and data communication in digital format. These digital networks use a bipolar signaling scheme for transmission over twisted-pair cable. A digital modem is therefore utilized to convert the unipolar signals generated by the terminal devices to a bipolar format used by the digitized PSTN.

Modems, depending on the type of datastreams they operate on, work in either an asynchronous or a synchronous mode. In an asynchronous mode of operation, often referred to as a *start/stop* transmission, each character is encoded into a series of pulses. The transmission is started by a start pulse followed by the encoded character (a series of pulses). The receiver is notified of the completion of the transmission of a character by the transmission of a stop pulse that may be equal to or longer than the start pulse depending on the transmission code being used.

In a synchronous mode of operation, a group of characters are transmitted in a continuous bitstream. Modems located at each end of the transmission medium normally provide a timing signal or clock to establish the data transmission rate and hence enable the devices attached to the modems to identify the appropriate characters as they are being transmitted or received. Before the data transmission is initiated, the transmitting and the receiving devices must establish synchronization between themselves. To keep the receiving clock in step with the transmitting clock for the duration of a bitstream representing a large number of consecutive characters, the data transmission is preceded by the transmission of a special set of synchronization characters. An error-free data transmission after the synchronization process is achieved by using an error detection scheme known as *cyclic redundancy check* (CRC). This mode of serial data transfer yields a much higher data rate at the expense of complex circuitry because the receiver must remain in phase with the transmitter for the duration of the transmitted group of characters.

2. MODEM OPERATION

A modem connects a PC (or a computing device) with the outside world through a series of cables. The connection between the modem and the PC is usually accomplished by a serial cable (in case of an external modem) or through the system bus itself (in case of an internal modem). The modem speaks with the outside world through a telephone line.

In the common scenario of a PC speaking with the outside world through a modem, the PC is referred to as a DTE (data terminal equipment) while the modem is referred to as a DCE (data communication equipment). An exception is the case of an internal modem wherein the concept of a DTE does not exist since the definitions of DTE and DCE are plausible with respect to a RS-232 connection (and an internal modem does not employ any RS-232-type connections since it is a bus-connected device and not a serial-interface-connected device).

The speeds at which a DTE can transmit information to a modem are usually much larger than the speeds at which

the modem can transmit that information to the outside world. This speed difference warrants the existence of a number of mechanisms to ensure the timely and accurate exchange of information.

In an effort to bridge the communication speed differences, various data compression schemes are employed. The data compression schemes currently in use are the MNP-5 and the V.42bis. MNP-5 is derived from Microcom Network Protocol standard (devised by Microcom, now acquired by Compaq) and yields a data compression rate of up to 2 : 1. However, MNP-5 accumulates a large amount of overhead while compressing data. Further, it is rather inefficient in transmitting precompressed files since it cannot sense the need for compression. Using this scheme with a precompressed file usually results in the transmission of a larger file as compared to the original compressed file.

V.42bis is a data compression standard approved by ITU-T (International Telecommunication Union — Telecommunication Sector) [1], which yields a compression ratio of up to 4 : 1. It builds on the advantages of the MNP-5 protocol and includes the ability to sense when compression is required. This makes it more efficient when transmitting precompressed files.

It should be noted that the two standards, MNP-5 and V.42bis, are both exclusive in nature; that is, they cannot be used at the same time. Further, for optimal performance, the DTE-DCE communication link (usually the terminal and the modem connection) should be able to sustain the data rate that these compression schemes afford. In the case of the MNP-5, since the compression rate is 2 : 1, for every bit that the modem sends out, the DTE is required to transmit 2 bits' worth of information. Therefore, the bit rate of the DTE should be 2 times that of the DCE. In the case of V.42bis, the DTE speed should be 4 times that of the DCE. For example, if your modem is operating at 14,400 bps and no data compression schemes are being used, the DTE speed would need to be set at 14,400 bps (bits per second). However, if MNP-5 were to be used, the DTE speed would need to be set at 28,000 bps (2 × DCE speed) since the MNP-5 supports a compression ratio of 2 : 1. In the case of V.42bis, the DTE speed would need to be 57,600 bps (4 ∗ DCE speed) as V.42bis supports a compression ratio of 4 : 1.

A convenient way in which these speed differences are quoted in modem terminology is through the concept of a baud. A *baud* is defined as the rate at which information is transmitted. In contrast, *bit rate* is defined as the rate at which bits are transmitted. Put another way, a baud could be defined as one pulse (or signal) interval in a carrier signal while bit rate is the number of bits that are transmitted per second (or signal). The relation between baud and bit rate is given below:

$$\text{Bit rate (in bits per second)} = \text{baud} \times \text{bits per baud}$$

With increasing transmitting speeds, the emphasis on better error control mechanisms has been increasing for the accurate exchange of information. Two distinct error control schemes that are used in most modern modems are MNP-4 and V.42. MNP-4 includes the functionalities of MNP-2 and MNP-3 while V.42 employs LAP-M (Link Access Protocol — Modem) protocol as the primary error control mechanism and reverts to MNP-4 as a backup. These error control schemes retransmit corrupted data using 16–32-bit CRCs. V.42 is an ITU-T specification that yields slightly better performance than MNP-4. In V.42, the primary error control mechanism is LAP-M. Data are grouped together in terms of frames, and these frames are transmitted over the communication channel along with a CRC header for error control. In case of LAP-M, each frame has a size of 128 bytes, and up to 15 frames (by default) can be sent without waiting for an acknowledgment from the receiver. This translates to a storage requirement of $128 \times 15 = 1920$ bytes at the transmitting end for accomplishing error-free transmission since the transmitter would need to retransmit all 15 frames if the transmitter receives a negative acknowledgment from the receiver. V.42 employs 16- or 32-bit CRC fields (although 32-bit CRC is more common these days.)

Because of the presence of many operating speeds, it is quite possible that the operating speed of one modem may be more than that of another modem involved in a typical end-to-end connection between two communicating terminals. In such cases, the receiving modem needs to be able to inform the transmitting modem to pause before it can fully process the data it received. This is accomplished using flow control mechanisms. Flow control mechanisms can be broadly classified as either software-based or hardware-based.

In *software-based* flow control mechanisms (also referred to as XON/XOFF mechanisms), the receiver modem sends a special signal (usually a Control-S character) to the transmitting end requesting the transmitter modem to pause for a while. The transmitter modem stops sending any new information to the receiver modem until it receives another special signal from the receiver modem (which is usually a Control-Q character) informing the transmitter modem that it can resume sending data. The advantage of this scheme is that no additional hardware support is required since the pause/resume signals are handled by the communication software (or the firmware in the modem). The disadvantage of this scheme is that the presence of noise in the transmission media can result in the loss of the pause/resume signals and therefore affect the operation of the modems. If the pause signal were lost, the transmitter modem would keep transmitting data at a rate that the receiver modem cannot handle, resulting in overruns at the receiving modem. If the resume signal is lost, the transmitter modem would never know when to resume transmission of data and therefore the transmission media would remain silent forever. Because of the transmission of special characters to signify pause/resume actions during transmission, XON/XOFF flow control mechanism should not be used for the transfer of binary data since the modems could falsely interpret the presence of Control-S character in the original binary file as a signal to pause data transmission.

In sharp contrast to software-based flow control schemes, *hardware-based* flow control schemes depend on special hardware support for ensuring proper control over

the flow of data. This is also referred to as RTS/CTS (ready to send/clear to send) flow control mechanism. In the case of external modems, a specific wire in the serial cable (that is used to connect the terminal to the modem) is used for exchanging flow control information. In internal modems, in the absence of a serial cable, flow control functionality is built into the modem itself.

Data transfer (more commonly known as *file transfer*) protocols exist for the transmission of binary data between two modems over a communication channel (which is usually a telephone line). Commonly used data transfer protocols include Xmodem, Xmodem-CRC, Xmodem-1K, Ymodem, and Zmodem. In the *Xmodem* transfer protocol, binary data are transmitted in 128-byte chunks and a checksum is appended to these 128-byte blocks so that the receiver can verify the integrity of the received block and intimate the transmitter of the status of its reception. If the receiver determines any errors in the reception, the transmitter modem would resend the entire 128-byte block.

Xmodem-CRC adds CRC functionality to the basic Xmodem protocol, while Xmodem-1K transfers data in terms of 1-kbyte data chunks as against the standard 128-byte data chunks used by the standard Xmodem protocol. The *Ymodem* protocol is quite similar to the *Xmodem-1K* protocol and is seldom used on noisy communication channels (like telephone lines) due to its inability to perform efficiently in such environments. Ymodem-G, an improvement on the original Ymodem protocol, yields slightly faster data transfer rates by eliminating software error control mechanisms and relying on the underlying hardware to perform the required error control operations.

Zmodem is the most widely used data transfer protocol over dialup connections because of its improved resilience to noisy environments and the higher data rates. It employs 32-bit CRC fields and does not wait for an acknowledgment from the receiver before it transmits the next block of data.

3. MODULATION TECHNIQUES AND MODEM STANDARDS

Because of the comparatively smaller bandwidths that are offered by present-day dialup connections, modems employ a modulation scheme to transmit more information at the same bit rate. This is accomplished by converting data into *symbols* and transmitting the symbols over the communication channel. The conversion of a datastream into a symbol stream is carried out by using an appropriate modulation scheme. The usage of these modulation schemes leads to a much better utilization of bandwidth because more information (in terms of data bits) can be inserted into specific *symbols* that are transmitted over the communication channel.

The most common modulation scheme is AM (amplitude modulation) [2]. This forms the basis for a few more advanced modulation schemes like QAM (quadrature amplitude modulation) [2]. In AM, symbols are defined in terms of the amplitude of the original signal and these symbols are transmitted over a carrier through the analog communication channel. A major disadvantage of AM is the fact that as the amplitude of the signal decreases, it becomes increasingly difficult to separate the signal from noise in the communication channel.

QAM is an improvement over AM in which information is encoded on the basis of the deviations in the phase and amplitude of the carrier wave. This so-called *two-dimensional* encoding leads to a greater encoding efficiency and, therefore, a higher data rate.

TCM (trellis-coded modulation) [3] is based on the QAM scheme. In TCM, additional bits are added to each *symbol* to accomplish *forward correction*. This leads to a better error control ability and bit errors introduced into the communication process can be effectively reduced.

FM (frequency modulation) [2] is the frequency counterpart for the AM scheme wherein the modulation is accomplished in terms of the frequency rather than the amplitude. While this modulation scheme is used more widely for radio broadcasts, its application in dialup connections is not widespread. A variation of FM modulation is FSK (frequency shift keying), which was designed primarily for transmission of data across a telephone line. In this scheme, the presence of bit "1" is represented by a specific frequency tone and the presence of bit "0" is represented by another specific frequency tone. To afford two-way communication, FSK allows the specification of two different sets of frequency tones.

PSK (phase shift keying) is similar in principle to FSK, with the sole difference that variations in phase of a constant frequency carrier are used to determine the bit values in the original datastream. A signal with unchanged phase is used to signify the presence of a bit whose value is the same as that of the previous bit. A 50% change in the phase is used to signify a bit value that is different from the previous bit. Differential PSK (DPSK) is a refinement of PSK wherein the changes in phase are determined by comparing the phase of the current state with that of its previous state.

PCM (pulse code modulation) is a modulation scheme wherein analog data are encoded into a specific number of bits (usually 8 bits) for transmission over a communication channel. This is, strictly speaking, not a true modulation scheme since a carrier is not employed at all. The encoding of analog information into digital format is accomplished using a quantizer and a sample-and-hold circuit.

Several modem standards have been introduced that allow modems to exchange data universally. While it is not possible to discuss every modem standard in this document, a few representative standards are discussed here.

Bell 103 is one of the older standards that allow data to be transmitted and received at a rate of 300 bps (bits per second). It employs FSK modulation and uses 1 bit to represent a baud (i.e., bit rate is equal to baud rate). Bell 202 is considered an improvement over the Bell 103 as it supports a 1200 bps data rate using the same FSK modulation technique as employed by Bell 103.

With the widespread usage of modems across the globe, the need arose for a set of modem standards that could facilitate modems throughout the world to communicate with each other. CCITT (later known as *ITU-T*) undertook to formulate a set of universally applicable standards to

this effect. One of the earlier CCITT standards for data communication was the CCITT V.21 [4], which allowed a data rate of 300 bps using FSK modulation (quite similar in operation to the Bell 103 standard). CCITT V.22 used DPSK to obtain data rates of 1200 bps at a baud rate of 600 baud (i.e., the number of bits per baud was set to be equal to 2). This standard is similar to the Bell 212A.

More advanced standards were later released by CCITT (viz., V.32 and V.32bis [5]) that allowed data rates of up to 14,400 bps using different modulation techniques such as TCM and QAM. ITU-T V.34 [6] is a more recently established standard that allows modems to transfer data at rates up to 33,600 bps.

V.34 is currently the fastest end-to-end analog modem standard. Because of the dependence of the more advanced standards such as V.90 on the V.34 standard, let us take a closer look at the details of the V.34 analog telephony standard.

Modems supporting V.34 can sustain data transmission capacities of 2400–28,800 bps. A feature referred to as *line probing* was introduced in this standard to allow modems to identify the capacities and quality of the phone landline and adjust themselves to allow, for each individual connection, the most optimal data transmission rate. V.34 also supports a synchronous auxiliary channel with a data rate of 200 bps that could be used in tandem with the primary data channel for signaling information.

The bandwidth offered by a phone line is around 3–4 kHz, and the maximum symbol rate that is supported by V.34 is 3429 symbols per second. The operation of V.34 near the theoretical limits of the phone line spurred the design engineers to incorporate a mechanism within V.34 to autonegotiate the available bandwidth on a phone line and adjust the data transmission rates accordingly. A new handshake protocol, called as V.8 mode negotiation handshake, was introduced to enable two V.34 compatible modems to exchange feature and mode negotiation information via V.21 standards (300 bps FSK modulated communication). V.8 mode is used by the two V.34 modems to identify them with other telephone network equipment. It is also used to determine whether the call is destined for a data or facsimile operation. Negotiation of the available modes of modulation is also accomplished along with ability to support V.42 and V.42bis standards [1]. During this handshake, the modems send a series of tones to each other, at specific frequencies and known signal levels. The received signal level is employed in the computation of the maximum possible available bandwidth for communication.

Line probing is employed immediately after V.8 handshake to determine parameters such as the optimal bandwidth and carrier frequency, preemphasis filters, and optimal output power level to be used during communication. Table 1 lists the symbol rates, carrier frequencies, and supported data rates as defined in V.34 that are implemented in the 3Com OfficeConnect series of modems. For every symbol rate [except 3429 symbols per second (sps)], the modem can select one of the two available carrier frequencies.

A preemphasis filter is usually employed in a modem to remove amplitude distortions that could creep into a phone line, by suitably shaping the transmitted signal's spectrum. V.34 supports 10 preemphasis filters, and the appropriate filter is selected during the line probe operation.

Table 1. V.34-Supported Symbol Rates, Carrier Frequencies, and Bandwidths (Compatible with 3Com OfficeConnect Series of Modems)

Symbol Rate	Minimum Bit Rate	Maximum Bit Rate	Carrier Frequency	Required Bandwidth
2400	2400	21,600	1600	400–2800
			1800	600–3000
2743	4800	24,000	1646	274–3018
			1829	457–3200
2800	4800	24,000	1680	280–3080
			1867	467–3267
3000	4800	26,400	1800	300–3300
			2000	500–3500
3200	4800	28,800	1829	229–3429
			1920	320–3520
3429	4800	28,800	1959	244–3674

Compared to the 2-D TCM schemes employed by V.32bis, V.34 lets the modems select any one of the three available 4-D TCM schemes. This allows for a more robust error-correction scheme for accurate data transmission.

When two modems lose synchronization with each other, a feature called a retrain is activated. During retraining, the modems do not send any data across the network since they suspend all operations and re-negotiate the connection once again. Unlike the earlier standards, retraining in V.34 is accomplished using the receiver modem's timer. This leads to a reduced (and variable) retrain time as compared to a large, fixed retrain time for the earlier standards.

With the emergence of faster terminals, the need arose for standards that would enable PCs to communicate more rapidly than the allowed 33,600 bps (as with ITU-T V.34). Two contemporary standards evolved to meet these requirements: the Rockwell/Lucent K56 standard and the USR X2 standard.

Traditionally, communication between modems is carried out on the assumption that both ends of a modem conversation have an analog connection to the telephone network. Therefore, data from a terminal are converted from digital format to analog format and transmitted over the PSTN (at speeds of up to 33,600 bps as supported by V.34). At the receiving end, a modem translates the analog signals into digital data that the receiving terminal can process.

In contrast, the Rockwell/Lucent K56 and USR X2 protocols assume that one end of the modem conversation has a digital connection to the telephone network (usually through a digital modem like ISDN). Since Internet service providers (ISPs) usually transmit data in digital format across networks, the digital end of the modem conversation is typically that of an ISP.

Since one end of the modem conversation is digital in nature, the downstream traffic (traffic from the ISP to a user's modem) is digitally modulated using PCM yielding data rates of up to 56,000 bps. Upstream traffic (traffic

from a user's modem to the ISP) still proceeds at a rate of 33,600 bps (using V.34 standards). Therefore, the K56 and the X2 standards are asymmetric in nature since they offer different data rates for different directions of data transfer.

Because of the incompatibility of the K56 and the X2 standards, ITU-T introduced a universally applicable standard called V.90 [7]. It incorporates the features of the K56 and the X2 standards and enables modems supporting V.90 to be universally compatible with the K56 and the X2 standards. The downstream modulation scheme is PCM and the upstream modulation scheme is carried out using V.34 standards. While these asymmetric data rates might not look all that lucrative, they provide the general feeling of a faster communication channel since most users deal with more downstream traffic than upstream traffic.

4. FAX TRANSMISSION

Facsimile (or fax) is defined as the process of sending a document from one terminal to another. In an effort to utilize the existing PSTN infrastructure, traditional fax transmission involved the usage of modems at either communicating terminals. The transmission of a document is preceded by an exchange of capabilities between the sender and the receiver and, at the end of the transmission, a confirmation of delivery sent from the receiver to the sender.

The earlier fax machines, also referred to as group 1 fax machines, were designed to handle fax transmission over an analog telephone network. These conformed to ITU-T T.2 standard for fax transmission and yielded a transmission rate of 6 min per page. With improvements in fax devices, ITU-T later introduced the T.3 standard that allowed for the transmission of up to 3 min per page. Group 3 fax machines, the most widely deployed category of fax machines, were standardized in 1980 for digital facsimile devices to communicate over analog telephone lines. Group 3 machines are based on ITU-T standards T.30 (for fax transmission) and T.4 (for fax file formats) [8] and yield a transmission rate of 6 to 30 seconds per page.

The procedures outlined in the T.30 recommendation comprise of 5 distinct phases: (1) call establishment, (2) control and capabilities exchange, (3) in-message processing and message transmission, (4) postmessage, processing, and (5) call release.

The call establishment phase consists primarily of establishing a connection between the calling and the called terminals and exchanging fax tones. The calling machine dials the telephone number of the called machine and the calling tone (referred to as CNG) is received at the called machine. The CNG tone beeps indicate the existence of a fax call as against a normal voice call. The called fax machine answers the ring signal by going off-hook. After a 1-s delay, the called fax machine sends a 3-s, 2100-Hz tone back to the calling machine.

In the premessage processing phase, the terminals carry out various identification procedures along with command procedures to establish a command set of capabilities for the successful transmission of facsimile data. During the identification phase, the terminals exchange information regarding, among others, the bit rate, page length, data compression format, telephone number, and name of the organization. The called machine sends its digital identification signal (DIS) at 300 bps identifying its capabilities (using V.21 protocol), including its optional features. For example, the called fax machine could send a DIS identifying its capability to support V.17 standard (14,400 bps data rate). On receipt of the DIS, the calling fax machine sends a digital command signal (DCS) locking the called unit into the selected capabilities. The calling machine sends a training check field (TCF) through the modem to ensure that the channel is suitable for transmission at the accepted data rate. The called fax machine sends a "confirmation to receive" (CTR) signal to confirm that the receiving modem is trained (adjusted for low-error operation).

The in-message processing phase takes place in parallel with the message transmission phase since the in-message processing phase handles the signaling required for the transmission of facsimile data. This includes control signals needed for error detection, error correction, and line supervision. The ITU-T T.4 recommendation governs the message transmission phase and addresses issues related to the dimension of the document, the transmission time per scanned line, the coding scheme, modulation/demodulation techniques, and similar. The modem standards that are supported for the transfer of facsimile data include V.27ter (4800/2400 bps), V.29 (9600/7200 bps), V.33 (14,400/12,000 bps), and V.17 (14,400/12,000/9600/7200 bps).

The postmessage processing phase consists of procedures to deal with tasks such as end-of-message signaling, multipage signaling, and confirmation signaling and is used as a precursor to the call release phase. The calling fax machine sends a "return to control" (RTC) command that effectively switches both modems to the 300 bps data rate condition (V.21 standard). The called fax machine sends a message confirmation (MCF) signal indicating the document was received successfully. If multiple pages exist, a multipage signal (MPS) is sent. The partial page signal (PPS) is sent for error correction of the transferred document. In the call release phase, the calling terminal transmits a disconnect (DCN) signal to the called terminal for the release of the call. It is important to note that no response is expected for the call release signal from the called terminal.

5. OTHER MODEMS

The typical bandwidth allocated to each individual user on a telephone line (also referred to as a *subscriber line*) is around 3400 Hz. This places an upper limit on the data rates that a voice band modem can achieve since the transmitting/receiving symbol rate cannot be higher than the available bandwidth. A digital subscriber line (DSL) overcomes this limitation by overlaying a data network onto the existing PSTN. This is accomplished by letting the data network use the same subscriber lines as the POTS (Plain Old Telephone System), with the only exception that the data signals use a different frequency band for data communication.

A device called a POTS splitter is responsible for splitting and recombining the two types of signals: voice signals and data signals, at both ends of the subscriber line. Since the data network and the voice network use the same subscriber line, the telephone companies need only upgrade their switching/terminal devices to handle the data signals. Therefore, the delivery of high-speed data services to customers without considerable investment in infrastructure is possible.

Depending on the data rates that are supported, there are different variations to DSL [9]. ADSL (asymmetric DSL) is characterized by a different data rate from the service provider to the customer (the downstream direction) as compared to the data rate from the customer to the service provider (the upstream direction). The upstream data rates are typically 10 times slower than the downstream data rates and range from 100 to 800 kbps. In sharp contrast, SDSL (symmetric DSL) offers a symmetric data rate; that is, the upstream and downstream data rates are equal. VDSL (very-high-speed DSL) is a new addition to the DSL family and is being developed to provide data rates as high as 25 Mbps in either direction (upstream or downstream).

In ADSL, the transceivers (transmitter and receiver units) are designed to carry more than one logical channel on a single physical channel to support different data rates. In addition, the transceivers support an embedded operations channel (EOC), ADSL overhead channel (AOC) etc that are used mostly for synchronization purposes. The logical data channels on an ADSL link are grouped together as either downstream channels or upstream channels. The downstream channels are simplex in nature and are designated AS0, AS1, AS2, and AS3. Each channel has an allowable data rate up to 8, 4, 3, and 1.5 Mbps, respectively. The duplex channels are named LS0, LS1, and LS3. These support different data rates in upstream and downstream directions and are usually configured as upstream channels by setting the downstream channel data rate equal to 0.

The presence of many different logical channels enables ADSL to support a wide variety of applications since different channels operate at different data rates. However, the total physical bandwidth available in the channel should be greater than the sum of bandwidths of all the logical channels (since a portion of the bandwidth will be consumed by control/synchronization signals).

Since different carrier frequencies are employed for traditional voice transmission, upstream data and downstream data, ADSL uses frequency-division multiplexing (FDM) to multiplex these different signals onto one physical channel. Voice transmission is carried out by letting POTS use the lower-frequency band (0–3400 Hz), while the downstream data is transferred in the higher-frequency band (138 kHz–1.104 MHz). A guard band of approximately 26 kHz is placed between the POTS band and the upstream band to reduce the possibility of interference between voice conversations and data transmission operations.

In sharp contrast to voice band or DSL technologies, cable modems capitalize on the existence of a network of coaxial cables (that are used primarily for video applications) to transmit data. Since the coaxial cables support high bandwidths, cable modems could be used for high-speed data transfers. This is a prime reason for cable technology offering formidable competition to DSL.

Cable systems, however, have traditionally supported only downstream traffic and the available bandwidth is shared among several users since a single line serves many cable subscribers. Therefore, the cable system had to be reconditioned to support upstream traffic from the subscriber back to the coaxial distribution point.

A typical cable system consists of an uplink site that encodes and modulates video content obtained from various sources (tapes, DVDs, live feeds, etc.). The modulated video content is transmitted to downlink site via a satellite. The video content is transmitted from the downlink site to the cable subscribers through a network of coaxial cables. This transfer of video information from the downlink site to the cable subscribers is accomplished with the aid of headend (HE). An HE is responsible for modulating and scrambling each channel that is supported on the cable system and transmitting the scrambled information onto the local hybrid fiber coaxial cable) (HFC) distribution box. Subscribers connect to the HFC through a series of local taps (where each HFC could service 500–2000 homes per fiber node). Data communication is enabled into the cable system by incorporating routing/switching functionality in the HE. This is accomplished by using a broadband router (like a Cisco uBR 7200 series router) for data connectivity. At the customer premise, an additional device would be needed to allow users to utilize the data communication facilities that are available on the same coaxial line that delivers audiovideo content. This additional device is referred to as a *cable modem* (CM).

The downstream communication (from the cable company to the subscriber) is carried out in the frequency range of 54–860 MHz. The upstream communication (from the subscriber to the cable company) is carried out in the frequency range of 5–42 MHz. The manner in which data are transferred over a cable network is specified by DOCSIS (data over cable service interface specification) [10]. Typically, a CMTS (cable modem termination system) is employed at the HE to modulate (or demodulate) the signals sent (or received) from the CM. The CM is associated with the customer premise equipment (CPE) and is responsible for communicating with the CMTS for data communication. CMs support DOCSIS defined connectors, namely, RJ45 ports for Ethernet, RJ11 ports for voice, and F-connector for video.

DOCSIS-compliant devices require the presence of a few servers that provide information regarding IP addresses through the dynamic host configuration protocol (DHCP), time of day timestamps (as defined in RFC 868) and CM configuration files through TFTP (Trivial File Transfer Protocol).

When a CM is powered up, it scans the downstream channel for a synchronizing clock signal. The HE continues broadcasting information regarding upstream channel descriptors, upstream channel frequencies, downstream channel descriptors, and other data that the CM can use for determining the details of the channels it can use

for upstream and downstream communication. Once the CM recognizes the upstream and downstream channels, it begins identifying the bandwidth that can be used for communication.

At this stage, the CM–HFC interface line protocol is considered to be up but the CM is not yet ready to start transferring data with other hosts on the global Internet since it does not yet have an IP address. A DHCP server provides the CM with an IP address, a default gateway, the address of the TFTP server, the CM configuration filename, and the address of a time-of-day (ToD) server that the CM can use for synchronizing its operations. The CM uses the address of the TFTP server to obtain the required files to configure itself as a network entity to handle data transfers across the cable network.

BIOGRAPHIES

Hyuck M. Kwon was born in Korea on May 9, 1953. He received his B.S. and M.S. degrees in electrical engineering (EE) from Seoul National University, Seoul, Korea, in 1978 and 1980, respectively, and his Ph.D. degree in computer, information, and control engineering from the University of Michigan at Ann Arbor, in 1984. From 1985 to 1989 he was with the University of Wisconsin, Milwaukee, as an assistant professor in EE and CS department. From 1989 to 1993 he was with the Lockheed Engineering and Sciences Company, Houston, Taxes, as a principal engineer, working for NASA space shuttle and space station satellite communication systems. Since 1993, he has been with the ECE department, Wichita State University, Kansas, where he is now a full professor. In addition, he held several visiting and consulting positions at communication system industries, and a visiting associate professor position at Texas A&M University, College Station, in 1997. His current research interests are in wireless, CDMA spread spectrum, smart antenna, space-time block code, and MIMO communication systems.

Ravi Bhagavathula received his B.E. degree in electronics and communication engineering in 1997 from Osmania University, Hyderabad, India, and his M.S degree in Electrical Engineering from the Wichita State University in 1998. He is currently working towards a Ph.D. degree in electrical engineering from Wichita State University. He was the recipient of the 2002 Electrical Engineering Outstanding Ph.D. student award. His areas of interest are memory hierarchy design, cache block replacement algorithms, router architectures, layer 3 mobility, and mobile IP extensions using MPLS and VPN architectures.

BIBLIOGRAPHY

1. Recommendation V.42bis, *Data Compression Procedures for Data Circuit Terminating Equipment (DCE) Using Error-Correcting Procedures*, ITU-T *(http://www.itu.int)*.
2. H. Taub and D. L. Schilling, *Principles of Communication Systems*, 2nd ed., McGraw-Hill, New York, 1996.
3. E. Biglieri, D. Divsalar, P. J. Mclane, and M. K. Simon, *Introduction to Trellis-Coded Modulation with Applications*, Macmillan, New York, 1991.
4. Recommendation V.21 (11/88)—*300 Bits per Second Duplex Modem Standardized for Use in the General Switched Telephone Network*, ITU-T *(http://www.int.int)*.
5. Recommendation V.32 (03/93), *A Family of 2-Wire, Duplex Modems Operating at Data Signalling Rates of up to 9600 Bit/s for Use on the General Switched Telephone Network and on Leased Telephone-Type Circuits*, ITU-T *(http://www.int.int)*.
6. Recommendation V.34 (02/98), *A Modem Operating at Data Signalling Rates of up to 33 600 Bit/s for Use on the General Switched Telephone Network and on Leased Point-to-Point 2-Wire Telephone-Type Circuits*, ITU-T *(http://www.itu.int)*.
7. Recommendation V.90 (09/98), *A Digital Modem and Analogue Modem Pair for Use on the Public Switched Telephone Network (PSTN) at Data Signalling Rates of up to 56,000 Bit/s Downstream and up to 33 600 Bit/s Upstream*, ITU-T *(http://www.itu.int)*.
8. Recommendation T.4 (04/99), *Standardization of Group 3 Facsimile Terminals for Document Transmission*, ITU-T *(http://www.itu.int)*.
9. D. J. Raushmayer, *ADSL/VDSL Principles*, Macmillan Technical Publishing, 1999.
10. G. Abe and A. Buckley, *Residential Broadband*, Cisco Press, Indianapolis, 1999.

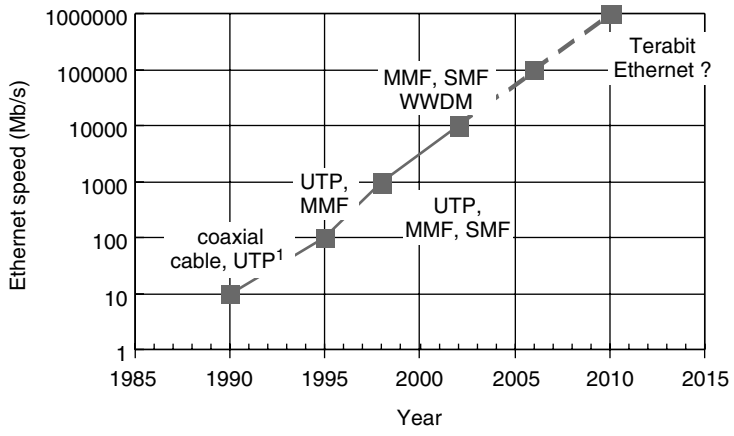
MODERN ETHERNET TECHNOLOGIES

CEDRIC F. LAM
Opvista Inc.
Irvine, California

1. INTRODUCTION

Ethernet was invented in 1973 at Xerox Labs in Palo Alto, California as a medium access control (MAC) protocol for local-area networks (LANs) [1]. Since its invention, Ethernet has gone through many changes in performance, architecture and the underlying technologies. The rapid adoption of the Internet since the early 1990s has made Ethernet the most popular network technology with very fast growth rate. Ethernet has become the ubiquitous means to network servers and desktop computers. Over 85% of the network traffic in today's Internet is generated as Ethernet packets. Not only has Ethernet been enjoying popularity in network computing; it is also becoming more and more popular for internetworking automated manufacturing systems and measurement equipment in factories and research labs. Figure 1 shows the development trend of Ethernet since 1982.

Ethernet has become the most popular network technology among many different competing technologies in the Internet era because of its low cost and simplicity. A 10/100BASE-T Ethernet Network Interface Card (NIC) costs as little as $15 nowadays. The cost for a 1000BASE-T (more commonly known as Gigabit Ethernet) Ethernet port [with 1000 Mbps (megabits per second) throughput] is around $350. Compared to Gigabit Ethernet, a SONET OC-12 intermediate reach interface with 622 Mbps throughput would cost about $3000.



UTP: unshielded twisted pair MMF: multi-mode fiber
SMF: single-mode fiber WWDM: wide wavelength division multiplexing
1. 10 Mb Ethernet using coaxial cables was developed in the 1980s. Ethernet becomes very popular after 10base-T was invented in 1990.

Figure 1. Development trend of Ethernet technology.

2. ETHERNET ARCHITECTURE

IEEE (Institute of Electrical and Electronics Engineers) 802.3 Standard Group charters the development and standardization of the Ethernet technology. The scope of Ethernet covers the physical layer and the data-link layer (layers 1 and 2) of the seven-layer OSI (Open System Interface) reference model. Thus, Ethernet consists of two major layers: (1) the MAC layer, which handles physical-layer-independent medium access control and (2) the PHY (physical) layer, which deals with different physical-layer technologies and various transmission media. In modern Ethernet, the MAC layer and the PHY layer are interconnected with a medium-independent interface (MII). This allows the same MAC design to be used with different transmission technologies. Figure 2 shows the architecture of Ethernet as defined in IEEE802.3 standard [2].

The design of Ethernet itself follows the layered architecture principle. So both the MAC layer and the PHY layer are further divided into sublayers with clearly defined functions.

2.1. Ethernet MAC Frame

Ethernet is a packet-switched technology. Ethernet data are transmitted in packets called *MAC frames*. We may use the terms "frame" and "packet" interchangeably in the following discussions. The format of an Ethernet MAC frame is shown in Fig. 3.

Even though Ethernet has gone through many different generations, the format of the Ethernet MAC frame has never changed. This invariant MAC frame defines the modern Ethernet. By keeping the MAC frame invariant, investment in the upper-layer software can be preserved as the network technology advances. This has tremendous impact on the success of Ethernet technology.

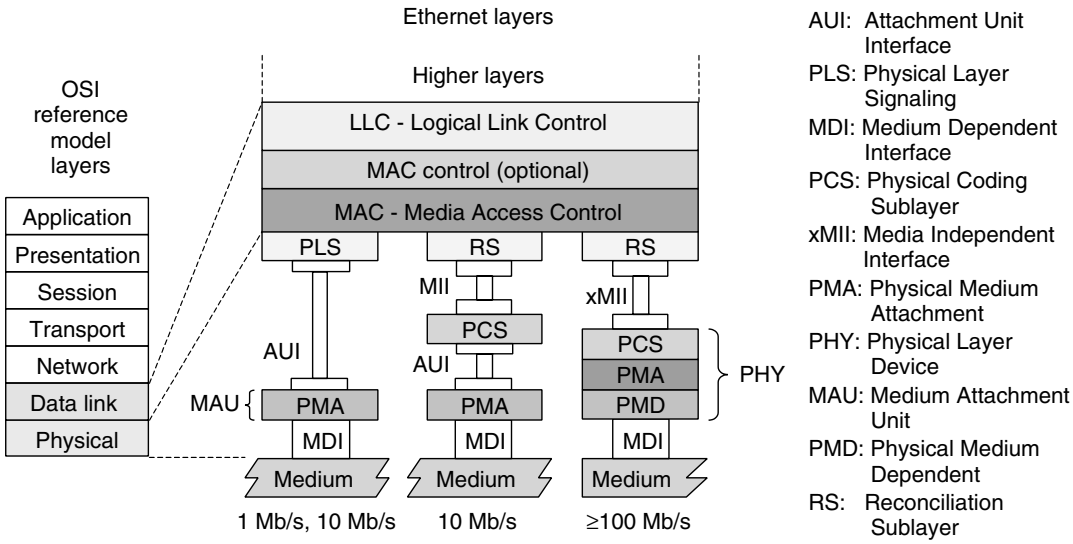


Figure 2. Architecture of Ethernet.

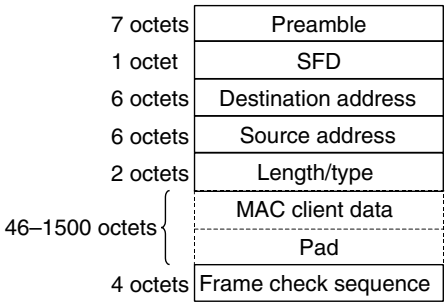


Figure 3. Ethernet MAC frame format.

Ethernet MAC frames have a very simple format with only eight fields (Fig. 3). The leading seven octets in a MAC frame are the preamble field with alternating 0s and 1s for clock recovery purposes at the receiver. This field was useful in early-generation Ethernet, where all the network stations share the same physical channel and data are transmitted in a "bursty" mode. We will see later that the importance of the preamble field diminishes in modern Ethernet, where all stations are joined together with point-to-point dedicated links.

Following the preamble is a one-octet *Start Frame Delimiter* (SLD) field, with the special bit pattern (10101011) to signify the beginning of the actual packet data in the next octet. The next two fields are the Destination and Source Addresses of the MAC frame. Each of them is six octets long. The field following the source address is the two-octet Length/Type field. Ethernet frames are variable length frames with a minimum size of 64 octets and maximum size of 1516 bytes.[1] This field is used to represent the Length of the payload data (from 1 to 1500 bytes) or the type of the MAC frame when its value lies in the range from 1536 to 65535 (2^{16}-1). The payload field follows the Length/Type field and has a size between 46 and 1500 octets. When the actual payload data is smaller than 46 octets, the payload field is padded with zeros to 46 bytes so that a minimum MAC frame of 64-octets is guaranteed.

The last field in the Ethernet MAC frame is a four-octet CRC (Cyclic Redundancy Check) field called frame-check sequence (FCS). Gigabit Ethernet frames may also have a Carrier Extension field following the FCS to extend the size of a packet to a minimum size of 512 bytes, so that the CSMA/CD MAC protocol (which will be explained later) can be supported with a reasonable distance at 1000 Mbps speed.

The simple MAC frame format is the key to the success of Ethernet because it simplifies MAC processing and makes Ethernet devices very cost-effective. On the other hand, Ethernet frames lack the overhead for network management, performance monitoring, fault detection, and localization. This makes large-scale deployment of native Ethernet services a challenging job.

[1] The size of an Ethernet frame does not include the Preamble field and the SFD field.

2.2. Ethernet Address Format

Ethernet uses six-octet-long addresses. The first bit in an Ethernet represents whether the address is a multicast (1) or unicast (0) address. The second bit indicates whether the address is globally administered (0) or locally administered (1). This give a total of 2^{47}-1 globally administered addresses and 2^{47}-1 locally administered addresses.

Ethernet address space is large enough that virtually every Ethernet device in the world can be assigned a globally unique address at the factory. IEEE is in charge of assigning blocks of globally administered addresses to manufacturers of Ethernet interfaces so that each globally administered address is unique in the whole universe. This has significant network implications: (1) there is no need to program the Ethernet MAC address after manufacturing — this reduces the possibilities of human errors; and (2) there is no need for address translation when Ethernet frames are forwarded from one subnetwork to another subnetwork. In some other technologies with very small network address space, the physical addresses of network interfaces in different subnetworks may have the same value. This adds burden to the internetworking devices called bridges (or switches) because address translation must be performed when forwarding data from one subnetwork to another subnetwork, incurring both performance and cost penalties.

2.3. Shared Ethernet: The CSMA/CD Protocol

Ethernet was invented as a medium access control (MAC) protocol for local-area networks (LANs). The first-generation Ethernet adopted a bus architecture with all the network stations sharing a common communication channel as shown in Fig. 4.

Channel arbitration is achieved by the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol [3]. In the CSMA/CD protocol, each station having packets to send first listens to the channel (carrier sense). If the channel is busy, it will wait until the channel becomes idle and clear. If the channel is clear, the station will send the data. However, it is possible for two stations to both sense the channel as clear and try to send data at the same time. In this case, a collision occurs and the stations that are transmitting will stop data transmission and backoff for a random amount of time before re-transmission (collision detection). The stations detecting a collision will also transmit a jamming signal for a certain period of time to ensure that all the stations in the network detect the collision and refrain from transmission.

The CSMA/CD protocol is an unacknowledged protocol; thus, if a station finishes the transmission of a MAC frame before it detects collision, it assumes that that frame is correctly transmitted. In the worst case as illustrated in Fig. 5, station (node) *A* at one end of the bus sends a frame on the bus. Another station, node *B*, at the other end of bus senses the channel as being idle just before the frame from *A* arrives at station *B*. So *A*'s frame will collide with *B*'s frame. By the time the collision propagates back to *A*, a round-trip propagating time has already elapsed since *A* started sending its frame. If *A* finishes transmitting its frame before the

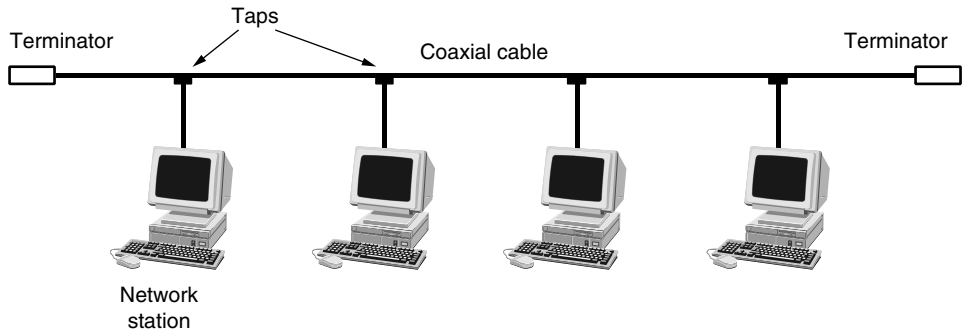


Figure 4. First-generation Ethernet using a shared bus architecture.

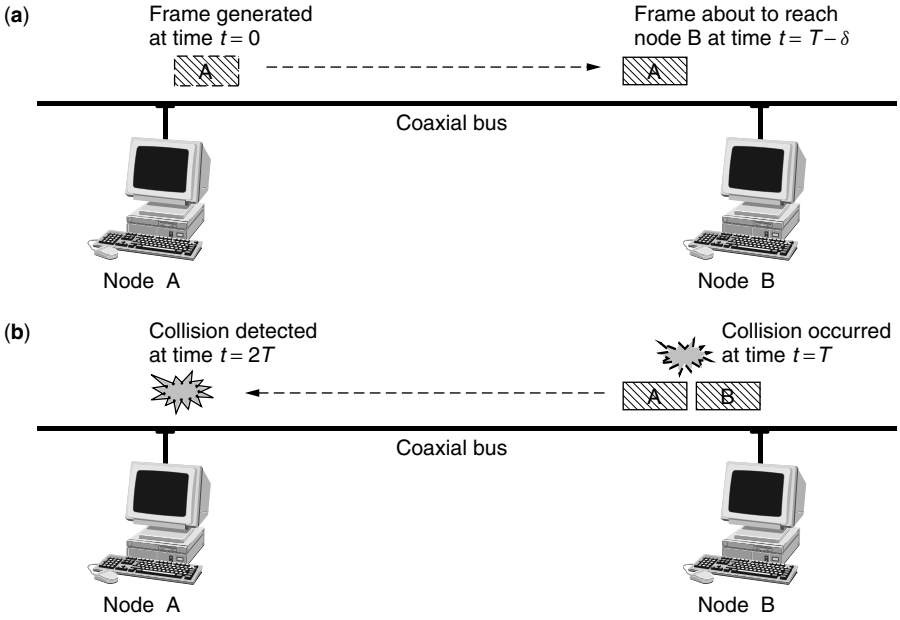


Figure 5. Collisions in a CSMA/CD protocol network.

collision propagates back to A, it will mistakenly think that its transmission was successful. Therefore the minimum packet length, the maximum network size and the transmission speeds are tightly coupled in the CSMA/CD protocol. Ethernet has chosen a fixed minimum packet size of 64 octets. As can be expected, as the transmission speed increases, the network size has to scale inversely for the CSMA/CD protocol to function properly. Thus the CSMA/CD-protocol-limited transmission distance is 2500 m and 250 m for 10-Mbps and 100-Mbps Ethernet, respectively. For Gigabit Ethernet, this would become a very limiting distance of only 25 m. In order preserve the CSMA/CD protocol for Gigabit Ethernet, IEEE 802.3 standard defined the Carrier Extension operation and Frame Bursting techniques to extend the protocol-limited transmission distance and retain the bandwidth efficiency. Since virtually no Ethernet device is implemented with Carrier Extension and Frame Bursting, we will not discuss it here. It should also be noted that Ethernet transmission distance is also limited by the physical technology such as available transmission power, receiver sensitivity, and signal attenuation and degradation as data propagates in the channel.

The CSMA/CD protocol is also called *half-duplex operation* as a station cannot transmit and receive at the same time. All the stations sharing the same bandwidth form a collision domain. Since only one station in a collision domain can be transmitting at a time, it can be expected that the average network performance degrades as the number of stations in a collision domain increases.

2.4. Repeaters

As explained before, Ethernet transmission distance is limited not only by the CSMA/CD protocol but also by signal degradation in the channel. For example, the collision domain size for 10-Mbps Ethernet is 2500 m. However, the physical limit for 10BASE-5 thick coaxial cable Ethernet is only 500 m. In order to extend the transmission distance, a repeater is required [2].

Repeaters usually have multiple ports. A repeater receives the signal from one port, recovers the MAC frame, and retransmits (broadcasts) the frame to every other port. If a repeater detects simultaneous transmission at more than one port, it will transmit a jamming signal to all the ports to cause collision detection by every workstation. Therefore, all the stations joined by a repeater belong to the same collision domain. Repeaters introduce signal delays. Such delays need to be taken into account when calculating the collision domain size. Because repeaters terminate the data at the MAC interface layer, it also enables Ethernet with different physical media (e.g., coaxial cable and optical fiber) to be interconnected. However, repeaters cannot be used to interconnect Ethernet with different speeds. "Bridges" (also called "switches") are needed in that case.

2.5. Modern Ethernet — Hubbed Architecture

In the late 1980s and early 1990s, structured wiring of category 5 unshielded twisted pairs became popular. In structured wirings, all the connections terminate at a central location (usually in a building wiring closet). Hubbed Ethernet was introduced to take the advantage of structured wiring.

In hubbed Ethernet, all the stations are connected to a hub (usually located at a wiring closet in a building) through point-to-point connections as shown in Fig. 6. There is no direct station-to-station communication. All the transmissions between stations have to go through the hub. Compared to the coaxial bus architecture, the hubbed topology has several advantages. First, the physical hub provides a convenient location where all the network connections can be centrally managed. In the bus architecture, the network connection is disrupted during the addition and removal of a workstation from the bus as the coaxial cable must always be properly terminated to prevent signal reflection. In the hubbed architecture, each port is individually terminated within the hub. Also, offending stations can be easily isolated from the network by disabling the corresponding hub port that the station is attached to, making the network more reliable.

It should be realized that there are two types of hubs used in Ethernet: a repeater hub and a switch hub. A repeater hub runs the CSMA/CD protocol. It allows only one station to transmit at a time in half-duplex mode. Repeater hubs are less costly and more commonly used for 10-Mbps (10BASE-T) Ethernet. Switched hubs with higher performance are more popular in modern Ethernet. More 100-Mbps Ethernet devices are running in the switched full-duplex mode than the half-duplex mode. Although the CSMA/CD protocol has been defined for Gigabit Ethernet, there is no Gigabit Ethernet devices built using half-duplex mode. Half-duplex operation is not defined in the 10-Gbps Ethernet standard, which will be finalized during the year of this writing (2002).

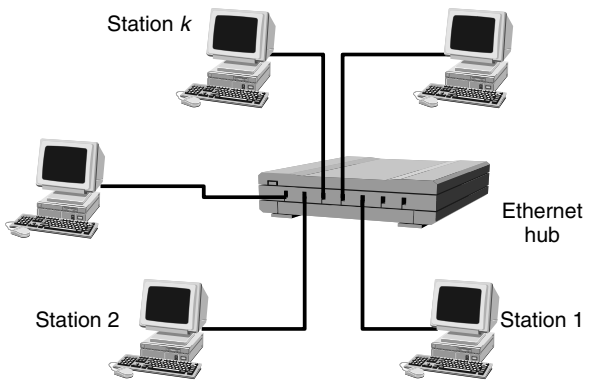


Figure 6. Hub Ethernet physical topology.

In full-duplex operation, the hub switch receives packets from stations attached to its ports and switches the packets to the appropriate output ports according to the destination address and an address table stored in the switch. Since all the connections are dedicated point-to-point links between switch ports and end stations, there is no multiple access and no need for the carrier sense operation. Of course, there is no collision to detect either. Furthermore, there is a separate transmitting path and a separate receiving path. This enables simultaneous transmission and reception, namely, full-duplex operation mode. It should also be realized that in a purely switched Ethernet environment, the transmission distance is limited only by physical signal impairments during transmission.

3. ETHERNET SWITCHES

Switches were initially called "bridges." Early bridges were mostly implemented in software. Bridges became switches when their functions were implemented in hardware, as a result of the development of silicon integrated-circuit technology.

A bridge is an internetworking device connecting different subnetworks [4]. Bridges enable networks with different speed, media, or even different protocols to be connected together. A bridge connecting different protocol subnetworks also needs to perform format translation and address mapping. We only cover Ethernet switches here. Ethernet switches are layer 2 devices switching at the level of Ethernet frames. Switches can be used to join networks running the CSMA/CD protocol. The use of switches enables the network to reach beyond the limit of collision domain size. In a shared Ethernet environment with a large number of stations, switches can also be used to improve the network performance by segregating the network into multiple interconnected collision domains with a smaller number of stations.

Ethernet switch operations and maintenance are defined in the IEEE 802.1D Standard [5]. Figure 7 shows the functional block diagram of an Ethernet switch. Switches have two major functions: packet forwarding and filtering. A switch interface to a workstation is called a port. Switch ports work in a "promiscuous" mode. A port examines all the input frames for the Destination Addresses. A switch uses a source address table (SAT) to determine whether the packet should be forwarded or filtered. The SAT stores MAC addresses and the associated switch ports. Figure 8 illustrates the operation of an Ethernet switch. If the Destination Address is found in

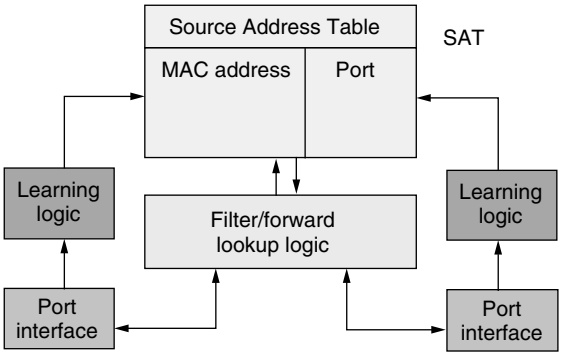


Figure 7. Functional block diagram of an Ethernet switch.

the SAT and its associated port is the same port where the packet arrives from, then the source node and destination node are attached to the same port of the switch and the packet is not forwarded; in other words, the arriving packet is filtered. If the destination node address is associated with another port in the SAT, then the packet is forwarded to that particular port. In the case when the destination node address is not in the SAT or if the packet is a multicasting packet, that packet is flooded to all the ports except the one it arrives from.

Switches use a switching fabric to route packets between ports. To achieve good performance, the switching fabric should have a capacity equal to the aggregate bandwidth of all the switch ports.

There are two ways that Ethernet switches use to populate the SAT. Static entries are entered through the management system and do not expire until updated by the system administrators. Dynamic entries are acquired through backward learning. In backward learning, switches examines the Source Address field of the arriving packets. If that address is not in the SAT, then it is entered into the SAT and associated with the arriving port. These addresses will time out if they are not active for a certain amount of time (300 s default value in IEEE 802.1D Standard). This allows the attached Ethernet interfaces to be moved from one location to another. Moreover, new entries will replace old entries when the SAT becomes full. To achieve good performance, the SAT size should be comparable to the expected number of stations connected to the switch to reduce the volume of broadcast traffic in the network.

4. ETHERNET PHYSICAL LAYER

4.1. Transmission Medium

Ethernet has gone through many different generations. Different physical (PHY)-layer technologies have been invented for different transmission media. The first-generation Ethernet (10BASE-5) uses thick coaxial cable as the transmission medium. Thick coaxial cables enable a transmission distance of 500 m. However, these cables are very inflexible and are used mostly in building risers. An *attachment unit interface* (AUI) has been devised to enable the analog transceiver called *medium access unit* (MAU) to be separated from the MAC digital processing unit (Fig. 2) and connected through a 50-m-long flexible cable with 15-pin IEC60807-2 connectors. As technology improves, the PHY layer and MAC layer become integrated into one circuit board and more flexible transmission media have been adopted. 10BASE-2 Ethernet using thin coaxial cable has a transmission distance of 185 m and unshielded twisted-pair (UTP) interfaces (10/100/1000BASE-T) with 100 m transmission distance have been subsequently introduced. A standard interface between the PHY layer and the MAC layer called *media-independent interface* (xMII) is adopted in modern Ethernet to separate the MAC design and the PHY design.

Category 5 UTP cables are by far the most popular medium for Ethernet connections to desktop computers for data rate up to 100 Mbps. Category 5 cables have the advantage of low cost and easy installation. The use of category 5 cable was enabled by modern signal processing techniques and silicon technology development.

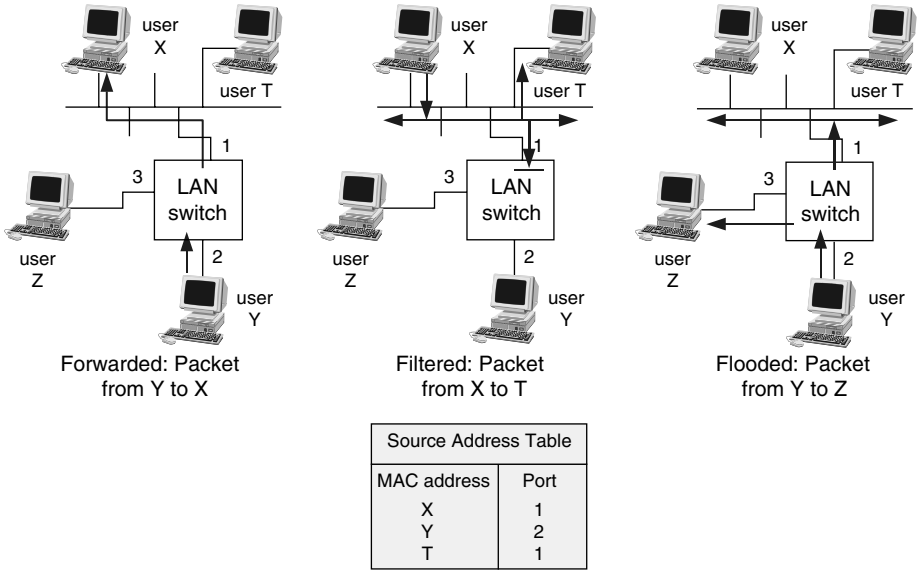


Figure 8. Ethernet switch operation.

As the transmission speed and transmission distance increased, the Ethernet community also moved from copper-based media to optical fiber–based media. Optical fiber has the advantages of virtually unlimited bandwidth and very low loss. An optical light source such as an LED (light-emitting diode) or laser diode is used as the transmitter and a photodetector is used as the receiver. Like copper cables, there are different kinds of optical fibers. Multimode fibers (MMFs) were commonly used in the past for short-distance communications. These fibers have large core diameters (55 μm or 62.5 μm are the two most commonly used). The large core diameter makes coupling light into the fiber an easier job compared to coupling light into single-mode fibers. However, in an MMF, optical signals can propagate in many modes with different speeds. This is called *modal dispersion*. Since light detection is mainly intensity-based, the received light pulses will get smeared after transmission in a multimode fiber. Modal dispersion limits the transmission distance as well as data speed.

Modern lightwave communication systems are increasingly using single-mode fiber (SMF). Standard SMF has a core diameter of 10 μm and therefore requires careful handling. As its name implies, optical signals can only propagate in one mode in an SMF. Therefore, SMF allows signals to propagate a longer distance. SMF suffer from chromatic dispersion. In a digital communication system, light signals are transmitted as pulses and have a finite frequency spectrum. Different frequencies of light travel at different speeds in an SMF. This incurs pulse broadening and limits transmission rate and transmission distance. Chromatic dispersion is a less severe effect compared to modal dispersion. It can be readily compensated if necessary (at a certain cost, of course). Single mode fiber has been adopted as one of the transmission media for Gigabit Ethernet. As the speed of Ethernet is increased to 10 Gbps, coarse wavelength division multiplexing (WDM) has also been adopted in one of the PHY designs. In the coarse WDM PHY design (called "10GBASE-LX4" in Ethernet Standard), four wavelengths separated by 20 nm in optical spectrum are used to carry the 10-Gbps payload in a parallel fashion. This reduces the requirement for high-speed electronics. The four wavelengths are combined into and separated from the same fiber using passive wavelength-division multiplexers.

Tables 1 and 2 specify the reach of Gigabit and 10 Gigabig Ethernets in different fiber media and the reach with different wavelength transmitters as specified in the IEEE 802.3 Standard. Signal attenuation and dispersion in optical fiber depends on the transmitter wavelength.

4.2. PHY Sublayers

4.2.1. MII and RS Sublayers. In modern Ethernet, the MAC layer and PHY layer are separated by the media-independent interface (MII). The MII uses parallel connections for control, timing, and data signals to reduce the digital processing speed requirements. The acronym MII is actually used for 100-Mbps Ethernet. For Gigabit and 10-Gigabit Ethernet, this interface is called "GMII" and "XGMII," respectively. Data are transmitted in units of 4 bits (called "nibbles") in MII, 8 bits (called "octet") in GMII, and 32 bits in XGMII.

The "reconciliation sublayer" (RS) is an abstract layer that defines the mapping of protocol primitives between the MAC layer and the PHY layer to the MII signal pins.

4.2.2. Physical Coding Sublayer (PCS). The PCS sublayer is responsible for line-coding the signals. There are

Table 1. Cable Length Specifications for 1000base-SX (850 nm Short-Wavelength Transmitter) and 100base-LX (1300 nm, Long-Wavelength Transmitter) Ethernet Interfaces

	1000base-SX		1000base-LX	
Fiber Type	Modal Bandwidth at 850 nm (MHz·km)	Range (ms)	Modal Bandwidth at 1300 nm (MHz·km)	Range (ms)
62.5-μm MMF	160	2–220	—	—
62.5-μm MMF	200	2–275	500	2–550
50-μm MMF	400	2–500	400	2–550
50-μm MMF	500	2–550	500	2–550
10-μm SMF	N/A	Not supported	N/A	2–5000

Table 2. Cable Length Specification for 10Gbase Ethernet Interfaces

	62.5 μm MMF		50 μm MMF			10 μm SMF	
Wavelength (nm)	850	850	850	850	850	1310	1550
Modal bandwidth (min; overfilled launch) (MHz · km)	160	200	400	500	2000	N/A	N/A
Operating distance	28 m	35 m	69 m	86 m	300 m	10 km	40 km
Channel insertion loss	1.61	1.63	1.75	1.81	2.55	6.5	13.0
Dispersion (ps/nm)	—	—	—	—	—	—	728

several reasons for line coding: (1) line coding provides enough transitions in the bit-stream for the receiver to recover signal clocks, (2) line coding provides redundant symbols for certain physical layer signaling purposes, and (3) line coding may encode multiple binary digits into a single transmitted symbol to reduce the transmission bandwidth requirement. This is especially important for the bandwidth-limited copper medium.

In 10-Mbps Ethernet, Manchester coding was used to embed the clock signal into the transmission data. Zeros and ones are respectively represented by high-low and low-high transitions in Manchester coding. Clock recovery is very easy. However, twice the bandwidth is required to transmit the data so that 20 MHz bandwidth is required for 10-Mbps signals. In 10-Mbps Ethernet with bus architecture, all the transceivers share the same physical bus, so transmission is bursty and a preamble in front of each packet frame is required for clock recovery at the receiver.

When Ethernet moved to the point-to-point architecture, a dedicated path is established between each hub port and the Ethernet station, and, hence, there is actually continuous physical signaling between each point-to-point transceiver pair. Burst-mode receiver operation is not required anymore. Even though the transmitted data may be bursty, idle periods between data packets are filled with idle symbols. Although the preamble has been defined for Ethernet frames, their importance has diminished in the point-to-point architecture except for backward compatibility.

Ethernet has been designed to operate on different media using different technologies. For example, 100-Mbps Ethernet has been designed to operate on both Category 3 and Category 5 cables and MMF. The most commonly used 100-Mbps Ethernet (100BASE-TX) uses two pairs of Category 5 cables and a line coding scheme called "4B/5B." The 4B/5B coding encodes 4 data bits into 5 bits and has also been used in FDDI (fiber distributed digital interface). 100base-T4 and 100base-T2 are defined for four pairs and two pairs of Category 3 cables, respectively. Since Category 3 cables have worse frequency responses than Category 5 cables, bandwidth efficient line coding schemes are used. In 100BASE-T4, an encoding scheme called "8B6T," which encodes eight binary digits into six ternary symbols, is used. 100BASE-T2 uses an encoding scheme called "PAM5 $\times$ 5." The transmitted data are encoded into two sets of five-level pulse-amplitude-modulated (PAM) symbols on two pairs of Category 3 cables. The resultant symbol rates of both 8B6T and PAM5$\times$5 encoding schemes are 25 Mbaud.

Gigabit Ethernet uses an 8B/10B encoding scheme, which encodes 8 binary digits into 10 bits. In the 8B/10B encoding scheme, there are no more than four continuous zeros or ones. The 8B/10B encoding scheme selects 256 patterns out of the 1024 possible 10-bit codes to represents the 8 data bits. Some of the extra codewords are used to represent idle symbols, and control sequences such as start of data and error conditions.

8B/10B encoding is very popular in high-speed data transmissions. It is also used for Fiber Channel systems invented by IBM. The price paid for 8B/10B encoding is a 25% overhead. So to transmit 1000 Mbps data, the physical layer needs to handle 1250 Msps (million symbols per second).

As Ethernet speed increases to 10 Gbps, the 25% overhead introduced by 8B/10B encoding makes the design of high-speed transceivers difficult. Therefore, 10-Gbps Ethernet adopted a new encoding scheme called "64B/66B," where 64 binary digits are encoded into 66 bits, with a 3% overhead. So instead of using 12.5-Gbps optical transceivers, 10.3-Gbps transceivers will suffice.

4.3. Physical Medium Attachment (PMA) Sublayer

The physical medium attachment sublayer transforms the PCS output codes into the physical symbols to be transmitted on the actual medium in the transmit path and performs the reverse operation in the receive path. For example, in Gigabit Ethernet, the PMA takes the 8B/10B-encoded PCS output from a parallel interface and transforms it into serial bit symbols to be transmitted on the actual fiber medium. In the receive path, it converts the received serial bits into 8B/10B codewords with parallel output. Therefore, the PMA layer is also commonly called "SERDES (SERializer-DESerializer)" in Gigabit and 10-Gigabit Ethernets.

4.4. Physical-Medium-Dependent (PMD) Sublayer

The physical-medium-dependent sublayer specifies the electrical and/or optical characteristics of the actual transceivers. These include properties such as output signal power, optical wavelength, modulation depth, receiver sensitivity, and saturation power. Table 3 shows typical characteristics of two different PMDs for Gigabit Ethernet. The 1000base-SX standard PMD uses short-wavelength (860-nm) lasers as transmitters and 1000base-LX standard PMD uses long-wavelength (1300-nm) lasers for transmission.

4.5. Medium-Dependent Interface (MDI)

The *medium-dependent interface* (MDI) specifies the electrical or fiber connectors used to connect Ethernet devices. For 10BASE-2 thin coaxial cable Ethernet, the BNC connector has been adopted. The most commonly seen MDI is the RJ45 connector for UTP cables. The fiber connector specified for Ethernet is the SC connector. Figure 9 shows some commonly used Ethernet media and their associated connectors (i.e., MDI).

5. 10-GIGABIT ETHERNET

We devote a section to 10-Gigabit Ethernet because it is the latest Ethernet standard. In fact, at the time of this writing, the 10-Gigabit Ethernet standard is still being finalized by the IEEE 802.3ae Working Group even though the standard has already been quite stable after many iterations of revisions [6].

Figure 10 shows the architecture of 10-Gigabit Ethernet. As shown in the figure, there are three types of 10-Gigabit Ethernet. 10 Gb/s transmission technology is still the state-of-the-art fiberoptic technology and is used

Table 3. Transmitter and Receiver Characteristics of 1000base-SX and 1000base LX

		1000base-SX	1000base-LX	
Medium:		MMF (50, 62.5 μm)	MMF (50, 62.5 μm)	SMF (10 μm)
Wavelength (λ):		770–860 nm	1270–1355 nm	
Transmitter	Spectral width	0.85 nm	4 nm	
	T_{rise}/T_{fall} (max: 20–80%)	λ > 830 nm: 0.26 ns; λ ≤ 830 nm: 0.23 ns	0.26 ns	
	Average launch power (max)	Lesser of class I safety limits or maximum receive power	−3 dBm	
	Average launch power (min)	−9.5	−11.5 dBm	−11 dBm
	Extinction ratio (min)	9 dB	9 dB	
	RIN (max)	−117 dB/Hz	−120 dB/Hz	
Receiver	Average receive power (max)	0 dBm	−3 dBm	
	Average receive power (min)	−17 dBm	−19 dBm	
	Return loss (min)	12 dB	12 dB	

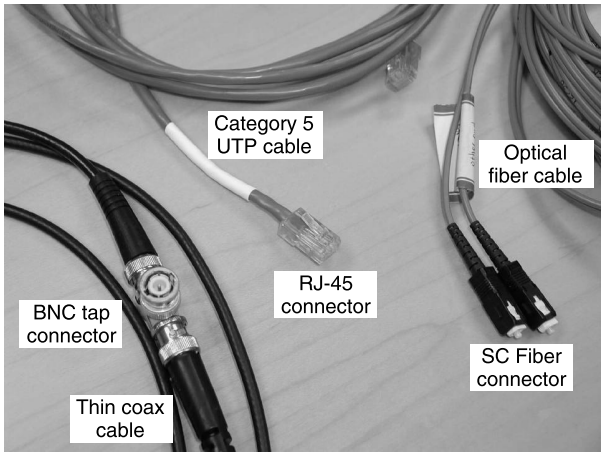


Figure 9. Some commonly seen Ethernet media and connectors (MDI).

mostly in wide-area backbone networks. In fact, 10-Gigabit Ethernet is targeted toward metropolitan-scale backbone network applications. Traditionally, wide-area networks (WANs) are dominated by SONET (Synchronous Optical NETwork) technology [7]. The 10Gbase-W standard is also called "WAN PHY." It includes a WAN Interface Sublayer (WIS) to encapsulate Ethernet frames into frames compatible with SONET Synchronous Payload Envelope (SPE). SONET framing includes extensive overhead bytes for operation, maintenance, and alarm signaling and performance monitoring. It should be noted that not all the SONET overhead fields are implemented by WIS. It should also be noted that the WAN PHY does not define SONET-compatible electrical and optical output, which is very stringent. The "SONET-Lite" framing introduced by the WAN PHY only makes it easier to map Ethernet traffic onto SONET equipment.

Both the 10Gbase-W and 10Gbase-R (Fig. 10) standards specify 64B/66B encoding in the PCS sublayer. Compared to the 8B/10B encoding used in Gigabit Ethernet, the overhead is reduced from 25% to 3%. This makes it easier to implement the high-speed optoelectronic front end.

The 10Gbase-X standard, however, continues to use the 8B/10B encoding scheme. The only PHY defined for 10Gbase-X is the 10Gbase-LX4 standard, which uses four coarse WDM wavelengths as shown in Fig. 11. Each wavelength is carrying a stream of symbols at 3.125 Gbaud/s speed.

10-Gigabit Ethernet will support both MMF and SMF fibers. SMF and 1.5-μm lasers will be used for connections up to 40 km without amplification. The 1.5-μm-wavelength signals have the lowest attenuation in optical fibers. They can also be easily amplified by mature Erbium-Doped Fiber Amplifiers (EDFAs) [8]. In fact, long-haul dense WDM (DWDM) system are designed mostly to operate around the 1.5 μm wavelengths. It should be noted that the 40-km transmission distance is limited by the low-cost transceivers specified in the IEEE 802.3 standard. There are a lot of commercially available nonstandard systems that enable native Gigabit and 10-Gigabit Ethernet signals to go beyond the distances specified in the IEEE 802.3 standard. Transmission of Gigabit Ethernet signals over 1000 km has been demonstrated in the MONET (Multiwavelength Optical NETwork) project that DARPA (Defense Advanced Research Program Agency) funded [9].

6. ETHERNET FOR THE FIRST MILE (EFM)

Ethernet in the first mile (EFM) is a new effort launched by the IEEE 802.3 committee in 2000 to work on the standards and technology required to provide Ethernet services by telecom service providers. The EFM study

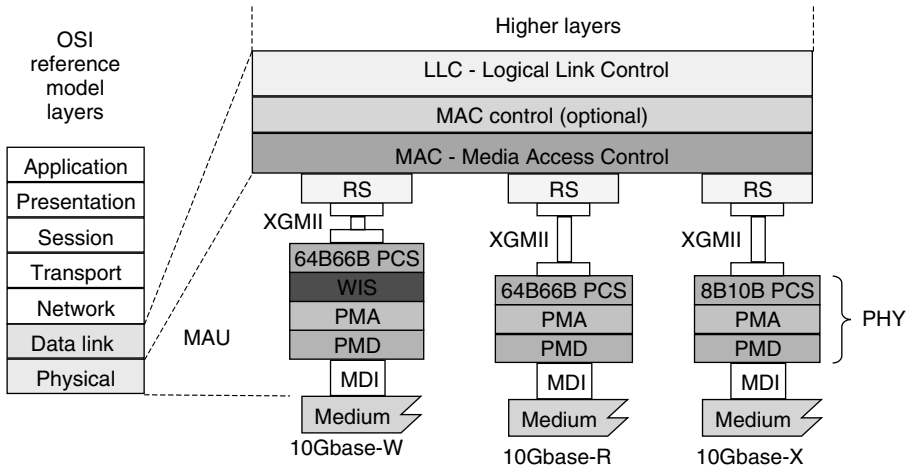
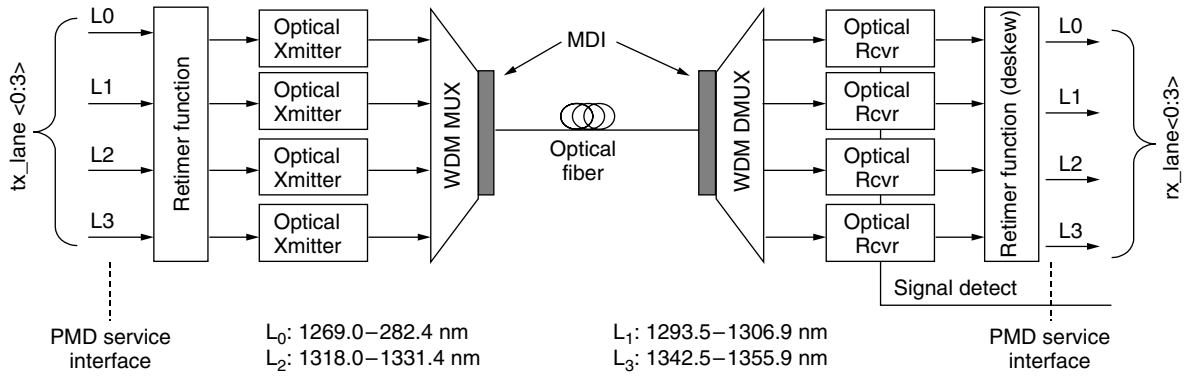


Figure 10. Architecture of 10-Gbps Ethernet.



Parameter	62.5 μm MMF	50 mm MMF		10 μm SMF	Unit
Modal bandwidth at 1300 nm (overfilled launch)	500	400	500	n/a	MHz•km
Link power budget	8.0	8.0	8.0	9.0	dB
Operating distance	300	240	300	10,000	m

Figure 11. 10Gbase-LX4 PMD.

group is registered as IEEE 802.3ah. The scope of the IEEE802.3ah is illustrated in Fig. 12. It covers the operation of Ethernet on a copper medium with extended reach and operation temperature, point-to-point Ethernet operation over single fiber, Ethernet passive optical network (EPON) and OAM (operation, administration, and management) issues and requirements for providing Ethernet services.

6.1. Copper PHY with Extended Reach and Operation Temperature

Ethernet has been defined to operate on Category 5 unshielded twisted pairs (UTP) up to 1000 Mbps and 100 m from a station to the hub. This group is studying the operation of Ethernet on a copper medium with longer reach, extended temperature range (for outdoor applications), and on lower-grade copper pairs. Advanced signal processing and error correction are implemented.

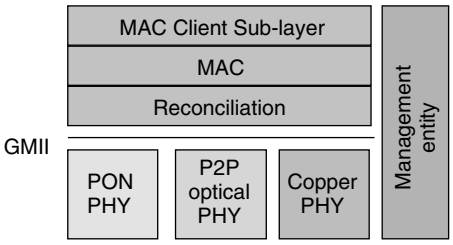


Figure 12. Scope of IEEE802.3ah (EFM) study group.

6.2. Single-Fiber Point-to-Point Operation

For this type of service, 100-Mbps Ethernet (100BASE-FX) has been defined to operate on MMF with a reach of 2 km. Gigabit Ethernet (1000base-LX) has been defined to operate on SMF with a reach of 5 km using 1300-nm lasers. For all the defined fiber-type Ethernet PHYs, full-duplex fiber pairs have been used, with one fiber for transmission and one fiber for reception. One job of the IEEE 802.3ah

study group is to define point-to-point (switched) Ethernet operation over a single fiber with extended reach at 100 and 1000 Mbps speeds. Fiber has the advantage of low loss and virtually unlimited bandwidth. However, fiber termination and connections are still more expensive compared to copper pairs. Therefore, single-fiber operation with longer reach has the advantage of saving cost and enabling service providers to cover a large area from a single central office. Wavelength-division duplex using widely separated 1.3/1.5 μm wavelengths and coarse WDM filters (called diplexers because they split and combine the 1.3 and 1.5 μm wavelengths) is the most popular method to achieve full-duplex operation in a single fiber.

6.3. Ethernet Passive Optical Network (EPON)

Passive optical networks (PONs) have been proposed as an economical way to provide future-proof broadband services. The architecture of an Ethernet PON is shown in Fig. 13. A PON consists of three major parts: an optical line termination (OLT) unit at a service provider's central office (CO), the distribution fiber plant itself, and an optical network unit (ONU) at each customer premise. PON is a distribution network. The signal from the OLT is distributed to ONUs at a remote node (RN). The RN can be either a power splitter (as in traditional PONs) or a wavelength router (as in WDM PONs). In both cases, the RN is a passive element. The term *passive optical networks* comes from the fact that the distribution fiber plant between the CO and the customer premise is passive, that is, there is no electrical power required in the field. This not only reduces operation cost but also improves the reliability.

Here we focus only on PONs using power splitters as the distribution mechanism. The OLT is basically an Ethernet switch that routes Ethernet packets between ONUs and also forms the gateway between the PON and the backbone network. Usually, the ONU outputs consists of T1 interface for legacy voice and data connections, PSTN (plain switched telephone network) interface for telephone services, and 10/100BASE-T interface for data applications.

6.3.1. Physical Design Considerations. The goal of EPON is to achieve transmission of Gigabit Ethernet with a distance of at least 10 km between the OLT and ONU, and a splitting ratio of 1:16 or higher. The available transmitter output power and receiver sensitivity will impose a limit on the transmission distance, remote node splitting ratio, and transmission speed. As an example, the ITU Standards (ITU Rec. G.983.1, G.983.2, G.983.3) [10] for an ATM PON (which carries ATM cells as opposed to Ethernet frames) specify 32-way split with 20-km transmission distances for an aggregate transmission speed of 622 Mbps (OC-12) in the fiber.

In a typical PON system, the downstream (from CO to customer premise) and the upstream (from customer premise to CO) signals are multiplexed on the same fiber using 1.5 and 1.3 μm wavelengths to avoid interference. Standard SMF has zero dispersion at 1.3 μm. This makes it possible to use low cost 1.3-μm-wavelength Fabry–Perot lasers for transmission without worrying about multimode output and chromatic dispersion. The non–zero dispersion at 1.5 μm wavelength will induce pulse broadening and hence will result in performance penalty. To support Gbps bit rate and beyond, DFBs (distributed feedback) lasers with single-mode output must be used at 1.5 μm wavelength. DFB lasers are more difficult to manufacture and therefore more expensive. Since the downstream laser is shared among all the ONUs, 1.5-μm lasers are used for downstream transmission, while 1.3-μm lasers are used at the more cost-sensitive ONUs.

6.3.2. Point-to-Multipoint Operation. In a PON system, the downstream signal is broadcast to all ONUs using the passive splitter. Each ONU detects its own packets by examining the destination address field as in the CSMA/CD protocol. The upstream transmission is achieved in a fashion similar to TDM. Time is divided into units of time quanta. Each ONU is granted permission to transmit at certain time instants for a specific number of time quantas by the OLT. The OLT may dynamically change the grant to transmit according to the network load and service-level agreement. This is called *Dynamic Bandwidth Allocation* (DBA).

Since each ONU is at a different distance from the OLT, it is necessary to align the logical time reference from each ONU to the OLT in order to avoid collision of upstream packets at the remote node. This is achieved through a ranging process illustrated in Fig. 14. The OLT periodically sends out ranging grant frames (or sync frames) to ONUs. An ONU will listen to the ranging grants when powered on. On receiving a ranging grant, it will send back a ranging request frame. The OLT will calculate the round-trip time between the OLT and ONU from the delay between the ranging grant and request frames. This information is sent back to the ONU so that the ONU can adjust its timing reference. If more than one ONU is trying to range at the same time, a backoff mechanism similar to the CSMA/CD protocol can be used to resolve collision.

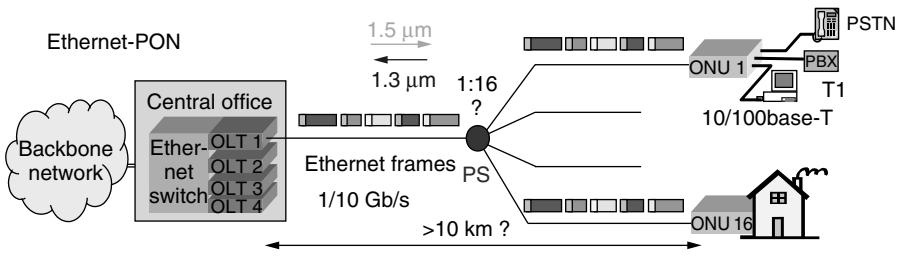


Figure 13. Ethernet PON architecture. A passive power splitter (PS) is used as the remote node.

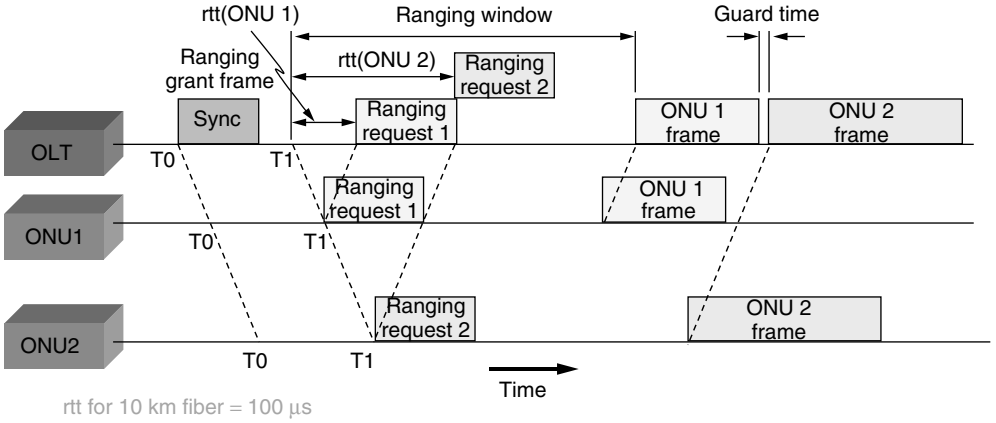


Figure 14. Ranging in a PON system.

We have seen that modern Ethernet assumes point-to-point operation between the hub and each station. The proper routing of Ethernet packets by the switches rely on this point-to-point architecture. However, a PON system is really a point-to-multipoint system. Therefore, although all the ONUs are on the same side of the OLT (which functions like a switch), if an ONU broadcasts a packet to other ONUs, this packet has to be relayed by the OLT. This also applies to other inter-ONU communications. The results of the point-to-multipoint nature of a PON system is that a point-to-point (P2P) emulation layer needs to be added to the OLT to function as the bridge between ONUs.

Previously, we also described that in modern Ethernet with point-to-point operation there is a continuous clock signal between two stations. Idle symbols are added by the PCS sublayer to keep the transmitter and receiver in sync when there are no data to transmit. This does not apply to PON systems. In a PON system, the downstream transmission from the OLT to ONUs is continuous. However, the reverse upstream transmission is bursty. A burst-mode receiver is required at the OLT and should have the capability to quickly synchronize with the transmitter with minimum preamble size. It should also be able to automatically adjust the threshold level for digital demodulation, as the received power from ONUs at different distances from the OLT will be different. At Gbps speed, these requirements are not easily achieved.

6.3.3. Other EPON Issues. There are other issues relating to OAMs covered by the EPON study group. Examples are loopback function to allow for link fault localization and detection and physical-layer device management. In fact, Ethernet PON is still an area of active research. Although there are proprietary prototypes developed by some system vendors, the IEEE 802.3ah study group still has a lot of work to do before a draft standard becomes available.

7. CONCLUSION

Ethernet has become the most popular technology for layer 2 transport. It has changed from a protocol for desktop LAN application to a transport technology for both LAN and MAN (metropolitan-area network) applications.

Some people believe that 10-Gbps Ethernet will also become a significant technology for long-haul wide-area network (WAN) applications. Traditionally, the high-bandwidth and long-distance backbone connections have been served using the SONET technology. The removal of the CSMA/CD protocol in high-speed Ethernet has enabled Ethernet to go long distances. Gigabit products are now widely available at much more affordable prices than SONET and ATM products [11] with comparable performance. Conventional SONET equipment has a very rigid time-division multiplexing (TDM) hierarchy. To compete with Ethernet, next-generation SONET boxes will implement TDM with statistical multiplexing to gain efficiency, and lightweight SONET functions to make them much more cost-effective.

The world of telecommunications is moving into packet switching, and Ethernet frames are really the dominating format in data communications. SONET and ATM equipment being designed for circuit switching does not provide an efficient platform for carrying the ubiquitous Ethernet traffic. Besides the cost and overhead efficiency, another advantage of end-to-end native Ethernet service is to simplify network management because there are no multiple platforms to manage. It should be noted that there are more people in traditional telecom companies who understand SONET better than Ethernet, and the reverse is true for data(communication) companies. The extensive SONET OAM overhead bytes are important for service providers to manage their systems. Ethernet, on the other hand, does not have many management capabilities because of its simplicity and cost-effectiveness. In order to provide end-to-end native Ethernet services, management functions need to be added to Ethernet technology. Standard bodies such as IEEE and IETF (Internet Engineering Task Force) are busy working on these issues. The important trend is that as data become the dominating network traffic, Ethernet will become the ubiquitous technology for both access and transport networks in LAN, MAN, and WAN environments.

8. USEFUL WEBSITES

1. IEEE 802 LAN/MAN Standard Committee, *http://www.ieee802.org/*.
2. Search for IEEE standards, *http://ieeexplore.ieee.org/lpdocs/epic03/standards.htm*.
3. IEEE 802.3 CSMA/CD (ETHERNET), *http://www.ieee802.org/3/*.
4. IEEE P802.3ae 10 Gb/s Ethernet Task Force, *http://grouper.ieee.org/groups/802/3/ae/index.html*.
5. Gigabit Ethernet Alliance, *http://www.gigabit-ethernet.org/*.
6. 10-Gigabit Ethernet Alliance, *http://www.10gea.org/index.htm*.
7. Online tutorial of Ethernet, *http://wwwhost.ots.utexas.edu/ethernet/ethernet-home.html*.
8. Ethernet for the first mile (IEEE802.3ah Task Force), *http://grouper.ieee.org/groups/802/3/efm/index.html*.

BIOGRAPHY

Cedric F. Lam obtained his B.Eng. in Electrical and Electronic Engineering with First Class Honors from the University of Hong Kong in 1993. He finished his Ph.D. degree in Electrical Engineering from the University of California, Los Angeles (UCLA) in 1999 and joined AT&T Labs — Research as senior technical staff member of the Broadband Access Research Department. He has worked on a range of research projects, including fiber to the home (FTTH), hybrid fiber coaxial (HFC) systems, optical regional access networks, and optical signal modulation techniques. More recently, he has devoted his energy to the development and application of high-speed Ethernet technology in optical networking. In 2002, Dr. Lam joined Opvista Inc., where he is now project leader.

Dr. Lam received the AT&T Research Excellence Award for his contribution to the Metro-DWDM project in 2000. He was a recipient of the Sir Edward Youde Fellowship from 1994 to 1997 and a recipient of the UCLA Non-Resident Fellowship from 1995 to 1999. Dr. Lam is technical program chair of the 2002 Wireless and Optical Communication Conference (WOCC 2002), program committee chair of the 2002 Asian Pacific Optical and Wireless Communication Conference (APOC 2002) and Associate Editor of the OSA *Journal of Optical Networking*. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE).

BIBLIOGRAPHY

1. R. M. Metcalfe and D. R. Boggs, Ethernet: Distributed packet switching for local computer networks, *Commun. ACM* **19**(7): 395–404 (July 1976).
2. IEEE Standard 802.3, *Carrier Sense Multiple Access with Collision Detection* (*CSMA/CD*) *Access Method and Physical Layer Specifications*, 2000 edition.
3. A. S. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, 1996.
4. J. J. Roese, *Switched LANs*, McGraw-Hill, 1999.
5. IEEE Standard 802.1D, *Information Technology — Telecommunications and Information Exchange Between Systems — Local Area Networks — Media Access Control* (*MAC*) *Bridges*, 1993.
6. IEEE Draft P802.3ae/D4.0, Dec. 2002.
7. U. Black and S. Waters, *SONET & T1: Architectures for Digital Transport Networks*, Prentice-Hall, 1997.
8. P. E. Green, Jr., *Fiber Optic Networks*, Prentice-Hall, 1993.
9. W. Xin, G. K. Chang, and T. T. Gibbons, Transport of Gigabit Ethernet directly over WDM for 1062 km in the MONET Washington DC network, *2000 Digest of IEEE LEOS Summer Topical Meeting on Broadband Optical Networks*, Aventura, FL, July 2000, pp. 9–10.
10. ITU-T Recommendations G.983.1, G.983.2, and G.983.3.
11. W. J. Goralski, *Introduction to ATM Networking*, McGraw-Hill, 1995.

MULTIBEAM PHASED ARRAYS

RANDALL G. SEED
MIT Lincoln Laboratory
Lexington, Massachusetts

1. INTRODUCTION

Multibeam phased arrays are used in wireless communications to establish RF links between one site and multiple other sites. The links may be either unidirectional or bidirectional. Multiple beams are formed from a phased-array antenna to provide directional point-to-point communications between nodes. The antenna beams from a phased array do not need to point in a constant fixed direction, as with a standard array antenna, or a reflector antenna. This property suggests that phased arrays are suitable not only for serving multiple users, but also for serving moving communications nodes. Multibeam phased arrays, in the current context, includes fixed and moving phased-array antenna beams (see Fig.1).

A phased-array antenna is composed of a large or small number of individual radiating antenna elements arranged in a one-dimensional (1D) or two-dimensional (2D) array. The 2D array is of more practical interest since it possesses the higher degree of freedom enabling greater beam agility. The 1D case is useful for mathematically demonstrating phased-array operation, and for illustration. Although generally the array is a flat plane, the elements can be arranged on a curved surface and thus made to conform to the shape of the vehicle that may host the antenna. Usually, all of the array's radiating elements are combined together to form one or more composite antenna beams from the array, and that will be assumed. This means that the multiple beams, producing the multiple channels, each emanate from the same single array, and are superimposed on each of the array elements. The signals from the elements are combined,

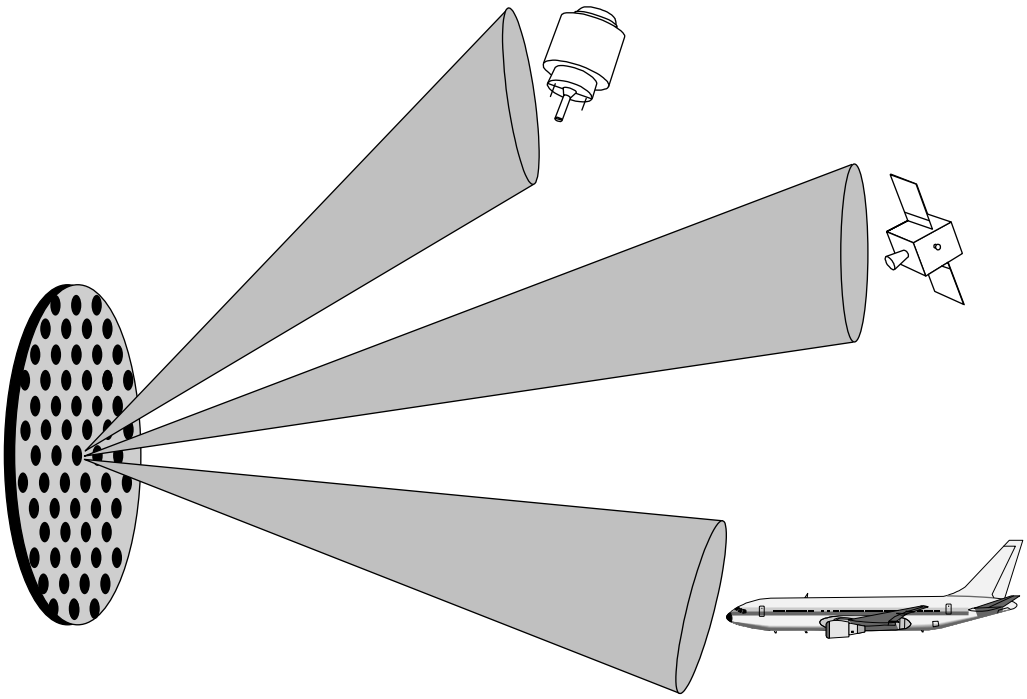


Figure 1. Single multibeam phased array antenna for simultaneous communications with multiple nodes.

or divided, depending on the direction of the signal, in a beam forming network. Each element can, and often does, radiate the distributed common RF signal with a different phase shift. The phase shift is defined relative to a reference element on the array. For convenience, the reference element is the nearest neighbor in a given direction. When the phase of the RF signal is altered on the individual radiating elements, the combined emanations form one or more directional beams in space. The beams are computed and measured in the "far field," also known as the *Fraunhofer region*. Nearer the array, in the "near field" or Fresnel region, the composite electromagnetic field structure is more complicated and generally more difficult to evaluate. The near-field analysis is not of relevance for communications link analysis, but rather for the antenna and antenna structural design.

A multibeam phased array antenna would generally be of interest in the following cases:

- Multiple users of undetermined angular position
- Broad area coverage with sustained high gain
- Multipath or jammer interference cancellation
- Multiple users, with node antenna volume and weight constraint
- Spectrum reuse

2. MULTIPLE-BEAM FORMATION

2.1. Single-Beam Linear Array

For a linear array, depicted in Fig. 2, the single beam array factor is

$$F(\theta) = \sum_{n=0}^{N-1} A_n e^{jkdn(\sin(\theta)-\sin(\theta_0))} = \sum_{n=0}^{N-1} A_n e^{jn(kd\sin(\theta)-\alpha)} \qquad (1)$$

when $A = 1$, then

$$|F(\theta)| = \frac{\sin\left[\frac{N}{2}kd(\sin(\theta)-\sin(\theta_0))\right]}{\sin[kd(\sin(\theta)-\sin(\theta_0))]} \qquad (2)$$

Example 1. Single beam from an unweighted linear array:

Array element spacing:	$d = \lambda/2$ m
Frequency:	$f = 1 \times 10^9$ Hz
Wavelength:	$\lambda = c/1 \times 10^9$ m
Number of elements:	$N = 16$ elements
Beam scan angle:	$\theta_0 = 7.18^\circ$ ($\alpha = 360^\circ/N$)
c is the speed of light	

The results are illustrated in Fig. 3.

Each 360° of total phase shift across the array, rotates the beam by one beamwidth. In this example, 7.18° is approximately one beamwidth.

2.2. Multibeam Linear Array

For a linear array producing two beams at the same frequency, the array factor is

$$F(\theta) = \sum_{n=0}^{N-1} A_{1n} e^{jkdn(\sin(\theta)-\sin(\theta_{10}))} + A_{2n} e^{jkdn(\sin(\theta)-\sin(\theta_{20}))} \qquad (3)$$

Since the two beams are assumed continuous wave, and are at the same frequency, the phases for each beam are additive and the signals to each beam position are correlated. As a result of this correlation, the beams will interfere unless suitable angular spatial isolation is provided. Frequency reuse may be obtained for modest

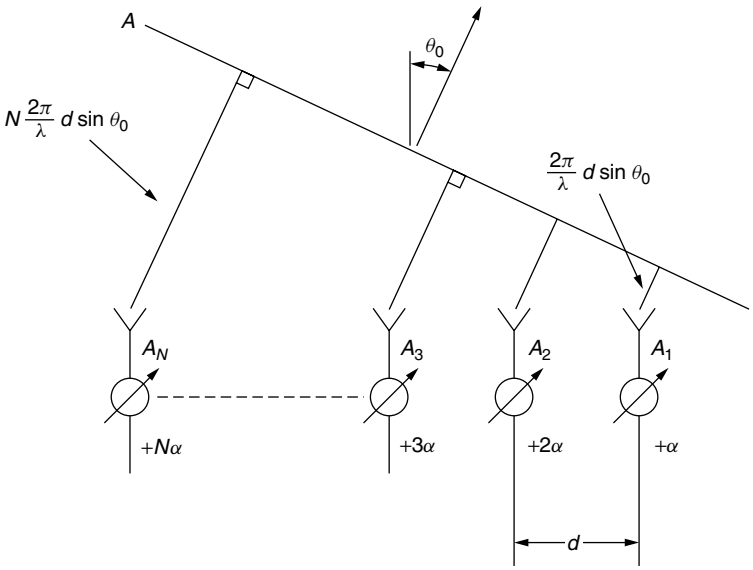


Figure 2. Geometry for linear array.

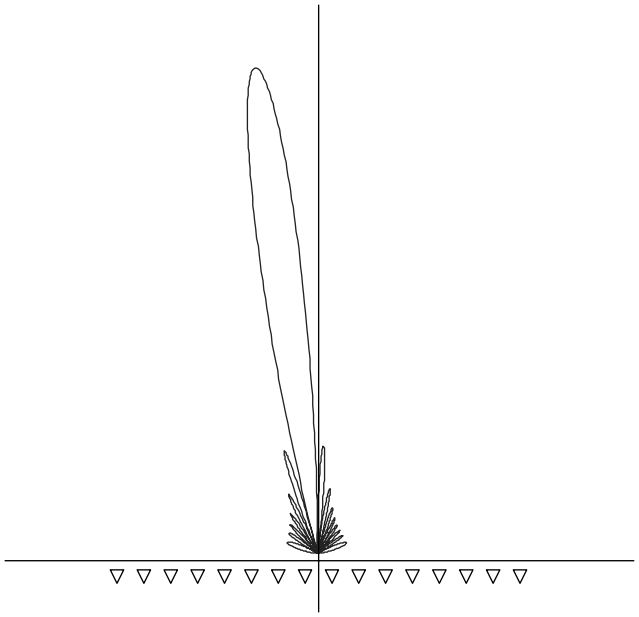


Figure 3. Single beam from a linear antenna array.

angle separation by providing suitable amplitude taper for each element, based on the amplitude taper for each beam, A_{1n}, and A_{2n}, such that the sidelobes from one beam negligibly impacts the other. Numerous amplitude weighting tapers are described in Chapter 2 of Hansen [1] for linear arrays, and Chapter 3 for planar arrays [1]. Spatially well separated beams will enable spectrum reuse. Finally, in the absence of an amplitude weighting taper, or spatial separation, then uncorrelated, or nearly uncorrelated signals will minimize the interference for arbitrarily spaced beams at the same frequency. Under this condition, frequency reuse channelization may be accomplished by generating uncorrelated signals within the same band, by means of polarization isolation, time-division multiplexing, or code-division multiplexing.

Example 2. Multiple beam, common frequency linear array.

$$d = \lambda/2$$
$$N = 16$$
$$\theta_{1,0} = 7.18^\circ \ (\alpha = 360/N)$$
$$\theta_{2,0} = -14.36^\circ \ (\alpha = -2 * 360/N)$$

In this example, two beams are generated and are separated by 3 beamwidths; the resulting pattern is illustrated in Fig. 4.

In design, it is best if signals are kept orthogonal, or minimally correlated for each beam, in order that spatial beams are separable, as in Fig. 5. In this case, the resulting beams are characterized by the individual array factors overlaid upon one another as in the graphic. Sidelobe levels are higher in Fig. 4 than in Fig. 5 because of phase interference.

2.3. Multibeam Two-Dimensional Array

For a two dimensional, planar rectangular array, the array factor is defined by

$$F(\theta) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} A_{mn} e^{\begin{array}{l} jk[md_x(\sin(\theta)\cos(\phi)-\sin(\theta_0)\cos(\phi_0)) \\ +nd_y(\sin(\theta)\sin(\phi)-\sin(\theta_0)\sin(\phi_0))] \end{array}}$$
$$= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} A_{mn} e^{jk[md_x(u-u_0)+nd_y(v-v_0)]} \qquad (4)$$

For uncorrelated signals, the antenna beams are non-interfering, and the multiple beam patterns are shown as graphical overlays of each array factor antenna pattern.

Example 3. Multibeam generation from planar rectangular array:

$$d_x = d_y = \lambda/2$$
$$N = M = 16$$

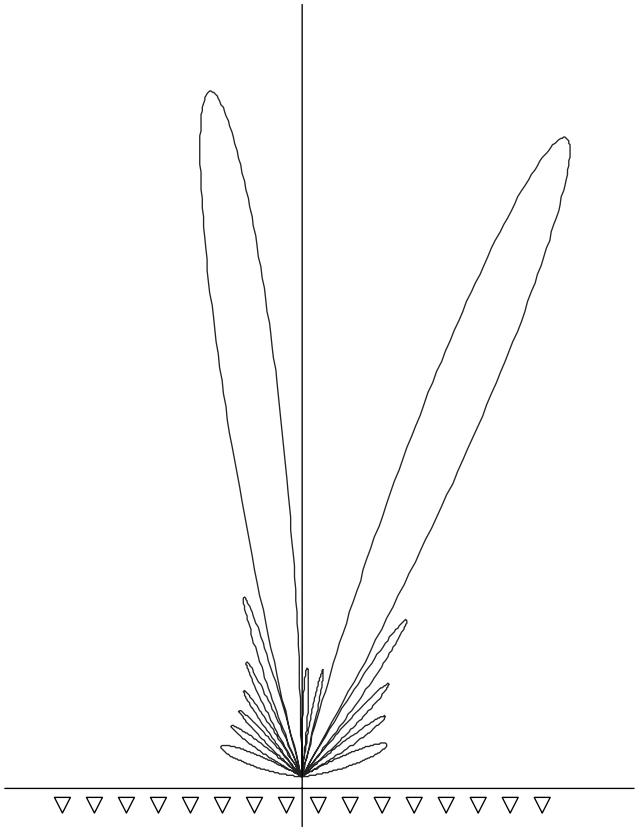


Figure 4. Multiple correlated beams from a single 16-element linear antenna array.

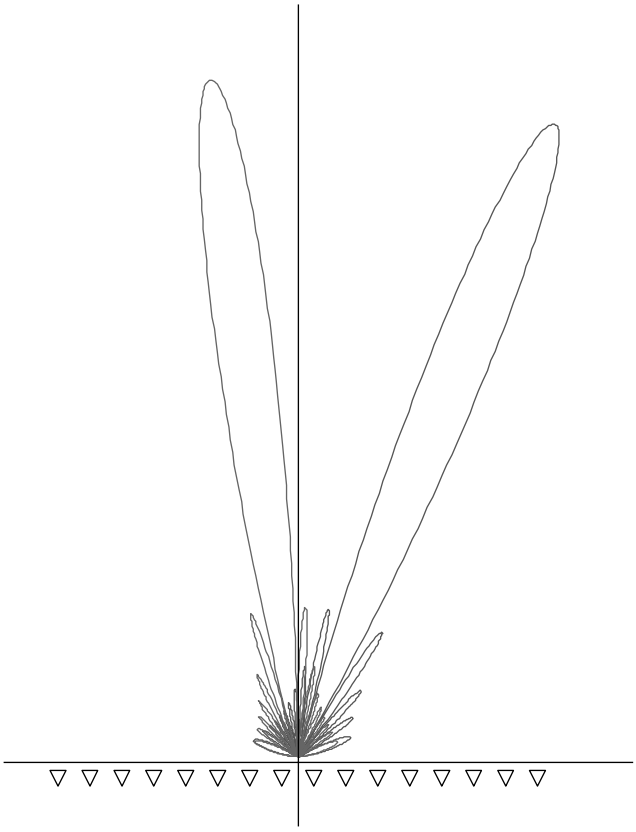


Figure 5. Multiple, uncorrelated beams, generated from a single linear antenna array.

$\theta_{1,0} = 0^\circ$
$\phi_{1,0} = 0^\circ$
$\theta_{2,0} = 20^\circ$
$\phi_{2,0} = 45^\circ$
$f = 1 \times 10^9$ Hz
$\lambda = c/1 \times 10^9$ m

A 3D representation of the resulting dual-beam pattern from the rectangular 2D array is plotted in u, v coordinates in Fig. 6.

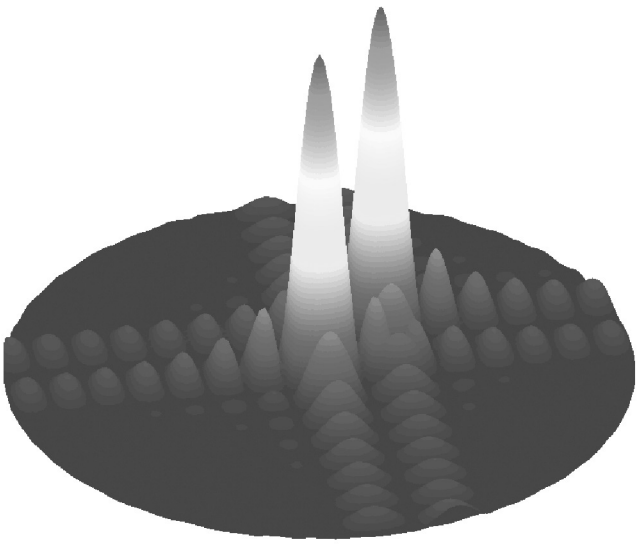


Figure 6. Multiple uncorrelated beams from a single two-dimensional phased array plotted in u, v coordinates.

There is a continuum of beam positions that can be obtained from the phased-array system. However, in practical terms the maximum number of beams required is the number of beams that fill the assigned angular volume, given an acceptable fractional beam overlap. The typical overlap is chosen as the angular distance at which adjacent beams are 3 dB below their peak gain value.

Tolerable antenna beam scan loss, and mutual coupling of antenna elements generally limit the designer's choice to scanning coverage of plus or minus 60°. Then, since the 3-dB beamwidth is

$$\theta_{3\ \mathrm{dB}} \cong \frac{0.886\lambda}{L} \tag{5}$$

for a linear array with uniform amplitude, the number of beams required to fill the angular space with 3-dB overlap per beam is

$$n \sin\left(\frac{0.886\lambda}{L}\right) = 2\sin(60^\circ) \tag{6}$$

For half-wavelength element spacing, we find that $n = N - 1$. By forcing the beam peaks for the end beams to

coincide with the edge of coverage, then the number of beams that fill the volume is equal to the number of elements, $n = N$.

Now, given this information, the choice to be made for beamforming is from:

1. Producing the logic and electronics to compute the phase shifts on each element to generate the desired number, M, of arbitrary beams
2. Hard-wiring, or hard-coding, the phase shifts per each element to form N beams from which the user selects M

A general beamformer used to produce N beams from an N element array is shown in Fig. 7. The number of electronic components used to form the beams is N attenuators and N phase shifters for each beam. This network requires $N \times N$ signal combinations, for a total of N^2 signal combinations.

On the other hand, with the second option for fixed beamforming, there are many applications in which it is convenient to form the fixed beams, and choose from among these N beams. The Butler matrix, for example, is used to form N beams using a minimum of phase shift and signal combining elements. The Butler beamforming matrix has been described as analogous to the FFT (fast Fourier transform) when N is a power of 2, and is as shown in Fig. 8 [1]. It requires only $N \times \log(N)$ combinations to form N beams. All lines are equal length, except for ones with phase shift.

2.4. Design Considerations for Multibeam Phased Arrays

The major design considerations for multibeam phased arrays for communications, are as follows:

Gain. Driven by the signal to noise ratio required by the two ends of the link. Drives array size, beamforming methodology.

Beamwidth. Driven by the requirement for spatial isolation of beams. Drives array size, frequency.

Scan Loss. Losses at maximum scan angle from array face normal. Drives array size.

Sidelobe Level. Driven by requirement for spatial isolation and interference rejection. Drives beamforming method, amplitude taper requirement.

Number of Simultaneous Users — Transmit. Driven by link requirements. Drives beamforming method, output power of array elements.

Number of Simultaneous Users — Receive. Driven by link requirements. Drives beamforming method.

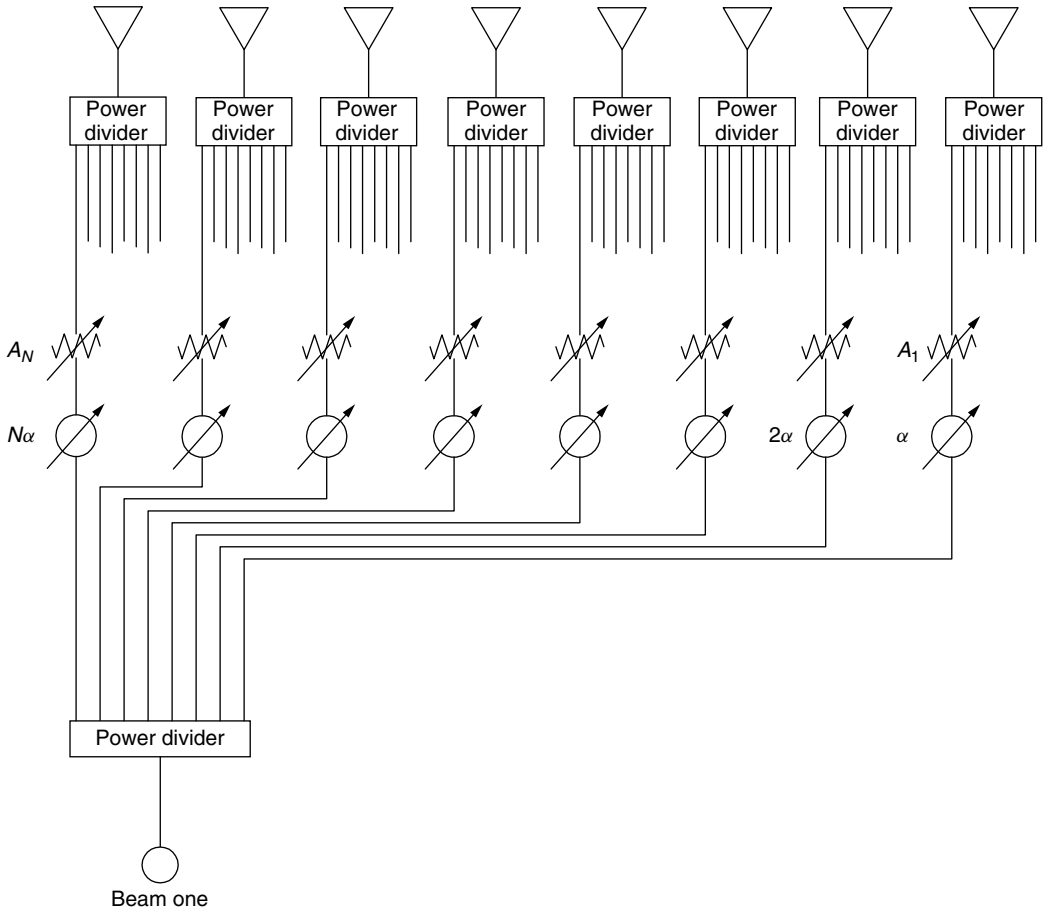


Figure 7. General beamformer for an N-element array.

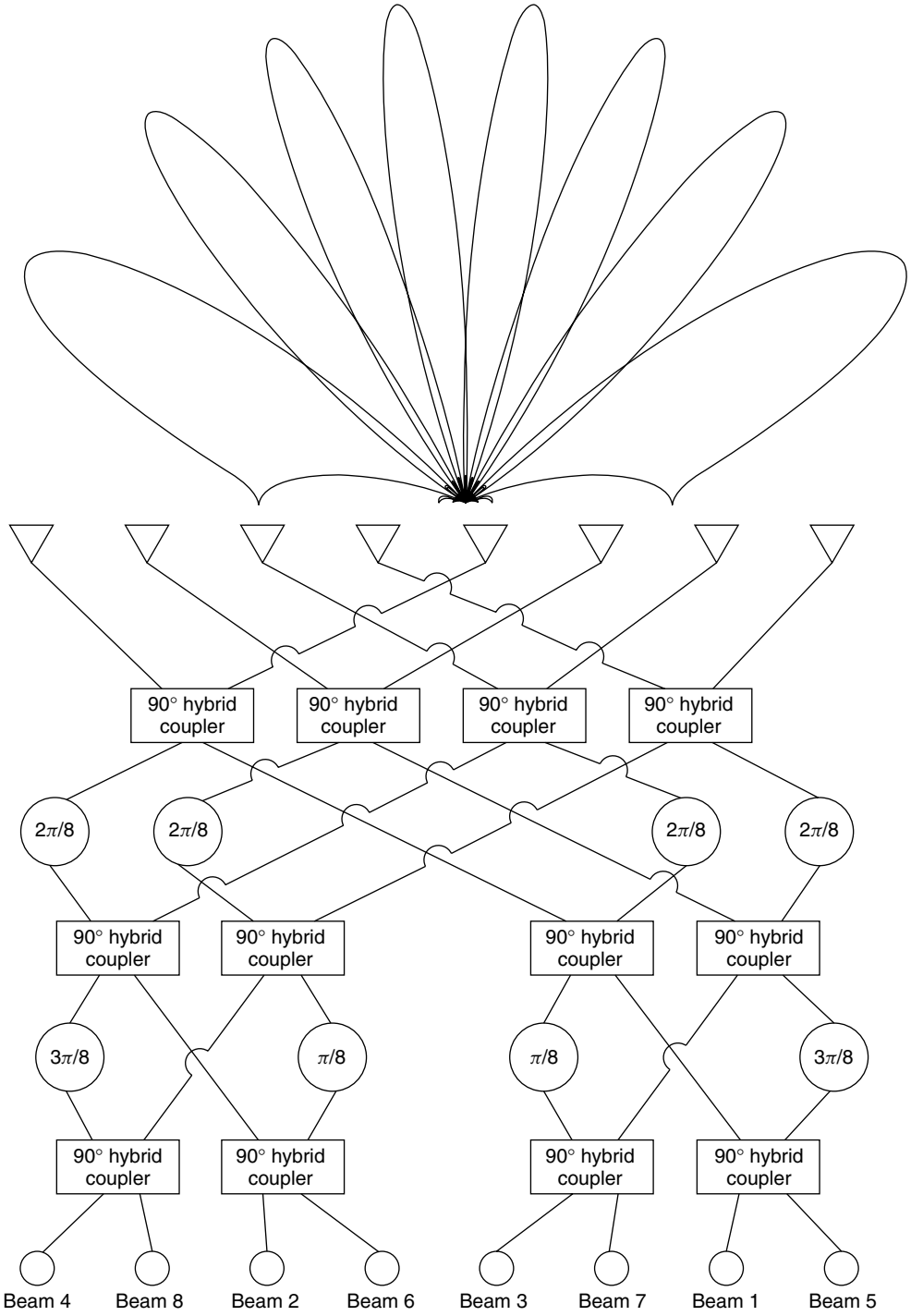


Figure 8. Butler beamformer for an 8-element linear array antenna.

Bandwidth. Driven by link requirement. Drives gain, scan loss, beamwidth, sidelobe level, and, beamforming method.

Generally, cost is the major issue with the phased-array antenna from an overall design perspective, versus a reflector antenna or fixed array. This is due to the number of amplifiers, transmit/receive switches, splitters, combiners, and the whole of the beamforming network. From an engineering point of view, the most significant issue is bandwidth. Since the array is driven by phase shifters, and not true time delay elements, a transmitted signal will undergo dispersion across the array face. The dispersion becomes a factor, and increases as the beam is scanned off of normal (see PHASED-ARRAY ANTENNAS).

Sensitivity on receive is set by the noise figure of the first-stage array element amplifier. The signal from an array element can be split as many times as desired after

the first stage without effective degradation in the SNR. This suggests the option of a large number of signals and channels with which to process the antenna beams. If the data are digitally sampled after the first stage, the signals can be combined in a computer.

Maximum transmit capability is defined by the maximum power output of the array element module. Since the signals are additive in the final-stage amplifier, fixed input levels are imposed. An option to take advantage of the full gain of the final-stage amplifier, is to time-division-multiplex the channels during transmission from the phased array. Both signal limiting and multiplexing of transmission are employed in practice.

3. EXAMPLES AND CURRENT APPLICATIONS

It is evident from the discussion that one advantage gained in developing a multibeam phased-array antenna is that one physical phased-array antenna with M beams, nearly performs the same task as M independent antennas. This advantage is appreciated, and exploited in the development of earth orbiting spacecraft, where size and weight become dominant constraints over complexity. The first communications systems to use multibeam phased arrays included several space telecommunications programs. Two of these are summarized, one U.S. government program, and one commercial program.

3.1. Government — TDRS System

The Tracking and Data Relay Satellite (TDRS) program, run by NASA, was an early user of the multiple beam phased array for communications. It is probably the best example available today, and certainly of its time, of multiple adaptive phased-array antenna beams used for communications.

The TDRS system is a series of spacecraft, operating independently from one another, that are positioned in geosynchronous orbit. Each spacecraft's set of antennas includes a phased array that points continuously towards the earth. It is designed to be able to acquire and communicate with other orbiting satellites or space vehicles, primarily those at low to medium earth orbit. The array has a field of view of plus or minus 13.5°, which allows it to follow satellites at up to 2300 mi altitude when not earth-eclipsed.

The TDRS multiple-access (MA) phased-array antenna is composed of 30 helical antenna array elements that, when combined, produce beams of nominal size $3.6^\circ \times 2.8^\circ$. All users transmit at the same frequency in 6 MHz bandwidth using QPSK (quadrature phase shift keying) at 2.2875 GHz. Their CDMA code key distinguishes users. Each TDRS user has a unique pseudonoise Gold code, yielding signals that are nearly uncorrelated. This minimizes antenna beam interference for unpredictable user position, and allows for the separation of individual user messages in the code domain [2].

The TDRS system avoids the need for complex beamforming circuitry in space, by transmitting the full bandwidth of each of the 30 elements to the TDRS ground terminal. Each element's 6 MHz information band is multiplexed into one of thirty 7.5-MHz channels and relayed to the ground in 225 MHz bandwidth. The ground system demultiplexes and digitally samples each channel. The digital antenna array element information is combined and reconstructed to form antenna receive beams in the direction of selected user satellites [3].

The TDRS MA beamforming process is executed in three modes:

1. *Beam steering* by knowing the position of the user beforehand, and setting the element phase shifts appropriately
2. *Adaptive beam steering*, determining the angles from measurements and signal processing
3. *Interference canceling*

One possible advantage to the adaptive method, is that a satellite's angular position may be determined by adaptively processing the returns employing the user's PN key, and thus, the vehicle may be closed-loop tracked.

In theory, any number of beams can be formed from the downlink data, since the raw measurements are available for each element. Beam formation limitations arise in the processing electronics at the ground station and, in this case, are reportedly limited to 10 beams per each of two ground terminals, for a total of 20 beams. Also, the number of elements, N, in this case 30, limits the ability to cancel co-channel interference. The array possesses $N - 1$ degrees of freedom [see Adaptive Arrays]. There are nominally four satellites in the constellation and so 5 beams are allocated to each spacecraft.

The second generation advanced TDRS system is under development now. The first and second of three satellites were launched in 2000 and 2001. Among other enhancements, the new TDRS system has moved the receive beamforming function of the multiple access phased-array antenna, onto the spacecraft itself [5].

3.2. Commercial — Iridium

The Iridium system is a constellation of 66 low-earth-orbit (LEO) satellites (the original system constellation was designed to contain 77 — the atomic number of iridium).

Each Iridium satellite has three main mission antenna (MMA) panels, each with approximately 108 patch antenna elements in a two-dimensional array. Sixteen simultaneous fixed beams are formed in two steps in the beamforming unit. First, 80 beamlets are formed using two-dimensional crossed Butler beamforming matrices. The Butler beamformer is composed of eight 16×16 Butler matrices, followed by ten 8×8 Butler matrices. Secondly, the sixteen beams are formed by performing power division/combining on groups of the 80 beamlets [6].

The communications channels are broken into 120 FDMA channels. Transmit and receive operations are effected by means of transmit/receive modules and separated using TDMA with four transmit and four receive channels for each FDMA channel [7]. Beam positions are switch selected from among the outputs of the fixed beamforming network.

4. DIGITAL BEAMFORMING

Multibeam phased array development trends are toward advanced techniques using digital beamforming, which enables adaptive beam steering, and adaptive nulling. This technique can be represented by a simple change to the block diagram: the amplitude and phase weights are performed digitally, therefore, the signal is digitally sampled at each element.

The technology driver for digital beamforming lies in the A/D (analog/digital) electronics. Although dynamic range requirements on digital sampling are less restrictive when conversion is performed at the element level, the A/D converter must operate at a rate of twice the full bandwidth of the entire link. For broadband, frequency-division multiple access communications, these rates can be formidable. Additionally, the datastream includes the samples from all of the N elements at once, and therefore, the internal data rate before signal processing becomes N times the sampling rate times the number of bits per sample.

Major advantages of digital beamforming include (1) full beamshape control, (2) ability to produce true time delay (vs. phase shift), by means of digital delay, (3) interference cancellation, and (4) adaptable calibration.

4.1. Adaptive Beam Steering

Adaptive beam steering is performed by using the least-mean-squares (LMS) beam-steering algorithm [8]. Beam steering can be performed for each of a multiple number of signals in a multiuser system. A signal, s, impinges the array face, producing the response, x, on each of 1 to N elements:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \tag{7}$$

Compute the signal-plus-noise covariance matrix for the excitation on the array elements, M:

$$\mathbf{M} = E[\mathbf{X}^*\mathbf{X}^T] \tag{8}$$

Compute a reference vector for the signal of interest as

$$\tilde{\mathbf{S}} = E[\mathbf{X}^* r] \tag{9}$$

The reference signal contains the representation, $r(t)$, of the desired signal, for example, that representation may contain the PN code for the channel of interest.

Now, the weight vector, W, containing the appropriate antenna element phase shifts to steer the beam in order to maximize the signal, is defined by

$$\mathbf{W} = \frac{\mathbf{M}^{-1}\tilde{\mathbf{S}}}{\tilde{\mathbf{S}}^{*T}\mathbf{M}^{-1}\tilde{\mathbf{S}}} \tag{10}$$

including a normalizing factor in the denominator and the best estimate signal is

$$\hat{s} = \mathbf{W}^T\mathbf{X} \tag{11}$$

The elements, W, form the array phase shifts, and amplitude weights, for the estimated beam.

Example 4. Single beam, linear array, adaptively steered. Let $N = 16$. Then

Signal at $\theta_{1,0} = 9.6^\circ$ ($\alpha = \pi/6$)
Signal power at each element = 1
Noise power at each element = 1
SNR at each element = 1.0 → 0 dB (see Fig. 9)

4.2. Adaptive Interference Cancellation

The Applebaum array is used for reducing, or canceling, the effects of interference, [4,8]. The procedure is analogous to adaptive beam steering, however, the covariance is computed slightly differently, as is the steering vector, **S**.

The noise covariance matrix is computed for the element response:

$$\mathbf{M}_n = E[\mathbf{X}_n^*\mathbf{X}_n^T] \tag{12}$$

Set the steering vector, **S**, to the phases of the elements for the current beam pointing position:

$$\mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} \tag{13}$$

Figure 9. One statistical representation of single beam adaptive beamforming of a 16-element linear array with 0 dB SNR per element. Adaptively formed beam — blue; ideal beam — red.

If it is a linear array, we obtain

$$\mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \begin{bmatrix} e^{-j\alpha} \\ e^{-j2\alpha} \\ \vdots \\ e^{-jN\alpha} \end{bmatrix} \tag{14}$$

Now, the weight vector, W, containing the appropriate antenna element phase shifts to modify the beam in order to maximize the signal in the presence of interference, is defined by

$$\mathbf{W} = \mathbf{M}_n^{-1}\mathbf{S}^* \tag{15}$$

and the maximized signal is

$$\hat{s} = \mathbf{W}^T\mathbf{X} \tag{16}$$

[see also Adaptive Arrays].

5. MULTIBEAM PHASED ARRAYS — FUTURE

Current trends in multibeam phased arrays indicate that future development is aligned toward digital beamforming. Key areas of focus include high-speed analog-to-digital converters. Impetus originates primarily from military users, including digital beamforming for radar and for multiuser spread-spectrum communications. The commercial world has its primary motivation for development of multibeam phased arrays for continued applications in space.

BIOGRAPHY

Randall Graham Seed received the B.S., M.S., Ph.D. degrees in electrical engineering from Northeastern University, Boston Massachusetts in 1990, 1992, and 1994, respectively. He was a Visiting Assistant Professor at Northeastern University from 1994 to 1995. In 1995 he was employed by TRW, Redondo Beach, California, where he worked on research and development in electromagnetic applications and space systems engineering. He joined Raytheon, Bedford, Massachusetts, in 1998, performing systems engineering on strategic ground-based radars. Since 2001, he has been with MIT Lincoln Laboratory, Lexington, Massachusetts, where he is engaged in research, design, analysis and development of novel airborne, sea-based, ground-based, and space based RF sensor systems. He has authored or co-authored approximately 20 papers, primarily in RF materials technology. Dr. Seed is a licensed Professional Engineer in the state of California, and is a member of IEEE. His areas of interest include advanced phased-array technology for radar and communication systems, sensor systems engineering, and active and passive materials technology applications to RF systems.

BIBLIOGRAPHY

1. R. C. Hansen, *Phased Array Antennas*, Wiley, New York, 1997.
2. R. Avant, B. Younes, D. Lai, and W.-C. Peng, STGT multiple access beamforming system modelling and analysis, *Military Communications Conf., 1992, IEEE MILCOM '92, Conf. Record, Communications — Fusing Command, Control and Intelligence*, 1992, Vol. 3, pp. 1028–1034.
3. R. Avant, B. Younes, G. Dunko, and S. Zimmerman, STGT multiple access beamforming equipment PC analysis system, *Military Communications Conf., 1992, IEEE MILCOM '92, Conf. Record, Communications — Fusing Command, Control and Intelligence*, 1992, Vol. 3, pp. 1035–1039.
4. S. P. Applebaum, Adaptive arrays, *IEEE Trans. Antennas Propag.* **24**(5): 585–598 (1976).
5. Space Network Online Information Center (no date), NASA Goddard Space Flight Center, *http://nmsp.gsfc.nasa.gov/tdrss/* (Jan. 2002).
6. J. J. Schuss et al., The IRIDIUM main mission antenna concept, *IEEE Trans. Antennas Propag.* **47**(3): 416–424 (1999).
7. R. A. Nelson (no date), Iridium: From Concept to Reality (online), Applied Technology Institute, *http://www.aticourses.com/news/iridium.htm* (Jan. 2002).
8. R. T. Compton, Jr., *Adaptive Antennas*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

FURTHER READING

Brookner E., ed., *Practical Phased Array Antenna Systems*, Lex Book, Lexington, MA, 1997.

Compton R. T., Jr., An adaptive array in a spread-spectrum communication system, *Proc. IEEE* **66**(3): 289–298 (1978).

Hansen R. C., ed., *Microwave Scanning Antennas*, Peninsula Publishing, Los Altos, CA, 1985.

Iridium Home, (2001). [Online]. Iridium Satellite LLC, *http://www.iridium.com/* (Jan. 2002).

Mailloux R. J., *Phased Array Antenna Handbook*, Artech House, Boston, 1994.

Zaghloul A. I., Y. Hwang, R. M. Sorbello, and F. T. Assal, Advances in multibeam communications satellite antennas, *Proc. IEEE* **78**(7): 1214–1232 (1990).

MULTICARRIER CDMA

DIMITRIS N. KALOFONOS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

The term *multicarrier CDMA* (MCCDMA) is used to describe multiple-access schemes that combine multicarrier modulation (MCM) and code-division multiple access (CDMA) based on direct-sequence spread spectrum (DSSS). Different ways of combining MCM and DSCDMA were proposed in 1993 by a number of researchers independently, and since then this idea has attracted significant attention because it combines two very successful techniques. Most of the proposed MCCDMA schemes use orthogonal carriers in overlapping subchannels for multicarrier transmission, also referred to as *orthogonal frequency-division multiplexing* (OFDM), but some MCCDMA schemes use few nonoverlapping subchannels. A MCCDMA scheme using nonoverlapping subchannels

is less bandwidth-efficient, but its implementation is a straightforward extension of existing DSCDMA systems and backward-compatible with them. For these reasons this type of MCCDMA has been selected as one of the options for digital cellular third-generation (3G) CDMA systems. On the other hand, MCCDMA schemes using OFDM are more bandwidth-efficient because they allow for minimum carrier separation and can be efficiently implemented in DSP using the fast Fourier transform (FFT). For these reasons OFDM-based MCCDMA schemes have attracted more intense research interest, although their adoption in practical systems is still limited.

MCCDMA schemes share many of the characteristics of both MCM and DSCDMA and attempt to make use of features of one component to overcome limitations of the other. The first component of MCCDMA, which introduces frequency-domain spreading in the signal design, is MCM [1]. MCM is based on the principle of transmitting data by dividing the stream into several parallel bitstreams, each of which has a much lower rate, and using these substreams to modulate several carriers. In this way, the available bandwidth is usually divided into a large number of subchannels, and each substream is transmitted in one of these subchannels. Since MCM is a form of frequency-division multiplexing (FDM), earlier MCM design borrowed from conventional FDM technology and used filters to completely separate the subchannels. This approach was soon abandoned since very sharp cutoff filters were needed and the number of subchannels that could be implemented was very small. The breakthrough in implementation of MCM came when the spectra of the individual subchannels were allowed to overlap, but the signals could still be separated at the receiver because they were mutually orthogonal. It was also found that in this approach both transmitter and receiver can be implemented using efficient fast Fourier transform (FFT) techniques. The orthogonality of the signals transmitted in different subchannels is achieved by using carrier separation of

$$\Delta f = \frac{1}{T_b} \tag{1}$$

where T_b is the MCM symbol duration. Thus, the carriers are located in frequencies

$$f_k = f_0 + k\Delta f, \quad k = 0, 1, \ldots, N-1 \tag{2}$$

where N is the total number of the subchannels and Δf is given in (1). Conceptually the transmitter and the receiver have the structure depicted in Fig. 1. In practice, MCM is implemented using the FFT, as depicted in Fig. 2, where block diagrams of the transmitter and the receiver are shown. With a proper selection of the guard interval, it can be shown that the effect of the channel on the transmitted symbols X_i is a multiplicative complex coefficient, which is the frequency response of the channel in the range of the respective subchannel. Note that by selecting the MCM symbol duration T_b large enough, the subchannels become narrow enough, so that the response of the channel in each of them can be considered approximately flat (non-frequency-selective channel)

$$Y_i = H_i X_i + N_i \tag{3}$$

where H_i is the complex channel coefficient for subchannel i, and N_i is the AWGN.

MCM enables high data rates and efficient bandwidth utilization and at the same time allows for large symbol duration. This large symbol duration is the most important characteristic of MCM, because it allows for almost intersymbol interference (ISI)-free transmission in both fixed and randomly fading frequency-selective channels. Other advantages of MCM are approximately flat-fading subchannels, which facilitate the inversion of the effects of the channel easier, and efficient and flexible implementation. On the other hand, the most serious drawbacks of MCM are sensitivity to carrier synchronization, nonlinear distortion because of the high peak-to-average values, and vulnerability to frequency selective fading, which can only be alleviated through signal diversity obtained with coding and interleaving in the frequency domain.

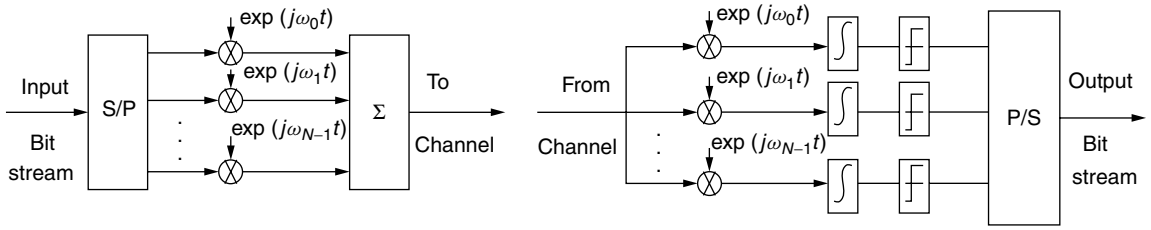


Figure 1. Conceptual structure of MCM transmitter (left) and receiver (right).

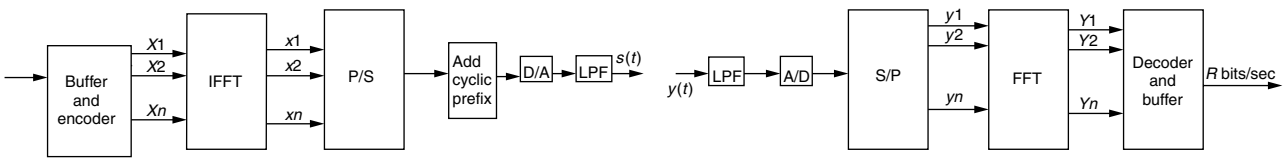


Figure 2. Implementation of MCM transmitter (left) and receiver (right) using FFT.

The other component of MCCDMA, which introduces time-domain spreading in the signal design, is DSSS [2]. DSSS signals are characterized by the fact that their bandwidth W is much greater than the information rate R. The redundancy added gives DSSS some important properties that make this technique popular for military and commercial applications. DSSS systems are used for combating interference due to jamming, other user interference, and self-interference caused by multipath propagation. The ability of DSSS systems to cope with interference caused by other users using the same channel makes DSSS the basis of a very effective multiple-access technique, namely, DSCDMA. In a DSCDMA system each user is coding its information symbols using a pseudorandom sequence, called a *spreading sequence*. Usually these sequences are selected such that they have an impulselike autocorrelation function and very low maximum cross-correlation values. In a DSCDMA system, each user i, $i = 1, \ldots, N_u$, spreads each information bit $b_i(k), k = 0, 1, \ldots$, by using a spreading sequence signal of length N_s, and the spread signals of all users are added when transmitted through the channel. If the channel is a frequency-selective, multipath fading channel, each station will receive multiple echoes of the transmitted signals with different attenuations arriving from N_p different paths. The optimal receiver of such a system involves a filter matched to the convolution of the channel impulse response with the transmitted signal. An approximation to that receiver is the RAKE receiver with N_p fingers. A simplified block diagram of a DS-CDMA system with a RAKE receiver, is depicted in Fig. 3.

DSSS-based CDMA is a very successful multiple-access technique, that has been selected as the basis of many contemporary wireless systems. The most important of its characteristics is its superb capability to resist interference: narrowband from intentional or unintentional jamming, wideband from other users using the same frequency band, and self-interference from transmission in dispersive multipath fading channels. DSSS also allows for hiding a signal from undesired listeners and achieving privacy. Cellular systems take advantage of the DSCDMA receiver structure to achieve soft handoff. On the negative side, practical DSCDMA receivers have a limited number of RAKE fingers and cannot exploit all useful signal energy, which may reduce their performance. Accurate synchronization and continuous tracking of signal arrivals from different paths is also necessary.

Depending on the system design, MCCDMA systems can share more characteristics with either of the two system components. Examples can range between the MCCDMA system selected as an option in some 3G CDMA cellular systems, which resembles more a conventional DS-CDMA system; and MCCDMA systems, where all the spreading takes place in the frequency domain, which resemble more a conventional MCM system. Different MCCDMA schemes attempt to a different degree to combine the interference rejection capabilities inherent in DSCDMA and the bandwidth efficiency and long symbol duration inherent in MCM.

2. MCCDMA SCHEMES

Because it combines MCM and DSSS, MCCDMA offers the unique possibility to spread the original data in both the time and the frequency domains. The different MCCDMA schemes that have been proposed in the literature cover the range between conventional DSCDMA, which offers maximum spreading in the time domain and no spreading in the frequency domain, and the scheme we will refer to as OFDM-CDMA, which offers no spreading in the time domain and maximum spreading in the frequency domain.

The first scheme we examine was proposed by DaSilva and Sousa [3]. In this scheme, the available bandwidth is divided into N subchannels that correspond to orthogonal carrier separation. Each user creates a block of $\mu = N$ symbols, and each of these symbols is spread using the user's spreading sequence. The chips corresponding to the spread symbol are then transmitted over one of the available subchannels using MCM. Note that in this way, each MCM block symbol contains one chip from each spread symbol, so that the transmission of each spread symbol is completed after N_s subsequent multicarrier blocks, where N_s is the length of the user's spreading sequence. The transmitter structure

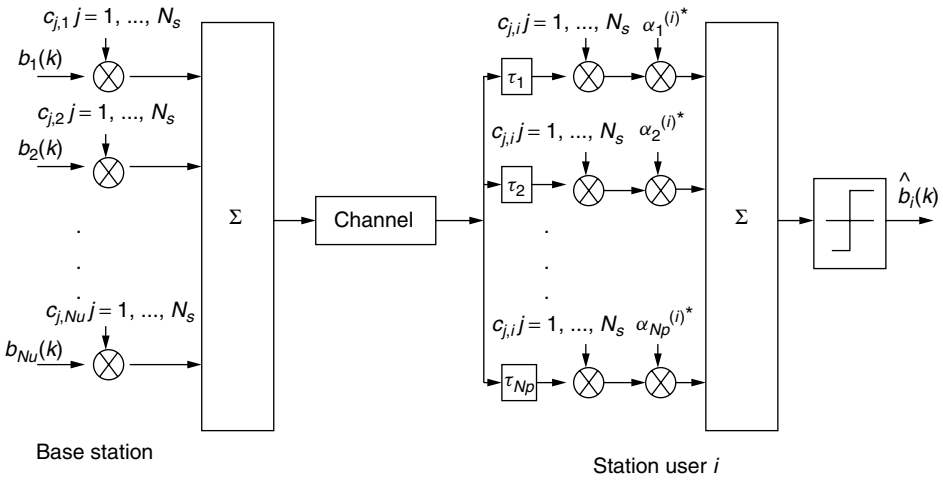


Figure 3. Block diagram of a DSCDMA system with a RAKE receiver.

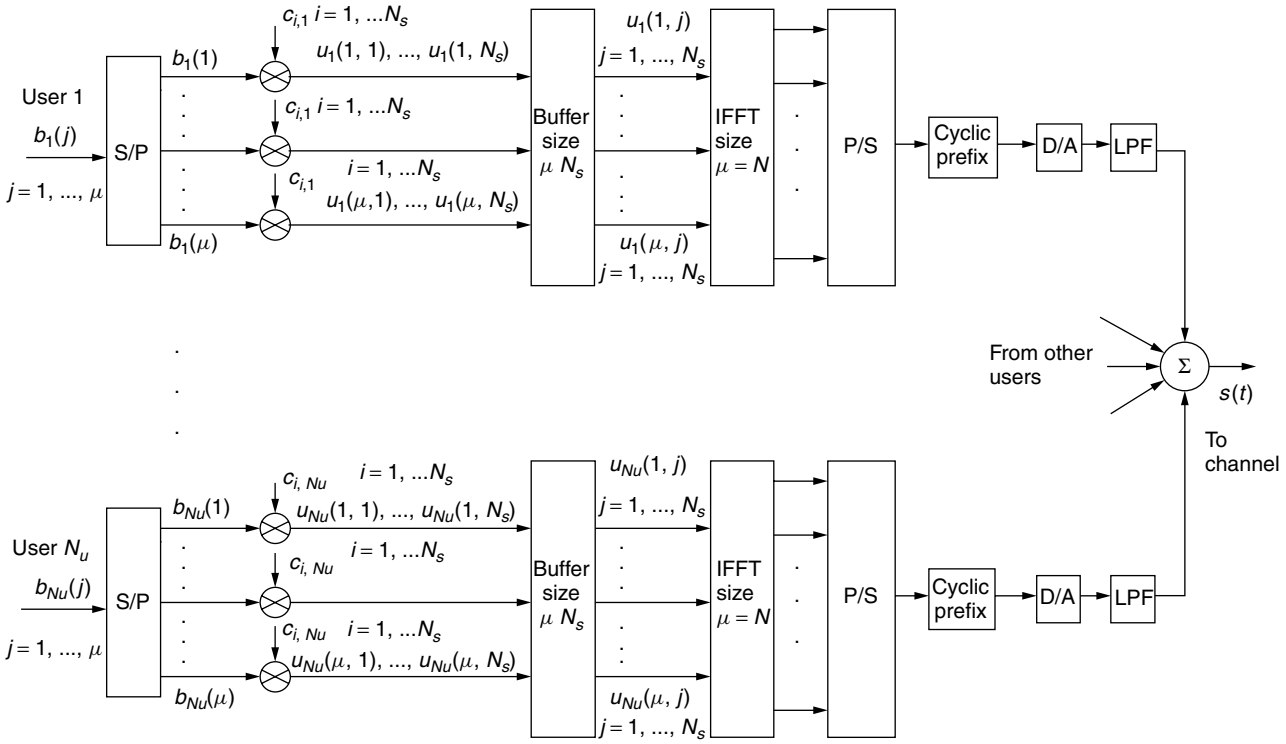


Figure 4. DaSilva and Sousa MCCDMA scheme: transmitter structure.

for this scheme is depicted in Fig. 4. This scheme uses multicarrier transmission as a way of maintaining the same data rate and spreading gain with a comparable DSCDMA system, while increasing the chip duration by a factor of N. This enables quasisynchronous transmission among different users and results in almost flat-fading subchannels, eliminating the need for RAKE receivers. The transmission of each symbol resembles that of a narrowband DSCDMA system over a flat-fading channel. Since each symbol is transmitted over only one subchannel, this scheme offers no frequency diversity. All the spreading is done in the time domain, and no spreading takes place in the frequency domain. Therefore, extensive coding of the symbols that are transmitted in parallel channels, often referred to in OFDM as *frequency-domain coding*, is needed to achieve acceptable performance [4]. Also, because of the large duration of the transmission of each symbol (N_s times the long MCM block symbol duration T_b), this scheme is more appropriate for slowly fading channels.

A variation of this scheme was proposed at the same time and independently by Kondo and Milstein [5]. The transmission of symbols in the available subchannels is done in a similar manner, but multiple copies of each symbol are transmitted in parallel in different subchannels as a means of introducing frequency diversity. The concept of the transmitter structure for this MCCDMA scheme is depicted in Fig. 5. In this scheme each symbol is spread in both the time and frequency domains. Although orthogonal carriers are considered in the original proposal, the authors later advocated the use of nonoverlapping subchannels as a practical design in realistic systems [6]. A scheme similar to that was later adopted as a multicarrier option for some 3G CDMA systems. A similar scheme that generalizes the idea of spreading symbols in both the time and frequency domains was proposed by Sourour and Nakagawa [7]. This scheme proposes the transmission of multiple copies of each symbol in orthogonal carriers and interleaving to maximize the achieved time and frequency diversity. This is a flexible system that allows the system designer to adjust the tradeoff between spreading in the frequency and time domains. Note that the term *multicarrier DSCDMA* (MCDSCDMA) is sometimes used [8] to distinguish these schemes, which allow for spreading in both the frequency and time domains, from other MCCDMA schemes.

The scheme proposed independently by Fazel [9], Chouly et al. [10], and Yee et al. [11] represents the other extreme in time–frequency spreading design. This scheme spreads each symbol entirely in the frequency domain, and there is no spreading in the time domain. The basic idea is to spread each symbol in the frequency domain by

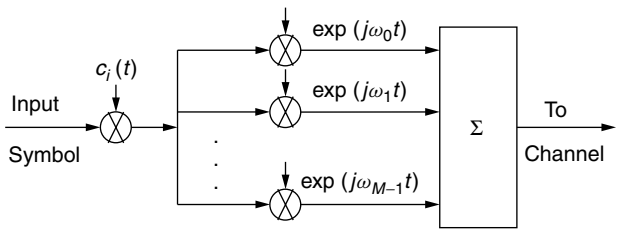


Figure 5. Kondo and Milstein MCCDMA scheme: transmitter structure.

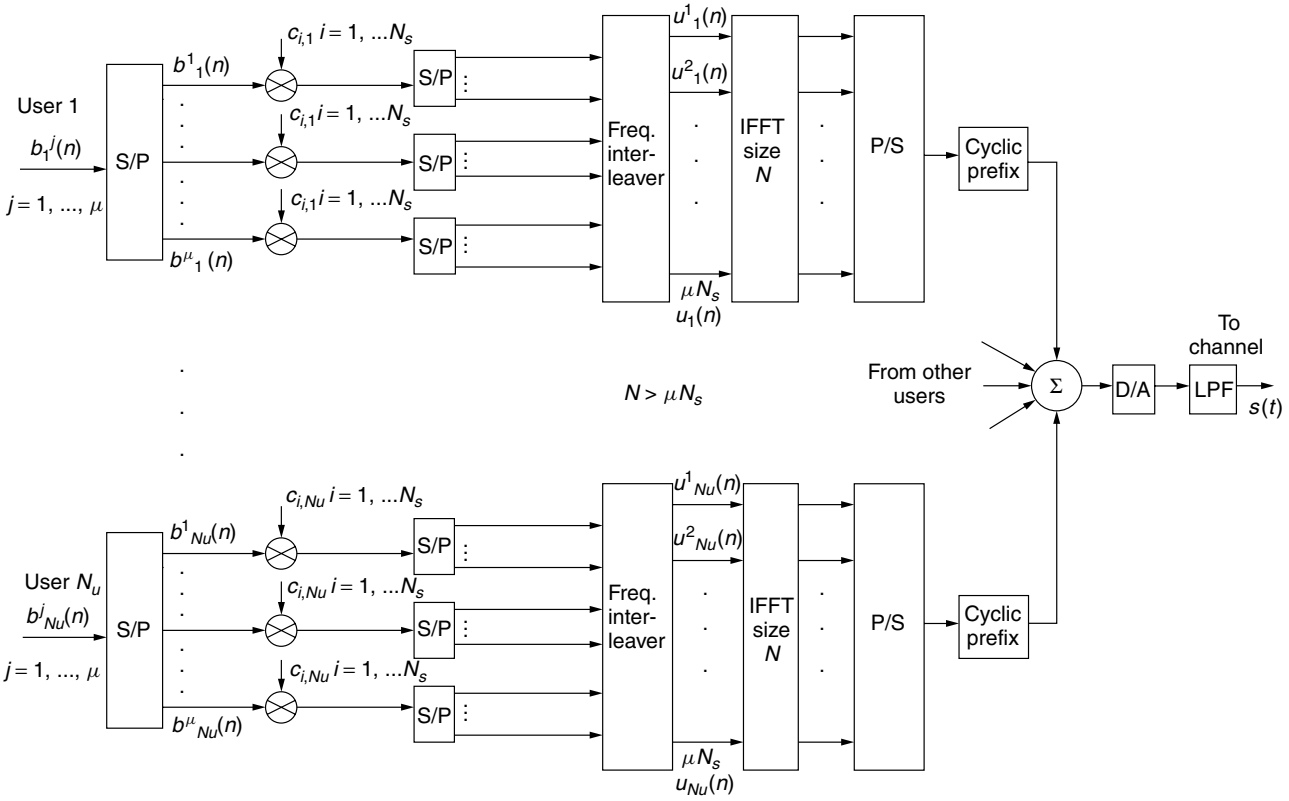


Figure 6. Fazel–Chouly–Yee [9–11] MCCDMA scheme: transmitter structure.

transmitting all the chips of a spread symbol at the same time, but in different orthogonal subchannels. The entire symbol is transmitted in one MCM block. This scheme offers the most frequency diversity among all MCCDMA schemes, since each symbol is transmitted in parallel over a large number of subchannels. To maximize the frequency diversity benefit, μ symbols of each user are transmitted in parallel in each MCM block symbol, where μ/T_b should be larger than the coherence bandwidth of the channel. Then with the addition of a frequency interleaver all chips corresponding to one data symbol are transmitted over subchannels undergoing approximately independent fading. The transmitter structure for this scheme is depicted in Fig. 6. This MCCDMA system has attracted the largest research interest to date and, in general, the acronym MCCDMA is used to describe this scheme [8]. A more appropriate acronym often used for this scheme, which helps avoid the confusion with other MCCDMA schemes, is OFDM-CDMA.

In all previous MCCDMA schemes, the symbols of each user were first spread using a spreading sequence according to DSSS, and then different techniques were used to send the resulting chips over the channel using multicarrier transmission. There is one MCCDMA scheme, however, proposed by Vandendorpe [12] and termed *multitone CDMA* (MTCDMA), which reverses this order. According to this scheme, first a MCM block symbol is formed using N symbols by each user, and then this signal is spread in the time domain by multiplying it with the spreading sequence. The idea behind this scheme is that for a given data rate and chip rate, the duration of each symbol can be much larger than in a corresponding DSCDMA system because of the effect of MCM. This longer duration, in turn, is translated in longer DSSS spreading sequences and higher processing gain. This approach has a potential drawback compared to other MCCDMA systems since it needs a RAKE receiver or some form of equalization at the receiver. The concept of the transmitter structure for this scheme is depicted in Fig. 7.

3. OFDM-CDMA

OFDM-CDMA has attracted the largest research interest to date among all MCCDMA systems. It is a scheme that offers a large degree of flexibility in system design, high bandwidth efficiency, and good performance with acceptable complexity. OFDM-CDMA is more appropriate for the forward link (base station to mobile users) of wireless systems, since its performance is best when all

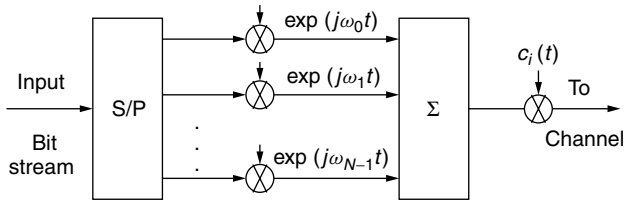


Figure 7. Vandendorpe MCCDMA scheme: transmitter structure.

users transmit in a synchronous manner. We will describe in more detail an OFDM-CDMA multiple-access system where N_u users are transmitting simultaneously in a synchronous manner using Walsh–Hadamard orthogonal codes of length N_s. Therefore up to N_s users can transmit at the same time. The nth MC block symbol (of duration T_b) for user i is formed by taking μ symbols $b_i^1(n), \ldots, b_i^\mu(n)$ in parallel, spreading them with the user's spreading sequence $\mathbf{c}_i = [c_{1,i} \ldots c_{N_s,i}]^T$, $c_{j,i} = \pm 1$, performing frequency interleaving, and placing the resulting $u_i^1(n), \ldots, u_i^{\mu N_s}(n)$ chips into the $N \geq \mu N_s$ available subchannels, each having width $\Delta f = 1/T_b$, by using an IFFT of size N. In practice N is larger than the number of subchannels μN_s required for the tranmission of the data in order to avoid frequency aliasing after sampling at the receiver. For that reason the data vector at the input of the IFFT is padded with zeros at its edges so that the $(N - \mu N_s)$ unmodulated carriers are split in both sides of the useful spectrum. The function of the identical frequency interleavers is to ensure that the N_s chips corresponding to each of the μ symbols are transmitted over approximately independently fading subchannels. As mentioned previously, this is possible only if μ/T_b is larger than the coherence bandwidth $(\Delta f)_c$ of the channel. After performing a parallel to serial conversion, a guard interval is added in the form of a cyclic prefix, and the signals of all the users are added and transmitted through the channel. The block diagram of this transmitter is depicted in Fig. 6. For simplicity we concentrate on only one of the μ symbols that each user transmits and we consider binary symbols $b_k(n) = \pm 1, k = 1, \ldots, N_u$ forming the data vector $\mathbf{b}(n) = [b_1(n), \ldots, b_{N_u}(n)]^T$, where n is the time index denoting the nth symbol interval. The transmitted signal during the nth MC block symbol period can be approximately written as follows:

$$s(t) = \sum_{k=1}^{N_u} \sum_{l=1}^{N_s} \sqrt{E_c} c_{l,k} b_k(n) e^{j[2\pi l(t-nT_G)/T_b]} \tag{4}$$

where $t \in [nT, (n+1)T]$, $T = T_b + T_G$, T_G is the guard interval chosen to be at least equal to the delay spread T_m of the channel, and E_c is the energy per chip.

Even in a frequency-selective, multipath fading channel, the fading of the narrow subchannels is approximately flat and is described by multiplicative complex channel coefficients $h_l(n), l = 1, \ldots, N_s$, which are samples of the channel frequency response at the center frequency f_l of the lth subchannel at $t = nT$. Because of the frequency interleaving function, the channel complex coefficient processes are approximately independent. Because of the existence of a guard interval with duration at least equal to the channel's delay spread, there is no intersymbol interference, and the signal received by user i can be approximately described by the following equation:

$$r(t) = ! \sum_{k=1}^{N_u} \sum_{l=1}^{N_s} \sqrt{E_c} h_l^{(i)}(n) c_{l,k} b_k(n) e^{j[2\pi l(t-nT_G)/T_b]} + \eta(t) \tag{5}$$

where $t \in [nT, (n+1)T]$, $h_l^{(i)}(n)$ are the complex channel coefficients that describe the channel between the transmitter and the user i, and $\eta(t)$ is the AWGN. At the receiver, the signal is sampled at a rate N/T_b, the samples that correspond to the cyclic prefix are discarded, an FFT of size N is performed, and frequency deinterleaving takes place. The vector $\mathbf{r}(n) = [r_1(n), \ldots, r_{N_s}(n)]^T$ at the output of the deinterleaver is given in matrix notation by the following equation:

$$\mathbf{r}(n) = \sqrt{E_c}\mathbf{H}(n)\mathbf{C}\mathbf{b}(n) + \boldsymbol{\eta}(n) \tag{6}$$

where $\mathbf{H}(n) = \text{diag}\{h_1(n), \ldots, h_{N_s}(n)\}$, matrix $\mathbf{C} = [\mathbf{c}_1 \mid \cdots \mid \mathbf{c}_{N_u}]$ is the $N_s \times N_u$ matrix whose columns are the spreading sequences of the users, $\mathbf{b}(n)$ is the data vector of the users, and $\boldsymbol{\eta}(n) = [\eta_1(n), \ldots, \eta_{N_s}(n)]^T$ is a vector containing zero mean, uncorrelated complex Gaussian noise samples, with variance $2\sigma^2$. The block diagram of an OFDM-CDMA receiver is depicted in Fig. 8.

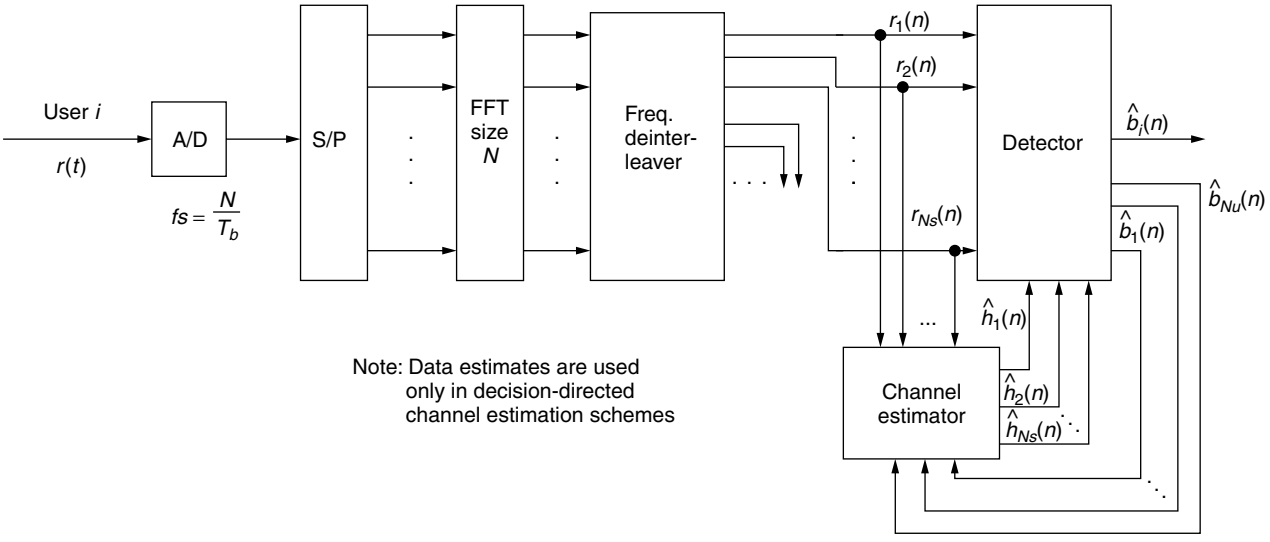


Figure 8. OFDM-CDMA receiver block diagram with channel estimation.

In the special case of an AWGN or flat-fading channel the channel coefficient matrix $\mathbf{H}(n)$ is within a constant the identity matrix. The optimal estimate of symbol $b_j(n)$ of user j is then obtained by correlating the received signal vector $\mathbf{r}(n)$ with the spreading sequence of user $\mathbf{c}_j$. This detector is also referred to as an *equal-gain combining* (EGC) detector. In this case, because of the orthogonality of the spreading codes used, there is no multiuser interference (MUI) and the performance is limited only by the AWGN. On the other hand, in the special case of a single-user system operating in a frequency-selective channel, the optimal symbol estimate is obtained by correlating the received signal vector $\mathbf{r}(n)$ with the spreading sequence of the user weighted by the complex conjugates of the channel coefficients. This detector is referred to as maximal ratio combining (MRC) detector. In the general case, however, of multiuser OFDM-CDMA systems operating over frequency-selective channels, the effect of the channel as demonstrated by matrix $\mathbf{H}(n)$ is to destroy the orthogonality among users and the performance is limited by the presence of both MUI and AWGN. The performance of both the EGC and MRC detectors in this case deteriorates significantly as the number of users increases and becomes unacceptable, even for relatively low system loads [13,14].

The optimal detector in the general case is the maximum-likelihood detector (MLD) [14,15]. This detector, however, may be impractical in many cases because its complexity grows exponentially with the number of users. For these reason, other low-complexity suboptimal detectors have been proposed, which attempt to combine good performance with reasonable implementation complexity. The simplest of such detectors, which corresponds to zero-forcing equalization, is termed *orthogonality restoring detector* (ORC). This detector inverts the effect of the channel by inverting matrix $\mathbf{H}(n)$, thus completely eliminating MUI. This causes, however, excessive noise enhancement due to the inversion of channel coefficients with very low magnitudes, which renders the performance of the ORC detector unacceptable [8,14]. More appropriate suboptimal detectors attempt to invert the effect of the channel without causing excessive noise enhancement. Examples of such detectors include the MMSE detector [10,14], the thresholded orthogonality restoring combining (TORC) detector [13,16] and the multistage or iterative OFDM-CDMA detector [15,16].

Most of the OFDM-CDMA detectors above require knowledge of the channel coefficient matrix $\mathbf{H}(n)$. For this reason a channel estimator is usually part of the receiver structure as shown in Fig. 8. The performance of practical OFDM-CDMA detectors operating in fast-fading channels deteriorates as a result of channel estimation errors. Examples of proposals for OFDM-CDMA detectors with channel estimation include pilot-symbol-based detectors [17] and decision-directed adaptive detectors [18].

4. COMPARISONS AND DISCUSSION

Few studies have been conducted to compare the performance of the different MCCDMA schemes. A comparison presented by Hara and Prasad [8] comparing OFDM-CDMA, MCDSCDMA, and MTCDMA shows that an OFDM-CDMA system with MMSE detection has the potential to outperform other schemes, with reasonable implementation complexity. The advantage inherent in OFDM-CDMA is higher frequency diversity. However, in designing an OFDM-CDMA system, calculation of the length of the spreading sequences, which determines the number of subchannels used to transmit each symbol, should be based on the maximum available frequency diversity as expressed by the ratio of total bandwidth over channel coherence bandwidth. If, for example, this ratio is in the order of 8, transmitting each symbol over $N_s \geq 64$ subchannels should not bring significant performance benefits. MCDSCDMA schemes using orthogonal carriers and introducing maximum frequency diversity based on the above ratio should demonstrate similar performance and capacity. On the other hand, MCDSCDMA schemes using nonoverlapping subchannels [6] have inherently lower bandwidth efficiency because of the necessary guardbands between adjacent subchannels. Other factors may favor the use of such schemes, however, such as the lack of intercarrier interference and ease of frequency synchronization. As a general comment, one has to be cautious before declaring one MCCDMA scheme as the "best" because implementation factors and the operational environment (e.g., the nature of interference sources and of the wireless channel) may impact the performance of each scheme differently and significantly.

Another topic that has attracted the interest of many researchers is the comparison between DSCDMA and MCCDMA. It has been reported [19] that there are cases where OFDM-CDMA systems with MMSE or MLD detectors significantly outperform DSCDMA systems, especially in the forward link of heavily loaded systems. Another argument in favor of OFDM-CDMA is that in practice RAKE receivers have a small number of fingers and thus may not be able to use all available signal energy, which may give some OFDM-CDMA systems with long symbol duration a certain advantage [8]. However, if all design aspects are taken into consideration (spreading with long random sequences, implementation issues such as channel estimation and carrier synchronization, channel coding), properly designed orthogonal MCCDMA and DSCDMA systems might plausibly demonstrate similar performance and system capacity [20]. Again, implementation issues and the operational environment can significantly affect the performance of each system and must be considered carefully before a system design is selected.

Finally, there has been research on the issue of comparing OFDM-CDMA with multiuser OFDM (MOFDM) systems [15,21,22]. OFDM can support multiuser systems without the use of spreading codes, either by assigning each user a subset of the available subchannels or by allowing only one user to transmit an MCM block at a given time as in TDMA systems. In general, OFDM-CDMA systems have higher complexity than do corresponding MOFDM systems; therefore their use is justified only when superior performance can be achieved. It was shown that in uncoded systems OFDM-CDMA significantly outperforms MOFDM because of the inherent

higher frequency diversity [21,22]. When channel coding is used, however, the performance gain due to increased frequency diversity introduced by frequency-domain coding is so much higher in MOFDM systems that their performance becomes similar to the more complex coded OFDM-CDMA [21,22]. Coded OFDM-CDMA systems can still outperform MOFDM at the expense of using more complex detectors (e.g., the MLD), especially for higher coding rates $>1/2$ [15]; however, coded MOFDM systems with lower coding rates ($<1/2$) have similar performance [15], and use of the more complex OFDM-CDMA is not justified in this case.

BIOGRAPHY

Dimitris N. Kalofonos received the Dipl.Ing. degree from the National Technical University of Athens (NTUA), Athens, Greece, in 1994, and the M.Sc. and Ph.D. degrees in electrical engineering from Northeastern University, Boston, Massachusetts in 1996 and 2001, respectively. From 1993 to 1994 he was with the Microwave Systems department in Intracom S.A., Athens, Greece, where he worked on DSP design for wireless systems. From 1996 to 2000 he was with the Wireless Systems department of GTE/Verizon Laboratories, Waltham, Massachusetts, where he conducted research on performance modeling of 2G and 3G CDMA cellular networks. From 2000 to 2001 he was with the Mobile Networking Systems Department of BBN Technologies, Cambridge, Massachusetts, working on adaptive waveform design for mobile ad hoc networks. He is currently a Senior Research Engineer at the Communication Systems Laboratory of Nokia Research Center in Boston, where he is conducting research on pervasive networking and mobile Internet technologies. His interests include wireless personal-area and local-area networks (PAN, LAN), ad hoc networks, and wireless integrated services networks. Dr. Kalofonos is a member of the Technical Chamber of Greece and a registered engineer in Greece.

BIBLIOGRAPHY

1. J. Bringham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* 5–14 (May 1990).
2. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, 1995.
3. V. DaSilva and E. Sousa, Performance of Orthogonal CDMA codes for quasi-synchronous communication systems, *Proc. IEEE Int. Conf. Universal Personal Communications* (*ICUPC'93*), 1993, Volume 2, pp. 995–999.
4. Q. Chen, E. Sousa, and S. Pasupathy, Performance of a coded multi-carrier DS-CDMA system in multi-path fading channels, *Wireless Pers. Commun.* **2**: 167–183 (1995).
5. S. Kondo and L. Milstein, On the use of multicarrier direct sequence spread spectrum systems, *Proc. IEEE MILCOM*, 1993, Vol. 1, pp. 52–56.
6. S. Kondo and L. Milstein, Performance of multicarrier DS-CDMA systems, *IEEE Trans. Commun.* **44**(2): 238–246 (Feb. 1996).
7. E. Sourour and M. Nakagawa, Performance of orthogonal multicarrier CDMA in a multipath fading channel, *IEEE Trans. Commun.* **44**(3): 356–367 (March 1996).
8. S. Hara and R. Prasad, Overview of multicarrier CDMA, *IEEE Commun. Mag.* 126–133 (Dec. 1997).
9. K. Fazel, Performance of CDMA/OFDM for mobile communication systems, *Proc. 2nd IEEE Int. Conf. Universal Personal Communications* (*ICUPC*), 1993, pp. 975–979.
10. A. Chouly, A. Brajal, and S. Jourdan, Orthogonal multicarrier techniques applied to direct sequence spread spectrum CDMA systems, *Proc. IEEE Global Communications Conf.* (*GLOBECOM'93*), 1993, pp. 1723–1728.
11. N. Yee J. Linnartz, and G. Fettweis, Multi-carrier CDMA in indoor wireless radio networks, *Proc. PIMRC*, Yokohama, Japan, 1993, pp. 109–113.
12. L. Vandendorpe, Multitone spread spectrum multiple access communications system in a multipath Rician fading channel, *IEEE Trans. Vehic. Technol.* **44**(2): 327–337 (May 1995).
13. T. Muller, H. Rohling, and R. Grunheid, Comparison of different detection algorithms for OFDM-CDMA in broadband Rayleigh fading, *Proc. IEEE Vehicular Technology Conf.* 1995, 835–838.
14. S. Kaiser, Analytical performance evaluation of OFDM-CDMA mobile radio systems, *Proc. 1 European Personal and Mobile Communications Conf.,* (*EPMCC'95*), Bologna, Italy, Nov. 1995, pp. 215–220.
15. S. Kaiser, *Multi-Carrier CDMA Mobile Radio Systems—Analysis and Optimization of Detection, Decoding, and Channel Estimation*, Ph.D. thesis, VDI-Verlag, Fortschittberichte VDI, Series 10, No. 531, 1998.
16. D. N. Kalofonos and J. G. Proakis, Performance of the multistage detector for a MC-CDMA system in a Rayleigh fading channel, *Proc. IEEE Global Communications Conf.* (*GLOBECOM'96*), Nov. 1996, Volume 3, pp. 1784–1788.
17. S. Kaiser and P. Hoeher, Performance of multi-carrier CDMA with channel estimation in two dimensions, *Proc. IEEE Symp. Personal Indoor and Mobile Radio Communications* (*PIMRC'97*), 1997.
18. D. N., Kalofonos, M. Stojanovic, and J. G. Proakis, Analysis of the impact of channel estimation errors on the performance of a MC-CDMA system in a Rayleigh fading channel, *Proc. IEEE Communications Theory Mini Conf.* (*CTMC*) *in Conjunction with GLOBECOM'97*, Nov. 1997, 213–217.
19. S. Kaiser, OFDM-CDMA vs DS-CDMA: Performance evaluation for fading channels, *Proc. IEEE Int. Conf. Communications*, 1995, pp. 1722–1726.
20. S. Hara and R. Prasad, Design and performance of multicarrier CDMA systems in frequency-selective Rayleigh fading channels, *IEEE Trans. Vehic. Technol.* **48**(5): 1584–1595 (Sept. 1999).
21. J.-P. Linnartz, Performance analysis of synchronous MC-CDMA in mobile Rayleigh channel with both delay and doppler spreads, *IEEE Trans. Vehic. Technol.* **50**(6): 1375–1387 (Nov. 2001).
22. C. Ibars and Y. Bar-Ness, Comparing the performance of coded multiuser OFDM and coded MC-CDMA over fading channels, *Proc. IEEE Global Communications Conf.* (*GLOBECOM'01*), 2001, 881–885.

MULTICAST ALGORITHMS

AIGUO FEI
MARIO GERLA
University of California at Los Angeles
Los Angeles, California

1. GROUP COMMUNICATION AND MULTICAST

Group or *multipoint communication* [12] refers to the type of communication in which information is exchanged among multiple (more than two) communication entities simultaneously. Many applications involve multipoint communication in nature, including videoconferencing, distance learning, distributed database synchronization, real-time distribution of news or stock quote, and multiplayer Internet gaming. On the other hand, as the fundamental method of telecommunication or computer communication, point-to-point communication is information exchange between two entities (although there may be many entities in between to help transport information). Modern communication networks have been very successful and efficient in supporting this type of communication service, while support for group communications is more a recent development and is likely to take many more years to mature.

Providing multipoint communication services is a multifacet problem: (1) addressing, group management, and membership management — how to identify and manage different communication groups and how to identify and manage members of a group; (2) session management — how to initiate a group communication session and how to control information transmission from one member to the others; (3) traffic control (e.g., not to let a sender overflow the network or a receiver); (4) reliability or data integrity — some applications may require any data sent from a source to be delivered to all other participants reliably, while some other applications may not have such requirement; and (5) data or information distribution — how data are distributed from their source to other participants.

The focus of this article is the last aspect of the problem: how to build a delivery structure to disseminate data for a communication group.

1.1. Group Communication in Shared-Medium Networks

There two fundamentally different types of networks: point-to-point networks and shared-medium networks. In a point-to-point network, every "link" in the network connects two stations and any data transmitted over that link by one station are received by and only by the station at the other end. In a shared-medium network (Fig. 1), any data transmitted by one station will be received by all others in the same network (or by those within a certain range as in wireless networks), although some stations may discard the data if they are not interested in them. Ethernet- or token-ring-based local-area networks (LANs) are examples of the shared-medium networks. Wireless LAN is another example. The nature of "shared medium"often limits this type of network to within a small range [i.e., LANs or metropolitan-area networks (MANs) at most].

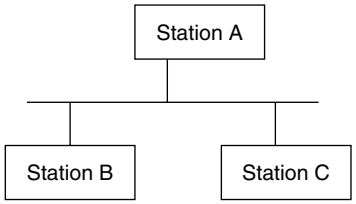


Figure 1. A shared-medium local-area network. Stations A, B, and C share a "same" wire; any data transmitted to the wire by one station will be seen by all others.

Here we discuss group communication in a shared-medium network using IEEE 802 LAN [31] as an example. In an IEEE 802 LAN network, each station can be identified by a unique 48-bit "individual" MAC (medium access control) address, as illustrated in Fig. 2. A data packet (i.e., usually called an "Ethernet frame" in an Ethernet network) targeted to a specific destination carries the address of the destination station. Some MAC addresses are "group" addresses that have the "group/individual" bit set to 1 (while the bit in an individual address is 0). One particular group address is called "broadcast" address, which is all 1s. Packets targeted for a specific group carry the corresponding group address (the broadcast address if targeted for all stations in the network). A station receives all packets transmitted in the network; however, except in some cases, it doesn't need to process every packet. The network interface device of a station can filter out packets except those that carry the station's unique address or addresses of groups that it is interested in. In some sense, group communication in a LAN comes (almost) "free"; a station can receive every packet transmitted — it needs to select only those in which it is interested. The set of groups a station is interested in is determined by the protocol layer above the MAC layer — we will discuss the case of IP.

In a more advanced switched Ethernet network [31] (or other type of switched LAN), the assumption that any packet transmitted by one station is received by all others is no longer true. However, the same address scheme is still used, and a switch delivers all packets with group

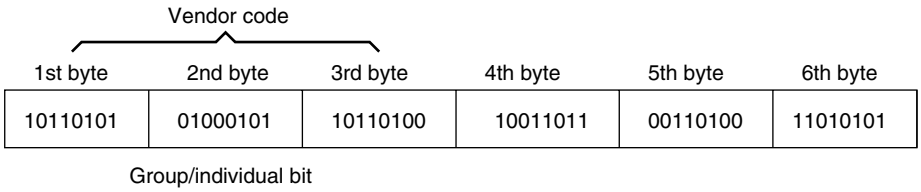


Figure 2. IEEE 802.11 MAC address. The higher 3 bytes are assigned to vendors to identify different vendors; a vendor then assigns a unique value of the lower 3 bytes to each network interface card it manufactures.

addresses to every other station in the network. Thus it makes no difference to a higher-layer protocol that utilizes the group communication capability of the IEEE 802 LAN.

1.2. Group Communication in Point-to-Point Networks

A point-to-point network consists of network nodes and point-to-point links connecting the nodes. Data transmitted by one node over a link are received only by the node at the other end of the link (i.e., a direct neighbor). Data destined to a single arbitrary node may travel a multihop path found by a routing protocol [19] to reach its destination. This type of network service is called *unicast*, and a point-to-point network can support it very efficiently. However, it takes more elaborate control to deliver data from one station to multiple destinations as required by group communications.

The simplest method to support multipoint communication is broadcasting. In this method, when a node receives a data packet, it sends the packet to every other neighbor except the one from which it receives the packet. This way, every packet is "broadcasted" from its source to all other nodes in the network. Conceptually broadcasting is very easy to implement; however, care must be taken to ensure that packets are "killed" somewhere so that they will not "loop" forever. This approach is extremely inefficient in network resource usage since there will be a lot of unwanted packets floating around in the network and bandwidth is wasted in transmitting them.

The second approach can be called a "naive unicast" approach, in which a source node sends a copy of the data to every other group member that is interested in receiving data from it, through unicast. This is illustrated in Fig. 3a; assuming that a group consists of nodes S, D, E, and F, among which S is the source node, when S wants to send data to the group, it sends a copy to each of them (D, E, and F). The third approach can be called a "server-based" unicast approach as illustrated in Fig. 3b; assume that node B is the server, S sends packets to B, and B forwards packets received to all other members (D, E, and F) through unicast. If another member node wants to send data to the group, it also sends the data to B, and B forwards it to everyone else. If there is only one single source, these two approaches would be the same if the server is placed at the source node. However, if any group member can be a traffic source, then the server-based approach has its advantages; every member node only needs to know what node is the server and doesn't need to know each other (which greatly simplifies group management). Clearly these two approaches are more efficient in terms of resource usage than broadcast.

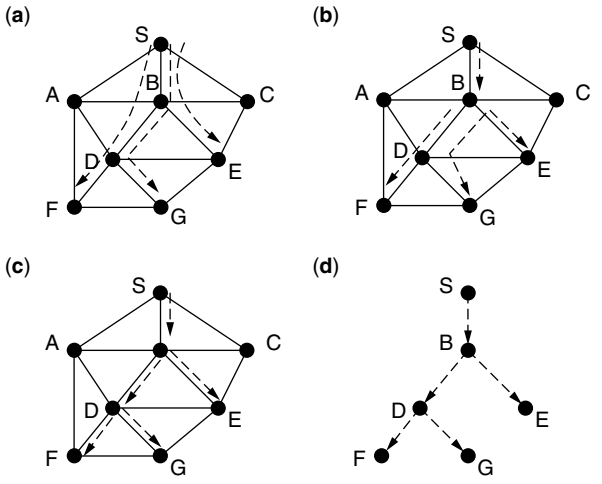


Figure 3. Different solutions for group communication: (**a**) naive unicast, (**b**) server-based unicast, (**c**) multicast, (**d**) a multicast tree.

Unicast-based approaches to group communication have some nice features: (1) they rely only on the unicast capability already provided by the network; no additional support is required for data delivery over the network (although other helper entities such as a server need to be introduced to help forward data among group members) and (2) they are conceptually easy to implement. Indeed, server-based solution is widely used to support group communication in the Internet today (Web-based chatroom, multiperson Internet gaming, etc.). In the public telephone network (including ISDN), services involving group communication [three-way conference call, multipoint videoconferencing using a multipoint videoconferencing control unit (MCU) [7]] are also implemented using unicast. However, they also suffer some clear drawbacks: scalability problems and resource efficiency. In both the naive unicast and server-based approaches, the scalability problem is a twofold problem: (1) a single node or a server has limited bandwidth to connect to the network, which limits the number of group participants that it can support; and (2) a single node or a server itself has limited local resource, which also limits the number of members it can support. The second disadvantages would only be seen in contrast of the multicast approach to group communication, which discussed next.

The fourth approach to group communication would be "multicast." This is illustrated in Fig. 3c; when node S wants to send data to the group, it sends a copy to node A, which forwards a copy to E and D, which then sends a copy to F and G. The paths taken by the data packets collectively form a tree delivery structure which is called a "multicast tree" as illustrated in Fig. 3d.

One advantage of the multicast approach is resource efficiency; for example, a data packet is forwarded only once over link ($B \rightarrow D$) to reach nodes F and G, but it is forwarded twice over that link (once by each unicast connection to F and G) in unicast-based approaches. Another advantage is better scalability with group size; if the network is large and a group has many members, any node in the multicast tree needs to forward a packet only to its neighbors in the tree (which are normally of a small number), while in unicast-based approaches, the source or center has to send many copies of a data packet to reach all other members.

Of course, the advantages of multicast are not achieved without a price — explicit extra support (besides unicast) from the network is required to do multicast forwarding. For example, when node B receives a packet from S, it must know that the packet is a multicast packet and that the packet should be forwarded to neighbors D and E (but

not *A* and *C*). This means that extra packet processing has to be done, and some extra state information regarding a multicast group has to be maintained at network nodes. The problem of constructing a multicast tree in a network and installing necessary state information is the multicast routing problem, which is accomplished by a multicast routing protocol, and at the center of that there is a multicast routing algorithm.

2. MULTICAST ROUTING IN THE INTERNET

A routing algorithm is part of a routing protocol. Before we go into any algorithm details, we give an overview of the multicast routing architecture in the Internet.

2.1. Overview

Conceptually the Internet can be modeled as a three-level hierarchy (see Fig. 4). At the lowest level, the computers (hosts) of end users are connected together to form a local-area network (LAN) such as an Ethernet network — this could be a single computer for a home user. Each LAN has a *designated router* that connects all the computers of that LAN to the Internet — the vast network consisting of all the computers interconnected together. For a home user, the designated router is normally the access router at the Internet service provider (ISP) side, which a home computer connects to through a dialup line or a cable/DSL modem. For business users, their LAN routers are also connected (e.g., through other routers within their own networks) to access routers at service providers.

Access routers of an ISP network are connected together through other routers to form an intradomain network, the second level. Some large corporations may also have multiple LANs connected together to form an intradomain network as well. At the highest level, different domains are interconnected together through border routers to form an interdomain network (i.e., a network of domains). Very often intradomain and interdomain networks are point-to-point networks. Although routers within a domain or border routers between domains are sometimes connected through a shared-medium network, they are often treated as a point-to-point network logically for routing purposes.

In the current IP multicast architecture, a multicast group is identified by an IP address of a special class — the class D IP addresses that have the first 4 bits as 1110 (thus a multicast address is within the range of 224.0.0.0–239.255.255.255). IP multicast follows a very simple model:[1] (1) a host that joins a group with a specific group address shall receive any packet sent to that group and (2) a packet sent to a group address from any host shall be received by all members of that group. Next we discuss how (1) and (2) would be implemented in IP networks.

2.2. Multicast at LAN Level

At LAN level, a host joins a multicast group by communicating with the designated router (DR) via the Internet Group Membership Protocol (IGMP [6,15]): (1) a host can send a message (membership report) to the DR to join a specific group; (2) the DR may periodically send query messages for a group, and a host can answer this query to express continuous interest in that group or silently drop the message to indicate that it is no longer interested in that group; or (3) a host may send "leave group" message to the DR to explicitly leave a group. A host operating system supporting multicast provides API (application programming interface) functions for applications to join or leave multicast groups.

The question now is how a host sends or receives multicast traffic. We use an Ethernet network as an example. To send IP packets for a group, a host just puts the group IP address as the destination address and sends it to the DR encapsulated in an Ethernet frame with DR's MAC address, which is the same as sending a unicast packet. A DR is responsible for sending it out to reach other group members (which we discuss next). A host doesn't need to join a group to send data to that group.

Receiving multicast packets is done through mapping of an IP multicast address to a MAC address. The IETF (Internet Engineering Task Force [1]) has a single 802 MAC address block (01-00-5E-00-00-00) with the lowest 24 bits assignable. A multicast IP address has 28 unique bits (the highest 4 bits are 1110), the lowest 23 bits of

[1] Some applications may enforce some kind of access control; thus these two rules may not always apply. Initially IP multicast was designed to support only the model at the network layer; thus such access control has to be implemented elsewhere. More recently, source-specific control was introduced into IP multicast [6]. Now, say, a host can join a group while specifying that it is interested only in receiving packets from a list of sources.

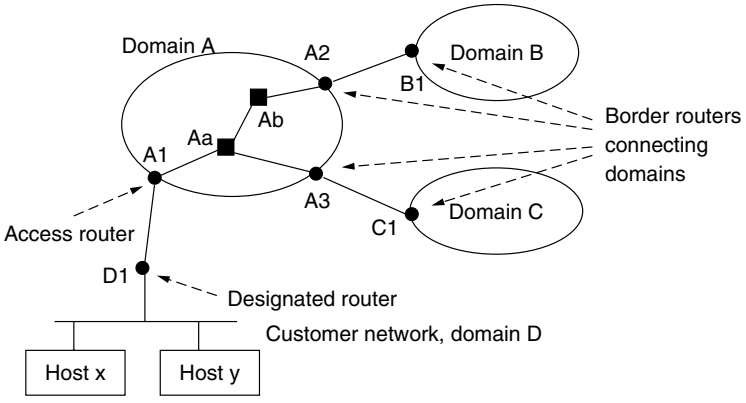


Figure 4. Internet hierarchy. The routers shown here together form a multicast tree.

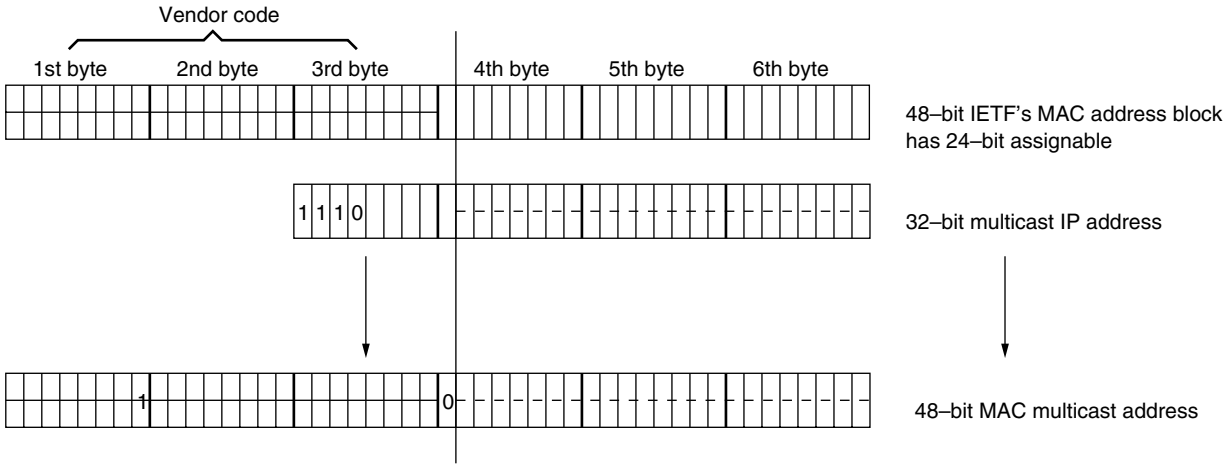


Figure 5. Map of a multicast IP address to a multicast MAC address. Half of IETF's address block (the 24th bit is 0) is used for multicast; the other half (the 24th bit is 1) is reserved.

that 28 bits are mapped into the lowest 23 bits of IETF's address space to form a MAC multicast address for that group, as illustrated in Fig. 5. When a host joins a group, it informs the network interface card to listen to the mapped MAC address. A DR sends multicast packets received from external routers for that group to the LAN network by encapsulating them in an Ethernet frame with the mapped address. One may note that, since an IP multicast address has 28 unique bits, there may be two or more different groups mapped into a single MAC address at the same time within a LAN. This won't be a problem since the original multicast address is carried in the IP header and the receiving host can just discard packets for groups from which it does not intend to receive data.

2.3. Multicast over Wide-Area Networks

Wide-area networks (WANs) are often point-to-point networks. IP multicast utilizes a tree structure to deliver multicast packets across WANs as described earlier. A tree consists of designated routers that have group members in their subnets and other intermediate routers that help transport multicast traffic in between. For example, as shown in Fig. 4, a multicast tree may consist of router $D_1, A_1, A_2, A_3, A_a, A_b, B_1, B_2$, and some other DR routers in domains B and C.

Once a router is in the tree for a specific group, it uses forwarding-state information to determine how to forward multicast packets received. In source-tree based multicast routing protocols, forwarding-state entries are intended for per group/source. For example, a forwarding entry at a router is like *group:g / source:s, expected in — interface:I, out — interface*(s):$O_1 \cdots O_n$. When this router receives a packet with destination address g and source address s from interface I, it will send the packet out to interface $O_1, \ldots, O_n$. In group shared-tree protocols, there is one entry for one group. For example, a forwarding entry will be *group:g, list of interfaces*: $I_1, I_2, \ldots, I_n$. A packet with destination address g received from interface I_{in} will be sent out to all other interfaces in the list except I_{in}.

Multicast routing protocols determine how a multicast tree is constructed. Matching the Internet hierarchy, routing protocols are divided into intradomain protocols and interdomain protocols. Normally a domain [often called an *autonomous system* (AS)] is controlled by a single administrative entity and can run an intradomain multicast routing protocol of its choice. An interdomain multicast routing protocol is deployed at border routers of a domain to construct multicast trees connecting to other domains. A border router capable of multicast communicates with its peer(s) in other domain(s) via interdomain multicast protocols and routers in its own via intradomain protocols, and forwards multicast packets across the domain boundary.

3. STEINER TREE PROBLEM

Multicast routing algorithms are closely related to the *Steiner tree problem* [20] in graph theory [21]. The Steiner tree problem is related to the *minimum spanning-tree* problem, which can be stated as follows. Given a connected graph $G(V, E)$, where V is a set of vertices(nodes) and E is a set of edges, and for each edge $e \in E$ it has a weight $w(e)$, a minimum spanning tree is a tree $T(V, E')$ $(E' \subset E)$, which contains all the nodes in G and its weight $W(T) = \sum_{e \in T} w(e)$ is minimal of all possible spanning trees. The minimum-weight spanning tree that covers a subset of V is a Steiner tree: tree $T(V', E')$ with minimal weight $W(T)$ for a given $V' \subset V$.

There are two classical algorithms for the minimum spanning-tree problem [33]: (1) *Prim's algorithm*, which starts with an arbitrary node and grows the tree by repeatedly adding the minimum-weight edge that connects an in-tree node to a node that is not yet in the tree until all nodes are connected; and (2) *Kruskal's algorithm*, which initially has each node as a separate tree and then constructs the Steiner tree through merging them into one by repeatedly adding the minimum-weight edge that connects two trees (into one) without creating a cycle.

The Steiner tree problem is NP-hard in general, which means that the time it takes to find exact solutions is exponential regarding graph size. A number of heuristic algorithms have been proposed [16,20] to find approximate

solutions in P time. A shortest-paths heuristic proposed by Takahashi and Matsuyama (TM algorithm) [20,34] is a simple heuristic that has a proven good performance bound. Based on a greedy strategy, it starts with a subtree T consisting of a single node arbitrarily chosen from V'. The tree T grows by adding the node from V' that has the shortest distance to nodes in T and is not covered by T, one by one until all nodes of V' are present in T.

In another heuristic proposed by Kou, Markowsky, and Berman (KMB algorithm) [20], first a complete graph $G'(D, F)$ is constructed with nodes $D = V'$, using the cost of the shortest path from i to j in G as the cost for edge $e_{ij} \in F$. Then the minimum spanning tree T' of G' is built. The spanning tree in G covering V' is obtained by replacing any edge (i, j) in G' with the shortest path $p(i \to j)$ in G.

In multicast routing, the goal is to construct a multicast tree that covers all the designated routers ("terminal nodes") for a multicast group so that all members can receive data sent to that group. On the other hand, one motivation for multicast is network resource (i.e., bandwidth) efficiency. Therefore we want to build a multicast tree that minimizes resource usage. For that purpose, we can assign to each link a weight that represents the cost to transport a unit of data over that link (i.e., the cost to use the bandwidth resource). Thus the problem of constructing a multicast tree that minimizes resource usage is to find a Steiner tree that covers all the terminal nodes of a group.

4. INTRADOMAIN ROUTING PROTOCOLS AND ALGORITHMS

From our presentation of the two Steiner tree algorithms, we can see that there are three elements in multicast routing: (1) network information [network nodes and how they are connected, i.e., $G(V, E)$], (2) group membership information (the set of nodes that we need to cover, V'), and (3) an algorithm to construct the tree.

None of the existing multicast protocols actually uses the TM algorithm or KMB algorithm in Section 3 because of their two requirements, as follows: (1) complete network topology information is needed [i.e., $G(V, E)$] and (2) all terminal nodes must be known in advance (i.e., V'). These two requirements make it difficult or impossible to implement either algorithm in a distributed routing environment like the Internet. At the same time, the IP multicast model assumes dynamic membership (i.e., members can join or leave at any time) and supports nonmember sending (i.e., a host can send data to a group without joining it). The implication of requirement 2 is that whenever there is a membership change, a multicast tree must be recomputed. To support this, it is more desirable to have a routing algorithm that would construct a multicast incrementally as members join a group and introduce minimal modification to the existing tree when a member leaves the group. At the same time, nonmember sending has to be supported. Nevertheless, optimal Steiner trees produced by good approximate algorithms are often used to compare with those constructed by IP multicast protocols in performance studies.

4.1. MOSPF: Shortest-Path Tree

MOSPF [27,28] is an extension of the Open Shortest Path First (OSPF) protocol to support multicast routing in an intradomain environment. OSPF is a link-state routing protocol [19,29], at the core of which there is a distributed and replicated link-state database at each router in an AS. This database is like a map of the network describing all routers within and their interconnections. It is constructed and updated through flooding of link-state advertisements (LSAs). Each LSA describes the links of a node to its neighbors (containing additional information such as *cost* of a link). For example, an LSA from router S may look like $\{S \to A: 2, S \to B: 1, S \to C: 3\}$, where the numbers are cost of the links, respectively. We also assume symmetric links here — for instance, there is also a link $(A \to S)$ with cost of $(A \to S)$ = cost of$(S \to A)$. Each router gathers LSAs from all other nodes in the network and constructs a complete topology of the network. On the basis of that topology, a router computes a routing table based on which to forward data packets. For example, an entry in the routing table could be like {131.169.96.0/24,[2] interface I}, which tells the router to forward a data packet received with destination address in the range from 131.169.96.0 to 131.169.96.255 to the interface I (which connects to a particular neighbor, the next hop for those packets). The algorithm used in OSPF is Dijkstra's algorithm [25], which computes the shortest paths from one node to all others. For example, assume that router F is the designated router for a subnetwork 131.169.96.0/24; node S computes the shortest path to F: $(S \to B \to D \to F)$; this tells S to forward packets destined to network 131.169.96.0/24 to neighbor B.

MOSPF extends OSPF to support multicast routing: (1) a new LSA is introduced to propagate group membership information, then (2) a router computes a multicast tree and determines the forwarding entry when it receives multicast packets for a group from a specific source. For example, when a host in subnet 131.169.96.0/24 sends an IGMP message to F to join a group g_x, F will flood the network with an LSA saying that subnetwork 131.169.96.0/24 is now a member of group g_x. We also assume that there are group members in router E's and G's subnets. Now a host in router S's subnet is a source sending traffic to group g_x. When S receives the first packet from that host to g_x, it computes all the shortest paths to all other nodes; then it recursively prunes routers that don't have any group members to get a multicast tree. This is illustrated in Fig. 6, where all nodes shown are routers; hosts are not shown since they are not involved in the routing process. From that tree, S knows that it should forward the packet to B. When B receives the packet, it computes a tree using its database of the network topology — the database is synchronized and the tree computed will be the same as S's, and B determines that it should forward the packet to D and C. Forwarding entries are cached. So when S receives the next packet from the same source to g_x, it knows how

[2] The number 24 means that the highest 24 bits are significant bits fixed as 131.169.96, while the lowest 8 bits can vary from 0 to 255 in value.

to forward it without computing the tree again. A cached entry also has a lifetime — it expires after a while and a router will compute it again; this way, a router doesn't waste bandwidth to forward multicast packets of group g_x to neighbors that no longer lead to group members. For example, after a while E is no longer a member of group g_x and wouldn't send out an LSA saying that it is; when the forwarding entry expires, B recomputes the tree and knows that it no longer needs to forward packets destined for g_x to E.

4.2. DVMRP: Broadcast and Prune

The Distance Vector Multicast Routing Protocol (DVMRP) [30] was developed to support multicast routing in an intradomain environment where a routing protocol of the *distance vector* protocol family is deployed. In the Internet, the Routing Information Protocol (RIP) [18,19] is a distance vector protocol that was widely deployed and used for intradomain routing before OSPF was developed. Instead of flooding link-state information, nodes exchange "distance" (to all other nodes) information with neighbors in RIP, and a distributed version of the Bellman–Ford shortest-path algorithm [9] is implemented for a node to figure out how to reach any other node. Unlike OSPF, in which each node maintains a complete topology of the network, in RIP a node only knows what next hop to go to reach a destination. For example, in the network shown in Fig. 6a, S knows that it should forward packets destined for node F's subnet to its neighbor B, while F knows that it should forward packets destined for node S's subnet to D. Similarly, B knows that it should forward packets destined for node S's subnet to S.

DVMRP employs a "broadcast and prune" approach to construct multicast trees. In MOSPF, when a node receives a multicast packet and doesn't know how to forward it [i.e., there is no forwarding entry for the (source, group) pair], it computes a shortest-path tree (rooted at the source) using the topology database and determines the forwarding entry. In DVMRP, when a node receives a multicast packet with source address s and destination (group) address g_x and it knows nothing about it, it does the following: (1) it checks whether the packet is received from the interface to which it normally forwards packets destined for s and (2) if not, the packet is dropped; otherwise the packet is forwarded to all other interfaces except the one from which the packet is received. For example, in Fig. 6, let's assume that the source host with IP s is in router S's subnet; when D receives a packet of (s, g_x) from node A, it will drop it because D always forwards packets destined for s to B instead of A; however, if that packet is received from B, D will forward it to nodes A, F, G, and E. For F and G, this is great, because both F and G have group members in their subnets. However, A and E don't have member in their subnets. What A or E does is to send a "prune" message to D to tell D not to forward multicast packets (s, g_x) to them anymore. D will remember that by adding a cache entry and will forward packets (s, g_x) to F and G only in the future. Similarly, when B receives a multicast packet with (s, g_x) from S, it forwards it to A, D, E, and C. After it receives prune messages from A and C, it remembers it and no longer forwards packets to them. But D will not send a prune message to B because D has nodes F and G at downstream and these nodes don't tell D that they don't want the packets (which translates into "they want those packets"). This way, eventually a multicast tree rooted at S and reaching F, G, and E is built. The cached prune information is periodically cleared. For example, after a while B will "forget" that it doesn't need to forward multicast packets (s, g_x) to A and C and again forwards them to all neighbors except S; then A and C will send prune messages again to "opt out" of the tree.

In DVMRP, a multicast tree is a reverse shortest-path tree rooted at the source compared with the shortest-path tree in MOSPF. In MOSPF, the path from source S to a destination F on the tree is the shortest path from S to F; in DVMRP, however, the path from F to S on the tree is the shortest path but the path from S to F is not necessarily the shortest (thus the term *reverse shortest-path tree*). In our example, they happen to be the same because we assume every link to be symmetric. In the modern Internet, "asymmetric" routing may happen (i.e., packets from F to S may travel a different path than packets from S to F) and the reverse shortest-path tree may not always be the same as the shortest-path tree.

4.3. PIM-SM: Reverse Shortest-Path Tree

There are some limitations with DVMRP and MOSPF. DVMRP is not very efficient because of the periodical flooding of multicast packets. This also leads to a scalability problem — when the network grows larger or when the number multicast groups grows larger, the efficiency problem becomes more severe. MOSPF also has a scalability problem; when the network size and/or the number of groups grows large, the processing requirement to compute trees may become excessive for a router. Another limitation is they are only

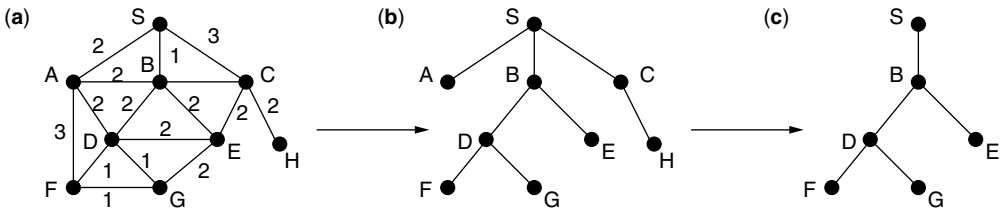


Figure 6. Construction of a shortest-path multicast tree: (**a**) network with link cost; (**b**) shortest-path tree from node S to all others; (**c**) get the multicast tree by recursively pruning nodes not interested in the group; remove node A and link (S, A), remove node H and link (C, H), remove node C and link (S, C).

applicable to intradomain routing. For these reasons, the Protocol Independent Multicast–Sparse Mode (PIM-SM) protocol [10,11,13] was developed.

PIM-SM also builds a reverse shortest-path tree. However, it is significantly different from DVMRP: (1) the tree is rooted at a central router called the *rendezvous point* (RP); (2) instead of broadcast-and-prune, it employs an explicit join mechanism to construct the tree. In PIM-SM, every multicast group has a central router (RP) responsible for that group. The RP is identified by its own unicast IP address and the RP for a group can be obtained or made known through a query or advertisement mechanism. Beyond that, similar to DVMRP, PIM-SM assumes that a router can provide the next hop information for any destination — this is easily and readily supported by any unicast routing protocol (this is why it is called protocol-independent; it doesn't rely on any specific unicast routing protocol). When a designated router (e.g., router S in Fig. 6) receives a membership report for a group from a host in its subnet, it sends a PIM-SM join request toward the RP router of the group. A PIM-SM capable (i.e., understands and supports PIM-SM) router will look up the next hop for the RP's address (a unicast address) and forward it to the next PIM-SM-capable router. That join request will stop when it reaches either (1) a router that already has multicast state (i.e., forwarding entry) for that group or (2) the RP router of the group. In either case, intermediate routers will install forwarding entries for the group and a new branch connecting to the new member is established. As an example, assume that router B is the RP for group g_x. When F first joins the group, it sends a join message destined for B to D (D is the next-hop router to which F should forward packets to reach B); D doesn't have a forwarding entry for g_x yet, it will create one and forward the request to B, which doesn't need to forward the message further. When node G wants to join the group, it sends a join request to D; D already has a forwarding entry for g_x, and will simply add G to the existing entry. When a host sends a packet to a group, its DR router will send it as unicast packet (through a technique called "encapsulation") to the RP router of the group, then the RP router sends the packet as a multicast packet along the tree established. For example, the host s sends a packet to group g_x; its DR router S sends the packet as a unicast packet to the RP router B. B has forwarding entry for g_x and forwards it to D and E, and D in turn forwards it to F and G.

PIM-SM is similar to a server-based solution to some degree. The main difference is that when packets reach the server (the RP router in this case), they are distributed to all members through a tree instead of many unicast connections. PIM-SM also has some other advanced features such as providing support to switch to a source-specific tree (i.e., a multicast tree rooted at the source for a particular source). We won't discuss these features here, and interested readers can refer to the related literature.

4.4. Core-Based Tree (CBT)

All the protocols discussed above build unidirectional multicast trees; at a tree router, multicast packets are expected to arrive from one interface and will be sent out to a list of outgoing interfaces. In PIM-SM, a tree is shared — all source nodes send packets to the RP and the RP sends them over the tree. However, DVMRP and MOSPF use source-specific trees. Thus, if there are multiple sending sources for a group, then multiple trees must be established. For example, if a host at A's subnet sends a packet to group g_x, a new tree rooted at A will be established in MOSPF or DVMRP. The implication is a scalability issue — the more trees, the more forwarding entries a router has to maintain and the more processing overhead for multicast forwarding. Another protocol, Core-Based Tree (CBT) [3,4] builds a single bidirectional shared tree for a group. Although this protocol hasn't been and may never be widely deployed, its idea has been shared in PIM-SM, and a newer interdomain multicast routing protocol called *Border Gateway Multicast Protocol* (BGMP) is based on it.

Similar to PIM-SM, CBT has a core router for a group and uses an explicit join mechanism for tree construction. When a router joins a group, it sends an explicit join request message toward the core until the message reaches the core or a node that is already in the tree, and then a new branch is established. Forwarding entry at a router is for per group (i.e., *g, list_of_interfaces*) (Section 2.3). When a node want to sends a packet to a group, it sends it toward the core until it reaches a node that is already in the tree, and the packet will travel in the multicast tree as a multicast packet from that point. The main difference compared with PIM-SM is that the packet doesn't need to reach the core.

5. INTERDOMAIN MULTICAST ROUTING PROTOCOLS AND ALGORITHMS

Two of the above routing protocols (DVMRP, MOSPF) are for intradomain multicast only. The other two (CBT and PIM-SM) were actually designed to support interdomain multicast as well — they both require only the underlying unicast routing protocol to provide a next hop to a core or RP router and that is readily supported by both intradomain routing protocols and interdomain routing protocols [e.g., Border Gateway Protocol (BGP) [19]]. However, because of some limitations and other emerging problems as Internet multicast grows out of the old experimental multicast backbone (MBone) [2], there is a pressing need to develop new protocols to better support multicast at the interdomain level.

Over the years, several protocols have been developed and considered by IETF to provide scalable hierarchical Internetwide multicast. The first step toward scalable hierarchical multicast routing is Multiprotocol Extensions to BGP4 (MBGP) [5], which extends BGP to carry multiprotocol routes (i.e., besides the "traditional" IP unicast routes). In the MBGP/PIM-SM/MSDP architecture [2], MBGP is used to exchange multicast routes and PIM-SM is used to connect group members across domains, while another protocol, Multicast Source Discovery Protocol (MSDP) [14], was developed to exchange information of active multicast sources among RP routers across domains.

The MBGP/PIM-SM/MSDP architecture has scalability problems and other limitations, and is recognized as

a near-term solution [2]. To develop a better long-term solution, a more recent effort is the MASC/BGMP architecture [24].

5.1. The MASC/BGMP Architecture

In the MASC/BGMP [24,37] architecture, border routers run Border Gateway Multicast Protocol (BGMP) to construct a bidirectional "shared" tree similar to a CBT tree for a multicast group. The shared tree is rooted at a "root domain" (instead of a single core router) that is mainly responsible for the group (e.g., the domain where the group communication initiator resides). To solve the difficult problem of mapping a multicast group to a RP or core router associated with PIM-SM or CBT, BGMP relies on a hierarchical multicast group address allocation protocol called the *Multicast Address-Set Claim Protocol* (MASC) to map a group address to a root domain and an interdomain routing protocol (BGP/MBGP) to carry "group route" information (i.e., how to reach the root domain of a multicast group).

MASC is used by one or more nodes of a MASC domain to acquire address ranges to use in a domain. Within the domain, multicast addresses are uniquely assigned to clients using an intradomain mechanism. MASC domains form a hierarchical structure in which a "child" domain (customer) chooses one or more "parent" (provider) domains to acquire address ranges using MASC. Address ranges used by top-level domains (domains that don't have parents) can be preassigned and can then be obtained by child domains. This is illustrated in Fig. 7, in which A, D, and E are backbone domains, B and C are customers of A, while B and C have their own customers F and G, respectively. A has already acquired address range 224.0.0.0/16 from which B and C obtain address ranges 224.0.128.0/24 and 224.0.1.1/25, respectively.

Using this hierarchical address allocation, multicast "group routes" can be advertised and aggregated much like unicast routes. For example, border router B_1 of domain B advertises *reachability* of root domains for groups in the range of 224.0.128.0/24 to A_3 of domain A, and $A_1(A_4)$ advertises the aggregated 224.0.0.0/16 to $E_1(D_1)$ in domain $E(D)$. Group routes are carried through MBGP and are injected into BGP routing tables of border routers. BGMP then uses such "group routing information" to construct shared multicast trees to distribute multicast packets.

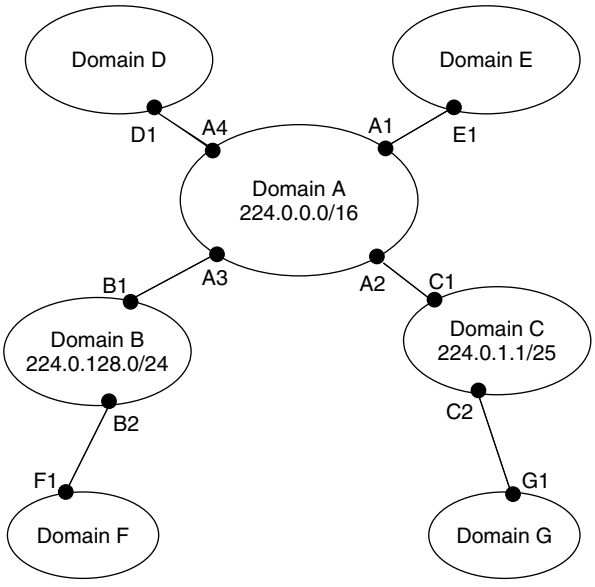


Figure 7. Address allocation using MASC, adopted from Ref. 24.

BGMP constructs a bidirectional shared tree for a group rooted at its root domain through explicit join/prune as in CBT. An example tree is illustrated in Fig. 8. A BGMP router in the tree maintains a *target list* that includes a *parent target* and a list of *child targets*. A parent target is the next-hop BGMP peer toward the root domain of the group. A child target is either a BGMP peer or an MIGP (Multicast Interior Gateway Protocol, i.e., MOSPF or PIM-SM) component of this router from which a join request was received for this group. For example, assume domain B is the root domain, then at node C_2, the parent target is node A_2 and a child target may be an interface of its own that connects to another tree router in its domain. Data packets received for the group will be forwarded to all targets on the list except the one from which the data packet originated. BGMP router peers maintain persistent TCP connections with each to exchange BGMP control messages (join/prune, etc.).

In the BGMP architecture, a source doesn't need to join the group in order to send data. When a BGMP router receives data packets for a group for which it doesn't have a forwarding entry, it will simply forward packets to the next-hop BGMP peer toward the root domain of the group. Eventually they will hit a BGMP router that has a forwarding state for that group or a BGMP router in the root domain. For example, if a node in domain D wants to send a packet to group 224.0.128.5, based on the group address (belonging to domain B), the packet will be sent to node D_1, then to A_4, and then A_3. Node A_3 is already in the tree; thus it will forward the packet over the multicast tree (to its interdomain peer B_1 and interior neighbor A_a). BGMP can also build source-specific branches, but only when needed (i.e., to be compatible with

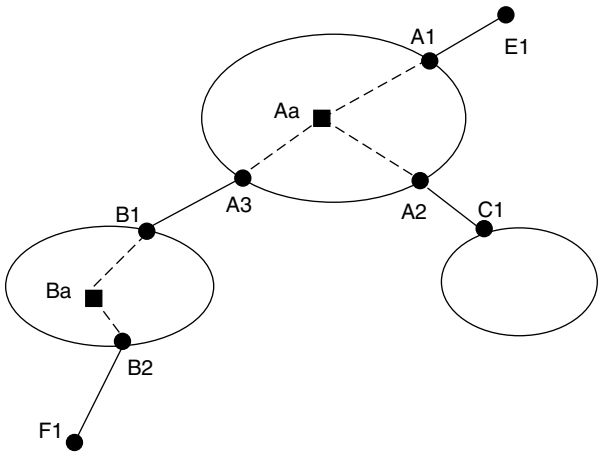


Figure 8. An interdomain multicast tree (solid lines are tree links). Within a domain, an intradomain multicast routing protocol builds an intradomain multicast (dashed lines).

source-specific trees used by some intradomain multicast protocols such as DVMRP and MOSPF), or to construct trees for source-specific groups.

6. CONCLUSIONS

Today, many applications involve group communications. Multicast was conceived as a mechanism to efficiently support such a communication need. However, multicast support at the network level is only rudimentary in traditional telephone networks. Though extensive work on multicast in IP networks has been started since the early 1980s, widespread availability of multicast service in the Internet is still not any time soon. Much research and engineering work is still to be done. This article focused mostly on multicast in IP networks and the Internet, especially routing protocols and algorithms that constitute the core of multicast support.

The multicast routing problem can theoretically formulated as the Steiner tree problem. However, because of the routing environment and some network requirements, traditional Steiner tree algorithms are not readily applicable for multicast routing. Several routing algorithms and protocols have been developed, and some are standardized by the IETF for the Internet. They include DVMRP, MOSPF, CBT, PIM-SM, and the more recent MBGP. In this article, we described a big picture of IP multicast and then gave an overview of those protocols and algorithms. Interested readers can refer to the references cited for more detailed specific information on any of them. This field is still under very active development, and many new advances are being made. Interested readers may refer to the most recent research literature and IETF documents for more recent developments.

BIOGRAPHIES

Aiguo Fei (afei@acm.org) received the B.S. degree in physics in 1995 from Fudan University, Shanghai, China; the M.S. degree in physics and the M.S. and Ph.D. degrees in computer science in 1996, 1998, and 2001, respectively, from University of California, Los Angeles. He joined a startup company in Silicon Valley, California in 2001 as a research engineer. His area of interests are multicast in IP networks, QoS support in next-generation IP networks, network and graph algorithms, and network intrusion detection and statistical anomaly.

Mario Gerla (gerla@cs.ucla.edu) was born in Milan, Italy. He received a graduate degree in engineering from the Politecnico di Milano, in 1966, and the M.S. and Ph.D. degrees in engineering from UCLA in 1970 and 1973, respectively. He joined the Faculty of the UCLA Computer Science Department in 1977. His research interests cover the performance evaluation, design, and control of distributed computer communication systems; high-speed computer networks; wireless LANs (Bluetooth); and ad hoc wireless networks. He has been involved in the design, implementation, and testing of wireless ad hoc network protocols (channel access, clustering, routing, and transport) within the DARPA WAMIS, GloMo projects and most recently the ONR MINUTEMAN project. He has also carried out design and implementation of QoS routing, multicasting protocols, and TCP transport for the next-generation Internet (see www.cs.ucla.edu/NRL for the most recent publications).

BIBLIOGRAPHY

1. Internet Engineering Task Force. *http://www.ietf.org/*.
2. K. Almeroth, The evolution of multicast: From the MBone to inter-domain multicast to Internet2 deployment, *IEEE Network* (Jan./Feb. 2000).
3. A. Ballardie, *Core Based Trees (CBT version 2) Multicast Routing: Protocol Specification*, IETF RFC 2189, Sept. 1997.
4. A. Ballardie, P. Francis, and J. Crowcroft, Core based trees (CBT), *Proc. ACM SIGCOMM'93*, Sept. 1993, pp. 85–95.
5. T. Bates, R. Chandra, D. Katz, and Y. Rekhter, *Multiprotocol Extensions for BGP-4*, IETF RFC 2283, Feb. 1998.
6. B. Cain et al., Internet group management protocol, version 3, *Internet draft: draft-ietf-idmr-igmp-v3-07.txt*, March 2001.
7. Ch.-H. J. Wu, and J. D. Irwin, *Emerging Multimedia Computer Communication Technologies*, Prentice-Hall, 1998.
8. D. E. Comer, *Computer Networks & Internets with Internet Applications*, 3rd ed., Prentice-Hall, 2001.
9. T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, The MIT Press, 1990.
10. S. Deering, D. Estrin, D. Farinacci et al., Protocol independent multicast-sparse mode (pim-sm): motivation and architecture, *IETF Internet draft: draft-ietf-idmr-pim-arch-05.txt{ps}*, Aug. 1998.
11. S. Deering et al., The pim architecture for wide-area multicast routing, *IEEE/ACM Trans. Network.* **4**(2): 153–162 (April 1996).
12. C. Diot, W. Dabbou, and J. Crowcroft, Multipoint communication: A survey of protocols, functions, and mechanisms, *IEEE J. Select. Areas Commun.* **15**(3): 277–290 (April 1997).
13. D. Estrin et al., *Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification*, IETF RFC 2362, June 1998.
14. D. Farinacci et al., Multicast source discovery protocol (msdp), *IETF Internet draft: draft-ietf-msdp-spec-06.txt*, 2000.
15. W. Fenner, *Internet Group Management Protocol, Version 2*, IETF RFC 2236, 1997.
16. Joe Ganley, *http://ganley.org/steiner/*.
17. M. Goncalves and K. Niles, *IP Multicasting: Concepts and Applications*, McGraw-Hill, 1999.
18. C. Hedrick, *Routing Information Protocol*, IETF RFC 1058, 1988.
19. C. Huitema, *Routing in the Internet*, Prentice-Hall, 1995.
20. F. Hwang, D. Richards, and P. Winter, *The Steiner Tree Problem*, Elsevier, 1992.
21. D. Jungnickel, *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics, Springer, 1999.
22. S. Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network*, Addison-Wesley, 1997.

23. D. Kosiur, *IP Multicasting*, Wiley, 1998.
24. S. Kumar et al., The MASC/BGMP architecture for inter-domain multicast routing, *Proc. ACM SIGCOMM'98*, Sept. 1998, pp. 93–104.
25. U. Manber, *Introduction to Algorithms: A Creative Approach*, Addison-Wesley, 1989.
26. C. K. Miller, *Multicast Networking and Applications*, Addison-Wesley, 1999.
27. J. Moy, Multicast routing extensions for ospf, *Commun. ACM* **37**: 61–66 (Aug. 1994).
28. J. Moy, *Multicast Routing Extensions to OSPF*, RFC 1584, March 1994.
29. J. Moy, *Ospf Version 2*, IETF RFC 2328, April 1998.
30. C. Partridge, D. Waitzman, and S. Deering, *Distance Vector Multicast Routing Protocol*, RFC 1075, 1988.
31. R. Perlman, *Interconnections: Bridges, Routers, Switches, and Internetworking Protocols*, 2nd ed., Addison-Wesley, 1999.
32. L. L. Peterson and B. S. Davie, *Computer Networks: A Systems Approach*, 2nd ed., Morgan Kaufmann, 1999.
33. S. S. Skiena, *The Algorithm Design Manual*, Springer, 1997.
34. H. Takahashi and A. Matsuyama, An approximate solution for the Steiner problem in graphs, *Math. Jpn.* **24**: 573–577 (1980).
35. A. S. Tanenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, 1996.
36. R. E. Tarjan, *Data Structures and Network Algorithms*, Society for Industrial and Applied Mathematics, 1983.
37. D. Thaler, D. Estrin, and D. Meyer, Border gateway multicast protocol (BGMP): Protocol specification, *IETF Internet draft: draft-ietf-bgmp-spec-02.txt*, Nov. 2000.
38. R. Wittmann and M. Zitterbart, *Multicast Communication: Protocols and Applications*, Morgan Kaufmann, Academic Press, San Francisco, 2001.

MULTIDIMENSIONAL CODES

JOHN M. SHEA
TAN F. WONG
University of Florida
Gainesville, Florida

1. INTRODUCTION

The term *multidimensional codes* is used in several different contexts relating to modern communications. For instance, trellis coding with multidimensional modulation [1–3] is sometimes referred to as *multidimensional coding*. Trellis coding with multidimensional modulation uses multiple modulation symbols that map to a symbol of greater than two dimensions. Certain algebraic geometry codes that are defined using projective algebraic curves over Galois fields [4] are also referred to as *multidimensional codes*. Other types of codes that are multidimensional in nature are the two-dimensional burst identification codes of Abdel-Ghaffar et al. [5] and the two-dimensional dot codes of van Gils [6].

In this article, we consider multidimensional codes in which the bits are encoded by a series of orthogonal parity checks that can be represented as coding in different dimensions of a multidimensional array. In Section 2, we provide a brief introduction to product codes and some of their properties. In Section 3, we discuss the properties of product codes constructed from single parity-check codes, and in Section 4 we discuss soft-decision decoding of these codes. Finally, in Sections 5–7, we present a class of multidimensional codes and provide performance results for two applications of these codes.

2. PRODUCT CODES

Product codes were introduced by Elias in 1954 [7] as a way to develop a code that could achieve vanishingly small error probability at a positive code rate. The scheme proposed by Elias uses an iterative coding and decoding scheme in which each decoder improves the channel error probability for the next decoder. In order to ensure that any errors at the outputs of one decoder appear as independent error events in each codeword input to the next decoder, Elias proposed encoding each information bit using a series of orthogonal parity checks. He termed this technique "iterative" coding. His coding scheme is now commonly referred to as a *product code*. In particular, the coding scheme he proposed is a systematic, multidimensional product code in which the number of dimensions can be chosen to achieve arbitrarily low bit error probability.

Product codes are the most common form of multidimensional code. A product code of dimension p is generated in such a way that each information bit is encoded p times. Product codes are typically formed using linear block codes [8]. Suppose that we have p (not necessarily different) block codes $C_1, C_2, \ldots, C_p$ with blocklength $n_1, n_2, \ldots, n_p$ and information length $k_1, k_2, \ldots, k_p$. Then the p-dimensional product of these codes is a block code C, with blocklength $n = n_1 n_2 \cdots n_p$ and information length $k_1 k_2 \cdots k_p$. The constituent codes $\{C_i\}$ are said to be subcodes of C [9].

For a two-dimensional product code, the code can be visualized as a rectangular array in which each column is a codeword in C_1 and each row is a codeword in C_2. Let n_i denote the blocklength of code C_i, and let k_i denote the number of information bits conveyed by each codeword of code C_i. We say that code C_i is a (n_i, k_i)-block code. Suppose that C_1 and C_2 are systematic codes. A diagram that illustrates the construction of Elias' systematic two-dimensional product code is shown in Fig. 1. One possible way to encode the information is as follows:

1. Place the information bits in the $k_1 \times k_2$ submatrix.
2. For each of the first k_2 columns, calculate $n_1 - k_1$ parity bits using code C_1 and append those to that column.
3. For each of the first k_1 rows, calculate $n_2 - k_2$ parity bits using code C_2 and append those to that row.
4. For each of the last $(n_1 - k_1)$ rows, calculate $(n_2 - k_2)$ parity bits from code C_2 and append those to that row. This last set of parity bits uses the parity bits from code C_1 and encodes them with code C_2, and are thus known as *parity-on-parity* bits.

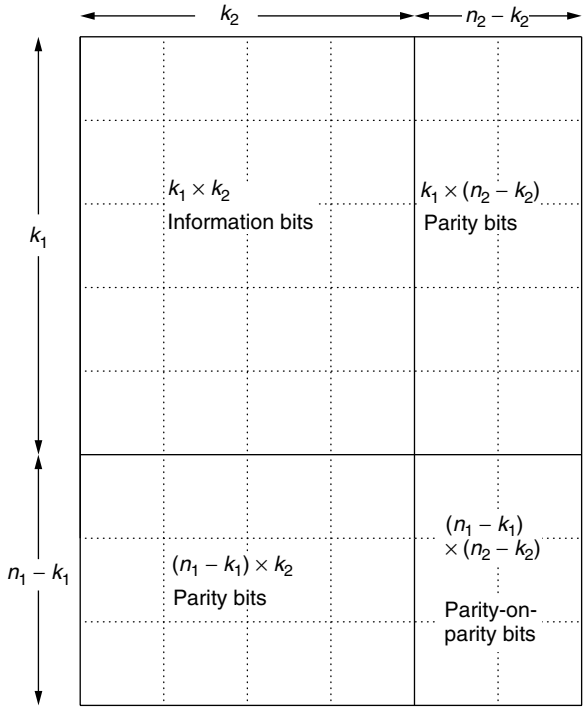


Figure 1. A two-dimensional product code.

Note that in step 4, the parity-on-parity bits have the same values if they are instead constructed using code C_1 on the parity bits from code C_2.

Product codes have many properties that are derived from their subcodes. Some of the most commonly used block codes are the cyclic codes [8]. If $\mathbf{v}$ is a codeword of a cyclic code C, then any cyclic shift of $\mathbf{v}$ is also a codeword of C. Peterson and Weldon [10] proved the following theorem for two-dimensional product codes constructed from cyclic subcodes.

Theorem 1. Suppose that C_1 and C_2 are cyclic codes with length n_1 and n_2, where n_1 and n_2 are relatively prime. Let $i_1 \equiv i \bmod n_1$, and let $i_2 \equiv i \bmod n_2$, for $i = 0, 1, 2, \ldots, n_1 n_2 - 1$. Then the product of C_1 and C_2 is a cyclic code if the codeword $\mathbf{v} = (v_0, v_1, \ldots, v_{n_1 n_2 - 1})$ is constructed such that the symbol v_i is the symbol in the (i_1, i_2)th position of the rectangular array representation.

The mapping from i to (i_1, i_2) results in a cyclic enumeration of i that wraps around the edges of the $n_1 \times n_2$ rectangular array. For example, consider $n_1 = 5$ and $n_2 = 3$. Then the positions of i in the rectangular array are as shown in Fig. 2a. Note that a right cyclic shift of the codeword $\mathbf{v}$ corresponds to a right cyclic shift and downward cyclic shift of the rectangular matrix, as is illustrated in Fig. 2b. Thus, a cyclic shift of the codeword $\mathbf{v}$ results in another valid codeword.

Products of more than two codes are also cyclic codes under similar constructions (see Corollary II of Ref. 9). Furthermore, the generator polynomial [8] for the product code is shown to be a simple function of the generator polynomials for the subcodes. Let $g_i(X)$ be the generator polynomial for the ith subcode, and let $g(X)$ be the

(**a**)

0	10	5
6	1	11
12	7	2
3	13	8
9	4	14

(**b**)

14	9	4
5	0	10
11	6	1
2	12	7
8	3	13

Figure 2. Bit ordering to convert 5×3 product code to a cyclic code: (**a**) original order; (**b**) order after right cyclic shift.

generator polynomial for the product code. The generator polynomial for the two-dimensional product code is given by (see Theorem III of Ref. 9)

$$g(X) = \text{GCD}\ \{g_1(X^{bn_2}) g_2(X^{an_1}), X^{n_1 n_2} - 1\}$$

where $\text{GCD}(y, z)$ is the greatest common divisor of y and z, and a and b are integers satisfying $an_1 + bn_2 \equiv 1 \bmod n_1 n_2$. For example, for the 5×3 code of Fig. 2, $a = 2$ and $b = 2$ will satisfy the equation $(2)(5) + (2)(3) \equiv 1 \bmod 15$. Note that these values come from the structure of the cyclic form of the product code. This is visible in Fig. 2, in which the separation $(\bmod n_1 n_2)$ between neighboring positions is $an_1 = 10$ in any row and $bn_2 = 6$ in any column.

Let $d_1, d_2, \ldots, d_p$ denote the minimum distances of subcodes $C_1, C_2, \ldots, C_p$, respectively. Elias [7], shows that the minimum distance d for the product code C is the product of the minimum distances of the subcodes, $d = d_1 d_2 \cdots d_p$. Thus, product codes offer a simple way to construct a code with a large minimum distance from a set of shorter codes with smaller minimum distances.

We present some results on the error correction capability of product codes that are based on the structure of the codes. We note that these results are not necessarily achievable with most hard-decision decoding algorithms. The random-error-correction capability, which is the maximum number of errors that a code is guaranteed to correct, is thus given by

$$t = \left\lfloor \frac{d-1}{2} \right\rfloor = \left\lfloor \frac{\left(\prod_{i=1}^{p} d_i \right) - 1}{2} \right\rfloor.$$

Some error-control codes are able to correct more than t errors if the errors occur in bursts. A single error burst of length B_p occurs if all the errors in the codeword are constrained to B_p consecutive symbols of the codeword. Cyclic product codes are particularly useful for burst error correction. Again, consider a two-dimensional cyclic product code. Let t_1 and t_2 denote the random error

correction capability of subcodes C_1 and C_2, respectively. Let B_1 and B_2 denote the maximum length of an error burst that is guaranteed to be corrected by subcodes C_1 and C_2, respectively. Then code C_i can correct all errors that are constrained to B_i consecutive positions, regardless of the weight of the error event. The value B_i is said to be the burst error correction capability of subcode C_i. Let B_p denote the burst error correction capability of the product code. Then it is shown in [9] that B_p satisfies the following bounds:

$$B_p \geq n_1 t_2 + B_1$$

and

$$B_p \geq n_2 t_1 + B_2.$$

Several researchers have investigated the burst error correction capability of product codes constructed from single parity-check codes. This research is discussed in the following section.

3. PRODUCTS OF SINGLE PARITY-CHECK CODES

A particular product code that has drawn considerable attention is the p-time product of single parity-check (SPC) codes. We will refer to these codes as product SPC codes. Single parity-check codes are $(k+1, k)$ codes for which one parity bit is added for each k input bits. Typically, the parity bit is computed using *even parity*, in which case the sum of all the bits in the codeword is an even number. The parity-check code is a cyclic code with generator polynomial $g(X) = X + 1$. Product-SPC codes appear in the literature at about the same time from two different groups of authors, Calabi and Haefeli [11] and Gilbert [12]. Interestingly, each of these authors further attributes the original code idea to other researchers. Calabi and Haefeli attribute the code to C. Hobbs of the Communications Laboratory, Air Force Cambridge Research Center. Gilbert attributes a two-dimensional form of the code to W. D. Lewis. Apparently, the original contributions by W. D. Lewis were unpublished, but some extensions were patented as U.S. patents 2,954,432 and 2,954,433. Product codes that use SPC codes as subcodes have been called Hobbs' codes [11] or Gilbert codes [13,14].

Let C_i be a single parity-check code of length n_i for $1 \leq i \leq p$, and let $n = n_1 n_2 \cdots n_p$. Then the p-dimensional product of C_i, $1 \leq i \leq p$, is denoted by C and has blocklength n. Note that each information bit participates in exactly one parity-check equation in each dimension. The number of parity-check equations in which a bit participates is referred to as its *density* [13,15]. Since each bit participates in exactly p checks, the parity-check matrix can be constructed to have exactly p ones in every column. Such a parity-check matrix is known as a *regular* low-density parity-check matrix [15]. Suppose that the 5×3 product code shown in Fig. 2 is constructed from even SPC codes. Let v_{ij} denote the (i,j)th code symbol in the rectangular array representation. Then the code symbols must satisfy the row parity-check equations

$$\sum_{j=0}^{2} v_{ij} = 0, i = 0, 1, 2, 3, 4$$

and the column parity-check equations

$$\sum_{i=0}^{4} v_{ij} = 0, j = 0, 1, 2$$

Thus, if the codeword $\mathbf{v}$ is constructed using the bit ordering described in Theorem 1, a low-density representation for the parity-check matrix $\mathbf{H}$ is given by

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1
\end{bmatrix}.$$

This low-density parity-check matrix can be used in implementing a soft-decision decoder for the code.

The SPC code is a cyclic code, so if $n_1, n_2, \ldots, n_p$ are relatively prime, then the product code will be cyclic. For the two-dimensional code with n_1 and n_2 relatively prime, the generator polynomial is given by [13]

$$\begin{aligned} g(X) &= \mathrm{LCM}\{X_1^n + 1, X_2^n + 1\} \\ &= \frac{(X^{n_1}+1)(X^{n_2}+1)}{X+1} \end{aligned}$$

where $\mathrm{LCM}(y, z)$ denotes the least common multiple of y and z. For example, for the 5×3 product SPC code, the generator polynomial is given by

$$\begin{aligned} g(X) &= \frac{(X^5+1)(X^3+1)}{X+1} \\ &= X^7 + X^6 + X^5 + X^2 + X + 1 \end{aligned}$$

Note that codes formed this way are not only cyclic, but are *palindromic* [13], in which the code is the same if each codeword is read backward. Thus, for the 5×3 code, the *reciprocal* [8] of $g(X)$ is

$$\begin{aligned} X^7 g(X^{-1}) &= 1 + X + X^2 + X^5 + X^6 + X^7 \\ &= g(X) \end{aligned}$$

A parity-check polynomial for a cyclic code can be found [8] using the parity-check polynomial $h(X) = (X^n + 1)/g(X)$. However, the parity-check matrix formed using $h(X)$ is seldom a low-density matrix.

The burst error correction capabilities of these codes have been investigated [11–14,16,17]. Burst error detection capabilities were also investigated [13,14]. We present a summary of some of the most important results here. Neumann [13] points out that some of the product SPC codes are "fire" codes [8], which are another class of block codes designed for burst error correction. We consider the maximum-length error burst that the code can correct, and we denote the length of such a burst by B_p. Consider first the case of a two-dimensional $n_1 \times n_2$ product code,

where n_1 and n_2 are relatively prime. Let π_i denote the smallest prime divisor of n_i, and define

$$b_i = \left(\frac{\pi_i - 1}{\pi_i}\right) n_i$$

Then the code can correct all single error bursts up to length $B_p = \min\{b_1, b_2, \lfloor (n_1 + n_2 + 2)/3 \rfloor\}$. Thus for the 5×3 product SPC code, the single-burst error correction capability is

$$\begin{aligned} B_p &= \min\left\{\frac{5-1}{5} \cdot 5, \frac{3-1}{3} \cdot 3, \left\lfloor \frac{3+5+2}{3} \right\rfloor\right\} \\ &= \min\{4, 2, 3\} \\ &= 2 \end{aligned}$$

A *solid* burst error is one in which every bit in the burst is received in error. Then it is shown [14] that the two-dimensional product SPC code can correct all solid burst errors of length $\min\{n_1, n_2\} - 1$.

If a two-dimensional cyclic product SPC code is used for single-burst error detection, then its error detection capability is easily derived from the properties of cyclic codes [8,10]. Any burst of length $n - k$ can be detected for an (n, k) cyclic code. Thus, any burst of length up to $n_1 + n_2 - 1$ can be detected for a $n_1 \times n_2$ cyclic product SPC code. Neumann also shows that the code has the capability to simultaneously correct B_p errors while detecting burst errors of length almost equal to $\max\{n_1, n_2\}$.

4. DECODING OF PRODUCT SPC CODES

Cyclic product codes can be decoded using a variety of hard-decision and soft-decision decoding algorithms. One advantage of product-SPC codes is that they have very simple and efficient soft-decision decoding algorithms that we discuss in this section. However, we first provide some references to hard-decision decoding algorithms for these codes, particularly for application to burst error correction. Neumann [13] presents a decoding algorithm for single- and double-burst error correction. Bahl and Chien present a threshold decoding algorithm for product SPC codes with multiple error bursts [16] and a syndrome-based decoding algorithm that provides better performance [17].

Several soft-decision decoding algorithms exist for block codes [18–20]. Product SPC codes can be decoded in different ways, but we focus on an iterative decoding process that uses optimal maximum a posteriori probability decoders on each subcode in each dimension. The resulting iterative decoding algorithm is typically not an optimal decoding algorithm but is usually significantly simpler than an optimal decoder. In order to understand the operation of this iterative decoding algorithm, we first focus on the optimal symbol-by-symbol maximum-likelihood decoding algorithm for binary linear block codes.

The useful notion of a *replica* was first introduced in the context of soft-decision decoding by Battail et al. [18]. Consider a codeword $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1})$. A replica for bit v_i is information about bit v_i that can be derived from the code symbols other than v_i. Consider the (7,4) Hamming code. The parity-check matrix for this code can be written as

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Let a codeword $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1})$. Then the parity-check equations that involve the first code symbol v_0 are

$$\begin{aligned} v_0 + v_2 + v_3 + v_4 &= 0 \\ v_0 + v_1 + v_2 + v_5 &= 0 \\ v_0 + v_3 + v_5 + v_6 &= 0 \\ v_0 + v_1 + v_4 + v_6 &= 0 \end{aligned}$$

where all sums are modulo 2. Note that these equations are linearly dependent, so not every one conveys unique information. Suppose that we use the first three equations, which are linearly independent. Then the first symbol can be written in terms of the other six symbols in either of three ways:

$$\begin{aligned} v_0 &= v_2 + v_3 + v_4 \\ v_0 &= v_1 + v_2 + v_5 \\ v_0 &= v_3 + v_5 + v_6 \end{aligned}$$

These three equations provide three *algebraic replicas* [18] for v_0 in terms of the other symbols in the code. Note that the parity-check matrix $\mathbf{H}$ is the generator matrix for the dual code [8]. Thus, the replicas can be defined using codewords of the dual code [18]. Suppose that the codeword $\mathbf{v}$ is transmitted using binary phase shift keying (BPSK). In the absence of noise, the symbols at the output of the demodulator can be represented by a vector $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})$, where x_i represents the binary symbol v_i. When noise is present, we denote the demodulator outputs by $\mathbf{y} = (y_0, y_1, \ldots, y_{n-1})$. Then, in terms of the demodulator outputs, there are three *received replicas* of v_0 in addition to y_0, which may be considered a trivial received replica. For example, for the (7,4) Hamming code described above, y_2, y_3, and y_4 provide one received replica of v_0.

With a random information sequence, let V_i be a random variable denoting the ith code symbol of a codeword $\mathbf{V} = (V_0, V_1, \ldots, V_{n-1})$. The codeword $\mathbf{V}$ is transmitted using BPSK over a memoryless channel. Let $\mathbf{X}$ be a vector that consists of the outputs of the demodulator in the absence of noise. For convenience, let $X_i = +1$ if $V_i = 0$, and let $X_i = -1$ if $V_i = 1$. The loglikelihood ratio (LLR) for X_i is defined by

$$L(X_i) = \log \frac{P(X_i = +1)}{P(X_i = -1)}$$

where the natural (base e) logarithm is used. This term is called the *a priori loglikelihood* ratio. In the presence of noise, the received sequence consists of demodulator outputs $\mathbf{Y} = (Y_0, Y_1, \ldots, Y_{n-1})$. The conditional loglikelihood

ratio for X_i given $Y_i = y_i$ is given by

$$\begin{aligned} L(X_i \mid y_i) &= \log \frac{P(X_i = +1 \mid Y_i = y_i)}{P(X_i = -1 \mid Y_i = y_i)} \\ &= \log \frac{P(Y_i = y_i \mid X_i = +1)}{P(Y_i = y_i \mid X_i = -1)} + \log \frac{P(X_i = +1)}{P(X_i = -1)} \\ &= L(y_i \mid X_i) + L(X_i). \end{aligned}$$

The LLRs are often referred to as "soft" values. The sign of the LLR corresponds to a hard-decision value, while the magnitude of the LLR corresponds to the reliability of the decision.

A symbolwise maximum a posteriori (MAP) probability decoder makes decision on X_i based on the larger of $P(X_i = +1 \mid \mathbf{Y} = \mathbf{y})$ and $P(X_i = -1 \mid \mathbf{Y} = \mathbf{y})$, or, equivalently, the sign of

$$L(X_i \mid \mathbf{Y} = \mathbf{y}) = \log \frac{P(X_i = +1 \mid \mathbf{Y} = \mathbf{y})}{P(X_i = -1 \mid \mathbf{Y} = \mathbf{y})}.$$

Note that in general, the a posteriori loglikelihood ratio for X_i depends not only on Y_i but also other symbols in $\mathbf{Y}$. This dependence corresponds to the other received replicas of X_i. Hence, we need to calculate the contribution of a replica of X_i to the above-mentioned a posteriori LLR. We follow the approach described by Hagenauer et al. [19]. Define $\oplus$ as the addition operator over Galois Field (2) with symbols $+1$ and -1, where $+1$ is the null element (since $+1$ corresponds to 0 in the original binary representation). Suppose that X_j and X_k form a replica of X_i via the relation

$$X_i = X_j \oplus X_k$$

We wish to find the loglikelihood ratio for the replica $X_j \oplus X_k$. By using

$$P(X_j = +1) = \frac{e^{L(X_j)}}{1 + e^{L(X_j)}}$$

with the relationship [19]

$$\begin{aligned} P(X_j \oplus X_k = +1) &= P(X_j = +1)P(X_k = +1) \\ &\quad + [1 - P(X_j = +1)][1 - P(X_k = +1)] \end{aligned}$$

we can write

$$P(X_j \oplus X_k = +1) = \frac{1 + e^{L(X_j)}e^{L(X_k)}}{(1 + e^{L(X_j)})(1 + e^{L(X_k)})}$$

Then, using $P(X_j \oplus X_k = -1) = 1 - P(X_j \oplus X_k = +1)$, the loglikelihood ratio for $X_j \oplus X_k$ can be written as

$$L(X_j \oplus X_k) = \log \frac{1 + e^{L(X_j)}e^{L(X_k)}}{e^{L(X_j)} + e^{L(X_k)}}$$

or equivalently

$$\begin{aligned} L(X_j \oplus X_k) = \log &\frac{[e^{L(X_j)} + 1][e^{L(X_k)} + 1] + [e^{L(X_j)} - 1]}{[e^{L(X_j)} + 1][e^{L(X_k)} + 1] - [e^{L(X_j)} - 1]} \\ &\times \frac{[e^{L(X_k)} - 1]}{[e^{L(X_k)} - 1]} \end{aligned}$$

Note that using the relationship $\tanh(x/2) = (e^x - 1)/(e^x + 1)$, this expression can be simplified to [18,19]

$$\begin{aligned} L(X_j \oplus X_k) &= \log \frac{1 + \tanh(L(X_j)/2)\tanh(L(X_k)/2)}{1 - \tanh(L(X_j)/2)\tanh(L(X_k)/2)} \\ &= 2\ \text{atanh}[\tanh(L(X_j)/2)\tanh(L(X_k)/2)] \end{aligned}$$

In general, if a replica of X_i is given by $X_{j_1} \oplus X_{j_2} \oplus \cdots \oplus X_{j_J}$, then the loglikelihood ratio for the replica is given by

$$2\ \text{atanh}\left[\prod_{k=1}^{J} \tanh \frac{L(X_{j_k})}{2}\right]$$

Note that for high signal-to-noise ratios, this can be approximated by

$$\left[\prod_{k=1}^{j} \text{sgn}(L(X_{j_k}))\right] \cdot \min_{k=1,\ldots,J} |L(X_{j_k})|$$

Thus, the reliability of a replica is generally determined by the smallest reliability of the symbols that make up that replica. More commonly, we wish to determine the conditional loglikelihood ratio for a replica of X_j given the received symbols $y_{j_1}, y_{j_2}, \ldots, y_{j_J}$. Then this conditional LLR can be written as

$$2\ \text{atanh}\left[\prod_{k=1}^{J} \tanh \frac{L(X_{j_k} \mid y_{j_k})}{2}\right]$$

For the $(k+1, k)$ even SPC code, the parity-check matrix is given by

$$\mathbf{H} = [111 \cdots 1]$$

Thus, there is one nontrivial (algebraic) replica for each code symbol. The algebraic replica for X_i is given by

$$\sum_{j=0, j\neq i}^{n-1} \oplus\ X_j = X_0 \oplus X_1 \oplus \cdots \oplus X_{i-1} \oplus X_{i+1} \oplus \cdots \oplus X_{n-1}$$

Thus, the conditional loglikelihood of this replica for X_i given $\mathbf{Y}$ is

$$2\ \text{atanh}\left[\prod_{k=1, k\neq j}^{n-1} \tanh \frac{L(X_k \mid y_k)}{2}\right]$$

Considering this algebraic replica and the trivial replica, the a posteriori loglikelihood ratio $L(X_i \mid \mathbf{Y} = \mathbf{y})$ for a code symbol X_i can be broken down into

1. The a priori loglikelihood ratio $L(X_i)$
2. The conditional loglikelihood ratio $L(y_i \mid X_i)$ of received symbol y_i given X_i
3. *Extrinsic information* $L_e(X_i)$ that is information derived from the replicas of X_i

For the SPC code, the a priori LLR for X_i is $L(X_i)$, and is set to zero initially. Consider transmission over an additive white Gaussian noise (AWGN) channel with code

rate R_c and bit energy-to-noise density ratio E_b/N_0. Then the LLR for the received symbol y_i given X_i is

$$L(y_i \mid X_i) = \log \frac{\exp[-\sigma^{-2}(y_i - 1)^2]}{\exp[-\sigma^{-2}(y_i + 1)^2]} = L_c \cdot y_i$$

where

$$L_c = \frac{2}{\sigma^2} = 4\frac{R_c E_b}{N_0}$$

The extrinsic information, that is, the conditional LLR of the replica of X_i given $\mathbf{y}$, is given by

$$L_e(X_i) = 2 \operatorname{atanh}\left[\prod_{k=0,k\neq i}^{n-1} \tanh\frac{L(X_k \mid y_k)}{2}\right] \approx \left[\prod_{k=0,k\neq i}^{n-1} \operatorname{sgn}(L(X_k \mid y_k))\right] \cdot \min_{k=0,\ldots,n-1,k\neq i} |L(X_k \mid y_k)|$$

Thus, the a posteriori loglikelihood ratio for X_i is given by

$$L(X_i \mid \mathbf{Y} = \mathbf{y}) = L(X_i) + L_c \cdot y_i + L_e(X_i)$$

The extrinsic information represents indirect information [19] about the symbol X_i from other code symbols. In the context of product codes, due to the orthogonal parity-check construction, the extrinsic information about X_i that is derived from a particular component subcode does not involve any received symbols (other than X_i) of any other subcode that involves X_i. Thus, this extrinsic information represents information that is not directly available to the decoders of the other subcodes. The extrinsic information can be used by other subcodes by exchanging extrinsic information between the subcodes in an iterative fashion. Each decoder treats the extrinsic information generated by other decoders as if it were a priori information in its decoding.

To illustrate this iterative decoding algorithm, we consider a two-dimensional product SPC code that is formed from the product of two identical (4,3) SPC codes. Conforming to the usual terminology in the iterative decoding literature [19], we refer to the conditional LLR of X_i given y_i, $L(X_i \mid y_i)$, as the *soft input* to a decoder of a SPC subcode. As discussed previously, this soft input is simply the sum of the *a priori* LLR of X_i, $L(X_i)$, and the conditional LLR of the received symbol y_i given X_i, $L_c y_i$. The decoder generates the extrinsic information of X_i, $L_e(X_i)$, based on the soft inputs as described above. This extrinsic information is then used as the a priori LLR X_i for the decoding of the other component subcode. Extrinsic information may be calculated for only the systematic symbols [19] or for all of the code symbols [21], which may offer some improvement in performance. For the results below, we calculate extrinsic information for all the code symbols. The decoding process alternates between the decoders for the two subcodes until a stopping criterion is met. Then the decoder outputs the a posteriori LLR of X_i, which is simply the sum of the soft input and the extrinsic information from each decoder for X_i. A hard decision on X_i is made based on the sign of this *soft output*.

For example, consider the example illustrated in Fig. 3. The transmitted symbols are shown in Fig. 3a. The rightmost column and the bottom row contain the parity-check symbols, and the other symbols carry information. The LLRs of the received symbols are given in Fig. 3b. We see that the decoder would make five errors (indicated by the shaded symbols) if decisions were made directly using these received values. Figure 3c shows the decoding results of the a posteriori decoding algorithm on the SPC defined along the rows. The number inside the upper triangle for each symbol denotes the soft input for that symbol, while the number inside the lower triangle is the extrinsic information generated by the decoder. Initially, we assume that the a priori LLRs of all the symbols are zero. For instance, the soft input for the symbol in the upper left corner is given by $2 + 0 = 2$. To obtain the soft output for a symbol, we simply need to add the values inside the upper and lower triangles corresponding to that symbol. For instance, the soft output of the symbol in the upper left corner is $2.0 - 0.5 = 1.5$. We see that three errors would result if the decoder were to make decisions based on the soft output after this decoder iteration. The decoding process continues for the SPC defined along the columns. The results are shown in Fig. 3d. We note that the soft input of a symbol is now obtained by adding the received LLR of that symbol (from Fig. 3b) to the extrinsic information generated in the previous decoding process (Fig. 3c). For instance, we obtain the soft input of the symbol in the upper left corner as $2.0 - 0.5 = 1.5$. We see that two errors would result if hard decisions were made at this time. The decoding process then returns to the decoder for the SPC along the rows (Fig. 3e) and then the SPC along the columns (Fig. 3e). We see that all five errors that were initially present are corrected by this iterative decoding process. It is also easy to see that any further decoder iterations will not result in any changes in the hard decisions. So for this example, the decoding process converges. Interested readers are referred to the article by Rankin and Gulliver [21] for additional discussion of the performance of iterative decoding with multidimensional product SPC codes.

5. MULTIDIMENSIONAL PARITY-CHECK CODES

In this section, we define a class of multidimensional parity-check (MDPC) codes [22]. These codes are punctured versions of the M-dimensional product SPC codes discussed above. In particular, the M-dimensional product SPC codes have code rates that decrease as M is increased. By puncturing the majority of the parity-check bits for $M > 2$, the MDPC codes have code rates that increase as M increases.

The parity bits for the MDPC code are determined by placing the information bits into a multidimensional array of size M, where $M > 1$. For a particular value of M, we refer to the M-dimensional parity-check code as an M-DPC code. For the special case of $M = 2$, this code is also referred to as a *rectangular parity-check code* (RPCC). In all that follows, we assume that the size of the array is the same

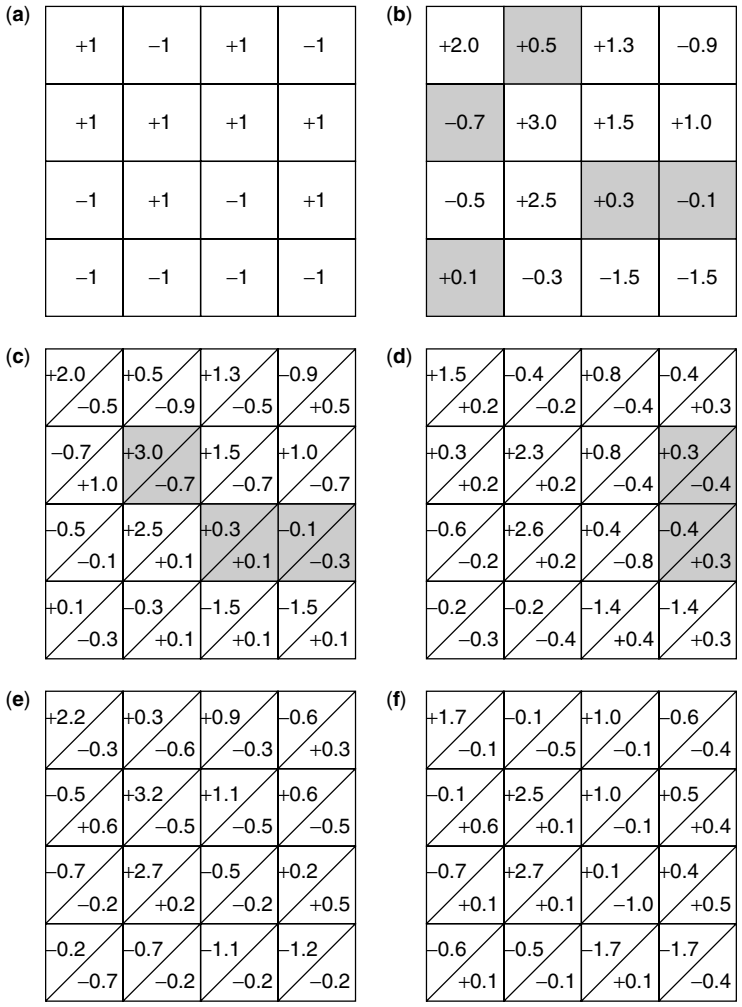


Figure 3. Iterative decoding example of a two-dimensional product SPC: (**a**) transmitted symbols; (**b**) LLRs of received symbols; (**c**) First iteration horizontal SISO decoding; (**d**) First iteration vertical SISO decoding; (**e**) Second iteration horizontal SISO decoding; (**f**) Second iteration vertical SISO decoding.

in each dimension. This is not a requirement in general, but it does minimize the redundancy (and hence the rate penalty) for a given block size and number of dimensions. Suppose that N is the blocklength (the number of input bits that are input to the code to create a codeword), where $N = D^M$. Then D is the size of the M-dimensional array in each dimension. Each M-DPC code is a systematic code in which a codeword consists of D^M information bits and MD parity bits. Then D parity bits are computed for each dimension, where the parity bits can be constructed as the even parity over each of the D hyperplanes of size D^{M-1} that are indexed by that dimension. This is illustrated in Fig. 4 for $M = 2$ and $M = 3$.

More formally, let $u_{i_1,i_2,\ldots,i_M}$ be a block of data bits indexed by the set of M indices $i_1, i_2, \ldots, i_M$. The data bits are arranged in the lattice points of an M-dimensional hypercube of side D. Then the MD parity bits satisfy

$$p_{m,j} = \sum_{i_1} \cdots \sum_{i_{m-1}} \sum_{i_{m+1}} \cdots \sum_{i_M} u_{i_1,\ldots,i_{m-1},j,i_{m+1}\ldots,i_M}$$

for $m = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, D$. Each sum above ranges over D elements, and modulo-2 addition is assumed. Since the M-DPC code produces MD parity bits, the code rate is $D^M/(D^M + MD)$, or equivalently, $(1 + MD^{1-M})^{-1}$. Clearly, as D is increased, the rate of the code becomes very high. For most values of N and M that are of interest, the rate of the M-DPC code increases as M is increased. The minimum code weight is $\min(M + 1, 4)$.

6. BURST ERROR CORRECTION CAPABILITY OF MDPCS WITH ITERATIVE SOFT-DECISION DECODING

MDPC codes can achieve close-to-capacity performance with a simple iterative decoding algorithm in additive white Gaussian noise as well as bursty channels. The iterative decoding algorithm described in Section 4 can be employed with a slight modification. Since the parity-on-parity bits in the product SPC code are punctured in the MDPC code, we do not update the soft inputs corresponding to the parity bits in the iterative decoding process of the MDPC code. Using iterative decoding and with a minimal amount of added redundancy, MDPC codes are very effective for relatively benign channels with possibly occasional long bursts of errors. For example, a three-dimensional parity-check code with a block size of 60,000 can achieve a bit error rate (BER) of 10^{-5} within 0.5 dB of the capacity limit in an AWGN channel while requiring only 0.2% added redundancy. The same code can get to within 1.25 dB of the capacity limit in a bursty channel.

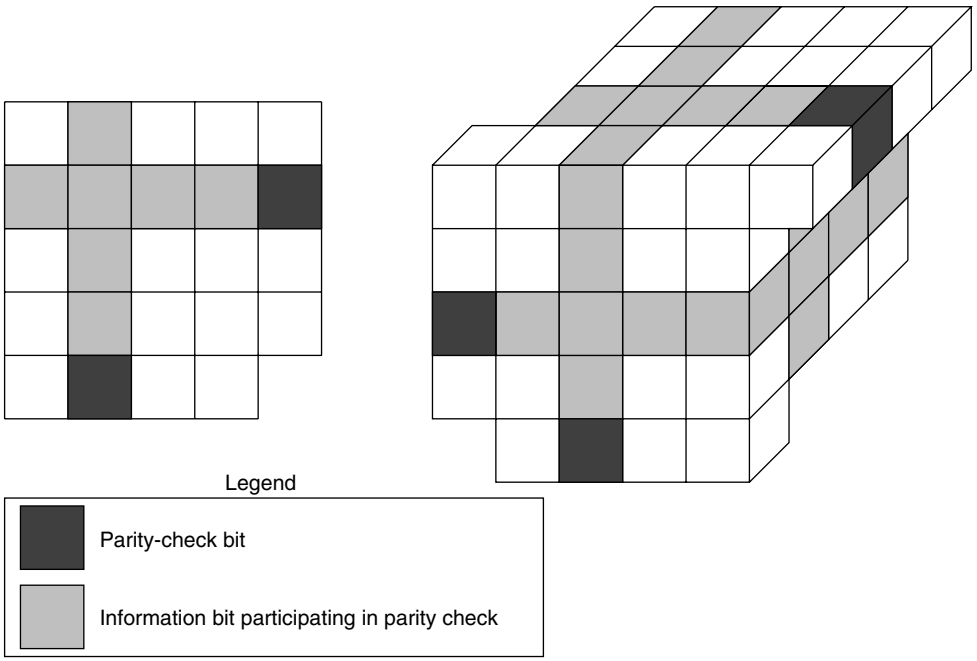


Figure 4. Determination of parity-check bits for $M = 2$ and $M = 3$ MDPC codes.

6.1. Additive White Gaussian Noise Channel

Simulation results of a number of MDPC codes with blocklengths of about 1000, 10,000, and 60,000 data bits over an AWGN channel with BPSK modulation are summarized in Table 1. The results are obtained after 10 iterations for all the codes. However, the decoding process essentially converges after five iterations for all of the MDPC codes that were considered. For instance, the convergence of the iterative decoding process for the 100^2 code is shown in Fig. 5. From Table 1, we conclude that with a block size of 1000 bits, the MDPC codes can achieve a BER of 10^{-5} within 2 dB of the capacity limit.[1] When the block size increases to 10,000 bits, the performance of the MDPC codes is within 1 dB of the capacity. These results are comparable to the ones reported in the article by Chen and McEliece [24], in which codes based on pseudorandom bipartite graphs obtained from computer searches are employed. In comparison, the MDPC codes considered here have much more regular structures, faster convergence rates, and a simpler decoding algorithm. With a block size of approximately 60,000 bits, the 3DPC code 39^3 can achieve a BER of 10^{-5} at 7.9 dB, which is 0.5 dB higher than the capacity limit. Moreover, significant coding gains over uncoded BPSK systems are achieved with very small percentages of added redundancy. For example, using the 21^3 code, a coding gain of 2.3 dB is obtained at 10^{-5} with less than 0.7% redundancy. This accounts for 80.4% of the maximum possible coding gain of 3.25 dB that is allowed by the capacity. It appears that the 3DPC codes are most efficient in terms of attaining the highest percentage of the maximum possible coding gain.

Using the union bound technique [25], we can obtain an upper bound on the bit error probability of the MDPC codes with maximum likelihood (ML) decoding as follows:

$$P_b \le \sum_{i=1}^{D^M} \frac{i}{D^M} \sum_{d=i}^{(M+1)i} W_{i,d} Q\left(\sqrt{\frac{2d E_b/N_0}{1 + M/D^{M-1}}}\right) \qquad (1)$$

where $W_{i,d}$ is the number of codewords with information weight i and codeword weight d. Figure 6 shows the union

[1] Symmetric capacity restricted to BPSK [23] is assumed here.

Table 1. Performance of MDPC Codes Over AWGN Channel

Code	Block Size	Code Rate	E_b/N_0 at 10^{-5} BER (dB)	Coding Gain at 10^{-5} BER (% of Possible Coding Gain)	E_b/N_0 at Capacity (dB)
32^2	1024	0.9412	6.3	3.3 (55.0%)	3.7
10^3	1000	0.9709	6.75	2.85 (63.1%)	4.75
4^5	1024	0.9808	7.25	2.35 (63.8%)	5.3
100^2	10,000	0.9804	6.75	2.85 (70.8%)	5.25
21^3	9261	0.9932	7.3	2.3 (80.4%)	6.35
10^4	10,000	0.9960	7.75	1.85 (80.4%)	6.8
245^2	60,025	0.9919	7.2	2.4 (79.4%)	6.2
39^3	59,319	0.9980	7.9	1.7 (89.1%)	7.4
9^5	59,049	0.9992	8.6	1.0 (88.1%)	8.05

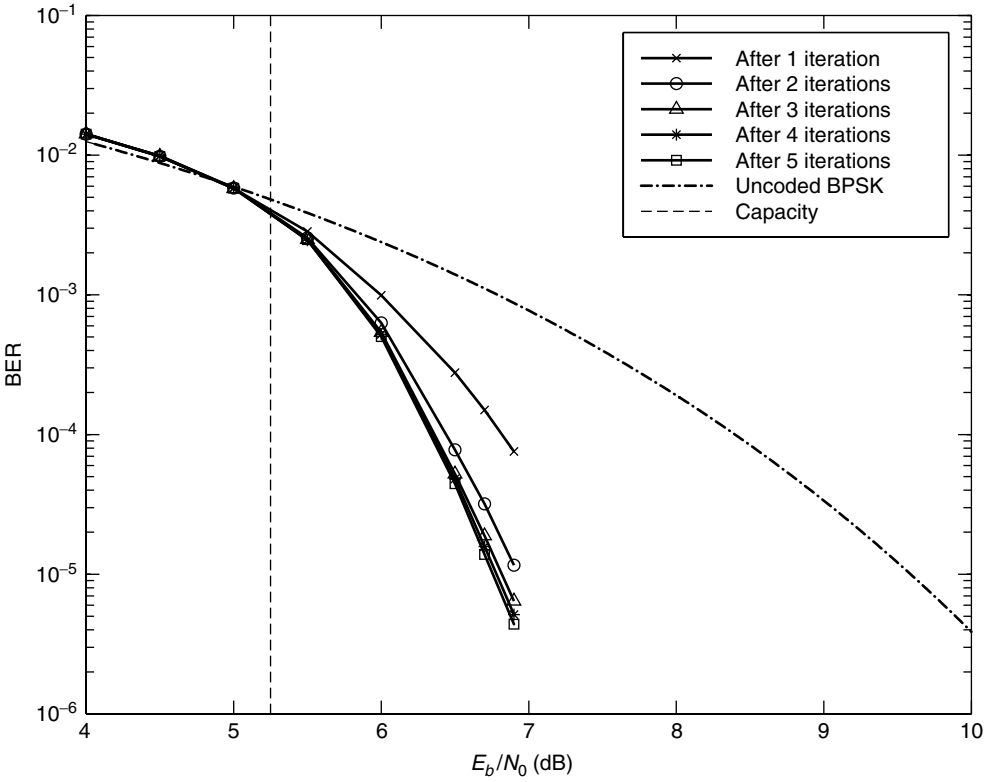


Figure 5. Convergence of iterative decoding process for 100^2 code over AWGN channel.

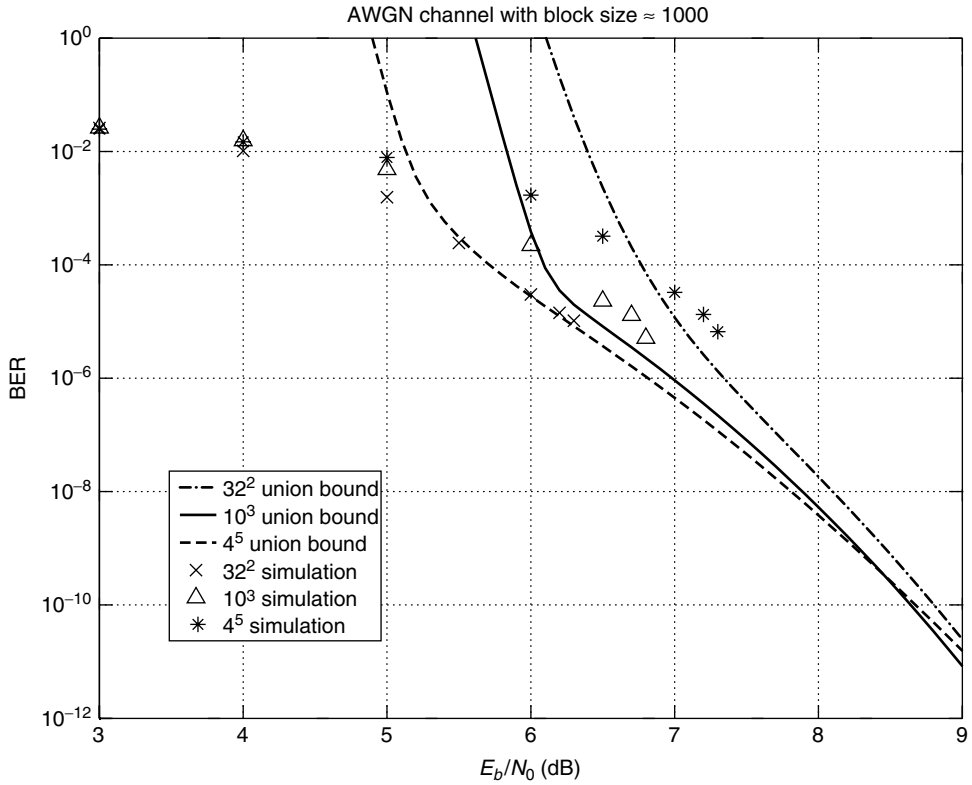


Figure 6. Union bounds and BER performance from simulations of 32^2, 10^3, and 4^5 codes over AWGN channel.

bounds obtained using Eq. (1) for the 32^2, 10^3, and 4^5 codes.[2] Also shown in Fig. 6 are the bit error probabilities of these three codes obtained by the iterative decoder from simulations. We observe from the figure that the BER performance obtained from simulations for the code 32^2 is very close to the corresponding union bound in the high E_b/N_0 region. This indicates that the performance of the iterative decoder is close to that of the ML decoder. For the three- and five-dimensional codes 10^3 and 4^5, the BERs obtained from simulations are poorer than the ones predicted by the respective union bounds. This implies that the iterative decoder becomes less effective as the dimension of the MDPC code increases. Nevertheless, iterative decoding can still provide good coding gains, as shown in Table 1, for MDPC codes with more than two dimensions.

6.2. Bursty Channels

Although the MDPC codes have small minimum distances, they can correct a large number of error patterns of larger weights because of their geometric constructions. With suitable interleaving schemes, the MDPC codes are effective for channels with occasional noise bursts. To examine this claim, we employ the simple two-state hidden Markov model, shown in Fig. 7, to model bursty channels [26]. The system enters state **B** when the channel is having a noise burst. In state **N**, usual AWGN is the only noise. In state **B**, the burst noise is modeled as AWGN with a power spectral density that is B times higher than that of the AWGN in state **N**. It is easy to check that the stationary distribution of the hidden Markov model is $\pi_b = \frac{1-P_n}{1-P_b+1-P_n}$ and $\pi_n = \frac{1-P_b}{1-P_b+1-P_n}$. Assuming that the noise is independent from symbol to symbol after the deinterleaver at the receiver, the conditional LLR $L(y_i \mid X_i)$ is given by

$$L(y_i \mid X_i) = 4R_c\frac{E_b}{N_0}\frac{x}{B} + \log\frac{\exp\left(-\frac{B-1}{B}R_c\frac{E_b}{N_0}(y_i-1)^2\right)+\frac{\pi_b}{\pi_n}}{\exp\left(-\frac{B-1}{B}R_c\frac{E_b}{N_0}(y_i+1)^2\right)+\frac{\pi_b}{\pi_n}} \quad (2)$$

Simulation results of a number of MDPC codes with block sizes of 10,000 and 60,000 bits are summarized in Table 2. In the simulation, $P_b = 0.99$, $P_n = 0.9995$, and $B = 10$ dB. This represents a case where long noise bursts occur occasionally. Random interleavers of sizes equal to the block size of the MDPC codes are employed. The results are obtained after 10 iterations for all the codes. The convergence of the iterative decoding process is similar to the AWGN case. From Table 2, with a block size of 10,000 bits, the 4DPC code 10^4 can achieve a BER of 10^{-5} within 2.4 dB of the capacity limit.[3] When the block size is increased to 60,000 bits, the 5DPC code 9^5 can achieve a BER of 10^{-5} within 1.2 dB of the capacity limit. Although the MDPC codes are not as effective in bursty channels as in AWGN channels, they do provide very significant coding gains with very reasonable complexity as no channel state estimation is needed [26]. In fact, the complicated conditional likelihood ratio calculation in Eq. (2) is not needed since simulation results show that the degradation on the BER performance is very small if the second term on the right-hand side of (2) is neglected.

Using the union bound and assuming that a perfect interleaver is employed so that channel state changes independently from bit to bit, we can obtain the following upper bound on the BER for an ML decoder with perfect channel state information:

$$P_b \le \sum_{i=1}^{D^M}\frac{i}{D^M}\sum_{d=i}^{(M+1)i}W_{i,d}\sum_{k=0}^{d}\binom{d}{k}\pi_b^k\pi_n^{d-k}Q \times\left(\sqrt{\frac{2E_b/N_0}{1+M/D^{M-1}}}\cdot\frac{d-k+k/\sqrt{B}}{\sqrt{d}}\right) \quad (3)$$

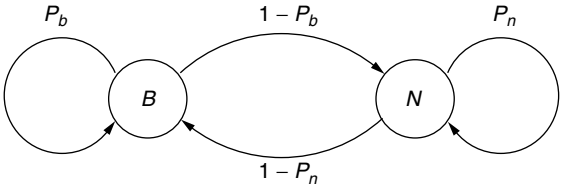


Figure 7. Hidden Markov model for bursty channel.

[2] Here, the weight enumerator coefficients $W_{i,d}$ are obtained approximately by the Monte Carlo method. The first 30 terms in Eq. (1) are used to approximate the union bound.

[3] The capacity here is obtained by averaging the symmetric capacities under the normal and bursty states based on the stationary distribution of the hidden Markov model. This corresponds to the case that a perfect interleaver is employed so that for a given bit, the channel state is independent of the channel states of the other bits, and perfect channel state information is available at the receiver [27].

Table 2. Performance of MDPC Codes Over Bursty Channel

Code	Block Size	Code Rate	E_b/N_0 at 10^{-5} BER (dB)	Coding Gain at 10^{-5} BER (% of Possible Coding Gain)	E_b/N_0 at Capacity (dB)
100^2	10,000	0.9804	13.7	4.15 (29.5%)	8.4
21^3	9261	0.9932	14.8	3.05 (52.5%)	12.0
10^4	10,000	0.9960	15.5	2.35 (57.5%)	13.1
245^2	60,025	0.9919	14.1	3.75 (55.0%)	11.5
39^3	59,319	0.9980	15.45	2.4 (75.0%)	14.2
9^5	59,049	0.9992	16.6	1.25 (75.9%)	15.4

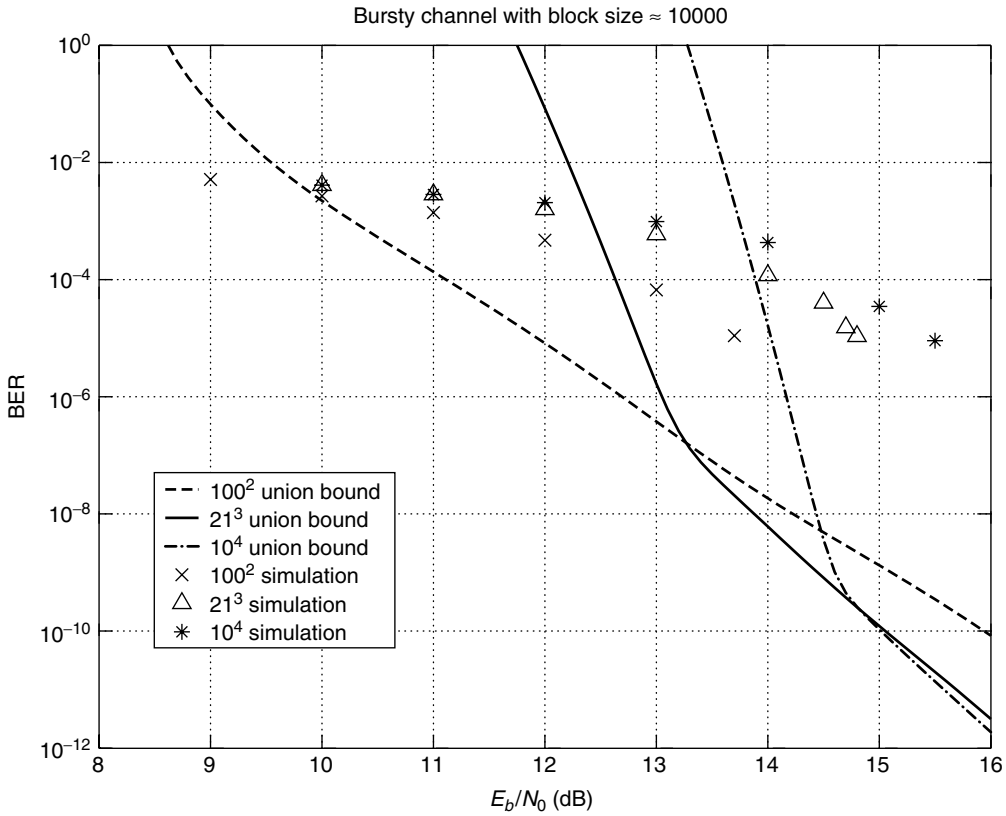


Figure 8. Union bounds and BER performance from simulations of 100^2, 21^3, and 10^4 codes over bursty channel with $P_b = 0.99$, $P_n = 0.9995$, and $B = 10$ dB.

Figure 8 shows the union bounds obtained by this equation for the 100^2, 21^3, and 10^4 codes. Also shown in Fig. 8 are the bit error probabilities of these three codes obtained by the iterative decoder from simulations. We observe from the figure that the BERs obtained from simulations are poorer than those predicted by the union bounds. The reason is threefold. First, the iterative decoder only approximates the MAP decoder. The second, and perhaps the most important, reason is that random interleavers are of the same size as the codewords. These interleavers are not good approximations to the perfect interleaver assumed in the union bound, since such an interleaver would require interleaving across multiple codewords. Third, no channel state information is assumed at the receiver. Nevertheless, the simple iterative decoder and the imperfect interleaver can still give large coding gains as shown in Table 2.

7. CONCATENATED MULTIDIMENSIONAL PARITY-CHECK CODES AND TURBO CODES

Turbo codes [28] are parallel-concatenated convolutional codes that have been shown to provide performance near the capacity limit when very large interleavers (and thus codeword lengths) are used. These codes suffer from an error floor that limits their performance for shorter blocklengths. The error floor is caused by error events that have very low information weight. Thus, these low-weight error-events can be corrected by even a simple outer code. Several authors have investigated the use of an outer code to deal with these low-weight errors. BCH codes have been considered [29–33], and Reed–Solomon codes were also considered [34,35]. However, these codes are typically decoded with algebraic decoders [8] because the complexity is too high for soft-decision decoders for these codes. An alternative approach is to use the multidimensional parity-check codes that are discussed in the previous two sections as outer codes with a Turbo inner code [36,37]. The MDPC codes have simple soft-decision decoders and typically have very high rates that result in less degradation to the performance of the turbo code than most of the other outer codes that have been used. Furthermore, the MDPC codes are good at correcting bursts of errors, such as those that occur at the output of a Turbo decoder.

The results in Fig. 9 illustrate the potential of these codes in regard to improving the error floor. The Turbo code used by itself and in concatenation with the multidimensional parity-check codes is the rate-$\frac{1}{3}$ turbo code with the constituent codes specified in the standards for the cdma2000 and WCDMA third-generation cellular systems [38,39]. The results indicate that the concatenated multidimensional parity-check codes can reduce the error floor by many orders of magnitude in comparison to a Turbo code by itself. These results have been verified by simulation, as illustrated by the results shown in Fig. 10. An alternate technique

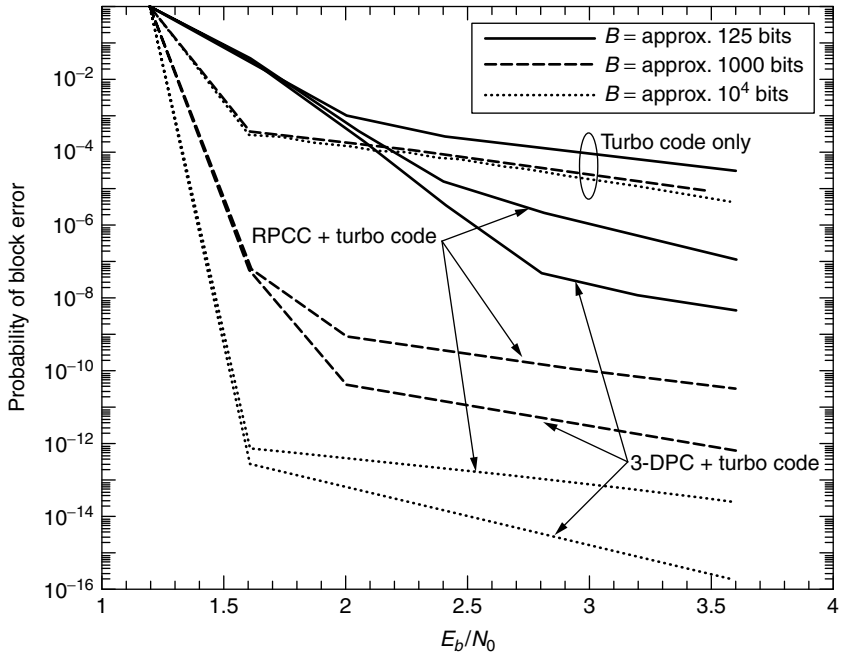


Figure 9. Bounds on the performance of concatenated outer multidimensional parity-check codes with inner rate-$\frac{1}{3}$ Turbo codes.

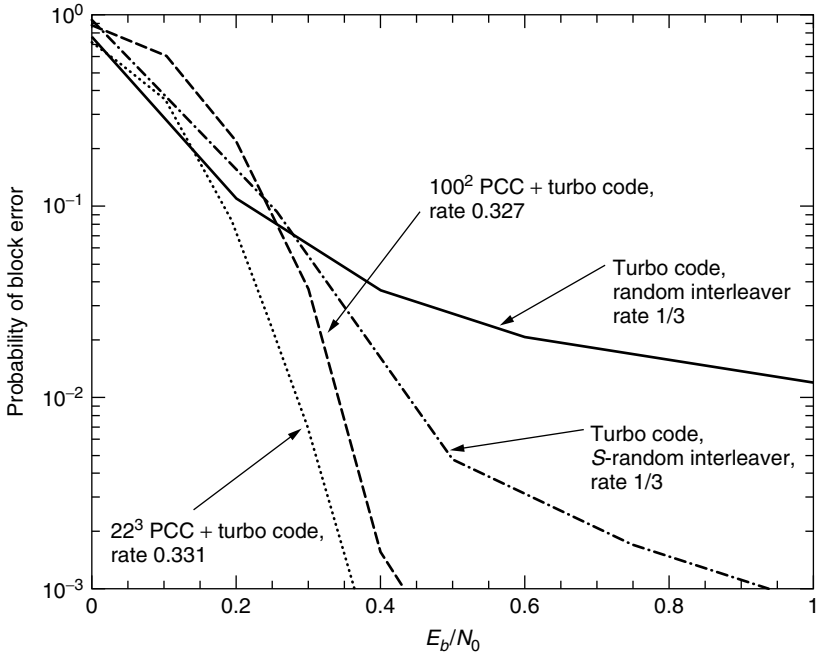


Figure 10. MDPC + Turbo codes and Turbo codes with blocklength of approximately 10^4 bits and rates approximately equal to $\frac{1}{3}$.

to improve the error floor is the use of *S*-random interleaving [40]. The simulation results presented in Fig. 10 show that the concatenated multidimensional parity-check and Turbo codes provide significantly better performance than do Turbo codes, even when *S*-random interleaving is used.

BIOGRAPHIES

John M. Shea (S'92-M'99) received the B.S. (with highest honors) in Computer Engineering from Clemson University in 1993 and the M.S. and Ph.D. degrees in Electrical Engineering from Clemson University in 1995 and 1998, respectively. In 1999 he joined the University of Florida, where he is currently an Assistant Professor of Electrical and Computer Engineering. Dr. Shea was a National Science Foundation Fellow from 1994 to 1998. He received the Ellersick Award from the IEEE Communications Society in 1996. "Dr. Shea serves as an Associate Editor for the *IEEE Transactions on Vehicular Technology*." He is currently engaged in research on wireless communications with emphasis on Turbo coding

and iterative decoding, adaptive signaling, and spread-spectrum communications.

Tan F. Wong received the B.Sc. degree (First Class Honors) from the Chinese University of Hong Kong, and the M.S.E.E., and Ph.D. degrees in Electrical Engineering from Purdue University in 1991, 1992, and 1997, respectively. He is currently an Assistant Professor of Electrical and Computer Engineering at the University of Florida. Prior to that, he was a research engineer working on the high-speed wireless networks project at the Department of Electronics at Macquarie University, Sydney, Australia. He also served as a Postdoctoral Research Associate in the School of Electrical and Computer Engineering at Purdue University. Dr. Wong serves as an Editor for the *IEEE Transactions on Communications* and as an Associate Editor for the *IEEE Transactions on Vehicular Technology*. His research interests include spread-spectrum communications, multiuser communications, error-control coding, and wireless networks.

BIBLIOGRAPHY

1. G. D. Forney, Jr., et al., Efficient modulation for bandlimited channels, *IEEE J. Select. Areas Commun.* **SAC-2**: 632–646 (Sept. 1984).
2. L.-F. Wei, Trellis-coded modulation with multidimensional constellations, *IEEE Trans. Inform. Theory* **IT-33**: 483–501 (July 1987).
3. E. Biglieri, D. Divsalar, P. J. McLane, and M. K. Simon, *Introduction to Trellis-Coded Modulation with Applications*, Macmillan, New York, 1991.
4. I. Blake, C. Heegard, T. Høholdt, and V. K.-W. Wei, Algebraic-geometry codes, *IEEE Trans. Inform. Theory* **44**: 2596–2618 (Oct. 1998).
5. K. A. S. Abdel-Ghaffar, R. J. McEliece, and H. C. A. van Tilborg, Two-dimensional burst identification codes and their use in burst correction, *IEEE Trans. Inform. Theory* **34**: 494–504 (May 1988).
6. W. J. van Gils, Two-dimensional dot codes for product identification, *IEEE Trans. Inform. Theory* **IT-33**: 620–631 (Sept. 1986).
7. P. Elias, Error-free coding, *IRE Trans. Inform. Theory* **IT-4**: 29–37 (Sept. 1954).
8. S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
9. H. O. Burton and E. J. Weldon, Jr., Cyclic product codes, *IEEE Trans. Inform. Theory* **IT-11**: 433–439 (July 1965).
10. W. W. Peterson and E. J. Weldon, Jr., *Error-correcting Codes*, MIT Press, Cambridge, MA, 1972.
11. L. Calabi and H. G. Haefeli, A class of binary systematic codes correcting errors at random and in bursts, *IRE Trans. Circuit* (special supplement to **CT-6**): 79–94 (May 1959).
12. E. N. Gilbert, A problem in binary encoding, *Proc. Symp. Applied Math.*, 1960, Vol. 10, pp. 291–297.
13. P. G. Neumann, A note on Gilbert burst-correcting codes, *IEEE Trans. Inform. Theory* **IT-11**: 377–384 (July 1965) (The reader should note that some of the results in this reference were corrected in Ref. 14, below).
14. L. R. Bahl and R. T. Chien, On Gilbert burst-error-correcting codes, *IEEE Trans. Inform. Theory* **IT-15**: 431–433 (May 1969).
15. R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.
16. L. R. Bahl and R. T. Chien, Multiple-burst-error correction by threshold decoding, *Inform. Control* **15**: 397–406 (Nov. 1969).
17. L. R. Bahl and R. T. Chien, Single- and multiple-burst-correcting properties of a class of cyclic product codes, *IEEE Trans. Inform. Theory* **IT-17**: 594–600 (Sept. 1971).
18. G. Battail, M. C. deCouvelaere, and P. Godlewski, Replication decoding, *IEEE Trans. Inform. Theory* **IT-25**: 332–345 (May 1979).
19. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**: 429–445 (March 1996).
20. R. M. Pyndiah, Near-optimum decoding of product codes: Block turbo codes, *IEEE Trans. Commun.* **46**: 1003–1010 (Aug. 1998).
21. D. M. Rankin and T. A. Gulliver, Single parity check product codes, *IEEE Trans. Commun.* **49**: 1354–1362 (Aug. 2001).
22. T. F. Wong and J. M. Shea, Multi-dimensional parity check codes for bursty channels, *Proc. 2001 IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 123.
23. R. E. Blahut, *Principles and Practices of Information Theory*, Addison-Wesley, Reading, MA, 1987.
24. J. Chen and R. J. McEliece, *Frequency-Efficient Coding with Low-Density Generator Matrices*, Technical Report, California Institute of Technology, available at *http://www.ee.caltech.edu/systems/jfc/publications.html*.
25. D. Divsalar, S. Dolinar, F. Pollara, and R. McEliece, *Transfer Function Bounds on the Performance of Turbo Codes*, Technical Report TDA Progress Report 42-122, NASA Jet Propulsion Laboratory, Aug. 1995.
26. K. Koike and H. Ogiwara, Application of turbo codes for impulsive noise channels, *IEICE Trans. Fund.* **E81-A**: 2032–2039 (Oct. 1998).
27. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: Information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* **44**: 2619–2692 (Oct. 1998).
28. C. Berrou, A. Galvieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding, *Proc. 1993 IEEE Int. Conf. Communications*, Geneva, Switzerland, Vol. 2, 1993, pp. 1064–1070.
29. J. D. Andersen, "Turbo" coding for deep space applications, *Proc. 1995 IEEE Int. Symp. Information Theory*, Whistler, British Columbia, Canada, Sept. 1995, p. 36.
30. J. D. Andersen, Turbo codes extended with outer BCH code, *IEE Electron. Lett.* **32**: 2059–2060 (Oct. 1996).
31. K. R. Narayanan and G. L. Stüber, Selective serial concatenation of turbo codes, *IEEE Commun. Lett.* **1**: 136–139 (Sept. 1997).
32. H. C. Kim and P. J. Lee, Performance of turbo codes with a single-error correcting BCH outer code, *Proc. 2000 IEEE Int. Symp. Information Theory*, Sorrento, Italy, June 2000, p. 369.
33. O. Y. Takeshita, O. M. Collins, P. C. Massey, and D. J. Costello, Jr., On the frame-error rate of concatenated turbo codes, *IEEE Trans. Commun.* **49**: 602–608 (April 2001).

34. D. J. Costello, Jr. and G. Meyerhans, Concatenated turbo codes, *Proc. 1996 IEEE Int. Symp. Information Theory and Applications*, Victoria, Canada, Sept. 1996, pp. 571–574.

35. M. C. Valenti, Inserting turbo code technology into the DVB satellite broadcast system, *Proc. 2000 IEEE Military Communications Conf.*, Los Angeles, Oct. 2000, pp. 650–654.

36. J. M. Shea, Improving the performance of turbo codes through concatenation with rectangular parity check codes, *Proc. 2001 IEEE Int. Symp. Information Theory*, Washington, DC, June 2001, p. 144.

37. J. M. Shea and T. F. Wong, Concatenated codes based on multidimensional parity-check codes and turbo codes, *Proc. 2001 IEEE Military Communications Conf.*, Washington, DC, Oct. 2001, Vol. 2, pp. 1152–1156.

38. 3rd Generation Partnership Project, *Technical Specification TS 25.212 v4.1.0: Radio Access Network: Multiplexing and Channel Coding* (*FDD*), Technical Report, available on the Web at *ftp://ftp.3gpp.org/Specs/2001-06/Rel-4/25_series/25212-410.zip* (June 2001).

39. 3rd Generation Partnership Project 2, *Physical Layer Standard for cdma2000 Spread Spectrum Systems-Release 0-version 3.0*, Technical Report, available on the Web at *http://www.3gpp2.org/Public_html/specs/C.S0002-0_v3.0.pdf* (July 2001).

40. S. Dolinar and D. Divsalar, *Weight Distributions for Turbo Codes Using Random and Nonrandom Permutations*, Technical Report TDA Progress Report 42-122, NASA Jet Propulsion Laboratory, Aug. 1995.

MULTIMEDIA MEDIUM ACCESS CONTROL PROTOCOLS FOR WDM OPTICAL NETWORKS

MOUNIR HAMDI
Hong Kong University of Science and Technology
Hong Kong

MAODE MA
Nanyang Technological University
Singapore

1. INTRODUCTION

Wavelength-division multiplexing (WDM) is the most promising multiplexing technology for optical networks. By using WDM, the optical transmission spectrum is configured into a number of nonoverlapping wavelength bands. In particular, multiple WDM channels could be allowed to coexist on a single fiber. As a result, the requirements on balancing the optoelectronic bandwidth mismatch could be met by designing and developing appropriate WDM optical network architectures and protocols.

WDM optical networks can be designed using one of two types of architecture: broadcast-and-select networks or wavelength-routed networks. Typically, the former is used for local-area networks, while the latter is used for wide-area networks. A local WDM optical network may be set up by connecting computing nodes via two-way fibers to a passive star coupler, as shown in Fig. 1. A node can send its information to the passive star coupler on one available wavelength by using a laser, which produces an optical stream modulated with information. Modulated optical streams from transmitting nodes are combined by the passive star coupler. Then the integrated stream is separated and transmitted to all the nodes in the network. A destination's receiver is an optical filter and is tuned to one of the wavelengths to receive its designated information stream. Communication between the source and destination nodes is implemented in one of the following two modes: *single-hop*, in which communication takes place directly between two nodes [1], or *multihop*, in which information from a source to a destination may be routed through the intermediate nodes of the network [2].

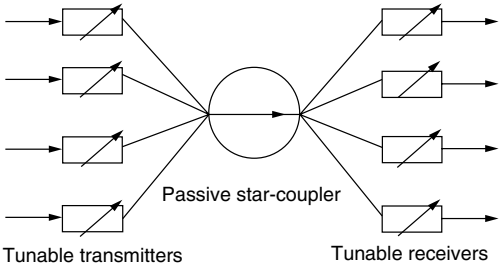


Figure 1. A single-hop passive star coupled WDM optical network.

On the basis of the architectures of WDM optical networks, medium access control (MAC) protocols are required to efficiently allocate and coordinate the system resources. The challenge of developing appropriate MAC protocols is to efficiently exploit the potential vast bandwidth of an optical fiber to meet the increasing information transmission demand under the constraints of the network resources and the constraints imposed on the transmitted information.

In this article, we will review state-of-the-art MAC protocols in passive star coupler–based WDM networks. Because the single-hop structure has more dynamics in nature than the multihop one, we will focus on the MAC protocols for the single-hop architecture. We will also discuss several protocols in some detail to show their importance in the development of MAC protocols for the single-hop passive star coupler–based WDM networks. According to the network service provided to the transmitted information, we roughly divide the MAC protocols into three categories as follows: MAC protocols for packet and variable-length message transmission, MAC protocols for real-time message transmission (including MAC protocol with QoS concerns), and MAC protocols for multimedia applications.

The remainder of this article is organized as follows. Section 2 reviews the MAC protocols of transmission service for packets and variable-length messages. Section 3 discusses the MAC protocols for real-time service. Section 4 provides an investigation on multimedia protocols. Section 5 concludes the article with a summary.

2. MAC PROTOCOLS FOR PACKET AND VARIABLE-LENGTH MESSAGE TRANSMISSION

Numerous MAC protocols for normal data transmission have been proposed since 1990 for the WDM single-hop network architecture. According to how the data are

presented and transmitted in the networks, the MAC protocols can be simply grouped as MAC protocols for fixed-size packet transmission and MAC protocols for variable-size message transmission.

2.1. MAC Protocols for Packet Transmission

The MAC protocols for packet transmission in single-hop passive star–coupled WDM networks are so called "legacy" protocols because they are dedicated for fixed-length packet transmission, and they are often adopted from legacy shared medium networks. In a single-hop network, a significant amount of dynamic coordination between nodes is required in order to access the network resources. According to the coordination schemes, the MAC protocols can be further classified into the following subcategories.

2.1.1. Nonpretransmission Coordination Protocols. Protocols with nonpretransmission coordination do not have to reserve any channels for pretransmission coordination. All the transmission channels are either preassigned to transmitting nodes or accessed by transmitting nodes through contest. These protocols can be categorized accordingly in the following subgroups.

2.1.1.1. Fixed Assignment. A simple approach, based on the fixed-wavelength assignment technique, is time-division multiplexing (TDM) extended over a multichannel environment [3]. It is predetermined that a pair of nodes is allowed to communicate with each other in the specified time slots within a cycle on the specified channel. Several extensions to the abovementioned protocol have been proposed to improve the performance. One approach, *weighted TDM*, assigns a different number of time slots to different transmitting nodes according to the traffic load on each node [4]. Another proposed approach is a versatile time–wavelength assignment algorithm [5]. Under the condition that a traffic demand matrix is given beforehand, the algorithm can minimize the tuning times and has the ability to reduce transmission delay. Some new algorithms based on the abovementioned algorithm [5] study problems such as the performance of scheduling packet transmissions with an arbitrary traffic matrix and the effect of the tuning time on the performance [6–8].

2.1.1.2. Partial Fixed Assignment Protocols. Three partial fixed assignment protocols have been proposed [3]. The first one is the destination allocation (DA) protocol. By using this protocol, the number of source and destination node pairs can be the same as the number of nodes. A source allocation (SA) protocol is also defined in which the control of access to transmission channels is further relaxed. Similar to the SA protocol, an allocation-free (AF) protocol has been proposed, in which all source–destination pairs of computing nodes have full rights to transmit packets on any channel over any time slot duration.

2.1.1.3. Random Access Protocols. Two slotted-ALOHA protocols have been proposed [9]. Using the first protocol, time is slotted on all transmission channels, and these slots are synchronized across all channels. Using the second protocol, each packet can have several numbers of minislots, and time across all channels is synchronized over minislots. In addition, two similar protocols appeared in the literature [10].

2.1.2. Pretransmission Coordination Protocols. Employing protocols that do require pretransmission coordination, transmission channels are grouped into control channels and data channels. These protocols can be categorized according to the ways to access the control channels into the following subgroups:

2.1.2.1. Random-Access Protocols. The architecture of the network protocols in this subgroup is as follows. In a single-hop communication network, a control channel is employed. Each node is equipped with a single tunable transmitter and a single tunable receiver.

Habbab et al. [11] describe three random-access protocols such as ALOHA, slotted ALOHA, and CSMA are proposed to access the control channel. ALOHA, CSMA, and the N-server switch scheme can be the subprotocols for the data channels. Under a typical ALOHA protocol, a node transmits a control packet over the control channel at a randomly selected time, after which it immediately transmits a data packet on a data channel, which is specified by the control packet.

Mehravari [12] has proposed an improved protocol, slotted-ALOHA/delayed-ALOHA. This protocol requires that a transmitting node delay transmitting data on a data channel until it gets the acknowledgment that its control packet has been successfully received by the destination node. The probability of data channel collisions can be decreased, and the performance in terms of throughput can be improved.

Sudhakar et al. [13] proposed one set of slotted-ALOHA protocols and one set of reservation-ALOHA protocols. The set of the slotted-ALOHA-based protocols are improvements over the protocols proposed by Habbab et al. [11].

A so-called multicontrol channel protocol has been proposed [14] that aims at improving reservation-ALOHA-based protocols. All channels are used to transmit control information as well as data information. Control packet transmission uses a contention-based operation; while data transmission follows it.

These protocols basically cannot prevent receiver collisions. A protocol that is especially designed to avoid receiver collision has been proposed [15].

2.1.2.2. Reservation Protocols. Using the dynamic time–wavelength-division multiple-access (DT-WDMA) protocol, a channel is reserved as a control channel and it is accessed only in a preassigned TDM fashion. It requires that each node have two transmitters and two receivers [16]. One pair of the transceivers is fixed to the control channel, while another pair is tunable to all the data channels. If there are N nodes in the network, N data channels and one control channel are required. Although this protocol cannot avoid receiver collisions, it ensures that exactly one data item can be successfully accepted when more than one data packet come to the same destination node simultaneously.

One proposal [17] to improve the TD-WDMA algorithm is to use an optical delay line to buffer the potential collided packets, when more than one node transmits data packets to the same destination node at the same time. Its effectiveness depends on the relative capacity of the buffer. Another protocol [18] also tries to improve the TD-WDMA algorithm by making transmitting nodes remember the information from the previous transmission of a control packet and combining this information into the scheduling of packet transmission.

Another two protocols [19,20] intended to improve the TD-WDMA algorithm are outlined. The first one is called the *dynamic allocation scheme* (DAS), where each node runs an identical algorithm based on a common random number generator with the same seed. The second protocol is termed *hybrid TDM*. Time on the data channels is divided into frames consisting of several slots. In a certain period of time, one slot will be opened for a transmitting node to transmit data packets to any destination receiver.

A reservation-based multicontrol channel protocol has been described [21]. Employing this protocol, x channels $[1 < x < (N/2)]$ can be reserved as control channels to transmit control information, where N is the number of channels in the network. The value of x is a system design parameter, which depends on the ratio of the amount of control information and the amount of actual data information. The objective to reserve multiple control channels in the network is to decrease the overhead of control information processing time as much as possible.

The properties of the "legacy" MAC protocols can be summarized based on the basis of the abovementioned survey as follows. Although the protocols using the fixed-channel assignment approach can ensure that data are successfully transmitted and received, they are sensitive to the dynamic bandwidth requirements of the network and are difficult to scale in terms of the number of nodes. The protocols using the contention-based channel assignment approach introduce contention on data channels in order to adapt to the dynamic bandwidth requirements. As a result, either channel collision or receiver collision will occur. The protocols with contention-based control channel assignment still have either a data channel collision or a receiver collision because contention is involved in the control channel. Some protocols [22,23] have the capability to avoid both collisions by continuously testing the network states. The reservation-based protocols, which use a fixed control channel assignment approach, can only ensure data transmission without collisions. However, by introducing some information to make the network nodes intelligent, it has the potential to avoid receiver collisions as well. It also has the potential to accommodate application traffic composed of variable-length messages.

2.2. MAC Protocols for Variable-Length Message Transmission

The "legacy" MAC protocols are designed to handle and schedule fixed-length packets. Using these MAC protocols, most of the application-level data units (ADUs) must be segmented into a sequence of fixed-size packets for transmission over the networks. However, as traffic streams in the real world are often characterized as bursty, consecutive arriving packets in a burst are strongly correlated by having the same destination node. An intuitive idea about this observation is that all the fixed-size packets of a burst should be scheduled as a whole and transmitted continuously in a WDM network rather than be scheduled on a packet-by-packet basis. Another way of looking at this is that the ADUs should not be segmented. Rather, they should be simply scheduled as a whole without interleaving. The main advantages of using a burst-based or message transmission over WDM networks are (1) to an application, the performance metrics of its data units are more relevant performance measures than ones specified by individual packets; (2) it perfectly fits the current trend of carrying IP traffic over WDM networks; and (3) message fragmentation and reassembly are not needed.

The first two MAC protocols proposed by Sudhakar et al. [13] for variable-length message transmission are protocols with contention-based control channel assignment. Another two reservation-ALOHA-based protocols [13] are presented in order to serve the long-holding-time traffic of variable-length messages. The first protocol aims to improve the basic slotted-ALOHA-based technique. The second protocol aims to improve the slotted-ALOHA-based protocol with asynchronous cycles on the different data channels. Data channel collisions can be avoided by the two reservation-ALOHA-based protocols.

Another protocol [24–26] tries to improve the reservation-based TD-WDMA protocol [16]. The number of nodes is larger than the number of channels; the transmitted data are in the form of a variable-length message rather than a fixed-length packet; data transmission can start without any delay. Both data collision and receiver collision can be avoided because any message transmission scheduling has to consider the status of the data channels as well as receivers.

Two other protocols, FatMAC [27] and LiteMAC [28], try to combine reservation-based and preallocation-based techniques to schedule variable-length message transmission. FatMAC is a hybrid approach that reserves access to preallocated channels through control packets. Transmission is organized into cycles where each of them consists of a reservation phase and a data phase. A reservation specifies the destination, the channel and the message length of the next data transmission. The LiteMAC protocol is an extension of FatMAC. Using the LiteMAC protocol, each node is equipped with a tunable transmitter and a tunable receiver rather than a fixed receiver as in FatMAC. LiteMAC has more flexibility than FatMAC because of the usage of a tunable receiver and its special scheduling mechanism. Hence, more complicated scheduling algorithms could be used to achieve better performance than FatMAC. Both FatMAC and LiteMAC have the ability to transmit variable-length messages by efficient scheduling without collisions. The performances of these two protocols have been proved to be better than that of the preallocation-based protocols, while fewer transmission channels are used than in reservation-based protocols. With these two protocols, low average message delay and high channel utilization can be expected.

2.2.1. A Reservation-Based MAC Protocol for Variable-Length Messages. Proposed [29], based on the protocol advanced by Bogineni and Dowd Jia et al. [24–26]; this was an intelligent reservation-based protocol for scheduling variable-length message transmission. The protocol employs some global information of the network to avoid both data channel collisions and receiver collisions while message transmission is scheduled. Its ability to avoid both collisions makes this protocol a milestone in the development of MAC protocols for WDM optical networks.

The network consists of M nodes and $W+1$ WDM channels. W channels are used as data channels. The other channel is the control channel. Each node is equipped with a fixed transmitter and a fixed receiver for the control channel, and a tunable transmitter and a tunable receiver to access the data channels. The time on the data channels is divided into data slots. It is assumed that there is a networkwide synchronization of data slots over all data channels. The duration of a data slot is equal to the transmission time of a fixed-length data packet. A node generates variable-length messages, each of which contains one or more fixed-length data packets. On the control channel, time is divided into control frames. A control frame consists of M control slots. A control slot has several fields such as address of destination node and the length of the message. A time-division multiple-access protocol is employed to access the control channel so that the collision of control packets can be avoided.

Before a node sends a message, it needs to transmit a control packet on the control channel in its control slot. After one round-trip propagation delay, all the nodes in the network will receive the control packet. Then a distributed scheduling algorithm is invoked at each node to determine the data channel and the time duration over which the message will be transmitted. Once a message is scheduled, the transmitter will tune to the selected data channel and transmit the scheduled message at the scheduled transmission time. When the message arrives at its destination node, the receiver should have been tuned to the same data channel to receive the message.

The data channel assignment algorithm determines the data channel and the time duration over which the message will be transmitted. The algorithm schedules message transmissions based on some global information in order to avoid the data channel collisions and the receiver collisions. The global information is expressed through two tables, which reside at each node. One table is the *receiver available-time table* (RAT). RAT is an array of M elements, one for each node. $\mathrm{RAT}[i]=n$, where $i=1,2,\ldots,M$, means that node i's receiver will become free after n data slots. If $n=0$, then node i's receiver is currently idle, and no reception is scheduled for it as yet. RAT is needed for avoiding receiver collisions. Another table is the *channel available-time table* (CAT). CAT is an array of W elements, one for each data channel. $\mathrm{CAT}[k]=m$, where $k=1,2,\ldots,W$, means that data channel k will be available after m data slots. If $\mathrm{CAT}[k]=0$, data channel k is currently available. CAT is needed to avoid collisions on data channels. Local and identical copies of these two tables are at each node. They contain consistent information on the messages whose transmissions have been scheduled but not yet transmitted. The contents of the tables are relative to current time. Three data channel assignment algorithms have been proposed. The fundamental one, the *earliest available-time scheduling* (EATS) algorithm. This algorithm schedules the transmission of a message by selecting a data channel, which is the earliest available.

This reservation-based protocol has been shown to have quite good performance while it can avoid data channel collisions and receiver collisions.

2.2.2. A Receiver-Oriented MAC Protocol for Variable-Length Messages. Some related protocols have been proposed to improve the performance of the network based on the same system architecture of Ref. 29. In Ref. 30, the proposed protocol tries to avoid the head-of-queue blocking during the channel assignment procedure by introducing the concept of "destination queue" to make each node maintain M queues, where M is the number of nodes in the network. Hamidzadeh et al. [31], notice that the performance of the network could be further improved by the way of exploiting more existing global information of the network and the transmitted messages. From this point of view, a general scheduling scheme, which combines the message sequencing techniques with channel assignment algorithms [29], is proposed to schedule variable-length message transmission. In Ref. 32, as an example of the general scheduling scheme in Ref. 31, a new scheduling algorithm is proposed. This algorithm, *receiver-oriented earliest available-time scheduling* (RO-EATS), decides the sequence of the message transmission using the information of the receiver's states to decrease message transmission blocking caused by avoiding collisions.

The RO-EATS scheduling algorithm employs the same system structure and network service as those of the protocol in [29] to form a receiver-oriented MAC protocol, which is an extension of the protocol [29]. The logic structure of the system model for the RO-EATS protocol can be expressed as in Fig. 2. Employing the new protocol, the management of messages' transmission and reception is the same as that of the Jia et al. protocol [29]. The difference between the two protocols is in the scheduling algorithm for message transmission. The RO-EATS algorithm works as follows. It first considers the

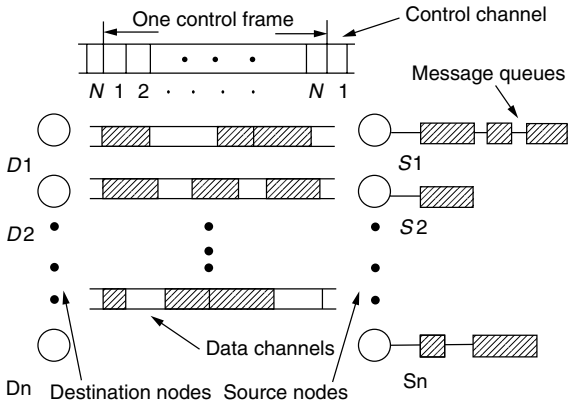


Figure 2. The network architecture for variable-length messages.

earliest available receiver among all the nodes in the network and then selects a message, which is destined to this receiver from those, which are ready and identified by the control frame. After that, a channel is selected and assigned to the selected message by the principle of the EATS algorithm. The scheme to choose a suitable message to transmit is based on the information of the states of the receivers presented in RAT. The objective of this scheme is to avoid lots of messages going to one or a few nodes at the same time and try to raise the channels utilization. The motivation of the algorithm comes from the observation that two consecutive messages with the same destination may not fully use the available channels when the EATS algorithm is employed. The new algorithm enforces the idea of scheduling two consecutive messages away from going to the same destination node. The RO-EATS algorithm always checks the table of RAT to see which node is the least visited destination and to choose the message, which is destined to this node to transmit. In this way, the average message delay can be shown to be quite low and channel utilization can be shown to be high.

3. MAC PROTOCOLS FOR REAL-TIME SERVICE

An important function of high-speed computer networks such as WDM optical networks is to provide real-time service to time-constrained application streams such as video or audio information. Most of the MAC protocols that provide real-time service on passive star–coupled WDM optical networks are protocols with reservation-based precoordination. According to the type of the real-time service provided to the transmitted messages, the MAC protocols for real-time service can be simply classified into two types: protocols with best-effort service and protocols with quality-of-service (QoS) capabilities.

3.1. MAC Protocols for Best-Effort Real-Time Service

A protocol termed *time-deterministic time- and wavelength-division multiple access* (TD-TWDMA) [33] provides services for both hard real-time messages and soft real-time messages for single destination, multicast, and broadcast transmissions. All channels can be accessed by a fixed-assignment method, which is a TDM approach. Using this approach, each channel is divided into time slots. Each node has a number of slots for hard real-time message transmission. Soft real-time messages can be transmitted if there is no hard real-time message requiring service. Each node is equipped with one fixed transmitter and tunable receivers. The transmitter is fixed to its assigned channel, while the receiver can be tuned over all channels in the network. Each node has a specified channel because the number of nodes, C, is equal to the number of channels, M, in the network. At each node, there are $2 * M$ queues, M queues for the hard real-time messages, and another M queues for the soft real-time messages. For each type of queue, one queue is for broadcast and $M - 1$ queues for the single destination. The messages in the broadcast queue can be either control information or data to be broadcast. The protocol works as follows. First it sends a broadcast slot containing the control information; then it invokes the slot allocation algorithm to determine the slots used to transmit the data information; Finally, each node tunes to the specified channel to receive the data. The slot allocation algorithm follows the static priority approach. The basic idea of the algorithm can be summarized as follows: (1) the M hard real-time message queues have higher priorities; while the M soft real-time message queues have lower priorities; (2) each queue in each group has a fixed priority, while the queues for broadcasting have the highest priority in each group; (3) message transmission scheduling is based on the queue priority; and (4) for the hard real-time messages, if transmission delay is over their deadlines, these messages will be dropped; while for the soft real-time messages, they will be scheduled whether they are beyond their deadlines or not.

A reservation-based MAC protocol for best-effort real-time service can be found in the literature [34]. This protocol is for the same network architecture as that in Jia et al. [29]. Both hard real-time and soft real-time variable-length message transmissions have been considered. The scheduling algorithms of the protocol are based on the time-related dynamic priority scheme, namely, *minimum laxity first* (MLF) scheduling. The principle of this dynamic scheduling scheme is that the most stringent message will get the transmission service first. This protocol employs global information of the network as well as the transmitted messages to ensure zero message loss rate caused by both data channel collisions and receiver collisions and decrease the message loss rate caused by network delay. This research work has confirmed that when real-time traffic is introduced in the networks, dynamic time-based priority assignment schemes as well as priority-based scheduling algorithms should be employed to improve the real-time performance of the networks as much as possible.

A novel reservation-based MAC protocol for real-time service has been proposed [35] that extends the functions of the protocol in Ma et al. [34] to provide differentiated service to benefit both real-time and non-real-time applications in one topology. This protocol considers the transmission of the variable-length messages with hard real-time constraints, soft real-time constraints, or non-real-time constraints. The scheduling algorithm, *minimum laxity first with time tolerance scheduling* (MLF-TTS), *algorithm*, of this protocol schedules real-time message transmission according to their time constraints. The basic minimum laxity first (MLF) scheduling policy is adopted for scheduling real-time traffic. For non real-time messages, the scheduling algorithm manages their transmission based on the following fact of the transmission of real-time messages. After the real-time messages have been scheduled for transmission on certain channels in certain time slots to their destination nodes, some of them could be blocked just because there may be more than two consecutive messages going to the same destination node in a very short time period. This fact causes the utilization of the transmission channels to be quite low, and succeeding messages will be blocked so that the average message delay for the non-real-time messages will be very high. The MLF-TTS algorithm seeks and takes a time period, in which the real-time messages are

being blocked to wait for their destinations to be free, to schedule the transmission of non-real-time messages under the condition that the transmission time of these messages should be less than the time that the blocked real-time messages are waiting for their destinations to be available. Since the global information of the receivers and channels in the network are available to every source node, this scheduling is feasible and can be easily implemented. Using the MLF-TTS algorithm, the average message delay for the messages without time constraints could be expected to decrease while the message loss rate or message tardy rate is kept as low as those of the simple MLF algorithms. In addition, the channel utilization could be expected to be high. Unlike the scheduling algorithms, which aim to only decrease the average message delay, the MLF-TTS could be expected to significantly increase the real-time performance of the WDM MAC protocols. As a result, a fairness transmission service to both real-time and non real-time traffics could be achieved.

3.2. MAC Protocols with Quality-of-Service Concerns

Quality of service (QoS) is an important issue when real-time applications demand a given network transmission service. It is obvious that the QoS provided by a network service to real-time applications indicates the degree to which the real-time applications can meet their time constraints. However, best-effort real-time network service cannot ensure QoS because it cannot guarantee that real-time applications can meet their time constraints to a certain degree when they are transmitted. It is necessary to develop MAC protocols with QoS capabilities so that the QoS guaranteed by the network service could be estimated and predicted. There are two types of MAC protocols with QoS capabilities: protocols with deterministically guaranteed service and protocols with statistically guaranteed service.

3.2.1. MAC Protocols with Deterministically Guaranteed Service. A preallocation-based channel access protocol has been proposed [36] to provide deterministic timing guarantees to support time-constrained communication in a single-hop passive star−coupled WDM optical network. This protocol takes a passive star−coupled broadcast-and-select network architecture in which N stations are connected to a passive star coupler with W different wavelength channels. Each W channel is slotted and shared by the N stations by means of a TDM approach. The slots on each channel are preassigned to the transmitters. A schedule specifies, for each channel, which slots are used for data transmission from node i to node j, where $1 \le i \le N,\ 1 \le j \le N,\ i \le j$. Each node of the network can be equipped with a pair of tunable transmitters and tunable receivers, which can be tuned over all the wavelengths. Each real-time message stream with source and destination nodes specified is characterized with two parameters, relative message deadline D_i and maximum message size C_i, which can arrive within any time interval of length D_i. A scheme called a *binary splitting scheme* (BSS) is proposed to assign each message stream sufficient and well-spaced slots to fulfill its timing requirement. Given a set of real-time message streams M specified by the maximum length of each stream C_i and the relative deadline of each stream D_i, this scheme can allocate time slots over as few channels as much as possible in such a way that at least C_i slots are assigned to M_i in any time window of size D_i slots so that the real-time constraints of the message streams can be guaranteed.

A modified preallocation-based MAC protocol has been proposed [37] to guarantee a reserved bandwidth and a constant delay bound to the integrated traffic. This protocol works as a centralized scheduler based on the star-coupled broadcast LAN topology, which is similar to that proposed by Jia et al. [29]. Each node in the LAN is equipped with a pair of tunable transceivers. The access to the transmission channels is controlled by the scheduler, which is based on the concept of computing maximal weighted matching, a generalization of maximal matching on an unweighted graph. According to this concept, several scheduling algorithms have been proposed. A *credit-weighted algorithm* is proposed to serve guaranteed traffic. A *bucket-credit-weighted algorithm* is designed to serve bursty traffic, and a *validated queue algorithm* is a modification of the *bucket-credit-weighted algorithm* to serve bursty traffic and keep throughput guarantee at the same time. It has been proved that these scheduling algorithms can guarantee the bandwidth reservation to a certain percentage of the network capacity and ensure a small delay bound even when bursty traffic exists.

A reservation-based MAC protocol for deterministic guaranteed real-time service has been proposed [38]. This protocol is for the same network structure as that in Jia et al. [29]. In this protocol [38], a systematic scheme is proposed to provide deterministic guaranteed real-time service for application streams composed of variable-length messages. It includes an admission control policy, traffic regularity, and message transmission scheduling algorithm. A traffic-intensity-oriented admission control policy is developed to manage flow-level traffic. A g-regularity scheme based on the max-plus algebra theory is employed to shape the traffic. An *adaptive round-robin and earliest available time scheduling* (ARR-EATS) algorithm is proposed to schedule variable-length message transmission. All of these are integrated to ensure that a deterministic guaranteed real-time service can be achieved.

3.2.2. MAC Protocols for Statistically Guaranteed Service. The MAC protocols with deterministically guaranteed service can normally guarantee specific transmission delays to real-time applications, or, under certain time constraints imposed to the real-time applications, a specific percentage of real-time messages, which can meet the time constraints, can be predicted. However, MAC protocols for statistically guaranteed service cannot provide a deterministic guaranteed QoS service. Only an estimated percentage of real-time messages, which can meet their time constraints, can be evaluated statistically. Most of the MAC protocols in this category consider the issue of providing differentiated service to both real-time and non-real-time applications. Using these protocols, statistical QoS to real-time applications can be expected by sacrificing the transmission service to non-real-time applications.

A reservation-based protocol has been proposed [39] to provide statistically guaranteed real-time services in WDM optical token LANs. In the network, there are M nodes and $W+1$ channels. One of the channels is the control channel, while the others are data channels. Different from the network structure presented by Jia et al. [29], the control channel in this network is accessed by token passing. At each node, there is a fixed receiver and transmitter tuned to the control channel. There is also a tunable transmitter, which can be tuned to any of the data channels. There are one or more receivers fixed to certain data channels. The protocol provides transmission service to either real-time or non-real-time messages. The packets in the traffic may have variable length but be bounded by a maximum value. At each node, there are W queues, each of which corresponds to one of the channels in the network. The messages come into one of the queues according to the information of their destination nodes and the information of the channels, which connect to the corresponding destination nodes. The protocol works as follows. A token exists on the control channel to ensure collision-free transmission on data channels. The token has a designated node K. Every node can read the contents of the token and updates its local status table by the information in the fields of the token. When node K observes the token on the network, it will check the available channels. If there are no channels available, node K gives up this opportunity to send its queued packets. Otherwise, the *priority index algorithm* (PIA) is invoked to evaluate the priority of each message queue on node K and then uses the *transmitter scheduling algorithm* (TSA) to determine the transmission channel. Also the *flying-target algorithm* (FTA) is used to decide the next destination of the control token. After all these have completed, node K's status and scheduling result will be written into the token. Then the token on node K will be sent out, and the scheduled packets will be transmitted.

A novel reservation-based MAC protocol has been proposed [40] to support statistically guaranteed real-time service in WDM networks by using a hierarchical scheduling framework. This work is developed for a network structure similar to that described by Jia et al. [29]. The major advantage of its protocol over that proposed by Yan et al. [39] is that it divides the scheduling issue into flow scheduling or VC scheduling and transmission scheduling. The former is responsible for considering the order of traffic streams to be transmitted. It schedules packets to be transferred from VC queues to the transmission queue. The latter is to decide the order of the packets transmission. The packets involved in the transmission scheduling are those selected from the traffic streams by the flow schedule scheme. A simple-round robin scheme is adopted in the VC scheduling, and a random scheduling with age priority is used in the transmission scheduling. Another good point of this protocol is that a rescheduling scheme is employed to compensate the failure scheduling result due to either output conflict or channel conflict. Using this scheme, if a scheduling fails, a decision has to be made as to whether rescheduling the same packet or scheduling a new packet from another VC is performed. If the failure is from a real-time traffic, it certainly makes sense to reschedule the very same packet as soon as possible. The very same real-time packet will be retransmitted immediately in the next control slot; thus no other new scheduling either from real-time traffic or non-real-time traffic of the same source node can be initiated. The more intriguing part of the scheduling algorithm is the rescheduling of non-real-time traffic. If real-time traffic has more stringent QoS requirements, the rescheduling scheme will ignore rescheduling the failed non-real-time packet to ensure that the real-time traffic meets its time constraints. Compared with the protocol proposed by Yan et al. [39], this protocol is expected to diminish the ratio of the packet, which are over their deadlines.

A protocol similar to that [39] has been proposed [41]. This protocol is based on the same network architecture as that described by Yan et al. [39], which is a WDM optical token LAN. The protocol tries to provide fairness transmission service to both real-time and non-real-time messages in the same network, while the QoS of the real-time messages could be adjusted to a reasonable level. The protocol separates the real-time and non-real-time messages into different queues at each transmission node. The real-time message queue has higher priority for transmission than does that for non-real-time messages. The outstanding point of this protocol is that the scheduling scheme of the protocol has set up a threshold on the queue length for the non-real-time message queue in order to balance the transmission service. The operation of the scheduling scheme works as follows. When the length of the non-real-time message queue has not reached the threshold, the real-time messages will be scheduled for transmission. However, when the threshold has been reached or exceeded, the non-real-time messages will be scheduled for transmission until the length of the lower-priority queue is under the threshold. The QoS of the real-time message transmission is measured by the loss rate. With setting of the threshold, the scheduling scheme can provide fair transmission service to both real-time and non-real-time traffic with certain QoS guarantee to real-time traffic. Alternatively, with a change of the threshold, the level of QoS guarantee to real-time traffic can be controlled.

A MAC protocol to provide statistical QoS service has been presented [42], that is based on a multichannel ring topology. This topology is somewhat different from that described by Yan et al. [39]. In this network, every node is equipped with a fixed receiver and a tunable transmitter. Every transmission channel is associated with each destination node. The transmitted information is in the form of fixed-size packets. A collision-free MAC protocol is proposed, known as *synchronous round-robin with reservation* (SR^3), to support both QoS guarantee to real-time traffic and best-effort service to non-real-time traffic. There are three components in the SR^3 protocol. The access strategy selects proper packets to transmit, the fairness control algorithm guarantees the throughput fairness among all the channels in the multiring network, and the reservation scheme allows the transmitting nodes to dynamically allocate a portion of available bandwidth. This protocol has been proved to have the capability to

provide quality of service to both real-time and non-real-time traffic in the multiring topology.

4. MAC PROTOCOLS FOR MULTIMEDIA APPLICATIONS

There has been a rapid growth in the number of multimedia applications. Different multimedia applications require various classes of transmission service, including the transmission of data, audio, and various types of video and images on WDM optical networks. High-speed protocols including the protocols at the medium access control layer are needed to cater for the different requirements of the transmission of various multimedia applications.

Multimedia applications contain a variety of media: data, graphics, images, audio, and video. The transmission of the multimedia applications is a kind of real-time and stream-oriented communication. The quality of service required of a stream communication includes guaranteed bandwidth (throughput), delay, and delay variation (jitter). However, the quality of service for different kinds of applications varies. On one hand, hard real-time traffic such as voice and video require stringent time delay and delay variance, but tolerates a small percentage of packet loss. On the other hand, soft or non-real-time traffic such as images, graphics, text, and data requires no packet loss, but tolerates time delay. Hence, the protocols that provide transmission service to multimedia applications should support and ensure the variety of QoS requirements of different types of media.

Support of multimedia applications by MAC protocols has become a hot topic in the field of research on WDM optical networks. Some research results have been generated from existing protocols for real-time service. However, some protocols are completely novel or based on new network architectures dedicated to multimedia traffic.

4.1. Modified Protocols for Multimedia Applications

The feasibility of several existing protocols based on WDM bus LAN architecture to support multimedia applications has been studied [43]. Using a simulation study, the authors point out that several currently existing MAC protocols such as FairNet, WDMA, and nDQDB are not satisfactory for supporting multimedia traffic in the sense that these protocols cannot guarantee that the total delay or jitter will not grow beyond the accepted value for different classes of multimedia applications. A further study on several MAC protocols to support multimedia traffic on WDM optical networks has been carried out [44]. These protocols, including distributed queue dual bus (DQDB), cyclic-reservation multiple access (CRMA), distributed-queue multipleaccess (DQMA), and fair distributed queue (FDQ) are distributed reservation access schemes for WDM optical networks, based on slotted unidirectional bus structures. The performance of these four protocols is studied to simultaneously support synchronous traffic (for various real-time multimedia applications) and asynchronous traffic (for interactive terminal activities and data transfers). The authors have pointed out, through extensive simulation results, that the reservation-based protocols are suitable for integrating real-time multimedia traffic with bursty data traffic in the WDM optical network when the delay constraint is somewhat relaxed. The FDQ protocol stands out for supporting heterogeneous traffic. The results from the two studies cited above imply that the reservation-based protocols have the potential to accommodate the multimedia transmission in the WDM optical network rather than the nonprecoordination-based protocols.

A video-on-demand system over WDM optical networks has been studied [45]. The video-on-demand (VOD) application is considered different from live video application or MPEG compressed video/audio application in that the live video traffic is a variable-bit-rate (VBR) traffic because the video/audio sources of the application should be captured, compressed and then transmitted in real-time fashion; while VOD traffic is a constant-bit-rate (CBR) traffic because the video/audio sources of the application are processed in advance, kept on the video server, and transmitted at a regular rate. The VOD traffic is desirable to be served by isochronous transmission service by the network. The network structure of the VOD system [45] is a passive star coupler–based WDM optical network. Each node in the network is equipped with multiple tunable transmitters and receivers for data transmission and one pair of fixed transceivers for the control channel access. A centralized medium access control scheduler is employed to schedule the isochronous and the asynchronous traffic demands. A scheduling algorithm, KT-MTR, is employed for scheduling the asynchronous traffic only. Another scheduling algorithm, IATSA-MTR, is presented for scheduling both isochronous and asynchronous traffic that coexist in the network. These scheduling algorithms are shown to be efficient for serving VOD applications in star coupler–based WDM optical networks.

In order to efficiently support various types of traffic streams with different characteristics and QoS requirements in a single WDM optical network and to dynamically allocate the network bandwidth to the different classes of traffic so that the network performance could be boosted, a novel approach to integrate different types of existing medium access control protocols into a single MAC protocol in the specified WDM network architecture has been proposed [46]. The network architecture for this unique MAC protocol is similar to that proposed by Jia et al. [29], which is a passive star coupler–based WDM optical LAN. The main difference between these architectures is that each transmitting node in the architecture described by Wang and Hamdi [46] has been equipped with three pairs of tunable transceivers for three types of traffic streams working in a pipeline fashion to reduce the tuning overhead. Three types of multimedia traffic streams, including a constant–bit-rate traffic, a variable-bit-rate traffic with large burstiness, and a variable-bit-rate traffic with longer interarrival times, are considered by the proposed MAC protocol, known as the *multimedia wavelength-division multiple access* (M-WDMA) scheme. The M-WDMA protocol consists of three subprotocols. One is the TDM subprotocol, which is an interleaved TDMA MAC protocol. The second is a reservation-based subprotocol, RSV, which controls the access to the data channels by using a

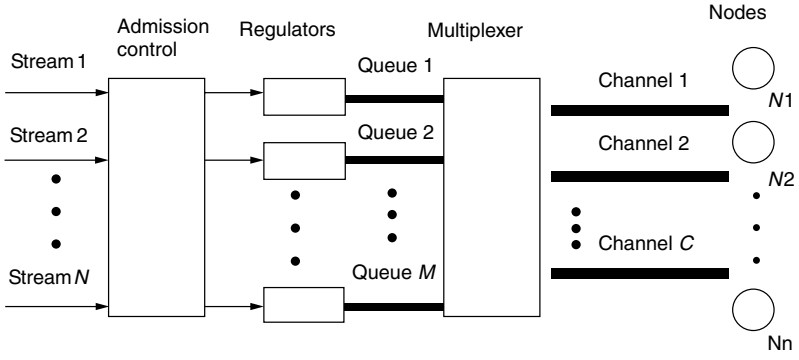


Figure 3. The logic model of the proposed multimedia systematic scheme.

multiple-token method. The third one is a random-access subprotocol, CNT, which works in a way similar to that of the interleaved slotted ALOHA. The outstanding point of this protocol is that a dynamic bandwidth allocation scheme is incorporated into the protocol to dynamically adjust the portions of the bandwidth occupied by the three types of traffic streams according to their QoS demands. The adaptation on the bandwidth allocation can be implemented by adjusting the segment sizes of the timeframe because the different classes of transmissions are grouped into a single timeframe. With knowledge of the size of each segment of a frame, bandwidth allocation can become flexible. A priority scheme is set up to allocate the bandwidth. The TDM requirements have the highest priority, while the requirements of RSV will be considered before the CNT requirements. Using an analytic model and simulation experiments, it has been proved that the performance of the M-WDMA is adequate for WDM optical networks in serving multimedia applications.

4.2. Novel Protocols for Multimedia Applications

A novel MAC protocol for providing guaranteed QoS service to multimedia applications in a WDM optical network [47]. The architecture considered in this network [47] is the same as that in Jia et al. [29]. The MPEG compressed video/audio applications are considered being transmitted in passive star–coupled WDM optical network. The QoS of transmission of the MPEG traffic is based on the frame size traces from the MPEG encoded real video sequences. A frame, which is considered as the basic element with variable size of the MPEG traffic streams, is scheduled and transmitted at one time. A systematic scheme is proposed to guarantee the deterministic delay to the transmission of the MPEG traffic. This scheme includes an admission policy, a traffic characterization mechanism, and a scheduling algorithm as a whole to ensure the QoS of the transmission of MPEG traffic. It is a distributed scheme in the sense that every function of the scheme is implemented at each transmitting node. The logical model of this scheme is shown in Fig. 3.

The admission control scheme is designed to basically limit the number of MPEG traffic entering the network. It is assumed that there are n MPEG traffic sources already connected to the network. There are another m new MPEG traffic sources requesting guaranteed bounded delay service. The admission control scheme employs a transmission bandwidth availability test algorithm to decide whether all the m new traffic sources can be accepted or rejected or which subset of them can be accepted. The traffic characterization scheme simply delays the transmission if the admitted traffic sources do not conform to their characterization to avoid excessive traffic entering into the network. The policy adopted for traffic characterization is based on the concept of the regularity of the marked point process, which is used to model the MPEG traffic. The scheduling algorithm, *adaptive round-robin and earliest available time scheduling* (ARR-EATS), is used as a virtual multiplexer to schedule the transmission of each message in the multiple MPEG traffic sources so that the deterministic delay to each frame of the MPEG streams can be guaranteed.

Analytic evaluation of the guaranteed deterministic delay bound for the proposed system service schemes is based on the theory of max-plus algebra. The deterministic delay bound is verified by intensive trace-driven simulations and modelled MPEG traffic simulations. It is proved that the proposed scheme is efficient and feasible in providing deterministic guaranteed QoS transmission service to the MPEG multimedia applications in the specified WDM optical networks. It is obvious that this protocol stands out as a state-of-the-art MAC protocol among most MAC protocols, which support QoS of the transmission of multimedia applications in WDM optical networks.

An interesting idea regarding the architecture of the WDM optical network has been proposed [48], to support varieties of traffic such as data, real-time traffic, and multicast/broadcast service. The proposed architecture, *hybrid optical network* (HONET), tries to combine the single-hop and multihop WDM optical network architectures into a synergy architecture based on the observation that a multihop network architecture does not efficiently support real-time traffic and multicast/broadcast service because of possible delay fluctuation. A single-hop network is not suitable for providing packet-switched service. The architecture of the HONET can be considered as a network that consists of the multihop network with an arbitrary virtual topology and a single-hop network based on a dynamically assigned T/WDMA MAC protocol. The virtual structured HONET can provide more flexibility in satisfying different application's traffic demands by allowing different network configurations. In this virtual network architecture, real-time traffic and other connection-oriented applications can

be supported by a single-hop network, while non-real-time data traffic, which can tolerate relatively large delay, is supported by a multihop network. The advantage of this virtual architecture is that it is flexible, capable of employing different topologies of the multihop network and different MAC protocols for the single-hop network to support varieties of traffic in the optical network according to the QoS of the traffic demands.

5. SUMMARY

This article has summarized state-of-the-art medium access control protocols for wavelength-division multiplexing (WDM) networks, especially for the passive star–coupled WDM optical networks. Depending on the characteristics, complexity, and capabilities of these MAC protocols, we have classified them as data and message transmission MAC protocols, MAC protocols for real-time transmission service, and MAC protocols for multimedia applications. Most of these protocols focus on local- and metropolitan-area environments. Architectural, qualitative, and quantitative descriptions of various protocols within each category have been provided. Some important or milestone protocols have been given quite detailed explanations to present their underlined significance. It is explicit that real-time message transmission with QoS demands and transmission service to multimedia applications with QoS requirements are currently needed to be supported by the MAC protocols on the WDM optical networks. This article can be used as a good starting point for researchers working on this area to give them an overview of the research efforts conducted since 1990. In addition, the article presents the fundamentals for further investigation into ways of coping with the current and the anticipated explosion of multimedia information transfer.

BIOGRAPHIES

Mounir Hamdi received the B.S. degree in computer engineering (with distinction) from the University of Louisiana in 1985, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh in 1987 and 1991, respectively. Since 1991, has been a faculty member in the Department of Computer Science at the Hong Kong University of Science and Technology, where he is now Associate Professor of Computer Science and the Director of the Computer Engineering Programme. In 1999 and 2000, he held visiting professor positions at Stanford University and the Swiss Federal Institute of Technology.

His general areas of research are in networking and parallel computing, in which he has published more than 130 research publications, and for which he has been awarded more than 10 research grants. He has graduated more than 10 MS and PhD students in the area of study. Currently, he is working on high-speed networks including the design, analysis, scheduling, and management of high-speed switches/routers, wavelengeh division multiplexing (WDM) networks/switches, and wireless networks.

Dr. Hamdi has been on the editorial board of *IEEE Transactions on Communications, IEEE Communication Magazine, Computer Networks, Wireless Communication* and *Mobile Computing*, and *Parallel Computing*, and has been on the program committees of mor than 50 international conferences and workshops. He was a guest editor of *IEEE Communications Magazine* and *Informatica*. He received the best paper award at the International Conference on Information and Networking in 1998 out of 152 papers. He received the best 10 lecturers award and the distinguished teaching award from the Hong Kong University of Science and Technology. He is a member of IEEE and ACM.

Maode Ma received the B.S. degree in automatic control in 1982 from Tsinghua University, Beijing, China; the M.s. degree in computer engineering in 1991 from Tianjin University, Tianjin, China; and the Ph.D. degree in computer science from Hong Kong University of Science and Technology in 1999. He joined the computer industry in 1982 as an engineer. In 1986, he was a system enginer at Tianjin University. Starting from 1991, he was an assistant professor in the department of computer engineering at Tianjin University. Since 2000, Dr. Ma has joined the school of electrical and electronic engineering at Nanyang Technological University in Singapore as an assistant professor. Dr. Ma has published approximately 20 academic papers in the areas of WDM optical networks. His areas of research interest are performance analysis of computer networks, optical networks, and wireless networks.

BIBLIOGRAPHY

1. B. Mukherjee, WDM-based local lightwave networks — Part I: Single-hop systems, *IEEE Network* 12–27 (May 1992).
2. B. Mukherjee, WDM-based local lightwave networks — Part II: Multi-hop systems, *IEEE Network* 20–32 (July 1992).
3. I. Chlamtac and A. Ganz, Channel allocation protocols in frequency-time controlled high speed networks, *IEEE Trans. Commun.* **36**(4): 430–440 (April 1988).
4. G. N. Rouskas and M. H. Ammar, Analysis and optimization of transmission schedules for single-hop WDM networks, *IEEE/ACM Trans. Network.* **3**(2): 211–221 (April 1995).
5. A. Ganz and Y. Gao, Time-wavelength assignment algorithms for high performance WDM star based systems, *IEEE Trans. Commun.* **42**(2–4): 1827–1836 (Feb.–April 1994).
6. G. R. Pieris and G. H. Sasaki, Scheduling transmissions in WDM broadcast-and-select networks, *IEEE/ACM Trans. Network.* **2**(2): 105–110 (April 1994).
7. M. S. Borella and B. Mukherjee, Efficient scheduling of nonuniform packet traffic in a WDM/TDM local lightwave network with arbitrary transceiver tuning latencies, *IEEE J. Select. Areas Commun.* **14**(6): 923–934 (June 1996).
8. M. Azizoglu, R. A. Barry, and A. Mokhtar, Impact of tuning delay on the performance of bandwidth-limited optical broadcast networks with uniform traffic, *IEEE J. Select. Areas Commun.* **14**(6): 935–944 (June 1996).
9. P. W. Dowd, Random access protocols for high speed interprocessor communication based on an optical passive star topology, *IEEE/OSA J. Lightwave Technol.* **9**(6): 799–808 (June 1991).

10. A. Ganz and Z. Koren, WDM passive star protocols and performance analysis, *Proc. IEEE INFOCOM'91*, April 1991, pp. 991–1000.

11. I. M. I. Habbab, M. Kavehrad, and C.-E. W. Sundberg, Protocols for very high speed optical fiber local area networks using a passive star topology, *IEEE/OSA J. Lightwave Technol.* **5**(12): 1782–1794 (Dec. 1987).

12. N. Mehravari, Performance and protocol improvements for very high-speed optical fiber local area networks using a passive star topology, *IEEE/OSA J. Lightwave Technol.* **8**(4): 520–530 (April 1990).

13. G. N. M. Sudhakar, M. Kavehrad, and N. Georganas, Slotted ALOHA and reservation ALOHA protocols for very high-speed optical fiber local area networks using passive star topology, *IEEE/OSA J. Lightwave Technol.* **9**(10): 1411–1422 (Oct. 1991).

14. G. N. M. Sudhakar, N. Georganas, and M. Kavehrad, Multicontrol channel for very high-speed optical fiber local area networks and their interconnections using passive star topology, *Proc. IEEE GLOBECOM'91*, Dec. 1991, pp. 624–628.

15. F. Jia and B. Mukherjee, The receiver collision avoidance (RCA) protocol for a single-hop lightwave network, *IEEE/OSA J. Lightwave Technol.* **11**(5–6): 1052–1065 (May/June 1993).

16. M.-S. Chen, N. R. Dono, and R. Ramaswami, A media access protocol for packet-switched wavelength division multiaccess metropolitan area networks, *IEEE J. Select. Areas Commun.* **8**(8): 1048–1057 (Aug. 1990).

17. I. Chlamtac and A. Fumagalli, Quadro-stars: High performance optical WDM star networks, *Proc. IEEE GLOBECOM'91*, Dec. 1991, pp. 1224–1229.

18. M. Chen and T.-S. Yum, A conflictfree protocol for optical WDM networks, *IEEE GLOBECOM'91*, Dec. 1991, pp. 1276–1291.

19. R. Chipalkatti, Z. Zhang, and A. S. Acampora, High-speed communication protocols for optical star networks using WDM, *Proc. IEEE INFOCOM'92*, May 1992, pp. 2124–2133.

20. R. Chipalkatti, Z. Zhang, and A. S. Acampora, Protocols for optical star-coupler network using WDM: Performance and complexity study, *IEEE J. Select. Areas Commun.* **11**(4): 579–589 (May 1993).

21. P. A. Humblet, R. Ramaswami, and K. N. Sivarajan, An efficient communication protocol for high-speed packet-switched multichannel networks, *IEEE J. Select. Areas Commun.* **11**(4): 568–578 (May 1993).

22. H. Jeon and C. Un, Contention-based reservation protocols in multiwavelength optical networks with a passive star topology, *Proc. IEEE ICC*, June 1992, pp. 1473–1477.

23. J. H. Lee and C. K. Un, Dynamic scheduling protocol for variable-sized messages in a WDM-based local network, *IEEE/OSA J. Lightwave Technol.* **14**(7): 1595–1600 (July 1996).

24. K. Bogineni and P. W. Dowd, A collisionless media access protocol for high speed communication in optically interconnected parallel computers, *Proc. SPIE* **1577**: 276–287 (Sept. 1991).

25. P. W. Dowd and K. Bogineni, Simulation analysis of a collisionless multiple access protocol for a wavelength division multiplexed star-coupled configuration, *Proc. 25th Annual Simulation Symp.* April 1992.

26. K. Bogineni and P. W. Dowd, A collisionless multiple access protocol for a wavelength division multiplexed star-coupled configuration: Architecture and performance analysis, *IEEE/OSA J. Lightwave Technol.* **10**(11): 1688–1699 (Nov. 1992).

27. K. M. Sivalingam and P. W. Dowd, A multilevel WDM access protocol for an optically interconnected multiprocessor system, *IEEE/OSA J. Lightwave Technol.* **13**(11): 2152–2167 (Nov. 1995).

28. K. M. Sivalingam and P. W. Dowd, A lightweight media access protocol for a WDM-based distributed shared memory system, *Proc. IEEE INFOCOM'96*, 1996, pp. 946–953.

29. F. Jia, B. Mukherjee, and J. Iness, Scheduling variable-length messages in a single-hop multichannel local lightwave network, *IEEE/ACM Trans. Network.* **3**(4): 477–487 (Aug. 1995).

30. A. Muir and J. J. Garcia-Luna-Aceves, Distributed queue packet scheduling algorithms for WDM-based networks, *Proc. IEEE INFOCOM'96*, 1996, pp. 938–945.

31. B. Hamidzadeh, Maode Ma, and M. Hamdi, Message sequencing techniques for on-line scheduling in WDM networks, *IEEE/OSA J. Lightwave Technol.* **17**(8): 1309–1319 (Aug. 1999).

32. M. Ma, B. Hamidzadeh, and M. Hamdi, A receiver-oriented message scheduling algorithm for WDM lightwave networks, *Comput. Networks* **31**(20): 2139–2152 (Sept. 1999).

33. M. Jonsson, K. Borjesson, and M. Legardt, Dynamic time-deterministic traffic in a fiber-optic WDM star network, *Proc. 9th Euromicro Workshop on Real Time Systems*, June 1997, pp. 25–33.

34. M. Ma, B. Hamidzadeh, and M. Hamdi, Efficient scheduling algorithms for real-time service on WDM optical networks, *Photon. Network Commun.* **1**(2): (July 1999).

35. M. Ma and M. Hamdi, An adaptive scheduling algorithm for differentiated service on WDM optical networks, *IEEE GLOBECOM'01*, 2001.

36. H.-Y. Tyan, J. C. Hou, B. Wang, and C. Han, On supporting temporal quality of service in WDMA-based star-coupled optical networks, *IEEE Trans. Comput.* **50**(3): 197–214 (March 2001).

37. A. C. Kam, K.-Y. Siu, R. A. Barry, and E. A. Swanson, A cell switching WDM broadcast LAN with bandwidth guarantee and fair access, *IEEE/OSA J. Lightwave Technol.* **16**(12): 2265–2280 (Dec. 1998).

38. M. Ma and M. Hamdi, Providing deterministic quality-of-service guarantees on WDM optical networks, *IEEE J. Select. Areas Commun.* **18**(10): 2072–2083 (Oct. 2000).

39. A. Yan, A. Ganz, and C. M. Krishna, A distributed adaptive protocol providing real-time services on WDM-based LANs, *IEEE/OSA J. Lightwave Technol.* **14**(6): 1245–1254 (June 1996).

40. B. Li and Y. Qin, Traffic scheduling in a photonic packet switching system with QoS guarantee, *IEEE/OSA J. Lightwave Technol.* **16**(12): 2281–2295 (Dec. 1998).

41. S. Selvakennedy, A. K. Ramani, M. Y. M-Saman, and V. Prakash, Dynamic scheduling scheme for handling traffic multiplicity in wavelength division multiplexed optical networks, *Proc. 8th Int. Conf. Computer Communications and Networks*, 1999, pp. 344–349.

42. M. A. Marsan et al., All-optical WDM multi-rings with differentiated QoS, *IEEE Commun. Mag.* 58–66 (Feb. 1999).

43. J. Indulska and J. Richards, A comparative simulation study of protocols for a bus WDM architecture, *Proc. Int. Conf. Networks*, 1995, pp. 251–255.

44. W. M. Moh et al., The support of optical network protocols for multimedia ATM traffic, *Proc. Int. Con. Networks*, 1995, pp. 1–5.

45. N.-F. Huang and H.-I. Liu, Wavelength division multiplexing-based video-on-demand systems, *IEEE/OSA J. Lightwave Technol.* **17**(2): 155–164 (Feb. 1999).

46. L. Wang and M. Hamdi, Efficient protocols for multimedia streams on WDMA networks, *Proc. 12th Int. Conf. Information Networking*, 1998, pp. 241–246.

47. M. Ma and M. Hamdi, Providing guaranteed deterministic performance service to multimedia applications on WDM optical networks, *Proc. IEEE GLOBECOM'00*, 2000, Vol. 2, pp. 1171–1175.

48. M. Kovacevic and M. Gerla, HONET: An integrated services wavelength division optical network, *Proc. IEEE ICC'94*, 1994, pp. 1669–1674.

MULTIMEDIA NETWORKING

ANDREA BIANCO
Politecnico di Torino
Torino (Turin), Italy

STEFANO GIORDANO
University of Pisa
Pisa, Italy

ALFIO LOMBARDO
University of Catania
Catania, Italy

1. INTRODUCTION

Multimedia networks are an evolution of integrated networks: networks designed to support different services, each of which is provided through the exchange of a single medium. The term *multimedia* itself denotes, in fact, the integrated manipulation of different media related to a single end-user application environment. The most widely used applications in multimedia networks are data transfer between computers, video/audiostreaming, interactive telephone applications, and voice and videoconferencing. Multimedia networking therefore deals with mechanisms to support real-time and non-real-time media over digital networks.

Distributed multimedia applications have several requirements with respect to the service that are offered by the communication network. These requirements can be classified as *functional* and *traffic requirements* [1].

Functional requirements are related mainly to the support of distributed cooperative services and therefore refer to either multicast transmission or the ability to define coordinated sets of unicast transmissions. Multicast support in the network is fundamental to provide the transmission of a single copy of data to multiple receivers with the purpose of reducing the network load and the processing load at the sender.

Traffic requirements are related to the user-perceived quality of service (QoS) and therefore refer to parameters such as bandwidth, delay, and reliability (or loss probability). *Bandwidth* specifies how much data are to be transferred over a given time period. *Reliability* pertains to the loss and corruption of data. *Delay* determines the temporal relationship between data transmission and reception.

Bandwidth requirements are fundamental for most applications, although some applications, such as data transfer, may be quite tolerant to even highly variable bandwidth. Loss requirements are stringent for most applications, with the exception of uncompressed voice and video applications, which may tolerate some losses. Delay is the main constraint for real-time interactive media, due to users' delay expectations and synchronization requirements. Users' delay expectations are due mainly due to the ability to interact in real time and enforce constraints on the end-to-end average delay. Synchronization is the preservation of time constraints within a multimedia stream at the time of playout. A multimedia stream is made up of multiple monomedia datastreams related to each other by proper timing relationships; for this reason, multimedia applications require preservation of both intramedia synchronization, to maintain the temporal order in each media evolution, and intermedia synchronization, to maintain the temporal order of correlated events in the application. The former is affected by delay jitter introduced in the network; the latter is affected by *skew*, that is, the difference between the delay jitter suffered at a given time t by two synchronized media.

Note that intermedia synchronization usually involves an increase in the QoS requirements of the low constraining media as well; for instance, in a slide show session, where time relationships exist between voice and image, still-picture transmission becomes delay-sensitive because it is related to the audio evolution, whereas in a monomedia stream still-picture transmission is seldom delay-sensitive.

All the QoS parameters mentioned above are closely related — the greater the overall bandwidth required over a link compared to the link capacity, the more messages will be accumulated and the larger the buffer needed to avoid losses; the more the buffer grows, the larger the delay experienced. Moreover, different user applications often require the enforcement of different and contrasting QoS parameters. Consider, for example, a voice phone application that is sensitive to end-to-end delay and quite tolerant to losses, and a data transfer application that, on the contrary, is interested in small losses and tolerant to delay variations. Clearly, it may be difficult to satisfy both requirements at the same time when the two applications share a link, as is the case in multimedia networks. Using large buffers, for example, makes it easier to control losses but typically increases delay. Thus, the design of flexible solutions to efficiently provide a compromise between different and contrasting traffic requirements is the main challenge for multimedia networking today. To satisfy the QoS requirements of different multimedia applications, suitable traffic control and resource management strategies have to be implemented in multimedia networks.

In this article, we will first discuss the possible approaches and key algorithms to efficiently support QoS requirements in a multimedia network. Then, an historical perspective of multimedia networks evolution will be presented. Finally, we will outline the challenging multimedia network paradigms currently being defined by the IETF to support QoS in the Internet arena.

2. QOS PROVISIONING APPROACHES

A multimedia network must be aware of the characteristics of the user traffic in order to manage different traffic flows and provide the required QoS parameters. For this reason traffic sources are usually classified as constant-bit-rate (CBR) sources and variable-bit-rate (VBR) sources. CBR traffic sources are described by means of their bandwidth requirements, as, for example, in the case of a 64 kbps (kilobits per second) digital telephone call. CBR traffic sources are exactly predictable given their bandwidth requirements, since the same amount of data is generated at fixed time instants. Conversely, VBR traffic sources, such as those generated by data transfer applications, are unpredictable because the number of packets and their emission time depend on terminal workload, software implementation and network access protocols. For this reason, VBR sources are described by a set of parameters that have to be representative of the traffic statistics. The most popular set of parameters are the average bandwidth, that is, the bit rate averaged over the flow duration, and peak bandwidth — the maximum bit rate over the flow duration. CBR traffic can be managed more easily by the network, because in this case deterministic rules may be used to assign network resources to the CBR source; this is the approach used, for example, in circuit-switched networks such as ISDN or telephone networks. Unfortunately, even CBR sources, when compressed to reduce their bit rate, become unpredictable sources of traffic. Thus, multimedia networks must deal primarily with VBR sources.

A first approach to supporting VBR traffic with the requested loss and delay requirements would be to overdesign network resources based on VBR source peak bandwidth demand; by so doing, in fact, VBR traffic is managed as CBR traffic simply ignoring VBR traffic fluctuations. Such an approach, often named *overprovisioning*, can be efficient in terms of resource utilization only if VBR traffic is characterized by a low ratio between peak bandwidth and average bandwidth, usually referred to as "burstiness"; if this is not the case, this approach results in low resource utilization and, therefore, high costs. The overprovisioning approach is used today to design access networks based on LAN technologies where no explicit mechanism to support QoS is provided. Several researchers believe that bandwidth costs will drop dramatically in the near future, thanks to the huge amount of bandwidth available on optical fibers. If this is the case, the overprovisioning approach might turn out to be the best one, since it does not require any effort in telecommunication network engineering.

If, on the contrary, killer applications that may cause network collapse are expected to always exist, then resource utilization is very important and resource management and allocation techniques have to be implemented in network nodes to jointly provide QoS and obtain high resource utilization. This approach requires knowledge of user traffic statistics on the basis of known parameters (traffic specification), knowledge of the path or the network portion where the flow will be routed, and implementation of the following key algorithms for traffic control and resources management:

- *Call admission control* (CAC) algorithms implement rules for accepting or refusing a user call (flow) according to the declared traffic parameters and the availability of the resources needed to meet the requested QoS without disrupting the QoS provided to already accepted calls.
- *Shaping* algorithms are used at network edge to make the user traffic compliant with the traffic specifications.
- *Traffic verification* or *policing* algorithms are used at the access node to ensure that the user traffic is compliant with the parameters declared by the source in the traffic specifications.
- *Resource allocation* or reservation algorithms implement functions for allocating, in all the nodes crossed by the traffic flows, the bandwidth and storage resources needed to provide the requested QoS. This allocation must consider resource sharing as a fundamental issue to obtain high network utilization. Bandwidth allocation is usually provided by *scheduling* algorithms that implement functions in each node to transmit at a given time the most urgent or important packet of the available packets so as to satisfy the application needs. *Buffer management* policies are the most important storage allocation techniques.
- *QoS routing* algorithms implement rules for QoS based routing. Unlike the minimum distance or hop count routing strategies used in traditional networks, QoS routing may ease the task of network dimensioning and provide higher network utilization.

An alternative or additional approach may be to introduce *scaling mechanisms* in the multimedia application to dynamically modify the characteristics of the transmitted data stream with the purpose of adapting the workload generated to the resources available in the network. These mechanisms usually use feedback information on the network congestion status as, for example, in the TCP protocol. Of course, the cost of scaling is a decrease in the user perceived quality; moreover, it is not easy to achieve a match between the workload and the available network resources.

In the following sections, we will first outline the main functions of traffic control algorithms and of resources management algorithms. Then scaling techniques for workload adaptation will be introduced.

2.1. Call Admission Control

Call admission results from a negotiation between the user and the network. For this aim the multimedia traffic source is requested to declare both a set of parameters

which characterize its traffic at the network ingress and the QoS it requires.

The task of the CAC algorithm is to determine whether there are enough resources to meet the new call's QoS requirements without violating the QoS provided to already accepted calls. The CAC algorithms run in each node selected by the routing algorithm to support the new call, and they are therefore related to routing algorithms. If enough resources exist in all the nodes, the call is accepted and the data transfer can start; otherwise the call is dropped.

Several CAC algorithms have been proposed in the literature. We can identify two broad families of possible approaches. The first is based on determining an "equivalent bandwidth" [2], that is, a bandwidth required to satisfy the call QoS needs given the traffic characterization. Only if a bandwidth greater than or equal to the equivalent bandwidth is available over each link of the path, will the call be accepted; in this case the equivalent bandwidth is reserved for the call in each node and it is subtracted from the available bandwidth of each link. This approach is very simple and efficient once the equivalent bandwidth is computed. To this end several sophisticated techniques have been proposed. In general, they are based on simulative or analytic paradigms modeling network node behavior, and compute the loss probability and delay experienced when a new call is multiplexed over the node output link together with other calls. The resulting equivalent bandwidth is typically greater than the mean rate of the call, and the more stringent the QoS parameters the closer the bandwidth is to the peak rate. Unfortunately the equivalent bandwidth approach works only if the traffic generated by the user is similar to the traffic model used for equivalent bandwidth computation. Moreover, the complexity of the computation may reduce the practical applicability of this approach.

The second approach is based on network measurements [3] and therefore does not need to assume any specific model for the source or for aggregation of sources. The available bandwidth is measured over each link and a call is accepted if the available bandwidth is greater than or equal to, for example, the peak rate of the call. Once the call has been accepted, the new bandwidth availability for further new calls will be obtained by successive measurements.

In this case CAC algorithms are usually simple since they have to compare a rate that can be very simply derived from the call traffic parameters with the measured available bandwidth over the link. Unfortunately, it is not easy to determine an effective measurement of the available bandwidth. A possible solution to this problem comes from some queuing theory results that link the effective bandwidth to queuing performances; by using these results, it is possible to propose some estimates of the effective bandwidth.

2.2. Traffic Shaping and Policing

A positive result of the CAC procedure commits the multimedia source to guaranteeing that the traffic profile emitted conforms to the declared traffic parameters. In this perspective, the effectiveness of both the shaping functions in the multimedia application and the policing functions performed at network edge constitutes a challenge to meet the requested QoS.

The set of parameters used to characterize user traffic at the network ingress is specific to each network technology. However, the key parameters that must be provided, regardless of network technology, are the average bit rate for CBR traffic, and the average and peak rate (or burstiness) for VBR traffic. It may be useful, and is often required, to provide information relating to the burst duration, that is, the time for which a given source may send data at the peak rate.

Shaping is usually achieved by means of buffers designed not to exceed a given maximum delay at the transmission side. In the case of real time data, particularly voice and video, in order to avoid or decrease the delay introduced, a feedback on the encoding parameters can be used in conjunction with a virtual buffer [4] that, without introducing a delay, can be used to monitor the source emission and to force the encoding process to maintain the parameters declared by the source.

The declared traffic parameters are usually policed by the network at its access point and sometime at network boundaries when data traffic crosses edges between two different network providers. The average rate and the burst duration are usually policed by means of a token bucket device [5]; the peak rate is policed by means of a controller that monitors the interarrival time of the incoming packets. The "token bucket" is a token pool in which the tokens are generated at a constant rate equal to the average rate declared by the multimedia source. The bucket size represents the maximum capacity of the pool and is related to the declared burst duration. When a packet arrives, a number of tokens equal to the packet dimension in bytes is drawn from the pool; if a packet arrives when the pool is empty, it is marked as nonconforming to the traffic specification declared by the multimedia source, and may be dropped.

2.3. Scheduling

Scheduling algorithms are run in nodes to support different levels of priority or urgency criteria for data belonging to different calls. They span from the very simple FIFO (first-in first-out) mechanisms, where data are sent over each output link in the order in which they were received, to strict priority mechanisms, where lower-priority data are sent if and only if no higher-priority data exist in the node, up to sophisticated scheduling algorithms such as weighted round robin (WRR) or weighted fair queuing (WFQ), which are able to provide each call with bandwidth guarantees [6].

Several aspects must be taken into account when considering scheduling algorithms: complexity, the ability to separate flow behavior (a property often referred to as "isolation"), the ability to provide bandwidth guarantees, and buffer sharing capability. It must be noted that buffer management techniques, and in particular the queue architecture adopted within nodes, are strictly connected to the scheduling algorithm. For the sake of conciseness, in this article we focus on traditional output queuing (OQ) node architectures. In this architecture packets arriving

at input links are immediately transferred and stored in buffers at output links; they provide the best performances but require very high speed in the internal switching fabric.

In traditional data networks, nodes usually employ a FIFO [sometimes called first-come first-served (FCFS)] scheduling technique, with a single buffer for each output link shared evenly among all calls. Data are transmitted on output links in the temporal order in which they were received at input links. When the queue becomes full, packets are dropped, regardless of the flow to which they belong, until a position in the queue becomes available. The FIFO scheduling does not provide any form of isolation among flows; all the flows obtain the same QoS, which depends on the behavior of the other flows. Moreover, no bandwidth guarantees can be provided. However, buffers are shared among all flows, and thus fully utilized, and the algorithm is very simple. This scheduling technique may be suited for monomedia networks, where all the flows are interested in the same QoS parameters, but not for multimedia networks.

The key assumption required to provide isolation among flows is to create several queues at each output link and assign each flow to a queue; this is also referred to as *queue partitioning*. If this is the case, all flows belonging to the same queue share the same QoS, whereas isolation among flows assigned to different queues is quite simple to obtain.

In priority-based scheduling algorithms, each queue is associated with a different level of priority. Arriving packets are stored in the queue with the same level of priority as flow to which they belong. Queues are served in strictly increasing order of priority. Packets are extracted from the highest-priority queue until it is empty; and only if this queue is empty does the scheduling look at the next highest-priority queue and so on for all defined priorities. No preemption is enforced on packets being transmitted; thus a higher-priority packet must wait until the transmission of the ongoing lower priority packet has ended. This scheduling is fairly simple and provides isolation among flows belonging to different level of priority. Unfortunately, no isolation exists among flows with the same priority; lower-priority flows may be starved by higher-priority ones and therefore bandwidth is not guaranteed to all flows.

The simplest bandwidth guaranteeing scheduling is the WRR scheduling [7]. In WRR, each queue is assigned a weight, typically related to flow bandwidth needs.[1] The scheduling defines a service cycle; during this cycle, flows are served a number of times proportional to flow weights. Although this idea can provide bandwidth guarantees and flow isolation quite simply, it has several drawbacks. First, the service cycle length depends on the ratio between the flows' bandwidth requests. If we imagine for the sake of simplicity that flows send fixed-size data, the length of the cycle is the MCD of the flow weights. This number can become fairly large. Moreover, calls may be closed when the scheduler is in the middle of its service cycle. In that case a new cycle should be defined, but the part of the previous cycle not served must also be taken into account to meet flow bandwidth requirements correctly. The same holds when a new call is accepted. Several implementations of WRR-like schedulers, based on counters associated to each flow to face the problem of WRR service cycle, have been proposed in the literature [see, e.g., an article on deficit round robin (DRR) [8]].

More complex algorithms derive from the generalized processor sharing (GPS) scheduler [9]. The GPS scheduler emulates the behavior of an ideal fluid system, where each flow is continuously served at a rate proportional to its fair bandwidth share, determined on the basis of the number of active flows and on their required bandwidth needs. The emulation must take into account the constraint that in the real system only complete packets can be sent and flows cannot be served continuously; this give rise to packet GPS (PGPS) scheduling. PGPS schedulers can be implemented by assigning tags to packets when they arrive at input links (tags are roughly inversely proportional to flow rates) and serving packets in order of increasing tags.

Other implementations that approximate the behavior of the ideal fluid system described above have been proposed in the literature starting from the weighted fair queuing (WFQ) scheduling algorithm [10–14]. The most important properties of WFQ schedulers are flow isolation and bandwidth guarantees and, most importantly, if flows are leaky-bucket-controlled and all the nodes implement WFQ scheduling, then bounded end-to-end delay can be computed and guaranteed, as demonstrated in [9].

It must be observed that many other schedulers have been proposed (e.g., EDD, jitter EDD, stop and go; see [6] for a summary they exhibit interesting properties, although for the moment they have been not extensively implemented.

2.4. Buffer Management

Buffer management techniques have always been used, even in traditional network architecture; they become a key issue for multimedia networks, however. The most important buffer management techniques exploit *buffer occupancy* measures, *buffer allocation and partitioning*, and *dropping techniques*.

Buffer occupancy may be used as status information by CAC procedures to assess the network capability to accept calls and is used as a congestion indication by network nodes, providing users with explicit congestion signals.

Buffer allocation and partitioning is a key element in a scheduling technique, as described above, to protect flows from the behavior of other flows. Moreover, in order to provide different flows with bandwidth guarantees, some authors propose controlling only buffer allocation in nodes, since the bandwidth provided on each link is clearly proportional to the number of packets stored for the relative flow.

Packet dropping techniques are fundamental because nodes have to deal with finite buffer capacity. Some dropping techniques have the goal of dropping an entire packet; in networks where packet fragmentation often

[1] To provide bandwidth guarantees, it is necessary either for each flow to declare its bandwidth needs (these may be computed by CAC algorithms as seen before) or for all the flows to agree to share the bandwidth fairly.

occurs as a result of small-size data units, nodes may try to drop fragments belonging to the same packet rather than different packets, given that a single fragment loss renders the remaining portion of the packet useless for the user (data loss or retransmission may occur depending on the application) [15]. Other dropping techniques, based on random choice, try to reduce either correlation among packet losses or synchronization in packet losses among calls. Reducing correlation is useful, since most protocols that deal with packet losses use retransmission techniques that are negatively affected by correlated losses. Reduction of packet loss synchronization among calls can be achieved by spreading packet losses for different calls over time; this improves the performance of TCP, the most widely used transport protocol: since TCP senders reduce the data transmission rate when experiencing losses to prevent network congestion, if too many TCP flows experience losses at the same time, network utilization may drop to very low values even when this is not really necessary. These techniques are generally known as active queue management (AQM) techniques and have received a great deal of attention. As an example, random early detection (RED) [16] and random exponential marking (REM) [17] have been proposed for the Internet arena.

2.5. QoS Routing

Routing is also a fundamental task in multimedia networks. Routing in a topology implies choosing a "best" path (sequence of edges) to connect two nodes according to some defined metric associated with edges (links), and then distributing network information if the link metric dynamically changes with time. Two routing aspects are peculiar to multimedia networks: (1) dealing with QoS requirements when selecting the best path for each flow and (2) providing support for multicast traffic.

QoS routing implies selecting network routes with sufficient resources to satisfy the requested QoS parameters while achieving high resource utilization. The difficulty lies in the fact that QoS routing implies satisfying multiple constraints (e.g., bandwidth, delay, and loss requirements); it is well known that finding a feasible path with even only two independent path constraints is NP-hard. Moreover, multiple constraints impose multiple metrics for each link, thus increasing the difficulty in gathering up-to-date information on network status.

Three broad classes of algorithms proposed for QoS routing exist [18,19]: source routing algorithms, distributed routing, and hierarchical routing algorithms. They can be used together to obtain better performance depending on the network architecture. In *source routing*, each source node keeps information about the network global state (topology and state information on each link) and computes a feasible path locally; each emitted packet contains, therefore, routing information that forces the forwarding action in each node along the path. In *distributed routing*, the path is computed using a distributed algorithm; control messages are exchanged among nodes to create network state information and search for a feasible path. In *hierarchical routing*, nodes are clustered into groups in a multilevel hierarchy. Each node maintains an aggregated global state (partial information) that contains detailed information about nodes in the same group and aggregated information about other groups.

Multicast routing implies finding the best tree connecting a source node with a set of destination nodes [20]. Finding either a least-cost tree or the least-cost tree with bounded delay are both NP-hard problems. Several heuristics have been proposed to solve the multicast routing problem. It is, however, important to point out that QoS routing and QoS multicast routing are definitely the less developed of the techniques to control multimedia traffic described in this article.

2.6. Scaling Mechanisms

Scaling mechanisms have been used up to now to prevent occurrence of congestion in TCP/IP-based networks [21]. They are also used today to provide users with CBR real-time applications even when the compression algorithms used intrinsically lead to VBR streaming. This is the case, for example, of the MPEG video encoder currently used for video transmission in circuit switched network environment such as the ISDN.

Since the late 1990s the use of scaling mechanisms has been proposed for VBR real-time data transmission over packet-switched networks even when no stringent QoS guarantees are provided [22]. Scaling mechanisms can, in fact, be used to adapt the traffic profile emitted by the VBR source according to the variable-bandwidth profile available in a multimedia network.

In a real-time VBR video source, for example, this can be achieved either by changing the video coding parameters runtime or by using hierarchical coding schemes, or again by lowering the frame rate.

To scale the source traffic efficiently, a feedback control loop is introduced to monitor the network status; by so doing, when changes occur in the available network bandwidth, the appropriate actions are taken in the source to change its throughput accordingly. All the control mechanisms proposed use the delay and loss experienced in the end-to-end transmission to compute the available bandwidth, that is, the amount of data that can be transmitted without incurring in QoS degradation.

The protocols defined for this purpose have the target of calculating the bandwidth that TCP congestion control mechanisms make available to the TCP data source, in each network condition. They can be classified into three groups:

- *Window-based congestion control protocols*, which use a congestion window; therefore all the congestion control mechanisms implemented in the different TCP versions belong to this group [21].
- *TCP-like congestion control protocols*, which adapt the transmission rate according to the additive increase multiplicative decrease (AIMD) strategy without using congestion windows [23].
- *Equation-based congestion control protocols*, which adapt the transmission rate according to a suitable equation which, usually, derives from a TCP throughput model [24].

3. MULTIMEDIA NETWORKS: HISTORICAL PERSPECTIVE

Since the 1970s, the challenge of a universal network capable of providing the transport of information related to different media has stimulated the development of several network architectures each characterized by a specific information transfer strategy [25]. The oldest and most widely diffused network architecture is, of course, the telephone network; its technical approach is more than a century old and at the current stages of evolution it supports integrated transmission and switching technologies [Integrated Digital Network (IDN)] and integrated access modalities for services provision (ISDN or narrowband ISDN). This architecture, based on circuit switching, surely represents the best solution for the provision of the QoS required by any end-to-end narrowband communication among CBR applications; it does not, however, possess the flexibility needed to support VBR applications, such as data transfer among computers, since the circuit-switching approach provides only deterministic resource allocation.

In order to support VBR data communication without any performance guarantees (best-effort data communication), the IP datagram approach soon became the de facto standard in all the enterprise LANs, and later on, in LAN and WAN interconnection. This approach, in fact, is based on packet switching techniques that are highly suitable for VBR traffic, thanks to their ability to provide statistical resource sharing. Unfortunately, the connectionless data transfer supported by the IP approach is unable to satisfy the specific QoS needs of different applications.

The design of broadband multimedia network architectures started from these two network realities: a ubiquitous circuit-switched telephone network and a widely diffused packet-switched datagram-based internetwork. In the 1990s in particular the standardization environment that defined the ISDN proposed broadband ISDN and the related asynchronous transfer mode (ATM) technology to support multimedia traffic and its QoS requirements. This choice is based on the virtual circuit approach, which represents a compromise between statistical packet switching based on datagram transmission and deterministic circuit switching based on the synchronous transmission of small fixed-size data units. However, the B-ISDN dream crashed for two main reasons:

- The difficulties of constructing applications directly on top of ATM APIs that require the handling of very powerful, but not user-friendly, ATM control and management facilities
- The need to interoperate seamlessly with LANs such as Ethernet for the provision of efficient IP packet transmission over the ATM.

So, ATM never reached the desktop, and it was relegated to the backbone, whereas the growing success of Ethernet from shared media to switching paradigm lead to *overprovisioning* as the solution for QoS provision in the local and corporate environment.

Most companies today seem to prefer high-speed switched LANs instead of an ATM infrastructure, and IP is clearly the most widely accepted network paradigm today, although the network does not guarantee the correct delivery of packets and may be subject to unpredictable delays. The key question today is therefore how to evolve IP networks toward a multimedia network providing worldwide support for multimedia traffic. This target requires that control and management functions are correctly designed and injected in IP technology.

4. NEXT-GENERATION INTERNET: A CHALLENGING APPROACH FOR MULTIMEDIA NETWORKING

The IETF (Internet Engineering Task Force) has defined two alternative frameworks to provide QoS in a datagram network matching the flexibility required by multimedia networking [26,27]. The first one, referred to as *Internet Integrated Service architecture*, can be considered as a relevant upgrade of the control plane of the classical IP network. New signaling protocols, such as Resource ReserVation Protocol (RSVP), has been introduced to distribute information about resources to be allocated within the network. Unlike ATM or ISDN networks where the signaling protocols produce "hard states", RSVP packets produce "soft state" within the devices (QoS-aware IP router), that is, proper configuration of resources that are automatically released if the "soft state" is not refreshed periodically.

Furthermore, the RSVP signaling paradigm was conceived from the beginning to take into account a multicast environment. For this reason, in fact, in RSVP it is the destination of the multimedia flows that claims for the resources to be allocated to the flow coming from the source; the reservation messages that flow from the destination to the source allow the nodes of a multicast tree to merge resources reservations related to the same multipoint session.

It is very relevant to notice that RSVP is just a signaling protocol able to transfer a request for proper resource allocation among nodes (including the end systems). To be effective, an IntServ network has to implement the functional components introduced in the previous sections in each of its routers. IntServ defines three service classes: the best-effort class, the controlled load service class, and the guaranteed service class.

The first is the present behavior of the Internet (no guarantees), whereas the last one corresponds to detailed (hard) guarantees for bandwidth, delay, and loss. The second class has not been defined in a rigorous manner but is related to behavior that is strictly better than best-effort. It corresponds to the best-effort performances that could be obtained on the network if it were lightly loaded (although these controlled load flows are passing through a network that is highly loaded). It is more a relative behavior that a service class, that is defined on the basis of specific performance parameters (as for the guaranteed service class).

Let us stress here that QoS is provided on a per flow basis in the IntServ architecture, that is

- Every flow is identifiable by means of a specific flow identifier or by other classifications such as the tu-pla consisting of IP-source, IP-dest, TCP-Port-num-source, TCP-Port-num-dest.

- Each end-to-end session in either a unicast or a multicast environment will be composed by several unidirectional flows that will be managed by the networks nodes with per flow queuing, per flow scheduling, shaping, discarding, and so on.

The traffic specification that makes it possible to accept or refuse each flow is very simple and based on a linear upper bound termed the linearly bounded arrival process (LBAP).

Per flow signaling and queuing, of course, produce an impressive increase in the complexity of the network due not only to the RSVP signaling burden on the network but in particular the processing power that is necessary in the router to manage the signaling components and the scheduling schemes. A backbone Internet router, at the time of writing, is loaded by more than 100,000 flows, and although not all of them will need resource reservations, the state and processing requirements within the nodes are prohibitive for scalability reasons.

Because of its scalability problem, IntServ could be considered a suitable architecture for QoS provisioning in IP networks only when there are no more than a 1000 reserved flows. This may be the case of a corporate or campus network, but within the backbone a new approach is needed.

The Internet Engineering Task Force faced the scalability problem of IntServ by defining the *Internet Differentiated Services* architecture, DiffServ for short. The basic idea of Diffserv is to consider an aggregate of flows (called *macroflows*) instead of single flows in an IP QoS domain. The complexity of the nodes is reduced since there is no need to manage the single flows, which lose their identity in an aggregation of many flows with similar requirements. Of course, this is questionable if the different flows have different requirements. Furthermore, the DiffServ architecture defines a core/edge approach were the most complicated activities are carried out by a border router (at the ingress point of the domain) while the core routers operate on the packets in a simpler way. At the ingress of the network the flows are classified as belonging to a certain macroflow by marking them with a marker (code) that is written in the header of the IP packet. To this end, the *type of service* field within the IPv4 header, renamed *differentiated services code point* (DSCP), is used. This field produces a different *per hop behavior* (PHB) in the router. The IETF defined three different possible PHBs: best effort, assured forwarding PHB, and expedited forwarding PHB. Again, the first one is the behavior of the present Internet architecture while the last corresponds to real-time traffic that requires hard guarantees on delay, jitter, loss, and bandwidth.

Assured forwarding is an intermediate class that is further divided into subclasses with different priorities and discarding privileges. Different network domains could associate their specific PHB with the same macroflow, and the relations between different domains have to be negotiated through a proper service-level agreement (SLA) which will initially be almost static and in the future could be set up in a dynamic fashion using proper observation of the state of each domain. For this purpose a possible approach could be based on the adoption of centralized devices called *bandwidth brokers*.

The main limitations of DiffServ are related to the uncertain provisioning of the end-to-end QoS. Associated with a macroflow, in fact, it is not certain whether a single flow will receive the correct performances, particularly if several domains are passed. DiffServ is certainly an oversimplification of the IntServ approach and tries as far as possible to avoid the requirement for signaling from end systems. However, a signaling scheme will definitely be necessary between the nodes and the bandwidth broker.

Whatever approach is used, multimedia will lead to significant problems with respect to the traffic forwarding on the network. *Forwarding* is the process of moving one packet from a certain input to a certain output, and this is done taking into account the information that is passed by the routing process. It is relevant to mention here that the classical routing process is carried out on the Internet via a "destination-based approach." The forwarding process, therefore, is quite complicated if a classless routing scheme is adopted and a "longest matching"[2] lookup has to be carried out for every packet. For this reason a fixed-size approach would be better. This was one of the reasons for the deployment of multiprotocol label switching (MPSL). At the entrance of the network a label is associated with each packet by a classification procedure. Similar to the label swapping techniques adopted in ATM and frame relay networks, this forwarding identification label simplifies the forwarding of variable-size packets to their destination. Furthermore, the path that a packet has to follow on the network can be established by the egress node (i.e., the node at the entrance of the network) and a label distribution procedure can update the forwarding tables, which can now be based on labels instead of the destination address field. In this way it is easy to provide load balancing between the different routes or build up recovery procedure for the traffic that was critically stopped by a fault.

MPLS can distribute and allocate these labels using different approaches, namely, Label Distribution Protocol (LDP), Constraint-based Routing using Label Distribution Protocol (CR-LDP), or RSVP. Again RSVP is used as a signaling protocols extending its functions to provide traffic control (RSVP–traffic Engineering).

The MPLS labels can be stacked (push operation) to funnel some traffic over specific paths and then bring back the identity of each macroflow at the egress nodes (pop operation). This helps in building up virtual private networks (VPNs) at the layer 3 level (by using MPLS it is also possible to build up level 2 VPNs).

It is relevant to mention that MPLS is a network control plane that is unable to provide QoS by itself. The constraint-based routing schemes (such as CSPF) adopted by MPLS can lead to a proper path on the network that takes into account the specific requirements of certain end-to-end sessions, but again it is scheduling, active queue management and shaping algorithms that are really

[2] Longest matching lookup aims to find the network address which best matches the destination written in the header of a packet, in a routing table.

responsible for the different forwarding behaviors offered by a node from its input to its output.

Because of its marking on entry approach, MPLS couples well with a DiffServ architecture, and this is, at the time of writing, the most promising architecture for a flexible, scalable, controllable, and manageable multimedia network.

BIOGRAPHIES

Andrea Bianco is Associate Professor at the Dipartimento di Elettronica of Politecnico di Torino, in Italy. He was born in Torino (Turin), Italy, in 1962. He holds a Dr. Ing. degree in Electronics Engineering (1986) and a Ph.D. in Telecommunications Engineering (1993), both from Politecnico di Torino. From 1987 to 1990 he was with S.S.B. in Torino, where he has been working on office automation projects based on database and distributed networking programming. Since 1994 he was an Assistant Professor at Politecnico di Torino, first in the Dipartimento di Sistemi di Produzione, and later in the Dipartimento di Elettronica. In 1993 he visited Hewlett-Pakard Labs in Palo Alto, California. In the summer 1998 he visited the Electronics Department at Stanford University, California. He has co-authored over 80 papers published in international journals and presented in leading international conferences in the area of telecommunication networks. His current research interests are in the fields of protocols for all-optical networks and switch architectures for high-speed networks. A. Bianco is a member of IEEE.

Stefano Giordano received the Laurea degree "cum laude" in Electronics Engineering from the University of Pisa in 1990 and the Ph.D. degree in Information Engineering in 1994. During 1988/89 he worked with CNR-CNUCE. Since year 1991 he was with the University of Pisa participating and coordinating several research activities sponsored (among others) by Saritel, Siemens, Italtel, CNR, RAI, HP, ASI, TILab, Marconi, and Finsiel. Since 2001 he has been an Associate Professor at the Department of Information Engineering of the University of Pisa, where he give lectures on telecommunication networks and design and simulation of telecommunication networks. His research and professional areas of interest are broadband multimedia communications and telecommunication networks analysis and design. He is responsible for the NetGroup at Consorzio Pisa Ricerche and the TLC NetLab of the Faculty of Engineering in Pisa. Stefano Giodano is participating to the Technical Committee of the Campus Network of the University of Pisa (SERRA), has been a member of the IEEE Communication Society since 1989, of the Internet Society since its foundation, and of the IFIP Working Group 6.3. He was referee of the projects of the European Union, the National Science Foundation, and the Ministry of Research in Italy in the area of telecommunications.

Alfio Lombardo received his degree in electrical engineering from the University of Catania, Italy, in 1983. Until 1987, he acted as consultant at CREI, the center of the Politecnico di Milano for research on computer networks, where he was involved in European research projects on protocol design (SEDOS and CTS-WAN projects). There he was the Technical Coordinator of the Formal Description Techniques (FDT) COST 11 TER project from 1986 to 1988. In 1988 he joined the University of Catania, where he is Full Professor of Telematics. There he was the leader of the University of Catania team in the European ACTS project DOLMEN (Service Machine Development for an Open Long-term Mobile and Fixed Network Environment). Presently, he is involved in the European IST Project VESPER (Virtual Home Environment for Service Personalization and Roaming Users) as leader of the University of Catania team. His research interests include distributed multimedia applications, multimedia traffic modeling and analysis, Internet2, and wireless networks. His email address is *lombardo@iit.unict.it*.

BIBLIOGRAPHY

1. L. C. Wolf, C. Griwodz, and R. Steinmets, Multimedia communication, *Proc. IEEE* **85**(12): (Dec. 1997).
2. R. Guerin, H. Ahmadi, and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE J. Select. Areas Commun.* **9**(7): (Sept. 1991).
3. S. Jamin, P. Danzig, S. Shenker, and L. Zhang, A measurement-based admission control algorithm for integrated service packet networks, *IEEE/ACM Trans. Network.* **5**(1): (Feb. 1997).
4. A. Lombardo, F. Cocimano, A. Cernuto, and G. Schembra, A queueing system model for the design of feedback laws in rate-controlled MPEG video encoders, *Trans. Circuits Syst. Video Technol.* (in press).
5. C. Partridge, *Gigabit Networking*, Addison-Wesley, Reading, MA, 1994.
6. H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, *IEEE Proc.* **83**(10): (Oct. 1995).
7. M. Katevenis, S. Sidiropoulos, and C. Courcoubetis, Weighted round-robin cell multiplexing in a general-purpose ATM switch chip, *IEEE J. Select. Areas Commun.* **9**(8): (Oct. 1991).
8. M. Shreedhar and G. Varghese, Efficient fair queuing using deficit round-robin, *IEEE/ACM Trans. Network.* **4**(3): (June 1996).
9. A. K. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The single-node case, *IEEE/ACM Trans. Network.* **1**(3): (June 1993).
10. A. Demers, S. Keshav, and S. Shenkar, Analysis and simulation of a fair queueing algorithm, *Internet Res. Exp.* **1**: (1990).
11. D. Varma and D. Stiliadis, Hardware implementation of fair queueing algorithms for asynchronous transfer mode networks, *IEEE Commun. Mag.* (Nov. 1997).
12. L. Zhang, Virtual clock: A new traffic control algorithm for packet switching networks, *ACM SIGCOMM'90*, Philadelphia, Sept. 1990.
13. J. C. Bennet and H. Zhang, WF^2Q: Worst-case fair weighted fair queueing, *INFOCOM'96*, March 1996.
14. J. C. Bennet and H. Zhang, Hierarchical packet fair queueing algorithms, *IEEE/ACM Trans. Network.* **5**(5): (Oct. 1997).

15. A. Romanow and S. Floyd, Dynamics of TCP traffic over ATM networks, *IEEE J. Select. Areas Commun.* **13**(4): (May 1995).

16. S. Floyd and Van Jacobson, Random early detection gateways for congestion avoidance, *IEEE/ACM Trans. Network.* (Aug. 1993).

17. S. Athuraliya, S. Low, V. Li, and Y. Qinghe, REM: Active queue management, *IEEE Network* **15**(3): (May–June 2001).

18. H. Chen and K. Nahrstedt, An overview of quality-of-service routing for the next generation high-speed networks: Problems and solutions, *IEEE Network Mag.* (*Special Issue on Transmission and Distribution of Digital Video*) **12**(6): (Nov.–Dec. 1998).

19. E. Crawley, R. Nair, B. Rajagopalau, and H. Sandick, A Framework for QoS-based routing in the Internet, Internet RFC 2386 (April 1998).

20. C. Diot, W. Dabbous, and J. Crowcroft, Multipoint communication: A survey of protocols, functions, and mechanisms, *IEEE J. Select. Areas Commun.* (April 1997).

21. R. Stevens, *TCP/IP Illustrated*, Addison-Wesley, Reading, MA, 1994.

22. S. Floyd and K. Frdl, (Promoting the use of end-to-end congestion control in the Internet), unpublished, Feb. 1998; *http://www-nrg.ee.lbl.gov/floyd/papers.htmlJend2end-paper.html*.

23. R. Rejaie, M. Handley, and D. Estrin, RAP: An end-to-end rate-based congestion control mechanism for realtime streams in the Internet, *Proc. INFOCOM'99*, New York, (March 1999).

24. S. Floyd, M. Handley, J. Padhye, and J. Widmer, *TCP-Friendly Rate Control* (*TRFC*)*: Protocol Specification*, IETF, (July 2001).

25. M. Decina and V. Trecordi, Convergence of telecommunications and computing to networking models for integrated services and applications, *Proc. IEEE* **85**(12): (Dec. 1997).

26. Z. Wang, *Internet QoS Architectures and Mechanism for Quality of Service*, Morgan Kaufmann, 2001.

27. B. Teitelbaum, Internet2 Qbone: Building a testbed for differentiated services, *IEEE Network* **13**(5): 8–16 (Sept.–Oct. 1999).

MULTIMEDIA OVER DIGITAL SUBSCRIBER LINES

HAITAO ZHENG
Bell Laboratories
Lucent Technologies
Holmdel, New Jersey

K. J. RAY LIU
University of Maryland
College Park, Maryland

1. INTRODUCTION

Multimedia communications has become one of the fastest-growing and yet most challenging fields for both academia and industry. Internet-enabled applications such as videoconferencing, multimedia mail, video-on-demand, HDTV broadcast — either by wirelines such as asymmetric digital subscriber lines (ADSLs), integrated services digital networks (ISDNs), or by wireless networks — present new problems of distinctive nature. The high volume of multimedia data can be handled efficiently only if all system resources are carefully optimized. The distinctive nature of multimedia data also requires existing transmission systems to be augmented with functions that can handle more than ordinary data.

Many advances in multimedia communications are made through interaction and collaboration between multimedia source coding and channel optimization. In this article, we present an approach to providing reliable and resource efficient multimedia services through ADSL by jointly considering compression/coding and channel optimization techniques.

We begin this article with a brief review of multimedia communications in general, focusing on multimedia compression/coding and joint source channel optimization. We then describe the channel characteristics and modulation procedure for ADSL. We then examine two transmission architectures for delivering multimedia content over ADSL. The important concept of resource allocation is discussed. The remainder of the article deals with technical details on underlying techniques for these two architectures to perform source/channel optimization and resource allocation. The performance is examined in some practical applications.

1.1. Multimedia Compression and Layered Coding

Multimedia communications benefits greatly from developments in source compression algorithms and hardware implementations. The objective of compression is to reduce the amount of data necessary to reproduce the original data while maintaining a desired level of signal quality and implementation complexity. Compression is necessary and important to reduce the bit rate for efficient transmission over normally bandwidth-limited networks. For digital data that cannot afford to lose any information, the compression schemes used are mainly lossless; that is, the reverse procedure can reproduce the exact original signal. Multimedia data, however, are subject to human perception. For image, video, speech, and audio, loss of some fidelity is tolerable as long as it is not perceivable. Therefore, compression may discard some information in order to achieve more compactness, which corresponds to the well-known lossy scheme.

In the past, the design of multimedia source coding was based mostly on the assumption of error free channels. The objective of compression was to reduce the information rate to the maximum extend without huge quality degradation. It was not recognized that compression, however, renders the compressed data highly vulnerable to channel errors and losses. A few bit errors could lead to severely disrupted media. Researchers have now realized the importance of joint consideration of the distortion induced by channel errors into source coding design. Among all the techniques, layered and scalable coding has been widely recognized and utilized since it can provide error resilience necessary for noisy channel transmissions. One major role of layered coding is to classify media signals in terms of importance and separate them into different layers.

The importance is often defined perceptually. The layers are compressed by different coding schemes and protected by different priority levels. The theme is to assign the highest priority to the most important layer. As such, layered coding achieves error resilience by preventing the loss of perceptually important information. Another advantage of layered coding is scalability. Data rate has a direct impact on transmission cost and media quality. If required, layered coding can alter the data rate to compromise any change in transmission cost and media quality requirements.

Some popular forms of layered coding are subband/wavelet coding and scalability options in H.263+, H.263++, and MPEG-4 [1–5]. It is worth pointing out that integrated services also display a level of scalability. Internet applications are often formulated as a mixture of data, speech and video services, each associated with different data rates and QoS (quality of service) requirements. They can be viewed as a set of layers of different importance. MPEG-4 also specifies audiovideo objects (AVOs) to represent integrated services.

1.2. Joint Source and Channel Optimization

Channel transmission inevitably introduces errors and losses. Multimedia communications systems, including both source and channel coding, have to be robust against channel errors so that media application will not be seriously disrupted by channel errors. In other words, multimedia communications requires efficient and powerful error control techniques.

In general, there are two approaches for error control and recovery. The first approach involves channel transmission design to reduce occurrence of channel errors. Powerful channel coding and Automatic Retransmission reQuest (ARQ) are two commonly used methods. However, the amount of channel coding is limited by bandwidth requirements, and the number of retransmissions is limited by delay requirements. Therefore, the channel transmission technique cannot fully remove channel errors. Given this fact, the second approach appends redundancy in compressed data so that channel error effects can be concealed and even become imperceptible. This is the so-called error concealment and recovery [6]. The design of a concealment strategy depends on the whole system design. More redundancy at source compression results in better concealment. However, it also implies increased bandwidth requirements on channel transmission. For a given bandwidth, increasing the source rate results in reduced channel coding gain and vice versa. As such, the optimal solution, namely, a joint source and channel optimization (JSCO) approach, would be to jointly optimize source and channel coding, and balance the amount of robustness between these two functions.

Both literature and practice have shown that joint optimization leads to substantial gain in performance [7]. For example, layered coding produces layers or classes with different error sensitivities. Unequal error protection, which assigns different loss rates to the layers, yields better media quality compared to a single-layer approach [8]. The joint optimization has two aspects: media quality and resource consumption. The ordinary source coding design only measures media quality, whereas the traditional channel optimization only takes into account resource consumption. The joint optimization aims to "minimize resource consumption while maintaining the desired media quality" or "achieve the best media quality for a set of preassigned resources." Optimization involves selecting source coding parameters (source coding rate, scalability format, etc.) and channel transmission parameters (such as transmit power, channel coding, and modulation). In the following discussion, we assume that the two end systems are connected by a direct link, namely, a circuit switch system. For a packet-switched system, the packetization strategy should also be considered in the optimization. Additional details on packet-based multimedia communications can be found in the literature [9,10].

2. FUNDAMENTALS OF ADSL

The exponential growth of Internet traffic has driven the demand for additional bandwidth and propelled the development of many new mechanisms of transmitting information. One of the most efficient mechanisms is digital subscriber lines (DSLs). DSLs can deliver megabit connectivity to the mass households using traditional phone lines. Mostly favored by Internet users, ADSL [11], asymmetric DSL, is specifically designed to support asymmetric data traffics to exploit the one-way nature of most multimedia applications where large amount of information flows toward the subscribers and only a small amount of interactive control information is transmitted in the upstream direction.

It is necessary to first understand the characteristics of the ADSL channel. Major channel impairments are attenuation and crossstack, which tend to increase as a function of frequency and distance. The resulting channel is spectrally shaped, as shown in Fig. 1. The wide variation in frequency leads to considerable difficulty and complexity for channel equalization, which is necessary for any single-carrier system. To avoid this problem, the ADSL

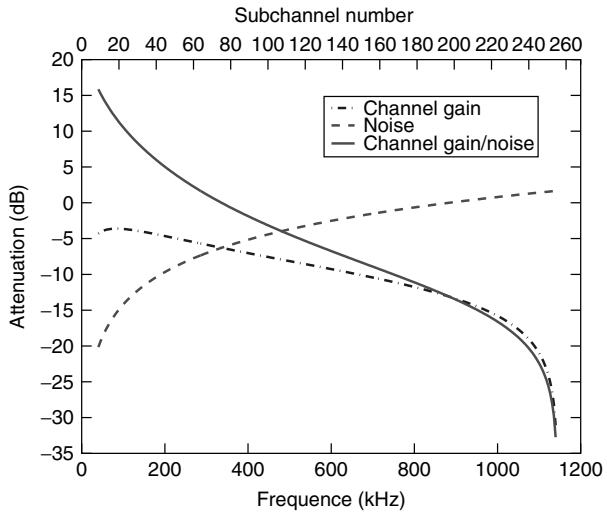


Figure 1. Typical spectrally shaped ADSL channels. The ADSL frequency band spans from 40 to 1397 kHz, with 256 subchannels of 4.3125 kHz each.

community adopted multicarrier modulation (MCM) [12] as the standard channel coding scheme. MCM partitions the ADSL frequency band into a set of independent subchannels, each corresponding to a smaller frequency band. When the number of subchannels is large enough, the subchannels are sufficiently narrow, and can be approximated as independent AWGN channels requiring little or no equalization. Mathematically, each subchannel can be described by

$$R_k = H_k S_k + N_k \tag{1}$$

where S_k and R_k represent the transmitted and the received signal at the kth subchannel and H_k and N_k represent the subchannel (complex-valued) gain and noise, respectively, approximated by the corresponding value at the center frequency of the kth subchannel.

One remarkable merit of MCM is that it allows the data rate, transmit power, and channel coding scheme at each subchannel to change independently. This flexibility provides optimality and fast adaptation to channel variations. As such, the optimal use of the entire channel can be achieved by making optimum use of each subchannel. Associated with subchannels are two transmission parameters: transmit power and data rate. Theoretically, waterfilling can achieve the ultimate system capacity, which allocates different transmit power levels to subchannels with different channel gains. The higher the channel gain, the larger the amount of transmit power. Accordingly, those subchannels with higher power level can transmit at a higher data rate to maximize the overall data rate. There has been extensive study on the allocation of power and data rate to the subchannels, known as the loading algorithm [13–17]. We will review these algorithms in the following sections. More information about ADSL can be found in the book by Starr et al. [18].

3. ARCHITECTURE DESIGN

A primary approach to resource efficiency in multimedia communications is to provide different priorities to layers of different perceptual importance. The architecture design for delivering multimedia over ADSL follows this approach. Although ADSL has the distinct characteristics of a multichannel structure, a rather simple solution — serial transmission — is to view the ADSL channel as a single transmission pipe and design source coding independent of multichannel optimization. Another solution — parallel transmission — combines ADSL channel partitioning and optimization into the system design, which was first proposed for image transmissions in 1998 [19] and extended to video transmissions in 1999 [20].

3.1. Serial Transmission

This approach ignores the multichannel structure within the source coding design. Therefore, joint source channel optimization involves a single transmission pipe and a source coder. Different priority levels can be achieved by transmitting source layers separately in time, and assigning different amount of channel resources. Precisely, the transmission is time-slotted based on where in each time slot only data from one source layer can be transmitted; that is, source layers are time-multiplexed. The optimization allocates time slots to source layers, and within each time slot, distributes channel resources among subchannels. Figure 2a depicts an example of serial transmission with three source layers. Layer 1 is transmitted in time slots 1 and 2, while layers 2 and 3 are transmitted in slots 3 and 4, respectively. Within time slots that a single source layer is transmitted, all the usable subchannels share the same error performance, thus the same priority. However, error performances across time slots in which different source layers are transmitted are completely different. This conclusion is reflected in Fig. 3a, where the error performance is represented by the bit error rate (BER) [20]. It also implies that the system has to optimize resource allocation for each layer. This requirement results in not only additional computational complexity but also extra difficulty for transmitter/receiver implementation due to frequent channel parameter changes. A detailed resource allocation scheme will be discussed in the following section.

3.2. Parallel Transmission

In general, ADSL channels vary slowly in time and may be considered as static. The channel gain and noise variations in the frequency domain are more severe compared to that in time domain. Serial transmission eliminates frequency variations through resource allocation. This variation, however, can be utilized to provide different error priorities to source layers. Such consideration leads to the invention of parallel transmission, which transmits source layers simultaneously in time but at different frequencies. As shown in Fig. 2b, each source layer occupies a certain number of subchannels, corresponding to frequency multiplexing [20]. The optimization requires a subchannel-to-layer assignment, which assigns subchannels to transmit each source layer. It is beneficial to assign subchannels with better channel gain and noise performance to transmit important layers. Such consideration could guarantee reliable base layer transmission with little power consumption — an important advantage of parallel transmission, especially under a low power constraint. Parallel transmission can also integrate traffic flows of various QoS requirements without frequently changing channel settings. Compared to serial transmission, parallel transmission has a completely different error performance distribution. It is observed from Fig. 3b that the error performance varies across a group of subchannels and remains static in time. The technical details of resource allocation will be discussed in next section.

4. SYSTEM OPTIMIZATION

Having described the architecture design for multimedia over ADSL, we now discuss the joint optimization procedure. Depending on the system goal, the optimization problem can be formulated in the two aspects discussed in Sections 4.1 and 4.2.

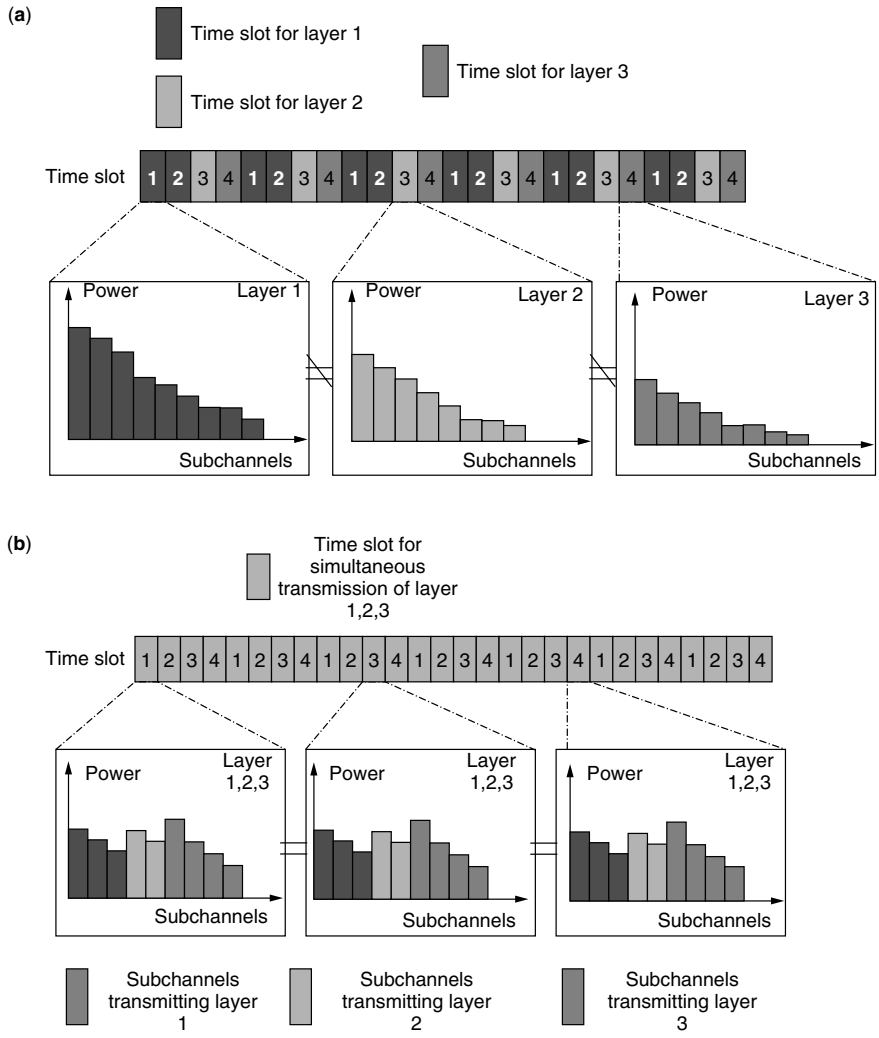


Figure 2. Example architecture for (**a**) serial transmission and (**b**) parallel transmission.

4.1. Optimization 1: Minimizing Cost

The viability of multimedia services business depends on a solution to provide desirable services at a low cost. Cost, within a communication system, is determined mainly by system resource consumption, such as bandwidth, transmit power, and hardware complexity. In this study, we assume fixed bandwidth and interpret power consumption as a major indicator of resource consumption. We intend to find a simple resource allocation scheme due to hardware complexity limitations. The optimization emphasizes subchannel resource allocation and time/subchannel to layer assignment, and intends to minimize transmit power and satisfy QoS requirements. We assume that QoS requirements are represented by the data rate R and the error rate BER.

The optimization consists of two loops. The first loop involves subchannel resource allocation, which is performed by a loading algorithm. On the basis of the QoS requirement, the system allocates transmit power and data rate to subchannels to minimize the overall power consumption. Most existing loading algorithms aim at achieving the same error performance on all the usable subchannels. Given the BER and R, theoretically, the amount of power allocated to subchannel k is proportional to $\Gamma(\text{BER})/g_k$, where g_k represents the channel gain to noise ratio (CGNR) at subchannel k [21]. $\Gamma(\ )$, a function of BER, measures the SNR distance from Shannon capacity. For uncoded QAM, a BER of 10^{-6} corresponds to a Γ of 8.8 dB; while at zero BER, $\Gamma = 10$ dB [18]. Once the amount of transmit power is derived, the corresponding rate at subchannel k can be computed and a modulation is selected accordingly. After loading, the sum of the subchannel transmit power is

$$E(R, \text{BER}, C) = \sum_{k=1}^{C} \frac{\Gamma(\text{BER})}{g_k}(2^{b_k} - 1) \tag{2}$$

where C represents the number of subchannels that are transmitting during this particular layer's transmission; b_k represents the bit rate at subchannel k where $R = \sum_{k=1}^{C_m} b_k$. More details about the loading algorithms can be found in the literature [13–16].

For a given time/subchannel to layer assignment, this inner loop optimization finds the optimal subchannel power and bit rate distribution for each layer. For each source layer, serial transmission performs power and bit

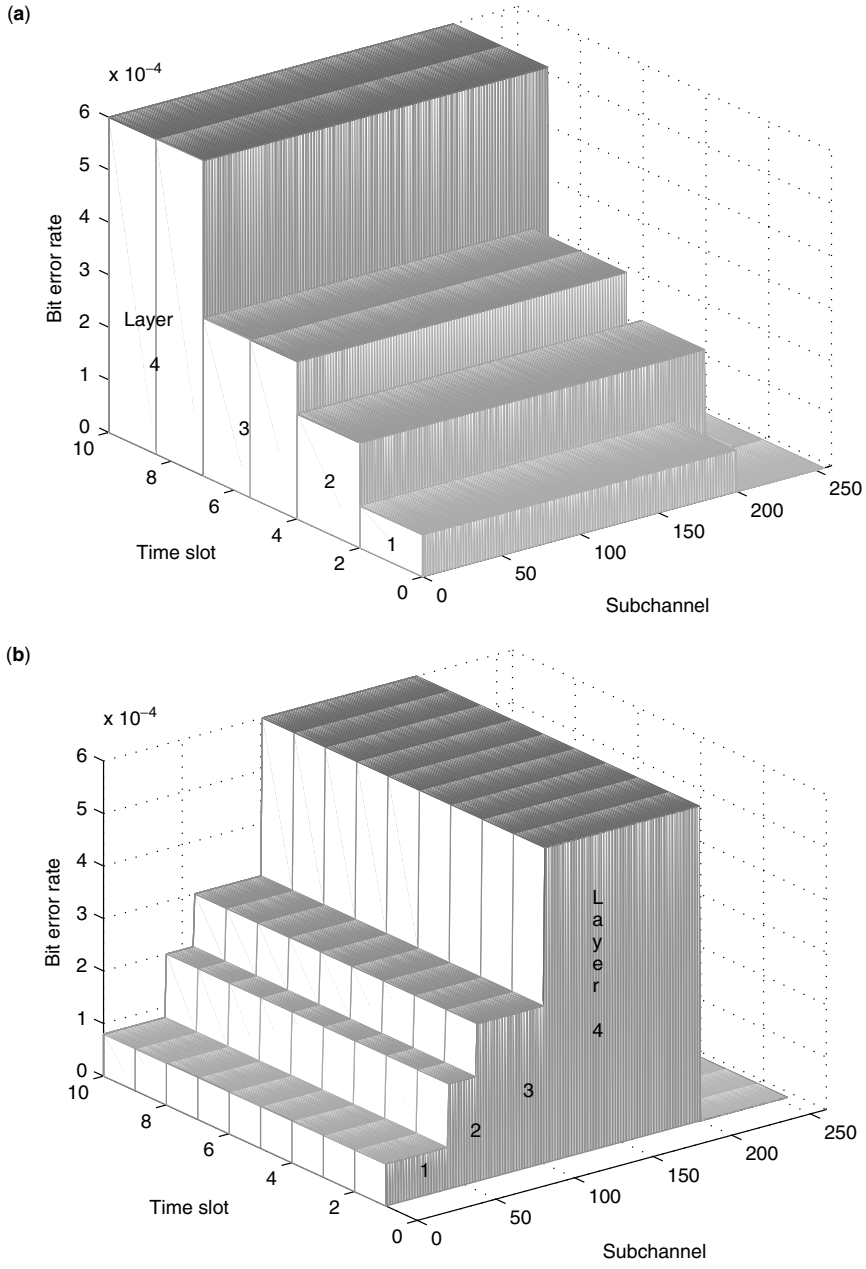


Figure 3. Error performance across the time slots and subchannels for (**a**) serial transmission and (**b**) parallel transmission.

loading using all the subchannels, while parallel transmission performs loading using a group of subchannels. The complexity depends on the number of subchannels and the number of source layers. The outer loop optimization finds the optimal time/subchannel to layer assignment.

- *Serial Transmission: Time-Slot Assignment*. A set of T time slots are grouped together into a time frame, where slot assignment is repeated frame by frame. The optimization can be represented mathematically as follows:

$$\text{Given} \quad \{R_m, \text{BER}_m\}_{m=1}^{N}$$

$$\text{Find} \quad \{T_m\}_{m=1}^{N} \quad \text{where} \quad \sum_{m=1}^{N} T_m = T \quad \text{to} \quad (3)$$

$$\text{Minimize} \quad E_T = \sum_{m=1}^{N} \frac{T_m}{T} E\left(\frac{T}{T_m} R_m, \text{BER}_m, C\right)$$

where T_m represents the number of time slots within a timeframe for layer m, R_m is the throughput requirement of layer m, C represents the number of subchannels, and E_T represents the power constraint. The optimal slot assignment that minimizes the overall transmit power E_T can be solved by an exhaustive search. The complexity depends on the number of slots per frame. A large number of slots results in fine granularity and better performance at the cost of higher complexity. The best solution is a compromise between quality and complexity.

- *Parallel Transmission: Subchannel-to-Layer Assignment.* Obviously, subchannels with higher CGNR

should transmit layers of higher importance. By sorting the subchannels in a decreasing CGNR order, the problem of subchannel-to-layer assignment can be reduced to finding the optimal number of subchannels for each source layer, $\{C_m\}_{m=1}^N$. After sorting, subchannels indexed 1 to C_1 are used to transmit layer 1 while subchannels indexed C_1+1 to C_1+C_2 are for layer 2, and so on. The loading algorithm derives the optimal power and rate distribution for each group of subchannels. Mathematically, the problem is equivalent to

$$\begin{aligned}
\text{Given} \quad & \{R_m, \text{BER}_m\}_{m=1}^N \\
\text{Find} \quad & \{C_m\}_{m=1}^N \quad \text{where} \quad \sum_{m=1}^N C_m \le C \quad \text{to} \\
\text{Minimize} \quad & E_T = \sum_{m=1}^N E(R_{m,T}, \text{BER}_m, C_m) \\
& = \sum_{m=1}^N \sum_{k=C_{m-1}+1}^{C_m} \frac{\Gamma(\text{BER}_m)}{g_k}(2^{b_k}-1) \\
\text{where} \quad & \sum_{k=C_{m-1}+1}^{C_m} b_k = R_m
\end{aligned} \tag{4}$$

Similarly, an exhaustive search can certainly lead to the optimal solution. The complexity depends on the number of source layers and more importantly, the number of subchannels. For ADSL, the number of subchannels is normally more than 256, which implies huge complexity. Using an efficiency measure, a successive search algorithm can quickly approach the optimal solution without examining all the subchannel-layer combinations [22].

4.2. Optimization 2: Quality Optimization

Sometimes, service providers aim to deliver the best service at a fixed cost budget. For most media applications, quality is measured by the distortion between the original and reconstructed data. Mathematically, the distortion can be approximated by the sum of source coding induced distortion D_s and channel transmission induced distortion D_c

$$D = D_s(R_s) + D_c = D_s(R_s) + \sum_{m=1}^N Pe_m W_m \tag{5}$$

where W_m represents the average distortion caused by a single bit error at layer m and Pe_m represents the BER for layer m. The source coding scheme determines the values of R_s, D_s, and $\{W_m\}_{m=1}^N$. When these are given, the goal of minimizing distortion is equivalent to finding the best BER distribution, $\{Pe_m\}_{m=1}^N$ for a given amount of power usage. Thus, BER in this problem, becomes an optimization parameter. Using the loading algorithm defined by Fisher and Huber [17], BER for layer m after loading can be computed as

$$Pe_m(R_m, E_{m,T}, C_m) \approx 4Q\left(\sqrt{\frac{3E_{m,T}G_m/C_m}{(2^{R_m/C_m}-1)}}\right) \tag{6}$$

where

$$G_m = \frac{C_m}{\sum_{i=C_{m-1}+1}^{C_m} 2^{(b_i-R_m)}/g_i} \tag{7}$$

represents the rate averaged CGNR for layer m, and

$$E_{m,T} = \sum_{i=C_{m-1}+1}^{C_m} E_i \tag{8}$$

represents the total power consumption of the subchannels assigned to layer m. We refer $E_{m,T}$ as layer power. $Q(x)$, the Q function, is defined by

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt \tag{9}$$

For a given R_s, the optimization problem focuses on D_c only

$$\begin{aligned}
\text{Given throughput} \quad & \{R_{m,T}, W_m\}_{m=1}^N \\
\text{Find} \quad & \{C_m, T_m\}_{m=1}^N \quad \text{to} \\
\text{Minimize} \quad & D_c = \sum_{m=1}^N Pe_m(R_m, E_m, C_m, T_m) W_m \\
\text{subject to} \quad & \sum_{m=1}^N \frac{T_m}{T} E_m \le E_T, \sum_{m=1}^N C_m \le C
\end{aligned} \tag{10}$$

where E_T is the total power constraint and C is the maximum number of subchannels [20].

- *Parallel Transmission*. For a given subchannel to layer assignment $\{C_m\}_{m=1}^N$, the loading algorithm optimizes each layer's power consumption to satisfy a total power contraight:

$$E_{m,T} = \Phi_{\alpha_m}^{-1}\frac{\lambda_{\text{opt}}}{W_m}, \qquad \text{where} \qquad \Phi_\alpha(x) = \sqrt{\frac{\alpha}{x}}\exp(-\alpha x) \tag{11}$$

and λ_{opt} satisfies

$$\sum_{m=1}^N \Phi_{\alpha_m}^{-1}\frac{\lambda_{\text{opt}}}{W_m} = E_T, \quad \text{where} \quad \alpha_m = \frac{3G_m}{2C_m(2^{R_{m,T}/C_m}-1)} \tag{12}$$

To find the optimal $\{C_m\}_{m=1}^N$ to minimize D_c, a successive search can be applied with reasonable computational complexity [20].

- *Serial Transmission*. The time-slot-to-layer assignment $\{T_m\}_{m=1}^N$ requires a power allocation at the source layer level to achieve different error performance during different time slots. We define the power consumption of all the subchannels during layer m's transmission to be e_m. For a given $\{T_m\}_{m=1}^N$, the optimal $\{e_m\}_{m=1}^N$ can be resolved by finding a λ such that [20]

$$E_T = \sum_{m=1}^N \frac{T_m}{T}\Phi_{\beta_m}^{-1}\left(\frac{\lambda}{(1-\rho_m)W_m + \rho_{m+1}W_{m+1}}\right) \tag{13}$$

where

$$e_m = \Phi_{\beta_m}^{-1}\left(\frac{\lambda}{(1-\rho_m)W_m + \rho_{m+1}W_{m+1}}\right)$$

$$\beta_m = \frac{3}{C_m(2^{R_{m,T}/C_m}-1)\dfrac{1}{C_m}\displaystyle\sum_{i=1}^{C_m}\frac{1}{g_i}2^{(R_i-R_{m,T}/C_m)}}$$

$$C_m \le C, m = \cdots N$$

5. SOME APPLICATIONS

Having studied the optimization algorithms for both serial and parallel transmissions, we now present some applications.

5.1. Image and Video

Today's Internet applications such as e-commerce require significant amount of image downloading. Subband/wavelet coding has been a well-known scheme for image compression [1]. The compressed image consists of a set of subbands with different level of perceptual importance. In this example, we consider the quality maximization problem, where image quality is measured by peak signal-to-noise ratio PSNR = 10 log (255^2/MSE), where MSE represents the mean-squared error between the original image and the reconstructed image, a widely used distortion measure. The power constraint is represented by the power usage averaged over the subchannels. Figure 4a depicts the image PSNR as a function of the power constraint E_{av}. A grayscale image "Lena" is subband-coded and quantized to achieve a source data rate of 0.1 and 0.5 bit per pixel (bpp). The ADSL channel consists of 256 subchannels, and supports QAM-64, QAM-32, QAM-16, QAM-8, and QAM-4 modulation types. This example proves that parallel transmission outperforms serial transmission by 4–10 dB in terms of PSNR [20].

Compared to image transmission, video transmission requires more sophisticated optimization because of its large volume, variable data rate, and sensitivity to channel errors. Recent advances in video coding emphasize its error resilience features, mostly in terms of scalability. Using H.263 low bit rate video as an example, low-frequency coefficients, particularly DC coefficients, reflect higher importance compared to high-frequency coefficients; motion vectors have more impact on the decoded video quality than DCT coefficients if corrupted. We use the error-resilient entropy code (EREC) to separate video signals into layers of different importance. EREC is widely recognized since it reorganizes variable-length blocks to fixed-length slots such that each block starts at a known position and the beginning of each block is more immune to error propagation than those at the end [23]. Figure 4b depicts the video quality in terms of averaged PSNR over 60 frames of QCIF (176 × 144 pels) color sequence "Miss America." Similarly, we observe 2–4 dB improvement by parallel transmission compared to serial transmission [20]. The gain is smaller than that of image transmission, since importance classification is much more difficult in video coding.

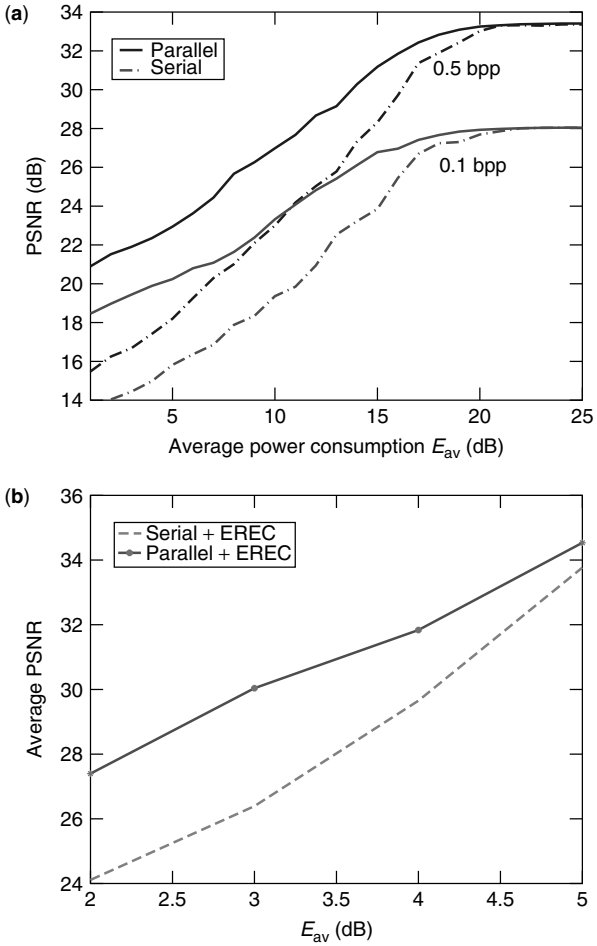


Figure 4. Received PSNR performances for (**a**) image "Lena" and (**b**) video "Miss America."

5.2. Integrated Service of Video, Speech, and Data

Many Internet applications involve integrated services. We select three services of 200, 64, and 10 kbps. We vary the BER distribution to achieve different QoS profiles. Table 1 illustrates the profile characteristics [22]. The outcome of cost/power minimization is shown in Fig. 5 in terms of the transmit power consumption of both serial and parallel transmissions. Similar to the previous examples, parallel transmission outperforms serial transmission by reducing power consumption by 0.5 to 1 dB. To examine the impact of modulation types, we compare the power usage when employing 5 and 11 modulation types. As shown in Fig. 5a, increasing the modulation type can further reduce power usage by 2–3 dB. During power and bit rate loading, the highest modulation type bounds the

Table 1. QoS Requirements

Requirement	Service 1 200 kbps	Service 2 64 kbps	Service 3 10 kbps
QoS1	10^{-6}	10^{-5}	10^{-3}
QoS2	10^{-5}	10^{-3}	10^{-6}
QoS3	10^{-3}	10^{-5}	10^{-6}

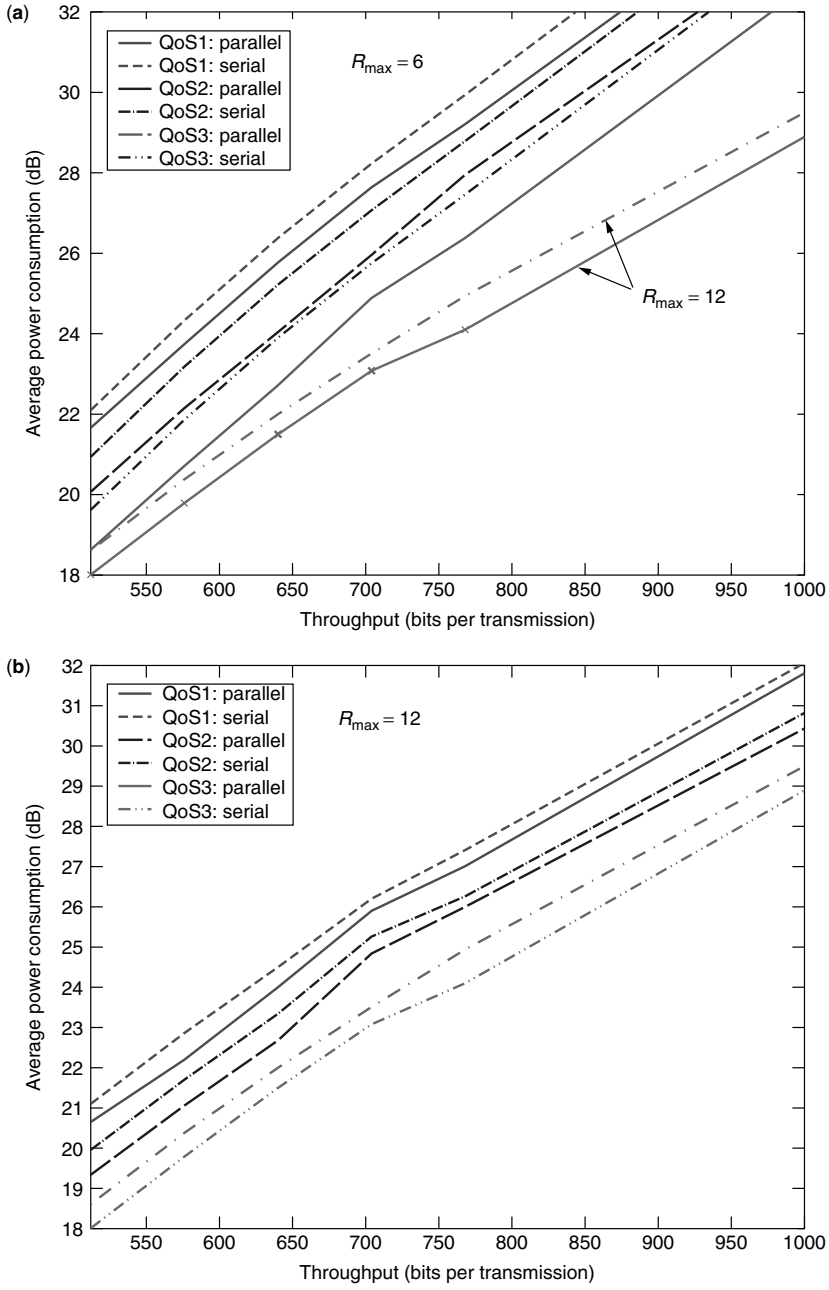


Figure 5. Transmitted Power Consumption vs. Data Throughput for (**a**) $R_{\max} = 6$ and (**b**) $R_{\max} = 12$. The ADSL system is configured to have 256 subchannels and using QAM modulations with $R_{\min} = 2$.

subchannel bit rate. If the allocated bit rate is higher than that offered by the highest modulation, the bit rate is set to that of the highest modulation and the remaining power is allocated to other subchannels. This certainly results in suboptimality. However, the number of modulations also reflects hardware complexity. Quality and complexity should be considered jointly on implementation.

6. CONCLUSION

In this article, we have discussed a key aspect of designing high-performance multimedia communications systems, the ability to perform efficient resource allocation. Most multimedia content consists of layers with different priorities. Unequal error protection, achieved either by source coding or during channel transmission, can effectively reduce resource consumption. We explored the concept of joint optimization in ADSL in terms of two transmission architectures: serial and parallel transmissions. We examined the resource allocation problem in two aspects: cost minimization and quality optimization. The performances of individual allocation algorithms are examined by several practical examples. The study indicates that the parallel transmission architecture, which utilizes the ADSL channel characteristics to provide unequal error protection to source layers, can effectively reduce resource consumption and achieve desirable media quality.

After reading this article, the readers should be able to understand the basic framework of a multimedia communication system, centering on joint source and channel optimization.

BIOGRAPHIES

Haitao Zheng received the B.S. degree in electrical engineering from Xian Jiaotong University, People's Republic of China, in 1995 and the MS and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, Maryland in 1998 and 1999, respectively.

From 1995 to 1998, she was an Institute for System Research Fellow at University of Maryland, College Park. She received the 1998–1999 George Harhalakis Outstanding Systems Engineering graduate Student Award in recognition of outstanding contributions in cross-disciplinary research from the University of Maryland, College Park. Since August 1999, she has been with Wireless Research Laboratory, Bell Labs, Lucent Technologies in Holmdel, New Jersey. Her research interests include design and performance analysis for wireless communications with an emphasis on MAC/PHY layer design, and signal processing techniques for multimedia communications.

K. J. Ray Liu received the B.S. degree from the National Taiwan University and the Ph.D. degree from UCLA, both in electrical engineering. He is a professor in the Electrical and Computer Engineering Department of the University of Maryland, College Park. His research interests span broad aspects of signal processing architectures; multimedia communications and signal processing; wireless communications and networking; information security; and bioinformatics, in which he has published over 230 refereed papers, of which more than 70 arc in archival journals.

Dr. Liu is the recipient of numerous awards, including the 1994 National Science Foundation Young Investigator, the IEEE Signal Processing Society's 1993 Senior Award, and the IEEE 50th Vehicular Technology Conference Best Paper Award, Amsterdam, 1999. He also received the George Corcoran Award in 1994 for outstanding contributions to electrical engineering education and the Outstanding Systems Engineering Faculty Award in 1996 in recognition of outstanding contributions in interdisciplinary research, both from the University of Maryland.

Dr. Liu is editor-in-chief of EURASIP *Journal on Applied Signal Processing* and has been an associate editor of *IEEE Transactions on Signal Processing*; a guest editor of special issues on Multimedia Signal Processing of Proceedings of the IEEE; a guest editor of a special issue on Signal Processing for Wireless Communications of the *IEEE Journal of Selected Areas in Communications*; a guest editor of a special issue on Multimedia Communications over Networks of the *IEEE Signal Processing Magazine*; a guest editor of a special issue on Multimedia over IP of *IEEE Transactions on Multimedia*; and an editor of the *Journal of VLSI Signal Processing Systems*.

BIBLIOGRAPHY

1. J. W. Woods and S. D. O'Neil, Subband coding of images, *IEEE Trans. Acoust. Speech Signal Process.* **34**: 1278–1288 (Oct. 1986).
2. M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
3. T. Gardos, H.263+: The new ITU-T recommendation for video coding at low bit rates, *Proc. IEEE ICASSP*, May 1998, Vol. 6, pp. 3793–3796.
4. J. D. Villasenor and D. S. Park, *Proposed Draft Text for the H.263 Annex V Data Partitioned Slice Mode for Determination at the SG Meeting*, ITU SG16, Proposal Q15 I14, Oct. 1999.
5. *Overall View of the MPEG-4 Standard*, ISO/IEC JTC1/SC29/WG11 N4030, March 2001; *http://mpeg.telecomitalialab.com/standards/mpeg-4/mpeg-4.htm.*
6. Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, Error resilient video coding techniques, *IEEE Signal Process. Mag.* **17**(4): 61–82 (July 2000).
7. S. B. Z. Azami, P. Duhamel, and O. Rioul, Combined source channel coding: Panorama of methods, *Proc. CNES Workshop on Data Compression*, Toulouse, France, Nov. 13–14, 1996.
8. M. Garrett and M. Vetterli, Joint source/channel coding of statistically multiplexed real-time services on packet networks, *IEEE Trans. Network.* **1**(1): 71–79 (Feb. 1993).
9. K. Stuhlmuller, M. Link, and B. Girod, Scalable Internet video streaming with unequal error protection, *Proc. Packet Video Workshop 99*, April 1999, New York.
10. H. Zheng and J. Boyce, An improved UDP protocol for video transmission over Internet-to-wireless networks, *IEEE Trans. Multimedia* **3**(3): 356–365 (Sept. 2001).
11. K. Maxwell, Asymmetric digital subscriber line: Interim technology for the next forty years, *IEEE Commun. Mag.* 100–106 (Oct. 1996).
12. J. A. C. Bingham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* 5–14 (May 1990).
13. M. Barton and M. L. Honig, Optimization of discrete multitone to maintain spectrum compatibility with other transmission systems on twisted copper pairs, *IEEE J. Select. Areas Commun.* **13**(9): (Dec. 1995).
14. P. S. Chow, J. M. Cioffi, and J. A. C. Bingham, A practical discrete multiton transceiver loading algorithm for data transmission over spectrally shaped channels, *IEEE Trans. Commun.* **43**(2): 773–775 (Feb.–April 1995).
15. B. S. Krongold, K. Ramchandran, and D. L. Jones, Computationally efficient optimal power allocation algorithm for multicarrier communication systems, *Proc. Int. Conf. Communications* (*ICC98*), Atlanta, GA, June 1998.
16. J. Campello de Souza, Optimal discrete bit loading for multicarrier modulation systems, *IEEE Symp. Information Theory*, Boston, 1998.
17. R. F. H. Fisher and J. B. Huber, A new loading algorithm for discrete multiton transmission, *Proc. GlobalCOM 96*, pp. 724–728.
18. T. Starr, J. M. Cioffi, and P. J. Silverman, *Understanding Digital Subscriber Line Technology*, Prentice-Hall, Englewood Cliffs, NJ, 1999.
19. H. Zheng and K. J. R. Liu, A new loading algorithm for image transmission over noisy channel, *Proc. 32nd ASILOMAR Conf. Signal, Systems and Computers*, Pacific Grove, CA, Nov. 1998.
20. H. Zheng and K. J. R. Liu, Robust image and video transmission over spectrally shaped channels using multicarrier

modulation, *IEEE Trans. Multimedia* **1**(1): 88–103 (March 1999).

21. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (1948).
22. H. Zheng and K. J. R. Liu, Power minimization for integrated multimedia service over digital subscriber line, *IEEE JSAC* (*Special issue on Error Robust Transmission of Images and Video*) **18**(6): 841–849 (June 2000).
23. D. W. Redmill and N. G. Kingsbury, The EREC: An error-resilient technique for coding variable-length blocks of data, *IEEE Trans. Image Process.* **5**(4): 565–574 (April 1996).

MULTIPLE ANTENNA TRANSCEIVERS FOR WIRELESS COMMUNICATIONS: A CAPACITY PERSPECTIVE

CONSTANTINOS B. PAPADIAS
Global Wireless Systems Research
Bell Laboratories, Lucent Technologies
Holmdel, New Jersey

1. INTRODUCTION

The use of multiple antennas for wireless systems, sometimes referred to as "the spatial frontier," is expected to affect considerably the operation of *wireless networks* [1]. Traditionally, arrays of multiple antenna elements ("antenna arrays") are employed at the base station, due to the associated cost, size, and power constraints that makes their use in wireless terminals more challenging. Used in conjunction with transmit processing on the downlink (base to terminal), or with receive processing on the uplink (terminal to base), *antenna arrays at the base station can offer important performance gains to wireless systems*. By synthesizing spatial beams at the uplink, the base station receiver amplifies the signal-to-noise ratio of a desired user, giving rise to the so-called antenna gain. This gain may in turn be cashed in different ways, such as in an increase of the throughput (data rate), quality, or range of the link. Further, the intelligent shaping of beams (to which the widely used term "smart antennas" is owed) allows one to attenuate the interference that comes from undesired users or undesired locations. An example of such interference mitigation with base station antennas is the use of sectorization in wireless systems; by spatially shaping multiple sectors in each cell, the in-cell interference experienced by each user drops, allowing the co-existence of more users in the cell (higher cell capacities).

On top of the SINR (signal-to-interference-plus-noise ratio) gains described above, antenna arrays may also offer protection against the temporal fluctuations of radio signals, commonly referred to as "fading." *Channel fading is one of the most severe impairments of radio channels*, and its successful handling impacts severely the performance of wireless systems. The best-known way to combat fading is the combining of a number of independently faded received replicas of the transmitted signal. When done cleverly, this (so-called diversity combining) reduces the fluctuation of the signal strength of the received signal against the background noise. As a result, the chance of a deep fade of the received signal is reduced, thus reducing the probability of an outage, specifically, of the situation wherein the link is dropped due to the poor quality of the received signal. As one would expect, the success of diversity combining relies on the degree of statistical independence (typically quantified through the cross-correlation) between the different diversity branches. The spatial dimension, that is, the availability of antenna elements in disjoint spatial locations, is a prime way of obtaining disjointly faded replicas of the transmitted signal for diversity combining. For example, a mobile user located in the proximity of two different base stations may communicate its information to both base antennas. Since the channel is likely to fade in an independent fashion on these two links, the combination of the two independent replicas can offer a substantial diversity gain.

Mechanisms similar to those described above for the realization of performance gains in the uplink can be used in the downlink. For example, if the base station knows the location of a user, it may direct to it a narrow spatial beam. By doing so, it increases the signal-to-noise ratio (SNR) of the user's terminal. Moreover, it helps reduce the interference directed toward users in different locations. To combat fading, the base station may use two antenna elements in order to transmit the downlink signal to a user. For example, by separating widely the two antenna elements, the SNRs of the two links to the user fluctuate in a rather independent fashion. If the base station happens to know at every time instant which of the two links toward the user experiences the best SNR, it may use the corresponding antenna to transmit the signal to the user. This will improve the fading statistics of the signal at the receiver, resulting again in improved reception.

More recently, a number of exciting novel results have given a new push to the field of multiple antennas. This recent revolution began with a research breakthrough by G. J. Foschini at Bell Laboratories in 1996 [2]. Up to that time, the ultimate limit of the spectral efficiency of a wireless link had been studied only for single-antenna systems (or, at most, for single-transmit multiple-receive antenna systems), and it is governed by Shannon's classic capacity formulas for noise-limited channels. In [2] the capacity of wireless links that are equipped with multiple antennas on both sides of the link were studied for the first time; quite astonishingly, the derived capacity formulas showed a spectacular increase in spectral efficiency with the number of transceiver antennas. Roughly speaking, it was shown in [2] that, *when the scattering environment between the multiantenna transmitter and receiver is rich enough, the capacity of the link is roughly proportional to the minimum of the number of antennas on each side* — that is, a doubling of the number of antenna elements on each side of the link is expected to roughly double its spectral efficiency! The capacities predicted by these formulas were unprecedented in the wireless community and have created a large amount of work in order to derive schemes that are capable of delivering them.

In the following, we will describe the main principles that govern the use of multiple antennas in wireless

communication systems. For a compact presentation, we will assume a simple (flat-fading) channel model, which allows us to describe the most important tradeoffs. Moreover, we will focus on link-level studies, for which the metric of Shannon capacity will provide good guidance about both the limitations and success of the described techniques.

2. BACKGROUND AND ASSUMPTIONS

In this section we will outline our assumptions about the considered multiple antenna systems and will provide a corresponding mathematical signal model. We will also briefly review Shannon's classical capacity formula, as it was formulated in the context of single-input/single-output (SISO) systems.

Figure 1 shows a generic architecture of a wireless communication link with M transmitter and N receiver antennas. Such a multiple-input/multiple-output (MIMO) system will be denoted in the remainder of the article as (M, N). As shown in the figure, an original information sequence $\tilde{b}(i)$ that is intended for wireless transmission, undergoes a demultiplexing into multiple data streams before being fed to the transmit antennas. It also typically undergoes forward error correction, interleaving, and spatial multiplexing before being transmitted. Moreover, these operations may happen in a different order. In this article we will be concerned mainly with the channel capacity that can be carried by MIMO channels. However, we will also mention some simple space–time transmission techniques that attempt to attain these capacity bounds.

The demultiplexing/encoding operations result into L data sequences (called *substreams*), denoted by $b_1(k), \ldots, b_L(k)$ (typically $L = M$). After being spatially multiplexed, they are converted into an ensemble of M transmit signals, which are then upconverted to radio frequencies and fed each on every transmit antenna. We will denote the baseband substream transmitted from the mth antenna by $\{s_m(k)\}$. We assume that the physical channel between the mth transmitter and the nth receiver antenna is flat-faded in frequency, so that it can be represented, at baseband, by a complex scalar h_{nm}. The baseband received signal at the receiver antenna array is then represented by a $N \times 1$ vector, denoted by $\mathbf{x}(k)$, that is related to the transmitted substreams as

$$\mathbf{x}(k) = \mathbf{H}\mathbf{s}(k) + \mathbf{n}(k) \tag{1}$$

where $\mathbf{s}(k) = [s_1(k) \cdots s_M(k)]^T : M \times 1$ vector snapshot of transmitted substreams, each assumed of equal variance σ_s^2
$\mathbf{H} = N \times 1$ channel matrix
$\mathbf{x}(k) = N \times 1$ vector of received signal snapshots
$\mathbf{n}(k) = N \times 1$ vector of additive noise samples, assumed i.i.d. and mutually independent, each of variance σ_n^2

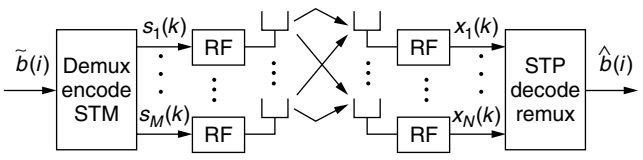


Figure 1. A generic (M, N) multiple antenna transceiver architecture.

We also denote by superscript $*$, T, $\dagger$ the complex conjugate, transpose, and Hermitian transpose, respectively, of a scalar or matrix.

Throughout the remainder of the article, we have chosen to use channel capacity, in the Shannon sense, as a metric for the evaluation of various wireless MIMO systems. Traditionally, smart antennas were viewed as a technology that can help increase the range of a wireless system, the quality of wireless voice calls, or the number of voice users that can be supported in the cell area. However, the advent of MIMO systems since 1996 or so has added a new dimension to the smart antenna technology. It allows us to increase the spectral efficiency of wireless links, without increasing the transmit power; that is, for the same power budget, higher data rates (information throughputs) can be attained.

Channel capacity, in the Shannon sense, is the prime theoretical tool for evaluating the maximum information throughput that can be supported by a communication link. We will hence use it for evaluating the capacity of MIMO systems. Moreover, we will show that, in hindsight, the use of the channel capacity metric offers interesting intuition about the performance of more conventional smart antenna systems. While it can be argued, and rightfully so, that channel capacity is not the sole tool for analyzing wireless systems, we believe that it sheds some light on a number of important issues in smart antenna systems.

As a background to the remainder of the article, we review briefly the channel capacity of single-transmit/single-receive (1, 1) systems. In this case, the (assumed flat-faded) channel is modeled through the complex scalar h, which represents its complex gain. The capacity of this channel, assuming additive white Gaussian noise (AWGN), independent from the transmitted signal, is given by Shannon's capacity formula:

$$C = \log_2(1 + \rho|h|^2) \quad \text{(bps/Hz)} \tag{2}$$

where ρ is the signal-to-noise ratio (SNR) defined as $\rho = \dfrac{\sigma_s^2}{\sigma_n^2}$, σ_s^2 and σ_n^2 being the signal and noise variance, respectively.

Note that the formula indicates that capacity grows logarithmically as a function of the SNR ρ. In other words, any increase in the link's SNR will only be reflected into a corresponding *logarithmic* increase in the channel's capacity. Consequently, each extra bit per second per Hertz (bps/Hz) of capacity, requires roughly a doubling of the link's SNR. This results in a high price for extra capacity — for a linear increase in spectral efficiency, the power needs to be increased exponentially!

3. THE NOTION OF RANDOM CAPACITY

Before discussing the extension of the (1, 1) capacity in Eq. (2) to MIMO systems, we believe that it is important to first underline the random character of wireless systems.

The capacity expression in (2) — similar to the MIMO capacity expressions that follow — is silent about one important issue that underlies the performance of wireless systems in practice, namely, their *statistical behavior*. This is due to the fact that it implicitly assumes the channel to be constant forever, that is, *static*. This point of view leaves out two important aspects of wireless systems:

1. The temporal variation of the channel: each user's (link) changes with time, mainly due to the user's and the environment's mobility
2. The spatial distribution of channels in a geographic area

In other words, in a wireless system, the channel is typically not static, when seen from the perspective of each user, and it is not uniformly distributed at the level of a user population. In one approach that is often taken in order to make capacity expressions relevant to practical systems, the user channel $\mathbf{H}$ is assumed to be *semistatic*. This means that, during time intervals that are of finite duration, but at the same time long enough to allow the desired benefit of error correction coding, the channel is considered static. However, from one such time interval to another, the channel is assumed to be different. *The capacity of such a semistatic channel can then be modeled as a random variable* C, where a pool (statistical ensemble) of channel realizations $\{\mathbf{H}\}$ corresponds to the different time intervals wherein the channel remains static. The capacities corresponding to each realization constitute an ensemble of capacities $\{C\}$ (capacity distribution). The attributes of this ensemble, contained in its cumulative density function (CDF) can be then evaluated in order to assess the statistical capacity of the system. The two most commonly used measures are

1. *Outage capacity* — the capacity point of the CDF that happens with probability higher than a certain target threshold:

$$C_o = \{C \text{ such that } \Pr\{C \geq C_o\} = P_o\} \tag{3}$$

   where P_o is the predetermined outage target. The quantity $1 - P_o$ is often called the *outage probability*. Typically, in cellular voice systems, P_o is chosen to be around 90%, corresponding to an outage probability of 10%. This means that there will be only a 10% probability of the system being in outage, that is, of a user not being able to reach the capacity target C_o.
2. *Average capacity* — the expected value of capacity over the entire CDF:

$$C_a = E(C) \tag{4}$$

   where E denotes statistical expectation. This measure is not traditionally used in voice systems, where it is important to guarantee good (low-latency) service with high probability. However, it appears that it may be a more relevant quantity in wireless data systems, where the bursty nature of data, combined with the higher tolerance to latency and support from higher network layers, makes average throughput a relevant quantity.

The statistical characterization of capacity outlined above does not apply only to time-varying wireless channels. Equally importantly, it can be used to characterize a wireless system in terms of *spatial user distribution*. To illustrate this point, consider a wireless system with static users that are randomly distributed in the geographic area surrounding the base station. Such a scenario applies for example with good accuracy to so-called fixed wireless systems, where the base station communicates with a number of static rooftop antennas. Another example would be that of very low-mobility users getting wireless service in a local-area network (LAN). In such cases, it is important to know what is the capacity as a function of user geographic location. This can be captured in a CDF for the entire ensemble of locations of interest. The notions of outage and average capacities apply here, too. The former conveys the percentage of spatial locations that can be supported with a certain capacity, whereas the latter relates to the total data throughput provided to the entire set of locations. For a service provider, spatial outage may be used as a metric to guarantee a certain quality of service to the customers, whereas the average throughput may relate more to the total revenue per time unit expected in a certain service area.

4. OPEN- AND CLOSED-LOOP MIMO LINK CAPACITIES

In this section, we will present the theoretical maximum spectral efficiencies (channel capacities) that are achievable in multiple antenna links. For a moment, we will neglect the random character of channel capacity; the capacity expressions that will be presented will be subject to specific channel instantiations (channel snapshots). Later on, though, these instantaneous capacity expressions will be numerically evaluated over statistical ensembles that correspond to either channel or user behavior, in order to capture the random aspect of capacity mentioned above.

In our treatment, we will consider the general case of an arbitrary number of antennas on each side of the link. As capacity depends on the knowledge of the MIMO channel response at the transmitter, we will treat separately the two extreme cases: the *open-loop capacity*, which assumes no channel knowledge at the transmitter; and the *closed-loop capacity*, which assumes full channel knowledge at the transmitter. As mentioned above, the presented formulas will be used in subsequent sections in order to analyze special cases and to evaluate the merits of different transmission/reception schemes.

4.1. Open-Loop Capacity

In the open-loop case, the Shannon capacity of the (M, N) flat-faded channel is given [2] by the now familiar (so-called "log-det") formula:

$$C = \log_2 \left\{ \det \left(I_N + \frac{\rho}{M} \mathbf{H}\mathbf{H}^{\dagger} \right) \right\} \quad \text{(bps/Hz)} \tag{5}$$

where the SNR is now defined as $\rho = M\sigma_s^2/\sigma_n^2$. A first important observation that can be made from this equation is that, for rich scattering channels, the MIMO channel capacity grows roughly proportionally to the minimum of the number of transmitter and receiver antennas [2–4].

This is an astonishing result, contrasted to the logarithmic capacity increase as a function of signal-to-noise ratio. Its consequences are quite dramatic as far as the spectral efficiency of a wireless link with a certain power budget is concerned. For example, consider a (1, 1) system that operates at 0 dB SNR. From Eq. (2), its average capacity over an ensemble of i.i.d. Gaussian (Rayleigh-amplitude) channel coefficients h with $E|h|^2 = 1$ is equal to $C_{a,1} = 0.85$ bps/Hz. At the same SNR, a (10, 10) system whose channel coefficients between any transmitter/receiver pair are again independently fading i.i.d. unit-variance Rayleigh variables, has an average capacity of $C_{a,10} = 8.38$ bps/Hz. This corresponds roughly to a 10-fold increase in capacity using 10 antennas on each side of the link and keeping the total transmission power constant ($C_{a,10}/C_{a,1} = 9.76$). Notice that a 10-fold capacity increase in the original (1, 1) system would require to double the power 9 times, that is, increase the power by $2^9 = 512$ times!

Figure 2 shows a graphical depiction of the linear capacity increase with the number of transmit antennas at different SNRs. We have plotted capacities at the 10% outage level for symmetrical MIMO systems that have up to 20 antennas on each side of the link. Each channel realization is a square matrix whose elements are chosen independently from a Rayleigh distribution of unit variance. Notice the extraordinarily high capacities achieved with many antennas (on the order of hundreds of bps/Hz). For comparison, it should be kept in mind that current cellular wireless systems operate typically at no more than 2 bps/Hz. It is also worth noting that, *in MIMO systems, the higher the SNR, the steeper the capacity increase*.

4.2. Closed-Loop Capacity

A generalization of the capacity formula (5) to the case where the transmitter has some knowledge of the channel characteristic $\mathbf{H}$ was derived in [3]:

$$C = \log_2 \left\{ \det \left(I_N + \frac{\rho}{P_T} \mathbf{H} \Phi \mathbf{H}^\dagger \right) \right\} \quad \text{(bps/Hz)} \qquad (6)$$

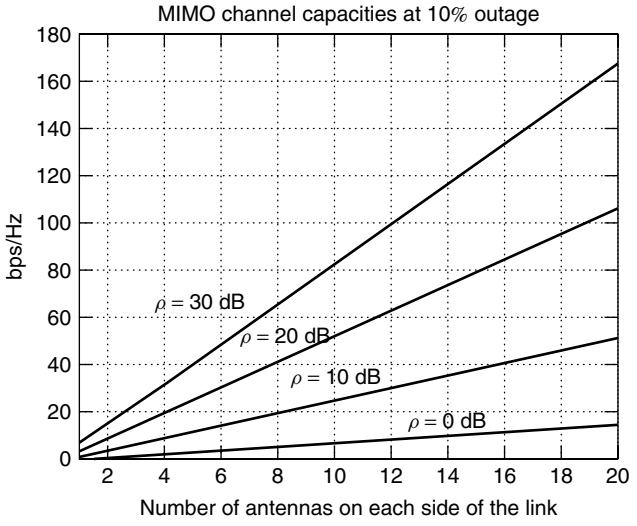


Figure 2. Channel capacity of MIMO systems as a function of the number of antennas.

where Φ is the $M \times M$ covariance matrix of the transmitted signal ($\Phi = E(\mathbf{s}\mathbf{s}^\dagger)$) and P_T is the total transmitted power from the M antennas [$P_T = M\sigma_s^2$, and $tr(\Phi) = P_T$, where $tr()$ denotes the trace of a matrix]. When the channel $\mathbf{H}$ is fully known at the transmitter, Φ in Eq. (6) can be optimized by so-called "spatial water filling." Spatial waterfilling is a notion similar to the one of frequency waterfilling (water pouring) used in orthogonal frequency division multiplexing (OFDM) systems in order to maximize capacity; it is based on the idea of distributing the available transmitted power in a nonuniform fashion across the different "modes" of a channel. In the case of narrowband MIMO systems that we are examining, these are *spatial* modes, defined as the eigendirections of the transmitted signal covariance matrix Φ. The use of a matrix Φ that is not a multiple of an identity matrix denotes a nondiagonal loading of the channel's spatial modes.

In the following, some special cases of (6) will help illustrate the concept of spatial waterfilling. For the moment, note that by choosing $\Phi = (P_T/M)\mathbf{I}_M$, (6) reduces to the open-loop capacity in (5). This represents the fact that when the transmitter has no channel knowledge, diagonal loading of the modes is used, resulting in the open-loop capacity (5).

5. (1, N) SYSTEMS

In this section, we will consider single-input/multiple-output (SIMO) systems. This case is frequently encountered in practice, for example, in the uplink (terminal-to-base) of cellular systems. For reasons of cost, complexity, power, and size, mobile terminals are typically equipped with a single antenna, whereas today's base stations already use two and will soon use even more antennas for reception. We should also note that SIMO systems are, in many ways, more conventional than both MISO (multiple-input/single output) and MIMO (multiple-input/multiple output) systems. This has to do with the fact that, when having a single transmitter antenna, the issue of spatial multiplexing of several substreams onto a number of antennas does not arise.

In this case, the signal model of (1) reduces to

$$\mathbf{x}(k) = \mathbf{h}s(k) + n(k) \qquad (7)$$

where $\mathbf{h} = [h_1 \cdots h_N]^T$ is of dimension $N \times 1$, and both $s(k)$, $n(k)$ are scalars. By evaluating the open-loop capacity (6) for $M = 1$, we obtain (after some algebraic manipulation) the following capacity expression for (flat faded) SIMO systems:

$$C = \log_2 \left(1 + \rho \sum_{n=1}^{N} |h_n|^2 \right) \quad \text{(bps/Hz)} \qquad (8)$$

By contrasting the expression in (8) with the one of (1, 1) systems given in (2), it is clear that, from a capacity perspective, the use of the extra $N - 1$ receiver antennas has resulted in an increase of the link's instantaneous SNR from $\rho|h|^2$ to $\rho \sum_{n=1}^{N} |h_n|^2$. The following observations can be made at this point:

5.1. Capacity Scaling

By rewriting the SNR gain in (8) as

$$\sum_{n=1}^{N} |h_n|^2 = N \times \left(\frac{1}{N} \sum_{n=1}^{N} |h_n|^2 \right)$$

we see that as N grows, the quantity $\frac{1}{N} \sum_{n=1}^{N} |h_n|^2$ converges to an average value. However, the multiplicative factor N keeps growing. In other words, for large N, $C_{1,N}$ grows approximately as

$$C_{1,N} \simeq \log_2(1 + \rho N) \tag{9}$$

that is, *the capacity of a* $(1, N)$ *system increases approximately logarithmically with the number of receiver antennas* N. Notice that in (9), we have assumed that $E|h_n|^2 = 1$ for all $n \in \{1, \ldots, N\}$, which gives $\lim_{N \to \infty} \left(\frac{1}{N} \sum_{n=1}^{N} |h_n|^2 \right) = 1$. The fact that the capacity does not saturate as more antennas are added to the receiver comes from the fact that more power is collected at the receiver for the same power budget at the transmitter.

5.2. Diversity Versus Power Gain

The SNR gain indicated by the multiplicative factor $\sum_{n=1}^{N} |h_n|^2$ is an instantaneous quantity. However, as mentioned above, statistical behavior is important in determining the performance of the system. The statistical behavior is in turn strongly dependent on the degree of correlation between the N receiver antenna elements. The following two extreme cases are typically considered:

- *Uncorrelated Antenna Elements.* This case arises (with good approximation) when the antenna elements are spaced far from each other (on the order of 10λ, where $\lambda = f/c$ is the carrier wavelength). Mathematically, this is expressed by the fact that $E(\mathbf{x}\mathbf{x}^\dagger)$ is a matrix close to diagonal. From a physical point of view, this property indicates that the signals received on different antenna elements fade independently. In this case, the SNR gain represented by the quantity $\sum_{n=1}^{N} |h_n|^2$ represents not only more collected power but also an improvement in the received signal's statistics against fading (it is a combined diversity/power gain).
- *Fully Correlated Antenna Elements.* This case arises when the antennas are closely spaced (on the order of $\lambda/2$). Then, in radiofrequencies, a wavefront impinging on the antenna array causes only phase differences on the signal received from the different elements. In other words, the time interval required by the radiowave to propagate across the antenna array is so small (compared to the inverse of the signal's bandwidth) that the antenna elements appear to be fully correlated.[1] As a result, all the antennas fade simultaneously. The gain $\sum_{n=1}^{N} |h_n|^2$ then is a pure SNR gain — it provides no diversity protection, in the sense that, if one antenna is faded, all other antennas will be faded as well. However, in any given channel realization (even if the channel never changes) the presence of N antennas increases the link's SNR by a factor of N (or $10 \log_{10}(N)$ decibels). This alludes to the classic "3 dB power gain" that is associated in coherent antenna systems with the doubling of antenna elements. Notice that in both extreme cases presented above, the average SNR gain (as well as the average capacity gain) is the same. However, the corresponding CDF's are different.

[1] This is sometimes referred to as the "narrowband assumption" in the antenna array literature [1].

5.3. Receiver Options for Capacity Attainment

Adhering to the assumed nondispersive signal model in (7), the optimal receiver is a linear combiner that operates as follows on the received signal to produce soft outputs for detection:

$$y(k) = \mathbf{h}^\dagger \mathbf{x}(k) \tag{10}$$

It is noteworthy that this simple linear receiver, in the case considered, allows the satisfaction of the following optimality criteria simultaneously:

- Maximum-likelihood detection
- Maximum conditional expectation
- Maximum output signal-to-noise ratio
- Minimum mean-squared error

(This is due partly to the assumed white and Gaussian character of the noise.) More importantly, *this linear receiver allows the attainment of the* $(1, N)$ *channel capacity* in (8). By this phrase we mean that, with the use of progressively stronger — such as Turbo (spatially one-dimensional) — encoding of the original input stream, the capacity in (8) will be attained asymptotically. This property is a direct consequence of the fact that, after the combining in (10) takes place, the $(1, N)$ system has been essentially converted to a (1, 1) system, for which state-of-the art encoding techniques that closely approximate channel capacity exist. In the next section, this property will be contrasted with the capacity attainment capabilities of $(M, 1)$ systems.

5.4. Open-Loop Versus Closed-Loop Operation

Another interesting property of SIMO systems is that their open- and closed-loop capacities coincide. This can be easily verified by comparing Eqs. (5) and (6) for $M = 1$. In this case, $\Phi = \phi$ is scalar. The constraint $tr(\phi) = P_T$ then results in $\phi = P_T$. This results in (6) coinciding with (5) in the case $M = 1$. The result is expected largely since with a single-transmitter antenna, the notion of spatial water filling does not apply. It also corroborates the fact, mentioned above, that $(1, N)$ systems can be represented by equivalent (1, 1) systems, for which there is no difference between open-loop and closed-loop capacities, either.

6. (M, 1) SYSTEMS

The narrowband signal model for systems with a single-receiver antenna is given by

$$x(k) = H\mathbf{s}(k) + n(k) \tag{11}$$

where $H^T = [h_1 \cdots h_M]^T$ and $\mathbf{s}(k) = [s_1(k) \cdots s_M(k)]^T$ are of dimension $M \times 1$ and $x(k)$, $n(k)$ are scalar. As mentioned above, in $(M, 1)$ systems, there are significant differences between the open- and the closed-loop cases. We will hence examine the two cases separately.

6.1. Open-Loop (M, 1) Systems

By evaluating (5) for $N = 1$, we obtain

$$C^{o}_{M,1} = \log_2\left(1 + \frac{\rho}{M}\sum_{m=1}^{M}|h_m|^2\right) \quad \text{(bps/Hz)} \tag{12}$$

By contrasting expression (12) to Eq. (8), we observe the trends described in the following paragraphs.

6.1.1. Diversity Gain. The SNR gain in (12) equals $1/M\left(\sum_{m=1}^{M}|h_m|^2\right)$. Assuming uncorrelated antenna elements, this is a pure *diversity gain*. This is to be contrasted with the joint power/diversity gain achieved in the $(1, N)$ case (see Section 5). This is immediately realized when evaluating the expected value of the SNR gain over many realizations. Assuming that $E|h_m|^2 = 1$ for all $m \in \{1, \ldots, M\}$, we obtain

$$E\left(\frac{1}{M}\sum_{m=1}^{M}|h_m|^2\right) = 1$$

In other words, adding more and more antennas at the transmitter side, without the benefit of more than one antenna at the receiver, does not change the average SNR at the receiver. This stems from the fact that the total transmitted power is kept constant at the transmitter, whereas in the $(1, N)$ case, each extra receiver antenna allows us to collect more signal power against the background noise.

The diversity gain, which is due to the assumption of uncorrelated antenna elements, is again expressed through a change in the CDF of the $(M, 1)$ system for each different M, which results in improved outage capacity (particularly at low outages). Moreover, the highest gain is achieved when going from 1 to 2 transmit antennas. As M keeps growing, it becomes smaller and eventually it saturates. Indeed, in the limit as $M \to \infty$ (and always keeping the total transmit power constant and equal to P_T), the capacity expression in (12) converges to the following expression:

$$C_{\infty,1} = \log_2(1 + \rho) \quad \text{(bps/Hz)} \tag{13}$$

[compare to (9)]. This means that, *in MISO systems, beyond a certain point, there is no benefit in adding more antennas at the transmitter*. Finally, we should note that, if the M antenna elements were fully correlated, there would be no benefit in having $M > 1$, since the received signal's CDF would be the same for all M.

6.1.2. Capacity Attainment. Unlike the simple type of processing required in $(1, N)$ systems to attain (in the sense mentioned above) their capacity in Eq. (8), the attainment of the open-loop capacity (12) in the $(M, 1)$ case seems to be a challenging task. More specifically, attaining the capacity in (12) requires sophisticated space–time coding (STC) techniques [5] at the transmitter. This means that the encoding/spatial multiplexing operations shown in Fig. 1 are nontrivial. So far, only the (2, 1) case seems to admit a straightforward STC technique that allows the attainment of its open-loop capacity [6,7]. This technique, which is briefly described below, relies on a smart (2, 1) space–time multiplexing idea developed by Alamouti [8].[2] In the case $M > 2$, no simple open-loop techniques are known that allow us to attain the capacity in Eq. (12). A (4, 1) technique presented in [9] allows us to get very close to the (4, 1) capacity (achieving on the order of 95% of it or so), and is also briefly discussed below. Despite the diminishing returns of $(M, 1)$ systems for M beyond four antennas, the quest for open-loop capacity attainment is ongoing.

6.1.3. Examples of (M, 1) Space–Time Transmission Schemes

6.1.3.1. (2, 1) Systems: The Alamouti Scheme. An ingenious transmit diversity scheme for the (2, 1) case was introduced by Alamouti [8], and remains to date the most popular scheme for (2, 1) systems. We denote by $\mathbf{S}$ the 2×2 matrix whose (i, j) element is the encoded signal going out of the jth antenna at odd $(i = 1)$ or even $(i = 2)$ time periods (the length of each time period equals the duration of one encoded symbol). In other words, one could think of the vertical dimension of $\mathbf{S}$ as representing "time" and of its horizontal dimension as representing "space." We also denote, as described in Section 2, by $\{b_l(k)\}$, $l = 1, 2$, the encoded version of the lth substream of the original signal $\{\tilde{b}(i)\}$ (see Fig. 1). The Alamouti scheme transmits the following signal every two encoded symbol periods:

$$\mathbf{S}(k) = [\mathbf{s}_1(k)\,\mathbf{s}_2(k)] = \begin{bmatrix} b_1(k) & b_2(k) \\ b_2^*(k) & -b_1^*(k) \end{bmatrix} \tag{14}$$

Having assumed, as noted earlier, the channel to be flat in frequency, the (2, 1) channel is characterized through $H = [\,h_1 \quad h_2\,]$. We group the odd and even samples of the received signal in a 2×1 vector $\mathbf{x}(k)$, which can be then expressed in baseband as

$$\begin{aligned}\mathbf{x}(k) = {} & (h_1(b_1(k)\mathbf{c}_1 + b_2^*(k)\mathbf{c}_2) \\ & + h_2(b_2(k)\mathbf{c}_1 - b_1^*(k)\mathbf{c}_2)) + \mathbf{n}(k)\end{aligned} \tag{15}$$

where $\mathbf{c}_1^T = [1 \mid 0]$, $\mathbf{c}_2^T = [0 \mid 1]$. After subsampling at the receiver and complex-conjugating the second output, we obtain

$$\begin{aligned} d_1(k) &= \mathbf{c}_1^T\mathbf{x}(k) = (h_1 b_1(k) + h_2 b_2(k)) + v_1(k) \\ d_2(k) &= (\mathbf{c}_2^T\mathbf{x}(k))^* = (-h_2^* b_1(k) + h_1^* b_2(k)) + v_2^*(k) \end{aligned} \tag{16}$$

[2] An extension of this technique to CDMA systems, called space–time spreading (STS), was presented in [11] and has been introduced in third-generation wireless standards for the (2, 1) case.

where $v_m(k) = \mathbf{c}_m^T \mathbf{n}(k)$, $m = 1, 2$. Equation (16) can be equivalently written as

$$\begin{aligned}\mathbf{d}(k) &= \begin{bmatrix} h_1 & h_2 \\ -h_2^* & h_1^* \end{bmatrix} \begin{bmatrix} b_1(k) \\ b_2(k) \end{bmatrix} + \mathbf{v}(k) \\ &= \mathbf{H}\mathbf{b}(k) + \mathbf{v}(k)\end{aligned} \tag{17}$$

where $\mathbf{v}^T(k) = [\, v_1(k) \quad v_2^*(k) \,]$ and $\mathbf{H}$ is a unitary matrix (up to a complex scalar). After match filtering to $\mathbf{H}$, we obtain

$$\begin{aligned}\mathbf{d}'(k) = \mathbf{H}^\dagger \mathbf{d}(k) = &\begin{bmatrix} |h_1|^2 + |h_2|^2 & 0 \\ 0 & |h_1|^2 + |h_2|^2 \end{bmatrix} \\ &\times \mathbf{b}(k) + \mathbf{v}'(k)\end{aligned} \tag{18}$$

where $\mathbf{v}'(k)$ remains spatially white. Because of the diagonal character of the mixing matrix in (18), the 2×1 space–time system has been now reduced to an equivalent set of two 1×1 systems! We call this the *decoupled property* of a space–time code. Moreover, each of these two equivalent single-dimensional systems has the same capacity. We call this the property of *balance* of a space–time code.

The total constrained capacity of this system equals the sum of the capacities of the two SISO systems (each SISO system operates at half the original information rate):

$$C_{2,1}^A = \log_2 \left(1 + \frac{\rho}{2}(|h_1|^2 + |h_2|^2)\right) \tag{19}$$

By contrasting (19) to (12) with $M = 2$, we see that

$$C_{2,1}^A = C_{2,1}^o \tag{20}$$

Hence, *the Alamouti scheme allows the attainment of the* $(2, 1)$ *open-loop capacity*. Moreover, this is possible with the use of conventional (spatially single-dimensional, such as Turbo) encoding. The Alamouti scheme remains, to the best of our knowledge, the only decoupled and balanced space–time code that allows the attainment of the system's full open-loop capacity.

6.1.3.2. (4, 1) Systems: A More Recently Proposed Scheme. As mentioned above, the attainment of the open-loop capacity in the general $(M, 1)$ case is a challenging problem. A scheme for the (4, 1) case that approaches the capacity closely was proposed in 2001 [9]. According to this scheme, the original information sequence $\tilde{b}(i)$ is first demultiplexed into four encoded substreams $b_m(k)$ $(m = 1, \ldots, 4)$. The four-dimensional transmitted signal is then organized in blocks of $L = 4$ (encoded) symbol periods and is represented by a 4×4 matrix $\mathbf{S}$, which is arranged as follows:

$$\mathbf{S} = \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \\ b_2^* & -b_1^* & b_4^* & -b_3^* \\ b_3 & -b_4 & -b_1 & b_2 \\ b_4^* & b_3^* & -b_2^* & -b_1^* \end{bmatrix} \tag{21}$$

where the time index k has been dropped for convenience. Similar to the (2, 1) case, the mth column of $\mathbf{S}$ in (21) represents a block of 4 symbols that are transmitted from the mth transmit antenna. The maximum capacity attainable by this transmission technique was computed in [9] and it is given by the following expression:

$$C_{4,1}^{\text{proposed,max}} = \frac{1}{2} \log_2 \det \left(\mathbf{I}_2 + \frac{\rho}{4} \Delta_1\right) \tag{22}$$

where

$$\Delta_1 = \begin{bmatrix} \gamma & \alpha \\ -\alpha & \gamma \end{bmatrix} \tag{23}$$

and

$$\begin{aligned}\gamma &= \mathbf{h}^H \mathbf{h} = \sum_{m=1}^{4} |h_m|^2 \\ \alpha &= 2j\text{Im}(h_1^* h_3 + h_4^* h_2)\end{aligned} \tag{24}$$

(where Im denotes the imaginary part of a complex scalar). Some quantitative results regarding the capacities of these techniques will be given in Section 7.

6.2. Closed-Loop $(M, 1)$ Systems

Unlike the open-loop case, the closed-loop capacity expression for $(M, 1)$ systems is given by the following expression:

$$C_{M,1}^c = \log_2 \left(1 + \rho \sum_{m=1}^{M} |h_m|^2\right) \quad \text{(bps/Hz)} \tag{25}$$

Notice that the SNR gain in Eq. (25) is similar to the $(1, N)$ case [see Eq. (8)]. The use of extra transmitter antennas now adds to the receiver power (and hence capacity keeps growing with M). For large M, the closed-loop capacity of $(M, 1)$ systems scales as follows:

$$C_{M,1} \simeq \log_2(1 + \rho M) \quad \text{(bps/Hz)} \tag{26}$$

[compare to Eq. (9)]. When the channel H is fully known at the transmitter, the closed-loop capacity in (25) is easily attainable through spatial waterfilling at the transmitter, as will be described below.

In conclusion, we observe that $(N, 1)$ and $(1, N)$ systems are not in general symmetric. They are, however, symmetric, when the $(N, 1)$ channel is perfectly known at the transmitter, allowing it to perform spatial maximal ratio combining (MRC) before transmission (spatial pre-equalization), as will be shown below.

6.2.1. Example Schemes

6.2.1.1. Transmit MRC. The optimal transmission approach in the $(M, 1)$ case, when the channel is flat and fully known at the transmitter, is to send the following signal out of the M antennas:

$$\mathbf{s}(k) = \left(\frac{1}{\sqrt{\sum_{m=1}^{M} |h_m|^2}}\right) \begin{bmatrix} h_1^* \\ \vdots \\ h_M^* \end{bmatrix} b(k) \tag{27}$$

(see [11]) which amounts to performing MRC at the transmitter. The operation is mathematically equivalent to the receive MRC performed at the receiver in $(1, N)$ systems, as described in (10). Notice that the same information sequence is sent simultaneously out of all the antennas, however it is multiplied by a different complex scalar on each antenna. As mentioned above, this

simple transmission scheme (provided that the channel is perfectly known at the transmitter), achieves the $(M, 1)$ closed-loop capacity for any M.

6.2.1.2. Switch Transmit Diversity (STD). In some cases, good knowledge of the channel coefficients at the base station is not feasible. This may be due, for example, to the excessive amount of feedback required in order to send back reliably the channel information from the terminal to the base station, the highly time-varying nature of the channel (high Doppler), errors in channel estimation, and errors in the feedback channel. In such cases, there is interest in exploring alternative methods that make use of *partial* channel state information at the transmitter. One very good and simple candidate is so-called selection transmit diversity (STD). This technique transmits, at each symbol period, only from one antenna, at full power. The symbol is transmitted from the antenna that experiences the highest SNR during that symbol period. Mathematically, STD transmission can be described as follows:

$$\mathbf{s}(k) = \delta_m(k) b(k) \tag{28}$$

where $\delta_m(k)$ is an $M \times 1$ vector, whose single nonzero entry is at position m, where $m \in \{1, \ldots, M\}$ is such that $|h_m|$ takes the highest value in that set during the kth symbol period.

From a practical point of view, this technique is appealing because it requires only the knowledge of which is the strongest antenna element ($\log_2(M)$ bits of information per symbol). From a capacity point of view, the STD scheme is capable of achieving the following capacity:

$$C_{M,1}^{\text{STD}} = \log_2(1 + \rho \max_{m \in \{1,\ldots,M\}} |h_m|^2) \quad \text{(bps/Hz)} \tag{29}$$

As it turns out, the capacity in (29) typically lies roughly midway between the $(M, 1)$ open-loop capacity (12) and the $(M, 1)$ closed-loop capacity (25).

6.2.2. Numerical Examples. Figure 3 shows the 10% outage capacities for some open- and closed-loop $(M, 1)$ cases. As predicted by expressions (12) and (25), the difference between the (4, 1) open- and closed-loop capacities is about 6 dB. Moreover, notice that the (4, 1) open-loop capacity is only about 2 dB better than the (2, 1) open-loop capacity (diminishing returns), whereas the (4, 1) STD capacity lies midway between the open- and closed-loop capacities.

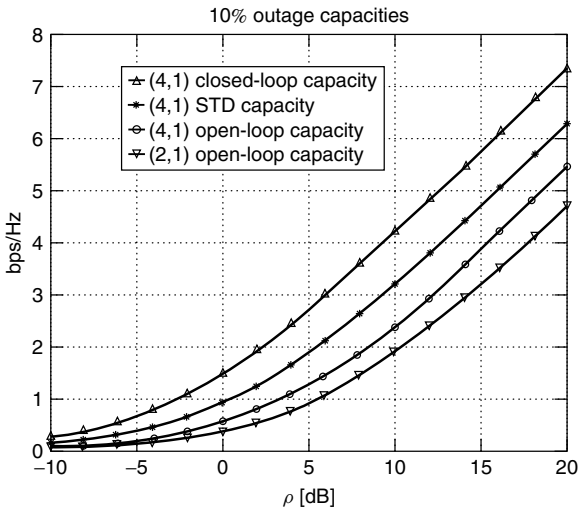


Figure 3. Open- and closed-loop $(M, 1)$ capacities.

7. (M, N) SYSTEMS

In this section we will present some practical existing techniques that attempt to approach the capacity of open-loop systems in the general (M, N) case.

7.1. Combined *Transmit* + *Receive* Diversity Systems

Given a certain $(M, 1)$ system, one straightforward way to design an (M, N) system is to simply

- Transmit as in the $(M, 1)$ system
- Receive on each antenna as in the $(M, 1)$ system
- Combine optimally the N receiver antenna outputs

The capacity quantification of these transmit/receive diversity systems is straightforward. The $M \times N$ (assumed flat) channel is represented through the $N \times M$ channel matrix:

$$\mathbf{H} = \begin{bmatrix} h_{11} & \cdots & h_{1M} \\ \vdots & \ddots & \vdots \\ h_{N1} & \cdots & h_{NM} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1 & \cdots & \mathbf{h}_N \end{bmatrix}$$

We first compute an upper bound for the capacity of such an (M, N) transmit/receive diversity system. With optimal ratio combining, and assuming that each $(M, 1)$ system takes no interference hit, the input/output relationship takes the form

$$\mathbf{d}(k) = \left(\sum_{m=1}^{M} \sum_{n=1}^{N} |h_{nm}|^2 \right) \mathbf{b}(k) + \mathbf{n}(k) \tag{30}$$

where $\mathbf{d}(k)$, $\mathbf{b}(k)$, and $\mathbf{n}(k)$ are all of dimension $M \times 1$. The corresponding capacity is given by

$$C_{M,N}^{\text{trd,max}} = \log_2 \left(1 + \frac{\rho}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} |h_{nm}|^2 \right) \tag{31}$$

It is clear that, when the attainable capacity of the corresponding $(M, 1)$ schemes is away from the $(M, 1)$ log-det capacity, the upper bound in (31) will not be attained either. Notice further that the expression in (31) is strictly smaller than the (M, N) log-det capacity in (5) for $N > 1$.

7.1.1. Example Schemes. To give some examples, the capacity of a $(2, N)$ system that uses the Alamouti (2, 1) scheme is

$$C_{2,N}^{A} = \log_2 \left(1 + \frac{\rho}{2} \sum_{n=1}^{N} (|h_{n,1}|^2 + |h_{n,2}|^2) \right) \tag{32}$$

thus, as expected, the upper bound in (31) is "attained" by the Alamouti scheme in the $(2, N)$ case. However, this still falls short of the $(2, N)$ log-det capacity (5).

It is also straightforward to compute the maximum attainable capacity of a $(4, N)$ system that uses the (4, 1) scheme of [9], which is given by

$$C_{4,N}^{\text{proposed,max}} = \frac{1}{2}\log_2 \det\left(\mathbf{I}_2 + \frac{\rho}{4}\mathbf{\Gamma}_{2N}\mathbf{\Gamma}_{2N}^{\dagger}\right) \tag{33}$$

where $\mathbf{\Gamma}_{2N} = [\Gamma_1^T \cdots \Gamma_N^T]^T$, with Γ_n defined from

$$\Delta_n = \Gamma_n \Gamma_n^{\dagger}$$

and where Δ_n is defined similarly to Δ_1 in (23) for the nth (as opposed to the first) receiver antenna.

7.1.2. Numerical Examples. Figures 4 and 5 show 10% outage capacities that were numerically evaluated for some schemes based on the capacity expressions that were presented above. All expressions were run over an ensemble of 10^4 (M, N) random Rayleigh-faded channel matrices (each entry of the matrix is chosen independently from any other entry from a complex i.i.d. Gaussian distribution of unit variance). The 10% outage value was then selected from the corresponding point of the CDF.

Figure 4 shows the 10% outage capacities for several $(M, 1)$ cases, as well as for the (2, 2) case. In the (2, 1) case, the plotted capacity corresponds both to the Alamouti scheme and to the maximum open-loop capacity, as indicated by Eq. (20). For the other $(M, 1)$ cases, we plot the capacity upper bounds corresponding to Eq. (12), and we use Eq. (13) for the asymptotic $(\infty, 1)$ case. We also use Eq. (31) with $N = 2$ for the capacity of a (2, 2) combined Alamouti/receive diversity scheme, and the log-det expression (5) for the (2, 2) maximum open-loop capacity. We observe that, at 10 dB, *the* $(2, 1)$ *system almost doubles the capacity of the* $(1, 1)$ *system*! However, as noted earlier, further increasing the number of transmit antennas in the $(M, 1)$ case offers diminishing returns. It is also worth noting that the (2, 2) combined transmit/receiver diversity scheme is capable of attaining

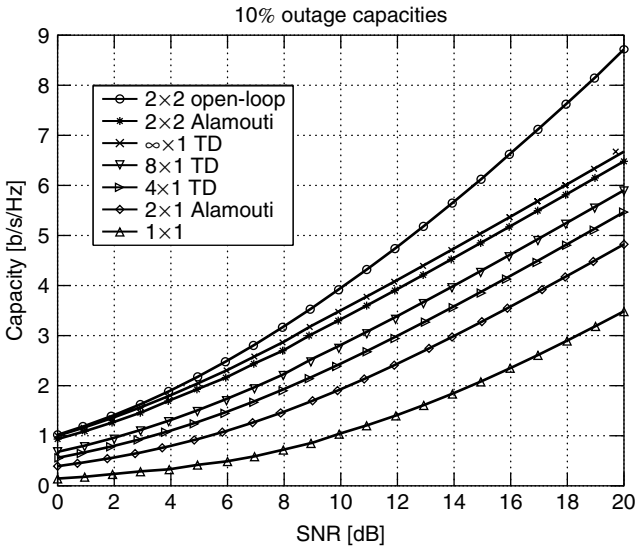


Figure 4. Outage capacities and bounds of $(M, 1)$ and $(M, 2)$ schemes.

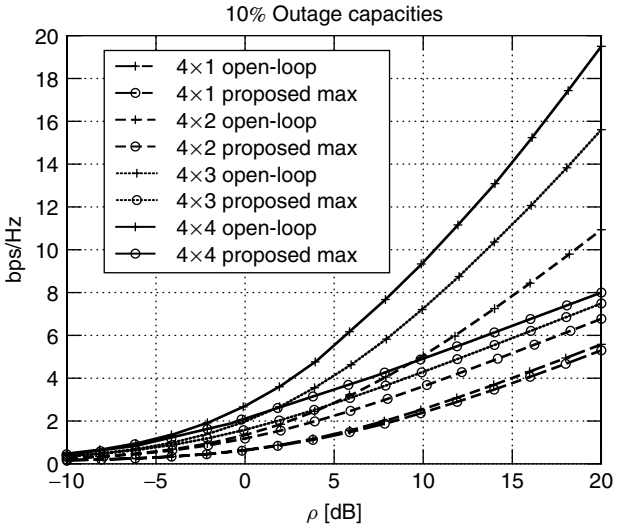


Figure 5. Outage capacities of the (4, 1) scheme described in Ref. 9, when used with up to four receiver antennas.

a quite significant fraction (particularly at low SNRs) of the maximum (2, 2) open-loop capacity. Finally, it is also interesting to note that a (2, 2) system achieves about the same capacity as an open-loop $(\infty, 1)$ system, which conveys again the message of the high value of adding extra antennas at the receiver.

In Fig. 5, we show the capacities of some combined transmit/receive diversity schemes for different $(4, N)$ cases. The circles represent the combined $(4, N)$ systems corresponding to the (4, 1) scheme of [9], in conjunction with optimal receiver diversity. When read from the bottom up, these four curves correspond to $N = 1, 2, 3, 4$, respectively. Similarly, the crosses represent the corresponding open-loop $(4, N)$ capacities. Notice that the proposed (4, 1) scheme is very close to the open-loop capacity; however, the gap gets increasingly larger as N grows from 1 to 4. In the (4, 2) case though, the scheme still performs well, particularly at low SNRs.

7.2. V-BLAST

A quite simple, from the transmitter's point of view, space–time transmission scheme was proposed in [10], and it is widely referred to as "V-BLAST," which stands for *Vertical Bell labs LAyered Space–Time*. In this architecture, $\{\tilde{b}(i)\}$ is first demultiplexed into M substreams, which are then encoded independently and mapped each on a different antenna:

$$s_m(k) = b_m(k), \qquad m = 1, \ldots, M$$

In other words, the original bit stream is converted into a vertical vector of encoded substreams (whence the term "vertical" BLAST), which are then streamed to the antennas through a 1–1 mapping. In [10], it was proposed to process the received signal with the use of a successive interference canceller. After determining the order into which the M substreams will be detected, the V-BLAST

receiver operates according to the following generic three-stage scheme, which is performed in a successive fashion for each substream:

1. Project away from the remaining interfering substreams.
2. Detect (after decoding, deinterleaving, and slicing) the substream.
3. Cancel the effect of the detected substream from subsequent substreams.

Mathematically, these operations can be described as follows for the k_mth substream:

$$\begin{aligned} z_{k_m}(k) &= W_{k_m}^{\dagger}\mathbf{x}^m(k) \\ \hat{z}_{k_m}(i) &= \mathrm{dec}(z_{k_m}(k)) \\ \mathbf{x}^{m+1}(k) &= \mathbf{x}^m(k) - \mathrm{enc}(\hat{z}_{k_m}(i))\mathbf{h}_{k_m} \end{aligned} \tag{34}$$

where $\mathbf{x}^1(k) = \mathbf{x}(k)$, $\{k_1, \ldots, k_M\}$ is a reordered version of the set $\{1, \ldots, M\}$ that determines the order in which the substreams will be detected, $\mathrm{dec}(\cdot)$ represents the decoding+detection operation, and $\mathrm{enc}(\cdot)$ represents the encoding operation. Finally, W_{k_m} represents the $N \times 1$ vector that operates on $\mathbf{x}^m(k)$ in order to project away from substreams $\{k_{m+1}, \ldots, k_M\}$. The operations in Eq. (34) are performed successively for $m = 1, \ldots, M$, after the ordering $\{k_1, \ldots, k_M\}$ has been determined.

We now discern between the following two cases for this linear operation, since they affect significantly the constrained capacity of the system:

7.2.1. Zero-Forcing Projection. In this case, at the mth stage, $W_{k_m}^{\dagger}$ nulls perfectly the interference from all the remaining (undetected) substreams. These are the substreams with indices $\{k_{m+1}, \ldots, k_M\}$. This nulling is represented mathematically as

$$W_{\mathrm{zf},k_m}^{\dagger}\mathbf{H} = [0 \cdots 0\ 1\ 0 \cdots 0] = \delta_{k_m}^T \tag{35}$$

where the unique non-zero element of the $1 \times M$ vector δ_{k_m} is in its k_mth position. As a result, the end-to-end model for the k_mth output is

$$d_{k_m}(k) = b_{k_m}(k) + W_{\mathrm{zf},k_m}^{\dagger}\mathbf{n}(k), m = 1, \ldots, M \tag{36}$$

where $\mathbf{n}(k) = [n_1(k) \cdots n_N(k)]^T$ is the receiver noise. Defining $\mathbf{d}(k) = [d_{k_1}(k) \cdots d_{k_M}(k)]^T$ and $\mathbf{b}(k) = [b_{k_1}(k) \cdots b_{k_M}(k)]^T$, (36) can be written in matrix form as

$$\mathbf{d}(k) = \mathbf{b}(k) + \mathbf{W}_{\mathrm{zf}}^{\dagger}\mathbf{n}(k) \tag{37}$$

where $\mathbf{W}_{\mathrm{zf}} = [W_{\mathrm{zf},k_1} \cdots W_{\mathrm{zf},k_M}]$. From (37), we observe that the ZF version of the V-BLAST superstructure is

- *Decomposable* — its capacity can be evaluated by computing separately each capacity of its M substreams.
- *Unbalanced* — all the substreams have different capacities, reflecting the fact that each "sees" a different SNR (this is also reflected in the fact that different columns of $\mathbf{W}_{\mathrm{zf}}$ have in general different square norms).

These attributes may be contrasted with the Alamouti (or STS) architecture mentioned in Section 6.1, which is both decomposable and balanced (each of its two substreams is capable of carrying exactly the same information rate). Regarding the capacity of the end-to-end system, it is important to emphasize that we have assumed that each substream is independently encoded, and that the transmitter has no way of knowing which is the maximum attainable rate for each antenna. As a result, it can at best transmit from all antennas the same rate. Hence, the capacity will equal M times the smallest of the M decomposed channel capacities:

$$C_{MN}^{\mathrm{VB-ZF}} = M \times \min_{m \in \{1,\ldots,M\}} \{\log_2(1 + \rho_{\mathrm{zf},k_m})\} \tag{38}$$

where ρ_{k_m} is the output SNR of the k_mth substream:

$$\rho_{\mathrm{zf},k_m} = \frac{\rho}{M|W_{\mathrm{zf},k_M}|^2} \tag{39}$$

It should finally be noted that the capacity in Eq. (38) can be optimized by choosing an optimal ordering for the set $\{k_1, \ldots, k_M\}$ [10].

7.2.2. MMSE Projection. In this case, at the mth stage, an optimal compromise between linear interference mitigation of the undetected substreams and noise amplification is sought. This is achieved through the following minimum mean squared error (MMSE) criterion:

$$\min_{W_{k_m}} E\|d_{k_m} - W_{k_m}^{\dagger}\mathbf{H}_{k_m}\|^2 \tag{40}$$

where $\mathbf{H}_{k_m}$ is derived from $\mathbf{H}$ by deleting its columns corresponding to indices $\{k_1, \ldots, k_{m-1}\}$. This gives for W_{k_m}:

$$W_{\mathrm{mmse},k_m}^{\dagger} = \left(\mathbf{H}_{k_m}\mathbf{H}_{k_m}^{\dagger} + \frac{M}{\rho}\mathbf{I}_N\right)^{-1}\mathbf{h}_{k_m} \tag{41}$$

where $\mathbf{h}_{k_m}$ is the k_mth column of $\mathbf{H}$. This end-to-end system has again been fully decomposed into four (1, 1) systems, which are, in general, not balanced (i.e., they do not have the same SINRs). Its capacity is computed again through the minimum of the four 1×1 capacities, and is given by a formula similar to (38):

$$C_{MN}^{\mathrm{VB-MMSE}} = M \times \min_{m \in \{1,\ldots,M\}} \{\log_2(1 + \rho_{\mathrm{mmse},k_m})\} \tag{42}$$

where now

$$\rho_{\mathrm{mmse},k_m} = \frac{\|W_{\mathrm{mmse},k_m}^{\dagger}\mathbf{H}_{k_m}\|^2}{M\|W_{\mathrm{mmse},k_m}\|^2/\rho + \sum_{l \neq k_m}\|W_{\mathrm{mmse},l}\|^2} \tag{43}$$

Again, the capacity in Eq. (42) can be maximized through optimal ordering.

7.2.3. Closed-Loop V-BLAST Operation. The fact that in both (38) and (42) the system capacity is a multiple of the

weakest rate is a direct result of the absence of knowledge of these rates at the transmitter. Had we assumed that, indeed, all M rates were known at the transmitter, the expressions (38) and (42) would use the sum of capacities as opposed to M times the minimum of capacities. A quite astonishing result that was reported in [12] (based on previous work in [13]) is that the sum of MMSE capacities in (42) equals the open-loop capacity of the MIMO channel in (5)! In other words

$$\sum_{m\in\{1,\ldots,M\}} \{\log_2(1+\rho_{\mathrm{mmse},k_m})\} = \log_2\left\{\det\left(I_N + \frac{\rho}{M}\mathbf{H}\mathbf{H}^\dagger\right)\right\} \tag{44}$$

This means, that *if the maximum attainable rates are known at the V-BLAST transmitter, then the system can attain the open-loop capacity with the use of linear MMSE* (+*subtractions*) *processing*. Moreover, the result holds irrespective of the ordering of the substreams. This is a quite nice tradeoff: a partially closed-loop technique (the receiver only feeds back to the transmitter a set of rates) allows to attain the system's open-loop capacity!

7.2.4. Numerical Results. In Fig. 6, we show a capacity CDF, at 10 dB SNR, of the ZF and MMSE V-BLAST architectures described above for the (4, 4) case. Notice that, at this SNR, the MMSE architecture is capable of attaining about 70% of the total open-loop capacity at 10% outage. However, the ZF architecture performs poorly, and it is even outperformed by a (1, 4) maximal ratio combining system at outages higher than 20%! The situation is more severe for lower SNRs such as 0 dB, as shown in Fig. 7. Now the V-BLAST MMSE architecture attains only about 50% of the (4, 4) open-loop capacity, whereas the ZF architecture is outperformed by the (1, 4) system across the board.

7.3. Other (M, N) Schemes

Similar to the (M, 1) case, several other schemes have been proposed in the literature for the general (M, N) case. For example, the use of a block space–time multiplexing whose

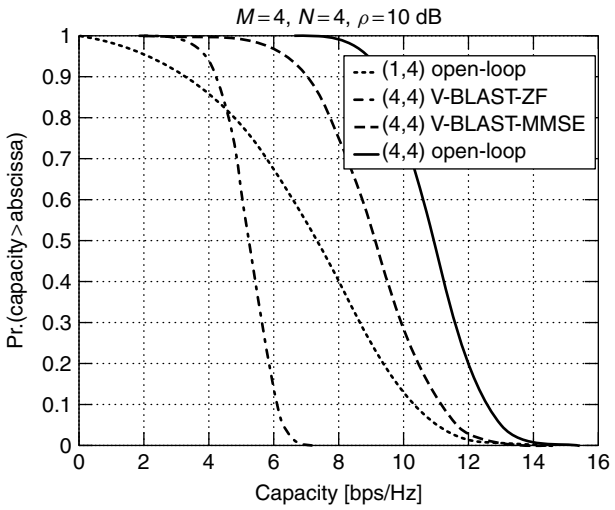


Figure 6. Outage capacity distribution of a V-BLAST MMSE architecture at 10 dB SNR.

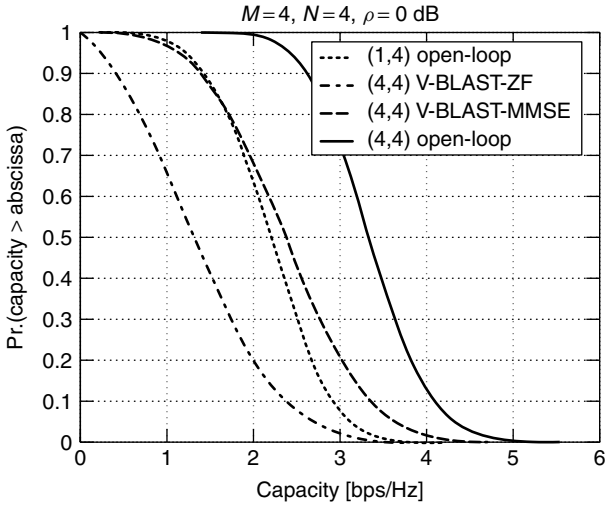


Figure 7. Outage capacity distribution of a V-BLAST MMSE architecture at 0 dB SNR.

mixing coefficients are optimized numerically according to a maximum *average capacity* criterion has been suggested in [14]. Another approach in [15] uses Turbo codes in the following way. The original substream is first demultiplexed into M substreams, which are separately encoded, each with a block code. Then, the M encoded outputs are space–time-interleaved in a random fashion, mapped onto constellation symbols, and sent out of the M antennas. At the receiver, the M substreams are separated through an iterative interference canceler, which uses MMSE for the linear (soft) part, and subtracts decisions made after (joint) deinterleaving and (separate) decoding of each interfering substream in the cancellation part.

These approaches have demonstrated encouraging performance in terms of bit/frame error rate at the receiver. However, their inherent capacity penalties are still unknown, mainly because of their apparent lack of structure and other properties such as the ones discussed above. The quantification of the capacity penalties of these and other emerging space–time transmission techniques remains an interesting open question.

8. CONCLUSIONS

In this article, we have taken a capacity view of multiple antenna wireless systems. We have outlined the fundamental channel capacity formulas that govern the spectral efficiencies of such multiple-input/multiple-output (MIMO) systems. We then described how these expressions reduce in a number of special cases. This has allowed us to draw interesting interpretations regarding the potential of different existing techniques to approximate the capacities promised by the formulas. Moreover, we believe that the capacity view has shed some new light on understanding the value of more conventional antenna combining techniques. Besides helping assessing the value of existing techniques, we believe that these results can be used to identify directions for future research in the field of MIMO systems.

Acknowledgments
The author would like to thank Dr. G. J. Foschini for many helpful and exciting discussions on the topic of MIMO systems, as well as his many colleagues from Bell Labs' Wireless Research Lab for numerous fruitful interactions on the topic.

BIOGRAPHY

Constantinos Papadias was born in Athens, Greece, in 1969. He received the diploma of electrical engineering from the National Technical University of Athens (NTUA), Greece, in 1991 and a Ph.D. degree in signal processing (highest honors) from the Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1995. From 1992 to 1995 he was a teaching and research assistant at the Mobile Communications Department, Eurécom, France. In 1995, he joined the Information Systems Laboratory, Stanford University, California, as a Postdoctoral researcher, working in the Smart Antennas Research Group. In November 1997 he joined the Wireless Research Laboratory of Bell Labs, Lucent Technologies, Holmdel, New Jersey, as a member of the technical staff. He is now a technical manager in Bell Laboratories Global Wireless Systems Research Department. His current research interests lie in the areas of multiple antenna systems (e.g., MIMO transceiver design and space-time coding), interference mitigation techniques, reconfigurable wireless networks, as well as financial evaluation of wireless technologies. He has authored several papers and patents on these topics. Dr. Papadias is a member of IEEE and a member of the Technical Chamber of Greece.

BIBLIOGRAPHY

1. A. Paulraj and C. Papadias, Space-time processing for wireless communications, *IEEE Signal Process. Mag.* **14**(6): 49–83 (Nov. 1997).
2. G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* **1**(2): 41–59 (1996).
3. E. Telatar, *Capacity of Multi-antenna Gaussian Channels*, AT & T Bell Laboratories Technical Memorandum, June 1995.
4. G. Foschini and M. Gans, On limits of wireless communications in a fading environment when using multiple antennas, *Wireless Pers. Commun.* **6**(6): 315–335 (1998).
5. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**(2): 744–765 (March 1998).
6. C. Papadias, On the spectral efficiency of space-time spreading schemes for multiple antenna CDMA systems, *33rd Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 24–27, 1999, pp. 639–643.
7. S. Sandhu and A. Paulraj, Space-time block codes: A capacity perspective, *IEEE Commun. Lett.* **4**(12): 384–386 (Dec. 2000).
8. S. Alamouti, A simple transmitter diversity scheme for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (Oct. 1998).
9. C. Papadias and G. J. Foschini, A space-time coding approach for systems employing four transmit antennas, *Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP 2001*), Salt Lake City, UT, May 7–11, 2001.
10. G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, Simplified processing for wireless communication at high spectral efficiency, *IEEE J. Select. Areas Commun.* **17**(11): 1841–1852 (Nov. 1999).
11. B. Hochwald, L. Marzetta, and C. Papadias, A transmitter diversity scheme for wideband CDMA systems based on space-time spreading, *IEEE J. Select. Areas Commun.* **19**(1): 48–60 (Jan. 2001).
12. S. T. Chung and A. Lozano, and H. C. Huang, Approaching eigenmode BLAST channel capacity using V-BLAST with rate and power feedback, *Vehicular Technology Conf.* (*VTC*) *Fall 2001*, Atlantic City, NJ, Oct. 2001.
13. M. K. Varanassi and T. Guess, Optimum decision-feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel, *1998 Asilomar Conf. Signals, Systems, and Computers*, 1998, pp. 1405–1409.
14. B. Hassibi and B. Hochwald, High-rate linear space-time codes, *Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP 2001*), Salt Lake City, UT, May 7–11, 2001.
15. M. Sellathurai and S. Haykin, Joint beamformer estimation and co-antenna interference cancelation for Turbo-BLAST, *Int. Conf. Acoustics, Speech, and Signal Processing* (*ICASSP 2001*), Salt Lake City, UT, May 7–11, 2001.

MULTIPROTOCOL LABEL SWITCHING (MPLS)

Pim Van Heuven
Steven Van Den Berghe
Filip De Turck
Piet Demeester
Ghent University
Ghent, Belgium

1. INTRODUCTION

The Internet Protocol (IP) is a connectionless networking layer protocol to route IP packets over a network of IP routers. Every router in the IP network examines the IP packet header and independently determines the next hop based on its internal routing table. A routing table contains information about the next hop and the outgoing interface for the destination address of each IP address. Multiprotocol label switching (MPLS) allows the setup of label-switched paths (LSPs) between IP routers to avoid the IP forwarding in the intermediate routers. IP packets are tagged with labels. The initial goal of label-based switching was to increase the throughput of IP packet forwarding. Label-based switching methods allow routers to make forwarding decisions based on the contents of a simple label, rather than by performing a complex route lookup according to the destination IP address. This initial justification for MPLS is no longer perceived as the main benefit, since nowadays routers are able to perform route lookups at sufficiently high speeds to support most interface types. However, MPLS brings many other benefits to IP-based networks, including

(1) traffic engineering, that is, the optimization of traffic handling in networks; (2) virtual private networks (VPNs), networks that offer private communication over the public Internet using secure links; and (3) the elimination of multiple protocol layers.

MPLS paths are constructed by installing label state in the subsequent routers of the path. Labels are fixed-length entities that have only a local meaning. Labels are installed with a label distribution protocol. MPLS forwarding of the IP packets is based on these labels. The IP and MPLS forwarding principles will be detailed first, followed by a description of the MPLS label distribution process.

1.1. Forwarding in IP and MPLS

In regular non-MPLS IP networks, packets are forwarded in a hop-by-hop manner. This means that the forwarding decision of a packet traversing the network is based on the lookup of the destination in the local routing table [also called *routing information base* (RIB)]. Figure 1a illustrates IP for a network consisting of four routers: nodes A, B, C and D. A simplified IP routing table of router B is shown. It consists of entries that map the destination network addresses of the IP packets to the IP addresses of the next hop and the router interface, which is connected to the next hop. When forwarding a packet, a router inspects the destination address of the packet (found in the IP header), searches through his local router table via a longest prefix match, and forwards it to the next hop on the outgoing interface.

The destination addresses in this table are aggregated in order to reduce the number of entries in this table. These entries are aggregated by indicating the length of the significant part of the destination addresses (from 0 to 32 bits). If n is the length of address a, then only the first n (most significant) bits of a are considered. The resulting partial address is called a *prefix* and is noted as a/n (e.g., 10.15.16.0/24). This aggregation of addresses has the drawback that searching through the table needs to be done with a *longest-prefix* match. A longest-prefix match is more complex than an exact match because the result of the search must be the entry with the longest prefix that matches the address [1].

An important characteristic of IP forwarding is that packets arriving at a router with the same destination prefix are forwarded equivalently over the network. A class of packets that can be forwarded equivalently is a forwarding equivalence class (FEC). Because of the destination-based forwarding of IP, FECs are usually associated with IP prefixes. The forwarding in an IP router can be restated as the partitioning of packets in FECs and assigning a next hop to the FECs. It is important to note that determination of the FEC needs to be done in every hop for every packet.

On the other hand, MPLS forwarding relies on labels, instead of prefixes to route packets through the network [2,3]. Labels are fixed-length entities that have only a local meaning. Because a label has only a local meaning, labels can be different at every hop and therefore must be adapted before forwarding the packet; this process is called *label switching*. The labels are distributed over the MPLS domain by means of a label distribution protocol. MPLS routers are called *label-switching routers* (LSR) (Fig. 1c) because they operate on labels rather than on IP prefixes when forwarding packets. The concatenation of these installed labels in the different LSRs is called a *label-switched patch* (LSP). An LSP is set up between the ingress LSR and the egress LSR; these edge LSRs are also called *label edge routers* (LERs). Packets belonging to a certain FEC are then mapped on an LSP. Determining the FEC of a packet is necessary only in the ingress of the LSP. The segregation of packets in FECs needs be done only once, in the ingress router, and this segregation can also be based on more than the destination prefix of the packet. For example, it is possible to take both the source and the destination into account. LSPs are unidirectional paths because they are based on FEC-to-label bindings.

In the case of label-switched routers, every router contains two tables: an incoming label map (ILM) table that contains all the incoming labels the router has allocated and a table that contains all the necessary information to forward a packet over an LSP (Fig. 1b). The latter table is populated with next-hop label-forwarding entries (NHLFE). There is a mapping between the ILM and an NHLFE mapping the incoming labels to an output label, the outgoing interfaces, and the next hop. The router inspects the incoming label and consults the ILM table to find the right NHLFE. This NHLFE contains the outgoing label, the next hop, and the outgoing interface. Before the packet is sent to the next hop, the label is switched to the outgoing label value.

1.2. LSP Setup

Two distinct cases can be distinguished: (1) hop-by-hop routed LSP setup and (2) explicit routed LSP setup. These two cases will be described in the following sections.

1.2.1. Hop-by-Hop Routed LSP Setup. To distribute labels over the network and consequently set up an LSP, a *label distribution protocol* is used. Path setup typically consists of two steps: (1) a request is sent to the egress of the LSP and (2) the response propagates back to the ingress. The first step is denoted by the generic term "label request," whereas the second step is denoted by the term "label mapping." Figure 2a illustrates the label distribution process. When LER A wants to set up an LSP to network *netD*, it will send a label request to its next hop toward netD (step a, Fig. 2). The intermediate nodes from the ingress towards the egress (like LSR B) will install state about the request and will forward the request toward netD according to their routing information bases (step b). When the request reaches the destination of the LSP, the egress node will allocate a label for this LSP and will store this information in the incoming label map (ILM). The LSR will then send a label mapping back to the previous hop. The "label mapping" message contains the label previously allocated by the LSR (step d). LER B will then receive the label mapping from node D. The label contained in the label mapping will be used to make a next-hop label-forwarding entry (NHLFE). Router B will then, in turn, allocate a label and store this label in its ILM. The

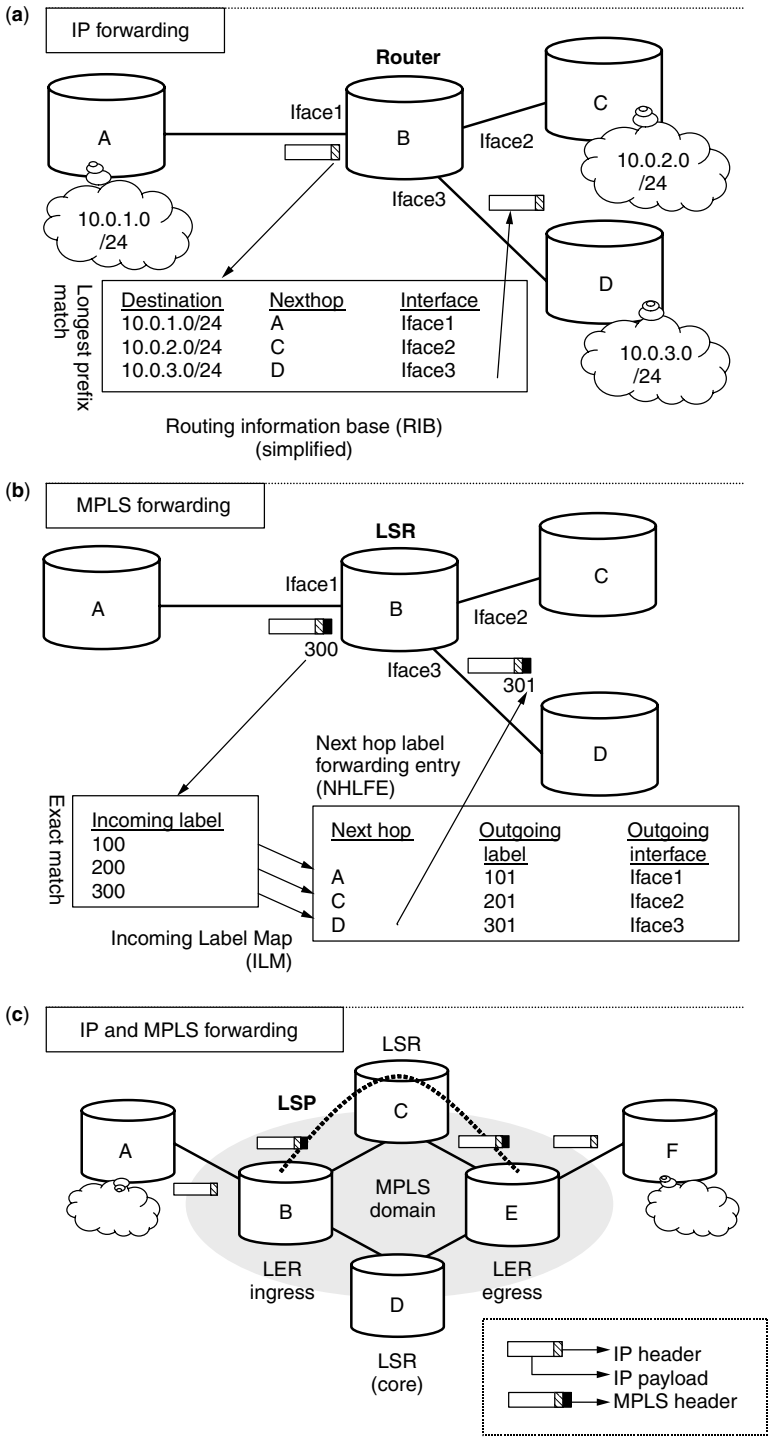


Figure 1. (**a**) IP forwarding for a network consisting of four nodes; (**b**) example of MPLS forwarding; (**c**) IP and MPLS forwarding.

information in the ILM (incoming label) and that in the NHLFE (outgoing label) are combined, effectively storing the information about the label switch (step *e*). After allocating the label and storing the relevant information, LSR *B* will send a label mapping to its previous hop (step *f*). Finally, the initiator of the LSP setup (node *A*) will receive the label mapping from its next hop. LSR *A* will store this information in a NHLFE. This ingress LER will then map traffic to the newly established LSP by mapping a class of packets (FEC) to the LSP, which implies that traffic that belongs to this traffic class will be forwarded over the LSP. The FEC is thus mapped on the NHLFE (step *g*). All the FEC-to-NHLFE mappings are stored in the FEC-to-NHLFE map (FTN). The FTN is used by the ingress of an LSP to forward the packets belonging to a certain FEC over the LSP (to find the outgoing label).

Because the request is forwarded according to the local RIB of the intermediate routers, the resulting LSP is called a *hop-by-hop routed LSP*. Another type of LSP is called an *explicit routed LSP* (ER-LSP).

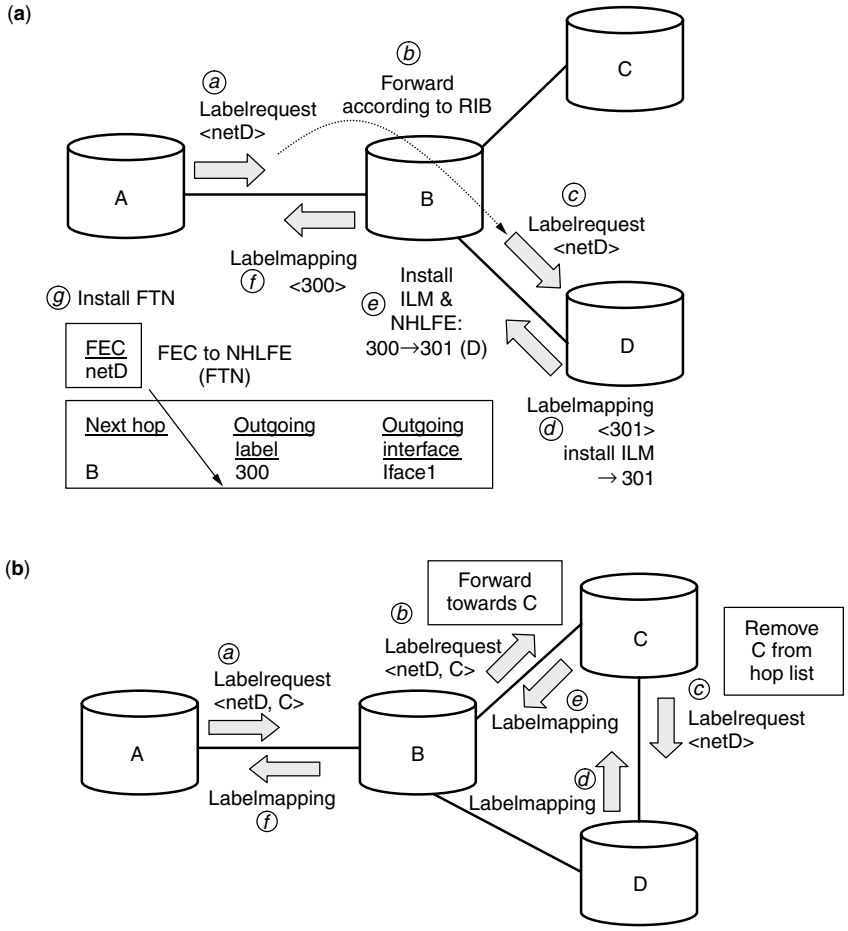


Figure 2. Examples of (**a**) hop-by-hop routed label distribution and (**b**) explicitly routed label distribution.

1.2.2. Explicit Routed LSP Setup. The real power of MPLS lies in the fact that paths can be set up with a great deal of flexibility. An example is an explicit routed LSP (ER-LSP). The term *explicit routed* means that some or all of the hops between the ingress and the egress of the LSP can be specified. Figure 2b illustrates this; in step *a* LSR *A* sends a label request for *netD* and the label request explicitly states that the LSP should be routed along node *C*. Node *B* will receive this label request and will forward it toward *C* along the shortest path (step *b*). When LSR *C* receives this label request, it removes itself from the hop list in the label request and forwards the label request toward the destination. From then on the LSP setup continues as detailed in Fig. 2. It is important to note that every node keeps state about the label request so that the label mappings are sent to the correct previous hop, that is, the hop it received the corresponding label request from.

1.3. Conclusion

In IP both the route calculation and the forwarding is based on the destination address. MPLS separates the forwarding from the route calculation by using labels to forward the packets. To distribute these labels over the domain and hence set up an LSP, MPLS uses a label distribution protocol. LSPs can be setup according to the IP routing tables or the hops to be traversed can be explicitly specified.

2. ARCHITECTURE

After the general overview of MPLS versus IP, this section describes MPLS architecture in greater detail [4].

MPLS architecture consists of a forwarding layer and a signaling layer. The functionality of the forwarding layer is to inspect the incoming label, look up the outgoing label(s), and forward the packet with the new label(s) to the correct outgoing interface (see Section 2.1). The signaling layer is responsible for setup of the MPLS paths (LSPs). The protocols responsible for the setup of LSPs are called *label distribution protocols* (see Section 3.2).

2.1. Forwarding Layer

This section begins with a summary of the most important MPLS forwarding concepts and then gives details on other important issues and terminology with respect to the MPLS forwarding.

2.1.1. MPLS Forwarding Concepts. MPLS architecture formalizes three concepts with respect to the forwarding plane:

1. The *next-hop label forwarding entry* (NHLFE) contains all the information needed in order to forward a packet in a MPLS router. It contains the packet's next hop and the outgoing label

operation. The NHLFE may also contain the data-link encapsulation and information on how to encode the label stack when transmitting the packet.

2. The *incoming label map* (ILM) defines a mapping between an incoming label and one or more NHLFEs. It is used when forwarding labeled packets. If the ILM maps a particular label to more than one NHLFE exactly, one NHLFE must be chosen before the packet is forwarded. Having the ILM map a label to more than one NHLFE can be useful to do, for instance, load balancing over a number of LSPs. Since the ILM is used to forward *labeled* packets in a LSR, it is typically used in a core LSR.
3. Finally, the *FEC-to-NHLFE map* (FTN) maps a FEC to one (or more) NHLFEs. It is used when forwarding packets that arrive unlabeled, but that are to be labeled before being forwarded. The FTN map is used in the ingress *label edge router*.

2.1.2. Label Encapsulation. It is apparent that "labels" constitute the center of MPLS architecture. However, the properties of the labels differ from the link layer on which MPLS is supported. Because of this close tie with the link-layer technology, MPLS is sometimes called a *layer 2.5 architecture*, situated between the link layer (layer 2) and the networking layer (layer 3). Two categories of data-link layers can be distinguished: (1) link layers that natively support fixed-length label entities and switch on them. Examples of [e.g., ATM — see Fig. 3c, virtual circuit identifier (VCI) or virtual path identifier/(VPI), (VPI/VCI) and frame relay — Fig. 3b, data-link circuit identifier (DLCI)] and (2) link-layer technologies that do not natively support labels but encapsulate the labels

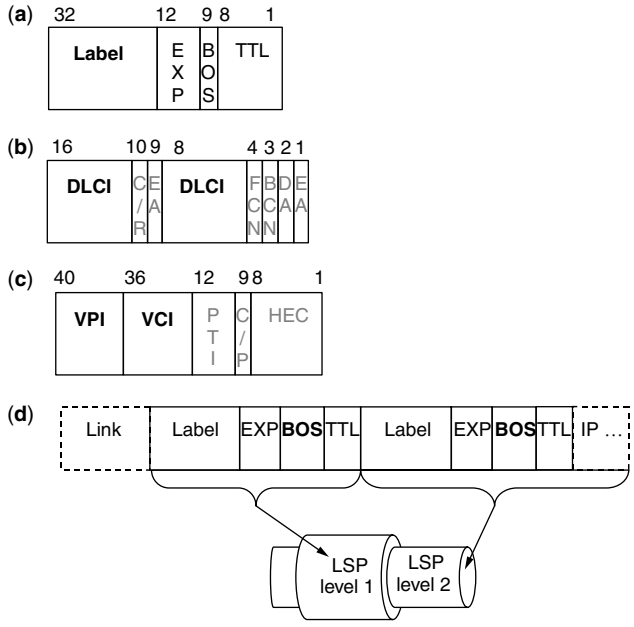


Figure 3. MPLS label encapsulation: (**a**) generic MPLS encapsulation with the shim header; (**b**) encapsulation in the frame relay DLCI; (**c**) encapsulation in the ATM VPI/VCI; (**d**) encapsulation of multiple labels with shim headers (label stacking).

by transmitting an additional header. This small header, called the *shim header*, is inserted between the link-layer header and the networking header. The former way of encapsulating the MPLS labels is called *link-layer-specific encapsulation*, whereas the latter is called *generic MPLS encapsulation*. The shim header contains a label, three experimental bits, a bottom-of-stack (BoS) indicator, and a TTL ("time to live") field (Fig. 3a). The *label* field (20 bits) is used to store the label value, the three experimental (*EXP*) can be used to support Diffserv over MPLS (Diffserv will be covered in detail in Section 2.3.2), and/or early congestion notification (ECN) or other experimental extensions to MPLS forwarding. The *BoS* bit is used to indicate the last shim header of the label stack. Finally, the *TTL* field is used to support the IP time to live mechanism (see Section 2.1.4).

An example of a link-layer technology that has a native label entity is ATM (asynchronous transfer mode). In ATM-based MPLS a label is a VPI, a VPI/VCI, or a VCI identifier. In ATM networks these identifiers are installed with user–network interface (UNI) or private network–node interface (PNNI) signaling. MPLS does not use ATM signaling because a label distribution protocol is used instead. When link-layer-specific label encapsulation is used, the label stack is still encoded in the shim header but the shim header cannot be processed by the intermediate LSRs (only at the ingress and the egress). Therefore the top label of the label stack is copied to the native label entity before the ATM or FR (Frame relay) segment. Similarly after the segment, the current label value from the native label is copied to the top label of the label stack.

2.1.3. Label Operations and Label Stacks. Only a limited number of operations are possible on MPLS labels: (1) replace the top label with a new label (label swap), (2) remove the top label (label pop), or (3) replace the top label and push a number of new labels. This means that multiple labels can be pushed on top of each other, leading to a label stack. This can be useful to aggregate multiple LSPs in one top-level LSP and hence reduce the LSP state. Label stacking is also use in MPLS VPNs (see Section 3.2).

Label operations are possible only on the top label of the label stack. In other words, "pop the label stack" means to remove the top label of the stack. Similarly at every node only the top label is considered when making the forwarding decision.

2.1.4. Support for the IP "Time to Live" (TTL) Field. In regular IP networks, the "time to live" (TTL) field in the IP header is decremented at every hop. When the TTL reaches zero, the packet is dropped. This mechanism is used to prevent packets from being forwarded forever in case of a network anomaly (e.g., a network loop). To support this mechanism in MPLS, the TTL information must be available to the LSR. Since the shim header contains a TTL field, the LSRs are able to decrement the TTL just as in regular IP forwarding.

When MPLS is supported by a link-layer technology that uses its own native label entities, the LSR can act only on these native labels. Unfortunately, these link-layer-specific MPLS labels do not have a TTL field. It

is therefore impossible to decrement the TTL at every hop. The solution is to compute the total TTL decrement for the non-TTL-capable segment (the switching segment) at LSP setup time. When a packet arrives at the first hop of this segment, the precomputed TTL decrement is subtracted from the current TTL. If the result is positive, then the TTL is written to the top entry of the label stack. A packet traveling through the LSP will have the correct TTL both before and after the non-TTL-capable segment. The TTL will have a constant value in the segment (the same value as immediately after the segment). If there is no information about the hop count of the non-TTL-capable segment, the current TTL is decreased by one and the resulting value is written to the label stack. In any case appropriate actions must be taken on this result (e.g., dropping the packet if the TTL reaches zero).

2.1.5. Label Merging. In order to reduce the number of outgoing labels, different incoming labels for the same FEC can use the same outgoing label (label merging).

A link layer is capable of merging labels if more than one incoming label can be mapped to a single outgoing label (all the incoming labels are merged to the same outgoing label). A merge-capable LSR needs to store only the label information of every FEC. However, some link-layer technologies are not capable of merging labels. In this case a one-to-one mapping between incoming and outgoing labels is required. This has scalability implications because an individual LSP state is needed for every ingress–egress pair traversing a given LSR (unless label stacking is be used).

2.1.6. Label Spaces. Another point of difference between some link-layer technologies is the scope of the labels. The MPLS architecture uses the term *label space*. Label spaces limit the scope of a label, which means that the label is valid and defined only for a given label space. Two labels match only if both the label value and the label space are equal. This has the consequence that the same label value can be reused in a different label space. ATM and frame relay have a label space per interface, which means that label A on interface 1 is will be interpreted differently from label A on interface 2. On the other hand, platformwide label spaces are not tied to a specific interface but are valid on the whole platform (i.e., router or host). Global label spaces have the advantage that if the incoming interface of an LSP changes (e.g., during rerouting), no action needs to be taken. Per interface label spaces reduces the number of incoming labels per interface (which might be useful if labels are a scarce resource).

2.1.7. Penultimate-Hop Popping. Since no forwarding decisions have to be made at the last hop of an LSP, the next-to-the-last hop (also referred to as the *penultimate hop*) can also pop the top label of the LSP. The penultimate hop pops the label and sends the packet without a label to the egress node, thereby eliminating the label overhead.

2.2. Signaling Layer

MPLS architecture allows for multiple methods for distributing labels. The reader is referred to Fig. 2 for a basic overview of the label distribution process. The subsequent sections describe the most important label distribution protocols. Some important terminology will be introduced first.

1. *Downstream versus Upstream Allocation*. Since labels only have a local meaning, these labels can be allocated decentralized by the switch controllers. For a given LSP, a core LSR has two neighbors, one upstream and one downstream. So it needs an incoming label and an outgoing label. With that in mind, there are two possible approaches: (1) the LSR supplies its upstream neighbor with a label and receives a label from his downstream neighbor or (2) vice versa. When the LSR receives the label from its downstream neighbor, this is called *downstream allocation*. MPLS typically uses downstream allocation.

2. *Unsolicited Distribution versus Distribution on Demand*. As described in paragraph 1, labels are chosen (allocated) by the downstream LSRs. When these labels are distributed spontaneously (without request), this is called *unsolicited label distribution*. When the upstream LSR always sends a request to its downstream neighbor in order to obtain a label, this is called *distribution on demand*.

3. *Independent versus Ordered Control*. In *independent control*, when an LSR recognizes a particular FEC, it makes an independent decision to map a label to that FEC and to distribute that mapping. In *ordered control*, an LSR maps a label to a particular FEC only if it is the egress for that LSP or if it has already received a label binding from its next hop.

4. *Liberal Retention versus Conservative Retention*.

Consider the situation where an upstream LSR has received and retained a label mapping from its downstream peer. When the routing changes and the original downstream peer is no longer the next hop for the FEC, there are two possible actions the LSR can take: (**a**) release the label, an action called *conservative retention*; or (2) keep the label for later use, which is called *liberal retention*.

Conservative retention (*release on change*) has the advantage that it uses fewer labels; liberal retention (*no release on change*) has the advantage that it allows for faster reaction to routing changes.

5. *Label Use Method*. Labels can be used as soon as a LSR receives them (*use immediate*) or the LSR can only use a certain label if the corresponding LSP contains no loop (*use loop-free*). MPLS supports both loop prevention (preventing an LSP with a loop from being set up) and loop detection mechanisms (detecting whether an LSP contains a loop).

2.2.1. The Label Distribution Protocol (LDP). The *label distribution protocol* (LDP) is the basic signaling protocol proposed by the IETF MPLS Working Group for hop-by-hop routed LSPs [5]. In LDP labels are distributed for a given forward equivalent class (FEC). In LDP a FEC can be either an IP prefix or an IP host address. LDP gives the user a great deal of freedom in how to set up the LSPs. LDP peers use TCP as the transport protocol for LDP messages. This ensures that these

messages are reliably delivered and need not be refreshed periodically (LDP is therefore called a *hard-state protocol*). Session management allows LDP peers (most of the time neighbors) to discover each other and to negotiate about session parameters (e.g., on-demand or unsolicited distribution). After this negotiation phase, a LDP session is set up and the distribution of the labels can start. LSP supports unsolicited and on-demand distribution of labels, liberal and conservative label retention, independent and ordered control, and the immediate or loop-free use of labels (Fig. 2a illustrates ordered control, downstream on-demand label distribution).

Information in LDP messages is encapsulated in type–length–value (TLV) structures. These TLVs are used for standard features but can also be used to extend LDP with experimental and/or vendor-private mechanisms. The constraint-based label distribution protocol (CR-LDP) is an extension to LDP and will be covered in Section 2.2.2.

2.2.2. Constraint-Based Label Distribution Protocol (CR-LDP). CR-LDP introduces a number of extensions to LDP in order to support MPLS traffic engineering (TE). CR-LDP supports only the downstream on-demand ordered label distribution and conservative label retention mode. The additional functionality of CR-LDP compared to LDP is (1) the possibility to setup constraint-based LSPs, (2) the support for traffic parameters, (3) preemption, and (4) resource classes [6].

1. *Constraint-Based Routes*. CR-LDP allows setup of LSPs that defer from the shortest path by explicitly indicating the hops that the LSP should traverse. There's a distinction between strict and loose hops. A "strict" hop on an LSP means that the next hop along the LSP should be that hop and that no additional hops may be present. A "loose" hop on an LSP simply requires that the hop be present on the path that the LSP traverses. Hops are ordered, which means that they should be traversed in the order they are specified. CR-LDP supports the notion of abstract nodes. An *abstract node* is a group of nodes whose internal topology is opaque to the ingress node of the LSP. An abstract node can, for example, be denoted by an IPv4 prefix, an IPv6 prefix, or an autonomous system (AS) number.

2. *Traffic Parameters*. The traffic parameters of an LSP are modeled with a peak and a committed rate. The *peak rate* is the maximum rate at which traffic should be sent over the LSP. The peak rate of an LSP is specified in terms of a token bucket with a rate and a maximum token bucket size. The committed rate is the rate that the MPLS domain commits to the LSP. The committed rate of an LSP is also specified in terms of a token bucket. The extent by which the offered rate exceeds the committed rate may be measured in terms of another token bucket that also operates at the committed rate, but the maximum size of the this token bucket is the maximum excess burst size. There can also be a weight associated with the LSP; this weight indicates the relative share of the available bandwidth the excess bandwidth of an LSP receives. Finally, the frequency of an LSP indicates the granularity at which the committed rate is made available. CR-LDP also provides support for DiffServ over MPLS, as will be described in Section 2.3.2.

3. *Preemption*. An LSP can have two priorities associated with it: a setup priority and a hold priority. The *setup* priority indicates the relative priority an LSP has to use resources (bandwidth) when set up. A LSP with a higher setup priority can be set up in favor of an existing LSP with a lower priority (it can preempt an existing LSP). The *holding* priority indicates how likely it is for an LSP to keep its resources after having been set up.

4. *Resource Classes*. In CR-LDP one can specify which of the resource classes an LSP can traverse. A resource class is usually associated with a link, enabling one to indicate which links are acceptable to be traversed by an LSP. Effectively, this information allows for the network's topology to be pruned; thus, certain links cannot be traversed by the LSP. For example, a provider might want to prevent continental traffic from traversing transcontinental links.

2.2.3. Extensions to RSVP for LSP Tunnels (RSVP-TE). The "extensions to RSVP for LSP tunnels (RSVP-TE)" protocol is an extension of the "resource reservation protocol" [7], a signaling protocol originally developed for Intserv reservations. RSVP-TE First extends RSVP with the possibility to set up LSPs and then adds traffic engineering functionality [8].

2.2.3.1. Setting up Paths with RSVP-TE. RSVP-TE is based on the RSVP protocol, which does not have support to set up LSPs. RSVP-TE has been extended to support this by introducing a new LSP session type and then defining two new objects: (1) a label request object, which is encapsulated in the downstream direction on the RSVP PATH messages; and (2) a label object, which is encapsulated in the upstream direction on the RSVP RESV messages. Labels are allocated in the upstream direction by the downstream nodes. In other words, RSVP implements a downstream on-demand label distribution protocol. RSVP does not have direct support to detect the failure of a neighboring node. To address this, the Hello protocol has been developed. This protocol allows RSVP to detect the liveliness of its neighbors. RSVP also has a loop detection protocol to prevent setting up an LSP that contain loops. This makes RSVP more or less, functionalitywise, equivalent to LDP in downstream on-demand mode, with the important difference that RSVP is a soft-state protocol. This means that the state with respect to the LSP has to be refreshed periodically in the network. The advantage of a soft-state protocol is that the protocol responds more naturally to network changes while it typically requires more signaling overhead.

2.2.3.2. Traffic Engineering with RSVP-TE. Other extensions to RSVP introduce the traffic engineering capabilities very similar to CR-LDP's functionality. Like CR-LDP, RSVP-TE supports the notion of explicitly routed paths whereby the (abstract) hops can be specified strict or loose. The approach to bandwidth and resource allocation differs fundamentally from the CR-LDP model. As mentioned before, RSVP is a signaling protocol for IntServ,

so support for IntServ in RSVP-TE is naturally inherited from the base RSVP protocol. As in CR-LDP, it is possible to indicate the setup and holding priority of the LSP. The resource class procedures for RSVP-TE are more powerful than those found in CR-LDP. In CR-LDP a link is eligible to be traversed by an LSP if the resource class of the link is part of the resource classes specified in the label request message. This leads to the procedure where any link can be used as long as the link is part of the resource class collection specified in the label request message. RSVP-TE also supports this include-any relationship between links and LSPs, but it also supports exclude-any and include-all relationships. It is not necessary to specify any of three relationships, but if set, they must match for the link to be taken into account.

2.2.4. Carrying Label Information in Border Gateway Protocol 4 (BGP4). The Border Gateway Protocol 4 (BGP4) is used to distribute routes across the Internet. These routes can be interdomain routes, making BGP the sole interdomain routing protocol. By piggybacking label information on the BGP route UPDATE messages, BGP can be used to distribute the label mapped to that route. A simple example of the use of BGP as a label distribution protocol is when two BGP peers are directly connected, in which case BGP can be used to distribute labels between them. A more important use of BGP as a label distribution protocol is the more common case where the BGP peers are not directly connected but belong to an MPLS domain that supports another label distribution protocol (e.g., LDP). BGP4 and another label distribution protocol is used to administer MPLS VPNs (see Section 3.2).

2.3. MPLS and Quality of Service

Although MPLS is not a quality-of-service (QoS) framework, it supports delivery of QoS. The following sections describe how the two major models for QoS in IP (IntServ and DiffServ) are implemented with MPLS [9].

2.3.1. Integrated Services. Integrated Services (IntServ) architecture has the goal to provide end-to-end QoS (in the form of services) to applications. The IntServ QoS model has defined two service types: *guaranteed service* (guaranteed delay and bandwidth) and *controlled load* (QoS closely approximating that of an unloaded network). The architecture uses an explicit setup mechanism to reserve resources in routers so that they can provide requested services to certain flows. RSVP is an example of such a setup mechanism, but the IntServ architecture can accommodate other mechanisms. RSVP-TE as an extension of RSVP has natural support for both IntServ service types. An LSP with IntServ reservation is created just like any other IntServ reservation but additionally the MPLS specific LABEL_REQUEST and LABEL objects are piggybacked on the PATH and RESV message, respectively.

CR-LDP does not have support for IntServ natively, but it can support (a number of) IntServ flows over an LSP by setting the appropriate traffic parameters of the LSP. In order to guarantee the service received on the LSP, admission control and policing on the ingress is required.

2.3.2. Differentiated Services. In order to solve the IntServ scalability problem, Differentiated Services (DiffServ) classifies packets into a limited number of *classes* and therefore does not need for per flow state or per flow processing. The identified traffic is assigned a value, a DiffServ code point (DSCP). A DiffServ *behavior aggregate* (BA) is a collection of packets with the same DiffServ codepoint (DSCP) crossing a link in a particular direction. A per hop behavior (PHB), the externally observable forwarding behavior, is applied to a behavior aggregate.

The classification is usually based on multiple fields in the IP header (multifield, MF classification) at the edge and on the DiffServ codepoint (behavior aggregate, BA classification) in the core of the network (see Fig. 4c). An example PHB is *expedited forwarding* (EF), which offers low loss, low delay, and low jitter with an assured bandwidth. This means that the collection of packets marked with the EF codepoint traversing a link in a certain direction (BA) will receive low loss, delay, jitter, and an assured bandwidth. The *assured forwarding* (AF) PHB group is a group of PHB. A PHB of the AF group is denoted as *AFxy*, where x is the class and y is the drop precedence. Packets belonging to a different AF class are forwarded separately. Usually more resources are allocated to the lower classes. Packets within a class that have a higher drop precedence will be dropped before packets with a lower drop precedence.

An *ordered aggregate* (OA) is the set of behavior aggregates that share an ordering constraint. This means that packets that belong to the same OA must not be reordered. When looking at DiffServ over MPLS, it immediately becomes apparent that packets that belong to a certain OA must be mapped on the same LSP; otherwise this ordering constraint cannot be enforced. This is trivial if only one PHB is applied to the ordered aggregate. However, PHBs can be grouped in a per hop behavior scheduling class (PSC). A PSC is the set of one or more PHB(s) that are applied to a given OA. For example, AF1y is a PSC comprising the AF11, AF12, and AF13 PHBs. Combining the notion of OA and PSC means that in DiffServ over MPLS, OA-PSC pairs will be mapped on LSPs. If the PSC contains more than one PHB, this means that it must be possible for an LSR to enforce different PHBs to the packets that belong to the same LSP. This, in turn, means that the LSR must have some information in the packet header to determine the PHB to be applied. This information must be encapsulated in a way that is accessible to the LSR and thus must be part of the label or shim header. We will now discuss how to encapsulate the PHB.

2.3.2.1. Encapsulating the PHB. In IPv4 the "type of service" (ToS) field or in IPv6 "traffic class" field is used to encapsulate the DSCP. With generic MPLS encapsulation there is a mapping from the IP DSCP space to the EXP field of the shim header [10]. The DSCP field uses the 6 most significant bits of the 8 bits of these IP header fields. Since the DSCP field is 6 bits wide, it can represent 64 different values. However, the EXP field of the shim header is only 3 bits wide, so it can represent only 8 different values. This means that the mapping from DSCP to EXP value cannot be a one-to-one mapping. This is quite a problem because currently there are more than eight defined DSCP values

(best effort, 12 AF values, and EF). If the DiffServ domain uses less than 8 different DSCP values, then the mapping between DSCP and EXP can be fixed over the domain. If the domain uses more than eight different codepoints, then the mapping must be explicitly defined on a per LSP basis.

When the EXP value is used to indicate the set of PHBs applied to an OA (the PSC), we call this an *EXP-inferred-PSC LSP* (E-LSP). This means that the PSC is inferred from the EXP value in the shim header (see Fig. 4b).

This works only for LSRs that support shim headers but link layer specific labels do not have an EXP field. The solution is to set up a distinct LSP for each FEC and ordered aggregate (FEC-OA) pair and signal the PSC during the LSP setup. When the PSC of an OA contains more than one PHB, these different PHBs still need to be enforced. The PHBs of the PSC differ only in drop precedence; thus we need to encapsulate the drop precedence in the link-layer specific label. In ATM only the cell loss priority (CLP) bit can be used to encapsulate this information. Similarly, the discard eligibility (DE) bit of frame relay can be used to encapsulated the drop precedence (shown in Fig. 3). An LSP where the PSC is inferred from the label value is called a *label-only-inferred-PSC LSP* (L-LSPs), meaning that the PSC is inferred from the label values as opposed to the EXP field value (see Fig. 4c).

The use of L-LSPs is not restricted to link-layer-specific label encapsulating LSRs; it can also be used with generic MPLS encapsulation. The drop precedence is then encapsulated in the EXP field of the shim header, and the PSC is still inferred from the label value.

2.3.2.2. Allocation of Bandwidth to L-LSP and E-LSPs. Bandwidth can be allocated to E-LSP and L-LSPs at setup time. When resources are allocated to an L-LSP, the bandwidth is allocated to the PSC of the LSP; when bandwidth is allocated to an E-LSP, then the bandwidth is associated to the whole LSP, that is, the set of PSCs of the LSP. Signaling bandwidth requirements of the LSPs can be useful in two ways: (1) associating bandwidth to an LSP can be used to admit the traffic to the LSP according to the availability of resources and (2), the bandwidth allocation information can also be used to shift resources from certain PSCs to others. It is important to note that allocating resources to an L-LSP or E-LSP does not lead to the necessity of having a per LSP forwarding treatment.

2.4. History

MPLS started out as a technique for IP over ATM interworking as a convergence of a number of "IP

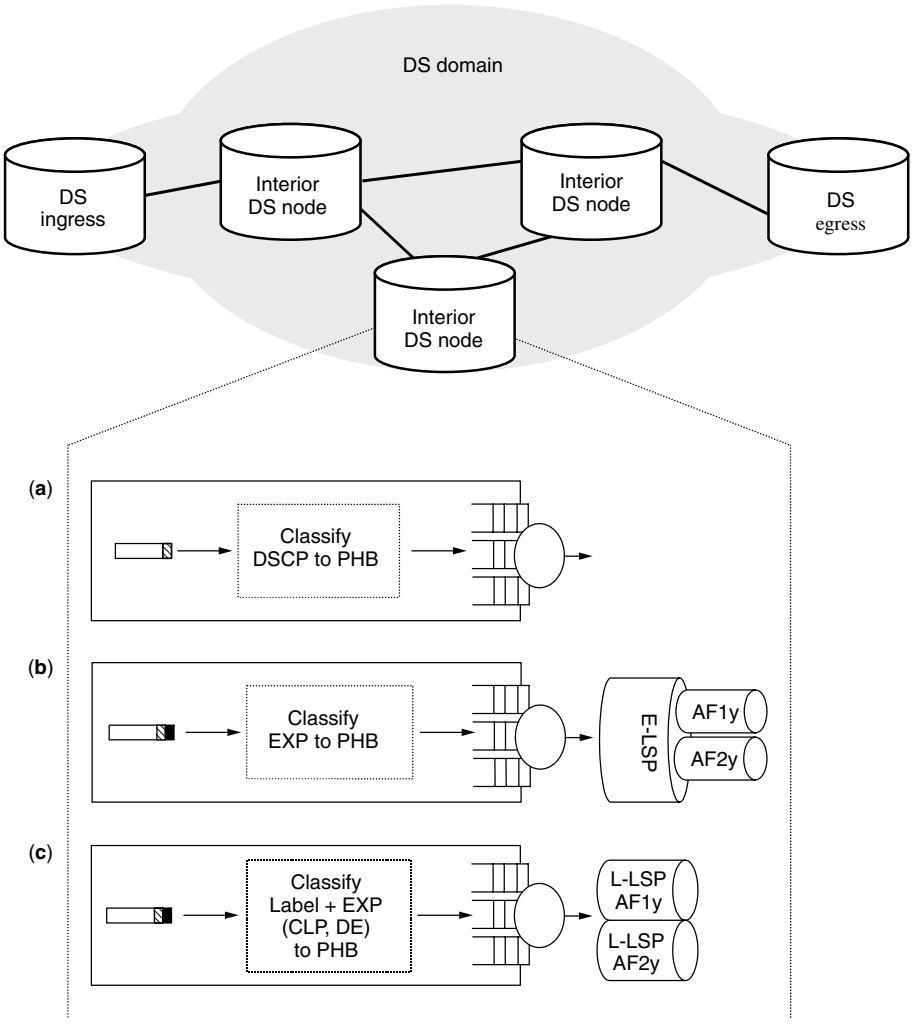


Figure 4. BA classification in DiffServ: (**a**) classification on the DSCP value of the IP header; (**b**) classification on the EXP bits of the shim header; (**c**) classification on the label and the EXP bits (or the ATM CLP or FR DE bit).

switching" schemes. IP switching is a technique that uses ATM hardware to forward of IP packets. In contrast to ATM networks, in MPLS networks the ATM hardware is administrated by IP and MPLS signaling protocols, and not by ATM signaling. There are a number of different IP switching implementations: Cisco Systems Tag Switching, IBM's Aggregated Route-based IP Switching (ARIS), Toshiba's Cell Switch Router (CSR), and NEC's Ipsofacto [11]. In order to standardize all these IP switching techniques, a new IETF work group came to life in 1997. The MPLS work group has since then been working on forming a common technology for IP switching.

MPLS contrasts with a number of other techniques for IP over ATM, which are overlay techniques. Examples of overlay techniques are multiprotocol over ATM (MPOA) and the work done in the IETF Internetworking Over Non-broadcast-capable networks (ION) working group. In the case of an ATM overlay network, there are two distinct networks: an ATM network and an IP network. This leads to disadvantage of having to administer the two networks and the fact that the scalability is limited due to full meshed peering [12].

3. APPLICATIONS OF MPLS

The following section will describe arguably the three most important applications of MPLS: traffic engineering, virtual private networks, and resilience (more specifically, fast rerouting).

3.1. Traffic Engineering

Traffic engineering (TE) is generally defined as the performance optimization of operational networks. (Other definitions focus more on the role of traffic engineering to offer efficient services to customers.) While TE is not strictly tied to multiservice networks and QoS, it is definitively more complex and mission-critical in multiservice networks. TE is probably considered as the most important application of MPLS networks.

In the following sections we will first address the applicability of TE and then discuss how these techniques are implemented in MPLS and how they relate to regular IP implementations.

3.1.1. Applicability. The optimization of operational networks is typically achieved by (1) the avoidance of congested routes, the (2) resource utilization of parallel links, and (3) routing policies using affinities.

1. *Avoiding Congested Routes*. When certain network segments are congested while others are underutilized, the network operator will want to route traffic away from the congested segments. In regular IP this can be done by modifying static link metrics [13] or using dynamic metrics, but this is difficult because of the destination-based forwarding of IP. In MPLS traffic can be routed away from the congested segments by setting up (explicit routed) LSPs. Traffic can be mapped on an LSP not only according to the destination, but virtually any classification can be used. A small number of LSPs can be used to route traffic away from the congested segments or a mesh of LSPs can be set up to distribute the traffic evenly over the network.

2. *Resource Utilization of Parallel Paths*. Regular IP calculates and uses only a single shortest path from point A to point B. This limitation is addressed by equal-cost multipath (ECMP) extensions to routing that takes paths of equal cost into account and spreads the traffic evenly over the available paths with the same cost. Even more advanced is the optimal multipath (OMP) extension, where paths with different cost values are used and the traffic is spread according to the relative cost (e.g., a path with a higher cost gets a lower share of the traffic). The cost metric of OMP can be dynamic; that is, it can be based on the actual load and length of the path. MPLS can be used to explicitly configure parallel paths. The calculation can be based on the online (routing) mechanisms such as ECMP or OMP, or alternatively an offline TE algorithm can compute the paths. Offline LSP calculations can be based on the measured and forecasted traffic between the edge nodes of the networks (the traffic matrix).

3. *Routing Policies*. A network operator might want to exclude some types of traffic from certain links or force traffic on certain links. In MPLS this can be achieved by using the resource class procedures of RSVP-TE or CR-LDP. In IP traffic engineering, extensions for OSPF or IS-IS have been defined to cope with resource affinity procedures.

3.1.2. Implementation. MPLS traffic engineering allows one to gain more network efficiency. But there's no such thing as a free lunch. Efficiency can be gained by introducing more LSPs in the network, but there's a tradeoff between the control granularity and the operational complexity associated with a large number of LSPs. In order for the traffic engineering to work properly, it is necessary to obtain detailed information about the behavior of LSPs (LSP monitoring). This is not a trivial task and can require significant resources. The assignment of resources to LSPs and the mapping of traffic to LSPs is another task that can be both time-consuming (for the network operator or in terms of computing power) and prone to errors due to inaccurate or outdated traffic matrices. Path calculation is another additional task to perform in comparison with non-traffic engineered networks. Finally, the signaling overhead introduced by traffic engineering can cause additional overhead.

An alternative for a traffic-engineered network is an overprovisioned network that always has enough capacity to transport the offer load. Even in overprovisioned networks, monitoring is necessary to determine when to upgrade the network capacity.

3.2. Virtual Private Networks

A *virtual private network* (VPN) is a network using secure links over the public IP infrastructure [14]. A VPN is a more cost-effective solution to a corporate extranet than a private network, which consists of private infrastructure. In order to create the extranet, the different sites have to be interconnected through the provider's network (ISP network). The access points between a customer's site and the provider are called *customer edge* (CE) and *provider edge* (PE), respectively. The internal routers are called *provider* (P) *routers* (see Fig. 5). A VPN consists of number

of sites that are connected through the ISP network. A participating site can be part of more than one VPN (e.g., site CE4). If two VPNs have no sites in common, then the VPNs can have overlapping address spaces. Since the addresses can overlap, the routers need to interpret the addresses on a persite basis by installing persite forwarding tables in the PE routers. MPLS VPNs are set up with the combination of BGP4 and another label distribution protocol (LDP, CR-LDP, or RSVP-TE) (see Section 2.2.4). The PE routers distribute labels associated with VPN routes to each other with BGP4. A VPN route is the combination of an IP prefix and a *router distinguisher* (RD). The RD allows one to distinguish between common prefixes of the different VPNs. The other label distribution protocol is used to create a mesh of LSPs between the PE routers. The VPN route labels and the labels distributed by the internal label distribution protocol are used by the PE routers to forward packets over the VPN. The internal LSRs (P routers) operate only on the top label, the label distributed by the internal label distribution protocol, so they don't need to be aware of the BGP routes.

Consider the example of a packet that needs to be forwarded from CE2 toward CE4 over VPN1. The ISP network consists of BGP peers on the edge (PE1 and PE2) and interior LSRs (P1 and P2). A BGP node sends a packet to a certain VPN by looking up the label it has received from his BGP next hop and pushes this label on the label stack. For example, PE1 pushes the label it has received from PE2 for VPN1. PE1 then looks up the label received from the internal label distribution protocol to PE2 and pushes this label on the label stack. The label stack then contains the BGP label for the VPN route (the VPN label) and on top of that the label for the BGP next hop (PE2). Then regular label switching is used to forward the packet to PE2. At PE2 the top label is popped and the VPN label pushed by PE1 is revealed. PE2 will then use this VPN label to look up the information needed to forward the packet to the next hop on VPN1, namely, CE4.

3.3. Resilience

Regular IP typically recovers from network (node and link) failures by rerouting. The routing protocol that is used to calculate the shortest paths in a network is able to detect network failures (or is notified of) and takes them into account when finding new routes after the failure. It typically takes some time before the routing protocol converges, that is, before the network reaches a stable state after the failure.

When using MPLS, IP routing can also be used to restore the shortest-path-routed LSPs. This rerouting in MPLS depends on the new paths calculated by the IP routing protocol, this means that MPLS rerouting based on IP rerouting is slower than IP rerouting.

However, MPLS allows the use of more advanced resilience schemes. *Protection switching* is a scheme where a recovery path is preestablished [15]. The recovery path can be node or link disjunct from the working path. (The working path is used as long as no failures are detected.) When a failure occurs, the traffic is switched from the working path to the recovery path (the protection switch). The use of protection switching leads to a much lower convergence time. An additional advantage of protection switching over rerouting is that resources can be allocated in advance so that even after a failure the traffic over the LSP can still be serviced according to the predefined traffic parameters. Rerouting typically does not offer such guarantees unless the network is carefully planned. The drawback of protection switching is that it requires recovery paths that are preestablished, leading to administrative and signaling overhead and a higher resource usage if the resources are dedicated (i.e., cannot be used when the recovery path is not in use).

BIOGRAPHIES

Pim Van Heuven graduated in computer science from Ghent University in 1998. At this same university

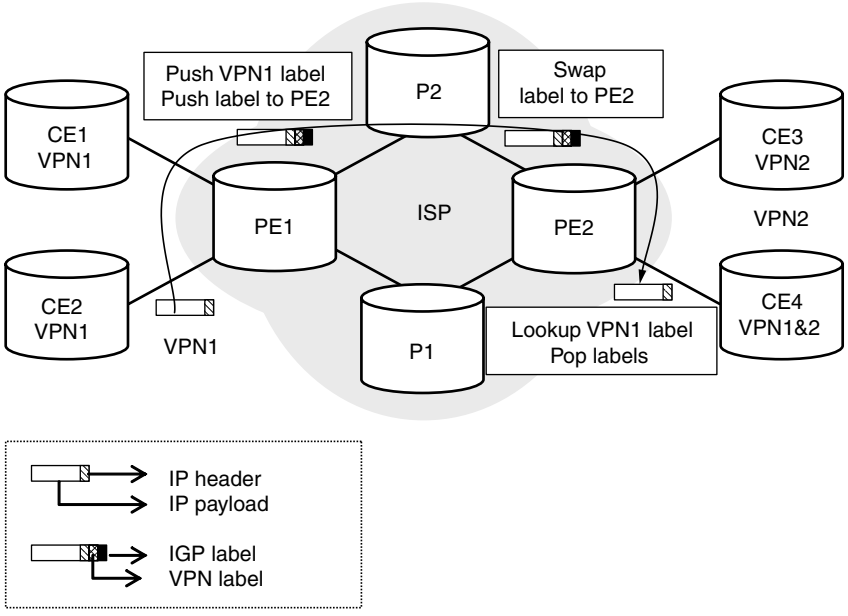


Figure 5. Example of an MPLS VPN and the forwarding of packets between the different interconnected sites.

he joined (August 1998) the Integrated Broadband Communications Networks Group (IBCN), where is now preparing a Ph.D. In January 1999 he was granted an IWT scholarship. He worked on the ACTS IthACI project. Since begin 2000 he has been working on the IST Tequila project. His research interests include MPLS, network resilience, and the areas of quality of service and traffic engineering. Since 2001, he has led the "RSVP-TE daemon for DiffServ over MPLS under Linux" project. He has published several papers on network resilience in IP, MPLS, and G-MPLS.

Steven Van den Berghe graduated in computer science from Ghent University in 1999. In July 1999, he joined the Broadband Communications Networks Group (IBCN) and now is preparing a Ph.D. In January 2001 he was granted an IWT scholarship. His research interests include mainly the areas of quality of service and traffic engineering in IP. He is focusing on measurement-based traffic engineering in a Diffserv/MPLS/multipath environment. He is active in the IST Tequila project and in development of DiffServ support for MPLS in the Linux community and has published, in addition to several papers, an Internet draft on the requirements for measurement architectures for use in traffic-engineered IP networks.

Filip De Turck received his M.Sc. degree in electronic engineering from the Ghent University, Belgium in June 1997. In May 2002, he obtained the Ph.D. degree in electronic engineering from the same university. From October 1997 to September 2001, Filip De Turck was Research Assistant with the Fund for Scientific Research — Flanders, Belgium (FWO-V). At the moment, he is affiliated with the Broadband Communications Networks Group (IBCN) of Ghent University as a Postdoctoral Researcher. His research interests include scalable software architectures for telecommunication networks and service management, performance evaluation and optimization of routing, admission control, and traffic management in telecommunication systems. He is the author of several research papers in this area and has served as a technical program committee member for several international conferences.

Piet Demeester (Senior Member IEEE) received his Ph.D. degree from Ghent University in the Department of Information Technology (INTEC) in 1988. He became Professor at the Ghent University, where he is currently responsible for the research on communication networks. He was involved in several European COST, ESPRIT, RACE, ACTS, and IST projects. He is a member of the editorial board of several international journals and has been a member of several technical program committees (ECOC, OFC, DRCN, ICCCN, IZS, etc.). His current interests are related to broadband communication networks (IP, MPLS, ATM, SDH, WDM, access, active, mobile) and include network planning, network and service management, telecom software, internetworking, and network protocols for QoS support. He has published over 200 papers in this field.

BIBLIOGRAPHY

1. C. Huitema, *Routing in the Internet*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
2. U. Black, *MPLS and Label Switching Networks*, Prentice-Hall, Englewood Cliffs, NJ, 2001.
3. B. Davie and Y. Rekhter, *MPLS Technology and Applications*, Morgan Kauffman, San Francisco, 2000.
4. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, IETF RFC (Online), *http://www.ietf.org/rfc/rfc3031.txt* (Jan. 2001).
5. L. Andersson et al., *LDP Specification*, IETF RFC (online), *http://www.ietf.org/rfc/rfc3036.txt* (Jan. 2001).
6. B. Jamoussi et al., *Constraint-Based LSP Setup Using LDP*, IETF RFC3212 (online), *http://www.ietf.org/rfc/rfc3212.txt* (Jan. 2002).
7. D. Durham and R. Yavatkar, *Inside the Internet's Resource reSerVation Protocol*, Wiley, New York, 1999.
8. D. Awduche et al., *RSVP-TE: Extensions to RSVP for LSP Tunnels*, IETF RFC (online), *http://www.ietf.org/rfc/rfc-3209.txt* (Dec. 2001).
9. Z. Wang, *Internet QoS: Architectures and Mechanisms for Quality of Service*, Morgan Kaufmann, San Francisco, 2001.
10. F. Le Faucheur et al., *Multi-Protocol Label Switching* (*MPLS*) *Support of Differentiated Services* (online), *http://www.ietf.org/rfc/rfc3270.txt* (May 2002).
11. I. Andrikopoulos et al., Experiments and enhancement for IP and ATM integration: The IthACI project, *IEEE Commun. Mag.* **39**(5): 146–155 (2001).
12. G. Armitage, MPLS: The magic behind the myths, *IEEE commun. Mag.* **38**(1): 124–131 (2000).
13. B. Fortz and M. Thorup, Internet traffic engineering by optimizing OSPF weights, *Proc. IEEE INFOCOM 2000*, Vol. 1, 2000, pp. 519–528.
14. J. Guichard and I. Pepelnjak, *MPLS and VPN Architectures: A Practical Guide to Understanding, Designing and Deploying MPLS and MPLS-Enabled VPNs*, Cisco Press, Indianapolis, 2000.
15. P. Van Heuven et al., Recovery in IP based networks using MPLS, *Proc. IEEE Workshop on IP-oriented Operations & Management IPOM'2000*, IEEE, 2000, pp. 70–78.

MULTIUSER WIRELESS COMMUNICATION SYSTEMS

ASHUTOSH SABHARWAL
BEHNAAM AAZHANG
Rice University
Houston, Texas

1. INTRODUCTION

The 1980s and 1990s witnessed the rapid growth and widespread success of wireless connectivity. The success of wireless systems is due largely to breakthroughs in communication theory and progress in the design of low-cost power-efficient mobile devices. Beyond the widespread use of voice telephony, new technologies are replacing wires

in virtually all modes of communication. For example, in addition to widely recognized outdoor connectivity via cellular wide-area networks (WANs), wireless local-area networks (LANs) and wireless personal-area networks (PANs) have also become popular. Wireless LANs (e.g., IEEE 802.11) provide high-speed untethered access inside buildings replacing traditional wired Ethernet, and wireless PANs (e.g., Bluetooth) are replacement for wires between common peripherals like mouse, keyboard, PDAs, and printers.

Providing ubiquitous mobile access to a large number of users requires solution to a wide spectrum of scientific and economic issues, ranging from low-power semiconductor design and advanced signal processing algorithms to the design and deployment of large cellular networks. In this article, we will highlight the challenges in the design of advanced signal processing algorithms for high-speed outdoor cellular access. The signal processing algorithms form the core of all wireless systems, and are thus critical for their success. In addition, the techniques and algorithms discussed here form a basis for most wireless systems, and thus have a wider applicability than outdoor wireless systems. To keep the discussion tractable, we will focus on baseband design for third-generation wireless cellular systems (e.g., WCDMA or CDMA2000) based on code-division multiple access (CDMA).

A wireless channel is a shared resource; multiple users in the same geographic locale have to contend for the common spectral resource and in the process interfere with other users. To allow meaningful and resource-efficient communication between different users, it is crucial that all participating users agree on a common protocol. The common protocol should enable fair access to the shared resource for all users. The three most commonly used multiple access protocols[1] are time-division (TDMA), frequency-division (FDMA), and code-division multiple access (CDMA). Among the three, direct-sequence CDMA (DS-CDMA) has been adopted as the access technique for all the third-generation wireless standards, and thus is the main focus of this article.

In outdoor cellular systems, the coverage area is divided into smaller regions called *cells*, each capable of supporting a subset of the users subscribing to the cellular system. The cellular structure exploits the fact that electromagnetic signals suffer loss in power with distance, thereby allowing reuse of the same communication channel at another spatially separated location. The reuse of communication channels allows a cellular system to support many more users as compared to a system that treats the whole geographic region as one cell. Each cell is served by a *base station* that is responsible for operations within a cell, primarily serving calls to and from users located in the respective cell. Figure 1 shows the components of a typical cellular system. The size and distribution of the cells [1] are dictated by the coverage area of the base station, subscriber density, and projected demand within a geographic region. As mobile users travel from cell to cell, their calls are *handed off* between cells in order to maintain seamless service. The base stations are connected to the *mobile telephone switching office* (MTSO), which serves as a controller to a group of base stations and as an interface with the fixed wired backbone.

Wireless networks, like typical multiple access networks, have a layered architecture [2,3]. The three main layers of each network are the physical layer, the network layer,[2] and the application layer. The physical layer is responsible for actual transport of data between the source and the destination points. The network layer controls the communication session, and the user applications

[1] We limit our discussion to circuit-switched networks and deterministic multiple-access schemes. In packet-switched networks, probabilistic multiple access is used; a good example is contention avoidance/resolution based protocol used in IEEE 802.11, and packet services used in EGPRS and 3G systems.

[2] The network layer consists of several layers that include the multiple-access layer (MAC), the data-link layer, and the transport layer.

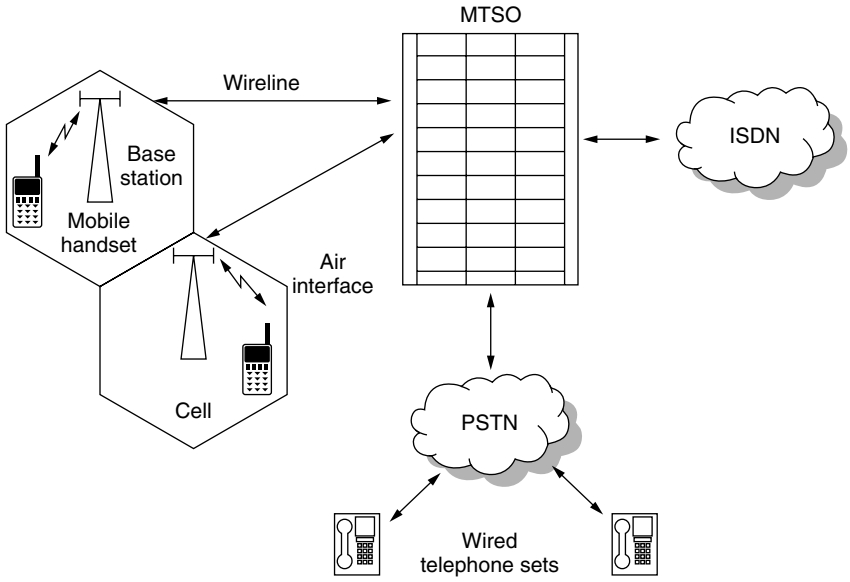


Figure 1. Components of a cellular wireless network.

operate in the application layer. Both network and application layer designs are critical in wireless networks, and are areas of active research. In this article, our focus will be on the design of physical layer for wireless networks.

The rest of the article is organized as follows. In Section 2, we will briefly discuss the three major challenges in the design of wireless systems and commonly used methods to combat them. Models for wireless channels are discussed in Section 3. In Section 4, we will introduce information-theoretic methods to analyze the limits of wireless systems. The core of the article is in Section 5, which discusses various aspects in the design of a typical transceiver. We conclude in Section 6.

2. CHALLENGES AND DESIGN OF WIRELESS SYSTEMS

In this section, we highlight the major challenges and techniques employed in wireless system design.

2.1. Time-Varying Multipath

Enabling mobility, which is the fundamental premise in designing wireless systems and is the major reason for their success, also presents itself as the most fundamental challenge. Because of the mobility of users and their surrounding environment, wireless channels are generally time-varying. Electromagnetic signals transmitted by base station or mobile users reach the intended receiver via several paths; the multiple paths are caused by reflections from man-made and natural objects (Fig. 2). Since the length of each path may be different, the resultant received signal shows wide fluctuations in its power profile (Fig. 3), thereby complicating the design of spectrally efficient systems.

To combat time-varying fading, a combination of time, spatial or frequency diversity is commonly used [4]. By using diversity techniques, the receiver obtains multiple copies of the transmitted signal, thereby increasing the chance that at least one of the copies is reliable. To exploit time diversity, error control codes are used in conjunction with an interleaver [4]. Spatial diversity can be obtained by using multiple antennas that are sufficiently separated.

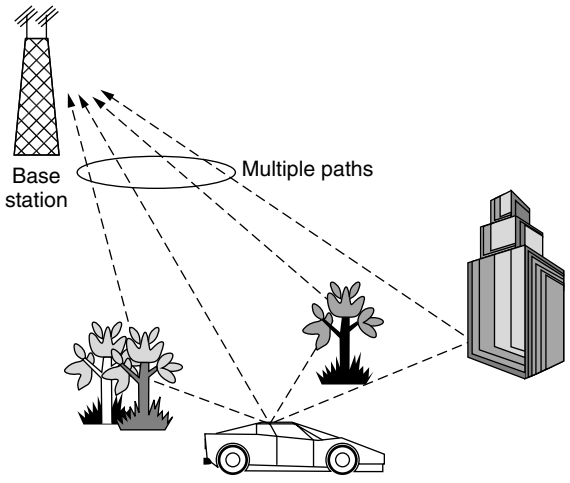


Figure 2. Multipath propagation.

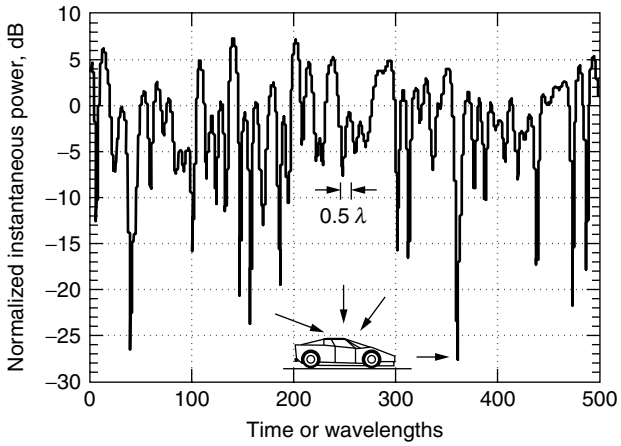


Figure 3. Time variations in the received signal power due to multipath and user mobility.

Spatial diversity can be tapped by using space-time codes [5] at the transmitter or signal combining [6] at the receiver. Spatial diversity techniques have received considerable interest because of their potential to support larger data rates on the same channels compared to current technology. Frequency diversity is analogous to spatial diversity where frequency selectivity due to multipath is used.

2.2. Shared Multiple Access

Unlike wired networks, where new bandwidth is "created" by adding additional physical resources (cables, servers, etc.), users in wireless systems have to share limited spectral resources. Although, the available spectrum for commercial wireless system has increased since 1980, it is clear that growth in demand will always outpace the available spectrum. Limited growth of resources immediately implies that the requirements of new data rate hungry wireless services can be sustained only by progress in efficiently using the available spectrum. An obvious way of increasing system capacity is to use smaller cells, but using smaller cells is undesirable for economic reasons; an increased number of base stations and the required wired backbone are the major reasons for the increased system cost. Further, smaller cells generally lead to increased intercell handoffs and out-of-cell interference, leading to diminishing returns with increasing cell partitioning.

The capacity of cellular systems can also be improved by cell sectorization [7,8], where each cell is further divided into sectors. Cell sectorization is achieved by using multiple directional antenna [9] at each base station, thereby reducing the intersector interference. Because of the directional antenna response, cell sectorization has also been shown to reduce the delay spread of the received signal leading to power savings [10]. Much like cell splitting, cell sectorization also has its limits. To achieve smaller sectors using directional antennas requires increasingly large antennas, which are both expensive and hard to deploy.

Information-theoretic results [11] for multiuser systems indicate that the optimal methods to share spectral

resources should not attempt to avoid intercell and intracell interference. The cochannel interference in wireless systems can be suppressed by using multiuser detection [12], leading to increased spectral efficiency [13,14]. Further improvements in system capacity can be obtained by the use of dynamic resource allocation among users, for example, adaptive channel assignment techniques [15] and dynamic spreading gain and power control [16].

2.3. Power Limitation for Mobile Users

Since most of the mobile devices are battery-operated, power efficiency is a crucial design parameter in wireless systems. The major consumers of power in wireless handsets are power amplifier used during transmission, silicon based computing units (A/D, D/A, and baseband processor) used in reception, and in some cases, the color display.

Power dissipation in the RF power amplifier can be reduced by using cells with smaller radii, better multiuser signal processing at the base-station, improved coding schemes, or receiver diversity. As pointed out earlier, cell splitting is not attractive because of increased system cost with diminishing returns. Advanced signal processing, multiuser channel estimation, and data detection have been shown to greatly reduce the power requirements to achieve a desired performance level [12]. More recent advances in channel coding, namely, Turbo coding [17], can lead to further reduction in power requirements for the transmitter to achieve a desired performance level. Reduction in power requirements of baseband processing units requires development of hardware-frugal algorithms and low-power CMOS circuits. Also, techniques that require more computation at the base-station to cut the complexity of handset are very effective in saving power at the mobile unit.

3. FADING CHANNEL MODELS

In this section, we will describe time-varying wireless channels and the statistical models used to capture their effect on transmitted signals. A detailed discussion of channel models can be found elsewhere [4,18]. A fading multipath channel is generally modeled as a linear system with time-varying impulse response[3] $h(t;\tau)$. The time-varying impulse response is assumed to be a wide-sense stationary random process with respect to the time variable t. Because of time variations of the channel, the transmitted signal is spread in frequency; the frequency spreading is called *Doppler spreading*. The transmitted signal also suffers time spreading as a result of multipath propagation. Thus, the received signal is spread both in time and frequency.

Two parameters are commonly used to characterize wide-sense stationary channels: *multipath delay spread* and *Doppler spread*. To define the multipath delay and Doppler spread, it is convenient to work with the scattering function $\mathcal{H}(\tau;\lambda)$, which is a measure of average power output[4] of the channel at delay τ and frequency offset λ relative to the carrier. The *delay power spectrum* of the channel is obtained by averaging $\mathcal{H}(\tau;\lambda)$ over λ:

$$\mathcal{H}_c(\tau) = \int_{-\infty}^{\infty} \mathcal{H}(\tau;\lambda)\, d\lambda \tag{1}$$

The multipath delay spread T_m is the maximum delay τ for which the delay power spectrum $\mathcal{H}_c(\tau)$ is nonzero. Similarly, the Doppler spread B_d is the maximum value of λ for which the following *Doppler power spectrum* $\mathcal{H}_c(\lambda)$ is nonzero:

$$\mathcal{H}_c(\lambda) = \int_{-\infty}^{\infty} \mathcal{H}(\tau;\lambda)\, d\tau \tag{2}$$

The reciprocal of the multipath delay spread is defined as *channel coherence bandwidth*, $B_{\text{coh}} = 1/T_m$ and provides an indication of the width of band of frequencies that are similarly affected by the channel. The Doppler spread provides a measure of how fast the channel variations are in time. The reciprocal of Doppler spread is called *channel coherence time* $T_{\text{coh}} = 1/B_d$. A large value of T_{coh} represents a slowly fading channel and a small values represents fast fading. If $T_m B_d < 1$, then the channel is said to be *underspread*; otherwise it is *overspread*. In general, if $T_m B_d \ll 1$, then the channel can be accurately measured at the receiver, which can aid in improving the transmission schemes. On the other hand, channel measurement is unreliable for the case of $T_m B_d > 1$.

An appropriate model for a given channel also depends on the transmitted signal bandwidth. If $s(t)$ is the transmitted signal with the Fourier transform $S(f)$, the received baseband signal, with the additive noise, is

$$\begin{aligned} z(t) &= \int_{-\infty}^{\infty} h(t;\tau)s(t-\tau)d\tau + \nu(t) \\ &= \int_{-\infty}^{\infty} H(t;f)S(f)e^{j2\pi ft}df + \nu(t) \end{aligned}$$

where $H(t;f)$ is the Fourier transform of $h(t;\tau)$ with respect to τ. If the bandwidth W of the transmitted signal $S(f)$ is much smaller than the coherence bandwidth, $W \ll B_{\text{coh}}$, then all the frequency components in $S(f)$ undergo the same attenuation and phase shift during propagation. This implies that within the bandwidth of the signal, the transfer function $H(t;f)$ is constant in f, leading to a *frequency nonselective* or *flat fading*. Thus, the received signal can be rewritten as

$$\begin{aligned} z(t) &= H(t;0)\int_{-\infty}^{\infty} S(f)e^{j2\pi ft}df + \nu(t) \\ &= H(t)s(t) + \nu(t) \end{aligned} \tag{3}$$

where $H(t) \in \mathbb{C}$ is the complex multiplicative channel. A flat fading channel is said to be *slowly fading* if the symbol

[3] A linear time-invariant system requires a single-variable transfer function. For a time-varying linear system, two parameters are needed; the parameter t in $h(t;\tau)$ captures the time-variability of the channel.

[4] Under the assumption that all different delayed paths propagating through the channel are uncorrelated.

time duration of the transmitted signal T_s is much smaller than the coherence time of the channel, $T_s \ll T_{\text{coh}}$. The channel is labeled as *fast fading* if $T_s \geq T_{\text{coh}}$.

If the signal bandwidth W is much greater than the coherence bandwidth of the channel, then the frequency components of $S(f)$ with frequency separation more than B_{coh} are subjected to different attenuations and phase shifts. Such a channel is called *frequency selective*. In this case, multipath components separated by delay more than $1/W$ are resolvable and the channel impulse response can be written as [4]

$$h(t;\tau) = \sum_{p=1}^{P} h_p(t)\delta\left(\tau - \frac{p}{W}\right) \tag{4}$$

Since the multipath delay spread is T_m and the time resolution of multipaths is $1/W$, the number of paths L is given by $\lfloor T_m W \rfloor + 1$. In general, the time-varying tap coefficients $h_p(t)$ are modeled as mutually uncorrelated wide-sense stationary processes. The random time variations of the channel are generally modeled via a probability distribution on the channel coefficients $h_p(t)$. The most commonly used probability distributions are Rayleigh, Ricean, and the Nakagami-m [4].

The main purpose of the channel modeling is to characterize the channel in a tractable yet meaningful manner, to allow design and analysis of the communication algorithms. Note that *all* models are approximate representations of the actual channel, and thus development of practical systems requires both theoretical analysis and field testing.

In the sequel, we will consider only slowly fading channels, where $T_s \ll T_{\text{coh}}$, that is, multiple consecutive symbols or equivalently, a block of symbols undergo the same channel distortion. Hence, these channels are also referred as *block fading channels* [19–22]. As a result of slow time variation of the channel, the time dependency of the channel will be suppressed; that is, $h(t)$ will be denoted by h and $h(t;\tau)$ by $h(\tau)$.

4. CAPACITY OF MULTIPLE-ACCESS CHANNELS

Developed in the landmark paper by Shannon [23], information theory forms the mathematical foundation for source compression, communication over noisy channels and cryptography. Among other important contributions [23], the concept of *channel capacity* was developed. It was shown that a noisy channel can be characterized by its capacity, which is the maximum rate at which the information can be transmitted reliably over that channel. Information theoretic methods provide not only the ultimate achievable limits of a communication system but also valuable insights into the design of practical systems.

Typically, a capacity analysis starts by using a simple model of the physical phenomenon. The simplified model captures the basic elements of the problem, such as time-varying fading wireless channel, shared multiple access, and power-limited sources. Information-theoretic analysis then leads to limits on reliably achievable data rates and provides guidelines to achieve those limits. Although information-theoretic techniques are rarely practical, information-theory-inspired coding, modulation, power control and multiple-access methods have led to significant advances in practical systems. Furthermore, the analysis techniques allow performance evaluation of suboptimal but implementation-friendly techniques, thereby providing a useful benchmarking methodology.

In this section, we will provide a brief sampling of results pertaining to time-varying fading wireless channels; the reader is referred to Ref. 19 for a detailed review. Our aim is to highlight basic single and multiuser results for fading channels to motivate the algorithms discussed in the sequel. In Section 4.1, we will first introduce two notions of channel capacity, Shannon theoretic capacity [23] and outage capacity [24]. Capacity of a channel characterizes its performance limits using *any* practical transmitter–receiver pair and is a fundamental notion in evaluating efficacy of practical systems. Single-user fading channels will be analyzed using the two capacity notions, motivating the importance of diversity techniques (e.g., spacetime coding and beamforming) and power control. In Section 4.2, the multiuser extensions will be discussed to motivate the use of power-controlled CDMA-based multiple access.

All results in this section will be given for flat fading channels. The results can be easily extended to frequency-selective fading by partitioning the channel into frequency bins of width B_{coh}, and then treating each bin as a separate channel.

4.1. Capacity of Single User Fading Channels

A channel is deemed *noisy* if it introduces random perturbations in the transmitted signals. In Ref. 23, the capacity of a noisy channel was defined as the highest data rate at which reliable communication is possible across that channel. *Communication reliability* is defined as the probability that the receiver will decode the transmitted message correctly; higher reliability means lower errors in decoding messages and vice versa. An information rate is *achievable* if there exists at least one transmission scheme such that any preset level of communication reliability can be achieved. To achieve this (arbitrary level of) reliability, the transmitter can choose any codebook to map information message sequences to channel inputs. If the rate of transmission R is no more than the channel capacity C, then reliable communication is possible by using codebooks that jointly encode increasingly longer input messages. This notion of channel capacity is commonly referred as *Shannon theoretic capacity*.

Besides providing a characterization of the channel capacity for a broad class of channels, Shannon [23] also computed the capacity of the following additive white Gaussian noise (AWGN) channel,

$$z(t) = s(t) + v(t) \tag{5}$$

as

$$C = W \log_2\left(1 + \frac{\mathcal{P}_{\text{av}}}{\sigma^2}\right) \text{ bits per second (bps)} \tag{6}$$

Note that the AWGN channel in (5) can be considered as a special case of fading channel (3) with $h(t) \equiv 1$. In (6), W represents the channel bandwidth (in hertz), $\mathcal{P}_{\text{av}} = \mathbb{E}_s\{|s(t)|^2\}$ is the average transmitted power over time,[5] and σ^2 is the variance of the additive noise $\nu(t)$. The fundamental formula (6) clarifies the role of two important system parameters: the channel bandwidth W and signal to noise ratio (SNR), $\mathcal{P}_{\text{av}}/\sigma^2$. The capacity result (6) claims a surprising fact that even for very small amount of power or bandwidth, information can be sent at a nonzero rate with vanishingly few decoding errors. To achieve this reliable communication, the transmitter *encodes* multiple information bits together using a channel code. The encoded bits are then jointly decoded by the receiver to correct errors introduced by the channel (5).

The capacity analysis in Ref. 23 forms the basis for deriving capacity of fading channels (3), which we review next. With an average transmitted power constraint, $\mathbb{E}_s\{|s(t)|^2\} \le \mathcal{P}_{\text{av}}$, the Shannon theoretic capacity of fading channels, with perfect channel information at the receiver, is given by [25]

$$C_{sc}^{r} = W\mathbb{E}_{\gamma}\left\{\log_2\left(1 + \frac{\mathcal{P}_{\text{av}}\gamma(t)}{\sigma^2}\right)\right\} \tag{7}$$

where σ^2 is the variance of the additive i.i.d. Gaussian noise $\nu(t)$ in (3), and $\gamma(t) = |h(t)|^2$ is the received instantaneous power. The expectation in (7) is computed with respect to the probability distribution of the variable $\gamma(t)$. If, in addition to perfect channel information at the receiver, the transmitter has knowledge of the instantaneous channel realization, then the transmitter can adapt its transmission strategy based on the channel. The optimal strategy, in this case, turns out to be "water-filling" in time [26]. To water-fill in time, the transmitter waits for the good channel conditions to transmit and does not transmit during poor channel conditions. Thus, the optimal transmission policy is a constant rate Gaussian codebook (see Ref. 11 for details on Gaussian codebooks) transmitted using an instantaneous channel SNR-dependent power. The optimal transmission power is given by [26]

$$\mathcal{P}_{sc}(\gamma(t)) = \begin{cases} \mathcal{P}_{\text{av}}\left(\dfrac{1}{\gamma_{sc}} - \dfrac{1}{\gamma(t)}\right), & \gamma(t) \ge \gamma_{sc} \\ 0, & \gamma(t) < \gamma_{sc} \end{cases} \tag{8}$$

where the threshold γ_{sc} is found to satisfy the power constraint $\mathbb{E}_{\gamma,s}\{\mathcal{P}_{sc}(\gamma(t))|s(t)|^2\} \le \mathcal{P}_{\text{av}}$. The achievable capacity is then given by

$$C_{sc}^{rt} = W\mathbb{E}_{\gamma}\left\{\log_2\left(1 + \frac{\mathcal{P}_{sc}(\gamma(t))\gamma(t)}{\sigma^2}\right)\right\} \tag{9}$$

Note that allocated power in (8) is zero for poor channels whose SNR is less than $\gamma_{sc}(t)$ and increases monotonically as channels conditions improve. Adapting the transmission power based on channel conditions is known as *power control*. Channel state information at the transmitter leads to only modest gains for most fading distributions [26] with a single transmitter and receiver; thus C_{sc}^{rt} is only marginally greater than C_{sc}^{r}. But the gains of transmitter information increase dramatically with multiple transmit and receive antennas. Using the extensions of (7) and (9) to multiple antennas [25,27], a representative example is shown in Fig. 4. Thus, building adaptive power control policies is more useful for multiple antenna systems; see Ref. 28 for practical methods to achieve a significant portion of this capacity in a practical system. The gain due to channel state information at the transmitter can also be achieved by using imprecise channel information [28–30]. The large gains promised by multiple antenna diversity, with or without channel information at the transmitter, have sparked the rich field of space-time coding [5,31,32].

In slow-fading channels, achieving Shannon theoretic capacity requires coding over exceedingly long input blocks. The long codewords are required to average over different fading realizations, which then allow the use of assumed ergodicity[6] of the fading process to prove the capacity theorem. The large delays associated with Shannon theoretic capacity directly translate into impractical delays in delay sensitive applications like voice and video. Thus, with a delay constraint, the Shannon

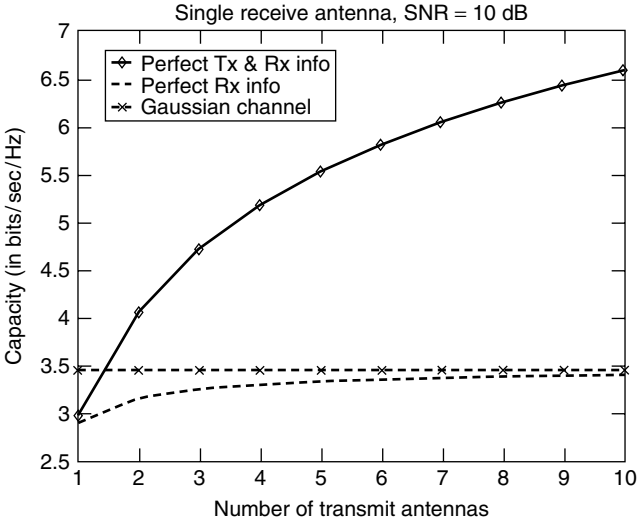


Figure 4. Capacity with multiple transmit antennas and single receive antenna, with different amount of channel state information at the transmitter.

[5] The expectation $\mathbb{E}_s\{|s(t)|^2\}$ represents an average computed over time (assuming that it exists) using the distribution of $s(t)$.

[6] A stochastic process $h(t)$ is called ergodic if its ensemble averages equal appropriate time averages [33]. The channel capacity theorem proved by Shannon [23] relied on the law of large numbers, where the time averages converge to their ensemble averages, which in turn motivated the idea of encoding increasingly long blocks of input messages. Ergodic channels are the most general channels with dependency across time for which the (strong) law of large numbers holds, thereby allowing a direct extension of capacity theorem [23] to ergodic channels. For a more general capacity theorem without any assumptions on channel structure, see Ref. 34.

theoretic capacity of slowly fading practical channels (more specifically, nonergodic channels) is zero [24]. In Ref. 24, the concept of capacity versus outage was introduced, which captures the effect of delay in slow fading channels. A block of transmitted data, which is assumed to undergo the same fading throughout, is in *outage* if the instantaneous capacity of the channel is less than the rate of transmission. The concept of outage provides a code-independent method (by using asymptotic approximations) to gauge the codeword error probability for practical codes. Assuming that the flat-fading channel h is constant for a block of transmitted data, the instantaneous capacity is given by[7] $W\log_2(1+\mathcal{P}_{\text{av}}\gamma(t)/\sigma^2)$. The outage probability, when only the receiver is aware of the channel state, is then given by

$$\Pi_{oc}^{r} = \text{Prob}\left(W\log_2\left(1+\frac{\mathcal{P}_{\text{av}}\gamma(t)}{\sigma^2}\right) < R\right) \tag{10}$$

where the probability is computed over the distribution of channel $h(t)$. Analogous to the preceding Shannon theoretic capacity analysis, the probability of outage can also be computed for different amount of channel state information at the transmitter. With perfect channel state information at the transmitter and receiver, the outage probability is given by

$$\Pi_{oc}^{rt} = \min_{\mathcal{P}_{oc}(\gamma(t))} \text{Prob}\left(W\log_2\left(1+\frac{\mathcal{P}_{oc}(\gamma(t))\gamma(t)}{\sigma^2}\right) < R\right) \tag{11}$$

The power allocation $\mathcal{P}_{oc}(\gamma(t))$ minimizing the outage is given by [27]

$$\mathcal{P}_{oc}(\gamma(t)) = \begin{cases} \dfrac{\sigma^2(2^{R/W}-1)}{\gamma(t)}, & \gamma(t) \geq \gamma_{oc} \\ 0, & \gamma(t) < \gamma_{oc} \end{cases} \tag{12}$$

The threshold γ_{oc} is chosen to meet the average power constraint, $\mathbb{E}_{\gamma,s}\{\mathcal{P}_{oc}(\gamma(t))|s(t)|^2\} \leq \mathcal{P}_{\text{av}}$. The *outage capacity*, which measures the total number of transmitted bits per unit time not suffering an outage, is given by

$$C_{oc}^{r} = (1-\Pi_{\text{out}}^{r})R$$
$$C_{oc}^{rt} = (1-\Pi_{\text{out}}^{rt})R$$

Because of the extra information at the transmitter, it immediately follows that $\Pi_{\text{out}}^{rt} < \Pi_{\text{out}}^{r}$ and hence $C_{oc}^{rt} > C_{oc}^{r}$. The gain in outage capacity due to transmitter information is much more substantial compared to Shannon capacity even for a single-antenna system [35]. Similar to the Shannon-capacity, outage capacity increases with the increasing number of transmit and receive antennas [25,36].

The differences in the objectives of achieving outage capacity versus achieving Shannon theoretic capacity can be better appreciated by the difference in the optimal power allocation schemes, $\mathcal{P}_{sc}(\gamma(t))$ and $\mathcal{P}_{oc}(\gamma(t))$. In the Shannon theoretic approach, the transmitter uses more power in the good channel states and less power during poor channel conditions. On the other hand, to minimize outage the transmitter employs *more* power as the channel gets *worse*, which is exactly opposite to the power allocation $\mathcal{P}_{sc}(\gamma(t))$. The difference in power allocation strategies, $\mathcal{P}_{sc}(\gamma(t))$ and $\mathcal{P}_{oc}(\gamma(t))$ can be attributed to optimization goals: Shannon theoretic capacity maximizes long-term throughput and hence it is not delay-constrained, and outage capacity maximizes short-term throughput with delay constraints.

Irrespective of the capacity notion, the main lesson learned from information-theoretic analysis is that diversity and channel information at the transmitter can potentially lead to large gains in fading channels. The gains promised by above information-theoretic results have motivated commonly used methods of space-time coding and power control to combat fading. Readers are referred to the literature [21,25,26,36–38] for detailed results on capacity of single user flat-fading channels. In the next section, we will briefly discuss the results for multiple access channels and their impact on the choice of multiple access protocols.

4.2. Multiple User Fading Channels

The primary question of interest in a multiuser analysis is the multiaccess protocol to efficiently share the spectral resources among several power-limited users. An accurate capacity analysis of a complete cellular system is generally intractable. Hence, the information-theoretic analysis relies on a series of simplifying assumptions to understand the dominant features of the problem. Our main emphasis will be on uplink communication in a single cell, where multiple users simultaneously communicate with a single receiver, the base station.

The sampled received baseband signal at the base station is the linear superposition of K user signals in additive white Gaussian noise, given by

$$y(t) = \sum_{i=1}^{K} h_i(t)s_i(t) + \nu(t) \tag{13}$$

The Gaussian noise $\nu(t)$ is assumed to be zero mean with variance σ^2. The channels for all users $h_i(t)$ are assumed to vary independently of each other and from one coherence interval to another. The fading processes for all users are assumed to be jointly stationary and ergodic. Furthermore, each user is subjected to an average power constraint, $\mathbb{E}_{s_i}\{|s_i(t)|^2\} \leq \mathcal{P}_i$.

Equivalent to the capacity of channel in the single-user case, a *capacity region* specifying all the rates that can be simultaneously and reliably achieved are characterized. Thus, the capacity region for K users is a set of rates defined as

$$\mathcal{R} = \{\underline{R} = (R_1, R_2, \ldots, R_K) : \text{ rates } R_i \text{ can be reliably achieved simultaneously}\} \tag{14}$$

When the base-station receiver is aware of all the fading realizations of all the users, $\{h_i(t)\}$, then the rate region

[7] Assuming that the transmitter is unaware of the instantaneous channel state and receiver has the perfect knowledge of $h(t)$ [25].

is described by the following set of inequalities (in the single-user case, there is only one inequality, $R \le C$)

$$\sum_{i\in\mathcal{B}} R_i \le \mathbb{E}_{\underline{\gamma(t)}} \log_2\left(1+\frac{\sum_{i\in\mathcal{B}}\gamma_i(t)\mathcal{P}_{\text{av}}}{\sigma^2}\right) \tag{15}$$

where it is assumed that each user has the same average power limit $\mathcal{P}_i = \mathcal{P}_{\text{av}}$. In (15), $\mathcal{B}$ represents a subset of $\{1, 2, \ldots, K\}$, $\gamma_i(t) = |h_i(t)|^2$ is the received power, and $\gamma(t) = [y_1(t)y_2(t)\ldots y_k(t)]$. The expectation of $\mathbb{E}_{\underline{\gamma(t)}}$ is over all the fading states $\{\gamma_i(t)\}_{i\in\mathcal{B}}$. A quantity of interest is the *normalized sum rate*, which is the maximum achievable equal rate per user and is obtained by taking $\mathcal{B}$ to be the whole set to yield [39]

$$R_{\text{sum}} = \frac{1}{K}\sum_{i=1}^{K} R_i = \mathbb{E}_{\underline{\gamma(t)}}\frac{1}{K}\log_2\left(1+\frac{\mathcal{P}_{\text{av}}\sum_{i=1}^{K}\gamma_i(t)}{\sigma^2}\right) \tag{16}$$

$$\underset{K\to\infty}{\to} \frac{1}{K}\log_2\left(1+\frac{K\mathcal{P}_{\text{av}}}{\sigma^2}\right) \tag{17}$$

The asymptotic result (17) shows an interesting phenomenon, that as the number of users increases, the effect of fading is completely mitigated because of the averaging effect of multiple users. The averaging effect due to increasing users is analogous to time or frequency [40] or spatial [25] averaging in single-user channels. Shamai and Wyner [39], using (16), showed that a nonorthogonal multiple-access scheme has a higher normalized sum rate R_{sum} than orthogonal schemes such as time (frequency) division multiple access.[8] By requiring orthogonality of users, an orthogonal multiple access scheme adds additional constraints on user transmission, which leads to a performance loss compared to optimal nonorthogonal method. Nonorthogonal CDMA is an example of the nonorthogonal multiple-access scheme. Spread signals, like CDMA signals, occupy more bandwidth than needed and were first conceived to provide robustness against intentional jamming [41]. The capacity–outage analysis also shows the superiority of CDMA schemes over orthogonal access methods [42].

A cellular multicell model [43] was introduced to study the effect of multiple cells. The model extends (13) to include intercell interference from users in neighboring cells. The cellular model [43] was extended to fading channels [39,44]. There again, it was concluded that CDMA, like wideband methods achieve optimal normalized sum rates even in the presence of multicell interference, for several important practical receiver structures. Even though the spread-spectrum signals occupy more bandwidth than needed for each signal, multiuser spread spectrum systems are spectrally efficient [13,14]. Motivated by the success of the second generation CDMA standard, IS-95, currently all third generation wireless systems (CDMA2000 and W-CDMA) use some form of spread spectrum technique. In addition to information theoretic superiority, CDMA-based multiple access provides other practical advantages [45]. First, CDMA signals allow finer *diversity combining* due to larger signal bandwidth, thereby providing robustness to multipath fading. In other words, combined with an interleaver, spread-spectrum signals naturally exploit both frequency and time diversity. Frequency diversity is not available in bandwidth-efficient TDMA systems. Moreover, CDMA allows a *frequency reuse* of one in contrast to TDMA/FDMA, which requires a higher reuse factor. A lower reuse factor immediately implies higher system capacity; a reuse factor of one also simplifies frequency planning. Finally, CDMA naturally exploits the *traffic activity factor*, the percentage of time during a two-way communication each channel is actually used. Most of the information theoretic analysis completely ignores the data burstiness, a property which is central to higher resource utilization in wired networking [46]; see Refs. 47 and 48 for insightful reviews.

The CDMA based systems allow communication without the need for a universal clock or equivalently synchronism among different users. The need for synchronism in TDMA requires the use of time guard bands between time slots and hence wastes resources. Finally, in long-code DS-CDMA systems, like the one used in IS-95 standard[9] assigning channels to users is straightforward because each user is given a unique fixed spreading code. In TDMA, time slots are granted adaptively as users hand off from one cell to another, thereby complicating resource management and requiring additional protocol overhead. Also, long-code CDMA leads to the same average performance for all users, and thus a fair resource allocation among users.

Although the area of multiuser information theory is rich and well studied, we maintain that many fundamental results are yet to be published. For instance, connections with queuing theory [47–49], which is the mathematical basis for networking, are far from well understood, but with the rise of Internet, it is more urgent than ever to unify the areas of data networking and wireless communications. Furthermore, with the growth of wireless services beyond voice communication, and advent of newer modes of communication like ad hoc networking,[10] current information-theoretic results should be considered as the beginnings of our understanding on the subject of multiuser communications.

5. TYPICAL ARCHITECTURE OF WIRELESS TRANSCEIVER

Most wireless systems transmit signals of finite bandwidth using a high-frequency carrier.[11] This immediately leads to the wireless transceiver with three major components: (1)

[8] In time (frequency)-division multiple access, each user transmits in its allocated time (frequency) slot such that no two users share a time (frequency) slot. Thus, the transmission of one user is orthogonal in time (frequency) to any other user.

[9] In long-code CDMA systems, unlike short repeating-code CDMA systems, each transmitted bit is encoded with a different spreading code.

[10] In ad hoc networking, mobile nodes can communicate with each other without the need for any infrastructure as in cellular systems; IEEE 802.11 and Bluetooth are examples of ad hoc networking.

[11] Carrierless systems include impulse radio [50].

an RF front end that performs the frequency conversion from passband to baseband and vice versa, (2) digital-to-analog converter (D/A) and analog to digital (A/D) converter, and (3) a baseband processing unit. In this section, we will discuss the signal processing algorithms used in the digital baseband unit. Wherever applicable, we will highlight the differences between the baseband unit at the mobile receiver and that at the base station.

We briefly note that the hardware receiver design for CDMA systems is generally more challenging than its TDMA counterparts. The design of A/D, D/A converters, and digital baseband processors require special effort. Higher chipping rates in CDMA systems require faster sampling and hence lead to higher computational throughout requirements and increased circuit power dissipation compared to their TDMA counterparts. Fortunately, advances in low-power high-speed complementary metal oxide semiconductor (CMOS) circuits have allowed implementation of sophisticated digital signal processing algorithms, and high-speed converters.

5.1. Transmitter

A simplified transmitter for DS-CDMA system is shown in Fig. 5. The data obtained from the higher layers is passed through a channel encoder, spread-spectrum modulator, digital to analog converter and finally through an RF unit.

5.1.1. Channel Encoding. The source data bits are first encoded using a forward error correction (FEC) code. A FEC code systematically adds redundant bits to the source bits, which are used by the receiver to correct errors in the received signal. Error correction coding is essential to achieve low bit error rates at the receiver and has a strong information theoretic foundation [23]. Following Shannon's work in 1948 [23], error control coding has seen tremendous growth since 1950; the readers are referred to the literature [51–54] for recent reviews on state of the art. Several excellent texts [55–58] on channel coding theory are available, hence we will keep our discussion in this section elementary.

The choice of code primarily depends on desired performance level, the specific channel under consideration and the complexity of the resulting receiver. The desired level of performance is based on the type of services to be provided. For instance, loss tolerant services such as speech can work with high packet loss probability, while data/email/fax requires a much higher error protection, thereby requiring FEC codes with different amount of error protection capabilities.[12] The complexity of decoding the received packets to correct errors is a major concern in the design of power-limited mobile handsets. Typically, stronger FEC codes are computationally harder to decode and, hence require more battery power for the baseband units; see Ref. 59 for a discussion.

The communication channel is a major factor in selection of FEC codes. For example, code design is different for slow- and fast-fading channels. To illustrate the concept of coding, our discussion will be limited to convolutional codes that are used in both telephone line modems, and both second and third generation digital wireless cellular standards. Further, we will highlight the interest in space-time coding by dividing this section into two parts: single-antenna systems and multiple-antennas systems. Our discussion on single-antenna systems will give a quick introduction to convolutional codes with a review of the most recent coding results for slow and fast fading channels. In the multiple antenna discussion, diversity techniques will be central to our discussion, with an emphasis on spatial and time diversity for wireless systems.

5.1.1.1. Single-Antenna Systems. The choice of convolutional codes is motivated by their simple optimal decoding structure, systematic construction of strong codes for large block lengths, and lower decoding delay compared to block codes. A convolutional code is generated by passing the information sequence through a linear finite-state shift register. In general, the shift register consists of S B-bit stages and m linear algebraic function generators; see Fig. 6 [4]. The input data to the encoder, assumed to be binary, are shifted into and along the shift register B bits at a time. The number of output bits for each B input bits is m bits. Consequently, the code rate is defined as $R_c = B/m$. The parameter S is called the *constraint length* of the convolutional code.

To understand the encoding procedure, consider the convolutional encoder for $S = 3$, $B = 1$, and $m = 3$ shown in Fig. 7 [4]. All the shift registers are assumed to be in

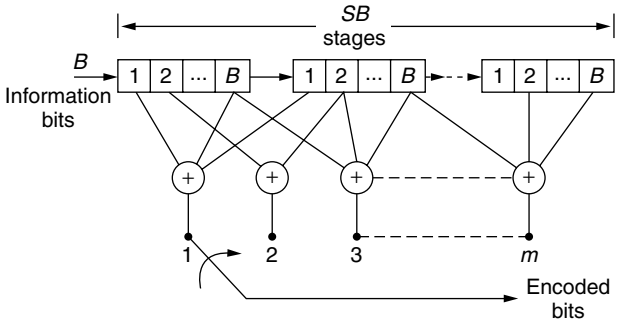


Figure 6. Convolutional encoder.

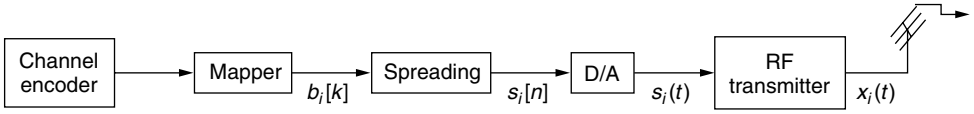


Figure 5. DS-CDMA handset transmitter components.

[12] Some of the networking layers use checksums for error detection and perform error correction by requesting retransmission of packets.

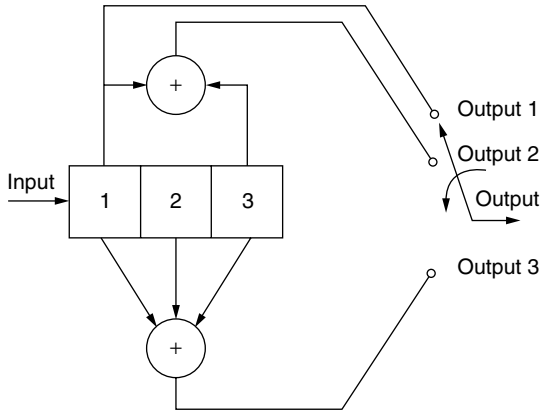


Figure 7. Convolutional encoder for a (3,1) code.

zero state initially. If the first input bit is a 1, the resulting output sequence of 3 bits is $[b[1]\ b[2]\ b[3]] = [1\ 1\ 1]$. Now, if the second input bit is a 0, the next three output bits are $[b[4]\ b[5]\ b[6]] = [0\ 0\ 1]$ (or else the output bits are [110] if the input bit is 1). If the third bit is a 1, the output is $[b[7]\ b[8]\ b[9]] = [1\ 0\ 0]$. The operation of a nonrecursive (Fig. 7) convolutional code is similar to that of a finite-impulse response (FIR) filter with all the operations done over a finite field; in Fig. 7, the finite field consists of only two elements $\{0, 1\}$ with binary addition. The convolutional code has one input and several outputs, equivalent to a single-input multiple-output FIR linear system. The equivalent of the impulse response of the filter is the *generator polynomial*, which succinctly describes the relation between output and shift register states for a convolutional code. For the example in Fig. 7, the generator polynomials are

$$\text{Output } 1 \rightarrow \mathbf{g}_1 = [1\ 0\ 0]$$

$$\text{Output } 2 \rightarrow \mathbf{g}_2 = [1\ 0\ 1]$$

$$\text{Output } 3 \rightarrow \mathbf{g}_3 = [1\ 1\ 1]$$

The generator polynomials of a convolutional code characterize its performance via different metrics, notably minimum distance and distance spectrum [60]. To design any code requires an appropriate metric space, which depends on the channel under consideration. For slowly block fading channels, the Euclidean distance between the codewords is the natural metric [60], while for fast fading channels, Hamming distance is the appropriate metric [61]; see discussion below for further discussion on diversity techniques.

Addition of redundant bits for improving the error probability leads to bandwidth expansion of the transmitted signal by an amount equal to the reciprocal of the code rate. For bandwidth constrained channels, it is desirable to achieve a coding gain with minimal bandwidth expansion. To avoid bandwidth expansion due to channel coding, the number of signal points over the corresponding uncoded system can be increased to compensate for the redundancy introduced by the code. For instance, if we intend to improve the performance of an uncoded system using BPSK modulation, a rate $\frac{1}{2}$ code would require doubling the number of signal points to quadrature phase shift keying (QPSK) modulation. However, increasing the number of signals leads to higher probability of error for the same average power. Thus, for the resultant bandwidth efficient scheme to provide gains over the uncoded system, it must be able to overcome the penalty due to increased size of the signal set.

If the modulation (mapping of the bits to channel signals) is treated as an operation independent of channel encoding, very strong convolutional codes are required to offset the signal set expansion loss and provide significant gains over the uncoded system [4]. On the other hand, if the modulation is treated as an integral part of channel encoding, and designed in unison with code to maximize the Euclidean distance between pairs of coded signals, the loss due to signal set expansion is easily overcome. The method of *mapping by set partitioning* [62] provides an effective method for mapping the coded bits into signal points such that the minimum Euclidean distance is maximized. When convolutional codes are used in conjunction with signal set partitioning, the resulting method is known as *trellis-coded modulation* (TCM). TCM is a widely used bandwidth efficient coding scheme with a rich associated literature; see Ref. 63 for a comprehensive in-depth review.

The fundamental channel coding theorem by Shannon [23] proved the existence of good codes, which can achieve arbitrarily small probability of error, as long as the transmission rate is lower than the channel capacity. The proof in Ref. 23 required creating codes that had continually increasing block sizes to achieve channel capacity. Another key component of the proof in Ref. 23 was the choice of codebooks, they were chosen at random. Random codes with large block sizes have no apparent structure to implement a physically tractable decoder. Proven optimality of random codes coupled with the inability to find good structured codes led to a common belief that the structured deterministic codes had a lower capacity than the channel capacity, often called the "practical capacity" [64,65]. The discovery of *turbo codes* [17] and the rediscovery of *low-density parity-check* (LDPC) codes [66] appears to have banished the abovementioned "practical capacity" myth. Both Turbo and LDPC codes have been shown to operate below the "practical capacity," within a tenth of a decibel of the Shannon capacity. Turbo codes have also been proposed for the third-generation wireless standards. The main ingredients of a turbo code are constituent codes (block or convolutional code) and a long interleaver. The long interleaver serves two purposes: lends codewords a "randomlike" structure, and leads to long codes that are easily and efficiently decoded using a (suboptimal yet effective) iterative decoding algorithm. Several extensions of turbo codes are areas of active research, notably, bandwidth-efficient Turbo codes [67,68], deterministic interleaver design [69] and spacetime Turbo codes [70].

We close the discussion on codes for slow fading Gaussian channels, by highlighting that none of the current codes come close to the lower bounds on the performance of codes [71]. Current codes require large block lengths to achieve small probability of decoded

message errors, but relatively short block lengths suffice to achieve the same level of performance for "good" codes [71]. Thus, the field of code design, although more than fifty years old, has still significant room to develop.

5.1.1.2. Multiple-Antenna Systems. The random time variations in the received signal provide diversity, which can be exploited for improved error performance. Typical forms of diversity include time, frequency, and spatial diversity. In Section 4.1, it was noted that diversity is important to improve the outage performance or achievable rates in fading channels. Although only spatial diversity using multiple transmit and receive antennas was studied in Section 4.1, similar benefits are also obtained by using time or frequency diversity or a combination of them. In time and frequency diversity, channel variations in time and across frequency are used to increase reliability of the received signal. In spatial diversity, multiple transmit and/or receive antennas exploit the random spatial time variations.

The codes designed for Gaussian channels can be used for slowly fading channels if an accurate channel estimator is available and all symbols of a codeword undergo the same channel fading. In the presence of medium to fast fading, where the coherence interval is shorter than a codeword, Hamming distance between the codewords should be maximized [61]. If channel variations are slower than a codeword, an interleaver is commonly used to induce time diversity. For interleaver-based schemes to be effective, the interleaver depths should be larger than the coherence interval; this implies that it is useful for fast-fading channels or for communications where large delay can be tolerated. For low-delay application, the interleaver-induced time diversity is not possible. In addition, if the channel is flat-fading (true for narrowband communications), then frequency diversity cannot be used, either. Irrespective of the availability of time and frequency diversity, the spatial diversity via multiple antennas is a promising method to achieve higher data rates.

Receiver diversity using multiple receive antennas is a well-understood concept [6] and often used in practice [72]. In contrast, using multiple antennas at the transmitter has gained attention only relatively recently due to discovery of space-time codes [5,31], motivated by encouraging capacity results [25,73]. Space-time coding exploits multiple independent channels between different transmit–receive antenna pairs in addition to time diversity (possibly interleaver induced). Later work [5] extended well-founded coding principles to spatial diversity channels, thereby simultaneously achieving coding gain and the highest possible spatial diversity. The space-time codes proposed there [5] have become a performance benchmark for all subsequent research in space-time coding [74–80]. The concept of transmitter diversity can be appreciated using the following elegant *Alamouti scheme* [81] for two transmit antennas.

In a given symbol period, two symbols are simultaneously transmitted from the two antennas. Denote the signal transmitted from antenna 1 as s_1 and from antenna 2 as s_2 (see Fig. 8). During the next symbol period, signal $-s_2^*$ is transmitted from antenna 1, and s_1^* is transmitted from antenna 2. Note that the encoding of symbols is done in both space and time. As is evident from Fig. 8, the received signal in any symbol interval is a linear combination of the signals transmitted from the two antennas. Thus, the spacetime channel is an interference channel. An analogous scenario exploiting frequency diversity would use nonorthogonal carrier frequency to send two symbols in each symbol period. The Alamouti scheme sends orthogonal signals over two time instants from the two antennas, where vector $[s_1 - s_2^*]$ transmitted from antenna 1 over two time symbols is orthogonal to the vector $[s_2 \;\; s_1^*]$ transmitted from antenna 2. If the channel stays constant over two consecutive symbol periods, then the orthogonality is maintained at the receiver. Since each symbol s_1 and s_2 is transmitted from both the antennas, they travel to the receiver from two different channels, which provides the desired diversity order of two. The orthogonality of the time signals helps resolution of the two symbols at the receiver without affecting the diversity order.

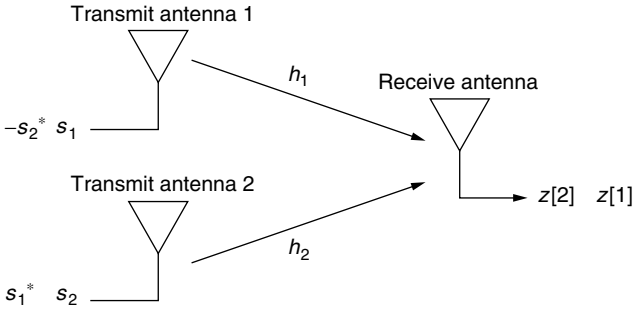


Figure 8. Alamouti encoder for two transmit and one receive antenna.

The Alamouti scheme can be extended to more than two transmit antennas using the theory of orthogonal designs [74]. The Alamouti scheme is a rate 1 code and thus requires no bandwidth expansion. But it provides a diversity order of two, which is twice that of any rate 1 single-antenna system. The Alamouti scheme has a very simple optimal receiver structure, thereby making it a prime candidate for practical implementations. In addition to its simplicity, the Alamouti scheme-based systems do not lose in their asymptotic performance. It has been shown [79] that orthogonal transmit diversity schemes are capacity-achieving, and this provided a motivation for the concatenated space-time coding methods [79,80]. The concatenated space-time codes decouple the spatial and temporal diversity to simplify the space-time code design.

All third-generation systems have adopted some form of transmit and receive diversity. Multiple antennas at the base station are relatively easier to implement in comparison to multiple antennas at the mobile handset, due to size limitation. Two cross-polarized antennas have been proposed and tested for mobile handsets [82].

5.1.2. Spreading and Modulation. The binary output of the error control encoder is mapped to either ± 1 to obtain the sequence $b_i[k]$, which is multiplied by a spreading sequence, $c_i[n] \in \{-1, 1\}$, of length N; the spreading operation is shown in Fig. 9. After spreading the

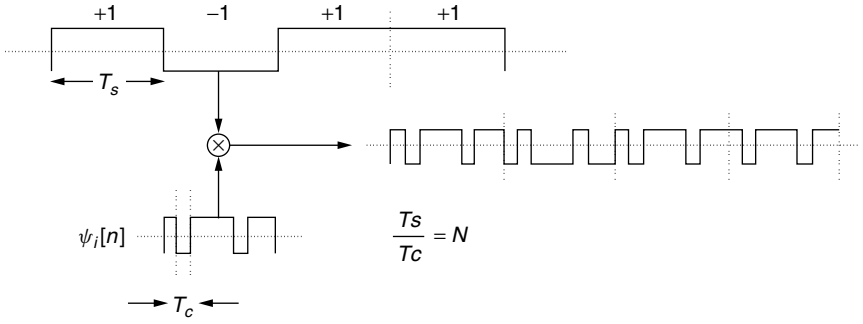


Figure 9. The spreading operation.

signal, the signal is passed through a digital pulseshaping filter, $\phi[n]$, which is typically a square-root raised-cosine filter [4]. The pulse shaping filter is chosen to limit the bandwidth of the transmitted signal to the available spectrum, while minimizing the intersymbol interference (ISI) caused by the filter. The digital signal for user i after pulse shaping can be written as

$$s_i[n] = \sum_{k=1}^{G} b_i[k]\psi_i[n - kNL] \tag{18}$$

where L is the number of samples per chip and $\psi_i[n] = \phi[n] * c_i[n]$, where $*$ represents linear convolution, and G is the number of bits in the packet.

After converting the digital signal to analog using a D/A converter, the RF upconverter shifts the baseband analog signal to the carrier frequency f_c. The upconverted signal is amplified by a power amplifier and transmitted via an antenna. The transmitted passband signal assumes the form

$$\begin{aligned} x_i(t) &= \sqrt{\mathcal{P}_i} e^{-j\omega_c t} \sum_{k=1}^{G} b_i[k]\psi_i(t - kT_s) \\ &= e^{-j\omega_c t} s_i(t) \end{aligned} \tag{19}$$

where T_s is the symbol period and $\mathcal{P}_i$ is the transmitted power. The bits $b_i[k]$ are the output of a suitable channel encoder discussed in Section 5.1.1. Since CDMA signals at the base station typically have large peak to average power ratios, the operating point of the power amplifier is kept low to avoid amplifier nonlinearities. The amplifier nonlinearities are avoided for several important reasons: (1) RF amplifier efficiency is lower in nonlinear region, which increases the power loss and hence total power consumed by the transmitter; (2) the nonlinearity introduces higher spectral components, which can cause increased interference in the neighboring frequency bands; and (3) the algorithm design for resulting nonlinear systems becomes intractable.

As discussed in Section 4.1, multiple antennas at the transmitter and receiver can lead to large gains in fading wireless channels [21,25,37]. If multiple transmit antennas are used, the vector transmitted passband signal is given by

$$\mathbf{x}_i(t) = \sqrt{\frac{\mathcal{P}_i}{M}} e^{-j\omega_c t} \sum_{k=1}^{G} \mathbf{b}_i[k]\psi_i(t - kT) \tag{20}$$

where M is the number of transmit antennas. The $M \times 1$ vectors, $\mathbf{x}_i(t)$ and $\mathbf{b}_i[k]$, represent the transmitted vector signal and spacetime-coded signal, respectively. In (20), we have assumed that the transmitter has no knowledge of the channel and hence uses the same average power on each transmitter. If the transmitter "knows" the channel, then the power across different antennas can be adapted to achieve an improved performance [25,83].

5.2. Base-Station Receiver

In cellular systems, the time and spectral resources are divided into different logical *channels*. The generic logical channels are broadcast, control, random-access, paging, shared, and dedicated channels [84,85]. All logical channels are physically similar and the distinction is solely made based on the purpose served by each channel. In the sequel, we will consider only the dedicated and shared channels, since they carry most of the user data and hence impose the biggest computational bottleneck. Implementation details of other channels can be found elsewhere [84,85].

As noted in Section 2, the unknown time-varying multipath is one of the biggest challenges in the design of wireless systems. Optimal transmission schemes that do not require knowledge of the wireless channel at the receiver can be designed using information theoretic tools (see Ref. 86 and the references therein), but are seldom employed. The primary reason for not using optimal strategies is their high computational complexity, and large latency of the resulting communication method. Hence, suboptimal and computationally efficient solutions are generally employed. The receiver estimates the unknown channel, and then uses the channel estimate to decode the data using a channel decoder.

A simplified illustration of the baseband receiver is shown in Fig. 10. The key components of the receiver are multiuser channel estimation, multiuser detection, and single-user channel decoding. Most systems also provide feedback from the receiver for power control and automatic repeat request (ARQ) to improve system reliability. The choice of algorithms used in each of the blocks is determined by their computational complexity, desired performance level, and the available side information. Mobile units are power- and complexity-constrained, and have little or no knowledge of the multiple access interference. On the other hand, the base stations are equipped with higher processing power and detailed

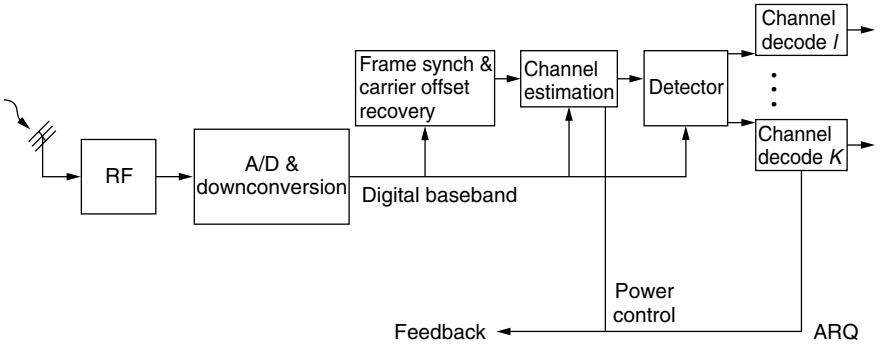


Figure 10. Base-station receiver structure.

information about all in-cell users, thereby allowing more sophisticated processing at the base stations. Our discussion will focus on base-station algorithms in the following section, with only bibliographic references to relevant counterparts for the mobile handset.

5.2.1. Received Signal. For each active user in a cell, the received signal at the base station consists of several unknown time-varying parameters. These parameters include propagation delay, amplitude, delay and number of paths, and residual carrier offset. The time variation in propagation delay is caused as users move closer or away from the base-station. The mobility of the users or the surrounding environment also causes time variation in the multipath environment. Finally, drift in the local oscillator frequencies of the transmitter and receiver leads to a residual carrier offset at the baseband.

Using the model (4) for the multipath channel impulse response and assuming that the channel coefficients for the ith user $h_{p,i}$ are constant over the observation interval, the received signal for a transmitted signal $x_i(t)$ without additive white noise is given by

$$\begin{aligned} z_i(t) &= \sum_{p=1}^{P} h_{p,i} x_i \left(t - \tau_i - \frac{p}{W}\right) \\ &= \sqrt{\mathcal{P}_i} e^{-j\omega_c t} \sum_{p=1}^{P} \underbrace{h_{p,i} e^{j\omega_c(\tau_i + p/W)}}_{a_{p,i}} \\ &\quad \times \sum_{k=1}^{G} b_i[k]\psi_i \left(t - kT - \tau_i - \frac{p}{W}\right) \\ &= \sqrt{\mathcal{P}_i} e^{-j\omega_c t} \sum_{k=1}^{G}\sum_{p=1}^{P} b_i[k] a_{p,i} \psi_i \left(t - kT - \tau_i - \frac{p}{W}\right) \end{aligned} \tag{21}$$

where τ_i is the propagation delay of the received signal. If the number of paths $P = 1$, then it is a flat-fading channel else a frequency selective channel. The received signal is amplified and downconverted to baseband. In practice, there is a small difference in the frequencies of the local oscillators at the transmitter and the receiver. The received baseband signal after downconversion (without additive noise) is given by

$$z_i(t) = \sqrt{\mathcal{P}_i} e^{-j\Delta\omega_i t} \sum_{k=1}^{G}\sum_{p=1}^{P} b_i[k] a_{p,i} \psi_i \left(t - kT - \tau_i - \frac{p}{W}\right) \tag{22}$$

where $\Delta\omega_i$ represents the residual carrier frequency offset. Assuming that the carrier offset $\Delta\omega_i$ is negligible or is corrected using a multiuser equivalent of digital phase-locked loop [4,87], the sampled baseband (without noise) with L samples per chip can thus be written as

$$z_i[n] = \sum_{k=1}^{G}\sum_{p=1}^{P} b_i[k] a_{p,i} \psi_i[n - kNL - \tau_i - p] \tag{23}$$

In general, the receiver components introduce thermal noise, which is generally modeled as additive noise. For K simultaneously active users, the received baseband signal in the presence of thermal noise at the base station is

$$z[n] = \sum_{i=1}^{K} z_i[n] + \nu[n] \tag{24}$$

The additive component $\nu[n]$ in (24) is generally modeled as white Gaussian noise. The received signal model in Eqs. (13) and (24) are similar; both consider a sum of all user signals in additive noise. The main difference is the assumption on the fading statistics; a flat-fading model is assumed in (13) compared to a multipath model in (24).

In the sequel, we will focus on estimating the unknown channel coefficients and subsequent detection of the data bits, $b_i[k]$ for all users $i = 1, \ldots, K$. The development of multiuser channel estimation and data detection is greatly simplified by using linear algebraic methods. We will write the received signal (24) using matrix–vector notation in two different forms. The first form will be used in multiuser channel estimation methods, and the second in multiuser detection.

5.2.1.1. Channel as Unknown. For simplicity, we will assume that all τ_i are multiple of sampling instants, $\tau_i = l_i$; for the general case, the reader is referred to Ref. 88. Let $u_i[n] = \sum_{k=1}^{G} b_i[k]\psi(n - kNL)$. Then the received signal $z_i[n]$

can be rewritten in matrix–vector notation [89] as

$$\mathbf{z}_i = \begin{bmatrix} u_i[1] & 0 & 0 & \cdots & 0 \\ u_i[2] & u_i[1] & 0 & & 0 \\ u_i[3] & u_i[2] & u_i[1] & & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \vdots & u_i[GLN + l_\phi] \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ a_{1,i} \\ \vdots \\ a_{P,i} \end{bmatrix}$$

$$= \mathbf{U}_i \mathbf{a}_i, \tag{25}$$

where there are l_i leading zeros in the channel vector $\mathbf{a}_i$ to account for the propagation delay, and l_ϕ is the length of the pulse ϕ (measured in number of samples). The total received signal can thus be written as

$$\mathbf{z} = [\,\mathbf{U}_1 \quad \mathbf{U}_2 \quad \cdots \quad \mathbf{U}_K\,] \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_K \end{bmatrix} + \boldsymbol{\nu}$$

$$= \mathbf{U}\mathbf{a} + \boldsymbol{\nu} \tag{26}$$

where we recall that N is the spreading gain, L is the number of samples per chip, and G is the number of bits in the packet. The signal model described above will be used to derive channel estimation algorithms in Section 5.2.2.

5.2.1.2. Data as unknown. Define $q_i[n] = \sum_{p=1}^{P} a_{p,i}\,\psi[n - kNP - l_i - p]$; $q_i[n]$ can be understood as the *effective* spreading waveform for the ith user. The waveform $q_i[n]$ is generally longer than one symbol period and hence causes interference between the consecutive symbols. To highlight the presence of intersymbol interference (ISI), we will write the received signal $z_i[n]$ for every symbol duration. For simplicity, we will assume that the length of $q_i[n], l_q$, is less than two symbol durations, namely, $l_q < 2NL$. Then the received signal $z_i[n]$ can be written as

$$\mathbf{z}_i[k] = \begin{bmatrix} 0 & q_i[1] \\ 0 & q_i[2] \\ \vdots & \vdots \\ q_i[NL+1] & q_i[2NL - l_q + 1] \\ \vdots & \vdots \\ q_i[l_q] & q_i[NL] \end{bmatrix} \begin{bmatrix} b_i[k-1] \\ b_i[k] \end{bmatrix}$$

$$= \mathbf{Q}_i \mathbf{b}_i \tag{27}$$

The total received signal can be written as

$$\mathbf{z}[k] = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 & \cdots & \mathbf{Q}_K \end{bmatrix} \begin{bmatrix} \mathbf{b}_1[k] \\ \mathbf{b}_2[k] \\ \vdots \\ \mathbf{b}_K[k] \end{bmatrix} + \boldsymbol{\nu} \tag{28}$$

$$= \mathbf{Q}\mathbf{b}[k] + \boldsymbol{\nu} \tag{29}$$

The received signal model in (29) clearly demonstrates the challenges in multiuser detection. Not only does the receiver have to cancel the multiple access interference, but also the ISI for each user introduced by the multipath channel. The ISI acts to increase the effective multiple access interference experienced by each bit. The multiuser detection methods aim to jointly make all bits decisions $\mathbf{b}[k]$.

In the following section, we will discuss channel estimation, multiuser detection and channel decoding algorithms for DS-CDMA systems.

5.2.2. Multiuser Channel Estimation. Most channel estimation can be divided into two broad classes: training based and blind methods. In each class, a further subdivision[13] is made on the basis of assumptions made regarding the multiple access interference: single-user channel estimation in the presence of multiple access interference or jointly estimating channels for all the users.

Most wireless systems add known symbols periodically to the data packets. The known data symbols are known as *training symbols* and facilitate coarse synchronization, channel estimation and carrier offset recovery. Training-based methods simplify estimation of unknown baseband parameters at the cost of throughput loss; symbols used for training could potentially be used to send more information bits. The amount of training depends on the number of simultaneous users, the number of transmit antennas [28], and the desired reliability of channel estimates. Given the training symbols and assuming perfect carrier offset recovery, multiuser channel estimation can be cast as a linear estimation problem [91], and admits a closed-form solution. The work in Ref. 91 also discusses extensions to multiple antennas.

A class of blind channel estimation procedures, collectively known as *constant modulus algorithms* (CMA), were first proposed [92,93] using the constant amplitude property of some of the communication signals such as BPSK. The CMA algorithms use a nonlinear (nonconvex) cost function to find the channel estimate, and hence can converge to poor estimates. An alternate procedure of blind estimation was proposed [94,95], which used the cyclostationarity of the communication signals. Motivated by another method [94], a single-user blind channel identification method, using only second-order statistics, was proposed [96]. The blind channel equalization exploits only the second (or higher)-order statistics without requiring periodic training symbols, with an assumption that the data symbols are independent and identically distributed. The assumption of i.i.d. data is rarely correct because of channel coding used in almost all systems. Hence, the results based on blind channel estimation should be interpreted with caution. Nonetheless, there is value in exploring blind channel identification methods. Blind estimation can improve the estimates based on training or completely avoid the use of training symbols; the reader is referred elsewhere [97,98] for results on single-user systems.

[13] Another possible subdivision can be based on linear and nonlinear algorithms. An example of feedback-based nonlinear algorithm is the decision-feedback-based equalization [90].

Single-user channel estimation in the presence of unknown multiple access interference has been addressed [99]. An approximate maximum-likelihood channel estimation for multiple users entering a system has been presented [100]; the estimate-maximize algorithm [101] and the alternating projection algorithm [102] in conjunction with the Gaussian approximation for the multiuser interference were used to obtain a computationally tractable algorithm. Blind multiuser channel estimation has also been addressed in several papers [103,104], with an assumption of coarse synchronization.

Most of the current work, with a few exceptions [105–108], assume square pulse shaping waveforms leading to closed-form optimistic results; see Ref. 107 for a detailed discussion. Furthermore, very little attention has been paid to carrier offset recovery in a multiuser system, except for the results reported in the paper [87]. In this section, we will only discuss channel estimation at the base-station, assuming coarse synchronization and perfect downconversion. For handset channel estimation algorithms, the reader is referred elsewhere [107,109]. Additionally, we restrict our attention to only training-based methods; blind techniques are rarely used in wireless systems.[14] The channel model assuming T training symbols for each user can be written as

$$\mathbf{z} = \mathbf{U}\mathbf{a} + \boldsymbol{\nu} \tag{30}$$

where the size of the vectors $\mathbf{z}$ and $\boldsymbol{\nu}$, and matrix $\mathbf{U}$ is appropriately redefined for an observation length of T symbols, using the definition in (26). The matrix $\mathbf{U}$ depends on the spreading codes, $\phi_i[n]$ and the training symbols $b_i[k]$, all of which are assumed known for all users. Thus, the matrix $\mathbf{U}$ is completely known. The maximum-likelihood estimate of the channel coefficients, $\mathbf{a}$, is given by the pseudoinverse [4,91]:

$$\hat{\mathbf{a}} = (\mathbf{U}^H\mathbf{U})^{-1}\mathbf{U}^H\mathbf{z}. \tag{31}$$

This solution retains several desirable statistical properties of the maximum-likelihood estimates for linear Gaussian problems [110], namely, consistency, unbiasedness and efficiency. Note that there are several leading zeros in $\mathbf{a}$. The variance of the maximum-likelihood estimator $\hat{\mathbf{a}}$ can be reduced by detecting the unknown number of leading (and possibly trailing) zeros in $\mathbf{a}$, which reduces the number of estimated parameters. The above channel estimation procedure can also easily be extended to long-code DS-CDMA systems [111]. In practice the additive noise $\boldsymbol{\nu}$ is better modeled as colored Gaussian noise with unknown covariance due to out-of-cell multiuser interference. The maximum-likelihood estimate of $\mathbf{a}$ requires estimation of the unknown covariance, thereby leading to more accurate results compared to (31) at the expense of increased computation [89].

[14] A notable exception is high-definition television (HDTV) transmission, where no resources are wasted in training symbols, and slow channel time variation permit the use of blind estimation techniques.

Having estimated the channel for all the users, the channel estimates are then used to detect the rest of the information bearing bits in the packet. For bit detection, the received signal representation in (29) is more appropriate, where the matrix $\mathbf{Q}$ is formed using the channel estimates $\hat{\mathbf{a}}$ and the user signature waveforms $\psi_i[n]$.

5.2.3. Multiuser Detection. As a result of channel-induced imperfections and time-varying asynchronism between the users, it is practically impossible to maintain orthogonality between the user signals. *Multiple-access interference* (MAI) is caused by the simultaneous transmission of multiple users, and is the major factor that limits the capacity and performance of DS-CDMA systems. In the second generation CDMA standards, the multiple-access interference is treated as part of the background noise and single-user optimal detection strategy is used. The single-user receiver is prone to the *near–far* problem, where a high-power user can completely drown the signal of a weak user. To avoid the near–far problem, CDMA-based IS-95 standard uses tight power control to ensure that all users have equal received power. Even with the equal received power, the output of the single-user detector is contaminated with MAI and is suppressed by using very strong forward error correcting codes.

The MAI is much more structured than white noise, and this structure was exploited [112] to derive the optimal detector that minimizes the probability of error. The optimal detector alleviates the near–far problem that plagues the single-user receiver. The optimal detector, thus, does not require fast power control to achieve a desired level of performance, thereby reducing the system overhead greatly. Further, as the number of users increases, the optimal receiver achieves significant gains over single-user receivers, even with perfect power control. Unfortunately, the optimal receiver is computationally too complex to be implemented for large systems [113]. The computational intractability of multiuser detection has spurred a rich literature on developing low-complexity suboptimal multiuser detectors.

Most of the proposed suboptimal detectors can be classified in one of two categories: linear multiuser detectors and subtractive interference cancellation detectors. Linear multiuser receivers linearly map the soft outputs of single-user receivers to an alternate set of statistics, which can possibly be used for an improved detection. In subtractive interference cancellation, estimates for different user signals are generated and then removed from the original signal.

To gain insight into different methods for multiuser detection, we will limit the discussion in this section to a simple case of no multipath and no carrier frequency errors. We further assume that the pulse-shaping introduces no ISI and all users are synchronous, thereby leading to simplification of (29) as

$$\mathbf{z}[k] = \mathbf{Q}\mathbf{b}[k] + \boldsymbol{\nu}[k] \tag{32}$$

where $\mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_K]$, $\mathbf{q}_i = [q_i[1]\ q_i[2]\ \cdots\ q_i[NP]]^T$, and $\mathbf{b}[k] = [b_1[k]\ b_2[k]\ \cdots\ b_K[k]]^T$. Note that this simplification

only eliminates ISI, not the multiple access interference, which is the primary emphasis of the multiuser detection. We quickly note that all the subsequently discussed multiuser detection methods can be extended to the case of asynchronous and ISI channels. The code matched-filter outputs, $\mathbf{y}[k] = \mathbf{Q}^H\mathbf{z}[k]$ can be written as

$$\mathbf{y}[k] = \mathbf{R}\mathbf{b}[k] + \boldsymbol{\nu}[k] \tag{33}$$

The $K \times K$ matrix $\mathbf{R} = \mathbf{Q}^H\mathbf{Q}$ is the correlation matrix, whose entries are the values proportional to the correlations between all pairs of spreading codes. The matrix $\mathbf{R}$ can be split into two parts, $\mathbf{R} = \mathbf{D} + \mathbf{O}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \mathcal{P}_i$. Thus (33) can be written as follows:

$$\mathbf{y}[k] = \mathbf{D}\mathbf{b}[k] + \mathbf{O}\mathbf{b}[k] + \boldsymbol{\nu}[k] \tag{34}$$

The matrix $\mathbf{O}$ contains the off-diagonal elements of $\mathbf{R}$, with entries proportional to the cross-correlations between different user codes. The first term in (34), $\mathbf{b}[k]$, is simply the decoupled data of each user and the second term, $\mathbf{O}\mathbf{b}[k]$, represents the MAI.

5.2.3.1. Matched-Filter Detector. Also known as *single-user optimal receiver*, the matched-filter receiver treats the MAI + $\boldsymbol{\nu}[k]$ as white Gaussian noise, and the bit decisions are made by using the matched-filter outputs, $\mathbf{y}[k]$. The hard bit decisions are made as

$$\hat{\mathbf{b}}_{\mathrm{MF}}[k] = \mathrm{sign}(\mathbf{y}[k]) \tag{35}$$

where sign$(\cdot)$ is a nonlinear decision device and outputs the sign of the input. The matched-filter receiver is extremely simple to implement and requires no knowledge of MAI for its implementation. However, the matched-filter receiver suffers from the near–far problem, where a nonorthogonal strong user can completely overwhelm a weaker user; in fading environments, power disparities are commonly encountered and perfect power control is generally impossible.

5.2.3.2. Maximum A Posteriori Probability (MAP) Detector. As the name suggests, the maximum–likelihood detector chooses the most probable sequence of bits to maximize the joint a posteriori probability, the probability that particular bits were transmitted having received the current signal: Prob($\mathbf{b}[k]|\mathbf{r}(t)$, for all t). The MAP detector minimizes the probability of error [112]. Under the assumption that all bits are equally likely, the MAP detector is equivalent to the maximum–likelihood detector, which finds the bits $\mathbf{b}[k]$ that maximize the probability Prob($\mathbf{r}(t)|\mathbf{b}[k]$).

For the case of K synchronous users in (32), there are 2^K possible transmitted bit combinations in each received symbol duration. Thus, the computation of the maximum–likelihood bit estimates requires number of operations proportional to 2^K. For large number of users, the number of operations to obtain maximum–likelihood estimates become prohibitive for real-time implementation.

In the general case of asynchronous users, if a block of $M \leq G$ bits per user is used to perform the detection, there are 2^{MK} possible bit decisions, $\{\mathbf{b}[k]\}_{k=1}^{M}$. An exhaustive search over all possible bit combinations is clearly impractical, even for moderate values of M and K. However, the maximum-likelihood detector can be implemented using the Viterbi algorithm [114]; the Viterbi implementation (see Section 5.2.4 for more details on Viterbi decoding) is similar to maximum-likelihood sequence detection for ISI channels [4]. The resulting Viterbi algorithm has a complexity that is linear in block length M and exponential in the number of users, of the order of $M2^K$.

The maximum-likelihood detector requires complete knowledge of all user parameters that include not only the spreading signatures of all users but also their channel parameters. The channel parameters are unknown a priori, and have to be estimated. Despite the huge performance and capacity gains of the maximum-likelihood detector, it remains impractical for real-time systems. The computational intractability of the ML detector has led to several detectors which are amenable to real-time implementation.

5.2.3.3. Linear Detectors. Linear detectors map the matched filter outputs, $\mathbf{y}[k]$, in Eqn. (33) into another set of statistics to reduce the MAI experienced by each user. Two of the most popularly studied matched-filter receivers are the decorrelating detector and minimum mean-squared error (MMSE) detector.

The **decorrelating detector** was proposed in 1979 and 1983 [115,116] and later analyzed [117,118]. The decorrelating detector uses the inverse of the correlation matrix, $\mathbf{R}^{-1}$, to decouple the data of different users. The output of the decorrelating detector before hard decision is given by

$$\hat{\mathbf{b}}_{dec}[k] = \mathbf{R}^{-1}\mathbf{y}[k] \tag{36}$$

$$= \mathbf{b}[k] + \mathbf{R}^{-1}\boldsymbol{\nu}[k] \tag{37}$$

$$= \mathbf{b}[k] + \boldsymbol{\nu}_{dec}[k] \tag{38}$$

The decorrelating detector completely suppresses the MAI at the expense of reduced signal power.[15] For nonmultipath channels and unknown user amplitudes, the decorrelating detector yields optimal maximum-likelihood estimates of the bits and the received amplitudes. The decorrelating detector leads to substantial performance improvements over the single-user detector [118] if the background noise is low compared to the MAI. In addition to the noise enhancement problem, the computational complexity of the decorrelating detector can be prohibitive to implement in real-time; however, dedicated application-specific integrated circuits (ASICs) can ameliorate the real-time implementation issues. The computational complexity of the decorrelating detector prohibits its use for long-code CDMA systems, since it requires recomputation of $\mathbf{R}^{-1}$ for every bit.

The **MMSE detector** [119] accounts for the background noise and the differences in user powers to suppress the MAI. The detector is designed to minimize the

[15] The decorrelating detector is very similar to the zero-forcing equalizer [4], which is used to completely suppress ISI.

mean-squared error between the actual data, $\mathbf{b}$ and the soft estimate of data, $\hat{\mathbf{b}}_{\text{mmse}}$. The MMSE detector hard limits the following transform of the received signal:

$$\hat{\mathbf{b}}_{\text{mmse}} = (\mathbf{R} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}[k]. \tag{39}$$

The MMSE detector[16] balances between the suppression of MAI and suppression of background noise. The higher the background noise level, the lesser is the emphasis on suppressing MAI and vice versa. The MMSE detector has been shown to have a better probability of error than the decorrelating detector [12]. It is clear that as the background noise goes to zero, the MMSE detector converges to the decorrelating detector. On the other hand, as the background noise becomes more dominant compared to MAI, the MMSE detector converges to a single-user detector. Unlike the decorrelator and single-user receiver, the MMSE detector requires an estimate of user amplitudes. Further, the complexity of the MMSE detector is similar to that of the decorrelator.

A blind extension of the MMSE detector, which does not require the knowledge of other user codes and parameters, has been presented [120]. The blind MMSE is similar to the commonly used beamformer in antenna array processing [121]. The probability of error performance of the MMSE detector was studied [122]. The MMSE estimator was extended to multiple data rate systems, like the third-generation standards [123,124].

5.2.3.4. Subtractive Interference Cancellation. The basic idea in subtractive interference cancellation is to separately estimate the MAI contribution of each user and use the estimates to cancel a part or all the MAI seen by each user. Such a detector structure can be implemented in multiple stages, where each additional stage is expected to improve the accuracy of the decisions. The bit decisions used to estimate MAI can be hard (after the sign$(\cdot)$ operation) or soft (before the sign$(\cdot)$ operation). The nonlinear hard-decision approach uses the bit decisions and the amplitude estimates of each user to estimate the MAI. In the absence of reliable estimates, the hard-decision detectors may perform poorly as compared to their soft-decision counterparts [125,126].

The **successive interference cancellation** (SIC) detector cancels interference serially. At each stage of the detector, bit decisions are used to regenerate a user signal and cancel out the signal of one additional user from the received signal. After each cancellation, the rest of the users see a reduced interference. The SIC detector is initialized by ranking all the users by their received power. For the following discussion, assume that the subscripts represent the user rank based on their received powers. The received signal corresponding to user 1 is denoted by $z_1[n]$ [cf. (32)], and its bit estimate is denoted by $b_1[n]$. The SIC detector includes the following steps:

1. Detect the strongest user bit, $b_1[k]$, using the matched-filter receiver.
2. Generate an estimate, $\hat{z}_1[n]$, of the user signal based on the bit estimate, $b_1[k]$, and the channel estimate.
3. Subtract $\hat{z}_1[n]$ from the received signal $z[n]$, yielding a signal with potentially lower MAI.
4. Repeat steps 1–3 for each of the successive users using the "cleaned" version of the signal from the previous stage.

Instead of using the hard bit estimates, $\hat{b}_i[k]$, soft bit estimates (without the sign operator) can also be used in step 3. If reliable channel estimates are available, hard-decision SIC generally outperforms the soft-decision SIC; the situation may reverse if the channel estimates have poor accuracy [125,126]. The reasons for canceling the signals in descending order of received signal strength are as follows: (1) acquisition of the strongest user is the easiest and has the highest probability of correct detection; (2) removal of the strongest user greatly facilitates detection of the weaker users — the strongest user sees little or no interference suppression, but the weakest user can potentially experience a huge reduction in MAI; and (3) SIC is information-theoretically optimal, that is, optimal performance can be achieved using SIC [127].

The SIC detector can improve the performance of the matched-filter receiver with minimal amount of additional hardware, but SIC presents some implementation challenges: (1) each stage introduces an additional bit delay, which implies that there is a tradeoff between the maximum number of users that are canceled and the maximum tolerable delay [128]; and (2) time variation in the received powers caused by time-varying fading requires frequent reordering of the signals [128]. Again, a tradeoff between the precision of the power ordering and the acceptable processing complexity has to be made.

Note that the performance of SIC is dependent on the performance of the single-user matched filter for the strongest users. If the bit estimates of the strongest users are not reliable, then the interference due to the stronger users is quadrupled in power (twice the original amplitude implies 4 times the original power). Thus, the errors in initial estimates can lead to large interference power for the weaker users, thereby amplifying the near-far effect. So, for SIC to yield improvement over the matched filter, a certain minimum performance level of the matched-filter is required.

In contrast to the SIC detector, the **parallel interference cancellation** (PIC) detector [129] estimates and cancels MAI for all the users in parallel. The PIC detector is also implemented in multiple stages:

1. The first stage of the PIC uses a matched-filter receiver to generate bit estimates for all the users, $\hat{\mathbf{b}}_{\text{MF}}[k]$.
2. The signal for the matched filter for user i in the next stage is generated as follows. Using the effective spreading codes and the bit estimates of all except the ith user, the MAI for user i is generated and subtracted from the received signal, $r[n]$.
3. The signal with canceled MAI is then passed to the next stage, which hopefully yields better bit estimates.

[16] The MMSE detector is similar to the MMSE linear equalizer used to suppress ISI [4].

4. Steps 1–3 can be repeated for multiple stages. Each stage uses the data from the previous stage and produces new bit estimates as its output.

The output of $(m+1)$st stage of the PIC detector can be concisely represented as

$$\hat{\mathbf{b}}^{(m+1)}[k] = \text{sign}\,(\mathbf{y}[k] - \mathbf{O}\hat{\mathbf{b}}^{(m)}[k]) = \text{sign}\,(\mathbf{D}\mathbf{b}[k] + \mathbf{O}(\mathbf{b}[k] - \hat{\mathbf{b}}^{(m)}[k]) + \boldsymbol{\nu}[k]) \quad (40)$$

The term $\mathbf{O}\hat{\mathbf{b}}^{(m)}[k]$ is the estimate of MAI after the mth stage. Since soft-decision SIC exploits power variation by canceling in the order of signal strength, it is superior in a non-power-controlled system. On the other hand, soft-decision PIC has a better performance in a power-controlled environment. Performance evaluation of soft-decision PIC can be found elsewhere [130,131], as well as comparison of the soft-decision PIC and SIC detectors [130].

The susceptibility of the PIC to the initial bit estimates has been discussed [129]. An improved PIC scheme, which uses a decorrelator in the first stage, has been proposed [132]. The decorrelator-based PIC detector provides significant performance gains over the original PIC scheme. Further improvements to PIC detector's performance can be obtained by linearly combining the outputs of different stages of the detector [133].

For long-code systems, multistage detection is best suited for its good performance–complexity tradeoff. Multistage detection requires only matrix multiplications in each processing window while other multiuser detectors such as the decorrelator and MMSE detector require matrix inversions during each processing window due to the time-varying nature of the spreading codes.

5.2.4. Channel Decoding. Following the multiuser detection, the detected symbols are decoded using a channel decoder to produce an estimate of the transmitted information bits. In this section we will review decoders for FEC coding when the sender uses either one or more than one transmit antenna. For single-antenna systems, we will consider Viterbi decoding [134] of convolutional codes and review its lower complexity approximations. For multiple antennas, the ML decoder for the Alamouti scheme is presented along with a discussion on complexity of decoding space-time trellis codes.

5.2.4.1. Single-Transmit Antenna. The detected bits after the multiuser detection can be treated to be free of multiple-access interference, and hence a single-user channel decoder can be used. Viterbi decoding for convolutional codes is an application of the dynamic programming principle, and allows efficient hard- or soft-decision decoding of convolutional codes. Furthermore, Viterbi decoding is amenable to VLSI implementation.

To understand the decoding of a convolutional code, an alternate representation for the encoding process, known as a *trellis diagram*, is better suited. A convolutional code is a finite-state machine, whose next state and output are completely determined by its current state and input. The states of a convolutional code can be depicted using a

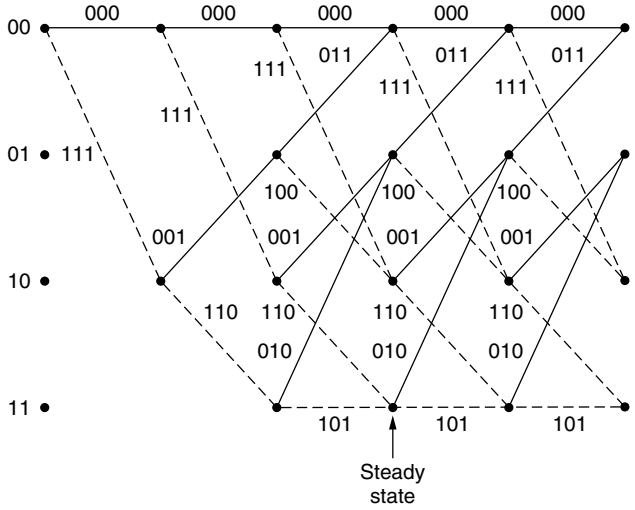


Figure 11. Trellis diagram for the example (3,1) convolutional code.

trellis diagram. The trellis diagram for the example code in Fig. 7 is given in Fig. 11. A close examination of the trellis diagram in Fig. 11 reveals that the diagram repeats itself after three stages, which is equal to the constraint length of the code, $S = 3$. In fact, the three outputs are completely determined by the first two states of the system and the input, which explains the four possible states (00, 01, 10, and 11) in the trellis and the two possible transitions from the current state to the next state based on the input (0 or 1). The solid transitions are due to input 0, and the dashed line shows transition due to input 1. The numbers along the transition describe the output of the decoder due to that transition.

Assume that κ encoded bits were sent using a rate R convolutional code; note that κ can be less than the packet length G if a training sequence is sent in the packet for channel estimation. The maximum a posteriori decoder chooses the information bit sequence that maximizes the posterior probability of the transmitted information symbols given the received noise corrupted signal. To compute the exact estimate of the transmitted information symbols, a total of $2^{\kappa R}$ bits should be considered. It was shown that, as a result of the encoding structure of the convolutional codes, the optimal decoder has a complexity which is linear in the codeword length κ [134]. In the Viterbi algorithm, a metric is associated with each branch of the code trellis. The metric associated with a branch at a particular stage or level i, is the probability of receiving r_i, when the output corresponding to that branch is transmitted. A *path* is defined as a sequence of branches at consecutive levels so that the terminal node of a branch ends in the source node of the next branch. The metric associated with a path is the sum of the metrics associated with the branches in the path. And the metric associated to a node is the minimum metric associated with any path starting from the start node to that node. With these associations, the MAP codeword corresponds to the path that has the lowest metric from the start node to the final node. If the decoder starts and ends in state 0, with start level labeled 0 and end level labeled κ, then for

all $0 < l < \kappa$, the defining equation in the optimization problem is

$$metric(0, \kappa) = \min_{m \in \text{states}} (metric(0, l_m) + metric(l_m, G)) \quad (41)$$

where $metric(i, j)$ is the minimum metric of any path originating from node i and ending in node j and l_m represent the mth node in level l. With additive Gaussian noise, the metric for each branch is the mean-squared error between the symbol estimate and the received data. Once we know the metric associated with all the nodes in level l, the metric associated with the mth node in level $l + 1$ can be calculated by

$$metric(0, (l+1)_m) = \min_{i \in \text{states}} (metric(0, l_i) + metric(l_i, (l+1)_m)) \quad (42)$$

If there is no branch between the node i in state l and node m in state $l + 1$, then the metric associated with that branch is assumed to be infinitely large.

This iterative method of calculating the optimal code reduces the complexity of the decoder to be linear in codeword length κ. However, at every stage of the trellis, the Viterbi algorithm requires computation of the likelihood of each state. The number of states is exponential in the size of the constraint length, S, of the code, thereby making the total complexity of the algorithm of the order of $\kappa 2^{(S+1)}$.

For large constraint lengths, the Viterbi decoding can be impractical for real-time low-power applications. As applications require higher data rates with increasing reliability, higher constraint lengths are desirable. There have been several low-complexity alternatives to Viterbi decoding proposed in the literature: sequential decoding [135], majority logic decoding [136], M algorithm or list decoding [137,138], T algorithm [139], reduced-state sequence detection [140,141], and maximal weight decoding [59].

As noted in the beginning of this section, most of the channel coding and decoding procedures are designed for single-user AWGN channels or fading channels. In the presence of multiaccess interference, joint multiuser detection and decoding [142–146] can lead to lower error performance at the expense of increased receiver complexity.

5.2.4.2. Multiple Transmit Antennas. The information symbols encoded using the Alamouti scheme in Fig. 8 admit a simple maximum-likelihood decoder. With two transmit and single receive antenna, the sampled received signal in two consecutive time symbols is given by

$$z[1] = h_1 s_1 + h_2 s_2 + n_1$$
$$z[2] = -h_1 s_2^* + h_2 s_1^* + n_1$$

where n_1 and n_2 are assumed to be independent instances of circularly symmetric Gaussian noise with zero mean and unit variance. The maximum-likelihood detector builds the following two signals:

$$\hat{s}_1 = h_1^* z[1] + h_2 z^*[2] = (|h_1|^2 + |h_2|^2) s_1 + h_1^* n_1 + h_2 n_2^*$$
$$\hat{s}_2 = h_2^* z[1] - h_1 z^*[2] = (|h_1|^2 + |h_2|^2) s_2 - h_0 n_1^* + h_1^* n_1 \quad (43)$$

followed by the maximum-likelihood detector for each symbol s_i, $i = 1, 2$. The combined signals in (43) are equivalent to that obtained from a two-branch receive diversity using maximal ratio combining (MRC) [6]. Thus, the Alamouti scheme provides an order two transmit diversity much like an order two receive diversity using MRC. Note that both the Alamouti and MRC schemes have the same average transmission rate, one symbol per transmission, but the Alamouti scheme requires at least two transmissions to achieve order two diversity, while MRC achieves order two diversity per transmission.

If a space-time trellis code is used, then the decoder is a simple extension of the decoder for the single-antenna case. As the number of antennas is increased to achieve higher data rates, the decoding complexity increases exponentially in the number of transmit antennas [5], thereby requiring power-hungry processing at the receiver. Though there is no work on reduced complexity decoders for space-time trellis codes, complexity reduction concepts for single-antenna trellis decoding should apply (see text above).

5.3. Power Control

Power control was amply motivated on the capacity grounds in Sections 4.1 and 4.2; in this section, we will only highlight some of the representative research on power control methods and its benefits. Power control is widely used in second- and third-generation cellular systems. For instance, in IS-95, transmit power is controlled not only to counter the near–far effect but also to overcome the time-varying fading. By varying the transmit power based on the channel conditions, a fixed received signal-to-noise ratio (SNR) can be achieved. A SNR guarantee implies a guarantee on the reliability of received information, through the relation between the packet error rate and the received SNR [4].

Information-theoretically optimal power control for a multiuser system was discussed elsewhere [147–150]. While providing a bound on the achievable capacity, the proposed power control algorithms assume perfect knowledge of the time-varying channel at the transmitter. Hence, the power control policies and the resultant system performance is only a loose bound for the achievable performance. Network capacity analysis with power control errors has appeared in Refs. 151,152, and references therein.

Significant research effort has been devoted to power control algorithms for data traffic [e.g., 153–160]. Most of the above work on power control has been for circuit-switched networks, where users are given a certain dedicated channel for their entire session. With the advent of services supporting bursty traffic, such as email and Web browsing, resource allocation for shared channels and packet networks becomes of importance. First steps in these directions can be found in the literature [158,159,161]. Lastly, we note that power control can also lead to gain in packet-switched networks, like IEEE 802.11 or ad hoc networks; preliminary results can be found elsewhere [162,163].

6. CONCLUSIONS

If the relentless advances in wireless communications since 1990 are an indicator of things to come, then it is clear that we will witness not only faster ways to communicate but also newer modes of communication. The fundamental information theoretic bounds hold as long as the assumed communication model holds. The capacity of the channel can be "increased," by introducing new capabilities such as multiple antennas and ad hoc networking. Thus, it will be safe to conclude that the actual physical limits of wireless communication are still unknown and it is for us to exploit that untapped potential with a mix of creativity and serendipity.

BIOGRAPHIES

Ashutosh Sabharwal received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, New Delhi, India, in 1993. He received his M.S. and Ph.D. degrees in electrical engineering in 1995 and 1999, respectively, from the Ohio State University, Columbus, Ohio. Since 1999, he has been a postdoctoral research associate at the Center for Multimedia Communication, Rice University, Houson, Texas, where he currently is a faculty fellow. He was the recipient of the 1999 Presidential Dissertation Fellowship sponsored by Ameritech. His current research interests include wireless communications, network protocols, and information theory.

Behnaam Aazhang received his B.S. (with highest honors), M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign in 1981, 1983, and 1986, respectively. From 1981, to 1985, he was a research assistant in the Coordinated Science Laboratory at the University of Illinois. In August 1985, he joined the faculty of Rice University, Houston, Texas, where he is now the J. S. Abercrombic Professor in the Department of Electrical and Computer Engineering and the Director of Center for Multimedia Communications. He has been a Visiting Professor at IBM Federal Syttems Company, Houston, Texas; and Laboratory for Communication Technology at Swiss Federal Insitute of Technology (ETH), Zurich, Switzerland; the Telexommunications Laboratory at University of Oulu, Oulu, Finland; and the U.S. Air Force Phillips Laboratory, Albuquerque, New Mexico. His research interests are in the arcas of communication theory, information theory, and their applications with emphasis on multiple access communications, cellular mobile radio communications, and optical communication networks. Dr. Aazhang is a Fellow is IEEE, a recipient of the Alcoa Foundation Award 1993, the NSF Engineering Initiation Award 1987–1989, and the IBM Graduate Fellowship 1984–1985, and is a member of Tau Beta Pi and Eta Kappa Nu. He currently is serving on Houston Mayor's Commission on Cellular Towers. He has served as the editor for spread Spectrum Networks of *IEEE Transactions on Communications* 1993–1998; the treasurer of the IEEE Information Theory Society 1995–1998; the technical area chair of the 1997 Asilomar Conference, Monterey, California; the secretary of the Information Theory Society 1990–1993; the publications chairman of the 1993 IEEE International Symposium on Information Theory, San Antonio, Texas; the co-chair of the Technical Program Committee of 2001 Multi-Dimensional and Mobile Communication (MDMC) Conference in Pori, Finland.

BIBLIOGRAPHY

1. R. Steele, J. Whitehead, and W. C. Wong, System aspects of cellular radio, *IEEE Commun. Mag.* **33**: 80–86 (Jan. 1995).
2. U. T. Black, *Mobile and Wireless Networks*, Prentice-Hall, 1996.
3. J. Geier, *Wireless LANs: Implementing Interoperable Networks*, Macmillan Technical Publishing, 1998.
4. J. G. Proakis, *Digital Communications*, McGraw-Hill, 1995.
5. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**: 744–765 (March 1998).
6. D. G. Brennan, Linear diversity combining techniques, *Proc. IRE*, 1959.
7. T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, 1996.
8. M. G. Jansen and R. Prasad, Capacity, throughput, and delay analysis of a cellular DS-CDMA system with imperfect power control and imperfect sectorization, *IEEE Trans. Vehic. Technol.* **44**: 67–75 (Feb. 1995).
9. A. Sabharwal, D. Avidor, and L. Potter, Sector beam synthesis for cellular systems using phased antenna arrays, *IEEE Trans. Vehic. Technol.* **49**: 1784–1792 (Sept. 2000).
10. E. S. Sousa, V. M. Jovanović, and C. Daigneault, Delay spread measurements for the digital cellular channel in Toronto, *IEEE Trans. Vehic. Technol.* **43**: 837–847 (Nov. 1994).
11. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.
12. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, 1998.
13. S. Verdú and S. Shamai (Shitz), Spectral efficiency of CDMA with random spreading, *IEEE Trans. Inform. Theory* **45**: 622–640 (March 1999).
14. S. Shamai (Shitz) and S. Verdú, The impact of frequency flat fading on the spectral efficiency of CDMA, *IEEE Trans. Inform. Theory* **47**: 1302–1327 (May 2001).
15. I. Katzela and M. Nagshineh, Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey, *IEEE Pers. Commun.* 10–31 (June 1996).
16. S. Jordan, Resource allocation in wireless networks, *J. High Speed Networks* **5**(1): 23–24 (1996).
17. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo codes, *Proc. 1993 Int. Conf. Communications*, Geneva, Switzerland, May 1993, pp. 1064–1070.
18. P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* **CS-11**: 360–393 (Dec. 1963).
19. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: Information-theoretic and communication aspects, *IEEE Trans. Inform. Theory* **44**: 2619–2692 (Oct. 1998).

20. E. Malkamäki and H. Leib, Coded diversity on block-fading channels, *IEEE Trans. Inform. Theory* **45**: 771–781 (March 1999).

21. T. L. Marzetta and B. M. Hochwald, Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading, *IEEE Trans. Inform. Theory* **45**(1): 139–157 (1999).

22. R. Knopp and P. A. Humblet, On coding for block fading channels, *IEEE Trans Inform. Theory* **46**: 189–205 (Jan. 2000).

23. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (Part I), 623–656 (Part II) (1948).

24. L. H. Ozarow, S. Shamai, and A. D. Wyner, Information theoretic considerations for cellular mobile radio, *IEEE Trans. Inform. Theory* **43**: 359–378 (May 1994).

25. I. E. Telatar, *Capacity of Multi-Antenna Gaussian Channels*, Technical Report, AT&T Bell Labs, 1995; [appeared in *Eur. Trans. Telecommun.* **10**(6): 585–595 (1999)].

26. A. J. Goldsmith and P. P. Varaiya, Capacity of fading channels with channel side information, *IEEE Trans. Inform. Theory* **43**: 1986–1992 (Nov. 1997).

27. G. Caire, G. Taricco, and E. Biglieri, Optimum power control over fading channels, *IEEE Trans. Inform. Theory* **45**: 1468–1489 (July 1999).

28. A. Sabharwal, E. Erkip, and B. Aazhang, On side information in multiple antenna block fading channels, *Proc. ISITA*, Honolulu, Hawaii, Nov. 2000.

29. A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, Efficient use of side information in multiple-antenna data transmission over fading channels, *IEEE-JSAC* **16**: 1423–1436 (Oct. 1998).

30. A. Narula, M. D. Trott, and G. W. Wornell, Performance limits of coded diversity methods for transmitter antenna arrays, *IEEE Trans. Inform. Theory* **45**: 2418–2433 (Nov. 1999).

31. J.-C. Guey, M. Fitz, M. Bell, and W. Y. Kuo, Signal design for transmitter diversity wireless communication systems over raleigh fading channels, *IEEE Trans. Commun.* **46**: 527–537 (April 1999).

32. V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criteria in the presence of channel estimation errors, mobility, and multiple paths, *IEEE Trans. Commun.* **47**: 199–207 (Feb. 1999).

33. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill International Editions, 1984.

34. S. Verdú and T. S. Han, A general formula for channel capacity, *IEEE Trans. Inform. Theory* **40**: 1147–1157 (July 1994).

35. H. Viswanathan, *Capacity of Fading Channels with Feedback and Sequential Coding of Correlated Sources*, PhD thesis, Cornell Univ., Aug. 1997.

36. E. Biglieri, G. Caire, and G. Taricco, Limiting performance of block-fading channels with multiple antennas, *IEEE Trans. Inform. Theory* (Aug. 1999).

37. G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* 41–59 (1996).

38. G. Caire and S. Shamai (Shitz), On the capacity of some channels with channel state information, *IEEE Trans. Inform. Theory* **45**: 2007–2019 (Sept. 1999).

39. S. Shamai and A. D. Wyner, Information theoretic considerations for symmetric, cellular, multiple-access fading channels — Part I, *IEEE Trans. Inform. Theory* **43**: 1877–1894 (Nov. 1997).

40. G. W. Wornell, Spread-signature CDMA: Efficient multiuser communication in presence of fading, *IEEE Trans. Inform. Theory* **41**: 1418–1438 (Sept. 1995).

41. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications — a tutorial, *IEEE Trans. Commun.* **COM-30**: 855–884 (May 1982).

42. E. Erkip and B. Aazhang, Multiple access schemes over multipath fading channels, *Proc. ISIT*, Cambridge, MA, Aug. 1998.

43. A. D. Wyner, Shannon-theoretic approach to Gaussian cellular multiple access channel, *IEEE Trans. Inform. Theory* **40**: 1713–1727 (Nov. 1994).

44. O. Somekh and S. Shamai (Shitz), Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading, *Proc. 1998 IEEE Int. Symp. Information Theory*, Cambridge, MA, Aug. 1998, p. 393.

45. A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, 1995.

46. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.

47. R. G. Gallager, A perspective on multiaccess channels, *IEEE Trans. Inform. Theory* **IT-31**: 124–142 (March 1985).

48. A. Ephremides and B. Hajek, Information theory and communication networks: An unconsummated union, *IEEE Inform. Theory* **44**: 2416–2434 (Oct. 1998).

49. I. E. Telatar and R. G. Gallager, Combining queuing theory with information theory for multiaccess, *IEEE J. Select. Areas Commun.* **13**: 963–969 (Aug. 1995).

50. M. Win and R. A. Scholz, Impulse radio: How it works, *IEEE Commun. Lett.* **2**: 36–38 (Feb. 1998).

51. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorders, *IEEE Trans. Inform. Theory* **44**: 2260–2299 (Oct. 1998).

52. J. D. J. Costello, J. Hagenauer, H. Imai, and S. B. Wicker, Applications of error-control coding, *IEEE Trans. Inform. Theory* **44**: 2531–2560 (Oct. 1998).

53. A. R. Calderbank, The art of signalling, *IEEE Trans. Inform. Theory* **44**: 2561–2595 (Oct. 1998).

54. I. Blake, C. Heegard, T. Høholdt, and V. Wei, Algebraic-geometry codes, *IEEE Trans. Inform. Theory* **44**: 2596–2618 (Oct. 1998).

55. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

56. G. D. Forney, *Concatenated Codes*, MIT Press, Cambridge, MA, 1966.

57. J. H. van Lint, *Introduction to Coding Theory*, Springer-Verlag, New York, 1992.

58. V. S. Pless, W. C. Huffman, and R. A. Brualdi, eds., *Handbook of Coding Theory*, Elsevier, New York, 1998.

59. S. Das, *Multiuser Information Processing in Wireless Communication*, PhD thesis, Rice Univ., Houston, TX, Sept. 2000.

60. A. Dholakia, *Introduction to Convolutional Codes with Applications*, Kluwer Academic Publishers, 1994.
61. D. Divsalar and M. K. Simon, The design of trellis coded MPSK for fading channels: Performance criteria, *IEEE Trans. Commun.* **36**: 1004–1012 (Sept. 1988).
62. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **28**: 55–67 (Jan. 1982).
63. G. D. Forney and G. Ungerboeck, Modulation and coding for linear Gaussian channels, *IEEE Trans. Inform. Theory* **44**: 2384–2415 (Oct. 1998).
64. J. L. Massey, Coding and modulation in digital communications, *Proc. 1974 Int. Zurich Seminar on Digital Communications*, Zurich, Switzerland, March 1974, pp. E2(1)–E2(4).
65. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.
66. R. G. Gallager, *Low Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1962.
67. S. L. Goff, A. Glavieux, and C. Berrou, Turbo-codes and high spectral efficiency modulation, *Proc. 1994 Int. Conf. Communications*, May 1994, Vol. 2, pp. 645–649.
68. W. Liu and S. G. Wilson, Rotationally-invariant concatenated (turbo) TCM codes, *Proc. 1999 Asilomar Conf. Signal System Comp.*, Oct. 1999, Vol. 1, pp. 32–36.
69. O. Y. Takeshita and J. D. J. Costello, New deterministic interleaver designs for turbo codes, *IEEE Trans. Inform. Theory* **46**: 1988–2006 (Sept. 2000).
70. Y. Liu, M. Fitz, and O. Y. Takeshita, QPSK space-time turbo codes, *Proc. 2000 Int. Conf. Communications*, June 2000, Vol. 1, pp. 292–296.
71. V. Tarokh, A. Vardy, and K. Zeger, Universal bound on the performance of lattice codes, *IEEE Trans. Inform. Theory* **45**: 670–681 (March 1999).
72. J. H. Winters, Smart antennas for wireless systems, *IEEE Pers. Commun.* **5**: 23–27 (Feb. 1998).
73. G. J. Foschini and M. J. Gans, On limits of wireless communication in a fading environment when using multiple antennas, in *Wireless Personal Communications*, Kluwer Academic Publishers, 1998.
74. V. Tarokh, H. Jafarkhani, and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* **45**: 1456–1467 (July 1999).
75. D. M. Ionescu, New results on space time code design criteria, *Proc. IEEE Wireless Communications and Networking Conf.*, New Orleans, Oct. 1999.
76. O. Tirkkonnen and A. Hottinen, Complex space-time block codes for four Tx antennas, *Proc. GLOBECOM*, 2000 pp. 1005–1009.
77. S. Baro, G. Bauch, and A. Hansmann, Improved codes for space-time trellis coded modulation, *IEEE Commun. Lett.* **4**: 20–22 (Jan. 2000).
78. A. R. Hammons and H. E. Gamal, On the theory of space-time codes for PSK modulation, *IEEE Trans. Inform. Theory* **46**: 524–542 (March 2000).
79. T. Moharemovic and B. Aazhang, Information theoretic optimality of orthogonal space-time transmission schemes and concatenated code construction, *Proc. Int. Conf. Communications* (*ICC*), Acapulco, Mexico, May 2000.
80. M. J. Borran, M. Memarzadeh, and B. Aazhang, Design of coded modulation schemes for orthogonal transmit diversity, *IEEE Trans. Commun.* (in press).
81. S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (Oct. 1998).
82. R. Prasad, Overview of wireless personal communications: Microwave perspectives, *IEEE Commun. Mag.* 104–108 (April 1997).
83. G. Taricco, E. Biglieri, and G. Caire, Limiting performance of block-fading channels with multiple antennas, *Proc. Inform. Theory Communication Workshop*, pp. 27–29, 1999.
84. *http://www.3gpp.org/*.
85. T. Ojanpera and R. Prasad, eds., *Wideband CDMA for Third Generation Mobile Communications*, Artech House Universal Personal Communications Series, 1998.
86. A. Lapidoth and P. Narayan, Reliable communication under channel uncertainty, *IEEE Trans. Inform. Theory* **44**: 2148–2177 (Oct. 1998).
87. K. Li and H. Liu, Joint channel and carrier offset estimation in CDMA communications, *IEEE Trans. Signal Process.* **47**: 1811–1822 (July 1999).
88. Z. Pi and U. Mitra, Blind delay estimation in multi-rate asynchronous DS-CDMA systems, *IEEE Trans. Commun.* (2000).
89. C. Sengupta, *Algorithms and Architectures for Channel Estimation in Wireless CDMA Communication Systems*, PhD thesis, Rice Univ., Dec. 1998.
90. Z. Tian, K. L. Bell, and H. L. V. Trees, A quadratically constrained decision feedback equalizer for DS-CDMA communication systems, in *IEEE Workshop on Signal Processing Advances in Wireless Communication*, May 1999, pp. 190–193.
91. C. Sengupta, J. Cavallaro, and B. Aazhang, On multipath channel estimation for DS-CDMA systems with multiple sensors, *IEEE Trans. Commun.* **49**: 543–553 (March 2001).
92. D. N. Godard, Self-recovering equalization and carrier tracking of two-dimensional data communication systems, *IEEE Trans. Commun.* **28**: 1867–1875 (Nov. 1980).
93. J. R. Treichler and B. G. Agree, A new approach to multipath correction of constant modulus signals, *IEEE Trans. Acoust. Speech Signal Process.* **31**: 459–472 (April 1983).
94. W. A. Gardner, A new method for channel identification, *IEEE Trans. Commun.* **39**: 813–817 (June 1991).
95. W. A. Gardner, Exploitation of spectral redundancy in cyclostationary signals, *IEEE Signal Process. Mag.* 14–36 (April 1991).
96. L. Tong, G. Xu, and T. Kailath, Blind identification and equalization based on second-order statistics: A time domain approach, *IEEE Trans. Inform. Theory* **40**: 340–349 (March 1994).
97. E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, Subspace methods for the blind identification of multichannel FIR filters, *IEEE Trans. Signal Process.* **43**: 516–525 (Feb. 1995).
98. H. Liu, G. Xu, L. Tong, and T. Kailath, Recent developments in blind channel equalization: From cyclostationarity to subspaces, *Signal Process.* **50**(1–2): 83–99 (1996).
99. S. E. Bensley and B. Aazhang, Subspace-based channel estimation for code division multiple access communications, *IEEE Trans. Commun.* **44**: 1009–1020 (Aug. 1996).

100. E. Ertin, U. Mitra, and S. Siwamogsatham, Maximum-likelihood based multipath channel estimation for code-division multiple-access systems, *IEEE Trans. Commun.* **49**: 290–302 (Feb. 2001).

101. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc. Ser. B* 1–38 (1977).

102. I. Ziskind and M. Wax, Maximum likelihood localization of multiple sources by alternating projection, *IEEE Trans. Signal Process.* **36**: 1553–1560 (Oct. 1988).

103. M. Torlak and G. Xu, Blind multiuser channel estimation in asynchronous CDMA systems, *IEEE Trans. Signal Process.* **45**: 137–147 (Jan. 1997).

104. M. K. Tsatsanis and G. B. Giannakis, Blind estimation of direct sequence spread spectrum signals in multipath, *IEEE Signal Process.* **45**: 1241–1252 (1997).

105. T. Östman and B. Ottersten, Near far robust time delay estimation for asynchronous DS-CDMA systems with bandlimited pulse shapes, *Proc. IEEE Vehicular Technology Conf.*, May 1998, pp. 1651–1654.

106. V. Tripathi, A. Mantravadi, and V. V. Veeravalli, Channel acquisition for wideband CDMA, *IEEE J. Select. Areas Commun.* **18**: 1483–1494 (Aug. 2000).

107. E. Aktas and U. Mitra, Single user sparse channel acquisition for ds/cdma, *Proc. CISS*, Princeton, NJ, 2000.

108. S. Bhashyam, A. Sabharwal, and U. Mitra, Channel estimation multirate DS-CDMA systems, *Proc. Asilomar Conf. Signal System Comp.*, Pacific Grove, CA, Oct. Nov. 2000.

109. T. P. Krauss and M. D. Zoltowski, Blind channel identification on CDMA forward link based on dual antenna receiver at handset and cross-relation, *Proc. 1999 Asilomar Conf. Signal System Communication*, Oct. 1999, Vol. 1, pp. 75–79.

110. C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, 1973.

111. S. Bhashyam and B. Aazhang, Multiuser channel estimation for long code CDMA systems, *Proc. 2000 Wireless Communication Networking Conf.*, 2000.

112. S. Verdú, *Optimum Multiuser Signal Detection*, PhD thesis, Univ. Illinois at Urbana — Champaign, Aug. 1984.

113. S. Verdú, Computational complexity of optimum multiuser detection, *Algorithmica* **4**: 303–312 (1989).

114. S. Verdú, Optimum multiuser asymptotic efficiency, *IEEE Trans. Commun.* **34**: 890–897 (Sept. 1986).

115. K. S. Schneider, Optimum detection of code division multiplexed signals, *IEEE Trans. Aerospace Electron. Syst.* **AES-15**: 181–185 (Jan. 1979).

116. R. Kohno, M. Hatori, and H. Imai, Cancellation techniques of co-channel interference in asynchronous spread spectrum multiple access systems, *Electron. Commun. Japan* **66-A**(5): 20–29 (1983).

117. R. Lupas and S. Verdú, Linear multiuser detectors for synchronous code-division multiple-access channels, *IEEE Trans. Inform. Theory* **35**(1): 123–136 (1989).

118. R. Lupas and S. Verdú, Near-far resistance of multi-user detectors in asynchronous channels, *IEEE Trans. Commun.* **38**: 496–508 (April 1990).

119. Z. Xie, R. T. Short, and C. K. Rushforth, A family of suboptimum detectors for coherent multiuser communications, *IEEE JSAC* **8**: 683–690 (May 1990).

120. M. Honig, U. Madhow, and S. Verdú, Blind adaptive multiuser detection, *IEEE Trans. Inform. Theory* **41**: 944–960 (July 1995).

121. R. Mailloux, *Phased Array Antenna Handbook*, Artech House, 1994.

122. H. V. Poor and S. Verdú, Probability of error in MMSE multiuser detection, *IEEE Inform. Theory* **43**: 858–871 (May 1997).

123. A. Sabharwal, U. Mitra, and R. Moses, Cyclic Wiener filtering based multirate DS-CDMA receivers, *Proc. IEEE WCNC*, New Orleans, Sept. 1999.

124. A. Sabharwal, U. Mitra, and R. Moses, Low complexity MMSE receivers for multirate DS-CDMA systems, *Proc. CISS*. Princeton, NJ, 2000.

125. H. Y. Wu and A. Duel-Hallen, Performance comparison of multi-user detectors with channel estimation for flat Rayleigh fading CDMA channel, *Wireless Pers. Commun.* (July/Aug. 1996).

126. S. D. Gray, M. Kocic, and D. Brady, Multiuser detection in mismatched multiple-access channels, *IEEE Trans. Commun.* **43**: 3080–3089 (Dec. 1995).

127. B. Rimoldi and R. Urbanke, A rate splitting approach to the Gaussian multiple-access channel, *IEEE Trans. Inform. Theory* **42**: 364–375 (March 1996).

128. K. I. Pederson, T. E. Kolding, I. Seskar, and J. M. Holzman, Practical implementation of successive interference cancellation in DS/CDMA systems, *Proc. IEEE Conf. Universal Personal Communication*, 1996, Vol. 1, pp. 321–325.

129. M. K. Varanasi and B. Aazhang, Multistage detection in asynchronous code-division multiple-access communications, *IEEE Trans. Commun.* **38**: 509–519 (April 1990).

130. P. Patel and J. Holzman, Performance comparison of a DS/CDMA system using a successive interference cancellation (IC) scheme and a parallel IC scheme under fading, *Proc. ICC*, New Orleans, May 1994, pp. 510–514.

131. R. M. Buehrer and B. D. Woerner, Analysis of adaptive multistage interference cancellation for CDMA using an improved Gaussian approximation, *Proc. IEEE MILCOM*, San Diego, CA, Nov. 1995, pp. 1195–1199.

132. M. K. Varanasi and B. Aazhang, Near-optimum detection in synchronous code-division multiple-access schemes, *IEEE Trans. Commun.* **39**: 725–736 (May 1991).

133. S. Moshavi, *Multistage Linear Detectors for DS-CDMA Communications*, PhD thesis, City Univ. New York, Jan. 1996.

134. A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**: 260–269 (April 1967).

135. J. M. Wozencraft and B. Reiffen, *Sequential Decoding*, MIT Press, Cambridge, MA, 1961.

136. J. L. Massey, *Threshold Decoding*, MIT Press, Cambridge, MA, 1998.

137. J. B. Anderson and S. Mohan, Sequential coding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**: 169–176 (Feb. 1984).

138. G. J. Pottie and D. P. Taylor, A comparison of reduced complexity decoding algorithms for trellis codes, *IEEE J. Select. Areas Commun.* **7**: 1369–1380 (Dec. 1989).

139. S. T. Simmons, Breadth-first trellis decoding with adaptive effort, *IEEE Trans. Commun.* **38**: 3–12 (Jan. 1990).

140. M. V. Eyuboglu and S. Qureshi, Reduced-state sequence estimation for coded modulation on interference channels, *IEEE J. Select. Areas Commun.* **35**: 944–955 (Sept. 1989).

141. P. R. Chevillat and E. Elephtheriou, Decoding of trellis-encoded signals in the presence of inter-symbol interference and noise, *IEEE Trans. Commun.* **37**: 669–676 (July 1989).

142. C. Schlegel, P. Alexander, and S. Roy, Coded asynchronous CDMA and its efficient detection, *IEEE Trans. Inform. Theory* **44**: 2837–2847 (Nov. 1998).

143. X. Wang and H. V. Poor, Iterative (turbo) soft interference cancellation and decoding for coded CDMA, *IEEE Trans. Commun.* **47**: 1046–1061 (July 1999).

144. H. E. Gamal and E. Geraniotis, Iterative multiuser detection for coded CDMA signals in AWGN and fading channels, *IEEE J. Select. Areas Commun.* **18**: 30–41 (Jan. 2000).

145. R. Chen, X. Wang, and J. S. Liu, Adaptive joint detection and decoding flat-fading channels via mixture Kalman filtering, *IEEE Trans. Inform. Theory* **46**: 2079–2094 (Sept. 2000).

146. L. Wei and H. Qi, Near-optimal limited-search detection on ISI-CDMA channels and decoding of long-convolutional codes, *IEEE Trans. Inform. Theory* **46**: 1459–1482 (July 2000).

147. D. N. C. Tse and S. V. Hanly, Multiaccess fading channels-Part I: polymatroid structure, optimal resource allocation and throughput capacities, *IEEE Trans. Inform. Theory* **44**: 2796–2815 (Nov. 1998).

148. S. V. Hanly and D. N. C. Tse, Multiaccess fading channels — Part II: Delay-limited capacities, *IEEE Trans. Inform. Theory* **44**: 2816–2831 (Nov. 1998).

149. P. Viswanath, V. Anantharam, and D. N. C. Tse, Optimal sequences, power control, and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers, *IEEE Trans. Inform. Theory* **45**: 1968–1983 (Sept. 1999).

150. S. Hanly and D. Tse, Power control and capacity of spread-spectrum wireless networks, *Automatica* **35**: 1987–2012 (Dec. 1999).

151. N. Bambos, Toward power-sensitive network architectures in wireless communications: concepts, issues, and design aspects, *IEEE Pers. Commun.* **5**: 50–59 (June 1998).

152. J. Zhang and E. K. P. Chong, CDMA systems in fading channels: admissibility, network capacity and power control, *IEEE Trans. Inform. Theory* **46**: 962–981 (May 2000).

153. J. Wu and R. Kohno, A wireless multimedia CDMA system based on transmission power control, *IEEE J. Select. Areas Commun.* **14**: 683–691 (May 1996).

154. J. Jacobsmeyer, Congestion relief on power-controlled CDMA networks, *IEEE J. Select. Areas Commun.* **14**: 1758–1761 (Dec. 1996).

155. A. Sampath and J. M. Holtzman, Access control of data in integrated voice/data CDMA systems: benefits and tradeoffs, *IEEE J. Select. Areas Commun.* **15**: 1511–1526 (Oct. 1997).

156. D. Ayyagari and A. Ephremides, Cellular multicode CDMA capacity for integrated (voice and data) services, *IEEE J. Select. Areas Commun.* **17**: 928–938 (May 1999).

157. Y. Lu and R. W. Broderson, Integrating power control, error correction coding, and scheduling for a CDMA downlink system, *IEEE J. Select. Areas Commun.* **17**: 978–989 (May 1999).

158. D. Kim, Rate-regulated power control for supporting flexible transmission in future cdma mobile networks, *IEEE J. Select. Areas Commun.* **17**: 968–977 (May 1999).

159. S. Manji and W. Zhuang, Power control and capacity analysis for a packetized indoor multimedia DS-CDMA network, *IEEE Trans. Vehic. Technol.* **49**: 911–935 (May 2000).

160. D. Goodman and N. Mandayam, Power control for wireless data, *IEEE Pers. Commun.* **7**: 48–54 (April 2000).

161. N. Bambos and S. Kandukuri, Power controlled multiple access (PCMA) in wireless communication networks, *Proc. INFOCOM 2000*, March 2000, Vol. 2, pp. 386–395.

162. P. Gupta and P. R. Kumar, The capacity of wireless networks, *IEEE Trans. Inform. Theory* **46**: 388–404 (March 2000).

163. J. Monks, Power controlled multiple access in *ad hoc* networks, *Proc. Multiaccess, Mobility and Teletraffic for Wireless Communications* (*MMT*), Duck Key, FL, Dec. 2000.

N

NETWORK FLOW CONTROL

STEVEN H. LOW
California Institute of Technology
Pasadena, California

1. INTRODUCTION

Flow and congestion control is a distributed algorithm to share network resources among competing users. Like a transportation network, congestion can build up in a telecommunications network when traffic load exceeds network capacity. When more and more automobiles enter a highway, their speed decreases until everyone is driving at, say, less than 10 km/h instead of 100 km/h, and the network throughput plummets. As load increases in a packet network, queues build up until packets are delayed by an excessive amount or even lost and need to be retransmitted. Packets that are transmitted multiple times or at upstream nodes only to be discarded at downstream nodes, waste network resources and intensify congestion. This has led to congestion collapse where throughput dropped to a small fraction of network capacity [1]. Flow control prevents congestion collapse by adapting user traffic to available capacity.[1]

We distinguish between two types of networks, circuit-switched networks, which we will abbreviate as circuit networks, and packet-switched networks, which we will abbreviate as packet networks. The most important difference between them is that, in a circuit network, when a connection is established between a source and a destination, network resources (e.g., time slots in time-division multiplexed systems, frequency slots in frequency-division multiplexed systems) are reserved along the path for its exclusive use for the duration of the connection. The traditional telephone network is an example of circuit network. The fixed rate allocation simplifies the control of the system and the provisioning of quality of service (QoS). Since network resources (called a "circuit") are dedicated to a connection, they are wasted when the information source of the connection occasionally has no information to send, even if other traffic can make use of these resources. Hence, circuit network is suitable to support applications that generate traffic at a fixed rate, such as uncoded voice. Flow control, that adapts source transmission rate to changes in the availability of network resources along its path, is unnecessary. Traffic is regulated at the connection level through *connection admission control*, which decides whether or not a new connection request is granted, depending on, for example, the availability of resources.

[1] Some authors use *flow* control to refer to mechanisms to avoid a source from overwhelming a receiver and *congestion* control to refer to mechanisms to avoid overloading the network. We make no such distinction and will refer to both as flow control in this article.

In a packet network, in contrast, a path may be established between a source and its destination during the connection setup phase, but no bandwidth or buffer resources are reserved. Rather, these resources are shared by all connections on demand. This is suitable for applications that generate bursty traffic, where periods of activity are interspersed with random idle periods. Sharing of resources dynamically by multiple traffic streams is referred to as *statistical multiplexing*. Statistical multiplexing improves efficiency, since a resource is never idle if there is traffic to be carried, unlike the situation in a circuit network. It, however, makes the control and provisioning of QoS harder. In a circuit network, each connection requires a fixed rate and therefore connection admission control can be easily implemented by checking whether this rate can be supported along the intended path between source and destination. It is difficult to characterize the resource requirements of a bursty source, and hence connection admission control is rarely implemented in packet networks. Since the number of connections in the network is not controlled, the source rates of these connections must be regulated to avoid overwhelming the network or the receiver. This is the purpose of flow control.

In this article, we describe the design objectives (Section 2) and implementation constraints (Section 3) of flow control mechanisms. There is no automatic way to synthesize a flow control scheme that satisfies these objectives, but we can analyze existing or proposed mechanisms through mathematical modeling and computer simulations and apply the understanding to enhance current schemes or design new ones. The goal of this article is to provide an introduction to recent mathematical models for understanding the equilibrium and stability properties of flow control mechanisms (Section 5). Our focus is on general properties that underlie a large class of flow control schemes and therefore many implementation details are abstracted out of the mathematical model. An explanation of network protocols in general is provided in the article [2] in this encyclopedia, and a detailed description of TCP, as well as further references, are provided in the article [3] also in this encyclopedia. To make our discussion concrete, we will use TCP Vegas with DropTail routers, explained in Section 4, for illustration throughout the article.

Our discussion centers around TCP both because it is pervasive — it is estimated that it carries 90% of traffic on the current Internet — and because its distributed, decentralized and asynchronous character allows a scalable implementation. For flow control schemes in asynchronous transfer mode (ATM) networks, see, Ref. 4. A breakthrough in TCP flow control is the algorithm proposed in Ref. 1, which was implemented in the Tahoe version of TCP, and later enhanced into TCP Reno and other variants. These protocols are widely deployed in the current Internet (see [3]). TCP Vegas is proposed in Ref. 5 as an alternative to TCP Reno. Even though it is not widely deployed, it possesses interesting fairness and scalability properties that make it potentially more suitable for

future high speed networks. It also has a simpler analytical structure than Reno that makes it more convenient for us to use as an illustration of the general principles.

In this article, we will use "sources" and "connections" interchangeably. With quantities z_i defined, $z = (z_i) = (z_1, z_2, \ldots)$ denotes the vector whose elements are z_i.

2. DESIGN OBJECTIVES

The objectives of flow control schemes are to share network resources fairly and efficiently, to provide good QoS, and to be responsive to changes in network (or receiver) congestion. What makes the implementation of these objectives challenging is the constraints imposed by decentralization. In this section we elaborate on the design objectives; in the next section we discuss decentralization constraints.

2.1. Fairness

There are many definitions of fairness. Consider a linear network consisting of links $1, 2, \ldots, N$ in tandem, each with a bandwidth capacity of 1 unit. The network is shared by sources $0, 1, 2, \ldots, N$. Source 0 traverses all the N links, while sources $i, i \geq 1$, traverses only link i. If we aim to equally share the bandwidth among the $N + 1$ sources, then each source should get 1/2; this is called *maxmin* fairness [6]. If we aim to maximize the sum of source rates, then each source $i, i \geq 1$, should receive a rate of 1 while source 0 gets 0, to achieve a total source rate of N, almost double that under maxmin fairness if N is large. A compromise, called *proportional* fairness [7], allocates $N/(N + 1)$ to each source $i, i \geq 1$, and $1/(N + 1)$ to source 0.

In general, we can associate a utility function $U_i(x_i)$ with each source i, as a function of the source rate x_i, say, in packets per second. The utility function measures how happy source i is when it transmits at rate x_i. It is usually a concave increasing function, with the interpretation that sources are happier the higher the rate allocation they receive but there is a diminishing return as rate allocation increases. We can then define a rate allocation vector $x = (x_i)$ as *fair*, with respect to utility functions $U = (U_i)$, if it maximizes the aggregate utility $\sum_i U_i$ subject to capacity constraints. For instance, $U_i(x_i) = x_i$ corresponds to maximizing aggregate source rate, and $U_i(x_i) = \log x_i$ corresponds to proportional fairness. We will come back to utility maximization in Section 6.1.

In summary, one of the objectives of flow control is to share network resources fairly, and this can be interpreted as maximizing aggregate utility with different fairness criteria corresponding to different source utility functions.

2.2. Utilization and Quality of Service

Different applications have different quality requirements. For our purposes, we will use QoS to mean packet loss or queueing delay. Packet loss and queueing delay will both be low if the queue length can be kept small.

Recall that packet networks typically do not restrict the number of concurrent sources. If these sources are not flow-controlled, then their aggregate load may exceed the available capacity. Packets will arrive at routers or switches faster than they can be processed and forwarded. Queues will build up, increasing queueing delay, and eventually overflow, leading to packet loss.

Of course, one way to maintain small queues is to underutilize the network, by restricting source rates to (much) less than network capacity. Indeed, if we model the network as an M/M/1 queue, then the model dictates that the input rate must be significantly smaller than the capacity if average queue length is not to be excessive. This suggests an inevitable tradeoff between utilization and QoS: we can achieve either high utilization or high QoS, but *not* both. This view, however, is flawed for it ignores the *feedback* regulation inherent in flow control. It implicitly assumes that the input process remains statistically unchanged as queue builds up indefinitely. With feedback, input rate will be reduced in response to queue buildup, and hence it is possible, with proper control strategy, to stabilize input rate close to the capacity without incurring a large queue.

Ideally, flow control should adapt external traffic load to available capacity, and in equilibrium match the arrival rate to capacity at every bottleneck link. Moreover, queues should then stabilize around a small value, achieving small loss and delay.

As we will explain later, however, it may be difficult to maintain a small queue when congestion information is fed back to traffic sources only implicitly. High utilization and high QoS can both be achieved in a decentralized manner if explicit feedback is available.

2.3. Dynamic Properties

Fairness, utilization, and QoS, such as packet loss and delay, typically are considered as "equilibrium" properties in that usually we only require flow control schemes to achieve these objectives in equilibrium (or stationary regime). Another important criterion by which a flow control scheme is evaluated is its dynamic properties, such as whether the equilibrium point is stable and whether the transition to a new equilibrium is fast.

An ideal flow control scheme should be stable, in the sense that after a disturbance (e.g., arrival or departure of connections), it always converges to a possibly new equilibrium. Moreover, it should converge rapidly, in the presence of network delays, and in an asynchronous environment.

Convergence to equilibrium is desirable because, under a properly designed scheme, fairness, utilization, and QoS objectives are achieved in equilibrium.

2.4. Scalability

Scalability refers to the property that a flow control scheme has a small implementation complexity (implementation scalability) and that it maintains its performance, with respect to the objectives previously discussed (performance scalability), as the network scales up in capacity, propagation delay (geographically reach), and the number of sources.

Flow control schemes that are practical and that are discussed in this article, must be distributed, decentralized, and easy to implement for it not to be a bottleneck itself. The implementation complexity of these schemes typically is low and scalable. Hence, in the rest of the article, we will focus only on performance scalability.

3. INFORMATION CONSTRAINTS

It is important to realize that flow control consists of two algorithms, one carried out by traffic sources to adapt their rates to congestion information on their paths and the other carried out by network resources, often implicitly, to update a measure of congestion whose value is fed back, often implicitly as well, to the sources. On the current Internet, the source algorithm is carried out by TCP and the link algorithm is carried out by a queueing discipline, such as DropTail or RED. Even though the link algorithms are often implicit and overlooked, they are critical in determining the equilibrium and dynamic properties of a network under flow control.

For example, the current TCP uses packet loss as a measure of congestion [3]. A link is considered congested if the loss probability at that link is high. As loss probability increases, a TCP source reduces its rate in response, which in turn causes the link to reduce its loss probability, and so on. The behavior of this feedback loop is determined by how TCP adjusts its rate and how a link implicitly adjusts its congestion measure.

Decentralization requires that the source and the link algorithms use only local information. In this section, we explain the local information available at sources and links for the class of flow control schemes we discuss.

TCP uses "window" flow control, where a destination sends acknowledgments for packets that are correctly received. The time from sending a packet to receiving its acknowledgment is called *round-trip time*. It can be measured at the source and thus does not need clock synchronization between source and destination. A source keeps a variable called window size that limits the number of outstanding packets that have been transmitted but not yet acknowledged. When the window size is exhausted, the source must wait for an acknowledgment before sending a new packet. By numbering the packets and the acknowledgments, the source can estimate the transfer delay of each packet and detect if a packet is lost. Two features are important. The first is the "self-clocking" feature that automatically slows down the source when a network becomes congested and acknowledgments are delayed. The second is that the window size controls the source rate: roughly one window of packets is sent every round-trip time. The first feature was the only congestion control mechanism in the Internet before Jacobson's proposal in 1988 [1]. Jacobson's idea is to *dynamically* adapt window size to network congestion.

These *end-to-end* delay and loss measurements succinctly summarize the congestion on a path. They are the only local information available for a source to adjust its rate, if the network provides no explicit congestion notification.[2] Moreover, the information is delayed in the sense that the observed delay and loss information at the source reflects the state of the path at an earlier time. Note that the source does not know the delay or loss at individual links in its path. It does not even know (or make use of) its own routing, network topology, or how many other sources are sharing links with it. It must infer congestion from the end-to-end measurements and adjust its rate accordingly.

Similarly, at the links, the local information that is available for the update of congestion measure is the arrival rates of flows that traverse the link. Again, no global information, such as flow rates, delays, or loss probabilities at other links, should be used by a link algorithm. In principle, individual flow rates can be measured and used in the adjustment of congestion measure. However, this would require per-flow processing which can be expensive at high speed. A link algorithm is simpler if it only uses the aggregate rate of all flows traversing the link. We restrict our discussion to this class of link algorithms. For instance, queueing delay and loss probability under first-in-first-out (FIFO) discipline are updated, implicitly, based only on aggregate rate.

[2] Even with ECN bit, RED still provides one-bit of congestion information on the end-to-end path, as DropTail does; see Section 4.2 below. The difference between RED and DropTail is how they generate packet losses, not their information carrying capacity.

4. EXAMPLE: TCP VEGAS

Before we describe mathematical models to understand the equilibrium and stability properties of flow control mechanisms, we briefly describe TCP Vegas, which will be used for illustration later.

4.1. TCP Vegas

Like TCP Reno, TCP Vegas also consists of three phases: slow start, congestion avoidance, and fast retransmit/fast recovery. A Reno source starts cautiously with a small window size of one packet (up to four packets have recently been proposed) and the source increments its window by one every time it receives an acknowledgment. This doubles the window every round-trip time and is called slow start. When the window reaches a threshold, the source enters the congestion avoidance phase, where it increases its window by the reciprocal of the current window size every time it receives an acknowledgment. This increases the window by one in each round-trip time, and is referred to as additive increase. The threshold that determines the transition from slow start to congestion avoidance is meant to indicate the available capacity in the network and is adjusted each time a loss is detected. On detecting a loss through three duplicate acknowledgments, the source sets the slow start threshold to half the current window size, retransmits the lost packet and halves its window size. This is called fast retransmit/fast recover; see Ref. 3 for more details. When the acknowledgment for the retransmitted packet arrives, the source re-enters congestion avoidance. In TCP Reno, slow start is entered only rarely when the source first starts and when a loss is detected by timeout rather than duplicate acknowledgments.

TCP Vegas [5] improves upon TCP Reno through three main techniques. The first is a modified retransmission mechanism where timeout is checked on receiving the first duplicate acknowledgment, rather than waiting for the third duplicate acknowledgment (as Reno would), and results in a more timely detection of loss. The second technique is a more prudent way to grow the window

size during the initial use of slow-start when a connection starts up and it results in fewer losses.

The third technique is a new congestion avoidance mechanism that corrects the oscillatory behavior of Reno. The idea is to have a source estimate the number of its own packets buffered in the path and try to keep this number between α (typically 1) and β (typically 3) by adjusting its window size. The window size is increased or decreased by one in each round-trip time according to whether the current estimate is less than α or greater than β. Otherwise the window size is unchanged. The rationale behind this is to maintain a small number of packets in the pipe to take advantage of extra capacity when it becomes available. Another interpretation of the congestion avoidance algorithm of Vegas is given in Ref. 8, in which a Vegas source periodically measures the round-trip *queueing* delay and sets its rate to be proportional to the ratio of its round-trip propagation delay to queueing delay, the proportionality constant being between α and β. Hence, the more congested its path is, the higher the queueing delay and the lower the rate. The Vegas source obtains queueing delay by monitoring its round-trip time the time between sending a packet and receiving its acknowledgment and subtracting from it the round-trip propagation delay.

4.2. DropTail

Congestion control of the Internet was entirely source-based at the beginning, in that the link algorithm was implicit. A link simply drops a packet that arrives at a full buffer. This is called DropTail (or Tail Drop) and the implicit link algorithm is carried out by the queue process. The congestion measure it updates depends on the TCP algorithm.

For TCP Reno and its variants, the congestion measure is packet loss probability. The end-to-end loss probability is observed at the source and is a measure of congestion on the end-to-end path. For TCP Vegas, the congestion measure turns out to be link queueing delay [8] when FIFO service discipline is used. The congestion measure of a path is the sum of queueing delays at all constituent links.

Random early detection (RED) is proposed in Ref. 9 as an alternative to DropTail. In RED, an arrival packet is discarded with a probability when the average queue length exceeds a minimum threshold, in order to provide early warning of incipient congestion before the buffer overflows. The dropping probability is an increasing function of average queue length. The rationale is that a large average queue length signifies congestion and should intensify the feedback signal. It has also been proposed that a bit in the IP header be used for explicit congestion notification (ECN), so that a link can mark a packet probabilistically (setting the ECN bit from 0 to 1) instead of dropping it.

5. DUALITY MODEL

In this section, we first describe an abstract model for general source and link algorithms. As an illustration; we then applied it to the Vegas/DropTail algorithms described in the last section.

5.1. General Source/Link Algorithms

A network is modeled as a set L of "links," indexed by l, with finite transmission capacities c_l packets per second. It is shared by a set of sources, indexed by s. Each source s is assigned a path along which data is transferred to its destination. A path is a subset of the links and is denoted by $L_s \subseteq L$. For convenience, denote by S_l the subset of sources that traverse link l. Hence, $l \in L_s$ if and only if $s \in S_l$. To understand the equilibrium and stability of the network, we assume for simplicity that the link capacities c_l, the set of links and sources, and the routes L_s are all fixed at the timescale of interest.

Each source s adjusts its transmission rate $x_s(t)$ at time t, in packets per second, based on the congestion on its path. Each link l maintains a measure of congestion $p_l(t)$ at time t. We will call $p_l(t)$ the link *price* for it can be interpreted as unit price for bandwidth at link l (see, [7,10]). A link is said to be congested if $p_l(t)$ has a large value. A path is said to be congested if the sum of link prices is high.

Each source s can observe a delayed version of the sum of link prices, summed over the links in its path. This path price is a measure of congestion in the path end-to-end. Suppose the backward delay from link l to source s is denoted by τ^b_{ls}. Then the path price that is observed by s at time t can be represented by [19]

$$q_s(t) := \sum_{l \in L_s} p_l(t - \tau^b_{ls}) \tag{1}$$

We assume each link l can observe a delayed version of the sum of source rates, summed over the sources that traverse the link. The aggregate rate is a measure of demand for bandwidth at link l. Suppose the forward delay from source s to link l is denoted by τ^f_{ls}. Then the aggregate rate that is observed by link l at time t can be represented by

$$y_l(t) := \sum_{s \in S_l} x_s(t - \tau^f_{ls}) \tag{2}$$

The decentralization requirement dictates that each source s can adjust its rate $x_s(t)$ based only on $q_s(t)$, in addition to its own rate $x_s(t)$. In particular, the rate adjustment cannot depend on individual link prices $p_l(t)$ nor path prices of other sources. This can be modeled as:

$$\dot{x}_s(t) = F_s(x_s(t), q_s(t)) \tag{3}$$

where F_s calculates the amount of rate adjustment. Similarly, each link l can adjust its price $p_l(t)$ based only on $y_l(t)$, in addition to its own price $p_l(t)$. In particular, the price adjustment cannot depend on individual source rates $x_s(t)$ nor aggregate rate at other links. This can be modeled as:

$$\dot{p}_l(t) = G_l(y_l(t), p_l(t)) \tag{4}$$

where G_l represents the (implicit or explicit) price adjustment algorithm at link l. In general, the link and source algorithms can also depend on some internal state variable.

In summary, a general congestion control scheme can be decomposed into two algorithms. The source algorithm that adapts the rate to congestion in its path can be modeled by Eq. (3). The link algorithm that updates the price based on aggregate rate can be modeled by Eq. (4). The information used by these algorithms is not only local, but also delayed as expressed by Eqs. (1) and (2).

5.2. Example: Vegas/DropTail

We now describe a model of Vegas/DropTail, developed in Ref. 8, as an illustration. The model ignores slow start and fast retransmit/fast recovery, and only captures the behavior of congestion avoidance.

The price at link l turns out to represent queueing delay whose dynamics is modeled as

$$\dot{p}_l(t) = \frac{1}{c_l}(y_l(t) - c_l) =: G_l(y_l(t), p_l(t)) \tag{5}$$

at bottleneck links. To model the TCP Vegas algorithm, let d_s be the round-trip propagation delay for source s and assume the Vegas parameters satisfy $\alpha = \beta$ for all sources s. Then the rate is adjusted according to:

$$\dot{x}_s(t) = \frac{1}{(d_s + q_s(t))^2} \operatorname{sgn}\left(1 - \frac{x_s(t)q_s(t)}{\alpha d_s}\right) =: F_s(x_s(t), q_s(t)) \tag{6}$$

where $\operatorname{sgn}(z)$ is -1 if $z < 0$, 0 if $z = 0$, and 1 if $z > 0$. Here, $q_s(t)$ is the (delayed) end-to-end queueing delay in the path of source s, $d_s + q_s(t)$ is the round-trip time observed at source s at time t, and $x_s(t)q_s(t)$ is the number of packets that are buffered in the queues in the path. Hence, Eq. (6) says that the window (rate $\times$ round-trip time) is incremented or decremented at a rate of 1 packet per round-trip time, according as the number $x_s(t)q_s(t)$ of packets buffered in the path is smaller or greater than the target αd_s. In equilibrium, each source s maintains αd_s packets in its path.

6. EQUILIBRIUM AND STABILITY PROPERTIES

Equilibrium properties, such as fairness, utilization and QoS, and dynamic properties, such as stability, of a flow control scheme can be understood by studying the mathematical model specified by Eqs. (1) and (4). In this section, we illustrate how to analyze the model Eqs. (1) and (4).

6.1. Equilibrium

Under mild assumptions on the source algorithm F_s, we can associate a utility function $U_s(x_s)$ with source x_s that is a concave increasing function of its rate x_s. As previously mentioned, this means sources are greedy and there is a diminishing return as rate increases.

Consider the following constrained utility maximization problem:

$$\max_{x \geq 0} \quad \sum_s U_s(x_s) \tag{7}$$

$$\text{subject to} \quad yl \leq c_l \text{ for all links } l \tag{8}$$

The constraint (8) says that the aggregate source rate at any link does not exceed the capacity. From optimization theory, we can associate with the primal problem (7, 8) the following dual problem

$$\min_{p \geq 0} \sum_s \max_{x_s \geq 0}(U_s(x_s) - x_s q_s) + \sum_l c_l p_l \tag{9}$$

where $q_s = \sum_{l \in L_s} p_l$ is the sum of link prices p_l in the path of source s.

Suppose (x^*, p^*) is an equilibrium point of the model (1–4). Then it is proved in [11] that x^* solves the primal problem (Eqs. (7) and (8)) if and only if for all links l

$$y_l^* \leq c_l \quad \text{with equality if } p_l^* > 0 \tag{10}$$

Moreover, in this case, p^* solves the dual problem (9). Note that the condition (10), called *complementary slackness*, says that every bottleneck link is fully utilized, that is, input rate is equalized to capacity.

This interpretation has several implications. First, we can regard the source rates $x_s(t)$ in Eq. (3) as primal variables, the prices $p_l(t)$ in Eq. (4) as dual variables, and a congestion control mechanism (Eqs. (3) and (4)) as a distributed asynchronous computation over a network to solve the primal problem (7–8) and the dual problem (9). The equilibrium rates can be interpreted as utility maximizing, and the equilibrium prices (delay or loss) as Lagrange multipliers that measure the marginal increase in optimal aggregate utility for each unit of increment in link capacities.

Second, different source and link algorithms are just different ways to solve the same prototypical problem (7–9), with different utility functions. Even though TCP algorithms were not designed to solve any optimization problem, they have implicitly chosen certain utility functions by adjusting the source rate in a particular way. Take TCP Vegas algorithm (6) for example: in equilibrium, we have $\dot{x}_s = 0$ and hence $x_s^* q_s^* = \alpha d_s$. This implies a utility function of

$$U_s(x_s) = \alpha d_s \log x_s \tag{11}$$

See Ref. 11 for details and for utility functions of Reno. Moreover, the TCP algorithm F_s alone determines the equilibrium rate allocation by defining the underlying optimization problem. The role of link algorithm G_l is to ensure the complementary slackness condition and to stabilize the equilibrium.

Third, the equilibrium properties are all determined by the underlying optimization problem. Fairness, a property of the optimal rate vector x^*, is determined by the utility functions in the utility maximization (7–8). For Vegas, the log utility function in Eq. (11) implies that it achieves proportional fairness. Hence, as mentioned earlier, we can define fairness through the corresponding utility function. Moreover, since the utility function depends only on source algorithm F_s, fairness is independent of the link algorithm, as long as link prices depend only on aggregate rates $y_l(t)$, not on individual source rates $x_s(t)$.

Fourth, if the link algorithm G_l achieves the complementary slackness condition (10), then the network will be maximally utilized in equilibrium. QoS however may not be properly controlled if prices are coupled with QoS.

In this case, the value of the Lagrange multiplier p_l^* is determined not by the update algorithm G_l, but by the underlying optimization problem. In particular, if the number of sources sharing a link is large, or if the link capacity is small, then p_l^* will be large. If p_l^* represents queueing delay, as in TCP Vegas, or loss probability, as in TCP Reno, QoS can be poor. It is however possible to design link algorithms that decouple congestion measure with QoS such as loss or delay, so that the link prices converge to their equilibrium values determined by the primal problem while queues are kept small; see, Refs. 12 and 13. In this case, price information is no longer embedded in end-to-end delay and must be fed back to sources explicitly.

In summary, equilibrium properties of general source and link algorithms can be understood by regarding them as distributed primal-dual iterations to solve the primal and dual problems (7–9), where the utility functions are determined by the source algorithm F_s.

6.2. Stability

In general, the stability of the distributed nonlinear system with delay, specified by Eqs. (1)–(4), is very difficult to analyze (but see [10]). We can however understand its stability in the presence of delay around an equilibrium by studying the linearized model. In this section, we briefly summarize the stability properties of TCP Reno and TCP Vegas; see, Refs. 14–16 for details.

It is well known that the queue length under TCP Reno can oscillate wildly, with either DropTail or RED link algorithm, and it is extremely hard to reduce the oscillation by tuning RED parameters, Refs. 17 and 18. The additive-increase-multiplicative-decrease (AIMD) strategy employed by TCP Reno (and its variants such as NewReno and SACK) and noise-like traffic that are not effectively controlled by TCP no doubt contribute to this oscillation. It is shown in Ref. 15 however that protocol instability can have much larger effect on the oscillatory behavior than these factors. By instability, we mean severe oscillation in *aggregate* quantities, such as queue length and average window size. The analysis of the linearized delayed model shows that the system becomes unstable when delay increases, and more strikingly, when network capacity increases! This agrees with empirical experience that the current TCP performs poorly at large window sizes. Moreover, even if we smooth out AIMD, that is even if window is not adjusted on each acknowledgment arrival or loss event, but is adjusted periodically by the same *average* amount AIMD would over the same period, the oscillation persists. In particular, this implies that equation-based rate control will not help if the equation mimics the Reno dynamics. This suggests that TCP Reno/RED is ill-suited for future networks where capacities will be large.

This motivates the design of new source and link algorithms that maintain linear stability for general delay and capacity [19–23]. The main insight from this series of work is to scale down source responses with their own round-trip times and scale down link responses with their own capacities, in order to keep the gain over the feedback loop under control.

It turns out that the implicit link algorithm (5) of Vegas has exactly the right scaling with respect to capacity as used in the scalable design of Refs. 19 and 22. This built-in scaling with capacity makes Vegas potentially scalable to high bandwidth, in stark contrast to the AIMD algorithm of Reno and its variants. The source algorithm of Vegas, however, has a different scaling with respect to delay from those in Refs. 19 and 22, making it susceptible to instability in the presence of large delay. It is possible however to stabilize it by slightly modifying the rate adjustment algorithm (6) of Vegas; see Ref. 16 for details.

7. CONCLUSION

Flow control schemes are distributed and asynchronous algorithms to share network resources among competing users. The goal is to share these resources fairly and efficiently, and to provide good QoS in a stable, robust, and scalable manner. There is no automatic method to synthesize flow control schemes that will achieve these objectives. We have provided an introduction to mathematical models that can help understand and design such schemes, and have illustrated these models using TCP Vegas.

Acknowledgments

This article is a gentle introduction to some of the recent literature on flow control. We gratefully acknowledge the contribution of authors of these papers, only some of which are cited here, and in particular, that of my collaborators Sanjeewa Athuraliya, Hyojeong Choe, John Doyle, Ki-baek Kim, David Lapsley, Fernando Paganini, Larry Peterson, Jiantao Wang, Limin Wang, Zhikui Wang. Finally, we acknowledge the support of US National Science Foundation through grant ANI-0113425, US Army Research Office, the Caltech Lee Center for Advanced Networking, and Ciso.

BIOGRAPHY

Steven. H. Low received his B.S. degree from Cornell University and PhD from the University of California–Berkeley, both in electrical engineering. He was with AT&T Bell Laboratories, Murray Hill, from 1992 to 1996, with the University of Melbourne, Australia, from 1996 to 2000, and is now an associate professor at the California Institute of Technology, Pasadena. He was a corecipient of the IEEE William R. Bennett Prize Paper Award in 1997 and the 1996 R&D 100 Award. He is on the editorial board of IEEE/ACM Transactions on Networking. He has been a guest editor of the IEEE Journal on Selected Area in Communications, on the program committee of major networking conferences. His research interests are in the control and optimization of communications networks and protocols. His home is netlab.caltech.edu and email is slow@caltech.edu.

BIBLIOGRAPHY

1. V. Jacobson, Congestion avoidance and control, *Proc. SIGCOMM'88, ACM*, August 1988. An updated version is available via *ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z*.
2. E. Varvarigos and T. Varvarigou, Computer communications protocols, in J. G. Proakis, ed., New York, *Encyclopedia of Telecommunications*, Wiley, 2002.

3. J. Aweya, Transmission control protocol, in John Proakis, ed. *Encyclopedia of Telecommunications*, New York, Wiley, 2002.

4. E. J. Hernandez-Valencia, L. Benmohamed, R. Nagarajan, and S. Chong, Rate control algorithms for the ATM ABR service, *Eur. Trans. Telecomm.* **8**: 7–20 (1997).

5. L. S. Brakmo and L. L. Peterson. TCP Vegas: end-to-end congestion avoidance on a global Internet, *IEEE J. Select. Areas Comm.* **13**(8): 1465–1480 (October 1995) *http://cs.princeton.edu/nsg/papers/jsac-vegas.ps.*

6. D. Bertsekas and R. Gallager. *Data Networks*, 2nd ed. Prentice-Hall, 1992.

7. F. P. Kelly, A. Maulloo, and D. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, *J. Operations Res. Soc.* **49**(3): 237–252 (March 1998).

8. S. H. Low, L. L Peterson, and L. Wang, Understanding Vegas: a duality model, *J. ACM* **49**(2): 207–235 (March 2002). *http://netlab.caltech.edu.*

9. S. Floyd and V. Jacobson, Random early detection gateways for congestion avoidance, *IEEE/ACM Trans. Networking* **1**(4): 397–413 (August 1993). *ftp://ftp.ee.lbl.gov/papers/early.ps.gz.*

10. S. H. Low and D. E. Lapsley, Optimization flow control, I: basic algorithm and convergence, *IEEE/ACM Trans. Networking* **7**(6): 861–874, (December 1999). *http://netlab.caltech.edu.*

11. S. H. Low, A duality model of TCP and queue management algorithms, In *Proc. ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management* (*updated version*) (September 18–20, 2000). *http://netlab.caltech.edu.*

12. S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin, REM: active queue management. *IEEE Network* **15**(3): 48–53 (May/June 2001). Extended version in *Proc. ITC17*, Salvador, Brazil, September 2001. *http://netlab.caltech.edu.*

13. C. Hollot, V. Misra, D. Towsley, and W. B. Gong, On designing improved controllers for AQM routers supporting TCP flows. In *Proc. IEEE Infocom* (April 2001). *http://www-net.cs.umass.edu/papers/papers.html*.

14. C. Hollot, V. Misra, D. Towsley, and W. B. Gong, A control theoretic analysis of RED. In *Proc. of IEEE Infocom* (April 2001). *http://www-net.cs.umass.edu/papers/papers.html*.

15. S. H. Low et al., Dynamics of TCP/RED and a scalable control. In *Proc. IEEE Infocom* (June 2002). *http://netlab.caltech.edu.*

16. H. Choe and S. H. Low. *Stabilized Vegas*, In *Proc. of 39th Annual Allerton Conference on Communication, Control, and Computing*, (October 2002). *http://netlab.caltech.edu.*

17. M. May, T. Bonald, and J.-C. Bolot, Analytic evaluation of RED performance, In *Proc. IEEE Infocom* (March 2000).

18. M. Christiansen, K. Jeffay, D. Ott, and F. D. Smith, Tuning RED for web traffic, In *Proc. ACM Sigcomm* (2000).

19. F. Paganini, John C. Doyle, and S. H. Low, *Scalable laws for stable network congestion control*. In *Proc. Conference on Decision and Control* (December 2001). *http://www.ee.ucla.edu/paganini.*

20. G. Vinnicombe, On the stability of end-to-end congestion control for the Internet. Technical report, Cambridge University, CUED/F-INFENG/TR.398, (December 2000).

21. G. Vinnicombe, Robust congestion control for the Internet. Submitted for publication, 2002.

22. F. Paganini, Z. Wang, S. H. Low, and J. C. Doyle. A new TCP/AQM for stability and performance in fast networks. In *Proc. of 39th Annual Allerton Conference on Communication, Control, and Computing* (October 2002).

23. S. Kunniyur and R. Srikant. A time-scale decomposition approach to adaptive ECN marking. *IEEE Trans. Automatic Control* (June 2002).

NETWORK RELIABILITY AND FAULT TOLERANCE

MURIEL MÉDARD
Massachusetts Institute of Technology
Cambridge, Massachusetts

STEVEN S. LUMETTA
University of Illinois
Urbana — Champaign
Urbana, Illinois

1. INTRODUCTION

The majority of communications applications, from cellular telephone conversations to credit card transactions, assume the availability of a reliable network. At this level, data are expected to traverse the network and to arrive intact at their destination. The physical systems that compose a network, on the other hand, are subjected to a wide range of problems, ranging from signal distortion to component failures. Similarly, the software that supports the high-level semantic interface often contains unknown bugs and other latent reliability problems. Redundancy underlies all approaches to fault tolerance. Definitive definitions for all concepts and terms related to reliability, and, more broadly, dependability, can be found in the book by Anderson et al. [1].

Designing any system to tolerate faults first requires the selection of a fault model, a set of possible failure scenarios along with an understanding of the frequency, duration, and impact of each scenario. A simple fault model merely lists the set of faults to be considered; the decision regarding inclusion in the set is based on a combination of expected frequency, impact on the system, and feasibility or cost of providing protection. Most reliable network designs address the failure of any single component, and some designs tolerate multiple failures. In contrast, few attempt to handle the adversarial conditions that might occur in a terrorist attack, and cataclysmic events are almost never addressed at any scale larger than a city.

The temporal characteristics of faults vary widely, but can be roughly categorized as permanent, intermittent, or transient. Failures that prevent a component from functioning until repaired or replaced, such as the destruction of a network fiber by a backhoe, are considered permanent. Failures that allow a component to function properly some of the time are called *intermittent*. Damaged connectors and electrical components sometimes produce intermittent faults, operating correctly until mechanical vibrations or thermal variations cause a failure, and recovering when conditions change again. The last category, transient

faults, is usually the easiest to handle. Transient faults range from changes in the contents of computer memory due to cosmic rays, or bit errors due to thermal noise in a demodulator, and are typically infrequent and unpredictable. The difference between an intermittent fault and a transient fault is sometimes solely one of frequency; for transient faults, a combination of error-correcting codes and data retransmission usually provides adequate protection.

Redundancy takes two forms, spatial and temporal. *Spatial redundancy* replicates the components or data in a system. Transmission over multiple paths through a network and the use of error correction codes are examples of spatial redundancy. Temporal redundancy underlies automatic repeat request (ARQ) algorithms, such as the sliding-window abstraction used to support reliable transmission in the Internet's Transmission Control Protocol (TCP). A reliable network typically provides both spatial and temporal redundancy to tolerate faults with differing temporal persistence. Spatial redundancy is necessary to overcome permanent failures in physical components, while temporal redundancy requires fewer resources and is thus preferable when dealing with transient errors.

Beyond the selection of a fault model, several additional problems must be considered in the design of a fault-tolerant system. A system must be capable of detecting each fault in the model, and must be able to isolate each fault from the functioning portion of the system in a manner that prevents faulty behavior from spreading. As a fault detection mechanism may detect more than one possible fault, a system must also address the process of fault diagnosis (or localization), which narrows the set of possible faults and allows more efficient fault isolation techniques to be employed. An error identified by a system need not necessarily be narrowed down to a single possible fault, but a smaller set of possibilities usually allows a more efficient strategy for recovery.

Fault isolation boundaries are usually designed to provide fail-stop behavior for the desired fault model. The term *fail stop* implies that incorrect behavior does not propagate across the fault isolation boundary; instead, failed components cease to produce any signals. Fail stop does not imply self-diagnosis; components adjacent to a failed component may diagnose the failure and deliberately ignore any signals from the failed component, but the physical system design must allow such a decision. In a router, for example, the interconnect between cards controlling individual links must provide electrical isolation to support fail-stop behavior for failed cards. A bus-based computer interconnect does not allow for fail stop, as nothing can prevent a failed card from driving the bus lines inappropriately. In modern, high-end servers, such buses have been replaced by switched networks with broadcast capability in order to enable such isolation. The eradication of similar phenomena in the move from shared to switched Ethernets in the mid-1990s was one of the main administrative advantages of the change, as failed hosts are much less likely to render a switched network unusable by flooding it with continuous traffic.

Two models of network service have dominated research and commercial networking. The first is the telephony network, or more generally a network in which quasipermanent routes called *circuits* deliver fixed data capacity from one point to another. In digital telephony, a voice circuit requires 64 kbps (kilobits per second); a single lightpath in a wavelength-division-multiplexed (WDM) optical network may deliver up to 40 Gbps, but is conceptually similar to the circuit used to carry a phonecall. The second network service model is the packet-switched data network, which evolved from the early ARPANET and NSFNET projects into the modern Internet. Packet-switched networks seldom provide strong guarantees on delivered data rate or maximum delay, but are typically more efficient than circuit-oriented designs, which must base guaranteed agreements on worst-case traffic load scenarios.

For the purposes of our discussion, the key difference between these two models lies in the fact that applications using packet-switched networks can generally tolerate more serious service disruptions than can those based on circuit-switched networks. The latter class of applications may assume that data rate, delay, and jitter guarantees provided by the network will be honored even when failures occur, whereas minor disruptions may occur even in normal circumstances on packet-switched networks because of fluctuations in traffic patterns and loads. Fault tolerance issues are thus addressed in markedly different ways in the two types of networks. In packet-switched networks like the Internet, users currently tolerate restoration times of minutes [2,3], whereas fault tolerance for circuit-switched networks can be considered a component of quality of service (QoS) [4,5], and is typically achieved in milliseconds, or, at worst, seconds.

The majority of this article focuses on fault tolerance issues in high-speed backbone networks, such as wide-area networks (WANs) and metropolitan-area networks (MANs). Such networks are predominantly circuit-based and carry heavy traffic loads. As even a short downtime may cause substantial data loss, rapid recovery from failure is important, and these networks require high levels of reliability. Backbone networks generally are implemented using optical transmission and, conversely, fault tolerance in optical networks is typically considered in the context of backbone networks [6,7]. In these networks, a failure may arise because a communications link is disconnected or a network node becomes incapacitated. Failures may occur in military networks under attack [8], as well as in public networks, in which failures, albeit rare, can be extremely disruptive [9, Chap. 8].

The next section provides an overview of fault detection mechanisms and the basic strategies available for recovery from network component failures. Sections 3 and 4 build on these basics to illustrate recovery schemes for high-speed backbone networks. Sections 5 and 6 examine simple and more complex topologies and discuss the relationship between topology and recovery. Section 5 highlights ring topologies, as they are a key architectural component of high-speed networks. Section 6 extends the concepts developed for rings by overlaying logical ring topologies over physical mesh topologies. We also discuss

some link- and node-based reliability schemes that are specifically tailored to mesh networks. Although the text focuses on approaches to fault tolerance in high-speed backbone networks, many of the principles also apply to other types of networks. In Section 7, we move away from circuit-switched networks and examine fault tolerance for packet-switched networks, and in particular the Internet. Finally, Section 8 discusses reliability issues for local-area networks (LANs).

2. FAILURE DETECTION AND RECOVERY

A wide variety of approaches have been employed for detection of network failures. In electronic networks with binary voltage encodings (e.g., RS-232), two nonzero voltages are chosen for signaling. A voltage of zero thus implies a dead line or terminal. Similarly, electronic networks based on carrier modulation infer failures from the absence of a carrier. Shared segments such as Ethernet have been more problematic, as individual nodes cannot be expected to drive the segment continuously. In such networks, many failures must be detected by higher levels in the protocol stack, as discussed later in this section.

The capacity of optical links makes physical monitoring a particularly important problem, and many techniques have been explored and used in practice. Optical encoding schemes generally rely on on/off keying; that is, the presence of light provides one signal, and its absence provides a second. With single-wavelength optics, information must be incorporated into the channel itself. One approach is to monitor time-averaged signal power, using an encoding scheme that results in a predictable distribution of ON and OFF frequencies. A second approach utilizes overhead bits in the channel, allowing bit error rate (BER) sampling at the expense of restricting the data format used by higher levels of the protocol stack. A third approach employs a sideband to carry a pilot tone. These approaches are complementary, and can be used in tandem.

A WDM system typically applies the single-wavelength techniques just mentioned to each wavelength, but the possibility of exploiting the multiplexing to reduce the cost of failure detection has given rise to new techniques. A single wavelength, for example, can be allocated to provide accurate estimates of BER along a link. Unfortunately, this approach may fail to detect frequency-dependent signal degradation. Pairing of monitoring wavelengths with data wavelengths reduces the likelihood of missing a frequency-dependent failure, but is too inefficient for most networks.

The approaches discussed so far have dealt with failure detection at the link level. With circuit-switched networks, the receiver on any given path can directly monitor accumulated effects along the entire path. The techniques discussed for a single wavelength can also be employed for a full path with optically transparent networks. With networks that perform optoelectronic conversion at each node, only in-band information is retained along the length of the path and overhead in the data format is typically necessary for failure detection. Path-based approaches are advantageous in the sense that they may cover a broader set of possible failures. They get to the root of the problem; something went wrong getting from the sender to the receiver. Link-based approaches, however, make fault localization simpler, an important benefit in finding and repairing problems in the network. In practice, most backbone networks use a combination of link and path detection techniques to obtain both benefits.

Additional fault tolerance is often included in higher levels of a network protocol stack. Most protocols used for data networking (as opposed to telephony), for example, include some redundancy coding for the purposes of error detection. Typically, feedback from these layers is not provided to the physical layer, although some exceptions do exist in LANs, such as the use of periodic packet transmissions and inference of failures when no packet arrives (see Section 8 for more detail). Instead, the error detection schemes allow the network to tolerate transient errors through temporal redundancy, namely, retransmission. Voice channels and other redundant forms of data also utilize error correction or other error tolerance techniques in some cases. A telephone circuit crossing an asynchronous transfer mode (ATM) network may lose an occasional cell to a cyclic redundancy check (CRC) failure. In such a case, the cell is discarded, and the voice signal regenerated by interpolation from adjacent cells. This interpolation suffices to make a single cell loss undetectable to humans; thus, as long as the transient errors occur infrequently, no loss is noticed by the people using the circuit.

The choice of failure detection methods used in a backbone network is intertwined with the choice of strategies for restoring circuits that pass through a failed element of the network. Path monitoring, for example, does not readily provide information for failure localization. Correlated failures between paths may help to localize failures, but typically a more careful investigation must be initiated to find the problem. Path monitoring also requires that failure information propagate to the endpoints of the path, delaying detection. Link monitoring allows more rapid and local response to failures, but does not require such an approach. Instead, failure information can be propagated to the ends of each path crossing a link, while the localized failure information is retained for initiating repairs and for dynamic construction of future paths. At the algorithmic level, circuit rerouting schemes can be broadly split into path-based and link- or node-based approaches.

Prompted by the increasing reliance on high-speed communications and the requirement that these communications be robust to failure, backbone networks have generally adopted self-healing strategies to automatically restore functionality. The study of self-healing networks is often classified according to the following three criteria [e.g., 10,11]: (1) the use of link (line) rerouting versus path (or end-to-end) rerouting, (2) the use of centralized computation versus distributed computation, and (3) the use of precomputed versus dynamically computed routes. A succinct comparison of the different options can be found in the book by Wu [12, pp. 291–294] and the paper by Johnson et al. [13]. For path recovery, when a failure leaves a node disconnected from the primary route, a backup route, which may or may not share nodes and links with the primary route, is used. *Link rerouting* usually refers

to the replacement of a link by links connecting the two end nodes of the failed link. When the rerouting is precomputed, the method is generally termed *protection*. Thus, *path protection* refers to precomputed recovery applied to connections following a particular path across a network. *Link or node protection* refers to precomputed recovery of all the traffic across a failed link or node, respectively. Figure 1 illustrates path and link rerouting. Protection routes are precomputed at a single location, and are thus centralized, although some distributed reconfiguration of optical switches may be necessary before traffic is restored. Restoration techniques, on the other hand, can rely on distributed signaling between nodes or on allocation of a new path by a central manager.

3. PATH-BASED SCHEMES

Protection schemes, in which recovery routes are preplanned, generally offer better recovery speeds than restoration approaches, which search for new routes dynamically in response to a failure and generally involve software processing [14,15]. The *Synchronous Optical* NET*work* (SONET) specification, for example, requires that recovery time with protection approaches be under 60 ms. Recovery can be achieved in tens of milliseconds using optomechanical add/drop multiplexers [16,17], and in a few microseconds using acoustooptical switches [18,19]. In contrast, dynamic distributed restoration using digital cross-connect systems (DCSs) for ATM or SONET [20–23] typically targets a 2-s recovery-time goal [17,24,25]. Dynamic centralized path restoration for SONET [26] may even take minutes [24,27]. The performance of several algorithms has been reviewed [28,29]. Restoration typically requires less protection capacity, however.

In this section, we focus on path protection, as the majority of current backbone networks utilize such techniques. Path protection trades longer recovery times for reduced capacity requirements relative to the link-based approaches discussed in the next section. These tradeoffs are discussed in more depth elsewhere [30–32]. Path protection involves finding, for each circuit, a backup route (or path). Figure 2 shows two primary routes and their corresponding backup routes. For each circuit, the two routes do not overlap on any links, implying that no single link failure can affect both a primary route and its backup.

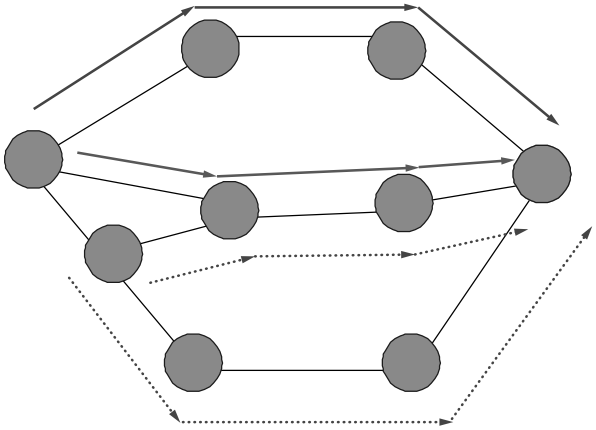


Figure 2. Path protection and associated bandwidth sharing on the backup.

Path protection can itself be divided into several categories: one-plus-one (written $1+1$), one-for-one (written $1:1$), and one-for-N ($1:N$). In the first case, all data are sent along two routes simultaneously, a primary and a backup. The two routes are link-disjoint for tolerance to link failures, or node-disjoint (excluding source and destination) for tolerance to node failures. The receiver monitors incoming traffic on the primary route; when a component along the primary route fails, the receiver switches to the backup signal. The backup route is typically the longer of the two, ensuring that no data are lost because of a single failure. Because both primary and backup routes carry live traffic, the $1+1$ approach is sometimes referred to as *live backup*. Recovery using live backup is extremely fast, as it is based on a local decision by a single node. Protection capacity requirements are high, however, as the backup channel cannot be shared among connections.

The other two approaches, together known as *event-triggered backup*, require less network capacity devoted to protection than does live backup. The penalty is loss of some data when a failure occurs as well as slower restoration times relative to live backup. With event-triggered backup, the backup path is activated only after a failure is detected. As with live backup, the receiver monitors the primary path, but rather than acting locally when it detects a failure, the receiver notifies the sender that a failure has occurred on the primary path, at

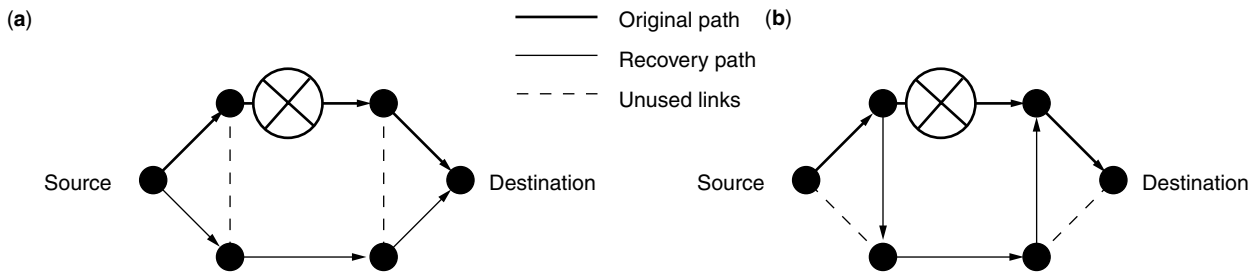


Figure 1. Path (**a**) and link (**b**) rerouting; the failure is marked with an ⊗.

which point the sender begins sending traffic on the backup path. All data transmitted or in flight between the time of the failure and the sender switching over to the backup route are lost. With $1:1$ protection, optical cross-connects are preconfigured for a particular route. Sharing and reuse of the backup route is therefore somewhat restricted. With $1:N$ protection, backup resources can be shared between any set of circuits for which the primary routes do not have resources in common, as illustrated by the two circuits in Fig. 2. Two primary routes passing through a single link, for example, cannot share backup resources; if that link should fail, only one of the two routes could be recovered. The need to configure optical cross-connects along the backup route adds further delay to restoration time and increases data loss with $1:N$ protection, but sharing of backup resources reduces protection capacity requirements by roughly 15–30% relative to $1+1$ protection in an all-optical network. Traffic grooming, in which traffic can be assigned to wavelengths in granularities smaller than those of whole wavelengths, can allow for much more effective sharing.

All forms of protection require adequate spatial redundancy in the network topology to allow advance selection of two disjoint routes for each circuit. For link (or node) protection, this requirement translates to a need for two-edge (-vertex)-redundant graphs; in other words, a graph must remain completely connected after removal of any single edge (vertex, along with all adjacent edges). For path protection, the same condition suffices, as shown by Menger's theorem [33,34], which states that, given a two-edge (-vertex)-redundant graph, two edge- (-vertex)-disjoint routes can be found between any two vertices in the graph. A variety of schemes based on Menger's theorem have been proposed, such as subnetwork connection protection (SNCP) and variations thereof [35–40], which establish two paths between every pair of nodes. However, Menger's theorem is only the starting point for designing a recovery algorithm, which must also consider rules for routing and reserving spare capacity. Path protection over arbitrary redundant networks can also be performed with trees, for example, which are more bandwidth-efficient for multicast traffic [41,42].

With ATM, path rerouting performed by the private network node interface (PNNI) tears down virtual circuit (VC) connections after a failure, forcing the source node to establish a new end-to-end route. Backup virtual paths (VPs) can be predetermined [43] or selected jointly by the end nodes [44]. Source routing, which is used by ATM PNNI, can be preplanned [45] or partially preplanned [46].

4. LINK- AND NODE-BASED SCHEMES

As with path rerouting, methods commonly employed for link and node rerouting in high-speed networks can be divided into protection and restoration, although some hybrid schemes do exist [23]. The two types offer a tradeoff between adaptive use of backup (or "spare") capacity and speed of restoration [25,46]. Dynamic restoration typically involves a search for a free path using backup capacity [47–49] through broadcasting of help messages [13,20,22,24,25,28]. The performance of several algorithms has also been discussed [28,29]. Overheads due to message passing and software processing render dynamic processing slow. For dynamic link restoration using digital cross-connect systems, a 2-s restoration time is a common goal for SONET [17,20–22,24,25]. Preplanned methods, or link protection, depend mostly on lookup tables and switches or add/drop multiplexers. For all-optical networks, switches may operate in a matter of microseconds or nanoseconds and propagation delay dominates switching time.

Link and node protection can be viewed as a compromise between live and event-triggered path protection. Although not as capacity-efficient as $1:N$ path protection [32,44], link protection is more efficient than live path backup, as backup capacity is shared between links. All traffic carried by a failed link or node is recovered independent of the circuits or end-to-end routes associated with the traffic. In particular, the two nodes adjacent to the failure initiate recovery, and only nodes local to the failure typically take part in the process. Backup is not live, but triggered by a failure. Overviews of the different types of protection and restoration methods and comparison of the tradeoffs among them can be found elsewhere in the literature [15,30,50–52].

The fact that link and node protection are performed independently of the particular traffic being carried does provide an additional benefit. In particular, these approaches are independent of traffic patterns, and can be preplanned once to support arbitrary dynamic traffic loads. Path protection does not provide this feature; new protection capacity may be necessary to support additional circuits, and routes chosen without knowledge of the entire traffic load, as is necessary when allocating routes online, are often suboptimal. This benefit makes link and node restoration particularly attractive at lower layers, at which network management at any given point in the network may not be aware of the origination and destination, or of the format [24] of all the traffic being carried at that location.

Link rerouting in ATM usually involves a choice of new routes by nodes adjacent to the failure [53,54].

5. RINGS

Rings have emerged as one of the most important architectural building blocks for backbone networks in the MAN and WAN arenas. While ring networks can support both path-based and link- or node-based schemes for reliability, rings merit a separate discussion because of the practical importance and special properties of ring architecture.

Rings are the most common means of implementing both path and link protection in SONET, which is the dominant protocol in backbone networks. The building blocks of SONET networks are generally self-healing rings (SHRs) and diversity protection (DP) [12,16,19,24,55–64]. SHRs are unidirectional path-switched rings (UPSRs) or bidirectional line-switched rings (BLSRs), while DP refers to physical redundancy in which a spare link (node) is assigned to one or several links (nodes) [12, pp. 315–322].

In SONET, path protection is usually performed by a UPSR, as illustrated by Fig. 3a, in which a route in the

clockwise direction is replaced by a route in the counterclockwise direction in response to a failure. Link rerouting is performed in a SONET BLSR using a technique known as *loopback*, in which traffic is sent back in the direction from which it came. Figure 3b illustrates this operation, in which a route in the clockwise direction is looped back (redirected) onto a counterclockwise route at the node adjacent to the failure. After traveling around the entire ring to the other end of the failed link, the route is looped back again away from the failed link, rejoining the original route. Note that, with the exception of the failed link, the final backup route includes all of the original route. The waste of bandwidth due to traversing both directions on a single link, known as *backhauling*, can be eliminated by looping back at points other than the failure location [65,66]. In case of failure of a node on a BLSR, the failure is handled in a manner similar to that for a link failure. The failure of a node is equivalent to the failure of a metalink consisting of the node and the two links adjacent to it. The only difference is that network management must be able to purge any traffic directed to the failed node from the network.

Loopback operations can be performed on entire fibers or on individual wavelengths. With fiber-based loopback, all traffic carried by a fiber is backed by another fiber, regardless of how many wavelengths are involved. If traffic is allowed in both directions in a network, fiber-based loopback relies on four fibers, as illustrated in Fig. 4. In WDM-based loopback, restoration is performed in a wavelength-by-wavelength basis.

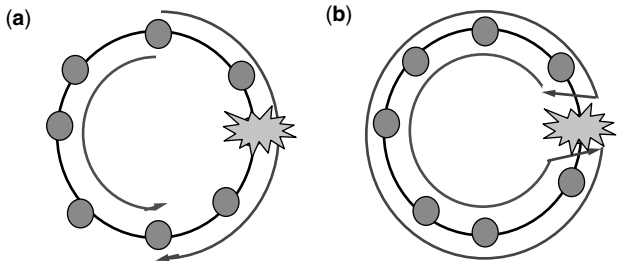


Figure 3. Path protection and node protection in a ring: (**a**) UPSR — automatic path switching; (**b**) BLSR — link/node rerouting.

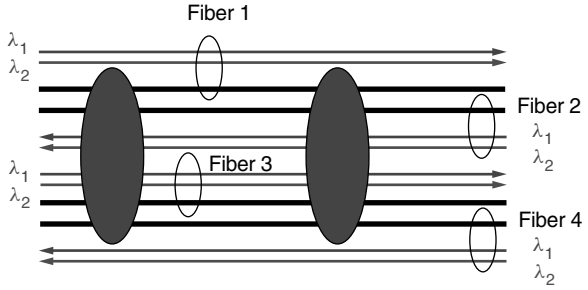


Figure 4. Four-fiber system with fiber-based loopback. Primary traffic is carried by fiber 1 and fiber 2. Backup is provided by fiber 3 for fiber 1 and by fiber 4 for fiber 2.

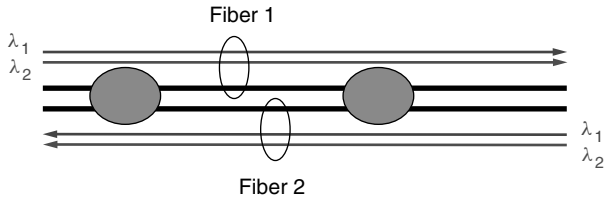


Figure 5. Two-fiber WDM-based loopback. Primary traffic is carried by fiber 1 on λ_1 and by fiber 2 on λ_2. Backup is provided by on λ_1 on fiber 2 for λ_1 on fiber 1. λ_2 on fiber 2 is backed up by λ_2 on fiber 1.

WDM-based loopback requires at least two fibers. Figure 5 illustrates WDM-based loopback. A two-fiber counterpropagating WDM system can be used for WDM-based loopback, even if traffic is allowed in both directions. Note that WDM loopback as shown in Fig. 5 does not require any change of wavelength; traffic initially carried by λ_1 is backed up by the same wavelength. Obviating the need for wavelength changing is economical and efficient in WDM networks. One could, of course, back up traffic from λ_1 on fiber 1 onto λ_2 on fiber 2, if there were advantages to such wavelength changing, for instance in terms of wavelength assignment for certain traffic patterns. We can easily extend the model to a system with more fibers, as long as the backup for a certain wavelength on a certain fiber is provided by some wavelength on another fiber. Moreover, we may change the fiber and/or wavelengths from one fiber section to another. For instance, the backup to λ_1 on fiber 1 may be λ_1 on fiber 2 on a two-fiber section and λ_2 on fiber 3 on another section with four fibers. Note, also, that we could elect not to back up λ_1 on fiber 1 and instead use λ_1 on fiber 1 for primary traffic. The extension to systems with more fibers, interwavelength backups and backups among fiber sections can be readily done.

The finer granularity of WDM-based recovery systems provides several advantages over fiber-based systems. First, if fibers carry at most half of their total capacity, only two fibers rather than four are necessary to provide recovery. Thus, a user need only lease two fibers, rather than paying for unused bandwidth over four fibers. On existing four-fiber systems, fibers could be leased by pairs rather than fours, allowing two leases of two fibers each for a single four-fiber system. The second advantage is that, in fiber based-systems, certain wavelengths may be selectively given restoration capability. For instance, half the wavelengths on a fiber may be assigned protection, while the rest may have no protection. Different wavelengths may thus afford different levels of restoration QoS, which can be reflected in pricing. In fiber-based restoration, all the traffic carried by a fiber is restored via another fiber. If each fiber is less than half full, WDM-based loopback can help avoid the use of counterpropagating wavelengths on the same fiber. Counterpropagating wavelengths on the same fiber are intended to enable duplex operation in situations that require a full fiber's worth of capacity in each direction and that have scarce fiber resources. However, counterpropagation on the same fiber is onerous and reduces the number of wavelengths that a fiber can carry with respect to unidirectional propagation. WDM-based loopback may make using two unidirectional

fibers preferable to using two counterpropagating fibers, for which one fiber is a backup for the other.

When more than one ring is required, rings must be interconnected. In SONET, the usual method to handle nodes shared between rings is called *matched nodes*. Figure 6 shows matched nodes under normal operating conditions. Consider traffic moving from ring 1 to ring 2; traffic in the reverse direction is handled similarly. Under normal operation, matched node 1 is responsible for all interring communications. Matched node 1 houses an add/drop multiplexer (ADM) that performs a drop and continue operation. The drop and continue operation consists of duplicating all traffic through matched node 1 and transmitting it to matched node 2. Thus, matched node 2 has a live backup of all the traffic arriving to matched node 1, and mirrors the operation of matched node 1. However, under normal operating conditions, ring 2 disregards the output from matched node 2. Failure of any node other than the primary matched node is handled by a single ring in a standalone manner. Failure of the secondary matched node treats intraring and interring traffic differently. Note that, depending on the failure mode of the primary matched node, the failure may be seen by both rings or by a single ring. Indeed, failures may occur only on access cards interfacing with one or the other ring, or a wholesale failure may be detected by both rings. Loopback is performed on all the rings that see the failure. Intraring traffic is recovered within the ring wherein it lies, and interring traffic is handled by the second node. How to extend matched nodes to cases other than simple extensions of the topology shown in Fig. 6 is generally unknown. For instance, the case in which a node is shared by more than one ring is difficult. Similarly, the case in which two adjacent rings share links without duplication of resources, as shown in Fig. 8, is complicated in the case of shared nodes.

Many schemes besides SONET exist or have been proposed to enable ring-based networks, usually using optical fiber as the medium. The proprietary protocol Fiber Distributed Data Interface (FDDI) [67–69] is such a scheme. Both FDDI and IEEE 802.5 control access to the ring by passing an electronic token from node to node. Only the node with the token is allowed to transmit. Multiple ring topologies may be interconnected through a hub [70], or rings may coexist in a logically interconnected fashion over a single physical ring [71–73], or rings may be arranged hierarchically [70,74,75].

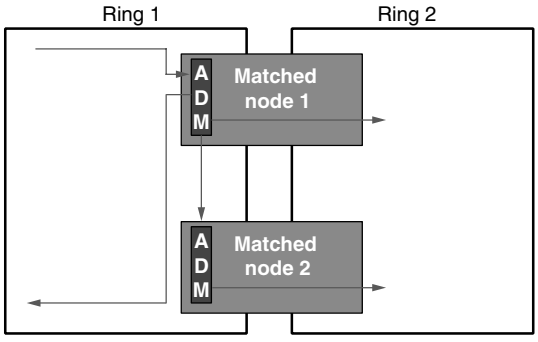


Figure 6. Matched nodes.

The IEEE 802.17 Resilient Packet Ring (RPR) Working Group has been set up to investigate the use, mainly at the MAN, of an optical ring architecture coupled with a packet-based MAC. The purpose of this project is to combine the robustness of rings with a flexible MAC that is well suited to current optical access applications [76].

6. MESH NETWORKS

In this section, we expand our exploration of topologies to redundant meshes. Restricting a network to use only DP and SHRs is a constraint that has cost implications for building and expanding networks [63]; see Stoer's book [34] for an overview of the design of topologies under certain reliability constraints. Ring-based architectures may be more expensive than meshes [63,77], and as nodes are added, or networks are interconnected, ring-based structures may be difficult to preserve, thus limiting their scalability [12,63,78]. However, rings are not necessary to construct fault-tolerant networks [79,80]. Mesh-based topologies can also provide redundancy [34,78,81]. Even if we constrain ourselves to always use ring-based architectures, such architectures may not easily bear changes and additions as the network grows. For instance, adding a new node, connected to its two nearest node neighbors, will preserve mesh structure, but may not preserve ring structure. Our arguments indicate that, for reasons of cost and extensibility, mesh-based architectures are more promising than interconnected rings.

Algorithmic approaches to general mesh restoration are often difficult, however, and implementations can be substantially more complex. To address this problem, many techniques attempt to find rings within the meshes. Overlays using rings are obtained by placing cycles atop existing mesh networks. Each such cycle creates a ring. Service protection or restoration is then generally obtained on each ring as though it were a physical ring. Covering mesh topologies with rings is a means of providing both mesh topologies and distributed, ring-based restoration. Numerous approaches ensure link restorability by finding covers of rings for networks. Many of these techniques have been proposed in the context of backbone networks in order to enable recovery over mesh topologies.

One such approach is to cover nodes in the network by rings [56]. In this manner, a portion of links are covered by rings. If primary routings are restricted to the covered links, link restoration can be effected on each ring in the same manner as in a traditional SHR, by routing backup traffic around the ring in the opposite direction to the primary traffic. Using such an approach, the uncovered links can be used to carry unprotected traffic; that is, traffic that may not be restored if the link that carries it fails. However, under some conditions it may not be possible to cover all nodes with a single ring, or the length of the resulting may be undesirable. A large ring forces long routes for many connections. Such long routes have several drawbacks, both from the point of view of routing (reduced wavelength-assignment efficiency) and from the point of view of communications (excessive jitter).

To allow every link to carry protected traffic, other ring-based approaches ensure that every link is covered by a

ring. One approach to selecting such covers is to cover a network with rings so that every link is part of at least one ring [82]. Several issues arise concerning the overlap and interconnection of such rings. Many of these issues are similar to issues encountered in SONET network deployment. The two main issues are management of links logically included in two rings and node management for ring interconnection.

The first issue concerns the case in which a single link is located on two rings. If that link bears a sufficient number of fibers or wavelengths, the two rings can be operated independently over that link, as shown in Fig. 7. However, the resources available to the overlay network may require sharing the resources over that link. Figure 8 shows such a network, in which only a single wavelength is available to the overlay network. In such a case, the logical fibers must be physically routed through available physical fibers, with network management acting to ensure that conflicts are avoided on the shared span. Such operations incur significant overhead.

The second issue relates to node interconnection among rings. Minimizing the amount of fiber required to obtain redundancy using ring covers is equivalent to finding the minimum cycle cover of a graph, an NP-complete problem [83,84], although bounds on the total length of the cycle cover may be found [85].

A more recent approach to ring covers, intended to overcome the difficulties of previous approaches, is to cover every link with exactly two rings, each with two fibers. The ability to perform loopback style restoration over any link in mesh topologies was first introduced in the late 1990s [86,87], using certain types of ring covers. In particular, Ellinas et al. [86] consider link failure restoration in optical networks with arbitrary two-link redundant arbitrary mesh topologies and bidirectional links. The approach is an application of the double-cycle ring cover [88–90], which selects cycles in such a way that each edge is covered by two cycles. For planar graphs, the problem can be solved in polynomial time; for nonplanar graphs, it is conjectured that double-cycle covers exist, and a counterexample must have certain properties unlikely to occur in network graphs [91]. Cycles can be used as rings to perform restoration. Each cycle corresponds to either a primary or a secondary two-fiber ring. Let us consider a link covered by two rings, rings 1 and 2. If we assign a direction to ring 1 and the opposite direction to ring 2, ring-based recovery using the double-cycle cover uses ring 2 to back up ring 1. This recovery technique is similar to recovery in conventional SHRs, except that the two rings that form four-fiber SHRs are no longer collocated over their entire lengths. In the case of four fiber systems, with two fibers in the same direction per ring, we have fiber-based recovery, because fibers are backed up by fibers. Extending this notion to WDM-based loopback, each ring is both primary for certain wavelengths and secondary for the remaining wavelengths. For simplicity, let us again consider just two wavelengths. Figure 9 shows that we cannot assign primary and secondary wavelengths in such a way that a wavelength is secondary or primary over a whole ring.

The use of double-cycle covers can also lead to asymmetric restoration times for a bidirectional connection. In particular, the links and nodes used to recover traffic crossing a link often depend on the direction of the traffic, with each direction being recovered by a separate cycle. The two directions on a link thus have different restoration times and timing jitter, which can lead to

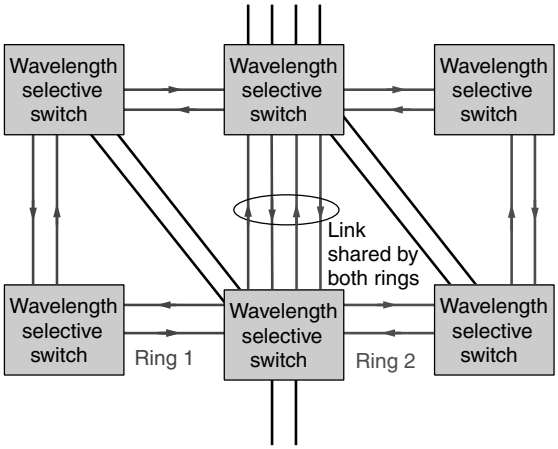


Figure 7. Two rings traversing separate resources over the same link.

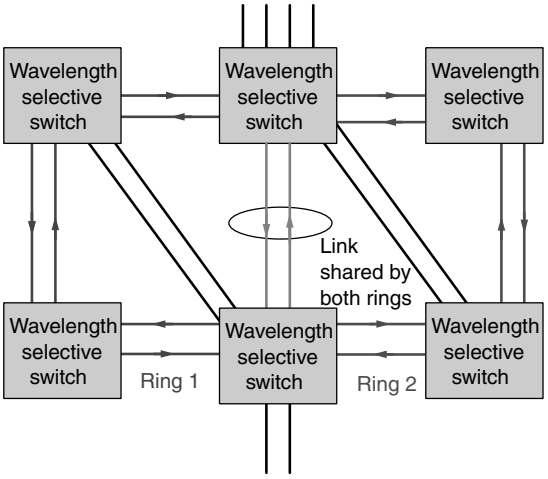


Figure 8. Two rings traversing shared resources over the same link.

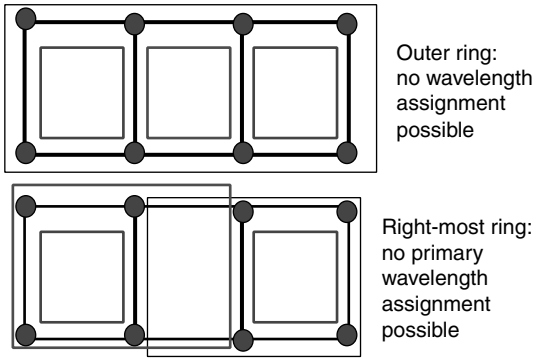


Figure 9. Example showing the problems of applying double cycle covers to wavelength recovery.

problems for bidirectional connections. In contrast, both SHRs and generalized loopback (discussed later in the section) avoid these problems by protecting bidirectional traffic with unique, bidirectional restoration paths.

Cycle covers work well for link failures, but have drawback for recovery from node failures, particularly when failures occur at nodes that are shared by one than one link. While node recovery can be effected with double-cycle ring covers, such restoration requires cumbersome hopping among rings. Moreover, if a link or node is added to a network, the cover of cycles can change significantly, limiting the scalability of double cycle covers. These drawbacks are a general property of ring embeddings, and are already found in SONET networks.

In order to avoid the limitations of ring covers, an approach using preconfigured protection cycles, or p-cycles, is given by Grover and Stamatelakis [92]. A *p-cycle* is a cycle on a redundant mesh network. Links on the p-cycle are recovered by using the p-cycle as a conventional BLSR. Links not on the p-cycle are recovered by selecting, along the p-cycle, a path connecting the nodes at either end of the failed link. Some difficulty arises from the fact that several p-cycles may be required to cover a network, making management among p-cycles necessary. A single p-cycle may be insufficient because a Hamiltonian circuit might not exist, even in a two-connected graph. Even finding p-cycles that cover a large number of nodes may be difficult. Some results [93–95] and conjectures [96,97] exist concerning the length of maximal cycles in two-connected graphs. The p-cycle approach is in effect a hybrid ring approach, which mixes link protection (for links not on the p-cycle) with ring recovery (for links on the p-cycle).

Another approach to link restoration on mesh networks, which we term generalized loopback, was first presented in 1999 [98]. The principle behind generalized loopback is to select a directed graph, called the *primary graph*, such that another directed graph, called the *secondary*, can be used to carry backup traffic for any link failure in the primary. Construction of a primary involves selection of a single direction for each link in the network. Loopback then occurs along the secondary graph in a manner akin to SONET BLSR. Figure 10 demonstrates generalized loopback for a simple network. In the figure, only two fibers per link are shown — one primary fiber and its corresponding secondary fiber. When the link $[Y, X]$ fails, traffic from the primary digraph floods onto the secondary digraph starting at Y. The secondary digraph carries this backup traffic from one endpoint of the failed node to the other endpoint, possibly along multiple paths. When traffic reaches X (along the first successful path), it is again placed on the primary fiber, as though no failure had occurred. Unnecessary backup paths are subsequently torn down. The fact that multiple paths may exist for restoration allows us to reclaim some arcs (fibers) from secondary digraphs to carry additional traffic. The capacity efficiency obtained in this manner is, for typical networks, in the order of 20% over methods, such as double-cycle cover, that require half of the network capacity to be devoted to recovery.

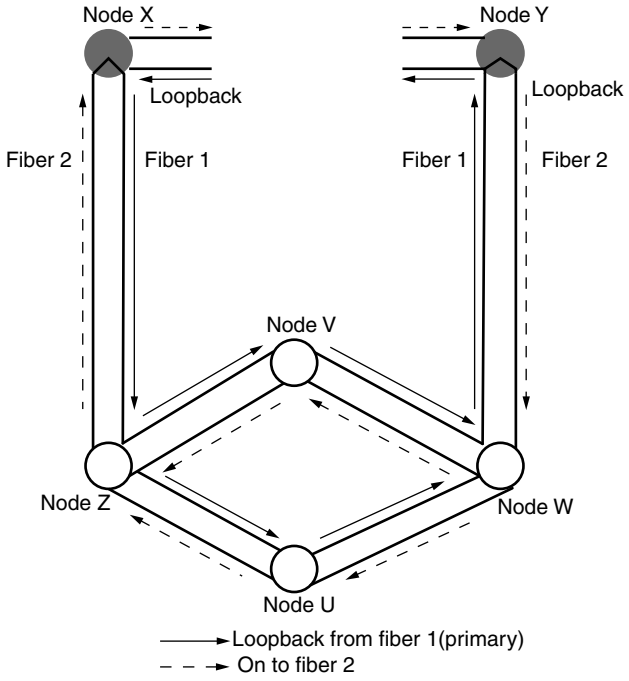


Figure 10. Generalized loopback.

7. PACKET-BASED APPROACHES

This section looks more closely at techniques particular to packet-based networks, giving more details for the failure detection techniques mentioned in Section 2 and the recovery schemes used when failures are detected. Some packet-based networks, such as FDDI, are based on ring topologies and use failure detection and recovery techniques nearly identical to those described in previous sections. For packet networks built with redundant mesh topologies, the approach is substantially different.

One protocol of particular interest and importance was developed in the mid-1980s [99] to allow redundant interconnection of LANs for reliability while avoiding problems of infinite routing loops. The model, known as an extended LAN, uses bridges to connect LANs and to forward traffic between LANs as necessary. The bridges cooperate in a distributed fashion to select a minimum spanning tree (MST) from the full-connectivity graph. All traffic is then routed along the MST, preventing cycles. Periodic configuration messages are sent from the root of the tree and forwarded over all LANs. Failure detection then relies on timeout mechanisms — if a bridge fails to hear the configuration message on any LAN to which it is attached, it acts to find a new route from the spanning tree to that LAN. The simplest method is to restart the search entirely, but some optimizations are possible.

The MST approach serves its purpose, allowing redundant topologies in packet-switched networks without allowing for routing loops. However, restricting traffic to flow along a tree can severely limit the capacity of the extended LAN, and can force traffic between two LANs that are close in the original mesh to follow a long path through the tree, consuming some of the capacity of many other LANs in the process.

Autonet, developed in the early 1990s, solved this problem to some extent by allowing the use of all links in the network [100]. Routing cycles and deadlocks in Autonet were prevented through the use of up*/down* (read "up star, down star," and alluding to regular expressions) routes. With this approach, all routers are assigned a unique number, and packets between two hosts in the network must follow a route that moves in a monotone sequence upward followed by a monotone sequence downward before exiting the network. Any route obeying this constraint can be used. Consider a cycle or self-cycle in routes, and label each link in the cycle as either up or down, depending on the identifiers assigned to the routers at the end of the link. Obviously, a cycle must contain both up and down links, and in particular must contain a two-link section with the first link down and the second up. No route can legally follow both links in this section, however, implying that deadlocks are impossible; thus, all up*/down* routes are mutually deadlock-free. Autonet relied on timeouts built into the switch hardware for detection of failed links and nodes, but was otherwise quite similar to the extended LAN approach in terms of reliability.

Since the early 1990s, many vendors of packet-based networks have recognized the importance of redundancy, and have introduced hardware and software support for combining physical channels into single logical channels between high-end switches. While this approach may seem fairly natural in a packet-based network, in which utilization is already based on statistical multiplexing of the links, some complexities must be addressed. These complexities arise from an assumption by higher-level protocols, in particular the Transmission Control Protocol (TCP), which packets sent through a network arrive in order of transmission. Use of multiple physical routes to carry packets from a single TCP connection often violates this assumption, causing TCP's congestion control mechanisms to drastically cut bandwidth. To address this issue, link aggregation schemes try to restrict individual TCP connections to specific links, and rely on the availability of many connections to provide good load balancing and capacity benefits from aggregation. Failure detection, as with Autonet, is generally handled in hardware, and results in routing reorganization. Unlike many backbone networks, the capacity of most packet-switched networks degrades in the presence of failures, encouraging network architects and managers to operate in somewhat risky modes in which inadequate capacity remains available after certain failures. This phenomenon can be observed even in high-end packet-based systems, including some SAN's backing bank operations.

In the wide area, fault tolerance in packet-switched networks relies on a combination of physical layer notification and rerouting by higher-level routing protocols. The Border Gateway Protocol (BGP) [101], a peer protocol to the Internet Protocol (IP), defines the rules for advertising and selecting routes to networks in the Internet. More specifically, it defines a homogeneous set of rules for interactions between *autonomous systems* (ASs), networks controlled by a single administrative entity. With each AS, administrators are free to select whatever routing protocol suits their fancy, but AS's must interact with each other in a standard way, as defined by BGP. BGP explicitly propagates failure information in the form of withdrawn routes, which cancel previously advertised routes to specific networks.

From the point of view of recovery in the context of optical backbone networks, the overhead required for packet-based systems depends critically on what functionalities are implemented in the optical domain. Restoration, in which, after a failure, excess bandwidth is claimed for the purpose of providing alternate routes to traffic around a failure, is challenging in the optical domain, since it requires operating on the whole datastream and possibly separating packets from a stream. In order to avoid packet-level operations, flow switching on a stream-by-stream basis, for instance using multiprotocol label switching (MPLS), is a promising alternative. Such stream-based operations are more amenable to optical processing. For recovery, stream-based processing reduces roughly to circuit-based recovery.

Packet-switched approaches for optical access seek to perform some subset of the functionalities of traditional opto-electronic packet-based networks optically [102]. These functionalities may be header recognition, buffering, packet insertion, packet reading, packet retrieval, and rate conversion. Performing such operations in the optical domain is challenging and no consensus has emerged regarding implementation. However, certain general statements can be made. Operations, such as buffering a stream, that involve significant timing issues or that introduce loss and distortion on the datastream, tend to be challenging. Replicating a stream, for instance, can be done using passive optical splitters and is therefore relatively straightforward. Merging streams, on the other hand, is challenging because of timing issues. The most challenging operations are the ones performed at the packet level. Again, different levels of difficulty arise. Reading signals from an optical datastream is possible by removing a fraction of the signal power and operating on that fraction. Retrieving a packet (reading the packet and removing it from the stream) is difficult because it involves performing an operation on the whole stream, as well as timing, phase, and polarization issues. Thus, operations such as packet-switching are also challenging because of issues of timing and speed of optical switches. Thus, fully optical packet-switched systems replicating the entire operations of electronic systems are still distant.

8. HIGH-SPEED LANS

The vast majority of the proposed architectures for LANs consist of star topologies or of networks built from combinations of star topologies, in which some type of switch, router, or other type of hub, is placed in the center of a topology and each node is directly connected to the hub [103]. The emergent 10 Gb/s standard (IEEE 802.3ae) for LANs and MANs also allows for optical stars and trees. From the point of reliability, stars present many weaknesses. In particular, a failure at the hub may entail failure of the whole network. However, other failures may occur even without outright failure of the hub. If the hub passively broadcasts, total failure of the hub is unlikely.

However, many partial failure scenarios exist: amplifier failures; port connection failures, at the access nodes or at the hub; transmitter or receiver failures at access nodes, for instance, because of laser failures; or cabling failures in the fiber itself. Such failures entail the failure of one or more arms of the star.

The center of a star topology is inherently a single point of failure, making complete replication of the system necessary to support recovery. Operation of a fully redundant system is difficult, however, as illustrated by existing reliable networks based on star topologies. Many enterprise networks and storage area networks (SANs), for example, are built as stars.

Such systems typically use single-wavelength optical connections rather than WDM, and rely on electronic switching. Enterprise networks are usually based on Gigabit Ethernet (GigE), while SAN's are based on Fibre Channel (FC). Network interface cards (NIC's), housing both a receiver and a transmitter, are optically connected to an electronic switch. The switch is closer to a traditional router than to the passive broadcast hubs or wavelength-selective switches discussed in the context of star-based WDM LANs. For such networks, redundancy is obtained by full duplication of all resources, as shown in Fig. 11.

In addition to replication, the two switches must be connected. Consider the case of failure of the primary NIC in server 1. Server 1 communicates via the secondary switch. Requiring other servers also to communicate via the secondary switch is undesirable. Indeed, although we show just two servers, such networks typically have many servers connected to them and reconfiguring so many connections simultaneously is difficult. Moreover, there is some delay involved in creating new connections through switch 2 owing to initialization overheads. To avoid reconfiguration at all servers, all servers other than server 1 continue to communicate with the primary switch and the two switches communicate with each other via the interswitch connection.

In the context of optical networks, an interswitch connection translates into connection between two hubs. In order to manage such an interhub connection, the hub needs to be equipped with far greater capabilities than simple optical broadcasting. Thus, it would appear that optical star dedicated networks will be difficult to deploy and that the means of providing robustness available in traditional star topologies cannot be easily extended to optical access networks.

The star topology for LAN's connected to a backbone is not limited to optical applications. Such an architecture has been proposed, for instance, for the Integrated Services LAN (ISLAN) defined by IEEE 802.9 using unshielded twisted pair.

While stars and topologies built from stars dominate in the LANs, LANs are also built using bus schemes. Bus schemes allow nodes to place and retrieve traffic using a shared medium. Figure 12 shows a folded bus and a dual bus. In a folded bus, a single bus, originating at a head end, serves all nodes. Typically, nodes use the bus first as a collection bus, onto which they place traffic (in the left-to-right direction in Fig. 12a). The last node folds back the bus to make it travel in the right-to-left direction. In the right-to-left direction, nodes collect traffic placed onto the bus. The traffic may be read only or read and removed. In the dual-bus architecture, two buses are used, each with its own headend. Folded and dual buses are simple options for LANs and certain types of MANs. In particular, they offer an effective way of sharing bandwidth among several users and are therefore attractive to allow nodes to access optical bandwidth, whether for a full fiber, a few wavelengths, or a single wavelength.

Folded and dual buses suffer from reliability drawbacks. Figure 13 shows a folded bus and a dual bus after a failure. Partial recovery can be effected by creating a bus on either side of the failure. For a dual-bus architecture, the node immediately upstream of the failure needs to be

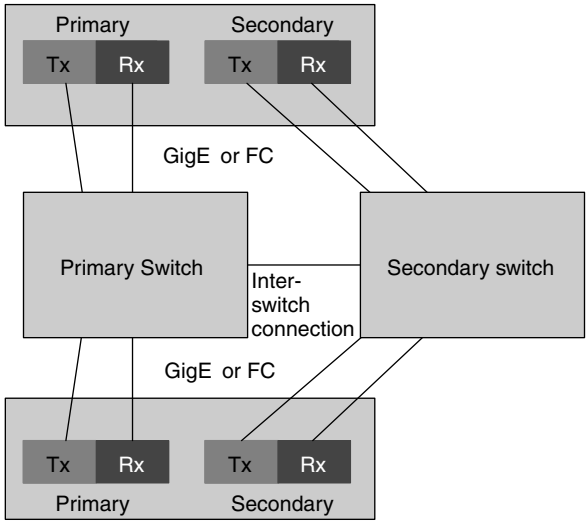


Figure 11. Redundant architecture in an enterprise network or LAN using traditional star topologies.

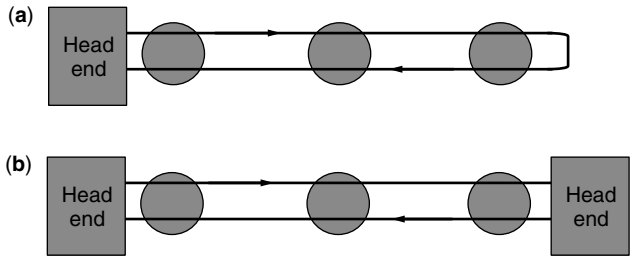


Figure 12. Folded (**a**) and dual (**b**) buses.

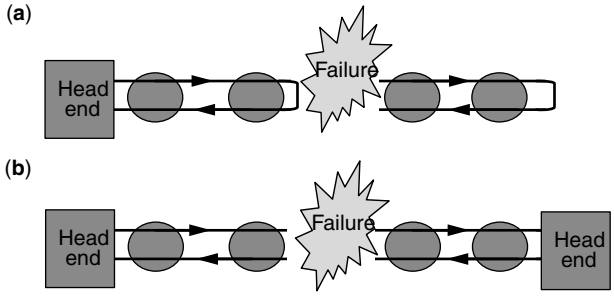


Figure 13. Folded (**a**) and dual (**b**) buses being restored as two folded buses and two dual buses, respectively, after a failure.

able to fold the bus. In order to reestablish full connectivity after a failure, the end nodes of the original buses must be able to connect outside the original buses to transmit traffic that was destined to traverse the cut.

BIOGRAPHY

Muriel Medard is an Assistant Professor in the Electrical Engineering and Computer Science (EECS) at MIT and a member of the Laboratory for Information and Decision Systems. She was previously an Assistant Professor at the Electrical and Computer Engineering Department and a member of the Coordinated Science Laboratory at the University of Illinois Urbana — Champaign. From 1995 to 1998 she was a Staff Member at MIT Lincoln Laboratory in the Optical Communications and the Advanced Networking Groups. Professor Medard received B.S. degrees in EECS and Mathematics in 1989, a B.S. degree in Humanities in 1990, a M.S. degree in Electrical Engineering 1991, and a Sc.D. degree in Electrical Engineering in 1995, all from the Massachusetts Institute of Technology (MIT), Cambridge. Medard's research interests are in the areas of reliable communications, particularly for optical and wireless networks. She received a 2001 NSF Career award. She was awarded the IEEE Leon K. Kirchmayer Prize Paper Award 2002 for her paper, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel."

BIBLIOGRAPHY

1. T. Anderson et al., *Dependability: Basic Concepts and Terminology*, Springer-Verlag, Wien (Vienna), Austria, 1992.
2. C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, Delayed Internet routing convergence, *Proc. ACM SIGCOMM Conf.*, 2000, pp. 175–187.
3. C. Labovitz, A. Ahuja, R. Watterhofer, and S. Venkatachary, The impact of Internet policy and topology on delayed routing convergence, *Proc. INFOCOM*, 2001.
4. K. Murakami and H. S. Kim, Virtual path routing for survivable ATM networks, *IEEE/ACM Trans. Network.* **4**: (1996).
5. D. J. Pai and H. L. Owen, An algorithm for bandwidth management with survivability constraints in ATM networks, *Proc. ICC*, 1997, Vol. 1, pp. 261–266.
6. O. Gerstel and R. Ramaswami, Optical layer survivability — an implementation perspective, *IEEE J. Select. Areas Commun.* 1885–1899 (2000).
7. D. Zhou and S. Subramanian, Survivability in optical networks, *IEEE Network* 16–23 (Nov./Dec. 2000).
8. C. J. Green, Protocols for a self-healing network, *Proc. Military Communications Conf.* (*MILCOM*), 1995, Vol. 1, pp. 252–256.
9. T. E. Stern and K. Bala, *Multiwavelength Optical Networks: A Layered Approach*, Prentice-Hall, Upper Saddle River, NJ, 2000.
10. A. Banerjea, C. J. Parris, and D. Ferrari, Recovering guaranteed performance service connections from single and multiple faults, *Proc. GLOBECOM*, 1994, Vol. 1, pp. 162–166.
11. R. Doverspike, A multi-layered model for survivability in intra-LATA transport networks, *Proc. GLOBECOM*, 1991, pp. 2025–2031.
12. T.-H. Wu, *Fiber Network Service Survivability*, Artech House, 1992.
13. D. Johnson et al., Distributed restoration strategies in telecommunications networks, *Proc. IEEE Int. Conf. Communications*, 1994, Vol. 1, pp. 483–488.
14. R. Nakamura, H. Ono, and K. Nishikawara, Reliable switching services, *Proc. GLOBECOM*, 1994, Vol. 3, pp. 1596–1600.
15. J. Veerasamy, S. Vemkatesan, and J. C. Shah, Effect of traffic splitting on link and path restoration planning, *Proc. GLOBECOM*, 1994, Vol. 3, pp. 1867–1871.
16. M. Tomizawa, Y. Yamabayashi, N. Kawase, and Y. Kobayashi, Self-healing algorithm for logical mesh connection on ring networks, **30**: 1615–1616 (Sept. 15 1994).
17. J. Sosnosky, Service application for sonet dcs distributed restoration, **12**: 59–68 (1994).
18. D. Edinger, P. Duthie, and G. R. Prabhakara. A new answer to fiber protection. 53–55, April 9, 1990.
19. T.-H. Wu and W. I. Way, A novel passive protected sonet bidirectional self-healing ring architecture, *jtl* **10**: (Sept. 1992).
20. W. D. Grover, The selfhealing^{TM} network, *Proc. GLOBECOM*, 1987, pp. 1090–1095.
21. C. H. Yang and S. Hasegawa, Fitness: Failure immunization technology for network service survivability, *Proc. GLOBECOM*, 1988, Vol. 3, pp. 47.3.1–47.3.6.
22. H. Fujii and N. Yoshikai, Double search self-healing algorithm and its characteristics, **77**: 975–995 (1994).
23. H. Sakauchi, Y. Okanoue, H. Okazaki, and S. Hasegawa, Distributed self-healing control in SONET, *J. Network Syst. Manage.* **1**(2): 123–141 (1993).
24. T.-H. Wu, A passive protected self-healing mesh network architecture and applications, *IEEE/ACM Trans. Network.* 40–52 (Feb. 1994).
25. H. Kobrinski and M. Azuma, Distributed control algorithms for dynamic restoration in dcs mesh networks: Performance evaluation, *Proc. GLOBECOM*, 1993, Vol. 3, pp. 1584–1588.
26. A. Gersht, S. Kheradpir, and A. Shulman, Dynamic bandwidth-allocation and path-restoration in SONET self-healing networks, *IEEE Trans. Reliability* **45**: 321–331 (June 1996).
27. M. Barezzani, E. Pedrinelli, and M. Gerla, Protection planning in transmission networks, *Proc. ICC*, 1992, pp. 316.4.1–316.4.5.
28. C. E. Chow, J. Bicknell, S. McCaughey, and S. Syed, A fast distributed network restoration algorithm, *Proc. 12th Int. Phoenix Conf. Computers and Communications*, March 1993, Vol. 1, pp. 261–267.
29. J. Bicknell, C. E. Chow, and S. Syed, Performance analysis of fast distributed network restoration algorithms, *Proc. IEEE GLOBECOM*, 1993, Vol. 3, pp. 1596–1600.
30. R. Doverspike and B. Wilson, Comparison of capacity efficiency of dcs network restoration routing techniques, volume 2, 1994.
31. T. Frisanco, Optimal spare capacity design for various protection switching methods in atm networks, *Proc. ICC*, 1997, Vol. 1, pp. 293–298.

32. N. Wauters, B. Van Caenegem, and P. Demeester, Spare capacity assignment for different restoration strategies in mesh survivable networks, *Proc. ICC*, 1997, Vol. 1, pp. 288–292.
33. K. Menger, *Zur allgemeinen Kurventheorie*, Fundamenta Mathematicae, 1927.
34. M. Stoer, *Design of Survivable Networks*, Springer-Verlag, 1992.
35. R. Bhandari, Optimal diverse routing in telecommunication fiber networks, *Proc. IEEE INFOCOM*, May 1994, Vol. 3, pp. 11.c.3.1–11.c.3.11.
36. S. Z. Shaikh, Span-disjoint paths for physical diversity in networks, *Proc. IEEE Symp. Computers and Communications*, 1995, pp. 127–133.
37. J. W. Suurballe, Disjoint paths in a network, pages 125–145, 1974.
38. W. T. Zaumen and J. J. Garcia-Luna Aceves, Dynamics of distributed shortest-path routing algorithms, *Proc. 21st SIGCOMM Conf.*, Sept. 3–6, 1991, ACM Press, 1991, Vol. 21, pp. 31–43.
39. J. S. Whalen and J. Kenney, Finding maximal link disjoint paths in a multigraph, *Proc. GLOBECOM*, 1990, pp. 403.6.1–403.6.5.
40. P. Mateti and N. Deo, On algorithms for enumerating all circuits of a graph, **5**: (March 1976).
41. A. Itai and M. Rodeh, The multi-tree approach to reliability in distributed networks, Number 79, 1988.
42. M. Médard, S. G. Finn, R. G. Gallager, and R. A. Barry, Redundant trees for automatic protection switching in arbitrary node-redundant or edge-redundant graphs, *Proc. ICC*, 1998.
43. R. Kawamura, H. Hadama, and I. Tokizawa, Implementation of self-healing function in ATM networks, *J. Network Syst. Manage.* **3**(3): 243–264 (1995).
44. N. D. Lin, A. Zolfaghari, and B. Lusignan, ATM virtual path self-healing based on a new path restoration protocol, *Proc. GLOBECOM*, 1994, Vol. 2, pp. 794–798.
45. D. K. Hsing, B.-C. Cheng, G. Goncu, and L. Kant, A restoration methodology based on preplanned source routing in ATM networks, *Proc. ICC*, 1997, Vol. 1, pp. 277–282.
46. R. S. K. Chng et al., A multi-layer restoration strategy for reconfigurable networks, *Proc. GLOBECOM*, 1994, Vol. 3, pp. 1872–1878.
47. S. Hasegawa, A. Kanemasa, H. Sakaguchi, and R. Maruta, Dynamic reconfiguration of digital cross-connect systems with network control and management, *Proc. GLOBECOM*, 1994, pp. 28.3.2–28.3.5.
48. A. Gersht and S. Kheradpir, Real-time bandwidth allocation and path restorations in SONET-based self-healing mesh networks, *Proc. IEEE Int. Conf. Communications*, 1993, Vol. 1, pp. 250–255.
49. J. E. Baker, A distributed link restoration algorithm with robust preplanning, *Proc. IEEE GLOBECOM*, 1991, Vol. 1, pp. 10.4.1–10.4.6.
50. S. Ramamurthy and B. Mukherjee, Survivable WDM mesh networks, part I — protection, *Proc. IEEE INFOCOM*, 1999, pp. 744–751.
51. J. Anderson, B. T. Doshi, S. Dravida, and P. Harshavardhana, Fast restoration of ATM networks, *IEEE J. Select. Areas Commun.* **12**(1): 128–136 (Jan. 1994).
52. Y. Xiong and L. G. Mason, Restoration strategies and spare capacity requirements in self-healing ATM networks, *IEEE J. Lightwave Commun.* **7**(1): 98–110 (Feb. 1999).
53. M. Azuma et al., Network restoration algorithm for multimedia communication services and its performance characteristics, *IEICE Trans. Commun.* **E78-B**(7): 987–994 (July 1995).
54. R. Kawamura, K. Sato, and I. Tokizawa, High-speed self-healing techniques utilizing virtual paths, *Proc. 5th Int. Network Planning Symp.*, May 1992.
55. M. Boyden T.-H. Wu, and R. H. Caldwell, A multi-period design model for survivable network architecture selection for sdh/sonet interoffice networks, volume **40**: 417–432 (Oct. 1991).
56. O. J. Wasem, An algorithm for designing rings for survivable fiber networks, volume **40**: (1991).
57. T.-H. Wu and M. E. Burrows, Feasibility study of a high-speed sonet self-healing ring architecture in future interoffice networks, pages 33–51, Nov. 1990.
58. C.-C. Shyur, Y.-M. Wu, and C.-H. Chen, A capacity comparison for sonet self-healing ring networks, *Proc. GLOBECOM*, 1993, pp. 1574–1578.
59. J. B. Slevinsky, W. D. Grover, and M. H. MacGregor, An algorithm for survivable network design employing multiple self-healing rings, *Proc. GLOBECOM*, 1993, Vol. 3, pp. 1568–1573.
60. J. Shi and J. Fonseka, Interconnection of self-healing rings, *Proc. ICC*, 1996, Vol. 1.
61. C.-C. Shyur, S.-H. Tsao, and Y.-M. Wu. Survivable network planning methods and tools in Taiwan, Sept. 1995.
62. L. M. Gardner et al., Techniques for finding ring covers in survivable networks, *Proc. GLOBECOM*, 1994, Vol. 3.
63. T.-H. Wu, D. J. Kolar, and R. H. Cardwell, High-speed self-healing ring architectures for future interoffice networks, *Proc. GLOBECOM*, 1989, Vol. 2, pp. 23.1.–23.1.7.
64. E. L. Hahne and T. D. Todd, Fault-tolerant multimesh networks, *Proc. GLOBECOM*, 1992, pp. 627–632.
65. R. B. Magill, A bandwidth efficient self-healing ring for b-isdn, *Proc. ICC*, 1997.
66. Y. Kajiyama, N. Tokura, and K. Kikuchi, An atm vp-based self-healing ring, **12**: (Jan. 1994).
67. F. E. Ross, An overview of FDDI: The fiber distributed data interface, *IEEE J. Select. Areas Commun.* **7**: 1043–1051 (Sept. 1989).
68. F. E. Ross, Fiber distributed data interface: An overview, *Proc. 15th Conf. Local Computer Networks*, 1990, pp. 6–11.
69. R. O. LaMaire, FDDI performance at 1 Gbit/s, *Proc. IEEE Int. Conf. Communications*, 1991, pp. 174–183.
70. T. S. Jones and A. Louri, Media access protocols for a scalable optical interconnection network, May 1998.
71. M. A. Marsan et al., An almost optimal MACprotocol for all-optical WDM multi-rings with tunable transmitters and fixed receivers, *Proc. IEEE Int. Conf. Communications*, May 1997, Vol. 1, pp. 437–442.
72. M. A. Marsan et al., All-optical WDMmulti-rings with differentiated qos., *IEEE Commun. Mag.* **37**: 58–66 (Feb. 1999).
73. M. A. Marsan et al., SR/sup 3/:a bandwidth-reservation MACprotocol for multimedia applications over all-optical WDM multi-rings, *Proc. INFOCOM '97*, 1997, Vol. 2.

74. A. Bianco et al., A-posteriori access strategies in all-optical slotted WDM rings, *Proc. Global Telecommunications Conf.*

75. A. Louri and R. Gupta, Hierarchical optical interconnection network HORN: Scalable interconnection network for multiprocessors and multicomputers, Jan. 1997.

76. Resilient Packet Ring Working Group.

77. A. Banerjea, C. J. Parris, and D. Ferrari, Recovering guaranteed performance service connections from single and multiple faults, *Proc. GLOBECOM*, 1994, Vol. 1, pp. 162–166.

78. T. H. Wu, D. J. Kolar, and R. H. Cardwell, Survivable network architectures for broad-band fiber optic networks: Model and performance comparison, *IEEE J. Lightwave Commun.* **6**(11): (Nov. 1988).

79. K. T. Newport and P. K. Varshney, Design of survivable communication networks under performance constraints, *IEEE Trans. Reliability* **40**: 433–440 (Oct. 1991).

80. T.-H. Wu and S. Fouad Habiby, Strategies and technologies for planning a cost-effective survivable network architecture using optical switches, *IEEE Trans. Reliability* **8**(2): 152–159 (Feb. 1991).

81. R.-H. Jan, F.-J. Hwang, and S. T. Cheng, Topologocial optimization of a communication network subject to a reliability constraint, *IEEE Trans. Reliability* **42**(1): (March 1993).

82. W. D. Grover, Case studies of survivable ring, mesh and mesh-arc hybrid networks, *Proc. GLOBECOM*, 1992, pp. 633–638.

83. C. Thomassen, On the complexity of finding a minimum cycle cover of a graph, *SIAM J. Comput.* **26**: 675–677 (June 1997).

84. A. Itai, R. J. Lipton, C. H. Papadimitriou, and M. Rodeh, Covering graphs with simple circuits, *SIAM J. Comput.* **10**: 746–750 (1981).

85. G. Fan, Covering graphs by cycles, *SIAM J. Comput.* **5**: 491–496 (Nov. 1992).

86. G. Ellinas, T. E. Stern, and A. Hailemarian, Link failure restoration in optical networks with arbitrary mesh topologies and bi-directional links, 1997.

87. G. Ellinas and T. E. Stern, Automatic protection switching for link failures in optical networks with bi-directional links, *Proc. GLOBECOM*, 1996.

88. F. Jaeger, A survey of the double cycle cover conjecture, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985.

89. P. D. Seymour, Sums of circuits, in U. S. R. Murty and J. A. Bondy, eds., *Graph Theory and Related Topics*, Academic Press, New York, 1979, pp. 341–355.

90. G. Szekeres, Polyhedral decomposition of cubic graphs, *J. Austral. Math. Soc.* **8**: 367–387 (1973).

91. L. Goddyn, A girth requirement for the double cycle cover conjecture, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 13–26.

92. W. D. Grover and D. Stamatelakis, Cycle-oriented distributed preconfiguration: Ring-like speed with mesh-like capacity for self-planning network reconfiguration, *Proc. IEEE Int. Conf. Communications*, 1998, Vol. 2, pp. 537–543.

93. I. Fournier, Longest cycles in 2-connected graphs of independence number α, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 201–204.

94. B. Jackson, Hamilton cycles in regular 2-connected graphs, *J. Comb. Theory Ser. B* **29**: 27–46 (1980).

95. Y. Zhu, Z. Liu, and Z. Yu, An improvement of Jackson's result on Hamilton cycles in 2-connected graphs, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 237–247.

96. R. H$\acute{}$aggkvist and B. Jackson, A note on maximal cycles in 2-connected graphs, in *Cycles in Graphs*, Annals of Discrete Mathematics, Vol. 115, North-Holland, 1985, pp. 205–208.

97. D. R. Woodall, Maximal circuits of graphs II, *Studia Sci. Math. Hungar.* **10**: 103–109 (1975).

98. M. Médard, S. G. Finn, and R. A. Barry, Wdm loopback recovery in mesh networks, *Proc. IEEE INFOCOM*, 1999.

99. R. Perlman, An algorithm for distributed computation of a spanning tree in an extended LAN, *Proc. 9th Symp. Data Communications,* (*SIGCOMM'85*), Whistler Mountain, British Columbia, Canada, 1985, pp. 44–53.

100. M. D. Schroeder et al., Autonet: A high-speed, self-configuring local area network using point-to-point links, *IEEE J. Select. Areas Commun.* **9**: 1318–1335 (Oct. 1991).

101. Y. Rekhter and T. Li, *A Border Gateway Protocol 4* (*BGP-4*), Internet Engineering Task Force RFC 1771, March 1995.

102. E. Modiano, Wdm-based packet networks, *IEEE Commun. Mag.* **37**: 130–135 (March 1999).

103. P. A. Humblet, R. Ramaswami, and K. N. Sivarajan, An efficient communication protocol for high-speed packet-switched multichannel networks, *IEEE J. Select. Areas Commun.* **11**: 568–578 (May 1993).

NETWORK SECURITY

ROLF OPPLIGER
eSECURITY Technologies
Rolf Oppliger
Bern, Switzerland

1. INTRODUCTION

According to Shirey [1], the term *computer network* (or *network*) refers to, "a collection of host computers together with the subnetwork or internetwork through which they can exchange data." Many different technologies and communication protocols can be used (and are in use) to build and operate computer networks. Examples include the IEEE 802 family of protocols for local area networking, the Point-to-Point Protocol (PPP) for dialup networking, and the TCP/IP protocol suite for internetworking.[1]

Almost all contemporary networking technologies and communication protocols are highly complex and have not been designed with security in mind. Consequently, they

[1] The acronym TCP/IP refers to an entire suite of communications protocols that center around the Transmission Control Protocol (TCP) and the Internet Protocol (IP). The emerging use of TCP/IP networking has led to a global system of interconnected hosts and networks that is commonly referred to as the Internet.

are inherently vulnerable and exposed to a variety of threats and corresponding attacks. To make things worse, there are a number of reasons why networked computer systems are inherently more vulnerable and exposed to threats and corresponding attacks than their standalone counterparts:

- More points exist from where an attack can be launched. Note that a computer system that is inaccessible or unconnectable to users cannot be attacked. Consequently, by adding more network connections (i.e., network connectivity) for legitimate users, more possibilities to attack the system are automatically added, as well.
- The physical perimeter of a networked computer system is artificially extended by having it connect to a computer network. This extension typically leads beyond what is actually controllable by a system administrator.
- Networked computer systems typically run software that is inherently more complex and error-prone. There are many network software packages that have a long bug record and that are known to be "buggy" accordingly (e.g., the UNIX sendmail daemon). More often than not, intruders learn about these bugs before system administrators do. To make things worse, intruders must know and be able to exploit one single bug, whereas system administrators must know and be able to fix all of them. Consequently, the workload between system administrators and potential intruders is asymmetrically distributed.

In essence, the aim of network security is to provide the technologies, mechanisms, and services that are required and can be used to protect the computational and networking resources against accidental and/or intentional threats. Mainly because of the importance of computer networks in daily life, network security is a hot topic today.

This article provides an overview and discussion about the current state -of- the -art and future perspectives in network security in general, and Internet security in particular. As such, it is organized as follows. Possible threats and attacks against computer networks and distributed systems are overviewed and briefly discussed in Section 2. The OSI security architecture is introduced in Section 3. The architecture provides a useful terminology that is applicable to a wide range of networking technologies and corresponding communication protocols. As a case study, Internet security is further addressed in Section 4. Finally, conclusions are drawn and an outlook is given in Section 5. Some parts of this article are taken from Ref. 2. Readers may refer to this reference book to get some further information about network security in general, and Internet security in particular.

2. THREATS AND ATTACKS

A threat refers to "a potential for violation of security, which exists when there is a circumstance, capability, action, or event that could breach security and cause harm [1]. That is, a threat is a possible danger that might exploit a vulnerability." A threat can be either accidental (e.g., caused by a natural disaster) or intentional (e.g., an attack). In the sequel, we focus only on intentional threats and corresponding attacks that may be launched either by legitimate users (i.e., insiders) or — more importantly — outside attackers (i.e., outsiders). All statistical investigations reveal (or confirm) the fact that most attacks are launched by insiders rather than outsiders. This is because insiders generally have more knowledge and possibilities to attack computer systems that store, process or transmit valuable information assets.

Again referring to Shirey [1], an attack is "an assault on system security that derives from an intelligent threat, i.e., an intelligent act that is a deliberate attempt (especially in the sense of a method or technique) to evade security services and violate the security policy of a system." There are many attacks that can be launched against computer networks and the systems they interconnect. Most attacks are due to vulnerabilities in the underlying network operating systems. In fact, the complexity of contemporary network operating systems makes it possible and very likely that we will see an increasingly large number of network-based and software-driven attacks in the future. What we experience today with macro viruses and network worms is only the tip of an iceberg.

With regard to telecommunications, it is common to distinguish between passive and active attacks:

- A *passive attack* attempts to learn or make use of information but does not affect system or network resources.
- An *active attack* attempts to alter system or network resources and affect their operation.

Passive and active attacks are typically combined to more effectively invade a computing or networking environment. For example, a passive wiretapping attack can be used to eavesdrop on authentication information that is transmitted in the clear (e.g., a username and password), whereas this information can later be used to masquerade another user and to actively attack the corresponding computer system. Passive and active attacks are further explored next.

2.1. Passive Attacks

As mentioned above, a passive attacker attempts to learn or make use of information but does not affect system or network resources. As such, a passive attack primarily threatens the confidentiality of data being transmitted. This data may include anything, including, for example, confidential electronic mail messages or usernames and passwords transmitted in the clear. In fact, the cleartext transmission of authentication information is the single most important vulnerability in computer networks today.

In regard to the intruder's opportunities to interpret and extract the information that is encoded in the transmitted data, it is common to distinguish between passive wiretapping and traffic analysis attacks:

- In a *passive wiretapping* attack, the intruder is able to interpret and extract the information that is encoded in the transmitted data. For example, if two parties communicate unencrypted, a passive wiretapper is trivially able to extract all information that is encoded in the data.
- In a *traffic analysis* attack, the intruder is not able to interpret and extract the information that the transmitted data encodes (because, e.g., the information is encrypted for transmission). Instead, *traffic analysis* refers to the inference of information from the observation of external traffic characteristics. For example, if an attacker observes that two companies — one financially strong, the other financially weak — begin to trade a large number of encrypted messages, he/she may infer that they are discussing a merger. Many other examples occur in military environments.

The feasibility of a passive attack primarily depends on the physical transmission media in use and their physical accessibility to potential intruders. For example, mobile communications is inherently easy to tap, whereas metallic transmission media at least require some sort of physical access. Lightwave conductors also can be tapped, but this is technically more challenging and expensive. Also note that the use of concentrating and multiplexing techniques, in general, makes it more difficult to passively attack data in transmission. Because of these difficulties, it is more likely that computer networks are passively attacked at the edge (e.g., local-area network segments that are connected to a wide-area network) than in its core or backbone.

It is, however, also important to note that a passive attacker does not necessarily have to tap a physical communications line. Most network interfaces can operate in a so-called promiscuous mode. In this mode, they are able to capture all frames transmitted on the local area network segment they are connected to, rather than just the frames addressed to the computer systems of which they are part. This capability has many useful purposes for network analysis, testing, and debugging (e.g., by utilities such as etherfind and tcpdump in the case of the UNIX operating system). Unfortunately, the capability also can be used by attackers to snoop on all traffic on a particular network segment. Several software packages are available for monitoring network traffic, primarily for the purpose of network management. These software packages are dual-use, meaning they can, for example, be effective in eavesdropping and capturing email messages or usernames and passwords as they are transmitted over shared media and communication lines.

A number of technologies can be used to protect a network environment against passive wiretapping attacks. For example, switched networks are more difficult to wiretap (because data are not broadcast and sent to all potential recipients). Consequently, the use of switched networks in the local area has had a very positive effect on the possibility to passively attack computer networks. Also, a few tools attempt to detect network interfaces that operate in promiscuous mode. For example, a tool named AntiSniff[2] implements a number of tricks to do so. One trick is to send an Ethernet frame with an invalid MAC address to a system and to encapsulate an Internet Control Message Protocol (ICMP) request packet with a valid IP header in the Ethernet frame (ICMP is the control protocol that complements IP). If the targeted system has a network interface that operates in promiscuous mode, it will grab the frame from the Ethernet segment decapsulate it and properly forward it to the local IP module. The IP module, in turn, will decapsulate and receive the ICMP echo request and eventually return a corresponding ICMP response. As a consequence, the targeted system reacts on something it should not have reacted (i.e., because of the invalid MAC address it should not have received the ICMP request in the first place). Obviously, it is simple to hide a network interface that operates in a promiscuous mode simply by not responding to ICMP requests that are encapsulated in Ethernet frames with invalid MAC addresses. What we are going to see in the future is that tools that can be used to passively wiretap network segments and tools that try to detect these tools play "hide and seek" on network segments. Last but not least, the use of data encryption is both effective and efficient against passive wiretapping attacks. In fact, it is the preferred technology and the technology of choice for network practitioners.

Contrary to passive wiretapping attacks, protection against traffic analyses is much more complicated and requires more sophisticated security technologies. Note that the use of encryption techniques does not protect against traffic analysis attacks. In fact, there are only a few technologies readily available to protect against traffic analysis attacks. Exemplary technologies include traffic padding (as discussed later) and a few privacy-enhancing technologies (PETs), such as onion routing (not addressed in this article). There is a lot of room for further research and development in this area.

2.2. Active Attacks

As mentioned above, an active attacker attempts to alter system or network resources and affect their operation. Consequently, an active attack primarily threatens the integrity or availability of data being transmitted. What this basically means is that the attacker can modify, extend, delete, or replay data units that are transmitted in computer networks and distributed systems.

The underlying reason why most active attacks are possible and fairly easy to launch in computer networks and distributed systems is that the data units that are sent and received are seldom protected in terms of authenticity and integrity. Examples of such data units include Ethernet frames, IP packets, User Datagram Protocol (UDP) datagrams, and Transmission Control Protocol (TCP) segments. Consequently, it is simple to do such things as flooding a recipient and cause a "denial of service" or "degradation of service," spoofing the source of data units or the identity of somebody else, or taking over and "hijacking" established network connections. Active attacks are very powerful and it is possible and very likely that we

[2] *http://www.securitysoftwaretech.com/antisniff/*.

will see many other active attacks being discovered and published in the future.

A number of technologies can be used to protect against some active attacks. Most of these technologies use cryptographic techniques to protect the authenticity and integrity of data units that are transmitted. There are, however, also a number of active attacks that are hard to protect against. Examples include denial-of-service and degradation-of-service attacks, as well as their distributed counterparts (i.e., distributed denial-of-service and degradation-of-service attacks). Similar to the real world, protection against this kind of attacks is very difficult to achieve in the digital world of computer networks. How would you, for example, protect your mailbox in the real world against somebody who fills it up with empty paper sheets? There seems to be no simple answer to this question, and the problem is difficult to address in either the real or digital world. In the digital world the problem is even more worrisome, simply because the corresponding attacks are much simpler (and less expensive) to launch. There are, for example, many tools that automatically fill up the mailboxes of particular victims.

3. OSI SECURITY ARCHITECTURE

According to Shirey [1], a *security architecture* refers to "a plan and set of principles that describe (a) the security services that a system is required to provide to meet the needs of its users, (b) the system elements required to implement the services, and (c) the performance levels required in the elements to deal with the threat environment." As such, a security architecture is the result of applying good principles of systems engineering and addresses issues related to physical security, computer security, communication security, organizational security (e.g., administrative and personnel security), and legal security. This is complicated and difficult to achieve, but it is very important. More often than not, systems and applications are designed, implemented and deployed without having an appropriate security architecture in mind.[3]

To extend the field of application of the reference model for open systems interconnection (OSI), the ISO/IEC JTC1 appended a security architecture as part two of ISO/IEC 7498 in the late 1980s [3]. The OSI security architecture is still valid and in use today. It provides a general description of security services and related security mechanisms, which may be provided by a computer network, and defines the positions within the OSI reference model where the services and mechanisms may be provided. Since its publication, the OSI security architecture has turned out to be a primary reference for network security professionals. In 1991, the ITU-T adopted the OSI security architecture in its recommendation X.800 [4] and the Privacy and Security Research Group (PSRG) of the Internet Research Task Force (IRTF) adopted the OSI security architecture in a corresponding Internet security architecture[4] in the early 1990s. In essence, ISO/IEC 7498-2, ITU-T X.800, and the Internet security architecture describe the same security architecture, and in this article we use the term OSI security architecture to collectively refer to all of them.

In short, the OSI security architecture provides a general description of security services and related security mechanisms and discusses their interrelationships. It also shows how the security services map onto a given network architecture and briefly discusses their appropriate placement within the OSI reference model. Having the abovementioned definition of a security architecture in mind, it is obvious that the OSI security architecture does not conform to it. In fact, the OSI security architecture rather refers to a (terminological) framework and a general description of security services and related security mechanisms than to a full-fledged security architecture. Nevertheless, we use it in this article to serve as a starting point for subsequent discussions.

3.1. Security Services

The OSI security architecture distinguishes among five complementary classes of security services. These classes comprise authentication, access control, data confidentiality, data integrity, and nonrepudiation services. Just as layers define functionality in the OSI reference model, so do security services in the OSI security architecture define various security objectives and aspects relevant for computer networks and distributed systems.

- *Authentication services* provide for the authentication of communicating peers or data origins:
  - A *peer entity authentication service* provides the ability to verify that a peer entity in an association is the one it claims to be. In particular, a peer entity authentication service provides assurance that an entity is not attempting to masquerade or perform an unauthorized replay of some previous association. Peer entity authentication is typically performed either during a connection establishment phase or, occasionally, during a data transfer phase.
  - A *data origin authentication service* allows the sources of data received to be verified to be as claimed. A data origin authentication service, however, cannot provide protection against the duplication or modification of data units. In this case, a data integrity service must be used in conjunction with a data origin authentication service. Data origin authentication is typically provided during a data transfer phase.

  Authentication services are important because they are a prerequisite for proper authorization, access control, and accountability. *Authorization* refers to the process of granting rights, which includes the granting of access based on access rights. Access control refers to the process of enforcing access rights, and accountability to the property that ensures that the actions of a principal may be traced uniquely to this particular principal.

[3] Refer to *http://www.esecurity.ch/security_architectures.pdf* for a white paper that describes the role and importance of having an appropriate security architecture.

[4] This work has been abandoned.

- *Access control services* provide for the protection of system or network resources against unauthorized use. As mentioned above, access control services are often closely tied to authentication services; For instance, a user or a process acting on a user's behalf must be properly authenticated before an access control service can effectively mediate access to system resources. In general, access control services are the most commonly used services in both computer and communication security.
- *Data confidentiality* refers to the property that information is not made available or disclosed to unauthorized individuals, entities, or processes. Thus, *data confidentiality services* provide for the protection of data from unauthorized disclosure:
  - A *connection confidentiality service* provides confidentiality of all data transmitted in a connection.
  - A *connectionless confidentiality service* provides confidentiality of single data units.
  - A *selective field confidentiality service* provides confidentiality of only certain fields within the data during a connection or in a single data unit.
  - A *traffic flow confidentiality service* provides protection of information that may otherwise be compromised or indirectly derived from a traffic analysis.

  The provision of a traffic flow confidentiality service requires fundamentally different security mechanisms than the other data confidentiality services.
- *Data integrity* refers to the property that information is not altered or destroyed in some unauthorized way. Thus, *data integrity services* provide protection of data from unauthorized modifications:
  - A *connection integrity service with recovery* provides integrity of data in a connection. The loss of integrity is recovered, if possible.
  - A *connection integrity service without recovery* provides integrity of data in a connection. In this case, however, the loss of integrity is not recovered.
  - A *selected field connection integrity service* provides integrity of specific fields within the data during a connection.
  - A *connectionless integrity service* provides integrity of single data units.
  - A *selected field connectionless integrity service* provides integrity of specific fields within single data units.

  Note that on a connection, the use of a peer entity authentication service at the start of the connection and a connection integrity service during the connection can jointly provide for the corroboration of the source of all data units transferred on the connection, the integrity of those data units, and may additionally provide for the detection of duplication of data units, for example, by using sequence numbers.
- *Nonrepudiation services* prevent one of the entities involved in a communication from later denying having participated in all or part of the communication. Consequently, they have to provide some sort of protection against the originator of a message or action denying that he/she has originated the message or the action, as well as against the recipient of a message denying having received the message. Consequently, there are two non-repudiation services to be distinguished:
  - A *nonrepudiation service with proof of origin* provides the recipient of a message with a proof of origin.
  - A *nonrepudiation service with proof of delivery* provides the sender of a message with a proof of delivery.

  Nonrepudiation services are becoming increasingly important in the context of electronic commerce (e-commerce) on the Internet [5]. For example, a non-repudiation service with proof of delivery may be important for secure messaging (in addition to any secure messaging scheme that employs digital envelopes and digital signatures). The corresponding service is sometimes also referred to as "certified mail." Certified mail is certainly a missing piece for the more professional use electronic mail.

The security services mentioned in the OSI security architecture can be complemented by anonymity or pseudonymity services. These services are not addressed in this article. Sometimes, the availability of anonymity or pseudonymity services directly contradicts the availability of other security services, such as authentication and access control services.

In either case, a security service can be implemented by one or several security mechanisms. The security mechanisms that are addressed in the OSI security architecture are briefly overviewed next.

3.2. Security Mechanisms

The OSI security architecture distinguishes between specific security mechanisms and pervasive security mechanisms.

3.2.1. Specific Security Mechanisms. The OSI security architecture enumerates the following eight specific security mechanisms:

- *Encipherment*, which refers to the application of cryptographic techniques to encrypt data and to transform it in a form that is not intelligible by an outsider (i.e., somebody not knowing a particular cryptographic key). As such, encipherment can be directly used to protect the confidentiality of data units and traffic flow information or indirectly to support or complement other security mechanisms. Many algorithms and standards can be used for encipherment. Examples include secret key cryptosystems, such as the Data Encryption Standard (DES) and the Advanced Encryption Standard (AES), and public key cryptosystems, such as RSA and ElGamal.

- *Digital signature mechanisms*, which can be used to provide an electronic analog of handwritten signatures for electronic documents. Like handwritten signatures, digital signatures must not be forgeable; a recipient must be able to verify it, and the signer must not be able to repudiate it later. But unlike handwritten signatures, digital signatures incorporate the data (or the hash of the data) that are signed. Different data therefore result in different signatures even if the signatory is unchanged. As of this writing, many countries have or are about the put in place laws for electronic or digital signatures (e.g., the U.S. Electronic Signatures in Global and National Commerce Act). In addition, there are many algorithms and standards that can be used for digital signatures. Examples include RSA, ElGamal, and the Digital Signature Standard (DSS).
- *Access control mechanisms*, which use the authenticated identities of principals, information about these principals, or capabilities to determine and enforce access rights and privileges. If a principal attempts to use an unauthorized resource, or an authorized resource with an improper type of access, the access control function (e.g., the reference monitor) must reject the attempt and may additionally report the incident for the purposes of generating an alarm and recording it as part of a security audit trail. Access control mechanisms and the distinction between discretionary access control (DAC) and mandatory access control (MAC) have been extensively discussed in the computer security literature. They are usually described in terms of subjects, objects, and access rights. A subject is an entity that can access objects. It can be a host, a user, or an application. An object is a resource to which access should be controlled and can range from a single data field in a file to a large program. Access rights specify the level of authority for a subject to access an object, so access rights are defined for each subject/object pair. Examples of UNIX access rights include read, write, and execute. More recently, the idea of role-based access controls (RBACs) has been proposed and adopted by operating system and application software developers.
- *Data integrity mechanisms*, which are used to protect the integrity of either single data units and fields within these data units or sequences of data units and fields within these sequences. Note that data integrity mechanisms, in general, do not protect against replay attacks that work by recording and replaying previously sent messages. Also, protecting the integrity of a sequence of data units and fields within these data units generally requires some form of explicit ordering, such as sequence numbering, timestamping, or cryptographic chaining.
- *Authentication exchange mechanisms*, which are used to verify the claimed identities of principals. In accordance with ITU-T recommendation X.509, the term "strong," is used to refer to an authentication exchange mechanism that uses cryptographic techniques to protect the messages that are exchanged, whereas the term "weak" is used to refer to an authentication exchange mechanism that does not do so. In general, weak authentication exchange mechanisms are vulnerable to passive wiretapping and replay attacks, and the widespread use of weak authentication exchange mechanisms is the single most important vulnerability of contemporary computer networks and distributed systems.
- *Traffic padding mechanisms*, which are used to protect against traffic analysis attacks. Traffic padding refers to the generation of spurious instances of communication, spurious data units, and spurious data within data units. The aim is not to reveal if data that are being transmitted actually represent and encode information. Consequently, traffic padding mechanisms can only be effective if they are protected by some sort of a data confidentiality service. Furthermore, traffic padding is effective in leased lines or circuit-switched networks. It is not particularly useful in packet-switched data networks, such as TCP/IP networks and the Internet.
- *Routing control mechanisms*, which can be used to choose either dynamically or by prearrangement specific routes for data transmission. Communicating systems may, on detection of persistent passive or active attacks, wish to instruct the network service provider to establish a connection via a different route. Similarly, data carrying certain security labels may be forbidden by a security policy to pass through certain networks or links.
- Notarization mechanisms, which can be used to assure certain properties of the data communicated between two or more entities, such as its integrity, origin, time, or destination. The assurance is provided by a trusted third party (TTP) in a testifiable manner.

There are many products that implement specific security mechanisms to provide one (or several) security service(s).

3.2.2. Pervasive Security Mechanisms. Pervasive security mechanisms are not specific to any particular security service and are in general directly related to the level of security required. Some of these mechanisms can also be regarded as aspects of security management. The OSI security architecture enumerates the following five pervasive security mechanisms.

- The general concept of *trusted functionality* can be used to either extend the scope or to establish the effectiveness of other security mechanisms. Any functionality that directly provides, or provides access to, security mechanisms should be trustworthy.
- System resources may have *security labels* associated with them, for example, to indicate sensitivity levels. It is often necessary to convey the appropriate security label with data in transit. A security label may be additional data associated with the data transferred or may be implicit (e.g., implied by the use of a specific key to encipher data or implied by the context of the data such as the source address or route).

- Security-relevant *event detection* can be used to detect apparent violations of security.
- A *security audit* refers to an independent review and examination of system records and activities to test for adequacy of system controls, to ensure compliance with established policy and operational procedures, to detect breaches in security, and to recommend any indicated changes in control, policy, and procedures. Consequently, a *security audit trail* refers to data collected and potentially used to facilitate a security audit.
- *Security recovery* deals with requests from mechanisms such as event handling and management functions, and takes recovery actions as the result of applying a set of rules.

The OSI security architecture can be used to discuss the security properties of any computer network. In the following section, it is used to discuss Internet security as a case study. Similar discussions could be held for any other networking technology, such as wireless networks (e.g., GSM, GPRS, and UMTS networks).

4. CASE STUDY: INTERNET SECURITY

Today, the Internet is omnipresent and issues related to network security are best illustrated using the Internet as a working example for an international information infrastructure. In the past, we have seen many network-based attacks, such as password sniffing, IP spoofing and sequence number guessing, session hijacking, flooding, and other distributed denial of service attacks, as well as exploitations of well-known design limitations and software bugs. In addition, the use and wide deployment of executable content, such as that provided by Java applets and ActiveX controls, for example, have provided new possibilities to attack hosts and entire sites.

There are basically three areas related to Internet security: access control, communication security, and intrusion detection and response. These areas are reviewed next.

4.1. Access Control

In days of old, brick walls were built between buildings in apartment complexes so that if a fire broke out, it would not spread from one building to another. Quite naturally, these walls were called *firewalls*. Today, when a private TCP/IP network (i.e., a corporate intranet) is connected to a public TCP/IP network (e.g., the Internet), its users are usually enabled to communicate with the outside world. At the same time, however, the outside world can interact with the private network and its computer systems. In this situation, an intermediate system can be plugged between the private network and the public network to establish a controlled link, and to erect a security wall or perimeter. The aim of the intermediate system is to protect the private network from network-based attacks that originate from the outside world, and to provide a single choke point where security (i.e., access control) and audit can be imposed. Note that all traffic in and out of the private network can be enforced to pass through this single, narrow choke point. Also note that this point provides a good place to collect information about system and network use and misuse. As a single point of access, the intermediate system can record what occurs between the private network and the outside world. Quite intuitively, these intermediate systems are called *firewall systems*, or *firewalls* in short.

In essence, a firewall system represents a blockade between a privately owned and protected network, which is assumed to be secure and trusted, and another network, typically a public network or the Internet, which is assumed to be nonsecure and untrusted. The purpose of the firewall is to prevent unwanted and unauthorized communications into or out of the protected network. Therefore, it is necessary to define what the terms "unwanted" and "unauthorized" actually mean. This is a policy issue and the importance of an explicitly specified network security or firewall policy is not readily understood today.

In addition to the physical firewall analogy mentioned above, there are many other analogies that may help to better understand and motivate for the use of firewalls. Examples include the tollbooth on a bridge, the ticket booth at a movie theater, the checkout line at a supermarket, the border of a country, and the fact that apartments are usually locked at the entrance and not necessarily at each door. These analogies illustrate the fact that it sometimes makes a lot of sense to aggregate security functions at a single point. A firewall is conceptually similar to locking the doors of a house or employing a doorperson. The objective is to ensure that only properly authenticated and authorized people are able to physically enter the house. Unfortunately, this protection is not foolproof and can be defeated with enough effort. The basic idea is to make the effort too big for an average burglar, causing the burglar to eventually go away and find another, typically more vulnerable, house. However, just in case the burglar does not go away and somehow manages to enter the house, we usually lock up our valuable goods in a safe. According to this analogy, the use of a firewall may not always be sufficient, especially in high-security environments in which we live these days.

Roughly speaking, a firewall is a collection of hardware, software, and policy that is placed between two networks to control data traffic from one network to the other (and vice versa). There are several technologies that can be used to build firewall systems. Examples include (static or dynamic) packet filters, circuit-level gateways (e.g., SOCKS servers), and application-level gateways (i.e., proxy servers). These technologies are usually combined in either a dual-homed firewall or screened subnet firewall configuration with one or several demilitarized zones (DMZs). You may refer to Ref. 2 for an overview and discussion of firewall configurations. In either case, a network security or firewall policy must specify what protocols and services are authorized to traverse the firewall. Typically, a firewall implements a policy that does not restrict outbound connections, but that requires inbound connection to be strongly authenticated by a corresponding application-level gateway. Strong authentication can

be based on one-time password systems, such as SecurID or S/Key, or challenge–response mechanisms.

Today, the market for firewalls is mature and the corresponding products start to differentiate themselves through the provision of additional functionality, such as network address translation (NAT), content screening (e.g., virus scanning), virtual private networking, and intrusion detection. As such, firewalls are likely to stay in corporate environments to provide basic access control and complementary security services to intranet systems.

4.2. Communication Security

According to Shirey [1], the term *communication security* refers to "measures that implement and assure security services in a communication system, particularly those that provide data confidentiality and data integrity and that authenticate communicating entities." The term is usually understood to include cryptographic algorithms and key management methods and processes, devices that implement them, and the lifecycle management of keying material and devices. For all practical purposes, key management is the Achilles heel of any communication security system, and it is certainly the point where an attacker would start with.

Several (cryptographic) security protocols have been developed, proposed, implemented, and partly deployed on the Internet:

- On the *network access layer*, several layer 2 tunneling protocols are in use. Examples include the Point-to-Point Tunneling Protocol (PPTP) and the Layer 2 Tunneling Protocol (L2TP). L2TP is often secured with IPsec as discussed next.
- On the *Internet layer*, several layer 3 tunneling protocols are in use. Most importantly, the IETF (IP security) IPSEC WG has developed and standardized an IPsec protocol with a corresponding key management protocol called Internet Key Exchange (IKE). The IPsec and IKE protocols are in widespread use for virtual private networking. Most firewalls implement the protocols to interconnect network segments or mobile systems.
- On the *transport layer*, several (cryptographic) security protocols layered on top of TCP are in widespread use. Examples include the Secure Sockets Layer (SSL) and the Transport Layer Security (TLS) protocols. These protocols are sometimes also referred to as "session layer security protocols." Unfortunately, there is currently no widely deployed transport layer security protocol layered on top of UDP. This is unfortunate, because SSL/TLS does not provide a solution for UDP-based applications and application protocols.
- On the *application layer*, there are a number of (cryptographic) security protocols that are either integrated into specific applications and application protocols or provide a standardized application programming interface (API). The first class leads to security-enhanced application protocols, such as secure Telnet and secure FTP, whereas the second class leads to authentication and key distribution systems, such as Kerberos. With its use in UNIX and Microsoft operating systems (e.g., Windows 2000 and Windows XP), Kerberos is the most widely deployed authentication and key distribution system in use today.

Above the application layer, there are a few (cryptographic) security protocols that can be used to cryptographically protect messages before they are actually transmitted in computer networks. Examples include Pretty Good Privacy (PGP) and Secure MIME (S/MIME). Another possibility is to use the eXtended Markup Language (XML) with its security features that are currently being standardized by the World Wide Web Consortium (W3C).

Given this variety of (cryptographic) security protocols for the various layers in the Internet model, one may ask what protocol is best or which layer that is best suited to provide security services for Internet applications and users. Unfortunately, both questions are difficult to address and may require different answers for different security services. For example, data confidentiality services can be provided at lower layers, whereas nonrepudiation services are more likely to be provided at higher layers. In either case, the end-to-end argument applies [6]. Roughly speaking, the end-to-end argument states that the function in question (e.g., a security function) can completely and correctly be implemented only with the knowledge of the application standing at the endpoints of the communications system. Therefore, providing that function as a feature of the communications system itself is not possible (sometimes an incomplete version of the function provided by the communications system may be useful as a performance enhancement). This argument should always be kept in mind when network providers argue that security functions can easily be outsourced.

4.3. Intrusion Detection and Response

An *intrusion* refers to a sequence of related actions by a malicious adversary that results in the occurrence of unauthorized security threats to a target computing or networking domain. Similarly, the term *intrusion detection* refers to the process of identifying and responding to intrusions. This process is not an easy one. Nevertheless, there is an increasingly large number of tools that can be used to automate intrusion detection. The tools are commonly referred to as *intrusion detection systems* (IDSs). Although the research community has been actively designing, developing, and testing IDSs for more than a decade, corresponding products have received wider commercial interest only relatively recently. Furthermore, the IETF has chartered an Intrusion Detection Exchange Format (IDWG) WG "to define data formats and exchange procedures for sharing information of interest to intrusion detection and response systems, and to management systems which may need to interact with them."

There are basically two technologies that can be used to implement IDSs: attack signature recognition and anomaly detection.

1. Using *attack signature recognition*, an IDS uses a database with known attack patterns (also known as attack signatures) and an engine that uses this database to detect and recognize attacks. The database can either be local or remote. In either case, the quality of the IDS is as good as the database and its attack patterns as well as the engine that makes use of this database. The situation is similar and quite comparable to the antivirus software (i.e., the database must be updated on a regular basis).
2. Using *anomaly detection*, an IDS uses a database with a formal representation of "normal" (or "normal-looking") user activities and an engine that makes use of this database to detect and recognize attacks. For example, if a user almost always starts up his/her email user agent after having successfully logged onto a system, the IDS's engine may get suspicious if this user starts a Telnet session to a trusted host first. The reason for this activity may be an attacker misusing the account to gain illegitimate access to a remote system. Again, the database can either be local or remote, and the quality of the IDS is as good as the database and its statistical material.

Obviously, it is possible and useful to combine both technologies in a single IDS. The design of IDSs is a new and very active area of research and development. Many technologies that had originally been developed under the umbrella of artificial intelligence (AI) are being reused and applied to the problem of how to reliably detect intrusions. Exemplary technologies include knowledge-based systems, expert systems, neural networks, and fuzzy logic. In fact, some of these technologies have experienced a revival in the field of intrusion detection.

Once an intrusion is detected, it is important to respond in appropriate ways. Therefore, large organizations usually establish and maintain an incidence response team (IRT). For smaller organizations, it is usually more efficient to outsource this task to a commercially operating IRT.

5. CONCLUSIONS AND OUTLOOK

Network security is a hot topic today. Like many other topics related to IT security, network security has many aspects and the OSI security architecture may serve as a primary reference to structure them. In this article, we introduced the OSI security architecture and elaborated on the current state-of-the-art and future perspectives of network security in general, and Internet security in particular. More specifically, we looked into three areas that are particularly important for Internet security: access control, communication security, and intrusion detection and response.

The network security industry has shifted away from an industry that focused primarily on the use of preventive security technologies to an industry that also takes into account the importance of detective and reactive security technologies (under the term "detection and response"). The major argument for this paradigm shift is the insight that preventive security technologies are not complete in the sense that they will always leave vulnerabilities that are still exploitable by attackers, and that administrators must know how to detect attacks and counteract on them. Consequently, detection and response is equally important to prevention and an increasingly large number of companies are providing (security) monitoring services to their customers [7]. The importance of detection and response is likely to continue in the future and we will see many companies starting to specialize in this particular field.

BIOGRAPHY

Rolf Oppliger studied computer science, mathematics, and economics at the University of Berne, Switzerland, where he received M.Sc. and Ph.D. degrees in computer science in 1991 and 1993, respectively. In 1999, he received the Venia legendi for computer science from the University of Zürich, Switzerland. The focus of his professional activities is information technology (IT) security in general, and network security in particular. He has authored nine books, including, for example, the second editions of *Internet and Intranet Security* (Artech House, 2002) and *Security Technologies for the World Wide Web* (Artech House, 2003), frequently speaks at security-related conferences, and regularly publishes papers and articles in scientific magazines and journals. He's the founder and owner of eSECURITY Technologies Rolf Oppliger (www.esecurity.ch), Gümligen, Switzerland works for the Swiss Federal Strategy Unit, Bern, Switzerland, for Information Technology (FSUIT), teaches at the University of Zürich, and serves as editor for the Artech House Computer Security Series and the Swiss digma magazine for data law and information security. He's a member of the Association for Computing Machinery (ACM), the IEEE Computer Society, and served as vice chair of the IFIP TC 11 working group on network security.

BIBLIOGRAPHY

1. R. Shirey, *Internet Security Glossary*, RFC 2828, May 2000.
2. R. Oppliger, *Internet and Intranet Security*, 2nd ed., Artech House, Norwood, MA, 2001.
3. ISO/IEC 7498-2, *Information Processing Systems—Open Systems Interconnection Reference Model—Part 2: Security Architecture*, 1989.
4. ITU X.800, *Security Architecture for Open Systems Interconnection for CCITT Applications*, 1991.
5. J. Zhou, *Non-repudiation in Electronic Commerce*, Artech House, Norwood, MA, 2001.
6. J. H. Saltzer, D. P. Reed, and D. D. Clark, End-to-end arguments in system design, *ACM Trans. Comput. Syst.* **2**(4): 277–288 (1984).
7. B. Schneier, *Secrets and Lies: Digital Security in a Networked World*, Wiley, New York, 2000.

NETWORK TRAFFIC MANAGEMENT

SONIA FAHMY
Purdue University
West Lafayette, Indiana

1. INTRODUCTION

Communication networks have experienced tremendous growth in size, complexity, and heterogeneity since the late 1980s. With the surging popularity of many diverse Internet applications and the increase in content distribution, a "tragedy of the commons" situation has arisen where access must be controlled and congestion must be avoided. Internet traffic can be classified according to the application generating it. A representative list of applications includes video, voice, image, and data in conversational, messaging, distribution, and retrieval modes. These applications are either inelastic (real time), which require end-to-end delay bounds, or elastic, which can wait for data to arrive. Real-time applications can be further subdivided into those that are intolerant to delay, and those that are more tolerant, called *delay-adaptive*. This chapter surveys the required network traffic management building blocks for both types of application traffic. We begin by discussing traffic management objectives and components, and then devote a section for each of these components. We also include a number of case studies that illustrate how these traffic management components can be used to compose services for Internet applications.

1.1. Traffic Management Objectives

Traffic management aims at delivering a negotiated quality of service (QoS) to applications and at controlling congestion. This implies that critical or real-time application traffic may be given better service at network nodes than less critical traffic. In addition, congestion must be controlled to avoid the performance degradation and congestion collapse that occur when network buffers overflow and packets are lost. The network load should not increase beyond a certain optimal operating point, commonly known as the "knee" of the delay throughput curves. This is the point beyond which increasing the load level on the network results in a dramatic increase in end-to-end delay, caused by network congestion and retransmissions. Therefore, the objectives of network traffic management include

1. *Fairness*. Traffic sources should be treated according to some fairness criteria, such as (weighted) max–min fairness (with or without minimum guarantees) [8,42,48], or proportional fairness, which can be tied to pricing through appropriate utility functions [21,52]. *Max–min fairness* gives equal (or weighted) shares to sources sharing a common bottleneck. This means that the share of the constrained (min) sources is maximized, and excess resources are distributed equally among unconstrained sources. Given a configuration with n contending sources, suppose that the ith source is allocated a bandwidth x_i. The allocation vector $\{x_1, x_2, \ldots, x_n\}$ is feasible if all link load levels are less than or equal to 100%. Given an allocation vector, the source with the smallest allocation is, in some sense, the "unhappiest source." We find the feasible vectors that give the maximum allocation to this unhappiest source (thus maximizing the minimum source, or max–min). Then, we remove this "unhappiest source" and reduce the problem to that of the remaining $n - 1$ sources operating on a network with reduced link capacities. We repeat this process until all sources have been allocated the maximum that they can obtain.
2. *Efficient Resource Utilization*. The available resources, such as network buffers, network link bandwidths, processing capabilities, proxy servers, should be efficiently utilized.
3. *Bounded Queuing Delay*. Queuing delay should be small to guarantee low end-to-end delay according to application QoS requirements, and to ensure buffers do not overflow and cause excessive packet loss. Guarantees made to an application can be either deterministic (given for all packets), or statistical. Statistical guarantees can be made in steady state or over specific intervals of time, for instance, over no more than $x\%$ of time intervals will have more than $y\%$ of the packet delays exceed 5 ms.
4. *Stability*. The transmission rates of the sources should not unnecessarily fluctuate in steady state.
5. *Fast Transient Response*. Traffic sources should react rapidly to changing network conditions, such as sudden congestion. Performance should be acceptable even when there is no steady state. Thus, traffic management operations should be robust.
6. *Simplicity*. "Occam's Razor" dictates that entities are not to be multiplied beyond necessity. Traffic management algorithms should have reasonable time and space complexity. This includes scaling to large numbers of users.

Note that traffic management and congestion avoidance are dynamic problems. Static solutions such as increasing buffer size, bandwidth and processing power [43–46], namely, overprovisioning, do not sufficiently address dynamic application needs, especially when some traffic sources are not well behaved.

1.2. Traffic Management Building Blocks

Network traffic management has witnessed a flurry of research activity, especially since the late 1970s. We will use the terms *microflow*, *connection*, and *session* to denote a data stream identified by fields in the Internet Protocol (IP) and the Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) headers, such as source and destination address, protocol identifier, and source and destination ports. The datastream may be unicast (point-to-point), multicast (point-to-multipoint, or multipoint-to-multipoint) from a sending application to a set of receiving applications. We will use *flow* to denote either a microflow or an aggregate of microflows (a macroflow). We will use *end system* to denote a sender,

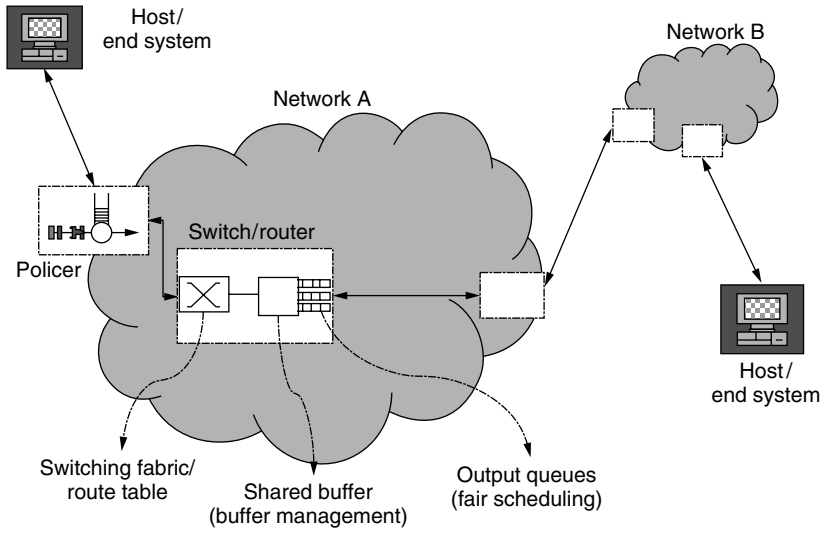


Figure 1. A network can use shaping or policing at the edge, and buffer management and scheduling at core routers and switches to meet guarantees. In combination with constrained routing, admission and policy control, and resource reservation, QoS is provided. Feedback-based congestion control is used to adapt to network conditions.

a receiver, or an edge router. An edge router can be an *ingress* router if it is at the entrance of a domain (with respect to a particular flow or connection); or an *egress* router if it is at the exit of a domain.

Figure 1 illustrates the traffic management components. Constrained routing, connection admission control, policy control, and resource reservations (not shown) are required to ensure that sufficient resources exist for QoS guarantees to be met. Once this is done, traffic shaping and policing, scheduling, and buffer management are required to control resource usage and provide QoS, as shown in the figure. Finally, traffic monitoring and feedback control are important to avoid congestion collapse in computer networks.

An interesting point to note is that the traffic management components bear some similarities to traffic management on transportation highways. For example, an analogy can be drawn between traffic shaping and stoplights that control traffic merging onto a highway on-ramp. Network traffic policing is also analogous to traffic patrol cars that stop speeding vehicles. Packet scheduling resembles several lanes merging onto one lane, or stop signs or lights that control traffic proceeding through an intersection. Buffer management is, in some sense, similar to traffic exiting the highway through various offramps. Finally, constrained routing and congestion control are in some sense similar to listening to traffic reports on your radio and deciding on alternative routes (constrained routing), or deciding not to leave your home yet, if critical highways are congested (congestion control).

The traffic management components in data networks are summarized in Table 1. Admission and policy control, resource reservation and constrained routing are typically performed at a coarse timescale, such as the connection setup time and during renegotiation. Thus we will refer to them as *session-level* (or *connection-level*) operations. Congestion control is invoked as a reaction to network state, and hence the response to network state is on the order of a burst or a round-trip time (time to send a packet from the source to the destination and time to receive feedback back to the source). The remaining traffic management components, namely, traffic shaping, policing, scheduling and buffer management, are performed at the packet level at network switches and routers. Other operations, such as capacity planning and pricing [18,58], operate on timescales on the order of connections, days, or even weeks, and we do not include those in our discussion.

As listed in the fourth column of Table 1, some of the traffic management mechanisms operate at end systems (sources or edge routers) such as traffic shaping; some operate at network switches or routers such as buffer management; and some require both end systems and network switches to cooperate such congestion control using explicit feedback from the network. Congestion control is a closed-loop form of control, while the other operations are typically open-loop, although they may sometimes use measurement to perform better decisions. In the case studies (Section 6), we will see how these building blocks are used in the Internet integrated (IntServ) and differentiated (DiffServ) services, in asynchronous transfer mode (ATM), and with traffic engineering (TE) for label-switched paths.

2. CONSTRAINT-BASED ROUTING

Several routing algorithms that base path selection decisions on policy or quality of service (QoS) have been proposed for the Internet since the mid 1990s. Constraint-based routing usually considers flow aggregates (also known as *macroflows* or *trunks*), rather than individual micro-flows (e.g., a single HTTP connection). Routing constraints may be imposed by administrative policies (Section 2.1), or by the application QoS requirements (Section 2.2).

2.1. Policy Routing

Policy-based routing chooses paths conformant to administrative rules and service agreements. With policy-based routing schemes, the administrators can base routing decisions not only on the destination location but also on factors such as the applications, the protocols used, the size of packets, or the identity of end systems. As the Internet

Table 1. Traffic Management Components and Their Timescales

Timescale	Component	Definition	Location	Open/Closed-Loop
Session level	Admission control	Determines if a new connection/flow requirements can be met without affecting existing connections	Routers/end systems	Open/measured
	Policy control	Determines if a new connection/flow has the administrative permissions to be admitted	Routers/end systems	Open
	Resource Reservation	Sets up resource reservations in network nodes for an admitted connection/flow	Routers/end systems	Open/measured
	Constrained routing	Selects a path based on requirements that are either administrative-oriented (policy-based routing) or service-oriented (QoS routing)	Routers/end systems	Open/measured
Round-trip time (burst level)	Congestion control	Controls the input load to the optimal operating point	End systems with or without router assistance	Closed
Packet level	Traffic shaping	Delays selected packets to smooth bursty traffic	End systems	Open
	Traffic policing	Drops selected packets to conform to a traffic profile	End systems	Open
	Packet scheduling	Determines which packet to transmit next onto the output link	Routers/end systems	Open
	Buffer management	Determines which packets to admit into a buffer	Routers/end systems	Open

continues to grow and diverse Internet services are offered, more stringent administrative constraints can ensure adequate service provisioning and safety from malicious users attempting to obtain services that do not conform to their service agreements or profiles without paying for such services. Policy-based routing can also provide cost savings, load balancing, and basic QoS. Policy constraints are applied before the application of the required QoS constraints (Section 2.2). Policy constraints may be exchanged by the routing protocols while updating route information, or simply provided manually during network configuration. In the latter case, the main problem that may occur is policy rule conflicts.

2.2. QoS Routing

QoS routing can be defined as "a routing mechanism under which paths for flows are determined based on some knowledge of resource availability in the network as well as the QoS requirement of flows" [20]. As most deployed Internet routing strategies are developed for the best-effort model, they are sometimes unsuitable for emerging real-time application requirements. QoS routing extends the best-effort paradigm by finding alternate routes for forwarding flows that cannot be admitted on the shortest existing path. Unlike connectionless best-effort schemes, QoS routing is connection-oriented with resource reservation. QoS routing, however, determines a path only from a source to a destination and does not reserve any resources on that path [80]. A resource reservation technique such as RSVP (Section 6.1) or ATM UNI must then be employed to reserve the required resources. After a path is found and resources are reserved, all packets of the QoS flow must be forwarded through that path. This means that the path must be fixed throughout the lifetime of the flow. This is called "route pinning."

QoS routing dynamically determines feasible paths that also optimize resource usage. Many factors affect the performance of QoS routing solutions, including the particular QoS routing scheme used, the accuracy of information that the QoS routing scheme uses, the network topology, and the network traffic characteristics [63]. A key problem that arises with QoS routing is tractability. Optimizing a path for two or more quality metrics is intractable if the quality metrics are independent and allowed to take real or unbounded integer values [64]. If all metrics except one take bounded integer values, or if all the metrics except one take unbounded integer values but the maximum constraints are bounded, then the problem can be solved in polynomial time [2]. More recent studies show the possibility of performing QoS routing with inaccurate information without suffering significant loss in performance. It was also shown that applying aggregation techniques for scalability does not always negatively impact performance.

3. ADMISSION CONTROL

The Internet protocol (IP) currently supports a best effort service, where no delay or loss guarantees are provided. This service is adequate for non-time-critical applications, or time critical applications under light-load conditions. Under highly overloaded conditions,

however, buffer overflows and queuing delays cause the real-time communication quality to quickly degrade. To support real time applications, a new service model was designed [12]. In this model, both real-time and non-real-time applications share the same infrastructure, thus benefiting from statistical multiplexing gains.

Applications specify their traffic characteristics and their quality of service requirements. Admission control is employed to determine whether these requirements can be met. If they can be met, reservations are made, as discussed in Section 5. Using different classification, policing, shaping, scheduling, and buffer management rules, different applications are serviced with different priorities to ensure that the quality of service requirements are met.

Therefore, admission control is the process where, given the current set of connections and the *traffic characteristics* of a new connection, a decision can be made on whether it is possible to meet the new connection *quality of service requirements*, without jeopardizing the performance of existing connections. Traffic characteristics are commonly described by traffic descriptors. These typically include a subset of the following components: a peak rate, an average rate, and the maximum burst size, that can be enforced by a token bucket or leaky bucket (described in Section 7). For example, in ATM networks the generic cell rate algorithm (GCRA) is used to enforce the peak rate (PCR/CDVT) and average (sustained) rate (SCR/BT) parameters, and the maximum burst size (MBS). QoS requirements are negotiated by the source with the network and used to define the expected quality of service provided by the network. The parameters typically include a maximum delay, delay variation (jitter) and packet loss ratio. For each service, the network guarantees the negotiated QoS parameters if the end system complies with the negotiated traffic contract. For noncompliant traffic, the network need not maintain the QoS objective.

Note that existing connections are accounted for in the admission control algorithm in several possible ways. The declared traffic characteristics of existing connections can be used in the QoS computations. Alternatively, measurement-based admission control (MBAC) can be used, where the new connection declared traffic characteristics are combined with the *measured* traffic characteristics of existing connections, in order to compute whether the QoS would be acceptable to the new connection [49]. This, however, assumes that past measurements are sufficiently correlated with future behavior, which may not always hold. More recently, endpoint admission control has been proposed, where the hosts (the endpoints) probe the network to detect the level of congestion. The host admits a new flow only if the level of congestion is sufficiently low [15].

4. POLICY CONTROL

It is important to firmly control which users are allowed to reserve resources, and how much resources they can reserve. Network managers and service providers must be able to monitor, control, and enforce use of network resources and services based on policies derived from criteria such as the identity of users and applications, traffic or bandwidth requirements, security considerations, or time of day or week. Figure 2 depicts the typical policy control architecture (standardized for the Internet). In this model, a protocol called *Common Open Policy Service* (COPS) is used for policy rule exchange between a policy server [referred to as a *policy decision point* (PDP)] and a network device [referred to as a *policy enforcement point* (PEP)] [11]. The *Lightweight Directory Access Protocol* (LDAP) is used for policy rule retrieval from a policy repository. A policy repository is a server dedicated to the storage and retrieval of policy rules. The Policy Management Console is the coordinator of the entire policy management process.

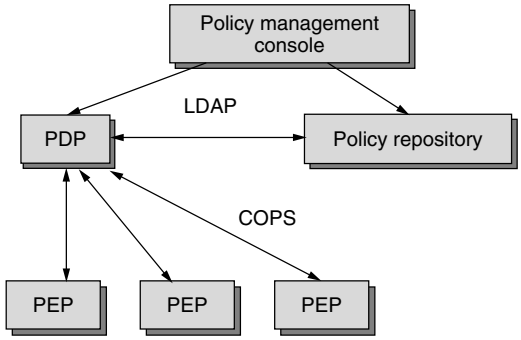


Figure 2. LDAP is used to retrieve the rules from the policy server, and COPS exchanges policy rules among the PDP and PEPs.

5. RESOURCE RESERVATION

A resource reservation protocol is the means by which applications communicate their requirements to the network in an efficient and robust manner. Applications that receive real-time traffic inform the network of their needs, while applications that send real-time traffic inform the receivers and network about their traffic characteristics. The reservation protocol is a "signaling" protocol (a term originating from telephone networks) that installs and maintains reservation state information at each router along the path of a stream. The reservation protocol does not provide any network service; it can be viewed as a "switch state establishment protocol," rather than just a resource reservation protocol. The protocol transfers reservation data as opaque data — it can also transport policy control and traffic control messages.

Resource reservation protocols interact with the *admission control process* to determine whether sufficient resources are available to make the reservation, and the *policy control process* to determine whether the user has permission to make the reservation. If the reservation process gets an acceptance indication from both the admission control and policy control processes, it sends the appropriate parameter values to the packet classifier and packet scheduler. The *packet classifier* determines the QoS class of packets according to the requirements, and the *packet scheduler* (Section 8) and *buffer manager* (Section 9) manage various queues to guarantee the required quality of service. For example, to guarantee the bandwidth and

delay characteristics reserved, a fair packet scheduling scheme can be employed. Fair scheduling isolates datastreams and gives each stream a percentage of the bandwidth on a link. This percentage can be varied by applying weights derived from the reservations [30].

In ATM networks, the User-Network Interface (UNI) protocol establishes resource reservations. In the integrated services framework, the RSVP protocol is used, as discussed in Section 6.1.

6. EXAMPLE ARCHITECTURES

Before we discuss the packet-level and burst-level traffic management components, we will look at four architectures to see how the connection-level building blocks are composed to provide various services.

6.1. Integrated Services and RSVP

An example of a multiservice network is the integrated services framework, which requires resources to be reserved a priori for a given traffic *microflow*. The integrated services framework maps the three application types (delay-intolerant, delay-adaptive, and elastic) onto three service categories: the guaranteed service for delay intolerant applications, the controlled load service for delay adaptive applications, and the currently available best-effort service for elastic applications. The guaranteed service gives firm bounds on the throughput and delay, while the controlled load service tries to approximate the performance of an unloaded packet network [12,81].

Figure 3 illustrates the components of an integrated services router. The Resource Reservation Protocol (RSVP) [13] is the signaling protocol adopted to establish resource reservations state for both unicast and multicast connections. An RSVP sender uses the PATH message to communicate with receiver(s) informing them of microflow characteristics. RSVP provides receiver-initiated reservation of resources, using different reservation *styles* to fit a variety of applications. RSVP receivers periodically alert networks to their interest in a data microflow, using RESV messages that contain the source IP address of the requester and the destination IP address, usually coupled with microflow details. The network then allocates

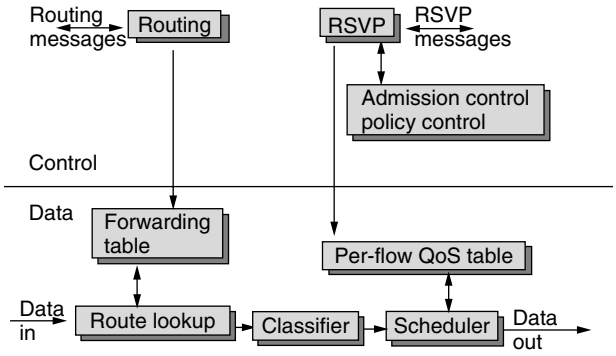


Figure 3. A router with QoS (integrated services) capabilities. RSVP interfaces with admission and policy control, and stores reservation information in a QoS table that is consulted when forwarding a flow.

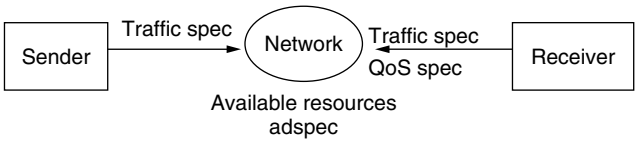


Figure 4. The RSVP sender announces its traffic specifications, and the receiver announces its QoS requirements. The network available resources are checked to see the requested QoS for this traffic can be supported.

the required bandwidth and defines priorities, as shown in Fig. 4. RSVP decouples the packet classification and scheduling from the reservation operation, transporting the messages from the source and destination as opaque data. Periodic renewal of state (soft state) allows networks to be self-correcting despite routing changes and loss of service. This enables routers to understand their current topologies and interfaces, as well as the amount of network bandwidth currently supported.

An RSVP reservation request consists of a **FlowSpec**, specifying the desired QoS, as well as a **FilterSpec**, defining the flow to receive the desired QoS. The FlowSpec is used to set parameters in the packet scheduler, while the FilterSpec is used in the packet classifier. The FlowSpec in a reservation request will generally include a service class and two sets of numeric parameters: (1) an **RSpec** (R for "reserve") that defines the desired QoS, and (2) a **TSpec** (T for "traffic") that describes the data flow. The basic FilterSpec format defined in the present RSVP specification has a very restricted form: sender IP address, and optionally the UDP/TCP source port number.

The main problem with the integrated services model has been its scalability, especially in large public IP networks, which may potentially have millions of concurrent microflows. RSVP exhibits overhead in terms of state, bandwidth, and computation required for each microflow. One of the solutions proposed to this problem is the aggregation of flows, and the simplification of core router state and computation, used in the differentiated services framework.

6.2. Differentiated Services

The differentiated services (DiffServ) framework provides a scalable architecture for Internet service differentiation [9,19]. The main DiffServ design principles are the separation of policy from mechanisms, and pushing complexity to the network domain boundaries, as illustrated in Fig. 5. For a customer to receive DiffServ from its Internet Service Provider (ISP), the customer should have a *service-level agreement* (SLA) agreed on with the ISP. Bandwidth brokers (BBs) [54,62] perform coarse-grained long-term admission and policy control and configure the edge (ingress and egress) routers.

DiffServ core routers are only responsible for forwarding based on the classification performed at the edge. The Differentiated Services Code Point (DSCP) (contained in the IP header DSFIELD/ToS) [61] is used to indicate the forwarding treatment a packet should receive (Fig. 5). DiffServ standardizes a number of per hop behaviors (PHBs) employed in the core routers, including a PHB, expedited

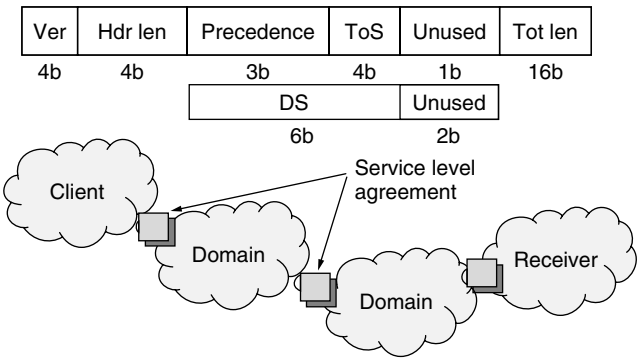


Figure 5. Differentiated Services networks use marking at the edge routers according to bilateral service level agreements, and simple forwarding in the core according to the 6 DS bits in the IP header.

forwarding (EF), and a PHB group, assured forwarding (AF) [37,41]. EF provides a low loss low delay service by employing strict admission control procedures. AF provides a set of better than best effort services for more bursty traffic, by using a number of levels of services, that use multiple queues and multiple drop priorities per queue.

6.3. Asynchronous Transfer Mode (ATM) Networks

Asynchronous transfer mode (ATM) networks were proposed to transport a wide variety of traffic, such as voice, video, and data, in a seamless manner. ATM transports data in fixed size 53 byte-long packets, called *cells*. End systems must set up virtual channel connections (VCCs) of appropriate service categories prior to transmitting information. Service categories are a small number of general ways to provide QoS, which are appropriate for different classes of applications. ATM service categories distinguish real-time from non-real-time services, and provide simple and complex solutions for each case. The added mechanisms in the more complex categories are justified by providing a benefit or economy to a significant subset of the applications [33].

ATM provides six service categories: constant bit rate (CBR), real-time variable bit rate (rt-VBR), non real-time variable bit rate (nrt-VBR), available bit rate (ABR), guaranteed frame rate (GFR), and unspecified bit rate (UBR) [32]. The *constant-bit-rate* (CBR) service category guarantees a constant rate called the *peak cell rate* (PCR). The network guarantees that all cells emitted by the source that conform to this PCR are transferred by the network at PCR. The *real-time variable-bit-rate* (VBR-rt) class is characterized by PCR, sustained cell rate (SCR), and maximum burst size (MBS), which control the bursty nature of traffic. The network attempts to deliver cells of these classes within fixed bounds of cell transfer delay (max-CTD) and cell delay variation (peak-to-peak CDV). *Non-real-time VBR* sources are also specified by PCR, SCR, and MBS, but the network does not specify the CTD and CDV parameters for VBR-nrt.

The *available-bit-rate* (ABR) service category is specified by a PCR as well as a minimum cell rate (MCR), which is guaranteed by the network. Excess bandwidth is shared in a fair manner by the network. We discuss ABR further in Section 10.5. The *unspecified bit rate* (UBR) does not support any service guarantees. UBR VCs are not required to conform to any traffic contract. PCR, however, may be enforced by the network. Switches are not required to perform any congestion control for UBR VCs. When queues become full, switches simply drop cells from UBR connections. Some improvements to UBR, known as UBR+, have been proposed. The *guaranteed-frame-rate* (GFR) service category is an enhancement of UBR that guarantees a minimum rate at the frame level. GFR is different from ABR because it does not use feedback control. The GFR class is intended to be a simple enhancement of UBR that guarantees some minimum rate to application frames.

6.4. Multiprotocol Label Switching

Multiprotocol label switching (MPLS) [73] uses fixed length labels, attached to packets at the ingress router. Forwarding decisions are based entirely on these labels in the interior routers of the MPLS path, as illustrated in Fig. 6. MPLS has made constraint-based routing a viable approach in IP networks [5,6]. Constraint-based routing can reduce manual configuration and intervention required for realization of traffic engineering objectives [6]. The traffic engineer can use administratively configured routes to perform optimizations. This enables a new routing paradigm with special properties, such as being resource-reservation-aware and demand-driven, to be merged with current Internet Gateway routing Protocols (IGPs), such as the Open Shortest Path First (OSPF), or the Intermediate System-Intermediate System (IS-IS) protocols. A constraint-based routing process incorporated in layer 3 and its interaction with MPLS and the current Internet Gateway Protocols (IGPs) is shown in Fig. 7. Constraint-based routing requires schemes for exchanging state information among processes, maintaining this state information, interaction with the current IGP protocols, and accommodating the adaptivity and survivability requirements of MPLS traffic trunks (aggregates of traffic flows) [20].

6.5. Interoperability Among Different Architectures

When QoS networks are deployed, typically only edge networks would be RSVP-enabled, and the core transit network would be DiffServ-enabled, use ATM, or use MPLS with constraint-based routing. In this scenario

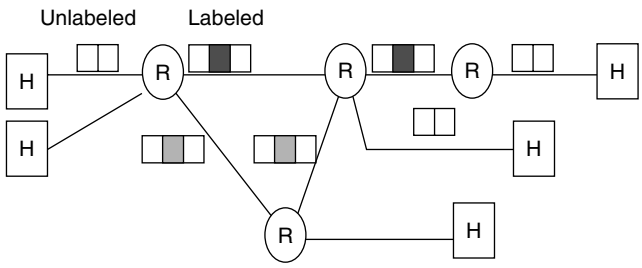


Figure 6. In MPLS, labels are attached to packets and used to perform switching decisions, until the labels are removed. Labels can also be nested.

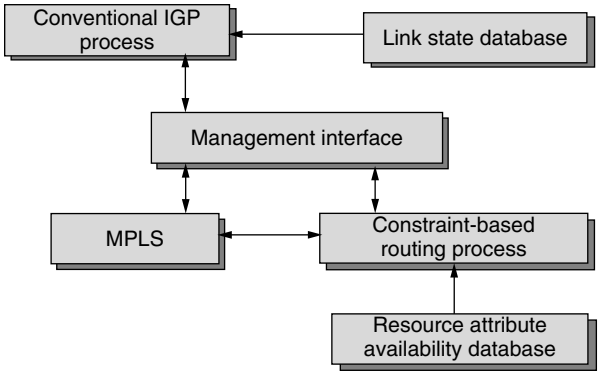


Figure 7. The constraint-based routing process interfaces with MPLS and resource databases.

the RSVP networks (at the edges) may be considered as customers of the transit DiffServ/ATM/MPLS network. The edge routers (at the edge of RSVP and DiffServ networks, for example) would be both RSVP- and DiffServ-capable. RSVP signaling messages are carried transparently through the DiffServ network; only the RSVP-enabled networks process RSVP messages. The DSCP marking can be done either at the host itself or at an intermediate router. RSVP reservations have to be converted into appropriate DiffServ PHBs for achieving end-to-end QoS. MPLS [73] may also be used for establishing label switched paths and trunks based on traffic engineering parameters, as discussed in Section 6.4.

7. POLICING AND SHAPING

In addition to the connection-level operations we have discussed so far, some traffic management operations must be performed for every packet. The end systems and edge routers are responsible for sending data that conforms to a negotiated traffic contract. As previously discussed, a traffic contract typically specifies the average rate, peak rate, and maximum burst size of a traffic flow. An incoming packet is checked against a traffic meter, and a decision is made on whether it is conforming (in profile) or non-conforming (out of profile). The shaping and policing functions (sometimes called usage parameter control (UPC)) have four possible choices when a packet is out of profile:

- *Dropping*. The nonconforming packet can be dropped to ensure that the traffic entering the network conforms to the contract, that is, that traffic is *policed* according to the profile.
- *Marking* (*Tagging*). The nonconforming packet can be marked as a low-priority packet by setting one or more bits in the packet header. In ATM, this is done by setting the value of the cell loss priority (CLP) bit in the ATM header to 1. In DiffServ networks, the DSCP is marked to reflect one of three priority levels, as discussed below. During congestion, the network may choose to discard low-priority packets in preference to high priority packets.
- *Buffering*. The nonconforming packet may be buffered and sent at a later time when it becomes conforming to the contract. This *traffic shaping* function reduces the traffic variation and burstiness and makes traffic smooth. It also provides an upper bound for the rate at which the flow traffic is admitted into the network, thus aiding in computing QoS bounds and buffer requirements [22,23].
- *No Action*. The nonconforming packet may be allowed into the network without any changes. This is typically an undesirable solution because it can cause congestion in the interior of the network.

The leaky-bucket and the token bucket algorithms have been designed for shaping and policing traffic. Partridge [69] describes the *leaky-bucket algorithm*, which was based on ideas discussed by Turner in 1986. The leaky bucket buffers incoming packets in a "bucket" that "leaks" at a certain rate. The algorithm has two input parameters: (1) the depth (size) of the bucket, b, and (2) the rate at which packets are drained out of the bucket, r. The generic cell rate algorithm (GCRA) [32] that is used in ATM is a variation of the leaky bucket.

Whereas the leaky bucket is filled with incoming packets and transmits them (if any are present) at a fixed rate (for shaping), the token bucket indicates whether traffic can be transmitted based on the availability of tokens. Tokens are added to the bucket at a fixed rate r, and can accumulate only up to the bucket depth b (where b controls the maximum burst size). The appropriate number of tokens are consumed when traffic is transmitted. Traffic may be allowed to be sent *in a burst* as long as a sufficient number of tokens is available in the bucket. This is the primary difference between a leaky bucket and a token bucket — a leaky bucket additionally controls the drain rate. Combinations of both are typically used to control the peak rate, average rate, and maximum burst size, according to the service provided.

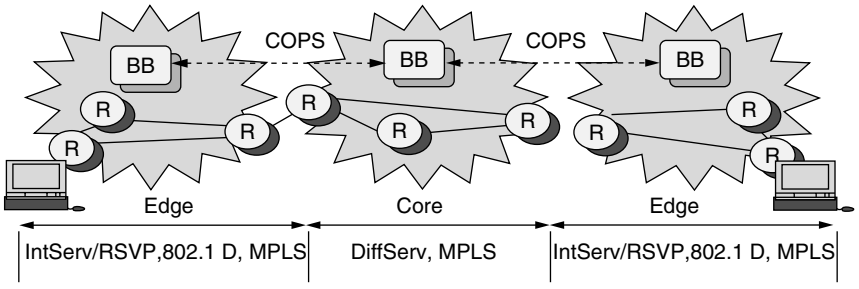


Figure 8. Networks with MPLS label-switched paths, ATM, or QoS (e.g., IntServ/RSVP or DiffServ) capabilities can interoperate. Bandwidth brokers (BBs) use the COPS protocol to exchange policy information among different domains.

For example, in the ATM variable-bit-rate (VBR) service, leaky buckets are used to control the peak rate, PCR, with tolerance CDVT, and to control the sustained (average) rate, SCR, with tolerance BT. The maximum burst size is limited to MBS. In DiffServ networks, the edge router contains meters, markers, droppers, and shapers, collectively referred to as *traffic conditioning functions*. A traffic conditioner may re-mark a traffic stream or discard or shape packets to alter the temporal characteristics of the stream and bring it into compliance with a traffic profile specified by the network administrator. As shown in Fig. 9, incoming traffic passes through a classifier, which is used to select a class for each traffic flow. The meter measures and sorts the classified packets into precedence (priority) levels. The decision (marking, shaping, or dropping) is based on the measurement result.

DiffServ-assured forwarding provides up to three drop precedences for each queue, as depicted in Fig. 10. Assume that the drop precedences are DP0 (green), DP1 (yellow) and DP2 (red), where DP0 means lower precedence to drop, and DP2 means higher (similar to colors in traffic stoplights). A three-color-marker (TCM) is used to mark packets with one of these three precedences. The DSCP is set to one of 3 values according to the DP. Traffic conditioners may also be TCP-aware and choose to protect "critical" TCP packets by marking them as DP0 [35].

8. SCHEDULING

Another packet-level traffic management function is packet scheduling. At every router or switch output port, a *scheduling discipline* must be used decide which packet to transmit next onto the output link (and onto the next hop). Scheduling is important because it resolves contention for a shared resource (the output link) and determines whether real-time applications can be given performance guarantees. Packet scheduling services different queues with different priorities to ensure that the quality of service requirements are met (recall Fig. 10). Multiple queues are needed because if packets from all flows share the same buffer using first-in first-out (FIFO) queuing, it is impossible to isolate packets from various flows. The degree of aggregation (how many microflows share a queue) determine the guarantees that can be made, as well as the state and time complexity of the scheduling discipline.

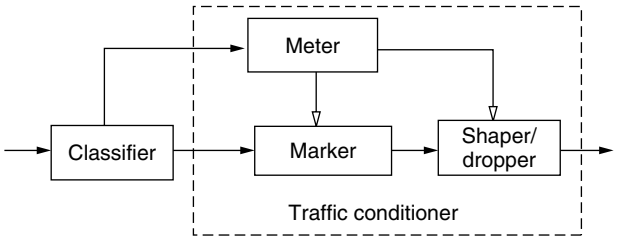


Figure 9. An edge router typically includes a classifier, a meter, a marker, a shaper, and a dropper.

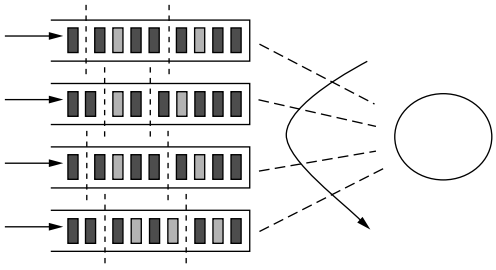


Figure 10. The assured forwarding service queues packets into four queues, with three drop precedences (denoted by three colors) per queue.

Scheduling disciplines aim at (1) meeting performance bounds for guaranteed services, including bandwidth, delay, delay variation (jitter), and loss; (2) providing some level of fairness and protection among flows; (3) being easily integrated with admission control algorithms; and (4) having low state and time complexity. A tradeoff among these goals must be achieved.

Scheduling algorithms may be work-conserving or non-work-conserving. A work-conserving scheduler is idle only when there is no packet awaiting service. Non-work-conserving schedulers may choose to be idle even if they have packets awaiting service, in order to make outgoing traffic more predictable and reduce delay variation (jitter). The best-known work-conserving scheduling discipline is the *generalized processor sharing* (GPS) discipline [67,68]. GPS serves packets as if they are in separate logical queues, servicing an infinitesimally small amount of data from each nonempty queue in turn. This intuitively resembles a *bit-by-bit* round-robin service. Connections can also be assigned weights and can be serviced in proportion to their weights. GPS cannot be directly implemented, but can be emulated as a weighted round robin or deficit round robin [36,75]. Deficit round robin can handle variable packet sizes without having to know the mean packet size of each connection in advance.

Weighted fair queuing (WFQ) approximates *packet-by-packet* GPS [7,24,82]. WFQ computes the time a packet would complete service in a GPS regime, and services packets in the order of these finishing times. A number of variants of WFQ have been developed, including self-clocked fair queuing (SCFQ), virtual clock (VC), and worst-case fair weighted fair queuing (WF^2Q). Other scheduling algorithms include delay and jitter earliest due date (EDD) and stop-and-go. The state overhead of scheduling algorithms can be alleviated if packets carry more information, as in the core-stateless fair queuing approach [77,78].

9. BUFFER MANAGEMENT

In most routers, packets are admitted into the router buffer as long as buffer space is still available. When the buffer is full, incoming packets have to be dropped — which is what is commonly referred to as the "drop tail" policy, since packets are dropped from the tail of the queue. This is the simplest policy to implement. Alternatively, packets may be dropped from the front of the queue, or from random locations within the queue. Such "pushout" mechanisms are typically more expensive.

Partial packet discard (PPD) schemes were first proposed to drop remaining segments, such as ATM cells of

an IP packet, if other segments of the packet have already been dropped. The intuition behind this is that the receiver will anyway discard segments of a packet if the complete packet cannot be reassembled. Therefore, these partial packets should not consume network resources (e.g., bandwidth) on the path between the router where a segment of the packet is discarded, and the receiver. Early packet discard (EPD) [72] has been proposed to extend this notion to drop complete packets when the buffer reaches a certain occupancy, say, 90%, in order to save the remaining capacity for partial segments of admitted packets.

Active queue management (AQM) in routers was later proposed to improve application goodput and response times by detecting congestion *early* and improving fairness among various flows. The main goal of AQM is to drop/mark packets before buffer overflow, in order to (1) give early warning to sources, (2) avoid synchronization among TCP congestion control phases of different flows (TCP congestion control is discussed in Section 10), (3) avoid bias against bursty connections, and (4) punish misbehaving sources.

Active queue management gained significant attention in the early 1990s with the design of the random early detection (RED) algorithm [29]. RED maintains a long-term average of the queue length (buffer occupancy) of a router using a lowpass filter. If this average queue length falls below a certain minimum threshold, all packets are admitted into the queue, as depicted in Fig. 11a. If the average queue length exceeds a certain maximum threshold, all incoming packets are dropped. When the queue length lies between the minimum and maximum thresholds, incoming packets are dropped/marked with a linearly increasing probability up to a maximum drop probability value, p_{max}. RED includes an option known as the "gentle" variant (Fig. 11b). With gentle RED, the packet drop probability varies linearly from p_{max} to 1 as the average queue size varies from th_{max} to twice th_{max}.

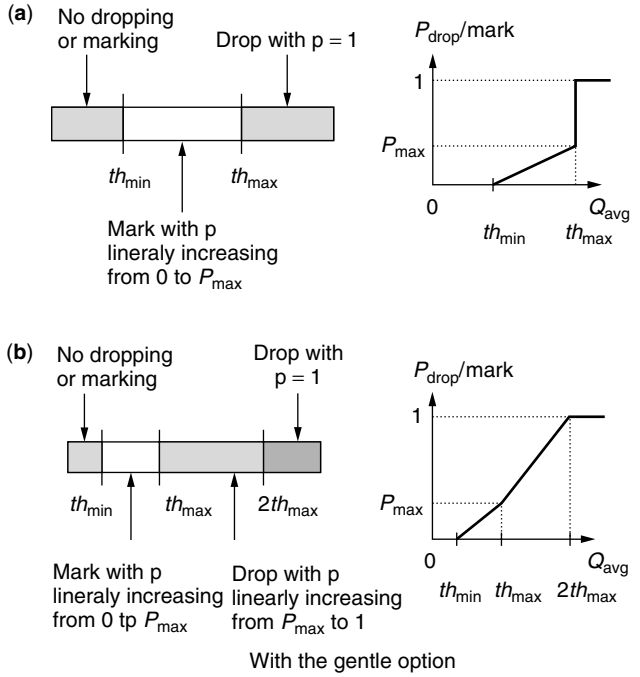


Figure 11. Random early detection (RED) probabilistically drops/marks packets based on average queue length without (**a**) and with (**b**) "gentle" option.

A number of RED variants have appeared, including flow-RED (FRED) [57], stabilized RED (SRED) [65], and BLUE [27]. Although FRED performs best among all RED variants, FRED maintains counts of the buffer occupancies for each flow in order to make better packet admission decisions. This provides the best isolation among flows, especially in the presence of misbehaving flows that send at high rates. However, maintaining per flow packet counts implies that FRED implementation complexity is higher than the other variants. Algorithms for the ATM guaranteed frame rate (GFR) service are also very similar in spirit to FRED. More recently, a number of other algorithms, including random early marking (REM) [4], adaptive virtual queue (AVQ) [55], and the proportional integrator (PI) controller [39], have been proposed. Although RED and these algorithms improve performance over simple drop-tail queues, it is difficult to configure their parameters, and some are complex to implement, so they are still under study.

If the network architecture supports marking packets with different drop precedence values, buffer management algorithms can provide differential drop. For example, in differentiated services networks, within each assured service queue, discrimination among packets can be performed using various mechanisms. The RIO (RED with IN and OUT) algorithm distinguishes between two types of packets, IN and OUT of profile, using two RED instances [19]. Each RED instance is configured with min_{th}, max_{th}, and P_{max} (recall Fig. 11a). Suppose the parameters for the IN profile packets are min_{in}, max_{in}, and P_{max_in}, and for the OUT-of-profile packets are min_{out}, max_{out}, and P_{max_out}. To drop OUT packets earlier than IN packets, min_{out} is chosen to be smaller than min_{in}. The router drops OUT packets more aggressively by setting P_{max_out} higher than P_{max_in}. To realize three drop precedences (red, yellow, and green), three REDs can be used.

10. CONGESTION CONTROL

In addition to connection-level and packet-level traffic management operations, network feedback can be used to control congestion at the burst level. This section discusses how closed-loop feedback operates, using the congestion control algorithms in TCP/IP and ATM networks as case studies.

10.1. TCP Congestion Control

In the year 2000, almost 90% of the Internet traffic used the TCP protocol, although multimedia traffic, especially RTP [74] and RTSP, have been slowly increasing. When congestion collapse was first experienced in the Internet in the 1980s [40,44,48], an investigation resulted into the design of new congestion control algorithms, now an essential part of the TCP protocol. Every TCP connection starts off in the "slow start" phase [40]. The slow-start algorithm uses a variable called *congestion window* (*cwnd*). The sender can only send the minimum of *cwnd* and

the receiver advertised window which we call *rwnd* (for receiver flow control). Slow start tries to reach equilibrium by opening up the window very quickly. The sender initially sets *cwnd* to 1 (or 2) and sending one segment. For each acknowledgment (ACK) the sender receives, *cwnd* is increased by one segment. Increasing by one for every ACK results in exponential increase of *cwnd* over round trips, as shown in Fig. 12. In this sense, the name "slow start" is a misnomer.

TCP uses another variable *ssthresh*, the slow-start threshold, to ensure *cwnd* does not increase exponentially forever. Conceptually, *ssthresh* indicates the "right" window size depending on current network load. The slow-start phase continues as long as *cwnd* is less than *ssthresh*. As soon as *cwnd* crosses *ssthresh*, TCP goes into "congestion avoidance." In the congestion avoidance phase, for each ACK received, *cwnd* is increased by 1/*cwnd* segments. This is *approximately* equivalent to increasing the *cwnd* by one segment in one round trip (an *additive* increase), if every segment (or every other segment) is acknowledged by the destination.

TCP maintains an estimate of the round-trip time (RTT), which is the time it takes for the segment to travel from the sender to the receiver plus the time it takes for the ACK (and/or any data) to travel from the receiver to the sender. The *retransmit timeout* (RTO) maintains the value of the time to wait for an ACK after sending a segment before assuming congestion, timing out and retransmitting the segment. When the TCP sender times out, it assumes the network is congested, and sets *ssthresh* to max(2, min(*cwnd*/2, *rwnd*)) segments, *cwnd* to one, and goes to slow start [1]. The halving of *ssthresh* is a *multiplicative* decrease. The *additive-increase multiplicative-decrease* (AIMD) system has been shown to be stable [17]. This basic version of TCP is called "TCP Tahoe."

10.2. TCP Flavors

Several variations on TCP congestion control have been designed, including TCP Reno [26], New-Reno [38], selective acknowledgments (SACK) [60], and forward acknowledgments (FACK) [59], to recover rapidly from one or multiple segment losses, detected through *duplicate* or *selective* acknowledgments, instead of only through timeouts as with the basic TCP (Tahoe). Other algorithms, such as TCP Vegas [14], use changes in round-trip time estimates, rather than packet loss, to adjust the TCP congestion window. Several TCP variations for wireless networks have also been proposed to operate in environments where bandwidth is limited, bandwidth may be asymmetric, and error rate is high. Because of possibly high error rates, losses are no longer assumed to be due to congestion in these proposals.

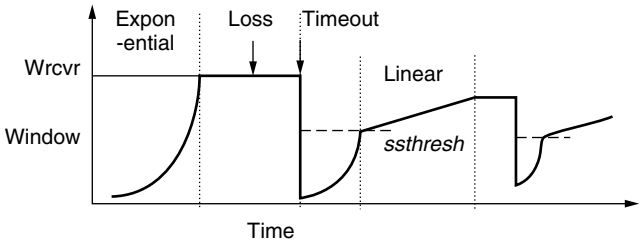


Figure 12. The TCP congestion window grows with ACKs and shrinks when packet loss is detected.

10.3. TCP-Friendly Rate Control

Clever techniques to discover the bottleneck bandwidth and control the source rate accordingly were designed in the early 1990s, including packet pair techniques [53]. Other techniques (proposed for congestion control at the application layer for both unicast and multicast) were used in the Internet Video Service [10], which uses network feedback obtained through the Real Time Control Protocol (RTCP) [74] to control the rate of sources in a video application. More recently, several researchers have investigated how applications can control their transmission *rates* (rather than *windows*) such that it approximates the behavior of TCP. This allows applications running on top of UDP (that do not require reliability) to coexist with TCP connections without starving the TCP connections. Different formulae have been developed that compute the precise "TCP-friendly" application rate. This rate is a function of the connection round trip time, and the frequency of packet loss indications perceived by the connection [66]. Example TCP-friendly protocols include the RAP protocol [71], and TCP-friendly rate control (TFRC) [31].

10.4. Explicit Congestion Indication

As discussed earlier, TCP assumes congestion when it times out waiting for an ACK, or it received duplicate or selective ACKs. This is *implicit feedback* from the network. The *explicit* congestion notification (ECN) option for TCP connections [28,70] allows active queue management mechanisms such as RED to probabilistically mark (rather than drop) packets when the average queue length lies between the two RED thresholds. This is only allowed if both the sender and receiver are ECN-capable (determined at connection setup time). In this case, the receiver echoes back to the sender the fact that some of its packets were marked, so the sender knows that the network is approaching a congested state (Fig. 13). The sender should therefore reduce its congestion window as if the packet was dropped, but need not reduce it drastically as long as it preserves TCP behavior in the long term [56]. The main advantages of ECN are that TCP does not have to wait for a timeout and some packet drops can be avoided.

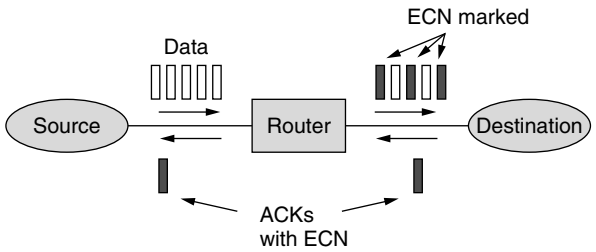


Figure 13. When congestion is incipient, AQM routers mark the ECN bit in TCP packets. The ECN bit is then marked in the ACK packets returning to the sender.

10.5. Explicit Rate Feedback

Instead of using only one bit for explicit feedback, the network can inform the sources of the precise rates they should transmit at. The ATM ABR service (designed before the ECN mechanism was proposed) allows the network to divide the available bandwidth fairly and efficiently among active sources. The ABR traffic management model is (1) "rate-based" because the sources transmit at a specified "rate," rater than using a window; (2) "closed-loop" because, unlike CBR and VBR, there is continuous feedback of control information to the source throughout the connection lifetime; and (3) "end-to-end" because control cells travel from the source to the destination and back to the source [47]. The key attractive features of the ATM ABR service are that it (1) gives sources low cell loss guarantees, (2) minimizes queuing delay, (3) provides possibly nonzero minimum rate guarantees, (4) utilizes bandwidth and buffers efficiently, and (5) gives the contending sources fair shares of the available resources.

The components of the ABR traffic management framework are shown in Fig. 14. To obtain network feedback, the sources send resource management (RM) cells every $Nrm - 1$ (Nrm is a parameter with default value 32) data cells. Destinations simply return these RM cells back to the sources. The RM cells contain the source rate, and several fields that can be used by the network to provide feedback to the sources. These fields are: the explicit rate (ER), the congestion indication (CI) flag and the no increase (NI) flag. The ER field indicates the rate that the network can support for this connection at that particular instant. The ER field is initialized at the source to a rate no greater than the PCR, and the CI and NI flags are usually reset. Each switch on the path *reduces* the ER field to the maximum rate it can support, and sets CI or NI if necessary. When a source receives a returning RM cell, it computes its allowed cell rate (ACR) using its current ACR value, the CI and NI flags, and the ER field of the RM cell [47].

Several algorithms have been developed to compute the ER feedback to be indicated by the network switches to the sources in RM cells [3,16,50,51,76]. The "explicit rate indication for congestion avoidance+" (ERICA+) algorithm [51] computes weighted max–min fair rates (with minimum guarantees) that result in high link utilization and small queuing delay in the network. The algorithm uses the measured load in the forward direction to provide feedback in the reverse direction. The rate is computed as a function of the connection load, the total load, the available capacity, the previous maximum allocation, the normalized weights, and the minimum guarantee. ABR queues can thus be pushed to the edge of a domain [34,79]. Extensions for supporting point-to-multipoint and multipoint-to-point connections are also provided [25].

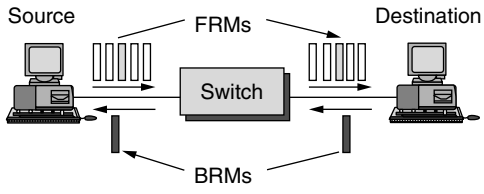


Figure 14. Resource management (RM) cells are sent every Nrm cells in the forward direction, and backward RM cells return to the source with explicit rate information.

BIOGRAPHY

Sonia Fahmy received her PhD degree at the Ohio State University in August 1999. Since then, she has been an Assistant Professor at the Computer Science Department at Purdue University. She has been very active in the Traffic Management working group of the ATM Forum, and has participated in several IETF working groups. Her work is published in over 40 journal and conference papers, and several ATM Forum contributions. She is a member of the ACM, IEEE, Phi Kappa Phi, Sigma Xi, and Upsilon Pi Epsilon, and is listed in the *International Who's Who in Information Technology*. She received the Schlumberger Foundation Technical Merit Award in 2000 and 2001. She has served on the program committees of several conferences, including IEEE INFOCOM, ICNP, and ICC; and co-chaired the SPIE Conference on Scalability and Traffic Control in IP Networks. Her research interests span several areas in the design and evaluation of network architectures and protocols. She is currently investigating multipoint communication, congestion control, and wireless networks. Please see *http://www.cs.purdue.edu/homes/fahmy/* for more information.

BIBLIOGRAPHY

1. M. Allman, V. Paxson, and W. Stevens, *TCP Congestion Control*, RFC 2581, April 1999; *http://www.ietf.org/rfc/rfc2581.txt*; see also *http://tcpsat.lerc.nasa.gov/tcpsat/papers.html.*
2. G. Apostolopoulos, R. Guerin, S. Kamat, and S. K. Tripathi, Quality of service based routing: A performance perspective, *Proc. ACM SIGCOMM*, Sept. 1998, pp. 17–28.
3. A. Arulambalam, X. Chen, and N. Ansari, Allocating fair rates for available bit rate service in ATM networks, *IEEE Commun. Mag.* **34**(11): (Nov. 1996).
4. S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin, REM: Active queue management, *IEEE Network* (May/June 2001) (*http://netlab.caltech.edu/netlab-pub/remaqm.ps.*).
5. D. Awduche et al., *Overview and Principles of Internet Traffic Engineering, Work in Progress*, Aug. 2001; *http://www.ietf.org/*.
6. D. Awduche et al., *Requirements for Traffic Engineering over MPLS*, RFC 2702, Sept. 1999; *http://www.ietf.org/rfc/rfc2702.txt.*
7. J. C. R. Bennett and H. Zhang, Hierarchical packet fair queueing algorithms, *IEEE/ACM Trans. Network.* **5**(5): 675–689 (Oct. 1997).
8. D. Bertsekas and R. Gallagher, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
9. S. Blake et al., *An Architecture for Differentiated Services*, RFC 2475, Dec. 1998; *http://www.ietf.org/rfc/rfc2475.txt.*

10. J.-C. Bolot, T. Turletti, and I. Wakeman, Scalable feedback control for multicast video distribution in the Internet, *Proc. ACM SIGCOMM*, Sept. 1994.
11. J. Boyle et al., *The COPS* (*Common Open Policy Service*) *Protocol*, RFC 2748; Jan. 2000; *http://www.ietf.org/rfc/rfc2478.txt*.
12. R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, June 1994; *http://www.ietf.org/rfc/rfc1633.txt*.
13. R. Braden et al., *Resource ReSerVation Protocol* (*RSVP*), RFC 2205, Sept. 1997; *http://www.ietf.org/rfc/rfc2205.txt*.
14. L. Brakmo, S. O'Malley, and L. Peterson, TCP vegas: New techniques for congestion detection and avoidance, *Proc. ACM SIGCOMM*, Aug. 1994, pp. 24–35; *http://netweb.usc.edu/yaxu/Vegas/Reference/vegas93.ps*.
15. L. Breslau et al., Endpoint admission control: Architectural issues and performance, *Proc. ACM SIGCOMM*, Stockholm, Sweden, Aug. 2000; *http://www.acm.org/sigcomm/sigcomm2000/conf/paper/sigcomm2000-2-2.pdf*.
16. A. Charny, D. Clark, and R. Jain, Congestion control with explicit rate indication, *Proc. ICC'95*, June 1995.
17. D. Chiu and R. Jain, Analysis of the increase/decrease algorithms for congestion avoidance in computer networks, *J. Comput. Networks ISDN Syst.* **17**(1): 1–14 (June 1989) (*http://www.cis.ohio-state.edu/~jain/papers/cong_av.htm*).
18. D. Clark, Internet cost allocation and pricing, in McKnight and Bailey, eds., *Internet Economics*, MIT Press, Cambridge, MA, 1997.
19. D. Clark and W. Fang, Explicit allocation of best effort packet delivery service, *IEEE/ACM Trans. Network.* (Aug. 1998).
20. E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, *A Framework for QoS-Based Routing in the Internet*, RFC 2386, Aug. 1998; *http://www.ietf.org/rfc/rfc2386.txt*.
21. J. Crowcroft and P. Oechslin, Differentiated end-to-end internet services using a weighted proportional fair sharing tcp, *ACM Comput. Commun. Rev.* **28**(3): (July 1998).
22. R. L. Cruz, A calculus for network delay, Part I: Network elements in isolation, *IEEE Trans. Inform. Theory* **37**(1): 114–131 (Jan. 1991).
23. R. L. Cruz, A calculus for network delay, Part II: Network analysis, *IEEE Trans. Inform. Theory* **37**(1): 132–141 (Jan. 1991).
24. A. Demers, S. Keshav, and S. Shenker, Analysis and simulation of a fair queueing algorithm, *J. Internetwork. Res. Exp.* **1**: 3–26 (1990).
25. S. Fahmy and R. Jain, ABR flow control for multipoint connections, *IEEE Network Mag.* **12**(5): (Sept./Oct. 1998).
26. K. Fall and S. Floyd, Simulation-based comparisons of Tahoe, Reno, and SACK TCP, *ACM Comput. Commun. Rev.* **26**(3): 5–21 (July 1996) (*ftp://ftp.ee.lbl.gov/papers/sacks.ps.Z.*).
27. W. Feng, D. Kandlur, D. Saha, and K. Shin, BLUE: A new class of active queue management algorithms, *Proc. NOSSDAV*, June 2001; also appears as technical report, Univ. Michigan, CSE-TR-387-99, April 1999.
28. S. Floyd, TCP and explicit congestion notification, *ACM Comput. Commun. Rev.* **24**(5): 8–23 (Oct. 1994) (*http://www.aciri.org/floyd/*).
29. S. Floyd and V. Jacobson, Random early detection gateways for congestion avoidance, *IEEE/ACM Trans. Network.* **1**(4): 397–413 (Aug. 1993) (*ftp://ftp.ee.lbl.gov/papers/early.ps.gz*).
30. S. Floyd and V. Jacobson, Link-sharing and resource management models for packet networks, *IEEE/ACM Trans. Network.* **3**(4): (Aug. 1995).
31. S. Floyd, M. Handley, J. Padhye, and J. Widmer, Equation-based congestion control for unicast applications, *Proc. ACM SIGCOMM*, Aug. 2000; multicast extension appears in SIGCOMM 2001.
32. *T*he ATM Forum, *The ATM Forum Traffic Management Specification Version 4.0*; *ftp://ftp.atmforum.com/pub/approved-specs/af-tm-0056.000.ps*, April 1996.
33. M. W. Garrett, Service architecture for ATM: from applications to scheduling, *IEEE Network.* **10**(3): 6–14 (May/June 1996).
34. R. Goyal et al., Per-vc rate allocation techniques for ATM-ABR virtual source virtual destination networks, *Proc. IEEE GLOBECOM* Nov. 1998; *http://www.cis.ohio-state.edu/~jain/papers/globecom98.htm*; see also: S. Kalyanaraman et al., Design considerations for the virtual source/virtual destination (VS/VD) feature in the ABR service of ATM networks, *J. Comput. Networks ISDN Syst.* **30**(19): 1811–1824 (Oct. 1998).
35. A. Habib, S. Fahmy, and B. Bhargava, Design and evaluation of an adaptive traffic conditioner for differentiated services networks, *Proc. IEEE ICCCN* 90–95 (Oct. 2001).
36. E. L. Hahne, Round-robin scheduling for max-min fairness in data networks, *IEEE J. Select. Areas Commun.* **9**(7): 1024–1039 (1991) (*citeseer.nj.nec.com/hahne91roundrobin.html*).
37. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, *Assured Forwarding PHB Group*, RFC 2597, June 1999; *http://www.ietf.org/rfc/rfc2597.txt*.
38. J. Hoe, Improving the start-up behavior of a congestion control scheme for TCP, *Proc. ACM SIGCOMM*, 270–280 (Aug. 1996) (*http://www.acm.org/sigcomm/ccr/archive/1996/conf/hoe.ps*).
39. C. V. Hollot, V. Misra, D. Towsley, and W.-B. Gong, On designing improved controllers for AQM routers supporting TCP flows, *Proc. IEEE INFOCOM'2001*, April 2001; *http://www.ieee-infocom.org/2001/*.
40. V. Jacobson, Congestion avoidance and control, *Proc. ACM SIGCOMM* **18**: 314–329, (Aug. 1988) (*ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z*).
41. V. Jacobson, K. Nichols, and K. Poduri, *An Expedited Forwarding PHB*, RFC 2598, June 1999; *http://www.ietf.org/rfc/rfc2598.txt*.
42. J. M. Jaffe, Bottleneck flow control, *IEEE Trans. Commun.* **COM-29**(7): 954–962 (July 1981).
43. R. Jain, A timeout-based congestion control scheme for window flow-controlled networks, *IEEE J. Select. Areas Commun.* **SAC-4**(7): 1162–1167 (Oct. 1986).
44. R. Jain, A delay-based approach for congestion avoidance in interconnected heterogeneous computer networks, *ACM Comput. Commun. Rev.* **19**(5): 56–71 (Oct. 1989).
45. R. Jain, Congestion control in computer networks: Issues and trends, *IEEE Network Mag.* 24–30 (May 1990).
46. R. Jain, Myths about congestion management in high-speed networks, *Internetwork. Res. Exp.* **3**: 101–113 (1992).

47. R. Jain et al., Source behavior for ATM ABR traffic management: An explanation, *IEEE Commun. Mag.* **34**(11): 50–57 (Nov. 1996) (*http://www.cis.ohio-state.edu/~jain/papers/src_rule.htm*).

48. R. Jain, K. K. Ramakrishnan, and D. M. Chiu, *Congestion Avoidance in Computer Networks with a Connectionless Network Layer*, Digital Equipment Corp., Technical Report DEC-TR-506, Aug. 1987; also in C. Partridge, ed., *Innovations in Internetworking*, Artech House, Norwood, MA, 1988, pp. 140–156.

49. S. Jamin, P. Danzig, S. Shenker, and L. Zhang, A measurement-based admission control algorithm for integrated services packet networks, *Proc. ACM SIGCOMM '95*, 1995, pp. 2–13.

50. L. Kalampoukas, A. Varma, and K. K. Ramakrishnan, An efficient rate allocation algorithm for ATM networks providing max-min fairness, *Proc. 6th IFIP Int. Conf. High Performance Networking*, Sept. 1995.

51. S. Kalyanaraman et al., The ERICA switch algorithm for ABR traffic management in ATM networks, *IEEE/ACM Trans. Network.* **8**(1): 87–98 (Feb. 2000) (*http://www.cis.ohio-state.edu/~jain/papers/erica.htm*).

52. F. Kelly, A. Maulloo, and D. Tan, Rate control in communication networks: Shadow prices, proportional fairness and stability, *J. Oper. Res. Soc.* **49**: 237–252 (1998).

53. S. Keshav, A control-theoretic approach to flow control, *Proc. ACM SIGCOMM '91*, 1991, pp. 3–15.

54. P. Key, Service differentiation: Congestion pricing, brokers and bandwidth futures, *Proc. NOSSDAV*, Basking Ridge, NJ, June 1999; *http://www.nossdav.org/1999/papers/75-1645029201.ps.gz*.

55. S. Kunniyur and R. Srikant, A time-scale decomposition approach to decentralized ECN marking, *Proc. IEEE INFOCOM'2001*, April 2001; *http://www.ieee-infocom.org/2001/*.

56. M. Kwon and S. Fahmy, TCP increase/decrease behavior for explicit congestion notification (ECN), *Proc. IEEE ICC*, April 2002; *http://www.cs.purdue.edu/homes/fahmy/*.

57. D. Lin and R. Morris, Dynamics of random early detection, *Proc. ACM SIGCOMM* **27**: 127–136 (Sept. 1997).

58. J. K. MacKie-Mason and H. R. Varian, *Pricing the Internet*, Dept. Economics, Univ. Michigan, Ann Arbor, 1993.

59. M. Mathis and J. Mahdavi, Forward acknowledgment: Refining TCP congestion control, *Proc. ACM SIGCOMM* (Aug. 1996) (*http://www.psc.edu/networking/papers/papers.html*).

60. M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, *TCP Selective Acknowledgement Options*, RFC 2018, Oct. 1996; *http://www.ietf.org/rfc/rfc2018.txt*.

61. K. Nichols, S. Blake, F. Baker, and D. Black, *Definition of the Differentiated Service Field (DS Field) in the IPv4 and IPv6 Headers*, RFC 2474, Dec. 1998; *http://www.ietf.org/rfc/rfc2474.txt*.

62. K. Nichols, V. Jacobson, and L. Zhang, *A Two-Bit Differentiated Services Architecture for the Internet*, RFC 2638, July 1999; *http://www.ietf.org/rfc/rfc2638.txt*.

63. A. Orda and A. Sprintson, QoS routing: The precomputation perspective, *Proc. IEEE INFOCOM*, March 2000.

64. A. Orda, Routing with end to end QoS guarantees in broadband networks, *IEEE INFOCOM'98*, April 1998.

65. Teunis J. Ott, T. V. Lakshman, and Larry H. Wong, SRED: Stabilized RED, *Proc. IEEE INFOCOM*, March 1999.

66. J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, Modeling TCP throughput: A simple model and its empirical validation, *Proc. ACM SIGCOMM* **28**: 303–314 (Sept. 1998) (*http://gaia.cs.umass.edu/*).

67. A. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The single-node case, *IEEE/ACM Trans. Network.* **1**(3): 344–357 (1993).

68. A. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The multiple node case, *IEEE/ACM Trans. Network.* **2**(2): 137–150 (1994).

69. C. Partridge, *Gigabit Networking*, Addison-Wesley, Reading, MA, 1993.

70. K. Ramakrishnan and S. Floyd, *A proposal to add explicit congestion notification (ECN) to IP*, RFC 2481, Jan. 1999; *http://www.ietf.org/rfc/rfc2481.txt*.

71. R. Rejaie, M. Handley, and D. Estrin, An end-to-end rate-based congestion control mechanism for realtime streams in the internet, *Proc. IEEE INFOCOM*, New York, March 1999; *http://www.ieee-infocom.org/1999/papers/09e_03.pdf*.

72. A. Romanow and S. Floyd, Dynamics of TCP traffic over ATM networks, *IEEE J. Select. Areas Commun.* **13**(4): 633–641 (May 1995).

73. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, Jan. 2001; *http://www.ietf.org/rfc/rfc3031.txt*.

74. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, 1996; *http://www.ietf.org/rfc/rfc1889.txt*.

75. M. Shreedhar and G. Varghese, Efficient fair queueing using deficit round robin, *Proc. ACM SIGCOMM* 231–243 (Sept. 1995).

76. K. Siu and T. Tzeng, Intelligent congestion control for ABR service in ATM networks, *Comput. Commun. Rev.* **24**(5): 81–106 (Oct. 1995).

77. I. Stoica, S. Shenker, and H. Zhang, Core-stateless fair queueing: A scalable architecture to approximate fair bandwidth allocations in high speed networks, *Proc. ACM SIGCOMM* (Sept. 1998).

78. I. Stoica and H. Zhang, Providing guaranteed service without per flow management, *ACM Comput. Commun. Rev.* **29**(4): 81–94 (Oct. 1999) (*http://redriver.cmcl.cs.cmu.edu/~hzhang-ftp/SIGCOM99.ps.gz*).

79. B. Vandalore et al., QoS and multipoint support for multimedia applications over the ATM ABR service, *IEEE Commun. Mag.* 53–57 (Jan. 1999) (*http://www.cis.ohio-state.edu/~jain/papers/multabr.htm*).

80. Z. Wang and J. Crowcroft, Quality-of-service routing for supporting multimedia applications, *IEEE J. Select. Areas Commun.* **14**(7): 1228–1234 (Sept. 1996).

81. P. P. White, RSVP and integrated services in the Internet: A tutorial, *IEEE Commun. Mag.* (May 1997).

82. H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, *Proc. IEEE* **83**(10): (Oct. 1995).

NETWORK TRAFFIC MODELING

MICHAEL DEVETSIKIOTIS
North Carolina State University
Raleigh, North Carolina

NELSON L. S. DA FONSECA
Institute of Computing
State University of Campinas
Campinas, Brazil

1. INTRODUCTION

Traffic, that is, data being transmitted, is what telecommunication networks are built to carry. The fascinating and unprecedented "Internet revolution" has led to an ever increasing need for larger amounts of data to be transmitted, as well as fast increasing expectations in terms of the diversity and quality of the transmitted data. Modern networks are expected to accommodate a very heterogeneous traffic mix, including traditional telephone calls, data services, World Wide Web browsing, and video or other multimedia information. In this context, network designers and telecommunication engineers are called on to design, control, and manage networks of increasing transmission speed (bandwidth), size, and complexity. Any effort in network design, control, or management requires decisions and optimization actions that in turn require accurate prediction of the performance of the system under design or control. This is why the science and "art" of traffic modeling has been playing a crucial role in the area of communication network design and operation [7].

The amount of traffic per unit time arriving at a network access point, the number of Internet access requests in an hour or the traffic *workload* through an Internet provider's nodes (routers or switches) is a real physical quantity, even though it consists of bits and bytes and *not* of atoms or molecules. This physical quantity is highly variable with time and space, and appears irregular or, *random*. Furthermore, network traffic usually exhibits visual clusters of activity separated by less active intervals, what is described in the telecommunications lingo, as *bursty* behavior. In order to predict the performance of networks carrying this variable and diverse traffic, researchers and telecommunication engineers utilize analysis (closed-form mathematics), numeric approximations, computer simulation, experimentation with real systems (in the laboratory or in the field), and heuristic or ad hoc projections based on past experience. All of these require, to a great or less degree, some representation or abstraction of *real-life* network traffic, that is, traffic "models."

Traffic modeling has a theoretical/analytic aspect, whereby suitable stochastic models are devised in the mathematical sense, and attributed to different types of data sources and network types. Each model has a number of parameters that determine specific aspects such as mean value, higher moments, autocorrelation function, and marginal density. Such models include [1]

- Renewal models
- Markov and semi-Markov processes
- Autoregressive processes (AR, ARMA, and ARIMA)
- Specially invented processes like Transform-Expand-Sample (TES), SRP, DAR and other
- Long-range dependent, self-similar and multifractal processes

There are also key *computational* and statistical aspects to traffic modeling: After deciding on or hypothesizing about a model (or model family) in the abstract, particular values have to be chosen for the parameters of the model. This usually means performing *matching* or *fitting* where parameter values are estimated statistically from the measured traffic data. Depending on the number of parameters involved, the type of model, and the nature of the data, this task may be far from straightforward and quite time-consuming. The moments to be estimated also depend on the type of network and traffic source, and represent an assumption in themselves. Typical traffic sources include

- Voice, very important for its dominant presence in telephone networks
- Video, especially digital, compressed video (e.g., MPEG)
- Data applications such as FTP, TELNET, SMTP, HTTP
- Traffic in local area and campus networks (LAN and MAN)
- Aggregated traffic on network "trunks" over wide-area networks (WANs)

In this article, we present the most common stochastic processes used for traffic modeling, in Section 2. Such processes can be used either to model the aggregate traffic of several sources (flows, connections, calls) on a network link or can be used to model individual sources, such as the stream generated by a phonecall. Models for specific sources are introduced in Section 3. Some special aspects of traffic modeling related to network performance, namely the concepts of *effective bandwidths* and *envelope processes* are discussed briefly in Section 4. Finally, conclusions and some current open and challenging issues are discussed in Section 5.

2. TRAFFIC MODELS

2.1. General Background

Traffic modeling starts usually by a researcher or telecom engineer collecting samples of traffic during a period of time ("traffic traces") from a specific source and/or at a specific point in the network (e.g., access point, router port, or transmission link). Before stochastic modeling is applied, care must be taken to remove *determinism* and identifying the "residual uncertainty" [16] so that what remains to be modeled is truly stochastic in nature, and *stationary* (i.e., does not have fundamental properties that change with time of the day or month). At a second step, the data are analyzed and a stochastic model is proposed so that a realization of the stochastic process matches the

data trace. A theoretical traffic model has to be checked against several data traces before one can be confident of its accuracy. In what follows, we present stochastic processes commonly used to describe traffic streams.

Network traffic can be *simple* or *compound*. Simple traffic corresponds to single arrivals of discrete data entities (e.g., "packets") and is typically described as a *point process* [9], that is, a sequence of arrival instants $T_1, T_2, \ldots, T_n, \ldots$, with $T_0 = 0$. Point processes can be described equivalently by counting processes and interarrival-time processes. A counting process $\{N(t)\}_{t=0}^{\infty}$ is a continuous-time, nonnegative integer-valued stochastic process, where $N(t)$ is the number of traffic arrivals in the interval $(0, t]$. An interarrival time process is a real-valued random sequence $\{A_n\}_{n=1}^{\infty}$, where $A_n = T_n - T_{n-1}$ is the length of the time interval separating the n-th arrival from the previous one.

Compound traffic consists of *batch arrivals*, that is, multiple units possibly arriving simultaneously at an instant T_n. In the case of compound traffic, we also need to know the real-valued random sequence $\{B_n\}_{n=1}^{\infty}$, where B_n is the (random) number of units in the batch.

In some cases, it is more appropriate or convenient to assume that time is *slotted*, which leads to *discrete-time* traffic models. This means that arrivals may take place only at integer times T_n and interarrival periods are also integer-valued. Furthermore, there are cases where the natural structure of the traffic is such that interarrival times are deterministic or periodic, with only the amount of arriving *workload* changing from arrival to arrival (e.g., compressed video "frames", arriving every $\frac{1}{30}$th of a second).

A simple way to represent a stochastic process is to give the moments of the process — particularly the first and the second moments, which are called the mean, variance, and autocovariance functions. The mean function of the process is defined by $\mu_t = E(X_t)$. The variance function of the process is defined by $\sigma_t^2 = E[(x_t - \mu_t)^2]$, and the autocovariance function between X_{t_1} and X_{t_2} is defined by $\gamma(t_1, t_2) = E[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})]$.

Another topic that is very relevant in traffic modeling is that of traffic "burstiness." Burstiness is present in a traffic process if the interarrival times process $\{A_n\}$ tends to give rise to runs for several short interarrival times followed by relatively long ones. With typical network traffic exhibiting patterns and bursts that coexist over many magnitudes of time scales (from minutes to hours to days) come the notion of timescale invariance. *Timescale* refers to the change or immunity to change of the process structure on scaling of the time axis. A process $\{X_t\}$ can be defined as scaling invariant if for some $\alpha \in [a, b]$ the process is equal in distribution to its scaled version $\{X_{\alpha t}\}$. If a traffic is not scale-invariant then when studying its behavior as time scales increase, it will show that the bursts and random fluctuations degenerate toward a white noise, nonbursty type of traffic.

The marginal distribution of a process $\{X_t\}$ captures the steady-state first-order distribution of X and is considered the primary characteristic in describing network traffic. Assuming that the process is wide-sense stationary (WSS), the marginal distribution becomes invariant to time and is then defined by the one-dimensional probability density function (PDF): $f_X(x) = f_{X_t}(x) = \frac{d}{dx} Pr[X_t \leq x]$. The PDF describes the probability that the data will assume a value within some given range at any instant of time.

The autocorrelation function of a process $\{X_t\}$ captures the second order measurement of the process and it is used as a supplement to the marginal distribution. The autocorrelation function for network traffic describes the general dependence of the values at another time. Assuming the process is WSS, then the autocorrelation between the data values at times t and $t + k$ is defined as follows:

$$\rho(k) = \frac{E[X_t X_{t+k}] - (E[X_t])^2}{E[(X_t - E[X_t])^2]}$$

where k is called the "lag," the difference or distance between timepoints under consideration. If the autocorrelation function $\rho(k)$ of $\{X_k\}$ is equal to zero for all values of $k \neq 0$, then $\{X_k\}$ is of the *renewal* type. Markov and other *short-range dependent* (SRD) models have a correlation structure that is characterized by an *exponential* decay, which leads to $\sum_k \rho(k) < \infty$.

On the other hand, many real traffic traces exhibit *long-range dependence* (LRD) and can be modeled by self-similar and multifractal models later in this article. For these processes, the autocorrelation function decays slowly (say, polynomially instead of exponentially) in a way that makes the autocorrelation nonsummable: $\sum_k \rho(k) \to \infty$ [23].

2.2. Short-Range Dependent Models

2.2.1. Renewal Models. Renewal models have been used for a long time because of their simplicity and tractability. For this type of traffic, the interarrival times are independent and identically distributed (i.i.d.), with an arbitrary distribution. The major modeling drawback of renewal processes is that the autocorrelation function of A_n is *zero* except for lag $n = 0$. Hence, renewal models seldom capture the behavior of high-speed network traffic in an accurate manner.

Within the renewal family, *Poisson* models are the oldest and most widely used, having been historically closely linked to traditional telephony and the work of A. K. Erlang. A Poisson process is a renewal process with *exponentially* distributed interarrival times with rate λ: $P[A_n \leq t] = 1 - e^{-\lambda t}$. It is also a counting process with $P[N(t) = n] = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$, and independent numbers of arrivals in disjoint intervals. Poisson processes are very appealing due to their attractive memoryless and aggregation properties.

2.2.2. Markov Models. Unlike renewal traffic models, Markov and Markov renewal traffic models [1,7] introduce dependence into the random sequence A_n. Consequently, they can potentially capture traffic burstiness, due to nonzero autocorrelations of A_n. Consider a Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with a discrete state space, where M behaves as follows. It stays in state i for an exponentially distributed holding time that depends on i alone; it then jumps to state j with probability p_{ij}, such that the matrix

$P = [p_{ij}]$ is a probability matrix. In a simple Markov traffic model, each jump of the Markov process corresponds to an arrival, so interarrival times are exponentially distributed, and their rate parameter depends on the state from which the jump occurred. Arrivals may be single, a batch of units or a continuous quantity.

Markov-modulated models constitute another important class of traffic models. Let $M = \{M(t)\}_{t=0}^{\infty}$ be a continuous-time Markov process, with state space of $1, 2, \ldots, m$. Now assume that while M is in state k, the probability law of traffic arrivals is completely determined by k. Thus, the probability law for arrivals is *modulated* by the state of M. The modulating process can be more complicated than a Markov process (so the holding times need not be restricted to exponential random variables), but such models are far less analytically tractable.

The most commonly used Markov modulated model is the *Markov modulated Poisson process* (MMPP) model, which combines a modulating (Markov) process with a modulated Poisson process. In this case, while in state k of M, arrivals occur according to a Poisson process of rate k. As a simple example, consider a 2-state MMPP model, where one state is an ON state with a positive Poisson rate, and the other is an OFF state with a rate of zero. Such models have been widely used to model voice traffic sources.

A semi-Markov process is a generalization of Markov processes, that allows the holding time to follow an arbitrary probability distribution. This destroys the Markov property since times are not exponentially distributed, however it allows for more general models of traffic. When values from a semi-Markov chain are generated, the next state is chosen first, followed by a value for the holding time. If the holding times are ignored, then the sequence of states will be a discrete time Markov chain, referred to as an *embedded* Markov chain.

2.2.3. Autoregressive Models. The autoregressive model of order p, AR(p), is a process $\{X_t\}$ whose current value is expressed as a finite linear combination of previous values of the process plus a white-noise process ε_t: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$ where ϕ_i are constants and X_{t-i} are past values of the process at time $t - i$. The recursive form of this model makes it a popular modeling candidate as it makes it straightforward to *generate* an autocorrelated traffic sequence, such as variable-bit-rate (VBR) video traffic [1,7]. However, autoregressive models cannot simultaneously match the empirical marginal distribution of arbitrary traffic such as video.

Another model of the same family is the autoregressive moving-average model of order (p, q), denoted by ARMA (p, q): $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$. Because of the larger number of parameters, ARMA models are more flexible than AR models and can be used in more cases. However, estimation of its parameters is more involved.

2.2.4. Transform–Expand–Sample (TES). Transform–expand–sample (TES) models represent another important class of models appropriate for modeling autocorrelated traffic streams. This family of models aims to capture *both* autocorrelation and *marginal distribution* of the empirical traffic trace; in fact, it was historically the first traffic model explicitly devised to accomplish exactly this dual purpose and specifically for network traffic data. TES models capture stationary, correlated time series and also allow one to generate synthetic streams of real-looking traffic streams to drive simulations of networks [7].

TES models include two types of TES processes: TES$^+$ and TES$^-$. TES$^+$ produces sequences with positive autocorrelation at lag 1, while TES$^-$ produces negative autocorrelation at lag 1. The TES$^+$ process is more suitable for modeling network traffic. To define the TES$^+$ process, we first introduce a modulo 1 operation. The modulo 1 of a real number x, denoted by $\langle x \rangle$, is defined as $\langle x \rangle = x - \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the maximum integer less than x. The recursive construction of the background TES$^+$ process is defined by

$$U_n^+ = \begin{cases} U_0^+ & n = 0 \\ \langle U_{n-1}^+ + V_n \rangle & n > 0 \end{cases}$$

where $\{V_n\}$ is a sequence of IID random variables referred to as *innovations* and U_0^+ is uniformly distributed on $[0, 1)$ and independent of $\{V_n\}$. The resulting sequence $\{U_n^+\}$ has a $[0, 1)$ uniform marginal distribution, and autocorrelation function determined by the probability density function $f_V(t)$ of V_n. The choice of $f_V(t)$ determines the correlation structure of the resulting process. From this background sequence the output process of the model referred to as the foreground sequence, $\{X_n^+\}$ is created by "distorting" each U_n^+ by $X_n^+ = F^{-1}(U_n^+)$, where F is the marginal distribution of the empirical data [9].

2.2.5. Other Short-Range Dependent Models. Another interesting model is the *spatial renewal process* (SRP), which efficiently models processes exhibiting arbitrary marginal distribution and aperiodically decaying autocorrelation (see the paper by Taralp et al. [20] and references cited therein).

A *discrete autoregressive model* of order p, denoted as DAR(p), generates a stationary sequence of discrete random variables with an arbitrary probability distribution and with an autocorrelation structure similar to that of an AR(p). DAR(1) is a special case of DAR(p) process; it has a smaller number of parameters than do general Markov chains, simpler parameter estimation, and can match arbitrary distributions. Moreover, the analytic queuing performance is tractable (see paper by Adas [1] and references cited therein).

2.3. Long-Range Dependent and Self-Similar Traffic Models

Measurements and statistical analysis of real traces performed during the 1990s revealed that traffic exhibits large irregularities (*burstiness*) both in terms of extreme variability of traffic intensities as well as persistent autocorrelation. Network traffic often looks extremely irregular at different timescales [12,17], and such extreme behavior is not exhibited by the traditional Poisson traffic, which smoothes out when aggregated at coarser timescales. If traffic were to follow a Poisson or Markov arrival process, it would have a characteristic burst length that would tend to be smoothed by averaging over a long enough timescale. Instead, measurements of real traffic indicate consistently

that significant traffic burstiness is present on a wide range of timescales.

This behavior is reminiscent of and has been modeled according to *self-similar* processes. Self-similar or *fractal* modeling has been used in a number of research areas such as hydrology, financial mathematics, telecommunications, and chaotic dynamics [4,24]. Internet traffic, and more generally broadband network traffic, is an area where fractal modeling has become popular more recently. Such modeling has also been related to the observation of ON–OFF traffic with "heavy-tailed" distribution [23].

2.3.1. Heavy-Tailed ON–OFF Models. The fractal nature of network traffic is consistent with and predicted by the behavior of the individual connections that produce the aggregate traffic stream. In WAN traffic, individual connections correspond to "sessions," where a session starts at a random point in time, generates packets or bytes for some time and then stops transmitting. On the other hand, in LAN traffic, individual connections correspond to an individual source–destination pair. Individual connections are generally described using simple traffic models such as ON–OFF sources.

Traditional ON–OFF models assume finite variance distributions for the duration of the ON and the OFF periods. The aggregation of a large number of such processes results in processes with very small correlations. On the other hand, a positive random variable Y is called "heavy-tailed with tail index α," if it satisfies: $P[Y > y] = 1 - F(y) \approx cy^{-\alpha}, y \to \infty, 0 < \alpha < 2$, where $C > 0$ is a finite constant independent of y. This distribution has infinite variance. Furthermore, if $1 < \alpha < 2$, then it has a finite mean. The superposition of many such sources was shown to produce aggregate traffic that exhibits long-range dependence and even self-similarity [12,22].

2.3.2. Monofractal Models. Self-similarity in a process indicates that some aspect of the process is *invariant* under scale-changing transformations, such as "zooming" in or out. In network traffic, this is observed when traffic becomes bursty, exhibiting significant variability, on many or all timescales. The appeal and modeling convenience of self-similar processes lies in the fact that the degree of self-similarity of a series can be expressed using only one parameter. The *Hurst* parameter, H, describes the speed of decay of the series autocorrelation function. For self-similar series the value of H is between 0.5 and 1. The degree of self-similarity increases as the Hurst parameter approaches unity.

A process $\{X_k\}$ whose autocorrelation function, $\rho(k)$, takes the form $\rho(k) \approx ck^{-\beta}, 0 < \beta < 1$, for large k and a constant $c > 0$, is said to be *long-range dependent*. This implies that the autocorrelation function decays slowly and is not summable, thus $\sum_k \rho(k) \to \infty$. Figure 1 shows the autocorrelation as a function of time for streams with different H values. Note that for streams with greater H the autocorrelation decays more slowly as a function of time.

In the case of traffic traces, self-similarity is used in the distributional sense: when viewed at varying

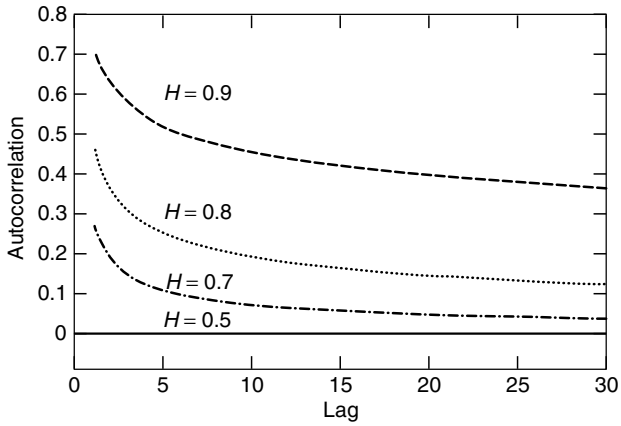


Figure 1. The autocorrelation as a function of time for different values of H.

scales, the object's distribution remains unchanged. Equivalence in distribution between X and Y is denoted by $X \stackrel{d}{=} Y$. We provide in the following certain common definitions of self-similar traffic processes, following Tsybakov and Georganas [22]. Let $X = \{X_t : t = 1, 2, 3, \ldots\}$ be a second-order stationary sequence with mean $\mu = E[X_t]$, variance $\sigma^2 = \text{var}(X_t)$, and autocorrelation function $r(k) = \dfrac{E[(X_{t+k} - \mu)(X_t - \mu)]}{\sigma^2}$. Let $X^{(m)}(t) = \frac{1}{m}(X_{tm-m+1} + \cdots + X_{tm}), m = 1, 2, 3, \ldots$, be the corresponding aggregated sequence with level of aggregation m, obtained by dividing the original sequence X into nonoverlapping blocks of size m and averaging over each block. The index t labels the block. For each $m = 1, 2, 3, \ldots$ let $\mathbf{X}^{(m)} = \{X^{(m)}(k) : k = 1, 2, 3, \ldots\}$ denote the averaged process with autocorrelation function $r^{(m)}(k)$.

- A process X is called *exactly second-order self-similar* with parameter $H = 1 - (\frac{\beta}{2}), 0 < \beta < 1$ if its correlation coefficient is $r(k) = \frac{1}{2}[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}], k = 1, 2, 3, \ldots$.
- A strict-sense stationary process X is called *strictly self-similar* with parameter $H = 1 - (\frac{\beta}{2}), 0 < \beta < 1$, if $X \stackrel{d}{=} m^{1-H}X^{(m)}$. If X is strictly self-similar, then it is also exactly second-order self-similar. The opposite is not true, except for Gaussian processes.
- A process X is called asymptotically second-order self-similar with parameter $H = 1 - (\frac{\beta}{2}), 0 < \beta < 1$, if $\lim_{m\to\infty} r^{(m)}(k) = \frac{1}{2}\delta^2(k^{2-\beta}), k = 1, 2, 3, \ldots$ where $\delta^2(f(x)) = f(x + \frac{1}{2}) - f(x - \frac{1}{2})$.
- A strict-sense stationary process X is called strictly asymptotically self-similar if $X^{(m)} \stackrel{d}{=} X, m \to \infty$. Note that a strictly asymptotically self-similar process is not necessarily asymptotically second-order self-similar.

2.3.3. Fractal Gaussian Noise and Fractal Brownian Motion. The fractal Brownian motion (FBM) is a self-similar process with Gaussian stationary increments [14]. The increment process is called *fractal Gaussian noise*, and its autocorrelation function is invariant under aggregation and is given by $r(k) = 1/2[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}]$.

The FBM process accurately models Ethernet, ATM, and FDDI traffic, as well as video sources. The aggregate of ON–OFF sources with heavy tails tends to an FBM.

The analysis of a queuing system with FBM input is quite challenging. However, it becomes manageable if the fractal Brownian traffic [15] process is used instead. The fractal Brownian traffic is defined as the fluid input in time interval $(s, t]$, and is given by $A(s,t) = m(t-s) + \sigma(Z_t - Z_s)$ where m is the mean input rate, σ^2 is the variance of traffic in a given time unit, and Z_x is a normalized fractal Brownian motion, defined as a centered Gaussian process with stationary increments and variance $E[Z_t^2] = t^{2H}$.

2.3.4. Distorted Gaussian. The distorted Gaussian (DGauss) model begins with a Gaussian process with a given autocorrelation structure and maps it into an appropriate marginal distribution. Examples of this popular traffic generation technique include the autoregressive-to-anything process [20] and the self-similar traffic model [8].

Many techniques exist to generate Gaussian time series (Gaussian in the marginal distribution) with a wide range of autocorrelation decay characteristics. A background Gaussian process Z_k is imparted with an autocorrelation structure $\rho'(t)$ and is run through a fitting function $X_k = F_X^{-1}(F_N(Z_k))$ to map its values into an appropriate distribution. Because of the background–foreground transformations, precompensation is applied to the background autocorrelation ρ' such that the resulting output autocorrelation ρ matches the desired specification [8].

2.3.5. Fractal Lévy Motion. Laskin et al. [11] introduced a teletraffic model that takes into account, in addition to the Hurst parameter $H \in [\frac{1}{2}, 1)$, the Lévy parameter $\alpha \in (1,2]$. This was the so-called *fractional Lévy motion* (fLm), mentioned by Mandelbrot [14]. Two important subclasses of Lévy motion exist: (1) the well-known ordinary Lévy motion (oLm), an α-stable process (distributed in the sense of P. Lévy) with independent increments, which is a generalization of the ordinary Brownian motion (the Wiener process); and (2) the fractional Lévy motion, a self-similar and stable distributed process, which generalizes the fractional Brownian motion (fBm), has stationary increments and an infinite "span of interdependence."

Several self-similar stable motions have been proposed for traffic modeling. These processes combine, in a natural way, both scaling behavior and extreme local irregularity.

2.4. Multifractal Models

Historically following self-similar models, researchers have been studying also the possibility of modeling network traffic with *multifractal* processes (see book by Park and Willinger [16] and references cited within). It appears that even though measured network traffic is consistent with asymptotic self-similarity, it also exhibits small timescaling features that differ from those observed over larger time scale. This small timescaling behavior has been related to communication protocol-specific mechanisms and end-to-end congestion control algorithms that operate at those small timescales (less than a few hundred milliseconds). Modeling network traffic with multifractals has the potential of capturing the observed scaling phenomena at large as well as small timescales and thus to naturally extend and improve the original self-similar models of measured traffic.

To quantify the local variations of traffic at a particular point in time t_0, let $Y = \{Y(t), 0 < t < 1\}$ denote the traffic rate process representing the total number of packets or bytes sent over a link in an interval $[t_0, t_0 + t]$. The traffic has a *local scaling component* $\alpha(t_0)$ at time t_0 if the traffic rate process behaves like $t^{\alpha(t_0)}$ as $t \to 0$. In this context, $\alpha(t_0) > 1$ relates to instants with low intensity levels or small local variations, and $\alpha(t_0) < 1$ is found in regions with high level of burstiness or local irregularities.

If $\alpha(t_0)$ is constant for all t_0, then the traffic is *monofractal*. Equivalently, if $\alpha(t_0) = H$ for all t_0, then the traffic is exactly self-similar, with Hurst parameter H. On the other hand, if $\alpha(t_0)$ is not constant and varies with time, the traffic is *multifractal*.

The multifractal appearance of WAN traffic is attributed to the existence of certain multiplicative mechanisms in the background. Multifractal processes are well modeled using multiplicative processes or "conservative cascades." The latter are a fragmentation mechanism, which preserves the mass of the initial set (or does so in the expected value sense). The generator of the cascade is called the fragmentation rule and the mathematical construct that describes the way mass is being redistributed is called the limiting object or multifractal. Modern data networks together with their protocols and controls can be viewed as specifying the mechanisms and rules of a process that fragments units of information at one layer in the networking hierarchy into smaller units at the next layer, and so on.

Multifractal processes are a generalization of self-similar processes. Hence, self-similar processes are also multifractal, but the reverse is not always true. This leads to the important modeling question: Which of the two types of models is more appropriate in a given case? A method for distinguishing between the two models has been proposed [19]. Their conclusion was that traffic traces from environments were well modeled using self-similar models and that more sophisticated models such as multifractals were not needed. On the other hand, in WAN environments, there were cases where self-similar models were not deemed adequate and where multifractal models appeared to be more appropriate.

2.5. Fluid Traffic Models

In fluid traffic modeling, individual units such as packets, are not explicitly modeled. Instead, traffic is viewed as a "stream of fluid" arriving at a certain *rate* that may be changing. Fluid models can simplify analysis due to their lower "resolution" or level of detail. More importantly, fluid models can make network simulation much more efficient, since the computer representation of the fluid traffic requires much fewer "events" (e.g., rate changes) that need to be tracked.

In modern high-speed networks such as asynchronous transfer mode (ATM) networks, the size of individual packets is often fixed and very small (e.g., 53 bytes), relative to the total transmission speed and aggregate volume of

information being transmitted (e.g., hundreds of megabits or gigabits per second). Therefore fluid modeling may be appropriate in such cases and, in general, whenever individual packets can be regarded as effectively insignificant with respect to the total traffic. The validity of this approximation depends heavily on the timescale involved as well as the point of interest inside the network (e.g., access points versus large routers in the middle of the network).

Fluid models [9] typically assume that sources are bursty, commonly of the ON–OFF type. In the OFF state, there is *no* traffic arriving, while in the ON state traffic arrives at a constant rate. To maintain analytic tractability, the durations of ON an OFF periods are assumed exponentially distributed and mutually independent (i.e., they form an alternating renewal process).

3. SOURCE MODELS

In this section, the modeling of different type of traffic sources is discussed. The flow generated by some network sources are regulated by the stack of protocols used in the network. Such type of sources is called *elastic sources*. Sources whose flow do not depend on network protocol are called *streaming sources*. First, streaming multimedia sources are introduced, followed by the modeling of elastic sources. This section concludes with a general characterization of traffic streams, called *effective bandwidth*.

3.1. Data

Datastreams were traditionally modeled by Poisson processes. The rationality behind it was that the superposition of several independent renewal processes tends to a Poisson process.

The nature of traffic changes as new applications becomes a significant part of the network traffic. SMTP, email, TELNET, and FTP were responsible for most of the traffic in pre-Web time. As the use of Web services became predominant, Internet traffic began to present new patterns. Most of today is network traffic is based on the Transmission Control Protocol (TCP). Internet traffic observed at long timescales exhibits self-similarity. However, at long timescales, typically shorter than a round–trip time, Internet traffic presents high variability. At short timescales, Internet traffic marginal distribution is non-Gaussian and the scaling exponent of the variance is smaller than the asymptotic exponent. In other words, at short timescales, Internet traffic exhibits multifractal scaling, with different moments of the traffic showing scaling described by distinct exponents. However, at long timescales it can be modeled as self-similar [4,5]. Such behavior is originated by the complex interaction between network protocols that governs the network flow and TCP sources.

3.2. Voice

The packetstream from a voice source can be characterized by an ON–OFF model; Thus, during silent periods no packet is generated and during "talkspurt" periods, packets are generated either at exponentially distributed intervals or at constant intervals depending whether compression algorithms are used. The residence time in each state is exponentially distributed.

A popular approach to analyze a multiplexer fed by several ON–OFF sources is to use Markov modulated processes to mimic the superposition process. The arrival rates and the transition probabilities of the underlying Markov chain are defined in a way that certain statistics of the Markov modulated process have the same numerical value of the corresponding statistics of the superposition process. The advantage of adopting a 2-state process is to keep the complexity of both the matching procedure and the queuing solution low. In a 2-state MMPP (Fig. 2) there are only four parameters to be determined: the arrival rate and the sojourn time in each state. Several procedures are available to set these four parameters. Most of procedures consider 2 superstates: the underloaded and the overloaded states [18]. In the overload state, the packet generation rate (due to the number of source in state ON)

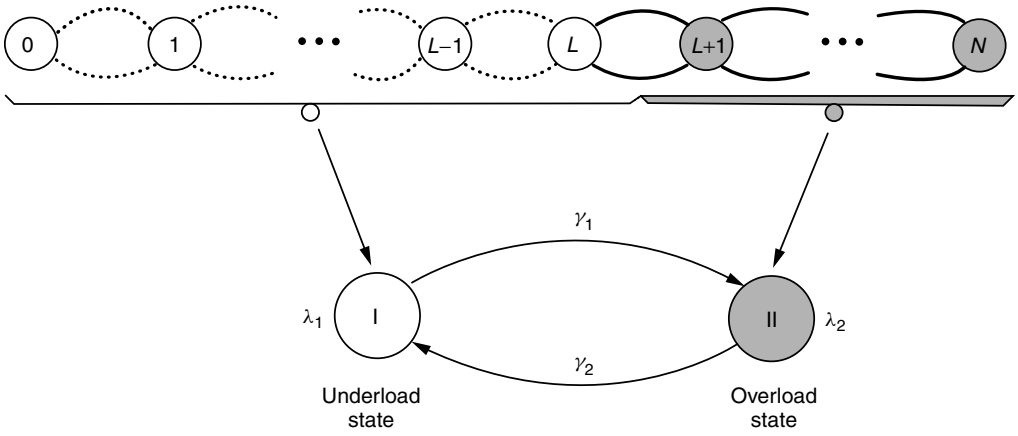


Figure 2. Modeling the superposition of voice sources as a 2-state MMPP. The states of the original Markov chain represent the number of sources in state ON and the arrival rate, in the nth state is n times the arrival rate in state ON. The superstates of the 2-state MMPP correspond to underload overload periods depending on whether the aggregated arrival rate surpasses the channel capacity.

exceeds the server capacity, whereas in the underload state it is below the server capacity.

3.3. Video

The bit rate of a videostream depends not only on the coding algorithm but also on the level of activity of a scene. Whenever there is a scene change, a new scene has to be encoded, generating a high number of bits to be transmitted, and, consequently high bit rates.

The MPEG coding scheme is widely used for several types of applications. MPEG streams consist of a series of frames. In MPEG-2, there are three types of frames: intracoded (I), predictive (P), and bidirectional (B). A periodic sequence of I, B, P frames is called *group of pictures* (GOP). An MPEG transmission consists of one GOP after the other. Typically I frames will have more bits than P and B frames, and B frames will have the least number of bits. The size of I frames can be approximated by a Normal distribution, whereas the size of B and P frames can be approximated either by a gamma or by a lognormal distribution. Figure 3 illustrates the bit rate profile of a typical videostream. The high peaks correspond to scene changes, whereas the low peaks correspond to the activity within a scene.

Video traffic exhibits long-range dependencies [2,8]. The repetitive pattern of GOPs introduces strong periodic components in the autocorrelation function (ACF). Video streams are usually modeled either by a fractal Brownian motion process or by a fractal ARIMA (0,d,0) process, which are LRD processes. However, some researchers advocate that, for finite buffer, long-term correlation have minor impact on queuing performance, and, therefore, Markovian models should be used, since only short term correlations impact the performance. The discrete first-order autoregressive model, DAR(1), is a popular Markovian process used for video modeling. Actually, Markovian models give rise to ACF of the form $\rho(k) \sim e^{-\beta k} (\beta > 0)$, whereas an LRD process exhibits ACF of the form $\rho(k) \sim k^{-\beta} = e^{-\beta \log^k} (\beta > 0)$ [10]. In fact, the performance of fractal models may be overly sensitive to the buffer size, and, consequently, may underestimate the actual performance. On the other hand, Markovian models provide good performance under heavy loads; however, they perform poorly under light loads [10].

3.4. Elastic Sources

The amount of data an application can pump into the network is often regulated by the network protocols and their congestion control mechanisms, which probe the available bandwidth to determine the amount of data that can be transmitted. Traffic sources whose transmission rate depend on network congestion status are called *elastic sources*. Examples of elastic sources are the available bit rate service (ABR) in ATM networks and the Transmission Control Protocol (TCP), largely deployed in the Internet.

TCP congestion control mechanism is a window-based one. Segments, or "packets" in TCP language, are transmitted and acknowledgments from the receiver are expected. Each segment has a sequence number, set at the sending end. Acknowledgments specify the sequence number of the acknowledged segment. Acknowledgments are cumulative; an acknowledgment notifies the transmitter that all the segments with a lower sequence number were properly received. The time from sending a packet to receiving its acknowledgment is called round–trip time (RTT). TCP controls a connection rate by limiting the number of transmitted-but-yet-to-be-acknowledged segments. In the beginning, the window size is set to one. Every time an acknowledgment is received, that is, at every RTT, the window size is doubled, and the window grows up to a threshold. After this threshold, the window is incremented by one segment.

Whenever an acknowledgment fails to arrive after a predefined interval, a timeout event occurs and the threshold is set to one-half the current congestion window and the congestion window is set to one. If the transmitter receives three consecutive acknowledgments for the same segment, it is assumed that the next segment was lost and the window is set to one-half its current value.

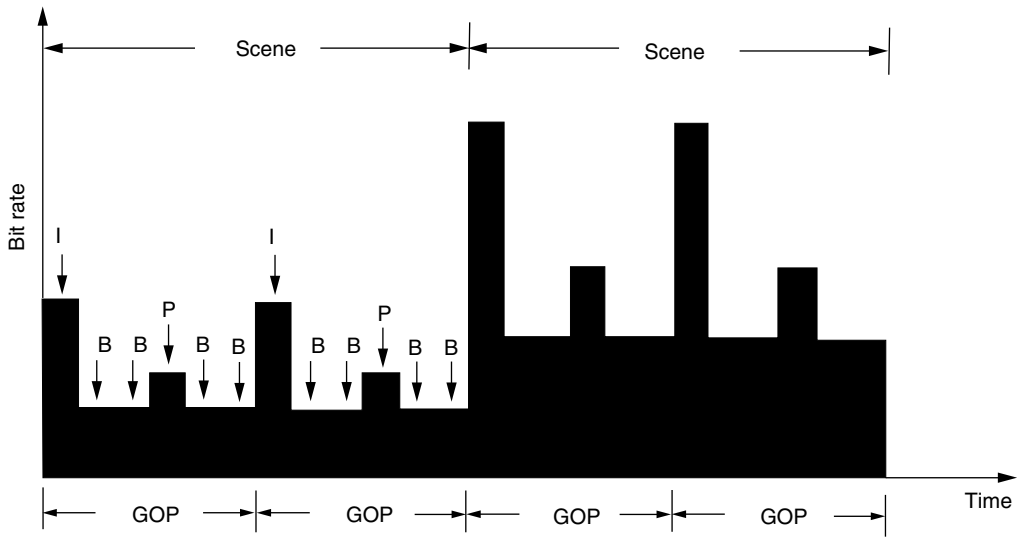


Figure 3. Bit rate of a video transmission.

The evolution of the window size between loss events can be analyzed in order to determine a TCP connection throughput, that is, the amount of data a TCP connection pumps into the network per unit of time by observing the window size evolution between loss events. The distribution of interloss periods as well as the distribution of the type of the loss event should be factored into this computation.

4. EFFECTIVE BANDWIDTHS AND ENVELOPE PROCESSES

Most communications services are subject to performance constraints designed to guarantee a minimal quality of service (QoS). Consider a general traffic stream offered to a deterministic server, and assume that some prescribed parameterized performance constraints are required to hold. The effective bandwidth of the traffic stream corresponds to the minimal deterministic service rate, required to meet these constraints. Queuing-oriented performance constraints include bounds on such statistics as queuing delay quantiles or averages, server utilization, and overflow probabilities. The effective-bandwidth concept serves as a compromise between two alternative bandwidth allocation schemes, representing a pessimistic and an optimistic outlook. The strict one allocates bandwidth based on the stream peak rate, seeking to eliminate losses, whereas the lenient one allocates bandwidth based on the stream average rate, merely seeking to guarantee stability.

Let us formally define effective bandwidth, Let $X[0,t]$ be the workload that arrived during time interval $[0,t]$ for a traffic stream. Effective bandwidth of the traffic stream is defined as $\alpha(s) = \lim_{t\to\infty} \frac{1}{st} \log E(e^{sX[0,t]})$, which is a function of $s > 0$, the so-called space parameter. In the effective bandwidth theory, s is the asymptotic exponential decay rate of queue size distribution tail probability with respect to queue size; that is, when the service rate is $\alpha(s)$, the queue size distribution tail probability with respect to queue size is $P(Q > B) \approx \exp(-sB)$, where Q denotes the queue size. This is why s is called the *space parameter*.

The notion of effective bandwidth provides a useful tool for studying resource requirements of telecommunications services and the impact of different management schemes on network performance. Estimates of effective bandwidths are called *empirical effective bandwidths*, while the analytic form are called *analytic effective bandwidths*.

4.1. Envelope Processes

An envelope process is a function which provides a bound for the amount of work generated in a traffic stream during a certain time interval. If $A(t_2 - t_1)$ is the amount of bits generated during the interval $t_2 - t_1$, then $\hat{A}(t)$ is an envelope process for $A(t)$ if and only if $\hat{A}(t_2 - t_1) > A(t_2 - t_1)$, for any $t_2 > t_1$. For any traffic stream, $A(t)$, there is a whole family of possible envelope processes; however, the lowest bound is the one of interest. Envelope processes are useful tools since, in general, they require a small number of parameters — that is, they are a *parsimonious* way of representing a stochastic process. However, dimensioning based on envelope processes may overestimate the required resources. Moreover, envelope processes are not appropriate for the study of phenomena at the cell timescale, such as cell discarding.

Network services can be either deterministic or statistical. Accordingly, deterministic and stochastic envelope processes are defined. A deterministic envelope process is a strict upper bound on the amount of work arriving during an interval for a traffic stream. A commonly used envelope process is $\int_0^t A(t) < \rho t + \sigma$, where ρ is the source mean arrival rate and σ is the maximum amount of work allowed in a burst [3]. $\rho t + \sigma$ is a model for the output of a leaky bucket regulator where ρ is the leaky rate and σ the bucket size.

In a stochastic envelope process, the amount of work generated in a certain interval may surpass a deterministic bound with a certain probability value. An accurate stochastic envelope process for a fractal Brownian motion process is $\rho t + k\sigma t^H$, where ρ is the mean arrival rate, σ is the standard deviation, and H is the Hurst parameter [6]. Note that the amount of work is not a linear function of time, it has a t^H which takes into account long periods of arrivals.

5. CONCLUSIONS

The aim of traffic modeling is to provide network designers with simple means to predict the network load, and consequently, the network performance. Since the early days of telephony networking, engineers have been engaged in understanding the nature of network traffic, and its impact on quality of service provisioning. Traffic models mimic the traffic patterns observed in real networks. The suitability of a traffic model is related to the degree of accuracy of the conclusions that can be drawn from studies using such a model. Therefore, there is no unique model for a certain type of traffic, but models with different degrees of accuracy.

With the advent of integrated networks, Poisson models for traffic streams were replaced by more sophisticated short range dependent models which considered the correlation pattern besides the mean arrival rate. By 1993, the seminal work of Leland et al. [12] demonstrated the fractal nature of LAN traffic. Several other works followed showing that other types of traffic such as video traffic were also fractal. Recent studies have shown that Internet traffic is not precisely fractal at small timescales, but can be represented well as fractal at larger timescales. The understanding of the impact of multifractality on network performance is still an open problem.

Traffic patterns are influenced by several factors such as the nature of file size, human think time, protocol fragmentation, and congestion control mechanisms. New challenging problems in traffic modeling will certainly exist when multimedia applications become a significant part of the whole network traffic.

BIOGRAPHIES

Mihail (Mike) Devetsikiotis was born in Thessaloniki, Greece. He received the Dipl. Ing. degree in Electrical Engineering from the Aristotle University of Thessaloniki, Greece, in 1988, and the M.Sc. and Ph.D. degrees in

Electrical Engineering from North Carolina State University, Raleigh, in 1990 and 1993, respectively. As a student, he received scholarships from the National Scholarship Foundation of Greece, the National Technical Chamber of Greece, and the Phi Kappa Phi Academic Achievement Award for a Doctoral Candidate at North Carolina State University. He is a member of the IEEE, INFORMS, and the honor societies of Eta Kappa Nu, Sigma Xi, and Phi Kappa Phi. In 1993 he joined the Broadband Networks Laboratory at Carleton University, as a Post-Doctoral Fellow and Research Associate. He later became an Adjunct Professor in the Department of Systems and Computer Engineering at Carleton University in 1995, an Assistant Professor in 1996, and an Associate Professor in 1999. Dr. Devetsikiotis joined the Department of Electrical and Computer Engineering at North Carolina State University as an Associate Professor, in 2000. He has served as an officer of the IEEE Communications Society Technical Committee on Communication Systems Integration and Modeling, and as Associate Editor of the journal *ACM Transactions on Modeling and Computer Simulation*.

Nelson Fonseca received his Electrical Engineer (1984) and M.Sc. in Computer Science (1987) degrees from The Pontificial Catholic University at Rio de Janeiro, Brazil, and the M.Sc. (1993) and Ph.D. (1994) degrees in Computer Engineering from The University of Southern California in Los Angeles. Since 1995 he has been affiliated to the Institute of Computing of the State University of Campinas, Brazil, where is currently an Associate Professor.

He is the recipient of Elsevier Editor of the Year 2000, of the 1994 USC International Book award, and of the Brazilian Computing Society First Thesis and Dissertations award. Mr. Fonseca is listed in Marqui's *Who's Who in the World* and *Who's Who in Science and Engineering*.

He served as Editor-in-Chief for the *IEEE Global Communications Newsletter* (1999–2002). He is an Editor for *Computer Networks*, an Editor for the *IEEE Transactions on Multimedia*, an Associate Technical Editor for the *IEEE Communications Magazine*, and an Editor for the *Brazilian Journal on Telecommunications*.

Dr. Fonseca was the chairman of the 7th IEEE Workshop on Computer-Aided Modeling, Analysis and Design of Communications Networks and Links (CAMAD'98), Vice-Chairman of IEEE GLOBECOM99 Symposium on Multimedia Services and Technology Issues, Vice-Chairman of CAMAD'2000, and Vice-Chairman of the International Teletraffic Congress'17, 2001.

BIBLIOGRAPHY

1. A. Adas, Traffic models in broadband networks, *IEEE Commun. Mag.* (July 1997).
2. J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, Variable-bit-rate video traffic and long range dependence, *IEEE Trans. Commun.* **43**(2–4): 1566–1579 (1995).
3. R. L. Cruz, A calculus for network delay, part I: Network elements in isolation, *IEEE Trans. Inform. Theory* **37**: 114–131 (Jan. 1991).
4. A. Erramilli, O. Narayan, A. Neidhardt, and I. Sanjee, Performance impacts of multi-scaling in wide area TCP/IP traffic, *Proc. INFOCOM'00*, 2000.
5. A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, Looking behind and beyond self-similarity: On scaling phenomena in measured WAN traffic, *Proc. 35th Annual Allerton Conf. Communications, Control and Computing*, 1997, pp. 269–280.
6. N. L. S. Fonseca, G. S. Mayor, and C. A. V. Neto, On the equivalent bandwidth of self similar sources, *ACM Trans. Model. Comput. Simul.* **10**(3): 104–124 (2000).
7. V. Frost and B. Melamed, Traffic modeling for telecommunications networks, *IEEE Commun. Mag.* (March 1994).
8. C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye, Modeling and simulation of self-similar variable bit rate compressed video: A unified approach, *Proc. SIGCOMM'95 Conf.*, 1995, pp. 114–125.
9. D. Jagerman, B. Melamed, and W. Willinger, Stochastic modeling of traffic processes, in J. Dshalalow, ed., *Frontiers in Queuing: Models, Methods and Problems*, CRC Press, Boca Raton, FL, 1996.
10. M. M. Krunz and A. M. Makowski, Modeling video traffic using M/G/∞ input process: A comparison between Markovian and LRD models, *IEEE J. Select. Areas Commun.* **16**(5): 733–745 (June 1998).
11. N. Laskin, I. Lambadaris, F. Harmantzis, and M. Devetsikiotis, Fractional Lévy motion and its application to traffic modeling, *Comput. Networks,* (*Special Issue on Long-Range Dependent Traffic Engineering*) **40**(3): (Oct. 2002).
12. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Network.* **2**(1): 1–15 (1994).
13. B. Liu et al., A study of networks simulation efficiency: Fluid simulation vs. packet-level simulation, *Proc. IEEE Infocom*, Alaska, April 2001.
14. B. B. Mandelbrot and J. W. Van Ness, Fractal Brownian motions, fractional noises and applications, *SIAM Rev.* **10**: 422–437 (1968).
15. I. Norros, A storage model with self-similar input, *Queuing Syst.* **16**: 387–396 (1994).
16. K. Park and W. Willinger, eds., *Self Similar Network Traffic and Performance Evaluation*, Wiley-Interscience, 2000.
17. V. Paxson and S. Floyd, Wide area traffic: The failure of Poisson modeling, *IEEE/ACM Trans. Network.* **3**(3): 226–244 (1995).
18. J. A. Silvester, N. L. S. Fonseca, and S. S. Wang, D-Bmap models for the performance analysis of ATM networks, in D. Kouvatsos, ed., *Performance Modeling of ATM Networks*, Chapman & Hall, 1995, pp. 325–346.
19. M. S. Taqqu, V. Teverovsky, and W. Willinger, Is network traffic self-similar or multifractal? *Fractals* **5**: 63–73 (1997).
20. T. Taralp, M. Devetsikiotis, and I. Lambadaris, In search of better statistics for traffic characterization, *J. Braz. Comput. Soc.,* (*Special Issue on Traffic Modeling and Control of Wired and Wireless Networks*) **5**(3): 5–13 (April 1999).
21. S. Tartarelli et al., Empirical effective bandwidths, *Proc. IEEE GLOBECOM 2000*, Vol. 1, 2000, pp. 672–678.
22. B. Tsybakov and N. D. Georganas, Self-similar processes in communication networks, *IEEE Trans. Inform. Theory* **44**(5): (Sept. 1998).

23. W. Willinger and V. Paxson, Where mathematics meets the Internet, *Notices Am. Mathe. Soc.* **45**(8): 961–970 (Sept. 1998).

24. W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level, *IEEE/ACM Trans. Network.* **5**(1): 71–86 (Feb. 1997).

NEURAL NETWORKS AND APPLICATIONS TO COMMUNICATIONS

ELIAS S. MANOLAKOS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Conventional (serial) computers have a processing unit that executes instructions one after the other, *sequentially*, at a rate that could be as high as 1 billion instructions per second. The way the human brain processes information seems, however, to be quite different. The brain has about 100 billion processing units, called *neurons*, that are highly interconnected. A neuron typically is connected to more than 1,000 other neurons, giving rise to more than 100 trillion connections. Even if only 0.1% of them are active at any given time, and each connection functions as a very slow processor performing one computation every 5 milliseconds, the brain can deliver *in parallel* 100 trillion operations per second (100 Teraops)! Exploiting vast amounts of low-level parallelism may explain why the human brain can perform high-level perceptual tasks, such as visual pattern recognition, speech understanding and so on, very efficiently even though its pattern-searching capabilities are very modest compared with today's fast computers.

Should we try to build computers that work like the human brain? Should we try to imitate closely what we know from biology? These questions have been long debated in the scientific and engineering community. At the one end, *computational neuroscientists* tell us that accurate modeling of the neuronal interactions will illuminate how the brain operates. At the other end, engineers, who are mostly interested in building intelligent machines, tell us that it may be sufficient to draw from the principles that led to successful designs in biology without trying to approximate the biological systems very closely. So starting from the same inquiry, the highly interdisciplinary field of *neural networks* (NN), or *neurocomputing*, has emerged and moved outward in several different directions, and it continues to evolve.

Artificial neural networks (ANNs) are highly interconnected distributed information processing architectures that are built by connecting many simple processing unit models. They come in many different flavors, but all are characterized by their ability to *learn*, that is, to adapt their behavior and structure to capture and form a representation of the process that generates information in the environment where they operate. It is typical to select first an appropriate NN model; then adequately *train* it, using data representative of the underlying environment; and finally present to it novel (unseen) data and let the network extract useful information, a process called *generalization*, or *recall*. Neural networks have become a mainstream information technology and are no longer considered exotic techniques. They are commonly used in data analysis (time series, forecasting, compression, etc.), in studying systems behavior (system identification, adaptive control, chaotic behavior etc.), and in modeling stochastic processes (for uncertainty characterization, pattern classification etc). Neural network solutions have been tried on all types of scientific and engineering problems, ranging from signal and image processing, communications, and intelligent multimedia systems, to data mining, credit card fraud detection, economic forecasting, modeling of ecological systems and detecting patterns in gene expression microarrays, and so on.

This article is organized as follows: In Section 2, we introduce basic knowledge NNs such as the structure of commonly used processing elements, a review of popular network architectures, and associated learning rules. In Section 3, we discuss why different NN architectures have found so many applications in communications by selecting some well-known subareas and discussing representative cases of their use.

2. BASIC NEURAL NETWORKS THEORY

2.1. The Neuron Node

Let us start by considering a simplified view of a bipolar biological neuron. At the "center" of the neuron there is the cell body, or *soma*, that contains the cell *nucleus*. One or more *dendrites* are connected to the nucleus. They are so called because they structurally resemble the branches of a tree. Dendrites form the receiving end of the nerve cell, and their role is to sense activity at their neighborhood and generate a proportional amount of electrical impulses that are sent toward the cell body. A long signal transmission line, called an *axon*, is emanating at the cell body and carries the accumulated activity (action potentials) away from the soma toward other neurons. At the other end of the axon there are *synaptic terminals*. There, the transmitted signal is translated to signals sensed by the dendrites of neighboring neurons through an electrochemical interface process called a *synapse*. A typical cortical neuron may have a few thousand synapses for receiving input, and its axon may have a thousand or so synaptic terminals for propagating activity to other neurons' dendrites.

The communication operations of the simplified neuron model can be summarized as follows: Signals (impulse trains) arrive at the various input synapses. Some signals are stronger than others and some synapses are less receptive than others, due to causes such as fatigue, exposure to chemical agents, and so on. So, as induced signals travel toward the cell body, they effectively are "weighted" (i.e., they are not all deemed equally important). The induced activity is integrated over time and if it exceeds a threshold, the neuron "fires" (produces an output). The generated output signal is transmitted along the axon and may induce signals at the dendrites of other neurons in the proximity.

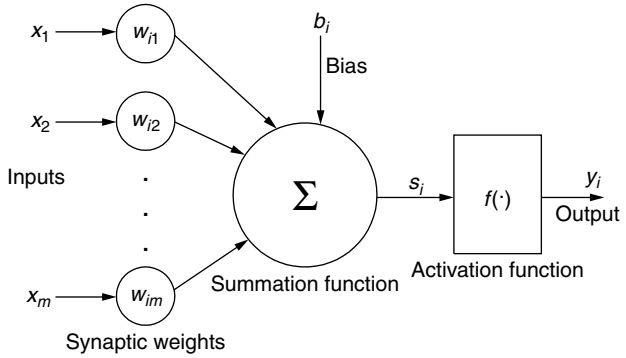


Figure 1. The basic neural network node model.

A neuron node, or *processing element* (PE), of an ANN is a simplistic model that mimics the sequence of the actions described above (see also Fig. 1). Each input $x_k, k = 1, 2, \ldots, m$, to PE_i is first multiplied by the corresponding *synaptic weight*, w_{ik}, that models the strength of the synapse between neurons k an i. Then all weighted inputs are summed. A node specific *bias*, or forcing term, b_i, is also added to the partial sum, providing an additional mechanism to influence node i directly and not through the outputs of other neurons. The *induced local field* s_i becomes the input of an *activation function* $f(\cdot)$ that determines the output y_i of the node. The whole processing can be described by the following two simple equations:

$$s_i = \sum_{k=1}^{m} w_{ik} x_k + b_i \tag{1}$$

$$y_i = f(s_i) \tag{2}$$

The activation, or *transfer function* $f(\cdot)$ usually is nonlinear and limits the range of the values of the neuronal output y_i. Among the most widely used functional forms are those shown in Fig. 2. The hard delimiter (left-most panel) produces a binary, 0 or 1, output when the induced local field has a negative value or positive value, respectively. The transition from level 0 to level 1 is gradual when using a linear ramp-like function (middle panel). Perhaps the most commonly used activation function is the *sigmoidal*

$$f(x) = \frac{1}{1 + e^{-ax}} \tag{3}$$

(see right-most panel). As a increases, the sigmoidal nonlinearity behaves like a hard delimiter. When a -1, $+1$, *antipodal* hard delimiter is used, the neuron node is also usually referred to as a *perceptron*.

2.2. Neural Network Architectures

As with biological NNs, individual neurons are not very useful when working in isolation. Neuron units are usually organized in layers. In a *feedforward* NN, the outputs of the nodes in one layer become inputs to the nodes of the next layer. This is shown in Fig. 3, where circles depict neuron units and arcs represent weighted connections among them. The network is *fully connected* because the output of a node feeds into every node of the next layer

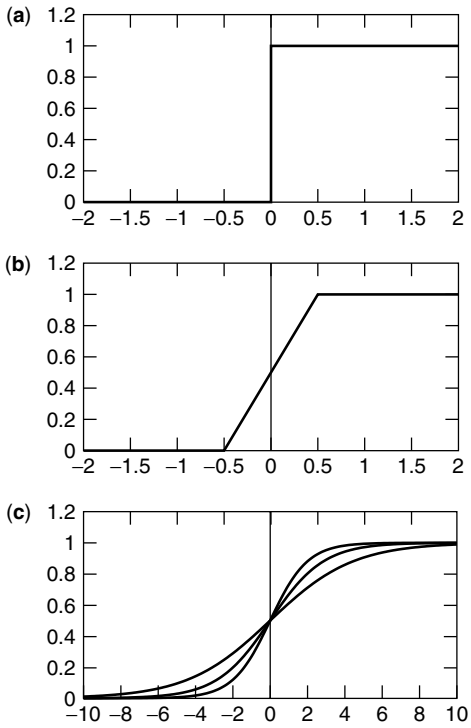


Figure 2. Neural network node activation functions. From left to right: (**a**) hard delimiter; (**b**) piecewise linear; (**c**) sigmoidal; it approaches the hard delimiter as a increases.

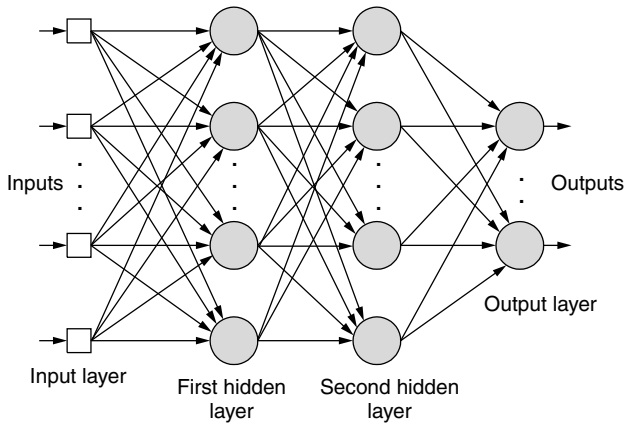


Figure 3. The multilayer feedforward neural network architecture.

via a weighted connection. It has one input layer, two *hidden* layers, and one output layer. The input layer is where the network receives stimuli from the environment, and the output layer is where it returns its response. The middle layers are called "hidden" because their nodes are not directly accessible from outside the network.

If the output of a neuron may become an input to itself, or to other neurons of the same or previous layers, the resulting architecture is called a *feedback*, or *recurrent*, NN. A single layer, fully recurrent NN with four nodes is shown in Fig. 4, where a small box denotes a synaptic weight. Each PE receives a weighted input from the output of every node, including itself, and from an external bias.

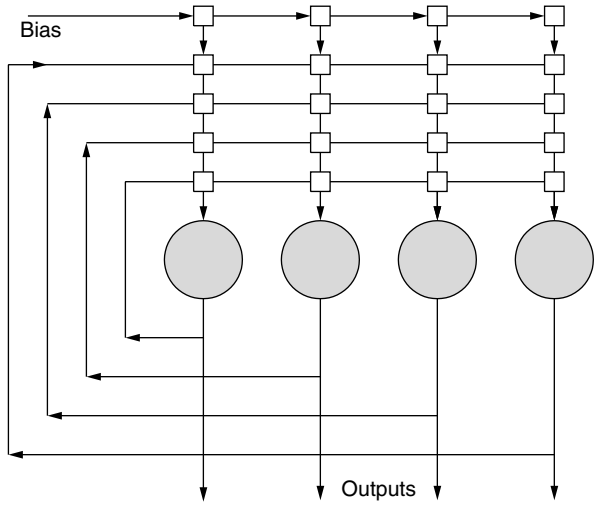


Figure 4. A single layer fully recurrent neural network architecture.

All weighted inputs are accumulated to produce the neuron activation that is passed through a nonlinear thresholding transfer function to produce the new output, as shown in Fig. 1.

Hopfield neural networks (HNNs) [1,2], are single-layer recurrent networks that can collectively provide good solutions to difficult optimization problems. A connection between two neuron processors (analog amplifiers) in established through a conductance weight T_{ij}, which transforms the voltage outputs of neuron unit j to a current input for neuron unit i. Externally supplied bias currents I_i are also feeding into every neuron processor.

It has been shown [3] that in the case of symmetric weight connections ($T_{ij} = T_{ji}$), the equations of motion for the activation of the HNN neurons always lead to convergence to a stable state, in which the output voltages of all the neurons remain constant. In addition, when the diagonal weights (T_{ii}) are zero and the width of the neuron amplifier gain curve is narrow, (i.e., the nonlinear activation function $f(\cdot)$ approaches the antipodal thresholding function), the stable states of a network with N neuron units are the *local* minima of the quadratic *energy function*

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}V_iV_j - \sum_{i=1}^{N} V_iI_i \tag{4}$$

If the small boxes in Fig. 4 represent conductance weights that meet the conditions discussed, the recurrent NN shown in the figure becomes an HNN with $N = 4$ neuron processors.

2.3. Learning and Generalization

2.3.1. Supervised Learning. Neural network learning amounts to adjusting the synaptic weights periodically and until an appropriate cost function is sufficiently minimized. There are two categories of learning methods. In *supervised* learning, also called "training with a teacher," it is assumed that a desired response $\mathbf{d}$ is provided for every input vector $\mathbf{x}$ in a training set $\mathcal{X}$. For every pair $(\mathbf{x}, \mathbf{d}) \in \mathcal{X}$ the network output $\mathbf{y}$ is compared to $\mathbf{d}$ and the difference (error) vector $\mathbf{e} = \mathbf{d} - \mathbf{y}$ is used to determine how the network weights will get updated, as suggested by the block diagram in Fig. 5 (a).

Supervised learning can be used to train a neural network to act as a *pattern classifier*. Let us assume for simplicity that the network input patterns (vectors) belong to one of two possible classes, C_1 or C_2. Then it suffices to use only one neuron in the output layer; its desired output can be $+1$ if it is known that the pattern belongs to class C_1, or -1 if the pattern belongs to class C_2. By adding more neurons in the output layer, neural classifiers that can discriminate among more than two categories can be built.

Supervised learning can also be used in the more general *pattern association* context. If the input and desirable output patterns are identical, the network is trained to act as an *associative memory*. After adequate training, when the network is presented with an unseen input pattern, it is expected to *recall* the learned pattern that most closely resembles the input. In general, input vectors can be n-dimensional and desirable output vectors m-dimensional, with $m \neq n$.

An example of a weight updating rule used in supervised learning is the so-called *delta rule*, or *Windrow-Hoff* rule. It states that the amount of weight adjustment of synapse k to neuron i, Δw_{ik}, should be proportional to the product of the observed error e_i and the input signal to

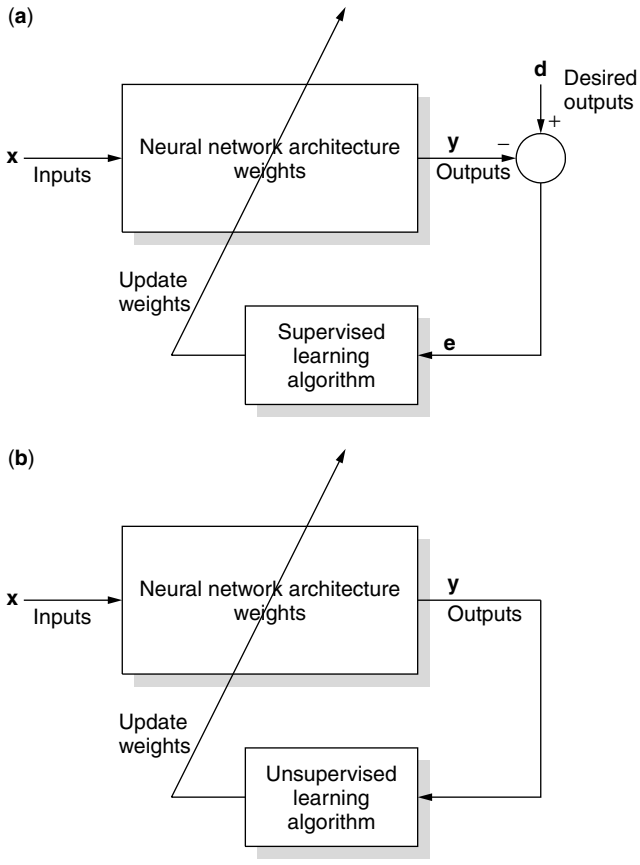


Figure 5. Learning methods: (**a**) supervised training; (**b**) unsupervised training.

neuron i from neuron k, that is

$$\Delta w_{ik} = \alpha e_i x_k \tag{5}$$

Constant α controls the *learning rate* and usually is a small number between 0 and 1. Therefore, after the presentation of input pattern n, the updated weight becomes

$$w_{ik}(n+1) = w_{ik}(n) + \Delta w_{ik}(n) \tag{6}$$

Note that the delta rule uses only information local to neuron i in updating its synaptic weights. However, it does assume that the error term is accessible (i.e., that a desired output d_i is provided for every neuron). It has been shown that if the input vectors in a training set can be separated by a linear hyperplane into two classes, this hyperplane (decision boundary) can be determined by using the weights, upon convergence, of a single layer perceptron trained using correction-based learning.

The error term is directly accessible only for the neurons in the output layer L of a multilayer perceptron (MLP). However, an elegant generalization of the delta rule has been derived that allows the supervised training of MLPs. It is widely known as the *back-propagation algorithm* [4,5] and works as follows. First, the network weights are initialized and an input pattern $\mathbf{x}(n)$ is presented to the input layer. The pattern propagates forward to produce output $\mathbf{y}(n)$ in the output layer L. The network error for pattern n can be defined as

$$E(n) = \frac{1}{2}\sum_{i \in L} e_i^2(n) \tag{7}$$

Adjusting the weight to minimize the total average error $E = 1/N \sum_n E(n)$ leads to a generalized delta rule

$$\Delta w_{ik}(n) = \alpha \delta_i(n) x_k(n) \tag{8}$$

where $\delta_i(n) = \dfrac{\partial E(n)}{\partial s_i}$ is the *local gradient* of the network error for pattern n with respect to neuron i. If neuron i is at the output layer L, this quantity is directly proportional to the error $e_i(n)$. On the other hand, if neuron i is at some level $l < L$, $\delta_i^l(n)$ is proportional to the sum of the $w_{ji}\delta_j^{l+1}(n)$ terms, taken over all neurons j in layer $l+1$ that are connected to neuron i in layer l. So the back-propagation algorithm starts from the output layer L, where vector δ^L is computed directly using the measured error vector $\mathbf{e}$ and is used to update the synaptic weights from layer $L-1$ to layer L. Then these weights and δ^L are used to compute δ^{L-1} and update the synaptic weights from layer $L-2$ to layer $L-1$, and so on. So the error correction propagates backward, toward the input layer, and the weights connecting neurons in two successive layers can be updated in parallel using only localized processing.

Supervised learning methods have been used extensively in communication applications such as channel modeling and identification, adaptive equalization, and so on, as is discussed in more detail in Section 3.

2.3.2. Unsupervised Learning. As with biological neurons, an ANN may have to learn without a "teacher" (i.e., without using a prelabeled training data set). In *unsupervised learning*, as input patterns are presented one after the other, the network should be able to build a meaningful representation of their distribution. Furthermore, it should detect changes in the important characteristics of this distribution and react by adapting its parameters accordingly, as suggested by the block diagram in Fig. 5 (b).

To *self-organize* the network relies on local processing rules that over time help it discover important features in the input patterns and track how they change. A stable representation eventually emerges, for example, a clustering or a topological map arrangement, that models quite well the input patterns' distribution.

Self-organization requires some form of competition, possibly followed by cooperation among neurons in the neighborhood of the winner. In *competitive learning*, neurons start competing when a new pattern is presented to the network. The neuron that is most excited is usually declared the "winner." Neurons in the winner's neighborhood are then cooperating and update their weights according to a local learning rule. For example, in the case of Kohonen's *self-organizing map* (SOM) [5,6], if neuron i wins in the competition for representing pattern $\mathbf{x}(n)$, only neurons j in its topological neighborhood (same region in the lattice of neurons) are allowed to update their weights according to the rule

$$\Delta \mathbf{w}_j(n) = \alpha h_{ji}(n)(\mathbf{x}(n) - \mathbf{w}_j(n)) \tag{9}$$

where function $h_{ji}(n)$ is defined over the winner's neighborhood and its value decreases as the distance of neuron j from the winner neuron i increases. Furthermore, the domain of $h_{ji}(n)$ may shrink over time. In essence, Kohonen's learning rule "moves" toward the input pattern vector, the weight vectors of neurons located in the winner's neighborhood.

A similar situation arises when training *radial basis functions* (RBF) [5,7]. RBFs are three-layer feedforward NNs. The first layer is used to input patterns; every neuron k in the middle (hidden) layer receives the input pattern vector $\mathbf{x}$ and produces an output based on a Caussian bell-shape nonlinear activation function

$$\phi_k(\mathbf{x}, \mathbf{c}_k) = \exp\left(-\frac{1}{2\sigma_k^2}\|\mathbf{x} - \mathbf{c}_k\|^2\right) \tag{10}$$

where $\mathbf{c}_k$ is the *center* associated with the kth neuron. Finally, each neuron j in the output layer is computing a different linear combination of the hidden layer neuron outputs, that is

$$y_j = \sum_k w_{jk}\phi_k(\mathbf{x}, \mathbf{c}_k) \tag{11}$$

A combination of unsupervised and supervised training can be used for RBF networks. The hidden neuron centers may be updated in an unsupervised manner as follows:

$$\Delta \mathbf{c}^* = \beta\,(\mathbf{x}(n) - \mathbf{c}^*) \tag{12}$$

where $\mathbf{c}^*$ is the center of the winning neuron [i.e., the one at smallest distance from the input pattern $\mathbf{x}(n)$]. The weights $\{w_{jk}\}$ may then be updated using error correction supervised learning

$$\Delta w_{jk} = \alpha\,(d_j(n) - y_j(n))\,\phi_k(\mathbf{x}, \mathbf{c}_k) \tag{13}$$

where $d_j(n)$, $y_j(n)$ are the desired and actual response of output layer neuron j to input pattern $\mathbf{x}(n)$ in the training set $\mathcal{X}$.

Unsupervised training is also used extensively in communication applications, such as coding and decoding, vector quantization, neural receiver structures, and blind equalization, to mention a few.

3. NEURAL NETWORK APPLICATIONS TO COMMUNICATIONS

Neural network techniques have traditionally been employed in solving complex problems where a conventional approach has not demonstrated satisfactory performance or has failed to adequately capture the underlying data-generation process. Although they often do improve the performance substantially, it sometimes is hard to understand the decisions they make and explain how these performance gains come about. This has been a point of criticism to the "black box" use of NNs. However, as the field matures and a clear link to Bayesian statistics is established [8], analyzing systematically and explaining how complex networks form successful representations is becoming more and more possible.

In the limited space of this article, we do not intend to provide a comprehensive survey of the numerous applications that NNs have found in communications. We will rather select a few well-known problems in communications and discuss some indicative neurocomputing solutions proposed for them. The selection of the problems to be discussed does not imply that they are more important than others. For a comprehensive review of the literature, the interested reader is referred to the recent article [9] and the edited collections of papers in Refs. 10 and 11.

3.1. Channel Modeling and Identification

Various feedforward neural network architectures, including the multilayer perceptrons and radial basis functions, are *universal approximators* [12,13]. This important property ensures that a neural network of appropriate size can approximate arbitrarily well any continuous nonlinear mapping from an n- to an to m-dimensional space, where $m \neq n$.

Communication channels often exhibit nonlinear, slowly time-varying behavior. So it is natural that NN techniques have been tried for the modeling and identification of communication channels that are used in receiver design, performance evaluation, and so on. Typical examples include satellite channels identification [14] and the modeling of nonlinear time-varying fading channels [15]. In this context, neural networks, with a moderate number of weights (free parameters) that can be updated in parallel, have been shown to provide an attractive alternative to classical nonlinear system identification methods, such as those based on Volterra series [16].

3.2. Channel Equalization

The demand for very high-speed transmission of information over physical communication channels has been constantly increasing over the past 20 years. Communication channels are usually modeled as linear filters having a low-pass frequency response. If the filter characteristics are imperfect, the channel distorts the transmitted signal in both amplitude and delay, causing what is known as *intersymbol interference* (ISI) [17]. As a result of this linear distortion, the transmitted symbols are spread and overlapped over successive time intervals. In addition to noise and linear distortion, the transmitted symbols are subject to other *nonlinear* impairments arising from the modulation/demodulation process, crosstalk interference, the use of amplifiers and converters, and the nature of the channel itself. All the signal processing techniques used at the receiver's end to combat the introduced channel distortion and recover the transmitted symbols are referred to as *adaptive equalization* schemes.

Adaptive equalization is characterized in general by the structure of the equalizer, the adaptation algorithm, and the use or not of training sequences [17]. Linear equalization employs a linear filter, usually with a finite impulse response (FIR) or lattice structure. A recursive least squares (RLS) algorithm or a stochastic gradient algorithm, such as the least mean squares (LMS), is used to optimize a performance index. However, when the channel has a deep spectral null in its bandwidth, linear equalization performs poorly because the equalizer places a highgain at the frequency of the null, thus enhancing the additive noise at this frequency band [17]. Decision feedback (DFE), in conjunction with a linear filter equalizer, can be employed to overcome this limitation. Although DFE and other methods, such as the maximum likelihood (ML) sequence detection [18], are nonlinear, the nonlinearity usually lies in the way the transmitted sequence is recovered at the receiver with the channel model being linear. If nonlinear channel distortion is too severe to ignore, the aforementioned algorithms suffer from a severe performance degradation. Among the many techniques that have been proposed to address the nonlinear channel equalization problem are those in Refs. 19–21, which rely on the Volterra series expansion of the nonlinear channel.

The authors in Ref. 22 have used an MLP feedforward NN structure for the equalization of linear and nonlinear channels. The network is trained to approximate the correct mapping from delayed channel outputs to originally transmitted symbols. It is demonstrated that significant performance improvements can be achieved. A functional-link NN-based DFE equalizer that exceeds the bit error rate performance of conventional DFE equalizers was reported in Ref. 23. For two-dimensional signaling, such as quadrature amplitude modulation (QAM) or phase shift keying (PSK) [17], NNs that can process complex numbers are needed. The authors in Ref. 24 have designed equalizers based on complex-valued radial basis functions and have shown that they can approximate well the decisions

of the optimal Bayesian equalizer. Combining in a loop a DFE equalizer and a self-organizing features map is discussed in Ref. 25. An interesting two-stage equalization strategy is proposed where the DFE compensates for dynamic linear distortions and the SOM compensates for nonlinear ones.

Fully recurrent neural networks (RNNs) have been used in Ref. 26 for both trained adaptation and blind equalization of linear and nonlinear communication channels. Since RNNs essentially model nonlinear infinite memory filters, they can accurately realize with a relatively small number of parameters the inverse of finite memory systems, and thus compensate effectively for the channel-introduced interferences. Their performance is shown to exceed that of traditional equalization algorithms and feedforward NN schemes, especially in the presence of spectral nulls and/or severe nonlinearities. Furthermore, due to the small number of neurons involved, the computational cost of their training may, in practice, be much smaller than that of the MLP-based equalizers of similar performance.

Blind equalization is a particularly useful and difficult type of equalization when training sequences are undesirable or not unfeasible, as, for example, in the case of multipoint communication networks and strategic communications. In the absence of a training sequence, the only knowledge about the transmitted signal is the constellation from which the symbols are drawn. A novel RNN training approach was introduced in Ref. 26 for the blind equalization of nonlinear channels using only a partial set of statistics of the transmitted signal. It is shown that simple RNN structures can equalize linear and nonlinear channels better than the traditional constant modulus algorithm.

3.3. CDMA Multiuser Detection

Code division multiple access (CDMA) is a spectrum-efficient method for the simultaneous transmission of digital information sent by multiple users over a shared communication channel. The spectral efficiency, as well as the antijamming and other attractive properties, make CDMA spread-spectrum techniques useful in a number of communication technologies, including mobile telephony and satellite communications. The wide bandwidth that spread-spectrum CDMA techniques employ enables them to exploit powerful low-rate error-correction coding to further enhance performance. The major limitation of the CDMA techniques however, is the so-called *near–far* problem. When the power of the signals transmitted by the users becomes very dissimilar, the conventional matched-filter detector exhibits severe performance degradation, so more complicated detectors have to be employed.

The optimum centralized demodulation of the information sent simultaneously by several users through a shared Gaussian multiple access channel is a very important problem arising in multipoint-to-point digital communication networks such as radio networks, local area networks, uplink satellite channels, uplink cellular communications, and so on. In CDMA, each transmitter modulates a different signature signal waveform, which is known to the receiver. At the receiver, the incoming signal is the sum of the signals transmitted by each individual user. To demodulate the received signal, we need to suppress the inherent channel noise, often modeled as an additive Gaussian process, and the multiple access interference (MAI).

It has been shown by Verdu et al. [27,28] that for the Gaussian channel, *optimal CDMA multiuser detection* (OMD) can be formulated as the solution of a quadratic integer programming problem that involves the sampled outputs of a bank of filters matched to the signature waveforms of the transmitting users as well as the cross-correlations of them. In Ref. 29, it is proved that, in both the *synchronous* and the *asynchronous* transmission cases, OMD is a computationally expensive problem for which polynomial time solutions most likely do not exist (NP-hard). Therefore, research efforts have concentrated on the development of suboptimal receivers that exhibit good near–far resistance properties, have low computational complexity, and achieve bit-error-rate (BER) performance that is comparable to that of the optimal receiver. Among the many suboptimal multiuser detectors proposed in the literature, we mention the *decorrelating detector* [28], which is linear in nature and complexity and achieves near-optimal performance, assuming that the users' signals form a linearly independent set and the spreading codes of all users are known. Another suboptimal detector is the *multistage detector* (MSD) [30], which relies on improving each stage's estimate by subtracting the estimate of the MAI obtained by the previous stage.

Feedforward NN-based multiuser detectors were first proposed in Refs. 31 and 32. While their performance is shown to be very good for a very small number of synchronous or asynchronous users, their hardware complexity (number of neurons and training time) appears to be *exponential* in the number of users, as conjectured in Ref. 31. Furthermore, it is only empirically possible to determine the number of neurons in the hidden layer as the number of users increases.

The well-known ability of HNNs to provide fast suboptimal solutions to hard combinatorial optimization problems has been exploited to implement efficiently the CDMA OMD in Ref. 33. Starting from the observation that the OMD problem's objective function can be put in the quadratic format of equation (4), an HNN multiuser detector is derived that is proven to be a generalization of the multistage detector. The HNN-based detector has been evaluated via extensive simulations and has been found to outperform the CD by orders of magnitude, exceed the performance of the MSD, and approach the performance of the OMD. Furthermore, the HNN-based detector has a hardware complexity (number of neurons) that is *linear* in the number of users K and does not require any training. Since it can be implemented directly using analog VLSI hardware, its computational cost per symbol is practically constant irrespective of the number of users.

However, as the number of simultaneous users increases, all NN-based receivers may become impractical. To address this severe limitation, a hybrid digital signal preprocessing–NN CDMA multiuser detection scheme was proposed in Ref. 34. An investigation on the nature of the local minima of the OMD's objective function led to the formulation of a computationally efficient digital

signal preprocessing stage that recursively reduces the size of the search space over which the original large-size OMD optimization problem has to be solved. After preprocessing, the remaining optimization problem has the same structure as the OMD but is much smaller and can be solved by an HNN implementable directly in hardware. The lesson learned is that combining conventional DSP with NN methods is worth considering because it often leads to optimized and practical solutions.

3.4. Networking Applications

Multimedia teleconferencing, video-on-demand, and distant learning are only a few examples of emerging applications that require high-speed communications. Each such application presents to the underlying network infrastructure different traffic characteristics and quality of service (QoS) requirements. An intelligent network should efficiently broker among applications with time-varying profiles to meet their short-term requirements while also maximizing the delivered long-term throughput.

The *asynchronous transfer mode* (ATM) is a widely accepted backbone networking technology due to its provisions for QoS delivery. ATM has builtin proactive and reactive mechanisms for effectively managing network resources (e.g., buffer space, etc.) according to traffic profiles and QoS requirements. By using them, it is possible to statistically time multiplex among sources with different burstiness and bit rate characteristics (e.g., voice, data, video) as they compete for the available bandwidth. It is not surprising that NNs, with their learning from examples, adaptation, and prediction capabilities, are among the artificial intelligence methods that have been employed extensively in this context.

The proactive mechanism that decides if a new call can be accepted given the current state of the network is termed *call admission control* (CAC). The integration of adaptive CAC and link capacity control for multimedia traffic over ATM is discussed in Refs. 35 and 36. Neural networks are trained to estimate the cell loss rate from link capacity and observed traffic. The link capacity assignment is then optimized according to the estimated cell loss rate. The application of NNs to the selective admission of a set of calls from a number of inhomogeneous call classes with differing rate and traffic variability characteristics is discussed in Ref. 37.

A good example of adaptive network traffic management is the dynamic allocation of video transmission bandwidth by taking into account the rate of scene changes. In Ref. 38 it is shown that it is feasible to predict scene changes online by employing low-complexity time-delay NNs and use these predictions in dynamic bandwidth allocation for the efficient transmission of real-time video traffic over ATM networks. Furthermore, in Ref. 39 a system consisting of pipelined recurrent NN models, where each RNN has a small number of neurons, is shown to be able to accurately predict the future behavior of MPEG video traffic, based on the ability of each RNN module to learn from previous traffic measurements. The system was used to guide control actions in real time and to prevent excess network loading.

An important task of adaptive network management is to ensure that applications honor their commitments and to respond if they violate their "contract" with the network. During call progress, the same parameters used for call admission may be utilized by a reactive network *policing* mechanism to ensure that the user's traffic remains within the prenegotiated values. A network policing architecture consisting of two interconnected NNs is introduced in Ref. 40. The first NN is trained to learn the probability density function (pdf) of the nonviolating traffic profile of the application. The second NN is trained to capture the characteristics of the pdf of the actual traffic as the call progresses. If the two distribution profiles start deviating considerably, an error signal is generated and used to "reshape" the accepted traffic [41]. The results show that this policing method can efficiently detect and properly react to peak and mean traffic violations. In Ref. 42 a closed loop, end-to-end, broadband network traffic control system is presented that employs different NN architectures for traffic characterization, call admission control, traffic policing, and congestion control.

3.5. Other Communications Applications

Other telecommunication areas where NNs have been used include, but are not limited to: cellular and mobile communications [43,44], mobile phone and credit card fraud detection [45,46], intelligent switching and routing [47,48], scheduling problems in radio networks [48,49], and so on. A large collections of application-related papers can be found in books [10,11].

Neural networks have also been extensively utilized in diverse application domains that are related to communications, such as electronic commerce over the Internet [50], modeling agents behavior [51], automatic language identification [52], text-independent speaker verification [53], character recognition and document analysis [54], image vector quantization [55], signal processing [56], and pattern recognition [8].

4. THE FUTURE?

The highly interdisciplinary field of NNs has always been an area where engineers, statisticians, neuroscientists, and cognitive scientists meet and interact. It is this cross-fertilization of ideas that has rejuvenated NN research over the past 10 years. At the one end, the field has established new links with areas of advanced statistics and machine learning, such as independent component analysis [57], support vector machines [58], graphical models [59], and so on. Coupled with developments in computational neuroscience and digital communications, these advances may soon lead to breakthroughs in novel fields such as that of *neurotechnology*, which promises the use of information technology to substitute for lost functionality in the human nervous system [60].

BIOGRAPHY

Elias S. Manolakos is leading the Parallel Processing and Architectures research group of the Communications

and Digital Signal Processing (CDSP) Center, a world-class Center for Research and Graduate Studies of the Electrical and Computer Engineering Department at Northeastern University, Boston, Massachusetts, where he is currently an associate professor. His research interests are in parallel and distributed computing; embedded systems; pattern recognition; neural networks; and applications in signal processing, communication, and biocomputing. He has served on the editorial boards of the *IEEE Transactions of Signal Processing, IEEE Computing in Science and Engineering, Journal of VLSI Signal Processing, IEEE Letters*, among others Manolakos has participated in the organization of several conferences and has chaired the technical program of the IEEE International Workshop on Signal Processing Systems Design and Implementation (SiPS, 1998) and the IEEE International Workshop in Neural Networks for Signal Processing (NNSP, 1995). He has authored or coauthored with his students more than 70 referenced publications and has coedited three books. He is a senior member of the IEEE and an elected advisory board member of the IEEE Signal Processing Society's Technical Committee on Neural Networks for Signal Processing.

BIBLIOGRAPHY

1. D. V. Tank and J. J. Hopfield, Simple "neural" optimization networks: an A/D converter, signal decision circuit, and a linear programming circuit, *IEEE Trans. Circuits and Systems* **33**: 533–541 (May 1990).
2. J. J. Hopfield, Neural networks and physical systems with emerging collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**: 2554–2558 (1982).
3. J. J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA* **81**: 3088–3092 (1984).
4. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning internal representation by error propagation. In D. E. Rumelhart and J. L. McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, MIT Press, Cambridge, MA, 1986, pp. 318–362.
5. S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 1999.
6. T. Kohonen, Self-organized formation of topological feature maps, *Biological Cybernetics* **43**: 59–69 (1982).
7. F. Girosi, M. Jones, and T. Poggio, Regularization theory and neural networks architectures, *Neural Computation* **7**(2): 219–269 (1995).
8. C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, UK, 1995.
9. M. Ibnkahla, Applications of neural networks to digital communications — a survey, *Signal Processing* **80**: 1185–1215 (2000).
10. N. Ansari and B. Yuhas, *Neural Networks in Telecommunications*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1994.
11. J. Alspector, R. Goodman, and T. X. Brown, eds., *Applications of Neural Networks to Telecommunications*, Lawrence Erlbaum, Hillsdale, NJ, 1993.
12. G. Cybenko, Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems* **2**(4): 303–314 (1989).
13. K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks* **2**: 359–366 (1989).
14. M. Ibnkahla, N. J. Bershad, J. Sombrin, and F. Castanié, Neural network modeling and identification of non-linear channels with memory: Algorithms, applications and analytic models, *IEEE Trans. Signal Proc.* **46**: 1208–1220 (May 1998).
15. M. Ibnkahla, J. Sombrin, F. Castanié, and N. J. Bershad, Neural network modeling non-linear memoryless communication channels, *IEEE Trans. Commun.* **45**: 1208–1220 (July 1997).
16. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, Wiley, New York, 1980.
17. J. Proakis, *Digital Communications*, Prentice Hall Inc., Cliffside Park, NJ, 1988.
18. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *Proc. IEEE, Trans. Infor. Theory* **IT-18**: 378–383 (May 1972).
19. S. Benedetto and E. Biglieri, Nonlinear equalization of digital satellite channels, *IEEE J. Select Areas Commun.* **SAC-1**: 57–62 (Jan. 1983).
20. E. Biglieri, A. Gersho, R. D. Gitlin, and T. L. Lim, Adaptive cancellation of nonlinear intersymbol interference for voiceband data transmission, *IEEE J. Select Areas Commun.* **SAC-2**: 765–777 (Sept. 1984).
21. D. D. Falconer, Adaptive Equalization of Channel Nonlinearities in QAM Data Transmission Systems, *Bell Syst. Tech. J.* **57**(7): 2589–2611 (1978).
22. G. J. Gibson, S. Siu, and C. F. N. Cowan, Application of multilayer perceptrons as adaptive channel equalizers, in *ICASSP Int. Conf. Accoustics, Speech and Signal Proc.*, Glasgow, Scotland, May 1989, pp. 1183–1186.
23. A. Husain, S. Soraghan, and T. Durrani, A new adaptive functional-link neural network-based dfe for overcoming co-channel interference, *IEEE Trans. Commun.* **45**: 1358–1362 (1997).
24. S. Chen, S. McLaughlin, and B. Mulgrew, Complex-valued radial basis function networks, Part I: Network architecture and learning algorithms, *Signal Proc.* **35**(1): 175–188 (Jan. 1994).
25. T. Kohonen, E. Oja, O. Simula, and A. Visa, Engineering application of the self-organizing map, *IEEE Proc.* 1357–1384 (Oct. 1996).
26. G. Kechriotis, E. Zervas, and E. S. Manolakos, Using recurrent neural networks for adaptive communication channel equalization, *IEEE Trans. Neural Networks* **5**(2): 267–278 (March 1994).
27. S. Verdu, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **32**: 85–96 (Jan. 1986).
28. R. Lupas and S. Verdu, Near-far resistance of multiuser detectors in asynchronous channels, *IEEE Trans. Commun.* **38**: 496–508 (Apr. 1990).
29. S. Verdu, Computational complexity of optimum multiuser detection, *Algorithmica* **4**: 303–312 (1989).

30. M. K. Varanasi and B. Aazhang, Multistage detection in asynchronous code-division multiple access communications, *IEEE Trans. Commun.* **38**: 509–519 (Apr. 1990).

31. B.-P. Paris, B. Aazhang, and G. Orsak, Neural networks for multi-user detection in CDMA communication, *IEEE Trans. Commun.* **40**: 1212–1222 (July 1992).

32. U. Mitra and H. V. Poor, Adaptive receiver algorithms for near-far CDMA, in *PIMRC 92*, pp. 639–644, Oct. 1992.

33. G. I. Kechriotis and E. S. Manolakos, Hopfield neural network implementation of the optimal CDMA multiuser detector, *IEEE Trans. Neural Networks* **7**(1): 131–141 (Jan. 1996).

34. G. I. Kechriotis and E. S. Manolakos, A hybrid digital signal processing — neural network CDMA multiuser detection scheme, *IEEE Trans. Circuits Systems II* **43**(1,2): 96–104 (Feb. 1996).

35. A. Hiramatsu, ATM communications network control by neural networks, *IEEE Trans. Neural Networks* **1**(1): 122–130 (March 1990).

36. A. Hiramatsu, Integration of ATM call admission control and link capacity control by distributed neural networks, *IEEE J. Select. Areas Commun.* **9**: 1131–1138 (Sept. 1991).

37. R. Morris and B. Samadi, Neural network control of communications systems, *IEEE Trans. Neural Networks* **5**(4): 639–650 (July 1994).

38. S. Chong and J. Ghosh, Predictive dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM, *IEEE J. Select. Areas Commun.* **13**(1): 12–23 (Jan. 1995).

39. P. R. Chong and J. T. Hu, Optimal nonlinear adaptive prediction and modeling of MPEG video in ATM networks using pipelined recurrent neural networks, *IEEE J. Select. Areas Commun.* **15**(6): 1087–1100 (Aug. 1999).

40. A. Tarraf, I. Habib, and T. Saadawi, A novel neural network enforcement mechanism for ATM networks, *IEEE J. Select. Areas Commun.* **12**(6): 1088–1096 (Aug. 1994).

41. I. Habib, A. Tarraf, and T. Saadawi, A neural network controller for congestion control in ATM multiplexers, *Computer Networks ISDN Systems* **29**(3): 325–334 (Feb. 1997).

42. A. Tarraf, I. Habib, and T. Saadawi, Intelligent traffic control for ATM broadband networks, *IEEE Commun. Mag.* **33**(10): 76–82 (Oct. 1995).

43. T. Fritsch, Cellular mobile communication design using self-organizing feature maps, in Ben Yuhas and Nirwan Ansari, ed., *Neural Networks in Telecommunications*, Kluwer, Dordrecht, Netherlands, 1994, pp. 211–232.

44. X. M. Gao, X. Z. Gao, J. M. A. Tanskanen, and S. J. Ovaska, Power prediction in mobile communication systems using an OPtimal neural-network structure, *IEEE Trans. Neural Networks* **8**(6): 1446–1455 (Nov. 1997).

45. Y. Moreau and J. Vandewalle, Detection of mobile phone fraud using supervised neural networks: A first prototype, in *International Conference on Artificial Neural Networks 97*, Springer, 1997, pp. 1065–1070.

46. J. R. Dorronsoro, F. Ginel, C. Sánchez, and C. Santa Cruz, Neural fraud detection in credit card operations. *IEEE Trans. Neural Networks* **8**(4): 827–834 (July 1997).

47. T. X. Brown, Neural networks for switching, *IEEE Commun. Mag.* **27**(11): 72–81 (1989).

48. Y. Takefuji, *Neural Network Parallel Computing*, Kluwer Academic Publishers, Boston, 1992.

49. L. Wei and R. Chang, Broadcast scheduling in packet radio networks by Hopfield neural networks, *Information Processing Letters* **63**(5): 271–276 (Sept. 1997).

50. C. Giraud-Carrier and M. Ward, Learning customer profiles to generate cash over the internet, in *Proceedings of the Third International Workshop on Applications of Neural Networks to Telecommunications* (*IWANNT'97*), Lawrence Erlbaum Associates, Publishers, June 1997, pp. 165–170.

51. A. E. Henninger, A. J. Gonzalez, M. Georgiopoulos, and R. F. DeMara, A connectionist-symbolic approach to modeling agent behavior: Neural networks grouped by contexts, *Lecture Notes Comput. Sci.* **2116**: 198 (2001).

52. R. A. Cole, J. W. T. Inouye, Y. K. Muthusamy, and M. Gopalakrishnan, Language identification with neural networks: a feasibility study, in *IEEE Pacific RIM Conference on Communications, Computers and Signal Processing*, Victoria, Canada, June 1989, Piscataway, NJ, 1989. IEEE, pp. 525–529.

53. A. Paoloni, S. Ragazzini, and G. Ravaioli, Predictive neural networks in text independent speaker verification: an evaluation on the SIVA database, in *Proc. ICSLP '96*, Vol. 4, Philadelphia, PA, Oct. 1996, pp. 2423–2426.

54. P. D. Gader et al., Neural and fuzzy methods in handwriting recognition, *Computer* **30**(2): 79–86 (Feb. 1997).

55. R. Lancini, Image vector quantization by neural networks, in Ben Yuhas and Nirwan Ansari, ed., *Neural Networks in Telecommunications*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1994, pp. 287–303.

56. B. H. Juang, S. Y. Kung, and C. A. Camm, eds., *Neural Networks for Signal Processing: Proceedings of the 1991 IEEE Workshop*, IEEE Press, 1991.

57. A. Hyvärinen and E. Oja, Independent component analysis: algorithms and applications, *Neural Networks* **13**(4–5): 411–430 (2000).

58. T. Evgeniou and M. Pontil, Support vector machines: theory and applications, *Lecture Notes Comput. Sci.* **2049**: 249–259 (2001).

59. Steffen L. Lauritzen, *Graphical Models*, Clarendon Press, Oxford, UK, 1996.

60. R. Eckmiller, Towards learning retina implants for partial compensation of retinal degenerations, in Dan Lundh, Bjorn Olsson, and Ajit Narayanan, eds., *Biocomputing and Emergent Computation*, World Scientific, 1997, pp. 271–281.

NONLINEAR EFFECTS IN OPTICAL FIBERS

ANDREW R. CHRAPLYVY
Bell Laboratories
Lucent Technologies
Holmdel, New Jersey

1. NONLINEAR EFFECTS IN OPTICAL FIBERS

The field of nonlinear optics in silica optical fibers originated in the late 1960s–early 1970s [1]. Initially nonlinear effects in single-mode silica fibers were laboratory curiosities requiring powerful lasers for their observation. The

discovery of erbium-doped fiber amplifiers for the 1.5-μm wavelength region [2,3] fundamentally altered the lightwave communication landscape by ushering in the era of wavelength-division multiplexing (WDM) and elevating optical nonlinearities to a primary systems consideration.

Before the early 1990s long-haul high-speed digital lightwave systems typically transmitted one wavelength channel on each optical fiber. These signals required frequent (every 40–50 km) 3R regeneration (reamplify, retime, reshape) that was accomplished using optoelectronic regenerators. In principle many information channels, each at a separate wavelength, could be transmitted over a single fiber. This is known as *wavelength-division multiplexing* (WDM). However, at each regenerator site the WDM channels must be optically demultiplexed, individually regenerated, and then optically multiplexed onto the next fiber span. For a large number of channels, this is prohibitively expensive. The advent of erbium-doped fiber amplifiers, which provide broadband optical gain in the wavelength region of minimum loss of silica fibers (1.55 μm), eliminated this problem. The broad amplifier gain bandwidth (~35 nm) allows simultaneous amplification of many WDM channels, thereby eliminating the need for demultiplexing at each repeater site. By replacing 3R regenerators with optical amplifiers, the distance between optoelectronic signal regeneration was increased from about 50 km to hundreds or even thousands of kilometers. However these major benefits of amplifiers increase the effects of optical nonlinearities. WDM increases the optical power propagating through fibers, and replacing regenerators with amplifiers increases the distances between signal regeneration. Increased optical powers and longer interaction lengths magnify the effects of nonlinearities. In WDM systems with several tens of channels propagating several hundreds of kilometers between regenerator sites various optical nonlinearities can be easily observed even though the power in the individual wavelength channels is on the order of 1 mW.

A number of nonlinearities in silica fibers can impact amplified lightwave systems [4]. They fall into two general categories. Stimulated scattering such as stimulated Brillouin scattering and stimulated Raman scattering are interactions between optical signals and acoustic or molecular vibrations in the fiber. Although both processes can produce exponential optical gain, they are qualitatively very different and affect lightwave systems in different ways. The second category of nonlinearities arises from modulation of the refractive index of silica by intensity changes in the signal. This gives rise to nonlinearities such as *self-phase modulation*, whereby an optical signal alters its own phase and spectrum; *cross-phase modulation* in WDM systems, where one optical signal affects the phases and spectra of all other optical signals and vice versa; and *four-photon mixing*, whereby WDM signals interact to produce mixing sidebands (as in intermodulation distortion).

1.1. Stimulated Scattering

1.1.1. Stimulated Brillouin Scattering. *Stimulated Brillouin scattering* (SBS) is the interaction between light and acoustic waves. In optical fibers SBS has the lowest threshold power of all the nonlinearities [5]. Light scatters from acoustic phonons and is downshifted in frequency. The magnitude of the downshift depends on the scattering angle (varying from a zero-frequency shift for forward scattering to a maximum frequency shift in the backward direction). In single-mode silica fibers the only frequency-shifted scattered light that continues to be guided is backward-propagating light. The Brillouin shift for backward scattering in silica is about 11 GHz. This backscattered light experiences exponential gain due to the forward-propagating light. System impairment occurs when the backscattered light level becomes comparable to the signal power and begins to deplete the signal. For typical fibers the threshold power for this process is ~10 mW for single fiber spans and correspondingly lower for concatenated amplified spans. However, optical amplifiers usually have optical isolators (otherwise long amplified systems could easily optically oscillate), which prevent backward SBS light from propagating through multiple fiber spans. Consequently the SBS impairment in amplified systems occurs at the same power levels as in regenerated systems. In addition, SBS impairments are not exacerbated in WDM systems, because each signal channel interacts with acoustic phonons having slightly different frequencies. Thus the nonlinearities accumulate individually for each channel. However, some systems will require individual signal powers greater than 10 mW. For example, to overcome fiber attenuation in extremely long single-span systems, higher signal powers, which may exceed the SBS threshold, are required.

Although SBS has the lowest threshold of all the fiber nonlinearities, it is also the easiest nonlinearity to counteract because of the lifetime of the acoustic phonons that give rise SBS. The phonon lifetime in silica fibers is about 15 ns, which corresponds to an optical linewidth of about 20 MHz. Optical sources with linewidths greater than 20 MHz will experience reduced SBS. Typical laser diodes have linewidths less than 10 MHz, but dithering the laser injection current can artificially broaden the effective linewidths because it causes a dithering of the optical frequency of the signal. The SBS gain is then reduced by the ratio of the magnitude of the frequency dither divided by 20 MHz. It is easy to increase the SBS threshold by an order of magnitude simply by dithering the diode laser frequency over a 200-MHz range. Typically this corresponds to a dither current of about 0.2 mA. The dithering technique is now an industrywide standard whenever SBS is a nuisance.

1.1.2. Stimulated Raman Scattering. *Stimulated Raman scattering* (SRS) in optical fibers is the interaction between light and the vibrational modes of silica molecules in the core of the fiber. Although both SRS and SBS are examples of stimulated scattering, there are a number of key differences between the two nonlinearities. Unlike SBS, SRS is an extremely broadband effect. The Raman transitions in silica glass are very broad and overlap into a continuous-gain curve such as that shown in Fig. 1. Note that the peak SRS gain occurs at a frequency about 15 THz lower than the input signal. Unlike SBS, SRS occurs in both forward and backward directions. Isolators at amplifier sites will not diminish forward SRS, and the effect accumulates with the number of amplified fiber spans.

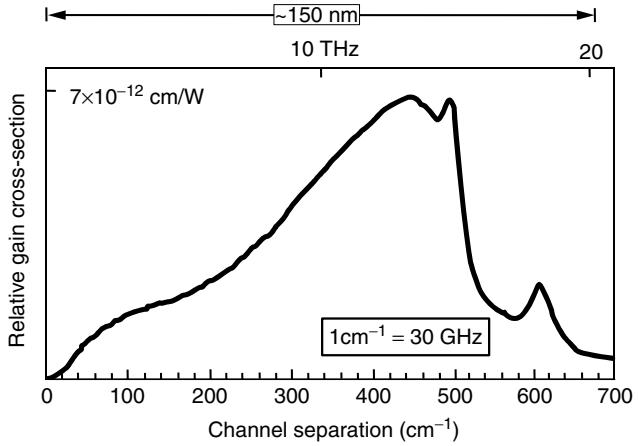


Figure 1. The relative SRS gain coefficient for fused-silica fibers at 1.5 μm.

In single-channel systems some of the spontaneously scattered Raman-shifted light, having amplitude given by the Raman gain curve in Fig. 1, will be guided in the core of the fiber in the forward direction. The copropagating signal light will amplify this scattered light. Because the peak SRS gain is over two orders of magnitude smaller than the peak SBS gain, significantly higher optical powers are required to exceed the SRS threshold for a single channel. The SRS threshold for single fiber spans is over 1 W and manifests itself by exponential amplification of wavelengths near the peak of the Raman gain curve about 15 THz lower in frequency than the signal. In amplified systems one might expect that the threshold is 1 W divided by the number of spans. However, the optical gain bandwidth of an erbium amplifier is roughly 4 times smaller than the bandwidth of the SRS gain profile. Consequently, only Raman light generated within the amplifier gain profile will propagate through the amplified chain of fibers. Since the SRS gain profile is roughly triangular, the peak Raman gain at 30 nm is roughly $\frac{1}{4}$th that of the maximum gain. It follows that in the worst case, the SRS threshold for an amplified chain will be about 4 W divided by the number of amplified spans between regenerators. SRS can be easily suppressed in single-channel amplified systems by periodically inserting optical bandpass filters that pass the signal and reject most of the SRS spectrum.

It is in WDM systems that SRS can be particularly vexing. Because the SRS gain is so broad, WDM channels will be coupled to each other for channels spaced up to 20 THz (150 nm). The short-wavelength channels will act as Raman pumps for long-wavelength channels [6,7]. The long-wavelength channels will be amplified at the expense of the short-wavelength channels, which will be attenuated. Impairments from such interactions will occur at powers much lower than 1 W. For example, for two channels separated by 15 THz (110 nm), unacceptable system degradations will occur at 50 mW in a single fiber span. For multiple channels and multiple spans the threshold powers for degradation will be proportionately smaller. Ultimately, SRS limits the number WDM channels that can be transmitted through single-mode fibers. The number of channels is inversely proportional to the overall transmission distance. Thus far the discussion is based on the assumption of continuous-wave (CW) power at all the signal wavelengths. However, digital systems typically transmit binary information in which a pulse of light represents a logical "one" (called a "mark") and absence of optical power represents a logical "zero" (called a "space"). A space can neither experience Raman gain from shorter wavelengths nor produce Raman gain for longer-wavelength signals. Only marks can be amplified or depleted by SRS, and the amount of gain or depletion will depend on the presence of marks in other channels. Since the occurrence of marks is a random process the amount of amplification or depletion for marks in a particular channel will vary from mark to mark. This is called *pattern-dependent SRS*. Marks with pattern-dependent amplitudes give rise to intersymbol interference (ISI), one of several types of degradations in digital systems. Pattern-dependent SRS is somewhat ameliorated by two effects: modulation statistics and chromatic dispersion. In a binary bit stream the probability of occurrence of marks and spaces is $\frac{1}{2}$. In a WDM system with many channels the occurrence of marks and spaces at a particular instant in time is a random variable. As the number of WDM channels increases, the pattern dependent variation of SRS decreases. This averaging effect is further increased due to chromatic dispersion; specifically, different wavelengths in an optical fiber propagate with different group velocities. Consequently a particular bit in a given channel "samples" multiple time slots (bits) of neighboring channels as it propagates through the fiber. This further diminishes pattern dependent SRS effects. Ultimately this leads to a very efficient method to combat SRS. Filters placed periodically along the fiber (conveniently located within the optical amplifiers themselves) with the inverse filter profile of the SRS gain curve in Fig. 1 can almost completely undo the effects of SRS; specifically, slightly more attenuation is provided to the long-wavelength channels than to the short-wavelength channels. This method of combating SRS is now routinely used in long WDM systems consisting of many channels.

In the early days of WDM systems, optical nonlinearities were viewed as detrimental phenomena that needed to be mitigated. With time clever applications of fiber nonlinearities have led to useful optical devices. An important example is the use of SRS to provide optical gain for the WDM signals. Injecting pump light of the appropriate wavelength(s) into a fiber can turn the fiber into a stimulated Raman amplifier; thus the transmission medium is also an amplification medium. In fact, systems can be designed so that the amount of optical gain produced by SRS exactly compensates the intrinsic attenuation of the transmission fiber. Such a system no longer requires discrete amplifiers every 80–100 km but requires only periodic injection of pump light. The noise figures of Raman-amplified systems are typically several decibels lower than the noise figures of conventional systems amplified by discrete amplifiers. Typically pump light is injected in the backward direction relative to the signals but there are examples of both counter and copropagating Raman pumping.

1.2. Nonlinear Refractive Index

The refractive indices of many optical materials are weakly intensity-dependent ($n = n_0 + n_2 I$). The intensity-dependent refractive index, n_2, of silica has a value of 2.6×10^{-20} m^2/W. Although the n_2 of silica is extremely small (many semiconductor materials have values of n_2 orders of magnitude larger than those of silica), in long amplified systems or in certain WDM systems the effects of the nonlinear refractive index can be quite prominent.

1.2.1. Self-Phase Modulation. *Self-phase modulation* (SPM) describes the effect of a pulse on its own phase [8]. The edge of an optical pulse represents a time-varying intensity. A time-varying intensity in a medium with an intensity-dependent refractive index will produce a time-varying refractive index, which, in turn, produces a time-varying phase that corresponds to a spectral broadening (Fig. 2). Therefore one of the consequences of the nonlinear refractive index of silica is that the spectral width of signal pulses will gradually increase as they propagate in a fiber. For example, a 1-mW pulse would exhibit a twofold broadening after propagating several thousand kilometers in an amplified system. However, a 10-mW pulse will experience the same spectral broadening in 10 times less interaction length.

Spectral broadening can degrade systems in several ways. In a densely spaced WDM system SPM can broaden the signals so that adjacent channels begin to partially overlap spectrally. This leads to optical crosstalk. Spectral broadening can also lead to pulseshape deformation. Because of the nonlinear refractive index, the peak of a pulse accumulates phase more quickly than the wings. This results in stretching of the wavelength on the leading edge of the pulse and compression on the trailing edge. Thus the trailing edge of a pulse acquires a blue shift and the leading edge acquires a red shift (Fig. 3). Recall that most fibers have finite chromatic dispersion. Thus two edges of a pulse will propagate at different speeds. In fibers

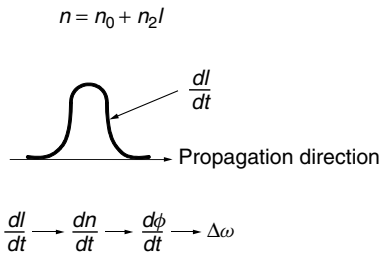


Figure 2. Source of spectral broadening due to nonlinear refractive index.

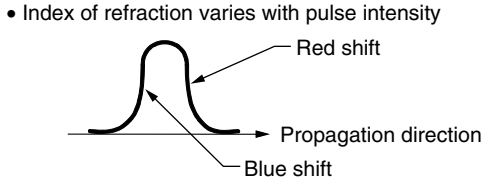


Figure 3. Effect of SPM on leading and trailing edges of a pulse.

with normal chromatic dispersion (red propagates faster than blue) the pulse will begin to temporally broaden. With sufficient temporal broadening pulses will begin to overlap neighboring time slots and interfere with neighboring bits. Bit errors can occur when a mark spreads into a neighboring space. In fibers with anomalous dispersion (blue propagates faster than red) the pulses initially become narrower in time. Ultimately the two edges pass through the center of the pulse, and with further propagation, the pulse will temporally broaden and bit errors will occur when pulses overlap neighboring time slots. System degradations due to the combined effects of SPM and chromatic dispersion are now mitigated by a technique known as *dispersion management*. An optical line system can be designed to consist of two different types of fibers, one having normal dispersion and the other having anomalous dispersion (the transmission fiber itself can be a concatenation of the two types of fiber, or in the more popular design there is one type of transmission fiber and the other type of fiber, called *dispersion compensating fiber*, is located within the optical amplifiers). The fibers are chosen so that the overall accumulated chromatic dispersion for the entire system is nearly zero. In such situations there is no net temporal broadening of the self-phase-modulated pulses, albeit during propagation the pulses "breathe" — broaden as a result of the dispersion of one fiber and then narrow to approximately the original widths due to the opposite-sign dispersion in the second type of fiber.

SPM, like SRS, is a nonlinearity that can be exploited to advantage. Even in the absence of nonlinearity (e.g., at very low powers), pulses propagating in fibers broaden as a result of chromatic dispersion. Each pulse intrinsically contains a spread of wavelengths determined by the pulsewidth and other factors. These wavelength components travel at different speeds, and this leads to pulse broadening. In anomalous dispersion fibers we have seen that the consequence of SPM is pulse narrowing. This tendency to narrow can exactly balance out the broadening due to linear chromatic dispersion and can produce pulses that do not change shape as they propagate. Such pulses are called *solitons* [4]. Soliton technology is now finding its way into commercial transmission systems.

1.2.2. Cross-Phase Modulation. In WDM systems the intensity variations in any signal channel will affect the phases of all the other signals [9]. The origin of this cross-phase modulation (CPM) is the same nonlinear refractive index that gives rise to SPM. If the fiber chromatic dispersion were zero for all channels (all pulses propagate in lock step), the effects of CPM due to each interfering channel would be exactly twice as strong as the SPM effect. However, there are no practical "dispersion-flattened" zero-dispersion fibers. Consequently, the group velocities of various channels in a WDM system are different and pulses in different channels will pass through each other while propagating in the fiber. Under some conditions these pulse collisions virtually eliminate spectral broadening due to CPM. Figure 4 schematically depicts pulses from two different channels passing through each other (the figure is shown in the reference frame of

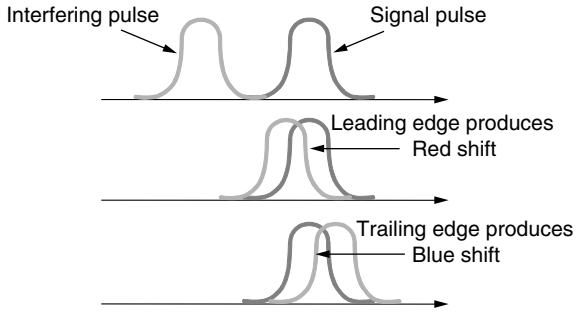


Figure 4. Effects of CPM on colliding pulses.

the signal pulse). Note that during the first half of the collision the interfering pulse produces a red shift in the signal pulse. In the second half of the collision the trailing edge of the interfering pulse produces a blue shift in the signal pulse. The blue shift exactly reverses the effects of the red shift if the intensities and shapes of the pulses have not significantly changed during the collision. Designing transmission systems around this cancellation effect is impractical for two main reasons:

1. A minimum wavelength spacing between neighboring channels is required for the pulse collisions to occur rapidly enough that their intensities do not appreciably change during the collision. This sacrifices precious spectral efficiency (channel bit rate divided by channel spacing) in many cases.
2. There is no way to avoid "partial" collisions. For example, two pulses in neighboring channels exit an optical amplifier partially overlapped. The leading-edge/trailing-edge symmetry is destroyed.

Partial collisions will frequency shift one part of a pulse relative to the remainder. This leads to different group velocities for different parts of a pulse. The occurrence of partial collisions is a random process depending on the presence or absence of marks in various channels. Consequently the arrival time of various marks in a particular signal channel will be random, leading to timing jitter at the receiver, another type of degradation in digital systems.

As with SPM, dispersion management can reduce the effects of CPM. Spectral broadening or frequency shifts do not give rise to dispersion penalties or timing jitter if the overall system dispersion is nearly zero.

1.2.3. Four-Photon Mixing. A third manifestation of the nonlinear refractive index in WDM systems is *four-photon mixing* (FPM). In the case of two signals (Fig. 5a) there exists an intensity modulation at the beat frequency that modulates the refractive index, producing a phase modulation at the difference frequency. This phase modulation creates two sidebands. These sidebands are called *two-tone products* because they were produced by the mixing of two signal waves. For three channels (Fig. 5b), in addition to the two-tone products created by each pair of signals, there are 3 three-tone products generated by all

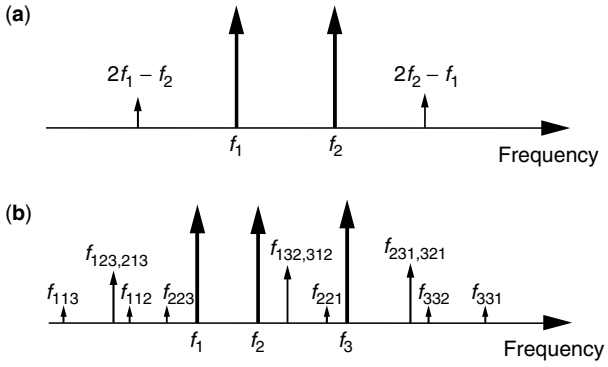


Figure 5. Phase modulation sidebands produced by FPM for (**a**) two signals and (**b**) three signals.

three signals (i.e., the beat frequency of one pair of signals produces sidebands on the third signal). Since there are two different ways of generating each three-tone electric field, the three-tone products are generated with four times the optical power of the two-tone products. For N channels there will be $N^2(N-1)/2$ mixing products generated. For example, in an 8-channel WDM system, 224 mixing products are generated.

Two different impairments are caused by FPM. The obvious and more benign degradation is depletion of signal power in creating the mixing products, especially in WDM systems with many channels. An even more serious degradation occurs if the signal channels are equally spaced. In this case many of the mixing products are produced with optical frequencies the same as the signal frequencies. These mixing products interfere coherently with the signals, either constructively or destructively depending on the (time-dependent) relative phases of the signals. Because it is the electric fields that interfere, small mixing products can produce severe degradations. For example, a mixing product with 1% of the power of one of the signals has an electric field magnitude 10% of the signal electric field and will produce 20% depletion in the signal channel if it destructively interferes with the signal.

To eliminate coherent mixing, a frequency allocation scheme for signals has been devised to ensure that no mixing product will have the same frequency as any signal [4]. This is accomplished by ensuring that the frequency spacing between any two signals is unique. This eliminates interference impairments and leaves only the less severe depletion effects with which to contend. The frequency allocation scheme requires control of signal frequencies to within several gigahertz, easily achievable with present transmitter lasers.

Since FPM is a phase-matched process, it depends strongly on chromatic dispersion [4]. In general, the intensity modulation generated by the beating between two or three signals propagates at a different speed than the signals themselves. If the difference in propagation velocities is large (large chromatic dispersion), the FPM generation efficiency becomes small (poor phase matching) and system degradations are inconsequential. In fibers with zero chromatic dispersion near the signal wavelengths, FPM is a very efficient nonlinear process and dramatic system degradations can occur in relatively short (≅20-km)

lengths of fiber. Dispersion management was initially invented in 1993 to combat the effects of FPM in high-speed systems that were sensitive to chromatic dispersion. The high local dispersion in any of the fiber segments suppresses FPM, but the overall dispersion is nearly zero, to avoid dispersion penalties for high-bit-rate signals. Subsequently dispersion management proved to be a useful technique in counteracting SPM and CPM effects.

FPM can also be exploited for useful purposes, namely, for parametric optical amplification [10]. Signals can be injected into a low-dispersion optical fiber along with a strong pump at a wavelength near the zero-dispersion wavelength of the fiber. FPM between the pump and the signals will produce strong FPM products located in frequency-reversed order (phase conjugation) at wavelengths on the opposite side relative to the pump. The noise figure (NF) of parametric amplifiers can be significantly smaller than the NF of conventional erbium-doped fiber amplifiers, and phase conjugation of amplified signals also reverses the effects of chromatic dispersion. In principle, a system based on parametric amplification would not need dispersion management.

2. CONCLUSION

In conclusion, silica optical fibers exhibit a rich collection of optical nonlinearities that have become important with the advent of practical optical amplifiers and the ultra-long-haul lightwave systems they enabled. Since the early 1990s an arsenal of techniques have been developed to mitigate the effects of nonlinearities. These techniques include dispersion management, frequency or phase modulation of sources, and optical filtering. However, the most tangible impact of optical nonlinearities is to provide gainful employment for systems engineers trying to maximize the ultimate information-carrying capacity of optical fibers by either counteracting or exploiting nonlinear effects in fibers. More advanced treatment of optical nonlinearities in fibers can be found in Chapter 8 of the article by Stolen and Lin [4] and references cited therein.

BIOGRAPHY

Andrew R. Chraplyvy received the B.S. degree in physics in 1972 from Washington University, St. Louis, Missouri, and the M.S. and Ph.D. degrees in physics from Cornell University in 1975 and 1977, respectively. He joined the Physics Department at General Motors Research Labs in 1977 as a Research Scientist. At GM he worked on ultra-high-resolution spectroscopy of gases and impurity modes in solids. Since 1980, he has been with Bell Laboratories, where he currently is Director of Lightwave Systems Research. Dr. Chraplyvy holds over 25 patents in the areas of lightwave systems and fiberoptics. He is the recipient of the 1999 Thomas Alva Edison Patent Award and the 1999 New Jersey Inventor of the Year Award. He is a Bell Labs Fellow, a member of the National Academy of Engineering and a Fellow of the Optical Society of America. His areas of interest are fiberoptics, lightwave communications systems, nonlinear optical interactions in fibers, fiber networks, and high-resolution spectroscopy of gases and solids.

BIBLIOGRAPHY

1. R. H. Stolen, E. P. Ippen, and A. R. Tynes, Raman oscillation in glass optical waveguide, *Appl. Phys. Lett.* **20**: 62–64 (1972).
2. R. J. Mears, L. Reekie, I. M. Jauncey, and D. N. Payne, Low-noise erbium-doped fiber amplifier operating at 1.54 μm, *Electron. Lett.* **23**: 1026–1027 (1987).
3. E. Desurvire, J. R. Simpson, and P. C. Becker, High-gain erbium-doped traveling-wave fiber amplifier, *Opt. Lett.* **12**: 888–890 (1987).
4. I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications IIIA*, Academic Press, San Diego, 1997.
5. D. Cotter, Observation of stimulated Brillouin scattering in low-loss silica fiber at 1.3 μm, *Electron. Lett.* **18**: 495–496 (1982).
6. A. R. Chraplyvy and P. S. Henry, Performance degradation due to stimulated Raman scattering in wavelength-division-multiplexed optical-fiber systems, *Electron. Lett.* **19**: 641–642 (1983).
7. A. R. Chraplyvy, Optical power limits in multichannel wavelength-division-multiplexed systems due to stimulated Raman scattering, *Electron. Lett.* **20**: 58–59 (1984).
8. R. H. Stolen and C. Lin, Self-phase modulation in silica optical fibers, *Phys. Rev. A* **17**: 1448–1453 (1978).
9. A. R. Chraplyvy and J. Stone, Measurement of crossphase modulation in coherent wavelength-division multiplexing using injection lasers, *Electron. Lett.* **20**: 996–997 (1984).
10. R. H. Stolen and J. Bjorkholm, Parametric amplification and frequency conversion in optical fibers, *IEEE J. Quantum Electron.* **QE-18**: 1062–1072 (1982).

NONUNIFORMLY SPACED TAPPED-DELAY-LINE EQUALIZERS FOR SPARSE MULTIPATH CHANNELS

FREDERICK K. H. LEE
PETER J. MCLANE
Queen's University
Kingston, Ontario, Canada

1. INTRODUCTION

The high data rate requirement in current and future broadband wireless communication systems has created a new regime of interesting and challenging problems for communication engineers of the new century. Of major concern in the physical layer is the growth of the lengths of the sampled channel impulse responses as a function of the transmission rate when measured in units of symbol intervals. Most commonly used equalization techniques for suppressing intersymbol interference (ISI) distortion caused by the multipath propagation phenomenon, such as the tapped-delay-line (TDL) equalizers, including the linear equalizers (LEs) and the decision feedback equalizers

(DFEs), and the maximum-likelihood sequence estimators (MLSEs), all exhibit structural and computational complexities that depend on the sampled channel lengths. An increase in transmission rate thus inevitably leads to an increase in complexity of these equalization techniques, which is particularly problematic for mobile receivers where resources are scarce because of constraints in cost, size, and battery power. While an impulse response can certainly be truncated to reduce its length, this would be undesirable if its tail portion contains a significant amount of energy. This is precisely the dilemma facing the family of wireless channels called *sparse multipath channels*.

2. SPARSE MULTIPATH CHANNELS

A sparse multipath channel is characterized by an impulse response consisting of only a few dominant multipath terms, but with any two terms separated by a large time delay, resulting in a long delay spread. Examples of sparse multipath channels include terrestrial broadcasting channels, such as those found in high-definition television (HDTV) systems; horizontal or vertical underwater acoustic channels, where reflections off the sea surface and the sea floor constitute the two main causes for the long reverberation of the multipath terms [1, Chap. 8]; as well as cellular land mobile radio channels encountered in hilly terrain environments. Combined with a high data rate, the lengths of the sampled channel impulse responses can range from several tens to hundreds of symbol intervals, depending on the specific application. For instance, in the proposed North American HDTV terrestrial broadcasting mode, 64-QAM (quadrature amplitude modulation) is used and the transmission rate is 5.38 Msps (megasymbols per second). With a typical delay spread of 20 μs, the sampled channel length spans 107.6 symbol intervals, but only a small number of the channel taps exhibit large magnitude due to the sparse nature of the channel [2]. An additional feature of these HDTV channels is the existence of strong precursor taps known as "ghost" signals, which adds to the difficulty of equalization. As another example, Fig. 1 shows the impulse response of a sparse underwater acoustic channel with delay spread that exceeds 80 ms. Hence, even with a modest transmission rate of 1.25 ksps, the sampled channel length is at least a hundred symbol intervals long.

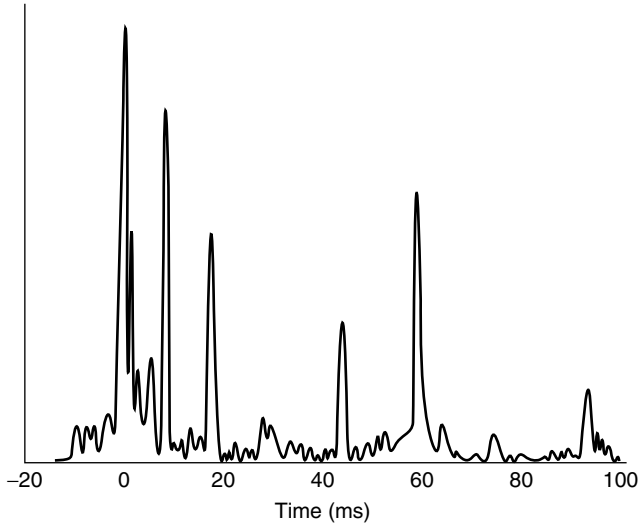


Figure 1. Impulse response of a sparse underwater acoustic channel. (*Source*: Ref. 7, Fig. 1(b).)

The description above clearly explains why existing equalization techniques cannot efficiently mitigate the ISI distortion intrinsic to a sparse multipath channel. Fortunately, as the majority of taps in a sampled sparse multipath channel contain zero or near-zero values, it is possible to develop new equalization methods with complexity associated with the number of large-magnitude taps instead of the entire channel length. One such feasibility is by using nonuniformly spaced TDL equalizers (NU-Es), which is the focus of this article. Interested readers can refer to the Further Reading list for references to other solutions that exploit the structure of sparse multipath channels.

3. NONUNIFORMLY SPACED TDL EQUALIZERS (NU-Es)

The distinguishing element of a NU-E is its variable spacings between taps, as opposed to fixed spacings in a uniformly-spaced TDL equalizer (U-E). For all practical purposes, however, a NU-E can be viewed as a U-E with a large number of zero-valued taps, as depicted in Fig. 2, since the spacings of the TDL in a NU-E are usually predetermined by a fixed-rate sampler. In other words, designing a NU-E is equivalent to choosing the best set of tap positions on a fixed-spaced TDL. There are a number

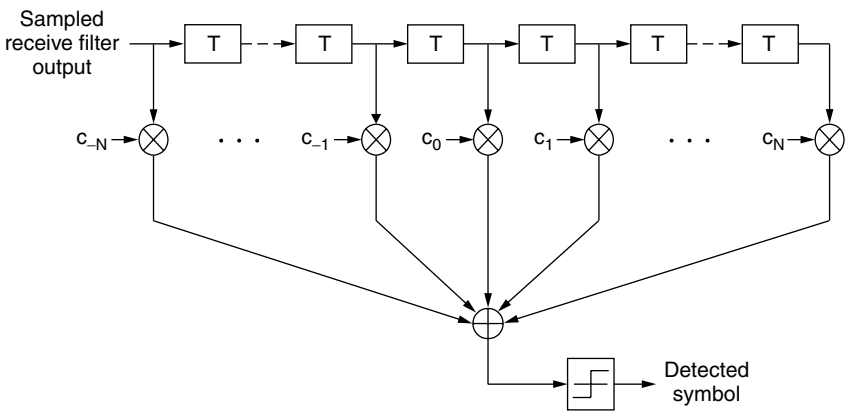


Figure 2. A T-spaced U-LE with $(2N + 1)$ taps. A T-spaced NU-LE can be viewed as a T-spaced U-LE with a large number of zero-valued taps.

of advantages to using a NU-E. First, an appropriately designed NU-E can achieve practically the same performance as a U-E but with fewer taps, hence reducing the computational load of the receiver. This is especially true for equalizing sparse multipath channels, where the number of taps required in a NU-E is proportional to the number of channel taps with large magnitude. In an adaptive NU-E, fewer taps implies faster convergence of the initial learning curve to the optimum tap values, which allows a shorter training sequence to be transmitted. The influence of noisy estimates at the output also tends to lessen without the need to train or track the near-zero-valued taps. For one case with a fractionally spaced TDL in a fixed-point implementation, tracking only a small set of taps decreases the degrees of freedom and thus minimizes the occurrence of the tap wandering phenomenon, an event that arises when tap values deviate from their exact optimum values due to noisy estimates and eventually accumulate to values that are too large and cause overflow.

Despite the many benefits of a NU-E, optimizing its tap positions is not an easy task. Closed-form solution to the optimum tap positions of a finite-length NU-E does not exist in general. Currently, two different approaches are available in the literature for finding these optimum tap positions. The first one, suggested by Raghavan et al. [3], uses a branch-and-bound algorithm to exhaustively search for the best combination of tap positions within a given span of the TDL. The second one, developed by Lee [4], involves numerically solving a set of nonlinear equations to obtain a NU-DFE with tap spacings and tap values that are locally optimum. Note that the normal convention of a fixed-space TDL is removed in the derivation of the nonlinear equations, which means that the resultant tap spacings in the feedforward filter (FFF) can be any real number. Unfortunately, due to the heavy computational burden and the long processing time, both methods are deemed unsuitable for real-time applications, and their usage is mainly restricted to off-line analysis for benchmarking purposes.

To circumvent the difficulty of finding the optimum tap positions, a number of suboptimum tap allocation schemes have also been proposed in the literature. A simple one is the strategy of thresholding, where taps of a U-E are first determined and only those with magnitude above a threshold are retained. Despite its simplicity, this method requires initial training of a large number of taps, which is inefficient. Moreover, the tap values are suboptimum with respect to the retained tap positions, though this problem can easily be corrected by re-optimizing the tap values after thresholding is completed, albeit at a further sacrifice of efficiency. Another easy method is to choose the positions of the channel taps with large magnitude as the positions of a NU-E, as has been adopted by Kocic et al. [5] to obtain a NU-DFE. Other more elaborate solutions include the one by Ariyavisitakul et al. [6], where the FFF of a NU-DFE is designed by selecting the set of taps that maximizes a simplified expression of the output signal-to-noise ratio (SNR), as well as an automated exchange-type algorithm by Lopez et al. [7] that allocates taps to the FFF and the feedback filter (FBF) of a NU-DFE alternatively in an iterative fashion until a desired level of performance is reached.

Similar to the optimum algorithms, the abovementioned suboptimum tap allocation schemes all share the common trait of attempting to find the tap positions of a finite-length NU-E directly. However, this is not the only way to tackle the problem. Given a sparse multipath channel, it turns out that certain infinite-length equalizers are inherently nonuniformly spaced, with the sparseness of those equalizers intimately related to the sparseness of the channel. Consequently, a logical alternative to designing finite-length NU-Es is to first identify the tap positions of an infinite-length NU-E, and then assign a subset of those positions to a finite-length NU-E. This design methodology is originally recognized by Geller et al. [8] and again by Berberidis and Rontogiannis [9], both using the infinite-length zero-forcing (ZF) LEs, and is later extended for designing infinite-length minimum mean-square error (MMSE) LEs and ZF/MMSE-DFEs by Lee and McLane [10]. The beauty of this approach lies in the application of the classical equalization theory to derive the infinite-length NU-Es, thus formally unifying the NU-Es with the well-known U-Es. In addition, provided that an infinite-length equalizer is nonuniformly spaced, the corresponding finite-length NU-E using its tap positions will be asymptotically optimum. Because of these advantages, this methodology is chosen over other existing techniques for designing finite-length NU-Es in this article. Details of the design process are described in the next two sections.

4. INFINITE-LENGTH SYMBOL-SPACED EQUALIZERS FOR SPARSE MULTIPATH CHANNELS

To begin, the infinite-length, symbol-spaced (T-spaced) LE and DFE under the ZF and the MMSE criteria are derived for any given sparse multipath channel. The goal is to determine which of these equalizers are nonuniformly spaced and, if so, their tap positions. Consider the baseband communication system shown in Fig. 3. The data source is an independent and identically distributed (iid) sequence where symbols can be taken from any QAM signaling scheme, and the impulse response of the transmit filter is a square-root raised-cosine (SRRC) pulse. The impulse response of a causal, M-ary sparse multipath channel is given by $h_c(t) = \sum_{i=0}^{M-1} a_i \delta(t - \tau_i T)$, where $\{\tau_i\}$ are restricted to be nonnegative integers

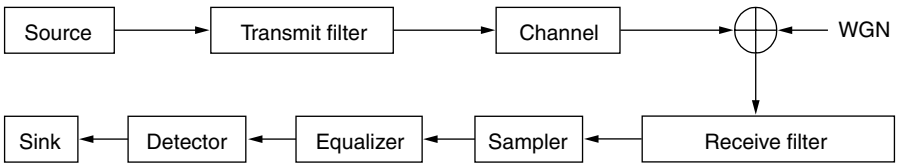


Figure 3. Block diagram of a baseband communication system model. (*Source*: Ref. 10.)

with $\tau_0 = 0$. For brevity, the channel is also denoted as $h_c = [a_0, \ldots, a_{\tau_i}, \ldots, a_{\tau_{M-1}}]$. Complex, zero-mean, white Gaussian noise $n(t)$ with two-sided power spectral density N_0 W/Hz is introduced at the output of the channel, and the noise process is assumed to be independent of the data sequence. At the receiving end, a matched filter (MF) and a SRRC receive filter are used alternatively. With T-spaced sampling at time instants $\{kT\}$, $k = 0, \pm 1, \ldots$, the sampled MF output is $f_s(t) = \sum_{i=-N}^{N} f_i \delta(t - \mu_i T)$, where $N \geq M - 1$ and $\{\mu_i\}$ are integers with $\mu_0 = 0$ and $\mu_i = \mu_{-i}$, and $f_s(t)$ remains a sparse impulse response. If the SRRC receive filter is used, the sampled output is simply $h_c(t)$.

4.1. Linear Equalizers (LEs)

According to classical equalization theory [11,12], the frequency response of the infinite-length, T-spaced ZF-LE under the MF system is

$$C_{\text{ZF-LE,MF}}(w) = \frac{1}{F_s(w)} \tag{1}$$

where $F_s(w) = \sum_{i=-N}^{N} f_i e^{-jw\mu_i T}$ is the (periodic) frequency response of $f_s(t)$. After some simple manipulations, Eq. (1) becomes

$$C_{\text{ZF-LE,MF}}(w) = \frac{k}{1 + \sum_{i=-N(i\neq 0)}^{N} g_i e^{-jw\mu_i T}} \tag{2}$$

where $k = 1/f_0$ and $g_i = k f_i (i \neq 0)$. Let $S(w) = \sum_{i=-N(i\neq 0)}^{N} g_i e^{-jw\mu_i T}$. If $|S(w)| < 1$, Eq. (2) can be expressed as

$$C_{\text{ZF-LE,MF}}(w) = k\left[1 + \sum_{n=1}^{\infty} (-1)^n S^n(w)\right] \tag{3}$$

by making use of the infinite series $1/1 + x = 1 - x + x^2 - x^3 + \cdots$ for $|x| < 1$. Expanding $S^n(w)$ term by term for each n and taking the inverse Fourier transform gives

$$c_{\text{ZF-LE,MF}}(t) = k\left[\delta(t) - \sum_{\substack{i=-N \\ i\neq 0}}^{N} g_i \delta(t - \mu_i T) + \sum_{\substack{i=-N \\ i\neq 0}}^{N} \sum_{\substack{j=-N \\ j\neq 0}}^{N} \times g_i g_j \delta\left(t - (\mu_i + \mu_j)T\right) - \cdots\right] \tag{4}$$

which shows that the tap positions of the ZF-LE are given by the nonnegative-integer-based linear combinations of the multipath delays of $f_s(t)$. Hence, the infinite-length, T-spaced ZF-LE under the MF system is nonuniformly spaced. In the sequel, this derivation method will be called the *infinite-series approach*.

The infinite-length, T-spaced MMSE-LE under the MF system can be determined similarly by using the frequency response

$$C_{\text{MMSE-LE,MF}}(w) = \frac{\sigma_I^2}{\sigma_I^2 F_s(w) + N_0} \tag{5}$$

where σ_I^2 denotes the variance of the data symbols. It is easy to verify that $c_{\text{MMSE-LE,MF}}(t)$ is exactly the same as $c_{\text{ZF-LE,MF}}(t)$ in Eq. (4) except $k = \sigma_I^2/(\sigma_I^2 f_0 + N_0)$, which implies that the ZF-LE and the MMSE-LE share the same tap positions. Hence, the infinite-length, T-spaced MMSE-LE under the MF system is also nonuniformly spaced, and it differs from the ZF-LE only in terms of their tap values. As $N_0 \to 0$, the values of k become the same for both equalizers, and the MMSE-LE converges to the ZF-LE, a well-known result. It is also interesting to note that N_0 actually helps satisfy the condition $|S(w)| < 1$ for the frequency response of the MMSE-LE to be written as an infinite series.

When the SRRC receive filter is used, the frequency response of the infinite-length, T-spaced ZF-LE is

$$C_{\text{ZF-LE,SRRC}}(w) = \frac{1}{H_c(w)} \tag{6}$$

where $H_c(w) = \sum_{i=0}^{M-1} a_i e^{-jw\tau_i T}$ is the frequency response of $h_c(t)$. Let τ_d denote the delay that corresponds to the multipath term with the largest magnitude. Equation (6) can then be expressed as

$$C_{\text{ZF-LE,SRRC}}(w) = \frac{1}{e^{-jw\tau_d T}[a_d + \sum_{i=0(i\neq d)}^{M-1} a_i e^{-jw(\tau_i - \tau_d)T}]} = \frac{k' e^{jw\tau_d T}}{1 + \sum_{i=0(i\neq d)}^{M-1} b_i e^{-jw\nu_i T}} \tag{7}$$

where $k' = 1/a_d$, $b_i = k' a_i$, and $\nu_i = \tau_i - \tau_d$ $(i \neq d)$. Using the infinite-series approach gives

$$c_{\text{ZF-LE,SRRC}}(t) = k'\left[\delta(t + \tau_d T) - \sum_{\substack{i=0 \\ i\neq d}}^{M-1} b_i \delta(t - (\nu_i - \tau_d)T) + \sum_{\substack{i=0 \\ i\neq d}}^{M-1} \sum_{\substack{j=0 \\ j\neq d}}^{M-1} b_i b_j \delta\left(t - (\nu_i + \nu_j - \tau_d)T\right) - \cdots\right] \tag{8}$$

which shows that the infinite-length, T-spaced ZF-LE under the SRRC receive filter system is nonuniformly spaced as well. In essence, its tap positions are obtained by time-shifting the nonnegative-integer-based linear combinations of the time-shifted multipath delays of $h_c(t)$, where the amount of time shift in both instances is τ_d. Note that for the MF system, the multipath term in $f_s(t)$ with

the largest magnitude is always found at $t = 0$. Hence, the time-shifting operations are not required.

For the infinite-length, T-spaced MMSE-LE, its impulse response under the SRRC receive filter system is given by

$$c_{\text{MMSE-LE,SRRC}}(t) = c_{\text{MMSE-LE,MF}}(t) \otimes h_c^*(-t)$$
$$= k \sum_{l=0}^{M-1} a_l^* \left[\delta(t + \tau_l T) - \sum_{\substack{i=-N \\ i\neq 0}}^{N} g_i \delta(t - (\mu_i - \tau_l)T) \right.$$
$$\left. + \sum_{\substack{i=-N \\ i\neq 0}}^{N} \sum_{\substack{j=-N \\ j\neq 0}}^{N} g_i g_j \delta(t - (\mu_i + \mu_j - \tau_l)T) - \cdots \right] \quad (9)$$

where $\otimes$ denotes convolution. This shows that the impulse response of the infinite-length, T-spaced MMSE-LE under the SRRC receive filter system is the sum of M scaled and time-shifted replicas of $c_{\text{MMSE-LE,MF}}(t)$, where the scale factors and delays are determined by $h_c^*(-t)$. In fact, $c_{\text{ZF-LE,SRRC}}(t)$ can also be expressed in the form of (9) with k, $\{g_i\}$ and $\{\mu_i\}$ being the variables of $c_{\text{ZF-LE,MF}}(t)$. Therefore, analogous to the MF system, the tap positions of the ZF-LE and the MMSE-LE under the SRRC receive filter system are identical, and they can be determined from either (8) or (9), even though the two equations appear to be different.

4.2. Decision Feedback Equalizers (DFEs)

To determine the infinite-length, T-spaced ZF-DFE under the MF system, the minimum-phase spectral factorization of $F_s(w)$, namely, $F_s(w) = A_P|P(w)|^2$, where A_P is a constant and $P(w)$ is the monic, causal, and minimum-phase (canonical) factor, is utilized. It is well known that [11] the FFF of the infinite-length, T-spaced ZF-DFE is a whitening filter with frequency response

$$C_{\text{ZF-FFF,MF}}(w) = \frac{1}{A_P P^*(w)} \quad (10)$$

which implies that the tap positions of $c_{\text{ZF-FFF,MF}}(t)$ can be obtained by invoking the infinite-series approach and are the nonnegative-integer-based linear combinations of the multipath delays of $p^*(-t)$. For the FBF, its impulse response is $c_{\text{ZF-FBF,MF}}(t) = p(t) - \delta(t)$, which means that its tap positions are merely the multipath delays of $p(t)$. It is obvious from these two relationships that the sparseness of the infinite-length, T-spaced ZF-DFE is directly dependent on the sparseness of $p(t)$. In fact, $p(t)$ is a sparse impulse response if and only if $h_c(t)$ is sparse and minimum-phase or maximum-phase. If $h_c(t)$ is a mixed-phase channel, $p(t)$ is nonsparse even if $h_c(t)$ is sparse [10]. Therefore, the infinite-length, T-spaced ZF-DFE under the MF system is nonuniformly spaced if and only if $h_c(t)$ is a sparse minimum-phase or maximum-phase channel.

By exploiting the minimum-phase spectral factorization of the received signal-plus-noise spectrum, i.e., $\sigma_I^2 F_s(w) + N_0 = A_Q|Q(w)|^2$, where A_Q is a constant and $Q(w)$ is the canonical factor, the infinite-length, T-spaced MMSE-DFE under the MF system can be derived in the same fashion as for the ZF-DFE. However, because of the noise term N_0, $q(t)$, the inverse Fourier transform of $Q(w)$, is nonsparse for any type of channel in general. As a result, the infinite-length, T-spaced MMSE-DFE under the MF system is uniformly spaced.

When the SRRC receive filter is used, the impulse response of the FFF of the infinite-length, T-spaced ZF-DFE is given by $c_{\text{ZF-FFF,SRRC}}(t) = c_{\text{ZF-FFF,MF}}(t) \otimes h_c^*(-t)$. Hence, similar to the ZF-LE under the SRRC receive filter system, $c_{\text{ZF-FFF,SRRC}}(t)$ is the sum of M scaled and time-shifted replicas of $c_{\text{ZF-FFF,MF}}(t)$ with the scale factors and delays determined by $h_c^*(-t)$. Note that if $h_c(t)$ is a minimum-phase channel, $c_{\text{ZF-FFF,SRRC}}(t)$ should reduce to a scalar. On the other hand, the FBF of the infinite-length, T-spaced ZF-DFE is identical to its counterpart under the MF system: $c_{\text{ZF-FBF,SRRC}}(t) = c_{\text{ZF-FBF,MF}}(t)$. Once again, both relationships hold true for the infinite-length, T-spaced MMSE-DFE. Therefore, conclusions regarding the sparseness of the both DFEs are the same as those under the MF system.

5. FINITE-LENGTH NU-Es FOR SPARSE MULTIPATH CHANNELS

Having derived the various infinite-length, T-spaced equalizers for sparse multipath channels, the next step is to exploit these results for designing finite-length NU-Es. Only the MMSE criterion is considered here, as MMSE equalizers are known to have better performance than ZF equalizers. The system model, notations, and assumptions of Section 4 remain unchanged throughout this section, except that the restriction on the channel will be relaxed later to allow its multipath delays to take on any real number.

For a finite-length, T-spaced NU-LE (i.e., a NU-LE implemented on a T-spaced TDL), its tap positions are obtained directly from its infinite-length counterpart as indicated in Eq. (4) or (8) for the MF system or the SRRC receive filter system, respectively. As a result, as long as the condition for expressing the frequency response of the channel as an infinite-series is valid (i.e., $|S(w)| < 1$), the finite-length, T-spaced NU-LE is asymptotically optimum. Priority is given to the low-order positions, as they correspond to taps with large magnitude. (In the tap allocation algorithms outlined below, this is achieved by choosing small positive values for the coefficients $\{r_i\}$ in determining the nonnegative-integer-based linear combinations.) However, the design procedure for a finite-length, T-spaced NU-DFE is not as straightforward, since its infinite-length counterpart is uniformly spaced and thus provides no useful information on how to select its tap positions. Fortunately, a simple suboptimum strategy can be employed for the MF system, which uses the tap positions on the anticausal side of a NU-LE as the tap positions for the NU-FFF. The NU-FFF under the SRRC receive filter system can then be designed by invoking the time-shifting property on the NU-FFF under the MF system. Once the NU-FFF is fixed, the tap positions of the NU-FBF can be obtained easily by taking the strictly causal portion of the convolution result between the impulse response at the receive filter

output and the impulse response of the NU-FFF. The tap allocation algorithms based on these principles are given below, one for each receive filter system. As an illustration on how to use the algorithms, the tap positions of the NU-LEs for a simple, artificial sparse multipath channel are listed in Table 1.

Algorithm 1 (MF System)

1. To design a finite-length NU-LE:
   a. Identify the positive multipath delays of the impulse response $h(t) = h_c(t) \otimes h_c^*(-t)$ and denote them by the set $\mathcal{S}_{\text{LE}}^1$.
   b. Assign taps at the positions that are the nonnegative-integer-based linear combinations of the elements in $\mathcal{S}_{\text{LE}}^1$ and denote them by $\mathcal{S}_{\text{LE}}^2$. Mathematically, $\mathcal{S}_{\text{LE}}^2 = \{\sum_{i=1}^{s_{\text{num}}} r_i s_i \mid r_i \in \mathcal{Z}^+, s_i \in \mathcal{S}_{\text{LE}}^1, s_{\text{num}} = |S_{\text{LE}}^1|\}$.
   c. Assign taps at the set of positions denoted by $\mathcal{S}_{\text{LE}}^3$, where $\mathcal{S}_{\text{LE}}^3 = \{-s \mid s \in \mathcal{S}_{\text{LE}}^2\}$.
2. To design the FFF of a finite-length NU-DFE:
   a. Assign taps at the set of positions denoted by $\mathcal{S}_{\text{FFF}}$, where $\mathcal{S}_{\text{FFF}} = \mathcal{S}_{\text{LE}}^3$.
3. To design the FBF of a finite-length NU-DFE:
   a. Assign taps at the set of positions denoted by $\mathcal{S}_{\text{FBF}}$, where $\mathcal{S}_{\text{FBF}} = \{s_1 - |s_2| \mid s_1 \in \mathcal{S}_{\text{LE}}^1, s_2 \in \mathcal{S}_{\text{FFF}}, s_1 - |s_2| > 0\}$.

Algorithm 2 (SRRC receive filter system)

1. To design a finite-length NU-LE:
   a. Identify the multipath delays of $h_c(t)$ and denote them by $\mathcal{T}_{\text{LE}}^1$.
   b. Denote as x the delay of $h_c(t)$ that corresponds to the multipath term with the largest magnitude.
   c. Define a new set of positions $\mathcal{T}_{\text{LE}}^2$ by renaming the positions in $\mathcal{T}_{\text{LE}}^1$ relative to x as follows: $\mathcal{T}_{\text{LE}}^2 = \{t - x \mid t \in \mathcal{T}_{\text{LE}}^1\}$.
   d. Define the set $\mathcal{T}_{\text{LE}}^3$ as the nonnegative-integer-based linear combinations of the elements in $\mathcal{T}_{\text{LE}}^2$. Mathematically, $\mathcal{T}_{\text{LE}}^3 = \{\sum_{i=1}^{t_{\text{num}}} r_i t_i \mid r_i \in \mathcal{Z}^+, t_i \in \mathcal{T}_{\text{LE}}^2, t_{\text{num}} = |\mathcal{T}_{\text{LE}}^2|\}$.
   e. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{LE}}^4$, where $\mathcal{T}_{\text{LE}}^4 = \{t - x \mid t \in \mathcal{T}_{\text{LE}}^3\}$.
2. To design the FFF of a finite-length NU-DFE:
   a. Define x as in 1(b). However, if there exists one or more multipath terms with magnitude comparable to that of the largest magnitude term, denote x as the one closest to position 0.
   b. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{FFF}}^1$, where $\mathcal{T}_{\text{FFF}}^1 = \{s - x \mid s \in \mathcal{S}_{\text{FFF}}\}$.
   c. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{FFF}}^2$, where $\mathcal{T}_{\text{FFF}}^2 = \{-t \mid t \in \mathcal{T}_{\text{LE}}^1, t < x\}$.
3. To design the FBF of a finite-length NU-DFE:
   a. Assign taps at the set of positions denoted by $\mathcal{T}_{\text{FBF}}$, where $\mathcal{T}_{\text{FBF}} = \{t_1 - |t_2| \mid t_1 \in \mathcal{T}_{\text{LE}}^1, t_2 \in \mathcal{T}_{\text{FFF}}^1 \cup \mathcal{T}_{\text{FFF}}^2, t_1 - |t_2| > 0\}$.

Table 1. Tap Positions of NU-LEs for Equalizing $\boldsymbol{h_{c_1} = [a_0, a_3, a_5]}$, Where $\boldsymbol{a_0 = 1}$, $\boldsymbol{a_3 = 0.2828 + j0.2828}$, and $\boldsymbol{a_5 = 0.1299 + j0.075}$[a]

13-tap NU-LE (MF)	$\mathcal{S}_{\text{LE}}^2 = \{0, \mathbf{2}, \mathbf{3}, \mathbf{5}, 6, 8, 9\}$; $\mathcal{S}_{\text{LE}}^3 = -\mathcal{S}_{\text{LE}}^2$
6-tap NU-LE (SRRC)	$\mathcal{T}_{\text{LE}}^4 = \{0, \mathbf{3}, \mathbf{5}, 6, 8, 9\}$

[a]The positions in bold are elements of $\mathcal{S}_{\text{LE}}^1$ or $\mathcal{T}_{\text{LE}}^1$, which are used to form the nonnegative-integer-based linear combinations to obtain the other positions.
Source: Ref. 10.

As stated in Section 3, closed-form solution to the optimum tap positions of a finite-length NU-E generally does not exist. However, for a special type of sparse multipath channels whose multipath components are evenly delayed by mT, where m is a positive integer, it can be proved that the only nonzero-valued taps of a finite-length, T-spaced U-LE or U-DFE are located at positions that are multiples of m. In other words, the finite-length, T-spaced U-Es are inherently nonuniformly spaced for such channels and the optimum tap positions are $\{c_0, c_{\pm m}, c_{\pm 2m}, \ldots\}$, which are also the positions assigned by the algorithms. Another feature of the algorithms is their applicability to channels with multipath delays that are nonnegative real numbers. To design a T-spaced NU-E for such channels, an additional step is needed to quantize the assigned tap positions, which are now real numbers, to their nearest integral positions. If a T/2-spaced NU-E is desired, which is usually actually more suitable than a T-spaced NU-E for such channels, the assigned tap positions can be quantized to the nearest positions that are multiples of $\frac{1}{2}$. Similarly, the idea can be extended to design any fractionally spaced equalizers, although fractional spacings smaller than T/2 are seldom used in practice. As an example, Table 2 lists the tap positions of the T- and T/2-spaced NU-DFEs for a sparse multipath channel with delays that are multiples of $\frac{1}{2}$.

While the emphasis has been on the traditional types of NU-Es so far, there exists a different form of NU-DFE, called the *decision-directed feedback equalizer* (NU-DDFE), which has become increasingly popular for equalizing sparse multipath channels [e.g., 13–15] and thus deserves some attention as well. As shown in Fig. 4, the DDFE cancels the postcursor ISI before feedforward filtering via a FBF with impulse response identical to the strictly causal portion of the impulse response of the sampled receive filter output. This feature enables the FBF to fully exploit the sparseness inherent to the output and enjoy a substantial tap reduction, in contrast to the conventional DFE, in which the amount of sparseness in the output is usually diminished after feedforward filtering and thus makes a sizable FBF tap reduction without incurring a significant performance loss difficult. To determine the tap positions of the FFF of a NU-DDFE, Algorithms 1 and 2 can be applied directly, since it has been proved that the FFFs of a DFE and a DDFE are equivalent provided that all postcursor ISI is eliminated [13]. For the FBF of a NU-DDFE, the large magnitude taps in the strictly causal portion of the sampled receive filter output that satisfy a constraint [16] are selected. Note that if a T/2-spaced FFF is employed

Table 2. Tap Positions of NU-DFEs for Equalizing $h_{c_2} = [a_0, a_{1.5}, a_9]$, Where $a_0 = 0.2598 + j0.15$, $a_{1.5} = 1$, and $a_9 = 0.7071 + j0.7071$[a]

(8, 6)-tap NU-DFE (MF)	$\mathcal{S}_{\text{FFF}} = \{0, \mathbf{-1.5}, \mathbf{-7.5}, \mathbf{-9}, -10.5, -15, -16.5, -22.5\}$	T-spaced	$\{0, -1, -2, -7, -8, -9, -15, -16\}$
		T/2-spaced	$\mathcal{S}_{\text{FFF}}$
	$\mathcal{S}_{\text{FBF}} = \{1.5, 6, 7.5, 9\}$	T-spaced T/2-spaced	$\{1, 2, 6, 7, 8, 9\}$
(8, 6)-tap NU-DFE (SRRC)	$\mathcal{T}^1_{\text{FFF}} = \mathcal{S}_{\text{FFF}} - 1.5$	T-spaced	$\{0, -1, -2, -3, -9, -10, -11, -16\}$
	$\mathcal{T}^2_{\text{FFF}} = \{0\}$	T/2-spaced	$\{0, -1.5, -3, -9, -10.5, -16.5, -18, -24\}$
	$\mathcal{T}_{\text{FBF}} = \{1.5, 6, 7.5, 9\}$	T-spaced T/2-spaced	$\{1, 2, 6, 7, 8, 9\}$

[a] A (k_1, k_2)-tap NU-DFE represents one with k_1 FFF taps and k_2 FBF taps.
Source: Ref. 10.

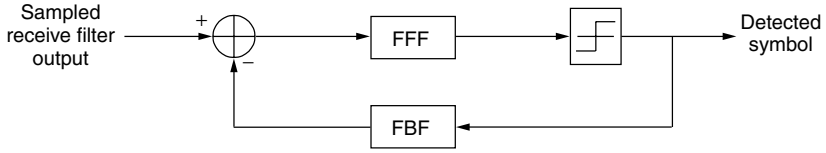


Figure 4. Structure of a DDFE.

in the NU-DDFE, the FBF of the NU-DDFE will be T/2-spaced as well.

Finally, it should be evident that some prior knowledge of the channel, from merely the positions of the taps with large magnitude to a complete estimate of the tap values, is required for most of the existing suboptimum tap allocation schemes discussed in Section 3. For Algorithms 1 and 2, knowing the positions of the taps with large magnitude and which one is the position of the largest magnitude tap are sufficient. Generally speaking, a key property of an appropriate channel estimation algorithm or detection method for the positions of the large magnitude taps is low complexity, so that the overall complexity when added together with a tap allocation scheme and an adaptive NU-E is still lower than the conventional approach of an adaptive U-E without the need of explicit channel estimation. Various channel estimation/detection methods tailored for sparse multipath channels are available in the literature; the more recent ones are specifically customized for use with subsequent equalization. While a discussion of these techniques is outside the scope of this article, a selected few are given in the Further Reading list for the interested reader. It should also be mentioned that, besides implementing an equalizer as an adaptive filter, the equalizer tap values can be computed directly from the channel estimates via fast algorithms with complexity of the order $O(n^2)$, where n is the number of equalizer taps. Unfortunately, such an approach may not be suitable for NU-Es, as these fast algorithms all utilize the Toeplitz nature of the input correlation matrices of U-Es and do not apply to those of NU-Es without the Toeplitz characteristic, which means that only standard algorithms with $O(n^3)$ complexity can be exploited to directly compute the tap values of a NU-E. Therefore, unless the reduction of taps is large, using a NU-E instead of a U-E in this manner may actually increase the computational load.

6. PERFORMANCE EXAMPLE

In this section, a representative example is selected to illustrate some fundamental properties of finite-length NU-Es. The system setup is the same as that shown in Fig. 3. The 4-QAM signaling scheme with constellations $\{\pm 1 \pm j\}$ is chosen and the excess bandwidth of the SRRC transmit and receive filters is set to 35%. Perfect channel knowledge is presumed at the receiver, and optimum tap values of the NU-Es are computed ideally using the matrix inversion approach once their tap positions are determined from the algorithms of Section 5. The bit error rate (BER) after equalization is evaluated analytically through the Beaulieu series as in Ref. 17, and perfect decision feedback is assumed in the DFEs.

Figure 5 shows the performance of the T- and T/2-spaced NU-DFEs for equalizing h_{TV}, one of seven test channels for HDTV systems [2] but with its precursor tap modified from −20 dB to −6 dB relative to the main response to increase the difficulty of equalization, as has been done in Ref. 13. An important result revealed in this plot is that the T-spaced NU-DFE under the MF system only performs better than the one under the SRRC receive filter system at low SNRs, with the crossover point of their BER curves around the SNR of 17.5 dB. This counterintuitive phenomenon is due to the incapability of an equalizer with a limited number of taps, such as a NU-E, to mitigate the extra ISI terms introduced by matched filtering. Although the MF maximizes the SNR before equalization, this gain is only minimal when the influence of noise is insignificant. As a result, if a NU-E is employed at a high SNR region, the benefit of using the MF to maximize the SNR will be insufficient for compensating the detrimental effect of the extra ISI terms introduced, and so the BER suffers. In fact, this phenomenon is even more apparent when the T/2-spaced NU-DFEs of the two systems are compared, with the one under the SRRC receive filter system having a lower BER curve than its counterpart under the MF system for the entire range of SNR shown. This is because a T/2-spaced equalizer can function as a MF, thus allowing the one under the SRRC receive filter system to enjoy the same benefit as the one under the MF system, but without the need to suppress the extra ISI terms that a preceding MF introduces. Therefore, the SRRC receive

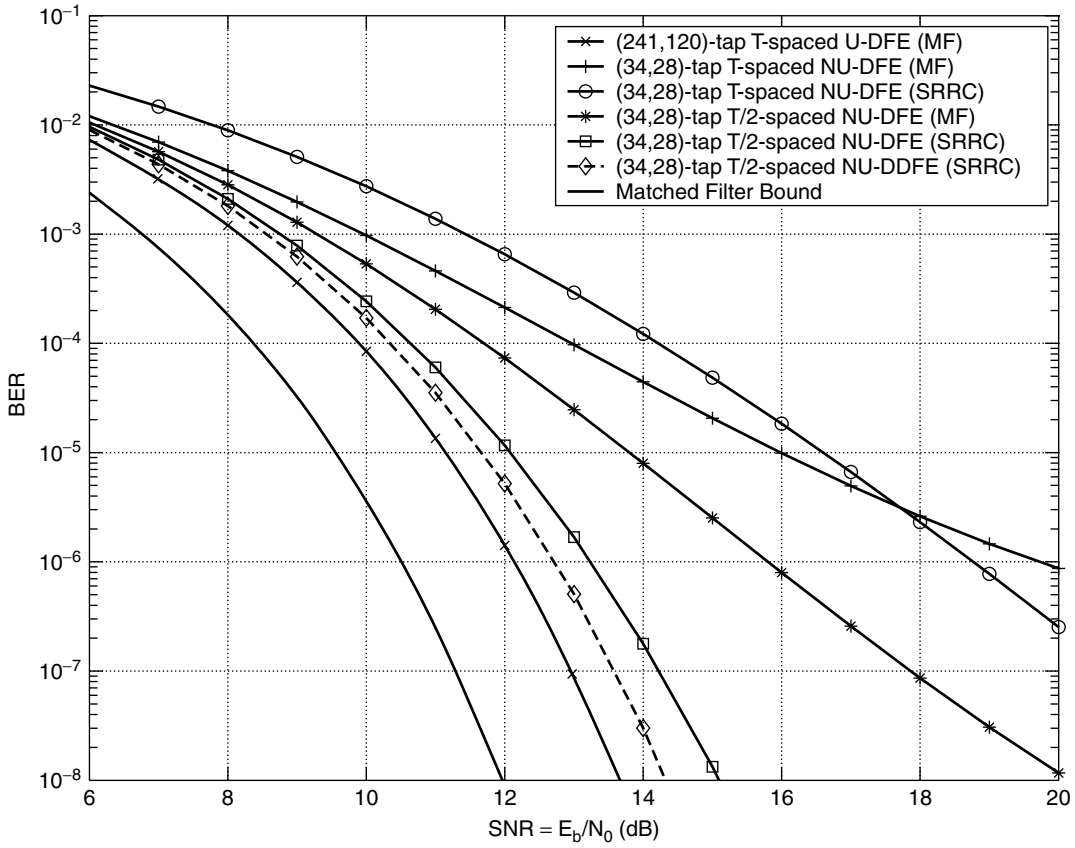


Figure 5. Performance of NU-DFEs for equalizing $h_{\mathrm{TV}} = [a_0, a_{9.68}, a_{10.49}, a_{19.37}, a_{40.35}, a_{106.52}]$, where $a_0 = 0.1549 - j0.4767$, $a_{9.68} = -1$, $a_{10.49} = 0.1$, $a_{19.37} = 0.0389 + j0.1197$, $a_{40.35} = -0.1614 + j0.1173$, and $a_{106.52} = -0.2558 - j0.1859$. (*Source*: Ref. 10.)

filter is a better front-end receive filter than the MF when followed by a NU-E. Note that the performance of the T/2-spaced NU-DDFE under the SRRC receive filter system is also included in Fig. 5. In accordance with the discussion of Section 5, it attains a lower BER curve than its traditional counterpart with only a SNR loss of about 0.7 dB relative to the reference U-DFE, which is used to approximate the performance of the infinite-length equalizer. However, for certain sparse multipath channels, the gain in using a NU-DDFE may not be as significant. More information regarding the strengths and weaknesses of the NU-DDFEs can be found in the paper by Lee and McLane [16].

7. CONCLUSIONS

The ever-increasing data rate in modern communication systems has prompted the quest for new equalization techniques to handle channels with long impulse responses efficiently. NU-Es are suggested in this article as a reduced-complexity solution for equalizing a special family of wireless channels called *sparse multipath channels* that exhibits this undesirable characteristic. Tap positions for the infinite-length, T-spaced ZF/MMSE-LEs/DFEs for such channels can be derived by expressing each frequency response as an infinite series, which in turn lead to simple tap allocation algorithms for designing finite-length NU-Es, including a modified form of the NU-DFE. A fundamental property associated with NU-Es is the nonoptimality of the MF as the front-end receive filter; a better substitute is the SRRC receive filter. Overall, good performance is attained by the NU-Es despite their large reduction in complexity. Together with an appropriate low-complexity channel estimation algorithm, it is conceivable that NU-Es can become a crucial component in future-generation mobile receivers that are expected to operate in a wide variety of channel conditions with ease.

BIOGRAPHIES

Frederick K. H. Lee received the B.Sc. degree in electrical and computer engineering in 1998 from Queen's University, Kingston, Ontario, Canada, where he is currently a Ph.D. candidate. During his postgraduate studies, he has been supported by two postgraduate Scholarships from the Natural Sciences and Engineering Research Council (NSERC) of Canada, a Fessenden Postgraduate Scholarship from the Communications Research Centre (CRC), Canada, and an Ontario Graduate Scholarship. His research interests are communications theory and signal processing algorithms for communications.

Peter McLane has been a Professor at Queen's University since 1969. He is a Fellow of the IEEE and served as the Chairman of the IEEE Communications Society Communication Theory Committee for 3 years. He has served as a Major Project Leader for both Communications and Information Technology Ontario (CITO) and for the Canadian Institute of Telecommunications Research (CITR). He has been a member of both Research and Scholarship Committees with NSERC. He jointly received two research awards: the 1994 Stentor Telecommunications Research Award and the TRIO Feedback Award in 1992. He is one of four authors of the books *Introduction to trellis Coded Modulation with Applications*, originally published by Macmillian in 1991. He has spent academic leaves at UBC, AT&T Bell Labs, Motorola, and Harris Canada.

BIBLIOGRAPHY

1. H. V. Poor and G. W. Wornell, eds., *Wireless Communications: Signal Processing Perspectives*, Prentice-Hall, Upper Saddle River, NJ, 1998.
2. W. F. Schreiber, Advanced television systems for terrestrial broadcasting: Some problems and proposed solutions, *Proc. IEEE* **83**: 958–981 (1995).
3. S. A. Raghavan, J. K. Wolf, L. B. Milstein, and L. C. Barbosa, Nonuniformly spaced tapped-delay-line equalizers, *IEEE Trans. Commun.* **COM-41**: 1290–1295 (1993).
4. I. Lee, Optimization of tap spacings for the tapped delay line decision feedback equalizer, *IEEE Commun. Lett.* **COMML-5**: 429–431 (2001).
5. M. Kocic, D. Brady, and M. Stojanovic, Sparse equalization for real-time digital underwater acoustic communications, *Proc. IEEE OCEANS'95*, 1995, pp. 1417–1422.
6. S. Ariyavisitakul, N. R. Sollenberger, and L. J. Greenstein, Tap-selectable decision-feedback equalization, *IEEE Trans. Commun.* **COM-45**: 1497–1500 (1997).
7. M. J. Lopez and A. C. Singer, A DFE coefficient placement algorithm for sparse reverberant channels, *IEEE Trans. Commun.* **COM-49**: 1334–1338 (2001).
8. B. Geller et al., Equalizer for video rate transmission in multipath underwater communications, *IEEE J. Ocean. Eng.* **OE-21**: 150–155 (1996).
9. K. Berberidis and A. A. Rontogiannis, Efficient decision feedback equalizer for sparse multipath channels, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, ICASSP'00*, 2000, pp. 2725–2728.
10. F. K. H. Lee and P. J. McLane, Design of nonuniformly-spaced tapped-delay-line equalizers for sparse multipath channels, *IEEE Trans. Commun.* (in press); see also *Proc. IEEE Global Telecommunications Conf., GLOBECOM'01*, 2001, pp. 1336–1343.
11. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, 2nd ed., Kluwer, Boston, MA, 1994.
12. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
13. I. J. Fevrier, S. B. Gelfand, and M. P. Fitz, Reduced complexity decision feedback equalization for multipath channels with large delay spreads, *IEEE Trans. Commun.* **COM-47**: 927–937 (1999).
14. P. De, J. Bao, and T. Poon, A calculation-efficient algorithm for decision feedback equalizers, *IEEE Trans. Consumer Electron.* **CE-45**: 526–532 (1999).
15. M. Stojanovic, L. Freitag, and M. Johnson, Channel-estimation-based adaptive equalization of underwater acoustic signals, *Proc. IEEE OCEANS'99*, 1999, pp. 985–990.
16. F. K. H. Lee and P. J. McLane, Comparison of two nonuniformly-spaced decision feedback equalizers for sparse multipath channels, *Proc. IEEE Int. Conf. Commun. ICC'02*, 2002, pp. 1923–1928.
17. J. E. Smee and N. C. Beaulieu, Error-rate evaluation of linear equalization and decision feedback equalization with error propagation, *IEEE Trans. Commun.* **COM-46**: 656–665 (1998).

FURTHER READING

Alternative Equalization Techniques for Sparse Multipath Channels

N. Benvenuto and R. Marchesani, The Viterbi algorithm for sparse channels, *IEEE Trans. Commun.* **COM-44**: 287–289 (1996).

N. C. McGinty, R. A. Kennedy, and P. Hoeher, Parallel trellis Viterbi algorithm for sparse channels, *IEEE Commun. Lett.* **COMML-2**: 143–145 (1998).

R. Cusani and J. Mattila, Equalization of digital radio channels with large multipath delay for cellular land mobile applications, *IEEE Trans. Commun.* **COM-47**: 348–351 (1999).

S. Chowdhury, M. D. Zoltowski, and J. S. Goldstein, Structured MMSE equalization for synchronous CDMA with sparse multipath channels, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, ICASSP'01*, 2001, pp. 2113–2116.

K. Chugg, A. Anastasopoulos, and X. Chen, *Iterative Detection: Adaptivity, Complexity Reduction and Applications*, Kluwer Academic Publishers, 2001, Chapter 3.

Selected Channel Estimation/Detection Methods for Sparse Multipath Channels

Y. F. Cheng and D. M. Etter, Analysis of an adaptive technique for modeling sparse systems, *IEEE Acoust. Speech Signal Process.* **ASSP-37**: 254–264 (1989).

M. Kocic and D. Brady, Complexity-constrained RLS estimation for sparse systems, *Proc. Conf. Information Science Systems, CISS'94*, 1994, pp. 420–425.

J. Homer, I. Mareels, R. R. Bitmead, B. Wahlberg, and F. Gustafsson, LMS estimation via structural detection, *IEEE Trans. Signal Process.* **SP-46**: 2651–2663 (1998).

I. Kang, M. P. Fitz, and S. B. Gelfand, Blind estimation of multipath channel parameters: A modal analysis approach, *IEEE Trans. Commun.* **COM-47**: 1140–1150 (1999).

I. Ghauri and D. T. M. Slock, Structured estimation of sparse channels in quasi-synchronous DS-CDMA, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, ICASSP'00*, 2000, pp. 2873–2876.

Y. Jin and B. Friedlander, On the performance of equalizers for sparse channels, *Proc. Asilomar Conf. Signals, Syst., Comput.* **34**: 1757–1761 (2000).

W. Sung, D. J. Shin, and I. K. Kim, Maximum-likelihood tap selection for equalization of land mobile radio channels, *Proc. IEEE Global Telecommunications Conf., GLOBECOM'01*, 2000, pp. 3326–3330.

S. F. Cotter and B. D. Rao, Sparse channel estimation via matching pursuit with application to equalization, *IEEE Trans. Commun.* **COM-50**: 374–377 (2002).

O

OPTICAL COUPLERS

GERD KEISER
PhotonicsComm Solutions, Inc.
Newton Center, Massachusetts

1. INTRODUCTION

Optical couplers play a key role in wavelength-division multiplexing (WDM) applications for combining and separating wavelength channels, tapping off power for monitoring purposes, or adding and dropping specific wavelengths at a particular point in an optical fiber communication link [1–3]. Most optical couplers are passive devices in the sense that they do not need to be powered externally to perform their function on optical signals. Fundamentally, optical couplers connect three or more fibers to combine, split, or redirect light signals.

Since optical couplers perform many different functions, they can be made in several configurations, as shown in Fig. 1. The T coupler, Y coupler, or 1×2 coupler is a three-port device that is mainly used to tap off a portion of the light from a throughput fiber into a second fiber. The relative optical power level in each output branch is usually given in percentages. The design can be tailored to achieve any coupling ratio between the two outputs. This coupler nominally is used for signal-monitoring applications. In this case, a tradeoff between coupling loss in the primary fiber and an adequate level of power required for the measurement threshold in the secondary branch shows that a 10% tap is the optimal configuration [4]. This means that 90% of the input optical power continues through the device and 10% is tapped off for signal monitoring purposes.

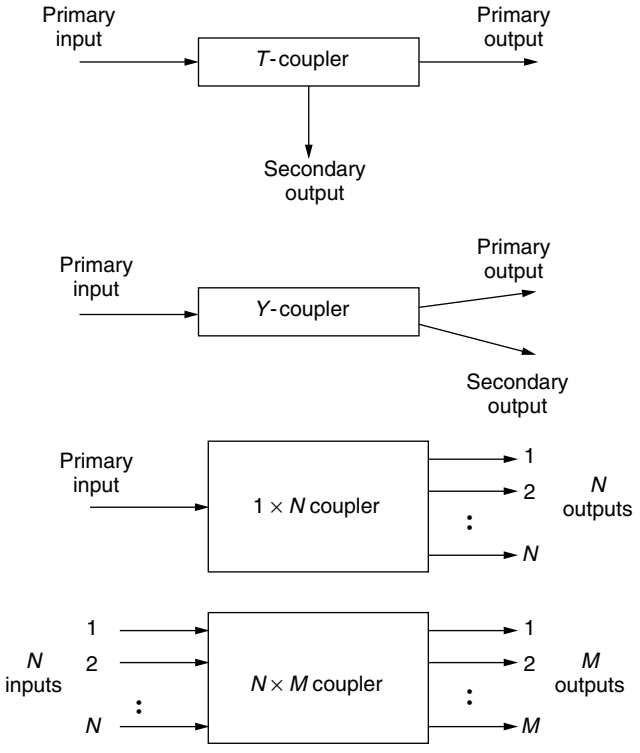


Figure 1. Example configurations for optical couplers.

The $1 \times N$ or tree coupler has one input fiber and N output fibers. In the most general case, this device is not wavelength dependent and it divides all the input optical power equally among the N output ports. Many of these devices are directional, which means that their function depends on the direction in which the light passes through it.

A more general configuration is the $N \times M$ or star coupler, which has N input ports and M output ports. In the broadest application, star couplers combine the light streams from two or more input fibers and divide them among several output fibers. In the general case, the splitting is done uniformly for all wavelengths, so that each of the M outputs receives $1/M$ of the power entering the device. A common fabrication method for an $N \times N$ coupler is to fuse together the cores of N single-mode fibers over a length of a few millimeters. The optical power inserted through one of the N fiber entrance ports gets divided uniformly into the cores of the N output fibers through evanescent power coupling in the fused region.

Wavelength-selective couplers form a more versatile category of devices for WDM applications. Among the technologies used for making these devices are 2×2 fused-fiber couplers, coupled planar waveguides, Mach–Zehnder interferometers, fiber Bragg gratings, and phased-array waveguide gratings.

2. COUPLER CONSTRUCTIONS

The 2×2 coupler is a simple fundamental device that we will use here to demonstrate the operational principles. These devices can be fabricated by microoptic, optical fiber, or integrated optic methods. A common construction is the fused-fiber coupler [3,5–7]. This is fabricated by twisting together, melting, and pulling two single-mode fibers so they get fused together over a uniform section of length W, as shown in Fig. 2. Each input and output fiber has a long tapered section of length L, since the transverse dimensions are gradually reduced down to that of the coupling region when the fibers are pulled during the fusion process. The total draw length is $2L + W$. This device is known as a fused biconical tapered coupler. Here P_0 is the input power, P_1 is the throughput power, and P_2 is the power coupled into the second fiber. The parameters P_3 and P_4 are extremely low signal levels (-50 to -70 dB below the input level) resulting from backward reflections and scattering due to bending in and packaging of the device.

As the input light P_0 propagates along the taper in fiber 1 and into the coupling region, an increasingly larger portion of the input field now propagates outside the core

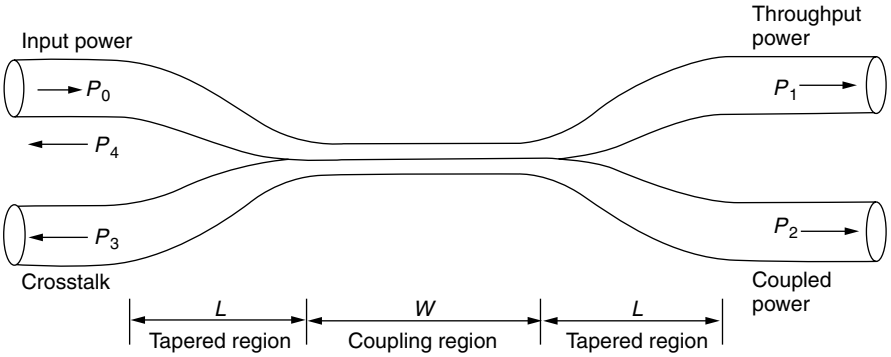


Figure 2. Cross-sectional view of a fused-fiber coupler having a coupling region W and two tapered regions of length L. The total span $2L + W$ is the coupler draw length.

of the input fiber and is coupled into the adjacent fiber. Depending on the dimensioning of the coupling region, any desired fraction of this decoupled field can be coupled into fiber 2. By making the tapers very gradual, only a negligible fraction of the incoming optical power is reflected back into either of the input ports. Thus these devices are also known as directional couplers.

Fused-fiber couplers have an intrinsic wavelength dependence in the coupling region. The optical power coupled from one fiber to another at a specific wavelength can be varied through three parameters: the axial length of the coupling region over which the fields from the two fibers interact; the size of the reduced radius r in the coupling region; and Δr, the difference in radii of the two fibers in the coupling region. In making a fused-fiber coupler, the coupling length W is normally fixed by the width of the heating flame that is used in the melting process, so that only L and r change as the coupler is elongated. Typical values for W and L are a few millimeters, the exact values depending on the coupling ratios desired for a specific wavelength, and $\Delta r/r$ is around 0.015. Assuming that the coupler is lossless, the expression for the power P_2 coupled from one fiber to another over an axial distance z is

$$P_2 = P_0 \sin^2(\kappa z) \tag{1}$$

where κ is the *coupling coefficient* describing the interaction between the fields in the two fibers. By conservation of power, for identical-core fibers we have

$$P_1 = P_0 - P_2 = P_0[1 - \sin^2(\kappa z)] = P_0 \cos^2(\kappa z) \tag{2}$$

Wavelength-dependent multiplexers can also be made using Mach–Zehnder interferometry techniques [8]. Figure 3 illustrates the constituents of an individual

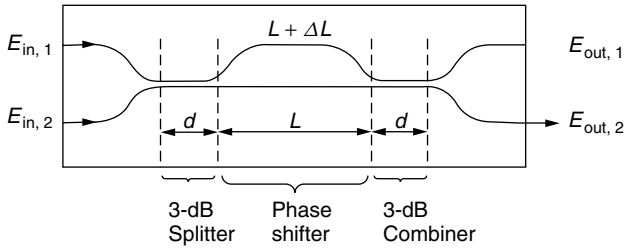


Figure 3. Constituents of an individual Mach–Zehnder interferometer (MZI). E_{in} and E_{out} are electric field intensities.

Mach–Zehnder interferometer (MZI). This 2×2 MZI consists of three stages: an initial 3-dB directional coupler which splits the input signals, a central section where one of the waveguides is longer by ΔL to give a wavelength-dependent phase shift between the two arms, and another 3-dB coupler which recombines the signals at the output. The function of this arrangement is that, by splitting the input beam and introducing a phase shift in one of the paths, the recombined signals will interfere constructively at one output and destructively at the other. The signals then finally emerge from only one output port.

A grating is an important element in WDM systems for combining and separating individual wavelengths. Basically a grating is a periodic structure or perturbation in a material. This variation in the material has the property of reflecting or transmitting light in a certain direction depending on the wavelength. Thus gratings can be categorized as either transmitting or reflecting gratings.

Figure 4 shows a simple concept of a demultiplexing function using a fiber Bragg grating [9,10]. To extract the desired wavelength, a circulator is used in conjunction with the grating. In a three-port circulator, an input signal on one port exits at the next port. For example,

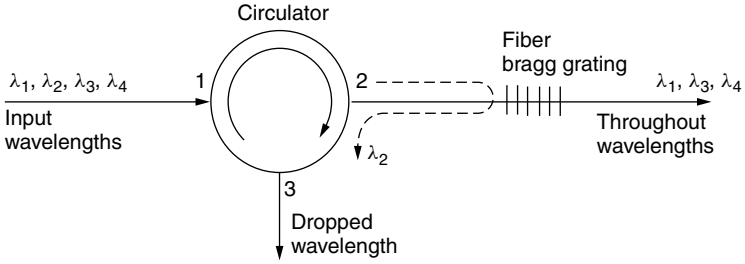


Figure 4. A simple wavelength demultiplexing function using a fiber Bragg grating.

an input signal at port 1 is sent out at port 2. Here the circulator takes the four wavelengths entering port 1 and sends them out port 2. All wavelengths except λ_2 pass through the grating. Since λ_2 satisfies the Bragg condition of the grating, it is reflected, enters port 2 of the circulator, and exits at port 3. More complex multiplexing and demultiplexing structures with several gratings and several circulators can be realized with this scheme.

A highly versatile WDM device is based on using an arrayed waveguide grating. This device can function as a multiplexer, a demultiplexer, a drop-and-insert element, or a wavelength router. A variety of design concepts have been examined [11,12]. The arrayed waveguide grating is a generalization of the 2×2 Mach–Zehnder interferometer multiplexer. One popular design consists of M_{in} input and M_{out} output slab waveguides and two identical focusing planar star couplers connected by N uncoupled waveguides with a propagation constant β. The lengths of adjacent waveguides in the central region differ by a constant value ΔL, so that they form a Mach–Zehnder-type grating, as Fig. 5 shows. For a pure multiplexer, we can take $M_{\text{in}} = N$ and $M_{\text{out}} = 1$. The reverse holds for a demultiplexer, that is $M_{\text{in}} = 1$ and $M_{\text{out}} = N$. In the case of a network routing application, we can have $M_{\text{in}} = M_{\text{out}} = N$.

3. PERFORMANCE CHARACTERISTICS

In specifying the performance of an optical coupler, one usually indicates the percentage division of optical power between the output ports by means of the splitting ratio or coupling ratio. Referring to Fig. 2, where P_0 is the input power and P_1 and P_2 the output powers, then

$$\text{Splitting ratio} = \left(\frac{P_2}{P_1 + P_2}\right) \times 100\% \tag{3}$$

By adjusting the parameters so that power is divided evenly, with half of the input power going to each output, one creates a 3-dB coupler. A coupler could also be made in which, for example, almost all the optical power at 1500 nm goes to one port and almost all the energy around 1300 nm goes to the other port.

In the analysis above, we have assumed for simplicity that the device is lossless. However, in any practical coupler there is always some light that is lost when a signal goes through it. The two basic losses are excess loss and insertion loss. The excess loss is defined as the ratio

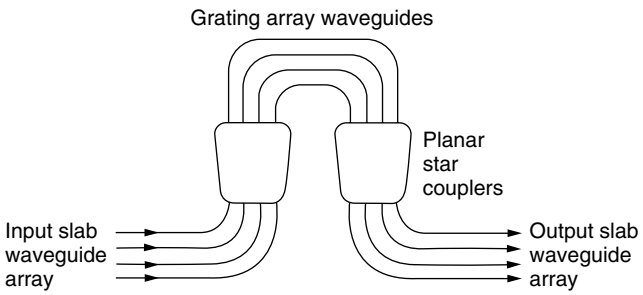


Figure 5. Adjacent waveguides in the central region differ in length to form a Mach–Zehnder-type grating.

of the input power to the total output power. Thus, in decibels, the excess loss for a 2×2 coupler is

$$\text{Excess loss} = 10 \log \left(\frac{P_0}{P_1 + P_2}\right) \tag{4}$$

The insertion loss refers to the loss for a particular port-to-port path. For example, for the path from input port i to output port j, we have, in decibels:

$$\text{Insertion loss} = 10 \log \left(\frac{P_i}{P_j}\right) \tag{5}$$

Another performance parameter is crosstalk, which measures the degree of isolation between the input at one port and the optical power scattered or reflected back into the other input port. That is, it is a measure of the optical power level P_3 shown in Fig. 2:

$$\text{Crosstalk} = 10 \log \left(\frac{P_3}{P_0}\right) \tag{6}$$

The principal role of any star coupler is to combine the powers from N inputs and divide them equally among M output ports. Techniques for creating star couplers include fused fibers, gratings, microoptic technologies, and integrated optics schemes. The fused-fiber technique has been a popular construction method for $N \times N$ star couplers. For example, 7×7 devices and 1×19 splitters or combiners with excess losses at 1300 nm of 0.4 and 0.85 dB, respectively, have been demonstrated. However, large-scale fabrication of these devices for $N > 2$ is limited because of the difficulty in controlling the coupling response between the numerous fibers during the heating–pulling process.

In an ideal star coupler the optical power from any input is evenly divided among the output ports. The total loss of the device consists of its splitting loss plus the excess loss in each path through the star. The splitting loss is given in decibels by

$$\text{Splitting loss} = -10 \log \left(\frac{1}{N}\right) = 10 \log N \tag{7}$$

For a single input power P_{in} and N output powers, the excess loss in decibels is given by

$$\text{Fiber star excess loss} = 10 \log \left(\frac{P_{\text{in}}}{\sum_{i=1}^{N} P_{\text{out},i}}\right) \tag{8}$$

The insertion loss and crosstalk can be found from Eqs. (5) and (6), respectively.

An alternative is to construct star couplers by cascading 3-dB couplers. Figure 6 shows an example for an 8×8 device formed by using twelve 2×2 couplers. This device could be made from either fused-fiber or integrated-optic components. As can be seen from this figure, a fraction $1/N$ of the launched power from each input port appears at all output ports. A limitation to the flexibility or modularity

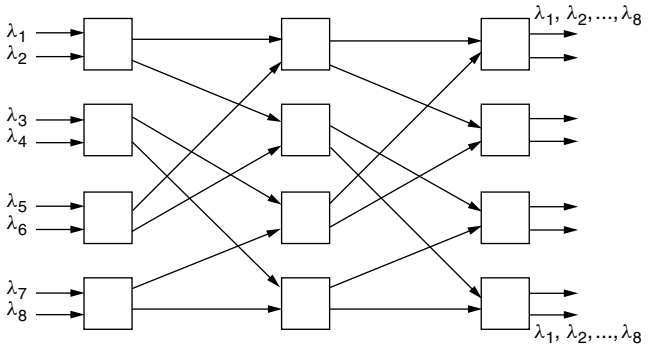


Figure 6. Example for an 8×8 device formed by using twelve 2×2 couplers.

of this technique is that N is a multiple of 2, that is, $N = 2^n$ with the integer $n \geq 1$. The consequence is that if an extra node needs to be added to a fully connected $N \times N$ network, the $N \times N$ star needs to be replaced by a $2N \times 2N$ star, thereby leaving $2(N-1)$ new ports being unused. Alternatively, one extra 2×2 coupler can be used at a port with the result that the outputs of the two new ports have an additional 3-dB loss.

As can be deduced from Fig. 6, the number of 3-dB couplers needed to construct an $N \times N$ star is

$$N_c = \frac{N}{2}\log_2 N = \frac{N}{2}\frac{\log N}{\log 2} \tag{9}$$

since there are $N/2$ elements in the vertical direction and $\log_2 N = \log N / \log 2$ elements horizontally.

If the fraction of power traversing each 3-dB coupler element is F_T, with $0 \leq F_T \leq 1$ (i.e., a fraction $1 - F_T$ of power is lost in each 2×2 element), then the excess loss in decibels is

$$\text{Excess loss} = -10\log(F_T^{\log_2 N}) \tag{10}$$

The splitting loss for this star is again given by Eq. (7). Thus, the total loss experienced by a signal as it passes through the $\log_2 N$ stages of the $N \times N$ star and gets divided into N outputs is, in decibels,

$$\begin{aligned}\text{Total loss} &= \text{splitting loss} + \text{excess loss} \\ &= -10\log\left(\frac{F_T^{\log_2 N}}{N}\right) = -10\left(\frac{\log N \log F_T}{\log 2} - \log N\right) \\ &= 10(1 - 3.322\log F_T)\log N \end{aligned} \tag{11}$$

This shows that the loss increases logarithmically with N.

BIOGRAPHY

Gerd Keiser is the founder and president of Photonics-Comm Solutions, Inc., Newton Center, Massachusetts, a firm specializing in consulting and education for the optical communications industry. He has 25 years experience at Honeywell, GTE, and General Dynamics in designing and analyzing telecommunication components, links, and networks. He is the author of the books *Optical Fiber Communications* (3rd ed. 2000) and *Local Area Networks* (2nd ed. 2002) published by McGraw-Hill. Dr. Keiser is an IEEE fellow and received GTE's prestigious Leslie Warner Award for work in ATM switch development. He earned his B.A. and M.S. degrees in mathematics and physics from the University of Wisconsin and a Ph.D. in solid state physics from Northeastern University, Boston, Massachusetts.

BIBLIOGRAPHY

1. G. Keiser, *Optical Fiber Communications*, 3rd ed., McGraw-Hill, Burr Ridge, IL, 2000, Chap. 10.
2. J. Hecht, *Understanding Fiber Optics*, 4th ed., Prentice-Hall, Upper Saddle River, NJ, 2002, Chap. 15.
3. V. J. Tekippe, Passive fiber optic components made by the fused biconical taper process, *Fiber Integr. Opt.* **9**(2): 97–123 (1990).
4. M. Hoover, New coupler applications in today's telephony networks, *Lightwave* **17**: 134–140 (March 2000) (see *http://www.light-wave.com*).
5. A. Ankiewicz, A. W. Snyder, and X.-H. Zheng, Coupling between parallel optical fiber cores — critical examination, *J. Lightwave Technol.* **4**: 1317–1323 (Sept. 1986).
6. E. Pennings, G.-D. Khoe, M. K. Smit, and T. Staring, Integrated-optic versus micro optic devices for fiber-optic telecommunication systems: A comparison, *IEEE J. Select. Top. Quant. Electron* **2**: 151–164 (June 1996).
7. R. W. C. Vance and J. D. Love, Back reflection from fused biconic couplers, *J. Lightwave Technol.* **13**: 2282–2289 (Nov. 1995).
8. R. Syms and J. Cozens, *Optical Guided Waves and Devices*, McGraw-Hill, New York, 1992.
9. Y. Fujii, High-isolation polarization-independent optical circulator coupled with single-mode fibers, *J. Lightwave Technol.* **9**: 456–460 (April 1991).
10. R. Ramaswami and K. N. Sivarajan, *Optical Networks*, 2nd ed., Morgan Kaufmann, San Francisco, 2002.
11. M. K. Smit and C. van Dam, PHASAR-based WDM devices: Principles, design and applications, *IEEE J. Select. Top. Quant. Electron* **2**: 236–250 (June 1996).
12. H. Takahashi, K. Oda, H. Toba, and Y. Inoue, Transmission characteristics of arrayed waveguide $N \times N$ wavelength multiplexers, *J. Lightwave Technol.* **13**: 447–455 (March 1995).

OPTICAL CROSSCONNECTS

Li Fan
OMM, Inc.
San Diego, California

1. INTRODUCTION

Massive information demand in the Internet is creating enormous needs for the capacity and communication bandwidth expansion in the service providers and the carriers. With the expectation that the new data traffic will have

exponential growth, the service providers are under pressure to find a new technology, which can dramatically reduce the cost of hardware and network management as well as solving the bandwidth bottleneck. Instead of chasing the growing bandwidth, service providers are desperate to keep ahead of the competition. In particular, they want to boost capacity by orders of magnitude. Optical networks are digital communications systems that use light waves in the fiber as a medium for the transmission or switching of data. This new optical layer is the future to providing cost-effective capacity. Dense wavelength division multiplexing (DWDM) is the ideal solution to dramatically increase bandwidth. With the enormous channels and traffic, optical crossconnects (OXC) is the emerging technology and component that will deliver and manage this new optical layer and the services running on it. Optical crossconnects can be used for protection and restoration in optical networks, bandwidth provisioning, wavelength routing, and network performance monitoring. It is one of the key elements for routing optical signals in an optical network or system for long-haul communication, metro area, and local access add-drop as shown in Fig. 1.

2. OPTICAL CROSSCONNECTS FABRIC

Currently, carrier backbones already carry information traffic with light over fiber. For long-distance fiber, the light has to be converted back into electrical and periodically regenerated depending on the fiber type. Most current OXC in fact use an electronic core for switching such as the synchronous optical network (SONET) system, fiber distributed data interface (FDDI) switches, asynchronous transfer mode (ATM) switches, and ethernet switches with fiber optic interfaces to convert the photons into electrical signals in order to switch them from one fiber to another at junction points. Sometimes the electronic switch is referred to as optical-electrical-optical (OEO) switch as shown in Fig. 2. In this OEO crossconnects, the input/output signals are optical, but receivers convert the input signals to electrical signals, then use electronic components to route the channels through the core. At the transceiver module, the electrical signals are converted back into photons. This solution is not future proof since when the data rate increases, the expensive transceivers and the electrical switch core have to be replaced.

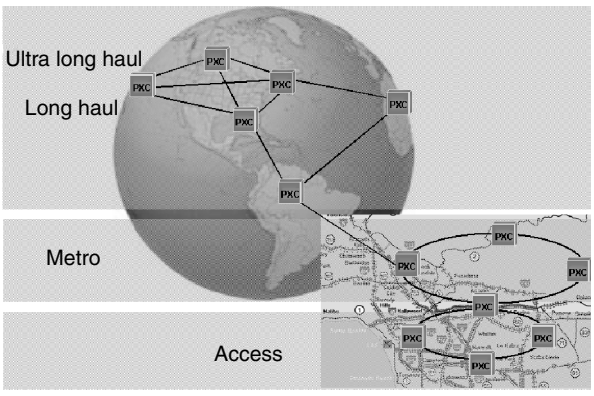


Figure 1. Use optical crossconnects to manage the optical network for wavelength routing, network performance monitoring, bandwidth provisioning, protection, and restoration in optical networks.

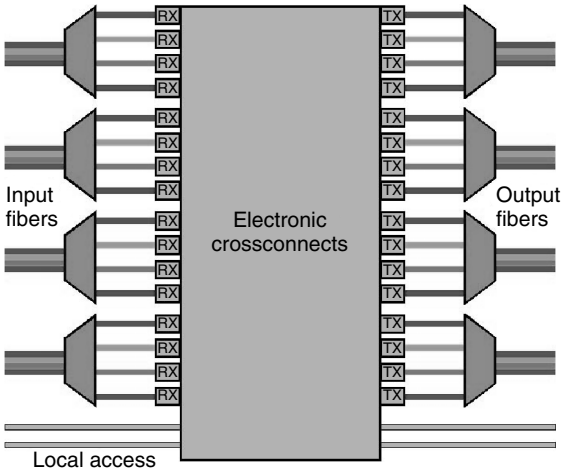


Figure 2. OEO electronic crossconnects. The wavelengths from DWDM fibers are converted into electrical signals by receivers. The switching and routing are done by the electronic core. The output signals are converted back into optical by transmitters.

All-optical crossconnects use light waves exclusively from end to end. The data is maintained in original optical format from the input fiber to the switch element. The data format is unchanged from switching elements to the output fiber. Sometimes the all-optical crossconnects is referred to as OOO crossconnects as shown in Fig. 3, which stands for optical-optical-optical. It is preferred to use the terminology "photonic crossconnects" for all-optical crossconnects rather than "optical crossconnects." This is to designate the fact that the data path of the switch is purely photonic, with no electrical conversions. The all-optical crossconnects are much more attractive because of the avoidance of the conversion stages and because the core switch is independent of data rate and data protocol, making the

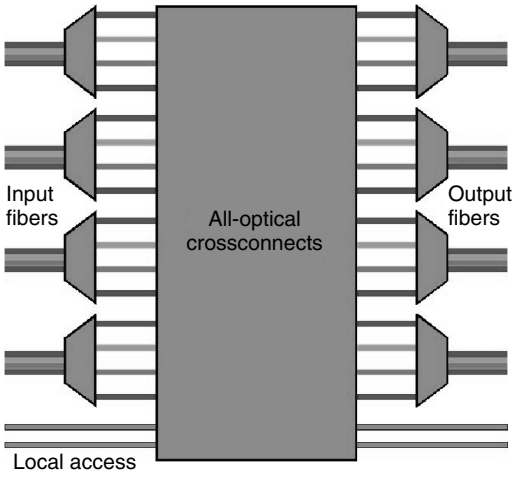


Figure 3. All-optical crossconnects. The data is maintained in original optical format from input fiber through the core switch to the output fiber. There is no need for expensive and high-power consumption high-speed electronics and transmitters and receivers.

cross-connect ready for future data rate upgrades. It provides a valuable capability for handling high bandwidth traffic, as there is no need for expensive and high-power consumption high-speed electronics, transmitters, and receivers. The system becomes less expensive in addition to the reduction of complexity. The all-optical crossconnects improve reliability and reduce the footprint when compared with OEO solutions. Another major benefit of all-optical devices is their greater scalability over OEOs.

Some advantages of the all-optical crossconnects are also disadvantages when we try to coexist this technology with current network. All-optical crossconnects maintain the original signal from input to output fibers without signal regeneration for cost saving. They use erbium doped fiber amplifiers (EDFA) to boost the signal, not regeneration. However, this approach also loses the advantages of signal regeneration. The network design would be challenging to route the same wavelength from the source to the destination and through the entire multirings or meshes network, eliminating the transponders and removing the capability of wavelength conversion. There is no visibility of bit error rate (BER) or monitoring. An all-optical network only has lambda (wavelength) level granularity and cannot perform sub-lambda mixing and grooming.

To combine the scalability of all-optical crossconnects and the wavelength regeneration and grooming, a compromise design is to integrate the all-optical crossconnects ports with an OEO switch as shown in Fig. 4. The majority of the fabric is a fully connected all-optical crossconnects. A small percent of the output ports and input ports are integrated with local add/drop switching through and OEO router. Since any input port can be connected to any output port and go through the OEO router, the data stream of the selected channel can be detected and processed by software at the individual packet level; the wavelength is capable of 3R (reshape/retime/regenerate) and wavelength translation if the network design is sufficiently advanced.

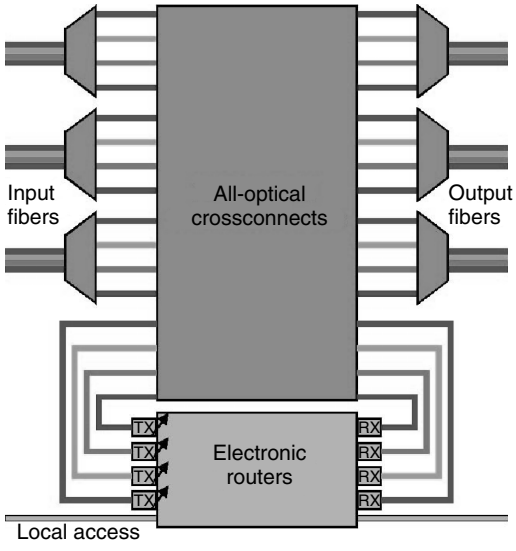


Figure 4. Compromised network design integrated all-optical crossconnects with electronic routers for 3R regeneration and wavelength conversion.

The more complex system may also use additional optical repeaters at each I/O module for 3R regeneration. The repeater is a receiver that can receive any wavelength, directly connected to a transmitter whose output wavelength matches the channel of the wavelength multiplexer.

3. OPTICAL CROSSCONNECTS ARCHITECTURE: 2D AND 3D

The architectures of OXC can be categorized into two approaches. The first configuration is the 2D approach. The devices are arranged in a crossbar configuration. Figure 5 shows the schematic diagram of an 8×8 crossconnects with switches arranged in a 2D array. Each mirror has a digital movement. The switch has only either on or off status, which makes the driving scheme very straightforward. The device can be controlled with a simple TTL signal and does not require feedback control. When a switch is set at the off position, the optical signal will pass through the switch with minimum insertion loss. When the switch is activated, it will bounce the optical signal by 90° and direct the light to the output fiber. The reflection can be achieved by total internal reflection from different refractive index or it can be reflected by a free-space micromirror. Additional functionality can be achieved for adding or dropping optical signals if plane 3 and plane 4 are utilized. Because of the crossbar arrangement, the required mirror number is not linear to port count. The mirror number is equal to N^2. This approach is ideal for small port count crossconnects. However, a large port count crossconnects such as 64×64 will require 4096 switching elements, which is still challenging for current technology.

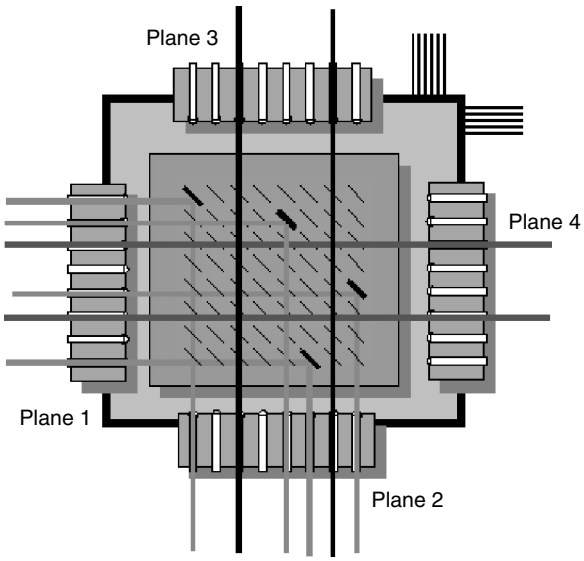


Figure 5. Schematic diagram of 2D digital crossconnects. It needs N^2 switches to configure an $N \times N$ crossconnects. Each switch has only two states: on and off. Optical signal will pass through the switch with minimum lose when the switch is at off position. The light will be reflected at 90° when the switch is activated.

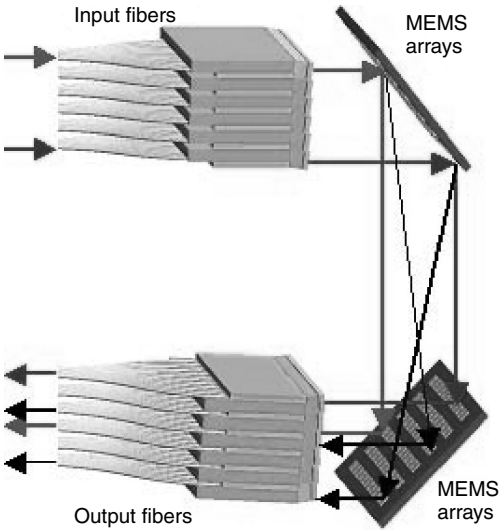


Figure 6. Schematic diagram of 3D crossconnects. The switch number is linearly proportional to channel number. It needs 2N switches to construct $N \times N$ crossconnects. Each mirror has N different states, which are able to point at all the output mirrors with high precision.

The second architecture for the crossconnects is the 3D approach as shown in Fig. 6. This approach has the advantage for switch number (2N), which is linearly proportional to channel number. The connection path is established by tilting two analog mirrors and directing the light from the input fiber to the output fiber. Each mirror has N different states, which need to be well controlled. The 3D approach is the most promising architecture for very large port counts, more than hundred or thousand. The analog fine-tuning of the mirror angle is another unique functionality from 3D that can be used to minimize the insertion loss and equalize uniformity. The drawback of this approach is the control loop and monitoring which can be very complex. The mirror stability, shock/vibration, and long-term drift need to be carefully controlled. The overall total states number for 2D and 3D are the same (Fig. 7). They both need $2N^2$ different states to complete the switching functionality. However, 2D puts more pressure on the switch number; 3D has a smaller switch number but puts more burden on mirror angle control.

Usually OXC are desired to be nonblocking; any input fiber can be switched to any output fiber. With 16 input fibers and 40 wavelengths, the crossconnects size can easily grow to several hundred or even thousand port count at the first installation. It is putting an incredible burden on current technology, especially a few years ago when there were only 2×2 switches commercially available. Even though the large port count crossconnects can be built from 2×2 switching elements, it is not practical or cost effective especially for the performance and scalability. Large nonblocking networks can be constructed with smaller switch fabric by using a multistage Clos network to reduce the number of crosspoints compared to simple matrices [1]. In blocking crossconnects, some connections cannot be established for certain choices or the switch paths are limited to certain zone area. However, the blocking switches can be used as an advantage to reduce the complexity of the crossconnects and enable a larger port count system from smaller modules. For example the wavelength-selective crossconnects (WSXC) is a stack of $N \times N$ switches, each dedicated to signals of the same wavelength as drawn in Fig. 8. For a network with sixteen input fibers, each carrying forty wavelengths, the crossconnects would need to be 640×640 if the system needs to be nonblocking. Because of the lack of wavelength conversion in the fabric, the wavelength is not interchangeable between different wavelengths. Therefore, crossconnects are needed only among the same wavelength. The same functionality network can be built with 40 packages of 16×16 switches, a pay-as-you-grow business model using smaller switches as building elements. New fabrics are added when new wavelengths are turned on. The total bandwidth capacity of a WSXC can be extremely large with low first-installed cost and scalability.

4. TECHNOLOGY

4.1. Planar Lightwave Circuit: Thermo-Optic and Electro-Optic

The planar lightwave circuits (PLC) are constructed with rectangular cross sections of different refraction index materials. The section that transmits the light has a slightly higher refractive index, so that total internal reflection acts to guide the light within the waveguides. The key elements of PLC switches are two directional couplers and the Mach-Zehnder interferometer (MZI). A directional coupler consists of two waveguides very close to each other, so that light waves can be coupled from one to the

	Array size	Switch number	States for each switch	Total states
2D	$N \times N$	N^2	2	$N^2 \times 2 = 2N^2$
3D	$N \times N$	2N	N	$2N \times N = 2N^2$

Figure 7. Two different approaches for OXC show different trade-off. However, the total states for each configuration are identical. Both 2D and 3D require total number of $2N^2$ different states in order to complete $N \times N$ crossconnects functionality. 2D approach has more weight on switch number, which will show bottleneck with larger port count. 3D approach only requires switch number linear proportional to channel number. However, the burden is shifted to the complexity of the switch design.

Figure 8. Wavelength-Selective Crossconnects uses small $N \times N$ switches as building block. The small switch port count (N) is equal to the input fiber number. The total package number (M) is increased when new wavelengths are turned on. The total bandwidth capacity of a WSXC is N times M, which can be extremely large with scalability and low first-install cost.

other. The MZI is a pair of waveguides with identical path lengths. They are separated far enough and will not couple energy between these two waveguides. The incoming light from the input waveguide is spit 50/50 in the first directional coupler. The upper branch goes through a controlled path while the lower branch goes through a reference pathway as illustrated in Fig. 9. The refractive index can be thermally controlled by locally heating the thin film heater above the waveguide [2], or it can be controlled by electro-optic lithium niobate technology [3]. Since the controlled branch and the reference branch of the MZI have identical path length, the light energy will be recombined in the second directional coupler and switch to the upper branch when the controller is off. The controller will change the refractive index and effectively change the path length and phase of the upper branch. The interference will switch the light wave to the lower branch when the controller is on.

Waveguide technology was among the first all-optical switches to be developed, typically in the 1×1, 1×2, and 2×2 range. Because of the planar technology, larger crossconnects can be formed by integrating basic 2×2 components on the same wafer. The optical performance parameters such as crosstalk and insertion loss could be unacceptable for optical network application. However, this technology is capable of integrating variable optical attenuators (VOA) optical switch and wavelength selective elements on the same substrate. It does not require free-space collimator alignment.

4.2. Microfluid

The microfluid and microbubble utilize the interfaces of different refractive index and cause total internal reflection to redirect the light beam. Bubble switches demonstrated the switching mechanism from intersecting waveguides [4]. At each intersection, a trench is etched into the waveguide. The trenches are filled with an index-matching fluid to direct the light in the through states. To direct the light, the thermal inkjet-like matrix-controller silicon chip element heats up the fluid and creates a microbubble in the liquid. The location of the bubble is at the intersection between the input waveguide and output waveguide. The light is reflected by total internal reflection from the liquid/bubble interface as shown in Fig. 10.

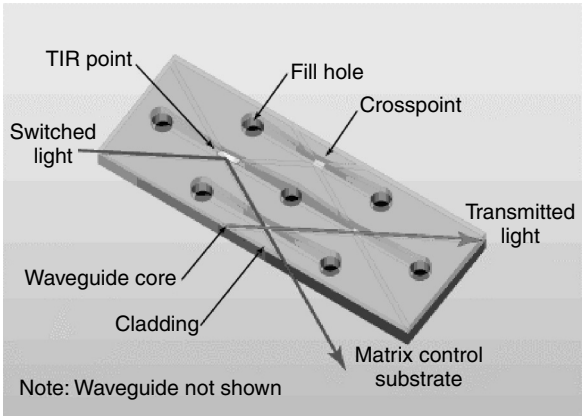


Figure 10. Bubble switch directs light from an input to an output, by using a thermal inkjet element to create a bubble in an index-matching fluid at the intersection between the input waveguide and the desired output waveguide. The light is redirected by means of total internal reflection. Courtesy of Agilent Technologies Inc.

A thermal-capillarity optical switch utilizes a similar total internal reflection concept [5]. The switch element consists of an upper substrate and an intersecting waveguide substrate that has a slit at each crossing point with refractive index matching oil in it and a pair of micro-heaters that produce a thermal gradient along the slit as shown in Fig. 11. The matching oil within the slit is driven by a decrease in interfacial tension of the air-oil interface caused by thermo-capillarity. This switch element also has bi-stable self-latching achieved by capillary pressure that depends on the slit width.

4.3. Liquid Crystal

Liquid crystal crossconnects uses liquid crystal to rotate the optical beam polarization by applying electric voltage to adjust the molecules orientation. Based on this rotation, a beam steering router displaces the signal to one of

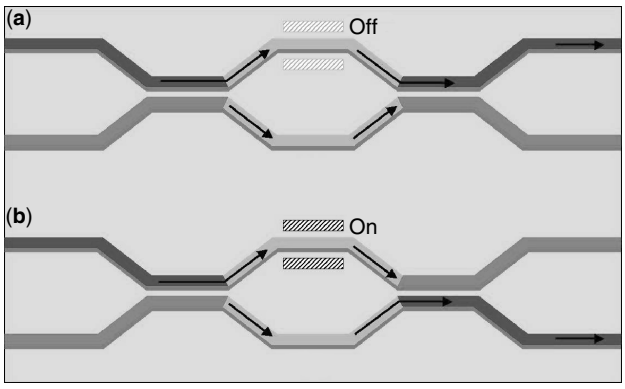


Figure 9. Light from input fiber enters a waveguide at the edge of the optical wafer and goes through a 50/50 split in a directional coupler. One branch goes through a refractive index controlled path (by thermal-optic or electro-optic) while the other goes through a reference pathway.

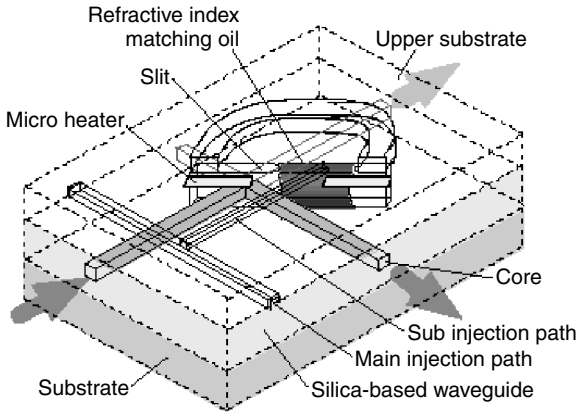


Figure 11. The basic structure of the thermo-capillarity optical switch element. The matching oil within the slit is driven by a decrease in interfacial tension of the air-oil interface caused by thermo-capillarity. Courtesy of NTT Electronics Corp.

two possible paths. Most vendors use this technology for variable optical attenuators rather than switches. This is because liquid crystals can be used to adjust the amount of light that passes through them, rather than simply deflect it. In the switch market, the technology is probably best suited to mid-size wavelength-selective devices. The absence of moving parts and low-power consumption make them a good candidate for test and measurement applications.

4.4. MEMS Micro Mirror

Micro-Electro-Mechanical Systems (MEMS) have been deployed for over a decade in a number of applications such as airbag sensors, projection systems, scanners, and microfluidics. Continued technical developments in the last five years has extended MEMS applications to include optical networking with devices such as all-optical switching fabrics and variable optical attenuators. The MEMS technology has opened up many new possibilities for free-space optical systems. The first commercial MEMS photonic crossconnects were made available in 1999. MEMS technology is using a batch-fabrication process, which is a similar process for making large scale integrated (VLSI) circuits. Wafer scale and chip scale batch process produced MEMS components with high-precision controlled movements. The micro-mechanical structures are smaller, lighter, faster, and more cost effective compared to traditional macro-scale components. The MEMS has become a very good candidate for optical applications, which require stringent reliability, precision, performance, and scalability. The dimension of the micromirror ranges from 0.1 mm to 1 mm for the "sweet spot" of OXC design space. The performance will be strongly limited by Gaussian beam diffraction if the dimension is too small. The large port count crossconnects will not be compatible or scalable with IC processes if the unit switch dimension is too large. The shock/vibration stability and the speed will also miss the SONET specs and Telcordia requirements. Figure 12 shows the OMM 2D digital mirror design and the 16×16 array by surface micromachining technology. The mirror is actuated by a simple TTL signal. The device is not sensitive to driving voltage fluctuations. Since a large force can be generated from this type of gap-closing actuator, the mechanical structure can be built more robust compared to a 3D scanning device. The MEMS array and collimators are hermetically sealed in the package. Maximum insertion loss as low as 1.7 db and 3.1 dB have been obtained for 8×8 and 16×16 2D crossconnects [6].

In the trend of building larger port-count systems and larger mirrors, the optics designs prefer to have a flat mirror surface. Bulk micromachining constructs MEMS devices from single crystal silicon. This technology also enables the possibility of using vertical comb drive to actuate the mirror with large force and more linear response. Figure 13 shows the Lucent LambdaRouter all-optical switch based on Bell Labs's innovative MicroStar 3D MEMS technology [7]. Each mirror angle can be continuously adjusted by voltage.

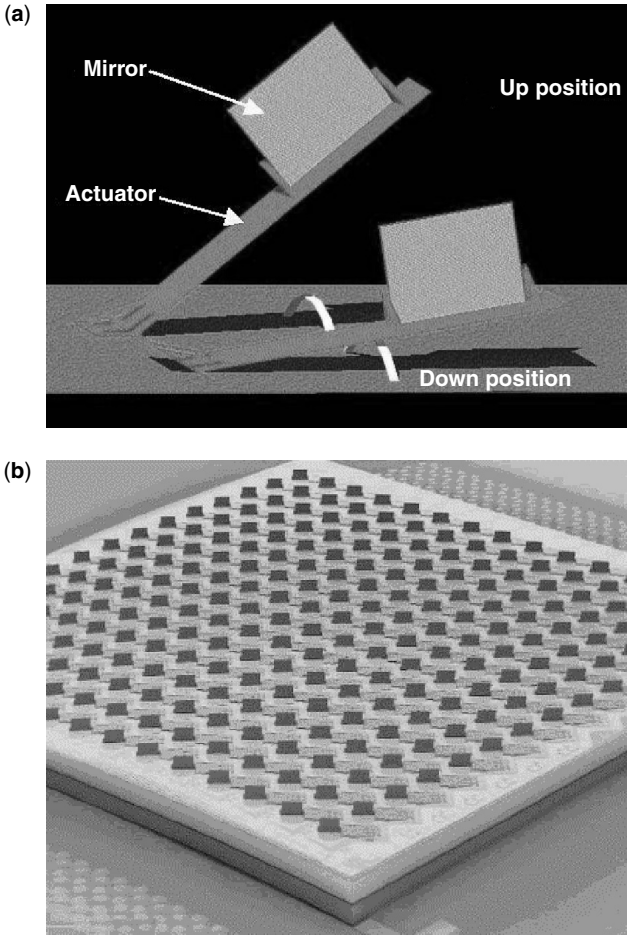


Figure 12. Digital mirrors design from OMM Inc.: (**a**) schematic of basic mirror/switch element. The mirror has bi-stable positions actuated by electrostatic force. (**b**) SEM image of a 16×16 crossconnects with fully populated 256 digital mirrors.

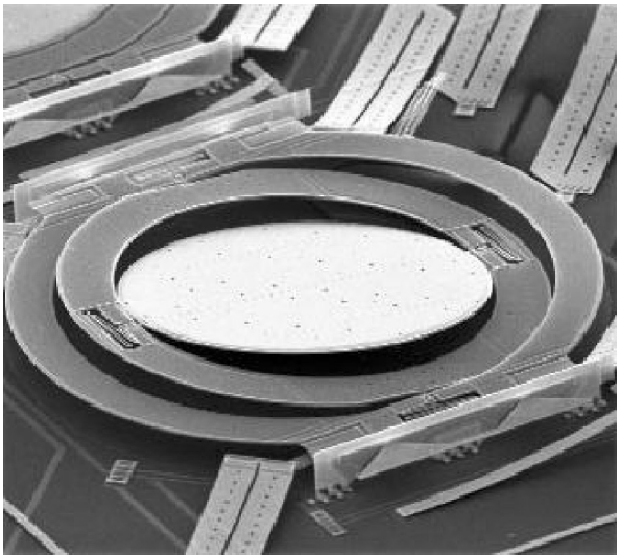


Figure 13. Lucent LambdaRouter based on 3D scanning mirror technology. Each gimbal-structural mirror is able to scan along X and Y axes simultaneously.

5. SUMMARY

Despite the overall slowdown in telecommunications since the year 2000, MEMS based crossconnects have proven to be a reliable and promising technology that can fulfill the stringent requirements of telecommunication industries. Optical crossconnects allow the carriers to add new flexibility and scalability to their optical network. 2D crossconnects are ideal for small port count, less than 64 channels. 3D crossconnects will be the candidate for large-scale networks such as a hundred or thousand channels. It is still not clear what is the optimized module size or approach for the future optical network. It is also not clear how the all-optical network is going to merge with existing opaque systems. The 2D array size has grown from 4×4, 8×8 to 16×16 and 32×32. There is high possibility that the 2D array could achieve larger size with new technology and design improvements. There is also potential that 3D crossconnects will lower the cost, which is competitive with 2D technology. The optical MEMS has gone through an explosive development period. It evolved from the first concept of a fixed stand-up mirror five years ago. Right now it has become a well-accepted technology for fiber communication. There is certainly development space for crossconnects once the demand is generated from the industry.

BIOGRAPHY

Li Fan is cofounder and chief technologist at OMM and is responsible for MEMS design and technology development. He has also successfully commercialized photonic crossconnect ranges from 4×4 to 32×32 based on his MEMS design. He received the Ph.D. in electrical engineering from UCLA in 1998. His research included Optical-Micro-Electro-Mechanical Systems (OMEMS), self-assembly micro-XYZ stage, fiber optical cross-connect, and beam-steering vertical cavity surface-emitting lasers (VCSEL).

BIBLIOGRAPHY

1. C. Clos, A study of nonblocking switching networks, *Bell Syst. Tech. J.* **32**: 406–424 (March 1953).
2. T. Goh et al., Low-loss and high-extinction-ratio silica-based strictly nonblocking 16×16 thermooptic matrix switch, *IEEE Photon. Technol. Lett.* **10**: 810–812 (June 1998).
3. E. J. Murphy, "Photonic switching," I. P. Kaminow and T. L. Koch, eds. *Optical Fiber Telecommunications III B*, Academic Press, New York, 1997, pp. 463–501.
4. J. E. Fouquet, Compact optical cross-connect switch based on total internal reflection in a fluid-containing planar lightwave circuit, *in Proc. Optical Fiber Communication* (*OFC*), TuM1, (2000).
5. M. Makihara, M. Sato, F. Shimokawa, and Y. Nishida, Micromechanical optical switches based on thermocapillary integrated in waveguide substrate, *J. Lightwave Tech.* **17**: 14–18 (1999).
6. P. D. Dobbelaere et al., Digital MEMS for optical switching, *IEEE Comm. Mag.* 88–95 (March 2002).
7. V. Aksyuk et al., Low insertion loss packaged and fiber-connectorized Si surface-micromachined reflective optical switch, *in Proc. Solid-State Sensor and Actuator Workshop*, Hilton Head Island, SC, June 1998, pp. 79–82.

OPTICAL FIBER COMMUNICATIONS

GERD KEISER
PhotonicsComm Solutions, Inc.
Newton Center, Massachusetts

1. INTRODUCTION

1.1. Overview

A major need in human society is the desire to send messages from one distant place to another. Some of the earliest communication systems were based on optical signaling. For example, in the eighth century B.C. the Greeks used a fire signal to send alarms, call for help, or announce certain events. Since then, many forms of communication methodologies have appeared. The basic motivation behind each new form was either to improve the transmission fidelity, to increase the data rate so that more information can be sent, to increase the transmission distance between relay stations, or a combination of these factors. The basic trend in these improvements was to move to higher and higher frequencies of the electromagnetic spectrum. The reason for this is that, in electrical systems, information is usually transferred over the communication channel by superimposing the data onto a sinusoidally varying electromagnetic wave, which is known as the *carrier*. Since the amount of information that can be transmitted is directly related to the frequency range over which the carrier operates, increasing the carrier frequency in turn increases the available transmission bandwidth, and, consequently, provides a larger information capacity.

Although these transmission links generally made use of radio, microwave, and copper-wire technologies, there has always been an interest in using light to communicate [1–4]. The reason for this is that, in addition to the optical fiber's inherently wide bandwidth capability, its dielectric nature renders it immune to electromagnetic interference and offers excellent electrical isolation, particularly in electrically hazardous environments. In addition, its low weight and hair-sized dimensions offer a distinct advantage over large, heavy copper cables, which is important not only for saving space in underground and indoor ducts but also for reducing the size and weight of cables on aircraft and in ships. Kao and Hockman first proposed the use of low-loss glass fiber in 1966, when they suggested that the intrinsic loss of silica-based glass could be made low enough to enable its use as a guiding channel for light [5]. The fabrication of a low-loss optical fiber by researchers at Corning in 1970 provided the key technology for finally realizing this in a practical way [6].

The optical spectrum ranges from about 50 nm (ultraviolet) to about 100 μm (far infrared), the visible region being the 400–700-nm band. Optical fiber communication systems operate in the 800–1600-nm wavelength

band. In optical systems it is customary to specify the band of interest in terms of wavelength, instead of frequency as in the radio region. However, with the advent of high-speed multiple-wavelength systems in the mid-1990s, researchers began specifying the output of optical sources in terms of optical frequency. The reason for this is that in optical sources such as mode-locked semiconductor lasers, it is easier to control the frequency of the output light, rather than the wavelength, in order to tune the device to different emission regions. Of course, the different optical frequencies ν are related to the wavelengths λ through the fundamental equation $c = \nu\lambda$. Thus, for example, a 1552.5-nm wavelength light signal has a frequency of 193.1 THz (193.1×10^{12} Hz).

1.2. Optical Fiber Link Applications

Communication networks composed of optical fiber links are sometimes referred to as *lightwave* or *photonic* systems. Network architectures using multiple wavelength channels per optical fiber can be utilized in local-area, metropolitan-area, or wide-area applications to connect hundreds or thousands of users having a wide range of transmission capacities and speeds. The use of multiple wavelengths greatly increases the capacity, configuration flexibility, and growth potential of this backbone. Moderate-speed regional networks attached to this backbone provide applications such as interconnection of telephone switching centers, access to satellite transmission facilities, and access to mobile-phone base stations. More localized, lower-speed networks offer a wide variety of applications such as telephony services to homes and businesses, distance learning, Internet access, CATV (cable television), security surveillance, and electronic mail (email). A major motivation for developing these sophisticated networks has been the rapid proliferation of information exchange desired by institutions such as commerce, finance, education, health, government, security, and entertainment. The potential for this information exchange arose from the ever-increasing power of computers and data-storage devices.

Once researchers showed in 1970 that it was possible to make low-loss fibers, the optical fiber communication field expanded rapidly to provide a broadband medium for transporting voice, video, and data traffic. In fact, optical fiber technology has been a key factor contributing to the extraordinary growth of global telecommunications. Optical fiber was being installed worldwide at the rate of 4800 km per hour by the year 2000, which is equivalent to a cable-laying rate of three time around the world every day [7,8]. Along with this high installation rate come numerous technological advances in photonic components. These advances permit more and more wavelengths to be transmitted at ever-increasing speeds on an individual optical fiber, which is resulting in an annual two-fold increase in the data-carrying capacity of an individual fiber strand. Table 1 illustrates this with a few of the many installations that have taken place since 1980. As shown in that table, in the year 2000, commercial systems were capable of transmitting 400 Gbps (gigabits per second) over distances of 640 km without regenerating the signal. To put this in perspective, this is equivalent to sending 12,000 encyclopedic volumes every second.

1.3. Basic Link Elements

Figure 1 shows typical components that are found within an optical fiber link. The key sections are a transmitter consisting of a light source and its associated drive circuitry, a cable offering mechanical and environmental protection to the optical fibers contained inside, and a receiver consisting of a photodetector plus amplification and signal-restoring circuitry. Additional components include optical amplifiers, connectors, splices, couplers, and regenerators (for restoring the signal shape characteristics). The cabled fiber is one of the most important elements in an optical fiber link. In addition to protecting the glass fibers during installation and service, it may contain copper wires for powering optical amplifiers or signal regenerators, which are needed periodically in long-distance links for amplifying and reshaping the signal.

Analogous to copper cables, the installation of optical fiber cables can be either aerial, in underground or indoor ducts, undersea, or buried directly in the ground. As a result of installation and/or manufacturing limitations, individual cable lengths will range from several hundred meters to several kilometers for terrestrial links. Cable lengths for oceanic links can be several tens of kilometers. Practical considerations such as reel size and cable weight determine the actual length of a single cable section. The shorter segments tend to be used when the cables are pulled through ducts. Longer lengths are used in aerial,

Table 1. Examples of Types of Multimode (MM) and Single-Mode (SM) Optical Fiber Systems Installed Since 1980

Year	Fiber Type	Wavelength (nm)	WDM Channels	Bit Rate per Channel	Bit Rate per Fiber	Regenerator Spans (km)
1980	MM	820	1	45 Mbps	45 Mbps	7
1985	SM	1300	1	417 Mbps	417 Mbps	50
1987	SM	1300	1	1.7 Gbps	1.7 Gbps	50
1992	SM	1300	1	2.5 Gbps	2.5 Gbps	50
1995	SM	1550	8	2.5 Gbps	20 Gbps	360
1997	SM	1550	16	2.5 Gbps	40 Gbps	360
1999	SM	1550	80	2.5 Gbps	200 Gbps	640
1999	SM	1550	40	10 Gbps	400 Gbps	640
2000	SM	1550	80	10 Gbps	800 Gbps	500

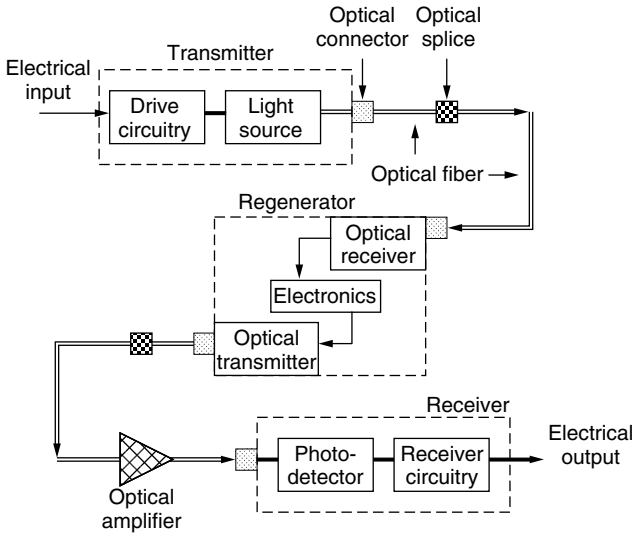


Figure 1. Typical components found within an optical fiber link.

directly buried, or undersea applications. Splicing together individual cable sections forms continuous transmission lines for these long-distance links. For undersea installations, the splicing and repeater-installation functions are carried out on board a specially designed cable-laying ship [9].

An optical fiber nominally is a thin cylindrical strand of two layers of glass surrounded by an elastic buffer coating, as shown in Fig. 2. The central cylinder has a radius a and an index of refraction n_1. This cylinder is known as the *core* of the fiber. The core is surrounded by a solid glass *cladding*, which has a refractive index n_2 that is slightly less than n_1. Since the core refractive index is larger than the cladding index, electromagnetic energy at optical frequencies can propagate along the fiber core through internal reflection at the core–cladding interface. Single-mode fibers, which sustain a single propagating mode along the core, have nominal core diameters of 8–12 μm.

One of the principal characteristics of an optical fiber is its attenuation as a function of wavelength, as shown in Fig. 3. Early technology made exclusive use of the 800–900-nm wavelength band, since in this region the fibers made at that time exhibited a local minimum in the attenuation curve, and optical sources and photodetectors operating at these wavelengths were available. This region is referred to as the *first window*. The large attenuation spikes in early fibers were due to absorption by water molecules (hydroxyl ions) in the glass. By reducing the concentration of hydroxyl ions and metallic impurities in the fiber material, in the 1980s manufacturers were able to fabricate optical fibers with very low loss in the 1100–1600-nm region. This spectral band is referred to

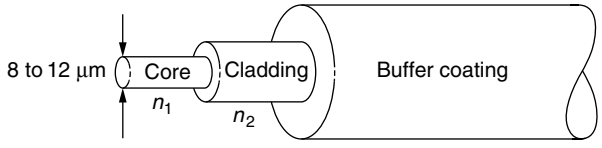


Figure 2. Physical configuration of an optical fiber.

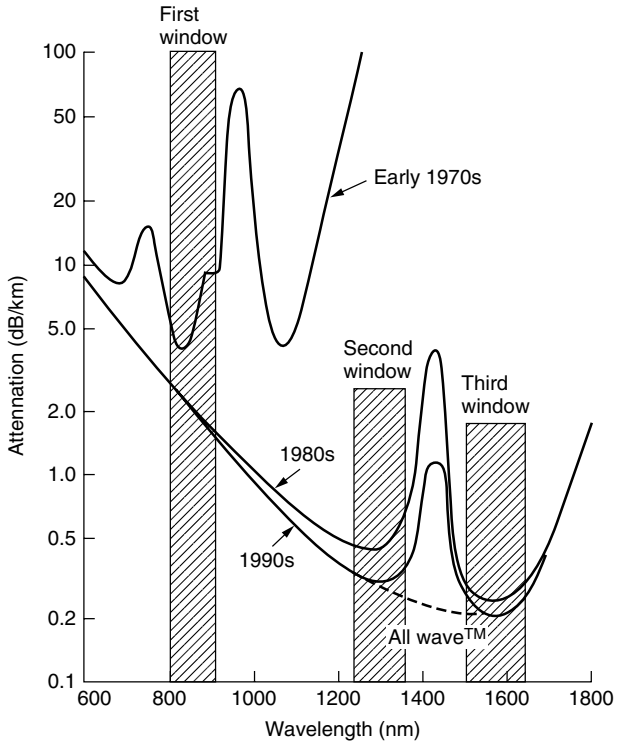


Figure 3. Attenuation as a function of wavelength.

as the *long-wavelength region*. As Fig. 3 shows, the two windows defined here are called the *second window* centered around 1310 nm and the *third window*, centered around 1550 nm.

In 1998 a new ultra-high-purification process patented by Lucent Technologies eliminated virtually all water molecules from the glass fiber material. By dramatically reducing the water attenuation peak around 1400 nm, this process opened the transmission region between the second and third windows to provide around 100 nm more bandwidth than in conventional single-mode fibers, as shown by the dashed line in Fig. 3. This particular AllWave fiber, which was specifically designed for metropolitan networks, gave local service providers the ability to cost-effectively deliver up to hundreds of optical wavelengths simultaneously.

Once the cable is installed, a light source that is dimensionally compatible with the fiber core is used to launch optical power into the fiber. Semiconductor light-emitting diodes (LEDs) and laser diodes are suitable for this purpose, since their light output can be modulated rapidly by simply varying the bias current at the desired transmission rate, thereby producing an optical signal. The electric input signals to the transmitter circuitry for the optical source can be either of an analog or digital form. For high-rate systems (usually greater than 1 Gbps), direct modulation of the source can lead to unacceptable signal distortion. In this case, an external modulator is used to vary the amplitude of a continuous light output from a laser diode source. In the 800–900-nm region the light sources are generally alloys of GaAlAs. At longer wavelengths (1100–1600 nm) an InGaAsP alloy is the principal optical source material.

After an optical signal is launched into a fiber, it will become progressively attenuated and distorted with increasing distance because of scattering, absorption, and dispersion mechanisms in the glass material. At the receiver a photodiode will detect the weakened optical signal emerging from the fiber end and convert it to an electric current (referred to as a *photocurrent*). Silicon photodiodes are used in the 800–900-nm region. The primary photodiode material in the 1100–1600-nm region is an InGaAs alloy.

The design of an optical receiver is inherently more complex than that of the transmitter, since it has to interpret the content of the weakened and degraded signal received by the photodetector. The principal figure of merit for a receiver is the maximum optical power necessary at the desired data rate to attain either a given error probability for digital systems or a specified signal-to-noise ratio for an analog system. The ability of a receiver to achieve a certain performance level depends on the photodetector type, the effects of noise in the system, and the characteristics of the successive amplification stages in the receiver.

1.4. Wavelength-Division Multiplexing

An interesting and powerful aspect of an optical communication link is that many different wavelengths can be sent along a fiber simultaneously in the 1300–1600-nm spectrum. The technology of combining a number of wavelengths onto the same fiber is known as *wavelength-division multiplexing* (WDM). Figure 4 shows the basic WDM concept [10,11]. Here N independent optically formatted information streams, each transmitted at a different wavelength, are combined with an optical multiplexer and sent over the same fiber. Note that each of these streams could be at a different data rate. Each information stream maintains its individual data rate after being multiplexed with the other streams, and still operates at its unique wavelength. Conceptually, the WDM scheme is the same as frequency-division multiplexing (FDM) used in microwave radio and satellite systems. WDM systems in which the channels are closely spaced are referred to as *dense WDM* (DWDM) systems.

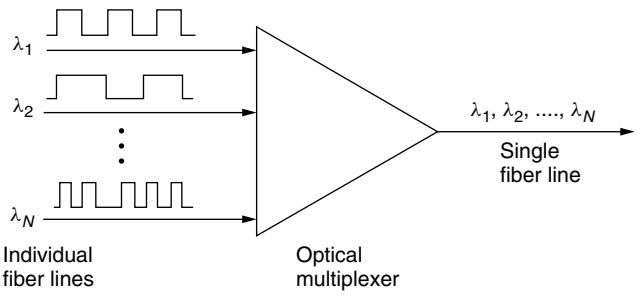


Figure 4. The basic WDM concept.

To realize this WDM scheme, one needs specialized components to efficiently multiplex an aggregate of wavelength channels into an optical fiber at one point and to divide them into their original individual channels at another location. Other WDM components are used to selectively add or drop one or more channels at specific points along a fiber link. These components include fiber Bragg gratings, arrayed waveguide gratings, dielectric thin-film interference filters, acoustooptic tunable filters, and Mach–Zehnder filters.

Figure 5 shows a simple concept of a demultiplexing function using a fiber Bragg grating. To extract the desired wavelength, a *circulator* is used in conjunction with the grating. In a three-port circulator, an input signal on one port exits at the next port. For example, an input signal at port 1 is sent out at port 2. Here, the circulator takes the four wavelengths entering port 1 and sends them out at port 2. All wavelengths except λ_2 pass through the grating. Since λ_2 satisfies the Bragg condition of the grating, it is reflected, enters port 2 of the circulator, and exits at port 3. More complex multiplexing and demultiplexing structures with several gratings and several circulators can be realized with this scheme.

1.5. Optical Amplifiers

Traditionally, when setting up an optical link, one formulates a power budget and adds repeaters when the path loss exceeds the available power margin. To amplify an optical signal with a conventional repeater, one performs photon-to-electron conversion, electrical amplification, retiming, pulseshaping, and then electron-to-photon conversion. Although this process works well for moderate-speed single-wavelength operation, it can be fairly complex and expensive for high-speed multiwavelength systems. Thus, a great deal of effort has been expended to develop all-optical amplifiers for the 1300-nm and the 1550-nm long-wavelength transmission windows of optical fibers. The main ones in use are *erbium-doped fiber amplifiers* (EDFAs) and *Raman fiber amplifiers* [12,13]. An EDFA can amplify optical signals in the 1530–1610-nm range, whereas Raman fiber amplifiers can amplify signals from 1270 to 1670 nm.

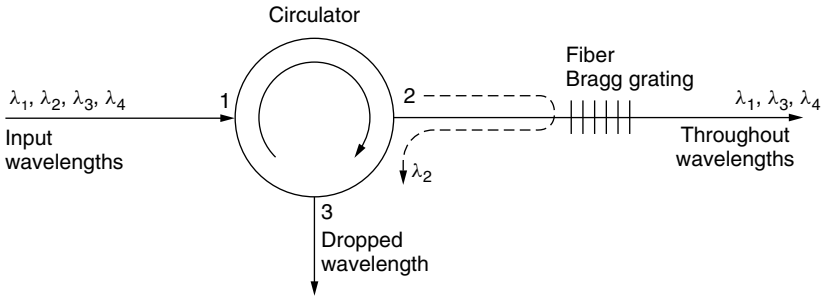


Figure 5. A simple wavelength demultiplexing function using a fiber Bragg grating.

Optical amplifiers have found widespread use in not only long-distance point-to-point optical fiber links but also in multiple-access networks to compensate for signal-splitting losses. The features of optical amplifiers has led to many diverse applications, each having different design challenges. Figure 6 shows general applications of optical amplifiers where the parameter G denotes gain.

In a single-mode link the effects of fiber dispersion may be small so that the main limitation to repeater spacing is fiber attenuation. Since such a link does not necessarily require a complete regeneration of the signal, simple amplification of the optical signal is sufficient. Thus an *inline optical amplifier* can be used to compensate for transmission loss and increase the distance between regenerative repeaters, as illustrated in Fig. 6a.

Figure 6b shows an optical amplifier being used as a front-end *preamplifier* for an optical receiver. In this way, a weak optical signal is amplified before photodetection so that the signal-to-noise ratio degradation caused by thermal noise in the receiver electronics can be suppressed. Compared with other front-end devices such as avalanche photodiodes or optical heterodyne detectors, an optical preamplifier provides a larger gain factor and a broader bandwidth.

Power or *booster amplifier* applications include placing the device immediately after an optical transmitter to boost the transmitted power, as Fig. 6c shows. This serves to increase the transmission distance by 10–100 km depending on the amplifier gain and fiber loss. As an example, using this boosting technique together with an optical preamplifier at the receiving end can enable repeaterless undersea transmission distances of 200–250 km. One can also employ an optical amplifier in a local-area network as a booster amplifier to compensate for coupler insertion loss and power-splitting loss. Figure 6d shows an example for boosting the optical signal in front of a star coupler.

2. LINK PERFORMANCE CHARACTERISTICS

The transmission characteristics are a major factor in determining what a signal looks like after it has propagated a certain distance. The three fundamental signal-distorting factors are attenuation, dispersion, and nonlinear effects in an optical fiber.

2.1. Attenuation

Attenuation of a light signal as it propagates along a fiber is an important consideration in the design of an optical communication system, since it plays a major role in determining the maximum transmission distance between a transmitter and a receiver. The basic attenuation mechanisms in a fiber are absorption, scattering, and radiative losses of the optical energy. Absorption is related to the fiber material, whereas scattering is associated both with the fiber material and with structural imperfections in the optical waveguide. Attenuation owing to radiative effects originates from perturbations (both microscopic and macroscopic) of the fiber geometry. As light travels along a fiber, its power decreases exponentially with distance. If $P(0)$ is the optical power in a fiber at the origin (at $z = 0$), then the power $P(z)$ at a distance z further down the fiber is

$$P(z) = P(0)e^{-\alpha_p z} \tag{1}$$

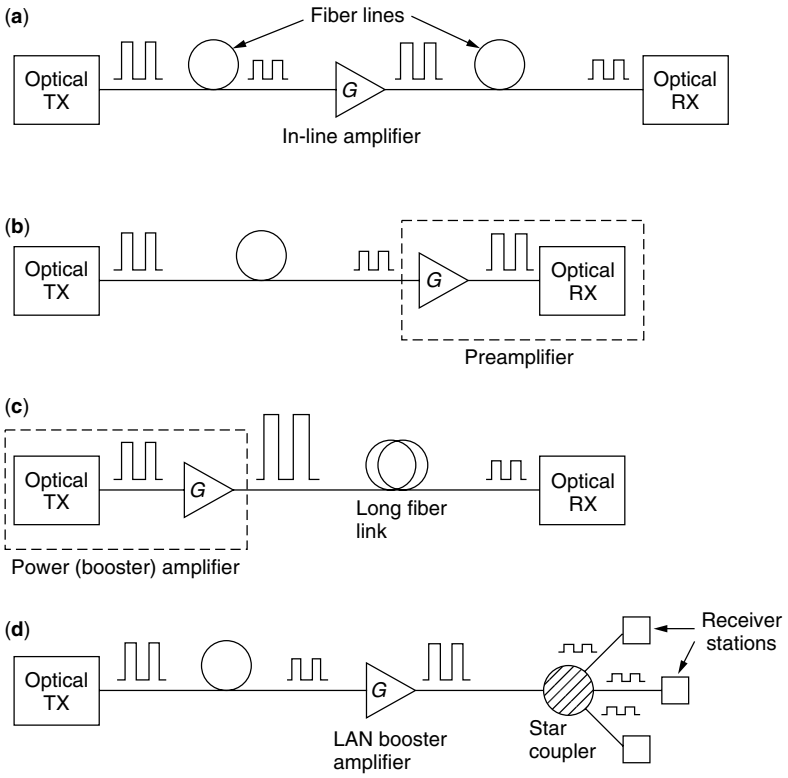


Figure 6. General applications of optical amplifiers.

where

$$\alpha_p = \frac{1}{z} \ln \left[\frac{P(0)}{P(z)} \right] \tag{2}$$

is the fiber *attenuation coefficient* given in units of, for example, reciprocal kilometers (km^{-1}). For simplicity in calculating optical signal attenuation in a fiber, the common procedure is to express the attenuation coefficient in units of *decibels per kilometer*, (dB/km). Designating this parameter by α, we have

$$\alpha(\text{dB/km}) = \frac{10}{z} \log \left[\frac{P(0)}{P(z)} \right] = 4.343\, \alpha_p(\text{km}^{-1}) \tag{3}$$

This parameter is generally referred to as the *fiber loss* or the *fiber attenuation*. Figure 3 illustrates the attenuation of optical power in a fiber as a function of wavelength.

2.2. Dispersion

Several different dispersion mechanisms in an optical fiber cause a light signal to become increasingly distorted as it travels along a fiber. These include material, waveguide, and polarization-mode dispersions [4,14]. *Material dispersion* arises because each optical pulse in a digital signal contains a small range of wavelengths. Since the refractive index n of silica glass varies slightly as a function of wavelength, the fundamental relationship for the velocity v in a material $v = c/n$, where c is the speed of light, shows that different parts of the pulse will travel at different speeds. Consequently, as a result of this material dispersion effect, a pulse will broaden as it travels along a fiber.

Waveguide dispersion occurs because a single-mode fiber confines only about 80% of the optical power to the core. The other 20% propagates in the cladding that surrounds the core. Dispersion arises since the light in the cladding sees a lower refractive index than in the core and thus travels faster than the light confined in the core. The amount of waveguide dispersion depends on the fiber design. Thus, through ingenious fiber construction, waveguide dispersion can be tailored to counteract the effects of material dispersion. In standard fiber designs, material and waveguide dispersions cancel at 1310 nm. To achieve zero total dispersion at 1550 nm, where the attenuation of a silica fiber is at its lowest point, the *dispersion-shifted fiber* was developed in the mid-1980s. This works well for single-wavelength operation, but is not desirable in WDM systems. Here nonlinear effects require different approaches, one of which is the dispersion management scheme described below.

Polarization-mode dispersion arises from the effects of fiber birefringence on the polarization states of an optical pulse [15]. *Birefringence* refers to slight variations in the indices of refraction along different axes of the fiber. This is particularly critical for high-rate, long-haul transmission links (e.g., 10 Gbps over tens of kilometers) that are designed to operate near the zero-dispersion wavelength of the fiber. Birefringence can result from intrinsic factors such as geometric irregularities of the fiber core or internal stresses on it. Deviations of less than 1% in the circularity of the core can already have a noticeable effect in a high-speed lightwave system. In addition, external factors such as bending, twisting, or pinching of the fiber can also lead to birefringence. Since all these mechanisms exist to some extent in any field-installed fiber, there will be a varying birefringence along its length.

A fundamental property of an optical signal is its polarization state. *Polarization* refers to the electric field orientation of a light signal, which can vary significantly along the length of a fiber. As shown in Fig. 7, signal energy at a given wavelength occupies two orthogonal polarization modes. A varying birefringence along the length of the fibre will cause all the polarization modes to travel at slightly different velocities and the polarization orientation will rotate with distance. The resulting difference $\Delta\tau$ in propagation times between the two orthogonal polarization modes will result in pulse spreading. This is known as *polarization-mode dispersion* (PMD).

2.3. Nonlinear Effects

Two different categories of optical nonlinear effects also have an adverse effect on signal quality [16–18]. The first category encompasses nonlinear inelastic scattering processes, which are interactions between optical signals and molecular or acoustic vibrations in a fiber. These

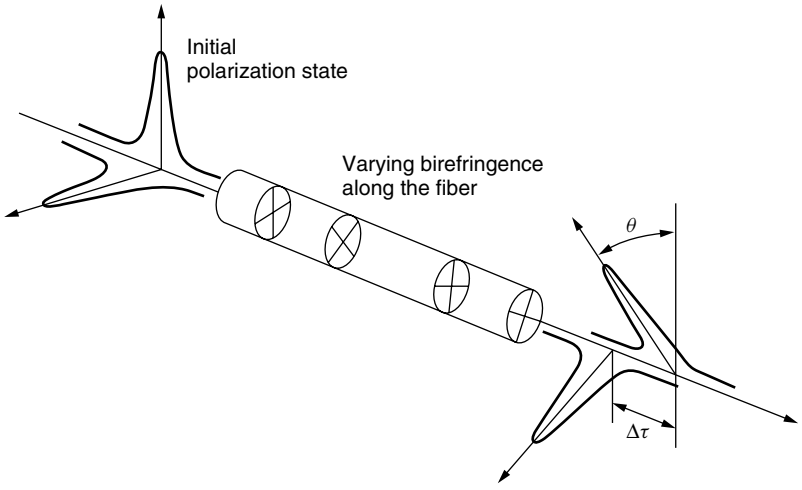


Figure 7. Signal energy at a given wavelength occupies two orthogonal polarization modes.

are stimulated Raman scattering (SRS) and stimulated Brillouin scattering (SBS). The second category involves nonlinear variations of the refractive index in a silica fiber that occur because the refractive index is dependent on intensity changes in the signal. This produces effects such as self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). In the literature, FWM is also referred to as *four-photon mixing* (FPM), and XPM is sometimes designated by CPM.

SRS, SBS, and FWM result in gains or losses in a wavelength channel that are dependent on the optical signal intensity. These nonlinear processes provide gains to some channels while depleting power from others, thereby producing crosstalk between the channels. SPM and XPM affect only the phase of signals, which causes chirping in digital pulses. This can worsen pulse broadening due to dispersion, particularly in very high-rate systems (>10 Gbps). When any of these nonlinear effects contribute to signal impairment, an additional amount of power will be needed at the receiver to maintain the same BER (bit error rate) as in their absence. This additional power (in decibels) is known as the *power penalty* for that effect.

Stimulated Raman scattering is an interaction between lightwaves and the vibrational modes of silica molecules. If a photon with energy $h\nu_1$ is incident on a molecule having a vibrational frequency ν_m, the molecule can absorb some energy from the photon. In this interaction the photon is scattered, thereby attaining a lower frequency ν_2 and a corresponding lower energy $h\nu_2$. The modified photon is called a *Stokes photon*. Because the optical signal wave that is injected into a fiber is the source of the interacting photons, it is called the *pump wave*, since it supplies power for the generated wave. This process generates scattered light at a wavelength longer than that of the incident light. If another signal is present at this longer wavelength, the SRS light will amplify it and the pump wavelength signal will decrease in power. Consequently, SRS can severely limit the performance of a multichannel optical communication system by transferring energy from short-wavelength channels to neighboring higher-wavelength channels. This is a broadband effect that can occur in both directions. Powers in channels separated by up to 16 THz (125 nm) can be coupled through the SRS effect, thereby producing cross talk between wavelength channels. SRS may be controlled using fiber dispersion management techniques, as described below.

Stimulated Brillouin scattering arises when lightwaves scatter from acoustic waves. The resultant scattered wave propagates principally in the backward direction in single-mode fibers. This backscattered light experiences gain from the forward-propagating signals, which leads to depletion of the signal power. In silica this interaction occurs over a very narrow *Brillouin linewidth* of 20 MHz at 1550 nm. This means that the SBS effect is confined within a single wavelength channel in a WDM system, and thus accumulates individually for each channel. System impairment starts when the amplitude of the scattered wave is comparable to the signal power. For typical fibers the threshold power for this process is around 10 mW for single-fiber spans. In a long fiber chain containing optical amplifiers, there are normally optical isolators to prevent backscattered signals from entering the amplifier. Consequently, the impairment due to SBS is limited to the degradation occurring in a single amplifier-to-amplifier span. Several schemes are available for suppressing the effects of SBS [4].

The refractive index n of the glass material in an optical fiber has a weak dependence on optical intensity (equal to the optical power per effective area in the fiber). Since the intensity varies at the leading and trailing edges of each optical pulse in a digital datastream, the nonlinearity in the refractive index produces a carrier-induced phase modulation of the propagating signal. Consequently, parts of the pulse undergo a frequency shift in a process called *self-phase modulation* (SPM). As a result of dispersion in the fiber, this shift is transformed into pulse distortion. The effects of SPM can be reduced by maintaining a low overall dispersion in a fiber link.

In WDM systems, the refractive index nonlinearity gives rise to *cross-phase modulation* (XPM), which converts power fluctuations in a particular wavelength channel to phase fluctuations in other copropagating channels. This can be greatly mitigated in WDM systems operating over standard nondispersion-shifted single-mode fiber, but can be a significant problem in WDM links operating at 10 Gbps and higher over dispersion-shifted fiber. To mitigate XPM effects, the dispersion should be high, since pulses in different channels travel at different speeds, so that they walk through each other quickly, thereby minimizing their interactions.

Four-wave mixing (FWM) is a third-order nonlinearity in silica fibers, which is analogous to intermodulation distortion in electrical systems. The FWM effect arises from the beating between two or more channels, which creates new tones at other frequencies. When these new frequencies fall in the transmission window of the original frequencies, it can cause severe crosstalk. For DWDM systems, FWM can cause the highest power penalty of all the nonlinear effects. The efficiency of four-wave mixing depends on fiber dispersion and the channel spacings. Since the dispersion varies with wavelength, the signal waves and the generated waves have different group velocities. This destroys the phase matching of the interacting waves and lowers the efficiency at which power is transferred to newly generated frequencies. The higher the group velocity mismatches and the wider the channel spacing, the lower the effects of four-wave mixing.

2.4. Dispersion Management

Using current fiber designs, high-speed WDM systems are limited by nonlinear effects and dispersion. To mitigate these effects, dispersion management techniques are being used to maintain moderate dispersion locally and near-zero dispersion globally across the entire link [19–21]. This needs to be implemented across the wide spectrum of wavelengths used in WDM systems. To achieve this, one may use passive *dispersion compensation*. This consists of inserting into the link a loop of fiber having a dispersion characteristic that negates the accumulated dispersion of the transmission fiber. The fiber loop is referred to as a *dispersion-compensating fiber* (DCF). If the transmission

fiber has a low positive dispersion [say, 2.3 ps/(nm · km)], then the DCF will have a large negative dispersion [say, −16 ps/(nm · km)].

With this technique, the total accumulated dispersion is zero after some distance, but the absolute dispersion per length is nonzero at all points along the fiber. The nonzero absolute value causes a phase mismatch between wavelength channels, thereby destroying the possibility of effective FWM production.

3. MEASUREMENT METHODOLOGIES

The design and installation of an optical fiber communication system require measurement techniques for verifying the operational characteristics of the constituent components [4,22]. In addition to optical fiber parameters, system engineers are interested in knowing the characteristics of passive splitters, connectors, and couplers, and electrooptic components, such as sources, photodetectors, and optical amplifiers. Furthermore, when a link is being installed and tested, the operational parameters of interest include bit error rate, timing jitter, and signal-to-noise ratio as indicated by the eye pattern. During actual operation, measurements are needed for maintenance and monitoring functions to determine factors such as fault locations in fibers and the status of remotely located optical amplifiers.

4. FURTHER INFORMATION

Many of the concepts covered in this article are described in more detail elsewhere in this encyclopedia. For example, see: articles on nonlinear effects in fibers, optical amplifiers, optical couplers, optical fiber dispersion, optical filters, optical networks, optical receivers, optical transmitters, standards, and wavelength-division multiplexing.

BIOGRAPHY

Gerd Keiser is the founder and president of PhotonicsComm Solutions, Inc., Newton Center, Massachusetts, a firm specializing in consulting and education for the optical communications industry. He has 25 years experience at Honeywell, GTE, and General Dynamics in designing and analyzing telecommunication components, links, and networks. He is the author of the books *Optical Fiber Communications* (3rd ed. 2000) and *Local Area Networks* (2nd ed. 2002) published by McGraw-Hill. Dr. Keiser is an IEEE fellow and received GTE's prestigious Leslie Warner Award for work in ATM switch development. He earned his B.A. and M.S. degrees in mathematics and physics from the University of Wisconsin and a Ph.D. in solid state physics from Northeastern University, Boston, Massachusetts.

BIBLIOGRAPHY

1. D. J. H. Maclean, *Optical Line Systems*, Wiley, Chichester, UK, 1996 (this book gives a detailed discussion of the evolution of optical fiber links and networks).
2. R. Ramaswami and K. N. Sivarajan, *Optical Networks*, 2nd ed., Morgan Kaufmann, San Francisco, 2002.
3. J. Hecht, *City of Light: The Story of Fiber Optics*, Oxford Univ. Press, 1999.
4. G. Keiser, *Optical Fiber Communications*, 3rd ed., McGraw-Hill, Burr Ridge, IL, 2000.
5. K. C. Kao and G. A. Hockman, Dielectric-fiber surface waveguides for optical frequencies, *Proc. IEE* **133**: 1151–1158 (July 1966).
6. F. P. Kapron, D. B. Keck, and R. D. Maurer, Radiation losses in glass optical waveguides, *Appl. Phys. Lett.* **17**: 423–425 (Nov. 1970).
7. S. Tsuda and V. L. da Silva, Transmission of 80 × 10 Gbps WDM channels with 50-GHz spacing over 500 km of LEAF fiber, *Tech. Digest IEEE/OSA Optical Fiber Commun. Conf.*, March 2000, pp. 149–151.
8. R. C. Alferness, H. Kogelnik, and T. H. Wood, The evolution of optical systems: Optics everywhere, *Bell Labs Tech. J.* **5**: 188–202 (Jan.–March. 2000).
9. Special issue on "Undersea Communications Technology," *AT&T Tech. J.* **74**: (Jan./Feb. 1995).
10. G. E. Keiser, A review of WDM technology and applications, *Opt. Fiber Technol.* **5**: 3–39 (Jan. 1999).
11. S. V. Kartalopoulos, *Introduction to DWDM Technology*, IEEE Press, New York, 2000.
12. E. Desurvire, *Erbium-Doped Fiber Amplifiers*, Wiley, New York, 1994.
13. H. Masuda, Review of wideband hybrid amplifiers, *Tech. Digest IEEE/OSA Optical Fiber Commun. Conf.*, March 2000, pp. 2–4.
14. P. Hernday, Dispersion measurements, in D. Derickson, ed., *Fiber Optic Test and Measurement*, Prentice-Hall, Upper Saddle River, NJ, 1998.
15. C. D. Poole and J. Nagel, Polarization effects in lightwave systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications — III*, Vol. A, Academic Press, New York, 1997, Chap. 6, pp. 114–161.
16. G. P. Agrawal, *Nonlinear Fiber Optics*, 2nd ed., Academic Press, New York, 1995.
17. F. Forghieri, R. W. Tkach, and A. R. Chraplyvy, Fiber nonlinearities and their impact on transmission systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Telecommunications — III*, Vol. A, Academic Press, New York, 1997, Chap. 8, pp. 196–264.
18. E. Iannone, F. Matera, A. Mecozzi, and M. Settembre, *Nonlinear Optical Communication Networks*, Wiley, New York, 1998.
19. B. Jopson and A. H. Gnauck, Dispersion compensation for optical fiber systems, *IEEE Commun. Mag.* **33**: 96–102 (June 1995).
20. L. Grüner-Nielsen et al., Dispersion compensating fibers, *Opt. Fiber Technol.* **6**: 164–180 (April 2000).
21. M. Murakami, T. Matsuda, H. Maeda, and T. Imai, Long-haul WDM transmission using higher-order fiber dispersion management, *J. Lightwave Technol.* **18**: 1197–1204 (Sept. 2000).
22. D. Derickson, ed., *Fiber Optic Test and Measurement*, Prentice-Hall, Upper Saddle River, NJ, 1998.

OPTICAL FIBER LOCAL AREA NETWORKS

MEHDI SHADARAM
VIRGILIO E. GONZALEZ-LOZANO
University of Texas at El Paso
El Paso, Texas

1. INTRODUCTION

Technological advances in the analog, digital, and photonic systems have transformed communications into a highly dynamic field. Nowadays, communication between computers is an essential part of modern living, and telecommunications is one of world's fastest-growing industries. Most advances in this field are driven by social, economical, political, and technological reasons. It is a well-known fact that without a reliable communication infrastructure, nations cannot retain a prosperous economy. The need for computer and communication engineers to implement new ideas and to satisfy the growing demand for higher bandwidth is apparent more than ever. The data rate of computer networks used in university campuses, hospitals, banks, and elsewhere is doubling almost every year. The data rate has reached a point where a single transmission medium such as twisted-pair or coaxial cable is not capable of transmitting the load. In order to avoid multiple-cables laying beneath the ground and overloading buildings with wires, the need for one single transmission medium that can convey up to several gigabits of information per second is necessary. Since the early 1970s optical fibers have been utilized as transmission media in long- and short-distance communication links. Typical optical fiber attenuation has decreased from several dB/km at 0.8 μm wavelength in the early seventies to about 0.1 dB/km at 1.55 μm wavelength as we enter the new millennium. During the same period, the capacity of optical fibers has increased from several Mbps/km (megabits per second per kilometer) to $\sim$300 Gbps/km. Because of gradual maturing of multiwavelength optical fiber systems, wavelength-division multiplexing (WDM), and development of zero-dispersion optical fibers (solitons), the capacity of a single optical fiber could reach thousands of Gbps in the near future. Small size, light weight, and immunity to electromagnetic interference noise also provide a crucial advantage for optical fibers over other media. Although cost of implementing fiber-based systems still exceeds the cost to deploy other systems such as coaxial or twisted pair, prices for optical fiber systems are dropping as rapidly as 30% per year.

2. WHAT IS AN OPTICAL FIBER LOCAL AREA NETWORK?

Organizations such as universities, hospitals, banks, or even small offices use computers for a variety of applications. Very often different users within these establishments need to share data. Thus, for a reliable and high-speed data transfer, computers are connected through a network of point-to-point communication links. These types of networks are usually referred to as *local-area networks* (LANs). The size of a LAN can vary anywhere from two computers connected to one another in a room to several thousand computers connected together in a large campus. The distance between nodes within a LAN can vary from a few meters to as much as 2 km. If computers are connected via optical fiber links, the LAN is referred to as *optical fiber LAN*. Optical fiber cables conduct light along a thin solid cylinder-shaped glass or plastic fiber, known as a *core*. The core is surrounded by *cladding*, which in turn is surrounded by a plastic sheath as shown in Fig. 1a. Depending on the fiber type, the core diameter can be in the range of 8–62 μm. The typical diameter of cladding for optical fibers used in telecommunication is about 125 μm. In order to confine the light within the core, as shown in Fig. 1b, the refractive index of the core is slightly higher than the refractive index of the cladding, as can be seen in Fig. 1c. The block diagram shown in Fig. 2 exhibits a typical optical fiber link used between two nodes within the network. The link typically consists of a laser diode (LD) or light-emitting diode (LED) in the transmitter unit, which launches an optically modulated signal into an optical fiber cable. At the receiver site, the cable is terminated by a photodiode, which converts the optical signal into an electric current.

In the early days of optical fiber technology, because of major developments in gallium arsenide devices, most fiber links were operating at short wavelengths ($\sim$0.8 μm). These links could carry bit rates of $\leq$100 Mbps with an attenuation of $\sim$10 dB/km. Nowadays, most optical fiber links operate at around 1.3 μm, minimum dispersion wavelength; or 1.55 μm, minimum attenuation wavelength. Development of dispersion-shifted fibers has made it possible for fibers to exhibit minimum dispersion and minimum attenuation at 1.55 μm. Links employing dispersion-shifted fibers are commonly used for long-distance communication purposes.

The main purpose of a LAN is to share resources such as files, printers, programs, and databases among the nodes in the network. The initial LAN designs were intended for communication between computers at short distances. However, the length of the links has grown from a few meters to a few kilometers.

To reach longer distances there are wide-area networks (WANs), which typically employ the infrastructure of a telecom service provider or carrier. A single network manager usually controls the LAN operation, but the WAN requires the coordination between the service provider and the LAN administrator. Very often, carrier networks require an initial setup of a connection between the end nodes, while LANs do not require that setup.

Through the years the capacity and the transmission distance of the networks have increased. The process has created a hybrid type of network between a LAN and a WAN. This type of network is called a *metropolitan-area network* (MAN). The MAN is capable of covering an entire city using the simplicity of the LAN protocols at 100 Mbps and above. On the other hand, there is a new specialized small network called *storage-area network* (SAN). The purpose of a SAN is to allow very-high-speed communication between computer processors and dedicated peripherals such as large disk arrays.

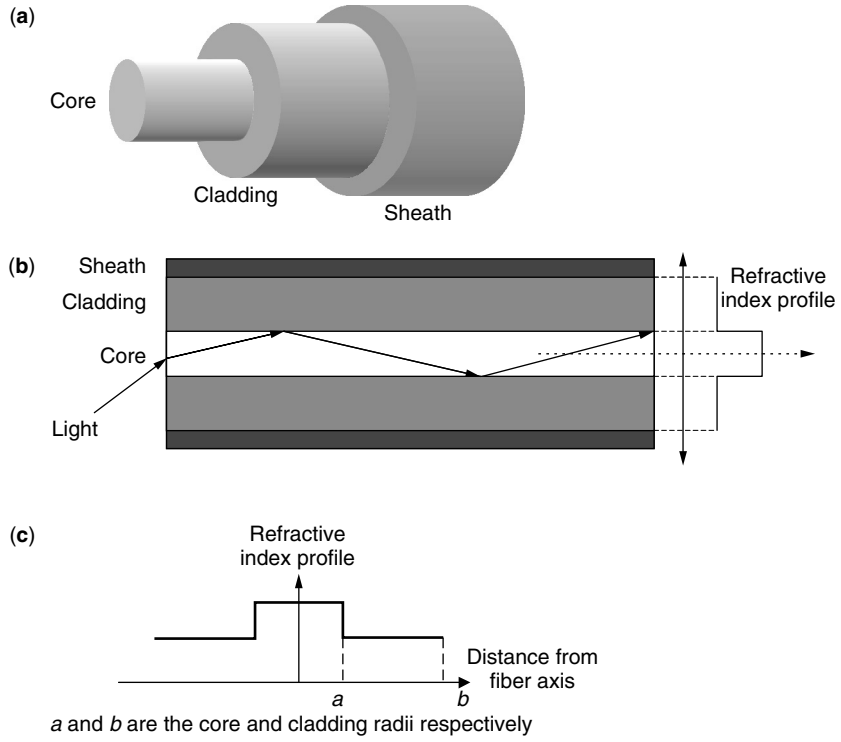


Figure 1. Optical fiber.

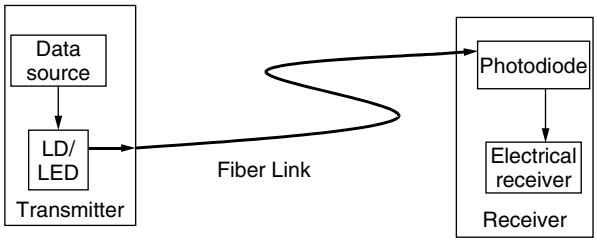


Figure 2. Typical optical fiber link.

3. NETWORK TOPOLOGIES

As pointed out previously, a network allows users to share information in an efficient, fast, and reliable manner. As images, audio, and video files become a regular part of the datastream, the need for networks with higher bandwidth becomes more apparent. Increasing the number of point-to-point links within a network or increasing the bandwidth of point-to-point links can increase the network bandwidth. Optical fiber is usually favored over other cables, since it is capable of very high data transfer rates. An optical link can replace an electric link as long as all electrical interface requirements such as voltage or current signal levels, timing, and control signals are met.

Implementation of an optical fiber LAN may follow one of two approaches. The first method involves the creation of a completely new network with optical fiber links. The second approach requires the replacement, within a conventional LAN, of electric links with optical links while meeting all original interface requirements. As a result, many LANs expand their reach by adding special repeaters with fiber segments.

There are two types of signal couplings to the fiber, known as active and passive. Figure 3a shows that, for active coupling, the network behaves like a series of point-to-point links and each node must be operational to maintain the network working. For passive coupling, stations broadcast the signals into a common fiber and each node captures only a small amount of their power. As seen in Fig. 3b, an inactive station does not affect the operation of the network. An alternative for inactive nodes is to include an optical bypass at each station, as shown in Fig. 4.

The network topology establishes a path for the flow of information between the transmitter and the rest of the network. The ring, star, and bus network topologies have become popular for most optical fiber LANs. For example, rings are used in the Fiber Distributed Data Interface (FDDI), the IEEE 802.6 Dual Queue Dual Bus (DQDB) standard uses a dual bus, and star configurations are encountered in switched systems. The LAN logical topology may differ from the physical cable layout. The

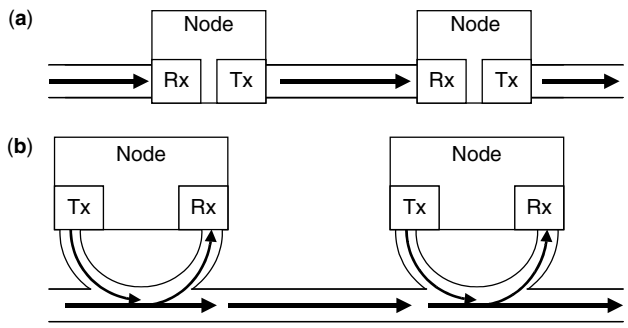


Figure 3. Types of coupling for optical links: (**a**) active and (**b**) passive coupling.

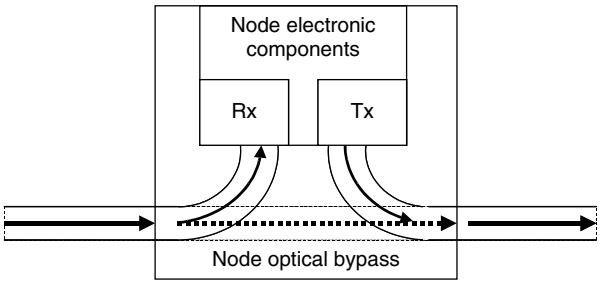


Figure 4. Optical bypass of a node.

budget, time, and resources usually dictate how the cable should be laid. For example, a ring topology can be folded to resemble two buses going in opposite directions using different fiber strands.

The technique to control the fair share of the transmission medium is defined by the medium access control (MAC) protocol and is dependant on the topology. The most common types of MAC protocols use token-passing, reservation, or collision avoidance techniques. The token-passing technique allows only the station holding the token to transmit at a time. The token is circulated through all the stations with a predefined mechanism that prevents unfair behavior and restores lost tokens. The reservation mechanism requires an arbiter that collects reservation requests from the stations and then allocates turns to transmit; this is commonly used in networks with asymmetric traffic needs. The collision mechanism allows stations to talk at any time, but two nodes trying to transmit simultaneously produce a collision. In that case, they cease transmission and wait a random time before retransmission. This method causes one of them to start before the other, reducing the probability of a new collision.

3.1. Ring Topology

Logic ring architecture, as can be seen in Fig. 5, establishes the circulation of information in a specific direction around the ring. The ring passes through all the stations in the LAN, enabling all of them to receive the same message until it completes the loop. The vulnerability of this network is that a single interruption in the ring disrupts the whole LAN. To maintain the reliability, the most common approaches are to create a second counterrotating ring for restoration, as shown in Fig. 6, or to add a bypass device for the damaged segment.

The most common mechanism to control access to the medium is the use of token-passing techniques. The ring nature simplifies the process to circulate the token among the stations. Only the station that possesses the token may transmit, and the rest remain silent. When the turn finishes, the station passes the token to the next on the ring, giving an opportunity to all nodes to transmit.

3.2. Bus Topologies

The bus topology establishes a sequence of nodes in line as exhibited in Fig. 7. The more commonly used access control mechanisms are collision detection and reservation of resources. Stations in a bus could use either active or passive coupling. The transmission into the bus

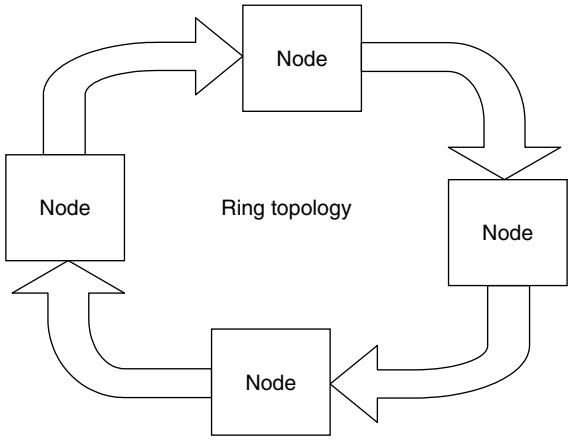


Figure 5. Ring topology.

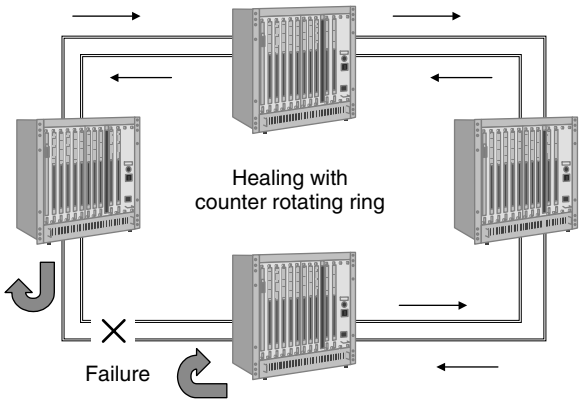


Figure 6. Self-healing ring.

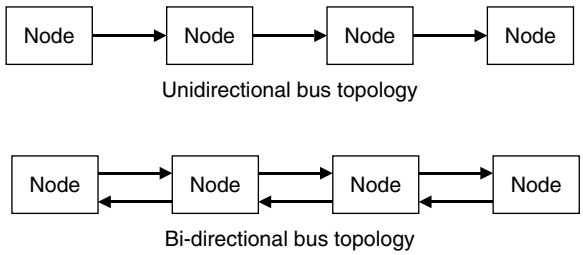


Figure 7. Bus topologies.

can be made in both directions; however, it is usually unidirectional for fiberoptic networks. For unidirectional fiber segments we need a return path, so typically there are two buses going in opposite directions. A station in a logic bus transmits information into the cable, and the remaining stations receive the transmission downward. The network will be divided into two independent segments if there is a break in the bus, unless there is an alternate path to rejoin the segments together.

3.3. Star Topologies

The star topology links several stations to a single central point. The connectivity in the network could be active

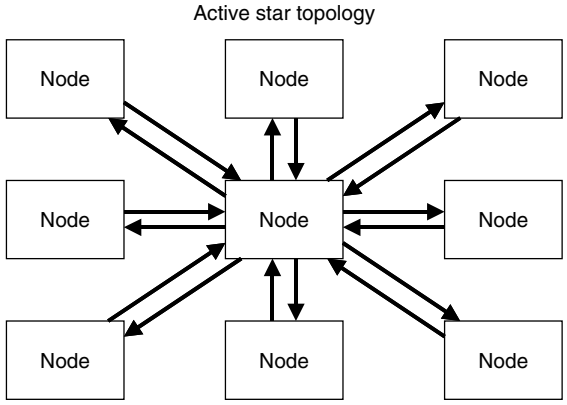


Figure 8. Active star topology.

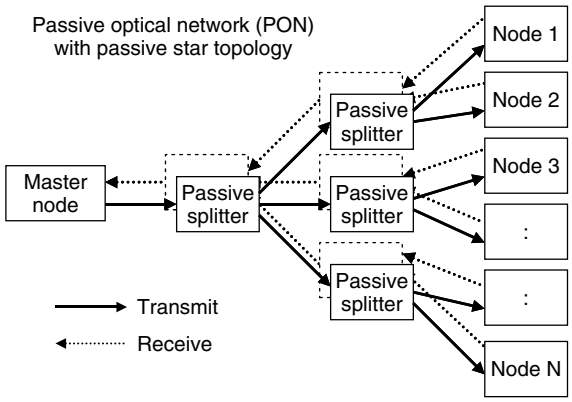


Figure 9. Passive star topology.

or passive, as illustrated in Figs. 8 and 9, respectively. Most configurations require the central station to be the controller for the rest of the network. It acts as a switch for messages between stations or performs as an arbiter for the allocation of resources. Fewer star networks distribute the MAC functionality to the nodes using token-passing or collision detection techniques. The more vulnerable point of the network is the central station because a break affects all the stations below the point of rupture.

The use of passive optical components, splitting the light among different fibers, allows several stations to receive the same broadcast from a single transmitter as shown in Fig. 10. These types of networks are commonly known as *passive optical networks* (PONS). The PON architecture normally requires a central station that controls the rest of the network. The access control is made by reservation of a resource, either a time slot or a whole optical channel.

Ethernet-type networks have collapsed the bus to a single-hub device, similar to the twisted-pair cabling option of Ethernet (known as *10BASE-T*). In addition, they have extended the individual links for each station using optical fiber. Therefore, the physical layout is a star and the links are point-to-point.

4. NETWORK DESIGN CONSIDERATIONS

Two important parameters need to be evaluated before designing a network: power loss and dynamic range. The assessment of these constraints depends on the network topology. In the bus topology, the optical signal is typically tapped by using an optical coupler at each node. In the ring topology, the coupling between the ring and the node can be either passive or active as illustrated in Fig. 3. In the star topology with N nodes, an $N \times N$ optical fiber star coupler is usually used to distribute the signal from one input to N outputs equally.

If we assume that couplers used in a bus topology couple C percent of the power from the bus to a node with a typical coupling loss of α dB, the total power coupling from the bus to a node will be $C \times 10^{-\alpha/10}$. In this case, the maximum power loss from one node to another node will be

$$L_{\text{bus}} = \frac{10^{\alpha N/10}}{(1-C)^{N-2}C^2}$$

$$L_{\text{bus,dB}} = 10(2-N)\log_{10}(1-C) - 20\log_{10} C + \alpha N \text{ dB} \quad (1)$$

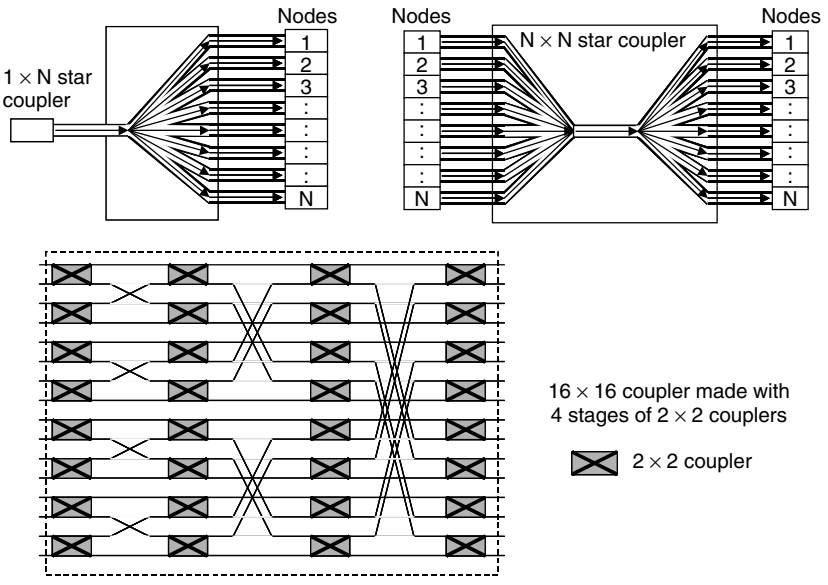


Figure 10. Optical star couplers.

where N is the total number of nodes. The maximum power loss occurs when the transmitting and the receiving nodes are at two opposite ends of the bus. As can be seen, the maximum power loss is a function of coupling ratio C and N. The optimum value of C, which minimizes the maximum power loss, is given by

$$C_{\text{opt}} = \frac{2}{N} \tag{2}$$

Substituting C_{opt} from Eq. (2) into Eq. (1) results in

$$L_{\text{bus,dB,min}} = 20\log_{10}(N-2) - 10N\log_{10}\left(\frac{N-2}{N}\right) + \alpha N - 6 \text{ dB} \tag{3}$$

The dynamic range (DR) is defined as the ratio of the maximum received power to the minimum received power by a node. In the single-bus topology, the maximum received power occurs when the signal comes from the adjacent node, and the minimum received power occurs when the receiving and transmitting nodes are farthest apart from each other. The DR for bus topology can be evaluated using the following procedure:

$$P_{\max} = P_0 C^2 10^{-2\alpha/10}$$
$$P_{\min} = P_0 C^2 (1-C)^{N-2} 10^{-N\alpha/10}$$
$$DR_{\text{bus}} = \frac{P_{\max}}{P_{\min}} = (1-C)^{2-N} 10^{(\alpha/10)(N-2)}$$
$$DR_{\text{bus,dB}} = \alpha(N-2) - 10N\log(1-C) + 20 \text{ dB}$$

where P_0 is the transmitted power.

An $N \times N$ star coupler typically consists of $\log_2 N$ stages of 2×2 couplers as shown in Fig. 10. In the star topology, the transmitted signal is equally distributed to N nodes. Thus, the power loss from a transmitter at one input to a receiver at one of the outputs can be evaluated using the following equation:

$$L_{\text{star}} = N10^{[(\alpha \log_2 N)/10]}$$
$$L_{\text{star,dB}} = (3+\alpha)\log_2 N \text{ dB} \tag{4}$$

where α is the coupling loss at each 2×2 stage. Since the transmitted power is equally divided among the receiving nodes, the DR in star topology is unity.

5. NETWORK PROTOCOLS

Most of the current LAN protocols are derived from the standards defined in the IEEE 802 series [1]. They specify the packet formats and medium access techniques for various types of transmission media. However, there are other types of networks defined by Industry associations and standards bodies. Table 1 shows the main characteristics of popular optical fiber LANs and we will describe the more relevant types below.

5.1. FDDI

In the late 1980s, The American National Standard Institute (ANSI) defined the protocol denominated Fiber Distributed Data Interface (FDDI) [2]. The accredited standards committee (ASC X3T9.5) had the objective to develop an inexpensive high-speed optical network functioning as the backbone of other slower LANs. Later,

Table 1. Characteristics of Common Optical LANs

Network Type	Fiber Type	Speed (Mbps)	Coding	Maximum Segment Distance	Topology	Access Method
FDDI	Multimode, 1300 nm	100	4B/5B NRZI	2 km (maximum 100 nodes)	Ring	Token passing
Ethernet 10BASE-T (fiber)	Multimode	10	Manchester	500 m	Active star	CSMA/CD
Ethernet 10BASE-FP	Multimode	10	Manchester	1 km (maximum 33 nodes)	Passive star	CSMA/CD
Ethernet 10BASE-FL/FB	Multimode or single-mode	10	Manchester	2 km	Active star	CSMA/CD
100BASE-FX/SX	Multimode	100	4B/5B NRZI	2 km	Active star	CSMA/CD
1000BASE-SX	Multimode	1000	8B/10B	275–550 m	Active star	CSMA/CD
1000BASE-LX	Multimode or single-mode	1000	8B/10B	550 m–5 km depending on fiber type	Active star	CSMA/CD
ATM LAN emulation	Single-mode	43, 155, 622	Any valid SONET, SDH, or DS3 line	Any valid SONET, SDH, or DS3 line	Active star	Switched central control
Fiber channel	Multimode or single-mode	100, 200, 400 & 800	8B/10B	175 m–10 km depending on fiber type	Active star	Switched central control

the International Standards Organization (ISO) published the same specifications under the ISO 9314 series.

FDDI specifies a 100-Mbps LAN using ring topology and token-passing access control. It employs multimode optical fiber and LEDs because they are more economical than laser diodes and single-mode fibers. The maximum distance between nodes is 2 km, and there is a maximum of 100 nodes per ring. The maximum frame size is 4500 bytes. The network is composed of two rings; the primary is for normal operation, and the secondary is reserved for restoration. Faults in the ring are resolved by sending the traffic through the secondary ring in the opposite direction, as shown in Fig. 6. Inactive nodes are optically bypassed, either internally or using a central hub with connection only to the primary ring.

5.2. IEEE 802.3 (Ethernet)-Type Protocols over Fiber

The Ethernet, developed by Xerox, is the basis for the IEEE 802.3 [1] standard using the CSMA/CD technique. It is the most popular protocol for LANs and has been improved to work at various speeds over many types of media. Since the first Ethernet implementations, there have been optical transceivers extending point-to-point connections [3]. However, there are newer standards to support higher speeds over several types of fiber. Most implementations use point-to-point fiber links in an active star configuration. The hubs convert the optical signals to electronic format and process them similar to other Ethernet LANs. The exception is a passive star defined in 10BASE-FP protocol. Lasers and single-mode fibers allow longer links for optical LANs; however, they are limited because they need to detect collisions. To overcome the problem, many networks employ bridges or switches in point-to-point links. Some equipment manufacturers have created proprietary solutions to reach distances of several hundreds of kilometers.

All the specifications employ the same frame format as 802.3. Only one station may transmit at a time, and if two nodes generate a collision, the hub sends a "collision presence" signal to the remaining stations. The standards 10BASE-FL and 10BASE-FB define 10-Mbps networks that may use multimode or single-mode fiber links. The difference between the standards is the retiming provided by the repeaters in 10BASE-FB. The 100-Mbps standards tried to reuse existing elements from other networks. For example, 100BASE-FX uses the same optical components similar to FDDI, and 100BASE-SX employs fiber compatible with the 10BASE-FL standard. The Gigabit Ethernet standard (1000BASE-SX/LX) specifies 1 Gbps transmission over multimode and single-mode fibers. There is an effort, by telecommunications service providers, to deploy 10-Gbps Ethernets. The standards are based on optical components similar to SONET OC-192 systems.

5.3. ATM Lan Emulation and Fiber Channel

5.3.1. ATM. The asynchronous transfer mode (ATM) protocol was designed by the telecom industry to handle all types of communication. It fragments the information into fixed-size packets, called cells, to work with real-time applications and bursty traffic. The definitions were made to support the backbone for carriers and the LANs in the corporations. ATM was designed to work principally on top of synchronous optical network (SONET) [4] links; the advantage is that the links can span very large distances using standard carrier equipment. ATM was expected to gradually replace LAN [5] protocols; however, the combination of Ethernet with TCP/IP dominated the LAN market.

ATM is a switching protocol that requires the establishment of virtual circuits between stations before a communication can be effected. The network administrators manually configure typical ATM implementations; however, LANs operate under a connectionless approach. The nodes in a LAN just send packets with enough information to reach the destination without the need to negotiate circuit establishment. ATM emulates a LAN relying in a signaling protocol that manages the connections.

The LAN emulation (LANE) standards define a mechanism in which the ATM stations behave like a LAN. The nodes employ a server that controls the establishment of the circuits and keeps track of the stations registered as members of the LAN. The individual nodes need to register their address with this server before requesting a connection to send packets. There is a second server used for broadcasting; it replicates a broadcast packet and sends it to all stations. This solution offers high-performance links between stations because there is no contention for resources in the switch. However, companies have limited its use because LAN switches are simpler and inexpensive compared to ATM switches.

5.3.2. Fiber Channel. The Fiber Channel standard was created to support communication between computers and intelligent distributed peripherals at short distances. When peripherals required higher transfer speeds, the interfaces using copper cable were more difficult to implement and the distances were reduced. The proposed solution for the problem defined a switched network with a star configuration. Each optical link is a point-to-point connection between switches and nodes. The standard was designed to encapsulate other types of protocols in point-to-point connections (ATM, SCSI, HPPI, IP, IEEE 802, etc). Fiber Channel supports multiple speeds, multimode or single mode fiber, and distances ranging from few meters up to 10 km. This protocol is normally used to communicate storage-area networks.

6. MULTIWAVELENGTH NETWORKS

WDM is a technique used to multiplex several optical channels through the same optical fiber. The operation is equivalent to combining and splitting light rays with different colors using a prism. A dense WDM system can be viewed as a parallel set of optical channels, each operating at a different wavelength, as illustrated in Fig. 11. This technology can increase the capacity of existing networks without the need for expensive additional cabling, reducing the cost of upgrades.

The bandwidth required for a channel transmission is in the order of gigahertz; however, the optical fiber

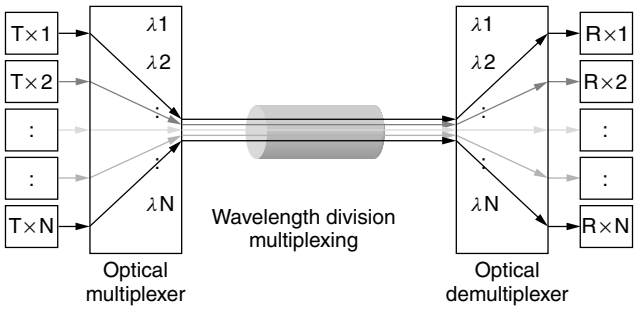


Figure 11. Wavelength-division multiplexing.

supports many terahertz in the available optical windows. WDM generally exploits the capacity of the optical fiber at 1300- and 1550-nm optical windows. The earlier WDM systems used only two channels, one at each window. The number of channels was increased later with the aid of optical filters, reaching separation between wavelengths in the range of tens of nanometers. More recently, the use of coherent techniques allowed separations of less than 1 nm. With this technology, the capacity of a single optical window expands to several dozens of channels known as *dense WDM* (DWDM).

There are several methods to allocate channels in WDM networks, where each wavelength corresponds to a channel. The major problem for these networks is to tune transmitters and receivers to different wavelengths.

Some methods propose fixed wavelengths assigned to each transmitter or each receiver. In order to establish communication, the other device will change its operating wavelength to match the desired station. This tuning process, in only one type of device, simplifies the network construction. However, it restricts the number of nodes to be the same as wavelengths available, and is subject to conflicts when two stations try to access the same destination.

In multihop [9] configurations the wavelength assignment is fixed for all the devices, and each station has several transmitters and receivers. A node can use other machines as repeaters, when it needs to talk to a station with an incompatible wavelength. This limits the number of nodes that can be reached from a particular source. However, the assignment of wavelengths is designed to create a sequence between stations forming a logical ring. This method simplifies the construction of the devices and reduces the conflict probability. On the other hand, it increases the traffic and the delay because it may require several retransmissions of the information.

Another technique enables all stations to tune the transmitters and receivers to any available wavelength. The method allows direct communication between all the stations and reduces the probability of blocking. Some disadvantages are complexity of the protocols required and more expensive devices. The nodes normally use a central control station to assign the wavelengths; however, they can employ techniques such as predefined sampling of channels, ALOHA protocol variants [6], and token passing. Those methods are not commonly used because they increase the delay and complexity of nodes.

In some networks, the signaling for call control is associated with the channel to use. This method blocks a resource just for the transmission of control information. A more efficient method for signaling employs a dedicated control channel that is independent of the information channels.

6.1. Phases for the Application of WDM Technology

The introduction of the WDM technology into optical fiber LANs, will follow an assimilation process similar to long-distance networks. At the beginning, it will serve the demand for increased bandwidth on some point-to-point links. Later, it will permeate to the backbone facilities, and finally will reach user interfaces. The evolution of DWDM technology seems likely to follow the stages shown below:

1. Increase the transmission capacity for congested point-to-point links.
2. Creation of an optical backbone that should add restoration capabilities.
3. Introduction of optical cross-connects allowing entire optical paths between end devices; however, the circuits are manually configured.
4. Usage of optical circuit switches. The stations will establish the optical circuits automatically, but will require sophisticated signaling protocols.
5. Introduction of optical packet routing. In the later stages the optical networks will be able to route individual packets without converting them to an electric form.

There is a trend, in the communications industry to create an all-optical network. The process will take several years to widely deploy commercial products. Meanwhile several experimental networks have been created to study their behavior. Some examples are Bellcore's Lambdanet [7], Lightning [8], Columbia's Telecomm Center TeraNet [9], IBM's RAINBOW [10,11]; All Optical Networking Consortium (AON) [12,13], Stanford's STARNET [14], and the European Advanced Communications Technologies and Services KEOPS [15] project.

6.2. DWDM System Components

Optical transmission systems may be divided into transmitters, receivers, amplifiers, and passive components. Each has to meet particular requirements to work in a WDM environment. The transmitters used for conventional optical fiber systems are light-emitting diodes or lasers. The link distance limits the available bandwidth, mainly due to a phenomenon called *dispersion*. It is a function of the type of fiber and transmitters used. Short-distance systems, in the order of few hundred meters, may use LEDs. For longer-distance transmission, the links require low dispersion; hence the use of lasers. All DWDM systems require a light source with a very narrow linewidth. This is accomplished with special types of lasers or with external filters.

The receiver must have good sensitivity and fast response for high-speed systems. The selection of a particular wavelength is obtained using optical filters.

There are several types of fixed filters manufactured for a specific optical channel; however, there are experiments to obtain a tunable receiver that is suitable for mass production.

For DWDM the most common optical amplifier is the erbium-doped fiber amplifier (EDFA) [16]. It amplifies the optical power of all the channels simultaneously, eliminating the need for several electronic devices. Mixing different wavelengths requires the use of optical couplers [16]. These devices combine different optical wavelengths at the inputs and distribute the mixed signal to all the outputs.

The International Telecommunications Union (ITU) proposed, through the Study Group 15 Work Project 4, a standard for wavelength assignment. It is based on a frequency reference of 193.1 THz, and a range of frequencies from 191.1 to 196.5 THz spaced 100 or 200 GHz each as shown in Table 2 (~0.8- or 1.6-nm spacing from 1568.77 to 1525.66 nm). It is ITU-T recommendation G.692 [17], "Optical interfaces for multichannel systems with optical amplifiers." This standard allows the creation of compatible devices from different vendors that operate on the same wavelengths.

7. CONCLUSION

Homes and corporations use LANs more and more each day, and the Ethernet family of protocols is the fastest-growing segment in LANs. Because of the high-bandwidth demand and wide range of applications, Ethernet is evolving into a hybrid network. The applications with lower requirements normally use conventional copper wiring; however, some of them have started to transform into wireless LANs. Users that need the higher data rates employ Gigabit Ethernet and 10-Gigabit Ethernet. This is the segment that needs more benefits from optical LANs. One lesson learned during this evolution is to reuse the existing technology, as much as possible, to make the products commercially viable. Fast Ethernet employs several elements from FDDI, and there is a trend for 10-Gigabit Ethernet to use the same optical technology as SONET OC-192 transport. Several optical LAN protocols are technically superior compared to Ethernet; however, the market has put them in disuse.

More recently the WDM technology has been utilized for the MANs. More likely in few years this technology will enter the LAN environment as the need for bandwidth increases. The major drivers for optical fiber networks are the multimedia applications and the communication between computers and peripherals in distributed computing environments.

The service providers have recognized that corporations need to link their LANs without the hassle of converting protocols. Therefore, they are currently offering LAN access ports to the customers. All the users share a high-capacity backbone segmented with virtual LANs (VLANs). This allows the corporations to reduce the costs to interconnect sites at high speeds; however, they are

Table 2. DWDM Grid Defined by Recommendation ITU-T G.692 with a Central Frequency of 193.10 THz and Separation of 100 GHz Between Channels

Number	Frequency THz	Wavelength nm	Number	Frequency THz	Wavelength nm
1	191.10	1568.77	29	193.90	1546.12
2	191.20	1567.95	30	194.00	1545.32
3	191.30	1567.13	31	194.10	1544.53
4	191.40	1566.31	32	194.20	1543.73
5	191.50	1565.50	33	194.30	1542.94
6	191.60	1564.68	34	194.40	1542.14
7	191.70	1563.86	35	194.50	1541.35
8	191.80	1563.05	36	194.60	1540.56
9	191.90	1562.23	37	194.70	1539.77
10	192.00	1561.42	38	194.80	1538.98
11	192.10	1560.61	39	194.90	1538.19
12	192.20	1559.79	40	195.00	1537.40
13	192.30	1558.98	41	195.10	1536.61
14	192.40	1558.17	42	195.20	1535.82
15	192.50	1557.36	43	195.30	1535.04
16	192.60	1556.55	44	195.40	1534.25
17	192.70	1555.75	45	195.50	1533.47
18	192.80	1554.94	46	195.60	1532.68
19	192.90	1554.13	47	195.70	1531.90
20	193.00	1553.33	48	195.80	1531.12
21	*193*.10	*1552*.52	49	195.90	1530.33
22	193.20	1551.72	50	196.00	1529.55
23	193.30	1550.92	51	196.10	1528.77
24	193.40	1550.12	52	196.20	1527.99
25	193.50	1549.31	53	196.30	1527.22
26	193.60	1548.51	54	196.40	1526.44
27	193.70	1547.72	55	196.50	1525.66
28	193.80	1546.92	56	193.80	1546.92

not exposed to the security risks involved in other public networks.

BIOGRAPHIES

Mehdi Shadaram received his B.S.E.E. degree from the University of Science and Technology in Tehran in 1976, his M.S. and Ph.D. degrees from the University of Oklahoma, both in electrical engineering, in 1980 and 1984, respectively. Currently, he is the Schellenger endowed professor and the chairman of the Department of Electrical and Computer Engineering at the University of Texas at El Paso. His research activities are focused in the field of optical fiber communications and photonic devices. During the last few years, he has investigated the performance of analog optical fiber links, WDM networks, and application of tapered single-mode optical fibers. NASA, Jet Propulsion Laboratory, National Science Foundation, Office of Naval Research, Department of Defense, Texas Instruments, Nortel Networks, and Lucent Technologies have funded his research projects. He has published more than 60 articles all in his area of research, most of them in refereed journals and conference proceedings. Dr. Shadaram is a registered professional engineer in the state of Texas. He is a senior member of IEEE, member of the International Society for Optical Engineering, Optical Society of America, and Eta Kappa Nu. He has received numerous awards for teaching and research excellence. He is cited in Marquis *Who's Who in America*.

Virgilio Gonzalez received his B.S. degree in Electrical Engineering in 1988 and M.S. degree in Industrial Engineering in 1991 from the Instituto Tecnologico y de Estudios Superiores de Monterrey (ITESM-CEM), Mexico. Later he obtained his Ph.D. degree in electrical engineering from the University of Texas at El Paso, in 1999, with the dissertation "Performance Analysis of a Fiber Optic Local Area Network Based in DWDM and ATM." From 1989 to 1993 he worked at the ITESM-CEM as telecommunications director developing the communications infrastructure for the different campus of ITESM University system. He worked from 1996 to 2001 in Alestra, the AT&T subsidiary Carrier in Mexico, as technology planning manager. His main responsibilities were the architecture design, technology development, and testing for all network functions in the carrier national network. In Mexico, he established the first Internet connection to Mexico City in 1989, set up the largest computer network in the country in 1991, and deployed the first DWDM network in 1998. Since 2001, he has been a professor at the University of Texas at El Paso. His areas of interest are high-speed optical communications, multiservice networks, and data communications protocols.

BIBLIOGRAPHY

1. IEEE Standards Assoc., *Welcome to Get IEEE 802™* (Oct. 31, 2001), Homepage (online): *http://standards.ieee.org/getieee802/* (Dec. 17, 2001).
2. W. Stallings, *Local and Metropolitan Area Networks*, 6th ed., Prentice-Hall, Upper Saddle River, NJ, 2000
3. International Engineering Consortium, *IEC Online Education — Optical Ethernet* (n.d), Homepage (online): *http://www.iec.org/online/tutorials/opt_ethernet/* (Dec. 17, 2001).
4. W Stallings, *ISDN and Broadband ISDN, with Frame Relay and ATM*, Prentice-Hall, Upper Saddle River, NJ, 1999
5. N. Kavak, Data communication in ATM networks, *IEEE Network* **9**(3): 28–37 (1995).
6. G. E. Keiser, *Local Area Networks*, McGraw-Hill, New York, 1989, pp. 205–214.
7. M. S. Goodman et al., The lambdanet multiwavelength network: Architecture, applications, and demonstrations, *IEEE J. Select. Areas Commun.* **8**(6): 995–1004 (Aug. 1990).
8. P. W. Dowd, K. Bogineni, K. A. Aly, and J. Perreault, Hierarchical scalable photonic architectures for high-performance processor interconnection, *IEEE Trans. Comput.* **42**(9): 1105–1120 (Sept. 1993).
9. R. Gidron and A. Temple, TeraNet: A multi-hop multichannel ATM lightwave network, *Conf. Records IEEE OFC 95*, 1995.
10. N. R. Dono et al., A wavelength division multiple access network for computer communication, *IEEE J. Select. Areas Commun.* **8**(6): 983–994 (Aug. 1990).
11. E. Hall et al., The Rainbow-II gigabit optical network, *IEEE J. select. Areas Commun.* **14**(5): 814–823 (June 1996).
12. S. B. Alexander et al., A precompetitive consortium on wideband all-optical networks, *J. Lightwave Technol.* **11**(5/6): 714–735 (May/June 1993).
13. Consortium (July 30, 1997), Homepage. (online): *AON All-Optical Networking*. *http://www.ll.mit.edu/aon/* (Dec. 17, 2001).
14. T. K. Chiang et al., Implementation of STARNET: A WDM computer communications network, *IEEE J. Select. Areas Commun.* **14**(5): 824–839 (June 1996).
15. M. Renaud, F. Masetti, C. Guillemot, and B. Bostica, Network and system concepts for optical packet switching, *IEEE Commun. Mag.* **35**(4): (April 1997).
16. J. Gowar, *Optical Communication Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1993.
17. ITU-T Recommendation G.692, *Optical Interfaces for Multichannel Systems with Optical Amplifiers*, Geneva: International Telecommunication Union, 1998.

OPTICAL FILTERS

LEON POLADIAN
University of Sydney
Eveleigh, Australia

1. INTRODUCTION

An optical filter introduces a wavelength- or frequency-dependent change in the amplitude or phase of an optical signal passing through it. They are important components in modern optical communications networks that exploit the ability to simultaneously transmit information on more than one wavelength along a single optical fiber or

link [wavelength-division multiplexed (WDM) networks]. These WDM networks require components that can manipulate, combine, change, and reroute information based on the wavelength of light; optical filters perform many of these functions [1–4].

Filters can operate as either selective or corrective devices, or often combining both attributes. *Selective* devices extract or separate an optical signal into separate components based on wavelength or frequency and are used in optical demultiplexing, add/drop filters, optical switching, and also to create narrow wavelength selective mirrors in various lasers or other optical cavity–based devices. *Corrective* devices are used to adjust the amplitude or phase of the optical signal to remove a distortion introduced by another component or part of the network. A common example of an amplitude-corrective filter is a gain-flattening filter designed to compensate for the strong wavelength dependence of optical amplifiers. An example of a phase-corrective filter is a dispersion compensator that is used to undo the undesirable effects of dispersion in long-distance communication links, or to remove the chirp from laser pulses.

Almost all optical filters rely on optical interference between two light waves for their filtering behavior. Optical information is carried on waves that oscillate with specific frequencies. Two different waves in the same location can combine to produce a locally higher intensity if their oscillations are in phase (constructive interference), or they can combine to produce a very low or zero intensity if their oscillations are out of phase (destructive interference). The mechanisms used to split the light into parts that can be interfered, and how the interference patterns or outputs are manipulated determine the type of filter. There are a few fundamental configurations or building blocks for optical filters. Each of these is described briefly here and in more detail later.

The Fabry–Perot interferometer (Fig. 1a) utilizes multiple traversals of the same path to produce interference. The signal is split into parts by partially reflecting mirrors at either end of a cavity and reflected back and forth: some parts of the signal traverse the path multiple times before interfering with the other parts of the signal. The fraction of the signal that emerges is determined by interference conditions that depend on the optical path length of the cavity formed by the mirrors, the angle of propagation and also on the reflectivity of the mirrors.

Bulk diffraction gratings act as filters by either reflecting or transmitting (refracting) light in a wavelength-dependent manner. The most common configuration in optical communications is reflection. A reflective diffraction grating consists of closely spaced parallel grooves on a reflective surface (Fig. 1b). The periodic structure of the surface produces an interference between the small fields reflected at each groove that enhances reflections in certain directions, and suppresses them in others. An incident wave is broken up into several orders on reflection from the grating. The directions of the diffracted waves depend on the wavelength of light. Such a device can be used to spatially separate and filter out various wavelengths.

A thin-film stack or dielectric interference filter (Fig. 1c) is a sequence of alternating layers with different refractive

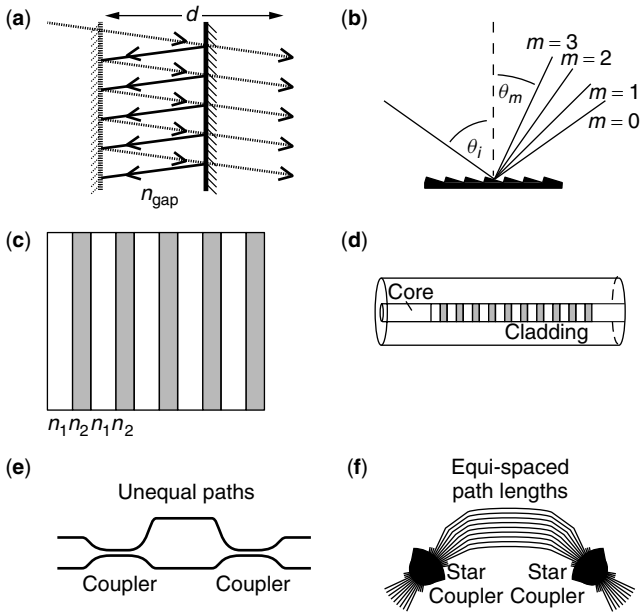


Figure 1. (**a**) A Fabry–Perot interferometer; (**b**) a planar diffraction grating; (**c**) a thin-film interference filter; (**d**) fiber Bragg grating; (**e**) a Mach–Zehnder interferometer; (**f**) an arrayed waveguide grating router.

indices (and sometimes different widths). Small amounts of energy are reflected at each interface between different refractive indices. In a similar way to the Fabry–Perot interferometer, interference between the various reflected waves results in a wavelength-dependent transmission through the stack. The major difference is that the thin-film stack utilizes multiple layers where each interface provides moderate reflectivity; the Fabry–Perot interferometer usually uses a single cavity bounded by a pair of reflecting elements with extremely high reflectivity. (Note that in a Fabry–Perot interferometer the reflecting elements either can be simple metallic mirrors or can themselves be a thin-film stack.)

Waveguide grating filters (Fig. 1d) are essentially dielectric interference filters that exist within the guiding region of an optical waveguide or fiber. Light traveling along the waveguide or fiber interacts with the layered structure and the light is reflected, coupled into different traveling modes of the guide or ejected in a wavelength-dependent manner. If the periodicity of the layers is on the order of the wavelength, the light is usually reflected at the resonant wavelength; these structures are called *Bragg gratings* or *counterpropagating gratings*. Long period gratings, however, usually couple the light to a different mode of the structure traveling in the same direction and are also called *mode-converting gratings*. If the planes of the layers are tilted with respect to the axis of the waveguide or fiber, then light will be coupled out of the guide; these gratings are called side-tap gratings.

The Mach–Zehnder interferometer (Fig. 1e) differs from the Fabry–Perot interferometer and the grating filters in that it acts as a filter by interfering two parts of the signal that have traveled along different paths. The incoming signal is equally split between the two alternate

paths by the input coupler, and on exit the signals in the two paths are recombined by a second coupler. The optical path difference between the two paths determines what fraction of each wavelength appears at each output port. When used as a demultiplexing device, it can be designed to send a particular wavelength completely to one output and another wavelength to the other output.

The arrayed waveguide grating router (Fig. 1f) is a generalization of the Mach–Zehnder interferometer to multiple arms or pathways. Incoming signals can arrive along several input ports. All of these are connected to an input star coupler. The input star coupler splits the signal equally between several pathways that are reconnected at their distant ends to another star coupler. Each pathway differs from its adjacent pathways by a fixed optical path delay. The output star coupler recombines the signals. Interference between the signals determines which output waveguide each wavelength emerges from. The structure is designed so that each wavelength from each input port is shuffled onto a different output port.

1.1. Tunable Filters and Switches

The wavelength(s) of operation of a filter depends (depend) on the optical and geometric properties of the structure. Changing any of these properties will affect the spectral properties of the filter, such as the peak value of transmission, the location of the wavelength at which maximum or minimum transmission occurs, or the extent of the band over which transmission is suppressed. This results in both undesirable effects such as the filter characteristics being temperature- or pressure-sensitive, and desirable effects in that the mechanism can be used to tune the filter.

Various controlling mechanisms (thermal, acoustic, electro-optic, nonlinear) can be used to alter one or more optical characteristics of the filter. If the filter characteristics can be changed sufficiently such that signals are completely diverted from their existing pathways to alternate pathways, the filter can be made to behave as a wavelength-dependent switch. Combining the ability to select wavelengths and to select pathways produces very powerful and useful optical components.

Reconfigurable and adaptive optical components are a current and highly active area of research, as communication networks continually move toward incorporating more flexibility, responsiveness, and intelligence into the optical layer [1,4]. It is far from clear which technologies will emerge as the leaders in this area, though it is likely to involve a hybrid of electronic, optical, micromechanical, and possibly chemical or biological technologies.

2. FILTER EXAMPLES AND APPLICATIONS

2.1. Fabry–Perot Filter

A basic Fabry–Perot filter or interferometer [5,6] consists of an optical cavity between two reflective mirrors. The incident light undergoes multiple reflections from the mirrors at either end of the cavity. The transmitted waves will constructively interfere to produce a maximum when the round-trip optical path length is an integral number of wavelengths:

$$2n_{\text{gap}}\, d \cos\theta = m\lambda \tag{1}$$

This resonance condition depends on the size of the cavity d, the refractive index in the cavity n_{gap}, and the angle of incidence θ.

The transmitted intensity for a cavity with equally reflective mirrors at both ends is given by

$$I_{\text{out}} = I_{\text{in}} \frac{1}{1 + \dfrac{4R}{(1-R)^2} \sin^2\left(2\pi n_{\text{gap}} \dfrac{d}{\lambda} \cos\theta\right)} \tag{2}$$

where R is the reflectivity of each mirror.

The transmission spectrum (Fig. 2) is periodic function of frequency. Each peak is associated with a different order m and the separation of successive peaks or orders is called the *free spectral range* (FSR) and is given by

$$\text{FSR} = \frac{c}{2n_{\text{gap}} d \cos\theta} \tag{3}$$

in frequency units. The sharpness of the transmission peaks is related to the reflectivity of the mirrors. The FWHM Δf (in frequency units) or $\Delta\lambda$ (in wavelength units) is given by the relationship

$$\frac{\Delta f}{f} = \frac{\Delta\lambda}{\lambda} = \frac{1-R}{m\pi\sqrt{R}} \tag{4}$$

The ability of a Fabry–Perot filter to resolve different signals is determined by the ratio of the FSR to the FWHM. This ratio is called the "finesse" $\mathcal{F}$ of a Fabry–Perot filter:

$$\mathcal{F} = \frac{\pi\sqrt{R}}{1-R} \tag{5}$$

Note that the expression for finesse is independent of order m. High finesse or high resolution is achieved for highly reflective mirrors. This expression is for the ideal or maximum value. In any real device the finesse will be smaller because it is limited by imperfections, such as a lossy medium and tilted or nonflat mirrors.

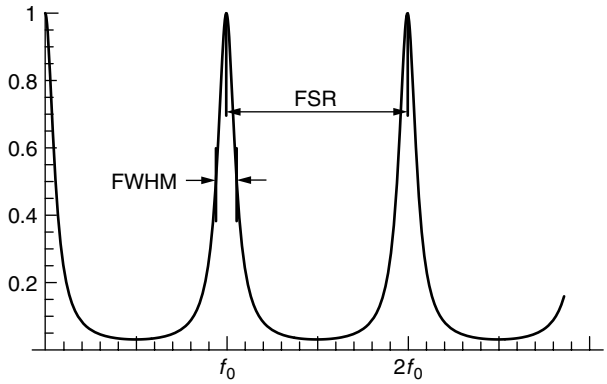


Figure 2. Transmission spectrum for a Fabry–Perot interference filter, depicting the free spectral range (FSR) and the full-width half-maximum (FWHM).

The contrast of the filter is a function of the reflectivity R:

$$\frac{T_{\max}}{T_{\min}} = \frac{(1+R)^2}{(1-R)^2} \tag{6}$$

In all the expressions above, the mirror reflectivity R has been taken to be a constant. In a real device, R is also potentially a function of wavelength and a function of angle, which complicates the analysis of the filtering characteristics. Nevertheless, the operation of the filter remains conceptually the same.

Ideally, one would like very high reflectivities suggesting the use of metallic mirrors. However, this is difficult to achieve in practice without simultaneously introducing excessive loss. Alternatively, these metallic mirrors can be replaced with dielectric stacks that have a broad reflectance peak around the spectral range of interest. The wavelength-dependent R introduced by using dielectric stacks depends on the index difference Δn of the stack. The analysis is not trivial, but as a general rule of thumb it will reduce the FWHM of the filter from the ideal value given above by a factor $\Delta n/\overline{n}$, where $\overline{n}$ is the average index in the stack.

2.1.1. Multipass and Multicavity Cascaded Fabry–Perot Filters. The finesse of a simple Fabry–Perot filter is limited by the reflectivity of the mirrors. Various improvements to the basic filter can be made by modifying the cavity structure [5,6].

A high finesse structure can be made by cascading multiple low finesse structures. There are two approaches. In the *multipass* method, the light passes twice (or multiple) times through the *same* cavity yielding a filter function which is the square (or higher power) of the original filter function. In the *multicavity* method several independent cavities are concatenated and the resulting filter function is the product of the filter functions of the individual cavities. In these configurations it is vital to keep the cavities isolated so that there is no resonating backreflection between the cavities. This can usually be done by slightly misaligning the orientations of each cavity so that spurious reflections will gradually deviate or walk away from the axis of the system.

If the cavities have identical free spectral ranges (FSRs) the cascaded system will have a transmission function with the same FSR but a much narrower FWHM, thus improving the finesse. The free spectral range of the cascaded system can be vastly increased by choosing the cavities to have different FSR (usually in the ratios of different integers). The Vernier principle can be exploited so that the FSR of the cascaded system will be determined by the coincidence of two different orders of the individual cavities (Fig. 3).

2.2. Bulk Diffraction Gratings

Bulk diffraction gratings are surfaces or plates with periodic grooves that can act as either reflective or transmissive structures [1,3]. Light incident on the grooves is split into components that are reflected in various directions depending on the angle of incidence, the wavelength, and the period of the grating. The zero-th order reflected ray is in the direction of specular reflection (equal angles of incidence and reflection); the higher order rays are referred to as diffracted rays and their direction is given below.

The basic diffraction grating equation in reflection is

$$\sin\theta_i + \sin\theta_m = m\frac{\lambda}{\Lambda} \tag{7}$$

where θ_i is the angle of incidence, θ_m is the angle of reflection of the mth-order diffracted ray, Λ is the period of the grating, and λ is the wavelength of the incident light. The fraction of light that ends up in each diffracted order is determined by the detailed shape of the diffraction grooves.

Diffraction gratings can be used in various configurations as wavelength selective elements; the most common is to combine a focusing element (either a lens or a concave mirror) with the grating. One configuration is shown in Fig. 4.

The light emerging from each waveguide hits the focusing mirror, which not only redirects the light onto the grating but also counteracts the diffractive spreading of the light as it exits the waveguide. Each wavelength

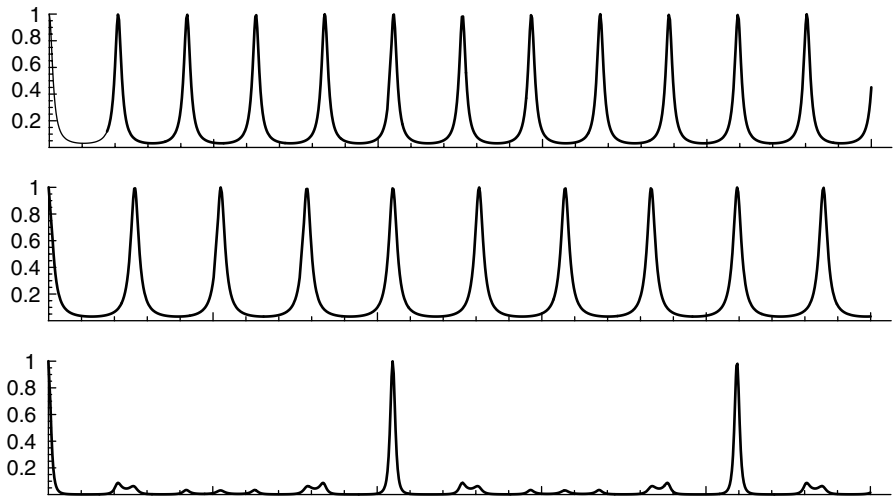


Figure 3. Transmission spectrum for a multicavity cascaded Fabry–Perot interference filter, exploiting the Vernier principle to improve the finesse.

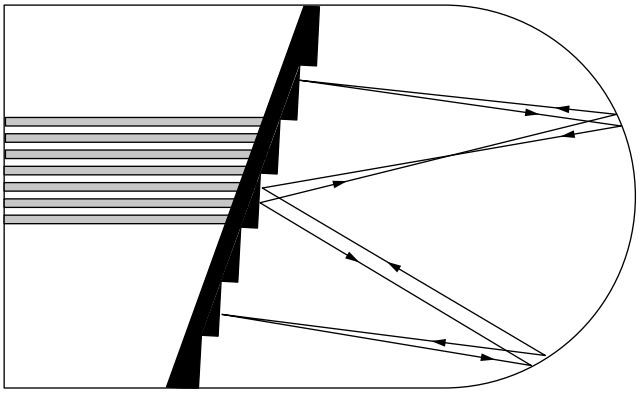


Figure 4. An example of a diffraction grating mounted onto a focusing lens or mirror.

is diffracted by a slightly different angle and therefore is reflected by the mirror to a different position, the subsequent reflections then refocus it onto a different exit waveguide.

2.3. Dielectric Thin-Film Stacks

The first historical observation of thin film interference was in 1817 by Fraunhofer, when he observed that glass with a thin layer of tarnish was less reflective than fresh glass. The simplest antireflection coatings consist of a thin layer of thickness d with an index n_1, which is intermediate between the indices n_0 and n_2 of the materials on either side. In the absence of the antireflective layer, the reflection for normal incidence is $R = r_{2,0}^2$, where

$$r_{i,j} = \frac{n_i - n_j}{n_i + n_j} \tag{8}$$

For an air–glass interface this reflectance is about 4%. In the presence of the thin layer, the reflectance becomes

$$R(\lambda) = \frac{r_{2,1}^2 + r_{1,0}^2 + 2r_{2,1}r_{1,0}\cos\left(\dfrac{4\pi n_1 d}{\lambda}\right)}{1 + r_{2,1}^2 r_{1,0}^2 + 2r_{2,1}r_{1,0}\cos\left(\dfrac{4\pi n_1 d}{\lambda}\right)} \tag{9}$$

When the individual interface reflections are low, this expression can be approximated by its numerator (which is equivalent to ignoring multiple or higher-order reflections). In either case, the minimum reflection occurs when the argument of the cosine function is an odd multiple of π. Thus, the minimum reflection occurs for wavelengths that satisfy $2n_1 d = (m + \frac{1}{2})\lambda$, where m is an integer. When $m = 0$, the thinnest possible antireflective coating is obtained with an optical thickness of $n_1 d = \lambda/4$. This is why these layers are also referred to as *quarter-waveplates*.

The value of the minimum reflection is

$$R_{\min} = \left(\frac{r_{2,1} - r_{1,0}}{1 + r_{2,1}r_{1,0}}\right)^2 \tag{10}$$

This minimum drops to zero if $n_1 = \sqrt{n_0 n_2}$ the geometric mean of the indices on either side.

A stack of thin-film layers of alternating refractive index each one quarter-wavelength thick will produce a very-high-contrast filter, with additional layers producing even greater contrast. A general thin-film filter or stack will consist of many layers of alternating refractive indices. A vast variety of filter characteristics can be designed by varying the layer thicknesses and refractive indices and incorporating more complicated patterns of alternating layers. The filters can be designed to be lowpass filters, highpass filters, and bandpass filters [7]. One general observation is that to obtain sharp cutoff filtering characteristics, the refractive index between the layers needs to be as high as possible. Some of the common materials used for visible light filters and their typical refractive indices are magnesium fluoride (1.39), zinc sulfide (2.35), cryolite (1.35), titanium dioxide (2.3), silicon dioxide (1.46), and various rare-earth oxides. In the infrared, silicon (3.5), germanium (4.0), and tellurium (5.1) can also be used in combinations with low index materials. Countless other oxides and fluorides are also used.

The wide variety of materials and deposition processes available for thin-film filters and the accuracy with which the layer thicknesses can be manufactured make thin-film filters among the most versatile and accurate of filtering devices. Thin-film filters can also be extremely narrow band and accurately manufactured to a precise wavelength; they are the ideal device to use in monitoring the wavelength drift of other devices. For example, a wavelength locker is used to monitor and control the wavelength of a laser source. The locker consists of two cascaded filters with equal bandwidths accurately located on either side of the desired operating wavelength. The optical signals transmitted through this pair of filters are compared and used to provide an electrical feedback signal to compensate for wavelength drift.

A set of thin-film filters can be cascaded to form a WDM demultiplexer (Fig. 5), with each filter either reflecting or transmitting a different specific wavelength channel. The simplest configuration utilizes each filter as a bandpass filter: transmitting a specific wavelength and rejecting all others. The filters can be conveniently deposited on both sides of a transparent dielectric slab, which also provides for accurate parallel alignment of the filters. The output wavelengths can be collimated and launched into fibers with a set of graded-index (GRIN) lenses. Note that since

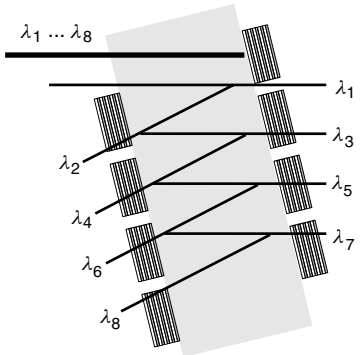


Figure 5. A cascaded set of thin-film filters used to demultiplex eight wavelengths.

all the filters are tilted with respect to the direction of the incident light, the filter layer thicknesses must be designed according to this angle. This type of cascaded filter is suitable for up to about 32 WDM channels.

2.4. Waveguide and Fiber Gratings

Gratings in waveguides and fibers are fundamentally similar to thin-film stacks. Both are structures with periodically varying optical properties [8,9]. The differences arise mainly because thin-film stacks are bulk-optic structures, whereas waveguide gratings are embedded in an existing guiding structure. The embedded gratings thus are restricted to a much smaller range of index differences since they are created by modifying the properties of a single material rather than interleaving different materials. Thus rather than use a moderate number of high contrast layers, embedded gratings use an extremely large number of low contrast layers.

The current importance of Bragg gratings is directly attributable to the photosensitive process whereby permanent index changes can be induced in glass materials using various specific wavelengths of light (mostly in the ultraviolet but also using two photon processes in the visible). An interference pattern is produced by combining two beams of light obtained from a diffraction phase mask or other source (Fig. 6). The periodic pattern of light in turn produces a periodic variation in the induced refractive index change. This is the preferred approach for writing gratings with micron and submicrometer periods. If the period of the grating is longer, then each layer of the grating can in principle be written individually. Those grating structures that can be produced by a holographic process may consist of many (thousands of) periods as opposed to manufactured thin-film stacks that have a much smaller number of layers.

Gratings can be used to selectively reflect light by coupling light from a forward traveling mode of the waveguide or fiber, to a backward-traveling mode. Such gratings are called *Bragg gratings* and have periods comparable to the wavelength of light. Gratings can also be used to selectively couple light between *different* modes traveling in the same direction. Such gratings are called *mode-converting* or *long-period gratings* and can have periods from several micrometers to centimeters.

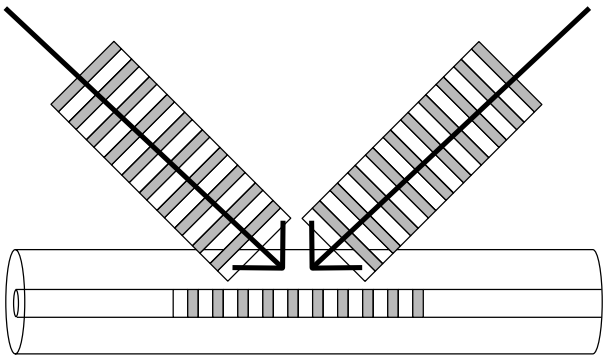


Figure 6. Two interfering beams of light used to produce a grating.

The relationship between the resonant wavelength (or Bragg wavelength) and the period of the gratings involves the effective refractive indices of the modes of the underlying waveguide. In any waveguide, each mode has its own characteristic phase velocity, which determines an effective refractive index (intermediate between the smallest and largest refractive indices occurring inside the waveguide structure). For a grating that couples between modes A and B the resonance condition is

$$\lambda_0 = (n_A \pm n_B)\Lambda \tag{11}$$

where Λ is the period of the grating, n_A and n_B are the effective indices of the modes and λ_0 is the free-space wavelength at which the coupling is most efficient. The positive sign is used for Bragg gratings (where the modes are traveling in opposite directions), and the negative sign is used for gratings that couple modes traveling in the same direction.

The actual efficiency of the grating and how it varies as the wavelength departs from the resonant value depend on other properties of the grating such as its length and the depth of modulation of the refractive index. This is explored briefly later in the section on coupled mode theory. In general, for simple gratings, Bragg gratings operate over a narrow range of wavelengths (a few nanometers) and long-period gratings operate over many tens of nanometers.

A wide variety of spectral characteristics can be obtained from gratings for a diverse range of applications [8,9]. Common applications of Bragg gratings are in optical add drop multiplexers (OADMs). Gratings can be easily designed to strongly reflect over a narrow range of wavelengths, referred to as the bandgap or reflection band of the grating. Unfortunately, most WDM applications of filters require a bandpass rather than a bandreject functionality and so various configurations have been developed to exploit reflective gratings.

Two different configurations are shown in Fig. 7. In the first configuration, the grating is located between two 3-port circulators. All incoming wavelength channels are sent by the first circulator toward the grating. All wavelength channels except one are passed by the grating and then sent by the second circulator back out to the network.

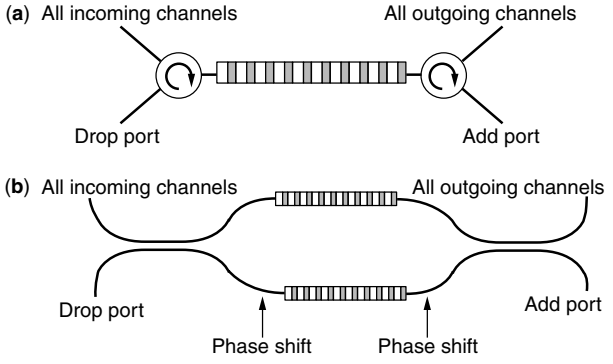


Figure 7. (**a**) Optical add-drop multiplexer formed by a grating and two circulators; (**b**) alternative configuration with two couplers.

The grating is designed to reflect the wavelengths corresponding to one of the channels: the drop channel. These wavelengths are reflected by the grating and after passing back through the first circulator are delivered to the drop port. Likewise, signals arriving at the add port are passed by the second circulator toward the grating. The add channel is reflected by the grating, and then, after passing back through the second circulator, these wavelengths join the other channels passing out to the network.

The high cost of circulators has stimulated alternative designs. The second configuration requires a pair of *identical* gratings located between two couplers. The drop channel wavelength is reflected by both gratings simultaneously, but picks up an extra π phase shift in one arm. As it passes back through the coupler, it is recombined onto the other port of the coupler, because of the phase shift. The analogous thing happens on the other side with the drop channel.

Another important application of gratings is in gain flattening. The output of an erbium-doped fiber amplifier (EDFA) varies strongly as a function of wavelength over its band of operation (Fig. 8). When a signal passes through many such amplifiers, the nonuniformity is accentuated, and this severely limits the usable bandwidth of the system. Gratings having a filter profile that is the opposite of the gain spectrum of the EDFA can be used to flatten the profile and thus extend the usable bandwidth. Various types of long-period gratings have been successfully used to modify the profile over a range of tens of nanometers.

The final important application of gratings considered here is the dispersion compensator. Signals traveling over long distances in standard telecommunications fiber near 1.55 μm experience dispersion-induced broadening since the longer wavelengths travel slightly more slowly than shorter wavelengths. This group velocity dispersion will lead to signal degradation. A chirped Bragg grating is one where the period of the grating varies along its length. If the grating period is longer at the front than at the back, the shorter wavelengths will travel further into the grating before being reflected. This introduces a wavelength-dependent delay on reflection that can be used to compensate for the delay induced by dispersion (see Fig. 9). The dispersive properties of gratings are a very important area of investigation [10].

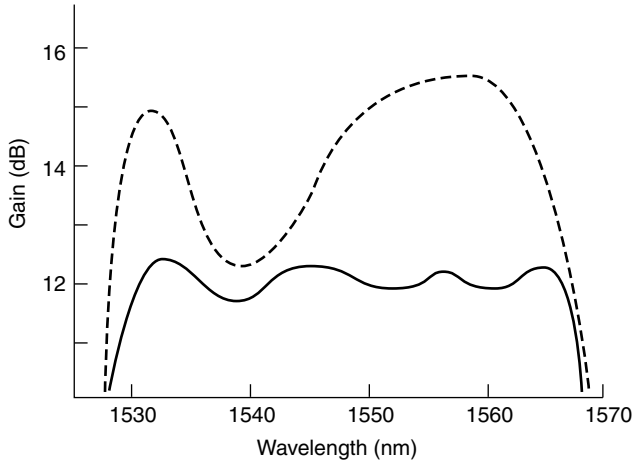


Figure 8. Output spectrum of a typical EDFA before (dashed) and after (solid) passing through a gain-flattening filter.

Figure 9. Dispersion-compensating Bragg grating.

A major problem associated with gratings is temperature dependence. Not only does the grating physically expand with temperature causing the period to shift, but a more important effect is the temperature-dependent refractive index of most glasses, which also results in a shift in the resonant or Bragg wavelength. Temperature compensation in grating packaging is a critical component of grating technology.

2.4.1. Coupled-Mode Theory. Coupled-mode theory is used extensively for the modeling and analysis of gratings. The coupling strength of the grating is encapsulated in a parameter κ, which is proportional to the index modulation. The frequency of the incoming signal is described by a detuning that is proportional to the frequency difference between the signal and the resonant or Bragg frequency.

More precisely, the coupling strength is defined by

$$\kappa = \eta \frac{\pi}{\lambda} \Delta n \tag{12}$$

where Δn is the index modulation and η is an efficiency factor describing how well the grating overlaps with the transverse intensity profile (or modal profile) of the light traveling along the waveguide; the detuning is defined by

$$\delta = (n_A \pm n_B)\frac{\pi}{\lambda} - \frac{\pi}{\Lambda} \tag{13}$$

where, as before, the positive sign is used when the modes are counterpropagating and the negative sign when they are copropagating.

A general rule of thumb for all gratings is that coupling is very efficient if the detuning δ is smaller than the grating strength κ and is very inefficient when the detuning is larger than κ. For Bragg gratings the detuning range between $\pm\kappa$ is also referred to as a *bandgap*.

The coupled-mode equations are the fundamental system of equations used to describe both Bragg (copropagating) gratings and long-period (counterpropagating) gratings [8,9]. The first mode (usually forward-traveling) has an amplitude that varies as $u(z)$; the second mode (either forward- or backward-traveling) has an amplitude varying as $v(z)$. The coupling of energy by the grating is represented by a pair of simple differential equations connecting these two amplitudes. For uniform gratings the equations are

$$iu'(z) + \delta u(z) + \kappa v(z) = 0 \tag{14}$$

$$-iv'(z) + \delta v(z) \pm \kappa u(z) = 0 \tag{15}$$

where the positive sign is used with counterpropagating modes and the negative sign is used with copropagating modes.

A typical spectrum for a uniform Bragg grating is shown in Fig. 10. The peak reflectivity is given by

$$R = \tan h^2(\kappa L) \tag{16}$$

where L is the length of the grating. The longer the grating, the stronger the reflection. The bandwidth of the spectrum is directly proportional to κ and can be represented in

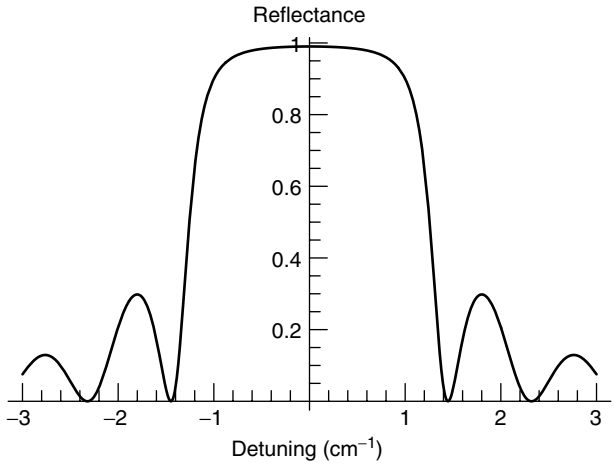


Figure 10. Reflection spectrum for a Bragg (counterpropagating) grating with a total grating strength $\kappa L = 3$.

various ways. As a rough rule

$$\frac{\Delta\lambda}{\lambda} \sim \frac{\Delta n}{n} \tag{17}$$

Very short gratings will have an apparent bandwidth wider than this, but as the grating becomes longer, the spectrum will stabilize to this fixed bandwidth. Making the grating extremely long will *not* narrow the reflection band any further.

For copropagating (or long period) gratings, it is still true that the most efficient energy coupling occurs at zero detuning. However, since both modes continue to travel in the same direction, if the grating is long enough, the coupling process will start to couple energy back into the original mode again. This phenomenon is called *overcoupling*. Figure 11 shows three typical spectra for a copropagating grating demonstrating ideally coupled and overcoupled gratings. The ideal condition for 100% coupling is first achieved for $\kappa L = \pi/2$.

2.5. Acoustooptic Filters

Acoustooptic filters exploit the interaction of light and sound in materials that are photoelastic [1,4]. The periodic compressions and expansions in the material produced by the presence of the sound wave result in corresponding variations in the refractive index via the photoelastic coefficient. A commonly used material with a high photoelastic coefficient is lead molybdate, $PbMoO_4$.

The period Λ of the index modulation produced by the photoelastic effect is equal to the wavelength of the

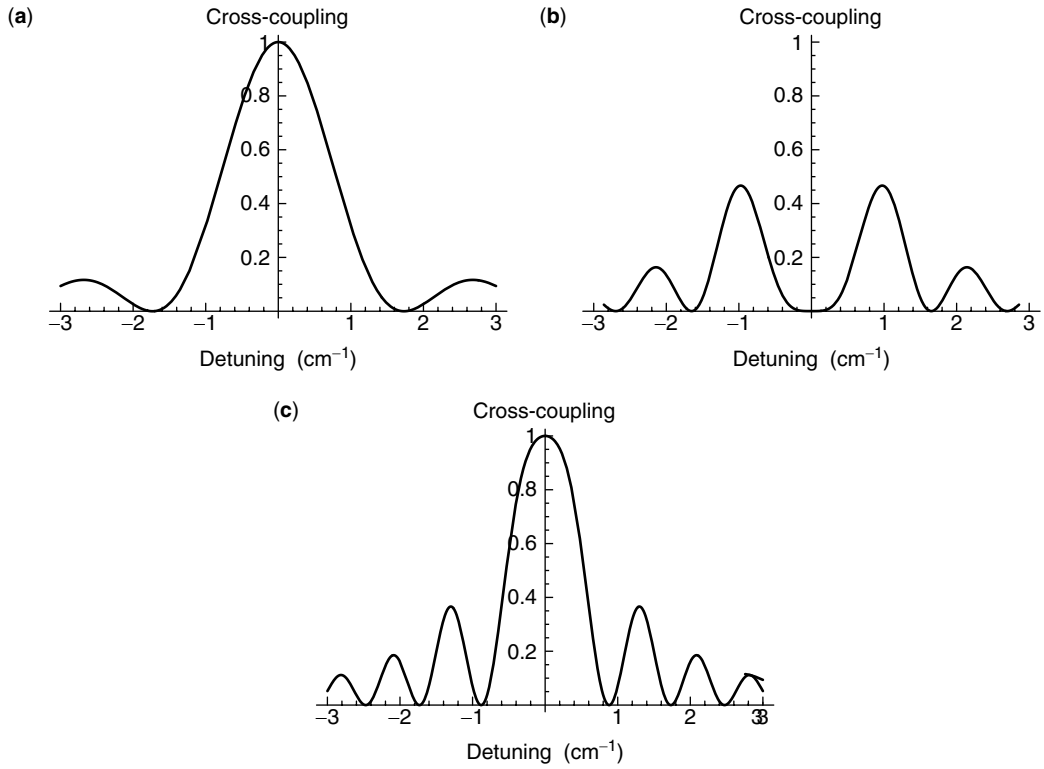


Figure 11. Cross-coupling spectra for copropagating gratings with three different values of total grating strength: (**a**) $\kappa L = \pi/2$; (**b**) $\kappa L = \pi$; (**c**) $\kappa L = 3\pi/2$.

sound wave in the material. Sound velocities in typical materials are on the order of several thousand meters per second. Thus acoustic frequencies of around 100 MHz will produce gratings with periods of several tens of micrometers. A common configuration is to use such gratings to couple energy between lightwaves traveling in the same direction but experiencing different refractive indices. The different refractive indices are obtained by using a birefringent material and exploiting the different refractive indices (ordinary n_o and extraordinary n_e) for the two polarizations. The wavelength at which maximum coupling of energy from one polarization to the other occurs is given by the condition

$$\lambda = (n_o - n_e)\Lambda \tag{18}$$

The bandwidth over which effective coupling occurs is roughly given by the relationship

$$\frac{\Delta\lambda}{\lambda} \sim \frac{\Lambda}{L_{\text{int}}} \tag{19}$$

where the interaction length L_{int} is the distance over which the acoustic and optical waves overlap and interact.

Acoustic gratings share many properties with waveguide gratings and thin-film stacks and can be analyzed by the same techniques. However, the most important difference is that the acoustic grating is a transient phenomenon, and can be easily controlled or modified by changing the properties of the acoustic wave, thus leading to various tunable filter configurations.

2.6. Mach–Zehnder Interferometer

The basic principle behind the Mach–Zehnder (MZ) interferometer is the interference of two parts of a signal that have traversed different optical paths [11]. In one of the simplest configurations (Fig. 1e) two waveguide arms of different optical lengths are connected by two 3-dB couplers. For simplicity, any wavelength-dependent properties of the couplers themselves are ignored (at least over the wavelength interval of interest). The light is split equally at the first coupler and recombined at the second coupler. The interference is produced by the phase difference between the waves traversing the two arms of the interferometer. The phase difference is given by

$$\Delta\phi = \frac{2\pi}{c} f n_{\text{eff}} \Delta L = \frac{2\pi n_{\text{eff}}}{\lambda} \Delta L \tag{20}$$

The intensities obtained from the two ports, respectively, are given by

$$I_1 = I_{\text{in}} \cos^2 \frac{\Delta\phi}{2} \tag{21a}$$

$$I_2 = I_{\text{in}} \sin^2 \frac{\Delta\phi}{2} \tag{21b}$$

Thus, interference between the two arms leads to a difference in power between the outputs of the second coupler.

The filtering characteristics are periodic in frequency and the channel spacing (also called *free-spectral range*) of this device is

$$\Delta f = \frac{c}{2n_{\text{eff}} \Delta L} \tag{22}$$

In more realistic devices, the perfect periodicity of the transmission spectrum will be modulated by the wavelength-dependent properties of the 3-dB couplers (which in turn depend on how the couplers are made).

If we contrive to have constructive interference for a specified frequency while having destructive interference for a second specified frequency, this will determine the length. For a 1.3/1.55-μm channel splitter, the length required is $L = 2.78$ μm (assuming a silica waveguide with $n_{\text{eff}} = 1.45$). The filtering characteristics are shown in Fig. 12. On the other hand, to create a device that separates 100-GHz channels (interleaving them) would require a length of $L = 1$ mm.

The filter can also be used in reverse as a multiplexer for combining wavelengths. For example, to combine a 980-nm-pump wavelength with a 1550-nm signal, the appropriate length would be $L = 0.92$ μm (for a pump at 1.48 μm the length becomes $L = 11.3$ μm). However, it is more common for this operation to be done by exploiting the wavelength-dependent coupling of a single simple directional coupler. Such a wavelength-dependent coupler can also be used to separate two wavelengths.

A cascaded series of MZ filters (sometimes called an MZ "chain") can be used to systematically separate or demultiplex a full-WDM channel set. Consider a set of equally spaced frequencies. The first MZ filter is designed to separate all the even channels from all the odd channels. Thus each output arm of the first device now carries a set of equally spaced frequencies with a spacing twice that of the original signal set. A second MZ filter that has a pathlength difference ΔL half that of the original will subsequently filter out every other channel in this reduced set and so on. Thus, for example, a cascaded chain of MZ filters 7 deep could demultiplex a 128-channel system.

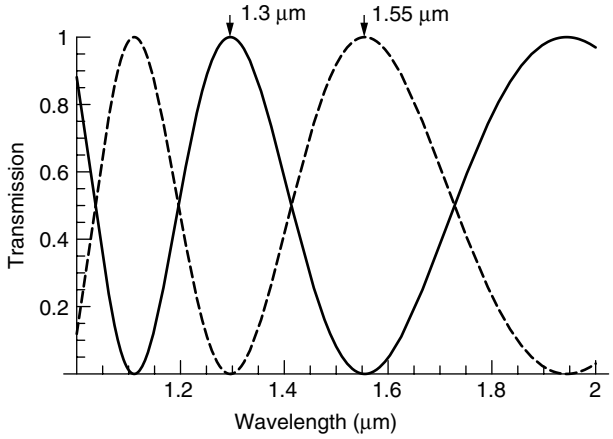


Figure 12. Filtering characteristics of the two output ports as functions of wavelength. If shown versus frequency, the characteristics would appear periodic.

2.7. Arrayed Waveguide Grating Router

The arrayed waveguide grating router (AWGR) can be regarded as a generalization of the Mach–Zehnder filter to multiple ports [3,4]. The input and output couplers are replaced by star couplers that have the property that the light entering at any port is split equally between all output ports. (Information about which specific input port the light came from is retained or encoded into the relative phases of the split signals.) If the multiple paths connecting the two star couplers were all of identical optical lengths (the relative phases would be preserved), then the second star coupler would just undo or reverse the action of the first star coupler; light entering a specific port on the left emerges from the corresponding port on the right (for all wavelengths). When the optical pathlengths of the arms are different, the signals arriving at the second star coupler will have different phases and interference will direct the output to a different port (which port will also depend on wavelength).

A useful analogy to understand the AWGR is to replace the star couplers with lenses and the array of arms with a triangular dispersive prism as in Fig. 13. The different input ports correspond to different point sources A, B, and so on. in the focal plane of the input lens. Light from each of these point sources after passing through the lens is transformed into *plane waves* traveling at different angles, each angle corresponding to a different input source. In the absence of any intervening structures, plane waves hitting the output lens are focused onto its focal plane, with the position of the image determined by the incident angle. Thus the arrangement of the output images corresponds precisely to the input sources (apart from a simple overall inversion).

The triangular prism refracts each incoming wave changing its direction, thus altering the location at which the output image is formed. Furthermore, because the amount of refraction will vary for different wavelengths, each wavelength will form an output image in a different location.

Using the notation $\lambda_i^{(j)}$ to indicate a channel with wavelength λ_i entering through port j, where $i = 0, 1, \ldots, N-1$ and $j = 0, 1, \ldots, N-1$, then such a signal emerges from port $j + i$(moduloN), (i.e., if $i + j$ is larger than $N - 1$, the port numbers wrap around.) The 4×4 example is shown explicitly in Fig. 14. The wraparound effect is obtained by ensuring that the frequency spacing between adjacent channels Δf and the optical pathlength difference $n\Delta L$ between adjacent arms satisfy $N \times \Delta f \times n\Delta L = 2\pi c$.

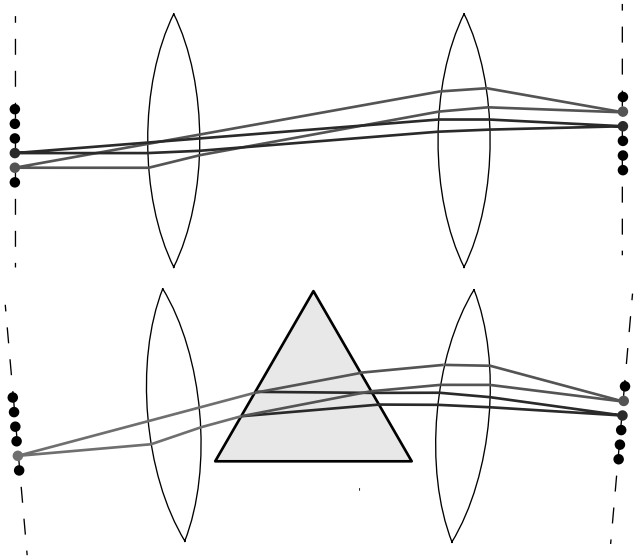


Figure 13. Analogy of an AWGR using two lenses and a triangular prism.

3. FILTER CHARACTERISTICS

Several important parameters characterize the spectrum of a filter, especially WDM filters designed for channel selection. The parameters are shown for a typical spectral profile in Fig. 15.

The peak wavelength is as the name suggests: the wavelength with the least loss. For an asymmetric spectrum, this will be different from the *center wavelength*, which is defined as the average of the upper and lower cut wavelengths or limits of the passband. Flexibility exists in the definition of these limiting wavelengths, as they are the wavelengths on either side of the peak for which the spectrum first falls to some predetermined level in decibels. For example, the 0.5- and 3-dB bandwidths are shown in Fig. 15. If the spectrum is significantly asymmetric, the bandwidth and centre wavelength will depend on the choice of level.

The isolation is given by the largest transmission levels (lowest loss) in the adjacent WDM channels. It is not necessary that this extreme value occur at the edge of the adjacent channels, nor that it even occur for the immediately adjacent channel. The figure of merit for

Figure 14. Input port and wavelength redistribution for a 4×4 AWGR.

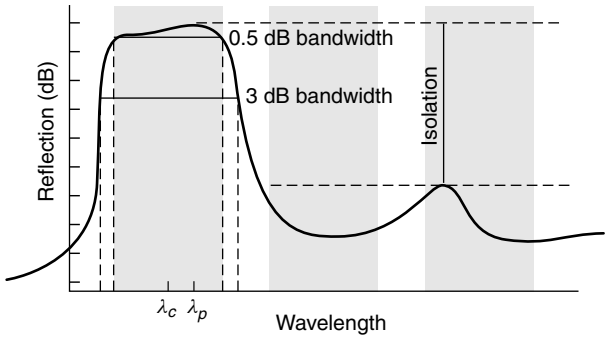


Figure 15. Typical WDM filter profile showing both the peak and central wavelengths, the bandwidth at various decibel levels, and the channel isolation or crosstalk.

a WDM filter can be defined by the ratio of utilizable bandwidth to the channel separation.

4. EVALUATION

Each filter technology or configuration has its own specific merits and disadvantages; the various designs are all competing for use in commercial optical network systems. Some of the important criteria are precision, temperature stability, low loss, low polarization dependence, manufacturing cost, packaging and pigtailing (i.e., connecting fibers to chip-based devices), scalability to large numbers of wavelength channels, isolation and crosstalk between channels, and design flexibility.

In terms of wavelength precision and stability, the thin-film filters are extremely attractive components for optical networks. The deposition process yields very accurate layers, and the variety of materials available make it easy to design for low loss and minimal temperature dependence [4] (as low as 5×10^{-1} nm/°C). The polarization dependence for these devices can also be kept low. However, scaling to larger numbers of channels usually requires the cascading of devices, which could impact both cost and overall loss.

An advantage of diffraction gratings is that the insertion loss is independent of the number of channels, making these devices more attractive for high-channel-count systems. Unfortunately, one problem with diffraction gratings is their polarization-dependence. These devices also do not fare too well on the cost criterion; the diffraction grating/concave reflector configuration is a bulk device that is not easy to fabricate.

Devices that have the best polarization-independent performance are the fiber Bragg grating filters. Passive temperature stabilization is possible for fiber gratings, giving stability similar to that of thin-film filters [4]. The holographic fabrication technique also provides this class of devices with an unrivaled ability to make unusual and customized filter responses by simply varying the chirp and apodization profiles of the grating. This technology does not excel in scaling with the number of wavelengths. It is possible to overlay a limited number of gratings with different center wavelengths, or to concatenate several gratings within the same length of fiber; however, most multiplexing schemes still require some form of cascading, which introduces the additional overheads of couplers, isolators, and circulators.

Certainly, arrayed waveguide grating routers appear to have the advantage of scalability, at least to moderate numbers of channels (from 16 to 64). However, because this is a planar or chip-based technology, polarization dependent loss may be a problem. A number of strategies for reducing the polarization dependence are available [3]; the simplest is to introduce a polarization-retarding plate half-way across the array waveguide region. The biggest disadvantage of the AWGR is the packaging and pigtailing cost of connecting fibers to the device.

The demand for greater bandwidth will continue to drive the development of components for WDM networks. There is a growing shift toward more intelligence within the optical layer of such networks, suggesting a trend toward devices with multiple functions. For example, early WDM filters concentrated mostly on filter shape, whereas later designs will also look closely at the implications of dispersive effects. Filters will be both selective and corrective at the same time, as well as combining other attributes such as switching and tunability.

The need to reduce packaging costs will also favor devices that can easily be integrated, and the incorporation of light sources, modulators, filters, and detectors onto the same substrate may be one possibility to achieve a scaling in production.

Although it has been possible to identify which attributes will become more important in future optical filter designs, it is still not possible to tell which of the existing technologies will best evolve to fit the needs of the next generation of optical networks. Almost certainly a diversity of designs and technologies will continue to exist, as no single solution will be able to meet all the needs of optical networking.

BIOGRAPHY

Leon Poladian received his B.Sc. degree in 1986 and a Ph.D. degree in theoretical physics in 1990 from the University of Sydney, Australia. His thesis was on the optical properties of periodic structures. He joined the Optical Sciences Centre at the Australian National University in 1990 as a postdoctoral fellow working on nonlinear fibre couplers and spatial solitons. Since 1992, he has been at the Optical Fibre Technology Centre at the University of Sydney, first as a Queen Elizabeth II fellow, then an Australian Research Council senior research fellow and currently as an Australian professorial fellow. Dr. Poladian has published over 80 journal and conference papers and holds five patents in the areas of fiber design and grating fabrication. His areas of interest are computational algorithms for novel fiber design; grating design, fabrication and characterization, and the optical properties of periodic and almost-periodic photonic structures in one, two, and three dimensions. Dr. Poladian also holds a graduate diploma in education from the University of New England, Armidale, Australia.

BIBLIOGRAPHY

1. J. Paul and E. Green, *Fiber Optic Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
2. K. Nosu, *Optical FDM Network Technologies*, Artech House, Boston, 1997.
3. H. J. R. Dutton, *Understanding Optical Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1998.
4. R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Perspective*, Morgan Kaufman, San Francisco, 1998.
5. W. H. Steel, *Interferometry*, Cambridge Univ. Press, Cambridge, UK, 1983.
6. J. M. Vaughan, *The Fabry-Perot Interferometer*, Adam Hilger, Bristol, UK, 1989.
7. H. A. Macleod, *Thin-Film Optical Filters*, Adam Hilger, London, 1969.

8. T. Erdogan, Fiber grating spectra, *J. Lightwave Technol.* **15**: 1277–1294 (1997).

9. R. Kashyap, *Fiber Bragg Gratings*, Academic Press, San Diego, 1999.

10. G. Lenz, B. J. Eggleton, C. R. Giles, C. K. Madsen, and R. E. Slusher, Dispersive properties of optical filters for WDM systems, *IEEE J. Quant. Electron.* **34**: 1390–1402 (Aug. 1998).

11. P. Hariharan, *Optical Interferometry*, Academic Press, Sydney, 1985.

OPTICAL MEMORIES

MICHAEL RUANE
Boston University
Boston, Massachusetts

1. INTRODUCTION

An optical memory is any system that stores and retrieves digital information using optical methods. Optical memories are mass storage devices, competing directly with magnetic hard drives and magnetic tape. Their high capacity, low cost per megabyte, reliability, and removability have made optical memories the preferred mass storage solution for many applications. They are standard components in personal computers and workstations, and support important consumer electronics. Digital versatile disc (DVD) players are now making significant inroads into the VHS tape market, and rewritable DVD (DVD-RW) is a candidate for the local storage of movies and multimedia distributed over the Internet. Optical memory is a natural component of an all-optical system for data retrieval, transmission, and storage, and will play an important role in advanced communications systems.

As of 2001, global production of content is expected to require 1–2 exabytes or roughly 1.5 billion gigabytes of storage. This is approximately 250 MB (megabytes) per person for every person on earth. This content in print, film, magnetic, and optical forms exceeds the production of content for all history before this year [1]! Consumer high bandwidth applications drive the need for inexpensive, removable memory, while high-performance military, industrial, and enterprise processing systems are investigating advanced optoelectronic devices and optical interconnections that will naturally interface with optical memories. Optical jukeboxes, for example, support enterprise-level storage-area networks (SAN), storing multiple terabytes in one system.

Sony and Phillips first developed optical media and players for the distribution of digital audio in the 1970s, building on laserdisc technology. The compact disc–digital audio (CD-DA or simply CD) was standardized in 1980 to provide a high-fidelity alternative to the conventional LP (long-playing) vinyl record. The computer industry quickly recognized that the CD, configured with stronger error correction as a data read-only memory (CD-ROM), allowed inexpensive distribution of large volumes of data [one CD-ROM holds about 450 HD (high-density) floppy disks]. In 1988 write-once optical memories CD-recordable (CD-R) allowed users to create small numbers of their own discs for storage, testing, or distribution. Fully rewritable systems [CD-rewritable (CD-RW)] followed in 1996, enabling removable optical RAM. Early CD-RW systems were much slower than comparable magnetic devices, but these problems have been largely overcome through higher rotation speeds and the availability of data compression.

DVD technology arrived in the late 1990s. DVD increased optical memory capacity and data transfer speed, making read-only optical memories suitable for full motion video and large data sets, such as high-resolution images. The 10-millionth DVD videoplayer was sold $3\frac{1}{2}$ years after introduction; it took 7 years to ship the 10-millionth CD audio player. Newer PCs increasingly have DVD drives, and DVD-R and DVD-RW units now entering the marketplace are providing high-capacity write-once and rewritable optical memories. Still in the laboratory are optical systems that further extend serial, disc-based optical memories, and new page-oriented optical systems based on holography. Such advanced systems seek capacities above 125 GB/disc and data transfer rates of 25 MBps (megabytes per second), about 25 times that of DVD. Figure 1 summarizes the growth of optical memory capacity.

Despite this success, optical memories face vigorous, application-dependent competition. Storage-area networks both compete with local removable optical memories and create demand for increased server memory. Magnetic tape and hard-disc systems promise significant improvements in capacity, data transfer rate, and access times [2]. Innovative data compression will also change the competitive memory scene, allowing tradeoffs among processor speeds, data transfer rate, and memory.

Optical memories are a specialized communications link (Fig. 2). The "transmitter" of an optical memory encodes, modulates, and writes digital data to the media "channel." Every media channel will introduce characteristic distortion, errors, and noise whose characteristics are determined by the interaction of the optical writing and reading systems with the storage medium. The "receiver" actively probes the channel to create a read data signal, which is detected and decoded to recreate the user's stored data. An important distinction is that stored data can remain "in the channel" indefinitely through the physical modification of the media during writing. The read system recovers stored data when a read beam interacts with the modified media, and experiences modulation of phase, amplitude, polarization, or frequency. The channel media can age, suffer abuse, or simply deteriorate from many cycles of the storage process. Channel characteristics also depend on the performance of control systems for maintaining rotation speed, tracking and focusing.

2. DISC-BASED SERIAL MEMORIES

This section discusses the general characteristics of available optical memories [3]. Commercial optical memories store data as an encoded serial binary bitstream on a

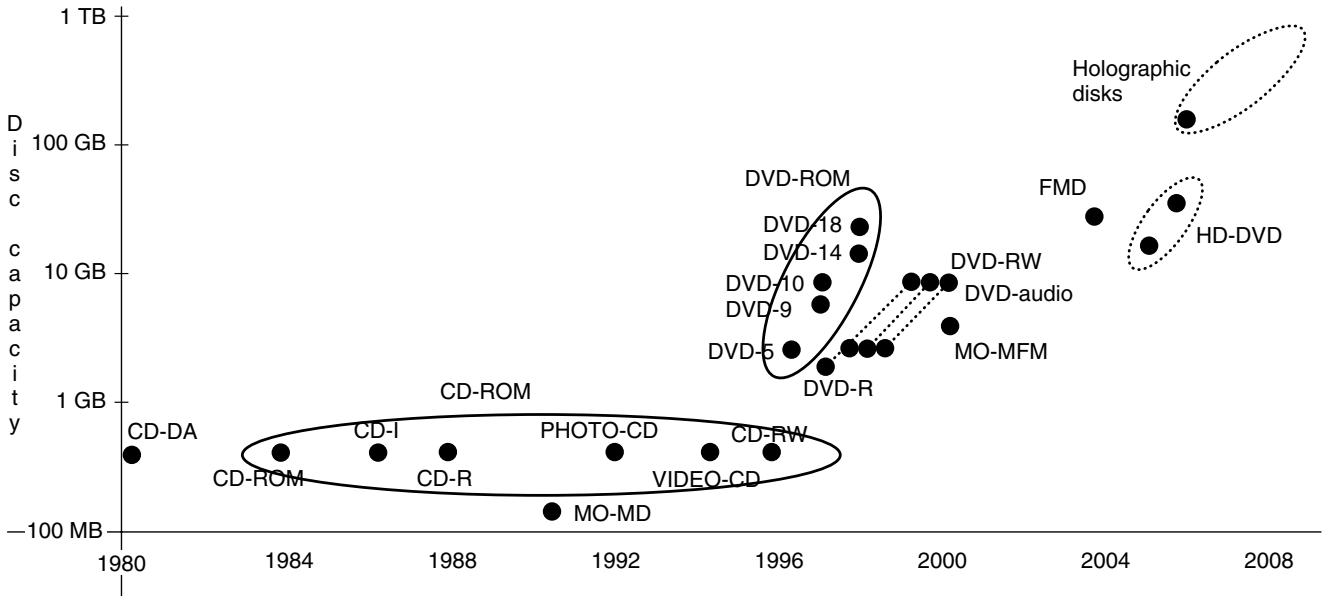


Figure 1. Evolution of optical memory capacity. Media and head constraints dictated initial CD-ROM capacity while different application formats evolved. DVD laser improvements enabled most of the 1990s increase in single layer systems, while disc capacities grew with multilayers. Continued growth will come from blue lasers in high-density DVD, fluorescent multilayers (FMD), and holographic discs. Other approaches are less established.

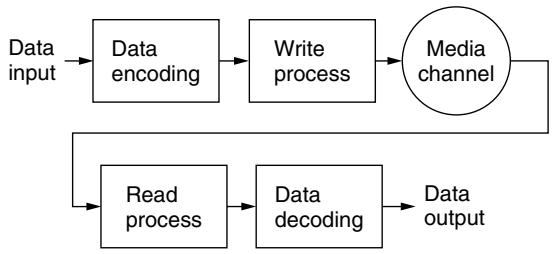


Figure 2. Block diagram of an optical memory. Optical memories are a form of communications link where the data reside in the media channel. User data undergo error coding and is organized into blocks and directories to produce a datastream that drives the write process. The "media channel" stores the data as an imperfect modulation of some property of the media. Readout recovers the media data, but with noise, bursty data loss, crosstalk, and other problems. Data decoding removes overhead, performs error checking, and regenerates the original data.

disc with one or more storage layers, accessible from one or both sides. Disc-based memories are standardized to a 120-mm-diameter disc, although minidiscs (80 mm), card discs, and other formats are specified under published standards.

CD-ROM and DVD-ROM media are mastered and injection molded. First a laser records data as exposed marks in photoresist spun on a polished glass substrate. When developed, the photoresist has holes where the data were written. Next, a conductive conformal nickel layer is deposited and thickened by electroforming to become the "father" stamper, from which "mother" and "daughter" stampers are produced. A daughter stamper is mounted in an injection molding system which is filled with molten polycarbonate under high pressure to form the final CD or DVD substrate. Injection molding takes about 10 s per disc; about 3000 CDs can be produced from one stamper. A reflective metallization layer, usually aluminum, is sputtered onto the molded substrate. Finally, a protective lacquer coating and screen printing or labels are applied.

Recordable and rewritable discs are more complex, with a mastered spiral tracking groove that usually has header marks, focusing marks, and even laser power calibration areas. Grooves sometimes have a deliberate mechanical timing wobble in their walls. An active layer stores the data by absorbing energy from the write beam and changing in some manner [4]. A reflective layer and possibly a complex optical stack enhance the active layer.

CD-ROM standards proliferated as optical memories attracted new applications that required better error control, storage of mixed media, management of multiple sessions, and extended capacities. All CD-ROMs have physical leadin and leadout areas to identify the start and end of data, and use physical data sectors of 2352 bytes, read at 75 sectors per second in a single-speed ($1\times$) drive. Different CD-ROM standards distinguish how the 2352 bytes are allocated to data, synchronization, headers, error detection, and error correction. Some standards require that entire discs must be written at one time, while others allow multiple sessions. Generally, data must be organized to meet all disc and data sector formatting requirements, and recorded without gaps (buffer underflow).

Data must be error encoded and framed to be reliable. Disc mastering has unavoidable defect rates of 10^{-4} or 10^{-5} while surface contamination often destroys or

obscures many adjacent marks. To combat these burst errors, interleaving and Reed–Solomon encoding are used. In CD audio, for example, stereo 16-bit samples are taken at 44.1 kHz, making four 8-bit symbols. A shortened Reed–Solomon (28,24) code operates on a frame of six stereo samples, or 24 8-bit symbols. The 28 output symbols are interleaved and encoded by a (32,28) Reed–Solomon code. Those 32 symbols are regrouped in even–odd groupings, extended by 8 bits for control and display, and then modulation encoded as "eight-to-fourteen modulation" (EFM), with 3 merge bits. This run-length encoding makes optical pickup and timing more reliable. An additional 27 synchronization and merging bits are then added, such that the initial frame of 192 user bits is expanded to 588 encoded bits; 32 such frames form one physical sector.

Encoded and run-length modulated data drive the laser that exposes the photoresist or modifies the active layer in a recordable medium. When read, the medium modulates the amplitude, polarization, or phase of the read beam. Frequency modulation, while possible, is not yet competitive. CD-ROM and CD-R modulate net reflectivity at the disc surface, changing read beam intensity by controlling diffraction from the physical relief of the mastered data pits or controlling the reflectivity of recordable dye materials. CD-RW switches a phase change material between crystalline and amorphous states, depending on its temperature rise under laser heating and subsequent cooling. Having different reflectivities, these amorphous and crystalline states are read as amplitude modulated data. In magnetooptical discs, thermomagnetic laser writing modifies the magnetic state of certain amorphous thin films. Read beam polarization is rotated by the magnetic state, recreating the stored data modulation.

Optical memory access time, the sum of track seek time, drive settling time, and track latency, is relatively slow. An optical head is complex (Fig. 3) and slow to accelerate. Constant-linear-velocity (CLV) drives must adjust rotation speed for each seek. For a 1× DVD, with constant linear velocity of 3.8 m/s, rpm (revolutions per minute) changes from 574 rpm at the outermost track to 1528 rpm at the innermost track. Access times are not greatly improved by higher disc speeds, but greater linear velocity improves data transfer rate. CD-ROMs are read at about 1.41 Mbps (megabits per second) (1×) while 1× DVDs have a user rate of 11.08 Mbps. Optical memories using zoned linear velocity (ZLV) eliminate settling time during localized head movements. Some advanced DVD drives, mimicking magnetic drives, use constant angular velocity (CAV).

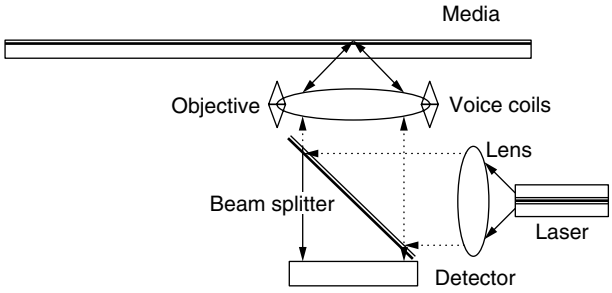


Figure 3. Schematic of a CD head. The laser diode emits elliptical light that is circularized and collimated by a lens. A thin plate directs the beam to the objective lens for focusing through the substrate and onto the pits, just below the lacquer layer. Reflected light, modulated by the effective surface reflectivity, is directed to a detector which reads the data and also senses tracking and focusing servo signals. Voice coils on the objective allow fine pitch tracking and focusing while the entire head moves for coarse tracking.

Optical memories have good lifetime and reliability. Heads fly millimeters above the media, avoiding catastrophic head crashes, while accelerated life testing predicts a shelf life of decades for aluminum-sputtered (silver) ROM media, and longer for gold media. Recordable and rewritable media have projected shelf lives of about 100 years. Mechanical handling damage is a concern, but most contamination is a minor problem because surface contaminants are out of focus. CDs can correct up to about 500 bytes (a 2.4-mm scratch); DVDs can correct about 2800 bytes (a 6-mm scratch).

The compact-laser diode "optical stylus" enables the high capacity of optical memories. A laser beam of wavelength λ in optics with numeric aperture NA can be focused to a diffraction-limited spot diameter [5] of $0.6\,\lambda/\mathrm{NA}$ and depth of focus $0.8\,\lambda/\mathrm{NA}^2$. Typical CD semiconductor infrared diodes have $\lambda = 780$ nm, and head $\mathrm{NA} = 0.5$, giving a full-width, half-maximum (FWHM) spot diameter of about 0.9 μm, and a depth of focus of 2.5 μm. Focus, track wobble, and disc runout pose demanding requirements for disc control servos at these dimensions. Table 1 compares existing disc memories.

3. COMPACT DISCS

3.1. Plain CDs

Compact discs remain the most common optical memories. CDs have mastered marks that are about 0.6 μm wide and from 0.83 to 1.7 μm long. Data are encoded in a non-return-to-zero-inverted (NRZI) format, so 1s (ones) occur at both

Table 1. Comparison of Commercial Optical Memories[a]

Type	CD-ROM	DVD-ROM	DVD-RAM	DVD-R	DVD-RW	MO
Capacity (GB)	0.68	4.7–17.1	4.7 or 9.4	4.7 or 9.4	4.7 or 9.4	2.6
Transfer rate (Mbps)	1.23	11.08	22.16	11.08	11.08	31.2
Rewrite Cycles	NA	NA	>100,000	NA	>1000	>1,000,000

[a]Transfer rates are for single speed drives. CD-ROM 40× drives can deliver up to Mbps transfer rate under constant angular velocity readout. DVD capacity ranges depend on layers in the media. Rewrite cycles are ultimately limited by thermally induced degradation of the active media layer. GB = 10^9 bytes.

edges of marks and 0s are clocked within marks and on the intervening lands. Marks, called "pits" because they are about 130 nm below the surface of the surrounding land, are approximately $\lambda/4$ deep for a 780-nm read laser in polycarbonate ($n = 1.58$). A diffraction pattern arises when the read beam overlaps a pit and its adjacent land regions (Fig. 4). Some diffracted modes do not reenter the read lens, so reading over a pit returns less light than a flat land area.

CD digital audio established the basis for the CD product family. CD-DA can hold up to 74 min of audio, played at 150 kbps, and organized in up to 99 "tracks" per disc. One continuous spiral of data is written. A single directory area after leadin locates tracks on the spiral. Multiple recording sessions or incremental writes are not possible.

The *Red Book* standard (ISO 10149) specifies CD formats [6,7]. Both level 0 (physical layer) and level 1 (sector and track formatting) specifications are given. Level 0 requires cross-interleaved Reed–Solomon code (CIRC) for error correction and EFM for establishing a (2,10) run-length-limited (RLL) disc format. RLL encoding improves mark edge detection, synchronization, and frequency management in playback.

3.2. CD-ROMs

The *Yellow Book* standard established CD-ROM. The CD-ROM sector provides 2048 data bytes (compatible with computer data structures), and uses the rest of the standard sector for additional error correction (mode 1 CD-ROM). The Yellow Book standard also specifies logical sector and logical file organization, and allows retention of a Red Book audio region with CIRC error correction on the disc (mode 2 CD-ROM). Mode 2 access times were slow, as players moved between computer applications (video clips) and the accompanying audio, in later sectors, making multimedia applications unsatisfactory. Mode 1 Yellow Book CD-ROMs can store computer data, compressed audio, and compressed video, and have stronger error correction than mode 2. A Yellow Book extension for multimedia, CD-ROM XA (extended architecture), stores compressed audio (adaptive differential PCM, ADPCM) close to the multimedia sectors to allow smooth access to images and sound. This also allows efficient compressed audio storage. Up to 18 h of monaural sound can be stored on one CD-ROM XA.

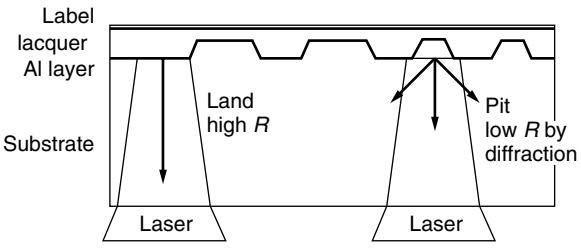


Figure 4. CD disc diffraction structure. As the laser read beam focuses on the flat land, most light is reflected back to the detector by the aluminum layer. Pits and their adjacent off-track land diffract the laser beam away from the objective, yielding a lower effective reflectivity. Polycarbonate substrates are 1.2 mm thick, while pits are only 130 nm deep, and lie just below the top surface.

Specialized standards evolved as CD-ROM was expanded into more demanding applications. The *Green Book* standard addressed efficient storage and access of a mix of data types, primarily for set-top interactive devices, such as Phillips Compact Disc Interactive (CD-I). The same track holds different data types, with each sector having a field that identifies its data type. For example, four levels of sound, from Red Book "CD quality" to various ADPCM levels of quality can be mixed with video, MPEG 1 and 2, and still pictures. CD-I has not made much market impact, and DVD is making the Green Book standard obsolete. The *White Book* standard addresses video CD (CD-V), CD-I, and PhotoCD exchange. While CD-V has been successful in China, Japan, and parts of Europe, it has had little impact in North America and is being replaced by DVD. PhotoCD is evolving its own hybrid standard from the Green Book and Orange Book standards.

The ISO 9660 standard, which evolved from the High Sierra File (HSF) format, addresses file structure and naming problems arising in cross-platform CD-ROM compatibility. ISO 9660 provides a common file structure to CD-ROM developers and is now used by most CD-ROM, CD-R, and CD-RW systems to convert the raw sector data to files, directories, and volumes. There continue to be modifications to the ISO 9660 standard, including Universal Disc Format (UDF) [8] for DVDs and UDF Bridge, which gives backward compatibility to ISO 9660 for readers that accept CD-ROM and DVD.

3.3. CD-R Media

CD-R media, initially called *CD write-once* (CD-WO) or *write-once read-many* (WORM) media, support permanent recording of a single CD. CD-R allows inexpensive, removable, archival optical memory, suitable for limited distribution of applications, images, video, audio, or data, testing of CDs before mass replication and mastering of data images for sending to a disc replicator. CD-R media store up to 650 MB of data and are compatible with all more recent drives under ISO 9660. Data must be streamed continuously throughout each session recording, or the directory information will be incorrect or missing. CD-R is not erasable, so errors in recording cannot be corrected. Typically an image of a complete session is created in a hard-drive partition or on another CD-R. Data buffering gives some security against short interruptions during writing.

On the physical layer, CD-R discs have an active dye layer of cyanine (greenish), phthalocyanine (gold-green), or metal azo (blue) between the polycarbonate substrate and the reflective metallic layer. During writing the dyes absorb heat under the laser spot and permanently change their local reflectivity. Dyes are designed to give high reflectivity change, high speed of response, good consistency, and long shelf life. Early ablative WORM recording thermally vaporized pits in a thin surface metallic film, giving us the term "burning" a CD. Some drive manufacturers wobble the mastered tracking groove to produce extended capacity CD-R discs storing about 700 MB, but these may not be compatible on playback

with all drives. Poor quality dyes on low-cost media may also prevent drives from reading CD-R discs.

The *Orange Book* standard specifies how the various CD application standards should be written onto recordable media, including not only CD-R but also CD-RW and magnetooptic (CD-MO and MO). Orange Book Part I discusses MO systems; Part II describes recording on CD-R, CD-WO and WORM devices; Part III discusses CD-RW. The Orange Book standard defines multisession recording, specifically how data should be stored in the data sectors, where track, session, and directory data are located, how disc and session lead-in and leadout are handled, and where write laser calibration regions occur. Each session of recording requires about 13 MB of overhead for its leadin/leadout and directory areas.

3.4. CD-RW

Rewritable CD-ROM, CD-RW, is the most versatile form of CD optical memory, allowing archiving, testing, distribution, and mastering for replication like the CD-R, with the addition of erasability. The physical data recording process, and the organization of the disc differ from CD-R, and drive requirements are more stringent.

CD-RW media have a polycarbonate substrate, with a mastered pregrooved spiral track. An optical stack manages heat absorption and diffusion during writing and facilitates reflectivity-based readout. The active layer is a metallic semiconductor, usually GST ($Ge_2Sb_{2.3}Te_5$), or AIST (AgInSbTe). These switch from a crystalline state to an amorphous state when the film is heated above its melt temperature. A weaker laser pulse, reaching only the glass transition temperature, returns an amorphous region to crystalline. Since these states have different reflectivity, marks can be written. CD-RW can perform direct overwrite (DOW), allowing faster data transfers. Phase change reflectivity signals are weaker than CD-ROM or CD-R signals, so older drives cannot read CD-RW media. Newer multiread drives have automatic gain control to adjust laser power to CD-RW signals.

In addition to Orange Book requirements, the Universal Disc Format packet writing scheme can be used to give CD-RW compatibility with DVD players. Under UDF, CD-RW discs must be formatted with logical sectors, a process that takes up to 30 min. Preformatted CD-RW media are entering the market. Depending on formatting, a CD-RW contains from 650 MB to 535 MB of user data. CD-RW phase change materials suffer from limited cycle lifetimes. CD-RW media must support at least 1000 write–read cycles; some media manufacturers claim 10,000 cycles.

4. DVD

DVD-ROM is the baseline digital data storage standard for all DVD applications, including DVD video and DVD audio [9]. Unlike CDs, where computer data standards evolved from the audio Red Book, DVDs were designed from the beginning for data memory use. DVD-ROM media store their data as pits that modulate readout intensity, similar to CDs.

Several engineering improvements increase bit density from about 480 Mb/in.^{2} (megabits per square inch) on CDs to over 2.2 Gb/in.^{2} on single-layer DVDs. Marks can be made smaller and closer because a higher NA (0.6), and a red (635-nm) laser yield a smaller diffraction-limited spot ($\approx$0.6 μm). The DVD spiraling data pattern is also more compact, with track pitch 0.74 μm and minimum mark spacing 0.4 μm. A DVD-5, which is a one-sided, one-layer DVD, holds approximately 7 times the capacity of a CD-ROM.

The second innovation of the DVD family is multilayered memory. DVD-ROMs are made from two 0.6-mm injection-molded polycarbonate discs, which are bonded together. Individual single-layer discs are mastered much like CDs. A single-sided disc has a data-bearing disc bonded with a spacer disc. For two-layer storage, two one-sided discs can be bonded, or two layers can be fabricated on one disc, which is then bonded with a spacer disc. In a single-sided two-layer disc, the first layer is injection-molded and covered with a semitransparent reflective layer. A second layer is bonded above the reflective layer, stamped with its own data, and coated with a full reflection layer. Two such 0.6-mm discs can be bonded to make a four-layer DVD (Fig. 5). During reading and writing the laser must focus on the appropriate layer. On a two-layer structure, the laser must read through the top layer of data to the deeper second layer, an additional distance of 55 μm. Two-sided discs must be flipped over or have two heads.

DVDs are defined by a set of application standards, also called "books," maintained by an industry consortium. The standards have not had the wide formal review of a standards organization like ISO/IEC, but nonetheless facilitate the spread of compatible media and players. The entire DVD family is based on the DVD-ROM (Book A), which supports computer applications. DVD-Video (Book B), DVD-Audio (Book C), DVD recordable (DVD-R, Book D), and DVD-rewritable (DVD-RAM, Book E) build on the Book A standard. The consortium (Hitachi, Ltd., Matsushita Electric Industrial Co., Ltd., Mitsubishi Electric Corporation, Phillips Electronics N.V., Pioneer Electronics Corporation, Sony Corporation, Thomson Multimedia, Time-Warner Inc., Toshiba Corporation, Victor Company of Japan, Ltd.) has established a

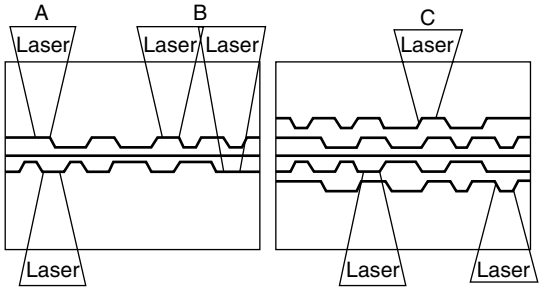


Figure 5. DVD disc structures. Two 0.6 mm CD-like discs are bonded together. A one-sided DVD (not shown) bonds a data disc to a blank substrate. Two-layered DVDs usually bond two data discs, and must be read from two sides (A). Careful reflectivity control allows two-layered DVDs that are read from one side only (B). If two, two-layered discs are bonded together, a four-layered DVD results. These require precise reflectivity control and two-sided reading (C).

jointly owned company (DVD Format/Logo Licensing Corporation) to license DVD formats and logos to third parties and to distribute the DVD Format Books [10].

DVD data are always organized in sectors of 2064 bytes, of which 2048 are data and the rest overhead for addressing, error correction, and copy protection. A combination of 16 byte Reed–Solomon code for the data columns and a 10 byte inner Reed–Solomon code for the rows are used. An $\frac{8}{16}$ run-length modulation is used to generate an acceptable physical sector (4836 bytes), and NRZI transitions produce the actual pit pattern on the disc.

4.1. DVD ROM

The simplest DVD-ROM is the DVD-5, which has a single-layer mastered disc and a blank spacer disc. It is read from one side, at one depth, and has the usual lacquer covering and printed label. Capacity is 4.7 GB per disc. A DVD-9 has two single-layer discs, but is read from one side only. DVD-9 capacity is 8.5 GB. DVD-10 has two single-side discs, each with fully reflective coatings. DVD-10 is read from both sides, and holds 9.4 GB. DVD-18 has two double-layer 0.6-mm discs, each with both partial and full reflection coating, and is read from both sides. Capacity is 17.1 GB. Raw bit densities for all DVD formats remain at about 2.2 Gb/in.^{2}.

A proliferation of DVD applications has created problems with compatibility among DVD memories. DVDs most often use the UDF file system and ISO 9660 file system. As data storage media, they could use other file systems, but this would increase the possibility of drive–media incompatibilities. The DVD Forum has created many application formats, and not all players support all formats. The MultiRead and MultiRead 2 specifications of the Optical Storage Technology Association, and the Multi logo of the DVD Format/Logo Licensing Corporation guarantee that drives will read certain classes of DVDs (and CDs).

4.2. DVD Application Standards

Application standards build on the physical layer specifications of DVD-ROM, file structures such as UDF and ISO 9660, and their extensions. DVD video is an application standard that delivers video, high-performance audio, presentation control information (PCI), and data search information (DSI). DVD video is supported worldwide by computer makers, movie studios, and content publishers. Typical data rates are about 4.7 Mbps, while peak application data rates are over 10 Mbps.

The basic DVD-ROM can accept a variety of video and audio streams, including MPEG-1 and MPEG-2 video, and MPEG-1, MPEG-2, PCM, and Dolby Digital audio, with playing times from one to 13 hours for DVD-5. DVD video media and players should conform to the UDF standard or to the MicroUDF format, which places additional constraints on file structures to simplify consumer electronics.

DVD video supports MPEG-2 video, a choice of multichannel audio formats, and extensive supplemental materials. Players output analog NTSC and PAL, digital interfaces for S-video and HDTV, and different video aspect ratios and display formats.

DVD audio shares many features with DVD video, including storing video and still pictures to accompany audio tracks. This has spawned multiple players, including DVD audio/video, video-capable audio players, and audio-only players. Audio is stored with linear PCM at 16, 20, or 24 bits/sample, with sampling frequency ranging from 44.1 to 192 kHz. Lossy and lossless compression are allowed in the specifications.

The market for DVD-ROM is led by DVD-5, with about 78% of disc releases. DVD-9 and DVD-10 each have about 10% of the total, while DVD-18 has been well under 1%. Almost all these releases have been video titles or games. Player sales rapidly in 2001, and may soon surpass VCR sales.

4.3. Recordable DVD Media

Recordable DVD-R media and players with 4.37-GB capacity began appearing widely in late 2000. The Book E standard is split into DVD-R(A) for authoring and DVD-R(G) for general or home use. These differ in laser wavelengths and land prepit addressing schemes. DVD-R(A) is single-sided, while DVD-R(G) is two-sided and must be flipped over to access both sides. Like CD-R, both DVD-R discs record data permanently by modifying a dye layer, and can support both disc-at-once and session recording. Multilayer DVD-R is not available. DVD-R cannot be erased.

Three rewritable versions of DVD are competing: DVD–RW, DVD–RAM, and DVD+RW. All use phase change materials and support 4.37-GB capacity. +RW and –RW can be rewritten about 1000 times before the phase change materials become unreliable, while RAM, which uses random shifts of the starting write position to reduce media stress, are supposed to withstand up to 100,000 rewrites. RAM and +RW use cartridges, while –RW is usually a bare disc.

DVD-RAM (random-access memory) is the closest product to a fully rerecordable optical memory, and has several technical innovations compared to other DVD rewritable systems. Its data transfer rate is 22.16 Mbps, equivalent to an 18× CD, and marks are recorded both in the land and in the premastered grooves. Zoned linear velocity is used to give good access times. A defect management scheme allows control of manufacturing and formatting defects for more reliable recording.

Rewritable media use *content protection for recordable media* (CPRM) to prevent content theft through unauthorized duplication of DVDs, a major concern of video content suppliers. CPRM places a unique 64-bit media ID in the substrate within the burst cutting area, a band just outside the clamping diameter. The media ID is used to encrypt and decrypt the disc data, such that a rerecorded disc, lacking the media ID, will be unplayable. Other security schemes are being pursued.

5. MAGNETOOPTIC DISCS

Magnetooptic discs offer rewritable, removable, high-capacity optical memory [11,12]. MO media have a shelf life projected to be over 100 years, do not require a

cartridge, and have been rewritten over a million times without losing reliability. One product, the Sony MiniDisc, has been moderately successful, but overall MO discs have not had a widespread market impact on optical memories. MO media and players have been significant in niche markets where high capacity, removability, and long archival lifetimes are important. For most applications MO memories face strong competition from improving magnetic drives and now DVD recordable media.

MO media record in an active layer containing a rare-earth transition metal amorphous alloy, such as TbFeCo, whose magnetic spins prefer to align perpendicularly to the film surface. During thermomagnetic writing, a high-power laser spot heats the active layer to about 250 °C. This elevated temperature reduces the film's magnetic coercivity, allowing a bias field of a few hundred oersted to flip the magnetization. On cooling, the reversed domain persists, creating a mark. Track pitch is typically 1.6 μm, and error-correction coding is similar to CD systems. Run-length encoding of data is used to enhance the readability of the magnetic marks. The MO layer is usually in an optical stack to enhance laser coupling to the metal. A reflective and heat-absorbing layer lies under the stack.

An alternate writing scheme uses magnetooptic magnetic field modulation (MO-MFM) to create domains in a continuously heated stripe under the moving laser beam. This can give higher along-track densities than laser power modulation, but is limited by the dynamics of the biasing magnet.

Readout uses the Kerr effect, in which the polarization of a laser read beam is rotated by the magnetic state of the film. Kerr effect rotation is less than one degree between mark and land, so a differential detection system is needed, with two detectors and polarizing elements in the head. This more complex head, and the bias magnet, require careful design if access speeds are to be maintained. Direct overwrite has also been difficult to achieve, although ingenious use of magnetic multilayers and careful control of pulse power and duration have demonstrated direct overwrite.

Preformatted grooves, including synchronization and header marks, are used to guide the writing and reading processes. The grooves are similar to CD and DVD pregrooves, and create a push–pull tracking and focusing signal. MO capacities are similar to DVD-5, but discs are typically 133 mm. The Orange Book standard applies to commercial MO media.

6. ADVANCED DISC MEMORIES

Near-term improvements in disc-based optical memories will require writing smaller marks to increase areal density, and better focusing and tracking to allow more layers. Several technologies are under development and show strong promise of continuing improvement in capacity and data transfer rate.

6.1. Blue-Violet Lasers

Blue laser diodes, at about $\lambda = 400$ nm, allow increased areal densities in phase change and MO media to over 6 Gb/in.^{2} or 15 GB per disc. Systems for mastering blue laser DVDs are available, but reasonably priced and reliable players must wait for improved blue semiconductor laser diodes. Research in this area is represented by Kondo et al. [13], who report a test on a 19.8-GB single-layer disc, mastered with a 351-nm krypton ion laser, and read with a blue-violet 405-nm laser diode (NA 0.70). Partial-response maximum-likelihood encoding and Viterbi decoding were used.

6.2. Solid Immersion Lens Technologies

Near-field optics [14] optically reduce the size of the coupled spot on the media. Solid immersion lens (SIL) heads incorporate a hemispherical lens that is cut or polished to give an effective NA well above 1.0. A conventional objective lens focuses onto the SIL, which couples a subwavelength spot to the surface. The major drawback is that the SIL must fly above the disc at a height less than 40–80 nm, similar to the flying height of a magnetic head. This compromises removability and robustness of the media head system. Areal density of 50 GB per disc and data transfer of 20 Mbps have been claimed; tracking control remains a problem.

6.3. Novel Laser Configurations

New structures for lasers and new laser-media configurations offer the possibility of higher capacities and data transfer rates. Vertical cavity surface-emitting lasers (VCSELs) are inexpensively manufactured in arrays. Present power levels and wavelengths are not impressive for optical memories, but their array structures suggest the possibility of highly parallel reading and writing. Novel apertured lasers are coated at their front facet, and small apertures are then ion-milled to release a near-field beam smaller than the wavelength. Flying close to the disc, these lasers write and read with a subdiffraction limit spot. A laser-media scheme by Aikio and Howe [15] uses the media itself as part of an external cavity to the read laser. As surface reflectivity is modulated by the data, the reflected light returning to the laser cavity varies and modifies the laser output power. Power can be monitored from the rear facet photodetector common on laser diodes.

6.4. Fluorescent Discs

Frequency multiplexing in the same media volume is possible if the read beam stimulates emission at different wavelengths. This concept underlies the fluorescent multilayered disc (FMD). Many thin active layers would be deposited on a single disc (structures with up to 500 layers have been proposed). With adequate focusing control a small band of layers could be excited by the read beam. If these layers fluoresced at different wavelengths, then one single layer could be filtered and read. A continuing AFOSR project has reported successful design of the head and detectors for this system [16,17].

6.5. Advanced Magnetooptic Systems

SIL near-field optics apply to MO systems, and can be combined with magnetic field modulation and superresolution

multilayered magnetic media to attain extremely high densities. The possibility of MO writing up to 100 Gb/in.^{2} has been claimed with these combined methods [18]. Using just magnetic superresolution with blue lasers has yielded 11 Gb/in.^{2} or 15 GB per disc.

7. HOLOGRAPHIC OPTICAL MEMORIES

After over four decades of effort, holographic memories have shown significant progress recently with impressive laboratory capacities and data rates [19,20]. These demonstrations, while not yet yielding commercial products, have benefited from improved enabling technologies [lasers, spatial light modulators, and CCD (charge-coupled device) cameras] and novel storage media. Holographic memories promise capacities of 125 GB per disc, data rates of 1–10 Gbps, and rapid access times. As a volumetric storage method, their capacities will grow as $1/\lambda^3$ when laser wavelengths get shorter, rather than as $1/\lambda^2$ for surface methods. Both stationary solid media and disc-based media are under study.

Holographic memories store information in a thick photosensitive medium as a phase modulation pattern (hologram) created by the interference of two coherent laser beams: a reference beam and the object or data beam (Fig. 6). The interference pattern of the two beams, captured by changes in the absorption, index of refraction or thickness of the photosensitive medium, can later be probed with a replica of the reference beam. An incident reference beam will be diffracted by the phase pattern to recreate the original data beam, traveling forward as that beam had originally done. It also creates a phase-conjugate beam that travels back toward the original data beam source. This allows reuse of the high-quality writing optics during readout.

Writing is not done bit-by-bit, but rather with a 2-D pattern of data, a page image, typically created by a spatial light modulator (SLM) similar to a LCD display.

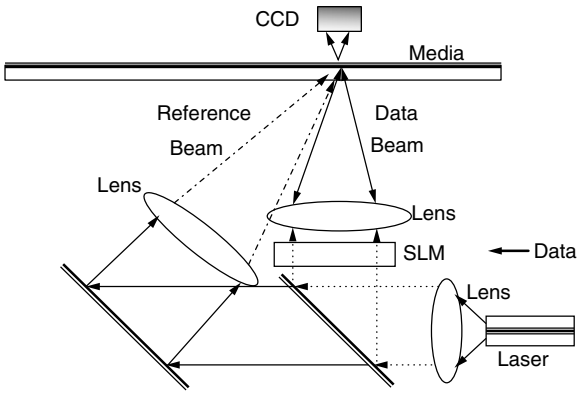


Figure 6. Schematic of a holographic disc system. The laser emits a beam that is collimated and split. One path goes through a spatial light modulator (SLM) that carries a data page, spatially modulating the beam. The second beam is directed as a reference beam. The beams interfere in the disc media creating a hologram indexed by the reference beam angle. On readout the reference beam creates a transmitted beam that is read as a page on the CCD device.

The recreated page image beam can be read by any 2-D array of detectors, e.g., a CCD camera. Multiple pages can be written in the same medium, since individual gratings can be distinguished by multiplexing the incident angle or tuning the laser to another wavelength. Focusing of the beams localizes the holographic patterns in the media. Theoretical capacities depend on the storage media, but exceed 100 Tb/in^3. Considerable technical problems must be overcome to make robust commercial systems based on holography. DARPA has supported the Photo Refractive Information Storage Materials (PRISM) consortium and the Holographic Data Storage System (HDSS) consortium to explore these problems and some progress was demonstrated, including 10 Gb/s data transfer rate [21]. An important tradeoff exists between data transfer rate and media storage density in holographic materials with fixed dynamic range.

Moving a laser probe beam is relatively simple compared to moving a massive read head. Access speeds could therefore be better in holographic memories. Since data is read a page at a time, and since multiple independent probe beams can be read at the same time, aggregate data rates also could be high. Finally, holographic memories offer an associative memory capability that is not possible with any other storage method. When the medium is probed with a search pattern (a *partial* replica of a desired data beam), reconstructed source beams will be created. Each will have intensity proportional to the overlap of their image with the search image. By reading the most intense beam's location with a reference beam, the original data page can be reconstructed.

BIOGRAPHY

Michael F. Ruane received his B.E.E. in 1969 from Villanova University, Villanova, Pennsylvania, and the S.M.E.E. and Ph.D. (Systems) from Massachusetts Institute of Technology, Cambridge, Massachusetts, in 1971 and 1980, respectively. He was a staff member at the MIT Energy Laboratory from 1971 until 1976, and joined Boston University, Boston, Massachusetts in the Electrical and Computer Engineering Department in 1980. At BU he was one of the developers of the Photonics Center, where he maintains the Magnetic and Optical Devices Laboratory, and is also the BU education coordinator for the Center for Subsurface Sensing and Imaging System (CenSSIS). His laboratory studies data storage media and systems, and has contributed to magneto-optical devices, phase change media, disc mastering, and conventional magnetic media. Dr. Ruane has two patents in ellipsometry for optically active media. His areas of interest are in optical systems, the optical storage channel, micromagnetic modeling, and image processing.

BIBLIOGRAPHY

1. H. Varian and P. Lyman, (Oct. 18, 2000) Project Home Page (online), *How Much Information?* Berkeley School of Information Management & Systems, Univ. of California at Berkeley, *http://www.sims.berkeley.edu/research/projects/how-much-info/*, March 30, 2001.

2. C. D. Mee, and E. D. Daniel, eds., *Magnetic Storage Handbook* (*Parts 1 and 2*), 2nd ed., McGraw-Hill, New York, 1996.
3. L. Purcell, *CD-R/DVD: Disc Recording Demystified*, McGraw-Hill, New York, 2000.
4. F. Yu and S. Jutamulia, eds., *Optical Storage and Retrieval*, Marcel Dekker, New York, 1996.
5. Marchant, *Optical Recording: A Technical Overview*, Addison-Wesley, Reading, MA, 1990.
6. International Standards Organization, list of ISO standards related to optical data storage, Search Engine Listing (online), *http://www.iso.ch/cate/3522030.html*, March 30, 2001.
7. D. Chen and J. Neumann, Status of international optical disc standards, *Proc. Recent Advances in Metrology, Characterization, and Standards for Optical Digital Data Discs*, SPIE 3806, 1999.
8. Optical Storage Technology Association, (2001, March 30) reference documents for the Universal Data Format standard and revisions (online), *http://www.osta.org/html/ostaudf.html*, March 30, 2001.
9. J. Taylor, *DVD Demystified*, 2nd ed., McGraw-Hill, New York, 2000.
10. Phillips International N.V., Systems Standards & Licensing, homepage for licensing of DVD technology (online), *http://www.licensing.philips.com*, March 30, 2001.
11. M. Mansuripur, *The Physical Principles of Magneto-optical Recording*, Cambridge Univ. Press, Cambridge, UK, 1995.
12. R. Gambino and T. Suzuki, eds., *Magneto-optical Recording Materials*, IEEE Press, Piscataway, NJ, 2000.
13. T. Kondo et al., 19.8-GB ROM disc readout using a 0.7-NA single objective lens and a violet laser diode, *Proc. Optical Data Storage 2000*, SPIE 4090, 2000, pp. 36–42.
14. T. D. Milster, Near field optics: A new tool for data storage, *Proc. IEEE* **88**(9): 1480–1490 (Sept. 2000).
15. J. Aikio and D. G. Howe, Direct semiconductor laser readout in optical data storage, *Proc. Optical Data Storage 2000*, SPIE 4090, 2000, pp. 56–65.
16. DARPA VLSI Photonics Program Summaries, (2001, January 18). overview page (online), *http://www.darpa.mil/MTO/VLSI/Overviews/Callrecall-4.html*, March 30, 2001.
17. H. Zhang et al., Single-beam two-photon-recorded monolithic multi-layer optical discs, *Proc. Optical Data Storage 2000*, SPIE 4090, 2000, pp. 174–178.
18. D. C. Karns et al., To 100 Gb/in.^{2} and beyond in magnetooptic recording, *Proc. Opt. Data Storage 2000*, SPIE 4090, 2000, pp. 238–245.
19. J. Ashley et al., Holographic data storage, *IBM J. Res. Devel.* **44**: 341–368 (May 2000).
20. H. J. Coufal, D. Psaltis, and G. Sincerbox, eds., *Holographic Data Storage*, Springer-Verlag, Heidelberg, Germany, 2000.
21. National Storage Industry Consortium, description of consortium projects for enhancing magnetic and optical data storage, home page (online), *http://www.nsic.org* March 30, 2001.

FURTHER READING

Annual meetings on optical memory and optical data storage are sponsored by IEEE/Lasers and Electro-Optics Society (LEOS), Optical Society of America (OSA), the International Society for Optical Engineering (SPIE) and other groups. Proceedings appear as SPIE volumes and provide the most convenient access to current research on more advanced optical memories. The most recent volumes include the following:

Mikaelian A. L. ed., *Proc. Optical Memory and Neural Networks*, SPIE 3402, 1998.

Mitkas P. A., and Z. U. Hasan, eds., *Proc. Advanced Optical Memories and Interfaces to Computer Storage*, SPIE 3468, 1998.

Petrov V. V., and S. V. Svechnikov, eds., *Proc. Int. Conf. Optical Storage, Imaging, and Transmission of Information*, SPIE 3055, 1997.

Sincerbox G. T., and J. M. Zavislan, *Selected Papers on Optical Storage*, SPIE Milestone Series, MS-49, 1992.

Sincerbox G. T. *Selected Papers on Holographic Storage*, SPIE Milestone Series, MS-95, 1994.

Proc. Joint Int. Symp. Optical Memory and Optical Data Storage 1999, SPIE 3864, 1999.

OPTICAL MODULATORS — LITHIUM NIOBATE

RANGARAJ MADABHUSHI
Agere Systems, Optical Core Networks Division
Breinigsville, Pennsylvania

1. INTRODUCTION

With the advent of the laser, a great interest in communication, at the optical frequencies, was created. A new era of optical communication was launched in 1970, when an optical fiber, having 20 dB/km attenuation, was fabricated at the Corning Glass Works. Dr. Kaminow and a team from Bell labs reported the concept of electrooptic light modulators [1]. At the same time, Miller [2] coined the term "integrated optics" and heralded the beginning of various efforts in a number of optical components including light sources, waveguide devices, and detectors. The demand for fiberoptic telecommunication systems and larger bandwidth requirements, has increased tremendously since the early 1990s, with the advent of time-division multiplexing (TDM) and wavelength-division multiplexing (WDM) systems. In these systems, the transmitter part basically consists of a laser, which provides the coherent optical (light)wave and the modulator (either external or the direct modulation of lasers), where the desired signal is modulated and is placed on the coherent lightwave.

The direct modulation of lasers is limited by the achievable bandwidth, chirp, or dispersion and the ability to be transmitted to longer distances. The advantages, for short-distance transmission applications include small device size and cost-effectiveness. On the other hand, external modulators are bulky and costly and increase the system requirements. But the advantages, such as large bandwidths and capability to propagate long distances, make these external modulators the winners in optical communication systems. The external modulators include

devices made of dielectric crystals, such as lithium niobate and lithium tantalite; semiconductors such as GaAs, InP, and InGaAs; and polymers such as PMMA. The lithium niobate–based modulators have the advantages of large bandwidth capabilities, low chirp characteristics, low insertion loss, better reliability, and improved manufacturing capabilities. The disadvantages include higher driving voltages, large size of the device, and high cost. The semiconductor modulators have the advantages of smaller size, low driving voltages, relatively low cost (for large volumes), and compatibility of future integration with other semiconductor devices. The disadvantages include large insertion loss, smaller transmission distances, chirp, and manufacturing yields. The polymers are just emerging, and although they can achieve large bandwidths and low driving voltages, the long-term reliability is still being investigated. The $LiNbO_3$ modulator technology, which started in late 1960s, advanced in terms of the material properties, fabrication process, and various modulation schemes in all these years [3–9]. Here, the lithium niobate external modulators are discussed.

1.1. Optical Modulation

It is possible to realize various optical devices, by controlling externally, the lightwave propagating in the optical waveguide. Optical modulators are the devices made of optical waveguides on some material with special properties, where the information is placed on the lightwave externally by imposing time-varying change on the lightwave. The information content is then related to the bandwidth of the imposed variation. Similarly, switches are devices that change the spatial location of the lightwave with respect to the switching signal. These modulators and switches are important components in most of the optical communication systems. The materials may have physical properties, such as electrooptic effect, acoustooptic effect, magnetooptic effect, and thermooptic effect [4].

The modulation types include intensity or amplitude modulation, phase modulation, frequency modulation, and polarization modulation. The intensity modulators are those in which the intensity or amplitude of the coherent lightwave varies according to a time-varying signal. In phase modulation, the phase of the lightwave responds to the applied signal. If the signal is time-varying, the phase change also varies with time. The amplitude of the first sideband and the carrier amplitude are related to the Bessel functions. In polarization modulation, using the electrooptic effect, the polarization states of the lightwave respond to the signal applied. In general, when there is no signal applied, the lightwave emerges as a linearly polarized light. The changes from linear to elliptical polarization, through the applied signal, are characteristics of polarization modulators that use the electrooptic effect. In case of magnetooptic polarization modulators, the light remains linearly polarized but rotated in directions as a function of the applied signal. These polarization modulators are usually used as switches. The last one is frequency modulation, in which the frequency or the wavelength is changed with the applied signal. The detection of such frequency shifts gives rise to more complicated heterodyne system applications.

1.2. Electrooptic Effect

The *electrooptic effect* is, in general, defined as the change of refractive index inside an optical waveguide in optical anisotropic crystals, when an external electric field is applied. If the refractive index changes linearly with the amplitude of the applied field, it is known as the *linear electrooptic effect* or the *Pockels effect*. This effect is the most widely used physical effect for the waveguide modulators. The details can be found in the existing literature [e.g., 4]. Some of the basic fundamentals are given here.

The linear change in the refractive index coefficients due to the applied electric field E_z is given by

$$\Delta n_o = 0.5 r_{13} n_o^3 E_z, \qquad \Delta n_e = 0.5 r_{33} n_e^3 E_z \tag{1}$$

where n_o is the ordinary refractive index, n_e is the extraordinary refractive index, and r_{ij} is the electrooptic constant. For $LiNbO_3$, $r_{33} = 30.8 \times 10^{-12}$ m/V, $r_{13} = 8.6 \times 10^{-12}$ m/V, $r_{22} = 3.4 \times 10^{-12}$ m/V, $r_{33} = 28.0 \times 10^{-12}$ m/V, $n_o = 2.2$, and $n_e = 2.15$ at $\lambda = 1.5$ μm.

2. BASIC STRUCTURE AND CHARACTERISTICS OF THE MODULATORS

In general, the Mach–Zehnder interferometer-type structure is used in the lithium niobate–based intensity modulators. The modulator basically consists of an input divider, an interferometer, and an output combiner. The input divider consists of a straight waveguide and an input Y-branch waveguide, which divides the incoming light into two parts. The interferometer consists of two arms, to which the signal can be applied in the form of voltage. The output combiner consists of an output Y-branch waveguide that combines the two waves from the interferometer arms and finally an output straight waveguide. When there is no signal/voltage applied ($V = 0$), the input wave (field) will be divided into two equal parts, E_A and E_B. At the interference arms they propagate with the same amplitude and phase and recombine at the output Y branch and propagate in the output waveguide without change in intensity (Fig. 1a).

When a voltage is applied, the two waves at the interferometer arms change the phase of the two waves and when the applied voltage, V, is equal to the voltage required, to achieve a π-phase shift, V_π, the output waves from the interferometer have the same amplitude, but a phase difference of π. The output light will become zero by destructive interference (Fig. 1b). For the values of the voltage between V and the V_π the output power varies as

$$P_{\text{out}} = 0.5\{(|E_A| - |E_B|)^2 + 2|E_A| \cdot |E_B| \cos^2 \Delta\varphi\} \tag{2}$$

$$= 0.5 P_{\text{in}} \cdot K_1 + K_2 \cos^2 \frac{\pi V}{2V_\pi} \tag{3}$$

where the phase shift is

$$2\Delta\varphi = \frac{\pi V}{V_\pi} \tag{4}$$

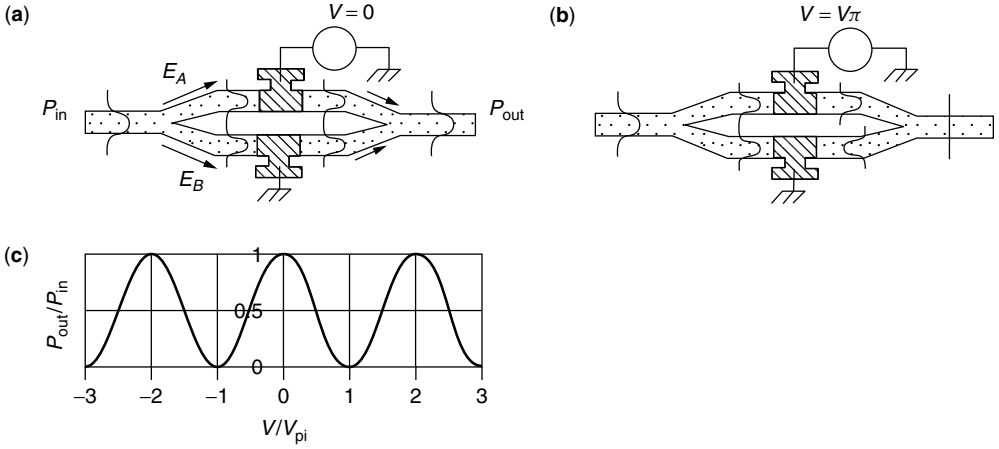


Figure 1. Basic principle of operation of a Mach–Zehnder-type optical modulator, (**a**) without, (**b**) with applied voltage; (**c**) the output intensity as a function of applied voltage.

Figure 1c shows the output intensity as the function of switching/driving voltage, which is represented by Eq. (3).

2.1. Driving Voltage

The change in the index as a function of voltage is

$$\Delta n(V) = \frac{n_e^3 r_{33} V \Gamma}{2G} \tag{5}$$

The phase difference in each arm of the interferometer will be φ, and as the voltage is applied on both arms, the push/pull effect can be used and the total phase difference will be 2φ, where

$$2\varphi = \frac{\pi V}{V_\pi}$$

The voltage length product is

$$V_\pi L = \frac{\lambda G}{2 n_e^3 r_{33} \Gamma} \tag{6}$$

where λ is the wavelength of operation (say, 1.5), n_e is the extraordinary refractive index of the $LiNbO_3$ waveguide (say, 2.15 at λ 1.5 μm), r_{33} is the electrooptic coefficient, 30.8×10^{-12} m/V, V is the voltage applied, Γ is the overlap integral between optical and electric (RF) fields (usually a value of 0.3–0.5), G is the gap between the electrodes, and L is the electrode length.

Depending on the crystal orientation (z-cut, x-cut, or y-cut), the electrodes configuration, whether the electrodes are placed on the waveguides or on the sides of the waveguide, will result in the use of vertical or horizontal fields (Fig. 2).

The overlap integral Γ is better for the z-cut modulator compared to that in the x-cutone. The driving voltage will be less in the case of the z-cut crystal orientation/vertical field, due to the large overlap factor. But there is a need to place a dielectric layer in between the electrode and the waveguides, to minimize the waveguide insertion loss for a TM mode propagation. This will increase the driving voltage. The parameters of the dielectric layer, usually the SiO_2 layer, can be used as a design parameter to achieve larger bandwidths.

2.2. Extinction Ratio and Insertion Loss

If I_o is the intensity at the output of the modulator, when no voltage is applied, $I_{\max}$ is the maximum intensity, and $I_{\min}$ is the minimum intensity when the voltage is applied, then the insertion loss is defined as

$$10 \log \frac{I_{\max}}{I_o} \tag{7}$$

and the extinction ratio (ER) is

$$10 \log \frac{I_{\min}}{I_{\max}}. \tag{8}$$

2.3. Chirp

In case of small-signal applications, the dynamic chirp $\alpha'(t)$ is the instantaneous ratio of the phase modulation to amplitude modulation of the transmitted signal and is expressed as

$$\alpha'(t) = \frac{\dfrac{d\Psi}{dt}}{\left[\left(\dfrac{1}{2I}\right)\left(\dfrac{dI}{dt}\right)\right]} \tag{9}$$

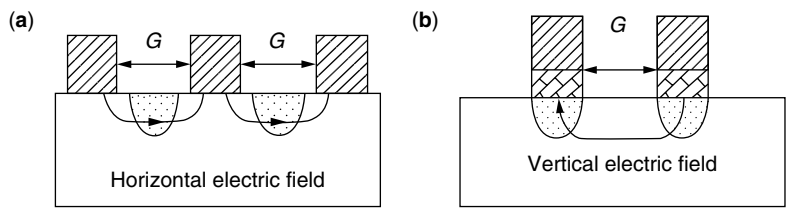


Figure 2. The normally used electrode configurations and the respective field conditions: (**a**) horizontal field used for the x- or y-cut crystal orientations; (**b**) vertical field used for the z-cut crystal orientation.

where Ψ and I are respectively the phase and intensity of the optical field and t denotes the time. In the case of the intensity modulator using the Mach–Zehnder type, the α' can be represented in a simplified form as

$$\alpha' \sim \frac{\Delta\beta_2 + \Delta\beta_1}{\Delta\beta_2 - \Delta\beta_1} = \frac{\Delta V_2 + \Delta V_1}{\Delta V_2 - \Delta V_1} \tag{10}$$

where $\Delta\beta_1$, $\Delta\beta_2$ are the electrooptically induced phase shifts and ΔV_1, ΔV_2 are the peak-to-peak applied voltages of the two arms of the interferometer. Although this expression can be applied in general to the small-signal region, it can also be applicable, to a large extent, for the large-signal region, due to the shape of the switching curve. Also, the value of α' can take minus $(-)$ or plus $(+)$ values, and the chirp can be used to the advantage, depending on the optical transmission system. For systems that operate away from the zero-dispersion wavelength region, and depending on the fiber used for transmission, a negative chirp can be advantageous to achieve low dispersion penalties [10]. In general, for a lithium niobate intensity modulator with a traveling-wave-type electrode, the value can be -0.7. Depending on the crystal orientation and the type of the electrode structure, the value can be zero or can be variable.

3. COMMON ELECTRODE STRUCTURES

A simple electrode structure, consisting of two symmetric electrodes on both interferometer waveguides, otherwise known as "lumped" electrode structure, is shown in Fig. 3a. As the bandwidth, in this case, is limited by the RC (load resistance and modulator capacitance), it is difficult to achieve large bandwidths.

The widely used electrode structure, for large bandwidths, is the traveling-wave electrode structure, where the modulator electrode structure is designed as an extension of the load resistance. Figure 3b shows the structure of a CPW, (coplanar electrode structure), which consists of a central signal electrode and two ground electrodes on both sides of the signal electrode. The two ground electrodes, have widths that are assumed to be sufficiently larger than the signal (or central) electrode structure. Figure 3c shows the asymmetric coplanar stripline (ACPS), or asymmetric stripline (ASL) electrode structure, which consists of a central signal and one ground electrode, where the ground electrode width is assumed to be sufficiently larger than that of the signal electrode. In both of these cases the bandwidth is not limited by the capacitance of the modulator but is dependent on the velocity matching and microwave attenuation of the electrode structures.

The other important characteristics include the following optical characteristics — wavelength of operation, optical return loss, maximum power, and polarization dependency, the following electrooptic and microwave characteristics — bandwidth (frequency response), microwave attenuation, characteristic impedance; and the following mechanical and long term stability — size, temperature, and DC drift stability, humidity, shock, and vibration stability, and fiber pull strength.

These characteristics need to be addressed by the modulator designer, from the initial stage. The waveguide technology is mature enough to satisfy most of the characteristics. The main characteristics that need special attention are the bandwidth and the driving voltage. The usual system requirements are larger bandwidths with lower driving voltages, due to the limitations of available low-driving-voltage drivers. Both bandwidth and driving voltage of lithium niobate modulators are in a tradeoff relationship; one has to be sacrificed for the other. For many years, modulator design has concentrated on optimizing various parameters and finding ways to achieve both larger bandwidths and lower driving voltages [11–19].

The bandwidth of a modulator is dependent on the velocity mismatch between the optical and microwave (RF) and the microwave attenuation of the electrode structure. The velocity mismatch can be controlled by the electrode/buffer-layer parameters. But once the electrode/buffer-layer parameters are fixed, the microwave attenuation (α) is also fixed. In other words, the microwave attenuation, which gets fixed by the electrode/buffer-layer parameters, limits the achievable bandwidth, even though perfect velocity matching is achieved. The driving voltage or V_π is also dependent on the electrode/buffer-layer parameters.

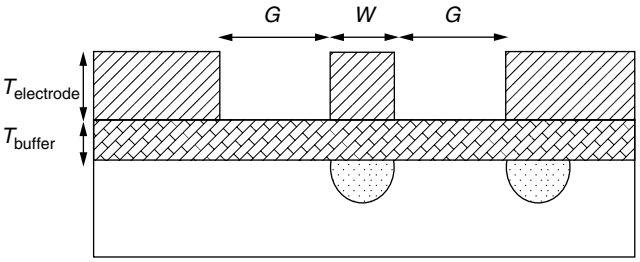


Figure 4. Cross section of a typical Mach–Zehnder optical modulator, with a CPW electrode structure.

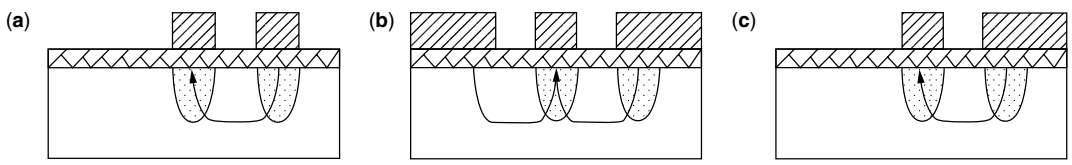


Figure 3. The most commonly used electrode structures: (**a**) lumped; (**b**) coplanar waveguide (CPW) (coplanar electrode structure); (**c**) asymmetric coplanar stripline (ACPS) electrode structure.

As the effective refractive indices of the optical wave (2.15, for TM mode, at 1.55 μm) and that of the microwave (4.2) are different, there exists the velocity mismatch between the two fields, which are propagating simultaneously. This mismatch limits the achievable optical bandwidth value. It is possible to reduce the microwave refractive index to that of the optical refractive index by optimizing the electrode/buffer-layer parameters. Figure 4 shows a cross-sectional view of the Ti-diffused $LiNbO_3$ Mach–Zehnder modulator with a CPW electrode structure. The parameters that are controlled and optimized are W, the width of the signal electrode; G, the gap between the signal and ground electrodes; $T_{\text{electrode}}$, the thickness of the electrode; T_{buffer}, the thickness of the buffer layer; and ε, the dielectric constant of the lithium niobate crystal.

Two-dimensional finite-element analysis can be used for microwave analysis to calculate the capacitance, effective microwave index, and characteristic impedance. The beam propagation method (BPM) or the propagation beam method (PBM) is used for optical field analysis.

The parameters used include the refractive index (TM modes) at 1.55 μm of wavelength, with $n_e = 2.15$, and the dielectric constants of the z-cut $LiNbO_3$ 28 for the z direction and 43 in other directions. The buffer layer is assumed to be SiO_2 with a dielectric constant of 3.9. Figure 5a,b shows the microwave refractive index n_m, and the characteristic impedance, Z, as functions of the electrode width : gap ratio, W/G, buffer-layer thickness, and electrode thickness. It can be observed that n_m decreases with increase in the buffer layer and electrode thickness. These design values depend on various experimental factors and fabrication conditions. Hence, care should be taken in incorporating the experimental values with the modulator design parameters and to ensure that the necessary optimization is performed.

The bandwidth of a modulator can be obtained from the optical response function, which can be defined as

$$H(f) = \frac{[1 - 2e^{-\alpha L}\cos 2u + e^{-2\alpha L}]^{1/2}}{[(\alpha L)^2 + (2u)^2]^{1/2}} \tag{11}$$

where

$$u = \frac{\pi f L(n_m - n_o)}{C} \tag{12}$$

$$\alpha = \frac{\alpha_0 f^{1/2}}{(20\log e)} \tag{13}$$

where α_0 = microwave attenuation constant
f = frequency
n_m = effective microwave index
n_o = effective optical index
$(n_m - n_o)$ = velocity mismatch
L = length of electrode
C = velocity of light

It is evident that even when a perfect velocity matching is achieved, the bandwidth is limited by the microwave attenuation. Thus, reduction of microwave attenuation is the key in achieving very large bandwidths.

The velocity matching using the thick electrodes and thick buffer layer is in the ACPS electrode structure reported by Seino et al. [12]. For an electrode length of 2 cm, a driving voltage of 5.4 V, a bandwidth of 20 GHz, and a microwave attenuation of 0.67 dB/[cm $(\text{GHz})^{1/2}$] were achieved. One problem of the ACPS structure is the resonance problem at higher frequencies, so there is a need to reduce the chip thickness and width. For CPW electrode structure, thick electrodes and buffer layer are utilized [13,14]. For an electrode length of 2.5 cm, a driving voltage of 5 V, and a bandwidth of 20 GHz with a microwave attenuation 0.54 dB/[cm $(\text{GHz})^{1/2}$] was achieved. The issue with a CPW electrode was higher microwave loss due to the higher-order mode propagation. Reduction of chip thickness is needed.

Reduction of microwave attenuation is the main factor in achieving very large bandwidths. The total microwave attenuation of the electrode structure can be reduced by reducing the stripline loss, higher-order mode propagation loss, losses due to bends/tapers, connector, connector-to-pad contact loss, and other package-related loss [20]. A potential reduction of the stripline electrode structure is the use of a two-stage electrode structure. For an electrode length of 4 cm, the driving voltage is 3.3 V, and the bandwidth is 26 GHz with a microwave attenuation of 0.3 dB/[cm $(\text{GHz})^{1/2}$] [21].

3.1. Driving Voltage Reduction

The driving voltage is given by Eq. (6). The driving voltage reduction can be realized mainly by increasing the electrode length or increasing Γ, the overlap integral, between the optical and RF waves, or decreasing G, the gap between the two arms of the interferometer. There is a limit to decrease of G. If the arms are too close, there is a problem of mode coupling between these two arms. This will cause a degradation of the extinction ratio. Also, G is

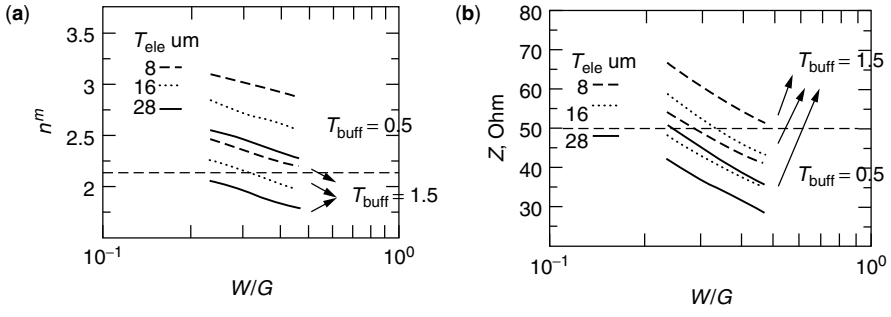


Figure 5. The calculated values of (**a**) microwave refractive index n_m, and (**b**) characteristic impedance, Z, as functions of the electrode width : gap ratio, W/G, buffer layer thickness, and electrode thickness.

the parameter that was fixed in earlier velocity matching design. Increasing the electrode length poses problems on the achievable bandwidth due to microwave attenuation problems. The driving voltage is dependent on the overlap integral between the optical and the microwave fields. The overlap integral needs to be as large as possible, and it depends on the waveguide fabrication parameters and diffusion parameters. The waveguide parameters include the titanium thickness, the titanium concentration, and the gap between the electrodes; the diffusion parameters include the diffusion time and temperature. All these parameters are to be optimized, in order to achieve strong mode confinement. Also, the position of electrodes vis-à-vis the waveguide position dictates the overlap integral value. The other important parameter is buffer-layer thickness, which increases with increase in driving voltage but decrease in overlap integral. Thicker buffer layers are needed to achieve the velocity matching, as explained above. Once the velocity matching condition is obtained, the buffer-layer thickness and the achievable driving voltage are fixed. Optimization of the waveguide/electrode parameters to achieve a strong confinement is usually the remaining issue to achieve the lower driving voltages.

Other methods to reduce the driving voltage include a dual-electrode structure, a ridge waveguide structure, and a controlled buffer-layer structure. In a dual-electrode structure [22], where the two arms of the interferometer are driven by two independent signal electrode structures, the driving voltage can be reduced by approximately half. This structure has the advantage of controlling the chirp value. By individually controlling the voltages applied to the two arms, it is possible to obtain a zero chirp or a negative/positive chirp. In the ridge waveguide structure, by etching ridges in the region, the overlap integral can be increased. At the same time, it is possible to design a modulator to achieve both the velocity matching and the required characteristic impedance. In the controlled buffer-layer structure, the thickness of the buffer layer across the waveguides to achieve both large bandwidth and low driving voltage has been reported [21,23]. The thickness is varied so that both the velocity matching condition and the low driving voltage are achieved at the same time. For the electrode lengths of 4 and 3 cm, driving voltages of 2.5 and 3.3 V, and bandwidths of 25 and 32 GHz were achieved, respectively.

3.2. Reliability

The long-term reliability was the main performance parameter that is vital for using these devices for commercial and practical systems. The DC drift and the temperature stability (and humidity drift) are the main long-term reliability issues [24,25].

3.3. DC Drift

DC drift is the optical output power variation under the constant DC bias voltage application.

Figure 6a shows the output power of the modulator as a function of the applied voltage. The dashed lines show the output power as a function of applied voltage when only AC voltage is applied (and no DC is applied, at $t = 0$), and the solid line shows the same, after $t = t_1$, when DC voltage is also applied in addition to the previous AC signal voltage. The shift between these two curves, ΔV, is the measure of the DC drift. When these types of modulator are used in practical systems, the signal is usually applied at the center of the switching curve (i.e., intermediate between maximum and minimum), which is known as the *driving point*. Once the shift due to DC drift occurs, driving point voltage must be brought back to the previous operating point, using an automatic bias control (ABC) circuit or feedback control (FBC) circuit. It is desirable to minimize this shift, and in most of the cases, a negative shift is more desirable as it facilitates a smaller voltages application through the ABC circuit. The cause of the DC drift can be attributed to the movement of ions, including OH ions, inside the lithium niobate substrate and that inside the buffer layer. It is influenced by the balance of the RC time constants, in both horizontal and vertical directions in the equivalent-circuit model as shown in Fig. 6b. It was also found that the DC drift is affected to a greater extent by the buffer layer. In the circuit model of Fig. 6b, all layers, the $LiNbO_3$ substrate, the Ti : $LiNbO_3$ optical waveguide, and the buffer layer are represented in resistances R, and capacitances C, in both vertical and horizontal directions.

It has been experimentally proved that DC drift can be reduced by decreasing the vertical resistivity of the buffer layer or by increasing the horizontal resistivity of the buffer layer (or that of the surface layer). The surface layer is the boundary layer between the buffer layer and the substrate. The reduction of the vertical resistivity is obtained by doping the SiO_2 buffer layer using TiO_2 and In_2O_3. The increase of the horizontal surface resistivity can be obtained by making a slit in the Si/SiO_2. In both

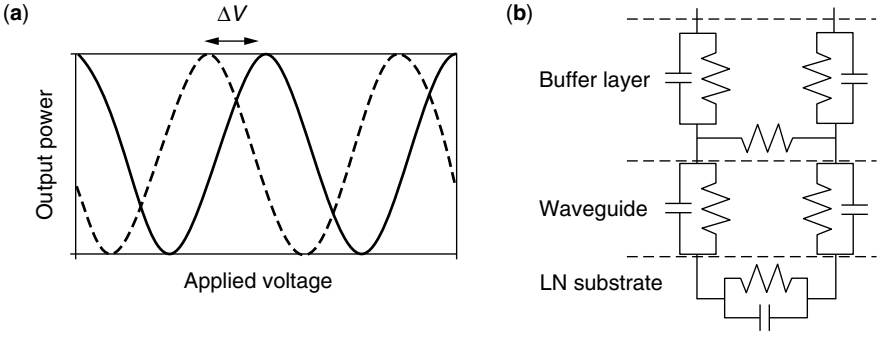


Figure 6. DC drift: (**a**) output power of the modulator as the function of the driving voltage, with and without the DC applied voltage; (**b**) equivalent RC circuit model of the structure, with vertical and horizontal components.

cases, the movement of ions, especially between the two interferometer arms (waveguides), is arrested.

3.4. Thermal Drift

Thermal drift is the optical output power variations as a function of changes in temperature. Once the temperature changes, the piezoelectric charges are induced on the surface of the LN substrate. This causes a surface charge distribution across the two arms of the interferometer, affecting the electric field. This results as a shift in the switching curve and the driving point/operation point shifts, similar to those in DC drift. Hence, in order to reduce this thermal drift, there is a need to distribute or dissipate the charges that are accumulated on the LN surface and between the electrodes. A method in which the charges are dissipated using the semiconductor layers such as Si [26] and other materials was proved to reduce the thermal drift. Another method, in which a Si double slit and a reduction of the resistivity by one order of magnitude, was reported [27].

BIOGRAPHY

Rangaraj Madabhushi received his B.S. and M.S. degrees in physics/mathematics and applied physics in 1974 and 1977, respectively, from Andhra University, Visakhapatnam, India, and a Doctor of Engineering in electronics in 1989 from Tohoku University, Sendai, Japan (optical waveguide devices based on $LiNbO_3$). After working in India, during 1977–1980 (project associate, Indian Institute of Technology, Madras, India) and 1980–1984 (senior scientific officer, Instruments Research and Development Establishment, Dehradun, India), he came to Japan in 1984 on a Japanese government scholarship. After completing a Japanese language course and a doctorate, he joined the NEC corporation, Kawasaki, Japan, in 1989. He worked at the NEC (Central research labs), in various capacities; researcher, assistant manager, manager, on the research, development, and management of $LiNbO_3$ devices, including switches, filters and high-speed modulators for optical communication. In 1999, he came to the United States and joined Lucent Technologies (now Agere Systems) as the technical manager and subsequently promoted as a director in 2000. Since then, he is managing the $LiNbO_3$ and SiWG product development at Breinigsville, Pennsylvania. Dr. Madabhushi holds over 15 Japanese and 10 U.S. patents in the area of $LiNbO_3$ devices and is the author of more than 40 papers in international conferences and journals. He is the senior member of IEEE/LEOS USA, OSA USA, and IEICE, Japan.

BIBLIOGRAPHY

1. I. P. Kaminow, T. J. Bridges, and E. H. Turner, Electrooptic light modulators, *Appl. Opt.* **5**: 1612–1614 (1966).
2. S. E. Miller, Integrated optics: An introduction, *Bell Syst. Tech. J.* **48**: 2059–2069 (1969).
3. H. F. Taylor and Y. Yariv, Guided wave optics, *Proc. IEEE* **62**: 1044–1060 (1974).
4. T. Tamir, ed., *Integrated Optics*, 2nd ed., Topics in Applied Physics, Springer-Verlag, New York, 1979.
5. R. C. Alferness, Waveguide electrooptic modulators, *IEEE Trans. Microwave Theory Tech.* **MT-30**: 1121–1137 (1982).
6. S. K. Korotky, J. C. Campbell, and H. Nakajima, Special issue on photonic devices and integrated optics, *IEEE J. Quant. Electron.* **QE-27**: 516–849 (1991).
7. K. Komatsu and R. Madabhushi, Gb/s range semiconductor and Ti : $LiNbO_3$ guided-wave optical modulators, *IEICE Trans Electron.* **E79-C**: 3–13 (1996).
8. F. Heismann, S. K. Korotky, and J. J. Veslka, Lithium niobate integrated optics: Selected contemporary devices and system applications, in *Optical Fiber Telecommunications*, Academic Press, New York.
9. R. Madabhushi, *High Speed Modulators for Coding and Encoding*, Short course, SPIE Photonics West, Int. Conf. Jan. 2001.
10. A. H. Gnauck et al., Dispersion penalty reduction using an optical modulator with adjustable chirp, *IEEE Photon. Technol. Lett.* **3**: 916–928 (1991).
11. S. K. Korotky et al., High-speed low-power optical modulator with adjustable chirp parameter, *Proc. Topical Meeting on Integrated Photonics Research*, Monterey, CA, paper TuG2, 1991.
12. M. Seino, N. Mekada, T. Namiki, and H. Nakajima, 33-GHz-cm broadband Ti : $LiNbO_3$ Mach-Zehnder modulator, *Proc. ECOC*, paper ThB22-5, 1989, pp. 433–435.
13. M. Rangaraj, T. Hosoi, and M. Kondo, A wide-band Ti : $LiNbO_3$ optical modulator with a conventional coplanar waveguide type electrode, *IEEE Photon. Technol. Lett.* **4**: 1020–1022 (1992).
14. G. K. Gopalakrishna et al., 40 GHz, low half-voltage Ti : $LiNbO_3$ intensity modulator, *Electron. Lett.* **28**: 826–827 (1992).
15. M. Seino et al., A low DC drift Ti : $LiNbO_3$ modulator assured over 15 years, *Proc. OFC'92*, Post Deadline papers, PD3, 1992.
16. D. W. Dolfi and T. R. Ranganath, 50 GHz velocity matched broad wavelength $LiNbO_3$ modulator with multimode active region, *Electron. Lett.* **28**: 1197–1198 (1992).
17. W. K. Burns, M. M. Hoverton, and R. P. Moeller, Performance and modeling of proton exchanged $LiTaO_3$ branching modulators, *J. Lightwave Technol.* **10**: 1403–1408 (1992).
18. K. Noguchi, O. Mitomi, K. Kawano, and M. Yanagibashi, Highly efficient 40-GHz bandwidth Ti : $LiNbO_3$ optical modulator employing ridge structure, *IEEE Photon. Technol. Lett.* **5**: 52–54 (1993).
19. S. K. Korotky and J. J. Vesalka, RC circuit model of long term Ti : $LiNbO_3$ bias stability, *Technical Digest Topical Meeting on Integrated Photonics Research*, San Francisco, paper FB3, 1994, pp. 187–189.
20. R. Madabhushi and T. Miyakawa, A wide band Ti : $LiNbO_3$ optical modulator with a novel low microwave attenuation CPW electrode structure, *Proc. IOOC'95*, Hong Kong, paper WD1-3, 1995.
21. R. Madabhushi, Y. Uematsu, and M. Kitamura, Wide-band Ti : $LiNbO_3$ optical modulators with reduced microwave attenuation, *Proc. IOOC'97/ECOC'97*, Tu1B, Edinburgh, UK, 1997.
22. S. K. Korotky et al., High-speed low-power optical modulator with adjustable chirp parameter, *Proc. of Topical Meeting on*

Integrated Photonics Research, Monterey, CA, paper TuG2, 1991.

23. R. Madabhushi, Y. Uematsu, K. Fukuchi, and A. Noda, Wideband Ti : $LiNbO_3$ optical modulators for 40 Gb/s applications, *Proc. ECOC'98*, Madrid, Spain, 1998, pp. 547–548.
24. S. Yamada and M. Minakata, DC drift Phenomenon in $LiNbO_3$ optical waveguide devices, *Jpn. J. Appl. Phys.* **20**: 733–737 (1981).
25. M. Seino, T. Nakazawa, M. Doi, and S. Taniguchi, The long term reliability estimation of Ti : $LiNbO_3$ modulator for DC drift, *Proc. IOOC'95*, Hong Kong, paper PD1-8, 1995, pp. 15–16.
26. I. Sawaki, H. Nakajima, M. Seino, and K. Asama, Thermally stabilized z-cut Ti : $LiNbO_3$ waveguide switch, *Proc. CLEO'86*, paper MF2, 1986, pp. 46–47.
27. T. Kambe et al., Highly reliable & high performance Ti : $LiNbO_3$ optical modulators, *Proc. LEOS'98*, Florida (USA), Orlando, paper ThI5, 1998, pp. 87–88.

OPTICAL MULTIPLEXING AND DEMULTIPLEXING

ALEXANDROS STAVDAS
National Technical University of Athens
Athens, Greece

1. INTRODUCTION

Optical multiplexing is a technique used in optical fiber networks for enhancing the capacity of point-to-point links as well as for simplifying the routing process within the optical layer. It is found in two forms. Borrowing the concept from its historic predecessor FDM, the optical domain equivalent [which is termed wavelength-division multiplexing (WDM)] is the predominant type of optical multiplexing for reasons to be explained in the following paragraphs. In WDM several information bearers, that is, optical carrier wavelengths, each modulated by a separate data pattern, are launched into (multiplexed) or decoupled from (demultiplexed) an optical fiber (Fig. 1a). The other technique first used in the PCM systems, namely, TDM, also has its optical equivalent, *optical time-division multiplexing* (OTDM). In OTDM, two (or more) pulses of equal energy from the same carrier wavelength are interleaved in time (Fig. 1b). To upgrade a system with bit rate B (pulse duration T) to a system with bit rate $2B$ (pulse duration $T/2$) using OTDM, the following steps are required: (1) generation of pulses with duration $T/2$ in time (at twice the initial power) and (2) delay, probably passive of one stream for $T/2$ with respect to the other before interleaving.

Wavelength multiplexing as a concept exists since the early days of the optical fiber revolution [1]. However, it emerged as a realistic solution only at the end of the 1980s, thanks to optical amplifiers. The advent of the erbium-doped fiber amplifier (EDFA) not only allowed viewing the optical fiber as a "lossless pipe" but also paved the way for collective power restoration of many

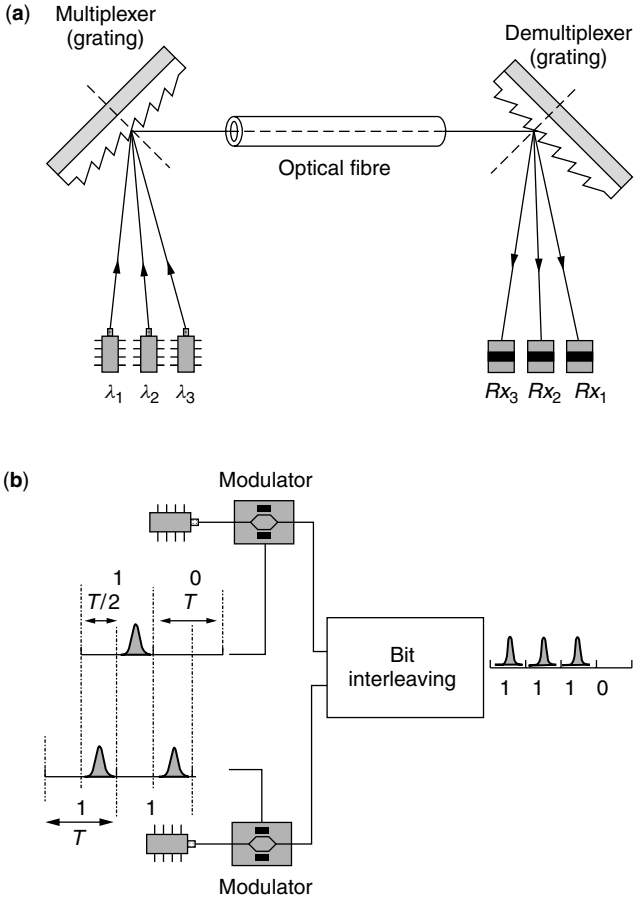


Figure 1. (**a**) Three wavelength channels are multiplexed, transmitted through an optical fiber, demultiplexed at the end and detected from the corresponding receivers (WDM); (**b**) to OTDM two bit streams, two pulses of half duration the initial time are derived (using active components) and a mutual time delay by a half time slot is introduced before interleaving (using passive components).

optical signals simultaneously. For comparison purposes it is mentioned that the first fiberoptic systems employed a single wavelength per fiber were expensive optoelectronic repeaters were used to compensate for the distortions (primarily power attenuation) due to transmission through the optical fiber. Capacity upgrade was achieved by deploying many single-wavelength fibers, something that increased the number of regenerators linearly. In contrast, the EDFA with the ability to restore the power of multiple wavelength channels when they are transmitted over the same single fiber made possible the replacement of these regenerators, leading to enormous cost savings. Wavelength multiplexing was first introduced in the field when it was realized that there is an economic incentive to upgrade from 2.5 to 10 Gbps (gigabits per second) using four wavelengths at 2.5 Gbps instead of one wavelength at 10 Gbps.

In addition, this form of optical multiplexing offers significant routing simplifications within the optical transport layer. Until the 1990s in all commercial telecommunications systems there were only electronic

switching fabrics, rendering electronic switching (and data processing) at the line rate mandatory. Given that the largest fraction of traffic at any node, like in a SDH ring, is transit traffic, processing of the entire traffic volume was becoming progressively more difficult, especially at increasingly higher bit rates. Adopting the principle of wavelength routing, where the final destination is uniquely identified by the wavelength (frequency) of the carrier, it is possible to isolate and process the information content of just the local traffic while the transit traffic, at a different carrier wavelength, could get through the node intact. In current commercial systems with capacity that is scaling up to Tbps (and even tens of Tbps in future systems), wavelength multiplexing is the indispensable technique for capacity upgrade.

On the component level there is a fundamental difference in the nature of the devices used for WDM and for OTDM. In OTDM, the pulses of two or more lower-speed sources are interleaved, generating a datastream with a speed equal to that of the aggregate rate. Thus, very fast "active" (i.e., electrically controlled) devices are needed. For optical processing of this stream in contrast, wavelength multiplexing makes use of "passive" devices. It is the passive nature of these devices that gives WDM all these desirable characteristics generally identified as "transparency": bit-rate independence as well as modulation format and protocol insensitivity. Because of its dominant role in real telecommunication systems, we will consider only wavelength multiplexing for the remaining part of this article.

2. OPTICAL (DE)MULTIPLEXING DEVICES

2.1. Physics of the Devices

When seen from the point of view of technical applications, the most important phenomena of light are *interference* and *diffraction*. Hence, the techniques used for optical (de)multiplexing (regardless of the form in which they appear) are primarily based on one of them. There is no satisfactory explanation of the difference between these two terms [2], but for any practical reason when two optical sources interfere, the result is called *interference*, while when there are a large number of them, the term *diffraction* is more appropriate. For optical (de)multiplexing purposes, the exploitation of two-beam interference is made through devices based on division of the amplitude of the incident beam before they are superposed again. Under this category are devices such as the Mach–Zehnder (MZI), Michelson (MMI), and Sagnac (SI) interferometers. An important family of (de)multiplexing devices are based on arrangements involving multiple divisions of the amplitude or multiple divisions of the wavefront of the incoming wave and they are classified as either (1) interference filters (Fabry–Perot interferometers, multilayer thin-film filters and fiber Bragg gratings) or (2) diffraction gratings (integrated optic, free-space or acoustooptic devices), respectively.

2.2. Functionality

The choice of the technology to be used strongly depends on the type of application under consideration. Hence, for low- to medium-capacity networks, that is, for up to 8-wavelength-channel WDM systems (with bit rates ranging from 644 Mbps to 10 Gbps per wavelength), all the aforementioned devices could be used indistinguishably (Fig. 2). When the total number of wavelength channels N is the predominant consideration for the choice of technology (in particular, when $N \geq 32$), the diffraction gratings are the primary candidates. Nevertheless, regardless of the technological platform, a higher wavelength channel count can be obtained by adding up groups of band-optimized devices. For example, (de)multiplexing devices with up to 60 channels are commercially available using interleaving of band-optimized interference filters in a parallel or cascadable configuration (Fig. 2c,d), while with band-optimized diffraction gratings, several hundred to thousands of channels could be produced.

In Fig. 2, the four different arrangements produce the same final result from a systems point of view. In Fig. 2b, the star coupler facilitates in distributing the same multiwavelength signal to all its N ports (each one will collect $1/N$ of the original optical power). Then a *thin-film* (or a *fiber-grating*) filter will select the requested wavelength. From a functionality point of view, the final outcome is the same as if a diffraction grating is used.* In any case, the diffraction grating-based devices are expected to dominate in the high-capacity systems and, therefore, will be dealt in more detail here. In Section 3 a more detailed comparison between the technological platforms will be provided.

2.3. Diffraction Gratings

2.3.1. Principle of Operation. A diffraction grating is any physical arrangement that is able to alter the phase (optical length) between two of its successive elements by a fixed amount. The impact of this progressive phase alteration becomes evident at the far-field intensity distribution. Consider the case of a plane reflection grating (Fig. 3). The incident plane wavefront PQ first reaches point A, which then becomes a source of secondary wavelets, and hence it advances point B. Finally the incident wavefront reaches B, which then becomes a source of secondary wavelets. These wavelets are exceeding those originating from A at the same time. Hence, the path difference from the corresponding points of the two neighboring grooves (spaced by d), as measured at a distant point of observation, is

$$AD - BC = d(\sin\alpha - \sin\beta) \tag{1}$$

If the incident beam is on the same side of the normal as the diffraction beam, then the sign in Eq. (1) should be replaced with a *plus*. For a more detailed presentation, the reader is referred to classic textbooks such as those by Born and Wolf [3] and Longhurst [4]. In any case it can be shown that the far-field intensity distribution of a planar diffraction grating is the same as that of N rectangular

* Nevertheless the performance in terms of crosstalk, losses [see below] might be different.

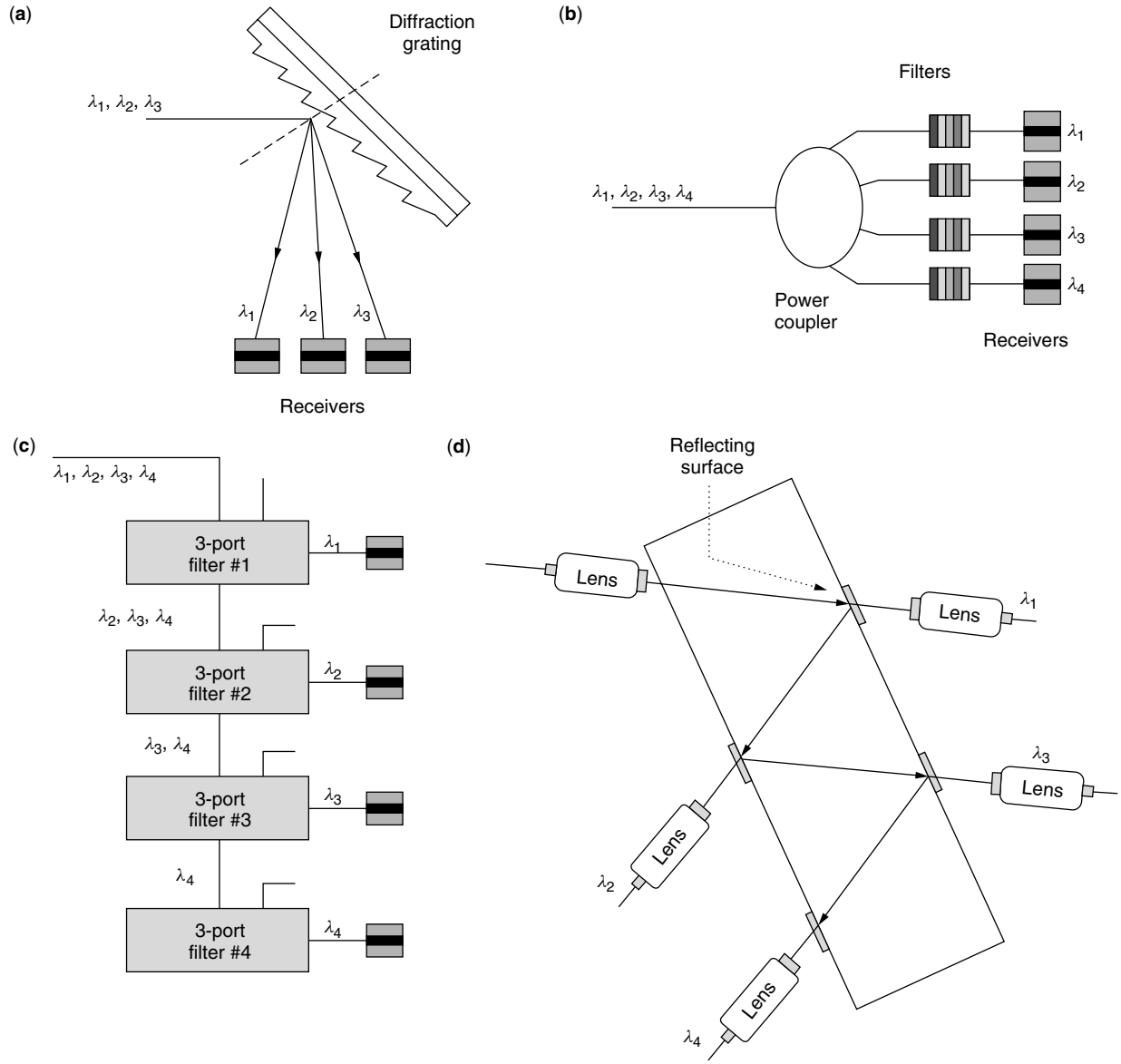


Figure 2. Four equivalent demultiplexing arrangements: (**a**) diffraction grating; (**b**) star coupler (for broadcasting) and fixed-wavelength filters; (**c**) a cascade (bus) of 3-port devices; (**d**) modified "cascade" configuration for two-port devices based on interference (thin-film) filters.

slits modulated by the diffraction envelope of a large slit. Constructive interference occurs when

$$d(\sin\alpha - \sin\beta) = m\lambda \tag{2}$$

where m is a constant called the *diffraction order* and λ is the wavelength (at free space) of the channel. The number N of the grooves/slits determines the sharpness of the principal maximum of the intensity distribution. For example, according to the Rayleigh criterion for the resolution limit, two equal-intensity wavelengths spaced by $\Delta\lambda$ are just resolved if the spatial distance between them is such that the two intensity distributions are crossing each other at 0.8 of their maximum value. The theoretical resolution limit for any diffraction grating is defined as

$$R = \frac{\lambda}{\Delta\lambda} = mN \tag{3}$$

2.3.2. Diffraction Grating Classification. The diffraction gratings could be either concave or planar, and they can operate either at a reflection or transmission mode. A planar diffraction grating — regardless the technological platform and the mode of operation — cannot be used as a standalone component when it is employed as a (de)multiplexer in optical communications. A practical (de)multiplexer based on a planar grating is always implemented in a spectrographic configuration employing two auxiliary optical components: one for collimating the incoming beam (i.e., for transforming the spherical wave to a plane wave) and a telescopic system for focusing the outgoing beam. The wavelength channels are diffracted at different angles determined by Eq. (2), and the telescopic system transforms this angle separation into a spatial separation at the image plane. This spatial

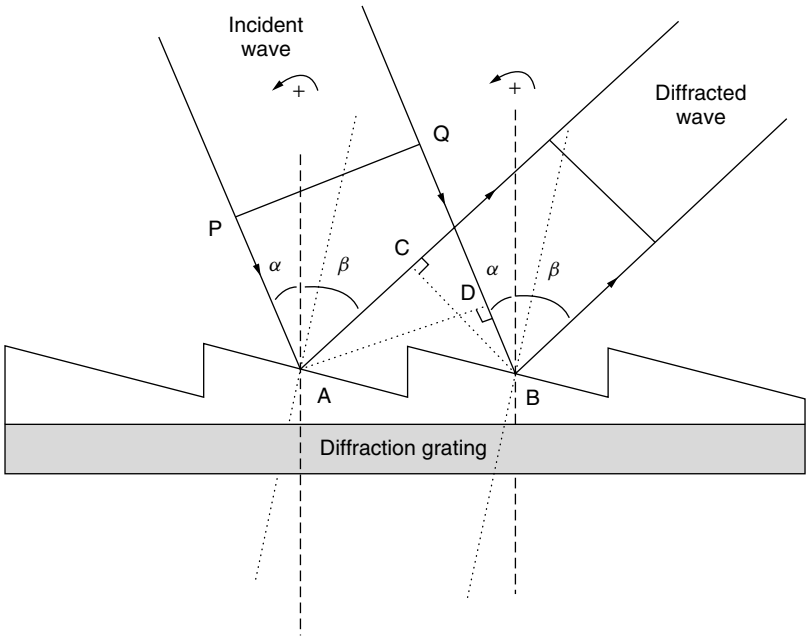


Figure 3. The principle of operation of a planar diffraction grating.

separation (Δx) between wavelength channels ($\Delta\lambda$) in the image plane is given by the reciprocal linear dispersion ($d\lambda/dx$ in nm/mm). Differentiation of Eq. (1) gives

$$\frac{d\lambda}{dx} = \frac{d\lambda}{fd\beta} = \frac{d\cos\beta}{mf} \tag{4}$$

where f is the focal length of the focusing part of the spectroscopic system. In practice, when the spectrograph is used as a (de)multiplexer, Δx is dictated by the minimum distance between the output waveguide/fiber cores.

2.3.3. Spectrograph Overview. The most important configurations for planar grating spectrographs are the *Ebert–Fastie* and the *Czerny–Turner* (Fig. 4a,b). In the former case, the spectrograph is constructed from a planar grating and a large concave mirror (or lens). The *Czerny–Turner* configuration offers the alternative of using two smaller concave mirrors instead of a single large one. The main drawback of these configurations is the use of the auxiliary optics off-axis, something that generates large aberrations. As a result, a point source is imaged as a geometric extended entity that degrades the performance of the optical system (Section 2.4).

A concave grating does not need auxiliary optics since it is a complete spectrograph. The corrugated surface provides the necessary diffraction for wavelength separation or recombination while the geometric properties of the concave surface allow focusing of the diffracted wavelengths. Since concave gratings operate off-axis, they also suffer from large geometric aberrations. However, there are specific geometric arrangements, called *focal curves*, which minimize the adverse effect of these aberrations. The best-known focal curve of concave gratings is the *Rowland circle* (Fig. 4c). For a concave substrate with radius of curvature R, the Rowland circle has a diameter R and is tangent to the apex of the substrate. The important characteristic of this geometric locus is that when a point source A is placed on it ($r_A = OA = R\cos\alpha$), an image free from second- and third- as well as reduced fourth-order Seidel meridional aberrations is produced at a location B on the Rowland circle ($r_B = OB = R\cos\beta$).

2.3.4. (De)multiplexer Performance Considerations. For assessing the quality of any (de)multiplexing device, a number of interrelated issues have to be considered. These include the spectral spacing between adjacent channels, the total number of wavelength channels, the passband flatness, the coupling losses, and the level of outband crosstalk. The best device is the one that allows the largest number of channels with the flattest bandpass, the smallest coupling losses per channel, and the largest optical isolation between adjacent channels with the minimum spectral separation between them.

From Eq. (3) it is concluded that the larger the size of the grating W (i.e., $W = Nd$), the sharper the intensity distribution is and the wider the spatial separation between two wavelength channels can be. Given that sufficiently large optical isolation is available between two adjacent channels, the number of grooves/slits should be considerably greater than that required for satisfying the Rayleigh criterion.

The *coupling loss* of any (de)multiplexer for a given wavelength channel is defined as the ratio of the incoming power to the outgoing power. This could vary with wavelength and depends on many parameters. For diffraction grating devices, these are the propagation material (free-space, Si, III–V semiconductor), the optical aberrations (that depend on the size of auxiliary optics for a planar grating and the clear aperture size for a concave grating), the mode mismatch between the device and the fiber (for integrated optic devices), and the type of the final receptor (e.g., detector, single-mode fiber, or multimode fiber with a clear aperture of 20, 10, and 50 μm, respectively).

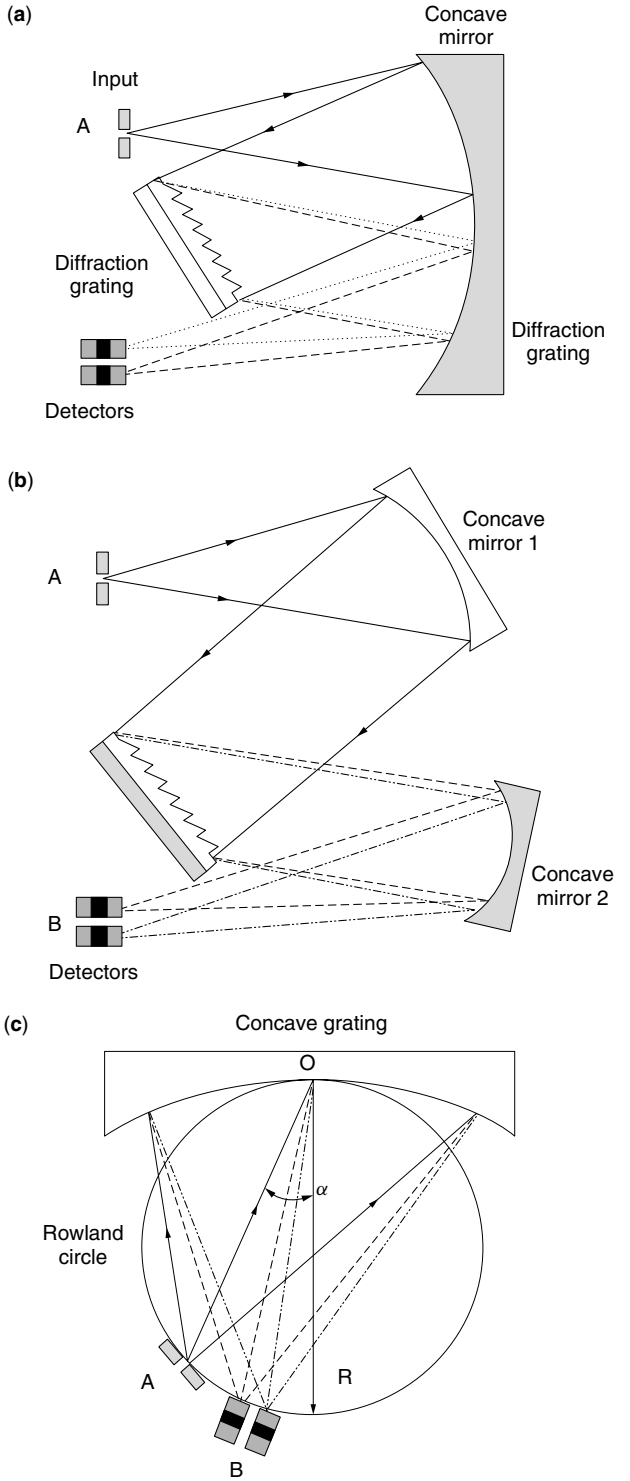


Figure 4. Spectrographs: (**a**) Ebert–Fastie; (**b**) Czerny–Turner; (**c**) Rowland circle.

A fraction of the optical power of a given wavelength that is not coupled to the corresponding outgoing receptor (detector, fiber) could be coupled to the adjacent-channel receptors generating *crosstalk*. In this way, crosstalk is the unintended coupling of signals from adjacent channels due to device imperfections. These phenomena can be better understood by considering the impulse response of any (diffraction grating) demultiplexer (Fig. 5). The impulse response has an intensity distribution with a Gaussian-like central part (due to the Gaussian intensity distribution emitted from a single-mode fiber) and a sinc-square function ($\sin^2 x/x^2$) distribution at the outer parts. As a practical rule of thumb, good optical isolation (low crosstalk) is achieved when the ratio of the optical signal bandwidth (measured at the $1/e^2$ point from its peak value) over the channel spacing is less than 0.25. The fact that the main part of the impulse response has a Gaussian intensity distribution profile leads to passband narrowing when many of these devices are cascaded. For this reason optical techniques are necessary in order to flatten the passband.

2.3.5. Practical Diffraction Grating Devices

2.3.5.1. Arrayed-Waveguide Grating (AWG). The most widely deployed (and studied) type of a grating-based (de)multiplexer is the arrayed-waveguide grating. This is a two-dimensional integrated-optic device (see Refs. 5,6, and references cited within). A special geometric arrangement of two slab waveguides and an array of single-mode waveguides forms a spectrographic setup based on a transmission grating.

The principle of operation, when the device is used as a demultiplexer, is as follows. The multiwavelength channel signal enters the device from the input slab waveguide (Fig. 6a), where it freely propagates. The input and output slab waveguides in most cases are constructed using the Rowland circle (Figs. 3c,6b). In principle, other geometric arrangements (generalized focal curves) are also possible. In any case, aberration-free focal curves are mandatory since the array of the single-mode waveguides is placed at the circumference of a spherical arc. The signal is coupled to the array of the waveguides probably via tapering for best coupling conditions. The array of waveguides plays the role of the grooves/slits in classic gratings [consider Eq. (3)]. Despite the use of the slabs on a Rowland circle, the entire setup has a planar grating configuration (a diffraction grating and two auxiliary optical systems). The length of the array waveguides is chosen such that the optical path length difference ΔL between adjacent waveguides is equal to an integer multiple of the central wavelength of the (de)multiplexer.

Because of the additional phase change introduced by the arrayed-waveguide length difference (that results in "hardwiring" all phases), the corresponding grating [Eq. (2)] is modified to

$$n_s d_0(\sin\alpha + \sin\beta) + n_c \Delta L = m\lambda \tag{5}$$

where n_s is the refractive index of the waveguide slab, n_c is the refractive index of the waveguides in the array (in the most general case, they are not the same), and d_0 is the distance between two successive waveguides in the array. To understand the reasons behind implementation of the AWG using this additional path length difference ΔL, one should consider the following.

Advances in integrated-optic fabrication techniques based on lithographic etching made possible the demonstration of AWGs on InP, Si/SiO_2, or $LiNbO_3$. Despite these

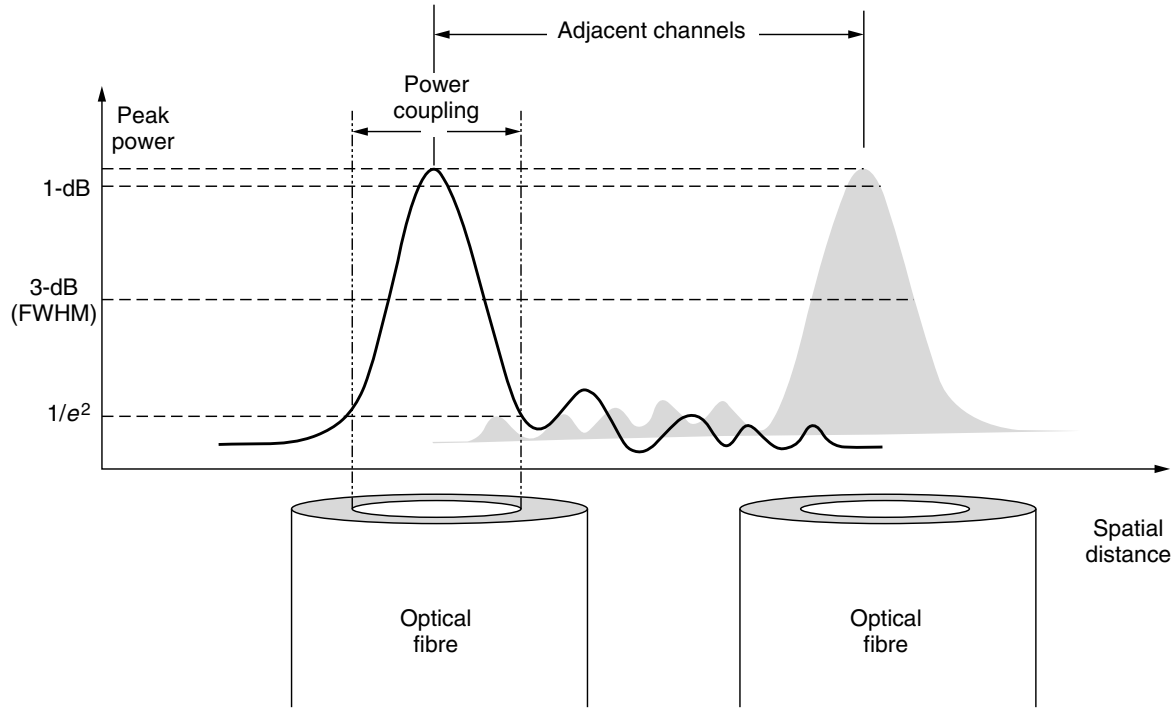


Figure 5. Coupling efficiency and crosstalk between adjacent channels. The part of the intensity not coupled to the destined output fiber smears with the wavelength signal destined to the adjacent fibers.

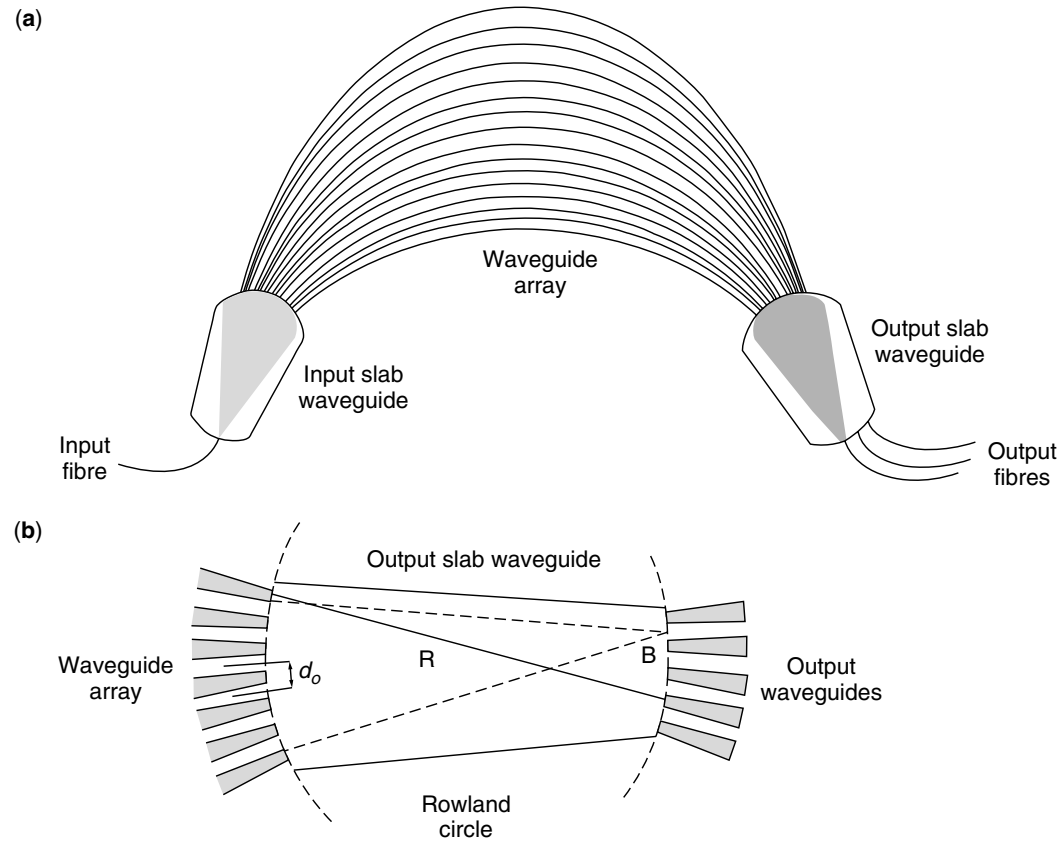


Figure 6. Schematic illustration of the arrayed-waveguide grating (AWG).

advances, there are limitations on the size of the wafers that could be practically used. Because of these size restrictions (resulting in small focal lengths), the required linear dispersion, Eq. (4), is obtained operating the AWG in high diffraction orders. From Eq. (5) the diffraction order equals to $\Delta L = m^{\mathrm{AWG}}\lambda/n_c$ and $m^{\mathrm{AWG}} + 1 = m$. However, the free spectral range (FSR) of the AWG is $\mathrm{FSR} = \Delta\lambda = \lambda/m$, and given that $50 < m < 100$, the FSR < 30 nm. This is one of the main limitations of this approach; specifically, the AWG can be used only in the context of limited spectral range. From an engineering point of view this problem can be solved, as mentioned earlier, by cascading coarse AWG demultiplexers followed by band-optimized fine-granularity AWGs. In this way, a 480-WDM-channel 10-GHz spaced (de)multiplexer has been reported, consisting of a WDM coupler and two 100-GHz-spaced AWGs, followed by 64 10-GHz-spaced AWGs [7]. The largest reported number of channels with a single AWG is 64 channels spaced by 50 GHz (0.4 nm) [8] or 128 channels spaced by 25 GHz [9], both developed on silica.

To improve the performance of the AWGs in terms of diffraction efficiency, modifications of known spectroscopic techniques have been applied [10,11]. The technique requires varying the optical path length (and hence the "hardwired" phase difference) between the waveguides in a nonuniform way, resulting in redistribution of the energy at the image plane. The effect is further improved after a suitable defocusing of the input/output waveguides. The former method is the integrated-optic equivalent of techniques used in aberration-corrected holographic concave gratings where the intensity distribution in the image plane is altered when the corresponding grooves are not equidistant parabolas.

Practical constraints in the fabrication of the AWGs are also attributed to the difficulty of the lithographic system to truly simulate a focal curve such as the Rowland circle. In addition, this mount requires the axis of the input/output waveguides to point toward the pole of the slab, implying that the waveguides have to be tilted with respect to each other. When this condition is not met, the consequent *vignetting* degrades the performance of the outer wavelength channels in the spectrum. Overall, the AWG is a very good candidate device for a demultiplexer operating within a single transmission band (like the *C* band of the EDFA), but it is problematic or impractical for wider optical bandwidth applications. Another issue, common to all integrated-optic devices, is the inherent birefringence of the materials used that leads to TE/TM polarization mode dispersion (PMD)-like problems. Also, temperature controllers are needed to thermally stabilize the operational conditions of the devices.

2.3.5.2. Other Integrated-Optic Spectrographs. Other practical integrated-optic spectrographs used as (de)multiplexers include a *two-dimensional concave grating* [12–15] and a modified *Czerny–Turner* configuration [16]. The former type has been implemented on both silica and III–V semiconductor compounds. Rowland circle or generalized focal curves have been used for producing aberration-free images [13]. The fabrication of this device, in contrast to the AWG, requires deeply etched grating facets, and this is achieved using ion-beam etching. A subsequent problem is the attainable degree of verticality of the grating wall. Another important consideration is associated with the rounding errors of the diffraction grating facets due to lithographic inaccuracies [14]. Again, because of the limited size of the wafers used, the requested linear dispersion [which in the current case is expressed as $d\lambda/dx = n_s d\cos\beta/(mf)$, where n_s is the refractive index of the slab] is achieved by operating the grating at high orders.

The difference between the 2D concave gratings and the AWGs is that in the arrayed-waveguide case the grating constant (pitch) equals to the distance between two successive waveguides. Given that the waveguide length is of the order of a millimeter, the waveguides need to be sufficiently apart to avoid exchange of energy between them (in Si-based devices the distance is at the order of 20 μm). This restriction does not apply to two-dimensional concave gratings. As a result, a grating pitch of few micrometers is feasible and the requested linear dispersion is attained operating the device at a lower order than the AWG (typically at 10th–30th order). Thus, these devices may operate in a wider spectral range compared to their AWG counterparts. With this technology, a device with 120 channels and 0.29-nm channel spacing has been reported [15]. A simple rule for identifying the tradeoff between grating pitch and diffraction order for the integrated optic concave gratings can be obtained by solving the equation for the linear dispersion and the grating equation that leads to

$$\frac{m}{d} = \frac{2\lambda n_s \sin\alpha + 2n_s\sqrt{(d\lambda/dx)^2 f^2 + \lambda^2 - (d\lambda/dx)^2 f^2 \sin^2\alpha}}{2((d\lambda/dx)^2 f^2 + \lambda^2)} \tag{6}$$

On the other hand, the *Czerny–Turner* mount consisting of a transmission grating and two parabolic mirrors used off-axis has been reported [16]. A paraboloid, although the spherical aberration, when it is operated off-axis. Introduces a significant amount of meridional coma. A *Czerny–Turner* spectrograph should be deployed using two spherical mirrors with different radii of curvature in order to compensate for meridional coma that degrade crosstalk.

In any case, both (de)multiplexer types manifest the same dependence for temperature control and compensation for the PMD-like dispersion due to material birefringence as the AWGs.

2.3.5.3. Free-Space Gratings. Practical free-space optical (de)multiplexers can be found in the form of either planar grating mounts or a holographic concave grating. Free-space grating multiplexers were the first to be tested in conjunction with WDM system experiments. The most established planar grating (de)multiplexer is implemented on the basis of a modification of the *Ebert–Fastie* configuration where the source and the image are almost collocated at the optical axis of a parabolic mirror. This is now a commercial product called STIMAX [17]. Operating an optical system on-axis results in an aberration-free image from second- and third-order Seidel aberrations. Fourth-order

Table 1. Performance of Diffraction Grating Devices

	Channel Spacing (nm)	Number of Channels	Losses (dB)	Crosstalk (dB)	Comments
AWG[a]	0.8	40	<6	< −20	Si/SiO_2
AWG[a]	0.8	≤40	<8	<−25 dB	1-dB band, ∼0.16 nm
AWG[a] flat passband	0.8	≤40	<9	<−24 dB	1-dB band, ∼0.32 nm
AWG [9][b]	0.2	128	3.5–5.9	<−16 dB	Si/SiO_2
Free-space planar[a]	0.8	≤40	<5.5	<−30 dB	1-dB band, >28 GHz
STIMAX[a]	0.8	≤64	<5.5	<−30 dB	—
Minilat[a]	0.8	≤92	<8	<−33 dB	1-dB band, ∼0.2 nm
Holographic concave[b] [24]	0.4	64	<8	<−30 dB	—
2D concave[b] [12]	1	50	16	<−19 dB	InP
2D concave[b] [15]	0.29	120	20–40	<−44 dB	Si-based

[a] Commercially available products.
[b] Laboratory results.

aberrations (spherical aberrations) do exist, and they are compensated by means of the parabolic mirror.

The STIMAX (de)multiplexer has a very high wavelength channel count (Table 1). The main limitation of this configuration is the rapid increase of third-order Seidel aberrations (coma) for the outer wavelength channels of the spectrum due to the use of the parabolic mirror off-axis. Mechanical and thermal stability issues have been successfully addressed. As a result of the free-space operation, PMD-like problems are inherently absent while the polarization dependence of the diffraction efficiency is an issue especially for wider bandwidth applications.

Another configuration uses a planar grating together with retroreflectors at the image plane that facilitate in producing a zero-dispersion focused light, reducing bandwidth narrowing due to cascaded (de)multiplexers [18]. In principle, this approach provides a bandwidth flattening technique that is still applicable even when the device is operated as a demultiplexer, something that is not the case with other solutions used in AWGs [19]. A variant of this technique is used to eliminate the polarization dependence of the diffraction efficiency [20].

Concave diffraction gratings are single-element optical systems that simultaneously provide dispersive and focusing properties. A single optical element is very attractive because of easier optical alignment and reduced packaging problems. For this reason concave gratings were used as (de)multiplexers even at the very early stages of WDM transmission [21]. However, it has been recognized that the main disadvantage of this optical system is the large inherent aberrations associated with the spherical concave substrate and, thus, aberration-corrected gratings need to be deployed [22].

The holographic concave gratings are the primary candidates for providing aberration corrected (de)multiplexers [23]. The principle of operation of the holographic concave grating is as follows (Fig. 7). The concave substrate is covered with a suitable photoresist material sensitive to a specific wavelength (e.g., Ar^+ laser). A laser beam is split and focused in two pinholes such that two coherent point sources (C and D) are created. In general, the resulting spherical waves interfere on the substrate, generating fringes that are neither equidistant nor straight. This is the recording phase. When the device is used as a demultiplexer, the single-mode fiber carrying a multiwavelength signal is placed at point A. Then a particular wavelength is imaged at point B, which, however, is not a point source, due to large optical aberrations. The aberrations can be eliminated by placing the two point sources C and D such as the generated fringes introduce a phase shift that cancel out the aberrations introduced by the spherical substrate. It has been demonstrated that up to 1000 wavelength channels can be (de)multiplexed using these devices, covering a spectral range of 200 nm [24]. Nevertheless, these devices are still available only in laboratories. As with all free-space devices, mechanical stability and polarization-dependent diffraction efficiency is an issue, and in particular when the number of wavelength channels are more than a hundred, the problem of an efficient fiber mount has to be tackled. The performance of all diffraction grating devices is summarized in Table 1.

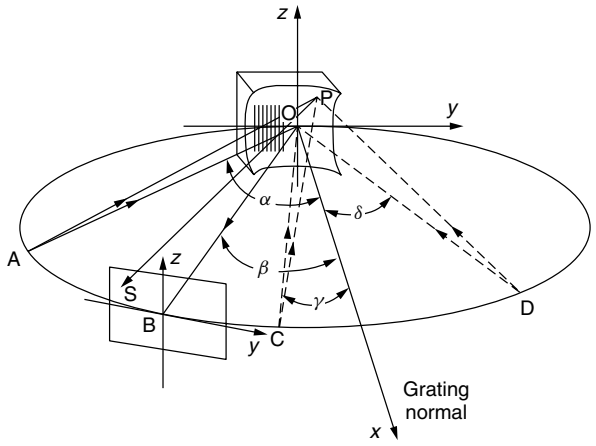


Figure 7. Schematic illustration of a holographic concave grating.

2.3.5.4. Acoustooptic Grating. This is an *active* device. The principle of operation of this device is based on the interaction of light with sound resulting in a transmission

grating. A sinusoidal sound wave travelling at the surface of an appropriate material generates periodic variations of the density (or strain) of the material according to the frequency of the wave. As a result, the macroscopic effect is a periodic change of the refractive index, and these periodic changes act as partially reflecting mirrors. Hence an incident plane wave, when specific conditions are met, will be diffracted at an angle according to its wavelength. The grating formed by the sound wave is a dynamic (time-varying) one.

The effect of the sound wave on the impinging plane wave can be understood in two ways. The distance between two "partially reflecting mirrors" depends on the frequency Ω of the sound wave. Conservation of energy and momentum require that $\omega_i = \omega_d + \Omega$, and $\mathbf{K}_\mathrm{i} = \mathbf{K}_d + \mathbf{K}$, respectively, where the index i indicates the incident wave while the index d the diffracted. $\mathbf{K}$ is the wavevector of the sound wave, and since $\Omega \ll \omega_i$ it is $\omega_i \cong \omega_d$. Thus, apart from a negligible frequency shift, the effect of the sound wave on a multiwavelength signal is to change the direction of propagation according to wavelength (demultiplexing) (Fig. 8a). Alternatively, it could be argued that an optical path length (phase) delay occurs between the two "partially reflecting mirrors" similar to what is produced by the grooves of a diffraction grating. When the distance between them satisfies the grating equation (which now is termed the *Bragg condition*), the elementary planes add constructively. The intensity distribution of the impulse response has the known sinc-square form and its sharpness will depend on the number of reflecting mirrors, namely, the total interaction length L (the sharpness will depend on the ratio L over the wave number determined by Ω).

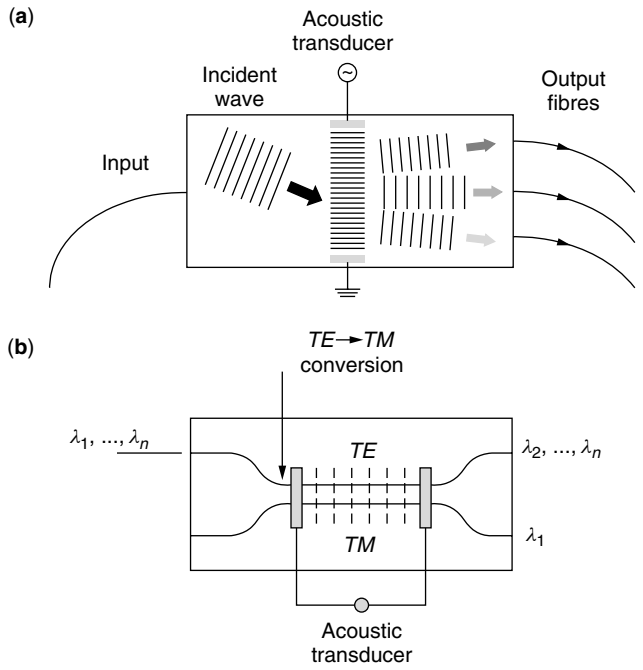


Figure 8. (**a**) The principle of operation of an acoustooptic grating; (**b**) a four-port acoustooptic filter.

2.4. Optical Filters

2.4.1. Acoustooptic Filters (AOFs). When an acoustic wave is applied on an acoustically active material, the induced birefringence (i.e., a dissimilar change of the refractive index for the ordinary and the extraordinary rays) of the medium alters the state of polarization of the incident wave. This principle is used for constructing acoustooptic filters (Fig. 8b). A multiwavelength signal enters a four-port device like a single coupler design. At the input stage, the incoming signal is 3-dB split by the directional coupler and the light at the lower branch also undergoes a phase delay of $\pi/2$. In other words, at the lower branch a polarization rotation is observed from the TE to the TM mode. When no acoustic wave is applied, another $\pi/2$ phase delay occurs at the output part and all channels exiting from the symmetric to the input port (e.g., upper-in, upper-out). When an acoustic wave is applied, the exact matching conditions are altered via the acoustooptic effect and the requested channel is selected from the lower output port. An interesting feature of the acoustooptic filter is that many acoustic frequencies could copropagate, allowing a simultaneous selection of more than one channel. Hence, the AOF can be used as a band-selecting filter in hierarchical (coarse/fine) WDM (de)multiplexing since it has a tunability of hundreds of nanometers. The crosstalk figure of the AOF is not as good as that of the other commercial diffraction gratings.

2.4.2. Interference Filters. These filters appear in the literature under many different names such as dielectric thin films, multilayer interferometric filters, and multistack thin-layer filters. Further, they are constructed from many different compounds ranging from liquid crystals to various oxides (SiO_2 or TiO_2) to multi-quantum-well (MQW) III–V semiconductors. Nevertheless, the principle of operation for all these structures is easily understood considering a Fabry–Perot etalon [25] that is an interferometer based on multiple divisions of the amplitude. Note that collimating optical devices [like bulk or GRID rod lenses] are mandatory at the input/output.

Let us assume that a material A with higher refractive index (n_H) compared to a material B (n_L) forms a cavity as shown in Fig. 9a. The reflected and transmitted intensities I_R and I_T of the Fabry–Perot etalon, respectively, normalized to the incident intensity, are given by

$$\frac{I_R}{I_i} = \frac{4R\sin^2(\delta/2)}{(1-R)^2 + 4R\sin^2(\delta/2)},$$
$$\frac{I_T}{I_i} = \frac{(1-R)^2}{(1-R)^2 + 4R\sin^2(\delta/2)} \tag{7}$$

where $\delta = (4\pi n_H \cos\theta_d L)/\lambda$ and R is the fraction of the intensity reflected at each interface like n_H/n_L and n_L/n_H .

Maximum power transfer to the reflected beam occurs when the thickness of the high-refractivity region is equal to a quarter-wavelength ($\delta/2 = m\pi/2$) while maximum transmission occurs when the thickness is one half-wavelength thick ($\delta/2 = m\pi$), assuming that $R \cong 1$. The former case is depicted in Fig. 9b. Many

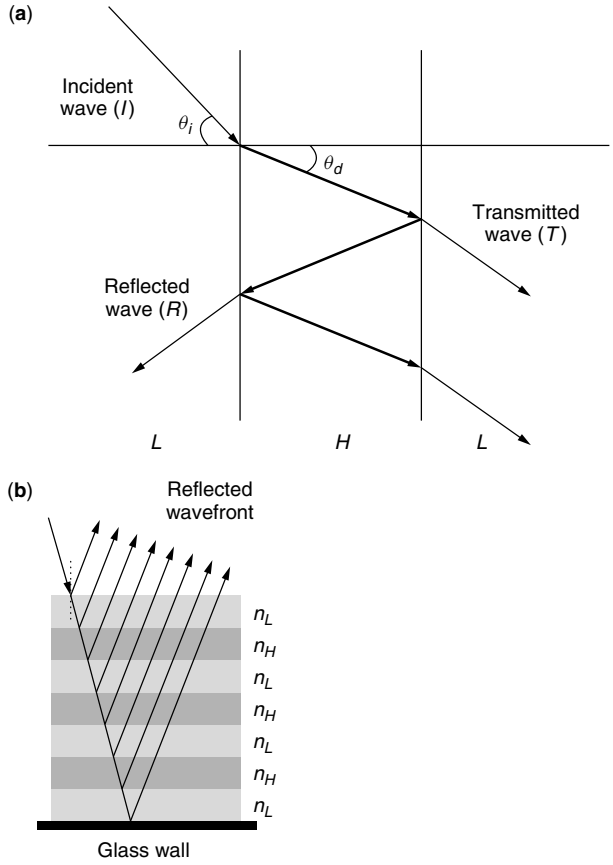


Figure 9. (**a**) A Fabry–Perot etalon; (**b**) a multilayer stack operated on a reflection mode.

different configurations could emerge by adding up such multistacks; for example, two structures as shown in Fig. 9b with a space layer between them form an additional Fabry–Perot cavity of the type HLH(2L)HLH, assuming that each layer has a quarter-wavelength thickness. It can be shown that sharper cutoff characteristics are obtained by increasing the number of layers or cavities. However, when all cavity lengths are the same, the overall structure produces a narrower passband. Hence, layers of unequal thickness are used, something that requires the addition of further layers for phase matching, leading to a complex optimization problem. Also, the passband increases with increasing values of n_H/n_L [26]. Overall, practical constructions lead to filters offering optical isolation up to -30 dB. The losses are range between 1 and 5 dB depending on the material, the number of channels, and other factors.

It is pointed out that these are the only filter devices that could be implemented using all the three configurations of Fig. 3b–d. Nevertheless, the reader should be aware that should the configurations of Fig. 3c,d be used in a practical system, power equalization techniques should be employed for compensating the loss variation (which could be a problem when these devices are cascaded if the number of channels is high).

2.4.3. Mach–Zehnder Filters. These filters are four-port devices that can be either *passive* or *active*. The

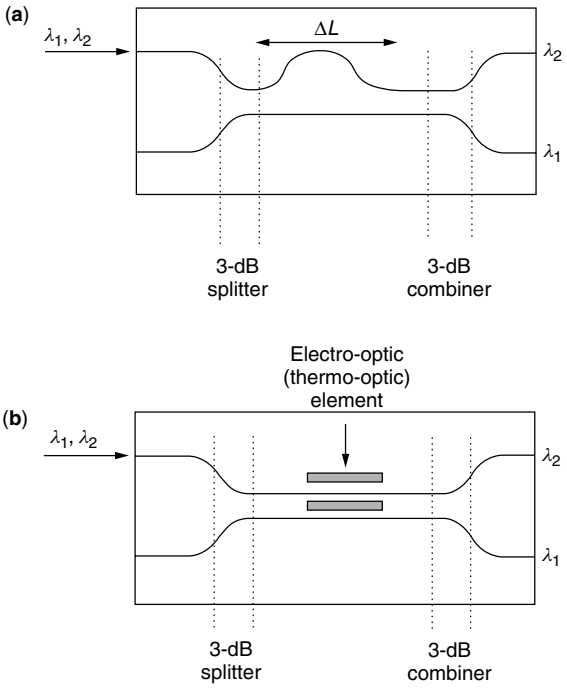


Figure 10. (**a**) The asymmetric Mach–Zehnder and (**b**) the symmetric configurations.

integrated-optic version of the Mach–Zehnder interferometer, illustrated in Fig. 10, consists of three parts. The input and the output parts are 3-dB couplers, while the central section has two waveguide arms (upper and lower) with different path lengths. The configuration is called *asymmetric*. For a signal entering the upper port, the overall phase difference due to the asymmetric length is

$$\Delta\phi = \frac{2\pi n_u}{\lambda}(L + \Delta L) - \frac{2\pi n_l}{\lambda}L = \frac{2\pi n}{\lambda}\Delta L \qquad (8)$$

where n_u and n_l are the refractive indices of the upper and lower waveguides, respectively. In this case they are assumed to be equal to n. The upper ports are labeled as 1 and $1'$ at the input/output, respectively, and likewise the lower ports as 2 and $2'$. The transmittance from port 1 to $1'$ is given by T_u and from 1 to $2'$, by T_l.

$$T_u = \sin^2(\Delta\phi), \quad T_l = \cos^2(\Delta\phi) \qquad (9)$$

when $(2\pi n/\lambda_1)\Delta L = (m + 1)\pi/2$ and $(2\pi n/\lambda_2)\Delta L = m\pi$, with m an integer, then T_u is unity for λ_1 and zero for λ_2 and vice versa (T_l is unity for λ_2 and zero for λ_1). In this way, the wavelengths λ_1 an λ_2 are collected from different ports. When the device is carefully designed, it can operate as a wavelength channel (de)interleaver. The *symmetric* configuration (Fig. 10b) has waveguide arms of equal length, so the phase difference occurs as a result of electrooptically or thermooptically [27] induced change of the refractive index, resulting in different n_u and n_l when the control signal is on.

The loss figure depends on the number of wavelengths that can be demultiplexed from a single module, as well as the host material (Si or III–IV semiconductor). In general,

the optical isolation between adjacent channels is not better than −30 dB.

3. OVERALL ASSESSMENT OF (DE)MULTIPLEXING TECHNIQUES

Having presented the main technological platforms currently used for optical (de)multiplexing, it would be interesting to highlight their pros and cons. As pointed out earlier, the main question when a technology is assessed is the type of application in mind. In general, the configurations illustrated in Fig. 2b,c have different drawbacks.

A layout like the one in Fig. 2b, based on a power coupler and optical filters, is a flexible solution up to approximately 8 wavelength channels. Beyond this point, splitting losses tend to be high that the grating solution is advised. Optical filters with tailormade spectral characteristics (e.g., through wavelength) can be used in this configuration leading to a (de)multiplexer construction, offering a wavelength comb with unequal channel spacing allocation (to combat, e.g., fiber nonlinearities such as four-wave mixing). Nevertheless, when systems with a large number of wavelength channels are desired, the cost of the system scales proportionally to channel count.

The "bus" architecture of Fig. 2c is implemented only via three-port or four-port devices. Thus, the loss performance is not uniform across the spectrum of interest; the first channel has the lowest losses while the final channel suffers from the worst losses. It is this loss figure that determines the maximum number of channels to be used per band. In general, the performance of the optical filters is good in terms of optical isolation. The loss performance of the device itself is good, but for practical applications a cascade of other optical components, such as band (de)multiplexers, is required.

With diffraction gratings the cost is not proportional to channel count and, indeed, devices with a very large number of channels have been demonstrated. AWG suffers from polarization-induced phenomena, while free-space gratings do not. Free-space gratings offer perhaps the best optical isolation from all (de)multiplexing devices and have no restrictions with respect to the total optical bandwidth they can handle. In principle, the free-space devices can explore the parallelism of optics to generate many (de)multiplexers in parallel or to be used in conjunction with other free-space devices such as microelectromechanical switches (MEMSs) in optical cross-connects.

BIOGRAPHY

Alexandros Stavdas received his B.Sc. in physics from the University of Athens, Greece, in 1988, his M.Sc. in optoelectronics and laser devices from Heriot-Watt University/St-Andrews University in 1990, and his Ph.D. from the University College of London, United Kingdom, in 1995 (supervisor, Professor J.E. Midwinter) in the field of wavelength routed WDM networks. He worked in the design of free-space and integrated optics demultiplexers, wavelength cross-connects and on issues related to optical switching and wavelength routing systems. In the past he worked on the ACTS COBNET Project on alternative ring architectures and in design considerations and scalability of WDM rings, and on ACTS PLANET in the area of generic architectures for the WDM upgrade of SuperPONs. Currently, he is leading the project ULTRA funded by Nortel Networks on ultra-wideband DWDM systems and he is the technical leader for NTUA on the IST-DAVID project dealing with packet-over-WDM in Metropolitan Networks. He served as chairman of the Optical Network Design and Modelling Conference (ONDM 2000). Current interests include physical layer modeling of optical networks, ultra-high capacity end-to-end optical networks, OXC architectures, WDM access networks, and optical packet switching.

BIBLIOGRAPHY

1. C. Koester, Wavelength multiplexing in Fiber optics, *J. Opt. Soc. Am.* **58**(1): 63–67 (1968).
2. R. Feynman, *Lectures in Physics*, Addison-Wesley, Reading, MA, 1983.
3. M. Born and E. Wolf, *Principles of Optics*, 6th ed., Pergamon Press, 1980.
4. R. Longhurst, *Geometrical and Physical Optics*, 3rd ed., Longman, 1986.
5. H. Takahasi, S. Suzuki, K. Kato, and I. Nishi, Arrayed waveguide grating for wavelength division multi/demultiplexing with nanometer resolution, *Electron. Lett.* **26**(2): 87–88 (1990).
6. M. Smit and C. van Dam, Phasar-based WDM-devices: Principles design and applications, *IEEE J. Select. Top. Quant. Electron.* **2**(2): 236–250 (1996); also, M. Smit, *Electron. Lett.* **24**(7): 385–386 (1988).
7. K. Takada, H. Yamada, and K. Okamoto, 480 channel 10 GHz spaced multi/demultiplexer, *Electron. Lett.* **35**(22): 1964–1966 (1999).
8. K. Okamoto, K. Moriwaki, and Y. Ohmori, Fabrication of a 64×64 arrayed-waveguide grating multiplexer on Si, *Electron. Lett.* **31**(3): 184–186 (1995).
9. K. Okamoto, K. Syuto, H. Takahashi, and Y. Ohmori, Fabrication of 128-channel arrayed-waveguide grating multiplexer with 25 GHz channel spacing, *Electron. Lett.* **32**(16): 1474–1476 (1996).
10. C. Doerr and C. H. Joyner, Double-chirping of the waveguide grating router, *IEEE Photon. Technol. Lett.* **9**(6): 776–778 (1997).
11. C. Doerr, M. Shirasaki, and C. H. Joyner, Chromatic focal plane displacement in the waveguide grating router, *IEEE Photon. Technol. Lett.* **9**(6): 776–778 (1997).
12. J. Soole et al., Monolithic InP-based grating spectrometer for wavelength-division multiplexed systems at 1.5 μm, *Electron. Lett.* **27**(2): 132–134 (1991).
13. K. McGreer, A flat-field broadband spectrograph design, *IEEE Photon. Technol. Lett.* **7**(4): 397–399 (1995).
14. R. Deri, J. Kallman, and S. Dijaili, Quantitative analysis of integrated optic waveguide spectrometers, *IEEE Photon. Technol. Lett.* **6**(2): 242–244 (1994).
15. Z. Sun, K. McGreer, and J. Broughton, Demultiplexing with 120 channels and 0.29-nm channel spacing, *IEEE Photon. Technol. Lett.* **10**(1): 90–92 (1998).

16. M. Gibbon et al., Optical performance of integrated 1.5 μm grating wavelength multiplexer on InP-based waveguide, *Electron. Lett.* **25**(16): 1441–1442 (1989).

17. *http://www.highwave-tech.com/products/*.

18. I. Nishi, T. Oguchi, and K. Kato, Broad passband multi/demultiplexer for multimode fibres using a diffraction grating with retroreflectors, *IEEE J. Lightwave Technol.* **LT-5**(12): 1695–1700 (1987).

19. J. B. Soole et al., Use of multimode interference couplers to broaden the passband of wavelength dispersive integrated WDM filters, *IEEE Photon. Technol. Lett.* **8**(10): 1340–1342 (1996).

20. *http://www.photonetics.com*.

21. R. Watanabe, K. Nosu, T. Harada, and T. Kita, Optical demultiplexer using concave grating in 0.7–0.9 μm wavelength region, *Electron. Lett.* **16**(3): 106–108 (1980).

22. T. Kita and T. Harada, Use of aberration corrected concave gratings in optical multiplexers, *Appl. Opt.* **22**(6): 819–825 (1983).

23. A. Stavdas, P. Bayvel, and J. E. Midwinter, Design and performance of concave holographic gratings for applications as multiplexers/demultiplexers for wavelength routed optical networks, *Opt. Eng.* (*SPIE*) **35**: 2816–2823 (1996).

24. A. Stavdas et al., The design of a free-space multi/demultiplexers for ultra-wideband WDM networks, *IEEE J. Lightwave Technol.* **19**(11): 1777–1784 (Nov. 2001); also, A. Stavda, *Design of Multiplexer/Demultiplexer for Dense WDM Wavelength Routed Optical Networks*, Ph.D. thesis, Univ. London, 1995.

25. A. Yariv, *Optical Electronics*, HRW International Edition, 3rd ed., 1985.

26. E. Hecht, *Optics*, 2nd ed., Addison-Wesley, Reading, MA, 1987.

27. B. J. Offrein et al., Wavelength tunable optical add-after-drop filter with flat passband for WDM networks, *IEEE Photon. Technol. Lett.* **11**(2): 239–241 (1999).

OPTICAL SIGNAL REGENERATION

GEERT MORTHIER
JAN DE MERLIER
Ghent University — IMEC
Ghent, Belgium

1. INTRODUCTION

It is well known that signals are significantly distorted when they propagate through optical fiber networks (see OPTICAL FIBER COMMUNICATIONS). In modern WDM communications, many channels at different wavelengths and each carrying signals of 10 or even 40 Gbps (gigabits per second) are sent through long stretches of fiber and are traversing several optical amplifiers and optical switches. The propagation of large-bandwidth signals through dispersive optical fiber gives rise to pulse broadening and distortion. The optical amplifiers add noise to the signals and the optical switches can give rise to crosstalk. In large optical networks, the resulting *deterioration* of the signals due to dispersion, noise, and crosstalk may be so large that errorless transmission is no longer possible unless the signals are regenerated at intermediate nodes. This is illustrated in Fig. 1, which schematically shows the increase in bit error rate (BER) as a function of the distance in an optical transmission link with and without *regeneration*.

Regeneration can be achieved by converting the optical signals into electronic signals (using optical receivers), regenerating the electronic signals, and by retransmitting the signals optically (i.e., using a laser). However, because of the involvement of power electronics, this *optoelectronic regeneration* becomes more and more costly as the bit rate increases. All-optical regeneration becomes a very interesting alternative to optoelectronic regeneration in this case. The fact that *all-optical regeneration* can in principle be extended to multichannel regeneration is, for WDM systems, especially attractive.

Different types of optical regenerators, based on fiber loops or on *photonic integrated circuits* (PICs), have been proposed so far in the literature. The regenerators based on PICs are in general much more stable than the ones based on fiber loops (which require, e.g., *polarization control*). The latter can be operated at higher speeds though. Fiber loop regenerators and high-speed regenerators in general are usually aiming at *return-to-zero* (*RZ*) *signals*. Regeneration of *non-return-to-zero* (*NRZ*) *signals* is often limited to lower bit rates.

2. FIGURES OF MERIT

One typically classifies regeneration into 1R, 2R, and 3R regeneration; 1R consists of simple reamplification, 2R includes *reamplification* and *reshaping*, and 3R includes reamplification, reshaping, and *retiming*. Figure 2 illustrates 3R regeneration. It is generally believed that 2R regeneration may suffice at bit rates below 10 Gbps. From 10 Gbps on, optical regeneration could consist of 3R regeneration at selected nodes and 2R regeneration at other nodes. At these very high bit rates, *timing jitter* or, in other words, fluctuations in the duration of the individual bits (see OPTICAL NOISE) can be a serious cause of degradation and must be alleviated by retiming techniques.

A number of properties determine the performance of regenerators. The quality of the *regeneration* itself is characterized by the extinction ratio improvement

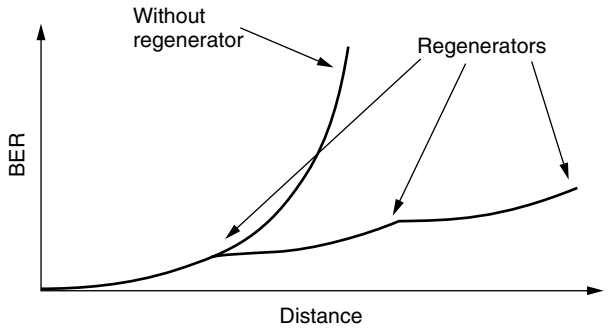


Figure 1. Increase of BER versus link distance with and without regeneration.

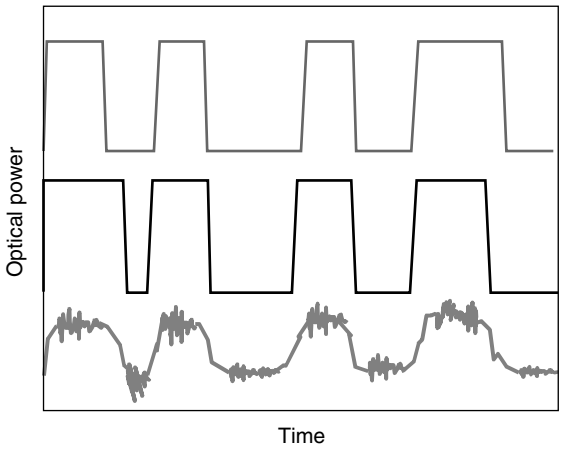


Figure 2. Schematic illustration of 3R optical regeneration; lower trace — signal before regeneration; middle trace — signal after 2R regeneration; upper trace — signal after 3R regeneration.

(see OPTICAL MODULATION) and by the *noise reduction* (determined by the flatness of the output vs. input power characteristic for both spaces and marks). An ideal 2R regeneration or *decision characteristic* as well as a more realistic characteristic are shown in Fig. 3. The *extinction ratio improvement* can be defined as $(P_{2,\text{out}}/P_{1,\text{out}})/(P_{2,\text{in}}/P_{1,\text{in}})$. In addition, the maximum bit rate should be as high as possible, while the required input power levels to the regenerator should be low or modest. The minimum required input power (for a digital one) is related to the decision threshold.

Finally, the ability to use the same device for regeneration over a broad wavelength range, even if one or more current settings have to be changed, is considered as a major advantage for WDM applications.

3. 2R REGENERATION

In order to obtain optical 2R regeneration, the nonlinear effects in optical waveguides (see OPTICAL WAVEGUIDES) are exploited. The incident light is used to influence the refractive index of the material through which it propagates. This happens via the interaction with the charge carriers in semiconductor waveguides. The refractive index variation in turn affects the propagation delay in the waveguides or, in other words, the phase of

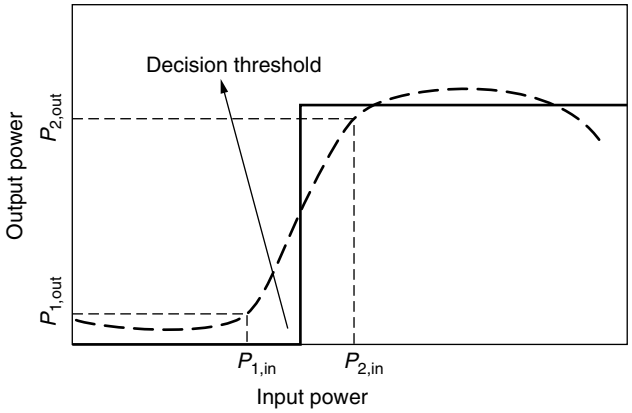


Figure 3. Ideal (full line) and realistic (dashed line) decision characteristic.

the optical field at the other end of the waveguide. The resulting phase changes can cause drastic changes in the output of a component when the waveguides are used in an *interferometric layout*.

3.1. SOA-Based Regenerators

The most promising results, in terms of regeneration, have been achieved with interferometric structures, such as the *Mach–Zehnder interferometer* (MZI) or Michelson interferometer (MI), containing a *semiconductor optical amplifier* (SOA) (see OPTICAL AMPLIFIERS) in each arm. With these interferometric structures, shown in Fig. 4, regeneration can be obtained through simultaneous *wavelength conversion* (see OPTICAL FREQUENCY CONVERSION) or without wavelength conversion, using a passthrough scheme.

When regeneration is performed using a passthrough scheme, the currents in both SOAs are chosen such that destructive interference is obtained at low input powers. Because of the difference in current, the saturation behavior of both SOAs will also differ. As a result, the phase difference between both arms will be modulated by the power variation in the data signal. This phase modulation is converted to an amplitude modulation at the output of the interferometer. Regeneration (although mainly for logical '1s') without wavelength conversion has been demonstrated at 40 Gbps using the Mach–Zehnder interferometer. When using a *Michelson interferometer*, input and regenerated signal have to be separated using a circulator.

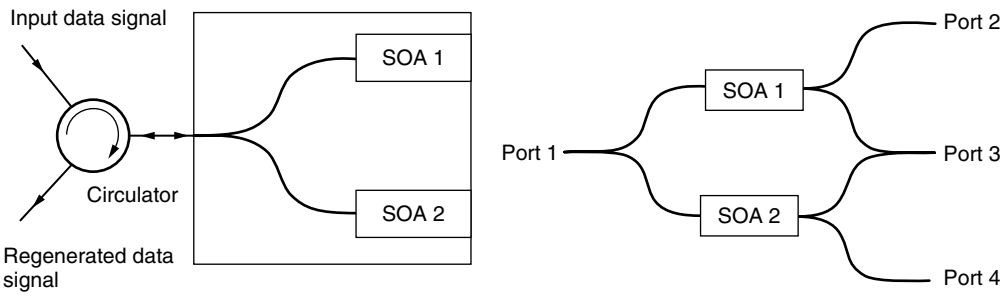


Figure 4. Michelson interferometer–based regenerator (left) and Mach–Zehnder interferometer–based regenerator (right).

The highest speed can theoretically be achieved with Mach–Zehnder interferometers and with simultaneous wavelength conversion. Wavelength conversion in the codirectional scheme (Fig. 4) is obtained by injecting a CW signal in port 3 and injecting the data signal in port 2. The electron density in SOA 1 is then modulated by the data signal. This causes a change in the phase difference between the arms, which causes modulation of the power in the CW signal at the output of the MZI. The CW signal can also be coupled into port 1, in a *counterdirectional scheme*. This allows the in- and output wavelengths to be the same and avoids the need for a filter at the output. Because of the sinusoidal dependence of the output on the phase difference between the arms, the polarity of the converted signal can be maintained or converted as compared to the incident data signal.

This scheme can be used at bit rates up to 40 Gbps for RZ signals. For the very high bit rates, however, a RZ input signal is converted to a NRZ signal [1], due to the finite lifetime of the electron density. Therefore other schemes are used at these very high bit rates.

Differential delay techniques make specific use of the fact that the interferometer output depends on the phase difference. In this scheme (Fig. 5), the data signals are injected in the SOAs in both arms of the MZI with a small delay between both pulses. The induced phase change in the arm where the data signal is injected first (e.g., ϕ_1), can be canceled out once the other pulse is injected into the other arm (giving a phase change ϕ_2). This differential control scheme has been shown feasible at bit rates of 40 Gbps. A filter is obviously required at the output to suppress the signal at λ_1.

Other SOA-based regenerators have been proposed on the basis of the nonlinear properties of, e.g., an active *multimode interference* (MMI) *coupler* and an active directional coupler. Both devices have been verified only by static measurements of the transfer characteristic, but for this static regime (and hence also for the low-bit-rate regime) they exhibited improved regeneration characteristics as compared to the MZI- and MI- based regenerators [e.g., 2].

3.2. Nonlinear Optical Loop Mirrors

Optical regenerators based on the *nonlinear optical loop mirror* (NOLM) make use of the ultrafast but weak Kerr nonlinearity (Fig. 6; see NONLINEAR EFFECTS IN FIBERS). This layout has the disadvantage that very high powers and several hundreds of meters of dispersion-shifted fiber (DSF) are required. The data signal is coupled in and out of the NOLM using a wavelength-dependent coupler (WDC). A CW-signal enters the NOLM through port 1 of a 50–50 coupler. Therefore, 50% of the CW signal propagates clockwise and 50% of it propagates counterclockwise in the ring. If a data pulse is present, the CW signal copropagating with the pulse experiences a nonlinear phase shift that is proportional to the data pulse intensity. At the output of port 2, one obtains interference between the phase-modulated clockwise and the counterclockwise propagating CW light. The light at the CW wavelength that is coupled out of the ring at port 2 is therefore a regenerated version of the data signal. The regenerative capabilities of the NOLM have been demonstrated at 80 Gbps, but this NOLM-based regenerator should be capable of regeneration at bit rates well over 100 Gbps [3].

3.3. Electroabsorption Modulators

Nonlinear behavior in a reverse-biased *electroabsorption modulator* (EAM) is achieved by using an intense input optical pulse to produce a large number of photogenerated charged carriers in the highly absorptive waveguide. Drift and diffusion of these carriers distort the electric field and cause a reduction of the field. As a result of the reduced electric field, the absorption decreases and a

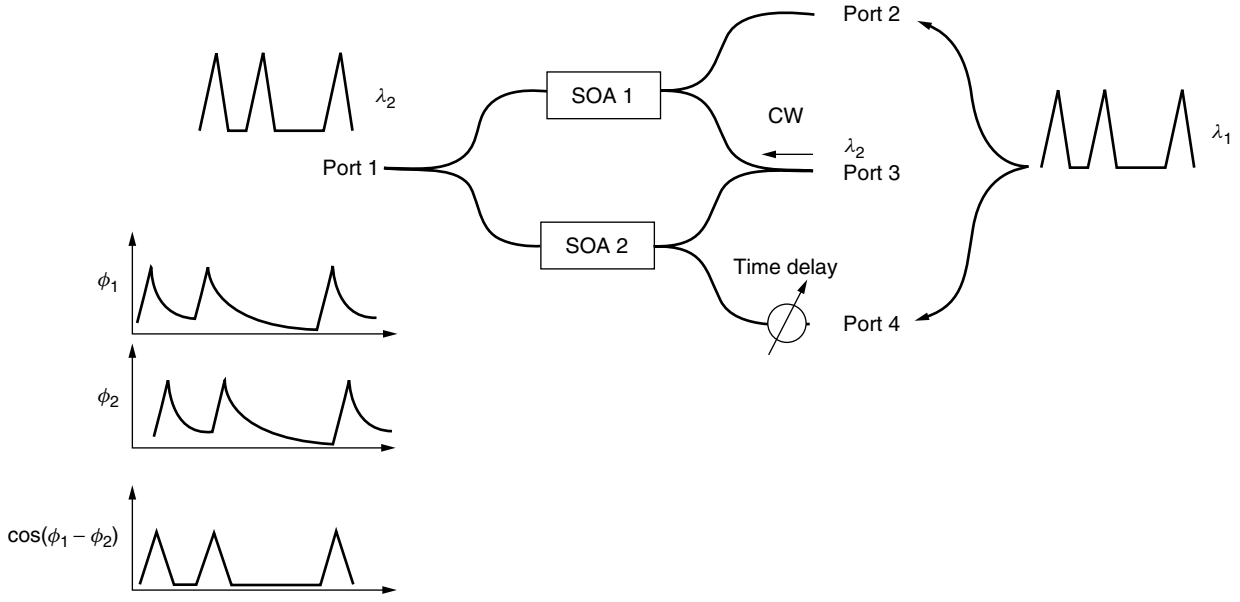


Figure 5. Differential delay scheme for a MZI-based regenerator.

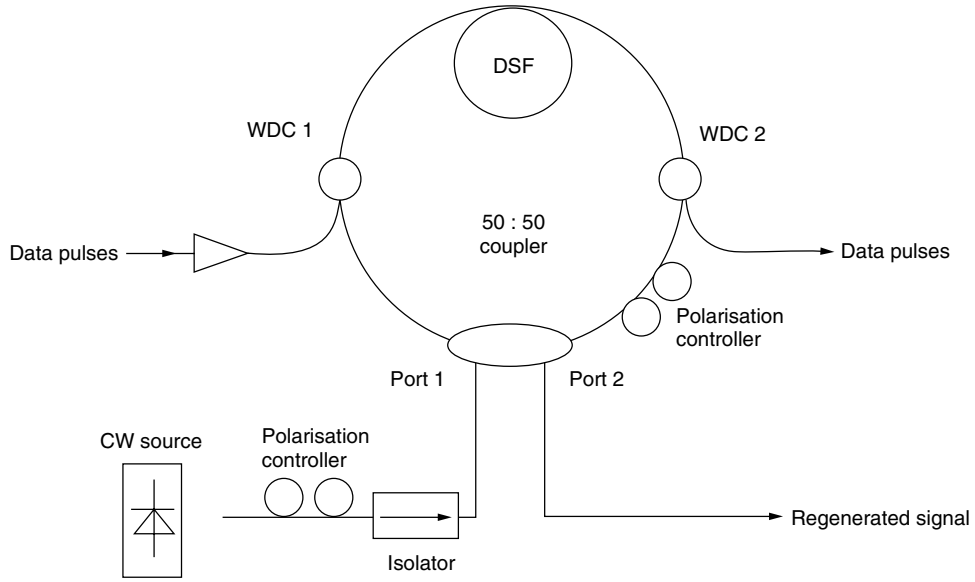


Figure 6. Scheme of nonlinear optical loop mirror (NOLM)-based regenerator.

transmission window is created for the pulse. When an intense optical pulse propagates through the EAM, a CW probe signal traversing the waveguide experiences the transient increase in transmission followed by a fast resumption of the absorption. As a result, a short pulse is generated at the wavelength of the probe signal. Weak input signals are absorbed without changing the transmission state of the EAM. This type has been demonstrated at bit rates of 40 Gbps.

3.4. Q-Switched Lasers

The *Q-switched laser* consists of three sections: two Bragg sections and one integrated phase tuning section Fig. 7. The second Bragg section is biased near transparency and is used as a dispersive reflector. The different sections are designed and biased such that lasing is achieved in the laser section only when a matched feedback is obtained from the reflector section.

Injection of a high-power data pulse changes the refractive index in the reflector section, which, in turn, causes a shift in the reflection band away from the lasing wavelength (change of Q factor), stops the required feedback and eventually ends the lasing in the laser section. Lasing starts again when the power in the signal falls again. Optimization of this scheme has led to successful tests up to 10 Gbps.

4. OPTICAL CLOCK EXTRACTION AND 3R REGENERATION

3R regenerators consist of 2R regenerators in combination with *optical clock extraction*. The periodic optical clock is then typically modulated in a more or less digital manner by the 2R-regenerated signal. Optical clock extraction itself is generally based on the use of self-pulsating laser diodes of which the self-pulsation frequency locks to the bitrate of the incoming signal. The self-pulsating laser diode can be either a *mode-locked laser* [4] or a multisection DFB laser [5]. Clock recovery up to 40 Gbps has been demonstrated with both types of devices.

The scheme for clock extraction based on a multisection DFB laser is the same as the scheme of Fig. 7, except that the regenerated signal is replaced by a clock signal. The DFB laser (see OPTICAL TRANSMITTERS AND RECEIVERS) now has to be designed such that lasing occurs in a self-pulsating regime and at a wavelength far away from the wavelength of the injected signal. The self-pulsation frequency of the free-running laser is determined by the injected currents

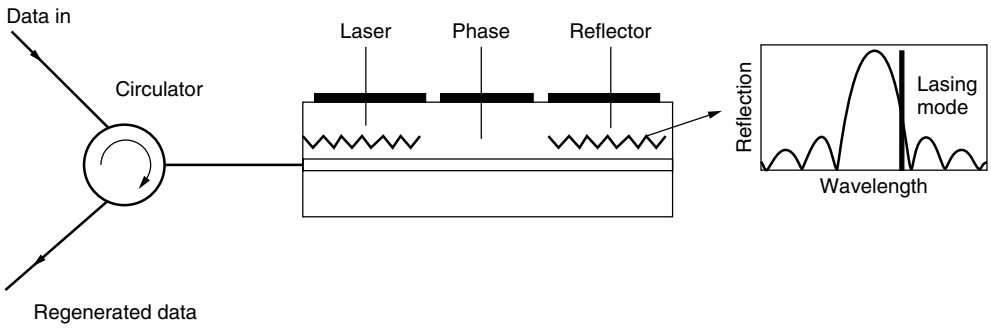


Figure 7. 2R regenerator based on a Q-switched laser.

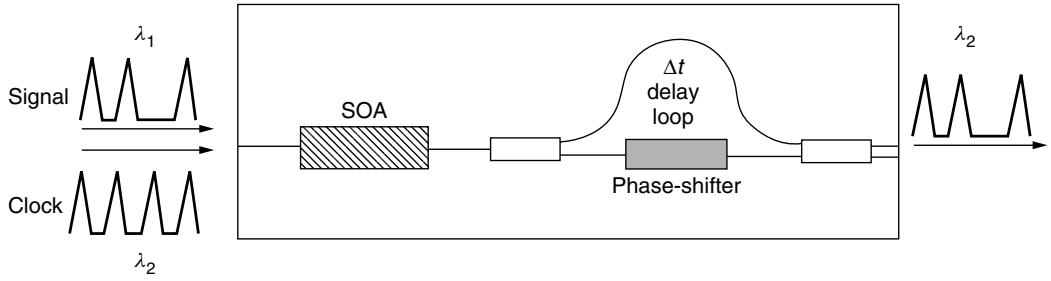


Figure 8. Scheme of semiconductor optical amplifier delayed interference device.

and can typically vary from 5 to over 40 GHz. When a data signal with a bit rate close to the repetition rate of the self-pulsation is injected, the self-pulsation synchronizes to the data signal and assumes a frequency (in GHz) that is exactly the bit rate (in Gbps).

All-optical 3R regeneration has been proposed by various groups [4–6]. In all cases, only a RZ signal format is considered. A recent idea is the use of a SOA *delayed interference device*, shown in Fig. 8. The delay Δt matches the period of the clock. The delayed interference coupler therefore produces at its output the interference of each clock pulse with the previous clock pulse. Pulses are produced in the lower arm of the coupler in the case of constructive interference and in the upper arm in the case of destructive interference. However, the phase and amplitude of the clock pulses are modulated in the SOA by the power of the data signal pulses. If the power of the data pulses is chosen carefully, a phase shift of π in the clock pulses can be obtained after each data pulse. Hence, each data pulse results in a constructive interference at the coupler and thus a clock pulse in the lower output of the coupler.

Optical 3R regeneration for NRZ signals has not been reported so far.

5. MULTICHANNEL REGENERATION

Multichannel optical regeneration is generally devised as a combination of phased-array multiplexers and demultiplexers and an array of single-channel regenerators. A possible concept for multichannel 2R regeneration, for instance, is depicted schematically in Fig. 9. The channels are demultiplexed by the *phased array* (see Optoelectronic Devices), and individual channels are fed to active Michelson interferometers [3]. After regeneration (and reflection), the channels are multiplexed again in the phased array. At the output waveguide of the phased array a circulator guarantees that the regenerated WDM signal is routed in another direction than that of the incoming signal.

Multichannel 3R regeneration requires simultaneous all-optical clock recovery of all WDM channels. This has been demonstrated recently using a module consisting of a demultiplexer, an array of SOAs (see Optical Amplifiers) and a multiplexer, all placed in an actively mode-locked fiber ring laser configuration [7]. The clock recovery module is shown in Fig. 10. The circuit of Fig. 10 forms an actively mode locked laser for each incoming channel. The mode locking results from the injection of a datastream into each SOA, which causes amplitude and

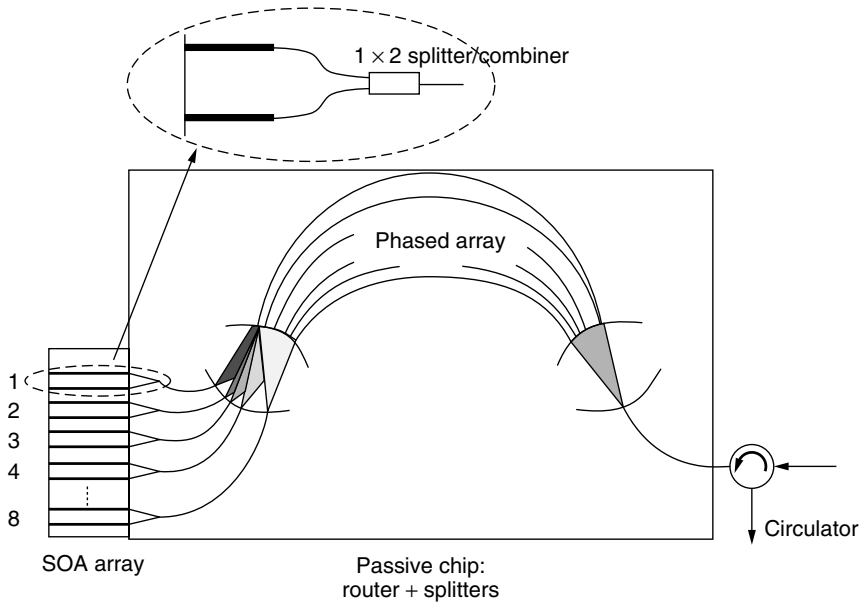


Figure 9. Top view of a multichannel 2R regenerator.

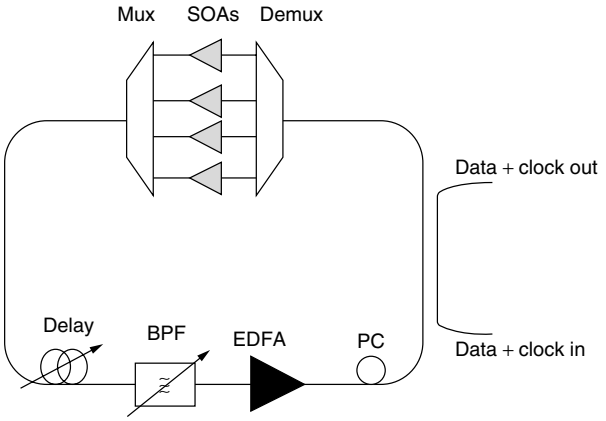


Figure 10. Multichannel clock recovery circuit [6] (PC — polarization control, EDFA — erbium-doped fiber amplifier, BPF — band-pass filter, Mux and Demux — phased arrays).

phase modulation of the light. The presence of the *tunable band-pass filter* allows the use of different phased array orders for the data and the extracted clock.

BIOGRAPHIES

Geert Morthier received the M.S. and Ph.D. degrees in electrical engineering from the University of Ghent, Belgium, in 1987 and 1991, respectively. He joined IMEC in 1991 and has been a group leader since 1992. Since 2001, he is also parttime professor at the University of Ghent. He has been author or co-author of approximately 100 publications and holds 5 patents. His areas of interest are DFB and tunable laser diodes and optical signal processing.

Jan De Merlier received the degree in physical engineering in 1998 at Ghent University, Belgium, and is currently working toward a Ph.D degree in electrical engineering at the Department of Information Technology, Ghent University. His main research interests are the dynamic properties of semiconductor optical amplifiers for use in all-optical regenerators.

BIBLIOGRAPHY

1. J. Leuthold et al., 100 Gbit/s all-optical wavelength conversion with integrated SOA delayed-interference configuration, *Electron. Lett.* **36**: 1129–1130 (2000).
2. J. De Merlier et al., Experimental demonstration of 15 dB extinction ratio improvement in a new 2R optical regenerator based on an MMI-SOA, *27th Eur. Conf. Optical Communication* (ECOC'2001), Amsterdam, Sept. 2001.
3. *J. K. Lucek, K. Smith, "All-optical signal regenerator" Optics Letters, vol 18, Aug. 1993, pp. 1226–1228.*
4. C. Bornholdt et al., Self-pulsating DFB laser for all-optical clock recovery at 40 Gbit/s, *Electron. Lett.* **36**: 327–328 (2000).
5. D. T. K. Tong et al., 160 Gbit/s clock recovery using electroabsorption modulator-based phase-locked loop, *Electron. Lett.* **36**: 1951–1952 (2000).
6. J. Leuthold et al., Novel 3R regenerator based on semiconductor optical amplifier delayed interference configuration, *IEEE Photon. Technol. Lett.* **13**: (2001).
7. V. Mikhailov and P. Bayvel, Multiwavelength all-optical clock recovery using an integrated semiconductor optical amplifier array, *Proc. ECOC'2000*, Munich, Sept. 2000, Vol. 3, pp. 63–64.

OPTICAL SOLITONS

Magnus Karlsson
Peter Andrekson
Chalmers University of Technology
Göthenburg, Sweden

1. INTRODUCTION

A traveling wave that is localized in the sense that it does not spread its energy while propagating, is defined as a *solitary wave*. A *soliton* is a solitary wave with the additional property that it can collide with other solitons and emerge unaffected with respect to energy, shape and momentum after the collision. Over the years, however, it has become common practice, especially in optics, to use "soliton" also in the less strict definition, although the mathematical differences between solitons and solitary waves are profound.

In 1834, the Scottish engineer John Scott Russel made the first scientific observation of a soliton, in the form of a single water wave that rolled forward in a canal with unchanged velocity and shape. The phenomenon was mathematically explained sixty years later in terms of a solution to a nonlinear partial differential equation that governs the motion of shallow water waves. Further progress on soliton physics came in the 1960s, when novel analytic methods were developed to exactly solve equations governing solitons.

It was realized in 1973, by Hasegawa and Tappert [1] that the weak Kerr nonlinearity present in optical fibers (which makes the refractive index increase in proportion to the optical intensity) might counteract the pulse broadening induced by group velocity dispersion (GVD). In fact, the two effects can form a stable balance in the form of a soliton pulse, which then propagates without changing shape.

Because of the lack of short-pulse laser sources at wavelengths above 1.3 μm, and low-loss silica fibers, it took another seven years for optical soliton pulses to be experimentally verified. In an experiment by Mollenauer et al. [2] in 1980, soliton pulse transmission over 700-m fiber was demonstrated. During the 1980s the soliton research aimed toward the use of solitons as information carriers in optical communications, and in 1990 the first data transmission experiment using solitons [2.8 Gbps (gigabits per second) over 23 km of fiber] were reported by Iwatsuki et al. [3].

During the 1990s, soliton-based communication systems have matured, an important reason being the development of the erbium-doped fiber amplifier (EDFA), which made high power levels commercially available. The most recent development has been toward the use of solitons in alternating dispersion maps (so called *dispersion management*), and such systems have reached performance levels near commercialization.

In the present review, we discuss both theoretical and experimental aspects of soliton transmission. We will distinguish between *conventional solitons*, which have constant dispersion during the transmission, and *dispersion managed solitons* (although solitary waves would be the proper mathematical name) for which the dispersion varies periodically during transmission.

2. WAVE PROPAGATION IN OPTICAL FIBERS

We restrict the treatment here to forward-going (in the z direction) waves along single-mode fibers. This means that the lightwave propagates according to $\exp(-i\beta z)$, where the wavenumber $\beta = n\omega/c$ is a function of the angular frequency ω, the refractive index n, and the vacuum speed of light c. In fused silica fibers the refractive index depends on the frequency and intensity of the light, and it is usually modeled as

$$n = n_0(\omega) + n_2|E|^2 \tag{1}$$

where $n(\omega)$ is the frequency-dependent part, n_2 is the nonlinear part, and E is the electric field of the wave. This nonlinear dependence on the wave intensity is known as the *optical Kerr effect*. In fused silica, the Kerr coefficient n_2 is very small; on the order of 10^{-20} m^2/V^2. The nonlinear part of the refractive index will enter as a nonlinear part in the dispersion relation, which can be written as $\beta = \beta_{\text{lin}}(\omega) + \beta_{\text{nonlin}}$. After averaging the nonlinear index over the mode profile, the nonlinear part becomes

$$\beta_{\text{nonlin}} = \frac{2\omega Z_0 n_2}{c n_0 A_{\text{eff}}}|u|^2 = \gamma|u|^2 \tag{2}$$

where $Z_0 \approx 120\pi\,\Omega$ is the wave impedance of vacuum, n_0 is the average refractive index, and A_{eff} is the effective mode area of the fiber. All these constants are contained in the nonlinear fiber coefficient γ, which has units of m^{-1} W^{-1}. We also use the wave amplitude u of the light, normalized so that the transmitted power is given by $|u|^2$. The value for γ varies from different fiber types, due to the dependence on the core area A_{eff}, but for standard single-mode fibers at $\lambda = 15.50$ nm it is approximately 2.2 m^{-1} W^{-1}.

2.1. Linear Dispersive Fiber Transmission

Linear transmission is straightforward to carry out in the frequency domain, where the wave amplitude spectrum $\tilde{u}(z,\omega)$ propagates according to $\tilde{u}(z,\omega) = \exp(-iz\beta_{\text{lin}}(\omega))\tilde{u}(0,\omega)$. This can also be expressed in the time domain, but then the function $\beta_{\text{lin}}(\omega)$ should be interpreted as the operator $\beta_{\text{lin}}(-i\partial/\partial t)$. In general we can Taylor expand the dispersion relation $\beta_{\text{lin}}(\omega)$ around the carrier frequency ω_0 to arbitrary orders. The first two terms of this expansion will correspond to the phase and group velocities of the wave, and can be removed with a suitable choice of coordinate system. Thus we can write

$$\beta_{\text{lin}}(\omega) = \frac{\beta_0''}{2}\omega^2 + \frac{\beta_0'''}{6}\omega^3 + \cdots \tag{3}$$

where now ω is measured from the carrier wavelength ω_0. The coefficient β_0'' is known as the *group velocity dispersion* (GVD) *coefficient*. The GVD is said to be *normal* if v_g decreases with frequency (i.e., $\beta_0'' > 0$), and *anomalous* if v_g increases with frequency (i.e., $\beta_0'' < 0$). Another common measure of the fiber dispersion is the *dispersion parameter*, D. The dispersion parameter D [ps/(km · nm)] is related to the GVD coefficient β_2 [ps^2/km] as $D = -\beta_2 2\pi c/\lambda^2$. In standard single-mode fibers (SMFs) the GVD is to a good approximation a linear function of the frequency, with slope $\beta_0''' \approx 0.1$ ps^3/km and zero at the *zero-dispersion wavelength* λ_0, at which $\beta_0'' = 0$. Above (below) this wavelength we have anomalous (normal) GVD. In standard single-mode fibers $\lambda_0 \approx 1.33\,\mu m$, whereas in dispersion-shifted fibers (DSFs) $\lambda_0 \approx 1.55\,\mu m$. The fiber manufacturing of today has reached a very mature level, and it is possible to tailor fibers to have a wide range of dispersion zeros and dispersion slopes.

It is instructive to consider the linear evolution of the Gaussian pulse $u(0,t) = u_0\exp(-t^2/2T_0^2)$, which is

$$u(z,t) = u_0\frac{T_0}{\sqrt{T_0^2 + i\beta_0'' z}}\exp\left(-\frac{t^2}{2(T_0^2 + i\beta_0'' z)}\right) \tag{4}$$

and where β_0''' has been neglected. The pulsewidth is determined by the real part of the exponent, which means that the width will broaden according to $T(z) = \sqrt{T_0^2 + (\beta_0'' z/T_0)^2}$. The imaginary part of the exponent corresponds to a phase modulation over the pulse, of the form $\exp(-iC(z)t^2)$, which does not affect the pulse width. Instead, this phase modulation (known as the *linear chirp* of the pulse) shows that the frequency components are redistributed within the pulse so that the slow and fast frequency components are put in the respective trailing or leading edges of the pulse. This is evident by considering the instantaneous frequency $\omega_i(t)$ of the pulse which is defined as $\omega_i(t) = d(\arg(u))/dt$, and for the Gaussian pulse it will be $\omega_i(t) = -2tC$, which reveals that the trailing (leading) part of the pulse is red (blue)-shifted for positive C (i.e., anomalous dispersion). For normal dispersion it is the other way around.

2.2. Nonlinear Fiber Transmission: Self-Phase Modulation

If we neglect the linear parts in the dispersion relation for the moment and concentrate on the nonlinear part, we see that an initial pulse $u(0,t)$ will propagate according to $u(z,t) = u(0,t)\exp(-i\gamma z|u(0,t)|^2)$. This is called *self-phase modulation* (SPM) as the pulse modulates its own phase. The corresponding instantaneous frequency will be $\omega_i(t) = -\gamma z d(|u|^2))/dt$, which is negative in the leading edge and positive in the trailing edge of the pulse, which is the same as for normal dispersion. As a result the nonlinearity will increase the dispersive spreading for normal dispersion, and decrease it for anomalous dispersion. In the case when the nonlinearity and the dispersion exactly cancel out, an optical soliton that propagates without dispersive spreading is formed.

3. CONVENTIONAL FIBER SOLITONS

3.1. The Nonlinear Schrödinger Equation

Taking both self-phase modulation and dispersion into account, we get the following propagation equation valid

for light in an optical fiber:

$$i\frac{\partial u}{\partial z} + \frac{\beta_0''}{2}\frac{\partial^2 u}{\partial t^2} - \gamma|u|^2 u = i\frac{\alpha}{2}u + i\frac{\beta_0'''}{6}\frac{\partial^3 u}{\partial t^3} \tag{5}$$

where $1/\alpha$ is the fiber loss length, which is of the order of 25 km at $\lambda = 15.50$ nm. For fiber lengths less than 1 km this can be ignored. Retaining only the terms on the left-hand side, gives the *nonlinear Schrödinger* (NLS) *equation*, which was derived for fibers originally by Hasegawa and Tappert in 1973 [1]. The NLS equation is a universal nonlinear propagation equation in the sense that it arises in many different fields of physics, and it has, for nonlinear partial differential equations the unusual and nice feature that it is *integrable*, which means that its initial-value problem can be solved exactly.

It is the anomalous dispersion case that is of most interest to us, and we will summarize the properties of the NLS equation in this case. For this it is convenient to work with the NLS equation in normalized form, and we introduce the "soliton normalizations"

$$q = u\sqrt{\frac{\gamma}{L_d}} \quad \tau = \frac{t}{t_0} \quad \xi = \frac{z}{L_d} \equiv \frac{z|\beta_0''|}{t_0^2} \tag{6}$$

where t_0 is the pulse width and L_d the dispersive length. In the case of anomalous dispersion, the NLS equation in these normalized units becomes

$$i\frac{\partial q}{\partial \xi} = \frac{1}{2}\frac{\partial^2 q}{\partial \tau^2} + |q|^2 q \tag{7}$$

The mathematical theory for the solution of this equation, which demonstrates the *inverse scattering transform* (IST) for the NLS, was given 1972 in an important paper by Zakharov and Shabat [4]. They found that a crucial role is played by the *soliton* solution to Eq. (7)):

$$q_{\mathrm{sol}}(\xi,\tau) = A\,\mathrm{sech}(A(\tau - V\xi)) \times \exp\left[-iV\tau + i\frac{(V^2 - A^2)\xi}{2}\right] \tag{8}$$

where A is an arbitrary soliton amplitude and V is an arbitrary frequency shift. Note that the soliton moves with the "velocity" V in the retarded reference frame due to this shift, which means that the soliton moves with the group velocity of the carrier wavelength. Because the soliton does not broaden dispersively, it seems very attractive for information transfer.

The IST reveals that all solutions to the NLS equation consist of solitons and dispersive radiation [i.e., dispersively broadening pulses like in Eq. (4)]. Early numerical simulations [1] also demonstrated the stable-attractor properties of the soliton. The theory also shows that any number (say, N) solitons can be present simultaneously, forming an "N soliton" solution. This can be either N well-separated soliton pulses, or if the pulses are clumped together, an oscillating N-soliton structure, called a *breather*.

One important implication of the IST is that we can already from the initial condition $q(0,\tau)$ conclude what kind of soliton will emerge for large values of ξ. For the initial condition $q(0,\tau) = A\mathrm{sech}(\tau)$, the emerging soliton is an N : th order soliton, where $A = N + \eta$, N is an integer and $|\eta| < \frac{1}{2}$. In the case where the initial pulse is given by $(1+\eta)\mathrm{sech}(\tau)$, an emerging soliton with $A = 1 + 2\eta$ is formed [5]. For an arbitrarily shaped real initial pulse, the condition for soliton creation is that the pulse area exceeds $\pi/2$ in normalized units. Note that the soliton area, $\int_{-\infty}^{+\infty} q_{\mathrm{sol}} d\tau = \pi/2$ is independent of the free parameters A and V, so that there is a critical *area* for soliton creation, rather than a critical power level. Therefore, solitons having all power levels $|q(0,0)|^2 = A^2$, and energies $\int |q_{\mathrm{sol}}|^2 d\tau = 2A$ exist. For a given pulse duration and shape, however, the condition for creation gives the necessary power to obtain a soliton, and this can be viewed as a critical power level.

Finally it is instructive to transform back to physical parameters and write the fundamental soliton of duration t_0 as

$$u_{\mathrm{sol}}(z,t) = \sqrt{\frac{|\beta_0''|}{\gamma t_0^2}}\mathrm{sech}\left(\frac{t}{t_0}\right)\exp[-iz\beta_{\mathrm{sol}}] \tag{9}$$

where the soliton wavenumber $\beta_{\mathrm{sol}} = |\beta_0''|/2t_0^2 = (2L_d)^{-1}$. It should be emphasized that the soliton wavenumber β_{sol} must lie in a regime where it cannot equal linear wavenumbers. For the NLS equation we have $\beta_{\mathrm{lin}} = -|\beta_0''|\omega^2/2 < 0$ and $\beta_{\mathrm{sol}} > 0$. Examples of cases when this condition is not fulfilled are when higher-order dispersion or periodic amplification is present. The soliton peak power P_s [W] and energy E_s [J] can be expressed as $P_s = |u_{\mathrm{sol}}(z,0)|^2 = |\beta_0''|/\gamma t_0^2$ and $E_s = 2P_s t_0 = 2|\beta_0''|/\gamma t_0$.

Next, we will review the effects of a few of the most important perturbations of solitons in fibers, and see how the right-hand side terms of Eq. (5) affect soliton propagation.

3.2. Solitons in Presence of Third-Order Dispersion

The significance of third-order dispersion (3OD) depends on the amount of soliton energy that lies in the normal-dispersion regime, which in practice depends on the pulsewidth and the carrier wavelength. The effects from 3OD are particularly important near the zero-dispersion wavelength, and the governing equation is then

$$i\frac{\partial q(\xi,\tau)}{\partial \xi} = \frac{1}{2}\frac{\partial^2 q}{\partial \tau^2} + |q|^2 q + i\varepsilon\frac{\partial^3 q}{\partial \tau^3} \tag{10}$$

where $\varepsilon = \beta_0'''/6|\beta_0''|t_0$. Using perturbation theory, it is possible to find the lowest-order (in ε) corrections to the soliton $q = A\mathrm{sech}(A\tau)\exp[-i\xi A^2/2]$, which reveals that the soliton will be spectrally shifted into the anomalous dispersion regime an amount εA^2, and as a result acquire a new group velocity. However this is not the whole story.

We note that the linear dispersion relation when third-order dispersion is present is $\beta_{\mathrm{lin}} = -|\beta_0''|\omega^2/2 + \beta_0'''\omega^3/6$, and the soliton wavenumber is $\beta_{\mathrm{sol}} = |\beta_0''|/2t_0^2$. Evidently, at a certain frequency β_{sol} equals β_{lin}, and for that frequency the soliton will act as a source for linear waves, which will then radiate and leave the soliton. The radiating frequency ω_r is approximately equal to $(2t_0\varepsilon)^{-1}$ when ε is small. This frequency lies in the normal dispersion regime. Since it is the soliton that acts as a

source for the radiation, the amplitude of the radiation will be proportional to the soliton spectral amplitude at ω_r, which is $u_{\rm sol}(t_0\omega_r) \sim \mathrm{sech}(\pi t_0\omega_r/2) \sim \exp(-\pi t_0\omega_u/2) = \exp(-\pi/(4\varepsilon))$ [e.g., 6]. Thus the radiation is exponentially small for small ε, but the fact that it cannot be expressed as a Taylor series in ε makes conventional perturbation analysis very difficult [7]. In communications, this kind of radiation must be avoided, and the solution is to make sure the soliton is sufficiently located to the anomalous dispersion regime, specifically, that ε is small enough. As a design criterion $\varepsilon < 0.04$ has been suggested [7].

3.3. Solitons in Presence of Amplification and Loss

The most important property that has been neglected in the derivation of the NLS equation for optical pulses is the effect of loss. We base the discussion on Eq. (5), retaining only the loss term $\alpha/2u$ on the right-hand side. For a soliton solution $q_{\rm sol} = A\mathrm{sech}(A\tau)\exp[-i\xi A^2/2]$, one can show that the amplitude and width are adiabatically modified according to $A(\xi) = A_0\exp[-\alpha L_d\xi]$. However, after an initial phase this rate of pulse broadening and amplitude decay will be replaced by conventional linear dispersive broadening and amplitude decay [8].

In soliton communication systems the effect of fiber loss must be compensated for by periodically spaced amplifiers. This will introduce a periodic perturbation on the soliton, with period equal to the amplifier distance L_a, and with a corresponding perturbation wavenumber $\beta_{\rm per} = 2\pi/L_a$. This wavenumber can make up for the wavenumber difference between the linear waves and solitons, and if the equation $\beta_{\rm sol} = \beta_{\rm per} + \beta_{\rm lin}(\omega)$ has any solutions, the corresponding frequencies will be unstable and radiate. From the condition above we find the radiating frequencies ω_r as $1 + (t_0\omega_r)^2 = 4\pi L_D/L_a$, and since the soliton spectral width is of the order t_0^{-1}, it suffices that $t_0\omega_r \gg 1$, and usually $L_D > L_a$ is adopted as a design criterion. This fact, that the dispersive length must be larger than the amplifier spacing is a serious obstacle for conventional soliton systems at high bit rates, and as a result solitons in the early 1990s showed most success in transoceanic systems, where the total system length is very long, but where the data rate is relatively moderate.

Finally it should be emphasized that the launched peak power of the soliton at each amplifier should be such that the path-average power between the amplifiers equals the soliton power, P_s. Since the peak power $P_{\rm peak}$ falls off as $\exp(-\alpha z)$, the average peak power over one amplifier span is $P_{\rm peak}(1-\exp(-\alpha L_a))/\alpha L_a = P_{\rm peak}(G-1)/G\ln(G)$, where $G = \exp(\alpha L_a)$ is the amplifier gain.

3.4. Sources of Timing Jitter

Another transmission obstacle is the various sources of random movement in the bit slot, namely, timing jitter of the pulses in the data transmission link. There are various sources of timing jitter, such as soliton interaction, Gordon–Haus, acoustic, and WDM–collision-induced types of jitter.

3.4.1. Soliton Interactions. Because solitons are nonlinear pulses, they will *interact* with adjacent pulses in the pulsetrain, as pointed out quite early [9]. Interaction between solitons of the same polarization and wavelength is *phase-sensitive*, so in-phase solitons will attract each other whereas out-of-phase solitons will repel each other. The interaction can be quantified via the collapse distance z_c at which in-phase solitons merge, and one can show that $z_c \approx \exp(T/2t_0)L_D\pi/4$, where T is the bit separation. The relative pulse separation is usually defined as T/T_{FWHM}, where $T_{FWHM} = 1.76t_0$ is the soliton width [in the full-width half-maximum (FWHM), sense]. A typical separation used in experiments is $T/T_{FWHM} \approx 5$.

Orthogonally polarized solitons interact substantially less, since it is then the intensity overlap that causes the interaction, rather than the amplitude overlap as for copolarized solitons. Polarization-multiplexed solitons, where adjacent pulses have orthogonal polarization can therefore be packed almost twice as dense, and $T/T_{FWHM} \approx 2.5$ is a commonly used separation.

3.4.2. Gordon–Haus Jitter. The noise from the inline amplifiers will give rise to a small jitter in the carrier frequency of each soliton, thereby changing the group velocity and hence affect the arrival time of each pulse. This is known as the *Gordon–Haus effect* after the authors of Ref. [10], and it has to be accounted for in long-distance systems. The variance of the timing jitter can be expressed as

$$\langle \delta t^2 \rangle = t_0^2 \frac{2n_{sp}(G-1)^2}{9N_s G\ln(G)} \frac{L^3}{L_d^2 L_a} \tag{11}$$

where $N_s = 2P_s t_0/h\nu$ is the number of photons in the soliton. The fact that Gordon–Haus jitter grows cubically with distance makes it particularly important at transoceanic lengths, typically exceeding 1000 Km.

3.4.3. Acoustic Jitter. The electrostriction nonlinearity in the fiber gives rise to a mechanical pressure proportional to the optical intensity, which in turn modifies the refractive index of the fiber. In particular, an intense optical pulse like a soliton will give rise to a pressure (acoustic) wave moving radially outwards from the fiber center. The pulses in the wake of this wave will experience a randomly changing local refractive index, and hence (just as for Gordon–Haus jitter) a randomly changing carrier frequency that transforms into a timing jitter. An approximate expression for the standard deviation of this jitter has been found as [11]

$$\langle \delta t^2 \rangle^{1/2} = 0.0138 D^2 B L^2 (B - 1.18)^{1/2} \tag{12}$$

where D is the dispersion in ps/(nm·km), L is the system length in million meters [Mm], and B is the bit rate in gigabits per second (Gbps). Here a soliton separation $T/T_{FWHM} = 5$ has been assumed. We see that the variance of the jitter scales with distance to the fourth power: more rapid than Gordon–Haus jitter at large distances. This is because the acoustic perturbation is constant along the fiber, whereas the amplifier noise (which is the source for Gordon–Haus jitter) increases linearly with the system length.

3.5. WDM Considerations

Wavelength-division multiplexing (WDM) is another way of increasing the bit rate of soliton systems, where several frequency bands are used for solitons transmission. This technique was pioneered by Olsson et al. [12] 1991. The problem with WDM transmission using solitons stems mainly from collisions between solitons from different wavelength channels. In a perfectly ideal NLS equation solitons would collide elastically without changing carrier wavelength. However, the presence of losses and amplification may cause the collisions to be asymmetric if they occur around an amplifier, and a result will be a frequency displacement and a concomitant timing jitter. One solution to this problem is to force the collisions to be sufficiently long, forcing the collision distance to be larger than several amplifier spans. Widely separated WDM channels, however, collide over a short distance of fiber (simply because of dispersion) so this condition will restrict the accessible optical bandwidth to WDM solitons.

3.6. Soliton Control

Soliton control is the common name for methods to control the soliton parameters such as wavelength and position. There are two different approaches: passive and active control.

Passive soliton control has been suggested in the form of filters that are inserted along the transmission path. This helps to keep the soliton wavelength fixed. In this way not only Gordon–Haus and acoustic jitter can be remedied, but also interaction jitter and WDM-collision-induced jitter. A problem with this kind of filtering is that it defines a spectral region with excess gain, in which amplifier noise will grow excessively. A way around that problem is to slightly shift the center wavelength of the filters along the transmission path. In that way the solitons will follow the frequency shift, but the linear noise will not. Such sliding filter experiments have demonstrated 8×10 Gbps WDM soliton transmission over 10000 Km [13].

Active control usually acts in the time domain by using phase or amplitude modulators to retime and reshape the solitons. Using this technique, 10 Gbps over unlimited distances [14] has been demonstrated. However, this kind of active reshaping of the pulses suffers from the same drawbacks as the conventional electronic regeneration, specifically, incompatibility with WDM, complexity, and high cost.

3.7. Polarization Effects

A single-mode fiber is not strictly single-mode, since it always allows for two polarization modes. Imperfections along the fiber cause birefringence that will accumulate randomly during propagation, and eventually cause a net birefringence that is nonnegligible. This effect, known as *polarization-mode dispersion* (PMD), is considered by several researchers as a fundamental limit for linear pulse propagation since it drifts randomly as the fiber cable is exposed to temperature and/or pressure changes. Soliton pulses, however, are found to be more robust to PMD [15] than linear pulses. The reason for this is that an attractive interaction force between the polarization states works to keep the pulses together. However, some amount of radiation is shed by the solitons as a result of the random birefringence of the fiber, and that gives rise to pulse broadening, although not as significant as that for linear pulses. The soliton robustness to PMD was recently verified by transmission on installed fibers [16].

PMD will also limit the benefit of using polarization multiplexing to suppress pulse interactions. It has been shown [17] that for high PMD, a copolarized pulsetrain will perform better than a polarization multiplexed one.

There is another timing jitter effect stemming from the birefringence of the fiber, the so-called birefringence-mediated timing jitter [18]. This effect arises from the fact that amplifier noise will give rise to a small uncertainty in the soliton polarization state, which via the birefringence is transformed to a timing jitter. Also WDM collisions of soliton trains will give rise to polarization changes, which will add to this timing jitter. This source of jitter can be reduced with inline synchronous modulation. Passive filters, however, will not work in this case, as no frequency jitter is associated with the timing displacement.

4. DISPERSION-MANAGED SOLITONS

4.1. Introduction

Since the very late 1990s, solitons have become substantially more attractive through the rapidly emerging strategy of improving the performance of soliton transmission with dispersion management (DM). While the DM strategy, which involves altering the local dispersion between a large positive and a large negative GVD (group velocity dispersion) such that the average GVD is small, has long been used in linear systems it was only relatively recently appreciated that the same technique, if properly implemented, gives rise to several very striking improvements over conventional soliton transmission systems. While DM solitons are clearly nonlinear pulses, they are by no means classical solitons.

From a commercial viewpoint, however, the most important benefit with using DM solitons is that they, in principle, can use the already installed conventional fibers (with zero dispersion at 1300 nm) allowing a much more cost-effective upgrade together with dispersion-compensating fibers (DCFs) or chirped fiber gratings (then likely collocated with inline EDFAs), at certain intervals in the link.

4.2. Properties

DM solitons are not real solitons (they are rather solitary waves) in the sense that they have an inverse scattering transform that can be used to find the most relevant properties. Instead they emerged from extensive simulation work, and it was quite surprising to many researchers that the simulations revealed such stable and strictly periodic pulses.

In Fig. 1 the evolution of the temporal and spectral widths for a DM soliton is shown. Also shown is the dispersion map, which is a plot of how the dispersion changes with propagation distance. A lossless model is assumed, which is applicable to dense amplifier

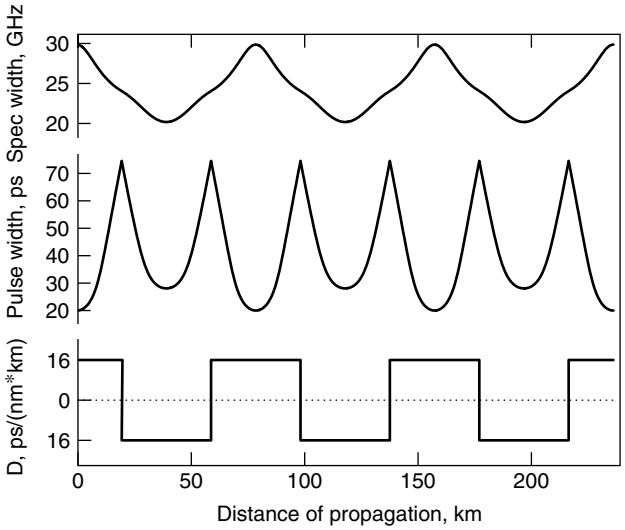


Figure 1. The evolution of the spectral width (top) temporal width (middle) of a DM soliton with zero average dispersion and map strength $S \approx 4$. The bottom plot shows the dispersion map. The symmetry of the evolution within the map period is due to the lossless model employed. (Figure contributed by A. Berntsson.)

spacing, but even when losses are included the qualitative conclusions and general properties will still hold. One difference in the lossy case is, however, that the symmetry of the propagation dynamics is removed, and the spectral dynamics will be concentrated to the positions in the dispersion map where the power is largest.

As a result of massive simulation work done by many groups the following properties of DM solitons have been found.

- The pulsewidth, chirp, and spectral width oscillates periodically in the dispersion map. There are two points within the period at which the pulse is chirp-free, and those correspond to local minima of the pulsewidth. One of those points is the global minimum width, referred to as the "shortest pulse" below.
- A central parameter that is useful for the characterization of DM solitons is the *map strength*, $S = (L_1|\beta_1''| + L_2|\beta_2''|)/T_{\mathrm{FWHM}}^2$, where L is length, β'' is dispersion, T_{FWHM} is the minimum pulsewidth in the full-width half-maximum sense, and subscripts 1,2 refer to the two fibers in the dispersion map. Physically the map strength is the number of dispersive lengths the pulse propagates in one period. DM solitons have been found for map strengths ranging from $S = 0$ (which is the same as conventional solitons) to $S \approx 12$, although this upper limit is a transition regime in which the pulses radiate and a perfect periodic evolution never arises.
- DM solitons have been found for anomalous, normal, and zero average dispersion $\overline{\beta}''$, defined as $\overline{\beta}'' = (L_1\beta_1'' + L_2\beta_2'')/(L_1 + L_2)$. Normal average dispersion is only possible for map strengths above 3.9.
- The shape of the shortest pulse ranges from hyperbolic secant at $S = 0$ to Gaussian for higher map strengths, and this is also evident from the time–bandwidth product, which increases with S from 0.32 (Sech) at $S = 0$ to 0.44 (Gaussian) and even higher for large values of S. In addition, the shortest pulse has oscillating tails in the pulse wings.
- The energy of a DM soliton pulse is enhanced relative to a soliton with the same average dispersion and pulsewidth.
- The interaction between DM solitons is less than that of conventional solitons, and an optimum map strength exists that minimizes the interaction.

The fact that DM solitons can work for an average net zero GVD and even for normal dispersion, was a striking and unexpected feature that would not work with conventional solitons. This can be understood by the fact that the spectral width of the pulses are larger when they propagate in the local anomalous dispersion regime than when they propagate in the normal dispersion regime (see Fig. 1). The pulses then effectively sense a net anomalous dispersion that is balanced by the nonlinearity as for conventional solitons.

The technical improvements with DM solitons over conventional ones are numerous. The signal-to-noise ratio is improved since the DM solitons have a larger peak power than do the corresponding conventional solitons. DM solitons have less Gordon–Haus and acoustic timing jitter, since the system average GVD is much smaller in these systems. A very important added benefit appears in wavelength-division-multiplexed (WDM) systems. Because of the alteration between large positive and negative GVD along the path, the jitter induced from WDM, soliton collisions is greatly reduced. This, in turn, allows for very dense WDM, which will improve the spectral efficiency substantially. The soliton PMD robustness is maintained or improved for DM solitons.

An important practical consequence of using DM solitons is that they reduce (or eliminate) the need for inline soliton control such as synchronous modulation or sliding filters. Yet soliton control methods are still applicable and will give improvement in terms of signal-to-noise ratio for DM solitons as well.

Quite impressive circulating loop experiments including WDM have been reported. In the study by Le Guen et al. [19] 51 very densely packed WDM channels each operating at 20 Gbps were transmitted over 1000 km with 100-km sections of standard fiber, clearly demonstrating the strength of the DM soliton technique.

4.3. Intrachannel Impairments

DM solitons have so many attractive features that they are likely to be implemented commercially in the near future. However, there are some novel transmission impairments, unique to DM solitons that need to be accounted for and analyzed in more detail. They are the so-called intrachannel effects; intrachannel four-wave mixing (ICFWM), and intrachannel cross-phase modulation (ICXPM) [20], and arise due to the nonlinear interaction between two neighboring pulses. Four-wave mixing (FWM) and cross-phase modulation (XPM) are

usually effects associated with WDM transmission. However, the fact that DM solitons are chirped and broadened, will cause different frequency components from neighboring pulses within the same wavelength channel to overlap in time, thereby causing FWM and/or XPM.

ICFWM arises for two neighboring pulses, that via four-wave mixing (FWM), creates new frequency components that in the time domain will give rise to a new pulse (commonly referred to as a "ghost pulse"), next to the two. The ghost pulse will then give rise to intersymbol interference and reduction of the eye opening. ICFWM is most prominent for large map strengths and high power.

ICXPM can be viewed as DM soliton interaction, and physically, it manifests as the frequency shift of one pulse induced by the presence of a neighboring pulse, which, by the dispersion transforms into a timing jitter. The effect can be minimized by selecting proper map strength and prechirp of the DM soliton.

As a rule, however, it seems that map strengths in the range 1–8 make best use of the unique features of DM solitons. This means that the use of installed standard fiber becomes difficult at very high bit rates (say, beyond 30–40 Gbps) as shorter pulses require a more rapidly varying dispersion map to maintain a proper S value. This is a limitation similar to the amplifier spacing limitation in conventional soliton systems, but it is much less severe. If an option is to install new dispersion-shifted fiber, on the other hand, this limitation becomes essentially unimportant.

5. EXPERIMENTS AND FIELD TRIALS

5.1. Soliton Pulse Sources

When doing soliton experiments, be it conventional or DM solitons, particular importance is placed on the properties of the pulse source, as it sets the lower limit of the system performance. A high-bit-rate soliton pulse source needs to produce low-chirp, low-timing-jitter pulses with proper duration (in the picosecond regime), repetition rate (10–40 GHz) and shape.

One possible choice is gain-switched (GS) laser diodes, possibly with an external cavity for tunability. However, they suffer the drawback of producing pulses that are strongly chirped, asymmetric and often too wide.

For laboratory experiments fiber ring lasers (FRLs) are very attractive, as they provide wavelength and pulse width tunability, besides meeting the above mentioned demands. Their drawback is that they are bulky, need active stabilization and sometimes also temperature control to achieve long-term stability.

Finally, it appears quite clear that electroabsorption modulators (EAMs) [whether integrated or not with a distributed feedback (DFB) laser] are very useful and simple sources for soliton transmission. While such sources were developed for linear NRZ (non-return-to-zero) systems, they have now proven to be near ideal in soliton systems as well. Although EAMs are not commercially available at 40 GHz yet, they are likely to be so within the near future.

Special considerations need to be taken in DM soliton systems, however, as the launch condition is different than for conventional solitons. The pulses should have a linear chirp such that it fits seamlessly in the periodically induced chirp variation along the link. This can be achieved by incorporating a proper length of fiber (or chirped fiber grating) in the transmitter once the overall dispersion map is known.

5.2. Loop Experiments

In order to investigate really long distances (megameters) of transmission, the loop experiments were developed in the early 1990s. This means that the data pulses are injected in a loop consisting of transmission fiber and amplifier, and then left to propagate a number of laps corresponding to a certain transmission distance. Acoustooptic switches are used to switch the pulse train in and out from the loop at proper time intervals. The drawback of loop experiments is that they may be poor models of reality when it comes to things like dispersion variation along the fiber, PMD or various kinds of drifts that may arise. In addition, a real system has more options to fine-tune, for example, amplifiers along the transmission line. However, as long as these drawbacks are recognized, loop experiments are very powerful indeed, and invaluable in lab evaluations of long-distance transmission.

5.3. Field Trials

In the field, many transmission link design restrictions and fiber properties make the systems far from optimal. The actual fiber parameters are nonperiodic with propagation distance as the systems are straight lines rather than relatively short loops. Both loss (in particular when including many contacts and splices along the link) and the PMD are typically much higher than in the laboratory. In particular, the PMD is higher since the fiber is no longer wound on a small drum making the mode coupling length longer, but also often simply because the installed fiber is old, being made before PMD was considered a real obstacle. The dispersion (or more specifically the zero-dispersion wavelength) might vary significantly along a fiber span. In addition, it may not be possible to tailor the dispersion map and amplifier locations to reach an optimal state. All these examples of nonidealities justify the need for field experiments.

Several soliton field experiments have been conducted in Japan by NTT [21–25], in the United States by MCI/Pirelli [26], and in Europe by the Advanced Communication Technologies & Services (ACTS) projects [27–30]. This is a good indication that solitons are indeed foreseen as very interesting candidates in commercial systems. Table 1 summarizes some data for the 10 soliton field experiments conducted from 1995 through 1999s. Much of the work has been at a bit rate of 40 Gbps, which is natural as this is expected to be the next standard trunk TDM rate. Again, it is not very easy to compare the results as the situation in each case differs. All the systems operated in the 1550 nm range, used optical time-division demultiplexing to the 10 Gbps electronic base rate, and the average loss/km ranged from 0.24 to 0.33 dB/km. Dispersion-shifted fiber was always used, apart from in

Table 1. Overview of the Soliton Field Experiments Conducted to Date[a]

Capacity (Gbps × 10^6 m)	Soliton Source	L_a (km)/ G (dB)	PMD (ps/km$^{1/2}$)/ DGD · T^{-1}	T/T_{FWHM}	Fiber	Ref./Year
10 × 2.7	EAM/GS	90/—	—	5	DSF	[25]/1995
10 × 2	FRL/GS	55/16	—	5	DSF	[21]/1995
10 × 0.3	—	50/12	0.04 / 0.7 %	2	SMF	[27]/1998
10 × 0.9	—	75/20	0.9 / 27 %	3.3	SMF+DCF	[26]/1998
4 × 10 × 0.45	—	75/20	0.9 / 27 %	3.3	SMF+DCF	[26]/1998
20 × 2	FRL	55/16	—	5	DSF	[22]/1995
40 × 1	FRL	55/16	selected	5	DSF/DCF	[24]/1998
40 × 1.4	FRL	55/16	selected	5	DSF/DCF	[25]/1998
40 × 0.4	FRL	57/19	0.3 / 24 %	2.5	DSF	[28]/1999
40 × 0.5	EAM	100/24	0.25 / 22 %	2.5	DSF	[29]/1999
80 × 0.2	FRL	57/19	0.11 / 12 %	2.8	DSF	[30]/1999

[a] Here, G and L_a denote the amplifier gain and separation; T, the bit separation; DGD, the differential group delay; and T/T_{FWHM}, the relative soliton separation.

two cases [26,27], where standard fiber was used. The study by Robinson et al. [26] deserves particular attention because DCF was used for dispersion compensation, which makes this the only DM soliton field experiment to date. In addition, that study [26] describes the transmission of four WDM channels (4 × 10 Gbps) but then over half the distance (450 km).

Polarization multiplexing was used in four experiments [27–30] and this serves mainly to allow the use of relatively wide pulses, which in turn allows for larger amplifier spans. Polarization multiplexing, however, is not as useful if the PMD of the system is high, as then the orthogonally polarized pulses would start to drift statistically in time relative to each other thereby creating intersymbol interference and increasing the soliton interaction. In the 40-Gbps cases and above, PMD was found to be the main capacity limiting factor. In two cases a special selection of low-PMD fiber was made [24,25].

Most of the more recent experiments used a mode-locked fiber ring laser (FRL) as a source, probably because these provide excellent pulse quality as well as tunability in terms of wavelength and pulse width. Other experiments used either gain-switched (GS) lasers or electroabsorbtion modulators (EAMs).

In only one case [25] was inline soliton control used (in the form of intensity modulation), and this experiment also achieved the highest capacity (54 Tbps · km).

Future soliton field trials are expected to (1) take advantage of the now well-known strategy of improving soliton transmission performance with dispersion management, as this method is very attractive for upgrading existing fiber plants; (2) implement dense WDM (in non-DS fiber lines) to boost aggregate capacity; (3) utilize different forms of inline control, particularly at high bit rates; and (4) further address the implications of PMD and techniques to combat it. The interesting WDM-TDM tradeoff for optimization of overall aggregate capacity will depend on the details of the fiber line parameters.

6. EVALUATION AND FUTURE OUTLOOK

To conclude, we note that the motivation for using solitons as information carriers have changed over the years. The property of being resistant to dispersive broadening was originally the main feature, but this was considered less important when the dispersion-compensating fibers became commercially available. Instead, this led to the development of the dispersion-managed soliton. The advantage of the DM soliton over linear transmission are features like the large power (which enables high signal-to-noise ratio) and PMD robustness. On the other hand, the difference between linear and nonlinear pulses are becoming increasingly fuzzy, and perhaps the distinction should be made between RZ (return-to-zero) and NRZ modulation rather than between linear and nonlinear transmission.

It is nevertheless interesting to note that solitons are now not only considered for oceanic systems but also for shorter terrestrial systems. There are still several challenges and opportunities remaining in order to take full advantage of solitons and to reach a better understanding. PMD remains an important topic that is not entirely understood when using solitons. Further work is also needed on WDM soliton and very-high-speed TDM soliton-systems. The use of DM solitons is a very recently established technique and there are thus many issues to consider. These include studies of robustness to deviations of optimum conditions, such as improper pulse launch condition, impact of nonperiodic dispersion maps and of PMD (both of which are difficult to study in loop experiments), and intrachannel effects. Nevertheless, solitons have now reached a level of maturity such that commercialization seems very near.

BIOGRAPHIES

Magnus Karlsson was born in Gislaved, Sweden, 1967. He received his M. Sc in engineering physics in 1991 and his Ph.D in electromagnetic field theory in 1994 from Chalmers University of Technology, Gothenburg, Sweden. The title of his Ph.D thesis is "Nonlinear propagation of optical pulses and beams." Since 1995, he has been with the Photonics Laboratory at Chalmers, first as assistant professor, and since 2001 as associate profesor in photonics. At the Photonics Lab his research has been devoted to fiber optic communication

systems, in particular transmission aspects such as fiber nonlinearities, solitons, four-wave mixing and polarization effects. He has authored or coauthored around 50 journal articles, 30 conference contributions, and two patents.

Peter A. Andrekson received his M.S. degree in electical engineering in 1984, and his Ph.D. degree in optoelectronics in 1988 from Chalmers University of Technology, Gothenburg, Sweden. During 1989–1992 he was with AT&T Bell Laboratories, Murray Hill, New Jerse, working on high speed fiber-optic transmission systems. He returned to Chalmers University in 1992 where he currently holds a professorship in photonics. Dr. Andrekson is the author and coauthor of over 200 technical publications and conference papers, and holds three patents. He also serves on several technical conference program committees. In 2000 he was awarded the Telenor Nordic Research Award for his contribution to optical technologies. Since 2000 he is on leave from Chalmers University, working as director of research at CENiX Inc. His interests cover essentially all aspects of high-capacity optical fiber communications.

BIBLIOGRAPHY

1. A. Hasegawa and F. Tappert, Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers. I. Anomalous dispersion, *Appl. Phys. Lett.* **23**(3): 142–144 (1973).
2. L. F. Mollenauer, R. H. Stolen, and J. P. Gordon, Experimental observation of picosecond pulse narrowing and solitons in optical fibers, *Phys. Rev. Lett.* **45**(13): 1095–1098 (1980).
3. K. Iwatsuki, S. Nishi, M. Saruwatari, and M. Shimizu, 2.8 Gbit/s optical soliton transmission employing all laser diodes, *Electron. Lett.* **26**(1): 1–2 (1990).
4. V. E. Zakharov and A. B. Shabat, Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media, *Sov. Phys. JETP* **34**: 62–69 (1972).
5. J. Satsuma and N. Yajima, Initial value problems of one-dimensional self-modulation of nonlinear waves in dispersive media, *Progr. Theor. Phys. Suppl.* **55**: 284–306 (1974).
6. N. Akhmediev and M. Karlsson, Cherenkov radiation emitted by solitons in optical fibers, *Phys. Rev. A* **51**(3): 2602–2607 (1995).
7. P. K. A. Wai, H. H. Chen, and Y. C. Lee, Radiations by "solitons" at the zero group-dispersion wavelength of single-mode optical fibers, *Phys. Rev. A* **41**(1): 426–439 (1990).
8. K. J. Blow and N. J. Doran, The asymptotic dispersion of soliton pulses in lossy fibres, *Optics Commun.* **52**(5): 367–370 (1985).
9. J. P. Gordon, Interaction forces among solitons in optical fibers, *Opt. Lett.* **8**(11): 596–598 (1983).
10. J. P. Gordon and H. A. Haus, Random walk of coherently amplified solitons in optical fiber transmission, *Opt. Lett.* **11**(10): 665–667 (1986).
11. E. M. Dianov, A. V. Luchnikov, A. N. Pilipetskii, and A. M. Prokhorov, Long-range interaction of picosecond solitons through excitation of acoustic waves in optical fibers, *Appl. Phys. B* **B54**(2): 175–180 (1992).
12. N. A. Olsson et al., Bit-error-rate investigation of two-channel soliton propagation over more than 10,000 km, *Electron. Lett.* **27**(9): 695–697 (1991).
13. L. F. Mollenauer and P. V. Mamyshev, Massive wavelength-division multiplexing with solitons, *IEEE J. Quant. Electron.* **34**(11): 2089–2102 (1998).
14. M. Nakazawa et al., Experimental demonstration of soliton data transmission over unlimited distances with soliton control in time and frequency domains, *Electron. Lett.* **29**(9): 729–730 (1993).
15. L. F. Mollenauer, K. Smith, J. P. Gordon, and C. R. Menyuk, Resistance of solitons to the effects of polarization dispersion in optical fibers, *Opt. Lett.* **14**(21): 1219–1221 (1989).
16. B. Bakhshi et al., Experimental observation of soliton robustness to polarisation dispersion pulse broadening, *Electron. Lett.* **35**(1): 65–66 (1999).
17. X. Zhang, M. Karlsson, P. A. Andrekson, and E. Kolltveit, Polarization-division multiplexed solitons in optical fibers with polarization-mode dispersion, *IEEE Photon. Technol. Lett.* **10**(12): 1742–1744 (1998).
18. L. F. Mollenauer and J. P. Gordon, Birefringence-mediated timing jitter in soliton transmission, *Opt. Lett.* **19**(6): 375–377 (1994).
19. D. Le Guen et al., Narrow band 1.02 Tbit/s (51*20 Gbit/s) soliton DWDM transmission over 1000 km of standard fiber with 100 km amplifier spans, *Proc. OFC/IOOC'99. Optical Fiber Communication Conf. and Int. Conf. Integrated Optics and Optical Fiber Communications*, 1999, pp. PD4–1–PD4–3.
20. R. J. Essiambre, B. Mikkelsen, and G. Raybon, Intra-channel cross-phase modulation and four-wave mixing in high-speed TDM systems, *Electron. Lett.* **35**(18): 1576–1578 (1999).
21. M. Nakazawa et al., Field demonstration of soliton transmission at 10 Gbit/s over 2000 km in Tokyo metropolitan optical loop network, *Electron. Lett.* **31**(12): 992–994 (1995).
22. M. Nakazawa et al., Soliton transmission at 20 Gbit/s over 2000 km in Tokyo metropolitan optical network, *Electron. Lett.* **31**(17): 1478–1479 (1995).
23. K. Iwatsuki et al., Field demonstration of 10 Gb/s-2700 km soliton transmission through commercial submarine optical amplifier system with distributed fiber dispersion and 90 km amplifier spacing, *Proc. 21st European Conf. Optical Communication, ECOC'95*, 1995, pp. 987–990.
24. A. Sahara et al., Single channel 40 Gbit/s soliton transmission field experiment over 1000 km in Tokyo metropolitan optical loop network using dispersion compensation, *Electron. Lett.* **34**(22): 2154–2155 (1998).
25. K. Suzuki et al., 40 Gbit/s soliton transmission field experiment over 1360 km using inline soliton control, *Electron. Lett.* **34**(22): 2143–2145 (1998).
26 N. Robinson et al., 4xSONET OC-192 field installed dispersion-managed soliton system over 450 km of standard fiber in the 1550 nm Erbium band, *Proc. Optical Fiber Communication Conf.* 1998, pp. PD19–1–PD19–4.

27. P. Franco et al., 10 Gbit/s alternate polarisation soliton transmission over 300 km step-index fibre link with no inline control, *Electron. Lett.* **34**(11): 1116–1117 (1998).

28. E. Kolltveit et al., Single-wavelength 40 Gbit/s soliton field transmission experiment over 400 km of installed fibre, *Electron. Lett.* **35**(1): 75–76 (1999).

29. F. Matera et al., Impact of polarisation mode dispersion in field demonstration of 40 Gbit/s soliton transmission over 500 km, *Electron. Lett.* **35**(5): 407–408 (1999).

30. J. Hansryd et al., 80 Gbit/s single wavelength soliton transmission over 172 km installed fibre, *Electron. Lett.* **35**(4): 313–315 (1999).

FURTHER READING

Agrawal G. P., *Nonlinear Fiber Optics*, 2nd ed., Academic Press, San Diego, 1995.

Hasegawa A. and Y. Kodama, *Solitons in Optical Communications*, Clarendon Press, Oxford, 1995.

Mollenauer L. F., J. P. Gordon, and P. V. Mamyshev, Solitons in high bit rate long-distance transmission, in I. P Kaminow and T. L. Koch (eds.), *Optical Fiber Telecommunications III A*, Academic Press, San Diego, 1997.

Taylor J. R. (ed.), *Optical Solitons-Theory and Experiment*, Cambridge Univ. Press, Cambridge, UK, 1992.

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 4

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 4

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at
http://www.mrw.interscience.wiley.com/eot

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor.
p. cm.
includes index.
ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G.
TK5102 .W55 2002
621.382′03 — dc21

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

OPTICAL SOURCES

JENS BUUS
Gayton Photonics
Gayton, Northants
United Kingdom

1. INTRODUCTION

Since the early 1980s an increasing fraction of the worlds telecommunication traffic has been carried on optical fiber, and since the late 1980s virtually all new trunk lines have been based on fiberoptics. Optical fibers are now the dominating medium for a variety of communication systems, ranging from short-distance data links to transoceanic telecommunication systems.

The dramatic increase in traffic brought about by the Internet has made optical fibers even more attractive because of the several terahertz of transmission bandwidth they offer. The wide bandwidth can be exploited by multiplexing a number of optical sources operating on different optical frequencies. This technology is known as *wavelength-division multiplexing* (WDM).

Different fiberoptic communication systems place different requirements on the optical sources used in the systems. The basic properties and characteristics of these sources are reviewed in this article.

2. HISTORICAL DEVELOPMENT

Since the late 1960s there has been an interesting interaction between the development of optical fibers and optical sources. Semiconductor lasers were first demonstrated in 1962, but it was only after the demonstration of room-temperature continuous-wave (CW) operation in 1970 that the practical use of these lasers became a reality. We note in passing that this improved laser performance was due to the introduction of the *heterostructure*, Alferov and Kroemer shared half the 2000 Nobel Prize in Physics for their work on this topic.

The use of optical fibers for long-distance communication was proposed in 1966, and the fiber loss was reduced to 20 dB/km in 1970. In the early fibers the minimum loss occurred at relatively short wavelengths, well suited to the emission wavelength of GaAs based lasers, which is in the 800–900-nm range. The fibers at that time were multimoded, and reduced fiber losses meant that transmission distances were limited mainly by modal dispersion (i.e., different fiber modes having different group velocity, thus leading to pulse distortion).

Further improvements resulted in fiber losses below 0.2 dB/km by 1980. As the fiber losses were reduced, the spectral range where the losses were lowest moved toward longer wavelengths. A wavelength region of particular interest was around 1300 nm, where the fiber dispersion was minimized. This new wavelength range was exploited by introducing lasers and LEDs based on InGaAsP compounds using InP substrates, this development started around 1980. These devices can cover the wavelength range from about 1100 nm to about 1700 nm. One of the first examples of a commercial long-distance transmission system at these "long wavelengths" was the London–Birmingham link in the early 1980s, based on LEDs operating around 1300 nm.

From the early 1980s single-mode fibers were being introduced, thus eliminating modal dispersion. The lowest loss for these fibers occurs at wavelengths around 1550 nm, which is within the range accessible by InGaAsP/InP lasers. However, the InP-based lasers have a tendency to operate simultaneously in several longitudinal modes. From the laser cavity length (typically about 300 μm), it follows that the longitudinal mode spacing (in frequency) is about 120 GHz, corresponding to a spectral spacing of about 1 nm (in wavelength). At a wavelength of 1550 nm a standard single-mode fiber has a chromatic dispersion of about 17 ps/(km $\cdot$ nm). Consequently, multimode laser operation will give rise to dispersion problems for high-speed systems operated over a long fiber length, and the development of single-frequency lasers (i.e., lasers operating in a single longitudinal mode) then became a priority. Single-frequency operation can be achieved by incorporating a wavelength selective element in the laser, typically a grating as in the DFB (distributed feedback) laser.

The development of the fixed wavelength DFB laser in turn made the efficient use of wavelength-division multiplexing (WDM) possible. In these systems the signals from a number of lasers, operating at different optical frequencies, are multiplexed together and transmitted over a single fiber, thus increasing the transmission capacity of the fiber significantly. As an example, a spectral range of 30 nm (around a wavelength of 1550 nm) corresponds to a frequency range of about 3800 GHz; using lasers spaced in frequency by 50 GHz will allow 76 separate channels, each of which can carry data at a rate of, for example, 10 Gbps (gigabits per second), thus giving an aggregate capacity of 760 Gbps. This capacity can be increased even more by the use of a wider spectral range, and/or a higher spectral efficiency (ratio of data rate to channel spacing).

The development of wavelength selective lasers and tunable lasers also opens new possibilities. Not only can these lasers be used as flexible spares or as "uncommitted" wavelength sources; they also allow the use of wavelength routing, where the path of a signal through a network is entirely determined by the wavelength of the signal. Ultimately the use of lasers that can switch fast between wavelengths opens the possibility for packet switching on the optical level.

Other interesting developments include lasers and LEDs specifically designed for (short-distance) data links, and pump lasers for optical amplifiers.

The reader is referred to Refs. 1 and 2 for more details on the history of the development of semiconductor lasers.

3. LIGHT-EMITTING DIODES

Red *light-emitting diodes* (LEDs) are well known from their use in displays and as indicator lights. LEDs based on GaAs or InP emit in the near infrared and are used for communication purposes. The basis for the operation

of an LED is that carriers (electrons and holes) are injected into a forward-biased *p-n* junction and recombine spontaneously, thereby generating photons. The photon energy (and hence the wavelength of the generated light) is determined mainly by the bandgap of the region where the recombination takes place. However, the carriers have a spread in energy, leading to a spectral width of the emitted light of a few times kT, where k is Bolzmann's constant and T is the operating temperature. Consequently the spectral width for an LED operating in the 1300-nm spectral region will be of the order 100 nm.

Not all the electrical power supplied to the LED is converted to light: (1) some power is lost because of electrical losses, (2) the radiative (spontaneous) recombination competes with various nonradiative recombination processes, and (3) only a finite fraction of the generated light is able to escape from the structure. This last effect is due to the process of total internal reflection; since the refractive index of the (semiconductor) LED structure is high compared to that of the surrounding medium (air), only light propagating at an angle nearly perpendicular to the surface will be able to escape from the structure. Additional optical losses occur when the light is coupled into a fiber. For coupling into a standard multimode fiber with a core diameter of 50 μm the overall efficiency (optical power in the fiber compared to electrical power supplied to the LED) is typically of the order of 1%, corresponding to coupled power levels of the order of 100 μW.

The light output from an LED can be modulated directly by varying the current passed through the device. In the absence of parasitics, the maximum possible modulation speed is approximately given by the inverse of the recombination time. With recombination times in the nanosecond range, it follows that LEDs can typically be used at data rates of up to a few hundred megabits per second (Mbps).

It should be noted that LEDs can be optimized for power levels of up to several milliwatts, and higher coupled power levels can be achieved by using large core fibers. Obviously, the wide spectral width of LEDs leads to chromatic dispersion, and the use of multimode (in particular large-core) fibers leads to modal dispersion. However, since LEDs are used at moderate data rates, they are an attractive simple and low-cost solution for links of a modest length (i.e., up to a few kilometers).

Finally, it should be mentioned that near infrared LEDs are also widely used for very short range free space communication between computers and peripheral equipment.

4. LASERS

4.1. Laser Basics

In order to understand the workings of a laser, we consider a system where the constituents (electrons, atoms, ions, or molecules) have two possible energy states. Transitions between these two states are accompanied by the absorption or the emission of photons, where the photon energy is equal to the difference in energy between the two states. In 1917 Einstein explained the relation between the energy distribution of a gas of molecules and that of the radiation field (Planck's law) by assuming that the following three processes occur:

1. *Spontaneous Emission*. Transition from the higher to the lower state accompanied by the emission of a photon.
2. *Absorption*. Transition from the lower to the higher state brought about by the absorption of a photon.
3. *Stimulated Emission*. In this process the transition from the higher to the lower state is triggered by incoming photons with energy equal to the transition energy; the additional photon emitted in the transition is in phase with the incoming photons. The transition probability is proportional to the number of incoming photons.

In thermal equilibrium the higher-energy state is less densely populated than the lower one, and it follows that an incoming stream of photons will be attenuated. However, if a situation is created where the higher-energy state is more densely populated than would be possible at the lower amplification level, this would be known as *population inversion*. Such a system is shown schematically in Fig. 1.

The basis for the acronym *LASER* (light amplification by stimulated emission of radiation) becomes clear from this description. The normal use of the word *laser*, however, refers to light generation (rather than amplification), and in order to construct an oscillator working at the lasing frequency, it is also necessary to provide feedback. A laser is usually constructed by placing material with an inverted population between a pair of partly reflecting mirrors. Light moving back and forth between the mirrors is amplified as a result of the stimulated emission process.

The combination of amplification and feedback from the mirrors forms an oscillator, and oscillation takes place if the amplification balances the loss caused by light escaping through the mirrors. With a gain factor g, a cavity length

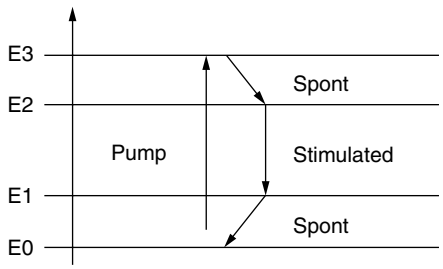


Figure 1. Schematic diagram of a four-level laser system. The upper energy level 3 is short-lived, giving a fast decay to the upper level 2 involved in the lasing transition, this level is long-lived. From the lower laser level 1 there is a fast decay to the ground level 0. "Pumping" from the ground level to the highest level is done by flashlamp, electric discharge, or the use of another laser, and population inversion can be achieved between levels 2 and 1.

L, and two mirrors both with a power reflectivity of R, this condition can be written as

$$\exp(gL)R\exp(gL)R = 1 \tag{1}$$

Since photons created by the stimulated emission process are emitted in phase with the incoming photons, the light emitted from the laser cavity is coherent.

A large number of laser types exist (gas lasers such as HeNe and CO_2, solid-state lasers such as Nd:YAG, etc.), and are being used in numerous fields (material processing, medical applications, etc.). The laser type of interest for optical fiber communication is the semiconductor laser. This laser type is also used in CD and DVD players, as well as in scanners and pointers.

4.2. Semiconductor Lasers

Nearly all semiconductor lasers are based on the double *heterostructure*. In this structure, a material with a relatively narrow bandgap — the *active layer* — (normally undoped) is sandwiched between a pair of n-type and p-type materials with wider bandgaps — the *confinement layers*. When this structure is under forward-biased *quasi-Fermi levels* are formed, and electrons and holes are injected into the active layer from the n-type and p-type materials, respectively. The Fermi level(s) determine the energy distribution of electrons in the conduction band and holes in the valence band. Population inversion, and thereby gain, is achieved when the quasi-Fermi level separation exceeds the bandgap of the active layer (see Fig. 2).

If the bandgap difference is sufficiently large, carriers injected into the active layer cannot escape over the heterobarrier, and carrier recombination can take place only in the active layer. Light generated in the active layer is not absorbed in the confinement layers since semiconductors are transparent to light with a photon energy lower than the bandgap. The photon energy is given by the product of Planck's constant h and the optical

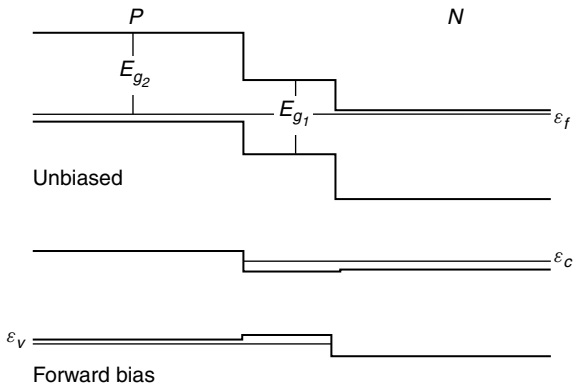


Figure 2. Double heterostructure formed with a material with a narrow bandgap E_{g1} placed between n-type and p-type materials with a wider bandgap E_{g2}. Carriers cannot have energies corresponding to states within the bandgap. Without bias the Fermi level ε_f is continuous. Under forward bias quasi-Fermi levels ε_c and ε_v are formed in the conduction and valence bands, respectively.

frequency ν, which is in turn equal to the speed of light in vacuum, c, divided by the wavelength λ, hence

$$h\nu = \frac{hc}{\lambda} \tag{2}$$

At a given optical frequency, a narrow bandgap semiconductor generally has a higher refractive index than a semiconductor with a wider bandgap. Consequently the structure shown in Fig. 2 also forms a planar dielectric waveguide with a high-index core between a pair of low-index cladding layers, analogous to an optical fiber. It is characteristic for a dielectric waveguide that only a part of the optical power is present in the core since the power distribution extends well into the cladding layers. The power fraction in the core is known as the *confinement factor*, and denoted by the symbol Γ.

The optical field distribution supported by the waveguide is known as a *mode*, and the structure is usually designed in such a way that only a single mode exists. The optical field propagates at a speed given by c/n_{eff}, where the *effective index* n_{eff} is higher than that of the cladding layers, but lower than that of the core.

The width of the optical power distribution is characterized by the *spotsize*, which is on the order of, or even less than, 1 μm. Since this is small compared to the wavelength, the output beam from a semiconductor laser is usually quite divergent. The optical mode in a fiber, on the other hand, has a spotsize in the 5–10 μm range. As the laser and fiber spotsizes are not compatible, lenses are required in order to ensure a reasonably efficient coupling of light from a semiconductor laser into a fiber.

The laser cavity forms a resonator, and the cavity length L and the effective index n_{eff} are related to the lasing wavelength λ by

$$n_{\text{eff}}L = \frac{M\lambda}{2} \tag{3}$$

which states that the optical length of the cavity is an integer number of half-wavelengths, where $M \gg 1$ is known as the (longitudinal) *modenumber*. The separation, *modespacing*, between two wavelengths (*longitudinal modes*) satisfying this condition (corresponding to modenumbers M and $M+1$) is

$$\Delta\lambda = \frac{\lambda^2}{2n_{\text{eff}}L} \tag{4}$$

A typical value for the modespacing (for a cavity length of about 300 μm) is about 1 nm. For a laser operating at a wavelength of about 1550 nm, this corresponds to a spacing between the optical frequencies of about 120 GHz.

As is indicated in Fig. 2, the lasing transition in a semiconductor laser is between *energy bands*, rather than between discrete energy levels. An important consequence of this is that the gain curve is quite wide, much wider than the modespacing given by Eq. (4). This wide gain can lead to simultaneous lasing in several longitudinal modes, thereby giving an effective spectral width of several nanometers. Such a wide spectral width will lead to

dispersion problems for systems operating at high data rates over a long length of dispersive fiber.

Another characteristic feature is that the gain levels can be very high. This means that the lasing condition, as expressed in Eq. (1), can be satisfied even for low values of the reflectivity R. Since semiconductors typically have refractive index values around 3.5, sufficient reflectivity (about 30%) can be obtained from a cleaved facet, and there is no need for special high reflectivity external mirrors as is the case for other laser types.

Two material systems are of particular importance for semiconductor lasers. The first is $Ga_{1-x}Al_xAs/GaAs$. Substituting a part of the group III element Ga by Al gives a material with a wider bandgap, but with nearly the same lattice constant. This means that structures containing varying amounts of Al can be grown on GaAs substrates without lattice mismatching. A band structure as shown in Fig. 2 is obtained by having a lower Al fraction in the active layer than in the confinement layers. Lasers based on this system usually operate in the 800–900-nm spectral region (the exact wavelength depends on the Al content in the active layer), and are widely used in CD players.

The material system of particular importance for fiberoptics is $In_{1-x}Ga_xAs_{1-y}P_y/InP$. By using two group III elements and two group V elements, there are 2 degrees of freedom in the composition. The first can be used to ensure lattice matching to an InP substrate, and the second can be used to adjust the bandgap. These materials are used in lasers for the important 1300-nm (minimum dispersion) and 1550-nm (minimum loss) fiber communication wavelength regions. Figure 3 shows a schematic of an InP-based communication laser.

The layers in a laser structure are normally grown by the *MOVPE* (metal-organic vapor-phase epitaxy) process, which allows the deposition of thin layers that are uniform in both thickness and material deposition. The active stripe is formed by a combination of photolithographic and etching processes followed by an overgrowth, and metallic contacts are formed to the n-type and p-type sides of the laser. Pumping of the upper laser level is performed simply by passing a current through the structure.

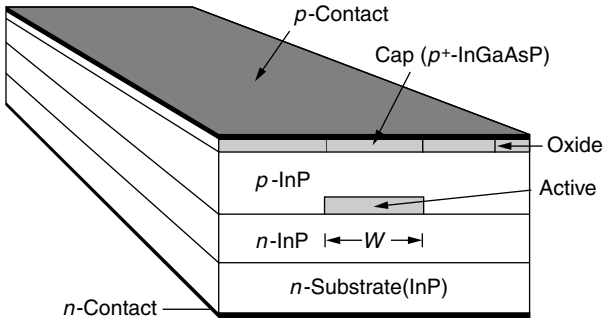


Figure 3. Diagram of communication laser. Typical dimensions are as follows: cavity length 300 μm, total width 100 μm, active region width (W) a few micrometers, substrate thickness 100 μm, active layer thickness a few tenths of a micrometer.

A semiconductor laser acts as a threshold device. For low values of the current, there is insufficient gain to satisfy the lasing condition, Eq. (1), and no laser light is emitted. Lasing starts as soon as the current is sufficient to give enough gain (this is known as the *threshold current,* I_{th}), and above this current level the output laser power increases in proportion with $(I - I_{\text{th}})$. Typically several milliwatts of optical power is emitted for a current in the range 10–100 mA. See Refs. 3–5 for more details on laser structures, and Ref. 6 for more advanced devices.

4.3. Laser Dynamics

As carriers (electrons and holes) are injected into the active region, they can recombine either spontaneously or by stimulated recombination brought about by the photon density in the active region. The photon density, on the other hand, is subject to both gain, due to the stimulated recombination of the carriers, and to losses, either internal losses or losses due to photons being emitted from the end facets of the laser. The interactions between the carrier density, N, and the photon density, S, is described by the so-called *rate equations* for the laser. In their simplest form these equations can be written as

$$\frac{dN}{dt} = \frac{I}{eV} - \frac{N}{\tau_s} - GS \tag{5}$$

$$\frac{dS}{dt} = GS - v_g\,(\alpha_{\text{int}} + \alpha_{\text{end}})\,S + \beta\frac{N}{\tau_s} \tag{6}$$

Equation (5) gives the time dependence of the carrier density. The first term on the right-hand side (RHS) is the pump term, where I is the current supplied to the laser, e is the unit charge, and V is the active volume of the laser. The second term accounts for spontaneous recombination, where τ_s is the spontaneous lifetime. Finally, the last term accounts for stimulated recombination, where G is the gain (per unit time). The second rate equation, Eq. (6), describes the time dependence of the photon density. The first term on the RHS is recognized as the stimulated recombination term. The second term accounts for losses, where α_{int} is the internal loss coefficient and α_{end} describes facet losses, both loss coefficients are losses per unit length, and v_g is the group velocity of the light in the laser. The final term occurs because a fraction β of the spontaneous emission events add a photon to the lasing mode.

The gain factor G is related to the gain per unit length in the active region, g_{act}, by

$$G = v_g \Gamma g_{act} \tag{7}$$

where the confinement factor Γ accounts for the fact that the laser active layer forms an optical waveguide with some of the power propagating outside the active region. The gain in the active region in turn is an increasing function of the carrier density N.

The facet loss α_{end} is caused by light being emitted from the ends of the laser. From Eq. (1) α_{end} can be found from

the gain required to balance the loss of photons through the facets

$$\alpha_{\text{end}} = \frac{1}{L} \ln\left(\frac{1}{R}\right) \tag{8}$$

It is a unique feature of a semiconductor laser that it can be modulated directly by varying the current [first term on the RHS of Eq. (5)], and the response of the laser can be found from the rate equations. Since these equations are nonlinear, due to the dependence of the gain on the carrier density, the rate equations cannot in general be solved analytically; however, a number of important results can still be derived from them. In the case of weak modulation, where the current I consists of a bias current plus a superimposed small-signal modulation current, the rate equations can be linearized. The result of this analysis shows that for low frequencies the optical output power will be modulated in proportion to the modulation current. For higher modulation frequencies the laser response has a resonance, with the resonance frequency increasing roughly in proportion to the square root of $(I - I_{\text{th}})$. The resonance frequency is typically in the gigahertz range. For modulation frequencies above the resonance frequency the laser response drops off rapidly. The resonance frequency provides a reasonable estimate on how fast the laser can be modulated directly, assuming that the laser response is not deteriorated by parasitic elements (such as the laser series resistance and parallel capacitance).

Other results that can be derived from the rate equations include harmonic distortion [7], and approximate expressions for the (large signal) turnon and turnoff times [8]. As spontaneous emission does not occur at a constant rate, but is a statistical process, it is possible to derive results on the laser intensity noise and its spectral distribution. Readers should consult Ref. 9 for more details on modulation and noise properties of semiconductor lasers.

4.4. Single-Frequency Lasers

In order to reduce the dispersion in optical fibers, it is necessary to restrict lasing to a single longitudinal mode. The conventional way to achieve this is by incorporating a periodic structure (*grating*) in the laser, as shown schematically in Fig. 4. In this structure the grating provides internal reflections at a wavelength determined by the grating period. This type of laser is known as a *distributed-feedback laser* (DFB).

The wavelength selected by the grating is given by

$$\lambda_{\text{DFB}} = 2n_{\text{eff}}\Lambda \tag{9}$$

where n_{eff} is the effective index and Λ is the grating period. Since the grating only provides efficient internal feedback for wavelengths very close to λ_{DFB}, any wavelength different from λ_{DFB} will have a higher rate of loss through the end facets, and as a result lasing will occur at λ_{DFB}. The discrete (and nonselective) reflections from the end facets will interfere with the distributed reflection from the grating, and usually the reflection from the front facet is suppressed by applying an *antireflection* (AR) coating.

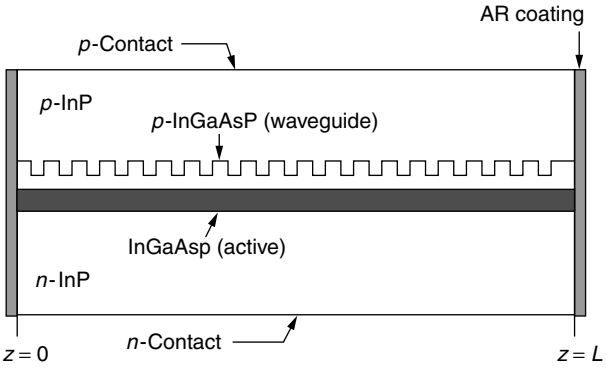


Figure 4. Outline of a DFB laser. The waveguide layer has a bandgap (and consequently a refractive index) between that of the active layer and InP.

The detailed theory for DFB lasers is rather involved, and readers are referred to Refs. 10 and 11 for more information on this topic.

Gratings can be fabricated by covering the waveguide layer with a photoresist, which is then exposed to an optical interference pattern, and the developed resist is used as an etch mask. After grating etching the remaining layers in the laser structure are grown.

The refractive index of a laser structure is quite sensitive to temperature; according to Eq. (9), this will lead to a temperature dependence of the lasing wavelength. A typical value is about 0.1 nm per degree (corresponding to a change in the optical frequency of about 10 GHz per degree). Whereas temperature tuning can be used to trim the wavelength to a given value, the temperature dependence also means that in order to ensure that the lasing frequency is within 10 GHz of a given value, the laser temperature has to be stabilized to within 1 degree.

In WDM systems signals from several lasers are multiplexed before transmission, and in order to ensure interoperability of equipment from different manufacturers, ITU has set a standard for optical transmission frequencies. This standard is based on a frequency grid with a 100-GHz spacing. Consequently the range from 192.1 THz (=1560.61 nm) to 195.9 THz (=1530.33 nm) consists of 39 channels.

4.5. Wavelength-Selectable Lasers

In order to achieve single-frequency lasing at a number of different wavelengths, arrays of DFB lasers can be formed. If these lasers are integrated with a combiner, several optical signals can be coupled into the same fiber. However, simultaneous operation of several closely spaced lasers will lead to crosstalk problems, and it may be more advantageous to consider *selectable* structures, where only one laser is operated at any time. An example of such a structure is shown in Fig. 5. This approach allows redundancy by having more than one laser per wavelength. The exact optical frequency is achieved by temperature tuning.

A different type of array is constructed by *cascading* of DFB lasers (several DFB lasers on a common optical

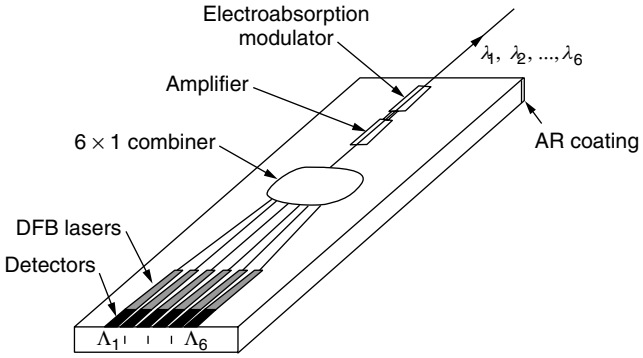


Figure 5. Selectable array with six DFB lasers and a combiner [5]. This *optoelectronic integrated circuit* also contains an amplifier to compensate for the combiner loss, a modulator for data encoding and monitor detectors for the lasers [12].

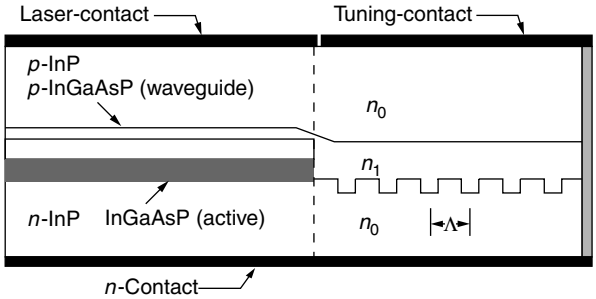


Figure 6. Two-section DBR laser. The output power is controlled by the laser current supplied to the active region. The tuning current supplied to the Bragg reflector region controls the carrier density in that region, and hence its refractive index. According to Eq. (9), this in turn tunes the wavelength at which the grating gives efficient reflection. Tuning ranges can be up to 10–15 nm [e.g., 14].

waveguide). By operating a single laser element above threshold and the other lasers close to threshold, the lasing wavelength is determined by the grating in the element operated above threshold. Several branches, each with several lasers, can be combined, and by using a high degree of temperature tuning, a wavelength range of 30 nm has been covered [13].

The wavelength range that can be covered by an array is limited by the number of array elements and by the degree of temperature tuning. For applications where many optical frequencies are required, or where a high degree of temperature tuning is undesirable, arrays may not be the best solution.

4.6. Tunable Lasers

The wide optical gain curve in a semiconductor laser makes it possible to achieve tuning of the lasing wavelength. As already mentioned, tuning is possible by changing the operating temperature. However, unless a large temperature variation is allowed, the tuning range will be limited to a few nanometers in wavelength (a few hundred gigahertz in frequency), and thermal tuning is comparatively slow (microsecond–millisecond range).

The fact that the refractive index depends on the carrier density can be applied for tuning. However, in a simple structure (such as Fig. 3 or 4), the carrier density is clamped to the value which is required to give sufficient gain to satisfy the lasing condition, and tuning by carrier density changes is not possible. This limitation can be overcome by using structures with two (or more) separate regions. One example is the *distributed Bragg reflector* (DBR) laser shown in Fig. 6. The tuning speed will be limited by the carrier lifetime in the tuning region (nanosecond range).

It should be noted that the tuning of a DBR laser is not continuous, but shows jumps between the wavelengths that satisfy the resonance condition given by Eq. (3).

Whereas the tuning range of a two-section DBR laser is limited by the extent to which the refractive index of the tuning section can be changed, wider tuning ranges can be achieved using somewhat more complicated structures.

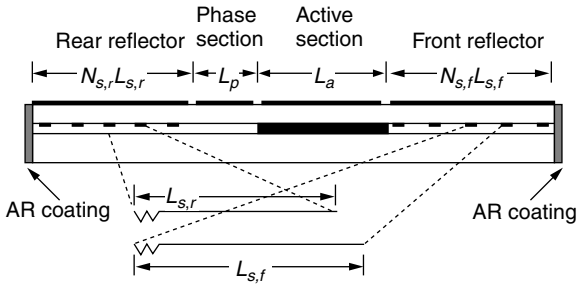


Figure 7. Sampled-grating DBR. Instead of a continuous grating the two reflectors have sampled gratings. These gratings give reflection spectra that have a comb of reflection peaks with a spacing determined by the sampling period. By using two different sampling periods, two reflection combs with different periodicities are obtained.

An example of such a structure is the *sampled-grating DBR* (SGDBR), which is shown in Fig. 7.

A small change of the refractive index of one of the tuning sections gives a large change in wavelength since a new pair of reflection peaks will coincide. This behavior is recognized as the *Vernier effect*; see Ref. 15 for more details. This principle leads to greatly enhanced tuning ranges; up to about 100 nm has been reported. Similar tuning ranges have also been achieved by combining a tunable codirectional coupler with a sampled grating [16].

A particular problem with the tunable laser structures described above is that several control currents are required for a specific combination of power and wavelength. Tunable lasers must therefore be characterized in sufficient detail to identify the current combinations required for various wavelengths, and the laser driver electronics must contain this information in such a form that tuning can be achieved in response to simple external instructions.

Other semiconductor laser structures capable of very wide tuning include external cavity lasers and vertical cavity lasers (see Section 4.7). Ultimately, the tuning range is of course limited by the width of the gain curve.

More details on various tunable laser types can be found in Ref. 17.

4.7. Vertical Cavity Surface-Emitting Lasers

In a *vertical cavity surface-emitting laser* (VCSEL), the direction of lasing is perpendicular to the active layer. Since this means that the active cavity is very short, it follows from Eq. (8) that very high end-reflectivities are required. Such high reflectivities can be achieved by having a stack consisting of a large number of layers with alternating high and low refractive index.

It is a considerable advantage of the short cavity that only one longitudinal mode exists because of the resulting wide modespacing [cf. Eqs. (3) and (4)]. This means that a VCSEL by its nature is a single frequency laser. Other major advantages include: the possibility of matching the laser spot size to that of a fiber, making coupling simpler and more efficient, and the use of on-wafer testing in the fabrication process. See Ref. 18 for more details and a review.

The various advantages of GaAs VCSELs make them highly suitable as relatively low-cost transmitters in short-distance systems operating at relatively short wavelengths, such as data links. The technology for VCSELs operating at the "telecoms" wavelength of 1300 and 1550 nm has proved to be considerably more difficult. One possible way of overcoming some of the problems is the use of 980-nm lasers for optical pumping as an alternative to electrical pumping.

Tunable VCSELs have been fabricated by incorporating an electrostatically deformable reflecting membrane at one end of the laser. A tuning range of up to 50 nm is then achieved by a simple voltage control of the cavity length [19].

4.8. Related Optical Components

A number of optical components are related to semiconductor lasers, because they are either of a similar structure or used together with semiconductor lasers:

Pump Lasers. Fiber amplifiers, used in long-haul linkage, require high-power optical pumping at specific wavelengths, usually 980 or 1480 nm. The pump power is supplied by specially designed high-power semiconductor lasers.

Semiconductor Optical Amplifiers (*SOAs*). These amplifiers are an alternative to fiber amplifiers and are very similar to lasers in structures. However, lasing is suppressed because SOA facet reflectivity is very low. Whereas SOAs can be integrated with other semiconductor components (see Fig. 5), the low coupling efficiency to fibers makes them less attractive for use as inline amplifiers in transmission systems.

Modulators. At data rates of 2.5 Gbps, directly modulated semiconductor lasers can be used, but direct modulation becomes increasingly problematic as the data rate increases, thus making dedicated modulators attractive in high data rate systems. Modulators are made from $LiNbO_3$ or semiconductors. Semiconductor-based electroabsorption modulators may be integrated with other elements, including lasers (see Fig. 5).

4.9. Packaging and Modules

In order to provide a fixed and robust coupling from a semiconductor laser to an optical fiber, the laser must be supplied in a suitably designed package. In addition to the *coupling optics*, a laser package may contain several of the following additional elements:

An *optical isolator* to prevent instabilities in the laser operation due to external reflection

A *thermoelectric element* to keep the laser temperature constant and prevent wavelength drift caused by variations in the ambient temperature

Drive electronics to provide bias current and modulation, in the case of external modulation a separate modulator may also be included

A *monitor detector* for control of the optical power from the laser

An example of a laser module is shown in Fig. 8.

BIOGRAPHY

Jens Buus was born in 1952 in Copenhagen, Denmark. He is an electrical engineer (MSc in electrophysics), graduated from the Technical University of Denmark (DTU) in August 1976. He also holds Lic. techn. (Ph.D.) and Dr. techn. (DSc) degrees from DTU. From 1979 to 1983 he was a postdoctoral fellow at DTU; from 1983 to 1992 he was with Marconi Caswell (formerly Plessey Research Caswell). Since January 1993 he has been a consultant at Gayton Photonics Ltd., United Kingdom. He

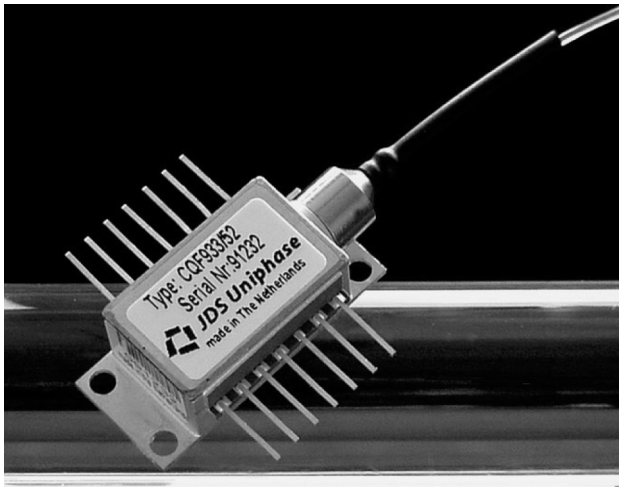


Figure 8. Packaged laser. This unit contains a fixed-frequency DFB laser (one of the standard ITU frequencies) and is designed for operation at a data rate of 2.5 Gbps. The package length is 30 mm, and the pins provide access to temperature control, monitor diode, as well as DC and AC input to the laser. (Courtesy of JDS Uniphase.)

has been project manager in the European RACE and ACTS programs and is currently project manager for a project under the IST program. Dr. Buus has served on several conference committees and given invited talks, tutorials, and short courses at several conferences; he has authored or coauthored about 60 papers, over 60 conference papers, and 2 books. During the academic years 1998–2000 he was a LEOS Distinguished Lecturer. He is a fellow of the IEEE, and a member of the Optical Society of America, the Institute of Electrical Engineers, and of the Danish Physical Society. His research has included contributions to the understanding of the properties of semiconductor lasers and optical waveguides, as well as contributions to work on gratings, integrated optics, and coherent optical communication.

BIBLIOGRAPHY

1. H. C. Casey and Jr., M. B. Panish, *Heterostructure lasers, Part A: Fundamental Principles*, Academic Press, Orlando, FL, 1978.
2. Special Issue on Semiconductor Lasers, *IEEE J. Quant. Electron.* **QE-23** (June 1987).
3. S. L. Chuang, *Physics of Optoelectronic Devices*, Wiley, Chichester, UK, 1995.
4. L. A. Coldren and S. W. Corzine, *Diode Lasers and Photonic Integrated Circuits*, Wiley, Chichester, UK, 1995.
5. G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*, Van Nostrand-Reinhold, New York, 1993.
6. P. S. Zory, ed., *Quantum Well Lasers*, Academic Press, Boston, 1993.
7. T. E. Darcie, R. S. Tucker, and G. J. Sullivan, Intermodulation and harmonic distortion in InGaAsP lasers, *Electron. Lett.* **21**: 665–666 (1985).
8. R. S. Tucker, Large-signal switching transients in index-guided semiconductor lasers, *Electron. Lett.* **20**: 802–803 (1984).
9. K. Petermann, *Laser Diode Modulation and Noise*, Kluwer, Dordrecht, The Netherlands, 1988.
10. G. Morthier and P. Wankwikelberge, *Handbook of Distributed Feedback Laser Diodes*, Artech House, Norwood, MA, 1997.
11. J. E. Carroll, J. E. A. Whiteaway, and R. G. S. Plumb, *Distributed Feedback Semiconductor Lasers*, IEE, Stevenage, UK, 1998.
12. M. G. Young et al., Six wavelength laser array with integrated amplifier and modulator, *Electron. Lett.* **31**: 1835–1836 (1995).
13. J. Hong et al., Matrix-grating strongly gain-coupled (MG-SGC) DFB lasers with 34 nm continuous wavelength tuning range, *IEEE Photon. Technol. Lett.* **11**: 515–517 (1999).
14. F. Delorme, S. Grosmaire, A. Gloukhian, and A. Ougazzaden, High power operation of widely tunable 1.55 μm distributed Bragg reflector laser, *Electron. Lett.* **33**: 210–211 (1997).
15. V. Jayaraman, Z.-M. Chuang, and L. A. Coldren, Theory, design, and performance of extended tuning range semiconductor lasers with sampled gratings, *IEEE J. Quant. Electron.* **29**: 1824–1834 (1993).
16. P.-J. Rigole et al., 114 nm wavelength tuning range of a vertical grating assisted codirectional coupler laser with a super structure grating distributed Bragg reflector, *IEEE Photon. Technol. Lett.* **7**: 697–699 (1995).
17. M.-C. Amann and J. Buus, *Tunable Laser Diodes*, Artech House, Norwood, MA, 1998.
18. K. Iga, Surface-emitting laser — its birth and generation of new optoelectronics field, *IEEE J. Select. Top. Quant. Electron.* **6**: 1201–1215 (2000).
19. D. Vakhshoori et al., 2 mW CW singlemode operation of a tunable 1550 nm vertical cavity surface emitting laser with 50 nm tuning range, *Electron. Lett.* **35**: 900–901 (1999).

OPTICAL SWITCHES

K. L. EDDIE LAW
University of Toronto
Toronto, Ontario, Canada

1. INTRODUCTION

Optical transport networks have been deployed around the world for many years. As an article [1] in *Nature* indicated, the theoretical maximum bandwidth of a typical optical fiber in access networks is estimated to be about 150 Tbps (terabits per second). Even though it may be hard to estimate if this "glass ceiling" of 150 Tbps will actually hold, the speed of commercial transport systems has already been reaching 40 Gbps (i.e., OC[1]-768) in synchronous optical networks (SONETs). On the other hand, dense wavelength-division multiplexing (DWDM) systems can deliver information in a number of wavelengths in one optical fiber. NEC[2] was successful in transmitting 10.92 Tbps in a 117-km-long fiber with 273 wavelength channels at 40 Gbps per channel data rate. With 40 Gbps channels, we need 3750 channels to reach this fiber bandwidth limit. In the case that if we will be able to build a futuristic 160-Gbps channel, 900 channels will suffice to reach this limit. The technology of the optical transport system has been evolving rapidly. Obviously, the switching nodes are the bottlenecks in today's optical networks. Without optical logic technology, the switching nodes need to undergo signal conversions from photons to electrons in order to switch packets to their respective outgoing ports. Thereafter, the packets will be converted and delivered in the form of photons. As of today, all SONET optical switching systems carry out optical–electrical–optical (OEO) conversions.

[1] OC-N stands for optical carrier digital signal rate of $N \times$ 53 Mbps in SONET.

[2] NEC announced the DWDM transmission capacity world record on March 22, 2001. Exactly one year later, Lucent announced the transmission distance world record of 4000 km with 64 channels running 40 Gbps on March 22, 2002.

Active research on optical switches has been carried out with the goal of constructing all-optical networks that do not require any OEO conversions. A switching core may need to switch information from L incoming fibers to L outgoing fibers. Each fiber carries multiple, W, wavelength channels of information with the DWDM technology. Therefore, the designs of the optical cross-connects (OXCs) are getting complicated with the rapidly increasing number of wavelength channels per fiber. The architectural designs of OXCs are in two dimensions that involve the space and wavelength-switching domains [2,4–6]. That is, an ultimate OXC design should switch information from one particular wavelength channel at an input port to a specific output port with a selected outgoing wavelength channel. Therefore, given an $N \times N$ optical cross connect design, then $N \geq L \cdot W$ is required to provide an internally nonblocking switching matrix for any wavelength channel to any fiber. Figure 1a shows a system design of an OXC

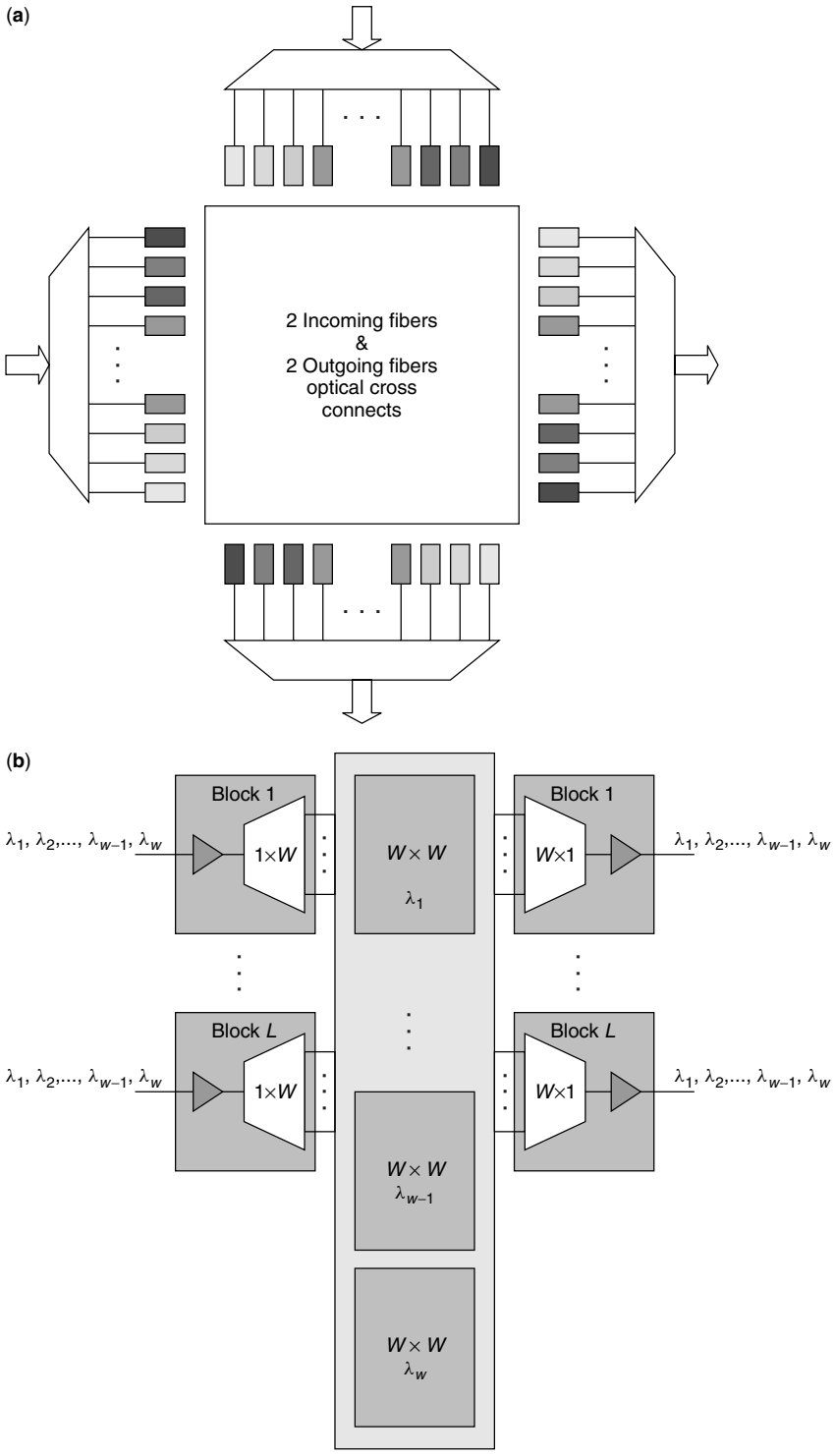


Figure 1. (**a**) Nonblocking OXC; (**b**) blocking 3-stage OXC.

with two incoming fibers and two outgoing fibers. We need to install wavelength filters and converters to produce a flexible design. Unfortunately, the cost of those converters is expensive. Traditionally, the Clos networks [3] are nonblocking multistage circuit-switched networks in the electronic domain. The alternative multistage design, as shown in Fig. 1b, has lower cost without any wavelength converters; however, it is not a nonblocking architecture. In this article, we discuss different architectural designs of the OXCs from a system perspective with consideration for device constraints.

2. DESIGN CONSTRAINTS

The performance of the next-generation all-optical networks relates to the optical properties and the functionalities of optical devices, and hence the throughput performance of the resulting OXC architectures. With different characteristics of optical devices, the selection criteria of the device components will definitely affect both the design architecture and the performance of the resulting optical cross-connects. Currently, the research on OXCs is still at an early phase. There are several technologies that are generally recognized to have the potential to construct the next-generation optical switches. They include the electrooptic, thermooptic, acoustooptic, thermobubble, liquid crystals, optomechanical, and beam-steering technologies. Among the initial investigations, some device level designs have already demonstrated that they are excellent candidates to be the basic building blocks for constructing the next-generation OXCs.

Before we describe the designs of optical switches, we consider several limiting factors of optical devices that may affect the architectural designs. With the wide bandwidth for transporting optical signals in fibers, the switching rate in OXC will be a limiting factor on the signal transfer rate. The switching time is determined through the rate of changing states on forwarding or detouring optical signals in devices. It may fall in the range from nanoseconds to milliseconds. The smaller the switching time, the better it is. It determines the information transfer rate in terms of bits, packets, and bursts. Moreover, it also indicates if the resulting optical networks can operate with circuit-switching, packet-switching, or burst-switching technology. Apart from switching time, port count, reliability, size, and cost are important design criteria with the space and wavelength switching in optical networks.

The overall loss budget is an important criterion that determines the power consumption and the placements of the optical amplifiers. There are different loss factors that include insertion loss, crosstalk, chromatic dispersion, polarization-dependent loss (PDL), and polarization-mode dispersion (PMD). Moreover, some architectures may be wavelength-dependent, and some may have a wide variation of losses between ports. In general, the maximum loss budget for an OXC should be around 25–30 dB. Therefore, it is seldom justifiable to create multistage networks if the per stage module has a high loss factor; otherwise, we need to provide signal amplification between stages. In the following, we focus on reviewing some important parameters: the insertion loss, crosstalk and switching time.

2.1. Insertion Loss

Whenever a photonic device is introduced in a lightpath, it introduces an insertion loss due to the mismatch at the interface. Ideally, it should be as small as possible in order to minimize the total loss budget, especially if the optical signal needs to travel through multiple OXCs. It is also important to determine a switching module that has uniform loss distribution with different interconnection patterns. If the insertion loss changes with different interconnection patterns, then a variable equalizer is required between switching stages. The resulting design is undesirable for it complicates the system control and increases the cost of the switch.

Free-space optical systems have the lowest insertion loss. For example, the microelectromechanical systems (MEMSs) belong to the optomechanical design class. The insertion loss of a MEMS OXC can reach as low as 1 dB. It is usually in the range from 1 to 6 dB depending on the size of the MEMS switch. Liquid crystal electrooptic is another switching technology for constructing OXCs. Its insertion loss is comparable to the MEMS switch as it can also reach 1 dB loss [23]. However, polarization loss may occur in the liquid crystal module. It relates to the Fresnel reflection on the glass–air interface and it can be as high as 3 dB. The Fresnel reflection occurs at a planar junction of two materials that have different refractive indices and is not a function of the angle of incidence. There are currently no reports on building large-scale liquid crystal switches.

For the other device technologies, for example, the insertion loss of thermooptic switches is usually in the range of 6.6–9 dB;[3] the lithium niobate ($LiNbO_3$) electrooptic switch[4] has an insertion loss of <9 dB. There has been some steady progress on improving both of these technologies. We can find moderately sized OXCs with these technologies in the market. However, the insertion loss is still considered to be comparatively high for next-generation large-scale OXCs. As a result, the $LiNbO_3$ switch is usually used in the external modulation rather than in the lightpath routing. This is because the external modulated information is usually amplified before it is transmitted.

2.2. Crosstalk

Crosstalk may be caused by either interference from signals on different wavelengths, the interband crosstalk, or interference from signals on the same wavelength on another source, the intraband crosstalk. Interband usually determines the channel spacing. Intraband crosstalk usually occurs in switching nodes where multiple signals

[3] The insertion loss of the 8×8 switch using a thermooptic Mach–Zehnder interferometer is <8 dB from NTT Electronics at *http://www.nel-dwdm.com/profile/profile.html*. Its switching speed is < 3 ms.

[4] The insertion loss of the 8×8 crossbar switch using $LiNbO_3$ planar lightwave circuit is <9 dB from Lynx Photonic Networks at *http://www.lynxpn.com/*. Its switching speed is <5 ns.

on the same wavelength are being switched from different inputs to different outputs. The degree of intraband crosstalk depends on the switch architectures.

In an optical device, crosstalk happens when a portion of the input signal "leaks" into another signal as they copropagate through the switch fabric. The ratio of the power at the unselected output port over the total input power in a switch element is referred to as the *crosstalk ratio* of the switch, since crosstalk is the noise usually introduced from the nearby connections. Therefore, crosstalk is usually more serious if the switch architecture design is complicated, especially if it has a large number of ports and connections. Since crosstalk measures the power of the loss signal to the input signal power [4–6], it is desirable if the value of crosstalk is as negative as possible in decibels.

With free-space optomechanical designs, MEMS optical switches provide the best crosstalk performance among all switching fabric technologies. Its crosstalk is in the range of −55 to −60 dB [29–33]. Besides, liquid crystal provides excellent insertion loss performance, and the crosstalk can reach −48 dB in general. There was a report on constructing an 8×8 crossbar liquid crystal switch with 1×8 switch arrays and the crosstalk could reach −59.5 dB [18,22,23]. This result is comparable to the MEMS switches. Unfortunately, there are still difficulties in building large-scale liquid crystal crossbar switches with the tradeoff between loss uniformity and the crosstalk level. Nevertheless, the liquid crystal switching technique is expected to improve with time, and is considered as a good candidate for building optical switching modules.

For the silica-based thermooptic switch using double-Mach–Zehnder interferometer (MZI) waveguide units, the crosstalk can reach −43 dB through a sophisticated hardware architecture design. There is a tradeoff between hardware complexity and the crosstalk level. To achieve this excellent crosstalk performance, we need to increase the hardware complexity by interconnecting 256 double-MZI units for a 16×16 silica thermooptic switch [12]. The resulting switch had an insertion loss and extinction ratio[5] of 17.5 and 32.9 dB, respectively. It will not be a cost-effective approach to construct an OXC with a large port count with thermooptic waveguide designs.

Some $LiNbO_3$ electrooptic switches are used to construct directional couplers by altering the refractive index of the waveguide with electric energy. These directional couplers were initially considered to have the potential to construct multi-stage switching networks. However, they cannot be used for OXCs because they have poor crosstalk isolation and a large insertion loss. On the other hand, there are $LiNbO_3$ acoustooptic tunable switches (AOTS). Acoustooptic switching technology uses surface acoustic waves to generate birefringence grating and alter the polarization of a lightbeam. Switching occurs at high speed, and it can reach as low as 3 μs.[6] An AOTS introduces about 5–6 dB insertion loss; however, its crosstalk ratio is about −20 dB [11–13]. The crosstalk of a single-channel acoustooptic 1×300 demultiplexer can reach about −35 dB. AOTS suffers both interband and intraband crosstalk. A double-stage devices or weighted coupling schemes may be required to reduce intraband crosstalk. Since interchannel interference may create intrinsic modulation of the transmitted signal, it affects the bit error rate (BIT) performance, and hence the device may not be working for long-haul optical systems. Both of these electrooptic and acoustooptic $LiNbO_3$ devices have yet to demonstrate that they can be used to build large-scale OXCs. Therefore, only small-scale OXCs [13] can be found in the market with these technologies. All in all, the optomechanical switch provides the best performance in crosstalk level compared to the other technologies.

2.3. Switching Time

The switching time describes the time it takes for a switch to establish an interconnection pattern. The desirable value must be as small as possible. As the data rate exceeds 10 Gbps per wavelength, a submicrosecond switching time is necessary to provide dynamic path provisioning, grooming, and path restoration on failure. This is an important parameter that determines the performance of future optical networks. Today's optical core networks are configured statically. When the optical device is able to switch states actively, future optical core networks are expected to provide dynamic path routing capability.

Among all the switch fabric technologies, the electrooptic switches have the fastest switching time compared to the other two technologies. The $LiNbO_3$ and semiconductor optical amplifier (SOA) switches are able to switch in the range of nanosecond response time. As discussed before, the $LiNbO_3$ device is suitable only for providing external modulations.

On the other hand, the thermooptic switch offers a switch time within the range of 1 ms. This response time is acceptable in optical path-switching applications. Unfortunately, the thermooptic switch's insertion loss is also considered to be too high when designing a large-scale switch. On the other hand, the mechanical switches, for example, fiber bundle switches [7], can achieve a good crosstalk level as well as low insertion loss. However, these fiber bundle mechanical switches usually have slow switching times. Fortunately, with the introduction of MEMS optical switching systems, apart from having excellent crosstalk ratio and low insertion loss, good mechanical design can also lead to good switching time, such as 700 μs [36]. At the moment, MEMS becomes the most appealing switch fabric technology for designing large-scale OXCs for future optical networks. In Table 1, we outline the characteristics of different available optical device technologies for building OXCs.

On concluding this part, we would like to outline the basic requirements for designing the OXCs as follows. The design should have (1) low insertion loss (typically <1 dB), (2) low crosstalk (typically <-50 dB), (3) low polarization-dependent loss (PDL), (4) switching time faster than or, at least, equal to millisecond range, (5) low power consumption, (6) long-term reliability, (7) small size, (8) low cost, (9) scalability to large port count, and

[5] The extinction ratio is defined as the ratio of the optical power transmitted for a bit "0" to the power transmitted for a bit "1."

[6] This is the 1×300 demultiplexer reported by the Light Management Group at *http://www.lmgr.net/*.

Table 1. Comparisons Between Different Optical Device Technologies

	Free-Space		Guided-Wave Integrated Optics		Guided-Wave Active Component
	MEMS	Liquid Crystal	Thermooptic, Bubble	Electrooptic	Semiconductor Optical Amplifier
Switching time	1–10 ms	2–5 ms[a]	1–10 ms	nsec	nsec
Insertion loss	Very good	Moderate	Moderate	Moderate	Acceptable
Crosstalk	Very good	Average	Average	Acceptable	Acceptable
Polarization-dependent loss	Good	Good	Good	Acceptable	Acceptable
Wavelength dependence	Good	Good	Average	Average	Acceptable
Bit rate transparency	Good	Good	Good	Good	Good
Power consumption	Good	Good	Bad	Good	Bad
Expandability/size	Large	Moderate	Small	Small	Small

[a]There was a report that could have a switching time of ~35 μs. However, the commercial products are usually in microseconds.

(10) self-holding or latching mechanism design. Since the total loss should be less than 30 dB, it then also limits the number of cascaded modules in the architectural design.

3. PROMISING SWITCHING FABRIC TECHNOLOGIES

Optical cross-connects can be classified into two broad classes: active and passive. Without optical logic devices, today's optical network can only offer a high-speed and large capacity transport system. Hence the optical routing paths are comparatively static and mostly preconfigured through the optical add/drop multiplexers (OADMs). In the near future, we expect higher deployments of the passive OXCs. These passive OXCs are usually designed with traditional doped waveguide technology. The switching characteristics are predefined and fixed; that is, an output signal pattern depends on the architectural design of a passive OXC and a specific arriving input signal pattern. In contrast, the switching points in active OXCs should be set according to the destination ports of the incoming signals. Most of the available technologies for constructing active OXCs are listed in Table 1. In Section 3.1, we will discuss the design of passive OXCs using arrayed waveguide gratings (AWGs). The technologies for constructing active OXCs is different from those for passive OXCs, and the designs can be found in Section 3.2.

3.1. Passive Optical Cross-Connects

Bulk optic or all-fiber filters and devices are used in a number of WDM applications. With the improvement of technology, the trend is to move toward monolithic integration of devices and components. One of those generic devices is the *arrayed waveguide grating* (AWG) multiplexer, also known as *waveguide grating router* (WGR). DWDM networks permit large capacity optical signal transfer. AWG can be used to split and combine optical signals of different wavelengths in the systems. The silica AWG allows the fusion splice of fiber to chip; however, it has low-contrast waveguide structure and its size is relatively large [8]. The AWG device can also be fabricated on indium phosphide (InP) [8] that provides high-index contrast and it is suitable for large-scale system integration. As shown in Fig. 2, an AWG [4–6,9] consists of two free-space couplers connected by a grating array. The first coupler has N inputs and N' outputs (where $N \ll N'$), while the second one has N' inputs and N outputs. For the first coupler, there is a regular angular distance between any adjacent input ports. Similarly, there is also another regular angular separation between any adjacent output arrayed waveguides. The setup of the second coupler is simply a mirror image of the first one. The grating array between couplers consists of N' waveguides, with lengths $l_0, l_1, \ldots, l_{N'-1}$, where $l_0 < l_1 < \cdots < l_{N'-1}$. The length difference between any two adjacent waveguides is constant. The constant difference in the lengths of the waveguides creates a phase difference in adjacent waveguides. This phase shift depends on the propagation constant in the waveguide, the effective refractive index of the waveguide, and the wavelength of the light. At the input of the second star coupler, the phase difference in the signal will be such that the signal will constructively recombine only at a single output port.

With this design, two signals of the same wavelength coming from two different input ports will not interfere with each other in the grating because of an additional phase difference created by the distance between any two input ports. The two signals will be combined in the grating but will be separated again in the second coupler and directed to different outputs. AWGs have been successfully demonstrated for a number of WDM enabling devices that include multiplexers, demultiplexers, channel dropping filters/equalizers, and tunable lasers.

Several important fundamental properties of AWGs enable the construction of passive OXCs [10]:

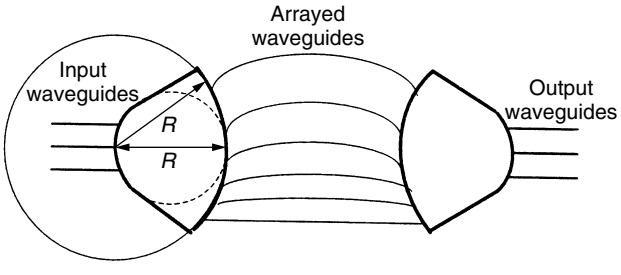


Figure 2. Arrayed waveguide grating (AWG).

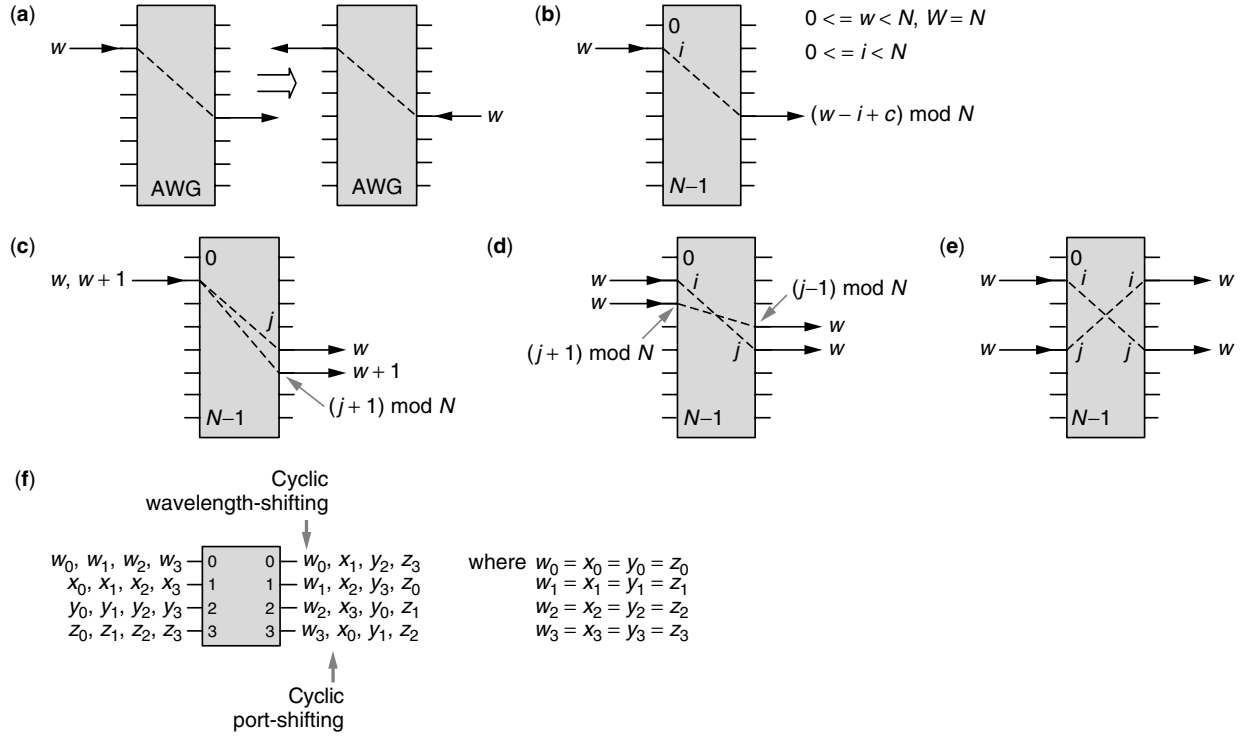


Figure 3. Routing properties of AWG.

1. *Reciprocity* (Fig. 3a). If a signal of a wavelength propagates from one input port to an output port, then any signal of the same wavelength injected at that output port will propagate backward to the input port in exactly the same way.
2. *Periodicity in Frequency* (Fig. 3b). A given frequency bandwidth may contain a number of wavelength channels. If all these wavelength channels in that frequency range follow the same transfer function in a device, then this frequency period is known as *free spectral range* (FSR). An $N \times N$ AWG has N input and output ports, and supports W wavelength channels, denoted by the sets $\mathcal{N} = \{0, 1, \ldots, N-1\}$, and $\mathcal{W} = \{0, 1, \ldots, W-1\}$ in an FSR, respectively. In general, $\mathcal{W} = \mathcal{N}$. Within an FSR, the wavelength channels have constant-frequency spacing instead of constant-wavelength spacing. For the periodicity property, a wavelength signal $w \in \mathcal{W}$ enters an input port $i \in \mathcal{N}$ is delivered to an output port $[(w - i + c) \bmod N]$. This c is an integer known as the FSR constant that depends on the selection of the FSR.
3. *Cyclic Wavelength Shifting* (Fig. 3c). If an input signal of wavelength w leaves AWG from port j, then any input signal of wavelength $w + 1$ entering the same port leaves the AWG from port $[(j + 1) \bmod N]$.
4. *Cyclic Port Shifting* (Fig. 3d). If an input signal enters port i and leaves AWG from port j, then any input signal of the same wavelength entering port $[(i + 1) \bmod N]$ leaves the AWG from port $[(j - 1) \bmod N]$.
5. *Symmetry* (Fig. 3e). If an input signal enters port i and leaves port j, then any input signal of the same wavelength entering from port j will leave the AWG from the port i.

As an example, the wavelength routing assignments of an AWG are shown in Fig. 3f. It is a 4×4 AWG with four wavelength channels in an FSR, specifically, $N = W = 4$. The four incoming ports are identified with w, x, y, and z from top to bottom. Assuming that c is zero and observing the top input port, then for an incoming wavelength numbered zero, w_0, the outgoing port is $[(w - i + c) \bmod N] = 0$, the top output port, with the periodicity property. With the cyclic wavelength-shifting property, we can arrange all four w wavelength channels sequentially at the outgoing ports as shown. Then observing the second top input port, the incoming wavelength numbered zero, x_0, goes to the outgoing port $[(-1) \bmod 4] = 3$ from the cyclic port-shifting property. The other wavelength channels can then be arranged with both the cyclic wavelength-shifting and cyclic port-shifting properties. The resulting wavelength assignment is shown in Fig. 3f. Moreover, it also satisfies the symmetry property.

There is a channel spacing concept in AWGs that allows more flexible wavelength assignments with the AWGs. In the following, a system with channel spacing of k is considered. From architectural point of view, two successive wavelengths that enter the same input port will be routed to two output ports x and $x + k$ with the wavelength-shifting property. From the setup of wavelength channels in an FSR, the two output ports are $[x = (w - i + c) \bmod N]$ and $[x + k = (w + k - i +$

c) mod N] with the periodicity property. This implies a spacing k between two successive wavelength channels at the input. Therefore, a wavelength set $\mathcal{S}$ is said to have a channel spacing of k if $\mathcal{S} = \{(s + ik) \bmod N : 0 \le i < |S|\} \subseteq \mathcal{W}$, for an $s \in \mathcal{S}$.

3.1.1. Compatible Ports. In order to have an AWG to perform as an OXC, we need to explore different routing properties that enable an AWG to do switching. Since one fiber can carry multiple wavelength channels, only some subsets of the N ports should be used for inputs and outputs. Given an $N \times N$ AWG with L incoming fibers, and if each fiber carries a set of wavelength channels $\mathcal{W}$, where $|\mathcal{W}| = W$, then we have $N \ge L \cdot W$ for the AWG to do proper switching.

A set of compatible ports $\mathcal{P}$ is defined with respect to $\mathcal{W}$ if for any pair of wavelengths in $\mathcal{W}$ that enter any two ports in $\mathcal{P}$ will be routed to two different output ports. With this compatible port concept, all incoming signals can be routed to some predetermined outgoing ports, and that will be useful for designing a passive OXC. In the following, two different cases for compatible ports will be examined. Given $\mathcal{N} = \{0, \ldots, N-1\}$, $|\mathcal{W}| = \mathcal{W}$ and $N \ge L \cdot W$ for both of them, $\mathcal{P} = \{p_i \in \mathcal{N} : 0 \le i \le L-1\}$ is defined as a set of L ports for L incoming fibers.

First, considering a channel spacing of 1, if $N = 8$, $W = 4$, and $L = 2$, then we have $\mathcal{P} = \{p_0, p_1 \in \mathcal{N}\}$. One possible arrangement of both the $\{p_0, p_1\}$, where $0 \le p_0 < p_1 < N$, can be found in Fig. 4a if $c = 0$. Mathematically, $\mathcal{P}$ is compatible with respect to $\mathcal{W}$ if and only if $p_i - p_{i-1} \ge W$, for all $1 \le i \le L-1$, and $p_{L-1} - p_0 \le N - W$. In particular, if $N = LW$, then $\mathcal{P}$ is compatible with respect to $\mathcal{W}$ if and only if $\mathcal{P}$ has an equal spacing of W. Second, if the channel spacing is L and N is a multiple of L, then the set of compatible port is $\mathcal{P} = \{p_i, p_j \in \mathcal{N} : p_i \bmod L \ne p_j \bmod L, \forall\, 0 \le i < j \le L-1\}$. The corresponding example can be found in Fig. 4b.

3.1.2. Self-Blocking Ports. With a selected set of compatible ports, only certain numbers of input and output ports are used. If some idle ports can be used for intermediate processing, then the number of AWGs may be possibly reduced to construct an OXC. For this purpose, a set of input ports $\mathcal{P}$ is defined as self-blocking with respect to $\mathcal{W}$ if they are mutually exclusive with their outgoing ports: $\mathcal{P} \cap \{(w - i + c) \bmod N : i \in \mathcal{P}, w \in \mathcal{W}\} = \emptyset$. This implies that the same set of self-blocking input ports of an AWG switch can be used as the output set and should not be used for any intermediate processing. An example shown in Fig. 4c has parameters $c = 1$, $W = 4$, and $N = 5$, and the top link is the self-blocking port.

It is desirable to construct a passive OXC with one AWG having such a set of input ports that is both compatible and self-blocking. For L incoming fibers, a necessary condition for the existence of L self-blocking and compatible ports with respect to $\mathcal{W}$ is $N \ge L \cdot (W + 1)$.

3.1.3. Compatible and Self-Blocking Ports. In order to design a preconfigured architecture with one AWG, it will be desirable to find a set of ports $\mathcal{P}$ that is both compatible and self-blocking with respect to $\mathcal{W}$:

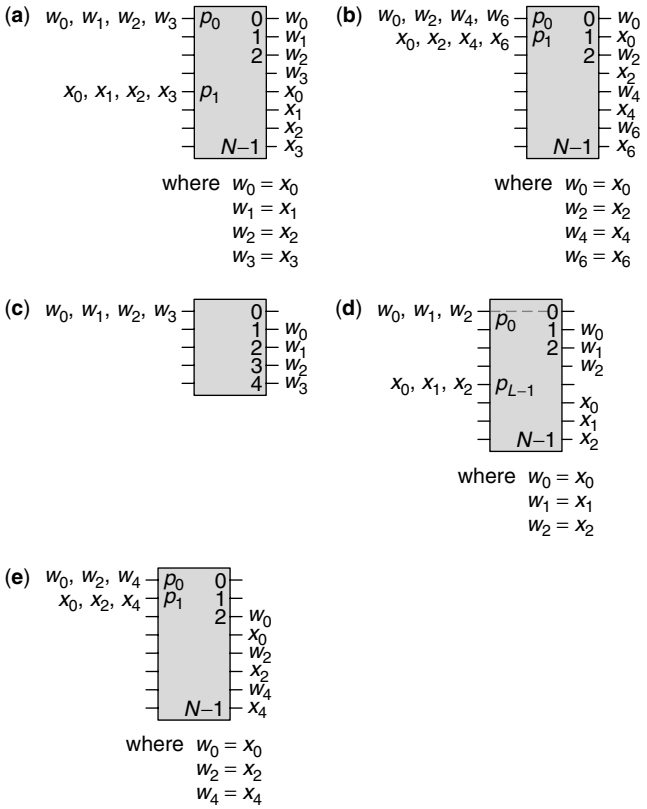


Figure 4. (**a**) Compatible ports with $k = 1$; (**b**) compatible ports with $k = 2$; (**c**) self-blocking port; (**d**) compatible and self-blocking ports with $k = 1$; (**e**) compatible and self-blocking ports with $k = 2$.

1. $\mathcal{P}$ is invariable with respect to some wavelength $w \notin \mathcal{W}$
2. $\mathcal{P}$ is compatible with respect to $\mathcal{W} \cup \{w\}$.

The invariable principle states that both the input and output ports are identical with respect to a specific wavelength w that is not in $\mathcal{W}$. The final design of the OXC includes the wavelength channels that are delivered to the set of compatible ports excluding the invariable ports with respect to w. These invariable ports are then the self-blocking ports with respect to $\mathcal{W}$.

As shown in Fig. 4d, a derived set of compatible and self-blocking ports is shown if the channel spacing is 1. This can be easily modified from the set of compatible ports shown in Fig. 4a. Mathematically, if N is a multiple of $(W+1)$, then the set of input ports can be described as $\mathcal{P} = \{i : 0 \le i < N, i \bmod (W+1) = q\}$ for some $0 \le q \le W$. If

$$[w - (2q+1) + c] \bmod (W+1) = 0$$

then $\mathcal{P}$ is compatible and self-blocking with respect to $\mathcal{W}$. On the other hand, when the channel spacing is L, then $\mathcal{W} = \{(w + Lx) \bmod N : 0 \le x < W\}$ for some integer w and $0 \le w \le N-1$. A possible configuration can be found in Fig. 4e, which can be easily obtained from Fig. 4d. Mathematically, if N is a multiple of L, then $\mathcal{P} = \{(q+i) \bmod N : 0 \le i \le L-1\}$, for some $0 \le q \le N$, is

compatible and self-blocking with respect to $\mathcal{W}$ if

$$[w - (2q + 2L - 1) + c] \bmod N = 0$$

or

$$[w + L(W - 1) - (2q - 1) + c] \bmod N = 0.$$

Detailed proofs can be found in Wan's treatise [10].

3.1.4. Implementations of Passive OXCs with AWGs. In this subsection, designs of different OXCs with AWGs will be discussed. It is assumed that there are L input fibers and each of them carries W wavelength channels. A design shown in Fig. 5a is equivalent to the one in Fig. 1b with $L = W$, a number of multiplexers, demultiplexers, and space switches. From Fig. 5b, only two AWGs are required to build a 2-line 4-wavelength OXC if the channel spacing is 1. The drawbacks are that some complicated crossover links must be set up between AWGs and the 2×2 space-switching modules. The designs of these 2×2 space-switching modules can be found in Section 3.2.

On careful investigation, there is a simple way to improve that design by selecting a wavelength set $\mathcal{W}$ with channel spacing L, where N is a multiple of L and L is even. As shown in Fig. 4b, a set of compatible ports, $\mathcal{P} = \{(q + i) \bmod N : 0 \le i < L\}$ for some $0 \le q < N$, is selected. Another desired requirement is to make sure that no waveguides are required to connect the top and bottom portions of an AWG to a space switch. This is achievable if any wavelength w from port $[(q + L - 1) \bmod N]$ is routed to a port that is a multiple of L, specifically, $\{[w - (q + L - 1) \bmod N + c] \bmod N\} \bmod L = 0$. An OXC can be built as long as q is selected such that $(w + c + 1 - q) \bmod L = 0$ is satisfied. If $w = c = 0$ and $q = 3$, then a cascaded configuration of a 2-line 4-wavelength OXC with a channel spacing of 2 is as shown in Fig. 5c.

So far, two AWGs are needed to build the designs shown in Fig. 5b,c. There are unused ports on the input sides in these designs. If a set of compatible and self-blocking ports with respect to $\mathcal{W}$ can be formulated as shown in Fig. 4d, then only one AWG with loopback links can be constructed as an OXC. Given $N \ge L\ (W + 1)$, the AWG operates as a multiplexer and demultiplexer in an OXC with the symmetric property. Furthermore, it may even reduce the insertion loss[7] in a single-AWG design. The AWG initially demultiplexes the L input fibers into $L \cdot W$ wavelength components. For those L identical wavelength channels, there is an $L \times L$ space switch. By taking advantage of the symmetric property, the signals can be rerouted to the input channels of the AWG. These

[7] The insertion loss of a 1-fiber 40-channel AWG from NTT Electronics is at worst 6 dB.

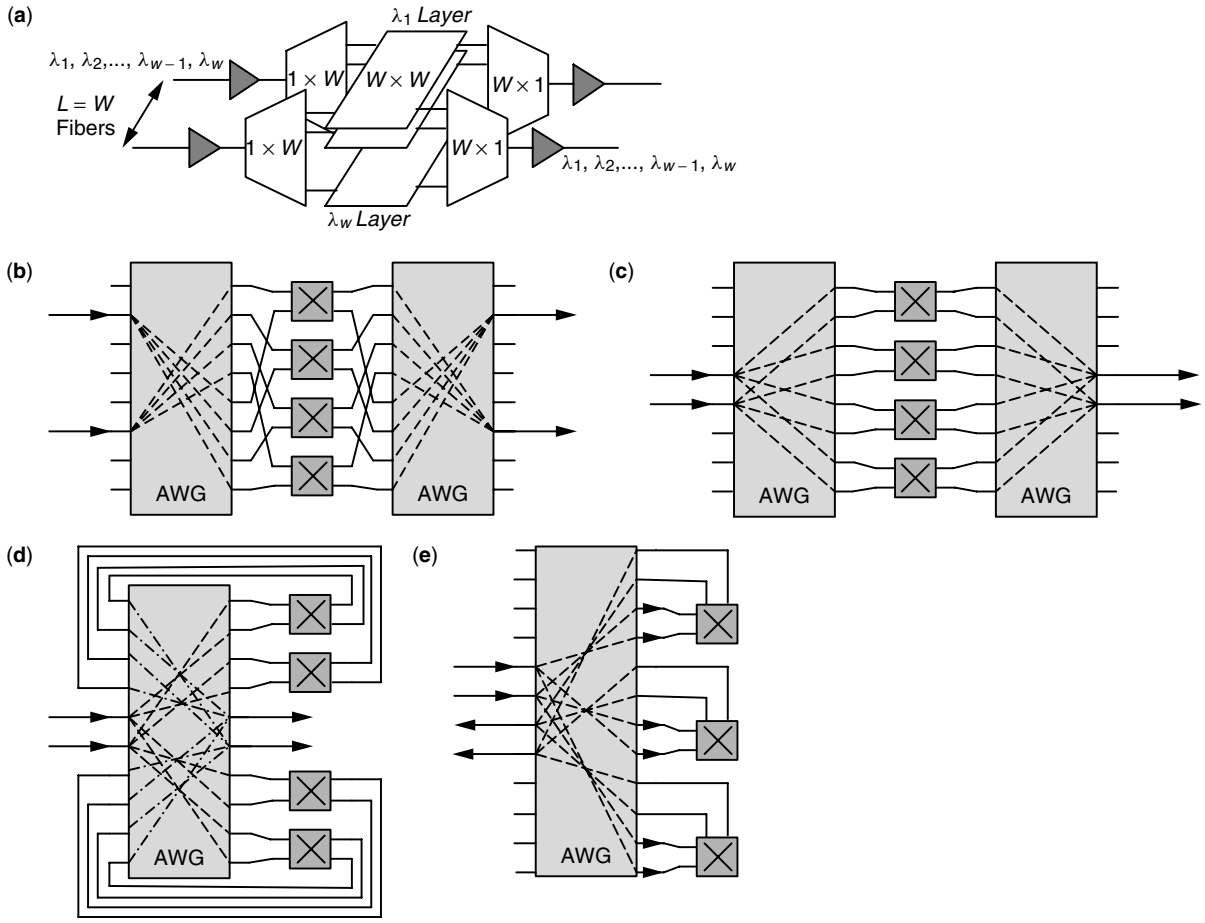


Figure 5. Optical cross-connects with AWGs.

multiple wavelength channels can then be remultiplexed at the L output ports. An example can be found in Fig. 5d with only one AWG. In this design, N is equal to $L \cdot (W+1)$ to minimize the cost, and the channel spacing is 1. To characterize the architecture mathematically, there is a set of compatible and self-blocking ports, $\mathcal{P} = \{0 \le i < N : i \bmod (W+1) = q\}$ with respect to the wavelength channels, $\mathcal{W} = \{(w+x) \bmod N : 0 \le x < W\}$. Provided $[w - (2q+1) + c] \bmod (W+1) = 0$ for some $0 \le w < N$ and $0 \le q \le W$, we obtain a passive loopback 2-line 4-wavelength OXC as shown in Fig. 5d with $w = 0, c = 1, q = 4$.

The design of a loopback architecture is effective on port utilization of an AWG. However, it may not be desirable to construct those recirculating fibers from switches to the input side of the AWG. Therefore, there is a foldback design as shown in Fig. 5e. In this design with $N \ge 2LW$, a set of double-sized compatible ports with respect to $\mathcal{W}$ is required. Suppose that $\mathcal{P}_1$ and $\mathcal{P}_2$ are the sets of L input and L output ports, respectively. It is desirable for $\mathcal{P}$ to be the set of compatible ports if $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. Similar to the last design of OXC, the AWG initially demultiplexes incoming signals into $L \cdot W$ wavelength components. Subsequently, an $L \times L$ space switch does the proper signal switching for each wavelength channel. In this design, a switch accepts the wavelength signals from a set of L input ports, $\{(w - i + c) \bmod N : i \in \mathcal{P}_1\}$, and delivers the signals to another set of L output ports, $\{(w - i + c) \bmod N : i \in \mathcal{P}_2\}$. The design makes use of both the reciprocal and symmetry properties by folding back wavelength signals onto the L output lines. The final foldback design without any crossovers is found in Fig. 5e with $w = 0$, $c = 1$, and $q = 4$ for a 2-line 3-wavelength OXC. However, there are certain input ports that are not used to construct an OXC. This is the only drawback in this design.

3.2. Active Optical Cross-Connects

Space-switching architectural designs have been created since the invention of the telephone. Currently, space-switching architectures are quite mature. Unfortunately, these space-switching designs cannot be applicable to the optical domain because of unavailable optical logic and storage devices. These architectures are still useful mostly in constructing large-scale electronic switches. They probably can provide only a limited number of stages on switch expansions because of the loss issues in the optical domain. In the following, we outline the designs of several optical cross-connects that are based on three of the latest popular optical device technologies. The designs of these OXCs are still closely related to the device technology. They are built with the liquid crystal, thermobubble/thermocapillary and MEMS technologies.

3.2.1. Liquid Crystal Switches. Despite the name given, liquid crystals [18–23] are not truly liquid. They exist in a state between liquid and solid called *mesophasic*. A liquid crystal molecule has an elongated shape, and is often represented as a rod. Under the proper conditions, the orientation of these molecules can be changed so that they face in a certain direction. The orientation can affect the optical properties of the liquid crystal, which in turn affects the polarization of the light passing through it.

There are several types of liquid crystals, including nematic, discotic, cholesteric, and various kinds of smectic phases, which can be characterized by different arrangements of the molecules. Utilizing a magnetic or electric field can often change the optical properties of a liquid crystal. Liquid crystal switches have low insertion loss and excellent performance at the same time. A major advantage with the liquid crystal is its ability to add and drop different colors of light without having to demultiplex all wavelengths of the incoming signal.

In constructing basic liquid crystal switching modules, an early result was reported [20]. The 1×2 splitter is shown in Fig. 6a. The polarization beamsplitter (PBS) is used for lowering crosstalk. It divides the input light into two linearly polarized lightbeams. The twisted nematic liquid crystal (TN-LC) provides polarization switching. It provides 90° polarization rotation without an applied voltage, but it keeps the original polarization with an applied voltage. The function of the birefringent crystal block (BRB) is for extraordinary wave walkoff. The switching operations from one input to one of the two output ports are shown in Fig. 6b,c.

In the following, an architectural design of multichannel liquid crystal switches is described. There are three different basic liquid crystal components that can be used to construct a multistage interconnection of optical switches. These components are a 2×2 polarization switch, an optical beam router, and a beam shifter.

The 2×2 polarization switch is shown in Fig. 7a,b. It is constructed with the transmission-type twisted nematic liquid crystal spatial light modulator (LC-SLM) arrays. A thin nematic liquid crystal layer at the center is surrounded by two glass plates that are bonded to transparent electrodes, such as indium tin oxide. When it is in OFF state, it rotates the light with a 90° polarization angle. When it is ON, then no polarization state will be changed. Therefore, the 2×2 polarization switch operates like a switch between two orthogonal polarizations. For the second component, the optical beam router, its operation model is shown is Fig. 7c. Its goal is to exchange one of the polarization components with a lightbeam that propagates along an adjacent path. The PBS operates differently on the two polarization orientations, the polarizations that are parallel and perpendicular to the surface, the P component and the S component. As shown in Fig. 7d, a basic beam router is composed of five PBSs. Each of these PBSs reflects only the S component of a projecting lightbeam. With the arrangement of these PBSs, the two adjacent S components will be exchanged on leaving the beam router. However, this architecture requires high accuracy in assembly in order to stabilize the coupling loss. The third component is a beamshifter. Its functional model and hardware construction are shown in Fig. 7e,f, respectively. A basic beamshifter consists of three PBSs. Similar to the optical beam router, each PBS reflects only the S component of the lightbeam. It uniformly displaces one of the polarization components, and takes in another S component from its adjacent beam path. After designing all these basic components, we can move forward to construct

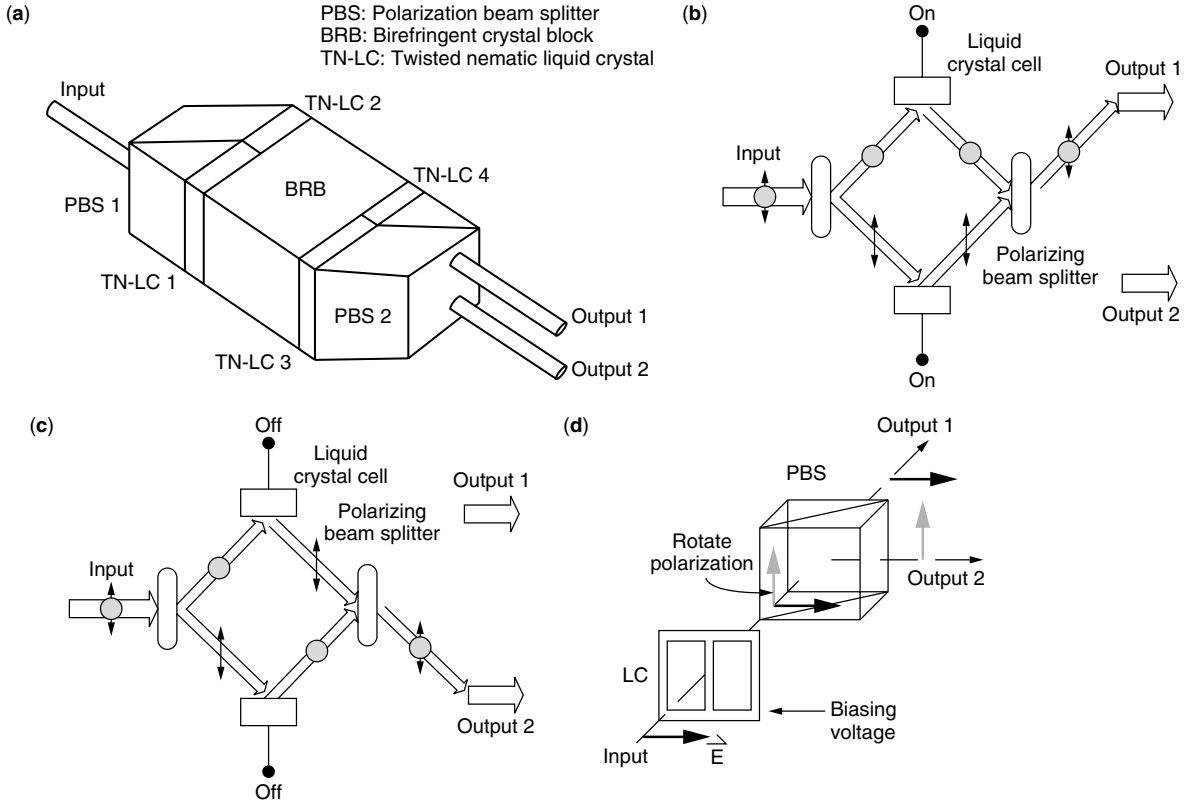


Figure 6. Liquid crystal switch and its operations.

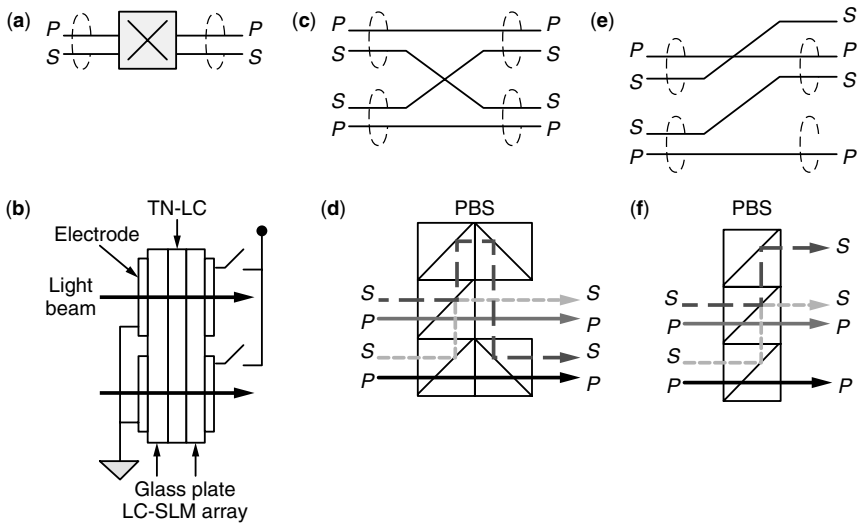


Figure 7. (**a**,**b**) 2×2 Polarization switch; (**c**,**d**) optical beam router; (**e**,**f**) beamshifter.

optical switch systems. Two optical switch architectures are as shown in Fig. 8. They demonstrate the feasibility of constructing multistage optical interconnection networks with the liquid crystal technology. Currently, the size of an optical liquid crystal switch will be limited because of the physical construction, alignment, and the signal loss issues.

In a relatively earlier design, the switching mechanism with the liquid crystal switch was based on the total internal reflection of the liquid crystal. A 2×2 switch consists of two glass prisms of equal refractive index and base angle [18]. The inside face of both prisms have been coated with a transparent electrode and thin polyamide layer. The prisms are bonded together using an epoxy edge seal loaded with spacers of the desired diameter in the range of few micrometers. Liquid crystal with OFF-state alignment is introduced between the prisms by vacuum filling, and the cell is sealed. The OFF-state alignment of the liquid crystal reflects light from the input port 1 back to the output port 1. The ON state is switched by applying a voltage to the electrode. It changes the alignment of the liquid crystal normal to the face of the prisms so that

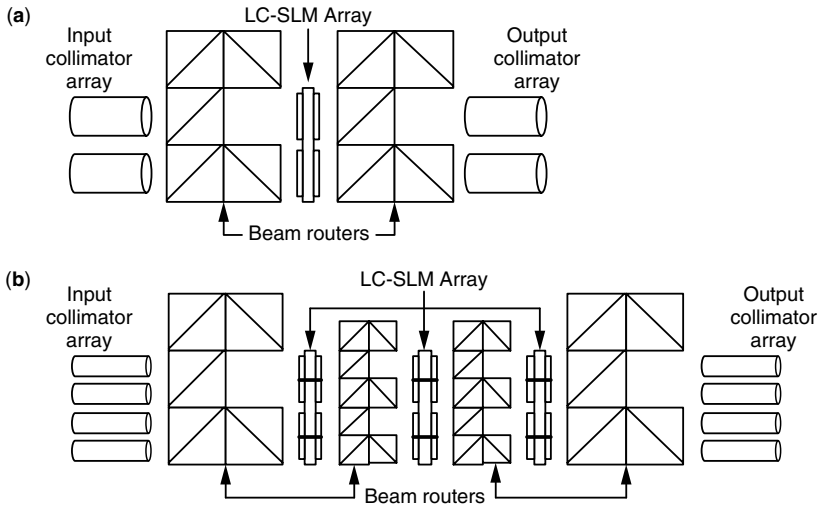


Figure 8. (**a**) 4×4 and (**b**) 8×8 optical switches.

light will be able to pass from input port 1 to output port 2 through the liquid crystal. However, this architecture is not easily extensible to construct large-sized optical switches. A design of 8×8 channels crossbar switch has been proposed [19]. The switch uses an array of eight 1×8 liquid crystal switches for signal divergence and eight 8×1 liquid crystal switches for signal convergence.

The switch performance of the 2×2 liquid crystal switch [18] has an average transmit state crosstalk of 33 dB, and the switching speed is <5 ms. The insertion loss obtained for the liquid crystal itself is 0.5 dB, excluding the Fresnel reflection at the glass–air interface and the 3 dB polarization loss [18]. For the 4×4 switch, it experiences an average crosstalk level of -22 dB and the insertion loss is 0.8 dB. The switching time should be longer because it is required to set up more electrodes along the switching path. The 8×8 liquid crystal crossbar switch described by Noguchi [19] has an average crosstalk level of -59.5 dB and an insertion loss of 3.44 dB. The improvement of the crosstalk level is due to the isolation of all the optical signal paths within the switch by a set of 1×8 and 8×1 liquid crystal array. The increase in insertion loss is due to the additional stage required within the switch body.

3.2.2. Bubble and Thermocapillary Switches. A bubble switch developed by Agilent is based on low-cost thermal inkjet bubble technology. The earliest report on bubble switching was provided by Jackel et al. [24]. The design concept of the basic bubble-switching component is simple. There are two core paths for light transmission. These two paths are crossing each other, and at the crosspoint, there is a trench holding refractive index-matching fluid. There is also a heater that makes the liquid boil, and forms a bubble. In the switch, this bubble is critical to reflecting the light onto a new path. When it forms, the bubble displaces the fluid from the trench, which makes the space more like air. It creates an interface between the glass and the bubble that shifts the light with the total internal reflection principle [25]. Even though there are no moving parts in the systems, the switching time for the device with the software control is around 10 ms as reported by Fouquet. The commercialized 32×32 two-dimensional crossbar switch from Agilent has a specification of -50 dB crosstalk.

Bubble switches are usually made on glass waveguides. The glass is etched to produce capillary channels that contain index-matching fluid and air bubbles. Usually, rough capillary walls are formed by reactive-ion etching or acid etching of glass, and they produce high scattering loss when the air bubble is located at the waveguide intersection. Typical losses of 2.2 dB have been reported [25] for glass waveguide devices. Switching in a bubble switch is shown in Fig. 9.

Bubble switch operations rely on the heating processes that create air bubbles in the index fluid. In order to keep the bubbles at their positions, continuous power has to be applied. On the other hand, there is a thermocapillary process being developed [26–28]. As reported by Sato et al., the structure of the thermocapillary switch is as shown in Fig. 10. A deep trench is formed at each cross

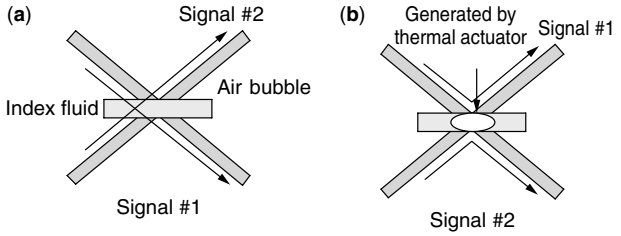


Figure 9. Optical bubble switch operations: (**a**) bar state; (**b**) cross state.

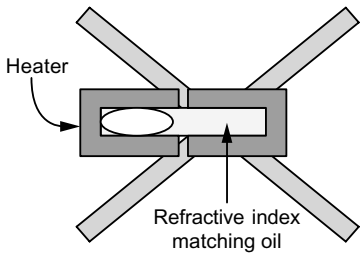


Figure 10. Thermocapillary switching module.

point of the waveguide cores. Refractive-index-matching oil is injected into the trench, which is sealed by a glass lid. The trench is half-filled with the oil; the other half, with a bubble. On top of the trench, there is a pair of microheaters to produce a thermal gradient along the trench. When the oil is heated, the surface tension of the bubble on the heater side is decreased, and then the bubble moves to the side of the actuating heater with a capillary force. This actuation mechanism with the thermal gradient is called thermocapillary. When the bubble is located at the cross point of the waveguides, the light is switched into the crossing waveguide because of the total internal reflection on the glass–air interface. This operation is identical to the operating principle in the bubble switch. However, the mechanism of the bubble action is completely different. For example, the bubbles in Agilent's switch are created through heating, while in the thermocapillary switch, the bubbles always exist and their motions are activated through the capillary force. On the performance side, this thermocapillary switch was reported to have an insertion loss of ~4 dB [27]. However, the switching time takes 50 ms to move the bubble for this specific design. The advantage of this design is that the bubble will latch on the wall and no extra power is required to keep it stationary at that position.

Further improvement can be made on both thermobubble and thermocapillary switches. In order to improve the performance, there are polymer-based thermocapillary switches. In this design, the fluid and air capillary is formed by precision laser ablation yielding a much smoother capillary wall. Optical surfaces achieved via laser ablation result in a polymer waveguide. This design has excellent insertion loss. The loss in the air bubble is less than 0.2 dB. When index fluid fills the capillary at the waveguide intersection, a loss of 0.1 dB is typical for polymer waveguides and glass waveguides.

As of today, there are several firms working on producing optical switches based on bubble technology. To date, only two-dimensional crossbar switches have been constructed with bubble technology.

3.2.3. Microelectromechanical Systems. We have already discussed the constructions of optical switches using the liquid crystal and bubble-switching technologies. Both of them show promising research results and they have the potential to be deployed for commercial use in the near future. However, as of today, the only OXC designs available on the market are made with the microelectromechanical systems (MEMS) technology [29–32], such as Lucent's LambdaRouter [33]. MEMS optical switches are different from the conventional mechanical switches, which are based on macroscopic bulk optics and utilize the advantages of free-space optics. These conventional mechanical switches suffer from large size and mass with slow switching time.

With the introduction of MEMS technology, MEMS optical switches not only retain their conventional advantages of free-space optics such as low losses and low crosstalk but also include additional advantages such as small size, small mass, and submillisecond switching times. Furthermore, MEMS fabrication techniques allow integration of microoptics, microactuators, complex micromechanical structures, and possibly microelectronics on the same substrate to realize integrated microsystems.

For the MEMS devices, their operating units are the micromirrors. The size of these mirrors may be as small as several hundreds of micrometers, and they are made with standard IC fabrication technology [32–38]. Since existing fabrication technology is quite mature, the cost is low for constructing MEMS devices. At the moment, there is a common fabrication process that is well accepted and is known as MUMPs (multiuser MEMS processes) from Cronos[8] for MEMS. The process is composed of both bulk and surface micromachining. Bulk micromachining, such as deep-silicon reactive-ion etching (DRIE), helps set up the overall outlook of a silicon system structure. The surface micromachining and LIGA[9] processes create the details of the final operating structure of a device. These MEMS optical switches consist of many moving mirrors. Therefore, there should be microactuators to drive or oscillate these moving parts in the MEMS device. For example, there may be moving micromirrors to switch laser beams from one fiber to another. There are currently a variety of methods to achieve these microactuation functions, for example, electrostatic, electromagnetic, piezoelectric, magnetostrictive, and thermal expansion. Currently, the electrostatic mechanism is the most common and best-developed method.

For the electrostatic actuators, several types of designs are suitable for constructing OXCs: parallel-plate capacitors, comb drives, and torsional bars. We roughly review the designs of the parallel-plate and comb drive mechanisms. The structure of a parallel-plate MEMS device is shown in Fig. 11a. In general, for all parallel-plate structures, the device stores some capacitance energy, that is, $W = CV^2/2$. When the plates move toward each other, the work done by the attractive force between them can be computed as a change in W with a displacement, x. Therefore, the force can be computed as $F = V^2(\partial C/\partial x)/2$. In the parallel-plate capacitor architecture, only attractive forces can be generated. The design will be more attractive if a larger force can be produced to carry out the heavier workload. It is desirable to offer a larger change in capacitance with respect to distance. This leads to the development of the electrostatic comb drives as shown in Fig. 11b. The comb drives consist of many interleaving fingers. When a voltage is applied, an attractive force is developed between fingers and they move toward each other. With this structural design, there is an increase in capacitance that is proportional to the number of fingers. Therefore, the larger number of fingers can generate larger forces. A potential problem is to carefully control the lateral gaps between fingers. A finger may swing and stick if the gaps are not identical on both sides.

As many reports [29–36] indicate, MEMS optical switches are able to demonstrate their superiority in

[8] Cronos is a division in JDS Uniphase.

[9] LIGA is a German acronym for lithography, electroplating, and molding.

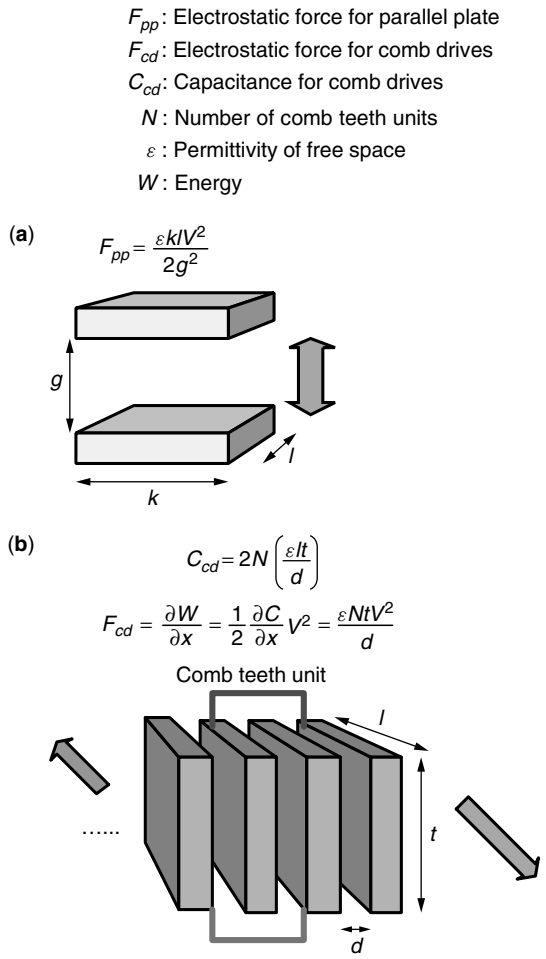


Figure 11. Electrostatic designs: (**a**) parallel-plate capacitor, (**b**) comb drives.

the areas of scalability, insertion loss, polarization-dependent loss (PDL), wavelength dependency, small size, low cost, crosstalk, switching speed, manufacturability, serviceability, and long-term reliability. Among these reports on system performance, the switching time is ~4 ms [33] with ~3 dB insertion loss. In general, we expect that the switching time may fall within the range 1–10 ms, and the insertion loss is between 1 and 6 dB. With steady advancement of the latest fabrication technology, the switching time will be performing even better in the future. For example, the switching time is less than 1 ms when scratch drive actuators are used [35].

In the following, we describe how MEMS technology helps us to construct optical switches. There are currently two broad approaches to implement MEMS optical switches: 2D and 3D MEMS optical switches. Even though both 2D and 3D MEMS optical switches operate on micromirrors in crossbarlike architectures, there are striking differences in terms of how the mirrors are controlled and their ability to redirect lightbeams. However, both of them have shown promise in finding their niche in telecommunication networks. There are already several large 2D MEMS optical switches in the market. Lucent/Agere's WaveStar LambdaRouter is the most sophisticated and the largest 3D MEMS optical switch available in the marketplace today. The delivery of the 1024 × 1024 WaveStar LambdaRouter is expected by the year 2002.

3.2.3.1. 2D MEMS Optical Switches. With the given MEMS technology, many two-dimensional crossbar optical switches were made. In this planar architecture, mirrors are always arranged in a crossbar configuration as shown in Fig. 12a. Each mirror has only two positions and is placed at the intersections of lightpaths between the input and output ports. They can be in either in the ON position to reflect light or in the OFF position to let light pass uninterrupted. The binary nature of the mirror positions greatly simplifies the control scheme. Typically, the control circuitry consists of simple transistor–transistor logic (TTL) gates and appropriate amplifiers to provide adequate voltage levels to actuate the mirrors. For an $N \times N$ switch, a total of N^2 mirrors are required to implement a strictly nonblocking optical crossbar switch fabric. For example, a 16×16-port switch will require 256 mirrors. Moreover, the capability of signal resynchronization within an optical switch is not possible,

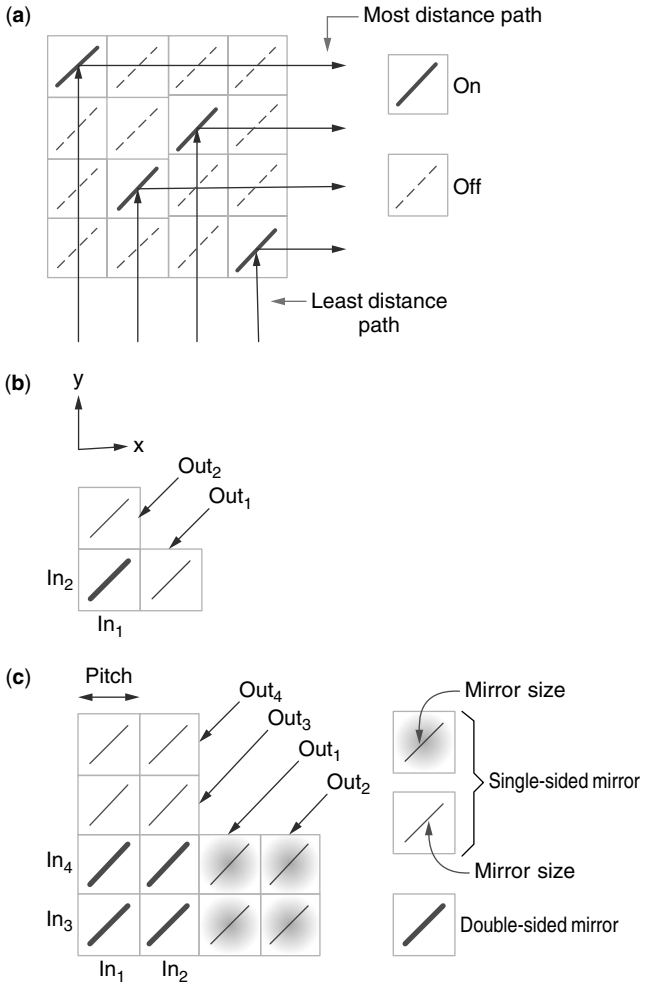


Figure 12. (**a**) Crossbar switch; (**b**) 2×2 and (**c**) 4×4 L-switching matrices.

and the free-space beam propagation distances among port-to-port switching are not constant. As a result, the insertion loss due to Guassian beam propagation is not uniform for all ports. Consequently, there are variations in losses among ports. The minimum and maximum insertion losses of AT&T's 8×8 switch with scratch drive actuators [36] are 3.1 and 3.5 dB, respectively. A simple way to improve the system performance of a 2D crossbar MEMS switch is to decrease the pitch size per mirror unit. It can then reduce the pathlength differences and signal loss as well as increase the port count in the 2D MEMS switch. Certainly, this also leads to smaller mirror size, which can cause signal loss due to the spreading of the Guassian beam. Therefore, more sophisticated and accurate fabrication may be needed to fabricate these systems.

An alternative approach to increasing port count is to interconnect smaller 2D MEMS switching modules to form multistage networks, for example, the three-stage Clos networks. However, this cascaded architecture typically requires up to thousands of complex interconnects between stages, thus decreasing serviceability of the overall switching system. Up to the current stage, extensive research is being performed on the device level. Yeow et al. [39] investigated double-sided mirror design to see if it can provide benefits in existing designs. A planar L-switching matrix design [39] was proposed. Figure 12b,c shows 2×2 and 4×4 L-switching matrices, respectively. The longest-distance path, l_{ldp}, in L-switching matrix is always 25% shorter than that of the regular two-dimensional crossbar switch; whereas the shortest-distance path in the L-switching matrix, l_{sdp}, is always about one-third of the l_{ldp}. As a result, when it is compared to the regular crossbar switch, the maximum path difference in the L-switching matrix grows slowly with the number of input or output ports. It helps slow down the impact of the loss nonuniformity issue [37,38] in 2D MEMS switches due to the pathlength difference problem. The current achievable port count of a 2D MEMS crossbar switch is 32×32, whereas we expect that the L-switching matrix should be able to scale to 64×64 without installing collimators with varying focal lengths for the system. Moreover, these L-switching modules can be used to construct larger-sized Clos networks. Comparison of the construction of three-stage Clos networks with that of the regular 2D crossbar MEMS switches reveals that the one with the L-switching matrix modules has substantially reduced accumulated insertion loss by almost 57% with the pathlength difference issue when only the pathlengths within the switches are counted [39]. However, there are shortcomings in the L-switching matrix; for instance, it may not be possible to establish a new connection without modifying existing connection configurations. Fortunately, the number of paths between an input and an output may have multiple possible paths. If the number of ports is N, the number of possible paths can be $N/2$ for some cases. As an example, you can find two setups for one set of connection requests in Fig. 13.

3.2.3.2. 3D MEMS Optical Switches. All lightbeams in a 2D MEMS switch reside on the same plane.

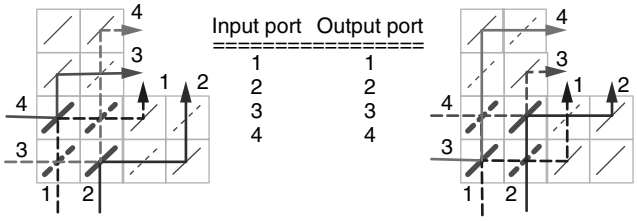


Figure 13. Two path configurations for one set of connection requests in a 4×4 L-switching module.

This arrangement usually results in unacceptable high and uneven loss for large port counts. The 3D MEMS switch [29,33] makes use of the three-dimensional space as an interconnection region that allows scaling far beyond 32 ports with acceptable optical losses. These analog or 3D MEMS switches have mirrors that can rotate freely on two axes as shown in Fig. 14, and light can be redirected precisely in space to multiple angles. The port count would be limited only by insertion loss that results from finite acceptance angle of fibers or lens. Another advantage is that the differences in free-space propagation distances among port-to-port switching are much less dependent on the scaling of the port count. Typically, the optical pathlength scales only as $\sqrt{N}$ instead of N, so port counts of several thousands are achievable with high uniformity in losses ($<$10 dB). Inevitably, much more complex switch design and continuous analog control are needed to improve stability and repeatability of the mirror angles.

To design 3D MEMS optical switches, N or $2N$ mirrors may be required. For example, Nortel Networks' 3D switching architecture [32,40] utilizes two sets of N mirrors. The first plane of N mirrors redirect light from N input fibers to the second plane of N mirrors. All mirrors on the second plane are addressable by each mirror on the first plane making nonblocking connections. In turn, mirrors on the second plane can each be actively and precisely controlled to redirect light into desired output fibers with minimum insertion loss. On the other hand, Lucent's WaveStar MEMS switches, shown in Fig. 14,

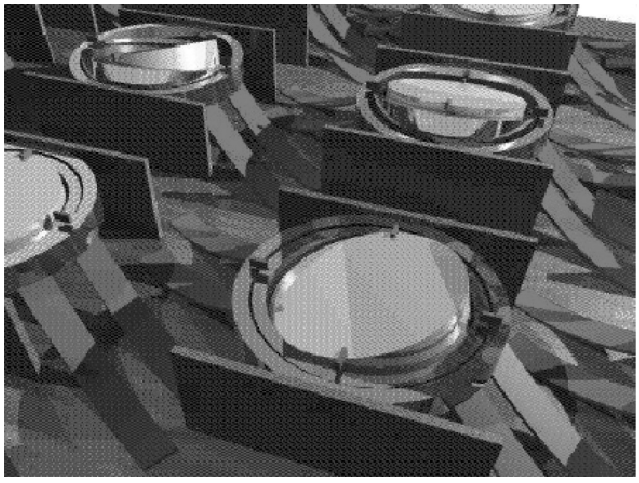


Figure 14. 3D MEMS mirrors.

use only N mirrors and a comparatively large and fixed reflective mirror tilted toward the mirrors. Light from an incoming fiber arrives at one 3D MEMS mirror and is reflected to the large reflective mirror. Then the light will return to another MEMS mirror on the same MEMS plane and be sent to an outgoing fiber. Typically, the mirror can rotate on two axes and is continuously controllable to tilt by at most $\pm 10^{\circ}$ [30]. Moreover, the reported switching module has a maximum insertion loss of 6 dB and a switching time of <10 ms. This 3D optical architecture clearly presents real hope for developing a scalable large-port-count OXC. A WaveStar LambdaRouter with more than one thousand ports is expected to be available in 2002.

4. CONCLUDING REMARKS

Optical switches are the most important components in the future all-optical networks. Numerous companies are producing the next-generation optical switches. Since the late 1990s, the properties of AWGs that allow predefined path setup in the optical domain have been more clearly understood. However, AWGs can provide a platform only for constructing passive OXCs. Therefore, the latest exciting research is on designing active OXCs. At the moment, multiple technologies have been showing promising results. Among them, 3D MEMS is the first one to show exciting and promising results for designing large-scale OXCs. Numerous issues still need to be investigated to further improve the performance of 3D MEMS switches, including the fabrication process, packaging, deposition uniformity on small mirror surfaces, and analog tilting open-loop and closed-loop control [30]. Until now, the research has focused on device-level technology. Novel architecture design on optical switches can also be investigated from the system level with the learnt hardware properties, for example, the L-switching matrix design [39] by integrating different component structures in one system design.

BIOGRAPHY

K. L. Eddie Law received the B.Sc.(Eng.) degree in electrical and electronic engineering from the University of Hong Kong, the M.S. degree in electrical engineering from Polytechnic University, Brooklyn, New York (USA), and the Ph.D. degree in electrical and computer engineering from the University of Toronto in Canada in 1995. From 1995 to 1999, he joined Nortel Networks in Ottawa, Canada, and worked in three different groups: Passport Research, Next Generation ATM Systems, and Computing Technology Lab. Since September 1999, he has been an Assistant Professor in the Communications Group in the Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto. His current research interests are on the active networks, policy-based management on the Internet, TCP/IP protocol design and development, reconfigurable network design, and photonic switch design.

BIBLIOGRAPHY

1. P. P. Mitra and J. B. Stark, Nonlinear limits to the information capacity of optical fibre communications, *Nature* **411**: 1027–1030 (June 28, 2001).
2. N. A. Jackson, S. H. Patel, B. P. Mikkelsen, and S. K. Korotky, Optical cross connects for optical networking, *Bell Labs Tech. J.* 262–281 (Jan.–March 1999).
3. C. Clos, A study of non-blocking switching network, *Bell Syst. Tech. J.* **32**: 406–424 (March 1953).
4. R. Ramaswami and K. N. Sivarajan, *Optical Networks: A Practical Perspective*, Morgan Kaufmann, San Francisco, 1998.
5. T. E. Stern and K. Bala, *Multiwavelength Optical Networks: A Layered Approach*, Addison-Wesley, Reading, MA, 1999.
6. A. Rogers, *Understanding Optical Fiber Communications*, Artech House, Boston, 2001.
7. J. E. Ford, D. J. DiGiovanni, and D. J. Reiley, $1 \times N$ Fiber Bundle, *Proc. Optical Fiber Commun. Conf.'98*, Feb. 1998, pp. 143–144.
8. M. K. Smit and C. van Dam, PHASAR-based WDM-devices: Principles, design and applications, *IEEE J. Select. Top. Quant. Electron.* **2**(2): 236–250 (1996).
9. Y. P. Li and C. H. Henry, Silica-based optical integrated circuits, *IEE Proc. Optoelectron.* **143**(5): 263–280 (Oct. 1996).
10. P.-J. Wan, *Multichannel Optical Networks*, Kluwer, 2002.
11. S. Morasca, D. Scarano, and S. Schmid, Application of $LiNbO_3$ acousto optic tunable switches and filters in WDM transmission at high bit rates, in G. Prati, ed., *Photonic Networks*, Springer-Verlag London Ltd., 1997, pp. 458–472.
12. A. Himeno, T. Kominato, M. Kawachi, and K. Okamoto, System applications of large-scale optical switch matrices using silica-based planar lightwave circuits, in G. Prati, ed., *Photonic Networks*, Springer-Verlag London Ltd., 1997, pp. 172–182.
13. T. Chikama, H. Onaka, and S. Kuroyanagi, Photonic networking using optical add drop multiplexers and optical cross-connects, *Fujitsu Sci. Tech. J.* **35**: 46–55 (July 1999).
14. N. Keil, H. Yao, C. Zawadzki, and B. Strebel, 4×4 polymer thermo-optic directional coupler switch at 1.55 μm, *Electron. Lett.* **30**(8): (April 1994).
15. N. Keil, H. Yao, C. Zawadzki, and B. Strebel, Rearrangeable nonblocking polymer waveguide thermo-optic 4×4 switching matrix with low power consumption at 1.55 μm, *Electron. Lett.* **31**(5): (March 1995).
16. C. Fernando et al., Thermo-optical switching in Si/Si1-xGex distributed Bragg reflectors, *Electron. Lett.* **30**(11): (May 1994).
17. T. Goh et al., Low loss and high extinction ratio 16×16 thermo-optic matrix switch using silica-based planar lightwave circuits, *Proc. Asia Pacific Conf. Communication'97*, Dec. 1997.
18. J.-C. Chiao, Liquid-crystal optical switches, *Proc. 2001 Optical Society of America Topical Meetings: Photonics in Switching*, June 11–15, 2001.
19. K. Noguchi, Transparent optical crossbar switch using liquid crystal optical light modulator arrays, *Integrated Opt. Opt. Fibre Commun.* (Sept. 1997).

20. Y. Fujii, Low-crosstalk 1×2 optical switch composed of twisted nematic liquid crystal cells, *IEEE Photon. Technol. Lett.* **5**(2): 206–208 (Feb. 1993).

21. A. Sneh and K. M. Johnson, High-speed continuously tunable liquid crystal filter for WDM networks, *J. Lightwave Technol.* (1996).

22. N. A. Riza and S. Yuan, Low optical interchannel crosstalk, fast switching speed, polarisation independent 2×2 fibre optic switch using ferroelectric liquid crystals, *Electron. Lett.* (June 25, 1998).

23. C. Mao et al., Liquid-crystal optical switches and signal processors, *Proc. 2001 Asia-Pacific Optical and Wireless Communications Conf.*, Nov. 2001.

24. J. L. Jackel and W. J. Tomlinson, Bistable optical switching using electrochemically generated bubbles, *Opt. Lett.* **15**(24): 1470 (1990).

25. J. E. Fouquet, Compact optical cross-connect switch based on total internal reflection in a fluid-containing planar lightwave circuit, *Proc. Optical Fiber Communications Conf. 2000*, 2000.

26. M. Makihara, M. Sato, F. Shimokawa, and Y. Nishida, Micromechanical optical switches based on thermocapillary integrated in waveguide substrate, *J. Lightwave Technol.* **17**: 14–18 (1999).

27. M. Sato et al., Thermo-capillary optical switch, *Hitachi Cable Rev.* (20): (Aug. 2001).

28. J. T. Gallo, B. L. Booth, C. A. Schuetz, and R. J. Furmanak, *Polymer Waveguide Components for Switched WDM Cross-Connects*, Optical CrossLinks, Inc. (online), *http://www.opticalcrosslinks.com*.

29. D. J. Bishop, C. R. Giles, and G. P. Austin, The Lucent LambdaRouter: MEMS technology of the future here today, *IEEE Commun. Mag.* **40**(3): 75–79 (March 2002).

30. P. B. Chu, S.-S. Lee, and S. Park, MEMS: The path to large optical crossconnects, *IEEE Commun. Mag.* **40**(3): 80–87 (March 2002).

31. P. De Dobbelaere et al., Digital MEMS for optical switching, *IEEE Commun. Mag.* **40**(3): 88–95 (March 2002).

32. T.-W. Yeow, K. L. E. Law, and A. Goldenberg, MEMS optical switches, *IEEE Commun. Mag.* **39**(11): 158–163 (Nov. 2001).

33. D. T. Neilson et al., Fully provisioned 112×112 micromechanical optical crossconnect with 35.8Tb/s demonstrated capacity, *Proc. Tech. Digest Optical Fiber Communications Conf.* (*OFC2000*), March 7–10, 2000 pp. 202–204.

34. R. Giles et al., Silicon micromachines in optical communications networks: Tiny machines for large system, in R. Rai-Choudhury, *MEMS and MOEMS Technology and Applications*, SPIE Press 2000, Chap. 6, pp. 301–329.

35. L. Y. Lin, E. Goldstein, and L. M. Lunardi, Integrated signal monitoring and connection verification in MEMS optical crossconnects, *IEEE Photon. Technol. Lett.* **12**(7): (July 2000).

36. L. Y. Lin, E. L. Goldstein, J. M. Simmons, and R. W. Tkach, High-density micromachined polygon optical crossconnects exploiting network connection symmetry, *IEEE Photonics Technol. Lett.* **10**: 1425–1427 (1998).

37. K. S. J. Pister, M. Judy, S. Burgett, and S. Fearing, Microfabricated hinges, *Sensors and Actuators* **33**(3): 249–256 (1992).

38. T. Akiyama and H. Fujita, A quantitative analysis of scratch drive actuator using buckling motion, *Proc. IEEE Workshop MEMS*, The Netherlands, Jan. 29–Feb. 2, 1995.

39. T.-W. Yeow, K. L. E. Law, and A. Goldenberg, Micromachined *L*-switching matrix, *Proc. IEEE ICC* (in press).

40. A. Neukermans and R. Ramaswami, MEMS technology for optical networking applications, *IEEE Commun. Mag.* **39**(1): 62–69 (Jan. 2001).

OPTICAL SWITCHING TECHNIQUES IN WDM OPTICAL NETWORKS

MYUNGSIK YOO
Soongsil University
Seoul, Korea

CHUNMING QIAO
SUNY at Buffalo
Buffalo, New York

1. INTRODUCTION

The major advances in optical technology have led to the development of optical communication systems and networks, beginning with long-haul communication systems to metropolitan-area networks, even to access networks, which are the final leg in communication systems, in the form of fiber to the curb (FTTC), fiber to the building (FTTB), and fiber to the home (FTTH).

There are a few reasons why *optical networks* are considered as a solution for transmission infrastructure:

1. Optical networks can provide vast bandwidth with low attenuation. Typically, optical systems use wavelengths in three ranges: previously 800–900 nm and 1280–1350 nm, and currently 1510–1600 nm. As a result of low attenuation, optical systems require fewer repeaters or amplifiers. The potential bandwidth of optical signals (wavelengths) is huge since the typical frequency is in the few hundred terahertz (10^{12}). This means that the data rate of optical systems can be much higher than that of communication systems using frequencies in the megahertz (10^6) or gigahertz (10^9) range. Currently, a single wavelength can operate at 10 Gbps (gigabits per second), and even at 40 Gbps. Furthermore, using dense wavelength division-multiplexing (DWDM) technology, it is possible to put many wavelengths into a single fiber. It is now possible to multiplex 80–100 wavelengths, creating Tbps capacity per fiber.
2. Optical networks provide a *transparency* to protocol, data format, and data rate. Thus, once a communication pipe is established between two client points, optical networks are easily able to accommodate any existing network protocols such as IP, ATM, and SONET/SDH. In addition, an optical network can carry data regardless of its format (e.g., analog or digital) and its data rate.
3. Because of WDM technology and transparency, optical networks can provide a cost-effective and futureproof way to building the transport network. As traffic demand grows, it is easy to upgrade

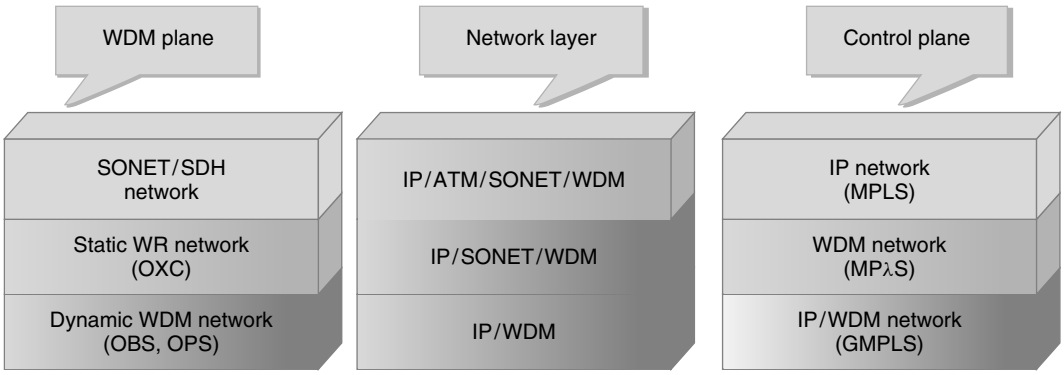


Figure 1. Evolution of optical networks.

network capacity using WDM technology. It is also easy to interface any existing or newly emerging network technology with optical networks due to service transparency.

4. Finally, optical networks become a more feasible solution as the internetworking technologies such as generalized multiprotocol label switching (GMPLS) [1] are actively developed and standardized.

Since it has been deployed in real systems (e.g., carrier systems), optical networking technology keeps evolving to seek better networking solutions in both cost and performance. Figure 1 shows the direction that optical networks have been evolving. We categorize the optical networking technology into three areas: WDM plane, network layer and control plane.

In the *WDM plane*, synchronous optical network (SONET)/synchronous digital hierarchy (SDH) systems were introduced as a first-generation (1G) optical network [2,3]. *SONET/SDH* are designed for carrying voice traffic over the optical fibers with very high transmission capacity. Currently, many carriers have buried many fibers and built SONET/SDH networks for transporting their voice traffic. Although SONET/SDH can be categorized as an optical network, it lacks the optical networking technology such as routing and switching. Thus, it can be said that SONET/SDH only takes advantage of the huge bandwidth of an optical transmission system. In order to enhance the networking capability in the optical network (or layer), *wavelength routing* (*WR*) *networks* [2,3] has been introduced. WR networks placed *optical cross-connect (OXC)* at the switching node, which switches individual wavelengths. In addition, WR networks perform a routing function in the optical domain, owing to the *control plane* such as MPλS (optical-domain MPLS) or GMPLS. On receiving a request from the clients (IP, ATM, SONET/SDH), WR networks set up an end-to-end *lightpath*, which is the process of assigning a wavelength to a particular path by routing and wavelength switching. Simply, WR networks provide a lightpath service to its client layers.

Depending on how frequently a network changes its virtual topology (or *lightpath topology*), an optical network can be classified either a static or dynamic. In a *static WDM optical network*, once established, a lightpath exists for a long period of time (e.g., years or months). This is the type of most optical networks deployed today. The drawback of a static network is that it results in inefficiency when traffic demand changes frequently. A *dynamic WDM optical network* can efficiently support bursty traffic by changing its virtual topology according to traffic demand. Thus, the lightpath in a dynamic WDM network reconfigures itself in much faster time scale. There are two optical switching technologies for dynamic WDM networks under active research and development (R&D), specifically, optical burst switching (OBS) [4–7] and optical packet switching (OPS) [8–10]. Both aim to switch optical packets (or optical bursts) in the optical domain as a conventional packet switching network does in the electronic domain.

In the network layer (layer in the OSI Reference Model), the evolution of optical networks is closely related to how to efficiently support IP traffic. Considering the unprecedented increase in Internet traffic since the mid-1990s, the network architecture should be optimized for the data traffic. Initially, the architecture includes ATM (asynchronous transfer mode) to carry IP packets due to its high-speed switching capability and QoS (quality-of-service) support. However, IP routers are improving their performance in capacity and forwarding speed, exceeding ATM's capability with the help of MPLS (multiprotocol label switching) technology. Thus, the architecture is simplified to IP over SONET over WDM, eliminating the ATM layer. In the next step, the IP layer is directly supported by the WDM layer without the SONET layer. Although the SONET layer provides fast restoration in the event of failure, the *IP over WDM architecture* without the SONET layer has few advantages: (1) SONET is designed for voice traffic, not for data traffic; (2) network control and management can be much simpler; and (3) it is more cost-effective, since SONET equipment increases in cost linearly with bit rate and the number of ports.

At the *control plane*, MPLS [11] was developed for IP networks. By introducing the concept of *label switching*, IP networks can forward packets much faster and overcome the shortcoming of connectionless service with a label-switched path (LSP). In addition, it is much easier to employ traffic engineering. While optical networks become a more attractive solution for transmission networks, MPλS was introduced, which is the application of MPLS

to the optical domain. The wavelengths and lightpaths in MPλS correspond to the labels and LSPs in MPLS, respectively. *GMPLS* [1] was developed for IP over WDM networks with a unified control plane. It is a generalized control plane in a sense that it encompasses packet-switch-capable (PSC), time-division-multiplex-capable (TDM), lambda-switch-capable (LSC), and fiber-switch-capable (FSC) interfaces. With GMPLS, it is possible to control different networks with a single unified control plane.

In the following discussion, we focus on optical switching techniques: wavelength routing, optical packet switching (OPS), and optical burst switching (OBS).

2. OPTICAL SWITCHING TECHNIQUES

Now, we look into the characteristics and issues in optical switching techniques. The general architecture of IP over WDM optical networks (optical Internet) is shown in Fig. 2 [12]. Multiple optical networks exist in the optical domain, where an ENNI (external network-to-network interface) is used for signaling between optical networks. A single optical network consists of multiple suboptical networks, where the INNI (internal network-to-network interface) is used for signaling between them. Again, a suboptical network has multiple optical nodes (e.g., OXCs or optical routers) interconnected with optical fibers. As clients, IP, ATM, and SONET networks are interfaced with optical networks via a UNI (user-to-network interface).

The optical switching techniques that we are interested in determine the service provided by optical networks to client networks. *Wavelength routing*, *optical burst switching*, and *optical packet switching* networks provide lightpath-level, burst-level, and packet-level services, respectively, to client networks.

2.1. Wavelength Routing Network

A *lightpath* is the service unit that the wavelength routing network provides. A *wavelength routing network* consists of multiple OXCs, which are interconnected with optical fiber links. At the network planning stage, for the estimated input traffic, network resources (e.g., number of wavelengths) are dimensioned so that the performance (e.g., blocking probability) can be satisfied.

When setting up a lightpath for a request from a client network, the edge node in the wavelength routing network sends out a lightpath request. In this process, the ingress edge node performs routing to find an optimal route to the egress edge node. The signaling mechanisms in GMPLS may be used in this routing process. Thus, the lightpath setup is done by the control plane. There are two ways of implementing the control plane: by centralized control or by distributed control. Both approaches have advantages and disadvantage. The interested reader may refer to two papers published in 1996 and 1997 [13,14].

When setting up a lightpath, OXCs in the wavelength routing network perform the key function: switching the wavelength to its destined port. The general *architecture of OXC* is shown in Fig. 3. There are M input fibers and M output fibers, each of which carries W wavelengths. The W wavelengths are demultiplexed and put into the optical switch, which has the dimension of $MW \times MW$. The OXC control is in charge of generating control signals to optical components such as optical switch and wavelength converters. The control signal is based on the decision made in the control plane where the lightpath setup requests are processed.

The OXC may have *wavelength converters* for the purpose of increasing performance. It is obvious that having wavelength converters decreases the blocking probability since without wavelength converters, the same wavelength should be used in every link (wavelength continuity), while with wavelength converters, any wavelength can be used. The wavelength conversion can be done either electrically or optically. If the conversion is performed electrically, the optical signal is terminated with O/E (optical to electrical) conversion, then transmitted over the different wavelength after E/O (electrical to optical) conversion. However, its cost linearly increases with the number of transponders per wavelength.

On the other hand, if the conversion is performed optically, the optical signal can be transparently transmitted without O/E/O conversion. However, the cost of optical wavelength converters is still high. Thus, it is desirable

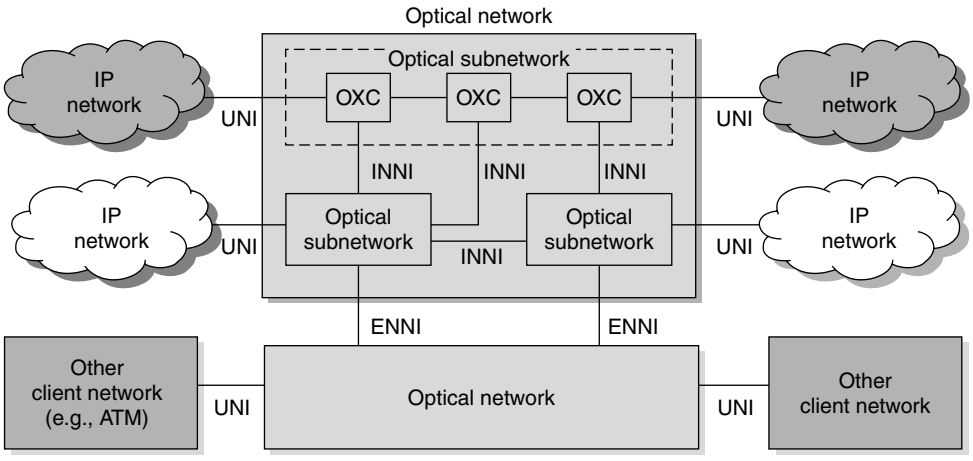


Figure 2. IP over WDM optical networks.

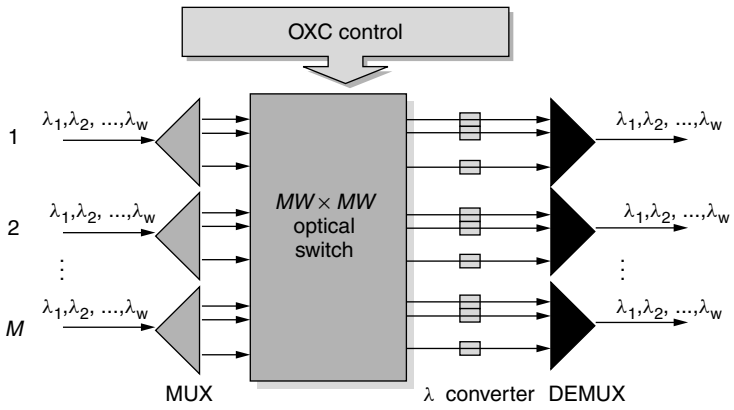


Figure 3. OXC architecture.

to keep the wavelength conversion capability minimum (no conversion or limited conversion). There are studies showing that the gain of full wavelength conversion over no wavelength conversion is a function of the number of wavelengths and the number of hops to traverse [15]. Thus, with proper network planning (e.g., short network diameter and enough wavelengths), it is possible to keep the conversion capability low.

Now, let us look at the characteristics of a wavelength routing network. A wavelength routing network is in the form of circuit switching, and thus we also call it an *optical circuit switching* (OCS) network. It requires *two-way reservation* for lightpath establishment. In other words, a lightpath is established when the acknowledgment comes back as a response to the setup request. This process introduces a setup delay, which is proportional to the round-trip propagation delay over the distance between the two points. Although the setup delay is ignored in a static OCS network with relatively long session time (e.g., years or months), the setup delay may affect the performance of a dynamic OCS network when the session time is of the same order of magnitude as the setup delay. It is another shortcoming of a wavelength routing network that the bursty traffic may result in poor performance due to the static nature of a wavelength routing network. Another important issue in a wavelength routing network is the routing and wavelength assignment (RWA). With an efficient RWA algorithm, it is possible to reduce the network resources for the given input traffic. However, in spite of intensive research on RWA [16], it is still a hard problem to solve, especially when RWA is performed online.

Although there are some disadvantages, a wavelength routing network is the feasible solution in the near term, due to immaturity of optical technology.

2.2. Optical Packet-Switching Network

In a *packet-switching network*, user data are segmented into packets. Each packet consists of two parts: a *header* containing the control information (e.g, routing information) and *data*. As an example, IP datagrams and ATM cells are the packets in IP and ATM networks, respectively. In a packet-switching network, there are two types of service: datagram service (or connectionless service) as in an IP network and virtual circuit service (or connection oriented service) as in an ATM network. While each packet takes the same path in the *virtual circuit service*, each packet may take a different path in a *datagram service*. In either case, each packet goes through intermediate packet switching nodes until it reaches its destination. The key functions performed by intermediate nodes is routing and forwarding, namely, deciding the next hop node and forwarding the packet. These are the characteristics of a packet switching network in the electrical domain.

The *optical packet switching* (*OPS*) *network* is to perform the same packet switching functions in the optical domain. Thus, an optical packet is transmitted and processed by the optical packet-switching nodes. The general architecture of an optical packet switching node is shown in Fig. 4.

An *OPS node* consists of the input and output processing units, a switching unit, a buffering unit, and control unit. The optical packets arrive at input processing unit, where synchronization and header extraction take place. If the system operates in time slots with a fixed size of packets, then synchronization is required for the alignment of multiple incoming packets, which may arrive at different times. The tunable delay, which can be implemented with fiber delay lines (FDLs) and 2×2 optical switches, may be used for synchronization.

When an optical packet arrives, its header is detached and sent to the control unit for processing. Currently, the implementation of the control unit using optical logic is very difficult. Thus, after being extracted, the header part goes through O/E conversion, while the data remain in optical form. To facilitate tapping the optical signal, optical packets are encoded using a technique such as subcarrier multiplexing (SCM). Once the header is decoded in the control unit, the routing process is performed to determine the output port. This process may be layer 3 IP routing or layer 2 label switching. In this process, the control unit generates the control signals to the switching unit (small and fast optical switch) and the buffering unit. The buffering unit resolves the contention problem when multiple packets are destined to the same output port. Since optical memory is not commercially available today, the buffering unit is usually implemented using FDLs, which provide only limited time of buffering.

Finally, the header part and data part are rejoined at the output processing unit before sending it out to the next switching node. If the information in the header needs to

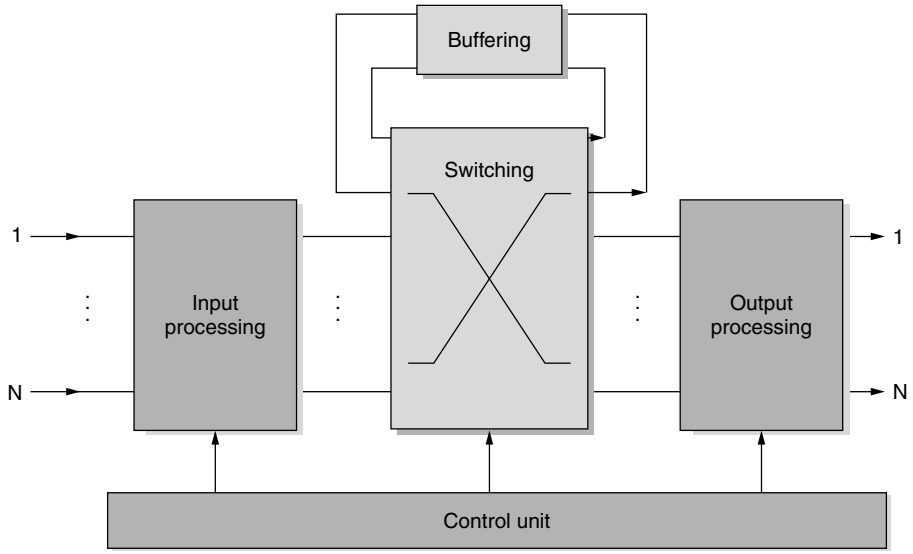


Figure 4. OPS node architecture.

be updated (e.g., by label update after label swapping), the rewriting process takes place.

Although an optical switching network overcomes some of the shortcomings of a wavelength routing network (e.g., inefficiency due to static nature), it is technically hard to implement packet-switching functions with current optical technology. The optical burst switching (OBS) technique attempts to effect balance between wavelength routing and optical packet switching by overcoming the disadvantages of both techniques, while preserving their merits.

3. OPTICAL BURST SWITCHING NETWORK

So far, we have described the characteristics of a wavelength routing network and an optical packet-switching network. Now, we focus on another alternative, *optical burst switching* (OBS), and look at its characteristics in some detail.

3.1. OBS Protocol

The distinction between OBS and OCS is that the former uses a one-way reservation while the latter uses a two-way reservation. It is called the two-way reservation when there must be a connection setup procedure before the data transmission takes place. It is called the *one-way reservation* when the data (which is called the data burst) follow the connection setup request immediately after waiting for some delay. This delay will be called an *offset time*, which will be explained later. Note that the connection setup request in the OBS will be called a *control packet* or a *burst control packet* (BCP).

Although OBS and OPS have similar characteristics (e.g., statistical multiplexing on the links), the distinction between the two is that the former has some unique features such as the offset time and the delayed reservation [4,5]. In addition, the payload in the OBS is much larger than that in the OPS. The payload unit in OBS networks will be referred as the data burst hereafter.

The operations of the *OBS protocol* are illustrated in Fig. 5a. When the source node S has a data burst to send, the burst control packet (BCP) is generated and transmitted first. At each node, the BCP takes δ time unit to be processed and makes the reservation for the following data burst. Meanwhile, the data burst, after waiting

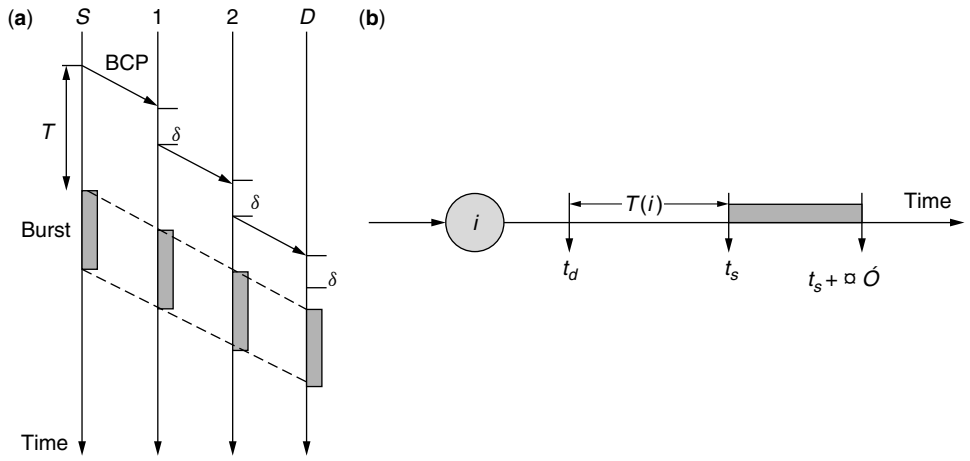


Figure 5. Offset time (**a**) and delayed reservation (**b**) in OBS.

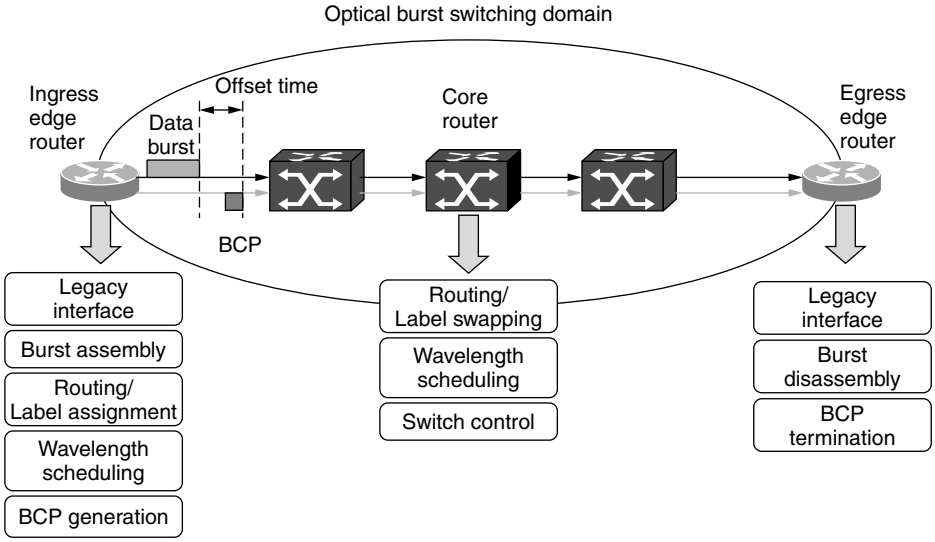


Figure 6. OBS network.

for an offset time, which is denoted as T, at the edge router, immediately follows the BCP. Since the BCP has already set up the necessary configurations, including the resource reservation, the data burst does not need to be delayed (buffered) for processing at the intermediate nodes (nodes 1 and 2), and just cuts-through to the destination node D. Thus, OBS networks can provide the data plane with the end-to-end transparency without going through any optical–electrical–optical (O/E/O) conversions.

The advantage of the offset time is to decouple the BCP from the data burst in time, which makes it possible to eliminate the buffering requirement at the intermediate nodes. In order to ensure that the data burst does not overtake the BCP, the offset time should be made long enough by considering the expected processing time at each intermediate node and the number of hops to be traversed by the BCP [4,5].

Another feature of OBS is the *delayed reservation* (DR), which makes it possible to statistically multiplex data bursts on the links. The DR is illustrated in Fig. 5b, where $T(i)$ is the offset time at node i, 1 is the transmission time of the data burst at node i, and t_a and t_s indicate the arriving time of the BCP and its corresponding data burst, respectively. According to DR, after being processed, the BCP reserves the resources at a future time (at the arrival time of the data burst), which can be obtained from the offset time $T(i)$. Also, the reservation is made for just enough time to finish the transmission (data burst duration 1). In OBS, two parameters — the offset time and the mean data burst size — need to be selected carefully in the design step since they have a great impact on performance.

3.2. OBS Network and OBS Routers

Now, we discuss architectural aspects of OBS [6]. For simplicity, we focus on a single OBS domain, where all nodes (or *OBS routers*) are well aware of the OBS protocols. Note that the terms *OBS domain* and the *OBS network* will be used interchangeably. Depending on the location in the OBS domain, the OBS routers are classified as *edge routers* and *core routers* as shown in Fig. 6. Note that the edge routers should be equipped with both capabilities as an *ingress edge router* (*IER*) and as an *egress edge router* (*EER*). Thus, the edge router functions as an ingress router when there are inbound data to the OBS domain, whereas the edge router functions as an egress router when there are outbound data from the OBS domain. In the following discussion, we describe the functions and general architectures for each type of OBS router.

The general *architecture of the IER* is shown in Fig. 7. The IER should provide the interface between the OBS network and other legacy networks. It also needs to provide the signaling interface between two different networks. When the IER receives incoming data (e.g., IP packets, ATM cells, or voicestreams) from the line interface cards, the burst assembly process takes place in which multiple packets are packed into a data burst. We will discuss the burst assembly process later. The arriving packets are switched to the appropriate assembly queues according to their destination and QoS. The first packet arrival in an assembly queue initiates the BCP generation where some BCP fields such as burst size, offset time, and label are to be determined and filled later when the *burst assembly* is completed.

When the burst is assembled long enough to meet the requirements, the BCP obtains the field information of burst size and offset time. On the basis of the routing decision, a label is assigned to establish the label-switched path (LSP). If the LSP already exists (has been set up by the previous BCP), the previously used label is assigned. Otherwise (i.e., if a new LSP needs to be set up or the existing LSP needs to be changed), a new label is assigned. The downstream nodes perform the label swapping, where the inbound label is mapped into the outbound label at the local node. The label information in the BCP should be updated accordingly. How the labels are distributed and assigned depends on the label distribution protocols such as RSVP-TE and CR-LDP [17,18].

According to OBS protocols, the BCP is transmitted to the core OBS router an offset time earlier than its corresponding data burst. The wavelength assignment

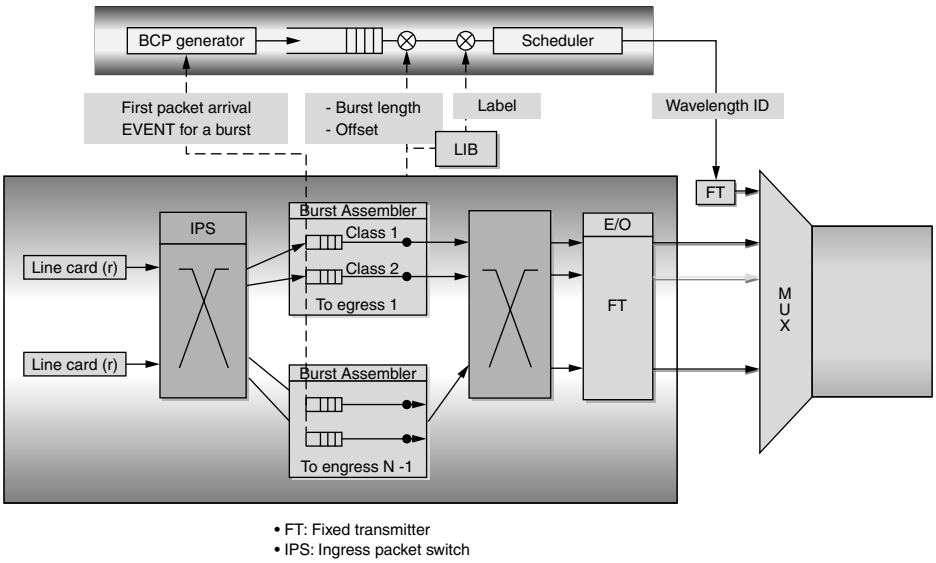


Figure 7. Architecture of ingress-edge router (IER).

and scheduling is required in this step. There are two kinds of wavelengths (or channels) in the OBS domain: the *control channels*, which carry the BCPs, called the *control channel group* (CCG); and the data burst channels, which carry the data bursts, called the *data burst channel group* (DCG). Thus, two wavelengths are assigned and scheduled for transmission of the BCP and its data burst. The wavelength scheduling on the DCG (for data bursts), which can enhance the utilization, is an interesting research area. There are a few scheduling algorithms proposed to date, such as first fit, horizon scheduling, and void-filling scheduling [6,19]. It is noted that while the labels carry the path (LSP) information, the wavelengths are meaningful only at the local node. In this way, the downstream nodes can assign the wavelengths dynamically, which combines with the label only for the duration of the data burst.

The *architecture for the EER* is shown in Fig. 8. The main functions of the EER are the BCP termination and the data burst disassembly. When a BCP arrives, the EER processes it and reserves the buffer space as required by the burst length field for the burst disassembly process. On arrival, the data burst, after being converted to an electrical signal, goes through the burst disassembly process. Then the disassembled packets are distributed to their destination ports.

The core router has the general architecture shown in Fig. 9. It consists of two parts: a switch control unit and a data burst unit. While the switch control unit is responsible for processing the BCPs on the CCG, the data burst unit is responsible for switching the data burst to the destined output port, which is controlled by the switch control unit. Most of the functions in the core router take place in the switch control unit, which includes the label swapping for establishing the LSP, local wavelength scheduling, burst contention resolution, and generation of the control signal to the data burst unit. The data burst unit consists roughly of demultiplexers, inlet fiber delay lines (FDLs) (for adjusting the timing jitter of offset time between the BCP and the data burst), optical switches, wavelength converters, FDL buffers (for

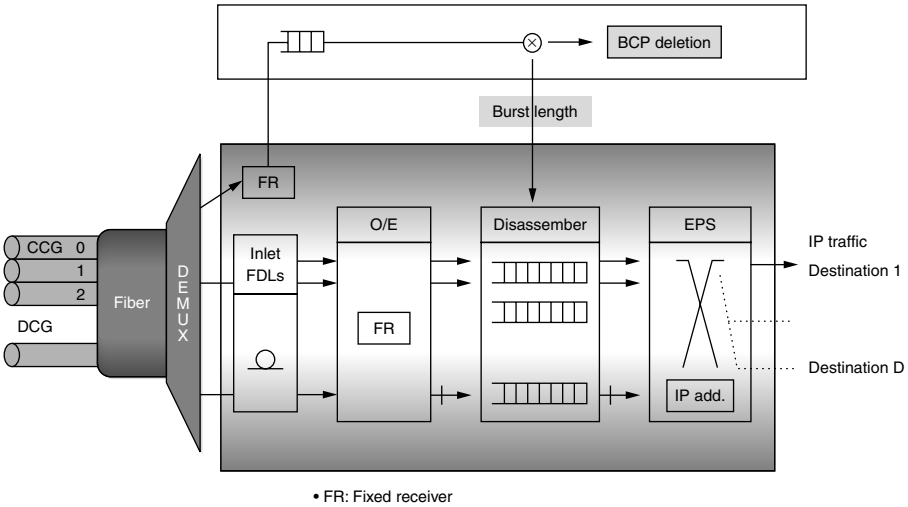


Figure 8. Architecture of egress edge router (EER).

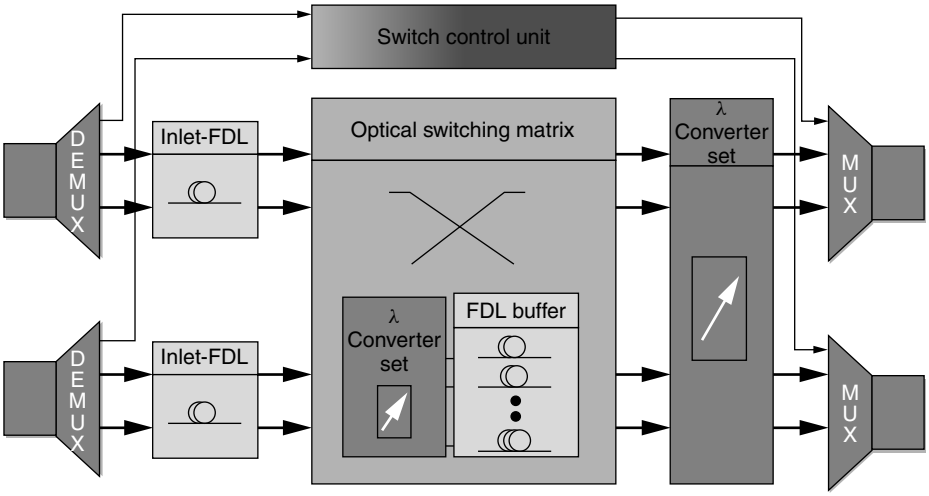


Figure 9. Architecture of core router.

contention resolution), and multiplexers. The components of the data burst unit are passively configured by the control signals from the switch control unit.

Before we conclude this description, of the architecture, it is worth mentioning some important design parameters such as the capacity of the core router, the switching speed of the optical switch, and the average burst size. The capacity of the router is determined by the number of incoming fibers and the number of wavelengths available in each fiber. For example, a core router of 10 Tbps capacity needs 32 incoming fibers, each of which has 32 wavelengths operating at 10 Gbps per wavelength. Of course, since some wavelengths are dedicated to the CCG, the router capacity is determined only by the number of wavelengths in the DCG.

The core router requires two switches; one is in the switch control unit, and the other is in the data burst unit. While the former switches the BCPs on the CCG, which can be implemented with a small electronic switch (depending on the number of wavelengths in the CCG), the latter switches the data bursts on the DCG, which can be implemented with an optical switch. The dimension of the optical switch depends on the number of wavelengths in the DCG. The architecture of the optical switch may be either a simple single stage if a large optical switch is available [e.g., MEMS (microelectromechanical systems) crossbar switch [20]] or multistage interconnection network (MIN) as in an ATM switch [21] if multiple small optical switches are used.

Next, consider the switching speed of an optical switch and the average burst size. The *switching time* of an optical switch is the delay that it takes to change from one state to another state. The switching speed imposes an overhead on the switch throughput. In other words, the slower the switching speed, the worse the throughput. However, the overhead caused by the switching speed can be reduced if the data burst is long compared to the switching time. Thus, for optimum switch throughput, the average burst size should be selected according to the switching speed of the optical switch in use. This average burst size is the requirement that the burst assembler should meet.

3.3. QoS in OBS Network

Given that some real-time applications (such as Internet telephony and videoconferencing) require a higher *QoS* than do non-real-time applications [such as electronic mail (email) and general Web browsing], the QoS issue should be addressed. Although QoS concerns related to optical (or analog) transmission (e.g., dispersion, power and signal-to-noise ratio) also need to be addressed, here, we focus on how to ensure that critical data can be transported in the OBS domain more reliably than noncritical data. Unlike the existing QoS schemes that differentiate the services using the buffer, the QoS scheme to be introduced in the following discussion takes advantage of the offset time, which was explained earlier. We call this an *offset-time-based QoS scheme*. For this purpose, we introduce a new offset time, which is called the *extra offset time*. Note that the offset time introduced previously, which is called the *base offset time* is different from the extra offset time.

We now explain how the offset-time-based QoS scheme works [22,23]. In particular, we explain how *class isolation* (or service differentiation) can be achieved by using an extra offset time in both cases with and without using fiber delay lines (FDLs) at an OBS node. Note that one may distinguish the following two different contentions in reserving resources (wavelengths and FDLs): the *intraclass contentions*, caused by requests belonging to the same class; and the *interclass contentions*, caused by requests belonging to different classes. In what follows, we focus on how to resolve interclass contentions using the offset-time-based QoS scheme.

For simplicity, we assume that there are only two classes of (OBS) services: classes 0 and 1, where class 1 has priority over class 0. In the offset-time-based QoS scheme, to give class 1 a higher priority for resource reservation, an extra offset time, denoted by t_o^1, is given to class 1 traffic (but not to class 0, i.e., $t_o^0 = 0$). In addition, we also assume that the base offset time is negligible as compared to the extra offset time, and will refer to the latter as simply the *offset time* hereafter. Finally, without loss of generality, we also assume that a link has only one wavelength for data (and an additional wavelength for control).

3.3.1. The Case Without FDLs. In the following discussion, let t_a^i and t_s^i be the arriving time and the service-start time for a class i request denoted by req(i), respectively, and let l_i be the burst length requested by req(i), where $i = 0, 1$. Figure 10 illustrates why a class 1 request that is assigned an (extra) offset time obtains a higher priority for wavelength reservation than does a class 0 request in the case of no FDLs. We assume that there is no burst (that arrived earlier) in service when the first request arrives. Consider the following two situations where contentions among two classes of traffic are possible.

In the first case as illustrated in Fig. 10a, req(1) comes first and reserves a wavelength using DR, and req(0) comes afterward. Clearly, req(1) will succeed, but req(0) will be blocked if $t_a^0 < t_s^1$ but $t_a^0 + l_0 > t_s^1$, or if $t_s^1 < t_a^0 < t_s^1 + l_1$. In the second case, as in Fig. 10b, req(0) arrives first, followed by req(1). When $t_a^1 < t_a^0 + l_0$, req(1) would be blocked had no offset time been assigned to req(1) (i.e., $t_o^1 = 0$). However, such a blocking can be avoided by using a sufficient offset time so that $t_s^1 = t_a^1 + t_o^1 > t_a^0 + l_0$. Given that t_a^1 may only be slightly behind t_a^0, t_o^1 needs to be larger than the maximum burst length over all bursts in class 0 in order for req(1) to completely avoid being blocked by req(0). With that much of offset time, the *blocking probability* of (the bursts in) class 1 becomes only a function of the offered load belonging to class 1, that is, independent of the offered load belonging to class 0. However, the blocking probability of class 0 is determined by the offered load belonging to both classes.

3.3.2. The Case with FDLs. Although the offset-time-based QoS scheme does not mandate the use of FDLs, its QoS performance can be significantly improved even with limited FDLs so as to resolve contentions for bandwidth among multiple bursts. For the case with FDLs, the variable B will be used to denote the maximum delay that a FDL (or the longest FDL) can provide. Thus, in the case of blocking, a burst can be delayed up to the maximum delay B.

Figure 11a,b illustrates class isolation at an OBS node equipped with FDLs where contention for both wavelength and FDL reservation may occur. In Fig. 11a, let us assume that when req(0) arrives at $t_a^0(t_s^0)$, the wavelength is in use by a burst that arrived earlier. Thus, the burst corresponding to req(0) has to be delayed (blocked) for t_b^0 units. Note that the value of t_b^0 ranges from 0 to B, and a FDL with an appropriate length that can provide a delay of t_b^0 will be chosen. Accordingly, if $t_b^0 < B$, the FDL is reserved for a class 0 burst as shown in Fig. 11b (the burst will be dropped if t_b^0 exceeds B), and the wavelength will

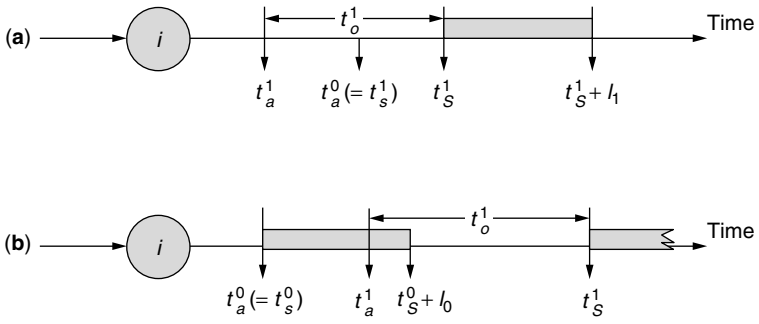


Figure 10. Class isolation in the case without FDLs.

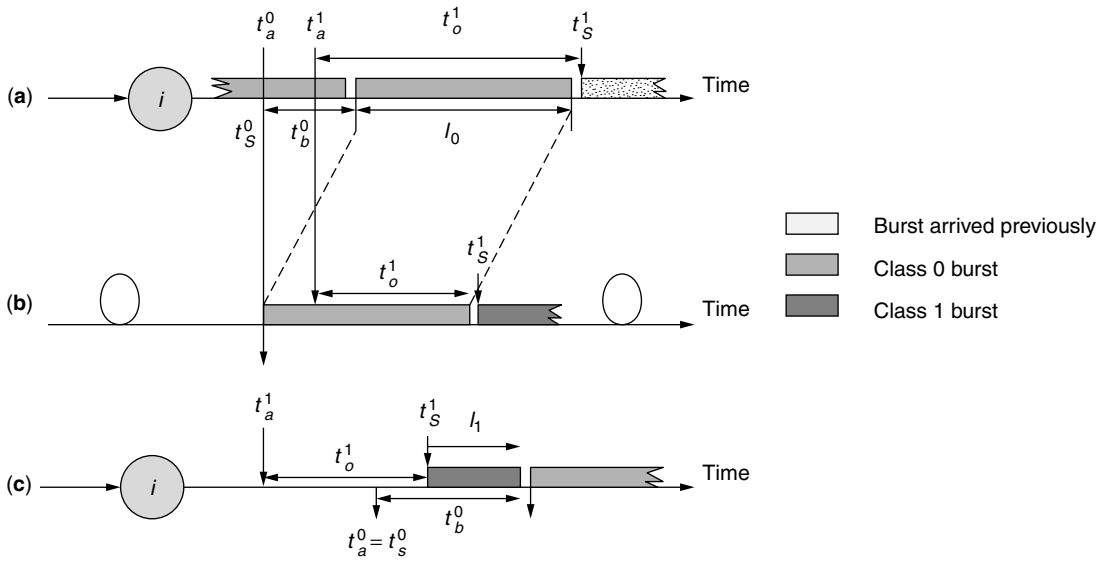


Figure 11. Class isolation in the case with FDLs.

be reserved from $t_s^0 + t_b^0$ to $t_s^0 + t_b^0 + l_0$ as shown in Fig. 11a. Now assume that req(1) arrives later at t_a^1 (where $t_a^1 > t_a^0$) and tries to reserve the wavelength. req(1) will succeed in reserving the wavelength as long as the offset time is so long that $t_s^1 = t_a^1 + t_o^1 > t_a^0 + t_b^0 + l_0$. Note that had req(1) arrived earlier than req(0) in Fig. 11a it is obvious that req(1) would not have interclass contention caused by req(0). This illustrates that class 1 can be isolated from class 0 when reserving a wavelength because of the offset time. Of course, without the offset time, req(1) would be blocked for $t_a^0 + t_b^0 + l_0 - t_a^1$, and it would be entirely up to the use of FDLs to resolve this interclass contention.

Similarly, Fig. 11b illustrates class isolation in FDL reservation. More specifically, let us assume that req(0) has reserved the FDLs as described earlier, and because t_o^1 is not long enough, req(1) would be blocked in wavelength reservation and thus needs to reserve the FDLs. In such a case, req(1) will successfully reserve the FDLs if the offset time is still long enough to have $t_s^1 = t_a^1 + t_o^1 > t_a^0 + l_0$. Otherwise (i.e., if $t_s^1 < t_a^0 + l_0$), req(1) would contend with req(0) in reserving the FDL and would be dropped.

As shown in Fig. 11c, if req(1) comes first and reserves the wavelength based on t_o^1 and delayed reservation (DR), and req(0) comes afterward, req(1) is not affected by req(0). However, req(0) will be blocked either when $t_a^1 < t_a^0 < t_s^1$ but $t_a^0 + l_0 > t_s^1$, or when $t_s^1 < t_a^0 < t_s^1 + l_1$. Similarly, if req(1) arrives first, it can reserve the FDL first regardless of whether req(0) succeeds in reserving the FDL. As mentioned earlier, this implies that class 1 can be isolated from class 0 in reserving both the wavelength and the FDL by using an appropriate offset time, which explicitly gives class 1 a higher priority over class 0. As a result of having a low priority on resource reservations, class 0 bursts will have a relatively high blocking and loss probability.

Although the offset-time-based QoS scheme does provide good service differentiation even when buffering is not possible, it has some disadvantages that need to be enhanced. For example, the offset-time-based QoS scheme introduces delay overhead caused by the extra offset time. Since the highest class suffers from the longest delay [22,23], the QoS provisioning will be strictly restricted in a given delay budget. It also lacks the controllability on the QoS, which can be enhanced by introducing the measurement based QoS provisioning and assigning the appropriate weight to each QoS class. In addition, it is worth considering how the offset-time-based QoS scheme can be integrated with the existing QoS domain such as DiffServ.

3.4. Burst Assembly

As we mentioned earlier, the *burst assembly* process takes place at the ingress edge router. The incoming data (e.g., IP packets) are assembled into a super packet, which is called the *data burst*. In the following discussion, we look into the issues of burst assembly.

It is obvious that IP packets would be the dominant traffic in future networks. Unlike the traffic from traditional telephone networks, Internet traffic is quite difficult to predict. This is because Internet traffic shows self-similarity [24,25]. *Self-similarity* means that even with a high degree of multiplexing, the burstiness of traffic still exists at all timescales, which makes the network resource dimensioning and traffic engineering harder.

One advantage of the burst assembly is that it may reduce the self-similarity of Internet traffic [26]. Figure 12 shows an example configuration of an OBS network for burst assembly [27]. IP routers in the IP domain inject IP packets into the OBS network. The OBS edge routers located at the boundary of the OBS network take IP packets and perform the burst assembly process.

The incoming IP packets are classified and queued into an assembly buffer according to their destination and QoS requirements. A simple burst assembly process is illustrated in Fig. 13. The burst assembler at the end of the assembly buffer runs a timer, which expires at a given time (i.e., the assemble time). Whenever the timer expires, the burst assembler takes the burst (multiple IP packets queued during a given period of time) out of the assemble buffer for transmission. The key parameter of the burst assembler is the *assemble time*, which controls the size of the data burst. The distribution of assemble time could be deterministic or random, in which case it is called either a constant assemble time (CAT) or a variable assemble time (VAT), respectively. The burst assembler waits for a constant time when using the CAT, whereas it waits for a random amount of time when using the VAT. Alternatively, the burst assembler adaptively controls the assemble time by monitoring both assemble time and the

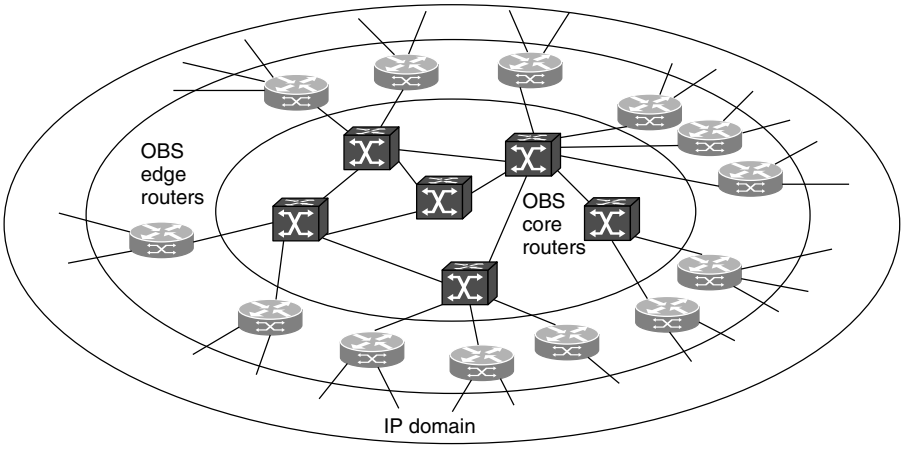


Figure 12. OBS edge routers for burst assembly.

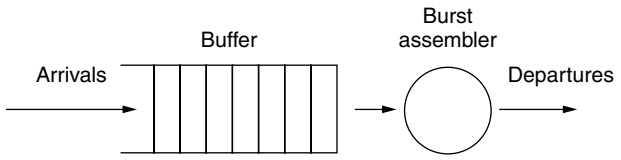


Figure 13. Burst assembly process.

size of the assembled IP packets. The optimization of the burst assembly process is a topic that requires further investigation.

4. SUMMARY

Because of its advantages over conventional communication networks, optical networks have been firmly positioned as a feasible solution for the next-generation networks. The optical networking technology keeps improving starting, from SONET to static wavelength routing network, and eventually, to dynamic optical networks. We have discussed optical switching techniques available today for optical networks: wavelength routing, optical burst switching, and optical packet switching.

Table 1 summarizes the qualitative comparison among these techniques. Wavelength routing is a simple and technically matured technology, but it may result in poor performance. On the other hand, optical packet switching may show its dynamicity in networking as in an electrical packet-switching network, but is technically immature for the implementation. Optical burst switching achieves balance between wavelength routing and optical packet switching by improving the inefficiency of wavelength routing and at the same time, by lessening the technical difficulty of optical packet switching. Thus, while the optical technology is mature enough to implement optical packet switching, optical burst switching appears to be a practical interim solution for optical networks.

Acknowledgments
This work was supported in part by the Korean Science and Engineering Foundation (KOSEF) through the OIRC project.

BIOGRAPHIES

Myungsik Yoo received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Korea, in 1989 and 1991, respectively, and the Ph.D. degree in electrical engineering from State University of New York at Buffalo (SUNY at Buffalo), Buffalo, New York (USA) in 2000. He was a Senior Research Engineer in Nokia Research Center, Burlington, Massachusetts. Since 2000, he has been an Assistant Professor in the School of Electronic Engineering, Soongsil University, Seoul, Korea. His current research interests are optical networks and optical Internet, including optical burst switching, protection/restoration, QoS support, and GMPLS.

Chunming Qiao is an Associate Professor at the University at Buffalo (SUNY). He has over 10 years of academic and industrial experience in optical networks. Dr. Qiao has published more than 100 papers in leading technical journals and conference proceedings, and several book chapters. He has given several keynote speeches, tutorials, and invited talks and is recognized for his pioneering research on optical Internet and in particular, the optical burst switching (OBS) paradigm. Dr. Qiao is the IEEE Communication Society's Editor-at-Large for optical networking and computing; an editor of several other journals and magazines, including *IEEE/ACM Transactions on Networking* (ToN), as well as a guest editor for *IEEE JSAC* and other publications. He has chaired and co-chaired many conferences and workshops on optical communications and networking, including the Optical Networks Symposium (ICC'03), and Opticomm 2002. Dr. Qiao is also the founder and Chair of the Technical Group on Optical Networks (TGON) sponsored by SPIE, and a Vice Chair of the IEEE Technical Committee on Gigabit Networking (TCGN).

Table 1. Comparison Between Optical Switching Paradigms

Switching Paradigm	Bandwidth Utilization	Implementation Difficulty	Switching Granularity
Wavelength routing	Poor	Low	Coarse
Optical packet switching	High	High (not mature)	Fine
Optical burst switching	Moderate	Moderate	Moderate

BIBLIOGRAPHY

1. B. Rajagopalan et al., A framework for generalized multi-protocol label switching (GMPLS), IETF Internet draft.
2. R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Approach*, Morgan Kauffman, San Francisco.
3. W. Goralski, *Optical Networking & WDM*, McGraw-Hill, New York.
4. C. Qiao and M. Yoo, Optical burst switching (OBS)—a new paradigm for an optical Internet, *J. High Speed Network* **8**(1): 69–84 (1999).
5. C. Qiao and M. Yoo, Choice, features and issues in optical burst switching, *Opt. Network Mag.* **1**(2): 36–44 (May 2000).
6. Y. Xiong, M. Vandenhoute, and H. C. Cankaya, Control architecture in optical burst-switched WDM networks, *IEEE J. Select. Areas Commun.* **18**(10): 1838–1851 (Oct. 2000).
7. J. Turner, Terbit burst switching, *J. High Speed Network* **8**(1): 1–18 (1999).
8. S. Yao et al., All-optical packet switching for metropolitan area networks: opportunities and challenges, *IEEE Commun. Mag.* **39**(3): 142–148 (March 2001).
9. M. J. O'Mahony et al., The application of optical packet switching in future communication networks, *IEEE Commun. Mag.* **39**(3): 128–135 (March 2001).
10. X. Lisong et al., Techniques for optical packet switching and optical burst switching, *IEEE Commun. Mag.* **39**(1): 136–142 (Jan. 2001).

11. E. Rosen et al., *Multiprotocol Label Switching Architecture*, IETF RFC 3031.

12. B. Rajagopalan et al., IP over optical networks: A framework, IETF Internet draft.

13. R. Ramaswami and A. Segall, Distributed network control for wavelength routed optical networks, *IEEE/ACM Trans. Network.* **5**(6): 936–943 (Dec. 1997).

14. C. Qiao and Y. Mei, Wavelength reservation under distributed control, *IEEE/LEOS Broadband Opt. Network.* 45–46 (1996).

15. Models of blocking probability in all-optical networks with and without wavelength changers, *IEEE J. Select. Areas Commun.* **14**(5): 858–867 (June 1996).

16. R. Ramaswami and K. Sivarajan, Routing and wavelength assignment in all-optical network, *IEEE/ACM Trans. Network.* 489–500 (Oct. 1995).

17. P. Ashwood-Smith et al., Generalized MPLS signaling—RSVP-TE extensions, IETF Internet draft.

18. P. Ashwood-Smith et al., Generalized MPLS signaling—CR-LDP extensions, IETF Internet draft.

19. J. S. Turner, WDM burst switching for petabit data networks, *Proc. OFC'2000*, 2000, Vol. 2, pp. 47–49.

20. L. Y. Lin and E. L. Goldstein, MEMS for free-space optical switching, *Proc. LEOS'99*, 1999, Vol. 2. pp. 483–484.

21. R. Awdeh and H. T. Mouftah, Survey of ATM switch architecture, *IEEE Commun. Mag.* **27**: 1567–1613 (Nov. 1995).

22. M. Yoo, C. Qiao, and S. Dixit, QoS performance of optical burst switching in IP-over-WDM networks, *IEEE J. Select. Areas Commun.* **18**(10): 2062–2071 (Oct. 2000).

23. M. Yoo, C. Qiao, and S. Dixit, Optical burst switching for service differentiation in the next generation optical Internet, *IEEE Commun. Mag.* **39**(2): 98–104 (Feb. 2001).

24. V. Paxon and S. Floyd, Wide area traffic: the failure of Poisson modeling, *IEEE Trans. Network.* **3**(3): 226–244 (1995).

25. W. Leland et al., On the self-similar nature of Ethernet traffic (extended version), *IEEE Trans. Network.* **2**(1): 1–15 (1994).

26. A. Ge, F. Callegati, and L. Tamil, On optical burst switching and self-similar traffic, *IEEE Commun. Lett.* **4**(3): 98–100 (March 2000).

27. A. Detti, A. Eramo, and M. Listanti, Performance evaluation of a new technique for IP support in a WDM optical network: Optical composite burst switching (OCBS), *J. Lightwave Technol.* **20**(2): 154–165 (Feb. 2002).

OPTICAL SYNCHRONOUS CDMA SYSTEMS

Tomoaki Ohtsuki
Tokyo University of Science
Noda, Chiba, Japan

Iwao Sasase
Keio University
Yokohama, Japan

1. INTRODUCTION

Optical communication systems in the optical fiber play a main part of the digital communications in backbone networks, high speed local-area networks (LANs) using a fiber distributed data interface (FDDI), metropolitan-area network (MAN), and a next-generation subscriber system such as a fiber to the home (FTTH). The main advantages of the optical fiber communications are the high speed, large capacity and high reliability by the use of the broadband of the optical fiber. A desirable feature for future optical networks would be the ability to process information directly in the optical domain for purposes of multiplexing, demultiplexing, filtering, amplification, and correlation. Optical signal processing would be advantageous since it can potentially be much faster than electrical signal processing and the need for photon–electron–photon conversion would be obviated.

Asynchronous multiple-access methods where network access is random and collisions occur, such as token passing and carrier-sense multiple-access, are well suited to LANs with low traffic demand. However, these asynchronous access methods suffer from cumulative delay as the traffic intensity increases. Also, contention protocols generally proposed for low traffic demands are not suitable if traffic delay is a major issue, as, in networks where information must be transmitted simultaneously. On the other hand, synchronous accessing methods where transmissions are perfectly scheduled provide more successful transmissions than asynchronous methods. As a typical synchronous protocol, time-division multiple access (TDMA) is an efficient multiple-access protocol in networks with heavy traffic demands, since it can accommodate higher traffic demands and do not suffer from cumulative delay. However, in situations where the channel is sparsely used, TDMA is inefficient.

As an alternate optical multiplexing technique, there is wavelength-division multiple access (WDMA). The term WDMA rather than the popular wavelength-division multiplexing (WDM) is used to indicate the access, routing and switching functionality in addition to the transmission multiplex. Tuned lasers are used as the optical source for each transmitter, and the modulated data are transmitted within its assigned band. At the receiver, an optical filter is tuned to the desired band, while all others are filtered out. A photodetector and a decoder are followed to obtain the data. WDMA technique partitions the available spectrum to different users, and offers a means of increasing capacity at minimal cost within the existing optical fiber infrastructure. The fundamental disadvantage in WDMA is that sophisticated hardware such as wavelength-controlled tunable lasers and high-quality narrowband tunable filters for each channel is required. Although WDMA can be used as a degree of design freedom with respect to routing and wavelength selection, the usable wavelength might be limited because of the crosstalk caused by the nonlinearity within the optical fiber. Wavelength routing can offer the switching function for dense WDMA networks; however, it may cause the crosstalk problem in the cross-connects based on space and wavelength, and thus, network reliability and flexibility are restricted.

Code-division multiple-access (CDMA) is a multiple access protocol that is efficient with low traffic and has zero access delay. Especially, direct detection optical CDMA

systems have been investigated widely to apply for high-speed LAN, because they allow multiple users to access the network simultaneously. In the case of data transfer where traffic tends to be bursty rather than continuous, CDMA can be used for contention-free, zero delay access. Compared with TDMA, CDMA is attractive in other points. Channel assignment is much easier with CDMA. CDMA isolates irregular channels so that they do not influence other channels, while with TDMA, even one irregular channel, such as continuous emission from a transmitter, causes the failure of all other channels.

In optical CDMA, incoherent systems using narrow pulse laser sources are mainly implemented, since optical links have vast bandwidth and the optical components can produce very narrow pulse precisely in time and offer extrahigh optical signal processing. In the transmitter and receiver, low-cost devices with high cost performance and high reliability, such as Fabry–Perot laser diode and avalanche and pin photodiodes, are available. Thus, in optical CDMA, intensity modulation/direct detection (IM/DD) is mainly used. In IM/DD systems, other arriving pulse sequences having positive pulses happen to overlap a pulse of the desired sequence, and produce correlation crosstalk. In optical CDMA, multiple user interference called multiple-access interference (MAI) is dominant compared to photodetector shot noise, dark current and thermal noise. Thus, the elimination or suppression of MAI is the key issue in optical CDMA. Most published optical CDMA systems are based on discrimination in the time domain to reduce the effects of pulse overlaps. This time-encoding process is most commonly implemented by encoding each data bit with a high-rate sequence; that is, a pulse laser source is intensity-modulated by electrical (0,1) data bit and the narrow pulse is emitted in the first chip in a slot. Here, data are usually modulated in on/off keying (OOK) or pulse position modulation (PPM) formats, and a slot is divided into chips where the number of chips in a slot equals to the length of the spreading code consisting of 1 and 0 allocated for users. Then, in the time encoder, a narrow pulse is time-spread into several chips within the slot according to each user's unipolar signature code. Thus, the time-encoding process relies on a simple, intensity-based, pulse time addressing process, and the sequence encoder and decoder in the time domain can be easily and cost-effectively implemented by using tapped optical delay lines. At the receiver, optical incoherent direct detection is done by an optical delay-line decoder matched to the encoder at the transmitter. After decoding, unwanted signals are time-spread over much larger time intervals than is the desired user's signal, and the crosstalk from adjacent chips are rejected to some extent by this time-despreading process.

For optical CDMA systems, both asynchronous and synchronous systems have been studied. In an asynchronous optical CDMA system, the synchronization among users is not required, and optical orthogonal codes (OOC) with good correlation properties [1] are widely used. However, in asynchronous CDMA systems, the available number of signature sequence codes is very small; hence the number of users is very limited. To solve this problem, synchronous CDMA systems, in which all users are synchronized in frame, is considered. With synchronous CDMA, the available number of signature codes is larger than that of asynchronous CDMA systems, because the same code can be reused with different phases. The modified prime sequence codes are known as typical "signature codes for optical synchronous CDMA," in which time-shifted versions of the prime sequence code can be used [2]. The cross-correlation peak between two time-shifted versions of the sequence code is as high as the autocorrelation peak; however, it always occurs either delayed or ahead of the autocorrelation peak. Since in the synchronized CDMA the receiver can be synchronized to the expected position of the autocorrelation peak, the autocorrelation peak can be distinguished from adjacent cross-correlation peaks. For a given value of bit error rate, synchronous CDMA systems can accommodate more simultaneous users than asynchronous CDMA systems. Furthermore, synchronous CDMA can be efficiently used in conjunction with TDMA and WDMA on multimedia communication networks where multiple services with different traffic requirements are to be integrated.

In Section 2, a family of good optical unipolar pseudoorthogonal (non-zero-cross-correlation) codes suitable for optimal CDMA IM/DD system with OOK and PPM signaling is described. In Section 3, the IM/DD systems with OOK and PPM signaling are described as typical optical synchronous CDMA systems. Also, as an alternative optical CDMA system, a frequency-encoded spread-time optical CDMA system utilizing bipolar codes is briefly introduced. Since the performance of the optical CDMA is degraded by the multiple-user interference, the interference cancellation is the key to realize the practical optical CDMA system. In Section 4, we describe two typical interference cancellation techniques for optical synchronous CDMA, based on the use of properties of modified prime sequence codes and optical hard-limiter.

2. SEQUENCES FOR OPTICAL SYNCHRONOUS CDMA

Many classes of binary signature sequences that are suitable for radio CDMA have been studied. In most of these codes, a strong autocorrelation peak and zero-cross-correlation function can be obtained through bipolar $(-1,+1)$ sequences. However, optical IM/DD systems can only accommodate unipolar $(0,+1)$ sequences, since incoherent systems use only positive narrow pulses emitted from laser sources. Therefore, codes intended for communication systems in which both positive and negative levels are available, are not necessarily optimal in a fiberoptic environment using optical signal processing. Compared to conventional electronic bipolar $(-1,+1)$ codes such as maximum-length sequence codes and Gold codes, the cross-correlation function of unipolar codes is high and the number of codes in the family is very low. The minimum value of the cross-correlation that unipolar codes can achieve is limited to be one, since at least one pulse is overlapped for asynchronous unipolar sequences. Thus, sets of sequences having no more than one pulse overlap in the pairwise cross-correlation are often called as *pseudoorthogonal codes* in optical CDMA.

The characteristics needed for the unipolar codes suitable for optical CDMA are good autocorrelation and cross-correlation properties. Sharp autocorrelation property is needed to achieve the synchronization as well as the discrimination between the time-shifted sequence codes in optical synchronous CDMA where the time-shifted codes are assigned to other users. The minimum value of the cross-correlation function should be as small as possible to discriminate between the desired user and others as well as to mitigate multiple user interference. In optical CDMA IM/DD systems, other arriving pulse sequences having positive pulses that happen to overlap a pulse of the desired sequence produce correlation crosstalk called *multiple access interference* (MAI), which might degrade the decoding performance. The code length and weight (the number of "1"s) also affect the system performance. Long codes and sparse codes comprising very few ones and narrow pulses are preferred to support a large number of users and higher transmission capabilities, respectively. On the other hand, short codes and a large weight are preferred to increase data rate and improve signal-to-interference ratio, respectively.

In this section, as good optical unipolar codes suitable for optical CDMA IM/DD system with OOK and PPM signaling, the prime code family is described. Especially for optical synchronous CDMA systems, the family of modified prime sequences is known to have good correlation properties.

2.1. Prime Code

Prime codes are defined as follows [3]: from the Galois field $\mathrm{GF}(P) = \{0, 1, 2, \ldots, P-1\}$, where P is a prime number, a set of P prime sequences $\{S_i^P : i = 0 \text{ to } P-1, \text{ and } j = 0 \text{ to } P-1\}$ is first generated, each with P elements given by

$$S_i^P(j) = [i,j]_P \quad \text{for } i = 0 \text{ to } P-1, \text{ and } j = 0 \text{ to } P-1 \tag{1}$$

where $[\ ]_P$ denotes reduction modulo P. For example, the family of prime sequences for $P = 7$ is shown in Table 1. For each sequence S_i^P, a binary code C_i^P of length P^2 chips is then constructed using the following rules:

$$C_i^P(n) = \begin{cases} 1, & \text{for } n = jP + S_i^P(j) \quad \text{for } j = 0 \text{ to } P-1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

This requires that the code C_i^P be divided into P frames, each consisting of P chips. Within the jth frame, the chip shifted relative to the start of the frame by $S_i^P(j)$ is a "1"; all other chips are zero. The code C_i^P is therefore a time-mapped, binary version of the sequence S_i^P. The set of prime codes for $P = 7$ is shown in Table 2, where the frames have been slightly separated for clarity.

Table 1. Prime Sequences for $P = 7$

	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$
Sequence S_0^7	0	0	0	0	0	0	0
Sequence S_1^7	0	1	2	3	4	5	6
Sequence S_2^7	0	2	4	6	1	3	5
Sequence S_3^7	0	3	6	2	5	1	4
Sequence S_4^7	0	4	1	5	2	6	3
Sequence S_5^7	0	5	3	1	6	4	2
Sequence S_6^7	0	6	5	4	3	2	1

The correlation functions arising in an IM/DD system, assuming on/off keyed data, are the aperiodic and periodic correlation functions, $C_{i,j}(l)$ and $\Theta_{i,j}(l)$, which are defined respectively as follows:

$$C_{i,j}(l) = \sum_{n=0}^{L-1} C_i^P(n) \cdot C_j^P(n+l) \tag{3}$$

$$\Theta_{i,j}(l) = \sum_{n=0}^{L-1} C_i^P(n) \cdot C_j^P([n+l]_L) = C_{i,j}([l]_L) + C_{i,j}([l]_L - L) \tag{4}$$

where $L = P^2$, $[\ ]_L$ denotes reduction modulo L, and $C_i^P(n) = 0$ for $n < 0$ and $n \geq L$, for all i. These are autocorrelation functions when $i = j$, and cross-correlation functions when $i \neq j$. The aperiodic form is generated at the matched filter output by an isolated "1" in the incoming datastream, while the periodic form is generated by adjacent "1"s; incoming "0"s produce no response. This periodic correlation is simply the number of positions where C_i^P and a cyclically shifted version of C_j^P both have "1"s. This means that the autocorrelation peak for any code (which occurs for $l = 0$, $i = j$) is equal to the number of "1"s it contains, or P in the case of a prime code:

$$\Theta_{i,j}(0) = C_{i,j}(0) = P \quad \text{for all } i \tag{5}$$

Note that the maximum number of coincidences of "1"s between two distinct prime codes C_i^P and C_j^P, having any

Table 2. Prime Codes for $P = 7$

	Frame 0	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6
Code C_0^7	1000000	1000000	1000000	1000000	1000000	1000000	1000000
Code C_1^7	1000000	0100000	0010000	0001000	0000100	0000010	0000001
Code C_2^7	1000000	0010000	0000100	0000001	0100000	0001000	0000010
Code C_3^7	1000000	0001000	0000001	0010000	0000010	0100000	0000100
Code C_4^7	1000000	0000100	0100000	0000010	0010000	0000001	0001000
Code C_5^7	1000000	0000010	0001000	0100000	0000001	0000100	0010000
Code C_6^7	1000000	0000001	0000010	0000100	0001000	0010000	0100000

relative shift, is 2, so that all periodic cross-correlation functions are bounded by

$$\Theta_{i,j}(0) \le 2 \quad \text{for all } l, \text{ and all } i, j \text{ such that } i \ne j \tag{6}$$

From Eq. (4), it is clear that $\Theta_{i,j}(l) \ge C_{i,j}(l)$ for all l (note that Θ and C are both positive functions); thus the above bound also applies to any interference contribution, regardless of the data it carries.

2.2. Quasiprime Code

Quasiprime codes are derived from prime codes as follows [4]: with the prime code C_i^P, a set of cyclically shifted versions of the code $\{C_{ik}^{SP} : k = 0 \text{ to } P-1\}$ is defined where C_{ik}^{SP} is obtained by cyclically shifting C_i^P left by k complete frames. The elements of C_{ik}^{SP} are thus given by

$$C_{ik}^{SP}(n) = C_i^P([n + kP]_L) \qquad \text{for } n = 0 \text{ to } P^2 - 1. \tag{7}$$

For each shifted code C_{ik}^{SP}, a quasiprime code C_{ik}^{SP} may then be defined for any positive integer Q, where

$$\begin{aligned} C_{ik}^{QP}(n) &= C_{ik}^{SP}([n]_L) \quad \text{for } n = 0 \text{ to } QP - 1 \\ &= C_i^P([n + kP]_L) \end{aligned} \tag{8}$$

Each quasi-prime code then comprises QP chips taken cyclically from a shifted prime code, and contains Q "1"s. For example, Table 3 shows two of the seven quasiprime codes in the set $\{C_{2k}^{87} : k = 0 \text{ to } 6\}$ together with prime code C_2^7. Both are derived from the same prime code C_2^7.

We have to take care to define correlation functions for quasiprime codes, because truncating or extending the shifted prime codes destroys the periodicity of the basic prime code, and this must be restored if the quasiprime codes are to show good periodic cross-correlation properties. Accordingly, when the length QP of the quasiprime codes lies between $\Lambda - 1$ and Λ lengths of the original prime codes, it is made up to Λ lengths by packing each code with "0"s. The aperiodic and periodic correlation functions are then defined as

$$C_{i,j}(l) = \sum_{n=0}^{\Lambda L - 1} C_{ik}^{QP}(n) \cdot C_{jl}^{QP}(n + l) \tag{9}$$

$$\Theta_{i,j}(l) = \sum_{n=0}^{\Lambda L - 1} C_{ik}^{QP}(n) \cdot C_{jl}^{QP}([n + l]_{\Lambda L}) \tag{10}$$

where $(\Lambda - 1)L < QP \le \Lambda L$, $[\ ]_{\Lambda L}$ denotes reduction modulo ΛL, and $C_{ik}^{QP}(n) = 0$ for $n < 0$ and $n \ge QP$, for all i, k.

Different quasiprime codes derived from the same prime code cannot act as distinct, orthogonal members of an asynchronous CDMA code set. This means that a quasiprime code set can contain a maximum of P codes (one for each code in the original prime code set). The correlation properties of such a set are as follows: first, each code contains Q "1"s, so that the autocorrelation peak is given by

$$\Theta_{i,j}(0) = C_{i,j}(0) = Q \quad \text{for all } i \tag{11}$$

In addition the periodic cross-correlation $\Theta_{i,j}$ between two distinct quasiprime codes C_{ik}^{QP} and C_{jl}^{QP} is bounded by

$$\Theta_{i,j}(l) \le 2\Lambda \quad \text{for all } l, \text{ and all } i, j, k, l \text{ such that } i \ne j \tag{12}$$

Considering (11) and (12), the interference probability obtained with quasiprime codes might be expected to be worse than that obtained with prime codes in general. For example, the number of interfering signals that can be accommodated without error is expected to be $\lfloor Q/2\Lambda \rfloor$, where $\lfloor x \rfloor$ is the integer part of x, and this is less than or equal to the prime code result $\lfloor P/2 \rfloor$. In fact, this is pessimistic, because when QP is only slightly greater than $(\Lambda - 1)L$, the bound $\Theta_{i,j}(l) \le 2(\Lambda - 1)$ can still apply, and in such cases the quasiprime codes would be able to achieve better interference probability.

2.3. 2^n Prime Code

Usually, the 2^n codes are defined as collections of binary N-tuples with weight 2^n [5]. Using the serial optical encoders, the distribution of the pulses in each generated codeword must be symmetric (i.e., the distribution of the current 2^m pulses highly depends on that of the previous 2^{m-1} pulses, where $1 < m \le n$) and results in a very restrictive pulse distribution constraint. Alternatively, it is sometimes more convenient to represent the pulse distribution in terms of delay distribution. Therefore, the constraint can be equivalently presented as a delay distribution constraint.

The delay distribution constraint functions as follows. For a given integer n, integers x, y, and z are assumed such that $x \ne y$, $0 \le x \le 2^n - 2$, $0 \le y \le 2^n - 2$, and $1 \le z \le n - 1$. If both x and y are divisible by 2^z, then adjacent relative cyclic delays $[t_0, t_1, \ldots, t_q, \ldots, t_{2^n-1}]$ of each codeword of the 2^n codes are related such that

$$t_{x \uplus (2^{z-1}-1) \uplus m} = t_{y \uplus (2^{z-1}-1) \uplus m} \tag{13}$$

for a given integer $m \in [0, 2^n - 1]$, where "$\uplus$" represents modulo-2^n addition. t_q denotes the adjacent relative cyclic delay (or simply the separation in chips) between the qth

Table 3. Quasiprime Codes for $Q = 8$, $P = 7$ Based on C_2^7

Code	Frame 0	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7
C_2^7	1000000	0010000	0000100	0000001	0100000	0001000	0000010	
C_{20}^{87}	1000000	0010000	0000100	0000001	0100000	0001000	0000010	1000000
C_{25}^{87}	0001000	0000010	1000000	0010000	0000100	0000001	0100000	0001000

and $(q+1)$th pulses. For the last delay t_{2^n-1}, the codeword is wrapped around to obtain the separation between the last and first pulses.

Example 1. Assuming that $n = 2$ (i.e., code weight of 4), we get $0 \le x \le 2$, $0 \le y \le 2$, $z = 1$, and $m \in [0, 3]$ from the above mentioned delay distribution constraint. The adjacent relative cyclic delays $[t_0, t_1, t_2, t_3]$ are related such that $t_0 = t_2$ for $m = \{0, 2\}$, or $t_1 = t_3$ for $m = \{1, 3\}$. Using sequence codes 100001010000100, 110000100010000, and 100010000100100 as examples, their corresponding adjacent relative cyclic delays are [5,2,5,3], [1,5,4,5], and [4,5,3,3], respectively. On the basis of the constraint, the first two are valid sequence codes, while the last one is not.

According to the symmetric property described above, the algebraic construction on the 2^n prime-sequence codes begins with Galois field $\mathrm{GF}(P) = \{0, 1, \ldots, P-1\}$ of a prime number P. As an example, the prime sequences in GF(13) are shown in Table 4 in a form different from that in Table 1. By inspection, the adjacent relative cyclic delays of the codeword generated by S_i^P can be found according to

$$t_j = \begin{cases} S_i^P(j+1) - S_i^P(j) + P, & \text{for } j = \{0, 1, \ldots, P-2\} \\ S_i^P(0) - S_i^P(j) + P, & \text{for } j = P-1 \end{cases} \tag{14}$$

for $i \in GF(P)$. Table 5 shows the adjacent relative cyclic delays for the prime sequences in GF(13).

The adjacent relative cyclic delays for each prime sequence S_i^P are then determined depending on whether the delay-distribution constraint is satisfied. If the constraint is satisfied, the prime sequence S_i^P will be modified: the elements $S_i^P(j)$ and $S_i^P(j+1)$ whose relative cyclic delay t_j satisfies the constraint are kept unchanged, while the remaining elements are replaced by Xs. Note that every X in the replaced prime sequences is simply mapped to P zeros. However, this S_i^P will be discarded if none of the delays satisfies the constraint.

Example 2. Using $2^n = 8$ and $i = 3$ as an example, the adjacent relative cyclic delays for S_3^{13} are [16, 16, 16, **16**, **3**, **16**, **16**, **16**, **3**, **16**, 16, 16, 3] as shown in Table 5. Those delays that satisfy the delay distribution constraint are boldfaced. From (13), S_3^{13} satisfies the two conditions $t_m = t_{m+2} = t_{m+4} = t_{m+6}$ and $t_{m+1} = t_{m+5}$ with $m = 3$. The remaining elements $S_3^{13}(0)$, $S_3^{13}(1)$, $S_3^{13}(2)$, $S_3^{13}(11)$, and $S_3^{13}(12)$ are then replaced by Xs as shown in Table 6, where the replaced prime sequences for $P = 13$ and $2^n = 8$ are tabulated. Note that the prime sequences S_5^{13} and S_8^{13} are discarded since the delay distribution constraint cannot be satisfied.

Finally, the codewords of the 2^n prime sequence codes are generated by mapping each replaced prime sequence S_i^P into a binary code sequence $C_i^P = (C_i^P(0), C_i^P(1), \ldots, C_i^P(k), \ldots, C_i^P(N-1))$ of length $N = P^2$ according to

$$C_i^P(k) = \begin{cases} 0, & \text{for } k = S_i^P(j) + jP \text{ and } S_i^P(j) \ne \mathrm{X} \\ 1, & \text{otherwise} \end{cases} \tag{15}$$

for $i, j \in \mathrm{GF}(P)$ and $k = \{0, 1, \ldots, N-1\}$.

2.4. Modified Prime Code

For prime sequence codes of length P^2, the number of sequence codes is limited to P; therefore, so is the number of total subscribers. In order to generate more sequence

Table 4. Prime Sequences in GF(13)

	j												
i	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9	10	11	12
2	0	2	4	6	8	10	12	1	3	5	7	9	11
3	0	3	6	9	12	2	5	8	11	1	4	7	10
4	0	4	8	12	3	7	11	2	6	10	1	5	9
5	0	5	10	2	7	12	4	9	1	6	11	3	8
6	0	6	12	5	11	4	10	3	9	2	8	1	7
7	0	7	1	8	2	9	3	10	4	11	5	12	6
8	0	8	3	11	6	1	9	4	12	7	2	10	5
9	0	9	5	1	10	6	2	11	7	3	12	8	4
10	0	10	7	4	1	11	8	5	2	12	9	6	3
11	0	11	9	7	5	3	1	12	10	8	6	4	2
12	0	12	11	10	9	8	7	6	5	4	3	2	1

Table 5. Adjacent Relative Cyclic Delays for the Prime Sequences in GF(13)

	j												
i	0	1	2	3	4	5	6	7	8	9	10	11	12
0	13	13	13	13	13	13	13	13	13	13	13	13	13
1	14	14	14	14	14	14	14	14	14	14	14	14	1
2	15	15	15	15	15	15	2	15	15	15	15	15	2
3	16	16	16	16	3	16	16	16	3	16	16	16	3
4	17	17	17	4	17	17	4	17	17	4	17	17	4
5	18	18	5	18	18	5	18	5	18	18	5	18	5
6	19	19	6	19	6	19	6	19	6	19	6	19	6
7	20	7	20	7	20	7	20	7	20	7	20	7	7
8	21	8	21	8	8	21	8	21	8	8	21	8	8
9	22	9	9	22	9	9	22	9	9	22	9	9	9
10	23	10	10	10	23	10	10	10	23	10	10	10	10
11	24	11	11	11	11	11	24	11	11	11	11	11	11
12	25	12	12	12	12	12	12	12	12	12	12	12	12

Table 6. Replaced Prime Sequences in GF(13) with $2^n = 8$

	j												
i	0	1	2	3	4	5	6	7	8	9	10	11	12
0	X	X	X	0	0	0	0	0	0	0	0	X	X
1	X	X	X	3	4	5	6	7	8	9	10	X	X
2	X	X	X	6	8	10	12	1	3	5	7	X	X
3	X	X	X	9	12	2	5	8	11	1	4	X	X
4	0	4	8	12	3	7	X	X	X	X	X	5	9
5													
6	X	X	X	5	11	4	10	3	9	2	8	X	X
7	X	X	X	8	2	9	3	10	4	11	5	X	X
8													
9	0	9	5	X	X	X	X	X	7	3	12	8	4
10	X	X	X	4	1	11	8	5	2	12	9	X	X
11	X	X	X	7	5	3	1	12	10	8	6	X	X
12	X	X	X	10	9	8	7	6	5	4	3	X	X

codes for the same length, that is, the same bandwidth expansion, at the expense of requiring synchronization among users, modified prime sequence codes have been proposed [2]. Modified prime sequence codes are time-shifted versions of prime sequence codes. Each original P prime sequence S_x^P is taken as a seed from which a group of new sequence codes can be generated. The sequence codes of the first group (i.e., $x = \{0\}$) are obtained by left-rotating the prime sequence code C_0^P. C_0^P can be left-rotated $P-1$ times before being recovered, so that $P-1$ new sequence codes can be generated from C_0^P. For the other $P-1$ groups (i.e., $x = \{1, \ldots, P-1\}$), the elements of the corresponding prime sequence S_x^P can be left-rotated $P-1$ times to create new prime sequences $S_{x,r}^P = (S_{x,r}^P(0), S_{x,r}^P(1), \ldots, S_{x,r}^P(P-1))$, where r represents the number of times S_x^P has been left-rotated. Therefore, P prime sequences per group are obtained. Each prime sequence $S_{x,r}^P$ is then mapped into a binary sequence code $C_{x,r}^P = (C_{x,r}^P(0), C_{x,r}^P(1), \ldots, C_{x,r}^P(i), \ldots, C_{x,r}^P(P^2-1))$ according to

$$\theta_{x,r}^P(i) = \begin{cases} 1, & \text{for } i = S_{x,r}^P(j) + jP, j = 0, 1, \ldots, P-1 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

The set of prime sequence S_x^P and their associated sequence codes C_x^P for GP(5) are tabulated in Table 7. The set of new prime sequences $S_{x,r}^P$ and their associated sequence codes $C_{x,r}^P$ for $GF(5)$ are tabulated in Table 8. Note that each new sequence code has P binary "1"s. Considering all groups, the total number of modified prime sequence codes is P^2. For a synchronous system, the cross-correlation between the modified prime sequence codes of the xth and the yth users can be written as [2]

$$\theta_{x,y} = \begin{cases} P, & x = y, \\ 0, & x \text{ and } y \text{ are in the same group} \\ 1, & x \text{ and } y \text{ are in the different groups} \end{cases} \tag{17}$$

The modified prime code has unique characteristics in that there is no correlation between any two users among the same group and the interference from other groups has the equal effect on the user in the same group. On the other hand, the optical orthogonal code (OOC) has the problem that it has to set the weights and sequence code independently and keep the number of spreading codes small to attain good correlation properties. Table 9 shows the OOCs with the code length $F = 32$ and weight $K = 4$. The total number of OOCs is given by the integer part of $(F-1)/(K^2-K)$, and there are only two codes when $K = 4$ and $F = 32$, because the maximum value of the cross-correlation should be 1. To satisfy this condition, the distance of any two "1"s should be different in all codes as shown in Table 9. Therefore, to increase the number of spreading codes in OOC, the frame length has to be increased or weights have to be smaller. In practice, in order to make 25 spreading codes with weight 5, the sequence code length needs to be 501. Therefore, OOC needs 20 times larger frame length compared to the modified prime sequence code, and the bit rate decreases when OOC is used. When the weights are decreased in OOC, the correlation property is degraded. Therefore, the modified prime sequence code whose weights are the same as those of OOC is more effective in a synchronous optical CDMA, because the modified prime sequence code can make more spreading codes with a shorter frame length.

Table 7. Prime Sequences S_x^P and Prime Sequence Codes C_x^P for GF(5)

	i		
x	01234	Sequence	Code
0	00000	S_0^5	$C_0^5 = 10000\ 10000\ 10000\ 10000\ 10000$
1	01234	S_1^5	$C_1^5 = 10000\ 01000\ 00100\ 00010\ 00001$
2	02413	S_2^5	$C_2^5 = 10000\ 00100\ 00001\ 01000\ 00010$
3	03142	S_3^5	$C_3^5 = 10000\ 00010\ 01000\ 00001\ 00100$
4	04321	S_4^5	$C_4^5 = 10000\ 00001\ 00010\ 00100\ 01000$

3. SYSTEM MODEL OF OPTICAL SYNCHRONOUS CDMA

3.1. Optical Time-Encoded CDMA Systems

In optical CDMA, intensity modulation/direct detection is used mainly in conjunction with on/off keying (OOK) and pulse position modulation (PPM) signaling formats. A pulse laser source is intensity-modulated by electrical (0,1) data bits. Data are modulated by emitting optical positive pulses at the first chips of the slots in OOK or PPM signaling. In optical time-encoded CDMA, a slot is divided into chips and the number of chips in a slot equals the spreading code length. In optical synchronous CDMA systems, synchronization between transmitter and receiver is required, and the synchronization is achieved by adding to the correlated signal the receiver clock signal delayed by a proper amount. The delay must be such that the peak of the autocorrelation function coincides with the optical clock pulse. Code synchronization can be realized through a two-stage process: a coarse alignment referred to as *code acquisition* and a subsequent fine alignment referred to as *code tracking*.

In OOK, a "1" information bit is transmitted by emitting a optical pulse at the first chip of the slot. When no pulse is emitted in chips within a slot, this means that "0" information bit is transmitted. Thus, one bit binary information is conveyed in a slot. At the decoder, threshold detection is used in OOK. Since the threshold value depends on the intensity level of received signal pulse, multiple-user interference, and noise, proper adjustment of the threshold level is required. In an M-ary PPM, a narrow pulse is emitted at the first chip of one of M slots in a PPM frame to represent data. Since the combination of selecting one slot among M slots in a frame is $\log_2 M$, $\log_2 M$ bits can be conveyed in a frame. This results in a low channel traffic in PPM compared to OOK in terms of the number of transmitted pulses, due to the pulse position multiplicity of PPM signaling. PPM can utilize maximum-likelihood detection in which the slot with largest intensity level in a frame is selected in maximum likelihood manner, and thus, no precise threshold adjustment is required.

Each user is assigned a signature sequence with length F, which serves as its address. In time encoding, the encoder consists of the tapped optical delay lines, and the

Table 8. Left-Rotated Prime Sequences $S^P_{x,r}$ and Modified Prime Sequence Codes $C^P_{x,r}$ for GF(5)

Group x	i 01234	Sequence	Code
0	00000	$S^5_{0,0}$	$C^5_{0,0} = 10000\ 10000\ 10000\ 10000\ 10000$
	44444	$S^5_{0,1}$	$C^5_{0,1} = 00001\ 00001\ 00001\ 00001\ 00001$
	33333	$S^5_{0,2}$	$C^5_{0,2} = 00010\ 00010\ 00010\ 00010\ 00010$
	22222	$S^5_{0,3}$	$C^5_{0,3} = 00100\ 00100\ 00100\ 00100\ 00100$
	11111	$S^5_{0,4}$	$C^5_{0,4} = 01000\ 01000\ 01000\ 01000\ 01000$
1	01234	$S^5_{1,0}$	$C^5_{1,0} = 10000\ 01000\ 00100\ 00010\ 00001$
	12340	$S^5_{1,1}$	$C^5_{1,1} = 01000\ 00100\ 00010\ 00001\ 10000$
	23401	$S^5_{1,2}$	$C^5_{1,2} = 00100\ 00010\ 00001\ 10000\ 01000$
	34012	$S^5_{1,3}$	$C^5_{1,3} = 00010\ 00001\ 10000\ 01000\ 00100$
	40123	$S^5_{1,4}$	$C^5_{1,4} = 00001\ 10000\ 01000\ 00100\ 00010$
2	02413	$S^5_{2,0}$	$C^5_{2,0} = 10000\ 00100\ 00001\ 01000\ 00010$
	24130	$S^5_{2,1}$	$C^5_{2,1} = 00100\ 00001\ 01000\ 00010\ 10000$
	41302	$S^5_{2,2}$	$C^5_{2,2} = 00001\ 01000\ 00010\ 10000\ 00100$
	13024	$S^5_{2,3}$	$C^5_{2,3} = 01000\ 00010\ 10000\ 00100\ 00001$
	30241	$S^5_{2,4}$	$C^5_{2,4} = 00010\ 10000\ 00100\ 00001\ 01000$
3	03142	$S^5_{3,0}$	$C^5_{3,0} = 10000\ 00010\ 01000\ 00001\ 00100$
	31420	$S^5_{3,1}$	$C^5_{3,1} = 00010\ 01000\ 00001\ 00100\ 10000$
	14203	$S^5_{3,2}$	$C^5_{3,2} = 01000\ 00001\ 00100\ 10000\ 00010$
	42031	$S^5_{3,3}$	$C^5_{3,3} = 00001\ 00100\ 10000\ 00010\ 01000$
	20314	$S^5_{3,4}$	$C^5_{3,4} = 00100\ 10000\ 00010\ 01000\ 00001$
4	04321	$S^5_{4,0}$	$C^5_{4,0} = 10000\ 00001\ 00010\ 00100\ 01000$
	43210	$S^5_{4,1}$	$C^5_{4,1} = 00001\ 00010\ 00100\ 01000\ 10000$
	32104	$S^5_{4,2}$	$C^5_{4,2} = 00010\ 00100\ 01000\ 10000\ 00001$
	21043	$S^5_{4,3}$	$C^5_{4,3} = 00100\ 01000\ 10000\ 00001\ 00010$
	10432	$S^5_{4,4}$	$C^5_{4,4} = 01000\ 10000\ 00001\ 00010\ 00100$

Table 9. Optical Orthogonal Codes: $F = 32, K = 4$

Number of Chips between the Subsequent 1s	OOC
9,3,15,5	10000000010010000000000000010000
4,7,19,2	10001000000100000000000000000010

laser pulse emitted in the first chip in a slot is time-spread in F chips within a slot corresponding to "1"s of the spreading codes. That is, ON information is transmitted as a sequence of F chipped pulses and OFF information is sent as an all-zero sequence. Note that the repetition rate of the light source limits the transmission rate, because the light source has difficulty to generate long successive optical pulse transmissions in time. Thus, time encoding is useful because there is no successive optical pulse transmission in adjacent chips. Optical pulse sequences transmitted from users are combined in the star coupler and then transmitted over the fiber to the desired destination. At the receiver, an optical incoherent passive filter-matched detection is done by an optical delay-line decoder matched to the encoder at the transmitter. It produces a peak in the correlation output for the intended bits. Data bits are discriminated in the chip duration using a photodiode followed by a threshold process.

The "near/far" problem is an essential issue in most CDMA systems, especially in wireless applications. Fortunately, the transceivers in an optical fiber networks are fixed, and to a certain extent, the optical path loss between the transmitter and the receiver can be predicted. Hence, a gain-clamped preamplifier can be used to compensate for optical power loss, so that all the transmitted optical signals originating from different locations can be received at similar optical power levels.

In the following subsection, the IM/DD systems with OOK and PPM signaling are described as typical optical synchronous CDMA systems. Also, as an alternative optical CDMA system, an optical frequency-encoded CDMA system utilizing bipolar codes is briefly introduced.

3.1.1. Optical Synchronous OOK CDMA. Figure 1, shows the transmitter block diagram of a direct-detection

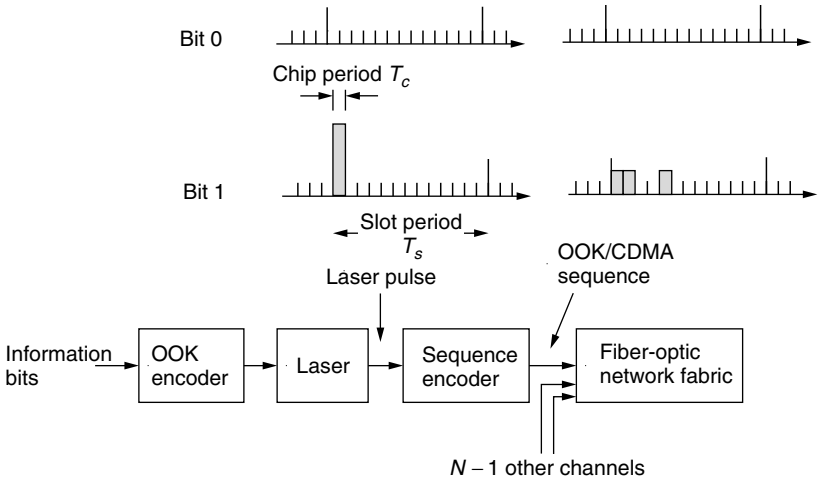


Figure 1. The transmitter block diagram of a direct-detection optical OOK CDMA system.

optical CDMA system. The output of the information source is fed into the OOK encoder where the information bitstream is directly converted into the OOK pulse sequences. When a logical "1" is to be conveyed, the laser is pulsed on at the first chip of the slot; when a logical "0" is to be conveyed, the laser is not pulsed on. Then the output laser pulse is converted into the assigned optical code sequence, that is, the signature sequence by a tapped optical delay line [6] shown in Fig. 2 that converts the initial laser pulse into a specific train of output pulses. When a logical "0" is to be conveyed, an all-zero sequence is transmitted. Light pulse sequences from all sources are combined in the fiber-optic network fabric and then transmitted over the fiber to the desired destination.

Figure 3 shows the receiver block diagram of the direct-detection optical OOK CDMA system. At the receiver, the matched optical correlator is used to recognize the arrival of the desired sequence. Figure 4 shows the optical correlator comprising a set of optical delay lines inversely matched to the pulse spacings. When the desired optical sequence passes through the correlator, the output light intensity traces out the correlation function of the sequence. At the last chip position, the sum of received optical intensity located in the same positions as the positions of "1" of the signature codes used for the desired channel is obtained. The correlator output is converted into an electrical signal by the photodetector and is then passed to the OOK decoder. The OOK decoder compares the correlator output voltage over the last chip position with the threshold level, then decides that "1" is sent if the output voltage is larger than the threshold level, and

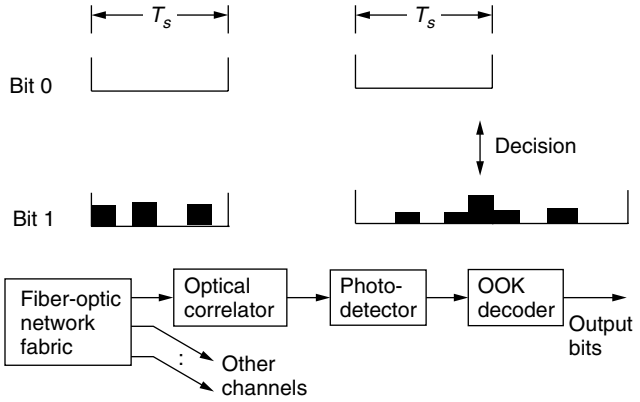


Figure 3. The receiver block diagram of a direct-detection optical OOK CDMA system.

"0" is sent otherwise. In this way each user can recover his own logical sequence.

3.1.2. Optical Synchronous PPM CDMA. Figure 5 shows the block diagram of an optical M-ary PPM CDMA transmitter and a signaling format. At the transmitter, a data bitstream is first blocked into words of length $\log_2 M$ bits in the PPM encoder and then each word is encoded into a M-ary PPM signaling format, that is, one of M slot positions. Every slot consists of F chips with the same chip period T_c, where F is the spreading factor of the CDMA signals. The laser is pulsed on at the first chip of the proper slot representing the word and the other slots

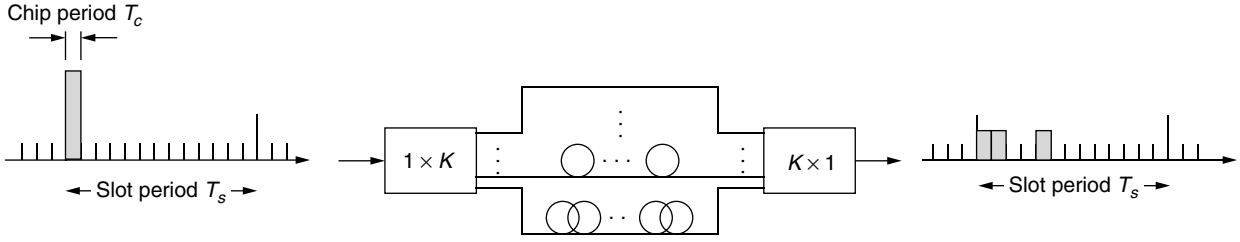


Figure 2. A sequence encoder consisting of tapped optical delay lines.

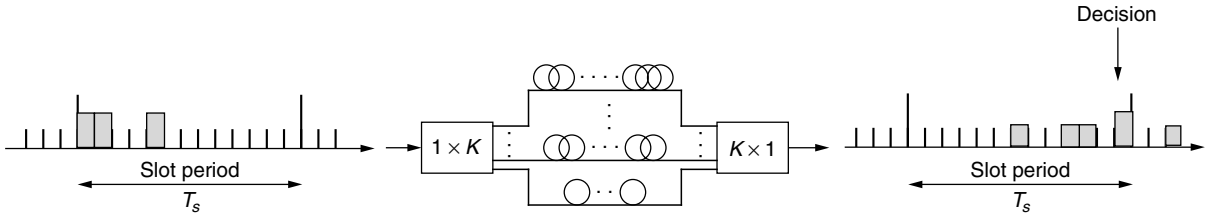


Figure 4. An optical correlator comprising tapped optical delay lines.

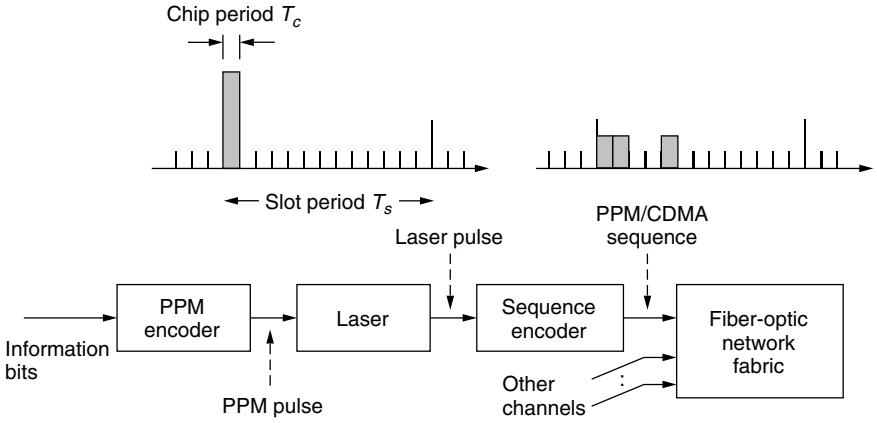


Figure 5. The transmitter block diagram of a direct-detection optical PPM CDMA system.

have no pulse. Then the output laser pulse is converted into the assigned optical code sequence, that is, the signature sequence by a set of tapped optical delay lines [6] that converts the initial laser pulse into a specific train of output pulses. The transmitted PPM CDMA signal in the pulsed slot have K pulses according to the assigned sequence code. Light pulse sequences from all the sources are combined in the fiberoptic network fabric and then transmitted over fiber to the desired destination.

The block diagram of an optical PPM CDMA receiver is shown in Fig. 6. At the receiver, a matched optical correlator is used to recognize the arrival of the desired sequence. The optical correlator is a set of optical delay lines inversely matched to the pulse spacings of the sequences. When the desired optical sequence passes through the correlator, the output light intensity traces out the correlation function of the sequence. In the photodetector the correlator output is converted into the electrical signal that is passed to the PPM decoder. The PPM decoder compares the output voltage over the last chips of all the slots and decides the slot having the highest voltage as the pulsed slot. Finally, the PPM decoder declares the corresponding word as the transmitted word.

There are two main advantages of synchronous M-ary PPM CDMA over OOK CDMA. The first advantage is that, under a bit error rate constraint, the maximum number

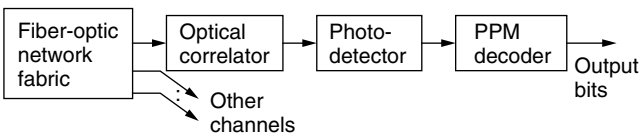


Figure 6. The receiver block diagram of a direct-detection optical PPM CDMA system.

of simultaneous users can be increased by increasing M and keeping the average power fixed. On the other hand, in the case of OOK CDMA, this number cannot be increased without increasing the average power. The second advantage is that any number of users can be accommodated by increasing M is the case of PPM CDMA. In the case of OOK CDMA, however, we may not be able to accommodate all the subscribers, even if the average power is increased. The reason is that, when the number of users is N, the average number of interfering optical pulses reduces to $(N-1)/M$ for PPM CDMA, whereas this number is equal to $(N-1)/2$ for OOK CDMA. Of course, these advantages of PPM CDMA over OOK CDMA are gained at the expense of increasing the system complexity [7].

3.2. Optical Frequency-Encoded CDMA System

In the previous two sections, we explained the time-encoded optical CDMA systems with OOK and PPM signaling and unipolar codes. Here, we introduce a frequency-encoded CDMA (FECDMA) system as an alternative optical CDMA using bipolar codes. Note that FE-CDMA can be used as both synchronous and asynchronous CDMA. The advantages of optical FECDMA are random-access, self-routing capability by the code itself, and the independence of the bit rate and processing gain, since coding and decoding are done by shifting phase without expanding the frequency band. In FECDMA, bipolar sequence codes such as pseudonoise (PN) sequences are assigned to each user, and the Fourier transform of the transmitted pulse for a given user is determined by encoding the phase of the desired transmitted spectrum by the user's PN sequence. The system model of the FE-CDMA is shown in Fig. 7. The FECDMA scheme is based on encoding and decoding of

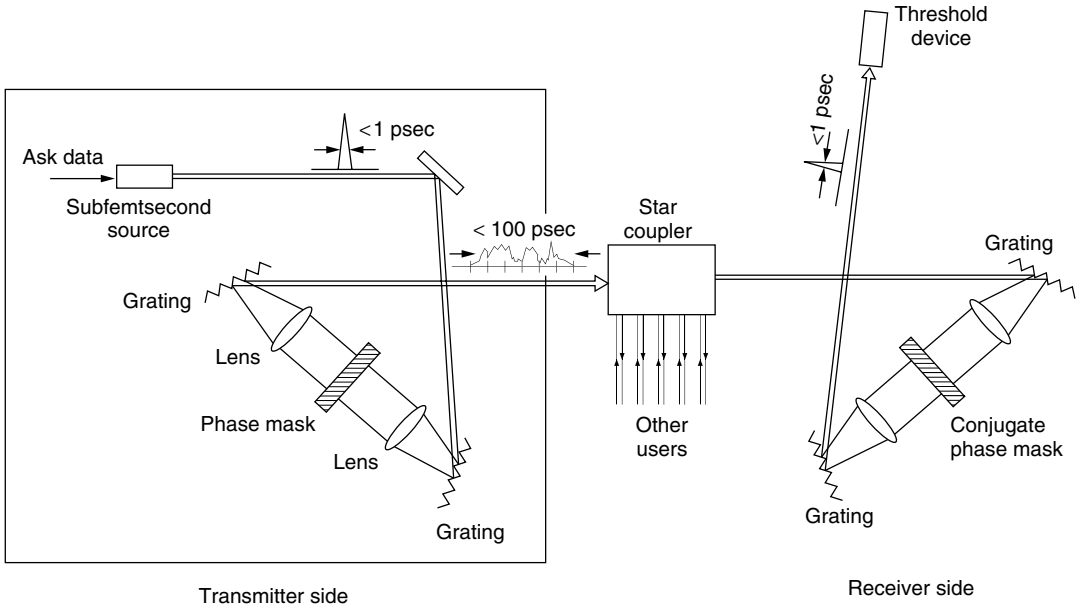


Figure 7. System model of FE-CDMA.

an ultrashort light pulse with duration τ_c and peak power P_0, and the operation is accomplished by using femtosecond pulseshapers. The pulseshapers offer high-resolution pulseshaping, programmability, and the flexibility to apply arbitrary phase codes of different code lengths. The operation of encoding and decoding is performed by using two fixed conjugate phase masks successively in the same pulse shaper. The liquid crystal modulator (LCM) is used to set the spectral phases to maximum-sequence phase. The LCM has a fully programmable linear array and individual pixels can be controlled by applying drive levels resulting in phase shifts (0 or π). By a phase mask, the dispersed bandwidth of a pulse is partitioned into N_c frequency chips, where each chip has the bandwidth W/N_c. Each chip is assigned a phase shift (i.e., 0 or π) depending on the user's PN sequence. The spectrum of the resulting pulse is reassembled by an inverse Fourier transform, and the encoded pulse is spread out within a time slot in synchronous CDMA as a low-intensity pseudonoise burst with average power P_0/N_c and duration T, as shown in Fig. 8. The transmitted data for a particular user can be recovered by sampling the output of a filter matched to the user's pulse. The phase mask and grating are implemented to perform the Fourier transform and matched filtering. At the receiver side, the spectral decoder consists of Fourier transformation of the time-windowed received signal followed by correlation with the PN sequence matched to the transmitter code. The spectral-phase code of the decoder is the complex conjugate of the encoder's spectral-phase code. The correctly decoded signal becomes a replica of the original short pulse with duration τ_c and peak power P_0, whereas the MAI signal remains a low intensity pseudonoise burst. The disadvantages of FECDMA are that the effects of MAI are large as the number of simultaneous user increases and a high-resolution phase mask as well as a narrow pulse are required to improve discrimination ability.

4. CHANNEL INTERFERENCE CANCELER FOR OPTICAL SYNCHRONOUS CDMA

Optical CDMA has several advantages over optical TDMA, including complete utilization of the entire time-frequency domain by each subscriber, flexibility in network design (because the quality depends on the number of active users), and security against interception. Synchronous CDMA has an additional advantage over asynchronous CDMA, where the number of available sequence codes (and in turn the number of subscribers) is much higher in the former under a given throughput constraint. The latter does not require, however, any time management as in the former. It follows that synchronous CDMA is suitable for very high speed networks with real time requirements (e.g., voice and digitized video). In contrast, asynchronous CDMA is suitable for bursty traffic with no stringent time requirements (e.g., data transmission). On the other hand, optical CDMA has a disadvantage over TDMA that is due to the multiple-user interference in the former. This leads in turn to a serious degradation in the bit error probability as the number of simultaneous users increases. This degradation cannot be overcome even for arbitrary high optical power. In fact, there will be an asymptotic error floor which limits the number of users that can communicate simultaneously and reliably. Several interference cancellation techniques have been proposed aiming at lowering these asymptotic error floors. These interference cancellation techniques are classified into two groups. One is based on the use of properties of modified prime sequence codes, and the other is based on the use of optical hard-limiters. The cancellation techniques using the properties of modified prime sequence codes have been proposed for both OOK CDMA and PPM CDMA systems [8–19]. These techniques estimate the amount of interference from a knowledge of some other users' sequence codes by using the correlation properties of modified prime sequence codes.

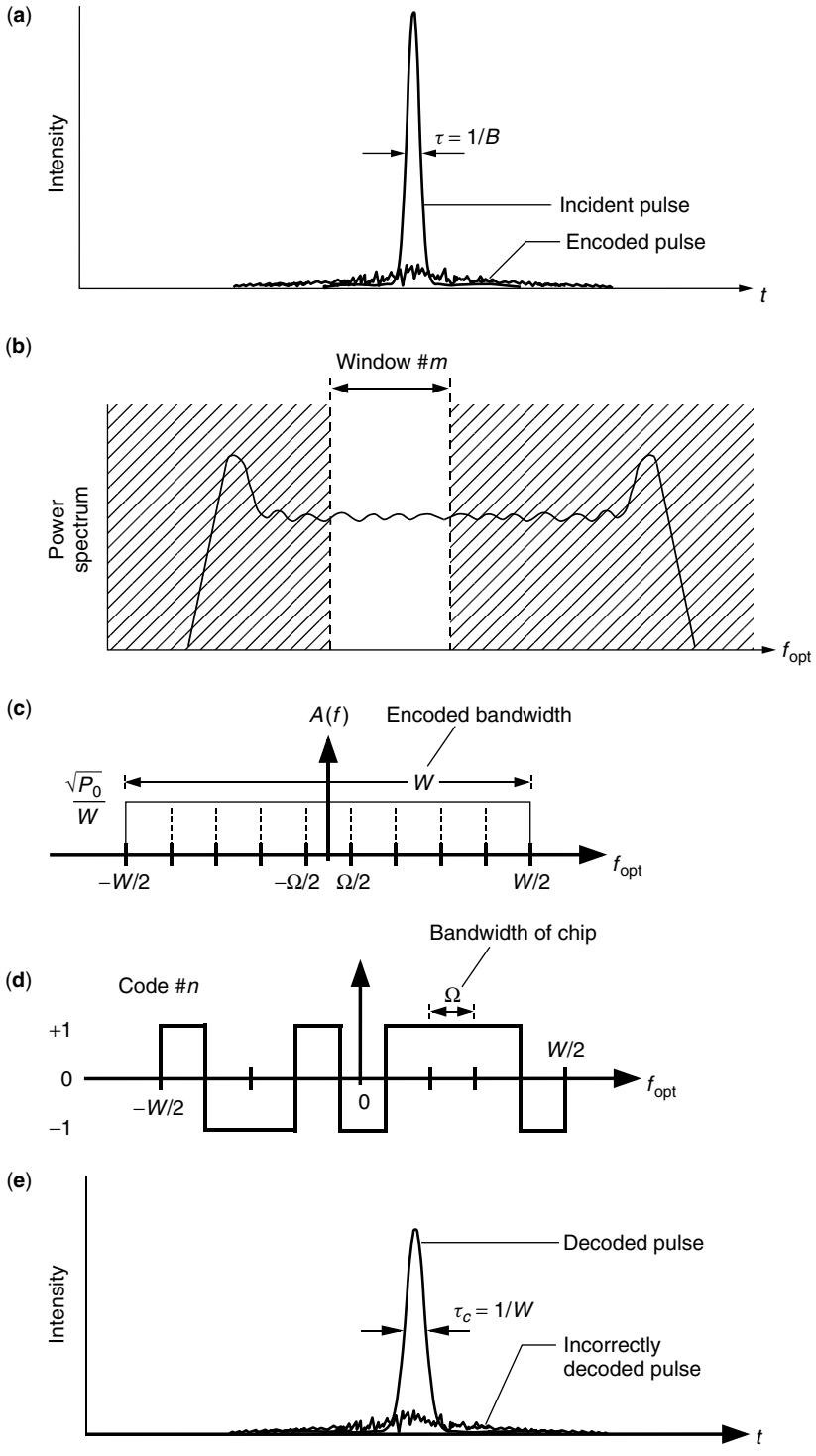


Figure 8. Frequency encoding.

As for the cancellation techniques based on the use of optical hard-limiters, Salehi and Brackett [20] used an optical hard-limiter that is placed before the optical correlator at the receiver side. This optical hard-limiter is shown to be able to remove some of the interference patterns. Ohtsuki et al. [21] proposed a synchronous optical CDMA system with double optical hard-limiters placed before and after the optical correlator. This system introduces an improvement in the performance over the system with a single optical hard-limiter as long as the number of users is not very large. In the case of asynchronous optical CDMA, Ohtsuki [22,23] showed that this improvement continues for all possible number of users. In Ohtsuki [24] was also able to reduce the error floor even lower than that of the system with double hard-limiters. Lin and Wu [25] suggested a synchronous optical CDMA system with an adaptive optical hard-limiter (or equivalently, a tunable optical attenuator) placed after

the correlator receiver. They were able to show that the performance can be improved as compared to the system with double hard-limiters.

We briefly review some cancellation techniques in the following sections.

4.1. Channel Interference Canceler Using Properties of Modified Prime Sequence Codes

We describe the channel interference canceller using a time-division reference signal [11,12] as an example of the channel interference cancellation technique using the properties of modified prime sequence codes.

Each user is assumed to be assigned a unique prime sequence code of length P^2 and weight P. In the system each user is allowed to access the network $P-1$ times out of P times, and $P-1$ users out of P users can access the network in each group simultaneously; that is, one user in each group is not allowed to access the network at each time; unallowable user's channel in each group at the time is used as a reference signal for other users in the same group at the time to cancel the effects of channel interference, because every code in the same group suffers the same amount of channel interference from other groups and every code in the same group does not interfere with each other.

Figure 9 shows an example of the access timing pattern for the first group in the system with the canceler where each user is not allowed to access the network at the marked time in the table. For instance, the first user is not allowed to access the network at time t_1, and the output of the first channel is used as a reference signal for other users in the same group at time t_1 to cancel the effects of channel interference. Notice that all the users in the same group suffer the same amount of channel interference from other groups. The bit rate of each user is thus

$$R_b = \frac{P-1}{PT} \tag{18}$$

At the same bit rate, the slot width of the proposed system is $(P-1)/P$ times as long as that of the system without canceler; thus the proposed system needs a slightly broader bandwidth.

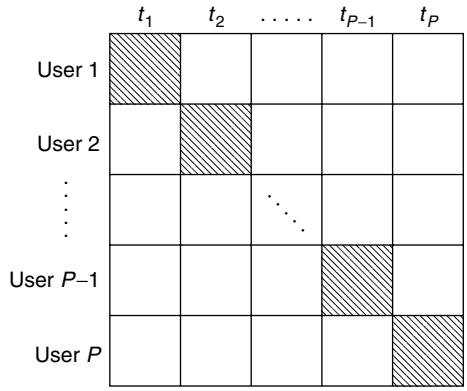


Figure 9. An example of the access timing pattern for the first group in the system with the canceler where each user is not allowed to access the network at the marked time in the table.

Figure 10 shows the receiver block diagram for the first user of the system. At the receiver, the matched optical correlators are used to recognize the arrival of the desired sequence. At the last chip position, the sum of the received optical intensity located in the same positions as the positions of "1" of the modified prime sequence code used for the desired channel is obtained. The correlator output is converted into an electrical signal by the photodetector. According to the table of the access timing pattern, the switch is connected to the reference channel that is not used at the time; the output of the reference signal is subtracted from the desired signal to cancel the effect of the channel interference. The signal after subtraction is passed to the OOK decoder. The OOK decoder compares the correlator output voltage over the last chip position with the threshold level and decides the data.

Figure 11 shows the bit error probability of OOK CDMA versus the received laser power P_W for some values of P where $N = P^2$, that is, the full-load case. As shown in Fig. 11, the received signal of each user is split into P

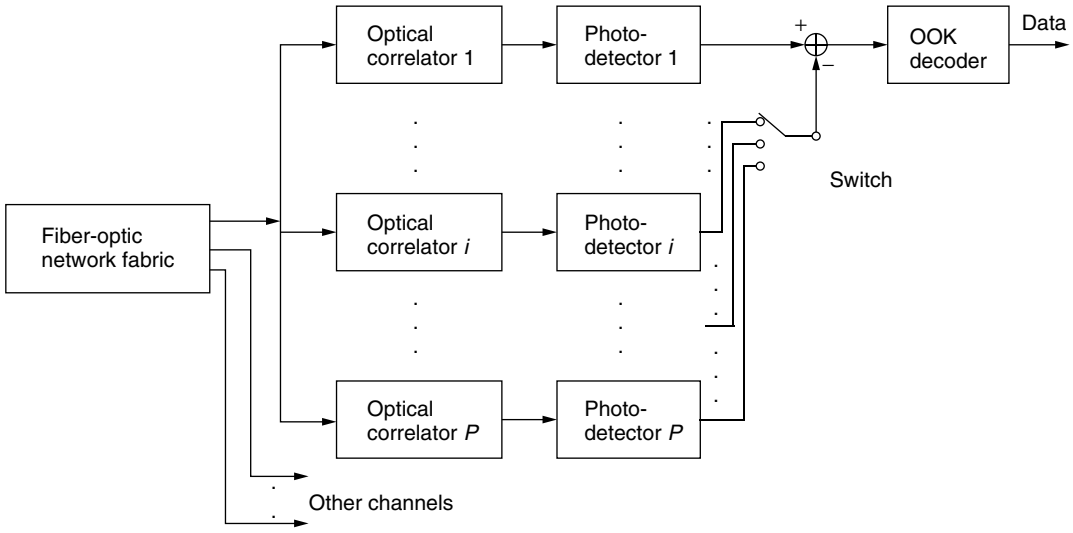


Figure 10. The receiver block diagram for the first user of the system with the canceler.

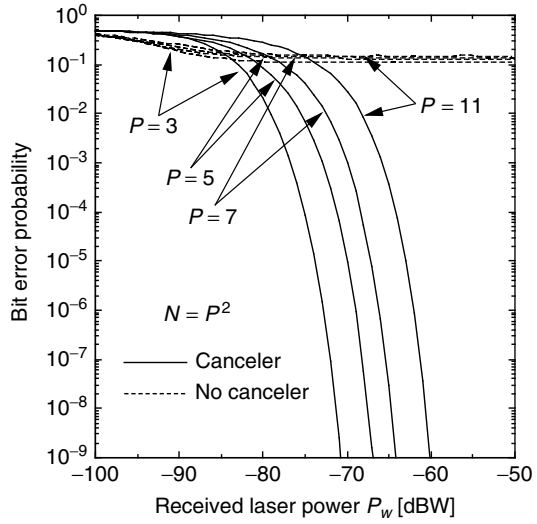


Figure 11. Bit error probability of OOK CDMA versus received laser power P_W for some values of P: $N = P^2$.

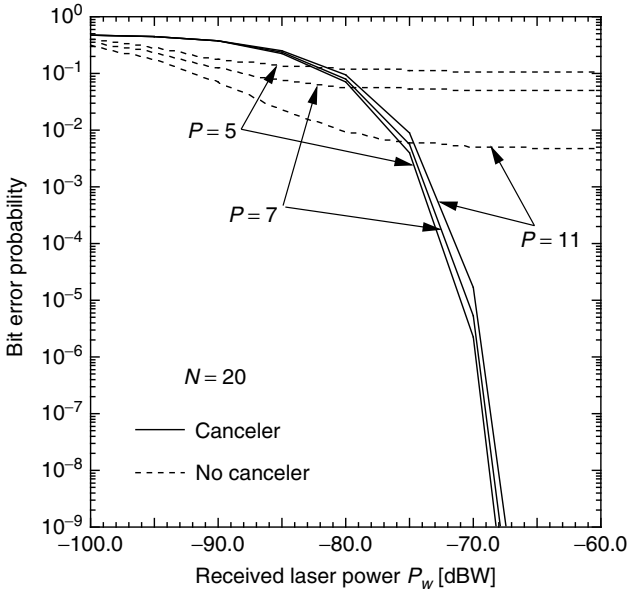


Figure 12. Bit error probability of OOK CDMA versus received laser power P_W for some values of P: $N = 20$.

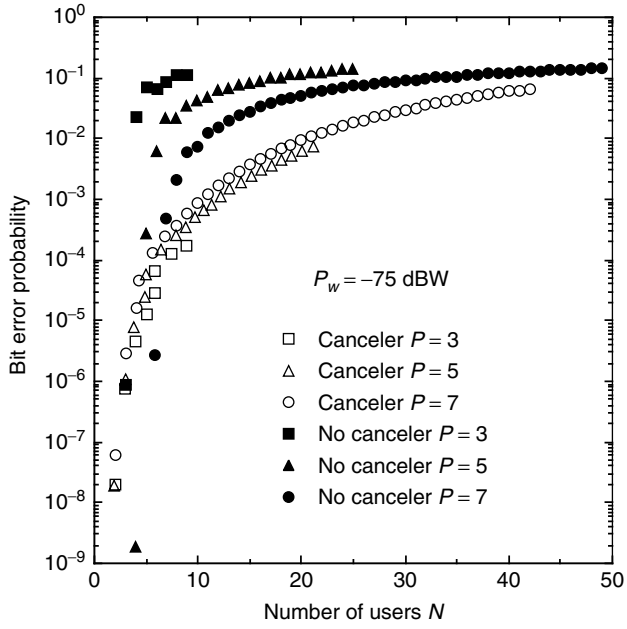


Figure 13. Bit error probability of OOK CDMA versus the number of users N for some values of P: $P_W = -75$ dBW.

branches at the receiver in the system with the canceler, and thus the unit received laser optical power in the delay-line of the optical correlator of the system with the canceler P_W is $1/P$ times as large as that of the system without the canceler. It can be seen that the system with the canceler has better performance than the conventional system without the canceler when P_W is not appreciably small; as P_W increases, the bit error probability of the system with the canceler is improved, while the error floor exists for the conventional systems without the canceler, because the effect of the channel interference is so large in the case of full load. Since the system with the canceler can reduce the effects of the channel interference, the error floor does not exist for the system with the canceler even in the case of full load. As P increases, the effect of the channel interference also increases in the case of full load, and thus the system with the canceler with larger P needs more power to have better performance.

Figure 12 shows the bit error probability of OOK CDMA versus the received laser power P_W for some values of P where $N = 20$. It can be seen that the system with the canceler has better performance than the conventional system without the canceller when P=5 and 7, and P_W is not appreciably small. Although the system with the canceler can reduce the effects of channel interference, it needs somewhat large power to have better performance than the system without the canceler, because the received signal of each user is split into P branches. It can be also seen that the system with the canceler has almost the same performance for any P, because when the number of simultaneous users is the same, the ratio of the signal power to the channel interference power is almost the same for any P. In addition, the system with the canceler can cancel the effects of the channel interference. Thus, the system with the canceler has almost the same performance for any P.

Figure 13 shows the bit error probability versus the number of users N for some values of P where $P_W = -75$ dBW. It can be seen that the system with the canceler has better performance when N is not appreciably small: when $N > 6$ for P=7, the system with the canceler has better performance. The effects of the channel interference are small when N is small. Thus the system with the canceler does not have better performance when N is small. It can be also seen that the system with the canceler has almost the same performance for any P, while the bit error probability of the conventional system is improved as P increases. In the conventional system without the canceler, the effect of channel interference is almost the same for any P at the same N, while the signal intensity becomes

larger as P increases even at the same N. In addition, the system with the canceler can reduce the effect of channel interference. Thus, the bit error probability of the conventional system is improved as P increases, while the system with the canceler has almost the same performance for any P.

4.2. Channel Interference Canceler Using Optical Hard-Limiters

Figure 14 shows the receiver block diagram of the direct-detection optical OOK CDMA system with a single optical hard-limiter [20]. In optical CDMA systems, the channel interference is the prime noise factor; it degrades the performance seriously and produces an asymptotic floor to the error probability. An optical hard-limiter is used to reduce the channel interference and to improve the system performance. An optical hard-limiter is defined as

$$g(x) = \begin{cases} v_f, & x \geq \text{Th} \\ 0, & 0 \leq x < \text{Th} \end{cases} \tag{19}$$

where v_f is the fixed value dependent on the signal intensity and Th is the threshold level. If an optical light intensity x is larger than or equal to the threshold level Th, the hard-limiter would clip the intensity back to v_f, and if the optical light intensity x is smaller than Th, the response of the optical hard-limiter would be zero. This optical hard-limiter would improve the system performance in the ideal link, because it would reduce the effect of the channel interference generated by some combinations of interference patterns. As shown in Fig. 14, the optical hard-limiter is placed before the optical correlator in the conventional optical CDMA systems. At the receiver, the matched optical correlator is used to recognize the arrival of the desired sequence. The optical correlator is a set of optical delay lines inversely matched to the pulse spacings. When the desired optical sequence passes through the correlator, the output light intensity traces out the correlation function of the sequence. At the last chip position, the sum of received optical intensity located in the same positions as the positions of "1" of the signature sequence code used for the desired channel is obtained.

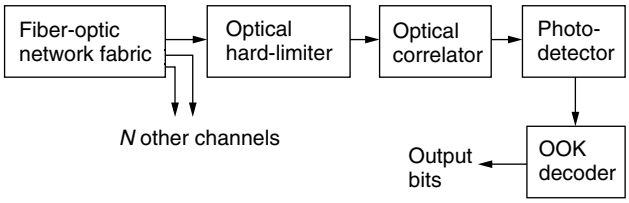


Figure 14. The receiver block diagram of a direct-detection optical OOK CDMA system with the single optical hard-limiter.

Figure 15 shows an example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends "1" where T_c is the chip duration: the second and the third marks of the desired user are hit by one undesired interferer's mark, respectively. The interference is removed by the first optical hard-limiter. Note that, however, the interference would contribute to the optical light intensity of the desired user when the desired user sends "1" if the optical hard-limiter is not used.

Figure 16 shows an example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends "0"; the first mark position of the desired user is hit by two undesired interferers' marks, and the second mark position of the desired user is hit by one undesired interferer's mark. As shown in this figure, there are some interference patterns that are not completely removed with the first optical hard-limiter when the desired user sends "0."

To improve the system performance by excluding some combinations of interference that cause incorrect bit decisions for "0" bit transmission, optical synchronous CDMA systems with double optical hard-limiters have been proposed. Figure 17 shows the receiver block diagram of the system with double optical hard-limiters. In the system, optical hard-limiters are placed before and after the optical correlator. The optical hard-limiters placed before and after the optical correlator are referred to as the *first* and the *second optical hard-limiters*, respectively.

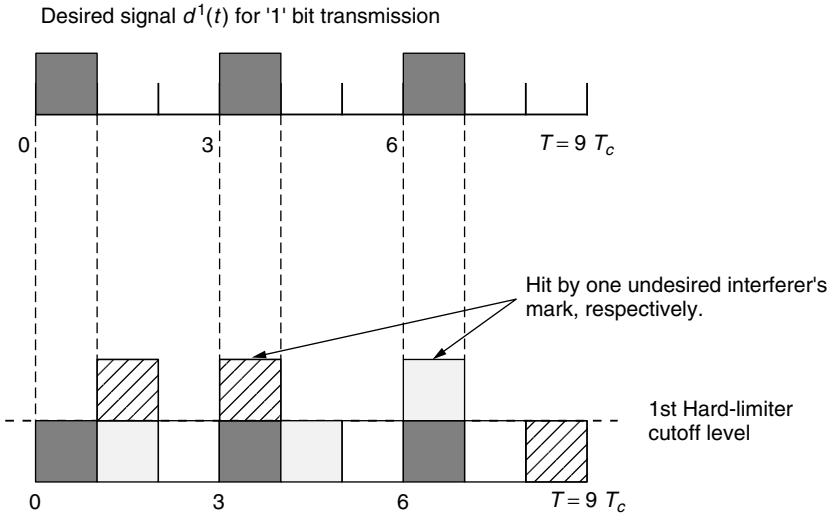


Figure 15. An example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends a "1"; the second mark of the desired user is hit by one undesired interferer's mark, and the third mark of the desired user is hit by one undesired interferer's mark.

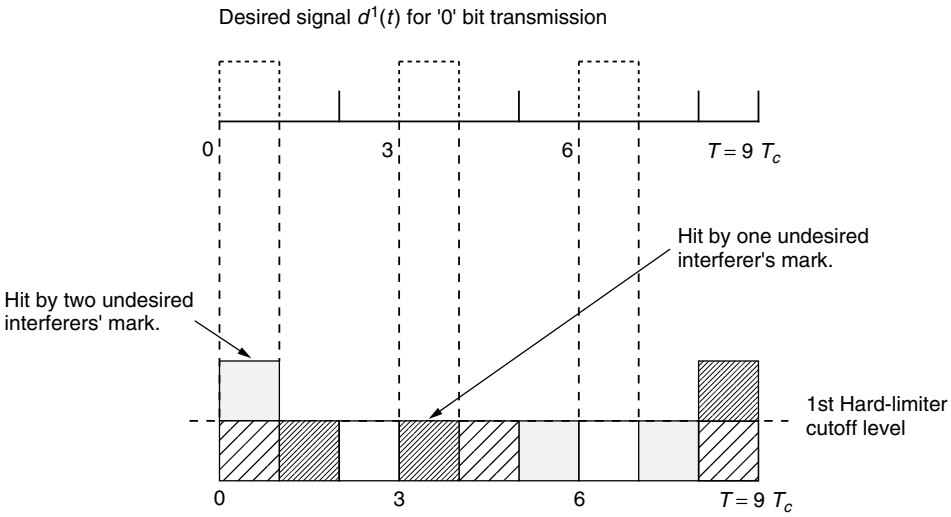


Figure 16. An example of an interference pattern on the desired signal over a sequence period $T = 9T_c$ when the desired signal sends "0"; the first mark position of the desired user is hit by two undesired interferes' mark, and the second mark position of the desired user is hit by one undesired interferer's mark.

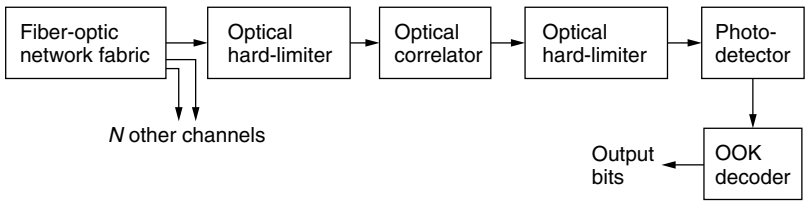


Figure 17. The receiver block diagram of a direct-detection optical CDMA system with double optical hard-limiters.

The first and the second optical hard-limiters are defined as the same as in the optical CDMA systems with the single optical hard-limiter given by Eq. (19). The first optical hard-limiter would clip the intensity back to v_f, and thus exclude or reduce the channel interference by other undesired interferers' marks. The second optical hard-limiter is used to exclude some interference patterns that are not completely removed with the first optical hard-limiter when the desired user sends "0" as shown in Fig. 16.

The second optical hard-limiter would clip the output intensity from the optical correlator back to zero if the optical light intensity x is smaller than Th. Therefore, the system using double optical hard-limiters would improve the system performance, because it would exclude the effect of the channel interference generated by some combinations of interference patterns as shown in Fig. 16.

Figure 18 shows the bit error probability versus the average received photocount in the last chip K_s for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with the double optical hard-limiters for some values of P. It can be seen that using the single optical hard-limiter placed before the optical correlator slightly degrades the performance of an optical synchronous CDMA system using modified prime sequence codes. Using the single optical hard-limiter excludes some combinations of interference patterns; however, it also excludes some combinations of interference patterns contributing to the optical intensity of the desired user. Thus, using the single optical hard-limiter slightly degrades the performance under the Poisson shot noise model for the receiver photodetector. It can be also seen that the

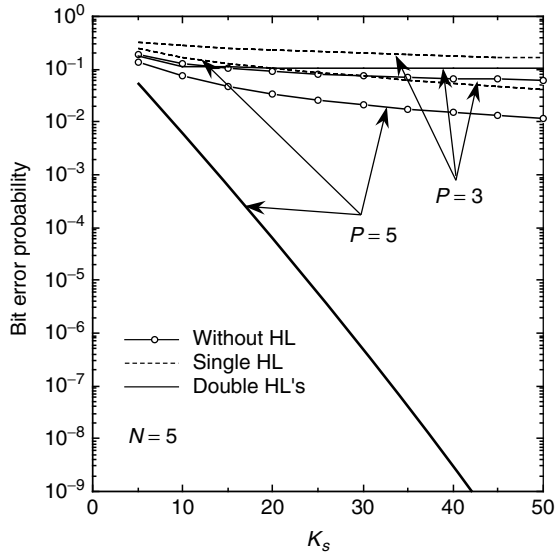


Figure 18. Bit error probability versus K_s for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with double optical hard-limiters for some values of P where $N = 5$.

optical CDMA system with double optical hard-limiters has better performance than do the other two systems when $P = 5$; when $P = 3$ using the system with double optical hard-limiters has a performance slightly worse than that of the system without the optical hard-limiter and slightly better that of than, the system with the single optical hard-limiter. The double optical hard-limiters can exclude combinations of interference patterns that the

single optical hard-limiter cannot exclude; all the mark positions of the desired user are not hit by undesired interference marks when the desired user sends "0." When $P = 3$ and $N = 5$, there are still some combinations of interference patterns that the double optical hard-limiters cannot exclude, while there is no combination of interference patterns that the double optical hard-limiters cannot exclude when $P = 5$ and $N = 5$. Thus, the error floor exists for the optical CDMA systems with double optical hard-limiters when $P = 3$ and $N = 5$ and does not exist when $P = 5$ and $N = 5$. Moreover, it can be seen that all the systems with $P = 5$ have better performance than do those with $P = 3$, respectively. This is because the probability that the undesired users' marks hit the desired user's mark positions is small when $P = 5$. Thus all the systems with $P = 5$ have better performance.

Figure 19 shows the bit error probability versus the number of users N for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with double optical hard-limiters with $P = 7$ and $K_s = 30$. It can be seen in the figures that the optical CDMA system with double optical hard-limiters has the constant low bit error probability when N is smaller than or equal to P: $N \leq 7$. This is because the optical hard-limiters can exclude all the combinations of interference patterns when $N \leq P$, that is, the number of interferers is smaller than P. It can be also seen that the performance of the optical CDMA system with double optical hard-limiters becomes worse than that of the optical CDMA systems without the optical hard-limiter when N is large. When N is large, the probability that all the mark positions of the desired user are hit by the undesired interference marks becomes high, and thus the number of combinations of interference patterns that the double optical hard-limiters cannot exclude becomes large and the performance is degraded. Therefore, using the double optical hard-limiters is effective in improving the performance of optical synchronous CDMA systems when the number of simultaneous users is not so large.

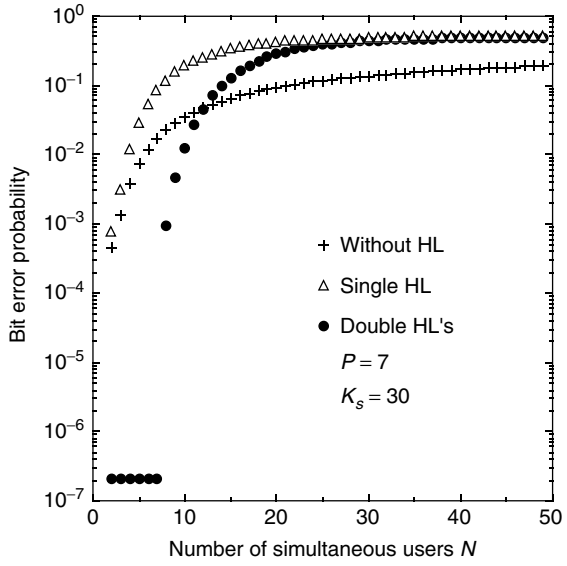


Figure 19. Bit error probability versus the number of simultaneous users N for the optical OOK CDMA systems without the optical hard-limiter, with the single optical hard-limiter, and with double optical hard-limiters: $P = 7$ and $K_s = 30$.

BIOGRAPHY

Tomoaki Ohtsuki received his B.S., M.S., and Ph.D. degrees in electrical engineering from the Keio University, Yokohama, Japan in 1990, 1992, and 1994, respectively. From 1994 to 1995 he was a postdoctoral Fellow and a visiting researcher of electrical engineering at the Keio University. From 1993 to 1995 he was a special researcher of fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. He joined the Tokyo University of Science in 1995 as an assistant professor. Since 2000 he has been a lecturer tenured of the Tokyo University of Science. He has been working on optical communication systems, wireless communication systems, and information theory. He was the recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, the Erricson Young Scientist Award 2000, and the 2002 Funai Information and Science Young Scientist Award.

Iwao Sasase received his B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1979, 1981, and 1984, respectively. He is currently a professor of information and computer science at Keio University. His research interests include mobile, satellite and optical communications, and information networks. He has authored more than 170 journal papers and 260 international conference papers. He was the recipient of the 1984 IEEE Communications Society Student Paper Award, 1986 Inoue Memorial Young Engineer Award, 1988 Hiroshi Ando Memorial Young Engineer Award and 1988 Shinohara Memorial Young Engineer Award, in 2002. He is the IEEE Communications Society Satellite and Space Communications Technical Committee Chair and Asia Pacific Region vice director.

BIBLIOGRAPHY

1. F. R. K. Chung, J. A. Salehi, and V. K. Wei, Optical orthogonal codes: design, analysis, and applications, *IEEE Trans. Inform. Theory* **IT-35**: 595–604 (May 1989).
2. W. C. Kwong, P. A. Perrier, and P. R. Prucnal, Performance comparison of asynchronous and synchronous code-division multiple-access, *IEEE Trans. Commun.* **COM-39**: 1625–1634 (Nov. 1991).
3. P. R. Prucnal, M. A. Santro, and T. R. Fan, Spread spectrum fiber-optic local area network using optical processing, *IEEE J. Lightwave Technol.* **LT-4**: 547–554 (May 1986).
4. A. S. Holmes and R. R. A. Syms, All-optical CDMA using "quasi-prime" codes, *IEEE J. Lightwave Technol.* **LT-10**: 279–286 (Feb. 1992).
5. W. C. kwong, G. C. Yang, and J. G. Zhang, 2^n prime-sequence codes and coding architecture for optical code-division multiple-access, *IEEE Trans. Commun.* **COM-44**: 1152–1162 (Sept. 1996).

6. K. P. Jackson et al., Optical fiber delay-line signal processing, *IEEE Trans. Microwave Theory Tech.* **MTT-33**: 193–210 (March 1985).

7. H. M. Shalaby, Performance analysis of optical synchronous CDMA communication systems with PPM signaling, *IEEE Trans. Commun.* **43**(2–4): 624–634 (Feb.–April 1995).

8. H. M. H. Shalaby and E. A. Sourour, Co-channel interference cancellation in optical synchronous CDMA communication systems, *Conf. Rec. ISSSTA'94*, Oulu, Finland, July 1994, pp. 579–583.

9. Y. Gamachi, T. Ohtsuki, H. Uehara, and I. Sasase, Optical synchronous PPM/CDMA systems using co-channel interference cancellation, *Proc. IEEE Global Telecommunications Conf.* (*GLOBECOM'95*), Singapore, Nov. 1995, pp. 2161–2165.

10. Y. Gamachi, T. Ohtsuki, H. Uehara, and I. Sasase, Performance analysis of optical synchronous PPM/CDMA systems with interference canceller under number-state light field, *Trans. IEICE*, **E79-B**(7): 915–922 (July 1996).

11. T. Ohtsuki, Channel interference cancellation using time division reference signal for direct-detection optical synchronous CDMA systems, *Proc. IEEE Int. Conf. Communications* (*ICC'96*), Dallas, TX, June 1996, pp. 187–191.

12. T. Ohtsuki, Direct-detection optical synchronous CDMA systems with channel interference canceller using time division reference signal, *Trans. IEICE* **E79-A**(12): 1948–1956 (Dec. 1996).

13. H. M. H. Shalaby, M. A. Mangoud, and S. E. El-Khamy, A new interference cancellation technique for synchronous CDMA communication systems using modified prime codes, *Conf. Rec. 2nd IEEE Symp. Computers and Communications*, 1997, pp. 556–560.

14. T. Ohtsuki, M. Takeoka, and E. Iwahashi, Performance analysis of direct-detection optical synchronous CDMA systems with co-channel interference canceller, *Trans. IEICE* **E80-A**(12): 2260–2263 (Nov. 1997) (letter).

15. T. Ohtsuki, Optical CDMA canceller systems with tunable prime code decoder, *Proc. Int. Symp. Information Theory and Its Applications* (*ISITA'96*), Victoria, Canada, Sept. 1996, pp. 766–769.

16. Y. Gamachi, H. Uehara, T. Ohtsuki, and I. Sasase, Upper bound of optical synchronous PPM/CDMA systems with interference canceller using reference signal, *Proc. Int. Conf. Telecommunications* (*ICT'97*), Melbourne, Australia, April 1997, pp. 909–914.

17. H. M. H. Shalaby, Cochannel interference reduction in optical synchronous PPM-CDMA systems, *IEEE Trans. Commun.* **COM-46**: 799–805 (June 1998).

18. H. Sawagashira, K. Kamakura, T. Ohtsuki, and I. Sasase, Direct-detection optical synchronous CDMA systems with interference canceller using group information codes, *IEEE Global Telecommunications Conf.* (*GLOBECOM'00*), San Francisco, Nov. 2000, pp. 1216–1220.

19. H. Sawagashira, K. Kamakura, T. Ohtsuki, and I. Sasase, Direct-detection optical synchronous CDMA systems with interference canceller using group information codes, *Trans. IEICE* **E83-A**(11): 2138–2142 (Nov. 2000) (letter).

20. J. A. Salehi and C. A. Brackett, Code division multiple-access techniques in optical fiber networks — Part II: Systems performance analysis, *IEEE Trans. Commun.* **COM-37**: 834–842 (Aug. 1989).

21. T. Ohtsuki, K. Sato, I. Sasase, and S. Mori, Direct-detection optical synchronous CDMA systems with double optical hard-limiters using modified prime sequence codes, *IEEE J. Select. Areas Commun.* **14**(9): 1879–1887 (Dec. 1996).

22. T. Ohtsuki, Performance analysis of direct-detection optical asynchronous CDMA systems with double optical hard-limiters, *IEEE J. Lightwave Technol.* **15**(3): 452–457 (March 1997).

23. T. Ohtsuki, Direct-detection optical asynchronous CDMA systems with double optical hard-limiters: APD noise and thermal noise, *Trans. IEICE* **E81-B**(7): 1491–1499 (July 1998).

24. T. Ohtsuki, Channel interference cancellation using electrooptic switch and optical hard-limiters for direct-detection optical CDMA systems, *IEEE J. Lightwave Technol.* **16**(4): 520–526 (April 1998).

25. C. L. Lin and J. Wu, A synchronous fiber-optic CDMA system using adaptive optical hardlimiter, *IEEE J. Lightwave Technol.* **16**(8): 1393–1403 (Aug. 1998).

OPTICAL TRANSMITTERS, RECEIVERS, AND NOISE

PETER J. WINZER
Bell Laboratories, Lucent Technologies
Holmdel, New Jersey

1. INTRODUCTION

Driven by the desire to meet the ever-growing bandwidth demand of our communication society while steadily reducing the cost per transmitted information bit, per-channel data rates in wavelength-division multiplexed (WDM) optical communication systems have continuously been increased, with 40-Gbit/s systems being commercially available today. Aggregate single-fiber transmission capacities on the order of 10 Tbit/s, as well as capacity-times-distance products exceeding several 10 Pbit/s km have been reported [1]. Conversely, tremendous advances in optical filter design and optical multiplexer technology have enabled channel spacings of some 10 GHz in dense WDM systems. Thus, 40 Gbit/s has become the data rate at which optics and electronics have met, and — for the first time in optical communications — has made *spectrally efficient modulation* a major issue.

As a consequence, the investigation of cost-effectively manufacturable transmitters for bandwidth-efficient optical modulation formats, as well as the optimization of high-speed optical receivers for dense WDM systems have become key topics of optical communications research and development. The quest is on for identifying combinations of modulation formats and receiver structures that can best cope with optical noise as well as with various linear and nonlinear signal-distortions accumulated along the fiber-optic transmission path [2], with the aim to trade high receiver sensitivities for longer transmission distances, relaxed component tolerances, or increased system margins.

Another field that asks for highly optimized optical transmitters and receivers is free-space optical communications. With applications both in terrestrial broadband access and in space-borne intersatellite links [3,5], free-space optical communications will enable future high-speed mobile data networking, bringing broadband data services to remote locations on the globe as well as to users on airplanes.

This section intends to open up the field of optical modulation and reception on an introductory level by discussing a selection of optical modulation techniques currently viewed as being most promising. Further, high-performance optical receiver structures and their noise properties are outlined, both for the fiber channel and for the free-space channel. Basic receiver design rules as well as important performance trade-offs are extracted. Frequently used concepts for quantifying receiver performance, such as *receiver sensitivity, quantum limit, Q-factor*, and *optical signal-to-noise ratio* (OSNR) are explained. To probe beyond the overview given in this chapter, and to acquire a more complete picture of the wide field of optical transmission, reception, and noise, the reader is kindly referred to the selection of excellent texts referenced at the end of the section.

2. OPTICAL MODULATION FORMATS AND THEIR IMPLEMENTATION

After giving a general classification of optical modulation formats, this section discusses the most important optical modulation techniques known today, with a particular view on their practical implementation by means of state-of-the-art high-speed opto-electronic components.

2.1. Classification

2.1.1. What to Modulate? The optical field[1] has three physical attributes that can be used to carry information: *Amplitude, phase* (including *frequency*), and *polarization*.

Depending on which of the three quantities is used to convey information, we distinguish between *amplitude-modulated, phase-modulated* (*frequency-modulated*), and *polarization-modulated* formats. Hybrid formats that simultaneously modulate two or more properties of the optical field (e.g., quadrature amplitude modulation (QAM)) have not yet made their way into high-speed optical communications. These formats are widely used in microwave communications, as well as in the related field of optical subcarrier-multiplexing, predominantly for cable-TV applications [6]. Here, several individually modulated signals on separate radio-frequency (RF) carriers are imprinted on an optical field by (linear) amplitude modulation.

[1] In optical communications, the term *optical field* is used to denote either one of the four electromagnetic field quantities observing the wave equation. It is usually expressed as a complex baseband quantity by eliminating the optical carrier frequency and is normalized such that its squared magnitude represents the optical power.

While amplitude and phase modulation have been widely used in high-speed optical communications, polarization modulation has received comparatively little attention so far [7]. This can primarily be attributed to the random polarization changes in optical fibers, necessitating active polarization control at the receiver. For amplitude- or phase-modulated formats and direct-detection receivers, however, polarization control is only required if polarization-mode dispersion (PMD) becomes an issue [8]. From a receiver point of view, this additional complexity would only be justifiable if polarization modulation offered significant baseline receiver sensitivity improvements over amplitude modulation, which it does not [9].

Note that our classification does not require a phase-modulated optical field to be constant-envelope, nor an amplitude-modulated field to have constant phase. It is the physical quantity from which information is extracted at the receiver that drives our classification. To give some examples: Differential phase shift keying (DPSK, cf. Section 2.3.1) is a phase-modulated format, regardless of whether it is transmitted constant-envelope or by means of phase-modulated optical pulses in the form of return-to-zero-DPSK (RZ-DPSK). Conversely, carrier-suppressed return-to-zero (CSRZ, cf. Section 2.2.3) is an amplitude-modulated format, regardless of the fact that the optical field's phase is additionally modulated in order to beneficially influence the spectrum.

2.1.2. How Many Symbols? The most widely used classes of optical receivers use *direct detection* (cf. Sections 3.2 and 3.6), that is they make use of the optical *power* $P = |E|^2$, the squared magnitude of the complex optical field amplitude E. If no optical phase-to-amplitude converting element is employed prior to detection, a direct-detection receiver is unable to distinguish between the two received symbols $E_{1,2} = \pm|E|$, since they both have the same optical power, $P_1 = P_2 = |\pm E|^2$. The additional degree of freedom gained by this ambiguity can be beneficially employed to shape the optical spectrum, or to make a format more resilient to distortions accumulated along the transmission line. Formats making use of this potential fall in the class of *pseudo-multilevel* or *polybinary* signals, depending on whether bit-correlations are introduced (as for duobinary formats, (M)DB, cf. Section 2.2.4) or not (as for carrier-suppressed return-to-zero, CSRZ, cf. Section 2.2.3). It is important to realize that these two classes of modulation formats use more than two symbols to encode a single bit of information, but transmit symbols at the bit rate R. For the formats of interest in optical communications today, the symbol alphabet $\{+|E|, -|E|, 0\}$ is used, which is mapped onto $\{0, |E|^2\}$ at the receiver. Pseudo-multilevel and polybinary signaling must not be confused with *multilevel* signaling, where $\log_2(M)$ bits are encoded on M symbols, and are then transmitted at a reduced symbol rate of $R/\log_2(M)$. Both multilevel amplitude shift keying [10] and (differential) quadrature phase shift keying (DQPSK, cf. Section 2.3.2) are multilevel optical modulation techniques. The difference between polybinary, pseudo-multilevel, and multilevel signaling is

Table 1. Symbol Encoding Examples for Multilevel, Pseudo-Multilevel, and Polybinary Signaling

Bit Sequence	0	0	1	0	1	1	1	0	0	1	0	1
pseudo-multilevel (CSRZ)	0	0	1	0	1	−1	1	0	0	−1	0	−1
polybinary (DB)	0	0	1	0	−1	−1	−1	0	0	−1	0	1
multilevel (DQPSK)	0		$+\pi/2$		π		$+\pi/2$		$-\pi/2$		$-\pi/2$	

visualized in Table 1, showing a data bit stream with three different symbol encodings.

2.1.3. Both Sidebands Needed? Apart from shaping (and compressing) the optical signal spectrum by means of (pseudo)-multilevel or polybinary signaling, it is possible for some modulation formats to additionally suppress half of their spectral content by appropriate optical filtering: Since the spectrum of real-valued baseband signals is symmetric around zero frequency, filtering out the redundant half of the spectrum (i.e., one of the two spectral '*sidebands*') preserves the full information content. This is exploited in *single-sideband* (SSB) signaling, where one sideband is completely suppressed, and in *vestigial-sideband* (VSB) signaling, where an optical filter with a gradual roll-off is offset from the optical carrier frequency to suppress major parts of one of the two sidebands, while at the same time performing some filter action on the other, desired sideband [11].

While broadband optical SSB is hard to generate in practice because of difficulties in implementing appropriate optical filter functions [10], optical VSB has been successfully demonstrated [12] on non-return-to-zero on/off keying (NRZ-OOK, cf. Section 2.2.1). Note that in fiber communications, VSB filtering is preferably done at the *receiver* instead of the transmitter, since if a sideband was suppressed at the transmitter, it would quickly reconstruct itself upon nonlinear fiber propagation. The advantage of using VSB comes from reduced WDM channel crosstalk for the desired sideband, if unequal channel spacings are employed. This situation is visualized in Fig. 1 [12], showing the composite spectrum of five wavelength-division multiplexed NRZ-OOK signals with alternating channel spacings of 1.2 and 1.7 times the data rate R. Severe WDM crosstalk is introduced for the sidebands on the closer-spaced sides (region A), making them useless for detection. Conversely, significantly less crosstalk is found for the sidebands on the larger-spaced sides (region B) than would be present if the channels were spaced on an equally-spaced frequency grid.

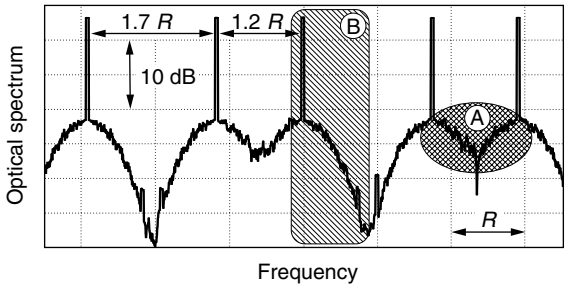


Figure 1. Vestigial sideband (VSB) transmission of optical signals on an unequally-spaced frequency grid to avoid WDM channel crosstalk [12].

2.2. Amplitude Modulation Formats

2.2.1. Non-Return-to-Zero On/Off Keying (NRZ-OOK). The simplest of all optical modulation formats is non-return-to-zero on/off keying (NRZ-OOK), often just called NRZ. This format imprints data on an optical carrier by switching light on and off. Historically, this was the first, and is still the most widely deployed optical modulation format. It has been used to *directly modulate*[2] both light-emitting diodes (LEDs) and lasers. Unfortunately, directly modulated laser light is highly chirped, that is, it exhibits strong residual phase modulation, which broadens the optical spectrum and degrades transmission performance through interaction with optical fiber dispersion in many important transmission scenarios. Thus, for modulation speeds above 2.5 Gbit/s and/or for long-haul fiber communication systems, *external modulation* has to be used. Here, the light of a continuously operating laser source is modulated by means of an external device, engineered for low chirp, or even designed for chirp-free operation. The two most important external modulators are semiconductor *electro-absorption modulators* (EAMs) and Lithium-Niobate ($LiNbO_3$) *Mach-Zehnder modulators* (MZMs). Both are commercially available for 40-Gbit/s modulation today.

EAMs [13] have the advantage of low-drive voltages (typ. 2 V), and are cheap in volume production. However, they still produce some residual chirp, have dynamic extinction ratios (maximum-to-minimum modulated light power) typically not exceeding 10 dB, and have limited optical power-handling capabilities (typ. 10 dBm). Their fiber-to-fiber insertion losses are about 10 dB, which has led to the integration with laser diodes, thus avoiding the input fiber-to-chip interface. *Electro-absorption modulated lasers* (EMLs) with output powers on the order of 0 dBm are widely available today. Another way of eliminating the high insertion losses of EAMs is the integration with semiconductor optical amplifiers (SOAs), which can even yield some net fiber-to-fiber amplification [14]. Figure 2a shows typical transmission characteristics of an EAM as a function of drive voltage. Note that the absorption of the EAM saturates at high drive voltages.

MZMs have excellent extinction performance (typ. 20 dB), can be made chirp free by balanced driving, and have lower insertion losses than EAMs (typ. 5 dB). The

[2] Using *direct modulation*, data are directly superimposed on a light-emitting device's drive current, which otherwise has biasing functionality only.

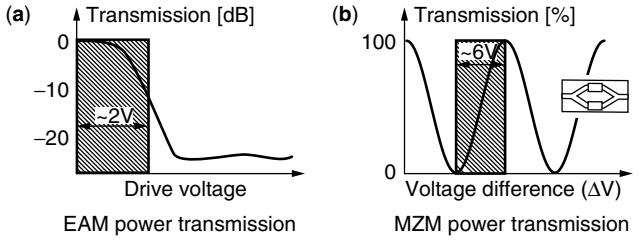


Figure 2. Typical optical power transmission characteristics of EAMs (**a**) and MZMs (**b**).

required (high-speed) peak-to-peak drive voltages of some 6 V, however, often represent serious practical problems. The MZM transfer characteristics is sinusoidal, owing to the Mach-Zehnder structure of the device (cf. inset to Fig. 2b): The incoming light is split into two paths at an input coupler. One (or both) paths are equipped with phase modulators that let the two fields acquire some phase difference relative to each other. Finally, the two fields interfere (destructively or constructively, depending on the modulated phases) at an output coupler. The optical field transfer function $T_E(V_1, V_2)$ thus reads

$$T_E(V_1, V_2) = \tfrac{1}{2}\{e^{j\phi(V_1)} + e^{j\phi(V_2)}\} = e^{j(\phi(V_1)+\phi(V_2))/2}\cos[(\phi(V_1) - \phi(V_2))/2] \quad (1)$$

where $\phi(V_{1,2})$ are the voltage-modulated optical phases of the two MZM arms. Since the phase modulation is a linear function of the drive voltage, the MZM *power* transfer function depends only on the drive voltage difference ΔV, $T_P(V_1, V_2) = |T_E(V_1, V_2)|^2 = T_P(\Delta V)$, which gives an additional degree of freedom in adjusting modulator chirp [15]. If the two modulator arms are driven by the same amount, but in opposite directions ($\phi(V_1) = -\phi(V_2)$), the phase term in Eq. (1) vanishes, resulting in purely real-valued transmission characteristics (i.e., in chirp-free operation). This driving condition is known as *balanced driving* or *push-pull operation*. Note that balanced driving cannot only be used to eliminate chirp, but also to reduce the output power requirements of the RF driver amplifiers by 6 dB (at the expense of having to use an additional amplifier, of course). Optical NRZ data signals are usually generated by driving the MZM from its minimum transmission to its maximum transmission, as visualized in Fig. 2b. Note that the non-linear parts of the MZM transfer function at high and low transmission can suppress overshoots and ripple on the electrical NRZ drive signal.

Figure 3 shows optical spectrum and optical eye diagram[3] of an idealized NRZ signal. The optical spectrum, defined as the bit-pattern-averaged squared magnitude of the optical field's Fourier transform, is composed of a continuous portion, which reflects the shape of the individual NRZ data pulses, and discrete tones at integer

[3] *Eye diagrams* are important means of visualizing the quality of digital signals. They are formed by plotting on top of each other copies of the same modulated bit pattern, shifted by integer multiples of the bit duration.

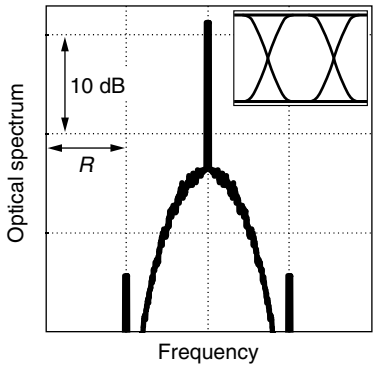


Figure 3. Optical spectrum and eye diagram of an NRZ signal.

multiples of the data rate. The weight of these tones is determined by the optical spectrum of the NRZ data pulses, and therefore depends on the NRZ rise/fall times. For ideal (rectangular) NRZ signals all tones vanish, apart from the one at zero frequency.

2.2.2. Return-to-Zero On/Off Keying (RZ-OOK). Regardless of the modulation format (phase or amplitude), NRZ often suffers from bandwidth-limitations, both at the transmitter and at the receiver, leading to the presence of intersymbol interference (ISI) in the bit sequence to be detected. ISI denotes the corruption of bits (most notably, of isolated '0'-bits) by their neighboring '1'-bits. In optical communications, ISI is particularly harmful, since detection noise often grows linearly with signal amplitude (cf. Sections 3.2, 3.4, and 3.6). Figure 4 shows typical NRZ and RZ electrical eye diagrams at the decision gate of a receiver for the same average optical input power and under the same filtering conditions, where the effect of ISI on NRZ becomes evident. In addition to ISI introduced by bandlimiting (optical or electrical) elements, NRZ formats degrade rapidly in many important fiber transmission scenarios. *Return-to-zero* (RZ, *impulsive coding*) coding mitigates these problems, and leads to enhanced system performance [16,17]. In the case of RZ-OOK, information is encoded on the presence or absence of optical pulses. By well centering the pulses in the bit slots, pattern effects coming from limited NRZ drive signal bandwidths are largely eliminated.

The advantages found for RZ coding come at the expense of higher optical transmission bandwidth requirements, as well as of more complicated transmitter structures, as shown in Fig. 5. Usually, RZ-OOK is generated from an optical NRZ signal by carving out pulses by

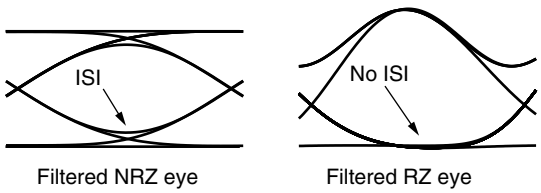


Figure 4. Eye diagrams for NRZ and RZ signals. Intersymbol interference (ISI) affects NRZ performance, while it is not seen for RZ.

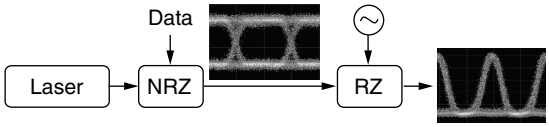


Figure 5. Structure of a typical RZ transmitter, consisting of a laser source, an external NRZ modulator, and a RZ pulse carver.

means of an additional modulator, termed *pulse carver*. Typically, pulse carvers are implemented as sinusoidally driven EAMs or MZMs.

Using an EAM, short (a few ps) optical pulses can be realized by biasing the modulator well in its absorption region, and letting only the peak portion of the sinusoidal drive signal reach appreciable transmission, as shown in Fig. 6. This technique is therefore widely used in optical time-division multiplexing (OTDM) transmitters [18,19].

If a MZM is used for pulse carving, three operating conditions have to be distinguished:

- Sinusoidally driving the MZM at the data rate between minimum and maximum transmission results in optical pulses with a full-width-half-maximum (FWHM) of 50% of the bit duration (a duty cycle of 50%), as shown in Fig. 7 (dashed). Decreasing the modulation swing while adjusting the modulator bias such as to still reach good extinction between pulses, the duty cycle can in principle be reduced to 36%, however, with significant excess insertion loss, since the modulator is then no longer driven to its transmission maximum. At a duty cycle of 40%, the excess insertion loss amounts to 2.2 dB. Increasing the drive voltage to reduce the pulse width (as can be done in the case of EAMs) is not possible with MZMs, owing to the periodic nature of the MZM transmission function.
- Sinusoidally driving the MZM at *half* the data rate between its transmission minima produces a pulse whenever the drive voltage passes a transmission maximum, as visualized in Fig. 7 (solid). This way, duty cycles of 33% can be realized, however, without the possibility for adjustments by varying the drive voltage. The doubled peak-to-peak drive voltage requirements usually pose little technical problems, since narrow-band RF amplifiers can be used.
- Sinusoidally driving the MZM at half the data rate between its transmission maxima results in pulses with 67% duty cycle and with alternating phase. The resulting format is called *carrier-suppressed RZ* (CSRZ), and will be discussed in Section 2.2.3.

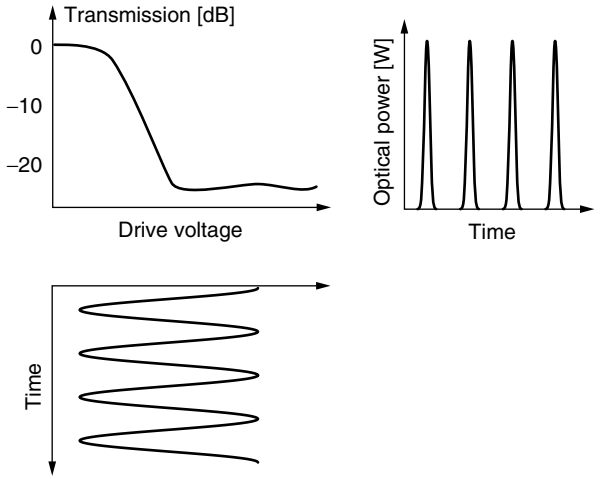


Figure 6. Sinusoidally driven EAM used as RZ pulse carver to attain short optical pulses, with duty cycles below 33%.

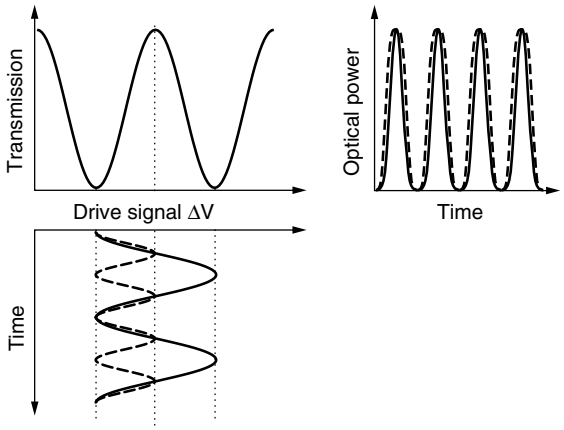


Figure 7. Sinusoidally driven MZM as pulse carver for 33%-duty-cycle RZ (solid) and 50%-duty-cycle RZ (dashed).

Other, less frequently used RZ-OOK modulation techniques include mode-locked lasers in combination with external NRZ-modulators to achieve very low duty cycle pulses for OTDM applications [20], single-step RZ-OOK modulation by means of an electrical RZ drive signal [21], and techniques employing the rising and falling edges of the electrical NRZ signal for RZ pulse generation [22,23]. Note that these methods make do with a single external optical modulator, without the need for a pulse carver.

Spectra and eye diagrams of 50% duty cycle RZ (gray) and 33% duty cycle RZ (black), as produced by a MZM in push-pull operation, are shown in Fig. 8.

2.2.3. Carrier-Suppressed Return-to-Zero (CSRZ). Carrier-suppressed return-to-zero (CSRZ) is a pseudo-multilevel modulation format, characterized by reversing

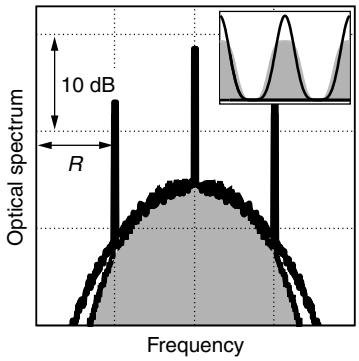


Figure 8. Optical spectra and eye diagrams for 50% duty cycle RZ (gray) and 33% duty cycle RZ (black), as produced by a MZM in push-pull operation.

the sign of the optical field at each bit transition. In contrast to the duobinary formats detailed in Section 2.2.4, the sign reversals occur at *every* bit transition, and are completely *independent* of the information-carrying part of the signal. CSRZ is most conveniently realized by sinusoidally driving a MZM pulse carver at half the data rate between its transmission maxima, as visualized in Fig. 9. Since the optical field transfer function $T_E(\Delta V)$ (dashed) of the MZM changes its sign at the transmission minimum (cf. Eq. (1) for push-pull operation), phase inversions between adjacent bits are produced. Thus, on average, the optical field of half the '1'-bits has positive sign, while the other half has negative sign, resulting in a zero-mean optical signal. As a consequence, the carrier at the optical center frequency vanishes, giving the format its name.

Using a MZM to generate CSRZ results in a duty cycle of 67%, which can be brought down to 50% at the expense of excess insertion loss by reducing the drive voltage swing. At a duty cycle of 55%, an excess insertion loss of 2 dB has to be accepted. It is important to note that, due to its most widely used practical implementation with MZMs, the duty cycle of CSRZ signals usually differs from the one of standard RZ. Thus, care has to be taken when comparing the two formats, since some performance differences result from the carrier-suppressed nature of CSRZ, while others simply arise from the different duty cycles.

Spectrum and eye diagram of 67%-duty cycle CSRZ, as generated by a MZM in push-pull configuration, are shown in Fig. 10. For comparison, the spectrum of a (hypothetical) 67%-duty cycle RZ signal is also given (gray). Note that the *only* difference between the two spectra is the location of the discrete tones.

2.2.4. Duobinary and Modified Duobinary (DB, MDB). Duobinary (DB) and modified duobinary (MDB) signals are polybinary signals, a subset of the partial response signaling format [11,24]. In optical communications they have also become known under the keywords *phase-shaped binary transmission* (PSBT) [25] and *phased amplitude-shift signaling* (PASS) [26]. Most conveniently, optical (M)DB signals, like CSRZ, employ

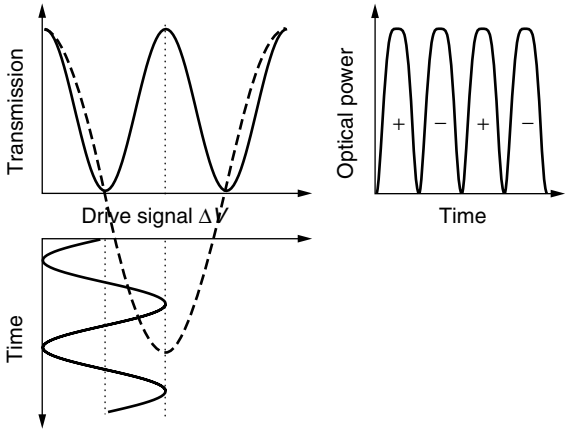


Figure 9. Sinusoidally driven MZM as pulse carver for 67%-duty-cycle CSRZ. The solid and dashed transmission curves apply for the optical power ($T_P(\Delta V)$) and field ($T_E(\Delta V)$), respectively.

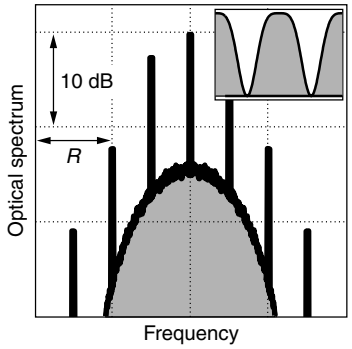


Figure 10. Optical spectrum and eye diagram of 67%-duty cycle CSRZ, as generated by a MZM in push-pull configuration (black). The spectrum of a 67%-duty cycle RZ signal without phase reversals is given for comparison (gray).

the signaling set $\{0, \pm|E|\}$, taking advantage of the power-detecting property of direct detection optical receivers that automatically converts the three optical symbols to the two electrical symbols $\{0, |E|^2\}$. However, unlike with CSRZ, the optical phases of the individual bits additionally depend on the bit pattern: For DB signaling, a phase change occurs whenever there is an odd number of '0's between two successive '1's, whereas for MDB the phase changes for each '1' (even for adjacent '1's), independent of the number of '0's inbetween (cf. also Table 1).

(M)DB signals are more tolerant than conventional binary signals with respect to chromatic dispersion, narrow-band optical filtering (thus allowing for closer WDM channel spacings), as well as to some non-linear transmission impairments. Explanations can be given both in the frequency domain [10,27] and in the time domain [25,26,28], the latter lending itself to a particularly intuitive interpretation: Consider the bit pattern ...0010100..., with the '1'-bits being represented by the tall, shaded pulses in Fig. 11. When transmitted through narrow-band optical filters or over dispersive optical fiber, the pulses broaden (hatched) to let some energy spill into the isolated '0'-bit. If the two pulses have the same optical phase, their optical fields add up constructively, leading to severe ISI (dashed). For (M)DB, however, two pulses separated by an isolated '0'-bit always have opposite phases, which lets them interfere destructively and thus reduces ISI (solid).

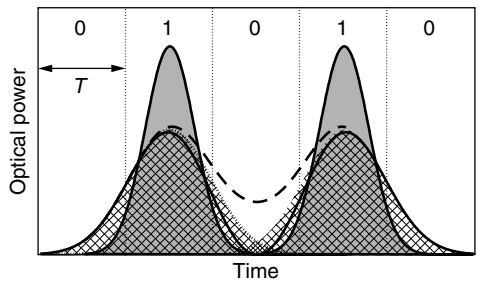


Figure 11. Initially narrow optical pulses (shaded) broaden through fiber dispersion or optical filtering (hatched). If the two pulses have the same optical phase, their optical fields add up constructively (dashed). For (M)DB, the two pulses have opposite phases, which lets them interfere destructively (solid) [25,26].

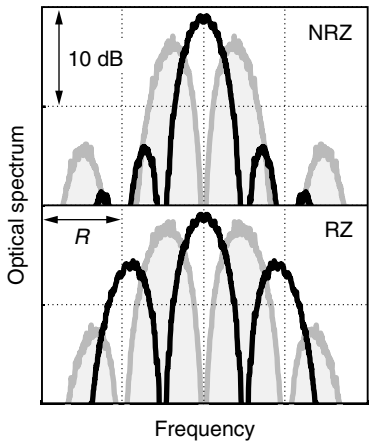


Figure 12. Optical spectra of duobinary (black) and modified duobinary (gray) signals in NRZ coding (upper) and RZ coding (lower).

Both DB and MDB can be implemented in RZ or NRZ format. Figure 12 shows the optical spectra of NRZ (upper) and RZ (lower) DB (black) and MDB (gray). A characteristic feature is the spectrally compressed main lobe as compared to NRZ-OOK (cf. Fig. 3). For NRZ-DB, the side lobes are filtered out for optimum performance [10]. Note that (M)DB spectra have no discrete spectral components, which helps to suppress stimulated Brillouin scattering (SBS) in optical fibers [29].

Duobinary transmitters are usually implemented using a three-level electrical drive signal $\{-1, 0, +1\}$ in combination with a MZM driven between its transmission maxima (like for CSRZ, Fig. 9, but using the data signal instead of a sinusoid). The methods resulting in chirp-free (M)DB signals operate the MZM in push-pull mode. As shown in Fig. 13a, the three-level electrical drive signal can be generated using analogue addition (DB) or subtraction (MDB) of the bit sequence with a 1-bit-delayed replica of itself, provided that appropriate precoding is performed on the data [11,27]. Since the required three-level (linear) RF driver electronics are hard to implement in practice, one usually resorts to method (b), taking a highly low-pass filtered version of the precoded data signal to drive the MZM. The filter bandwidth B has to be chosen on the order of one fourth the data rate R [27]. A third realization (c) uses a MZM as a phase modulator (cf. Section 2.3.1) to generate an intermediate DPSK signal, which is transformed to (M)DB using an optical delay interferometer [30]. By reducing the optical delay to values less than the bit duration T, variable duty cycle RZ-MDB can be generated without the need for a pulse carver [23,31].

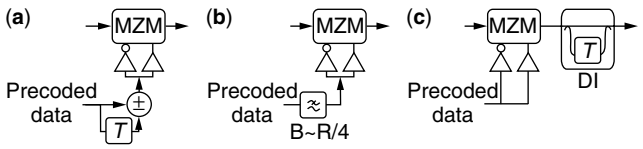


Figure 13. (Modified) duobinary signals are either generated by driving a MZM around its transmission minimum (cf. Fig. 9) using a 3-level electrical drive signal [(a) and (b)]. Alternatively, (M)DB can be generated using a MZM as phase modulator, and passing the resulting DPSK signal through a delay interferometer (DI) (c).

2.2.5. Chirped Return-to-Zero (CRZ). Chirped return-to-zero (CRZ) is predominantly used for ultra-long-haul fiber communication, as found in transoceanic (submarine) systems [32,33]. CRZ signals are generated by sinusoidally modulating the phase of a RZ signal at the data rate, using a separate phase modulator. The intentionally introduced chirp on the one hand beneficially influences nonlinear fiber transmission performance, but on the other hand broadens the signal spectrum. In WDM systems, the amplitude of the sinusoidal phase modulation has thus to be optimized by trading the gain due to enhanced nonlinear propagation performance against WDM channel crosstalk. Typically, the optimum phase modulation amplitude amounts to $\sim$1 rad [32].

The benefits of CRZ obviously come at the expense of more complex transmitter architectures, comprising a total of three external modulators whose drive signals have to be carefully synchronized. Integrated GaAs/AlGaAs modulators for CRZ, combining NRZ data modulator, RZ pulse carver, and CRZ phase modulator in one module, have been reported [34].

2.3. Phase Modulation Formats

The most widely used class of optical receivers employs direct detection, that is, the receiver is only sensitive to optical *power* variations (cf. Sections 3.2 and 3.6). To detect modulation of the optical field's *phase*, phase-to-amplitude converting elements therefore have to be inserted into the optical path prior to detection. Since these elements are unable to offer an absolute optical phase reference, the phase reference has to be provided by the signal itself: Each bit acts as a phase reference for another bit, which is at the heart of all *differential phase shift keying* formats.

2.3.1. Differential Phase Shift Keying (DPSK). Binary differential phase shift keying (BDPSK, or simply DPSK) encodes information on a binary phase change between adjacent bits. A logical '1' is encoded onto a π phase change, whereas a logical '0' is represented by the absence of a phase change. Thus, like for (M)DB, an appropriate precoding circuit has to be employed at the transmitter prior to modulation. Like OOK, DPSK can be implemented in RZ and NRZ format. The main advantage from using DPSK instead of OOK comes from a 3-dB sensitivity improvement at the receiver, provided that balanced detection is employed [9,35]. This enhanced sensitivity directly translates into increased transmission distance [36].

An optical (N)RZ-DPSK transmitter is shown in Fig. 14. The phase of the optical field of a narrow-linewidth laser source is flipped between 0 and π using the precoded (differentially encoded) version of the NRZ data signal, as visualized by the two bit patterns in the figure. If a straight-line phase modulator (PM) is used, the speed of the phase transitions is limited by the combined bandwidth of driver amplifier and phase modulator, while the

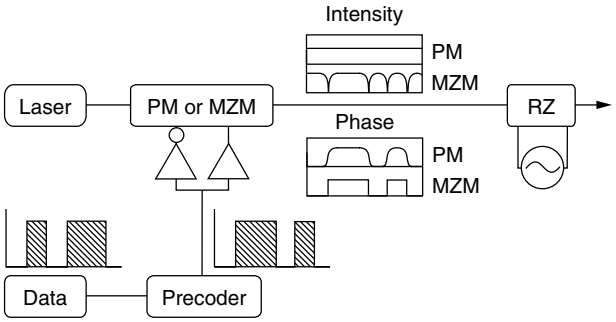


Figure 14. Setup of a RZ-DPSK transmitter. Phase modulation can either be achieved using a MZM, or by means of a straight-line phase modulator (PM), resulting in different amplitude and phase waveforms.

intensity of the phase-modulated light is constant. Instantaneous π phase jumps can be realized at the expense of some residual intensity modulation of the phase modulated light by using a dual-drive MZM, symmetrically driven around zero transmission [37], in analogy to the (M)DB transmitter of Fig. 13(a) and (b) and the CSRZ transmitter of Fig. 9. Typical intensity and phase waveforms of the two modulation techniques are shown at the output of the phase modulator in Fig. 14, where the upper traces apply to PM and the lower traces to MZM phase modulation. Like for OOK, a subsequent pulse carver converts the NRZ-DPSK signal to RZ-DPSK, if desired.

Figure 15 shows spectra and eye diagrams for RZ-DPSK (black) and NRZ-DPSK (gray), as generated by a MZM operated as a phase modulator. The carrier-free nature of the spectra, like for (M)DB, owes to the balance of $-|E|$ and $+|E|$ amplitude levels. Note the absence of a '0'-bit rail in the eye diagrams, which is characteristic for phase-coded formats. The deep amplitude dips between two bits

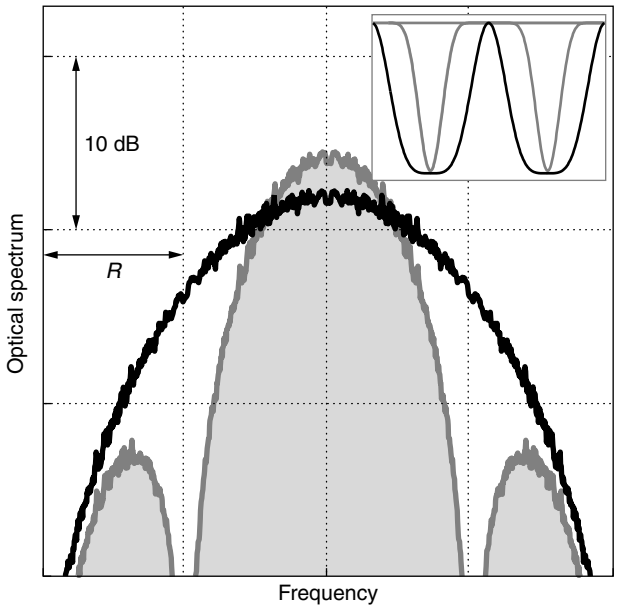


Figure 15. Optical spectra and eye diagrams of NRZ-DPSK (gray) and 33%-duty cycle RZ-DPSK (black), as produced by a MZM as phase modulator.

in the NRZ-DPSK eye represent the residual amplitude modulations of the MZM caused by the finite NRZ drive signal bandwidth.

Since DPSK cannot directly be received with direct-detection techniques, an optical delay interferometer (DI) is inserted in the optical path at the receiver to convert the differential phase modulation into amplitude modulation. As shown in Fig. 16, a DI splits the phase modulated signal into two paths, into one of which a delay equal to the bit duration T is introduced. At the DI's output coupler, the phase modulated optical field thus interferes with its one-bit-delayed replica to produce destructive interference at port $\mathcal{A}$ whenever there is *no* phase change, and constructive interference whenever there *is* a phase change, in agreement with the DPSK-coding rule described above. To exploit the 3-dB sensitivity advantage of DPSK over OOK, a *balanced receiver* has to be employed, where the second DI-output port $\mathcal{B}$, yielding the inverted data pattern, is also made use of, and the difference signal is detected [9,35].

2.3.2. Differential Quadrature Phase Shift Keying (DQPSK). Instead of using two phase levels (DPSK), one can use four phase levels $\{0, +\pi/2, -\pi/2, \pi\}$ to produce differential quadrature phase shift keying (DQPSK). DQPSK is a true four-level signaling format, transmitting symbols at *half* the aggregate bit rate (cf. Table 1). While the receiver sensitivity benefit over OOK that is gained for DPSK is largely lost for DQPSK, the transmission bandwidth is significantly reduced, potentially allowing for higher spectral efficiency in WDM systems, as well as for increased tolerance to chromatic dispersion and PMD [38].

A DQPSK transmitter is best implemented by taking advantage of the *exact* π-phase shifts produced by a MZM operated as a phase modulator (cf. Fig. 9). Figure 17 shows the corresponding transmitter setup [38], consisting of a continuously operating laser source, a splitter to divide the light into two paths of equal intensity, two MZMs operated as phase modulators, a $\pi/2$-phase shifter in one of the paths, and a combiner to produce a single output signal. While the modulated field E_1 of the upper path can take on the values $\pm E_0/\sqrt{2}$, the lower path produces $E_2 = \pm E_0 e^{j\pi/2}/\sqrt{2}$, leading to the four symbols $E_0/\sqrt{2} \cdot \{e^{j\pi/4}, e^{j3\pi/4}, e^{j5\pi/4}, e^{j7\pi/4}\}$ after the combiner. A pulse carver can optionally be added to yield RZ-DQPSK.

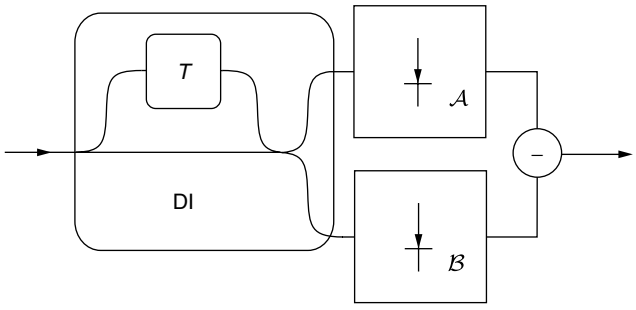


Figure 16. Balanced DPSK receiver using an optical delay interferometer (DI) to convert the phase modulation to amplitude modulation.

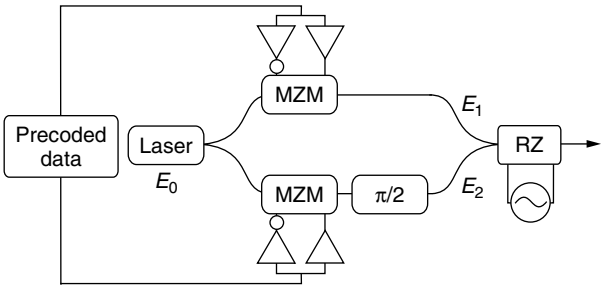


Figure 17. Structure of a DQPSK transmitter. Two MZMs are used as phase modulators, and the two separately modulated fields are combined with a $\pi/2$ phase shift [38].

The *shape* of the DQPSK optical spectra is identical to those of DPSK (Fig. 15). However, the DQPSK spectrum is compressed in frequency by a factor of two due to the halved symbol rate for transmission at the same bit rate.

At the receiver, the DQPSK signal is split, and *two* balanced receivers of the form depicted in Fig. 16 are used in parallel to simultaneously demodulate the two data streams contained within the DQPSK signal [38]. Note that the DI delay has to equal the *symbol* duration for DQPSK demodulation, which is *twice* the bit duration. Due to the DQPSK phase shifts of $\pi/2$ (instead of π for DPSK), the two DIs cannot simultaneously be operated to full destructive and to full constructive interference, which results in a reduced demodulated eye opening for DQPSK.

3. OPTICAL RECEIVER CONCEPTS AND NOISE

3.1. The *Q*-Factor

Before embarking on optical receiver concepts, we will briefly discuss the *Q-factor* as an important parameter that is widely used in optical communications to describe receiver performance. Although occasionally frowned upon by theoreticians, since its derivation relies on (sometimes hard to justify) approximations, the *Q*-factor allows for intuitive interpretations, reasonably accurate quantitative predictions, and has also become an indispensable tool for experimentalists [32].

The *Q*-factor was first introduced by Personick in 1973 [39] to relate mean and variance of the electrical signal at the receiver's decision gate to a bit-error ratio (BER), the quantity of ultimate interest when assessing the performance of digital communication systems. Leaving the derivation of the *Q*-factor to more comprehensive texts [35,40,41], we restrict ourselves to its definition,

$$Q = \frac{|s_1 - s_0|}{\sigma_1 + \sigma_0}, \tag{2}$$

where $s_{0,1}$ are the mean electrical signal amplitudes for a logical '0' and '1' at the decision gate, and $\sigma_{0,1}$ are the associated noise standard deviations. Under the assumption of Gaussian detection statistics, which is sufficiently accurate in most situations of practical interest [42], the BER is related to the *Q*-factor via

$$\mathrm{BER} = 0.5\ \mathrm{erfc}[\mathrm{Q}/\sqrt{2}], \tag{3}$$

where $\mathrm{erfc}[x] = (2/\sqrt{\pi}) \int_x^\infty \exp(-\xi^2)\, d\xi$ denotes the complementary error function. For $\mathrm{BER} = 10^{-9}$, which is often taken as a baseline for specifying receiver sensitivities, we have $Q \approx 6$.

Note that σ_0 and σ_1 may differ from each other, since many important noise terms encountered in optical communications are *signal-dependent*, that is, the noise variance is a function of the optical signal power. For purely *signal-independent* noise ($\sigma_1 = \sigma_0 = \sigma$), Eqs. (2) and (3) reduce to $\mathrm{BER} = 0.5\ \mathrm{erfc}[|s_1 - s_0|/(2\sqrt{2}\sigma)]$, a well-known expression in classical communication theory [11].

3.2. *Pin*-Receiver

The *pin*-receiver depicted in Fig. 18 is the simplest optical receiver structure. It consists of a *pin*-photodiode,[4] some postdetection electronic amplification and filtering with (single-sided) bandwidth B_e (electronics impulse response $h(t)$), and a sampling-and-decision device that restores the digital data. Detection of the filtered signal $s(t)$ is corrupted by two types of noise in a pin-receiver, *shot noise* and *electronics noise*.

Shot noise is a direct consequence of the quantum nature of light: Interactions of light and matter can only take place in discrete energy quanta (*photons*), governed by the rules of quantum statistics. Thus, a discrete, random number of electron-hole pairs is generated in a semiconductor diode when light impinges on it, causing the photocurrent to leave the diode in individual elementary impulses, each carrying the elementary charge $e \approx 1.602 \cdot 10^{-19}$. As, as visualized in Fig. 19. This fine structure of the electrical signal is perceived as shot noise [43,45]. On average, a fraction η of the incoming optical power $p(t)$ is converted to an electric current, leading to an average electrical signal amplitude of

$$\langle s(t) \rangle = (R_T) \cdot S(p * h)(t) \tag{4}$$

where $S = \eta e/(hf)$ [A/W] is the receiver's responsivity. The symbol * denotes a convolution, and $h(t)$ is normalized to let the low-frequency portion of its spectrum equal unity. Depending on whether the electrical signal $s(t)$ is specified in terms of current or of voltage, a resistance R_T has to be taken into account that converts the current-output of the *pin*-photodiode into a voltage. This resistance is

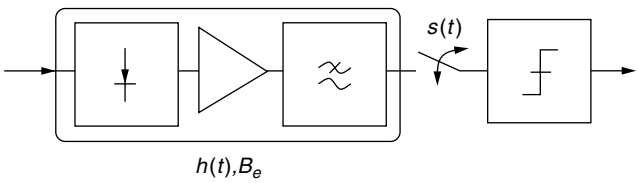


Figure 18. Setup of a *pin*-receiver, incorporating a *pin*-photodiode, an electrical preamplifier, electrical low-pass filtering, and a sampling-and-decision device.

[4] The abbreviation *pin* stands for *p-doped/intrinsic/n-doped*, and describes the basic layer structure of the associated semiconductor device.

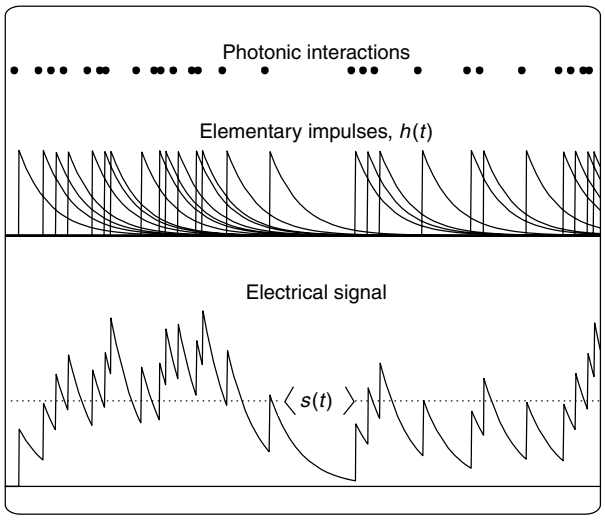


Figure 19. Photons arrive at random, dictated by quantum statistics. Each photonic interaction produces an elementary electronic impulse. The impulses add up to produce the overall electrical signal, whose fluctuations are known as shot noise [43].

frequently referred to as *transimpedance; hf* denotes the photon energy. ($h \approx 6.626 \cdot 10^{-34}$ Js is Planck's constant, and f stands for the optical carrier frequency of light.) The shot noise variance associated with photodetection is given by [43]

$$\sigma^2_{shot}(t) = (R_T^2) \cdot eS(p * h^2)(t) \approx (R_T^2) \cdot 2eSp(t)B_e \qquad (5)$$

The approximation in Eq. (5) applies for optical power variations that are slow compared to the speed of the electronics. Note that shot noise is a *nonstationary*, non-Gaussian noise process in general; its statistical parameters, most notably its variance, change with time.

Electronics noise is the sum of all stationary noise sources generated within the opto-electronic circuitry, *independent* of the optical signal, such as thermal noise, transistor shot noise, $1/f$-noise, or dark-current shot noise. The design of the receiver front-end electronics significantly impacts its noise performance, and is detailed in numerous excellent texts [39,46].

On a system level, electronics noise is often characterized by an *equivalent noise current density* $i_n[\mathrm{A}/\sqrt{\mathrm{Hz}}]$, which can be converted to an electronics noise variance σ^2_{elec} at the decision gate by

$$\sigma^2_{elec} = (R_T^2) \cdot i_n^2 B_e \qquad (6)$$

In the Gbit/s-regime, i_n is typically on the order of some 10 $\mathrm{pA}/\sqrt{\mathrm{Hz}}$. Alternatively, the noise performance of receivers can be specified using the *noise equivalent power* (NEP) $[\mathrm{W}/\sqrt{\mathrm{Hz}}]$, which is usually defined as the average optical power per square-root electrical receiver bandwidth that would be required to make the average electrical power equal to the electronics noise variance,

$$\sigma^2_{elec} = (R_T^2) \cdot S^2 \mathrm{NEP}^2 B_e \qquad (7)$$

3.3. Receiver Sensitivity and Quantum Limit

Instead of using equivalent noise current density or NEP, optical receiver front-ends are sometimes also characterized in terms of their *receiver sensitivity*, defined as the average optical power that is required at the receiver input to obtain a certain BER (typ.10^{-9}) at a certain data rate and for a certain modulation format (typ. NRZ-OOK). While the receiver sensitivity is undoubtedly of high interest in optical receiver design, it comprises not only the degrading effects of noise, but also encompasses essential properties of the received signal, such as extinction ratio, signal distortions and intersymbol interference (ISI), generated either within the transmitter or within the receiver itself. Thus, knowledge of the receiver sensitivity alone does not allow trustworthy predictions on how the receiver will perform for other formats (e.g., for RZ-OOK).

Although electronics noise usually dominates shot noise, it can in principle be engineered to zero. Shot noise, however, is fundamentally present. The limit, when *only* fundamental noise sources determine receiver sensitivity is called *quantum limit* in optical communications. The existence of quantum limits makes optical receiver design an exciting task, since there is always a fundamental measure against which practically implemented receivers can be compared, much like the Shannon-limit in information theory. Note, however, that each class of receivers in combination with each class of modulation formats has its own quantum limit.

The quantum limit for the *pin*-receiver using OOK is obtained by ignoring thermal noise, assuming a perfect ($\eta = 1$) receiver, and evaluating the BER for the Poissonian photon statistics of perfect laser light [43,44]. Leaving the derivation to more detailed texts [9,35,41], we merely cite the result,

$$\mathrm{BER} = 0.5 \exp[-2\overline{n}] \qquad (8)$$

where $\overline{n}$ is the average number of photons/bit at the receiver input. For $\mathrm{BER} = 10^{-9}$, the *quantum limited receiver sensitivity* of the *pin*-receiver is $\overline{n} = 10$ photons/bit (average power), or $n_1 = 20$ photons per '1'-bit. Note that specifying the receiver sensitivity in terms of *photons / bit* leads to more fundamental statements than specifying it in terms of an average optical power $\overline{P}$ ([W] or [dBm]), since both wavelength dependence and bit-rate dependence of receiver performance are eliminated. The two measures are related via

$$\overline{P} = \overline{n} hfR \qquad (9)$$

where R denotes the data rate. The intriguingly low-receiver sensitivity of *pin*-receivers, however, does not apply to practically implementable receivers, since in reality electronics noise by far dominates shot noise. As a consequence, receiver sensitivities achieved by *pin*-receivers are typically 20–30 dB off the quantum limit: Assuming an equivalent noise current density of 10 $\mathrm{pA}/\sqrt{\mathrm{Hz}}$ and a 10-Gbit/s receiver with some 7-GHz bandwidth operating at a wavelength of 1550 nm, the electronics noise variance amounts to $7 \cdot 10^{-13}\mathrm{A}^2$, while the '1'-bit shot noise variance going with the detection of 10 photons/bit comes to about $4 \cdot 10^{-17}\mathrm{A}^2$, four orders of magnitude below electronics noise. For realistic BER, the Q-factor is thus entirely determined by electronics noise, and for $Q = 6$ we arrive at a receiver sensitivity of some $\overline{n} \approx 5000$ photons/bit, 27 dB above the quantum limit.

To achieve higher receiver performance, more advanced receiver types must be employed. There are basically three ways to proceed: *Avalanche photodetection, coherent detection*, and *optically preamplified detection*. These rather diverse techniques, which will be discussed in the following sections, have still one common attribute: They all amplify the received signal before or at the stage of photodetection, while at the same time introducing additional noise. In the limit when the newly introduced noise terms dominate electronics noise, receiver performance becomes independent of electronics noise, leading to the respective quantum limits. In that limit, any further increase of the respective gain mechanism does *not* affect receiver performance any more. In contrast to *pin*-receivers, the quantum limits can be closely approached with these receiver types in experimental reality.

3.4. Avalanche Photodiode (APD) Receiver

An avalanche photodiode (APD) is the semiconductor equivalent to a photomultiplier tube. The incoming light generates primary electron-hole pairs (like in a *pin*-diode), which are then accelerated in a high-field region to launch an avalanche multiplication process through ionizing collisions [43]. The average number of the resulting secondary electron-hole pairs relative to the primary electron-hole pairs is called *avalanche gain* M_{APD}. The avalanche multiplication process is by itself a random process, since the exact number of secondary electron-hole pairs generated by a primary pair varies randomly. The unavoidable shot noise present for primary photodetection (cf. Fig. 19) is thus enhanced. This increase in detection noise is quantitatively captured in the APD's *noise enhancement factor* $F_{APD} > 1$ via the *multiplied shot noise* relationship [43]

$$\begin{aligned}\sigma^2_{shot,APD}(t) &= (R_T^2)\cdot eSM^2_{APD}F_{APD}(p * h^2)(t)\\ &\approx (R_T^2)\cdot 2eSM^2_{APD}p(t)F_{APD}B_e \end{aligned} \tag{10}$$

where the approximation, again, holds for optical power variations slow compared to the detection electronics' speed. In addition to multiplied shot noise, an APD also generates multiplied dark current shot noise through avalanche multiplication of dark current charge carriers, which is stationary and independent of the optical power, and can thus be added to the electronics noise variance.

In the desired limit when multiplied shot noise dominates electronics noise, the Q-factor for high-signal extinction ratios ($s_0 \ll s_1$) approaches[5]

$$\begin{aligned}Q &= \frac{(R_T)\cdot SM_{APD}P_1}{\sigma_{elec}+\sqrt{\sigma^2_{shot,APD}+\sigma^2_{elec}}}\\ &\xrightarrow{\sigma^2_{shot,APD}\gg\sigma^2_{elec}} \sqrt{\frac{\eta\overline{n}R}{F_{APD}B_e}} \sim \sqrt{2\overline{n}/F_{APD}}\end{aligned} \tag{11}$$

[5] Note that while Eq. (11) reveals general trends, care has to be taken with quantitative predictions, since the Gaussian assumption of detection statistics breaks down for shot-noise limited direct detection: Specializing equation (11) for *pin*-reception ($F_{APD} = 1$), we arrive at a quantum limit of $\overline{n} =$ 18 photons/bit, which is off its correct value by 2.6 dB.

with $P_1 = 2\overline{P}$ equal to the '1'-bit optical signal power for NRZ-OOK. Thus, the excess noise factor F_{APD} takes the role of a noise figure in degrading detection performance. Note that optimum performance of an APD receiver is in general *not* attained at the highest possible multiplication M_{APD}, since F_{APD} is a complicated and highly technology-dependent function of M_{APD}, necessitating joint optimization of M_{APD}, F_{APD}, and σ^2_{elec} [40,43,46].

Good APDs ($M_{APD} \sim 100, F_{APD} \sim 5$) for operation up to 1 Gbit/s are available in Silicon technology, which limits their operating range to wavelengths below $\sim$ 1100 nm. Receiver sensitivities of 200 photons/bit have been achieved at 50 Mbit/s [47]. InGaAs or InAlAs-based APDs for use in the 1550-nm wavelength region, however, exhibit fairly low multiplication ($M_{APD} \sim 10$) for 10 Gbit/s detection. Receiver sensitivities of 1000 photons/bit have been demonstrated at 10 Gbit/s [48].

3.5. Coherent Receiver

Another way of amplifying the signal and boosting the accompanying noise above the electronics noise floor is known as *coherent detection* [9,35,41,49]. A coherent receiver, as depicted in Fig. 20 a, combines the signal with a *local oscillator* (LO) laser by means of an optical coupler. Upon detection, the two fields beat against each other, and the average electrical signal reads

$$\begin{aligned}\langle s(t)\rangle = (R_T)\cdot S\{&\varepsilon P_s(t) + (1-\varepsilon)P_{LO}\\ &+ 2\mu\sqrt{\varepsilon(1-\varepsilon)}\sqrt{P_s(t)P_{LO}}\cos(2\pi f_{IF}t + \phi_s(t))\}\end{aligned} \tag{12}$$

where $P_s(t)$ and $\phi(t)$ denote the modulation-carrying received signal's power and phase, respectively, P_{LO} stands for the LO power, and f_{IF}, the beat frequency between signal and LO, is called intermediate frequency, since the IF signal is usually mixed down to baseband after photodetection, using standard microwave techniques. The parameter ε captures the splitting ratio of the optical coupler, which has to be chosen as high as possible such as not to waste too much signal power, and as low as acceptable to let sufficient LO power reach the detector to achieve shot-noise limited performance (see explanation below). The heterodyne efficiency μ accounts for the degree of spatial overlap as well as for the polarization match between LO field and signal field. If both LO and signal

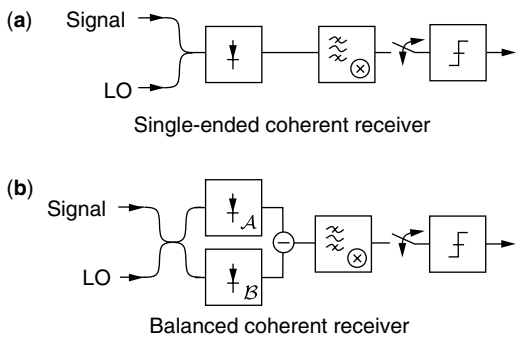


Figure 20. Single-ended (a) and balanced (b) coherent receiver. After photodetection, various kinds of electronic signal-processing can be performed.

are provided copolarized in single-mode optical fibers, μ equals unity.

If the frequency of the LO differs from the signal frequency, we speak of a *heterodyne* receiver. If LO and signal have the same frequency, such that $f_{IF} = 0$, we speak of a *homodyne* receiver. Homodyne detection strictly requires optical phase locking between the LO and signal optical fields, which implies significant technological effort.[6] In the desired range of operation, the LO power is chosen much stronger than the signal power $[(1-\varepsilon)P_{LO} \gg \varepsilon P_s(t)]$, and the first term in Eq. (12) can be neglected compared to the second and third. Filtering out the temporally constant second term in Eq. (12), we are then left with an exact replica of the received *optical field's amplitude* $\sqrt{P_s(t)}$ and *phase* $\phi_s(t)$ (as compared to the optical *power* accessible in direct-detection receivers). Thus, any amplitude or phase modulation scheme can directly be employed in combination with coherent receivers.

Due to the high LO power reaching the detector, the main noise contribution in a coherent receiver is the shot noise produced by the LO, $\sigma^2_{LO-shot} = 2eS(1-\varepsilon)P_{LO}B_e$. If this noise term dominates electronics noise ($\sigma^2_{LO-shot} \gg \sigma^2_{elec}$), optimum receiver performance is achieved. This limit is known as the *shot noise limit* in the context of coherent receivers. The highest receiver sensitivity with the potential of practical implementation that is known today can be achieved using homodyne detection of phase shift keying[7] (PSK), where the data bits are directly mapped onto the phase $\phi_s(t)$ of the optical signal, $\{0, 1\} \rightarrow \{0, \pi\}$. Without going into the derivations [9], the quantum limit for homodyne PSK can be shown to equal only 9 photons/bit, with a reported receiver sensitivity record of 20 photons/bit at 565 Mbit/s [51]. Using OOK instead of PSK, the sensitivity degrades by 3 dB. Going to heterodyne detection results in an additional loss of 3 dB in terms of receiver sensitivity. A detailed discussion of quantum limits for coherent receivers can be found in [9].

An alternative implementation of coherent receivers is shown in Fig. 20b. It makes use of *balanced detection* with $\varepsilon = 1/2$. While a balanced coherent receiver has *exactly* the same quantum-limited sensitivity as its single-ended equivalent, it offers the advantage of utilizing the full optical signal and LO power, and of being more robust to LO relative intensity noise (RIN).

Although still seriously considered for inter-satellite link applications due to the high achievable sensitivities, the interest in coherent receivers has vanished for fiber-optic systems with the availability of Erbium-doped fiber amplifiers (EDFA) in the early 1990s. To understand this evolution, let us look at the main advantages of coherent receivers, and how they have become outdated:

- Receiver sensitivities of coherent receivers by far outperform those achieved with pin-receivers and APDs, thus allowing for increased transmission distances in unamplified optical links. BUT: Optically preamplified receivers (cf. Section 3.6) exhibit similar receiver sensitivities to coherent receivers, are polarization-insensitive, and take less serious hits in performance if inline amplification is present.
- The possibility of correcting for chromatic dispersion in the microwave regime is offered in coherent detection, since both amplitude *and* phase of the optical field are converted to an electronic signal. BUT: Efficient and adaptive broadband phase corrections can only be performed on RF *bandpass* signals, asking for heterodyne detection. Since f_{IF} has to be chosen about 3 times the data rate [41], unrealistically high receiver front-end bandwidths would be needed for the high data rates used today. Conversely, all-optical dispersion compensators and adaptive optical filters are quickly advancing technologies [52], allowing for efficient phase corrections in the *optical* regime.
- Coherent receivers allow for the separation of closely spaced WDM channels by means of RF bandpass filters with sharp roll-offs. BUT: Optical filter technology has advanced dramatically, thus enabling channel spacings on the order of 10 GHz with sharp optical filter roll-offs, which opens up the possibility of *optical* channel filtering even for ultradense WDM applications.

But even if coherent reception is highly unlikely to reenter the high-speed fiber-optical communications market, the technique is far from being history: With coherent receiver terminals in their final product development phases [53], coherent receivers will soon find applications in nonfiber optical communications scenarios, most notably in free-space optical communications, where the enormous link distances of up to 80,000 km for geostationary intersatellite links ask for utmost receiver performance, and make homodyne PSK an attractive candidate.

3.6. Optically Preamplified Receiver

The historically youngest class of highly sensitive optical receivers uses *optical preamplification* to boost the weak received signal to appreciable optical power levels prior to detection, as shown in Fig. 21. At the same time, and fundamentally unavoidable, amplified spontaneous emission (ASE) with a power spectral density per (spatial and polarization) mode of

$$N_{ASE} = hfGF/2 \tag{13}$$

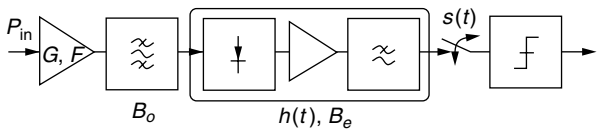


Figure 21. An optically preamplified receiver uses an optical amplifier in combination with an optical bandpass filter prior to detection with a *pin*-receiver.

[6] Using square-law detection of the electrical signal, one can also build a quasi-homodyne receiver without phase-locking and $f_{IF} \approx 0$ [50].

[7] Because the LO provides an optical phase reference, true PSK can be used instead of DPSK in direct detection receivers.

is introduced by the amplification process [43,54,55]; G and F denote the optical amplifier's gain and noise figure,[8] respectively. Today's high-performance Erbium-doped fiber amplifiers (EDFAs) exhibit gains between 15 dB and 30 dB, while having noise figures closely approaching their fundamental limit of 3 dB. These intriguing performance characteristics of EDFAs have made optically preamplified receivers the most important detection technique known today. Since the gain spectrum (and thus also the ASE spectrum) is much broader than the signal spectrum (typ. 30 nm in the 1550-nm wavelength band), an optical bandpass filter is employed to suppress out-of-band ASE. Upon detection, the random ASE field beats against the signal field (as well as against itself), leading to *signal-ASE beat noise* and *ASE-ASE beat* noise[9] after electrical filtering [42,16,17],

$$\sigma^2_{\mathrm{s-ASE}}(t) = (R_T^2)\cdot 2S^2 N_{ASE}\mathrm{Re}\left\{\iint_{-\infty}^{\infty} e(\tau)e^* \times(\tilde{\tau})r_n(\tau-\tilde{\tau})h(t-\tau)h(t-\tilde{\tau})\,d\tau d\tilde{\tau}\right\} \quad (14)$$

and

$$\sigma^2_{\mathrm{ASE-ASE}} = (R_T^2)\cdot M_{pol}S^2N_{ASE}^2\int_{-\infty}^{\infty}|r_n(\tau)|^2 r_h(\tau)\,d\tau \quad (15)$$

where $r_n(\tau)=\langle n^*(\tau)n(\tilde{\tau})\rangle$ is the autocorrelation of the optically filtered ASE field $n(t)$, and $r_h=\int h(\tau)h(\tau-t)\,d\tau$ is the autocorrelation of the detection electronics. The number of ASE modes reaching the detector is denoted M_{pol}. In a single-mode fiber system, M_{pol} usually equals 2, since polarization filtering typically is not done. The optically filtered signal field is denoted $e(t)$. In the limit of rectangular filters and constant input power P_{in} to the optical preamplifier, the above relations simplify to [56]

$$\sigma^2_{\mathrm{s-ASE}} \approx 4S^2GP_{in}N_{ASE}B_e \quad (16)$$

and

$$\sigma^2_{\mathrm{ASE-ASE}} \approx M_{pol}S^2N_{ASE}^2B_e(2B_o-B_e) \quad (17)$$

where B_o stands for the optical filter bandwidth. From Eqs. (14) through (17), we see that the signal-ASE beat noise variance is *nonstationary*, grows linearly with signal power, and is independent of B_o, as long as the optical filter does not significantly influence the signal spectrum. Also, we see that the ASE-ASE beat noise variance is *stationary*, grows linearly with B_o, and linearly depends on M_{pol}. Owing to the linear dependence of N_{ASE} on G (cf. Eq. (13)), both beat noise standard deviations as well as the detected electrical signal amplitude SGP_{in} grow linearly with G. Thus, the Q-factor becomes independent of G as soon as the beat noise terms starts to dominate electronics noise. This limit is called *optical noise limit* or *beat noise limit*, and forms the usual operating condition of optically preamplified receivers. Idealizing the beat noise limit ($F=3$ dB, $B_e=R/2$), we arrive at a quantum limit of 38 photons/bit for OOK, and of 20 photons/bit for DPSK [9]. Experimentally, sensitivities of 43 photons/bit at 5 Gbit/s [57] (52 photons/bit at 10 Gbit/s [58]) have been achieved for OOK, and 30 photons/bit at 10 Gbit/s [59] (45 photons/bit at 40 Gbit/s [36]) have been demonstrated for DPSK using balanced detection (cf. Fig. 16).

3.7. Bandwidth Optimization

Once the gain of the optical amplifier is chosen high enough to let optical beat noise dominate electronics noise, the main impact on receiver performance comes from optical and electrical filter characteristics. In the frame of a single-pulse theory (i.e., neglecting ISI), optimum receiver performance is achieved if the *optical* filter is matched to the data pulses,[10] and if the *electrical* filter is made sufficiently broadband to have no influence on the detected signal [41]. By constructing a receiver of this type, the same performance could in principle be achieved for NRZ and RZ signaling formats. However, the electrical bandwidth is usually upper-bounded by technological constraints and cost considerations, especially at data rates in the multi-Gbit/s regime, and ISI often *does* have a significant impact on detection. Thus, other than matched filter characteristics frequently turn out to be superior in practice, and the equality of NRZ and RZ formats is eliminated: NRZ formats usually take a noticeable hit in performance with respect to RZ '0'-bit ISI (cf. Figs. 4 and 23).

Figure 22 [60,61] shows a typical dependence of receiver performance on B_o and B_e for NRZ-OOK and 33%-duty cycle RZ-OOK. The contours give the dB-penalty relative to the quantum limit. RZ can be seen to have a better optimum sensitivity than NRZ, and, by the wider spacing of the contour lines, to be more tolerant to suboptimum choices of optical and electrical filters. For both formats, the performance decrease at larger-than-optimum filter bandwidths can be attributed to increased detection noise. At lower-than-optimum filter bandwidths, RZ is affected by pulse amplitude reductions due to filtering, while NRZ is predominantly affected by ISI. The different degrading effects for NRZ and RZ at low bandwidths become evident in Fig. 23, showing the dB-sensitivity penalty relative to the quantum limit for RZ and NRZ as a function of electrical filter bandwidth for fixed $B_o=2R$ (left) and as a function of optical filter bandwidth for fixed $B_e=0.8R$ (right). The solid curves apply for a pseudo-random bit sequence (PRBS) of length 2^7-1, and are section lines of Fig. 22. The dashed curves represent the results obtained for a single optical pulse, thus eliminating the effect of

[8] It has become common to call F a *noise figure*, although this terminology is somewhat sloppy, since it only considers the influence of signal-ASE beat noise [55].

[9] Note that the "beating"-picture is only correct in the frame of a classical consideration. Using quantum mechanical reasoning, the signal-ASE beat noise turns out to result from the fact that each photon is amplified by a random, integer number within the amplifier, similar to the random multiplication of electron-hole pairs in APDs [54].

[10] The impulse response of a *matched filter* is identical to the temporally reversed pulse to be detected [11].

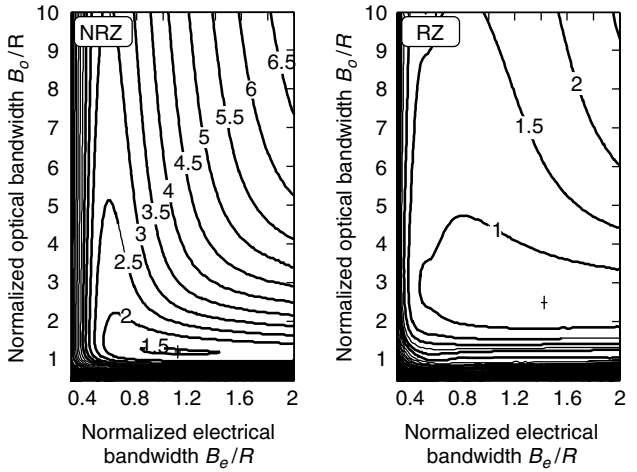


Figure 22. Dependence of receiver sensitivity on optical and electrical filter bandwidths for NRZ-OOK and 33%-duty cycle RZ-OOK. The contours are labelled in terms of dB-penalties relative to the quantum limit [60,61].

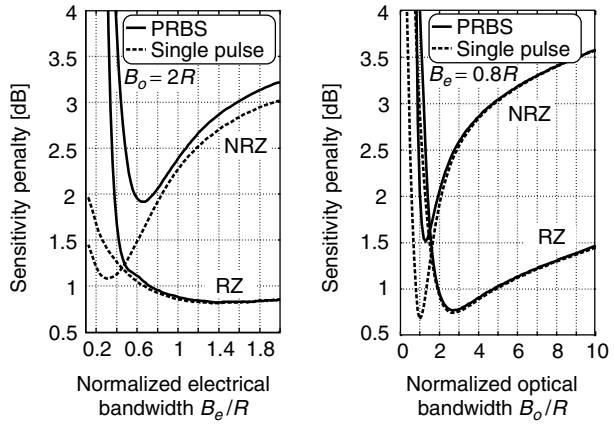


Figure 23. Sensitivity penalty to the quantum limit for NRZ and RZ as a function of electrical filter bandwidth (left) and optical filter bandwidth (right). Solid lines apply to ISI-corrupted detection, while dashed curves represent the ISI-free case [61].

ISI. For RZ, the ISI-free curves and the PRBS-curves run in parallel until well below the optimum bandwidth constellations, indicating that the RZ bandwidth optima are *not* influenced by ISI. For NRZ, however, the two curves depart for $B_e \lesssim 0.8R$ and $B_o \lesssim 1.5R$, which clearly shows that the optimum NRZ-receiver bandwidths are determined by trading ISI against detection noise.

In addition to the above bandwidth considerations, when optimizing operational, cost-effective receivers for WDM systems, one further has to take into account the effects of WDM channel crosstalk, optical source frequency offsets and drifts, filter concatenations effects due to a large number of optical add/drop multiplexers, as well as technological constraints on high-speed receiver bandwidths and receiver imperfections, such as jitter of the sampling phase.

3.8. Required Optical Signal-to-Noise Ratio (OSNR)

Specifying an optical receiver in terms of its *receiver sensitivity* dates back to pre-EDFA times, when the ultimate limit to fiber-optic link distances was given by the lowest possible receive power at which a specified receiver performance could still be guaranteed. With the deployment of in-line optical amplifiers this situation has changed, and optical signals can be transmitted over much longer distances through periodic optical reamplification. Since each amplifier fundamentally introduces ASE according to Eq. (13), it is now the *total* ASE N_{tot} accumulated along the transmission line per polarization mode rather than the received signal power level that sets limits on the maximum transmission distance, and the ability of a receiver to cope with ASE determines its performance in a system.

Figure 24 visualizes the situation of beat-noise limited detection in an in-line amplified transmission system. It shows the receiving end of a transmission line carrying a WDM signal with average per-channel power $\overline{P}_s^{\lambda_i}$, onto which the total ASE accumulated along the line is added. A WDM demultiplexer simultaneously acts to separate the WDM channels and to suppress out-of-band ASE. Comparing Fig. 24 to Fig. 21, we notice equivalence with

$$\overline{P}_s^{\lambda_i} \longleftrightarrow GP_{in} \quad \text{and} \quad N_{ASE} \longleftrightarrow N_{tot} \tag{18}$$

These substitutions are most conveniently captured in the definition of the *optical signal-to-noise ratio* (OSNR) as the ratio of the average optical signal power $\overline{P}_s^{\lambda_i}$ to the (unpolarized) ASE power within some reference bandwidth B_{ref},

$$\text{OSNR} = \frac{\overline{P}_s^{\lambda_i}}{2N_{tot}B_{ref}} \tag{19}$$

The bandwidth B_{ref} is typically (but not exclusively) chosen to be 0.1 nm at a wavelength of 1550 nm, that is, $B_{ref} \approx 12.5$ GHz. Using the relations (18), the OSNR required at a beat-noise limited receiver to attain a certain BER can be directly related to the input sensitivity of a preamplified receiver that is operated with a 'clean' input signal $\overline{P}_{in}$ (more precisely, with an input signal satisfying $GN_{in} \ll N_{ASE}$, where N_{in} denotes the ASE power spectral density at the optical amplifier input). It thus makes sense to also define the *quantum limited OSNR* as the minimum OSNR an ideal beat-noise limited direct detection receiver has to have at its input to produce a certain BER, for example, BER $= 10^{-9}$. The quantum-limited OSNR is then connected to the quantum limit of an optically preamplified receiver by

$$\text{OSNR} = \frac{\overline{n}R}{2B_{ref}}. \tag{20}$$

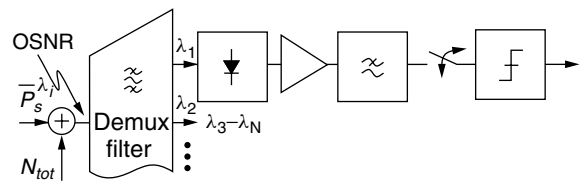


Figure 24. Beat-noise limited detection of optical WDM signals corrupted by noise accumulated along the transmission line through optical in-line amplification.

3.9. Photonic Integrated Receiver

A newly emerging class of optical receivers with a high potential for deployment is called *photonic integrated receiver* [62]. It is basically an optically preamplified receiver *without* optical bandpass filtering, and typically consists of a *pin*-photodiode following an on-chip integrated semiconductor optical amplifier (SOA). This receiver type is used to boost the optical signal power prior to detection to improve upon receiver performance, thus eliminating the need for an external optical preamplifier. Due to the absence of optical filtering, the bandwidth of the ASE generated by the SOA is solely determined by the amplifier's gain bandwidth, letting the ASE-ASE beat noise reach appreciable values. However, depending on the relationship of the ASE-ASE beat noise due to ASE from the SOA to the other receiver noise terms, significant improvements in receiver performance can be achieved. For example, if the signal-ASE beat noise originating from ASE produced along the transmission line is well above the ASE-ASE beat noise produced by the SOA, no receiver degradation will be noticed.

4. SUMMARY

In this article we discussed modulation formats with the potential of being used in high-speed fiber-optic communications. We distinguished between amplitude-modulated and phase-modulated formats, discussed the role of the number of signaling levels, and showed how the optical spectrum can be influenced to achieve high-spectral efficiency. We outlined optical receiver concepts, gave an introduction to their performance evaluation by means of frequently used performance measures, and discussed important receiver design trade-offs.

Acknowledgments
I would like to acknowledge many valuable discussions on modulation formats and receiver design with my present and former colleagues René-Jean Essiambre, Jake Bromage, Alan H. Gnauck, S. Chandrasekhar, Hoon Kim, Herwig Kogelnik, Martin Zirngibl, Klaus H. Kudielka, A. Kalmar, Martin M. Strasser, Martin Pfennigbauer, Martin Pauer, and Walter R. Leeb.

BIOGRAPHY

Peter J. Winzer was born in Vienna, Austria, in 1973. He studied electrical engineering/communications engineering at the Vienna University of Technology, and received his Dipl.-Ing. (M.S.) and Dr.techn. (Ph.D.) degrees in 1996 and 1998, respectively. His work, largely supported by the European Space Agency (ESA), was related to the analysis and modeling of noise in Doppler wind lidar and space-borne optical communication systems. Following his assistant professorship at the Vienna University of Technology, Dr. Winzer joined Bell Laboratories in Holmdel, New Jersey, in 2000, where he has since been working on fiber-optic communications, with an emphasis on Raman amplification, 40-Gbit/s optical transmitter and receiver optimization, and spectrally efficient optical modulation formats. Dr. Winzer has authored and co-authored some 60 papers and holds several patents. His present areas of interest include transmission and reception aspects in both fiber-optic and free-space optical communication systems.

BIBLIOGRAPHY

1. Record transmission distances and capacities are reported in the Post-Deadline Sessions of the annual conferences *Optical Fiber Communication* (OFC) and *European Conference on Optical Communication* (ECOC).
2. I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IV B*. Academic Press, 2002.
3. *Proc. Free-Space Laser Communication Technologies I* (1988) through *XIV* (2002), Proc. SPIE vols. 0885, 1218, 1417, 1635, 1866, 2123, 2381, 2699, 2990, 3266, 3615, 3932, 4272, and 4635; D. L. Begley, *Selected Papers on Free-Space Laser Communications I* (1991) and *II* (1994), Proc. SPIE vols. MS30 and MS100.
4. V. W. S. Chan, Optical space communications, *IEEE J. Selec. Top. Quantum Electron*. **6**: 959–975 (2000).
5. P. J. Winzer and W. R. Leeb, Space-borne optical communications — a challenging reality, *Proc. 15th Annual Meeting of the IEEE Lasers and Electro-Optics Society* (LEOS'02), 2002.
6. X. Lu and O. Sniezko, The evolution of cable TV networks, in [2], (2002).
7. A. S. Siddiqui et al., Dispersion-tolerant transmission using a duobinary polarization-shift keying transmission scheme, *IEEE Photon. Technol. Lett.* **14**: 158–160 (2002).
8. H. Kogelnik et al., Polarization-mode dispersion, in [2], (2002).
9. G. Jacobsen, *Noise in Digital Optical Transmission Systems*, Artech House, 1994.
10. J. Conradi, Bandwidth-efficient modulation formats for digital fiber transmission systems, in [2], (2002).
11. R. D. Gitlin et al., *Data Communications Principles*, Plenum Press, 1992.
12. S. Bigo et al., 10.2 Tbit/s (256×42.7 Gbit/s PDM/WDM) transmission over 100 km TeraLight^{TM} fiber with 1.28 bit/s/Hz spectral efficiency, *Proc. Optical fiber communication Conference* (OFC'01), paper PD25, (2001); W. Idler et al., Vestigial side band demultiplexing for ultra high capacity (0.64 bit/s/Hz) transmission of 128×40 Gb/s channels, *Proc. Optical fiber communication Conference* (OFC'01), paper MM3, 2001; Y. Frignac et al., Transmission of 256 wavelength-division and polarization-division-multiplexed channels at 42.7 Gb/s (10.2 Tb/s capacity) over 3×100 km of TeraLight (*TM*) fiber, *Proc. Optical fiber communication Conference* (OFC'02), paper FC5, 2002.
13. D. A. Ackerman et al., Telecommunication lasers, in I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IVA*, Academic Press, 2002.
14. A. Ougazzaden et al., 40Gb/s tandem electro-absorption modulator, *Proc. Optical fiber communication Conference* (OFC'01), paper PD14, 2001.
15. H. Kim and A. H. Gnauck, Chirp characteristics of dual-drive Mach-Zehnder modulator with a finite DC extinction ratio, *IEEE Photon. Technol. Lett.* **14**: 298–300 (2002).
16. L. Boivin and G. J. Pendock, Receiver sensitivity for optically amplified RZ signals with arbitrary duty cycle, *Proc. Optical*

Amplifiers and their Applications (OAA'99), paper ThB4, 106–109, (1999).

17. P. J. Winzer and A. Kalmar, Sensitivity enhancement of optical receivers by impulsive coding, *J. Lightwave Technol.* **17**: 171–177 (1999).
18. M. Suzuki et al., Transform-limited 14 ps optical pulse generation with 15 GHz repetition rate by InGaAsP electroabsorption modulator, *Electron. Lett.* **28**: 1007–1008 (1992).
19. R. -J. Essiambre, B. Mikkelsen, and G. Raybon, Pseudo-linear transmission of high-speed TDM signals: 40 and 160 Gb/s, in [2], (2002).
20. P. B. Hansen et al., 5.5-mm long InGaAsP monolithic extended-cavity laser with an integrated Bragg-reflector for active mode-locking, *IEEE Photon. Technol. Lett.* **4**: 215–217 (1992).
21. N. M. Froberg et al., Generation of 12.5Gbit/s soliton data stream with an integrated laser-modulator transmitter, *Electron. Lett.* **30**: 1880–1881 (1994).
22. J. J. Veselka et al., A soliton transmitter using a CW laser and an NRZ driven Mach-Zehnder modulator, *IEEE Photon. Technol. Lett.* **8**: 950–952 (1996).
23. P. J. Winzer and J. Leuthold, Return-to-zero modulator using a single NRZ drive signal and an optical delay interferometer, *IEEE Photon. Technol. Lett.* **13**: 1298–1300 (2001).
24. A. Lender, The duobinary technique for high-speed data transmission, *IEEE Trans. on Commun. Electronics* **82**: 214–218 (1963).
25. D. Penninckx et al., The phase-shaped binary transmission (PSBT): A new technique to transmit far beyond the chromatic dispersion limit, *IEEE Photon. Technol. Lett.* **9**: 259–261 (1997); D. Penninckx et al., Relation between spectrum bandwidth and the effects of chromatic dispersion in optical transmissions, *Electron. Lett.* **32**: 1023–1024 (1996).
26. J. B. Stark, J. E. Mazo, and R. Laroia, Phased amplitude-shift signaling (PASS) codes: Increasing the spectral efficiency of DWDM transmission, *Proc. European Conf. on Optical Communication* (ECOC'98): 373–374, 1998; J. B. Stark, J. E. Mazo, and R. Laroia, Line coding for dispersion tolerance and spectral efficiency: Duobinary and beyond, *Proc. Optical Fiber Communication Conference* (OFC'99), paper WM46, 1999.
27. T. Ono et al., Characteristics of optical duobinary signals in Terabit/s capacity, high-spectral efficiency WDM systems, *J. Lightwave Technol.* **16**: 788–797 (1998).
28. K. S. Cheng and J. Conradi, Reduction of pulse-to-pulse interaction using alternative RZ formats in 40-Gb/s systems, *IEEE Photon. Technol. Lett.* **14**: 98–100 (2002).
29. T. Franck, T. N. Nielsen, and A. Stentz, Experimental verification of SBS suppression by duobinary modulation, *Proc. European Conf. on Optical Communication* (ECOC'97): 71–74, (1997).
30. X. Wei et al., 40 Gb/s duobinary and modified duobinary transmitter based on an optical delay interferometer, *Proc. European Conf. on Optical Communication* (ECOC'02): paper 09.6.3, 2002.
31. Y. Miyamoto et al., S-band 3×120-km DSF transmission of 8×42.7-Gbit/s DWDM duobinary-carrier-suppressed RZ signals generated by novel wideband PM/AM conversion, *Proc. Optical Amplifiers and their Applications* (OAA'01), paper PD6, 2001.
32. N. S. Bergano, Undersea communication systems, in [2] (2002).
33. C. R. Menyuk et al., Dispersion managed solitons and chirped RZ: What is the difference?, in [2] (2002).
34. R. A. Griffin et al., Integrated 10 Gb/s chirped return-to-zero transmitter using GaAs/AlGaAs modulators, *Proc. Optical Fiber Commun. Conf.* (OFC'01), paper PD15, 2001.
35. G. Einarsson, *Principles of Lightwave Communications*, John Wiley & Sons, 1996.
36. A. H. Gnauck et al., 2.5 Tb/s (64×42.7 Gb/s) transmission over 40×100 km NZDSF using RZ-DPSK format and all-Raman-amplified spans, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FC2, 2002.
37. T. Chikama et al., Modulation and demodulation techniques in optical heterodyne PSK transmission systems, *J. Lightwave Technol.* **8**: 309–321 (1990).
38. R. A. Griffin and A. C. Carter, Optical differential quadrature phase-shift key (oDQPSK) for high capacity optical transmission, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper WX6, 2002; R. A. Griffin et al., 10Gb/s optical differential quadrature phase shift key (DQPSK) transmission using GaAs/AlGaAs integration, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FD6, 2002.
39. S. D. Personick, Receiver design for digital fiber optic communication systems, I, *Bell Syst. Tech. J.* **52**: 843–874 (1973).
40. G. P. Agrawal, *Fiber-Optic Communication Systems*, 3rd edition, John Wiley & Sons, 2002.
41. L. Kazovsky, S. Benedetto, and A. Willner, *Optical Fiber Communication Systems*, Artech House, 1996.
42. P. J. Winzer, Receiver noise modeling in the presence of optical amplification, *Proc. Optical Amplifiers and their Applications* (OAA'01), paper OTuE16, 2001; P. J. Winzer, Performance estimation of receivers corrupted by optical noise, in J. D. Minelly, and Y. Nakano, eds., *OSA Trends in Optics and Photonics* (TOPS) vol. 60, N. Jolley, 268–273, 2001.
43. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, John Wiley & Sons, 1991.
44. B. E. A. Saleh, *Photoelectron Statistics*, Springer-Verlag Berlin Heidelberg, New York, 1978.
45. L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, 1995.
46. S. D. Personick, Receiver design, in S. E. Miller and A. G. Chynoweth (eds.), *Optical Fiber Telecommunications* Academic Press, 1979; B. L. Kasper, Receiver design, in S. E. Miller and I. P. Kaminow (eds.), *Optical Fiber Telecommunications II*, Academic Press, 1988; K. Ogawa et al., I. P. Kaminow and T. L. Koch (eds.), Advances in high bit-rate transmission systems, in *Optical Fiber Telecommunications IIIA*, Academic Press, 1997; B. L. Kasper, O. Mizuhara, and Y. -K. Chen, High bit-rate receivers, transmitters, and electronics, in I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IVA*, Academic Press, 2002; T. V. Muoi, Receiver design for high-speed optical-fiber systems, *J. Lightwave Technol.* **2**: 243–267 (1984). J. N. Hollenhurst, Fundamental limits on optical pulse detection and digital communication, *J. Lightwave Technol.* **13**: 1135–1145 (1995); S. B. Alexander, *Optical Communication Receiver Design*, SPIE tutorial texts in Optical Engineering, vol. TT22, 1997.
47. G. Planche et al., SILEX final ground testing and in-flight performance assessment, *Proc. SPIE* **3615**: 64–77 (1999).

48. K. Sato et al., Record highest sensitivity of -28 dBm at 10 Gb/s achieved by newly developed extremely-compact superlattice-APD module with TIA-IC, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FB11, 2002.

49. S. Betti, G. De Marchis, and E. Iannone, *Coherent optical communication systems*, Wiley-Interscience, 1995; S. Ryu, *Coherent Lightwave Communication Systems*, Artech House, 1995.

50. L. G. Kazovsky, P. Meissner, and E. Patzak, ASK multiport optical homodyne receivers, *J. Lightwave Technol.* **5**: 770–790 (1987).

51. B. Wandernoth, 20 photon/bit 565 Mbit/s PSK homodyne receiver using synchronisation bits, *Electron. Lett.* **28**: 387–388 (1992).

52. C. R. Doerr, Planar lightwave devices for WDM, in I. Kaminow and T. Li (eds.), *Optical Fiber Telecommunications IVA*, Academic Press, 2002.

53. K. Kudielka and K. Pribil, Transparent Optical Intersatellite Link Using Double-Sideband Modulation and Homodyne Reception, *Int. J. Electron. Commun.* (AE), **56**: 254–260 (2002).

54. E. Desurvire, *Erbium-Doped Fiber Amplifiers*, John Wiley & Sons, 1994.

55. H. A. Haus, *Electromagnetic Noise and Quantum Optical Measurements*, Springer Verlag, 2000.

56. N. A. Olsson, Lightwave systems with optical amplifiers, *J. Lightwave Technol.* **7**: 1071–1082 (1989).

57. D. O. Caplan, and W. A. Atia, A quantum limited optically-matchedcommunication link, *Proc. Optical Fiber Communication Conference* (OFC01), paper MM2, 2001.

58. M. M. Strasser, M. Pfennigbauer, M. Pauer, and P. J. Winzer, Experimental verification of optimum filter bandwidth in direct-detection (N)RZ receivers limited by optical noise, *Proc. LEOS 2001 Annual Meeting* (LEOS'01): 485–486, 2001.

59. W. Atia and R. S. Bondurant, Demonstration of return-to-zero signaling in both OOK and DPSK formats to improve receiver sensitivity in an optically preamplified receiver, *Proc. 12th annual meeting of LEOS*: 2244–225, 1999.

60. P. J. Winzer et al., Optimum bandwidths for optically preamplified RZ and NRZ reception, *J. Lightwave Technol.* **9**: 1263–1273 (2001).

61. M. Pfennigbauer et al., Performance optimization of optically preamplified receivers for return-to-zero and non return-to-zero coding, *Int. J. Electron. Commun.* (AE) **56**: 261–268 (2002).

62. B. Mason et al., 40Gb/s photonic integrated receiver with -17dBm sensitivity, *Proc. Optical Fiber Commun. Conf.* (OFC'02), paper FB10, 2002.

OPTICAL TRANSPORT SYSTEM ENGINEERING

Milorad Cvijetic
NEC America
Herndon, Virginia

1. INTRODUCTION

In the time officially deemed as "the information era," we are witnessing the insatiable demand for high information capacity and distance-independent connectivity. Optical networking has been the most efficient solution in satisfying this ongoing demand for bandwidth and connectivity. Optical fiber has been laid down all the way to the curb, building, home, and desk. In general, all optical networks can be considered as part of a global optical network; they are all owned either by private enterprises or by telecommunication carriers.

Several logical parts in a global optical network can be identified, as illustrated in Fig. 1:

- *The core optical network*, which is a long-haul network interconnecting big cities or major communication hubs. Connections between big cities on different continents are made by submarine optical cables. The *core network* is a generic name, but very often we refer to the core network as a wide-area network (WAN) if it belongs to an enterprise, or as the interchange network if it is operated by a telecommunication carrier.
- *The edge optical network*, which covers a smaller geographical area, usually a metropolitan area. Again, we can refer to the edge network either as a metropolitan-area network (MAN) if it belongs to an enterprise, or as a local exchange network if it is operated by telecommunication carriers.
- *The access optical network*, which is the part of the network related to last-mile access and bandwidth distribution to individual end users (in corporate, government, medical, entertainment, scientific, and private sectors). Both the enterprise local-area networks (LANs) and the distribution part of the carrier network connecting the central office with individual users belong to the access network.

The physical network topology that best supports traffic demand is generally different for different parts of a global optical network, as presented in Fig. 1. It could vary between mesh, ring, or star topology. In spite of different network topologies, the main consideration of optical transport engineering is always an optical (lightwave) path, since an optical network is just the means of supporting an end-to-end connection via the lightwave path. Optical transport engineering is related to the physical layer of an optical network, and takes into account the optical signal propagation length, characteristics of optical elements used (fibers, lasers, amplifiers etc.), modulation bit rate, networking impact and transmission requirements.

In this article we will introduce fundamentals of optical transport system engineering. Although an optical signal can take on either a digital or an analog form, our focus will be on digital signals, since the majority of modern applications are related to digital signal transmission.

2. OPTICAL TRANSMISSION PARAMETERS

The simplest optical transmission system is a point-to-point connection on a single optical wavelength, which propagates through an optical fiber. An upgrade to this is the deployment of WDM (wavelength-division multiplex) technology, where multiple optical wavelengths are

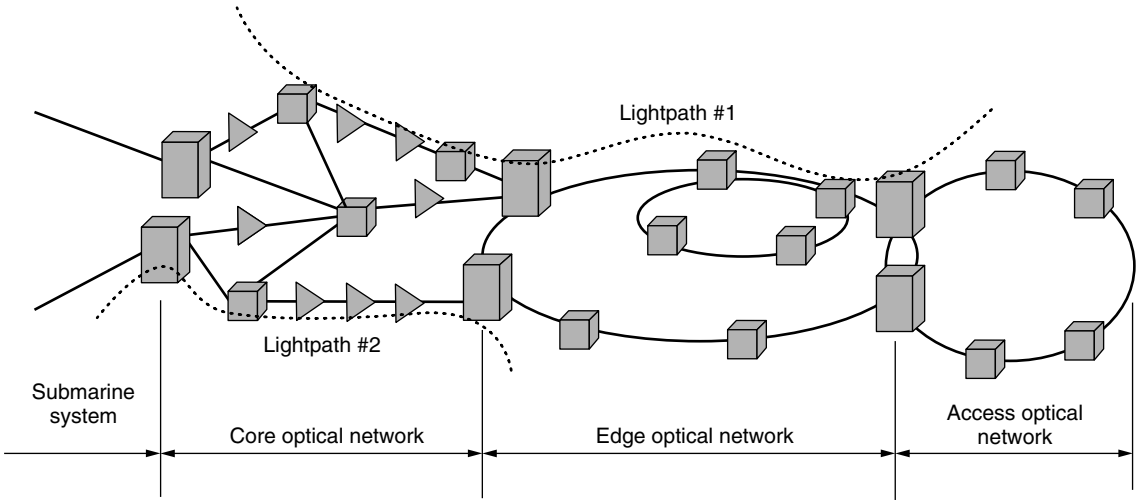


Figure 1. Optical networking structure.

combined to travel over the same physical path. WDM technology originally served to increase the bandwidth of already installed fiber, but it has quickly become the foundation of optical networking by combining optical signal transport over arbitrary distances with wavelength routing and optical protection.

The general scheme of an optical transport system is shown in Fig. 2. Several optical channels, carrying independent modulation signals, have been multiplexed by WDM technology, and sent to the optical fiber line. The aggregated signal is then transported over some distance before it is demultiplexed and detected (converted back to an electrical level). The optical signal transmission path can include a number of optical amplifiers, crossconnects, and optical add/drop multiplexers. The illustrated set of parameters, related either to enabling technologies or to transmission and networking issues, can be attached to Fig. 2.

Providing stable and reliable operation of an optical transport system over time requires proper design and engineering. Optical transport systems engineering involves accounting for all effects that can alter an optical signal on its way from the source (laser) on through photodetection by photodiode, and then to the threshold decision point. Different impairments will degrade and compromise the integrity of the signal before it arrives to the decision point to be recovered from corruptive additives (noise, crosstalk, and interference). The transmission quality is measured by the received signal-to-noise-ratio (SNR), which is defined as the ratio of the signal level to the noise level at the threshold point. The other parameter used to measure signal quality is the bit error rate (BER). BER is interrelated with SNR and defines the probability that a signal space (or a logic 0) will be mistaken for a signal mark (a logic 1), and vice versa. The main goal in optical signal transport is to achieve the required BER between end-to-end users, or between two specified points. Evaluating the BER requires determining the received signal level at the threshold point, calculating the noise power, and quantifying and including the influence of various relevant impairments.

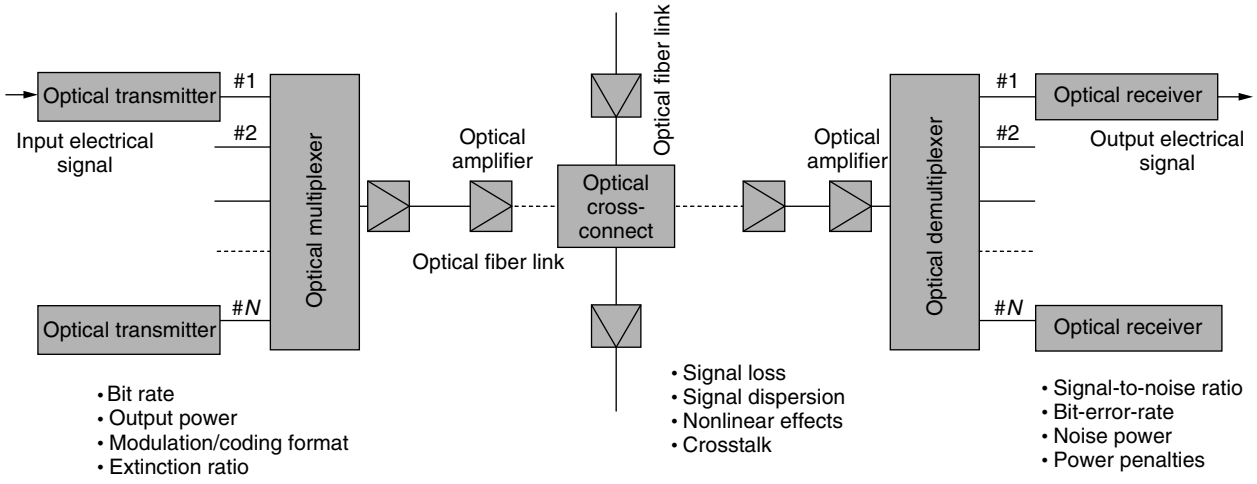


Figure 2. Optical DWDM system.

2.1. Optical Signal Path

The end-to-end signal path from Fig. 1 includes both the electrical and optical path portions. Conversion from the electrical to the optical level is done in the optical transmitter, while conversion from the optical level to an electrical signal takes place in the optical receiver. The key elements on an optical signal path are

1. *Semiconductor lasers* that convert an electrical signal to optical radiation. The bias current flows through the laser p-n junction and stimulates the recombination of electrons and holes, producing photons. If the current is higher than a certain threshold, recombination occurs in an organized way, with strong correlation in phase, frequency, and direction of radiated photons that form the output optical signal (called "stimulated emission of radiation.") Semiconductor lasers could either be directly modulated by an electrical signal or simply biased by a DC voltage and operate in combination with an external optical modulator. Each laser generates a specified optical wavelength, but some spectral linewidth is associated with the generated optical signal as well. These lasers are known as single-mode lasers (SMLs), characterized by a distinguished single longitudinal mode in the optical spectrum. If a set of separated longitudinal modes can be recognized under the optical spectrum envelope, the lasers are called *multimode lasers* (MMLs).
2. *Optical fibers* that transport an optical signal to its destination. The combination of low signal loss and extremely wide transmission bandwidth allows high-speed optical signals to be transmitted over long distances before regeneration becomes necessary. There are two groups of optical fibers. The first group, called *multimode optical fibers*, transfer light through multiple spatial or transversal modes. Each mode, defined through a specified combination of electric and magnetic field components, occupies a different cross-sectional area of the optical fiber core, and takes a slightly distinguished path along the optical fiber. The difference in mode pathlengths causes a difference in arrival times at the receiving point. This phenomenon is known as multimode dispersion and causes signal distortion and limitations in transmission bandwidth. The second group of optical fibers effectively eliminates multimode dispersion by limiting the number of modes to just one through a much smaller core diameter. These fibers, called *single-mode optical fibers*, do however introduce another signal impairment known as *chromatic dispersion*. Chromatic dispersion is caused by a difference in velocities among different wavelength components within the same pulse. There are several methods to minimize chromatic dispersion at a specified wavelength, involving either the introduction of new single-mode optical fibers, or the utilization of different dispersion compensation methods.
3. *Optical amplifiers* that amplify weak incoming optical signals through the process of stimulated emission, without conversion back to the electrical level. Optical amplifiers should provide enough gain to amplify a specified number of optical channels. There are different types of optical amplifiers currently in use, such as semiconductor optical amplifiers (SOAs), erbium-doped fiber amplifiers (EDFAs), or Raman amplifiers. Amplifier parameters are gain, gain flatness over amplification bandwidth, output power, bandwidth, and noise power. The noise generated in an optical amplifier occurs due to a spontaneous emission process that is not correlated with the signal. All amplifiers degrade the SNR of the output signal because of amplified spontaneous emission (ASE) that adds itself to the signal during its amplification. SNR degradation is measured by the noise figure. Optical amplifiers can take several positions along the optical path, as indicated in Fig. 2. The output power could be enhanced by a booster amplifier within a transmitter, on the transmission line (inline amplifier), or before the receiver (to act as a preamplifier to increase the receiver sensitivity).
4. *Photodiodes* that convert an incoming optical signal back to the electrical level through a process just opposite to the one that takes place in lasers. Photodiodes can be classified into PIN or avalanche photodiodes (APD). The process within the PIN photodiodes is characterized by *quantum efficiency*, which is the probability that each photon will generate an electron–hole pair. In the avalanche photodiode each primary electron–hole pair is accelerated in a strong electric field, which can cause the generation of several secondary electron-hole pairs through the effect of impact ionization. This process is random in nature and avalanchelike.

A more elaborate analysis of advanced optical transport systems can be found in the bibliography [1–4].

2.2. Optical Signal Parameters

There are a set of parameters along the lightwave path that determines the received signal power:

- *Output power* from the laser/modulator coupled to the fiber pigtail. Optical power is defined per individual wavelength and depends on the lasers/modulators used. The output optical power is usually expressed in decibels per milliwatt (dB_m), defined as $dB_m = 10\log(P)$, where the output power P is expressed in milliwatts.
- *The extinction ratio*, which is the ratio between the optical power related to a logic 1 (mark) to the power related to a logic 0 (space). By increasing the extinction ratio, the signal to noise ratio is increased as well, but at the cost of additional penalties in the modulation speed and laser chirp.
- *Optical amplifier gain*, which determines the level of an optical signal that is being amplified. Optical amplifier gain is correlated to noise parameters,

which means that higher gain will generate more noise and vice versa.

- *Photodiode responsivity*, which defines the ratio between the number of electrical carriers produced and the number of incoming photons.

2.3. Noise Parameters

The total noise under consideration in optical transport system engineering is generated along the lightwave path and during the photodetection process. There are some additive noise components (the noise components remaining even if the signal is not present), and some multiplicative noise components (which are produced only if the signal is present). The additive noise components are

- *Dark-current noise* generated in photodiodes due to the thermal process.
- *Amplified spontaneous emission* (ASE) noise generated by any optical amplifier along the lightwave path.
- *Crosstalk*, which occurs in multichannel systems. Components introducing crosstalk in WDM systems are optical filters, optical multiplexers and demultiplexers, optical switches, semiconductor optical amplifiers, and optical fibers through nonlinearities. Crosstalk can either be intrachannel (occurs when another signal of the same wavelength interferes with the signal in question), or interchannel (occurs when some portion of a neighboring channel has been spread out, and detected by the specified signal's receiver).
- *Thermal noise*, which is created in the resistive part of the input impedance of an electrical preamplifier that follows the photodiode. The noise created in the electronic amplification stages following the preamplification process is thermal noise in nature as well.

The multiplicative noise components are

- *Avalanche shot noise*, caused by the random nature of the amplification of primary electron–hole pairs through the effect of impact ionization in avalanche photodiodes.
- *Laser intensity noise*, which occurs as a result of microvariations in the laser output power intensity. This noise is characterized through the relative intensity noise (RIN) parameter, and is more relevant for analog transmission systems.
- *Laser phase noise*, which is related to microvariations in phase of generated photons. The output optical signal, as a collection of individual photons, exhibits finite nonzero spectral width.
- *Modal noise*, which arises in multimode fibers through the random process of excitation of transversal modes.

2.4. Impairment Parameters

Impairment parameters relevant to optical transport system engineering are either optical power related or optical wavelength related. They can also be constant or time-dependent. Each of them results in a *signal power penalty*, which means that a higher signal power is required at the receiver to keep the BER at a level that would exist if the impairment were negligible.

Optical power-related impairments are

- *Optical fiber attenuation*, or fiber loss, which is the ratio between the output and input optical power at the defined optical fiber section. Optical attenuation is characterized by an attenuation coefficient α, usually expressed in decibels per kilometer. The decibel is defined as $\mathrm{dB} = 10\log(P_2/P_1)$, where P_2 and P_1 are the output and input power respectively.
- *Insertion losses* in different optical components along the lightwave path, such as optical connectors, optical splices, optical couplers, optical multiplexers, and optical filters. These losses are sometimes added to the optical fiber loss and are considered together.

Impairments that are optical power related, but are also functions of time are

- *Polarization mode dispersion* (PMD), a stochastic process that appears in real fibers, caused by variations in the shape of their core along the fiber length. The light in an optical fiber can be considered as a superposition of two polarized components. If they travel at the same speed, no polarization dispersion occurs, but if they travel at different speeds, as in real fibers, the light will separate into its faster and slower components, leading to a difference in propagation of the two polarization states and to pulse spreading. Both mechanical stresses and temperature effects contribute to PMD.
- *Polarization-dependent loss* (PDL), which is similar in nature to PMD, but this time the difference between polarization states is in transmission losses, rather than in arrival times. These differential losses accumulate in the system, since there might be many components having polarization-dependent loss. Since polarization fluctuates with time, the SNR at the end of the lightwave path will fluctuate as well, causing a power penalty.

Impairments that are dependent on both the optical power and optical wavelengths are

- *Four-wave mixing* (FWM), a nonlinear effect where a new optical frequency is generated when three frequencies mutually interact. It causes crosstalk noise in channels, since the newly generated optical frequency might coincide with one of the original channels.
- *Stimulated Raman scattering* (SRS), a nonlinear effect that occurs when a propagating optical power interacts with glass molecules in the fiber undergoing a wavelength shift. The result is a power transfer from one wavelength to another, which causes crosstalk between channels.

Impairments that are dependent on optical wavelength, but are also functions of time are

- *Chromatic dispersion* caused by the dependence of the fiber refractive index on the wavelength. Since a laser is not an ideal monochromatic source, each pulse in its time domain contains different spectral components that travel at different velocities through an optical fiber. Chromatic dispersion induces pulse broadening when the neighboring pulses cross their allotted time slot borders, which can severely limit system transmission rates. Chromatic dispersion is also a cumulative effect that increases with optical fiber length.
- *Laser chirp*, which is the modulation of optical frequency (or wavelength) when the optical signal, is intensity-modulated by a specific electrical waveform. The change in the frequency causes laser spectral linewidth broadening and, in the interaction with chromatic dispersion, leads to optical pulse distortion and the intersymbol interference effect [see Eq. (4)]. Chirp can be reduced by decreasing the extinction ratio, but this decrease would introduce additional power penalties, requiring some compromise [5].

Finally, there are some impairments that are dependent on both the optical power and optical wavelength, and are also functions of time:

- *Stimulated Brillouin scattering* (SBS), a nonlinear effect that occurs when a high optical power reflects off the grating formed by acoustic vibrations, downshifts in optical frequency, and comes back. The SBS can cause signal attenuation if the launched power is higher than a certain threshold. Optical signal dithering with low frequencies helps to increase the SBS threshold and effectively suppress SBS effect.
- *The self-phase modulation* (SPM) *effect*, which results from the fact that a higher fiber refractive index causes wavelengths at the center of the pulse to accumulate phase more quickly than at the wings. This stretches the wavelengths at the leading edge (called "red shift") of the pulse and compresses wavelengths at the trailing edge ("blue shift"). This phase modulation effect broadens the spectrum, which causes pulse spreading. If combined with positive dispersion in the optical fiber under controlled conditions, it can lead to suppression of the chromatic dispersion effect. This is the basis for soliton transmission, where return-to-zero (RZ) soliton pulses propagate over very long distances without optoelectronic regeneration.
- *Cross-phase modulation* (XPM), which has the same nature as SPM, occurs following the interaction between multiple optical frequencies.

3. ASSESSMENT OF THE OPTICAL TRANSPORT LIMITATIONS AND PENALTIES

3.1. Attenuation

A silica-based optical fiber is the central point of an optical signal transmission, offering wider available bandwidth, lower signal attenuation, and smaller signal distortion than other wired physical media. The output power P_2 from an optical fiber can be calculated from the input power P_1 and the optical attenuation coefficient α. For the lightwave path with length L, $P_2 = P_1 \exp(-\alpha L)$. If parameters P_2, P_1, and α are expressed in decibels, then the relation becomes $P_2 = P_1 - \alpha L$.

Four low-attenuation bands can be recognized within the usable optical bandwidth of silica-based optical fibers. They are usually referred as U, S, C, and L bands, although this nomenclature is not standardized yet. The C band occupies wavelengths from 1530 to 1560 nm, while the L band includes wavelengths between 1580 and 1610 nm. Both these bands have been considered as the most suitable bands for high-channel-count WDM transmission. The S band (sometimes called the S+ band) and U band (sometimes referred to as the S− band) cover shorter wavelengths down to approximately 1230 nm, where optical fiber attenuation is slightly higher than in the wavelength region covered by the C and L bands.

3.2. Noise

Detected photocurrent is the sum of signal and noise contributions after the photodetection process has taken place. It can be expressed as $I_p = I + i_s + i_{\text{th}}$, where I is the signal current calculated as a product of incoming optical power P and photodiode responsivity R (R is expressed in amperes per watt). Noise components, expressed by currents $i_s + i_{\text{th}}$, correspond to the quantum (shot) and thermal noise, respectively. The power of the total noise that appears after photodetection is equal to the product of the sum of noise spectral density components and the noise electric bandwidth Δf. In the case where direct detection without optical preamplification takes place, the total noise power can be expressed as

$$\langle i^2 \rangle_{tot} = \left[2qM^2F(M)I + \frac{4kT}{R_L} \right] \Delta f \tag{1}$$

where the first term in brackets describes the quantum noise, while the second is related to the thermal noise generated at a load resistance R_L. In the previous equation q represents the electron charge ($q = 1.6 \times 10^{-19}$ C), M is the avalanche amplification factor, $F(M)$ is the avalanche excess noise factor, k is Boltzmann's constant ($k = 1.38 \times 10^{-23}$ J/K), and T is absolute temperature in kelvins. If the PIN photodiode is used, the factor $M^2F(M)$ becomes unity.

In case an optical amplifier precedes the photodiode the major noise contribution comes from ASE noise. The spectral density of ASE noise is

$$S_{\text{sp}} = \frac{(G-1)N_f h\nu}{2} \tag{2}$$

where G is amplifier gain, N_f is the amplifier noise figure, h is Planck's constant ($h = 6.63 \times 10^{-34}$ J/Hz), and ν is optical frequency in hertz. The total noise power in this case becomes

$$\langle i^2 \rangle_{\text{tot}} = 2qR[GP + S_{\text{sp}}B_{\text{op}}]\Delta f + 4R^2 \text{GPS}_{\text{sp}}\Delta f + 2R^2 S_{\text{sp}}^2 [2B_{\text{op}} - \Delta f]\Delta f + \frac{4kT}{R_L}\Delta f \tag{3}$$

Parameter B_{op} refers to the optical filter bandwidth. A standard deviation $\sigma = [\langle i^2 \rangle_{tot}]^{1/2}$ is usually used in signal-to-noise ratio and bit-error-rate calculations [see Eq. (12)].

3.3. Chromatic Dispersion

Recall that chromatic dispersion is the cause of pulse spreading and the occurrence of intersymbol interference. Pulse spreading is proportional to the fiber dispersion parameter D, expressed in picoseconds per nanometer and kilometer (ps/nm.km). The dispersion parameter is an ascending linearlike, wavelength-dependent function, characterized by its zero value cross-point and a dispersion slope. There are several fiber types that differ in their dispersion parameter profile [6]:

- *Standard single-mode fibers* (SMFs), where the parameter D has zero value at the 1310 nm wavelength and the dispersion slope of approximately 0.072 ps/km.nm^2. With this, the chromatic dispersion parameter reaches the value of 17–20 ps/nm.km in the wavelength region belonging to C and L bands.
- *Dispersion-shifted fiber* (DSF), where the parameter D has zero value at the 1550 nm wavelength and the dispersion slope of approximately 0.09 ps/km.nm^2. The chromatic dispersion parameter can take on both negative and positive values in the wavelength region belonging to C and L bands. This fiber type has been good for single-wavelength transmission, but is not suitable for WDM applications because of high penalties due to nonlinear effects.
- *Non-zero dispersion-shifted fibers* (NZDSF), where the parameter D has zero value shifted from the 1550 nm wavelength and the dispersion slope of approximately 0.03 ps/km.nm^2. The chromatic dispersion parameter has some minimal value in the wavelength region belonging to C and L bands, thus minimizing penalties due to nonlinear effects.

The influence of chromatic dispersion and the penalties related to it can be evaluated by assuming that the pulse spreading due to dispersion should be less than a fraction δ of the bit period T. For a 1-dB power penalty, $\delta = 0.306$; for a 2-dB penalty, $\delta = 0.491$. For a signal having a bit rate $B = 1/T$ and spectral linewidth $\Delta\lambda$, and transmitted over a distance L, this condition can be expressed as

$$\begin{aligned} \Delta\lambda |D| LB &< \delta \qquad \text{for direct modulation} \\ B\lambda [|D| L/2\pi c]^{1/2} &< \delta \qquad \text{for an external modulation} \end{aligned} \tag{4}$$

The influence of chromatic dispersion is a critical factor for higher bit rates and longer distances and should be suppressed by a proper dispersion compensation scheme. The dispersion compensation process is based on the following observation. While in single-mode optical fiber longer wavelengths impose more delay than shorter wavelengths, the dispersion compensating modules do just the opposite. As a result, signal delays over a specified wavelength band have been equalized. As for dispersion compensating modules, using dispersion compensation fibers (DCFs) with a negative dispersion coefficient is the most common method for dispersion compensation. Since there is an insertion loss introduced by DCF, the figure of merit, defined as the ratio of the absolute amount of dispersion divided by the insertion loss, is used to characterize DCF. Generally it is good if the figure of merit is better than 150 ps/nm/dB.

Optical fiber Bragg gratings can be used for chromatic dispersion compensation as well. The grating reflects different wavelengths at different points along its length, introducing different delays at different wavelengths. Delay introduced by the length of 10 cm is approximately 100 ps. Dispersion is inversely proportional to the bandwidth; that is, a large dispersion occurs over smaller bandwidth and vice versa. For example, 1000 ps/nm occurs over a 1 nm bandwidth, while 100 ps/nm occurs over a 10 nm bandwidth. Future applications, however, will require adaptive dispersion compensation modules that allow for adjustment of both the dispersion compensation value and the dispersion slope.

3.4. Polarization Mode Dispersion

Polarization mode dispersion (PMD) is characterized by two coefficients, D_{p1} and D_{p2}, reflecting so-called "first- and second-order" polarization mode dispersion, respectively. The extent of pulse broadening D_t is governed by the following relation:

$$D_t = D_{p1} L^{1/2} + D_{p2} L \tag{5}$$

where the coefficient D_{p1} presents the average differential group delay (DGD) between the two orthogonal states of polarization over length L, while D_{p2} measures the wavelength dependence of PMD. The contribution of D_{p2} is much smaller than the contribution of D_{p1}, and very often just the first term in Eq. (5) is considered. The value of the coefficient D_{p1} can vary from 0.01 ps/km$^{1/2}$ for new optical fibers to over 1 ps/km$^{1/2}$ for older fibers.

PMD is a stochastic process described by the Maxwellian distribution, which complicates the process of its control and compensation. The probability that actual delay will be 3 times larger than the average delay calculated by Eq. (5) is 4×10^{-5}. This is why we correlate the average delay expressed by Eq. (5) to the actual delay equal to three times the average delay. For differential delay equal to $0.3T$, the power penalty due to PMD will be less than 1 dB.

3.5. Nonlinear Effects

Nonlinear effects in an optical fiber are neither design nor manufacturing defects, but can occur regardless and can cause severe transmission impairments (unexpected loss and interference in the network). On the other hand, in some cases, they may be used to improve transmission characteristics (such as in soliton transmission).

Nonlinear effects are cumulative in nature and proportional to the lightwave pathlength L. Since signal power decreases with increasing lightwave pathlength, an effective length L_{eff} has been introduced to help with calculations. The effective length is defined as

$$L_{eff} = \frac{1 - \exp(-\alpha l)}{\alpha} M \tag{6}$$

where M is the number of fiber spans, each of length l. (Recall that one span is the distance between two amplifiers, therefore $l = L/M$.) In the wavelength region around 1.55 μm, and for links where $L > 1/\alpha$, L_{eff} is ~20 km (α is ~0.046 km^{-1}, or 0.2 dB/km). From the previous relation, it is clear that the effective length can be reduced by increasing the span length and by decreasing the number of amplifiers on the line. But what matters most is the product of the power launched from the amplifier, P, and the effective length L_{eff}. If amplifier spacing is increased, the launched power needs to be increased as well to compensate for additional fiber losses. This increase will be exponential: $P = \exp(\alpha l)$. Since the product increases with span length l, reducing the amplifier spacing can reduce the effect of nonlinearities.

The effects of nonlinearity are inversely proportional to the area of the fiber core. It is convenient to use an effective core area A_{eff} since the power is not uniformly distributed within the core section. This effective area is about 50 μm for a single-mode fiber with a core diameter of 8 μm, but for a dispersion compensating fiber (DCF) it is smaller (thus DCF tends to exhibit higher nonlinearities).

Nonlinear effects can be divided into two categories: the effects due to variations in the fiber refractive index, and the effects due to light scattering. Agraval has given a more detailed explanation of nonlinear effects [7].

Variations in the refractive index at high signal power are at the root of nonlinear effects classified as refractive-index phenomena: self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). At low optical powers, an optical fiber's refractive index n is pretty constant [i.e., it is $n = n_1(\lambda)$ for specified wavelength λ]. Higher optical powers, however, cause a refractive index change as follows:

$$n(\lambda, E) = n_1(\lambda) + \frac{n_2 P}{A_{\text{eff}}} \tag{7}$$

where n_2 is the nonlinear refractive index ($n_2 \sim 3 \times 10^{-8}$ μm^2/W), and P is the optical signal power. Both SPM and XPM affect the optical signal phase in proportion to the nonlinear part of the refractive index and generate spectral broadening. Spectral broadening, in combination with chromatic dispersion, will contribute to signal distortion.

In *four-wave mixing* (FWM), new optical frequencies $\nu_{ijk} = \nu_i + \nu_j - \nu_k$ are generated whenever three wavelengths with frequencies ν_I, ν_j, and ν_k propagate through the fiber. The power of a resultant new wave is calculated as presented by Shibata et al. [8]

$$P_{ijk} = \frac{\alpha^2}{\alpha^2 + \Delta\beta^2}\left[1 + \frac{4\exp(-\alpha l)\sin^2(\Delta\beta l/2)}{[1 - \exp(-\alpha l)]}\right] \times \left(\frac{2\pi \nu_{ijk} n_2 d_{ijk}}{3cA_{\text{eff}}}\right)^2 P_i P_j P_k L_{\text{eff}}^2 \tag{8}$$

where $P_{ijk}(i, j, k = 1 \cdots N)$ is the power of the generated wave, n_2 is the nonlinear refractive index, and d_{ijk} is the so-called "degeneracy" factor. The value $\Delta\beta = \beta_i + \beta_j - \beta_k - \beta_{ijk}$ defines a phase condition or relationship among the propagation constants of the optical waves involved (a propagation constant is defined as $\beta = 2\pi n\lambda/c$, where n is the refractive index, λ is the wavelength, and c is the speed of light in a vacuum).

The total crosstalk due to FWM in a given channel is the sum of all generated waves according to Eq. (8) and can be analyzed as interchannel crosstalk [see Eq. (17)]. To alleviate the penalty introduced by FWM, the following measures could be taken: using unequal channel spacing, increasing channel spacing, using dispersion, or reducing the power of interacting channels. The most effective means is to use some amount of dispersion; this clarifies the need to shift the zero dispersion point from the 1550-nm-wavelength region.

Nonlinear effects that occur due to light scattering include *simulated Raman scattering* (SRS) and *stimulated Brillouin scattering* (SBS). For SBS, the acoustic phonons are involved in an interaction that occurs over a very narrow linewidth $\Delta\nu_{\text{SBS}}$($\Delta\nu_{\text{SBS}} = 20$ MHz at 1.55 μm). There is no such interaction if channel spacing is greater than 20 MHz. The SBS process depletes the signal and creates a strong backward signal if the incident power per channel is higher than some threshold value P_{th} expressed as

$$P_{\text{th}} = \frac{21 b A_{\text{eff}}}{g_B L_{\text{eff}}}\left(1 + \frac{\Delta\nu_L}{\Delta\nu_{\text{SBS}}}\right) \tag{9}$$

where g_B is the SBS gain coefficient equal to approximately 4×10^{-11} m/W, $\Delta\nu_L$ is the laser linewidth, while parameter b takes on a value between 1 and 2 depending on relative polarization of pump and Stokes waves. The worst case leads to $P_{\text{th}} \sim 1.3$ mW, since $\Delta\nu_L$ is approximately 20 MHz. The SBS penalty can be reduced by either keeping the power per channel below the SBS threshold or broadening the linewidth of the source using signal dithering. This method is commonly deployed in high-bit-rate systems.

Stimulated Raman scattering is a broadband effect, and its gain coefficient is a function of wavelength spacing. The gain coefficient peak is $g_R \sim 6 \times 10^{-14}$ m/W, which is much smaller than the gain coefficient peak for SBS. Channels up to 125 nm apart will be coupled by SRS, possibly in both directions. SRS coupling occurs only if both channels are at a logic 1 at that moment. The fraction of the power leaking from a particular channel to all other channels is given by

$$P_{\text{SRS}} = \frac{g_R \Delta\lambda_s P L_{\text{eff}} N(N-1)}{4\Delta\lambda_c A_{\text{eff}}} \tag{10}$$

where $\Delta\lambda_s \sim 125$ nm, $\Delta\lambda_c$ is the optical channel spacing, P is the power per channel, and N is the number of channels [7]. The SRS effect is reduced by dispersion, since different channels travel with different velocities and the probability of an overlap between pulses at different wavelengths is reduced. The penalty introduced by SRS can be alleviated by proper channels spacing and/or postequalization of optical channel powers. Some special techniques, such as polarization interleaving between the neighboring optical channels, can help as well.

4. OPTICAL TRANSPORT SYSTEM ENGINEERING

4.1. BER, Signal-to-Noise Ratio, and the *Q* Factor

The bit error rate (BER) is the most important parameter for measuring a digital signal transmission quality. It is

defined as

$$\text{BER}\,(Q) = \frac{1}{2\pi}\int_{Q}^{\infty} e^{-y^2/2}\,dy \approx \frac{1}{Q\sqrt{2\pi}} e^{-Q^2/2} \tag{11}$$

The so-called Q factor, introduced above, corresponds to the electrical signal-to-noise ratio:

$$Q = \frac{R(P_1 - P_0)}{\sigma_1 + \sigma_0} = \text{SNR} \tag{12}$$

where P_0 is the optical power during a "space bit," P_1 is the optical output power during a "mark bit," R is the responsivity of the photodiode, while σ_1 and σ_0 are standard deviations of the noise current during the 1 and 0 bits, respectively. The following practical values are mutually related: $\text{BER} = 10^{-15}$ with $Q = 8$, $\text{BER} = 10^{-12}$ with $Q = 7$, and $\text{BER} = 10^{-9}$ with $Q = 6$. In optical transport systems with cascades of optical amplifiers along the lightwave path, the following important relation between the Q factor and an optical signal-to-noise ratio can be established:

$$\text{OSNR} = \frac{P_1 + P_0}{4S_{\text{sp}}B_{\text{opt}}} \approx \frac{2Q^2\Delta f}{B_{\text{op}}} \tag{13}$$

Equations (12) and (13) are the basic ones since the impact of various impairments is not included.

4.2. Power Penalty Handling

If there are impairments involved, signal, noise-related values will be Q', P_0', P_1', σ_1', and σ_0', rather than Q, P_0, P_1, σ_1 and σ_0 respectively. Each impairment will contribute to a power penalty to the transport system. The total optical power penalty can be calculated as

$$\Delta P = -10\log\left(\frac{Q'}{Q}\right) \tag{14}$$

The biggest contribution to the total power penalty comes from the nonideal extinction ratio, imperfect dispersion compensation, nonlinear effects, and crosstalk:

- The power penalty due to a nonfinite extinction ratio r, defined as $r = P_1/P_0$, can be calculated in decibels as
$$\Delta P_{\text{ER}} = 10\log\left[\frac{r+1}{r-1}\right] \tag{15}$$
- The power penalty due intrachannel crosstalk involving N interfering signals is
$$\Delta P_{\text{intra}} = C\log\left(1 - 2\sum_{i=1}^{N}\sqrt{\delta_i}\right) \tag{16}$$
where the coefficient C takes on the value 10 for direct detection, and the value 5 forAPD/preamp detection, while δ_i is the crosstalk portion divided by the power of specified channel signal. If we allow a 1-dB crosstalk penalty, then the intrachannel crosstalk level should be just 1%, or 20 dB, below the specified channel signal.
- The power penalty due interchannel crosstalk involving N interfering signals is
$$\Delta P_{\text{inter}} = C\log\left(1 - \sum_{i=1}^{N}\delta_i\right) \tag{17}$$
with the same coefficient C as in relation (17). If we allow a 1-dB crosstalk penalty, then the intrachannel crosstalk level should be 13.5 dB below the desired signal. A more thorough treatment of crosstalk can be found in the article by Zhou et al. [9].
- As for power penalties due to imperfect dispersion compensation or nonlinear effects (FWM and SRS), Eq. (17) can be applied. The portion δ for a particular case can be calculated by Eqs. (4), (8), and (10).

4.3. Noise Accumulation

Recall that in an optical transport system a lightpath contains a number of optical amplifiers spaced l km apart. The length l defines the span length. If fiber attenuation is α, the span loss between two amplifiers is $\alpha_{\text{span}} = \exp(-\alpha l)$. Each optical amplifier amplifies an incoming optical signal to compensate for the loss at the previous span. At the same time, however, it generates some spontaneous emission noise. Both the signal and the spontaneous emission noise are then amplified by the following optical amplifiers.

If the gain G of an optical amplifier is larger than the span loss α_{span} the signal power will increase gradually throughout the amplifier chain. However, the output power from an optical amplifier is physically limited to a saturated value P_{sat}, which means that as input power increases the amplifier gain drops. Consequently, after some number of spans, amplifiers will enter into the saturation regime and the total gain will drop from its initial value G to a saturated value G_{sat}. Further along the lightwave path, a spatial steady-state condition will be reached, in which both the saturated output power P_{sat} and the gain G_{sat} remain the same from span to span. If there are N optical channels, the saturated output power will be equally divided among them. Therefore, the output power per channel will be $P_{\text{out}} = P_{\text{sat}}/N$.

The OSNR gradually decreases along the chain, since the accumulated ASE noise gradually makes up a more significant portion of the limited total power from an amplifier. The steady-state gain will be slightly smaller than the span loss, due to added noise at each amplifier point. Thus, the best engineering approach is to choose a saturated gain that is very close to the span loss. If we prescribe the OSNR for a lightwave path with total length L, and M amplifiers on the line ($M = L/l$), the following relation can be established:

$$\text{OSNR} = P_{\text{out}} - \alpha l - \Delta P - 10\log(M) - 10\log(F_n h\nu B_{\text{op}}) \tag{18}$$

where all except the last two terms are expressed in decibels. If we did not need to worry about impairments, we would neglect the power penalty term and either maximize

the power or decrease the span length to increase the OSNR in Eq. (18). However, the story is different when impairments cannot be neglected, since a power margin equal to ΔP needs to be allocated in advance to compensate for impairment power penalties.

5. OPTICAL TRANSPORT ENABLING TECHNOLOGIES AND TRADEOFFS

5.1. Enabling Technologies

Enabling technologies will continue to provide the means of increasing both transmission capacity and lightwave pathlength. There is a number of enabling technologies that are helpful in resolving the before-mentioned issues in optical transport systems, and in approaching a transmission capacity predicted by Mecozzi and Shtaif [10]. These include optical amplifiers, forward error correction, advanced coding techniques, and advanced dispersion compensators.

Optical amplifiers should provide enough gain for a specified number of optical channels, which suggests that an aggregate optical power should be >22 dB for systems with more than 100 optical channels. Next, the noise figure should approach its theoretical value of 3 dB. In addition, the gain profile should be equalized along the entire wavelength band, and this gain equalization should be dynamically adjustable.

Fiber doped optical amplifiers, such as erbium-doped fiber amplifiers (EDFAs) and thulium-doped fiber amplifiers (TDFAs) can serve to cover the entire low-loss optical fiber bandwidth. In addition, Raman amplifiers can be used to cover a wide range of wavelengths as well. The Raman based amplification is a newer technology based on the SRS nonlinear effect. In the Raman amplifier, the pump light is launched into the fiber at inline amplifier sites (or optical receiver sites), opposite the signal direction. As a result, the forward-propagating optical signals get some energy, and a low-noise preamplifier has been created. By combining several pumps, a fairly flat gain profile over a wide range of optical wavelengths can be achieved.

Advanced dispersion-compensating techniques are necessary to take full advantage of the improved optical amplifiers. First, the chromatic dispersion-compensating device (DCM) should not only incorporate dispersion compensation ability but slope compensation as well. Secondly, polarization mode dispersion compensation will be needed more often, but an efficient PMD compensator still needs to be introduced.

Forward error correction (FEC) is needed to push systems reach even further in future optical transport networks. For now, so-called Reed–Solomon 239–255 coding scheme remains widely used, but some other coding methods have been introduced as well. The final result of FEC application is BER enhancement for a lower signal-to-noise ratio.

Advanced modulation/transmission methods are emerging for use in high-speed optical transport systems. Some of them are

- *Solitons* or return-to-zero distortionless pulses, where spectral disorder at the trailing and leading edges that occurs due to self-phase modulation is corrected by another disorder caused by chromatic dispersion. The technique has shown very promising results when used in combination with fibers having a "prescribed dispersion map."
- *Advanced coding*, such as duobinary coding, where special filtering reduces the spectral width. Thus, the slowest and the fastest components are eliminated; the rest is more resistive to chromatic dispersion influence.
- *Coherent detection* where the weak input optical signal is mixed with a much stronger signal from the local laser. With this method the signal current takes on the value $I = 2RP_sP_{\mathrm{LO}}\cos[|\omega_s - \omega_L|t + \phi(t)]$, where R, P_s, and P_{LO} relate to receiver responsivity, incoming optical power, and local laser power, respectively. Because of P_{LO} contribution, the coherent receiver sensitivity is considerably enhanced. The detection process can either be heterodyne or homodyne. In the heterodyne detection scheme, the incoming optical signal frequency ω_s and the frequency ω_L of the local laser are slightly different, while in homodyne detection scheme both the optical frequencies and the phases of the signal and the local laser are completely matched. The demodulation process applied to the signal current depends on the modulation format of the incoming optical signal; this format could be amplitude shift keying (ASK), frequency shift keying (FSK), or phase shift keying (PSK). The realization of the coherent detection scheme involves the necessity of having a stable relationship between the phases and frequencies of an incoming optical signal and the local laser signal, which brings an additional complexity to the system design.

5.2. Transmission System Engineering Tradeoffs

Proper design of a transmission system is a big challenge, and some tradeoffs must be done. Generally speaking, the longer the distance, the smaller the transmission capacity, and vice versa. Overall design tradeoffs include

- *Fiber type selection* is applicable just for new fiber deployment. It is obvious that single-channel transmission favors DSF fibers, while NZDSF fiber should be, generally speaking, favorable for long-distance DWDM systems. Standard single-mode fibers (SMFs) might be the best choice in certain cases where either chromatic dispersion is not critical, or where SMF-based systems are less vulnerable to the influence of nonlinearities. Even low-cost multimode optical fibers can be considered for some short-reach and/or lower-bit-rate applications.
- *Spectral efficiency* versus the total optical bandwidth occupied is an important design issue in some cases. By increasing spectral efficiency with dense wavelength spacing, designers expose the system to greater susceptibility to degradation from nonlinear

effects. On the other hand, increasing the optical bandwidth occupied would lead to system cost increase.

- *Chromatic dispersion management* is essential for high-speed optical transmission systems. A novel transmission line design with proper dispersion management will provide conditions for ultra-high-capacity systems.
- *Optical power level* per optical channel is dependent on the output power level from optical amplifiers, nonlinear effects, crosstalk, and safety issues. It is always desirable to increase the power level from the SNR point of view, but the power penalty due to an increase in nonlinearity and amplifier noise will limit the signal power level, leading to an optimal in-between value.
- *Optical pathlength* is very important in transmission systems engineering. Design is more complex on longer optical paths since optical networking issues, such as crosstalk, wavelength misalignment, and cascaded filter effects, accumulate with increasing pathlength. Optical signal powers and the SNR among different paths that come together at the input of an optical amplifier or receiver should be equalized.

BIOGRAPHY

Milorad Cvijetic received his Ph.D. degree in electrical engineering from University of Belgrade in 1984. Dr. Cvijetic has experience in both academia (teaching at University of Belgrade and Carleton University), and industry (work in the area of high-speed optical transmission systems and optical networks). His research work related to quasi-single mode optical fibers, BER evaluation in soliton-based systems, and system performance evaluation in high-speed optical systems with external modulation, has been widely recognized.

He currently serves as the chief technology strategist for Optical Network Products with NEC America, Herndon, Virginia. Previously, he has been with Bell Northern Research (later NORTEL Technologies) in Ottawa, Canada, working in the Advanced Technology Laboratory. Dr. Cvijetic has published more than 40 technical papers and two books titled *Digital Optical Communications* and *Coherent and Nonlinear Lightwave Communications*. He has taken part in numerous telecommunications conferences and symposiums, in some as a session/conference chairman, technical committee member, or invited speaker. He is a member of IEEE Communications Society and LEOS.

BIBLIOGRAPHY

1. I. P. Kaminov and T. L. Koch, eds., *Optical Fiber Telecommunications*, Academic Press, San Diego, CA, 1997.
2. J. A. Buck, *Fundamentals of Optical Fibers*, Wiley, New York, 1994.
3. E. Desurvire, *Erbium Doped Fiber Amplifiers*, Academic Press, New York, 1994.
4. M. Cvijetic, *Coherent and Nonlinear Lightwave Communications*, Artech House, Boston, 1996.
5. M. Cvijetic, Performance Evaluation of externally modulated high bit rate lightwave systems, *IEEE Photon. Technol. Lett.* **9**: 687–689 (1997).
6. *ITU-T Recommendations on Optical Fibers*, G.652/G. 653/G. 655, Geneva, 1993.
7. G. P. Agraval, *Nonlinear Fiber Optics*, 2nd ed., Academic Press, San Diego, CA, 1995.
8. N. Shibata, R. P. Brown, and R. G. Waarts, Phase mismatch dependence of efficiency of wave generation through four wave mixing in a single mode optical fiber, *IEEE J. Quant. Electron.* **QE-23**: 1205–1210 (1987).
9. J. Zhou et al., Crosstalk in multiwavelength optical crossconnect networks, *IEEE/OSA J. Lightwave Technol.* (Special Issue on Multiwavelength Technology and Networks) **JLT-14**: 1423–1435 (1996).
10. A. Mecozzi and M. Shtaif, On the capacity of intensity modulated systems using optical amplifiers, *IEEE Photon. Technol. Lett.* **13**: 1029–1031 (2001).

OPTICAL WIRELESS LASER COMMUNICATIONS: FREE-SPACE OPTICS

Dennis Killinger
University of South Florida
Tampa, Florida

1. INTRODUCTION

Free-space optics (FSO) communication involves the use of modulated optical beams to send telecommunication information through the atmosphere from one location to another location, and has been the subject of a series of several conferences on FSO communication [1–3]. The concept of FSO light communication is not new, having been used by the Romans to transmit information via mirror reflected optical beams from one hill to another during ancient times. Indeed, Alexander Graham Bell in his photophone patent dated 1880 showed the use of an intensity modulated optical beam to transmit telephone signals 200 m through the air to a distant receiver. More recently, however, the tremendous growth in Internet traffic due to the use of high-bandwidth optical fiber transmission networks and the development of low-cost and high-power diode lasers has greatly increased the utility of transmitting information on an optical laser beam from one location to another through the air. Since 1996, the use of free-space optics has grown exponentially in the commercial market since it offers the potential to help connect the millions of telecom users within the "last mile" and at extremely high bandwidths of 1–10 gigabits per second (Gbps) or more. While cable (coax) offers 1-Gbps capability, it must be shared in bandwidth among different users and channels within a neighborhood or hub. T1 lines into offices carry a 1 Mbps (megabits per second) bandwidth, but are usually fiberoptic-coupled to a main hub. The more recent RF wireless 802.11b capability to link office and home computers wirelessly

to a common hub provides bandwidths of 11 Mbps, but can become crowded in capacity when many (say, 20–100) notebook computers are being used at the same time, such as within a campus library room. As such, future need and growth is anticipated for the development of individual higher bandwidth (0.1–1 Gbps per user) connectivity within all offices and homes in a metro (metropolitan) market. Optical access and FSO in particular may offer the optimal technical solution for such connectivity in the future since only optics offers such large bandwidths for each individual user. As the National Academy/NRC Committee on Optical Science and Engineering (COSE) report recently stated, "The Tera-bit/s era for information technology . . . includes the need for cost-effective networks of virtually unlimited bandwidth with local area networks operating at tens of gigabits/s" [4].

1.1. Historical Background and Current Technology Perspective

Historically, a significant amount of laser telecommunication and laser atmospheric propagation studies were conducted in the 1970s and 1980s as part of the development of military electrooptic instruments, laser radar systems, and secure communication data links. Much of the optical and laser science underlying these systems can be found in current optical handbooks. [5,6] Early Department of Defense (DoD) work involved the detailed analysis of the attenuation and scatter of a laser beam transmitted through the atmosphere, the common physical and parametric analysis of different types of visible and infrared detectors, and the intensity modulation and wavelength control of a wide range of lasers. For example, several laser FSO communication systems were developed in the 1980s for secure ship-to-ship communication and ground-to-aircraft applications. In addition, since the early 1990s, several secure laser communication systems for use between the ground and satellite-to-satellite have been developed and launched [7–9]. Most of these early or DoD FSO systems were designed to be used for long-range (5–1000 km) communication links and often used high-power (1–200-W) 10-μm-wavelength CO_2 lasers, 1.06-μm Nd:YAG, 0.85 μm GaAs, or 1.5-μm diode/erbium:fiber amplifier lasers. They often involved complex tracking systems, multiple detector receivers or adaptive optics to compensate for atmospheric turbulence, external modulators for the high-power lasers to place the communication signal on the laser intensity bitstream, and seldom were considered to eye-safe. Many of these systems had development costs on the order of several millions of dollars, although vehicle mounted systems have typical costs on the order of $100,000. The number of systems deployed is relatively small, ranging from one-of-a-kind satellite-to-satellite or ground-to-airborne systems, to unit numbers in the hundreds for small-size mobile systems. These DoD FSO systems and laser atmospheric studies are emphasized here since they provide much of the scientific groundwork in laser, propagation, signal processing, and detector technology in the context of current commercial FSO systems and for the design of future optical wireless communication systems.

Since 1996 there has been an explosion in commercial development of FSO systems that has been driven in part by several considerations from both technology and business viewpoints [10]. First and most importantly, the demand for high-bandwidth Internet connectivity within the last-mile market has driven the use of FSO systems in places where optical fiber is too expensive to use, especially within the urban metro market. Here, the price to lay fiber from one building to another building just across the street may cost $300,000 and take 6 months time to obtain a permit, if allowed at all. Second, the advent of directly modulated, moderate power diode lasers and light-emitting-diodes (LEDs) that are inexpensive and compact have allowed the development of low–moderate-cost FSO systems for short to moderate ranges. Initially, FSO commercial systems developed in the 1990s used higher-power (1–10-W) lasers to transmit a 0.1–1-GHz bandwidth signal at distances of 5–10 km. The systems were designed to provide point-to-point connectivity over a long distance, often used active tracking to compensate for building sway and atmospheric turbulence, and cost upwards of $100,000 [1]. Since the mid-1990s, the design of many commercial systems has gravitated toward shorter-range and lower-cost systems that use moderate power (10–100 mW) diode lasers or LEDs, and operate at shorter ranges (100–500 m) [3]. There are about 15 commercial companies currently selling FSO systems, ranging in price from $1000 per unit for 10-Mbps systems to about $20,000–$100,000 per system for advanced capabilities (1–10 Gbps). Several of these companies have installed or sold nearly 5000 FSO systems each, while some are implementing a complete networking capability in a point-to-multipoint or mesh-net configuration [3]. It is interesting to note, however, that most of the current commercial FSO systems operate at wavelengths of 0.8 or 1.5 μm, and use laser or optical technology derived mainly from the fiberoptic telecommunication community as opposed to the previously mentioned DoD laser sensor and atmospheric propagation community. This is due to two reasons: (1) the development of inexpensive and reliable laser and receiver/detectors near 0.8 μm (fiberoptic telecommunication in the 1980s) and 1.5-μm diode lasers with erbium:fiber amplifiers (fiberoptic telecommunication in the 1990s), and (2) the wish to remain compatible in optical format to the burgeoning fiberoptics telecommunications field and the wavelength and information requirements imposed by the standards adopted by the industry. FSO technology and deployment in the commercial sector may be viewed as being just in its infancy, with several marketing studies indicating that the current $200 million/year market could grow to $2 billion/year before the year 2007. As such, the reader should remember that the current growth in FSO may continue with advances along the current wavelength and devices used, or it may branch out to other wavelength regions (say, for instance, 3.5 or 9 μm) according to the technical needs of the market. If the last-mile, metro, or home market becomes the main focus of the telecommunications field in the future (say, 10 years from this writing), and if FSO plays a major role in its development, then the optimization of FSO last-mile technology at other wavelengths may

be more important than the demands to interface and use optical technology from the fiberoptic network legacy. It is beyond the scope of this article to speculate on the future of this area in light of some of these considerations, but such thoughts focus on the importance of covering the basic optical physics and technology behind FSO in this so that the reader can appreciate the different optical tradeoffs, technical limitations, and capabilities of FSO.

1.2. Technical Emphasis of Tutorial Overview

This article presents a brief overview of the optical science and technology involved in free-space optics communication. The emphasis will be on the basic physics and engineering aspects of FSO design mainly with an eye toward point-to-point communication for the commercial market. As such, the physics of laser atmospheric propagation, laser specifications, detector criteria, telescope design, eye safety, laser-beam tracking, and atmospheric turbulence and beam wander will be discussed. These are the main criteria for the design of current FSO systems and will probably determine the design of future systems. Most FSO systems are "modulation-tolerant" or "protocol-agnostic," which means that they faithfully reproduce the input modulation codes of a communication link. As such, they can be placed inside a wide range of different communication networks (Ethernet, FDDI, SONET, etc.) without changes being required as the system communication schemes are modified and updated in future years. Although some traditional copper data link formats such as T-1 and T-3 may require some data conversion. In addition, their use within a traditional network ring or multipoint configuration may also be independent of other communication parameters, however, this may not be the case if the FSO is not an add-on to an existing net but develops as the main net or "mesh-net" component. Since many of these latter topics on communication network topology and modulation or coding schemes are covered elsewhere, they will not be covered in this article on FSO except only as needed. The interested reader is directed toward these topics within this encyclopedia. It should be noted that there are several excellent overviews and technical articles on FSO that the reader may want to review, including the book on free-space optics communication by Willebrand and Ghuman that covers technical, marketing, and network issues; in-depth papers presented at several SPIE conferences; and specific research papers [1–3,10].

2. OVERVIEW OF A TYPICAL FREE-SPACE OPTICS COMMUNICATION SYSTEM

A typical free-space optics (FSO) communication system is depicted Fig. 1. It often consists of a small laser source that can be directly modulated in intensity at fairly high data rates, a beamshaping–transmitting telescope lens to transmit the laser through the atmosphere toward a distant point, a receiving lens or telescope to collect and focus the intercepted laser light onto a photodetector, and a receiver amplifier to amplify and condition the received communication signal. Figure 1 also shows the transmitted laser beam passing through the atmosphere and being partially collected by the receiver telescope. The laser or optical beam can be absorbed, scattered, or displaced by the atmosphere, depending on the atmospheric conditions and wavelength/linewidth of the laser source. If the laser beam has to transverse distances less than about 200–500 m or so, then finite movement and sway of the local buildings attached to the system may move the transmitted beam away from the receiving telescope aperture and outside the angular acceptance angle of the system. In this case or the case of high atmospheric turbulence, an active tracking device may have to be used to align the beam onto the receiver using a small gimbal mirror, lens translation stage, or detector/laser translation stage; active tracking may be eliminated if sufficient power is available by expanding the divergence of the beam or if the building and alignment is stable.

To give the reader a perspective of the size and shape of a typical FSO system, Fig. 2 is a photograph of a FSO indoor unit from PlainTree, Inc. (model 340) that uses a low-power (40-mW) 0.85-μm-wavelength light-emitting diode (LED) nonlaser (noncoherent) source for optical communication at short to moderate ranges (100–200 m) [11]. The transmitter lens is about 6.5 cm in diameter, the receiver telescope lens is about 13 cm

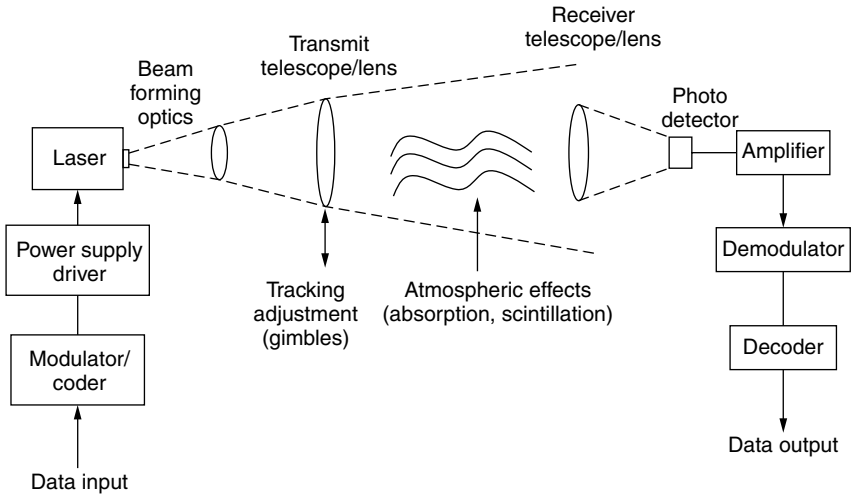


Figure 1. Pictorial schematic of free-space optics laser communication system for point-to-point applications. A typical FSO system has two of these optical channels for bidirection flow of the communication link.

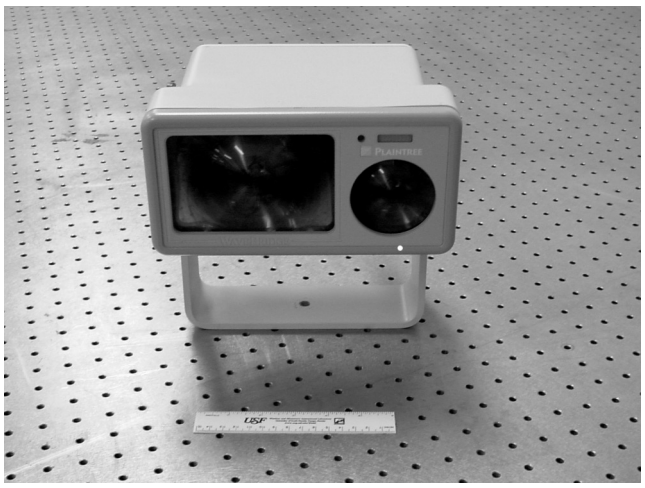


Figure 2. Photo of 0.85-μm noncoherent LED-based FSO system from PlainTree used for short/moderate indoor applications. Photo is that of a PlainTree Model 340 system.

Figure 3. Photo of 0.785-μm-wavelength diode laser FSO system from Optical Access that uses Four laser beams to reduce speckle fading and atmospheric turbulence effects.

in diameter, and the beam divergence is about 1° (0.0175 radians). The data rate is about 10 Mbps. Figure 3 shows a FSO system from Optical Access (model 155) that operates in the same near-IR wavelength region, but uses a 0.785-μm-diode laser and four laser beams to reduce fading and atmospheric turbulence effects. Here the laser output power is about 7 mW and the data rate is 155 Mbps [12].

Figure 4 shows a higher power and different wavelength system from fSONA, Inc. (model 622-M) that uses four separate 4-cm-diameter beams from 100-mW diode lasers operating at 1.55 μm that is able to transmit 622 Mbps information at ranges up to 2.5 km [13]. Here, the receiver telescope size is 20 cm and the four transmitted beams are used to reduce the effect of increased FSO signal fluctuations due to the effects of atmospheric turbulence at longer ranges and interference/speckle fluctuations associated with the use of a coherent laser. The four transmitter lasers also offer redundancy for the link.

Figure 4. Photo of 1.55-μm high-power diode laser FSO system from fSONA that uses Four laser beams to reduce signal fluctuations.

More sophisticated systems have also been developed that have been able to transmit data at rates beyond 40 Gbps at ranges of 5 km or more using multiple laser wavelengths, active atmospheric tracking, and multiple detectors and beams [14].

The detected FSO optical signal is usually converted to an electrical signal as shown in Fig. 1 and then sent to the communication network or individual hub. However, the optical detected signal can be redirected via mirrors to another location, or received by a router that will redirect it into a fiberoptic communication system. Of course, the optical signal received could also be amplified by another laser amplifier, as in the case of a 1.5-μm laser signal and an Er:YAG fiber amplifier. This is discussed later in Section 3 in this article.

As can be appreciated from Fig. 1, several important optics and laser spectroscopic issues need to be considered in the design and development of a FSO system. These include the availability and wavelength coverage of lasers and LED sources, the interaction and attenuation of the FSO optical beam as it traverses the atmosphere, the received light intensity collected by the receiver telescope, the sensitivity of the optical detectors, and the influence of atmospheric turbulence. These factors influence the relative SNR of the received laser signal and are depicted in the FSO range equation. These aspects are covered in the following sections.

3. LASER AND OPTICAL SOURCES FOR FSO

Most current FSO systems use either 0.8- or 1.5-μm lasers or LED sources. However, there are a wide range of other lasers and LED sources that have potential for use in a FSO system. For FSO applications, it is appropriate to look only at continuous-wave (CW) lasers as opposed to pulsed lasers since they can more often be intensity modulated at MHz or GHz rates. In this regard, Fig. 5 shows a plot of the output power of several CW lasers as a function of the wavelength covered [15–17]. As can

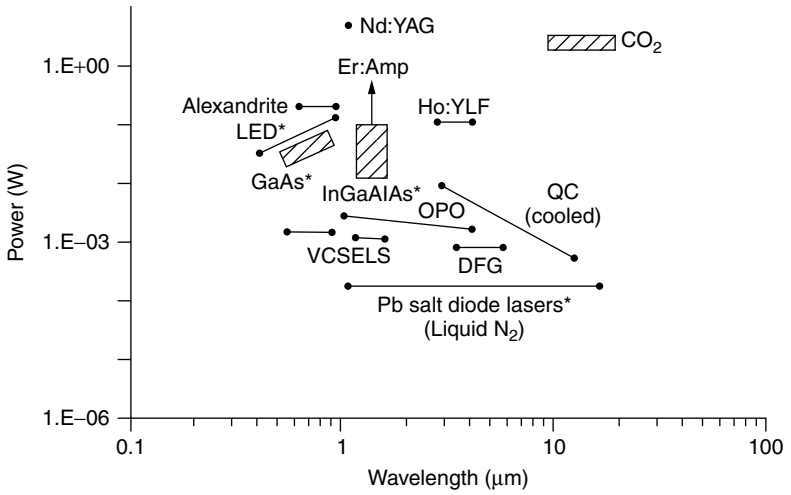


Figure 5. Typical laser output power in Watts as a function of wavelength for current lasers. Asterisks (∗) indicate direct modulation (10 MHz to 10 GHz); otherwise, external cavity modulation is used.

be seen, several high-power lasers such as the CO_2 and Nd:YAG laser operate only over a narrow wavelength range, while several lower-power lasers such as the optical parametric oscillator (OPO) and difference frequency generation (DFG) laser operate over a wider tuning range but at lower output power. The GaAs(Al) diode lasers operating near 0.8–0.9-μm wavelengths have CW power levels in the order of 0.01–0.1 W, while InGaAs(P) lasers near 1.5 μm operate with tens of mW power; the latter can be boosted by the use of Er:fiber laser amplifiers to levels of 1–10 W; governmental laboratory fiber lasers have reached levels of 100–200 W and higher, although problems with spectral mode hopping and spontaneous background emission require stringent cavity design and injection seeder laser-beam isolation. The vertical-cavity surface emitting lasers (VCSELs) are vertical layered semiconductor lasers that have output powers on the order of 1–10 mW and are tunable in some cavity arrangements. The quantum cascade (QC) laser offers potential for future FSO usage, especially with the development of 9-μm room-temperature lasers [18]. Also shown in Fig. 5 is the output power for a LED, which is a noncoherent light source but has wide utility as a FSO source. Of the lasers shown in Fig. 5, only the semiconductor diode lasers are directly modulated at rates up to 10 Gbps using the drive current of the laser or an internal loss material. The other lasers have to use external modulators (electrooptic, acoustooptic, or traveling-wave modulators) to reach Mbps–Gbps modulation rates. The LED modulation rate is generally 1–10 MHz, but newer models, including laboratory quasicavity LEDs, have modulation rates on the order of 100 MHz.

Some important output characteristics of these different CW lasers are tabulated in Table 1 [15–17]. As can be seen, the GaAs lasers and InGaAs lasers offer a significant combination of high output power and can be directly modulated via their drive currents. Some of the lasers have linewidths that are either single frequency (single longitudinal cavity mode) or consist of several laser modes within a group of lines, while an LED has a broad noncoherent emission spectrum. This can be seen in Fig. 6, which shows the spectral output measured by the author from three different optical sources as a function of wavelength or frequency. As can be seen in the figure, the output spectrum from the LED is broad, covering a range of 50 nm (about 500 cm^{-1} or 15,000 GHz), while the typical 1.33-μm diode laser shows multiple longitudinal modes covering a range of about 3 nm. The bottom portion of the figure shows the 50-kHz linewidth output spectrum of a single-frequency 1.55-μm distributed feedback (DFB) laser whose laser output has been controlled through use of Bragg reflection from imposed index variations along

Table 1. Output Characteristics of Several Currently Available CW Laser Sources[a]

Laser	λ (μm)	Power	Temperature	Modulation Rate	LineWidth[a]
GaAs	0.8	10–100 mW*	Room	100–500 MHz*	Multi/SF
VCSELs	0.8, 1.5	1–10 mW	Room	1 GHz*	SF
InGaAs	1.5	10–100 mW	Room/TE[b]	10 GHz*	Multi/SF
LED (nonlaser)	0.8	10–100 mW	Room	10–100 MHz*	50 nm
CO_2	10	1 W	Room	200 kHz*/20 GHz	$0.1cm^{-1}$/SF
Nd:YAG	1.06	1 W	Room	1 GHz	$0.1cm^{-1}$/SF
Quantum cascade	3–9	0.1–1 mW	77 K–room	?	$0.1cm^{-1}$/SF
Pb salt	2–10	0.1 mW	77 K	?	0.1–1 cm^{-1}
Ho:YAG	2.06	100 mW	Room	1 MHz ?	$0.1cm^{-1}$/SF

[a] Linewidths may have multilongitudinal modes or be single frequency (SF).
[b] Thermoelectrically cooled.
*indicates direct modulation of the laser intensity while other lasers use external cavity modulation.

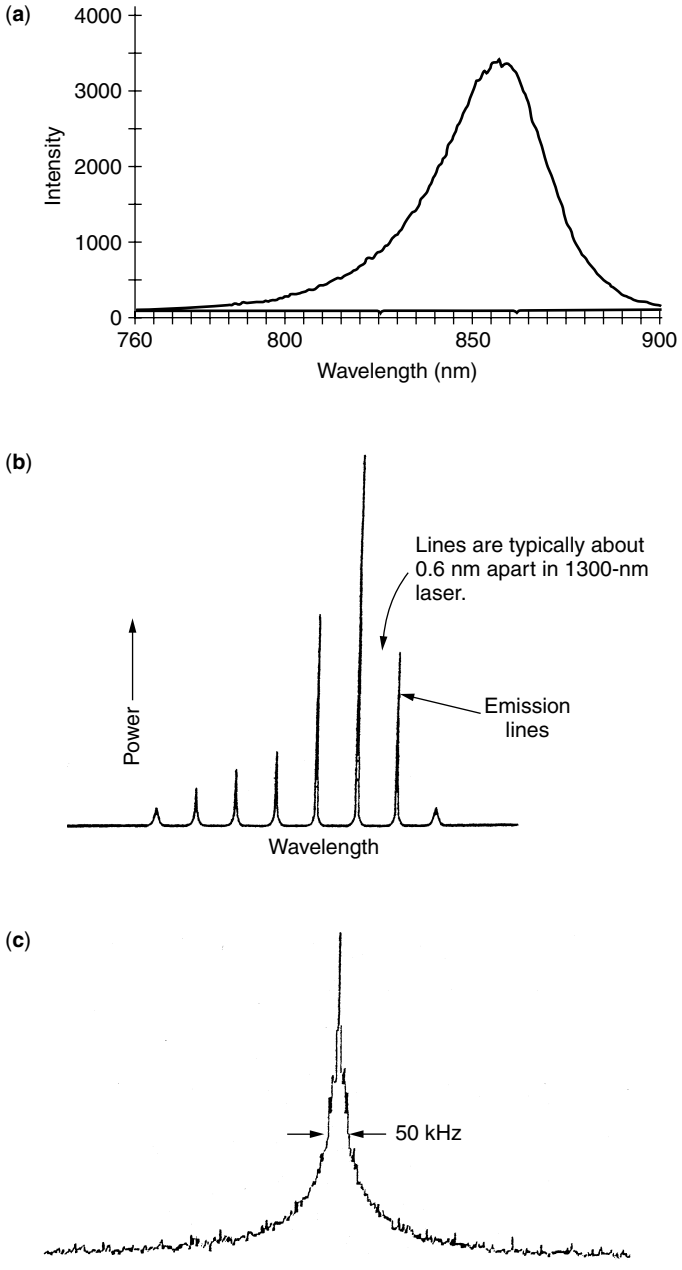


Figure 6. Spectral plots of the output power as a function of wavelength showing the spectral linewidth of a typical 0.86-μm LED source (**a**), 1.3-μm multimode diode laser (**b**), and a 1.55-μm single-frequency DFB diode laser (**c**).

the semiconductor laser cavity. Figure 6 is important in that the exact wavelength and linewidth of the laser can determine the absorption and scatter properties of a FSO laser beam as it propagates through the atmosphere. This will be seen more clearly in the next section.

It should be noted that many laser sources do not operate in a single spatial mode, but may have a divergence that is greater than the lowest-order (Gaussian) mode. A measure of this deviation is the M^2 parameter value, or mode structure parameter. For a Gaussian spatial beam, M^2 is equal to 1. However, for many short cavity lasers, the M^2 value may be closer to 2 or 3 for diode lasers, and as high as 20 to 50 for short cavity OPO lasers. An alternative to specifying the M^2 parameter value is to directly specify or measure the divergence of the laser in terms of milliradians. The divergence of the laser beam is equal to M^2 times the divergence of a Gaussian beam [19]. It is common in a FSO system for the divergence of the transmitted beam to be made larger than the diffraction minimum value so that the projected beam size is larger than the receiver optics and more tolerant of misalignment of the beam. This is discussed in Sections 6.1 and 6.2.

Finally, it should be added that Fig. 5 does not show an important third axis, which would be related to the cost of the laser system and modulation scheme. Such information is very important especially for commercial systems and influences the engineering trade-off design of the FSO system. To give the reader a ballpark value, typical costs range from a few dollars for a LED or low power GaAs laser to several $100K for an externally modulated CO_2 laser. Typical costs for moderate power 100 MHz 0.8 μm lasers are near $0.1K

to \$1K, with somewhat higher costs for 1.5 μm diode lasers, and approaching \$10K to \$20K for higher power Er doped amplifier lasers. Of course, these values are only approximate values amd will be reduced as technical progress is made in this area.

4. ATMOSPHERIC ATTENUATION AND SCATTER OF THE FSO BEAM

The attenuation of an optical beam as it propagates through a medium is given by the Beer–Lambert law as

$$I(x) = I_0 e^{-\alpha x} \tag{1}$$

where I_0 is the initial optical intensity in watts, $I(x)$ is the intensity after the beam has traveled a distance x meters, and α is the attenuation coefficient of the medium in reciprocal meters. The attenuation of the atmosphere can be due to several factors, including absorption of the beam via molecules in the atmosphere and scatter of the beam due to Rayleigh, Mie, and resonant scatter with molecules or aerosol particles in the air [20]. For most applications, the Mie Scatter (especially due to fog) is dominant.

The optical transmission of the normal atmosphere can be shown as in Fig. 7, which shows the low-resolution transmission spectrum of the atmosphere for a 2-km path near ground level [21]. The spectrum is that for a low-resolution spectrometer with a spectral resolution of about 20 cm^{-1}. As can be seen, there are several regions where water vapor and other gases absorb the optical beam, while there are large optical windows in the visible (0.4–0.7 μm) and at 1.5 μm and near 9–13 μm, where the beam is hardly absorbed. Of interest for FSO applications is the higher-resolution spectra for the atmosphere at regions that appear almost opaque in Fig. 7. Figure 8 is a calculated transmission spectrum of the atmosphere for a U.S. standard atmosphere for a path of 500 m over three different spectral regions of potential FSO interest near 0.85 μm, near 1.55 μm, and near 9 μm. As can be seen, the spectrums show individual absorption lines due to the vibrational–rotational absorption lines of water vapor, CO_2, CH_4, and other gases in the atmosphere. The individual lines all have a pressure-broadened linewidth of about 0.1 cm^{-1}, so that a tunable laser beam that has a linewidth on the order of 0.1 cm^{-1} or less can be absorbed if it is tuned online, or not absorbed if it is tuned in wavelength to the offline position. It is because of this close connection between the laser or optical source linewidth and wavelength and that of the atmosphere absorption lines, that careful selection of the laser wavelength can have a significant influence on the performance of a FSO system. It should be noted that the spectra in Fig. 8 were calculated using the HITRAN database and HITRAN-PC computer program since it has a spectral resolution better than 0.01 cm^{-1}, and usually produces spectral plots that have line centers with an accuracy of 0.001 cm^{-1} and line intensity accuracy of a few percent [22–24]. Other Air Force atmospheric spectral codes such as MODTRAN have a resolution of 2–20 cm-1 and may be valid for wide-linewidth LEDs and regions of little spectral absorption. However, in general, it is best to use the high-resolution capability of the Air Force FasCode program or HITRAN-PC, which uses the HITRAN spectral line database, and then convolute the overlap of the laser spectrum with that of the atmosphere. The HITRAN database has been developed by the U.S. Air Force since 1971 and is the compilation of over one million individual spectral lines of over 32 molecules in the atmosphere.

In addition to the absorption of molecules in the atmosphere, there is also the attenuation due to the scatter from aerosols and particles in the atmosphere. In this case, the fogs, clouds, and dust particles can add to the attenuation of the optical beam. For molecules and spatial scale changes in the index of refraction that are much smaller than the wavelength of light, the scatter is called Rayleigh scatter, named after Lord Rayleigh (1842–1919), who first quantified the effect. Rayleigh scatter attenuation coefficient can be given approximately for the standard atmosphere as [24,25]

$$\begin{aligned}\alpha_{\text{ray}} &= N\sigma_\pi = N\frac{8\pi}{3}\sigma_\pi \\ &= \frac{8\pi}{3}1.18 \times 10^{-8}\ [550\ \text{nm}/\lambda\ (\text{nm})]^4\ \text{cm}^{-1} \\ &= 1.1 \times 10^{-5}\ [550\ \text{nm}/\lambda\ (\text{nm})]^4\ \text{m}^{-1}\end{aligned} \tag{2}$$

where N is the number of molecules ($\sim 2.55 \times 10^{19}$ molecules/cm^3) in air and σ_π is the backscatter Rayleigh scatter cross section. Equation (2) is normalized to that for a wavelength of 550 nm. As can be seen, Rayleigh scatter increases in the short-wavelength or blue-wavelength regions, which is why sunsets appear red (most of the blue light has been scattered away from the viewing angle).

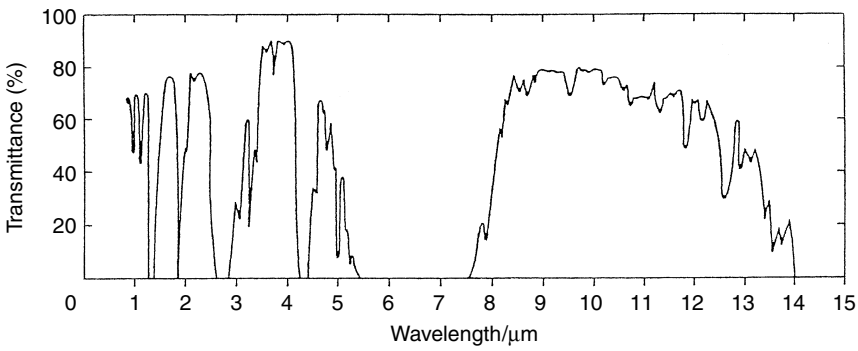


Figure 7. Calculated transmission spectrum of the standard atmosphere for a pathlength of 2 km. Strong absorption regions near 2.8μ m and 6–7 μ m are due mostly to water vapor.

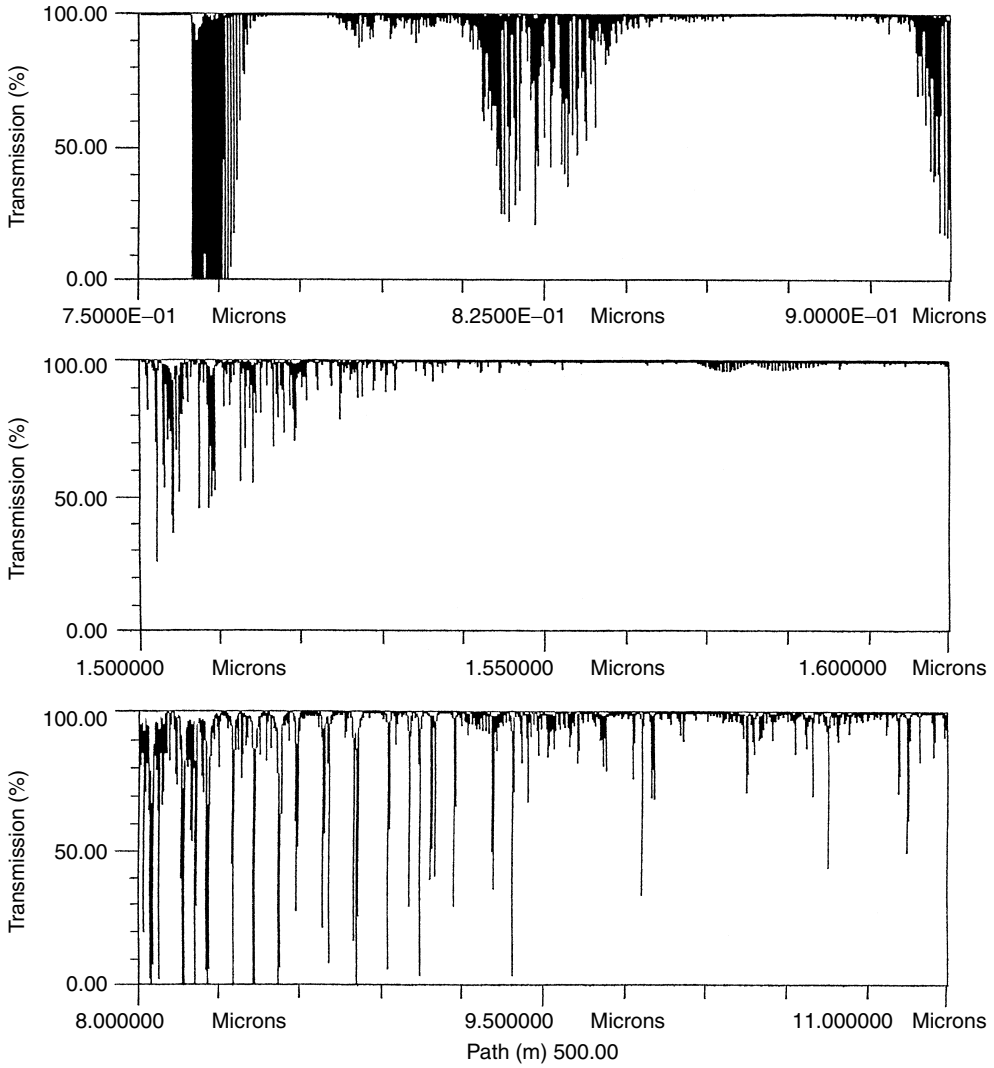


Figure 8. Calculated transmission spectrum of the atmosphere for a 500-m path near the wavelengths of 0.8, 1.55, and 9.5 μm. Most of the strong rotational–vibrational absorption lines seen are due to water vapor, CO_2, ozone, and oxygen.

When the scatter site is large or on the order of the wavelength of light then the scatter is a complex interference phenomenon with destructive and constructive interference lobes emanating outward from the particle. Such scatter is called *Mie scatter* and is highly dependent on angle, polarization, and wavelength/particle size. In theory, Mie scatter can be calculated for known particle sizes and orientation. However, it cannot be calculated a priori for complex shapes and orientations of particles such as those often found in the atmosphere. As such, the Mie scatter for the atmosphere is usually measured experimentally. Figure 9 shows the measured attenuation or extinction coefficient of the atmosphere as a function of wavelength for several different atmospheric conditions [25]. The values shown in Fig. 9 have error bars on the order of an order of magnitude dependent on atmospheric conditions.

Comparison of Figs. 8 and 9 suggests that at many laser wavelengths, the attenuation due to a strong absorption line in the atmosphere may be more dominant than that due to the normal background attenuation of the atmosphere such as that due to urban haze. In this case, FSO design is such that one chooses a laser wavelength that is offline of any strong absorption line in the atmosphere. After this choice, then the next dominant attenuation consideration is that due to clouds or heavy fog. For example, at a wavelength of about 1.51 μm, the extinction coefficient due to urban haze is about 0.9×10^{-4} m^{-1}, a value much smaller than that possible due to the molecular lines in Fig. 8. However, the attenuation due to absorption lines at 1.56 μm is negligible so that the Mie/Rayleigh or haze attenuation dominates.

Of more concern for a FSO system is the attenuation due to rain, snow, and fog. The Air Force MODTRAN and LOWTRAN computer programs have excellent attenuation calculations for rain and snow [22,23]. Under most cases with short ranges (<500 m), rain and snow attenuation may not be severe. However, fog can cause severe degradation in the signal. Figure 10 is a plot of the attenuation due to fog, rain, and snow as a function

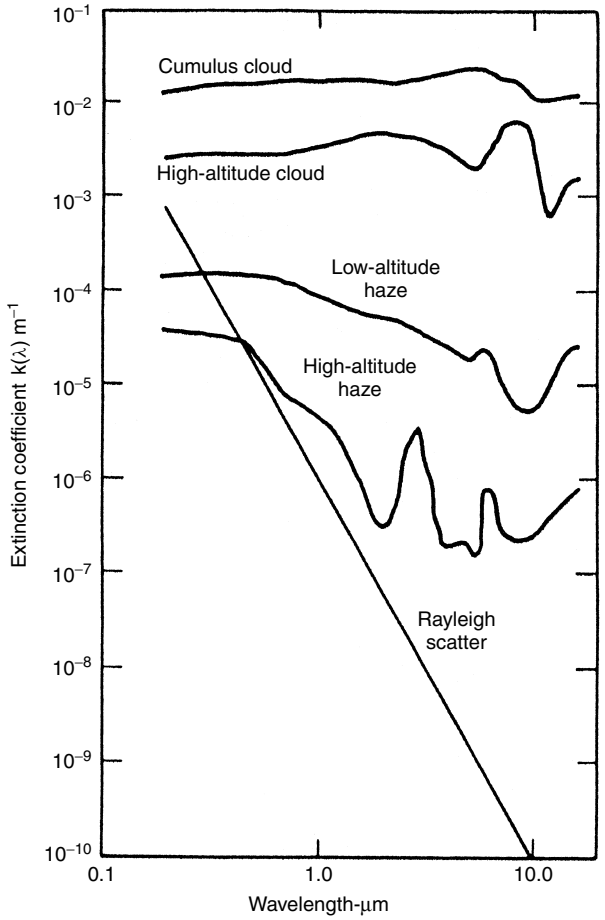


Figure 9. Approximate attenuation coefficient of the atmosphere as a function of wavelength for different atmospheric conditions. (Reproduced by permission of R. Measures, *Laser Remote Sensing*, John Wiley & Sons, New York, 1984.)

of the visibility [26]. As can be seen, thick fog can cause attenuation of up to 200 dB/km, or a reduction factor of 10^{-20} for a km path. Recent studies by Kim, McArthur, and Korevaar at 0.8 and 1.5 μm have indicated a refined equation for an approximation of the attenuation value, α, given by [26]

$$\alpha = \left(\frac{3.91}{V}\right)\left(\frac{\lambda}{550\text{ nm}}\right)^{-q}\ \text{km}^{-1} \qquad (3)$$

where V is the visibility (in km), λ is the wavelength in nm, and q depends on the size of the scattering particles, but is equal to 1.3 for average visibility and 0 for fog. In Eq. (3) α is given in units of km^{-1}.

It should be noted that in many reported studies, the attenuation values given in dB/km were extrapolated from short range studies on the order of several centimeters to meters, and were unable to take into account multipath scatter. The inclusion of multipath effects may decrease the overall attenuation value but also may spread out in time the modulated intensity waveform of the multiscattered beam [24].

Under normal conditions the first decision criterion for a FSO wavelength design is to reduce the atmospheric line spectra as shown in Fig. 8. Then the next consideration should be the reduction due to fog-type aerosols in the path. As can be seen in Fig. 9, the latter consideration may indicate that longer wavelengths near 9–10 μm may offer less attenuation due to fog and snow. As such, they may be considered for a backup system for a 0.8- or 1.5-μm FSO system, as opposed to the use of a microwave or RF backup system.

5. OPTICAL DETECTORS AND NOISE

Most current commercial FSO systems use the direct detection of the intensity-modulated laser beam. The optical detectors used are usually a small, high-bandwidth photodetector, either a Si photodiode or Si avalanche photodiode (APD) for wavelengths up to 1.1 μm, or a InGaAs photodiode or APD for the 1.5 μm wavelengths. To obtain the high speed required of 10 MHz up to 10 GHz, the size of the detector is kept small, on the order of 20–100 μm, to reduce capacitance and RC time constants. Table 2 shows a sampling of several optical detectors and some of their performance parameter values, including their size, electrical bandwidth, and noise equivalent power (i.e., minimum signal detected) [5,6,27]. As can be seen, they are small in size and have detection sensitivities ranging from a microwatt down to tens of nanowatts.

The overall science of optical detectors is covered in several excellent books, including that of R. Kingston, which covers photon counting, amplifier and background noise, and signal-dependent noise-limited performance of a FSO optical communication system [27,28]. In general, the detectors used in the visible to near-IR spectral region

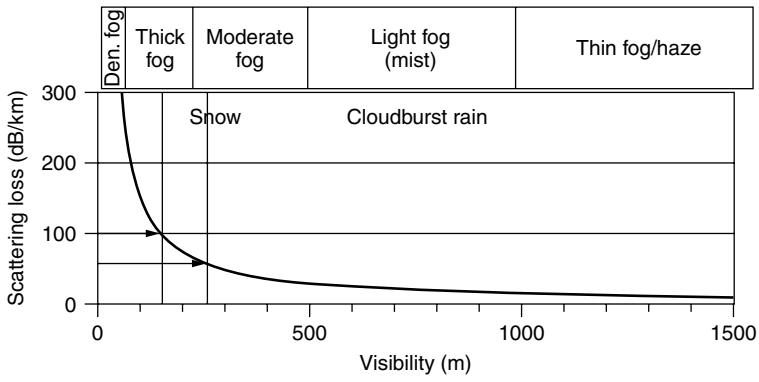


Figure 10. Attenuation/scattering loss as a function of visibility for different haze and fog conditions. (Reproduced by permission from I. I. Kim, B. McArthur, and E. Korevaar, SPIE Vol. 4214, 2001.)

Table 2. Typical NEP Values of Selected Detectors for Nighttime Conditions

	λ (μm)	Size (μm)	Bandwidth	Nighttime NEP (nW)
Si photodiode	0.9	2000	10 Mbps	200
Si APD	0.9	200	155 Mbps	20
InGaAS APD	1.5	50	2.5 Gbps	50
InGaAs APD	1.5	200	100 Mbps	20

[a]Daytime usage may require narrowband optical blocking filters to reduce background light. Typical daytime increase in NEP is ~2–8×.

are shot-noise-limited; that is, the dominant noise is the statistical fluctuations of the signal photons, which is essentially the square root of the number of photons in the signal. As such, the noise of the detector is usually stipulated in terms of a minimum detectable signal in decibels referenced to a milliwatt, or as a background current in the case of Johnson noise of the detector or amplifier combination. In the infrared spectral region, however, the background radiation or thermal emission of the 300-K world is often the dominant noise source. In the latter case, the detectors are background limited in their performance [i.e., background-limited infrared performance (BLIP)], and a different parameter related to the intrinsic sensitivity of the photodetector material is used to determine the system noise level. In this case, the detectivity, D^*, in units of Jones (i.e., cm $Hz^{1/2}$ W^{-1}), is a universal parameter for a particular material and is related to the noise equivalent power (NEP) in watts by

$$\text{NEP} = \frac{(A_D B)^{1/2}}{D^*} \tag{4}$$

where A_D is the area of the detector and B is the electrical bandwidth of the detector/amplifier combination [29]. The NEP is where the SNR in voltage equals 1 at the output terminals of the detector. The NEP is related to the term "sensitivity" as expressed in watts and used more often in the visible–near IR. "Sensitivity" is often used for a specified bit-error-rate and is about equal to 6 times NEP; this is explained in a later section.

With these concepts, one can compare different types of detectors at different wavelength ranges. Figure 11 shows a plot of the measured D^* for a range of infrared detectors along with that due to a S-20 photocathode photomultiplier tube (PMT), Si photovoltaic (pv) photodetector, and liquid nitrogen cooled HgCdTe for 10 μm wavelengths [30]. The influence of the 300-K background theoretical maximum thermal noise level shown by the doted line is given for both a photovoltaic detector and a photoconductor detector. As can be seen, the BLIP performance limit, due to the 300-K thermal background radiation that peaks near 8–10 μm in wavelength, is dominant for wavelengths greater than ~2 μm. Here, the D^* value was calculated for a bandwidth of 1 Hz for the PMT so that a comparison between the different detectors could be made. What are not shown in Fig. 11, however, are the signal bandwidth, lifetime, and cost of each detector. For example, many of the infrared detectors shown in Fig. 11 can operate at bandwidths

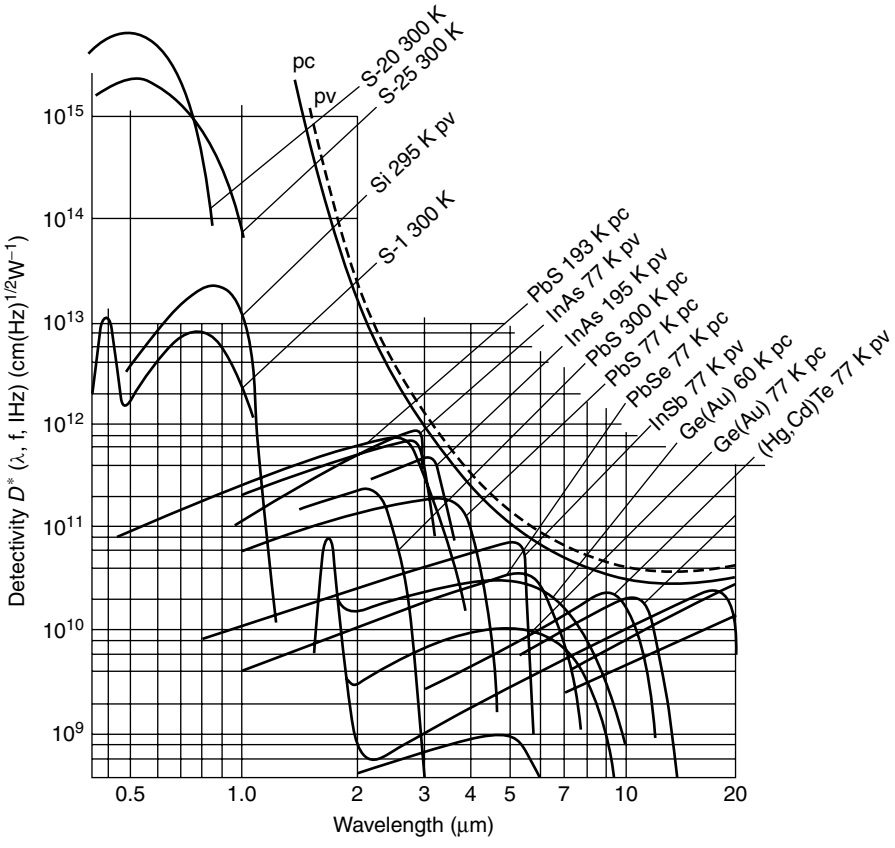


Figure 11. Detectivity, D^*, for selected detectors as a function of wavelength. (Adapted from P. R. Norton, *Handbook of Optics*, McGraw-Hill, New York, 1995, Chap. 15.)

of only 1–10 MHz or slower. On the other hand, cooled HgCdTe detectors have been operated at a wavelength of 10 μm in a heterodyne detection mode at speeds of ≤60 GHz.

Most communication systems use the bit error rate (BER) as a measure of the system sensitivity and level of signal-to-noise ratio in determining the probability of correctly decoding the bitstream in the signal [28,31]. The BER can be related to the signal-to-noise ratio (SNR) of the communication link and is a measure of the percentage of bits that are in error within a large ensemble of bits received. It is common for current FSO links to have a BER on the order of 10^{-9} to 10^{-10} for the case where no error correcting codes are used. The optical BER can be calculated as the integral of 1 minus the cumulative normal distribution function with argument $(\mathrm{SNR})_v/2$, where the SNR_v is the peak voltage SNR for a detector, which is the same as the returned peak power divided by the NEP of the detector. The average SNR power is half the peak SNR value. The "sensitivity" of a detector is defined for optical communication purposes as the average power required for a BER of 10^{-9}.

A plot of the BER value as a function of the voltage signal-to-noise ratio (returned optical power divided by NEP) is shown in Fig. 12 [28]. As can be seen, a BER of 10^{-9} requires a peak SNR of 12, which corresponds to a value of 6 for the average SNR value [28]. As such, the formula $\mathrm{SNR} = P_r/\mathrm{NEP} = 6$ is often used as the detection threshold for a FSO communication link; here, P_r is the average detected laser beam power by the receiver.

6. FSO RANGE EQUATION

The FSO range equation combines the attenuation and geometrical aspects of FSO in order to calculate the received optical power as a function of range and telescope aperture size. Before introducing the FSO range equation, some parameter values need to be defined and discussed.

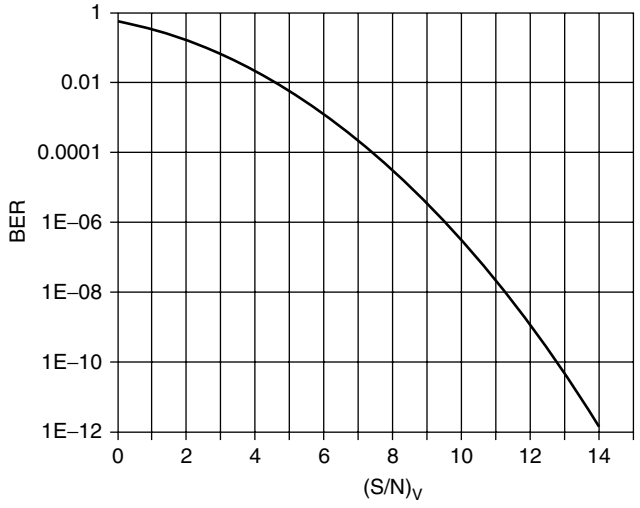


Figure 12. Bit error rate (BER) as a function of peak voltage signal-to-noise ratio (SNR), which is equal to the received power P_r, divided by the NEP of the detector. (Reproduced by permission of R. H. Kingston, *Optical Sources, Detectors, and Systems*, Academic Press, New York, 1995.)

6.1. Laser-Transmitted Beam Divergence

In the most simplistic case, the transmitted laser beam is divergent as a result of optical diffraction, where the angular spread, $\Delta\theta_1$ is equal to λ/D_1, where D_1 is the size of the initial laser beam. This is an approximate equation for the divergence of the lowest-order Gaussian spatial TEM_{00} mode for a Fabry–Perot laser cavity. Laser beams with higher-order spatial modes have greater diffraction than a Gaussian mode and will have a mode structure parameter, M^2, value greater than 1. In these cases, the divergence of the beam in one direction is equal to $M^2\Delta\theta_1$, or [19]

$$\Delta\theta_1 = \frac{(M^2)\lambda}{D_1} \tag{5}$$

The size of the projected laser beam at a distance of R meters will be equal to $D_1 + R\Delta\theta_1$. For example, a $M^2 = 10$ and 1 μm wavelength laser beam collimated through a 1 cm aperture will have an angular divergence of 10^{-2} radians (i.e., 0.57°) and will have a width of 10 m at a range of 1000 m.

Often, the beam divergence of a FSO system is made intentionally larger than the diffraction limit so that the projected beam size is larger than several times the size of the receiver telescope. This facilitates alignment of the two transmitter and receiver telescope optical axes. Beam divergence or beam spread is often made to be 0.1–1 mrad by slight defocusing of the transmitter telescope, as opposed to the normal Gaussian diffraction minimum of, say, 0.001–0.01 mrad.

Finally, the divergence of a semiconductor laser is often shaped by beamforming optics near the output facet of the diode laser. Normally, the output from a semiconductor laser has a beam divergence of, say, $3^\circ \times 15^\circ$. The beamshaping optics use a cylindrical lens to bring it down to a milliradian or so, in both axes.

6.2. Receiver Telescope Field of View

The receiver telescope field of view is the beam collection angle of the detector and telescope combination. Only light that falls or originates within this cone about the telescope optical axis will be focused onto the detector. The receiver telescope field-of-view angle, $\Delta\theta_r$, is given by

$$\Delta\theta_r = \frac{D_d}{f} \tag{6}$$

where D_d is the size of the detector and f is the focal length of the receiver telescope. Equation (6) does not usually influence the optical performance unless the optical axis of the receiver telescope–detector combination is aligned outside the receiver field of view, $\Delta\theta_r$. It should be noted that just because the receiver telescope intercepts a portion of the transmitted beam, the light collected would not be focused onto the detector unless the receiver telescope axis is pointed toward the transmitter location within $\Delta\theta_r$. For a typical detector size of 300 μm and a telescope focal length of 0.3 m, the field of view is about 10^{-3} radians.

6.3. FSO Range Equation Analysis

The FSO range equation can be given by inspection of Fig. 1, and the use of the Beer–Lambert law and Eq. (5). Under these simplifying assumptions, the FSO range equation is

$$P_R = P_T \frac{A_r}{(D_1 + R\Delta\theta_1)^2} T K e^{-\alpha R} \tag{7}$$

where P_R is the received optical signal power, P_T is the transmitted optical laser power, A_r is the area of the receiver telescope or collection lens, T is the transmission or efficiency of the receiver optical system, and the area of the beam at a range R is given by $(D_1 + R\Delta\theta_1)^2$. In Eq. (7) K is another loss factor that deviates from a normal value of 1 when a noncoherent light source is used, such as an LED. This latter parameter is equal to 1 for a laser source, and has a value equal to 1 or less for an LED source as

$$K = \frac{A_{\text{det}}}{A_{\text{LED}}} \tag{8}$$

if $A_{\text{det}} < A_{\text{LED}}$, and $K = 1$ otherwise, where A_{det} is the area of the detector and A_{LED} is the area of the LED source. The K factor takes into account the fact that a noncoherent optical source cannot be focused to an area smaller than that from which it originated due to thermodynamic reciprocity (brightness) considerations.

Equation (8) can be used to generate FSO SNR or power detection curves as a function of range. For example, Fig. 13 shows the calculated received power as a function of range for a case of a 10-Mbps bandwidth, low-power 0.85-μm LED FSO system with 40 mW power, 13 cm receiver, $T = 0.2$, and $K = A_{\text{det}}/A_{\text{LED}} = (0.28\text{ mm})^2/(0.5\text{ mm})^2 = 0.3$, divergence of $1^\circ = 0.0175$ radians, and NEP of the Si detector of 300 nW for daytime operation; these specifications are appropriate for a moderate power 0.85 μm LED FSO system and serves as a "strawman" for illustrative purposes only. Two atmospheric cases are shown for low α attenuation (10^{-4}/m, or 0.2 dB/km) due to low-altitude haze, and for moderate attenuation due to clouds (10^{-2}/m, or 20 dB/km) similar to light/moderate fog. As can be seen, the returned signal follows a $1/R^2$ dependence at close-in ranges, and follows the Beer–Lambert exponential decay at longer ranges. A threshold for the NEP of the detector at 300 nW is also shown, along with a value of the required SNR of 6 times greater than the NEP corresponding to a BER of 10^{-9}. As can be seen, the system should have good FSO communication capability at ranges out to 600 m for hazy conditions and out to 200 m under moderate/light fog conditions. These ranges would be even greater under nighttime conditions when the NEP of the detector would be less, or when operating under dry/no-fog conditions.

Another example is given in Fig. 14 for a 622-Mbps-bandwidth FSO "strawman" system appropriate for a higher power multi-beam 1.55 μm laser based FSO system. Here, the FSO parameters are approximately 1.55 μm wavelength, 400 mW power diode laser, $T = 0.2, K = 1$, 20 cm receiver telescope size, transmitted 1 mrad beam divergence, and a daytime NEP of 150 nW using two solar filters. Again, two attenuation cases are shown of 1.1×10^{-2}/m for light/moderate haze (cloud) and 0.7×10^{-4}/m for low-altitude haze. As can be seen the communication range is beyond 3 km for light haze and about 600 m under light/moderate fog. However, what is not shown in Fig. 14 are the deep fades due to constructive/destructive interference and atmospheric fluctuations in the beam. This will be mentioned in a later section of this article.

It should be added that the use of the FSO range equation is an approximation to give the reader some idea of the parameters and importance of there values. The calculated ranges are theoretical values, however, and are approximations subject to atmospheric conditions, which can cause errors in the attenuation as large as an order of magnitude or more. As such, it is important that the reader understand the usefulness and limitations in using the FSO range equation. Often, the values have to be modified by direct measurements under specific atmospheric conditions. This is why in so many cases extensive field tests of a FSO optical link have to be made under a wide range of weather conditions in order to accurately measure the system performance.

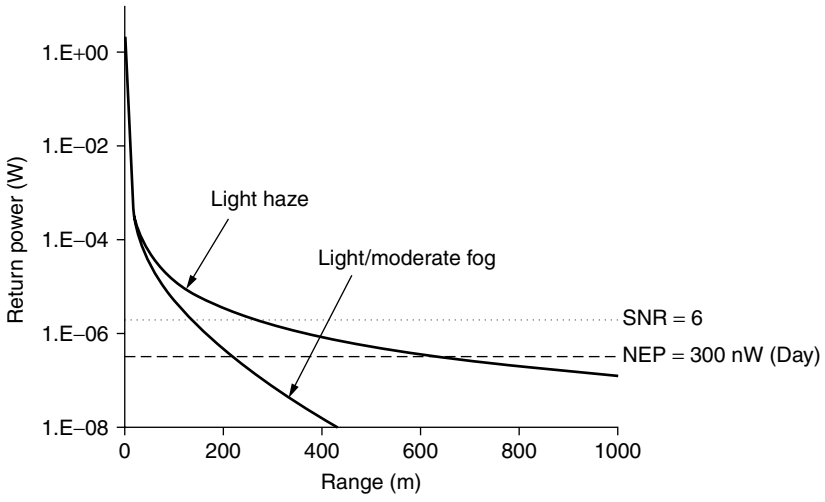


Figure 13. Calculated received optical signal as a function of range for "strawman" 0.85-μm LED FSO system. FSO range equation parameters included 40 mW power, 17 mrad beam divergence, and 13 cm telescope aperture.

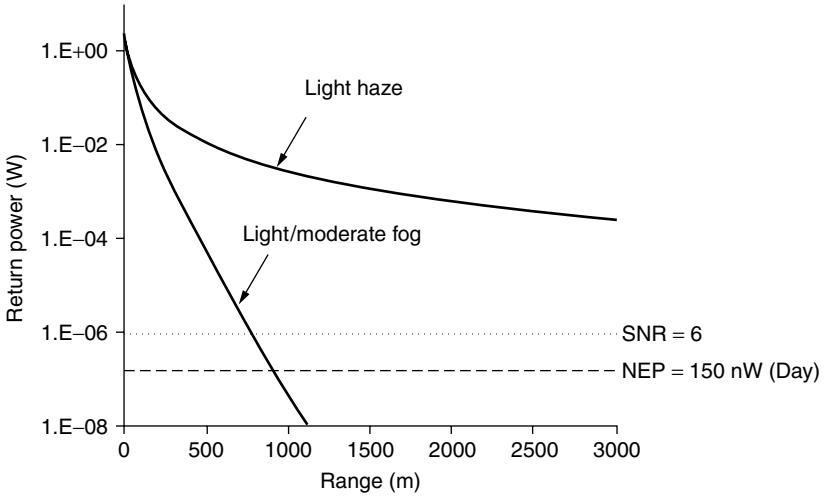


Figure 14. Calculated received optical signal as a function of range for "strawman" 1.55-μm high-power diode laser FSO system. FSO range equation parameters included 400 mW power, 1 mrad beam divergence, and 20 cm telescope aperture.

7. ATMOSPHERIC REFRACTIVE TURBULENCE

The most familiar effect of refractive turbulence in the atmosphere is the twinkling of the stars and the shimmer of the horizon on a hot day. The first of these is due to the random fluctuations in amplitude of the light, also known as *scintillation*. The second effect is the random change in the optical phase of the lightbeam that leads to a reduction in the resolution of an image. Other atmospheric effects are large-scale beam wander and breakup of the optical beam into smaller phase fronts or speckles. In the visible and near IR, these fluctuations are caused by small fluctuations in the temperature (0.01–0.1°) of the atmosphere on the spatial scale of 0.1 cm–10 m, which cause changes in the index of refraction of the atmosphere. These small-scale fluctuations can distort and break the laser beam into small turbulent cells. In the far IR spectral region, the influence of these temperature fluctuations is diminished, but large-scale spatial changes in the background absorption and concentration of water vapor can also cause large beam wander and fluctuations.

Extensive work by NOAA and DoD since the early 1960s, starting with the pioneering work of David Freed and Tatarskii, has been able to successfully clarify the phenomena of atmospheric refractive turbulence and yield predictive equations [24,32–34]. These atmospheric turbulence studies are valid for energy-conserving fluctuations in the atmosphere and have resulted in well-understood equations relating the optical beam fluctuations and the refractive-turbulence structure parameter, C_n^2. Figure 15 shows values of C_n^2 measured during a sunny day in Florida as a function of time using an optical beam intensity scintillometer instrument from NOAA [35]. As can be seen, the value of C_n^2 varies by several orders of magnitude, becoming largest during the middle part of the day.

The variation of C_n^2 with height above the ground is given approximately by [24]

$$C_n^2(h) = C_n^2(0)\, h^{-4/3} \tag{9}$$

where h is the height in meters above the ground and $C_n^2(0)$ is the value at ground level.

Fluctuations in the intensity or irradiance of the optical beam can be expressed approximately (for weak turbulence) as [24]

$$\sigma_r^2 = \exp(0.5\, k^{7/6} R^{11/6} C_n^2) - 1 \tag{10}$$

where σ_r^2 is the irradiance variance (normalized by the mean irradiance value), k is the optical wavenumber ($2\pi/\lambda$), and R is the range. This expression is modified slightly when the inner scale of the turbulence (smallest spatial fluctuation size) is greater than the square root of λR. There are several other expressions for σ_r^2 depending on the approach to saturation and the value of the inner

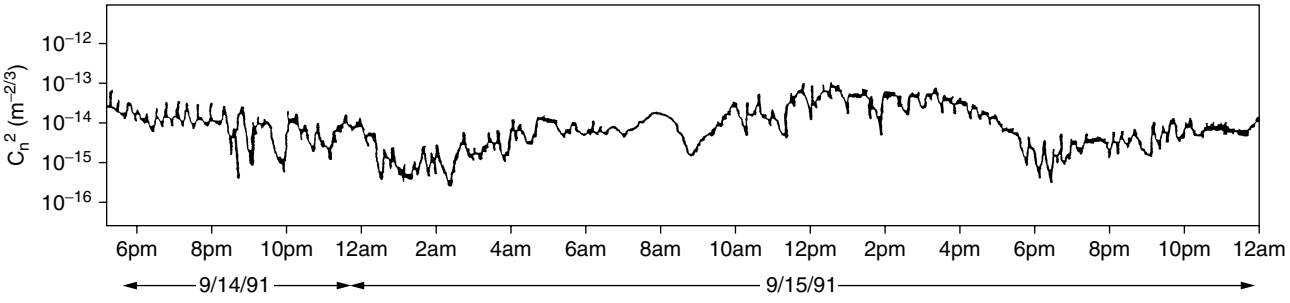


Figure 15. Measured atmospheric refractive-turbulence structure parameter, C_n^2, as a function of time throughout the day.

and outer scale values. The interested reader should consult the references for more detail [24,32–34].

The autocorrelation spectrum or power spectral density of the optical beam fluctuations gives the frequency or speed of the fluctuations. Experiments have shown that the decorrelation time of atmospheric refractive turbulence fluctuations is on the order of 1–10 ms, so that the frequency of the fluctuations is on the order of a few hundred hertz or less [36,37].

The influence of refractive turbulence is to break a Gaussian mode laser beam into smaller speckles or individual coherent wavefront groups. The effect can be given by the atmospheric turbulence field coherence length, ρ_0, which indicates the approximate size of the interference speckles within the laser beam, and is given by [24]

$$\rho_0 = (1.09\, k^2 R C_n^2)^{-3/5} \tag{11}$$

Equation (9) yields the size of the speckles within a beam front. It is an important parameter for a heterodyne detection or photon counting detection system since it places a limit as to the effective telescope aperture size (approximately $3\rho_0$) that can be used in a system [38]. However, for most current FSO systems that use direct detection of the beam, this will affect only the number of speckle modes to be aperture-averaged within the receiver, which will then affect the aperture-averaged SNR. The total system SNR will be a combination of all the fluctuation SNR values, power-limiting SNR considerations, and averaging effects within the communication decision time. Additional research is required in this area to better understand the tradeoffs in the area of signal averaging and atmospheric effects, including mitigation through the use of multiple wavelengths, beams, detectors, and temporal samples.

The beam wander due to refractive turbulence can be given by

$$\sigma_d^2 = 0.97 C_n^2 D^{-1/3} R^3 \tag{12}$$

where σ_d^2 is the variance in the displacement of the beam axis in m^2 and D is the diameter of the initial beam [39,40].

The preceding equations can be used to estimate the approximate level of fluctuations of a projected laser beam under controlled or laboratory conditions. Usually, the values for the standard deviation of the fluctuations σ are on the order of 0.05–0.7 (i.e., 5–70%), depending on the values of C_n^2 used. However, experience has shown that in many experimental and field-site cases complex windflow patterns exist (which are not energy-conserving) along with other atmospheric inhomogeneities that can cause beam drift and local absorption. As such, it is usually hard to accurately predict the fluctuation level in a particular setup. In this case, one has to resort to actual measurements of the fluctuation variance levels σ^2, in order to compare different experimental systems.

Under normal circumstances, the fluctuation variance level can be related to the information content signal-to-noise ratio (SNR) by [41]

$$\mathrm{SNR} = \frac{1}{\sigma^2} \tag{13}$$

where σ^2 represents the averaged or processed normalized variance measured over the time interval used to determine SNR of a decision period (possibly of one data bit or multiple bits for a coded word). Usually, if signal averaging is used, this has the effect of reducing the estimate of the variance by the square root of the number of samples averaged:

$$\mathrm{SNR}_n = \mathrm{SNR}_1\, n^{1/2} = \frac{n^{1/2}}{\sigma^2} \tag{14}$$

where n is the number of samples integrated, SNR_n is the SNR for n samples, and SNR_1 is the SNR for a single sample. Equation (14) is valid for an ergodic process that has random noise, but has to be reduced if nonrandom noise or processes are present [41]. This is true for large-scale attenuation processes in the infrared, where long-term temporal drifts in attenuation due to water vapor and other phenomena may be present. However, in the visible and near IR, the major noise sources are usually random.

Equation (14) is the general relationship for a signal detection process and helps relate the influence of increased fluctuation levels on SNR, which then directly affects the BER via Fig. 12. However, in the case of a FSO system, the fluctuation levels should more properly be measured only over the decision time of a bit. As such, the short-term variance measured over a time period of a data bit period may have to be used.

As can be seen in Eq. (14), the SNR can be improved through averaging multiple signals within the information decision period. This can be seen in an excellent research study by Kim et al., who studied the effect of refractive-turbulence fluctuations through the use of multiple laser beams [42]. Figure 16 shows the fluctuation levels measured over a 20-s time period for a 1.5-μm, 1.2-km FSO system that used one, two, and three separate laser beams. As can be seen, the use of several laser beams reduced the fluctuation levels that had the effect of increasing the measured SNR.

While the preceding equations indicate the level of fluctuations, it is customary in the FSO community for such an effect to be accounted for by introducing a fluctuation, fade, or link margin that is derived more from experimental measurements than from theory. Such a link margin may be on the order of 15–20 dB, although such a high value often includes the desired SNR for a specified BER compared to that for a SNR of 1. For example, for a system with a BER of 10^{-9} and a NEP of the detector of 20 nW, the BER value is related to a SNR of 6 [28]. In this case, a fade link margin could be 7 dB (factor of 5), so that the required SNR would be 6 times that due to 7 dB, specifically, a value of 30, or 15 dB. This is the equivalent of solving the range equation for a SNR of 1, but using a link margin of 15 dB. Both analysis types are used in the literature and are equivalent.

Finally, another atmospheric phenomena related to deep fades in the communication link has received considerable study [43,44]. For a FSO laser-based system transmitting over moderate to long ranges of 2–10 km, the fluctuations observed have the traditional refractive turbulence frequency fluctuations of up to a few hundred hertz, but long-term fades lasting 0.1 s to a few seconds are also observed. These are difficult to extract from similar effects due to tracking drifts or from building sway.

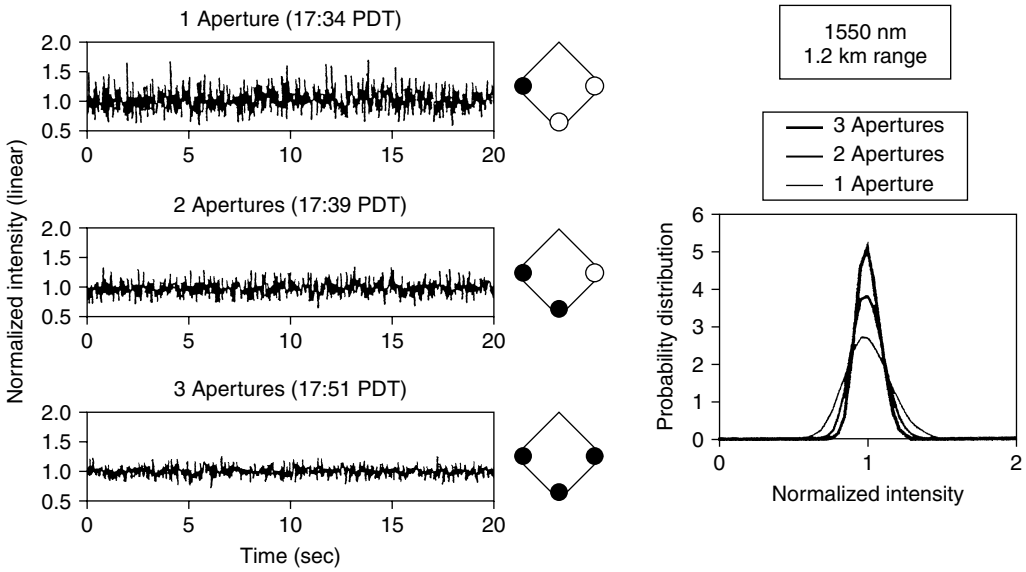


Figure 16. Intensity and fluctuation distribution measured for a 1.2-km FSO 1.55-μm system using one, two, and three laser beams. The reduction in signal fluctuations using three beams is easily seen. (Reproduced with permission of I. I. Kim, M. Mitchell, and E. Korevaar, SPIE Vol. 3850, 1999.)

However, some studies suggest that they may also be due to beam bending due to the presence of spatial localized concentrations of water vapor that move into the beam path. Such water vapor spatial clouds have been observed with high resolution Raman lidar (laser IR radar) systems, and indicate water vapor "clouds" of 10–100 m in diameter and movement times on the order of 1–100 s [45]. On the other hand, long-term fading can occur if the coherent beam produces a single speckle at the receiver and the optical alignment is such that destructive interference occurs (due to long-term building sway, beam wander, etc.). In this case, the intensity received is close to zero. Sometimes, these fades have values of up to 10–30 dB, suggesting partial destructive interference of the speckle. The effects of these fades can be mitigated through the use of multiwavelength, multispatial, or multitemporal samples within the link bit decision period in order to produce independent samples of the transmitted bit.

8. TELESCOPE DESIGN, TRACKING/ALIGNMENT DETECTORS, AND ENVIRONMENT

The telescope and receiver lens used in a FSO system is usually a compromise between the use of wide apertures for greater light gathering, short focal length for ease of handling, moderate field of view for ease of alignment, and optical coatings for narrow spectral filters for daytime use. This can be seen in Fig. 17, which shows several typical telescope configurations. Common configurations for a FSO system is either the use of a single lens and a

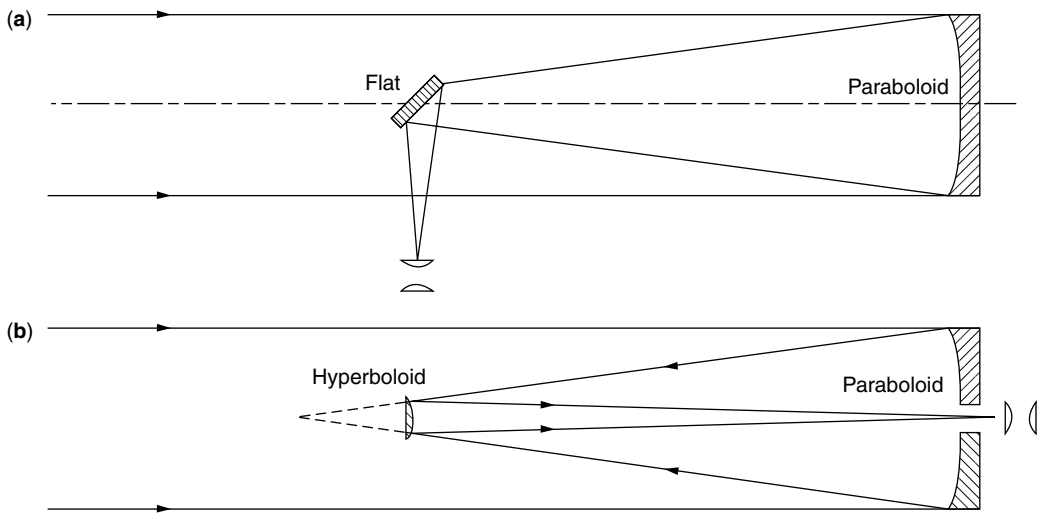


Figure 17. Schematic of Newtonian (**a**) and Cassegrainian (**b**) telescopes.

detector or the use of a Cassegrain telescope and a detector system that folds the beam path so that the telescope is shorter than that of a Newtonian configuration. An example of a short Cassegrain-type telescope used for FSO is one developed by Kaiser Electro-Optics in their f#/0.67 Hyperscope transceiver with an 8 in. aperture and 4 in. overall length. In this telescope a coated front reflecting window is used as the secondary mirror as in a Mangin mirror telescope configuration, which shortens the overall length of the telescope by a factor of ~2 [46,47].

Another method is to use a holographic lens or mirror arrangement to also ensure spectral and position placement that is similar to combining a diffraction grating with a lens [48]. Such a holographic lens is flat. Another possibility is that the holographic lens can be configured to displace or rotate the focused light around the perimeter of the lens as a function of input direction.

A common configuration for an inexpensive telescope systems is to use either a conventional optical lens or a flat Fresnel lens. In the latter case, the focal volume and resolution are not as good as with a conventional lens, but may be sufficient since in many cases the beam divergence and field of view of the transceiver have been enlarged to ease optical alignment of the system.

Alignment and tracking of the FSO system can be a difficult problem for longer-range systems. For short ranges under 200–500 m, the alignment and beam wander of the communication beam may be small enough that only coarse mechanical alignment is required during initial setup. For longer ranges, however, usually some form of active alignment and tracking is required. Two axis gimbals on the telescopes or the entire FSO unit are often used in this regard, as well as active positioning of the detector in the x–y image plane of the telescope. Fluctuation times on the order of 1–10 ms are required for atmospheric fluctuation, while building sway may have resonance times on the order of 1–10 s. Temperature- and wind-driven gross movement of buildings will have times on the order of hours. It is common to use quad (4) detectors and/or CCD cameras to monitor a separate alignment beam in order to actively track and compensate such movement.

Weather/environment protection of a FSO system is determined by the indoor or outdoor use of the system. Several systems are made for indoor use and require little environmental design. Some units are made for use behind the windows of an office and use the window and benign office conditions to separate them from harsher outside conditions. Outdoor use often entails hermetically sealing the FSO unit, with the placement of sunshields, rain canopies, and heaters on the telescope lens to reduce frost or condensation.

9. LASER EYE SAFETY

Any laser beam can cause damage to the human eye if it is operating with an irradiance (W/cm^2) that is above a certain level. The minimum permissible exposure level (MPE) is tabulated in the ANSI standards [49]. The standards are written for direct ocular view by the eye, and are given as a function of wavelength. This latter part is important because for wavelengths less than ~1.4 μm (~1400 nm), the optical radiation that enters the eye is focused and increased in irradiance onto the retina. For wavelengths longer than ~1.4 mm, the light is absorbed by the cornea and vitreous humor inside the eye. The reader should remember that it is possible for a coherent laser beam to be focused by a lens down to the diffraction limited spot size given by $f\#\,\lambda$, where the f-number, f#, of the lens is equal to the diameter of the lens (or pupil of the eye) divided by the focal length of the lens; as can be appreciated, for the human eye this spot size on the retina is close to the wavelength of light: about 1 μm in diameter. Such a focusing effect is already taken into effect by the ANSI standards.

While the reader should use the ANSI standards for each explicit situation, a general idea of the eye-safety values can be shown as in Table 3. Table 3 shows the direct ocular MPE values, and the approximate calculated maximum transmitted laser power levels for a FSO transmitter that delivers a 30 × 30-cm beam at a distant receiver where the ocular viewing would occur; this corresponds to the size of a 1° divergence laser beam after being transmitted 200 m through the atmosphere. The MPE is shown for a 10-s exposure. The maximum transmitted laser power is seen to be about 1 W for the 0.7-μm laser and about 90 W for the 1.55-μm system. Of course, these values are much lower if the eye is at a closer range and intercepts the beam where the beam is smaller and the irradiance is higher. In addition, the effect of atmospheric turbulence may increase the fluctuating intensity levels so that the values would need to be reduced appropriately. The eye safety for a LED is higher since it is a noncoherent source and will not be focused to a small diffraction-limited spot on the retina.

Most FSO systems are designed to be eye-safe, or to operate where a human will not intercept the beam. In the case where humans may intermittently intercept the beam, other warning systems such as the use of an inexpensive microwave radar system (e.g., marine radar) could be used to detect the presence of a human within the FSO beam path and shut down the system temporarily.

Finally, laser safety regulations and standards have been instituted as to the manufacturing classification of the laser, such as class IV or class IM, which covers the power levels and operational safety standards for the laser. The international organizations such as the International Electrotechnical Commission (IEC) and U.S.

Table 3. Approximate Maximum Permissible Exposure (MPE) Power in W/cm^2 for Direct Ocular (Eye) View for Several Different Wavelengths

λ (μm)	MPE (mW/cm^2)	P_t (Maximum Transmitted Power) – 200 m away (w)[a]
0.7	1	0.9
0.9	25	22
1.55	100	90
10	100	90

[a]The maximum transmitted laser power calculated is for the case of a beam size of 30 cm in diameter at the position of the eye (appropriate for a propagation distance of 200 m and beam divergence of 1 degree).

agencies (FDA, Laser Institute of America, and ANSI) have helped coordinate these classification schemes. For example, the IEC has expanded its coverage of TC 76/60825 part 12, Working Group 5 (WG5) to cover safety and transmission issues associated with laser and LED FSO communication [50]. The WG5 committee has been working on a draft titled *Part 12: Safety of Free Space Optical Communications Systems Using Directed Beams* to be released in 2003. The interested reader is encouraged to review this information in the references. They are not of concern for the scientific analysis of a FSO system, but are very important for the manufacturing and proper use of commercial FSO systems.

10. FSO SYSTEM AND ENGINEERING TRADE-OFFS

As can be seen from the above discussions, there are many scientific aspects of FSO design. In addition, there are a considerable number of system and engineering trade-offs that also have to be looked at. It is beyond the scope of this article to discuss this in detail, but some general aspects can be listed. Some of the trade-off parameters that need to be taken into account include (1) modulation of the laser or LED (direct modulation through the power supply or the need for an expensive external modulator), (2) detector bandwidth and cooling requirements in the case of IR detectors, (3) increased laser beam divergence and possible need to increase laser power versus increased cost of using active alignment of a narrow laser beam, (4) cost of laser or LED system at different wavelengths versus advantages of availability of cheaper detector components versus penetration of beam through fog or rain, and (5) eye safety versus laser beam size versus divergence of beam and beam size at detector telescope. As can be seen, there are a significant number of engineering trade-offs that have to be made in any FSO system. Although the above list of trade-offs seems formidable, it is not really as bad as it first appears. This is because there are many different ways to build a successful FSO system for a specified operating condition. As such, there is no "one" or ultimate maximized system, but rather several that are sufficient to provide the communication link required. The most basic, first-level trade-off studies are often conducted to provide a high level of reliability and link BER for the specified atmospheric conditions and system environmental conditions. Then, within these broad constraints, one finds that there are usually several approaches that will meet the requirements.

11. FUTURE TECHNICAL AND INDUSTRY CONSIDERATIONS

FSO is just starting to impact the Internet "last- mile" interconnectivity problem. It is felt that it may offer the unlimited bandwidth solution for this problem within the metro urban core involving downtown building-to-building communication, but may also be a major technology for home-to-home and office-to-office connectivity. As stated earlier, if the home/business last-mile connectivity becomes the main technical driving force for communication, then the optimization of the technical specifications for FSO may become more important than using laser wavelengths and transceivers as part of the current fiberoptic communication legacy. In that case, wavelength issues and tradeoffs between long-range point-to-point versus short-range mesh-net connectivity will be addressed and operational standards will be set by the industry. It is common to hear that FSO solves a technical problem at present, but that the industry does not yet recognize or fully understand the potential that FSO offers. As such, some of the current problems in the deployment of FSO is in the industry perception and marketing of this technology. FSO systems have now shown that they are reliable (99.9% to 99.999%) communication channels that have fast bandwidth, are easy to set up, and provide cost-effective solutions. The FSO community recognizes these concerns, and has launched the Free Space Optics Alliance organization [51]. The FSO Alliance currently consists of about 25 companies, and has a mission to educate and promote FSO technical information to the communication industry as a whole and the print and journal media sector. It is believed that through such industry wide education, that industry standards and proper growth of FSO technology will occur within the communication carrier industry.

It is believed by the author that FSO is on the verge of changing the basic communication medium and technology of the metro and last-mile network market. The challenges and importance of setting standards for FSO are becoming increasingly clear in order to help the field become a major component in the whole communication network. It is hoped that the reader has gained an appreciation of some of the technical challenges and opportunities that FSO offers in this regard.

Finally, the author would like to acknowledge several helpful discussions and suggestions from Drs. Issac Kim, David Rockwell, and John Schuster.

BIOGRAPHY

Dennis K. Killinger received the B.A. degree from the University of Iowa, M.A. degree from De Pauw University, and Ph.D. degree in Physics from the University of Michigan. He has conducted research on radar analysis and microwave atmospheric propagation while employed as a Research Physicist at the Naval Avionics Facility, and joined the research staff in Quantum Electronics at Lincoln Laboratory, Massachusetts Institute of Technology in 1978 conducting research in the development of new solid-state lasers and their application as spectroscopic lidar probes of the atmosphere. Since 1987 he has been a Professor of Physics at the University of South Florida and is a Distinguished University Professor and Director of the Laboratory for Atmospheric Lidar and Laser Communication Studies. Dr. Killinger is a Fellow of the Optical Society of America, Senior Member of the IEEE, past associate editor of *Applied Optics and Optics Letters*, past member of the NAS/NRC Committee on Optical Science and Engineering, and has served as chairman of several international conferences on lasers and applied spectroscopy. He has published over 200 technical papers, reports, and conference papers in laser remote sensing/lidar, applied laser spectroscopy, laser

physics, laser atmospheric propagation, and free-space optics (FSO) laser communication.

BIBLIOGRAPHY

1. E. J. Korevaar, ed., *Optical Wireless Communication II*, SPIE Vol. 3850, 1999.
2. G. S. Mecherle, ed., *Free-Space Laser Communication Technologies XII*, SPIE Vol. 3932, 2000.
3. E. J. Korevaar, ed., *Optical Wireless Communications III*, SPIE Vol. 4214, 2000.
4. *Harnessing Light: Optical Science and Engineering for the 21st Century*, Committee on Optical Science and Engineering (COSE) NRC Report, National Academy Press, Washington, DC, 1998.
5. W. Wolfe and G. Zissis, eds., *The Infrared Handbook*, 3rd ed., Environmental Research Institute of Michigan (ERIM) and SPIE, 1989.
6. M. Bass, E. Van Stryland, D. Williams, and W. Wolfe, eds., *Handbook of Optics*, 2nd ed., Optical Society of America, McGraw-Hill, 1995.
7. D. J. Petrovich, R. A. Gill, and R. J. Feldmann, US Air Force development of a high-altitude laser crosslink, in SPIE Vol. 4214, 2000, pp. 14–25.
8. I. I. Kim et al., Preliminary results of the STRV-2 satellite to ground lasercom experiment, in SPIE Vol. 3932, 2000, pp. 21–43.
9. S. Lee, J. W. Aleander, and M. Jeganathan, Pointing and tracking subsystem design for optical communications link between the international space station and ground, in SPIE Vol. 3932, 2000, pp. 150–157.
10. H. Willebrand and B. S. Ghuman, *Free-Space Optics: Enabling Optical Connectivity in Today's Networks*, Sams Publications, Indianapolis, 2002.
11. PlainTree, Inc., Ottawa, Ontario, Canada, *www.plaintree.com*.
12. Optical Access, Inc., San Diego, CA, *www.opticalaccess.com*.
13. fSONA, Inc., Richmond, BC, Canada, *www.fsona.com*.
14. G. Nykolak et al., A 40 Gb/s DWDM free space optical transmission link over 4.4 km, in SPIE Vol. 3932, 2000, pp. 16–20.
15. J. Hecht, *The Laser Guidebook*, 2nd ed., McGraw-Hill, New York, 1992.
16. *Laser Focus World Buyers' Guide*, Pennwell Publications, 2002; *www.laserfocusworld.com*.
17. *Photonics Spectra Buyers' Guide*, Laurin Publication, 2002.
18. G. Scamarcio et al., High power infrared (8 μm wavelength) superlattice lasers, *Science* **276**: 773–776 (1997).
19. P. Mamidipudi and D. Killinger, Optimal detector selection for a 1.55 micron KTP OPO atmospheric lidar, in SPIE Vol. 3707, 1999, pp. 327–335, and references cited therein; A. E. Seigman, *Lasers*, University Science Books, Mill Valley, CA, 1986.
20. R. M. Goody and Y. L. Young, *Atmospheric Radiation*, Oxford Univ. Press, 1989.
21. R. T. Menzies and D. K. Killinger, IR Lasers tune into the environment, *IEEE Circuits Devices* **10**: 24–29 (1994).
22. L. S. Rothman et al., The HITRAN molecular database: Editions of 1991 and 1992, *J. Quant. Spectrosc. Radiat. Transfer* **48**: 734 (1992).
23. HITRAN, FasCode, HITRAN-PC, and PCTRAN computer programs; ONTAR Corp., 9 Village Way, North Andover, MA 01845-2000; Website *www.ontar.com*.
24. D. K. Killinger, J. H. Churnside, and L. S. Rothman, Atmospheric Optics, in M. Bass, ed., *OSA Handbook of Optics*, 1995, Chap. 44.
25. R. Measures, *Laser Remote Sensing*, Wiley-Interscience, New York, 1984.
26. I. I. Kim, B. McArthur, and E. Korevaar, Comparison of laser beam propagation at 785 nm and 1550 nm in fog and haze for optical wireless communication, in SPIE Vol. 4214, 2001, pp. 26–37.
27. E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*, Wiley, New York, 1996.
28. R. H. Kingston, *Optical Sources, Detectors, and Systems: Fundamentals and Applications*, Optics and Photonics; Academic Press, New York, 1995; R. H. Kingston, *Detection of Optical and Infrared Radiation*, Springer, New York, 1978.
29. J. S. Accetta and D. L. Shumaker, eds., *The Infrared and Electro-Optical Systems Handbook*, Environmental Research Institute of Michigan, Ann Arbor, MI, 1993.
30. P. R. Norton, Photodetectors, in *OSA Handbook of Optics*, McGraw-Hill, New York, 1995, Chap. 15, pp. 15–16.
31. B. R. Strickland, M. J. Lavan, E. Woodbridge, and V. Chan, Effects of fog on the bit error rate of a free-space laser communication system, *Appl. Opt.* **38**: 424–431 (1999).
32. D. L. Fried and J. B. Seidman, Laser beam scintillation in the atmosphere, *J. Opt. Soc. Am.* **57**: 181–185 (1967).
33. V. I. Tatarskii, *The Effects of the Turbulent Atmosphere on Wave Propagation*, Israel Program for Scientific Translations, Jerusalem, 1971.
34. L. C. Andrews and R. L. Phillips, *Laser Beam Propagation through Random Media*, SPIE Press, 1998.
35. W. E. Wilcox, Jr., *Diurnal measurements of atmospheric optical turbulence with application to coherent lidar*, master's thesis, Dept. Physics, Univ. South Florida, Tampa, 1991.
36. G. Nykolak et al., Update on 4 × 2.5 Gb/s, 4.4 km free-space optical communications link: Availability and scintillation performance, in SPIE Vol. 3850, 1999, pp. 11–19.
37. N. Menyuk and D. Killinger, Temporal correlation measurements of pulsed dual CO_2 lidar returns, *Opt. Lett.* **6**: 301–303 (1981).
38. K. P. Chan and D. K. Killinger, Enhanced detection of atmospheric-turbulence distorted 1 micron coherent lidar returns using a two-dimensional heterodyne detector array, *Opt. Lett.* **16**: 1219–1221 (1991).
39. J. H. Churnside and R. J. Lataitis, Wander of an optical beam in the turbulent atmosphere, *Appl. Opt.* **29**: 926–930 (1990).
40. I. I. Kim et al., Wireless optical transmission of fast Ethernet, FDDI, ATM, and ESCON protocol data using the TerraLink laser communication system, *Opt. Eng.* **37**: 3143–3155 (1998).
41. N. Menyuk, D. K. Killinger, and C. R. Menyuk, Limitations of signal averaging due to temporal correlation in laser remote sensing measurements, *App. Op.* **21**: 3377–3383 (1982).
42. I. I. Kim, M. Mitchell, and E. Korevaar, *Measurement of scintillation for free-space laser communication at 785 nm and 1550 nm*, in SPIE Vol. 3850, 1999, pp. 49–62.

43. P. Polak-Dingels, P. R. Barbier, D. W. Rush, and M. L. Plett, Long-term fading statistics measurements of an atmospheric optical communication channel, in SPIE Vol. 3850, 1999, pp. 40–48.

44. C. C. Davis et al., Characterization of a liquid filled turbulence simulator, SPIE Conf. Artificial Turbulence and Wave Propagation, July 1998.

45. D. N. Whiteman, S. H. Melfi, and R. A. Ferrare, Raman lidar system for the measurement of water vapor and aerosols in the earth's atmosphere, *Appl. Opt.* **31**: 3068–3082 (1992).

46. *Hyperscope FSO Telescope*, Kaiser Electro-Optics, Rockwell-Collins, *www.keo.com*.

47. T. Carbonneau and G. S. Mecherle, *SONAbeam optical wireless products*, in SPIE Vol. 3932, 2000, pp. 45–51.

48. *Holographic Rotating Telescope: HARLIE Lidar Technology Program*, NASA Goddard, Greenbelt, MD; *http://bll.gsfc.nasa.gov/harlie*.

49. *American National Standard for Safe Use of Lasers*, ANSI Z136.1 - 2000, Laser Institute of America, Orlando, FL, 2000; *http://www.laserinstitute.org*.

50. D. Britz, *Free Space Optical Communication: A New Broadband Access Technology with Implications for Laser Safety in the Public Sector*, Laser Institute of America Newsletter, Jan./Feb. 2002, pp. 1–7; *http://www.laserinstitute.org*.

51. Free Space Optics Alliance, *http://www.fsoalliance.com*.

ORTHOGONAL FREQUENCY-DIVISION MULTIPLEXING

LEONARD J. CIMINI JR.
LARRY J. GREENSTEIN
AT&T Labs-Research
Middletown, New Jersey

1. INTRODUCTION

The permissible data rate of a digital communications link is limited by the available bandwidth and also by power and noise. The data rate can also be limited by phenomena in the communications medium (channel) between the transmitter and the receiver, especially by intersymbol interference (ISI) caused by time dispersion of the transmission medium, such as occurs on the multipath radio channel and the frequency-selective telephone channel.

As a general rule, the effects of ISI are small as long as the time extent of the channel impulse response is significantly shorter than the duration of a transmitted symbol. This implies that the symbol rate transmitted over a dispersive channel is practically limited by the channel's memory. However, mechanisms exist for countering ISI and thus extending symbol rates. These include receiver equalization, transmitter preequalization, and some forms of radio diversity. All are aimed at permitting the transmission of datastreams with symbol periods comparable to, or even smaller than, the channel's memory.

An alternative approach employs multiple carriers. In multicarrier transmission, the datastream to be transmitted is split into multiple parallel datastreams of reduced rate, and each of them is transmitted on a separate frequency (or *subcarrier*). Each subcarrier is modulated at a rate so low (or, equivalently, has a symbol period so long) that dispersion does not cause a problem. A given subcarrier, with its associated data signal, constitutes a *subchannel*. Ideally, the bandwidth of a subchannel would be so narrow as to preclude any ISI. More realistically, there will be *reduced* ISI on each subchannel, which can be either tolerated or easily corrected. Since the system's data throughput is the sum of the throughputs of the parallel subchannels, the data rate per subchannel is only a fraction of that of a single-carrier system having the same throughput. Thus we see that multicarrier transmission permits high data rates while maintaining symbol durations much longer than the channel's memory.

At the same time, the subchannels must be spaced, and spectrally shaped, to ensure that they do not interfere with each other. Such precautions can limit spectral efficiency, defined as the total bit rate divided by the total bandwidth. *Orthogonal frequency-division multiplexing* (OFDM) [1–4] is a special case of multicarrier transmission that permits the subchannels to overlap in frequency without mutual interference. In addition to improved spectral efficiency, this technique exploits digital signal processing technology to obtain a cost-effective means of implementation. Our primary aim here is to detail the theory and practice of this form of multicarrier transmission.

Before proceeding with the basics, we pause to note the numerous communications systems, past and present, that have used some form of multicarrier transmission. The first systems using this technique were designed in the late 1950s and early 1960s for military high-frequency radio applications [5,6]. These included the Kineplex and Kathryn systems. Since these early systems, multicarrier transmission (and in particular OFDM) has been used over many different communications media. Practical interest has increased partly as a result of enabling advances in signal processing and microelectronics, and partly because of the demand for ever-higher data-rate services over dispersive channels. Multicarrier modems have been standardized in different parts of the world for both wireline and wireless data applications, including digital audio/video broadcasting (DAB/DVB) [7,8]; digital transmission over copper wire, for example, digital subscriber loops (DSLs) [9]; wireless local-area networks (WLANs) [10]; and have been proposed for mobile radio applications [11].

2. BASIC CONCEPTS

2.1. Multicarrier Transmission

There are several techniques for realizing a multicarrier link. In the conceptually simplest approach, the total signal frequency band is divided into N ideally nonoverlapping (band-limited) frequency subchannels, employing N independent transmitter–receiver pairs. A block diagram description of how this can be done is given in Fig. 1. In the transmitter (Fig. 1a), an input stream of data, at rate R bits per second (bps), is divided into N

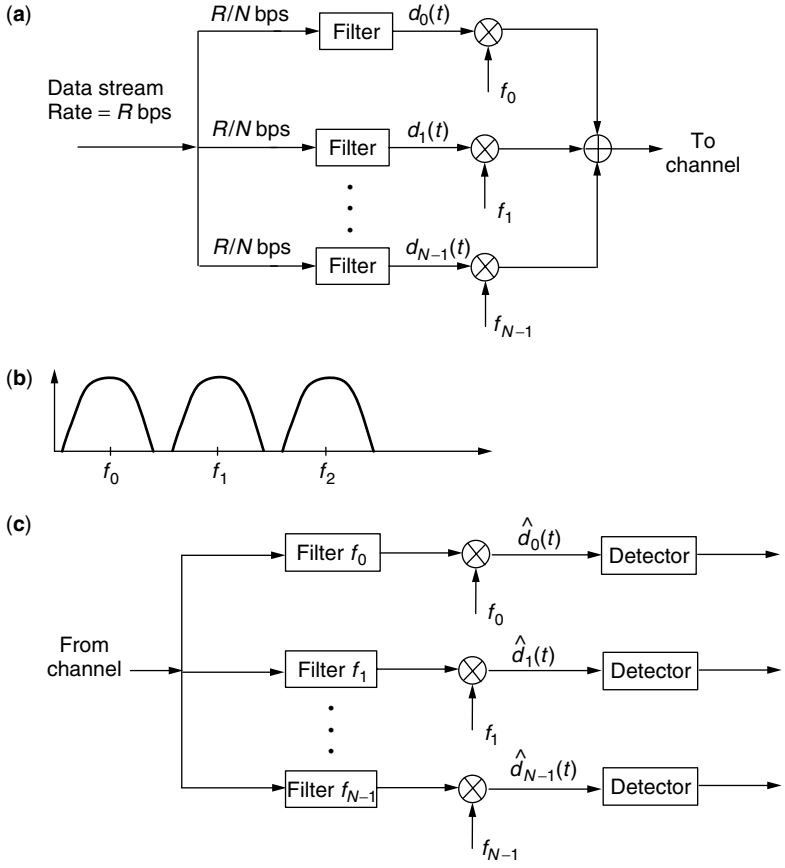


Figure 1. Multicarrier transmission system: (**a**) multicarrier transmitter; (**b**) transmit spectrum; (**c**) multicarrier receiver.

parallel substreams, each at data rate R/N bps. (The data values in the mainstream and the substreams are, in general, complex, and the real and imaginary components can be binary or multilevel.) Each substream is passed through a baseband pulseshaping circuit ("filter"), where we assume identical filters for all substreams. The kth filter output ($k = 0, 1, \ldots, N-1$) is then upconverted by a balanced mixer to frequency f_k. The result is a subcarrier with *quadrature amplitude modulation* (QAM). The N-QAM signals are combined (frequency-multiplexed) and sent over the channel. An example of the output signal spectrum is given in Fig. 1b. In the receiver, Fig. 1c, a set of bandpass filters centered on f_k, $k = 0, 1, \ldots, N-1$, is used to frequency-demultiplex the N subchannels, after which each subchannel is downconverted to baseband by a balanced mixer. Each substream is then applied to a detector, and the output data values are sent on for possible further processing. The spectral guard bands shown between subchannels in the figure are introduced so that easily realizable filters can be used in the receiver.

While the advantages of multicarrier transmission in terms of reduced sensitivity to dispersion are obvious, there are two major disadvantages to this particular realization. First, it is spectrally inefficient, since the signals must be sufficiently spaced in frequency to facilitate separation at the receiver. Second, a receiver with a large bank of filters may be prohibitive in terms of complexity and cost. The alternative approach (OFDM), using overlapping subchannels (to improve the spectral efficiency) and efficient digital signal processing techniques (to reduce the complexity and cost), is described next.

2.2. Basic OFDM

Orthogonal frequency-division multiplexing provides a solution to the disadvantages of conventional multicarrier transmission. In particular, a more efficient use of bandwidth can be obtained if the spectra of the individual subchannels are permitted to overlap, with specific orthogonality constraints imposed to facilitate separation of the subchannels at the receiver. Figure 2 shows the spectra for the two alternative forms of multicarrier transmission.

To analyze either form of multicarrier signal, we denote the symbol rate of the original data sequence by f_s, where $T_s = 1/f_s$ is the original symbol period. After serial-to-parallel conversion, there are N parallel data sequences, each with symbol rate f_s/N and symbol period $T = NT_s$. Thus, each subchannel is tolerant of N times as much time dispersion as would be the original datastream.

Now assume that, in a given symbol period $[0, T]$, the N subchannels carry data values $D_0, D_1, \ldots, D_k, \ldots, D_{N-1}$. We assume that D_k is two-dimensional, that is, $D_k = A_k + jB_k$, where A_k and B_k are real numbers representing the in-phase and quadrature data components, respectively. The set of discrete values possible for each component depends solely on the chosen data constellation, for example, $A_k = \pm 1$ and $B_k = \pm 1$ for four-level

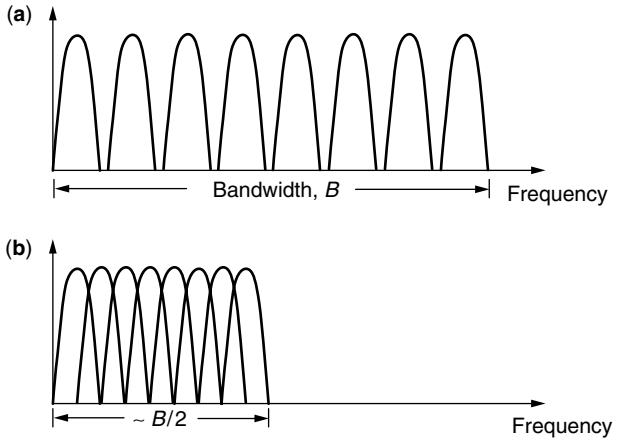


Figure 2. Transmit spectra of multicarrier and OFDM signals: (**a**) multicarrier spectrum; (**b**) OFDM spectrum.

quadrature amplitude modulation (4-QAM), also called *quadrature phase shift keying* (QPSK). Finally, assume initially that the data values are carried by rectangular pulses, that is, that the kth subchannel data value is carried by a pulse that is 1 on $[0, T]$ and 0 elsewhere. Then, the multicarrier signal transmitted on the given symbol interval can be represented as

$$s(t) = \mathrm{Re}\left\{\sum_{k=0}^{N-1} D_k e^{j\omega_k t}\right\}, \quad 0 \le t \le T \tag{1}$$

$$= \sum_{k=0}^{N-1} [A_k \cos \omega_k t - B_k \sin \omega_k t], \quad 0 \le t \le T \tag{2}$$

where the subcarrier radian frequency is $\omega_k = 2\pi f_k$, with $f_k = f_0 + k\Delta f$. The offset frequency, f_0, could represent the carrier frequency in a passband transmission system, such as one using a wireless channel, or could be adjusted for baseband transmission. Also, for baseband transmission, the data could be chosen in a symmetric fashion to guarantee a real output. This latter situation is discussed in Section 2.4. The parameter Δf represents the subcarrier spacing, which we discuss next.

The structure in Fig. 3 represents a general form of a multicarrier transmitter. For OFDM, the subchannels are permitted to spectrally overlap. To enable separation of these channels at the receiver, the data pulses for every pair of subchannels must be mutually orthogonal. For rectangular pulses, this can be achieved by relating the subcarrier spacing and the symbol duration via $\Delta f = 1/T$. Under these conditions, a simple correlation for each subchannel (i.e., multiplication by the appropriate waveform followed by integration over the symbol period) can separate out the subchannels. This receiver structure is shown in Fig. 4.

The power spectral density of the transmitted OFDM signal is the sum of the power spectral densities of N separate QAM signals at N subcarrier frequencies separated by the signaling rate. For rectangular symbol pulses, the Fourier transform of the symbol in each subchannel is a shifted version of $\sin x/x = \mathrm{sinc}(x)$, with nulls at the centers of the other subchannels. These and other spectral properties of an OFDM signal with rectangular symbol pulses are illustrated in Fig. 5. The Fourier transform of a single pulse in a single subchannel is shown in Fig. 5a; a set of Fourier transforms corresponding to eight subchannels ($N = 8$) is shown in Fig. 5b, and the OFDM power spectral density is shown in Fig. 5c for $N = 64$ and 256. (For convenience, we display the out-of-band portion in each case as the *envelope* of the actual lobe structure.) For large N, the total power spectral density is essentially flat in the bandwidth containing the subcarriers, and only the subchannels near the band edge contribute to the out-of-band power. Therefore, as the number of subcarriers becomes large, the spectral compactness approaches that of single-carrier modulation with rectangular bandpass filtering.

We now express the above ideas mathematically. Using the complex envelope of the transmitted signal, a single OFDM symbol can be represented as

$$S(t) = \sum_{k=0}^{N-1} D_k e^{j\omega_k t} \times \mathrm{rect}\left(\frac{t}{T}\right) \tag{3}$$

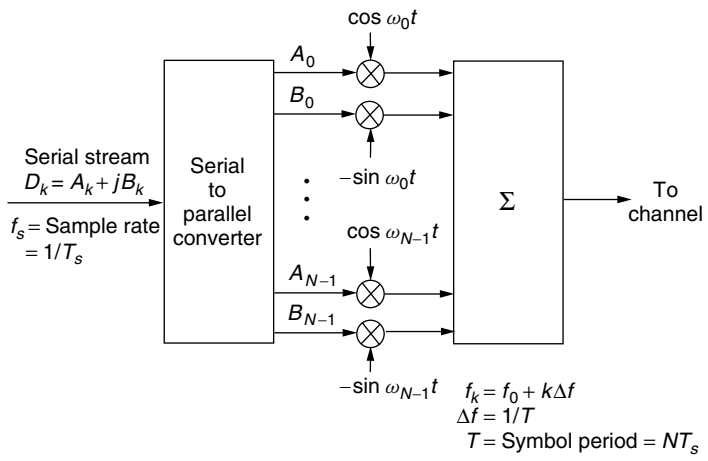


Figure 3. OFDM transmitter.

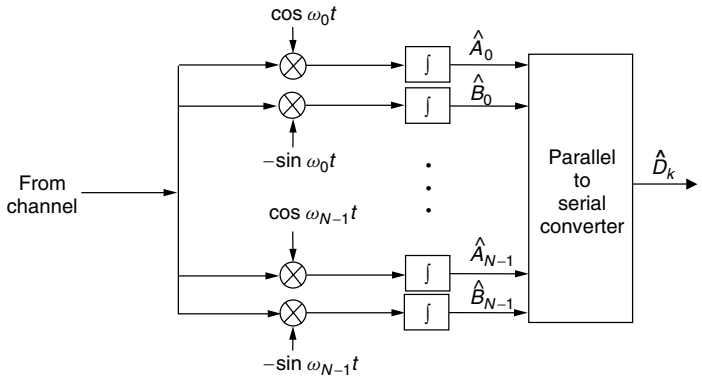


Figure 4. OFDM receiver.

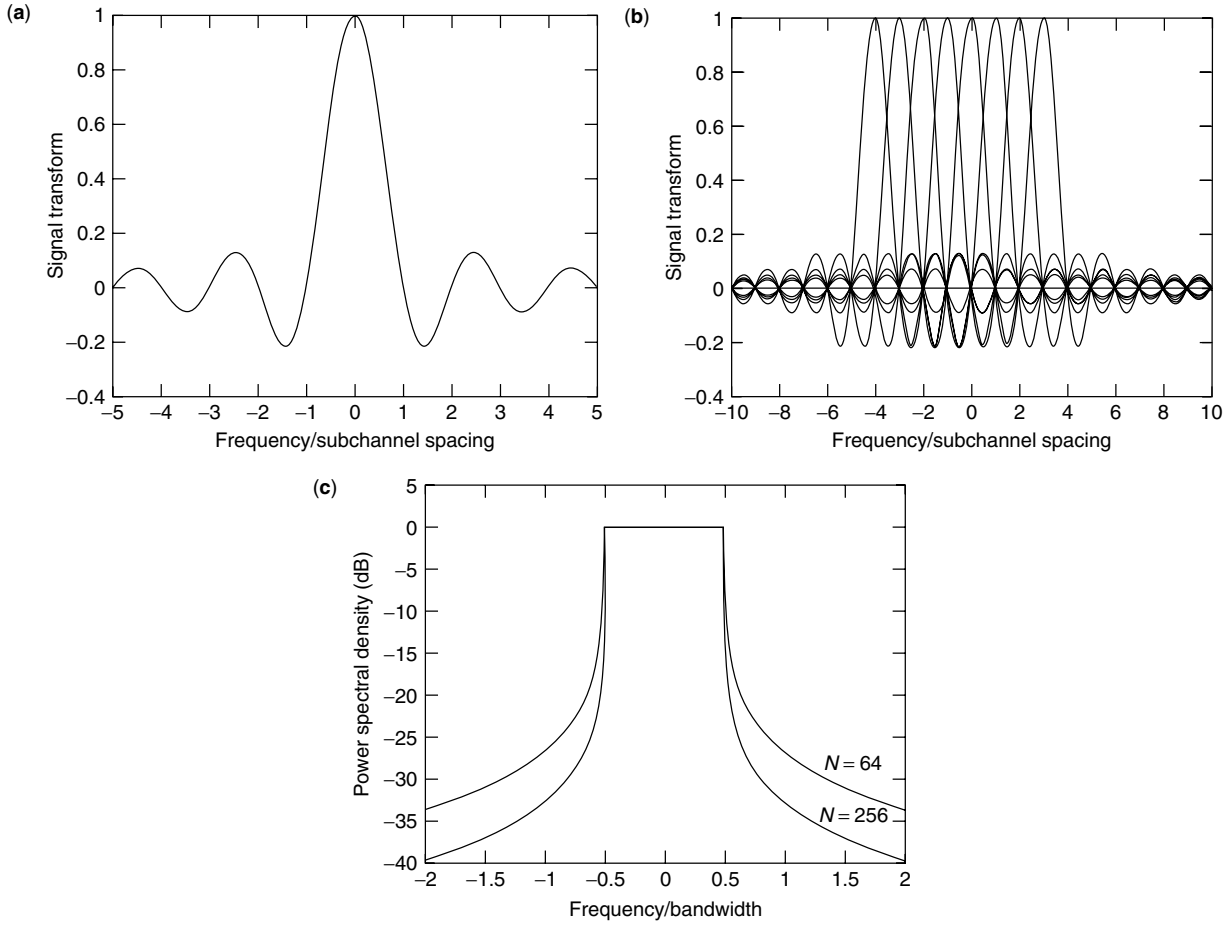


Figure 5. Spectral properties of an OFDM signal: (**a**) single channel; (**b**) eight individual subchannels; (**c**) Spectral power density (referred to center frequency).

where rect (x) is defined as having the value 1 on [0,1] and 0 elsewhere. The Fourier transform of $S(t)$ is thus given by

$$P(f) = T \sum_{k=0}^{N-1} D_k \frac{\sin \pi \left(\dfrac{f - f_k}{\Delta f} \right)}{\pi \left(\dfrac{f - f_k}{\Delta f} \right)} \tag{4}$$

If the data symbols are mutually independent (both among symbols and among subchannels), the power spectral density of the OFDM signal is $\overline{|P(f)|^2}/T$, where the overbar denotes averaging over the data. This formulation was used to obtain the results in Fig. 5c, where $f = 0$ corresponds to the center frequency of the OFDM spectrum. Note that the sharpness of the spectral falloff outside the main bandwidth increases with N.

We now show the orthogonality of the N transmitted pulses. Assuming, at this point, a perfect and noiseless channel, we can also regard $S(t)$ in Eq. (3) as the received signal. To detect the kth data value, D_k, $S(t)$ is multiplied by $e^{-j\omega_k t}$ and then integrated over $[0, T]$. The received data

symbols at the output of the kth correlator are

$$\hat{D}_k = \int_0^T S(t)e^{-j\omega_k t}\,dt = \sum_{l=0}^{N-1} D_l \int_0^T e^{j2\pi\Delta f(l-k)t}\,dt \tag{5}$$

For the case $\Delta f = 1/T$, it is easily shown that

$$\int_0^T e^{j2\pi\Delta f(l-k)t}dt = \delta(l-k) \tag{6}$$

where we use the Kronecker delta function, $\delta(l-k) = 1$ when $l = k$ and 0 otherwise. Therefore, $\hat{D}_k = D_k$, and, so, even though the subchannels overlap, they can be separated at the receiver with no interference among subchannels; that is, the subchannels are orthogonal.

We note that deriving the multicarrier transmitted signal from the data sequence, Eq. (5), and detecting that sequence from the received signal, Eq. (6), involves operations that resemble Fourier transforms. We will show more formally that the orthogonality that arises from setting $\Delta f = 1/T$ allows the use of the discrete Fourier transform (DFT) at both ends and thus the use of very efficient digital signal processing [12]. The combination of orthogonal pulses and efficient DFT processing constitutes the essence of OFDM. There are, of course, many details. The most important of these have to do with practical impairments in the transmission medium (notably, time dispersion and time variations) and in the system hardware (notably, frequency and timing errors in the receiver and amplifier nonlinearities in the transmitter). Discussions of basic implementation, channel and system impairments, and their remedies occupy most of the remaining sections.

2.3. DFT Implementation

We show here how the DFT and the inverse DFT (IDFT) can be used to implement OFDM. In most cases, these transforms can be done very efficiently by using the fast Fourier transform (FFT) algorithm. In this discussion, the number of subchannels and the FFT size are the same, N. (Later, we show why the FFT size is generally greater.) If N is a power of 2, the number of operations is on the order of $N\log_2 N$, as opposed to N^2 for conventional DFTs, leading to substantial savings for large N. For example, the number of FFT operations for $N = 1024$ is about 10^4, as opposed to about 10^6 with conventional processing, for a reduction of 100 to 1. Thus, completely digital implementations can be built around special-purpose hardware performing the FFT and its inverse (IFFT), replacing the banks of oscillators, mixers, and filters shown in Fig. 1.

Consider a discrete-time version of the complex envelope of the transmitted OFDM symbol in Eq. (3). Assuming $f_0 = 0$, without loss of generality, and sampling at times $t_n = nT_s$, Eq. (3) becomes

$$S_n = \sum_{k=0}^{N-1} D_k e^{j2\pi k\Delta f n T_s}, \quad 0 \le n \le N-1 \tag{7}$$

With the imposed orthogonality condition, $\Delta f = 1/T = 1/NT_s$, this becomes

$$S_n = \sum_{k=0}^{N-1} D_k e^{j2\pi kn/N}, \quad 0 \le n \le N-1 \tag{8}$$

which is simply the IDFT of the input data sequence, $D_0, D_1, \ldots, D_{N-1}$. With N suitably chosen, the transmitted signal samples can then be generated using the efficient IFFT algorithm.

The receiver correlation operations can also be performed in this fashion. Specifically, suppose that the block of received signal samples, $\{S_n\}$, is transformed in the receiver using a DFT. This yields

$$\begin{aligned}\hat{D}_k &= \frac{1}{N}\sum_{n=0}^{N-1} S_n e^{-j2\pi kn/N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1} D_l e^{j2\pi n(l-k)/N} \\ &= \frac{1}{N}\sum_{l=0}^{N-1} D_l \sum_{n=0}^{N-1} e^{j2\pi n(l-k)/N} \\ &= \frac{1}{N}\sum_{l=0}^{N-1} D_l N\delta(l-k) \\ &= D_k\end{aligned}$$

Thus, the correlation operations can also be efficiently implemented using the FFT algorithm.

We should mention at this point that there are several alternative forms of OFDM, that is, orthogonality can be achieved in various ways [13–15]. In particular, several of the early forms of OFDM were based on band-limited signaling, using specially designed pulses or special signaling patterns to guarantee orthogonality. Nevertheless, the form of OFDM described here is the most popular and the one proposed or implemented for all OFDM-based standards.

Finally, it is important to note how the sequence of IDFT samples in the transmitter, $\{S_n\}$, is converted into a continuous analog signal for transmission over the medium. The N samples, spaced in time by $1/N\Delta f = T_s$, are passed through a digital-to-analog converter (DAC) and then applied to a band-limiting filter. The spectrum of a discrete-time waveform such as the S_n sequence is periodic, with period $N\Delta f$. The purpose of the band-limiting filter is to pass one such period (the *primary spectrum*) and suppress all others. The burden on this filter can be severe if the primary spectrum has significant content at the band edges, as is generally the case (Fig. 5). This situation is avoided, and the filtering problem eased, by "padding" the original block of data values, $\{D_k\}$, with zeros before and/or after the actual data values; that is, zero-valued subcarriers are added to the data-carrying subcarriers. Thus, the band-limiting filter does not require as sharp a cutoff characteristic. This padding creates a difference between the number (N) of data-carrying subcarriers and the total number (N') of subcarriers

processed by the FFT. The FFT size, N', can readily be chosen to be a power of 2; we will assume hereafter that this is the case, so that the central transmitter and receiver processings are IFFTs and FFTs, respectively. Note that the bandwidth to be transmitted is still $N\Delta f$, but the samples in time, S_n, are now spaced by $1/N'\Delta f$ (oversampling) and the spectrum period is $N'\Delta f > N\Delta f$.

2.4. Baseband versus Passband Representations

For a given block of data values $\{D_k\}$, the complex envelope of the OFDM signal is $S(t)$ in Eq. (5). For passband transmissions (as, e.g., in wireless applications), a stage of modulation is needed to convert $S(t)$ to a real passband signal. For baseband transmission (as in wireline applications like DSL), no modulation is needed but $S(t)$ must be a real baseband signal. This can be done by converting the N-symbol stream $\{D_k\}$ to a stream of $2N$ symbols according to the following rule:

$$D'_k = \begin{cases} \mathrm{Re}(D_0) & k = 0 \\ D_k & k = 1, 2, \ldots, N-1 \\ \mathrm{Im}(D_0) & k = N \\ D^*_{2N-k} & k = N+1, \ldots, 2N-1 \end{cases} \tag{9}$$

where $*$ denotes complex conjugate. It is easy to show that a DFT or an IDFT applied to this sequence will produce a sequence of real numbers. This requires that D'_0 and D'_N be real, so these symbols are used in the above rule to carry the real and imaginary parts of D_0. Note that all the data are contained in the first $N+1$ terms (D'_0 through D'_N). The rest are used to ensure a real baseband signal at the IDFT output in the transmitter.

Since the input to the IDFT in the baseband case has twice as many samples, so must the output. For the same OFDM symbol length, T, this means the samples of $S(t)$ are now spaced by $T/2N$ (not T/N), thus requiring a baseband bandwidth of about $2N/T$ Hz. This is the same as the bandwidth required for the passband case.

2.5. Guard Interval Considerations

Even with a large symbol duration, channel time dispersion will cause consecutive symbols (also called *OFDM blocks*) to overlap, resulting in some residual ISI that could degrade performance. This residual ISI can be eliminated, at the expense of spectral efficiency, by using guard time intervals, between OFDM symbols, that are at least as long as the maximum extent of the channel impulse response. Samples of the received signal lying in the guard interval are discarded in the receiver and the demodulated OFDM symbol is generated from the remaining N samples.

The guard interval could be filled at the transmitter with null signal samples (zeros). However, in the case where there is dispersion, the receiver FFT processing will truncate the spread signal, so that each detected data value, $\hat{D}_k$, will consist of D_k plus interchannel interference (ICI) from the other data values.[1] In particular, if the signal has length NT_s and the impulse response of the channel is of length LT_s, the signal at the output of the channel is the linear convolution of the channel and the transmitted signal and is therefore of length $(N+L)T_s$. In the receiver, however, the FFT processes only N samples; this is the truncation that causes ICI. Putting it an alternative way, an FFT preserves orthogonality between tones only when the convolution in time is a *cyclic* convolution, rather than the linear convolution that occurs in a real channel.

One widely used solution to this problem is to cyclically extend the OFDM block by an amount longer than the expected time extent of the channel impulse response [16]. Specifically, to create a periodic received signal for the FFT to process (and thereby eliminate ICI), M' time samples are copied from the end of the original OFDM sequence and appended as a prefix; and, M'' time samples are copied from the beginning of the original OFDM sequence and appended as a suffix, where $M = (M' + M'') \geq L$. In some systems only a prefix is used ($M'' = 0$) and the processing window position is adjusted accordingly. An example is shown in Fig. 6. At the receiver, the samples of the cyclic extension are discarded before FFT processing. Clearly, the need for a cyclic extension in time-dispersive environments reduces the efficiency of OFDM transmissions by a factor of $N/(N+M)$. In most OFDM designs, a guard interval of not more than 10% to 20% of the symbol duration is employed.

2.6. Windowing

In some OFDM applications, compactness of the transmitted spectrum is important. A case in point is wireless systems, where spectrum is precious and multiple systems are closely spaced in frequency. The sharpness with which the signal spectrum falls off outside the allocated bandwidth is then of great interest.

In the OFDM systems described so far, a rectangular symbol pulse has been assumed. In other words, all samples of the IFFT output and the cyclic extension (if used)

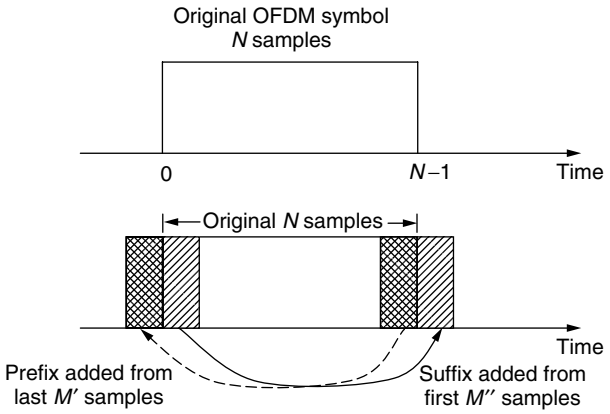


Figure 6. Cyclic extension.

between symbols in the same subchannel. Other causes of ICI, besides the one cited above, are channel time variations, frequency offset, and phase noise (Section 3).

[1] *ICI* refers, generally, to the interference between subchannels in the same symbol period. By contrast, *ISI* refers to interference

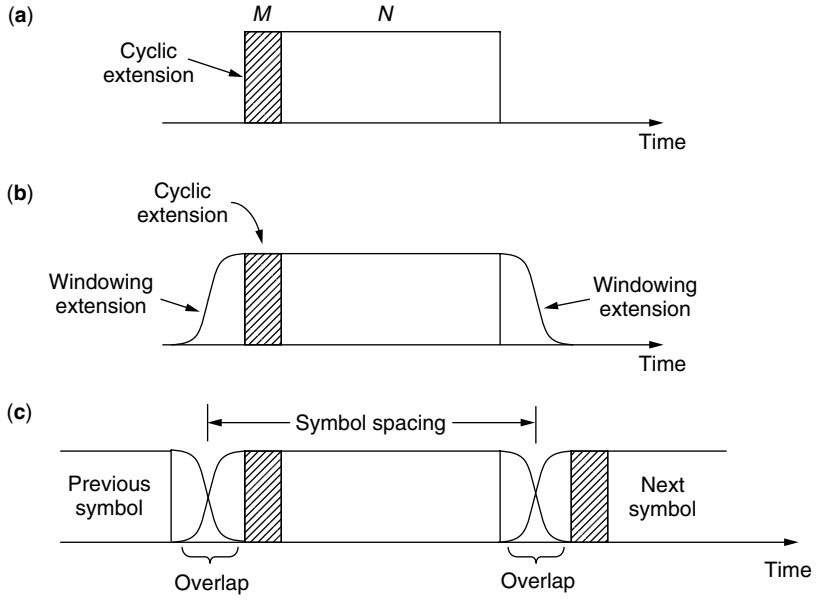


Figure 7. Windowing: (**a**) OFDM symbol; (**b**) windowed OFDM symbol; (**c**) sequence of windowed OFDM symbols.

are unweighted, which corresponds to having a rectangular symbol pulse, at each tone, of length $(N + M)T_s$. This is depicted in Fig. 7a. The spectral properties of the rectangular pulse shape (high sidelobes that decay slowly) lead to poor out-of-band spectral falloff [see Fig. 5c for $M = 0$, $N \leq 64$ or $(M + N) \leq 64$].

A simple way to improve the spectrum is to increase the periodic extension of $\{S_n\}$ even further and to taper the additional extension. This is called *windowing*. An example is shown in Fig. 7b. A commonly used shape is the cosine rolloff function. Although the total symbol duration is thus enlarged, the symbol *spacing* can be smaller than this duration, because adjacent symbols can overlap in the (unprocessed) rolloff region. This is shown in Fig. 7c.

To see the improvement possible with even a small extension, note the power spectral density plots of Fig. 8 for $N = 64$ ($M = 0$ for these computations). The parameter in the plots is the cosine rolloff factor, β, and the fractional

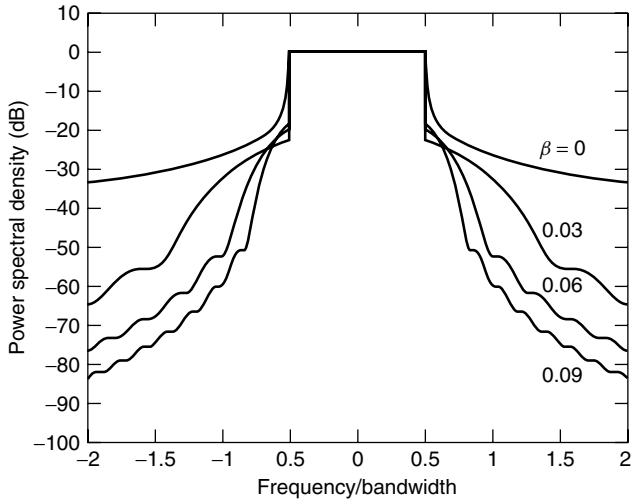


Figure 8. Power spectral density for windowed OFDM signal.

increase in symbol spacing can be shown to be $\beta/(1 - \beta)$. Thus, the curves show that an increase of only 3% in symbol spacing can produce dramatic benefits in out-of-band spectral falloff.

2.7. Coding

Because of the frequency-selective nature of the typical wideband channel (which is the main motivation for using OFDM), the OFDM subchannels generally have different received powers. Variations in the channel gain with frequency may cause some groups of received subcarriers to be much weaker than others, or even completely lost. Therefore, even though most subcarriers may be detected without errors, overall performance will be dominated by the performance of the few subcarriers with the lowest SNR (signal-to-noise ratio). As a result, satisfactory performance cannot be achieved without the addition of some form of error correction coding. By using coding across the subcarriers, errors in weak subcarriers can be corrected, up to a limit that depends on the code and the channel. Coding in OFDM systems has an additional dimension: it can be implemented in both the time and frequency domains, so that both dimensions can be utilized to achieve immunity against channel variations.

2.8. A Sample Design

The processing steps discussed above are all reflected in the simplified block diagrams of Fig. 9. We will present a

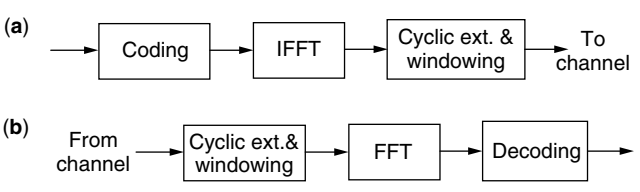


Figure 9. OFDM transmitter and receiver: (**a**) OFDM transmitter processing; (**b**) OFDM receiver processing.

sample design here, highlighting the factors influencing the choices of key parameters. It will be seen that OFDM system design involves tradeoffs among various, often conflicting, requirements. For example, to minimize the effects of time dispersion, a long symbol duration is required, meaning a large number of subchannels. However, if the channel is time-varying, as in a mobile radio environment, the variations during a long symbol period could be significant, causing possible ICI. The design parameters of interest are (1) the number of subcarriers, N; (2) the size of the FFTs, N'; (3) the guard time, T_g; the OFDM symbol duration, T; and (4) the subcarrier spacing, Δf. These are influenced by the assigned bandwidth, the desired bit rate, the time extent of the channel impulse response, and the rate of the channel time variations.

As an example, consider a wireless system that requires a bit rate of 1.2 Mbps (megabits per second) in a bandwidth of 800 kHz. Assume that the system must operate in an environment with a channel delay span of 20 μs, corresponding to wide-area transmission. A guard time, T_g, of 40 μs should be more than enough to guarantee that there is no ISI. (In this example, the guard time is assumed to be sufficient to handle the channel dispersion as well as any additional extension for windowing.) The OFDM symbol duration, T, is then chosen large enough to ensure that the efficiency loss due to the guard interval is small and to guarantee that the subchannel bandwidth is narrow enough to suffer only flat fading. In this case, we consider an OFDM symbol interval $T' = T + T_g = 200$ μs. This is five times the size of the guard interval, resulting in a 20% guard time overhead. The subcarrier spacing is then $\Delta f = 1/T = 6.25$ kHz. At a carrier frequency of 2 GHz and assuming a vehicle speed of no more than 100 km/h, the maximum Doppler spread is about 200 Hz, which is small enough compared to 6.25 kHz that ICI should be acceptably small. This choice of spacing allows for at most 128 subchannels in the 800 kHz bandwidth. Assuming QPSK modulation (i.e., 2 bits per symbol) and four guard subchannels at either end of the OFDM spectrum (to facilitate filtering), the resulting bit rate is

$$R_b = \frac{120 \text{ data subchannels} \times 2 \text{ bits per subchannel}}{200\ \mu s} = 1.2 \text{ Mbps}$$

With a half-rate code, this results in an information rate of 600 kbps. Finally, as discussed in Section 2.3, zero-valued subcarriers can be added to the data set $\{D_k\}$, to facilitate transmit filtering, so that the FFT size, N', is greater than the number of subcarriers, N. A typical choice for this design example might be $N' = 4N = 512$.

3. CHANNEL AND SYSTEM IMPAIRMENTS

3.1. Introduction

Noise and the channel frequency (or impulse) response largely determine the performance of an OFDM system. In addition, several phenomena can significantly degrade the performance. Time variations in the channel, as well as frequency offset, phase noise, and timing errors, can impair the orthogonality of the subchannels [17]. Also, the large amplitude fluctuations characteristic of a multicarrier signal can be a serious problem when transmitting through a nonlinearity, such as the transmit power amplifier. We discuss all these issues here.

3.2. Noise and Interference

The ultimate limit on system performance, even without other impairments, is the combination of thermal noise and interference. In the case of OFDM, we can assume that there are N independent subchannels, each with its own signal-to-interference-plus-noise ratio (SINR). The usual methods of analysis can be used to compute the performance (bit error rate, block error rate, etc.) of each subchannel as a function of SINR. Typical approaches for maximizing the performance of a given subchannel are power control and coding, as in other systems. The difference in OFDM is that power control and coding can be applied across subchannels as well as within subchannels.

In the case of an additive white Gaussian noise (AWGN) channel (no frequency or time selectivity), all subchannels have the same performance. Moreover, the total system performance is identical to that of a single-carrier system having the same modulation, bandwidth, and power.

3.3. Channel Time Dispersion

Channel time dispersion can produce deep fades at one or more subchannel frequencies, causing performance degradation. However, the problems of ISI and ICI due to dispersion can be avoided using guard times and a cyclic extension (see Section 2.5). To put this mathematically, let the channel impulse response be expressed in discrete-time form by the finite set $\{h_l, 0 \le l \le L\}$, where T_s is the spacing between samples and LT_s is the maximum delay. The channel response at the subcarrier frequency $f_k = k/N$ is

$$H_k = H\left(\frac{2\pi k}{N}\right) = \sum_{l=0}^{L} h_l e^{-j2\pi lk/N} \tag{10}$$

Assuming that the channel is time-invariant, each h_l is constant. Given suitable choices for the length of the OFDM symbol and the guard time, and the use of a cyclic extension to avoid ICI, the demodulated sequence may be expressed as

$$X_k = H_k D_k + \eta_k \tag{11}$$

where η_k is additive Gaussian noise in the kth subchannel. Note that the noise components for different subcarriers are generally uncorrelated, that is, $E[\eta_k \eta_l^*] = \sigma_k^2 \delta(k - l)$.

If the communication channel is time-invariant, its effect on each subchannel is seen to be represented by a single complex-valued coefficient. Therefore, correcting for the channel response can be accomplished by following the receiver FFT with a single complex gain adjustment at each subcarrier frequency. Estimation to correct for the channel is discussed in Section 4, along with the possibility of matching (adapting) the transmitted signal to the channel frequency response.

3.4. Channel Time Variations

Channel and system time variations over a symbol result in spectral spreading of the individual subchannels, which causes ICI. We now show this analytically for one type of variation. Specifically, assume that the composite effect of the channel and system time variations can be represented as a multiplicative complex factor, so that the received signal's complex envelope is $\gamma(t)S(t)$. The factor $\gamma(t)$ could represent a frequency-independent gain variation, as might be encountered in a narrowband mobile radio channel. Let the nth received sample be $R_n = \gamma_n S_n$. Then, we find the kth data value at the receiver output is

$$\begin{aligned}\hat{D}_k &= \frac{1}{N}\sum_{n=0}^{N-1}\gamma_n S_n e^{-j2\pi kn/N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1}\gamma_n D_l e^{j2\pi n(l-k)/N} \\ &= \sum_{l=0}^{N-1} D_l \frac{1}{N}\sum_{n=0}^{N-1}\gamma_n e^{j2\pi n(l-k)/N} \\ &= D_k\Gamma_0 + \sum_{\substack{l=0\\ l\neq k}}^{N-1} D_l\Gamma_{l-k} \end{aligned} \tag{12}$$

where the sequence $\{\Gamma_l\}$ is the DFT of the sequence $\{\gamma_n\}$. If $\gamma_n = 1$ for all n (a time-invariant channel), then $\Gamma_{l-k} = \delta(l-k)$ and $\hat{D}_k = D_k$; otherwise, there is ICI, namely, a complex-weighted average of the other data values. In addition, the desired signal term is attenuated and rotated via the complex factor Γ_0. In a channel that is both time-dispersive and time-varying, the mathematics is more complicated, but the basic concept is the same.

3.5. Frequency Offset

Before an OFDM receiver can demodulate subcarriers, it has to perform at least two synchronization tasks: (1) it must locate the symbol boundaries and derive the optimal timing instants, so as to minimize the effects of ICI and ISI; and (2) it must estimate and correct for carrier frequency errors due to frequency offset and phase noise. We discuss these in the next three subsections, starting with frequency offset. A number of techniques have been devised for estimating and correcting timing and carrier frequency errors at the OFDM receiver, and these are discussed in Refs. 1 and 3.

The usual source of frequency offset in OFDM is a static frequency recovery error in the receiver. To analyze the impact, we can use Eq. (12), where $\gamma(t)$ can now be modeled simply as $e^{j2\pi\delta ft}$, with δf representing the difference between the transmitter and receiver carrier frequencies. In this case, the received data symbol again suffers from ICI, as in Eq. (12), with

$$\Gamma_0 = \frac{\sin\pi\left(\dfrac{\delta f}{\Delta f}\right)}{\pi\left(\dfrac{\delta f}{\Delta f}\right)} e^{j\pi\delta f/\Delta f} \tag{13}$$

and

$$\Gamma_{l-k} = \frac{\sin\pi\left(l-k-\dfrac{\delta f}{\Delta f}\right)}{\pi\left(l-k-\dfrac{\delta f}{\Delta f}\right)} e^{j\pi[l-k-\delta f/\Delta f]} \tag{14}$$

If the frequency error is a multiple, I, of the subcarrier spacing, then the received subcarriers are shifted in frequency by $\delta f = I\Delta f$. The subcarriers remain orthogonal in this case (all still have an integer number of cycles within the FFT processing window), but the recovered data have the wrong index values. This can be seen from Eqs. (12)–(14); if $\delta f = I\Delta f$, with $I \neq 0$, then Γ_0 will be 0 and so will every Γ_{l-k} except for $l = k + I$. Thus, the detected data for the kth subchannel will be $\hat{D}_k = D_{k+I}$, meaning that all data values are detected but are associated with the wrong subchannels. In general, *all* offsets of magnitude $\Delta f/2$ or more will lead to subchannel ambiguity, where the strongest component of $\hat{D}_k$ is that of a subchannel other than the kth. The first task of receiver frequency correction, then, is a coarse acquisition that brings δf within the range $\pm\Delta f/2$.

Assuming that δf lies within this range following initial acquisition, the number of cycles within the processing window will be a noninteger for all subchannels, and ICI will result, [Eq. (14)]. (This is analogous to the ISI in a single-carrier system caused by timing offset.) Also, the desired component will be reduced in magnitude by a factor $\text{sinc}(\delta f/\Delta f)$, as given by Eq. (13).

3.6. Phase Noise

A problem related to frequency offset is phase noise: a practical oscillator does not produce a carrier at exactly one frequency, but rather a carrier that is phase modulated by random noise. As a result, the receiver's recovered frequency, which is the time derivative of its phase, is never perfectly constant. Thus, phase noise produces a *dynamic* frequency error, whereas frequency offset is a static one. The result, in both cases, is ICI. The problem is more serious in OFDM than in a single-carrier system because the subchannels are so close in frequency and, in addition, their spectra overlap.

Although OFDM is more susceptible to phase noise and frequency offset than are single-carrier systems, there are techniques for keeping this degradation to a minimum. First, phase noise in the local oscillator is common to all subcarriers. If the oscillator linewidth (i.e., the spread of the oscillator tone due to phase noise) is much smaller than the OFDM symbol rate, which is usually the case, the common phase error is strongly correlated from symbol to symbol and from tone to tone; thus, tracking or differential detection (Section 4.2) can be used to minimize its effects. Second, the impact of phase noise grows monotonically with the ratio of the linewidth to the subcarrier spacing. Therefore, control of this ratio in choosing oscillators and subcarrier spacings can control the ICI.

3.7. Timing Errors

To achieve time synchronization (as well as frequency synchronization) with a minimum of processing at the

receiver, and also a minimum of redundant information added to the data signal, the synchronization process is normally split into an acquisition phase and a tracking phase. This is possible if the general characteristics of the timing (and frequency) errors are known. In the acquisition phase, an initial estimate of the errors is acquired, perhaps using more complex algorithms and more overhead; then, the follow-on tracking algorithms only have to correct for small short-term deviations.

With respect to timing offsets, OFDM is relatively robust; in fact, the symbol timing offset may vary over an interval equal to the guard time without causing ISI or ICI. This is because, for timing offsets smaller than the guard interval, the impact is just a phase shift; that is, for a timing offset δt, the received sample for the kth subcarrier is

$$\hat{D}_k = D_k e^{j2\pi f_k \delta t} \tag{15}$$

Thus, no ICI results; just a phase error which grows with f_k. If differential detection between tones is used, the impact of the timing error can be controlled by just ensuring that the root mean square (RMS) value of $\delta t/T$ is sufficiently small, say, 0.01 or less. The precise requirement depends on the modulation, the target bit error rate, and other such factors. Of course, if δt exceeds the guard time, the receiver's FFT window spans samples from two consecutive OFDM symbols and ISI occurs.

3.8. Transmitter Nonlinearities

An OFDM signal is the superposition of many modulated subcarrier signals and thus may exhibit a high signal peak relative to the average signal level. If the transmitter processing is not linear over the full range of the signal variation, nonlinear distortion will occur. This is manifested in two ways: (1) in-band intermodulation products, causing interference to each subchannel; and (2) out-of-band spectral spreading, potentially causing adjacent-channel interference (ACI) to other systems. Avoiding these problems requires a degree of transmitter linearity that can be costly.

One possible metric for characterizing signal peaking is the ratio of the peak signal power to the average signal power, or the *peak-to-average power ratio* (PAPR). This quantity can be taken over an OFDM symbol, in which case it varies from symbol to symbol, or over all time, in which case it is a single number. Either way, this metric must be used with care. The most extreme peaking occurs when all of the subcarrier signals line up in their peak amplitudes at the same time instant. It is easy to show that, for N subcarriers having equal average powers and using BPSK or QPSK (binary and quadrature phase shift keying) modulation, the PAPR defined as above (and taken over all time) is N. Thus, the PAPR would be 12 dB for $N = 16$ and 21 dB for $N = 128$. It may therefore seem that signal peaking progressively worse as N increases. However, worst-case signal peaking becomes less probable as N increases, so it is necessary to look at peaking in a *statistical* way. For N sufficiently large, the complex envelope converges to a complex Gaussian process, meaning that the squared envelope approximates an exponential variate. This approximation is used in Fig. 10, which shows the complementary cumulative distribution function for the ratio of the instantaneous power to the average power, taken over all time. If we redefine PAPR as the value not exceeded more than 0.001% of the time, the proper value to use is about 10.6 dB. This result, which holds for all realistic N and all modulations, is large enough to raise concerns about transmitter nonlinearities.

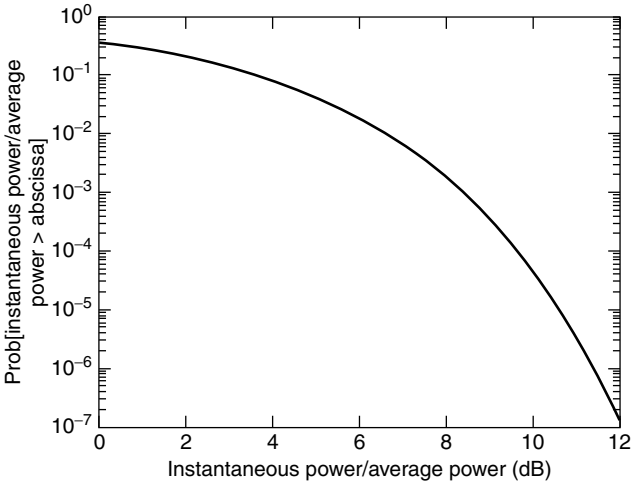


Figure 10. Complementary cumulative distribution function of PAPR of an OFDM signal.

To transmit signal peaks without major distortion, the transmitter's DAC must use a sufficient number of bits to accommodate these peaks, which is a cost/technology issue. More importantly, the power amplifier must remain linear over an amplitude range that includes the peaks, which leads to both high amplifier cost and high power consumption (low power efficiency). Several techniques have been proposed to mitigate the peaking problem, and they divide basically into three categories: (1) *signal distortion* techniques, which reduce the peak amplitudes by nonlinearly distorting the OFDM signal at or around the peaks. (e.g., clipping and filtering, peak windowing, peak cancellation); (2) *coding* techniques, involving special codes that exclude OFDM symbols with high peaking; and (3) *scrambling* techniques, that is, scrambling each OFDM symbol with different sequences and selecting the one that gives the least peaking. Details and the relative performances of these techniques can be found in Refs. 1, 2, 18, and 19.

4. OTHER MAJOR ISSUES

4.1. Introduction

We have seen that, to get the most value out of OFDM, special techniques have been devised such as cyclic extension and windowing. These relate primarily to how the signal is prepared at the transmitter to be sent over the channel. Equally important are methods of data detection and channel estimation at the receiving end and methods for adapting both transmission and reception to the frequency selectivity of the channel so as to maximize data efficiency. We explore these topics here.

4.2. Detection Techniques

In general, the data constellation of each subcarrier will show a random phase shift and amplitude change. These are caused by carrier frequency offset, timing recovery offset, and the frequency selectivity of the channel, as discussed in the previous section. To cope with these unknown changes, two classes of detection techniques exist. The first is *coherent detection*, using estimates of the channel response to derive reference values for the amplitude and phase correction for each subchannel. Spectrally efficient use of this approach requires reliable techniques for channel estimation that, at the same time, do not require excessive overhead, as discussed in the next section.

The second technique is *differential detection*, which does not require absolute reference values but accounts only for the phase and/or amplitude differences between two data symbols. In OFDM, differential detection can be done in the time domain or in the frequency domain. In the first case, each subcarrier is compared with the same subcarrier of the previous OFDM symbol; in the second case, each subcarrier is compared with the adjacent subcarrier within the same OFDM symbol. In contrast to coherent detection, differential detection does not require channel estimation, thereby saving complexity and gaining overhead efficiency. The cost is a degraded performance because of the noisy references that are effectively being used.

If differential detection is used within each subchannel, symbols must be highly correlated in time; performance can thus degrade if the channel response has significant time variation. Similarly, if differential detection is done between subchannels, symbols must be highly correlated in frequency; performance can thus degrade if the channel response has significant frequency variation.

4.3. Channel Estimation and Correction for Coherent Detection

In the kth OFDM subchannel, the data component appears at the detector input with a complex amplitude scaling, H_k, plus Gaussian noise, η_k, as in Eq. (11). Coherent detection of this sample amounts to comparing it against all points in the data constellation and choosing the point closest to it. To do this optimally, it is necessary to undo the amplitude scaling $|H_k|$ and the phase rotation $\mathrm{Arg}(H_k)$. Doing this individually for all frequency components of the FFT output is called *frequency-domain equalization*; broadly speaking, it consists of scaling each subchannel with a complex multiplier $1/\hat{H}_k$, where $\hat{H}_k$ is an approximation to H_k.

A conventional approach to implementing this equalization is to initially estimate the subchannel gains (e.g., by transmitting a known modulated sequence in each subchannel) and to then handle time variations via either periodic updates or decision-directed tracking. An alternative approach, ideally suited to OFDM, is pilot-aided estimation. Pilots are unmodulated tones, lasting for one or more symbols at a time, that are inserted by the transmitter and processed by the receiver to estimate channel gains. They can be distributed in time and frequency in any number of ways, one example of which is shown in Fig. 11. The two competing goals in placing pilots are that they should occupy a small fraction of the time–frequency slots, and their frequency of occurrence in each direction should be high enough to adequately sample the channel.

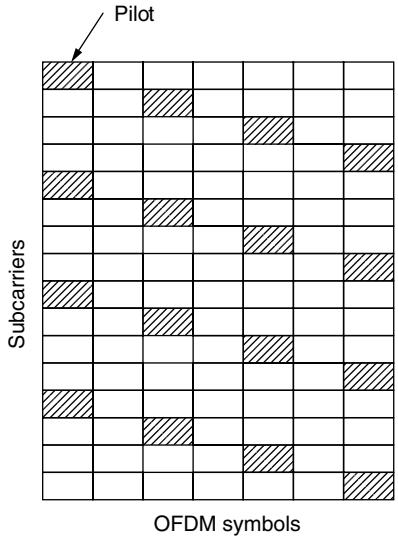


Figure 11. Typical pilot pattern.

Pilots are used for channel correction as follows. First, the receiver estimates channel gains at all times and frequencies where pilots have been transmitted. Then the channel gains at all other time–frequency positions can be estimated using two-dimensional interpolation filtering. Equalization then consists of setting the scaling to $1/\hat{H}_k$ for each data-carrying subchannel, where $\hat{H}_k$ is the gain estimate or some modification to account for additive noise.

To accurately interpolate the channel estimates from the available pilots, the pilot spacing in each dimension must satisfy the Nyquist sampling theorem. This means that there exist both a minimum necessary subcarrier spacing and a minimum necessary symbol spacing between pilots. To determine these spacings, two quantities must be known or estimated, namely, the double-sided bandwidth, $B_{\max}$, of the channel gain's time variations; and the full time extent, $\tau_{\max}$, of the channel's impulse response. The requirements for the pilot spacings in time and frequency, Δt_p and Δf_p, are then $\Delta t_p < 1/B_{\max}$ and $\Delta f_p < 1/\tau_{\max}$. In order to get a degree of noise reduction by filtering, the pilot symbol spacing should be smaller than half these values (oversampling) but not so small that the fraction of pilots is excessive.

Many solutions based on both pilot-aided estimation and decision-directed scaling are described in the references (e.g., see Refs. 1 and 3). The proper choice among pilots, training sequences, and decision-directed tracking, and the "best" design of whichever methods are used, depend on such factors as channel variability, type of traffic (e.g., continuous voice, packet data), and performance and cost objectives. For example, in the case of high-speed packet transmission to low-mobility users, as in wireless LAN applications, the most appropriate approach seems to be the use of a preamble consisting of one or more known OFDM symbols. The choice of the number of training

symbols is a tradeoff between short training time (better spectral efficiency) and good estimation performance.

4.4. Adaptive Loading

The frequency selectivity of an OFDM channel provides both challenges and opportunities. One way to address the problem of weak subchannels is to code across tones (thereby exploiting the frequency selectivity that causes the problem), as noted in Section 2.7. Another is to adaptively turn off weak subchannels, that is, to send no data at frequencies where the received SNR is below some threshold. To better exploit frequency selectivity and realize a spectral efficiency benefit, the data constellation used in each subchannel can be adaptively sized to its SNR [20]. This process, called *adaptive loading*, recognizes the fact that, in media where some subchannels are weaker than average, others are stronger. Thus, for example, each subchannel could use QPSK, 16-QAM, or 64-QAM (equivalently, 2, 4, or 6 bits per symbol), depending on the frequency response (gain) for that subchannel. This optimum form of OFDM is used for DSL applications and is called *discrete multitone* (DMT) [21]. A sample variation of the channel frequency response across the OFDM subchannels is illustrated in Fig. 12.

For the case of fixed transmit power, P, in each subchannel, the SNR in the kth subchannel would be

$$SNR_k = \frac{P}{N_0/T}|H_k|^2 \tag{16}$$

where N_0 is the noise power density at the receiver input. Assume that SNR is accurately measured in the receiver for every subchannel and communicated to the transmitter over a feedback channel. For a specified bit error rate, each such measurement can be used to select a data constellation size; specifically, the bits per symbol for the kth subchannel can be matched to the subchannel gain, $|H_k|^2$. To be effective, this approach requires an accurate SNR measurement in the receiver, a reliable feedback channel to the transmitter, and a means for changing constellations at the transmitter and efficiently notifying the receiver. If the same carrier frequency is used for both transmit and receive, as in Time-Division Duplexing, the over-the-air feedback channels is not required.

An additional degree of freedom is power control, that is, adaptively changing the transmit power, P_k, in the kth subchannel in accordance with its gain, subject to a total power constraint. If the power and constellation size are jointly distributed among subchannels in the most optimal way, the overall spectral efficiency of OFDM matches that of a single-carrier system with ideal decision feedback equalization [22].

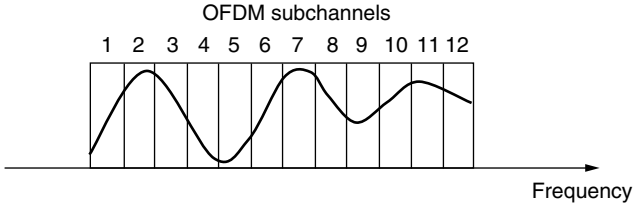


Figure 12. A channel response that justifies adaptive loading.

5. CONCLUSION

OFDM is a very attractive technique for high-bit-rate data transmission over a dispersive communication channel. However, to realize an OFDM system, many practical issues must be addressed, including high signal peaks, frequency offset, timing mismatch, and channel variation. Promising solutions have been devised for all these problems, but most solutions contribute to a "nibbling away" of the spectral efficiency. Even so, OFDM can attain the same spectral efficiency as an equalized single-carrier system and has the virtue of flexibility and a processing complexity that grows gracefully with channel dispersion.

Multicarrier modulation has been used in modems for both radio and telephone channels, as well as for digital audio and video broadcasting. The digital audio broadcasting (DAB) standard was, in fact, the first OFDM-based standard. The main reasons to choose OFDM for this system, which also applies to digital video broadcasting (DVB), are the possibility to provide a single frequency network and the efficient handling of multipath delay spread. A particularly suitable application of multicarrier modulation is in digital transmission over copper wire subscriber loops, as exemplified in high-speed digital subscriber loop (DSL) systems. In addition, the major high-bit-rate wireless LAN standards (IEEE 802.11a, HIPERLAN/2, and MMAC) use OFDM to overcome the bit-rate limitations caused by delay spread.

While much effort on OFDM is focused on critical implementation issues, there is also research on new variations and applications. Examples include *orthogonal frequency-division multiple access* (OFDMA); OFDM combined with *code-division multiple access* (CDMA); and OFDM combined with *multiple-input/multiple-output* (MIMO) antenna techniques. In OFDMA, multiple access is realized by providing each user with a fraction of the total number of subcarriers [23]. It is similar to conventional FDMA, except that it avoids the usual guard bands and exploits the power of FFT processing. OFDM-CDMA techniques (of which there are several variants) provide alternative ways to achieve multiple access while still combating frequency selectivity with moderate processing complexity [24]. OFDM-MIMO techniques exploit the power of array processing to increase wireless system capacity [25]. In all these applications, the need for time-domain equalization or RAKE reception is avoided because of the use by OFDM of narrow subchannels. Thus, powerful signal processing techniques like the FFT and adaptive arrays can be used instead to achieve high levels of performance.

BIOGRAPHIES

Leonard J. Cimini, Jr., received his B.S., M.S., and Ph.D. degrees in electrical engineering from the University

of Pennsylvania in 1978, 1979, and 1982, respectively. Over a 20-year AT&T career, starting at Bell Labs and then at AT&T Labs, he conducted research on lightwave and wireless communications systems. His main emphasis has been on devising techniques for overcoming the bit-rate limitations imposed by communications channels. In this context, he pioneered the application of Orthogonal Frequency Division Multiplexing to the emerging field of wireless communications. Dr. Cimini has been very active within the IEEE, including serving on several editorial boards and on the board of governors of the IEEE Communications Society. He was also the founding editor in chief of the IEEE J-SAC: Wireless Communications Series. Dr. Cimini is an adjunct professor in the Electrical Engineering Department of the University of Pennsylvania where he teaches a graduate-level course in wireless systems. He was elected a fellow of the IEEE in 2000 for contributions to the theory and practice of high-speed wireless communications.

Larry J. Greenstein received his B.S., M.S., and Ph.D. degrees in electrical engineering from Illinois Institute of Technology, Chicago, Illinois, in 1958, 1961, and 1967, respectively. From 1958 to 1970 he was with IIT Research Institute, working on radio frequency interference and anti-clutter airborne radar. He joined Bell Laboratories, Holmdel, New Jersey, in 1970. Over a 32-year AT&T career, he conducted research in digital satellites, point-to-point digital radio, lightwave transmission techniques, and wireless communications systems. For 21 years during that period (1979–2000), he led a research department renowned for its contributions in these fields. His research interests in wireless communications have included measurement-based channel modeling, microcell system design and analysis, diversity and equalization techniques, and system performance analysis and optimization. He recently retired from AT&T Labs–Research, Middletown, New Jersey, as a technology leader. Dr. Greenstein is an AT&T fellow and an IEEE fellow, has won two best paper awards, and has been a guest editor, senior editor, and editorial board member for numerous publications.

BIBLIOGRAPHY

1. R. Van Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*, Artech, 2000.
2. A. Bahai and B. Saltzberg, *Multi-Carrier Digital Communications — Theory and Applications of OFDM*, Kluwer, 1999.
3. L. Hanzo, W. Webb, and T. Keller, *Single- and Multi-carrier Quadrature Amplitude Modulation*, Wiley, 2000.
4. J. A. C. Bingham, *ADSL, VDSL, and Multicarrier Modulation*, Wiley, 2000.
5. M. L. Doelz, E. T. Heald, and D. L. Martin, Binary data transmission techniques for linear systems, *Proc. IRE* **45**: 656–661 (May 1957).
6. M. S. Zimmerman and A. L. Kirsch, The AN/GSC-10 (KATHRYN) variable rate data modem for HF radio, *IEEE Trans. Commun.* **COM-15**: 197–205 (April 1967).
7. M. Alard and R. Lasalle, Principles of modulation and channel coding for digital broadcasting for mobile receivers, *EBU Tech. Rev.* 168–190 (1987).
8. U. Reimers, DVB-T: The COFDM-based system for terrestrial television, *Electron. Commun. Eng. J.* **9**: 28–32 (Feb. 1997).
9. P. S. Chow, J. C. Tu, and J. M. Cioffi, A discrete multitone transceiver system for HDSL applications, *IEEE J. Select. Areas Commun.* **SAC-9**: 909–919 (Aug. 1991).
10. R. van Nee et al., New high rate wireless LAN standards, *IEEE Commun. Mag.* **37**: 82–88 (Dec. 1999).
11. L. J. Cimini, Jr., Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing, *IEEE Trans. Commun.* **COM-33**: 665–675 (June 1985).
12. S. B. Weinstein and P. M. Ebert, Data transmission by frequency-division multiplexing using the discrete fourier transform, *IEEE Trans. Commun. Technol.* **COM-19**: 628–634 (Oct. 1971).
13. U.S. Patent 3,488,445 (filed Nov. 14, 1966; issued Jan. 6, 1970), R. W. Chang, Orthogonal frequency division multiplexing.
14. B. R. Saltzberg, Performance of an efficient data transmission system, *IEEE Trans. Commun. Technol.* **COM-15**: 805–813 (Dec. 1967).
15. B. Hirosaki, An orthogonally multiplexed QAM system using the discrete fourier transform, *IEEE Trans. Commun.* **COM-29**: 982–989 (July 1981).
16. A. Peled and A. Ruiz, Frequency domain data transmission using reduced computational complexity algorithms, *Proc. ICASSP'80*, April 1980, pp. 964–967.
17. T. Pollet, M. van Bladel, and M. Moeneclaey, BER Sensitivity of OFDM systems to carrier frequency offset and Wiener phase noise, *IEEE Trans. Commun.* **43**: 191–193 (Feb.–April 1995).
18. X. Li and L. J. Cimini, Jr., Effects of clipping and filtering on the performance of OFDM, *IEEE Commun. Lett.* **2**: 131–133 (May 1998).
19. S. Müller and J. Huber, A comparison of peak power reduction schemes for OFDM, *Electron. Lett.* **33**: 3680–3689 (Feb. 1997).
20. I. Kalet, The multitone channel, *IEEE Trans. Commun.* **37**: 119–124 (Feb. 1989).
21. P. S. Chow, J. M. Cioffi, and J. A. C. Bingham, A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels, *IEEE Trans. Commun.* **43**: 773–775 (Feb.–March 1995).
22. N. A. Zervos and I. Kalet, Optimized decision feedback equalization versus orthogonal frequency division multiplexing for high speed data transmission over the local cable network, *Proc. ICC'89*, Sept. 1989, pp. 1080–1085.
23. M. Suzuki, R. Boehnke, and K. Sakoda, BDMA — band division multiple access — a new air interface for third generation mobile system, UMTS, in Europe, *Proc. ACTS Mobile Commun. Summit*, Oct. 1997, pp. 482–488.
24. N. Yee, J.-P. Linnartz, and G. Fettweis, Multi-carrier CDMA in indoor wireless networks, *Proc. IEEE PIMRC'93*, Sept. 1993, pp. 109–113.
25. G. G. Raleigh and J. M. Cioffi, Spatio-temporal coding for wireless communications, *IEEE Trans. Commun.* **46**: 357–366 (March 1998).

ORTHOGONAL TRANSMULTIPLEXERS: A TIME-FREQUENCY PERSPECTIVE

ALI N. AKANSU
HUSREV T. SENCAR
New Jersey Institute of Technology University Heights
Newark, New Jersey

1. INTRODUCTION

Orthogonality of carriers has been widely utilized in communications as the way to share available common resources by multiple users [1–3]. The most popular multiuser communication systems use one of the three modulation techniques. Namely, the frequency division multiple access (FDMA), the time division multiple access (TDMA), and code division multiple access (CDMA). The orthogonality of the user signature functions or carriers in a multiuser communication scenario is the underlying feature. The basic difference between these modulation techniques comes from the domain where the orthogonality conditions of the carrier functions are emphasized. In other words, an FDMA system aims to minimize the interaction of its multiple carriers in the frequency domain. Similarly, a TDMA system tries to reduce the time domain overlaps or correlations of its carrier or modulation functions. In contrast, a CDMA system prefers maximized overlaps of its signature functions both in the time and frequency domains while keeping their orthogonality features as the most vital requirement.

The fundamentals of signal and transform theories help us to better understand the multiple user or multicarrier communication systems where the time-frequency and orthogonality properties of the carrier functions are the defining issues. Therefore, we will briefly describe them in the following section.

2. MATHEMATICAL PRELIMINARIES

2.1. Time-Frequency Measures for a Discrete-Time Function

The time and frequency domain energy spread of a function has been a classical topic in signal processing field. The celebrated "uncertainty principle" states that no function can be concentrated simultaneously in both the time and frequency domains [4]. The time domain spread of a discrete-time function $\{h_0(n)\}$ is defined as [5],

$$\sigma_n^2 = \frac{1}{E}\sum_n (n-\overline{n})^2 |h_0(n)|^2 \tag{1}$$

The energy E and time center $\overline{n}$ of the function $\{h_0(n)\}$ are expressed as

$$E = \sum_n |h_0(n)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_0(e^{jw})|dw, \tag{2}$$

$$\overline{n} = \frac{1}{E}\sum_n |h_0(n)|^2 \tag{3}$$

where its Fourier transform is given as

$$H_0(e^{jw}) = \sum_n h_0(n)e^{-jwn}x \tag{4}$$

Similarly, we define the frequency domain spread of a discrete-time function with $\overline{w} = 0$ as follows

$$\sigma_w^2 = \frac{1}{2\pi E}\int_{-\pi}^{\pi} (w-\overline{w})^2 |H_0(e^{jw})|^2 dw \tag{5}$$

where its center in the frequency domain is given as

$$\overline{w} = \frac{1}{2\pi E}\int_{-\pi}^{\pi} w|H_0(e^{jw})|^2 dw \tag{6}$$

Figure 1 illustrates time-frequency properties of a discrete-time function $\{h_0(n)\}$ using spreading measures defined above. This representation is also called time-frequency tile of a function. The shape and location of the tile can be defined by properly designing the time and frequency features of the function under construction. This interpretation of functions can be further extended in the case of orthogonal basis design. In addition to shaping time-frequency tiles, the orthogonality requirements are also imposed on the basis functions.

For any real signal with $\overline{w} = 0$ and $\overline{n} = 0$ the lower bound for the product of *time-frequency spread* $\sigma_n\sigma_w$ is given as [6]

$$\sigma_n\sigma_w \geq \frac{|1-\mu|}{2} \tag{7}$$

where

$$\mu = \frac{|H_0(e^{jw})_{w=\pi}|^2}{E} \tag{8}$$

Similar time-frequency spreading measures for band-pass signals with peak frequency responses $\overline{w} \neq 0$ are also introduced in Ref. 6.

2.2. Orthogonal Function Sets

2.2.1. Orthogonal Block Transforms. Orthogonal block transforms like discrete Fourier transform (DFT) and

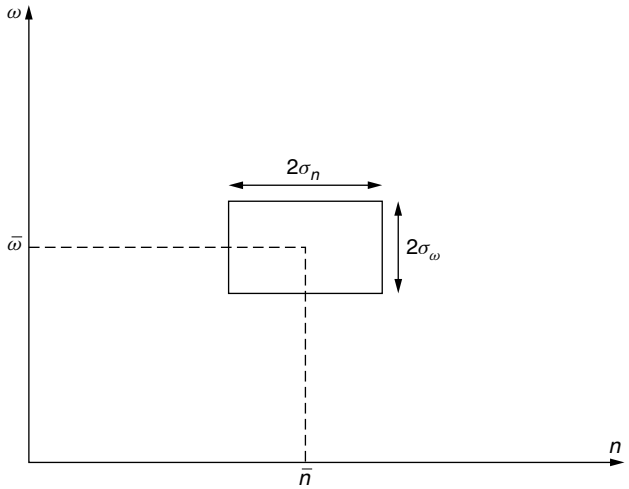


Figure 1. Time-frequency plane illustrating time-frequency properties of discrete time function $\{h_0(n)\}$.

discrete cosine transform (DCT) have been widely used in many engineering applications. The basis of an orthogonal block transform consists of functions $\{h_k(n)\}$ with the orthogonality property as

$$\sum_{n=1}^{N} h_k(n)h_l(n) = \delta_{k-l} \tag{9}$$

where δ_{k-l} is the Kronecker delta sequence given as

$$\delta_{k-l} = \begin{cases} 1, & k-l=0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Note that the length of basis functions and number of functions in the set are the same in block transforms. Therefore, the main emphasis in block transforms has been the orthogonality requirements along with implementation efficiency since there is not much freedom in the design to adjust the time and frequency properties of the functions in the set.

Figure 2 displays DCT basis functions in the time and frequency domains for $N = 8$. Due to the short time durations of these functions it is observed that their frequency selectivity is somehow limited and they overlap significantly. They perform like a filter bank with poor frequency selectivity. The time-frequency measures of these functions are presented in Table 1.

The only way to improve the frequency localizations of orthogonal basis functions is to increase their durations in the time domain. Due to the time-frequency duality property of functions, the localization of a function in one domain can be improved at the expense of its localization in the other domain. This property paved the way for filter banks and subband transforms that we introduce in the next section.

One can use orthogonal transforms for analysis of a given function or signal through an operation called forward transform. Let real orthonormal sequences $\{h_r(n)\}$ be the rows of a transformation matrix, $H(r, n)$,

$$H = [H(r, n)], \quad r, n = 1, \ldots, N \tag{11}$$

Orthonormality of the matrix H assures that its inverse matrix

$$H^{-1} = H^T \tag{12}$$

where ^{T} indicates a matrix transpose. Hence,

$$HH^T = I \tag{13}$$

where I is an $N \times N$ identity matrix.

The forward transform of an input vector $\underline{x}$ of size N is written in a matrix notation as

$$\theta = Hx \tag{14}$$

where θ is the transform coefficient vector of size N. Therefore, x can be perfectly reconstructed through the inverse transform operation as

$$x = H^T\theta \tag{15}$$

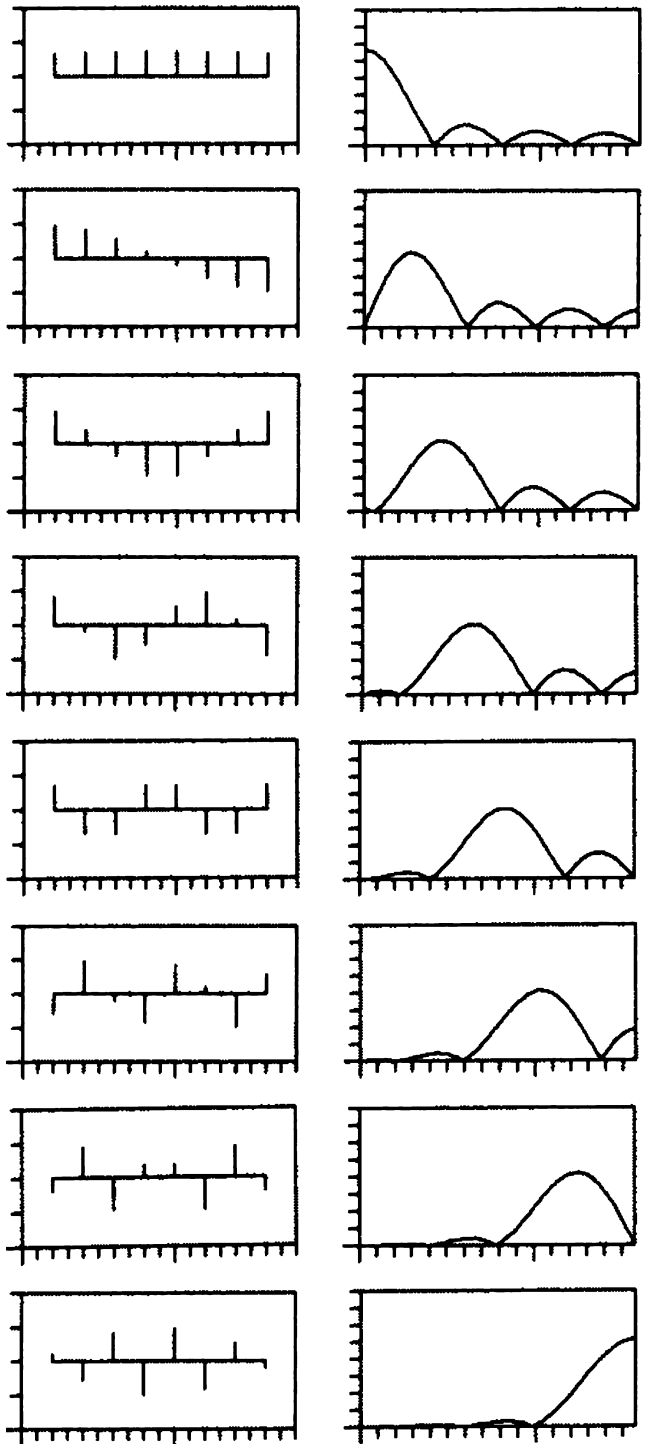


Figure 2. DCT basis functions in time and frequency domains for $N = 8$.

This defines a pair of forward/inverse transform operators where the input x is mapped onto the transform space as vector θ and perfectly recovered from θ through the inverse transform operator.

In contrast, one could synthesize the signal vector x by transforming the given input signal θ onto the inverse transform space as

$$x = H^T\theta \tag{16}$$

Table 1. Time-Frequency Localization of DCT for $N = 8$

$\overline{w}$	$\overline{n}$	σ_w^2	σ_n^2	$\sigma_w\sigma_n$
0	3.5	0.3447	5.25	1.3452
0.74	3.5	0.3021	8.4054	1.5935
1.02	3.5	0.2413	5.9572	1.1989
1.36	3.5	0.1957	5.4736	1.0350
1.71	3.5	0.1488	5.25	0.8839
2.08	3.5	0.1206	5.0263	0.7786
2.45	3.5	0.0797	4.5428	0.6017
π	3.5	0.1388	2.0955	0.5393

where H^T is the inverse transform matrix. It is a straightforward operation to perfectly reconstruct θ from x through a forward transform operation on x as

$$\theta = Hx \tag{17}$$

This is a sequence of inverse/forward transform operators that serves as the foundation for *orthogonal transmultiplexers* in multiple access communications. Fourier transform basis has been widely used in telecommunication applications for many decades utilizing the concept of transmultiplexers. As mentioned earlier, the frequency selectivity of these carrier functions, DFT basis, are not very good although their implementation in a real-time transmultiplexer structure is efficient. Therefore, they have been quite popular [7].

2.2.2. M-Band Filter Banks with Perfect Reconstruction. A maximally decimated M-band finite impulse response (FIR) perfect reconstruction quadrature mirror filter (PR-QMF) bank analysis/synthesis configuration is displayed in Fig. 3a. The output of this filter bank is the delayed version of its input as

$$y(n) = x(n - n_0) \tag{18}$$

where n_0 is a delay constant. In a paraunitary filter bank solution, the synthesis and analysis filters are related as

$$g_r(n) = h_r^*(p - n) \tag{19}$$

where p is a time delay. Hence, the PR-QMF conditions can be imposed on the analysis filters in the time domain

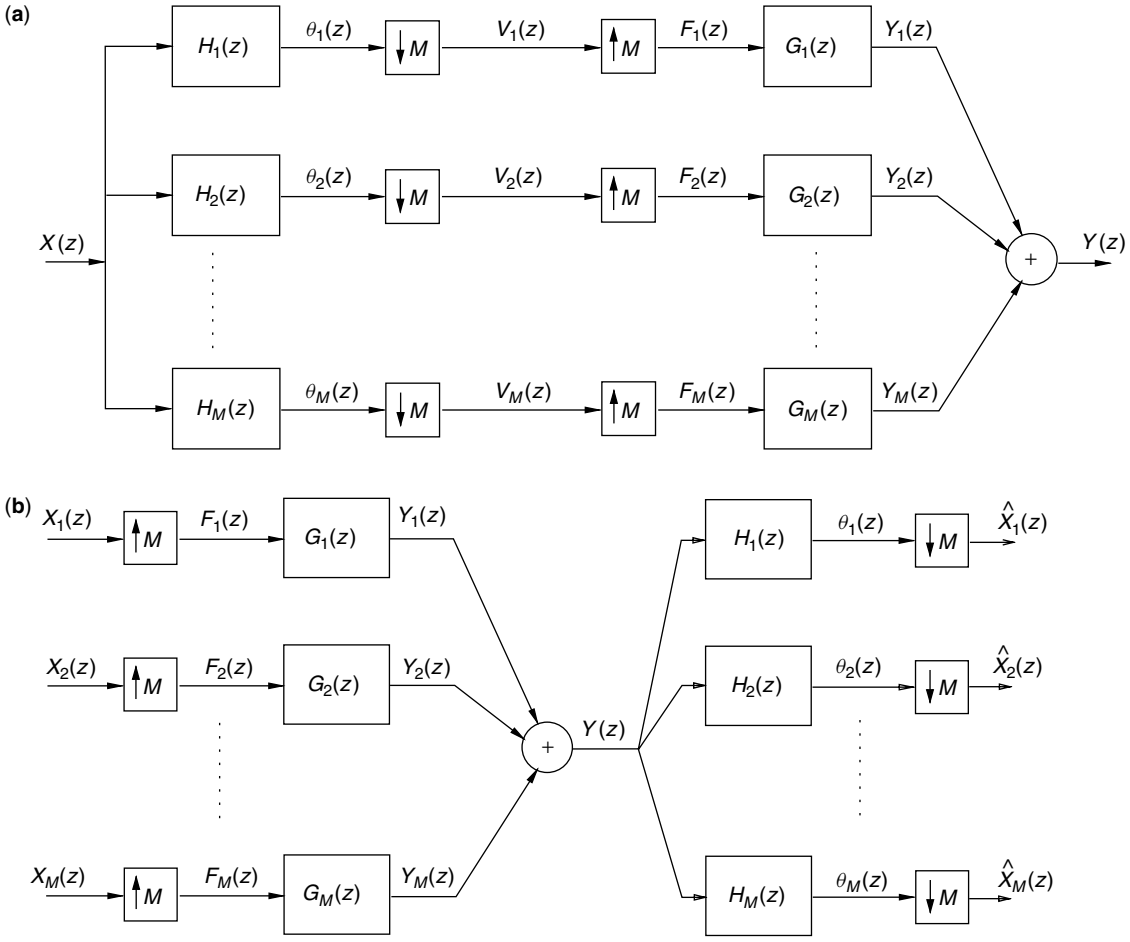


Figure 3. (**a**) Maximally decimated M-band FIR PR-QMF bank structure (analysis/synthesis filter bank configuration). (**b**) M-band Transmultiplexer structure (synthesis/analysis filterbank structure).

as [6]

$$\sum_n h_r(n)h_r(n+Ml) = \delta(l), r = 1, \ldots, M \quad (20)$$

$$\sum_n h_r(n)h_s(n+Ml) = 0, \text{ for all } l \quad (21)$$

Analysis/synthesis filter bank configurations are widely used in image or video processing, speech or audio processing, interference cancellation, and other applications [8–9].

In contrast, Fig. 3b depicts a synthesis/analysis filter bank where there are M inputs and M outputs of the system. It is shown that if the synthesis $\{g_r(n)\}$ and analysis $\{h_r(n)\}$ filters satisfy the PR-QMF conditions of Eqs. (19)–(21), synthesis/analysis filter bank configuration gives equal input and output for its all branches as

$$\hat{x}_r(n) = x_r(n - n_0) \quad r = 1, \ldots, M \quad (22)$$

where n_0 is a time delay. The synthesis/analysis PR-QMF bank with equal inputs and outputs at all branches has been used as orthogonal transmultiplexers in communications applications for single user and multiuser scenarios [6,10].

3. COMMUNICATION APPLICATIONS OF ORTHOGONAL TRANSMULTIPLEXERS

The unified framework for orthogonal transmultiplexers along with time-frequency tools allowing some design flexibilities for the application at hand was given in the previous sections of the article. The main engineering challenge in this context is to design the most suitable transform basis $\{h_r(n)\}$ for a given application. Applications using orthogonal multiplexers vary from single-user communication scenarios to multiuser communication systems. These applications might require to utilize frequency localized or time-localized orthogonal carriers depending on the system requirements including channel properties. The orthogonal block transforms like DFT has been widely used in a synthesis/analysis filter bank configuration (inverse/forward block transform sequence of operators) as a transmultiplexer in multicarrier (single and multiuser) communication systems. Although the frequency selectivity of DFT basis functions is not very good, except at the bin frequencies of the orthogonal carrier functions, the ease of its implementation has been very attractive for real-time applications.

An examination of M-band orthogonal transmultiplexer structure displayed in Fig. 3b helps us to interconnect time-frequency properties of carrier (modulation) basis with the type of communication system under consideration. The most popular types are FDMA, TDMA, and CDMA. We discuss these orthogonal modulation types further from a time-frequency perspective in the following sections.

3.1. FDMA

Figure 4a displays an ideal filter bank that consists of brick-wall frequency functions without any interbrand energy leakage. The frequency localizations of these orthogonal carriers are perfect although their time-localization is extremely poor. Since they are noncausal functions with infinite time durations they are not implementable. In practice, finite length (FIR) orthogonal carriers are used. Therefore, interbrand (cross-carrier) energy leakage (interference) is of a great concern in communication applications. As mentioned earlier, DFT basis has been used in many applications including the popular digital subscriber line (DSL) communications. The other applications like digital audio broadcasting (DAB) and low probability of intercept (LPI) communication also employ orthogonal transmultiplexers with properly selected filter banks or carrier basis [8,9].

3.2. TDMA

Similarly, Fig. 4b displays the ideal orthogonal carriers (basis functions) for a TDMA configuration. Note that each carrier is a unit sample function in the time domain with a perfect localization. In contrast, those functions completely overlap in the frequency domain. In this case, a perfect orthogonality is imposed in the time domain. Practical TDMA systems use nonideal time pulses or symbols where intercarrier energy leakage (interference) is inevitable.

3.3. CDMA

CDMA is a marked departure from the traditional FDMA and TDMA systems where the spreading of orthogonal carrier functions (signatures) in both domains (time and frequency) is aimed. Note that the orthogonal transmultiplexer configuration of Fig. 3b is still applicable even for the CDMA systems. The optimal code design problem for CDMA can also be handled by the PR-QMF requirements of Eqs. (19)–(21) with the addition of maximizing the joint time-frequency spread, $\sigma_n\sigma_w$, of the codes in the design. Figure 5 displays frequency spectra of a 32-length spread spectrum PR-QMF along with 31-length Gold Codes. It is observed from this figure that the functions used in orthogonal transmultiplexers for CDMA are not selective in either domain. Although filter banks have been mostly used for spectral analysis/synthesis problems the underlying theory is also applicable for any time-frequency shaped function sets.

4. CONCLUSIONS

The synthesis/analysis configurations of filter banks (transmultiplexers) have been widely employed in many popular communications applications. Conversely, the time-frequency shaping of orthogonal functions has been well tied to the optimal filter bank design in the signal processing literature. We highlighted those developments in the context of orthogonal transmultiplexers that serve as the foundation in single and multiple user communication systems.

BIOGRAPHIES

Ali N. Akansu received the B.S. degree from the Technical University of Istanbul in 1980 and the M.S. and Ph.D.

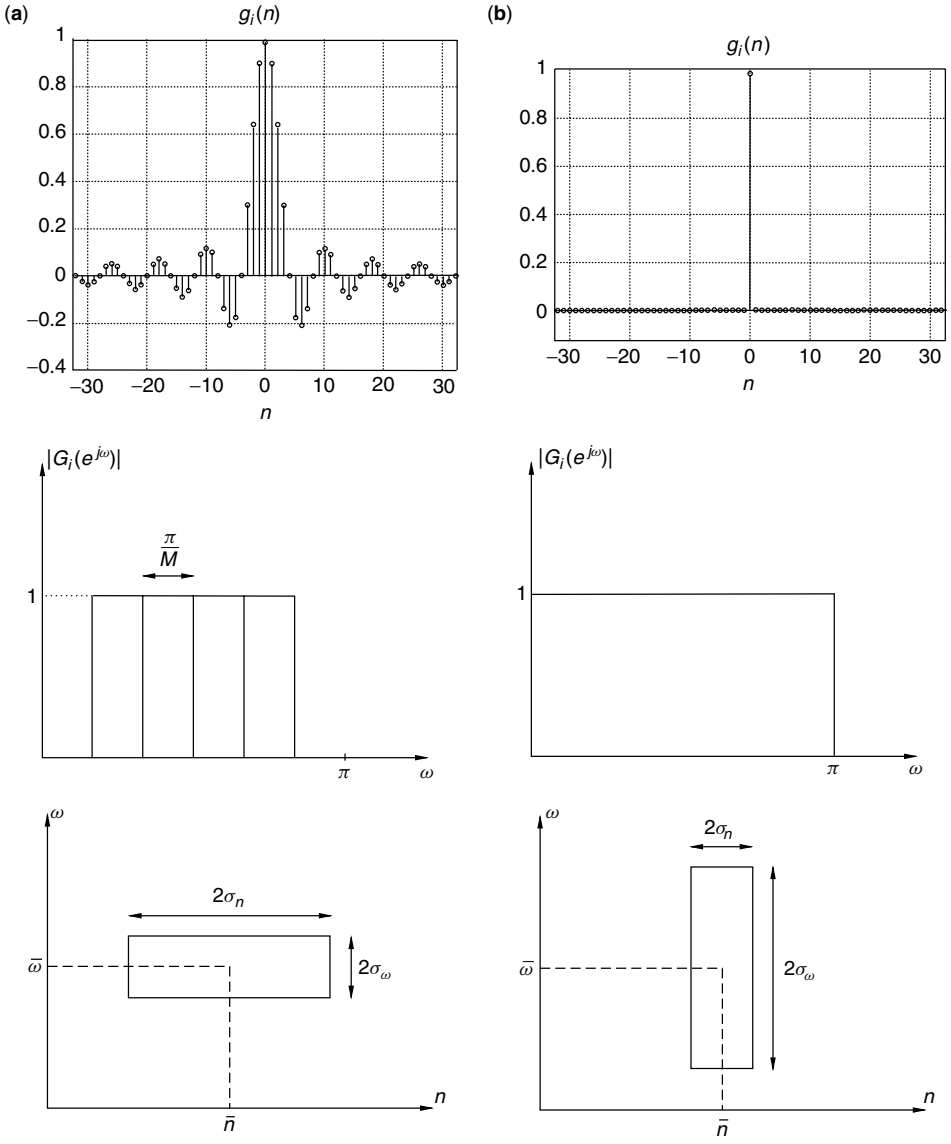


Figure 4. Ideal filter banks (orthogonal carriers) for the cases of (**a**) FDMA (brick-wall shaped in frequency) and (**b**) TDMA (unit sample function in time) scenarios.

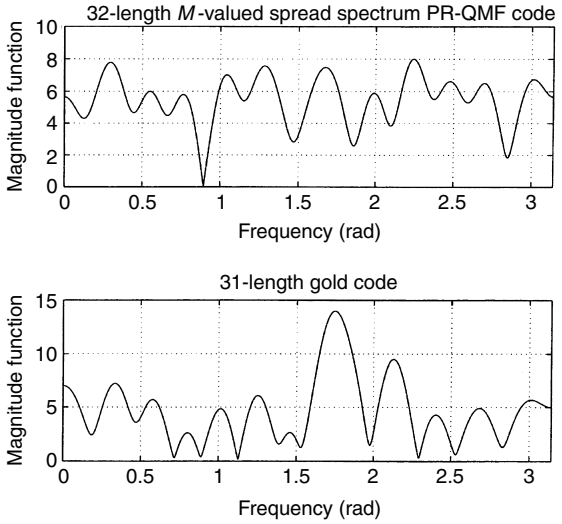


Figure 5. Frequency spectra of a 32-length spread spectrum PR-QMF and 31-length Gold codes.

degrees from the Polytechnic University in 1983 and 1987, respectively, all in electrical engineering. Since 1987, he has been with the New Jersey Institute of Technology, where he is a professor of electrical and computer engineering. He was a co-founder and director of the New Jersey Center for Multimedia Research between 1996 and 2000. He was the vice president for R&D of the IDT Corporation from 2000 to 2001. He has been the founding president of PixWave Inc. His industrial affiliations also include his visits to IBM T.J. Watson Research Center and GEC-Marconi Electronic Systems Corp. during the summers of 1989 and 1996, and 1992, respectively. He serves as a consultant to the industry and sits on the boards of a few Internet startup companies. He co-authored and co-edited three books and many papers on his research. His current research is on signal and transform theories with applications in multimedia and communications.

Husrev T. Sencar received the B.Sc. and M.Sc. degrees from Middle East Technical University, Ankara,

Turkey, and Baskent University, Ankara, Turkey, in 1996 and 1998, respectively, all in electrical and electronics engineering. He currently pursues a Ph.D. in electrical engineering at New Jersey Institute of Technology, Newark, New Jersey. His research interests include signal processing for communications and multimedia with emphasis on information hiding and video/image processing.

BIBLIOGRAPHY

1. B. R. Saltzberg, Performance efficient parallel data transmission system, *IEEE Trans. Commun.* **Com-15**: 805–811, (Dec. 1967).
2. *Special Issue on Transmultiplexers*, *IEEE Trans. Commun.* **Com-30**: (7, July 1982).
3. *Special Issue on Multicarrier Communications*, *Wireless Personal Communications* **2**(1–2): (July 1995).
4. A. Papoulious, *Signal Analysis*, McGraw-Hill, New York, 1977.
5. R. A. Haddad, A. N. Akansu, and A. Benyassine, Time-frequency localization in M-band filter banks and wavelets: A critical review, *J. Opt. Eng.* **32**: 1411–1429 (July 1993).
6. A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands and Wavelets*, 2nd Ed., Academic Press, 2001 347–349.
7. M. G. Bellanger and D. L. Daguet, TDM-FDM multiplexer digital polyphase and FFT, *IEEE Trans. Commun.* **Com-22**: 1199–1205 (Sep. 1974).
8. A. N. Akansu and M. J. T. Smith (eds.), *Subband and Wavelet Transforms: Design and Applications*, Kluwer, 1995.
9. A. N. Akansu and M. J. Medley (eds.), *Wavelet, Subband and Wavelet Transforms in Communication and Multimedia*. Kluwer, 1999.
10. *Special Issue on Theory and Application of Filter Banks and Wavelet Transforms*, *IEEE Trans. Signal Processing* **SP-46**(4): (Apr. 1998).

P

PACKET-RATE ADAPTIVE RECEIVERS FOR MOBILE COMMUNICATIONS

STELLA N. BATALAMA
State University of New York
at Buffalo
Buffalo, New York

1. INTRODUCTION

The effectiveness of a receiver designed for a rapidly changing multiple-access (multiuser) communications environment depends on the following design attributes: (1) system adaptivity under limited data support, (2) multiple-access-interference resistance, and (3) low computational complexity. Short-data-record adaptive designs appear as the natural next step for a matured discipline that has extensively addressed the other two design objectives, 2 and 3, in ideal setups (perfectly known or asymptotically estimated statistical properties). System adaptivity based on short data records is necessary for the development of practical adaptive receivers that exhibit superior signal-to-interference-plus-noise ratio (SINR) or bit error rate (BER) performance when they operate in rapidly changing communications environments that limit substantially the input data support available for adaptation and redesign.

In modern packet data transmission systems where the basic information flow unit is the packet (a group of bits that includes the actual information bits as well as other coding and network control bits), the main measure of link quality is the throughput (either packet throughput or information throughput) that which is directly related to the packet error rate (PER). Real-time voice communications impose stringent delay constraints and require a guaranteed upper bound on PER of about 10^{-2}. On the other hand, data packets can tolerate reasonable delays but may require a lower PER bound [1,2]. Packet throughput improvements can be achieved as a result of BER improvements. On the other hand, BER improvements can be achieved by means of advanced receiver designs that exploit both the characteristics of the transmitted signal and the current state of the environment (these "raw" BER values can be further improved through channel coding (forward error correction)). In dynamic environments, adaptive receiver designs can react to variations as opposed to static receivers that remain unchanged regardless of the changes in the environment. Inherently, a major consideration in the design of successful adaptive receivers is the fact that their adaptation rate must be commensurate to the rate of change of the environment.

An example of a system that can benefit from modern advanced adaptive receiver technology is the direct-sequence code-division multiple-access (DSCDMA) radiofrequency (RF) system. In such a system, the transmitted signal is a spread-spectrum (SS) signal obtained by multiplying each information bit by a unique code (or signature) waveform dedicated to each user. The SS characteristics of the transmitted signal allow intelligent temporal (code) processing at the receiver (unmasking of the signature). During RF transmission, the signal in general undergoes a process known as *multipath-fading* dictated by the physical characteristics of the communication channel. As a result, the received signal consists of multiple faded and delayed copies of the transmitted signal. At the receiver, the multiple copies, instead of being discarded as interference, can be processed in an advantageous manner (a procedure known as RAKE processing). Further performance improvements can be obtained by exploiting the spatial characteristics of the transmitted signal; such processing requires that antenna-array ("smart antenna") technology is employed at the receiver. DSCDMA systems equipped with antenna arrays offer the opportunity for jointly effective spatial (array) and temporal (code) noise suppression. Primary noise sources include additive white Gaussian noise (AWGN) usually due to the receiver front-end electronics as well as interference from other users who transmit similar signals at the same time and in the same frequency spectrum [CDMA systems allow such channel accessing as opposed to time-division multiple-access (TDMA) or frequency-division multiple-access (FDMA) systems]. This general DSCDMA signal model example will be revisited many times throughout our discussion, and a complete adaptive antenna-array DSCDMA receiver will be developed as an illustration.

Returning to the main topic of our discussion, an adaptive receiver consists of a set of building blocks that are a reevaluated (or estimated) every time there is a significant change in the statistics of the environment. The design of each building block is initially formulated mathematically as a solution to an optimization problem under the assumption that all statistical quantities are perfectly known. This is known as the *ideal* or *optimum* solution. Then, the statistical quantities that are present in the optimum solution are substituted by corresponding estimates that are based on the actual received data (observations). This is known as an *estimate* of the optimum solution. It is the latter estimates that need to be adapted according to the changes of the environment, justifying this way the term "adaptive receiver." For example, a popular class of DSCDMA receivers utilizes minimum mean-square-error (MMSE) linear (discrete-time) filters as a means to suppress multiple-access interference (MAI) and AWGN. In other words, the receiver consists of a linear filter that operates on a discrete sequence of spacetime (ST) received signal samples and the optimum filter solution is found by minimizing the mean-square error between the output of the linear filter and a pilot information bit sequence. Several adaptive MMSE filtering algorithms are known and include the sample matrix inversion (SMI), the least-mean-square (LMS) and the recursive-least-square (RLS) algorithms, which will be discussed in detail in subsequent sections. As a general comment, these algorithms/adaptive

filters outperform significantly the popular static ST RAKE filter when a sufficiently large number of data is made available to them [4–9]. Unfortunately, the time-varying nature of the channel frequently necessitates fast (short-data-record) adaptive ST optimization through the use of small input data sets that can "catch up" with the channel variations.

To motivate the developments presented in this article, let us consider a DSCDMA system with 5-element antenna-array reception, system processing gain 64, and, say, 3 resolvable multipaths for the user signal of interest (usually the number of resolvable multipaths is between 2 and 4 including the direct path if any [10]). For such a system, we will see later that jointly optimal S-T processing at the receiver under the MMSE criterion requires processing in the $5(64+3-1)=330$ space-time product space. That is, filter optimization needs to be carried out in the complex $\mathbb{C}^{330}$ vector space. We know that adaptive SMI implementations of the MMSE filter solution require data samples many times the space-time product to approach the performance characteristics of their ideal counterpart (RLS/LMS implementations behave similarly) [11,12]. In fact, theoretically, system optimization with data samples less than the spacetime product may not even be possible, as we will explain later in our discussion. With CDMA chip rates at 1.25 MHz [10], processing gain 64, and typical fading rates of $\geq$70 Hz for vehicle mobiles [13], the fading channel fluctuates decisively at least every 280 data symbols. In this context, conventional SMI/RLS/LMS adaptive filter optimization in the $\mathbb{C}^{330}$ vector space becomes an unrealistic objective.

The goal of our presentation is to first introduce and then elaborate on the underlying principles of short-data-record adaptive filter estimation. Through illustrative examples from the mobile communications literature, we will observe that short-data-record (e.g., packet-rate) filter estimation results in improved channel BER, which translates to higher packet success probability and higher user capacity for a given PER upper bound quality-of-service (QoS) constraint. This ensures an improvement in terms of packet throughput and delay characteristics of a network system that satisfies the QoS constraint. Additional performance improvements can and must be pursued through synergistic use of channel coding (FEC).

While our target applications are all time-critical communications problems, the theoretical developments that will be presented herein may touch many aspects of multidisciplinary engineering that are hampered by the "curse of dimensionality" and could benefit from adaptive filtering and/or adaptive system optimization through limited input data.

2. BASIC SIGNAL MODEL

For illustration purposes, we consider throughout this presentation a multiuser communications system where binary antipodal information symbols from user0, user1, ..., user $Q-1$ are transmitted at a rate $1/T$ by modulating (being multiplied by) a signal waveform $d_q(t)$ of duration T, $q=0,\ldots,Q-1$, that uniquely identifies each user and is assumed to be approximately band-limited or have negligible frequency components outside a certain bandwidth. If H_1 (H_0) denotes the hypothesis that the information bit $b_0=+1$ ($b_0=-1$) of the user of interest, say, user 0, is transmitted during a certain bit period T, then the corresponding equivalent lowpass composite received waveform over the bit interval T may be expressed in general as

$$\begin{aligned} H_1&: x(t)=(+1)\sqrt{E_0}v_0(t)+z(t) \quad \text{and} \\ H_0&: x(t)=(-1)\sqrt{E_0}v_0(t)+z(t), \quad 0\leq t\leq T \end{aligned} \tag{1}$$

With respect to the user of interest, user 0, E_0 denotes transmitted energy per bit, $v_0(t)$ represents the channel processed version of the original waveform $d_0(t)$ [w.l.o.g. the signal waveform $d_0(t)$ is assumed to be normalized to unit energy over the bit period T], and $z(t)$ identifies comprehensively the channel disturbance and includes one, some, or all of the following forms of interference: (1) MAI, (2) intersymbol interference (ISI), and (3) additive Gaussian or non-Gaussian noise.

The continuous-time waveform $x(t)$ is "appropriately" sampled and the discrete samples are grouped to form vectors of "appropriate" length P (both the sampling method and the length value P are pertinent to the specifics of the application under consideration). Let $\mathbf{x}$ denote such a discrete-time complex, in general, received signal vector in $\mathbb{C}^P$:

$$\begin{aligned} H_1&: \mathbf{x}=+\sqrt{E_0}\mathbf{v}_0+\mathbf{z} \quad \text{and} \\ H_0&: \mathbf{x}=-\sqrt{E_0}\mathbf{v}_0+\mathbf{z}, \quad \mathbf{x},\ \mathbf{v}_0,\ \mathbf{z}\in\mathbb{C}^P \end{aligned} \tag{2}$$

where P identifies the dimension of the discrete-time complex observation space, $\mathbf{v}_0$ is the signal vector that corresponds to $v_0(t)$, and $\mathbf{z}$ denotes the discrete-time comprehensive disturbance vector [14]. Our objective is to detect b_0 (i.e., to decide in favor of H_1 or H_0) by means of a linear filter $\mathbf{w}$ as follows:

$$\hat{b}_0=\text{sgn}(\text{Re}\{\mathbf{w}^H\mathbf{x}\}) \tag{3}$$

where $\text{sgn}(\cdot)$ is the ±1 hard-limiter, $\text{Re}\{\cdot\}$ extracts the real part of a complex number, and $(\cdot)^H$ denotes the Hermitian operation. In other words, we fix the structure of the receiver to that given by Fig. 1. Our discussion will be focused on the design of the linear filter $\mathbf{w}$ according to the MMSE or minimum-variance distortionless response (MVDR) optimization criteria we present in the section that follows.

The specific illustrative example of a multipath fading AWGN DS-CDMA packet data communication link with narrowband linear antenna-array reception that we considered earlier, is certainly covered by the above general basic signal model. The transmitted signal waveform of a particular user is obtained as follows. The user is assigned a unique binary antipodal signature (code)

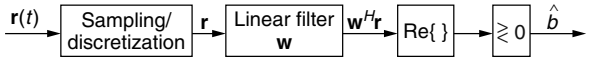


Figure 1. General receiver structure for the (one-shot) detection of binary antipodal information symbols of the user of interest.

sequence, that is a sequence with elements $+1$ or -1 of length L (L is also called *system processing gain*). The bits of the user code multiply a basic signal pulse (e.g., square pulse or raised cosine) of duration T_c, known as *chip*. This way we obtain the signature waveform $d_q(t)$ of duration $T = LT_c$. The transmitted signal waveform that corresponds to a single information bit is the product of the information bit itself and the signature waveform. The corresponding received waveform is the convolution of the transmitted waveform and the impulse response of the multipath fading channel (when the latter is modeled as a linear filter) and is assumed to be band-limited by the chip rate. Discretization of the continuous received waveform at each antenna element of the array can be achieved by chip-matched filtering of the received waveform and sampling at the chip rate (or by lowpass filtering, commensurate Nyquist sampling, and chip-rate accumulation) over the multipath-extended symbol period. The discrete vector outputs from all antenna elements are stacked together (one on top of the other) to create a supervector known as the *spacetime* (ST) received vector. In this way the data are prepared for processing by the linear filter $\mathbf{w}$, extraction of the real part of the filter output and finally sign detection as shown in Fig. 1, a process termed "one shot" detection: detection on a symbol-by-symbol (information bit) basis, as opposed to simultaneous detection of all information bits of the user of interest. If M is the number of antenna elements of the array, L is the length of the signature (code) vector, and N is the number of resolvable multipaths (w.l.o.g. we assume that N is the same for all users) then the discretetime, ST complex received vector $\mathbf{x}$ is of dimension $M(L+N-1)$; where $L+N-1$ is exactly the length of what we referred to earlier as the multipath-extended symbol period. "Inside" $\mathbf{x}$ in (2), $\mathbf{v}_0$ corresponds to the channel-processed (also known as "effective") ST signature vector of the user of interest, while $\mathbf{z}$ corresponds to the discrete-time disturbance vector that accounts for MAI, ISI, and AWGN. Specifically, $\mathbf{v}_0$ can be expressed as a function of the transmitted signal, channel and receiver structure parameters:

$$\mathbf{v}_0 = \sqrt{\frac{E_0}{L}} \sum_{n=0}^{N-1} c_{0,n} \left[\underbrace{0 \ldots 0}_{n} \quad \mathbf{d}_0^T \quad \underbrace{0 \ldots 0}_{N-n-1} \right]^T \odot \mathbf{a}_{0,n} \tag{4}$$

where $c_{0,n}, n = 0, \ldots, N-1$, denote the path coefficients of the channel of the user of interest. The coefficients $c_{0,n}, n = 0, \ldots, N-1$, are frequently modeled as independent zero-mean complex Gaussian random variables (that exhibit Rayleigh distributed amplitude and uniformly distributed phase that fits experimental measurements) and are assumed to remain constant over the entire packet duration. In a realistic environment the coefficients may vary approximately every 300 symbols [15]. Thus, keeping the packet size less than 300 validates the assumption of constant multipath coefficients over the duration of a packet. In (4), $\mathbf{d}_0 = [\mathbf{d}_0[0], \ldots, \mathbf{d}_0[L-1]]^T$ is the binary signature vector (spreading sequence) of the user of interest, $\mathbf{d}_0[l] \in \{\pm 1\}, l = 0, \ldots, L-1$, $\mathbf{a}_{0,n}$ is the array response vector that corresponds to the nth path of the user of interest, and $\odot$ denotes the Krönecker tensor product. The array response vector of the nth path of the user of interest is defined by

$$\mathbf{a}_{0,n}(m) = e^{j2\pi(m-1)\frac{d}{\lambda}\sin\theta_{0,n}}, \quad m = 1, 2, \ldots, M \tag{5}$$

where $\theta_{0,n}$ identifies the angle of arrival of the corresponding path, λ is the carrier wavelength, and d is the element spacing of the antenna array (usually $d = \lambda/2$). More details on the DSCDMA ST received signal model in (4) and the operational characteristics of an antenna-array system can be found elsewhere in the literature [5,16]. Finally, the noise vector $\mathbf{z}$ represents the comprehensive disturbance effect of AWGN and all other user signal contributions that are again of the form of (4), yet with different in general energy values, signature vectors, multipath coefficients, and angles of arrival.

3. FILTERING WITH KNOWN INPUT STATISTICS

3.1. Optimum MMSE/MVDR Filter

Minimum-variance distortionless response (MVDR) *receiver design* refers to the problem of identifying a linear finite-impulse response filter that minimizes the variance at its output, while at the same time the filter maintains a "distortionless" response toward a specific input vector direction of interest. In mathematical terms, if $\mathbf{x}$ is a random, *0*-mean (without loss of generality) complex input vector of dimension P, $\mathbf{x} \in \mathbb{C}^P$, that is processed by a P-tap filter $\mathbf{w} \in \mathbb{C}^P$, then the filter output variance is $E\{|\mathbf{w}^H\mathbf{r}|^2\} = \mathbf{w}^H\mathbf{R}\mathbf{w}$, where $\mathbf{R} = E\{\mathbf{x}\mathbf{x}^H\}$ is the input autocorrelation matrix ($E\{\cdot\}$ denotes the statistical expectation operation). The MVDR filter minimizes $\mathbf{w}^H\mathbf{R}\mathbf{w}$ and simultaneously satisfies an equation of the form $\mathbf{w}^H\mathbf{v}_0 = \rho$, where $\mathbf{v}_0$ is the given input signal vector direction to be protected. In this setup, MVDR filtering is a standard linear constraint optimization problem and the conventional Lagrange multipliers constraint optimization technique leads to the solution (the Lagrange multipliers optimization technique is presented in detail elsewhere [16])

$$\mathbf{w}_{\text{MVDR}} = \rho^* \frac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0} \tag{6}$$

where $(\cdot)^*$ denotes conjugation. Extensive tutorial treatments of MVDR filtering can be found in many sources [e.g., 16,17], along with historical notes on the early work by Capon [18] and Owsley [19].

MVDR filtering has long been a workhorse for blind (unsupervised) communications and signal processing applications where a desired (pilot) scalar filter output $y \in \mathbb{C}$ cannot be identified or cannot be assumed available for each input $\mathbf{x} \in \mathbb{C}^P$. Prime examples include radar and array processing problems where the constraint vector $\mathbf{v}_0$ is usually referred to as the "target" or "look" direction of interest. It is interesting to observe the close relationship between the MVDR filter and the MMSE ("Wiener") filter. Indeed, if the constraint vector $\mathbf{v}_0$ is chosen to be the statistical cross-correlation vector between the desired output y and the input vector $\mathbf{x}$; that is, if $\mathbf{v}_0 = E\{\mathbf{x}y^*\}$,

then the MMSE filter obtained by minimizing the mean-square (MS) error between the filter output $\mathbf{w}^H\mathbf{x}$ and the desired output y is given by

$$c\mathbf{R}^{-1}\mathbf{v}_0, \quad c > 0 \tag{7}$$

that is, the MMSE filter becomes a positive scaled version of the MVDR filter and exhibits identical output SINR performance. For this reason in the rest of our discussion we refer comprehensively to both filters as MMSE/MVDR filters as [16,17].

Conventionally, the computation of the MMSE/MVDR filter in (6) or (7) begins with the calculation of the inverse of the ideal input autocorrelation matrix $\mathbf{R}^{-1}$ (assuming that the Hermitian matrix $\mathbf{R}$ is strictly positive definite, hence invertible). The calculation of the inverse is usually based on numerical iterative diagonalization linear algebra procedures [20]. Then, the matrix $\mathbf{R}^{-1}$ is used for the linear transformation (left multiplication) of the constraint vector $\mathbf{v}_0$, followed by $\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0$ normalization, as necessary.

Linear transformations that involve the inverse of a high-dimension matrix are computationally intensive. In addition, and most importantly, severe complications arise at the adaptive implementation stage when the estimate of such a high-dimension matrix is inverted (particularly when the estimate is based on a small set of data/observations and is obtained, possibly, by some form of sample averaging). One extreme example of such a complication is the fact that the inverse may not even exist. Thus, when the data that are available for adaptation and redesign are limited, use of inverses of (sample average) estimated high-dimension matrices is not viewed favorably (this issue will be discussed in detail in the next section). In such cases, it is preferable to proceed with alternative methods that *approximate* the optimum solution and, hopefully, avoid implicit or explicit use of inverses. Then, at the adaptive implementation stage, we may utilize estimates of the approximate solutions. Algorithmic designs that aim at approximating the optimum MMSE/MVDR filter include (1) the generalized sidelobe canceler (GSC) and its variations, (2) the auxiliary vector (AV) filter, and (3) the orthogonal multistage filter (also "called nested Wiener filter"). The relative performance of these methods in limited data support environments is examined in Section 4.

3.2. Generalized Sidelobe Canceler (GSC)

For a given (not necessarily normalized) constraint vector $\mathbf{v}_0 \in \mathbb{C}^P$, any "distortionless" linear filter $\mathbf{w} \in \mathbb{C}^P$ that satisfies $\mathbf{w}^H\mathbf{v}_0 = \rho$ can be expressed/decomposed as $\mathbf{w} = (\rho^*/\|\mathbf{v}_0\|^2)\mathbf{v}_0 - \mathbf{u}$ for some $\mathbf{u} \in \mathbb{C}^P$ such that $\mathbf{v}_0^H\mathbf{u} = 0$, as shown in Fig. 2 (this decomposition is an application of the projection theorem in linear algebra). There are two general approaches for the design of the filter part $\mathbf{u}$: (1) eigen-decomposition-based approaches and (2) non-eigendecomposition-based approaches.

Algorithmic eigendecomposition-based designs that focus on the MMSE/MVDR filter part $\mathbf{u}$, which is orthogonal to the constraint vector, or "look" direction $\mathbf{v}_0$, include the Applebaum–Howells arrays, beamspace partially adaptive processors, or generalized sidelobe cancelers

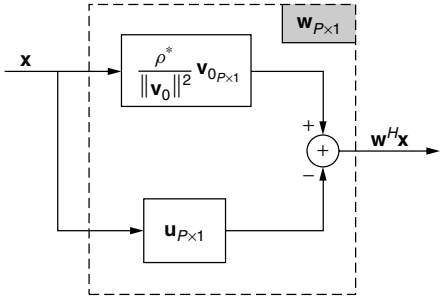


Figure 2. General decomposition of a linear filter $\mathbf{w}$ that satisfies $\mathbf{w}^H\mathbf{v}_0 = \rho$ to two orthogonal components ($\mathbf{u}^H\mathbf{v}_0 = 0$).

(GSCs). More recent developments have been influenced by principal-components analysis (PCA) reduced-rank processing principles. The general goal of these designs is to approximate the MMSE/MVDR filter part $\mathbf{u}$ by utilizing different rank-reducing matrices as explained below [16,17,21–26]. The approximation is of the general form (Fig. 3)

$$\mathbf{u}_{P\times1} \simeq \mathbf{B}_{P\times(P-1)}\mathbf{T}_{(P-1)\times p}\mathbf{w}_{p\times1}^{\mathrm{GSC}} \tag{8}$$

where $\mathbf{B}$ is a matrix that satisfies $\mathbf{B}^H\mathbf{v}_0 = \mathbf{0}_{P-1}$ and is, thus, called "blocking matrix" since it blocks signals in the direction of $\mathbf{v}_0$ ($\mathbf{B}$ is a full column-rank matrix that can be derived by Gram–Schmidt orthogonalization of a $P \times P$ orthogonal projection matrix such as $\mathbf{I} - (\mathbf{v}_0\mathbf{v}_0^H/\|\mathbf{v}_0\|^2)$, where $\mathbf{I}$ is the identity matrix). $\mathbf{T}$ is the rank-reducing matrix with $1 \le p < P-1$ columns that have to be selected and $\mathbf{w}^{\mathrm{GSC}}$ is a vector of weights of the p columns of $\mathbf{T}$ that is designed to minimize the variance at the output of the "overall" filter $\mathbf{w}$, $E\left\{\left|\left(\frac{\rho^*}{|\mathbf{v}_0|^2}\mathbf{v}_0 - \mathbf{u}\right)^H\mathbf{x}\right|^2\right\}$. The solution to the latter optimization problem (assuming that $\mathbf{T}$ is given) is

$$\mathbf{w}^{\mathrm{GSC}} = \frac{\rho^*}{\|\mathbf{v}_0\|^2}[\mathbf{T}^H\mathbf{B}^H\mathbf{R}\mathbf{B}\mathbf{T}]^{-1}\mathbf{T}^H\mathbf{B}^H\mathbf{R}\mathbf{v}_0 \tag{9}$$

We note that the rank-reducing matrix $\mathbf{T}$ "reduces" the dimension of the linear filter (number of parameters to be designed) from P (filter $\mathbf{w}$) to p (filter $\mathbf{w}^{\mathrm{GSC}}$), $1 < p < P-1$. The p columns of the rank-reducing matrix $\mathbf{T}$ can be chosen in various ways. We can choose the p columns to be the eigenvectors that correspond to the P maximum eigenvalues of the *disturbance-only* autocorrelation matrix [27]. This choice is mean-square (MS) optimum under the assumption that the disturbance-only eigenvectors are not rotated by the

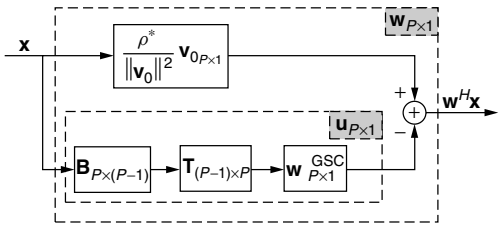


Figure 3. Generalized sidelobe canceler structure.

blocking matrix being used (i.e., when the disturbance subspace is orthogonal to the constraint vector $\mathbf{v}_0$), which is not valid in general. We can address this concern by choosing alternatively the p columns of $\mathbf{T}$ to be the eigenvectors that correspond to the p maximum eigenvalues of the *blocked* data autocorrelation matrix $\mathbf{B}^H\mathbf{R}\mathbf{B}$ [28,29]. If, however, the columns of the rank-reducing matrix $\mathbf{T}$ *have to be* eigenvectors of the blocked-data autocorrelation matrix (there is no documented technical optimality to this approach), then the best way in the minimum output variance sense is to choose the p eigenvectors $\mathbf{q}_i$ of $\mathbf{B}^H\mathbf{R}\mathbf{B}$ is to choose the p eigenvectors $\mathbf{q}_i$ with corresponding eigenvalues λ_i that maximize the ratio $\dfrac{|\mathbf{v}_0^H\mathbf{R}\mathbf{B}\mathbf{q}_i|^2}{\lambda_i}$, $i = 1, \ldots, p$ [30]. This algorithm is also known as "cross-spectral metric" reduced-rank processing [31]. Non-eigendecomposition-based alternatives for the synthesis of $\mathbf{u}$ include the auxiliary vector (AV) filters and the orthogonal multistage filters (also called "nested Wiener filters") [5,32–41].

3.3. Auxiliary Vector (AV) filters

Auxiliary vector (AV) filters are non-eigendecomposition-based filters that approximate the optimum MMSE/MVDR solution [5,32–38]. The AV algorithm is a statistical optimization procedure that generates a sequence of filters (AV filters). Each filter in the sequence has the general structure described in Fig. 2, where the vector $\mathbf{u}$ is approximated by a weighted sum of auxiliary vectors that maintain orthogonality *only* with respect to the distortionless direction $\mathbf{v}_0$ (and they are, in general, *nonorthogonal* to each other). The number of auxiliary vectors used to approximate the filter part $\mathbf{u}$ in Fig. 2 is increasing with the filter index in the sequence. Both the auxiliary vectors and the corresponding weights are subject to design (they are designed according to the maximum cross-correlation and minimum-variance criteria, respectively, as explained in detail below). An important characteristic of the AV algorithm (besides the nonorthogonality of the auxiliary vectors) is that it is a conditional optimization procedure; that is, each filter in the sequence is a function of the previously generated filter. Furthermore, AV filters do not require any explicit or implicit matrix inversion, eigendecomposition, or diagonalization. Finally, under ideal setups (perfect known input autocovariance matrix) the AV filter sequence converges to the MMSE/MVDR optimum solution, [33,34].

A pictorial presentation of generation of the sequence of AV filters is given by Fig. 4a. The sequence is initialized at the appropriately scaled constraint vector $\mathbf{w}_0 = \dfrac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$, which is MMSE/MVDR optimum only when the vector inputs are white (i.e., when $\mathbf{R} = \sigma^2\mathbf{I}$, $\sigma > 0$). Next, we incorporate in $\mathbf{w}_0$ an "auxiliary" vector component $\mathbf{g}_1$ that is orthogonal to $\mathbf{v}_0$, and we form $\mathbf{w}_1 = \mathbf{w}_0 - \mu_1\mathbf{g}_1$, where $\mathbf{g}_1 \in \mathbb{C}^P - \{\mathbf{0}\}$, $\mu_1 \in \mathbb{C}$, and $\mathbf{g}_1^H\mathbf{v}_0 = 0$. We assume for a moment that the orthogonal auxiliary vector $\mathbf{g}_1$ is arbitrary but nonzero and fixed, and we concentrate on the selection of the scalar μ_1. The value of μ_1 that minimizes the variance of the output of the filter $\mathbf{w}_1$ can be found by direct differentiation of the variance $E\{|\mathbf{w}_1^H\mathbf{x}|^2\}$ or simply as the value that minimizes the MS error between $\mathbf{w}_0^H\mathbf{x}$ and $\mu_1^*\mathbf{g}_1^H\mathbf{x}$. This leads to $\mu_1 = \mathbf{g}_1^H\mathbf{R}\mathbf{w}_0/\mathbf{g}_1^H\mathbf{R}\mathbf{g}_1$.

Since $\mathbf{g}_1$ is set to be orthogonal to $\mathbf{v}_0$, the expression of μ_1 shows that if the vector $\mathbf{R}\mathbf{w}_0$ happens to be "on $\mathbf{v}_0$" [i.e., if $\mathbf{R}\mathbf{w}_0 = (\mathbf{v}_0^H\mathbf{R}\mathbf{w}_0)\mathbf{v}_0$ or equivalently $(\mathbf{I} - \mathbf{v}_0\mathbf{v}_0^H)\mathbf{R}\mathbf{w}_0 = \mathbf{0}$], then $\mu_1 = 0$. Indeed, if $\mathbf{R}\mathbf{w}_0 = (\mathbf{v}_0^H\mathbf{R}\mathbf{w}_0)\mathbf{v}_0$, then $\mathbf{w}_0$ is *already* the MMSE/MVDR filter. To avoid this trivial case and continue with our presentation, we assume that $\mathbf{R}\mathbf{w}_0 \neq (\mathbf{v}_0^H\mathbf{R}\mathbf{w}_0)\mathbf{v}_0$. By inspection, we also observe that for the MS-optimum value of μ_1 the product $\mu_1\mathbf{g}_1$ is independent of the norm of $\mathbf{g}_1$. Hence, so is $\mathbf{w}_1$. At this point, we set the auxiliary vector $\mathbf{g}_1$ to be a normalized vector that maximizes the magnitude of the cross-correlation between $\mathbf{w}_0^H\mathbf{x}$ and $\mathbf{g}_1^H\mathbf{x}$ (i.e., $\mathbf{g}_1 = \arg\max_{\mathbf{g}} |\mathbf{w}_0^H\mathbf{R}\mathbf{g}|$) subject to the constraint that $\mathbf{g}_1^H\mathbf{v}_0 = 0$ and $\mathbf{g}_1^H\mathbf{g}_1 = 1$. For the sake of mathematical accuracy, we note that both the criterion function $|\mathbf{w}_0^H\mathbf{R}\mathbf{g}|$ to be maximized as well as the orthogonality constraint are phase-invariant. Without loss of generality, to avoid any ambiguity in our presentation and to have a uniquely defined auxiliary vector, we choose the one and only auxiliary vector $\mathbf{g}_1$ that satisfies the maximization problem and places the cross-correlation value on the positive real line ($\mathbf{w}_0^H\mathbf{R}\mathbf{g}_1 > 0$).

The general inductive step is as follows. At step $k+1$, we define the AV filter $\mathbf{w}_{k+1} = \mathbf{w}_k - \mu_{k+1}\mathbf{g}_{k+1}$, where $\mathbf{g}_{k+1}$ and μ_{k+1} are to be *conditionally* optimized given the previously identified AV filter $\mathbf{w}_k$. The auxiliary vector $\mathbf{g}_{k+1}$ is chosen as the vector that maximizes the magnitude of the cross-correlation between the output of the previous filter $\mathbf{w}_k$ and the output of $\mathbf{g}_{k+1}$ (Fig. 4a), again subject to $\mathbf{g}_{k+1}$ being orthonormal to $\mathbf{v}_0$ *only* (we note that the choice of the norm does not affect the solution since $\mu_k\mathbf{g}_k$, $k = 1, 2, \ldots$, is $\mathbf{g}$-norm-invariant; we also emphasize that $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \mathbf{g}_4 \ldots$, *are not* necessarily orthogonal to each other). The value of μ_{k+1} minimizes the output variance of $\mathbf{w}_{k+1}$ given $\mathbf{w}_k$ and $\mathbf{g}_{k+1}$ (or equivalently minimizes the MS error between $\mathbf{w}_k^H\mathbf{x}$ and $\mu_{k+1}^*\mathbf{g}_{k+1}^H\mathbf{x}$). The solution for $\mathbf{g}_{k+1}$ and μ_{k+1} is given below, while the iterative algorithm for the generation of the infinite sequence of AV filters $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots$ is presented in Fig. 4b (we note that in Fig. 4b we dropped the unnecessary normalization of $\mathbf{g}_1, \mathbf{g}_2, \ldots$ since $\mu_k\mathbf{g}_k$ is independent of the norm of $\mathbf{g}_k$):

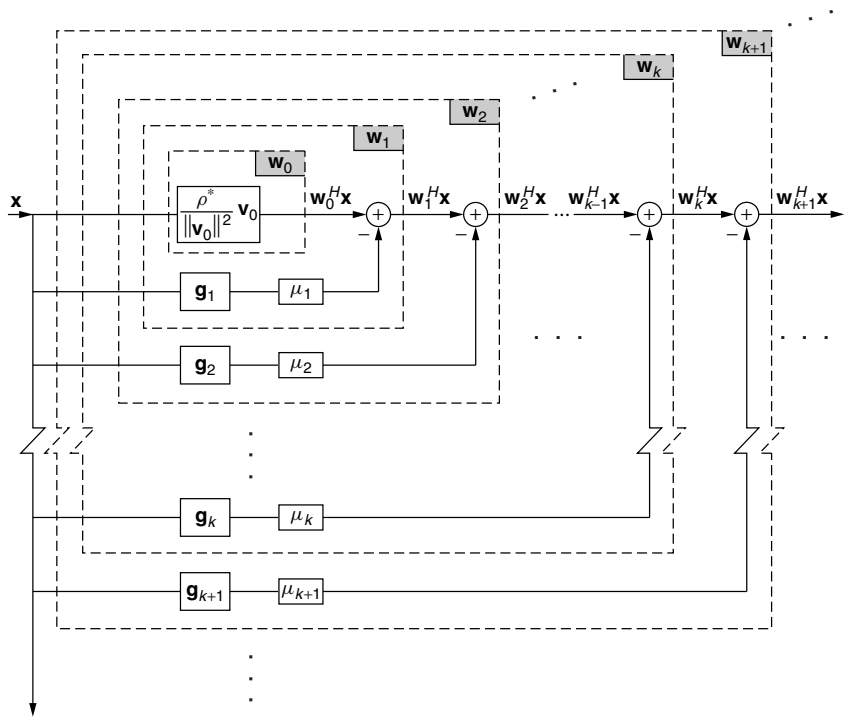
1. The scalar μ_{k+1} that minimizes the variance at the output of $\mathbf{w}_{k+1}$ or equivalently minimizes the MS error between $\mathbf{w}_k^H\mathbf{x}$ and $\mu_{k+1}^*\mathbf{g}_{k+1}^H\mathbf{x}$ is

$$\mu_{k+1} = \frac{\mathbf{g}_{k+1}^H\mathbf{R}\mathbf{w}_k}{\mathbf{g}_{k+1}^H\mathbf{R}\mathbf{g}_{k+1}}, \quad k = 0, 1, 2, \ldots \tag{10}$$

2. Suppose that $\left(\mathbf{I} - \dfrac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\mathbf{R}\mathbf{w}_k \neq \mathbf{0}$ ($\mathbf{w}_k \neq \mathbf{w}_{\text{MVDR}}$). Then, the auxiliary vector

$$\mathbf{g}_{k+1} = \frac{\mathbf{R}\mathbf{w}_k - \dfrac{\mathbf{v}_0^H\mathbf{R}\mathbf{w}_k}{\|\mathbf{v}_0\|^2}\mathbf{v}_0}{\left\|\mathbf{R}\mathbf{w}_k - \dfrac{\mathbf{v}_0^H\mathbf{R}\mathbf{w}_k}{\|\mathbf{v}_0\|^2}\mathbf{v}_0\right\|}, \quad k = 0, 1, 2, \ldots \tag{11}$$

(**a**)



(**b**)

Auxiliary-Vector (AV) algorithm

Input:

Autocovariance matrix $\mathbf{R}$, constraint vector $\mathbf{v}_0$,

desired response $\mathbf{w}^H\mathbf{v}_0 = \rho$.

Initialization:

$$\mathbf{w}_0 := \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0.$$

Iterative computation:

For $k = 1, 2, \ldots$ do

begin

$$\mathbf{g}_k := \left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\mathbf{R}\,\mathbf{w}_{k-1}$$

if $\mathbf{g}_k = 0$ then EXIT

$$\mu_k := \frac{\mathbf{g}_k^H\mathbf{R}\,\mathbf{w}_{k-1}}{\mathbf{g}_k^H\mathbf{R}\,\mathbf{g}_k}$$

$$\mathbf{w}_k := \mathbf{w}_{k-1} - \mu_k\mathbf{g}_k$$

end

Output:

Filter sequence $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots$.

Figure 4. (**a**) Block diagram representation and (**b**) algorithmic description/generation of the auxiliary vector (AV) filter sequence $\mathbf{w}_1, \mathbf{w}_2, \ldots$.

maximizes the magnitude of the cross-correlation between $\mathbf{w}_k^H\mathbf{x}$ and $\mathbf{g}_{k+1}^H\mathbf{x}$ (which is equal to $|\mathbf{w}_k^H\mathbf{R}\mathbf{g}_{k+1}|$), subject to the constraints $\mathbf{g}_{k+1}^H\mathbf{v}_0 = 0$ and $\mathbf{g}_{k+1}^H\mathbf{g}_{k+1} = 1$. In addition, $\mathbf{w}_k^H\mathbf{R}\mathbf{g}_{k+1}$ is real positive ($\mathbf{w}_k^H\mathbf{R}\mathbf{g}_{k+1} > 0$).

With respect to the convergence of the filter sequence $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots$ to the MVDR filter $\rho^*\dfrac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0}$, we can show that [33]

1. The generated sequence of auxiliary vector weights $\{\mu_k\}, k = 1, 2, \ldots,$ is real-valued, positive, and bounded: $0 < \dfrac{1}{\lambda_{\max}} \le \mu_k \le \dfrac{1}{\lambda_{\min}}$, $k = 1, 2, \ldots$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the corresponding maximum and minimum eigenvalues of $\mathbf{R}$
2. The sequence of auxiliary vectors $\{\mathbf{g}_k\}, k = 1, 2, \ldots,$ converges to the $\mathbf{0}$ vector: $\lim_{k\to\infty}\mathbf{g}_k = \mathbf{0}$

3. The sequence of AV filters $\{\mathbf{w}_k\}, k = 1, 2, \ldots,$ converges to the MVDR filter: $\lim_{k\to\infty} \mathbf{w}_k = \rho^* \dfrac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H \mathbf{R}^{-1}\mathbf{v}_0}$.

3.4. Orthogonal Multistage Filters

An alternative mechanism to approximate the optimum MMSE/MVDR solution can be obtained through the use of the orthogonal "multistage" filter decomposition procedure [39,40] (also called "nested Wiener filter"). It can be shown theoretically that the l-stage filter in [39,40], $\mathbf{w}_{l-\text{stage}}$, $0 \le l \le P-1$, is equivalent to the following structure. First, change the auxiliary vector generation recursion in (11) or Fig. 4b to impose orthogonality with respect not only to the constraint vector $\mathbf{v}_0$ but also to *all previously defined* auxiliary vectors that we denote now as $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{k-1}, k \le P-1$:

$$\mathbf{y}_k = \left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2} - \sum_{i=1}^{k-1} \frac{\mathbf{y}_i\mathbf{y}_i^H}{\|\mathbf{y}_i\|^2}\right)\mathbf{R}\mathbf{w}_{k-1} \tag{12}$$

Next, terminate the recursion at $k = l$, $0 \le l \le P-1$, and organize the l (orthogonal to each other and to $\mathbf{v}_0$) vectors $\mathbf{y}_1, \ldots, \mathbf{y}_l$ in the form of a blocking matrix $\mathbf{B}_{P\times l} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_l]$. Then

$$\mathbf{w}_{l-\text{stage}} = \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0 - \mathbf{B}_{P\times l}\vec{\boldsymbol{\alpha}}_{l\times 1} \tag{13}$$

where

$$\vec{\boldsymbol{\alpha}} = \frac{\rho^*}{\|\mathbf{v}_0\|^2}[\mathbf{B}^H\mathbf{R}\mathbf{B}]^{-1}\mathbf{B}^H\mathbf{R}\mathbf{v}_0 \tag{14}$$

is the MS *vector*-optimum set of weights of the vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_l$ [5,37]. We note that "vector-optimum" implies that the elements of the vector $\vec{\boldsymbol{\alpha}}$ (weights of the columns of $\mathbf{B}$) are designed/optimized jointly (and *not* in a conditional one-by-one manner). The multistage decomposition algorithm [39,40] is a computationally efficient procedure for the calculation of the weight vector $\vec{\boldsymbol{\alpha}}$ tailored to the particular structure of $\mathbf{B}^H\mathbf{R}\mathbf{B}$ (tridiagonal matrix); the calculation incorporates an implicit matrix inversion operation [in view of (14)]. The same computational savings can be achieved by the general forward calculation algorithm of Liu and Van Veen [42] that returns all intermediate stage filters along the way, up to the stage of interest l.

We conclude this section with a few comments on the relative merits and characteristics of the structures presented so far. From a general input space synthesis/decomposition point of view, the main distinguishing features of the AV algorithm with respect to the multistage algorithm are the *non-orthogonal* AV synthesis approach and the *conditional statistical optimization* procedure. Nonorthogonal synthesis allows the designer to grow an *infinite* sequence of AV filters on a "best effort" basis that takes into account the whole interference space at every step. Conditional optimization results in estimators that do not require any implicit or explicit matrix inversion or decomposition operation and, thus, plays a key role in developing superior adaptive filtering schemes in short-data-record environments, as illustrated in the next section.

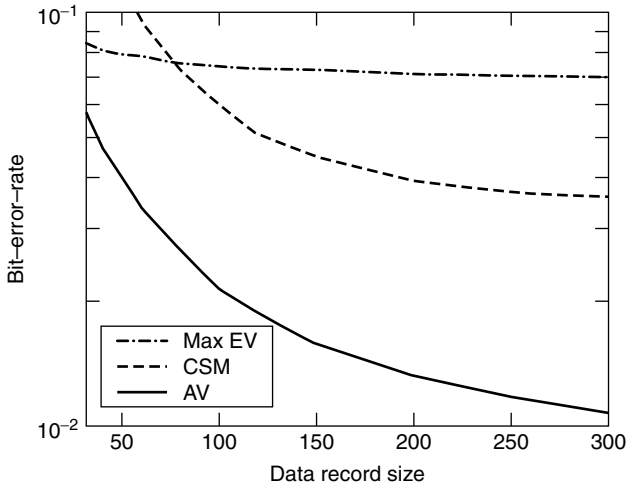


Figure 5. BER as a function of the data record size for the AV, "maximum eigenvector" (MaxEV), and "cross-spectral metric" (CSM) receivers of the same order (3 auxiliary vectors, 3 eigenvectors). *Operational environment*: synchronous DS-CDMA system, user of interest at 12 dB, 12 interferers at 10–14 dB, processing gain $L = 32$, and arbitrary normalized signatures (cross-correlation with the signature of the user of interest ~0.2).

Figure 5 presents an illustrative example of the relative merits of various receiver designs in terms of BER. The example is based on a simple single-path synchronous DSCDMA signal model ($P = L$). The BER performance of the receiver $\mathbf{w}_3$ (that utilizes three auxiliary vectors $\mathbf{g}_1, \mathbf{g}_2$, and $\mathbf{g}_3$) is compared with the BER performance of the "maximum eigenvector" (Max EV) [28,29] receiver and the "cross-spectral-metric" (CSM) [30,31] receiver (both of which use three eigenvectors). As a numerical example that illustrates the convergence of the AV-filter sequence to the ideal MMSE/MVDR solution under perfectly known (ideal) autocorrelation matrix $\mathbf{R}$, in Fig. 6 we plot the squared norm error between the AV filter of the user of interest $\mathbf{w}_k$ and $\mathbf{w}_{\text{MMSE/MVDR}}$ as a function of k (i.e., the number of auxiliary vectors used or equivalently the index of the AV filter in the sequence).

4. ADAPTIVE FILTER ESTIMATION

4.1. Known Channel

4.1.1. SMI, GSC, Auxiliary Vector, and Multistage Estimators. We recall that the MMSE/MVDR filter is a function of the *true* input autocorrelation matrix $\mathbf{R}$ and the *true* constraint vector, $\mathbf{v}_0$. However, in almost every practical adaptive filtering application neither $\mathbf{R}$ nor $\mathbf{v}_0$ is known to the receiver. In this section (4.1) we present various estimates of the optimum MMSE/MVDR filter when $\mathbf{R}$ is unknown and sample average estimated from a data packet (data record) of size J, that is, $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$:

$$\hat{\mathbf{R}}(J) = \frac{1}{J}\sum_{j=0}^{J-1} \mathbf{x}_j\mathbf{x}_j^H \tag{15}$$

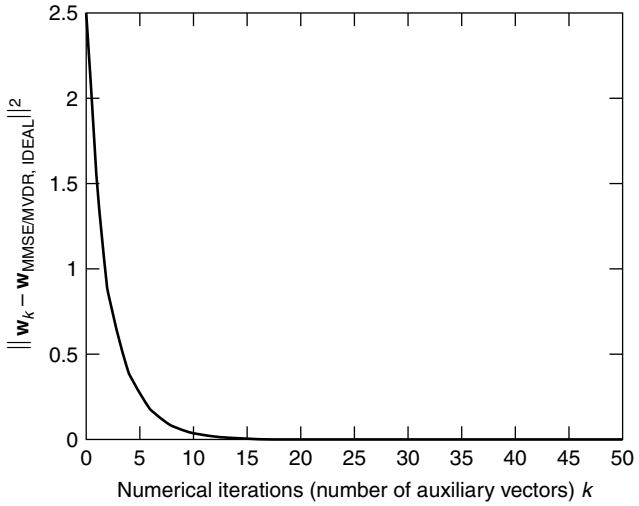


Figure 6. Convergence of the AV filter sequence to the ideal MMSE/MVDR solution as a function of the number of iterations k in Fig. 4b. The sequence of *conditionally optimized* AV filters (that utilize *nonorthogonal* auxiliary vectors) converges to the $\mathbf{w}_{\mathrm{MMSE/MVDR}}$ optimum solution for a perfectly known input autocorrelation matrix $\mathbf{R}$.

Throughout this (4.1) section, $\mathbf{v}_0$ is assumed to be known (since $\mathbf{v}_0$ is a function of the channel parameters we label this section as the "known channel" case); a procedure for the estimation of $\mathbf{v}_0$ from the same data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$ will be presented in the Section 4.2 ("unknown channel").

When $\mathbf{R}$ is unknown, the most widely used MMSE/MVDR filter estimator is obtained from (6) by using the sample average estimate $\hat{\mathbf{R}}(J)$ in place of $\mathbf{R}$. This estimator is known as the sample matrix inversion (SMI) filter. If we choose to work with the approximate solutions presented in Section 3 and utilize the sample average estimate of the autocorrelation matrix $\hat{\mathbf{R}}(J)$ instead of $\mathbf{R}$ in Eqs. (9)–(14), we obtain a GSC, AV, or multistage-type estimator of the MMSE/MVDR solution, respectively. We note that, for Gaussian inputs, $\hat{\mathbf{R}}(J)$ is a maximum-likelihood (ML), consistent, unbiased estimator of $\mathbf{R}$. On the other hand, the inverse of $\hat{\mathbf{R}}(J)$, which is utilized explicitly by the SMI filter and implicitly by both the GSC and the orthogonal multistage decomposition estimator, is not always defined. We can guarantee (with probability 1) that $\hat{\mathbf{R}}(J)$ is invertible only when the number of observations J is greater than or equal to the dimension of the input space (or filter dimension) P and the input distribution belongs to a specific class of multivariate elliptically contoured distributions that includes the Gaussian [43–46]. On the basis of the convergence properties of the AV filter sequence discussed in the previous section we can show that the corresponding sequence of AV filter estimates $\hat{\mathbf{w}}_k(J)$ converges, as $k \to \infty$, to the SMI filter [33]:

$$\hat{\mathbf{w}}_k(J) \xrightarrow[k\to\infty]{} \hat{\mathbf{w}}_\infty(J) = \hat{\mathbf{w}}_{\mathrm{SMI}} = \rho^* \frac{[\hat{\mathbf{R}}(J)]^{-1}\mathbf{v}_0}{\mathbf{v}_0^H[\hat{\mathbf{R}}(J)]^{-1}\mathbf{v}_0} \tag{16}$$

4.1.2. Properties of the Sequence of AV Estimators. The AV filter sequence of estimators begins with $\hat{\mathbf{w}}_0(J) = \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$, which is a zero-variance, fixed-value, estimator that may be severely biased ($\hat{\mathbf{w}}_0(J) \neq \mathbf{w}_{\mathrm{MMSE/MVDR}}$) unless the input is white (i.e., $\mathbf{R} = \sigma^2\mathbf{I}$, for some $\sigma > 0$). In the latter trivial case, $\hat{\mathbf{w}}_0(J)$ is already the perfect MMSE/MVDR filter. Otherwise, the next filter estimator in the sequence, $\hat{\mathbf{w}}_1(J)$, has a significantly reduced bias due to the optimization procedure employed, at the expense of nonzero estimator (co)variance. As we move up in the sequence of filter estimators $\hat{\mathbf{w}}_k(J)$, $k = 0, 1, 2, \ldots$, the bias decreases rapidly to zero[1] while the variance rises slowly to the SMI [$\hat{\mathbf{w}}_\infty(J)$] levels [cf. (16)]. To quantify these remarks, we plot in Fig. 7 the norm-square bias $\|E\{\hat{\mathbf{w}}_k(J)\} - \mathbf{w}_{\mathrm{MMSE/MVDR}}\|^2$ and the trace of the covariance matrix $E\{[\hat{\mathbf{w}}_k(J) - E\{\hat{\mathbf{w}}_k(J)\}][\hat{\mathbf{w}}_k(J) - E\{\hat{\mathbf{w}}_k(J)\}]^H\}$ as a function of the iteration step (filter index) k, for the same signal model as in Fig. 6 and data packet (data record) size $J = 256$. Bias and covariance trace values are calculated from 100,000 independent filter estimator realizations for each iteration point k; that is, we generate 100,000 independent data packets (J received random vectors per packet). For each packet we evaluate $\hat{\mathbf{w}}_1(J), \hat{\mathbf{w}}_2(J), \ldots$. Then, we evaluate expectations as sample averages over 100,000 data packets.

Formal, theoretical statistical analysis of the generated estimators $\hat{\mathbf{w}}_k(J), k = 0, 1, 2, \ldots$, is beyond the scope of this presentation. We do note, however, that for multivariate Gaussian input distributions, an analytic expression for the covariance matrix of the SMI estimator $\hat{\mathbf{w}}_\infty(J)$ can be found in [46]

$$E\{[\hat{\mathbf{w}}_\infty(J) - E\{\hat{\mathbf{w}}_\infty(J)\}][\hat{\mathbf{w}}_\infty(J) - E\{\hat{\mathbf{w}}_\infty(J)\}]^H\} = \frac{|\rho|^2}{(\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0)(J-P+1)}\left(\mathbf{R}^{-1} - \frac{\mathbf{R}^{-1}\mathbf{v}_0\mathbf{v}_0^H\mathbf{R}^{-1}}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0}\right) \tag{17}$$

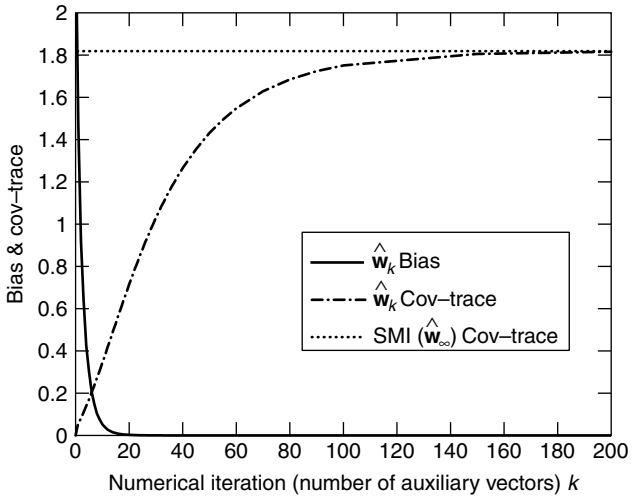


Figure 7. Norm-square bias and covariance trace for the sequence of estimators $\hat{\mathbf{w}}_k(J), k = 0, 1, \ldots$. The signal model is as in Fig. 6; data record size $J = 256$.

[1] The SMI estimator is unbiased for multivariate elliptically contoured input distributions [46,47]: $E\{\hat{\mathbf{w}}_\infty(J)\} = \mathbf{w}_{\mathrm{MMSE/MVDR}} = \rho^*\frac{\mathbf{R}^{-1}\mathbf{v}_0}{\mathbf{v}_0^H\mathbf{R}^{-1}\mathbf{v}_0}$.

Since under these input distribution conditions $\hat{\mathbf{w}}_\infty(J)$ is unbiased, the trace of the covariance matrix is the MS filter estimation error. It is important to observe that the covariance matrix and, therefore, the MS filter estimation error depend on the data record size J, the filter length P, as well as the specifics of the signal processing problem at hand (actual $\mathbf{R}$ and $\mathbf{v}_0$). It is also important to note that when the input distribution is not Gaussian (e.g., for the CDMA signal model example considered earlier in our discussion, the input is Gaussian-mixture-distributed), then the analytic result in (17) is not directly applicable and can be thought of as only an approximation (a rather close approximation for DSCDMA systems). From the results in Fig. 7 for $J = 256$, we see that the estimators $\hat{\mathbf{w}}_1(J), \hat{\mathbf{w}}_2(J), \ldots,$ up to about $\hat{\mathbf{w}}_{20}(J)$ are particularly appealing. In contrast, the estimators $\hat{\mathbf{w}}_k(J)$ for $k > 20$ do not justify their increased covariance trace cost since they have almost nothing to offer in terms of further bias reduction.

We emphasize that since the AV filters $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots$ can be considered as approximations of the MMSE/MVDR optimum filter under idealset-ups, the AV-filter estimates $\hat{\mathbf{w}}_1(J), \hat{\mathbf{w}}_2(J), \hat{\mathbf{w}}_3(J), \ldots$ have been viewed so far not only as estimates of the filters $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots$ but also, and most importantly, as estimates of the MMSE/MVDR optimum filter in (6) and (7). In this context, the mean-square estimation error expression $E\{\|\hat{\mathbf{w}}_k(J) - \mathbf{w}_{\mathrm{MMSE/MVDR}}\|^2\}$ captures the bias/variance balance of the individual members of the estimator sequence $\hat{\mathbf{w}}_k(J)$, $k = 0, 1, 2, \ldots$. In Fig. 8 we plot the MS estimation error as a function of the iteration step k (or filter index) for the same signal model as in Fig. 6, for $J = 256$ [part (a)] and $J = 2048$ [part (b)]. As a reference, we also include the MS-error of the constraint LMS estimator and the RLS estimator.

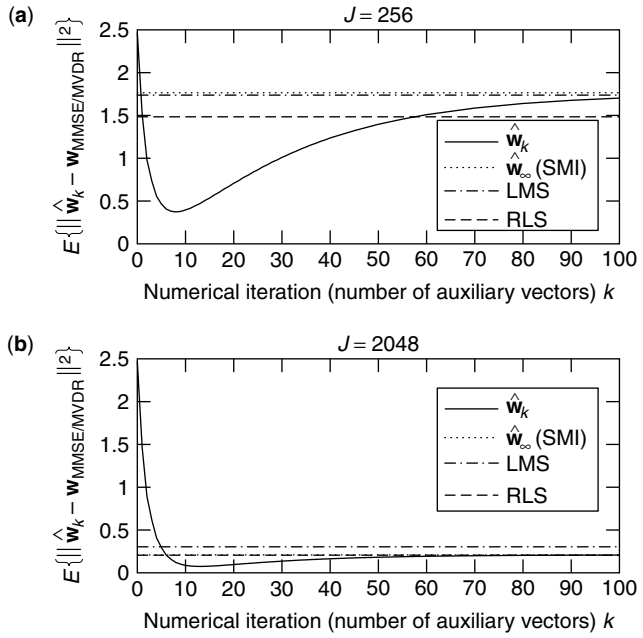


Figure 8. MS estimation error for the sequence of estimators $\hat{\mathbf{w}}_k(J)$, $k = 0, 1, \ldots$: (**a**) data record size $J = 256$; (**b**) $J = 2048$.

The constraint LMS estimator is given by the following recursion:

$$\hat{\mathbf{w}}_{\mathrm{LMS}}(j) = \left(\mathbf{I} - \frac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\left[\hat{\mathbf{w}}_{\mathrm{LMS}}(j-1) - \mu\mathbf{x}_j\mathbf{x}_j^H\hat{\mathbf{w}}_{\mathrm{LMS}}(j-1)\right] + \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0, \quad j = 1, \ldots, J \tag{18}$$

with $\hat{\mathbf{w}}_{\mathrm{LMS}}(0) = \dfrac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$ and some $\mu > 0$. The recursion of the RLS estimator can be obtained from the SMI formula in (16) by utilizing the following iterative estimation of $\mathbf{R}^{-1}$ that is based on the matrix inversion lemma:

$$\hat{\mathbf{R}}^{-1}(j) = \hat{\mathbf{R}}^{-1}(j-1) - \frac{\hat{\mathbf{R}}^{-1}(j-1)\mathbf{x}_j\mathbf{x}_j^H\hat{\mathbf{R}}^{-1}(j-1)}{1 + \mathbf{x}_j^H\hat{\mathbf{R}}^{-1}(j-1)\mathbf{x}_j}, \quad j = 1, \ldots, J \tag{19}$$

where $\hat{\mathbf{R}}^{-1}(0) = \dfrac{1}{\varepsilon_0}\mathbf{I}$ for some $\varepsilon_0 > 0$. Theoretically, the LMS gain parameter $\mu > 0$ has to be less than $\dfrac{1}{2 \cdot \lambda_{\max}^{\mathrm{blocked}}}$, where $\lambda_{\max}^{rmblocked}$ is the maximum eigenvalue of the "blocked data" autocorrelation matrix $\left(\mathbf{I} - \dfrac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)\mathbf{R}\left(\mathbf{I} - \dfrac{\mathbf{v}_0\mathbf{v}_0^H}{\|\mathbf{v}_0\|^2}\right)$. While this is a theoretical upper bound, practitioners are well aware that empirical, data-dependent "optimization" or "tuning" of the LMS gain $\mu > 0$ or the RLS initialization parameter $\varepsilon_0 > 0$ is necessary to achieve acceptable performance (in our study we set $\mu = \dfrac{1}{200 \cdot \lambda_{\max}^{\mathrm{blocked}}}$ and $\varepsilon_0 = 20$, respectively) [8,9,48–51]. This data-specific tuning frequently results in misleading, overoptimistic conclusions about the short-data-record performance of the LMS/RLS algorithms. In contrast, when the AV filter estimators $\hat{\mathbf{w}}_k(J)$ generated by the algorithm of Fig. 4b are considered, tuning of the real-valued parameters μ and ε_0 is virtually replaced by an integer choice among the first several members of the $\{\hat{\mathbf{w}}_k(J)\}$ sequence. In Fig. 8a, for $J = 256$ all estimators $\hat{\mathbf{w}}_k(J)$ from $k = 2$ up to about $k = 55$ outperform in mean-square error (MSE) or their RLS, LMS, and SMI $[\hat{\mathbf{w}}_\infty(J)]$ counterparts. $\hat{\mathbf{w}}_8(J)$ ($k = 8$ auxiliary vectors) has the least MSE of all (best bias/variance tradeoff). When the data record size is increased to $J = 2048$ (Fig. 8b), we can afford more iterations and $\hat{\mathbf{w}}_{13}(J)$ offers the best bias/variance tradeoff (lowest MSE). All filter estimators $\hat{\mathbf{w}}_k(J)$ for $k > 8$ outperform the LMS/RLS/SMI $[\hat{\mathbf{w}}_\infty(J)]$ estimators. For such large data record sets ($J = 2048$), the RLS and the SMI $[\hat{\mathbf{w}}_\infty(J)]$ MSE are almost identical. Figure 9 offers a three-dimensional plot of the mean-square estimation error as a function of the sample support J used in forming $\hat{\mathbf{w}}_k(J)$ and the number of auxiliary vectors k (or filter index). The dark line that traces the bottom of the MS estimation error surface identifies the best number of auxiliary vectors (or the index of the best filter) for any given data record size J.

4.1.3. How to Choose the Best AV Estimator. We recall that, when the autocovariance matrix is sample-average-estimated, the sequence of AV estimators converges

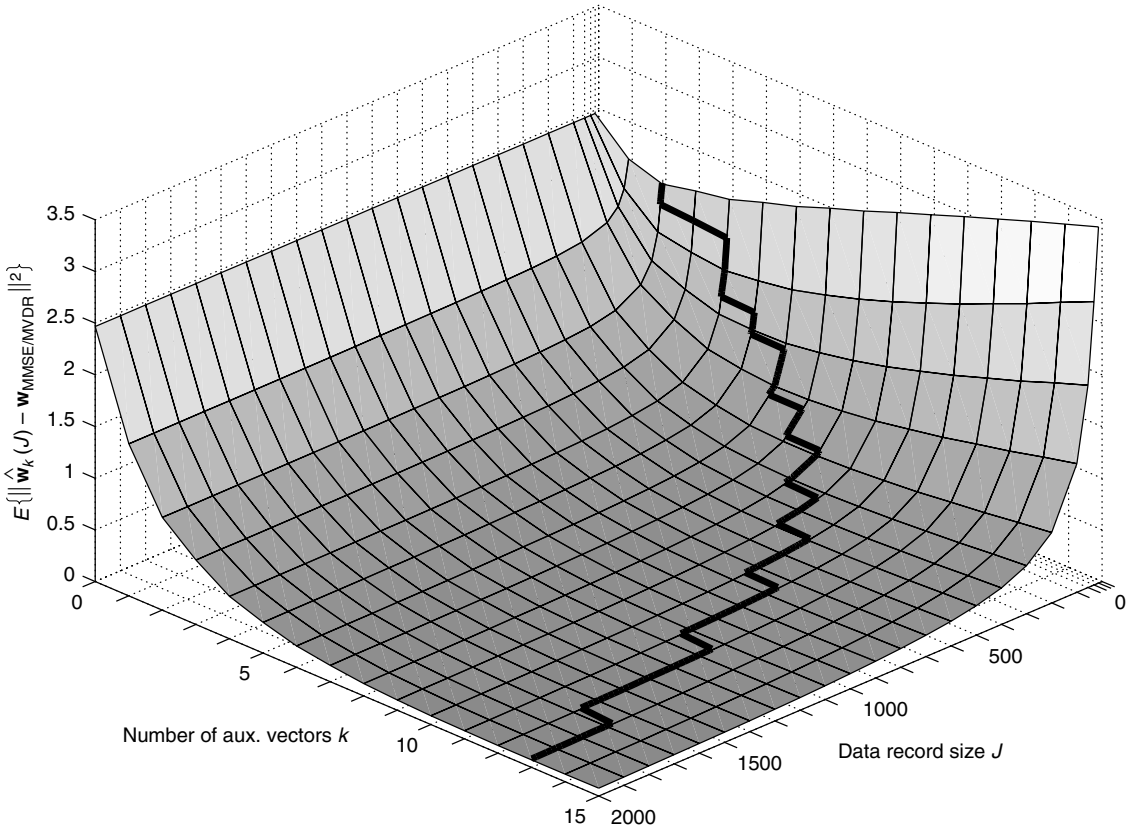


Figure 9. MS estimation error versus number of auxiliary vectors k and sample support J (the signal model is the same as in Fig. 6).

to the SMI filter. Evidently, the early, nonasymptotic elements of the sequence offer favorable bias/variance balance characteristics and outperform in mean-square filter estimation error, as we have seen in Figs. 7–9, the unbiased SMI filter estimator as well as the (constraint) LMS, RLS. We will later see that they also outperform the orthogonal multistage decomposition, and diagonally loaded (DL) SMI filter estimators. In the context of digital wireless communication receivers, superior mean-square filter estimation error translates to superior BER performance under short-data-record receiver adaptation. Selecting the most successful (in some appropriate sense) AV estimator in the sequence for a given data record is a critical problem. Below we present two data-dependent selection criteria [52,53]. The first criterion minimizes the cross-validated sample average variance of the AV filter output and can be applied to general filter estimation problems; the second criterion maximizes the estimated $\mathcal{J}$ divergence of the AV filter output conditional distributions and is tailored to general hypothesis testing (detection) applications.

In particular, the *cross-validated minimum output variance* (CVMOV) rule is motivated by the fact that minimization of the output variance of filters that are constrained to be distortionless in the vector direction of a signal of interest is equivalent to maximization of the output SINR. Cross-validation is a well-known statistical method. In the context of AV filtering, cross-validation is used to select the filter parameter of interest (number of auxiliary vectors k) that minimizes the output variance, which is estimated on the basis of the observations (training data) that have not been used in the process of building the filter itself. A particular case of this general method used in this presentation is the "leave one out" method [54]. The following criterion outlines the CVMOV AV filter selection process.

Criterion 1. For a given data packet (data record) of size J, the cross-validated minimum-output-variance AV filter selection rule chooses the AV filter estimator $\hat{\mathbf{w}}_{k_1}(J)$ that minimizes the cross-validated sample average output variance:

$$k_1 = \arg\min_k \left\{ \sum_{j=1}^{J} \hat{\mathbf{w}}_k(J\backslash j)^H \mathbf{x}_j \mathbf{x}_j^H \hat{\mathbf{w}}_k(J\backslash j) \right\} \tag{20}$$

where $(J\backslash j)$ identifies the AV filter estimator that is evaluated from the available data record after removing the jth sample.

While the CVMOV criterion can be applied to general filter estimation problems, the second criterion, the *maximum $\mathcal{J}$-divergence* criterion, is tailored to applications that can be formulated as binary hypothesis testing problems on AV-filtered data. For any scalar binary hypothesis testing problem, if f_0 and f_1 denote the conditional distributions of the detector input under

hypothesis H_0 and H_1, respectively, then the $\mathcal{J}$-divergence distance between f_0 and f_1 is defined as the sum of the Kullback–Leibler (KL) distances between f_0 and f_1 [55]

$$\mathcal{D}(f_0, f_1) \triangleq \mathcal{KL}(f_1, f_0) + \mathcal{KL}(f_0, f_1) \tag{21}$$

where the KL distance of f_1 from f_0 is defined as $\mathcal{KL}(f_1, f_0) \triangleq \int_{-\infty}^{\infty} f_1(x) \log \frac{f_1(x)}{f_0(x)}\, dx$.

The choice of the output $\mathcal{J}$ divergence as one of the underlying rules for the selection of the AV filter is motivated by the fact that the probability of error of the optimum (Bayesian) detector for any scalar binary hypothesis testing problem is lower bounded by

$$P_e \geq \pi_0 \pi_1 \exp\left\{\frac{-\mathcal{D}(f_0, f_1)}{2}\right\} \tag{22}$$

where π_0 and π_1 are the a priori probabilities of H_0 and H_1, respectively. The right-hand side of (22) is a monotonically decreasing function of the $\mathcal{J}$ divergence between the conditional distributions of the detector input. When the conditional distributions under H_0 and H_1 are Gaussian with the same variance, (22) is satisfied with equality. The latter implies that the larger the $\mathcal{J}$ divergence, the smaller the probability of error or, equivalently, the larger the $\mathcal{J}$ divergence, the easier the detection problem. Thus, maximization of the $\mathcal{J}$ divergence implies minimization of the probability of error. Because of the above mentioned properties and their relationship to the probability of error of the optimum detector, $\mathcal{J}$ divergence has been extensively used in the detection literature as a hypothesis discriminant function. In the context of AV filtering, we denote the AV scalar filter output conditional distributions under H_0 and H_1 by $f_{0,k}(\cdot)$ and $f_{1,k}(\cdot)$, respectively, where the index k indicates the dependence of the distributions on the specific AV filter $\hat{\mathbf{w}}_k$ used from the available sequence $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots$. Then, the $\mathcal{J}$ divergence between $f_{0,k}(\cdot)$ and $f_{1,k}(\cdot)$ is also a function of k; for this reason, in the rest of our presentation it will be denoted as $\mathcal{D}(k)$. To the extent that the conditional distributions of the AV filter output under H_0 and H_1 are approximated by Gaussian distributions with opposite means and equal variances (which is a reasonable, in general, assumption), we can show in a straightforward manner that

$$\mathcal{D}(k) \approx \frac{4E^2\{b_0 \mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}]\}}{\mathrm{Var}\{b_0 \mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}]\}} \tag{23}$$

where $\mathrm{Var}(\cdot)$ denotes variance. The following criterion outlines the $\mathcal{J}$-divergence AV-filter selection process.

Criterion 2. For a given data packet (data record) of size J, the $\mathcal{J}$ divergence AV filter selection rule chooses the AV filter estimator $\hat{\mathbf{w}}_{k_2}(J)$ that maximizes the estimated $\mathcal{J}$ divergence $\hat{\mathcal{D}}(k)$ between the AV filter output conditional distributions:

$$k_2 = \arg\max_k \{\hat{\mathcal{D}}(k)\} \tag{24}$$

If we substitute b_0 in (23) by $\hat{b}_0 = \mathrm{sgn}(\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}])$ and evaluate expectations by sample averaging, then we can obtain a blind estimate of the $\mathcal{J}$ divergence:

$$\hat{\mathcal{D}}_B(k) = \frac{4\left[\frac{1}{J}\sum_{j=1}^{J} |\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}_j]|\right]^2}{\frac{1}{J}\sum_{j=1}^{J} |\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}_j]|^2 - \left[\frac{1}{J}\sum_{j=1}^{J} |\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}_j]|\right]^2} \tag{25}$$

where the subscript B identifies the blind version of the $\mathcal{J}$-divergence function. Then, we can evaluate $k_2 = \arg\max_k \{\hat{\mathcal{D}}_B(k)\}$. We recall that in (25) $\mathbf{x}$ denotes the received signal vector of the general form $\mathbf{x} = b_0\sqrt{E_0}\mathbf{v}_0 + \mathbf{z}$ where [cf. (2)] $\mathbf{v}_0 \in \mathbb{C}^P$ is a known deterministic signal vector, $E_0 > 0$ represents the unknown energy scalar, $\mathbf{z} \in \mathbb{C}^P$ is a zero-mean disturbance vector (i.e., it may incorporate ISI, MAI, and additive noise effects), and b_0 is $+1$ or -1 with equal probability. We also recall [cf. (3)] that the decision on $H_0(b_0 = -1)$ or $H_1(b_0 = +1)$ is based on the real part of the AV filter output $\mathrm{Re}[\hat{\mathbf{w}}_k^H(J)\mathbf{x}]$, where $\hat{\mathbf{w}}_k(J)$ is the AV estimator that utilizes k auxiliary vectors.

4.1.4. Properties of the Multistage and the DL-SMI Estimators. A *finite set* of P filter estimators with varying bias/variance balance can be obtained through the use of the orthogonal "multistage" filter decomposition procedure [40]. In the context of filter estimation from a data record of size J, $\hat{\mathbf{w}}_{0-\mathrm{stage}}(J)$ is the matched filter and $\hat{\mathbf{w}}_{(P-1)-\mathrm{stage}}(J)$ is the SMI estimator. In Fig. 10b we plot the MS estimation error of $\hat{\mathbf{w}}_{l-\mathrm{stage}}(J)$ as a function of l, $0 \leq l \leq P-1 = 31 (J = 60)$. We identify the *best* multistage estimator ($l = 3$ stages), and in Fig. 10c we compare against the AV estimator sequence. We see that all AV estimators $\hat{\mathbf{w}}_k(J)$ from $k = 3$ to 8 outperform in MSE the best multistage estimator ($l = 3$ stages).

An alternative bias/variance trading mechanism through real-valued tuning is the diagonally-loaded (DL) SMI estimator obtained by adding an amount (Δ) to each element of the diagonal of $\hat{\mathbf{R}}(J)$ in the SMI formula (16) [56] [in this way we ensure the invertibility of $\hat{\mathbf{R}}(J)$]

$$\hat{\mathbf{w}}_{\mathrm{DL-SMI}}(\Delta) = \rho^* \frac{[\hat{\mathbf{R}}(J) + \Delta\mathbf{I}]^{-1}\mathbf{v}_0}{\mathbf{v}_0^H[\hat{\mathbf{R}}(J) + \Delta\mathbf{I}]^{-1}\mathbf{v}_0} \tag{26}$$

where $\Delta \geq 0$ is the diagonal loading parameter. We observe that $\hat{\mathbf{w}}_{\mathrm{DL-SMI}}(\Delta = 0)$ is the regular SMI estimator, while $\lim_{\Delta\to\infty} \hat{\mathbf{w}}_{\mathrm{DL-SMI}}(\Delta) = \frac{\rho^*}{\|\mathbf{v}_0\|^2}\mathbf{v}_0$, which is the properly scaled matched filter. In Fig. 10a we plot the MS estimation error of the DL-SMI estimator as a function of the diagonal loading parameter $\Delta (J = 60)$. We identify the *best possible* diagonal loading value $\Delta \simeq 3.45$ (at significant computational cost), and in Fig. 10c we compare the best DL-SMI estimator against the AV estimator sequence for which *no* diagonal loading is performed. Interestingly, $\hat{\mathbf{w}}_k(J)$ from $k = 4$–7 outperform in MSE the best possible DL-SMI estimator ($\Delta \simeq 3.45$). In Fig. 11 we plot the MS error of the $\Delta = 3.45$ DL-SMI estimator together with the MS error of the *best* multistage and AV estimators over the data support

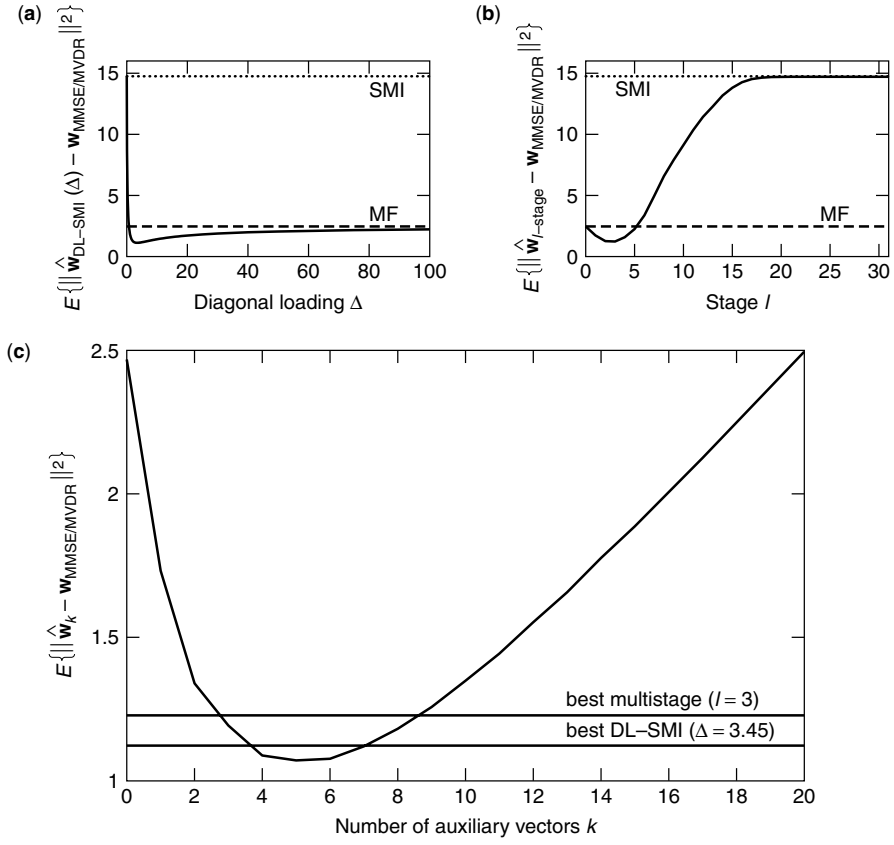


Figure 10. MS estimation error studies for (**a**) diagonally loaded SMI, (**b**) multistage (also called "nested Wiener"), and (**c**) auxiliary vector estimators (the signal model is the same as in Fig. 6 and $J = 60$).

(packet size) range $J = P/2 = 16$ to $J = 3P = 96$. The total computational complexity of the multistage algorithm is of order $O((J + l)P^2)$. The AV algorithm has computational complexity $O((J + k)P^2)$, where k is the desired number of auxiliary vectors $\mathbf{g}_1, \ldots, \mathbf{g}_k$. All intermediate AV filters are returned. The computational complexity of DL-SMI is of order $O((J + P)P^2)$. Estimators of practical interest have $l \ll J$ or $k \ll J$. Therefore, the complexity of all algorithms is dominated by $O(JP^2)$, which is required for the computation of $\hat{\mathbf{R}}(J)$ (the computational complexity of the RLS estimator is, similarly, of the order $O(JP^2)$, the complexity of the GSC estimator is $O(JP^2 + P^3)$, while the complexity of LMS estimator — with no data recycling — is of order $O(JP)$). In terms of performance, the *explicit* or *implicit* matrix inversion that the SMI and the multistage algorithm (at any given stage) performs, respectively, affects adversely their behavior under short-data-record adaptation.

4.1.5. Performance Illustrations. We illustrate the overall short-data-record adaptive filter performance in Figs. 12 and 13 for a multipath fading DSCDMA system that employs antenna array reception. We consider processing gain 31, 20 users, 5 antenna elements, and 3 resolvable multipaths per user with independent zero-mean complex Gaussian fading coefficients of variance 1. The maximum cross-correlation between the assigned user signatures reaches 30%. The total SNR's (over the three paths) of the 19 interferers are set at $\mathrm{SNR}_{2-6} =$ 6 dB, $\mathrm{SNR}_{7-8} = 7$ dB, $\mathrm{SNR}_{9-13} = 8$ dB, $\mathrm{SNR}_{14-15} = 9$ dB, $\mathrm{SNR}_{16-20} = 10$ dB. The spacetime product (filter length) is

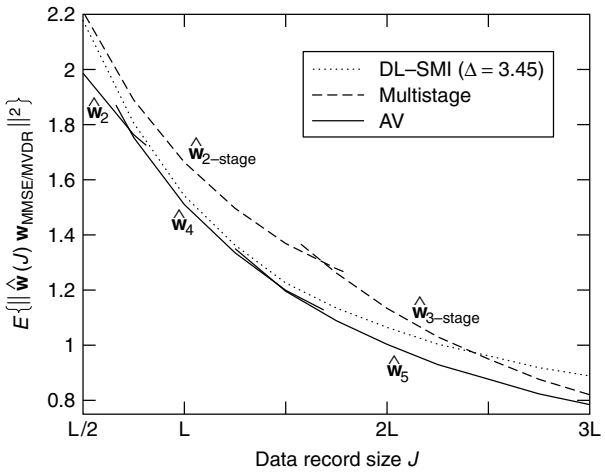


Figure 11. MS estimation error for the *best* multistage and AV estimators over the data support range $J = P/2 = 16$ to $J = 3P = 96$. The MS estimation error of the $\Delta = 3.45$ DL-SMI estimator is also included as a reference (the signal model is the same as in Fig. 6).

$P = (31 + 2)5 = 165$. The experimental results are averages over 1000 runs (100 different channel realizations and 10 independent data record generations per channel). In Fig. 12, we plot the BER[2] of the AV estimators

[2] The BER of each filter under consideration is approximated by $Q(\sqrt{\mathrm{SINR}_{\mathrm{out}}})$, since the computational complexity of the BER

$\hat{\mathbf{w}}_{k_1}(J)$ and $\hat{\mathbf{w}}_{k_2}(J)$ as a function of the SNR of the user of interest for data records of size $J = 230$. We also plot the BER curve of the "genie" assisted BER-optimum filter $\hat{\mathbf{w}}_{k_{\text{opt}}}(J)$ as well as the corresponding curves of the ideal MMSE/MVDR filter $\mathbf{w}_{\text{MMSE/MVDR}}$, the SMI filter estimator $\hat{\mathbf{w}}_{\infty}(J)$, the S-T RAKE matched-filter (MF) $\hat{\mathbf{w}}_0(J)$, and the multistage filter (with the preferred number of stages[3] $l = 7$). We observe that both $\hat{\mathbf{w}}_{k_1}(J)$ and $\hat{\mathbf{w}}_{k_2}(J)$ are very close to the "genie" BER-optimum AV filter estimator choice and outperform significantly the SMI filter estimator, the multistage filter estimator, and the matched filter. We also observe that for moderate to high SNR of the user of interest, the $\mathcal{J}$-divergence selection rule is slightly superior to the CVMOV selection rule. Figure 13 repeats the study of Fig. 12 as a function of the data record size. The SNR of the user of interest is fixed at 8 dB.

Concluding our discussion in this section (4.1), we note that the key for a successful solution (in the sense of superior filter output SINR or BER performance) to the problem of adaptive receiver design under limited data support is to employ receivers with varying bias/variance characteristics and to effectively control these characteristics in a data-driven manner. For this reason, operations and filter design/optimization criteria that suffer from "data starvation" (e.g., implicit or explicit matrix inversion and/or eigendecomposition) should be avoided. From a general input space synthesis/decomposition point of view, the *nonorthogonal* synthesis and the conditional statistical optimization are two principles that allow to grow an *infinite* sequence of AV filters on a "best effort" basis that takes into account the whole interference space at each step. These two features play a key role, leading to superior adaptive filtering performance in short-data-record environments. In particular, under

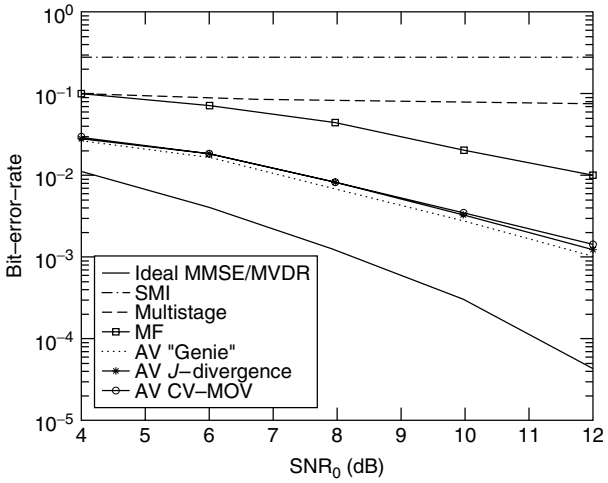


Figure 12. BER versus SNR for the user signal of interest for a multipath fading antenna array received signal model ($L = 31$, $K = 20, M = 5, N = 3$) with $P = 165$ and $J = 230$.

expression for this antenna-array CDMA system prohibits exact analytic evaluation [57].

[3] Honig and Xiao [41] argued that $l = 7$ ($D = 8$ in their notation [41]) stages is "nearly optimal over a wide range of loads and SNRs."

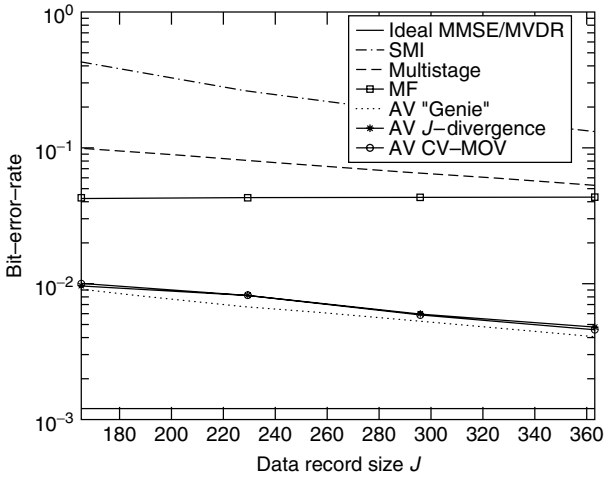


Figure 13. BER versus data record size (the signal model is the same as in Fig. 12, and $\text{SNR}_0 = 8$ dB).

short-data-record adaptation, the early, non-asymptotic elements of the sequence of AV estimators are mildly biased but exhibit much lower variance than other alternatives (for digital communications applications, the latter implies superior BER performance). As the available data record increases we can afford to go higher and higher in the sequence of generated estimators. In the limit, if we are given infinitely many input data, we can go all the way up to the convergence point of the algorithm, which is the ideal MMSE/MVDR receiver. The significant role of conditional statistical optimization in short-data-record adaptive filtering is evident even when *orthogonal* vectors are utilized. For example, it has been seen that the nonconditional (vector-optimum) scheme that utilizes orthogonal vectors with vector-optimum weights (12)–(14) (or, equivalently, the filter obtained by the algorithm of Goldstein et al. [40] that performs implicitly a matrix inversion) exhibits inferior short-data-record performance than does the structure that utilizes orthogonal vectors and conditionally optimized weights [5]. As a few concluding notes, an online version of the AV algorithm has been presented [58,59]. Application of AV filtering to the problem of rapid synchronization and combined demodulation of DSCDMA signals has been considered [60–62]. Detailed results on data record size requirements to achieve a given output SINR (or BER) performance level can be found in earlier studies [63,64].

4.2. Unknown Channel

The second part of this section is devoted to the estimation of the channel-processed constraint vector $\mathbf{v}_0$ from the same data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$. To be consistent with our discussion and illustrative studies presented in the previous sections, we consider the general case of ST DSCDMA signal model described by (1)–(5). We recall that Q, L, N, M, and J denote the number of active users in the system, the processing gain, the number of paths experienced by the transmitted signal of each user, the number of antenna elements, and the data packet (data record) size, respectively. We also recall that $\mathbf{v}_0$ is the ST

RAKE filter of the user of interest (user 0), defined by $\mathbf{v}_0 \triangleq E_{b_0}\{\mathbf{x}b_0\}$, where the statistical expectation operation $E_{b_0}\{\cdot\}$ is taken with respect to the bit of the user of interest b_0 only. Clearly, $\mathbf{v}_0$ consists of shifted versions of the ST matched filter multiplied by the corresponding channel coefficients [cf. (4)]:[4]

$$\mathbf{v}_0 = \sum_{n=0}^{N-1} c_{0,n} \begin{bmatrix} \underbrace{0 \ldots 0}_{n} & \mathbf{d}_0^T & \underbrace{0 \ldots 0}_{N-n-1} \end{bmatrix}^T \odot \mathbf{a}_{0,n} \qquad (27)$$

Hence, $\mathbf{v}_0$ is a function of the binary signature vector (spreading sequence) of the user of interest $\mathbf{d}_0$, the channel coefficients $c_{0,0}, c_{0,1}, \ldots, c_{0,N-1}$, and the corresponding angles of arrival $\theta_{0,0}, \theta_{0,1}, \ldots, \theta_{0,N-1}$ [cf. (5)]. While the spreading sequence is assumed to be known to the receiver, the channel coefficients and the angles of arrival are in general unknown.

4.2.1. Subspace Channel and Angle-of-Arrival Estimation. In this section we explain how the channel coefficients $\mathbf{c}_0 \triangleq [c_{0,0}, c_{0,1}, \ldots, c_{0,N-1}]^T$ and the angles of arrival $\boldsymbol{\theta}_0 \triangleq [\theta_{0,0}, \theta_{0,1}, \ldots, \theta_{0,N-1}]^T$ for the user of interest, user 0, can be estimated by subspace-based techniques from the ST data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$ [65,66]. We note that while adaptive subspace (eigendecomposition)-type MMSE/MVDR filtering is not a favorable approach under limited data support (the resulting estimates exhibit high variance), subspace-type channel estimation techniques do not suffer from "data starvation" as illustrated later.

Let the binary data of each user be organized in identically structured packets of J bits. The channel estimation procedure that we employ utilizes J_p pilot bits (bits that are known to the receiver). Thus, the qth user data packet, $\{b_q(0), b_q(1), \ldots, b_q(J-1)\}, q = 0, 1, \ldots, Q-1$, contains $J - J_p$ information bits and J_p pilot bits. The J_p known bits will be utilized later for the supervised recovery of the phase of the subspace channel estimates since blind second-order channel estimation methods return phase-ambiguous estimates. An example of the data packet structure is shown in Fig. 14, where the J_p pilot bits appear as a *midamble* in the transmitted packet [67]. Without loss of generality, we assume that each user transmits one data packet per slot and the slot duration is T_s seconds. Therefore, the data packet size J is the number of information bits transmitted by each user in one time slot, $T_s = JT$, where T is the duration of each information bit transmission.

The rank r_s of the *signal subspace* of the received data vectors $\mathbf{x}$ can be controlled by considering one-sided or two-sided truncations of $\mathbf{x}$ (the latter eliminates ISI). The possible values of r_s, depending on the data format of choice, are as follows:

1. *No truncation*: data dimension $= M(L+N-1)$, $2Q+1 \le r_s \le 3Q$.
2. *One-sided truncation*: data dimension $= ML$, $2Q \le r_s \le 3Q-1$.
3. *Two-sided truncation*: data dimension $= M(L-N+1)$, $Q \le r_s \le 2Q-1$.

To have a guaranteed minimum rank of the *noise subspace* of $M(L-N+1)-(2Q-1)$, we choose to truncate $\mathbf{x}$ from both sides (case (3)) as shown in Fig. 15, and we form the "truncated" received vector $\mathbf{x}^{\text{tr}}$ of length $M(L-N+1)$ as follows:

$$\mathbf{x}^{\text{tr}} = \begin{bmatrix} \mathbf{x}((N-1)T_c) \\ \mathbf{x}(NT_c) \\ \vdots \\ \mathbf{x}((L-1)T_c) \end{bmatrix}$$

Then, with respect to the jth information bit of user 0, $\mathbf{x}_j^{\text{tr}}$ can be expressed as

$$\mathbf{x}_j^{\text{tr}} = b_0(j)\frac{\sqrt{E_0}}{L}\mathbf{A}_0\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{c}_0 + \text{MAI}_j + \mathbf{n}_j \qquad (29)$$

where MAI_j accounts comprehensively for multiple-access interference of rank $r_s - 1$, $\mathbf{B}(\boldsymbol{\theta}_0)$ is a block diagonal matrix of the form $\mathbf{B}(\boldsymbol{\theta}_0) \triangleq \text{diag}(\mathbf{a}_{0,0}, \mathbf{a}_{0,1}, \ldots, \mathbf{a}_{0,N-1})$, and $\mathbf{A}_0 = \mathbf{A}_0^s \odot \mathbf{I}_M$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix, and

$$\mathbf{A}_0^s = \begin{bmatrix} d_0[N-1] & d_0[N-2] & \ldots & d_0[0] \\ d_0[N] & d_0[N-1] & \ldots & d_0[1] \\ \vdots & \vdots & & \vdots \\ d_0[L-1] & d_0[L-2] & \ldots & d_0[L-N] \end{bmatrix} \qquad (30)$$

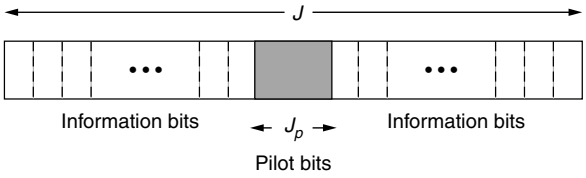


Figure 14. Data packet structure of total length J bits that contains a midamble of J_p pilot bits.

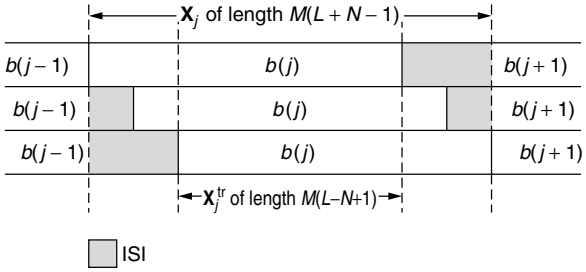


Figure 15. Data collection and ISI trimming.

[4] For the sake of mathematical accuracy

$$\mathbf{v}_0 = \sqrt{\frac{E_0}{L}} \sum_{n=0}^{N-1} c_{0,n} \begin{bmatrix} \underbrace{0 \ldots 0}_{n} & \mathbf{d}_0^T & \underbrace{0 \ldots 0}_{N-n-1} \end{bmatrix}^T \odot \mathbf{a}_{0,n}$$

The positive multiplier $\sqrt{\frac{E_0}{L}}$ is dropped in (27) as inconsequential.

Let $\mathbf{R}_{\text{tr}} = E\{\mathbf{x}^{\text{tr}}\mathbf{x}^{\text{tr}H}\}$ be the autocorrelation matrix of $\mathbf{x}^{\text{tr}}$. We form a sample-average estimate

$$\hat{\mathbf{R}}_{\text{tr}} = \frac{1}{J}\sum_{j=0}^{J-1}\mathbf{x}_j^{\text{tr}}\mathbf{x}_j^{\text{tr}H} \tag{31}$$

based on the truncated J available input vectors $\mathbf{x}_j^{\text{tr}}, j = 0, 1, \ldots, J-1$. If $\hat{\mathbf{R}}_{\text{tr}} = \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^H$ represents the eigendecomposition of $\hat{\mathbf{R}}_{\text{tr}}$, where the columns of $\hat{\mathbf{Q}}$ are the eigenvectors of $\hat{\mathbf{R}}_{\text{tr}}$ and $\hat{\mathbf{\Lambda}}$ is a diagonal matrix consisting of the eigenvalues of $\hat{\mathbf{R}}_{\text{tr}}$, then we use the eigenvectors that correspond to the $M(L-N+1)-(2Q-1)$ smallest eigenvalues to define our estimated noise subspace. Let the matrix $\hat{\mathbf{U}}_n$ of size $[M(L-N+1)] \times [M(L-N+1)-(2Q-1)]$ consist of these "noise eigenvectors." We estimate $\mathbf{c}_0$ and $\boldsymbol{\theta}_0$ indirectly through an estimate of the $MN \times 1$ vector

$$\mathbf{h}_0 \triangleq \mathbf{B}(\boldsymbol{\theta}_0)\mathbf{c}_0 \tag{32}$$

We estimate $\mathbf{h}_0$ as the vector that minimizes the norm of the projection of the signal of the user of interest, user 0, $\mathbf{A}_0\mathbf{h}_0$, onto the estimated noise subspace $\hat{\mathbf{U}}_n$:

$$\hat{\mathbf{h}}_0 = \arg\min_{\mathbf{h}_0} \|(\mathbf{A}_0\mathbf{h}_0)^H\hat{\mathbf{U}}_n\| \quad \text{subject to} \quad \|\hat{\mathbf{h}}_0\| = 1 \tag{33}$$

The solution to this constrained minimization problem is the eigenvector that corresponds to the minimum eigenvalue of $\mathbf{A}_0^H\hat{\mathbf{U}}_n\hat{\mathbf{U}}_n^H\mathbf{A}_0$. After obtaining $\hat{\mathbf{h}}_0$, we may extract the desired vectors $\hat{\mathbf{c}}_0$ and $\hat{\boldsymbol{\theta}}_0$ by applying least-squares (LS) fitting to $\hat{\mathbf{h}}_0$. Then, the estimate $\hat{\mathbf{v}}_0$ is completely defined by (27).

Since the channel estimation method described above is based on a blind second-order criterion, the phase information is absorbed, which means that the estimate $\hat{\mathbf{v}}_0$ is phase-ambiguous. Inherently, adaptive filter estimators that utilize a phase-ambiguous estimate of $\mathbf{v}_0$ are also phase-ambiguous. Next, we consider the recovery (correction) of the phase of linear filters when the vector $\mathbf{v}_0$ is known within a phase ambiguity.

4.2.2. Phase Recovery. Without loss of generality, let $\tilde{\mathbf{v}}_0$ denote a phase ambiguous version of $\mathbf{v}_0$

$$\tilde{\mathbf{v}}_0 e^{j\psi} = \mathbf{v}_0 \tag{34}$$

where ψ is the unknown phase. We consider the class of linear filters $\mathbf{w} \in \mathbb{C}^{M(L+N-1)}$ that are functions of the ST RAKE vector $\mathbf{v}_0$ and share the following property:

$$\mathbf{w}(\mathbf{v}_0) = \mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi} \tag{35}$$

Such filters include (1) the ST RAKE filter itself, $\mathbf{v}_0$, (2) the ST MMSE/MVDR filter of (6), and (3) the auxiliary vector sequence of ST filters $\{\mathbf{w}_k\}$.

As seen in (35), for this class of filters the phase ambiguity of $\tilde{\mathbf{v}}_0$ leads to a phase ambiguous linear filter $\mathbf{w}(\tilde{\mathbf{v}}_0)$. Phase ambiguity in digital communications can be catastrophic since it may result in receivers that exhibit BER equal to 50%. Given $\tilde{\mathbf{v}}_0$, we attempt to correct the phase of $\mathbf{w}(\tilde{\mathbf{v}}_0)$ as follows. As a selection criterion for the phase correction parameter ψ we consider the minimization of the mean-square error (MSE) between

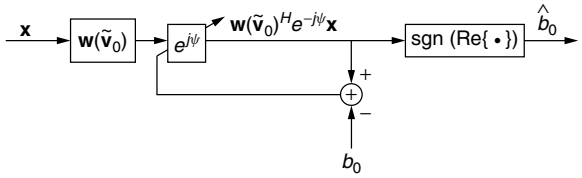


Figure 16. Supervised (pilot-assisted) phase correction for the spacetime linear filter $\mathbf{w}(\tilde{\mathbf{v}}_0)$.

the output of the phase corrected filter $[\mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi}]^H\mathbf{x}$ and the desired information bit b_0 (Fig. 16):

$$\hat{\psi} = \arg\min_{\psi} E\{|[\mathbf{w}(\tilde{\mathbf{v}}_0)e^{j\psi}]^H\mathbf{x} - b_0|^2\}, \quad \psi \in [-\pi, \pi) \tag{36}$$

The optimum phase correction value according to this criterion is given by

$$\hat{\psi} = \text{angle}\{\mathbf{w}(\tilde{\mathbf{v}}_0)^H E\{\mathbf{x}b_0\}\} \tag{37}$$

Essentially, (37) suggests to project the phase ambiguous $\mathbf{w}(\tilde{\mathbf{v}}_0)$ filter onto the ideal ST RAKE filter $\mathbf{v}_0 = E\{\mathbf{x}b_0\}$. However, $E\{\mathbf{x}b_0\}$ is certainly not known. Since we have assumed that a pilot information bit sequence of length J_p is included in each packet, the expectation $E\{\mathbf{x}b_0\}$ can be sample-average-estimated by $\frac{1}{J_p}\sum_{j=1}^{J_p}\mathbf{x}_j b_0(j)$, where $b_0(j), j = 1, 2, \ldots, J_p$, is the jth *pilot* information bit and $\mathbf{x}_j$ is the corresponding input data vector. Then, the phase-corrected adaptive filter estimate is given by

$$\mathbf{w}(\hat{\mathbf{v}}_0, \hat{\mathbf{R}})e^{j\hat{\psi}}, \quad \hat{\psi} = \text{angle}\left\{\mathbf{w}(\hat{\mathbf{v}}_0, \hat{\mathbf{R}})^H\left[\sum_{j=1}^{J_p}\mathbf{x}_j b_0(j)\right]\right\} \tag{38}$$

Since J represents the packet size of the DSCDMA system and J_p is the number of midamble pilot information bits per packet, then the ratio J_p/J quantifies the wasted bandwidth due to the use of the pilot bit sequence. Ideally, J_p/J is to be kept small. As we will see in the next section, a few pilot bits (on the order of 5 bits) are sufficient for effective recovery of the filter phase. As a numerical example, when the packet size is set at $J = 256$ and $J_p = 5$ is chosen, then $J_p/J \simeq 2\%$ only.

5. PACKET ERROR RATE, CAPACITY, AND THROUGHPUT ANALYSIS

5.1. Packet Error Rate

So far, we have concentrated on the design/estimation of the receiver filter $\mathbf{w}$ in (3). The filter estimate is generated adaptively on an individual packet-by-packet basis. All J received vectors of the packet are utilized for the design of $\mathbf{w}$ (estimation of $\mathbf{R}$, $\mathbf{v}_0$, and the number of auxiliary vectors k_1 or k_2), while J_p of them are also used for supervised phase correction (estimation of ψ). Then, the $J - J_p$ information bits of user 0 associated with the remaining $J - J_p$ received vectors are detected by (3).

The effectiveness of the filter is characterized statistically by the probability distribution of the number of bit errors in a packet:

$$p(i) \triangleq \Pr\{i \text{ bit errors in the packet}\}, i = 0, 1, \ldots, J - J_p. \tag{39}$$

Without loss of generality, we define the bit error rate (BER) as the probability of erroneous detection of $b_0(0)$ (the first bit of user 0 in the packet):

$$\begin{aligned}\text{BER} &\triangleq \Pr\{\hat{b}_0(0) \neq b_0(0)\} \\ &= \sum_{i=0}^{J-J_p} \Pr\{\hat{b}_0(0) \neq b_0(0) \mid i \text{ bit errors in the packet}\} p(i) \\ &= \sum_{i=0}^{J-J_p} \frac{i}{J-J_p} p(i) = \frac{1}{J-J_p} \sum_{i=0}^{J-J_p} i p(i)\end{aligned} \tag{40}$$

It is interesting to note at this point that if the filter $\mathbf{w}$ were independent of the information bit stream $b_0(0), b_0(1), \ldots$, then the BER could have been expressed analytically as a weighted sum of the value of the error function $Q(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_x^\infty \mathrm{e}^{t^2/2}\, dt$ evaluated at 2^{K-1} different points (interfering bit combinations) weighted by the probability of each point $2^{-(K-1)}$. However, this independence assumption does not hold true in our system since the data packet (data record) $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{J-1}$, which includes the information bits $b_0(0), b_0(1), \ldots, b_0(J-1)$ to be detected, is directly utilized for the calculation of $\mathbf{w}$. Therefore, performance cannot be evaluated using the analytic BER expression; instead, we can only rely on (40).

A data packet is received successfully within a single time slot if the number of errors in the detection of the $J - J_p$ information bits is less than or equal to the maximum number of (correctable) bit errors allowed by the forward error correction (FEC) module. If no FEC is present, then the packet is successfully received when all $J - J_p$ information bits are detected correctly. The packet error rate (PER) is defined as the probability of receiving an uncorrectable packet within a single time slot and is given by

$$\begin{aligned}\text{PER}(h) &\triangleq \Pr\{\text{more than } h \text{ bit errors in the packet}\} \\ &= \sum_{i=h+1}^{J-J_p} p(i) = 1 - \sum_{i=0}^{h} p(i)\end{aligned} \tag{41}$$

where h is the maximum number of correctable bit errors per packet. By setting $h = 0$ we obtain the PER of a system without FEC:

$$\text{PER}(0) = \sum_{i=1}^{J-J_p} p(i) = 1 - p(0) \tag{42}$$

To examine the PER performance of the general DSCDMA system described in (1)–(5) equipped with the packet-rate adaptive ST AV filter receiver, we proceed with an illustration. We consider a Q-user system with Gold signatures of length $L = 31$. We fix the packet size at $J = 256$ bits and use $J_p = 5$ of them as pilot midamble bits. Each user signal experiences $N = 3$ independent Rayleigh fading paths with equal average received energy per path and independent angles of arrival uniformly distributed in $\left(-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right)$. We consider averages over 20,000 independently drawn multipath Rayleigh fading ST channels. The receiver antenna array consists of $M = 4$ elements. With these numbers, the multipath extended ST product (or, equivalently, the length of the adaptive filter) is $M(L + N - 1) = 132$. The total received predetection SNRs of each user, namely, the sum of the received SNRs over all paths defined as $2E_q \displaystyle\sum_{n=0}^{N-1} E\{|c_{q,n}|^2\}/N_0$, $q = 0, 1, \ldots Q-1$, is set to 11 dB (we recall that $\dfrac{N_0}{2}$ is the AWGN power spectral density assumed to be identical for every spatial channel/antenna element).

In Fig. 17a, we plot the PER as a function of the number of active users Q using (41) for various receivers: (1) ST RAKE, (2) SMI, (3) auxiliary vector, (4) and the orthogonal multistage decomposition filter (also known as "nested Wiener filter") with the preferred number of stages $l = 7$. We also add to this study the multistage filter with $l = 1$ stages, which we found empirically to be the best number of stages for this specific problem. No FEC is assumed ($h = 0$). In a rather more interesting study, Fig. 17b shows the PER under $h = 4$ FEC.

5.2. Capacity

System BER/PER performance improvements due to the use of the AV receiver allow us to accommodate more users for a set quality-of-service (QoS) PER constraint (or reduce the transmitting power of the handset or increase the range/coverage of the base-station transceiver for a preset maximum number of active users). The QoS constraint is based strictly on the specific user application requirements and determines the *user capacity* of the system. Let us express the PER as a function of both the FEC capabilities h and the number of active users Q, $\text{PER}(h, Q)$. Then, we define the user capacity $C(h)$ as the maximum number of users under which the QoS constraint is not violated:

$$C(h) \triangleq \max\{Q: \text{PER}(h, Q) \leq \lambda_{\text{QoS}}\} \tag{43}$$

where λ_{QoS} is the QoS constraint threshold.

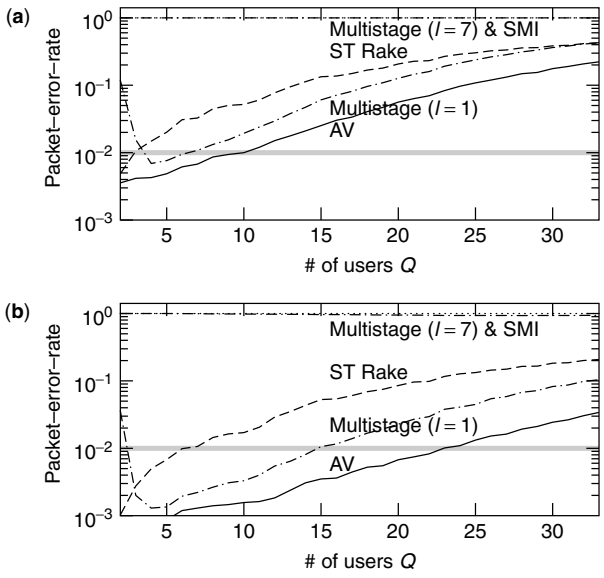


Figure 17. PER versus number of users Q for a system with (**a**) no FEC and (**b**) 4-bit FEC.

We return to Fig. 17 to examine the user capacity of the system for the adaptive receivers under consideration. We set the QoS constraint threshold to $\lambda_{\text{QoS}} = 10^{-2}$. From Fig. 17a we conclude that in the absence of FEC the user capacity of the AV system is 10 and the capacity of the RAKE system is 2. Significant capacity improvement is achieved with 4-bit FEC (Fig. 17b). The AV system supports 23 users, while the RAKE scheme supports only 6 users. Neither the $l = 7$ multistage nor the SMI receiver can meet the QoS constraint for any number of users $Q \geq 1$, with or without FEC. The best multistage ($l = 1$) receiver can support 4–6 users when no FEC is considered and 3–14 users with 4-bit FEC. However, it cannot meet the QoS requirement when $Q \leq 2$. We conclude that we have two viable solutions: the AV and plain ST RAKE receiver systems. The AV system allows up to $\frac{23}{31} \simeq 74\%$ loading for this Gold-coded system with processing gain $L = 31$ and 4-bit FEC (with $M = 4$ antenna elements and multipath fading reception with 11 dB total predetection SNR per user).

5.3. Throughput

If a packet is received with only correctable errors, a positive acknowledgment (ACK) is sent back to the user over a different downlink channel (FDD). Once an uncorrectable error is detected in the packet, a negative acknowledgment (NAK) is sent back to the mobile which then retransmits after waiting for a random number of time slots. If packet arrival is modeled as Poisson-distributed, then the probability of the arrival of Q new messages during a time-slot interval T_s is given by

$$A(Q) = \frac{{G_N}^Q e^{-G_N}}{Q!}, \qquad Q \geq 1 \tag{44}$$

where $G_N \triangleq (\lambda T_s / L)$ is the normalized offered traffic load (we recall that L is the system processing gain) and λ is the packet arrival rate. *Packet throughput* is defined as a measure of the ratio of the average number of successful transmissions to the number of transmission attempts made in a particular time slot. Let $\text{PSR}(h, Q)$ be the packet success rate (that is the probability that a packet is received successfully within a single time slot) with h-bit FEC in the presence of Q users. Then, $\text{PSR}(h, Q) = 1 - \text{PER}(h, Q)$ and the *normalized* packet throughput of the system S_N under slotted ALOHA accessing can be expressed as

$$\begin{aligned} S_N(h) &= \frac{1}{L} \sum_{Q=1}^{C(h)} Q A(Q) \text{PSR}(h, Q) \\ &= \frac{1}{L} \sum_{Q=1}^{C(h)} Q A(Q)(1 - \text{PER}(h, Q)) \end{aligned} \tag{45}$$

In Fig. 18a we plot the normalized packet throughput S_N versus the offered traffic G_N without FEC ($h = 0$). As a reference, we add the familiar packet throughput

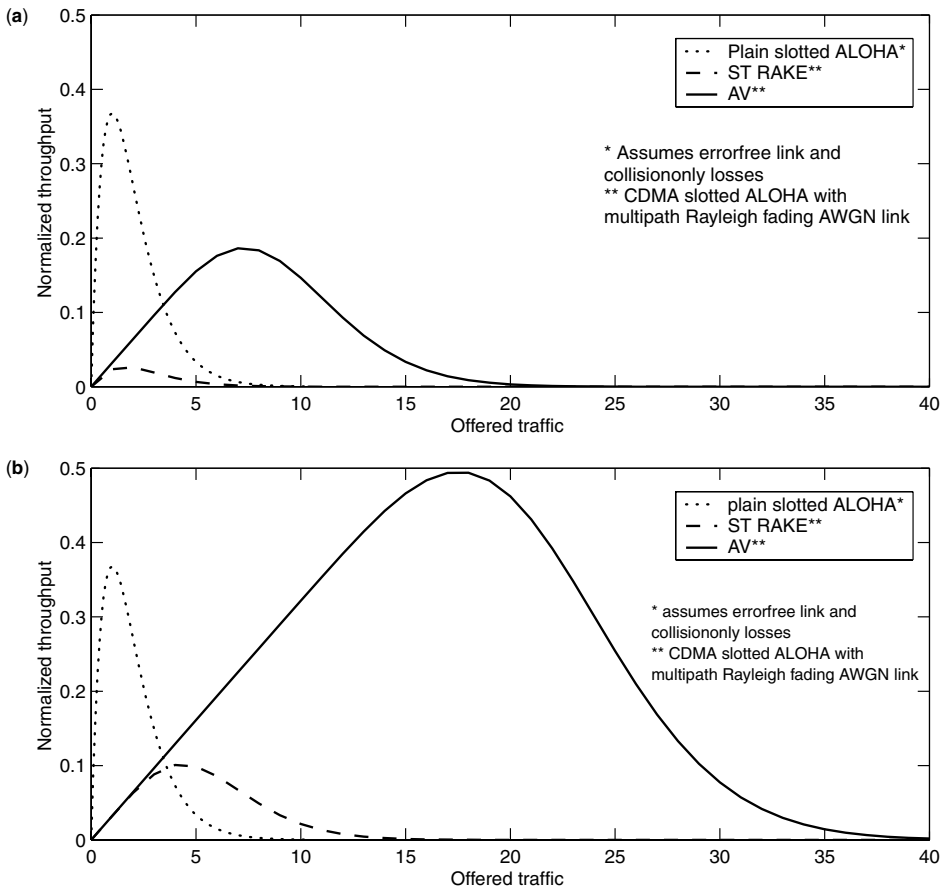


Figure 18. Normalized throughput versus offered traffic for a system with (**a**) no FEC and (**b**) 4-bit FEC.

curve for a plain slotted ALOHA system that assumes an *ideal error-free* link and *collision-only* packet losses. As one might expect, the CDMA slotted ALOHA (RAKE or AV) system with a multipath fading AWGN link and no FEC exhibits a lower maximum achievable throughput value than the plain ALOHA system with an error-free link. However, enhancing the system with FEC (Fig. 18b) gives rise to a significant improvement of the AV system whose throughput performance at higher traffic loads now overshoots the ideal-link plain slotted ALOHA which allows only one user per slot for successful transmission. We observe that with the adaptive AV filter receiver we can achieve up to 0.4938 throughput when we have FEC capability of $h = 4$ bits.

6. CONCLUDING REMARKS

Wireless cellular and personal communication service (PCS) networks have experienced significant growth driven by a strong market interest for highly mobile, widely accessible, two-way voice and data communications. Current research efforts are focusing on system improvements to meet future demand and quality of service requirements. User capacity increase may be sought in the form of a synergy of effective multiple accessing schemes and advanced receiver technology (e.g., code-division multiple access with adaptive antenna arrays). Manageable complexity (hardware and software) at the physical layer can be achieved by means of linear equalizers as opposed to more complex structures. Improved receiver output SINR and BER performance may be sought in the form of intelligent modulation techniques as well as intelligent signal processing at the receiver end of the communications link. However, realistically, receiver output SINR and BER improvements in rapidly changing channel environments can be achieved only by means of adaptive short-data-record-optimized receiver designs (as opposed to designs based on ideal asymptotic optimization solutions).

This article focused on linear MMSE/MVDR adaptive filtering with applications to adaptive receiver designs for mobile communications. We presented three alternative methods that approximate the optimum solution under perfectly known input statistics (input autocorrelation matrix and input/desired-output cross-correlation vector): (1) The generalized sidelobe canceler, (2) the auxiliary vector filters, and (3) the multistage filters. When the input statistics are unknown and estimated, these three approximate solutions provide estimates of the optimum solution with varying performance levels (output SINR and BER). When estimation is based on a short data record, that is, when system adaptation and redesign has to be performed with limited data support (which is the case for most systems of practical interest), then the performance differences become even more pronounced. In mobile packet data communications, for example, the size of the data record that is available for receiver adaptation and redesign is limited by the coherence time of the communication link and may be of the order of 300 data symbols or less in practical situations. In this context, for a given system transmission bit rate, the packet size may be designed to be sufficiently small to conform with the coherence time of the link. Then, packet-rate adaptive receiver designs may be pursued.

A viable solution for adaptive MMSE/MVDR system designs under limited data support is provided by the auxiliary vector (AV) algorithm. AV estimators exhibit varying bias/covariance characteristics — the bias of the generated estimator sequence decreases rapidly to zero while the estimator covariance trace rises slowly from zero (for the initial, fixed-valued, matched-filter estimator) to the asymptotic covariance trace of the SMI filter. Sequences of practical estimators that offer such control over favorable bias/covariance balance points are always a prime objective in the estimation theory literature. Indeed, under quasistatic fading over the duration of a packet and packet-rate adaptation, members of the generated sequence of AV estimators outperform in MS estimation error LMS/RLS-type, SMI and diagonally loaded SMI, and orthogonal multistage decomposition filter estimators. In addition, the troublesome, data-dependent tuning of the real-valued LMS learning gain parameter, the RLS initialization parameter or the SMI diagonal loading parameter, is replaced by an integer choice among the first several members of the estimator sequence. In that respect, we presented two data-driven criteria for the selection of the best AV filter estimator in the sequence.

As a representative application throughout this article we considered a wireless multiuser multipath fading AWGN link with direct-sequence spread-spectrum signaling and slotted ALOHA accessing. We developed a complete adaptive antenna-array CDMA linear filter receiver design that adapts itself and detects the transmitted information bits on an individual packet-by-packet basis. The receiver incorporated seamlessly packet-rate blind subspace-based spacetime channel estimation and supervised recovery of the spacetime channel phase through the use of a few packet midamble pilot bits. Illustrative examples showed that very limited midamble pilot signaling (on the order of $\frac{5}{256} \simeq 2\%$) can be sufficient for phase recovery and effective adaptive receiver design. Therefore, differential modulation to overcome the phase ambiguity problem is not absolutely necessary.

Bit error rate, packet error rate, and user capacity studies and comparisons were also included. Through the development of the probability mass distribution of the bit errors in a packet, we can translate these findings to packet throughput results. Interestingly, the adaptive AV receiver designed for a CDMA system in multipath Rayleigh fading and AWGN that assumed a modest 11 dB total received predetection SNR per user, four antenna elements, and 4-bit FEC, offers a $\frac{0.4938 - 0.3679}{0.3679} 100\% \simeq 34\%$ improvement in terms of normalized maximum packet throughput over plain (non-CDMA) slotted ALOHA with an *error-free* link. In this context, the packet-rate adaptive receiver design using the AV filtering principles coupled with FEC techniques seems to provide a viable solution in improving the performance of DS-CDMA mobile communication links.

BIOGRAPHY

Stella N. Batalama received the Diploma degree in computer engineering and science from the University of Patras, Greece in 1989 and the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, in 1994.

From 1989 to 1990 she was with the Computer Technology Institute, Patras, Greece. From 1990 to 1994 she was a Research Assistant in the Communication Systems Laboratory, Department of Electrical Engineering, University of Virginia. In 1995 she joined the Department of Electrical Engineering, State University of New York at Buffalo, where she is presently an Associate Professor. During the summers of 1997–2002 she was Visiting Faculty in the U.S. Air Force Research Laboratory, Rome, New York. Her research interests include small sample support adaptive filtering and receiver design, adaptive multiuser detection, robust spread-spectrum communications, supervised and unsupervised optimization, and distributed detection.

Dr. Batalama is currently an associate editor for the *IEEE Transactions on Communications* and the *IEEE Communications Letters*.

BIBLIOGRAPHY

1. T. K. Liu and J. A. Silvester, Joint admission congestion control for wireless CDMA systems supporting integrated services, *IEEE J. Select. Areas Commun.* **16**: 845–857 (Aug. 1998).
2. H. Bischl and E. Lutz, Packet error rate in the non-interleaved Rayleigh channel, *IEEE Trans. Commun.* **43**: 1375–1382 (April 1995).
3. R. D. J. van Nee, R. N. van Wolfswinkel, and R. Prasad, Slotted ALOHA and code-division multiple-access techniques for land-mobile satellite personal communications, *IEEE J. Select. Areas Commun.* **13**: 382–388 (Feb. 1995).
4. X. Wu and A. Haimovich, Space-time processing for CDMA communications, *Proc. Conf. Information Science and Systems*, Baltimore, March 1995, pp. 371–376.
5. D. A. Pados and S. N. Batalama, Joint space-time auxiliary-vector filtering for DS/CDMA systems with antenna arrays, *IEEE Trans. Commun.* **47**: 1406–1415 (Sept. 1999).
6. I. S. Reed, J. D. Mallet, and L. E. Brennan, Rapid convergence rate in adaptive arrays, *IEEE Trans. Aerospace Electron. Syst.* **10**: 853–863 (Nov. 1974).
7. B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, Adaptive antenna systems, *Proc. IEEE* **55**: 2143–2158 (Dec. 1967).
8. R. L. Plackett, Some theorems in least squares, *Biometrika* **37**: 149 (1950).
9. R. A. Wiggins and E. A. Robinson, Recursive solution to the multichannel filtering problem, *J. Geophys. Res.* **70**: 1885–1891 (1965).
10. J. S. Thompson, P. M. Grant, and B. Mulgrew, Smart antenna arrays for CDMA systems, *IEEE Personal Commun.* 16–25 (Oct. 1996).
11. L. E. Brennan and I. S. Reed, Theory of adaptive radar, *IEEE Trans. Aerospace Electron. Syst.* **9**: 237–252 (March 1973).
12. E. J. Kelly, An adaptive detection algorithm, *IEEE Trans. Aerospace Electron. Syst.* **22**: 115–127 (March 1986).
13. L. C. Godara, Applications of antenna arrays to mobile communications, Part I: Performance improvement, feasibility, and system considerations, *IEEE Proc.* **85**: 1031–1060 (July 1997).
14. J. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
15. E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, UMTS/IMT-2000 based on wideband CDMA, *IEEE Commun. Mag.* **36**: 70–80 (Sept. 1998).
16. S. Haykin, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1991.
17. V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
18. J. Capon, High-resolution frequency-wavenumber spectrum analysis, *Proc. IEEE* **57**: 1408–1418 (Aug. 1969).
19. N. L. Owsley, A recent trend in adaptive spatial processing for sensor arrays: Constraint adaptation, J. W. R. Griffiths et al., eds., *Signal Processing*, Academic Press, New York, 1973, pp. 591–604.
20. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1990.
21. S. P. Applebaum and D. J. Chapman, Adaptive arrays with main beam constraints, *IEEE Trans. Antennas Propag.* **24**: 650–662 (Sept. 1976).
22. P. W. Howells, Explorations in fixed and adaptive resolution at GE and SURC, *IEEE Trans. Antennas Propag.* **24**: 575–584 (Sept. 1976).
23. B. D. Van Veen and R. A. Roberts, Partially adaptive beamformer design via output power minimization, *IEEE Trans. Acoust. Speech, Signal Process.* **35**: 1524–1532 (Nov. 1987).
24. L. J. Griffiths and C. W. Jim, An alternative approach to linearly constrained adaptive beamforming, *IEEE Trans. Antennas Propag.* **30**: 27–34 (Jan. 1982).
25. P. Strobach, Low-rank adaptive filters, *IEEE Trans. Signal Process.* **44**(12): 2932–2947 (Dec. 1996).
26. P. A. Thompson, An adaptive spectral analysis technique for unbiased frequency estimation in the presence of white noise, *Proc. 13th Asilomar Conf. Circ. Systems Computers*, Nov. 1980, pp. 529–533.
27. N. L. Owsley, in S. Haykin, ed., *Array Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.
28. B. D. Van Veen, Eigenstructure based partially adaptive array design, *IEEE Trans. Antennas Propag.* **36**: 357–362 (March 1988).
29. A. M. Haimovich and Y. Bar-Ness, An eigenanalysis interference canceler, *IEEE Trans. Signal Process.* **39**: 76–84 (Jan. 1991).
30. K. A. Byerly and R. A. Roberts, Output power based partially adaptive array design, *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, 1989, pp. 576–580.
31. J. S. Goldstein and I. S. Reed, Reduced-rank adaptive filtering, *IEEE Trans. Signal Process.* **45**: 492–496 (Feb. 1997).
32. S. N. Batalama, M. J. Medley, and D. A. Pados, Robust adaptive recovery of spread-spectrum signals with short data records, *IEEE Trans. Commun.* **48**: 1725–1731 (Oct. 2000).

33. D. A. Pados and G. N. Karystinos, An iterative algorithm for the computation of the MVDR filter, *IEEE Trans. Signal Process.* **49**: 290–300 (Feb. 2001).

34. D. A. Pados and G. N. Karystinos, Short-data-record estimators of the MMSE/MVDR filter, *Proc. ICASSP 2000*, Istanbul, Turkey, June 2000, Vol. 1, pp. 384–387.

35. D. A. Pados and S. N. Batalama, Low-complexity blind detection of DS/CDMA signals: Auxiliary-vector receivers, *IEEE Trans. Commun.* **45**: 586–1594 (Dec. 1997).

36. A. Kansal, S. N. Batalama and D. A. Pados, Adaptive maximum SINR rake filtering for DS-CDMA multipath fading channels, *IEEE J. Select. Areas Commun.* 1965–1973 (Dec. 1998).

37. D. A. Pados and S. N. Batalama, Joint space-time auxiliary-vector filtering for antenna array DS/CDMA systems, *Proc. 1998 Conf. Information Science and Systems*, Princeton, NJ, March 1998, Vol. 2, pp. 1007–1013.

38. D. A. Pados, T. Tsao, J. H. Michels, and M. C. Wicks, Joint domain space-time adaptive processing with small training data sets, *Proc. IEEE Radar Conf.*, Dallas, TX, May 1998, pp. 99–104.

39. J. S. Goldstein, I. S. Reed, P. A. Zulch, and W. L. Melvin, A multistage STAP CFAR detection technique, *Proc. IEEE Radar Conf.*, Dallas, TX, May 1998, pp. 111–116.

40. J. S. Goldstein, I. S. Reed, and L. L. Scharf, A multistage representation of the Wiener filter based on orthogonal projections, *IEEE Trans. Inform. Theory* **44**: 2943–2959 (Nov. 1998).

41. M. L. Honig and W. Xiao, Performance of reduced-rank linear interference suppression, *IEEE Trans. Inform. Theory* **47**: 1928–1946 (July 2001).

42. T.-C. Liu and B. Van Veen, A modular structure for implementation of linearly constrained minimum variance beamformers, *IEEE Trans. Signal Process.* **39**: 2343–2346 (Oct. 1991).

43. N. E. Nahi, *Estimation Theory and Applications*, R. E. Krieger, Huntington, NY, 1976.

44. C. D. Richmond, Derived PDF of maximum likelihood signal estimator which employs an estimated noise covariance, *IEEE Trans. Signal Process.* **44**: 305–315 (Feb. 1996).

45. R. L. Dykstra, Establishing the positive definiteness of the sample covariance matrix, *Ann. Math. Stat.* **41**(6): 2153–2154 (1970).

46. C. D. Richmond, PDF's, confidence regions, and relevant statistics for a class of sample covariance-based array processors, *IEEE Trans. Signal Process.* **44**: 1779–1793 (July 1996).

47. A. O. Steinhardt, The PDF of adaptive beamforming weights, *IEEE Trans. Signal Process.* **39**: 1232–1235 (May 1991).

48. B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, Adaptive antenna systems, *Proc. IEEE* **55**: 2143–2158 (Dec. 1967).

49. L. C. Godara and A. Cantoni, Analysis of constrained LMS algorithm with application to adaptive beamforming using perturbation sequences, *IEEE Trans. Antennas Propag.* **34**: 368–379 (March 1986).

50. V. Solo, The limiting behavior of LMS, *IEEE Trans. Acoust. Speech, Signal Process.* **37**: 1909–1922 (Dec. 1989).

51. J. M. Cioffi and T. Kailath, Fast recursive-least-squares transversal filters for adaptive filtering, *IEEE Trans. Acoust. Speech, Signal Process.* **32**: 304–337 (April 1984).

52. H. Qian and S. N. Batalama, Data-record-based criteria for the selection of an auxiliary-vector estimator of the MVDR filter, *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Oct. 2000, pp. 802–807.

53. H. Qian and S. N. Batalama, Data-record-based criteria for the selection of an auxiliary-vector estimator of the MMSE/MVDR filter, *IEEE Trans. Commun.* (in press).

54. C. R. Rao, *Handbook of Statistics 9*. New York, NY: Elsevier, 1993.

55. D. Kazakos and P. Papantoni-Kazakos, *Detection and Estimation*, Computer Science Press, New York, 1990.

56. B. D. Carlson, Covariance matrix estimation errors and diagonal loading in adaptive arrays, *IEEE Trans. Aerospace and Electron. Syst.* **24**: 397–401 (July 1988).

57. H. V. Poor and S. Verdú, Probability of error in MMSE multiuser detection, *IEEE Trans. Inform. Theory* **43**: 858–871 (May 1997).

58. I. N. Psaromiligkos and S. N. Batalama, Interference-plus-noise covariance matrix estimation for adaptive space-time processing of DS/CDMA signals, *Proc. IEEE VTC 2000—Vehicular Technology Conf.*, Boston, Sept. 2000, Vol. 5, pp. 2197–2204.

59. I. N. Psaromiligkos and S. N. Batalama, Recursive AV and MVDR filter estimation for maximum SINR adaptive space-time processing, *IEEE Trans. Commun.* (in press).

60. I. N. Psaromiligkos and S. N. Batalama, Blind self-synchronized demodulation of DS-CDMA communications, *Proc. IEEE ICC 2000—Int. Conf. Communications*, New Orleans, LA, June 2000, pp. 2557–2560.

61. I. N. Psaromiligkos, M. J. Medley, and S. N. Batalama, Rapid synchronization and combined demodulation for DS/CDMA communications. Part I: Algorithmic developments, *IEEE Trans. Commun.* (in press).

62. I. N. Psaromiligkos and S. N. Batalama, Rapid synchronization and combined demodulation for DS/CDMA communications. Part II: Finite data-record-size performance analysis, *IEEE Trans. Commun.* (in press).

63. S. N. Batalama and I. N. Psaromiligkos, Data record size requirements of MVDR-optimized adaptive antenna arrays, *Proc. IEEE ICASSP 2000—Int. Conf. Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000, Vol. V, pp. 3069–3072.

64. I. N. Psaromiligkos and S. N. Batalama, Data record size requirements for adaptive space-time DS/CDMA signal detection and direction-of-arrival estimation, *IEEE Trans. Commun.* (in press).

65. S. Gopalan, G. N. Karystinos, and D. A. Pados, Capacity, throughput, and delay of slotted ALOHA DS-CDMA links with adaptive space-time auxiliary-vector receivers, *IEEE Trans. Wireless Commun.* (in press).

66. S. E. Bensley and B. Aazhang, Subspace-based channel estimation for code division multiple access communication systems, *IEEE Trans. Commun.* **44**: 1009–1020 (Aug. 1996).

67. P. Chaudhury, W. Mohr, and S. Onoe, The 3GPP proposal for IMT-2000, *IEEE Commun. Mag.* **37**: 72–81 (Dec. 1999).

PACKET-SWITCHED NETWORKS

DIMITRIOS STILIADIS
Bell Laboratories
Lucent Technologies
Holmdel, New Jersey

1. INTRODUCTION

Packet-switched networks are becoming the dominant method of communication, replacing earlier schemes based on the telephone-type circuit-switched networks. With the increasing popularity of the World Wide Web, electronic mail (email), and multimedia applications, the traditional role of a data network as a means of transmitting data between computers is expanding. The same integrated network is now used by applications such as teleconferencing, distance education, real-time video and voice, email, Facsimile (fax), and distributed systems. At the same time, link speeds are experiencing dramatic increases, and the number of users is growing exponentially.

Telephone networks are based on the idea of *circuit switching* (Fig. 1). The dominant application supported by the circuit-switching architecture is voice, and it is assumed that the bandwidth required between two users is determined by the bandwidth needed for transmitting good-quality analog voice, or 8 kHz. The network establishes a dedicated bidirectional path between end nodes (i.e., telephones), and while the call is active, end users can continuously use this path to transmit information. The network is responsible for allocating enough resources throughout the communication path so that data can be transmitted as a continuous flow. Once the resources are allocated, they cannot be reused by another user, until the call is complete. This allocation is usually done by using either time-division multiplexing (TDM) or frequency-division multiplexing (FDM).

One of the most important properties of circuit-switched networks is that the sum of the capacities required by all the active communication paths cannot exceed the capacity of the link. The network must perform *admission control* and cannot accept new connections once the sum of the capacities needed by the active connections reaches a maximum threshold. When the network cannot admit any more connections, new requests by end users receive a "busy signal" similar to the one encountered in phone networks (or the *call is blocked*). Because of the requirement for admission control, establishing a connection and allocating resources (i.e., initiating a new phone call) is a relatively expensive operation that is tackled by a set of distributed computers.

The ideas of voice switching can be extended to the context of data communications by assuming that the two end nodes are computers exchanging data. The network offers a dedicated communication path between the two computers, and once a path is established, it may remain active for a long time period. However, nodes do not constantly transmit data during this period, but they may remain idle for some long intervals of time. Consider, for example, the case where one of the nodes is a desktop computer and the other is a file server. The computer will issue a request to the file server for some data, and once the data are received, the computer will begin to process them. When processing is complete, it will write the data back to the file server. While the computer is processing the data, no information is exchanged on the communication path.

Several data applications may exhibit a similar *bursty* behavior, where they use the bandwidth for some intervals of time and remain idle for others. File access and the World Wide Web (WWW) are prominent examples of such applications. Furthermore, the notion of burstiness can be extended to real-time applications such as voice or video. During a voice call, communicating parties do not always talk, but might have long periods of silence, during which a voice signal corresponding to silence does not need to be transmitted over the network.

If multiple users reserve bandwidth that is needed only during short intervals of time, valuable resources are wasted. In the example of Fig. 1, let us assume that only one session can be active at any time at the central link. Assume that user A places a phone call to user B, and they start talking. Although all the bandwidth of the critical link is reserved for this communication, user A only talks half of the time, and thus the link is 50% utilized. Now let us assume that user C tries to place a phone call to user D. The network cannot accept this call, since the bandwidth

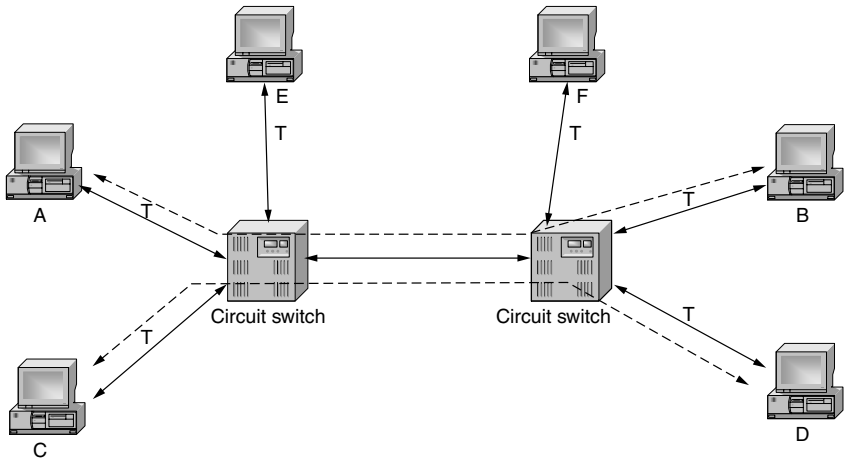


Figure 1. Circuit-switched network.

of the critical link is already allocated to the other phone call. Therefore, when an dedicated path is set up between bursty applications, the bandwidth might not be properly utilized.

Now assume that we could have a network where the path is set up when the link is actually needed and the resources are released when no information is transmitted. In this case, user A would use the critical link while talking to user B. The rest of the time the critical link would be free and user C could also transmit information. Both calls would take place at the same time, and resources would be shared in a more efficient manner.

Unfortunately, this is not easy to achieve in a circuit-switched network, since establishing a communication path requires the configuration of several nodes in the network. However, it is exactly this observation that led to the idea of packet-switched networks, also referred to as *store-and-forward networks*.

2. PACKET-SWITCHED NETWORKS

Communication between two users can be thought of as an exchange of *messages*. For example, when two people talk to each other, they form sentences, and each sentence can be considered as a message. When an end user browses the Web, he/she reads several pages of information, and each page can be considered as a message. A message can have variable sizes and the size of any given message is not bounded. For this reason, the concept of a *packet* is introduced. A packet can be part of a message or it can encapsulate several messages. However, depending on the technology, the packet size is either constant or bounded. For example, packets in a local network based on Ethernet technology are no more than 1500 bytes [1].

When two users communicate, they form a series of messages (and thus packets) and transmit them to one another. Referring to the network of Fig. 2, when user A talks, a packet is created and it is transmitted over the access link. The packet arrives at the intermediate node (referred to as *packet switch* or *router*), and assuming that the link is free, it is transmitted to user B. At a later time a packet from node C is transmitted to the intermediate node and through the critical link to user D. Information from the two users is *multiplexed* on the link on a packet-by-packet basis.

A problem arises when packets from both users arrive at a packet switch at the same time, but only one of the packets can be transmitted at the core link. The packet switch must store in some local memory the packet received from one of the two users (let us say user A) and transmit the packet from the other user. When the transmission is complete, it will retrieve the packet that originated from user A and transmit this packet over the link. It is exactly this concept, that is the foundation of *packet-switched* or *store-and-forward* networks.

Although packet networks may operate efficiently in most cases, *congestion* may degrade their performance during some time periods. Congestion occurs when during a period of time, the bandwidth needed for transmitting all arriving packets exceeds the capacity of the outgoing link. When this happens, the memory in the packet switches may overflow and messages might get lost. In order for packet switched networks to avoid or prevent congestion, they must support a range of mechanisms that control the end user behavior.

3. STATISTICAL MULTIPLEXING

Let us consider again the network of Fig. 2 and assume that the capacity of all the links between users and network nodes is equal to 2 packets per second, and thus, the time to transmit a packet is equal to half a second. Let us also assume that users transmit information half of the time. For example, user A transmits 10 packets and then waits for 5 s before transmitting more packets. If both users transmit their packets at exactly the same time, some of these packets must be buffered at the packet switch.

Let us now assume that the capacity of the core link is equal to the sum of the capacities of the input links, or 4 packets per second. Note, that this is also equal to the capacity needed from a circuit-switched network if both calls are active at the same time. In this scenario, both packets can be transmitted on the link without any delay (Fig. 3). A packet is transmitted over this link within 0.25 s and the link remains idle for half of the time. The maximum queueing delay of any packet is bounded by 0.25 s.

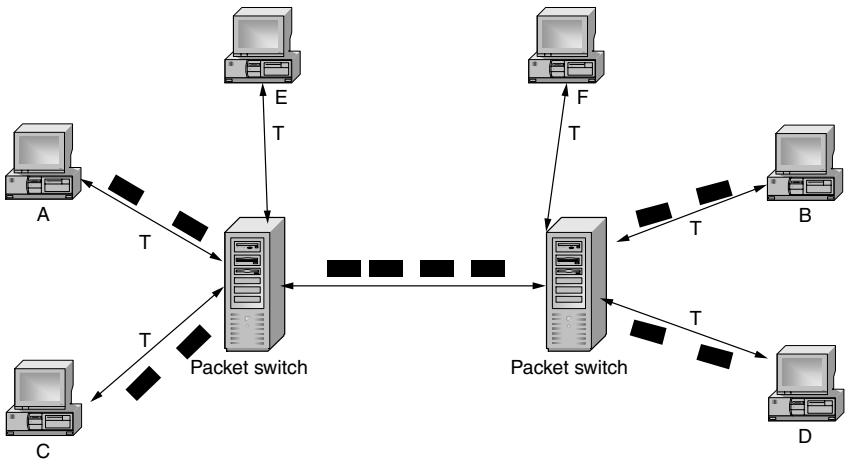


Figure 2. Packet-switched network; information transmitted in terms of packets.

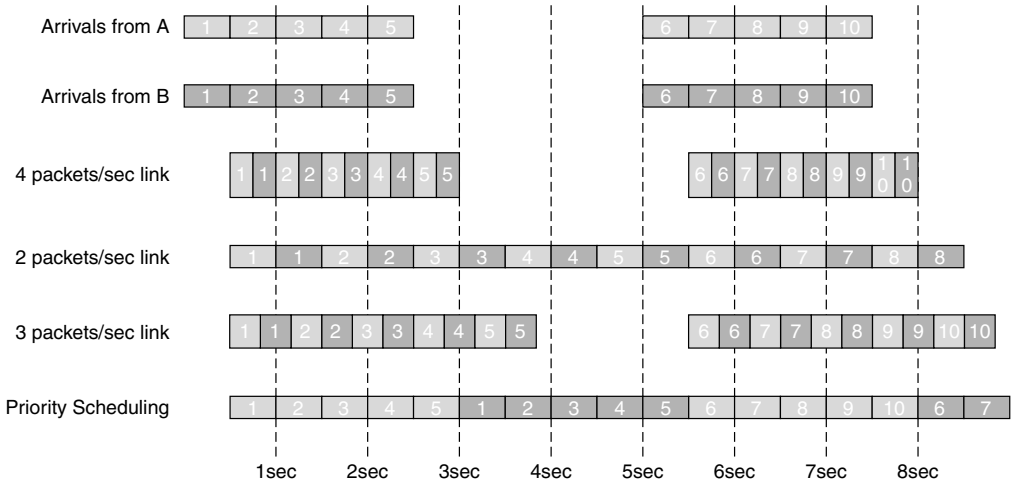


Figure 3. Sequence of packet arrivals and transmissions by a packet switch for different link capacities and scheduling disciplines.

If we set the capacity of the core link to 2 packets per second, some packets are delayed by as much as 2 s (or 4 packet transmission times) and the packet switch must maintain a buffer of at least 4 packets. However, the utilization of the core link is 100%. If we set the capacity of the core link equal to 3 packets per second, no packet will see a delay longer than 0.33 s and the utilization of the core link is 66%.

We can observe in the example above that by modifying the capacity of the core link, we can trade off link utilization for delays and buffers in the core switches. This is exactly the concept of statistical multiplexing. The ratio derived as the sum of the incoming link capacities divided by the outgoing link capacity is referred to the *statistical multiplexing gain* [2]. For example, when we set the core link capacity equal to the access link capacity, the statistical multiplexing gain is $\frac{4}{2}$ or 2.

The concept of statistical multiplexing can be viewed from a different perspective. If we assume that the buffers and link capacities at the core node are fixed (and thus, the maximum delay we can afford is fixed), we need to determine whether traffic from a particular source will not encounter excessive losses. This leads to the *effective bandwidth theory* [3]. If some statistical model for the arrival traffic is available, we can determine the loss probability of different sources.

3.1. Timescales

The concept of statistical multiplexing is not unique to data networks, but it was actually developed within the context of circuit switching. The main difference between the two approaches is in terms of times of interest or, *timescales*. In circuit switching it is assumed that end nodes use the network mostly for phone-calls. The main concept behind engineering voice networks is that not all users will initiate a call at the same time. Thus, a large number of users can share the same resources on a *call-by-call* basis. Statistical models are used to describe how often users actually initiate a call (*call arrival rate*) and how long is the call duration (*call holding time*). Based on these parameters, one can estimate the link capacities required to reduce the probability that the network will reject a call because of lack of resources. This was the original concept of statistical multiplexing of resources among a large number of users.

When we move to packet-switched networks, multiplexing is done on a *packet-by-packet* basis as opposed to a call-by-call basis. The lengths of times (timescales) of interest are much shorter. The reason for our interest in these shorter timescales is derived from the nature of data applications. Users would like to have a constant high-speed connection to the network and transmit very fast only for short intervals of time. Consider the operation of the Web (WWW). Users access Webpages, process the received data (i.e., read the context), and make decisions about accessing more data (i.e., follow links). The response time of the network is critical, and the duration of the connection to the Internet might be very long.

During the late 1990s, when many users were connecting to the Internet using modems and their telephone lines, the phone network was constantly overloaded. The traffic models used to engineer the network assumed that the call holding time is around 3 min. These models were failing to capture the actual behavior of end users however, since an increasing number of users kept their phone lines busy for hours when connected to the Internet. Although the network would reject calls, the bandwidth of the lines was not fully utilized.

3.2. Traffic Scheduling

Since packet-switched networks are used for a variety of end-user applications, it is reasonable to expect that different applications might have different requirements from the network. For example, email messages might be stored in the routers for several seconds without any problem. On the other hand, messages that carry voice must be delivered immediately since they are part of an interactive communication. Similarly, some applications might be able to afford information loss, whereas for other applications the network must support mechanisms that will guarantee that all information is delivered without packet losses.

Let us consider the network of Fig. 2, and let us assume that the core link capacity is 2 packets per second. In the previous case (Fig. 3), we assumed that packets are served in an first come–first served order (FCFS or FIFO) (where packets are transmitted in the same order as they were received). If traffic from user A is real-time traffic and requires minimum possible delays, whereas traffic from user C is email traffic, we can modify the way packets are selected for transmission. When packets from both users are buffered in the switch, it will always transmit packets from node A first (*static priority* order). The sequence of departures is also shown in Fig. 3. Real-time packets see no delay, whereas email packets see a maximum delay of 2 s. Thus, the method used for selecting packets for transmission (or *traffic scheduling discipline*) can determine the maximum and average queueing delays of different users. Notice however, that if the scheduler always transmits packets when packets are available in its queues, the average delay of *all* packets does not depend on the traffic scheduling discipline. Interested readers are referred to Zhang [4] for an overview of various traffic scheduling mechanisms.

4. CONNECTION-ORIENTED VERSUS CONNECTIONLESS NETWORKS

There is one main taxonomy of packet-switched networks, which is based on the method used to decide how packets are forwarded (or *routed*), and how resources are allocated [5].

In the previous examples we concentrated on a switch where traffic from all input links is multiplexed on a specific output link. This type of a switch is also known as a *multiplexer*. In the more general case, however, a packet switch might receive traffic from several interfaces and forward the packets to different interfaces. A large network will consist of multiple packet switches as shown in Fig. 4. When a packet arrives in such a switch a decision must be made as to which is the outgoing interface.

There are two main philosophies or methods used to resolve this question:

1. *Connection-Oriented* or *Virtual Circuit Switched Networks*. In these networks a virtual circuit is established between the source and destination nodes before any communication starts. The establishment of the virtual circuit does not necessarily lead to an explicit resource allocation as in the case of circuit-switched networks. However, some state information is associated with each switch that determines how all packets that belong to this connection must be forwarded. When a packet is created, a unique connection identifier is attached to it. All intermediate nodes will use this connection identifier to determine the output interface for this packet. The path that all packets of the connection follow is determined a priori during the connection setup phase. Technologies such as *asynchronous transfer mode* (ATM) [6] and *multiprotocol label switching* (MPLS) [7] are based on this principle.
2. *Connectionless Networks*. In these networks no explicit path is established a priori. Every end node in the network is associated with some address, and every switch or router has a "view" of the topology of the network. In other words, a *forwarding table* is stored in every switch with an entry corresponding to every other node. This table determines on a packet-by-packet basis, the interface that must be

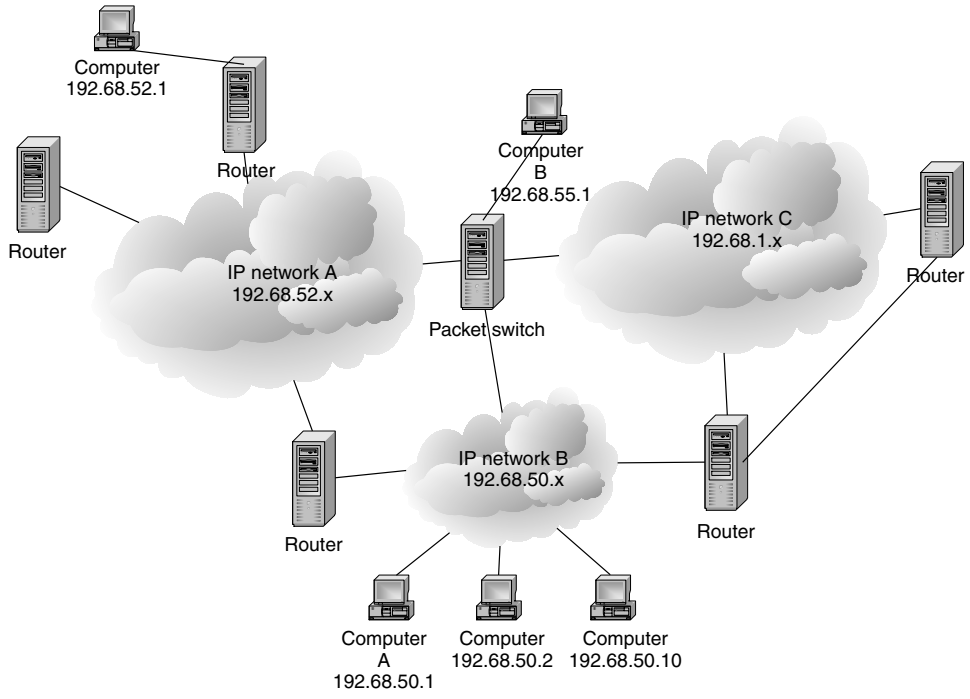


Figure 4. Global Internet architecture as an interconnection of multiple networks.

used in order to reach specific end nodes. Packets carry a *destination address* that identifies the end nodes, and switches use only this information to decide how to forward them. The network supports the mechanisms that allow individual switches to discover the topology of the network and create this forwarding table. Networks based on the *Internet Protocol* use this approach.

A connectionless network operates very much like regular mail and the post office. Senders address letters to specific recipients and the post office uses the address to forward the letters from one city to another and finally to the recipient. The originating post office cares only about the destination town or state. Local post offices make decisions about street addresses and apartment numbers. Similarly, in a packet network, routers in the middle of the network need to know only the router that is closest to the destination user. Only the final (or *edge*) router needs to know about specific users.

An important feature of connectionless networks is that packets between two end nodes can be routed through multiple paths of the network at the same time. For example, subsequent packets from workstation A to workstation B in Fig. 4 might take different paths. It is exactly this feature of connectionless networks, however, that makes them extremely reliable. If one of the nodes fails, packets can be still forwarded to their destination through a different path. This is known as the *shelf-healing* property of IP networks. On the other hand, in a connection-oriented network if a node or a link fails, a completely new path must be established from scratch. Since establishing a path might take a long period of time, traffic may be interrupted during this period and packets may get lost. For this reason connection oriented networks support techniques where multiple paths are reserved a priori.

Note that, depending on the path delays of a connectionless network, it is possible for packets that use different paths arrive to their destination out of order (i.e., if packet p1 is transmitted before packet p2 from node (workstation) A, it is not guaranteed to arrive at node B before packet p2). For this reason, the network must support mechanisms that will reorder packets at the end node.

In addition to these differences, which are related mainly to how the forwarding decisions are made, there is another fundamental distinction between the two philosophies. In connection-oriented networks, that are very similar to traditional phone networks, the end user must notify the network in advance of the bandwidth and the type of service he/she wants to use. The network uses this information to *admit* the user, avoid congestion, and offer different services. In connectionless networks the user transmits the packet to the network and hopes that enough capacity is available for the packet to be delivered to its destination. The network does not promise to any node that packets will be delivered after a predetermined delay, but it does its best (*best-effort networks*). The user is responsible for adapting his/her bandwidth requirements based on feedback received by the network, in order to prevent congestion.

5. PROTOCOLS AND LAYERING

During the design of networks (both packet-switched and circuit-switched), it became apparent that a large number of technologies and architectures must work together. For example, packet networks might work on top of Ethernet, and Ethernet is defined to work either over copper using electrical signals or over fiber using optical signals. The Ethernet network can also interchange information with a wireless network or a network using strictly optical signals [like a SONET (synchronous optical network)].

In order to allow a variety of transmission media to interwork with a variety of networking architectures and physical links, the notion of protocol layering was introduced [8]. First, a *protocol* can be considered as a set of rules, messages, and behaviors, that allows two end nodes to communicate to each other. Networks are built using a layered architecture of protocols.

To understand the concept of layering, let us consider the case of the voice network. End users know only about telephone numbers. When they want to communicate with other users, they dial a number on their phone, which can be considered as a *module*. The telephone will interact with the telephone network, that is another module, to establish a connection and the end users can begin talking. The user does not need to know how the telephone works, and the telephone does not to know how exactly the telephone network establishes a connection. However, the user expects a behavior from the phone, and the phone expects a behavior from the network.

In other words, each module in the network follows a set of rules to communicate with another module and expects a prespecified behavior. In the case of packet-networks, there are different modules in the network architecture that decide how packets are formed and how delivery is guaranteed. These modules expect a specific behavior from modules downstream, and finally from the links that are used to interconnect network nodes.

Figure 5 illustrates the most commonly accepted layer architecture as defined by the reference model of the Open Systems Interconnection (OSI) model developed by the International Standards Organization (ISO). Most networks follow this model. A description of the different layers follows.

5.1. Physical Layer

The physical layer defines the electrical, optical, electromagnetic or other properties of actual physical links. Examples are electrical signal voltages, signal frequencies, coding schemes, clock recovery and distribution schemes, and error correction schemes. Layers above the physical layer do not need to worry about such details, and they expect usually a low loss transmission of information among nodes.

5.2. Data-Link Layer

The data-link layer provides a variety of mechanisms that offer a logical communication path between two nodes. Mechanisms defined are packet formats, error checking, and retransmission mechanisms. A commonly used data-link layer is encountered in Ethernet networks. The basic

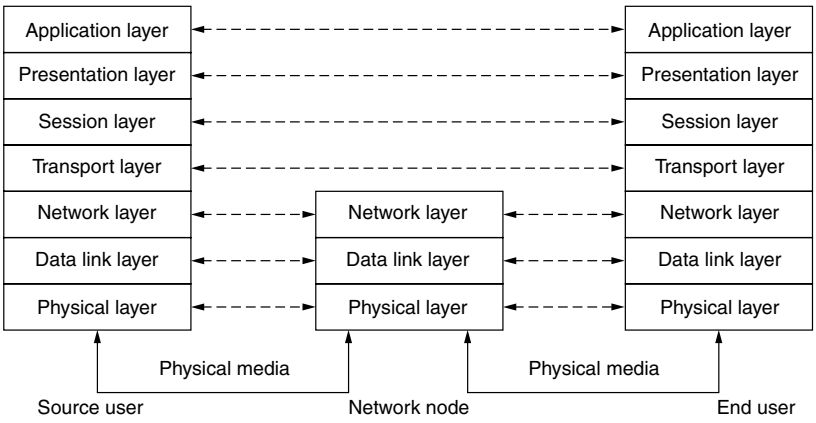


Figure 5. Network protocol layers.

principle of Ethernet (or any *multiple-access network* in general) is that several nodes are connected through the same physical media. Each node broadcasts information on the media, and only the destination node actually uses this information. The data-link layer will designate the source and destination nodes, and it will resolve contentions that appear when multiple nodes try to use the media at exactly the same time. The operation of this layer might involve retransmission of packets if communication fails. The data-link layer is limited to communication between two neighboring network nodes only.

5.3. Network Layer

The main function of the network layer is to determine where and how packets must be forwarded. When a packet arrives to this layer, the network layer uses local information to determine whether this packet is destined for some local process or it must be transmitted to another node. For packets arriving from higher layers, the network layer will format the packet in such a way that it will be recognized by other nodes in the network. The network layer does not care for the method used to transmit a packet to a neighboring node, and it assumes that the data-link and physical layers will provide this communication path. The *Internet Protocol* (IP) is an example of a network-layer protocol.

In some networks, the network layer might also assist with flow control. It might provide mechanisms that will avoid or reduce the probability of congestion. For example, if a downstream node is congested, information might be provided to the upstream nodes to reduce their transmission rate. The nodes might pick a different path to transmit traffic or they might propagate this information all the way back to the end user.

5.4. Transport Layer

The transport layer is responsible for establishing a logical connection between two end nodes. Connections might have properties such as bidirectional or unidirectional, or error-free. The transport layer will receive information from higher layers, packetize it, and pass it to the network layer. The transport layer might also verify that information is delivered correctly to the other application, by adding error checking mechanisms and defining the end-to-end protocol for retransmissions that will fix errors. In connectionless networks the transport layer will also guarantee in-order delivery of packets. The Transport Control Protocol (TCP) and User Datagram Protocol (UDP) encountered in the Internet are examples of transport-layer protocols.

5.5. Session Layer

The session layer is the user–network interface. It translates a user request to a request from the network. Users do not need to know how the network establishes connections and how it guarantees packet delivery. The session layer maps user requirements to network functions. If the application needs an error-free communication, the session layer will interface to a transport protocol like TCP. If the application needs an one-way transmission of information, where error-free delivery is not required, the session layer will interface to a transport protocol like UDP.

The session layer might also assist with resolving names to addresses that are recognized by the network. For example, end users do not really know about numeric addresses used by the transport and network layers (like IP addresses), but mostly remember mnemonics (like World Wide Web addresses).

5.6. Presentation Layer

The presentation layer might apply specific transformations on the data when they are delivered across the network. For example, data might be encrypted to prevent any other users from receiving the information. Sometimes data might be compressed in order to provide a speedier delivery. The presentation layer will also determine issues like ordering of bytes in big-endian or little-endian formats.

5.7. Application Layer

The application layer consists of end-to-end applications that use the network as a means of providing a service. An example of an application layer protocol is HTTP, which is the protocol used between a Web browser and a Web server.

5.8. Assigning Tasks to Layers

Although in the previous sections we gave some definitions of layers, depending on the network technology some of the functions can be moved to different layers. Flow control

is a clear example of such a function. In the case of ATM networks, flow control or congestion control is a part of the network layer definition, and *all* nodes in the network are responsible for assisting in this task. On the other hand, in IP networks, congestion control is a transport-layer function, and intermediate nodes do not interfere. The basic concept of TCP is that end nodes will detect packet losses, and will use these losses as an indication of congestion (i.e., several packets were queued on the same node and some packets had to be dropped). When congestion is detected the rate of data transmission of the end nodes is decreased. One can easily notice that if the data-link and physical layers do not guarantee error-free transmission, when a packet is lost because of link quality, it will be misinterpreted by TCP as an indication of congestion. This can lead to performance problems that have been the focus of several studies [9].

6. INTERNET AND TCP/IP

In this section we briefly discuss the principles of the most popular packet-switched network, the Internet. The basic architecture follows the principles of connectionless networks as described earlier [10] (Fig. 4). Data are forwarded in terms of packets of variable size, and the maximum packet size is 64K bytes (64 kilobytes). A header is associated with every packet that includes among other fields the addresses for the source and destination nodes of the packet. Routers use this header to determine how packets must be forwarded.

One can imagine the Internet as being composed of a large number of independently controlled networks without centralized control. Each network consists of three basic components:

1. A global assignment of addresses to nodes
2. Routers that forward packets to different interfaces and run the protocols that allow them to decide how to forward packets (*routing protocols*)
3. End nodes that use the TCP/IP protocol stack

6.1. Addressing

The Internet architecture provides a global means for assigning addresses to end nodes or routers, and once an address is assigned to a node, it is unique across the Internet. Whenever another user wants to communicate with this node, it must forward packets toward this address. IP version 4[1] addresses are 32 bits or 4 bytes. A decimal notation of four numbers separated by a dot is used to represent these addresses. For example 192.68.52.1 is an IP address that maps to the 32-bit number *C*0443401 (Fig. 6).

In order to minimize administrative overheads a hierarchical method of address assignment, was developed. Several end nodes that can be accessed through similar paths are assigned addresses within a specific range. Take for example, the networks in Fig. 4. Networks A and B

[1] We will refer to IP version 4 address within this text. There are newer versions of the IP protocol (IPv6) that use a slightly different address space.

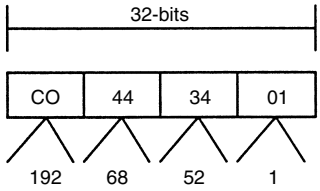


Figure 6. Representation of IP addresses.

are two global networks that can span the same country or continent. There are only two routers connecting networks A and B, and the IP addresses assigned to nodes on network A can be all aggregated together in a single number 192.68.52.x. This means that every node in network A has an address between 192.68.52.0 and 192.68.52.255. Similarly, IP addresses within network B are aggregated as 192.68.50.x. When any node in network B wants to communicate with a node in network A, the packets must go through one of the paths that interconnect networks A and B.

Note, that even though two nodes might be located in the same area they might connect to the Internet using a different network. Their IP address is determined by the network they connect to and not by their physical location. This is contrary to the method used in the telephone network, where address assignment is based on geographic boundaries (i.e., all telephone numbers within the same area have the same area code).

The IP addresses in these examples can be thought of as consisting of two components: (1) the *network number*, which is the 24 most significant bits, and identifies the network where a host resides; and (2) the *host or interface address*, which is the 8 less significant bits, which identify individual hosts within the network. The maximum number of hosts in a given network depends on the number of bits that are used to identify the network.

Originally, the IP address space was partitioned into four *static* classes of addresses (classes A–D) with different number of bits assigned to the network number. For example, a class A address uses 8 bits for the network number and 24 bits for hosts. Only a small number of networks (127) can have a class A address, and they were assigned to large organizations. The problem with this approach is that if an organization does not need all these addresses, a significant portion of the address space is wasted. For example, if a network is assigned a class A address, it can have up to 2^{24} interface addresses. If the network needs only a quarter of them, the rest will remain unused.

As the popularity of the Internet increased, it became apparent that the address space must be utilized properly, and that wasting address space resources can limit the scalability of the network. The new version of IP (version 6) is designed to address this problem, and it introduces 128-bit addresses. In the meantime, however, the concept of classless interdomain routing (CIDR) was introduced to allow a more efficient use of the address space [11]. Each network is now identified by two numbers: (1) an IP address and (2) a mask that determines the number of most significant bits of the address that identify the network. For example, a network might be identified with an address of 192.128.0.0/9, which means that the 9 most

significant bits correspond to the network number, and the 23 less significant bits identify hosts within this network. This method allows a variable number of hosts in any given network, and the address space is utilized more efficiently.

6.2. Routing

The second component is routing. This is a set of protocols run by the IP routers that allow them to construct a local view of the topology of the network. The routers must know which interface to use in order to reach a specific IP address in the network. However, routers do not need to maintain a distinct entry for each individual address. Because of the aggregation method described earlier, routers in network A can maintain a single entry for all nodes in network B, since they are all reachable through the same path [12]. This is actually the main benefit of the hierarchical address assignment. Routers at the core of the network must only maintain state information that determines how different networks are reached. Routers at the edge of the network must maintain state information on individual hosts.

In order for the routers to discover the network topology, they exchange information. Such information might include the list of neighboring routers or information learned from other routers in the network. For example, router R1 knows that one of its interfaces is connected to network A. It will thus notify the routers or end nodes in network B that if they want to reach network A, they must forward packets toward R1.

Once the network topology is discovered, most routing protocols assume that the network is represented by a graph, and routing is the task of finding the shortest path between any two nodes in this graph. Other criteria, such as finding the least-congested path, might be used in order to optimize network performance. Routing protocols are distributed among the various routers, and there is no centralized place that decides how packets must be forwarded. Each router makes independent decisions, and it is crucial that the routing algorithms lead to a stable network configuration and do not create loops. Assume two routers A and B and a destination C. Let us assume that router A views that in order for a packet to reach destination C, it must send the packet to node B. If router B has also an entry that says that if a packet is destined to destination C, it must be forwarded to node A, then any packet destined to node C will be constantly transmitted between nodes A and B.

As we mentioned earlier, the Internet can be considered as an interconnection of various networks. Each of these individually managed networks is called an *autonomous system* (AS). Routers that belong to the same AS use an *interior gateway protocol* (IGP) to exchange information, and forwarding is mostly based on shortest-path criteria. Routing between ASes uses an *exterior gateway protocol* (EGP), that is based mainly on policies determined by contracts between network operators.

6.3. TCP/IP Stack

The third component is that all end nodes must follow the TCP/IP protocol stack. This means that they must encapsulate packets using the IP headers and use the correct destination IP addresses. Most IP traffic uses one of two basic transport protocols. When a reliable end-to-end communication is required without any packet losses, the TCP protocol is used. TCP allows end users to open a logical "connection" to another end user. Once the connection is established, TCP enables the transmission of data between the end users and guarantees reliable delivery. For example, if a packet is lost in the network, TCP will take care of retransmissions.

TCP is also an important part of the congestion control mechanisms of IP networks. The basic principle is that end users must adapt their transmission rate to the bandwidth that is available from the network. Users start with a low transmission rate and increase the rate periodically. When losses are detected, they are interpreted as an indication of congestion and the rate is dropped. This process is constantly repeated, and allows TCP to calibrate the transmission rate to the available resources. One can consider this as a *closed-loop* flow control protocol, since information from the network is used to adjust the rate of the transmitter [13].

When nonreliable communication is sufficient, the User Datagram Protocol (UDP) can be used. UDP establishes an simplex connection between two users and allows delivery of data without flow control or error recovery. UDP is especially useful in real-time streaming applications, like video or voice, where small data losses are acceptable.

6.4. Domain Name Servers

There is an additional layer, handled by the *domain name servers*, that maps mnemonic addresses to IP addresses and can be regarded as a directory service. Most users are aware of World Wide Web (WWW) addresses such as *www.wiley.com.* These addresses must be translated to actual IP addresses before a communication can start. End nodes first send a request to a DNS asking for the IP address of the mnemonic address of the destination. The DNS will translate the WWW address to an IP address. When the address is resolved, the end node can use the IP protocol to communicate with the desirable destination. Note that the DNS address itself is statically configured by the end user.

7. HISTORY OF PACKET-SWITCHED NETWORKS

The first theoretical work in packet switching appeared in L. Kleinrock's Ph.D. thesis in 1962, and it is still considered as forming the theoretical foundation [14]. Around the same time Baran invented the fundamental concepts behind store-and-forward switching [15]. Among others, Baran's work developed the concepts of packets or messages, adaptive routing based on failures, and decoupling between logical and physical addresses. Similar concepts were developed independently in the Cyclades packet-switched network in France [16].

The telecommunications industry did not pay much attention to these concepts until the U.S. Department of Defense Advanced Research Project Agency (DARPA) sponsored a research program for the development of ARPAnet. The ARPAnet was mainly developed in order to provide a reliable and low cost communication network

among timesharing systems scattered throughout the country. This can be considered as the first wide-area packet-switched network, that was demonstrated by 1969. By 1972 the ARPAnet had 4 hosts, expanding to 23 by 1973.

From that point on the expansion of the Internet has been exponential, and it entered our everyday lives with the development of the World Wide Web in the early 1990s. Currently millions of hosts are interconnected in the Internet through a maze of networks without any centralized control, and data traffic is increasing exponentially.

BIOGRAPHY

Dimitrios Stiliadis received his Ph.D and M.S. degrees in computer engineering from the University of California at Santa Cruz, in 1996 and 1994, respectively. Prior to that he received the Diploma in computer engineering from the University of Patras, Greece, in 1992. Since 1996, he has been with the High-Speed Networks Research Department of Bell Laboratories, where he is currently a distinguished member of technical staff. During these years he has been leading the architecture of several generations of packet switching equipment. His recent research has been in issues related to traffic management, switch scheduling, and applications of optical technologies to packet networks. He is a corecipient of the 1998 IEEE Fred W. Ellersik Award.

BIBLIOGRAPHY

1. W. Stallings, *Handbook of Computer-Communications Standards*, Vol. 2, *Local Network Standards*, Macmillan, 1987.
2. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.
3. R. Guerin, H. Ahmadi, and M. Nagshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE J. Select. Areas Commun.* **9**(7): 968–981 (Sept. 1991).
4. H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, *Proc. IEEE* **83**(10): 1374–1396 (Oct. 1995).
5. S. Keshav, *An Engineering Approach to Computer Networking*, Addison-Wesley, 1997.
6. D. E. McDysan and D. L. Spohn, *ATM Theory and Applications*, McGraw-Hill, 1998.
7. B. Davie and Y. Rekhter, *MPLS: Technology and Applications*, Morgan Kaufmann Publishers, 2000.
8. H. Zimmerman, OSI reference model — the ISO model of architecture for open systems interconnection, *IEEE Trans. Commun.* **28**(4): 425–432 (April 1980).
9. H. Balakrishnan, V. N. Padmanabhan, S. Sheshan, and R. Katz, Comparison of mechanisms for improving TCP performance over wireless links, *Proc. ACM SIGCOMM '96*, Sept. 1996.
10. W. R. Stevens, *TCP/IP Illustrated Volume 1, 2, 3*, Addison-Wesley, 2000.
11. V. Fuller et al., *Classless Inter-Domain Routing*, RFC 1519, *ftp://ds.internic.net/rfc/rcf1519.txt*, June 1993.
12. R. Perlman, *Interconnections: Bridges, Routers, Switches, and Internetworking Protocols*, Addison-Wesley, 2000.
13. V. Jacobson, Congestion avoidance and control, *Proc. ACM SIGCOMM 88*, Aug. 1998, pp. 314–329.
14. L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964.
15. P. Baran, On distributed communications networks, *IEEE Trans. Commun. Syst.* (March 1964).
16. B. M. Leiner et al., *A Brief History of the Internet*, Internet Society *http://www.isoc.org/internet/history/brief.shtml*.

PAGING AND REGISTRATION IN MOBILE NETWORKS

CHRISTOPHER ROSE
Rutgers WINLAB
Piscataway, New Jersey

1. INTRODUCTION

A communications network routes messages from senders to recipients. In *fixed* networks where units are not mobile, the terminal (such as a telephone handset, computer, video display, or a host of other possible devices) resides at a fixed physical location which rarely changes. In contrast, a *mobile communication network* routes messages between senders and receivers who may often change location. The seemingly simple addition of terminal mobility to the networking problem complicates it in both obvious and subtle ways.

We will start by exploring conceptually simple methods for accommodating mobility and then examine the deeper implications of these methods on network organization and design. However, rather than plunging directly into what could easily become an opaque technicalese treatise, it is easiest to use an analogy that we will expand as needed. So, consider a postal address, 536 West 145th Street, Apartment 21 in New York City — the author's childhood residence. Any messages for the author would be delivered to this address via a mail carrier or through a specific set of copper wires running from a central office to this address. The "address" of the telephone was (212) AU1-4676, which could be reached from anywhere in the "developed" world at that time. This basic fixed scenario was the dominant communications network structure — static identifiers corresponding to fixed physical locations for the terminal equipment, and tacitly, for the reams of equipment necessary to carry traffic between arbitrary addresses. For the postal service this would include postoffices and the city streets along which mail carriers traveled. For telephone service this would mean central switching offices and the cables that threaded their way under and around the city between central offices and homes.

Of course, people move about in their day to day lives, and the fixed scenario could only awkwardly accommodate mobility. A mail carrier arriving before August 6, 1970 would have been able to properly deliver mail to the author. After that date, without a *forwarding address*, the mail would remain undelivered or marked "Return to sender." Likewise, a telephone call that arrived while the

author was in school across the street at PS 186 would be missed, or the caller would have to be given a new number corresponding to the new physical location. At the time there was no (inexpensive) method to guarantee delivery of messages to a recipient in motion.

This simple analogy illustrates that having fixed terminals associated with fixed physical locations makes routing information through any sort of network a relatively straightforward task in principle. Certainly there are details such as communications link congestion/availability that the service provider must handle, but the basic notion of network topology stasis remains intact.

Now, suppose that units require messages to be delivered wherever they happen to be. The network needs to know exactly where the unit is — or in the parlance of the mobility management field — the unit's *point of network attachment*. There are two ways the network can ascertain the unit's location: (1) the network can search for the unit in likely places (paging) or (2) the unit can tell the network where he/she is (registration) — and we have thus provided an en passant definition of paging and registration as basic building blocks (or *primitives*) for handling mobility in communications networks.

The basic ideas of paging and registration lead to a variety of techniques that we will explore more carefully in Section 2. However, before proceeding we pause here to note that "units" need not be rigidly defined. In the Internet age where computer programs such as *Web crawlers* might autonomously search the Web for information, a *unit* could in principle be a *program*. Conceptually, paging and registration for such programs is not a big leap; however, from the network management standpoint one must consider that such programs could change their physical location *much* more rapidly than any "physically realized" unit such as a person or a piece of hardware and this could lead to unique stresses on the network.

Finally, suppose that in addition to keeping track of mobile units, the actual network topology were labile. By analogy, imagine the perplexity of a mail carrier who found not only that had the author moved from 536 West 145th Street but also that the connecting avenues and streets had changed relative locations as well! The analog to this somewhat nightmarish scenario might be the rule rather than the exception in some types of ad hoc network architectures where the communication infrastructure is *composed* of mobile nodes communicating wirelessly as opposed to the usual sets of fixed location equipment such as cables, microwave towers, and switching centers.

Interestingly (and perhaps obviously), all these scenarios can be handled by suitably abstracted versions of the basic paging and registration paradigm. We therefore carefully describe paging and registration along with their associated costs for conventional mobile networks in following sections and later apply the concepts to the more exotic scenarios which will almost certainly arise in the future.

2. PAGING AND REGISTRATION

The key idea behind paging and registration is that routing messages to a unit is a cooperative affair where the network makes an effort to find the unit through paging and the unit, who needs to be found, registers its location with the system. The optimization problem arises since both paging and registration require some sort of communication, and communication bandwidth is a valuable commodity — thus, there are *costs* associated with paging and registration.

Specifically, it is obvious that a unit could always be immediately located by the system if that unit always registered each change in location, and in response the system updated its global routing databases. However, the network cost of such updates, especially for units whose point of network attachment changed often could be prohibitive [1–6]. Thus, mobile networks trade unit localization delay against the cost of perfect location information available through constant registration. In other words, some uncertainty in the actual unit point of attachment is tolerated and is resolved by seeking the unit through paging — which requires time and some cost in communications bandwidth. In addition, the unit concurrently offers periodic updates on location within cost constraints, or alternatively, the system actively seeks a unit when paging cost exceeds some threshold. Finally, if a unit never receives calls, then there is little need to ever register location since the system will never need to find that unit. All these issues combined form the paging/registration problem.

2.1. Paging in a Typical Cellular System

Consider a cellular telephone system where units can reside in one of N possible locations as in Fig. 1. These locations might be associated with cellular base stations or be *location areas* composed of many cells associated with a mobile telephone switching office (MTSO). If the unit location is uncertain, then the system must find the unit when an incoming call arrives. To do so, in a wireless system, paging messages are sent to possible unit locations. The paging message uniquely specifies the desired mobile terminal and thus requires passage of information, which further implies use of bandwidth — the most precious resource in a wireless system. The system could page a unit in all possible locations simultaneously and could in principle find a unit almost immediately.

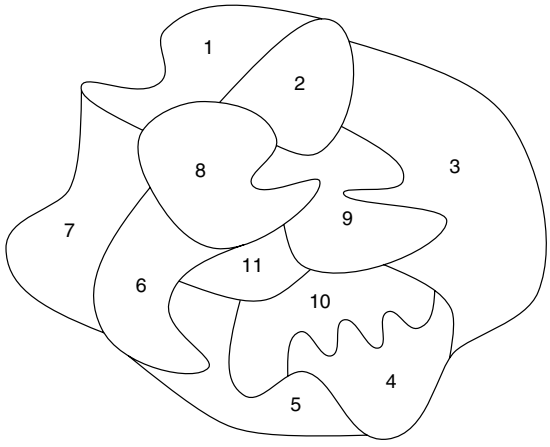


Figure 1. Possible locations associated with a mobile unit.

However, paging a single unit in all locations precludes paging more than one unit simultaneously, and in systems where signaling channels and traffic channels are shared, consumes resources everywhere, which can actually increase the average delay in finding a unit [7,8] and/or degrade quality of service (QoS).

Thus, if some delay can be tolerated in finding a unit, sequential paging of a unit starting with the most likely location will minimize the expected amount of paging bandwidth used. If some delay bound must be met, then groups of locations can be paged simultaneously and in this way various points on a paging delay/cost performance curve can be achieved. Regardless, in all such paging problems, the key result is that locations should be searched in order of the *a priori* probability of unit residence [9].

For such *unit-centric* approaches to mobility management, the unit location probability distribution plays a key role in minimizing the system cost of paging; the problem of assembling and maintaining such information has been carefully examined [10,11]. Specifically, it was shown how the Lempel–Ziv empirical sequence coding method could be applied to constructing and maintaining compact mobility profiles for different users.[1] In addition, the tradeoff between paging group size and paging delay for multiple unit systems has been considered [7,8].

Of course, some practical issues must be mentioned.[2] There is the possibility that a unit, owing to the propagation environment, will not receive a page. Thus, paging messages may be repeated up to some system-specified number of retries before a failure is declared at a given location. This somewhat complicates the analysis, but the basic premise (search most likely first) still holds. It should also be noted that most current systems do not bother to maintain a dynamic register of likely unit locations, even though it could significantly reduce paging channel use [12] since the complexity of implementation is thought to exceed the benefit.

2.2. Registration

Classically, registration strategies have used the previously mentioned concept of a *location area* — groups of (usually) contiguous locations. These areas can be global in the sense that they are identical for all mobile units, or personal in that each unit has its own location area. The two concepts are illustrated in Figs. 2 and 3.

Under the classical scenario, a service area is partitioned into groups of locations. Incoming calls result in page requests at all locations in the appropriate location area. When a mobile unit crosses a group boundary, a registration is mandated. Location area boundaries are chosen based on incoming call rates, aggregate mobility patterns, and the relative cost of paging and registration.

[1] The papers by Bhattacharya and Das [10,11] received top prize at ACM MobiCom'99 and are required reading for anyone who studies mobility management

[2] Details about paging message structure and protocols for current systems can be found in (for example) the paper by Goodman et al. [7].

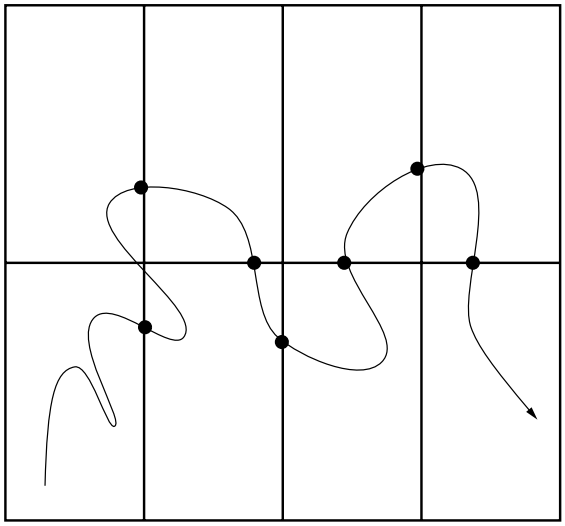


Figure 2. Illustration of classical registration strategy with geographically fixed location areas. Registration occurs at location area boundary crossings (denoted by solid dots). The arrival of an incoming call triggers polling requests over all locations contained in the current location area.

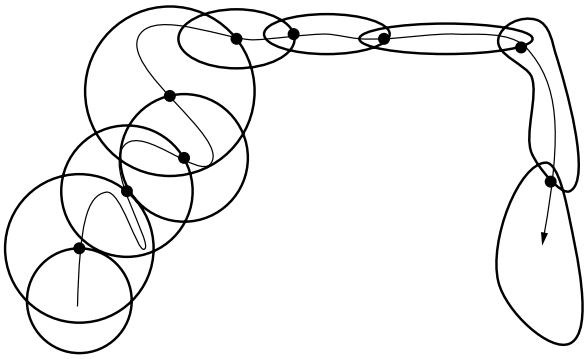


Figure 3. Illustration of personal location areas, recalculated after every contact with system. The location area shape may change with time depending on the mobility characteristics of the unit. For example, the mobile unit in this example might first be driving around a city and then enter a high-speed thoroughfare. The location areas change shape to accommodate the changed mobility. Registration points are denoted by solid dots.

Unfortunately the simple classical strategy suffers from spurious registrations since mobile units that dwell near boundaries can "pingpong" between location areas. Furthermore, since mobile units may have different mobility characteristics, designing location areas for the aggregate can be inefficient.

A key element of the registration problem is the rate at which calls arrive to the mobile unit. If calls never arrive, then the unit need never register since paging cost is always zero. Likewise, if calls arrive regularly then registration might not often be necessary because the network has reasonably accurate location information. Thus, the procedure used depends both upon the mobility process and the incoming call rate as well as the relative cost of registration.

Using these ideas, the concept of personal location areas was proposed [13,14] where each mobile unit contact with

the system results in the calculation of a new location area, centered around the current location. Since the boundaries are reset after each registration, this method does not suffer from spurious location updates, and since location areas can be designed for each unit, the problem of designing for the aggregate does not exist.

However, as with the classical scheme, all locations in a location area are paged in response to an incoming call even if the call arrives shortly after a previous call has terminated. Furthermore, since the boundaries of these location areas are fixed, high-velocity mobile units must register more frequently or require larger location areas, both of which tend to increase cost.

2.3. Paging/Registration Using Optimal Paging and the Mobility Index

Suppose a unit follows a brownian motion pattern[3] in which case the cost of paging is proportional to $(\mathcal{D}t)^{n/2}$, where n is the number of degrees of freedom in unit motion, τ is the time since the unit location was known exactly, and $\mathcal{D}$ is a constant [15]. It is useful to further define a *mobility index* as $\rho = \mathcal{D}/\lambda$, where λ is the incoming call rate. The inverse of ρ is known as the *call to mobility ratio* [16,17]. The paging cost can then be written as $(\rho t)^{n/2}$, where the new variable t, the time since last sighting, is measured in units of average call interarrival times $1/\lambda$. The utility of the mobility index lies in its explicit inclusion of the frequency with which a unit is contacted. For example, if the unit is often paged, then its whereabouts are better known to the system, thereby decreasing the unit mobility index.

These basic notions have been applied to registration strategies based on time and/or place [18]. Other work has also considered versions of this basic problem, for example, the paper by Bar-Noy et al. [19]. It is assumed that the mobile unit knows the paging method employed by the network and thereby knows the cost (or expected cost) to be incurred at the current point in (time, place). Assuming Poisson arrivals of paging requests makes for a relatively straightforward optimization.

When only time information is used by the mobile unit, a deadline registration policy is appropriate. Thus, the mobile unit should register at time τ^* after the time of last contact with the network, where τ^* depends on the mobility index and the relative costs of registration and paging. Intervening call initiations or received calls reset the timer. With both time and place information, however, optimality is an open question.

Application of the timer-only registration method leads to an improvement over simpler place-based methods [13], especially at higher mobile unit velocities since the optimal paging algorithm is affected only by location uncertainty and not directly by the unit velocity. Suboptimal registration methods based on time *and* place have been explored [20,21], and similar to an optimal method based on distance from last sighting derived for the random walk with drift velocity zero [22], threshold rules for registration were obtained. It was found that the (time, place)-based method performed only slightly better than the timer-alone method over a range of conditions. The primary improvement afforded by spatial information was a reduction in cost variance.

[3] Brownian motion is the simplest and most often used mobility model. In certain ways it constitutes a worst-case mobility scenario since for finite location variance (a surrogate for average energy expended in moving an object), the location uncertainty (entropy) is maximum.

3. PAGING, REGISTRATION, AND INFORMATION THEORY ABSTRACTIONS

3.1. Entropy and Paging Cost

Since paging amounts to issuing queries about unit location and these queries require signaling messages, two obvious questions arise:

How much information does the network need to resolve a mobile unit location at a given point in time?

What is the relationship of this information to the amount of signaling required?

The obvious simple answer to the first question is *the entropy of the location distribution*. Likewise, considering that locations vary with time, we see that the average information *rate* necessary to completely specify unit location for all time is the entropy rate of the motion process [23,24]. However, the structure of allowable queries about location profoundly affects the relationship between information content and the necessary signaling in the paging problem.

For example, consider the game of "Guess my number," where a number n is chosen between 1 and J according to some probability distribution p_j [25,26]. Using the standard information-theoretic formulation, the minimum average number of yes/no questions necessary to identify n is the entropy of the probability distribution. However, in a mobile communications system, each polling event at a given location requires signaling in the network, and possibly use of a radio channel as well in the case of a wireless system. Thus, from a signaling standpoint, the appropriate queries are of the form "Is your *number* (*location*) *k*?" This leads to problems with relying solely on entropy as a measure of location uncertainty. Specifically, one can easily describe distributions whose entropies exist and are finite, but for whom the paging cost is infinite [27].

3.2. Beyond Paging and Registration

We have so far considered only the effort associated with finding a given mobile unit through paging and registration; that is, the unit location was not known exactly. Here we take a slightly different view and assume that the location of every unit is known somewhere. Now we ask how much signaling effort is necessary to efficiently disseminate the location information over the network.

Consider then the *universal phone number* (UPN), where a single number is used to identify each mobile unit and route the call appropriately. At the heart of the UPN problem is a question similar to that which arose in the paging/registration context: *Who needs to know?* Thus, the frequency with which a given UPN routing entry is used is the primary index of the importance of maintaining its accuracy.

Let us assume that the answer to this question is given by the rate, r_{ij}, at which unit i calls unit j. This concept of calling rate, coupled to a view of mobility as a set of time-varying location distributions, allows a simple lower bound on the amount of information that must be disseminated over the network.

Specifically, let a random variable τ_{ij} describe a renewal process for the time between calls from unit i to unit j. We assume that τ_{ij} has a density function $f_{\tau_{ij}}(t)$ with mean $1/r_{ij}$. At the initiation of a call at time t, from the perspective of unit i, the location uncertainty of unit j is given by the location probability distribution $p_j(t, t_0, x_j(t_0))$, where $x(t_0)$ was the position of unit j at time t_0.[4] In information-theoretic terms, however, the entropy of this distribution represents the average amount of information required to exactly specify the location of unit j to unit i at time t.

Define $H_{ij}(t)$ with $t \geq t_0$ as the entropy in bits of the location distribution $p_j(t, t_0, x_j(t_0))$. The average number of bits needed by unit i to specify the location of unit j just before the next call is

$$\overline{H}_{ij} = \int_0^\infty f_{\tau_{ij}}(t') H_{ij}(t')\, dt' \tag{1}$$

and by renewal theory [28], the *absolute minimum* average number of bits per second needed by unit i to determine the location of unit j is simply $\overline{H}_{ij}/E[\tau_{ij}] = r_{ij}\overline{H}_{ij}$.

The minimum necessary aggregate network signaling load can then be determined by considering the average number of hops [29]; that is, location information from unit j must be routed to unit i and must traverse some number of nodes. As a simple example, assume M mobile units. With uniformly distributed mobile unit locations, we define γ as the mean number of hops from any node to the rest of the network. Therefore, the aggregate signaling rate associated with location information dissemination is

$$\mathcal{R} = \gamma \sum_{ij} r_{ij}\overline{H}_{ij} \tag{2}$$

The generality of this approach allows connection rates between *any* two entities to be defined; thus, not only "mobile units" but possibly routing tables as well — harkening back to the perplexed postman from the introduction who must travel streets that rearrange themselves from day to day. Possibly, this basic method can be extended to derive lower bounds for signaling costs associated with any mobility management scheme on any given network.

4. MOBILITY MANAGEMENT AND THE FUTURE

When considering mobility in present-day communications systems, what usually comes to mind is a person, conveyance, or mobile computer. However, in future networks, programs as well as physical objects might also be mobile. For example, suppose that you are an investment banker and seek financial market information distributed over the network. Specifically, you might wish to exploit small price differences over numerous "cybermarkets" — analogous to but more extensive than current arbitrage practices.

If the processing and communication capacity of the machines and network were infinite or there were no competing units, it would be a relatively simple task to spawn search processes on all machines in the network and have them report back information in some fashion. However, capacity constraints allow only a small number of programs to be launched into the network and run on other machines. For efficient search or to execute trades, the programs should be able to communicate and modify their behavior as information is gathered. For effective search, the programs should be able to relocate themselves to appropriate databases or markets. Such migrant programs charged with ferreting out information are generically called *mobile agents* [30–33].

The need for communication between agents implies a need for registration and paging since the agents are effectively mobile units to which calls may be routed. In motion processes that involve movement of mass, there are inherent constraints on motion between physical locations. Mobile agents, however, suffer few such constraints since their motion processes are influenced primarily by the location of contacts relevant to the search. Thus, one instant an agent could be active at a host in Los Angeles, and when done relocate to a suggested host in Madras.

For a simple random walk in n dimensions, paging cost is proportional to $(\rho t)^{n/2}$, where, once again, ρ is the mobility index and t is the time elapsed since last contact with the system (measured in units of intercall arrival). The intrinsic lack of constraints on both mobility index (how fast and far) and motion dimensionality (how many choices) for mobile agents suggests that groups of intercommunicating agents moving rapidly over a network of effective dimension $n \geq 3$ could severely stress signaling resources. Under this not-too-futuristic scenario, efficient mobility management could easily become the principal issue in network design.

5. CONCLUSIONS

The concept of mobility management based on time-varying mobile unit location probability distributions was introduced. Mobile units could be cellular telephone units, mobile computers or even mobile computer programs such as mobile agents. Regardless, the problem of finding mobile units boils down to the intuitively pleasing problem of searching for units in the most likely places as characterized by a unit location probability distribution.

Since location uncertainty is at the heart of the mobility management problem and information theory is the lingua franca of uncertainty, it is tempting to apply information theory to the paging problem, but sometimes unilluminating since even if the entropy of the location distribution is finite, it could still require infinite paging effort to find the unit on average. Thus, it is safest to rely directly on the ordered location probability function (ordered from most likely to least likely locations).

[4] Paging and registration are done by the *system* as opposed to the caller. Thus, the only a priori information available to the caller is the location distribution $p_j(t, t_0, x_j(t_0))$, which is independent of the registration/paging method used to track unit j.

We outlined various simple paging/registration procedures based on location probability distributions and compared them to more classical methods. A byproduct of this study was the definition of a *mobility index*, which is a useful reification of mobile unit location uncertainty and its growth as a function of time since the last contact. We then showed how paging/registration cost varies as a function of the mobility index.

We also showed how this cost varies with the dimensionality of the motion process which led to a consideration of non-classical mobility: groups of mobile programs ranging over a network in a coordinated way seeking out information. One could imagine agents that roam the network for information [33] or to buy and sell goods [31]. The results suggest that severe stress could be placed on a communication network by widespread use of such programs.

In the networks context, we then suggested ways in which time-varying location probability distributions for mobile units might be used to underbound the amount of signaling traffic necessary for distributing location information. In this case, an information theoretic approach *is* helpful.

BIOGRAPHY

Dr. Christopher Rose received the B.S. (1979), M.S. (1981), and Ph.D. (1985) degrees all from the Massachusetts Institute of Technology in Cambridge, Massachusetts, Dr. Rose joined AT&T Bell Laboratories in Holmdel, New Jersey, as a member of the Network Systems Research Department in 1985 and in 1990 moved to Rutgers University, where he is currently an Associate Professor of Electrical and Computer Engineering and Associate Director of the Wireless Networks Laboratory. He is Editor for the *Wireless Networks* (ACM), *Computer Networks* (Elsevier), and *Transaction on Vehicular Technology* (IEEE) journals and has served on many conference technical program committees. Dr. Rose was Technical Program Co-Chair for MobiCom'97 and Co-Chair of the WINLAB Focus'98 on the U-NII, the WINLAB Berkeley Focus'99 on Radio Networks for Everything and the Berkeley WINLAB Focus 2000 on Picoradio Networks. Dr. Rose, a past member of the ACM SIGMobile Executive Committee, is currently a member of the ACM MobiCom Steering Committee and has also served as General Chair of ACM SIGMobile MobiCom 2001 (Rome, July 2001). In December 1999 he served on an international panel to evaluate engineering teaching and research in Portugal.

His current technical interests include mobility management, short-range high-speed wireless (Infostations), and interference avoidance methods for unlicensed band networks.

BIBLIOGRAPHY

1. K. Meier-Hellstern, E. Alonso, and D. O'Neill, The use of SS7 and GSM to support high density personal communications, in J. M. Holtzman and D. J. Goodman, eds., *Wireless Communications: Future Directions*, Kluwer Academic, 1993.
2. M. J. Beller, E. H. Lipper, and M. P. Rumsewicz, Switching system impacts of PCS traffic, *2nd Int. Conf. Universal Personal Communications*, Ottawa, Canada, Oct. 1993.
3. G. Columbo, L. DeMartino, C. Eynard, and L. Gabrielli, Mobility load control in future personal communications networks, *2nd Int. Conf. Universal Personal Communications*, Ottawa, Canada, Oct. 1993.
4. C. N. Lo, S. Mohan, and R. S. Wolff, An estimate of network database transaction volumes to support voice and data personal communications services, *Proc. 8th ITC Specialist Seminar on Universal Personal Communications*, Santa Margherita, Italy, Oct. 1992.
5. E. H. Lipper and M. P. Rumsewicz, Teletraffic considerations for widespread deployment of PCS, *IEEE Networks* (*Special Issue on Nomadic Personal Communications*) (Sept./Oct. 1994).
6. E. H. Lipper, Switching system performance problems for universal personal communications, *Computer Networks and ISDN Systems*, 1995 (P. Enslow, Editor-in-Chief).
7. D. Goodman, P. Krishnan, and B. Sugla, Minimizing queueing delays and number of messages in mobile phone location, *ACM-Mobile Networks Appli.* (*MONET*) **1**(1): 39–48 (1996).
8. C. Rose and R. Yates, Ensemble polling strategies for increased paging capacity in mobile communications networks, *ACM Wireless Networks* **3**(2): 159–167 (1997).
9. C. Rose and R. Yates, Minimizing the average cost of paging under delay constraints, *ACM Wireless Networks* **1**(2): 211–219 (1995).
10. A. Bhattacharya and S. K. Das, LeZi-update: An information-theoretic approach to track mobile users in PCS networks, *Proc. ACM Mobicom'99*, Seattle, Aug. 1999.
11. A. Bhattacharya and S. K. Das, LeZi-update: An information-theoretic approach for personal mobility tracking in PCS networks, *ACM/Kluwer J. Wireless Networks* **8**(2): 121–135 (March 2002).
12. C. U. Saraydar and C. Rose, Minimizing the paging channel bandwidth for cellular traffic, *ICUPC'96*, Boston, Oct. 1996, pp. 941–945.
13. H. Xie, S. Tabbane, and D. J. Goodman, Dynamic location area management and performance analysis, *Proc. IEEE Vehicular Technology Conf. VTC'93*, Secaucus, NJ, May 1993.
14. S. Tabbane, An alternative strategy for location tracking, *IEEE J. Select. Areas Commun.* **13**(5): 880–892 (June 1995).
15. C. Rose, Minimizing the average cost of paging and registration: A timer-based method, *ACM Wireless Networks* **2**(2): 109–116 (June 1996).
16. R. Jain, Y.-B. Lin, C. Lo, and S. Mohan, A caching strategy to reduce network impacts of PCS, *IEEE J. Select. Areas Commun.* **12**(8): 1434–1444 (Oct. 1994).
17. R. Yates, C. Rose, S. Rajagopalan, and B. Badrinath, Analysis of a mobile-assisted adaptive location management strategy, *ACM Mobile Networks Appli.* (*MONET*) **1**(2): 105–112 (1996).
18. C. Rose, State-based paging/registration: A greedy approach, *IEEE Trans. Vehic. Technolo.* **48**(1): 166–173 (Jan. 1999).
19. A. Bar-Noy, I. Kessler, and M. Sidi, To update or not to update? *ACM Wireless Networks* **1**(2): 175–186 (1995).
20. C. Rose, *State-based paging/registration: A greedy technique*. Winlab-TR 92, Rutgers Univ., Dec. 1994.
21. C. Rose, A greedy method of state-based registration, *IEEE Int. Conf. Communications ICC'96*, Dallas, TX, June 1996.

22. U. Madhow, M. L. Honig, and K. Steiglitz, Optimization of wireless resources for personal communications mobility tracking, *IEEE Trans. Network.* **3**(6): 698–707 (Dec. 1995).

23. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.

24. R. E. Blahut, *Data Principles and Practice of Information Theory*, Addison-Wesley, Reading, MA, 1988.

25. J. L. Massey, Guessing and entropy, *IEEE Int. Symp. Information Theory*, Trondheim, Norway, 1994, p. 204.

26. E Arikan, An inequality on guessing and its application to sequential decoding, *IEEE Trans. Inform. Theory* **42**(1): 99–105 (Jan. 1996).

27. C. Rose and R. Yates, Location uncertainty in mobile networks: A theoretical framework, *IEEE Commun. Mag.* **35**(2): 94–101 (Feb. 1997).

28. S. M. Ross, *Stochastic Processes*, Wiley, New York, 1983.

29. C. Rose, Mean internodal distance in multihop store & forward networks, *IEEE Trans. Commun.* **40**(8): 1310–1318 (1992).

30. P. Maes, Agents that reduce work and information overload, *Commun. ACM* **37**(7): 31–40 (1994).

31. A. Chavez and P. Maes, Kasbah: An agent marketplace for buying and selling goods, *1st Int. Conf. Practical Application of Intelligent Agents and Multi-Agent Technology*, London, 1996.

32. P. Maes, Intelligent programs, *Sci. Am.* **273**(3): 84–88 (Sept. 1995).

33. H. Lieberman, Letizia: An agent that assists Web browsing, *1995 Int. Joint Conf. Artificial Intelligence*, Montreal, CA, 1995.

PARABOLIC ANTENNAS

ALESSANDRO ORFEI
CNR, Istituto di
Radioastronomia
Bologna, Italy

1. INTRODUCTION

The aim of this article is to introduce the important tutorial matters related to the specific field of parabolic antennas and to give an overview as complete as possible on the key parameters and factors to understand the operation of this widely used tool to transmit and receive radiowaves. Parabolic antennas can be designed in various ranges of radiofrequencies, spanning from ~100 MHz to 100 GHz, and a many applications take advantage of this well-established and versatile technology. Whether very small parabolas are used for commercial links or large reflectors are needed to detect faint signals coming from the sky, the fundamentals are the same; the particular application will address which aspects among others have to be taken into account.

2. ANTENNA PARAMETERS AND CHARACTERISTICS

There are many parameters characterizing the operation and performance of parabolic antennas. Their importance in designing a system is dependent on the application; for instance, large antennas have to be carefully designed with respect to structural and mechanical constraints, because induced deformations, arising from gravity, wind, and temperature, dictate most of the performances. On the other hand, simple small parabolas used in receiving satellite TV signals need reasonable design, but the cost and ease of effective mass production are of primary importance.

2.1. Geometry of Used Configurations, Primary Focus-Fed Parabola

The simplest way to collect electromagnetic energy is to use a paraboloidal mirror only. Exploiting the geometric definition, a plane wave incoming at different points of the parabolic surface will be focused at the focus (called *primary focus*) of the parabola. If the phase center of a *feed* (or *illuminator*) coincides with the focus, the energy will be collected by the waveguide and then transformed in an electric signal (Fig. 1). A paraboloidal surface is obtained simply by rotating a parabola about its *focal axis*, that is the axis joining the *vertex* of the paraboloid and the primary focus. The distance between these two points is called *focal length*.

2.2. Geometry of Used Configurations, Cassegrain/ Gregorian System

Collection of electromagnetic energy could also be achieved by adding a portion of a hyperbolic shape as a second mirror. Exploiting the geometrical definition of hyperbola the energy will be focused at one of the hyperbole foci if the other one coincides with the paraboloidal primary focus. In this case the first one is called a *secondary focus* (F2) of the paraboloid and the phase center of the feed is placed to coincide with it. Again, the hyperboloid is obtained by revolution of an hyperbola about its axis. This kind of configuration is called *Cassegrain* system (Fig. 2a). The goal is also achieved if the second mirror is a portion of an ellipsoid (Fig. 2b). In this case the configuration is called a *Gregorian* system. In both systems the *secondary mirror* is also called a *subreflector*.

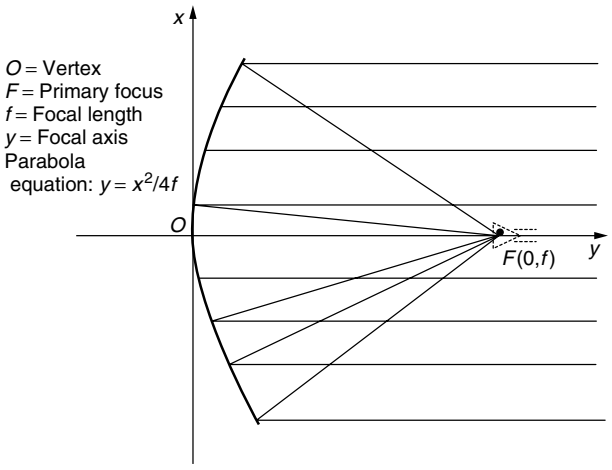


Figure 1. Primary focus-fed system.

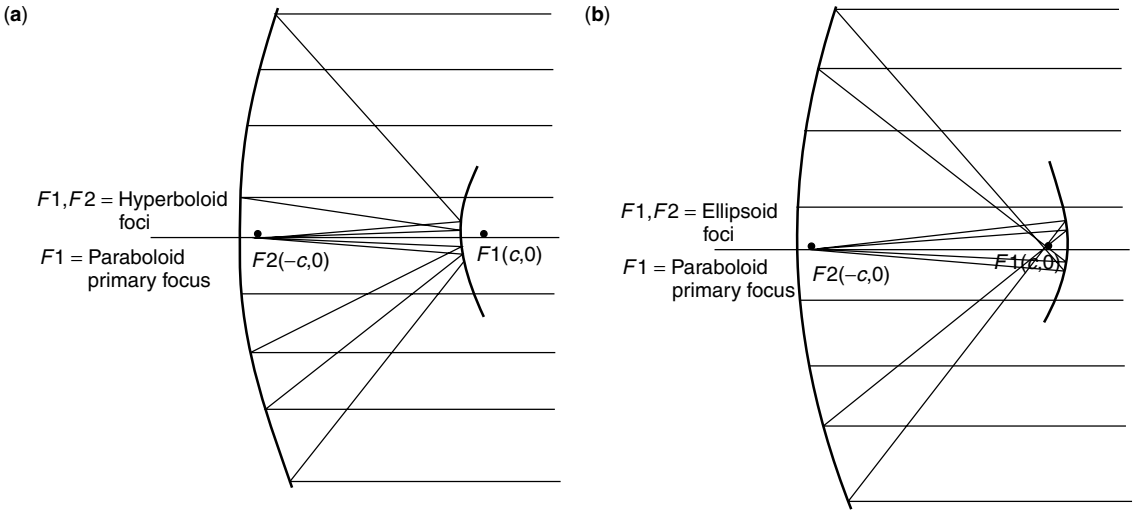


Figure 2. (**a**) Cassegrain system; (**b**) Gregorian system.

In applications where very high antenna efficiency and very particular antenna illumination are necessary, both mirrors have a shape that slightly deviates from a perfect parabolic/hyperbolic or parabolic/elliptic pair. In this case the configuration is called a *shaped* system. A shaped form of the illumination function can have a dip in the central portion of the mirror because that surface is obscured by the primary focus arrangement and so it is not useful in picking up a signal. The central dip also improves the return loss of the system, avoiding a standing wave.

2.3. Geometry of Used Configurations, Offset System

The *offset* system can be built starting with one of the three classic configurations shown in Figs. 1 and 2. Let's imagine that we remove part of the paraboloidal surface, keeping only the subsurface that collects the electromagnetic energy that doesn't interfere with either the feed or the secondary mirror (Fig. 3). This particular and very difficult-to-design configuration avoids the efficiency loss for the blockage effect due to the obstruction of the feed or secondary mirror. An impressive realization of this configuration is the GBT (Green Bank Telescope, Green Bank WV), a radiotelescope with the primary dish of 100 m in diameter.

2.4. Geometric Parameters

Referring to Fig. 4, seven parameters are used to describe and design the optics of the parabolic system:

D = diameter of the primary mirror
d = diameter of the secondary mirror
f = focal length
Φ_p = primary mirror edge half-angle
Φ_s = secondary mirror edge half-angle
L = secondary mirror depth
F = secondary mirror focal length

For describing the Cassegrain/Gregorian system, only four of them have to be fixed/the other ones are dependent by the following three equations:

$$\tan\frac{\Phi p}{2} = \pm\frac{D}{4f}\begin{cases}+\text{Cassegrain;}\\ -\text{Gregorian}\end{cases} \tag{1}$$

$$\frac{1}{\tan\Phi p}+\frac{1}{\tan\Phi s}=\frac{2F}{d} \tag{2}$$

$$1-\frac{\sin\left(\frac{\Phi p-\Phi s}{2}\right)}{\sin\left(\frac{\Phi p+\Phi s}{2}\right)}=\frac{2L}{F} \tag{3}$$

One of the most representative parameter for the design is the ratio f/D, because many of the electromagnetic characteristics have mathematical dependence on it.

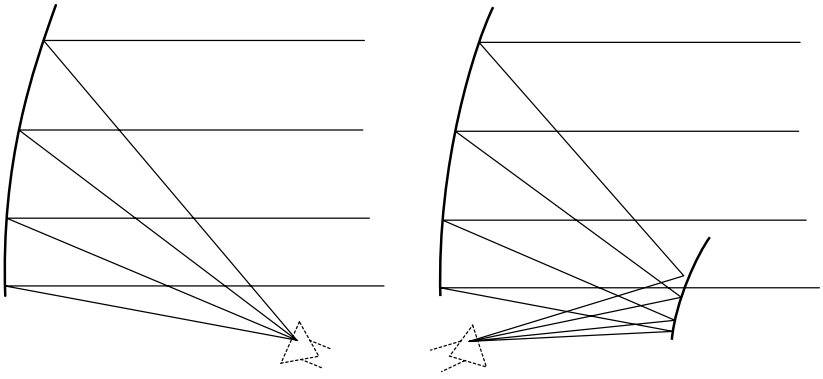


Figure 3. Offset configurations.

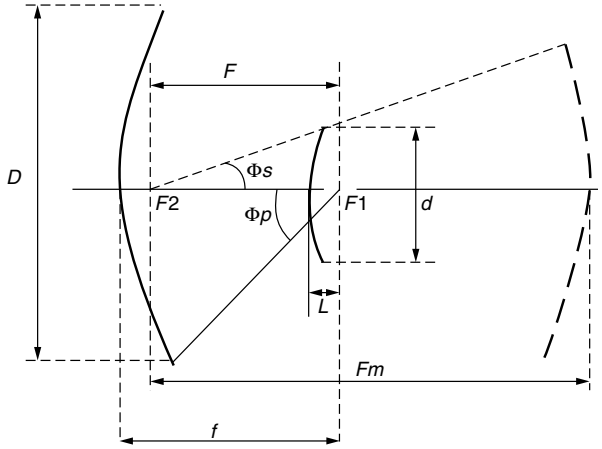


Figure 4. Antenna parameters.

Figure 4 also shows the concept of an *equivalent parabola*, which is useful in introducing the term known as *magnification* M, defined as the ratio between the focal length of the equivalent parabola and the focal length of the real system

$$M = \pm\frac{Fm}{f} = \frac{\tan\dfrac{\Phi p}{2}}{\tan\dfrac{\Phi s}{2}} (+\text{Cassegrain}, -\text{Gregorian}) \qquad (4)$$

The equivalent parabola has the same diameter and feed of the real system but is a paraboloid of different curvature. Its property is to focalize the incoming rays at the same point of the real dual-mirror system. This concept is useful in appreciating that a dual-mirror system gives a longer focal length, avoiding any lengthening of the mechanical structure in front of the primary mirror.

2.5. Electromagnetic Characteristics of Parabolic Antennas

2.5.1. Pattern. The combination of each geometry previously described and the illuminator gives the properties by means of which the antenna can irradiate or receive an electromagnetic signal. The first characteristic to be considered is the *pattern*. The antenna pattern relates the spatial distribution of the transmitted or received power [power pattern, $P(\theta, \phi)$]. Similarly, the pattern gives the value of the electric field at every point in space [field pattern, $E(\theta, \phi)$]. Usually the patterns are functions of spherical coordinates and often are normalized values with respect to the maximum of the function. The simplest pattern refers to an isotropic antenna, namely, an antenna that radiates the same amount of power in all directions. However, an isotropic antenna is inappropriate in cases where the antenna must pick up signals from a specific direction at a time while avoiding spurious signals coming from the ground (increasing *antenna noise temperature*) or unwanted other transmitters (interference). If the useful signal is to be transmitted to or received from different directions, the antenna can be pointed in the desired direction. In this way the antenna is directive and the *antenna gain* will be much higher than an isotropic one, but only at specific desired directions in space. Therefore, from Fig. 5a–c, many electromagnetic parameters can be defined.

The range of angles of maximum propagation are referred as the *main lobe* and numerically are within the half-power beamwidth (HPBW), namely, all the directions between the maximum power received (or transmitted) and its half-power. HPBW is often called the antenna *beam* and the following equation holds:

$$\text{HPBW} = k_i \frac{\lambda}{D} \qquad (5)$$

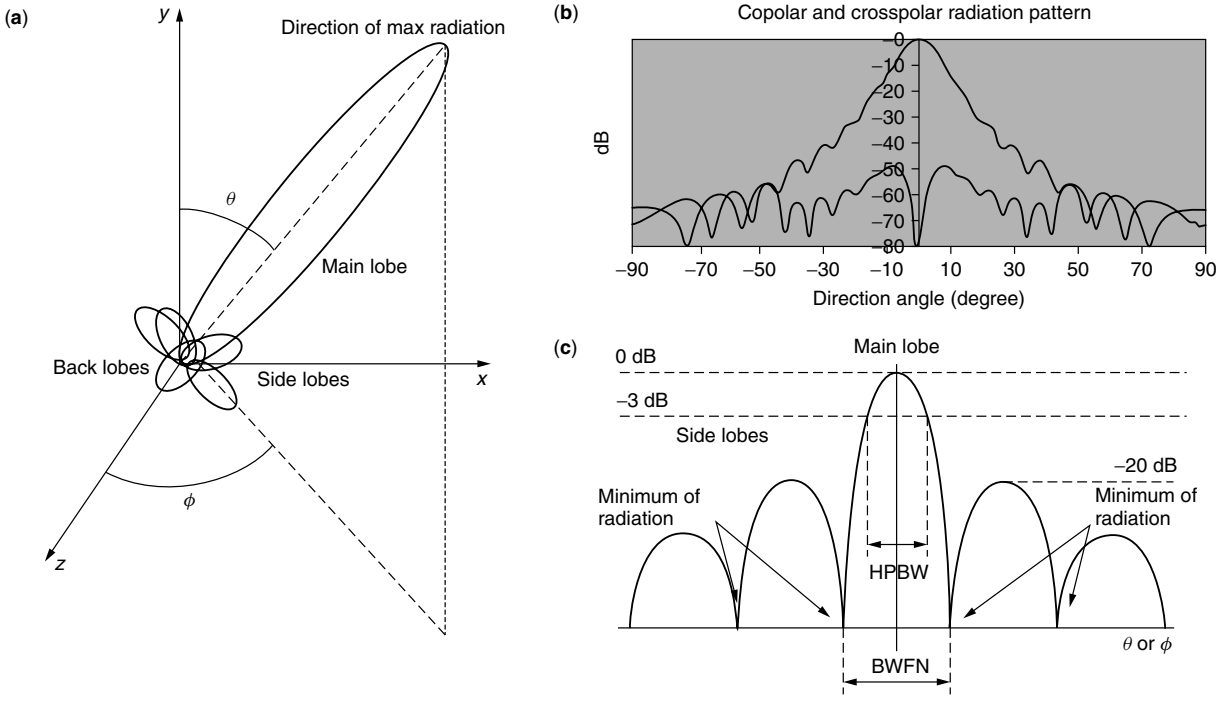


Figure 5. Antenna pattern.

where λ is the operating wavelength, D is the diameter of the antenna aperture, and k_i is a factor, always near unity, depending on which kind of function the feed illuminates in the mirror of the antenna. Often HPBW is also considered a measure of the antenna *resolution*; that is, its capability to discriminate two closely spaced objects in space. To clarify this statement, in Fig. 5c, note that the distance between the direction of max pick up and the first minimum is half of beamwidth between first nulls (BWFN). This means that two objects spaced at these locations are distinguishable because from one point the pickup is maximum while the other one is null. It could be shown that HPBW is approximately equal to BWFN/2.

Figure 5c represents a cut in either the θ or Φ plane (see Fig. 5a) and shows how low the near sidelobes are with respect to the main lobe. This is indicated as -20 dB. This illustration could be viewed as a zoom of the more complete polar pattern in Fig. 5b. In this figure the relative amount of power radiated (or received) as a function of angle is shown (polar or *copolar pattern*) together with the radiation in a perpendicular direction where, ideally, the radiation should be zero (*cross-polar pattern*).

2.5.2. Gain and Aperture Efficiency. It can be shown that the antenna gain has a direct relation to the collecting area of the antenna:

$$G = \eta \frac{4\pi}{\lambda^2} Ag \tag{6}$$

Although G is a dimensionless quantity, often given in decibels by simply taking $10 \log_{10}(G)$, there are application fields, such as radio astronomy, in which the antenna gain assumes the meaning of how much the antenna temperature is increased when the antenna surface receives a given amount of power density per unit bandwidth. In this case the dimension of gain is kelvins over jansky, where jansky is a unit defined as

$$1 \text{ Jy} = 10^{-26} \frac{W}{\text{m}^2 \cdot \text{Hz}} \tag{7}$$

and the antenna gain can be calculated in the following way:

$$G_R = 10^{-26} \eta \frac{Ag}{2k} \tag{8}$$

where k is the Boltzmann constant. The two ways to express the antenna gain are equivalent; what simply changes is the measurement unit. If G_R is known, we can calculate η and use Eq. (6) to get G; conversely, if G is available, we can calculate η and then get G_R from (8).

Ag is the geometric area of the aperture of the parabolic antenna. *Aperture* is the cross section subtended by the dish, and for parabolic antennas it is a circle with diameter D. λ is the wavelength at which the gain is to be calculated or measured, and η is called the aperture efficiency. The *aperture efficiency* is a number less than one and acts in reducing the real area that is effective in collecting the electromagnetic energy coming into the aperture. η originates from many causes, each of them described by an appropriate efficiency parameter, and in general depends on a lot of variables, including the frequency, direction of pointing, structural deformations of the antenna due to gravity, temperature and wind, and type of function by which the feed illuminates the dish. Trying to clarify as simply as possible, we could start by stating (9), which relates the causes that most affect the efficiency:

$$\eta = \eta_b \eta_x \eta_{\text{ph}} \eta_{\text{sp}} \eta_{\text{diff}} \eta_{\text{ill}} \eta_{\text{surf}} \eta_{f\text{loss}} \eta_{\text{vswr}} \eta_{s\text{loss}} \tag{9}$$

η_b is the *blocking efficiency*. It comes from the obstruction of the feed or subreflector, and the supporting legs raise at the incoming electromagnetic energy. Practical values span from 0.85 to 0.95, except for the antenna offset solution. A rule of thumb to get the order of magnitude of η_b is to compute the ratio of the total blocked area A with the area of the antenna aperture Ag, then

$$\eta_b = \left(1 - \frac{A}{Ag}\right)^2 \approx 1 - 2\frac{A}{Ag} \tag{10}$$

η_x is the *cross-polarization efficiency*. It arises when the polarization of the incoming wave doesn't match the polarization of the antenna. The extreme example should be a dipole sensitive at the horizontal polarization: η_x is zero for vertical polarized waves, so they cannot be detected. Parabolic antennas have a high degree of symmetry, so they are sensitive to both linear and circular polarizations and η_x has very high values (0.99 or better), particularly if the feed is circular and designed so that the amplitude and phase patterns are equal in the two orthogonal planes of maximum radiation (E and H planes). The situation worsens if the antenna components are not perfectly aligned to each other or the feed is not symmetric. In the case that the antenna is an offset system, a very careful design must be developed if a low cross-polarization level is needed.

η_{ph} is the *phase efficiency*. If all rays (see Figs. 1–3) don't arrive in phase (e.g., because the mirror deforms under structural loads or because the feed is not able to perfectly illuminate in phase all directions), a small amount of power can be lost. This term is generally negligible for classic parabolic antennas, 0.99 or so, but it could be very low for shaped antennas when used with primary focus receivers, because the equalizing effect of the subreflector is absent. However, in this case, the effect is very frequency-dependent, putting a limit on the highest usable frequency of receivers placed at the primary focus.

η_{sp} is the *spillover efficiency*. When a feed illuminates the primary or secondary mirror of the antenna, the energy at the edges cannot go sharply to zero. Thus, a fraction of the total illumination energy will be lost: the ratio between this fraction and the total energy is called "spillover efficiency". This term is generally the result of a compromise, as the illumination is far from uniform, the higher is η_{sp} but the lower is the effective area. Normally antennas have an illumination function different from uniform, so a certain amount of tapering is used. The *taper* is how much the function is lower at the edge with respect to its maximum value. This is one of the design parameters for the feed and is evaluated at the angle Φp or Φs depending on which geometry is chosen, and usually the tapering is higher in the primary focus configuration than the secondary. The reason is that a primary mounting of the feed "sees" the ground at angles greater than Φp,

so it picks up an unwanted noise temperature at about 300 K that must be attenuated. Instead, by illuminating the subreflector, the feed "sees" the atmosphere at angles greater than Φs, that it is cooler than the ground.

η_{diff} originates from *diffraction* due to edge effects for both primary and secondary mirrors.

η_{ill} is the *illumination efficiency* and accounts for the fact that the illumination function is not uniform over the aperture, so it is a measure of the reduction of gain due to tapering.

η_{surf} is the *surface efficiency*. The rays colliding on the antenna surface find a nonideal shape. The subreflector and primary mirror surfaces have roughness. Furthermore, large-diameter antennas have a primary mirror consisting of a lot of aluminum panels drawn close, which also means that their relative alignment is a concern. Sometimes larger subreflectors are made by panels as well. Antennas suffer important deformation due to gravitational and wind effects as their dimensions increase; also temperature-induced deformation of the surfaces must be taken into account. The net result is that incident rays are reflected by nonperfect surfaces, resulting in phase errors that reduce the gain. Manufacturing errors of the panels and surface, temperature, and wind effects on the mirrors and on the antenna supporting structure, and also gravitational deformations, are treated like random errors. Therefore the parameter indicating the departure from the ideal shape is the RMS (root-mean-squared) value of the real surface with respect to the ideal one. Because of the many the causes, the RMS values must be combined to obtain the *total surface* RMS σ. Usually σ is the RSS (root sum squared) of the RMS values:

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2} \tag{11}$$

The surface efficiency follows a Gaussian function depending on the σ/λ ratio:

$$\eta_{\text{surf}} = e^{-[4\pi(\sigma/\lambda)]^2} \tag{12}$$

$\eta_{f\text{loss}}$ is the *feed loss efficiency*. When the electromagnetic energy passes through the feed, the waveguide attenuates a small amount of power, so the power for illuminating the aperture is slightly less than the supplied one. The insertion losses are generally low — let's say in the range 0.1–0.4 dB, which means that $\eta_{f\text{loss}} = 0.91 \div 0.98$.

η_{vswr} is the feed *return loss efficiency*. Together with the insertion loss mentioned previously, the feed loses power because a part goes back due to a nonperfect impedance matching. Return losses less than -15 dB are easily obtained, so $\eta_{\text{vswr}} \geq 0.97$.

$\eta_{s\text{loss}}$ is the *surface loss efficiency*. The surface of the mirrors are conducting electric currents; thus ohmic losses due to material resistivity arise. The effect is low, $\eta_{s\text{loss}} = 0.99$.

To conclude this section, it is worthwhile to mention other possible causes that reduce the antenna gain. Up to now a perfect geometric alignment of the three antenna components — the primary mirror, the secondary mirror, and the feed — were assumed. In the case that either the feed or the subreflector are not well positioned on the focus in the direction of its axes, a *defocusing* effect occurs. Defocusing exists in a movable antenna also if it were properly aligned. In fact, the alignment holds for a single position of the antenna, because the antenna deformation, as the elevation changes, will move the focus so that the feed or subreflector will be defocused. If a proper tracking of the focus movement is necessary to recover that amount of loss, a mechanical facility must be added in the antenna design in order to move the feed or subreflector.

Another useful concept is the *field of view* (*FOV*). It is related to a displacement of the feed with respect to the focus position outside the focal axis. If this is the case, a reduction of gain, together with an increase of sidelobe level and cross-polar pattern, will be experienced. The field of view could be defined as the space region where the feed can be displaced losing no more than a fixed amount of gain. Of course, FOV is not an absolute parameter, but it depends on the amount of gain loss that is considered acceptable. Generally it has an angular dimension, but it could also be expressed as HPBW times or as a multiple of wavelength. The FOV concept suggests that in the real world the focus of an antenna must not be viewed as a point just outside of which the system doesn't completely work. Instead a "focal surface" exists where the performance of the antenna worsens with the distance that the feed is placed with respect to the focus. The FoV can be exploited to use the antenna with more than one frequency by placing receivers working at different wavelengths, or using feed arrays (many identical feeds working at the same frequency).

2.6. Structural and Mechanical Aspects of Parabolic Antennas

Small parabolic and stationary dishes don't give particular structural and mechanical problems. They are mounted on a lattice or on a mast, pointed in a fixed direction, and their performances are not affected by temperature, wind stress, or operating environment. On the other hand, some applications call for the antenna to be protected from all these causes, and thus it is completely enclosed in a *radome*, a microwave transparent dielectric housing protecting the antenna from adverse environmental conditions.

Between these two extremes a lot of applications use parabolic antennas without radomes, that experience all weather conditions and, because of their dimension and weight, also gravitational deformations both for the dish and the supporting framework. The general characteristics of a movable and large reflector antenna are described in the following text. The following structural and mechanical elements are shown in Fig. 6.

- A *concrete foundation* with pillars, which is embedded some meters in the ground.
- A *track* over which the *wheels* move. This is a very common solution to allow the rotation of the antenna in the azimuth direction.
- The *alidade*, which is the supporting structure of the antenna.
- The *elevation wheel*, which allows the antenna to rotate in the elevation direction.
- The *backup structure*, which supports the primary mirror.

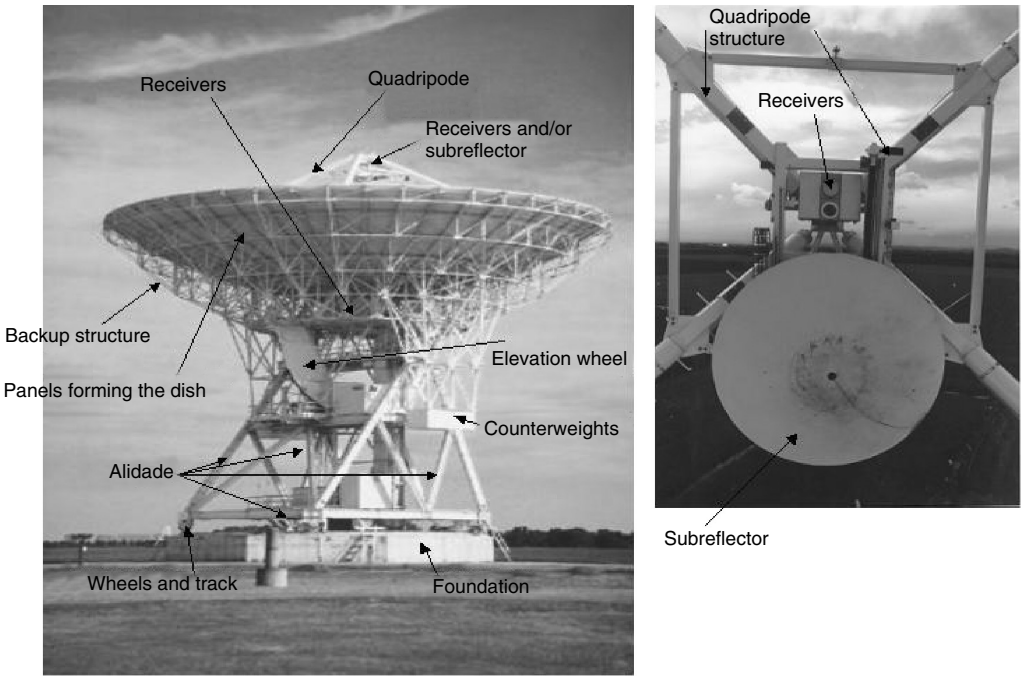


Figure 6. Elements of a parabolic antenna (Medicina observatory radiotelescope, Bologna, Italy).

The *quadripode*, usually three or four legs, which supports the secondary mirror and primary focus receivers.

The *panels*, which form the primary mirror of the antenna.

The right side of Fig. 6, a zoom of the primary mirror location, shows a very particular arrangement of the mechanical coexistence of the subreflector and the primary focus receivers. This allows a fast change among receivers placed at both the primary and secondary focus.

The wheels and track are made of iron alloy, and this is interfaced with the concrete foundation by means of either a suitable grout solution or, better, iron plates over which the track is bolted and grout under the plates. The plates are bolted in the concrete by anchor bolts. The alidade, elevation wheel, and backup and quadripode structure consist of large steel beams, trusses, frames, and brackets. The panels are made of aluminum, covered by a highly reflective white paint, the dimension and quantity of which are determined by the diameter of the dish; to give an idea, a mirror with a diameter around 30 m needs 200 or more panels, with an area of 3–5 m^2 each. The specifications regarding the panels are given in terms of manufacturing error and deformation under gravity, temperature, and wind. All of these are RMS values with respect to the ideal parabolic contour and they contribute, together with the RMS value of the subreflector surface, the RMS deformation of the structure and the mirrors alignment error, to the total surface accuracy σ [see formula (11)]. σ is said to determine the minimum operating wavelength of the antenna by means of the conventional relation

$$\sigma = \frac{\lambda \min}{20} \tag{13}$$

The subreflector can be made of fiberglass or, if the dimension is too large, by aluminum panels. Usually, the dimension for the subreflector is around $D/10$, where D is the diameter of the primary dish. It results as a compromise by taking the blockage at minimum values together with an efficient illumination of the paraboloid. The shaping of the antenna may instead require a subreflector diameter higher than $D/10$.

Sun exposure heats in a nonuniform way all the elements of the structure, both deforming the shape of the mirror and changing the direction of pointing. Further, in these heavy antennas (a 30-m antenna can weigh 200 tons or more), gravity takes a great role in deforming the backup structure. The departure from a true parabolic shape changes with respect to the elevation. To overcome the gravitational effect, some antennas adopt a structural design called *homology*; the mirror maintains a parabolic shape, changing the focal length only (i.e., the mirror opens or closes, maintaining the symmetry). By focusing the subreflector, the focal length can be tracked for each elevation. This design results in a much heavier antenna and significantly increases the cost.

2.7. Pointing

2.7.1. Overview. This subject is rarely taken into account in most applications and books, but it is worthwhile to mention both for completeness and for those applications that need a precise tracking of a target.

In Section 2.5 the term *beam* and the acronym HPBW were introduced, indicating a measure of the angular size where most of the radiation is contained. If the antenna has to point at a target and, above all, has to track it, the response of the antenna servosystem to the target coordinates must be within the pointing performance, to maintain the target inside the antenna beam. It is easily

recognized here that a pointing error can be viewed as a loss of power. It acts like the antenna efficiency terms, multiplying by a factor of <1 the amount of power received (transmitted) from (to) the target. To give a quantitative example, suppose that the antenna main lobe is Gaussian so it can be expressed in terms of HPBW by the following equation (see also Fig. 7):

$$\eta_{\text{point}} = e^{-(1.665*\theta/\text{HPBW})^2} \tag{14}$$

η_{point} is unity for perfect pointing, but rapidly decreases if errors occur. If the pointing error $\theta = \text{HPBW}/2$, half of the power is lost and an error of equal to one-fifth of the beam is enough to loose 10% of power.

Pointing errors can be divided into two classes: systematic and nonsystematic. *Systematic* errors are repeatable errors, and a mathematical model can be predicted. Gravitational deformations induce a pointing error. The erection of the antenna leaves unavoidable errors such as a slight nonorthogonality between the azimuth and elevation axis, nonperfect horizontal azimuth plane, and a mechanical axis of the mirror different from the electromagnetic direction of maximum pickup. All these factors induce systematic pointing errors that can be measured or derived from the best-fitting technique.

Nonsystematic errors are random errors and are due to temperature effects on all the elements of the antenna structure, for example unevenly expanding alidade trusses, and wind forces acting so that average wind and its gusts slightly move the antenna. These are not predictable and environment-dependent, in the sense that the pointing accuracy is a function of the amount of wind, absolute temperature, and its drift. Generally the antenna is said to work in three possible conditions — precision, normal, and extreme operation, indicating worsening of performance as wind and temperature effects increase.

2.7.2. Beam Deviation Factors. In Section 2.5 the case of a displaced feed was reported, listing the effects on the antenna pattern. A feed displacement gives pointing displacement as well. Generally speaking, the movements of all elements forming the antenna geometry, primary or secondary focus feed, subreflector, or primary mirror, cause pointing errors. These movements can be the translation or rotation of the element that originates a misalignment angle α, causing a pointing angle error β: the ratio between these angles is called the *beam deviation factor* (BDF), which is a function of the antenna parameters, namely, focal length f, diameter D, magnification M, and secondary mirror focal length F and for most cases its value ranges from 0.7 to 0.9. In the following, a survey of possible situations is presented,

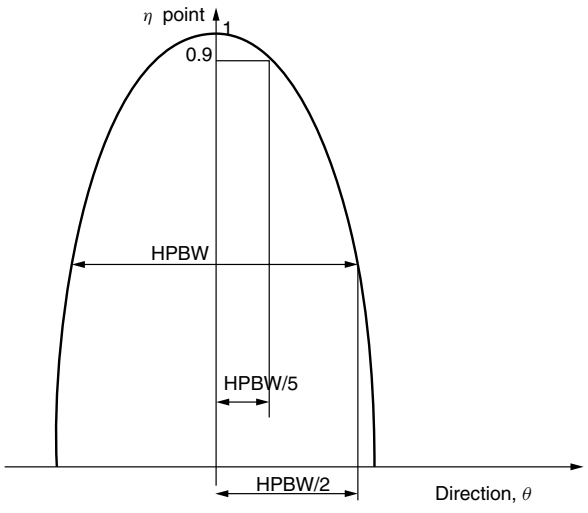


Figure 7. Pointing error.

- Primary feed lateral displacement (Fig. 8a):

$$\text{BDF} = \frac{\beta}{\alpha_p} = \frac{\beta}{\tan^{-1} d/f} = \frac{\sin^{-1}\left[\dfrac{d*\left(1+k_i(D/4f)^2\right)}{f*\left(1+(D/4f)^2\right)}\right]}{\tan^{-1} d/f} \tag{15}$$

k_i is the same factor appearing in Eq. (5). In the case $d \ll f$, so that $\tan^{-1}$ and $\sin^{-1}$ are equal to their argument, a more practical and usual relation can be used:

$$\text{BDF} = \frac{1 + k_i * (D/4f)^2}{1 + (D/4f)^2} \tag{16}$$

- Secondary feed lateral displacement (Fig. 8b):

$$\text{BDF} = \frac{\beta}{\tan^{-1}\left(\dfrac{d}{M*F}\right)} \tag{17}$$

- Subreflector rotation (Fig. 8b):

$$\text{BDF} = \frac{\beta}{\tan^{-1}\left(\dfrac{F}{f}*\dfrac{2\alpha_s}{M+1}\right)} \tag{18}$$

- Subreflector displacement (Fig. 8b):

$$\text{BDF} = \frac{\beta}{\tan^{-1}\left(\dfrac{h}{f}*\dfrac{M-1}{M}\right)} \tag{19}$$

The situation is such that by moving an element to one side, the beam is deviated to the opposite side (as shown in Fig. 8a), so to recover the right pointing, the antenna must be moved according to the element. BDF values allow us to understand the pointing error sensitivity of the antenna with respect to misalignments or effects due to the stiffness of the antenna elements.

2.8. Antenna Noise Temperature

Regardless of whether the antenna is receiving, or transmitting the wanted signal, a certain amount of noise power is picked up by its pattern. The amount of noise power is expressed as an equivalent resistor at temperature T_a, called the *antenna noise temperature*, which, when matched at the antenna receiver input in place of the antenna, gives the same amount of noise power. It is recalled here that the relation between power and temperature is

$$P = kT_aB \tag{20}$$

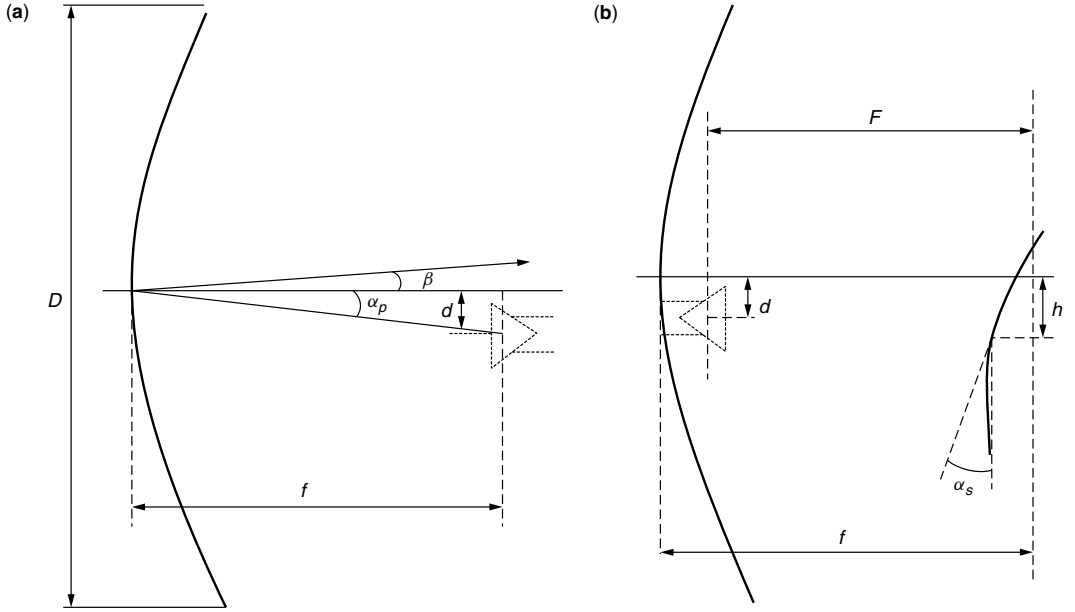


Figure 8. Beam deviation factors: (**a)** primary and (**b)** secondary feed lateral displacement, subreflector rotation and lateral displacement.

where k is the Boltzmann constant, B is the bandwidth, and T_a is the equivalent temperature.

The noise power comes from different origins:

- Noise power picked up from ground (conventionally at a physical temperature equal to 290 K) by backlobes or by antenna spillover.
- Noise power coming from cosmic sources, from the sky or planets and received by either the mainlobe or sidelobes. At least the antenna receives the big bang remnant, 2.73 K, which is uniformly distributed over all the sky and not dependent on frequency. Galaxies or other sky objects emit signals in the form of noise. The same could be said for planets and the sun of our solar system. The amount of noise power picked up depends strongly on frequency, on the object, and on how the antenna lobes point toward the object.
- Noise power from the atmosphere. The atmosphere is an absorbing medium, so it acts like an attenuator producing noise. The atmosphere noise varies with frequency and with the elevation angle at which the antenna is pointing. Low elevation angles give a higher amount of noise because the radio path is longer than that at high elevation.

2.9. G/T

By adding the antenna noise temperature to the receiver noise temperature, the so called *system temperature* T is obtained; that is, the overall noise of the complete antenna receiving system proportional to the inferior limit on the detection of signals. In any case, the detection of signals also depends on the antenna gain; the larger the reflector, the higher is the amount of power received, but this also holds for antenna noise. The ratio between T and the antenna gain G, the term G/T, gives a figure on the capability of the antenna to detect small signals. G/T can also be used to compare different antenna receiving systems; if a larger antenna has a very noisy receiver, its performance can be worse than a smaller antenna having a receiver, at the state of the art, with very low noise.

3. APPLICATIONS

Parabolic antennas have many applications in various fields requiring very different performance. Depending on the application, the design will encompass all or part of the characteristics described in Section 2.

3.1. Radio Astronomy

This relatively young science uses many large reflector antennas located all over the world to receive signals from many different natural sky objects such as galaxies, quasars, stars, and planets. The nature of the received signals is generally noise or plenty of spectral lines. If an object has to be resolved, the interferometry technique is used. This means that many parabolic antennas are used in observing the same object at the same time. In that case the value D of the interferometer is the distance among the antennas that can span from some kilometers to thousands of kilometers. This last case is called a *very-long-baseline-interferometry* (VLBI) technique and allows us to reach angular resolution as low as milliarcseconds at microwave frequencies [by applying Eq. (5)].

The parabolic antennas used in radio astronomy are of all types described in Sections 2.1 to 2.3, and operative frequencies span from about 300 MHz to 30 GHz for centimetric wave antennas to hundreds of gigahertz for millimeter and submillimeter antennas.

This field of application often needs a very careful design for all the antenna characteristics described in Section 2.

3.2. Microwave Relay Link

A microwave relay link is intended as a link that transmits and receives the radio signal (e.g., the broadcasting of

analog and/or digital signals) between parabolic antennas many tens of kilometers in distance and on a line-of-sight path. The diameters of the antennas used are a few meters, and the operative frequency range is in the microwave region from about 2 to 20 GHz. The Friis formula addresses the design of the link:

$$\mathrm{Pr} = \frac{\mathrm{Pt} * \mathrm{Gt} * \mathrm{Gr}}{(4 * \pi * r)^2} * \lambda^2 \tag{21}$$

P_r = received power
P_t = transmitted power
G_t, G_r = gain of transmitting and receiving antenna
r = distance of the link
λ = operating wavelength in free space of the transmitted and received signal

3.3. Satellite Communication

Most of the earth stations in a satellite communication system are parabolic antennas of all types described in Section 2. Both single feed and feed array configurations are used. In this last case, displaced location of the feeds with respect to the reflector are used to get different pointing directions in order to receive signals from different satellites. The used frequencies span from a few gigahertz to over 12 GHz. The diameter of the antennas can range well over 10 m.

3.4. Remote Sensing

Parabolic antennas can be used in the microwave range for radiometric measurements of the atmosphere parameters and earth and sea surface characteristics. Also in this case, scanning beam techniques and multifrequency measurements call for use of a feed array.

BIOGRAPHY

Alessandro Orfei received his degree in 1983 from Bologna University, Department of Electronic Engineering. He worked for three years at the G. Marconi Foundation Laboratories in the field of the fiber optics, and then for three years in a private company as a design engineer in the field of telecommunication. Since 1989, he has been a researcher at the Istituto di Radioastronomia–Consiglio Nazionale delle Ricerche (Italy). He is in charge of all work concerning the VLBI (very-long-baseline interferometry) 32m antenna at the Medicina Radio Observatory (Bologna, Italy).

PARTIAL-RESPONSE SIGNALS FOR COMMUNICATIONS

Apostolos Rizos
AWARE, Inc.
Bedford, Massachusetts

1. INTRODUCTION

Partial-response signals are those where intentional controlled intersymbol interference (ISI) is introduced between successive symbols, either for channel-matching or bandwidth-efficiency purposes. In contrast to a partial response signal, a *full-response signal* does not introduce any intentional ISI between successive symbols.

To explain these definitions in more detail, let us remember the basic representation of a linear digital modulation technique [1]

$$s(t) = \sum_{n=0}^{\infty} I_n g_T(t - nT) \tag{1}$$

A binary sequence d_n with values $\{0, 1\}$ is mapped to an information-bearing sequence I_n, which is typically an amplitude value that will scale the transmitting filter output $g_T(t)$ (the subscript T refers to the transmitter and not the symbol rate $1/T$). For binary PAM (pulse amplitude modulation), I_n will take values $A, -A$, so that a binary 1 will be transmitted using pulse $A \cdot g_T(t)$ and a binary 0 will be transmitted using the inverse pulse $-A \cdot g_T(t)$.

Note that for higher-order modulations (e.g., 4-PAM), k bits of the binary sequence are mapped into a 2^k-level information symbol I_n. The information sequence I_n may also be complex-valued, which corresponds to a QAM (quadrature amplitude modulation) scheme that needs to be modulated onto a carrier; the real part of the information signal will modulate the cosine of the carrier, while the imaginary part will modulate the sine (which is orthogonal to the cosine, and hence, can be distinguished from it). Without loss of generality, we will assume the information sequence to take values $\{1, -1\}$, and the cascade of the information sequence and the transmitting filter $g_T(t)$ to give a total energy $\mathcal{E}_b$ for each transmitted bit.

Assuming a linear transmission channel with impulse response $c(t)$, and a receiving filter with impulse response $g_R(t)$, then the output of the receiver filter will be

$$r(t) = \sum_{n=0}^{\infty} I_n x(t - nT) + v(t) \tag{2}$$

where $v(t)$ is the noise from the channel and $x(t)$ is the combined effect of $g_T(t), c(t), g_R(t)$. Its frequency representation (Fourier transform) will be the product of the Fourier transforms of its components

$$X(f) = G_T(f)C(f)G_R(f) \tag{3}$$

By sampling the output of the receiving filter every T seconds, we obtain the decision variables that are used for the estimation of the information sequence

$$r_m = x(0)I_m + \sum_{n=0, n\neq m}^{\infty} I_n x(mT - nT) + v(mt) \tag{4}$$

We see that the mth received sample r_m depends on the mth information symbol I_m, but also [depending on the values of $x(nT), n \neq 0$] on the adjacent symbols $I_n, n \neq m$. This is called *intersymbol interference* (ISI). A full-response

signal is designed in such a way that the combination of transmit and receiver filters will give no ISI

$$x(n) = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \tag{5}$$

while a partial-response signal is designed in such a way that the combination of transmit and receive filters will give controlled ISI, $x(n) \neq 0$ (but equal to predetermined values) for more than one sample n. The controlled ISI amount in a PR system is chosen in such a way so as to satisfy the system requirements, such as small bandwidth or no DC spectral component.

An example of a full-response signal is shown in Fig. 1. The basic transmitting pulse is a rectangular pulse $p(t)$ of duration T, and we assume that both the channel and the receiver filter introduce no change to the signal: $C(f) = G_R(f) = 1, \forall f$. We see that each symbol does not interfere with adjoining symbols.

There is a problem with a signal such as this rectangular pulse; it is not bandwidth-limited; that is, its Fourier transform $P(f)$ is nonzero for a nonbounded frequency range. Nonband-limited signals pose problems in communications for two reasons. First, the characteristics of the transmission medium (and the transmitter and receiver components) usually result in some form of nonideal frequency response and bandwidth limitation for the signal — the signal has to be of limited bandwidth to be able to be transmitted without significant frequency content loss or distortion from the channel. The second reason is that in order to accommodate many channels in large-capacity trunks, most telecommunication standards impose some form of frequency-division multiplexing (FDM), whereas each individual channel has to satisfy a bandwidth constraint in order to be multiplexed and not interfere with adjacent channels.

For the remainder of this discussion we will assume that the channel characteristics are such that it is an ideal channel of bandwidth W, with

$$C(f) = \begin{cases} 1, & |f| \leq W \\ 0, & |f| > W \end{cases} \tag{6}$$

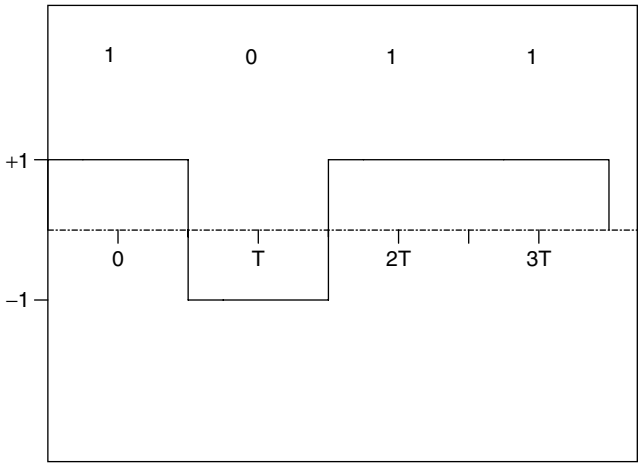


Figure 1. An example of a full-response signal.

Dealing with a channel that, besides the bandwidth limitations, also introduces amplitude or phase distortion to the transmitted signal is a subject of channel equalization.

Hence, the problem of designing full-response signals [i.e., signals that satisfy Eq. (5)] that also have a limited bandwidth W, having $G_T(f)G_R(f) = 0$ for $|f| \geq W$, arises. The pioneering work on this was done by Nyquist [2]. He established the condition (*Nyquist condition*) that $X(f)$ has to satisfy in order to have zero ISI in the received samples.

The major result of the Nyquist condition is that in order to transmit without ISI a signal with symbol rate $1/T$, there is a minimum bandwidth requirement of $W \geq 1/2T$. The resulting maximum symbol rate of $1/T = 2W$ is called *Nyquist rate*. It is interesting to note the duality with sampling theory, where it is known (again by Nyquist) that in order to uniquely sample a bandlimited signal, one has to use a sampling rate (frequency) of at least $2W$; a higher rate results in correlation (i.e., ISI, from a communications point of view) between successive samples.

A family of pulses that satisfy the Nyquist criterion for zero ISI is the one with a *raised cosine spectrum*. The bandwidth occupancy B of this pulse is determined by the rolloff factor α, which takes values $0 \leq \alpha \leq 1$, giving

$$B = \left(\frac{1}{2T}\right) \cdot (1 + \alpha) \tag{7}$$

An example of two pulses with raised-cosine spectrum, with rolloff factors $\alpha = 0, 0.5$ is shown in Fig. 2. We notice that the pulses have zero value at multiple integers of the symbol interval $t = nT, n \neq 0$, which is the non-ISI criterion of Eq. (5). Thus, superimposing shifted versions (by nT) of these pulses leads to a combined signal where only one constituent pulse contributes in the value of the signal at a specific sampling instant $t = mT$.

The limiting case of pulses with the raised-cosine spectrum is when the rolloff factor is $\alpha = 0$. Then the spectrum has an ideal rectangular characteristic with the smallest possible bandwidth $B = 1/2T$ for no ISI. The resulting pulse is the sinc( ) function

$$x(t) = \frac{\sin(\pi t/T)}{\pi t/T} = \text{sinc}\,\frac{t}{T} \tag{8}$$

However, filters that have such a sharp frequency response are practically nonrealizable. For the filter to be causal (i.e., to have its impulse response to the right of the $t = 0$ axis in Fig. 2a), the truncation of its impulse response must be of a reasonable length and a plus a delay function is required. Since the tails of the sinc( ) function decay quite slowly (proportionally to $1/t$) the truncation length has to be extremely large, to avoid the significant loss of the signal characteristics, and this makes the filter realization very difficult. Another detrimental effect of the heavy tails of the sinc( ) pulse is that a mistiming error (sampling at a time slightly off the multiples of T) results in a nonconverging (due to the $1/t$ decay) series of ISI components.

Hence, usually for a realizable full-response system a raised-cosine spectrum characteristic with $\alpha >$

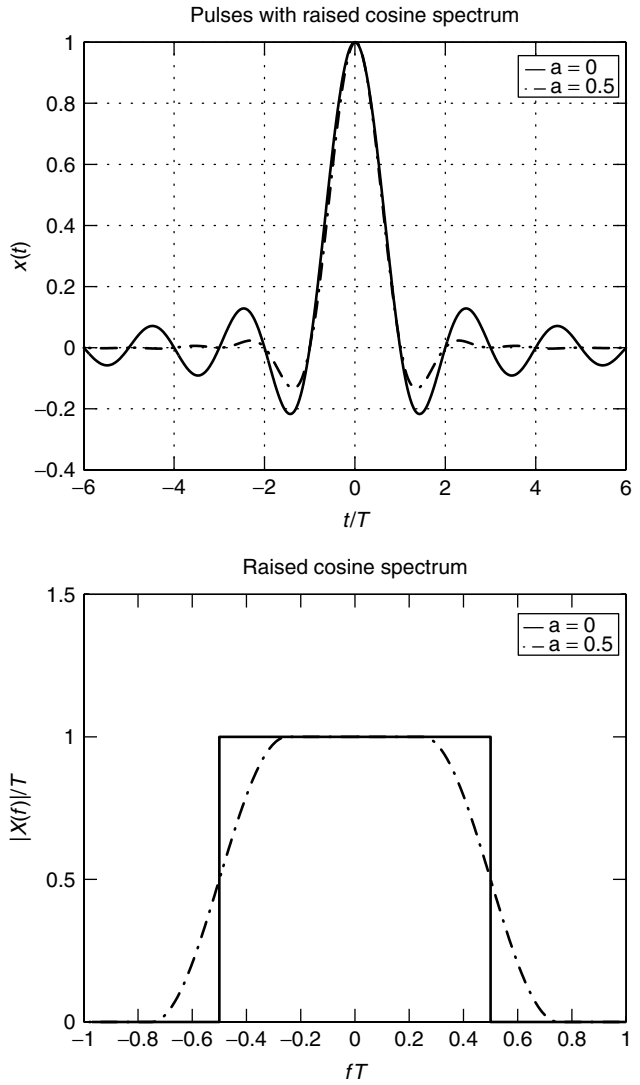


Figure 2. Pulses with raised-cosine spectrum: (**a**) $x(t)$ in the time domain; (**b**) frequency response $X(f)$.

0 is chosen. This leads to a filter with a more realistic truncation length requirement. However, this also results to a signal bandwidth occupancy that is larger than the optimum $(1/2T)$ one, typically by 15–25%.

2. SIMPLE PARTIAL-RESPONSE SIGNALS

In the previous section we noticed that the Nyquist limit on the transmission rate over band-limited channels $1/T = 2W$ is practically nonrealizable with full-response signaling. However, if one relaxes the zero-ISI condition and allows for a controlled amount of ISI, then one can obtain realizable filters that have the Nyquist bandwidth $W = 1/2T$. This was first observed by Lender [3] and later extended by Kretzmer [4], Kobayashi [5], and Pasupathy [6].

Let's examine the simplest case of a partial-response signal, one that has a composite response $x(n)$ [as given by Eq. (3)] with values[1]

$$x(n) = \begin{cases} 1, & n = 0 \\ 1, & n = 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

It is easily shown that for $W = 1/2T$ the actual corresponding pulse is

$$x(t) = \text{sinc}\left(\frac{t}{T}\right) + \text{sinc}\left(\frac{t}{T} - \frac{1}{T}\right) \tag{10}$$

with frequency response

$$X(f) = 2Te^{-j\pi fT}\cos(\pi fT),\ |f| < \frac{1}{2T}(0 \text{ otherwise}) \tag{11}$$

This is the first partial-response pulse that was examined and is called a *duobinary* pulse. We notice that it is equivalent to a digital FIR filter with coefficients [1 1] followed by a filter with an ideal rectangular frequency response [to which $\text{sinc}(t/T)$ corresponds]. We mentioned above that the ideal rectangular filter is not practically realizable on its own; however, the pulse given by Eq. (9) and shown in Fig. 3a is much more easily realizable since its tails decay rapidly and its frequency response (shown in Fig. 3b) decays smoothly toward zero.

Another simple partial-response scheme is the *modified duobinary pulse*, which is characterized by the following composite response:

$$x(n) = \begin{cases} 1, & n = -1 \\ -1, & n = 1 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

The modified duobinary pulse and its spectrum are given by the following relationships and are shown in Fig. 4; also

$$x(t) = \text{sinc}\left(\frac{t+T}{T}\right) - \text{sinc}\left(\frac{t-T}{T}\right) \tag{13}$$

$$X(f) = j \cdot 2T\sin(\pi f 2T),\ |f| < \frac{1}{2T}(0 \text{ otherwise}) \tag{14}$$

We notice that the modified duobinary spectrum has a null at DC ($f = 0$), which makes it suitable for channels that don't pass DC (e.g., circuits with transformer couplings) or for SSB modulation. Currently, the most significant application of the modified duobinary pulse is in magnetic recording systems. From Fig. 4a, we notice that the modified duobinary pulse is similar to the read-back signal response to a pulse in a magnetic recording system. On the basis of this observation, one can shape with minimal equalization the combined magnetic channel into a modified duobinary system response, and use maximum-likelihood decoding (explained in the next subsection) to estimate the data sequence. The use of these ideas in magnetic recording systems [7] (labeled as PRML, from

[1] One should notice that the given sample values result in a higher energy per bit with respect to a zero-ISI system with $x(0) = 1$, and this should be taken into account when one calculates the SNR of the signal and its bit error probability.

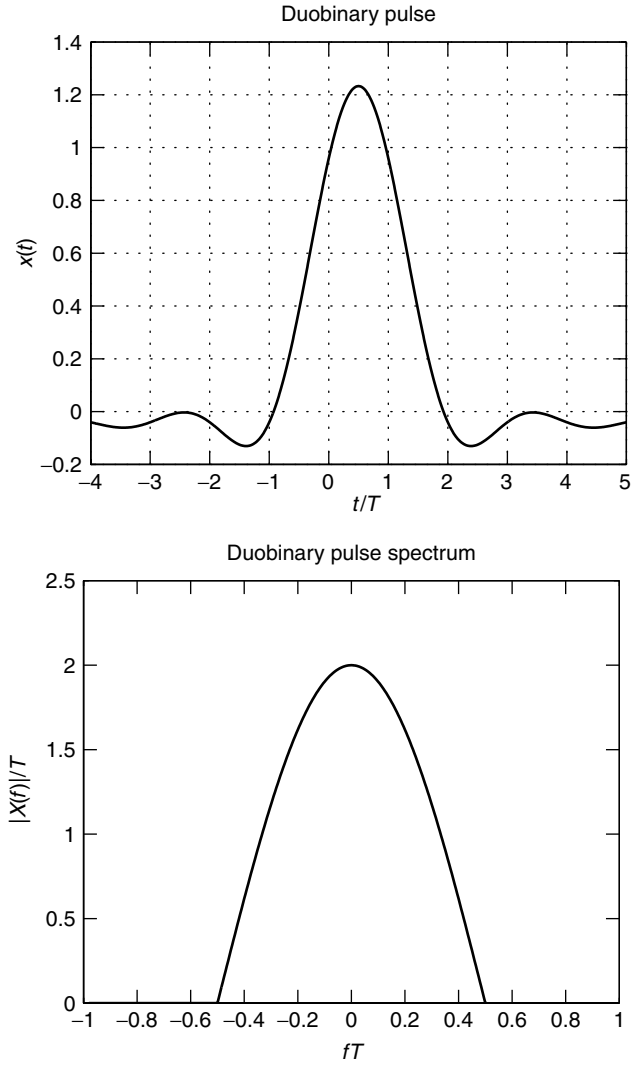


Figure 3. Duobinary pulse: (*a*) $x(t)$ in the time domain; (*b*) frequency response $X(f)$.

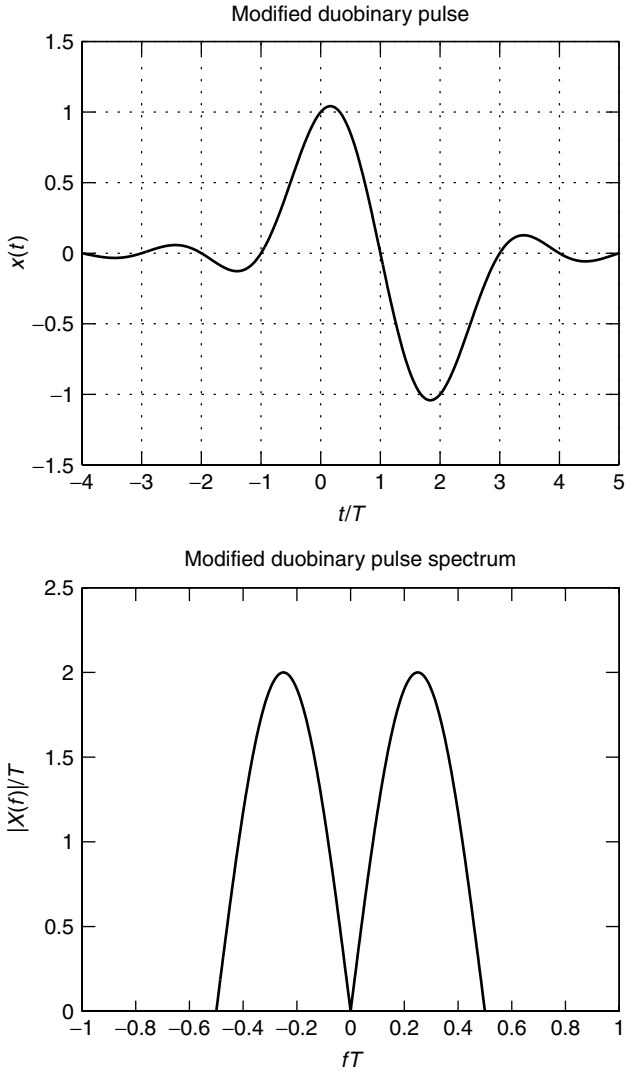


Figure 4. Modified duobinary pulse: (**a**) $x(t)$ in the time domain; (**b**) frequency response $X(f)$.

partial-response maximum likelihood) was one of the main reasons in the impressive increase of recording density, and hence capacity, of the hard disks in the last decade.

2.1. Detection of Partial-Response Signals

Let's examine the detection methods that may be employed for the simplest duobinary partial response signal. From Eqs. (4) and (9) we see that the output of the sampler, which is used as the decision variable, is

$$r_m = I_m + I_{m-1} + v_m$$

If I_{m-1} has already been detected with value I'_{m-1}, then its effect on r_m may be eliminated through subtraction, and hence I_m may be estimated from

$$I'_m = \text{sgn}(r_m - I'_{m-1})$$

where sgn() denotes the sign function, since in a binary scheme the detection rule for a signal with possible values $\{1, -1\}$ is $x > 0?1 : -1$, for zero-mean Gaussian noise. The detected value I'_m can then be used for the detection of I_{m+1} and so on. However, the use of this method can lead to serious error propagation: an error (because of the noise) in the estimate of a particular symbol will adversely affect the estimate of the next symbol and so on.

A technique that may be used to eliminate the dependence on the previous symbol estimate is *precoding*. The binary source sequence d_n is transformed to a new binary sequence p_n through the precoding operation

$$p_n = d_n \ominus p_{n-1} \tag{15}$$

where $\ominus$ denotes modulo-2 subtraction (equivalent to modulo-2 addition). Then the precoded sequence p_n is mapped into the information sequence, $I_n = 2p_n - 1$, that is used for transmission using the duobinary pulse. Using the relationship between d_n, p_n, I_n, and r_n, one can show

Table 1. Example of Precoding for Duobinary Signal

Binary sequence d_n		0	1	1	1	0	0	1	0
Precoded sequence p_n	1	1	0	1	0	0	0	1	1
Information sequence I_n	1	1	−1	1	−1	−1	−1	1	1
Duobinary signal r_n		2	0	0	0	−2	−2	0	2
Estimate d'_n		0	1	1	1	0	0	1	0

that the estimate of d'_m may be obtained as

$$d'_m = \frac{r_m}{2} - 1(\text{modulo } 2) \tag{16}$$

directly from r_m, without any use of the previous symbols. The (noiseless) example presented in Table 1 demonstrates this detection operation. We notice that, according to Eq. (16), values of $r_m = \pm 2$ (which occur with probability $\frac{1}{4}$ each) map to a binary 0 in the original binary source sequence, and the value $r_m = 0$ (which occurs with probability $\frac{1}{2}$) maps to a binary 1. In the case where noise affects the sample r_m the corresponding decision regions will be $|r_m| \geq 1$ and $-1 < r_m < 1$ respectively.

In a similar way, the modified duobinary signal can be detected on a symbol-by-symbol basis, without any dependence on the previous decisions, and hence without any error propagation, by using the precoding operation of

$$p_n = d_n \oplus p_{n-2}$$

The abovementioned precoding techniques may be extended directly from a binary to an M-ary ($M = 2^k$) modulation scheme, by using modulo-M instead of modulo-2 operations.

It can be shown [8] that the performance of a duobinary or modified duobinary scheme that employs a symbol-by-symbol detector is approximately 2.1 dB worse than the performance of a full-response scheme with the same constellation size. However, the symbol-by-symbol detector does not exploit a significant property of the partial-response signal: its memory. As an example, we notice that two consecutive received samples of a valid duobinary signal cannot have the values $2, -2$ or $-2, 2$. However, the symbol-by-symbol detector will not factor this in and will decode the preceding received samples according to the rule of Eq. (16), which will lead to at least one binary error.

There is an estimation method that does factor in the memory present in communication signals and gives the sequence of symbols with the minimum probability of error. This is the maximum-likelihood sequence estimator (MLSE), which can be implemented efficiently using the Viterbi algorithm [9]. This algorithm was first proposed for the decoding of convolutional codes, where the output depends on a finite number of previous inputs, and its use was later extended for systems with finite ISI, either unintentional (channel distortion) or intentional (partial response signals). It uses the squared distance[2] between the received samples and the possible valid sequences, and declares as detected sequence the one having the smallest total squared distance from the received one.

It can be shown that a duobinary or modified duobinary signal with MLSE detection at the receiver has the same performance as a full-response signal. So the MLSE detector does not have the 2.1-dB performance degradation that characterizes the symbol-by-symbol detector. As a side note, we should mention that there were research efforts after the introduction of partial response signaling to evaluate the free Euclidean distance of a *faster-than-Nyquist* scheme [10]. This name has been used for transmission schemes that employ a sinc( ) pulse with zero crossings at $1/T$ (and a bandwidth of $W = 1/2T$), but which are spaced at $T' < T$ (i.e., have a higher baud rate than does the pulse designed for zero ISI). This closer spacing of the sinc pulses leads to ISI, since the sample points for successive symbols no longer correspond to the zero crossing of the pulse. However, it was proved that the minimum Euclidean distance of such a scheme is the same as that of a full-response (no-ISI) system, as long as $T' > 0.802 \cdot T$. Since the use of the ideal rectangular spectrum is not practically realizable and the receiver will have to deal with an infinite series of ISI terms in order to estimate the received sequence, these ideas were just a form of mathematical exercise. However, the underlying principle of the existence of PR schemes that have bandwidth less than the Nyquist rate with a free distance that is the same as a full-response system led to additional research for bandwidth-efficient implementable PR techniques.

3. PARTIAL RESPONSE FOR BANDWIDTH EFFICIENCY

The duobinary and modified duobinary systems that we outlined above were designed with the goal of achieving the Nyquist rate of $W = 1/2T$, with practically implementable filters. Still, one can achieve even better bandwidth efficiency by using a partial response scheme. Since partial response can be viewed as just one form of filtering, one may use a longer filter $x(n)$ to achieve better bandwidth characteristics.

The disadvantage of this approach is the increased intersymbol interference between symbols, due to the longer $x(n)$ that spans $L + 1$ successive samples, where by L we denote the memory of the partial-response scheme.

As noted in a previous section, for received sequences coming from a channel/system with finite memory, the MLSE detector is the optimum one. We noted that the major determining factor for its performance is the minimum Euclidean distance $d_{\min}$ between any two valid sequences.

It can be shown [8] that any partial response system has minimum distance $d_{\min}$ less or equal to a full-response system of the same constellation size and transmitted energy level. Hence, the goal of the partial response scheme is to lose as little performance as possible, while maximizing the bandwidth efficiency. Said and Anderson [8] offered an optimization framework for finding the best partial-response scheme for a certain memory length L; given a constraint on the bandwidth B,

[2] For the case of linear modulation in AWGN the likelihood of a received sample r depends on its squared distance $|r - s|^2$ from the nominal value s to be received.

the partial-response scheme $x(n)$ with the best (largest) $d_{\min}$ would be found; or, given a certain $d_{\min}$ (i.e., a given performance level) the partial-response scheme $x(n)$ with the best bandwidth characteristics B would be found.

Besides the potential loss in $d_{\min}$, there is a second disadvantage associated with such a partial-response scheme, namely, the complexity of its receiver. An MLSE, implemented through a Viterbi algorithm (VA), has complexity quite bigger than the slice-and-quantize operation of a full-response receiver for an M-ary PAM. The MLSE complexity grows proportionally to the number of states in the finite memory system to which the PR scheme is equivalent. The number of states is equal to 2^{ML}, where L is the memory of the PR scheme and M is the constellation size. So for large constellation sizes (where the advantages of a smaller-bandwidth scheme are more easily exhibited), or for longer filters $x(n)$ (which give better bandwidth efficiency), the complexity issue makes implementation problematic. To overcome this disadvantage, reduced-complexity schemes, such as *reduced-state sequence estimation* (RSSE) [11,12], were proposed. These schemes were originally proposed for the traditional channel ISI problem, of which partial response is a special case. It was shown that with careful design and state reduction choices these schemes can offer performance very close to MLSE with significant computational cost savings.

Using these ideas, practical PR schemes were found [13], with bandwidth occupancy $B \approx 0.35 \cdot 1/T$ (roughly half of the bandwidth of a realizable full-response system), which exhibits a performance penalty of $<2-3$ dB compared to an equivalent (same constellation size) FR scheme. However, because of its much better bandwidth characteristics, the symbol (baud) rate of the PR scheme can be increased w.r.t. (with respect to) to the FR scheme, while still satisfying spectral occupancy constraints. If the bandwidth of the full-response scheme is $B_{\mathrm{FR}} = 0.7 \cdot 1/T_{\mathrm{FR}} = 1.4 \cdot 1/2T_{\mathrm{FR}}$ (raised cosine with 40% rolloff), and the bandwidth of the PR scheme is $B_{\mathrm{FR}} = 0.35 \cdot 1/T_{\mathrm{PR}}$, then the symbol rate of the PR scheme can be increased to twice that of the full response scheme: $1/T_{\mathrm{PR}} = 2 \cdot 1/T_{\mathrm{FR}}$. This will allow the use of PR scheme with a constellation size of $M^{1/2}$, if the full-response scheme is M-ary, which for moderate to large M more than compensates — due to the increased distance, and hence noise immunity, between energy levels — for the performance penalty paid for the ISI between symbols.

4. LINE CODING: MODULATION SIGNALS WITH MEMORY

A more generalized form of partial-response signaling is employed in modulation systems that are used mainly for high-speed baseband transmission, and their purpose is to shape the spectrum of the transmitted signal to match the spectral characteristics of the channel. The way to shape the spectrum is typically through the introduction of restrictions, namely, memory, on the generator pulse. These techniques are usually called *transmission line coding* or *modulation coding*. We should make the distinction here between the above type of coding, where correlation between transmitted symbols is introduced, and the more usual notion of coding that involves an increase in the bit rate between the source data sequence and the line sequence. In the latter case, the code typically encodes k source bits into n transmitted bits (code rate $R = k/n < 1$), and the receiver use this redundancy to detect and correct errors. These are called *channel coding* or *error correction* techniques, and they are different from modulation coding techniques.

Focusing on transmission line codes, we depict in Fig. 5, some of the most commonly employed ones.

The first, and probably simplest, modulation scheme for baseband transmission is the *non-return-to-zero* (NRZ) method. In NRZ a binary $d_k = 1$ is mapped into a square waveform of amplitude +A, and a binary $d_k = 0$ is mapped into a square waveform of amplitude −A. Actually, this scheme is a memoryless system, equivalent to binary PAM transmission. It does not pose any demodulation problems, but it is not suitable for spectrum shaping because of its lack of correlation between symbols. Furthermore, it does not offer another desirable property for baseband transmission — it does not guarantee a minimum rate for pulse transitions, since a sequence of multiple 1s (or 0s) leads to a constant voltage signal throughout its duration. Pulse transitions are necessary for deriving timing and synchronization from the received signal at the receiver.

Numerous slight variations of the NRZ scheme exist. The system that is described above is also called by some authors a *(bi)polar NRZ scheme*, to distinguish it from a unipolar NRZ scheme, where the two amplitude levels are $[+A, 0]$ instead of $[+A, -A]$ of bipolar NRZ. The unipolar NRZ scheme has a more pronounced DC component than the bipolar one, and is used less in practise.

Another memoryless system is the *return-to-zero* (RZ) method. In this system, the basic square pulse (of amplitude $[+A, -A]$, for binary $d_k = 1$, 0 respectively) returns to 0 for the second half of the symbol duration, so it has a 50% duty cycle. The return to 0 guarantees voltage transitions at a rate of $2/T$, thus eliminating

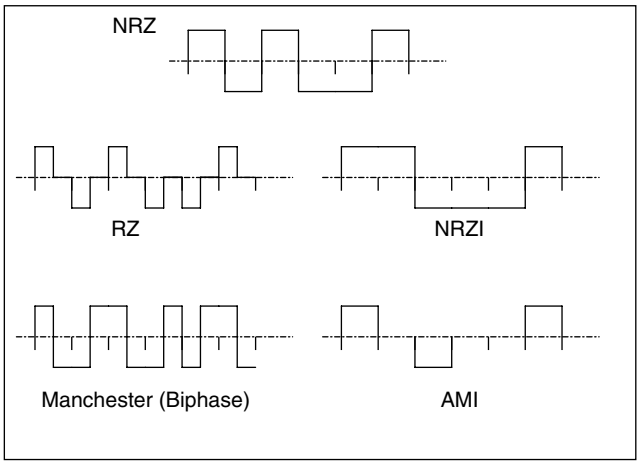


Figure 5. Examples of line coding techniques: NRZ, RZ, Manchester, NRZI, AMI (all mapping the binary information sequence [1, 0, 1, 0, 0, 1]).

the timing/synchronization problems that an NRZ scheme might have. However, its narrower pulses occupy a larger bandwidth, and this is a serious disadvantage for an RZ scheme.

Similar to the RZ scheme is biphase (or Manchester) encoding, which employs a square wave of period T that transitions from $+$A to $-$A (instead of to 0, like the RZ scheme) at the middle of the symbol interval. The abovementioned pulse is used for a binary $d_k = 1$, while its antipodal pulse is used for 0. Although relatively bandwidth-inefficient, the Manchester scheme is used in some practical systems, including some Ethernet and TTY applications, because of its lack of DC, and timing information provision.

A line code that employs memory in its waveform generation is the *non-return-to-zero, inverse* (NRZI) scheme. This employs the same basic pulse as the NRZ scheme, but with the transitions from one amplitude level to the other happening when the information bit is $d_k = 1$, and the previous symbol level is retained when $d_k = 0$. NRZI can be directly generated from NRZ, if the actual data sequence is passed through a precoder of the form $p_k = p_{k-1} \oplus d_k$, and then p_k is used to generate the transmitted waveform through NRZ. NRZI is used in magnetic recording systems, since it has good bandwidth characteristics and no DC component. However, it presents the usual problem of no guaranteed timing information, if many consecutive 0's appear in the data sequence.

Finally, another line coding technique with memory is *alternate mark inversion* (AMI). An AMI waveform employs zero voltage/amplitude for a binary 0, while it employs a square pulse of alternating amplitude for binary 1s, so if the previous 1 bit were sent using $-$A voltage, the next 1 bit would be sent using a pulse with $+$A voltage. AMI has good bandwidth characteristics and no DC component, and its memory offers some elementary error detection capabilities, since two consecutive pulses of the same sign (with any number of zero-voltage symbols in between them) mean that an error has occurred. Although AMI might present the receiver with timing/synchronization problems when multiple consecutive 0s exist in the transmitted sequence, it is a widely employed scheme, especially in T1/T3 trunklines.

The problem of the transmitted waveform being constant for many symbols, if multiple consecutive 0s are being transmitted, that practical NRZI and AMI schemes face is addressed in two ways.

1. The first one is to introduce codes that operate on the binary source sequence and generate a channel sequence with a restriction in the maximum number of successive zeros. Typically this technique is used in conjunction with NRZI in magnetic recording systems. These codes usually impose an additional restriction in the minimum number of 0s between two 1s in a sequence, and this is used to increase the distance (and, hence, reduce interference) between pulse transitions. These codes are known as *runlength-limited codes* (RLL codes), and have a code rate of less than 1; that is, they introduce redundancy. A good tutorial paper on them has been written by Immink [22].
2. The second technique to address the consecutive 0s problem is used in conjunction with AMI. We have seen that AMI offers some basic error detection, since two successive (ignoring intermediate 0s) pulses of the same polarity do not correspond to a valid input sequence. The idea for the second technique is to replace a string of consecutive zeros with a string that violates the AMI principle (and contains pulse transitions). The receiver will detect the "error" condition, and, since the received string has a predetermined pattern, it will substitute the original string of all zeros in its place. This way both sufficient timing information is available, and the original data are not lost.

The class of these codes is usually denoted as *binary bipolar with n zero substitution* (BnZS). A common one is B8ZS: a pattern of 8 zeros is replaced by a string where the bits 4 and 7 violate the bipolar principle. Let's take as an example the string $+(00000000)0-\cdots$, where $+$ stands for amplitude A, $-$ stands for amplitude $-$A (both correspond to 1s in the information sequence), and 0 is zero amplitude (0s in the information sequence). This string is a valid bipolar sequence, since the two consecutive nonzero pulses have opposite polarity. The B8ZS code replaces the constant string of 8 zeros (inside parentheses) by the string with the two bipolar violations (BPV) resulting in the transmitted string of $+(000+-0-+)0-\cdots$. The receiver, on detecting the two BPVs at the specified locations, replaces the 8 bits with all zeros, and the original sequence is restored. If the most recent pulse sign before the 8-zero string is negative, then the replacement string will be $(000-+0+-)$ in order to give the two BPVs at bits 4 and 7. The B8ZS scheme is employed by many commercial carriers. A similar, but simpler, scheme is the B6ZS code, where 6 consecutive 0s are replaced by a string with bipolar violation at the second and fifth bits (i.e., the B8ZS scheme, without the two first initial zero bits).

Another two members of the BnZS family are the B3ZS and B4ZS codes, which are very similar. Let's examine B4ZS [which is widely used, and is also called *high-density bipolar 3* (HDB3)]. Here, a string of four zeros is replaced by the string B00V or 000V, where B stands for a pulse that satisfies the bipolar rule, while V is a pulse violating the rule. The choice between these two strings is made such that the number of pulses satisfying the bipolar rule between violations is odd. As an example, suppose that we have the string $+(0000)0-0+\cdots$, and the number of valid pulses after the last (not shown) violation is even. Then, the marked four 0s will be replaced by B00V, because the first bit (valid pulse) will make for an odd number of valid pulses between the last violation and the current one. Hence, the transmitted string will be $+(-00-)0+0-\cdots$. Note that, contrary to B8ZS, the B4ZS code does affect the polarity of the subsequent nonzero pulses. The B3ZS scheme is the same as B4ZS, but with three consecutive zeros replaced by either B0V or 00V patterns (again, with an odd number of valid B pulses between successive V violations).

5. NOTES ON THE LITERATURE

Partial-response systems were first proposed by Lender [3], and later extended by Kretzmer [4], Pasupathy [6], and others. Faster-than-Nyquist signaling was investigated by Mazo [15], Foschini [10], and Hajela [16].

The combination of partial response with TCM encoding was investigated by Wolf and Ungerboeck [17], Ketchum [18], and Forney and Calderbank [19].

The use of partial-response shaping in magnetic recording systems is covered in various papers. As an example we mention the papers by Cideciyan et al. [7], and Tyner and Proakis [20], where additional references in the evolution of PRML systems can be found. Typically nowadays, the target response in a PRML system is a *generalized* partial-response scheme, which is the product of the modified duobinary with a system of the form $(1+D)^n$. A tutorial paper on the general issue of coding for magnetic recording channels was written by Immink et al. [21].

A paper that surveys the use of modulation coding in copper wire subscriber lines was published by Lechleider [22], while a good reference for BnZS and other currently employed line coding techniques is the one by Bellamy [23].

The correlation between a PR system and precoding techniques to combat ISI in a decision feedback equalization scheme can be found in the paper by Forney and Eyuboglu [24] and references cited therein. A paper that covers the state of partial response signaling is the one by Said and Anderson [8], while a typical implementation of a PR system for a copper cable channel is given in [25].

BIOGRAPHY

Apostolos D. Rizos received the B.S. degree in physics from the University of Athens, Athens, Greece, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Northeastern University, Boston. Since 1999, he has been with AWARE Inc., Bedford, Massachusetts, where he is a Senior DSP Engineer, working on algorithms for DSL modems. Before that, he was a technical consultant with Delphi Communication Systems, Maynard, Massachusetts, working on underwater acoustic modems. His interests lie in modulation and coding for communication systems, and computationally efficient algorithm implementations.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
2. H. Nyquist, Certain topics in telegraph transmission theory, *AIEE Trans.* **47**: 617–644 (1928).
3. A. Lender, The duobinary technique for high-speed data transmission, *IEEE Trans Commun. Electron.* **82**: 214–218 (May 1963).
4. E. R. Kretzmer, Generalization of a technique for binary data communication, *IEEE Trans. Commun. Technol.* **14**: 67–68 (Feb. 1966).
5. H. Kobayashi, Correlative level coding and maximum likelihood decoding, *IEEE Trans. Inform. Theory* **17**: 586–594 (Sept. 1971).
6. S. Pasupathy, Correlative coding: A bandwidth-efficient signaling scheme, *IEEE Commun. Soc. Mag.* **15**: 4–11 (July 1977).
7. R. D. Cideciyan et al., A PRML system for digital magnetic recording, *IEEE J. Select. Areas Commun.* **10**: 38–55 (Jan. 1992).
8. A. Said and J. B. Anderson, Bandwidth-efficient coded modulation with optimized linear partial-response signals, *IEEE Trans. Inform. Theory* **44**: 701–713 (March 1998).
9. G. D. Forney Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 363–378 (May 1972).
10. G. D. Foschini, Contrasting performance of faster binary signaling with QAM, *AT&T Bell Labs Tech. J.* **63**: 1419–1445 (Oct. 1984).
11. V. Eyuboglu and S. Qureshi, Reduced-state sequence estimation with set partitioning and decision-feedback, *IEEE Trans. Commun.* **36**: 13–20 (Jan. 1988).
12. P. Chevillat and E. Eleftheriou, Decoding of trellis-encoded signals in the presence of intersymbol interference and noise, *IEEE Trans. Commun.* **37**: 669–676 (July 1989).
13. A. D. Rizos and J. G. Proakis, Reduced-complexity sequence detection approaches for PR-shaped, coded linear modulations, *Proc. IEEE GLOBECOM'97*, Phoenix, AZ Nov. 1997.
14. K. A. S. Immink, Runlength-limited codes, *Proc. IEEE* **78**: 1745–1759 (Nov. 1990).
15. J. E. Majo and H. J. Landau, On the minimum distance problem for faster-than-Nyquist signaling, *IEEE Trans. Inform. Theory* **34**: 1420–1427 (Nov. 1988).
16. D. Hajela, On computing the minimum distance for faster than Nyquist signaling, *IEEE Trans. Inform. Theory* **36**: 289–295 (March 1990).
17. J. Wolf and G. Ungerboeck, Trellis coding for partial-response channels, *IEEE Trans. Commun.* **34**: 765–772 (Aug. 1986).
18. J. Ketchum, Performance of trellis-codes for M-ary partial-response, *Proc. IEEE GLOBECOM'87*, 1987, pp. 1720–1724.
19. G. D. Forney Jr. and A. R. Calderbank, Coset codes for partial response channels; or, coset codes with spectral nulls, *IEEE Trans. Inform. Theory* **35**: 925–943 (Sept. 1989).
20. D. J. Tyner and J. G. Proakis, Partial response equalizer performance in digital magnetic recording channels, *IEEE Trans. Magn.* **29**: 4194–4208 (Nov. 1993).
21. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorder, *IEEE Trans. Inform. Theory* **44**: 2260–2299 (Oct. 1998).
22. J. W. Lechleider, Line codes for digital subscriber lines, *IEEE Commun. Mag.* **27**: 25–32 (Sept. 1989).
23. J. C. Bellamy, *Digital Telephony*, 3rd ed., Wiley, New York, 2000.
24. G. D. Forney Jr. and V. Eyuboglu, Combined equalization and coding using precoding, *IEEE Commun. Mag.* **29**: 25–34 (Dec. 1991).
25. G. Cherubini, S. Olcer, and G. Ungerboeck, A quaternary partial-response class-IV transceiver for 125 Mbit/s data transmission over unshielded twisted-pair cables: principles of operation and VLSI realization, *IEEE J. Select. Areas Commun.* **13**: 1656–1669 (Dec. 1995).

PATH LOSS PREDICTION MODELS IN CELLULAR COMMUNICATION CHANNELS

H. L. BERTONI
Polytechnic University
Brooklyn, New York

1. INTRODUCTION

Modern wireless services involve two-way transmission of individual signals, rather than one-way broadcast of the same signal to many listeners. In order to accommodate many users for such services in an allocated frequency band, the concept of spatial frequency reuse was developed. The simplest implementation of frequency reuse involves spatial separation of the simultaneous use of the same frequency channel linking subscribers with their nearest access point (or base station). By keeping the interference just low enough to achieve a desired quality of service, the subscribers can be accommodated with a minimum of physical infrastructure. In other words, using the smallest distance possible between cochannel cells allows for the most channels per base station. The conflicting requirements of achieving radio coverage, while maintaining the desired limit to interference, places a premium on accurate methods for prediction of the received radio signal strength and other channel characteristics. A survey of many of the characteristics can be found in Refs. 1 and 2.

One approach to understanding the channel characteristics is to make measurements over a wide range of system parameters, such as frequency and bandwidth, antenna height, and distance between antennas [3–5]. The measurement results can be reduced to best-fit formulas that give the system parameter dependence in a way that is simple to use in prediction software [6–8]. Because most subscribers live in cities, it is important to measure these quantities in different building environments. An alternative approach is to use theoretical methods for predicting radiowave propagation; the most common are ray optics and the uniform theory of diffraction (UTD). In this article we discuss only the theoretical methods for predicting the path loss (or path gain), and compare them with measurement-based models. The prediction of other channel characteristics can be found elsewhere [1,9].

Path gain PG is defined as the ratio of the received signal strength to the total radiated power. The commonly used term "path loss" refers to the reciprocal of path gain, and is usually expressed in decibels. As an example, for antennas located in free space the path gain is given by

$$PG_0 = \left(\frac{\lambda}{4\pi R}\right)^2 g_1 g_2 \tag{1}$$

where R is the separation between antennas, λ is the wavelength, and g_1 and g_2 are the directive antenna gains. It is often convenient to consider the path gain between idealized isotropic antennas for which $g_1, g_2 = 1$. Provided that the carrier frequency is the same, reciprocity of Maxwell's equations implies that the path gain is the same, no matter which end of the wireless link is the transmitter and which is the receiver.

2. RAY CONCEPTS FOR UNDERSTANDING AND PREDICTING PROPAGATION

Modern wireless systems use UHF (300 MHz–3 GHz) and microwave radiolinks to connect base station antennas and subscribers located between the buildings, or even inside buildings. Thus the buildings have a major influence on the received signal. Since the wavelength λ at these frequencies is small compared to the building size, it is appropriate to think of the radiowaves as traveling along rays radiated in all directions by the source. The rays travel along straight lines until they encounter the buildings or ground where they are reflected according to the laws of geometric optics (GO). A ray incident on a building edge or corner creates a family of diffracted rays propagating away from the edge in a cone whose half-angle is equal to the angle between the incident ray and the edge, as described by the uniform theory of diffraction (UTD) [10]. The influence of each interaction on the ray fields is contained in a reflection or diffraction coefficient, together with an algebraic dependence on the length of the ray segments that conserves energy within a tube of neighboring rays. Subsequent ray encounters with buildings act in cascade. Thus, rays connecting base station antenna and subscriber may be multiply reflected by the building walls, and/or diffracted at the corners and rooftops, as suggested in Fig. 1 for propagation from the base station antenna to the subscriber.

The ray segments reaching the subscriber come from all directions in the horizontal plane, and a wedge of angles in the vertical direction. Reciprocity of Maxwell's equations implies that the same ray paths, with arrows reversed, apply for propagation from the subscriber to the base station antenna. For an elevated base station, as shown in Fig. 1, the arriving rays come from limited wedge of angles. However, base station antennas located below the rooftops will receive rays coming from all directions in the horizontal plane.

Each ray carries a copy of the transmitted symbol bit $p(t)$ that is delayed by the path length L_j divided by the speed of light c, and has complex amplitude A_j. The received voltage at location x along a street is the sum of

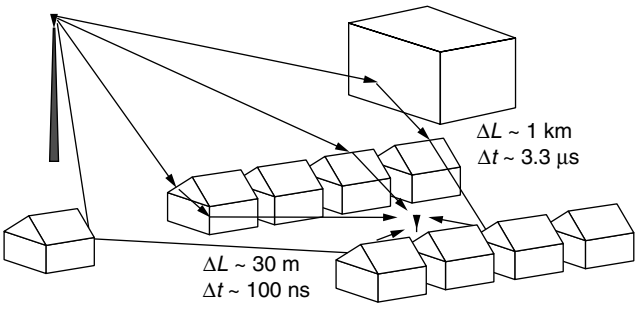


Figure 1. Multiple ray paths by which signals propagate from an elevated base station to a subscriber at street level. (©2002 by H. L. Bertoni.)

the ray contributions

$$V(x)e^{j\omega t} = \sum_j A_j p(f - L_j/c)e^{j\omega(t-L_j/c)} \qquad (2)$$

where c is the speed of light and $\omega = 2\pi f$. Because the differences in path length ΔL of the various rays reflected from objects near to the mobile are on the order of the street width, which is on the order of 30 m, the time difference for this cluster of rays is on the order of 100 ns. In addition, rays from the base station may be reflected from large structures at a greater distance before arriving at the building in the vicinity of the subscriber. Such paths may have delays on the order of 1 μs, followed by additional delays due to scattering in the vicinity of the subscriber, which results in another cluster of rays arriving at the subscriber. Figure 2 is an example of the time-delay profile of the signal received during wideband pulsed measurements made in Paris at 890 MHz with an omnidirectional receiving antenna and a system time resolution of 0.1 μs [11]. Echos with significant amplitude are received with time delays up to about 5 μs.

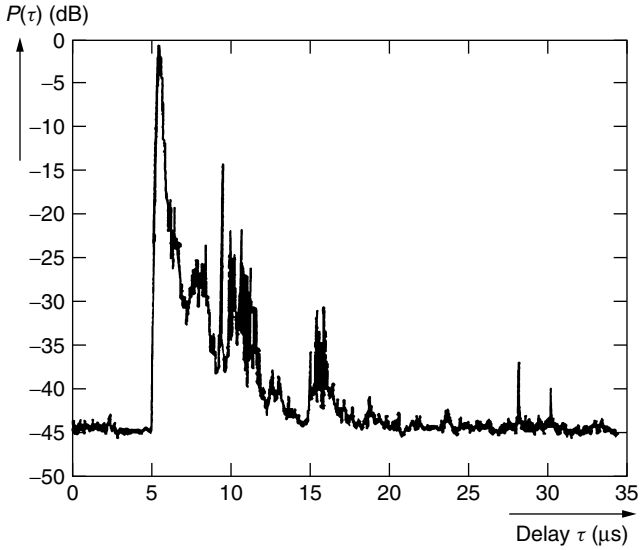


Figure 2. Time-delay profile of a pulsed signal in the 890-MHz band measured in Paris showing significant arrivals at up to 5 μs. (From Ref. 11, with permission.)

For systems whose bandwidth BW is small enough so that individual ray signals overlap in time ($1/BW \gg \Delta L/c$), all terms in (2) can be approximated using $p(t - L_j/c) \approx p(t - R/c)$, where R is the distance from the base station. In this case the rays arriving in all directions at the subscriber add coherently and generate a standing wave or interference pattern in space. The rapidly varying curve in Fig. 3 is a plot of the total received power as a function of the position x of a vehicle that is non-line-of-sight (NLoS) of the base station [12]. The amplitude is seen to undergo variations of up to 20 dB over distances on the order of one half the wavelength λ, which at 910 MHz is about $\frac{1}{3}$ m. In a moving vehicle, this spatial variation is perceived as a rapid time variation, which has led to the term "fast fading." Taking a sliding average of the received power over a distance of about 10–20λ smooths out the rapid fluctuation, as shown in Fig. 3, and is known as the small-area average.

The small average in Fig. 3 is seen to vary by about ±6 dB from the overall average, and variation has a scale length of 5–10 m. This variation is often referred to as "slow fading" or "shadow loss" since it results from shadowing by buildings, trees, and other objects. Finally, for outdoor or indoor links, the signal amplitude shows a systematic dependence on the distance or range R between base station and subscriber. Typically, measurements of the small-area average are plotted in dB against $\log R$, and a linear regression line is fit to the measurements. The regression fit gives the range dependence, as discussed in a later section, while the deviation from the regression line is interpreted as the shadow fading.

The sliding average of power is proportional to the spatial average $\langle |V(x)|^2 \rangle$ of the voltage in Eq. (2). Since the ray amplitudes A_j are slowly varying functions of distance x, and since the pulse duration ($1/BW$) is typically long compared to the time difference over the averaging interval of about $20\lambda/c = 20/f$, the sliding average is therefore

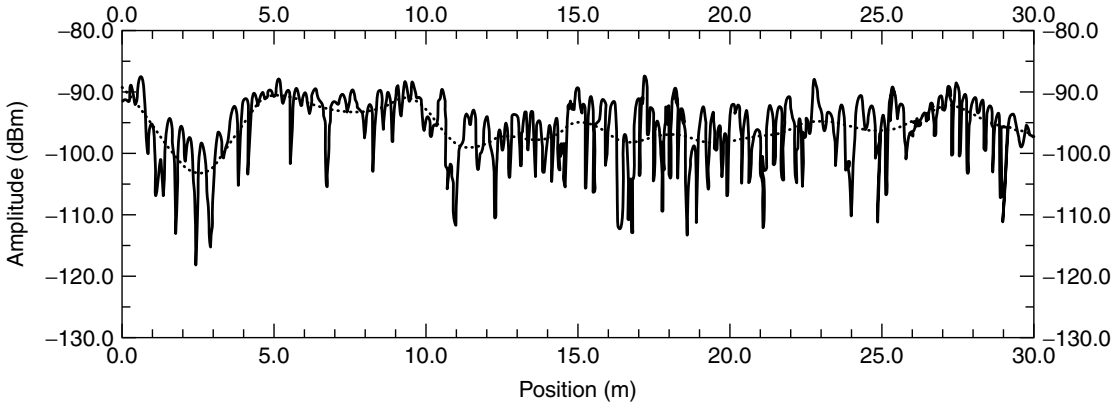


Figure 3. Narrowband [continuous-wave (CW)] measured signal variation, and the sliding average, as a function of distance along a street for non-LOS conditions. (From Ref. 12, with permission.)

given by

$$\langle |V(x)|^2 \rangle = \sum_{j,k} A_j A_k^* p(t - L_j/c) p^*(t - L_k/c) \langle e^{-j\omega(L_j - L_k)/c} \rangle$$
$$\approx \sum_j |A_j|^2 |p(t - L_j/c)|^2 \quad (3)$$

The final approximation in this equation is obtained by recognizing that the phase differences $\omega(L_j - L_k)/c$ for $j \neq k$ go through 2π variations due to changes in the pathlengths L_j with distance x along the street. As a result, the spatial average of the exponential vanishes for $j \neq k$. Thus the spatial average power is equal to the sum of the ray powers. For each symbol bit, the total received energy is found by integrating (3) over time, and is seen to be proportional to $\sum |A_j|^2$. Thus the total bit energy is proportional to the sum of the ray powers. Again for narrow band systems $|p(t - L_j/c)|^2 \approx |p(t - R/c)|^2$ and the time variation can be taken outside of the summation in (3).

3. TWO-RAY MODEL FOR FLAT EARTH

The simplest propagation environment occurs when there is only flat earth between the base station and the subscriber. In this case the received signal can be computed from the two-ray model consisting of a direct ray and a ground-reflected ray, as shown in Fig. 4. Because the two ray paths are of nearly equal length, it is necessary to add the ray fields coherently, and not simply add the ray powers [13]. For isotropic antennas the path gain of the two-ray model is given by [13]

$$PG = \left(\frac{\lambda}{4\pi}\right)^2 \left| \frac{e^{-jkr_1}}{r_1} + \Gamma(\theta) \frac{e^{-jkr_2}}{r_2} \right|^2 \quad (4)$$

where r_1 is the direct distance from the transmitter to the receiver, r_2 is the distance through reflection point. The Fresnel reflection coefficient $\Gamma(\theta)$ depends on the angle of incidence θ and the polarization, and is given by

$$\Gamma(\theta) = \frac{\cos\theta - a\sqrt{\varepsilon_r - \sin^2(\theta)}}{\cos\theta + a\sqrt{\varepsilon_r - \sin^2(\theta)}} \quad (5)$$

Here $a = 1/\varepsilon_r$ for vertical polarization and $a = 1$ for horizontal polarization, where ε_r is the relative dielectric constant of the ground. For average ground, the relative dielectric constant is $\varepsilon_r = 15 - i60\sigma\lambda$, and the conductivity σ is around 0.005 mho/m [13]. As the distance between the transmitter and receiver increases, the angle θ approaches 90°, the reflection coefficient Γ approaches -1 and r_2 approaches r_1.

Measured path loss between vertically polarized dipoles and between vertically polarized bicones is shown in Fig. 5 when the antennas are located along a flat road whose only features are low vegetation and wooden telephone poles [13]. For comparison, the dashed curve in Fig. 5 is a plot of the signal predicted by (4) for vertical polarization. For small horizontal separation $R < 10$ m, the antenna patterns have an influence on the measurements. However, for $R > 10$ m, only the antenna gains are important and they result in a vertical offset of the curves (the signal for dipoles is few decibels greater than for isotropic antennas, and for bicones it is a few decibels smaller).

Using a logarithmic scale for the horizontal separation R, as in Fig. 5, the received power is seen to vary about straight lines having two distinct slopes separated by a breakpoint R_B. Before the R_B, the radio signal oscillates severely as a result of alternating regions of destructive and constructive combination of the two rays, while after the R_B it decreases more rapidly with distance. The breakpoint lies near the last peak in the two-ray model, at a distance $R_B = 4h_1h_2/\lambda$, where h_1, h_2 are the antenna heights [1]. Well beyond the breakpoint distance the path gain of (4) reduces to

$$PG = \frac{h_1^2 h_2^2}{4R^4} \quad (6)$$

so that the received signal decreases more rapidly than the $1/R^2$ dependence of free space.

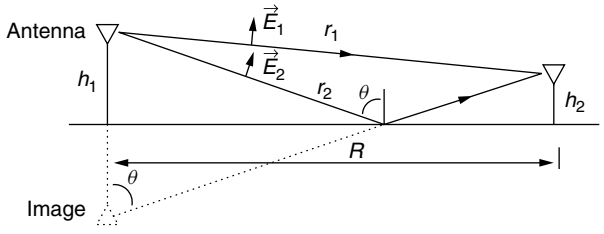


Figure 4. Propagation over flat earth as described by the two ray model. The ground reflected ray appears to come from the image of the source. For large R, the fields of both rays are nearly parallel.

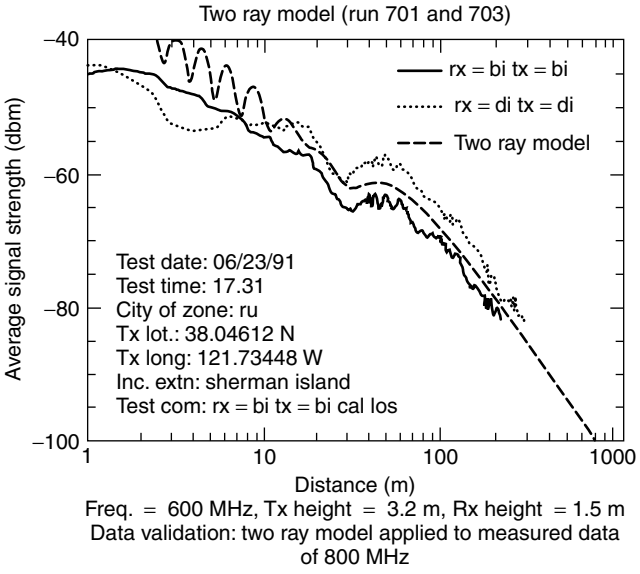


Figure 5. Measured and predicted variations of the received signal for propagation over flat earth for antennas of height 3.2 and 1.6 m at 800 MHz. (From Ref. 13, with permission.)

The two-ray model also serves as a basis for understanding the received signal when the two antennas are located within a line of sight along a street in an urban environment. In this case the direct and ground reflected rays give the dominant contributions, while additional contributions come from rays that are reflected by the buildings lining the streets. Building-reflected rays result in additional rapid variations about the simple two-ray model, but do not change the overall variation. Accounting for a single reflection in the building walls and ground reflection leads to the six-ray model, which has been used to obtain the plot of Fig. 6. Similar results are obtained from measurements in urban environments [1,14].

4. PROPAGATION OVER BUILDINGS IN RESIDENTIAL ENVIRONMENTS

In residential sections of cities, and in suburban regions, the buildings are of more or less uniform height, and the propagation may take place past many rows of buildings between the base station and subscriber. For base station antennas near to or above the rooftops, the radiowaves to a subscriber will propagate primarily over the rooftops, except for subscribers on the few streets aligned with the base station. Signals propagating through the buildings are highly attenuated by the exterior and interior walls. Except in the distant suburbs, the gaps between buildings are small and are seldom aligned with the base station, or aligned from row to row.

To predict the range dependence of the spatial average path gain for macrocells, the individual buildings in a row are replaced by a continuous smooth obstacle, as seen in the end view in Fig. 7. All rows are assumed to have the same height, and each row of buildings is separated by the same distance d. Using this model of the buildings, the mean path gain PG is the product of three factors [1]:

$$PG = (PG_0)(Q^2)(PG_2) \tag{7}$$

Here, PG_0 is the free-space path gain given by Eq. (1). The term Q in (7) is the reduction of the fields arriving at the buildings near to the mobile as a result of diffraction past the previous rows of buildings. The term PG_2 represents the signal resulting from diffraction from the rooftops down to ground level. These two terms are discussed in more detail below.

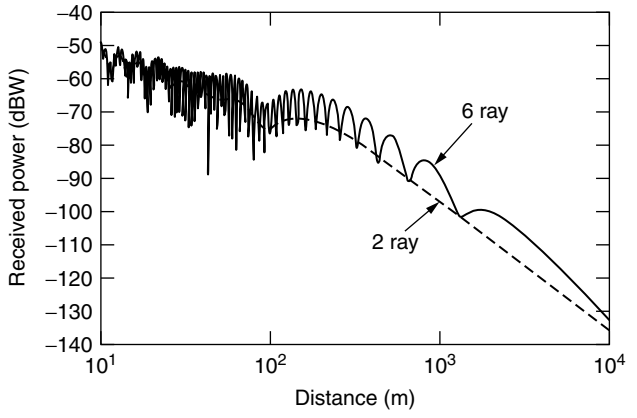


Figure 6. Comparison of the predictions of the two-ray model and the six ray model accounting for reflections form the buildings lining a street, as well as reflections from the ground. (© 1999 by H. L. Bertoni.)

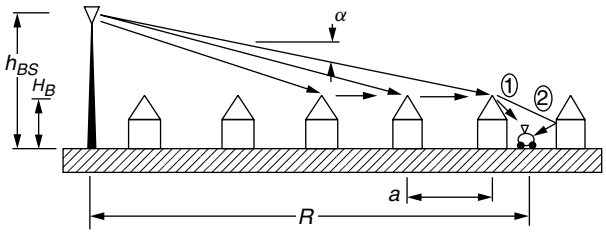


Figure 7. Side view of propagation over the rooftops to the last building before the subscriber and subsequent diffraction down to street level.

4.1. Q: Reduction of the Rooftop Fields

Except close to the base station, the horizontal distance from the base station to the buildings around the subscriber is large compared to the elevation of the base station antenna above the average building height. As a result, the glancing angle α shown in Fig. 7 is given by

$$\alpha = \tan^{-1}\left(\frac{h_{BS} - H_B}{R}\right) \approx \frac{h_{BS} - H_B}{R} \tag{8}$$

where h_{BS} is the base station height and H_B is the average building height. Provided that α is small, the radiowave propagating from the base station to the rooftops near the subscriber will undergo a cascade of multiple diffraction events at all of the previous rows of buildings. The mathematical treatment of multiple diffraction can be found in Ref. 1, while the simpler results are presented here.

A simple case occurs when the rooftops are all of the same height, the rows of buildings are all separated by the same distance d, and the base station antenna is at the same height as the rooftops. In this case the additional loss of the field reaching the Mth row of buildings from the base station is $Q = 1/M$ [1]. Because $M \approx R/d$, when (1) and $Q = 1/M$ are substituted into (7), the path gain is found to vary with distance as $PG \propto 1/R^4$, which is like the dependence for large separations of antennas above a flat earth. However, the proportionality constant for PG, and its dependence on frequency and antenna height will be different in the two cases.

When the base station antenna is well above the rooftops, as is the case for macrocellular applications, relatively simple expressions can again be found for Q. In this case the number of rows of buildings crossed by the radiowave is large and the glancing angle α is small. For example, in older cities the row separation is $d = 50$ m, so that 40 rows are crossed when the signal propagates to a distance of $R = 2$ km. In this case the reduction of the rooftop fields due to diffraction by previous rows of buildings can be expressed in terms of the parameter g_p, which is defined by

$$g_p = \alpha\sqrt{\frac{d}{\lambda}} \tag{9}$$

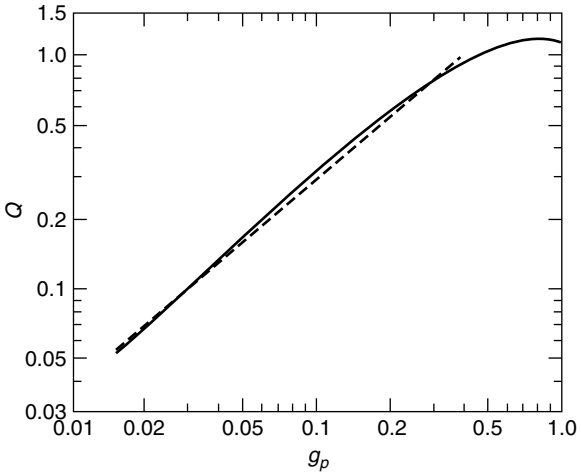


Figure 8. Dependence of the reduction of the rooftop fields Q on the dimensionless parameter g_p. The dashed line gives the simple approximation. (From Ref. 15, with permission.)

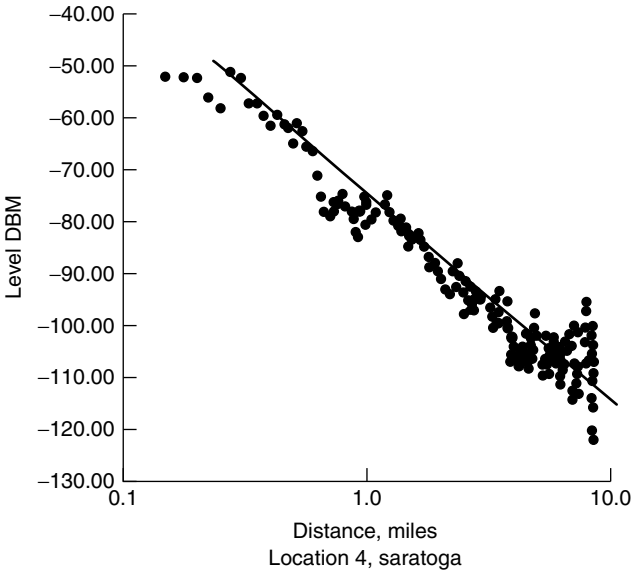


Figure 9. Measured small-area average power in dBm (dots) plotted versus R on a logarithmic scale. The solid line represents the theoretical prediction. (From Ref. 15, with permission.)

for propagation perpendicular to the rows of buildings. The variation of $Q(g_p)$ with g_p is plotted in Fig. 8 [15], and can be expressed in terms of a third order polynomial.

For large angles α, such as on satellite links, $g_p > 1$ and $Q(g_p) \approx 1$ so that only the last row of buildings before the mobile affects the received signal. A simple approximation to $Q(g_p)$ is given by the straight line shown dashed in Fig. 8. This approximation is given by

$$Q(g_p) = 2.35\, g_p^{0.9} \tag{10}$$

and is accurate to within 0.8 dB over the range $0.01 < g_p < 0.4$ [1].

4.2. PG_2: Diffraction from Rooftop to Ground Level

Many diffraction paths exist whereby the waves above the buildings reach ground level. In Fig. 7 the two rays giving the major contribution are as shown. The first of these is diffracted from the rooftop of the building nearest the mobile in the direction of the base station, while the second is reflected from the face of the building across the street. The field resulting from diffraction at the building edge is in the form of a cylindrical wave, with the edge acting as an equivalent line source. Because of the rapid spatial variation resulting from the interference of the two waves, the spatial average power will be the sum of the individual ray powers. With the foregoing assumptions, the spatial average received power is given by

$$PG_2 = \left[\frac{1}{\rho_1}|D(\theta_1)|^2 + |\Gamma|^2\frac{1}{\rho_2}|D(\theta_2)|^2\right] \tag{11}$$

where Γ is the reflection coefficient of the building opposite to the mobile, $k = 2\pi/\lambda$ and $D(\theta_i)$ is the diffraction coefficient.

For a receiver in the middle of the street, the distances ρ_1 and ρ_2 in from the diffracting edge (Fig. 9) are given by

$$\begin{aligned}\rho_1 &= \sqrt{(H_B - h_m)^2 + x^2}\\ \rho_2 &= \sqrt{(H_B - h_m)^2 + (2d - w - x)^2}\end{aligned} \tag{12}$$

while the angles θ_i for $i = 1, 2$ are

$$\theta_i = \arcsin\frac{(H_B - h_m)}{\rho_i} \tag{13}$$

In these expressions, H_B is the building height, h_m is the mobile height, w is the front-to-back dimension of the building, and x is the distance of the receiver from the diffracting edge of the building just before the mobile.

The diffraction coefficient $D(\theta_i)$ in Eq. (11) depends on the boundary condition at the rooftop edge of the building, which is rarely known. However, for diffraction angles θ_i away from 90°, the diffraction coefficient is not very sensitive to the boundary conditions. Thus we may use the diffraction coefficient for an absorbing edge, which for propagation perpendicular to the rows of buildings is [1]

$$D(\theta_i) = \frac{1}{\sqrt{2\pi k}}\left(\frac{1}{\theta_i - \alpha} + \frac{1}{2\pi + \theta_i - \alpha}\right) \approx \frac{1}{\sqrt{2\pi k}}\frac{1}{\theta_i} \tag{14}$$

We can simplify (11) by accounting for Γ, which for common building materials is $\Gamma \approx 0.3$. The value of Γ compensates for the differences in sizes of $D(\theta_1)/\sqrt{\rho_1}$ and $D(\theta_2)/\sqrt{\rho_2}$, so that the second term is close to the first term. The near equality of the two terms is seen from the deep fades observed in the fast fading pattern. Thus, the path gain for diffraction down to street level can be rewritten as

$$PG_2 \approx \frac{2}{\rho_1}[D(\theta_1)]^2 \approx \frac{1}{\pi k}\frac{1}{\rho_1\theta_1^2} \approx \frac{1}{\pi k}\frac{d/2}{(H_B - h_m)^2} \tag{15}$$

The variation of (11) with x has been validated by measurements in Japan [16] and England [17].

4.3. Path Gain for Macrocells

Macrocells in cities have $h_{BS} - H_B \sim 10$ m and 1 km $< R <$ 10 km. Since $d \approx 50$ m, the value of g_p falls in the range

0.015–0.15 at 900 MHz and 0.021–0.21 at 1800 MHz, so that we may use expression (12) for $Q(g_p)$. Combining expressions (1), (8), (9), (12), and (15) into (7), the path gain for isotropic antennas can be expressed in decibels as

$$PG_{dB} = 10\log\left(\frac{\lambda}{4\pi R}\right)^2 + 10\log\left[(2.35)^2\left(\frac{h_{BS}-H_B}{R}\right)^{1.8} \times \left(\frac{d}{\lambda}\right)^{0.9}\right] + 10\log\left[\frac{d/(2\pi k)}{(H_B-h_m)^2}\right] \quad (16)$$

Substituting $k = 2\pi/\lambda$ and $\lambda = c/f$, combining the various constant terms in (14) and expressing the frequency f_M in megahertz, the range R_k in kilometers, the path gain, can be written as

$$PG_{dB} = -92.5 - 21\log f_M + 10\log\left[\frac{d^{1.9}(h_{BS}-H_B)^{1.8}}{(H_B-h_m)^2}\right] - 38\log R_k \quad (17)$$

It is seen from (15) that the R dependence of Q combines with the free-space path to give the overall range dependence of 38 log R_k, corresponding to a range index $n = 3.8$ that is close to values measured in North American cities [4]. As a result of the near cancellation of the frequency dependence in Q^2 and PG_2, the path gain is seen to vary inversely with frequency to the 2.1 power, which is nearly that of the free-space path gain.

The predictions given by (17) for the received signal are shown in Fig. 9 superimposed on the small area average received power (dots) measured in Philadelphia [4]. The horizontal range is plotted on a logarithmic scale, for which (17) plots as a straight line. Excellent agreement is seen with the slope index of propagation and the average signal level. The deviations of the small area averages in Fig. 9 from the straight line correspond to the shadow fading. This deviation can be modeled in terms of the differences in building height along the rows; gaps between buildings, including street intersections; and the presence of trees [1,17].

4.4. Measurement-Based Models

In designing the original CMR systems, extensive measurements of the small-area average power versus R were made by various groups around the world. Okumura et al. [3] made an extensive set of measurements in and around Tokyo. Their work examined the effects of base station antenna height, frequency, building environment, terrain roughness, and other factors on the range dependence of the received signal, which was presented as curves of median received field strength (proportional to voltage) versus R for various parameters. Subsequently, Hata [4] fitted curves with simple formulas based on the slope intercept form $L = -10\log A + 10n\log R$ for the path loss L in decibels between isotropic antennas. Recall that the path loss in decibels is the negative of PG_{dB}. Hata's formulas were made to fit the measurements over the range of parameters: $150 \le f_M \le 1{,}500$ MHz, $1 \le R_k \le 20$ km, $30 \le h_{BS} \le 200$ m, and $1 \le h_m \le 10$ m. Over this range the result for urban areas is

$$L = 69.55 + 26.16\log f_M - 13.82\log h_{BS} - a(h_m) + (44.9 - 6.55\log h_{BS})\log R_k \quad (18)$$

The term $a(h_m)$ gives the dependence of path loss on subscriber antenna height, and is defined such that $a(1.5) = 0$.

To compare the predictions of (17) with (18), assume that $f_M = 1000$ MHz, $H_B = 10$ m (3 stories), $h_m = 1.5$ m, and $d = 50$ m. If we further assume that $h_{BS} = 20$ m, which is somewhat below the range of the Hatta model but consistent with practice, then the path loss obtained from (17) and (18) is

$$\text{Theory:}\quad L = 123.8 + 38\log R_k$$
$$\text{Hatta:}\quad L = 130.9 + 36.4\log R_k \quad (19)$$

The close agreement of the theory and measurements shown in this equation is a further substantiation of diffraction as a key process in the propagation from an elevated base station to subscribers at street level.

4.5. Range Dependence for Microcells

Microcellular systems make use of base station antennas located at about the height of three-story buildings, or on lampposts, to cover cells of radius 1 km or less. In a high-rise building environment, this placement is well below the building so that propagation is around the buildings rather than over them. For residential areas, this base station antenna height will be near to, or below the rooftops. In both environments, the location of the base station antenna relative to the buildings needs to be taken into accounted. Over microcells, the street grid likely to be rectangular, and line-of-sight (LoS) streets are a more significant fraction of all the streets in a cell, calling for their separate treatment. Measurement models appropriate to such antenna placement have been proposed for high-rise and residential environments [7,18]. Theoretical models have also been evaluated for predicting path loss, as discussed, for example, in Ref. 1.

5. 3D RAYS FOR SITE-SPECIFIC PREDICTIONS

Computer codes have been written to compute the ray paths and the ray fields working from databases of buildings and ground elevation. An example of a building database is shown in Fig. 10, which represents a simplified view of the high rise section of Rosslyn, Virginia [19]. In creating a building database there are tradeoffs that limit useful fidelity. The cost of creating the database increases with the detail included, as does the running time of the ray code. However, accuracy of the ray predictions does not continue to increase as more detail is added. The Rosslyn database of Fig. 4 shows the shape of the major geometric components of the buildings, but omits many smaller features, such as windows, balconies and decorative masonry.

Because almost all buildings have vertical sides, building databases are usually constructed with this assumption in order to reduce computation time. It is common to take the roofs to be flat, and some computer codes assume the ground to be flat. The most inexpensive databases describe each building using a polygon to represent its footprint, and a single building height. The more elaborate database of Fig. 10 stacks individual building elements, each of which has a polygonal base, while the height at each corner of the polygon can be different to accommodate slanting roofs.

In practice it is difficult to create a building database with position accuracy better than 0.5 m, or to include the architectural detail that can introduce phase shift in the reflection coefficients. For these reasons it is not possible to accurately compute the phases of the individual rays that would be needed to predict the spatial interference pattern for narrowband signals. Although the exact fading pattern cannot be predicted, its statistical properties can. The most important parameter is the small-area average received power, which corresponds to the sliding average in Fig. 2. This average, which is found by spatially averaging the power or field magnitude squared, can be computed by adding ray powers, as discussed in the text following Eq. (3).

The primary problem when using GO and UTD is to find the rays connecting the transmitter and the receiver. Even in simple environments there are a very large, and possibly infinite, number of such rays to be found. Geometric optical (GO) rays that undergo only reflections at the building walls and the ground are found using either the image method or the shooting–bouncing ray approach, which is most commonly employed for outdoor environments. It has been found that five to seven reflections must be accounted for to ensure accurate predictions [20].

In the shooting–bouncing ray (SBR) approach, rays starts from the transmitter in all directions over a sphere at incremental angular separations. For each ray from the transmitter, the first intersection with a building surface is computed. Using the laws of geometric reflection, the reflected ray is traced to the point of intersection with the first building surface, and so on. Because the rays have finite angular separation, there is a vanishing probability that a ray will pass through a predefined receiver point. To overcome this problem, the receiver is given a finite cross section whose diameter is equal to the product of the angular separation and the total pathlength. This procedure replaces the actual ray to the receiver by a single neighboring ray that undergoes reflections at the same building surfaces.

Diffracted rays are found by tracing the GO rays to the edge, and then treating the diffracting edge as a secondary source of rays that are traced using GO. An example of rays diffracted at a corner of one building and at the roof of another is shown in Fig. 11. The ray incident at one point along the edge excites diffracted rays that leave the edge in a Keller cone whose half-angle is equal to the angle between the incident ray and the edge [10]. Since the cone angle will vary along the edge, the edge is divided into small segments and a secondary ray trace is carried out from each segment. Moreover, each ray incident on an edge segment will, in general, have a different cone angle, hence requiring a separate ray trace.

Because each edge initiates a series of ray traces for each ray family that illuminates the edges, the three-dimensional (3D) ray trace is very time consuming, and can only accommodate rays that undergo no more than two diffraction events. In order to speed the code and to account for more diffraction events at the rooftops, the vertical plane launch (VPL) approximation has been proposed [19,21]. This approximation is involves replacing the Keller cone for diffraction at horizontal edges by the vertical planes. For the common case when the horizontal displacement of the rays is larger than the vertical displacement, the distortion of the ray path will be small with this approximation. When viewed from above, the rays diffracted in the forward direction lie in the plane of incidence, while rays diffracted backwards lie in the plane of reflection.

The VPL approximation allows the ray paths to be constructed by first carrying out a 2D ray trace in

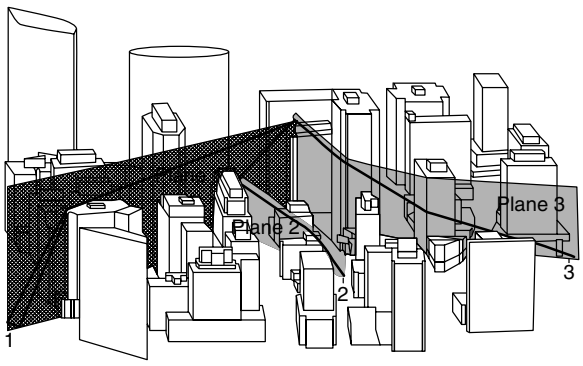


Figure 10. Building database of Rosslyn, Virginia, showing the vertical planes launched from a base station atop a building. The ray paths in the planes are those found from the VPL approximation. (From Ref. 19, with permission.)

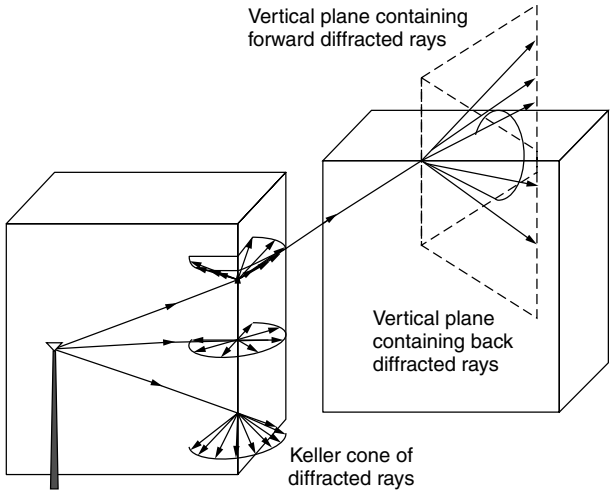


Figure 11. Rays diffracted at an edge lie in the Keller cone, whose half-angle is equal to that between the incident ray and the edge. In the VPL approximation, the cone for diffraction at horizontal edges is unrolled into the vertical planes. (© 2002 by H. L. Bertoni.)

the horizontal plane, followed by a simpler analytic determination of the ray paths in the vertical dimension. The procedure, in effect, traces vertical planes launched by the transmitter as they pass over and are reflected from building surfaces, as suggested in Fig. 10. These planes may also diffract at vertical building corners. After unfolding the vertical plane for each ray, the path in the vertical plane is determined by accounting for possible reflection or diffraction at horizontal edges. Because the shooting and bouncing ray method needs be used in only two dimensions, much less computer time is required. Moreover, by using analytic methods in the vertical plane, many more diffraction events at horizontal edges can be accommodated. An example of the VPL prediction of the spatial average received power in Rosslyn is shown in Fig. 12 [22]. The transmitter is located atop the building shown in Fig. 10, while the receiving locations are spaced approximately 5 m apart along a 2-km drive path on six different streets. For comparison with these predictions, measurements were made as the receiving van was driven along all the streets. The predictions are generally in good agreement with the measurements, except for 1360 receiver numbers and higher, which are on a street at the edge of the database. When compared to measurements, the error of ray predictions typically has an average of 1–2 dB and a standard deviation of 6–8 dB.

6. SUMMARY

Measurements and theory give complementary ways of predicting path loss and other statistical properties of the radio channel. Given the highly variable nature of the path loss, theoretical predictions are in good agreement with measurements. When validated against a set of measurements, the theoretical models allow for the variation of system parameters, such as frequency and antenna height, and building geometry. Thus a good theoretical model increases the value of a set of measurements by carrying it to a much larger range of parameters and building environments. Although we have discussed only a few aspects of path loss prediction in this article, theoretical models can be used to predict other channel characteristics. For example, computation of the ray fields via GO and UTD directly gives the directions of departure and arrival at both ends of the link, time delay of the ray, and the contribution of the ray fields to the received voltage. By making predictions for many mobile locations, the ray codes can be used for Monte Carlo simulation of statistical channel parameters, such as delay spread or angle spread.

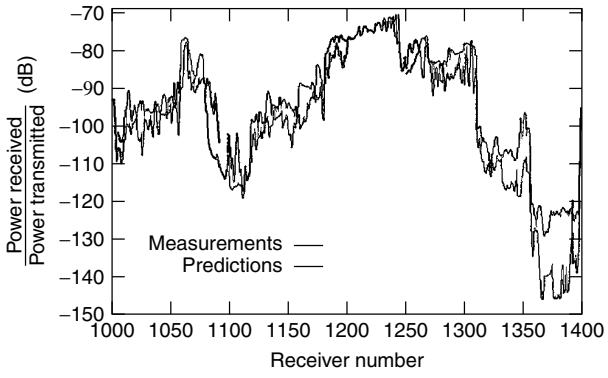


Figure 12. Comparison of measurements and predictions of the small-area path gain from a rooftop base station and mobiles at approximately 5-m intervals along six different streets in Rosslyn for a frequency of 900 MHz.

BIOGRAPHY

Henry L. Bertoni is on the faculty of Polytechnic University in Brooklyn, serving as head of the ECE Department (1990–95, 2001–present), and as vice provost of graduate studies (1995–96). His research has dealt with theoretical aspects of wave phenomena in electromagnetics, ultrasonics, acoustics, and optics. He has authored or coauthored over 80 journal papers and nine book chapters on these topics. Four journal articles have received best paper awards. His current research deals with characterizing the radio channel for modern wireless application, and the theoretical prediction of these characteristics. He and his students were the first to explain the physical mechanisms underlying characteristics observed in the measurements of cellular mobile radio signals. Much of this work is described in his recent book *Radio Propagation for Modern Wireless Systems*, Prentice Hall PTR, 2000. Dr. Bertoni is a fellow of the IEEE. He was the first chairman of the Technical Committee on Personal Communications of the IEEE Communications Society, and was IEEE representative to, and chairman of, the Hoover Medal Board of Award. He is a member of the International Scientific Radio Union and the Radio Club of America. From 1998 to 2001 he was a distinguished lecturer of the IEEE Antennas and Propagation Society.

BIBLIOGRAPHY

1. H. L. Bertoni, *Radio Propagation for Modern Wireless Applications*, Prentice-Hall PTR, Englewood Cliffs, NJ, 2000.
2. H. L. Bertoni, Radio channel characteristics observed for cellular and microcellular links: A tutorial review, *J. Commun. Networks* **1**: 249–265 (1999).
3. Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, Field strength and its variability in VHF and UHF land-mobile radio service, *Rev. Electric. Commun. Lab.* **16**: 825–873 (1968).
4. G. D. Ott and A. Plitkins, Urban path-loss characteristics at 820 MHz, *IEEE Trans. Vehic. Technol.* **VT-27**: 189–197 (1978).
5. H. H. Xia et al., Microcellular propagation characteristics for personal communications in urban and suburban environments, *IEEE Trans. Vehic. Technol.* **43**: 743–752 (1994).
6. M. Hata, Empirical formula for propagation loss in land mobile radio service, *IEEE Trans. Vehic. Technol.* **29**: 317–325 (1980).
7. D. Har, H. H. Xia, and H. L. Bertoni, Path loss prediction model for microcells, *IEEE Trans. Vehic. Technol.* **48**: 1453–1462 (1999).
8. L. J. Greenstein, V. Erceg, Y. S. Yeh, and M. V. Clark, A new path-gain/delay-spread propagation model for digital cellular channels, *IEEE Trans. Vehic. Technol.* **46**: 477–485 (1997).

9. C. Cheon, G. Liang, and H. L. Bertoni, Simulating radio channel statistics for different building environments, *IEEE J. Select. Areas Commun.* **19**: 2191–2200 (2001).

10. D. A. McNamara, C. W. I. Pistorius, and J. A. G. Malherbe, *Introduction to the Uniform Geometrical Theory of Diffraction*, Archtech House, Norwood, MA, 1990.

11. J. Fuhl, J.-P. Rossi, and E. Bonek, High-resolution 3-D direction-of-arrival determination for urban mobile radio, *IEEE Trans. Antennas and Propag.* **45**: 672–682 (1997).

12. M. Lecours, I. Y. Chouinard, G. Y. Delisle, and J. Roy, Statistical modeling of the received signal envelope in a mobile radio channel, *IEEE Trans. Vehic. Technol.* **37**: 204–212 (1988).

13. H. H. Xia et al., Radio propagation characteristics for line-of-sight microcellular and personal communications, *IEEE Trans. Antennas Propag.* **41**: 1439–1447 (1993).

14. A. J. Rustako, Jr., N. Amitay, g. J. Owens, and R. S. Romano, Radio propagation at microwave frequencies for line-of-sight microcellular mobile and personal communications, *IEEE Trans. Vehic. Technol.* **40**: 203–210 (1991).

15. J. Walfish and H. L. Bertoni, A theoretical model of UHF propagation in urban environments, *IEEE Trans. Antennas Propag.* **36**(10): 1788–1796 (1988).

16. F. Ikegami, S. Yoshida, T. Takeuchi, and M. Umehira, Propagation factors controlling mean field strength on urban streets, *IEEE Trans. Antennas Propag.* **32**: 822–829 (1984).

17. L. R. Maciel and H. L. Bertoni, Theoretical prediction of slow fading statistics in urban environments, *Proc. IEEE ICUPC'92 Conf.*, 1992, pp. 1–4.

18. E. Damosso, ed., *COST Action 231: Digital Mobile Radio; towards Future Generation Systems*, European Commission, Directorate G, Brussels, 1999.

19. G. Liang and H. L. Bertoni, A new approach to 3-D ray tracing for propagation prediction in cities, *IEEE Trans. Antennas Propag.* **46**: 853–863 (1998).

20. G. E. Athanasiadou, A. R. Nix, and J. P. McGeehan, Microcellular ray tracing propagation model and evaluation of its narrow-band and wide-band predictions, *IEEE J. Select. Areas Commun.* **18**: 322–335 (2000).

21. J. P. Rossi et al., A ray-launching method for radio-mobile propagation in urban area, *Digest of IEEE APS Symp.* London, Ontario, Canada, June 1991, pp. 1540–1543.

22. G. Liang, private communication, Jan. 2002.

PEAK-TO-AVERAGE POWER RATIO OF ORTHOGONAL FREQUENCY-DIVISION MULTIPLEXING

CHINTHA TELLAMBURA
Monash University
Clayton, Victoria, Australia

MATTHEW G. PARKER
University of Bergen
Bergen, Norway

1. INTRODUCTION

OFDM techniques have been proposed for digital TV broadcasting and high-speed wireless networks over multipath channels [1]. OFDM is commonly implemented using discrete Fourier transform (DFT) techniques and has been adopted, or is being investigated, for wireless LANs, wireless ATM, digital audio broadcasting [2], terrestrial digital videobroad-casting [3], and the broadband wireless local loop. OFDM offers many advantages such as resistance to multipath and excellent performance under noisy conditions.

With OFDM, the single carrier wave is replaced by simultaneous transmission of the signal on multiple, equally spaced subcarriers [4]. A baseband version of OFDM is called *discrete multitone transmission* (DMT). OFDM systems require the calculation of discrete Fourier transforms (DFTs). It is the availability of technology that allows for the implementation of fast transform (FFT) algorithms on integrated circuits at a reasonable price, that has made OFDM and DMT the modulation method of choice for the commercial applications given above [5].

Unfortunately, there are a number of difficulties with implementing OFDM and DMT:

- When the sinusoidal signals of the N subcarriers add constructively, the peak power can be N times the mean power; that is, the peak-to-average power ratio (PAR) of the transmitted signal can as large as N.
- Radiofrequency amplifiers are used to achieve the linearity over the entire signal. This causes high battery demand in mobile/wireless applications.
- The allocation of the radio spectrum for radio LANs limits the isotropically radiated peak envelope power. If the output peak is clipped, this generates out of band radiation due to intermodulation distortion as well as in-band distortion.

The first difficulty listed above is in fact the cause of the second and third. These problems limit the usefulness of OFDM for some applications.

2. PEAK-TO-AVERAGE POWER RATIO OF OFDM

The complex baseband OFDM signal may be represented as

$$s(t) = \frac{1}{\sqrt{N}} \sum_{n=-\infty}^{\infty} \sum_{k=0}^{N-1} a_{k,n} e^{j2\pi k \Delta f t} g[t - k(T + T_g)] \qquad (1)$$

where $j^2 = -1$, N is the total number of subcarriers, and $a_{k,n}$ is the data symbol for the k subcarrier and the nth OFDM symbol (i.e., N subcarrier OFDM system transmits a block of N data symbols per OFDM symbol). The frequency separation between any two adjacent subcarriers is $\Delta f = 1/T$. The unit rectangular pulse $g(t)$ is of duration $T + T_g$, where T_g is known as the "guard interval."

Because there is no overlap between different OFDM symbols, for the PAR problem it is sufficient to consider a single OFDM symbol ($n = 0$). In practice, filtering can cause some degree of intersymbol interference, which will be neglected here. The guard interval is used to repeat parts of each OFDM signal, but has no effect on the PAR. Therefore we may set $T_g = 0$. Since only $n = 0$ is sufficient

for the problem at hand, $a_{k,n}$ may be replaced by a_k. Thus, for PAR considerations, the complex baseband signal may be represented as

$$s(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k e^{j2\pi k \Delta f t}, \quad 0 \le t < T \tag{2}$$

Note that all the subcarriers are mutually orthogonal. Each modulated symbol a_k is chosen from the set $F_q = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ of q distinct elements. The set F_q is called the *signal constellation* of the q-ary modulation scheme. While several modulation schemes are in use it should be noted that the statistical distribution of the PAR is largely independent of the signal constellation. In most applications, one uses phase shift keying (PSK) signaling in which

$$F_q = \{1, \zeta, \zeta^2, \ldots, \zeta^{q-1}\} \tag{3}$$

where $\zeta = e^{j2\pi/M}$. For example, for binary PSK (BPSK) $F_q = \{1, -1\}$ and for quaternary PSK (QPSK), $F_q = \{1, j, -1, -j\}$. Another popular modulation technique is quadrature amplitude modulation (QAM), for which

$$F_q = \{m + jn\} \tag{4}$$

where m and n are selected integers.

Note that for any PSK constellation F_q for any $u \in F_q$, $|u|^2 = 1$, and this condition does not hold for QAM constellations. In general, all elements in F_q occur with equal probability $1/q$.

We shall write an ordered N-tuple $\mathbf{a} = (a_0, a_1, \ldots, a_{N-1})$ and let $(F_q)^N$ denote the set of all ordered N-tuples where each $a_k \in F_q$. We shall refer to any member of the set $(F_q)^N$ as a *data frame*. Note that each a_k carries $\log_2 q$ data bits. Normal values are $q = 2, 4, 8$, and so on. Each OFDM symbol thus carries $N \log_2 q$ data bits. The instantaneous envelope *power* of the signal is the real-valued function $P_a(t) = |s(t)|^2$. We define the peak-to-average power ratio (PAR) as

$$\mathrm{PAR}(\mathbf{a}) = \frac{\max_t P_a(t)}{E(P_a(t))} \tag{5}$$

where $E(\cdot)$ denotes the time average. For a PSK constellation, this value is unity. Strictly speaking, this definition should be called the peak-to mean envelope power ratio (PMEPR), because $s(t)$ is the envelope but not the transmitted signal itself. As such, this is also called the *baseband* PAR. The actual transmitted signal is modeled as

$$S(t) = \Re(s(t)e^{j2\pi f_c t}) \tag{6}$$

where f_c is the carrier frequency and $\Re(z)$ denotes the real part of z. The definition of PAR would now be

$$\mathrm{PAR}(\mathbf{a}) = \frac{\max_t |\Re(s(t)e^{j2\pi f_c t})|^2}{E(|S(t)|^2)} \tag{7}$$

which is also known as the *passband* PAR. It is often easier to work with definition (5) rather than (7). Further, if f_c is large (i.e., $f_c \gg N/T$), which is the case in practice, this is approximately 3 dB higher than the baseband PAR. This difference is more or less fixed. Consequently, we will be using the baseband PAR without any loss of generality or applicability.

The PAR is a function of the data frame, and recall that there are q^N distinct data frames. For any input data frame, we have

$$1 < \mathrm{PAR}(\mathbf{a}) \le N \tag{8}$$

For example, for $N = 256$ the PAR can be as high as 24 dB $[10 \log_{10}(256)]$. Fortunately, very high PAR values are very rare. For example, with BPSK, only four sequences $0000\ldots, 1111\ldots, 0101\ldots$, and $1010\ldots$ achieve $\mathrm{PAR}(\mathbf{a}) = N$. For randomly distributed data, the probability of an occurrence of this is $4/2^N = 2^{2-N}$. This probability is negligible when N is large - as is the case in practice.

The PAR of a sequence is closely related to its out-of-phase aperiodic autocorrelation (APA) values. The APA coefficients of $\mathbf{a}$ are

$$\rho(k) = \sum_{n=0}^{N-1-k} a_{n+k} a_n^* \quad \text{for} \quad k = 0, \ldots, N-1 \tag{9}$$

The PAR is bounded as

$$\mathrm{PAR}(\mathbf{a}) \le 1 + \frac{2}{N} \sum_{k=1}^{N-1} |\rho(k)| \tag{10}$$

This shows that binary or polyphase sequences with low out-of-phase APA values [i.e., small $\rho(k)$ for $k \ge 1$] can be used to construct low PAR signals. Conversely, Schroeder, [6] notes that sequences that have low PAR also have low APA values. The problem of constructing sequences with low APA values (i.e., similar to an impulse function) is a longstanding problem. The general problem of finding sequences that minimize the PAR seems just as difficult.

Example 1. Consider $\mathbf{a} = (1, 1, 1, -1, 1)$, which is a Barker sequence. Its APA is $\{5, 0, 1, 0, 1\}$. Hence applying (10) gives $\mathrm{PAR}(\mathbf{a}) \le 1 + \frac{4}{5}$; thus, the PAR is less than 2.55 dB. By computing (5), the PAR is exactly 2.55 dB. In this case, the upper bound coincides with the exact.

Example 2. Using (10), we can immediately device a simple PAR reduction code. For an information sequence $p_0 = (m_0, m_1, \ldots, m_{n-1})$, where $m_k \in \{1, -1\}$, the encoder output is given by $p_e = (m_0, \ldots, m_{n-1}, -m_{n-2}, m_{n-3}, -m_{n-4}, \ldots, -m_0)$. This code rate is $n/(2n-1)$ and length $N = 2n - 1$. For k odd, $\rho(k)$ of p_e is zero. For example, when $n = 3$, then $\rho(1) = \rho(3) = 0$ and $|\rho(2)| \le 3$ and $|\rho(4)| = 1$. Thus our bound gives PAR ≤ 2.6, almost a 3-dB reduction. For large N, the coding rate is almost $\frac{1}{2}$. The sum of $|\rho(k)|$ for k even is bounded by $(N-1)^2/4$. Thus, the peak is bounded as $\mathrm{PAR} \le 1 + (N-1)^2/(2N)$, a 3 dB reduction.

3. STATISTICAL PROPERTIES OF PAR

3.1. CCDF of the PAR

Since the input data are randomly distributed in many applications (if not, they can be made so by the use of a

suitable scrambling operation), PAR(**a**) itself is a random variable. The complementary cumulative distribution function (CCDF), the probability that the PAR of an OFDM symbol exceeds a certain threshold, is useful for many purposes. The CCDF is defined as

$$F(\zeta) = \Pr(\text{PAR}(\mathbf{a}) \geq \zeta) \tag{11}$$

The CCDF is shown for $N = 32$, 64, 128, and 256 in Fig. 1. For 256 subcarriers, the maximum PAR is 24 dB. To reach this, however, all the subcarriers need to be in phase at some instant in time, and therefore produce an amplitude peak equal to the sum of the amplitudes of the individual subcarriers. This occurs with extremely low probability for large N. For example, the PAR exceeds 12.5 dB for 1 in 10^5 of all the possible transmitted OFDM symbols.

3.2. Gaussian Approach

An exact expression for $F(\zeta)$ is not yet known because computing PAR(**a**) requires the time instances for which the derivative of the envelope power $P_a(t)$ equals zero. This is a root-finding problem for a nonlinear equation and as such there are no analytic formulas for the roots. However, as a way around this difficulty, we make the following assumptions:

A1. $s(t)$ is a complex Gaussian process.
A2. N samples of $s(t)$ given by $s_n = s(nT/N)$ for $n = 0, 1, \ldots, N-1$ are independent and identically distributed complex Gaussian random variables.
A3. The maximum of $|s_n|^2$ is equal to $\max_t = P_a(t)$.

A1 becomes quite accurate as N increases. The independence assumption in A2 is never exact because we know that the N samples must satisfy Parseval's theorem. A3 is never exactly true. Nevertheless, we define random variables (RVs):

$$Y_n = |s_n|^2 \quad n = 0, \ldots, N-1 \tag{12}$$

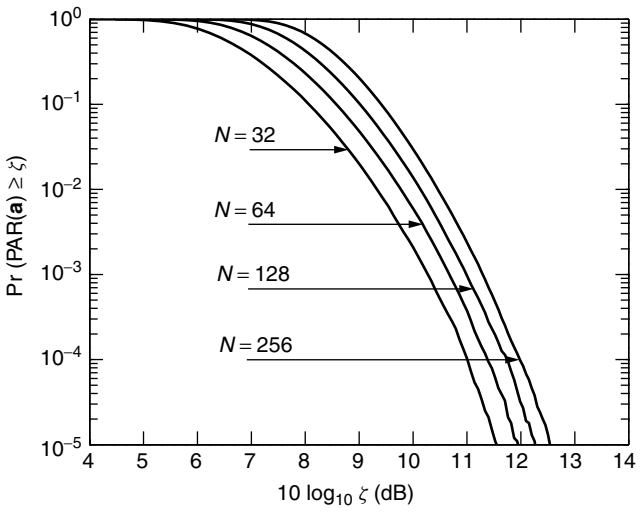


Figure 1. The CCDF for N QPSK subcarriers.

Since the real and imaginary parts of s_n are independent (provided no oversampling is used), with mean zero and the same variance, Y_n approaches a chi-squared distribution with two degrees of freedom. The Y_n values hence are i.i.d. and exponential RVs. Their cumulative density function (CDF) is given by

$$F(y) = \Pr(Y_n \leq y) = 1 - e^{-y} \tag{13}$$

The statistical properties of the maximum of Y_n can be readily derived. We see that the CDF of the maximum is given by

$$\begin{aligned} F_{\max}(y) &= \Pr\{Y_{\max} \leq y\} \\ &= \Pr\{\text{all } Y_n \leq y\} \\ &= (1 - e^{-y})^N \end{aligned} \tag{14}$$

The CDF of the PAR is then obtained as

$$F(\zeta) = 1 - (1 - e^{-\zeta})^N \tag{15}$$

Despite the three assumptions that may not always hold, this result is useful for quick analysis of the statistics of PAR and for determining the achievable PAR reduction for some schemes.

3.3. Asymptotic Results

The statistical behavior of PAR(**a**) for large N is important, and in some practical applications, N can be as large as 2048 or more. We can show that PAR(**a**) grows as $\ln N$; that is, for a randomly picked data sequence, PAR(**a**) is unlikely to be significantly less than $\ln N$. From Eq. (14), we get

$$\begin{aligned} \Pr\{\text{PAR}(\mathbf{x}) \leq \alpha \ln N + h\} &= (1 - e^{-(\alpha \ln N + h)})^N \\ &\simeq e^{-N^{(1-\alpha)}e^{-h}} \text{ for } N \to \infty \end{aligned} \tag{16}$$

where $\alpha \geq 1$. This follows readily from fact that $\lim_{n\to\infty}(1 - \theta/n)^n = e^{-\theta}$. If $\alpha = 1$ and $h = -h$ in (16), we see that

$$\Pr\{\text{PAR}(\mathbf{x}) \leq \ln N - h\} \simeq e^{-e^h} \tag{17}$$

This is the formal mathematical equivalent of our statement above. This also elicits information about clipping and coding for PAR reduction. *Clipping* is a method used to deal with high peak amplitude excursions at the transmitter output. This is necessary because the D/A converter has a limited resolution (i.e., the number of bits) and the power amplifier cannot be linear over an amplitude range that includes the peak amplitudes. A clip occurs when the signal amplitude exceeds a predefined threshold, and hence clipping is described as follows:

$$s_c(t) = \begin{cases} s(t) & \text{if } |s(t)| \leq s_{\text{clip}} \\ s_{\text{clip}} e^{j\angle s(t)} & \text{if } |s(t)| > s_{\text{clip}} \end{cases} \tag{18}$$

where $s_{\text{clip}} > 0$ is the clipping threshold. The probability of clipping is the number of clips per unit time. Of course, each clip introduces symbol errors and out-of-band noise.

Equation (17) shows that for a normal OFDM system, if the clipping threshold is set below $\ln N$, then the clipping probability will be unity for large N. Likewise, (16) suggests if the clipping threshold is set above αN, then the clipping probability can be arbitrarily reduced. Here the right-hand side (RHS) of (16) explicitly shows that the decay rate depends on both α and h.

Additionally, coding seems unnecessary for PAR reduction of $\alpha \ln N$ or higher as clipping will not occur very often. In contrast, keeping the PAR at a level significantly below $\ln N$ using clipping will introduce significant distortion. At this point coding becomes interesting. Consider coding to reduce the PAR to h below $\ln N$, where $|h|$ is small compared to $\ln N$. Assume a binary modulation [i.e., $x_k \in (+1, -1)$]. From (17), the achievable coding rate is

$$R(h) = \frac{\log_2(2^N e^{-e^h})}{N} = 1 - \frac{e^h}{N \log_2 e} \tag{19}$$

As h increases, the achievable coding rate tends to zero. Similarly, if h is negative, the achievable coding rate tends to one, suggesting that the PAR can be limited to a level above $\ln N$ with very little redundancy. Moreover, the required amount of redundancy decays exponentially with the difference between the target PAR and $\ln N$. It appears that any family of good codes (i.e., of nonvanishing coding rate) must have a PAR bound of around $\ln N$.

3.3.1. Code Rate. Consider QPSK modulated N subcarriers, which can accommodate $2N$ information bits at most. Suppose that we need a code rate $R = 1 - K/(2N)$ to limit the PAR to ζ. Thus we have

$$\Pr(\text{PAR} \le \zeta) = \frac{2^{2N-K}}{2^{2N}} \tag{20}$$

which can be rearranged as

$$R = 1 + \frac{1}{2N} \log_2 \Pr(\text{PAR} \le \zeta) \tag{21}$$

This probability term can be estimated by simulation. Figure 2 shows the required code rate to limit the PAR. For $N = 128$, to reduce PAR to 7 dB from 21 dB for the uncoded case, the required code rate is 0.98. This suggests that the PAR can be much reduced by a small amount of redundancy. Unfortunately, it has thus far not been possible to discover such a code.

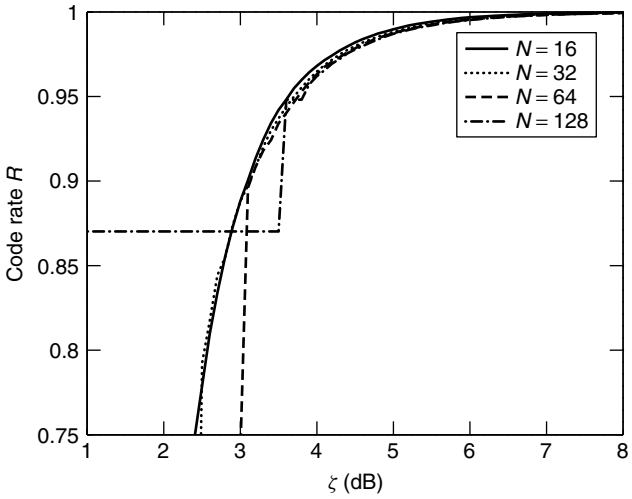


Figure 2. Code rate for limiting the PAR.

4. COMPUTATIONAL METHODS FOR PAR

4.1. Discrete-Time PAR

In order to compute PAR($\mathbf{a}$) exactly, we need all the roots of the equation

$$\frac{dP_{\mathbf{a}}(t)}{dt} = 0 \tag{22}$$

which is difficult to solve, especially for higher-order modulation formats. Most PAR reduction techniques are concerned with reducing PAR($\mathbf{a}$) [Eq. (5)]. However, since most systems employ discrete-time signals, the maximum amplitude of LN samples of $s(t)$ is used to approximate it, where L is the oversampling factor. The sampling can be implemented by an inverse discrete Fourier transform (IDFT). Hence consider the IDFT of length LN of $\mathbf{a}$ expressed as

$$A_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k e^{i2\pi nk/LN} \quad n = 0, 1, \ldots, LN-1 \tag{23}$$

It is seen that A_n are the samples of the waveform [Eq. (2)]. The discrete-time PAR is thus defined as

$$\text{PAR}(\mathbf{a})_{\text{dis}} = \max_{0 \le n < LN} |A_n|^2 \tag{24}$$

$L > 1$ corresponds to oversampling. Of course, if $L \gg 1$, the discrete-time PAR should approach the (continuous-time) PAR. It is therefore clear that

$$\text{PAR}(\mathbf{a})_{\text{dis}} \le \text{PAR}(\mathbf{a}) \tag{25}$$

Moreover [7] has shown that $\text{PAR}(\mathbf{a}) \le 2\text{PAR}(\mathbf{a})_{\text{dis}}$ if $L = \lceil 2\pi \rceil$. In practice, samples of (2) are generated by means of an inverse fast Fourier transform (IFFT), and fed to a digital-to-analog converter followed by an antialiasing lowpass filter. Quite often an oversampling factor of 4 is sufficiently accurate. For BPSK modulated subcarriers, this fact can be verified because the continuous-time PAR can be computed exactly. The most common method to compute PAR($\mathbf{a}$) is to use the discrete-time PAR. Note that we will simply use the term PAR, except when we are concerned about the difference between the discrete- and continuous-time values.

4.2. Using the Infinity Norm

This method was suggested by Van Eetvelt. The peak of a continuous function $s(t)$ is given by the L_∞ norm defined as

$$\max |s(t)| = L_\infty(s(t)) = \lim_{n \to \infty} \left[\frac{1}{T} \int_0^T |s(t)|^n dt \right]^{1/n} \tag{26}$$

To compute the L_∞ norm, we can use the result that for increasing p the L_p norm is nondecreasing. So in practice, taking a sufficiently large power allows the peak value to be approximated as closely as required. However, as computing this integral is not that easy, the use of the discrete-time PAR is much more convenient.

4.3. Computation of Continuous-Time PAR: BPSK case

To compute the continuous-time PAR, the roots of the derivative of the envelope power function (EPF) are required. At first, finding the required roots appears very difficult, since this derivative consists of sinusoidal functions. As such, the problem suggests a general root finding algorithm for nonlinear functions. Fortunately, this difficulty can be avoided for the BPSK case. Using an inverse cosine based transformation, the EPF can be converted to a sum of Chebysev polynomials. Moreover, the required roots are now trapped within the interval from 0 to 1. So the original root finding problem is reduced to a root finding problem for a polynomial. Reliable algorithms for finding all roots of a polynomial (a polynomial of order n will have n roots) are well known. Consequently, using this approach, the absolute peak of the EPF can be evaluated exactly. In this case, we have $a_k \in \{1, -1\}$. It is easy to show that [8,9]

$$P_a(t) = \sum_{k=0}^{N-1} \beta_k \cos(2\pi k t) \tag{27}$$

where

$$\beta_k = \begin{cases} 1 & k = 0 \\ \dfrac{2}{N} \displaystyle\sum_{n=0}^{N-1-k} a_n a_{n+k} & k = 1, 2, \ldots, N-1 \end{cases} \tag{28}$$

To compute PAR($\mathbf{a}$) exactly, the roots of $\dfrac{dP_a(t)}{dt} = 0$ are needed. Let us define

$$Q_a(t) = P_a\left[\frac{\cos^{-1} t}{2\pi}\right] = \sum_{k=0}^{N-1} \beta_k T_k(t) \tag{29}$$

where $T_k(t) = \cos(k \cos^{-1} t)$ is the kth order Chebysev polynomial (Ref. 10, p. 1054). Note that $T_0(t) = 1, T_1(t) = t, T_2(t) = 2t^2 - 1$ and so on (explicit expressions for the coefficients of $T_n(t)$ for any N are available). Since $Q_{\mathbf{a}}(t)$ is a polynomial of degree $(N-1)$, its first derivative is a polynomial of degree $(N-2)$. Recall that a polynomial of degree n will have n roots. These roots can be real or complex and many algorithms exist with which one can find all the roots of a polynomial. For example, a companion matrix can be constructed whose eigenvalues are the desired roots. Since

$$\frac{dP_{\mathbf{a}}(t)}{dt} = \frac{dQ_{\mathbf{a}}(\cos(2\pi t))}{dt} = -2\pi \sin(2\pi t) \frac{dQ_{\mathbf{a}}[\cos(2\pi t)]}{d\cos(2\pi t)} \tag{30}$$

the derivative vanishes both at $t = 0, \frac{1}{2}$ and at the transforms of the real roots of $\dfrac{dQ_a(t)}{dt}$ that lie between -1 and $+1$. Let those roots be $\xi_1, \xi_2, \ldots, \xi_M$ where $M \le N-2$. Define the set

$$\Lambda = \left\{0, \frac{1}{2}, \frac{\cos^{-1}\xi_1}{2\pi}, \ldots, \frac{\cos^{-1}\xi_M}{2\pi}\right\} \tag{31}$$

It is clear that all the required roots of the derivative of $P_a(t)$ are in this set. Note that $P_{\mathbf{a}}(t)$ is a periodic function and only the roots between 0 to 1 need to be considered. Therefore, the continuous-time PAR is obtained by

$$\text{PAR}(\mathbf{a}) = \max_{t \in \Lambda} P_{\mathbf{a}}(t) \tag{32}$$

5. PAR REDUCTION METHODS

We next describe some of the techniques that have been proposed to reduce the effects of high PAR.

5.1. Multiple Signal Generation

The basic idea behind this approach is to generate multiple, independent OFDM symbols to represent an input data frame and select the minimum PAR symbol for transmission. There are several techniques based on this idea and these primarily differ in the way they generate the multiple symbols. Another issue with this approach is the need for side information to tell the receiver which one of the signals has been used. Suppose that $M \ge 1$ independent OFDM symbols are generated for an information sequence. The CCDF of the minimum of these is given by

$$\Pr(\text{PAR}_{\min} \ge \zeta) = [1 - (1 - e^{-\zeta})^N]^M \tag{33}$$

Figure 3 shows this CCDF for $M = 1, 2, \ldots, 16$ and $N = 128$. $M = 1$ is the ordinary OFDM. Even for $M = 4$, a PAR reduction of about 4 dB occurs at a probability of 10^{-6}.

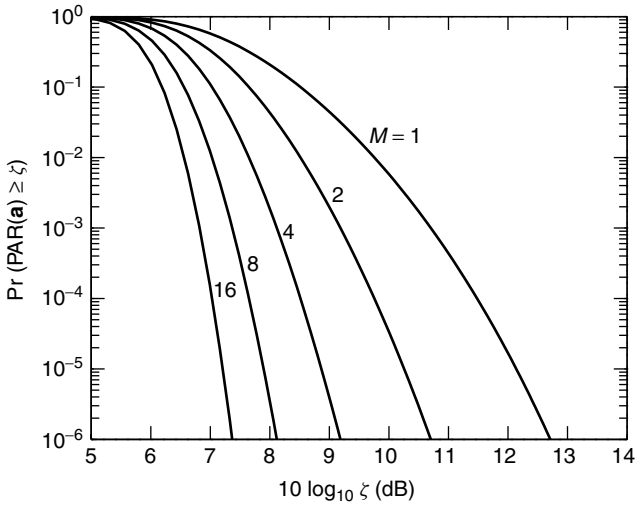


Figure 3. CCDF for the lowest PAR out of M statistically independent signals.

5.1.1. Partial Transmit Sequences. We shall write the input data block as a vector, $\mathbf{X} = [X_0, \ldots, X_{N-1}]^T$. For the PTS approach, the input data vector $\mathbf{X}$ is partitioned into disjoint subblocks, as $\{\mathbf{X}_m | m = 1, 2, \ldots, M\}$, and these are combined to minimize the PAR. While several subblock partitioning schemes do exist, we assume the simplest scheme for which the subblocks consist of a contiguous set of subcarriers and are of equal size. Now, suppose that for $m = 1, \ldots, M$, $\mathbf{A}_m = [A_{m1}, A_{m2}, \ldots, A_{m,LN}]^T$ is the zero-padded IFFT of $\mathbf{X}_m$. These are the partial transmit sequences. The objective is thus to combine these with the aim of minimizing the PAR. The signal samples at the output of the PTS combiner can be written as

$$\mathbf{S} = \begin{bmatrix} A_{11} & A_{21} & \ldots & A_{M1} \\ A_{12} & A_{22} & \ldots & A_{M2} \\ \ldots & \ldots & \ldots & \ldots \\ A_{1LN} & A_{2LN} & \ldots & A_{MLN} \end{bmatrix} \begin{bmatrix} e^{j\phi_1} \\ e^{j\phi_2} \\ \vdots \\ e^{j\phi_M} \end{bmatrix} \tag{34}$$

where $\mathbf{S} = [S_1(\Phi), \ldots, S_{LN}(\Phi)]^T$ contains the optimized signal samples. We shall write the phase factors as a vector, $\Phi = [\phi_1, \phi_2, \ldots, \phi_M]^T$. The phase factors $\{\phi_k\}$ are chosen to minimize the peak of the signal samples $|S_k(\Phi)|$. So the minimum PAR is related to the problem

$$\begin{aligned} &\text{Minimize} \quad \max_{0<k\le LN} |S_k(\Phi)| \\ &\text{subject to} \quad 0 \le \phi_m < 2\pi, m = 1, \ldots, M \end{aligned} \tag{35}$$

Suppose $\hat{\phi}_m$ to be the global optimal solution to this problem. Unfortunately, there appears to be no simple way to obtain $\hat{\phi}_m$ analytically. For coherent demodulation, it is necessary to send $\hat{\phi}_m$ to the receiver as side information. When $\hat{\phi}_m$ is a continuous value, an infinite number of bits will be required as side information. The solution to this problem is to limit $\hat{\phi}_m$ to a level from a finite number of predetermined levels (quantization). For differential demodulation, it is not necessary to send $\hat{\phi}_m$ to the receiver, but $M - 1$ subcarriers at the subblock boundaries have to be set aside as reference carriers.

The phase factors are restricted to a finite set of values and hence (35) is approximated by the problem

$$\begin{aligned} &\text{Minimize} \quad \max_{0<k\le LN} |S_k(\Phi)| \\ &\text{subject to} \quad \phi_m \in \left\{ \frac{2\pi l}{W} | l = 0, \ldots, W-1 \right\} \end{aligned} \tag{36}$$

If the number of rotation angles W is "sufficiently" large, the solution of (36) will approach that of (35). Furthermore, ϕ_1 can be fixed without any performance loss. Now, there are only $M - 1$ free variables to be optimized and hence W^{M-1} distinct phase vectors, Φ_i, need to be tested. As such, (36) is solved using W^{M-1} iterations; the ith iteration involves computing LN signal samples, each of which is denoted by $S_k(\Phi_i)$, using (34) and choosing the maximum $|S_k(\Phi_i)|$ value. At the end of each iteration, the phase vector is retained if the current value of max $|S_k(\Phi_i)|$ is less than the previous maximum. The phase vector that is retained after all the iterations are completed will be an approximation to the global optimal solution of (35).

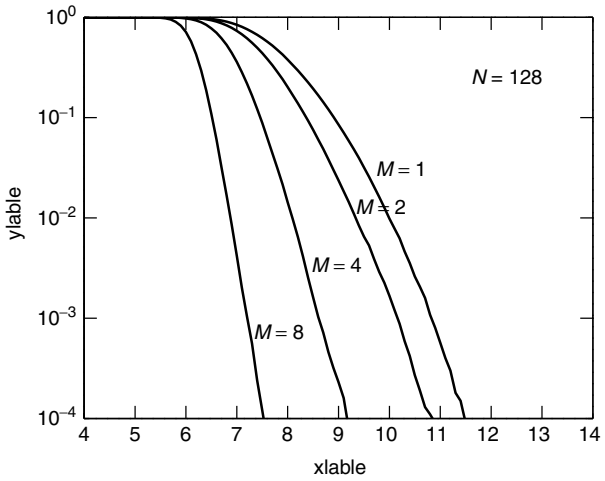


Figure 4. PTS performance for several M for $N = 128$ QPSK subcarriers.

In most reported studies, $W = 2$. In some cases, the use of more rotation angles ($W > 2$) has been found to yield diminishing returns. Figure 4 shows the PAR distribution for this method for varying M, with normal OFDM being $M = 1$. For $M = 8$, the PAR can be reduced by about 4 dB at a probability of 10^{-6}.

5.1.2. Selected Mapping (SLM) Approach. For a given M-PSK sequence, one generates M independent M-PSK sequences by multiplying by M fixed vectors and choosing the sequence with lowest PAR for transmission. This method is simple and very impressive. However, it needs M FFTs to select the best sequence among L. Suppose that we have M fixed phase vectors $\underline{P}_k = (p_k^0, p_k^1, \ldots, p_k^{N-1})$ for $k = 1, \ldots, M$, where $p_k^n \in \{0, 1, \ldots, M-1\}$ for $\forall n, k$. Without loss of generality, $p_1^n = 0$ $\forall n$. For an input data sequence $\mathbf{a}$, we generate M independent' sequences

$$A_k = \mathbf{a} \oplus \underline{P}_k \quad k = 1, \ldots, M \tag{37}$$

where $\underline{u} \oplus \underline{v}$ is the componentwise modulo M addition of $\underline{u}$ and $\underline{v}$. Originally, it was suggested to use the following selection function [9,11]: transmit A_l for $1 \le l \le L$ if

$$\text{SF} = \begin{cases} \text{PAR}(A_l) & \text{Baüml} \\ |W_H - N|^2 + |R_1|^2 & \text{Van Eetvelt} \end{cases} \tag{38}$$

is minimized. Here

$$R_1 = \sum_{k=1}^{N-1} A_{l:k} A_{l:k+1}^*$$

and W_H is the binary Hamming weight of the length N binary sequence. The performance of Baüml's SF is quite good, but requires multiple FFTs per input data frame.

5.1.3. Interleaving Approach. In this approach $K - 1$ interleavers are used at the transmitter [12]. These interleavers produce $K - 1$ permuted frames of the input

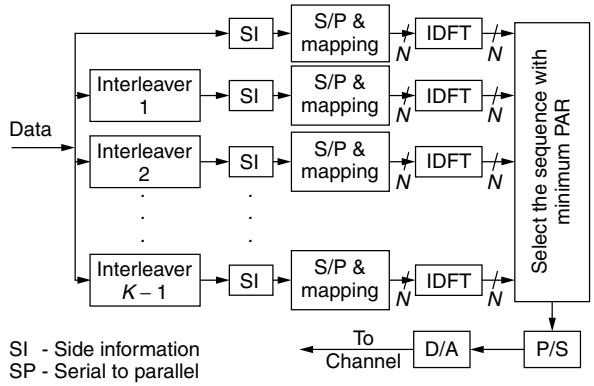


Figure 5. System model.

data before mapping into QPSK symbols. The 4 times oversampled IDFT of each frame (including the uncoded frame) is used to compute its PAR. The minimum PAR frame of all the K frames is selected for transmission. The identity of the corresponding interleaver is also sent to the receiver as side information. Figure 5 describes an OFDM transmitter with interleavers to reduce the PAR. The PAR reduction achievable with this method is similar to that of the PTS method.

5.2. Coding Techniques

5.2.1. Constructing Sequence Families with Low PAR and High Distance. Some definitions: A M-ary code $\mathcal{C}$ is a given set of sequences of symbols where each symbol is chosen from a set $F_M = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ of M distinct elements. The set F_M is often taken to be the set $Z_M = \{0, 1, 2, \ldots, M-1\}$, with $M = 2^h$ for positive integer h. We will denote a codeword as an N-tuple $(b_0, \ldots, b_{N-1})$ by $\mathbf{b}$. The Hamming distance between two sequences or codewords is defined as

$$d_H(\mathbf{a}, \mathbf{b}) = \sum_{n=0}^{N-1} \delta(a_n - b_n)$$

where $\delta(\cdot)$ is the Krönecker delta function. A critical parameter of a code $\mathcal{C}$ is the minimum Hamming distance, or just minimum distance, which measures how good it is at error-correcting. This is defined to be the smallest of the distances between any two distinct codewords:

$$d_{\min} = \min\{d_H(\mathbf{a}, \mathbf{b}) | \mathbf{a}, \mathbf{b} \in \mathcal{C}, \mathbf{a} \neq \mathbf{b}\}$$

The PAR of a code $\mathcal{C}$ is defined as

$$\mathrm{PAR}_{\max} = \max\{\mathrm{PAR}(\mathbf{a}) | \forall \mathbf{a} \in \mathcal{C}\}$$

Finally, the code rate of a code $\mathcal{C}$ is defined as

$$R = \frac{\log_2 \#(\mathcal{C})}{N \log_2 M}$$

where $\#(S)$ denotes the number of elements of set S. The following notation will be used where necessary. An $[n, k, d_{\min}, \eta]$ code is a code of length n, containing k information symbols, with minimum distance $d_{\min}$ and $\mathrm{PAR}_{\max} \eta$.

5.2.2. Golay Complementary Sequences and Reed–Muller Codes. One of the earliest low-PAR code constructions was that proposed by Jones et al. [13], with parameters [4,3,2,1.75]. This was based on a table of low-PAR codewords. Several authors noted that low-PAR codes can be constructed using complementary sequences. A pair of sequences $\mathbf{a}$ and $\mathbf{b}$ is complementary if

$$\rho_{\mathbf{a}}(k) + \rho_{\mathbf{b}}(k) = 2\delta(k) \qquad k = 0, 1, \ldots, N-1$$

Taking the Fourier transform of this, we have

$$P_{\mathbf{a}}(t) + P_{\mathbf{b}}(t) = 2.$$

Hence it is clear the PAR of $\mathbf{a}$ and $\mathbf{b}$ must be less than or equal to 2. Recently, Davis and Jedwab made a significant breakthrough by identifying the relationship between complementary sequences and Reed–Muller codes [14]. Further important results have been found [7,15]. The rth-order Reed–Muller code $\mathrm{RM}(r, m)$ has length $n = 2^m$, minimum Hamming distance $d = 2^{m-r}$, and the number of information bits $k = \sum_{i=0}^{r} \binom{m}{i}$. For example, RM(1,5) is (32,6,16), a low rate linear code.

5.2.3. Constructing Single Sequences with Low PAR. Although OFDM requires the transmission of a large family of sequences with low PAR, it is helpful to also consider the special case where the family size is 1. These single sequences can form the kernel of larger families of sequences with low PAR. The periodic autocorrelation (PA) of a length N sequence, $\mathbf{a}$, is given by

$$\rho_p(k) = \sum_{n=0}^{N-1} a_{n+k} a_n, \quad \text{for } k = 0, 1, \ldots, N-1$$

where all indices are taken, mod N. We can upper bound the PAR of the N-point DFT of $\mathbf{a}$ using,

$$\mathrm{PAR}_p(\mathbf{a}) \leq 1 + \frac{2}{N} \sum_{k=1}^{(N-1)/2} |\rho_p(k)|, \; N \text{ odd} \tag{39}$$

and a similar expression for N even.

Similarly, the negaperiodic autocorrelation (NA) of a length N sequence, $\mathbf{a}$, is given by

$$\rho_n(k) = \sum_{n=0}^{N-1} (-1)^{\left\lfloor \frac{n+k}{N} \right\rfloor} a_{n+k} a_n, \quad \text{for } k = 0, 1, \ldots, N-1$$

where all indices are taken, mod N. We can upper bound the PAR of the N-point negaperiodic DFT of $\mathbf{a}$ using

$$\mathrm{PAR}_n(\mathbf{a}) \leq 1 + \frac{2}{N} \sum_{k=1}^{(N-1)/2} |\rho_n(k)|, \; N \text{ odd} \tag{40}$$

and a similar expression for N even.

Figure 6 shows the continuous power spectrum of an m-sequence, where the N periodic DFT points are bounded by the PA, and interleaved with the N negaperiodic DFT points that are bounded by the NA. There are numerous constructions for sequences with low PA in the literature (e.g., m sequences, trace sequences, Legendre sequences,

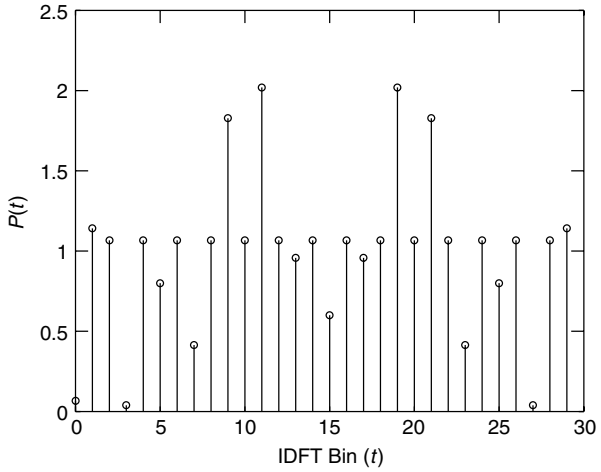


Figure 6. 30-point IDFT power spectrum for length $N = 15$ binary m-seq: 001001101011110.

and cyclotomic constructions). For instance, a binary m sequence guarantees $\rho_p(k) = -1, \forall k, k \neq 0$. Substituting back into (39) gives an upper bound on PAR_p (**a**) of $\dfrac{2N-1}{N}$, which is not particularly tight as the true PAR_p of a binary m sequence is $\dfrac{N+1}{N}$. Sequences with low PA are often proposed for code-division multiple access, but the spectral power peak of a sequence in between N IDFT points can rise considerably. Figure 6 illustrates this fact for a length 15 m sequence which has a low N-point IDFT, but not such a low $2N$-point IDFT. The periodic and negaperiodic N-point IDFTs are bins 0, 2, 4, ..., 26, 28, and bins 1, 3, ..., 27, 29, respectively, of Fig. 6.

Unlike the periodic case, the NA of a sequence has not been studied in such great detail. This is partly because there is a certain overlap between construction techniques for sequences with low PA and low NA, respectively. For instance, for BPSK, if **a** has odd length and low PA, then $\mathbf{a} \oplus 010101\cdots$ has equally low NA. However, the constructions are distinct for the even-length case. The APA and associated PAR upper bound were given in Eqs. (9) and (10). The APA can be viewed as the sum of PA and NA as follows:

$$\rho(k) = \frac{\rho_p(k) + \rho_n(k)}{2}, 0 \leq k < N,$$

$$\rho(k) = \frac{\rho_p(N-k) - \rho_n(N-k)}{2}, -N < k < 0$$

It follows that if a sequence has low PA *and* low NA, then it has low APA and a low upper bound on PAR. Finding constructions that have both low PA and low NA is a difficult problem.

Sequences with low APA are often parameterized by their merit factor (MF), where

$$\mathrm{MF}(\mathbf{a}) = \frac{N^2}{2\sum_{k=1}^{N-1} |\rho(k)|^2} \tag{41}$$

and high MF implies low APA. Finding sequences with highest MF is closely related to the APA problem, and has been studied by a few authors [16,17]. The BPSK sequence with highest known MF is of length 13 and has an MF of 14.08. This sequence has $\max_{k>0} |\rho(k)| = \max_{k>0} |\rho_p(k)| = \max_{k>0} |\rho_n(k)| = 1$. In words, the optimum periodic and negaperiodic properties of **a** guarantee its aperiodic optimality. Although BPSK sequences have been found with MF $\simeq$ 9.0 up to length 117, these sequences are the result of optimized computer search. *Very few* infinite constructions for high MF sequences are known. The best-known are the offset Legendre, twin prime, and (modified) Jacobi constructions. These constructions generate sequences with optimally low PA, moderately low NA, and with MF $\to$ 6.0 as $N \to \infty$, and this is the highest asymptote known for BPSK. A more recent Legendre-type infinite construction has generated sequences with optimally low NA, moderately low PA, and again with MF $\to$ 6.0 as $N \to \infty$. In contrast, both the m sequence and length 2^m complementary sequence have an asymptotic MF of 3.0 [17]. Finally we note that if an infinite construction could be found such that $|\rho_p(k)| \leq c_p, |\rho_n(k)| \leq c_n$, where c_p, c_n are constants, and $k > 0$, then, as $N \to \infty$, the construction would have asymptotically infinite MF. The existence of such a sequence construction is extremely unlikely.

5.2.4. Golay–Davis–Jedwab Codes. Many of the early code proposals were comprehensively generalized by Davis and Jedwab when they proposed an infinite family of binary Golay complementary sequences with parameters $[2^m, \log_2(m!) + m, 2^{m-2}, 2.0]$, and defined them as certain Reed–Muller (RM) RM(2, m) cosets of RM(1, m). We call this family *DJ*, where *DJ* comprises codewords, $c(\mathbf{x})$, with algebraic normal form

$$c(\mathbf{x}) = \sum_{i=0}^{m-2} x_{\pi(i)} x_{\pi(i+1)} \oplus \sum_{i=0}^{m-1} g_i x_i \oplus h\mathbf{1}$$

where $g_i, h \in (0, 1)$, $\mathbf{1}$ is the all-ones vector, π is a permutation of the integers $\{0, 1, \ldots, m-1\}$ and the x_i are binary variables representing length 2^m binary sequences such that element t of x_i is $\left\lfloor \dfrac{t}{2^i} \right\rfloor$, mod 2. The DJ construction was also generalized to higher alphabets and to PARs that are a multiple of 2. The construction is optimal for low N. For instance, for $m = 3$ and binary sequences we can construct an optimal $[8, \log_2(48), 2, 2.0]$ DJ code. There are only 16 more sequences with PAR $\leq$ 2.0, which are not included in the DJ set, and the inclusion of any of these sequences would reduce d from 2 to 1. Unfortunately the rate, k/N, of the DJ construction vanishes rapidly for $N > 32$. Therefore the DJ construction is practically useful only in the context of OFDM for systems requiring no more than 32 subcarriers. It remains an open problem to discover low PAR error-correcting code constructions for $N > 32$ with an acceptable rate. Unfortunately, many OFDM systems require anything from 8 to 8192 subcarriers.

5.2.5. Rudin–Shapiro Recursion and Its Generalizations. The DJ construction can be viewed as Rudin–Shapiro recursion [18]. Let length N sequences, $\mathbf{a_i}$ and $\mathbf{b_1}$ satisfy $P_{\mathbf{a}}(t) + P_{\mathbf{b}}(t) \leq Q$ so that both **a** and **b** have PAR $\leq Q$. Then

it can be shown that sequence concatenations $\mathbf{a_{i+1}} = \mathbf{a_i} \mid \mathbf{b_i}$ and $\mathbf{b_{i+1}} = \mathbf{a_i}|\overline{\mathbf{b_i}}$ both have PAR $\leq Q$, where "|" means concatenation, and "$\overline{*}$" means negation of every component of *. When $Q = 2$, the recursive definition, along with swapping the places of $\mathbf{a}$ and $\mathbf{b}$, and varying the position of the negation, gives the complete DJ binary codeset. This is the case when $\mathbf{a_0}, \mathbf{b_0} \in \{1, -1\}$. More generally, $\mathbf{a_0}$ and $\mathbf{b_0}$ can be any length N_1 pair of sequences, where we wish Q to be as close to 2.0 as possible. Then we can construct a Rudin–Shapiro-type code from this 'seed pair' with parameters $[N_1 2^m, \log_2(m!) + m, d_s 2^{m-2}, Q]$, where $d_s = \min(d_H(\mathbf{a_0}, \mathbf{b_0}), d_H(\mathbf{a_0}, \overline{\mathbf{b_0}}))$.

For N_1 small these codes are roughly the same size as the DJ code with marginally worse PAR. Moreover one can form the union of codes constructed this way, where the distance of the combined codeset is dependent on the set of seed pairs, and the PAR is governed by the constituent code with highest Q. However, as Rudin–Shapiro recursion generates a quadratic extension of an arbitrary-degree seed, the rate of the construction still vanishes as N increases, as quadratics constitute a vanishingly small part of the whole space of 2^N sequences. We can devise higher-degree constructions by viewing the parameters of the Rudin–Shapiro construction as arising from orthogonality of the matrix $\mathbf{R} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, where $\mathbf{v_{i+1}} = \begin{pmatrix} \mathbf{a_{i+1}} \\ \mathbf{b_{i+1}} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \mathbf{a_i} \\ \mathbf{b_i} \end{pmatrix}$, where addition is replaced by concatenation (more generally *tensor sum*). Let $[N_i, k_i, d_i, \eta_i]$ be the code parameters at iteration i of the recursion. Let r and c be the number of rows and columns of the associated matrix, $\mathbf{R}$, respectively, and δ be the number of rows and columns of the associated matrix, $\mathbf{R}$, respectively, and δ be the Hamming distance between rows of $\mathbf{R}$. Then, using matrix $\mathbf{R}$ along with permutation and negation to define the recursion, we get

$$\begin{aligned} N_{i+1} &= cN_i & k_{i+1} &= c\tbinom{r}{c}k_i \\ d_{i+1} &= \min(\delta(N_i - d_i) + (c-\delta)d_i, \delta d_i \\ &\quad + (c-\delta)(N - d_i)) & \eta_{i+1} &= f(\eta_i, i) \end{aligned}$$

where $N_0 = T, k_0 = r, d_0 = d_T, \eta_0 = \eta_T$, and where f is some function determined by the closeness to orthogonality of the matrix $\mathbf{R}$. If $\mathbf{R}$ is orthogonal then $f(\eta_i, i) = \eta_i$. The rate can be improved (as is the case with the DJ code) at the price of distance by including all tensor permutations of the construction. It is possible to increase the rate still further, at the price of distance, by including further permutations of v_i in between iterations. Thus there are a vast number of currently unexplored recursive constructions based on orthogonal and near-orthogonal matrices of various dimensions. Near-orthogonal constructions are of particular interest as they provide a moderate increase in PAR while also providing a large number of rows to maintain rate as N increases.

5.3. Other Techniques

The tone reservation approach has been proposed for the reduction of PAR of OFDM signals by Tellado and Cioffi [19]. In this method, both the transmitter and receiver agree on reserving a small subset of tones for generating PAR reduction signals. The transmitter does not send data on these reserved tones. The complex baseband signal may now be represented as

$$s(t) = \sum_{k \in I_{\text{info}}}^{N-1} c_k e^{j2\pi k \Delta f t} + \sum_{k \in I_{\text{tones}}}^{N-1} b_k e^{j2\pi k \Delta f t}, 0 \leq t \leq T \qquad (42)$$

where I_{info} and I_{tones} are two disjoint tone-index sets such that

$$I_{\text{info}} \cup I_{\text{tones}} = \{0, 1, \ldots, N-1\}$$

The values $b_k, k \in I_{\text{tones}}$, will be called PAR reduction tones (PRTs). The redundancy for this method is

$$R = \frac{|I_{\text{tones}}|}{N} \qquad (43)$$

Parallel combinatory OFDM [20] reduces the PAR without reducing the bandwidth efficiency and without increasing the bit error probability. An OFDM system with N subcarriers using q-PSK can transmit q^N different OFDM symbols. PC-OFDM is based on expanding the q-PSK signal constellation with one extra, zero-amplitude, point. With this expanded signal constellation, the number of different OFDM symbols increases to $(q+1)^N$. A subset of these modified OFDM symbols may be chosen with lower PAR. The authors show that this method will have at least the same bandwidth efficiency, and lower bit error probability, when compared to the original OFDM system.

Another PAR reduction method, and the simplest, is to deliberately clip the OFDM signal before amplification. In particular, since the large peaks occur with very low probability, clipping could be an effective technique for PAR reduction. However, clipping is a nonlinear process and may cause significant in-band distortion, which increases the bit error rate, and out-of-band noise, which reduces the spectral efficiency. Filtering after clipping can reduce the spectral splatter but may also cause some peak regrowth. Peak regrowth can however be reduced by oversampling the OFDM signal and clipping [21,22].

Henkel and Wagner [23] develop a trellis shaping technique for PAR reduction. In this method, a valid code sequence of a convolutional code is added (modulo 2) to a data sequence. The code sequence is chosen to reduce the PAR. If H is the parity-check matrix of the convolutional code, then for a valid code sequence y we have $yH^T = 0$. This property can be used to eliminate the added code sequence at the receiver side. However, it is necessary to precode the information sequence with the left inverse $(H^T)^{-1}$.

BIOGRAPHIES

C. Tellambura received his B.Sc. degree with honors from the University of Moratuwa, Sri Lanka, in 1986, his M.Sc. in electronics from the King's College, UK, in 1988, and his Ph.D. in electrical engineering from the University of Victoria, Canada, in 1993. He was a postdoctoral research fellow with the University of Victoria and the University of Bradford. Currently, he is a senior lecturer at Monash University, Australia. He is an editor for the *IEEE Transactions on Communications* and the *IEEE Journal on Selected Areas in Communications* (Wireless

Communications Series). His research interests include coding, communications theory, modulation, equalization, and wireless communications.

Matthew G. Parker received a B.Sc. in electrical and electronic engineering in 1982 from University of Manchester Institute of Science and Technology, U.K. and, in 1995, a Ph.D. in residue and polynomial residue number systems from University of Huddersfield, U.K. From 1995 to 1998 he was a postdoctoral researcher in the Telecommunications Research Group at the University of Bradford, U.K., researching into coding for peak factor reduction in OFDM systems. Since 1998 he has been working as a postdoctoral researcher with the Coding and Cryptography Group at the University of Bergen, Norway. He has published on residue number systems, number-theoretic transforms, complementary sequences, sequence design, quantum entanglement, coding for peak power reduction, factor graphs, linear cryptanalysis, and VLSI implementation of Modular arithmetic.

BIBLIOGRAPHY

1. J. A. C. Bingham, Multicarrier modulation for data transmission: An idea whose time has come, *IEEE Commun. Mag.* 5–14 (1990).
2. P. Shelswell, The COFDM modulation system: the heart of digital audio broadcasting, *Electron. Commun. Eng. J.* 127–136 (June 1995).
3. G. K. H. Sari and I. Jeanclaude, Transmission techniques for digital terrestrial TV broadcasting, *IEEE Commun. Mag.* **33**: 100–109 (Feb. 1995).
4. S. B. Weinstein and P. M. Ebert, Data transmission by frequency division multiplexing using the discrete Fourier transform, *IEEE Trans. Commun.* **19**: 628–634 (Oct. 1971).
5. C. Reiners and H. Rohling, *Multicarrier transmission technique in cellular mobile communications systems*, *Proc. IEEE Vehicular Technology Conf.* IEEE, 1994, pp. 1645–1649.
6. M. R. Schroeder, Synthesis of low-peak-factor signals and binary sequences with low autocorrelation, *IEEE Trans. Inform. Theory* **IT-13**: 85–89 (1970).
7. K. G. Paterson, Generalized Reed-Muller codes and power control in OFDM modulation, *IEEE Trans. Inform. Theory* **46**: 104–120 (Jan. 2000).
8. C. Tellambura, Upper bound on the peak factor of N-multiple carriers, *IEE Electron. Lett.* **33**: 1608–1609 (Sept. 1997).
9. P. Van Eetvelt, G. Wade, and M. Tomlinson, Peak to average power reduction for OFDM schemes by selective scrambling, *IEE Electron. Lett.* **32**: 1963–1964 (Oct. 1996).
10. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed., Academic Press, 1994.
11. R. W. Bäuml, R. F. H. Fischer, and J. B. Huber, Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping, *IEE Electron. Lett.* **32**: 2056–2057 (Oct. 1996).
12. A. D. S. Jayalath and C. Tellambura, Reducing the peak-to-average power ratio of an OFDM signal through bit or symbol interleaving, *IEE Electron. Lett.* **36**: 1161–1163 (June 2000).
13. A. E. Jones, T. A. Wilkinson, and S. K. Barton, Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes, *IEE Electron. Lett.* **30**: 2098–2099 (1994).
14. J. A. Davis and J. Jedwab, Peak-to-mean power control in OFDM, Golay complementary sequences, and Reed-Muller codes, *IEEE Trans. Inform. Theory* **45**: 2397–2417 (Nov. 1999).
15. K. G. Paterson and V. Tarokh, On the existence and construction of good codes with low peak-to-average power ratio, *IEEE Trans. Inform. Theory* **46**: 1974–1987 (Sept. 2000).
16. M. J. E. Golay, A new search for skewsymmetric binary sequences with optimal merit factors, *IEEE Trans. Inform. Theory* **36**: 1163–1166 (Sept. 1990).
17. T. Høholdt, *Difference Sets, Sequences and Their Correlation Properties*, vol. 542 of *Series C: Mathematical and Physical Sciences*, Kluwer, 1999.
18. T. Høholdt, H. E. Jensen, and J. Justesen, Aperiodic correlations and the merit factor of a class of binary sequences, *IEEE Trans. Inform. Theory* **31**: 549–552 (July 1985).
19. J. Tellado, *Multicarrier Modulation with Low PAR*, Kluwer, 2000.
20. P. K. Frenger and N. A. B. Sevensson, Parallel combinatory OFDM signalling, *IEEE Trans. Commun.* **47**: 558–567 (April 1999).
21. X. Li and L. J. Cimini, Jr., Effects of clipping and filtering on the performance of OFDM, *IEEE Commun. Lett.* **2**(5): 131–133 (1998).
22. H. Ochiai and H. Imai, Performance of the deliberate clipping with adaptive symbol selection for strictly band-limited OFDM systems, *IEEE J. Select. Areas Commun.* **18**: 2270–2277 (Nov. 2000).
23. W. Henkel and B. Wagner, Another application of trellis shaping: PAR reduction for DMT (OFDM), *IEEE Trans. Commun.* **48**: 1471–1476 (Sept. 2000).

PERMUTATION CODES

EMANUELE VITERBO
Politechnico di Torino
Torino (Turin), Italy

1. INTRODUCTION

Permutation codes were proposed by David Slepian in 1965 [12]. In the quest for efficient codes for the band-limited Gaussian channel, permutation codes are among the first attempts to solve the problem taking into account both coding gains and decoding efficiency. Permutation codes are multidimensional spherical signal constellations with the desirable property that they possess a very simple maximum-likelihood (ML) decoding algorithm. Slepian used the term *permutation modulation* for his permutation codes.

A *variant I* permutation modulation is the set of codewords obtained by taking all permutations of an initial vector in the n-dimensional Euclidean space. A *variant II* permutation modulation is the set of codewords obtained by taking all permutations and sign changes of the components of an initial vector in the n-dimensional Euclidean space. Trivial examples of variant I and variant II modulations are orthogonal and biorthogonal codes,

respectively. Also pulse code modulation (PCM), pulse position modulation (PPM), and simplex codes can be viewed as permutation modulations.

Good permutation modulations may be designed by appropriately selecting the initial vector. Permutation modulations may be very efficiently decoded by essentially applying a sorting algorithm to the received signal vector. Permutation modulations are very special cases of group codes for the band-limited channel proposed by Slepian [10,13].

Karlof later proposed the term *permutation code* for a generalization of permutation modulations [6]. In particular, he studied the group codes obtained from subgroups of the symmetric group (the group of permutations of n objects). The resulting spherical codes may be seen as particular subsets of the corresponding permutation modulation. We can think of permutation modulation as a "full rate" permutation code. In this case the initial vector selection problem and the decoding algorithm are much harder [6–8].

Here, we will focus on permutation codes according to Slepian's definition. This article is organized as follows. The next section introduces the notation and gives detailed definition of permutation modulation. Performance in terms of error probability and the ML decoding algorithm are also presented. Section 3 gives some examples of permutation codes and discusses some application issues.

2. THEORY

Digital transmission over the band-limited additive white Gaussian noise (AWGN) channel is commonly modeled in the n-dimensional Euclidean space $\mathbf{R}^n$ as $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where $\mathbf{y}$ is the received signal vector, $\mathbf{x}$ is the transmitted signal vector (or codeword) taken from a finite signal constellation (or codebook) $\mathcal{S}$, and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is a real Gaussian random vector with i.i.d. (independent, identically distributed) components. The space dimension n is related to the time–bandwidth product through the sampling theorem, namely, if T is the signal duration and W the occupied bandwidth, then $n = 2WT$.

Letting $M = |\mathcal{S}|$ be the number of points in the constellation, we define the spectral efficiency as

$$\frac{R}{W} = \frac{2}{n}\log_2 M \text{ bps/Hz} \tag{1}$$

Let $r = \log_2 M$. If we are interested in transmitting binary information, we simply label $2^{[r]}$ codewords by distinct binary vectors of length $[r]$ and disregard the remaining $M - 2^{[r]}$ codewords.[1]

The average signal power of $\mathcal{S} = \{\mathbf{x}_i\}_0^{M-1}$ is given by

$$\mathcal{P} = \frac{1}{nM}\sum_{i=0}^{M-1}\|\mathbf{x}_i\|^2 \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbf{R}^n$.

The maximum-likelihood (ML) receiver gives an estimate $\hat{\mathbf{x}}$ of the transmitted codeword $\mathbf{x}$ according to the minimum distance criterion

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}_i \in \mathcal{S}} \|\mathbf{y} - \mathbf{x}_i\|^2 \tag{3}$$

[1] Where $[x]$ denotes the greatest integer smaller than x.

We note that the complexity of the ML receiver depends greatly on the structure of the code $\mathcal{S}$. In the worst case a total of M Euclidean distances must be computed. For large values of M this may be impractical; hence it is common to trade some of the performance for a reduced decoding complexity. Many classical forward error-correcting (FEC) codes have been selected for applications because they have simple decoding algorithms.

The average codeword error probability with ML detection is given by

$$P(e) = \frac{1}{M}\sum_{i=0}^{M-1} P(e \mid \mathbf{x}_i) = \frac{1}{M}\sum_{i=0}^{M-1}\int_{\overline{\mathcal{R}_i}} \frac{e^{-\|\mathbf{x}-\mathbf{x}_i\|^2/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}}\,d\mathbf{x} \tag{4}$$

where $\overline{\mathcal{R}_i} = \mathbf{R}^n \backslash \mathcal{R}_i$ is the complement of the decision region corresponding to the codeword $\mathbf{x}_i$, defined as

$$\mathcal{R}_i = \{\mathbf{z} \in \mathbf{R}^n : \|\mathbf{z} - \mathbf{x}_i\| < \|\mathbf{z} - \mathbf{x}_j\|, \quad \forall j \neq i\} \tag{5}$$

These regions are also known as *minimum-distance* or ML regions.

Good codes should be designed in order to minimize $P(e)$, given the parameters M, n, $\mathcal{P}$, and σ. Shannon showed that the codeword error probability decreases exponentially with n and gave the famous asymptotic result that, for given R, W, and $\mathcal{P}/\sigma^2$, the $P(e)$ can be made arbitrarily small as $n \to \infty$, provided that $R < C$, where

$$C = W\log_2\left(1 + \frac{\mathcal{P}}{\sigma^2}\right) \tag{6}$$

On the contrary, if $R > C$, then $P(e) \to 1$ as $n \to \infty$.

Explicit construction of optimal codes is an open problem and has been analyzed in some very special cases only [1,15].

2.1. Definitions

Let $\{\mu_1, \ldots, \mu_k\}$ be a set of distinct real numbers with $\mu_1 < \mu_2 < \cdots < \mu_k$, and let $\{m_1, \ldots, m_k\}$ be a set of positive integers such that

$$n = \sum_{j=1}^{k} m_j \tag{7}$$

Consider the initial vector with components sorted in ascending order:

$$\mathbf{x}_0 = (\underbrace{\mu_1, \ldots, \mu_1}_{m_1}, \underbrace{\mu_2, \ldots, \mu_2}_{m_2}, \ldots, \underbrace{\mu_k, \ldots, \mu_k}_{m_k}) \tag{8}$$

A variant I permutation code consists of the set of vectors obtained by permuting the components of the initial vector $\mathbf{x}_0$. The total number of codewords in such a code is

$$M_I = \frac{n!}{m_1!m_2!\cdots m_k!} \tag{9}$$

The variant I code with $k = 2$, $m_1 = n - 1$, $m_2 = 1$, and $\mu_1 = 0$ is the well-known PPM or orthogonal modulation.

A variant II permutation code consists of the set of vectors obtained by permuting and applying all possible sign changes to the components of the initial vector $\mathbf{x}_0$. Without loss of generality, we may assume

$$0 \leq \mu_1 < \mu_2 < \cdots < \mu_k \tag{10}$$

The total number of codewords in this code is

$$M_{\mathrm{II}} = \frac{2^h n!}{m_1! m_2! \cdots m_k!} \tag{11}$$

where $h = n - m_1$, if $\mu_1 = 0$ and $h = n$, if $\mu_1 > 0$.

The variant II code with $k = 2$, $m_1 = n - 1$, $m_2 = 1$, and $\mu_1 = 0$ results in the well-known biorthogonal modulation. The variant II code with $k = 1$, $m_1 = n$, and $\mu_1 \neq 0$ yields an n-bit PCM. In this case the points of $\mathcal{S}$ correspond to the 2^n vertices of an n-dimensional hypercube of edge length $2\mu_1$.

It is clear that all codewords of both variant I and II codes lie on a hypersphere of radius $\sqrt{n\mathcal{P}}$ centered at the origin and

$$\mathcal{P} = \frac{1}{n} \sum_{j=1}^{k} m_j \mu_j^2 \tag{12}$$

2.2. Decoding

Let us consider ML decoding of variant I codes. We need to find the minimum of the quantities

$$\|\mathbf{y} - \mathbf{x}_i\|^2 = \|\mathbf{y}\|^2 + \|\mathbf{x}_i\|^2 - 2(\mathbf{y}, \mathbf{x}_i) = \|\mathbf{y}\|^2 + n\mathcal{P} - 2(\mathbf{y}, \mathbf{x}_i), \quad i = 0, \ldots, M-1 \tag{13}$$

where $(\mathbf{y}, \mathbf{x}_i)$ denotes the scalar product of the two vectors. Since $\|\mathbf{y}\|$ is independent of i, the ML decoder may simply maximize the scalar product between the received vector and the codewords:

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \in \mathcal{S}} \sum_{k=1}^{n} x_k y_k \tag{14}$$

This maximization problem may be solved as follows. Given the received vector $\mathbf{y}$, replace the smallest m_1 components by the values μ_1, replace the smallest m_2 remaining components with μ_2, and so on until all the components have been replaced.

It is interesting to show the very simple and elegant proof of the optimality of this decoding algorithm given by Slepian. In particular, we want to show that the sum

$$x_{i_1} y_1 + x_{i_2} y_2 + \cdots + x_{i_n} y_n \tag{15}$$

is maximized by the permutation of indices $(i_1, i_2, \ldots, i_n)$, which pairs the largest x to the largest y, the second largest x to the second largest y, and so forth.

For $n = 1$ it is trivially true; then we proceed by induction. For some $n > 1$, let $\overline{x}$ and $\overline{y}$ denote the largest x and the largest y. If $\overline{x}$ is not paired with $\overline{y}$ in (15), then the sum contains the two terms $\overline{x}y' + \overline{y}x'$ for some $x' \leq \overline{x}$ and $y' \leq \overline{y}$. If we swap x' with $\overline{x}$, the sum (15) will decrease; in fact

$$(\overline{x}\overline{y} + x'y') - (\overline{x}y' + \overline{y}x') = (\overline{x} - x')(\overline{y} - y') \geq 0 \tag{16}$$

Hence, pairing $\overline{x}$ with $\overline{y}$ maximizes the sum (15). We now delete $\overline{x}\overline{y}$ from the sum and proceed by induction on the $n - 1$ terms; pairing the second largest x to the second largest y does not reduce the $n - 1$ term sum, and so on.

For variant II codes ML decoding can be performed as follows:

1. Take the absolute value of the components of the received vector $\mathbf{y}$; that is, let
$$\mathbf{y}' = (|y_1|, |y_2|, \ldots, |y_n|)$$
2. Apply the decoder of variant I codes to $\mathbf{y}'$ to make a first decision $\mathbf{x}'$.
3. The final decision is given by
$$\hat{\mathbf{x}} = (\mathrm{sgn}(y_1)x'_1, \mathrm{sgn}(y_2)x'_2, \ldots, \mathrm{sgn}(y_n)x'_n)$$
where $\mathrm{sgn}(x) = +1$, if $x \geq 0$ and $\mathrm{sgn}(x) = -1$, if $x < 0$.

It can be shown that this algorithm is equivalent to solving the maximization problem (14).

The complexity of these decoding algorithms is rather small if compared to the brute-force exhaustive search. In particular, it is enough to perform a sorting algorithm on the n components of the received vector and to keep track of the final index permutation. This permutation uniquely identifies the ML decoded codeword, and the corresponding information bit label may be easily recovered. Sorting can be performed with a complexity of $O(n \log(n))$, whereas exhaustive decoding requires Mn multiplications and $M(n-1)$ additions.

2.3. Decision Regions and Error Probability

Evaluation of the average codeword error probability [4] for permutation codes can be simplified by the following argument. Consider the collection $\mathcal{C}$ of all $n \times n$ permutation matrices, that is, matrices having a single entry equal to one in each row and column and zeros in the remaining positions. When a permutation matrix is applied to the vectors of the codebook of a variant I code, it simply maps the codebook back into itself. In $\mathcal{C}$ we can find a permutation matrix A_{ij} that maps any codeword $\mathbf{x}_i$ into the codeword $\mathbf{x}_j$.

The permutation matrices are also orthogonal matrices so that, when they operate on $\mathcal{S}$, they preserve the distances between the points. Since the decision regions are defined in (5) in terms of distances, the permutation matrix A_{ij} also sends $\mathcal{R}_i$ into $\mathcal{R}_j$. Thus all the decision regions of a variant I code are congruent, and (4) reduces to $P(e) = P(e \mid \mathbf{x}_i)$, which is independent of i.

For variant II codes, a similar argument also enables us to conclude that $P(e) = P(e \mid \mathbf{x}_i)$ is independent of i. In particular, it is enough to replace the collection $\mathcal{C}$ by the collection $\mathcal{O}$ of $2^n n!$, $n \times n$ matrices having a single nonzero entry equal to $+1$ or -1 in each row and column.

We note that this simplification is similar to the one that can be used in evaluating the codeword error probability of linear codes, where it is convenient to consider the case where the all-zero codeword is transmitted. For permutation codes we will focus on $P(e \mid \mathbf{x}_0)$.

Let us now consider in detail the average codeword error probability of variant I codes. First observe that the received vector components have independent Gaussian distributions. The first m_1 components have mean μ_1 and variance σ^2, the next m_2 components have mean μ_2 and variance σ^2, and so on.

Assume that the codeword $\mathbf{x}_0$ was transmitted. To understand when a decoding error appears, let us split the received vector components into k runs of length m_j each:

$$\mathbf{y} = (y_1^{(1)}, y_2^{(1)}, \ldots, y_{m_1}^{(1)}, y_1^{(2)}, y_2^{(2)}, \ldots, y_{m_2}^{(2)}, \ldots, y_1^{(k)}, y_2^{(k)}, \ldots, y_{m_k}^{(k)}) \tag{17}$$

The correct decision will be made if the first m_1 components are smaller than the following m_2 components, the next m_2 components are smaller than the following m_3 components, and so forth. Then we can write

$$P_{\mathrm{I}}(e) = P_{\mathrm{I}}(\mu_1, \mu_2, \ldots, \mu_k) = 1 - P\{\eta_1 \le \xi_2 \le \eta_2 \le \xi_3 \le \cdots \le \eta_{k-1} \le \xi_k\} \tag{18}$$

where, for $j = 1, \ldots, k$

$$\xi_j = \min(y_1^{(j)}, y_2^{(j)}, \ldots, y_{m_j}^{(j)})$$

$$\eta_j = \max(y_1^{(j)}, y_2^{(j)}, \ldots, y_{m_j}^{(j)})$$

and the n independent Gaussian random variables $y_i^{(j)}$ have mean μ_j and variance σ^2.

Note that $P_{\mathrm{I}}(e)$ depends only on the differences of the μ values; thus, for all δ

$$P_{\mathrm{I}}(\mu_1 + \delta, \mu_2 + \delta, \ldots, \mu_k + \delta) = P_{\mathrm{I}}(\mu_1, \mu_2, \ldots, \mu_k) \tag{20}$$

Let $\beta_1 = \mu_1$ and $\beta_i = \mu_i - \mu_{i-1}$, for $i = 2, \ldots, k$. Let

$$\phi(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/(2\sigma^2)} \tag{21}$$

be the probability distribution function of a zero-mean Gaussian random variable with variance σ^2 and

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)\, dz \tag{22}$$

the corresponding cumulative distribution function.

Slepian [12] has shown that for both variants I and II it is possible to bound $P(e)$ as

$$B - \frac{B^2}{2} \le P(e) \le B \tag{23}$$

For variant I codes, we obtain

$$B = B_{\mathrm{I}} = \sum_{i=2}^{k} P_{m_i, m_{i-1}}(\beta_i) \tag{24}$$

and

$$P_{m,n}(\alpha) = P_{n,m}(\alpha) = n \int_{-\infty}^{\infty} \phi(z)[1 - \Phi(z)]^{n-1} \times [1 - \Phi^m(z + \alpha)]\, dz \tag{25}$$

For variant II, when $\mu_1 = 0$, we have

$$B = B_{\mathrm{II}} = \overline{P}_{m_1, m_2}(\beta_i) + \sum_{i=3}^{k} P_{m_i, m_{i-1}}(\beta_i) \tag{26}$$

where

$$\overline{P}_{m,n}(\alpha) = 2m \int_{0}^{\infty} \phi(z)[1 - 2\Phi(-z)]^{m-1} \times \{1 - [1 - \Phi(z - \alpha)]^n\}\, dz \tag{27}$$

and when $\mu_1 \neq 0$

$$B = B_{\mathrm{II}'} = 1 - [1 - \Phi(-\beta_1)]^{m_1} + \sum_{i=2}^{k} P_{m_i, m_{i-1}}(\beta_i) \tag{28}$$

3. EVALUATION

We are now able to consider the problem of selecting good permutation codes for a fixed dimension n. We want to optimize the choice of k, the μ values, and the m values. From (1), (9), and (11) we see that R is independent of the μ values and depends only on k and the m values.

It is natural to fix k and the m values (i.e., fix R) and choose the μ values to minimize $\mathcal{P}$ for some fixed value of $P(e)$.

For variant I codes, the first optimization step to reduce the average signal power is to center the signal constellation $\mathcal{S}$ around its barycenter, by selecting

$$\sum_{j=1}^{k} m_j \mu_j = 0 \tag{29}$$

By imposing this condition to PPM, we obtain the simplex modulation. For variant II codes, $\mathcal{S}$ is already centered around its barycenter.

The optimization problem has been solved numerically [12], and some optimal codes are presented for various n, m values, and k for two values of $P(e) = 10^{-3}$ and $P(e) = 10^{-5}$. A simplified version of the code optimization problem was solved analytically by Biglieri and Elia [2] and independently by Ingemarson [9]. They selected the μ values in order to maximize the minimum Euclidean distance $d_{\min}$, among the points of $\mathcal{S}$, for a fixed average power $\mathcal{P}$. Here, we report the optimal codes for $P(e) = 10^{-5}$ in Table 1.

Figures 1–4 show the codeword error probability of the codes in Table 1 as a function of E_b/N_0, where $E_b = n\mathcal{P}/r$ and $N_0 = 2\sigma^2$. These figures may be interpreted as follows. Given the system constraints T (maximum acceptable

Table 1. Optimized Codes at $P(e) = 10^{-5}$ Found by Slepian [12]

No.	n	m Values	μ Values	$[r]$	γ_{dB}
1	5	I (4,1)	−1.2908, 5.1631	2	1.6
2	5	II (4,1)	0, 6.6558	3	2.2
3	5	II (3,2)	0, 6.7720	5	1.2
4	5	II (2,3)	0, 6.7720	6	0.2
5	5	II (4,1)	4.6540, 11.4152	7	−1.4
6	5	II (2,2,1)	0, 6.7748, 13.3491	7	−2.0
7	5	II (1,3,1)	0, 6.6712, 13.4039	8	−2.3
8	10	I (9,1)	−0.6691, 6.0220	3	2.7
9	10	II (9,1)	0, 6.8907	4	3.3
10	10	II (7,3)	0, 7.1240	9	2.2
11	10	II (4,5,1)	0, 7.1747, 14.1284	16	−0.6
12	10	II (3,5,2)	0, 7.1293, 14.1828	18	−1.6
13	10	II (3,3,3,1)	0, 7.0814, 14.0512, 21.0215	21	−3.7
14	10	II (2,4,2,2)	0, 7.0672, 14.0365, 20.9950	22	−4.4
15	7	II (6,1)	0, 6.7720	3	2.8
16	7	II (5,2)	0, 6.9169	6	2.0
17	25	II (14,11)	0, 7.6409	33	1.8
18	50	II (25,22,3)	0, 8.0046, 15.7001	83	0.6
19	50	II (20,25,5)	0, 7.9966, 15.7951	92	0.0
20	51	II (19,24,6,2)	0, 8.0255, 15.8067, 23.2291	105	−1.5
21	50	II (17,23,8,2)	0, 7.9926, 15.8064, 23.3280	108	−1.7
22	7	II (4,3)	0, 6.9694	8	1.3
23	31	II (23,8)	0, 7.6873	30	2.9

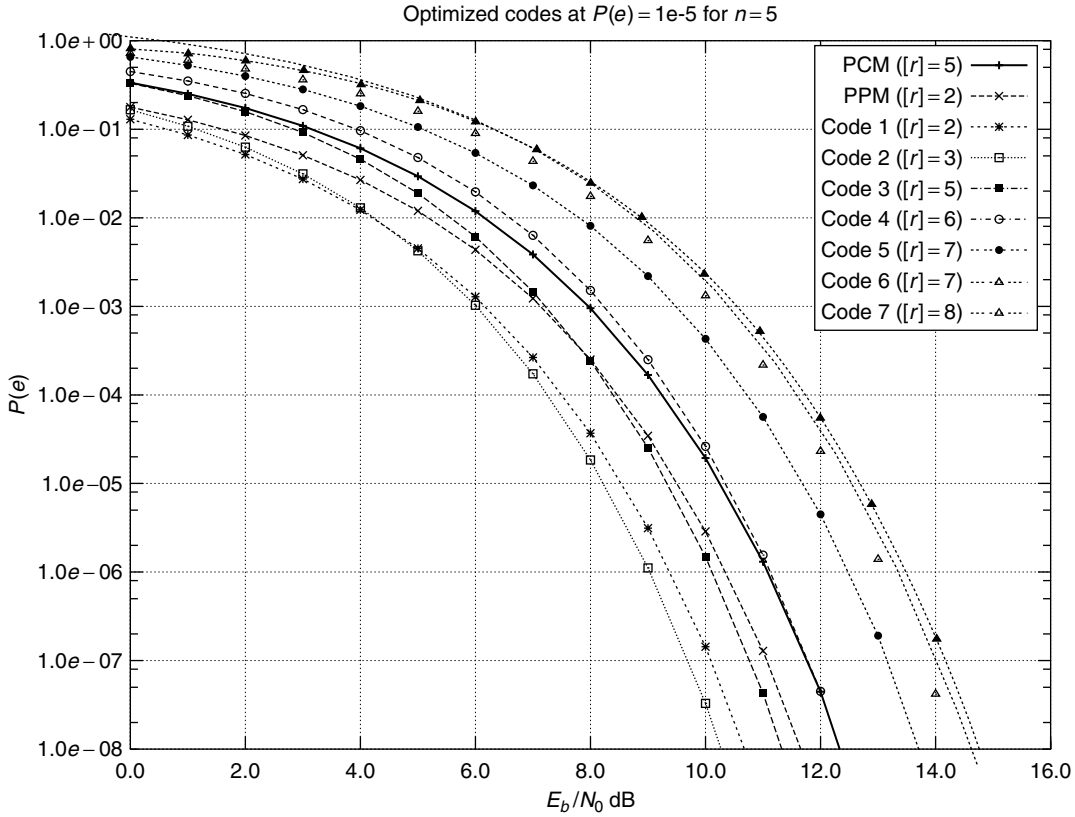


Figure 1. Optimized codes at $P(e) = 10^{-5}$ for $n = 5$.

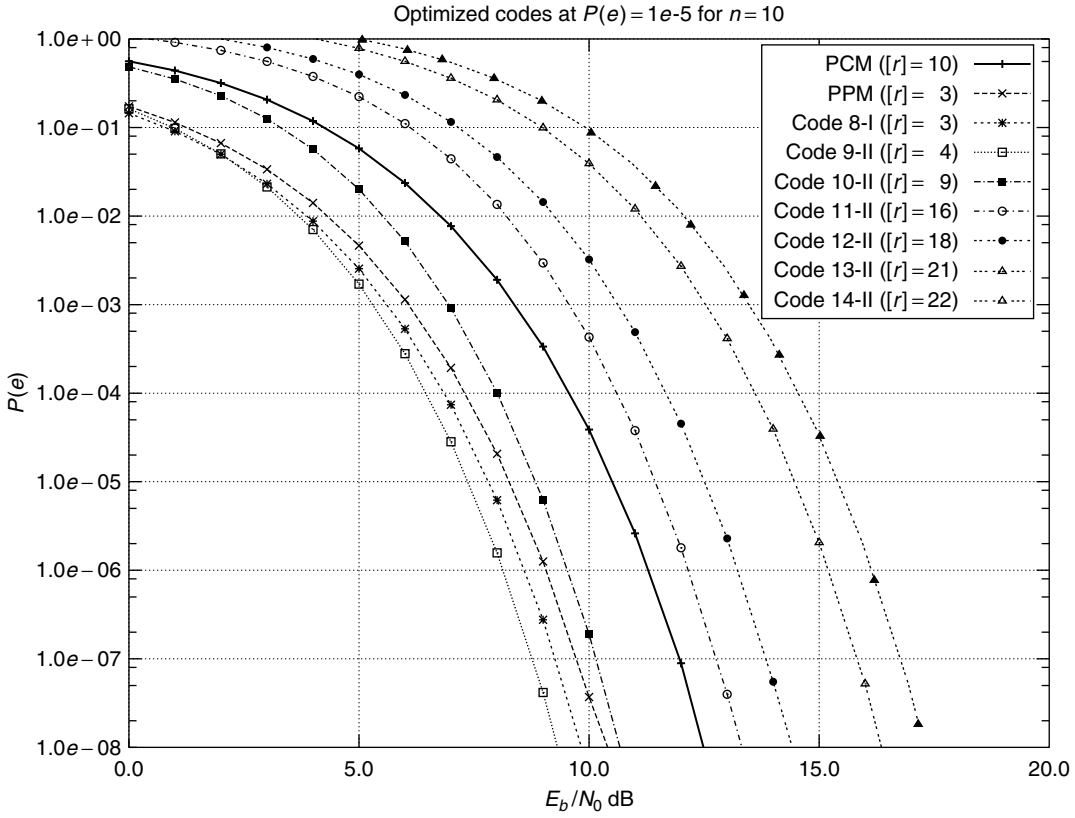


Figure 2. Optimized codes at $P(e) = 10^{-5}$ for $n = 10$.

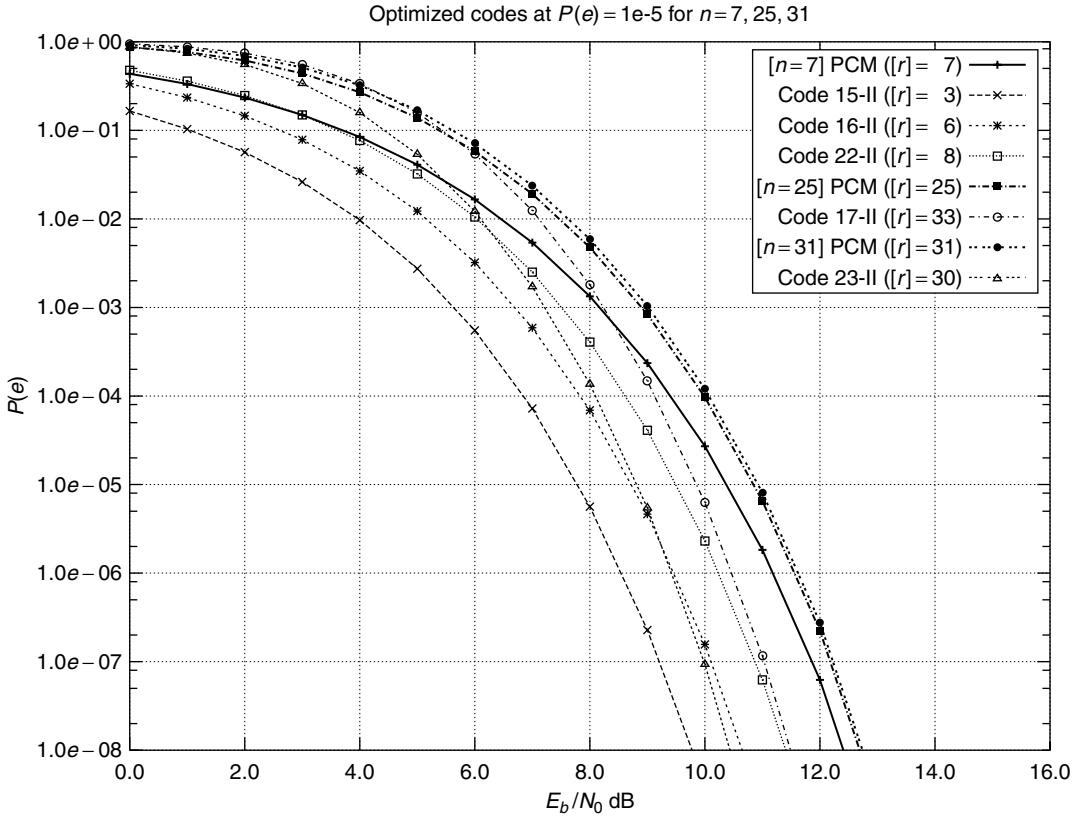


Figure 3. Optimized codes at $P(e) = 10^{-5}$ for $n = 7, 25, 31$.

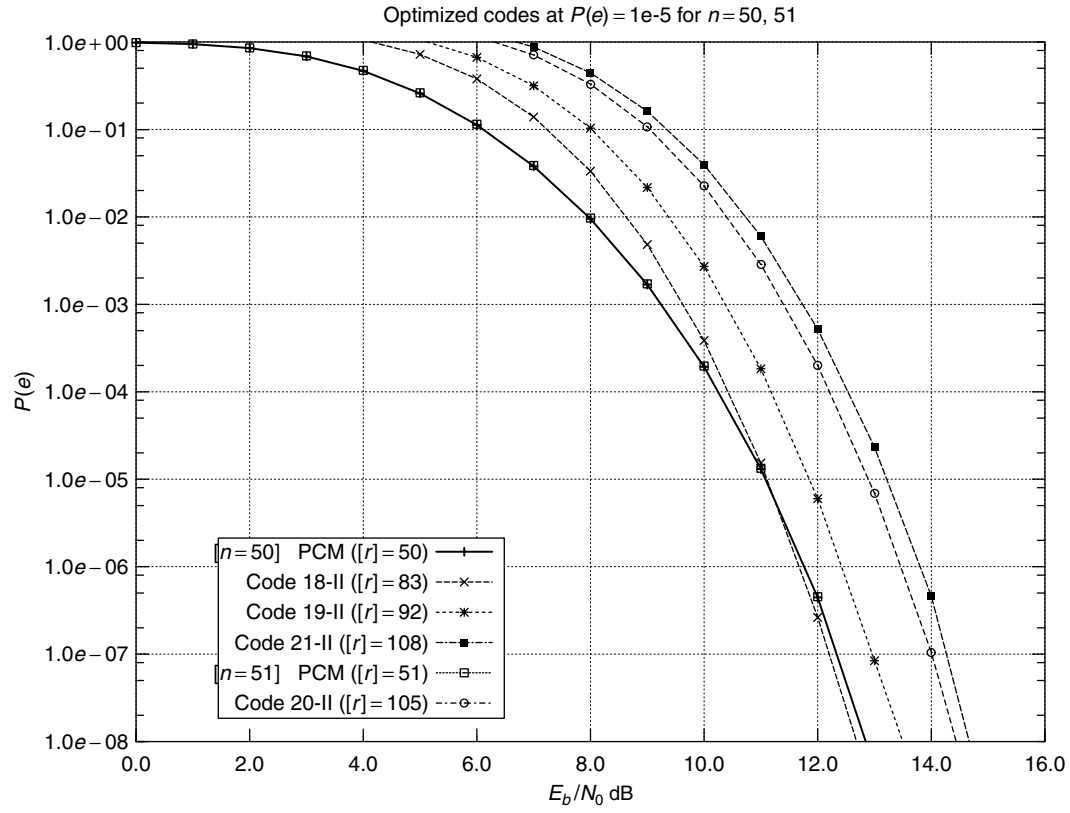


Figure 4. Optimized codes at $P(e) = 10^{-5}$ for $n = 50, 51$.

delay) and W (available bandwidth), we have n; then we can chose among the codes of different rate the one that satisfies our packet error rate $P(e)$ requirement.

For example, code 3 for $n = 5$ enables us to transmit at the same rate of the 5-bit PCM with a lower $P(e)$, corresponding to an asymptotic gain of 1.2 dB (see Fig. 1). Code 10 enables us to transmit at 0.9 the rate of PCM with an asymptotic gain of 2.2 dB (see Fig. 2). Code 23 enables us to transmit at almost at the same rate of PCM with an asymptotic gain of 2.9 dB (see Fig. 3).

Figure 5 shows the performance of the optimal codes given by Slepian [12] at $P(e) = 10^{-3}$. Comparison with Fig. 1 shows that their performance in almost identical in both cases, so we may conclude that the optimization is quite insensitive to the value $P(e)$.

Given an n-dimensional code, we define its asymptotic coding gain with respect to an n bit PCM as

$$\gamma = \frac{d^2_{\min}/E_b}{d^2_{\min,\mathrm{PCM}}/E_{b,\mathrm{PCM}}} \tag{30}$$

We report γ in decibels in the last column of Table 1. These asymptotic values can be approximately verified in all figures at $P(e) = 10^{-8}$, with an accuracy of about 0.5 dB.

We conclude with a few comments on the possible application of permutation codes. In order to implement the transmitter, we need to define an orthonormal basis $\{\psi_j(t)\}_1^n$ of the signal space so that each transmitted signal is given by

$$x(t) = \sum_{j=1}^{n} x_j \psi_j(t) \tag{31}$$

The basis functions must be approximately time- and band-limited. For example, we can select the strictly bandlimited functions

$$\psi_j(t) = \frac{\sin(2\pi Wt - j\pi)}{2\pi Wt - j\pi} \quad j = 1, \ldots, n \tag{32}$$

or the strictly time-limited functions for $0 \le t \le T$

$$\psi_{2j-1}(t) = \sin\left(\frac{2\pi jt}{T}\right)$$
$$\psi_{2j}(t) = \cos\left(\frac{2\pi jt}{T}\right) \quad j = 1, \ldots, \frac{n}{2} \tag{33}$$

Other choices are also possible [e.g., 3–5,11]. The receiver can be implemented using a bank of matched filters, matched to the basis functions in order to obtain the components of the received vector $\mathbf{y}$.

Although permutation codes have a very long history and some very promising coding gains, to the author's knowledge, they have never been used in applications. Nevertheless, we can be expect that they may be exploited in the future for high-speed transmission, due to their very simple decoding algorithm.

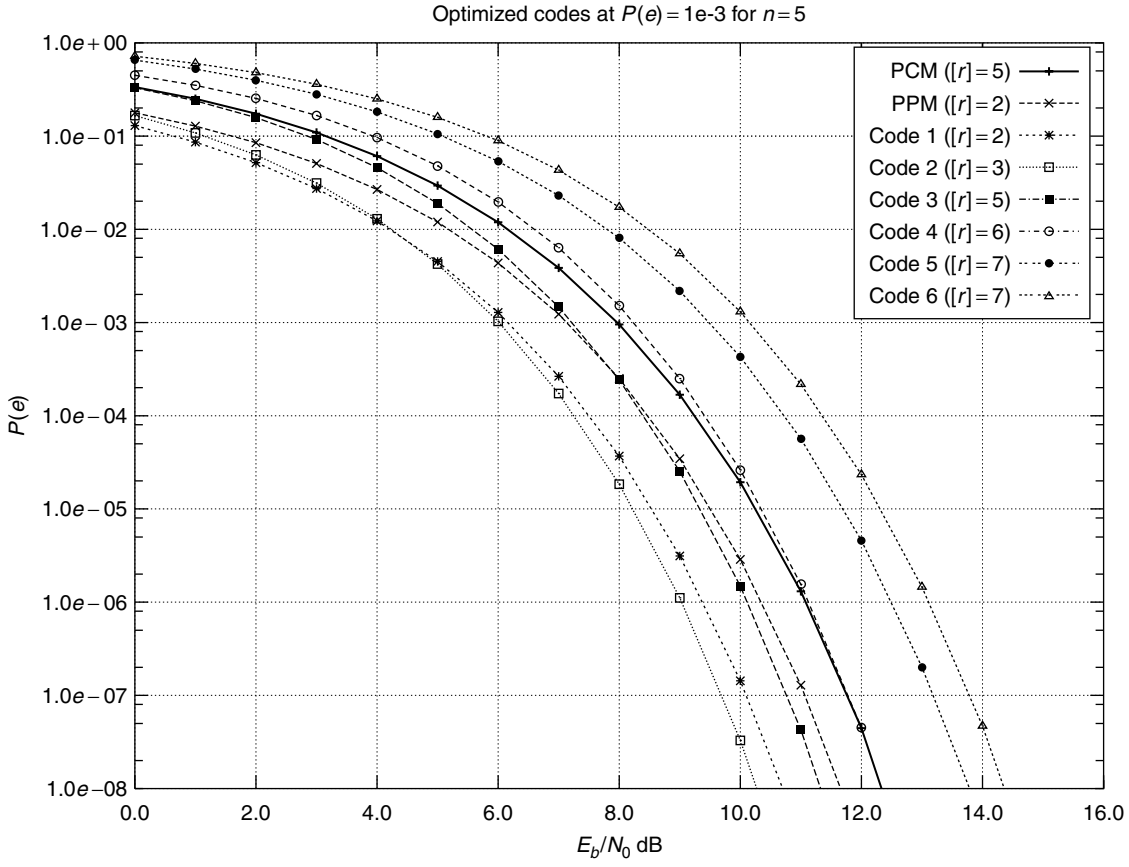


Figure 5. Optimized codes at $P(e) = 10^{-3}$ for $n = 5$.

BIOGRAPHY

Emanuele Viterbo was born in Torino (Turin), Italy, in 1966. He received his baccalaureate degree (Laurea) in Electrical Engineering in 1989 and his Ph.D. in 1995 in Electrical Engineering, both from the Politecnico of Torino, Torino, Italy. From 1990 to 1992 he was with the European Patent Office, The Hague, The Netherlands, as a patent examiner in the field of dynamic recording and in particular in the field of error-control coding. Between 1995 and 1997 he held a postdoctoral position in the Dipartimento di Elettronica of the Politecnico di Torino in Communications Techniques over Fading Channels. Between 1997 and 1998 he was Visiting Researcher in the Information Sciences Research Center of AT&T Research, Florham Park, New Jersey. Since 1998, he has been Assistant Professor at Politecnico di Torino, Dipartimento di Elettronica. Dr. Emanuele Viterbo was awarded a NATO Advanced Fellowship in 1997 from the Italian National Research Council. His current research interests are in lattice codes for the Gaussian and fading channels, algebraic coding theory, digital terrestrial television broadcasting, and digital magnetic recording.

BIBLIOGRAPHY

1. A. V. Balakrishnan, A contribution to the sphere-packing problem of communication theory, *J. Math. Anal. Appl.* **3**: 485–506 (Dec. 1961).
2. E. Biglieri and M. Elia, Optimum permutation modulation codes and their asymptotic performance, *IEEE Trans. Inform. Theory* 751–753 (Nov. 1976).
3. M. Elia, G. Taricco, and E. Viterbo, Optimal energy transfer over bandlimited communication channels, *IEEE Trans. Inform. Theory* **45**(6): 2020–2029 (Sept. 1999).
4. H. J. Landau and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty II, *Bell Syst. Tech. J.* **40**: 65–84 (1961).
5. H. J. Landau and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty III, *Bell Syst. Tech. J.* **41**: 1295–1336 (1962).
6. J. K. Karlof, Permutation codes for the Gaussian channel, *IEEE Trans. Inform. Theory* **35**(4): 726–732 (July 1989).
7. J. K. Karlof, Decoding spherical codes for the Gaussian channel, *IEEE Trans. Inform. Theory* **39**(1): 60–65 (Jan. 1993).
8. J. K. Karlof and Y. O. Chang, Optimal permutation codes for the Gaussian channel, *IEEE Trans. Inform. Theory* **35**(4): 726–732 (July 1989).
9. I. Ingemarsson, Optimized permutation modulation, *IEEE Trans. Inform. Theory* **36**(5): 1098–1100 (Sept. 1990).
10. D. Slepian, Some further theory of group codes, *Bell Syst. Tech. J.* 1219–1252 (Sept. 1960).
11. D. Slepian and H. O. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty I, *Bell Syst. Tech. J.* **40**: 43–64 (1961).
12. D. Slepian, Permutation modulation, *Proc. IEEE* 228–236 (March 1965).
13. D. Slepian, Group codes for the Gaussian channel, *Bell Syst. Tech. J.* **47**(4): 575–602 (April 1968).
14. D. Slepian, On neighbor distances and symmetry in group codes, *IEEE Trans. Inform. Theory* 630–632 (Sept. 1971).
15. M. Steiner, The strong simplex conjecture is false, *IEEE Trans. Inform. Theory* 721–731 (May 1994).

PHOTONIC ANALOG-TO-DIGITAL CONVERTERS

BARRY L. SHOOP
United States Military Academy
West Point, New York

1. INTRODUCTION

Analog-to-digital (A/D) conversion is the process by which an analog signal that is continuous in both time and amplitude is converted to a digital signal that is discrete in time and amplitude. The A/D converter is an important component in any electronic or photonic system that senses the natural environment and processes, stores, displays, or communicates the information using digital techniques. Since the vast majority of signals encountered in nature are analog and the preferred method of processing, storing, and transmitting signals is digital, this interface is generally considered to be the most critical and challenging part of the overall signal acquisition and processing system. Recent advances in both electronics and telecommunication markets as well as continued improvements in sensor resolution has renewed interest in the pursuit of high-speed, high-resolution A/D converters and has focused attention on photonic techniques to provide improvements in this technology area.

In general, the process of A/D conversion can be characterized by the four distinct functional blocks shown in Fig. 1. The analog input signal $x(t)$ is first band-limited to the range $0 \le f_x \le f_B$ (Hz) by a lowpass analog filter with cutoff frequency f_B to protect against aliasing that could occur during the subsequent sampling operation. The sampling operation in a conventional Nyquist rate A/D converter is chosen to satisfy the minimum Nyquist criterion: $f_S = f_N = 2f_B$, where f_S is the sampling frequency, f_N is the Nyquist frequency, and f_B is the constrained signal bandwidth. There are also other alternatives to sampling frequency depending on the specific application. Subsampling below the Nyquist rate is acceptable if the input signal is known to be periodic. Oversampling is another alternative in which $f_S \gg f_N$ and signal processing techniques are subsequently used to achieve improved amplitude resolution through a technique called *spectral noise shaping*. The output from the sampler is $x_n \equiv x(nT_S)$, where T_S is the uniform sampling period $T_S = 1/f_S$. Scalar quantization then maps each continuous-amplitude input x_n to one value in a discrete-amplitude ensemble $q_n \equiv q(nT_S)$. On the basis of the results of this mapping, the digital processor generates a digital codeword that most closely approximates the input analog signal value. The output $y_n \equiv y(nT_S)$ is then the multibit, digital word representing the input analog input value.

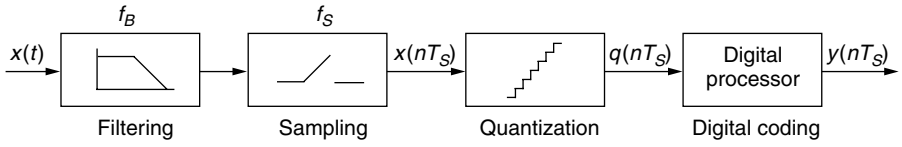


Figure 1. Generic analog-to-digital converter block diagram.

A complete coverage of the subject of photonic A/D conversion can be found in the 2001 book *Photonic Analog-to-Digital Conversion* [1]. For additional background and information on the general subject of A/D conversion, Refs. 2–5 provide an excellent overview.

1.1. Motivation for Photon-Based A/D Conversion

Photonic approaches to A/D conversion provide some distinct performance advantages over their electronic counterparts. Nonlinear optoelectronic devices have demonstrated ultrahigh switching speeds that can be applied to the quantization functionality required in the A/D converter. Mode-locked lasers provide high-speed and accurate optical pulses in the range of 0.5 ps–50 fs that can be used for precision clocks and sampling. These optical clocks provide the capability for ultrafast sampling with extremely low clock jitter. Another advantage specifically associated with optical sampling techniques is the decoupling of the sampled and sampling signals, achieved when the sampling signal is optical and the sampled signal is electronic. Mach–Zehnder interferometers that can be used for optical sampling and modulation have been demonstrated with modulation bandwidths of up to 120 GHz. Photonic techniques also bring the potential for utilization of the full two-dimensional (2D) nature of optics. Many other applications have successfully converted from temporal approaches to spatial approaches and, applying the 2D nature of optical processing, extended the performance of the specific application. Many current approaches to photonic A/D conversion are leveraging this higher dimensionality in an effort to extend converter performance bounds. Many photonic approaches to A/D conversion also produce output digital codes that are Gray codes directly, eliminating the need for additional hardware to produce these coding schemes.

2. HISTORICAL PERSPECTIVE

The application of photonic techniques to the problem of A/D conversion began with optical sampling. In 1970, Steigman and Kuizenga [6] first proposed the use of a mode-locked laser for optical sampling of an electronic signal. Later, in 1975, Taylor [7] applied optical sampling using mode-locked laser pulses to the first optical A/D converter. This optical A/D converter integrated Mach–Zehnder interferometers, avalanche photodetectors, and electronic comparators. In the early 1980s this basic approach was further developed by a group at the Massachusetts Institute of Technology (MIT) Lincoln Laboratory, demonstrating a 4-bit electrooptic A/D converter operating at a sampling frequency of 1 GHz [8]. Jalali and Xie extended this electrooptic A/D converter using a folding architecture [9].

In the early 1990s, Shoop and Goodman proposed the first optical approach to oversampling A/D conversion [10]. This work resulted in a proof-of-concept experimental demonstration of an optoelectronic oversampled A/D converter based on multiple quantum well modulators [11]. Later this work was extended to low-resolution photonic A/D conversion for digital image halftoning based on spatial oversampling and error diffusion using smart pixel technology [12]. A newer approach to photonic A/D has been developed that integrates spatial oversampling, an error diffusion neural algorithm, and spectral noise shaping for high-resolution A/D conversion applications [13,14]. Pace has investigated an alternate approach to photonic oversampled A/D conversion based on a fiber lattice accumulator [15,16].

In the late 1990s, there was renewed interest in photonic A/D conversion, particularly for high-speed A/D applications. The majority of these architectures relied on channelization or interleaving techniques that partition the wide-bandwidth input into N-parallel channels, each operating at $1/N$ of the original sampling rate. This approach allows integration of high-speed optical sampling and lower-speed conventional electronic quantization, taking advantage of each technology's strengths. Twichell and colleagues at MIT Lincoln Laboratory have applied phase-encoded optical sampling and a time-interleaving architecture to wideband photonic A/D conversion, demonstrating a 505-MS/s (megasample/second) system providing 8 bits of resolution [17,18]. Clark and colleagues at the Naval Research Laboratory have investigated time- and wavelength-interleaving for photonic A/D conversion [19,20]. Another interesting approach that falls within the context of a channelized architecture is based on time-stretch preprocessing using an optically dispersive element [21].

Since the mid-1970s, other approaches to photonic A/D converters have also been investigated. Most were novel approaches that were ultimately limited by speed, complexity, or resolution. The interested reader can find details of several of these other approaches to optical A/D conversion in the literature [22–29].

3. REPRESENTATIVE PHOTONIC A/D CONVERTER ARCHITECTURES

There are a variety of different approaches to photonic A/D conversion differing in the fundamental architectures as well as the photonic components used. The following examples of photonic A/D converters are not intended to be inclusive but rather provide an introduction to several representative approaches to the application of photonic architectures and devices to the problem of A/D conversion.

3.1. Electrooptic A/D Conversion

Probably the best known optical A/D conversion technique to date was developed by Taylor in 1975 [7]. He was the

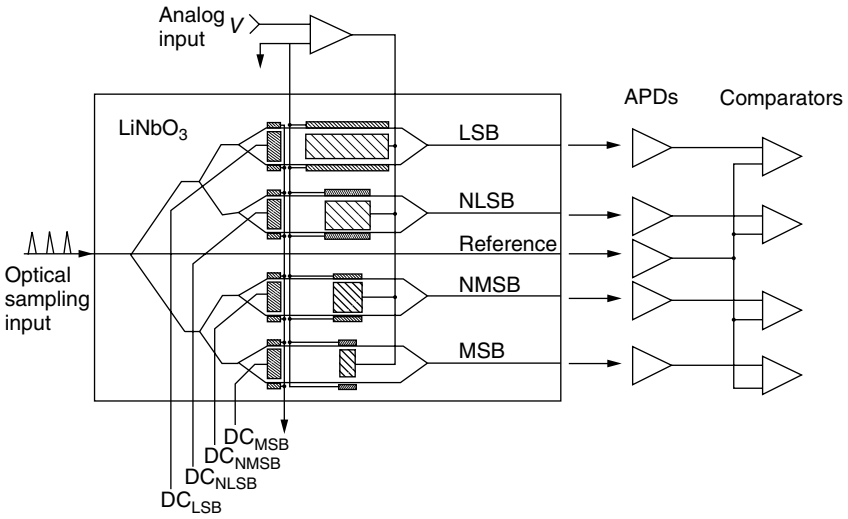


Figure 2. Schematic diagram of a 4-bit electrooptic A/D converter. (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 5.1, p. 124. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

first to recognize the relationship between the periodicity of the output of an interferometric electrooptic modulator with applied voltage and the periodic variation of a binary representation of an analog quantity. A 4-bit implementation of this concept is shown in Fig. 2 [8].

The basic optical element used in this architecture is a planar waveguide version of a Mach–Zehnder interferometric modulator. The interferometer consists of an electrooptic crystal containing a single-mode input optical waveguide that branches at a "Y" to split the optical power into two equal components. The light in the two paths then travels an equal distance before recombining at the second Y and exiting the crystal. The input analog voltage is applied to one arm of the interferometer through coplanar electrodes. In the absence of an applied electric field, the light from the two paths recombines in phase and produces a maximum in the output intensity. With an electric field applied to the electrode, the phase velocity of the light propagating in that arm is changed as a result of the linear electrooptic effect. The output intensity of a single interferometer can be shown to vary as

$$I = I_0 \cos^2\left(\frac{\phi}{2} + \frac{\psi}{2}\right) \tag{1}$$

where ψ is the static phase difference between the two paths and ϕ is the net electrooptic phase difference between the light propagating in the two guides

$$\phi = 2\pi L\left(\frac{\Delta n}{\lambda}\right) = kLV \tag{2}$$

Here, Δn is the refractive index change, V is the applied voltage, L is the modulator length, and k is a constant that depends on the electrooptic parameters of the crystal, the electrode spacing, and optical wavelength. The use of a three-electrode configuration, one between the two waveguides and two outside the waveguides produces an opposite phase shift in the two waveguides and therefore doubles the magnitude of the parameter k. An important parameter of a Mach–Zehnder interferometer is the voltage, which yields a phase shift of $\phi = \pi$

$$V_\pi = \frac{\pi}{kL} \tag{3}$$

In Fig. 2, the analog input signal V is applied in parallel to one arm of each of the four modulators, one for each bit of resolution. The sampling of the analog signal is performed optically, using a series of short optical pulses derived from a pulsed laser source. The optical output from each waveguide modulator is detected by an avalanche photodiode (APD) which converts the optical signal to an electronic signal and also provides amplification. The electronic signal from each modulator is then compared to a reference signal, obtained from the common light source. The output of each comparator is either a binary "1" or "0", depending on whether the modulator output intensity is greater or less than $I_0/2$, respectively. The output of the top modulator represents the least significant bit (LSB) in the digital word, and that of the bottom modulator is the most significant bit (MSB). The output intensity, the threshold, and the corresponding binary representation for each modulator are shown in Fig. 3. The Gray-code representation in Fig. 3 is achieved by controlling the static phase difference in each modulator by applying the appropriate DC biases, labeled as DC_{LSB}, DC_{NLSB}, DC_{NMSB}, and DC_{MSB} in Fig. 2.

In this Gray-code approach, the voltage quantization step size V_Q is equal to one-half the value of V_π for the LSB channel, or

$$V_Q = \frac{\pi}{2kL_N} \tag{4}$$

for an N-bit comparator where L_N is the electrode length for the LSB channel. For each subsequent significant bit, the value of V_π increases by a factor of 2 and therefore the electrode length L_n decreases by a factor of 2. Therefore

$$L_n = 2^{n-N} L_N \tag{5}$$

where $n = 1$ corresponds to the MSB and $n = N$ corresponds to the LSB. The maximum input analog voltage

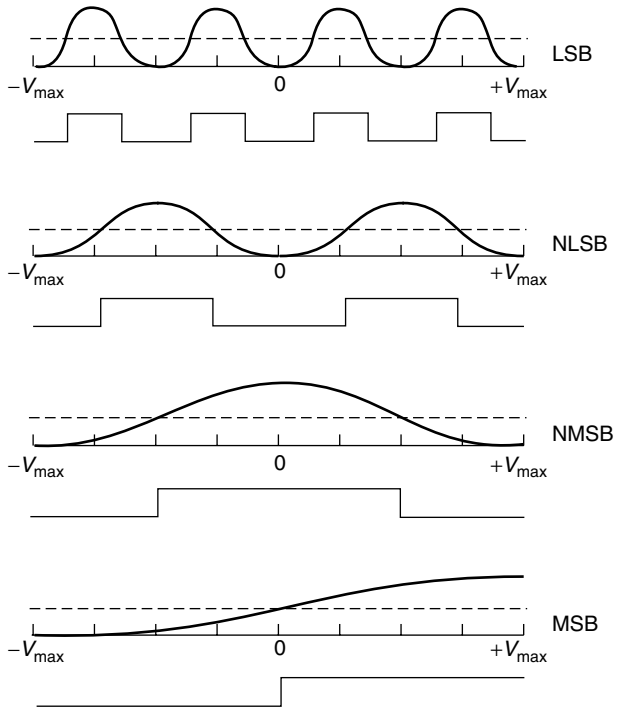


Figure 3. Intensity versus voltage for a 4-bit electro-optic A/D converter with a Gray-code output (the dashed-lines represent the threshold level in each of the four channels). (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 5.2, p. 125. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

that can be converted in this architecture is

$$V_{\max} = 2^{N-1} V_Q \tag{6}$$

This electrooptic A/D converter provides several distinct advantages. This photonic A/D converter is linear in complexity, requiring one additional Mach–Zehnder interferometer for each additional bit of resolution. Another important advantage is the decoupling of the analog sampled signal from the optical sampling signal. This eliminates the distortion effects common to diode bridge sampling circuits, which tend to couple the sampling signal into the converter circuitry. A limitation of this type of converter is that each additional bit of resolution requires a doubling of the electrode length of the LSB modulator. In $LiNbO_3$, this produces a transit-time limitation on performance of approximately 6 bits at 1 GHz. Other electrooptic crystals exist that produce larger refractive index changes and therefore could improve the performance of this transit-time limit. However, most of these crystals also have larger loss mechanisms and therefore would produce loss-limited performance instead.

3.1.1. An Optical Folding-Flash A/D Converter. Jalali and Xie [9] proposed an extension to the electrooptic A/D converter based on a folding architecture. This approach eliminates the geometrical scaling of V_π with amplitude resolution by incorporating an analog encoding technique. Figure 4 is a block diagram of a 4-bit optical folding-flash A/D converter. In contrast to the previous electrooptic A/D converter, the electrode lengths of all the Mach–Zehnder interferometers are identical. An analog encoding scheme is used in this approach to bias all these individual Mach–Zehnder interferometers at different points on the interferometer transfer characteristic.

The optical folding-flash A/D converter provides a solution to the V_π and subsequent electrode length scaling problem of the original electrooptic A/D converter; however, it also introduces some alternate challenges. In this approach, the MSB is susceptible to noise for extreme values of the input. Furthermore, this approach relies on the ability to accurately bias each interferometer in the architecture at several different points on the sinusoidal interferometer transfer characteristic. Hardware complexity is another limitation in this approach. The optical folding-flash A/D converter architecture is exponential in complexity with the number of interferometers

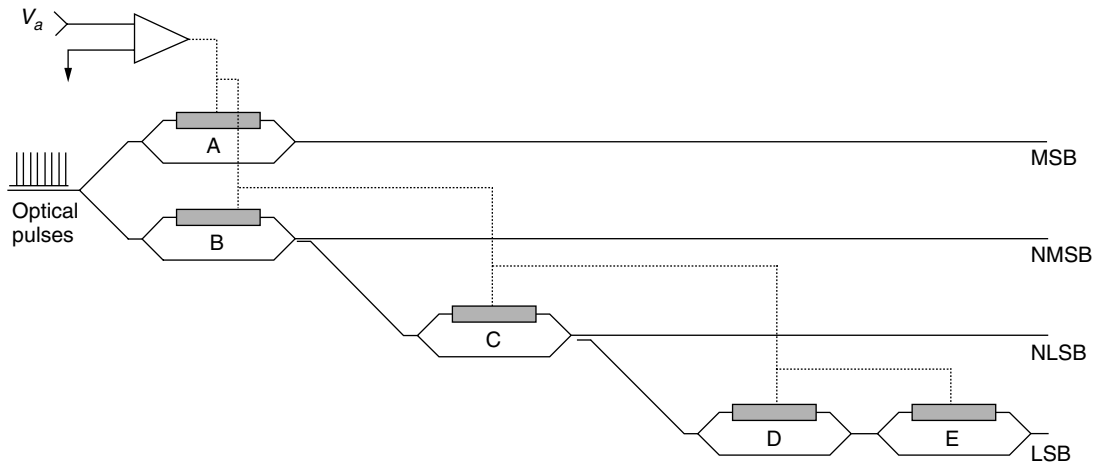


Figure 4. Block diagram of the optical folding-flash A/D converter. (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 5.3, p. 127. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

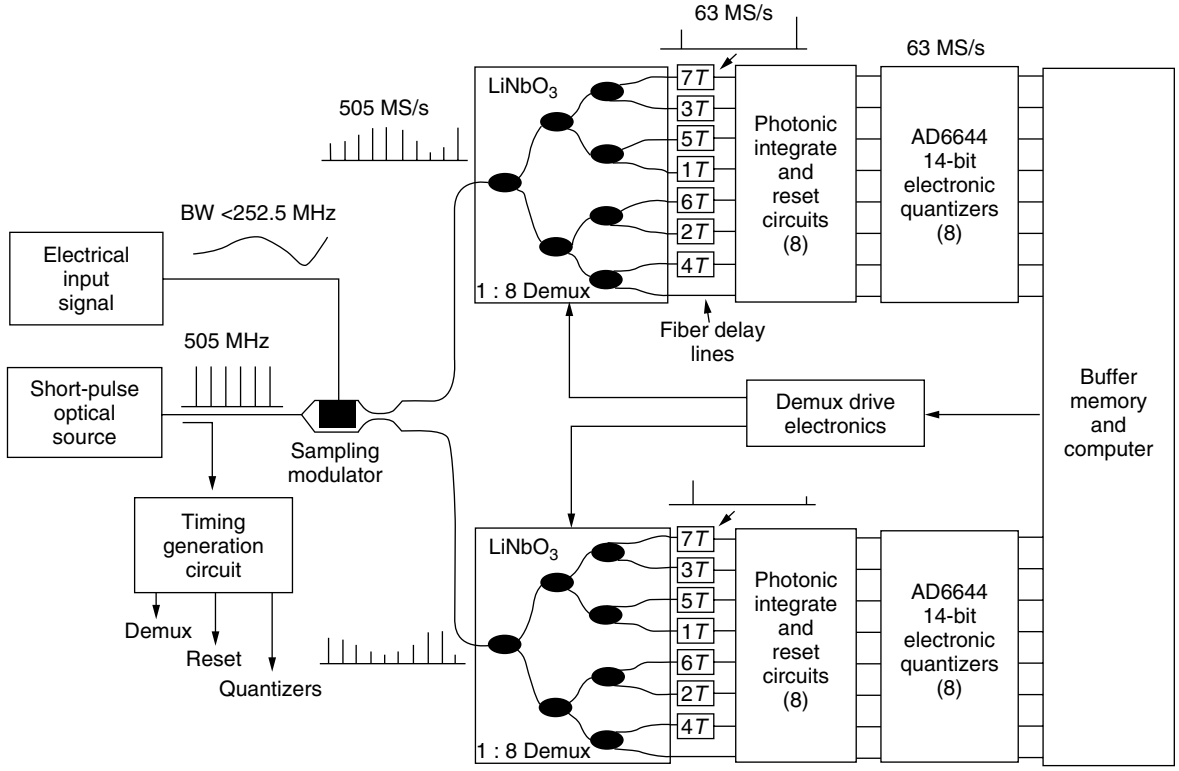


Figure 5. Block diagram of 505-MS/s optically sampled A/D converter.

scaling as $2^{(b-2)} + 1$, where b is the number of bits in the desired resolution.

3.2. Channelized Photonic A/D Conversion

Channelized or interleaved A/D converter architectures partition the input sampled signal into N-parallel channels, each of which operates at $1/N$ of the original sampling rate. In this approach, high-speed optical sampling can be integrated with N-electronic A/D converters in a common architecture that extends the bandwidth performance of the overall photonic A/D converter beyond that of the individual electronic A/D converters.

Mach–Zehnder interferometers can be used for wideband electrooptic modulation and as electrooptic switches for optical demultiplexing. One popular approach to achieving optical time interleaving is to use a tree-structured Mach–Zehnder switching architecture in which an array of Mach–Zehnder interferometric switches is used to provide the demultiplexing functionality. Using 1×2 electrooptic switches, each subsequent stage of the optical time-division demultiplexer is driven by a clock with a switching frequency that is one-half that of the previous stage. Unlike classic Mach–Zehnder interferometers that have a single output, the 1×2 electrooptic switch has a two-channel output. By modulating the refractive index of the interferometer, the output can be switched between the two output channels. Each output channel of the photonic demultiplexer can then be followed by a photodetector stage and an electronic quantizer stage to complete the time-interleaved A/D architecture.

An 8-channel time-interleaved photonic A/D converter operating at 505 MS/s has been demonstrated that employs a wideband electrooptic modulator and 1×2 electrooptic switches to accomplish the 1 : 8 optical demultiplexing [30]. Figure 5 is block diagram of this system [18]. Because the range over which linear modulation occurs in a Mach–Zehnder interferometer is relatively small, this architecture uses a new phase encoding technique [31] to produce linear modulation over a much wider dynamic range. This new approach combines the complimentary outputs from a dual-output electrooptic modulator to improve the linearity of the output.

In the architecture in Fig. 5, the optical source produces 25-ps pulses at the 505-MHz sampling rate that are used to sample the analog electrical input signal. Phase-encoded optical sampling is performed by the dual-output $LiNbO_3$ Mach–Zehnder modulator. The complimentary output of the sampling modulator is then fed to a pair of $LiNbO_3$ 1 : 8 optical time division demultiplexers to distribute the output pulsestreams to an array of photonic integrate and reset circuits and subsequently to electronic quantizers. In this architecture, the 14-bit electronic quantizers operate at 63 MS/s, which is one-eighth of the sampling rate. Fiber delay lines are used to time-align the pulses at the input to the photonic integrate and reset circuits to simplify timing.

The 1 : 8 optical demultiplexers were Ti-indiffused $LiNbO_3$ and employed a high-extinction design to minimize crosstalk between the parallel channels. Each demultiplexing stage consisted of a 1×2 electrooptic switch with an additional extinction modulator at each

output arm. The extinction for a single stage ranged between 36 and 46 dB. The bandwidth of a stage was 600 MHz and the insertion loss for a channel ranged between 6.8 and 8.4 dB. This photonic A/D architecture demonstrated SNR in the Nyquist bandwidth of 51 dB or 8.2 bits of resolution. The spur-free dynamic range (SFDR) was limited to 61 dB by channel-to-channel mismatch and crosstalk errors [32].

Although this approach holds promise for high-speed and high-resolution photonic A/D conversion, a number of technical challenges remain. Incomplete isolation between the individual demultiplexed channels can lead to channel-to-channel crosstalk, which, in turn, will result in undesirable spectral tones in the output frequency spectrum [33]. Consequently, each switch must be accurately designed and fabricated to ensure that optical crosstalk between channels is minimized. Also, the individual clock signals that control the switching of each electrooptic switch must be accurate and synchronous. Inaccuracies in clock timing between channels produce jitter in the demultiplexer which will also contribute to additional mismatch distortion.

3.3. Oversampling Photonic A/D Conversion

Oversampled A/D converters derive their name from the fact that the sampling rate is routinely chosen to be much higher than that required to satisfy the Nyquist criterion. This type of A/D converter trades bandwidth for improved amplitude resolution. In this type of converter, a low-resolution quantizer is embedded in a feedback architecture in an effort to reduce the quantization noise through spectral noise shaping. Here, a large error associated with a single sample is diffused over many subsequent samples and then linear filtering techniques are applied to remove the spectrally shaped noise, thereby improving the overall performance of the converter. Figure 6 shows a generalized block diagram of an oversampled A/D converter.

The analog signal $x(t)$ is again bandlimited to the range $0 \le f_x \le f_B(Hz)$ by an antialiasing filter and is then sampled at a rate $f_S \gg f_N$, where f_S is the sampling frequency, $f_N = 2f_x$ is the Nyquist frequency of the sampled signal, and $f_B \le f_S/2$ is the constrained signal bandwidth. The output of the sampler is then applied to a modulator that provides coarse amplitude quantization and spectral shaping of the quantization noise. The digital postprocessor, which consists of a digital lowpass filter (LPF) and a decimator, removes the quantization noise that was spectrally shaped by the modulator, provides antialiasing protection, and reduces the rate to the original sampled signal's Nyquist rate by trading word rate for word length.

3.3.1. The Modulator. The function of the modulator is to quantize the analog input signal and reduce the quantization noise within the signal baseband. This is accomplished through the use of a low-resolution quantizer, oversampling, negative feedback, and linear filtering.

One common type of modulator used in photonic oversampled A/D architectures is the recursive error diffusion modulator, shown in Fig. 7. Here, $H(z)$ represents the z transform of a causal, unity-gain filter and z^{-1} is a unit sample delay.

For a first-order modulator in which $H(z) = 1$, the modulator output can be shown to be

$$q(u_n) = \underbrace{x_n}_{\text{signal}} + \underbrace{\varepsilon_n - \varepsilon_{n-1}}_{\text{quantization error}} \tag{7}$$

Here, the quantity ε_n is the quantization error that would be seen at the modulator output *if there were no feedback loop*. However, as a result of the negative feedback, the first-order difference or discrete-time derivative of the error, $\varepsilon_n - \varepsilon_{n-1}$, appears at the output instead. By design, this difference signal is concentrated at high frequencies and can be removed by the digital LPF in the postprocessor.

3.3.2. The Digital Postprocessor. The function of the digital postprocessor is to digitally filter and decimate the output of the modulator so that the quantization noise that was spectrally shaped by the modulator is removed through lowpass filtering and the output digital signal is decimated to the Nyquist rate of the original sampled signal. An ideal LPF with cutoff frequency f_B is generally used to characterize upper-bound performance. Since the ideal LPF can be only approximated in practice, other practical filters have been investigated for this application [34]. With more recent advances in microlithography and silicon VLSI (very-large-scale integration technology), the digital postprocessor has

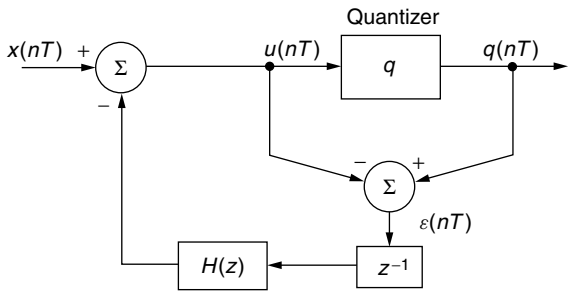


Figure 7. Block diagram of recursive error diffusion modulator.

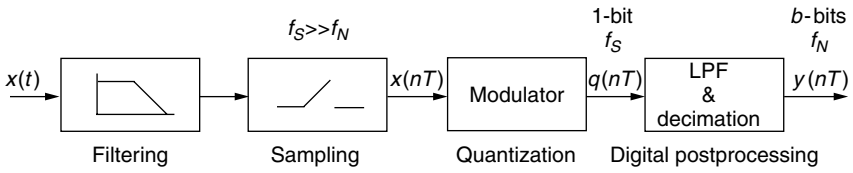


Figure 6. Generalized block diagram of an oversampled A/D converter.

become extremely sophisticated, routinely allowing the realization of filters with orders in excess of 100th-order.

3.3.3. Performance Analysis. Assuming white quantization noise characteristics, an ideal LPF with cutoff frequency f_B, and an Nth-order modulator with a noise shaping characteristic described by

$$H_{\text{ns}}(z) = (1 - z^{-1})^N \tag{8}$$

the spectral distribution of the quantization noise after shaping is the product of the filter shaping function and the spectral density of the quantizer error

$$S_\varepsilon(f) = \underbrace{[|H_{\text{ns}}(z)|^2_{z=e^{j2\pi fT}}]}_{\text{noise shaping}} \cdot \underbrace{\left[\frac{S_q}{f_S}\right]}_{\text{noise power density}} \tag{9}$$

If the quantizer step size is assumed to be Δ and the input signal is sinusoidal with a full-scale input range of $\pm\Delta/2$, the maximum signal-to-quantization noise ratio ($\text{SQNR}_{\max}$) can be computed as

$$\text{SQNR}_{\max}(M, N) = \frac{3}{2} \cdot \left[\frac{2N+1}{\pi^{2N}}\right] \cdot M^{2N+1} \tag{10}$$

where M is the oversampling ratio

$$M = \frac{f_S}{f_N} \tag{11}$$

For comparison, the $\text{SQNR}_{\max}$ of a conventional Nyquist rate uniform quantizer with b-bits resolution can be shown to be

$$\text{SQNR}_{\max}(b) = 3 \cdot 2^{2b-1} \tag{12}$$

Figure 8 shows the theoretical $\text{SQNR}_{\max}(M, N)$ and equivalent resolution for first- through fourth-order oversampled modulators as a function of oversampling ratio. The case of no noise shaping represents the $\text{SQNR}_{\max}$ that can be expected if the same quantizer, embedded in the feedback loop of the oversampled modulator, were simply oversampled and digitally filtered. The slope of this curve is 3 dB per octave and is included only for comparison. The slope of the $N = 1$ curve is 9 dB per octave and that of the $N = 2$ curve is 15 dB per octave, showing the significant advantage achieved by using a noise shaping modulator.

3.3.4. Spectral Characteristics. To better understand the relationship between the spectral shaping of the quantization noise and the input signal, the power spectrum for a fourth-order oversampling modulator output is shown in Fig. 9. This power spectrum assumes a half-scale sinusoidal input and an oversampling ratio of $M = 32$. Here, spectral noise shaping in the low frequencies of the power spectrum improves the A/D converter SNR performance.

3.3.5. The Error Diffusion Neural Network. The concept of spectral noise shaping can be extended to spatial dimensions and can be formulated in the context of a 2D symmetric error diffusion architecture. This extension requires each state variable in Fig. 7 to be represented in matrix–vector notation and can be described in terms of a specialized neural network.

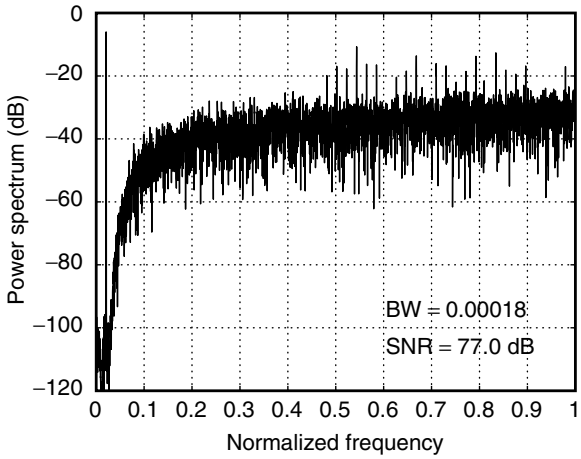


Figure 9. Power spectrum of the output data sequence for a fourth-order oversampling modulator.

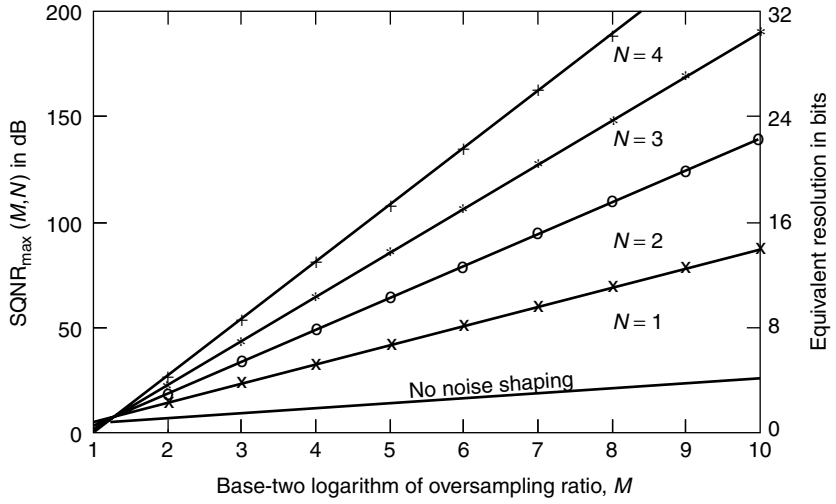


Figure 8. Maximum signal-to-quantization noise ratio of Nth-order modulators.

In equilibrium, the error diffusion neural network satisfies

$$\mathbf{u} = \mathbf{W}(\mathbf{y} - \mathbf{u}) + \mathbf{x} \tag{13}$$

For an $N \times N$ image, $\mathbf{W}$ is an $N^2 \times N^2$ sparse, "circulant without wrapping" matrix derived from the error diffusion filter weights $w_{i,j}$.

An equivalence to the classic Hopfield network can be described by

$$\mathbf{u} = \mathbf{A}(\mathbf{W}\mathbf{y} + \mathbf{x}) \tag{14}$$

where $\mathbf{A} = (\mathbf{I} + \mathbf{W}^{-1})$ and $\mathbf{I}$ is the identity matrix. Effectively, the error diffusion neural network includes a prefiltering of the input image $\mathbf{x}$ by the matrix $\mathbf{A}$ while still filtering the output image $\mathbf{y}$ but now with a new matrix, $\mathbf{AW}$.

The energy function of the error diffusion neural network can be shown to be a quadratic function

$$E(\mathbf{x}, \mathbf{y}) = [\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}]^T[\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}] \tag{15}$$

where $\mathbf{y} \in \{-1, 1\}$ is the output vector of quantized states with one element per pixel and $\mathbf{A} = \mathbf{B}^T\mathbf{B}$. As the error diffusion neural network converges and the energy function is minimized, so, too, is the error between the output and input images.

The error diffusion neural network represents an important class of A/D converter that applies to digital image halftoning. In digital halftoning [35], a continuous-tone input image is converted to a binary output image for purposes of printing, storage, or display.

3.3.6. A Distributed Mesh Feedback Approach to Photonic A/D Conversion. A relatively new approach to photonic A/D conversion uses spatial oversampling techniques, an error diffusion neural network, and a smart pixel [12] hardware implementation. This approach converts a 1D temporal signal to a 2D spatial representation in an effort to leverage the 2D nature of a photonic A/D architecture. In this approach, the input signal is first sampled at a rate higher than that required by the Nyquist criterion and then presented spatially as the input to a 2D error diffusion neural network consisting of $N \times N$ neurons, each representing a pixel in the image space. The neural network processes the input oversampled analog image and produces an $N \times N$ pixel binary or halftoned output image. Decimation and lowpass filtering techniques, common to conventional 1D oversampling A/D converters, digitally sum and average the $N \times N$ pixel output binary image using high-speed digital electronic circuitry. By employing a 2D neural approach to oversampling A/D conversion, each pixel constitutes a simple oversampling modulator, thereby producing a distributed A/D architecture. Spectral noise shaping across the array diffuses the quantization error, thus improving overall SNR performance. The matrix $\mathbf{A}$ in Eq. (14) describes the interconnectivity of the network resulting from local connections through the error diffusion filter. Since $\mathbf{A}$ is full-rank, the network is fully connected. Therefore each quantizer within the $N \times N$ network is embedded in a fully connected, distributed mesh feedback loop that spectrally shapes the overall quantization noise, thereby significantly reducing the effects of component mismatch typically associated with parallel or channelized A/D approaches.

3.3.7. Spectral Characteristics. A representative power spectrum of the output halftoned image of this fully connected distributed mesh feedback architecture is shown in Fig. 10. The input signal used was an $N \times N$ sinusoid with period 2 in both spatial dimension, and the error diffusion filter was a LPF with a 25×25 region of support designed using a Kaiser window with $\alpha = 5$. Approximately 30 dB of low-frequency noise suppression is achieved in this specific approach resulting in approximately a 54-dB SNR. Larger in-band noise suppression has also been achieved with other filter designs.

3.3.8. A Photonic Implementation of the Error Diffusion Neural Network. The functionality necessary to implement the error diffusion neural network consists of a one-bit quantizer, two differencing nodes, and the interconnection and weighting of the error diffusion filter. One photonic implementation of the error diffusion neural algorithm uses smart pixel technology [36]. Smart pixels integrate both electronic processing and individual optical devices on a common chip to take advantage of the complexity of electronic processing circuits and the speed of optical devices. Arrays of these smart pixels bring the advantage of parallelism that optics provides.

The electronic circuitry, used for quantization, weighting, and summation functions, was fabricated in a 0.5 μm complementary metal oxide semiconductor (CMOS) process. In a subsequent processing step, self-electrooptic effect device (SEED) multiple-quantum-well (MQW) modulators are integrated with the CMOS VLSI circuitry using a flip-chip bonding process. The MQW modulators are used in this implementation to provide optical input and output

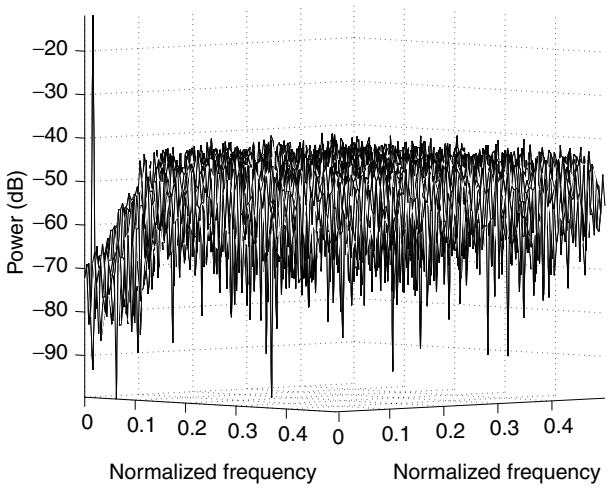


Figure 10. 3D perspective plot of one quadrant of the power spectrum of the fully connected distributed mesh feedback architecture generated with a 2D sinusoidal input of period 2 and a Kaiser error diffusion filter.

functionality. Using this approach, a 5×5 proof-of-concept photonic error diffusion neural network was fabricated.

The electronic circuitry for this CMOS-SEED error diffusion neural network consists of six distinct circuits associated with a single neuron. The standard neuron contains a total of 152 transistors with 23 dedicated to the neural functionality and 3 used for the output SEED driver. The error weighting circuitry for weights 1–5 contain the remaining 126 transistors with weights 1–5 accounting for 22, 20, 20, 28, and 36 transistors, respectively. This photonic implementation of the error diffusion neural network has been experimentally characterized and performed as predicted by both theory and simulation.

Figure 11 shows a photomicrograph of a single neuron of the CMOS-SEED photonic error diffusion neural network. The long rectangular features are the SEED MQW modulators and the background features are the CMOS VLSI transistors and interconnect metallization.

The challenges associated with this distributed approach to photonic A/D conversion include achieving large 2D arrays of smart pixel devices, the speed of convergence of the neural algorithm, and the integration with the electronic digital postprocessing circuitry.

3.4. Time Stretching Using Dispersive Optical Elements

Time stretching utilizes linear group velocity dispersion, most often in optical fibers, to frequency-downshift a signal that has been modulated onto optical pulses. In many of the more recent demonstrations, two long fiberoptic spools of lengths L_1 and L_2 are used with the modulator positioned between the two spools. The stretch factor M is defined as the width of the pulse exiting L_2 compared to that exiting L_1. If the pulsewidth exiting the source is defined as τ_0, and if δ_{τ_1} and δ_{τ_2} are defined as the additional broadening resulting from fiber spools L_1 and L_2, respectively, the stretch factor can be shown to be [37]

$$M = \frac{\tau_0 + \delta_{\tau_1} + \delta_{\tau_2}}{\tau_0 + \delta_{\tau_1}} \approx 1 + \frac{L_2}{L_1} \tag{16}$$

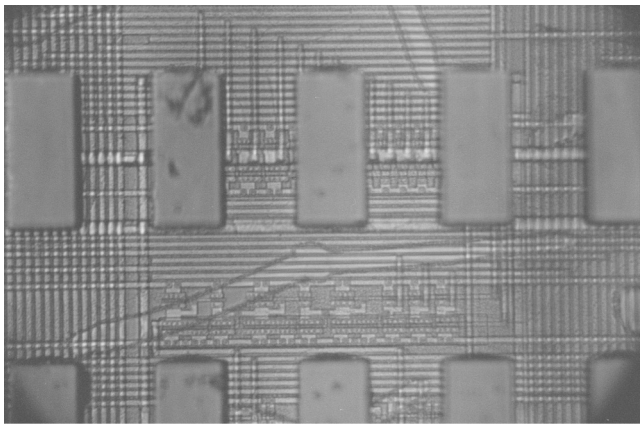


Figure 11. Photomicrograph of a single neuron of the 5×5 error diffusion neural network. (Reprinted with permission from B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.4, p. 220. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

if $\tau_0 \ll \delta_{\tau_1}$. In experimental demonstrations to date, L_1 and L_2 have generally been several kilometers in length.

Photonic time-stretch preprocessing techniques have also been proposed as a method to extend the performance of electronic A/D converters [21]. This technique is based on the premise that if an analog signal can be stretched in time, then the effective sampling rate and consequently the input bandwidth of a conventional electronic A/D converter can be increased. This specific technique is best suited to the class of *time-limited signals* such as those used in pulsed radar applications. Figure 12 shows the general concept of time stretching. Here, a modulated optical carrier is introduced to an optically dispersive element that subsequently stretches the modulated envelope of the signal. In Fig. 12, the vertical lines represent individual samples and the sampling interval T that would be required of the original signal is modified by the stretch factor M after the signal passes through the dispersive element. The optically dispersive element can be in a waveguide structure or a fiber. Time stretching using an array waveguide that is a wavelength dispersive element has been applied to photonic A/D conversion [37] and a 30-Gsps, 4-bit time-stretched photonic A/D converter using an 8-Gsps electronic digitizer has been reported [38].

One of the challenges associated with this particular approach to wide-bandwidth A/D conversion is distortion introduced by the nonuniform spectrum of the nonlinear dispersion. Nonuniformities in the spectrum result in temporal modulation of the carrier, which is mixed with the input signal during modulation. This nonuniformity results in a broadband distortion that limits the bandwidth and resolution of this particular approach to A/D conversion.

4. TRENDS IN PHOTON-BASED A/D CONVERSION

In general, trends in photonic A/D conversion can be categorized within the context of individual device and component development and investigation of new and novel architectures.

Individual photonic device development continues to be an active area of research interest. Improving the linearity, dynamic range, insertion loss, and speed of photonic components such as the electrooptic interferometers used for modulation and switching in current photonic A/D converters will directly improve the performance of these converters. Low-noise clock sources are an important part of any photonic A/D converter architecture. In high-speed, high-resolution applications, these sources require low timing and amplitude jitter, high repetition rate, and

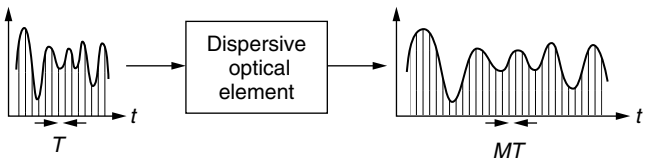


Figure 12. Concept of time stretching using a dispersive optical element.

short pulsewidth. For WDM applications, the clock source also requires large spectral width. For high-speed, high-resolution A/D applications, the clock source specifications are particularly challenging. Because of the importance of low-noise optical clock sources to both photonic A/D converters and other high-speed sampling applications, these sources also continue to be the focus of numerous research efforts.

New approaches and novel architectures continue to be developed to improve both the electronic and photonic contributions to the overall photonic A/D converter performance. Various forms of interleaving architectures will continue to be investigated and improved for application to wideband photonic A/D converters. As optical sampling speeds continue to increase, oversampling architectures will undoubtedly be considered for high-resolution applications. Distributed approaches to photonic A/D conversion can also be expected to be the focus of continued development because of their potential modularity and tolerance to fabrication errors.

A number of other promising approaches are currently under investigation. Details of several of these can be found in Refs. 39–41.

Photonics technology has generally been described as an enabling technology, supporting many of the functions provided by electronics technology. As both electronic and photonic technologies continue to mature, we can expect continued application and improvements to photonic A/D converter technology.

BIOGRAPHY

Barry L. Shoop is an associate professor in the Department of Electrical Engineering and Computer Science at the U.S. Military Academy, West Point, New York. He received his B.S. from the Pennsylvania State University, University Park, Pennsylvania, in 1980, an M.S. from the U.S. Naval Postgraduate School, Monterey, California, in 1986, and a Ph.D. from Stanford University, Stanford, California, in 1992, all in electrical engineering. His research interests are in the area of photonic A/D conversion, optical information processing, image processing, and smart pixel technology. He is a fellow of the OSA, a senior member of the IEEE, and a member of SPIE, Phi Kappa Phi, Eta Kappa Nu, and Sigma Xi.

BIBLIOGRAPHY

1. B. L. Shoop, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences Vol. 81, Springer-Verlag, Berlin, 2001.
2. D. F. Hoeschele, *Analog-to-Digital and Digital-to-Analog Conversion Techniques*, 2nd ed., Wiley, New York, 1994.
3. B. Razavi, *Principles of Data Conversion System Design*, IEEE Press, Piscataway, NJ, 1995.
4. M. J. Demler, *High-Speed Analog-to-Digital Conversion*, Academic Press, San Diego, CA, 1991.
5. R. van de Plassche, *Integrated Analog-to-Digital and Digital-to-Analog Converters*, Kluwer, Dordrecht, The Netherlands, 1994.
6. A. E. Steigman and D. J. Kuizenga, Proposed method for measuring picosecond pulsewidths and pulse shape in CW more-locked lasers, *IEEE J. Quant. Electron.* **6**: 212–219 (1970).
7. H. F. Taylor, An electrooptic analog-to-digital converter, *Proc. IEEE* **63**(10): 1524–1525 (1975).
8. R. A. Becker, C. E. Woodward, F. J. Leonberger, and R. W. Williamson, Wide-band electrooptic guided-wave analog-to-digital converters, *Proc. IEEE* **72**(7): 802–819 (1984).
9. B. Jalali and Y. M. Xie, Optical folding-flash analog-to-digital converter with analog encoding, *Opt. Lett.* **20**: 1901–1903 (1995).
10. B. L. Shoop and J. W. Goodman, Optical oversampled analog-to-digital conversion, *Appl. Opt.* **31**(26): 5654–5660 (1992).
11. B. L. Shoop and J. W. Goodman, A first-order error diffusion modulator for optical oversampled A/D conversion, *Opt. Commun.* **97**(4): 167–172 (1993).
12. B. L. Shoop, A. H. Sayles, D. A. Hall, and E. K. Ressler, A smart pixel implementation of an error diffusion neural network for digital halftoning, *Int. J. Optoelectron.* **11**: 217–228 (1997).
13. B. L. Shoop, Photonic A/D converters, *Proc. SPIE* **3490**: 252–255 (1998).
14. B. L. Shoop et al., A highly-parallel mismatch tolerant photonic A/D converter, *Proc. Conf. Lasers and Electro-Optics*, OSA Technical Digest, Optical Society of America, Washington, DC, 2001, pp. 64–65.
15. P. E. Pace, S. J. Ying, J. P. Powers, and R. J. Pieper, Integrated optical sigma-delta modulators, *Opt. Eng.* **35**: 1826–1836 (1996).
16. P. E. Pace, S. A. Bewley, and J. P. Powers, Fiber-lattice accumulator design considerations for optical $\sigma\delta$ analog-to-digital converters, *Opt. Eng.* **39**: 1517–1526 (2000).
17. J. C. Twichell and R. J. Helkey, Phase-encoded optical sampling for analog-to-digital converters, *IEEE Photon. Technol. Lett.* **12**: 1237–1239 (2000).
18. P. W. Juodawlkis et al., 505 MS/s photonic analog-to-digital converter, *Proc. Conf. Lasers and Electro-Optics*, OSA Technical Digest, Optical Society of America, Washington, DC, 2001, pp. 63–64.
19. T. R. Clark, J. U. Kang, and R. D. Esman, Performance of a time- and wavelength-interleaving photonic sampler for analog-digital conversion, *IEEE Photon. Technol. Lett.* **11**: 1168–1170 (1999).
20. T. R. Clark and M. L. Dennis, Toward a 100-G sample/s photonic analog-to-digital converter, *IEEE Photon. Technol. Lett.* **13**: 236–238 (2001).
21. A. S. Bhushan, F. Coppinger, and B. Jalali, Time-stretched analog-to-digital conversion, *Electron. Lett.* **34**: 839–840 (1997).
22. Y. Tsunoda and J. W. Goodman, Combined optical A/D conversion and page composition for holographic memory applications, *Appl. Opt.* **16**(10): 2607–2609 (1977).
23. H. K. Liu, Coherent optical analog-to-digital conversion using a single halftone photograph, *Appl. Opt.* **17**(14): 2181–2185 (1978).
24. K. Takizawa and M. Okada, Analog-to-digital converter: A new type using an electrooptic light modulator, *Appl. Opt.* **18**(18): 3148–3151 (1979).

25. N. N. Evtikheiv, D. I. Mirovitskii, N. V. Rostovtseva, and O. B. Serov, Multilayer holographic functional element in an analog-to-digital converter, *Sov. J. Quant. Electron.* **16**(9): 1180–1184 (1986).

26. J. A. Bell et al., Extension of electronic A/D converters to multi-gigahertz sampling rates using optical sampling and demultiplexing techniques, *Proc. 23rd Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 1989.

27. A. D. McAulay, Optical analog-to-digital converter using optical logic and table look-up, *Opt. Eng.* **29**(2): 114–120 (1990).

28. Y. Li and Y. Zhang, Optical analog-to-digital conversion using acoustooptic theta modulation and table lookup, *Appl. Opt.* **30**(30): 4368–4371 (1991).

29. J. U. Kang, M. Y. Frankel, and R. D. Esman, Highly parallel pulsed optoelectronic analog-digital converter, *IEEE Photon. Technol. Lett.* **10**: 1626–1628 (1998).

30. P. W. Juodawlkis et al., Time-interleaved optical sampling for analog-to-digital converters (in press), *IEEE J. Lightwave Technol.*

31. J. C. Twichell and R. Helkey, Optical sampling for analog-to-digital converters, in *Lincoln Laboratory, Solid State Research, Quarterly Technical Report*, Lexington, MA, 1996, Vol. ESC-TR-96-096, pp. 28–33.

32. R. C. Williamson et al., Effects of crosstalk in demultiplexers for photonic analog-to-digital converters, *IEEE J. Lightwave Technol.* **19**: 230–236 (2001).

33. R. C. Williamson et al., Effects of crosstalk in demultiplexed photonic analog-to-digital converters, *Proc. Conf. Lasers and Electro-Optics,* OSA Technical Digest, Optical Society of America, Washington, DC, 2000, pp. 625–626.

34. E. B. Hogenauer, An economical class of digital filters for decimation and interpolation, *IEEE Trans. Acoust. Speech Signal Process.* **29**(2): 155–162 (1981).

35. R. A. Ulichney, *Digital Halftoning*, MIT Press, Cambridge, MA, 1987.

36. C. DeCusatis, D. Clement, and R. Lasky, eds., *Handbook of Fiber Optic Data Communication*, Academic Press, San Diego, CA, 1997.

37. F. Coppinger, A. S. Bhushan, and B. Jalali, Optoelectronic time-stretch and its application to analog-to-digital conversion, *IEEE Trans. Microwave Theory Tech.* **47**: 1309–1314 (1999).

38. A. S. Bhushan, P. Kelkar, F. Coppinger, and B. Jalali, 30 G sample/s 4-bit time-stretch analog-to-digital converter, *Proc. Conf. Lasers and Electro-Optics,* OSA Technical Digest, Optical Society of America, Washington, DC, 2000, pp. 623–624.

39. R. Urata et al., High-speed sample and hold using low temperature grown GaAs MSM switches for photonic A/D conversion, *IEEE Photon. Technol. Lett.* **13**: 717–719 (2001).

40. J. Cai and G. W. Taylor, Optoelectronic thyristor-based photonic smart comparator for analog-to-digital conversion, *IEEE Photon. Technol. Lett.* **10**: 1295–1297 (1999).

41. H. Sakata, Photonic analog-to-digital conversion by use of nonlinear Fabry-Perot resonators, *Appl. Opt.* **40**(2): 240–248 (2001).

POLARIZATION MODE DISPERSION MITIGATION

HENNING BÜLOW
Optical Systems
Stuttgart, Germany

With the increase of channel bit rate to 10 and 40 Gbps (gigabits per second) and in conjunction with unrepeated link lengths of hundred of kilometers and beyond, polarization mode dispersion (PMD) might become visible as a degrading property of the transmission link. Since PMD effects are drifting with time and differ for each wavelength channel, various dynamically adapting PMD compensators (PMDC) have been proposed that process either the optical signal field at link output or the electrical signal in the receiver after detection.

1. IMPACT OF PMD ON TRANSMISSION

The PMD of the optical transmission link arises mainly as a result of the residual optical birefringence varying along the fiber length, which is induced during the production of or cabling of the fiber. Nevertheless, other optical components or subsystems within the optical signal path such as optical amplifiers, dispersion-compensating fiber modules, or wavelength multiplexers might also add to the link PMD.

The effect of all these cascaded birefringences can again be regarded as an optical birefringence having a pair of principal states of polarization (PSP) e_- and e_+, which are orthogonally polarized and exhibit different group delays τ_- and τ_+. Unlike the scenario with fixed birefringence, they vary with the optical frequency ω. The difference of the group delay between fast and slow PSP is denoted by the differential group delay (DGD) $\Delta\tau = \tau_+ - \tau_-$ (see Fig. 1). The PMD is described by the dispersion vector $\Omega(\omega) = \Delta\tau e_-$, which has the length $\Delta\tau(\omega)$ and is oriented into the direction of the Stokes vector $e_-(\omega)$, which represents the state of polarization on the Poincaré sphere of the fast PSP at the output of the link [1]. Evaluation of Ω around the signal center frequency ω_0 exhibits that, as long as the signal bandwidth $\Delta\omega_S$ is sufficiently narrow with respect to the link PMD, the transmitter signal spreads among fast and slow PSP and thus suffers from a dual-path propagation due to the DGD (see Fig. 1). This leads to a temporal broadening of the bit beyond a bit period T and thus to intersymbol interference (ISI).

Since the DGD as well as the PSP vary with both wavelength and time (see Fig. 2), PMD is quantified by

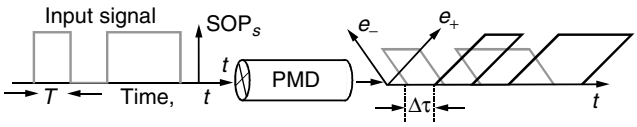


Figure 1. Optical data signal at input and output of a fiber having a first-order PMD with a DGD $\Delta\tau$.

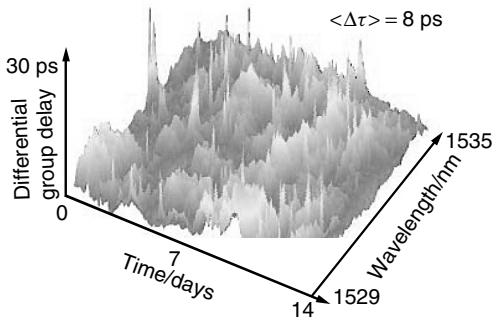


Figure 2. DGD of an installed fiber (246 km length, 8 ps PMD) measured over a timespan of 14 days and a wavelength range of 6 nm.

statistical means: the mean value of the DGD $\langle\Delta\tau\rangle$ and the probability density function of the actual DGD $\Delta\tau$, which is given by the Maxwellian distribution

$$\rho(\Delta\tau) = \sqrt{\frac{6}{\pi}}\frac{3\Delta\tau^2}{\langle\Delta\tau\rangle^3}\exp\left[-\frac{3\Delta\tau^2}{2\langle\Delta\tau\rangle^2}\right]$$

Thus, an actual DGD $\Delta\tau$ of beyond 3.1 $\langle\Delta\tau\rangle$ occurs with a probability of 10^{-5}. As indicated by the autocorrelation function (ACF) of PMD, which decays below 0.5 for $\Delta\omega > 2/\langle\Delta\tau\rangle$, the PMD cannot be considered as constant anymore within increasing signal bandwidth $\Delta\omega_S$ and increasing link PMD $\langle\Delta\tau\rangle$. Then the probability of a signal distortion due to second- and higher-order PMD increases. Second-order PMD denotes a constant variation $d\Omega/d\omega$ of the PMD vector, and higher orders $(n+1)$ denote nonvanishing derivatives $d^n\Omega/d\omega^n$ at the signal center frequency. The impact of second-order PMD on the signal is a depolarization proportional to $de_-/d\omega$ and to $\Delta\tau$ [2], which denotes a cross coupling of signal fields between the orthogonal PSPs, and the PMD-induced chromatic dispersion proportional to $\pm 0.5 d\Delta\tau/d\omega$. The induced dispersion exhibits opposite signs in both PSPs and adds to the chromatic dispersion of the link. A signal distortion induced by PSP rotation $de_-/d\omega$ has the highest likelihood of the two second-order contributions. It generates distortions similar to over- and undershoots in the data signal detected by the receiver photodiode [2].

Variations of $\Delta\tau$ and e_- with time are induced mainly by environmental temperature changes that lead to a variation of the signal distortion and thus to a fluctuating bit error rate (BER). The BER limit that is tolerated, such as 10^{-12} or for forward error correction (FEC) support 10^{-4},

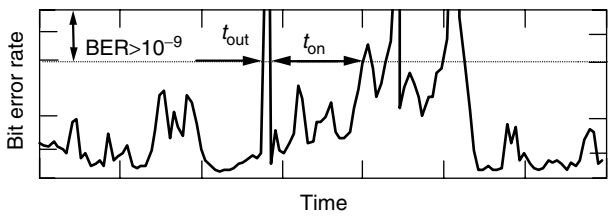


Figure 3. Temporal BER evolution for the transmission over a fiber link with PMD (14-day timespan).

might be exceeded (see Fig. 3). Times of poor BER mean an outage for the transmission system. An outage event has a duration t_{out}. Because of the statistical nature of PMD, the impact on a transmission system is quantified by the cumulative probability (CP) for a PMD-induced outage. CP can be defined by the ratio of accumulated outage time and system operation time for long observation timespans. Besides the outage probability, the mean outage duration $\langle t_{out}\rangle$ and the mean time between outages $\langle t_{on}\rangle$ have been proposed to describe the impact of PMD on a system. That is 56 min for $\langle t_{out}\rangle$ and 2.56 years (outage rate 0.39 yea^{-1}) for $\langle t_{on}\rangle$ have been deduced from PMD measurements of buried fiber [3]. On the other hand, PMD fluctuations, also on the timescale of a few milliseconds, have been observed for aerial cables or fiber exposed to mechanical vibrations. If these fast fluctuations need to be compensated, they determine the upper speed limit of a PMDC.

Alternatively, the robustness of a transmission system to PMD can be quantified by the PMD limit $\langle\Delta\tau\rangle_{lim}$, which is the maximum PMD that leads to a specific outage probability CP given that the system operates with a certain optical signal-to-noise ratio (OSNR) margin. Typical values for CP and margin are 10^{-5} and 2 dB, respectively. $\langle\Delta\tau\rangle_{lim}$ amounts to about ~15% of the bit period T for a non-return-to-zero (NRZ) signal.

In order to assess the relevance of PMD for a transmission system, the value of the length-related PMD, the PMD coefficient τ', needs to be related to $\langle\Delta\tau\rangle_{lim}$ and the link length L. Fiber PMD $\langle\Delta\tau\rangle_{fiber}$ and all component and subsystem PMDs $\langle\Delta\tau\rangle_i$ contribute to the total link PMD$_{tot}$ according to $\langle\Delta\tau\rangle_{tot}^2 = \langle\Delta\tau\rangle_{fiber}^2 + \Sigma\langle\Delta\tau\rangle_i^2$. The maximum link length is $L < (\langle\Delta\tau\rangle_{lim}^2 - \Sigma\langle\Delta\tau\rangle_i^2)/\tau'^2$. With more recently manufactured fiber having a low PMD coefficient of $\leq$0.08 ps/$\sqrt{}$km, and assuming a dispersion compensation fiber PMD coefficient of 0.15 ps/$\sqrt{}$km, an amplifier PMD of 0.5 ps and 80 km amplifier spacing, 40 Gbps transmission over long-haul distances is affected by PMD beyond 1400 km. In fiber production PMD received only minor attention until the mid-1990s. Therefore links incorporating older fibers might exhibit PMD coefficients of $\geq$0.5 ps/$\sqrt{}$km. This means that a few hundreds of kilometers of link might reach the PMD limit even at 10 Gbps.

2. PMD COMPENSATION

A slight improvement in the robustness of PMD can be obtained by keeping the receiver threshold in the middle of the eye diagram degraded by ISI. This automatic threshold adjustment leads to the abovementioned PMD limit of $0.15T$. Moreover, specific modulation formats such as return to zero (RZ) tolerate a slightly higher bit broadening due to PMD and thus increase the limit to approximately $0.18T$. The tolerance to PMD can further be increased if there is room to allocate a power margin (OSNR margin) for PMD of much more than 1 or 2 dB. This margin can be obtained by reducing span loss or by incorporating forward error correction (FEC).

However, in general only a limited margin will be reserved for PMD, and a higher PMD limit must be overcome. Therefore several active compensation

techniques have been proposed to mitigate the degrading effect of PMD. The majority of these approaches can be classified as either optical or electrical compensation techniques that apply optical signal processing within an optical PMD compensator unit, or postdetection electronic signal processing of the photodiode signal by an electrical equalizer within the receiver. Commonly the dynamic adaptation of the compensator unit to the drifting signal distortion is accomplished by three elements arranged in a feedback scheme (see Fig. 4): a signal processing element to reduce the distortion, an element at its output to measure the signal quality and to provide a feedback signal, and an adaptation control algorithm implemented in an electronic processor that tunes the parameters of the signal processor into direction of optimum feedback signal and thus to reduced signal distortion.

3. OPTICAL COMPENSATION

The signal processing element of the simple optical PMD compensator is formed by an electrically tunable polarization controller (PC) based, for example, on the electrooptic effect in lithium niobate, on the elastooptic effect used in fiber squeezer devices, or on liquid crystal technology. A constant optical birefringence [e.g., polarization-maintaining fiber (PMF)] is attached at its output (see Fig. 4). A typical value of the PMF's DGD is $\Delta\tau_C = 0.6T$. The principle of operation of this basic PMDC is to modify the dispersion vector of the total PMD $\Omega_{\text{tot}} = \Omega_L + \Omega_C$ formed by the link PMD Ω_L and the compensator PMD Ω_C (see Fig. 5), by tuning of the PC in such a way that the signal degradation at compensator output is a minimum [4]. The PC modifies the orientation of Ω_C ($|\Omega_C| = \Delta\tau_C$). Factoring in these dispersion vectors and the Stokes vector SOP_S representing the signal input state of polarization on the Poincaré sphere, all

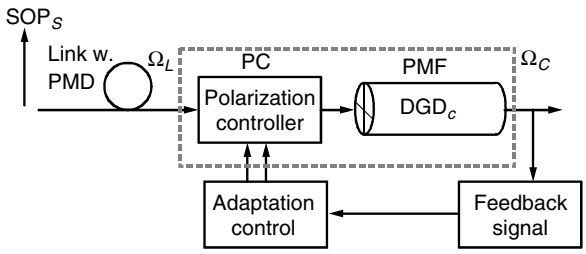


Figure 4. Basic optical PMD compensator.

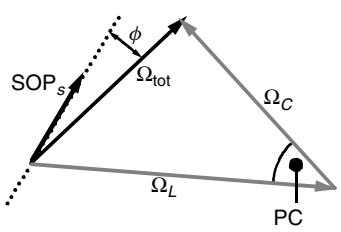


Figure 5. PMD dispersion vectors Ω and the signal state of polarization SOP_S observed at the output of the basic PMD compensator. Ω_L and Ω_C represent the link PMD and compensator PMD, respectively, at signal center frequency.

transformed to PMDC output, the PMD conditions for compensation of first-order PMD at the center frequency can be explained. During operation the compensator will either minimize the total DGD $|\Omega_{\text{tot}}|$ (case A) or the angle ϕ between SOP_S and the total PMD Ω_{tot} (case B). In both cases the degradation effects are minimized: in case A due to minimum delay $|\Omega_{\text{tot}}|$ and in case B due to single-path propagation in one PSP only. A more detailed analysis shows that by detuning of the PC from these ideal first-order conditions — to some extent — higher-order PMD can be mitigated too.

During operation of a PMDC under conditions of drifting link PMD, the adaptation control keeps the signal processor in an extremum of the feedback signal. In this simple PMD compensator structure the decay of a global maximum of the feedback signal to a local maximum limits the efficiency.

Calculations indicate that with the basic PMD compensator, the PMD limit can be extended to about $0.30T$ (2 dB power margin, 10^{-5} outage) for non-return-to-zero (NRZ) signals.

A feedback signal (FS) must satisfy the following the requirements: (1) good correlation with the receiver BER and (2) sufficient sensitivity to detect a detuning of the signal processor from the optimum setting or a variation of the link PMD. As a feedback signal the measurement of signal properties can be used, either in the optical domain such as the degree of polarization (DoP) or, after detection, in the electrical domain, including "spectral line" or eye opening. The DoP is measured at the compensator output by a polarimeter setup. The PMD-induced pulse splitting results in a time-varying state of polarization along the bit pattern, which reduces the measured DoP [5]. The spectral line feedback denotes the analysis of the electrical spectrum at a photodiode illuminated by the compensator output signal. Since PMD generates a notch in the electrical spectrum at the frequency $0.5/\Delta\tau$, the maximization of spectral power measured at, for instance, $0.5/T$ or $0.25/T$ or the weighted sum of both indicate a decreasing residual PMD distortion after compensation and thus an improved receiver BER [6]. A very good correlation between BER and the feedback signal is provided by an electronic performance monitor, also referred to as an "eye monitor" [7]. It extracts a quality measure of the actual eye diagram by bit-synchronous sampling at decision time t_0 and is thus strongly correlated with the BER. The sampling demands a valid clock that might not always be extractable from the signal during times of strong signal distortion as might occur during startup of the compensator when the accommodation to the actual distortion is not completed.

4. HIGH-ORDER PMD COMPENSATION

In basic PMD compensator with a PMD limit of about $0.30T$, the adaptation control has to optimize the 2 degrees of freedom (DoF) of the polarization controller. An improvement of this basic structure can be achieved by replacing the fixed DGD by an continuously tunable DGD that is also tuned by the adaptation control. With this increase of the number of DOFs to 3, the PMD limit

is slightly improved [4]. The main advantage of a variable DGD is that the decay of the global optimum to a less efficient local optimum can be avoided during tracking of the drifting link PMD. Realizations of a variable group delay are based on beam optics with mechanically variable gaps or stacks of birefringent elements with electrically controlled polarization switches in between.

In order to address higher orders of PMD and increase the PMD limit of the compensator, improved optical signal processing can be achieved by a cascade of two or more basic compensators [6]. This leads to structures with ≥ 4 DoF controlled by maximization of one single feedback signal. The increase of the PMD limit beyond $0.40T$ was calculated for a two stage structure [4,8]. However, in general an increasing number of DoFS improves the effectiveness of a PMDC, but it seems that the sensitivity of the feedback signal to the variation of an individual tuning parameter decreases. This slows down the accommodation time for a changing PMD. Moreover, with increasing number of DoFs, the danger of being trapped in suboptimal local maxima increases. Therefore the search for efficient high-order PMDC is continued [8].

5. ANALYSIS OF COMPENSATION

For experimental or numerical assessment of compensator performance, the statistics of the system-relevant quality parameter degradation (BER, power penalty or OSNR penalty) have to be determined. Several hundreds or some thousands of statistically independent samples of actual PMD values of a given PMD $\langle \Delta\tau \rangle$ are applied at a transmission system with PMDC and the BER or power penalties are determined. These PMD samples are generated either in a PMD emulator or in a computer model with all relevant orders. Their occurrence should obey the PMD statistics of a real link (or it should be possible to convert them to the appropriate statistics). The set of BER or penalty values exhibits the same statistics that one would obtain from measurements at a real link after many months or years. For analysis, the relative occurrence that a specific BER or penalty value is exceeded is plotted against this limit (see Fig. 6). Thus, the resulting curves show that with PMDC, specific degradations are exceeded with a lower likelihood than without it and that an improved PMDC (two stages) leads to further reductions of the likelihood. The curves can be extrapolated to more relevant 10^{-5} outage, and a set of curves for different PMDs $\langle \Delta\tau \rangle$ allows us to interpolate the PMD limit $\langle \Delta\tau \rangle_{\text{lim}}$.

6. ELECTRICAL COMPENSATION

Postdetection electronic signal processing is also discussed and applied for PMD mitigation. Adaptive electrical signal processing schemes, well established in lower-rate communications for compensation of channel degradation, are also implemented in an optical receiver. The equalization schemes that are in the scope of application are a feedforward equalizer (FFE) followed by a decision feedback equalizer (DFE). The FFE superimposes differently delayed and weighted replica of the input signal and the

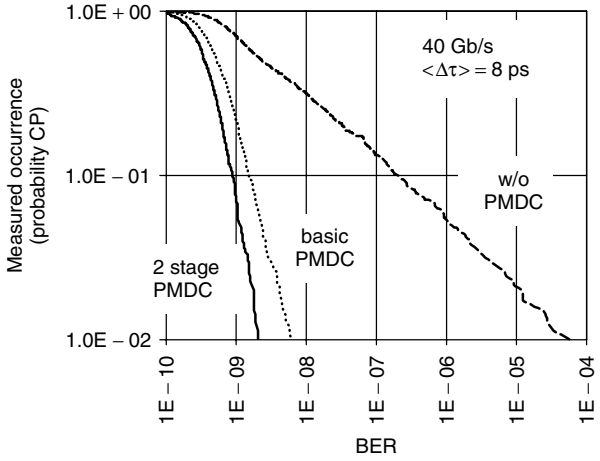


Figure 6. Relative occurrence (y axis; ordinate) of measured system bit error ratios beyond a BER limit (x axis; abscissa) for some hundred transmissions over a statistically changing PMD emulator. The experiments were performed without and with optical PMD compensators.

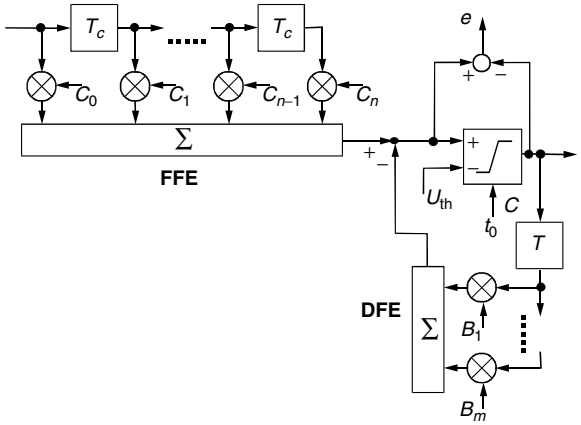


Figure 7. Electronic equalizer for electrical PMD compensation. The electronic signal processing is performed by a feedforward equalizer (FFE) and a decision feedback equalizer (DFE), which also provides an error signal e for the adaptation control.

DFE subtracts the weighted and delayed signal appearing at the decision gate output from the input signal provided by the FFE (see Fig. 7). Tuning to the actual PMD distortion is achieved by adjusting the FFE weights (C_i), also referred to as "tap coefficients," and the DFE feedback weights (B_i). FFE and DFE suitable for 10-Gbps operation have been realized as integrated analog signal processing circuits in high-speed technologies such as SiGe, InP, or GaAs [9]. An alternative realization in a digital signal processing (DSP) scheme based on digitally processing of the signal in an integrated electronic circuit [complementary metal oxide semiconductor integrated circuit (CMOS IC)] after analog-to-digital conversion of bit-synchronous signal samples is also discussed.

Besides these equalization techniques that perform the decision on a bit-by-bit basis, there also exists a concept of optimum detection based on data decisions that are most likely to be correct when utilizing the entire

received signal and the knowledge of the characteristics of the transmission channel. Since in the case of PMD the characteristics are neither known nor stable in time, adaptive mechanisms also have to be applied. The performance limits of this maximum-likelihood sequence estimation (MLSE) under realization constraints for the optical channel is currently under investigation [10]. Constraints are high-speed electronic realization for 10-Gbps signals with reduced complexity, processing of truncated sequences, nonstationary noise, and nonlinear channel due to optical nonlinearity of the fiber and to square-law detection of the photodiode. Preliminary results confirm the potential for an improved performance for PMD compensation compared to the FFE and a DFE equalizer (see Fig. 8). The MLSE detector will be realized as DSP scheme. The processing is based on the Viterbi algorithm, which is an efficient way to organize the computations of the signal samples.

Different adaptation schemes for the FFE, DFE, and clock-phase alignment are under consideration. Similar to the optical compensator, the eye monitor can be used at the FFE output in conjunction with a gradient algorithm that consecutively tunes the equalizer taps. The implementation of the least-mean-square (LMS) algorithm, well established at lower rates, has also been discussed. An error signal e at the decision time t_0 is generated for adaptation purpose by subtracting the signal at the decision gate output, serving as reference, from the signal at the equalizer output (see Fig. 7). The multiplication of this error signal with samples of signals at different positions within the equalizer generates different independent feedback signals for each FFE tap or DFE feedback tap. This allows for simultaneous adaptation of all tap weights. Moreover, the feedback signals are bipolar and thus automatically indicate the direction of tuning to the optimum. Therefore very fast adaptation within some hundreds of bits is possible in principle. An alternative realization of the LMS scheme is possible in conjunction with channel coding used for forward error correction (FEC). By comparing the uncorrected and the corrected data sequences launched into and appearing at the output of the FEC decoder, respectively, the observed errors can be correlated with specific bit constellations. The error count for a specific constellation provides a measure for the error e.

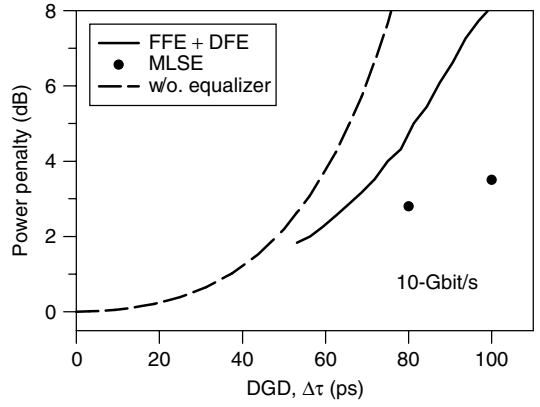


Figure 8. Power penalty (OSNR penalty) of a 10-Gbps optical receiver versus a first-order PMD distortion quantified by the DGD. The penalties are determined numerically for a receiver both without (dashed line) and with (solid line) an equalizer for electrical PMD compensation.

7. OPTICAL VERSUS ELECTRICAL COMPENSATION

Optical PMD compensation by a cascade of two ($0.4T$ PMD limit) or more basic compensators has been demonstrated experimentally. Theoretically, zero penalty can be achieved for an ideal exact compensation of the PMD and operation at bit rates of $\leq$160 Gbps has already been shown in the laboratory. In contrast to these findings, electrical equalization exhibits a residual penalty in the presence of strong PMD distortion and operation has been demonstrated at $\leq$10 Gbps as a result of the speed limits of electronic realization. The remaining penalty is due to the loss of polarization and phase information after square-law detection by the receiver photodiode (see Fig. 8). Mitigation by FFE and DFE is accompanied by a residual penalty of 7 dB for a first-order PMD distortion with a DGD of one bit period and beyond (equal excitation of both PSPs) and in the presence of signal-dependent optical noise (signal independent thermal noise leads to approximately half-residual penalty). This is too high for target applications such as long-distance transmission, which is strongly limited by noise and where only approximately $\leq$2 dB can be allocated for PMD. In this case the FFE and DFE exhibit a reasonable low penalty only if the DGD remains well below a bit period. This value corresponds to a PMD limit in the range of $0.25T$. An improvement is possible by applying MLSE schemes, with the first numerical results indicating a residual penalty in the range of 4 dB [9,10].

Nevertheless, postdetection electronic signal processing improves the receiver performance not only in the presence of PMD but also for ISI stemming from chromatic dispersions or optical nonlinearity such as self-phase modulation (SPM). In addition, it has the potential for effecting a seamless and cost-effective integration in the receiver electronics.

More recent numerical and experimental results have indicated that the efficiency of PMD compensation might be degraded when being used in a WDM environment [11]. As a result of the birefringence induced by other wavelength channels via the Kerr effect of the fiber, a fast polarization modulation can occur at high channel power that is not compensated by current PMDC schemes.

BIOGRAPHY

Henning Bülow received the Dipl.-Ing. degree in electrical engineering in 1985 from the University of Dortmund, Germany, and the Dr. degree in electrical engineering from the University of Berlin in 1988. He joined the Research Center of Alcatel in Stuttgart, Germany, in 1990 as a Research Engineer. At Alcatel he has been studying different aspects of optical communication systems, focusing mainly on erbium-doped fiber amplifiers, optical end electrical time-multiplexed 40-Gbps transmission systems, and assessment of data

transmission in the presence of polarization mode dispersion of the fiber. Since 1998, Dr. Bülow has headed a research team investigating the dynamic mitigation of transmission distortions at 10, 40, and 160 Gbps by adaptive electrical and optical signal processing. He holds over 20 patents in the area of optical communications, has authored or co-authored more than 60 journal and conference publications, and serves on the technical program committee of the Optical Fiber Communication Conference (OFC).

BIBLIOGRAPHY

1. C. D. Poole and J. Nagel, Polarization effects in lightwave systems, in I. P. Kaminow and T. L. Koch, eds., *Optical Fiber Communications IIIA*, Academic Press, San Diego, 1997, Chap. 6.
2. C. Francia et al., PMD second-order effects on pulse propagation in single-mode fibers, *IEEE Photon. Technol. Lett.* **10**: 1739–1741 (1998).
3. R. Caponi et al., WDM design issues with highly correlated PMD spectra of buried optical cables, *Technical Digest OFC 2002*, Anaheim, CA (USA), ThI5, 2002.
4. H. Sunnerud et al., A comparison between different PMD-compensation techniques, *IEEE J. Lightwave Technol.* **20**(3): 368–378 (2002).
5. S. Lanne, W. Idler, J.-P. Thiéry, and J.-P. Hamaide, Demonstration of adaptive PMD compensation at 40Gb/s, *Technical Digest OFC 2002*, Anaheim, CA, TuP3, 2001.
6. R. Noé et al., Polarization mode dispersion compensation at 10, 20, and 40 Gb/s with various optical equalizers, *J. Lightwave Technol.* **17**(9): 1602–1615 (1999).
7. F. Buchali et al., A 40 Gb/s eye monitor and its application to adaptive PMD compensation, *Technical Digest OFC 2002*, Anaheim, CA, Proc. WE6, 2002.
8. J. Poirrier, F. Buchali, and H. Bülow, Optical PMD compensation performance: numerical assessment, *Technical Digest OFC 2002*, Anaheim, CA, WI3, 2002.
9. H. Bülow, Electronic equalization of transmission impairments, *Technical Digest OFC 2002*, Anaheim, CA, TuE4, 2002.
10. H. F. Haunstein et al., Design of near optimum electrical equalizer for optical transmission in the presence of PMD, *Technical Digest OFC 2001*, Anaheim, CA, WAA4, 2001.
11. J. H. Lee et al., Impact of nonlinear crosstalk on optical PMD compensation, *Technical Digest OFC 2002*, Anaheim, CA, ThI2, 2002.

POLYPHASE SEQUENCES

So Ryoung Park
Iickho Song
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea

1. INTRODUCTION

Sequences have played an important role in the history of communication systems, especially as the spreading sequences in spread-spectrum (SS) code-division multiple access (CDMA) systems used widely for the most recent personal cellular communications. Among the systems using the SS technique, the direct-sequence (DS) CDMA system expands the bandwidth of a signal by directly multiplying an information symbol by a spreading sequence uniquely assigned to each user, so that a number of users in a cell can share the same frequency band simultaneously. The DSCDMA system is usually preferred over the other SS techniques because it has low implementation cost, can be used with coherent demodulation, and can provide a large capacity in addition to such usual benefits of the SS systems as interference rejection/suppression, multipath mitigation, and security [1–5].

Cellular DSCDMA systems have adopted the multiple or two-layered spreading sequence allocation for flexible system deployment and operation [6]. Orthogonality can be achieved by first multiplying each user's information symbol by a short spreading sequence that is orthogonal to that of any other user in the same cell. This first spreading is followed by a multiplication of a long spreading sequence, which is cell-specific but common to all users in the same cell in the forward link and user-specific in the reverse link. It is possible to provide waveform orthogonality among all users in the same cell while maintaining only mutual randomness between users in different cells. The short spreading sequences are called *channelization sequences* and the long spreading sequences, *scrambling sequences*.

Since different base stations (different mobile users) use different timeshifts of the same sequence in the forward (reverse) link in intercell synchronous operation, the long scrambling sequences are required to have good autocorrelation (AC) properties. The AC property of a sequence is said to be perfect when the AC value is N for the in-phase component and zero for the out-of-phase components, where N is the length of the sequence. This property can support fast symbol synchronization, low multipath interference (MPI), and low intercell interference (ICI). In the intercell asynchronous operation, each cell is assigned to a unique long sequence, which is thus required to have good AC property for fast symbol synchronization and low MPI and to have good crosscorrelation (CC) property for low ICI.

The short spreading sequences in a cell play an important role not only in spreading and despreading but also in the identification of a desired user from interfering users. This is the reason why the short spreading sequence is alternatively called the *signature sequence*. Because the whole frequency band is being used all the time, the bandwidth can be utilized more efficiently (i.e., with narrower equivalent bandwidth per user) in the DSCDMA system than in the conventional systems. The number of users that a system can accommodate is determined by the required signal-to-noise ratio (SNR), which is in turn determined by system design requirements. Note that there is a stringent limitation on the maximum number of users in the time-division multiple access (TDMA) and

frequency-division multiple access (FDMA) systems. On the other hand, the capacity of a CDMA system is softly limited; that is, the maximum number of users is not a clear-cut number. Instead, as more users share the same CDMA channel, the signal quality degrades gradually until it is unacceptable. The acceptable number of users depends on many aspects, including the CC property of the signature sequences in DSCDMA systems. The CC property of signature sequences is linked directly with the multiple access interference (MAI), and consequently with the capacity of the cell. Thus, for low MAI and a large channel capacity, it is desired that the CC value is always (close to) zero.

Traditionally, the cellular DSCDMA systems have utilized the binary maximal length (m) and Gold sequences as the scrambling sequences and Walsh sequences as the channelization sequences. The m sequences have been chosen because they possess the desired randomness and are easily generated via linear feedback shift registers. Gold sequences are an important subclass of the m sequences that can provide good periodic (even) CC. The maximum magnitudes of the periodic AC and CC of Gold sequences are 1 and $1 + 2^{\lfloor (m+2)/2 \rfloor}$, respectively, which is optimum in the sense of the Sidelnikov lower bound for binary sequences. Here, $m \geq 3$ is an integer not equal to a multiple of 4, the length of the sequence is $N = 2^m - 1$, and $\lfloor x \rfloor$ denotes the largest integer less than or equal to x [7]. The Walsh sequences are generated by mapping $\{0, 1\}$ onto $\{-1, 1\}$ for codeword rows of square Hadamard matrices.

In order to obtain sequences having better correlation properties, a number of polyphase sequences have been suggested [8–29]. Using the phase diversity in a chip, polyphase sequences can be so designed as to be more suitable for than binary (two-phase) sequences channelization and scrambling sequences in cellular DSCDMA systems [30,31].

2. SOME PRELIMINARIES

Figure 1 shows the correlation of two sequences in asynchronous DSCDMA systems when the chip synchronization is assumed to be perfect, where k_1 and k_2 denote the user index and $x_j^{(k_i)}$ denotes the jth chip of the k_i(th) sequence. For M-ary phase shift keying (MPSK) systems, the general correlation function [32–34] between two polyphase sequences $\mathbf{x}^{(k_1)}$ and $\mathbf{x}^{(k_2)}$ of length N is

$$\theta_\gamma(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) + \gamma C^*(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l - N) \tag{1}$$

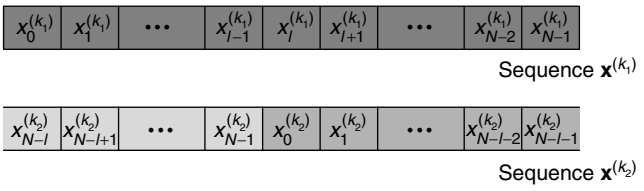


Figure 1. Correlation of two sequences.

where

$$C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = \begin{cases} \sum_{i=0}^{N-1-l} x_i^{(k_1)} \{x_{i+l}^{(k_2)}\}^*, & 0 \leq l \leq N-1 \\ \sum_{i=0}^{N-1+l} x_{i-l}^{(k_1)} \{x_i^{(k_2)}\}^*, & 1-N \leq l < 0 \\ 0, & |l| \geq N \end{cases} \tag{2}$$

is the partial correlation

$$\gamma \in \{W_M^t \mid t = 0, 1, \ldots, M-1\} \tag{3}$$

$$W_A^k = e^{2\sqrt{-1}\pi k/A} \tag{4}$$

and A is a natural number with $*$ denoting the complex conjugate. For binary ($M = 2$) phase shift keying (BPSK) systems, we have two important functions, the *even correlation* (EC)

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) + C^*(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l - N) \tag{5}$$

and the *odd correlation* (OC)

$$\hat{\theta}(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = C(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) - C^*(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l - N), \tag{6}$$

from Eq. (1). When $k_1 = k_2$, Eqs. (5) and (6) are called the "even AC (EAC)" and "odd AC (OAC)," respectively, and when $k_1 \neq k_2$, Eqs. (5) and (6) are called the "even CC (ECC)" and "odd CC (OCC)," respectively.

In order to completely analyze the performance of asynchronous DSCDMA systems using a particular class of sequences in an environment of multiple simultaneous users, we should consider not only the EC properties but also the OC properties of sequences. The OC function affects the output of the correlator when the information symbols change over one integration interval, while the EC function affects the output when the information symbols do not change. Thus, when the binary information symbols are equiprobable, both the EC and OC functions are equally important in the system design and performance analysis. Designing sequences with good OC properties is a difficult problem, as observed by other authors [35–37].

3. POLYPHASE SEQUENCES FOR CELLULAR DSCDMA SYSTEMS

3.1. Polyphase Sequences with Perfect EAC

We first consider several polyphase sequences that have perfect EAC properties. Since the good AC properties of such sequences guarantee low interference among the different timeshifts, they are useful as the long scrambling sequences in intercell synchronous cellular DSCDMA systems:

Frank sequence [22] — the sequence of length $N = M^2$ is defined by

$$x_{nM+k} = W_M^{nk}, \quad n, k \in \{0, 1, \ldots, M-1\} \tag{7}$$

Golomb sequence [18,23] — the Golomb sequence of length N is defined by

$$x_n = W_N^{n(n+1)/2}, \quad n = 0, 1, \ldots, N-1 \tag{8}$$

P1 sequence [11] — the P1 sequence of length $N = M^2$ is defined by

$$x_{nM+k} = W_{2M}^{-(M-2n-1)(nM+k)}, \quad n, k \in \{0, 1, \ldots, M-1\} \tag{9}$$

Px sequence [25] — the Px sequence of length $N = M^2$ is defined by

$$x_{nM+k} = \begin{cases} W_{4M}^{(M-2n-1)(M-2k-2)}, & M \text{ even}, \\ W_{4M}^{(M-2n-1)(M-2k-1)}, & M \text{ odd}, \end{cases}$$
$$n, k \in \{0, 1, \ldots, M-1\}. \tag{10}$$

P3 sequence [12] — the P3 sequence of length N is defined by

$$x_n = W_{2N}^{n^2}, \quad n = 0, 1, \ldots, N-1 \tag{11}$$

P4 sequence [13] — the P4 sequence of length N is defined by

$$x_n = W_{2N}^{n(n-N)}, \quad n = 0, 1, \ldots, N-1 \tag{12}$$

In Eqs. (7–12), M and N are natural numbers. The EAC properties of all the sequences listed above are perfect:

$$\theta(\mathbf{x}, \mathbf{x})(l) = \begin{cases} N, & l = 0 \\ 0, & l \neq 0 \end{cases} \tag{13}$$

For example, we obtain Table 1 when $N = 16$. When the length of sequences is 16, the numbers of phases of the Frank, P1, Golomb, Px, P3, and P4 sequences are 4, 8, 16, 16, 32, and 32, respectively. For the sequences shown in Table 1, we have calculated the normalized EAC and OAC functions as shown in Fig. 2; thus we can clearly confirm the perfect EAC property of the sequences. Although each of the six sequences has its own unique defining equation, the EAC properties of the sequences are all the same (at least in the sense of magnitude) and the OAC properties are quite similar to each other. The Frank sequence has the same AC function as the P1 sequence, whose normalized maximum magnitude of the out-of-phase OAC value is 0.1768. The Golomb sequence has the same AC function as do the P3 and P4 sequences, whose normalized maximum magnitude of the out-of-phase OAC value is 0.2310. The maximum magnitude of the out-of-phase OAC value of the Px sequence is the same as that of the Frank sequence, although the OAC functions of these two sequences are different. The peak : second-peak ratio of the Frank, P1, and Px sequences is better than those of the Golomb, P3, and P4 sequences. In addition, the mainlobe : total sidelobe energy ratio of the Px sequence is the lowest among the six sequences.

3.2. Polyphase Sequences with Optimum or Near-Optimum EC

Let $\theta_a = \max_{\mathbf{x}} \max_{l \neq 0} \theta(\mathbf{x}, \mathbf{x})(l)$ and $\theta_c = \max_{\mathbf{x} \neq \mathbf{y}} \max_l \theta(\mathbf{x}, \mathbf{y})(l)$, where $\mathbf{x}$ and $\mathbf{y}$ denote members in a class of polyphase sequences. Then, we have [35]

$$\max\{\theta_a, \theta_c\} \geq \sqrt{N} \tag{14}$$

Thus, if a set of sequences satisfies $\theta_a \leq \sqrt{N}$ and $\theta_c \leq \sqrt{N}$, the sequence is called *optimum* in the sense of the lower bound for polyphase sequences. We now consider some polyphase sequences whose EC functions are (nearly) optimum. These (near) optimum sequences are useful as the long scrambling sequences in intercell asynchronous cellular DSCDMA systems.

3.2.1. Four-Phase Sequence. The generating polynomial of four-phase sequences [16] is a primitive basic irreducible polynomial in $\mathbb{Z}_4[t]$, whose modulo 2 projections are primitive irreducible polynomials in $\mathbb{Z}_2[t]$. If $g(t)$ is a primitive basic irreducible polynomial of degree m, the set of four-phase sequences has period $N = 2^m - 1$ and size $N + 1$. Let the generating polynomial of degree m for the four-phase sequence be

$$g(t) = g_m t^m + g_{m-1} t^{m-1} + \cdots + g_1 t + g_0 \tag{15}$$

where $g_i \in \{0, 1, 2, 3\}$, $g_m \neq 0$, and $g_0 \neq 0$. Then, the recurrence condition of the quaternary sequence $\mathbf{s} = [s_0\, s_1\, \cdots\, s_{N-1}]$ is

$$s_{i+m} = g_{m-1} s_{i+m-1} + g_{m-2} s_{i+m-2} + \cdots + g_1 s_{i+1} + s_i \pmod 4, \text{ for } i \geq 0 \tag{16}$$

and the four-phase sequence $\mathbf{x} = [x_0\, x_1\, \cdots\, x_{N-1}]$ can be obtained from $x_i = W_4^{s_i}$. For example, let $m = 3$ and $g(t) = t^3 + 3t^3 + 2t + 3$. Then, $N = 2^m - 1 = 7$ and the recurrence condition is $t^3 = -3t^2 - 2t - 3 = t^2 + 2t + 1 \pmod 4$. There are $4^m - 1 = 63$ possible nonzero initial

Table 1. Examples of Some Sequences with Perfect EAC When Sequences Length Is 16

Sequence	Value of A in W_A^i (number of phases)	Value of i in W_A^i															
Frank	4	0	0	0	0	0	1	2	3	0	2	0	2	0	3	2	1
P1	8	0	5	2	7	4	3	2	1	0	1	2	3	4	7	2	5
Golomb	16	0	1	3	6	10	15	5	12	4	13	7	2	14	11	9	8
Px	16	9	3	13	7	3	1	15	13	13	15	1	3	7	13	3	9
P3	32	0	1	4	9	16	25	4	17	0	17	4	25	16	9	4	1
P4	32	0	15	28	7	16	23	28	31	0	31	28	23	16	7	28	15

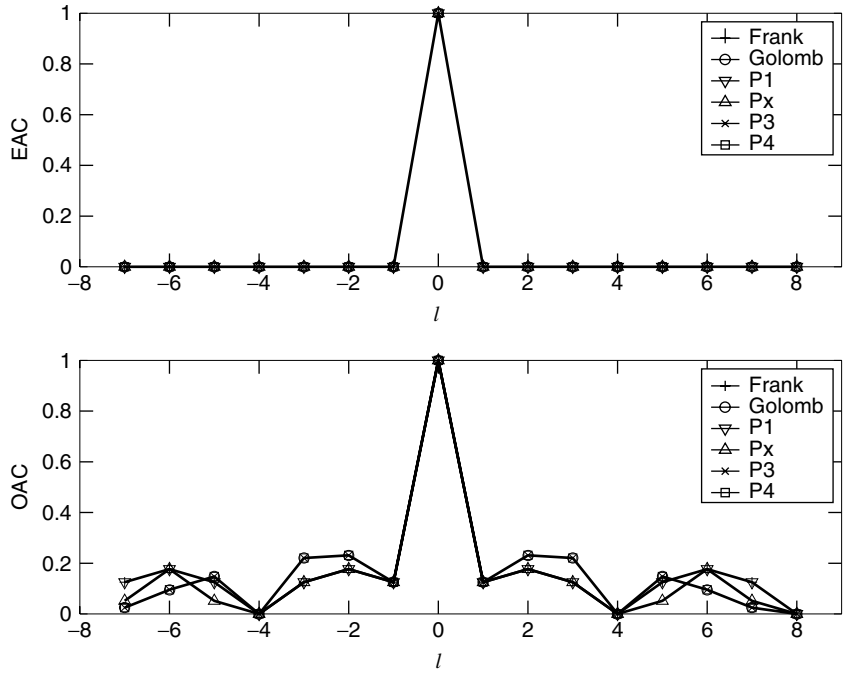


Figure 2. Normalized magnitudes of the AC functions of the sequences in Table 1.

loadings for the mod 4 shift register circuit, but only a collection of $(4^m - 1)/N = N + 2 = 9$ of them will yield cyclically distinct sequences. One of the 9 quaternary sequences is $\mathbf{s} = [1001132]$, and the corresponding four-phase sequence is

$$\mathbf{x} = [W_4^1\ W_4^0\ W_4^0\ W_4^1\ W_4^1\ W_4^3\ W_4^2] \tag{17}$$

The maximum nontrivial correlation magnitude of the four-phase sequence is

$$\max\{\theta_a, \theta_c\} \le 1 + \sqrt{N+1} \tag{18}$$

As in the case of the (binary) Gold sequence, the exact distribution of EC values of the four-phase sequence is known.

3.2.2. Frank–Zadoff–Chu (FZC) Sequence. Let p denote the smallest prime divisor of an odd number N and M_k denote the multiplicative inverse of k mod N, $k = 1, \ldots, p-1$. Then the set $\{\mathbf{x}^{(k)}; k = 1, \ldots, p-1\}$ of $p-1$ FZC sequences [8,10] is defined by

$$\begin{aligned} x_n^{(k)} &= (-1)^{nM_k} W_{2N}^{M_k n^2} \\ &= W_{2N}^{nM_k(n+N)}, \quad n = 0, \ldots, N-1 \end{aligned} \tag{19}$$

When N is prime, there can be as many as $N-1$ FZC sequences. An example of the FZC sequence, when $N = 7$ and $k = 1$, is

$$\mathbf{x}^{(1)} = [W_7^0\ W_7^4\ W_7^2\ W_7^1\ W_7^1\ W_7^2\ W_7^4] \tag{20}$$

The correlation properties of the FZC sequence include

$$\theta(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})(l) = \begin{cases} N, & l = 0 \\ 0, & l \ne 0 \end{cases} \tag{21}$$

for the EAC function, and

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) \le \sqrt{N}, \quad k_1 \ne k_2 \tag{22}$$

for the ECC function.

3.2.3. Generalized Chirplike (GCL) Sequence. Let $\{b_0, b_1, \ldots, b_{m-1}\}$ be a set of m complex numbers all having absolute value 1, and let $\{a_0, a_1, \ldots, a_{N-1}\}$ be a set of $N = sm^2$ numbers defined by

$$a_n = \begin{cases} W_N^{-n^2/2-qn}, & N \text{ even} \\ W_N^{-n(n+1)/2-qn}, & N \text{ odd} \end{cases} \tag{23}$$

where q is an integer and m and s are natural numbers. Then, the GCL sequences [17–21] $\mathbf{x}^{(k)}$, $k = 0, 1, \ldots, N-1$, are defined by

$$x_n^{(k)} = W_N^k a_n b_{n \bmod m}, \quad n = 0, 1, \ldots, N-1. \tag{24}$$

For example, when $s = 2$, $m = 2$ (i.e., $N = 8$), $q = 0$, $k = 0$, and $b_0 = b_1 = 1$, we have

$$\mathbf{x}^{(0)} = [W_{16}^0\ W_{16}^{15}\ W_{16}^{12}\ W_{16}^7\ W_{16}^0\ W_{16}^7\ W_{16}^{12}\ W_{16}^{15}] \tag{25}$$

The EAC function of the GCL sequence is

$$\theta(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})(l) = \begin{cases} N, & l = 0 \\ 0, & l \ne 0 \end{cases} \tag{26}$$

The ECC between two GCL sequences of odd length N, obtained from the two different primitive Nth roots $W_N^{k_1}$ and $W_N^{k_2}$ of unity, is constant if $k_1 - k_2$ is relatively prime to N:

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = \sqrt{N}, \quad N \text{ odd} \tag{27}$$

3.2.4. Lüke Sequence. The construction of the family of Lüke sequences [15] starts with a q-ary m sequence (see Section 3.4) $\mathbf{c} = [c_0\, c_1 \cdots c_{N-1}]$ of length $N = q^r - 1$, where q is prime and $r \geq 2$ is an integer. The Lüke sequence can then be generated by

$$x_n^{(k)} = W_{qN}^{Nc_n + knq}, \quad n, k \in \{0, 1, \ldots, N-1\} \tag{28}$$

For example, when $q = 3$, $r = 2$, $N = 8$, $k = 1$, and $\mathbf{c} = [12022101]$, we have

$$\mathbf{x}^{(1)} = [W_{24}^{8}\ W_{24}^{19}\ W_{24}^{6}\ W_{24}^{1}\ W_{24}^{4}\ W_{24}^{23}\ W_{24}^{18}\ W_{24}^{5}] \tag{29}$$

The EAC function of the Lüke sequence is two-level in magnitude and nearly perfect:

$$\theta(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})(l) = \begin{cases} N, & l = 0 \\ 1, & l \neq 0 \end{cases} \tag{30}$$

Interestingly, the ECC function is also two-level in magnitude:

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = \begin{cases} 0, & l = 0, \\ \sqrt{N+1}, & l \neq 0, \end{cases} \tag{31}$$

for $k_1 \neq k_2$.

3.3. Polyphase Sequences with Perfect ECC

We now consider some polyphase sequences that have perfect ECC properties and are useful as the signature (channelization) sequences in cellular DSCDMA systems.

3.3.1. Park–Park–Song–Suehiro (PS) Sequence. Let $\mathbf{b} = [b_0\, b_1 \cdots b_{N_b-1}]$ be a sequence of length N_b with elements $b_i \in \{W_M^0, \ldots, W_M^{M-1}\}$, $i = 0, 1, \ldots, N_b - 1$, where M is a natural number. Then, the PS sequence [28] $x^{(k)}$ of length $N = KN_b^2$ is defined by

$$x_n^{(k)} = W_N^{nk} \sum_{p=0}^{N_b-1} b_p W_{N_b^2}^{np} \delta(R(n + mp, N_b)) \qquad n = 0, 1, \ldots, N-1, \quad k = 0, 1, \ldots, K-1 \tag{32}$$

where $\delta(\cdot)$ is the Kronecker delta function, K is a natural number, $R(a, b)$ is the remainder when a is divided by b, and m is a natural number less than N_b. If N_b is prime, we have

$$x_n^{(k)} = b_{p_s} W_N^{n(k + Kp_s)} \tag{33}$$

where p_s is the number in $\{0, 1, \ldots, N_b - 1\}$ such that $R(n + mp_s, N_b) = 0$. For example, when $N_b = 2$, $K = 2$, $m = 1$, $k = 1$, and $\mathbf{b} = [W_2^0\ W_2^1]$, we have

$$\mathbf{x}^{(1)} = [W_8^0\ W_8^7\ W_8^2\ W_8^5\ W_8^4\ W_8^3\ W_8^6\ W_8^1] \tag{34}$$

3.3.2. Song–Park (SP) Sequence. The SP sequence [29] of length $N = 2(L + 1)$ is defined by

$$x_n^{(k)} = (-1)^n W_{L+1}^{nk}, \quad n = 0, 1, \ldots, N-1, \tag{35}$$

where the even integer L is the size of the SP sequence. For example, when $L = 2$ and $k = 1$, we have

$$\mathbf{x}^{(1)} = [W_6^0\ W_6^5\ W_6^4\ W_6^3\ W_6^2\ W_6^1] \tag{36}$$

The ECC functions of these two sequences [Eqs. (32) and (35)] are perfect:

$$\theta(\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)})(l) = 0, \quad k_1 \neq k_2 \tag{37}$$

In order to compare the OCC properties of the PS and SP sequences, we show the cumulative frequencies of the normalized maximum magnitude of the OCC (MMO) values in Fig. 3 when the lengths of sequences are almost the same (the lengths of the PS sequence are 36, 126, and 513 and those of the SP sequence are 30, 126, and 510). The distribution curves are obtained by evaluating the MMO for all possible ${}_SC_2$ pairs of sequences, where S is the number of sequences (which depends on N). For example, when $N = 126$, the cumulative frequency for the PS and SP sequences in Fig. 3b is obtained from ${}_{14}C_2$ and ${}_{62}C_2$ values of the MMO, respectively. In this figure, we can clearly see that the MMO of the SP sequence is smaller than that of the PS sequence with high probability. In addition, the cumulative frequency for the SP sequence converges to 1 at a much faster rate than that for the PS sequence.

3.4. Other Polyphase Sequences

Among the other interesting polyphase sequences are the q-phase m sequences, equal OC and EC (EOE) sequences, and generalized Barker sequences.

3.4.1. *q*-Phase *m* Sequences. q-phase m sequences are obtained by expanding the number of phases of binary m sequences. The q-phase m sequence $\mathbf{x}$ of period $N = q^m - 1$ can be defined by $x_i = W_q^{c_i}$, where $\mathbf{c} = [c_0\, c_1 \cdots c_{N-1}]$ is called the "q-ary m sequence." When q is prime, the generating polynomial

$$g(t) = g_m t^m + g_{m-1} t^{m-1} + \cdots + g_1 t + g_0 \tag{38}$$

for q-ary m sequences is a primitive polynomial of degree m. Here, $g_i \in \{0, 1, \ldots, q-1\}$, $g_m \neq 0$, and $g_0 \neq 0$. The recurrence condition of q-ary m sequences $\mathbf{c}$ is

$$c_{i+m} = g_{m-1} c_{i+m-1} + g_{m-2} c_{i+m-2} + \cdots + g_1 c_{i+1} + c_i \pmod q, \text{ for } i \geq 0 \tag{39}$$

For example, let $q = 3$, $m = 2$, and $g(t) = t^2 + 2t + 2$. Then, the recurrence becomes $t^2 = -2t - 2 = t + 1 \pmod 3$. Using the initial loading of 21 in the registers yields the ternary sequence of period $N = 3^2 - 1 = 8$ as $\mathbf{c} = [c_0\, c_1 \cdots c_{N-1}] = [12022101]$, and a three-phase m sequence can be obtained as

$$\mathbf{x} = [W_3^1\ W_3^2\ W_3^0\ W_3^2\ W_3^2\ W_3^1\ W_3^0\ W_3^1] \tag{40}$$

The EAC function of q-phase m sequences is

$$\theta(\mathbf{x}, \mathbf{x})(l) = \begin{cases} N, & l = 0 \\ -1, & l \neq 0 \end{cases} \tag{41}$$

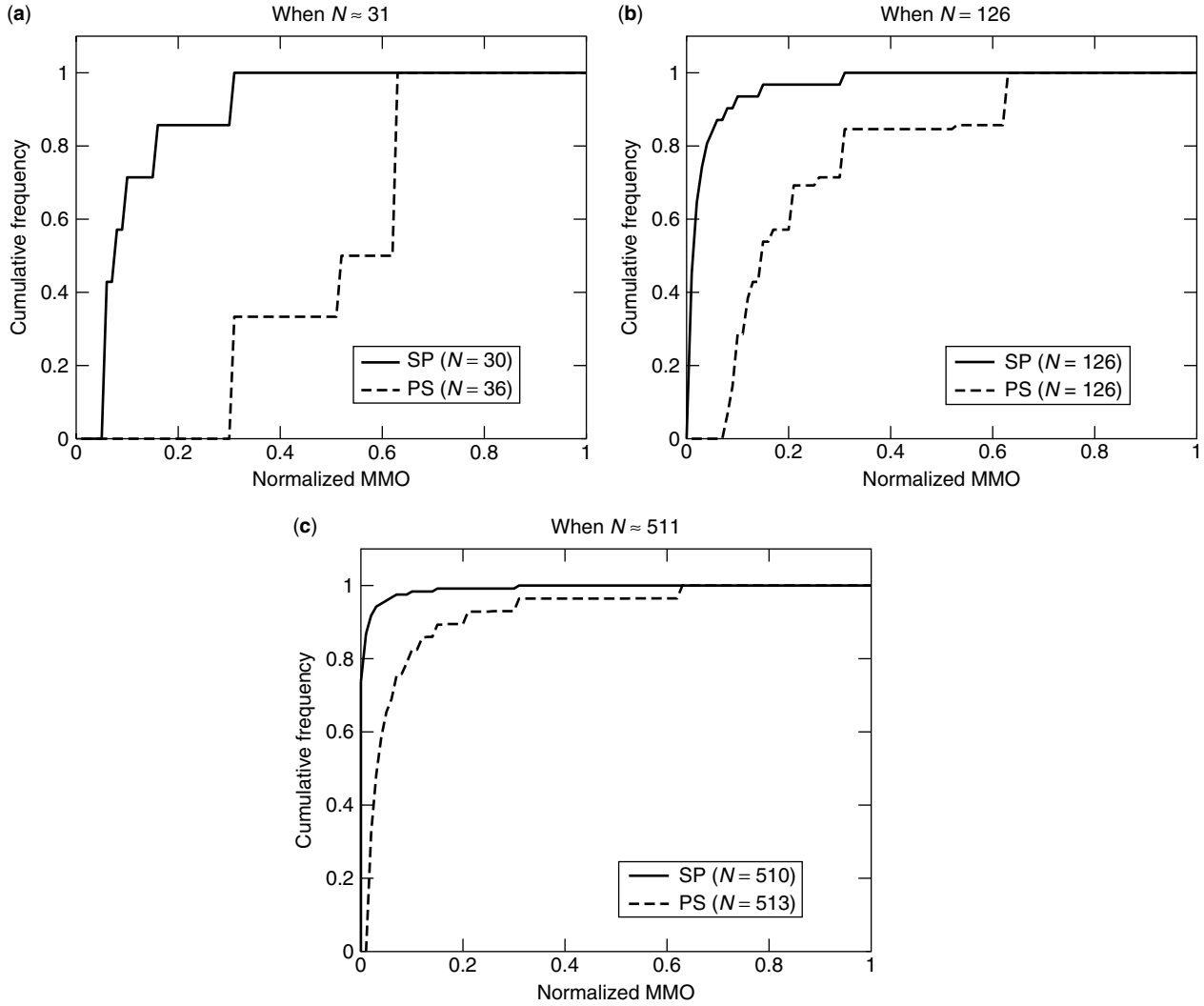


Figure 3. Cumulative frequency of the normalized MMO values of the PS and SP sequences.

3.4.2. EOE Sequences. When $\mathbf{u} = [u_0\, u_1 \cdots u_{N-1}]$ is an arbitrary real-valued sequence of period N,p is an odd integer and β is a real constant satisfying $0 \le \beta < 2\pi$, the complex-valued sequence $\mathbf{x} = [x_0\, x_1 \cdots x_{N-1}]$ defined by

$$x_n = u_n e^{\sqrt{-1}\beta} W_{4N}^{pn}, \quad n = 0, 1, \ldots, N-1 \tag{42}$$

is an EOE sequence [20]. For example, when $N = 7, p = 7$, $\beta = 0$, and $\mathbf{u} = [1\ -1\ -1\quad 1\ -1\ 1\ 1]$ is a Gold sequence, an EOE–Gold sequence with length 7 can be obtained as

$$\begin{aligned}\mathbf{x} &= [W_4^0\ -W_4^1\ -W_4^2\ W_4^3\ -W_4^0\ W_4^1\ W_4^2] \\ &= [W_4^0\ W_4^3\ W_4^0\ W_4^3\ W_4^2\ W_4^1\ W_4^2]\end{aligned} \tag{43}$$

The magnitudes of the OC and EC functions of these sequences are equal:

$$|\theta(\mathbf{x}, \mathbf{y})(l)| = |\hat{\theta}(\mathbf{x}, \mathbf{y})(l)| \tag{44}$$

In addition, we have

$$|\theta(\mathbf{x}, \mathbf{y})(l)| = |\hat{\theta}(\mathbf{x}, \mathbf{y})(l)| \le \max\{\theta(\mathbf{u}, \mathbf{v})(l), \hat{\theta}(\mathbf{u}, \mathbf{v})(l)\} \tag{45}$$

where $\mathbf{y} = [y_0\, y_1 \cdots y_{N-1}]$ denotes an EOE sequence defined with $\mathbf{v} = [v_0\, v_1 \cdots v_{N-1}]$:

$$y_n = v_n e^{\sqrt{-1}\beta} W_{4N}^{pn}, \quad n = 0, 1, \ldots, N-1 \tag{46}$$

3.4.3. Generalized Barker Sequences. A polyphase sequence $\mathbf{x}$ is called a "generalized Barker sequence" [9,24] if the partial AC function satisfies $C(\mathbf{x}, \mathbf{x})(l) \le 1$ for $l \ne 0$. Generalized Barker sequences are widely used as a synchronization sequence because the partial AC property is so good. Let $\mathbf{y} = [y_0\, y_1 \cdots y_{N-1}]$ be a p-phase sequence of length N. Then the q-phase generalized Barker sequence $\mathbf{x}$ of length N is defined by

$$x_n = y_n W_M^n \tag{47}$$

where M is a nonzero integer and q is the least common multiple of p and M. It is easy to see that the partial AC of $\mathbf{x}$ is

$$C(\mathbf{x}, \mathbf{x})(l) = W_M^{-n} C(\mathbf{y}, \mathbf{y})(l), \quad \text{for all } l \tag{48}$$

In particular, we have $|C(\mathbf{x},\mathbf{x})(l)| = |C(\mathbf{y},\mathbf{y})(l)|$ for all l as well as $|x_n| = |y_n|$ since $|W_M^{-n}| = 1$. As a special case, if $|C(\mathbf{y},\mathbf{y})(l)| \le 1$ for all l and $y_n \in \{1,-1\}$, we can obtain a polyphase generalized Barker sequence of length N from a binary Barker sequence of length N. For example, using a binary Barker sequence $\mathbf{y} = [1\ 1\ 1\ -1\ -1\ 1\ -1]$ with length 7, we obtain a four-phase generalized Barker sequence with length 7 as

$$\mathbf{x} = [W_4^0\ W_4^1\ W_4^2\ W_4^1\ W_4^2\ W_4^1\ W_4^0] \tag{49}$$

It has been shown that the binary Barker sequences with odd length $N > 13$ and even length $4 < N < 189884$ do not exist: on the other hand, polyphase sequences of length $N > 13$ satisfying the Barker condition can be found; for example, the 180-phase generalized Barker sequence with length 36 has been reported [24].

4. CONCLUSION

In this article, we have reviewed some polyphase sequences and described their applications to cellular DSCDMA systems based on the correlation properties. The Frank, Golomb, P1, Px, P3, and P4 sequences have a perfect EAC property so that they can be used as scrambling sequences in intercell synchronous cellular DSCDMA systems. The four-phase, FZC, GCL, and Lüke sequences, whose EC functions are (near) optimum in the sense of the lower bound for polyphase sequences, are useful as the scrambling sequences in intercell asynchronous cellular DSCDMA systems. Next, the PS and SP sequences are suitable for the short signature (channelization) sequences in cellular DSCDMA systems because of their perfect ECC properties. Some other polyphase sequences such as the q-phase m, EOE, and generalized Barker sequences are also described briefly.

Acknowledgments
This research was supported by Korea Science and Engineering Foundation (KOSEF) under Grant R01-2000-000-00259-8, for which the authors would like to express their thanks.

BIOGRAPHIES

So Ryoung Park received the B.S. degree in Electronics Engineering from Yonsei University, Seoul, Korea, in 1997, and the M.S.E. and Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1999 and 2002, respectively. She was a Research Assistant at the Department of Electrical Engineering, KAIST, from 1997 to 2001 and is now a Research Scientist at the Statistical Signal Processing Laboratory, Department of Electrical Engineering and Computer Science, KAIST. She was the recipient of a Gold Prize at the Samsung Humantech Paper Contest in 2001. Her current research interests are in mobile communications and statistical signal processing with emphasis on spread-spectrum communications.

Iickho Song received the B.S.E. (magna cum laude) and M.S.E. degrees in Electronics Engineering from Seoul National University, Seoul, Korea, in 1982 and 1984, respectively, and the M.S.E. and Ph.D. degrees in Electrical Engineering from University of Pennsylvania, Philadelphia (USA), in 1985 and 1987, respectively. In 1988, he joined KAIST, as an Assistant Professor, where he has been a full Professor since 1998. He is a Fellow and Chartered Engineer of the Institution of Electrical Engineers, and a Senior Member of the Institute of Electrical and Electronics Engineers. He has coauthored a book (Advanced Theory of Signal Detection, Springer, 2002) and more than 150 papers in the area of signal detection and estimation and of code design and acquisition for CDMA systems. He received a number of awards, including the Young Scientists Award presented by the President of the Republic of Korea and four Academic Awards from the Korean Institute of Communication Sciences. He has been listed in many "who's who" books, including *Marquis Who's Who in the World* and *Marquis Who's Who in Science and Engineering*. His research interest include detection and estimation, statistical signal processing, and CDMA communications.

BIBLIOGRAPHY

1. R. A. Scholtz, The origins of spread-spectrum communications, *IEEE Trans. Commun.* **COM-30**: 822–854 (1982).
2. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications: a tutorial, *IEEE Trans. Commun.* **COM-30**: 855–884 (1982).
3. H. Ochsner, Direct-sequence spread-spectrum receiver for communication on frequency-selective fading channels, *IEEE J. Select. Areas Commun.* **JSAC-5**: 188–193 (1987).
4. E. Geraniotis and B. Ghaffari, Performance of binary and quaternary direct-sequence spread-spectrum multiple-access systems with random signature sequences, *IEEE Trans. Commun.* **COM-39**: 713–724 (1991).
5. P. G. Flikkema, Spread-spectrum techniques for wireless communication, *IEEE Signal Process. Mag.* **14**: 26–36 (1997).
6. E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* **36**: 48–54 (1998).
7. R. Gold, Maximal recursive sequences with 3-valued recursive crosscorrelation functions, *IEEE Trans. Inform. Theory* **IT-14**: 154–156 (1968).
8. R. L. Frank and S. A. Zadoff, Phase shift pulse codes with good periodic correlation properties, *IRE Trans. Inform. Theory* **IT-8**: 381–382 (1962).
9. S. W. Golomb and R. A. Scholtz, Generalized Barker sequences, *IEEE Trans. Inform. Theory* **IT-11**: 533–537 (1965).
10. D. C. Chu, Polyphase codes with good periodic correlation properties, *IEEE Trans. Inform. Theory* **IT-18**: 531–532 (1972).
11. B. L. Lewis and F. F. Kretschmer, Jr., A new class of polyphase pulse compression codes and techniques, *IEEE Trans. Aerospace. Electron. Syst.* **AES-17**: 364–372 (1981).
12. B. L. Lewis and F. F. Kretschmer, Jr., Linear frequency modulation derived polyphase pulse compression codes, *IEEE Trans. Aerospace. Electron. Syst.* **AES-18**: 637–641 (1982).

13. B. L. Lewis and F. F. Kretschmer, Jr., Doppler properties of polyphase coded pulse compression waveform, *IEEE Trans. Aerospace. Electron. Syst.* **AES-19**: 521–531 (1983).

14. N. Suehiro and M. Hatori, Modulatable orthogonal sequences and their application to SSMA systems, *IEEE Trans. Inform. Theory* **IT-34**: 93–100 (1988).

15. H. D. Lüke, Families of polyphase sequences with near-optimal two-valued auto- and cross-correlation functions, *Electron. Lett.* **28**: 1–2 (1992).

16. S. Boztas, R. Hammons, and P. V. Kumar, 4-phase sequences with near-optimum correlation properties, *IEEE Trans. Inform. Theory* **IT-38**: 1101–1113 (1992).

17. B. M. Popović, Generalized chirp-like polyphase sequences with optimum correlation properties, *IEEE Trans. Inform. Theory* **IT-38**: 1406–1409 (1992).

18. N. Zhang and S. W. Golomb, Polyphase sequence with low autocorrelations, *IEEE Trans. Inform. Theory* **IT-39**: 1085–1089 (1993).

19. B. M. Popović, GCL polyphase sequences with minimum alphabets, *Electron. Lett.* **30**: 106–107 (1994).

20. H. Fukumasa, R. Kohno, and H. Imai, Design of pseudonoise sequences with good odd and even correlation properties for DS/CDMA, *IEEE J. Select. Areas Commun.* **JSAC-12**: 828–836 (1994).

21. B. M. Popović, Efficient matched filter for the generalized chirp-like polyphase sequences, *IEEE Trans. Aerospace. Electron. Syst.* **AES-30**: 769–777 (1994).

22. P. Z. Fan and M. Darnell, Aperiodic autocorrelation of Frank sequences, *IEE Proc. Commun.* **142**: 210–215 (1995).

23. E. M. Gabidulin, P. Z. Fan, and M. Darnell, Autocorrelation of Golomb sequences, *IEE Proc. Commun.* **142**: 281–284 (1995).

24. S. W. Golomb and M. Z. Win, Recent results on polyphase sequences, *IEEE Trans. Inform. Theory* **IT-44**: 817–824 (1998).

25. P. B. Rapajic and R. A. Kennedy, Merit factor based comparison of new polyphase sequences, *IEEE Commun. Lett.* **2**: 269–270 (1998).

26. S. I. Park et al., A noise reduction method for a modulated orthogonal sequence under impulsive noise environments, *IEICE Trans. Fund.* **E82A**: 2259–2265 (1999).

27. S. I. Park, S. R. Park, I. Song, and S. Yoon, On the generation and analysis of a modulated orthogonal sequences, *Signal Process.* **80**: 451–464 (2000).

28. S. I. Park, S. R. Park, I. Song, and N. Suehiro, Multiple access interference reduction for QS-CDMA systems with a novel class of polyphase sequences, *IEEE Trans. Inform. Theory* **IT-46**: 1448–1458 (2000).

29. S. R. Park, I. Song, S. Yoon, and J. Lee, A new polyphase sequence with perfect even and good odd crosscorrelation functions for DS/CDMA systems, *IEEE Trans. Vehic. Technol.* **VT-51** (in press).

30. T. M. Lok and J. S. Lehnert, An asymptotic analysis of DS/SSMA communication systems with random polyphase signature sequences, *IEEE Trans. Inform. Theory* **IT-42**: 129–136 (1996).

31. S. R. Park, I. Song, S. Yoon, and S. Y. Kim, A statistical analysis of random polyphase signature sequences in multipath fading DS-CDMA channels, *Signal Process.* **81**: 2461–2477 (2001).

32. M. B. Pursley, Performance evaluation for phase-coded spread-spectrum multiple-access communication, *IEEE Trans. Commun.* **COM-25**: 795–803 (1977).

33. D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudo random and related sequences, *Proc. IEEE* **68**: 593–618 (1980).

34. F. W. Sun and H. Leib, Optimal phases for a family of quadriphase CDMA sequences, *IEEE Trans. Inform. Theory* **IT-43**: 1205–1217 (1997).

35. D. V. Sarwate, Bounds on crosscorrelation of sequences, *IEEE Trans. Inform. Theory* **IT-25**: 720–724 (1979).

36. K. G. Paterson and P. J. G. Lothian, Bounds on partial correlations of sequences, *IEEE Trans. Inform. Theory* **IT-44**: 1164–1175 (1998).

37. S. R. Park, I. Song, and H. Kwon, DS/CDMA signature sequences based on PR-QMF banks, *IEEE Trans. Signal Process.* **SP-50** (in press).

POWER CONTROL IN CDMA CELLULAR COMMUNICATION SYSTEMS

MATTI RINTAMÄKI
Helsinki University of Technology
Helsinki, Finland

1. INTRODUCTION

Transmitter power control (TPC) is vital for capacity and performance in cellular communication systems, where high interference is always present as a result of frequency reuse. The basic intent is to control the transmitter powers in such a way that the interference power from each transmitter to other cochannel users (users that share the same radio resource simultaneously) is minimized while preserving sufficient quality of service (QoS) among all users. Cochannel interference management is important in any system employing frequency reuse. However, in CDMA there are interfering users both inside and outside a cell, which makes CDMA interference limited. Thus efficient TPC is essential in CDMA,[1] especially in the uplink (from mobile to base station communication).

Consider the situation depicted in Fig. 1. Mobile stations MS1 and MS2 share the same frequency band, and their signals are separable at the base station (BS) by their unique spreading codes. The link attenuation of MS2 a particular time instant might be several tens of decibels greater than that from MS1 to BS. If power control is not applied, the signal of MS1 will overpower the signal of MS2 at the base station. This is called the *near–far effect*. To mitigate this effect, power control aims to set the

[1] This applies to direct-sequence CDMA (DS-CDMA). In frequency-hopping CDMA (FHCDMA) the intracell interference can be made very small. In this article we concentrate on DS-CDMA, where transmitter power control is more critical. Thus throughout the rest of the text, DS-CDMA is referred to simply as CDMA.

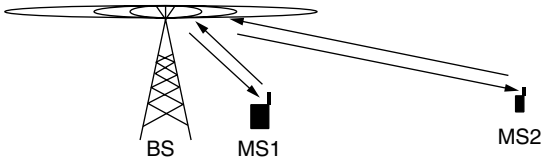


Figure 1. Near–far effect.

transmitter powers of MS1 and MS2 so that both signals are received at the same power level at the base station.

1.1. Radio-Link Attenuation

The attenuation of a radio signal can be modeled as a product of three effects: path loss, shadowing, and multipath fading. *Path loss* is the large-scale distance-dependent attenuation in the average signal power, which can be modeled as r^{-d}, where r is the distance in meters and d is the path loss exponent with typical values ranging from 2 to 5. *Shadowing* is the medium-scale attenuation, which is caused by diffraction and shielding phenomena emanating from terrain variations. This results in relatively slow variations in the mean signal power over a distance of a few tens of wavelengths, caused by reflections, refractions, and diffractions of the signal from buildings, trees, rocks, and other objects. It can be modeled as a lognormally distributed random variable with zero decibels mean and a standard deviation of typically 4–12 dB. *Multipath fading* is the rapid fluctuation in the received signal power that is caused by the constructive and destructive addition of the signals that propagate through different paths with different delays from the transmitter to the receiver. This is usually modeled as a complex Gaussian process, resulting in a Rayleigh-distributed envelope and a classic Doppler spectrum. If a line-of-sight signal is present, the Rice distribution is a better model. Also Nakagami distribution has been widely used to model multipath fading.

With all the three effects included, the attenuation g of the signal power in linear scale can be modeled as

$$g = r^{-d} \times 10^{\zeta/10} \times A_f \tag{1}$$

where ζ is a Gaussian random variable with zero mean and a standard deviation of 4–12 and A_f is a random variable such that $(A_f)^{1/2}$ is either Rayleigh-, Rice- or Nakagami distributed with a classic Doppler spectrum.

As can be understood from above, the signal attenuation is a random variable. Thus, when power control is applied, it must adapt to the changing attenuation of the desired signals, as well as the changing interference conditions, since the attenuations of the cochannel users' signals are also changing, and those signals are power-controlled as well.

1.2. Uplink Versus Downlink Power Control

In CDMA the uplink transmission creates a near–far situation if power control is not used. This occurs because the signals of the different mobile stations propagate through different paths before reaching their serving base station. The task of power control is thus to vary the transmitter powers in order to compensate for the varying channel attenuations, so that the signals from the different mobile stations are received with equal powers at the base station. The requirement of the dynamic range of uplink power control can be of the order of 80 dB.

In downlink there is no near–far situation, since all signals transmitted by a base station propagate through the same path before reaching a mobile station. These signals can be made essentially orthogonal by using proper spreading codes. However, unnecessary high powers can be provided to mobile users near the cell border, thus creating unnecessarily high interference to the neighboring cells. Therefore, downlink power control is used to reduce this interference. The dynamic range of downlink power control is usually much smaller than in uplink, typically on the order of 20–30 dB. This limitation is because a high dynamic range could produce a near–far situation for the mobile stations, since the signals cannot be made perfectly orthogonal.

1.3. Quality Measures for Power Control

A great deal of the work on power control in CDMA cellular systems has focused on how to set the transmitter powers so that all users in the system have acceptable *bit energy-to-interference spectral density ratios* (E_b/I_0). This approach is based on the fairly reasonable assumption that the bit error probability (BEP) at a receiver is a strictly monotonically decreasing function of E_b/I_0. For instance, BEP in an *additive white Gaussian noise* (AWGN) channel decreases very quickly with increasing E_b/I_0. E_b/I_0 is closely related with another measure, namely, the *signal-to-interference ratio* (SIR), such that

$$\frac{E_b}{I_0} = \mathrm{SIR}\frac{W}{r} \tag{2}$$

where W is the transmission bandwidth in hertz and r is the data rate in bits per second (bps). The quantity W/r is called the *processing gain*. When the data rate is fixed, the SIR differs from E_b/I_0 by merely a scaling factor.

In digital communication systems the information to be transmitted is arranged in strings of bits called *frames*, and error correction coding is applied to each frame to further decrease the BEP after decoding. A frame is useless if there are still bit errors in the frame after decoding, and it must be discarded. Hence, depending on the service, a sufficiently low frame erasure rate (FER) must be guaranteed. However, long time delays are needed to obtain reliable estimates of BEP or FER. Since the channel conditions can change very rapidly, these delays might be unacceptably long in practice. Hence most of the attention in the power control field has been on SIR-based algorithms. Also algorithms based on received signal strength have been used, but it has been shown that SIR-based power control offers better performance [1]. The FER information can be used to adjust the target SIR, which a fast TPC algorithm is trying to achieve. This increases system capacity, since a worst-case setting of the SIR target is not required.

In addition to the SIR and FER requirements, the delay or *latency* requirements must be taken into account. For instance, a voice service tolerates a certain amount of data loss but is delay-critical, whereas file downloads do not tolerate bit errors at all (erroneous frames must be retransmitted), but the transmission need not be continuous and must only satisfy some delay limit on the average. The delay tolerances can be taken advantage of in the design of power control algorithms for non-real-time services.

2. THE SIR BALANCING PROBLEM

A widely studied approach to transmitter power control is the SIR balancing problem, namely, how to set the transmitter powers so that all users in the system have equal SIRs. This method is applicable for circuit-switched real-time services such as voice, where the data rate is fixed.

Figure 2 illustrates a simple two-cell CDMA system. We consider only uplink, but the same analysis applies in downlink. The mobile stations MS1 $\cdots$ MS3 share a common frequency band, and their signals are separable at the base stations by their unique spreading codes. The link attenuation from mobile j to base station i is denoted by g_{ij} and is assumed to be fixed in the analysis. In this "snapshot" approach [2,3], it is assumed that the link attenuations change slowly enough compared to the power control dynamics, and can thus be assumed constant. This assumption is reasonable if multipath fading is neglected, but generally leads to optimistic results.

Define a base station assignment function $b(i)$ so that $k = b(i)$ if mobile i is served by base station k. Then the uplink SIR requirement for user i can be expressed as

$$\gamma_i = \frac{g_{b(i)i}p_i}{\sum_{j=1, j\neq i}^{N} g_{b(i)j}p_j + \eta_i} = \frac{p_i}{\sum_{j=1, j\neq i}^{N} \frac{g_{b(i)j}}{g_{b(i)i}}p_j + \frac{\eta_i}{g_{b(i)i}}} \geq \gamma_i^t \qquad (3)$$

where γ_i is the SIR at the receiver of mobile station i, p_i is the transmission power of mobile i, N is the number of mobile stations using the same channel including the intracell and intercell users, η_i is the receiver noise power, and γ_i^t is the uplink SIR requirement for user i. Define the vectors $\mathbf{p} = \{p_i\}$ and $\boldsymbol{\eta} = \{\gamma_i^t \eta_i / g_{b(i)i}\}$ and matrix $\mathbf{H} = \{H_{ij}\}$ with elements $H_{ij} = \gamma_i^t g_{b(i)j}/g_{b(i)i}$ when $i \neq j$ and $H_{ii} = 0$. Now we can put (3) in matrix form:

$$(\mathbf{I} - \mathbf{H})\mathbf{p} \geq \boldsymbol{\eta} \qquad (4)$$

where $\mathbf{I}$ denotes the identity matrix and the inequality holds componentwise. A minimum-power solution corresponds to the case where (4) is satisfied with equality. This is desirable, as it saves energy and thus prolongs the mobile station battery life.

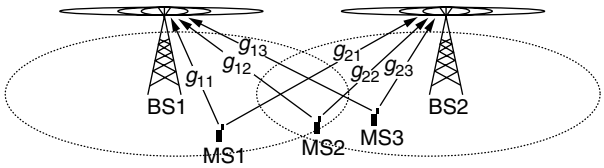


Figure 2. Illustration of a two-cell CDMA system with three mobile stations.

Definition 1. The problem is said to be feasible if there exists a nonnegative power vector $\mathbf{p}$ such that condition in (4) is satisfied.

It has been shown that balancing the SIRs and making the balanced SIR as high as possible maximizes the minimum SIR in all links. Consider the power control problem in (4) with $\gamma_i^t = \gamma^t$. Define matrix $\mathbf{A}$ so that $\mathbf{H} = \gamma^t \mathbf{A}$. If the interference in the receiver is sufficiently high that the receiver noise can be neglected, the problem can be identified as an eigenvalue problem:

$$\mathbf{p} = \mathbf{H}\mathbf{p} \Leftrightarrow \frac{1}{\gamma^t}\mathbf{p} = \mathbf{A}\mathbf{p} \qquad (5)$$

The maximum possible value for γ^t (denoted by γ^*) is equal to the inverse of the maximum (real) eigenvalue of $\mathbf{A}$. The corresponding positive eigenvector $\mathbf{p}^*$ is the power vector achieving this maximum.

If receiver noise cannot be neglected, the optimal power vector is a solution to a set of linear equations:

$$\mathbf{p}^* = \mathbf{H}\mathbf{p}^* + \boldsymbol{\eta} \qquad (6)$$

Proposition 1 [5]**.** The power control problem in (6) is feasible if the largest eigenvalue of the matrix $\mathbf{H}$, denoted by $\rho(\mathbf{H})$, is less than or equal to one.

Note that the case $\rho(\mathbf{H}) = 1$ can only be met if the receiver noise is zero, otherwise infinite transmitter power would be required. Moreover, in practice there always exists an upper limit for the transmitter power.

Using the optimal power vector $\mathbf{p}^*$ results in all users in the system having the same SIR. If the power control problem is not feasible, this results in the disastrous case that none of the users achieve the SIR requirement. To prevent this, a removal strategy must be employed, which removes transmitters from the channel until the power control problem becomes infeasible. The problem is to determine which transmitters to remove. An intuitive approach is to remove those transmitters that produce largest interference, that is, transmitters having the worst link quality. Several removal strategies have been investigated in the literature [4]. Note that a removal of a transmitter from a channel does not necessarily mean that the connection is broken, but it can be handed over to another channel.

3. DISTRIBUTED POWER CONTROL

Solving (6) directly is a centralized method, since it requires the information of all the link attenuations and receiver noise values in the system. This is generally not suitable in real implementations, since it would require extensive signaling overhead. However, it is valuable in determining upper bounds for the performance of distributed algorithms that can be implemented in practice.

A distributed algorithm uses only local measurements to update the transmitter powers. Hence it is more suitable for practical implementation than a centralized algorithm. Since in this case a user does not know the link attenuations, the problem must be iteratively solved. It is thus necessary to find an iteration that depends only on local measurements, and converges to the optimal solution reasonably soon (sooner than the link gains change). Fast convergence can be achieved in two ways: by making the iteration time step smaller, and by designing an iteration with faster convergence property.

A general iterative algorithm to solve the problem in (6) can be found from numerical linear algebra, and is given by [5]

$$\mathbf{p}(k+1) = \mathbf{M}^{-1}\mathbf{N}\mathbf{p}(k) + \mathbf{M}^{-1}\boldsymbol{\eta}, \quad k = 0, 1, \ldots \tag{7}$$

where $\mathbf{M}$ and $\mathbf{N}$ are matrices such that $\mathbf{p}^* = \mathbf{M}^{-1}\mathbf{N}\mathbf{p}^* + \mathbf{M}^{-1}\boldsymbol{\eta}$. By selecting $\mathbf{M}$ and $\mathbf{N}$ properly, the iteration in (7) will converge so that $\lim_{k\to\infty} \mathbf{p}(k) = \mathbf{p}^*$. For example, if we select $\mathbf{M} = \mathbf{I}$ and $\mathbf{N} = \mathbf{H}$, we get

$$\mathbf{p}(k+1) = \mathbf{H}\mathbf{p}(k) + \boldsymbol{\eta}, \quad k = 0, 1, \ldots \tag{8}$$

Looking at this equation componentwise, we have

$$\begin{aligned} p_i(k+1) &= \gamma_i^t \left(\sum_{j=1, j\neq i}^{N} \frac{g_{b(i)j}}{g_{b(i)i}} p_j(k) + \frac{\eta_i}{g_{b(i)i}} \right) \\ &= \frac{\gamma_i^t}{\gamma_i(k)} p_i(k), \quad k = 0, 1, \ldots \end{aligned} \tag{9}$$

where $\gamma_i(k)$ and $p_i(k)$ are the received SIR and transmission power of transmitter i at iteration k, respectively. This is referred to as the *distributed power control* (DPC) algorithm. It was proposed by Foschini and Miljanic [6]. It is totally distributed, since it depends only on local measurements of the SIR.

3.1. Convergence of the Iterative Algorithm

A necessary and sufficient condition for the iteration in (7) to converge is the following [5]. Let $\alpha_1, \alpha_2, \ldots$ be the eigenvalues of the matrix $\mathbf{M}^{-1}\mathbf{N}$. Then the iteration converges if and only if $\max_k |\alpha_k| < 1$. Consider the DPC algorithm in (9). In this case we have $\mathbf{M}^{-1}\mathbf{N} = \mathbf{H}$. Hence the dominant eigenvalue of $\mathbf{H}$, $\rho(\mathbf{H})$, should be less than one. By proposition 1, this ensures that the SIR requirements are satisfied for all users. Hence DPC converges to $\mathbf{p}^*$ whenever the SIR requirements can be satisfied.

The speed of the convergence is very important, since the link attenuations are changing all the time. For the iteration in (7), it can be shown that the smaller the $\rho(\mathbf{M}^{-1}\mathbf{N})$, the faster the convergence. Hence the task is to find $\mathbf{M}$ and $\mathbf{N}$ such that $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$ and as small as possible.

3.2. Convergence Using Standard Interference Functions

A different way of proving convergence of iterative algorithms was developed by Yates [7] and extended by Huang and Yates [8]. There the iteration is formulated by defining an *interference function* $I(\mathbf{p})$ such that

$$\mathbf{p}(k+1) = I(\mathbf{p}(k)) \tag{10}$$

The standard interference function framework gives a *sufficient* but not *necessary* condition for convergence of the iteration in (7). The following definition and proposition summarize this framework.

Definition 2 [7]**.** An interference function $I(\mathbf{p})$ is called "standard" if for all nonnegative power vectors

$$\begin{aligned} &I(\mathbf{p}) > 0 \\ &\mathbf{p} \geq \mathbf{p}' \Rightarrow I(\mathbf{p}) \geq I(\mathbf{p}') \\ &\forall \alpha > 1, \quad \alpha I(\mathbf{p}) > I(\alpha\mathbf{p}) \end{aligned} \tag{11}$$

Proposition 2 [7]**.** If the power control problem is feasible and $I(\mathbf{p})$ is a standard interference function, then for any initial nonnegative power vector $\mathbf{p}$ the iteration in (10) converges to the unique nonnegative fixed point $\mathbf{p}^*$.

3.3. Distributed Constrained Power Control

In all practical systems the transmitter powers are limited so that

$$\mathbf{0} \leq \mathbf{p} \leq \mathbf{p}_{\max} \tag{12}$$

where $\mathbf{0}$ is a vector with all-zero elements and $\mathbf{p}_{\max} = [p_i^{\max}, p_2^{\max}, \ldots, p_N^{\max}]^T$ denotes the maximum transmitter power of each transmitter. To take these limitations into account, the distributed constrained power control (DCPC) algorithm was suggested by Grandhi et al. [9]. It is defined by

$$p_i(k+1) = \min\left(p_i^{\max}, \frac{\gamma_i^t}{\gamma_i(k)} p_i(k) \right), \quad k = 0, 1, \ldots \tag{13}$$

where $p_i^{\max}$ is the maximum allowed transmitter power of transmitter i. With DCPC some transmitters can transmit with the maximum power, thus producing maximum interference to other users, but still not achieving their SIR target. Therefore it might be beneficial to lower the transmission power when link quality is poor. With this in mind, a following more general algorithm has been proposed (see Ref. 5 and the references cited therein) that has DCPC as a special case:

$$p_i(k+1) = \begin{cases} \dfrac{\gamma_i^t}{\gamma_i(k)} p_i(k) & \text{if, } \dfrac{\gamma_i^t}{\gamma_i(k)} p_i(k) \leq p_i^{\max} \\ p_i' & \text{if, } \dfrac{\gamma_i^t}{\gamma_i(k)} p_i(k) > p_i^{\max} \end{cases} \tag{14}$$

where $0 \leq p_i' \leq p_i^{\max}$. It can be shown that this algorithm converges to the optimal power vector $\mathbf{p}^*$ provided that the system in (6) has the optimal solution $\mathbf{p}^*$ in the power range given by (12).

3.4. A Two-User Example

Consider a system with only two mobile stations MS1 and MS2 and two base stations BS1 and BS2. Assume that

MS1 is connected to BS1 and MS2 is connected to BS2. In this case the power control problem is the following:

$$\begin{cases} \gamma_1 = \dfrac{g_{11}p_1}{g_{12}p_2 + \eta_1} \geq \gamma_1^t \\ \gamma_2 = \dfrac{g_{22}p_2}{g_{21}p_1 + \eta_2} \geq \gamma_2^t \end{cases} \Rightarrow \begin{cases} p_1 \geq \gamma_1^t \left(\dfrac{g_{12}}{g_{11}} p_2 + \dfrac{\eta_1}{g_{11}} \right) \\ p_2 \geq \gamma_2^t \left(\dfrac{g_{21}}{g_{22}} p_1 + \dfrac{\eta_2}{g_{22}} \right) \end{cases} \quad (15)$$

This situation is depicted in Fig. 3. The feasible region is shaded in the figure, and the optimal (minimum power) solution $\mathbf{p}^*$ is in the intersection of the two lines. Since $\mathbf{p}^*$ is within the maximum power limits, the problem is feasible. Consider that user 1 raises its target SIR γ_1^t while the link attenuations remain unchanged. It is then necessary for it to also raise its transmitter power as seen from (15). This, in turn, forces user 2 to raise its transmitter power. Thus it can happen that the optimal point $\mathbf{p}^*$ moves outside the maximum power limits, thereby making the problem unfeasible.

4. PRACTICAL ISSUES ON POWER CONTROL

A number of practical issues limit the implementation of the theoretical algorithms:

- *SIR Estimation*. The received SIR is not known exactly at a receiver, but it must be estimated, and thus there will always be some estimation error.
- *TPC Update Rate*. In CDMA one has to deal with the near–far situation, and thus the update rate of the power control algorithm must be sufficiently high that the variations in the link attenuation can be tracked. Typical update rates are from 800 Hz (used in the IS95 system [11]) to 1500 Hz (used in WCDMA [10]).
- *Feedback Information Accuracy*. The information of the SIR at the receiver should somehow be communicated to the transmitter. An accurate representation of the SIR measurement requires several bits, but this requires more signaling overhead. A usual case in practice is that only one bit is used to inform the transmitter to either increase or decrease its transmitter power by a fixed amount, typically by 1 dB (e.g., in IS95).
- *Loop Delay*. The loop delay greatly affects the performance of a power control algorithm. This delay comes from the SIR measurement process, the transmission of the SIR information over the radio channel, the processing of the SIR information to calculate and adjust the transmission power, and the propagation time after which the new transmission power affects the next SIR measurement. Therefore the power update is based on outdated information of the received SIR. If the channel variation is fast in comparison to the loop delay, the TPC algorithm cannot track the variations.
- *Errors in the Transmission of Feedback Information*. To minimize the loop delay, the TPC command bits are sent without error correction coding. Hence the probability of receiving an erroneous command can be relatively high (e.g., 5–10%).

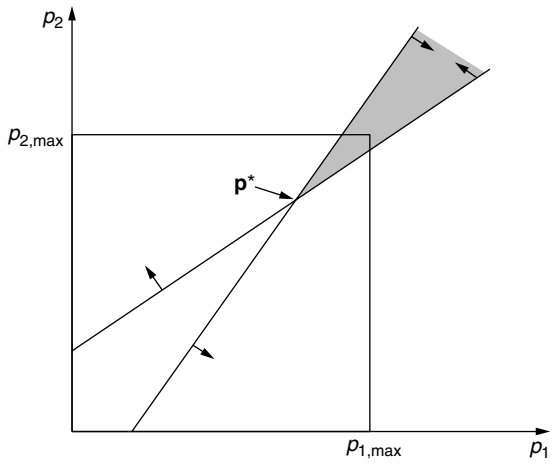


Figure 3. Power control problem for two-user case.

4.1. Open-Loop, Closed-Loop, and Outer-Loop Power Control

Because of the limitations mentioned above, TPC in CDMA is implemented in a slightly different way than the theoretical algorithms.

An intuitive way to compensate for the channel attenuation in the uplink would be to measure the strength of a pilot signal from the downlink, and adjust the transmitter power proportionally to the inverse of this measurement. Since the pilot signal is transmitted at constant power, the variation of its strength gives information of the downlink link attenuation. This is called *open-loop power control*. Unfortunately the center frequencies allocated to up- and down-link transmissions are usually widely separated, and thus the correlation between up- and down-link attenuations is generally weak. Therefore, the transmitter power update of a mobile must be based on feedback information of the received SIR at the base station, forming a closed loop between them. This *closed-loop power control* aims to keep the received uplink signal power level at a specified target. Moreover, the target must also be varied, because the SIR requirement for a given BER is not constant, but depends on the radio propagation conditions. This is the task of the *outer-loop power control*.

Open-loop TPC is generally used for initial power setting, when the two-way communication link is not yet established and closed loop is not possible. *Closed-loop TPC* aims to keep the received SIR at a target value. This is what is happening in the DPC algorithm given in Eq. (9). However, in practice only one bit is used to signal the received SIR information at a fast rate to track the channel variations. The transmitter is commanded to increase its power by a fixed step if the received SIR is below the target, and decrease otherwise. This kind of algorithm is used in IS95. In WCDMA there are some more degrees of freedom, for instance, the possibility to signal a "no change" command when the received SIR is reasonably close to the target, thus reducing the "bang-bang"-effect around the target.

The outer-loop control adjusts the SIR target so that a desired FER is guaranteed. A typical way to do this in practice is to raise the target by a larger step Δ_{up} when a frame is discarded, and to decrease the target with a smaller step Δ_{down} when a frame is correctly received. The relation between the step sizes gives the resulting average FER as $\text{FER} = \Delta_{\text{down}}/(\Delta_{\text{up}} - \Delta_{\text{down}})$ [10].

5. POWER CONTROL IN REAL-TIME VERSUS NON-REAL-TIME AND MULTIRATE SERVICES

The SIR balancing concept has the goal of maximizing the number of users in the system. If the users are real-time users requiring constant bit rates, this is a good strategy for maximizing capacity. However, as the cellular systems evolve to the next generation, the variety of services will be considerably different from those in the previous systems. In addition to the familiar real-time voice, there will be both real-time and non-real-time services with different data rates. Hence, maximizing the data throughput instead of the number of users might be more interesting from the operator's point of view. Of course the delay requirements must be fulfilled, as merely maximizing throughput would be achieved by just letting the user with the best instantaneous link quality to transmit.

Since CDMA is interference-limited, any decrease in the transmission power of one user is directly advantageous for other users because of decreased interference. If the link quality between a transmitter and a receiver is poor, high transmitter power is needed to satisfy the SIR requirements. This produces high interference to other receivers, and their serving transmitters must also increase their powers in order to cope with the increased interference.

In non-real-time services the data rate must only be satisfied on the average sense, and therefore the instantaneous data rate can be considerably varied. This allows the TPC algorithm even to cut off the transmission when the link quality is bad, and to transmit at a high data rate when the link quality is good. Thus, the situation as compared to conventional TPC designed for real-time services is reversed — the transmission power should be small when link quality is low, and vice versa. Since the time dimension can be utilized in the optimization, there is potential for significant capacity gain by minimizing the total transmitted *energy* instead of power. This can be accomplished by scheduling the data transmissions properly [12].

To elaborate, consider a set of users requiring individual data rates. Using (2) and (3), we write the effective data rate of user i as

$$r_i = \frac{W}{(E_b/N_0)_i}\gamma_i(\mathbf{p}) \tag{16}$$

where $\gamma_i(\mathbf{p})$ is the received SIR of user i with the power vector $\mathbf{p}$ and $(E_b/N_0)_i$ is the E_b/N_0 requirement for user i for achieving the data rate r_i. Let the maximum transmission power vector for the users be $\mathbf{p}_{\max} = (p_1^{\max}, p_2^{\max}, \ldots, p_N^{\max})$.

Definition 3 [5,12]**.** A rate vector $\mathbf{r}(\mathbf{p}_{\max}) = (r_1, r_2, \ldots, r_N)$ is instantaneously achievable if there exists a nonnegative power vector $\mathbf{p} \le \mathbf{p}_{\max}$ such that $r_i \le \dfrac{W}{(E_b/N_0)_i}\gamma_i(\mathbf{p})$ for all $1 \le i \le N$.

Definition 4 [5,12]**.** A rate vector $\mathbf{r}^*(\mathbf{p}_{\max}) = (r_1^*, r_2^*, \ldots, r_N^*)$ is achievable in the average sense if it may be expressed as $\mathbf{r}^* = \sum_k \lambda_k r_k$, where $\lambda_k \in [0, 1]$, $\sum_k \lambda_k = 1$, and all the r_k are instantaneously achievable rate vectors.

Thus, different rate vectors can be assigned a fraction of time (or frequency) yielding the required rate vector on the average. Assume that each link i requires a minimum data rate $r_i^{\min}$. Any excess data rate is potentially consumed, and thus paid for, by the user. It is the interest of the operators then to provide as much excess data rate as possible. For non-real-time services, therefore, the following optimization problem is of interest:

$$\begin{aligned} &\max \sum_{i=1}^{N} r_i^*(\mathbf{p}_{\max}) \\ &\text{subject to } r_i^*(\mathbf{p}_{\max}) \ge r_i^{\min}, \forall i \end{aligned} \tag{17}$$

6. POWER CONTROL AND OTHER RADIO RESOURCE MANAGEMENT (RRM)

Optimizing power control alone is not always the best way to enhance capacity. By understanding the relations between power control and other RRM functions, one can design more efficient algorithms by combining them in an ingenious way. For instance, base station assignment is closely related to the power control problem. For the SIR balancing problem, if the mobile stations could always be connected to the base station to which the link quality is optimal, the total transmission power would be minimized. Combined power and rate control is interesting for services with heterogeneous bit rates and quality requirements as discussed in the last section. Other methods that have been considered jointly with power control include smart antennas and beamforming, where there are more degrees of freedom in the optimization of the algorithms. An interested reader is directed to the articles in the Further Reading list for more details.

7. DISCUSSION AND VIEWS INTO THE FUTURE

The ultimate goal of radio resource management is to maximize the network capacity without unduly sacrificing the satisfaction of the users. Efficient transmitter power control is essential in CDMA cellular systems for achieving these goals. The efficiency of TPC depends on its ability to control the interference inherent in wireless multiuser systems. However, there are other methods for combating the multiple access interference in CDMA. One of these methods is *multiuser detection* (MUD). An optimal MUD-based receiver would theoretically eliminate the need for power control completely! In practice,

however, an optimal MUD receiver would be too complex, and suboptimal solutions must be used that are not completely resistant to interference, and TPC can still provide additional gain. Using only TPC provides a much cheaper way of controlling interference. This situation might change in the future, as the microtechnology evolves very rapidly and more efficient chips become available.

BIOGRAPHY

Matti J. Rintamäki received an M.S. degree in electrical engineering from Helsinki University of Technology, Finland, in 2000. Since then he has been a research scientist at Signal Processing Laboratory at HUT, where he has been working on power control algorithms for CDMA systems. His areas of interest are adaptive control and signal processing algorithms, and the design of radio resource management algorithms for wireless communications.

BIBLIOGRAPHY

1. S. Ariyavisitakul, Signal and interference statistics of a CDMA system with feedback power control — part II, *IEEE Trans. Commun.* **42**(2–4): 597–605 (Feb.–April 1994).
2. J. Zander, Performance of optimum transmitter power control in cellular radio systems, *IEEE Trans. Vehic. Technol.* **41**(1): 57–62 (Feb. 1992).
3. S. A. Grandhi, R. Vijayan, D. J. Goodman, and J. Zander, Centralized power control in cellular radio systems, *IEEE Trans. Vehic. Technol.* **42**(4): 466–468 (Nov. 1993).
4. M. Andersin, Z. Rosberg, and J. Zander, Gradual removals in cellular PCS with constrained power control and noise, *ACM/Baltzer Wireless Networks J.* **2**: 27–43 (1996).
5. J. Zander, S.-L. Kim, M. Almgren, and O. Queseth, *Radio Resource Management for Wireless Networks*, Artech House, Norwood, MA, 2001.
6. G. J. Foschini and Z. Miljanic, A simple distributed autonomous power control algorithm and its convergence, *IEEE Trans. Vehic. Technol.* **42**(4): 641–646 (Nov. 1993).
7. R. D. Yates, A framework for uplink power control in cellular radio systems, *IEEE J. Select. Areas Commun.* **13**(7): 1341–1347 (Sept. 1995).
8. C. Y. Huang and R. Yates, Rate of convergence for minimum power assignment in cellular radio systems, *ACM/Baltzer Wireless Networks J.* **1**: 223–231 (1998).
9. S. A. Grandhi, J. Zander, and R. Yates, Constrained power control, *Wireless Pers. Commun.* **1**: 257–270 (1995).
10. H. Holma and A. Toskala, *WCDMA for UMTS, Radio Access for Third Generation Mobile Communications*, Wiley, Chichester, UK, 2000.
11. TIA/EIA Interim Standard-95, *Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System*, Telecommunications Industry Assoc., 1993.
12. F. Berggren, S.-L. Kim, R. Jäntti, and J. Zander, Joint power control and intracell scheduling of DS-CDMA nonreal time data, *IEEE J. Select. Areas Commun.* **19**(10): 1860–1870 (Oct. 2001).

FURTHER READING

Bambos N., Toward power-sensitive network architectures in wireless communications: Concepts, issues, and design aspects, *IEEE Pers. Commun.* **5**(3): 50–59 (June 1998) (this article contains a nice review on power control).

Gilhousen K. S. et al., On the capacity of a cellular CDMA system, *IEEE Trans. Vehic. Technol.* **40**(2): 303–312 (May 1991) (an early paper on the capacity of CDMA as a multiple access technology; the need for power control is discussed).

Jäntti R. and S.-L. Kim, Second-order power control with asymptotically fast convergence, *IEEE J. Select. Areas Commun.* **18**(3): 447–457 (March 2000) (a distributed power control algorithm with faster convergence than with the DPC algorithm).

Kim D., On the convergence of fixed-step power control algorithms with binary feedback for mobile communication systems, *IEEE Trans. Commun.* **49**(2): 249–252 (Feb. 2001) (in this paper the author proves the convergence of fixed-step binary-feedback power control algorithms into a specific range of values).

Ulukus S. and R. D. Yates, Stochastic power control for cellular radio systems, *IEEE Trans. Commun.* **46**(6): 784–798 (June 1998) (in this paper the authors take the stochastic nature of the link attenuations and signal measurements into account and develop some power control algorithms whose convergence is proved stochastically).

Viterbi A. J., *Principles of Spread spectrum communication*, Reading, MA: Addison-Wesley, 1995 (a comprehensive book on spread-spectrum technology for commercial wireless applications).

Yener A., R. D. Yates, and S. Ulukus, Interference management for CDMA systems through power control, multiuser detection, and beamforming, *IEEE Trans. Commun.* **49**(7): 1227–1239 (July 2001) (in this paper the authors combine intelligent techniques to achieve higher performance than using the techniques alone).

POWER SPECTRA OF DIGITALLY MODULATED SIGNALS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

In the design of digital communication systems for transmitting digital information through a channel, the spectral characteristics of the information-bearing signals is an important element. The communication system designer must ensure that the signals used for transmitting the information do not violate the bandwidth constraint imposed by the channel or some governmental agency. In this article, the spectral characteristics of several types of digitally modulated signals are described.

In describing the spectral characteristics of digitally modulated signals, it is desirable to classify such signals into two different categories, namely, linearly modulated signals and nonlinearly modulated signals. The class of linearly modulated signals include pulse amplitude modulation or amplitude shift keying (PAM or ASK),

phase shift keying (PSK), and quadrature amplitude modulation (QAM). The class of nonlinearly modulated signals includes continuous-phase modulation (CPM) and the special form of CPM called *continuous-phase frequency shift keying* (CPFSK). CPM and CPFSK are constant-amplitude signals; hence, they are well suited for radiocommunications, where the transmitted signals can be amplified by power amplifiers that can be driven into saturation without introducing nonlinear distortion in the signals.

The spectral characteristics of linearly modulated signals are considered in the next section. Nonlinearly modulated signals are treated in the subsequent section.

2. POWER SPECTRA OF LINEARLY MODULATED SIGNALS

Linear digital modulation methods that include PAM (ASK), PSK, and QAM can be treated in a unified manner. The digitally modulated signals for these methods are usually generated as lowpass signals, which may be expressed mathematically as

$$v(t) = \sum_n I_n g(t - nT) \tag{1}$$

and then translated to bandpass, for transmission over the bandpass channel, by a frequency translation. Hence, the bandpass signal is

$$s(t) = \mathrm{Re}[v(t)e^{j2\pi f_c t}] \tag{2}$$

where f_c is the carrier frequency. In the expression for the lowpass signal given by Eq. (1), the data sequence $\{I_n\}$ has values taken from either a PAM, PSK, or QAM signal-point constellation. In particular, if the modulated signal is M-level PAM, the sequence $\{I_n\}$ takes on values from the set $\{\pm 1, \pm 3, \pm 5, \ldots, \pm(M-1)\}$. When the modulated signal is M-phase PSK, the data sequence $\{I_n\}$ takes values from the set $\{\exp(2\pi m/M), m = 0, 1, \ldots, M-1\}$. A QAM signal is basically a combined amplitude/phase-modulated signal, so the data sequence takes an values from the set $\{A_m e^{j\theta_m}, m = 0, 1, \ldots, M-1\}$. Finally, the signal $g(t)$ in Eq. (1) is a signal pulse that is used to shape the spectrum of the transmitted signal. The rate at which data symbols are transmitted is $1/T$, where T defines the symbol interval.

The digitally modulated signal $v(t)$ is a random process because the data sequence $\{I_n\}$ is random. The power spectrum of $v(t)$ will be determined by first obtaining the autocorrelation function of $v(t)$ and then computing its Fourier transform. The power spectrum of the bandpass signal $s(t)$ is simply obtained from the power spectrum of $v(t)$ by a simple frequency translation. Thus, the autocorrelation function of the bandpass signal $s(t)$, denoted as $\phi_{ss}(\tau)$ is related to the autocorrelation function of the equivalent lowpass signal $v(t)$ through the expression

$$\phi_{ss}(\tau) = \mathrm{Re}[\phi_{vv}(\tau)e^{j2\pi f_c \tau}] \tag{3}$$

where $\phi_{vv}(\tau)$ is the autocorrelation function of the equivalent lowpass signal $v(t)$. The Fourier transform of Eq. (3) yields the desired expression for the power density spectrum $\Phi_{ss}(f)$ in the form

$$\Phi_{ss}(f) = \tfrac{1}{2}[\Phi_{vv}(f - f_c) + \Phi_{vv}(-f - f_c)] \tag{4}$$

where $\Phi_{vv}(f)$ is the power density spectrum of $v(t)$. It suffices to determine the autocorrelation function and the power density spectrum of the equivalent lowpass signal $v(t)$.

The autocorrelation function of $v(t)$ is defined as

$$\begin{aligned}\phi_{vv}(t + \tau; t) &= \frac{1}{2}E[v^*(t)v(t + \tau)] \\ &= \frac{1}{2}\sum_{n=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} E[I_n^* I_m]g^*(t - nT)g(t + \tau - mT)\end{aligned} \tag{5}$$

It is assumed that the sequence of information symbols $\{I_n\}$ is wide-sense stationary with mean μ_i and autocorrelation function

$$\phi_{ii}(m) = \tfrac{1}{2}E[I_n^* I_{n+m}] \tag{6}$$

Hence Eq. (5) can be expressed as

$$\begin{aligned}\phi_{vv}(t + \tau; t) &= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\phi_{ii}(m - n)g^* \\ &\quad \times (t - nT)g(t + \tau - mT) \\ &= \sum_{m=-\infty}^{\infty}\phi_{ii}(m)\sum_{n=-\infty}^{\infty} g^*(t - nT)g \\ &\quad \times (t + \tau - nT - mT)\end{aligned} \tag{7}$$

The second summation in Eq. (7), namely

$$\sum_{n=-\infty}^{\infty} g^*(t - nT)g(t + \tau - nT - mT) \tag{8}$$

is periodic in the t variable with period T. Consequently, $\phi_{vv}(t + \tau; t)$ is also periodic in the t variable with period T:

$$\phi_{vv}(t + T + \tau; t + T) = \phi_{vv}(t + \tau; t) \tag{9}$$

In addition, the mean value of $v(t)$, which is

$$E[v(t)] = \mu_i \sum_{n=-\infty}^{\infty} g(t - nT) \tag{10}$$

is periodic with period T. Therefore, $v(t)$ is a random process having a periodic mean and autocorrelation function. Such a process is called a *cyclostationary process* or a *periodically stationary process in the wide sense*.

In order to compute the power density spectrum of a cyclostationary process, the dependence of $\phi_{vv}(t + \tau; t)$ on the t variable must be eliminated. This can be

accomplished simply by averaging $\phi_{vv}(t+\tau;t)$ over a single period:

$$\begin{aligned}\overline{\phi}_{vv}(\tau) &= \frac{1}{T}\int_{-T/2}^{T/2}\phi_{vv}(1+\tau;t)\,dt \\ &= \sum_{m=-\infty}^{\infty}\phi_{ii}(m)\sum_{n=-\infty}^{\infty}\frac{1}{T}\int_{-T/2}^{T/2}g^*(t-nT)g \\ &\quad\times(t+\tau-nT-mT)\,dt \\ &= \sum_{m=-\infty}^{\infty}\phi_{ii}(m)\sum_{n=-\infty}^{\infty}\frac{1}{T}\int_{-T/2-nT}^{T/2-nT}g^*(t)g(t+\tau-mT)\,dt\end{aligned} \tag{11}$$

We interpret the integral in this equation as the time autocorrelation function of $g(t)$ and define it as

$$\phi_{gg}(t) = \int_{-\infty}^{\infty}g^*(t)g(t+\tau)\,dt \tag{12}$$

Consequently Eq. (11) can be expressed as

$$\overline{\phi}_{vv}(\tau) = \frac{1}{T}\sum_{m=-\infty}^{\infty}\phi_{ii}(m)\phi_{gg}(\tau-mT) \tag{13}$$

The Fourier transform of the relation in Eq. (13) yields the (average) power density spectrum of $v(t)$ in the form

$$\Phi_{vv}(f) = \frac{1}{T}|G(f)|^2\Phi_{ii}(f) \tag{14}$$

where $G(f)$ is the Fourier transform of $g(t)$, and $\Phi_{ii}(f)$ denotes the power density spectrum of the information sequence, defined as

$$\Phi_{ii}(f) = \sum_{m=-\infty}^{\infty}\phi_{ii}(m)e^{-j2\pi fmT} \tag{15}$$

The result in Eq. (14) illustrates the dependence of the power density spectrum of $v(t)$ on the spectral characteristics of the pulse $g(t)$ and the information sequence $\{I_n\}$. Thus, the spectral characteristics of $v(t)$ can be controlled by design of the pulseshape $g(t)$ and by design of the correlation characteristics of the information sequence.

Whereas the dependence of $\Phi_{vv}(f)$ on $G(f)$ is easily understood on observation of Eq. (14), the effect of the correlation properties of the information sequence is more subtle. First, we note that for an arbitrary autocorrelation $\phi_{ii}(m)$ the corresponding power density spectrum $\Phi_{ii}(f)$ is periodic in frequency with period $1/T$. In fact, the equation [Eq. (15)] relating the spectrum $\Phi_{ii}(f)$ to the autocorrelation $\phi_{ii}(m)$ is in the form of an exponential Fourier series with the $\{\phi_{ii}(m)\}$ as the Fourier coefficients. As a consequence, the autocorrelation sequence $\phi_{ii}(m)$ is given by

$$\phi_{ii}(m) = T\int_{-1/2T}^{1/2T}\Phi_{ii}(f)e^{j2\pi fmT}\,df \tag{16}$$

Second, let us consider the case in which the information symbols in the sequence are real and mutually uncorrelated. In this case, the autocorrelation function $\phi_{ii}(m)$ can be expressed as

$$\phi_{ii}(m) = \begin{cases}\sigma_i^2+\mu_i^2 & (m=0)\\ \mu_i^2 & (m+0)\end{cases} \tag{17}$$

where σ_i^2 denotes the variance of an information symbol. When Eq. (17) is used to substitute for $\phi_{ii}(m)$ in Eq. (15), we obtain

$$\Phi_{ii}(f) = \sigma_i^2+\mu_i^2\sum_{m=-\infty}^{\infty}e^{-j2\pi fmT} \tag{18}$$

The summation in Eq. (18) is periodic with period $1/T$. It may be viewed as the exponential Fourier series of a periodic train of impulses with each impulse having an area $1/T$. Therefore Eq. (18) can also be expressed in the form

$$\Phi_{ii}(f) = \sigma_i^2+\frac{\mu_1^2}{T}\sum_{m=-\infty}^{\infty}\delta\left(f-\frac{m}{T}\right) \tag{19}$$

Substitution of Eq. (19) into Eq. (14) yields the desired result for the power density spectrum of $v(t)$ when the sequence of information symbols is uncorrelated:

$$\Phi_{vv}(f) = \frac{\sigma_i^2}{T}|G(f)|^2+\frac{\mu_1^2}{T^2}\sum_{m=-\infty}^{\infty}\left|G\left(\frac{m}{T}\right)\right|^2\delta\left(f-\frac{m}{T}\right) \tag{20}$$

The expression in Eq. (20) for the power density spectrum is purposely separated into two terms to emphasize the two different types of spectral components. The first term is the continuous spectrum, and its shape depends only on the spectral characteristic of the signal pulse $g(t)$. The second term consists of discrete frequency components spaced $1/T$ apart in frequency. Each spectral line has a power that is proportional to $|G(f)|^2$ evaluated at $f=m/T$. Note that the discrete frequency components vanish when the information symbols have zero mean: $\mu_i=0$. This condition is usually desirable for the digital modulation techniques under consideration, and it is satisfied when the information symbols are equally likely and symmetrically positioned in the complex plane. Thus, the system designer can control the spectral characteristics of the digitally modulated signal by proper selection of the characteristics of the information sequence to be transmitted.

As an example that illustrates the spectral shaping from $g(t)$, consider the rectangular pulse shown in Fig. 1. The Fourier transform of $g(t)$ is

$$G(f) = AT\frac{\sin\pi fT}{\pi fT}e^{-j\pi fT}$$

Hence

$$|G(f)|^2 = (AT)^2\left(\frac{\sin\pi fT}{\pi fT}\right)^2 \tag{21}$$

This spectrum is illustrated in Fig. 1. Note that it contains zeros at multiples of $1/T$ in frequency and that it decays inversely as the square of the frequency variable. As a consequence of the spectral zeros in $G(f)$, all except one of

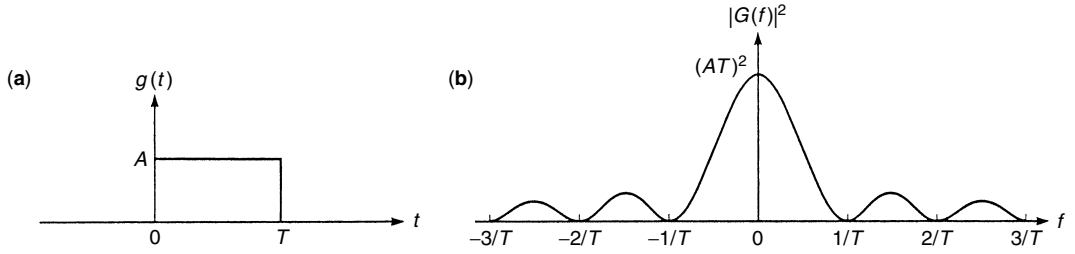


Figure 1. Rectangular pulse and its energy density spectrum.

the discrete spectral components in Eq. (20) vanish. Thus, on substitution for $|G(f)|^2$ from Eq. (21), Eq (20) reduces to

$$\Phi_{vv}(f) = \sigma_i^2 A^2 T \left(\frac{\sin \pi fT}{\pi fT}\right)^2 + A^2 \mu_i^2 \delta(f)$$

To illustrate that spectral shaping can also be accomplished by operations performed on the input information sequence, we consider a binary sequence $\{b_n\}$ from which we form the symbols

$$I_n = b_n + b_{n-1}$$

The $\{b_n\}$ are assumed to be uncorrelated random variables, each having zero mean and unit variance. Then the autocorrelation function of the sequence $\{I_n\}$ is

$$\begin{aligned}\phi_{ii}(m) &= E(I_n I_{n+m}) \\ &= \begin{cases} 2 & (m = 0) \\ 1 & (m = \pm 1) \\ 0 & (\text{otherwise}) \end{cases}\end{aligned}$$

Hence, the power density spectrum of the input sequence is

$$\begin{aligned}\Phi_{ii}(f) &= 2(1 + \cos 2\pi fT) \\ &= 4 \cos^2 \pi fT\end{aligned}$$

and the corresponding power density spectrum for the (lowpass) modulated signal is

$$\Phi_{vv}(f) = \frac{4}{T}|G(fg)|^2 \cos^2 \pi fT$$

Since $\cos^2 \pi fT$ has its first null at $f = 1/2T$, the effect of multiplying $|G(f)|^2$ by $\cos^2 \pi fT$ is to narrow the width of the mainlobe of the signal spectrum.

2. POWER SPECTRA OF CONTINUOUS-PHASE MODULATED SIGNALS

A CPM signal is described mathematically as

$$s(t) = A \cos[2\pi f_c t + \phi(t; \mathbf{I})] \tag{22}$$

where A is the signal amplitude, f_c is the carrier frequency, and $\phi(t; \mathbf{I})$ is the phase of the signal that carries the information. The phase function may be defined as

$$\phi(t; \mathbf{I}) = 2\pi h \sum_{k=-\infty}^{n} I_k q(t - kT), \quad nT \le t \le (n+1)T \tag{23}$$

where $\{I_k\}$ is the data sequence selected from the M-level amplitude alphabet $\{\pm 1, \pm 2, \ldots, \pm(M-1)\}$, h is the modulation index, and $q(t)$ is defined as the integral of a pulse $g(t)$:

$$q(t) = \int_0^t g(\tau)\, d\tau \tag{24}$$

The pulse $g(t) = 0$ for $t < 0$ and $t > LT$, where L is an integer and T is the symbol interval. When $L = 1$, the pulse $g(t)$ is nonzero over a single signal interval, and the CPM signal is called *full-response CPM*. When $L \ge 2$, the pulse $g(t)$ is nonzero over two or more signal intervals and the CPM signal is called partial response. Furthermore $g(t)$ is normalized in area so that

$$q(LT) = \int_0^{LT} g(\tau)\, d\tau = \frac{1}{2}$$

The phase continuous signal may be viewed as having been generated as an FM signal. Suppose that the lowpass data signal $d(t)$ is a PAM signal of the form

$$d(t) = \sum_k I_k g(t - kT) \tag{25}$$

where $\{I_m\}$ is the data sequence of symbols selected from the alphabet $\{\pm 1, \pm 3, \ldots, \pm(M-1)\}$ and $g(t)$ is a pulse with unit area that is nonzero over the interval $0 \le t \le LT$. If $d(t)$ is used to frequency modulate the carrier f_c, then, the phase of the carrier is

$$\begin{aligned}\phi(t; \mathbf{I}) &= 4\pi f_d T \int_{-\infty}^{t} d(\tau)\, d\tau \\ &= 2\pi h \sum_{k=-\infty}^{n} I_k q(t - nT), \quad nT \le t \le (n+1)T\end{aligned} \tag{26}$$

where f_d is the peak frequency deviation and the modulation index is defined as $h = 2f_d T$.

Some examples of the pulseshapes $g(t)$ and $q(t)$ are illustrated in Fig. 2. When $L = 1$ and $g(t)$ is a rectangular pulse, as shown in Fig. 2a, the CPM signal reduces to the special case called *continuous-phase frequency shift keying* (CPFSK). Furthermore, if the modulation index is selected as $h = \frac{1}{2}$, the CPFSK signal is called minimum-shift keying (MSK). The signal pulse $g(t)$ shown in Fig. 2b is called a *raised-cosine pulse*, and the resulting CPM is a full-response CPM signal. On the other hand, the signal pulses $g(t)$ shown in Fig. 2c,d extend over two signal intervals, and hence the CPM signals are partial response.

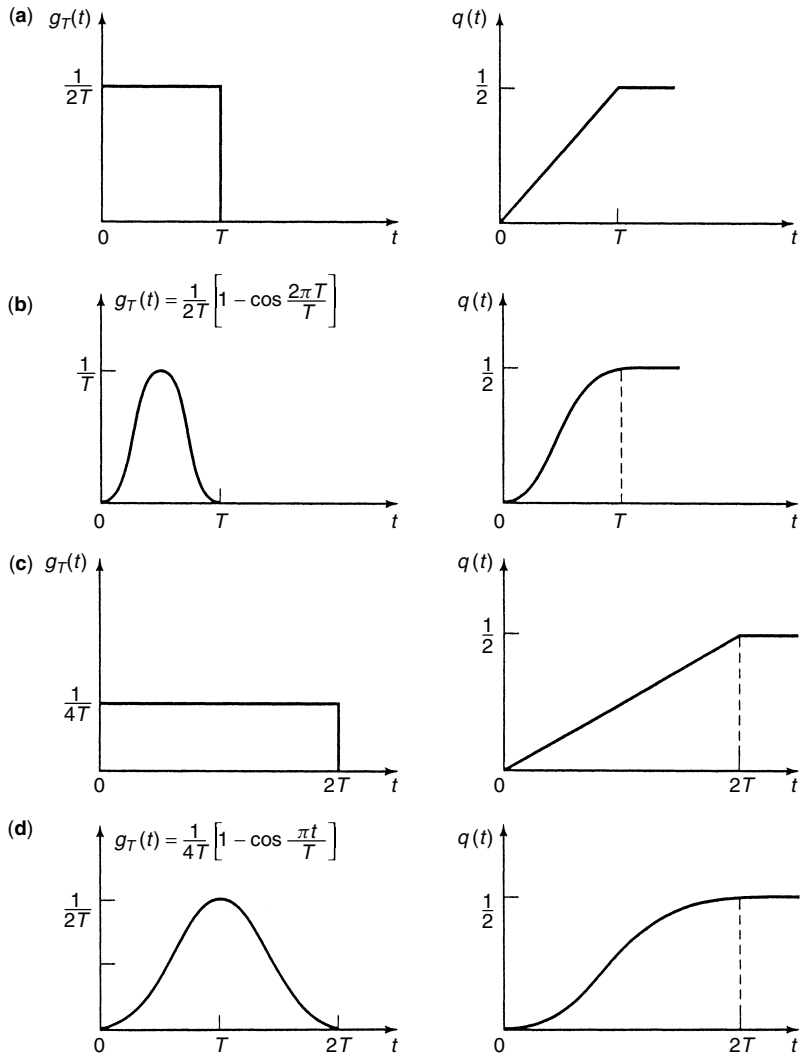


Figure 2. Pulseshapes for full-response (**a**,**b**) and partial-response (**c**,**d**) CPM.

To determine the power spectrum of the CPM signal, the general procedure described in the preceding section may be used. One begins with the equivalent lowpass signal

$$v(t) = e^{j\phi(t;\mathbf{I})} \tag{27}$$

where $\phi(t;\mathbf{I})$ is given by Eq. (26) with $n = \infty$. The autocorrelation function of $v(t)$ is easily shown to be

$$\phi_{vv}(t+t;t) = \frac{1}{2}E\left[\exp\left(j2\pi h \sum_{k=-\infty}^{\infty} I_k[q(t+\tau-kT) - q(t-kT)]\right)\right] \tag{28}$$

The sum of exponents may also be expressed as a product of exponents:

$$\phi_{vv}(t+\tau;t) = \frac{1}{2}E\left(\prod_{k=-\infty}^{\infty} \exp\{j2\pi h I_k[q(t+\tau-kT) - q(t-kT)]\}\right) \tag{29}$$

Although Eq. (29) implies that there are an infinite number of factors in the product, the pulse $g(t) = 0$ for $t < 0$ and $t > LT$, and $q(t) = 0$ for $t < 0$. Consequently only a finite number of terms in the product have nonzero exponents. The next step is to perform the expectation over the data symbols $\{I_k\}$ and then to average the periodic autocorrelation function over the period $(0, T)$:

$$\overline{\phi}_{vv}(\tau) = \frac{1}{T}\int_0^T \phi_{vv}(t+\tau;t)\,dt \tag{30}$$

The final step is to compute the Fourier transform of $\overline{\phi}_{vv}(t)$ to obtain the power spectrum.

Unfortunately, these computations do not yield a closed-form solution except in special cases. When the information symbols are equally likely, the average autocorrelation function is given as

$$\overline{\phi}_{vv}(\tau) = \frac{1}{2T}\int_0^T \prod_{k=1-L}^{[t/T]} \frac{1}{M} \times \frac{\sin 2\pi h M[q(t+\tau-kT) - q(t-kT)]}{\sin 2\pi h[q(t+\tau-kT) - q(t-kT)]}\,dt \tag{31}$$

and the corresponding expression for the power density spectrum is

$$\Phi_{vv}(f) = 2\left[\int_0^{LT} \overline{\phi}_{vv}(\tau)\cos 2p\pi f\tau\, d\tau + \frac{1-\psi(jh)\cos 2\pi fT}{1+\psi^2(jh)-2\psi(jh)\cos 2\pi fT} \times \int_{LT}^{(L+1)T} \overline{\phi}_{vv}(\tau)t\cos 2\pi f\tau\, d\tau - \frac{\psi(jh)\sin 2\pi fT}{1+\psi^2(jh)-2\psi(jh)\cos 2\pi fT} \times \int_{LT}^{(L+1)T} \overline{\phi}_{vv}(\tau)\sin 2\pi f\tau\, d\tau\right] \tag{32}$$

where $\psi(jh)$ is the characteristic function of the information symbols $\{I_k\}$, which is given as

$$\begin{aligned}\psi(jh) &= \sum_{\substack{n=-(M-1)\\ n \text{ odd}}}^{M-1} p_n e^{j\pi hn} \\ &= \frac{1}{M}\sum_{\substack{n=-(M-1)\\ n \text{ odd}}}^{M-1} e^{j\pi hn} \\ &= \frac{1}{M}\frac{\sin M\pi h}{\sin \pi h}\end{aligned} \tag{33}$$

where $p_n = 1/M$ is the probability of each of the M levels. The expression for the power density spectrum given in Eq. 32 must be evaluated numerically.

2.1. Power Density Spectrum of CPFSK

A closed-form expression for the power density spectrum can be obtained from Eq. (32) when the pulseshape $g(t)$ is rectangular and zero outside the interval $[0, T]$. In this case, $q(t)$ is linear for $0 \le 1 \le T$. The resulting power spectrum may be expressed as

$$\Phi_{vv}(f) = T\left[\frac{1}{M}\sum_{n=1}^{M} A_n^2(f) + \frac{2}{M^2}\sum_{n=1}^{M}\sum_{m=1}^{M} B_{nm}(f)A_n(f)A_m(f)\right] \tag{34}$$

where

$$\begin{aligned}A_n(f) &= \frac{\sin \pi[fT - \frac{1}{2}(2n-1-M)h]}{\pi[fT - \frac{1}{2}(2n-1-M)h]} \\ B_{nm}(f) &= \frac{\cos(2\pi fT - \alpha_{nm}) - \psi\cos\alpha_{nm}}{1+\psi^2-2\psi\cos 2\pi fT} \\ \alpha_{nm} &= \pi h(m+n-1-M) \\ \psi \equiv \psi(jh) &= \frac{\sin M\pi h}{M\sin \pi h}\end{aligned} \tag{35}$$

The power density spectrum of CPFSK for $M = 2, 4$ is plotted in Figs. 3 and 4 as a function of the normalized frequency fT, with the modulation index $h = 2f_dT$ as a parameter. Note that only one-half of the bandwidth occupancy is shown in these graphs. The origin corresponds to the carrier f_c. The graphs illustrate that the spectrum of CPFSK is relatively smooth and well confined for $h < 1$. As h approaches unity, the spectra become very peaked and, for $h = 1$ when $|\psi| = 1$, we find that impulses occur at M frequencies. When $h > 1$, the spectrum becomes much broader. In communication systems where CPFSK is used, the modulation index is designed to conserve bandwidth, so that h is selected to be less than unity.

The special case of binary CPFSK with $h = \frac{1}{2}$ (or $f_d = 1/4T$) and $\psi = 0$ corresponds to MSK. In this case, the spectrum of the signal is

$$\Phi_{vv}(f) = \frac{16T}{\pi^2}\left(\frac{\cos 2\pi fT}{1-16f^2T^2}\right)^2 \tag{36}$$

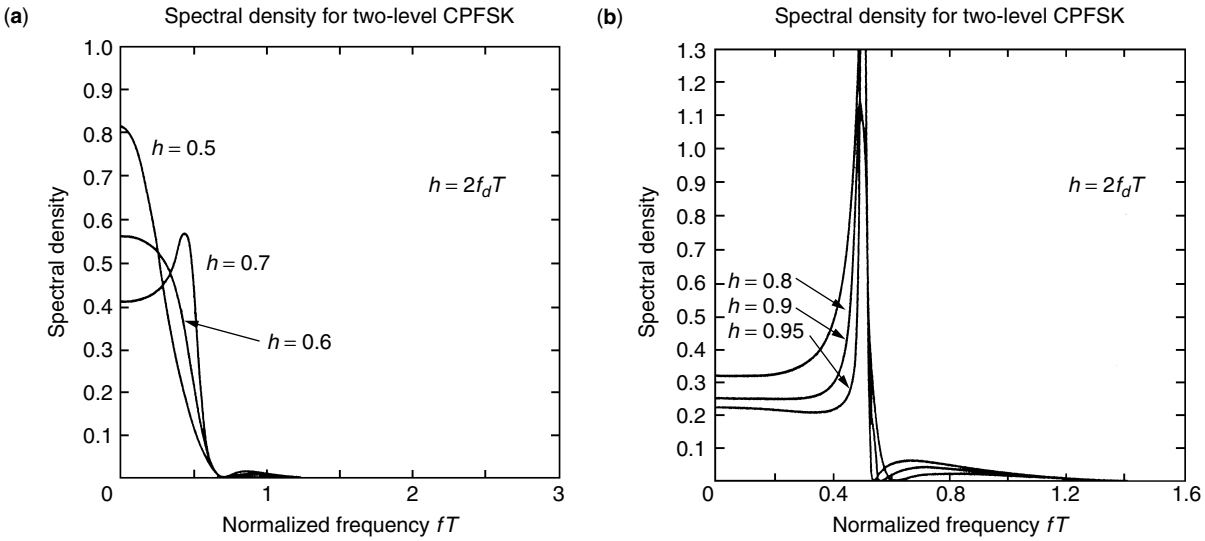


Figure 3. Power spectra for binary CPFSK.

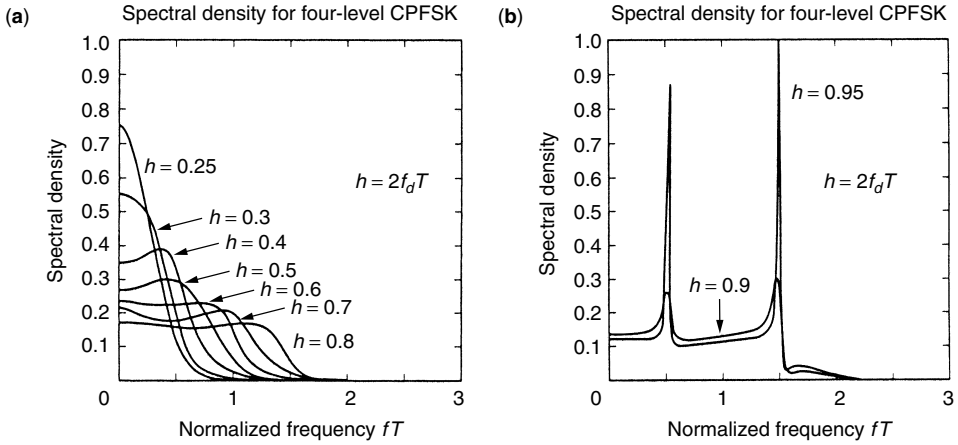


Figure 4. Power spectra for quaternary CPFSK.

2.2. Power Density Spectrum of CPM

The use of smooth pulses such as raised-cosine pulses of the form

$$g(t) = \begin{cases} \dfrac{1}{2LT}\left(1 - \cos\dfrac{2\pi t}{LT}\right) & (0 \le t \le LT) \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

where $L = 1$ for full response and $L > 1$ for partial response, result in smaller bandwidth occupancy and, hence, greater bandwidth efficiency than the use of rectangular pulses. For example, Fig. 5 illustrates the power density spectrum for binary CPM with different partial-response raised-cosine (LRC) pulses when $h = \frac{1}{2}$. For comparison, the spectrum of (MSK) binary CPFSK is also shown. Note that as L increases the pulse, $g(t)$ becomes smoother and the corresponding spectral occupancy of the signal is reduced.

The effect of varying the modulation index in a CPM signal is illustrated in Fig. 6 for the case of $M = 4$ and a raised-cosine pulse of the form given in Eq. (37) with $L = 3$. Note that these spectral characteristics are similar to the ones illustrated previously for CPFSK, except that these spectra are narrower because of the use of a smoother pulseshape.

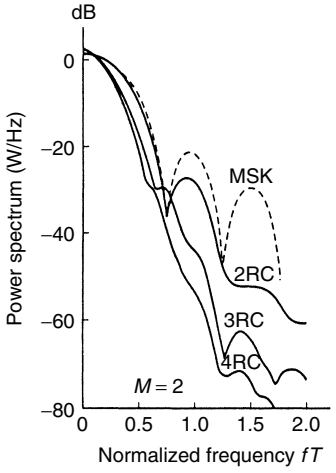


Figure 5. Power density spectrum for binary CPM with $h = \frac{1}{2}$ and different pulseshapes. [From Aulin et al. (1981); © 1981 IEEE.

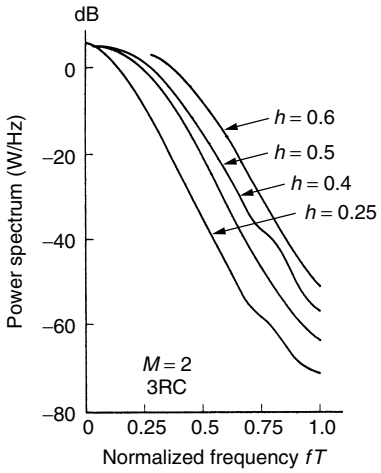


Figure 6. Power density spectrum for $M = 4$ CPM with $3RC$ and different modulation indices. [From Aulin et al.(1981); © 1981 IEEE.]

3. CONCLUDING REMARKS

PAM, PSK, QAM, and CPM signals are described in greater detail in other articles of this encyclopedia as well as in Ref. 1. Additional numerical results on the spectral characteristics of CPM signals can be found in Refs. 2 and 3.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College

of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
2. J. B. Anderson, T. Aulin, and C. W. Sundberg, *Digital Phase Modulation*, Plenum, New York, 1986.
3. R. R. Anderson and J. Salz, Spectra of digital FM, *Bell Syst. Tech. J.* **44**: 1165–1189 (July–Aug. 1965).

POWERLINE COMMUNICATIONS

Halid Hrasnica
Abdelfatteh Haidine
Ralf Lehnert
Dresden University of Technology
Dresden, Germany

1. INTRODUCTION

Powerline communications (PLC) uses the standard electrical power distribution network in parallel with the energy distribution for the transmission of digital data. This avoids the installation of a new telecommunications infrastructure, where a power grid already exists. PLC with low data rates of ~1 kbps (kilobit per second) has been in use since 1990 or 1991. This narrowband technology has been standardized in Europe by CENELEC EN 50065 [1]. It operates in the frequency band from 3 to 148.5 kHz and uses simple digital modulation schemes, such as frequency shift keying (FSK).

With the advent of high-speed digital signal processors, advanced digital modulation techniques can be implemented. As an example, OFDM not only allows a high data rate but is also able to cope with a time-varying transmission channel. This presently allows transmission data rates over the powerline of $\leq$4 Mbps. PLC is therefore able to compete with, For example, digital subscribers lines (DSLS) on the telephone two-wire copper line or with wireless LAN techniques in the home area.

PLC offers an advantage for developing countries, because the power distribution network is already in place and a second network need not be built. In developed countries PLC also opens an opportunity for prospective network operators, who plan to compete with the incumbent operator, the earlier monopolist. Here again, an existing infrastructure can be used, and therefore the often prohibitive costs of new cabling can be avoided.

Within the power grid, PLC technology may be used in the backbone links in the high-voltage area, in the medium-voltage plane in urban areas, in the low-voltage plane for the access to the customer's premises, and, finally, within a household to reach the subscriber's communications terminal. In modern backbone's underground cables, there is usually an integrated fiber offering a nearly unlimited capacity, much higher than current PLC technology. Also the traffic of a large number of subscribers cannot be transported by current PLC technology. The medium voltage feeder network may be a candidate for PLC, but the bandwidth needed for a large number of high-speed customers cannot be supported by state-of-the-art technology. Also many medium- to low-voltage transformers are now reached by a fiber.

PLC has a promising application area and business case in the low-voltage area of the power grid. Here the number of customers per distribution section is limited to approximately 10–50, such that today's gross bandwidth still allows sufficient bandwidth per subscriber. This access network is usually limited in its size, the so-called last mile. In the case of in-home PLC, the size is even further limited (<100 m) and also the number of terminals may correspond to the number of persons in a home.

PLC operates as a shared medium. This means that the total capacity is shared in a statistical manner by all users on the same distribution section. Therefore a MAC protocol is needed to coordinate the sharing of the bandwidth fairly (see Section 5.3).

A major challenge for PLC is electromagnetic compatibility (EMC) (see Section 4.4). This is a bidirectional problem. On one hand, PLC is disturbed by electric noise (spikes, etc.; see Section 4.4). PLC systems also generate

radiation that disturbs other wireless communications. For this reason, countries have issued radiation limits, which constitute the main reason for capacity limits in PLC. In Germany, for example, these limits are defined by the regulatory body RegTP in regulation NB30 [2]. Currently under discussion is the replacement of the actual "chimney" regulation of forbidden frequency bands by a limiting curve within the entire frequency range of 1–30 MHz.

2. APPLICATION OF PLC TECHNOLOGY

2.1. Overview

The application of electrical supply networks in telecommunications has been known since the beginning of the twentieth century. The first *carrier frequency systems* (CFSs) have been operated in high-voltage electrical networks that were able to span distances over 500 km using 10 W signal power [3]. Such systems have been used for internal communication of electrical utilities and realization of remote measurements and control tasks. Also, communication over medium- and low-voltage electrical networks has been implemented by ripple carrier signaling (RCS) systems for realization of load management in electrical supply systems. Internal electrical networks have been used mostly for realization of various automation services within buildings or private houses.

The electrical supply systems consist of three network levels that can be used as a transmission medium for the realization of PLC networks (Fig. 1):

- *High-voltage networks* (110–380 kV) usually connect the power stations with supply regions or very big customers. They span very long distances, allowing the power exchange within a continent. Electrical supply grids of high-voltage networks are realized mostly as overhead lines.
- *Medium-voltage networks* (10–30 kV) supply large regions, cities, and big industrial or commercial complexes. The spanned distances are significantly shorter than those with high-voltage networks. These networks are realized by overhead lines or by underground cables.
- *Low-voltage networks* (230/400 V; in the United States, 110 V) directly supply the end users that are connected either as individual customers or single users belonging to a big customer (e.g., within a business building). Low-voltage networks are usually realized by overhead lines in rural areas and by underground cables within cities. The cable length between the last transformer unit and the customers is varying, but normally not longer than a few hundred meters.

In-home electrical installations belong to the low-voltage network level (Fig. 1). However, internal installations are usually owned by the users. They are connected to the supply network over an electrical power meter unit (M) (see Fig. 2). On the other hand, the rest of the low-voltage network (outdoor) belongs to the electrical supply utilities.

2.2. Narrowband PLC

As a successor to the ANSI X-10 standard, CENELEC has standardized PLC in the frequency range 3–148.5 kHz. This allows mainly private home or business building automation applications with data rates up to 1200 bps. This is the PLC variant of the European installation bus, named Powernet EIB. It became a de facto industry standard with FSK modems and a simple MAC protocol.

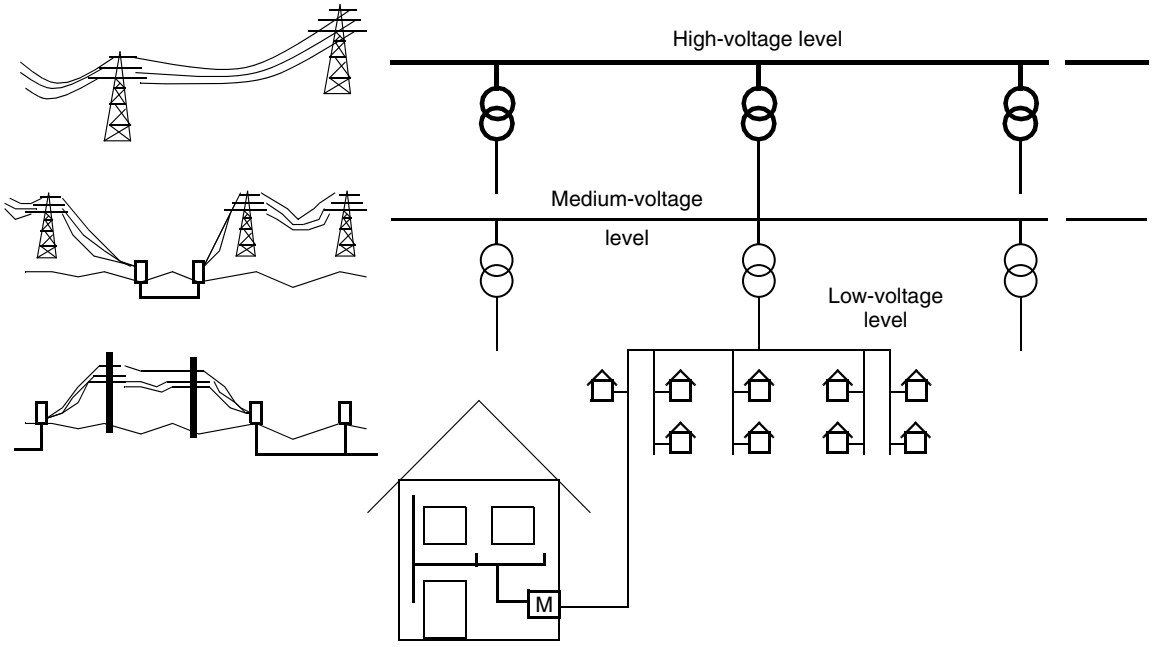


Figure 1. Structure of electrical supply networks.

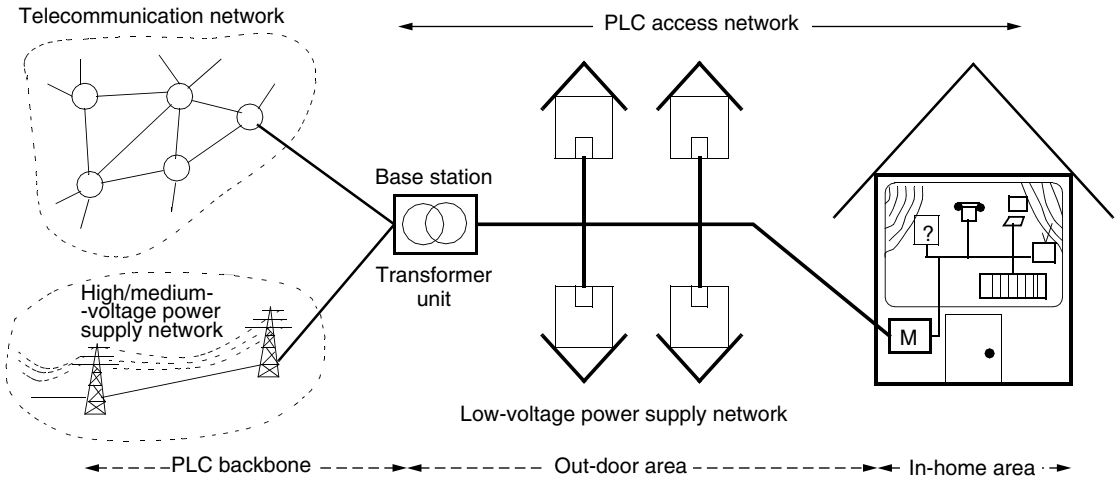


Figure 2. Structure of PLC access networks.

Loss of data may occur especially under high load. This is circumvented by ARQ mechanisms at the higher layers.

2.3. Broadband PLC

Broadband PLC systems provide significantly higher data rates (>2 Mbps) than do narrowband PLC systems. Where the narrowband networks can realize only a small number of voice channels and data transmission with very low bit rates, broadband PLC networks offer realization of more sophisticated telecommunication services: multiple voice connections, high-speed data transmission, transfer of video signals, and narrowband services as well. Therefore, PLC broadband systems are also considered as a capable telecommunication technology.

The realization of broadband communication services over powerline grids offers a great opportunity for cost-effective telecommunication networks without laying of new cables. However, electrical supply networks are not designed for information transfer, and there are some limiting factors in the application of broadband PLC technology. Therefore, the distances that can be covered are limited, as well as the data rates that can be realized by PLC systems. A further very important aspect for application of broadband PLC is its electromagnetic compatibility. For realization of broadband PLC, a significantly wider frequency spectrum is needed (≤30 MHz) than is provided within CENELEC bands. On the other hand, a PLC network acts as an antenna becoming a noise source for other communication systems working in the same frequency range (e.g., various radio services). For this reason, broadband PLC systems have to operate with a limited signal power, which decreases their performance (data rates, distances).

In contrast to narrowband PLC systems, no specified standards apply to broadband PLC networks. The standardization process is currently led by several international bodies. PLCforum [4] is an international association formed to unify and represent the interests of all players engaged in PLC (utility, providers, manufacturers, developers, researchers) and to expedite standardization and regulation on PLC technology worldwide as well as to support its commercialization. HomePlug Powerline Alliance [5] is a similar organization that is oriented to in-home PLC technology, providing both narrowband and broadband applications. Concrete work on technical standardization is done within ETSI (European Telecommunications Standards Institute) and CENELEC.

Current broadband PLC systems provide data rates beyond 2 Mbps in the outdoor arena, which includes middle- and low-voltage supply networks (Fig. 1), and up to 12 Mbps in the in-home area. Some manufacturers have already developed product prototypes providing much higher data rates (~40 Mbps). Middle-voltage PLC technology is usually used for realization of point-to-point connections bridging distances up to several hundred meters. Typical application areas of such systems is connection of LAN networks between buildings or campuses and connection of antennas and base stations of cellular communication systems to their backbone networks. Low-voltage PLC technology is used for realization of the so-called last mile of telecommunication access networks. Because of the importance of telecommunication access, current development of broadband PLC technology is directed mostly to applications in access networks, including the in-home area.

3. PLC ACCESS NETWORKS

3.1. PLC Alternative for Communication Over the "Last Mile"

The access networks are very important for network providers because of their high costs and the possibility for the realization of a direct access to the end users and subscribers. Typically, about 50% of all investments in the telecommunication infrastructure is needed for the realization of the access networks. Following the deregulation of the telecommunication market in a large number of countries, the access networks are still owned by the former monopolistic companies (incumbent providers). New network providers build up their wide-area networks (WANs), but they still have to use the access infrastructure owned by incumbent providers.

Consequently, the new network providers are trying to find a solution to realize their own access network.

Building new access networks (cable or fiberoptic networks, mobile and fixed wireless access systems, satellite networks) is the best way to implement the newest communication technology, which allows realization of attractive services and applications. On the other hand, the realization of new access networks is expensive and, in the case of laying new fiber or cable, takes a long time. The expensive buildup of new communication networks can be avoided by using existing infrastructure. In this case, already existing wireline networks are candidates for connection of subscribers to the telecommunication transport networks. This is possible by using the following infrastructure:

- Telephone networks
- TV cable networks (CATV)
- Power supply networks

Telephone networks usually belong to the former monopolistic companies, and this is a major disadvantage for new network providers to use them to offer services such as ADSL. That is very often the case with CATV networks, too. Additionally, CATV networks have to be made capable for bidirectional transmission, which results in extra costs. Therefore, the usage of power supply networks for communications seems to be reasonable.

3.2. Network Structure

A low-voltage supply network consists of a transformer unit and a number of power supply cables connecting the end users/subscribers (Fig. 2). The transformer unit connects the low-voltage supply network to the medium- and high-voltage levels. A PLC system applied to a low-voltage network uses the power grids as a communication medium and is connected to the backbone communication networks (WAN) via a base station that is usually placed within the transformer unit. The base station may also be located elsewhere on the powerline, such as at a subscriber's premises.

Many utilities supplying electrical power have their own telecommunication networks that can be used as backbone networks for PLC access systems. If this is not the case, a PLC access network can be connected to a conventional telecommunication network. In any case, the transmitted signal from the backbone has to be converted into a form that makes possible its transmission over a low-voltage power supply network. The conversion takes place at the base station of a PLC system.

PLC subscribers are connected to the network via a PLC modem, placed in the electrical power meter unit (Fig. 2, M). The modem converts the signal received from the PLC network in a standard form that can be processed by the conventional communication systems. On the user side, standard communication interfaces (Ethernet, ISDN S_0, etc.) are offered. Within a house, the transmission can be realized via a separate communication network or via an internal electric installation (in-home PLC solution). In this way, a number of communication devices within a house can be directly connected to a PLC access network.

3.3. In-Home PLC Networks

In-home PLC (indoor) systems use the internal electrical infrastructure as a transmission medium. This makes it possible to setup PLC local networks that connect the typical devices used in private homes, including telephones, computers, printers, and video devices. In the same way, small offices can be interconnected by LAN systems realized with PLC technology. Automation services have become more and more popular not only for their application in the industrial and business sector within large buildings but also for private households. The systems providing automation services such as security observation, heating control, and automatic light control, have to connect a big number of end devices, such as sensors, cameras, electromotors, and lights. Therefore, in-home PLC technology seems to be a reasonable solution for the realization of such networks with a large number of end devices, especially within older houses and buildings that do not have an appropriate internal communication infrastructure.

Basically, the structure of in-home PLC networks is not much different from PLC access systems using low-voltage supply networks. There is also a base/main station that controls the in-home PLC network and connects it to the outdoor area. The base station can be placed within the meter unit (Fig. 2), but also in any other suitable place in the in-home PLC network. All devices of an in-home network, are connected via PLC modems like subscribers of a PLC access network.

An in-home PLC network can exist as an independent network covering only a house or a building. However, it excludes usage and control of in-home PLC services from a distance and also access to WAN services from each electrical socket within a house. In-home PLC networks can be connected to a PLC access system and also to an access network realized by any other communication technology.

3.4. Network Elements

PLC systems consists of the following network elements:

- Basic PLC network elements, which exist in every PLC network
  - PLC modem
  - PLC base/master station
- Additional network elements
  - PLC repeater
  - PLC gateway

A PLC modem connects standard communication equipment, used by PLC subscribers, to a powerline transmission medium. The user-side interface can provide various standard interfaces for different communication devices (e.g., Ethernet and USB interfaces for realization of data transmission and S_0 and *a/b* interfaces for telephony). On the other side, the PLC modem is connected to the power grid using a specific coupling method [3] that allows the feeding of communication signals to the powerline medium and its reception. The coupling has to ensure a safe galvanic separation and to act as a highpass filter dividing the

communication signal (>9 kHz) from the electrical power (50 or 60 Hz). To reduce electromagnetic emission from the powerline (Section 4.4), the coupling is realized between two phases in the access area and between a phase and the neutral conductor in the indoor area. The PLC modem implements all functions of the physical layer including modulation and coding. The second communication layer (link layer) is also implemented within the modem, including its MAC (media access control) and LLC (logical link control) sublayers.

A PLC base station (master station) connects a PLC access system to its backbone network (Fig. 2). It realizes the connection between the backbone communication network and the powerline transmission medium. However, the base station does not connect individual subscriber devices, but it may provide multiple network communication interfaces. Usually, the base station controls the operation of a PLC access network. However, the realization of network control or its particular functions can be realized in a distributed manner.

Additional network elements are needed in some PLC systems to provide signal conversion between different network segments. Repeaters (R) make it possible to realize longer network distances, consisting of several network segments (Fig. 3). The segments are separated by using different frequency bands or by different time slots. In the second case, a time slot is used for the transmission within the first network segment, and another slot for the second segment. A repeater does not modify the contents of the transmitted information.

A gateway is used to interconnect a PLC access network and an in-home PLC network. Similar to a repeater, a gateway also converts the signal between the access and in-home frequencies or time slots. It is usually placed near the house meter unit (Fig. 2) providing the division of the access and in-home areas on the logical network level, too. In this case, an in-home network is fully controlled by the gateway and operates independently form the access network. Therefore, the gateway acts as a subscriber of the access network realizing the connection between the in-home and the access areas. Generally, a gateway can be also placed anywhere in a PLC access network to provide both network segmentation (repeater functionality) and network separation on the logical level.

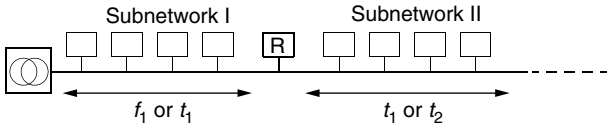


Figure 3. PLC network with repeaters.

4. PLC SYSTEM CHARACTERISTICS

4.1. Network Topology

Low-voltage supply networks are realized by various technologies (different types of cable, transformer units, etc.) according to the existing standards, which differ from country to country. The topologies of low-voltage power supply networks are also different and depend on several factors:

- *Network location* — a PLC network can be installed in a residential, industrial or business area. Furthermore, there is a difference between rural and urban residential areas.
- *Subscriber density* — the number of users/subscribers as well as the user concentration vary from network to network. The subscribers can be placed in family houses (low subscriber density), within houses with several individual customers, in buildings with a larger number of flats or offices, and within apartment or business towers (very high subscriber density).
- *Network span* — the longest distance between the transformer unit and a customer also varies.
- *Network structure* — low-voltage networks usually consist of several network sections/branches, and this number also varies.

Figure 4 shows a possible PLC network structure. There are generally several network sections and branches from the transformer station to the end users. Each branch can have a different topology and connects a variable number of users.

Some characteristic values describing a typical European PLC network, given in Refs. 6 and 7, are listed below — note that the number of users, the number of

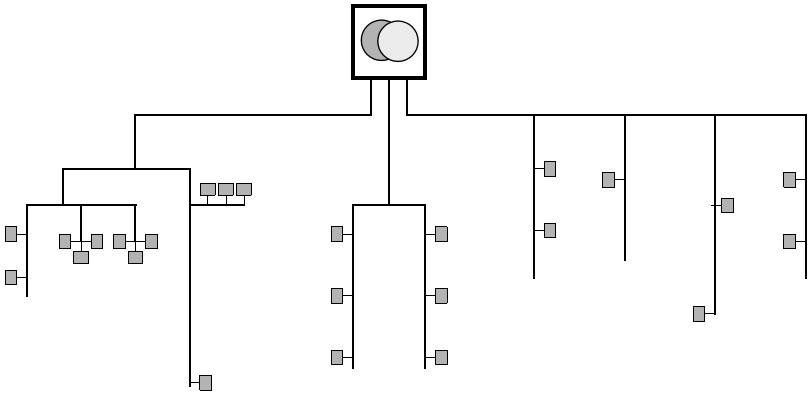


Figure 4. Possible topology of a low-voltage supply network.

potential PLC subscribers, in the United States is significantly lower:

- Number of users in the network: 250–400
- Number of network sections: ~5
- Number of users in a network section: 50–80
- Network span: ~500 m

The PLC base station is usually located within the transformer unit (Fig. 4). However, it could also be placed somewhere else within the network depending on the appropriate interconnection point to the backbone. Each network section/branch may form an individual PLC subnetwork, including a base station and a number of subscribers. Furthermore, a network can be segmented into multiple PLC subnetworks by using repeaters or gateways (Section 3.4). A large PLC network can also cover multiple low-voltage networks. In this case, a number of electrical networks are connected, but only for data transmission (passing only the high-frequency communication signal). Independently of the position of the base station and the kind of network segmentation or concentration, PLC access networks as well as PLC in-home systems physically maintain a tree network topology.

In any network configuration, the communication between PLC subscribers and the WAN is carried out over a base station. This is also the case for internal communication between PLC subscribers. A signal sent by any network station to the base station (uplink transmission direction) reaches all network stations. This is also the case for a signal transmitted in the downlink. Thus, a PLC access network has a star topology with a base station in its center, which leads to a logically bus network structure representing a shared transmission medium.

4.2. Characteristics of the PLC Transmission Channel

The electrical power grid has not been designed to support telecommunication services. To use these lines as an information transmission medium, we need to determine the transfer characteristics, including the attenuation and the impedance. From these parameters, we construct a mathematical channel model that is used in the design of the PLC systems.

The propagation of signals over powerlines introduces an attenuation, which increases with the length of the line and frequency. The attenuation is a function of the powerline characteristic impedance Z_L and the propagation constant γ [8]. These two parameters are defined by the primary resistance R' per unit length, the conductance G' per unit length, the inductance L' per unit length and the capacitance C' per unit length, which are generally frequency dependent, as

$$Z_L(f) = \sqrt{\frac{R'(f) + j2\pi f \cdot L'(f)}{G'(f) + j2\pi f \cdot C'(f)}} \tag{1}$$

and

$$\gamma(f) = \sqrt{(R'(f') + j2\pi f L'(f)) \cdot (G'(f') + j2\pi f C'(f))} \tag{2}$$

$$\gamma(f) = \alpha(f) + j\beta(f) \tag{3}$$

Considering a matched transmission line, which is equivalent to regarding only the wave propagation from source to destination, the transfer function of a line with length l can be expressed by

$$H(f) = e^{-\gamma(f)\cdot l} = e^{-\alpha(f)\cdot l} \cdot e^{-j\beta(f)\cdot l} \tag{4}$$

Investigations and measurements of the fundamental properties of power cables have revealed that $R'(f) \ll 2\pi f L'(f)$ and $G'(f) \ll 2\pi f C'(f)$ is valid in the frequency range from 1 to 30 MHz. Consequently, the dependence of L' and C' on frequency is neglected so that the characteristic impedance Z_L and the propagation constant γ in this frequency range can be expressed as [8]

$$Z_L = \sqrt{\frac{L'}{C'}} \tag{5}$$

and

$$\gamma(f) = \frac{1}{2} \cdot \frac{R'(f)}{Z_L} + \frac{1}{2} \cdot G'(f) \cdot Z_L + j2\pi f \cdot \sqrt{L'C'} \tag{6}$$

$$\gamma(f) = k_1 \cdot \sqrt{f} + k_2 \cdot f + jk_3 \cdot f \tag{7}$$

$$\gamma(f) = \alpha(f) + j\beta(f) \tag{8}$$

where k_1, k_2, and k_3 are constants.

The real part of the propagation constant, describing the cable losses, can be approximated by the equation

$$\alpha(f) = a_0 + a_1 \cdot f^k \tag{9}$$

and with a suitable selection of the attenuation parameters a_0, a_1, and k, the powerline attenuation, representing the amplitude of the channel transfer function, can be defined by the formula [9]

$$A(f, l) = e^{-\alpha(f)\cdot l} = e^{-(a_0 + a_1 \cdot f^k)} \cdot l \tag{10}$$

where l represents the length of the path for the signal wave propagation and k is the exponent of the attenuation factor.

4.3. PLC Channel Model

In addition to the frequency-dependent attenuation that characterizes the powerline channel, deep narrowband notches occur in the transfer function, which may be spread over the whole frequency range. These notches are caused by multiple reflections at impedance discontinuities. The length of the impulse response and the number of the occurred peaks can vary considerably depending on the environment. This behavior can be described by the "echo model" [10]. Transfer characteristics of powerline channels can be regarded as quasistationary, as their changes occur only as a result of changes in the topology and changes in the load situation. Load changes are caused mainly by the connecting or switching of electrical appliances.

Complying with the echo model, each transmitted signal reaches the receiver on N different paths (see Fig. 5). Each path i is defined by a certain delay τ_i and

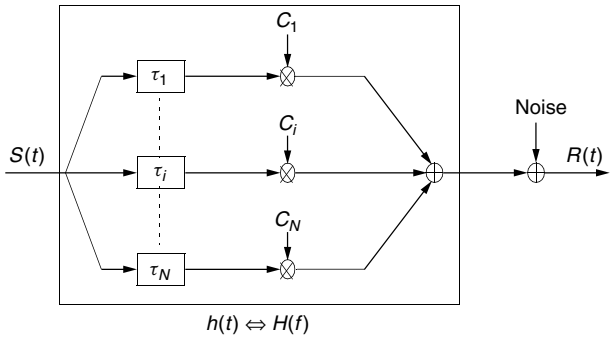


Figure 5. The echo model representing the PLC channel.

a certain attenuation factor C_i. The PLC channel can be described by means of a discrete-time impulse response:

$$h(t) = \sum_{i=1}^{N} C_i \cdot \delta(t - \tau_i) \Leftrightarrow H(f) = \sum_{i=1}^{N} C_i \cdot e^{-j2\pi f \tau_i} \qquad (11)$$

Factoring in the formula of the channel attenuation, the transfer function in the frequency domain can be written

$$H(f) = \sum_{i=1}^{N} g_i \cdot A(f, l_i) \cdot e^{-j2\pi f \tau_i} \qquad (12)$$

where g_i is a weighting factor representing the product of the reflection and transmission factors along the path. The variable τ_i, representing the delay introduced by the path i, is a function of the pathlength l_i.

By replacing the medium attenuation $A(f, l_i)$ by the expression given in Eq. (10), we obtain the final equation defining the PLC channel model, encompassing the parameters of its three main characteristics: attenuation, impedance fluctuations, and multipath effects. This equation is composed of a weighting term, an attenuation term, and a delay term:

$$H(f) = \sum_{i=1}^{N} g_i \cdot e^{-(a_0 + a_1 \cdot f^k) \cdot l_i} \cdot e^{-j2\pi f \tau_i} \qquad (13)$$

4.4. Electromagnetic Compatibility

Broadband PLC systems operate in the high-frequency range from 1.6 to 30 MHz, in order to achieve data rates above 1 Mbps and to avoid the high noise level in the low-frequency range. On the other hand, experiments have shown that PLC systems must overcome an attenuation of about 70 dB, in order to reach from the transformer unit to the subscribers premises [11]. Unlike other communications media, powerlines are electrically asymmetric, as they were not built to transmit information. As a consequence, electromagnetic fields emitted by these lines are high. A solution has to be found to guarantee the coexistence of PLC systems and the radio systems operating in the same spectrum. To solve the problems of electromagnetic compatibility, two solutions are proposed [3]:

1. According to the regulating administration for telecommunications and post (RegTP) in Germany, at present a total spectrum of approximately 7.5 MHz in the frequency range between 0 and 30 MHz may be used principally for PLC. This spectrum is not contiguous, as schematically represented in Fig. 6. It represents some gaps of different width and distributed arbitrarily in the frequency band depicted. This is to secure certain public frequency bands.

At a first glance, it appears feasible to assign the open gaps to PLC services, permitting an increased transmission power spectral density within these chimneys. From such a solution, multicarrier modulation schemes are attractive, particularly those that are able to use narrow gaps with high spectral efficiency, such as orthogonal frequency division multiplex (OFDM). Unfortunately this chimney approach (Fig. 6) encounters some serious problems that render this solution less practical: (a) these gaps are not really free, but they are already dedicated to certain primary users for wireless services, who reserved them for future use; and (b) if the chimneys are very narrow, their use will require the implementation of complex filters.

2. Another proposed solution to accomplish EMC without the chimney approach is a general limitation of radiated fields from powerlines. In March 1999, the RegTP in Germany issued a limitation for the "radiation of telecommunications services in and alongside of cables" (including CATV, xDSL, and PLC). These radiation limitations are part of a plan for the frequency allocation in Germany and are known as NB30.

In comparison with the FCC Part 15, which is the EMC American standard for wire communications, the disadvantages imposed by the German regulations become clearly obvious. At ~2 MHz, for example, the allowed American limits are ~30 dB above the NB30. This means a factor of 1000 in terms of transmitted power or a factor of ~10 in data rate.

4.5. Noise Behavior

Generally, noise in PLC networks can be defined as a superposition of five types, distinguished by their origin, time duration, and spectrum occupancy (see Fig. 7) [12]:

- *Colored background noise* — whose power spectral density (PSD) is relatively low and decreasing with

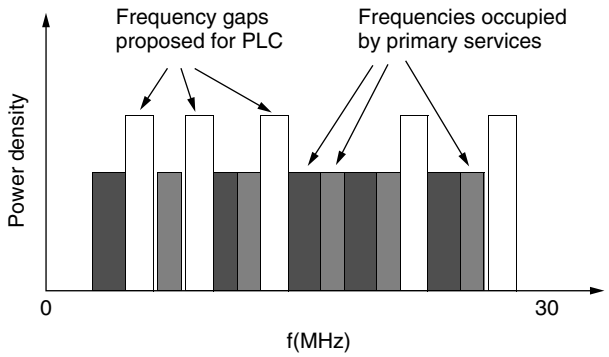


Figure 6. The "chimney approach" for PLC frequency allocation guaranteeing EMC.

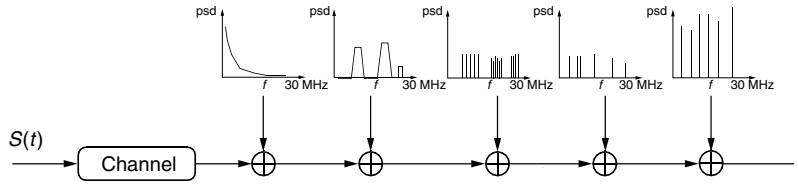


Figure 7. Noise types present in the PLC environment.

frequency. This type of noise is caused mainly by a superposition of numerous noise sources with lower power. Its parameters vary over time in terms of minutes and hours.

- *Narrowband noise* — mostly sinusoidal signals, with modulated amplitudes. This type of noise is caused mainly by ingress of broadcast stations in the medium- and short-wave broadcast bands. Their amplitude generally varies during the daytime.
- *Periodic impulsive noise*, *asynchronous to the main frequency* — impulses that usually have a repetition rate between 50 and 200 kHz, which results in a spectrum with discrete lines with a frequency spacing according to the repetition rate. This type of noise is caused mostly by switching power supplies.
- *Periodic impulsive noise*, *synchronous with the main frequency* — impulses that have a repetition rate of 50 or 100 Hz and are synchronous with the main powerline frequency. They are of short duration (some microseconds) and have a PSD that decreases with frequency. This type of noise is caused by power supplies operating synchronously with the main frequency.
- *Asynchronous impulsive noise* — a type of impulsive noise caused by switching transients in the networks. The impulses have durations of some microseconds up to a few milliseconds with arbitrary arrival times. The PSD of this type of noise can reach values of more than 50 dB above the background noise, which makes it the main cause of occurrence of error in digital transmission.

4.6. PLC Services

PLC access networks together with their backbone provide a bearer service ensuring realization of different teleservices that allow the use of various communication applications [7]. Accordingly, the specification of a PLC transmission system has to include the definition of specific bearer network layers: the physical layer, including modulation and coding methods as well as the data-link layer specifying MAC and LLC sublayers. PLC networks have to offer a large palette of telecommunications services with certain quality requirements to be able to compete with other communications technologies applied to the access area. Therefore, the following four groups of teleservices have to be provided by PLC access networks:

- Connectionless services without QoS guarantees
- Connection-oriented constant-bit-rate (CBR) services, such as telephony
- Connectionless services with QoS guarantees
- Specific PLC services

Present PLC systems provide the connectionless services offering high-speed Internet access to customers. However, PLC networks should also support the classical telephone service, because of its significant penetration in the communications world. Therefore, the manufacturers of PLC equipment have already released systems supporting telephony service, based mostly on VoIP solutions. A special emphasis has to be given to the specific PLC services (home automation, energy management, security, remote functions, etc.), which usually require low data rates and do not require very low transmission delays. So, most of these services can be realized by the connectionless services class without guarantees.

The support of teleservices mentioned above can ensure a competitive position of PLC networks toward other access technologies. However, further development of PLC access networks leads to realization of CBR services with higher data rates and connectionless data services achieving higher QoS requirements.

5. REALIZATION OF PLC NETWORKS

5.1. Specific Performance Problems

The regulatory bodies specify the limits for electromagnetic radiation, which is allowed to be produced by PLC systems operating out of the frequency range defined by the CENELEC standard. In Germany, NB30 directions define very low radiation limits for systems like PLC, which operate in the frequency range up to 30 MHz (see Section 4.4). Accordingly, PLC networks have to operate with a limited signal power to stay within the NB30 limits. Because of the limited signal power, PLC networks are more sensitive to disturbances and are not able to span longer distances.

Well-known error-handling mechanisms can also be applied to PLC systems to reduce the problem of transmission errors caused the disturbances [e.g., forward error correction (FEC) and ARQ]. However, the application of FEC consumes an additional part of the transmission capacity because of the overhead needed for the error detection and correction. ARQ retransmits defective data units, which also consumes a part of the transmission capacity. Additionally, the powerline is a transmission medium that has to be shared by all PLC subscribers.

PLC systems have to compete with other access technologies and to offer a satisfactory QoS and sufficient data rates, but at the same time, to be economically efficient. Therefore, broadband PLC systems have to be provided with the following features:

- Application of efficient modulation schemes ensuring a good utilization of used frequency spectrum and a certain robustness against disturbances

- Realization of suitable coupling methods to reduce electromagnetic radiation
- Implementation of efficient media access control (MAC) protocols to achieve near-maximal utilization of limited PLC network capacity and realization of needed QoS
- Application of optimal error-handling methods to deal with an unfavorable noise scenario consuming a minimum of network resources

5.2. Modulation and Transmission Schemes

Within the PLC systems, the communication is supposed to occur in a channel characterized by frequency-selective phenomena, presence of echoes, and impulsive and colored noise with the superposition of narrowband interferences. This requires that the modulation scheme adopted for PLC effectively face such a hostile environment.

Direct-sequence code-division multiple access (DSCDMA) and orthogonal frequency-division multiplexing (OFDM) are considered as candidates for future broadband PLC [13,14] (see Table 1). As DSCDMA and OFDM permit the separation of the overall transmitted data in many parallel independent substreams, flexible resource management strategies can be implemented. This characteristic is very important in order to cope with channel impairments and to provide fine granularity. This fine granularity is necessary for multimedia services and to achieve a high utilization.

The CDMA technique has the advantages of robustness to narrowband interference and multiple access with low power spectrum density, thus reducing EMC problems. On the other hand, the OFDM technique allows for significant reduction of channel equalizer complexity and increased resistance to narrowband and impulsive noise. Moreover, bit-loading techniques make it possible to achieve a capacity very near the theoretical limit, but at the cost of an increased system complexity, [14].

Overall OFDM seems to be advantageous compared to DSCDMA:

1. The main advantage is obtained by the fact that the channel (the spectrum) is divided into many narrow subchannels. Therefore, equalizing in OFDM is a simple procedure, compared with wideband equalizing.
2. OFDM inherently solves one essential problem associated with high-speed PLC: intersymbol interference (ISI) caused by the multipath delay spread. This is achieved by the introduction of the "guard interval," which is filled by a cyclic prefix.

For channel coding, two variants have been investigated: Reed–Solomon coding [15] and Hamming codes [3]. Several types of interleavers can be implemented with different variants of OFDM, such as OFDM with diversity or adaptive OFDM [15].

5.3. MAC Layer

The MAC layer specifies a multiple-access scheme and a resource-sharing strategy (MAC protocol). Particularly in PLC systems, the MAC layer has to be robust against disturbances and must allow for the use of various telecommunication services [16].

The most widely applied access scheme in PLC networks is TDMA. Because of the disturbances, data packets (e.g., IP packets) are usually segmented into smaller data units whose size is chosen according to the length of a time slot specified by the TDMA scheme. So, if a disturbance occurs, only erroneous data segments are retransmitted. This consumes a smaller network capacity than does retransmission of the entire data packets. The data segmentation ensures a fine network granularity and an easier provision of QoS guarantees.

An effective solution to avoid the influence of narrowband disturbances is to apply FDMA methods, where particular carrier frequencies can be switched off, if they are affected by narrowband disturbances. Therefore, a TDMA/FDMA combination seems to be a reasonable solution for PLC networks (realized by Ascom [17]). PLC networks using OFDM modulation can also provide transmission channels distributed in a frequency spectrum. Unlike FDMA, the transmission channels are realized by a number of subcarriers, which leads to an OFDM access scheme (OFDMA). A further division of the transmission channels in the time domain can also be done according to a slotted nature of OFDM transmission systems combining both multiple-access schemes (OFDMA/TDMA) [18].

MAC protocols for PLC systems have to achieve a maximum utilization of the limited network capacity and to realize time-critical telecommunication services. This can be ensured by reservation of bandwidth that allows particular QoS guarantees needed for various services [16]. In the case of reservation protocols, a part of the network resources is reserved for the reservation procedure: signaling, organized according to a random or dedicated access principle as well as by application of various hybrid solutions [18]. Figure 8 presents the signaling delay caused by the reservation protocols with a signaling procedure organized according to random and dedicated access methods. The protocols based on random access achieve significantly shorter signaling delay, if the transmission requests are relatively infrequent. However, in the case of frequent requests, protocols with dedicated

Table 1. Comparison of PLC Modulation Candidates

	Spectral Efficiency (bps/Hz)	Maximum Data Rate (Mbps)	Robustness against Channel Distortions	Flexibility and Adaptive Features	System Costs (Including Equalizers and Repeaters)	EMC
DSCDMA	<0.1	~ 0.5	Poor	Very poor	Very poor	Excellent
OFDM	$\gg 1$	>10	Excellent	Excellent	Poor	Good

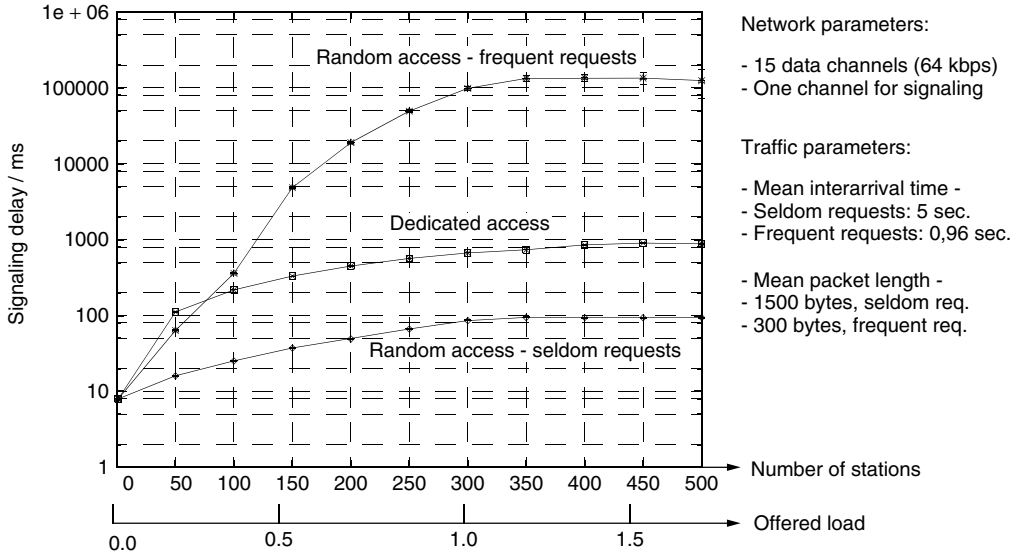


Figure 8. Average signaling delay of different reservation protocols.

access behave much better. So, an optimal solution for the organization for the reservation/signaling procedure can be achieved by a hybrid protocol changing dynamically from random to dedicated access depending on the current situation in the network (network load, number of active stations, etc.). However, the reservation procedure has to be particularly protected against disturbances [19]. Besides reservation MAC protocols, variations of the CSMA/CA protocol ensuring realization of multiple priorities for different services [20] are also widely applied to PLC systems (e.g., Itran technology [21]).

Because of the asymmetric and changing nature of data traffic in the access area, dynamic duplex schemes are used in PLC access networks. This allows the optimal utilization of the network resources, in both downlink and uplink transmission directions according to the current load situation. However, the relatively small PLC network capacity makes it difficult for the simultaneous provision of a required QoS for a high number of subscribers. Therefore, PLC systems have to implement traffic scheduling strategies, including connection admission control (CAC), to limit the number of active subscribers ensuring a satisfactory QoS for currently admitted connections. In the same way, a part of the network resources has to be reserved for capacity reallocation in case of disturbances.

5.4. Error Handling

Because of the special disturbance characteristic, error handling within different network layers warrants careful attention [19]. After a correct dimensioning, it is assumed that the signal-to-noise ratio (SNR) is sufficient to avoid any influence of the background noise. However, the impulsive noise makes it difficult to ensure error-free transmission (Section 4.5). Its influence is reduced by the following methods:

- Setting a sufficient duration of the transmitted symbol within the physical network layer (e.g., duration of an OFDM symbol) eliminates disturbances that are shorter than the symbol duration.
- Channel coding using FEC mechanisms and interleaving allow for the correction of a number of erroneous bits in the case of various kind of disturbances (e.g., single, periodic, and burst errors). However, FEC mechanisms provide overhead, and interleaving increases the transmission delay. Because of the limited network capacity and the demand for low delays, channel coding in PLC networks is organized in a dynamic manner, providing a changing level of protection according to the current noise conditions in the network.
- If the reduction of the BER by channel coding is not sufficient and delay requirements are not too hard (e.g., data traffic), even ARQ mechanisms (retransmission of erroneous data) can be used.

Application of ARQ can improve network performances significantly (Fig. 9). Network utilization in a network without ARQ (only a simple packet retransmission is provided) can achieve a maximum of 50%. Application of go-back-N ARQ retransmitting smaller portions of erroneous packets (disturbed data segments) improves the network utilization by 23%. If the ARQ is additionally adjusted to a per packet reservation method (e.g., ARQ + mechanism [19]), utilization achieves 81%.

In the case of long-term disturbances, a part of the network capacity is unavailable for a longer time. This cannot be improved by FEC and ARQ. Therefore, this part of the transmission capacity (a frequency spectrum) is temporarily not usable (is switched off).

6. SUMMARY

Present powerline communications systems, using electrical power grids as transmission media, provide relatively high data rates (>2 Mbps). PLC can be applied to high-,

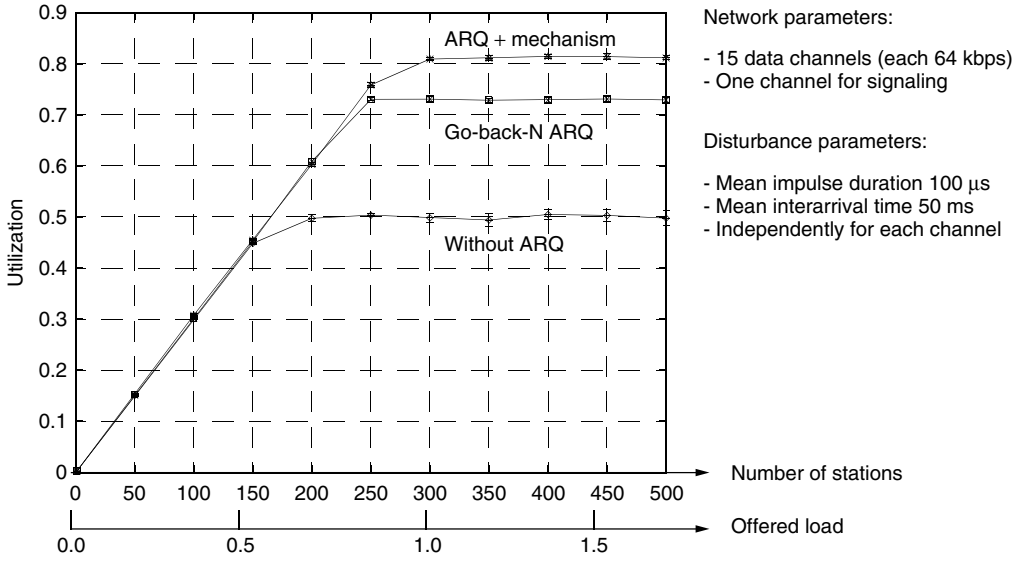


Figure 9. Performances of different ARQ mechanisms — average network utilization.

medium-, and low-voltage supply networks as well as within buildings. PLC technology is now used mainly for access networks and in-home communication networks. This is because of the high cost of the access networks (about 50% of the investments in network infrastructure are needed for the access area) and the liberalization of the telecommunication market in many countries.

Power supply networks are not designed for communications, and they do not present a favorable transmission medium. The PLC transmission channel is characterized by a large, frequency-dependent attenuation, changing impedance, and fading, as well as a strong influence of noise. On the other hand, broadband PLC networks have to operate in a frequency spectrum up to 30 MHz, which is used by various radio services, too. Therefore, the regulatory bodies specify very strong limits regarding the electromagnetic emission from PLC networks to the environment. As a consequence, PLC networks have to operate with a limited signal power, which reduces network distances and data rates, and also increases sensitivity to disturbances.

To reduce the negative impact of powerline transmission medium, PLC systems apply efficient modulation, such as OFDM, which is able to avoid restricted frequency bands. OFDM is also robust against narrowband and impulsive noise and can deal with fading. The problem of longer noise spikes can be solved by well-known error-handling mechanisms (e.g., FEC, ARQ). However, their application consumes a certain portion of the PLC network capacity due to overhead and retransmission. The PLC bandwidth is shared by the subscribers, and therefore any reduction of capacity due to protocol overhead should be minimized. At the same time, PLC systems have to compete with other access technologies and to offer a wide palette of telecommunication services with a satisfactory QoS. Both good network utilization and provision of QoS guarantees can be achieved by an efficient MAC layer. We have shown that reservation-based MAC protocols can achieve these goals efficiently.

ACRONYMS

ADSL	Asymmetric digital subscriber line
ARQ	Automatic repeat request
BER	Bit error rate
CAC	Connection admission control
CATV	Cable TV
CBR	Constant bit rate
CFS	Carrier frequency systems
CSMA/CA	Carrier sensing multiple access with collision avoidance
DS-CDMA	Direct sequence code division multiple access
DSL	Digital subscriber line
EMC	Electromagnetic compatibility
ETSI	European Telecommunications Standards Institute
FDD	Frequency-division duplex
FDMA	Frequency-division multiple access
FEC	Forward error correction
FSK	Frequency shift keying
IP	Internet Protocol
ISDN	Integrated Services Digital Network
LAN	Local-area network
LLC	Logical link control
MAC	Media access control
OFDM	Orthogonal frequency-division multiplexing
OFDMA	OFDM access
PLC	Powerline communications
psd	Power spectral density
QoS	Quality of service
RCS	Ripple carrier signaling
RegTP	Regulating Administration for Telecommunications and Post
TDD	Time-division duplex

TDMA	Time-division multiple access
VBR	Variable bit rate
VoIP	Voice over IP
WAN	Wide-area network

BIOGRAPHIES

Halid Hrasnica, graduated in 1993 at the Faculty for Electrical Engineering — Department for Telecommunications, at the University of Sarajevo — Bosnia and Herzegovina. From 1993 to 1995 he was working in Energoinvest Communications in Sarajevo as developing software engineer for the telephone exchange systems. In 1995 he joined the Chair for Telecommunications at Dresden University of Technology, Germany, as visitor scientist. Since 1996 he is research assistant at Dresden University of Technology and he is currently working toward his Ph.D. He has been involved in several research projects: development of least cost routing strategy, performance analysis and simulation of broadband communications networks. His current research interest is performance analysis of powerline communication networks and investigation on PLC MAC layer and protocols.

Abdelfatteh Haidine received the B.S. in electronics and telecommunications from the University Cadi Ayyad in Marrakech and the MSc in telecommunications from University Chouaib Doukkali El Jadida in 1999, in Morocco. Since 2000, he joined the University of Technology Dresden in Germany as research assistant. He worked in different European and German projects about power line communication networks. His actual research domain is planning and optimization of the access network based on optical fiber and Very high bit Digital Subscriber Line (VDSL) technology.

Ralf Lehnert received both his 1972 diploma degree and the 1979 Ph.D. degree in electrical engineering from Aachen University, Germany. Since 1980 he has been with the basic development department at Philips Communication Systems in Nuremberg, Germany, as the head of a group on applied research in performance evaluation of communication networks, traffic engineering and network planning. In July 1994 he took over the chair for telecommunications, as a full professor in the department of electrical engineering at Dresden University, Germany. His current research interests are in the field of performance evaluation of telecommunication systems, including modeling of B-ISDN networks, network planning and optimization. He has been involved in RACE (Parasol, Exploit, Tribune) and ACTS projects (Expert) all on the subject of ATM.

BIBLIOGRAPHY

1. Cenelec (online) *http://www.cenelec.org* (June 2002).
2. Reg TP — Regulierungsbehörde für Telekommunikation und Post (online), *http://www.regtp.de* (June 2002).
3. K. Dostert, *Powerline Communications*, Prentice-Hall, 2001.
4. PLCforum (online), *http://www.tech.ascom.ch/preview/plc/index.htm* (June 2002).
5. HomePlug Powerline Alliance (online), *http://www.homeplug.org/index_basic.html* (June 2002).
6. O. G. Hooijen, On the channel capacity of the residential power circuit used as a digital communications medium, *IEEE Commun. Lett.* **2**(10): (Oct. 1998).
7. H. Hrasnica and R. Lehnert, *Powerline Communications in Telecommunication Access Area*, VDE World Microtechnologies Congress, MICRO.tec 2000 ETG-Fachtagung und -Forum: Verteilungsnetze im liberalisierten Markt, Sept. 25–27, 2000. Expo 2000, Hannover, Germany.
8. M. Zimmerman and K. Dostert, The low voltage power distribution network as last mile access network — signal propagation and noise scenario in the HF-range, *Int. J. Electron. Commun.* **54**(1): 13–22 (2000).
9. M. Zimmerman, *Energieverteilnetze als Zugangsmedium fuer Telekommunikationsdienste*, Ph.D. dissertation, Karlsruhe Univ. Technology, Shaker Verlag 2000 (in German).
10. H. Philipps, Development of a statistical model for powerline communication channels, *Proc. Int. Symp. Power-Line Communications and Its Applications*, Limerick, Ireland, 2000.
11. H. Dalichau, *EMV-Aspekte von Inhome-PLC-Anlagen, Vergleich des kHz-Bereiches mit dem MHz-Bereich*, EMC Kompendium, 2002 (in German).
12. M. Zimmerman and K. Dostert, An analysis of the broadband noise scenario in powerline networks, *Proc. Int. Symp. Power-Line Communications and Its Applications*, Limerick, Ireland, 2000.
13. S. Tachikawa, M. Nari, and M. Hamamura, Power line data transmission using OFDM and DS/SS systems, *Proc. 6th Int. Symp. Power-Line Communications and Its Applications*, Athens, Greece, 2002.
14. E. Del Re, R. Fantacci, S. Morosi, and R. Seravalle, Comparison of CDMA and OFDM systems for broadband downstream communications on low voltage power grid, *Proc. 5th Int. Symp. Power-Line Communications and Its Applications*, Malmö, Sweden, 2001.
15. T. Waldeck, *Einzel- und Mehrträgerverfahren für die störresistente Kommunikation auf Energieverteilnetz*, Ph.D. dissertation, Karlsruhe Univ. Technology, 1999, Logos Verlag, Berlin, 2000 (in German).
16. H. Hrasnica, A. Haidine, and R. Lehnert, Reservation MAC protocols for powerline communications, *Proc. 5th Int. Symp. Power-Line Communications and Its Applications* (*ISPLC2001*), Malmö, Sweden, April 4–6, 2001.
17. Ascom Powerline Communications, *Powerline System Description*, version 1.2 (March 2002).
18. H. Hrasnica and R. Lehnert, Performance analysis of polling based reservation MAC protocols for broadband PLC access networks, *Proc. 14th Int. Symp. Services and Local Access* (*ISSLS2002*), Seoul, Korea, April 14–18, 2002.
19. H. Hrasnica and R. Lehnert, Performance analysis of error handling methods applied to a broadband PLC access network, *Proc. SPIE Int. Symp. ITCom2002*, Boston, MA, July 29–Aug. 1, 2002.
20. T. Langguth et al., *Performance study of access control in power line communications, Proc. 5th Int. Symp. Power-Line Communications and Its Applications* (*ISPLC2001*), Malmö, Sweden, April 4–6, 2001.
21. Itran Communications Ltd. (online), *http://www.itrancomm.com/index1.html* (June 2002).

PRODUCT CODES

FRANK R. KSCHISCHANG
University of Toronto
Toronto, Ontario, Canada

1. INTRODUCTION

The *product construction*, introduced by Peter Elias in 1954 [1], is one of the simplest ways to combine simple error control codes to get more powerful ones, and any code that results from the product construction is called a *product code*. The codewords of a product code are defined as matrices that are constrained, like crossword puzzles, so that rows and columns form valid codewords in some constituent codes. Decoding typically also proceeds in crossword-puzzle fashion, alternating between the horizontal and vertical constraints in an attempt to find a valid codeword near the received word. Although more complicated codes may offer superior performance, product codes are often attractive for practical implementation — they can be designed to provide excellent performance with reasonably low decoding complexity, and their structure leads to hardware decoder implementations with natural parallelism that is easily exploited.

This article describes product codes, derives some of their basic properties, discusses methods to decode them, and provides a number of examples.

1.1. Linear Codes

We consider only *linear block codes* in this article, so, to establish notation and to be somewhat self-contained, we will start with a brief review of their properties. The basic theory of linear block codes is described in every textbook on coding theory; see, [e.g., 2,3]. While all of the constructions we describe will apply to codes defined over any finite field $\mathbb{F}$, for simplicity we will confine all of our examples to codes defined over the binary field $\mathbb{F}_2 = \{0, 1\}$, with all scalar arithmetic (addition and multiplication) computed modulo 2. For use throughout this article, it is convenient to denote the vector space of $m \times n$ matrices over $\mathbb{F}$ as $\mathbb{F}^{m\times n}$, where m and n are positive integers. The space of $1 \times n$ row vectors is denoted as $\mathbb{F}^n$. If $\mathbf{M}$ is a matrix in $\mathbb{F}^{m\times n}$, then $[\mathbf{M}]_{i,j}$ denotes the component of $\mathbf{M}$ in row i, column j, where $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$.

A block code C of length n over $\mathbb{F}$ is any nonempty subset of $\mathbb{F}^n$. More specifically, C is *linear* if it forms a subspace of $\mathbb{F}^n$, that is, if C satisfies all the axioms that define a vector space, including the requirement that every $\mathbb{F}$-linear combination of codewords is itself a codeword. A code of length n and dimension k is referred to as an $[n, k]$ code.

The simplest example of a binary linear code is the $[k + 1, k]$ single-parity-check (SPC) code, which consists of all binary vectors of length $k + 1$, each having an even number of ones. For example, the $[3, 2]$ SPC code has four codewords: $\{000, 011, 101, 110\}$. It is easy to check that any linear combination of codewords is a codeword; for example, $011 + 101 = 110$.

An *encoder* for a linear code C maps a message of k (input) symbols to a codeword of $n \geq k$ (output) symbols. The ratio of input to output symbols, namely, k/n, is called the *rate* of the code. An encoder is called *systematic* if the message symbols appear directly as the first k symbols of each codeword. A systematic encoder for the binary SPC code would take in k message bits and append a single *check bit*, chosen to make the total number of ones in the codeword an even number.

1.2. Generator and Parity-Check Matrices

An $[n, k]$ code C is a k-dimensional vector space; accordingly, it is always possible to find k linearly independent vectors $g_1, \ldots, g_k$ in C. Any such set is a basis or *generating set* for C, since every codeword v of C may be expressed (uniquely) as an $\mathbb{F}$-linear combination of the generators

$$v = \sum_{i=1}^{k} u_i g_i, \tag{1}$$

where $u_i \in \mathbb{F}$ for all $i \in \{1, \ldots, k\}$. Writing $u = (u_1, \ldots, u_k)$ as a (row) vector with k components, and collecting $g_1, \ldots, g_k$ as the rows of a matrix $\mathbf{G} \in \mathbb{F}^{k\times n}$, we may express (1) in matrix form as $v = u\mathbf{G}$. The matrix $\mathbf{G}$ is referred to as a *generator matrix* for C, and C is equal to the *row space* of $\mathbf{G}$. Since C can have many different bases, it follows that C can have many different generator matrices (all of which, however, must have row space C).

Suppose that C has a generator matrix of the form $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$, where $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{P} \in \mathbb{F}^{k\times(n-k)}$ is arbitrary. Then multiplication of an information vector u by $\mathbf{G}$ results in the systematic encoding $u\mathbf{G} = (u, u\mathbf{P})$ of u. In this case, $\mathbf{G}$ is said to be in *systematic form*. Although not every code has a generator matrix in systematic form, the $[3, 2]$ SPC code *does* have a generator matrix in systematic form given by

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

In general, since multiplication by $\mathbf{G}$ effects a one-to-one mapping from $\mathbb{F}^k$ to $\mathbb{F}^n$, it follows that $\mathbf{G}$ always has at least one (in general more than one) right-inverse $\mathbf{G}^{-1} \in \mathbb{F}^{n\times k}$, with the property that $\mathbf{G}\mathbf{G}^{-1} = \mathbf{I}_k$. In other words, from every codeword $c = u\mathbf{G}$, it is always possible to recover a unique message $u = c\mathbf{G}^{-1}$.

An alternative description of a code C arises as follows. Define the scalar ("inner" or "dot") product between two vectors $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$ as $\langle v, w\rangle := \sum_{i=1}^{n} v_i w_i$. For every $[n, k]$ code C, one can always find a set of linearly independent n vectors $h_1, \ldots, h_{n-k}$ such that the scalar product $\langle h_i, v\rangle = 0$ for all $i \in \{1, \ldots, n - k\}$ if and only if v is a codeword of C. Collecting $h_1, \ldots, h_{n-k}$ as the rows of an matrix $\mathbf{H} \in \mathbb{F}^{(n-k)\times n}$, we see that a given vector v is a codeword of C if and only if v satisfies the *parity-check equation* $v\mathbf{H}^T = 0$. The matrix $\mathbf{H}$ is called a *parity-check matrix for* C. As is the case with generator matrices, a code C can have many different parity-check matrices. One possible parity-check matrix

for a code with systematic generator matrix $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{P}]$ is $\mathbf{H} = [-\mathbf{P}^T \mid \mathbf{I}_{(n-k)}]$. For example, the $[3, 2]$ SPC code has parity-check matrix $\mathbf{H} = [111]$. In general, it is useful to allow *any* matrix $\mathbf{H}$ with n columns and rank $n - k$ satisfying $v\mathbf{H}^T = 0$ for all $v \in C$ to serve as a parity-check matrix for C, even if $\mathbf{H}$ contains some linearly dependent rows (redundant checks).

1.3. Minimum Hamming Distance

The *Hamming weight* $wt(v)$ of a vector v is defined as the number of positions in which v is nonzero. For example, $wt(110) = 2$. The *Hamming distance* $d(v, w)$ between two vectors v and w of the same length is defined as the Hamming weight of their difference: $d(v, w) := wt(v - w)$. For example, $d(011, 101) = wt(011 - 101) = wt(110) = 2$. It is easily seen that the Hamming distance between two vectors is the number of positions in which the two vectors differ. Thus, since 011 and 101 differ in their first two positions, but not in their last position, we have $d(011, 101) = 2$.

The minimum Hamming distance between pairs of *distinct* codewords in a code C is called the *minimum distance of* C, and is denoted $d_{\min}(C)$. The minimum distance of a code has traditionally been regarded as a parameter of fundamental importance, since this parameter determines the error correction radius $t = \lfloor (d_{\min} - 1)/2 \rfloor$ within which all error patterns are guaranteed correctable by any decoder that maps a received word to the nearest (in the sense of Hamming distance) codeword. In general, it is desirable to make the minimum distance d as large as possible for a given n and k.

For linear codes, since $d(v, w) = wt(v - w)$, the Hamming distance between two codewords v and w is always equal to the weight of some codeword, namely, $v - w$. In other words, every distance (between two codewords) is a weight (of some codeword). On the other hand, since $wt(v) = d(v, 0)$, every weight (of some codeword) is a distance (between two codewords). This equivalence between weight and distance implies that the minimum distance of a linear code C is given by the minimum weight of its nonzero codewords. The $[3, 2]$ SPC code, for example, has minimum distance 2. In general, an $[n, k]$ code with minimum distance d is often referred to as an $[n, k, d]$ code.

For many values of n and k, there are methods known to construct $[n, k]$ codes with large minimum distance, (e.g., see the articles BCH CODES, CYCLIC CODES, FINITE GEOMETRY CODES, HADAMARD CODES, REED–MULLER CODES, and REED–SOLOMON CODES in this encyclopedia).

2. THE PRODUCT CONSTRUCTION

2.1. Definition and Basic Properties

Let A and B be codes over $\mathbb{F}$ of length n_A and n_B, respectively. Whereas the codewords of A and B are vectors, we will now define a code of length $n_A n_B$, the *direct product* of A and B, whose codewords may be viewed as $n_A \times n_B$ matrices over $\mathbb{F}$, namely, as elements of $\mathbb{F}^{n_A \times n_B}$. Every such matrix is readily converted to a vector of length $n_A n_B$ simply by ordering the matrix elements in some way; see Section 2.3.

Definition 1. The direct product $A \otimes B$ is the code consisting of all $n_A \times n_B$ matrices with the property that each matrix column is a codeword of A and each matrix row is a codeword of B.

When A and B are general nonlinear codes, there is no guarantee that $A \otimes B$ is even nonempty. When A and B are linear codes, however, $A \otimes B$ is also linear, with parameters given in the following theorem.

Theorem 1. If A is an $[n_A, k_A, d_A]$ linear code over $\mathbb{F}$ and B is an $[n_B, k_B, d_B]$ linear code over $\mathbb{F}$, then $A \otimes B$ is an $[n_A n_B, k_A k_B, d_A d_B]$ linear code over $\mathbb{F}$.

Before proving this theorem, we give a small example. When A and B both are $[3, 2]$ SPC codes, $A \otimes B$ consists of all the 3×3 matrices in which each row and column contains an even number of ones. The codewords of $A \otimes B$ are listed as follows:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$
$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$
$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

There are $2^{2\times 2} = 16$ codewords of length $3 \times 3 = 9$ with minimum nonzero weight $2 \times 2 = 4$, exactly as predicted by Theorem 1. Even though the SPC codes themselves cannot correct errors, their direct product is single-error correcting, since a single error that occurs in row i, column j causes a parity-check violation in that row and column, a situation that is easily detected and corrected by the decoder.

Given any two codewords $a = (a_1, \ldots, a_{n_A}) \in A$ and $b = (b_1, \ldots, b_{n_B}) \in B$, define the $n_A \times n_B$ matrix $a \otimes b$ as the "outer product" $a^T b$ of a and b, specifically, where $[a \otimes b]_{i,j} = a_i b_j$. It is easy to see that each column of $a \otimes b$ is a scalar multiple of a and each row of $a \otimes b$ is a scalar multiple of b. Thus $a \otimes b$ is in fact a codeword of $A \otimes B$. A codeword in $A \otimes B$ of the form $a \otimes b$ is said to be *separable*. In general $A \otimes B$ contains $1 + (|\mathbb{F}|^{k_A} - 1)(|\mathbb{F}|^{k_B} - 1)$ separable codewords, where $|\mathbb{F}|$ denotes the number of elements in $\mathbb{F}$. When k_A and k_B are large, the separable codewords represent a tiny fraction of

all possible codewords, yet, as we will see, the separable codewords generate the entire product code.

Proof of Theorem 1 The code $C = A \otimes B$ contains the zero matrix, and so is clearly a nonempty code of length $n_A n_B$. The linearity of C follows directly from the linearity of codes A and B and the fact that every $\mathbb{F}$-linear combination of codewords gives rise to a corresponding $\mathbb{F}$-linear combination of rows (or columns).

To see that C has minimum distance $d_A d_B$, let v be a nonzero codeword of C. Then v has a nonzero column $a \in A$, whose weight must be at least d_A. Each nonzero component of a also participates in a nonzero row, and each such row is a nonzero codeword of B of weight at least d_B. Thus v has weight at least $d_A d_B$, and so (1) the minimum distance of C is at least $d_A d_B$; on the other hand, if $a \in A$ and $b \in B$ are nonzero codewords of weight d_A and d_B, respectively, then $a \otimes b$ is a codeword of C of weight $d_A d_B$, so (2) the minimum distance of C is at most $d_A d_B$. Statements (1) and (2) together imply that C has minimum distance exactly $d_A d_B$.

The fact that the dimension of $A \otimes B$ is given by $k_A k_B$ follows directly from Theorem 2 (below).

2.2. Encoding of Product Codes

Given generator matrices for A and B, a convenient encoder for $A \otimes B$ is obtained via the following theorem.

Theorem 2. Let code A have generator matrix $\mathbf{G}_A \in \mathbb{F}^{k_A \times n_A}$, let code B have generator matrix $\mathbf{G}_B \in \mathbb{F}^{k_B \times n_B}$, and let $\mathbf{M}$ be an arbitrary matrix in $\mathbb{F}^{n_A \times n_B}$. Then $\mathbf{M} \in A \otimes B$ if and only if $\mathbf{M} = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$, for some $\mathbf{U} \in \mathbb{F}^{k_A \times k_B}$.

Before proving Theorem 2, we remind the reader of the following simple property of matrix multiplication.

Lemma 1. Let $\mathbf{X}$ and $\mathbf{Y}$ be matrices conformable for the product $\mathbf{XY}$. Then each column of $\mathbf{XY}$ is in the column space of $\mathbf{X}$ and each row of $\mathbf{XY}$ is in the row space of $\mathbf{Y}$.

Proof of Theorem 2 Let $\mathbf{U}$ be an arbitrary $k_A \times k_B$ matrix over $\mathbb{F}$, and consider the matrix product $\mathbf{M} = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$. From Lemma 1 (setting $\mathbf{X} = \mathbf{G}_A^T \mathbf{U}$ and $\mathbf{Y} = \mathbf{G}_B$) we see that every row of $\mathbf{M}$ is in the row space of $\mathbf{G}_B$, and hence is a codeword of B. Likewise (setting $\mathbf{X} = \mathbf{G}_A^T$ and $\mathbf{Y} = \mathbf{U}\mathbf{G}_B$) we see from Lemma 1 that every column of $\mathbf{M}$ is in the column space of $\mathbf{G}_A^T$ or equivalently is in the row space of $\mathbf{G}_A$, and hence is a codeword of A. Thus $\mathbf{M}$ is a codeword of $A \otimes B$.

Conversely, let $\mathbf{M}$ be a codeword of $A \otimes B$. Then, since every row of $\mathbf{M}$ is a codeword of B, it follows that $\mathbf{M} = \mathbf{W}\mathbf{G}_B$ for some $n_A \times k_B$ matrix $\mathbf{W}$. Now, since $\mathbf{W} = \mathbf{M}\mathbf{G}_B^{-1}$, where $\mathbf{G}_B^{-1}$ is a right-inverse of $\mathbf{G}_B$, from Lemma 1 it follows that every column of $\mathbf{W}$ is in the column space of $\mathbf{M}$, and hence is a codeword of A. Thus $\mathbf{W} = \mathbf{G}_A^T \mathbf{U}$ for some $\mathbf{U} \in \mathbb{F}^{k_A \times k_B}$, and $\mathbf{M} = \mathbf{W}\mathbf{G}_B = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$.

Observe that the encoder mapping $m: \mathbb{F}^{k_A \times k_B} \to \mathbb{F}^{n_A \times n_B}$ defined by $m(\mathbf{U}) = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B$ is linear. The kernel of this mapping [i.e., the set of matrices $\mathbf{U}$ such that $m(\mathbf{U}) = \mathbf{0}$] contains just the zero matrix $\mathbf{U} = \mathbf{0}$. Since the kernel of m is trivial, the mapping m is one-to-one, and C, the image of $\mathbb{F}^{k_A \times k_B}$ under the mapping m, has dimension equal to that of $\mathbb{F}^{k_A \times k_B}$, namely, $k_A k_B$, as claimed in Theorem 1.

When $\mathbf{G}_A = [\mathbf{I}_{k_A} \mid \mathbf{P}_A]$ and $\mathbf{G}_B = [\mathbf{I}_{k_B} \mid \mathbf{P}_B]$, so that A and B have systematic encoders, then the encoder mapping m takes the information matrix $\mathbf{U}$ to the codeword

$$m(\mathbf{U}) = \mathbf{G}_A^T \mathbf{U} \mathbf{G}_B = \begin{bmatrix} \mathbf{U} & \mathbf{U}\mathbf{P}_B \\ \mathbf{P}_A^T \mathbf{U} & \mathbf{P}_A^T \mathbf{U} \mathbf{P}_B \end{bmatrix}$$

as illustrated in Fig. 1. The matrix block $\mathbf{U}$ is the information matrix; the block $\mathbf{U}\mathbf{P}_B$ contains "checks on rows," namely, parity-check symbols corresponding to the rows of the message matrix $\mathbf{U}$; the block $\mathbf{P}_A^T \mathbf{U}$ contains "checks on columns," specifically, parity-check symbols corresponding to the columns of the message matrix $\mathbf{U}$; and the block $\mathbf{P}_A^T \mathbf{U} \mathbf{P}_B$ contains "checks on checks," namely, parity-check symbols corresponding to other parity-check symbols. Since $m(\mathbf{U}) = \mathbf{G}_A^T(\mathbf{U}\mathbf{G}_B) = (\mathbf{G}_A^T \mathbf{U})\mathbf{G}_B$, checks on rows and checks on columns may be computed in either order while yielding the identical codeword.

2.3. The Krönecker Product

Let $\mathbf{E}_{i,j}$ be the $k_A \times k_B$ matrix with a one in row i, column j, and zeros in all other positions. Since the set $\{\mathbf{E}_{i,j}: 1 \le i \le k_A, 1 \le j \le k_B\}$ is a basis for $\mathbb{F}^{k_A \times k_B}$, and the encoder mapping m is one-to-one, it follows that $\{m(\mathbf{E}_{i,j}): 1 \le i \le k_A, 1 \le j \le k_B\}$ is a basis for $A \otimes B$. Denote the ith row of $\mathbf{G}_A$ as g_i^A and the jth row of $\mathbf{G}_B$ as g_j^B. Then $m(\mathbf{E}_{i,j}) = \mathbf{G}_A^T \mathbf{E}_{i,j} \mathbf{G}_B = (g_i^A)^T (g_j^B) = g_i^A \otimes g_j^B$. Thus, given a basis $\{g_i^A: 1 \le i \le k_A\}$ for A and a basis $\{g_j^B: 1 \le j \le k_B\}$ for B, the set of outer products $\{g_i^A \otimes g_j^B: 1 \le i \le k_A, 1 \le j \le k_B\}$ is a basis for $A \otimes B$. This proves that separable codewords do indeed generate the entire product code.

It is often useful to "flatten" a product code by converting codewords that are $n_A \times n_B$ matrices into a vectors of length $n_A n_B$. This may be accomplished by establishing a one-to-one correspondence between matrix and vector components. For example, we may define the mapping $s: \mathbb{F}^{n_A \times n_B} \to \mathbb{F}^{n_A n_B}$ that maps a matrix $\mathbf{M}$ to a vector $s(\mathbf{M})$ be assigning $[\mathbf{M}]_{i,j}$ to the $[(i-1)n_B + j]$th component of $s(\mathbf{M})$. In effect, $s(\mathbf{M})$ is obtained by concatenating together the consecutive rows of $\mathbf{M}$; for example

$$s\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}\right) = (110011101)$$

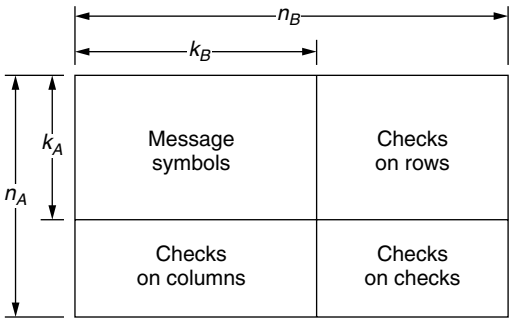


Figure 1. The array formed by the direct product of systematic $[n_A, k_A]$ and $[n_B, k_B]$ codes.

When a and b are vectors in codes A and B, respectively, it is interesting to flatten their outer product. We have

$$\begin{aligned} s(a \otimes b) &= (a_1 b, a_2 b, \ldots, a_{n_A} b) \\ &= a_1 b_1, \ldots, a_1 b_{n_B}, a_2 b_1, \ldots, a_2 b_{n_B}, \\ &\quad \ldots, a_{n_A} b_1, \ldots, a_{n_A} b_{n_B}) \end{aligned}$$

which, in fact, is the *Krönecker product* of the vectors a and b. Recall that the Krönecker product of an $m \times n$ matrix $\mathbf{X}$ with a $p \times q$ matrix $\mathbf{Y}$ is the $mp \times nq$ matrix $\mathbf{X} \otimes \mathbf{Y}$ given in block form as

$$\begin{bmatrix} [\mathbf{X}]_{1,1}\mathbf{Y} & [\mathbf{X}]_{1,2}\mathbf{Y} & \cdots & [\mathbf{X}]_{1,n}\mathbf{Y} \\ [\mathbf{X}]_{2,1}\mathbf{Y} & [\mathbf{X}]_{2,2}\mathbf{Y} & \cdots & [\mathbf{X}]_{2,n}\mathbf{Y} \\ \vdots & \vdots & \ddots & \vdots \\ [\mathbf{X}]_{m,1}\mathbf{Y} & [\mathbf{X}]_{m,2}\mathbf{Y} & \cdots & [\mathbf{X}]_{m,n}\mathbf{Y} \end{bmatrix}.$$

Note that the symbol $\otimes$ has been burdened to designate the direct product of codes, the outer product of vectors, and now the Krönecker product of matrices. When a and b are vectors, the interpretation of $a \otimes b$ either as an outer product or as a Krönecker product must be made clear.

Observe that the rows of the Krönecker product of $\mathbf{X}$ and $\mathbf{Y}$ are precisely the Krönecker products of all possible pairs of rows, one drawn from $\mathbf{X}$, the other from $\mathbf{Y}$. In light of our observation that the set of all possible outer products $\{g_i^A \otimes g_j^B : 1 \le i \le k_A, 1 \le j \le k_B\}$ of the rows of $\mathbf{G}_A$ and $\mathbf{G}_B$ is a basis for $A \otimes B$, it follows that "flattened code" $s(A \otimes B)$ is generated by the Krönecker product $\mathbf{G}_A \otimes \mathbf{G}_B$. For example, the "flattened" $[9, 4, 4]$ product code from Section 2.1 has generator matrix

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}$$

3. ITERATED PRODUCT CODES AND THE ELIAS CONSTRUCTION

The product construction may be *iterated*, that is, applied to more than two codes. For example, given linear codes A, B, and C of length n_A, n_B, and n_C, respectively, and with generator matrices $\mathbf{G}_A$, $\mathbf{G}_B$, and $\mathbf{G}_C$, respectively, one can define a code $A \otimes B \otimes C$ of length $n_A n_B n_C$. Although one might visualize such codes in terms of structures of higher order than matrices, it is often easier to deal with the "flattened" versions of such codes, by defining $A \otimes B \otimes C$ as the code generated by the Krönecker product $\mathbf{G}_A \otimes \mathbf{G}_B \otimes \mathbf{G}_C$. The Krönecker product is associative:

$$(\mathbf{G}_A \otimes \mathbf{G}_B) \otimes \mathbf{G}_C = \mathbf{G}_A \otimes (\mathbf{G}_B \otimes \mathbf{G}_C)$$

Thus the order in which the Krönecker products are formed does not affect the final outcome. The Krönecker product is *not* commutative, however, so the code generated by $\mathbf{G}_A \otimes \mathbf{G}_B \otimes \mathbf{G}_C$ is not, in general, equal to that generated by $\mathbf{G}_B \otimes \mathbf{G}_A \otimes \mathbf{G}_C$, although they are equivalent.

If $A_1, \ldots, A_L$ is a sequence of codes, where A_i is a linear $[n_i, k_i, d_i]$ code over $\mathbb{F}$, then it follows directly from Theorem 1 that $A_1 \otimes \cdots \otimes A_L$ is a linear $[n_1 n_2 \ldots n_L, k_1\, k_2 \ldots k_L, d_1 d_2 \ldots d_L]$ code.

For example, in his 1954 paper [1], Elias constructed a family of binary codes, each member of which is the direct product of a sequence of extended Hamming codes. Recall that for $m \ge 2$, the extended binary Hamming code H_m of length 2^m is a $[2^m, 2^m - m - 1, 4]$ code. Let $C_{m,L} = H_m \otimes H_{m+1} \otimes \cdots \otimes H_{m+L-1}$. Then $C_{m,L}$ has rate

$$R(m, L) = \prod_{i=m}^{m+L-1} (1 - (i+1)2^{-i})$$

It can be shown that, even when L approaches infinity, the rate $R(m, \infty)$ approaches a nonzero limit. Table 1 tabulates the numerical value of $R(m, \infty)$ for some small values of m. The key point is that the Elias product codes all have finite nonzero rate.

Assuming transmission through a binary symmetric channel with crossover probability p, Elias considered a straightforward strategy for decoding $C_{m,L}$. In the first stage of decoding, decoders for H_m are employed, corresponding to the "rows" of the product code. In the second stage, decoders for H_{m+1} are employed, corresponding to the "columns" of the product code, and so on.

A decoder for the extended Hamming code will successfully correct any single error in a given block. If an even number of errors occur, the Hamming decoder makes no changes to the block. If an odd number of errors greater than 1 occurs, the Hamming decoder will map the received word to a codeword by flipping some bit. This may increase the number of errors (if the flipped bit was correct), or decrease the number of errors (if the flipped bit was in error). Although the exact relationship between the expected number of output errors per block, e_o, and the expected number of input errors per block, e_i, is complicated, it can be shown that if e_i is sufficiently small, then the Hamming decoder for H_m actually *reduces* errors: $e_o < e_i$. In particular, as shown in [2, Sect. 14.83], for large values of m, if $e_i < 2.1779$, then $e_o < e_i$, and if $e_i < 0.6246$, then $e_o < e_i/2$.

Although the ith-stage Hamming decoder introduces statistical dependency among the symbols *within* the decoded word, since each symbol supplied to the $(i+1)$th-stage Hamming decoder is drawn from the output of an independent decoder at stage i, decoders at all stages are supplied with words in which errors in the bits are statistically independent. Provided that at each stage of

Table 1. Limiting Rates for Elias Product Codes

m	2	3	4	5	6	7	8	9
$R(m, \infty)$	0.054	0.215	0.431	0.627	0.772	0.866	0.924	0.958

decoding $e_i < 0.6246$, the expected number of errors per block supplied to the next decoder stage will also satisfy $e_i < 0.6246$, and in fact the number of errors per block will approach zero as the number of stages approaches infinity. Berlekamp [2, Sect. 14.84], provides a table of "threshold values" for the channel crossover probability p that will result in a chain of decreasing error probability at successive decoding stages. When $m = 3$, for example, the threshold is 0.0828, indicating that if $p < 0.0828$, then the family of codes $C_{3,L}$ can be decoded at arbitrarily small error rates when L is large enough. For $m = 4$, the threshold is 0.0402, and for $m = 5$, the threshold is 0.0192.

According to Table 1, $R(3, \infty) = 0.215$. A code operating at the Shannon limit on a binary symmetric channel would be able to tolerate a crossover probability $p = \mathcal{H}^{-1}(1 - 0.215) = 0.234$, while still providing error-free transmission, a value considerably larger than the threshold value 0.0828. [Here $\mathcal{H}$ denotes the binary entropy function $\mathcal{H}(p) := -p \log_2 p - (1-p) \log_2(1-p)$.] Similar conclusions hold for other values of m. Thus, one concludes that the Elias product codes are not capacity-achieving. Nevertheless, they were the first explicitly known family of codes to achieve asymptotically zero error probability at positive code rates. While certain families of irregular low-density parity-check codes (see the article LOW-DENSITY PARITY-CHECK CODES in this encyclopedia) are now known to contain codes with better threshold performance under message-passing decoding than the Elias product codes, it is remarkable indeed that the underlying principles — the use of simple constituent codes with intercommunicating decoders — had already been discovered by Elias as early as 1954.

4. ITERATIVE DECODING OF PRODUCT CODES

The Elias scheme for decoding product codes is strictly sequential. Decoders from one stage pass decoding results on to the next stage, without feedback. In crossword-puzzle terms, this decoding strategy is analogous to looking at the "across" clues, filling in the letters as well as possible, and then looking at the "down" clues, filling in any remaining letters, and then stopping. This leads to a tractable analysis for the decoder; however, as every aficionado of crossword puzzles knows, a far more effective decoding strategy is to alternate or *iterate* between the "across" and "down" clues, since solutions to the "down" clues provide information that are helpful in solving the "across" clues, and vice versa. While the statistical dependence between these interacting decoders makes an exact analysis of the decoder very difficult, simulations of decoder performance show that such iterative decoding can be extremely effective. In fact, it is precisely this sort of iterative decoding that underlies many of the most powerful practical coding schemes known (see, for example, the articles TURBO CODES and LOW-DENSITY PARITY-CHECK CODES in this encyclopedia).

To explain the iterative decoding of product codes, it is helpful to describe these codes in graph-theoretic terms, an approach pioneered by R. M. Tanner [4]. Figure 2 shows a *Tanner graph* for the direct product of codes of length n_A and n_B. A Tanner graph is a bipartite graph containing

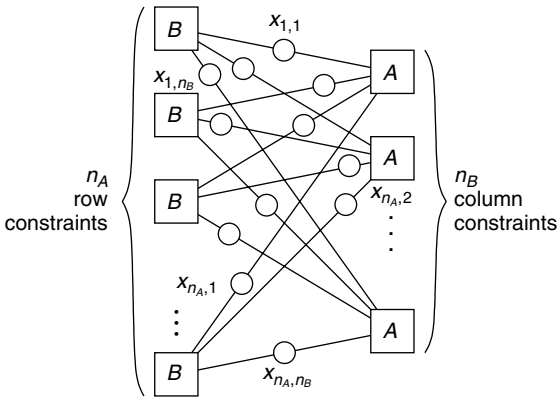


Figure 2. A Tanner graph for the direct product of $[n_A, k_A]$ and $[n_B, k_B]$ codes. The codeword symbol appearing in row i, column j is denoted $x_{i,j}$.

symbol nodes (shown as circles in Fig. 2) and check nodes (shown as squares in Fig. 2). Two nodes are *adjacent* if they are connected by an edge in the graph. In true bipartite fashion, symbol nodes can be adjacent only to check nodes (not to other symbol nodes), and similarly for check nodes.

The symbol nodes represent codeword symbols; in the binary case, each symbol node represents a bit. The purpose of the check nodes is to define the set of *valid configurations* or codewords of the code. They do this locally: each check node imposes a constraint only on the adjacent symbol nodes. While many global configurations may satisfy one or more local configurations, only those global configurations that satisfy *all* the local constraints are deemed to be globally valid configurations.

An example should make this clear. In the product code $A \otimes B$, the symbols in each column of a valid codeword must be a codeword of A and the symbols of each row of a valid codeword must be a codeword of B. Thus there are naturally n_B column constraints (constraining adjacent symbols to codewords of A) and n_A row constraints (constraining adjacent symbols to codewords of B), giving $n_A + n_B$ check nodes in the Tanner graph, as shown in Fig. 2. Each codeword symbol participates in exactly one row and one column, and hence each variable node is adjacent to exactly one column check node and one row check node.

Iterative decoding of this code may be interpreted as a process of *message passing* on the edges of the Tanner graph. We provide here only an intuitive description of the decoding procedure, and refer the reader to, for example, Ref. 5 for a more precise treatment.

The "messages" passed by the algorithm can be interpreted as probabilities or "beliefs" concerning the value of the corresponding symbol node. For example, the beliefs are often expressed as loglikelihood ratios (LLRs), defined as $\lambda(x) = \log(P[x = 0]/P[x = 1])$. When $\lambda(x)$ is large and positive, this indicates a strong belief that x has value 0; when $\lambda(x)$ is large and negative, this indicates a strong belief that x has value 1. A $\lambda(x)$ value that is close to zero indicates that $P[x = 0]$ and $P[x = 1]$ are nearly equal. The initial beliefs are determined by the received word.

Decoding proceeds iteratively. Each symbol node x transmits $\lambda(x)$ to, say, the adjacent-row check nodes.

Each row check node receives n_B incoming messages, and produces n_B outgoing messages that *update* the $\lambda(x)$ values, taking into account the constraints imposed by the structure of B. This updating procedure is sometimes referred to as "soft-in/soft-out" decoding, since the inputs and outputs of the decoder are beliefs, not hard decisions about the values of the bits.

For example, suppose that B is a $[3, 2]$ SPC code, and that a particular row check node receives three messages: two of which represent strong beliefs that the symbols x_1 and x_2 have value 1, and one representing a relatively weaker belief that symbol x_3 is a 1. Since 111 is not a valid local configuration, the most likely explanation is that x_3 is in fact 0. Thus the decoder would update the beliefs, reinforcing the belief that x_1 and x_2 are 1, and suggesting that x_3 is in fact zero.

These updated beliefs are then passed to the column decoders. Each column decoder receives n_A incoming messages, and produces n_A outgoing messages that update the belief values, taking into account the constraints imposed by the structure of A.

These newly updated beliefs could then be passed back to the row decoders, and the whole decoding process would repeat. The process would normally halt when the updated beliefs suggest a valid configuration, or when some predetermined upper limit on the number of iterations is reached.

It is important to note that different row decoders operate independently of one another. In a hardware decoder implementation, it would be possible to operate these decoders fully in parallel. The same is, of course, true for the column decoders. The local soft-in/soft-out decoders are typically implemented using the forward/backward algorithm [5] operating on a trellis representation of the local code. Any codes for which an efficient trellis representation exists would therefore be candidate constituent codes for such a product code. Since the different decoders are interconnected by a simple row/column structure, message passing can be implemented in hardware in straightforward fashion.

The performance of practically decodable product codes can be quite good, with performance on an additive white Gaussian noise channel with BPSK modulation that is within $\sim$2 dB of the Shannon limit at blocklengths on the order of 4000 bits [6]. While Turbo codes and low-density parity-check codes can approach somewhat closer to the Shannon limit, the simplicity of product codes and the prospect of high-speed decoder implementations in hardware make product codes very suitable for a wide variety of practical applications.

BIOGRAPHY

Frank R. Kschischang is a Professor in the Department of Electrical and Computer Engineering at the University of Toronto, where he has been a faculty member since 1991. He received the B.A.Sc. degree with honors from the University of British Columbia in 1985, and the M.A.Sc. and Ph.D. degrees from the University of Toronto in 1988 and 1991, respectively, all in electrical engineering. From October 1997 to October 2000, Dr. Kschischang served as *IEEE Transactions on Information Theory* Associate Editor for Coding Theory. In April 1999, he received the Ontario Premier's Research Excellence Award, and in December 2000, he was named a Canada Research Chair. His research interests lie in the general area of channel coding techniques, particularly in iterative decoding of codes defined on graphs.

BIBLIOGRAPHY

1. P. Elias, Error-free coding, *IRE Trans. Inform. Theory* **PGIT-4**: 29–37 (1954); reprinted in E. R. Berlekamp, ed., *Key Papers in The Development of Coding Theory*, IEEE Press, New York, 1974, pp. 39–47.
2. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
3. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
4. R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **IT-27**: 533–547 (1981).
5. F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, Factor graphs and the sum-product algorithm, *IEEE Trans. Inform. Theory* **47**: 498–519 (2001).
6. R. M. Pyndiah, Near-optimum decoding of product codes: Block turbo codes, *IEEE Trans. Commun.* **46**: 1003–1010 (Aug. 1998).

PROPAGATION MODELS FOR INDOOR COMMUNICATIONS

F. Landstorfer
G. Woelfle
R. Hoppe
Institut Fuer Hochfrequenztechnik
University of Stuttgart
Stuttgart, Germany

1. INTRODUCTION

The performance of wireless communication systems depends in a fundamental way on the mobile radio channel. In contrast to wired channels that are stationary and easy to design, radio channels show a time-variant behavior which complicates their analysis. Inside buildings the transmission path between transmitter and receiver can vary from simple line-of-sight to one severely obstructed by walls and furniture. As a consequence, predicting the propagation characteristics between two antennas still belongs to the most important tasks for the design and installation of wireless indoor communication systems, ranging from low-bit-rate cordless telephone and cellular systems to high-bit-rate wireless local-area networks (WLANs).

This article reviews and discusses a variety of methods for modeling wave propagation in indoor scenarios. The basic requirements necessary for predicting path loss and other relevant parameters are discussed. Apart from well known and widely used propagation models, new

approaches with minimized computational complexity are also introduced.

1.1. The Mobile Indoor Radio Channel

The mobile radio channel concerning transmission within buildings is characterized by a multipath scenario as shown in Fig. 1. The signal from the transmitting antenna (usually only the downlink is considered as the principle of reciprocity applies) propagates along different paths to the antenna of the (mobile) receiver. In many cases there is no direct line of sight and the only paths connecting transmitter and the receiver penetrate several walls and are reflected, diffracted, and scattered at a number of different obstacles. Since the phases of the waves are randomly distributed, the superposition of these contributions causes constructive and destructive interference (i.e., small-scale fading), which leads to rapidly fluctuating signal levels over very short distances. Figure 2 illustrates this small-scale fading and the slower large-scale signal variation for an indoor radiocommunication system. While the small-scale fading is random, the large-scale variations occur as a result of fundamental changes of the propagation paths (e.g., larger distances, different obstacles). Typically, the local average of the received power is computed by averaging signal measurements over an interval of 10λ to 20λ, which corresponds to movements of the receiver of 1.5–3 m at a frequency of 2 GHz.

1.2. Radiowave Propagation within Buildings

With decreasing wavelength, that is, increasing frequency, wave propagation becomes more and more similar to the propagation of light. A radio ray is assumed to propagate essentially along a straight line and is influenced only by the given obstacles. For a criterion for this type of modeling to be successful, the wavelength should be much smaller than the extensions of the partitions of the building structure. At the frequencies used for wireless indoor communication networks, this criterion is sufficiently fulfilled.

The phenomena that influence radiowave propagation can generally be described by four basic mechanisms: penetration, reflection, diffraction, and scattering. For the practical usage of propagation models in real

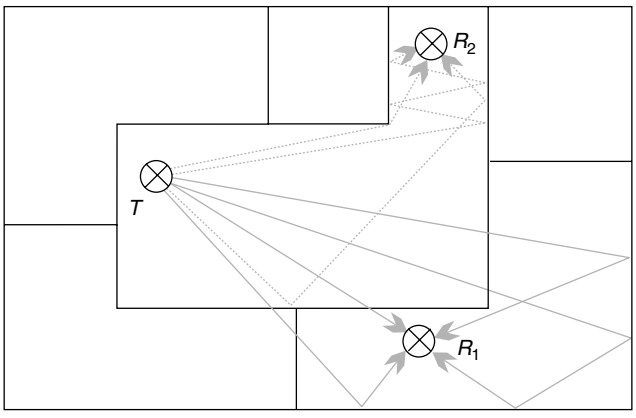


Figure 1. Multipath propagation within buildings.

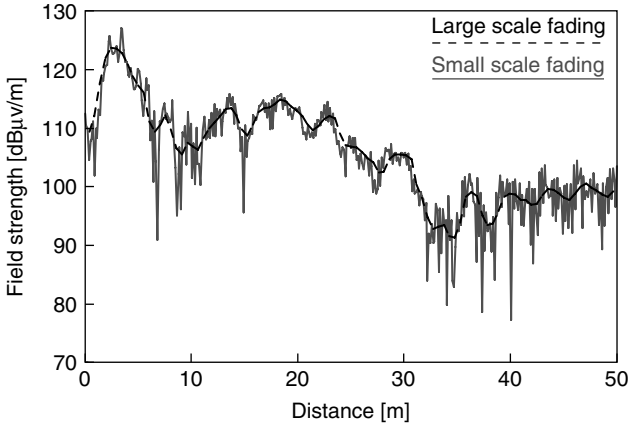


Figure 2. Large-scale and small-scale fading.

scenarios, these mechanisms must be introduced by approximations. This requires a multistage modeling process [1], introduced in the following paragraphs.

In the first step the propagation environment has to be digitized, leading to a database that describes the considered scenario in an adequate way. The second step includes the definition of mathematical approximations for the physical propagation phenomena. On the basis of the solutions for the basic problems, both empirical and deterministic approaches have been developed, forming the third modeling step.

The indoor radio channel differs considerably from that of an outdoor scenario [2,3]. The transmitter–receiver distance is shorter in order to compensate for the high attenuation along the path caused by internal walls and furniture and also because of the lower transmitter power. The short distance implies shorter delay of echoes and consequently a lower delay spread. The temporal variations of the channel are slower than those of mobile antennas moving with an automobile. Nevertheless, the conditions within buildings are time-variant even when transmitter and receiver are fixed; for instance, whether interior doors are open or closed can change the propagation scenario considerably. In general, the propagation within buildings is strongly influenced by the local environment, namely, the layout of the particular building under consideration and the construction materials used for walls, floors, and ceilings. According to the type of building and the corresponding layout, four different categories of indoor environments can be defined as listed in Table 1 [1,4].

As it is the case in outdoor scenarios, there are several important propagation parameters to be predicted. The path loss and the statistical characteristics of the received signal strength are most important for coverage planning applications. The wideband characteristics (delay spread, impulse response) and the time variation are essential for the evaluation of the system performance.

1.3. Properties of Materials

The buildings taken into account within a planning process have a wide variety of walls and obstacles that form the internal and external structure. Hard partitions and soft

Table 1. Different Categories of Indoor Environments [1]

Environment Category	Description	Typical Values for the Delay Spread (ns)
Corridor	Transmitter and receiver along the same corridor (LoS)	≤20
Dense	Small rooms; typically an office where each employee has his own room; mostly non-line-of-sight (NLoS) conditions	10–30
Open	Large rooms; typically an office where one room is shared by several employees; mostly line-of-sight (LoS) or obstructed-line-of-sight conditions (OLoS)	20–50
Large	Environments consisting of very large rooms; typically a factory hall, shopping center, or airport building; mostly LoS or OLoS conditions	50–80

partitions can be distinguished according to the magnitude of the penetration loss for electromagnetic waves [2]. While hard partitions are formed as bearing parts of the building structure (e.g., walls constructed from reinforced concrete or brick, thickness >10 cm), soft partitions may be moved and therefore show lower losses (e.g., plasterboard or plywood, thickness <10 cm). In order to get a more accurate modeling of wave propagation, a detailed description of the electrical properties for all building elements considered is necessary. However, in most cases the partitions are made out of several layers and are not homogenous. Nevertheless, a lot of researchers have collected databases for a great amount of different materials. While the physical behavior is described by the complex permittivity ε, the resulting partition losses as given in Table 2 are of more interest from a practical point of view.

The total loss experienced by electromagnetic waves when penetrating walls can be divided into two independent parts. The first part refers to the penetration of the wall surface and shows no explicit frequency dependence, whereas the second part belongs to the in depth penetration of the wall material. The magnitude of the latter depends on the wavelength and leads to higher losses with increasing frequency (see Table 2).

1.4. Databases for Buildings

The basis for any propagation model is a database that describes the propagation environment. For the purpose of propagation modeling, each building element should be categorized into classes (wall, floor, door, window, etc.) and specified by its coordinates and finally its material properties (thickness and electrical characteristics).

Modern planning tools [5,6] store the building data in a 3D-vector format including all walls, doors, and windows. Usually all elements inside the building are described in terms of plane elements; for example, every wall is represented by a plane and its extent and location is defined by its corners as indicated in Fig. 3. Through the use of

Table 2. Partition Losses of Different Construction Materials

Material Type	Frequency (MHz)	Loss (dB)	Ref.
Reinforced concrete	900	10	1
	1700	16	1
Brick	900	6	1
	1800	10	1
Plywood	900	1.5	1
	1800	2	1
Glass	900	3	1
	1800	4	1
Metal	815	26	2

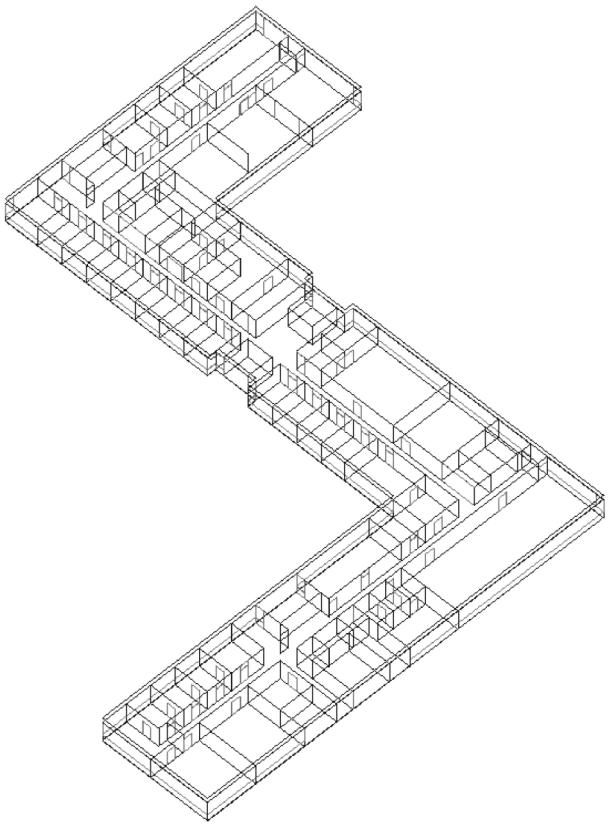


Figure 3. Example of an indoor database.

such building databases, which may be drawn or digitized utilizing standard graphical CAD (computer-aided design) software packages, wireless system designers are able to include accurate representations of building features. With respect to an efficient use it is often possible to import drawing interchange files (DXFs), a very common CAD data format in architecture. Figure 3 shows an example for a three dimensional building data base utilized within a commercial planning tool [6].

2. PROPAGATION MODELS

Propagation models focus on the prediction of the averaged received signal strength, as well as the variability of the signal strength in close vicinity of the particular location. This leads to the distinction between large-scale propagation models that predict the mean signal strength and small-scale propagation models that characterize the rapid fluctuations of the received signal strength over very short distances (a few wavelengths) or short time durations (in the order of seconds) [2].

Similar to this distinction, the propagation models presented here can be divided into three groups: empirical narrowband models, empirical wideband models, and deterministic ones. Empirical narrowband models are expressed in form of simple mathematical equations that give the path loss as an output result. The equations are obtained by fitting the model parameters to measurement results. The empirical wideband models allow the prediction of the wideband characteristics of the channel (e.g., impulse response, delay spread). Deterministic models simulate the propagation of radiowaves in a more physical way. These models predict both narrowband and wideband information of the mobile radio channel inside buildings. Additionally, directional channel properties such as angular spread are readily available, a fact that may be very important for the planning of future systems [7]. Propagation models for the prediction of field strength levels will generally provide only mean or median values, as the small-scale fading within indoor environments is adequately represented by Rayleigh distributions for NLoS (see Table 1) conditions and Rice distributions for clear LoS conditions with K factors up to 15 dB. The long-term fading, which describes the fluctuation of the mean value, can be approximated by a lognormal distribution with standard deviations between 2.7 and 5.3 dB [1].

2.1. Empirical Narrowband Models

Figure 4 shows the two basic approaches to the prediction of the field strength inside buildings. There are empirical models that analyze the direct path between the transmitter and the receiver. A calibration of these empirical models is mandatory, and their computation times are very small. All models of this type are based on the free-space propagation model.

2.1.1. Free-Space Propagation Model. This model is utilized to predict the received power if transmitter and receiver have a clear, unobstructed line-of-sight path. Generally this is not the case within indoor scenarios; nevertheless, this model gives a valuable insight into propagation modeling. The received power of an antenna at a distance d from the radiating transmitter antenna is calculated by the Friis equation:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2} \tag{1}$$

where P_t is the transmitted and P_r is the received power, G_t and G_r are the antenna gains of the transmitter (Tx) and receiver (Rx), respectively, already mention before equation and λ is the free space wavelength. The free space path loss L_{FS}, which represents the signal attenuation as a positive quantity measured in decibels (dB), is defined as difference (in dB) between the transmitted and the received power level:

$$L_{FS} = 10 \log \frac{P_t}{P_r} = 10 \log \left(\frac{4\pi d}{\lambda} \right)^2 - 10 \log G_t - 10 \log G_r \tag{2}$$

Obviously the path loss increases both with distance and with frequency at a rate of 20 dB/decade. With the

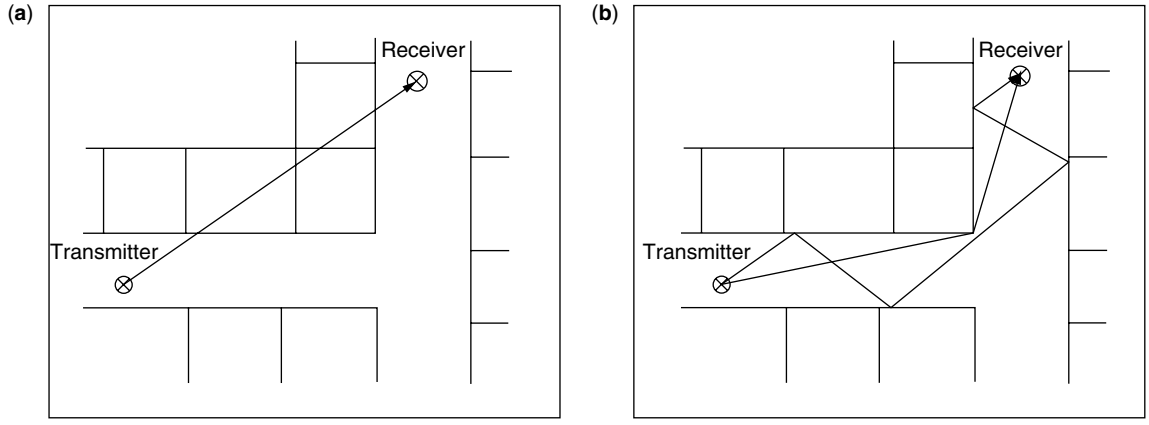


Figure 4. Basic approaches to the modeling of indoor propagation: (**a**) empirical models; (**b**) deterministic models.

introduction of the default loss L_0 (including the antenna gains) at distance d_0, Eq. (2) can be simplified to

$$L_{FS} = L_0 + 2 \cdot 10 \log\left(\frac{d}{d_0}\right) \tag{3}$$

2.1.2. One-Slope Model. The one-slope model analyses the building concerning distances between walls and penetration losses of the walls, but the individual positions of the walls and their material properties are not considered (see Fig. 5). Therefore this model computes the path loss L_{OS} (in dB) similar to the free space loss with adaptable power decay factor n and offset L_0.

$$L_{OS} = L_0 + n \cdot 10 \log\left(\frac{d}{d_0}\right) \tag{4}$$

The walls of the building are not taken into account by the one-slope model; thus no building database is required. With constant values for n and L_0 for every receiver location the prediction leads to path loss values increasing in concentric circles around the transmitter. Consequently, the prediction results are fairly inaccurate and suited only for a rough estimation. For line-of-sight the values of n are in the range between 1.4 and 1.8; in non-line-of-sight scenarios values up to 5.0 are possible [1,2].

2.1.3. Motley–Keenan Model. The model according to Motley and Keenan [8] computes the path loss L_{MK} based on the direct path between transmitter and receiver. In contrast to the one-slope model, in this model the exact locations of the walls, floors, and ceilings are considered. Additional factors for the attenuation of the direct path by partitions account for the shadowing effects.

$$L_{MK} = L_{FS} + k \cdot L_W \tag{5}$$

As shown in Fig. 6, parameter k describes the number of walls intersected by the direct path between transmitter and receiver. A uniform penetration loss L_W (in dB) of all partitions is taken for the computation (see Table 2); thus the individual material properties of the different walls are not considered. This uniform penetration loss as well

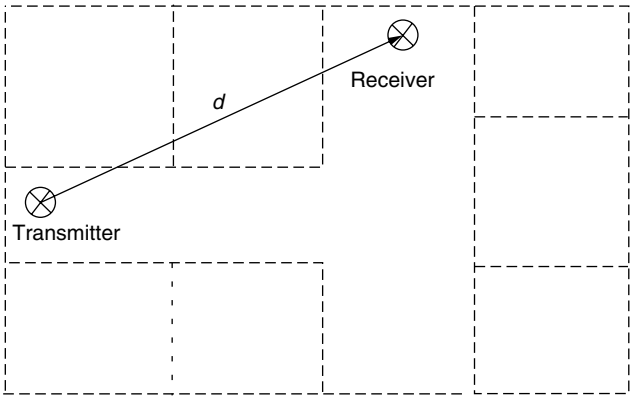


Figure 5. Principle of the one-slope model.

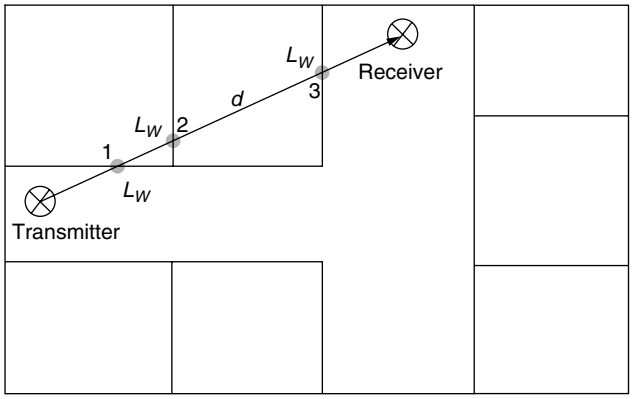


Figure 6. Principle of the Motley–Keenan model.

as the default loss L_0 which is part of the free-space path loss L_{FS} can be calibrated by measurements.

2.1.4. Multiwall Model. The multiwall model [1] gives the path loss L_{MW} as the free-space loss added to losses introduced by the walls and floors penetrated by the direct path between transmitter and receiver (see Fig. 6). In contrast to the Motley–Keenan model, for the multiwall model, individual penetration losses of the walls (depending on their electrical characteristics) are considered for the prediction of the path loss. It has been observed that the total floor loss is a nonlinear function of the number of floors penetrated. This effect is taken into account by introducing an empirical factor b. Hence the multiwall model can be expressed as follows:

$$L_{MW} = L_{FS} + \sum_{i=1}^{N} k_{Wi} \cdot L_{Wi} + k_f^{[(k_f+2)/(k_f+1)-b]} \cdot L_f \tag{6}$$

where k_{Wi} represents the number of penetrated walls of type i and k_f the number of penetrated floors. The parameter L_{Wi} describes the penetration loss of wall type i and L_f the loss between adjacent floors, and N denotes the number of different wall types.

The default loss L_0 as a part of the free-space path loss L_{FS} may be calibrated according to measurement results by evaluating a multiple linear regression technique. It is important to note that the loss factors in the Eq. (6) do not represent physical wall losses but model coefficients that are optimized by the measured path-loss data. Consequently, the loss factors implicitly include the effect of furniture. However, hard partitions such as concrete walls are overemphasized, which leads to pessimistic values of predicted field strength behind these elements as indicated in Fig. 7. Additionally, waveguiding effects are not taken into account within this model; thus its accuracy is only moderate. Nevertheless, this more refined empirical narrowband model has a low dependency on database accuracy and, because of the simple approach, a very short computation time (in the order of seconds).

Within the European cooperation of COST 231, measurements in different environments from several institutions have been evaluated [1] in order to optimize the coefficients of the empirical narrowband models (as given

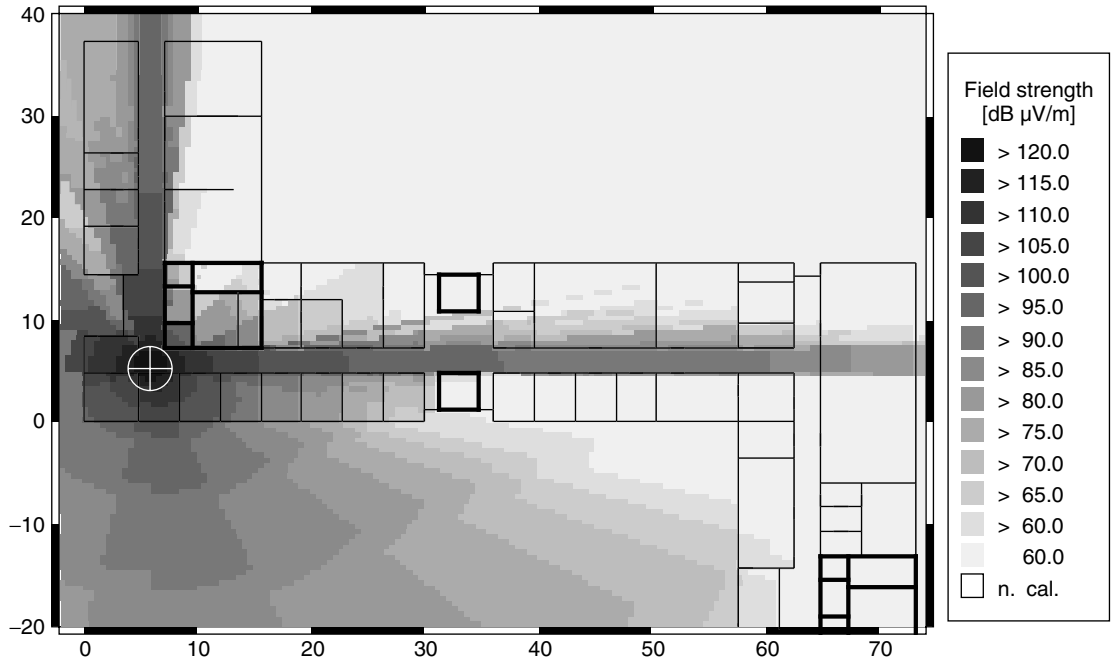


Figure 7. Prediction of the field strength by evaluating the multiwall model at 1800 MHz for a transmitter with 100 mW and omnidirectional Tx-antenna pattern.

Table 3. Investigated Coefficients at 1800 MHz for Empirical Narrowband Models [1]

	One-Slope Model		Multiwall Model			
Environment	L_0 (dB)	n	L_{W1} (dB)	L_{W2} (dB)	L_f (dB)	b
Corridor	39.2	1.4	3.4	6.9	18.3	0.46
Dense						
One floor	33.3	4.0	3.4	6.9	18.3	0.46
Two floors	21.9	5.2				
Multifloor	44.9	5.4				
Open	42.7	1.9	3.4	6.9	18.3	0.46
Large	37.5	2.0	3.4	6.9	18.3	0.46

in Table 3). For the multiwall model two different wall types have been taken into account corresponding to the previous distinction between hard and soft partitions. The coefficients of the multiwall model have been optimized for the category "dense." However, these values can also be utilized within other environments, leading to results close to the free-space model. According to measurements at 900 MHz, the values of the multiwall model should be reduced by 1.5 dB for the soft partition loss and by 3.5 dB for the floor loss. Concerning the one-slope model, the default loss L_0 should be reduced about 10 dB, while the values for the decay index remain constant.

2.2. Empirical Wideband Models

Beyond the planning and deployment of wireless communication networks, propagation models are also utilized for studying and evaluating new radio systems. The propagation models provide a valuable input for so-called channel models. These models reproduce the behavior of the radio channel in order to estimate the performance and the capacity of different system implementations concerning access, modulation, or coding schemes.

Empirical wideband models provide average impulse responses and power delay profiles (PDP) for this purpose. This kind of data can be derived from wideband measurements (e.g., channel sounder or network analyzer) as well as from deterministic propagation models. On the basis of wideband measurements, the delay spread values as given in Table 1 have been determined [1], indicating lowest values in dense environments and increasing values in open and large environments.

Impulse response models are usually defined as tapped delay lines, including a limited number of paths with individual amplitude and delay (both referred to the dominant path). The ITU channel models [9] as given in Table 4 contain six different paths and describe typical impulse responses for large office buildings with an open layout and a Tx-Rx separation less than 100 m. While the tap delay-line models represent typical impulse responses for unlimited bandwidth, the power delay

Table 4. Impulse Response Model for Indoor Office Environments [9]

	Channel A			Channel B		
	Delay Spread = 35 ns			Delay Spread = 100 ns		
Tap #	Delay (ns)	Loss (dB)	Doppler Spectrum	Delay (ns)	Loss (dB)	Doppler Spectrum
1	0	0.0	Flat	0	0.0	Flat
2	50	−3.0	Flat	100	−3.6	Flat
3	110	−10.0	Flat	200	−7.2	Flat
4	170	−18.0	Flat	300	−10.8	Flat
5	290	−26.0	Flat	500	−18.0	Flat
6	310	−32.0	Flat	700	−25.2	Flat

profiles characterize the radio channel with respect to the limited bandwidth of a specific radio system. Typical averaged power delay profiles within buildings show a logarithmic (for LoS and OLoS conditions) to linear (for NLoS conditions) decay on dB scale depending on the specific environment [1].

Because of the slow movements of indoor terminals (in most cases they are more portable than mobile) the Doppler spectrum has maximal values of about 10 Hz for frequencies utilized within personal communication systems (about 2 GHz). For simplicity a flat spectrum is often considered (see Table 4).

2.3. Deterministic Models

Deterministic models utilize physical phenomena in order to describe the propagation of radiowaves. Here, the effect of the actual environment is taken into account more accurately than within empirical models.

2.3.1. Basic Mechanisms. The mobile radio channel in indoor environments is characterized by multipath propagation (see Fig. 1). Dominant propagation phenomena in these scenarios are reflection, transmission, diffraction, and scattering. Because of the multiple reflections, waveguiding in corridors can be observed. Deterministic propagation models are generally based on ray optical techniques. With such an approach it is possible to consider the abovementioned effects as well as the multipath situation within the propagation model (as indicated in Fig. 4). The ray optical models determine all relevant rays between the transmitter and the receiver. Calibration of the deterministic models is not necessary because the predicted values are computed with the Fresnel equations for reflection and transmission and with the universal theory of diffraction (UTD) for diffracted rays [10]. Alternatively, there are empirical equations available for the calculation of the reflection, diffraction, and transmission loss [11].

2.3.1.1. Reflection and Transmission. When an electromagnetic wave propagating in one medium impinges on another medium with different electrical properties, the wave is partially reflected and partially penetrates the medium. Although this phenomenon is valid for the incidence on a perfect dielectric, all the incident energy is reflected back to the first medium if the second medium is a perfect conductor. If there is a smooth boundary between the two materials, the impinging wave is reflected specularly. Concerning the penetration of a dielectric plate, the direction of the penetrated radiowave corresponds to the incident direction if there is the same medium in front of and behind the dielectric plate. The electric field intensities of the reflected and transmitted waves are related to the incident wave through a reflection coefficient. The most common mathematical description of the reflection is the Fresnel reflection coefficient, which is valid for an infinite boundary between two media. The coefficients for reflection and transmission are a function of the material properties, and generally depend on the polarization and the angle of incidence of the propagating wave [10].

2.3.1.2. Diffraction. The diffraction process in ray modeling is the *propagation phenomenon*, which explains the transition from the lit region to the shadowed regions behind obstacles. Although the electrical field strength decreases as a receiver moves deeper into the shadowed region, the diffraction field still exists and often has enough strength to guarantee a sufficient signal. According to the principle of Huygens, the diffraction field is the vector sum of the electric field components of all secondary wavelets in the space around the obstacle. Diffraction by a single wedge can be solved in various ways: empirical formulas, perfectly absorbing wedge, geometric theory of diffraction (GTD) or universal theory of diffraction (UTD) [10]. The advantages and disadvantages of using either of these formulations is difficult to address since it may not be independent of the investigated environments. Indeed, reasonable results are possible with each formulation. The various expressions differ mainly in the approximations being made on the surface boundaries of the wedge considered. However, diffraction around a corner is commonly modeled using the heuristic UTD formulas since they behave well in the lit/shadow transition region and account for the polarization of the incident wave as well as the wedge material. Generally, the wedge diffraction coefficient is inversely proportional to the square root of the frequency, specifically, the coefficient decreases with increasing frequency. Therefore the effect of diffraction can be neglected for millimeter waves ($f > 30$ GHz).

2.3.1.3. Scattering. Rough surfaces and finite surfaces (i.e., surfaces with small extensions in comparison to the wavelength) scatter the incident energy in all directions according to a radiation pattern which depends on the roughness and size of the surface or volume. The dispersion of energy through scattering means a decrease of the energy reflected in specular direction, which can be

taken into account by reducing the reflection coefficient. The consideration of the true dispersion of radio energy in various directions is much more difficult and may be described by radar cross-sectional models. Surface roughness is often tested using the Rayleigh criterion, which defines a critical height of surface protuberances for a given angle of incidence [10].

2.3.2. Path Finding. For the determination of valid rays between transmitter and receiver, which is the most time-consuming part of the ray optical approach, two different principles can be utilized: ray launching and ray tracing.

2.3.2.1. Ray Launching. This approach launches rays in discrete angular increments from the transmitter and determines their path through the database (as indicated in Fig. 8). If there is an intersection between the ray and a wall, the ray is split into a penetrating and a reflecting part and both rays are further launched independently from each other. When the ray impinges on a wedge, new rays are launched on the diffraction cone, which leads to multiple rays. Every time when a ray intersects the prediction plane, the field strength of this ray is added to the already computed field strength at the specific receiver point. The consideration of the path ends if the number of interactions is higher than a given number or if the field strength at the end of the ray is smaller than a given threshold. In order to keep the resolution of the rays independent of the distance to the transmitter, a splitting algorithm may be introduced (ray splitting).

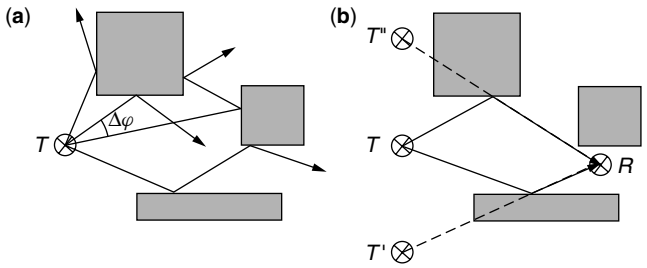


Figure 8. Algorithms for path finding: (**a**) ray launching; (**b**) ray tracing.

2.3.2.2. Ray Tracing (Ray Imaging). The ray tracing algorithm determines valid rays between the transmitter and a given receiver point (as indicated in Fig. 8). For coverage predictions the receiver has to be assigned subsequently to all pixels in the prediction area. Obviously the computation time increases linearly with the number of receiver points and shows an exponential dependence on the number of walls and the number of interactions, respectively. In comparison to the ray launching there is a higher computational effort, but on the other hand, ray tracing obtains a constant resolution and a high degree of accuracy.

Figure 9 shows a prediction with a rigorous 3D ray tracing for the situation already presented in Fig. 7. The differences between empirical and ray optical predictions are obvious especially the waveguiding in corridors due to multiple reflections, the coupling of the waves into the rooms and the diffraction around corners are responsible for the high accuracy of the ray optical models.

2.3.3. Advanced Ray Optical Model Based on Database Preprocessing. The main disadvantage of the deterministic prediction models is their excessive computation time (in the order of hours). Different authors presented ideas to accelerate the path finding, and some of the methods lead to considerable acceleration factors. A further disadvantage of the ray optical propagation models is the abovementioned dependence on the accuracy of the database, in which even small errors in the positions and

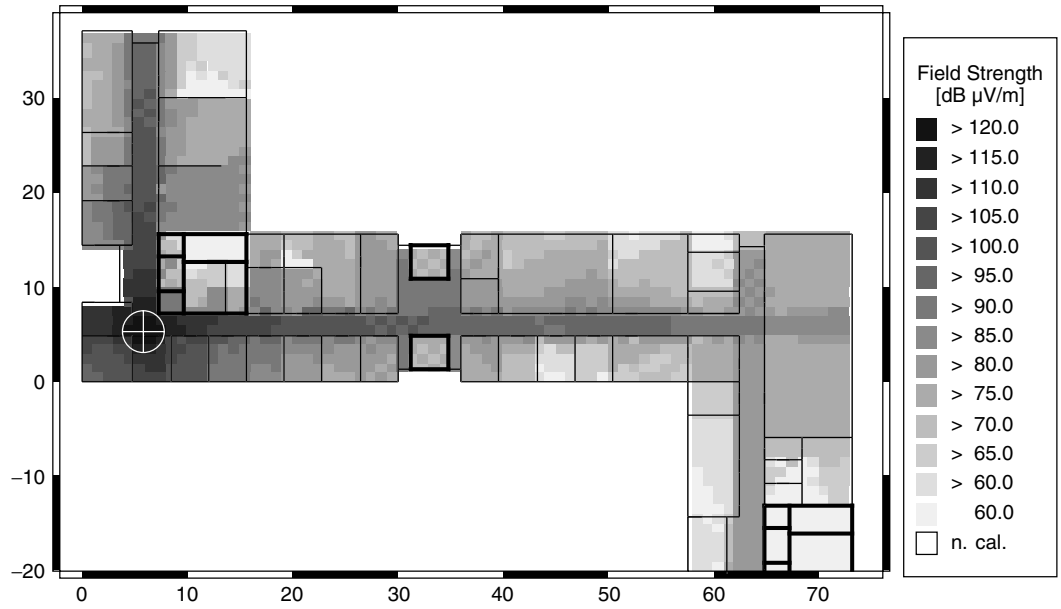


Figure 9. Prediction of the field strength by evaluating the 3D ray tracing at 1800 MHz for a transmitter with 100 mW and omnidirectional Tx-antenna pattern.

materials of the walls or missing parts (e.g., furniture) influence the predicted results. The advanced model [11] presented in this section neglects these disadvantages and is therefore well suited for the planning and deployment of indoor wireless communication networks.

One major application of propagation models is to evaluate the degree of coverage that can be achieved in a radio cell depending on the position of the base station. While the database of the building in question remains the same and only the position of the transmitter changes, the overwhelming part of the different rays remains unchanged; only the rays between the transmitting antenna and primary obstacles or receiving points in line of sight change.

This is the basis for "intelligent database preprocessing." In a first step the walls of the building (or other obstacles) are divided into tiles (reflections and penetrations) and the edges (diffractions) into horizontal and vertical segments. After this, the visibility conditions between these different elements (possible rays) are determined and stored in a file. Figure 10 shows the visibility relation between a central tile and a receiving point. The result of this preprocessing can be represented in the shape of a "visibility tree" as shown in Fig. 11. For a different transmitter location only the uppermost branches in this tree must be computed again. Consequently, all branches below the first interaction layer have to be computed only once, which can be done prior to optimizing the location of the transmitter. The remaining computation time after the preprocessing is many orders of magnitude lower than that needed for the conventional analysis without preprocessing. As a consequence, 3D deterministic models, with their supreme accuracy, can be utilized for all practical applications with computation times in the order of those found with empirical models.

When analyzing rays that contribute to the field strength at the receiving point of a typical indoor situation (as given in Fig. 1), it is obvious that a number of different rays reach the receiver after passing the same sequence of rooms and penetrating the same walls. These rays can be summarized into several dominant paths, each of them characterizing the propagation of a bundle of waves [12]. There is generally more than one dominant path between transmitter and receiver. Figure 12 shows dominant paths for the scenario depicted in Fig. 1. The dominant paths can be deduced using simple algorithms that consider the arrangement of the rooms within the building relative to the transmitter and the receiver [12].

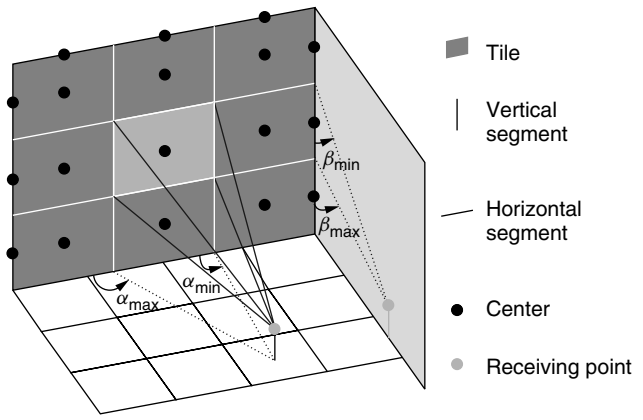


Figure 10. Tiles and segments of a wall with visibility relation indicated.

3. PLANNING TOOLS

With the increasing computational and visualization capabilities of computers, new methods for predicting radio signal coverage have been established, which are based on site specific propagation models and building databases as described in Section 1.4. As planning tools become prevalent, wireless system designers will be able to design and deploy indoor wireless networks for covering the buildings adequately without performing extensive radio measurements. Generally, indoor planning tools support graphical user interfaces in order to generate and visualize building databases, including

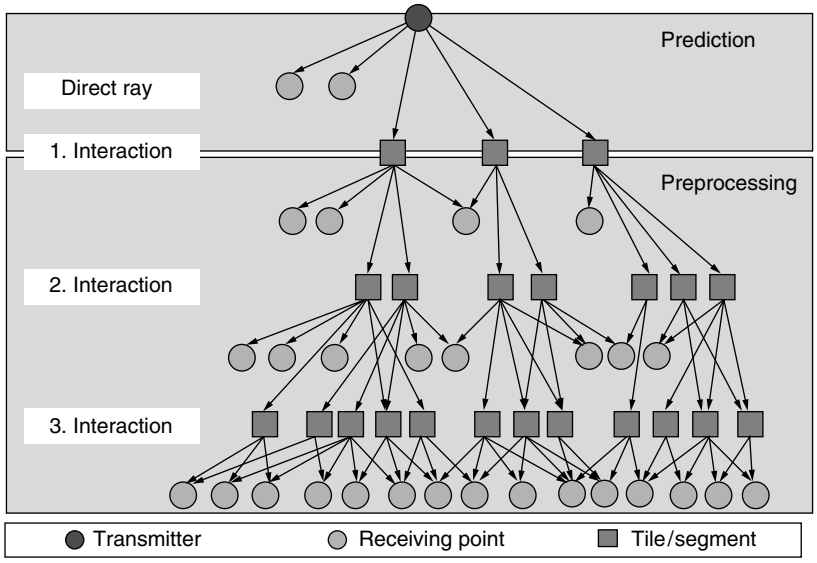


Figure 11. Tree structure of the visibility relations.

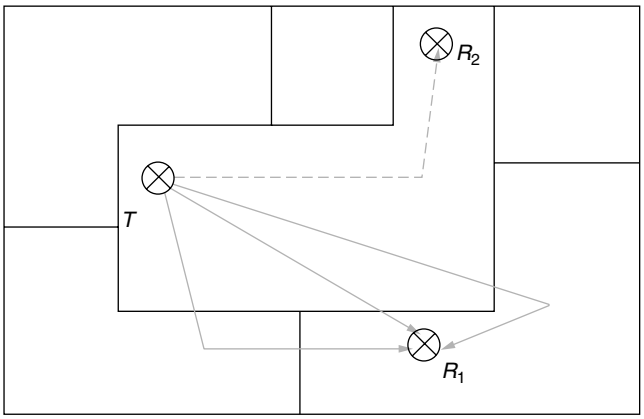


Figure 12. Dominant paths concept.

the material properties as well as to define the characteristics of transmitters and receivers (e.g., antenna patterns). Beyond this, most planning tools provide deterministic and empirical propagation models (as presented in Section 2) for predicting the large-scale path loss in a wide range of building structures and capabilities to visualize and analyze the corresponding results [5,6].

BIOGRAPHIES

Friedrich M. Landstorfer received a Dipl.-Ing. degree in 1964 and a Dr.-Ing. degree in 1967 in electrical engineering from the Technical University of Munich, Germany. In 1971 he became lecturer at the same institution and professor in 1976. In 1986, he moved to Stuttgart to become professor and head of the RF-Institute at the University of Stuttgart, Germany. His research interests include antennas, microwaves, electromagnetic theory, wave propagation in connection with mobile communications, navigation, and electromagnetic compatibility. He received the award of the NTG (Nachrichtentechnische Gesellschaft in VDE, now ITG) in 1977 for the optimization of wire antennas and was chairman of the 21st European Microwave Conference in Stuttgart in 1991. Professor Landstorfer is a member of URSI, was awarded honorary professor of Jiaotong University in Chengdu, China, in 1993, and became fellow of the IEEE in 1995.

Gerd Woelfle received his Dipl.-Ing. and Ph.D. degrees in electrical engineering from the University of Stuttgart, Germany, in 1994 and 1999, respectively. In his Ph.D. thesis he analyzed the indoor multipath propagation channel and developed adaptive models for indoor propagation. In 1999, he joined AWE Communications as a R&D Engineer, where he has been working on propagation models for wireless communication networks. Dr. Woelfle has received several best paper awards on wireless communication conferences. His areas of interest are empirical and deterministic propagation models, network planning tools, and channel modeling. Since 1999, he has been a lecturer for mobile communications at the University of Stuttgart.

Reiner A. Hoppe received his Dipl.-Ing. degree in electrical engineering in 1997 from the University of Stuttgart, Germany. Since this time, he has been a research scientist at the Institute of Radio Frequency Technology (University of Stuttgart), where he is working towards his Ph.D. degree. His research interests include wave propagation modeling and radio network planning especially within urban and indoor scenarios.

BIBLIOGRAPHY

1. E. Damosso, ed., *Digital Mobile Radio towards Future Generation Systems*, Final Report of the COST Action 231, Bruxelles: European Commission, 1998.
2. T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, Upper Saddle River, NJ, 1996.
3. D. Parsons, *The Mobile Radio Propagation Channel*, Pentech Press, London, 1992.
4. D. Molkdar, Review on radio propagation into and within buildings, *IEE Proc. H* (*Microwaves, Antennas and Propagation*) **138**(1): 61–73 (Feb. 1991).
5. S. J. Fortune, *WiSE—a Wireless System Engineering Tool* (online), *http://www.bell-labs.com/innovate98/wireless/wiseindex.html*, July 4, 2001.
6. AWE Communications, *WinProp—Software for Radio Network Planning Within Terrain, Urban and Indoor Scenarios*, (online) *http://www.awe-communications.com* (July 4, 2001).
7. L. M. Correia, ed., *Wireless Flexible Personalised Communications*, Final Report of the COST Action 259, Wiley, Chichester, UK, 2001.
8. A. J. Motley and J. M. Keenan, Personal communication radio coverage in buildings at 900 MHz and 1700 MHz, *IEE Electron. Lett.* **24**(12): 763–764 (June 1988).
9. ITU-R Recommendation M.1225, *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000*, International Telecommunication Union, 1997, pp. 24–28.
10. C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.
11. G. Woelfle, R. Hoppe, and F. M. Landstorfer, A fast and enhanced ray optical propagation model for indoor and urban scenarios, based on an intelligent preprocessing of the database, *Proc. 10th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications* (*PIMRC*), F5-3, Osaka, Japan, Sept. 1999.
12. G. Woelfle and F. M. Landstorfer, Dominant paths for the field strength prediction, *Proc. 48th IEEE Int. Conf. Vehicular Technology* (*VTC*), Ottawa, May 1998, pp. 552–556.

PULSE AMPLITUDE MODULATION

BJØRN A. BJERKE
Qualcomm, Inc.
Concord, Massachusetts

1. INTRODUCTION

Pulse amplitude modulation (PAM) is a modulation method used in digital communication systems to facilitate

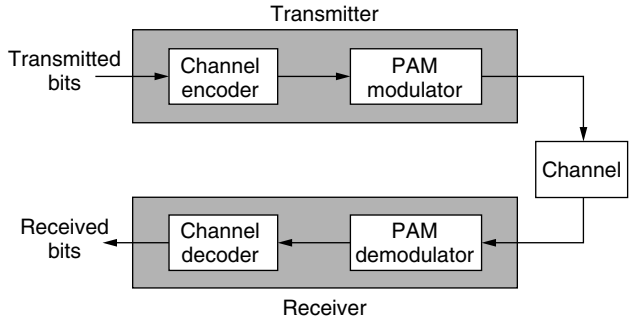


Figure 1. The basic elements of a digital communication system that uses pulse amplitude modulation.

transmission of digital information from a source (the transmitter) to a destination (the receiver). In such systems, the digital information is represented by a sequence of binary digits, or bits. Figure 1 shows the basic elements of a communication system that uses PAM. The transmitter consists of two blocks, namely, a channel encoder and a digital modulator. The channel encoder adds redundancy in a controlled manner to the transmitted bits in order to enable detection and possibly also correction of bit errors. Such errors may occur as a result of disturbances in the transmission system.

The output sequence from the encoder is passed to the PAM modulator, which transforms the discrete-time sequence into a signal that conforms to the limitations of the transmission channel. The channel is the physical medium through which transmissions occur. Most channels are analog in nature, so the modulator outputs analog waveforms that correspond to the particular coded sequence. The physical channel can, for example, be a cable, an optical fiber, or a wireless radio channel. In each of these channels, the signal is affected by disturbances such as thermal noise and interference from various sources. These disturbances corrupt the signal in a number of ways that ultimately may lead to bit errors at the receiver.

The role of the PAM demodulator is to recover the sequence of coded information bits from the corrupted received signal. In so doing, the demodulator often must compensate for the various channel disturbances. Finally, the coded sequence is passed to the channel decoder. Using knowledge of the channel code and the redundancy contained in the received sequence, the decoder attempts to reconstruct the original information bit sequence with as few bit errors as possible.

In this article, we shall concentrate on the modulator/demodulator pair, commonly referred to as the *modem*. The remainder of the article is organized into three sections, each dealing with different aspects of PAM. First, in Section 2, we introduce a convenient mathematical representation of PAM signals. This representation is used in the subsequent sections where we investigate the properties and characteristics of such signals. In Section 3, we discuss the spectral characteristics of PAM signals, and, finally, in Section 4, we discuss the various aspects of signal demodulation and detection, including carrier recovery and symbol synchronization.

2. PAM SIGNAL REPRESENTATION

Many digital communication systems are *bandpass* systems, where the digital information is transmitted over the communication channel by means of carrier modulation. At the transmitter, the information-bearing *lowpass* signal is impressed on the carrier signal by the modulator, thus translating the information signal from *baseband* frequencies to the carrier frequency. At the receiver, the demodulator performs the reverse process of recovering the digital lowpass signal from the modulated carrier, translating the signal from the frequency band of the carrier down to baseband. When representing bandpass signals mathematically, it is often convenient to reduce them to equivalent lowpass signals so that they may be characterized independently of the particular carrier frequency or frequency band in use.

A bandpass signal may be represented as

$$s(t) = \mathrm{Re}\{x(t)e^{j2\pi f_c t}\} \tag{1}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part of a complex-valued quantity and $x(t)$ is the equivalent lowpass signal that is impressed on the carrier signal with center frequency f_c. In general, $x(t)$ is a complex-valued signal and it is often referred to as the *complex envelope* of the real signal $s(t)$, but in PAM the lowpass signal is real-valued.

A PAM modulator maps digital information (bits) into analog, finite energy waveforms that differ only in amplitude. The mapping is usually performed by taking groups of k bits at a time and selecting one out of a total of $M = 2^k$ possible such waveforms for transmission over the channel. The signal waveforms may be represented as

$$s_m(t) = \mathrm{Re}\{A_m g(t)e^{j2\pi f_c t}\} = A_m g(t)\cos 2\pi f_c t,$$
$$m = 1, 2, \ldots, M, \quad 0 \le t \le T \tag{2}$$

where $\{A_m, m = 1, 2, \ldots, M\}$ denotes the set of M possible amplitudes, $g(t)$ is a real-valued signal pulse, and T is the duration of a symbol interval. The shape of $g(t)$ may be tailored to achieve a certain spectral shaping of the transmitted signal so that it matches the spectral characteristics of the channel.

The energy of the PAM signal is dependent on the energy in the signal pulse, and is given by

$$E_m = \int_0^T s_m^2(t)\,dt = \int_0^T \left[A_m g(t)\cos 2\pi f_c t\right]^2 dt$$
$$= A_m^2 \int_0^T g^2(t)\cos^2 2\pi f_c t\,dt = \frac{1}{2}A_m^2 E_g \tag{3}$$

where $E_g = \displaystyle\int_0^T g^2(t)\,dt$ denotes the signal pulse energy. When $g(t)$ has a rectangular shape, as shown in Fig. 2, the resulting modulation is also known as *amplitude shift keying* (ASK). In this case, the pulseshape is given as

$$g(t) = \begin{cases} a, & 0 \le t \le T \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

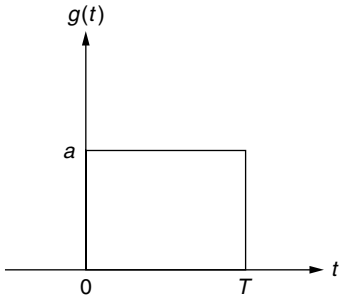


Figure 2. Rectangular signal pulse $g(t)$.

The reciprocal of the symbol interval, $R_s = 1/T$, is known as the symbol rate, namely, the rate at which the modulated carrier changes amplitude. The symbol rate is related to the bit rate $R_b = 1/T_b$ as

$$R_s = \frac{R_b}{k} \tag{5}$$

where k is the number of bits transmitted per symbol and T_b is the duration of a bit interval. The equivalent lowpass representation of the PAM waveform is real-valued and given by

$$x_m(t) = A_m g(t), \quad m = 1, 2, \ldots, M, \quad 0 \le t \le T \tag{6}$$

The signal amplitude can take on the discrete values

$$A_m = (2m - 1 - M)\, d, \quad m = 1, 2, \ldots, M \tag{7}$$

where $2d$ is the distance between adjacent signal amplitudes. Figure 3 illustrates the signal amplitude diagram for binary ($M = 2$) and quaternary ($M = 4$) PAM signals, respectively, as well as examples of possible bit mappings. In the examples shown in the figure, *Gray mapping* is used, where only a single bit differs in adjacent constellation points. Since the signal amplitudes are confined to the real line, we refer to these signals as *one-dimensional*.

Digital PAM may also be used in baseband transmission systems, that is, systems that do not require carrier modulation. In this case, the signals are represented by the equivalent lowpass formulation of Eq. (6). Figure 4 illustrates a four-level baseband PAM signal using rectangular signal pulses with amplitude $a = 1$.

3. SPECTRAL CHARACTERISTICS OF PAM

Now that we have introduced a mathematical representation of pulse-amplitude-modulated (PAM) signals, we are

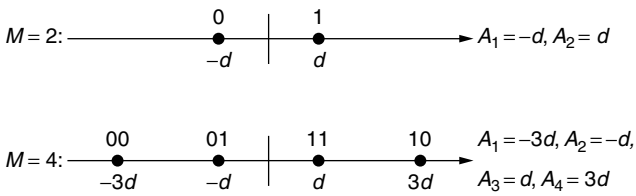


Figure 3. Signal amplitude diagram for binary and quaternary PAM signals.

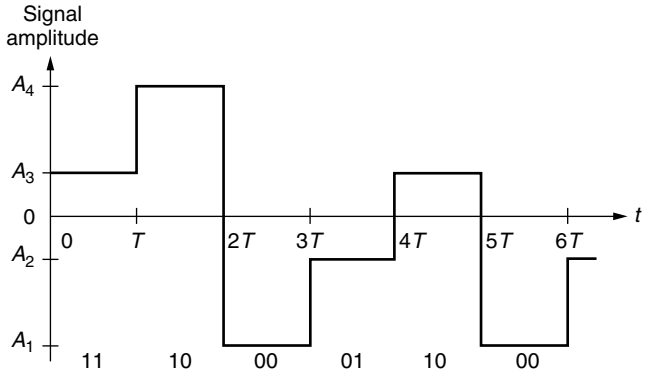


Figure 4. Four-level baseband PAM signal.

ready to discuss their spectral characteristics. Most transmission channels offer a limited bandwidth, due to either physical limitations or regulatory constraints. It is therefore important to determine the spectral content of PAM signals, or any modulated signals, for that matter, to make sure that the bandwidth of the signal does not exceed that of the channel. Also, since bandwidth is such a precious resource, it is important that it is utilized efficiently.

Let us again consider the bandpass signal $s(t)$ introduced in (1). The transmitted information sequence is a random sequence. Consequently, the corresponding PAM signal is a stochastic process whose spectral characteristics are described by its power density spectrum. The power density spectrum, denoted by $\Phi_{ss}(f)$, is obtained by Fourier transforming the autocorrelation function of $s(t)$, which is given by

$$\phi_{ss}(\tau) = \mathrm{Re}\{\phi_{xx}(\tau) e^{j2\pi f_c \tau}\} \tag{8}$$

The Fourier transformation results in

$$\Phi_{ss}(f) = \tfrac{1}{2}[\Phi_{xx}(f - f_c) + \Phi_{xx}(-f - f_c)] \tag{9}$$

where $\Phi_{xx}(f)$ is the power density spectrum of the lowpass equivalent signal $x(t)$. It is evident from (9) that in order to investigate the spectral characteristics of $s(t)$ it is sufficient to consider the power density spectrum of $x(t)$.

The lowpass signal has the general form

$$x(t) = \sum_{n=-\infty}^{\infty} A_n g(t - nT) \tag{10}$$

where the subscript n is a time index. The mean of $x(t)$ and its autocorrelation function are both periodic with period T. Hence, $x(t)$ is a *cyclostationary process*. The autocorrelation function averaged over a single period can be shown to be [1]

$$\overline{\phi}_{xx}(\tau) = \frac{1}{T} \sum_{m=-\infty}^{\infty} \phi_{AA}(m) \phi_{gg}(\tau - mT) \tag{11}$$

where $\phi_{AA}(m)$ is the autocorrelation of the information sequence represented by the amplitudes $\{A_n\}$, and $\phi_{gg}(\tau)$ is the autocorrelation function of the pulse $g(t)$, defined as

$$\phi_{gg}(\tau) = \int_{-\infty}^{\infty} g^*(t) g(t + \tau)\, dt \tag{12}$$

where $*$ denotes the complex conjugate. By taking the Fourier transform of (11), we obtain the power density spectrum

$$\Phi_{xx}(f) = \frac{1}{T}|G(f)|^2 \Phi_{AA}(f) \tag{13}$$

where $G(f)$ is the Fourier transform of $g(t)$ and $\Phi_{AA}(f)$ is the power density spectrum of the information sequence. From (13) we realize that the power density spectrum $\Phi_{xx}(f)$ depends directly on the spectral characteristics of both the information sequence $\{A_n\}$ and the pulse $g(t)$. Consequently, we may shape the spectral characteristics of the PAM signal by manipulating either the transmitter pulseshape or the correlation properties of the transmitted sequence. In the latter case, dependence between signals transmitted in different symbol intervals is introduced in a process known as *modulation coding*, resulting in a signal with memory. However, standard pulse amplitude modulation is a *memoryless* modulation since the mapping of bits into symbol waveforms is performed independently of any previously transmitted waveforms.

We shall assume here that the information symbols in the transmitted sequence are uncorrelated and have zero mean. In this case, the power density spectrum of the transmitted sequence is simply $\Phi_{AA}(f) = \sigma_A^2$. Thus, spectral shaping of the PAM signal is accomplished exclusively by selecting the pulseshape $g(t)$. Let us consider two examples that illustrate the spectral shaping that results from two different pulses. The first is the rectangular pulse introduced earlier in Fig. 2, which is used in amplitude shift keying. The energy density spectrum of the rectangular pulse, given as the square of the magnitude of the Fourier transform $G(f)$, is

$$|G(f)|^2 = (aT)^2 \left(\frac{\sin \pi f T}{\pi f T}\right)^2 \tag{14}$$

The energy spectrum is shown in Fig. 5. We note that the spectrum has a main lobe with a width of $2/T$ and zeros at $f = n/T$, $n = \pm 1, \pm 2, \ldots$. Its tail decays inversely as f^2. The resulting power density spectrum of the lowpass signal is given by

$$\Phi_{xx}(f) = \sigma_A^2 a^2 T \left(\frac{\sin \pi f T}{\pi f T}\right)^2 \tag{15}$$

In the second example, the pulse is a *raised-cosine* pulse as shown in Fig. 6 and mathematically represented by

$$g(t) = \frac{a}{2}\left[1 + \cos \frac{2\pi}{T}\left(t - \frac{T}{2}\right)\right], \quad 0 \le t \le T. \tag{16}$$

The corresponding energy density spectrum is given by

$$|G(f)|^2 = \frac{(aT)^2}{4}\left(\frac{\sin \pi f T}{\pi f T(1 - f^2T^2)}\right)^2 \tag{17}$$

In this case, the main lobe has a width of $4/T$, which is twice the width of the main lobe of the rectangular pulse. Zeros occur at $f = n/T$, $n = \pm 2, \pm 3, \ldots$. However, the tails of the spectrum decay inversely as f^6, which means that the energy outside the main lobe is virtually zero. This is illustrated in Fig. 7, which shows a heavily magnified version of the energy density spectrum.

These two examples serve to illustrate how the spectral shape of the transmitted signal can be tailored to match the spectral characteristics of the channel. The raised-cosine pulse requires a greater bandwidth, but in return the out-of-band energy is much lower than with the rectangular pulse.

4. DEMODULATION AND DETECTION OF PAM SIGNALS

In this section, we describe the various elements of the optimal receiver for PAM signals transmitted over the additive white Gaussian noise (AWGN) channel. Furthermore, we evaluate the performance of this receiver in terms of the symbol error and bit error probabilities. The AWGN channel is the most benign of channels, and the performance that can be achieved on this channel is

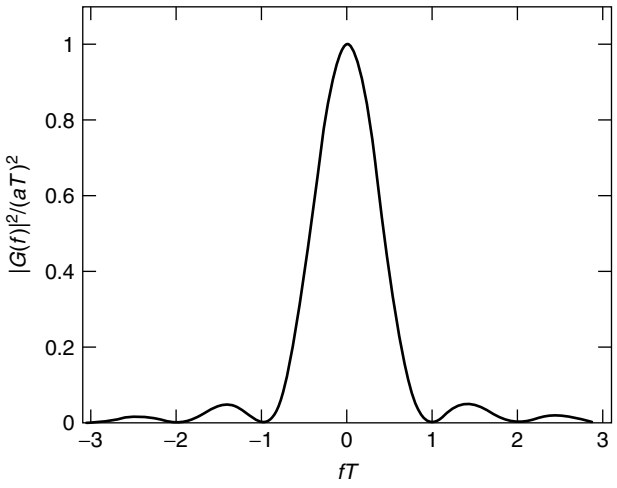


Figure 5. Energy density spectrum $|G(f)|^2$ for a rectangular pulse.

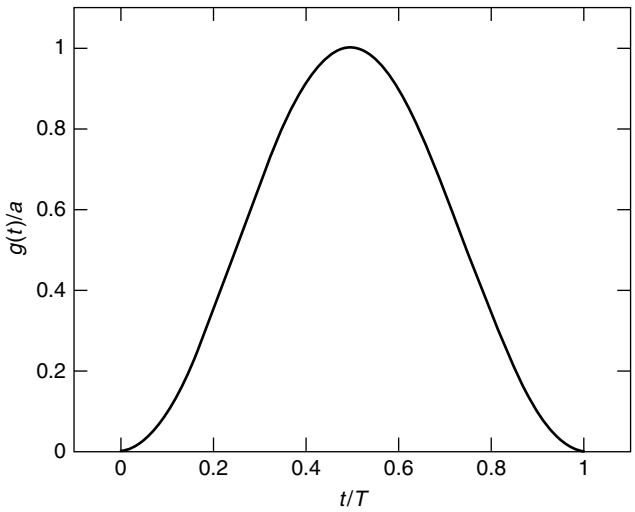


Figure 6. Raised-cosine pulse.

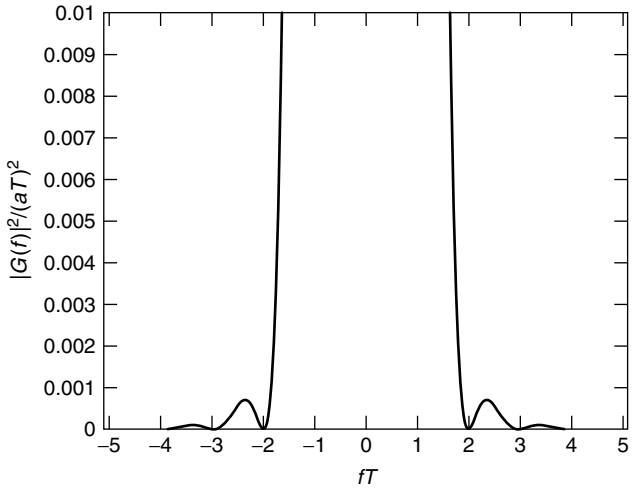


Figure 7. Energy density spectrum $|G(f)|^2$ for a raised-cosine pulse.

often used as a benchmark for evaluating a particular modulation scheme and receiver structure on other types of channels (e.g., fading channels, multipath channels, and channels with various kinds of interference). At the end of this section, we discuss carrier-phase recovery and symbol synchronization for PAM signals.

Let us first consider the input to the receiver, that is, the received signal, which consists of the original transmitted signal as well as white Gaussian noise. As described earlier, in each signaling interval the transmitted signal consists of one of M possible signal waveforms $\{s_m(t), m = 1, 2, \ldots, M\}$. The received signal can therefore be represented as

$$r(t) = s_m(t) + n(t), \quad 0 \le t \le T \tag{18}$$

where $n(t)$ is a sample function of the stochastic AWGN process with power spectral density $N_0/2$. At this point it is useful to introduce the concept of *signal space*. Signal waveforms can be given an equivalent vector representation, where, in general, the M finite energy waveforms are represented by weighted linear combinations of $N \le M$ orthonormal functions $f_n(t)$. Thus, any signal can be represented as a point in the N-dimensional signal space spanned by the basis functions $\{f_n(t), n = 1, 2, \ldots, N\}$. The *Euclidean distance* between these points is a measure of the similarity of the signal waveforms, and consequently dictates how well the receiver will be able to determine which waveform was transmitted. As noted earlier, PAM signals are one-dimensional ($N = 1$) and can therefore be represented simply by the general form

$$s_m(t) = s_m f(t) \tag{19}$$

where the basis function $f(t)$ is defined as the unit energy signal waveform

$$f(t) = \sqrt{\frac{2}{E_g}}\, g(t) \cos 2\pi f_c t \tag{20}$$

and s_m is a point on the real line given as

$$s_m = A_m \sqrt{\frac{E_g}{2}}, \quad m = 1, 2, \ldots, M \tag{21}$$

The Euclidean distance between any pair of signal points is

$$d_{lk} = \sqrt{(s_l - s_k)^2} = d\sqrt{2E_g}\,|l - k|, \quad l, k = 1, 2, \ldots, M, l \ne k \tag{22}$$

where $2d$ is the distance between adjacent signal amplitudes, as defined earlier. The distance between a pair of adjacent signal points is known as the minimum Euclidean distance, and is given by $d_{\min} = d\sqrt{2E_g}$.

A block diagram of an M-ary PAM receiver is shown in Fig. 8. The receiver consists of a demodulator and a detector as well as circuits for carrier recovery and symbol synchronization. The task of the demodulator is to compute the projection r of the received waveform $r(t)$ onto the basis function spanning the one-dimensional signal space. Based on r, the detector then decides which one of the M possible waveforms was transmitted. Since we assume a memoryless modulated signal, the detector can make its decisions separately for each signaling interval. In

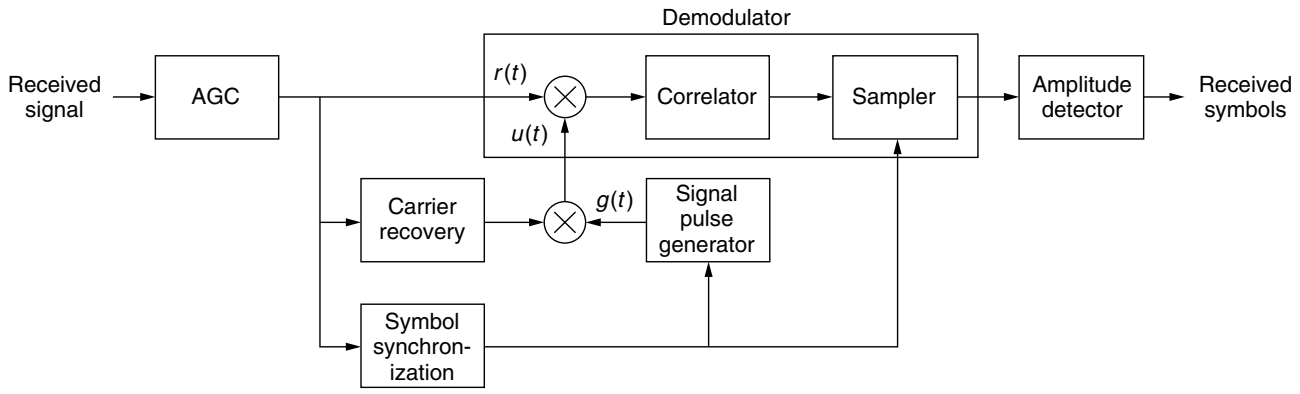


Figure 8. Block diagram of M-ary PAM receiver.

order to perform these tasks, the demodulator output must be sampled periodically, once per symbol interval. Symbol synchronization is therefore required and must be established before demodulation and detection can be performed. Because of the unknown propagation delay from the transmitter to the receiver, the symbol timing must be extracted from the received signal. Carrier-phase offset is another consequence of the unknown propagation delay. In the case of *phase-coherent detection*, this offset must be estimated, and this is the task of the carrier-phase recovery circuit. In the subsequent sections, we examine in more detail the functions of signal demodulation, signal detection, carrier recovery, and symbol synchronization.

4.1. Signal Demodulator

Let us first assume that the carrier phase has been accurately estimated and symbol synchronization has been properly established. The demodulator computes the projection of the received waveform onto the signal space by passing the received signal $r(t)$ through a *correlator*, as shown in Fig. 8. The result of this projection is

$$r = \int_0^T r(t)f(t)\,dt = s_m + n \tag{23}$$

where

$$s_m = \int_0^T s_m(t)f(t)\,dt \tag{24}$$

and n is the projection of the noise onto the signal space, given as the zero-mean random variable

$$n = \int_0^T n(t)f(t)\,dt \tag{25}$$

with variance $\sigma_n^2 = N_0/2$. The correlator output is a Gaussian random variable with mean $\mathrm{E}[r] = s_m$ and variance equal to the noise variance, i.e., $\sigma_r^2 = \sigma_n^2 = N_0/2$. Thus, the probability density function of r, conditional on the mth signal waveform being transmitted, is the Gaussian *likelihood function* given by

$$p(r|s_m) = \frac{1}{\sqrt{\pi N_0}} \exp\left[-\frac{(r-s_m)^2}{N_0}\right] \tag{26}$$

Instead of using a correlator for signal demodulation, a *matched filter* may be used in its place. In general, the matched filter is a bank of N filters whose impulse responses are matched to the N basis functions $\{f_n(t)\}$. In our case, where $N = 1$, the impulse response of the single filter is given by

$$h(t) = f(T-t), \quad 0 \le t \le T \tag{27}$$

The output of the matched filter is

$$\begin{aligned} y(t) &= \int_0^t r(\tau)h(t-\tau)\,d\tau \\ &= \int_0^t r(\tau)f(T-t+\tau)\,d\tau \end{aligned} \tag{28}$$

Sampling this output at $t = T$ yields

$$y(T) = \int_0^T r(\tau)f(\tau)\,d\tau = r \tag{29}$$

which is identical to the output of the correlator. An important property of the matched filter is that it maximizes the output signal-to-noise ratio (SNR) in an AWGN channel. The output SNR is given by

$$\mathrm{SNR}_{\mathrm{out}} = \frac{(\mathrm{E}[r])^2}{\sigma_n^2} \tag{30}$$

4.2. Signal Detector

The correlator (or matched-filter) output contains all the relevant information, namely, the *sufficient statistic*, about the transmitted signal that the detector needs in order to make a decision, assuming that the carrier phase is known at the receiver and symbol synchronization has been established. The detector applies a decision rule that seeks to maximize the probability of a correct decision. More specifically, it decides in favor of the signal which has the maximum a posteriori probability

$$\Pr\{s_m|r\}, \quad m = 1, 2, \ldots, M \tag{31}$$

of the M possible transmitted signals. This decision rule is known as the *maximum a posteriori probability* (MAP) criterion. The a posteriori probabilities can be expressed as

$$\Pr\{s_m|r\} = \frac{p(r|s_m)\Pr\{s_m\}}{p(r)} \tag{32}$$

where $p(r|s_m)$ was given in (26) and $\Pr\{s_m\}$ is the a priori probability that the mth signal will be transmitted. The denominator is independent of which signal is transmitted. When all the M transmitted signals are equally likely, the MAP criterion is reduced to maximizing the likelihood function (26) over the M possible signals, and this is called the *maximum-likelihood* (ML) criterion. For ease of computation, it is common to use the *loglikelihood function* given by

$$\ln p(r|s_m) = -\frac{1}{2}\ln(\pi N_0) - \frac{1}{N_0}(r-s_m)^2 \tag{33}$$

Maximizing (33) is equivalent to finding the signal that minimizes the Euclidean distance metric

$$d(r, s_m) = (r-s_m)^2 \tag{34}$$

In our case, the detector is an amplitude detector, as shown in Fig. 8. An *automatic gain control* (AGC) circuit [2] is added to the front end of the receiver to eliminate short-term channel gain variations that would otherwise affect the amplitude detector. The AGC maintains a fixed average SNR at its output.

4.3. Detector Performance

Armed with the knowledge gained in the previous two sections, we may evaluate the performance of the optimal

receiver for M-ary PAM signals in terms of the symbol error and bit error probabilities. As before, we assume equally probable transmitted signals. First we need to define some quantities that we will use in our calculations. The average energy of the transmitted signals, using the representations given in (3) and (7), is

$$E_{\text{av}} = \frac{1}{M}\sum_{m=1}^{M} E_m = \frac{d^2 E_g}{2M}\sum_{m=1}^{M}(2m-1-M)^2 = \frac{1}{6}(M^2-1)d^2 E_g \tag{35}$$

The average power is given as $P_{\text{av}} = E_{\text{av}}/T$, and the average energy per transmitted bit is $E_b = P_{\text{av}} T_b$.

Now let us consider the Euclidean distance metric given by (34). The detector compares the projection r with all the M signal candidates and decides in favor of the signal with the smallest metric. Equivalently, the detector compares r with a set of $M-1$ thresholds located at the midpoints between adjacent signal amplitude levels. A decision is made in favor of the level that is closer to r. An erroneous decision will occur if the magnitude of the noise is significant enough to make r extend into one of the two regions belonging to the neighboring amplitude levels. This is true for all but the two outermost levels $\pm(M-1)$, where an error can be made in only one direction. Hence, the average probability of a symbol error, P_M, is equal to the probability that the noise component $n = r - s_m$ exceeds one-half of the distance between two thresholds, $d_{\min}/2$:

$$P_M = \frac{M-1}{M}\Pr\left\{|n| > \frac{d_{\min}}{2}\right\} = \frac{2(M-1)}{M}\Pr\left\{n > \frac{d_{\min}}{2}\right\} \tag{36}$$

where

$$\Pr\left\{n > \frac{d_{\min}}{2}\right\} = \frac{1}{\sqrt{\pi N_0}}\int_{d_{\min}/2}^{\infty} e^{-x^2/N_0}\,dx \tag{37}$$

It is straightforward to show that the symbol error probability is equal to

$$P_M = \frac{2(M-1)}{M} Q\left(\sqrt{\frac{d_{\min}^2}{2N_0}}\right) \tag{38}$$

where $Q(\cdot)$ is the standard error function [1]. It is customary to express the error probability as a function of the average SNR per bit, $\gamma_b = E_b/N_0$, and using the definition of $d_{\min}$ and (35), we can easily show that the symbol error probability can be expressed as

$$P_M = \frac{2(M-1)}{M} Q\left(\sqrt{\frac{6k}{M^2-1}\gamma_b}\right) \tag{39}$$

where $k = \log_2 M$. Figure 9 shows the symbol error probability as a function of γ_b for $M = 2, 4, 8, 16$. When Gray coding is used in the mapping of bits into symbols, adjacent symbols differ in only a single bit and in most

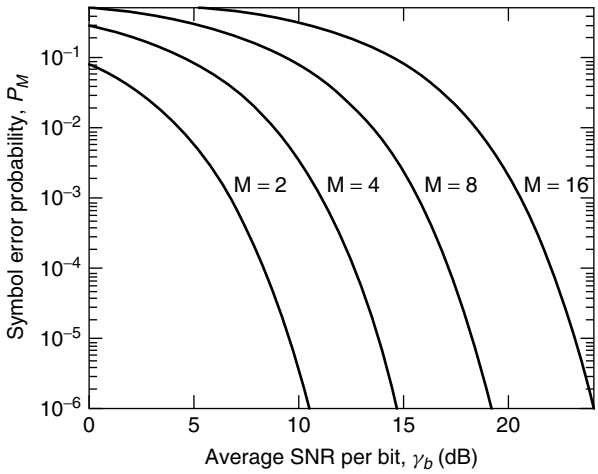


Figure 9. Symbol error probability for M-ary PAM.

cases a symbol error will result in only a single bit error. The bit error probability P_b is therefore approximated as

$$P_b \approx \frac{1}{k} P_M \tag{40}$$

4.4. Carrier-Phase Recovery

If phase-coherent detection is employed, the carrier phase needs to be known or accurately estimated. In most practical systems the phase is estimated directly from the modulated signal. Let us assume that the received signal has the form

$$r(t) = x(t)\cos(2\pi f_c t + \phi) + n(t) \tag{41}$$

where $x(t)$ is the information-bearing lowpass signal and ϕ is the carrier phase. For simplicity, we assume here that the propagation delay is known and we set it equal to zero. As shown in Fig. 8, the carrier recovery circuit computes a phase estimate $\hat{\phi}$, which is then used to generate the reference signal $u(t) = g(t)\sin(2\pi f_c t + \hat{\phi})$ for the correlator.

A *phase-locked loop* (PLL) [2] may be used to provide the maximum-likelihood estimate of the carrier phase. The PLL consists of a multiplier, a loop filter, and a *voltage-controlled oscillator* (VCO), as shown in Fig. 10. Let us for a moment assume that the input to the PLL is an unmodulated carrier signal represented by the sinusoid $\cos(2\pi f_c t + \phi)$. The output of the VCO is another sinusoid

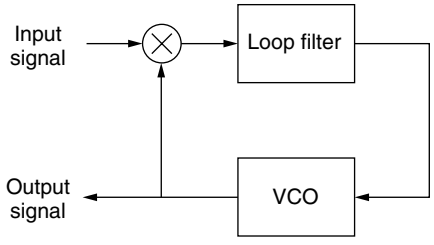


Figure 10. Phase-locked loop.

given by $\sin(2\pi f_c t + \hat{\phi})$. The output of the multiplier is the product of these two signals, given by

$$\cos(2\pi f_c t + \phi)\sin(2\pi f_c t + \hat{\phi}) = \tfrac{1}{2}\sin(\hat{\phi} - \phi) + \tfrac{1}{2}\sin(4\pi f_c t + \phi + \hat{\phi}) \quad (42)$$

The loop filter is a lowpass filter that effectively removes the double-frequency ($2f_c$) component. The output of this filter is a voltage signal that controls the VCO. The resulting PLL tracks the phase of the incoming carrier, seeking to minimize the phase error $\Delta\phi = \hat{\phi} - \phi$. In normal operation, the phase error is small and therefore $\sin(\hat{\phi} - \phi) \approx \hat{\phi} - \phi$.

We now consider two specific methods for estimating the carrier phase of a PAM signal. One widely used method involves employing a square-law device to square the received signal and generate a double-frequency signal that can then be used to drive a PLL tuned to $2f_c$ [1]. Figure 11 illustrates such a *squaring loop*, where, for the sake of clarity, we have ignored the noise term of the received signal. The output of the squarer is

$$r^2(t) = x^2(t)\cos^2(2\pi f_c t + \phi) = \tfrac{1}{2}x^2(t) + \tfrac{1}{2}x^2(t)\cos(4\pi f_c t + 2\phi) \quad (43)$$

This signal is passed through a bandpass filter tuned to $2f_c$, resulting in a periodic signal without the sign information contained in $x(t)$, namely, a phase-coherent signal at twice the carrier frequency. The filtered frequency component at $2f_c$ is used to drive the PLL. Finally, the output signal $u'(t) = \sin(4\pi f_c t + 2\hat{\phi})$ from the VCO is passed through a frequency divider to provide the output signal $u(t) = \sin(2\pi f_c t + \hat{\phi})$. This output has a phase ambiguity of 180° relative to the phase of the received signal, so binary data must be differentially encoded at the transmitter and, consequently, differentially decoded at the receiver. The squaring operation leads to some noise enhancement that results in an increase in the variance of the phase error $\Delta\phi$.

Another well-known carrier-phase estimation method is the *Costas loop* [1], which is illustrated by the block diagram shown in Fig. 12. As before, we ignore the noise term of the received signal for the sake of clarity. In this method, the received signal is multiplied by two versions of the VCO output, one phase-shifted 90° relative to the other. The multiplier outputs are

$$y_{\cos}(t) = r(t)\cos(2\pi f_c t + \hat{\phi}) = \tfrac{1}{2}x(t)\cos(\hat{\phi} - \phi) + \tfrac{1}{2}\cos(4\pi f_c t + \phi + \hat{\phi}) \quad (44)$$

and

$$y_{\sin}(t) = r(t)\sin(2\pi f_c t + \hat{\phi}) = \tfrac{1}{2}x(t)\sin(\hat{\phi} - \phi) + \tfrac{1}{2}\sin(4\pi f_c t + \phi + \hat{\phi}) \quad (45)$$

These signals are passed through lowpass filters that reject the double-frequency terms. An error signal is generated by multiplying the filtered outputs of the multipliers, yielding

$$e(t) = \tfrac{1}{8}x^2(t)\sin(2(\hat{\phi} - \phi)) \quad (46)$$

The error signal is then filtered by the loop filter, and the resulting signal is a voltage signal that controls the VCO.

As in the squaring loop, some noise enhancement occurs that causes the variance of the phase error to increase. Also, the VCO output has a 180° phase ambiguity which necessitates the use of differential encoding and decoding of the binary data.

4.5. Symbol Synchronization

As mentioned at the beginning of Section 4, the output of the correlator or matched filter must be sampled once per symbol interval. Because of the unknown propagation delay τ from the transmitter to the receiver, symbol synchronization must first be established in order to perform the sampling at the right time instants. This is usually accomplished by extracting a clock signal directly from the received signal itself and using it to control the sampling time instants $t_k = kT + \hat{\tau}$, where $\hat{\tau}$ denotes an estimate of the propagation delay.

The maximum-likelihood estimate of the propagation delay can be obtained by maximizing its likelihood

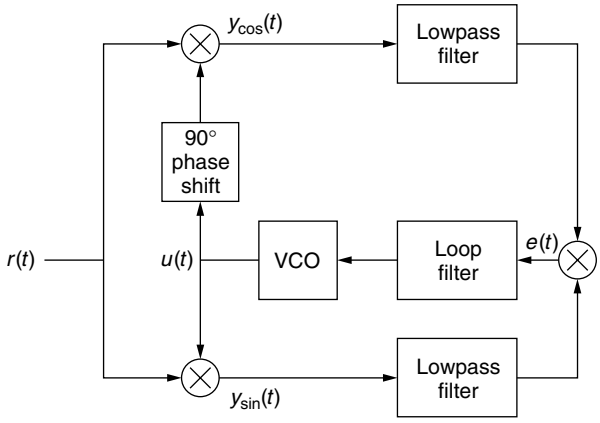


Figure 12. Carrier-phase recovery using a Costas loop.

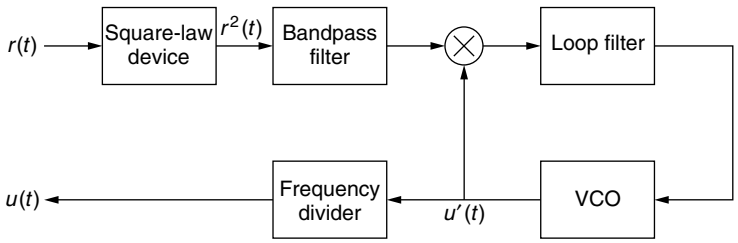


Figure 11. Carrier-phase recovery using a squaring loop.

function. Let us, for simplicity, consider baseband PAM since the procedure is easily extended to carrier-modulated PAM. The received signal is represented by

$$r(t) = s(t;\tau) + n(t) \tag{47}$$

where $s(t;\tau)$ is the (delayed) transmitted signal

$$s(t;\tau) = \sum_k x(t - kT - \tau) = \sum_k A_k g(t - kT - \tau) \tag{48}$$

and $\{A_k\}$ is the transmitted symbol sequence. The likelihood function has the form

$$\begin{aligned}\Lambda(\tau) &= C \int_{T_0} r(t)s(t)\,dt \\ &= C \sum_k A_k y_k(\tau)\end{aligned} \tag{49}$$

where $y_k(\tau)$ is the output of the correlator given by

$$y_k(\tau) = \int_{T_0} r(t)g(t - kT - \tau)\,dt \tag{50}$$

and T_0 is the integration interval. The maximum-likelihood estimate is obtained by averaging $\Lambda(\tau)$ over the probability density function (PDF) of the information symbols, $p(A_k)$, and differentiating the result. In the case of binary PAM, where $A_k = \pm 1$ with equal probability, the PDF is given as

$$p(A_k) = \tfrac{1}{2}\delta(A_k - 1) + \tfrac{1}{2}\delta(A_k + 1) \tag{51}$$

In the case of M-ary PAM, we may approximate the PDF by assuming that the symbols are continuous random variables with a zero-mean, unit-variance Gaussian distribution. In this case, the PDF is given by

$$p(A_k) = \frac{1}{\sqrt{2\pi}} e^{-A_k^2/2} \tag{52}$$

It can be shown that averaging the likelihood function over the Gaussian PDF and taking the natural logarithm yields a loglikelihood function of the form

$$\overline{\Lambda}_L(\tau) \approx \frac{1}{2} C^2 \sum_k y_k^2(\tau) \tag{53}$$

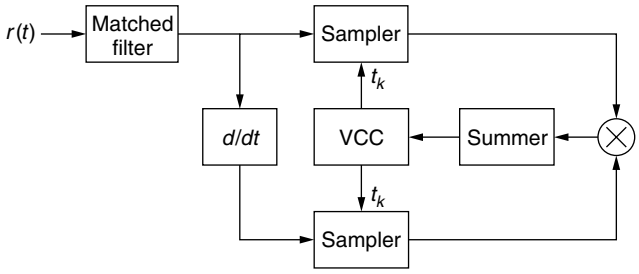


Figure 13. Non-decision-directed symbol synchronization for baseband PAM.

Next, we find the derivative of (53) and set it equal to zero:

$$\frac{d}{dt} \sum_k y_k^2(\tau) = 2 \sum_k y_k(\tau) \frac{dy_k(\tau)}{d\tau} = 0 \tag{54}$$

Figure 13 shows a symbol synchronization circuit that performs timing estimation by using a tracking loop based on Eq. (54). The circuit attempts to move the sampling instant until the derivative is zero, which occurs at the peak of the signal. The summation element serves as the loop filter that drives the voltage-controlled clock (VCC). We note that the structure of the circuit is quite similar to that of the Costas loop discussed earlier in the section on carrier-phase estimation.

A related technique, illustrated in Fig. 14, is known as the *early–late gate* synchronizer. In this technique, the output of the matched filter (or correlator) is sampled 2 times extra per symbol interval, once prior to the proper sampling instant (i.e., $T_- = T - \delta$) and once after the proper sampling instant (i.e., $T_+ = T + \delta$). As an example, let us consider the output of the filter matched to the rectangular symbol waveform introduced in Section 2. The matched-filter output attains its peak at the midpoint $t = T$, as shown in Fig. 15, and this time instant is naturally the optimal sampling time. Since the output is even with respect to the optimal sampling instant, the magnitudes of the samples taken at T_- and T_+ are equal. Hence, the proper sampling instant can be determined by adjusting the early and late sampling instants until the absolute values of the two samples are equal. An error signal is formed by taking the difference between the absolute values of the two samples, and the filtered error signal is used to drive the VCC. If the timing is off, the error signal will be nonzero and the timing signal is either advanced or retarded, depending on the sign of the error.

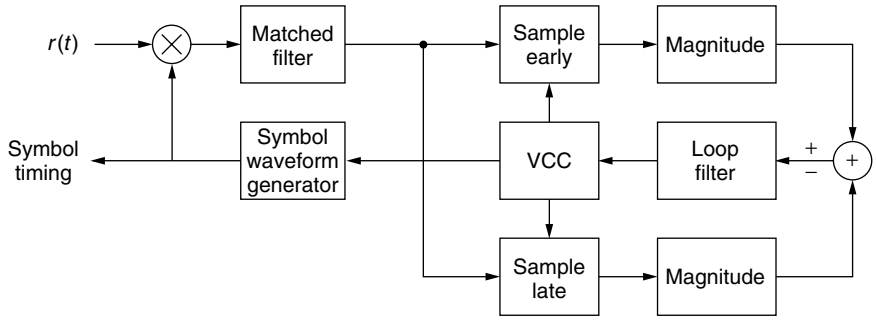


Figure 14. Symbol synchronization using an early–late gate synchronizer.

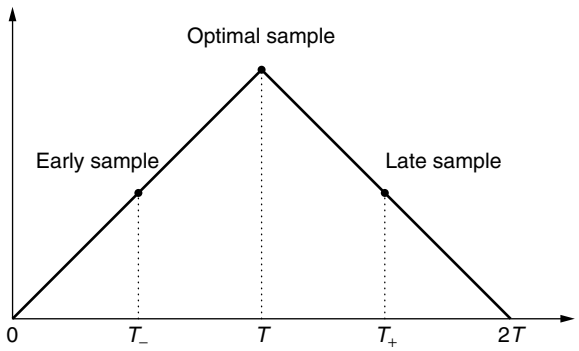


Figure 15. Matched-filter output for a rectangular signal pulse.

5. CONCLUSION

In this article, we have given an overview of the modulation method known as *pulse amplitude modulation*. We started out by introducing a mathematical representation of PAM signals and discussing their spectral characteristics. In particular, we noted that the spectrum of memoryless PAM signals can be manipulated to match the spectral characteristics of the transmission channel by selecting an appropriate signal pulse. Next, we discussed the various elements of PAM receivers for use in AWGN channels, including signal demodulation and detection, carrier-phase recovery, and symbol synchronization. We also discussed the performance of PAM detectors in terms of symbol and bit error probabilities. PAM is a well-established and mature modulation technique and has been treated in numerous articles and textbooks over the years. The interested reader will find comprehensive treatments of the topic in the works listed in the Bibliography.

BIOGRAPHY

Bjørn A. Bjerke received his Siv. Ing. degree in electrical engineering from the Norwegian Institute of Technology (NTH), Trondheim, Norway, in 1995, and his M.S. and Ph.D. degrees from Northeastern University, Boston, Massachusetts, in 1997 and 2001, respectively, both in electrical engineering. He joined Qualcomm, Inc. in 2001 and is currently a senior systems Engineer in their Concord, Massachusetts, R&D unit. His research focuses on physical layer algorithms and architectures for multi-antenna wireless communications, including channel coding/decoding, adaptive modulation, interference cancellation and channel equalization. Dr. Bjerke is a member of Eta Kappa Nu, IEEE, the Norwegian Signal Processing Society and the Norwegian Society of Chartered Engineers.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, 2001.
2. H. Meyr and G. Ascheid, *Synchronization in Digital Communications*, Vols. 1, 2, Wiley, 1990.
3. S. G. Wilson, *Digital Modulation and Coding*, Prentice-Hall, 1996.
4. J. G. Proakis and M. Salehi, *Contemporary Communication Systems Using Matlab*, Brooks/Cole, 2000.
5. B. Sklar, *Digital Communications Fundamentals and Applications*, Prentice-Hall, 1988.
6. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, 1965.
7. J. G. Proakis and M. Salehi, *Communication Systems Engineering*, Prentice-Hall, 1994.
8. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer, 1988.

PULSE POSITION MODULATION

G. CARIOLARO
T. ERSEGHE
Università di Padova
Padova, Italy

1. INTRODUCTION

This article deals with *pulse position modulation* (PPM), which is now introduced in the more general context of *pulse modulations*. In any communication system we deal with some kind of information that we wish to send through a medium (or channel) separating the transmitter from the receiver. The information is rarely in a form that is suitable for direct transmission. In this context we call *modulation* the process of transforming the original signal in such a way that the resulting waveform can be efficiently transferred through the medium and that the original signal can be efficiently reconstructed at the receiver's end.

In *pulse modulation* systems the information is usually carried over a series of regularly recurrent pulses on which we intervene by varying parameters such as amplitude, duration, shape and, of course, temporal position. If the signal that we wish to transmit is a continuous-time function $s(t), t \in \mathbb{R}$, it will be sampled to be conformable to the discrete nature of pulses. In considering the feasibility of pulse modulations, it is important to recognize that the continuous transmission is unnecessary, provided that the continuous function $s(t)$ is band-limited and the pulses occur often enough [1]. That is, pulse modulations are inherently discrete modulations.

The necessary conditions for their feasibility are thus given by the sampling theorem stating, in its basic formulation, that any analog signal $s(t)$ with limited bandwidth range B can be uniquely expressed by samples $s_n = s(nT), n \in \mathbb{Z}$ taken at regular intervals T, where the sampling frequency $F = 1/T$ is at least twice the bandwidth range. The original signal can then be exactly reconstructed from its samples by an interpolating filter. Fortunately, in any physically realizable transmission system this condition is always satisfied and so pulse modulations constitute a feasible way to transfer information.

Nowadays, PPM is mainly used for digital transmission (which does not exclude that the original message may have an analog format converted to digital just on the basis of the sampling theorem), where the discrete sequence s_n takes the values from a finite set of amplitudes (alphabet) and where the size is typically a power of 2. So we have 2-PPM, 4-PPM, 8-PPM, and so on.

The article is organized as follows. In Section 2 we introduce PPM (both analog and digital) in the more general framework of pulse modulations, and in Section 3 we consider the problem of the generation of PPM signals. In these preliminary sections, the signals may be interpreted as deterministic or random functions as well. However, following the lines of modern communication theory, in the subsequent parts a probabilistic methodology becomes mandatory. Thus, in Section 4 we formulate the spectral analysis where the signal is modeled in terms of random processes. In the two final sections, Section 5 and Section 6, we evaluate performances of digital PPM systems in the presence of noise and also their achievable information rates. The article concludes with a brief record of modern applications, mainly in optical communication systems.

We have deliberately omitted the performance evaluation of analog PPM systems. To this regard we suggest to the interest reader the optimal lectures of [2], [3].

2. VARIETIES OF PULSE MODULATIONS

In pulse modulations the unmodulated carrier is usually a periodic repetition of a given pulse $q(t)$

$$v_0(t) = \sum_{n=-\infty}^{+\infty} q(t - nT) \tag{1}$$

and the message is always a discrete-time signal, $s_n = s(nT)$, with sampling spacing T equal to the period of the carrier. The fundamental parameters of the carrier are the *pulse shape* $q(t)$, usually rectangular with duration limited to a fraction of T and the period T that uniquely determines the sampling instants nT, $-\infty < n < +\infty$, which express the repetitiveness of pulses, that is the *synchronism*.

The message s_n can be a sampled version of an analog signal $s(t)$, $t \in \mathbb{R}$, as previously discussed, in which case we will talk of *analog* modulation. Conversely, the message can be digital, in which case the modulation becomes *digital*. The way to reconstruct the original signal is pretty different in these two cases.

Modulation of the carrier is obtained by varying the nth pulse in dependence of the value s_n, that is, the reference pulse $q(t - nT)$ is replaced by a pulse $q(t - nT, s_n)$ dependent on s_n. The modulated signal can thus be written as

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT, s_n) \tag{2}$$

The transmitter has a double role: to generate the modulated pulses $q(t, s_n)$ and to position them at the synchronism instant nT. In the digital case, if $\mathcal{A}$ is the finite set (alphabet) of values that s_n can assume, the set of modulated pulses

$$\{q(t, a) \mid a \in \mathcal{A}\} \tag{3}$$

must be in a one-to-one correspondence with $\mathcal{A}$ to guarantee demodulation. For example, if the alphabet $\mathcal{A}$ has 64 values, the modulator must be capable of generating 64 different pulses. In the analog case, s_n belongs to a continuous set, usually given by a finite interval to assure that pulses do not overlap.

The simplest pulse modulation, and maybe the most common approach, is pulse amplitude modulation (PAM), which varies the pulse amplitude proportionally to the value of s_n, that is, we have $q(t, s_n) = s_n q(t)$ (Fig. 1). Another classical modulation, the one of interest to us, is PPM which modifies the *position* of the pulses, that is, we have

$$q(t, s_n) = q(t - K\, s_n) \tag{4}$$

where K is a suitable constant with the constraint $-\frac{1}{2}T \le K\, s_n < \frac{1}{2}T$ or, alternatively, $0 \le K\, s_n < T$ to guarantee that modulated pulses do not overlap. In PPM the modulated signal has the expression

$$\boxed{v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT - K\, s_n)} \tag{5}$$

A closely related alternative, called pulse duration modulation (PDM) or pulse width modulation (PWM), would instead consider to modify the duration of pulses, that is,

$$q(t, s_n) = q(t/(K\, s_n)) \tag{6}$$

In this case the constraint is in the duration of the modulated pulses, which must be positive and not bigger

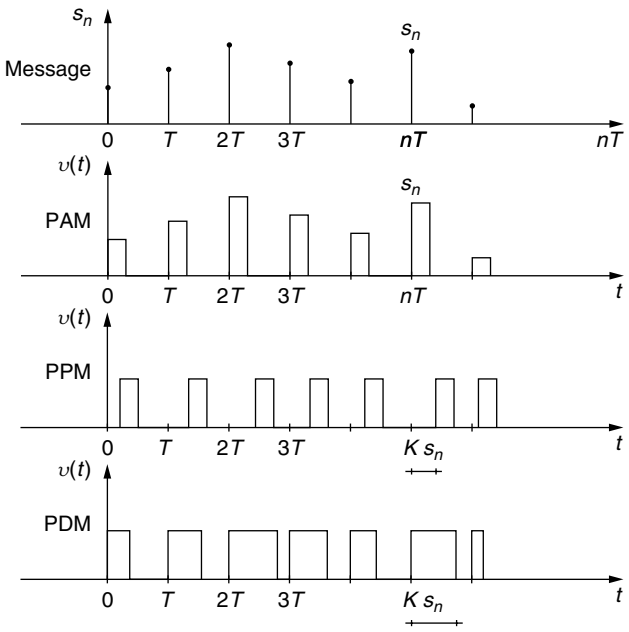


Figure 1. Classical pulse modulations.

than T, so that in (6) we have the further constraint $s_n > 0$; if s_n is not strictly positive, then the expression $K\,s_n$ must be substituted by something of the form of $T_0 + K\,s_n$.

Another worth mentioning modulation format related to PPM is pulse interval modulation (PIM), also known as differential PPM where the information is contained in the relative-distance between successive pulses. The relation to PPM can be seen in Fig. 2, where it is underlined that PIM discards the void portion of the frame after the PPM pulse, thus resulting more efficient in terms of transmission capacity and bandwidth requirements (but not in terms of transmitted-power).

3. GENERATION AND MODELS OF PPM SIGNALS

In a PPM signal the information is confined to the sequence of instants (positions) $nT + K\,s_n$ and the shape of the fundamental pulse $q(t)$ is, in some respects, irrelevant. In the ideal case the pulse may be a delta function, that is

$$v_\delta(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\,s_n) \tag{7}$$

The transition from the ideal PPM signal $v_\delta(t)$ to a PPM with a given pulse shape $q(t)$ is simply obtained by a shaping filter, that is

$$v(t) = v_\delta * q(t) = \sum_{n=-\infty}^{+\infty} q(t - nT - K\,s_n)$$

where $*$ denotes convolution.

2.1. Generation of Analog PPM Signals (general approach)

The traditional generation of analog PPM signals is based on the preliminary generation of a PDM signal. In practice, the principal use of PDM is for the generation and detection of PPM because the latter is superior for message transmission.

A sequence of operations to produce a PPM signal starting from the sampling sequence $s_n = s(nT)$ is depicted in Fig. 3. It is assumed that s_n is unipolar and bounded, that is $0 \le s_n < S_0$. The first operation is a holding of the value s_n for the corresponding time-slot $\mathcal{I}_n = [nT, (n+1)T)$, thus producing a PAM signal $\tilde{s}(t)$ with

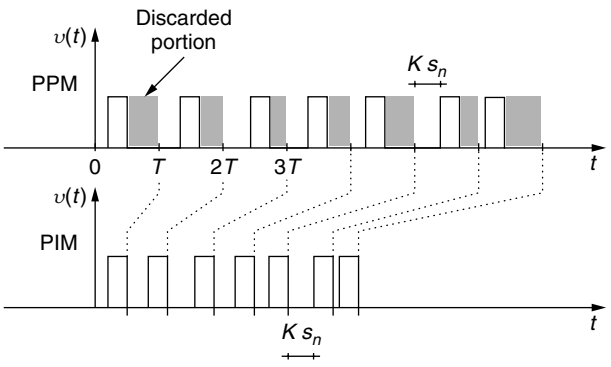


Figure 2. From PPM to PIM.

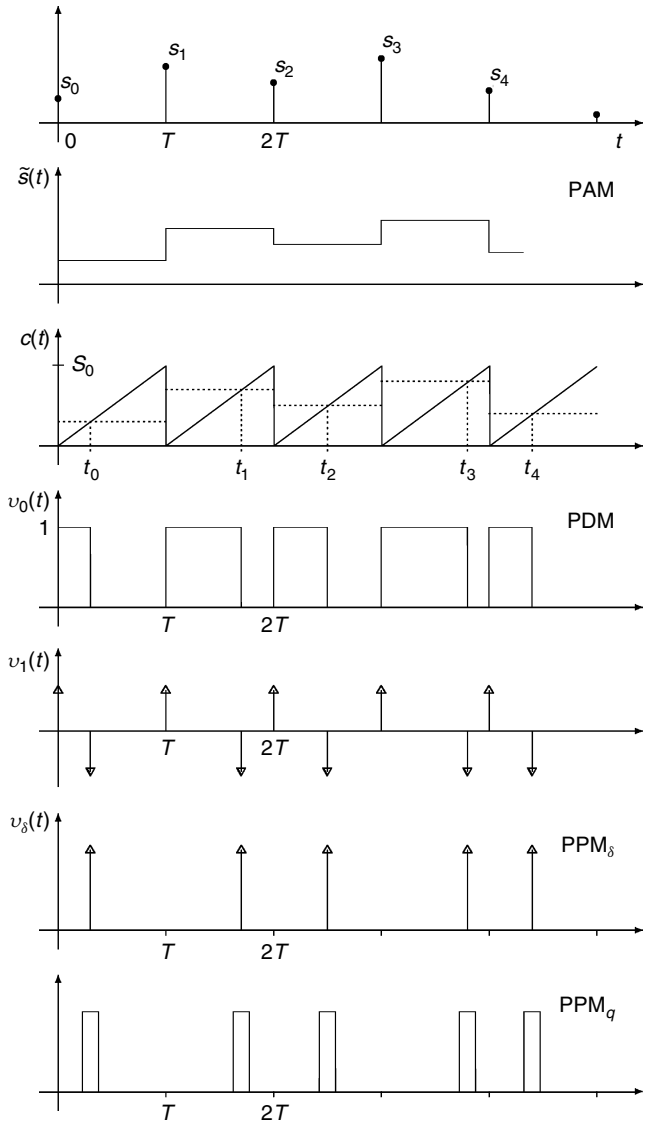


Figure 3. Generation of a PPM signal.

a full duty cycle. In the second step, a triangular carrier $c(t)$, running linearly from $c(nT) = 0$ to $c((n+1)T) = S_0$, is compared with $\tilde{s}(t)$ and a unitary output is produced as long as $\tilde{s}(t)$ is above $c(t)$. In this way, a sequence of rectangular pulses $v_0(t)$ is produced with the nth pulse starting at time nT with a duration proportional to the value of $\tilde{s}(t)$ in $\mathcal{I}_n$, that is $K\,s_n$. Hence, $v_0(t)$ is just a PDM signal. The derivative of $v_0(t)$ produces a sequence of paired delta functions $\delta(t - nT) - \delta(t - nT - K\,s_n)$. With an inverse half-wave rectification which removes the upward delta functions, the PPM signal

$$v_\delta(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\,s_n) \tag{8}$$

is obtained. Finally, a reshaping of these *spikes* gives the standard PPM format.

It is clear that an ideal receiver can recover from the PPM wave the modulating sequence s_n. The natural way

is to first obtain a PDM signal from (8), which can be done in the presence of synchronization which marks the beginning nT of the time slot. Next, the PDM signal is integrated starting from nT and with a finite integration time ($< T$). Hence, a uniform sampling at time just before $(n+1)T$ gives a sequence proportional to s_n. If s_n was obtained by sampling an analog signal $s(t)$, an interpolation finally provides the full recovery of $s(t)$. For digital PPM other more efficient recovery systems are used (see Section 5).

2.2. Analog PPM Signals with Nonuniform Sampling

In the standard analog PPM the modulating sequence s_n is obtained from a continuous-time message $s(t)$ with a sampling at the equally-spaced instants nT (uniform sampling), as remarked by writing $s_n = s(nT)$. There is another format of PPM, which is obtained by a nonuniform sampling. Historically, this form came first for its implementation is easier, at least with analog circuitry.

As a matter of fact, if in the sequence of operations of Fig. 3 we avoid the sampling, that produces the sequence s_n, and the holding, which gives $\tilde{s}(t)$ from s_n, and we feed the comparator directly with the analog message $s(t)$, the subsequent operations work as well to produce a PPM format. But, the times t_n, in correspondence of which the signal values $s_n = s(t_n)$ are taken, are no more equally spaced (Fig. 4). In fact, in the nth time-slot the carrier ramp is given by $c(t) = (t - nT)S_0/T, t \in \mathcal{I}_n$ and it is compared directly with $s(t)$. Hence, the coincidence $s(t) = c(t)$ happens at the instant

$$t_n = nT + K\, s(t_n), \qquad K = S_0/T \tag{9}$$

and the final PPM signal becomes

$$v\delta(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\, s(t_n)) \tag{10}$$

with nonuniform sampling instants determined by (9).

Although more easy to generate (with analog circuitry!), the analysis of this PPM format is very difficult because the instants t_n are determined by an implicit equation, as (9) is. Fortunately, Rowe [2], using some powerful properties of the delta function, was able to obtain a very interesting expression of (10), namely

$$\begin{aligned} v_\delta(t) &= F|1 - K\, s'(t)| \sum_{n=-\infty}^{+\infty} e^{j2\pi nF(t - K\ s(t))} \\ &= F|1 - K\, s'(t)| \left\{ 1 + 2 \sum_{n=1}^{+\infty} \cos 2\pi nF[t - K\, s(t)] \right\} \end{aligned} \tag{11}$$

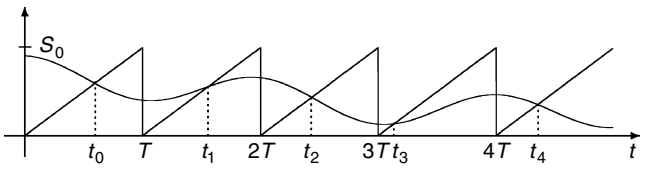


Figure 4. PPM with nonuniform sampling.

where $F = 1/T$. In Appendix A, we give an alternative and easier deduction of this expression. According to (11) the PPM signal $v\delta(t)$ consists of a baseband term $F|1 - K\ s'(t)| = A_0(t)$ and of bandpass terms around the frequencies nF,

$$2\, A_0(t) \cos 2\pi\, [nFt + \varphi_n(t)]$$

which exhibit both an amplitude modulation and a phase modulation with phase deviation $\varphi_n(t) = -nFT_0 s(t)$. Note that with a little signal processing it is possible to obtain a phase or a frequency modulated signal starting from a PPM signal with a nonuniform sampling. As a matter of fact, this possibility was concretely used in the past under the name of *serrasoid* technique.

We remark that (11), although suggestive, should be used with caution, especially in spectral analysis, because in the frequency domain the terms in (11) may have a strong overlapping.

2.3. Generation of Digital PPM Signals

For the generation of digital PPM signals, the general method can be used, but more specific methods, which rely upon the consideration that the permitted positions are finitely many, are possible. Here, we outline a general method, which is valid for all pulse modulations and gives the PPM as a special case.

Let $v(t)$ be the pulse modulated signal

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT, s_n)$$

where the data sequence $s_n = s(nT)$ belongs to an M-ary alphabet $\mathcal{A}_M = \{0, 1, \ldots, M-1\}$. The modulation format is specified by M distinct pulses $q(t, \alpha) = q_\alpha(t)$, which we store in a vector $\mathbf{q}(t) = [q_0(t), q_1(t), \ldots, q_{M-1}(t)]$. In particular, for PPM we have

$$q(t, \alpha) = q_\alpha(t) = q(t - \alpha T_0), \qquad T_0 = T/M$$

so that the vector $\mathbf{q}(t)$ collects M replicas of a same pulse uniformly distributed in the time slot $[0,\ T)$. The key of the method lies on a representation of the M-ary data s_n by a binary word of length M, $\mathbf{b}_n = [b_n(0), b_n(1), \ldots, b_n(M-1)]$, where

$$b_n(\alpha) = \delta_{s_n,\alpha} = \begin{cases} 1 & \text{for } s_n = \alpha \\ 0 & \text{for } s_n \neq \alpha \end{cases} \tag{12}$$

$\delta_{s_n,\alpha}$ being the Kronecker delta function. For instance, for $M = 4$, we find

$$\begin{aligned} \mathbf{b}_n &= [1\ 0\ 0\ 0] \quad \text{when } s_n = 0 \\ &= [0\ 1\ 0\ 0] \quad \text{when } s_n = 1 \\ &= [0\ 0\ 1\ 0] \quad \text{when } s_n = 2 \\ &= [0\ 0\ 0\ 1] \quad \text{when } s_n = 3 \end{aligned}$$

Clearly, the word sequence $\mathbf{b}_n = \mathbf{b}(nT)$ brings the same information of the original data sequence s_n without

ambiguity, and in fact one can uniquely determine s_n from $\mathbf{b}_n$ because

$$s_n = \alpha \quad \Longleftrightarrow \quad b_n(\alpha) = 1$$

that is, the value of s_n at the time nT is given by the position of the 1 in the vector $\mathbf{b}_n$.

Next, consider that the words $\mathbf{b}_n$ feed a bank of interpolating filters with impulse responses $q_\alpha(t)$ (Fig. 5). The input-output relationship of the αth filter is

$$v_\alpha(t) = \sum_{n=-\infty}^{+\infty} b_n(\alpha) q_\alpha(t - nT)$$

and in the nth time slot the output is $q_\alpha(t - nT) = q(t - nT, s_n)$ if $s_n = \alpha$ and 0 otherwise. Hence, the contribution of all filters

$$\sum_{\alpha \in \mathcal{A}_M} v_\alpha(t) = \sum_{n=-\infty}^{+\infty} \sum_{\alpha \in \mathcal{A}_M} b_n(\alpha) q_\alpha(t - nT) = v(t) \tag{13}$$

provides the desired modulated signal $v(t)$. In other words, the word sequence $\mathbf{b}_n$ gives the signaling to the filter-bank to turn on the right filter (one per time) in every time slot.

Note that the scheme of Fig. 5 provides a powerful representation of a pulse modulator because it explicitly separates the nonlinear part of the modulator, given by the word formatting operation, from the linear part, given by the filter bank $\mathbf{q}(t)$. Finally, we remark that (13) can be written in the elegant form

$$v(t) = \sum_{n=-\infty}^{+\infty} \mathbf{b}_n \mathbf{q}(t - nT)$$

provided that $\mathbf{b}_n$ is regarded as a row vector (matrix $1 \times M$) and $\mathbf{q}(t)$ is a column vector (matrix $M \times 1$).

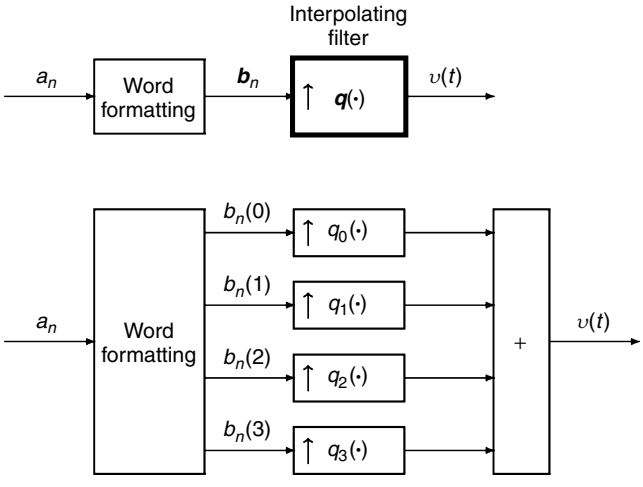


Figure 5. Generation of digital pulse modulated signals.

2.4. Digital PPM Signals Through PAM Signals

In the case of PPM, where the interpolating filters have impulse responses $q_\alpha(t) = q(t - \alpha T_0)$ that are uniformly distributed delayed versions of a reference pulse $q(t)$, the scheme of Fig. 5 can be further simplified to express PPM in terms of PAM modulations.

In particular, we obtain the scheme of Fig. 6, where the word sequence $\mathbf{b}_n$ is mapped by a parallel-to-serial conversion, called de-framing, to the binary sequence c_m (with rate $1/T_0$), where

$$c_{nM+\alpha} = b_n(\alpha)$$

The binary sequence c_m is then interpolated by a filter with impulsive response $q(t)$ to obtain the PPM signal

$$\sum_{m=-\infty}^{+\infty} c_m q(t - mT_0) = \sum_{n=-\infty}^{+\infty} \sum_{\alpha=0}^{M-1} c_{nM+\alpha} q(t - (nM + \alpha)T_0)$$
$$= \sum_{n=-\infty}^{+\infty} \sum_{\alpha=0}^{M-1} b_n(\alpha) q_\alpha(t - nT)$$

which, by use of (13), gives $v(t)$. Again, in Fig. 6 the nonlinear operations are contained in the word formatting, while the final PPM signal is obtained from the binary sequence c_m by a simple PAM modulation.

Simple modifications of the scheme of Fig. 6 let us easily define the discrete version of modulations related to PPM. For example, PDM with $M = 4$ is obtained by redefining the words

$$\begin{aligned} \mathbf{b}_n &= [1\ 0\ 0\ 0] \quad \text{when } s_n = 0 \\ &= [1\ 1\ 0\ 0] \quad \text{when } s_n = 1 \\ &= [1\ 1\ 1\ 0] \quad \text{when } s_n = 2 \\ &= [1\ 1\ 1\ 1] \quad \text{when } s_n = 3 \end{aligned}$$

Similarly we have for PIM, for which the word mapping becomes a variable-length operation,

$$\begin{aligned} \mathbf{b}_n &= [1] \quad \text{when } s_n = 0 \\ &= [0\ 1] \quad \text{when } s_n = 1 \\ &= [0\ 0\ 1] \quad \text{when } s_n = 2 \\ &= [0\ 0\ 0\ 1] \quad \text{when } s_n = 3 \end{aligned}$$

and de-framing must take into account for the variable length of words.

4. SPECTRAL ANALYSIS

As for every other modulation format, spectral analysis is a fundamental tool for understanding the band occupancy of the PPM modulation but also for several other reasons. Surely, PPM produces a *baseband* signal, that is with a

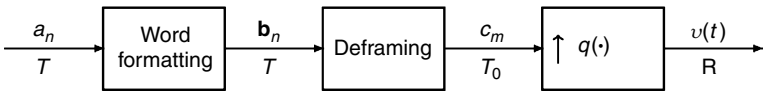


Figure 6. Digital PPM signals generated by a PAM filter.

spectrum displayed around the frequency origin, which is essentially determined by the fundamental pulse $q(t)$, and we also have to expect that the bandwidth is of the order of $1/T_0$, where T_0 is the duration of $q(t)$. However, the exact evaluation of the spectrum is not trivial for two main reasons: 1) PPM is a nonlinear modulation and 2) PPM contains a periodic component, given by the mean value, which determines the presence of spectral lines and that causes a nonstationary behavior.

4.1. Formulation of Spectral Analysis

In spectral analysis signals are modeled as random processes. In particular, we assume that the sample sequence $s_n = s(nT)$ is a discrete-time *stationary* random process. Then, it is easy to show that the PPM signal

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t - nT - s_n) \tag{14}$$

is not stationary, rather it is *cyclostationary*, that is with a statistical description which is periodic with respect to a reference time, being the period given by the sampling period T. We start by showing this for the mean value $m_v(t) = \mathrm{E}[v(t)]$, where $\mathrm{E}[\cdot]$ is the expectation operator. Considering that s_n is stationary, we have

$$m_v(t) = \sum_{n=-\infty}^{+\infty} \mathrm{E}[q(t - nT - s_n)] = \sum_{n=-\infty}^{+\infty} \overline{q}(t - nT)$$

where $\overline{q}(t) = \mathrm{E}[q(t - s_n)]$ is the *average pulse*, which is independent of n because of stationarity. Hence, $m_v(t)$ is the periodic repetition of the average pulse and has period T.

The spectral analysis starts from the *correlation* function of $v(t)$, defined as $\tilde{r}_v(t, \tau) = \mathrm{E}[v(t)v(t + \tau)]$, which is periodic with respect to the reference time t. Then, taking the time average

$$r_v(\tau) = \frac{1}{T} \int_0^T \tilde{r}_v(t, \tau)\, dt \tag{15}$$

the dependence on t is removed. The (average) power spectral density (PSD), that is the quantity of interest in spectral analysis, is then obtained as the Fourier transform of the average correlation (15), namely

$$R_v(f) = \int_{-\infty}^{+\infty} r_v(\tau) e^{-j2\pi f\tau}\, d\tau \tag{16}$$

This is the most common approach to determine the PSD for cyclostationary processes.[1]

In general, the PSD $R_v(f)$ can be decomposed in a *continuous* part $R_v^{(c)}(f)$ and in a *discrete* part $R_v^{(d)}(f)$, which exhibits *spectral lines* (Lebesgue decomposition). Spectral lines are due to the periodic component of the PPM signal, that is by the mean value $m_v(t)$. In fact, it can be shown that $m_v(t)$ has PSD $R_v^{(d)}(f)$ and the deviation $v(t) - m_v(t)$ has PSD $R_v^{(c)}(f)$. For a correct spectral analysis it is important to evaluate the two components separately.

4.2. Evaluation of the PSD

Because PPM is a nonlinear modulation, the evaluation of the PSD requires to know the second-order statistics of the modulating sequence s_n. More specifically, let $\Psi_s(f)$ and $\Psi_s(f_1, f_2; \tau)$ be the characteristic functions of first and second order of s_n, written in terms of "frequency" in place of the usual z variables, namely

$$\begin{aligned} \Psi_s(f) &= \mathrm{E}[e^{j2\pi f s_n}] \\ \Psi_s(f_1, f_2; kT) &= \mathrm{E}[e^{j2\pi(f_1 s_n + f_2 s_{n+k})}] \end{aligned} \tag{17}$$

Let also $\Phi_s(f_1, f_2; f)$ be the (discrete) Fourier transform of $\Psi_s(\cdot; kT)$ with respect to kT, that is

$$\Phi_s(f_1, f_2; f) = \sum_{k=-\infty}^{+\infty} \Psi_s(f_1, f_2; kT) e^{-j2\pi fkT} \tag{18}$$

Then, as shown in Appendix B, the PSD of the PPM signal $v(t)$ is given by

$$\boxed{R_v(f) = F\Phi_s(-f, f; f)|Q(f)|^2} \tag{19}$$

where $F = 1/T$ and $Q(f)$ is the Fourier transform of the fundamental pulse $q(t)$, that is

$$Q(f) = \int_{-\infty}^{+\infty} q(t) e^{-j2\pi ft}\, dt$$

This result holds in general as soon as the modulating sequence s_n is a stationary process. For the spectral separation of the continuous and the discrete parts, further assumptions should be done concerning the behavior of the characteristic function $\Psi(f_1, f_2; kT)$ for $k \to \infty$. Here, we consider the simplest case in which s_n is statistically independent and refer to Ref. [2] for a more general situation.

If s_n and s_{n+k} are independent for $k \neq 0$, then the second part of (17) gives

$$\Psi_s(f_1, f_2; kT) = \begin{cases} \Psi_s(f_1 + f_2) & k = 0 \\ \Psi_s(f_1)\Psi_s(f_2) & k \neq 0 \end{cases} \tag{20}$$

Moreover, at the level of the characteristic function the separation between the discrete and continuous part is such that $\Psi_s(f_1)\Psi_s(f_2)$ determines the discrete part and the difference $\Psi_s(f_1 + f_2) - \Psi_s(f_1)\Psi_s(f_2)$ for $k = 0$ determines the continuous part. So, from (18) we obtain

$$\begin{aligned} \Phi_s^{(c)}(f_1, f_2; f) &= \Psi_s(f_1 + f_2) - \Psi_s(f_1)\Psi_s(f_2) \\ \Phi_s^{(d)}(f_1, f_2; f) &= \Psi_s(f_1)\Psi_s(f_2) \sum_{k=-\infty}^{+\infty} e^{-j2\pi fkT} \\ &= \Psi_s(f_1)\Psi_s(f_2) F \sum_{k=-\infty}^{+\infty} \delta(f - kF) \end{aligned} \tag{21}$$

[1] An equivalent way is the introduction of a *stationary version* of the cyclostationary signal, defined as $v_\vartheta(t) = v(t + \vartheta)$ where ϑ is a random variable uniformly distributed in $[0, T)$ and statistically independent on $v(t)$. It can be shown that $v_\vartheta(t)$ is stationary with correlation given by (15) and PSD given by (16).

where in the last row we used a widely known identity between sequences of exponentials and sequences of delta functions.

Finally, by substituting (21) in (19) we obtain

$$\begin{aligned} R_v^{(c)}(f) &= F|Q(f)|^2(1-|\Psi_s(f)|^2) \\ R_v^{(d)}(f) &= F^2|Q(f)|^2|\Psi_s(f)|^2 \sum_{k=-\infty}^{+\infty} \delta(f-kF) \\ &= F^2 \sum_{k=-\infty}^{+\infty} |Q(kF)|^2|\Psi_s(kF)|^2\delta(f-kF) \end{aligned} \tag{22}$$

where we see the presence of spectral lines at the frequencies that are multiples of the sampling rate $F = 1/T$, at least for those instances where $Q(kF) \neq 0$.

4.3. Spectrum of Digital PPM

In digital PPM the "positions" s_n are the discrete and equally spaced instants $0, T_0, \ldots, (M-1)T_0$ with $T_0 = T/M$ and each position is taken with a given probability $p_i = \mathrm{P}[s_n = iT_0]$. Then, the characteristic function of the first order becomes

$$\Psi_s(f) = \sum_{i=0}^{M-1} p_i e^{j2\pi fiT_0}$$

We now assume equally likely positions, that is $p_i = 1/M$ and a rectangular pulse $q(t) = 1$ for $0 < t < \alpha T_0$, where α is the duty cycle ($0 < \alpha < 1$). Then, the result will be expressed by the very well-known function $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$ and by its (less known) periodic version

$$\mathrm{sinc}_M(x) = \frac{1}{M}\frac{\sin(\pi x)}{\sin(\frac{\pi}{M}x)} \tag{23}$$

which has period M for M odd and period $2M$ for M even. Note that $\mathrm{sinc}^2(k) = 0$ for k an integer that is not a multiple of M and that $\mathrm{sinc}^2(kM) = 1$. The periodic sinc allows us to express the characteristic function in the form

$$\Psi_s(f) = \frac{1}{M}\frac{1-e^{-j2\pi MfT_0}}{1-e^{-j2\pi fT_0}} = e^{-j2\pi(M-1)fT_0}\mathrm{sinc}_M(fMT_0)$$

Conversely, the Fourier transform of the pulse is given by

$$Q(f) = \alpha T_0 e^{-j\pi\alpha T_0 f}\mathrm{sinc}(\alpha T_0 f)$$

Hence, by substitution in (22) we have

$$\begin{aligned} R_v^{(c)}(f) &= F(\alpha T_0)^2\,\mathrm{sinc}^2(\alpha T_0 f)(1-\mathrm{sinc}_M^2(fT)) \\ R_v^{(d)}(f) &= F^2(\alpha T_0)^2 \sum_{k=-\infty}^{+\infty} \mathrm{sinc}^2(k\alpha/M)\,\mathrm{sinc}_M^2(k)\delta(f-kF) \end{aligned}$$

The plot of this PSD is given in Fig. 7 for $M = 8$ and for two values of α. Note that $R_v^{(c)}(f)$ is always zero at $f = 0$ and other zeros fall at frequencies multiple of $F_0 = MF$. An indication of the bandwidth may be given by the side lobe determined by the factor $\mathrm{sinc}^2(f\alpha T_0)$, that is $1/\alpha T_0 = F_0/\alpha$. The spectral lines may be present at frequencies multiples of F_0 whenever $\mathrm{sinc}^2(k\alpha) \neq 0$. Note, in particular, that for $\alpha = 1$ there is only a line at $f = 0$; this is in agreement with the fact that in this case the mean value $m_v(t)$ degenerates to a constant value.

4.4. Spectrum of Analog PPM

When the sampling sequence s_n is continuous the characteristic function is given by

$$\Psi_s(f) = \int_{-\infty}^{+\infty} e^{-j2\pi fa} f_s(a)\,da$$

where $f_s(a)$ is the probability density function of s_n. Assuming a rectangular pulse with duration T_0, we consider two cases: a) s_n is uniform in $[0, T-T_0)$, not extended to $[0, T)$ to avoid collisions, and b) s_n is Gaussian with mean $m_s = (T-T_0)/2$ and variance σ_s^2 with $\sigma_s << T$ so that the probability of collision is negligible.

With s_n uniform the characteristic function is

$$\Psi_s(f) = e^{-j\pi(T-T_0)f}\mathrm{sinc}(f(T-T_0))$$

and hence

$$\begin{aligned} R_v^{(c)}(f) &= FT_0^2\,\mathrm{sinc}^2(fT_0)[1-\mathrm{sinc}^2(f(T-T_0))] \\ R_v^{(d)}(f) &= (FT_0)^2 \sum_{k=-\infty}^{+\infty} \mathrm{sinc}^2(kFT_0) \\ &\quad \times \mathrm{sinc}^2(k(1-FT_0))\delta(f-kF) \end{aligned}$$

which are illustrated at the top of Fig. 8 for $T_0 = \frac{1}{8}T$.

With s_n Gaussian we find that

$$\Psi_s(f) = e^{-j2\pi fm_s}e^{-(2\pi f\sigma_s)^2/2}$$

and hence

$$\begin{aligned} R_v^{(c)}(f) &= FT_0^2\,\mathrm{sinc}^2(fT_0)[1-e^{-(2\pi f\sigma_s)^2}] \\ R_v^{(d)}(f) &= (FT_0)^2 \sum_{k=-\infty}^{+\infty} \mathrm{sinc}^2(kFT_0)e^{-(2\pi kF\sigma_s)^2}\delta(f-kF) \end{aligned}$$

which are illustrated below in Fig. 8 for $T_0 = \sigma_s = \frac{1}{8}T$. We recall that these results hold when the samples s_n are statistically independent.

5. DIGITAL PPM DEMODULATION AND ERROR PROBABILITY

Digital PPM is the common approach in practical applications, mainly because of the digital nature of the information available for transmission. So, in the following we will concentrate on this "digital" format, for which the PPM signal is

$$v(t) = \sum_{n=-\infty}^{+\infty} q(t-nT-s_n)$$

where s_n is an M-ary data sequence with alphabet $\mathcal{A}_M = \{0, T_0, \ldots, (M-1)T_0\}$ with $T_0 = T/M$. The receiver

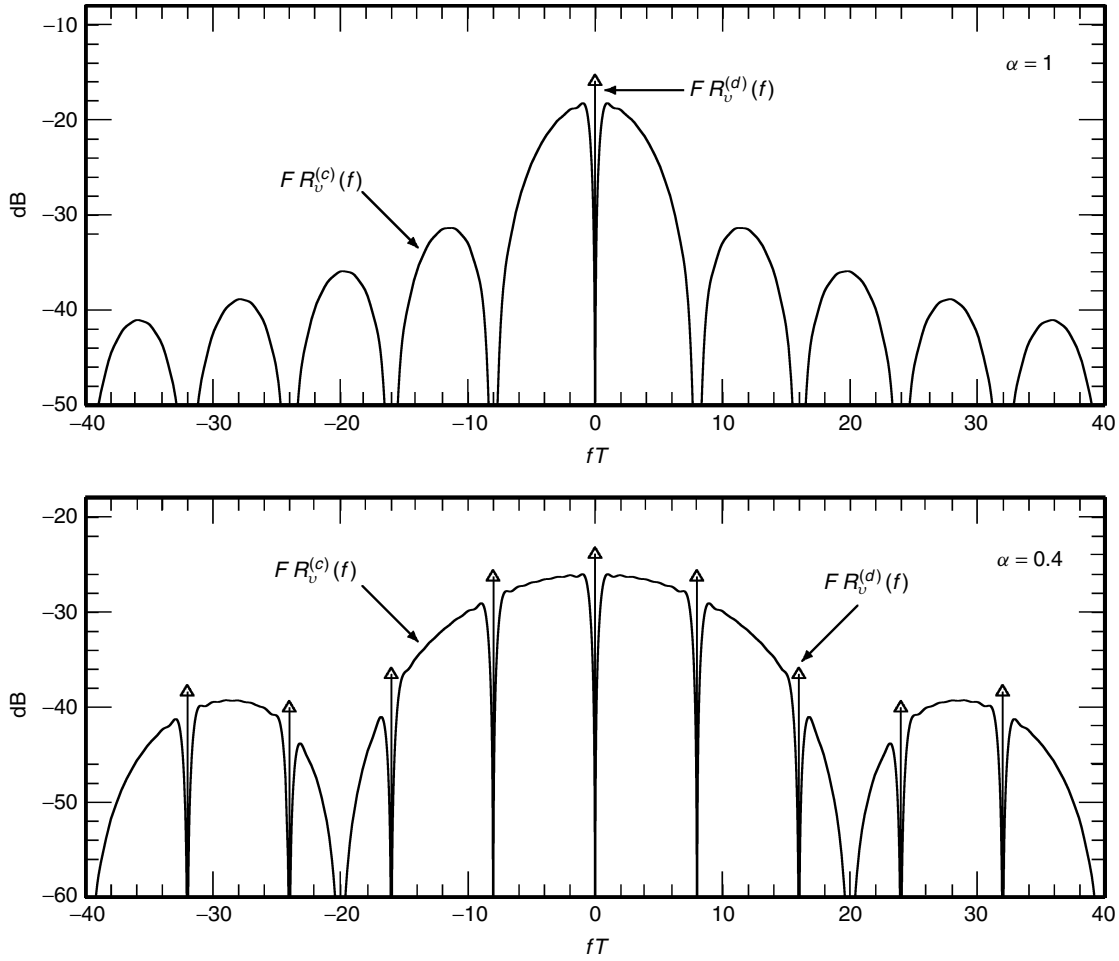


Figure 7. PSD examples for digital PPM and $M = 8$. Spectral lines are highlighted by vertical arrows.

structure is based upon optimal detection strategies and relies on banks of correlators. Fundamental performance measures of the system are given by the error probability evaluation, but also by more sophisticated capacity measures (see Section 6), which set a target for the maximum amount of information that can be sent through the PPM channel with an arbitrarily small probability of error.

5.1. Optimal Detection of Digital PPM Signals

Optimal strategies for detecting digital PPM signals rely on the assumption of ideal propagation and additive white Gaussian noise (AWGN), where the received signal can be modeled as

$$r(t) = Av(t) + \eta(t)$$

with A the attenuation and $\eta(t)$ a stationary Gaussian noise with zero mean and power spectral density $R_\eta(f) = N_0/2$. Optimal detection, that is the one that minimizes the symbol error probability (also known as maximum likelihood), is achieved by use of a matched-filter bank where the demodulator selects the value resulting in the largest cross-correlation between the received signal $r(t)$ and each of the pulses $q_\alpha(t - nT) = q(t - nT - \alpha), \alpha \in \mathcal{A}_M$.

The implementation is illustrated in Fig. 9, where the M-tuple $\varphi_n(\alpha), \alpha \in \mathcal{A}_M$ is generated by a bank of sampling filters that constitute the dual of the interpolating filter bank of Fig. 5. In particular, we have

$$\varphi_n(\alpha) = \int_{-\infty}^{+\infty} r(t) q_\alpha(t - nT)\, dt = \int_{-\infty}^{+\infty} r(t) q(t - nT - \alpha)\, dt$$

where the integration can be limited to a finite region by taking into account that, in practice, $q(t)$ always has limited extension. This extension is usually constrained to $[0, T_0)$ to overcome the superposition of adjacent pulses and, in the following, we consider this assumption (see [4] for more general cases). Finally, the demodulated sequence becomes

$$\hat{s}_n = \arg \max_\alpha \varphi_n(\alpha)$$

The statistical properties of the M-tuple $\varphi_n(\alpha), \alpha \in \mathcal{A}_M$ are usually given under the condition that the transmitted sample s_n is known. By setting $s_n = \beta$, with a little effort one easily derives that

$$\varphi_n(\alpha \mid \beta) = \varphi_n(\alpha \mid s_n = \beta) = \underbrace{AE_q \delta_{\alpha,\beta}}_{\text{expected signal}} + \underbrace{\eta_n(\alpha)}_{\text{Gaussian noise}} \qquad (24)$$

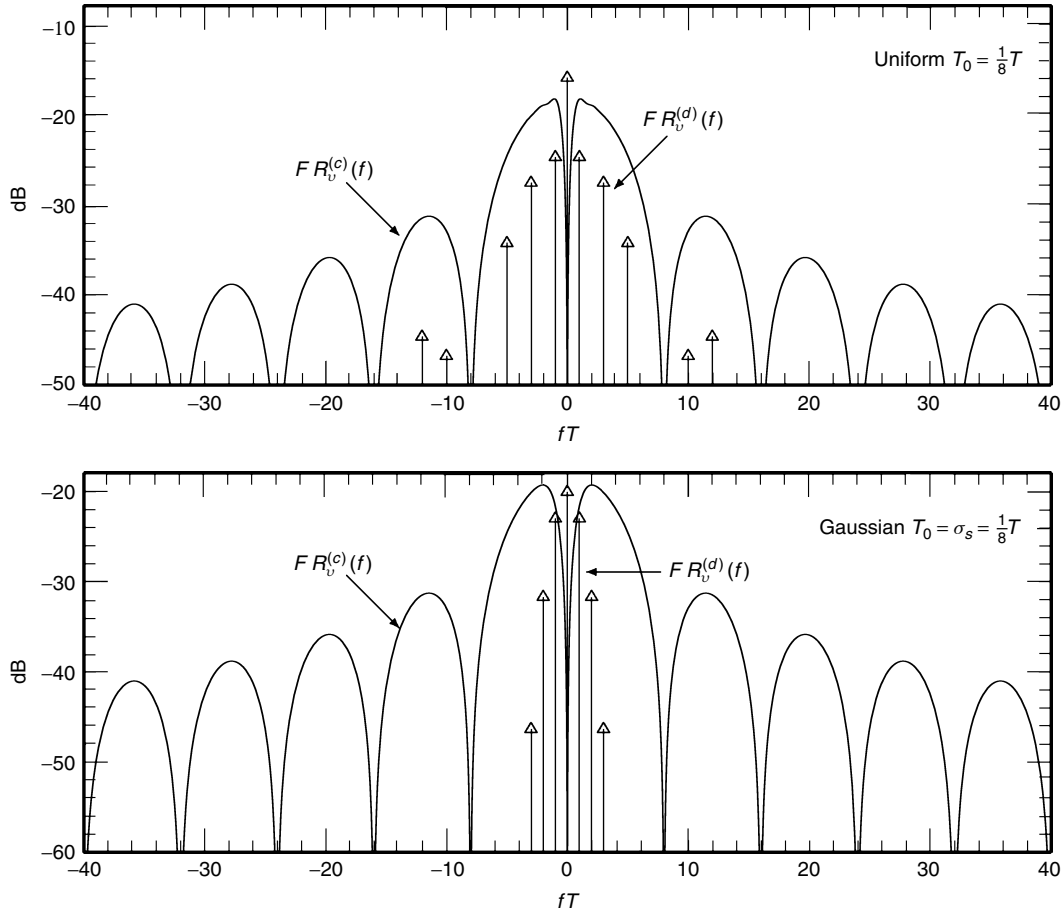


Figure 8. PSD examples for analog PPM: uniform (above) and Gaussian (below). Spectral lines are highlighted by vertical arrows.

where $E_q = \int |q(t)|^2\, dt$ is the energy of the pulse and $\delta_{\alpha,\beta}$ is the Kronecker delta function, giving 1 when $\alpha = \beta$ and 0 otherwise. Note that, the Gaussian noise is the only random term of (24) and, moreover, it is independent on the value of s_n. Because we are dealing with an AWGN channel, for any given instant nT, the random variables $\varphi_n(\alpha \mid \beta)$ are Gaussian and are thus completely specified by their means and covariances. In Appendix C we prove that these Gaussian random variables are independent with mean and variance

$$\begin{aligned} m_{\alpha|\beta} &= \mathrm{E}\,[\varphi_n(\alpha \mid \beta)] = AE_q\delta_{\alpha,\beta} \\ \sigma^2 &= \mathrm{E}\,[(\varphi_n(\alpha \mid \beta) - m_{\alpha|\beta})^2] = \tfrac{1}{2}N_0E_q \end{aligned} \tag{25}$$

where only the mean depends on the transmitted value.

5.2. Symbol Error Probability Evaluation

In the above context, the symbol error probability is given by

$$\begin{aligned} P_e = \mathrm{P}\,[\hat{s}_n \neq s_n] &= \sum_{\alpha\in\mathcal{A}_M} \mathrm{P}\,[\hat{s}_n \neq \alpha \mid s_n = \alpha]\,\mathrm{P}\,[s_n = \alpha] \\ &= \sum_{\alpha\in\mathcal{A}_M} (1 - \mathrm{P}\,[\hat{s}_n = \alpha \mid s_n = \alpha])\,\mathrm{P}\,[s_n = \alpha] \end{aligned}$$

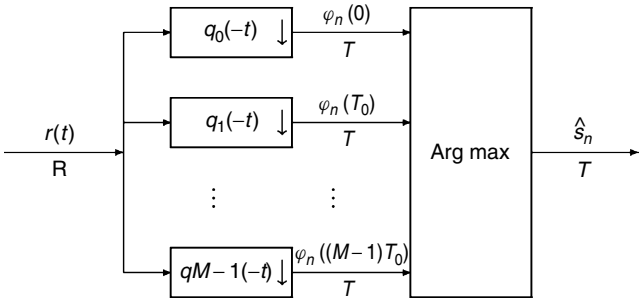


Figure 9. Optimal detection of digital PPM signals.

Because the diagonal transition probabilities

$$\mathrm{P}[\hat{s}_n = \alpha \mid s_n = \alpha] = \mathrm{P}[\alpha = \arg\max_{\gamma} \varphi_n(\gamma \mid \alpha)] = P_c$$

are independent of the value of α, as can be easily derived from the symmetric statistical properties of the M-tuple $\varphi_n(\gamma \mid \alpha)$, $\gamma \in \mathcal{A}_M$, the symbol error probability becomes

$$P_e = 1 - P_c = 1 - \mathrm{P}[\varphi_n(0 \mid 0) = \max_{\gamma} \varphi_n(\gamma \mid 0)] \tag{26}$$

where we deliberately set $\alpha = 0$.

The probability of a correct decision P_c in (26) may be further expressed in compact form as

$$\begin{aligned} P_c &= \int \mathrm{P}[\varphi_n(T_0 \mid 0) < x, \varphi_n(2T_0 \mid 0) < x, \ldots, \varphi_n \\ &\quad \times ((M-1)T_0 \mid 0) < x \mid \varphi_n(0 \mid 0) = x] f_0(x)\, dx \\ &= \int \mathrm{P}[\varphi_n(T_0 \mid 0) < x] \mathrm{P}[\varphi_n(2T_0 \mid 0) < x] \\ &\quad \times \ldots \mathrm{P}[\varphi_n((M-1)T_0 \mid 0) \times < x] f_0(x)\, dx \\ &= \int (\mathrm{P}[\varphi_n(T_0 \mid 0) < x])^{M-1} f_0(x)\, dx \end{aligned}$$

where $f_0(x)$ is the probability density of $\varphi_n(0 \mid 0)$ and where we have taken into account that all $\varphi_n(\alpha \mid 0), \alpha \neq 0$ are statistically independent and identically distributed. Then, by considering the Gaussian distribution of all $\varphi_n(\alpha \mid 0)$, we finally find

$$P_e = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (1 - \Phi^{M-1}(x)) \exp\left[-\frac{1}{2} \left(x - \sqrt{\frac{2\mathcal{E}_s}{N_0}} \right)^2 \right] dx \tag{27}$$

where $\mathcal{E}_s = A^2 E_q$ is the received energy per symbol and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp[-t^2/2]\, \mathrm{d}\, t$ is the normalized cumulative Gaussian distribution function.

Note that, in general, equation (27) must be evaluated numerically, except for the case $M = 2$ for which the exact expression is known to be $P_e = Q(\sqrt{2\mathcal{E}_s/N_0})$ with $Q(x) = 1 - \Phi(x)$ the complementary Gaussian cumulative distribution function.

5.3. Bit Error Probability

Sometimes, it is also desirable to convert the probability of a symbol error into the equivalent probability of a binary digit error. This is a welcome operation if we wish to compare performances of different PPM alphabets and is usually done by considering that M is a power of 2, say $M = 2^k$. So, provided that the M-ary symbol α is mapped into the k-bit word $\mathbf{d}_\alpha = (d_{\alpha,0}, \ldots, d_{\alpha,k-1})$, the bit error probability becomes

$$P_b = \sum_{\alpha \in \mathcal{A}_M} \mathrm{P}[s_n = \alpha] \sum_{\substack{\beta \in \mathcal{A}_M \\ \beta \neq \alpha}} \mathrm{P}[\hat{s}_n = \beta \mid s_n = \alpha] \frac{\mathrm{dist}(\mathbf{d}_\alpha, \mathbf{d}_\beta)}{\log_2 M}$$

where $\mathrm{dist}(\mathbf{d}_\alpha, \mathbf{d}_\beta)$ expresses the Hamming distance. Note that nondiagonal transition probabilities are independent on α and β, that is, $\mathrm{P}[\hat{s}_n = \beta \mid s_n = \alpha] = P_e/(M-1)$ for $\alpha \neq \beta$, because of the symmetric behavior of the channel. So, by further considering equally likely symbols, $\mathrm{P}[s_n = \alpha] = 1/M$, we obtain

$$P_b = \frac{2^{k-1}}{2^k - 1} P_e \tag{28}$$

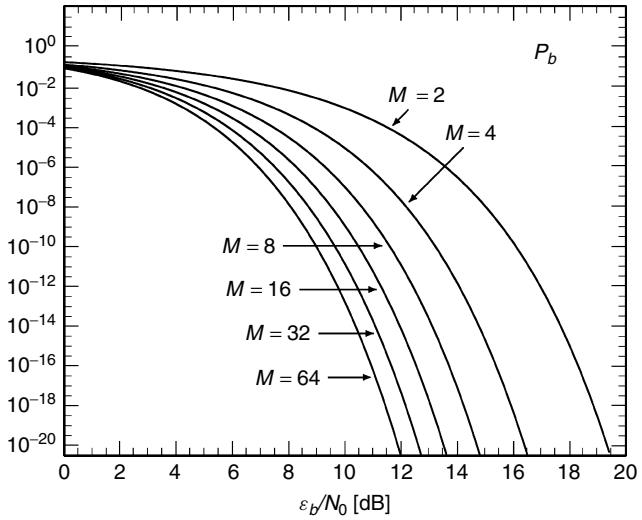


Figure 10. Bit error probability P_b versus SNR per bit $\mathcal{E}_b/N_0$ for different values of M.

that, for large values of M, reaches the limit $P_b \to P_e/2$.

In Fig. 10 we show bit error rate curves for various values of M as a function of the signal-to-noise ratio (SNR) per bit $\mathcal{E}_b/N_0$, where $\mathcal{E}_b = \mathcal{E}_s/\log_2 M$ is the energy-per-bit. Note that, for high values of the SNR, the bit error probability can be drastically reduced by increasing the cardinality M of the PPM alphabet. In particular, it can be demonstrated that as M approaches infinity, the probability of error approaches zero exponentially, provided that $\mathcal{E}_b/N_0 > \ln 2 (= -1.6\ \mathrm{dB})$.

6. INFORMATION RATES

Another fundamental measure in communications is given by the concept of *capacity* (or information rate) which expresses the maximum amount of information that can be sent through a channel with arbitrarily small error probability by use of appropriate coding techniques [5]. One of the main advantages of capacity is that it does not require us to specify the coding technique and represents a target to be (hopefully) met in the construction of error correcting codes. In the following section we investigate the capacity of digital PPM.

6.1. Formulation of Capacity

The basic concept in capacity evaluation is that of *entropy* $H(A)$ of a symbol source A, expressing the amount of information (in bits) carried by each of the symbols of the source. Entropy takes different expressions in dependence on the form of the source output (analog, digital, vectorial).

In digital PPM the source A outputs the digital symbol sequence s_n. We consider s_n to be stationary and with independent and identically distributed symbols with probabilities $p_A(\alpha) = \mathrm{P}[s_n = \alpha]$, in which case the entropy of the PPM source is

$$H(A) = \mathrm{E}[-\log_2 p_A(s_n)] = -\sum_{\alpha \in \mathcal{A}_M} p_A(\alpha) \log_2 p_A(\alpha) \quad [\text{bit/symbol}] \tag{29}$$

with the property $0 \le H(A) \le \log_2 M$, where $H(A) = \log_2 M$ if and only if the symbols are equally likely, that is $p_A(\alpha) = 1/M$.

The symbols s_n are then transmitted through the PPM channel, which includes modulation and demodulation. The output of the PPM channel is thus represented by an output source B that, in the case of optimal detection of each PPM pulse (as in the previous section), produces the symbol sequence $\hat{s}_n$. This approach is called *hard-detection*. Alternatively, one could use the correlation-measures $\varphi_n(\alpha)$ as weights in a Viterbi trellis, in which case we will talk of soft-detection and consider the real-valued vector sequence $\boldsymbol{\varphi}_n = (\varphi_n(0), \varphi_n(1), \ldots, \varphi_n(M-1))$ as the output of the source B. For ease of mathematical treatment, in the following we consider the hard-detection case.

The second quantity of interest is not represented by the entropy of B, rather on the conditional entropy $H(A \mid B)$ expressing the uncertainty on the output value of A once the output value of B is known (e.g., on the value of s_n once $\hat{s}_n$ is known). In hard-detection, the conditional entropy is expressed as

$$H(A \mid B) = \mathrm{E}[-\log_2 p_{A|B}(s_n \mid \hat{s}_n)] = -\sum_{\alpha,\beta \in \mathcal{A}_M} p_{AB}(\alpha, \beta) \log_2 p_{A|B}(\alpha \mid \beta) \qquad (30)$$

where we used the conditional probabilities $p_{A|B}(\alpha \mid \beta) = \mathrm{P}[s_n = \alpha \mid \hat{s}_n = \beta]$ and the joint probabilities $p_{AB}(\alpha, \beta) = \mathrm{P}[s_n = \alpha, \hat{s}_n = \beta]$.

From the source entropy $H(A)$ and the conditional entropy $H(A \mid B)$ it is customary to define the *average information flow* $I(A;B) = H(A) - H(A \mid B)$ expressing the difference between the information $H(A)$ carried by the source and the information $H(A \mid B)$ lost during transmission. In this context, capacity is defined as

$$C = \max_{p_A(\alpha)} I(A;B) = \max_{p_A(\alpha)} H(A) - H(A \mid B) \qquad \text{[bit/symbol]} \qquad (31)$$

that is the maximum value, taken with respect to the source symbol probabilities $p_A(\alpha), \alpha \in \mathcal{A}_M$, of the information flow. Note that, because we always have $H(A \mid B) \le H(A)$, capacity is a positive quantity.

When further taking into account that the symbol rate is $F = 1/T$, we can introduce the related concept of *information rate*

$$R = CF \qquad \text{[bit/s]} \qquad (32)$$

expressing the maximum amount of bit per second that can be sent through the channel with arbitrarily small error probability.

6.2. Information Rates for Digital PPM

In order to derive a compact expression for digital PPM capacity, it is appropriate to use the well-known identity for conditional probabilities $p_{A|B}(\alpha \mid \beta)p_B(\beta) = p_{AB}(\alpha, \beta) = p_{B|A}(\beta \mid \alpha)p_A(\alpha)$, where the meaning of the two new measures $p_{B|A}(\beta \mid \alpha)$ and $p_B(\beta)$ is obvious. With a little effort, this property lets us express the information flow as a function of the transition probabilities $p_{B|A}(\beta \mid \alpha)$ and of the source probabilities $p_A(\alpha)$, and we have

$$I(A;B) = \sum_{\alpha,\beta \in \mathcal{A}_M} p_{B|A}(\beta \mid \alpha)p_A(\alpha) \log_2 \frac{p_{B|A}(\beta \mid \alpha)}{p_B(\beta)} \qquad (33)$$

where

$$p_B(\beta) = \sum_{\gamma \in \mathcal{A}_M} p_{B|A}(\beta \mid \gamma)p_A(\gamma)$$

Moreover, from the results of the previous section we know that the transition probabilities satisfy

$$p_{B|A}(\beta \mid \alpha) = \begin{cases} 1 - P_e & \alpha = \beta \\ P_e/(M-1) & \alpha \ne \beta \end{cases}$$

where P_e is given by (27).

In the present situation, it can be proved that the maximization of (33) occurs when equally likely symbols are used, that is $p_A(\alpha) = 1/M$ in which case we also have $p_B(\beta) = 1/M$, and the capacity thus becomes

$$C = \log_2 M + P_e \log_2 \frac{P_e}{M-1} + (1 - P_e)\log_2(1 - P_e) \qquad (34)$$

while the information rate is expressed by $R = C/T$.

For a fair comparison between modulations employing different alphabet cardinalities M, it is convenient to assume that the source bit rate is fixed to $R_s = \log_2(M)/T$, and in Fig. 11 we show the normalized information rate R/R_s (efficiency) as a function of the bit-to-noise ratio $\mathcal{E}_b/N_0$. We note that for higher values of M the curves display higher efficiency and saturate faster. However, it is perhaps worth recalling that increased values of M correspond to a decrease of the pulse width T_0, and in fact $T_0 = \log_2(M)/(MR_s)$, in which case the available pulse technology could be a limiting factor for the choice of M.

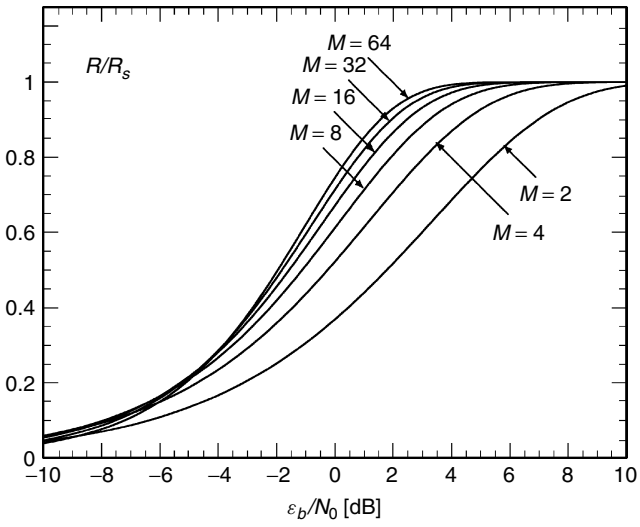


Figure 11. Efficiency R/R_s for hard-detected PPM versus SNR per bit $\mathcal{E}_b/N_0$ for different values of M.

7. APPLICATIONS (MAINLY IN WIRELESS OPTICAL COMMUNICATIONS)

PPM is perhaps the most widely used form of pulse modulation for its efficiency. We have to make a clear distinction between the analog era (the past) and the digital era (the present and the future).

In the past, analog PPM was normally used in the form of nonuniform sampling, also in connection with other modulations, since this approach resulted a simplified circuitry. To this regard the interested reader can refer to [2,3].

More recently, the digital PPM format found remarkable applications mainly in the field of optical communications, where the technology of generating very short pulses perfectly matches with the unipolar nature of PPM. It is worth recalling that PPM found some applications in fiber optics during the transition from copper-cable, where transmission required line encoding strategies with multilevel formats; in place of using multilevel PAM, in fiber optics it was more convenient to use multilevel PPM.

Nowadays, the interest on PPM is mainly restricted to wireless optical communications using infrared links, which require high average-power efficiency to minimize ocular hazards and power consumption. PPM is a technique that achieves very good power efficiency and it is largely used in these applications. For example, the Infrared Data Association (IrDA) has designated 4-PPM as the standard modulation technique for 4-Mb/s optical wireless links working over very short distances (1 m or less) [6]. These are envisioned to connect laptop computers, PDAs, palmtops, printers calculators, and mobile phones. NASA has proposed use of PPM in various free-space applications.

PPM also enters the IEEE 802.11 standard for infrared communications over local area networks (LANs), using 16-PPM schemes for 1 Mb/s links and 4-PPM for 2 Mb/s links. These are diffuse systems for which PPM is the technique that offers the best characteristics for transmission. We may thus emphasize the definitive success of PPM because of its use in many important standards in the digital format.

APPENDIX

Appendix A. Expression of PPM with Nonuniform Sampling

To prove (11) we can follow the sequence of operations of Fig. 3 with $\tilde{s}(t)$ replaced by $s(t)$, as requested in the case of nonuniform sampling.

The PDM signal can be written in the form

$$v_0(t) = u[s(t) - c(t)] \tag{35}$$

where $u(x)$ is the step function: $u(x) = 1$ for $x > 0$ and $u(x) = 0$ for $x < 0$. In fact, (35) gives 1 whenever $s(t) > c(t)$ and 0 otherwise. Considering that the derivative of the step function is the delta function, $du(x)/dx = \delta(x)$, the derivative of $v_0(t)$ is given by

$$v_1(t) = [s'(t) - c'(t)]\delta(s(t) - c(t))$$

Considering that $c(t) = (t - nT)S_0/T, t \in \mathcal{I}_n$ and recalling the property $\delta(\pm Ax) = (1/|A|)\delta(x)$, in the nth time slot we find

$$v_1(t) = [-1 + K\,s'(t)]\delta(t - nT - K\,s(t)), \quad t \in \mathcal{I}_n$$

The inverse half-wave rectification inverts the sign and we find

$$v_\delta(t) = |1 - K\,s'(t)|\delta(t - nT - K\,s(t)), \quad t \in \mathcal{I}_n$$

Hence, the expression of $v_\delta(t)$ for all t is

$$v_\delta(t) = |1 - K\,s'(t)| \sum_{n=-\infty}^{+\infty} \delta(t - nT - K\,s(t))$$

and (11) follows after use of the well-known identity

$$\sum_{n=-\infty}^{+\infty} \delta(t - nT) = F \sum_{n=-\infty}^{+\infty} e^{j2\pi nFt}.$$

Appendix B. Proof of the PSD Expression (19)

We first introduce the PPM expression in Eq. (14) in the correlation definition

$$\begin{aligned}\tilde{r}_v(t,\tau) &= \mathrm{E}[v(t)v^*(t+\tau)] \\ &= \sum_{m,k=-\infty}^{+\infty} \mathrm{E}[q(t - mT - s_m) \\ &\quad \times q^*(t + \tau - (m+k)T - s_{m+k})]\end{aligned}$$

where the conjugate is irrelevant (because the pulse function $q(t)$ is real valued) but useful in the following. By next expressing $q(t)$ as an inverse Fourier transform, $q(t) = \int Q(f)e^{j2\pi ft}df$, and by introducing the characteristic functions (17) and the function defined by (18), we obtain

$$\begin{aligned}\tilde{r}_v(t,\tau) &= \sum_{m,k=-\infty}^{+\infty} \mathrm{E}\left[\int df_1 \int df\; Q(f_1)Q^*(f)\right. \\ &\quad \left. \times e^{j2\pi[f_1(t-mT-s_m)-f(t+\tau-(m+k)T-s_{m+k})]}\right] \\ &= \sum_{m,k=-\infty}^{+\infty} \int df_1 \int df\; Q(f_1)Q^*(f)\Phi_s(-f_1, f; kT) \\ &\quad \times e^{j2\pi[f_1(t-mT)-f(t+\tau-(m+k)T)]} \\ &= \sum_{m=-\infty}^{+\infty} \int df_1 \int df\; Q(f_1)Q^*(f)\Psi_s(-f_1, f; f) \\ &\quad \times e^{j2\pi[f_1(t-mT)-f(t+\tau-mT)]}\end{aligned}$$

which clearly shows that $\tilde{r}_v(t,\tau)$ has period T in t. At this point we evaluate (15)

$$r_v(\tau) = \frac{1}{T}\int df_1 \int df\; Q(f_1)Q^*(f)\Psi_s(-f_1, f; f)A(f_1, f)e^{-j2\pi f\tau}$$

where

$$A(f_1,f)=\sum_{m=-\infty}^{+\infty}\int_0^T e^{j2\pi(f_1-f)(t-mT)}\,dt$$
$$=\int_{-\infty}^{+\infty}e^{j2\pi(f_1-f)u}\,du=\delta(f_1-f)$$

The presence of this delta function allows us to remove the integral with respect to f_1 setting $f_1=f$ elsewhere. Hence,

$$r_v(\tau)=\frac{1}{T}\int df\; Q(f)Q^*(f)\Psi_s(-f,f;f)e^{-j2\pi f\tau}$$

which expresses $r_v(\tau)$ as the inverse Fourier transform of the quantity defined in Eq. (19). Because the inverse Fourier transform is unique, the proof is complete.

Appendix C. Mean and Covariances of $\varphi_a(nT)$

We derive mean and cross-correlation for the Gaussian random variables $\eta_n(\alpha), \alpha\in\mathcal{A}_M$ of (24). The mean value gives

$$\mathrm{E}[\eta_n(\alpha)]=\mathrm{E}\left[\int \eta(t)p(t-nT-\alpha)\,dt\right]$$
$$=\int \mathrm{E}[\eta(t)]p(t-nT-\alpha)\,dt=0$$

because $\eta(t)$ is a zero-mean Gaussian process. This proves the first of Eq. (25) since the expected signal term in Eq. (24) is a deterministic quantity. Covariance of (24) is instead completely determined by the noise terms $\eta_n(\alpha)$ and we have

$$\mathrm{E}[\eta_n(\alpha)\eta_n(\beta)]=\mathrm{E}\left[\int \eta(t)p(t-nT-\alpha)\right.$$
$$\left.\times dt\int \eta(x)p(x-nT-\beta)\,dx\right]$$
$$=\int\int \mathrm{E}[\eta(t)\eta(x)]p(t-nT-\alpha)$$
$$\times p(x-nT-\beta)\,dt\,dx$$
$$=\frac{N_0}{2}\int\int \delta(t-x)p(t-nT-\alpha)$$
$$\times p(x-nT-\beta)\,dt\,dx$$
$$=\frac{N_0}{2}\int p(t-nT-\alpha)p(t-nT-\beta)\,dt \tag{36}$$

where we used the property $\mathrm{E}[\eta(t)\eta(x)]=\frac{1}{2}N_0\delta(t-x)$. By further considering that pulses in the last of (36) are non-colliding for $\alpha\neq\beta$, the Gaussian random variables $\eta_n(\alpha)$ become statistically uncorrelated (hence independent) and the second of (25) follows straightforwardly.

BIOGRAPHIES

Gianfranco Cariolaro (M'66) was born in 1936. He received the degree in Electrical Engineering from the University of Padova, Italy, in 1960, and the Libera Docenza degree in Electrical Communications in 1968 from the same university. He was appointed full professor in 1975 and is currently Professor of Electrical Communications and Signal Theory at the University of Padova. His main research is in the fields of data transmission, images, digital television, multicarrier modulation systems (OFDM), cellular radios, deep space communications, and the fractional Fourier transform. He is author of several books including "Unified Signal Theory" (Torino, Italy: UTET, 2nd Edition, 1996).

Tomaso Erseghe was born in Valdagno, Italy, in 1972. He received the laurea degree in Telecommunication Engineering from the University of Padova, Italy, in 1996, with a thesis on the fractional Fourier transform, and the Ph.D. in Telecommunication Engineering from that same university, in 2002, with a thesis on ultra-wide-band communications. From 1997 to 1999 he worked as an R&D Engineer at Snell & Wilcox a British broadcast equipment manufacturer, in the areas of image restoration and motion compensation. He is now a Post Doc at the University of Padova. His research interests include fractional Fourier transforms, the theory of symmetries with application to the DFT, ultra-wide-band impulse radio, time-hopping constructions. He has also been involved in some research projects sponsored by the European Community.

BIBLIOGRAPHY

1. *Transmission Systems for Communications*, Bell Telephone Laboratories Inc., 4th ed., Feb., 1970.
2. H. E. Rowe, *Signals and Noise in Communication Systems*, D. Van Nostrand Company, Princeton, New Jersey, 1965.
3. H. Schwartz, W. R. Bennet, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill Inc., New York, 1966.
4. J. G. Proakis, *Digital Communications* 3rd ed., McGraw-Hill, New York, 1995.
5. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (July, Oct. 1948).
6. IrDA, Serial Infrared Link Access Protocol (IrLAP)—Version 1.1, 1996.
7. IEEE Std. 802.11, IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, September 1999.

Q

QUADRATURE AMPLITUDE MODULATION

ISRAEL KORN
University of New South Wales
Sydney, Australia

JOHN P. FONSEKA
University of Texas at Dallas
Richardson, Texas

1. INTRODUCTION

Quadrature amplitude modulation (QAM), or quadrature amplitude shift keying (QASK), is a linear method of digital modulation in which M-ary symbols are transmitted by varying the amplitude of two carriers in quadrature. QAM requires coherent detection and linear amplifiers. QAM has an excellent bandwidth efficiency (i.e., the ratio of bit rate to occupied bandwidth is high), a moderate energy efficiency (i.e., moderate values of energy-to-noise ratio per bit for a given bit error probability), and a high degree of complexity at the receiver because there is a need to track the amplitude, phase, and frequency of the carrier as well as the clock of the symbols. QAM is used mainly in modems over telephone lines but its application is growing in satellite communications, coaxial cables and line-of-sight microwave systems. Numerous papers have been published about various aspects of QAM. Reference 1 is a book totally devoted to QAM. Most books (both undergraduate and graduate level) on digital communications contain chapters of various degree of depth on QAM. A partial list of these books is presented in Refs. 2–7. The first paper about QAM was published in September 1960 and at the time of writing this article (January 2002), papers on this topic still appear in the professional literature.

2. QAM SYSTEM

A model of a typical QAM system is shown in Fig. 1.

A sequence of M-ary independent and equiprobable symbols ($M = 2^\mu$, $\mu =$ integer) $\boldsymbol{a} = (a_0 a_1 \ldots)$ is generated by a source at the rate R bauds (symbols/s) so that symbol a_i is produced at time $t = iT$, where $T = 1/R$. Symbol a_i takes values from the set

$$\Omega = \{a(m): m = 1, 2, \ldots, M\} \tag{1}$$

In QAM the symbols are two-dimensional or complex

$$a(m) = a_I(m) + j a_Q(m) \tag{2}$$

where subscripts I and Q denote the in-phase (real) and quadrature-phase (imaginary) components. The symbol set forms a symbol constellation. Several typical but simple symbol constellations are shown in Fig. 2 for $M = 8, 12, 16$. These are examples of (a) square, (b) rectangular, (c) cross, and (d) star constellations.

For rectangular constellations with $M = M_I M_Q$, $a_x(m) = \pm 1, \pm 3, \pm(M_x - 1)$, $x = I, Q$, and for square constellations $M_I = M_Q = \sqrt{M}$. At the destination, we receive the sequence $\hat{\mathbf{a}} = (\hat{a}_0 \hat{a}_1 \ldots)$ where $\hat{a}_i \in \Omega$ and $\hat{a}_i$ may differ from a_i. Hence, the symbol error probability (SEP) is

$$P(e) = P(\hat{a}_i \neq a_i) \tag{3}$$

Each symbol represents μ bits. For example, if $\mu = 4$, the 4 bits represented by a_i are $(b_{\mu i+1}, b_{\mu i+2}, b_{\mu i+3}, b_{\mu i+4})$, $b_j \in \{0, 1\}$. The bit error probability (BEP) is

$$P_b(e) = P(\hat{b}_j \neq b_j) \tag{4}$$

and the bit rate is

$$R_b = \frac{R}{\mu} \tag{5}$$

The BEP and SEP are related by the particular representation or mapping between the symbols and bit sequences. For example, in Gray coding two neighboring symbols in the symbol constellation differ by only 1 bit. In Fig. 3 we illustrate for the case of $M = 4$ two mappings of which (a) is a Gray code and (b) is not. The aim in designing a communication system is to minimize $P(e)$ and $P_b(e)$ when all other factors remain equal.

More complicated and optimal symbol constellation can be found in Refs. 8 and 9. By *optimal* we mean the one producing the minimum SEP or BEP for a given average energy-to-noise ratio. In the process of transmitting the symbols through a physical channel, we impose the sequence of symbols $\boldsymbol{a}$ on a signal that can pass through the channel. For a bandpass channel with transfer function $\tilde{H}_C(f)$ and bandwidth $\tilde{B}_c$ around the frequency f_c (illustrated in Fig. 4 for an ideal channel with rectangular frequency response) the impulse response is

$$\tilde{h}_C(t) = \mathrm{Re}\{2h_C(t)e^{jw_c t}\} = 2h_{CI}(t)\cos(w_c t) - 2h_{CQ}(t)\sin(w_c t) \tag{6}$$

where Re { } denotes the real part of the term in the braces, $\omega_c = 2\pi f_c$ and

$$h_C(t) = h_{CI}(t) + j h_{CQ}(t) \tag{7}$$

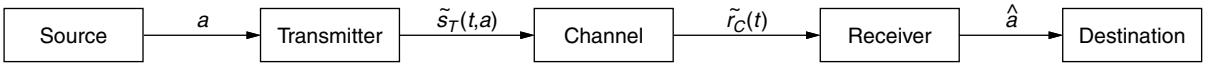


Figure 1. Typical QAM system.

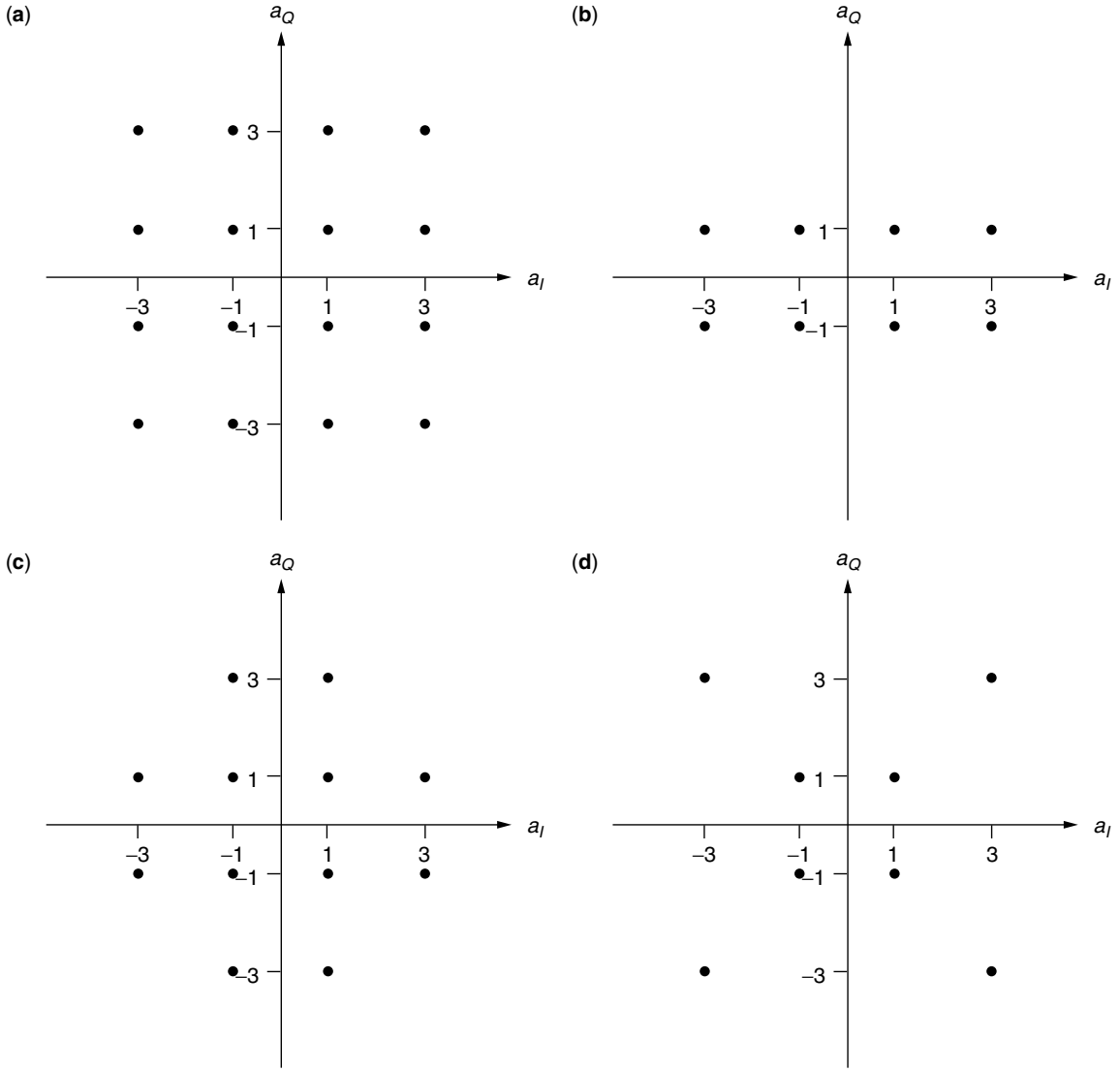


Figure 2. QAM symbol constellations: (**a**) square; (**b**) rectangular; (**c**) cross; (**d**) star.

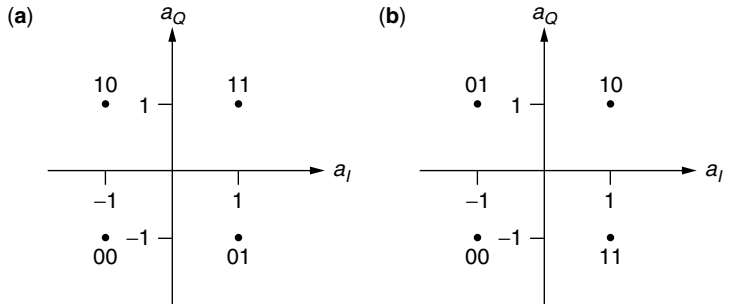


Figure 3. (**a**) Gray code; (**b**) arbitrary code.

is called the *baseband equivalent* or *complex envelope* of $\tilde{h}_C(t)$ with transfer function

$$H_C(f) = \tilde{H}_C(f+f_c)u(f+f_c), \quad u(f) = \begin{cases} 1 & f \geq 0 \\ 0 & f < 0 \end{cases} \tag{8}$$

and bandwidth $B_C = \tilde{B}_C/2$.

The terms $H_c(f)$ and $h_c(t)$ form a Fourier transform pair. Similar notation will be used for other bandpass filters, signals, and noise. Since the channel is bandpass the transmitted signal has to be also bandpass, and is generated in three stages as shown in Fig. 5.

At the first stage, we produce a bandpass signal with a certain desirable baseband shaping pulse, $h_S(t)$ and

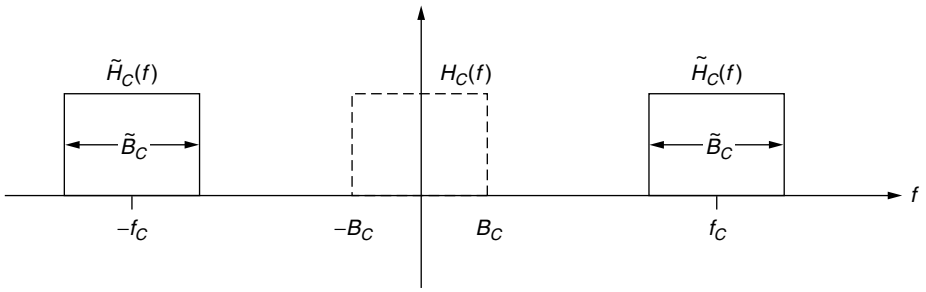


Figure 4. Transfer function of bandpass and baseband equivalent channel.

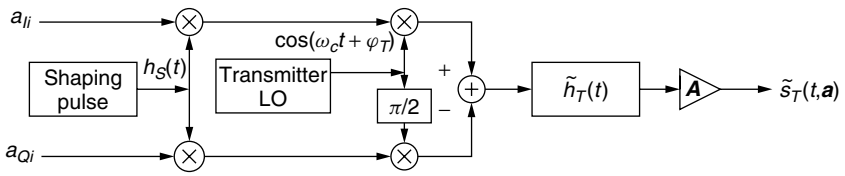


Figure 5. Transmitter of QAM.

two carriers in quadrature $\cos(w_c t + \varphi_T)$, $\sin(w_c t + \varphi_T)$ with carrier frequency f_c and phase φ_T generated by the transmitter local oscillator. This signal is

$$\begin{aligned}\tilde{s}_S(t,\boldsymbol{a}) &= \sum_i [a_{Ii} h_S(t - iT)\cos(w_c t + \varphi_T) \\ &\quad - a_{Qi} h_S(t - iT)\sin(w_c t + \varphi_T)] \\ &= \mathrm{Re}\left\{\sum_i a_i h_S(t - iT) e^{j(w_c t + \varphi_T)}\right\} \end{aligned} \tag{9}$$

with complex envelope

$$s_S(t,\boldsymbol{a}) = \sum_i a_i h_S(t - iT) e^{j\varphi_T} \tag{10}$$

The frequency upconversion from baseband to bandpass is called *modulation*. At the second stage the signal in (9) is filtered by the transmitter bandpass filter with impulse response $\tilde{h}_T(t)$. At the third stage the signal is amplified by a linear amplifier to the required power level. The total amplification of the transmitter is combined into one amplitude A. The transmitted signal is similar to those in (9) and (10) with $H_S(f)$ replaced by the transmitter transfer function $G_T(f) = H_S(f)H_T(f)$ and subscript S replaced by T.

The selection of the pulse shape is of paramount importance in the design of the communication system because it determines the power spectral density (PSD) and hence the bandwidth of transmitted signal. It can be shown [1,2] that the PSD of the signal in (10) is

$$\begin{aligned} S_S(f) &= 0.5A^2\sigma_a^2 |H_S(f)|^2, \\ \sigma_a^2 &= \frac{1}{M}\sum_{m=1}^{M} |a(m)|^2 = \frac{2}{3}(M-1) \end{aligned} \tag{11}$$

where the last equality represents a square constellation. Similarly, the PSD of the signal in (9) is a shifted version of (11) centered at $\pm f_c$.

We shall see later that in order to eliminate intersymbol interference (ISI), which increases the SEP and BEP, only certain pulses (called *Nyquist pulses*) are desirable. Many Nyquist pulses differ in duration and bandwidth. For example a rectangular pulse of duration T

$$h_s(t) = u_T(t) = u(t) - u(t - T) \tag{12}$$

is simple to generate however the resulting bandwidth is very large and more severe filtering is required by $h_T(t)$. In fact the 99% bandwidth, B_{99} of this pulse defined by

$$\int_0^{B_{99}} |H_S(f)|^2\, df = 0.99 \int_0^{\infty} |H_S(f)|^2\, df \tag{13}$$

and that contains 99% of the energy or power of the signal is $B_{99} = 8.65R$. On the other hand the infinite duration pulse $h_S(t)$ (called the raised cosine in frequency) with

$$H_S(f) = \begin{cases} 1 & 0 \le |f| \le (1-\alpha)R/2 \\ \cos\left(\dfrac{\pi}{4\alpha}\left(\left|\dfrac{2f}{R}\right| - 1 + \alpha\right)\right) & (1-\alpha)R/2 \le |f| \\ & \quad \le (1+\alpha)R/2 \\ 0 & (1+\alpha)R/2 \le |f| \le \infty \end{cases} \tag{14}$$

has a 100% bandwidth

$$B = B_{100} = \frac{(1+\alpha)R}{2} \tag{15}$$

where the parameter $0 \le \alpha \le 1$ determines the excess bandwidth beyond the minimum bandwidth of $R/2$. These two pulses are shown in Fig. 6.

For the raised-cosine pulse, no additional filtering is required as long as $B \le B_c$. In designing the system we usually select $B = B_c$ unless guard bands [to reject adjacent-channel interference (ACI) signals with carrier frequencies $f_c \pm \Delta f_c$] are required.

The bandwidth efficiency is defined by the ratio of bit rate and bandpass bandwidth:

$$\eta = \frac{R_b}{\tilde{B}} = \frac{\mu R}{2B} = \frac{\mu}{1+\alpha} \tag{16}$$

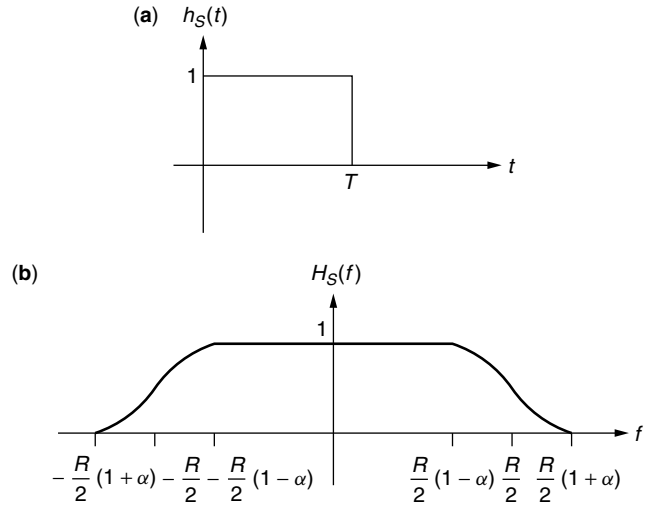


Figure 6. Nyquist pulses: (**a**) rectangular pulse; (**b**) raised cosine pulse in frequency domain with bandwidth $(1+\alpha)R/2$.

Thus we increase the bandwidth efficiency by decreasing α. In practice values of $\alpha \geq 0.15$ are achievable. The bandwidth efficiency as a function of α for $\mu = 2, 4, 6, 8, 10$ (corresponding to $M = 4, 16, 64, 256, 1024$, respectively) is shown in Fig. 7.

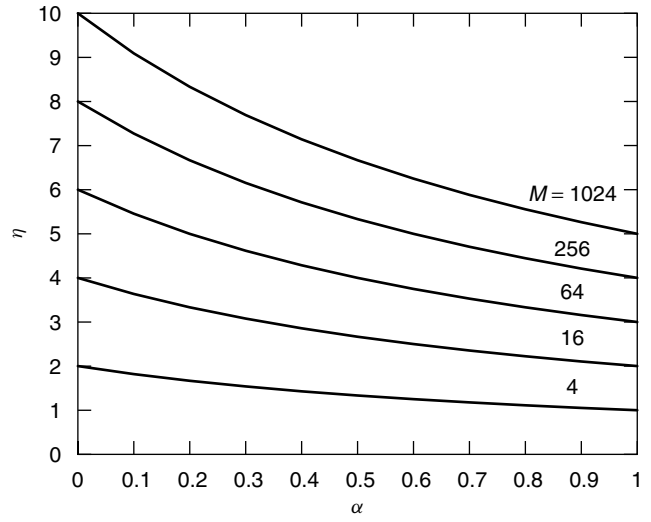


Figure 7. The bandwidth efficiency of QAM as a function of excess bandwidth.

There is a version of QASK called OQASK (offset QASK) or SQASK (staggered QASK) in which the quadrature term in (9) is delayed by $T/2$ [instead of $h_S(t-iT)$, we have $h_S(t-iT-T/2)$]. There is a whole range of shaping pulses, $h_S(t)$ for which we obtain in OQASK with $M=4$ a constant envelope signal for all time: $|\tilde{s}_T(t,\boldsymbol{a})| = c$. This enables the usage of nonlinear amplifiers, which are more efficient than linear amplifiers. In QASK there is only one shaping pulse, namely, a rectangular pulse that gives a constant envelope; however, the resulting bandwidth is large. In OQASK there are many pulses available with a reduced bandwidth. The application of nonlinear amplifiers to QAM with $M > 4$ has been presented in many papers and a comprehensive review can be found in Chap. 7 of Ref. 3.

The channel, a model of which is shown in Fig. 8, is composed of the channel filter, $\tilde{h}_C(t)$ and additive white Gaussian noise (AWGN), $\tilde{n}_C(t)$ with PSD $N_0/2$ for all frequencies. Such a channel is called a AWGN channel.

The channel filter takes into account the channel attenuation and phase shift. The channel output is thus

$$\tilde{r}_C(t) = \tilde{s}_C(t,\boldsymbol{a}) + \tilde{n}_C(t) = \mathrm{Re}\left\{A\sum_i a_i g_C(t-iT)e^{j(w_c t+\varphi_T)}\right\} + \tilde{n}_C(t), \quad G_C(f) = G_T(f)H_C(f) \tag{17}$$

The receiver, a model of which is shown in Fig. 9, is also composed of several stages.

First there is a receiver bandpass filter, $\tilde{h}_R(t)$, whose main task is to eliminate the noise beyond the signal bandwidth and to reduce ACI or spurious interference. The output of this filter is

$$\begin{aligned}\tilde{r}_R(t) &= [\tilde{s}_C(t,\boldsymbol{a}) + \tilde{n}_C(t)] * \tilde{h}_R(t) = \tilde{s}_R(t,\boldsymbol{a}) + \tilde{n}_R(t)\\ &= \mathrm{Re}\left\{\left[A\sum_i a_i g_{CR}(t-iT) + n_R(t)\right]e^{j(w_c t+\varphi_T)}\right\}\\ &= [s_{RI}(t,\boldsymbol{a}) + n_{RI}(t)]\cos(w_c t+\varphi_T)\\ &\quad - [s_{RQ}(t,\boldsymbol{a}) + n_{RQ}(t)]\sin(w_c t+\varphi_T),\\ G_{CR}(f) &= G_C(f)H_R(f)\end{aligned} \tag{18}$$

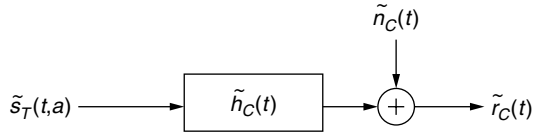


Figure 8. Model of AWGN with filter.

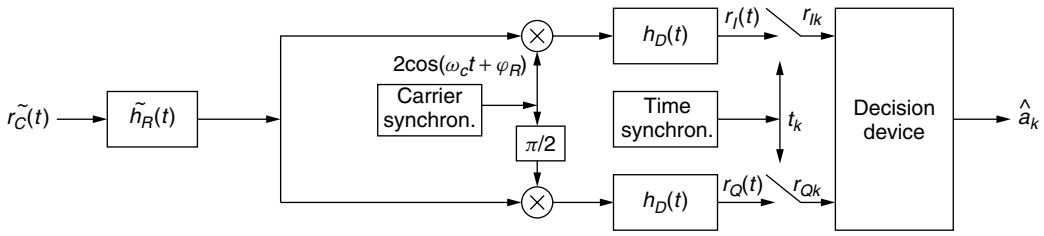


Figure 9. Receiver of QAM.

where $n_R(t)$ is the complex envelope of $\tilde{n}_R(t)$ and is a zero mean complex baseband Gaussian noise process with PSD

$$S_{n_R}(f) = N_0|H_R(f)|^2 \tag{19}$$

In the second stage the output of the receiver filter is multiplied by two carriers in quadrature, $2\cos(w_c t + \varphi_R)$, $2\sin(w_c t + \varphi_R)$ generated by the receiver local oscillator where φ_R tracks φ_T. There is a random phase error

$$\Delta\varphi = \varphi_T - \varphi_R \tag{20}$$

which also takes into account a carrier frequency error. This receiver is called a coherent receiver because it tracks the phase and frequency of the transmitter. Since

$$\begin{aligned} &2\cos(w_c t + \varphi_T)\cos(w_c t + \varphi_R) \\ &\quad = \cos\Delta\varphi + \cos\ (2w_c t + \varphi_T + \varphi_R) \\ &2\sin(w_c t + \varphi_T)\sin(w_c t + \varphi_R) \\ &\quad = \cos\Delta\varphi - \cos\ (2w_c t + \varphi_T + \varphi_R) \\ &2\sin(w_c t + \varphi_T)\cos(w_c t + \varphi_R) \\ &\quad = \sin\Delta\varphi + \sin\ (2w_c t + \varphi_T + \varphi_R) \end{aligned} \tag{21}$$

the multiplier output is the sum of a baseband signal and a bandpass signal with a carrier frequency of $2f_c$, which is easily eliminated by the baseband demodulator filter, $h_D(t)$. The frequency downconversion from bandpass to baseband is called *demodulation* or *detection* and is the inverse of the modulation. The outputs of the detector filters are

$$\begin{aligned} r_I(t) &= s_I(t,\boldsymbol{a}) + n_I(t) = \mathrm{Re}\left\{A\sum_i a_i g(t - iT)e^{j\Delta\varphi}\right\} + n_I(t) \\ r_Q(t) &= s_Q(t,\boldsymbol{a}) + n_Q(t) = \mathrm{Im}\left\{A\sum_i a_i g(t - iT)e^{j\Delta\varphi}\right\} + n_Q(t) \end{aligned} \tag{22}$$

where the transfer function of $g(t)$

$$G(f) = G_R(f)H_D(f) = H_S(f)H_T(t)H_C(f)H_R(f)H_D(f) \tag{23}$$

is the combined effect of all filters in the system on the shaping pulse and $n_I(t)$, $n_Q(t)$ are zero mean, real, baseband, independent, Gaussian noises with identical PSDs

$$S_n = N_0|G_R(f)|^2, \quad G_R(f) = H_R(f)H_D(f) \tag{24}$$

and power or variance

$$P_n = \sigma_n^2 = N_0\int_{-\infty}^{\infty}|G_R(f)|^2\,df \tag{25}$$

In (22) we left the noise unchanged by $\Delta\varphi$ because $n(t)$ and $n(t)e^{j\Delta\varphi}$ are identical, Gaussian processes. The signal in (22)

$$r(t) = r_I(t) + jr_Q(t) = A\sum_i a_i g(t - iT)e^{j\Delta\varphi} + n(t) \tag{26}$$

is sampled at times $t_k = t_0 + kT$. The timing is generated by a time synchronizer from the incoming signal therefore random variations in sampling time are expected. From $r(t_k)$ the decision circuit produces an estimate of symbol a_k, $\hat{a}_k$. We can obtain the result of (26) from the baseband equivalent block diagram shown in Fig. 10, which is composed only of baseband filters and in which the carrier frequency is irrelevant. It can be shown that the receiver presented in Fig. 9 is an optimal receiver provided $g(t)$ is a Nyquist pulse and $g_R(t)$ is a filter matched to $g_C(t)$, namely, $g_R(t) = g_C^*(t_0 - t)$, where x^* denotes the complex conjugate of x.

3. COMPUTATION OF ERROR PROBABILITY

3.1. Computation of SEP

Using the notation

$$r = r(t_k), \quad s = s(t_k,\boldsymbol{a}), \quad n = n(t_k), \quad g_k = g(t_k) \tag{27}$$

we obtain from (26)

$$r = s + n, \quad s = Ag_o e^{j\Delta\varphi}a_k + \mathrm{ISI}, \quad \mathrm{ISI} = A\sum_{i\neq 0} a_{k-i}g_i e^{j\Delta\varphi} \tag{28}$$

Under ideal conditions both $\Delta\varphi = 0$ and $\mathrm{ISI} = 0$. We may also assume that g_0 is real and nonnegative and there is a matched filter; hence

$$r = s + n, \quad s = A_0 a_k, \quad A_0 = Ag_0, \quad s(m) = A_0 a(m) \tag{29}$$

In deciding which symbol is transmitted, we divide the plane of $r = r_I + jr_Q$ into M nonoverlapping regions $\{R_m\}$

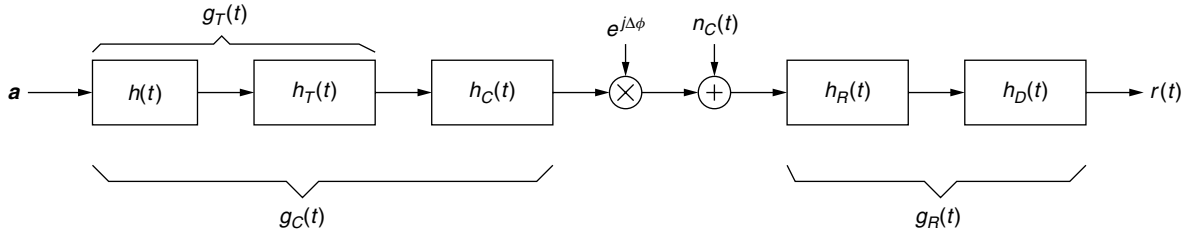


Figure 10. Baseband equivalent of QAM system.

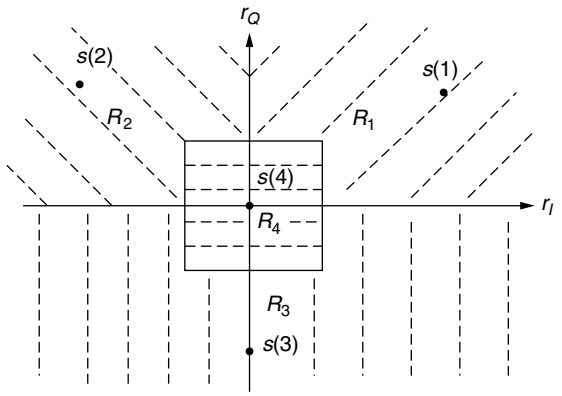


Figure 11. Arbitrary regions of the r plane.

such that if $r \in R_m$ the decision is $\hat{a}_k = a(m)$. This is illustrated in Fig. 11.

The conditional error probability that $\hat{a}_k = a(\hat{m})$ given that $a_k = a(m)$ is computed from

$$P(\hat{a}_k = a(\hat{m}) \mid a_k = a(m)) = P(r \in R_{\hat{m}} \mid a_k = a(m)) = P(\{s(m) + n\} \in R_{\hat{m}}) \tag{30}$$

For equiprobable symbols the SEP is

$$P(e) = P(\hat{a}_k \neq a_k) = \frac{1}{M} \sum_m \sum_{\hat{m} \neq m} P(\hat{a}_k = a(\hat{m}) \mid a_k = a(m)) \tag{31}$$

It can be shown that for optimum decisions we select $r \in R_m$ only if $|r - A_0 a(m)|$ is minimum. These optimum regions are illustrated for two constellations in Fig. 12.

For the square constellations the decision regions are

$$R_m = \{r : s_I(m) - C \leq r_I \leq s_I(m) + C, s_Q(m) - C \leq r_Q \leq s_Q(m) + C\} \tag{32}$$

where C is either A_0 or ∞ depending on whether the regions are finite in both dimensions, finite in one direction or infinite in both dimensions. This is illustrated in Fig. 12a. For square QAM the SEP is [2,5–7]

$$P(e) = 4\left(1 - \frac{1}{\sqrt{M}}\right) Q\left(\sqrt{\frac{3}{M-1}\gamma}\right) - 4\left(1 - \frac{1}{\sqrt{M}}\right)^2 Q^2\left(\sqrt{\frac{3}{M-1}\gamma}\right) \tag{33}$$

where

$$\gamma = \frac{E}{N_0}, \quad E = \frac{A^2 \sigma_a^2 g_0^2}{2} \tag{34}$$

is the energy-to-noise ratio per symbol and

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-0.5y^2}\, dy = \frac{1}{\pi} \int_0^{\pi/2} e^{-0.5x^2/\sin^2\phi}\, d\phi \tag{35}$$

is the standard Q function illustrated in Fig. 13 with lower and upper bounds. $Q^2(x)$ in (33) can be computed from (35) with $\pi/2$ replaced by $\pi/4$ in the limit of the second integral.

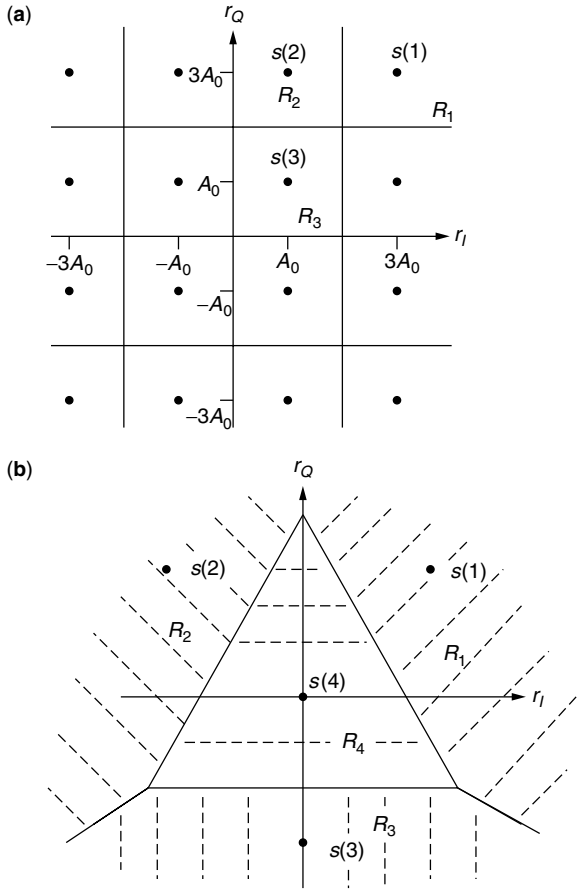


Figure 12. Optimal decision regions: (**a**) square constellation for $M = 16$; (**b**) signal constellation as in Fig. 11.

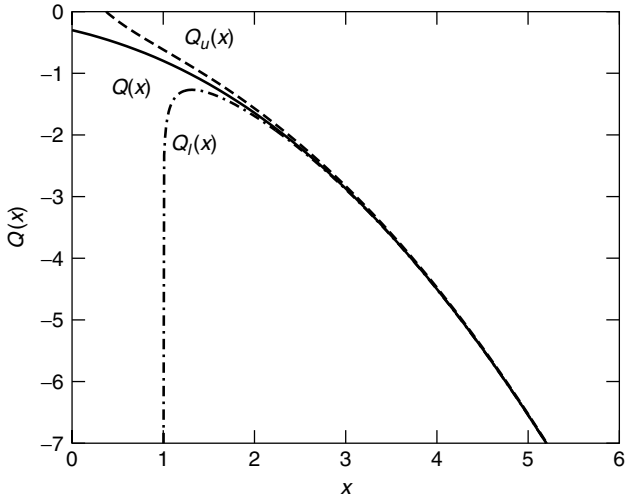


Figure 13. The variations of $Q(x)$, $Q_u(x) = \frac{1}{\sqrt{2\pi x^2}} e^{-(x^2/x)}$ and $Q_1(x) = \frac{1}{\sqrt{2\pi x^2}} \left(1 - \frac{1}{x^2}\right) e^{-(x^2/2)}$.

The SEP as a function of γ is shown in Fig. 14.

The square constellation is optimal for $M = 4$ and very close to optimal [8,9] for other M. For example, to achieve

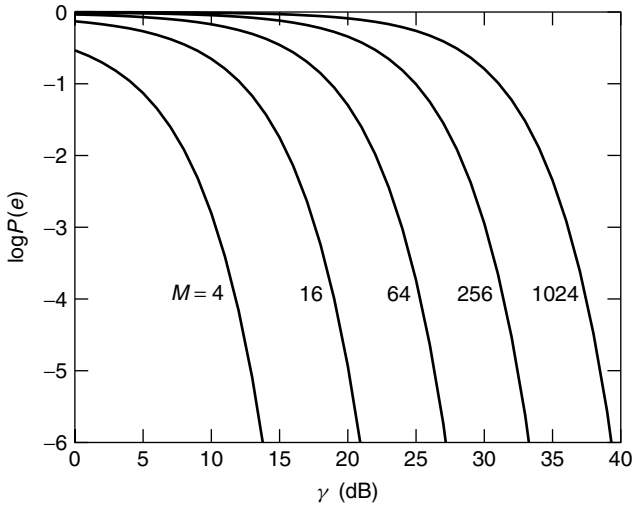


Figure 14. The SEP as a function of energy-to-noise ratio per symbol.

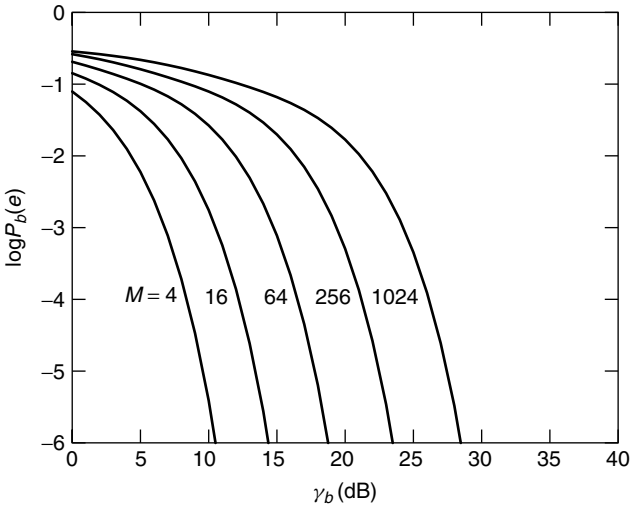


Figure 15. The BEP as a function of energy-to-noise ratio bit.

a SEP of 10^{-5}, γ increases by only 0.1 dB for $M = 16$ and by 0.4 dB for $M = 64$ relative to optimum constellation [2].

3.2. Computation of BEP

When computing the BEP, we have to take into account the mapping between symbols and bits. Let $w_{m\hat{m}}$ be the number of bits by which symbols $a(m)$ and $a(\hat{m})$ differ. Thus, even if the symbols are in error only $w_{m\hat{m}}$ out of the μ bits are in error. The BEP is

$$P_b(e) = \frac{1}{M}\sum_m \sum_{\hat{m}\neq m} \frac{w_{m\hat{m}}}{\mu} P(\hat{a}_k = a(\hat{m}) \mid a_k = a(m)) \geq \frac{P(e)}{\mu} \tag{36}$$

and the lower bound is obtained by substituting $w_{m\hat{m}} = 1$. For square constellations and Gray coding, we can compute $P_b(e)$ precisely:

$$P_b(e) = \sum_{i=1}^{I_M} c_i Q(\sqrt{d_i \gamma_b}), \quad \gamma_b = \frac{E_b}{N_0} \tag{37}$$

For example, for $M = 4$, $I_M = c_1 = 1$, $d_1 = 2$ and for $M = 16$, $I_M = 3$, $c_1 = 0.75$, $c_2 = 0.5$, $c_3 = -0.25$, $d_1 = 0.8$, $d_2 = 7.2$, $d_3 = 20$. An excellent approximation is given by [16]

$$P_b(e) \cong 4\left(1 - \frac{1}{\sqrt{M}}\right) \frac{Q\left(\sqrt{\frac{3\mu\gamma_b}{M-1}}\right) + Q\left(\sqrt{\frac{27\mu\gamma_b}{M-1}}\right)}{\mu} \tag{38}$$

In Fig. 15 we show the BEP as a function of γ_b.

3.3. Optimal Receiver and Transmitter Filters

It follows from (29) that there is no ISI if

$$g_i = \begin{cases} g_0 \neq 0 & i = 0 \\ 0 & i \neq 0 \end{cases} \tag{39}$$

where using the Inverse Fourier transform and assuming $t_0 = 0$ (there is no loss in generality in that because the delay t_0 can be introduced in the final stage so that the matched filter is physically realizable)

$$\begin{aligned} g_i = g(iT) &= \int_{-\infty}^{\infty} G(f) e^{j2\pi fiT}\, df = \sum_{k=-\infty}^{\infty} \int_{-f_N+kR}^{f_N+kR} G(f) e^{j2\pi fiT}\, df \\ &= \sum_{k=-\infty}^{\infty} \int_{-f_N}^{f_N} G(f' + kR) e^{j2\pi(f'+kR)iT}\, df' \\ &= \int_{-f_N}^{f_N} G_{\sum}(f) e^{j2\pi fiT}\, df \end{aligned} \tag{40}$$

where $f_N = R/2$, $R = 1/T$, and

$$G_{\sum}(f) = \sum_k G(f + kR) = T \sum_{i=-\infty}^{\infty} g_i e^{-j2\pi fiT} \tag{41}$$

is a periodic function of f with period R. Thus, if there is no ISI, then

$$G_{\sum}(f) = g_0 T \tag{42}$$

A function $G(f)$ which satisfies this (Nyquist) condition is called a *Nyquist function* $G_N(f)$. No function $G(f)$ with bandwidth less than $R/2$ can satisfy (42). The raised-cosine family with bandwidth $f_N(1+\alpha)$ satisfies this condition [2]. The square root of $G_N(f)$ is denoted by

$$H_N(f) = \sqrt{G_N(f)} \tag{43}$$

The pulse in (14) is $H_N(f)$ for the raised cosine with $g_0 T = 1$.

For a matched filter we can select

$$\begin{aligned} G_C(f) &= H_C(f) G_T(f) = H_N(f), \\ G_R(f) &= H_R(f) H_D(f) = H_N(f) \end{aligned} \tag{44}$$

Knowing one of the filters in the receiver or transmitter, we can find the other filters, which are now the optimal filters. In deriving (44) (although not stated explicitly), we

optimized the filters under the condition of fixed energy, E, at receiver input. If we optimize the filters with the condition of fixed energy at the transmitter the results are different [7] if $H_C(f) \neq 1$ in the range $|f| \leq R_N(1+\alpha)$.

4. QAM IN FADING CHANNELS

There are many fading channels, particularly in mobile communications. The simplest fading channel is a flat and slow fading channel [6]. In such a channel the random channel filter does not distort the signal and the variations in the channel are slow relative to the symbol rate. In such a channel only the amplitude A and phase φ_T are random. Since A is random, the energy-to-noise ratios and their square roots

$$\gamma = \frac{E}{N_0}, \quad \gamma_b = \frac{E_b}{N_0}, \quad \alpha = \sqrt{\gamma}, \quad \alpha_b = \sqrt{\gamma_b} \tag{45}$$

are random variables with probability density function (PDF) $p_x(x)$, where $x = \gamma, \gamma_b, \alpha, \alpha_b$. The SEP and BEP computed in Section 3 are based on the assumption that the receiver tracks the amplitude and phase and knows their values. Without knowledge of A, for example, the decision regions $\{R_m\}$ are useless. Thus the formulas for the SEP and BEP in Section 3 are conditional on the value of γ and γ_b, $P_S(e \mid \gamma)$, and $P_b(e \mid \gamma_b)$. In fading channels these formulas have to be averaged over γ and γ_b.

4.1. The Error Probability without Diversity

The SEP follows from (32) and (33)

$$\begin{aligned} P(e) &= \int_0^\infty P(e \mid \gamma) p_\gamma(\gamma)\, d\gamma = 4C_1 \int_0^\infty Q(\sqrt{C_2\gamma}) p_\gamma(\gamma)\, d\gamma \\ &\quad - 4C_1^2 \int_0^\infty Q^2(\sqrt{C_2\gamma}) p_\gamma(\gamma)\, d\gamma \\ &= \frac{4C_1}{\pi} \int_0^\infty \int_0^{\pi/2} e^{-0.5C_2\gamma/\sin^2\phi} p_\gamma(\gamma)\, d\gamma\, d\phi \\ &\quad - \frac{4C_1^2}{\pi} \int_0^\infty \int_0^{\pi/4} e^{-0.5C_2\gamma/\sin^2\phi} p_\gamma(\gamma)\, d\gamma\, d\phi \end{aligned} \tag{46}$$

where $C_1 = 1 - 1/\sqrt{M}$, $C_2 = 3/(M-1)$.

Applying the Laplace transform (LT)

$$\hat{p}_\gamma(s) = \int_0^\infty e^{-s\gamma} p_\gamma(\gamma)\, d\gamma \tag{47}$$

and combining with (46), we obtain

$$\begin{aligned} P(e) &= \frac{4C_1}{\pi} \int_0^{\pi/2} \hat{p}_\gamma\left(\frac{0.5C_2}{\sin^2\phi}\right) d\phi \\ &\quad - \frac{4C_1^2}{\pi} \int_0^{\pi/4} \hat{p}_\gamma\left(\frac{0.5C_2}{\sin^2\phi}\right) d\phi \end{aligned} \tag{48}$$

Instead of the LT, we can use the moment generating function $M_\gamma(s) = \hat{p}_\gamma(-s)$.

There are many models for the fading channel [6]. A popular and versatile model is the Nakagami $-m$ channel (the m here is a parameter of fading and is unrelated to the m used previously in symbol counting) for which

$$\begin{aligned} p_\gamma(\gamma) &= \frac{m^m \gamma^{m-1}}{(\overline{\gamma})^m \Gamma(m)} e^{-m\gamma/\overline{\gamma}} u(\gamma), \; p_\alpha(\alpha) = \frac{2m^m \alpha^{m-1}}{(\overline{\gamma})^m \Gamma(m)} \\ &\quad \times e^{-m\alpha^2/\overline{\gamma}} u(\alpha), \quad 0.5 \leq m \leq \infty \end{aligned} \tag{49}$$

$$\hat{p}_\gamma(s) = \left(\frac{1+s\overline{\gamma}}{m}\right)^{-m} \tag{50}$$

where

$$\Gamma(m) = \int_0^\infty t^{m-1} e^{-t}\, dt \tag{51}$$

is the gamma function, which for integer m is

$$\Gamma(m) = (m-1)! \tag{52}$$

and $\overline{\gamma}$ is the average value of γ. Note that for $m = 1$, $p_\alpha(\alpha)$ is a Rayleigh PDF, for $m = 0.5$ $p_\alpha(\alpha)$ is a half-Gaussian PDF (because $\Gamma(0.5) = \sqrt{\pi}$), and for $m = \infty$ there is no fading, $p_\alpha(\alpha) = \delta(\alpha - \sqrt{\overline{\gamma}})$ where $\delta(x)$ is impulse function. The fading becomes less severe when m increases. Thus the SEP follows from (48) and (50) as

$$\begin{aligned} P(e) &= \frac{4C_1}{\pi} \int_0^{\pi/2} \left(1 + \frac{0.5C_2\overline{\gamma}}{m \sin^2\phi}\right)^{-m} d\phi \\ &\quad - \frac{4C_1^2}{\pi} \int_0^{\pi/4} \left(1 + \frac{0.5C_2\overline{\gamma}}{m \sin^2\phi}\right)^{-m} d\phi \end{aligned} \tag{53}$$

There is a closed-form formula for integer m that can be found in Eq. (8.108) of Ref. 6. In Fig. 16 we present the SEP as a function of $\overline{\gamma}$ of 16-QAM for several values of m. We can see from this figure that to obtain a SEP of 10^{-2} for a Rayleigh channel, we need an average energy-to-noise ratio per symbol of ~28 dB instead of about ~15 dB with no fading.

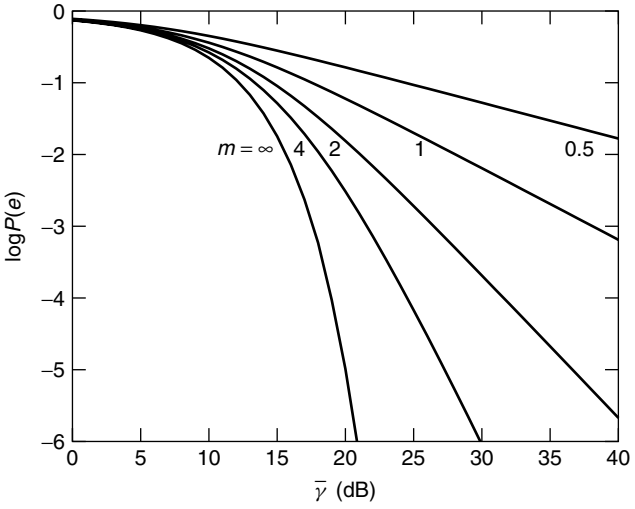


Figure 16. The SEP of 16-QAM as a function of average energy-to-noise ratio per symbol in Nakagami $-m$ fading channel.

The BEP is computed similarly. We cannot use the approximation of (38) unless $\overline{\gamma}_b$ is large. To be precise, we have to now use the conditional BEP (37):

$$P_b(e \mid \gamma_b) = \sum_{i=1}^{I_M} c_i Q(\sqrt{d_i \gamma_b}) \tag{54}$$

After averaging with respect to $p_{\gamma_b}(\gamma_b)$, we obtain

$$P_b(e) = \sum_{i=1}^{I_M} \frac{c_i}{\pi} \int_0^{\pi/2} \hat{p}_{\gamma_b}\left(\frac{0.5 d_i}{\sin^2 \phi}\right) d\phi \tag{55}$$

For $M = 16$ and Nakagami $-m$ fading, the result is

$$\begin{aligned} P_b(e) = {} & \frac{3}{4\pi} \int_0^{\pi/2} \left(1 + \frac{0.4\overline{\gamma}_b}{m \sin^2 \phi}\right)^{-m} d\phi \\ & + \frac{1}{2\pi} \int_0^{\pi/2} \left(1 + \frac{3.6\overline{\gamma}_b}{m \sin^2 \phi}\right)^{-m} d\phi \\ & - \frac{1}{4\pi} \int_0^{\pi/2} \left(1 + \frac{10\overline{\gamma}_b}{m \sin^2 \phi}\right)^{-m} d\phi \end{aligned} \tag{56}$$

There is a closed-form for integer m [see Eq. (5.17a) of Ref. 6]. In Fig. 17 we show the BEP as a function of $\overline{\gamma}_b$ for QAM with $M = 16$ for several values of m.

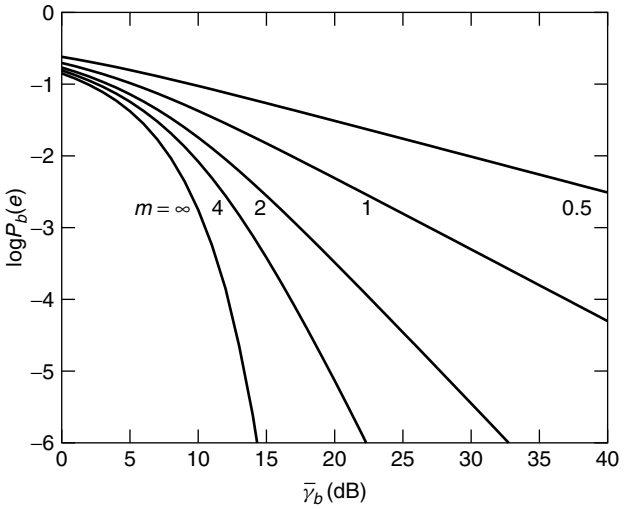


Figure 17. The BEP of 16-QAM as a function of average energy-to-noise ratio per bit for Nakagami $-m$ fading channel.

4.2. The Error Probability with Diversity

In order to reduce the error probability in fading channels, we use *diversity*, where the signal is received simultaneously via K fading channels (by using, e.g., L antenna), which hopefully fade independently so that at least one of the signals has sufficient energy. The received signals are (using complex envelope notation)

$$r_{Cl}(t) = A_l e^{j\varphi_l} \sum_l a_l \cdot g_T(t - \tau_l) + n_{Cl}(t), \quad l = 1, 2, \ldots, L \tag{57}$$

where the amplitudes $\{A_l\}$, phases $\{\varphi_l\}$, and delays $\{\tau_l\}$ are independent random variables that can be tracked by the receiver and $n_{Cl}(t)$ are independent, zero mean, Gaussian, white noises. These signals are combined in an optimal way, called *maximal ratio combining* (MRC) [6] by first aligning the signals in time, using delays $\tau_L - \tau_l$ (τ_L has the maximal delay), then multiplying by $A_l e^{-j\varphi_l}$ and finally processing by $g_R(t)$, which should be a matched filter matched to $g_T(t)$. We have not included the channel filter $h_C(t)$ because we assume flat fading. The block diagram of MRC receiver is shown in Fig. 18.

The decision is based on

$$r(t) = \sum_{l=1}^{L} A_l e^{-j\varphi_l} r_{Ck}(t - \tau_L + \tau_l) \tag{58}$$

For MRC the energy-to-noise ratio per symbol is the sum of the energy-to-noise ratios in each channel [6]

$$\gamma_{\text{MRC}} = \sum_{l=1}^{L} \gamma_l, \quad \gamma_l = \frac{E_l}{N_0} \tag{59}$$

and $\{\gamma_l\}$ are independent random variables. The SEP is

$$\begin{aligned} P(e) = \overline{P(e \mid \gamma_{\text{MRC}})} = {} & \frac{4C_1}{\pi} \int_0^{\pi/2} \overline{e^{-0.5C_2 \gamma_{\text{MRC}}/\sin^2 \phi}}\, d\phi \\ & - \frac{4C_1^2}{\pi} \int_0^{\pi/4} \overline{e^{-0.5C_2 \gamma_{\text{MRC}}/\sin^2 \phi}}\, d\phi \end{aligned} \tag{60}$$

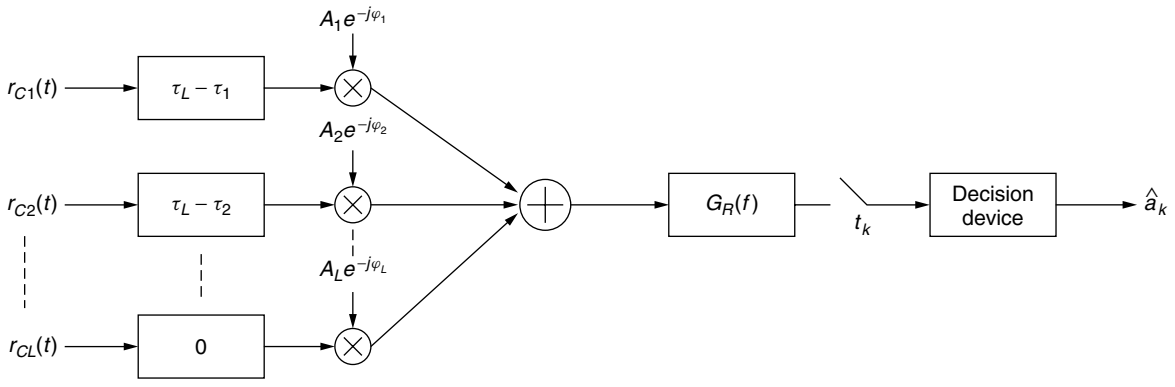


Figure 18. Maximal ratio combining of L fading signals.

where the overbar (vinculum) denotes the average operation. For independent random variables

$$\overline{e^{-C\gamma_{\mathrm{MRC}}}} = \overline{e^{-C\sum_{l=1}^{L}\gamma_l}} = \prod_{l=1}^{L}\overline{e^{-C\gamma_l}} = \prod_{l=1}^{L}\hat{p}_{\gamma_l}(C) = \hat{p}_{\gamma}^{L}(C) \quad (61)$$

where $C = 0.5\,C_2/\sin^2\phi$, $\hat{p}_{\gamma_l}(s)$ is the LT of $p_{\gamma_l}(\gamma_l)$, and the last equality is for identical PDFs for all γ_l,. We thus have

$$P(e) = \frac{4C_1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\hat{p}_{\gamma_l}\left(\frac{0.5C_2}{\sin^2\phi}\right)d\phi - \frac{4C_1^2}{\pi}\int_0^{\pi/4}\prod_{l=1}^{L}\hat{p}_{\gamma_l}\left(\frac{0.5C_2}{\sin^2\phi}\right)d\phi \quad (62)$$

which is a generalization of (48). For Nakagami $-m$ fading with m_l in channel l (62) turns into

$$P(e) = \frac{4C_1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\left(1+\frac{0.5C_2\overline{\gamma}_l}{m_l\sin^2\phi}\right)^{-m_k}d\phi - \frac{4C_1^2}{\pi}\int_0^{\pi/4}\prod_{l=1}^{L}\left(1+\frac{0.5C_2\overline{\gamma}_l}{m_l\sin^2\phi}\right)^{-m_k}d\phi \quad (63)$$

If all $m_l = m$ and $\overline{\gamma}_l = \overline{\gamma}$, we obtain

$$P(e) = \frac{4C_1}{\pi}\int_0^{\pi/2}\left(1+\frac{0.5C_2\overline{\gamma}}{m\sin^2\phi}\right)^{-mL}d\phi - \frac{4C_1^2}{\pi}\int_0^{\pi/4}\left(1+\frac{0.5C_2\overline{\gamma}}{m\sin^2\phi}\right)^{-mL}d\phi \quad (64)$$

which is identical to (53) for $L = 1$ (no diversity). For integer m, the SEP in (64) can be written in a closed form (without integrals). In Fig. 19 we show the SEP of 16-QAM as a function of $\overline{\gamma}$ (energy-to-noise ratio per symbol in one channel) for $m = 1$ (Rayleigh fading) and several values of L. We see from this figure that diversity significantly reduces the SEP.

The corresponding BEP can be calculated in terms of the respective bit quantities

$$\gamma_{b\mathrm{MRC}} = \sum_{l=1}^{L}\gamma_{bl}, \quad \gamma_{bl} = \frac{E_{bl}}{N_0} \quad (65)$$

and replacing (55) by

$$P_b(e) = \sum_{i=1}^{I_M}\frac{c_i}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\hat{p}_{\gamma_{bl}}\left(\frac{0.5d_i}{\sin^2\phi}\right)d\phi = \sum_{i=1}^{I_M}\frac{c_i}{\pi}\int_0^{\pi/2}\hat{p}_{\gamma_b}^{L}\left(\frac{0.5d_i}{\sin^2\phi}\right)d\phi \quad (66)$$

where the last equality is for identical PDFs for all γ_{bl}.

For Nakagami $-m$ fading with identical $m_l = m$ and $\overline{\gamma}_{bl} = \overline{\gamma}_b$, we obtain

$$P_b(e) = \sum_{i=1}^{I_M}\frac{c_i}{\pi}\int_0^{\pi/2}\left(1+\frac{0.5d_i}{m\sin^2\phi}\right)^{-mL}d\phi \quad (67)$$

which can be written in a closed form for integer m.

For 16-QAM, we obtain an equation similar to (56) with the exponential $-m$ replaced by $-mL$. In Fig. 20 we show the BEP of 16-QAM as a function of $\overline{\gamma}_b$ (energy-to-noise ratio per bit in one channel) for $m = 1$ and several values of L.

We conclude that the BEP is also reduced by diversity.

5. CARRIER AND SYMBOL SYNCHRONIZATION

At the receiver of QAM we have to estimate the incoming carrier frequency and phase (this is called *carrier synchronization* or *carrier recovery*) as well as the symbol rate and time delay (this is called *time* or *symbol synchronization* or *clock recovery*). Several books and many papers [11–15] as well as a special (August 1980) issue of

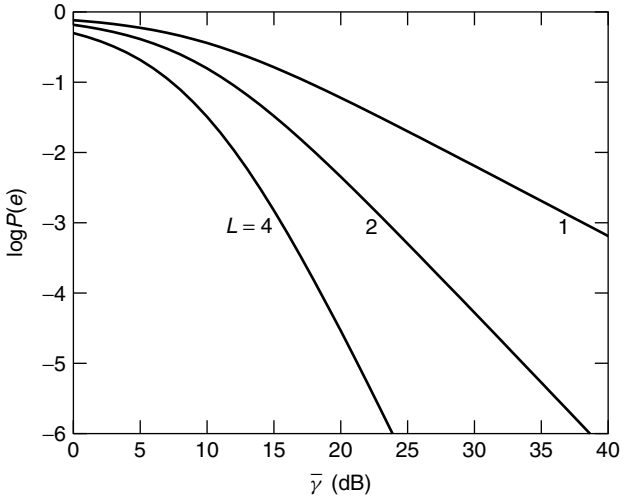


Figure 19. The SEP of 16-QAM with MRC diversity in Rayleigh fading channel.

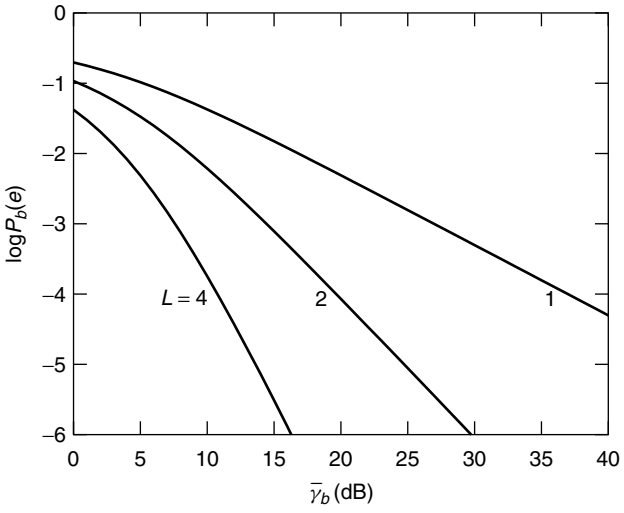


Figure 20. The BEP of 16-QAM with MRC diversity in Rayleigh fading channel.

the *IEEE Transactions on Communications* are dedicated to this problem.

The estimation is performed in two stages: an acquisition stage and a tracking stage. At the acquisition stage we can either use a known sequence of symbols (called *pilot symbols*), which form a preamble to the random data sequence, or we use the unknown data symbols, in which case the acquisition is called blind. At the tracking stage (when the SEP and BEP are already low), we can use the detected symbols $\{\hat{a}_k\}$ instead of the pilot symbols in which case the tracking is called decision directed. The basic device of every synchronizer is a phase-locked loop (PLL) [16].

5.1. Carrier Recovery with Pilot Tone

A block diagram of a PLL is shown in Fig. 21.

We shall assume that the input is a pilot signal corrupted by bandpass noise with PSD $N_0/2$ in the vicinity of the carrier frequency:

$$\tilde{r}_R(t) = A\cos(w_c t + \varphi(t)) + \tilde{n}_R(t) \tag{68}$$

The voltage controlled oscillator (VCO) produces a carrier whose phase is controlled by the input voltage

$$\tilde{s}_v(t) = A_v\cos(w_c t + \hat{\varphi}(t)), \quad \hat{\varphi}(t) = K_v\int_0^t v(t')\,dt' \tag{69}$$

The multiplier output is

$$e(t) = K_m\tilde{s}_v(t)\tilde{r}_R(t) = A_L[\sin(\varphi(t) - \hat{\varphi}(t)) + n(t)] + \text{HFT},$$
$$A_L = \frac{AA_vK_m}{2} \tag{70}$$

where $n(t)$ is zero mean, Gaussian noise with PSD, N_0/A^2, and HFT is a high-frequency term that is eliminated by the loop filter, which is a lowpass filter with transfer function $H_L(f)$. In terms of the phases we obtain the nonlinear circuit in Fig. 22.

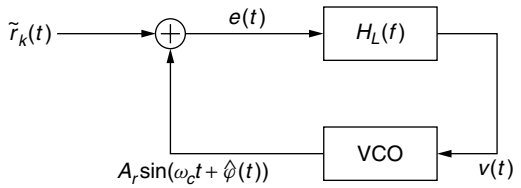


Figure 21. A phase-locked loop.

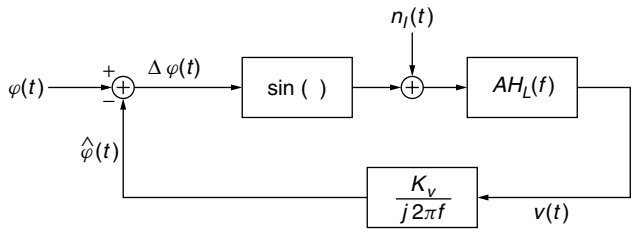


Figure 22. Equivalent circuit of PLL.

The nonlinearity is caused by the term $\sin(\Delta\varphi)$ where

$$\Delta\varphi(t) = \varphi(t) - \hat{\varphi}(t) \tag{71}$$

Note that as long as $|\Delta\varphi(t)| \le \pi/2$, when $\varphi(t)$ increases so does $\Delta\varphi(t)$ and the control signal $v(t)$, hence also $\hat{\varphi}(t)$ which thus tracks $\varphi(t)$. When $\Delta\varphi(t) = 0$, the PLL is locked and $\hat{\varphi}(t) = \varphi(t)$. For small values of $\Delta\varphi(t)$, $\sin(\Delta\varphi) \approx \Delta\varphi$; hence we obtain the linear PLL shown in Fig. 23.

If the input signal has a carrier frequency offset Δf_c, the input phase

$$\varphi(t) = 2\pi f_c t + \varphi(0) \tag{72}$$

may become very large and the PLL may not be able to lock. Therefore during the acquisition stage the frequency of the VCO is swept over a large range until it is locked to $f_c + \Delta f_c$. After locking, the PLL tracks the slow variations in the input phase. There are several methods of sweeping or equivalent operations to obtain frequency locking. These include (1) two switched loop filters, a wideband for acquisition and a narrowband for tracking; (2) nonlinear loop filter with greater sensitivity near lock; and (3) a frequency detector in parallel with a phase detector that is switched off after locking. An example of (3) can be found in Ref. 13.

For the linear PLL in Fig. 24 we can write the equation

$$\hat{\varphi}(t) = [\varphi(t) - \hat{\varphi}(t) + n(t)] * h(t) \tag{73}$$

where

$$H(f) = \frac{KH_L(f)}{j2\pi f} \quad K = A_LK_L \tag{74}$$

is the open-loop transfer function of the PLL. The closed loop transfer function of the PLL is [setting $n(t) = 0$]

$$G_L(f) = \frac{\hat{\Phi}(f)}{\Phi(f)} = \frac{H(f)}{1 + H(f)} = \frac{KH_L(f)}{j2\pi f + KH_L(f)} \tag{75}$$

The noise term that affects $\hat{\varphi}(t)$ has a PSD of $(N_0/A^2)|G_L(f)|^2$ and a variance of

$$\sigma_{\hat{\varphi}}^2 = \frac{2N_0}{A^2}B_L, \quad B_L = \int_0^\infty |G_L(f)|^2\,df \tag{76}$$

where $2B_L$ is the noise bandwidth of the PLL. The signal to noise ratio of the input signal of (68) when taken within the bandwidth $2\tilde{B}_L = 4B_L$ is

$$\text{SNR} = \frac{0.5A^2}{0.5N_0 2\tilde{B}_L} = \frac{0.5A^2}{2N_0B_L} \tag{77}$$

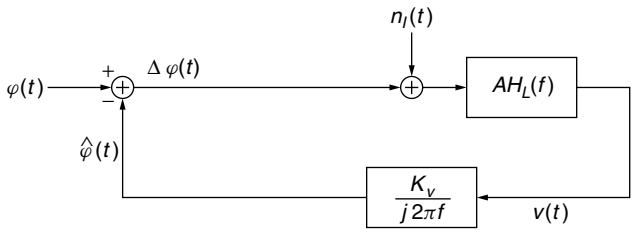


Figure 23. Linear approximation of PLL.

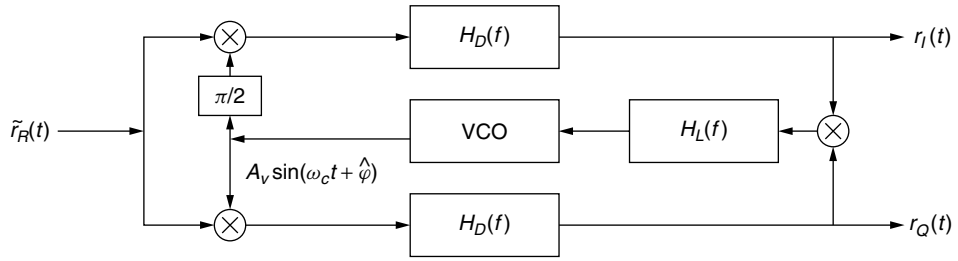


Figure 24. Costas loop.

therefore

$$\sigma_{\hat{\varphi}}^2 = \frac{1}{2\text{SNR}} \tag{78}$$

The variance in (78) is also the variance of the phase $\Delta\varphi$. The nonlinear PLL has been analyzed for the case $H_L(f) = 1$ in Ref. 16, and it has been found that the resulting variance is greater than that of the linear PLL. An alternative to the PLL is the Costas loop, shown in Fig. 24.

For the input in (68), the output of the multipliers are

$$e_I(t) = A_L[\cos(\Delta\varphi) + n_I(t)], \quad e_Q(t) = A_L[\sin(\Delta\varphi) + n_Q(t)] \tag{79}$$

with additional high-frequency terms that are eliminated by the filters. Thus the control voltage (assuming the filters do not change) follows from (79) as

$$v(t) = 0.5K_v A_L^2 \sin\ (2\Delta\varphi) + \text{noise terms} \tag{80}$$

and again $\hat{\varphi}(t)$ will track $\varphi(t)$.

5.2. ML Carrier Recovery with Pilot Symbols or Decision Directed

Here we assume that K symbols $\{a_i\}$ are known. The received signal after the detector filters is

$$r(t) = A\sum_i a_i g(t - iT)e^{j\varphi} + n(t) \tag{81}$$

Taking samples and assuming no ISI, we have

$$r_k = A_0 a_k e^{j\varphi} + n_k, \quad k = 1, 2, \ldots, K, \quad A_0 = Ag_0 \tag{82}$$

The ML estimate of the phase is the minimum of

$$\Lambda_L(\varphi) = \sum_{k=1}^{K} |r_k - A_0 a_k e^{j\varphi}|^2 \tag{83}$$

or equivalently the maximum of

$$\Lambda_L(\varphi) = \text{Re}\left\{\sum_{k=1}^{K} r_k^* a_k e^{j\varphi}\right\}$$

$$= z_I \cos\varphi - z_Q \sin\varphi, \quad z = \sum_{k=1}^{K} r_k^* a_k \tag{84}$$

The maximum is achieved when the derivative is zero, and the solution is

$$\hat{\varphi} = -\tan^{-1}\frac{z_Q}{z_I} \tag{85}$$

It can be shown that $\overline{\hat{\varphi}} = \varphi$; hence the estimate is unbiased. Since z is Gaussian, we can also compute the PDF and variance of $\hat{\varphi}$, which is also $(2KE/N_0)^{-1}$. The block diagram of the QAM system with pilot symbols is shown in Fig. 25, which is very similar to the Costas loop. In the tracking stage we replace the pilot symbol with the estimated symbols $\{\hat{a}_k\}$ which is also shown in the figure.

5.3. Blind Carrier Recovery of QAM Using PLL

We have to create a pilot tone from the QAM signal

$$\tilde{r}_R(t) = A_I(t)\cos(w_c t + \varphi) - A_Q(t)\sin(w_c t + \varphi) + \tilde{n}_R(t) \tag{86}$$

where [assuming for simplicity that $g_{CR}(t)$ is real]

$$A_x(t) = A\sum_i a_{xi} g_{CR}(t - iT), \quad x = I, Q \tag{87}$$

For a symmetric constellation we have

$$\overline{a_{Ii}^p} = \overline{a_{Qi}^p}, \quad \overline{a_{Ii}a_{Qi}} = \overline{a_{Ii}a_{Qi}}, \quad \overline{a_{Ii}} = \overline{a_{Qi}} = 0 \tag{88}$$

therefore

$$\overline{A_I^p(t)} = \overline{A_Q^p(t)}, \quad \overline{A_I(t)A_Q(t)} = \overline{A_I(t)A_Q(t)} = 0 \tag{89}$$

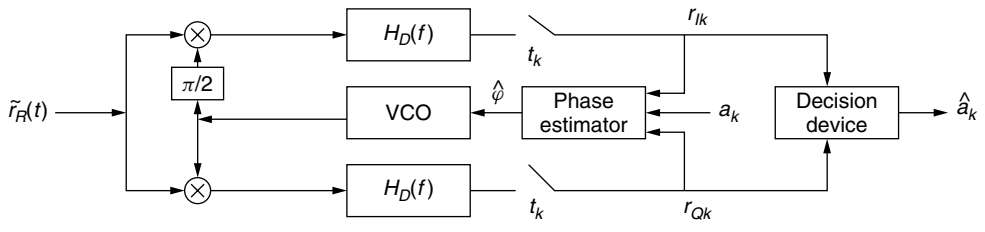


Figure 25. QAM with pilot symbols or decision directed.

The minimum value of p for which we have a pilot tone is 4, namely, $\tilde{s}_R^4(t)$. The pilot tone component is

$$A(t)\cos(4w_c t + 4\varphi(t)), \quad A(t) = 0.25[\overline{A_I^4(t)} - 3(\overline{A_I^2(t)})^2] \tag{90}$$

This component is tracked by the PLL, however to obtain the original carrier we need a frequency divider by 4. The corresponding PLL for QAM is shown in Fig. 26.

Because only a small part of $\tilde{s}_R^4(t)$ contains energy at frequency $4f_c$ and the additional noise terms in the vicinity of this frequency, the variance of the estimated phase is

$$\sigma_{\hat{\varphi}}^2 = \frac{S_L}{2\text{SNR}}, \quad S_L \le 1 + \frac{9}{\text{SNR}} + \frac{6}{\text{SNR}^2} + \frac{1.5}{\text{SNR}^3} \tag{91}$$

where S_L is called a *power loss*. The estimate of the phase can also be computed [14] from discrete samples as in Fig. 25:

$$\hat{\varphi} = 0.25 \arg\left[\overline{a_k^4} \sum_{k=1}^{K} r_k^4\right] \tag{92}$$

with a variance which can be approximated by

$$\sigma_{\hat{\varphi}}^2 = \left(\frac{c_1}{2\gamma} + c_2\right)\frac{1}{K}, \quad \gamma = \frac{E}{N_0} \tag{93}$$

The values of c_1 and c_2 can be computed and are shown in Table 1 for various QAM square and cross-constellations. Note that there is a self-noise represented by c_2 and that both c_1 and c_2 are large for cross-constellations. A reduced constellation power law algorithm [14] in which only r_k with amplitudes that exceeds a certain threshold gives a better estimate.

Note that if $2n\pi$ is added to $4\hat{\varphi}(t)$, the VCO output is unchanged however after the frequency divider $\hat{\varphi}(t)$ has an uncertainty of $n\pi/2$. To resolve this uncertainty, we need again pilot symbols or we have to use differential phase modulation, in which the phase of the symbols in each quarter of the constellation is transmitted as a phase difference instead of an absolute phase. The first 2 bits of each symbol can be represented by this phase.

5.4. Blind ML Carrier Recovery for QAM

The blind ML estimate of the phase maximizes

$$\Lambda_L(\varphi) = \exp -\overline{\frac{1}{2\sigma_n^2}\sum_{k=1}^{K}\left|r_k - A_0 a_k e^{j\varphi}\right|^2}$$

$$= \prod_{k=1}^{K} \exp -\overline{\frac{1}{2\sigma_n^2}\left|r_k - A_0 a_k e^{j\varphi}\right|^2} \tag{94}$$

where the average is over the symbols. For the square constellations, this is equivalent to maximizing

$$\Lambda_L(\varphi) = \sum_{k=1}^{K} \ln\left\{\sum_{i=1}^{\sqrt{M}-1} \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right) \times \cosh\frac{r_{Ik}\cos\varphi(2i-1)A_0}{\sigma_n^2}\right\}$$

$$+ \sum_{k=1}^{K} \ln\left\{\sum_{i=1}^{\sqrt{M}-1} \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right) \times \cosh\frac{r_{Qk}\sin\varphi(2i-1)A_0}{\sigma_n^2}\right\} \tag{95}$$

The Cramer–Rao lower bound (CELB) [15] to any estimate is

$$\sigma_{\hat{\varphi}}^2 \ge \text{CRLB}(\hat{\varphi}) = -\left(\frac{d^2\Lambda_L(\varphi)}{d\varphi^2}\right)^{-1} \tag{96}$$

The CRLB of phase and also of frequency has been derived for QAM [15]. The results are

$$\text{CRLB}(\hat{\varphi}) = \left[2K\gamma F\left(\frac{1}{2\gamma}\right)\right],$$

$$\text{CRLB}(\hat{f}_c) = \left[2K\gamma\frac{K^2-1}{12}F\left(\frac{1}{2\gamma}\right)\right]^{-1} \tag{97}$$

where $F(x)$ is an integral that is evaluated numerically. The values of CRLB $(\hat{\varphi})$ and of $\sigma_{\hat{\varphi}}^2$ for several

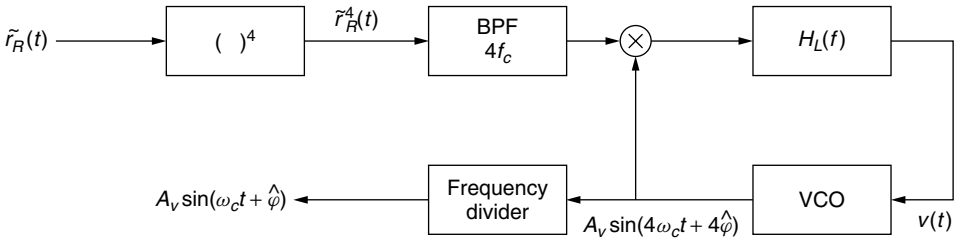


Figure 26. PLL with 4th power signal.

Table 1. Values of Constants Needed in Computation of Phase Error Variance for QAM

M	4	16	32	64	128	256	512	1024	2048	4096
c_1	1	4.24	52.63	5.81	62.07	6.27	64.68	6.39	63.39	6.42
c_2	0	0.06	3.14	0.17	3.79	0.20	3.98	0.21	4.03	0.21

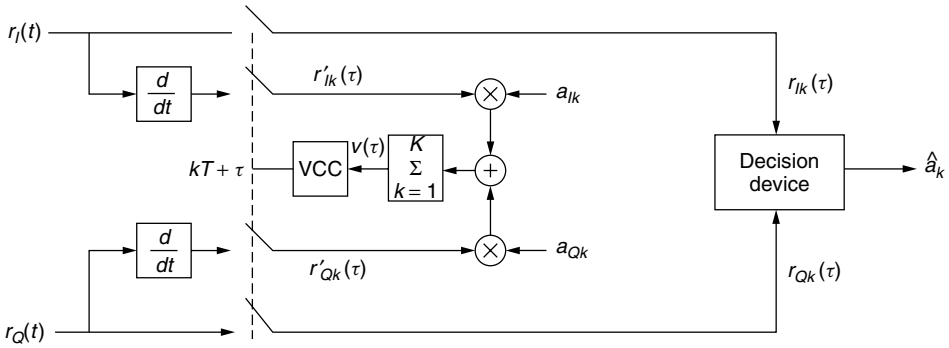


Figure 27. Standard deviation of phase, $\sigma_{\hat{\phi}}$, for 64-QAM and $K = 200$ symbols. (*Source*: F. Riou, B. Cowley, B. Morgan, and M. Rice, Cramer–Rao lower bounds for QAM phase and frequency estimation, *IEEE Trans. Commun.* **COM-49**: 1582–1591, © 2001 IEEE.)

practical systems [histogram algorithm (HA), two-stage conjugate algorithm (2SC), MDE-minimum-distance algorithm (MDA), and power-law estimate (PLE) ($P = 4$)] are shown in Fig. 27 (which is Fig. 8 of Ref. 15) as a function of γ for 64-QAM and $K = 200$. The CWCRLB is $(2K\gamma)^{-1}$. CRLB $(\hat{f}_c)$ is presented in Fig. 28 (which is Fig. 3 of Ref. 15) for several QAM constellations. More figures on this matter can be found in Refs. 14 and 15.

5.5. Time Recovery with Pilot Symbols or Decision Directed

The ML estimator of τ maximizes

$$\Lambda_L(\tau) = \mathrm{Re}\left\{\sum_{k=1}^{K} a_{ki}^{*} r_k(\tau)\right\} = \sum_{k=1}^{K} r_{Ik}(\tau)a_{Ik} + r_{Qk}(\tau)a_{Qk}, r_k(\tau) = r(kT+\tau) \tag{98}$$

Taking the derivative, we have

$$v(\tau) = \frac{d\Lambda_L(\tau)}{d\tau} = \sum_{k=1}^{K} r'_{Ik}(\tau)a_{Ik} + r'_{Qk}(\tau)a_{Qk}, r'_{Xk}(\tau) = \left.\frac{dr(t)}{dt}\right|_{t=kT+\tau} \tag{99}$$

The maximum is achieved when $v(\hat{\tau}) = 0$. The implementation of this equation is shown in Fig. 29. The VCC (voltage controlled clock) is a VCO that produces rectangular pulses instead of a sinusoid. The summer is the digital equivalent of an integrator and forms the loop filter.

In the tracking stage estimated symbols are reliable and can be used instead of the pilot symbols. This is the decision directed mode of the clock recovery.

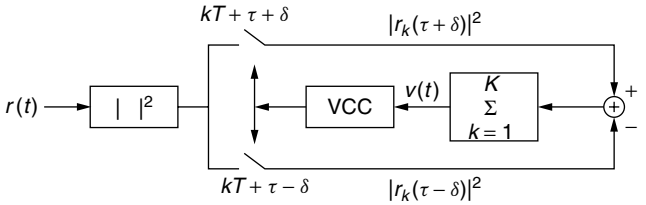


Figure 28. Standard deviation of frequency, $\sigma_{\hat{f}_c}$, for $K = 200$ symbols. (*Source*: F. Riou, B. Cowley, B. Morgan, and M. Rice, Cramer–Rao lower bounds for QAM phase and frequency estimation, *IEEE Trans. Commun.* **COM-49**: 1582–1591, © 2001 IEEE.)

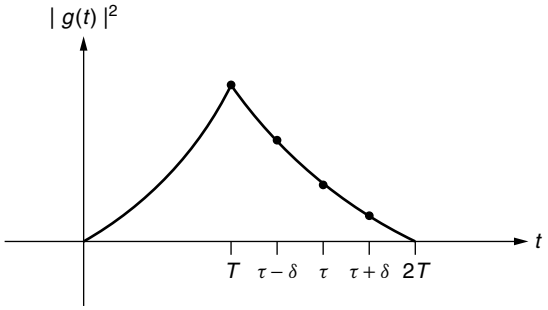


Figure 29. ML clock recovery with pilot symbols or decision directed.

5.6. Blind ML Clock Recovery

Here we maximize an equation similar to (94)

$$\Lambda(\tau) = \exp -\overline{\frac{1}{2\sigma_n^2}\sum_{k=1}^{K}|r_k(\tau) - A_0 a_k|^2} = \prod_{k=1}^{K}\exp -\overline{\frac{1}{2\sigma_n^2}|r_k(\tau) - A_0 a_k|^2} \tag{100}$$

where the average is over the symbols. For QAM with square constellation, this is equivalent to maximizing

$$\Lambda_L(\tau) = \sum_{lk=1}^{K}\ln\left\{\exp -\frac{|r_k(\tau)|^2}{2\sigma_n^2}\sum_{i=1}^{\sqrt{M}-1} \times \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right)\cosh\frac{r_{Ik}(\tau)(2i-1)A_0}{\sigma_n^2}\right\} + \sum_{k=1}^{K}\ln\left\{\exp\left(-\frac{|r_k(\tau)|^2}{2\sigma_n^2}\right)\sum_{i=1}^{\sqrt{M}-1} \times \exp -\left(\frac{A_0^2(2i-1)^2}{2\sigma_n^2}\right)\cosh\frac{r_{Qk}(\tau)(2i-1)A_0}{\sigma_n^2}\right\} \tag{101}$$

The optimal $\tau$is the solution of

$$\frac{d\Lambda_L(\tau)}{d\tau} = 0\ |_{\tau=\hat{\tau}} \tag{102}$$

An approximation to the ML synchronizer is the E-L (early–late) gate synchronizer, which computes

$$\Delta\Lambda(\tau) = \frac{\Lambda_L(\tau+\delta) - \Lambda_L(\tau-\delta)}{2\delta}, \quad \Lambda_L(\tau) \approx \sum_{k=1}^{K} |r_k(\tau)|^2 \tag{103}$$

A block diagram of this scheme is shown in Fig. 30.

The E-L algorithm is based on the idea that the pulse at the receiver has a peak at the correct sampling time and is symmetric around this point as shown in Fig. 31. Note that the correct sampling time is T and $\Delta\Lambda(T) = 0$ while $\Delta\Lambda(\tau) > 0$ if $\tau < T$ and $\Delta\Lambda(\tau) < 0$ if $\tau > T$. The E-L has to be modified for QAM because certain sequences of symbols will result in false values of $\Delta\Lambda(\tau)$ (see Ref. 1 for elaborations).

A simple blind recovery of the symbol rate is based on the autocorrelation of the signal

$$\begin{aligned} R_s(\tau) &= 0.5\overline{s(t)s^*(t-\tau)} \\ &= 0.5A^2\sigma_a^2 \sum_i g(t-iT)g(t-iT-\tau) \end{aligned} \tag{104}$$

which is a periodic function of t with period T. Therefore a bandpass filter with center frequency $R = 1/T$ or nR can produce a clock frequency directly or after frequency division by n. This circuit is implemented in Fig. 32. The best results are obtained for $\tau = T/2$.

6. SUMMARY

We have described a QAM system and computed the error probability in both Gaussian and fading channels with and without diversity. We presented various methods of carrier and clock recovery. Many topics such as equalization, effect of errors in carrier and clock recovery on the error probability, the error probability of differential phase detection, combined differential phase and differential amplitude for mobile communication, joint phase, frequency and time synchronization, and QAM in spread-spectrum systems have been omitted for lack of space.

BIOGRAPHIES

John P. Fonseka received B.Sc.(Hons.) degree in Electronic and Telecommunication Engineering from University of Moratuwa, Sri Lanka in 1980, M.Eng. in Electrical

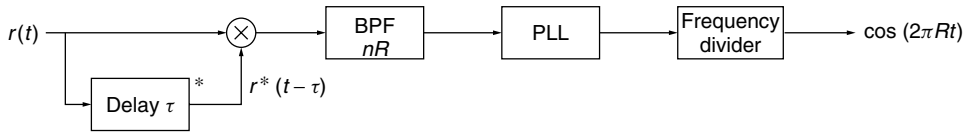


Figure 30. Early–late gate clock recovery.

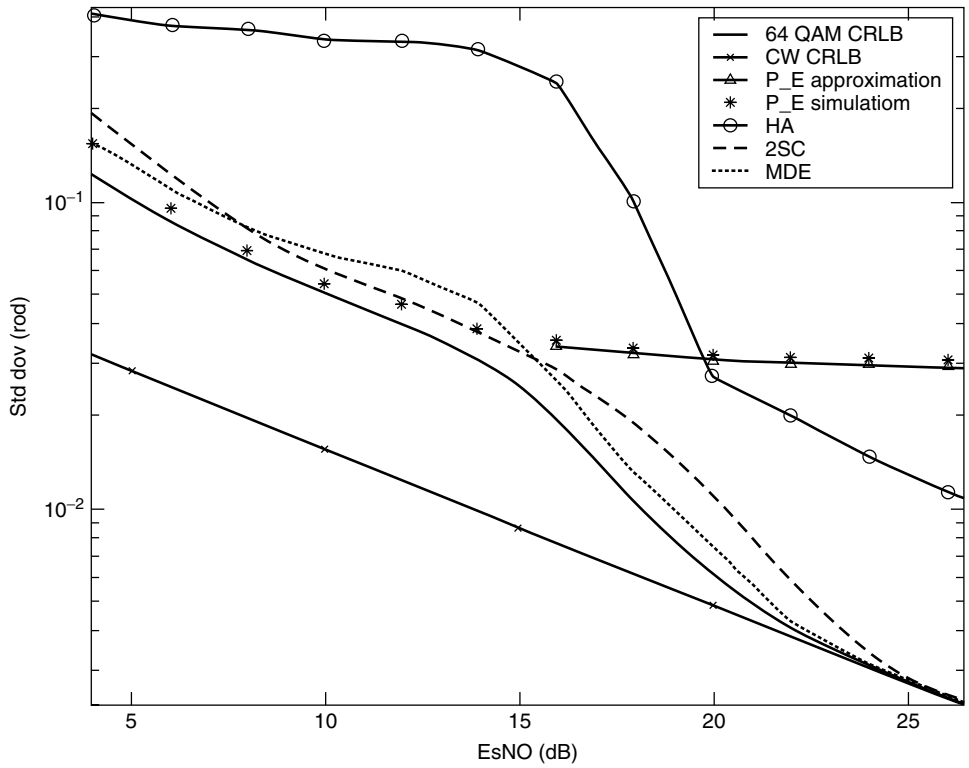


Figure 31. Pulse for early–late clock recovery.

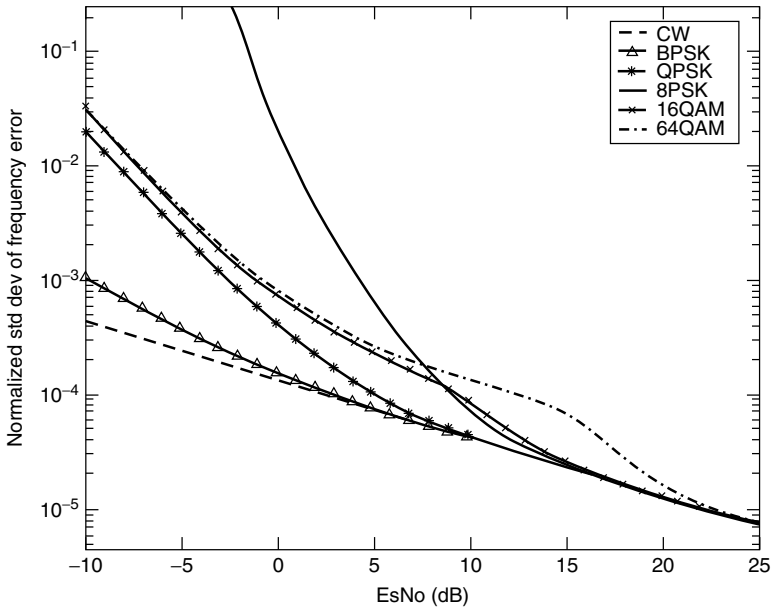


Figure 32. Blind recovery of symbol rate.

Engineering from Memorial University of Newfoundland, Canada in 1985, and Ph.D. in Electrical Engineering from Arizona State University in 1988.

He joined University of Texas at Dallas in August of 1988 as an Assistant Professor, where he is currently serving as a Professor in Electrical Engineering. His research interests include combined coded modulation, signaling through narrowband fading channels, coding theory, telecommunication networks, and optical communications.

Israel Korn received his BSc, MSc, and DSc degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 1962, 1964 and 1968 respectively. Since 1978 he has been with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, New South Wales, Australia, where he is now a Visiting Professor or Honorary Visiting Fellow. He was a Visiting Professor at various universities and institutions in the USA, Germany, Denmark Spain and Australia. His research and teaching interests are in the area of Digital Communication with applications to mobile, wireless and personal communications. He has published 87 papers in refereed journals, 45 papers in conference proceedings and two books one of which is, "Digital Communications," Van Nostrand, NY 1985. He has taught various undergraduate and postgraduate course in the general area of Communications at various universities. He was an editor of the IEEE Transactions on Communications (1992–1995) and of Wireless Personal Communications (1992–). Since 1994 he is a Fellow of IEEE.

BIBLIOGRAPHY

1. W. T. Webb and L. Hanzo, *Modern Quadrature Amplitude Modulation*, IEEE Press, New York, Pentech Press, London, 1994.
2. I. Korn, *Digital Communication Systems*, Van Nostrand Reinhold, New York, 1985.
3. K. Feher, *Advanced Digital Communication and Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
4. R. E. Ziemer and R. L. Peterson, *Introduction to Digital Communications*, Macmillan, New York, 1992.
5. M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques. Signal. Design and Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
6. M. K. Simon and M. S. Alouini, *Digital Communication over Fading Channels*, Wiley, New York, 2000.
7. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, Boston, 2001.
8. G. J. Foscini, R. D. Gitlin, and S. B. Weinstein, Optimisation of the two-dimensional signal constellations in the presence of Gaussian noise, *IEEE Trans. Commun.* **COM-22**: 28–38 (1974).
9. C. M. Thomas, M. Y. Weidner, and S. H. Durrani, Digital amplitude-phase keying with *M*-ary alphabets, *IEEE Trans. Commun.* **COM-22**: 168–180 (1974).
10. J. Lu, K. B. Letayef, J. C.-I. Cuang, and M. L. Liou, *M*-PSK and *M*-QAM BER computation using signal-space concepts, *IEEE Trans. Commun.* **COM-47**: 181–184 (1999).
11. H. Meyr, M. Moenclaey, and S. A. Fechtel, *Digital Communication Receivers*, Wiley, New York, 1998.
12. U. Mengali and D. D'Andrea, *Synchronization Techniques for Digital Receivers*, Plenum Press, New York, 1997.
13. H. Sari and S. Moridi, New phase and frequency detectors for carrier recovery in PSK and QAM systems, *IEEE Trans. Commun.* **COM-36**: 1035–1043 (1988).
14. C. N. Georghiades, Blind carrier phase acquisition for QAM constellation, *IEEE Trans. Commun.* **COM-45**: 1477–1486 (1997).
15. F. Riou, B. Cowley, B. Moran, and M. Rice, Cramer-Rao lower bounds for QAM phase and frequency estimation, *IEEE Trans. Commun.* **COM-49**: 1582–1591 (2001).
16. A. J. Viterbi, *Principles of Coherent Communications*, McGraw-Hill, New York, 1966.

R

RADIO PROPAGATION AT LF, MF, AND HF

JOHN C. H. WANG
Federal Communications Commission
Washington, District of Columbia

1. INTRODUCTION

Electromagnetic (or radio) waves can be transmitted from a transmitting site to a receiving site by a number of different mechanisms. At the frequencies under study (between about 30 kHz and 30 MHz), the most important ones are ground wave and sky wave. At higher frequencies, the space (or tropospheric) wave becomes more important. The ground (or surface) wave exists when the transmitting and receiving antenna are on or near the surface of the earth. Standard broadcast signals received in daytime are all ground waves.

The sky wave represents energy that travels from the transmitting antenna to the receiving antenna as a result of a "bending" by the earth's upper atmosphere called the *ionosphere*. The ionosphere, which consists of several different layers, begins about 50 km above the earth's surface. Shortwave signals and nighttime medium-wave signals are examples of sky waves. Under certain conditions, the ground-wave and sky-wave components from the same source may be comparable in amplitude but arrive at slightly different times, resulting in interference.

The space (or tropospheric) wave represents energy that travels from the transmitting antenna to the receiving antenna in the earth's troposphere. The troposphere is the lower portion of the earth's atmosphere in which the temperature decreases with increasing altitude. This part of the atmosphere extends to an altitude of about 9 km at the earth's poles and 17 km at the equator. TV and FM signals are examples of space waves. Space wave propagation is beyond the scope of this article.

In the subsequent sections factors affecting propagation will be described. Methods of predicting field strength will also be analyzed and compared. Definitions of the most frequently used terms are given in Section 6.

2. GROUND-WAVE PROPAGATION

2.1. Early Pioneering Studies

At frequencies between about 10 kHz and 30 MHz, ground-wave propagation is possible because the surface of the earth is a conductor. The ground wave is vertically polarized. Any horizontal component of an electric field on the surface of the earth is shorted-circuited by the earth. The earliest work on ground-wave propagation was carried out by Summerfield [1]. His flat-earth theory states that ground-wave field strength, E_g, can be expressed in the form

$$E_g = A\frac{E_0}{d} \tag{1}$$

where E_0 = field strength of wave at the surface of the earth at unit distance from the transmitting antenna, neglecting earth's losses
d = distance to transmitting antenna
A = factor taking into account the ground losses

The field strength E_0 at unit distance in Eq. (1) depends on the power radiated by the transmitting antenna and the directivity of the antenna in the vertical and horizontal planes. If the radiated power is 1 kW and the short vertical antenna is omni-directional in the horizontal plane, then, $E_0 = 300$ mV/m when the distance is 1 km. The reduction factor A is a complicated function of electrical constants of the earth, frequency, and the distance to the transmitter in wavelengths. The reduction is highly frequency-dependent; it increases with increasing frequency. Thus, at LF and MF ground-wave signals can be sufficiently strong for broadcasting service. On the other hand, at HF, ground-wave signals are usually too weak for broadcasting purposes. The Summerfield flat-earth approach, the subsequent Watson transformation [2], and the Bremmer residue series [3] were the important milestones and theoretical advances on which the modern ground-wave theory is still based.

2.2. The Development of Ground-Wave Curves

Intensive efforts to convert the theoretical advances to simple and practical field strength curves took place between 1930 and 1940 [4]. Extensive measurement programs were conducted by many organizations including the Federal Communications Commission. In 1939, the FCC released a complete set of ground-wave curves as an appendix to the Standards for Good Engineering Practice Concerning Standard Broadcasting Stations [5]. These curves and a comprehensive discussion were included in a paper by Norton [6]. Similar but not identical ground-wave curves can also be found in ITU-R Recommendation PN.368-7 [7]. The current FCC curves cover the frequency range of 535–1705 kHz. The ITU-R curves cover a much wider range of frequency, from 10 kHz to 30 MHz. [*Note*: ITU-R, which appears frequently in this article, is the abbreviated name of the *Radiocommunication Study Groups of the International Telecommunication Union*, formerly known as the CCIR (see Section 6). PN denotes propagation in nonionized media. PI, which will appear in the subsequent sections, denotes propagation in ionized media. Numeral 7 after a dash means it is the 7th revised edition. In 2000, ITU-R decided to use the prefix P to replace both PN and PI.]

2.3. Available Software

Currently, there are three computer programs available for calculating ground-wave field strengths. The first

program is called *ITSGW* [8]. The second program, GRWAVE [9], is in ITU-R Recommendation PN.368-7. This program, which takes into account the effects of refraction in an exponential atmosphere, is available from ITU Sales Service, Place des Nations, 1211 Geneva, Switzerland. The third program is called *FCCGW* [10]. FCCGW has been used to generate the metric version of the FCC curves. The FCC curves take into account the effect of refraction by using an effective radius that is $\frac{4}{3}$ times the actual radius of the earth. Refraction is insignificant at distances less than about 100 km. At greater distances, it becomes progressively more significant. It should also be mentioned that GRWAVE is designed for personal computers while FCCGW is designed for mainframe computers. A comparison of these programs has been done [10,11]. It is reported that the three methods give ground-wave field strength predictions sufficiently close in value that they could be considered identical for frequency management purposes.

2.4. Ground-Wave Propagation over Inhomogeneous Terrain

For predicting ground-wave field strengths over paths composed of successive sections of terrain of different conductivities (e.g., land and sea), there are two basic methods available. These are the equivalent-distance (or Kirke) method [12] and the equivalent-field (or Millington) method [13]. The Kirke method has the advantage of simplicity, but in cases where the successive sections show considerable differences in conductivities, it can lead to large error. On the other hand, the Millington method does not suffer from this problem. Furthermore, the Millington method is now no longer as difficult to apply as before, because a simplified graphical solution has been developed by Stokke [14]. The Millington method and the Stokke approximation are presented in ITU-R Recommendation PN.368-7.

2.5. Ground Conductivity

Ground-wave propagation can be considered a reasonably well understood topic. In one area, however, more work is needed. Ground conductivity is a very important factor in calculating ground-wave field strengths. Accurately measured data should be used. Although several maps are available, they present estimates and are not very accurate. A map showing the estimated ground conductivities of the continental United States has been published by the FCC [15]. An atlas of ground conductivities in different parts of the world can be found in ITU-R Recommendation PN.832-2 [16].

Conductivity of seawater is typically 5 siemens per meter (5 S/m) while that of freshwater is about 10 mS/m. Conductivities of rocky land, hills, and mountains vary between 1 and 2 mS/m. Conductivity of rich agricultural land is typically 10 mS/m. Cities and residential areas have a conductivity of about 2 mS/m. In industrial areas, it is even less.

3. THE SKY-WAVE PROPAGATION MECHANISM

3.1. The Solar–Terrestrial System

In 1901, Guglielmo Marconi (1874–1937), a young Italian engineer, succeeded in sending a Morse code message from Cornwall, England, across the Atlantic Ocean to Newfoundland. It is generally believed that the frequency Marconi used was about 1.6 MHz. This revolutionary wireless experiment not only brought him a Nobel Prize later in 1909 but also created a new frontier in the scientific world. Perhaps Oliver Heaviside, an English physicist, gave the earliest satisfactory explanation of his experiment. He theorized that in the earth's upper atmosphere, there is a sufficiently conducting layer [17]. This conducting layer is known as the *ionosphere*, so called because it consists of heavily ionized molecules. To understand sky-wave propagation, it is essential to study the entire solar–terrestrial system, not just the ionosphere alone. In this article, we shall discuss this subject only briefly. For more details, see books by Davies [18] and by Goodman [19]. It should be mentioned that the discussion in this section applies to LF, MF, and HF (low, medium, and high frequency).

3.2. The Ionosphere

The ionosphere consists of three regions (or layers). They are the D, E, and F regions, respectively, in increasing order of altitude.

The D region spans the approximate altitude range of 50–90 km. It exists only in daytime and disappears shortly after sunset. The D region can be considered as an absorber, causing significant signal attenuation. The absorption is frequency-dependent; it decreases with increasing frequency.

The E region spans the approximate altitude range of 90–130 kilometers. This region is important for nighttime low- and medium-frequency propagation at distances greater than about 200 km. The E region exhibits a solar cycle dependence with maximum electron density occurring at solar maximum.

Sporadic E (Es), which has very little relationship with solar radiation, is an anomalous form of ionization embedded within the E region. It can have significant effects on propagation at HF and VHF.

The F region extends upward from about 130 kilometers to about 600 kilometers. The lower and upper portions of the F region display different behaviors at daytime, resulting in a further subdivision into F1 and F2 layers. The F1 layer is the region between 130 and 200 km above the surface of the earth. The F2 layer, which is the highest and the most prominent one, is the principal reflecting region for long-distance high-frequency communications. At night, the F1 layer merges with the F2 layer and the average height of the combined layer (still called the "F2 layer") is about 350 km.

3.3. Solar Activity

The existence of the ionosphere is a direct result of the radiation from the sun, both electromagnetic and corpuscular. The electromagnetic radiation travels toward

the earth at the speed of light, and the entire journey takes about 8.3 min. The ionization process is linked with the intensity of the solar radiation, which in turn varies with factors such as time of day, latitude, and season. Solar activity changes drastically from time to time. Sunspot number is a reasonably good index of the state of solar activity, although several other indices are also available. Sunspots are dark areas on the surface of the sun. Sunspots were, perhaps, first observed by the Chinese in March of 20 A.D. [20] during the Han Dynasty (206 B.C.–220 A.D.). Sunspots appear dark because the temperature is low, only about 3000 K, while the average temperature of the sun is about 6000 K. Sunspots tend to group together and display an 11-year cyclic nature. The astronomic records of the Jin Dynasty (265–418 A.D.) of China [21] indicate that for quite a while in the fourth century, sunspots were observed every 11 years (e.g., 359, 370, 381, and 393 A.D.). Sunspot numbers vary from day to day and year to year. Routine observations have been made since 1749. The cycle beginning in 1755, a year of minimum sunspot number, is considered cycle 1. The ascending portion of a cycle (on average, 4.5 years) is usually much shorter than the descending one (6.5 years). The Zurich (or Wolf) number R is given by

$$R = k(10\,g + s) \tag{2}$$

where g is the number of sunspot groups, s is the number of observed small spots, and k is a correction factor, approximately unity, used to equalize the results from different observations and equipment. The sunspot number is subject to wide variations from month to month and is of little usefulness. Furthermore, it is known that the characteristics of the ionosphere do not follow the short-term variations. In order to achieve a better correlation, some kind of "smoothing" technique is desirable. Consequently, the 12-month smoothed sunspot number (R_{12}) has been adopted and is the most widely used index in ionospheric work today:

$$R_{12} = \frac{1}{12}\left[(0.5)(R_{n+6} + R_{n-6} + \sum_{n-5}^{n+5}(R_n)\right] \tag{3}$$

Thus, by definition, the value of R_{12} is known only 7 months after the recorded observation. R_{12} varies from a minimum of about 10 to a maximum generally of 100–150, although in December 1957 it reached a record high of 239.4.

3.4. Atmospheric Radio Noise

In order to estimate the performance to be expected in a communication system, it is insufficient to consider field strength alone. Equally important is radio noise. There are many different sources of noise: the atmosphere, the receiving system, human activity, the sun, and galaxies. In this article, we emphasize atmospheric noise. For an excellent discussion on all types of noise, see ITU-R Recommendation PI.372-6 [22].

Atmospheric noise is produced mainly by lightning discharges in thunderstorms. Discharges take place between 2 and 4 km above ground. The power released is very great, typically greater than 10 GW [23]. Atmospheric radio noise obeys the same propagation laws as do sky-wave signals. Thus, it travels to distances several thousands of kilometers away. The noise level thus depends on the time of day, season of the year, weather, geographic location, and frequency. In general, atmospheric noise is the highest (1) when the receiver is located near a thunderstorm center, (2) during local summer, (3) during the night, or (4) when the frequency is low. There are three major thunderstorm (hence, noise) centers in the world: the Caribbean, Equatorial Africa, and Southeast Asia. Maps showing the atmospheric noise levels for different parts of the world corresponding to different seasons of the year and different hours of the day have been developed by the CCIR since 1964. The most recent maps can be found in ITU-R Recommendation PI.372-6 [22].

3.5. Magnetic Coordinates

There are several definitions of latitude connected with the geomagnetic field. The centered dipole latitude, or simply the dipole latitude, is an approximation and has been used for ionospheric work for decades. Corrected geomagnetic latitude more accurately represents the real geomagnetic field and should be used when accuracy is desired. Conversion tables from geographical coordinates to corrected latitude are readily available [24].

4. SKY-WAVE PROPAGATION AT LF AND MF

4.1. Results of Six Decades of Worldwide Efforts

This section presents a brief summary of the historical background behind the development of the four currently used LF and MF sky-wave propagation models. For a more detailed presentation, see a paper by Wang [25].

4.1.1. The Cairo Curves. The earliest worldwide efforts to study LF/MF sky-wave propagation began in 1932. At its meeting held in Madrid, the CCIR established a committee, chaired by Balthasar van der Pol of Holland, to study propagation at frequencies between 150 and 2000 kHz. Measurements were made on 23 propagation paths between 1934 and 1937. Consequently, two sky-wave propagation curves were developed. One of the curves is for paths far away from the earth's magnetic poles; this is better known as the *north–south curve*. The other curve is for paths approaching the earth's magnetic poles and is better known as the *east–west curve*. Actually, the former should have been called the *low-latitude curve*; the latter, the *high-latitude curve*. The two curves were formally adopted by the CCIR at the 1938 International Radio Conference, Cairo. Therefore, these curves are known as the *Cairo curves*. In1974 the LF/MF Conference adopted the north–south curve for use in the Asian part of Region 3 [26].

4.1.2. The Region 2 Method (the FCC Clear-Channel Curve). Recognizing the needs for a set of sound engineering standards, the FCC, under the leadership of the late Ken Norton, carried out a sky-wave field

strength measurement program in the spring of 1935, a year of moderate sunspot number. Nighttime signals of all eight clear-channel stations were monitored at 11 widely scattered receiving sites. From these measurements, the FCC clear-channel curve was derived. The 1980 Broadcasting Conference for Region 2 adopted this method for applications in Region 2 [27]. Hence, this method is also known as the Region 2 method. The staff of the FCC has since developed a newer and more accurate method. The newer method, which is being used by the FCC for domestic applications, will be discussed in the subsequent sections.

4.1.3. Comparison of the Two Graphical Methods. Both the Cairo and the FCC curves present field strength as a function of distance only. They do not take into account effects of latitude, sunspot number, and so on. When converted to the same conditions, the two Cairo curves and the FCC clear-channel curve are similar for distances up to about 1400 km. At 3000 km, the north–south curve is about 8 dB greater than the east–west curve; at 5000 km, the difference is about 18 dB. The FCC clear-channel curve falls between the two Cairo curves. The Cairo north–south curve offers good results when applied to low latitudes. When applied to higher latitudes, it tends to overpredict field strength levels [25]. The FCC curve offers reasonable results when applied to temperate latitudes. Neither of these methods is a true worldwide method. The Cairo east–west curve, because it often underestimates field strength levels, has virtually been disregarded.

4.1.4. The Development of Recommendation 435 (the Udaltsov–Shlyuger Method). Recognizing the need for a simple but accurate field strength prediction method for worldwide applications, and in anticipation of a broadcasting conference, the CCIR at its Xth Plenary Assembly (Geneva, 1963) established International Working Party (IWP) 6/4 to undertake such a task. This IWP was first chaired by J. Dixon (Australia), succeeded by G. Millington, P. Knight (UK), and J. Wang (USA). In the late 1960s and early 1970s a number of administrations and scientific organization were very active and collected valuable data. For example, the European Broadcasting Union (EBU), which started its sky-wave studies soon after World War II, reactivated its efforts. Its counterpart in Eastern Europe, the International Organization of Radio and Television (OIRT) was also active. The administration of the former USSR also collected a significant amount of measurements. Results of their analysis together with a new propagation model were published in 1972 [28], although their data have not been made available to the public.

Three international organizations jointly planned and carried out a measurement campaign in Africa between 1963 and 1964. They are the EBU, the OIRT, and the Union of National Radio afd Television Organizations (URTNA). Later, the British Broadcasting Corporation set up seven receiving stations in Africa and signals from two transmitters on the British Ascension Islands were monitored. The BBC project was intended to study polarization coupling loss and sea gain. The Max Planck Institute also conducted measurements at Tsumeb, southwest Africa.

Administrations in ITU Region 3 (parts of Asia, Australia, and New Zealand), in cooperation with the Asian-Pacific Broadcasting Union (ABU), were equally active and productive. Furthermore, the Japanese administration carried out a number of mobile experiments in different parts of the Pacific [29].

While the administrations in the Eastern Hemisphere were busy preparing for the 1974/75 Regional LF/MF Conference for ITU Regions 1 and 3, IWP 6/4 was actively developing a propagation model to be used as part of the technical bases for such a conference. After extensive studies and lengthy deliberations, the IWP was able to agree on the following: the method developed by Udaltsov and Shlyuger [28] was recommended together with the Knight sea gain formula [30] and the Phillips and Knight polarization coupling loss term [31]. Later in 1974, this method was adopted by the CCIR as Recommendation 435 [32]. In the subsequent sections, this method will be called the *Udaltsov–Shlyuger method*. This method, which includes a sound treatment of latitude, appeared to be very promising at that time. When applied to one-hop intra-European paths, good results were obtained [33]. After years of extensive testing against measured data from other parts of the world, however, some major limitations have surfaced. For example, when applied to paths longer than, say, 4000 km, the method has a strong tendency to underestimate field strengths, in many cases by more than 20 dB [25]. Furthermore, Region 2 data do not seem to corroborate the frequency term of this method [34]. Although this method is a great step forward from the two previous methods, it is something short of a true worldwide method.

4.1.5. The Development of Recommendation 1147-1 (the Wang Method). Knowing the clear-channel curve has some limitations and the need for more field-strength data, the FCC initiated a long-term and large-scale measurement program in 1939. Data from more than 40 propagation paths were collected. The measurement lasted for one sunspot cycle. Unfortunately, the midpoint geomagnetic latitude of the majority of the paths range from 45 °N to 56° North, a narrow window of 11°. Recognizing the need for additional data from the low- and the high-latitude areas, the FCC later initiated two separate projects. In 1980, the FCC and the Institute for Telecommunication Sciences of the Department of Commerce jointly began collecting low-latitude data at Kingsville, Texas, and at Cobo Rojo, Puerto Rico. In 1981, the FCC awarded a contract to the Geophysical Institute, University of Alaska. This project called for the acquisition of sky-wave data from the high-latitude areas. The Alaskan project lasted for about 7 years; data representing different levels of solar activity have been successfully collected. On the basis of the enlarged data bank, a new field strength prediction method was developed by Wang [34]. In 1990 the FCC adopted this method for domestic applications. In 1999 the ITU-R adopted this method as Recommendation P.1147-1 [35], replacing the Udaltsov–Shlyuger method. Like the Udaltsov–Shlyuger method, the Wang method

also contains a similar latitude term. This method has essentially linked the Cairo and the FCC clear-channel curves together mathematically. The special case corresponding to geomagnetic latitude of 35° in the Wang method is extremely close to the Cairo curve. The special case corresponding to 45° is very similar to the FCC curve. It works well for long paths and short paths alike.

4.2. Data Bank

The work of IWP 6/4 is now being handled by Working Party 3L (Ionospheric Propagation) of Study Group 3 (Propagation). The most recent version of the data bank consists of field strengths from 417 propagation paths. Great-circle lengths range from 290 to 11,890 km. Signals of the few very short paths have been verified to be sky waves. Frequencies range from 164 to 1610 kHz. Control-point geomagnetic latitudes range from 46.2° south to 63.8° north. This data bank is available to the general public from the ITU Website (*www.itu.int/brsg/sg3/databanks*).

4.3. Characteristics of LF/MF Sky-Wave Propagation

4.3.1. Amplitude Distribution. Nighttime sky-wave field strengths vary greatly from minute to minute and night to night. The within-the-hour short-term variation usually takes the form of Rayleigh distribution. Night-to-night median values of field strengths for a given reference hour are lognormally distributed. For frequency management purposes, the yearly median value of field strength at six hours after sunset is usually used to determine sky-wave (or secondary) service area of a station while the yearly upper-decile value is used to determine interference level. The difference between the annual upper-decile and the median value varies with latitude, from 6 dB in tropical areas [18] to 12 dB or more at high latitudes [36].

4.3.2. Diurnal Variation. At LF, the transition from daytime condition to nighttime condition in winter is very gradual, and field strength does not reach its maximum value until about 2 h before sunrise. The change at sunrise is more rapid. In summer, field strength increases more rapidly at sunset.

At MF, field strength changes very rapidly at sunset as well as sunrise. Field strength reaches its maximum value shortly after midnight or 6 h after sunset. U.S. data suggest that field strength is highly frequency-dependent during transition hours. For example, the signal of a 1530-kHz station is about 15 dB stronger than that of a 700-kHz station at sunrise. At 6 h after sunset, the difference is only about 3 dB in favor of the higher frequency [35].

4.3.3. Seasonal Variation. At LF and in daytime, sky waves propagating in winter are at least 20 dB stronger than in summer. At night, LF signals are strongest in summer and winter and are weakest in spring and autumn. The summer maximum is more pronounced.

At MF and in daytime, sky waves are strongest in winter months. The seasonal variation may exceed 30 dB. At night, MF sky waves propagating at temperate latitudes are strongest in spring and autumn and are weakest in summer and winter. The summer minimum is more pronounced. The overall variation may be as much as 15 dB at the lowest frequencies in the MF band, decreasing to about 3 dB at the upper end of the band. The variation is much smaller in tropical latitudes.

Nighttime high-latitude field strength data collected in Alaska show a pronounced summer minimum and a consistent maximum in April [36]. In a year of minimum sunspot number, the nighttime monthly median field strength of April is typically 10–15 dB greater than the annual median value.

4.3.4. Effect of Sunspots. At LF, the effect of sunspots is virtually nonexistent. At MF, sunspots greatly reduce sky-wave field strength levels. The reduction (L_r) is a function of sunspot number, latitude, distance, and, to a lesser degree, frequency [37].

The effect of sunspots is clearly latitude-dependent. In low-latitude areas (e.g., Central America, Mexico), annual median values of field strengths vary slightly (less than 3 dB) within a sunspot cycle and there is no detectable pattern. A pattern of correlation begins to surface at higher latitudes (e.g., central USA). For example, measured field strengths from a path in the southern parts of the United States (San Antonio, TX to Grand Island, NE; 1200 kHz, 1279 km, 45.1°N) decreased by about 3 dB when the sunspot number reached from minimum to maximum in cycle 18. The correlation becomes more pronounced at still higher latitudes. For example, measured field strengths of a path in the northern United States (Chicago, IL to Portland, OR, 890 kHz, 2891 km, 54°N) decreased by 15 dB in the same cycle. In Alaska, in a year of maximum solar activity, there are virtually no sky waves from northern-tier U.S. stations, although signals can be very strong in a year of low or moderate solar activity [36].

The effect of sunspots has a diurnal variation of its own. In other words, L_r is different at different hours of the night. At 6 h after sunset, L_r is considerably smaller than that at two hours after sunset. For example, consider a path from Cincinnati, OH to Portland, OR (700 kHz, 3192 km, 53.2°N). From 1944 (a year of minimum sunspot number) to 1947 (a year of maximum sunspot number), field strength for the sixth hour after sunset decreased by 7.3 dB; that for the fourth hour after sunset decreased by 13.3 dB, and that for the second hour after sunset decreased by 16.9 dB [37].

4.3.5. Effect of the Magnetic Field. When a radiowave enters the ionosphere in the presence of a magnetic field, it is split into two components: the ordinary and the extraordinary waves. Both are elliptically polarized. The extraordinary wave is then absorbed. Further loss occurs when the wave leaves the ionosphere. Because of the nature of elliptical polarization, only its vertical component normally couples with the receiving antenna. This process is known as polarization coupling loss (L_p). At LF, L_p is negligible. At MF, L_p is negligible in high and temperate latitudes. In tropical areas, however, it can be very large and depends on the direction of propagation relative to that of the earth's magnetic field. In some

extreme cases (e.g., east–west paths in equatorial Africa), polarization coupling losses of more than 20 dB have been observed. This phenomenon is not yet fully understood, and more data are needed. An interim formula, however, has been developed by Phillips and Knight [31].

4.3.6. Influence of Seawater. When at least one terminal of a path is situated near the sea and a significant portion of the path is over seawater, the received signal is significantly stronger than otherwise. This is commonly called sea gain (G_s). Sea gain is a complicated function of several factors, including pathlength (i.e., elevation angle), distance from antenna to the sea, and frequency. Under ideal conditions (elevation angle = 0, antenna is on the coast), sea gain is about 4 dB at LF and about 10 dB at MF. For a more detailed discussion, see a paper by Knight [30].

4.3.7. Propagation at Daytime. Daytime measurements from more than 30 paths are believed to be sky waves and have been studied by Wang [38]. Some trends are briefly stated as follows:

LF Cases. LF sky-wave field strengths at noon can be surprisingly strong, particularly in winter months. Daytime annual median field strength is typically 20 dB lower than its counterpart at night. Daytime upper-decile value is about 13 dB stronger than the median value.

MF Cases. MF sky-wave field strengths at noon display a consistent seasonal variation pattern with maximum occurring in winter months. The average winter-month field strength is about 10 dB stronger than the annual median value. The winter : summer ratio can exceed 30 dB. The annual median value of field strength at noon is about 43 dB lower than its counterpart at 6 h after sunset. Field strength exceeded for 10% of the days of the year is about 13 dB stronger than the median value.

4.4. Comparison of Predicted and Measured Field Strengths

An extensive comparative study using the most recent data bank has been carried out [25]. For each and every propagation path in the data bank, field strengths have been calculated by using the four most popular methods discussed in this article. Polarization coupling loss and sea gain, if applicable, have been included. In a very small number of cases where measurements were taken in a year of maximum sunspot number, the Wang formula [37] for sunspot losses has also been used. Calculated results are compared with measured data, and prediction errors are thus obtained and analyzed. Prediction errors are analyzed from four different viewpoints (path length, latitude, frequency, and geographic regions) and tabulated in detail [25]. In this article, only pathlength and latitude are considered.

4.4.1. Pathlength and Path Accuracy. We arbitrarily define a short path as one whose length is shorter than 2500 km, a medium path between 2500 and 4999 km, and a long path greater than 5000 km. Then there are 267 short paths, 85 medium paths, and 65 long paths. On the RMS (root-mean-square) basis, the prediction errors of the Cairo curve are 6.6 dB for short paths, 7.5 dB for medium paths, and 10.1 dB for long paths. The corresponding errors of the Region 2 method are 6.6, 7.5, and 12.0 dB, respectively. Those of the Udaltsov–Shlyuger method are 5.7, 7.8, and 16.4 dB, respectively. The errors of the Wang method are 5.5, 6.5, and 6.7 dB, respectively. The Wang method is the only method that offers good to excellent results in all distance ranges.

4.4.2. Latitude and Accuracy. In this section we arbitrarily define low latitudes as those between 0° and 35° (geomagnetic), temperate latitudes as those between 35.1° and 50°, and high latitudes as those greater than 50°. Then there are 203 paths in the low-latitude areas, 152 paths in the temperate-latitude areas, and 62 paths in the high latitudes. On the RMS basis, the prediction errors of the Udaltsov–Shlyuger method are 7.9 dB in low-latitude areas, 6.5 dB in temperate-latitude areas, and 14.0 dB in high-latitude areas, respectively. On the other hand, the corresponding errors of the Wang method are 5.7, 5.6, and 4.4 dB, respectively.

In summary, the Wang method is the only method that offers good to excellent results for short and long paths alike, at all frequencies in the LF/MF bands, at all latitudes, and in all regions.

4.5. The Recommended Propagation Model

In this section we recommend and present the Wang method. For a complete step-by-step procedure, see ITU-R Recommendation P.1147-1 [35]. This section is not meant to be self-contained. Only the most important equations are given here.

4.5.1. Annual Median Field Strength. According to the Wang method, the annual median sky-wave field strength at 6 h after sunset, E (in decibels above 1 μV/m), is given by

$$E = P + G + (A - 20\log p) - k\left(\frac{p}{1000}\right)^{0.5} - L_p + G_s - L_r \quad (4)$$

$$k = 2\pi + 4.95\tan^2(\Phi) \quad (5)$$

where P = radiated power in dB above 1 kW
G = transmitting antenna gain (dB)
A = a constant (at LF, $A = 110.2$; at MF, $A = 110$ in Australia and New Zealand; and $A = 107$ in all other places)
p = actual slant distance of the path under study, in kilometers, assuming that average height of E layer is 100 km
Φ = geomagnetic latitude of the midpoint of the path under study in degrees
L_p = polarization coupling loss (dB) [31]
G_s = sea gain (dB) [30]
L_r = loss of field strength due to solar activity (dB) [37]

4.5.2. Upper Decile Field Strength. Field strength exceeded for 10% of the nights of a year $E(10)$, is greater than the annual median value by Δ dB. Then

$$\Delta = 0.2|\Phi| - 2 \tag{6}$$

where Δ is limited between 6.0 and 10 dB.

5. SKY-WAVE PROPAGATION AT HF

5.1. General Description of HF Propagation

HF sky-wave propagation may be represented by rays between the ground and the ionosphere. In the ionosphere, the radiowaves experience dispersion and changes in polarization. The propagation is affected by, among other factors, ionization, operating frequency, ground conductivity and elevation angle. HF waves in the ionosphere undergo continuous refraction (i.e., bending of the ray path). At any given point, refraction is less at lower electron densities, for higher frequencies, and for higher elevation angle. For a given elevation angle, there exists a certain frequency below which the rays will be reflected back to earth. At a higher frequency, the refraction is too low for the rays to be returned to earth. Waves launched vertically may be reflected, if their frequency is below the "critical frequency" (see Section 6).

The apparent height of reflection varies between about 100 and 300 km. Radiowaves that are launched more obliquely travel to greater range. The maximum range attained after one hop arises for rays launched at grazing incidence. For typical E, F1, and F2 layers, it is about, 2000, 3400, and 4000 km, respectively. In HF communication, several propagation paths are often possible between a given transmitter and a given receiver, such as a single reflection from the E region (1E mode), a single reflection from the F region (1F mode), and double reflection from the F region (2F mode). Mode 2F is said to have higher "order" than mode 1F in propagation terms. This feature is known as *multipath*.

At frequencies above the critical frequency, there is an area surrounding the transmitter defined by "skip distance" in which sky wave cannot be received because the elevation angle is too high. The maximum usable frequency (MUF), a very important concept in HF propagation, may be defined as the frequency that makes the distance from the transmitter to a given reception point equal to the skip distance (see also Section 6). The MUF increases with pathlength and decreases with the height of the ionospheric layer. The MUF also undergoes diurnal, seasonal, solar cycle, and geographic variations. The MUF tends to be high during the day and low during the night. Also, the MUF is higher in summer than in winter during the night. Furthermore, the MUF tends to increase with increasing sunspot number. The F2-layer MUF may increase as much as 100% from sunspot minimum and sunspot maximum. The MUF has a very complex geographic variation. The most authoritative presentation of MUF is undoubtedly the CCIR Report 340, *Atlas of Ionospheric Characteristics* [39], which presents world maps of MUF for the F2 layer corresponding to different month of the year, solar activity levels, and distance ranges.

5.2. Fading

5.2.1. Interference Fading. Interference fading results from interference between two or more waves, which travel by different paths to arrive at the receiving point. This type may be caused by interference between multiple reflected sky waves, sky wave, and ground wave. This type of fading may last for a period of a fraction of a second to a few seconds, during which time the resultant field intensity may vary over wide limits.

5.2.2. Polarization Fading. Polarization fading occurs as a result of changes in the direction of polarization of the downcoming wave, relative to the orientation of the receiving antenna, due to random fluctuations in the electron density along the path of propagation. This type of fading also lasts for a fraction of a second to a few seconds.

5.2.3. Absorption Fading. Absorption fading is caused by variation in the absorption due to changes in the densities of ionization, and it may sometimes last longer than one hour.

5.2.4. Skip Fading. Skip fading may be observed at receiving locations near the skip distance at about sunrise and sunset, when the basic MUF for the path may oscillate around the operating frequency. The signal may decrease abruptly when the skip distance increases past the receiving point (or increase with a decrease in the skip distance).

5.3. Regional Anomalies

5.3.1. Tropical Anomalies. In the tropical zone, sky-wave propagation is characterized by the presence of equatorial sporadic E and the spread F. Equatorial sporadic E (Es-q), which appears regularly during daytime in a narrow zone near the magnetic equator, is the principal cause for fading at daytime. In the equatorial zone after local sunset, some irregularities develop in the F-region ionization and are called *spread F*. Under these conditions, the F region increases markedly in height and seems to break up into patchy irregular regions. As a result, a peculiar type of rapid fading, called flutter fading, usually occurs after sunset. Flutter fading is one of the most important factors in the degradation of HF broadcast service in tropical areas. Flutter fading is most pronounced following the equinoxes. Flutter fading correlates negatively with magnetic activity. On magnetically quiet days, it is usually evident, whereas on magnetically disturbed days, it is absent. The fading rate is proportional to the wave frequency and may range between 10 and 300 per minute [19].

5.3.2. High-Latitude Anomalies. At high latitudes, the ionosphere is exposed to the influence of disturbances in interplanetary space and in the magnetosphere, since the magnetic field lines extend far from the earth. Electrically charged particles can move easily along the field lines and perturb the high-latitude ionosphere. The

absorption is inversely proportional to frequency squared. The absorption may be preceded by a sudden ionospheric disturbance (SID) on the sunlit side of the earth, at all latitudes, caused by X rays from solar flares. At HF, absorption can be greater than 100 dB [40,41]. The magnetic storm-related absorption in the sunlit part of the polar cap is much stronger than in the dark side. The average duration of the event is about 2 days, but may be as long as 4 days. It may spread to lower latitudes too.

5.4. Predicting HF Sky-Wave Field Strength

The calculation of HF field strengths is a very complicated process. It requires a computer. In the succeeding section, a survey of existing programs will be presented. In this section, only a brief outline of the calculation procedure is given. The purpose is to illustrate the general procedures and terms involved. For a more detailed presentation, see, for example, ITU-R Recommendation PI.533-4 [42].

The median value of sky-wave field strength for a given mode of propagation, in dB (μV/m), is given by

$$E_{ts} = 136.6 + P_t + G_t + 20\log f - L_{bf} - L_i - L_m - L_g - L_h - 12.2 \quad (7)$$

where P_t = transmitter power in dB relative to 1 kW
G_t = transmitting antenna gain (dB)
f = transmitting frequency (MHz)
L_{bf} = basic free space transmission loss = $32.45 + 20\log f + 20\log p$ (8)
p = slant distance (km)
L_i = absorption loss (dB)
L_m = above MUF loss (dB)
L_g = ground reflection loss (dB)
L_h = auroral and other signal losses

5.5. Performance Prediction Software

A large number of computer programs have been developed for predicting HF circuit performance. The following is a brief list of the programs that are widely used today. For an excellent discussion on this topic, see a paper by Rush [43].

5.5.1. IONCAP. Ionospheric Communications Analysis and Prediction Program (IONCAP) was developed by staff of the Institute for Telecommunication Sciences (ITS) of the National Telecommunications and Information Administration (NTIA), Department of Commerce [44]. The propagation features include refraction bending, scattering on frequencies above the MUF, and sporadic E propagation. The predicted field strength and noise levels can help the designer to determine optimum frequencies, correct antennas, required transmitter powers. IONCAP is available from NTIA/ITS, Department of Commerce, Boulder, CO (USA).

5.5.2. VOACAP. At the request of the Voice of America, IONCAP has been modified and improved [45]. The resultant program, VOACAP, is available from ITS Website *http://elbert.its.bldrdoc.gov/hf.html*.

5.5.3. ITU-R Recommendation 533-4 (REC533). In preparation for the 1984 HF World Administrative Radio Conference (WARC HFBC-84), the CCIR established Interim Working Party 6/12. After extensive deliberations, it adopted the following. For paths shorter than 7000 km, IWP 6/12 adopted a simplified version of the method described in CCIR Report 252-2, similar to IONCAP. For paths longer than 9000 km, the FTZ method [46] was adopted. For in-between paths, a linear interpolation scheme is used. The FTZ method has been known for its simplicity and accuracy when applied to very long paths. Results of the work of IWP 6/12 are documented in Recommendation PI.533-4 [42]. This software is known as REC533, available from the ITU, Geneva, Switzerland; also available from the abovementioned ITS Website.

5.5.4. Input Data and Results of Calculations. In order to use any of the aforementioned programs, the following required input information is usually needed for each given circuit: (1) time of day, month, and year; (2) expected sunspot number; (3) antenna type; (4) geographic locations of the transmitter and receiver; (5) human-made noise level; (6) required reliability; and (7) required signal-to-noise ratio. The results of calculations usually include the following: (1) great-circle and slant distances, (2) angles of departure and arrival, (3) number of hops, (4) time delay of the most reliable propagation mode, (5) the virtual height, (6) MUF and the probability that the frequency exceeds the predicted MUF, (7) median system loss in dB, (8) median field strength in dB above 1 μV/m, (9) median signal power in dBW, (10) median noise power in dBW, (11) median signal/noise ratio in dB, and (12) LUF (the lowest usable frequency).

6. GLOSSARY

CCIR. French acronym for International Radio Consultative Committee (now ITU-R).

Critical frequency fo. The highest frequencies at which a radio wave is reflected by a layer of the ionosphere at vertical incidence. There is usually one such frequency for each ionospheric component (e.g., foE, foF2). The critical frequency is determined by the maximum electron density in that layer. Waves with their frequency below fo will be reflected. As the frequency is increased beyond this, the ray will penetrate the layer.

Fading. The temporary and significant decrease of the magnitude of the electromagnetic field or of the power of the signal due to time variation or the propagation conditions.

Free-space propagation. Propagation of an electromagnetic wave in a homogeneous ideal dielectric medium, which may be considered of infinite extent in all directions.

Frequency band. Continuous set of frequencies in the frequency spectrum lying between two specific limiting frequencies; generally includes many channels.

Low-frequency (LF) band. The part of the spectrum between 30 and 300 kHz. This band is also known as band 5 because the center frequency is 1×10^5 Hz. The corresponding waves are sometimes called the *kilometric* or *long waves*.

Medium-frequency (MF) band. The part of the spectrum between 300 and 3000 kHz. This band is also known as band 6. The corresponding waves are sometimes called the *hectometric* or *medium waves*.

High-frequency (HF) band. The part of the spectrum between 3 and 30 MHz. This band is also known as band 7. The corresponding waves are sometimes called *decametric* or *short waves*.

ITU. International Telecommunication Union.

ITU-R. Radiocommunication Study Groups of the ITU.

ITU Region 1. Africa, Europe, the entire territory of the former USSR, Outer Mongolia, and Turkey.

ITU REGION 2. The Americas and Greenland.

ITU REGION 3. Australia, New Zealand, and all other Asian countries.

Ionosphere. The ionized region of the earth's upper atmosphere.

MUF. Maximum usable frequency.

Basic MUF. The highest frequency by which a radiowave can propagate between given terminals, on a specific occasion, by ionospheric refraction alone. Where the basic MUF is restricted to a particular propagation mode, the values may be quoted together with an indication of that mode (e.g., 2F2 MUF, 1E MUF). Furthermore, it is sometimes useful to quote the ground range for which the basic MUF applies. This is indicated in kilometers following the indication of the mode type [e.g., 1F2(4000) MUF].

Operational MUF (or simply MUF). The highest frequency that would permit acceptable performance of a radio circuit by signal propagation via the ionosphere between giving terminals at a given time under specific working conditions.

Median values of field strengths—yearly median. The median of daily values for the year, usually for a given reference hour.

Multipath propagation. Propagation of the same radio signal between a transmission point and a reception point over a number of separate propagation paths.

Noise

- **Atmospheric noise**. Radio noise produced by natural electric discharges below the ionosphere and reaching the receiving point along the normal propagation paths between the earth and the lower limit of the ionosphere.
- **Human-made noise**. Radio noise having its source in synthetic (human-made) devices.
- **Galactic noise**. Radio noise arising from natural phenomena outside the earth's atmosphere.

Propagation. Energy transfer between two points without displacement of matter.

Reliability. Probability that a specific performance is achieved.

- **Basic reliability**. The reliability of communications in the presence of background noise alone.
- **Overall reliability**. The reliability of communications in the presence of background noise and of known interference.

Skip distance. The minimum distance from the transmitter at which a sky wave of a given frequency will be returned to earth by the ionosphere.

Solar activity. The emission of electromagnetic radiation and particles from the sun, including slowly varying components and transient components caused by phenomena such as solar flares.

Sudden ionospheric disturbance (SID). A sudden marked increase in electron density of the lower ionosphere during daylight hours. This is caused by X-ray emission from the sun.

Transmission loss. The ratio, usually expressed in decibels, for a radio link between the power radiated by the transmitting antenna and the power that would be available at the receiving antenna output.

Basic free-space transmission loss (L_{bf}). The transmission loss that would occur if the antennas were replaced by isotropic antennas located in a perfectly dielectric, homogeneous, isotropic, and unlimited environment [see also Eq. (8)].

Zenith angle. The angle between the sun and the zenith (i.e., directly overhead) at a given geographic location.

BIOGRAPHY

John C. H. Wang received his B.S. degree in electrical engineering in 1959 from the University of Maryland, and his M.S. degree in electrical engineering in 1968 from the University of Pittsburgh, Pennsylvania. He has been with the technical research staff of the Federal Communications Commission in Washington D.C. since 1969. His main area of research is ionospheric propagation and has published more than 30 technical papers. He is also active in the Radiocommunication Study Groups of the Telecommunication Union (ITU-R) and chairs its Working Party on Ionospheric Propagation. His other interests include astronomy and Chinese history. With that unique combination, he has written papers on the sighting of sunspots in ancient China and on the subject of unearthing the star of Bethlehem from Chinese history. He is a fellow of the IEEE.

BIBLIOGRAPHY

1. A. Summerfield, The propagation of waves in wireless telegraphy, *Ann. Physik* **28**: 665–736 (1909).
2. G. N. Watson, The diffraction of radio waves by the earth, *Proc. Roy. Soc. A* **95**: 83–99 (1918).
3. H. Bremmer, *Terrestrial Radio Waves*, Elsevier, Amsterdam, 1949.

4. K. A. Norton, Propagation of radio waves over the surface of the earth and in the upper atmosphere part I, *Proc. IRE* **24**(10): 1367–1387 (1936).
5. Federal Communications Commission, Standards for good engineering practice concerning standard broadcast stations, *Fed. Reg.* (4FR 2862) (July 8, 1939).
6. K. A. Norton, The calculation of ground-wave field intensity over a finitely conducting spherical earth, *Proc. IRE* **29**(12): 623–639 (1941).
7. ITU-R, *Ground-Wave Propagation Curves for Frequencies between 10 kHz and 30 MHz*, Recommendation PN.368-7, Geneva, ITU, 1994.
8. L. A. Berry, *User's Guide to Low-Frequency Radio Coverage Program*, Office of Telecommunications TM 78-247, 1978.
9. S. Rotheram, Ground wave propagation, part 1: theory for short distances; Part 2: Theory for medium and long distances, *Proc. IEEE* **128**(5): 275–295 (1981).
10. R. P. Eckert, *Modern Methods for Calculating Ground-Wave Field Strength over Smooth Spherical Earth*, FCC Report OET R 8601, 1986.
11. E. Haakinson, S. Rothschild, and B. Bedford, *MF Broadcasting System Performance Model*, NTIA Report 88-237, Dept. Commerce, Washington, DC, 1988.
12. H. L. Kirke, Calculation of ground-wave field strength over a composite land and sea path, *Proc. IRE* **37**(5): 489–496 (1949).
13. G. Millington, Ground wave propagation over an inhomogeneous smooth earth, *Proc. IEE* **96**(39) (Pt. III): 53–64 (1949).
14. K. N. Stokke, Some graphical considerations on Millington's method for calculating field strength over inhomogeneous earth, *Telecommun. J.* **42**(Pt. III): 157–163 (1975).
15. H. Fine, An effective ground conductivity map for continental United States, *Proc. IRE* **49**: 1405–1408 (1954).
16. ITU-R, *World Atlas of Ground Conductivities*, Recommendation P.832-2, Geneva, ITU, 1999.
17. O. Heaviside, The theory of electric telegraphy, in *Encyclopedia Britannica*, 10th ed., 1902.
18. K. Davies, *Ionospheric Radio*, Peregrinus, London, 1990.
19. J. Goodman, *HF Communications: Science and Technology*, Van Nostrand, New York, 1992.
20. Ban Ku, *Book of Han*, 99, published for the first time about 92 A.D., reedited and republished under the supervision of Emperor Chien Lung in 1736; available from many publishers, including Yee Wen Press, Taipei.
21. Fang Shyuan Ling, *Book of Jin*, 12, published for the first time about 640 A.D., reedited and republished under the supervision of Emperor Chien Lung in 1736, available from many publishers including Yee Wen Press, Taipei.
22. ITU-R, *Radio Noise*, Recommendation PI.372-6, ITU, Geneva, 1994.
23. F. Horner, Analysis of data from lightning flash counters, *Proc. IEE* **114**: 916–924 (1967).
24. G. Gustafsson, A revised corrected geomagnetic coordinate system, *Arkiv Geofysik* **5**: 595–617 (1970).
25. J. C. H. Wang, An objective evaluation of available LF/MF sky-wave propagation models, *Radio Sci.* **34**(3): 703–713 (1999).
26. International Telecommunication Union, *Final Acts of the Regional Administrative LF/MF Broadcasting Conference (Regions 1 and 3), ITU, Geneva, 1975*, Geneva, 1976.
27. International Telecommunication Union, *Final Acts of the Regional Administrative MF Broadcasting Conference (Region 2) Rio de Janeiro, 1981*, ITU, Geneva, 1982.
28. A. N. Udaltsov and I. S. Shlyuger, Propagation curves of the ionospheric wave at night for the broadcasting range, *Geomag. i Aeronom.* **10**: 894–896 (1972).
29. C. Nemeto, N. Wakai, M. Ose, and S. Fujii, Integrated results of the mobile measurements of MF field strength along the Japan-Antarctica sailing course, *Rev. Radio Res. Lab.* **33**(168): 157–182 (1987).
30. P. Knight, *LF and MF Propagation: An Approximate Formula for Estimating Sea Gain*, BBC Report RD 1975/32, 1975.
31. G. J. Phillips and P. Knight, Effects of polarisation on a medium-frequency sky-wave service, including the case of multihop paths, *Proc. IEE* **112**: 31–39 (1965).
32. CCIR, *Sky-Wave Field Strength Prediction Method for Frequency Range 150 to 1600 kHz*, Recommendation 435, 1974.
33. J. C. H. Wang, P. Knight, and V. K. Lehtoranta, A study of LF/MF skywave data collected in ITU Region 1, in J. M. Goodman, ed., *Proc. 1993 Ionospheric Effects Symp.*, Alexandria, VA, 1993.
34. J. C. H. Wang, Prudent frequency management through accurate prediction of skywave field strengths, *IEEE Trans. Broadcast.* **35**(2): 208–217 (1989).
35. ITU-R, *Prediction of Sky-Wave Field Strength at Frequencies between about 150 and 1700 kHz*, Recommendation P.1147-1, 1999.
36. R. D. Hunsucker, B. S. Delana, and J. C. H. Wang, Medium-frequency skywave propagation at high latitudes: Results of a five-year study, *IEEE Trans. Broadcast.* **35**(2): 218–222 (1989).
37. J. C. H. Wang, Solar activity and MF skywave propagation, *IEEE Trans. Broadcast.* **BC-35**(2): 204–207 (1989).
38. J. C. H. Wang, LF/MF skywave propagation at daytime, *IEEE Trans. Broadcast.* **BC-41**(1): 23–27 (1995).
39. CCIR, *Atlas of Ionospheric Characteristics*, Report 340, ITU, Geneva, 1983.
40. R. D. Hunsucker, B. S. Delana, and J. C. H. Wang, Effects of the 1986 magnetic storm on medium frequency skywave signals recorded at Fairbanks, Alaska, in J. Goodman, ed., *Proc. IES'87*, 1987, pp. 197–204.
41. R. D. Hunsucker, Anomalous propagation behavior of radio signals at high latitudes, in H. Soicher, ed., *AGARD Conf. Proc. P-332*, 1982.
42. ITU-R, *HF Propagation Prediction Method*, Recommendation PI.533-4, 1994.
43. C. M. Rush, Ionospheric radio propagation models and predictions: A mini review, *IEEE Trans.* **AP-34**: 1163–1170 (1986).
44. L. R. Teters, J. L. Lloyd, G. W. Haydon, and D. L. Lucas, *Estimating the Performance of Telecommunication Systems Using the Ionospheric Transmission Channel-Ionospheric Communications Analysis and Predictions Programs User's Manual*, NTIA Report 83-127, NTIS Access No. PB84-111210, 1983.

45. G. Lane, *Signal-to-Noise Prediction Using VOACAP*, Rockwell Collins, Cedar Rapids, IA, 2000.

46. A. Ochs, The forecasting system of the Fernmeldetechischen Zentralamt (FTZ), in V. Agy, ed., *AGARD Conf. Proc. P-49*, 1970.

RATE-DISTORTION THEORY

TOBY BERGER
Cornell University
Ithaca, New York

Coding, in the sense of C. E. Shannon's information theory [1], divides naturally into channel coding and source coding. Source coding exhibits a further dichotomy into lossless source coding and lossy source coding. Lossless source coding is treated in a separate article in this encyclopedia. Here, we discuss the theory of lossy source coding, also widely known as *rate-distortion theory*.[1]

Shannon's principal motivation for including "Section V: The Rate for a Continuous Source" in his celebrated paper was that it provided a means for extending information theory to analog sources. Since all analog sources have infinite entropy, they cannot be preserved losslessly when stored in or transmitted over physical, finite-capacity media.

Shannon's famous formula for the capacity of an ideal band-limited channel with an average input power constraint and an impairment of additive, zero-mean, white Gaussian noise reads

$$C = W \log_2(1 + P/N) \quad \text{bits/s} \tag{1}$$

In this formula — the most widely known and the most widely abused of Shannon's results — P is the prescribed limitation on the average input power, W is the channel bandwidth in positive frequencies measured in Hz, and N is the power of the additive noise. Since the noise is white with one-sided power spectral density N_0 or two-sided power spectral density $N_0/2$, we have $N = N_0 W$. Common abuses consist of applying (1) when

1. The noise is non-Gaussian.
2. The noise is not independent of the signal and/or is not additive.
3. Average power is not the (only) quantity that is constrained at the channel input.
4. The noise is not white across the passband and/or the channel transfer function is not ideally bandlimited.

Abuse (1) is conservative in that it underestimates capacity because Gaussian noise is the hardest additive noise to combat. Abuse (2) may lead to grossly underestimating or grossly overestimating capacity. A common instance of abuse (3) consists of failing to appreciate that it actually may be peak input power, or perhaps both peak and average input power, that are constrained. Abuse (4) leads to an avoidable error in that the so-called "water pouring" result [8], generalizing (1), yields the exact answer when the noise is not white, the channel is not bandlimited, and/or the channel's transfer function is not flat across the band. (See also [3,4].) There is a pervasive analogy between source coding theory and channel coding theory. The source coding result that corresponds to (1) is

$$R = W \log_2(S/N) \quad \text{bits/s} \tag{2}$$

It applies to situations in which the data source of interest is a white Gaussian signal bandlimited to $|f| \leq W$ that has power $S = S_0 W$, where S_0 denotes the signal's one-sided constant power spectral density for frequencies less than W. The symbol, N, although often referred to as a "noise," is actually an estimation error. It represents a specified level of mean squared error (MSE) between the signal $\{X(t)\}$ and an estimate $\{\hat{X}(t)\}$ of the signal constructed on the basis of data about $\{X(t)\}$ provided at a rate of R bits/s. That is,

$$N = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt\, E[\hat{X}(t) - X(t)]^2$$

It was, and remains, popular to express MSE estimation accuracy as a "signal-to-noise ratio," S/N, as Shannon did in Eq. (2). It must be appreciated, however, that $\{\hat{X}(t) - X(t)\}$ is not noise in the sense of being an error process that is independent of $\{X(t)\}$ that nature adds to the signal of interest. Rather, it is a carefully contrived error signal, usually dependent on $\{X(t)\}$, that the information/communication theorist endeavors to create in order to conform to a requirement that no more than R bits/s of information may be supplied about $\{X(t)\}$. In modern treatises on information theory, the symbol "D," a mnemonic for average distortion, usually is used in place of N. This results in an alternative form of (2), namely

$$R(D) = W \log_2(S/D) \quad \text{bits/s} \tag{3}$$

which is referred to as the MSE rate-distortion function of the source.

Equation (3) also suffers widespread abuse which takes the form of applying it to situations in which

1. The signal is non-Gaussian.
2. Distortion does not depend simply on the difference of $\hat{X}(t)$ and $X(t)$
3. Distortion is measured by a function of $\hat{X}(t) - X(t)$ other than its square.
4. The signal's spectral density is not flat across the band.

Again, abuse (1) is conservative in that it results in an overestimate of the minimum rate R needed to achieve a specified MSE estimation accuracy because white Gaussian sources are the most difficult to handle in the sense of bit rate versus MSE. Abuses (2) and (3),

[1] This article is drawn in considerable measure from the first half of a survey paper on lossy source coding by T. Berger and J. Gibson [2] that appeared in the 50th Anniversary Issue of the *IEEE Transactions on Information Theory* in October 1998. There have been some additions, some corrections, some enhancements, and many deletions.

which often stem in practice from lack of knowledge of a perceptually appropriate distortion measure, can result in gross underestimates or overestimates of R. Abuse (4) can be avoided by using a water-pouring generalization of (3) which we shall soon discuss. Toward that end we recast (3) in the form

$$R(D) = \max[0, W \log(S_0 W/D)] \tag{4}$$

This explicitly reflects the facts that (2) the signal spectrum has been assumed to be constant at level S_0 across the band of width W in which it in nonzero, and (2) $R(D) = 0$ for $D \geq S_0 W$, because one can achieve a MSE of $S = S_0 W$ without sending any information simply by guessing that $X(t) = m(t)$, the signal's possibly time-varying deterministic mean value function. The base of the logarithm in Eq. (4) determines the information unit — bits for $\log_2$ and nats for $\log_e$. When we deal with continuously distributed quantities, it is more "natural" to employ natural logs. When no log base appears, assume henceforth that base e is intended.

A basic inequality of information theory is

$$D \geq R^{-1}(C) \tag{5}$$

which we shall refer to as the *information transmission inequality*. It says that if you are trying to transmit data from a source with rate-distortion function $R(D)$ over a channel of capacity C, you can achieve only those average distortions that exceed the distortion-rate function evaluated at C. (The distortion-rate function is the inverse of the rate-distortion function; denoted $D(R)$, it always exists because $R(D)$ always is convex.)

Suppose, for example, that we wish to send data about the aforementioned band-limited white-Gaussian process $\{X(t)\}$ over an average-input-power-limited, ideally band-limited AWGN channel and then construct on the basis of the channel output an approximation $\{\hat{X}(t)\}$ that has the least possible MSE. The source and the channel have the same frequency band $|f| \leq W$. Since $R(D) = W \log_2(S/D)$, the distortion-rate function is

$$D(R) = S 2^{-R/W}$$

so Eqs. (1) and (4) together tell us that

$$D \geq D(C) = S \exp\left[-\frac{W \log(1 + P/N)}{W}\right]$$

or

$$D/S \geq (1 + P/N)^{-1} \tag{6}$$

This tells us that the achievable error power per unit of source power (i.e., the achievable normalized MSE) is bounded from below by the reciprocal of one plus the channel SNR. There happens to be a trivial scheme for achieving equality in Eq. (5) when faced with the communication task in question. It consists of the following steps:

Step 1. Transmit $X(t)$ scaled to have average power P; that is, put $\sqrt{P/S}X(t)$ into the channel.

Step 2. Set $\hat{X}(t)$ equal to the MMSE estimate of $X(t)$ based solely on the instantaneous channel output $\sqrt{P/S}X(t) + N(t)$ at time t.

Since the signal and the channel noise are jointly Gaussian and zero mean, the optimum estimate in Step 2 is simply a linear scaling of the received signal, namely

$$\hat{X}(t) = \alpha[\sqrt{P/S}X(t) + N(t)]$$

The optimum α, found from the requirement that the error of the optimum estimator must be orthogonal to the data, is $\alpha = \sqrt{PS}/(P + N)$. The resulting minimized normalized MSE is easily computed to be

$$D/S = (1 + P/N)^{-1} \tag{7}$$

which means we have achieved equality in Eq. (5).

Thus, the simple two-step scheme of instantaneously scaling appropriately at the channel input and output results in an end-to-end communication system that is optimum. No amount of source and/or channel coding could improve upon this in the MSE sense for the problem at hand. This fortuitous circumstance is attributable to a double coincidence. The first coincidence is that the source happens to be the random process that drives the channel at capacity. That is, the given source, scaled by $\sqrt{P/S}$, is that process of average power not exceeding P which maximizes the mutual information between the input and output of the channel. The second coincidence is that the channel just happens to provide precisely the transition probabilities that solve the MSE rate-distortion problem for the given source. That is, when the channel is driven by the scaled source, its output minimizes mutual information rate with the source over all processes from which one can calculate an approximation to the source that achieves a normalized MSE not in excess of $(1 + P/N)^{-1}$.

We are operating at a saddle point at which the *mutual information* rate is simultaneously maximized subject to the average power constraint and minimized subject to the average distortion constraint. The slightest perturbation in any aspect of the problem throws us out this saddle — unequal source and channel bandwidths, non-Gaussianness of the source or channel, an error criterion other than MSE, and so on. The result of any such perturbation is that, in order to recover optimality, it is in general necessary to code both for the source and for the channel. From 1949 to 1958 no research was reported on rate-distortion theory in the United States or Europe. However, there was a stream of activity during the 1950s at Moscow University by members of Academician *A. N. Kolmogorov's* probability seminar. Kolmogorov saw in Shannon's entropy rate an ideal candidate for the long-sought "invariant" in the challenging isomorphism problem of ergodic theory. Kolmogorov and Sinai [5,7] succeeded in showing that equal entropy rates were a necessary condition for isomorphism of ergodic flows. Years later, Ornstein [9] proved sufficiency within an appropriately defined broad class of random stationary processes comprising all finite-order Markov sources and

their closure in a certain metric space that will not concern us here. With the Moscow probability seminar's attention thus turned to information theory, it is not surprising that some of its members also studied Shannon's Section V, The Rate for a Continuous Source. Pinsker, Dobrushin, Iaglom, Tikhomirov, Oseeyevich, Erokhin, and others made contributions to a subject that has come to be called ε-entropy, a branch of mathematics that is intimately intertwined with what information theorists today call rate-distortion theory. Thus, when invited to address an early information theory symposium, Kolmogorov [8] reported without attribution the exact answer for the ε-entropy of a stationary Gaussian process with respect to the squared L_2-norm. That result, and its counterpart for the capacity of a power-constrained channel with additive colored Gaussian noise, have come to be known as the "water pouring" formulas of information theory. In this generality the channel formula is attributable to [12] and the source formula to Pinsker [13,14].

Accordingly, we shall call them the Shannon–Pinsker water pouring formulas. They generalize the formulas given by Shannon in 1948 for the important case in which the spectrum of the source or of the channel noise is flat across a band and zero elsewhere.

The Shannon–Pinsker water pouring formula for the MSE information rate of a Gaussian source can be derived by using the fact that processes formed by bandlimiting a second-order stationary random processes to nonoverlapping frequency bands are uncorrelated with one another. In the case of a Gaussian process, this uncorrelatedness implies independence. Thus, we can decompose a Gaussian process $\{X(t)\}$ with one-sided spectral density $S(f)$ into independent Gaussian processes $\{X_i(t)\}, i = 0, 1, \ldots$ with respective spectral densities $S_i(f)$ given by

$$S_i(f) = \begin{cases} S(f), & \text{if } i\Delta \le f < (i+1)\Delta; \\ 0, & \text{otherwise} \end{cases}$$

Let us now make Δ sufficiently small that $S_i(f)$ becomes effectively constant over the frequency interval in which it is nonzero, $S_i(f) \approx S_i, i\Delta \le f < (i+1)\Delta$. It is easy to see that the best rate-distortion tradeoff we can achieve for subprocess $\{X_i(t)\}$ is

$$R_i(D_i) = \max[0, \Delta \log(S_i\Delta/D_i)]$$

By additively combining said approximations over all the subprocesses, we get an approximation to $\{X(t)\}$ that achieves an average distortion of

$$D = \sum_i D_i$$

and requires a total coding rate of

$$R = \sum_i R_i(D_i) = \sum_i \max[0,\ \Delta \log(S_i\Delta/D_i)]$$

In order to determine the MSE rate-distortion function of $\{X(t)\}$, it remains only to select those D_i's summing to D which minimize this R. Toward that end we set

$$d(R + \lambda D)/dD_i = 0, i = 0, 1, 2, \ldots,$$

where λ is a Lagrange multiplier subsequently selected to achieve a desired value of D or of R. It follows that the D and R values associated with parameter value λ are

$$\begin{aligned} D_\lambda &= \sum_{\{i:S_i\Delta > (\lambda\Delta)^{-1}\}} (\lambda\Delta)^{-1} + \sum_{\{i:S_i\Delta \le (\lambda\Delta)^{-1}\}} S_i\Delta \\ &= \sum_i \min[(\lambda\Delta)^{-1}, S_i\Delta] \end{aligned}$$

and

$$R_\lambda = \sum_i \max[0, \Delta \log(S_i\Delta/(\lambda\Delta)^{-1})]$$

If we use

$$\gamma = (\lambda\Delta^2)^{-1}$$

as our parameter instead of λ and then let $\Delta \to 0$, a parametric expression for the rate-distortion function emerges. Casting the result in terms of the two-sided spectral density $\Phi(f)$, an even function of frequency satisfying $\Phi(f) = S(f)/2, f \ge 0$ and replacing the parameter γ by $\theta = \gamma/2$, we obtain

$$D_\theta = \int_{-\infty}^{\infty} \min[\theta, \Phi(f)]df \tag{8}$$

$$R_\theta = \int_{-\infty}^{\infty} \max\left[0, \frac{1}{2}\log(\Phi(f)/\theta)\right]df \tag{9}$$

[Some practitioners prefer to use angular frequency $\omega = 2\pi f$ as the argument of $\Phi(\cdot)$; of course, df then gets replaced in Eqs. (8) and (9) by $d\omega/(2\pi)$.]

The parametric representation (8) of the MSE rate-distortion function of a stationary Gaussian source is the source coding analog of the Shannon-Pinsker "water pouring" result for the capacity of an input-power-limited channel with additive stationary Gaussian noise. In a rate-distortion function of a time-discrete Gaussian sequence provided we limit the range of integration to $|f| \le 1/2$ or to $|\omega| \le \pi$. In such cases $\Phi(\omega)$ is the discrete-time power spectral density, a periodic function defined by

$$\Phi(\omega) = \sum_{k=-\infty}^{\infty} \phi(k)\exp(j\omega k)$$

where $\phi(k) = EX_jX_{j\pm k}$ is the correlation function of the source data. Note that when the parameter θ assumes a value less than the minimum[2] of $\Phi(\cdot)$, which minimum we shall denote by D^*, (8a) reduces to $D_\theta = \theta$, which eliminates the parameter and yields the explicit expression

$$R(D) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log[\Phi(\omega)/D]d\omega,\ D \le D^*$$

This may be recast in the form

$$R(D) = \tfrac{1}{2}\log(Q_0/D),\ D \le D^*$$

[2] More precisely, less than the essential infimum.

where

$$Q_0 = \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \Phi(\omega) d\omega\right]$$

is known both as the entropy rate power of $\{X_k\}$ and as its optimum one-step prediction error. In 1959 Shannon delivered a paper at the IRE Convention in New York City entitled "Coding Theorems for a Discrete Source with a Fidelity Criterion" [15]. This paper not only introduced the term "rate-distortion function" but also put lossy source coding on a firmer mathematical footing. Major contributions of the paper are:

- Definition and properties of the rate-distortion function.
- Calculating and bounding of R(D).
- Coding theorems.
- Insights into source-channel duality.

A discrete information source is a random sequence $\{X_k\}$. Each X_k assumes values in a discrete set $\mathcal{A}$ called the source alphabet. The elements of $\mathcal{A}$ are called the letters of the alphabet. We shall assume until further notice that there are finitely many distinct letters, say M of them, and shall write $\mathcal{A} = \{a(0), a(1), \ldots, a(M-1)\}$. Often we let $a(j) = j$ and hence $\mathcal{A} = \{0, 1, \ldots, M-1\}$; the binary case $\mathcal{A} = \{0, 1\}$ is particularly important.

The simplest case, to which we shall restrict attention for now, is that in which:

1. The X_k are independent and identically distributed (i.i.d.) with distribution $\{p(a), a \in \mathcal{A}\}$.
2. The distortion that results when the source produces the n-vector of letters $a = (a_1, \ldots, a_n) \in \mathcal{A}^n$ and the communication system delivers the n-vector of letters $b = (b_1, \ldots, b_n) \in \mathcal{B}^n$ to the destination as its representation of a is

$$d_n(a, b) = n^{-1} \sum_{k=1}^{n} d(a_k, b_k) \tag{10}$$

Here, $d(\cdot, \cdot) : \mathcal{A} \times \mathcal{B} \to [0, \infty)$ is called a single-letter distortion measure. The alphabet $\mathcal{B}$—variously called the reproduction alphabet, the user alphabet, and the destination alphabet—may be but need not be the same as $\mathcal{A}$. We shall write $\mathcal{B} = \{b(0), b(1), \ldots, b(N-1)\}$, where $N < M, N = M$ and $N > M$ all are cases of interest. When Eq. (9) applies, we say we have a *single-letter fidelity criterion* derived from $d(\cdot, \cdot)$.

Shannon defined the rate-distortion function $R(\cdot)$ as follows. First, let $Q = \{Q(b \mid a), a \in \mathcal{A}, b \in \mathcal{B}\}$ be a conditional probability distribution over the letters of the reproduction alphabet given a letter in the source alphabet.[3] Given a source distribution $\{p(j)\}$, we associate with any such Q two nonnegative quantities $d(Q)$ and $I(Q)$ defined by

$$d(Q) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) Q(b|a) d(a, b)$$

and

$$I(Q) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a) Q(b|a) \log\left(\frac{Q(b|a)}{q(b)}\right)$$

where

$$q(b) = \sum_{a \in \mathcal{A}} p(a) Q(b|a)$$

The quantities $d(Q)$ and $I(Q)$ are, respectively, the average distortion and the average Shannon mutual information associated with Q.

The rate-distortion function of the i.i.d. source $\{X_k\}$ with letter distribution $\{p(a) = P[X_k = a]\}$ with respect to the single-letter fidelity criterion generated by $d(\cdot, \cdot)$ is defined by the following minimization problem:

$$R(D) = \min_{Q: d(Q) \le D} I(Q) \tag{11}$$

Since the generally accepted object of communication is to maximize mutual information, not to minimize it, many people find the definition of the rate-distortion function counterintuitive. In this regard it often helps to interchange the independent and dependent variables, thus ending up with a distortion-rate function defined by

$$D(R) = \min_{Q: I(Q) \le R} d(Q) \tag{12}$$

Everyone considers that minimizing average distortion is desirable, so no one objects to this definition. Precisely the same curve results in the (D, R)-plane, except that now R is the independent variable instead of D. Distortion-rate functions are more convenient for certain purposes, and rate-distortion functions are more convenient for others. One should become comfortable with both.

Properties of the rate-distortion function include:

(a) $R(D)$ is well defined for all $D \ge D_{\min}$, where

$$D_{\min} = \sum_{a \in \mathcal{A}} p(a) \min_{b \in \mathcal{B}} d(a, b)$$

The distortion measure can be modified to assure that $D_{\min} = 0$. This is done via the replacement $d(a, b) \leftarrow d(a, b) - \min_b d(a, b)$, whereupon the whole rate-distortion curve simply translates leftward on the D-axis by $D_{\min}$.)

(b) $R(D) = 0$ for $D \ge D_{\max}$, where

$$D_{\max} = \min_b \sum_a p(a) d(a, b)$$

$D_{\max}$ is the maximum value of D that is of interest, since $R(D) = 0$ for all larger D. It is the value of D associated with the best guess at $\{X_k\}$ in the absence of any information about it other

[3] Such a Q often is referred to as a test channel. However, it is preferable to call it a test system because it functions to describe a probabilistic transformation from the source all the way to the user—not just across the channel. Indeed, the rate-distortion function has nothing to do with any channel per se. It is a descriptor of the combination of an information source and a user's way of measuring the distortion of approximations to that source.

than *a priori* statistical knowledge. For example, $D_{\max} = 1 - \max_a p(a)$ when $\mathcal{A} = \mathcal{B}$ and $d(a, b) = 1$ if $b \neq a$ and 0 if $b = a$.

(c) $R(D)$ is nonincreasing in D and is strictly decreasing at every $D \in (D_{\min}, D_{\max})$.

(d) $R(D)$ is convex downward. It is strictly convex in the range $(D_{\min}, D_{\max})$ provided $N \leq M$, where $N = |\mathcal{B}|$ and $M = |\mathcal{A}|$. In addition to the ever-present straight-line segment $R(D) = 0, D \geq D_{\max}$, if $N > M$ then $R(D)$ can possess one or more straight line segments in the range $D_{\min} < D < D_{\max}$.

(e) The slope of $R(D)$ is continuous in $(D_{\min}, D_{\max})$ and tends to $-\infty$ as $D \downarrow D_{\min}$. If there are straight-line segments in $(D_{\min}, D_{\max})$ (see (d) above), no two of them share a common endpoint.

(f) $R(D_{\min}) \leq H$, where

$$H = -\sum_{a \in \mathcal{A}} p(a) \log p(a)$$

is the source entropy. If for each $a \in \mathcal{A}$ there is a unique $b \in \mathcal{B}$ that minimizes $d(a, b)$, and each $b \in \mathcal{B}$ minimizes $d(a, b)$ for at most one $a \in \mathcal{A}$, then $R(D_{\min}) = H$.

Some of these properties were established by Shannon [15], including the essential convexity property (d). For proofs of the others see Gallager [3], Jelinek [16], and Berger [17].

In the special case of a binary equiprobable source and an error-frequency (or Hamming) distortion measure, calculations reveal that

$$R(D) = 1 - h(D) = 1 + D \log_2 D + (1 - D) \log_2 (1 - D), \ 0 \leq D \leq 1/2 = D_{\max}$$

where $h(\cdot)$ is Shannon's binary entropy function,

$$h(x) = -x \log_2 x - (1 - x) \log_2 (1 - x)$$

The desired end-to-end system behavior then becomes that of a binary symmetric channel (BSC) with crossover probability D. It follows that, if one seeks to send a Bernoulli(1/2) source over a BSC that is available once per source letter, then optimum performance with respect to the single-letter fidelity criterion generated by $d(a, b) = 1 - \delta_{a,b}$ can be obtained simply by connecting the source directly to the BSC and using the raw BSC output as the system output. There is need to do any source and/or channel coding. The average distortion will be $D = \varepsilon$, where ε is the crossover probability of the BSC.

This is another instance of a double "coincidence." This time the first coincidence is that a Bernoulli(1/2) source drives every BSC at capacity, and the second coincidence is that BSC(ε) provides precisely the end-to-end system transition probabilities that solve the rate-distortion problem for the Bernoulli(1/2) source at $D = \varepsilon$. Again, their combination represents a precarious saddle point. If the channel were not available precisely once per source symbol, if the Bernoulli source were to have a bias $p \neq 1/2$ if the channel were not perfectly symmetric, or if the distortion measure were not perfectly symmetric (i.e., if $d(0, 1) \neq d(1, 0)$), it would become necessary to employ source and channel codes of long memory and high complexity in order to closely approach performance that is ideal in the sense of achieving equality in the information transmission inequality (5). To enhance appreciation for the fragility of the double-coincidence saddle point, let us replace the Bernoulli(1/2) source with a Bernoulli(p) source, $p \neq \frac{1}{2}$. Calculations (see [17], p. 46–47) reveal that the rate-distortion function then becomes

$$R(D) = h(p) - h(D), \ 0 \leq D \leq \min(p, 1 - p) = D_{\max}$$

Although the optimum backward system transition probabilities $P(a|b)$ remain those of BSC(D), the optimum forward transition probabilities become those of a binary asymmetric channel. Hence, it is no longer possible to obtain an optimum system simply by connecting the source directly to the BSC and using the raw channel output as the system's reconstruction of the source. Not only does the asymmetric source fail to drive the BSC at capacity, but the BSC fails to provide the asymmetric system transition probabilities required in the $R(D)$ problem for $p \neq 1/2$. For example, suppose $p = 0.25$ so that $R(D) = 0.811 - h(D)$ bits/letter, $0 \leq D \leq 0.25 = D_{\max}$. Further suppose that $\varepsilon = 0.15$ so that the channel capacity is $C = 1 - h(0.15) = 0.390$ bits/channel use. Direct connection of the source to the channel yields an error frequency of $D = \varepsilon = 0.15$. However, evaluating the distortion-rate function at C in accordance with Eq. (5) shows that a substantially smaller error frequency of $R^{-1}(0.390) = 0.0855$ can be achieved using optimum source and channel coding. It is noteworthy that, even when treating continuous-amplitude sources and reconstructions, Shannon always employed discrete output random variables. "Consider a finite selection of points $z_i (i = 1, 2, \ldots, l)$ from the B space, and a measurable assignment of transition probabilities $q(z_i|m)$"[15]. One reason for why he did this is the he appreciated that the representation of the source would always have to be stored digitally; indeed, his major motivation for Section V in 1948 had been to overcome the challenge posed by the fact that continuous-amplitude data has infinite entropy. But, an even better explanation is that it turns out that the output random variable $\hat{X}$ that results from solving the rate-distortion problem for a continuous-amplitude source usually is discrete! In retrospect, it seems likely that Shannon knew this all along. (For more about this discreteness, see the excellent article by Rose [18] and also work of Fix [19] dealing with cases in which X has finite support.) Shannon did not state or prove any lossy source coding theorems in his classic 1948 paper. He did, however, state and sketch the proof of an end-to-end information transmission theorem for a communication system, namely his Theorem 21. Since the notation $R(D)$ did not exist in 1948, Shannon's theorem statement has v_1 in place of D and R_1 in place of $R(D)$. It reads:

Theorem 21. If a source has a rate R_1 for a valuation v_1 it is possible to encode the output of the source and

transmit it over a channel of capacity C with fidelity as near v_1 as desired provided $R_1 \leq C$. This is not possible if $R_1 > C$.

In 1959, however, Shannon proved many lossy source coding theorems, information transmission theorems, and their converses in the quite general case of stationary and ergodic sources. These theorems have since been considerably generalized by various authors; see, for example, the work of Gray and Davisson [20], Bucklew [21], and Kieffer [22]. It is not our purpose here to enter into the details of proofs of source coding theorems and information transmission theorems. Suffice it to say that at the heart of most proofs of positive theorems lies a random code selection argument, Shannon's hallmark. In the case of sources with memory, the achievability of average distortion D at coding rate $R_n(D)$ is established by choosing long-code words constructed of concatenations of "super-letters" from $\mathcal{B}^n$. Each super-letter is chosen independently of all the others in its own code word and in the other code words according to the output marginal $q(b)$ of the joint distribution $p(a)Q(b|a)$ associated with the solution of the variational problem that defines $R_n(D)$. Shannon concluded his 1959 paper on rate-distortion theory with some memorable, provocative remarks on the duality of source theory and channel theory. He mentions that, if costs are assigned to the use of its input letters of a channel, then determining its capacity subject to a bound on expected transmission cost amounts to maximizing a mutual information subject to a linear inequality constraint and results in a capacity-cost function for the channel that is concave downward. He says, "Solving this problem corresponds, in a sense, to finding a source that is just right for the channel and the desired cost." He then recapitulates that finding a source's rate-distortion function is tantamount to minimizing a mutual information subject to a linear inequality constraint and results in a function that is convex downward. "Solving this problem," Shannon says, "corresponds to finding a channel that is just right for the source and allowed distortion level." He concludes this landmark paper with the following two provocative sentences:

> This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.

Gallager [3] introduced the following dual to the convex mathematical programming problem that defines $R(D)$: Let λ denote a vector with components $\lambda(a)$ indexed by the letters of the source alphabet, $\mathcal{A}$. Given any real s and any $\lambda 0$ let c denote the vector with components $c(b)$, $b \in \mathcal{B}$ defined by

$$c(b) = \sum_{a \in \mathcal{A}} \lambda(a) p(a) \exp[sd(a,b)]$$

Let

$$\Lambda_s = \{\lambda \geq 0 : c \leq 1\}$$

Gallager proved that

$$R(D) = \max_{s \leq 0, \lambda \in \Lambda_s} \left[sD + \sum_{a \in \mathcal{A}} p(a) \log \lambda(a) \right]$$

Expressing $R(D)$ as a maximum rather than a minimum allows one to generate lower bounds to $R(D)$ readily. Just pick any $s \leq 0$ and any $\lambda \geq 0$. Then evaluate c. If the largest component of c exceeds 1, form a new λ by dividing the original λ by this largest $c(b)$. The new λ then belongs to Λ_s. It follows that the straight line $sD + \Sigma_a p(a) \log \lambda(a)$ in the (D, R)-plane underbounds $R(D)$. Not only are lower bounds to $R(D)$ produced aplenty this way, but we are assured that the upper envelope of all these lines actually *is* $R(D)$. This dual formulation is inspired by and capitalizes on the fact that a convex downward curve always equals the upper envelope of the family of all its tangent lines. It turns out that all known interesting families of lower bounds to $R(D)$ are special cases of this result. In particular, choosing the components of λ such that $\lambda(a)p(a)$ is constant yields Shannon's lower bound [15] for cases in which the distortion measure is balanced (i.e., every row of the distortion matrix is a permutation of the first row and every column is a permutation of the first column) and yields a generalization of the Shannon lower bound when the distortion measure is not balanced. The families of lower bounds introduced by Wyner and Ziv also can be shown to be obtainable via dual problem investigations.

Although it may have appeared in the early 1970s that the then 25-year-old subject of rate-distortion theory was reaching maturity, this has not turned out to be the case. Rate-distortion theory thrived at Stanford under Gray, at Cornell under Berger, who wrote a text devoted entirely to the subject [17], at JPL under Posner, at UCLA under Omura and Yao, and at Bell Labs under Wyner.[4]

Attending a seminar on the mathematics of population genetics and epidemiology somehow inspired Blahut to work on finding a fast numerical algorithm for the computation of rate-distortion functions. He soon thereafter found that the point on an $R(D)$ curve parameterized by s could be determined by the following iterative procedure [23]:[5]

Step 0. Set $r = 0$. Choose any probability distribution $q_0(\cdot)$ over the destination alphabet that has only positive components, for example, the uniform distribution $q_0(b) = 1/|\mathcal{B}|$.

Step 1. Compute $\lambda_r(a) = (\Sigma_b q_r(b) \exp[sd(a,b)])^{-1}$, $a \in \mathcal{A}$.

[4] Centers of excellence in rate-distortion emerged in Budapest under Csiszar, in Tokyo under Amari, in Osaka under Arimoto, in Israel under Ziv and his "descendants," in Illinois under Pursley and at Princeton under Verdu.

[5] Blahut and, independently, Arimoto [24] found an analogous algorithm for computing the capacity of channels. Related algorithms have since been developed for computing other quantities of information-theoretic interest. For a treatment of the general theory of such max-max and min-min alternating optimization algorithms, see Csiszar and Tusnady [25].

Step 2. Compute $c_r(b) = \Sigma_a \lambda_r(a) p(a) \exp[sd(a,b)]$, $b \in \mathcal{B}$. If $\max_b c_r(b) < 1 + \varepsilon$, halt.

Step 3. Compute $q_{r+1}(b) = c_r(b) q_r(b)$. $r \leftarrow r+1$. Return to Step 1.

Blahut proved the following facts.

1. The algorithm terminates for any rate-distortion problem for any $\varepsilon > 0$.
2. At termination, the distance from the point (D_r, I_r) defined by

$$D_r = \sum_{a,b} p(a)\lambda_r(a) q_r(b) \exp[sd(a,b)] d(a,b)$$

and

$$I_r = sD_r + \sum_a p(a) \log \lambda_r(a)$$

to the point D, $R(D)$ parameterized by s (i.e., the point on the $R(D)$-curve at which $R'(D) = s$) goes to zero as $\varepsilon \to 0$. Moreover, Blahut provided upper and lower bounds on the terminal value of $I_r - R(D_r)$ that vanish with ε.

Perhaps the most astonishing thing about Blahut's algorithm is that it does not explicitly compute the gradient of $R + sD$ during the iterations, nor does it compute the average distortion and average mutual information until after termination. In practice, the iterations proceed rapidly even for large alphabets. Convergence is quick initially but slows for large r; Newton-Raphson methods could be used to close the final gap faster, but practitioners usually have not found this to be necessary. The Blahut algorithm can be used to find points on rate-distortion functions of continuous-amplitude sources, too; one needs to use fine-grained discrete approximations to the source and user alphabets. See, however, the so-called "mapping method" recently introduced by Rose [18], which offers certain advantages especially in cases involving continuous alphabets; Rose uses reasoning from statistical mechanics to capitalize on the fact, alluded to earlier, that the support of the optimum distribution over the reproduction alphabet usually is finite even when $\mathcal{B}$ is continuous. Following his seminal work on autoregressive sources and certain generalizations thereof, Gray joined the Stanford faculty. Since rate-distortion is a generalization of the concept of entropy and conditional entropy plays many important roles, Gray sensed the likely fundamentality of a theory of conditional rate-distortion functions and proceeded to develop it [26] in conjunction with his student, Leiner [27,28]. He defined

$$R_{X|Y}(D) = \min I(X; \hat{X}|Y)$$

where the minimum is over all r.v., $\hat{X}$ jointly distributed with (X, Y) in such a manner that $E_{X,Y,\hat{X}} d(X, \hat{X}) \le D$. This not only proved of use per se but also led to new bounding results for classical rate-distortion functions. However, it did not treat what later turned out to be the more challenging problem of how to handle side-information $\{Y_k\}$ that was available to the decoder only and not to the encoder. That had to await groundbreaking research by Wyner and Ziv [29].

Gray also began interactions with the mathematicians Ornstein and Shields during this period. The fruits of those collaborations matured some years later, culminating in a theory of sliding block codes for sources and channels that finally tied information theory and ergodic theory together in mutually beneficial and enlightening ways. Other collaborators of Gray in those efforts included Neuhoff, Omura, and Dobrushin [30–32]. The so-called process definition of the rate-distortion function was introduced and related to the performance achievable with sliding block codes with infinite window width (codes in the sense of ergodic theory). It was shown that the process definition agreed with Shannon's 1959 definition of the rate-distortion function $\liminf_{n\to\infty} R_n(D)$ for sources and/or distortion measures with memory. More importantly, it was proved that one could "back off" the window width from infinity to a large, finite value with only a negligible degradation in the tradeoff of coding rate versus distortion, thereby making the theory of sliding block codes practically significant.

Seeing that Slepian and Wolf [33] had conducted seminal research on lossless multiterminal source coding problems analogous to the multiple access channel models of Ahlswede [34] and Liao [35], Berger and Wyner agreed that research should be done on a lossy source coding analog of the novel Cover-Bergmans [36,37] theory of broadcast channels. Gray and Wyner were the first to collaborate successfully on such an endeavor, authoring what proved to the first of many papers in the burgeoning subject of multiterminal lossy source coding [38]. The seminal piece of research in multiterminal lossy source coding was the paper by Wyner and Ziv [29], who considered lossy source coding with side information at the decoder. Suppose that in addition to the source $\{X_k\}$ that we seek to convey to the user, there is a statistically related source $\{Y_k\}$. If $\{Y_k\}$ can be observed both by the encoder and the decoder, then we get conditional rate-distortion theory a la Gray. The case in which neither the encoder nor the decoder sees $\{Y_k\}$, which perhaps is under the control of an adversary, corresponds to Berger's source coding game [39]. The case in which the encoder sees $\{Y_k\}$ but the decoder does was long known [40] to be no different from the case in which there is no $\{Y_k\}$. But the case in which the decoder is privy to $\{Y_k\}$ but the encoder is not proves to be both challenging and fascinating. For the case of a single-letter fidelity criterion and (X_k, Y_k)-pairs that are i.i.d. over the index k, Wyner and Ziv showed that the rate-distortion function, now widely denoted by $_{WZ}(D)$ in their honor, is given by

$$R_{WZ}(D) = \min_{Z \in cal Z_D} I(X; Z \mid Y) \tag{13}$$

where $\mathcal{Z}_D$ is the set of auxiliary r.v. $Z \in \mathcal{Z}$ jointly distributed with a generic (X, Y) such that:

1. $Y - X - Z$ is a Markov chain; i.e., $p_{Y,X,Z}(y,x,z) = p_Y(y) p_{X|Y}(x|y) p_{Z|X}(z|x)$.
2. There exists $g : \mathcal{Z} \times \mathcal{Y} \to \mathcal{X}$ such that $Ed(X, g(Z, Y)) \le D$.

3. The cardinality of the alphabet $\mathcal{Z}$ may be constrained to satisfy $|\mathcal{Z}| \leq |\chi| + 1$.

In the special case in which $\{X_k\}$ and $\{Y_k\}$ are Bernoulli(1/2) and statistically related as if connected by a BSC of crossover probability $p \leq 1/2$ and $d(a, b) = 1 - \delta_{a,b}$,

$$R_{WZ}(D) = \begin{cases} h(p) - h(p * D), & \text{if } 0 \leq D \leq D_{ci} \\ \text{straight line from } (D_c, h(p) - h(p * D_c) \text{ to } (p, 0) & \text{if } D_c \leq D \leq p \end{cases} \tag{14}$$

where $p * d = p(1 - D) + (1 - p)D$ and D_c is such that the straight-line segment for $D \geq D_c$ is tangent to the curved segment for $D \leq D_c$. Berger had used Bergmans [37] theory of "satellites and clouds" to show that Eq. (17) was a upper bound to $R(D)$ for this binary symmetric case. The major contribution of Wyner and Ziv's paper resided in proving a converse to the unlikely effect that this performance cannot be improved upon, and then generalizing to Eq. (17) for arbitrary (X, Y) and $d(\cdot, \cdot)$.

The advent of Wyner-Ziv theory gave rise to a spate of papers on multiterminal lossy source coding, codified, and summarized by Berger in 1977 [41]. Contributions described therein include works by Korner and Marton, [42–44], Berger and Tung [45,46], Chang [47], Shohara [48], Omura and Housewright [49], Wolfowitz [50], and Sgarro [51]. In succeeding decades further strides have been made on various side-information lossy coding problems [52–58]. Furthermore, challenging new multiterminal rate-distortion problems have been tackled with considerable success, including the multiple descriptions problem [59–68], the successive refinements problem [69], and the *CEO problem* [70–72]. Applications of multiple descriptions to image, voice, audio and video coding are currently in development, and practical schemes based on successive refinement theory are emerging that promise application to progressive transmission of images and other media. In order for rate-distortion theory to be applied to images, video, and other multidimensional media, it is necessary to extend it from random processes to random fields (i.e., collections of random variables indexed by multidimensional parameters or, more generally, by the nodes of a graph). The work of Hayes, Habibi, and Wintz [73] extending the water-table result for Gaussian sources to Gaussian random fields already has been mentioned. A general theory of the information theory of random fields has been propounded [74], but we are more interested in results specific to rate-distortion. Most of these have been concerned with extending the existence of critical distortion to the random field case and then bounding the critical distortion for specific models. The paper of Hajek and Berger [75] founded this subfield. Work inspired thereby included Bassalygo and Dobrushin [76], Newman [77], Newman and Baker[78] in which the critical distortion of the classic Ising model is computed exactly, and several papers by Berger and Ye [79,80]. For a summary and expansion of all work in this arena, see the monograph by Ye and Berger [81].

Work by Fitingof, Lynch, Davisson, and Ziv in the early 1970s showed that lossless coding could be done efficiently without prior knowledge of the statistics of the source being compressed, so-called universal lossless coding. This was followed by development of Lempel-Ziv coding [82,83], arithmetic coding [84–86] and context-tree weighted encoding [87,88], which have made universal lossless coding practical and, indeed, of great commercial value.

Universal lossy coding has proven more elusive as regards both theory and practice. General theories of universal lossy coding based on ensembles of block codes and tree codes were developed [89–95], but these lack sufficient structure and hence require encoder complexity too demanding to be considered as solving the problem in any practical sense. Recent developments are more attractive algorithmically [96–103]. The paper by Yang and Kieffer [100] is particularly intriguing; they show that a lossy source code exists that is universal not only with respect to the source statistics but also with respect to the distortion measure. Though Yang–Kieffer codes can be selected a priori in the absence of any knowledge about the fidelity criterion, the way one actually does the encoding does, of course, depend on which fidelity criterion is appropriate to the situation at hand. All universal lossy coding schemes found to date lack the relative simplicity that imbues Lempel–Ziv coders and arithmetic coders with economic viability. Perhaps as a consequence of the fact that approximate matches abound whereas exact matches are unique, it is inherently much faster to look for an exact match than it is to search a plethora of approximate matches looking for the best, or even nearly the best, among them. The right way to tradeoff search effort in a poorly understood environment against the degree to which the product of the search possesses desired criteria has long been a human enigma. This suggests it is unlikely that the "holy grail" of implementable universal lossy source coding will be discovered soon.

We have several times noted how sources can be matched to given channels and channels can be matched to given sources. Indeed, we even quoted Shannon's 1959 comments about this. Doubly matched situations that require no coding in order to for optimum performance to be achieved have been stressed. Whereas these situations are rarely encountered in the context of man-made communication systems, there is growing evidence that *doubly matched* configurations are more the norm that the exception in sensory information processing by living organisms. A growing community of biologists and information theorists are engaged in active collaboration on mathematical and experiments treatments of information handling within living systems. These bioinformation theorists are finding that chemical pathways and neural networks within organisms not only exhibit double matching but do so robustly over a range of data rates and energy consumption levels. The evolutionist's explanation for how this comes to pass is that, over eons, natural selection evolves channels within successful organisms that are matched to the information sources that constitute the environment. These channels extract just enough information to provide a representation of each facet of the environment that is sufficiently accurate for the organism's purposes. That is, lossy coding that is nearly optimal in the sense of rate-distortion theory is routinely performed by living

organisms. Moreover, these channels evolve so as to reside at saddle points which correspond not only to minimizing information flow subject to fidelity constraints but also simultaneously to maximizing the mutual information between their inputs and outputs subject to constraints on rates at which they consume bodily resources, especially metabolic energy. Further information about lossy coding in living systems can be found in the works of Levy and Baxter [104,105] and especially in the Shannon Lecture recently delivered by Berger [106].

BIOGRAPHY

Toby Berger received the B.E. degree in electrical engineering from Yale in 1962 and the M.S. and Ph.D. degrees in applied mathematics from Harvard in 1964 and 1966. From 1962 to 1968, he was a senior scientist at Raytheon Company. In 1968, he joined the faculty of Cornell University where he is presently the Irwin and Joan Jacobs professor of engineering. His research interests include information theory, random fields, communication networks, wireless communications, video compression, voice and signature compression and verification, biological information theory, quantum information theory, and coherent signal processing. Berger has been a Guggenheim Fellow, a Fulbright Fellow, a Fellow of the Japan Association for Advancement of Science, and a Fellow of the Ministry of Education of the PRC. An IEEE Fellow, he has served as editor-in-chief of the IEEE Transactions on Information Theory and as president of the IEEE Information Theory Society. Berger received the 1982 Frederick E. Terman Award from the American Society for Engineering Education and the 2002 Shannon Award from the IEEE Information Theory Society.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**: 379–423, 623–656, (July and Oct. 1948). (Also in N. J. A. Sloane and A. D. Wyner, eds., *Claude Elwood Shannon: Collected Papers*, IEEE Press, Piscataway, NJ, 1993, 5–83.)
2. T. Berger and J. D. Gibson, Lossy Source Coding, Invited paper for Special 50th Anniversary Issue, *IEEE Trans. Inform. Theory* **44**(6): 2693–2723 (Oct. 1998). (Reprinted in *Information Theory: 50 Years of Discovery* IEEE Press, Piscataway, NJ 1999.)
3. R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.
4. J. L. Holsinger, Digital Communication over Fixed Time-Continuous Channels with Memory — with Special Application to Telephone Channels. Sc. D. dissertation, M.I.T., Cambridge, MA (TR No. 366, Lincoln Labs, Lexington, MA), 1968.
5. A. N. Kolmogorov, A new metric invariant of transitive dynamic systems and automorphisms in Lebesgue spaces, *Dokl. Akad. Nauk. SSSR* **119**: 861–864 (1958).
6. A. N. Kolmogorov, The theory of transmission of information, Plenary Session of the Academy of Sciences of the USSR on the Automization of Production, Moscow, 1956. *Iz. Akad. Nauk SSSR* 66–99 (1957).
7. Ya. G. Sinai, On the concept of entropy of a dynamical system, *Dokl. Akad. Nauk. SSSR* **124**: 768–771 (1959).
8. A. N. Kolmogorov, On the Shannon theory of information transmission in the case of continuous signals, *IRE Transactions on Information Theory* **IT-2**: 102–108 (1956).
9. D. S. Ornstein, Bernoulli shifts with the same entropy are isomorphic, *Advances in Math.* **4**: 337–352 (1970).
10. E. C. Posner and E. R. Rodemich, Epsilon entropy and data compression, *The Annals of Math. Statist.* **42**: 2079–2125 (1971).
11. R. J. McEliece and E. C. Posner, Hiding and covering in a compact metric space, *Annals of Statistics* **1**: 729–739 (1973).
12. C. E. Shannon, Communication in the presence of noise, *Proc. IRE* **37**: 10–21 (1949).
13. M. S. Pinsker, Mutual information between a pair of stationary Gaussian random processes, *Dokl. Akad. Nauk. USSR* **99**(2): 213–216 (1954).
14. M. S. Pinsker, Computation of the message rate of a stationary random process and the capacity of a stationary channel, *Dokl. Akad. Nauk. USSR* **111**(4): 753–756 (1956).
15. C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion. *IRE Convention Record* **7**: 142–163 (1959). (Also in R. E. Machol, ed., *Information and Decision Processes*, McGraw-Hill, Inc. New York, 1960, 93–126, and in N. J. A. Sloane and A. D. Wyner, eds., *Claude Elwood Shannon:Collected Papers*, IEEE Press, Piscataway, NJ, 1993, 325–350.)
16. F. Jelinek, *Probabilistic Information Theory*, McGraw-Hill, New York, 1968.
17. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
18. K. Rose, A mapping approach to rate-distortion computation and analysis, *IEEE Trans. Inform. Theory* **IT-42**: 1939–1952 (1996).
19. S. L. Fix, Rate distortion functions for squared error distortion measures, In *Proc. 16th Annual Allerton Conference on Comm., Contr. and Comput.*, Monticello, IL, (1978).
20. R. M. Gray and L. D. Davisson, Source coding theorems without the ergodic assumption, *IEEE Trans. Inform. Theory* **IT-20**: 625–636 (1974).
21. J. A. Bucklew, The source coding theorem via Sanov's theorem, *IEEE Trans. Inform. Theory* **IT-33**: 907–909 (1987).
22. J. C. Kieffer, A survey of the theory of source coding with a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-39**: 1473–1490 (1993).
23. R. E. Blahut, Computation of channel capacity and rate distortion functions, *IEEE Trans. Inform. Theory* **IT-18**: 460–473 (1972).
24. S. Arimoto, An algorithm for calculating the capacity of an arbitrary discrete memoryless channel, *IEEE Trans. Inform. Theory* **IT-18**: 14–20 (1972).
25. I. Csiszar and G. Tusnady, Information Geometry and Alternating Minimization Procedures, in *Statistics and Decisions / Supplement Issue*, R. Oldenbourg Verlag, Munich, Germany, E. J. Dudewicz, D. Plachky and P. K. Sen, Eds., No. 1, 205–237, 1984. (Formerly entitled On Alternating Minimization Procedures, Preprint of the Mathematical

Institute of the Hungarian Academy of Sciences, No. 35/1981, 1981.)

26. R. M. Gray, A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions, *IEEE Trans. Inform. Theory* **IT-19**: 480–489 (1973).
27. B. M. Leiner, *Rate-Distortion Theory for Sources with Side Information*, Ph.D. dissertation, Stanford University, Calif., August 1973.
28. B. M. Leiner and R. M. Gray, Rate-distortion for ergodic sources with side information, *IEEE Trans. Inform. Theory* **IT-20**: 672–675 (1974).
29. A. D. Wyner and J. Ziv, The rate-distortion function for source coding with side-information at the receiver, *IEEE Trans. Inform. Theory* **IT-22**: 1–11 (1976).
30. R. M. Gray, D. L. Neuhoff and J. K. Omura, Process definitions of distortion rate functions and source coding theorems, *IEEE Trans. Inform. Theory* **IT-21**: 524–532 (1975).
31. R. M. Gray, D. L. Neuhoff and D. S. Ornstein, Nonblock source coding with a fidelity criterion, *Annals of Probability* **3**: 478–491 (1975).
32. R. M. Gray, D. S. Ornstein and R. L. Dobrushin, Block synchronization, sliding-block coding, invulnerable sources and zero error codes for discrete noisy channels, *Annals of Probability* **8**: 639–674 (1975).
33. D. Slepian and J. K. Wolf, Noiseless coding of correlated information sources, *IEEE Trans. Inform. Theory* **IT-19**: 471–480 (1973).
34. R. Ahlswede, Multi-way Communication Channels, In *Proc. 2nd. Int. Symp. Information Theory* (*Tsahkadsor, Armenian SSR*), 23–52, 1971.
35. H. Liao, Multiple Access Channels, Ph.D. dissertation, University of Hawaii, Honolulu, HI, 1972.
36. T. M. Cover, Broadcast channels, *IEEE Trans. Inform. Theory* **IT-18**: 2–14 (1972).
37. P. Bergmans, Random coding theorem for broadcast channels with degraded components, *IEEE Trans. Inform. Theory* **IT-19**: 197–207 (1973).
38. R. M. Gray and A. D. Wyner, Source coding for a simple network, *Bell Syst. Tech. J.* **58**: 1681–1721 (1974).
39. T. Berger, The Source coding game, *IEEE Trans. Inform. Theory* **IT-17**: 71–76 (1971).
40. T. J. Goblick, Jr., Coding for a Discrete Information Source with a Distortion Measure, Ph.D. dissertation, M.I.T., Cambridge, MA, 1962.
41. T. Berger, Multiterminal Source Coding, In *The Information Theory Approach to Communications*, CISM Courses and Lectures No. 229, 171-231, Springer-Verlag, Wien–New York, 1977.
42. J. Korner and K. Marton, The Comparison of Two Noisy Channels, *Trans. Keszthely Colloq. Inform. Theory*, Hungarian National Academy of Sciences, August 8–12, Keszthely, Hungary, 411–423.
43. J. Korner and K. Marton, Images of a set via two channels and their role in multi-user communications, *IEEE Trans. Inform. Theory* **IT-23**: 751–761 (1977).
44. J. Korner and K. Marton, How to encode the modulo-two sum of binary sources, *IEEE Trans. Inform. Theory* **IT-25**: 219–221 (1979).
45. T. Berger and S. Y. Tung, Encoding of Correlated Analog Sources, *Proc. 1975 IEEE-USSR Joint Workshop on Information Theory*, IEEE Press, 7–10, December 1975.
46. S. Y. Tung, Multiterminal Rate-Distortion Theory, Ph.D. dissertation, Cornell University, Ithaca, NY, 1977.
47. M. U. Chang, Rate-Distortion with a Fully Informed Decoder and a Partially Informed Encoder, Ph.D. dissertation, Cornell University, Ithaca, NY, 1978.
48. A. Shohara, Source Coding Theorems for Information Networks, Ph.D. dissertation, University of California at Los Angeles, Tech. Rep. UCLA-ENG-7445, 1974.
49. J. K. Omura and K. B. Housewright, Source Coding Studies for Information Networks, *Proc. IEEE 1977 International Conference on Communications*, IEEE Press, 237–240, Chicago, Ill., June 13–15, 1977.
50. T. Berger et al., An upper bound on the rate-distortion function for source coding with partial side information at the decoder, *IEEE Trans. Inform. Theory* **IT-25**: 664–666 (1979).
51. A. Sgarro, Source coding with side information at several decoders, *IEEE Trans. Inform. Theory* **IT-23**: 179–182 (1977).
52. A. D. Wyner, The rate-distortion function for source coding with side information at the decoder — II: General sources, *Information and Control* **38**: 60–80 (1978).
53. H. Yamamoto, Source coding theory for cascade and branching communication systems, *IEEE Trans. Inform. Theory* **IT-27**: 299–308 (1981).
54. H. Yamamoto, Source coding theory for a triangular communication system, *IEEE Trans. Inform. Theory* **IT-42**: 848–853 (1996).
55. A. H. Kaspi and T. Berger, Rate-distortion for correlated sources with partially separated encoders, *IEEE Trans. Inform. Theory* **IT-28**: 828–840 (1982).
56. C. Heegard and T. Berger, Rate distortion when side information may be absent, *IEEE Trans. Inform. Theory* **IT-31**: 727–734 (1985).
57. T. Berger and R. W. Yueng, Multiterminal source encoding with one distortion criterion, *IEEE Trans. Inform. Theory* **IT-35**: 228–236 (March 1989).
58. T. Berger and R. W. Yueng, Multiterminal source encoding with encoder breakdown, *IEEE Trans. Inform. Theory* **IT-35**: 237–244 (1989).
59. A. Gersho and A. D. Wyner, The Multiple Descriptions Problem, Presented by A. D. Wyner, IEEE Information Theory Workshop, Seven Springs Conference Center, Mt. Kisco, NY, September 1979.
60. H. S. Witsenhausen, On source networks with minimal breakdown degradation, *Bell Syst. Tech. J* **59**: 1083–1087 (1980).
61. J. K. Wolf, A. D. Wyner and J. Ziv, Source coding for multiple descriptions, *Bell Syst. Tech.* **59**: 1417–1426 (1980).
62. L. H. Ozarow, On the source coding problem with two channels and three receivers, *Bell Syst. Tech. J* **59**: 1909–1922 (1980).
63. H. A. Witsenhausen and A. D. Wyner, Source coding for multiple descriptions II: A binary source, *Bell Syst. Tech. J.* **60**: 2281–2292 (1981).

64. A. El Gamal and T. M. Cover, Achievable rates for multiple descriptions, *IEEE Trans. Inform. Theory* **IT-28**: 851–857 (1982).

65. T. Berger and Z. Zhang, Minimum breakdown degradation in binary source encoding, *IEEE Trans. Inform. Theory* **IT-29**: 807–814 (1983).

66. R. Ahlswede, The rate-distortion region for multiple descriptions without excess rate, *IEEE Trans. Inform. Theory* **IT-31**: 721–726 (1985).

67. New results in binary multiple descriptions, *IEEE Trans. Inform. Theory* **IT-33**: 502–521 (1987).

68. Z. Zhang and T. Berger, Multiple description source coding with no excess marginal rate, *IEEE Trans. Inform. Theory* **IT-41**: 349–357 (1995).

69. W. E. Equitz and T. M. Cover, Successive refinement of information, *IEEE Trans. Inform. Theory* **IT-37**: 269–275 (1991). (See also W. E. Equitz and T. M. Cover, Addendum to Successive refinement of information, *IEEE Trans. Inform. Theory* **IT-39**: 1465–1466 (1993).

70. T. Berger, Z. Zhang and H. Viswanathan, The CEO problem, *IEEE Trans. Inform. Theory* **IT-42**: 887–903 (May 1996).

71. H. Viswanathan and T. Berger, The quadratic gaussian CEO problem, *IEEE Trans. Inform. Theory* **43**: 1549–1561 (1997).

72. Y. Oohama, The rate distortion problem for the quadratic gaussian CEO problem, *EEE Trans. Inform. Theory* (1998).

73. J. F. Hayes, A. Habibi, and P. A. Wintz, Rate-distortion function for a gaussian source model of images, *IEEE Trans. Inform. Theory* **IT-16**: 507–509 (1970).

74. T. Berger, S. Y. Shen, and Z. Ye, Some communication problems of random fields, *International Journal of Mathematical and Statistical Sciences* **1**: 47–77 (1992).

75. A decomposition theorem for binary markov random fields, *Annals of Probability* **15**: 1112–1125 (1987).

76. L. A. Bassalygo and R. L. Dobrushin, ε-Entropy of the random field, *Problemy Peredachi Informatsii* **23**: 3–15 (1987).

77. C. M. Newman, Decomposition of binary random fields and zeros of partition functions, *Annals of Probability* **15**: 1126–1130 (1978).

78. C. M. Newman and G. A. Baker, Decomposition of Ising Model and the Mayer Expansion, In S. Albeverio et al., eds., *Ideas and Methods in Mathematics and Physics — In Memory of Raphael Hoegh-Krohn* (*1938–1988*), Cambridge University Press, Cambridge, UK, 1991.

79. T. Berger and Z. Ye, ε-Entropy and critical distortion of random fields, *IEEE Trans. Inform. Theory* **IT-36**: 717–725 (1990).

80. Z. Ye and T. Berger, A new method to estimate the critical distortion of random fields, *IEEE Trans. Inform. Theory* **IT-38**: 152–157 (1992).

81. Z. Ye and T. Berger, *Information Measures for Discrete Random Fields*, Chinese Academy of Sciences, Beijing, 1998.

82. J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans. Inform. Theory* **IT-23**: 337–343 (1977).

83. J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Trans. Inform. Theory* **IT-24**: 337–343 (1978).

84. R. Pasco, Source Coding Algorithms for Fast Data Compression, Ph.D. dissertation, Stanford University, California, 1976.

85. J. Rissanen, Generalized kraft inequality and arithmetic coding, *IBM J. Res. Devel.* **20**: 198 (1976).

86. J. Rissanen, Universal coding, information, prediction and estimation, *IEEE Trans. Inform. Theory* **IT-30**: 629–636 (1984).

87. F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, The context-tree weighting method: Basic properties, *IEEE Trans. Inform. Theory* **IT-41**: 653–664 (1995).

88. F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, Context weighting for general finite-context sources, *IEEE Trans. Inform. Theory* **IT-42**: 1514–1520 (1996).

89. D. L. Neuhoff, R. M. Gray, and L. D. Davisson, Fixed rate universal block source coding with a fidelity criterion, *IEEE Trans. Inform. Theory* **21**: 511–523 (1975).

90. K. M. Mackenthum, Jr. and M. B. Pursley, Strongly and Weakly Universal Source Coding, In *Proc. 1977 Conference on Information Science and Systems*, 286–291, Johns Hopkins University, 1977.

91. M. B. Pursley and K. M. Mackenthum, Jr., Variable-rate source coding for classes of sources with generalized alphabets, *IEEE Trans. Inform. Theory* **IT-23**: 592–597 (1977).

92. K. M. Mackenthum, Jr. and M. B. Pursley, Variable-rate universal block source coding subject to a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-24**: 349–360 (1978).

93. H. H. Tan, Tree coding of discrete-time abstract alphabet stationary block-ergodic sources with a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-22**: 671–681 (1976).

94. J. Ziv, Coding of sources with unknown statistics — part II: Distortion relative to a fidelity criterion, *IEEE Trans. Inform. Theory* **IT-18**: 389–394 (1972).

95. T. Hashimoto, *Tree Coding of Sources and Channels*, Ph.D. dissertation, Osaka University, Japan, 1981.

96. Y. Steinberg and M. Gutman, An algorithm for source coding subject to a fidelity criterion based on string matching, *IEEE Trans. Inform. Theory* **IT-39**: 877–886 (1993).

97. Z. Zhang and V. K. Wei, An on-line universal lossy data compression algorithm by continuous codebook refinement, *IEEE Trans. Inform. Theory* **IT-42**: 803–821 (1996).

98. Z. Zhang and V. K. Wei, An on-line universal lossy data compression algorithm by continuous codebook refinement, part two: Optimality for ϕ-mixing source models, *IEEE Trans. Inform. Theory* **IT-42**: 822–836 (1996).

99. I. Sadeh, Universal Compression Algorithms Based on Approximate String Matching, *Proc. 1995 IEEE International Symposium on Information Theory*, Whistler, British Columbia, September 17–22, 1995, p. 84.

100. E. H. Yang and J. Kieffer, Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm, *IEEE Trans. Inform. Theory* **IT-42**: 239–245 (1996).

101. E. H. Yang, Z. Zhang, and T. Berger, Fixed-slope universal lossy data compression, *IEEE Trans. Inform. Theory* **IT-43**: 1465–1476 (1997).

102. I. Kontoyiannis, An implementable lossy version of the Lempel-Ziv algorithm — part I: Optimality for memoryless sources, NSF Technical Report no. 99, Department of Statistics, Stanford University, April 1998.

103. I. Kontoyiannis, Asymptotically Optimal Lossy Lempel-Ziv Coding. To be presented at 1998 IEEE International Symposium on Information Theory, MIT, Cambridge MA, August 16–21, 1998.

104. W. B. Levy and R. A. Baxter, Energy efficient neural codes, *Neural Computation* **8**(3): 531–543 (1996).

105. W. B. Levy and R. A. Baxter, Energy efficient neural computation via quantal synaptic failures, *J. Neurosci.* **22**: 4746–4755 (2002).

106. T. Berger, "Living Information Theory," The Shannon Lecture, IEEE International Symposium on Information Theory, Lausanne, Switzerland, July 4, 2002. (Visit *http://www.itsoc.org* to download the slides of this presentation.)

REFLECTOR ANTENNAS

CAREY RAPPAPORT
Northeastern University
Boston, Massachusetts

1. APERTURE ANTENNA FUNDAMENTALS

1.1. Introduction

A critical component in a high-performance wireless telecommunications system is the antenna, which couples the signals carried by wires into and from waves propagating through space. For point-to-point communications links, specific antennas are used to constrain radiated power in a prescribed manner. In satellite and terrestrial applications, the antenna concentrates waves in an angular region of high intensity in desired directions and minimizes the power that would be wasted elsewhere. This key measure of antenna performance, the gain, depends directly on the effective antenna aperture area. Common high-gain antennas with large effective apertures include lenses, reflectors, and phased arrays.

The first two devices increase the effective aperture area by a purely geometric transformation, whereas the phased array electrically transforms the aperture. The lens antenna is a heavy and often complex structure, which, like the phased array, is (in most cases) frequency-dependent. The reflector antenna is the simplest, cheapest, and lightest alternative and has been the primary means of providing high-gain microwave beams for over half a century. With each type of antenna, the input signal flowing on cables is distributed in a prescribed way across the outer radiating surface. The particular phase and amplitude distribution of field (or equivalently, current density) at this aperture governs the radiation pattern shape of the antenna.

1.2. Far-Field Radiation Concepts

Most telecommunications applications involve links with distant stations. When the distance r to these stations is great compared with the antenna size D, $r \gg 2D^2/\lambda$, the stations are in the far-field of the antenna. The wavelength is given by $\lambda = c/f$, for frequency f and speed of light c. In the farfield, or Fraunhofer region, the spatial field distribution pattern is independent of the radial distance from the source. Indeed the electromagnetic field falls off as $\exp(-jkr)/kr$, with wavenumber $k = 2\pi f/c$, while the radiation pattern is a function of polar angle from boresight (direction the antenna is aimed) θ, and circumferential angle φ. The radiation pattern is proportional to the two-dimensional spatial Fourier transform of the aperture current phase and amplitude distribution [1–9]. Thus, the spatial current distribution across a finite aperture $A(x, y)$ is mapped to the angular radiation pattern $P(k_x, k_y)$, with angular wavenumbers, $k_x = k\cos\theta\cos\varphi$, and $k_y = k\cos\theta\sin\varphi$. Clearly, a larger aperture (bigger antenna) produces a higher-intensity beam, with a narrower beamwidth. A rectangular aperture of dimension $2a$ by $2b$ — illuminated by a wave with uniform amplitude and constant phase — produces a beam proportional to $\sin(k_x a)/(k_x a) \bullet \sin(k_y b)/(k_y b)$, while a circular aperture of radius R produces a radiation pattern proportional to $J_1(u)/u$, where $u = k_\rho R$, $k_\rho = \sqrt{(k_x^2 + k_y^2)}$ and J_1 is the first-order Bessel function of first kind. This radiation pattern of radiated power as a function of angle is shown in Fig. 1 for a 20-wavelength-diameter aperture. Figure 2 displays the same pattern in 3D polar format. The antenna pattern is very sensitive to the current distribution at the aperture. Beam direction, maximum gain, power outside the mainlobe, and associated sidelobe levels depend strongly on the aperture phase function [10,11].

The gain of an antenna is the measure of field intensity in a particular direction relative the average intensity over all directions. The peak gain, or directivity, specifies the ratio of the power density at boresight in the middle of the antenna's main beam at a distance r to the average power density (total radiated power divided by $4\pi r^2$). An aperture antenna with area A has maximum peak gain when the current amplitude and phase are uniform across the aperture. Such an aperture is referred to as having 100% illumination efficiency. In this case the peak gain is

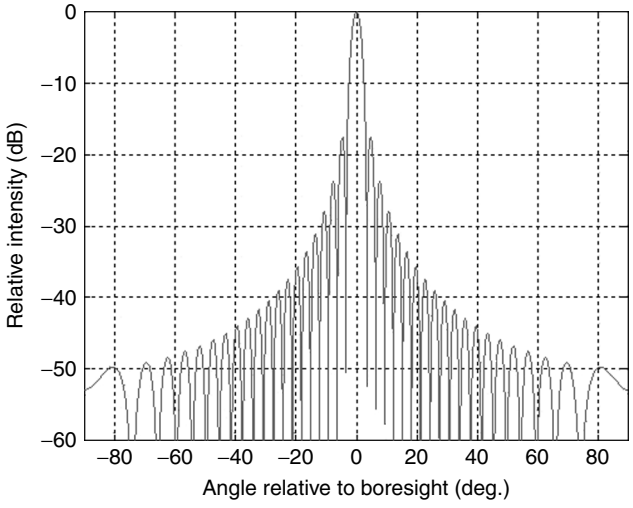


Figure 1. Far-field radiation pattern of an ideal, uniformly illuminated 20-wavelength-diameter circular aperture.

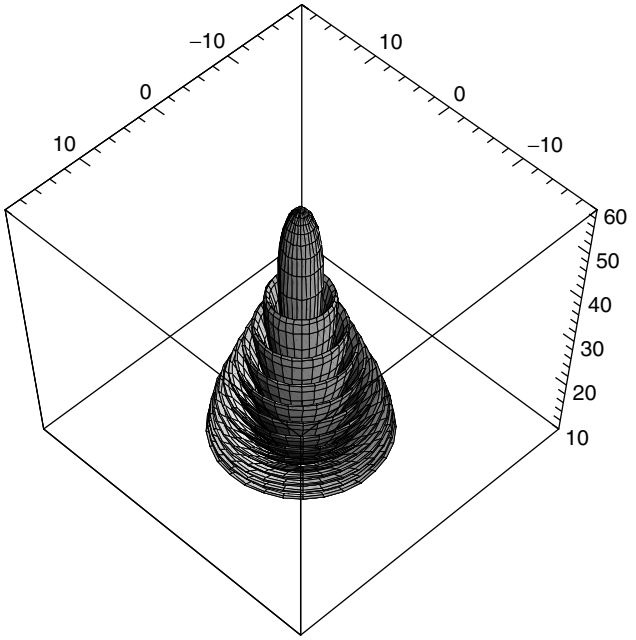


Figure 2. Three-dimensional polar plot of the far-field antenna pattern of field intensity (in dB) produced by a uniformly illuminated 20-wavelength-diameter circular aperture.

given by [2]

$$G = \frac{4\pi A}{\lambda^2}$$

and for circular apertures of diameter D

$$G = \left(\frac{\pi D}{\lambda}\right)^2$$

or in decibels

$$G(dB) = 20 \log \frac{\pi D}{\lambda}$$

Doubling the antenna diameter — or doubling its frequency of operation — leads to a 6-dB increase in antenna gain (as long as no aberrations are introduced in the electrically larger aperture).

The 3-dB, or half-power, antenna beamwidth can be approximated for large apertures as [1]

$$BW = \lambda / D$$

and the beamwidth between first nulls is double this.

The first sidelobe level is another important antenna specification, as it gives a measure of crosstalk intensity outside the main beam. For circular apertures it is −17.3 dB, and for rectangular apertures it is −13.5 dB [9]. By altering the amplitude distribution, it is possible to reduce these first sidelobe levels; but in doing so, the peak gain decreases as well. For example, a same-size rectangular aperture with a triangular amplitude distribution in the x direction could be thought of as the spatial convolution of two half-size uniform amplitude distributions, which would produce a radiation pattern with dependence $\sin^2(k_x a/2)/(k_x a/2)^2 \bullet \sin(k_y b)/(k_y b)$. The first sidelobe level in the x direction is lowered by 6 dB to −19.5 dB, but since the effective area is halved, the gain is also reduced by 3 dB.

Variations in the current phase across the aperture have an even greater effect on the radiation pattern than amplitude variations. A linear phase variation steers or "scans" the beam in the direction perpendicular to a plane of constant phase, without changing the beam pattern shape. Any quadratic or higher-order phase variations across the aperture are aberrations and lead to beam pattern degradation. It is essential to understand the effects of phase aberrations when designing antenna systems [12].

The simplest aperture antennas provide a single high-gain beam in the boresight direction. More sophisticated antennas produce specifically tailored beam shapes, or multiple simultaneous or independently excited beams. To maintain good beam quality, each radiated beam must have only linear phase variation. Although maintaining a flexible linear phase function while minimizing higher-order optical aberration terms is generally difficult, a phased-array antenna can accomplish these goals fairly simply by applying the required phase shifts to successive elements. With enough elements, a linear variation can be approached as closely as desired.

For single fixed-beam antennas, the aperture conditions depend only on the reflector surface placement and geometry. There is only one source input, the antenna feed, with fixed position and orientation. However, a scanning antenna must be able to produce a variable linear phase distribution that depends on varying source conditions.

One important issue with reflector antennas is blockage of the aperture with the feed or subreflector structure. There is no blockage with the transmission-based lens antennas, because the feed structure is on the other side of the antenna from the radiating aperture. However, surface impedance-matching requirements and dissipative transmission losses, as well as weight and bandwidth limitations, render lenses inferior to reflectors in most satellite applications.

2. REFLECTOR DESIGN FUNDAMENTALS

2.1. Geometric Optics Analysis

While the radiation pattern of a reflector antenna is determined by diffraction analysis of the fields across its radiating aperture, the conventional method for reflector shape synthesis and optimization makes use of geometric optics. Geometric optics follows from the eikonal equation [13,6] $|\nabla L(\boldsymbol{r})|^2 = n^2$, which is the lowest order (infinite frequency) approximation of wave equation, in which all field variation is assumed to be contained in the phase $\Phi = \omega/c\ L(r)$. This approximation assumes that all waves in uniform media propagate along straight rays, where each wavefront is perpendicular to these rays, occurring at the same distance from a given source. The rays are reflected by metal reflector surfaces as if they encountered piecewise perfectly conducting planes with the same surface normal at the reflection point [14–17]. Geometric optics cannot be used to find the

far-field radiation pattern of a well-formed beam, because this pattern is entirely determined by diffraction [18,19]. However, since a focused reflector antenna must receive plane waves from distant sources, it can be thought of as a single surface — or collection of multiple surfaces — that converts parallel incoming rays into rays that converge on a feed element. Conversely, when transmitting, reflectors *collimate* rays by ideally converting spherically diverging rays from a point-source feed to a set of parallel rays. Furthermore, the pathlength from the source point (or any starting wavefront) to any given wavefront along each ray must be the same regardless of one or more redirections of the rays by reflector surfaces.

The behavior of rays incident on metal surfaces is governed by Snell's law of reflection, which ensures that incident and reflected rays and the surface normal at the point of intersection are coplanar, and that the angle of incidence equals the angle of reflection. Snell's law and the constant-pathlength condition are sufficient to synthesize a focused reflector system. In general, two additional constraints are applied to the reflector surface synthesis formulation: surface continuity and field amplitude concentration.

Although phased arrays and some lens antennas divide the aperture into separate physical regions [20–22], reflectors tend to be continuous surfaces, with few steps or cusps. Discontinuities in the surface or the surface normal lead to diffraction effects that are unpredictable with geometric optics. Thus, the surface reflection point and its normal are coupled, and specifying the pathlengths of the set of reflected rays along with their directions greatly restricts the family of focusing surfaces. For a single reflector, only the circular paraboloid satisfies the Snell's law, constant-pathlength, and continuous-surface constraints. For multireflector systems, Snell's law must be specified on each reflector, and the pathlength condition applies to the full path from source to aperture plane. In designing reflector antenna systems, usually a two-dimensional profile is generated first; then this cross section is rotated about the central system axis line of symmetry.

The other useful addition to geometric optics analysis is the assumption that each ray represents a constant differential power flow [9]. The power within a bundle of rays must remain constant throughout the optical system. Power density on any two wavefronts is inversely proportional to the ratio of areas on those wavefronts bounded by a given set of rays. This property of conservation of energy is useful in many synthesis applications.

Additional parameters must be specified if the energy-conservation constraint is imposed. This constraint merely describes how the feed phase and amplitude distribution are transformed to an aperture distribution. Thus it is necessary to specify the input and output distributions. In many cases discussed in the literature, the input is assumed to be a tapered feed pattern with amplitude of the form $\cos^n \theta$, and with a spherical phase distribution [4]. In many cases, attempts are made to ensure that the aperture has high illumination efficiency, with the output field as close as possible to a uniformly amplitude plane wave. Occasionally the exact pathlength constraints are relaxed in favor of a particular aperture amplitude distribution.

Many attempts have been made to alter the geometry of dual reflectors to improve their performance [23–32]. Unlike with the single reflector, there are an infinite number of pairs of surfaces that produce a plane-wave output for spherical wave input at the focus. Although it is possible to generate a symmetric dual reflector with various arbitrary phase and amplitude distributions at the aperture [33,34], most efforts have been directed toward simply improving the main reflector illumination efficiency.

2.2. Single Reflectors

The simplest reflector antenna systems are single reflectors with parabolic cross section, as shown in Fig. 3. Often referred to as *prime focus reflectors*, these paraboloidal reflectors are specified entirely by their focal length F, aperture diameter D, and vertex position z_0. The paraboloid shape is given by [35]

$$z = \frac{x^2 + y^2}{4F} + z_0$$

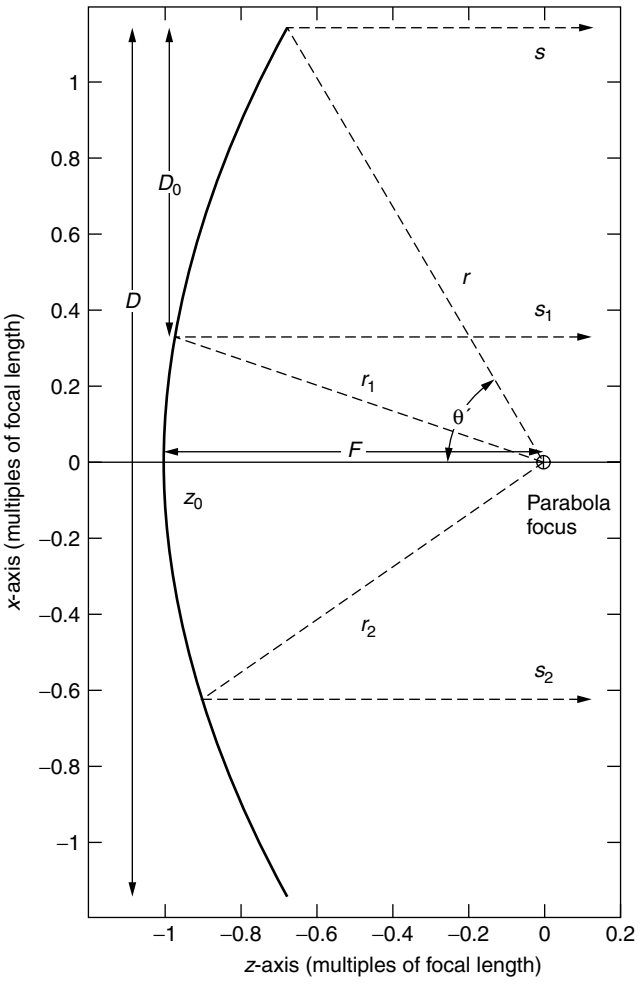


Figure 3. Parabolic single-reflector geometry, with offset section shown.

or equivalently, in terms of distance from the focus to the reflector surface r and angle from the negative axis of symmetry θ' [7,8]:

$$r = \frac{2F}{1+\cos\theta'} = F\sec^2\frac{\theta'}{2}$$

The feed is positioned at the parabola focus, facing the paraboloid at a distance F from its vertex. The pathlength condition demands that for any combination of rays, the total distance to the planar wavefront $r+s$, or r_1+s_1, or r_2+s_2 be the same. Since the feed structure lies in the path of outwardly reflected rays, it tends to block the central portion of the aperture.

An important characteristic of this single-reflector antenna is the ratio of focal length to diameter F/D [also known as the *f number* (or *f stop*; aperture) of camera lenses]. Reflectors with smaller F/D have greater surface curvature, and are more physically compact, with closer feed structures than those with larger F/D ratios, but are more sensitive to the precise feed position and beam pattern.

Offset single reflectors avoid the blockage problem by using only a portion of a larger symmetric paraboloid that reflects unblocked rays [36,37]. In Fig. 3, a possible offset parabolic section would be bounded by rays s and s_1, and have diameter D_0. While offset designs enhance both radiated power and prevent some beam distortion, the lack of perfect symmetry introduces differential effects for waves polarized parallel and perpendicular to the offset direction.

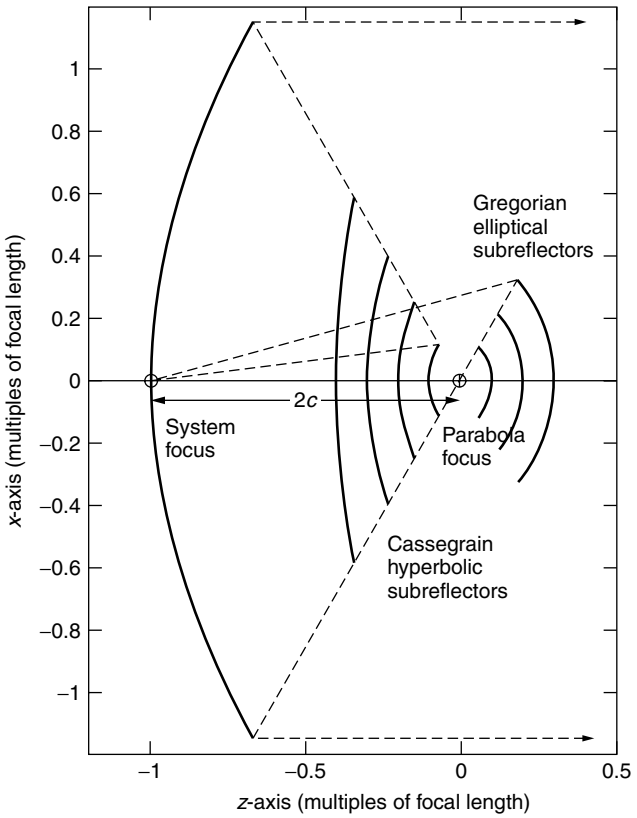


Figure 4. Cassegrain and Gregorian dual-reflector profiles with sample ray paths.

2.3. Multiple Reflectors

Conventional dual reflectors are of two major types: the Cassegrain and the Gregorian. Based on the optical telescope designs first introduced in 1672, they consist of a paraboloidal main reflector and, respectively, either a hyperboloidal or ellipsoidal secondary or subreflector [38]. As long as the paraboloid focus coincides with one hyperboloid (ellipsoid) focus, and the feed is positioned at the other hyperboloid (ellipsoid) focus, the dual reflector will collimate outgoing rays. For a given paraboloidal main reflector, there are infinitely many possible subreflectors with differing size and position relative to the main reflector. A family of possible Cassegrain and Gregorian subreflector profiles for a parabola with unit focal length and system focus at the parabola vertex is shown in Fig. 4. All rays originating at the system focus reflect from a subreflector as if they had originated at the common focal point. Since the difference of segments from each hyperbola focus to a point on the hyperbola is constant and equal to the major hyperbola axis length $2a$, the pathlength from feed to an aperture plane is again constant, $2a$ greater than for the main reflector if it were used as a prime focus single reflector. The same argument applies for the ellipse, with constant sum of segments from each focus equal to $2a$.

The equation for the subreflector profile is given by

$$\frac{(z-z_c)^2}{a^2} + \frac{x^2+y^2}{c^2-a^2} = 1$$

where the central point is found by associating the common foci $z_c = z_0 + F - c$. The equivalent polar form of the subreflector equation is [5]

$$r_s = \frac{c^2-a^2}{a+c\cos\theta'}$$

where the distance between subreflector foci $2c = F$ for the profiles of Fig. 4, with positive values of r_s corresponding to hyperbolas and negative values corresponding to ellipses.

It should be noted that the system focus need not be positioned at the parabola vertex. The subreflector size to chosen so that the extreme ray from feed to the edge of the paraboloid reflects from its edge. Generally, subreflectors are kept as small as possible to minimize blockage without requiring too narrow a taper of the feed radiation pattern.

Dual-reflector systems offer an improvement over paraboloidal single reflectors in both packaging efficiency and design flexibility. Because the subreflector redirects rays to the main reflector, the actual length involved in the system is much less than its equivalent focal length. Also, when a subreflector is used, the feed can be positioned near or behind the main reflector, so that feedlines are shortened, and the placement of the associated electronics is simplified. Servicing and maintenance of the feed is simplified as well. Another mechanical advantage is the ease of supporting a relatively lightweight subreflector.

There are several disadvantages to symmetric (nonoffset) dual reflectors. The subreflector blocks the center

of the aperture, thereby reducing efficiency and increasing the sidelobe levels. The fact that two surfaces must intercept the rays also introduces the requirement for minimizing power missing the surfaces (spillover) while keeping the illumination efficiency high. Alignment is crucial in dual reflectors. Unlike single reflectors, where the feed concentrates power over the wide angle subtended by the main reflector, the dual-reflector feed must illuminate a much smaller subreflector. Thus larger feedhorns are required, and their placement must be precise. Also, the subreflector placement relative to the main reflector is critical [39–42]. Larger feedhorns are less preferable for multibeam antennas, since their size may prevent their phase centers from being close enough to generate adjacent component beams.

With symmetric dual reflectors, subreflector blockage can be regarded as the removal of a circle of uniform, constant-amplitude power from the center of the aperture. Thus one subtracts a lower intensity, much wider $J_1(u')/u'$ pattern from the original far-field amplitude pattern. This effect lowers the mainlobe power level, lowering gain and increasing sidelobe magnitude. Power spilling past the subreflector is sufficiently large and sufficiently close to the axis to also become apparent on the combined pattern. With single reflectors, the spillover is almost always at least 150° away from the boresight axis, so its effect is negligible for communications antennas.

Methods of making the subreflector less obstructive have been proposed [43–47], such as serrating or perforating the surface; or fabricating it from a linearly polarized material, illuminating it with radiation of the orthogonal polarization, and using a twist reflection material for the main reflector. A much more feasible concept is the offset configuration, which consists of a (usually circular) section of the main reflector and the corresponding section of the subreflector of a symmetric dual reflector. Blockage is eliminated, and the VSWR (reflection back into the feed) is reduced, but the focal length : diameter ratio (F/D) increases by the ratio of the parent main reflector diameter to the offset section diameter. The lack of symmetry also tends to increase cross-polarization and make analysis much more difficult.

Offset designs often make use of the Gregorian configuration. In the symmetric case, having a concave subreflector increases the reflector separation without much improvement in performance. For an offset geometry, however, the Gregorian subreflector can be placed entirely below the antenna axis, where it can still illuminate the entire main reflector above the axis. Blockage is avoided while still using a large offset section of the main reflector. By properly tilting the axis of the ellipsoidal subreflector with respect to the main reflector focal axis, cross-polarization can be significantly reduced [48,49].

Several methods are used for analyzing Cassegrain antennas. One very powerful approximate technique is that of defining the equivalent parabola (Fig. 5), which has the same aperture and feed illumination angle as the Cassegrain [50,51]. The equivalent focal length can be found using the maximum subreflector angle θ_m to be

$$F_e = F\frac{c+a}{c-a}$$

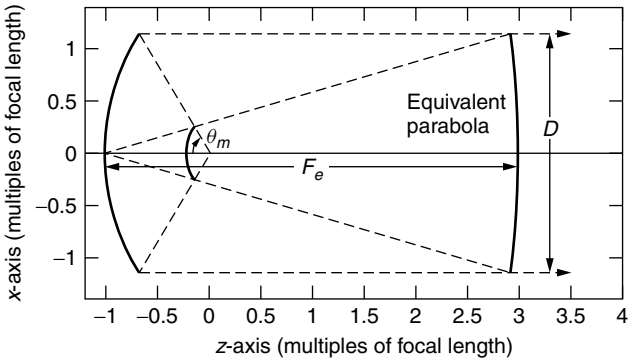


Figure 5. Equivalent parabola for Cassegrain dual reflector.

The factor $m = (c + a)/(c - a)$ is called the *magnification factor*. The F/D ratio of the equivalent parabola is m times larger than the physical Cassegrain system. The equivalent parabola concept can be used to determine the amplitude distribution across the aperture by replacing the real dual-reflector system with a single reflector.

Scattering theory and geometric theory of diffraction have been used for a more detailed analysis of dual reflectors [40–42,52–54]. The mathematics is complicated, and numerical methods are often required. Dual reflectors have less frequency dependence because of their relatively long effective focal length for their given packaging volume. Also, illumination efficiency is directly dependent on the system's geometry, and this tends to have the dominant effect in scanning.

2.4. Dual-Reflector Shaping

Much effort has been spent on dual-reflector designs that keep one of the standard Cassegrain surfaces fixed (usually the main reflector) and shape the other. Green first showed that by shaping the subreflector and then by making small alterations to the main reflector to preserve the planar phase front, uniform illumination could be attained [28]. The resulting beam is not exactly focused, but the improvement in performance is appreciable. Other systems consider subreflectors that correct for main reflector deformations or errors, or generate particular aperture distributions and polarizations [29,32,55,57].

Simultaneous shaping of both surfaces of dual-offset reflectors has been explored [26,27,34,35,43–47]. The problem involves solving a pair of partial nonlinear differential equations. These simplify to ordinary differential equations for symmetric systems. Approximate methods of solving the equations show reasonable numerical results [46], but the processes involve considerable trial and error and significant amounts of numerical computation. However, Westcott et al. [47], claim to have arrived at an exact solution.

3. BEAM SCANNING WITH REFLECTOR ANTENNAS

3.1. Analysis of Reflector Beam Scanning

For reflectors and lenses, beam steering, or "scanning" is accomplished by transversely displacing the feed from

the antenna focus, or unscanned, aberration-free source position [58,59]. For systems with a single perfect focal point, this displacement always produces higher-order phase terms. The absence of higher-order terms at any other position would imply the existence of a second perfect focal point there.

Scanning is of great concern for antennas used in satellite applications. When high resolution is important, as is the case with point-to-point networking and direct broadcast communications, the antenna must provide a large number of high-gain, well-formed spot beams over its entire field of view [60–67]. Well-formed beams are also required in multibeam antennas. High-efficiency shaped coverage contours are useful for illuminating specific geographic regions. Because they are produced by coherently combining component beams, each beam must have low-intensity sidelobes and well-defined nulls to prevent undesirable interference.

The optical aberrations, caused by lateral feed displacement, cause an asymmetric effect in the plane of scan [13–15,68]. The beam broadens, the first null in the scanned direction fills, and the first sidelobe on the axial side, the coma lobe, rises. Several authors have addressed the problem of finding the best focal locus of a paraboloidal antenna [58,59,61,69]. The most general solution appears to be given in Balling [62].

For small scan angles, the equivalent parabola concept is very useful [51] for predicting the performance degradation in dual reflectors. The scanning tradeoff analysis of dual reflectors is much more complex than it is for paraboloids. The effective focal length increases as subreflector curvature increases; however, the subreflector is smaller and spillover is of greater concern. For larger feed displacements and scan angles, spillover and distorted imaging prevents acceptable modeling of the dual reflector as an equivalent paraboloid [70,71]. Choosing the optimum focal point for a particular scan angle is strongly dependent on the particular Cassegrain geometry. Changing the surface curvature of either reflector surface will greatly change the position of the focal surface.

Scanning with an offset antenna system is slightly different from that in the symmetric case. The optimum focal locus is no longer symmetrically disposed about the antenna axis. Instead, the best focal positions for scanned beams, to the first approximation, are on the plane perpendicular to the offset axis [36,37,64,72]. The precise feed position depends on spillover and illumination efficiency as well as pathlength and phase error considerations.

3.2. Reflectors Designed for Scanning

Several antennas have been developed specifically for efficient scanning applications. The spherical cap [73] is the simplest scanning antenna. With a source positioned approximately halfway between the sphere and its center, the surface resembles a paraboloid over a limited angular region. In fact, if the parabola's curvature value at its vertex is inverted and used as the radius of the osculating circle at that vertex, this circle corresponds to the profile of the spherical cap. Several feeds can be positioned on the spherical focal surface (with radius $F/2$), each generating a beam in a different scanned direction.

Because it is not a perfect paraboloid, the spherical cap introduces phase errors in the reflected wave in the form of spherical aberration. This optical aberration increases with larger angular illumination from the feed or with higher frequencies. Since the symmetry of the sphere produces identical patterns for all scan directions, the entire reflector is underilluminated for wide-field-of-view ($>5^\circ$) applications; that is, only a fraction of the reflector is illuminated by each feed for any given beam. A compromise is made frequently between scanning perfection and reflector inefficiency.

Subreflector correctors have been designed to eliminate the spherical aberration in the cap [74]. The Gregorian type corrector makes use of caustics formed by rays reflected by the spherical surface to collimate incoming rays. The corrector applies only for a single scan direction; therefore, it must be moved mechanically if the beam is to be redirected. This characteristic makes scanning arrays impractical and shaped beams impossible for corrected spherical caps.

One improvement to the basic spherical cap is the torus [3,75]. The torus is a double-curved surface with a circular profile in the plane of scan and a parabolic profile in the orthogonal plane. Reflected waves suffer from spherical aberration in the scan plane, but are focused without aberration in the orthogonal plane. The torus can scan in only one plane, so it is well suited for communication with multiple stations situated on a single arc, such as a group of satellites in geostationary orbit. Like the spherical cap, the torus has poor illumination efficiency in the scan direction.

An alternative to the torus is a specifically shaped single-reflector surface that balances the aperture phase errors with illumination efficiency [76,77]. This surface is described by a polynomial whose coefficients minimize the variations across a symmetric pair of specified subapertures of the pathlengths from a symmetric pair of specified focal points to a pair of scanned aperture planes. In addition to minimizing the optical aberrations for these extreme scan angles, a third beam direction — that of boresight — is considered as well. The feed is positioned along the axis of symmetry, at the point that minimizes phase errors across a central subaperture, and the surface polynomial coefficients are adjusted to reduce the errors for the boresight beam without greatly worsening the errors for the scanned beams. It has been shown that for equivalent beam quality across a 60° wide field of view, this type of reflector can be made about 40% smaller than the torus.

Dual offset reflectors have been designed specifically for scanning and multiple beam applications [60,63–67,78–85]. Acceptable multibeam systems are achievable for large F/D conventional Cassegrain systems with oversized main and subreflectors. Unconventional designs include the confocal paraboloid [64,79], which has a feed array in the near field of a small offset parabolic subreflector with focal point coinciding with that of a parabolic main reflector. This type of antenna has been shown to scan effectively up to 3°.

The bifocal dual reflector [82] is a totally redesigned antenna system, with two perfect focal points, one for each extreme scanned beam. Since it consists of two reflectors, each tracing of rays through the system has 2 degrees of freedom. The two Snell's law conditions allow for two independent collimating pathlength conditions. The design principle for the bifocal is based on this idea by insisting that each point on each reflector will lie on the path from one focal point to its corresponding aperture plane, as well as on the path from the second point to its aperture plane. First, the focal points are selected, an initial subreflector point and it normal are chosen, and the total pathlength to the first scanned aperture plane is specified. Ray tracing uniquely determines the corresponding main reflector point and its normal. Next, this main reflector point is used with ray tracing from the second focal point to the second scanned aperture to determine the subsequent subreflector point. This process continues until a full set of reflector profile points are generated. These points are joined by spline fitting, and then the profiles are rotated about the axis of symmetry to generate a pair of reflector surfaces. The focal points are also smeared out into a focal ring, which unfortunately introduces aberrations.

The need for an oversized subreflector to simultaneously redirect both positively and negatively scanned rays to the main reflector causes severe blockage in the symmetric bifocal. The offset bifocal reflector antenna system [83] overcomes this deficiency by choosing asymmetric focal points, limiting the illuminated sections of the main reflector to only an unblocked aperture, and synthesizing each entire reflector surfaces from the ray-tracing procedure rather than by rotating the profiles about the axis of symmetry. The offset bifocal surfaces only approximate the continuous surface condition, but the resulting dual-reflector system produces insignificant phase errors. Performance results indicate that it is possible to design a compact dual-reflector system that can radiate well-formed 1° beams across the 17° field of view subtended by the earth as seen from geostationary orbit.

One last type of scanning dual-reflector antenna system incorporates the ideas of the shaped single reflector [76,77] with a mechanically tilting subreflector [84–86]. This configuration is driven by the need to minimize antenna cost, by using only a single feed along with a single front end (amplifier, filter, polarizer), yet providing the capability to scan a single beam across a wide field of view. Having the smaller subreflector rotate instead of the entire antenna structure provides mechanical and cost advantages for mass-market communication systems such as direct broadcast television. The subreflector can be planar or specially shaped [86] for additional packaging benefits.

BIOGRAPHY

Carey M. Rappaport (M., S.M. 1996) received five degrees from the Massachusetts Institute of Technology: the S.B. in Mathematics, the S.B., S.M., and E.E. in Electrical Engineering in June 1982, and the Ph.D. in Electrical Engineering in June 1987. He is married to Ann W. Morgenthaler and has two children, Sarah and Brian.

Professor Rappaport has worked as a teaching and research assistant at MIT from 1981 until 1987, and during the summers at COMSAT Labs in Clarksburg, Maryland, and The Aerospace Corp. in El Segundo, California. He joined the faculty at Northeastern University in Boston, in 1987. He has been Professor of Electrical and Computer Engineering since July 2000. During fall 1995, he was Visiting Professor of Electrical Engineering at the Electromagnetics Institute of the Technical University of Denmark, Lyngby, as part of the W. Fulbright International Scholar Program. He has consulted for Geo-Centers, Inc., PPG, Inc., and several municipalities on wave propagation and modeling, and microwave heating and safety. He is Principal Investigator of an ARO-sponsored Multidisciplinary University Research Initiative on Demining and Co-Principal Investigator of the NSF-sponsored Center for Subsurface Sensing and Imaging Systems (CenSSIS) Engineering Research Center.

Professor Rappaport has authored over 190 technical journal and conference papers in the areas of microwave antenna design, electromagnetic wave propagation and scattering computation, and bioelectromagnetics, and has received two reflector antenna patents, two biomedical device patents, and three subsurface sensing device patents. He was awarded the IEEE Antenna and Propagation Society's H.A. Wheeler Award for best applications paper, as a student in, 1986. He is a member of Sigma Xi and Eta Kappa Nu professional honorary societies.

BIBLIOGRAPHY

1. S. Silver, ed., *Microwave Antenna Theory and Design*, Dover Publications, New York, 1965.
2. D. Staelin, A. Morgenthaler, and J. Kong, *Electromagnetic Waves*, Prentice-Hall, Englewood Cliffs, NJ, 1994.
3. R. E. Collin and F. J. Zucker, *Antenna Theory*, McGraw-Hill, New York, 1969.
4. J. D. Kraus and R. Marhefka, *Antennas*, McGraw-Hill, New York, 2002.
5. R. S. Elliot, *Antenna Theory and Design*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
6. J. Kong, *Electromagnetic Wave Theory*, Wiley, New York, 1999.
7. W. Stutzman and G. Thiele, *Antenna Theory Design*, Wiley, New York, 1981.
8. C. Balanis, *Antenna Theory Analysis and Design*, Wiley, New York, 1998.
9. R. Collin, *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985.
10. D. K. Cheng, Effect of arbitrary phase error on the gain and beamwidth characteristics of radiation pattern, *IRE Trans. Antennas Propag.* **AP-3**: 145–147 (July 1965).
11. T. B. Vu, The effect of phase errors on the forward gain, *IEEE Trans. Antennas Propag.* **AP-13**: 981–982 (Nov. 1965).
12. K. S. Kelleher, *Antenna Wavefront Problems*, Naval Research Lab., Washington, DC, Sept. 1949.

13. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, New York, 1970.
14. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill, New York, 1957.
15. A. E. Conrady, *Applied Optics and Optical Design, Part 1*, Dover Publications, New York, 1957.
16. P. S. Holt, *Application of Geometric Optics to the Design and Analysis of Microwave Antennas*, AFCRL, Bedford, MA, AFCRL-67-0501, Sept. 1967.
17. G. A. Fry, *Geometrica1 Optics*, Chilton, Philadelphia, 1969.
18. S. Cornbleet, *Microwave Optics*, Academic Press, New York, 1976.
19. R. S. Longhurst, *Geometrica1 and Physical Optics*, Wiley, New York, 1967.
20. F. S. Holt and A. Mayer, A design procedure for dielectric microwave lenses of large aperture ratio and large scanning angle, *IRE Trans. Antennas Propag.* 25–30 (Jan. 1957).
21. R. M. Brown, Dielectric bifocal lenses, *IRE National Convention Record*, 1956, pp. 180–187.
22. C. Rappaport and A. Zaghloul, Optimized three dimensional lenses for wide-angle scanning, *IEEE Trans. Antennas Propag.* 1227–1236 (Nov. 1985).
23. S. P. Morgan, Some examples of generalized Cassegrainian and Gregorian antennas, *IEEE Trans. Antennas Propag.* **AP-12**: 685–691 (Nov. 1964).
24. P. Brickell and B. S. Westcott, Reflector design as an initial-value problem, *IEEE Trans. Antennas Propag.* (Communication) **AP-24**: 531–533 (July 1976).
25. B. S. Westcott and A. P. Norris, Reflector synthesis for generalized far fields, *J. Phys. A, Math. Nucl. Gen.* **8**: 521–532 (1975).
26. G. W. Collins, Shaping subreflectors in Cassegrainian antennas for maximum aperture efficiency, *IEEE Trans. Antennas Propag.* **AP-21**: 309–313 (May 1973).
27. W. F. Williams, High efficiency antenna reflector, *Microwave J.* **8**: 79–82 (July 1965).
28. K. A. Green, Modified Cassegrain antenna for arbitrary aperture illumination, *IRE Trans. Antennas Propag.* (Communication) **AP-11**: 589–590 (Sept. 1963).
29. T. Kitsuregawa and M. Mizusawa, Design of the shaped reflector Cassegrainian antenna in consideration of the scattering pattern of the subreflector, *IEEE Group; Antennas and Propagation, Int. Symp. Digest*, Sept. 9–11, 1968, pp. 391–396.
30. P. Rouffy, Design of dual reflector antennas, *URSI. 1968 Symp. Digest*, Sept. 10–12, 1968, p. 88.
31. S. Von Hoerner, The design of correcting secondary reflectors, *IEEE Trans. Antennas Propag.* **AP-24**: 336–340 (May 1976).
32. M. O. Millner and R. H. T. Bates, Design of subreflectors to compensate for Cassegrain main reflector deformations, *Proc. IEEE, Microwaves, Optics and Antennas*.
33. V. Galindo, Design of dual reflector antennas with arbitrary phase and amplitude distributions, *IEEE Trans. Antennas Propag.* **AP-12**: 403–408 (July 1964).
34. B. Y. Kinber, On two reflector antennas, *Radio Eng. Electron. Phys.* **6**: 914–921 (June 1962).
35. W. V. T. Rusch and P. O. Potter, *Analysis of Reflector Antennas*, Academic Press, New York, 1970.
36. P. G. Ingerson and W. C. Wong, Focal region characteristics of offset fed reflectors, *1974 Int. IEEE/AP-S Symp. Program and Digest*, June 10–12, 1974, pp. 121–123.
37. A. W. Rudge and N. A. Adatia, Offset-parabolic-reflector antennas: A review, *Proc. IEEE* **66**: 1592–1618 (Dec. 1978).
38. P. W. Hannan, Microwave antenna derived from the Cassegrain telescope, *IRE Trans. Antennas Propag.* **AP-9**: 140–153 (March 1961).
39. A. M. Isber, Obtaining beam pointing accuracy with Cassegrain antennas, *Microwaves* 40–44 (Aug. 1967).
40. W. V. T. Rusch, Scattering from a hyperboloidal reflector in a Cassegrainian feed system, *IEEE Trans. Antennas Propag.* **AP-11**: 414–421 (July 1963).
41. P. D. Potter, Application of spherical wave theory to Cassegrainian-fed paraboloids, *IEEE Trans. Antennas Propag.* **AP-15**: 727–736 (Nov. 1967).
42. P. D. Potter, Aperture illumination and gain of a Cassegrain system, *IEEE Trans. Antennas Propag.* **AP-71**: 373–375 (May 1963).
43. G. Bjontegaard and T. Pettersen, A shaped offset dual reflector antenna with high gain and low sidelobe levels, *IEE 2nd Int. Conf. Antennas and Propagation*, April 1981, York, UK, pp. 163–167.
44. R. Mittra and V. Galindo-Israel, Shaped dual reflector synthesis, *IEEE Antennas Propag. Newsl.* 5–9 (Aug. 1980).
45. V. Galindo-Israel, R. Mittra, and A. Cha, Aperture amplitude and phase control of offset dual reflectors, *IEEE Trans. Antennas Propag.* **AP-27**: 159–164 (March 1979).
46. J. J. Lee, L. I. Parad, and R. S. Chu, Shaped offset-fed dual reflector antenna, *IEEE Trans. Antennas Propag.* **AP-27**: 165–171 (March 1979).
47. B. S. Westcott, F. A. Stevens, and P. Brickell, GO synthesis of offset dual reflectors, *IEEE Proc.* **128**: 11–18 (Feb. 1981).
48. T. Mizuguchi, M. Akagawa, and H. Yokoi, Offset Gregorian antenna, *Electron. Commun. Jpn.* **61-B**(3): 58–66 (1978).
49. M. Tanaka and M. Mizusawa, Elimination of cross polarization in offset dual-reflector antennas, *Electron. Commun. Jpn.* **58-B**(12): 71–78 (1975).
50. W. D. White and L. K., DeSize, Focal length of a Cassegrain reflector, *IRE Trans. Antennas Propag.* **AP-9**: 412 (Jan. 1961).
51. W. C. Wong, On the equivalent parabola technique to predict the performance characteristics of a Cassegrainian system with an offset feed, *IEEE Trans. Antennas Propag.* **AP-21**: 335–339 (May 1973).
52. L. K. DeSize, D. J. Owen, and G. K. Skahill, *Investigation of multibeam antennas and wide-angle optics*, Airborne Instruments Laboratory Report 7358–1, Jan. 1960.
53. O. Sorensen and W. V. T. Rusch, Application of the geometric theory of diffraction to Cassegrain subreflectors with laterally defocused feeds, *IEEE Trans. Antennas Propag.* **AP-73**: 698–701 (Sept. 1975).
54. C. A. Mentzer and L. Peters, A GTD analysis of the far-out sidelobes of Cassegrain antennas, *IEEE Trans. Antennas Propag.* **AP-23**: 702–709 (Sept. 1975).
55. Q. Ji-zen, Equivalent phase center of the sub-reflector in the shaped Cassegrain antenna, *IEE 2nd Int. Conf. Antennas and Propagation*, April 1981, York, UK, pp. 204–206.
56. S. K. Buchmeyer, An electrically small Cassegrain antenna with optically shaped reflectors, *IEEE Trans. Antennas Propag.* **AF-25**: 346–351 (May 1977).

57. P. J. B. Clarricoats, Some recent developments in microwave reflector antennas, **126**(1): 9–25 (Jan. 1979).

58. J. Ruze, Lateral feed displacement in a paraboloid, *IEEE Trans. Antennas Propag.* **AP-13**: 660–665 (Sept. 1965).

59. W. A. Imbriale, P. G. Ingerson, and W. C. Wong, Large lateral feed displacements in a parabolic reflector, *IEEE Trans. Antennas Propag.* **AP-22**: 742–743 (Nov. 1974).

60. E. A. Ohm, A proposed multiple-beam microwave antenna for Earth stations and satellites, *Bell Syst. Tech. J.* **53**: 1657–1665 (Oct. 1974).

61. A. W. Rudge, Multiple-beam antennas: Offset reflectors with offset feeds, *IEEE Trans. Antennas Propag.* **AP-23**: 234–239 (May 1975).

62. P. Balling, R. Jorgensen, and K. Pontoppidan, *Study of techniques for design of high gain antennas with Contoured Beams*, Final Report ESTEC Contract 3371/77/NL/AK, Dec. 1978.

63. T. S. Bird, J. L. Boomars, and P. J. B. Clarricoats, Multiple-beam-dual-offset reflector antenna with an array feed, *Electron. Lett.* **14**(14): 439–440 (July 1978).

64. K. Woo, Array-fed reflector antenna design and applications, *IEE 2nd Int. Conf. Antennas and Propagation*, April 1981, York, UK, pp. 209–213.

65. E. A. Ohm and M. J. Gans, Numerical analysis of multiple-beam offset Cassegrainian antennas, *AIAA/CASI 6th Communications Satellite Conf.*, Apri1 5–8, 1976.

66. E. A. Ohm, System aspects of a multibeam antenna for full U.S. coverage, *Int. Conf. Communications*, June 10–14, 1979, pp. 49.2.1–49.2.5.

67. D. C. Chang and W. V. T. Rusch, Transverse beam scanning for an offset dual reflector system with symmetric main reflector, *IEE 2nd Int. Conf. Antennas and Propagation*, Apri1 1981, York, UK, pp. 207–208.

68. J. R. Cogdell and J. H. Davis, Astigmatism in reflector antennas, *IEEE Trans. Antennas Propag.* **AP-21**: 565–567 (July 1973).

69. W. V. T. Rusch and A. C. Ludwig, Determination of the maximum scan-gain contours of a beam-scanning paraboloid and their relation to the Petzval surface, *IEEE Trans. Antennas Propag.* **AP-21**: 141–147 (March 1973).

70. M. Akagawa and D. P. DiFonzo, Beam scanning characteristics of offset Gregorian antennas, *APS Symp. Digest*, June 1979, pp. 262–265.

71. W. D. White and L. K. DeSize, Scanning characteristics of two-reflector systems, *1962 IRE Convention Record*, Part I, pp. 44–70.

72. N. A. Adatia and A. W. Rudge, Beam squint in circularly polarized offset-reflector antennas, *Electron. Lett.* **11**(21): 513–515 (Oct. 1975).

73. T. Li, A study of spherical reflectors as wide-angle scanning antennas, *IRE Trans. Antennas Propag.* **47** (July 1959).

74. F. S. Holz and E. L. Bouche, A Gregorian corrector for a spherical reflector, *IEEE Trans. Antennas Propag.* **AP-12**: 44–47 (Jan. 1964).

75. C. Sletten, *Reflector and Lens Antennas*, Artech House, Norwood, MA, 1988.

76. C. Rappaport and W. Craig, High aperture efficiency, symmetric reflector antennas with up to 60 degrees field of view, *IEEE Trans. Antennas Propag.* **AP-39**(3): 336–344 (March 1991).

77. W. Craig, C. Rappaport, and J. Mason, A high aperture efficiency, wide-angle scanning offset reflector antenna, *IEEE Trans. Antennas Propag.* **41**(11): 1481–1490 (Nov. 1993).

78. G. Tong, P. J. B. Clarricoats, and G. L. James, Evaluation of beam-scanning dual reflector antennas, *Proc. IEEE* **124**(12): 1111–1113 (Dec. 1977).

79. W. D. Fitzgerald, *Limited Electronic Scanning with an Offset Feed Near-Field Gregorian System*, MIT Lincoln Laboratory, Technical Report 486, (Sept. 1971, DDC AP-736029.

80. M. Kumazawa and M. Karikomi, Multiple-beam antenna for domestic communication satellites, *IEEE Trans. Antennas Propag.* **AP-21**: 876–877 (Nov. 1973).

81. M. Karikomi, A limited steerable dual reflector antenna, *Electron. Commun. Jpn.* **55-B**(10): 62–68 (1972).

82. B. L. J. Rao, Bifocal dual reflector antenna, *IEEE Trans. Antennas Propag.* **AP-22**: 711–714 (Sept. 1974).

83. C. Rappaport, An offset bifocal reflector antenna design for wide angle scanning, *IEEE Trans. Antennas Propag.* 1196–1204 (Nov. 1984).

84. A. Garcia Pino, C. Rappaport, J. Rubinos, and E. Lorenzo, A shaped dual reflector antenna with a tilting flat subreflector for scanning applications, *IEEE Trans. Antennas Propag.* **43**(10): 1022–1028 (Oct. 1995).

85. E. Lorenzo, A. Pino, and C. Rappaport, An inexpensive scanning dual offset reflector antenna with rotating flat subreflector, *1995 Antennas and Propagation Society/URSI Symposium Digest*, June 20, 1995, pp. 1178–1181.

86. E. Lorenzo, C. Rappaport, and A. Pino, Scanning dual reflector antenna with rotating curved subreflector, *Progress Electromagn. Res. Symp.*, July 1998, p. 347.

RADIO RESOURCE MANAGEMENT IN FUTURE WIRELESS NETWORKS

JENS ZANDER
Royal Institute of Technology
Stockholm, Sweden

1. INTRODUCTION

The rapid increase of the size of the wireless mobile community and their demands for high speed, multimedia communications stands in clear contrast to the rather limited spectrum resource that has been allocated in international agreements. Efficient spectrum or Radio Resource Management (RRM) is of paramount importance due to these increasing demands. Figure 1 illustrates the principles of wireless network design. The network consists of a fixed network part and a wireless access system. The fixed network provides connections between base stations or Radio Access Ports (RAP), which in turn provide the wireless "connections" to the mobiles. The RAPs are distributed over the geographical area where we wish to provide the mobile users with communication services. We will refer to this area simply as the service area. The area around a RAP where the transmission conditions are favorable enough to maintain a connection of the required quality between a mobile and the RAP, is denoted the coverage area of the RAP. The transmission quality and thus

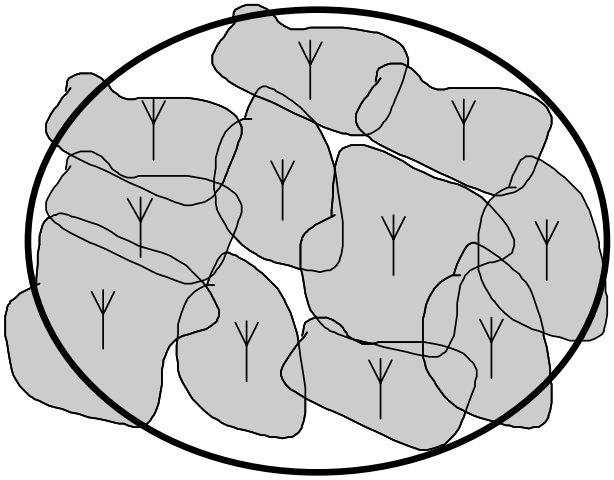


Figure 1. Wireless area communication system.

the shape of these regions will, as we may expect, depend heavily on the propagation conditions and the current interference from other users in the system. The coverage areas are therefore usually of highly irregular shape.

The fraction of the service area where communication with some required quality of service (QoS) is possible is called the coverage or the area availability of the system. In two-way communication systems (such as mobile telephone systems), links have to be established both from the RAP to the mobile (down- or forward link) and between the mobile terminal and the RAP (up- or reverse link). At first glance these two links seem to have very similar properties, but there are some definite differences from a radio communication perspective. The propagation situation is quite different, in particular in wide area cellular phone systems, where the RAP (base station) usually has its antennas at some elevated location free of obstacles. The terminals, however, are usually located amidst buildings and other obstacles creating shadowing and multipath reflections. Also, the interference situation in the up- and downlink will be different since there are many terminals and varying locations and quite few RAPs at fixed locations.

For obvious economical reasons, we would like our wireless network to provide ample coverage with as few RAPs as possible. Clearly this would not only minimize the cost of the RAP hardware and installation, but also limit the extent of the fixed wired part of the infrastructure. Coverage problems due to various propagation effects

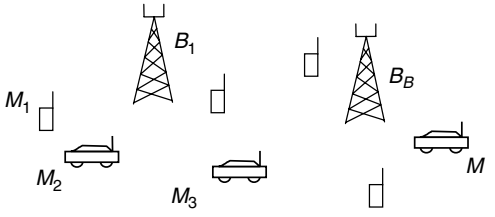


Figure 2. Resource management problem formulation.

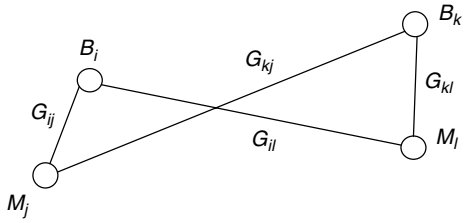


Figure 3. Link gains.

puts a lower limit to the number of RAP that are required. However not quite correct, one could say that the range of the RAPs is too small, compared to the inter-RAP distance. Such a system where this type of problem is dominant is called a range limited system. As the number of transmitters in the system becomes large within some fixed chunk of available RF-spectrum, the number of simultaneous connections (links) will become larger than the number of orthogonal signals that the available bandwidth may provide. In order to provide service for such a large population of users, it is obvious that the bandwidth used by the RAPs and terminals has to be reused in some clever way at the cost of mutual interference. The system is said to be bandwidth or interference-limited. Absolutely vital to the study of any resource management problem is a thorough understanding of the user requirements, that is, the required QoS and the traffic characteristics. All resource management schemes are designed (or optimized) using some model for the traffic. The resulting performance will clearly be a function of not only how well our design has been adapted to the traffic model, but also how accurate the traffic model is. Most wireless systems of today use circuit switched speech as the main design model (e.g., GSM). This does not prevent such systems to carry other types of traffic, but they always do this at a performance penalty. Future wireless access systems are expected to carry both large bandwidths as well as a mixture of services with very different and often conflicting service requirements. Particularly in these scenarios, accurate modeling is imperative for efficient resource utilization. Unfortunately, however, future user applications are little known and most work to derive realistic traffic models for these kinds of applications lies still ahead of the telecommunication community.

In the remainder of the article we present a more rigorous formulation of the radio resource management and review some of the ideas and results from the literature. Finally, we give an outlook on how these results can be applied to future wideband systems and which are the key problems that should be addressed in further studies.

2. RADIO RESOURCE MANAGEMENT — A GENERAL PROBLEM FORMULATION

Assume that M mobiles $(M_1, M_2, \ldots, M_M)$ are served by access ports (base stations), numbered from the set

$$\mathbf{B} = \{1, 2, 3, \ldots, B\}.$$

Now, let us assume that there are C (pairs of) waveforms (in conventional schemes these can be seen as orthogonal channels (channel pairs)) numbered from the set

$$\mathbf{C} = \{1, 2, 3, \ldots, C\}$$

available for establishing links between access ports and mobile terminals. To establish radio links, to each mobile the system has to assign

(a) an access port from the set **B**,
(b) a waveform (channel) from the set **C**,
(c) a transmitter power for the access port and the terminal.

This assignment (of access port, channel, and power) is performed according to the resource allocation algorithm (RAA) of the wireless communication system. The assignment is restricted by the interference caused by the access ports and mobiles as soon as they are assigned a "channel" and when they start using it. Another common restriction is that access ports are in many cases restricted to use only a certain subset of the available channels. Good allocation schemes will aim at assigning links with adequate SIR to as many (possibly all) mobiles as possible. Note that the RAA may well (should) opt for not assigning a channel to an active mobile if this assignment would cause excessive interference to other mobiles.

Let us now study the interference constraints on resource allocations in somewhat more detail. We now may compute the signal and interference power levels in all access ports and mobiles, given the link (power) gains, G_{ij}, between access port i and mobile terminal j. For the sake of simplicity, we will here consider only rather wideband modulation schemes that will make the link gains virtually independent of the frequency. Collecting all link gains in matrix form, we get a BxM rectangular matrix — the link gain matrix $\boldsymbol{G}$. The link gain matrix describes the (instantaneous) propagation conditions in the system. Note that in a mobile system, both the individual elements of the matrix (due to mobile motion) and the dimension of the matrix (due to the traffic pattern) may vary over time.

The task of the resource allocation scheme is to find assignments for the QoS that is sufficient in as many links as possible (preferably all). Providing a stringent definition to the QoS for a practical communication service is a complex and multifaceted problem. In this treatment we will confine ourselves to a simple measure the signal-to-interference ratio (SIR), or actually, to be precise, the signal-to-interference+noise ratio. This measure is strongly connected with performance measures as the bit or message error probability in the communication link. We require the SIR in link i to exceed a given threshold γ_i which is determined by both the QoS requirements for that particular link as well as by the modulation and coding formats of the system. This means that the following inequality must hold for both the up-(mobile-to-access port) and down-(access port-to-mobile) link of the connection:

$$\Gamma_i = \frac{P_j G_{ij}}{\sum_m P_m G_{im}\theta_{jm} + N} \geq \gamma_i \tag{1}$$

where Γ_j denotes the SIR at the receiver and N denotes the receiver (thermal) noise power at the access port. P_j denotes the transmitter power used by terminal j. The quantity θ_{jm} is the normalized crosscorrelation between the signals from mobiles j and m at the access port receiver, that is, the effective fraction of the received signal power from transmitter m contributes to the interference when receiving the signal from access port j. If the waveforms are chosen to be orthogonal (as in FDMA and TDMA) these correlations are either zero or one depending on if the station has been assigned the same frequency (time slot) or not. In nonorthogonal access schemes (e.g., DS-CDMA) the θ_{jm} take real values between zero and one. Note that we may not be certain that it is possible to comply with all the constraints (2) for all the M mobiles, in particular if M happens to be a large number. As system designers, we may have to settle for finding resource allocation schemes that assign channels with adequate quality to as many mobiles as possible.

In the classical single service (e.g., mobile telephony) case, the service requirements are identical in all links, that is, $\gamma_i = \gamma_0$ for all i. In this case the *system capacity* can be measured by the largest number of users that may be successfully handled by the system. Since the number of mobiles is a random quantity and the constraints (2) depend on the link matrix, that is, on the relative position of the mobiles, such a capacity measure is not a well-defined quantity. The classical approach for telephone type of traffic is to use as capacity measure the maximal relative arrival rate of calls ρ for which the blocking probability (the probability that a newly arrived session request is denied) can be kept below some predetermined level. Due to the mobility of the mobiles this is not an entirely satisfying measure. A call or session may be lost due to adverse propagation conditions. To include such phenomena into our capacity would require detailed specification of call-handling procedures (e.g., handling of new vs. old calls, hand-off procedures as a mobile moves from one access port to another, etc.). It may therefore be practical to choose a simpler and more fundamental capacity measure that will reflect the performance of the resource allocation scheme as such. For this purpose, the assignment failure probability ν (or assignment failure rate [14,15]) has been proposed. The instantaneous capacity $\omega * (\nu_o)$ of a wireless

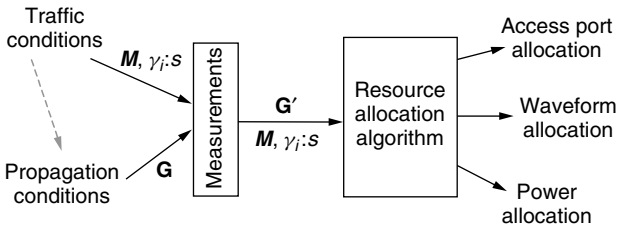


Figure 4. A resource allocation algorithm.

system is the maximum allowed traffic load in order to keep the assignment failure rate below some threshold level ν_o, that is,

$$\omega * (\nu_o) = \{\max \omega: \nu \leq \nu_o\} \quad (2)$$

As we have seen above, finding the optimum resource allocation for each mobile determining

(i) a waveform assignment (determining the θ_{jm})
(ii) an access port assignment (of one or more (!) ports)
(iii) a transmitter power assignment

that maximizes Y for a given link gain matrix, is a formidable problem. No efficient general algorithm that is capable of doing such an optimal assignment for arbitrary link gain matrices and mobile sets is known. Instead, partial solution and a number of more less complex heuristic schemes have been proposed (and are used in the wireless systems of today). These schemes are usually characterized by low complexity and by using simple heuristic design rules. The capacity $\omega*$ achieved by these schemes is, as expected, often considerably lower that can be expected to be achieved by optimum channel assignment.

3. CURRENT APPROACHES TO RESOURCE ALLOCATION STRATEGIES

The subproblem that has attracted most of the interest in the literature so far, is the choice and allocation of waveforms. Orthogonal waveforms such as frequency division multiplexing (FDMA) and time division (TDMA), which provide a "channelization" of the spectrum, have no doubt been the most popular ones, although considerable interest has recently been devoted to nonorthogonal waveforms, such as the IS-95 DS-CDMA waveforms [25]. Given the set of signaling waveforms **C**, the next problem is the allocation of waveforms to the different terminal-access port links. This allocation can be done numerous ways depending on the amount and quality of the information available regarding the matrix **G** and the traffic situation (activity of different terminals). Another important issue is the time scale on which resource (re-)allocation is feasible.

Channel allocation in early FDMA cellular radio systems operates on a long-term basis. Based on average type statistical information regarding **G** (i.e., large scale propagation predictions), frequencies are on a more or less permanent basis assigned to different access ports. Such a "cell plan" provides a sufficient reuse distance between RAPs providing a reasonably low probability of outage (to low SIR) [1]. Inhomogenuities in the traffic load can also be taken care of by adapting the number of channels in each RAP to the expected traffic carried by that access port. To minimize the planning effort, adaptive cell planning strategies (e.g., "channel segregation" [2]) have been devised using long-term average measurements of the interference and traffic to automatically allocate channels to the access port. These "static" (or "quasi-static") channel allocation schemes work quite well when employed in macrocellular systems with high-traffic loads. In short range (microcellular) systems propagation conditions tend to change more abruptly. Since each of the RAPs tend to carry less total traffic in small microcells, the relative traffic variations are also large, particularly in multimedia traffic scenarios. Employing "static" channel allocation schemes in the situations require considerable design margins. Large path loss variations are countered with large reuse distances, unfortunately at a substantial capacity penalty. In the same way microcellular traffic variations are handled by assigning excess capacity to handle traffic peaks. In recent years two principally different methods to approach this problem have been devised: Dynamic channel allocation (DCA) and Random Channel Allocation (RCA).

In dynamic (real-time) channel allocation(DCA), real-time measurements of propagation and/or traffic conditions are used to (re-)allocate spectrum resources. Early graph theoretic schemes, adapted only to traffic variations [4,5] yielding only moderate capacity gains (<50%) compared to static systems in microcellular environments. Other schemes adapt their channel allocation to the received wanted signal strength. One example of the latter type of schemes is the class of reuse-partitioning schemes [6,8]. Here, several overlaid cell plans with different reuse distances are used. Terminals with a high received signal level are tolerant to interference and can be allocated a channel from a dense reuse cell plan, whereas the "weaker" terminals get channels with a large reuse distance and lower interference levels. Capacity gains in the order of up to 100% have been reported for these schemes [7]. Also, schemes directly estimating the C/I and thereby in a distributed way finding channels with adequate quality have been proposed [2,9]. Similar gains as in the reuse partitioning schemes are found in the literature. A comprehensive survey of different DCA-schemes is provided in Ref. 46.

The performance of the DCA schemes is critically dependent on the rate at which allocation or reallocation occurs. Purely traffic adaptive schemes act on incoming user requests and users releasing capacity. Channel reallocation has to occur at these rates to fully utilize the potential of such a DCA scheme. For speech traffic this means that reallocations typically occur at second rates.

Path loss and interference adaptive schemes that "track" (at least slow fading) signal level variations and reallocation rates in the 10's of millisecond range may be required. An alternative class of allocation schemes are the random channel allocation schemes. The principle is most easily explained using Fig. 5. Figure 5a shows a typical set of C/I-trajectories of five terminals in a cellular system. As we can see, 4 of the 5 terminals achieve an adequate C/I, corresponding to an (ensemble) outage rate of 20%. Compare this situation to the one in Fig. 5b exhibiting the same outage rate. In contrast to the situation in Fig. 5a where 20% of the terminals are experiencing too low C/I, here each terminal will experience insufficient quality 20% of the time. In Fig. 5a channel coding is a waste of capacity since four terminals have sufficient quality and the last unlucky terminal is probably "beyond salvage." In Fig. 5b however, there

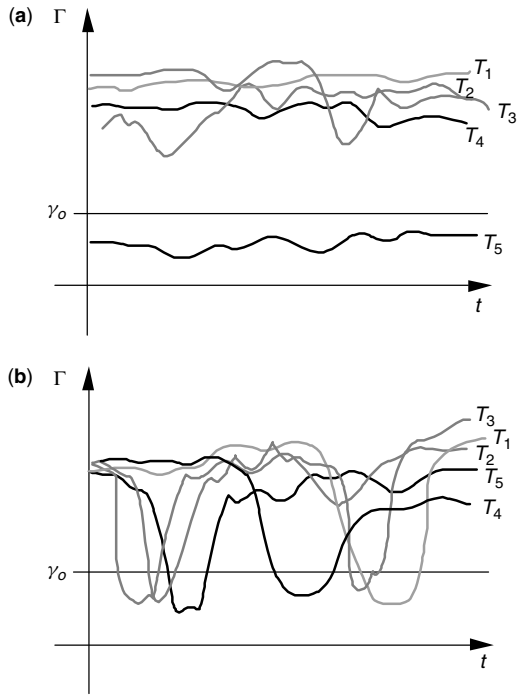


Figure 5. Typical realizations of terminal SIR:s in cellular systems with slowly moving terminals (**a**), and with rapidly moving terminals (**b**).

are probably a sufficient number of reliable channel symbols in all terminals to make reception possible, provided suitable constraint lengths and interleaving are used. The obvious way to achieve the latter situation, regardless of the mobile speed, is to permute channel allocations in a random fashion. The simplest way is to use (orthogonal) frequency hopping which can be seen as a static channel allocation where terminals allocated to a certain access port swap channels with each other [28,33]. Frequency hopping occurs typically 100–1000 times/second. Also nonorthogonal waveforms can be used as in the DS-CDMA based IS-95 scheme [25]. Effectively, a new random waveform is used for every transmitted bit. DS-CDMA schemes require only a very low level of synchronization and no cell planning, which has made them attractive. Regarding capacity the comparison between DS and FH schemes is not obvious although orthogonal schemes seem to have advantages in mixed cell environments [26]. Comparing the performance of (deterministic) DCA to the performance of the random allocation schemes is even more complex and stands out as one of the more fundamental research topics of the near future. Another quite different situation where similar interference conditions as in frequency hopping prevail are certain packet communication systems. Here the "randomness" is mainly induced by the random arrivals of packets triggering transmission events.

The selection of the proper transmitter power in terminals and access ports is another topic that has attracted considerable interest in recent years. There can be several objectives for this: to suppress adjacent channel (cross correlation) interference in nonorthogonal schemes, to minimize power consumption in order to extend terminal battery life, and to control cochannel interference (in schemes with orthogonal waveforms). In the resource allocation problem context, it can be shown that the maximum number of terminals is supported under a power control (PC) regime that balances the C/I of all terminals that can be supported and shuts off the rest [10].

Figure 6, showing the cumulative distribution function (CDF) of the received C/I in an (orthogonal signaling) cellular system under three power control regimes, illustrates why this is so. As we can see, the uncontrolled system exhibits a rather flat CDF with a high outage probability (at the threshold C/I). A received signal strength based algorithm, such as the constant received power scheme, reduces some of the variations in the C/I by limiting the variations in the "C" component. The variations in the interference part ("I") are however now larger than before and the figure shows a typical net result CDF. The outage probability is now slightly lower. In the C/I balancing scheme all stations have the same C/I, here slightly over the threshold, leaving only a small fraction of terminals without support. Finding this optimum set of nonsupported terminals is a problem closely related to the design of DCA schemes. Distributed implementations and different implementational constraints [11,12] have been studied. Results show that very robust near-optimum power control schemes can be devised at very low complexity. Performance results indicate that in static channel allocations substantial (>100%) capacity gains can be achieved using optimum power control. These gains are, of course, not additive with the gains obtained by DCA schemes. However, preliminary results regarding combined DCA/PC schemes show substantial capacity gains [13,15].

For packet communication with short messages or in frequency hopping environments, power control as described above may not work properly due to the fact that the feedback delay in the power control loop may in fact be longer than the time required to transmit the message (or in FH the chip/burst duration). The bursty interference caused by other users compounds to the problem of accurately measuring and predicting the C/I. Several approaches to this problem have been proposed. In systems utilizing mainly forward error correction, C/I-balancing power control strategies involving estimating statistical parameters of the C/I, such as the average C/I (measured over many packet/chip durations), and as

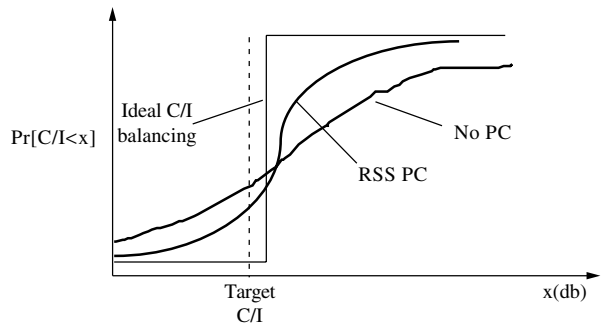


Figure 6. Outage probability minimization-CDF of received C/I.

well as fast C/I estimation/tracking schemes have been proposed [34–37].

In random access systems utilizing retransmissions (ARQ), however, the picture is quite different. Two important differences between these systems and the continuous ("circuit switched") systems can be noted. Providing equal transmission quality to all packet transmission (e.g., letting all packets be received with equal powers), turns out to be a disastrous strategy, since it guarantees that no packet can be received properly when several packets happen to be transmitted simultaneously (a "collision"). It has clearly been demonstrated that a received power spread, such as the one caused by near-far effects or Rayleigh/Shadow fading actually improves the capacity (throughput) of such systems [39]. Results show that power control, which creates an even larger power spread, can produce even better results [40–42].

4. MANAGING CHANGE — THE DYNAMICS OF RESOURCE ALLOCATION

As terminals move about in the service area, the propagation and interference situation may turn such that the terminal cannot be supported by the same access port on any waveform. New terminals may enter the service area requiring services, while others are terminating their communication sessions. As most of the basic resource allocation strategies described above deal mainly with static or quasi-static situations that are encountered on a microscopic, short term, time scale, we have to devise resource management schemes capable of handling these variations.

In the first case, where we have signal quality variations as a result of the terminal moving during a communication session some kind of resource reallocation may become necessary. This may be a waveform reallocation, or an "intra-port" handoff, which in principle involves a reexecution of the basic channel allocation scheme, or an inter-port ("inter-cell") handoff. In early cellular systems which are mainly noise-limited systems these handoffs were basically triggered by too low received signal levels. The handoff mechanism has, for these cases, often been modeled as a selection (macro-) diversity scheme where the terminals are assigned to the access port with the highest received signal level. This situation where hand-offs occur at more or less well-defined "cell-borders" has been extensively studied.

Maximizing the instantaneous received signal level may, however, not be neither very practical nor produce the best results. In high density wireless systems, the coverage areas of the access port overlap to a large extent. Low signal levels is rarely a problem since normally several access ports provide sufficient signal levels. In these cases the variations in the interference and not cell boundary crossings is the most probable cause of a handoff. When several access ports may provide sufficient C/I, the system is also able to handle traffic variations by means of *load sharing*, that is, by letting less loaded access ports support terminals even though they are providing less C/I than the best (often the closest) port [20]. Combinations of power control and access port selection also show promising results [21,22]. Instead of conventional handoff schemes ("switching diversity"), continuous combining schemes ("soft handoff") have been studied quite extensively [19].

Keeping track of the mobile terminals in a large (possibly global) wireless system, the mobility management, is a formidable task. Although this is handled mainly on the fixed network end, there are important implications to the resource management. The tradeoff between the capacity required for the air signaling to monitor the whereabouts of the terminals (the "locating" procedures) and the capacity required for finding, or paging, a terminal when a communication request comes from the network end, has received quite some attention [23,24] in CDMA schemes.

Handling arriving and departing terminals poses a slightly different problem. Whenever a new terminal arrives (a new request for service or an inter-port handoff) the RRM system has to decide if this particular terminal may be allowed into the system. An algorithm making these decisions is called an admission control algorithm. Since the exact terminal population and gain matrix may not be tracked exactly at all times and due to the complexity of the RRM-algorithms, determining the success of an admission decision may not be possible beforehand without physically executing the admission itself. The admission procedure may fail in two ways:

1. ("False admission") A terminal is admitted giving rise to a situation where one or more terminals cannot be supported (not necessarily including the admitted terminal).
2. ("False rejection") A terminal is rejected when successful resource allocation actually was possible.

Traditional approaches involving static channel allocation normally use simple thresholding strategies on the available channels in each cell. Access ports have been assigned a fixed set of channels that "guarantee" to provide a certain low-outage probability. Such a system is what we call "blocking limited," whenever a call arrives, we may check if there are channels available or not. Since we may choose to give priority to an ongoing session experiencing an inter-port handoff, new arriving calls are admitted only up to the point where there is only some small fraction of the resource remaining. This spare capacity is "reserved" for calls entering a cell due to an inter-cell handoff.

In systems using dynamic channel allocation or random allocation there is no clear limit on the number of channels/waveform that can be used. In such "interference-limited" systems, the feasibility of admitting new users will depend on the current interference situation. In particular in systems utilizing C/I-balancing power control this is complicated by the fact that already active terminals will react to the admission of a new terminal by adjusting (raising) their transmitter powers. It is therefore quite possible that the admission of yet another user may cause several of the original users (possibly all!) to no longer be supported at the required C/I-level. Admission control schemes can be grouped into noninteractive schemes and interactive schemes [43]. The noninteractive schemes proposed are mainly using

different types of interference or transmitter power thresholds [43], that is, when the measured interference (or the currently used power) on some channel (cell) is too high, admission is denied. The interactive schemes involve the gradual increase of the power of new terminals until they are finally admitted. Such a procedure to protect the already established procedure connection is referred to as "Soft-and Safe (SAS)" admission [44] or channel probing/active link protection [45].

5. MULTIPLE QUALITY-OF-SERVICE RESOURCE MANAGEMENT

Looking at the more general problem definition, most of the capacity definitions provided above fail if the users have different service requirements. In our problem definition this is reflected in the SIR requirements γ_i. Most of the techniques discussed above, properly generalized, lend themselves readily also for the nonspeech and multiservice cases. It has, however, to be understood that (for nonspeech services) the mapping of user perceived performance on technical parameters, such as the SIRs, in the wireless system is certainly a very complex task. With the proliferation of "best-effort"-type backbone networks, the user experience of the service provided is influenced not only by the shortcomings of the wireless system but also by other factors. As indicated by Fig. 7, such factors include the performance of the wireline backbone and switching, the service providers application software and hardware, and even the user interface provided by the service provider and the terminal manufacturer. In order to allow rational design of telecommunication systems, these overall (end-to-end) requirements are broken down to specific (sub-)service requirements for the individual building blocks. Here the interest is focused on studying the behavior of the radio network part of the "transport system." The "services" provided at this level have been coined bearer services in the UMTS/3G standardization process. These services have been divided into four classes as outlined in Table 1 [47] and are mainly distinguished by their delay requirements ranging from very strict delay requirements in "conversational class" (e.g., voice services) to the very relaxed requirements in the "background-best effort" class. In more technical terms, the services in the different classes are characterized by means of sets of service parameters, forming a QoS profile. Some of the QoS parameters (service attributes) in the 3G systems are found in Table 2.

In principle, an infinite set of QoS-profiles and thus different bearer services could be defined by varying these parameters — possibly one combination for each user. In practice, limitations on the number of modulation waveforms, codes, and so on, restrict the number of service

Table 1. 3G (UMTS) Bearer Service Classes [47]

Service Class	Typical Applications	Service Functional Characteristics
Conversational		
Real Time(RT) Streaming RT	Voice Video/Audio streams	• Preserve time relations between entities • Stringent preservation of conversational patterns (low delay)
Interactive		Preserve time relations between entities
Best effort (BE) Background BE	Web-browsing File transfer, E-mail	• Request-response pattern • Preserve payload (low error rate) • Not time critical • Preserve payload (low error rate)

offerings. Most systems will therefore offer a finite set of bearer services, where each parameter will be allowed to take one out of a few discrete values. Table 2 provides an indication of what ranges these service parameters can take in a 3G wireless system. In addition to these service parameters, the *availability* of the services has to be considered since it may vary over time and over the user services. Clearly, users with QoS-profiles with large "resource consumption" (e.g., high bit rate, poor location) or with low-relative priority will more often experience that the system is not capable of accommodating their service request.

Looking at a popular example to illustrate the problems we face, we take a web browsing session. Such a session consists of an irregular sequence of file transfers (using the TCP/IP protocol stack). Typical very short messages are transmitted in the uplink from the terminals (corresponding to a mouse-click) a random instants to request rather large files (web-pages, pictures, etc.) to be downloaded into the terminal. This can be seen as a service of the "interactive" class. The critical QoS characteristic to the user is the response-time (i.e., delay between request and the complete reception of the requested page). For large requested files the delay is dominated by the transfer delay of the files, that is, the average data rate is in fact the critical QoS parameter that will determine the user delay. The undetected error rate at the user level has to be below 10^{-8} corresponding to 1 error in about 10 MB. The radio bearer may however have a larger undetected error probability since the TCP/IP protocol provides end-to-end error control of its own. Classical models for data traffic of this type are based on Poisson distributions — both for

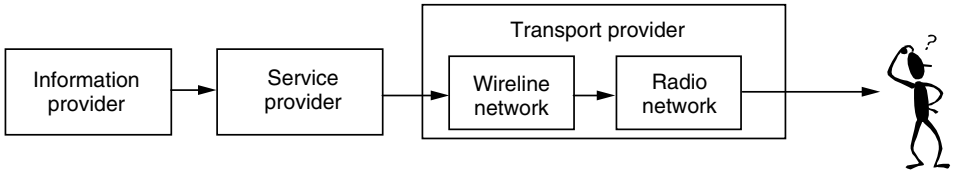


Figure 7. Service provisioning in modern information and communication systems.

Table 2. Some 3G (UMTS) Service Attribute/Parameter Ranges [47]

Traffic Class	Conversational	Streaming	Interactive	Background
Max bit rate(kbps)	<2000	<2000	<2000-overhead	<2000-overhead
Max SDU size(byte)	<1500	<1500	<1500	<1500
Guaranteed bit rate	<2000	<2000		
Transfer delay(ms)	80-max value	500-max value		
Priority	1,2,3	1,2,3	1,2,3	1,2,3
Residual BER	$5*10^{-2}$, 10^{-2}, 10^{-3}, 10^{-6}	$5*10^{-2}$, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}	$4*10^{-3}$,10^{-5}, $6*10^{-8}$	$4*10^{-3}$,10^{-5}, $6*10^{-8}$

the interarrival time of packets as well as the size of the requested files. Recent studies of Internet traffic have however shown that these models tend to underestimate the time between packets and the packet sizes. As an alternative, the Pareto distributions, that is, stochastic variables with density function

$$f(x) = \frac{\beta a^{\beta}}{x^{\beta+1}} a, \beta \geq 0, x \geq a$$

have been proposed [49]. These distributions are "heavy-tailed," they assign higher probability to large values of x compared to the exponential distributions found in Poisson point processes. Another approach for simulation purposes, is to use a more detailed model of the actual TCP/IP protocol [48].

After having discussed the user performance perspective, let us now briefly return to the discussion of the network performance. Defining a capacity performance measure for such a single service system is not obvious. Examples of popular measures (for non–real-time service systems as in the example above) found in the literature are the total throughput, that is, the sum data rate of all connections in the systems (possibly per area unit) and the user Circuit switched equivalent Bit Rate (CBR, i.e., the constant data rate would provide the same user perceived performance). In non–real-time, "best-effort," systems efficient resource management makes use of the extra degree of freedom provided by not fixing the delay. Efficient scheduling and adapting the transmission rate to the current propagation and SIR conditions may be used to maximize throughput. In a fading channel (as illustrated by Fig. 5b), the channel conditions may be bad for a certain user during one time frame. That particular user could thus postpone its transmission, allowing some other users with favorable data rates to transmit — potentially at a much higher data rate. Typically, such schemes maximize the total throughput by favoring terminals with good propagation conditions as they generally have more time slots at their disposal and can use higher data rates [52,53].

When mixing services, the situation becomes even more complex. In order for any capacity definition to be precise in the general (multiservice) case requires a model, not only for the number of users, but also for their behavior. What will be the QoS-profile requested for a certain user (which is then mapped to the individual γ_i), and what is the required service availability? Typically, random models will be used for this purpose. A user will, with some given probability, belong to a certain class of users with an identical QoS-profile. The probability distribution of class membership is usually referred to as the service mix.

Determining which service mix should be used for the capacity definition is indeed difficult. One approach has been to look at the pricing strategy and maximizing the revenue of the operator. This leads to both an optimal service mix and maximal revenue derived from the network. A difficulty is that this type of model disregards the fact that if the demand for the service provided is finite, operators are prone to competition from either other similar operators or alternative technical solutions. In these situations the pricing strategy clearly affects the demand for the services.

Given a certain service mix, several capacity definitions have been tried in the literature, but a generally acceptable single definition has still to be found. Looking at the total throughput or the total operator revenue is one approach. The main difficulty with this is to set the price parameters to reasonable values due to the interdependence of the pricing and the service mix. More promising is to use the (total) number of users satisfied with their respective service, this is one approach that leads to a capacity definition very close to the one discussed above [50,51]. In this approach, the capacity regions, that is, the permissible combinations of numbers of users in the different service classes play an important role in the cases where no price structure can be determined [50].

Another problem caused by mixing traffic, in particular circuit-switched and bursty best-effort traffic, is that it makes link-quality assessment more difficult. The main result from traffic theory tells us to dynamically share all the available spectrum for all types of traffic. A sideeffect of this is, however, that also the interference experienced by different users will exhibit the same wide span in character [38]. In particular if we would like to estimate the link quality for a high quality circuit switched service, the link will be subject to both quasi-constant as well as intermittent interference (from packet service users). Reliably estimating, for instance, using the C/I as a basis for RRM decisions, will be considerably more difficult.

6. DISCUSSION

We have previously presented a formulation of the radio resource problem based on the three basic allocation decisions: waveform, access port, and transmitter power. As the reader may have realized, these are closely related. Most of the recent work indicates that good results are achieved when these decisions are coordinated. Combinations of power control and DCA [13,15], base station assignment and power control [21,22], as well as power control assisted admission schemes, have provided

interesting results. Another area where research is just in its preliminary stages is the combination of detailed modulation waveforms/channel coding and its interaction with DCA and power control. Although the current mobile telephony systems are rather easily modeled in the terms described above, it seems clear that also most of the RRM problems expected in the future can be mapped onto the framework presented here. A key problem in bursty and mixed traffic is the tradeoff between maximizing instantaneous resource utilization (transmit only when data is available) and obtaining reliable quality measurements to facilitate the efficient adaptation of the radio resources to the needs of the users.

Traditionally, we consider the frequency spectrum to be the resource to be shared. Since there, in fact, does not exist any upper limit on the capacity that can be provided (with a dense enough infrastructure), it is important that we widen the resource management perspective. Parameters such as infrastructure density costs and terminal power consumption play important roles. One could easily identify tradeoffs such as where the signal processing load should be put in a wireless system — in the terminal where power is scarce or in the fixed infrastructure. The key question here is: Should the access port infrastructure be very dense (and costly) allowing for dumb, cheap, low-power terminals, or should terminals be more complex allowing for the rapid deployment of a cheap infrastructure at the expense of battery life and terminal cost ?

BIOGRAPHY

Jens Zander (S'82–M'85) received the M.S degree in electrical engineering (Y) and the Ph.D degree (in datatransmission) from Linköping University, Sweden, in 1979 and 1985, respectively.

From 1985 to 1989 he was a partner of SECTRA, a high-tech company in telecommunications systems and applications. In 1989 he was appointed professor and head of the Radio Communication Systems Laboratory at the Royal Institute of Technology, Stockholm, Sweden. Since 1992 he also serves as Senior Scientific Advisor to the Swedish National Defence Research Institute (FOI). He currently is the Scientific Director of the Center for Wireless Systems (Wireless@KTH) at the Royal Institute of Technology, Stockholm.

Dr. Zander has published numerous papers in the field of radio communication, in particular on resource management aspects of personal communication systems. He has also coauthored four textbooks on radio communication systems, including the English textbooks *Principles of Wireless Communications* and *Radio Resource Management for Wireless Networks*. He was the recipient of the IEEE Veh. Tech. Soc. Jack Neubauer Award for best systems paper in 1992.

Dr. Zander is a member of the Royal Academy of Engineering Sciences. He is the chairman of the IEEE VT/COM Swedish chapter. He is associate editor of the ACM *Wireless Networks* Journal and area editor of *Wireless Personal Communications*.

His current research interests include future wireless infrastructures, in particular related resource allocation and economic issues.

BIBLIOGRAPHY

1. W. C. Y. Lee, *Mobile Communication Fundamentals*, Wiley, New York, 1993.
2. Y. Furuya, Y. Akaiwa, Channel segregation — A distributed adaptive channel allocation scheme for mobile communication systems, *Proc DMR-II*, Stockholm, 1987.
3. R. Beck and H. Panzer, Strategies for handover and dynamic channel allocation in micro-cellular mobile radio systems, *IEEE Veh Tech Conf VTC89*, May 1989.
4. D. C. Cox and D. O. Reudink, Dynamic channel assignment in high capacity mobile communication systems, *Bell Syst Tech J.* **50**(6): (July–Aug. 1971).
5. D. Everitt and D. Manfield, Performance analysis of cellular communication systems with dynamic channel allocation, *IEEE Trans. Sel. Areas Comm.* **7**(8): (Oct. 1989).
6. S. W. Halpern, Reuse partitioning in cellular systems, *IEEE Veh. Tech. Conf. VTC85*, May 1985.
7. J. Zander and H. Eriksson, Asymptotic bounds on the Performance of a class of dynamic channel Assignment Algorithms, *IEEE Trans. Sel. Areas Comm.* **11**(3): (Aug. 1993).
8. T. Kanai, Autonomous reuse partitioning in cellular systems, *IEEE Veh. Tech. Conf. VTC92*, May 1992.
9. D. J. Goodman, S. A. Grandhi, and Vijayan, Distributed dynamic channel assignment schemes, *IEEE Veh. Tech. Conf. VTC93*, May 1993.
10. J. Zander, Performance of optimum transmitter power control in radio systems, *IEEE Trans. Veh. Tech.* **41**(1): (Feb. 1992).
11. G. J. Foscini and Z. Mijanec, A simple distributed power control algorithm and its convergence, *IEEE Trans. Veh. Tech.* **42**(4): (Nov. 1993).
12. S. A. Grandhi, J. Zander, and R. Yates, Constrained power control, *Wireless Personal Communications*, (Kluwer) **2**(3): (Aug. 1995).
13. G. J. Foscini and Z. Mijanec, Distributed autonomous wireless channel assignment algorithm with power control, *IEEE Trans. Veh. Tech.* **44**(4): (Nov. 1995).
14. M. Frodigh, Bounds on the performance of DCA-algorithms in highway micro cellular radio systems, *IEEE Trans. Veh. Tech.* **43**(3): (Aug. 1994).
15. M. Frodigh, Performance bounds for power control supported DCA-algorithms in highway micro cellular radio systems, *IEEE Trans. Veh. Tech.* **44**(2): (May 1995).
16. S. Tekinay and B. Jabbari, Handover and channel assignment in mobile cellular network, *IEEE Comm Mag.* **29**(11): (Nov. 1991).
17. M. Austin and G. Stüber, Cochannel interference modelling for signal-strength based handoff, *Electronics Letters* **30**: 1914–1915 (Nov. 1994).
18. N. Zhang and J. Holtzman, Analysis of handoff algorithms using both absolute and relative measurements, *IEEE 44th Veh Tech Conf VTC94*, June 1994.
19. N. Zhang and J. Holtzman, Analysis of a CDMA soft handoff algorithm, *Proc PIMRC 95*, Toronto, Sept. 1995.
20. B. Eklundh, Channel utilization and blocking probability in cellular mobile telephone systems with directed retry, *IEEE Trans. Comm.* **34**(4): (Apr. 1986).
21. R. Yates and C-Y. Huang, Integrated power control and base station assignment, *IEEE Trans. Veh. Tech.* **44**(4): (Nov. 1995).

22. S. V. Hanly, An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity, *IEEE Trans. Sel. Areas Comm.* **13**(7): (Sep. 1995).

23. J. G. Markoulidakis and E. D. Sykas, Model for location updating and handover rate estimation in mobile telecommunications, *Electronics Letters* **29**(17): (Aug. 1993).

24. G. Morales-Andes and Villen-Altamirano, An approach to modelling subscriber mobility in cellular radio networks, *Forum Telecom 87*, Geneva, 1987.

25. A. Salmasi and K. S. Gilhousen, On the system design aspects of code division multiple access(CDMA) and personal communication networks, *IEEE 42ndVeh Tech Conf VTC92*, May 1992.

26. H. Eriksson, G. Gudmundson, J. Sköld, and J. K. Ugland et al., Multiple access options for cellular based personal communications, *IEEE 43rdVeh Tech Conf VTC93*, May 1993.

27. E. Anderlind, Resource allocation for heterogenous traffic in a wireless network, *Int. Symp. on Personal, Indoor and Mobile Radio Comm. PIMRC 95*, Toronto, Sept. 1995.

28. G. Gudmundson, J. Sköld, and J. K. Ugland, A comparison between CDMA and TDMA systems, *IEEE 42ndVeh Tech Conf VTC92*, May 1992.

29. A. Acampora, Wireless ATM: A perspective on issues and prospects, *IEEE Personal Comm. Mag.* (Aug. 1996).

30. C. Robool and C. Roobol, Message delay in 1-D indoor packet radio systems with diversity reception, *IEEE First Symposium on Communications and Vehicular Technology in the Benelux*, Delft, The Netherlands Oct. 1993.

31. F. Borgonovo, Zorzi & Acampora, Capture division packet access, *IEEE Comm. Mag.* **34**(9): (Sept. 1996).

32. G.Q. Maguire, B. Ottersten, H. Tenhunen, and J. Zander, Future wireless computing & communication, *Nordisk Radioseminarium*, NRS-94, Linköping, Sweden, Oct. 1994.

33. H. Olofsson, Interference diversity as means for increasing capacity in GSM, *Proc EPMCC'95*, Bologna, Italy, Nov. 1995.

34. Z. Rosberg, Fast power control in cellular networks based on short term correlation of Rayleigh fading, *Proc 6th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ, March 1997.

35. Z. Rosberg and J. Zander, Power control in wireless networks with random interferers, *Internal Report*, Radio Communication lab, Royal Institute of Technology, (Dec. 1995) (to be published, *http://www.s3.kth.se/s3/radio/PUBLICATIONS/documents.html)*.

36. M. Andersin and Z. Rosberg, Time-variant power control, *Proc PIMRC-96*, Taipeh Oct. 1996.

37. D. Mitra and J. A. Morrisson, A distributed power control algorithm for bursty transmissions in cellular CDMA networks, *Proc 5th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ, 1995.

38. D. Mitra and J. A. Morrisson, A novel distributed Power Control Algorithm for classes of service in cellular spread spectrum wireless networks, *Proc 6th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ March 1997.

39. J. A. Arnbak and W. van Blitterswijk, Capacity of slotted ALOHA in Rayleigh fading channels, *IEEE J Sel. Areas Commun.* **SAC-5**(2): (Feb. 1987).

40. J. J. Metzner, On improving utilization in ALOHA Networks, *IEEE Trans. Comm.* **COM-24**: (Apr. 1976).

41. C. Leung and V. Wong, A transmit power control scheme for improving performance in a mobile packet radio system, *IEEE Trans. Veh. Tech.* **VT-43**(1): (Feb 1994).

42. C. Roobol, On the Packet Delay in wireless local area networks with access port diversity and power control, *Proc PIMRC-95*, Toronto Sept. 1995.

43. M. Andersin, *Power Control and Admission Control in Cellular Radio Systems dissertation*, Ph.D., Radio Comm Systems Lab, Royal Institute of Technology, June 1996.

44. M. Andersin, Z. Rosberg, and J. Zander, Soft admission in cellular PCS with constrained power control and noise, *Proc 5th WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ, 1995.

45. N. Bambos, S. C. Chen, and D. Mitra, Channel probing for distributed access in wireless communication networks, *Proc GLOBECOM '95*, Singapore Nov. 1995.

46. I. Katzela and M. Naghshneh, Channel allocation schemes for cellular mobile telecommunication systems: A comprehensive survey, *IEEE Personal Commun.* 10–31 (June 1996).

47. 3GPP Specification, TS23.107 Dec. 1999.

48. E. Anderlind and J. Zander, A traffic model for non real-time data users in a wireless radio network, *IEEE Comm. Letters* **1**(2): (Mar. 1997).

49. V. Paxson and S. Floyd, Wide Area Traffic:The Failure of Poisson Modeling, *IEEE/ACM Trans. Networking* **3**(3): (June 1995).

50. A. Furuskär, P. de Bruin, C. Johansson, and A. Simonsson, Managing mixed services with controlled QoS in GERAN—The GSM/EDGE Radio Access Network, *IEE 3G Conference on Mobile Communication Technologies* 147–151, 2001.

51. A. Furuskär, P. de Bruin, C. Johansson, and A. Simonsson, Mixed Service Management with QoS Control for GERAN—The GSM/EDGE Radio Access Network, *IEEE Vehicular Technology Conference 2001* Spring, May 2001.

52. J. Zander, Performance Bounds for Joint Power Control & Link Adaption for NRT bearers in Centralized (Bunched) Wireless Network *PIMRC 99*, Osaka, Japan, Sept. 1999.

53. F. Berggren, S-L Kim, R. Jäntti, and J. Zander, Joint Power Control and Intracell Scheduling of DS-CDMA Nonreal Time Data, *IEEE J. Sel. Areas Commun.* **19**(10): 1860–1870 (2001).

54. J. Zander and S.-L. kim, *Resource Management in Wireless Networks*, Artech House, 2001.

ROUTING AND WAVELENGTH ASSIGNMENT IN OPTICAL WDM NETWORKS

GEORGE N. ROUSKAS
North Carolina State University
Raleigh, North Carolina

1. INTRODUCTION TO OPTICAL WDM NETWORKS

A basic property of a single-mode optical fiber is its enormous low-loss bandwidth of several tens of terahertz. However, because of dispersive effects and limitations in optical device technology, single-channel transmission is limited to only a small fraction of the fiber capacity. To take full advantage of the potential of fiber, the use of wavelength-division multiplexing (WDM) technology

has become the option of choice. With WDM, a number of distinct wavelengths are used to implement separate channels [1]. An optical fiber can carry several channels in parallel, each on a particular wavelength. The number of wavelengths that each fiber can carry simultaneously is limited by the physical characteristics of the fiber and the state of the optical technology used to combine these wavelengths onto the fiber and isolate them off the fiber. With currently available commercial technology, a few tens of wavelengths can be supported within the low-loss window at 1550 nm, but this number is expected to grow rapidly in the near future. Therefore, optical fiber links employing WDM technology have the potential of delivering an aggregate throughput in the order of terabits per second, enough to satisfy the ever-growing demand for more bandwidth per user on a sustained, long-term basis.

Unfortunately, because of the mismatch between aggregate fiber capacity and peak electronic processing speeds, simply upgrading existing point-to-point fiber links to WDM creates the well-known *electrooptic bottleneck* [2]: rather than achieving the multiterabit-per-second throughput of the fiber, one has to settle for the multigigabit-per-second throughput that can be expected of the electronic devices where the optical signals terminate. Overcoming the electrooptic bottleneck, therefore, involves the design of properly structured architectures to interconnect the fiber links. An optical WDM network is a network with optical fiber transmission links and with an architecture that is designed to exploit the unique features of fibers and WDM. Such networks offer the promise of an all-optical information highway capable of supporting a wide range of applications that involve the transport of massive amounts of data and/or require very fast response times. Such applications include video on demand and teleconferencing, telemedicine applications, multimedia document distribution, remote supercomputer visualization, and many more to come. Consequently, optical WDM networks have been a subject of extensive research both theoretically and experimentally [3,4].

The architecture for wide-area WDM networks that is widely expected to form the basis for a future all-optical infrastructure is built on the concept of *wavelength routing*. A wavelength routing network, shown in Fig. 1, consists of two types of nodes: *optical cross-connects* (OXCs), which connect the fibers in the network, and *edge nodes*, which provide the interface between non-optical end systems (such as IP routers, ATM switches, or supercomputers) and the optical core. Access nodes provide the terminating points (sources and destinations) for the optical signal paths; the communication paths may continue outside the optical part of the network in electrical form.

The services that a wavelength routed network offers to end systems attached to edge nodes are in the form of *logical* connections implemented using *lightpaths*. Lightpaths (also referred to as λ-*channels*), are clear optical paths between two edge nodes, and are shown in Fig. 1 as red and green directed lines. Information transmitted on a lightpath does not undergo any conversion to and from electrical form within the optical network, and thus, the architecture of the optical network nodes can be very simple because they do not need to do any signal processing. Furthermore, since a lightpath behaves as a literally transparent "clear channel" between the source and destination edge node, there is nothing in the signal path to limit the throughput of the fibers.

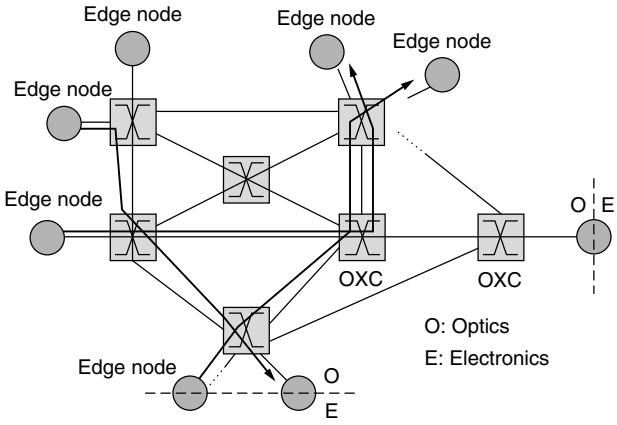


Figure 1. A wavelength-routed WDM network.

The OXCs provide the switching and routing functions for supporting the logical connections between edge nodes. An OXC takes in an optical signal at each wavelength at an input port, and can switch it to a particular output port, independent of the other wavelengths. An OXC with N input and N output ports capable of handling W wavelengths per port can be thought of as W independent $N \times N$ switches. These switches have to be preceded by a wavelength demultiplexer and followed by a wavelength multiplexer to implement an OXC, as shown in Fig. 2. Thus, an OXC can cross-connect the different wavelengths from the input to the output, where the connection pattern of each wavelength is independent of the others. By appropriately configuring the OXCs along the physical path, a logical connection (lightpath) may be established between any pair of edge nodes.

A unique feature of optical WDM networks is the tight coupling between routing and wavelength selection. As can be seen in Fig. 1, a lightpath is implemented by selecting a path of physical links between the source and destination edge nodes, and reserving a particular wavelength on each of these links for the lightpath. Thus, in establishing an optical connection we must deal with both routing (selecting a suitable path) and wavelength assignment (allocating an available wavelength for the connection). The resulting problem is referred to as the *routing and wavelength assignment* (RWA) problem [5], and is significantly more difficult than the routing problem in electronic networks. The additional complexity arises from the fact that routing and wavelength assignment are subject to the following two constraints:

1. *Wavelength continuity constraint* — a lightpath must use the same wavelength on all the links along its path from source to destination edge node. This constraint is illustrated in Fig. 1 by representing

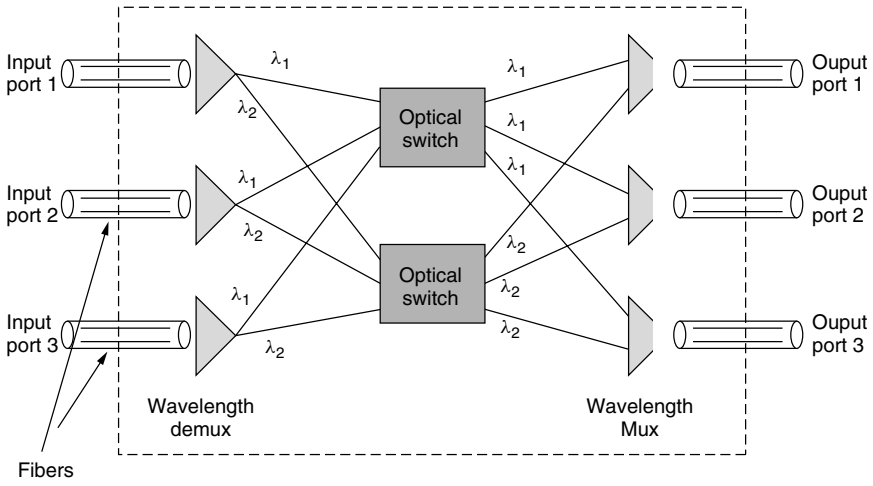


Figure 2. A 3×3 optical cross-connect (OXC) with two wavelengths per fiber.

each lightpath with a single color (wavelength) along all the links in its path.

2. *Distinct wavelength constraint* — all lightpaths using the same link (fiber) must be allocated distinct wavelengths. In Fig. 1 this constraint is satisfied since the two lightpaths sharing a link are shown in different colors (wavelengths).

The RWA problem in optical networks is illustrated in Fig. 3, where it is assumed that each fiber supports two wavelengths. The effect of the wavelength continuity constraint is represented by replicating the network into as many copies as the number of wavelengths (in this case, two). If wavelength i is selected for a lightpath, the source and destination edge node communicate over the ith copy of the network. Thus, finding a path for a connection may potentially involve solving W routing problems for a network with W wavelengths, one for each copy of the network.

The wavelength continuity constraint may be relaxed if the OXCs are equipped with *wavelength converters* [6]. A wavelength converter is a single input/output device that converts the wavelength of an optical signal arriving at its input port to a different wavelength as the signal departs from its output port, but otherwise leaves the optical signal unchanged. In OXCs without a wavelength conversion capability, an incoming signal at port p_i on wavelength λ can be optically switched to any port p_j, but must leave the OXC on the same wavelength λ. With wavelength converters, this signal could be optically switched to any port p_j on some other wavelength λ'. Thus, wavelength conversion allows a lightpath to use different wavelengths along different physical links.

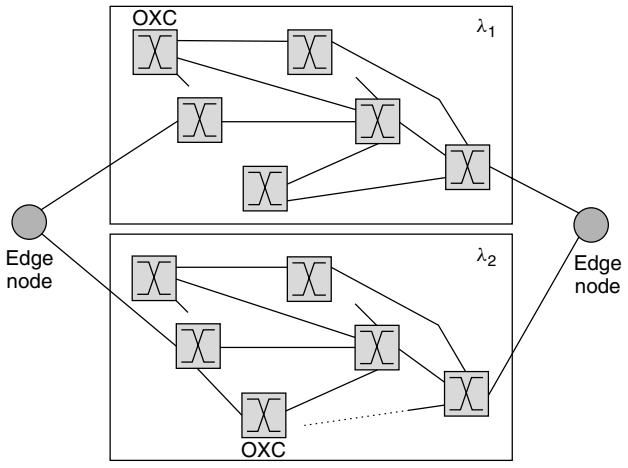


Figure 3. The RWA problem with two wavelengths per fiber.

Different levels of wavelength conversion capability are possible. Figure 4 illustrates the differences for single-input and single-output port situations; the case for multiple ports is more complicated but similar. *Full wavelength conversion* capability implies that any input wavelength may be converted to any other wavelength. *Limited wavelength conversion* [7] denotes that each input wavelength may be converted to any of a specific set of wavelengths, which is not the set of all wavelengths for at least one input wavelength. A special case of this is *fixed wavelength conversion*, where each input wavelength can be converted to exactly one other wavelength. If each wavelength is "converted" only to itself, then we have no conversion.

The advantage of full wavelength conversion is that it removes the wavelength continuity constraint, making it possible to establish a lightpath as long as each link along the path from source to destination has a free wavelength (which could be different for different links). As a result, the RWA problem reduces to the classical routing problem, that is, finding a suitable path for each connection in the

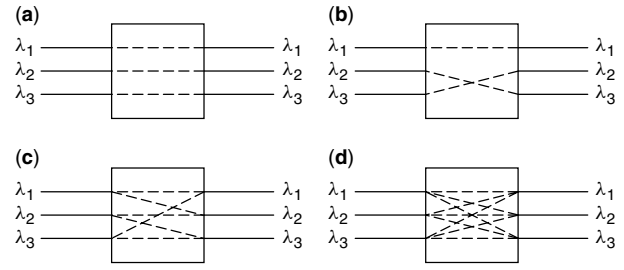


Figure 4. Wavelength conversion: (**a**) no conversion; (**b**) fixed conversion; (**c**) limited conversion; (**d**) full conversion.

network. Referring to Fig. 3, full wavelength conversion collapses the W copies of the network into a single copy on which the routing problem is solved. On the other hand, with limited conversion, the RWA problem becomes more complex than with no conversion. To see why, note that employing limited conversion at the OXCs introduces links between *some* of the network copies of Fig. 3. For example, if wavelength λ_1 can be converted to wavelength λ_2 but not to wavelength λ_3, then links must be introduced from each OXC in copy 1 of the network to the corresponding OXC in copy 2, but not to the corresponding OXC in copy 3. When selecting a path for the connection, at each OXC there is the option of remaining at the same network copy or moving to another one, depending on the conversion capability of the OXC. Since the number of alternatives increases exponentially with the number of OXCs that need to be traversed, the complexity of the RWA problem increases accordingly.

Wavelength conversion (full or limited) increases the routing choices for a given lightpath (i.e., makes more efficient use of wavelengths), resulting in better performance. Since converter devices increase network cost, a possible middle ground is to use *sparse conversion*, that is, to employ converters in some, but not all, OXCs in the network. In this case, a lightpath must use the same wavelength along each link in a segment of its path between OXCs equipped with converters, but it may use a different wavelength along the links of another such segment. It has been shown that implementing full conversion at a relatively small fraction of the OXCs in the network is sufficient to achieve almost all the benefits of conversion [8,9].

Routing and wavelength assignment is the fundamental control problem in optical WDM networks. Since the performance of a network depends not only on its physical resources (OXCs, converters, fibers links, number of wavelengths per fiber, etc.) but also on how it is controlled, the objective of an RWA algorithm is to achieve the best possible performance within the limits of physical constraints. The RWA problem can be cast in numerous forms. The different variants of the problem, however, can be classified under one of two broad versions: a static RWA, whereby the traffic requirements are known in advance, and a dynamic RWA, in which a sequence of lightpath requests arrive in some random fashion. Sections 2 and 3 discuss the static and dynamic versions, respectively, of the RWA problem, and present some algorithms to solve them. Finally, Section 4 presents the multicast RWA problem and algorithms to build *light trees* that connect a source edge node to multiple destinations.

2. STATIC ROUTING AND WAVELENGTH ASSIGNMENT

If the traffic patterns in the network are reasonably well known in advance and any traffic variations take place over long timescales, the most effective technique for establishing optical connections (lightpaths) between edge nodes is by formulating and solving a static RWA problem. For example, static RWA is appropriate for provisioning a set of semipermanent connections. Since these connections are assumed to remain in place for relatively long periods of time, it is worthwhile to attempt to optimize the way in which network resources (e.g., physical links and wavelengths) are assigned to each connection, even though optimization may require a considerable computational effort.

A solution to the static RWA problem consists of a set of long-lived lightpaths that create a *logical* (or *virtual*) topology among the edge nodes. This virtual topology is embedded onto the physical topology of optical fiber links and OXCs. Accordingly, the static RWA problem is often referred to as the *virtual topology design* problem [10]. In the virtual topology, there is a directed link from edge node s to edge node d if a lightpath originating at s and terminating at d is set up (refer also to Fig. 1), and edge node s is said to be "one hop away" from edge node d in the virtual topology, although the two nodes may be separated by a number of physical links. The type of virtual topology that can be created is usually constrained by the underlying physical topology. In particular, it is generally not possible to implement fully connected virtual topologies: for N edge nodes this would require each edge node to maintain $N-1$ lightpaths and the optical network to support a total of $N(N-1)$ lightpaths. Even for modest values of N, this degree of connectivity is beyond the reach of current optical technology, in terms of both the number of wavelengths that can be supported and in the optical hardware (transmitters and receivers) required at each edge node.

In its most general form, the RWA problem is specified by providing the physical topology of the network and the traffic requirements. The physical topology corresponds to the deployment of cables in some existing fiber infrastructure, and is given as a graph $G_p(V, E_p)$, where V is the set of OXCs and E_p is the set of fibers that interconnect them. The traffic requirements are specified in a traffic matrix $\mathbf{T} = [\rho p_{sd}]$, where ρp_{sd} is a measure of the long-term traffic flowing from source edge node s to destination edge node d [11]. Quantity ρ represents the (deterministic) total offered load to the network, while the p_{sd} parameters define the distribution of the offered traffic.

Routing and wavelength assignment are considered together as an optimization problem using mixed-integer programming (MIP) formulations. Usually, the objective of the formulation is to minimize the maximum congestion level in the network subject to network resource constraints [10,12]. While other objective functions are possible, such as minimizing the average weighted number of hops or minimizing the average packet delay, minimizing network congestion is preferable since it can lead to mixed-integer linear programming (MILP) formulations. While we do not present the RWA problem formulation here, the interested reader may refer to the literature [11–12]. These formulations turn out to have extremely large numbers of variables, and are intractable for large networks. This fact has motivated the development of heuristic approaches for finding good solutions efficiently.

Before we describe the various heuristic approaches, we note that the static RWA problem can be logically decomposed into four subproblems. The decomposition is approximate or inexact, in the sense that solving the subproblems in sequence and combining the solutions may

not result in the optimal solution for the fully integrated problem, or some later subproblem may have no solution given the solution obtained for an earlier subproblem, so no solution to the original problem may be obtained. However, the decomposition provides insight into the structure of the RWA problem and is a first step towards the design of effective heuristics. Assuming no wavelength conversion, the subproblems are as follows:

1. *Topology subproblem* — determine the logical topology to be imposed on the physical topology; that is, determine the lightpaths in terms of their source and destination edge nodes.
2. *Lightpath routing subproblem* — determine the physical links which each lightpath consists of; that is, route the lightpaths over the physical topology.
3. *Wavelength assignment subproblem* — determine the wavelength each lightpath uses; that is, assign a wavelength to each lightpath in the logical topology so that wavelength restrictions are obeyed for each physical link.
4. *Traffic routing subproblem* — route packet traffic between source and destination edge nodes over the logical topology obtained.

A large number of heuristic algorithms have been developed in the literature to solve the general static RWA problem discussed here or its many variants. Overall, however, the different heuristics can be classified into three broad categories: (1) algorithms that solve the overall MILP problem suboptimally, (2) algorithms that tackle only a subset of the four subproblems, and (3) algorithms that address the problem of embedding regular logical topologies onto the physical topology.

Suboptimal solutions can be obtained by applying conventional tools developed for complex optimization problems directly to the MILP problem. One technique is to use LP relaxation followed by rounding [13]. In this case, the integer constraints are relaxed creating a nonintegral problem which can be solved by some linear programming method, and then a rounding algorithm is applied to obtain a new solution that obeys the integer constraints. Alternatively, genetic algorithms or simulated annealing [14] can be applied to obtain locally optimal solutions. The main drawback of these approaches is that it is difficult to control the quality of the final solution for large networks: simulated annealing is computationally expensive and thus, it may not be possible to adequately explore the state space, while LP relaxation may lead to solutions from which it is difficult to apply rounding algorithms.

Another class of algorithms tackles the RWA problem by initially solving the first three subproblems listed above; traffic routing is then performed by employing well-known routing algorithms on the logical topology. One approach for solving the three subproblems is to maximize the amount of traffic that is carried on one-hop lightpaths, where traffic that is routed from source to destination edge node directly on a lightpath. A greedy approach taken is to create lightpaths between edge nodes in order of decreasing traffic demands as long as the wavelength continuity and distinct wavelength constraints are satisfied [15]. This algorithm starts with a logical topology with no links (lightpaths) and sequentially adds lightpaths as long as doing so does not violate any of the problem constraints. The reverse approach is also possible [16]; starting with a fully connected logical topology, an algorithm sequentially removes the lightpath carrying the smallest traffic flows until no constraint is violated. At each step (i.e., after removing a lightpath), the traffic routing subproblem is solved in order to find the lightpath with the smallest flow.

The third approach to RWA is to start with a given logical topology, thus avoiding the need to directly solve the first of the four subproblems listed above. Regular topologies are good candidates as logical topologies since they are well understood and results regarding bounds and averages (e.g., for hop lengths) are easier to derive. Algorithms for routing traffic on a regular topology are usually simple, so the traffic routing subproblem can be trivially solved. Also, regular topologies possess inherent load balancing characteristics which are important when the objective is to minimize the maximum congestion.

Once a regular topology is decided on as the one to implement the logical topology, it remains to decide which physical node will realize each given node in the regular topology (this is usually referred to as the *node mapping* subproblem), and which sequence of physical links will be used to realize each given edge (lightpath) in the regular topology (this *path mapping* subproblem is equivalent to the lightpath routing and wavelength assignment subproblems discussed earlier). This procedure is usually referred to as *embedding* a regular topology in the physical topology. Both the node and path mapping subproblems are intractable, and heuristics have been proposed in the literature [16,17]. For instance, a heuristic for mapping the nodes of shuffle topologies based on the gradient algorithm has been developed [17].

Given that all the algorithms for the RWA problem are based on heuristics, it is important to be able to characterize the quality of the solutions obtained. To this end, one must resort to comparing the solutions to known bounds on the optimal solution. A comprehensive discussion of bounds for the RWA problem and the theoretical considerations involved in deriving them can be found in [10]. A simulation-based comparison of the relative performance of the three classes of heuristic for the RWA problem has been presented [12]. The results indicate that the second class of algorithms discussed earlier achieve the best performance.

3. DYNAMIC ROUTING AND WAVELENGTH ASSIGNMENT

Under a dynamic traffic scenario, edge nodes submit to the network requests for lightpaths to be set up as needed. Thus, connection requests are initiated in some random fashion. Depending on the state of the network at the time of a request, the available resources may or may not be sufficient to establish a lightpath between the corresponding source–destination edge node pair. The network state consists of the physical path (route) and

wavelength assignment for all active lightpaths. The state evolves randomly in time as new lightpaths are admitted and existing lightpaths are released. Thus, each time a request is made, an algorithm must be executed in real time to determine whether it is feasible to accommodate the request, and, if so, to perform routing and wavelength assignment. If a request for a lightpath cannot be accepted because of lack of resources, it is blocked.

Because of the real-time nature of the problem, RWA algorithms in a dynamic traffic environment must be very simple. Since combined routing and wavelength assignment is a hard problem, a typical approach to designing efficient algorithms is to decouple the problem into two separate subproblems: the routing problem and the wavelength assignment problem. Consequently, most dynamic RWA algorithms for wavelength routed networks consist of the following general steps:

1. Compute a number of candidate physical paths for each source–destination edge node pair and arrange them in a path list.
2. Order all wavelengths in a wavelength list.
3. Starting with the path and wavelength at the top of the corresponding list, search for a feasible path and wavelength for the requested lightpath.

The specific nature of a dynamic RWA algorithm is determined by the number of candidate paths and how they are computed, the order in which paths and wavelengths are listed, and the order in which the path and wavelength lists are accessed.

Let us first discuss the routing subproblem. If a *static* algorithm is used, the paths are computed and ordered independently of the network state. With an *adaptive* algorithm, on the other hand, the paths computed and their order may vary according to the current state of the network. A static algorithm is executed offline and the computed paths are stored for later use, resulting in low latency during lightpath establishment. Adaptive algorithms are executed at the time when a lightpath request arrives and require network nodes to exchange information regarding the network state. Lightpath setup delay may also increase, but in general adaptive algorithms improve network performance.

The number of path choices for establishing an optical connection is another important parameter. A *fixed* routing algorithm is a static algorithm in which every source–destination edge node pair is assigned a single path. With this scheme, a connection is blocked if there is no wavelength available on the designated path at the time of the request. In *fixed–alternate* routing, a number k, $k > 1$, of paths are computed and ordered off-line for each source–destination edge node pair. When a request arrives, these paths are examined in the specified order, and the first one with a free wavelength is used to establish the lightpath. The request is blocked if no wavelength is available in any of the k paths. Similarly, an adaptive routing algorithm may compute a single path, or a number of alternate paths at the time of the request. A hybrid approach is to compute k paths offline, however, the order in which the paths are considered is determined according to the network state at the time the connection request is made (e.g., least to most congested).

In most practical cases, the candidate paths for a request are considered in increasing order of pathlength. *Pathlength* is typically defined as the sum of the weights assigned to each physical link along the path, and the weights are chosen according to some desirable routing criterion. Since weights can be assigned arbitrarily, they offer a wide range of possibilities for selecting path priorities. For example, in a static (fixed–alternate) routing algorithm, the weight of each link could be set to 1, or to the physical distance of the link. In the former case, the path list consists of the k minimum-hop paths, while in the latter the candidate paths are the k minimum-distance paths (where *distance* is defined as the geographic length). In an adaptive routing algorithm, link weights may reflect the load or "interference" on a link (i.e., the number of active lightpaths sharing the link). By assigning small weights to least loaded links, paths with larger number of free channels on their links rise to the head of the path list, resulting in a *least-loaded* routing algorithm. Paths that are congested become "longer" and are moved further down the list; this tends to avoid heavily loaded bottleneck links. Many other weighting functions are possible.

When pathlengths are sums of link weights, the k shortest path algorithm [18] can be used to compute candidate paths. Each path is checked in order of increasing length, and the first that is feasible is assigned the first free wavelength in the wavelength list. However, the k shortest paths constructed by this algorithm usually share links. Therefore, if one path in the list is not feasible, it is likely that other paths in the list with which it shares a link will also be infeasible. To reduce the risk of blocking, the k shortest paths can be computed so as to be pairwise link-disjoint. This can be accomplished as follows. When computing the ith shortest path, $i = 1, \ldots, k$, the links used by the first $i - 1$ paths are removed from the original network topology and Dijkstra's shortest path algorithm [19] is applied to the resulting topology. This approach increases the chances of finding a feasible path for a connection request.

Let us now discuss the wavelength assignment subproblem, which is concerned with the manner in which the wavelength list is ordered. For a given candidate path, wavelengths are considered in the order in which they appear in the list to find a free wavelength for the connection request. Again, we distinguish between the static and adaptive cases. In the *static* case, the wavelength ordering is fixed (e.g., the list is ordered by wavelength number). The idea behind this scheme, also referred to as *first-fit*, is to pack all the in-use wavelengths toward the top of the list so that wavelengths toward the end of the list will have higher probability of being available over long continuous paths. In the *adaptive* case, the ordering of wavelengths is typically based on usage. Usage can be defined either as the number of links in the network in which a wavelength is currently used, or as the number of active connections using a wavelength. Under the *maximum-reuse* method, the most used wavelengths are considered first (i.e., wavelength are considered in order of decreasing usage). The rationale

behind this method is to reuse active wavelengths as much as possible before trying others, packing connections into fewer wavelengths and conserving the spare capacity of less used wavelengths. This is turn makes it more likely to find wavelengths that satisfy the continuity requirement over long paths. Under the *minimum-reuse* method, wavelengths are tried in the order of increasing usage. This scheme attempts to balance the load as equally as possible among all the available wavelengths. However, minimum-reuse assignment tends to "fragment" the availability of wavelengths, making it less likely that the same wavelength is available throughout the network for connections that traverse longer paths.

Both reuse schemes introduce communication overhead because they require global network information in order to compute the usage of each wavelength. The first-fit scheme, on the other hand, requires no global information, and since it does not need to order wavelengths in real time, it has significantly lower computational requirements than either maximum reuse or minimum reuse. Another adaptive scheme that avoids the communication and computational overhead of maximum reuse and minimum reuse is *random* wavelength assignment. With this scheme, the set of wavelengths that are free on a particular path is first determined. Among the available wavelengths, one is chosen randomly (usually with uniform probability) and assigned to the requested lightpath.

We note that in networks in which all OXCs are capable of wavelength conversion, the wavelength assignment problem is trivial; since a lightpath can be established as long as at least one wavelength is free at each link and different wavelengths can be used in different links, the order in which wavelengths are assigned is not important. On the other hand, when only a fraction of the OXCs employ converters (i.e., a sparse conversion scenario), a wavelength assignment scheme is again required to select a wavelength for each segment of a connection's path that originates and terminates at an OXC with converters. In this case, the same assignment policies discussed above for selecting a wavelength for the end-to-end path can also be used to select a wavelength for each path segment between OXCs with converters.

The performance of a dynamic RWA algorithm is generally measured in terms of the *call blocking probability*, that is, the probability that a lightpath cannot be established in the network because of lack of resources (e.g., link capacity or free wavelengths). Even in the case of simple network topologies (such as rings) or simple routing rules (such as fixed routing), the calculation of blocking probabilities in WDM networks is extremely difficult. In networks with arbitrary mesh topologies, and/or when using alternate or adaptive routing algorithms, the problem is even more complex. These complications arise from both the link load dependencies (due to interfering lightpaths) and the dependencies among the sets of active wavelengths in adjacent links (due to the wavelength continuity constraint). Nevertheless, the problem of computing blocking probabilities in wavelength-routed networks has been extensively studied in the literature, and approximate analytical techniques that capture the effects of link load and wavelength dependencies have been developed in [8,9,20]. A detailed comparison of the performance of various wavelength assignment schemes in terms of call blocking probability can be found in Zhu et al. [21].

Although important, average blocking probability (computed over all connection requests) does not always capture the full effect of a particular dynamic RWA algorithm on other aspects of network behavior, in particular, fairness. In this context, *fairness* refers to the variability in blocking probability experienced by lightpath requests between the various edge node pairs, such that lower variability is associated with a higher degree of fairness. In general, any network has the property that longer paths are likely to experience higher blocking than shorter ones. Consequently, the degree of fairness can be quantified by defining the *unfairness factor* as the ratio of the blocking probability on the longest path to that on the shortest path for a given RWA algorithm. Depending on the network topology and the RWA algorithm, this property may have a cascading effect that can result in an unfair treatment of the connections between more distant edge node pairs — blocking of long lightpaths leaves more resources available for short lightpaths, so that the connections established in the network tend to be short ones. These shorter connections "fragment" the availability of wavelengths, and thus the problem of unfairness is more pronounced in networks without converters since finding long paths that satisfy the wavelength continuity constraint is more difficult than without this constraint.

Several studies [8,9,20] have examined the influence of various parameters on blocking probability and fairness, and some of the general conclusions include the following:

- Wavelength conversion significantly affects fairness. Networks employing converters at all OXCs sometimes exhibit orders of magnitude improvement in fairness (as reflected by the unfairness factor) compared to networks with no conversion capability, despite the fact that the improvement in overall blocking probability is significantly less pronounced. It has also been shown that equipping a relatively small fraction (typically, 20–30%) of all OXCs with converters is sufficient to achieve most of the fairness benefits due to wavelength conversion.
- Alternate routing can significantly improve the network performance in terms of both overall blocking probability and fairness. In fact, having as few as three alternate paths for each connection may in some cases (depending on the network topology) achieve almost all the benefits (in terms of blocking and fairness) of having full wavelength conversion at each OXC with fixed routing.
- Wavelength assignment policies also play an important role, especially in terms of fairness. The random and minimum-reuse schemes tend to "fragment" the wavelength availability, resulting in large unfairness factors (with minimum-reuse having the worst performance). On the other hand, the maximum-reuse assignment policy achieves the best performance in terms of fairness. The first-fit scheme exhibits a

behavior very similar to maximum-reuse in terms of both fairness and overall blocking probability, and has the additional advantage of being easier and less expensive to implement.

4. MULTICAST ROUTING AND WAVELENGTH ASSIGNMENT

In Sections 2 and 3, we considered static and dynamic RWA algorithms, respectively, for establishing lightpaths in optical networks. In Ref. 22, the concept of a lightpath was generalized into that of a *light tree*, which, like a lightpath, is a clear channel originating at given source node and implemented with a single wavelength. But unlike a lightpath, a light tree has multiple destination nodes, hence it is a point-to-multipoint channel. The physical links implementing a light tree form a tree, rooted at the source node, rather than a path in the physical topology, hence the name. That study [22] focused on virtual topology design (i.e., static RWA) for point-to-point traffic and observed that, since a light tree is a more general representation of a lightpath, the set of virtual topologies that can be implemented using light trees is a superset of the virtual topologies that can be implemented only using lightpaths. Thus, for any given virtual topology problem, an optimal solution using light trees is guaranteed to be at least as good and possibly an improvement over the optimal solution obtained using only lightpaths. Furthermore, it was demonstrated that by extending the lightpath concept to a light tree, the network performance (in terms of average packet hops) can be improved while the network cost (in terms of the number of optical transmitters and receivers required) decreases.

Light trees are implemented by employing optical devices known as *power splitters* [2] at the OXCs. A power splitter has the ability to split an incoming signal, arriving at some wavelength λ, into up to m outgoing signals, $m \geq 2$; m is referred to as the *fanout* of the power splitter. Each of these m signals is then independently switched to a different output port of the OXC. Note that, because of the splitting operation and associated losses, the optical signals resulting from the splitting of the original incoming signal must be amplified before leaving the OXC. Also, to ensure the quality of each outgoing signal, the fanout m of the power splitter may have to be limited to a small number. If the OXC is also capable of wavelength conversion, each of the m outgoing signal may be shifted, independently of the others, to a wavelength different than the incoming wavelength λ. Otherwise, all m outgoing signals must be on the same wavelength λ.

While the authors of Ref. 22 considered mainly point-to-point traffic, another attractive feature of light trees is the inherent capability for performing multicasting in the optical domain (as opposed to performing multicasting at a higher layer, e.g., the network layer, which requires electrooptic conversion). Such wavelength-routed light trees are useful for transporting high-bandwidth, real-time applications such as high-definition TV (HDTV). Therefore, OXCs equipped with power splitters will be referred to as *multicast-capable* OXCs (MC-OXCs). Note that, just like with converter devices, incorporating power splitters within an OXC is expected to increase the network cost because of the large amount of power amplification and the difficulty of fabrication.

With the availability of MC-OXCs and the existence of multicast traffic demands, the problem of establishing light trees to satisfy these demands arises. We will call this problem the *multicast routing and wavelength assignment* (MC-RWA) problem. MC-RWA bears many similarities to the RWA problem discussed in Sections 2 and 3. Specifically, the tight coupling between routing and wavelength assignment remains, and even becomes stronger; in the absence of wavelength conversion the same wavelength must be used by the multicast connection not only along the links of a single path but also along the links of the whole light tree. Since the construction of optimal trees for routing multicast connections is by itself a hard problem [23], the combined MC-RWA problem becomes even harder. Depending on the nature of traffic demands, we also distinguish between static and dynamic MC-RWA problems. As we already know, optimal solutions for the point-to-point RWA problems are not practically obtainable, and with a more general construct (the light tree) and hence a much larger search space, this will be even more true for the MC-RWA problems. In general, the approaches to tackling the static and dynamic MC-RWA problems are similar to the ones we described for the static and dynamic RWA problems, respectively. The challenge in this case is to design heuristics that can cope with the increased complexity of the problem and yet produce good solutions. In the following we summarize the most recent work on multicasting in optical networks, but the reader should keep in mind that this is an area of current research.

The benefits of multicasting in wavelength routed optical networks were first demonstrated by Malli et al. [24]. Specifically, it was shown that using light trees (spanning the source and destination nodes) rather than individual parallel lightpaths (each connecting the source to an individual destination) requires fewer wavelengths and consumes a significantly lower amount of bandwidth. In another paper [25] both the static and the dynamic MC-RWA problems were studied. A MILP formulation that maximizes the total number of multicast connections was presented for the static MC-RWA problem. Rather than providing heuristic algorithms for solving the MILP, bounds on the objective function were presented by relaxing the integer constraints. The dynamic MC-RWA problem, on the other hand, was solved by decoupling the routing and wavelength assignment problems. A number of *alternate* trees were constructed for each multicast connection using existing routing algorithms. When a request for a connection arrives, the associated trees are considered in a fixed order. For each tree, wavelengths are also considered in a fixed order (i.e., the first-fit strategy). The connection is blocked if no free wavelength is found in any of the trees associated with the multicast connection.

Finally, the problem of constructing trees for routing multicast connections was studied [26] independently of wavelength assignment, under the assumption that not all OXCs are multicast capable, that is, that there is a limited number of MC-OXCs in the network. Four new algorithms were developed for routing multicast

connections under this *sparse light splitting* scenario. While the algorithms differ slightly from each other, the main idea to accommodate sparse splitting is to start with the assumption that all OXCs in the network are multicast capable and use an existing algorithm to build an initial tree. Such a tree is infeasible if a non-multicast-capable OXC is a branching point. In this case, all but one branches out of this OXC are removed, and destination nodes in the removed branches have to join the tree at a MC-OXC.

BIOGRAPHY

George N. Rouskas received his degree in electrical engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1989, and M.S. and Ph.D. degrees in computer science from the College of Computing, Georgia Institute of Technology, Atlanta, Georgia, in 1991 and 1994, respectively. He joined the Department of Computer Science, North Carolina State University, Raleigh, North Carolina, in August 1994, where he is currently an associate professor. During the 2000–2001 academic year he spent a sabbatical term at Vitesse Semiconductor, Morrisville, North Carolina, and in May and June 2000 he was an invited professor at the University of Evry, Evry, France. He is a recipient of a 1997 NSF Faculty Early Career Development (CAREER) Award, and a coauthor of a paper that received the Best Paper Award at the 1998 SPIE conference on all-optical networking. He also received the 1995 Outstanding New Teacher Award from the Department of Computer Science, North Carolina State University, and the 1994 Graduate Research Assistant Award from the College of Computing, Georgia Tech. He was a coguest editor for the *IEEE Journal on Selected Areas in Communications*, Special Issue on Protocols and Architectures for Next Generation Optical WDM Networks, October, 2000, and is on the editorial boards of the *IEEE/ACM Transactions on Networking, Computer Networks, and Optical Networks*. Dr. Rouskas' interests include network architectures and protocols, optical networks, multicast communication, and performance evaluation.

BIBLIOGRAPHY

1. T. E. Stern and K. Bala, *Multiwavelength Optical Networks*, Prentice-Hall, Upper Saddle River, NJ, 2000.
2. B. Mukherjee, *Optical Communication Networking*, McGraw-Hill, New York, 1997.
3. O. Gerstel et al., eds., Special issue on protocols and architectures for next generation optical WDM networks, *IEEE J. Select. Areas Commun.* **18**(10) (Oct. 2000).
4. G.-K. Chung, K.-I. Sato, and D. K. Hunter, eds., Special issue on optical networks, *J. Lightwave Technol.* **18**(12) (Dec. 2000).
5. H. Zang, J. P. Jue, and B. Mukherjee, A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks, *Opt. Networks* **1**(1): 47–60 (Jan. 2000).
6. B. Ramamurty and B. Mukherjee, Wavelength conversion in WDM networking, *IEEE J. Select. Areas Commun.* **16**(7): 1061–1073 (Sept. 1998).
7. V. Sharma and E. A. Varvarigos, Limited wavelength translation in all-optical WDM mesh networks, *Proc. INFOCOM'98 IEEE* 893–901 (March 1999).
8. Y. Zhu, G. N. Rouskas, and H. G. Perros, A path decomposition algorithm for computing blocking probabilities in wavelength routing networks, *IEEE/ACM Trans. Network.* **8**(6): 747–762 (Dec. 2000).
9. S. Subramaniam, M. Azizoglu, and A. Somani, All-optical networks with sparse wavelength conversion, *IEEE/ACM Trans. Network.* **4**(4): 544–557 (Aug. 1996).
10. R. Dutta and G. N. Rouskas, A survey of virtual topology design algorithms for wavelength routed optical networks, *Opt. Networks Mag.* **1**(1): 73–89 (Jan. 2000).
11. R. Ramaswami and K. N. Sivarajan, Design of logical topologies for wavelength-routed optical networks, *IEEE J. Select. Areas Commun.* **14**(5): 840–851 (June 1996).
12. E. Leonardi, M. Mellia, and M. A. Marsan, Algorithms for the logical topology design in WDM all-optical networks, *Opt. Networks* **1**(1): 35–46 (Jan. 2000).
13. D. Banerjee and B. Mukherjee, A practical approach for routing and wavelength assignment in large wavelength-routed optical networks, *IEEE J. Select. Areas Commun.* **14**(5): 903–908 (June 1996).
14. B. Mukherjee et al., Some principles for designing a wide-area WDM optical network, *IEEE/ACM Trans. Network.* **4**(5): 684–696 (Oct. 1996).
15. Z. Zhang and A. Acampora, A heuristic wavelength assignment algorithm for multihop WDM networks with wavelength routing and wavelength reuse, *IEEE/ACM Trans. Network.* **3**(3): 281–288 (June 1995).
16. I. Chlamtac, A. Ganz, and G. Karmi, Lightnets: Topologies for high-speed optical networks, *J. Lightwave Technol.* **11**: 951–961 (May/June 1993).
17. S. Banerjee and B. Mukherjee, Algorithms for optimized node placement in shufflenet-based multihop lightwave networks, *Proc. INFOCOM'93 IEEE*, March 1993.
18. E. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart, Winston, 1976.
19. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
20. E. Karasan and E. Ayanoglu, Effects of wavelength routing and selection algorithms on wavelength conversion gain in WDM optical networks, *IEEE/ACM Trans. Network.* **6**(2): 186–196 (April 1998).
21. Y. Zhu, G. N. Rouskas, and H. G. Perros, A comparison of allocation policies in wavelength routing networks, *Photon. Network Commun.* **2**(3): 265–293 (Aug. 2000).
22. L. H. Sahasrabuddhe and B. Mukherje, Light-trees: Optical multicasting for improved performance in wavelength-routed networks, *IEEE Commun.* **37**(2): 67–73 (Feb. 1999).
23. S. L. Hakimi, Steiner's problem in graphs and its implications, *Networks* **1**: 113–133 (1971).
24. R. Malli, X. Zhang, and C. Qiao, Benefit of multicasting in all-optical networks, *Proc. SPIE* **3531**: 209–220 (Nov. 1998).
25. G. Sahin and M. Azizoglu, Multicast routing and wavelength assignment in wide-area networks, *Proc. SPIE* **3531**: 196–208 (Nov. 1998).
26. X. Zhang, J. Y. Wei, and C. Qiao, Constrained multicast routing in WDM networks with sparse light splitting, *J. Lightwave Technol.* **18**(12): 1917–1927 (Dec. 2000).

S

SAMPLING OF ANALOG SIGNALS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Many communication signals that are transmitted from a source to a destination are analog signals. Examples of such signals are speech, images, and video. Since the middle of the twentieth century, the trend has been to convert such analog signals to digital form and to transmit the digital signal using a digital modulation technique. At the receiver, the digital signal can be converted back to an analog signal for the user.

An analog signal is converted to a digital signal through the process of sampling the analog signal periodically and quantizing the samples to obtain a digital signal (a sequence of binary digits), which is then transmitted by a digital communication system. The sampling and quantization processes are generally performed by an analog-to-digital (A/D) converter, whose basic functions are illustrated in Fig. 1. Thus, conceptually, we may view A/D conversion of an analog signal as a three-step process:

1. *Sampling*. This is the conversion of a continuous-time signal $x_a(t)$ into a discrete-time signal $x_a(nT) = x(n)$, where $\{x_a(nT)\}$ are samples of $x_a(t)$ taken at times $t = nT$, and where T is the *sampling interval* and n takes integer values.
2. *Quantization*. This is the conversion of the continuous-valued signal samples $\{x(n)\}$ into discrete-valued signal samples $\{x_q(n)\}$, where the values $\{x_q(n)\}$ are selected from a finite set of possible values. The difference $x(n) - x_q(n)$ is called the *quantization error*.
3. *Coding*. This is the process of representing each quantized value $x_q(n)$ by a b-bit binary sequence.

Although we model the A/D conversion process as shown in Fig. 1, in practice the A/D conversion is performed by a single device whose input is the analog signal $x_a(t)$ and whose output is a sequence of b-bit values representing the quantized samples $x_q(n)$.

2. SAMPLING OF ANALOG SIGNALS

As indicated above, an analog signal $x_a(t)$ that is sampled periodically at $t = nT$ results in the sampled sequence

$$x(n) \equiv x_a(nT), \quad -\infty \le n \le \infty \tag{1}$$

The sampling process is illustrated in Fig. 2. The time interval T between successive samples is called the *sampling interval*, and its reciprocal $1/T$ is called the *sampling frequency*, denoted as

$$F_s = \frac{1}{T} \tag{2}$$

The choice of the sampling frequency is governed by the frequency content of the analog signal $x_a(t)$. To establish this basic relationship, let us consider the sampling of the sinusoidal signal

$$x_a(t) = A \cos 2\pi F t \tag{3}$$

Its sampled values are

$$\begin{aligned} x(n) \equiv x_a(nT) &= A \cos 2\pi F n T \\ &= A \cos 2\pi n \frac{F}{F_s} \end{aligned} \tag{4}$$

We note that $x(n)$ is a discrete-time sinusoidal signal having a normalized frequency $f = F/F_s$. We also observe that discrete-time sinusoids whose normalized frequencies are separated by an integer multiple of unity (or 2π radians) are identical. That is

$$x(n) = A \cos 2\pi n(f + k) = A \cos 2\pi n f \tag{5}$$

where k is any integer. We further observe that the highest rate of oscillation in a discrete-time sinusoid is attained

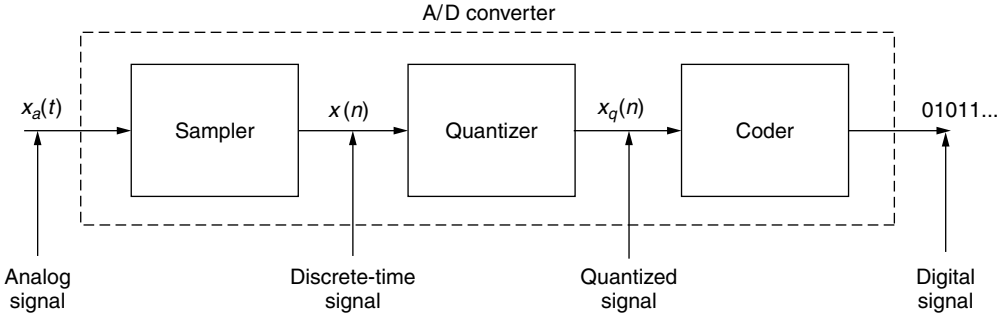


Figure 1. Elements of an analog-to-digital (A/D) converter.

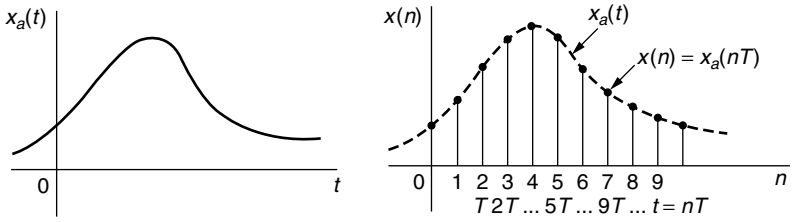


Figure 2. Periodic sampling of an analog signal.

when the normalized frequency $f = \pm\frac{1}{2}$. Any discrete-time sinusoid with frequency $f > \frac{1}{2}$ takes values that are identical to a discrete-time sinusoid whose frequency is contained in the *fundamental range* $-\frac{1}{2} \le f \le \frac{1}{2}$.

The implications of these observations can be fully appreciated by considering the sampling of the two analog sinusoidal signals

$$\begin{aligned} x_1(t) &= \cos 2\pi(10)t \\ x_2(t) &= \cos 2\pi(50)t \end{aligned} \tag{6}$$

at a rate $F_s = 40$ Hz (samples per second). The corresponding discrete-time signals are

$$\begin{aligned} x_1(n) &= \cos 2\pi \left(\frac{10}{40}\right) n = \cos \frac{\pi}{2} n \\ x_2(n) &= \cos 2\pi \left(\frac{50}{40}\right) n = \cos \frac{5\pi}{2} n = \cos \frac{\pi}{2} n \end{aligned} \tag{7}$$

Hence, $x_1(n) = x_2(n)$, so the two sampled signals are identical and, consequently, indistinguishable.

If we are given the sampled values obtained from $\cos \pi n/2$, there is ambiguity as to whether these sampled values correspond to $x_1(t)$ or $x_2(t)$. Since $x_2(t)$ yields exactly the same values as $x_1(t)$ when sampled at $F_s = 40$ samples per second, we say that the frequency $F_2 = 50$ Hz is an alias of the frequency $F_1 = 10$ Hz at the sampling rate of $F_s = 40$ Hz. We also note that F_2 is not the only alias of F_1. At the sampling rate of $F_s = 40$ Hz, the frequencies $F_3 = 90$ Hz, $F_4 = 130$ Hz, and so on, are all aliases of $F_1 = 10$ Hz; that is, the sinusoids $\cos\ 2\pi(F_1 + 40k)t$, $k = 1, 2, \ldots$ sampled at $F_s = 40$ Hz yield identical values.

The preceding observations lead to the conclusion that the highest frequency that can be represented uniquely by sampling an analog signal $x_a(t)$ at a rate F_s is $F_{\max} = F_s/2$. Hence, to avoid aliasing, an analog signal that is to be sampled must be band-limited. This is usually accomplished in practice by prefiltering the analog signal prior to sampling. Thus, the frequency content of the analog signal will be confined to a well-defined frequency band with highest frequency $F_{\max}$. Such a band-limited signal is then sampled at a rate $F_s \ge 2F_{\max}$, so as to avoid aliasing. The critical rate $2F_{\max} = F_N$ is called the *Nyquist sampling rate*.

For example, speech signals that are transmitted over telephone channels are limited to approximately 3200 Hz: $F_{\max} = 3200$ Hz. In this case, the Nyquist sampling rate is $F_N = 6400$ Hz. Such signals are typically sampled at a nominal rate of $F_s = 8000$ Hz. If the speech signal is to be transmitted by pulse code modulation (PCM), for example, each sample is typically (quantized) represented as a 7-bit binary word.

3. FREQUENCY-DOMAIN RELATIONSHIPS

If $x_a(t)$ is an aperiodic signal with finite energy, its (voltage) spectrum $X_a(f)$ is related to $x_a(t)$ by the inverse Fourier transform:

$$x_a(t) = \int_{-\infty}^{\infty} X_a(F) e^{j2\pi Ft} dF \tag{8}$$

The sampled signal sequence is

$$\begin{aligned} x(n) \equiv x_a(nT) &= \int_{-\infty}^{\infty} X_a(F) e^{j2\pi FnT} dF \\ &= \int_{-\infty}^{\infty} X_a(F) e^{j2\pi fF/F_s} dF \end{aligned} \tag{9}$$

The integration range of this integral can be subdivided into an infinite number of intervals of width F_s. Thus, we obtain

$$\begin{aligned} x(n) &= \sum_{k=-\infty}^{\infty} \int_{(k-1/2)F_s}^{(k+1/2)F_s} X_a(F) e^{j2\pi nF/F_s} dF \\ &= \sum_{k=-\infty}^{\infty} \int_{-F_s/2}^{F_s/2} X_a(F - kF_s) e^{j2\pi nF/F_s} dF \\ &= \int_{-F_s/2}^{F_s/2} \left[\sum_{k=-\infty}^{\infty} X_a(F - kF_s) \right] e^{j2\pi nF/F_s} dF \end{aligned} \tag{10}$$

Changing variables in Eq. (10) from F to the normalized frequency $f = F/F_s$ yields

$$x(n) = \left[\int_{-1/2}^{1/2} F_s \sum_{k=-\infty}^{\infty} X_a[(f - k)F_s] \right] e^{j2\pi nf}\, df \tag{11}$$

This equation is simply the inverse Fourier transform that relates the discrete-time signal $x(n)$ to its spectrum $X(f)$. Hence, the spectrum $X(f)$ of the discrete-time signal $x(n)$ is related to the spectrum $X_a(F)$ of the analog signal $x_a(t)$ by the expression

$$X(f) = F_s \sum_{k=-\infty}^{\infty} X_a[f - k)F_s] \tag{12}$$

or, equivalently

$$X\left(\frac{F}{F_s}\right) = F_s \sum_{k=-\infty}^{\infty} X_a(F - kF_s) \tag{13}$$

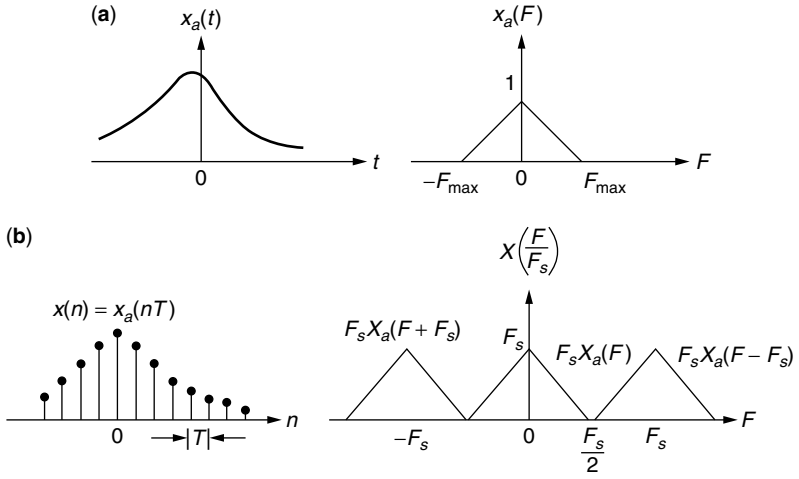


Figure 3. Time-domain and frequency-domain relationships between $x_a(t)$ and $x(n)$.

Figure 3 illustrates the time-domain and frequency-domain relationships between the analog signal $x_a(t)$ and its discrete-time sampled version. We observe that the spectrum of the discrete-time signal $x(n)$ is periodic with period F_s. With $F_s \geq 2F_{\max}$, the spectrum of the discrete-time signal within the fundamental range $|F| \leq F_s/2$ is simply

$$X\left(\frac{F}{F_s}\right) = F_s X_a(F) \tag{14}$$

Hence, the spectrum of the sampled signal is identical (within the scale factor F_s) to the spectrum of the analog signal. This implies that we can reconstruct the analog signal $x_a(t)$ from its samples $x(n)$.

If the sampling frequency F_s is selected such that $F_s < 2F_{\max}$, the periodic continuation of $X_a(f)$, given by Eq. (13) results in spectral overlap. Hence, the spectrum of the sampled signal $x(n)$ contains aliased frequency components of the analog signal spectrum $X_a(F)$. As a consequence, the spectrum of the discrete-time signal $x(n)$ is no longer equal to $F_s X_a(F)$ in the fundamental frequency range $|F| \leq F_s/2$. Therefore, we are unable to reconstruct the analog signal $x_a(t)$ from its samples $x(n)$.

4. THE SAMPLING THEOREM

Given the discrete-time signal sequence $x(n)$ with its spectrum $X(F/F_s)$, as illustrated in Fig. 3, with no aliasing, it is possible to reconstruct the original analog signal $x_a(t)$. This can be accomplished by passing the discrete-time sequence $x(n)$ through an ideal lowpass filter with frequency response

$$G(F) = \begin{cases} T, & |F| \leq \dfrac{F_s}{2} \\ 0, & |F| > \dfrac{F_s}{2} \end{cases} \tag{15}$$

The input to this filter may be expressed as

$$v(t) = \sum_{n=-\infty}^{\infty} x_a(nT)\delta(t - nT) \tag{16}$$

where $T = 1/F_s$ and $\delta(t)$ represents a unit impulse. With $F_s = F_N = 2F_{\max}$, the output of the ideal lowpass filter is

$$x_a(t) = \sum_{n=-\infty}^{\infty} x_a(nT)\frac{\sin(\pi/T)(t - nT)}{(\pi/T)(t - nT)} \tag{17}$$

This equation provides the formula for the ideal reconstruction of the analog signal $x_a(t)$ from its samples $x_a(nT)$. We observe that the ideal reconstruction involves the interpolation function

$$g(t) = \frac{\sin(\pi t/T)}{\pi t/T} \tag{18}$$

and its time-shifted versions (time shifts of nT, $n = \pm 1, \pm 2, \ldots$) which are multiplied by the corresponding samples $x_a(nT)$. We also observe that the interpolation function $g(t - nT)$ is zero at $t = kT$, except at $k = n$. Consequently, $x_a(t)$ evaluated at $t = kT$ is simply the sample $x_a(kT)$. At all other time instants, the weighted sum of the time-shifted versions of the interpolation function combine to yield $x_a(t)$.

The reconstruction formula given by Eq. (17) forms the basis for the *sampling theorem*, which can be stated as follows.

Sampling Theorem. A band-limited continuous-time signal $x_a(t)$ with highest-frequency (bandwidth) $F_{\max}$ can be uniquely recovered from the samples $x_a(nT)$, provided the sampling rate $F_s \geq 2F_{\max}$ samples per second.

5. SAMPLING OF BANDPASS SIGNALS

Suppose that a real-valued analog signal $x(t)$ has a frequency content concentrated in a narrow band of frequencies in the vicinity of a frequency F_c, as shown in Fig. 4. The bandwidth of the signal is defined as $B = B_2 - B_1$ and, usually, $F_c \gg B$. Such a signal is usually called a (narrowband) bandpass signal. Clearly, the highest frequency contained in this signal is $F_{\max} = B_2$. If we blindly apply the sampling principle described in the previous sections, we would sample such a signal at a rate

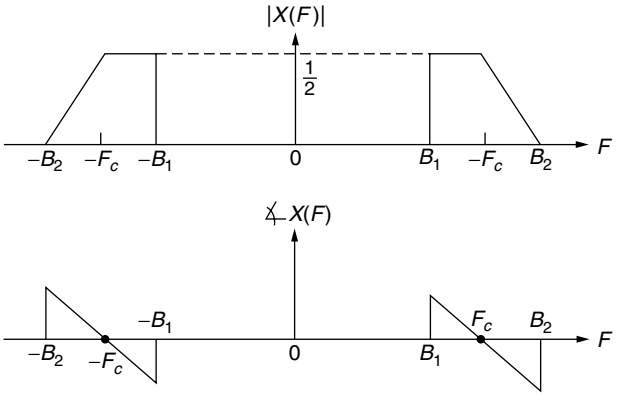


Figure 4. Spectrum of a bandpass signal.

$F_s \geq 2B_2$. However, it is not necessary to sample $x(t)$ at such a high rate.

It is easily shown that any (narrowband) bandpass signal (see, e.g., Ref. 1) can be represented by an equivalent lowpass signal. This representation may be expressed as

$$\begin{aligned} x(t) &= u_c(t)\cos 2\pi f_c t - u_s(t)\sin 2\pi F_c t \\ &= \mathrm{Re}\{[u_c(t) + ju_s(t)]e^{j2\pi F_c t}\} \end{aligned} \tag{19}$$

where $u_c(t)$ and $u_s(t)$ are called the *quadrature components* of the bandpass signal. The complex-valued signal

$$x_l(t) = u_c(t) + ju_s(t) \tag{20}$$

is called the *equivalent lowpass signal*. Its spectrum $X_l(F)$ is illustrated in Fig. 5, and it corresponds to the spectrum obtained by a frequency translation of F_c of the bandpass signal spectrum $X(F)$, shown in Fig. 4. We observe that the spectrum of the bandpass signal is related to the spectrum of the equivalent lowpass signal by the formula

$$X(f) = \tfrac{1}{2}[X_l(f - F_c) + X_l^*(-F - F_c)] \tag{21}$$

Hence, the spectrum of the bandpass signal $x(t)$ can be obtained from the spectrum of the equivalent lowpass signal $x_l(t)$ by a frequency translation.

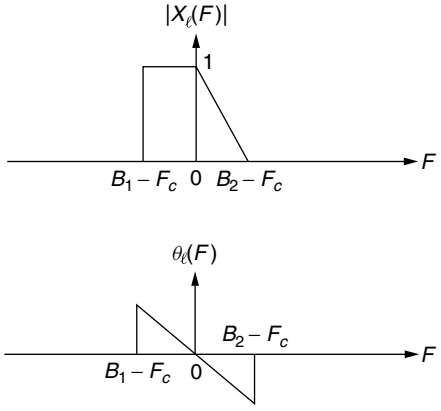


Figure 5. Spectrum of equivalent lowpass signal.

Given the equivalence between the bandpass signal $x(t)$ and the lowpass signal $x_l(t)$, it is advantageous to perform a frequency translation of the bandpass signal by an amount

$$F_c = \frac{B_1 + B_2}{2} \tag{22}$$

and sampling the equivalent lowpass signal. Such a frequency shift can be achieved by multiplying the bandpass signal as given in Eq. (19) by the quadrature carriers $\cos 2\pi F_c t$ and $\sin 2\pi F_c t$ and lowpass-filtering the products to eliminate the signal components at $2F_c$. Clearly, the multiplication and the subsequent filtering are first performed in the analog domain, and then the outputs of the filters are sampled. The resulting equivalent lowpass signal has a bandwidth $B/2$, where $B = B_2 - B_1$. Therefore, it can be represented uniquely by samples taken at the rate of B samples per second for each quadrature component. Thus the sampling can be performed on each lowpass filter output at the rate of B samples per second, as indicated in Fig. 6. Therefore, the resulting rate is $2B$ samples per second.

In view of the fact that frequency conversion to lowpass allows us to reduce the sampling rate to $2B$ samples per second, it should be possible to sample the bandpass signal at a comparable rate. In fact, it is.

Suppose that the upper frequency $F_c + B/2$ is a multiple of the bandwidth B (i.e., $F_c + B/2 = kB$), where k is a positive integer. If we sample $x(t)$ at the rate $2B = 1/T$ samples per second, we have

$$\begin{aligned} x(nT) &= u_c(nT)\cos 2\pi F_c nT - u_s(nT)\sin 2\pi F_c nT \\ &= u_c(nT)\cos\frac{\pi n(2k-1)}{2} - u_s(nT)\sin\frac{\pi n(2k-1)}{2} \end{aligned} \tag{23}$$

where the last step is obtained by substituting $F_c = kB - B/2$ and $T = 1/2B$.

For n even, say, $n = 2m$, Eq. (23) reduces to

$$x(2mT) \equiv x(mT_1)\cos\pi m(2k-1) = (-1)^m u_c(mT_1) \tag{24}$$

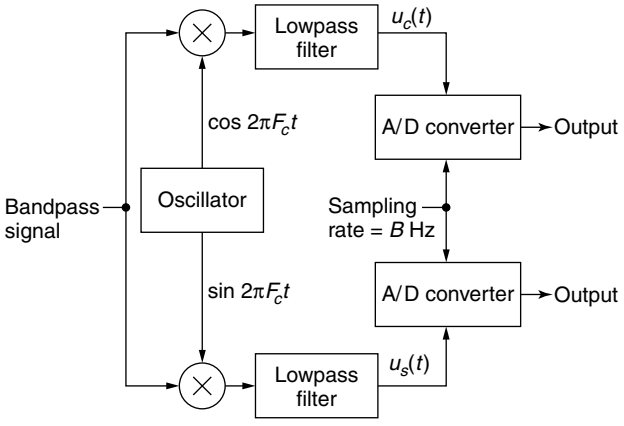


Figure 6. Sampling of a bandpass signal after converting it to a lowpass signal.

where $T_1 = 2T = 1/B$. For n odd, say, $n = 2m - 1$, Eq. (23) reduces to

$$x(2mT - T) \equiv x\left(mT_1 - \frac{T_1}{2}\right) = u_s\left(mT_1 - \frac{T_1}{2}\right)(-1)^{m+k+1} \tag{25}$$

Therefore, the even-numbered samples of $x(t)$, which occur at the rate of B samples per second, produce samples of the lowpass signal component $u_c(t)$. The odd-numbered samples of $x(t)$, which also occur at the rate of B samples per second, produce samples of the lowpass signal component $u_s(t)$.

Now, the samples $\{u_c(mT_1)\}$ and the samples $\{u_s(mT_1 - T_1/2)\}$ can be used to reconstruct the equivalent lowpass signals. Thus, according to the sampling theorem for lowpass signals with $T_1 = 1/B$, we obtain

$$u_c(t) = \sum_{m=-\infty}^{\infty} u_c(mT_1)\frac{\sin(\pi/T_1)(t - mT_1)}{(\pi/T_1)(t - mT_1)} \tag{26}$$

$$u_s(t) = \sum_{m=-\infty}^{\infty} u_s\left(mT_1 - \frac{T_1}{2}\right) \times \frac{\sin(\pi/T_1)(t - mT_1 + T_1/2)}{(\pi T_1)(t - mT_1 + T_1/2)} \tag{27}$$

Furthermore, the relations in Eqs. (24) and (25) allow us to express $u_c(t)$ and $u_s(t)$ directly in terms of samples of $x(t)$. Now, since $x(t)$ is expressed as

$$x(t) = u_c(t)\cos 2\pi F_c t - u_s(t)\sin 2\pi F_c t \tag{28}$$

substitution from Eqs. (27), (26), (25), and (24) into Eq. (28) yields

$$x(t) = \sum_{m=-\infty}^{\infty}\left\{(-1)^m x(2mT)\frac{\sin(\pi/2T)(t - 2mT)}{(\pi/2T)(t - 2mT)} \times \cos 2\pi F_c t + (-1)^{m+k}x((2m-1)T) \times \frac{\sin(\pi/2T)(t - 2mT + T)}{(\pi/2T)(t - 2mT + T)}\sin 2\pi F_c t\right\} \tag{29}$$

But

$$(-1)^m \cos 2\pi F_c t = \cos 2\pi F_c(t - 2mT)$$

and

$$(-1)^{m+k}\sin 2\pi F_c t = \cos 2\pi F_c(t - 2mT + T)$$

With these substitutions, Eq. (29) reduces to

$$x(t) = \sum_{m=-\infty}^{\infty} x(mT)\frac{\sin(\pi/2T)(t - mT)}{(\pi/2T)(t - mT)}\cos 2\pi F_c(t - mT) \tag{30}$$

where $T = 1/2B$. This is the desired reconstruction formula for the bandpass signal $x(t)$, with samples taken at the rate of $2B$ samples per second, for the special case in which the upper-band frequency $F_c + B/2$ is a multiple of the signal bandwidth B.

In the general case, where only the condition $F_c \geq B/2$ is assumed to hold, let us define the integer part of the ratio $F_c + B/2$ to B as

$$r = \left\lfloor \frac{F_c + B/2}{B} \right\rfloor \tag{31}$$

While holding the upper cutoff frequency $F_c + B/2$ constant, we increase the bandwidth from B to B' such that

$$\frac{F_c + B/2}{B'} = r \tag{32}$$

Furthermore, it is convenient to define a new center frequency for the increased bandwidth signal as

$$F_c' = F_c + \frac{B}{2} - \frac{B'}{2} \tag{33}$$

Clearly, the increased signal bandwidth B' includes the original signal spectrum of bandwidth B.

Now the upper cutoff frequency $F_c + B/2$ is a multiple of B'. Consequently, the signal reconstruction formula in Eq. (30) holds with F_c replaced by F_c' and T replaced by T', where $T' = 1/2B'$:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT')\frac{\sin(\pi/2T')(t - mT')}{(\pi/2T')(t - mT')}\cos 2\pi F_c'(t - mT') \tag{34}$$

This proves that $x(t)$ can be represented by samples taken at the uniform rate $1/T' = 2Br'/r$, where r' is the ratio

$$r' = \frac{F_c + B/2}{B} = \frac{F_c}{B} + \frac{1}{2} \tag{35}$$

and $r = \lfloor r' \rfloor$.

We observe that when the upper cutoff frequency $F_c + B/2$ is not an integer multiple of the bandwidth B, the sampling rate for the bandpass signal must be increased by the factor r'/r. However, note that as F_c/B increases, the ratio r'/r tends toward unity. Consequently, the percent increase in sampling rate tends to zero.

The derivation given above also illustrates the fact that the lowpass signal components $u_c(t)$ and $u_s(t)$ can be expressed in terms of samples of the bandpass signal. Indeed, from Eqs. (24)–(27), we obtain the result

$$u_c(t) = \sum_{n=-\infty}^{\infty}(-1)^n x(2nT')\frac{\sin(\pi/2T')(t - 2nT')}{(\pi/2T')(t - 2nT')} \tag{36}$$

and

$$u_s(t) = \sum_{n=-\infty}^{\infty}(-1)^{n+r+1}x(2nT' - T)\frac{\sin(\pi/2T')(t - 2nT' + T)}{(\pi/2T')(t - 2nT' + T)} \tag{37}$$

where $r = \lfloor r' \rfloor$.

In conclusion, we have demonstrated that a bandpass signal can be represented uniquely by samples taken at a rate

$$2B \leq F_s < 4B$$

where B is the bandwidth of the signal. The lower limit applies when the upper frequency $F_c + B/2$ is a multiple

of B. The upper limit of F_s is obtained under worst-case conditions when $r = 1$ and $r' \approx 2$.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.

SATELLITES IN IP NETWORKS

ABBAS JAMALIPOUR
University of Sydney
Sydney, Australia

Satellites have been an important element of telecommunications networks for many years in providing long-distance telephony and television broadcasting. The involvement of satellites in IP networks is a direct result of new trends in global telecommunications where the Internet traffic will have a dominant share of the total network traffic and special features of high-capacity satellite channels to be discussed in this article. The large geographical coverage of the satellite footprint and its unique broadcasting capabilities as well as its high-capacity channel keep the satellite as an irreplaceable part of communications systems, despite the high cost and long development and launching cycle of a satellite system.

In this article, we will review the satellite communications systems and introduce a new era for satellite communications toward broadband satellite systems and satellite-for-the Internet systems. This means that the satellite is changing its traditional role as being simply a relay in space into becoming an active element similar to a switch or a router in terrestrial networks. We will start the article with a short historical overview of satellite communications and then provide up-to-date information on new broadband and Internet satellite systems. We will briefly review third-generation wireless cellular systems, where Internet access is considered, in order to show the role and contribution of satellites in these systems and thus in the mobile Internet. Satellite applications within the third-generation wireless terrestrial systems as well as in the global Internet will be discussed. Several implementation topics, including mobility and location management that are common in satellite and terrestrial mobile networks, will also be discussed. We will then open the topics in satellite transport of Internet traffic and challenging issues that need to be resolved. Finally we will conclude the article with a concise but complete discussion on satellite future perspectives to the global Internet connectivity problem.

1. AN OVERVIEW OF SATELLITE COMMUNICATIONS

A satellite is one of the oldest components in telecommunications systems. For almost half a century, satellite networks have provided long-distance communications services to the public switching telephone network (PSTN) as well as television broadcasting. These services are particularly best justified by the large footprint coverage of the satellite, and to this date, there is no substitute for satellites in this field. In these types of service, a satellite acts as a communications repeater or relay (according to whether the transmitted signal is digital or analog) that communicates with ground stations and solves the problem of transmission of electromagnetic waves between different parts of the world that are not in line of sight of each other. A noteworthy achievement in satellite communications is

the formation of INTELSAT (International Telecommunications Satellite Organization), which in 1964 established a means of fixed-satellite service among nations [1].

In the 1980s, satellites were being deployed for the first time in mobile telecommunications by providing direct communications to maritime vessels and aircrafts. The first major development in this area was the INMARSAT satellite system. INMARSAT started a new era of satellite communications, called *mobile satellite services* (MSSs) in 1982. The INternational MARitime telecommunication SATellite organization used a geostationary satellite system using L-band (1.5–1.6 GHz) to provide telecommunication services mainly to ships. In the first generation of MSS, INMARSAT defined five standards: Standard A (1982), Standard B (1993), Standard C (1991), Standard M (1992/93), and the Aeronautical Standard (1992). Different worldwide telecommunication services, including voice, facsimile, and data were considered in these standards. While INMARSAT A and B are mostly considered the service to ships, INMARASAT C is planned to provide service to small craft, fishing boats, and land mobiles. INMARSAT continues its worldwide services as one of the most reliable satellite communication systems.

Although INMARSAT remains as the most distinguishable satellite system of its kind, there were other MSSs developed during the *first-generation mobile satellite systems*, such as QUALCOMM in North America (1989), ALCATEL QUALCOMM for Europe (1991), and the Japanese NASDA system (1987).

Reduction in size and cost of user terminals was the motive for second-generation MSSs started around 1985. In this generation, interconnection of satellite systems with terrestrial wireless systems has also been considered. INMARSAT defined its mini-M standard in 1995 with worldwide voice, data, facsimile, and telex service at 2.4 kbps (kilobits per second). American Mobile Satellite Corporation (AMSC), NSTAR of Japan, European mobile satellite (EMS), and several others are included in the second-generation MSSs.

Satellite systems have always being faced with unavoidable long propagation delay and large transmission power requirements. Consideration of small-size user terminals and direct radiocommunications between users and satellites (i.e., without using a ground station) led to the idea of using satellites in orbits lower in altitude than the geostationary orbit. Among possible orbit selections, low-earth-orbit (LEO) satellites with altitudes of 500–1500 km and medium-earth-orbit (MEO) satellites of altitude 5000–13000 km were considered [1]. The altitude selection given above assures that the satellites reside outside the two Van Allen belts to avoid the radiation damage to electronic components installed in satellites. The use of these nongeostationary satellite systems for commercial purposes started a new era in mobile satellite communications. Use of spot-beam antennas in these satellites produces a cellular-type structure within coverage areas, and hence a frequency reuse scheme can be applied, making the system a high-capacity cellular-like network on the ground with satellites as the base stations in space.

LEO and MEO satellite systems, because of their shorter distance to the earth, solve the problem of long propagation delay and high power consumption, but they introduced new challenges to the communications industry. Since the satellite is closer to the earth, compared to a geostationary satellite, it is not possible to employ just three satellites to cover all parts of the world as in case of geostationary satellite systems. Therefore, for LEO and MEO, a constellation of satellites in order of tens of satellites is required. This means more complexity and, of course, higher cost to the satellite system, which eventually must be passed to the users. Many LEO and MEO satellite systems were proposed in the early 1990s in North America and around the world and obtained frequency spectrum licenses. Only a few of these systems were completed and became operational, including IRIDIUM (1998) with 66 satellites and GLOBALSTAR (2000) with 48 satellites. However, financial problems associated with the high-cost of LEO systems forced IRIDIUM to cease its operation in 2000. Besides higher network complexity and more expensive control management requirements of nongeostationary satellite systems, a LEO or MEO satellite itself has a shorter lifetime than in geostationary systems. This means more frequent satellite launch requirements and higher maintenance cost to the satellite system.

The operational failure of the advanced but complex IRIDIUM satellite system revealed that although the technology for implementing a mobile satellite phone system is available, it is not possible to compete in the cost and services of such a system with the rapidly growing terrestrial cellular systems and new Internet services using LEO satellites. The roaming capability between different second-generation (2G) cellular networks in different countries and those considered in the third-generation (3G) wireless systems are quite adequate to provide telecommunications services to the majority of world population at lower cost and better quality (e.g., delay) than are those that can be achieved through satellite systems. The new trends in the telecommunications industry in transmitting data traffic and Internet traffic at high speed over wireless channels could not be matched by satellite systems. The IRIDIUM system, for example, could provide short data services at the very low data rate of only 2.4 kbps.

Satellite systems, however, maintain their unique feature of broadcasting. *Satellite broadcasting* has been a success for a long time and continues its dominance for long-distance coverage and service to highly populated telephony networks. If this unique feature of satellite systems can be incorporated into the new trends in telecommunications industry toward high-speed Internet access, then a new era of satellite communications technology will have begun. Broadband satellites are being developed for this market.

Broadband satellites [2–5] are recognized as systems that can provide high data rate transmission in the order of $\geq$1 Mbps. Digital video broadcasting (DVB) systems such as Eutelsat, SES, and INTELSAT; proprietary systems such as Spaceway, Astrolink, and iPSTAR; and proposed systems such as Teledesic, SkyBridge, WEST, and Celestri are among such broadband satellite systems. Standardizations of these satellite systems are

ongoing [6] in order to reduce the cost and increase applicability, similar to the way in which terrestrial cellular systems have developed and become successful. In this standardization, multicasting is also considered as a strong feature. Geostationary satellite systems are becoming of main interest for these services. Some of the applications of broadband satellite systems are shown in Fig. 1, which illustrates how satellites can interconnect geographically distant networks through land gateways. The system shown in this figure is designed to provide access to the Internet contents in one place by users of many other networks. Each network usually includes a caching system for fast local multicast to its Internet users.

Broadband satellite systems can be categorized according to their specifications and capabilities. This could be based on the frequency bands of operation (C band 4–8 GHz, Ku band 10–18 GHz, Ka band 18–31 GHz, and higher bands V and Q); the orbit altitude and hence the satellite lifetime; power requirements and antenna size; usage of bent-pipe or onboard processing (OBP) technologies; global or regional coverage of the system; satellite total capacity and user capacity; use of intersatellite links (ISL); number of supported terminals and required gateways; protocol used in the satellite system such as TCP/IP, DVB, and ATM; use of open or proprietary standards; total number of satellites in the system; and the total cost. Most new designs of satellite systems include onboard processing (OBP) and on-board switching (OBS) facilities so that the satellite node changes its simple role of relaying into being an active element in the network. An example of a functional satellite system that includes onboard switching and an ATM-switch-like satellite is shown in Fig. 2.

There are many regulatory and standard bodies currently involved in development of issues related to the satellite communications industry. Regulatory bodies include WRC, FCC, ITU, ERO, and many national and regional regulatory organizations. Standards are developed mainly in IETF, ETSI, TIA, and ITU.

To conclude this section, we must say that satellite systems have long development cycles compared to terrestrial systems, and usually less funding and fewer engineers are involved in their development. Moreover, they compete over a more limited market than cellular systems. Most satellite systems are still proprietary, and interfaces are not public, which, in turn, prevents competition. The standards for satellites will ensure interoperability and real competition and will be required for a broader consumer market. Therefore, in order to see further development in satellite systems, a widely acceptable standardization is vital.

2. SATELLITES FOR THE GLOBAL INTERNET

New multimedia and Internet services demand more cost-effective high-quality and high-speed telecommunication technologies and architectures. The primary issue is how the current global Internet infrastructure can be expanded so that the quality of service can be improved from current best-effort service and that high-speed access can be achieved. In this context, satellites can play an important role in expanding the Internet infrastructure using the large coverage area feature and in providing high-speed data transmission through a high-bandwidth-capacity channel. Satellites, however, would not perform this task as an isolated network but rather use an efficient integration with current terrestrial networks. So instead of having an IP network in the sky, as suggested in earlier proposals on some satellite IP networks, a combination of terrestrial and satellite networks would be a solution to the future high-speed Internet.

In a global Internet infrastructure, satellites can be used for many purposes. They can be used to connect geographically distant segments of the network or interconnect heterogeneous networks. Satellites can provide direct telecommunication service to aircraft, ships, and isolated local networks on the ground and even to individual users. Flexible and quick deployment of bandwidth by satellite systems, make them easily approachable by densely wired networks when required, as a good backup and supporting network.

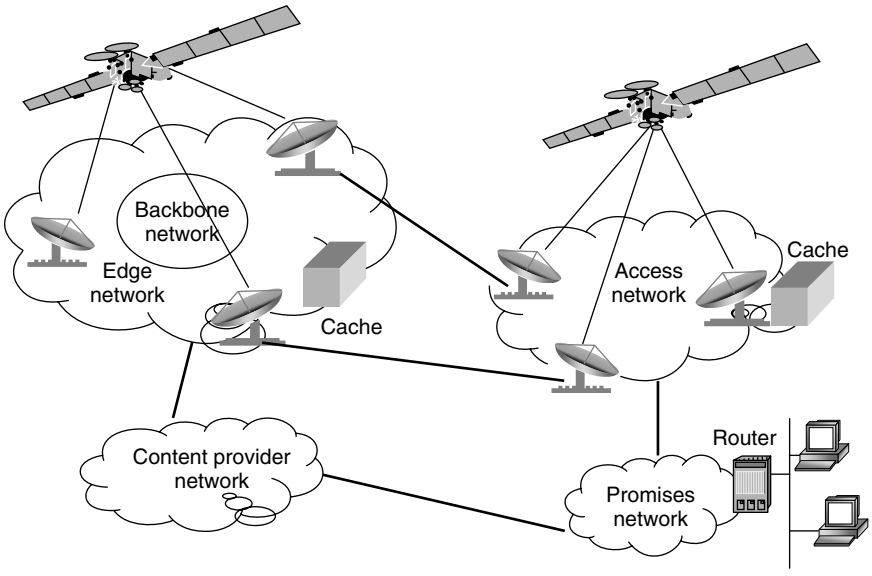


Figure 1. Applications of broadband satellites in interconnecting different networks.

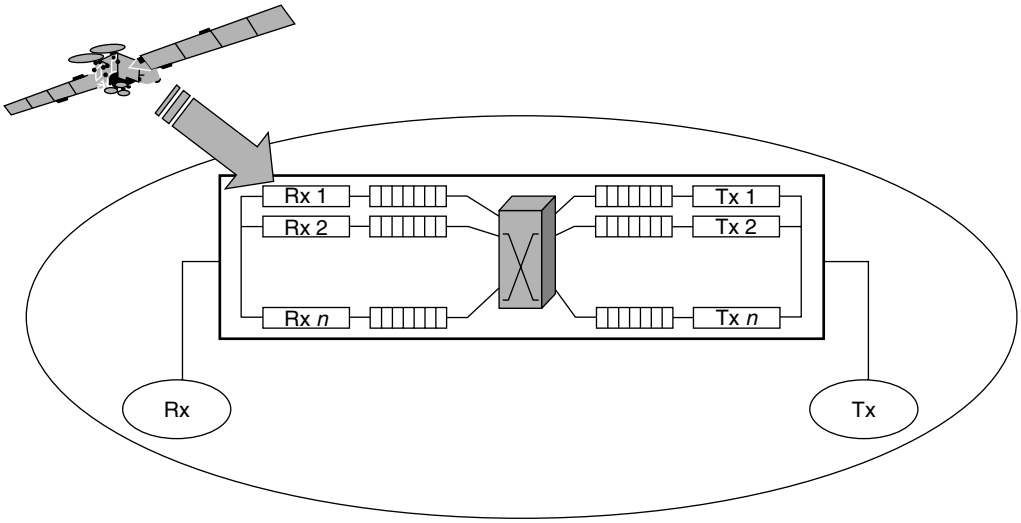


Figure 2. An example of satellite with onboard processing and switching facilities.

Figure 3 shows two different options for the satellite payload that can be used in satellite-based Internet architectures. In Fig. 3a the satellite is used as a reflector in space connecting separate network segments through ground gateways. In Fig. 3b, however, the satellite acts as an active component of the network that can utilize routing and switching processing. The satellites used in Fig. 3 can be on any of the altitudes explained in Section 1; that is, geostationary or nongeostationary (LEO or MEO) or a combination of different altitudes. The satellites shown in Fig. 3b, in addition to having connection to the ground gateways, are also employed in intersatellite links so that network connectivity can be created in the sky independently. This method should be considered as an important option in a future satellite-based Internet architecture. The method requires higher cost for the

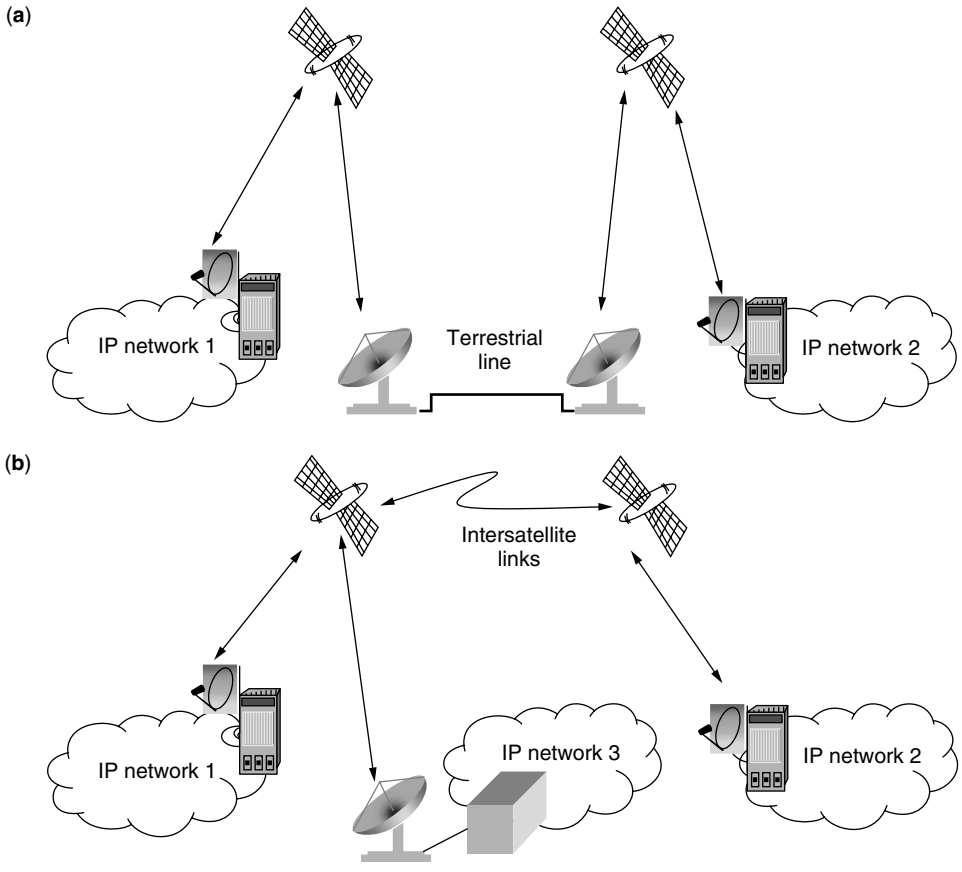


Figure 3. Two different payload options for satellite-based IP architectures: (**a**) bent-pipe architecture; (**b**) onboard processing satellites.

system and more complicated routing management. If special facilities are included in the mobile stations on the earth, both methods can provide direct Internet connectivity to remote users without any other alternative terrestrial telecommunications infrastructure. Hu and Li have summarized the satellite-based Internet architecture described above and current proposals of this kind [7].

A satellite node can also be used as a high-speed downlink for home Internet access. In this method, a home or office user with a satellite receiver, usually used for satellite television, can download the Internet contents at a very high data rate through the satellite downlink channel. A simple architecture of such a satellite–ground high-speed Internet is shown in Fig. 4. In the architecture shown in this figure, the user first connects to its Internet service provider (ISP) using a normal dialup connection. The dialup connection forms a low-speed data communication (e.g., a typical 56-kbps connection) mainly in order to send requests to the Internet servers at the local ISP site. All Internet contents can then be forwarded to the customer through the high-speed satellite downlink on receiving the request. The downlink can send the data to the user at speeds of 1 to a few Mbps using digital videobroadcasting satellites or other types of satellites. This method is especially appropriate for video-on-demand and other type of real-time Internet applications where many users located in the same region want to retrieve the same contents over the Internet. The asymmetry in the Internet traffic that usually results in up to 10 times more data traffic on a typical Internet downlink connection compared to the uplink makes this method of special interest and application. Currently this method is competing with other high-speed Internet access for home users, including cable modems and ADSL technologies. Some prototypes of these satellite Internet systems for home users have already been developed and demonstrated in Europe and other parts of the world [8,9]. With some modifications it is possible to extend the coverage of this type of Internet access to mobile users on the ground and also during long-distance flights and to ships.

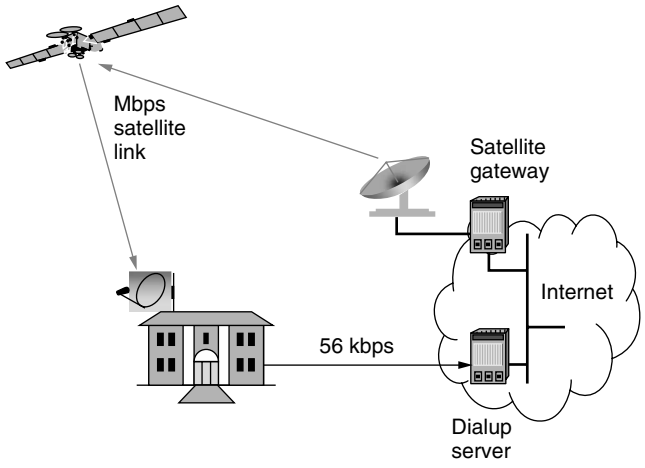


Figure 4. An architecture for satellite high-speed Internet access.

3. SATELLITES IN THIRD-GENERATION WIRELESS NETWORKS

With the increasing popularity of portable computers and the expanding Internet capabilities of mobile phone handsets, a large demand for mobile computing has been generated. Thus, instead of restricting data connections to be maintained always at a fixed position in the network, mobile users will be provided with equivalent multimedia and IP services. There is no doubt that the trend is toward a global mobile networking environment. In such a network, broadband satellites can be considered as an integral part of the network interconnecting the fast-growing terrestrial cellular and wired networks.

Broadband satellite networks for Internet access are the new generation of satellite networks in which Internet-based applications and services will be provided to users regardless of their degree of geographic mobility [2,3]. The main distinction from conventional satellite networks will be that the new satellite networks will support high-data-rate transmission and broadband services and in particular the Internet. The Internet is the most rapidly growing technology, and many new applications such as electronic commerce find their way through the Internet. Therefore, it is not surprising that broadband satellite networks focus on the Internet-based applications for their primary services, although voice and low-bit-rate applications still remain in the list of network services. Asynchronous transfer mode (ATM) will be the envisaged switching mode for future broadband satellites because of its support of a variety of traffic such as constant and variable bit rate and quality of service (QoS) support [2]. Nevertheless, IP routing is considered as another alternative for these satellites due to lower cost and "friendliness" toward the Internet traffic.

In the sphere of terrestrial networks, there are at present two possible approaches for establishing the task of mobile computing: cellular-based and IP-based solutions [3]. Intuitively, while a cellular-based solution enhances the current mobile communications by extending capacity for data and multimedia transmissions, an IP-based solution allows for user mobility by maintaining all ongoing Internet connections even in the presence of frequent handoffs or changes in the network point of attachments. In the forefront of these technologies, third-generation wireless systems are being considered.

Third-generation (3G) wireless communications systems evolve by orienting the integration of three essential domains: broadband, mobile, and Internet (IP). In such a milieu, the increasing feasibility of virtual connections allows mobile users not only to roam freely between heterogeneous networks but also to remain engaged in various forms of multimedia communications. Whether it is geographic coverage, bandwidth, or delay, it would then be up to the users to decide when and how to switch from one access network to another depending on the availability and appropriate cost/performance considerations and, thus, advancing toward an era of all-IP-based communications. Consequently, it will be necessary to implement the 3G system as a universal solution that prompts transparent user roaming (among different wireless networks)

while delivering the widest possible range of cost-effective services [10].

IMT-2000 (International mobile Telecommunications) is a unified 3G mobile system that supports both packet-switched and circuit-switched data transmissions with high spectrum efficiency, making the vision of anywhere, anytime communications a reality. Basically, it is a collection of standards that provides direct mobile access to a range of fixed and wireless networks. Among all, the three most significant developments are UMTS, cdma2000, and UWC-136, which are the 3G successors to the main 2G technologies of GSM, IS95, and AMPS, respectively [11]. The general idea was to make the development of 3G wireless technologies a gradual process from circuit-switched to packet-switched. Take GSM (Global System for Mobile communications) for example. In order to have the system enhanced with improved services (by means of increased capacity, coverage, quality, and data rates), the evolution to 3G was made possible through the incorporation of an intermediate stage called GPRS, the general packet radio services.

Based on the enhanced core network of GPRS, UMTS (Universal Mobile Telecommunications System) is designed to be the backward-compatible 3G standard for GSM. UMTS is the European proposal for a 3G mobile system aiming to support multimedia services with extended intelligent network features and functions. As a first step of the integration, UTRAN (the UMTS terrestrial access radio network) will coexist with GSM access networks. The idea was to develop the UMTS core network by gradually incorporating the desired UMTS features to the GSM/GPRS core network. At this stage, UTRA supports time-division duplex (WCDMA-DSTDD) and frequency-division duplex (WCDMA-DSFDD) modes with the combined operation offering an optimized solution to coverage areas of all sizes. A further multicarrier (MCFDD) mode is to be established at a later date intended mainly for the use in cdmaOne/cdma2000 evolutions [12].

For satellite systems, the situation is somehow different. The most apparent is that the market for satellite systems is much more limited than their cellular counterparts. Therefore, it would be difficult to assume the same approaches for satellite systems. Instead satellite systems can incorporate their global coverage feature for enhancement of the 3G terrestrial networks. Satellites can establish a high-speed backbone network to support the terrestrial networks and also to use their broadcast nature to deliver Internet content at high speed directly to a group a users.

Satellite UMTS (S-UMTS), for example, is considered as a component of 3G networks [13]. The satellite segment of the network connects through appropriate interworking units (IWUs) to the ground segments. An illustration of this incorporation of satellites in providing mobile Internet connectivity is shown in Fig. 5. IWU for the satellite has similar functionality as the gateways used to interconnect 2G and 3G networks for interoperation of these networks during the transition period from 2G to 3G as well as the gateways used for interconnection of different operator networks of the same kind (e.g., GSM). Such a concept is depicted in Fig. 6.

4. TECHNICAL ISSUES FOR SATELLITE-BASED INTERNET IMPLEMENTATION

After the overviews on satellite communications and third-generation wireless networks and introducing the role of satellites within the 3G network architecture in previous sections, in this and the following section we look at some specific but important issues for mobile satellite networks that also apply to terrestrial and cellular networks.

4.1. Mobility Management

It is widely agreed that to allow seamless user mobility, several considerations are necessary to ensure smooth

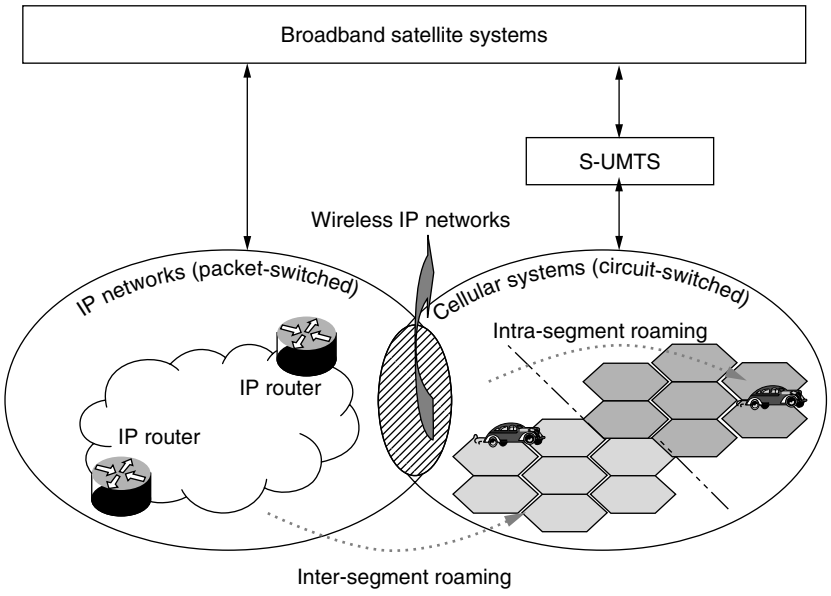


Figure 5. Satellite applications in global communications networks.

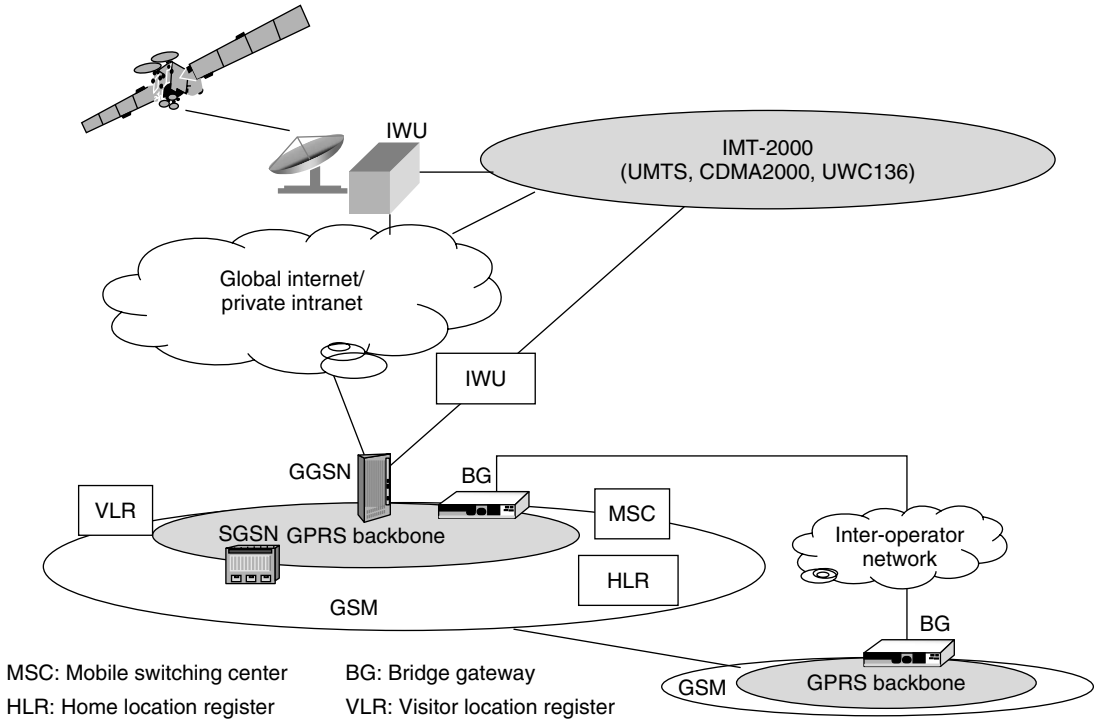


Figure 6. Interconnection of different terrestrial and satellite networks through interworking units.

transitions between different wireless technologies. Ultimately, mobility management is the key in enabling successful convergence between wireless communications and computing. Often, mobility management is interpreted as a process that simply routes packets from one point (the source) to the other (the intended destination). However, such an assumption becomes inadequate as more and more unsolvable issues gradually become apparent. Given its complexity, there seems to be an inevitable need to redefine the previously overlooked issue—*the mobility problem*.

Very briefly, mobility implies adaptability: the capability of maintaining any established network connections by accommodating different system characteristics when a mobile user roams within and/or between networks. More specifically, mobility refers to the initiation of a handoff process, not only when moving between cells (as in a cellular wireless environment) but also when roaming from one wireless network (e.g., satellite) to another (e.g., GPRS). Depending on the level of the network stack from which a movement is considered, mobility can be classified into three categories: air-interface mobility, link-level mobility, and network-level mobility.

Air-interface mobility is perhaps the most common case where a handoff takes place between two adjacent base stations (BSs) or access point (APs) within a radio access network. One can envisage this scenario as a pedestrian walking across microcell boundaries while being engaged in a conversation whether through voice or data transmissions. Link-level mobility goes one level up in the network hierarchy, and is concerned with maintaining a point-to-point protocol (PPP) context across multiple radio access networks. The transitions, however, would still be within the same domain and technology. On the highest network level (among the three categories), network (or IP) mobility provides network level mobility between different access networks (including wireless). Basically, this involves a change in the mobile's domain- (or location-) related IP address due to either (1) a change in radio access technologies or (2) a transition from one network operator to another. Note that in the latter case, the two networks involved might be implemented by the same access technology. Figure 7 attempts to illustrate the differences by listing the hierarchy of concern in the three cases. Note that the overall structure of a subnet can be considered to consist of three distinct stacks, with each stack developed depending on the specific technology (or network architecture) under consideration.

Most issues related to mobility are associated either directly or indirectly with service delivery. By and large, it involves, but is not limited to, the process of routing management, handoff management (including resource management), and also QoS management. Having said that, each operation has its own set of predefined actions. Given that it might be possible to incorporate some of the specific techniques used in satellite systems to enhance the complete operations in a 3G network, the purpose of the brief discussion in the next few paragraphs is to frame this part of the mobility problem.

4.2. Location Management

Obviously, mobility is managed largely through the process of location management. This action involves not only storing and/or retrieving information from the location database but also sending paging signals whenever necessary to locate a roaming user. Although

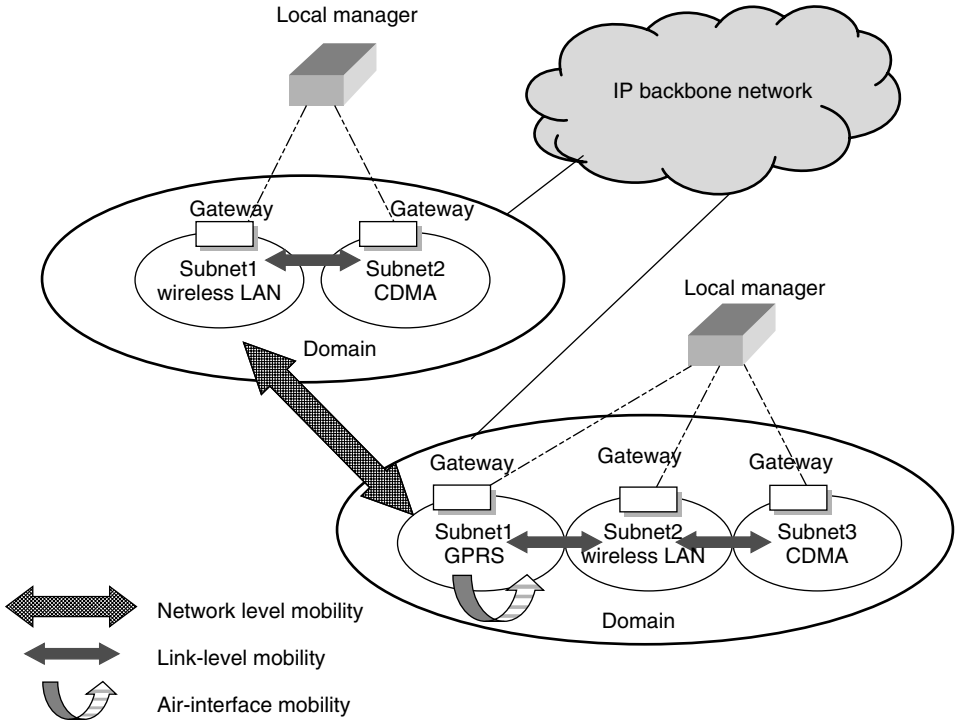


Figure 7. Different levels of mobility.

the need to notify the home network of mobile users' current locations is always present, it is questionable whether accurate location information is essential for mobile nodes (MNs) that are not committed to any data transmissions. Somehow, it seems logical to have separate mobility management techniques for idle and active mobile nodes and thus to allow variations in adjusting the updating frequency at which MNs' movement notifications are sent [14].

4.3. Routing Management

One issue that relates closely to the provision of global roaming is the establishment (or management) of roaming agreements between various networks. As a mobile user roams across various geographic and/or network boundaries, appropriate global roaming agreements between networks (or among ISPs) will also have to be established [15]. In fact, the incorporation of satellites is a particularly significant example of such operations. In scenarios where the limited users population (or terrain conditions) has made it infeasible to implement wireless terrestrial technologies, the availability of satellite systems would instantly form an alternative access medium for a dual-mode handheld terminal.

4.4. Handoff Management

A quality handoff management is particularly important for any real-time transmissions. As continuity plays a crucial part in grading the service quality for such a communicating purpose, it is worth exploring the factors that would have impact on handoff performance.

Mobility implies the necessity of MN handoffs. It is the process of reassociating the roaming mobile with a designated entity (in the new network) to maintain the much needed service continuity. Handoffs are performed at two layers; in the link layer (OSI layer 2, which maintains link connectivity) and also in the IP layer (OSI layer 3, which maintains network access). Physical connection to a base station can often be assumed to be seamless and almost instantaneous. The term network access denotes a MN's capability to exchange traffic in its current subnet without compromising its permanently allocated identity (IP address). A MN is said to have network access when its current subnet location corresponds with its registered network access point [16].

Successful handoffs are also crucial to ensure continuity of ongoing data connections (hence the provision of seamless roaming). However, to appreciate various coexisting network standards, decisions on the necessity of a handoff should be based on underlying network characteristics and specific application scenarios. This is particularly the case for intertechnology roaming (e.g., overlaying networks of Wireless LAN and GPRS), where the system characteristics might be significantly different. Essentially, a tradeoff analysis among the resultant throughput, data rate, latency, and disruption measures would be beneficial in such instances, to determine the actual essentiality of a handoff action. In this respect, the availability of resources has a direct influence on the number of successful handoffs. Thus, specific issues with regard to resource management, such as channel allocation schemes and call admission controls, should all be carefully considered.

4.5. Quality of Service Management

Quality of service management is another significant area that has gained tremendous research interest at both

academic and commercial levels. In particular, with the gradual replacement of voice traffic by data traffic (and multimedia transmissions), a proper implementation of suitable bandwidth allocation mechanisms is crucial to allow successful provision of satisfying customer services. Even apart from this, the possibility of having a network capable of maintaining multiple, concurrent QoS flows (for various applications running simultaneously) would also be advantageous. In addition, the users should be given the opportunity to alter or renegotiate (when desired) the predefined service-level specifications (SLS) with the corresponding providers through a continuous monitoring process of customer requirements [17].

Finally, there is a need to set up necessary SLS between networks. This is important as it allows user mobility to be managed independently of the access and backbone networks (e.g., B-ISDN or GSM), while maintaining a certain quality level that has been specified by any one particular mobile user.

5. MOBILITY MANAGEMENT IN MOBILE SATELLITE NETWORKS

In the previous section we discussed some important aspects of mobility management. Mobile satellite systems using nongeostationary orbits such as LEO and MEO have characteristics very similar to those of cellular networks, and thus those issues are common between both networks. The main similarity is that most of these systems use the cellular concept of increasing the total system capacity through a cellular-like coverage area arrangement introducing similar handoff issues as those involved in cellular systems. There is, however, a major difference between the proposed scheme in satellite systems and what exists in the sphere of cellular networks. The key operational concept that differentiates the specific operations of mobility management in the two systems is the "entity" being considered as the moving object. While the former encounters *mobile* movements within a fixed network architecture, the latter incurs *satellite* movements in reference to fixed mobile nodes. Essentially, this suggests that the mobility management technique will be different.

Furthermore, while the prediction of mobile's deterministic location is relatively easy to obtain (given the traveling characteristics, particularly speed of the LEO satellites), such certainty is not guaranteed in cellular systems. In other words, the difficulties encountered in cellular networks do not seem applicable to its satellite–mobile network counterparts. As a result, appropriate modification of the existing ground based solutions will be necessary before similar operations are suitable for applications on satellite-incorporated third-generation systems. Generally, it would be conceptually a lot easier to anticipate satellite movements than mobile movements. Besides, the rate of call arrival would not be essential in the latter case (i.e., paging seems to be less of a problem for satellite applications).

The use of LEO satellites is most favorable for its high traffic capacity and reduced user power requirements in both satellites and the ground terminal. However, as individual LEO satellites rotate relatively fast along the earth's surface, handoff becomes a particular concern in coping with the nonstationary nature of the coverage area. Depending on the relations of the two satellites involved in the handoff operations, three types of handoff are classified. Basically, *intrasatellite handoffs* are used to describe changes between spot beams under the management of the same satellite. With the increasing involvement of network management, *intersatellite handoffs* indicate handoffs between satellites and link handoffs incur due to changes in the connection pattern of satellite footprints (or satellite network topology). As an example of the latter scenario, such handoffs occur when links to adjacent orbits are turned off when the concerned satellite moves near to the polar region. Thus, the task is not only to utilize the available frequency spectrum efficiently but also to minimize unnecessary forced termination of connections due to handoff failures. In other words, it would be essential to at least attempt to anticipate user motions and to reserve the resources accordingly for the predicted residual time. A brief literature survey indicates that two major prioritization techniques were designed specifically for such purposes: (1) use of guard channels and (2) queuing of handoff requests when the resources were not currently available. Consequently, the operation of call admission control also becomes important as it decides whether sufficient resources would be available to accommodate the newly arrived transmission requests [18].

Cellular networks, on the other hand, focused more on the efficient operations of location management specifically for idle mobile users. Thus, although the basic problem of managing mobility is the same, the actual emphasis on the system developments is different for satellite and cellular systems. In fact, because of the movement of the LEO satellites, definitions of location area (LA) cannot be fixed even for the duration of a connection. Consequently, it is difficult to reapply some of the existing solutions from one system to the other (i.e., from cellular to satellite and vice versa). However, there are certain aspects where the "approaches" might be useful to seek alternative solutions for the open issues identified. For example, in terms of routing [19], a protocol has been developed in an attempt to reduce the frequency rerouting attempts during a link handoff.

Basically, *target probability* is defined to quantify the estimated duration of residency of a mobile terminal on one particular intersatellite link (ISL). During the route establishment of a new call, only links that can demonstrate a lifetime of greater than the target probability will be considered to form segments of the route. Although it might not seem obvious, this idea is similar to the predicting method used in location management operations in cellular networks, specifically, in the sphere of sequential paging where the optimal sequence is selected according to the probability of residence in individual subsections (or subareas). Thus, while acknowledging the fact that the prediction method would have been easier in the satellite systems (because its motion is deterministic and predictable), the goal

of determining efficient operations in both systems is the same.

Consequently, it becomes necessary to more closely identify the differences (or relations) between the operations of handoff and location management. Clearly, handoff management is significant only when the mobile unit is active; specifically, it is about the appropriate reservation of resources (such as bandwidth) along the roaming path of a mobile user while engaging in a call connection. Its efficient operation is important to ensure that the various aspects of the QoS requirements (e.g., throughput versus forced call termination) are satisfactorily complied. Location management, on the other hand, is intended mainly for users who are currently idle but are expected to receive calls (or become active) while they frequently change their point of attachment to the network. In essence, only sufficient location information (about the mobile) is maintained so that the network could loosely track the mobile's movement and subsequently incur a minimal paging (or searching) load when the precise residency is required. Consequently, it would be correct to conclude that the predicted information for handoff needs to be more reliable than that desired for efficient location management. On the basis of this observation, it seems potentially viable to combine (at least to some extent) the operations of the two management processes of handoff and location.

6. SATELLITE TRANSPORT OF INTERNET TRAFFIC

Broadband satellite networks are being developed to transport high-speed multimedia and in particular Internet traffic through high-capacity satellite channels to network segments as well as to individual users. As in the case of any other wireless network designed to deliver Internet traffic, broadband satellite networks need to connect to the backbone wired Internet on the ground. TCP (Transmission Control Protocol) is the most commonly used protocol at the transport layer of the network stack in the Internet, originally developed in wired networks with low bit error rate (BER) in the order of less than 10^{-8}. In this context, any wireless network with Internet service needs to be compatible with the protocol used in the wired network: mainly the TCP/IP protocol. There are, however, some design issues in the TCP/IP protocol, which make it difficult to use it efficiently over the wireless and satellite links. There have been many research activities comparing the performance of TCP in high-BER and high-latency channels and modification proposals to improve its performance in terrestrial and satellite wireless networks [20–24].

TCP has been designed and tuned for networks in which segment losses and corruption of performance are due mainly to network congestion. This assumption might be invalid in many of the emerging networks such as wireless networks. The flow control mechanism used in TCP is based on timeout and window-size adjustment, which can work with high utilization in wired networks with low BER, in the order of 10^{-8}. However, when the wireless channel is used (partially or totally) as the physical layer with a BER as high as 10^{-3}, it may perform inefficiently. The reason is that in the wireless channels the main cause for packet loss is the high BER and not congestion as it is in wired networks. The low efficiency of the TCP in a wireless channel is a result of the fact that the TCP misinterpreted the packet loss because of high error rate and congestion. On the other hand, in high-latency networks (such as satellite networks) adjustment of the window size could take a long time and reduce the system throughput.

TCP has the ability to probe the unused network bandwidth by a mechanism called *slow start* and also to back off the transmission rate on detection of congestion through the *congestion avoidance* mechanism. At the connection startup, TCP initializes a variable called *congestion window* to a value of one segment. This variable determines the transmission rate of TCP. The window size is doubled at every round-trip period until a packet loss is experienced. At this time, the congestion avoidance phase commences, the window size is halved, and the lost packet is retransmitted. During this phase of TCP, the window size is increased only linearly by one segment at each round-trip period and might be halved again on detection of another packet loss. If the retransmitted packet is lost, the timeout mechanism employed in TCP reduces the window size to one. Since all these procedures are performed at the periods equal to round-trip delay of the channel, the system throughput could be degraded significantly where high-latency channels such as geostationary satellites are involved. Therefore, the high-latency satellite channel, combined with the slow increase of the TCP congestion window size, results in underutilization of the satellite high-capacity channel.

Some modifications to the basic TCP can be made to ensure more efficient performance in high-latency satellite networks with Internet services (e.g., see Refs. 21–23 and their reference lists). *Selective acknowledgment* (SACK) TCP (RFC 2018), for example, is a method in which multiple losses in a transmission window can be recovered in one round-trip period instead of two in the basic TCP. *TCP for transactions* (T/TCP) also reduces the user perceived latency to one round-trip delay for short transmissions (RFC 1644). In *TCP spoofing* (cited by partridge and shepard [21]), a router close to the satellite link is considered that sends back acknowledgments for the TCP data. The responsibility of any segment loss in this method belongs to the router. In another method, called *split TCP*, a TCP connection is divided into multiple TCP connections and a special *satellite TCP* connection is employed for the satellite link part.

Another alternative for delivering Internet traffic through broadband satellite networks and simultaneously providing quality of service is to use IP-over-ATM or ATM protocols. In this regard, the IP protocol will provide the availability of various Internet applications, whereas the ATM protocol supports connection between two end-user terminals with a guaranteed end-to-end quality of service. An example of such protocol combination has been proposed for the Astra Return Channel System (ARCS), a geostationary multimedia satellite system using the Ka band on the return channels and the Ku bands on the forward channels [25].

In conclusion, we can say that the use of basic TCP in future broadband satellite networks will impose significant problems, especially in the case of short transmissions (compared with the channel delay–bandwidth product). For the geostationary satellite links the major problem with TCP is the long round-trip time, whereas in the case of non-geostationary satellite networks, the round-trip delay variation or jitter becomes more dominant. In both situations, the burst error nature of the satellite channel and the high BER require more sophisticated flow and congestion control mechanisms that can separate the segment loss because of network congestion or because of high channel error rate.

7. SUMMARY AND CONCLUSIONS

In this article, we summarized the satellite communications from a networking point of view in order to see the role of satellites in future mobile and fixed IP networks. The historical summary provided in the first section of this article revealed that despite the high initial investment and maintenance cost of satellite systems, satellites will remain as an irreplaceable component for long-distance communications and multimedia broadcasting. With the progress in optical communications and increasing number of transoceanic cables, it may be mistakenly thought that cable will replace the satellite for long-distance communications. However, the satellite's easy and quick deployment of additional capacity in any part of the world provides a distinct advantage over the deployment of cable systems. Improvement in cable television also could not replace satellite's broadcasting feature, especially due to the satellite's large footprint and simpler deployment.

When it comes to high-speed Internet access to the home, office, ships, aircraft, and mobile users, again satellite systems show its unique features. The global Internet needs expansion in both the geographic domain and data transport capacity. Satellites would be the main telecommunications component, if not the only one, as in many terrain circumstances, that can promise such expansion. The satellite huge onboard channel capacity and large coverage area are sufficient to provide future deployment of new systems. A very handy example would be the efforts toward realization of in-flight Internet access to passengers using satellite networks by major aircraft companies and airlines [26]. Satellites will soon bring inexpensive Internet access to long-distance flights and using voice-over-IP techniques make a huge reduction in the cost of phonecalls from and to airplanes.

For high-speed Internet access to the home and small-office users, currently ADSL and cable modem are the two leading technologies. With new digital videobroadcasting satellite systems in North America and Europe, however, these technologies found a need to compete with the satellites. The number of subscribers to satellite high-speed Internet access is increasing and close to the other two technologies, and this number is expected to increase even more rapidly by introduction of inexpensive satellite receivers in the near Future.

Satellites have an even larger contribution in IP networks than to the individual access discussed above. A satellite node can be an intelligent ATM switch or IP router in the sky interconnecting segments of the backbone networks on the earth. Similar to the conventional usage of satellites in public switching telephony networks, satellites can play an important role in future packet-switched networks, including the public Internet. The third-generation wireless networks and beyond consider Internet and multimedia traffic to have the dominant share of the network traffic load, and satellites have already shown their role in completion of any terrestrial mobile network. An example of satellite UMTS was given in this article to outline the role of satellites in future mobile communication systems. Satellite ground stations acting as an interworking unit can solve the roaming issue between heterogeneous wired and wireless terrestrial networks, expanding the telecommunications to its ultimate universal stage.

Some important technical implementation issues concerned with a global IP network have been discussed in this article. Mobility management has been revisited and redefined and location, handoff, routing, and quality of service managements have been discussed. All these issues are current research topics in mobile and satellite communications. For the high-latency satellite channel, as well as the error-prone wireless channel (including both satellite and terrestrial), the need for improvement in transport protocols currently employed in the Internet has been discussed and state-of-the art research activities toward improvement of TCP protocols has been reviewed. Note that other researchers are also currently working to improve the error probability of the wireless channel using forward error correction (FEC) schemes and sophisticated coding algorithms. Although these works are of great importance in the establishment of a better-quality wireless channel, we should not forget there are always situations in which the wireless signal-to-noise ratio is too low and no coding scheme can improve it. Therefore, a better solution would lie in the higher layers of the network, including the transport and network layers, where enhanced flow control algorithms speed up the data rate and the throughput of the wireless channel.

BIOGRAPHY

Abbas Jamalipour has been with the School of Electrical and Information Engineering at the University of Sydney, Australia, where he is responsible for teaching and research in wireless data communication networks and satellite systems, since 1998. He holds a Ph.D. degree in Electrical Engineering from Nagoya University, Japan. His current areas of research include wireless broadband data communications and wireless IP networks, mobile and satellite communications, traffic modeling, and congestion control. He is a recipient of a number of technology and paper awards and has authored two technical books and coauthored two others. He has authored numerous publications in these areas, and given short courses and tutorials in major international conferences. He has served on several major conferences technical committees, and organized and chaired many

technical sessions and panels in international conferences, including a symposium on satellite IP in IEEE Globecom 2001. He is the Vice Chair to the Satellite and Space Communications Committee of the IEEE ComSoc and has served as a guest editor to two special issues on 4G networks in IEEE magazines. He is a technical editor to the *IEEE Wireless Communications Magazine* (formerly, *Personal Communications Magazine*) and a Senior Member of IEEE.

BIBLIOGRAPHY

1. A. Jamalipour, *Low Earth Orbital Satellites for Personal Communication Networks*, Artech House, Norwood, MA, 1998.
2. A. Jamalipour, Broadband satellite networks — the global IT bridge, *Proc. IEEE* (special issue on multidimensional broadband wireless technologies and services) **89**(1): 88–104 (Jan. 2001).
3. A. Jamalipour and T. Tung, The role of satellites in global IT: Trends and implications, *IEEE Pers. Commun. Mag.* (special issue on Multimedia Communications over Satellites) **8**(3): 5–11 (June 2001).
4. J. Farserotu and R. Prasad, A survey of future broadband multimedia satellite systems, issues and trends, *IEEE Commun. Mag.* **38**(6): 128–133 (June 2000).
5. P. Chitre and F. Yegenoglu, Next-generation satellite networks: Architectures and implementations, *IEEE Commun. Mag.* **37**(3): 30–36 (March 1999).
6. J. Neale, R. Green, and J. Landovskis, Interactive channel for multimedia satellite networks, *IEEE Commun. Mag.* **39**(3): 192–198 (March 2001).
7. Y. Hu and V. O. K. Li, Satellite-based Internet: A tutorial, *IEEE Commun. Mag.* **39**(3): 154–162 (March 2001).
8. I. Minei and R. Cohen, High-speed Internet access through unidirectional geostationary satellite channels, *IEEE J. Select. Areas Commun.* **17**(2): 345–359 (Feb. 1999).
9. H. D. Clausen, H. Linder, and B. Collini-Nocker, Internet over direct broadcast satellites, *IEEE Commun. Mag.* **37**(6): 146–151 (June 1999).
10. M. Zeng, A. Annamalai, and V. Bhargava, Harmonization of global third-generation mobile systems, *IEEE Commun. Mag.* **38**(12): 94–104 (Dec. 2000).
11. W. Mohr and W. Konhauser, Access network evolution beyond third generation mobile communications, *IEEE Commun. Mag.* **38**(12): 122–133 (Dec. 2000).
12. R. Steele, C. C. Lee, and P. Gould, *GSM, cdmaOne and 3G Systems*, Wiley, Chichester, UK, 2001.
13. F. Priscoli, UMTS architecture for integrating terrestrial and satellite systems, *IEEE Multimedia* **6**(4): 38–44 (Oct.–Dec. 1999).
14. V. Wong and V. Leung, Location management for next-generation personal communications networks, *IEEE Network* **14**(5): 18–24 (Sept./Oct. 2000).
15. J. Solomon, *Mobile IP: The Internet Unplugged*, PTR Prentice-Hall, Englewood Cliffs, NJ, 1998.
16. N. Fikouras, K. Malki, S. Cvetkovic, and C. Smythe, Performance evaluation of TCP over Mobil IP, *Proc. Int. Conf. PIMRC'99*, Osaka, Japan, Sept. 1999.
17. A. Mehrotra and L. Golding, Mobility and security management in the GSM system and some proposed future improvements, *Proc. IEEE* **86**(7): 1480–1497 (July 1998).
18. I. Akyildiz, J. NcNair, J. Ho, H. Uzunalioglu, and W. Wang, Mobility management in next generation wireless systems, *Proc. IEEE* **87**(8) (August 1999).
19. H. Uzunalioglu, Probabilistic routing protocol for low earth orbit satellite networks, *Proc. Int. Conf. ICC'98*, Atlanta, GA, June 1998.
20. R. Goyal et al., Traffic management for TCP/IP over satellite ATM networks, *IEEE Commun. Mag.* **37**(3): 56–61 (March 1999).
21. C. Partridge and T. J. Shepard, TCP/IP performance over satellite links, *IEEE Network* **11**(5): 61–71 (Sept./Oct. 1997).
22. T. R. Henderson and R. H. Katz, Transport protocols for Internet-compatible satellite networks, *IEEE J. Select. Areas Commun.* **17**(2): 326–344 (Feb. 1999).
23. I. Minei and R. Cohen, High-speed Internet access through unidirectional geostationary satellite channels, *IEEE J. Select. Areas Commun.* **17**(2): 345–359 (Feb. 1999).
24. G. Xylomenos, G. C. Polyzos, P. Mahonen, and M. Saaranen, TCP performance issue over wireless links, *IEEE Commun. Mag.* **39**(4): 52–58 (April 2001).
25. *ASTRA Return Channel System, System Description Documentation*, Societe Europeenne des Satellites, Document ARCS.240.DC-Eoo1-0.2, issue 0.2, May 1998.
26. S. Karlin, Take off, plug in, dial up, *IEEE Spectrum* 52–59 (Aug. 2001).

SCALAR AND VECTOR QUANTIZATION

TIMO KAUKORANTA
Turku Centre for Computer Science (TUCS)
University of Turku
Turku, Finland

1. INTRODUCTION

Although the term *quantization* is not commonly used in our everyday life, the phenomenon itself is familiar to everyone. There the word *quantization* is known as *rounding*. We round the ages of the people to exact numbers of years even if the accurate age is a real number, which can have infinite decimal places. We apply rounding to several measurements in our life such as weights, distances, and periods of time. The purpose of rounding, or quantization, is to make these measurements easier to understand and handle. The measurements in these everyday examples are *scalars*, namely, single numbers, which present the amount of some quantity. This technique of rounding scalars is in fact known as *scalar quantization* (SQ).

Quantization can be considered as a technique for analog-to-digital conversion and data compression. By *quantization*, an originally analog signal, for example audio or video, is transformed into a digital form. This is very important operation because of the well-known benefits that digitization offers for signal processing. As

to the compression, speech and image data demand huge amounts of storage space so that many applications, such as in the fields of medicine or data communication, suffer from the lack of efficiency in coding of data.

Analog-to-digital conversion can be performed by SQ. There the original signal is sampled at some frequency f producing a sequence of samples x_i. These samples are then quantized so that they attain values from a set of predetermined *reproduction values* or *points* c_j. While the set of sample values may be unlimited, the size of the reproduction values is small and fixed. A unique binary word (code) w_j, which can simply be the index j of the reproduction value, is assigned to each reproduction value. This codeword is stored or transmitted instead of the original sample x_i.

In many applications we must store and transmit structured collections of data instead of scalars. If these data can be treated as vectors of fixed dimensionality, we can, instead of using the exact values of the vectors, handle their approximations. This is much like in the rounding process for individual scalars, but now we speak about *vector quantization* (VQ). Our hope is again to get advantage in transmission times and the storage space.

Generally, on the basis on the ability to preserve the information, compression techniques can be classified to *lossy* and *lossless*. A lossless compression technique reconstructs information exactly as it was before compression. Instead, a lossy compression technique can produce distortion to information, but the idea is that original information can be reproduced at sufficient accuracy, that is, that no necessary information is missing. The benefit of lossy compression is that it is possible to reach better compression ratios than by using lossless techniques.

Both scalar and vector quantization use fixed number of reproduction levels. Because the number M of these levels is normally much smaller than the number of all different inputs, the quantizer can be interpreted as a compression technique. The number of bits used to represent the binary word w_j may be a fraction of the number of bits needed to represent the corresponding sample of the original signal. On the other hand, quantization clearly loses some information; all possible input values cannot be reproduced. Therefore, quantization is a lossy compression technique.

It is typical that the input signal of a vector quantizer has already been scalar quantized. Single sample values (from the point of view of the vector quantizer) are organized as a sequence, and a fixed number of successive samples are grouped together to form vectors. The order of the samples in the sequence depends on the original source signal; in the case of one-dimensional signals such as audio, it is convenient to use time order. From images, which are two-dimensional inherently, $k \times k$-pixel blocks are often extracted and their pixels are assigned to vector elements in row-major order. In the case of color images, the RGB (red-green-blue) pixels can be interpreted as three-dimensional vectors.

VQ applies the same principles as SQ. In the encoding phase, the *vector quantizer* groups the values of the source signal into K-dimensional input vectors x_i. An appropriate transformation can be applied to the vectors. Then, for a given input vector x_i, the nearest *representative vector* c_j is searched from the *codebook* C of M K-dimensional *codevectors*. We suppose that the codebook, which is known by both the encoder and decoder, has been constructed before quantization as a separate process. The index, or binary word w_j, of the selected codevector c_j is stored or transmitted to the receiver. Compression rate can be improved by applying entropy coding to the indices. In the decoding phase, the receiver selects the representative vector from the codebook C using the transmitted and decoded index. A new reproduced signal is then constructed from the representative vectors. Compression is achieved because the cardinality of the set of the representative vectors is much smaller than the cardinality of the whole vector set. The encoding/decoding process is described as a sequence of successive steps in Fig. 1.

It is impossible to cover all the rich topics in scalar and vector quantization in a single article. We therefore restrict ourselves to topics, which we feel are important to a practitioner or a student in algorithmics who is interested in understanding the basic ideas in quantization methods. Our focus is more biased to the implementation issues whereas the theoretical consideration is largely omitted. This does not mean that the value of theory, including, for example, Shannon's source coding theory or Bennett's approximation on the quantization distortion, could be bypassed when evaluating different quantizers. These are dealt in depth, for example, in textbook of Gersho and Gray [1].

2. SCALAR QUANTIZATION

In scalar quantization we are given a set of numbers X drawn from a known distribution $f(x)$, and our task is to encode the numbers with a lower number of bits in such a way that the average distortion caused by this action is as small as possible. By the *distortion* we mean the difference of the original data x and the corresponding reproduction values $Q(x)$ obtained by the encoding and decoding sequence. When the original numbers are from $\Re$ (the set of real numbers), a natural choice of the distortion is the squared one-dimensional Euclidean distance $d(x, Q(x)) = (x - Q(x))^2$. Now it is clear that the distribution $f(x)$ has a profound influence on the proper choice of the quantization scheme. We should make more accurate quantization at the regions of high density than at the low-density regions.

The simplest way of thinking is to use a fixed number of bits for expressing the indices of the reproduction values $Q(X)$. If this number is M, we need $\log M$ bits for expressing each of them. This kind of method is called *fixed-rate coding*. Now one can fix M and ask for the best way (in distortion sense) to quantize the source data. This corresponds to minimization of the *distortion-rate function*. Another way of looking at the situation is to minimize M when the maximal distortion of the scalar quantizer has been fixed. Nothing restricts the coding to fixed-length codes. We can as well code the indices of the reproduction points by a *variable-length code*, in which case we may spend fewer bits to indices occurring more

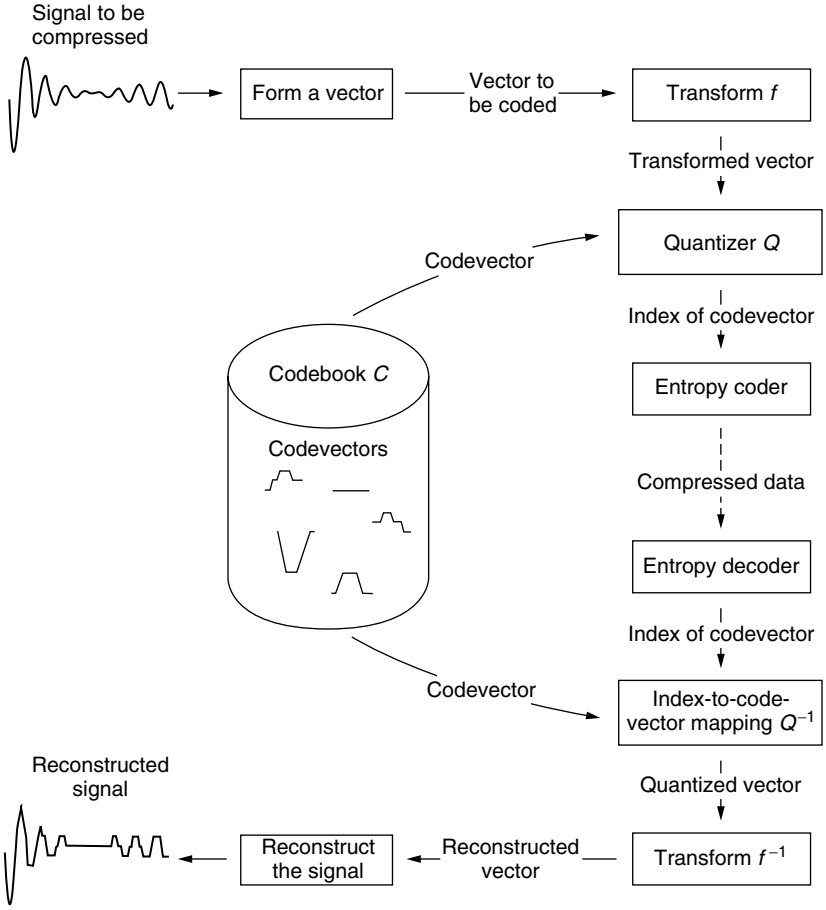


Figure 1. An example of a compression system based on (vector) quantization.

frequently. Note that here the coding must naturally be lossless (like Huffman code); compare with the encoding phase of the scalar quantizer, which is a lossy process. While the basic idea in SQ is the independent quantization of individual scalar data, there is no reason why we could not use entropy coding with some suitable context for the lossless compressor of the indices. By this way, we utilize the knowledge of the dependencies of individual samples on the history of the data when predicting the current quantization index. Then we code (e.g., by arithmetic coding) the difference between the actual and predicted index values.

2.1. Uniform Versus Nonuniform Quantizer

Quantizers can be divided into *uniform* and *nonuniform* depending on whether the reproduction values are at equal distances from their neighbors. A uniform scalar quantizer has an equal space $\Delta \in \Re$ between successive reproduction levels $c_i (i = 1, 2, \ldots, M)$, namely, $c_{i+1} = c_i + \Delta$. Correspondingly, the cell boundaries have the same space between them: $t_{i+1} = t_i + \Delta$. An example of a uniform quantizer is presented in Fig. 2. For this kind of quantizers one can show that the Euclidean squared distortion is of the size [2]

$$D_{un} \cong \frac{\Delta^2}{12} \tag{1}$$

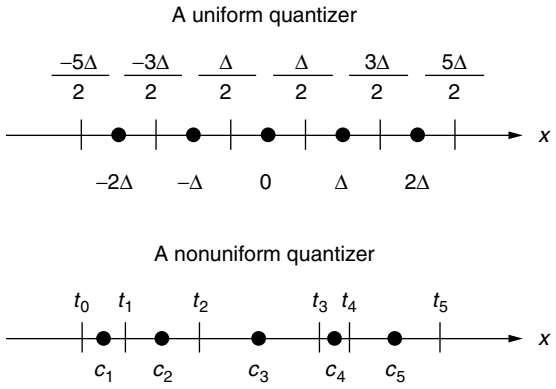


Figure 2. Illustration of a uniform and nonuniform quantizer.

The intervals of a nonuniform quantizer are of different lengths (see Fig. 2). This reflects the varying density of the samples in different regions of $\Re$, giving the overall distortion for samples at the range $[t_0, t_n]$

$$D_n = \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} d(x, c_i) f(x)\, dx \tag{2}$$

Several options for SQ are provided in the JPEG 2000 standard [3].

2.2. Lloyd's Conditions and Algorithm

Lloyd [4] showed two necessary optimality conditions for a scalar quantizer with fixed-length codewords to be optimal: (1) for a fixed setting of reproduction levels the cell boundaries must be selected optimally and (2) for a fixed setting of cell boundaries the reproduction levels should be selected optimally. This important observation can be turned to an algorithm (*Lloyd's algorithm*):

1. Initialize M cells and reproduction levels.
2. Repeat until the distortion converges:
   a. For each sample define the nearest centroid and place the sample to the set of samples mapping to this interval.
   b. Calculate new centroids for each interval by using the sets defined above.

In certain cases (e.g., for Gaussian and Laplacian densities of the samples), Lloyd's algorithm gives a globally optimal quantizer whereas there are applications where a local optimum will be found. Lloyd's algorithm is generalized and considered more closely in Section 4.1.

3. VECTOR QUANTIZATION

The goal and principle of vector quantization (VQ) are the same as in scalar quantization (SQ). The difference is in the nature of the samples. These are for VQ structured collections of scalars organized as vectors. One can separate four main tasks in the design of a vector quantizer: (1) *selection of a training set*, (2) *codebook generation* on the basis of the training set, (3) *encoding* of the source vectors with the aid of the codebook, and (4) *decoding* of the compressed vectors. The decoding commonly is the easiest of these tasks because it can often be performed extremely rapidly and simply by using a lookup table. The two main problems in the design of a vector quantizer are the construction of the codebook and efficient search from the codebook at the encoding phase. These problems are related because fast and accurate search methods are needed in codebook generation also.

A vector quantizer can be defined as a mapping Q of K-dimensional (Euclidean) space $\Re^K$ into a finite subset C of $\Re^K$:

$$Q: \Re^K \rightarrow C \tag{3}$$

where $C = (c_i; i = 1, 2, \ldots, M)$ is the set of *reproduction vectors* (i.e., the *codebook*) and M is the number of codevectors in C (see Fig. 3). Thus, a vector quantizer can be seen as a combination of two functions. The *encoder* produces an index i for input vector x. The *decoder* transforms the index i back to the reproduction vector c_i. As in the case of SQ, we may still have a lossless compressor of the indices. This may again produce a variable-length code, in which case we have a variable-rate VQ. One can also integrate the compression technique to the vector quantizer itself. Entropy-constrained VQ is one example of this kind (see Section 3.4), but other approaches exist also.

VQ has several benefits as a compression technique. The most important of these is the high decoding speed. By changing the number of codevectors in the codebook, the bit rate can be controlled quite easily. A drawback of VQ is that the vector quantizer has to be trained over a representative set of samples of the signal. Thus, a vector quantizer does not perform well for input vectors emitted by a source, which has characteristics very different from those of the original training set. The method suffers also from relatively high computational complexity both in the codebook generation and in the encoding of a signal.

3.1. Transforms

A suitable transform can be applied to vectors before the use of a vector quantizer. The purpose of the transform is to organize data in a vector in such way that important information is concentrated on some elements of the vector. This transformed representation is probably more suitable for quantization and the search for the nearest vector. For example, it may be reasonable to normalize the vectors by subtracting from each element their average. Another way to normalize a vector is to divide its elements by the norm of the vector. Also other transformations, such as the *discrete*

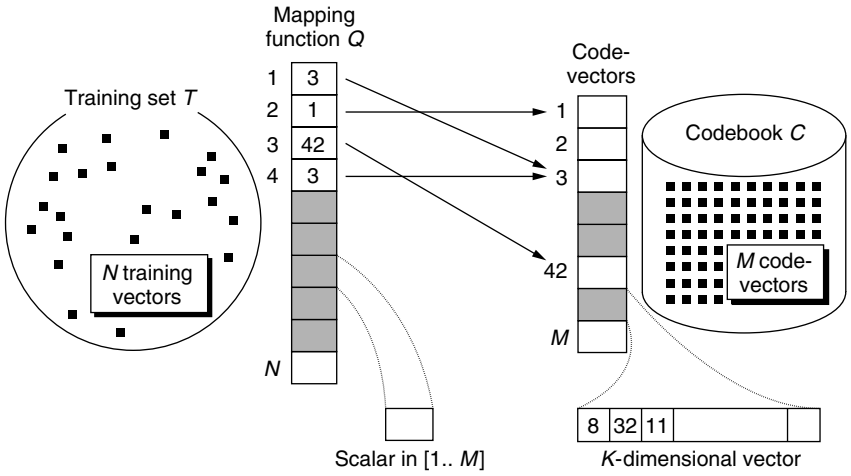


Figure 3. Roles of a training set and a codebook in vector quantization.

cosine transform and the *Walsh–Hadamard transform*, are used in applications. Transforms are often used in *product codevector quantization* (see Section 3.4). For general properties of transform based vector quantizers, see Ding's paper [5].

In some cases, the vectors to be quantized are sequences of the coefficients produced by a signal transformation technique, while in other cases, they may be simply blocks of individual pixels of a digitized image. (Note that this latter case clearly demonstrates the structural dependence of the elements in a vector.) Because of its fast decoding phase, VQ is suitable for compression situations demanding fast reproduction and therefore it has been popular in video codecs.

3.2. Quality and Resolution

The quality of a vector quantizer Q can be measured by the distortion between the original input vector x_i and the reproduction vector $y_j = Q(x_i)$. The smaller the distance, the better the quality. The most widely used distortion measure is the squared error or squared Euclidean distance defined by

$$d(x_i, y_j) = \|x_i - y_j\|^2 = \sum_{k=1}^{K} (x_{i,k} - y_{j,k})^2 \tag{4}$$

where $x_{i,k}$ and $y_{j,k}$ denote the kth element of the vectors x_i and y_j, respectively. Some other distortion measures have been proposed [1,6,7]. The quality of the quantizer is the average distance between the sample vectors and the corresponding reproduction vectors. This average can also be weighted for different components.

The distortion (quality) D of a codebook C can be measured in relation to the training set X for which the codebook has been constructed:

$$D(C, X) = \sum_{i=1}^{N} d(x_i - Q_C(x_i)) \tag{5}$$

Here N is the number of training vectors in the training set X and $Q_C(x_i)$ gives the nearest codevector in the codebook C. In other words, $Q_C(x_i)$ is an optimal mapping that minimizes the distance $\mathrm{d}(x_i, y_j)$, where $y_j \in C$.

There is one question largely omitted in VQ literature, namely, the representativeness of the training set. The VQ codebook is constructed on the basis of a training set of N samples, which are supposed to have the same statistical properties, the density, as the data to which VQ is later applied. A normal way to assert this is to evaluate the operation efficiency of the VQ codebook by an independent test set. Even here caution should be taken to the later changes of the source distribution.

The *resolution* (*code rate* or simply *rate*) of a fixed-rate vector quantizer is

$$r = \frac{\log_2 M}{K} \tag{6}$$

which measures the number of bits per vector element used to represent the K-dimensional input vector. The resolution gives an approximation of the quality of the reproduction of the vector quantizer. One problem of VQ is the *complexity barrier*, which is described as follows. If we limit the resolution r to a fixed value, we can increase the performance of the vector quantizer only by increasing the vector dimensionality K. The reason for this improvement of the performance is that long vectors express the statistical dependencies of the signal more extensively than short vectors. On the other hand, the amount of memory needed to store the codebook and the search complexity (operations per vector element in exhaustive search) are both relative to KM. Thus the space complexity, which is given by

$$KM = K2^{rK} \tag{7}$$

grows exponentially with dimension K. If we suppose that the encoding phase includes an exhaustive search of the codebook, the time complexity of the encoding is exponential on K, too.

3.3. Encoding and Decoding in Unstructured Vector Quantization

In *unstructured vector quantization*, the codebook is just a set of codevectors (reproduction vectors). In *structured vector quantization*, a fast encoding process is pursued by utilizing a special structure of the quantizer. Several structured vector quantizers are discussed in the next section. Here we concentrate on unstructured vector quantizers.

The decoding in unstructured VQ is performed by a table lookup and therefore is very fast and simple. Unfortunately, the encoding is not an easy task. In the encoding, the representative vector for the K-dimensional input vector is searched from the codebook of M codevectors. This *nearest-neighbor problem* is known in other applications, too. Search techniques can be classified into two groups: *exact methods* and *approximate methods*. The former ones always find the nearest codevector, whereas the approximate methods select a vector that is reasonably close and can be found quickly.

Several fast exact search techniques have been introduced to replace the exhaustive search. These techniques typically rely on the properties of the Euclidean vector space. Because the exhaustive search takes $O(MK)$ operations, it is therefore natural to try to reduce the effect of either M or K, or even both. *Partial distortion search* [8], *mean-distance-ordered partial search* [9,10], and *triangular inequality elimination* [11] are widely known methods for the exact search. The importance of efficient search techniques is emphasized by the fact that these techniques are also needed in the codebook generation algorithms, which are discussed later.

In an exact search, the approximate nearest-neighbor search is its own research field also. These techniques often guarantee some property, for example, that the selected codevector is at most at the distance ε from the input vector. One example of approximate techniques is the *tree-structured search*, in which the codebook is organized as a (binary) search tree.

3.4. Vector Quantization Structures

In the case of unstructured VQ, the codebook is just an array of codevectors. In addition, it is typical that the training vectors have been formed from the raw signal data without any special transformations. However, one should keep in mind that there are several alternatives for this basic organization of VQ [1,12,13]. We briefly discuss some of those structures in the following paragraphs.

In several compression methods, *prediction techniques* are utilized. The value of the next sample is predicted by the sample values, which have been coded thus far. The prediction is subtracted from the original pixel value giving a *prediction error* or *residual*, and then the residual is coded by a suitable technique. The same idea is exploited in *predictive vector quantization* (PVQ) [1] by predicting the next vector. The prediction is based on the previously coded vectors so that the decoder is capable of forming the same prediction. The prediction vector is then subtracted from the input vector to form a difference vector, which is finally quantized. Thus, the difference to basic schema of VQ is that here VQ is applied to residual vectors instead of pure vectors in the source domain. The design of a predictive vector quantizer [14–16] includes the design of the predictor in addition to the design of the codebook. These two are dependent on each other, which makes the whole process difficult. Basic approaches are *open-loop*, *closed-loop*, and *semi-closed-loop* designs, [1].

Finite-state vector quantization [1] uses several separate codebooks. The method is based on the assumption that the value of the next input vector depends on the values of the previous vectors. Therefore, a finite-state automaton is constructed to model the dependencies of the vectors. The encoding starts from an initial state of the automaton. In each state, a separate codebook is used for the quantization of the current input vector. The next state is selected on the basis of the reproduction vector. This approach allows us to construct a suitable quantizer for each particular vector context.

In *classified vector quantization* [17], an L-level classifier is used to select a subcodebook for the quantization of the input vectors. First, the encoder transmits the index of the proper subcodebook to the decoder and then the actual index of the selected codevector. The subcodebooks are designed separately and they can be of different sizes. There are several alternatives for the design of the classifier. One possibility is that the classifier is a common vector quantizer, which performs L-level clustering of the vector space. Statistical properties of the input vector can be utilized also. For image compression, the classifier can recognize the pixel blocks as *shade*, *midrange*, *edge*, or *mixed* blocks [17]. This approach has been inspired by the observation that there are too few edge vectors in a single codebook, which has been optimized according to the Euclidean distance. Therefore, the classified vector quantizer tries to guarantee a sufficient number of vectors for those image areas where errors are most annoying (cf. edge areas of the image). Other classified vector quantizers have been discussed [6,14,18].

Operating with high-dimensional vectors slows down both the encoding and the construction of the codebook. *Product code techniques* [1,13] are one way to relieve this problem. The idea is to divide the input vector into a collection of *subvectors*. Each of these describes a certain property of the vector, and therefore one can design a separate codebook for each property. The assumption is that the subvectors are easier to quantize because they take values from a smaller range of the K-dimensional space or have lower dimensionality. In particular, the dimension of a subvector can be one, in which case the property is a scalar. The encoder divides an input vector into a set of subvectors, which are quantized separately, and the corresponding indices are sent to the decoder. The decoder reproduces the subvectors from the indices and constructs a joint reproduction vector from these subvectors. This technique is called *product code vector quantization* because the whole (effective) reproduction codebook is a Cartesian product of all codebooks of the subvectors. Thus, when we have V subcodebooks with M_i codevectors in each, the number of possibilities to construct a reproduction vector is as large as

$$M = \prod_{i=1}^{V} M_i \tag{8}$$

but the storage space and encoding complexities are only of the magnitude

$$\tilde{M} = \sum_{i=1}^{V} M_i \tag{9}$$

The mean-removed vector quantizer [1] is one example of product codevector quantizers. It divides the input vector into two subvectors: a scalar that contains the average of the elements of the *input vector*, and a *difference vector*, where the average has been subtracted from each element of the input vector. *A shape–gain vector quantizer* [1,19] divides the input vector into the *gain* and *shape vectors* similarly. The gain is a scalar, whose value is the Euclidean norm (i.e., length) of the input vector. The shape vector consists of the elements of the input vector divided by the gain.

In *residual* or *multistage vector quantization* [16,20,21], the encoding process is divided into successive applications of separate vector quantizers. In the beginning, the input vector is coded by the first-stage quantizer. A residual vector is formed from the difference between the input vector and its reproduction vector. The second-stage quantizer is then applied to the residual vector. Again, a new residual vector is formed and the same procedure is iterated. By repeating this operation L times, we have an *L-stage vector quantizer*. A separate codebook is generated for each stage. The decoder constructs the whole reproduction vector by summing up the codevector of each quantizer. An upper bound for the number of separate reproduction vectors is the same as for the product codes [see Eq. (8)]. The complexity of the storage and encoding are also the same [see Eq. (9)].

A *lattice vector quantizer* [1,13] is simply a vector quantizer whose codebook is constructed from a lattice. Here the codevectors have a regular arrangement in the K-dimensional vector space. In fact, the lattice vector quantizer can be interpreted as a generalization of a

uniform scalar quantizer. An essential parameter in the design of a lattice vector quantizer is the density of the lattice, which defines how many lattice points per unit volume of the space are contained in the codebook. The higher the density, the higher the bit rate and the smaller the average distortion. The benefit of the lattice vector quantization is that special search methods can be applied because of the regular structure of the codevectors. In addition, the codebook can be expressed by its parameters instead of listing all the codevectors separately.

The output of the vector quantizer can be compressed further by applying entropy coding. Since the entropy codes can reduce the average length of the final code to the entropy of the vector, the vector quantizer should be designed to minimize its output entropy. In the design of an *entropy-constrained vector quantizer* [20,22–24], the task is to generate a codebook, which minimizes the average distortion between the source vectors and reproduction vectors subject to a constraint on the index entropy. Thus the distance between a training vector x_i and a codevector c_j is defined as

$$d(x_i, c_j) = \|x_i - c_j\|^2 + \lambda r_j \tag{10}$$

where r_j is the entropy of the index of codevector c_j, and λ is a weighting parameter for the entropy.

A vector quantizer is typically designed for a particular distribution of source vectors. However, it is possible that the input vectors given to the encoder are from a different distribution of vectors, and therefore the vector quantizer is unable to code these vectors well. In *adaptive vector quantization*, this problem is relieved by modifying the vector quantizer during the coding process. The predictive and finite-state vector quantizers could be considered as adaptive vector quantizers, because they change their encoding rule in time. However, usually, methods that change the codebook in order to match the local properties of input vectors are called *adaptive vector quantizers*.

4. CODEBOOK GENERATION

In the codebook generation problem, the task is to construct a codebook C that minimizes distortion D of Eq. (5) for a given training set X. If the VQ system transfers the codebook to a receiver as a part of compressed data, it is natural to form the training set from the source signal to be compressed. Instead, when the system uses a *static codebook*, which is known by both the encoder and the decoder, the training set is formed from a large set of similar signal samples as the signal source to be compressed. When generating a static codebook, its quality has to be determined against input vectors outside the training set. This process and the construction of training sets have been studied [25,26]. To exclude trivial problem instances, we suppose that the size M of the codebook is smaller than the size N of the training set (in practice we assume that $M \ll N$).

In addition to the distortion D of the produced codebook, the quality of a codebook generation algorithm can be characterized by other attributes. *Robustness* of the algorithm describes how independent the output of the algorithm is from the initial conditions and parameter setups. Another important property of the algorithm is small *running time*; especially in online applications. A *scalable* algorithm is able to improve the quality of the codebook with additional computing resources. *Memory requirements* should be reasonable, so that the algorithm would also work for large problem instances.

The codebook generation problem resembles the *clustering problem* [27], in which the task is to classify N input vectors into M clusters. However, the quality of the clustering is measured both by evaluating the *similarity* of objects (vectors) in the same cluster and *dissimilarity* of the objects in different clusters. Instead, in codebook generation, the task is to find a good set of *representatives* for the input vectors. In addition, the number of codevectors M is usually given in the setup of the codebook construction, whereas in clustering the search for the right number of clusters may be a vital part of the problem. The codebook generation problem also resembles the *P-median problem*, which differs from VQ in that the solution vectors of the P-median problem are limited to be vectors of the training set. A similar approach has been also applied to codebook generation in VQ by modifying the generalized Lloyd algorithm (GLA). Codebook generation for RGB-tuples is known as a *palette generation problem* in color image quantization [28,29].

Construction of an optimal codebook is a combinatorial optimization problem and it is NP-hard [30]. In other words, there is no known polynomial time algorithm for finding the globally optimal solution. However, reasonable suboptimal solutions are typically obtained by *heuristic algorithms* [31,32]. Methods to solve the codebook generation problem can be divided into *problem specific methods* and *general optimization methods*. Problem-specific methods have been developed particularly to solve the codebook generation problem (and the clustering problem). General optimization methods, however, are suitable for all optimization problems on the condition that the problem has been formulated properly for a particular optimization method.

In the following, we concentrate on codebook construction methods for unstructured vector quantizers. However, after minor modifications, these general methods are suitable for design of other vector quantizers also. By *iterative method*, we mean a method that attempts to improve the quality of an existing solution (codebook). Therefore, these methods need an initial codebook, which is feasible for the given problem. Examples of iterative methods are GLA and genetic algorithms, which are discussed later. *Hierarchical methods* are problem-specific and are widely used in clustering problems. Hierarchical methods can be classified as *divisive* and *agglomerative* types, and they repeatedly split or merge clusters until a clustering of the desired size has been reached. The splitting method and the pairwise nearest-neighbor (PNN) algorithm are two examples of hierarchical methods, which are discussed later. In the following three sections we briefly describe problem-specific methods. After that, we also briefly discuss general optimization methods.

4.1. Generalized Lloyd Algorithm

The *generalized Lloyd algorithm* (GLA) [33] is perhaps the most widely known and used method for codebook generation. The algorithm is based on the iterative use of the codebook modification operation generalized from the Lloyd's algorithm for SQ. The GLA is popular also in the context of clustering and pattern recognition, where it is known as a *C-means algorithm* [34]. C refers here to the number of clusters (codevectors).

The codebook modification operation is based on two optimality conditions: *nearest-neighbor condition* and *centroid condition*. These conditions describe how to generate an optimal clustering for a given codebook, and vice versa, how to generate the optimal codebook for a given clustering.

- *Nearest-Neighbor Condition*. For a given codebook, the optimal clustering of the training set is obtained by mapping each training vector to its nearest codevector in the codebook with respect to the distortion function.
- *Centroid Condition*. For a given cluster, the optimal codevector is the *centroid* (average vector) of the vectors within the cluster.

The GLA applies the two optimality conditions in turn. In the *clustering step*, the training set is grouped into clusters according to the existing codebook. The optimal clustering is obtained by mapping each training vector to the nearest codevector. In the *codebook step* a new codebook is constructed by calculating the centroids of the clusters defined in the clustering step. The two optimality conditions guarantee that the new solution is always equal to or better than the previous one. However, it should be noted that although these two steps are locally optimal, the whole process does not necessary produce an optimal codebook. Most of the computational burden of the GLA originates from the clustering step, which can be expedited by several techniques [35].

After the clustering step, it may happen that the codebook contains some codevectors to which no training vectors have been mapped, that is, that do not represent any of the training vectors. This so *empty-cluster problem* is usually solved by replacing the codevectors of empty clusters by some existing training vectors [1].

The GLA can be iterated until the quality of the codebook does not improve anymore. Another *stopping condition* is to require the relative improvement to be higher than a given threshold. The relative improvement can be measured by

$$\Delta D = \frac{D(C^{(t)}, X) - D(C^{(t+1)}, X)}{D(C^{(t)}, X)} \tag{11}$$

where $C^{(t)}$ refers to the codebook on the tth iteration round and $D(C, X)$ is the distortion of coding training set X with C. The process can also be limited to a certain number of iterations.

The GLA is a *descent* algorithm, which means that each iteration decreases (or at least never increases) the distortion. Because of the deterministic nature of the GLA, the process leads to a *local optimum*, which depends on the *initial codebook*. Because of this, the GLA is unable to locate a *globally optimal codebook*. The GLA can be seen as a *fine-tuner* of a codebook, and it has been integrated in many other algorithms.

The GLA tries to improve the quality of an existing codebook, and therefore it needs an initial codebook to start with. Techniques for generating an initial codebook are typically fast and simple; at least faster than the GLA itself. One such technique, which is often satisfactory, is to use M random training vectors as the codevectors. Several other techniques have been proposed in the literature [1,33].

4.2. Splitting Method

The splitting method [36] starts from a *singular codebook* containing a single codevector, which is the centroid of the whole training set. The codebook is then enlarged hierarchically by splitting a cluster (represented by a codevector) into two new clusters. This procedure is repeated until the codebook reaches the desired size. The selection of a codevector to be split is based on the properties of its cluster. For example, this property may be the *variance* of training vectors in the cluster. The problem of selecting the cluster is avoided in *binary splitting* [33], where all current clusters are split. However, this approach may allocate too many codevectors to sparse areas of the vector space.

Splitting of a given cluster is a special codebook generation problem ($M = 2$), where the task is to find two new representative codevectors for a given subcluster. We have two approaches to this problem. The *partition-based* approach divides the training vectors of the cluster into two subclusters and uses the two centroids of the subclusters as new codevectors. In the *codevector-based* approach, two codevectors are selected by some heuristic and the training vectors are mapped to them.

The overall structure of the splitting method is easy to understand. However, internal components of the method can be rather complicated. The selection of these components gives us a versatile way to generate codebooks of varying quality with controllable running time. The top–down process of the splitting method has the benefit that it can be used also for the construction of a *tree-structured vector quantizer* [37].

4.3. Pairwise Nearest Neighbor

The pairwise nearest-neighbor (PNN) algorithm [38] belongs to the class of *agglomerative clustering methods*. In the field of clustering research the PNN is often called *Ward's method* [39]. Because the cluster centroids can be easily used as codevectors of the codebook, the PNN is suitable for codebook construction. The algorithm starts by constructing an initial codebook in which each training vector is considered as its own codevector. Two nearest codevectors, whose merging increases the total distortion least, are merged at each step, and the process is repeated until the desired size of the codebook has been reached.

The algorithm is straightforward to implement in its basic form, and in comparison to the GLA, it gives good-quality results (i.e., codebooks with small distortion). The PNN has also the advantage that the hierarchical approach produces codebooks of differing sizes as a side product. Thus, the PNN can be easily applied to joint minimization of distortion and entropy of codevector indices [23,24,40]. The algorithm can also be used as a part of hybrid methods such as a genetic algorithm [41], or an iterative split-and-merge method [42].

A drawback of the PNN is the relatively high running time in its exact form [43]. There is a large number of steps to be performed by the algorithm, because typically we have $M \ll N$, and at each step, all pairwise distances are calculated for finding the pair of vectors to be merged. This makes the algorithm very slow for large training sets. It should also be noted that although a single merge operation is always performed optimally (i.e., two nearest clusters are merged), the whole process does not guarantee an optimal codebook of the size M.

Approximate PNN variants [44,46] reduce the number of distance calculations by relieving the condition of merging two nearest clusters. In the *K-d tree* variant [38], the nearest cluster is searched from a (small) subset of clusters only, and several cluster pairs are merged at the same iteration round. This approach is clearly faster than the original one, but the quality of the codebooks is worse. Fast *exact PNN variants* have been proposed [45].

4.4. General Optimization Methods

Although the codebook generation problem is NP-hard [30], good solutions can be produced by general heuristic optimization methods [1,47]. In this section, we describe briefly some methods, which have been observed to work well for several other difficult combinatorial problems and have been successful for codebook optimization also. Their application to the generation of unstructured codebooks is straightforward.

Genetic algorithms (GAs) are optimization techniques derived from the model of the natural selection in real-life evolution. The idea is that a *population* of possible solutions (codebooks) called *individuals* is first (randomly) generated. The next generation consists of the survivors and of new individuals generated from the individuals of the former population by genetic operations such as *crossover* and *mutation*. As in the real life, the genes of the best-fitted individuals survive.

A problem of several optimization methods is that the methods may stick in a local minimum. *Tabu search* [48,49] tries to avoid this problem by keeping the previous solutions in a *tabu list*. The method starts from an initial solution by generating a set of candidate solutions. Solutions, that appear in the tabu list are discarded. The best of the remaining candidates is selected as a new solution, and it is inserted in the tabu list. The procedure is repeated until a stopping condition is fulfilled. The purpose of the tabu list is to prevent the search from returning to solutions that have been evaluated recently. This forces the search to proceed into new directions instead of sticking in a local minimum and in its neighborhood.

Stochastic relaxation (SR) is a family of optimization techniques, which add perturbation to the current solution at each iteration. The amount of perturbation decreases with time, making convergence possible. Because the method allows an increase of the distortion, it can continue the search after reaching a local minimum. A popular variant of SR algorithms is known as *simulated annealing* (SA) [50,51]. The solutions, which decrease the distortion, are always accepted in the SA. However, the acceptation of a solution, which increases the distortion, is determined probabilistically. This probability depends on the change of the distortion value and the parameter called *temperature*, whose decrease rate is given by a *cooling schedule*.

Neural network algorithms used in learning applications are proposed for the codebook generation problem in Refs. 52 and 53. The idea is that the neural network learns the properties of the training set and the codebook is the result of this learning process. A popular method is the *self-organization feature maps* (SOM). The method starts from an initial codebook. One training vector is used to modify the codebook at the time. This is done by moving the training vector's nearest codevector and its neighboring codevectors toward the training vector. The size of the neighborhood reduces during the process and finally only the nearest codevector is moved. Because the method is based on the update of the existing codebook, it is also suitable for adaptive vector quantization.

5. SUMMARY

Scalar quantization is a basic technique for analog-to-digital signal transformation. It has an extensive theoretical background in addition to practical usefulness. One can augment the method with variable-rate compression techniques in order to decrease the storage space.

In theory, vector quantization offers very good possibilities for lossy signal compression. The method has the great advantage that its decoding runs extremely fast. Unfortunately, it includes other complexity problems. To obtain good compression results, one has to use codebooks with high dimensionality with respect to both the vector dimension and the number of reproduction vectors. This makes the encoding process slow and the memory consumption of the compression system unpractical.

Acknowledgment

The author would like to thank Professor O. Nevalainen for his helpful comments and support during this work.

BIOGRAPHY

Timo Kaukoranta received his M.Sc. and Ph.D. degrees in computer science from the University of Turku, Finland, in 1994 and 2000, respectively. He has been a researcher at the University of Turku from 1994 to 2000. Since 2001 he has been a postdoctoral researcher at Turku Centre for Computer Science (TUCS). His primary research interests are in distributed interactive simulations, multiplayer computer games, and vector quantization.

BIBLIOGRAPHY

1. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, Dordrecht, 1992.
2. W. R. Bennett, Spectra of quantized signals, *Bell Syst. Tech. J.* **27**: 446–472 (1948).
3. M. W. Marcellin et al., An overview of quantization in JPEG 2000, *Signal Process. Image Commun.* **17**: 73–84 (2002).
4. S. P. Lloyd, *Least Squares Quantization in PCM*, unpublished Bell Laboratories Technical Note; portions presented at the Institute of Mathematical Statistics Meeting, Atlantic City, NJ, 1957; published in special issue on quantization, *IEEE Trans. Inform. Theory* **28**: 129–137 (1982).
5. W. Ding, Optimal vector transform for vector quantization, *IEEE Signal Process. Lett.* **1**(7): 110–113 (1994).
6. B. Marangelli, A vector quantizer with minimum visible distortion, *IEEE Trans. Signal Process.* **39**(12): 2718–2721 (1991).
7. V. J. Mathews and P. J. Hahn, Vector quantization using the L_∞ distortion measure, *IEEE Signal Process. Lett.* **4**(2): 33–35 (1997).
8. C.-D. Bei and R. M. Gray, An improvement of the minimum distortion encoding algorithm for vector quantization, *IEEE Trans. Commun.* **33**(10): 1132–1133 (1985).
9. S.-W. Ra and J.-K. Kim, A fast mean-distance-ordered partial codebook search algorithm for image vector quantization, *IEEE Trans. Circuits Syst.-II: Analog Digital Signal Process.* **40**(9): 576–579 (1993).
10. S. Baek, B. Jeon, and K.-M. Sung, A fast encoding algorithm for vector quantization, *IEEE Signal Process. Lett.* **4**(12): 325–327 (1997).
11. S.-H. Chen and W. M. Hsieh, Fast algorithm for VQ codebook design, *IEE Proc.-I* **138**(5): 357–362 (1991).
12. N. M. Nasrabadi and R. A. King, Image coding using vector quantization: A review, *IEEE Trans. Commun.* **36**(8): 957–971 (1988).
13. R. M. Gray and D. L. Neuhoff, Quantization, *IEEE Trans. Inform. Theory* **44**(6): 2325–2384 (1998).
14. K. N. Ngan and H. C. Koh, Predictive classified vector quantization, *IEEE Trans. Image Process.* **1**(3): 269–280 (1992).
15. S. A. Rizvi and N. M. Nasrabadi, Predictive vector quantizer using constrained optimization, *IEEE Signal Process. Lett.* **1**(1): 15–18 (1994).
16. S. A. Rizvi and N. M. Nasrabadi, Predictive residual vector quantization, *Proc. IEEE Int. Conf. Image Processing*, 1994, pp. 608–612.
17. B. Ramamurthi and A. Gersho, Classified vector quantization of images, *IEEE Trans. Commun.* **34**(11): 1105–1115 (1986).
18. K. L. Oehler and R. M. Gray, Combining image compression and classification using vector quantization, *IEEE Trans. Pattern Anal. Mach. Int.* **17**(5): 461–473 (1995).
19. K. L. Oehler and R. M. Gray, Mean-gain-shape vector quantizer, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Minneapolis, MN, 1993, Vol. V, pp. 241–244.
20. F. Kossentini, M. J. T. Smith, and C. F. Barnes, Image coding using entropy-constrained residual vector quantization, *IEEE Trans. Image Process.* **4**(10): 1349–1357 (1995).
21. C. F. Barnes, S. A. Rizvi, and N. M. Nasrabadi, Advances in residual vector quantization: A review, *IEEE Trans. Image Process.* **5**(2): 226–262 (1996).
22. P. A. Chou, T. Lookabaugh, and R. M. Gray, Entropy-constrained vector quantization, *IEEE Trans. Acoust. Speech Signal Process.* **37**(1): 31–42 (1989).
23. D. P. de Garrido, W. A. Pearlman, and W. A. Finamore, Vector quantization of image pyramids with the ECPNN algorithm, *SPIE Proc. Visual Commun. Image Process.* **1605**: 221–232 (1991).
24. F. Kossentini and M. J. T. Smith, A fast PNN design algorithm for entropy-constrained residual vector quantization, *IEEE Trans. Image Process.* **7**(7): 1045–1050 (1998).
25. D. Cohn, E. A. Riskin, and R. Ladner, Theory and practice of vector quantizers trained on small training sets, *IEEE Trans. Pattern Anal. Mach. Int.* **16**(1): 54–65 (1994).
26. D. S. Kim, T. Kim, and S. U. Lee, On testing trained vector quantizer codebooks, *IEEE Trans. Image Process.* **6**(3): 398–406 (1997).
27. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
28. M. T. Orchard and C. A. Bouman, Color quantization of images, *IEEE Trans. Signal Process.* **39**(12): 2677–2690 (1991).
29. P. Scheunders, A comparison of clustering algorithms applied to color image quantization, *Pattern Recogn. Lett.* **18**: 1379–1384 (1997).
30. M. R. Garey, D. S. Johnson, and H. S. Witsenhausen, The complexity of the generalized Lloyd-Max problem, *IEEE Trans. Inform. Theory* **28**(2): 255–256 (1982).
31. C.-M. Huang and R. W. Harris, A comparison of several vector quantization codebook generation approaches, *IEEE Trans. Image Process.* **2**(1): 108–112 (1993).
32. N. Akrout, R. Prost, and R. Goutte, Image compression by vector quantization: A review focused on codebook generation, *Image Vision Comput.* **12**(10): 627–637 (1994).
33. Y. Linde, A. Buzo, and R. M. Gray, An algorithm for vector quantizer design, *IEEE Trans. Commun.* **28**(1): 84–95 (1980).
34. J. B. McQueen, Some methods of classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Math. Statist. Probability* 1, Univ. of California, Berkeley, 1967, Vol. 1, pp. 281–296.
35. T. Kaukoranta, P. Fränti, and O. Nevalainen, A fast exact GLA based on code vector activity detection, *IEEE Trans. Image Process.* **9**(8): 1337–1342 (2000).
36. P. Fränti, T. Kaukoranta, and O. Nevalainen, On the splitting method for VQ codebook generation, *Opt. Eng.* **36**(11): 3043–3051 (1997).
37. J. Lin and J. A. Storer, Design and performance of tree-structured vector quantizers, *Inform. Process. Manage.* **30**(6): 851–862 (1994).
38. W. H. Equitz, A new vector quantization clustering algorithm, *IEEE Trans. Acoust. Speech Signal Process.* **37**(10): 1568–1575 (1989).
39. J. H. Ward, Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.* **58**: 236–244 (1963).
40. D. P. de Garrido, W. A. Pearlman, and W. A. Finamore, A clustering algorithm for entropy-constrained vector quantizer design with applications in coding image pyramids, *IEEE Trans. Circuits Syst. Video Technol.* **5**(2): 83–95 (1995).

41. P. Fränti, J. Kivijärvi, T. Kaukoranta, and O. Nevalainen, Genetic algorithms for large scale clustering problem, *Comput. J.* **40**(9): 547–554 (1997).

42. T. Kaukoranta, P. Fränti, and O. Nevalainen, Iterative split-and-merge algorithm for VQ codebook generation, *Opt. Eng.* **37**(10): 2726–2732 (1998).

43. J. Shanbehzadeh and P. O. Ogunbona, On the computational complexity of the LBG and PNN algorithms, *IEEE Trans. Image Process.* **6**(4): 614–616 (1997).

44. T. Kurita, An efficient agglomerative clustering algorithm using a heap, *Pattern Recogn.* **24**(3): 205–209 (1991).

45. P. Fränti, T. Kaukoranta, D.-F. Shen, and K.-S. Chang, Fast and memory efficient implementation of the exact PNN, *IEEE Trans. Image Process.* **9**(5): 773–777 (2000).

46. T. Kaukoranta, P. Fränti, and O. Nevalainen, Vector quantization by lazy pairwise nearest neighbor method, *Opt. Eng.* **38**(11): 1862–1868 (1999).

47. C. R. Reeves, ed., *Modern Heuristic Techniques for Combinatorial Problems*, McGraw-Hill, UK, 1995.

48. K. Al-Sultan, A tabu search approach to the clustering problem, *Pattern Recogn.* **28**(9): 1443–1451 (1995).

49. P. Fränti, J. Kivijärvi, and O. Nevalainen, Tabu search algorithm for codebook generation in vector quantization, *Pattern Recogn.* **31**(8): 1139–1148 (1998).

50. J. K. Flanagan et al., Vector quantization codebook generation using simulated annealing, *Proc. ICASSP*, 1989, pp. 1759–1762.

51. S. Z. Selim and K. Alsultan, A simulated annealing algorithm for the clustering problem, *Pattern Recogn.* **24**(19): 1003–1008 (1991).

52. N. M. Nasrabadi and Y. Feng, Vector quantization of images based upon the Kohonen self-organizing feature maps, *Proc. IEEE Int. Conf. Neural Networks*, 1988, pp. 101–108.

53. N. B. Karayiannis and P.-I. Pai, Fuzzy algorithms for learning vector quantization, *IEEE Trans. Neural Networks* **7**(5): 1196–1211 (1996).

SECURE ULTRAFAST DATA COMMUNICATION AND PROCESSING INTERFACED WITH OPTICAL STORAGE

BAHRAM JAVIDI
University of Connecticut
Storrs, Connecticut

OSAMU MATOBA
University of Tokyo
Tokyo, Japan

1. INTRODUCTION

Pulseshapers of femtosecond laser pulses [1–6] are attractive in wide research fields such as optical communication, information processing, and spectroscopy. In an ultrafast data communication system, spatial data can be sent to remote users at an ultrafast rate, faster than terabit/s, via optical fiber communications. Because there is no device to handle temporal data directly at such an ultrafast rate, the pulseshaper is one of the most promising methods to send or process data. To send a large amount of data, it is also required to use optical storage systems that can readout data in parallel and at an ultrafast rate.

In this article, we present a secure ultrafast data communication system that can link remote users to an encrypted optical database with ultrafast transfer rate [7]. It is well known that holographic memories potentially have a large storage capacity with a fast data transfer rate [8,9]. Security communication is achieved by use of encrypted holographic memory. In the encrypted holographic memory, each data frame to be stored is encoded by optical encryption techniques, such as double-random-phase encryption [10,11] or by an exclusive-OR method. Spatial-temporal converters enable us to send the stored data in the encrypted holographic memory to the receiver at an ultrafast rate. At the remote users, the correct key is required to reconstruct the original data. Without the key information, it is very difficult to decrypt the data because optics can provide a high level of security. In the optical security system more than the one-dimensional nature of light and many physical parameters of the lightwave can be used to encode the data. We show the operation of the secure ultrafast data communication system and present some numerical results of encryption and decryption.

Section 2 describes a secure ultrafast data communication system. Section 3 numerically evaluates the performance of the secure data communication system.

2. SECURE ULTRAFAST DATA COMMUNICATIONS

Figure 1 shows a block diagram and a schematic of a secure data communication system. An encrypted holographic memory is linked to remote users via optical fibers. The secure holographic memory works as an encrypted database. All the data frames to be sent to the remote users were already encrypted by optical encryption techniques and then were recorded in the encrypted holographic memory. The data frame readout from the database is sent to the remote user via an optical fiber by spatio-temporal converters. After the transmission, an original data frame is decrypted at the remote users by using the correct key. The present system consists of four subsystems: the secure holographic memory system, a transmitter based on a space-to-time converter, a receiver based on a time-to-space converter, and a decrypting system to recover the original data. The following subsections describe the operation of each subsystem.

2.1. Secure Holographic Memory

Optical encryption opens new research fields in optical information processing [12–15]. An optical encryption technique, called *double random-phase encryption*, described by Rèfrègier and Javidi in 1995 [10], has been proposed and extensively investigated. The double-random-phase encryption technique can convert an original data frame into a white-noise distribution by using two random phase functions at the input and the Fourier planes. The double random phase encryption technique is easy to introduce into a holographic memory [16–19]. We

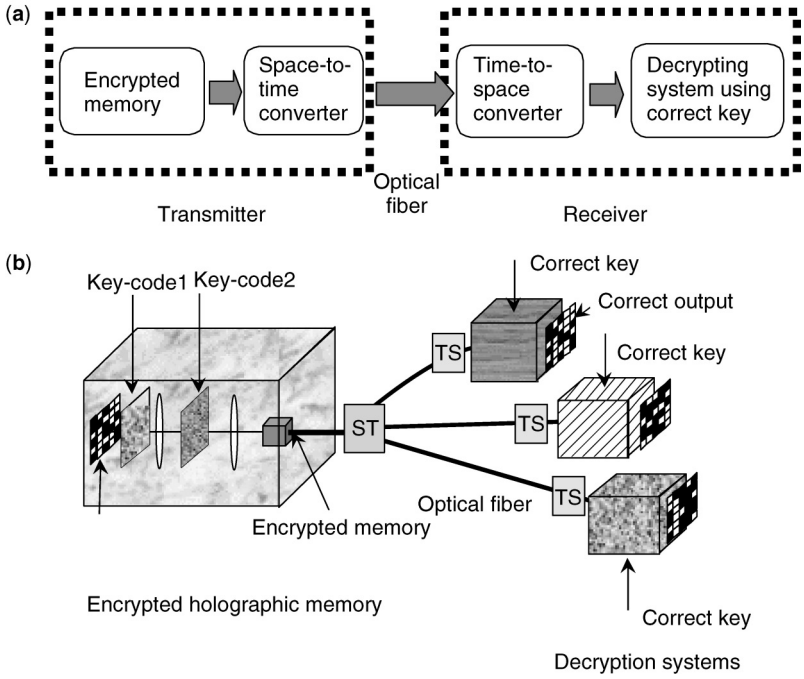


Figure 1. (**a**) Block diagram and (**b**) schematic of secure data communication system. ST and TS denote the space-to-time and time-to-space converters, respectively.

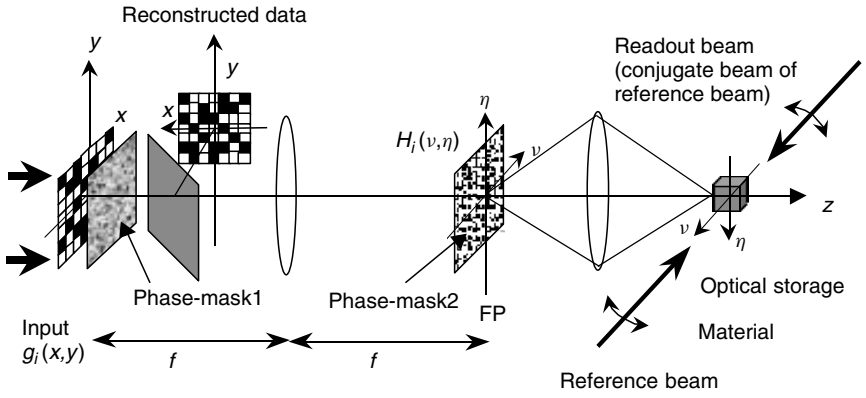


Figure 2. Encrypted holographic memory system.

briefly describe the encrypted holographic storage system, as shown in Fig. 2, using double-random-phase encryption.

Let $g_i(x)$ denote the ith positive and real data to be encrypted. We consider one-dimensional description because the following spatio-temporal converter uses one-dimension signals in the temporal domain. In the encryption process, the ith original data, $g_i(x)$, is multiplied by a random-phase mask (RPM1), $\exp\{-jn_i(x)\}$. This modulated input is Fourier-transformed and then is multiplied by a second random-phase mask (RPM2), $\tilde{H}_i(\nu) = \exp\{-h_i(\nu)\}$. Here x and ν denote the spatial domain and the Fourier domain coordinates, respectively. Two independent white sequences, as $n_i(x)$ and $h_i(\nu)$, are uniformly distributed on the interval $[0, 2\pi]$. After taking another Fourier transform, we obtain the encrypted data, $e_i(x)$:

$$e_i(x) = g_i(x) \exp\{-jn_i(x)\} \otimes F[\exp\{-jh_i(\nu)\}] \qquad (1)$$

where $\otimes$ denotes convolution and $F[\cdot]$ denotes the Fourier transformation. Equation (1) shows that encrypted data are white-noise-like data because of the convolution of two independent white noises.

In a holographic memory, as shown in Fig. 2, the Fourier transformed pattern of the encrypted data described by Eq. (1) is stored holographically together with a reference beam as a *plane wave*. Photorefractive materials are used to record volume holograms. In photorefractive materials, the intensity distribution is recorded as a refractive-index distribution. In order to record many data frames, angular multiplexing is employed. The total intensity distribution, $\phi(\nu)$, stored in the photorefractive material is given by

$$\phi(\nu) = \sum_{i=1}^{M} |\tilde{E}_i(\nu) + \tilde{R}_i(\nu)|^2 \qquad (2)$$

where M is the total number of stored images, $\tilde{E}_i(\nu)$ is the Fourier transform of ith input encrypted data described in Eq. (1), and $\tilde{R}_i(\nu)$ is a reference beam with a specific angle used to record the ith encrypted data.

A sufficient separation angle prevents crosstalk between adjacent stored data in the reconstruction.

Before we present the security of data communication, we show some experimental results in holographic encrypted memory systems as shown in Fig. 2. In the encrypted holographic memory, all stored images are encrypted. In the decryption process, we have to eliminate the phase modulation caused by the two random-phase modulations. Therefore we use the phase conjugate reconstruction in the decryption process. A readout beam is the conjugate beam of the reference beam used in the recording. The conjugate readout can eliminate the phase distortions of the optical field due to the random phase masks and the aberrations of optical elements. The data of the ith stored image can be reconstructed only when the readout beam is incident at a correct angle. The reconstructed data at the photorefractive material is written as $\tilde{E}_i^*(\nu)$, where the asterisk denotes complex conjugate. When we use a phase key, $\tilde{K}_i(\nu) = \exp\{-j2\pi k_i(\nu)\}$ in the Fourier plane, the reconstructed data in the Fourier plane is written as

$$\tilde{S}_i'(\nu) = F[g_i * (x)\exp\{jn_i(x)\}]\tilde{H}_i * (\nu)\tilde{K}_i(\nu) \tag{3}$$

The ith reconstructed image can be obtained by taking another Fourier transform of Eq. (3):

$$s_i'(x) = [g_i * (x)\exp\{jn_i(x)\}] \otimes C_i(x) \tag{4}$$

where

$$C_i(x) = F[\exp\{-jh_i(\nu)\}] \oplus F[\exp\{-jk_i(\nu)\}] \tag{5}$$

In Eq. (5), $\oplus$ denotes correlation. When a correct phase key, $k_i(\nu) = h_i(\nu)$, is used in the Fourier plane, the original data are successfully recovered because Eq. (5) becomes a delta function. The random phase function in the input plane is removed by detecting with an intensity-sensitive device such as a CCD (charge-coupled device) camera. When an incorrect phase key, $k_i(\nu) \neq h_i(\nu)$, is used, the reconstructed data frame is still a white-noise-like image.

Figure 3 shows the experimental setup. An Ar^+ laser at a wavelength of 514.5 nm is used as recording and readout beams. A light beam emitted from the Ar^+ laser was divided into an object beam and a reference beam by a beamsplitter, BS1. The reference beam was again divided into two reference beams by a beamsplitter BS2: one for recording holograms and one for the conjugate readout. All the beams were ordinarily polarized due to the creation of an interference pattern. We use a $10 \times 10 \times 10$ mm^3 $LiNbO_3$ crystal doped with 0.03 mol% Fe as a photorefractive material. The crystal was placed at the Fourier plane and was mounted on a rotary stage for angular multiplexing. The c axis is on the paper and is at 45° with respect to the crystal faces.

An input binary image is displayed on a liquid crystal display controlled by a computer. The input image is illuminated by a collimated beam and is then Fourier-transformed by lens L1. Two random phase-masks, RPM1 and RPM2, are located at the input and the Fourier planes, respectively. A reduced size of the Fourier-transformed image is imaged into the $LiNbO_3$ crystal by lens L2. This Fourier-transformed image and a reference beam interfere with an angle of 90° in the $LiNbO_3$ crystal. In holographic recording, shutters SH1 and SH2 are opened, and SH3 is closed. The encrypted image is observed by a CCD camera (CCD1) after the Fourier transform was taken by lens L3. The focal lengths of L1, L2, and L3 were 400, 58, and 50 mm, respectively.

In the decryption process, the readout beam is the conjugate beam of the reference beam used in the recording. In the experiments, a pair of counterpropagating plane waves is used as the reference and readout beams. Shutters SH1 and SH2 are closed, and SH3 is opened. If the same mask used in the recording is located at the same place, the original image is reconstructed successfully. The intensity information of the reconstructed image is obtained at the CCD camera (CCD2).

We present an example of the experimental results as shown in Fig. 4. In the experiments, angularly multiplexed recording of four binary images was demonstrated. One of the four original binary images is shown in Fig. 4a.

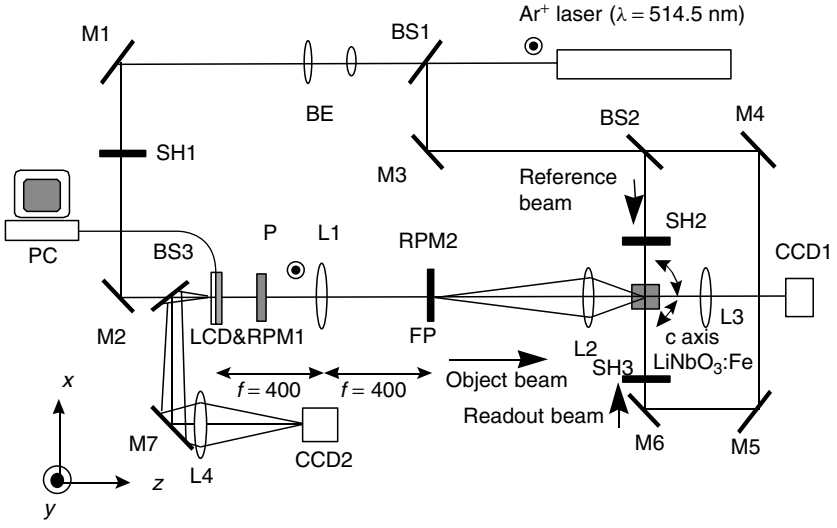


Figure 3. Experimental setup: LCD — liquid crystal display; RPM — random phase masks; BS — beamsplitters; L — lenses; M — mirrors; BE — beam expander; SH — shutters; CCD — CCD cameras; FP — Fourier plane; P — Polarizer; PC — personal computer.

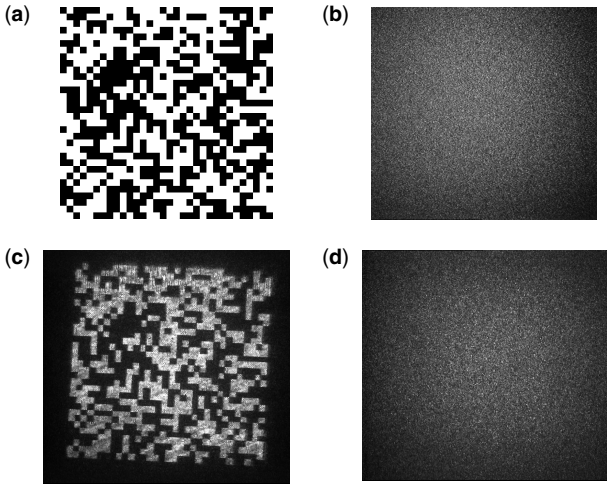


Figure 4. An example of encrypted holographic memory: (**a**) original binary image to be encrypted; (**b**) encrypted image; (**c**) reconstructed image by using correct key; (**d**) reconstructed image by using incorrect key.

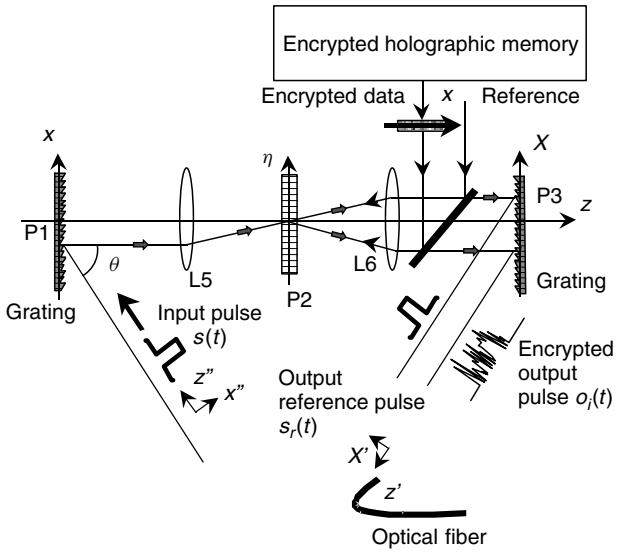


Figure 5. Transmitter based on space-to-time converter.

The image consists of 32×32 pixels and is randomly generated by a personal computer. Two diffusers are used as the random phase-masks, RPM1 and RPM2. Figure 4b shows the intensity distributions of the encrypted images. We can see that random-noise-like image was observed. In the recording process, the optical powers of the object and the reference beams were 78 and 1.4 W/cm^2, respectively. The exposure time was 60 s. Angular multiplexing was achieved by rotating the $LiNbO_3$ crystal in the plane of Fig. 3. The angular separation between adjacent stored images was 0.2°. Figure 4c shows the reconstructed images obtained by using the correct phase keys. These keys are the same as the random-phase masks in the Fourier plane used in the recording. Figure 4c shows that the stored image was reconstructed successfully. After the binarization, we confirmed that there is an error-free reconstruction in the four reconstructed images. Figure 4d shows the reconstructed image when an incorrect key was used. Here the incorrect key was generated by shifting the correct key. The reconstructed image was still a random-noise-like image. The average bit error rate for the reconstructed images with the incorrect keys was 0.502.

In the ultrafast secure communication system as described in Fig. 1, the encrypted data are transmitted to remote users by spatio-temporal converters via optical fibers and then the decryption is implemented at the remote users. To readout the encrypted data from the memory, we use the reference beam as the readout beam. In this case, the reconstructed data are given by Eq. (1) and then are converted into temporal data using the space-to-time converter.

2.2. Transmitter

Figure 5 shows an optical transmitter based on a space-to-time converter with ultrashort pulses. The space-to-time converter converts spatial data to temporal data by controlling the amplitude and phase of the spectra of the input pulse at the Fourier plane, P2 in Fig. 5. Sun and Fainman have extensively investigated the space-to-time converter [5,6]. Using their analysis, we briefly derive an output temporal signal to be sent to the receivers. Because the encrypted data readouts from the secure holographic memory are complex-valued, we introduce here a complex value notation to describe the operation of the space-to-time converter.

An input ultrashort pulse, $s(t)$, is described as

$$s(t) = p(t - t_0)\exp(j\omega_0 t) \tag{6}$$

where $p(t)$ is the envelope of the pulse, t_0 is the time of peak intensity, and ω_0 is the central temporal frequency of the pulse. The temporal frequency distribution of the input pulse, $\tilde{S}(\omega)$, is obtained, by taking the temporal Fourier transform of Eq. (6):

$$\tilde{S}(\omega) = \tilde{P}(\omega - \omega_0)\exp\{-j(\omega - \omega_0)t_0\} \tag{7}$$

where $\tilde{P}(\omega)$ is the temporal Fourier transform of $p(t)$. In the same pulseshaper as described in Fig. 5, each temporal frequency distribution is converted into the spatial distribution at the Fourier plane by using a grating and Fourier-transform lens, L5.

We consider the temporal frequency response of the space-to-time converter in Fig. 5 and then obtain an output temporal pulse. The space-to-time converter consists of two pairs of grating and Fourier transform lens. Suppose that the grating diffracts the light pulse with a temporal frequency of ω_0 into the direction parallel to the optical axis (z axis). When a monochromatic plane wave with a temporal frequency of ω is incident on the grating at an angle of θ, the diffracted optical field at plane P1 is given by

$$\psi_1(x;\omega) = \exp\left\{-j\frac{\omega - \omega_0}{c}\alpha x\right\} w(x) \tag{8}$$

where $\alpha = sin\theta$, w(x) is a pupil function of the grating, and c is the speed of light in vacuum. After taking the Fourier

transform of Eq. (8) by lens L5, the optical field at P2 is given by

$$\tilde{\psi}_2(\eta;\omega) = \tilde{W}\left\{\frac{\omega\eta}{2\pi cf} + \frac{\omega-\omega_0}{2\pi c}\alpha\right\} \tag{9}$$

where $\tilde{W}(\eta)$ is the spatial Fourier transform of $w(x)$, η is the Cartesian coordinate in the plane P2, and f is the focal length of lens L5. Equation (9) shows that the spectral distribution of the temporal delta function is projected into the spatial distribution as a function of the temporal frequency of light. When the pupil function $w(x)$ is infinite [i.e., $\tilde{W}(\eta) = \delta(\eta)$], the relation between η and $(\omega - \omega_0)$ is given by

$$\eta = -f\alpha\frac{\omega-\omega_0}{\omega} \tag{10}$$

The spatially spread spectra described in Eq. (9) are modulated by a hologram that is the interference pattern between the Fourier-transform of the encrypted signal described in Eq. (1) and a reference plane wave.

We suppose that the encrypted signal $e_i(x)$ is spatially sampled at an interval of Δ as follows:

$$\begin{aligned} r_i(x) &= \sum_n e_i(x)\delta(x-n\Delta) \\ &= \sum_n A_i(x)\exp\{j\phi_i(x)\}\delta(x-n\Delta) \\ &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \end{aligned} \tag{11}$$

where $A_i(x) = |e_i(x)|$, $\exp\{j\phi_i(x)\} = e_i(x)/|e_i(x)|$, and Δ is the sampling period. This data sampling is required to avoid the overlap between adjacent data in the reconstructed spatial data at the receiver. In Section 2.3, we show that the spatial data become wide at the receiver after the transmission of spatio-temporal converters. The encrypted data are Fourier-transformed by lens L6. The spatial Fourier transform of the encrypted data at P2 is given by

$$\tilde{R}_i(\eta) = \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}\exp\left(-j\frac{n\Delta\omega'}{cf}\eta\right) \tag{12}$$

where $\omega' = 2\pi c/\lambda'$, f is the focal length of lens L6, and λ' is the wavelength of the light beam used to write the hologram. This signal is recorded as a real-time hologram together with a reference plane wave in an intensity-sensitive medium, such as a multiple-quantum-well photorefractive device with sufficient spectral response. Here we assume that the hologram works as a grating with the transmittance of

$$t_i(\eta) = \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}\exp\left(-j\frac{n\Delta\omega'}{cf}\eta\right) \tag{13}$$

where we neglect the coefficient for normalization of transmittance and the effect of the carrier frequency of the hologram caused by the angle between the encrypted data and the reference beam. This hologram modulates the temporal signal described in Eq. (9). The optical field after the diffraction through the hologram is expressed by

$$\begin{aligned} \tilde{\psi}_3(\eta;\omega) &= \tilde{\psi}_2(\eta;\omega)t_i(\eta) \\ &= \tilde{W}\left\{\frac{\omega\eta}{2\pi cf} + \frac{\omega-\omega_0}{2\pi c}\alpha\right\} \times \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \\ &\quad \times \exp\left(-j\frac{n\Delta\omega'}{cf}\eta\right) \end{aligned} \tag{14}$$

This modulated field is then Fourier-transformed by lens L6:

$$\begin{aligned} \tilde{\psi}_4(X;\omega) &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \\ &\quad \times \exp\left\{j\frac{\omega-\omega_0}{c}\alpha\left(X + n\Delta\frac{\omega'}{\omega}\right)\right\} w\left(-X - n\Delta\frac{\omega'}{\omega}\right) \end{aligned} \tag{15}$$

This optical field is diffracted again by a grating and is given by

$$\begin{aligned} \tilde{\psi}_5(X';\omega) &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\} \\ &\quad \times \exp\left\{j\frac{\alpha n\Delta}{c}\frac{(\omega-\omega_0)\omega'}{\omega}\right\} w'\left(-X' - n\Delta\frac{\omega'}{\omega}\right) \end{aligned} \tag{16}$$

where X' is the coordinate as shown in Fig. 5, $w'(X')$ is the pupil function of the grating projected onto the X' coordinate. Equation (16) denotes the temporal frequency response of the system. Using Eqs. (7) and (16), we can obtain the output temporal signal by taking an inverse temporal Fourier transform of $\tilde{\psi}_5(X',\omega)\tilde{S}(\omega)$:

$$\begin{aligned} o_i(X',t) &= \int_{-\infty}^{\infty}\tilde{\psi}_5(X';\omega)\tilde{S}(\omega)\exp(-j\omega t)d\omega \\ &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}w'\left(-X' - n\Delta\frac{\omega'}{\omega_0}\right) \\ &\quad \times p(t - t_0 + n\delta t)\exp(j\omega_0 t) \end{aligned} \tag{17}$$

where $\delta t = (\alpha\Delta/c) \times (\omega'/\omega_0)$. To derive Eq. (17) we used the approximation, $1/\omega \approx 1/\omega_0$. This is valid in case of $\Delta\omega = (\omega - \omega_0) \ll \omega_0$ in a few hundreds femtosecond pulse.

Another light pulse, which is passing through P2 without diffraction due to the hologram, is also sent to remote users. This pulse has the same envelope as the input pulse; thus it can be written as

$$s_r(t) = p(t - t_0)\exp(j\omega_0 t) \tag{18}$$

This pulse is used as a reference pulse at the remote users to eliminate the phase distortion in the optical fiber. Both the temporally encrypted signal described in Eq. (17) and the reference pulse in Eq. (18) are sent to the receiver through a single mode fiber. Both pulses travel along the same line with a sufficient temporal interval by using a delay line before the fiber.

2.3. Receiver

At the receiver as shown in Fig. 6, the temporally encrypted data are converted again into spatially

encrypted data using a time-to-space converter. Because a single-mode fiber is used to send the temporally encrypted pulse and the reference pulse, the spatial information of both pulses should be dropped. The time-to-space converter consists of a pair of grating and Fourier transform lens. As shown in Fig. 6, the two light pluses are incident on the grating at an angle of θ and then are diffracted by the grating. These two pulses create the interference pattern after taking the spatial Fourier transform by lens L7. Here the temporal frequency response is considered. Using Eqs. (9), (17), and (18), the intensity distribution of the interference pattern at the Fourier plane P5 is described by

$$\begin{aligned}\tilde{I}_i(\eta,\omega) &= \left|\tilde{O}_i(\omega)\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right. \\ &\quad \left.+\,\tilde{S}_r(\omega)\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &= |\tilde{O}_i(\omega)|^2\times\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &\quad +|\tilde{S}_r(\omega)|^2\times\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &\quad +\tilde{O}_i*(\omega)\tilde{S}_r(\omega)\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \\ &\quad +\tilde{O}_i(\omega)\tilde{S}_r*(\omega)\left|\tilde{W}\left[\frac{\omega\eta}{2\pi cf}+\frac{\omega-\omega_0}{2\pi c}\alpha\right]\right|^2 \quad (19)\end{aligned}$$

where $W(\eta)$ is the spatial Fourier transform of $w(x)$ and $w(x)$ is a pupil function of the grating at P4. In Eq. (19)

$$\begin{aligned}\tilde{O}_i(\omega) &= \sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}P(\omega-\omega_0) \\ &\quad\times\exp\{-j(\omega-\omega_0)(t_0-n\delta t)\} \quad (20)\end{aligned}$$

and

$$\tilde{S}_r(\omega)=\tilde{P}(\omega-\omega_0)\exp\{-j(\omega-\omega_0)t_0\} \quad (21)$$

Equations (20) and (21) denote the Fourier transforms of Eqs. (17) and (18), respectively.

We use the third term in Eq. (19) to reconstruct the spatially encrypted signal. A continuous-wave (CW) laser beam is incident on the hologram to read out the stored information. The reconstructed optical field is then spatially Fourier transformed by lens L7. When the pupil function of the grating and the beam width are large enough [$\tilde{W}(\eta)=\delta(\eta)$], the reconstructed signal at plane P6 is expressed by

$$\begin{aligned}\xi_i(x') &= F\left[\sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\}|P(\omega-\omega_0)|^2\right. \\ &\quad\left.\times\ \exp\{-j(\omega-\omega_0)n\delta t\}\right] \quad (22)\end{aligned}$$

Note that F[·] denotes the spatial Fourier transform. Equation (22) is calculated as follows:

$$\begin{aligned}\xi_i(x') &= \int_{-\infty}^{\infty}\sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\}\left|P\left(\frac{-\omega_0\eta}{f\alpha}\right)\right|^2 \\ &\quad\times\exp\left\{\frac{j\omega_0\eta n\delta t}{f\alpha}\right\}\exp\left\{\frac{-j2\pi x'\eta}{\lambda'' f}\right\}d\eta \\ &= \sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\} \\ &\quad\times\exp\left\{-\frac{\alpha^2\left(\frac{1}{\lambda''}x'+n\frac{\Delta}{\lambda'}\right)^2}{4\omega_0^2\tau^2}\right\} \quad (23)\end{aligned}$$

To derive Eq. (23), we use a Gaussian-shaped input pulse envelope written as $p(t)=\exp(-t^2/2\tau^2)$, τ is a pulse width, and λ'' is the wavelength of the CW laser beam. Equation (23) shows that each pixel becomes wide by convolving a Gaussian function with a $1/e^2$ width of $w_d=4\sqrt{2}\omega_0\tau\lambda''/\alpha$. This reconstructed signal is used in the following decryption system as shown in Fig. 7 to reconstruct the original data.

2.4. Decryption System

When the sampling interval, Δ, is larger than w_d in Eq. (23) and $\lambda''=\lambda'$, the reconstructed data do not overlap each other. We sample again Eq. (23) at $x=n\Delta$. The sampled data are given by

$$\begin{aligned}\xi_i'(x) &= \sum_n \xi_i(x)\delta(x-n\Delta) \\ &= \sum_n A_i(n\Delta)\exp\{-j\phi_i(n\Delta)\} \\ &= \left[\sum_n A_i(n\Delta)\exp\{j\phi_i(n\Delta)\}\right]^* \quad (24)\end{aligned}$$

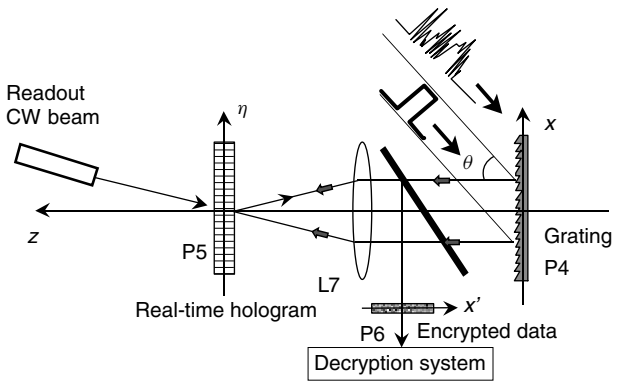


Figure 6. Receiver using time-to-space converter.

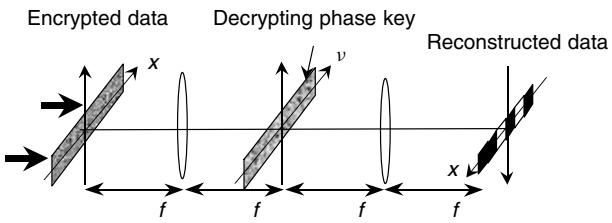


Figure 7. Decrypting system.

where the asterisk denotes complex conjugation. This equation shows that the complex conjugate of the encrypted signal in Eq. (11) is reconstructed. To decrypt the data, the spatial Fourier transform of Eq. (24) is multiplied by the phase key $\tilde{H}_i(\nu) = \exp\{-jh_i(\nu)\}$ as shown in Fig. 7. This phase key is the same random phase mask as that used in the encryption system. In the Fourier plane, the reconstructed data are written by

$$\tilde{\Psi}_i(\nu) = \left\{ \tilde{G}_i^*(\nu) \otimes F[\exp\{-jn_i(x)\}] * \tilde{H}_i^*(\nu) \otimes \exp\left(j\frac{2\pi}{\lambda f} n\Delta\nu\right) \right\} \tilde{H}_i(\nu) \tag{25}$$

where ν denotes the coordinate in the Fourier plane, and $\otimes$ denotes convolution. Finally the reconstructed data are obtained by taking another Fourier transform of Eq. (25):

$$I_{\text{out}}(x) = \sum_n g_i(x) \exp\{-jn_i(x)\} \otimes F * [\exp\{-jh_i(\nu)\}] \cdot \delta(x - n\Delta) \otimes F[\exp\{-jh_i(\nu)\}] \tag{26}$$

This equation shows that the reconstructed data are not exactly the same as the original data, $g_i(x)$ due to the spatial sampling of the encrypted data. In the following section, we numerically evaluate the reconstructed data.

3. NUMERICAL EVALUATIONS

We numerically evaluate the error between the original data and the reconstructed data by use of Eq. (26). When $\omega_0 = 6\pi \times 10^{14}$, $\lambda'' = 1\ \mu\text{m}$, $\alpha = 1/\sqrt{2}$, and $\tau = 50$ fs, the sampling interval, sufficient to avoid the overlap, is 754 μm. If the sampling interval, Δ is smaller than the width of the Gaussian distribution, w_d, the reconstructed spatial data overlap each other. Thus, the original data cannot be reconstructed when the overlap is large, even if we use the correct phase key in the decryption process. To avoid the overlap at the receiver, we have to use a large sampling interval at the transmitter. The sampling results in some loss of the encrypted data and leads to the error in the reconstructed data. In the following calculations, we evaluate the bit error rate in a binary data transmission using undersampled data.

When the encrypted data are undersampled by a factor of 2, half of the encrypted data are lost. This loss of information causes the error in the decrypted data. Using binary data, however, it is possible to reduce the noise in the decrypted data by thresholding. A mean squared error is used as the performance criterion

$$e = E\{|g(x) - m \times g'_\Delta(x)|^2\} \tag{27}$$

where $E\{\cdot\}$ denotes statistical average, $g(x)$ is the original data, m denotes the coefficient to compensate for the loss of total power due to undersampling the encrypted data, and $g'_\Delta(x)$ is the reconstructed data using Eq. (26) when the encrypted data are sampled at the interval of Δ.

The original binary data are 32-bit data, where each bit consists of 64 pixels. Thus the input data have 2048 pixels. This redundancy of the original data is introduced to recover the original binary data. Two random phase masks in the input and Fourier planes consist of 2048 pixels. The original binary data and the two random-phase masks are randomly generated in a computer. The average mean squared error was calculated over 1000 different trials. An example of the original data, encrypted data, and decrypted data is shown in Fig. 8. Here the decrypted data are lowpass-filtered. The lowpass filtering was performed by locally averaging the data with an 11-pixel window. We can see that the encrypted data are random, but the decrypted data have the same structure of the original data. This means that it is possible to recover the original binary data without bit error.

We calculated the bit error rate as a function of the sampling interval, Δ. The bit error rate is defined as a ratio of the number of error bits to the total number of bits. By using the reconstructed analog data as shown in Fig. 8, we calculated the energy of each cell of the reconstructed data. After the calculation of energy, each datum is binarized. To determine the threshold value, all energies are rank ordered. The threshold value is the Nth largest energy of the cells, where N is the number of bright pixels(ones) in the original binary data. Figure 9 shows the bit error rate as a function of the sampling interval Δ. For the binary data used here, the original binary data were reconstructed without error in the case of $\Delta = 1$ and 2. When the sampling interval becomes large, the loss of encrypted data causes a large number of bit errors.

4. CONCLUSION

We have presented a secure data communication system that links remote users to a secure holographic memory system at ultrafast transfer rate. In the encrypted holographic memory each data frame is recorded as a hologram after encrypting the original data by using double-random-phase encryption. The recorded data can be readout in parallel at a fast rate. The spatio-temporal converters based on ultrashort pulse shaper enable us to send the encrypted data at ultrahigh speed via optical fibers. At the remote users, only an authorized user can reconstruct the original data by using the correct keys. We expect that the system can provide a high level of security and ultrafast data communication.

BIOGRAPHIES

Bahram Javidi, distinguished Professor of Electrical and Computer Engineering at University of Connecticut, received the B.S. degree (1980) in electrical engineering from George Washington University, Washington, D.C., and the M.S. (1982) and Ph.D. (1986) degrees in electrical engineering from the Pennsylvania State University, University Park. He is fellow of Institute Of Electrical and Electronics Engineers, fellow of the Optical Society of America, and fellow of the International Society for Optical Engineering (SPIE). In 1990, he was named a Presidential Young Investigator by the National Science Foundation. He has been awarded the University of

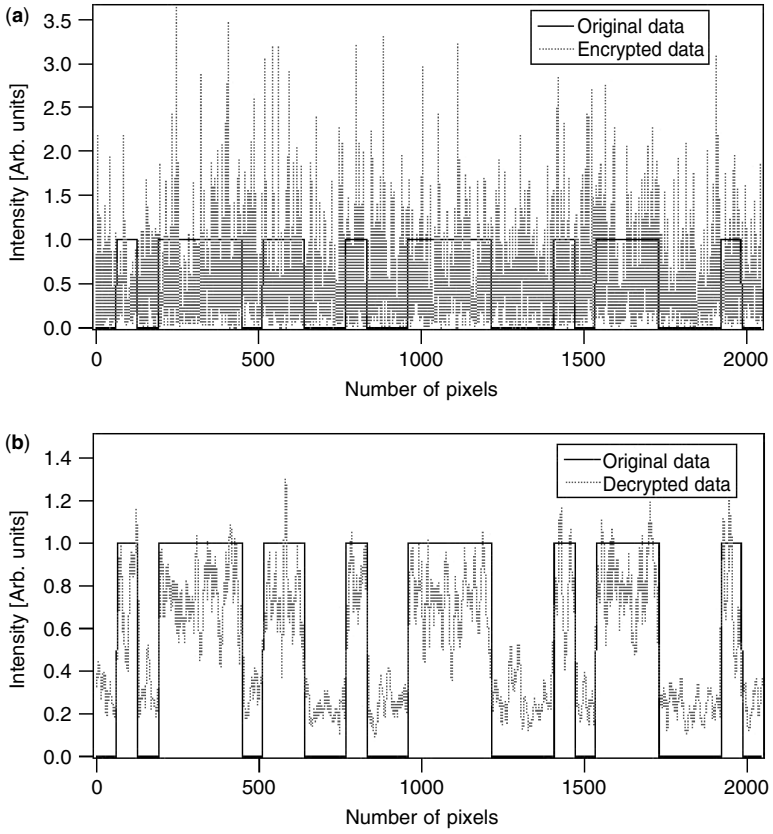


Figure 8. An example of original, encrypted, and decrypted data: (**a**) original and encrypted data; (**b**) original and decrypted data.

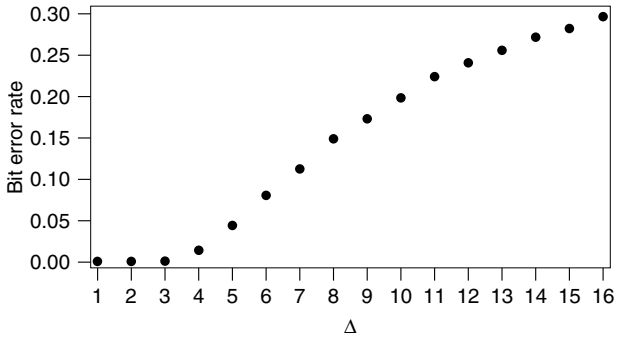


Figure 9. Bit error rate as a function of sampling period, Δ.

Connecticut Alumni Association Research Excellence Award, and the first Electrical and Computer Engineering Department Outstanding Research Award. He is the editor of several books including, "Image Recognition: Algorithms, Systems, and Applications," published by Marcel-Dekker, New York, 2002; "Three Dimensional Television, Video, and Display Technologies," Springer Verlag in 2002; "Smart Imaging Systems," SPIE Press in 2001; and "Real-time Optical Information Processing," Academic Press, 1994. In addition, he has published over 260 technical articles in major journals and conference proceedings, including over 40 invited papers. He has served on the editorial boards for Springer-Verlag, Marcel Dekker, the Optical Engineering Journal, and IEEE/SPIE Press.

His email address is bahram.javidi@uconn.edu

Osamu Matoba received the B.Eng., M.Eng., and D.Eng. degrees in applied physics from Osaka University, Osaka, Japan, in 1991, 1993, and 1996, respectively. Since 1996, he has been a research associate at the Institute of Industrial Science, University of Tokyo, Japan. His current research interests include optical information processing, optical security, three-dimensional display, digital holography, and the development of photorefractive materials. Dr. Matoba is a member of the Japanese Society of Applied Physics, the Optical Society of Japan, the Laser Society of Japan, OSA, SPIE, and IEEE LEOS.

BIBLIOGRAPHY

1. A. M. Weiner, D. E. Leaird, D. H. Reitze, and E. G. Paek, Femtosecond spectral holography, *IEEE J. Quant. Electron.* **QE-28**: 2251–2261 (1992).
2. Y. T. Mazurenko, Holography of wave packets, *Appl. Phys. B* **50**: 101–114 (1990).
3. A. W. Weiner and A. M. Kan'an, Femtosecond pulse shaping for synthesis, processing, and time-to-space conversion of ultrafast optical waveforms, *IEEE J. Select. Topics Quant. Electron.* **4**: 317–331 (1998).
4. Y. Ding, D. D. Nolte, M. R. Melloch, and A. W. Weiner, Time-domain image processing using dynamic holography, *IEEE J. Select. Topics Quant. Electron.* **4**: 332–341 (1998).
5. P. C. Sun et al., All-optical parallel-to-serial conversion by holographic spatial-to-temporal frequency encoding, *Opt. Lett.* **20**: 1728–1730 (1995).

6. D. M. Marom, P. C. Sun, and Y. Fainman, Analysis of spatial-temporal converters for all-optical communication links, *Appl. Opt.* **37**: 2858–2868 (1998).
7. O. Matoba and B. Javidi, Secure ultrafast communication with spatial-temporal converters, *Appl. Opt.* **39**: 2975–2981 (2000).
8. J. F. Heanue, M. C. Bashaw, and L. Hesselink, Volume holographic storage and retrieval of digital data, *Science* **265**: 749–752 (1994).
9. H. J. Coufal, D. Psaltis, and G. T. Sincerbox, eds., *Holographic Data Storage*, Springer, New York, 2000.
10. P. Rèfrègier and B. Javidi, Optical image encryption based on input plane and Fourier plane random encoding, *Opt. Lett.* **20**: 767–769 (1995).
11. B. Javidi, Encrypting information with optical technologies, *Phys. Today* **50**(3): (March 1997).
12. J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, New York, 2000.
13. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York, 1991.
14. B. Javidi, ed., *Smart Imaging Systems*, SPIE Press, Bellingham, WA, 2001.
15. B. Javidi and J. L. Horner, eds., *Real-Time Optical Information Processing*, Academic Press, Boston, 1994.
16. O. Matoba and B. Javidi, Encrypted optical memory system using three-dimensional keys in the Fresnel domain, *Opt. Lett.* **24**: 762–764 (1999).
17. O. Matoba and B. Javidi, Encrypted optical storage with wavelength-key and random phase codes, *Appl. Opt.* **38**: 6785–6790 (1999).
18. O. Matoba and B. Javidi, Encrypted optical storage with angular multiplexing, *Appl. Opt.* **38**: 7288–7293 (1999).
19. O. Matoba and B. Javidi, Encrypted optical memory systems based on multidimensional keys for secure data storage and communications, *IEEE Circuits Devices Mag.* **16**: 8–15 (Sept. 2000).

SEQUENTIAL DECODING OF CONVOLUTIONAL CODES

YUNGHSIANG S. HAN
National Chi Yan University
Taiwan Republic of China

PO-NING CHEN
National Chi Tung University
Taiwan Republic of China

1. INTRODUCTION

The convolutional coding technique is designed to reduce the probability of erroneous transmission over noisy communication channels. The most popular decoding algorithm for convolutional codes is perhaps the Viterbi algorithm. Although widely adopted in practice, the Viterbi algorithm suffers from a high decoding complexity for convolutional codes with long constraint lengths. While the attainable decoding failure probability of convolutional codes generally decays exponentially with the code constraint length, the high complexity of the Viterbi decoder for codes with a long constraint length to some extent limits the achievable system performance. Nowadays, the Viterbi algorithm is usually applied to codes with a constraint length no greater than nine.

In contrast to the limitation of the Viterbi algorithm, sequential decoding is renowned for its computational complexity being independent of the code constraint length [1]. Although simply suboptimal in its performance, sequential decoding can achieve a desired bit error probability when a sufficiently large constraint length is taken for the convolutional code. Unlike the Viterbi algorithm that locates the best codeword by exhausting all possibilities, sequential decoding concentrates only on a certain number of likely codewords. As the sequential selection of these likely codewords is affected by the channel noise, the decoding complexity of a sequential decoder becomes dependent on the noise level [1]. These specific characteristics make sequential decoding useful in particular applications.

Sequential decoding was first introduced by Wozencraft for the decoding of convolutional codes [2,3]. Thereafter, Fano developed the sequential decoding algorithm with a milestone improvement in decoding efficiency [4]. Fano's work subsequently inspired further research on sequential decoding. Later, Zigangirov [5], and independently, Jelinek [6], proposed the *stack algorithm*.

In this article, the sequential decoding will not be introduced chronologically. Rather, Algorithm A [7] — the general sequential search algorithm — will be introduced first because it is conceptually more straightforward. The rest of the article is organized as follows. Sections 2 and 3 provide the necessary background for convolutional codes and typical channel models for performance evaluation. Section 4 introduces Algorithm A, and then defines the general features of sequential decoding. Section 5 explores the Fano metric and its generalization for use to guide the search of sequential decoding. Section 6 presents the stack algorithm and its variants. Section 7 elucidates the well-known Fano algorithm. Section 8 is devoted to the trellis variants of sequential decoding, especially on the proposed maximum-likelihood sequential decoding algorithm (MLSDA). Section 9 examines the decoding performance. Section 10 discusses various practical implementation issues regarding sequential decoding, such as buffer overflow. For completeness, a section on the code construction (Section 11) is included at the end of the article. Section 12 concludes the article.

For clarity, only binary convolutional codes are considered throughout discussions on sequential decoding. Extension to nonbinary convolutional codes can be carried out similarly.

2. CONVOLUTIONAL CODE AND ITS GRAPHICAL REPRESENTATION

Denote a binary convolutional code by a 3-tuple (n, k, m), which corresponds to an encoder for which n output bits are generated whenever k input bits are received, and for which the current n outputs are linear combinations of the present k input bits and the previous $m \times k$ input

bits. Because m designates the number of previous k-bit input blocks that must be memorized in the encoder, m is called the *memory order* of the convolutional code. A binary convolutional encoder is conveniently structured as a mechanism of shift registers and modulo-2 adders, where the output bits are modulo-2 additions of selective shift register contents and present input bits. Then n in the 3-tuple notation is exactly the number of output sequences in the encoder, k is the number of input sequences (and hence, the encoder consists of k shift registers), and m is the maximum length of the k shift registers (i.e., if the number of stages of the jth shift register is K_j, then $m = \max_{1 \le j \le k} K_j$). Figures 1 and 2 exemplify the encoders of binary $(2, 1, 2)$ and $(3, 2, 2)$ convolutional codes, respectively.

During the encoding process, the contents of shift registers in the encoder are initially set to zero. The k input bits from the k input sequences are then fed into the encoder in parallel, generating n output bits according to the shift register framework. To reset the shift register contents at the end of input sequences so that the encoder can be ready for use for another set of input sequences, m zeros are usually padded at the end of each input sequence. Consequently, each k input sequence of length L bits is padded with m zeros, and these k input sequences jointly induce $n(L + m)$ output bits. As illustrated in Fig. 1, the encoder of the $(2, 1, 2)$ convolutional code extracts two output sequences, $\mathbf{v}_1 = (v_{1,0}, v_{1,1}, v_{1,2}, \ldots, v_{1,6}) = (1010011)$ and $\mathbf{v}_2 = (v_{2,0}, v_{2,1}, v_{2,2}, \ldots, v_{2,6}) = (1101001)$, due to the single input sequence $\mathbf{u} = (u_0, u_1, u_2, u_3, u_4) = (11101)$, where u_0 is fed in the encoder first. The encoder then interleaves $\mathbf{v}_1$ and $\mathbf{v}_2$ to yield

$$\begin{aligned}\mathbf{v} &= (v_{1,0}, v_{2,0}, v_{1,1}, v_{2,1}, \ldots, v_{1,6}, v_{2,6}) \\ &= (11\ 01\ 10\ 01\ 00\ 10\ 11)\end{aligned}$$

of which the length is $2(5 + 2) = 14$. Also, the encoder of the $(3, 2, 2)$ convolutional code in Fig. 2 generates the output sequences of $\mathbf{v}_1 = (v_{1,0}, v_{1,1}, v_{1,2}, v_{1,3}) = (1000)$, $\mathbf{v}_2 = (v_{2,0}, v_{2,1}, v_{2,2}, v_{2,3}) = (1100)$, and $\mathbf{v}_3 = (v_{3,0}, v_{3,1}, v_{3,2}, v_{3,3}) = (0001)$, due to the two input sequences $\mathbf{u}_1 = (u_{1,0}, u_{1,1}) = (10)$ and $\mathbf{u}_2 = (u_{2,0}, u_{2,1}) = (11)$, which in turn generate the interleaved output sequence

$$\begin{aligned}\mathbf{v} &= (v_{1,0}, v_{2,0}, v_{3,0}, v_{1,1}, v_{2,1}, v_{3,1}, v_{1,2}, v_{2,2}, v_{3,2}, v_{1,3}, v_{2,3}, v_{3,3}) \\ &= (110\ 010\ 000\ 001)\end{aligned}$$

of length $3(2 + 2) = 12$. Terminologically, the interleaved output $\mathbf{v}$ is called the *convolutional codeword* corresponding to the combined input sequence $\mathbf{u}$.

An important subclass of convolutional codes is the *systematic codes*, in which k out of n output sequences retain the values of the k input sequences. In other words, these outputs are directly connected to the k inputs in the encoder.

A convolutional code encoder can also be viewed as a linear system, in which the relation between its inputs and outputs is characterized by generator polynomials. For example, $g_1(x) = 1 + x + x^2$ and $g_2(x) = 1 + x^2$ can be used to identify $\mathbf{v}_1$ and $\mathbf{v}_2$ induced by $\mathbf{u}$ in Fig. 1, where the appearance of x^i indicates that a physical connection is applied to the $(i + 1)$th dot position, counted from the left. Specifically, putting $\mathbf{u}$ and $\mathbf{v}_i$ in polynomial form as $\mathbf{u}(x) = u_0 + u_1 x + u_2 x^2 + \cdots$ and $\mathbf{v}_i(x) = v_{i,0} + v_{i,1} x + v_{i,2} x^2 + \cdots$ yields that $\mathbf{v}_i(x) = \mathbf{u}(x) g_i(x)$ for $i = 1, 2$, where addition of coefficients is based on modulo-2 operation. With reference to the encoder depicted in Fig. 2, the relation between the input sequences and the output sequences can be formulated through matrix operation as

$$[\mathbf{v}_1(x)\ \mathbf{v}_2(x)\ \mathbf{v}_3(x)] = [\mathbf{u}_1(x)\ \mathbf{u}_2(x)] \times \begin{bmatrix} g_1^{(1)}(x) & g_2^{(1)}(x) & g_3^{(1)}(x) \\ g_1^{(2)}(x) & g_2^{(2)}(x) & g_3^{(2)}(x) \end{bmatrix}$$

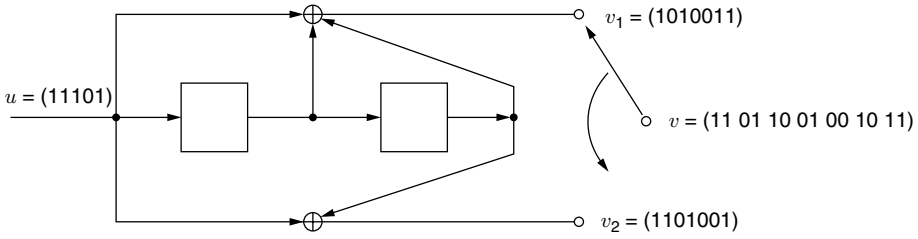


Figure 1. Encoder for the binary $(2, 1, 2)$ convolutional code with generators $g_1 = 7$ (octal) and $g_2 = 5$ (octal), where g_i is the generator polynomial characterizing the ith output.

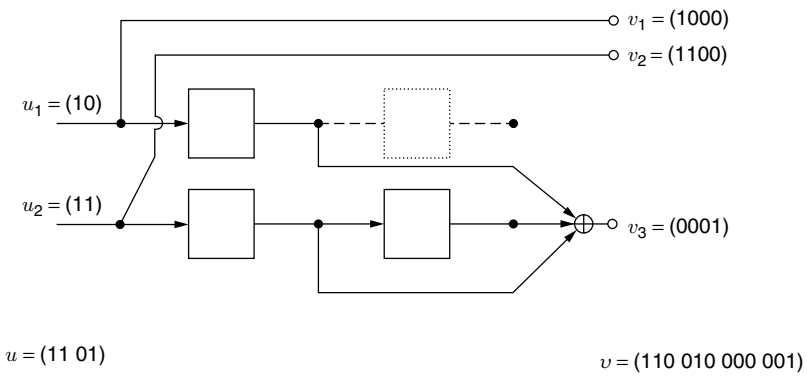


Figure 2. Encoder for the binary $(3, 2, 2)$ systematic convolutional code with generators $g_1^{(1)} = 4$ (octal), $g_1^{(2)} = 0$ (octal), $g_2^{(1)} = 0$ (octal), $g_2^{(2)} = 4$ (octal), $g_3^{(1)} = 2$ (octal), and $g_3^{(2)} = 3$ (octal), where $g_i^{(j)}$ is the generator polynomial characterizing the ith output according to the jth input. The dashed box is redundant and can actually be removed from this encoder; its presence here is only to help demonstrating the derivation of generator polynomials. Thus, as far as the number of stages of the jth shift register is concerned, $K_1 = 1$ and $K_2 = 2$.

where $\mathbf{u}_j(x) = u_{j,0} + u_{j,1}x + u_{j,2}x^2 + \cdots$ and $\mathbf{v}_i(x) = v_{i,0} + v_{i,1}x + v_{i,2}x^2 + \cdots$ define the jth input sequence and the ith output sequence, respectively, and the generator polynomial $g_i^{(j)}(x)$ characterizes the relation between the jth input and the ith output sequences. For simplicity, generator polynomials are sometimes abbreviated by their coefficients in octal number format, led by the least significant one. Continuing the example in Fig. 1 gives $g_1 = 111$ (binary) $= 7$ (octal) and $g_2 = 101$ (binary) $= 5$ (octal). A similar abbreviation can be used for each $g_i^{(j)}$ in Fig. 2.

An (n, k, m) convolutional code can be transformed to an equivalent linear block code with *effective code rate*[1] $R_{\text{effective}} = kL/[n(L+m)]$, where L is the length of the information input sequences. By taking L to infinity, the effective code rate converges to $R = k/n$, which is referred to as the *code rate* of the (n, k, m) convolutional code.

The *constraint length* of an (n, k, m) convolutional code has two different definitions in the literature: $n_A = m + 1$ [8] and $n_A = n(m + 1)$ [1]. In this article, the former definition is adopted, because it is more extensively used in military and industrial publications.

Let $\mathbf{v}_{(a,b)} = (v_a, v_{a+1}, \ldots, v_b)$ denote a portion of codeword $\mathbf{v}$, and abbreviate $\mathbf{v}_{(0,b)}$ by $\mathbf{v}_{(b)}$. The Hamming distance between the first rn bits of codewords $\mathbf{v}$ and $\mathbf{z}$ is given by

$$d_H(\mathbf{v}_{(rn-1)}, \mathbf{z}_{(rn-1)}) = \sum_{i=0}^{rn-1} v_i \oplus z_i$$

where "$\oplus$" denotes modulo-2 addition. The Hamming weight of the first rn bits of codeword $\mathbf{v}$ thus equals $d_H(\mathbf{v}_{(rn-1)}, \mathbf{0}_{(rn-1)})$, where $\mathbf{0}$ represents the all-zero codeword. The *column distance function* (CDF) $d_c(r)$ of a binary (n, k, m) convolutional code is defined as the minimum Hamming distance between the first rn bits of any two codewords whose first n bits are distinct

$$d_c(r) = \min\{d_H(\mathbf{v}_{(rn-1)}, \mathbf{z}_{(rn-1)}) : \mathbf{v}_{(n-1)} \neq \mathbf{z}_{(n-1)} \text{ for } \mathbf{v}, \mathbf{z} \in \mathcal{C}\}$$

where $\mathcal{C}$ is the set of all codewords. Function $d_c(r)$ is clearly nondecreasing in r. Two cases of CDFs are of specific interest: $r = m + 1$ and $r = \infty$. In the latter case, the input sequences are considered infinite in length.[2] Terminologically, $d_c(m + 1)$ and $d_c(\infty)$ (or d_{free} in general) are called the *minimum distance* and the *free distance* of the convolutional code, respectively.

The operational meanings of the minimum distance, the free distance and the CDF of a convolutional code are as follows. When a sufficiently large codeword length is taken, and an optimal (i.e., maximum-likelihood) decoder is employed, the error-correcting capability of a convolutional code [9] is generally characterized by d_{free}. In case a decoder figures the transmitted bits only on the basis of the first $n(m + 1)$ received bits (as in, e.g., the majority-logic decoding [10]), $d_c(m + 1)$ can be used instead to characterize the code error-correcting capability. As for sequential decoding algorithm that requires a rapid initial growth of column distance functions (to be discussed in Section 9), the decoding computational complexity, defined as the number of metric computations performed, is determined by the CDF of the code being applied.

Next, two graphical representations of convolutional codewords are introduced. They are derived from the graphs of *code tree* and *trellis*, respectively. A *code tree* of a binary (n, k, m) convolutional code presents every codeword as a path on a tree. For input sequences of length L bits, the code tree consists of $(L + m + 1)$ levels. The single leftmost node at level 0 is called the *origin node*. At the first L levels, there are exactly 2^k branches leaving each node. For those nodes located at levels L through $(L + m)$, only one branch remains. The 2^{kL} rightmost nodes at level $(L + m)$ are called the *terminal nodes*. As expected, a path from the single origin node to a terminal node represents a codeword; therefore, it is named the *code path* corresponding to the codeword. Figure 3 illustrates the code tree for the encoder in Fig. 1 with a single input sequence of length 5.

In contrast to a code tree, a code *trellis* as described by Forney [11] is a structure obtained from a code tree by merging those nodes in the same *state*. The *state* associated with a node is determined by the associated shift register contents. For a binary (n, k, m) convolutional code, the number of states at levels m through L is 2^K, where $K = \sum_{j=1}^{k} K_j$ and K_j is the length of the jth shift register in the encoder; hence, there are 2^K nodes on these levels. Due to node merging, only one terminal node remains in a trellis. Analogous to a code tree, a path from the single origin node to the single terminal node in a trellis also mirrors a codeword. Figure 4 exemplifies the trellis of the convolutional code presented in Fig. 1.

3. TYPICAL CHANNEL MODELS FOR CODING SYSTEMS

When the $n(L + m)$ convolutional code bits encoded from kL input bits are modulated into respective waveforms (or signals) for transmission over a medium that introduces attenuation, distortion, interference, noise, and other parameters, the received waveforms become "uncertain" in their shapes. A "guess" of the original information sequences therefore has to be made at the receiver end. The "guess" mechanism can be conceptually divided into two parts: demodulator and decoder.

The demodulator transforms the received waveforms into discrete signals for use by the decoder to determine the original information sequences. If the discrete demodulated signal is of two values (i.e., binary), then the demodulator is termed a *hard-decision demodulator*. If the demodulator passes analog (i.e., discrete-in-time but continuous-in-value) or quantized outputs to the decoder, then it is classified as a *soft-decision demodulator*.

[1] The effective code rate is defined as the average number of input bits carried by an output bit [1].

[2] Usually, $d_c(r)$ for an (n, k, m) convolutional code reaches its largest value $d_c(\infty)$ when r is a little beyond $5 \times m$; this property facilitates the determination of $d_c(\infty)$.

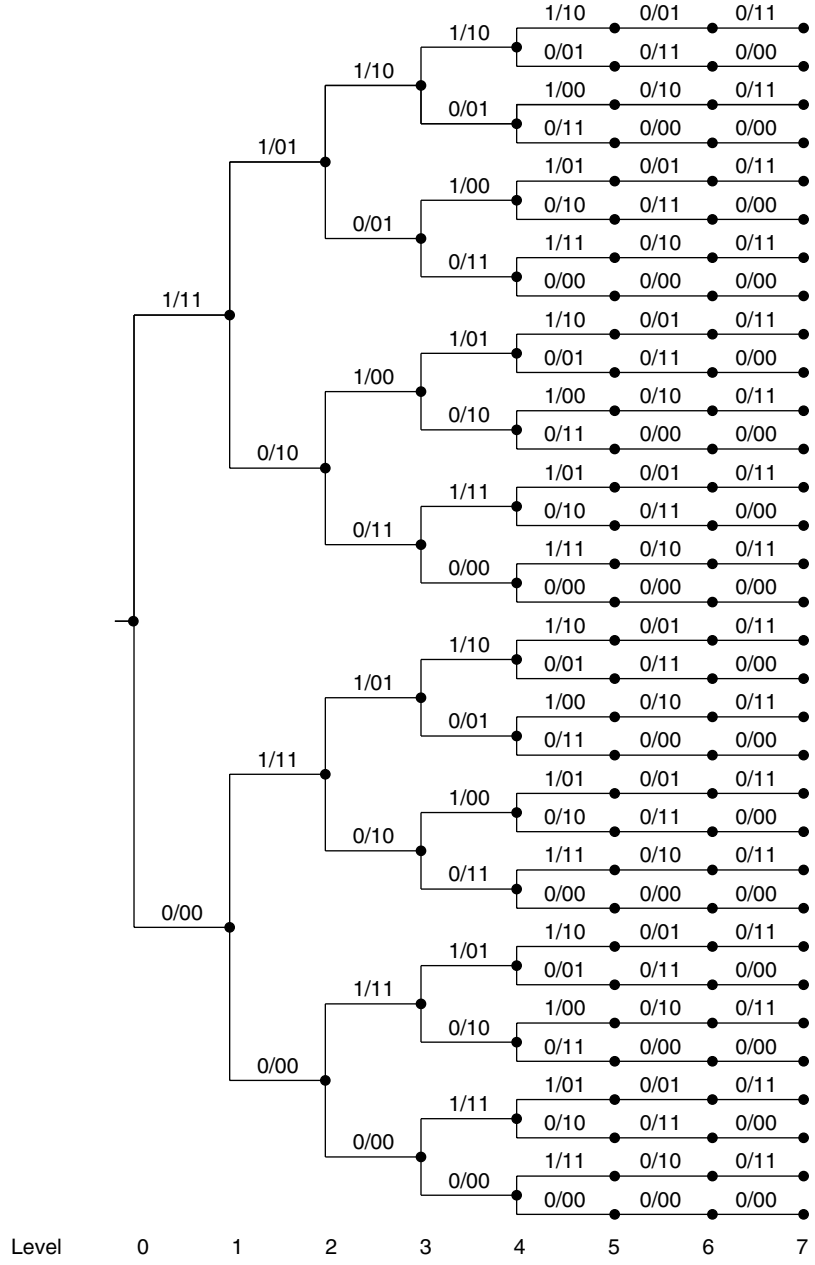


Figure 3. Code tree for the binary $(2, 1, 2)$ convolutional code in Fig. 1 with a single input sequence of length 5. Each branch is labeled by its respective "input bit/output code bits." The code path indicated by the thick line is labeled in sequence by code bits 11, 01, 10, 01, 00, 10 and 11, and its corresponding codeword is $\mathbf{v} = (11\ 01\ 10\ 01\ 00\ 10\ 11)$.

The decoder, on the other hand, estimates the original information sequences on the basis of the $n(L + m)$ demodulator outputs, or equivalently a received vector of $n(L + m)$ dimensions, according to some criterion. One frequently applied criterion is the *maximum-likelihood decoding* (MLD) rule, under which the probability of codeword estimate error is minimized subject to an equiprobable prior on the transmitted codewords. Terminologically, if a soft-decision demodulator is employed, then the subsequent decoder is classified as a *soft-decision decoder*. In a situation in which the decoder receives inputs from a hard-decision demodulator, the decoder is called a *hard-decision decoder* instead.

Perhaps, because of their analytic feasibility, two types of statistics concerning demodulator outputs are of general interest. They are respectively induced from the *binary symmetric channel* (BSC) and the *additive white Gaussian noise* (AWGN) *channel*. The former is a typical channel model for the performance evaluation of hard-decision decoders, while the latter is widely used in examining the error rate of soft-decision decoders. They are introduced after the concept of a channel is elucidated.

For a coding system, a *channel* is a signal passage that mixes all the intermediate effects onto the signal, including modulation, upconversion, medium, downconversion, and demodulation. The demodulator incorporates these aggregated channel effects into a widely adopted additive channel model as $r = s + n$, in which r is the demodulator output, s is the transmitted signal that is a function of encoder outputs, and n represents the aggregated signal distortion, simply termed *noise*. Its extension to multiple

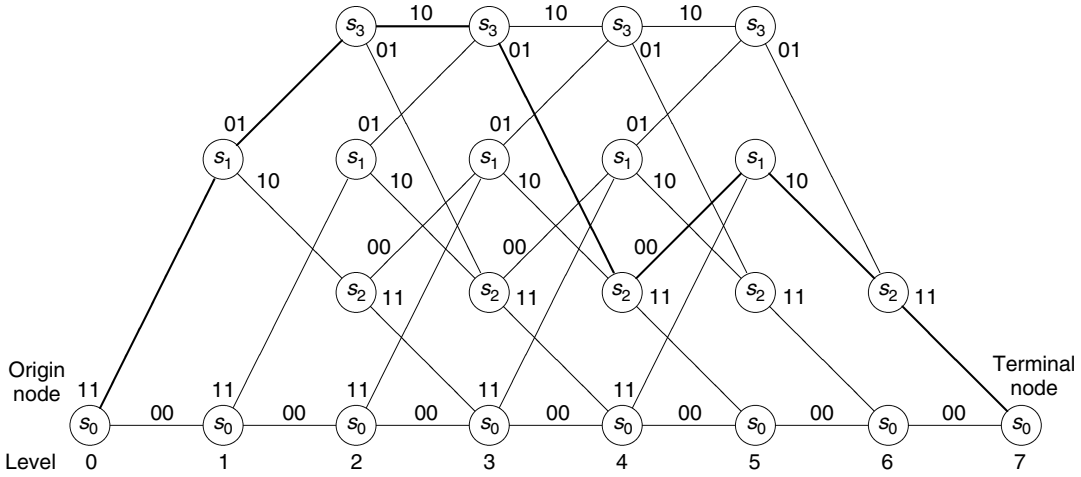


Figure 4. Trellis for the binary $(2, 1, 2)$ convolutional code in Fig. 1 with a single input sequence of length 5. States S_0, S_1, S_2, and S_3 correspond to the states of shift register contents that are 00, 01, 10, and 11 (from right to left in Fig. 1), respectively. The code path indicated by the thick line is labeled in sequence by code bits 11, 01, 10, 01, 00, 10 and 11, and its corresponding codeword is $\mathbf{v} = (11\ 01\ 10\ 01\ 00\ 10\ 11)$.

independent channel usages is given by

$$r_j = s_j + n_j \text{ for } 0 \le j \le N-1$$

which is often referred to as the *time-discrete channel*, since the *time* index j ranges over a *discrete* integer set. For simplicity, independence with common marginal distribution among noise samples $n_0, n_1, \ldots, n_{N-1}$ is often assumed, which is specifically termed *memoryless*. In situation where the power spectrum (i.e., the Fourier transform of the noise autocorrelation function) of the noise samples is a constant, which can be interpreted as the noise contributing equal power at all frequencies and thereby imitating the composition of a white light, the noise is dubbed *white*.

Hence, for a time-discrete coding system, the AWGN channel specifically indicates a memoryless noise sequence with a Gaussian distributed marginal, in which case the demodulator outputs $r_0, r_1, \ldots, r_{N-1}$ are independent and Gaussian distributed with equal variances and means $s_0, s_1, \ldots, s_{N-1}$, respectively. The noise variance exactly equals the constant spectrum value $N_0/2$ of the white noise, where N_0 is *the single-sided noise power per hertz* or $N_0/2$ is *the doubled-sided noise power per hertz*. The means that $s_0, s_1, \ldots, s_{N-1}$ are apparently decided by the choice of mappings from the encoder outputs to the channel inputs. For example, assuming an *antipodal* mapping gives $s_j(c_j) = (-1)^{c_j}\sqrt{E}$, where $c_j \in \{0, 1\}$ is the jth code bit. Under an implicit premise of equal possibilities for $c_j = 0$ and $c_j = 1$, the second moment of s_j is given by

$$E[s_j^2] = \tfrac{1}{2}(\sqrt{E})^2 + \tfrac{1}{2}(-\sqrt{E})^2 = E$$

which is commonly taken to be the average signal energy required for its transmission.

A conventional measure of the noisiness of AWGN channels is the *signal-to-noise ratio* (SNR). For the time-discrete system considered, it is defined as the average signal energy E (the second moment of the transmitted signal) divided by N_0 (the single-sided noise power per hertz). Notably, the SNR ratio is invariable with respect to scaling of the demodulator output; hence, this noisiness index is consistent with the observation that the optimal error rate of guessing c_j based on the knowledge of $(\lambda \cdot r_j)$ through equation

$$\lambda \cdot r_j = \lambda \cdot (-1)^{c_j}\sqrt{E} + \lambda \cdot n_j$$

is indeed independent of the scaling factor λ whenever $\lambda > 0$. Accordingly, the performance of the soft-decision decoding algorithms under AWGN channels is typically given by plotting its error rate against the SNR.[3]

The channel model can be further simplified to that for which the noise sample n and the transmitted signal s (usually the code bit itself in this case) are both elements of $\{0, 1\}$, and their modulo-2 addition yields the hard-decision demodulation output r. Then a binary input/binary output channel between convolutional encoder and decoder is observed. The channel statistics can be defined using the two crossover probabilities: $\Pr(r = 1 \mid s = 0)$ and $\Pr(r = 0 \mid s = 1)$. If the two crossover probabilities are equal, then the binary channel is *symmetric*, and is therefore called the *binary symmetric channel*.[4]

[3] The SNR per information bit, denoted as E_b/N_0, is often used instead of E/N_0 in picturing the code performance in order to account for the code redundancy for different code rates. Their relation can be characterized as $E_b/N_0 = (E/N_0)/R_{\text{effective}} = (E/N_0) \times n(L+m)/(kL)$ because the overall energy of kL uncoded input bits equals that of $n(L+m)$ code bits in principle.

[4] The BSC can be treated as a quantized simplification of the AWGN channel. Hence, the crossover probability p can be derived from $r_j = (-1)^{c_j}\sqrt{E} + n_j$ as $p = (1/2)\text{erfc}(\sqrt{E/N_0})$, where $\text{erfc}(x) = (2/\sqrt{\pi})\int_x^\infty \exp\{-x^2\}\,dx$ is the complementary error function. This convention is adopted here in presenting the performance figures for BSCs.

The correspondence between the transmitted signal s_j and the code bit c_j is often isomorphic. If this is the case, $\Pr(r_j \mid c_j)$ and $\Pr(r_j \mid s_j)$ can be used interchangeably to represent the channel statistics of receiving r_j given that c_j or $s_j = s_j(c_j)$ is transmitted. For convenience, $\Pr(r_j \mid v_j)$ will be used at the decoder end to denote the same probability as $\Pr(r_j \mid c_j)$, where $c_j = v_j$, throughout the article.

4. GENERAL DESCRIPTION OF SEQUENTIAL DECODING ALGORITHM

Following the background introduction to the time-discrete coding system, the optimal criterion that motivates the decoding approaches can now be examined. As a consequence of minimizing the codeword estimate error subject to an equiprobable codeword prior, the MLD rule, on receipt of a received vector $\mathbf{r} = (r_0, r_1, \ldots, r_{N-1})$, outputs the codeword $\mathbf{c}^* = (c_0^*, c_1^*, \ldots, c_{N-1}^*)$ satisfying

$$\Pr(\mathbf{r} \mid \mathbf{c}^*) \geq \Pr(\mathbf{r} \mid \mathbf{c}) \text{ for all } \mathbf{c} = (c_0, c_1, \ldots, c_{N-1}) \in \mathcal{C}$$

where $\mathcal{C}$ is the set of all possible codewords, and $N = n(L + m)$. When the channel is memoryless, the MLD rule can be reduced to

$$\prod_{j=0}^{N-1} \Pr(r_j \mid c_j^*) \geq \prod_{j=0}^{N-1} \Pr(r_j \mid c_j) \quad \text{for all} \quad \mathbf{c} \in \mathcal{C}$$

which in turn is equivalent to

$$\sum_{j=0}^{N-1} \log_2 \Pr(r_j \mid c_j^*) \geq \sum_{j=0}^{N-1} \log_2 \Pr(r_j \mid c_j) \quad \text{for all} \quad \mathbf{c} \in \mathcal{C} \quad (1)$$

A natural implication of (1) is that by simply letting $\sum_{j=n(\ell-1)}^{n\ell-1} \log_2 \Pr(r_j \mid c_j)$ be the metric associated with a branch labeled by $(c_{n(\ell-1)}, \ldots, c_{n\ell-1})$, the MLD rule becomes a search of the code path with the maximum metric, where the metric of a path is defined as the sum of the individual metrics of the branches of which the path consists. Any suitable graph search algorithm can then be used to perform the search process.

Of the graph search algorithms in artificial intelligence, Algorithm A is one that performs priority-first (or metric-first) searching over a graph [7,12]. In applying the algorithm to the decoding of convolutional codes, the graph G undertaken becomes either a code tree or a trellis. For a graph over which Algorithm A searches, a link between the origin node and any node, either directly connected or indirectly connected through some intermediate nodes, is called a *path*. Suppose that a real-valued function $f(\cdot)$, often referred to as the *evaluation function*, is defined for every path in the graph G. Then Algorithm A can be described as follows:

Algorithm A

Step 1. Compute the associated f-function value of the single-node path that contains only the origin node. Insert the single-node path with its associated f-function value into the stack.

Step 2. Generate all immediate successor paths of the top path in the stack, and compute their f-function values. Delete the top path from the stack.

Step 3. If the graph G is a trellis, check whether these successor paths end at a node that belongs to a path that is already in the stack. Restated, check whether these successor paths merge with a path that is already in the stack. If it does, and the f-function value of the successor path exceeds the f-function value of the subpath that traverses the same nodes as the merged path in the stack but ends at the merged node, redirect the merged path by replacing its subpath with the successor path, and update the f-function value associated with the newly redirected path.[5] Remove those successor paths that merge with some paths in the stack. (Note that if the graph G is a code tree, there is a unique path connecting the origin node to each node on the graph; hence, it is unnecessary to examine the path merging.)

Step 4. Insert the remaining successor paths into the stack, and reorder the stack in descending f-function values.

Step 5. If the top path in the stack ends at a terminal node in the graph G, the algorithm stops; otherwise go to step 2.

In principle, the *evaluation function* $f(\cdot)$ that guides the search of Algorithm A is the sum of two parts, $g(\cdot)$ and $h(\cdot)$; both range over all possible paths in the graph. The first part, $g(\cdot)$, is simply a function of all the branches traversed by the path, while the second part, $h(\cdot)$, called the *heuristic function*, help to predict a future route from the end node of the current path to a terminal node. Conventionally, the heuristic function $h(\cdot)$ equals zero for all paths that end at a terminal node. Additionally, $g(\cdot)$ is usually taken to be zero for the single-node path that contains only the origin node.

A question that follows is how to define $g(\cdot)$ and $h(\cdot)$ so that Algorithm A performs the MLD rule. Here, this question is examined by considering a more general problem of how to define $g(\cdot)$ and $h(\cdot)$ so that Algorithm A locates the code path with maximum metric over a code tree or a trellis. Suppose that a metric $c(n_i, n_j)$ is associated with the branch between nodes n_i and n_j. Define the metric of a path as the sum of the metrics of those branches contained by the path. The g-function value for a path can then be assigned as the sum of all the branch metrics experienced by the path. Let the h-function value of the same path be an estimate of the maximum cumulative metric from the end node of the path to a terminal node. Under such a system setting, if the heuristic function satisfies certain optimality criteria, such as whether it always upper-bounds the maximum cumulative metric

[5] The redirect procedure is sometimes time-consuming, especially when the f-function value of a path is computed based on the branches the path traverses. Section 8 introduces two approaches to reduce the burden of path redirection.

from the end node of the path of interest to any terminal node, then Algorithm A guarantees the finding of the maximum-metric code path.[6]

Following the discussion in the previous paragraph and the observation from Eq. (1), a straightforward definition for function $g(\cdot)$ is

$$g(\mathbf{v}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} \log_2 \Pr(r_j \mid v_j) \tag{2}$$

where $\mathbf{v}_{(\ell n-1)}$ is the label sequence of the concerned path and the branch metric between the end nodes of path $\mathbf{v}_{(\ell(n-1)-1)}$ and path $\mathbf{v}_{(\ell n-1)}$ is given by $\sum_{j=\ell(n-1)}^{\ell n-1} \log_2 \Pr(r_j \mid v_j)$. Various heuristic functions with respect to the g function defined above can then be developed. For example, if the branch metric defined above is nonpositive, which apparently holds when the received demodulator output r_j is discrete for $0 \le j \le N-1$, a heuristic function that equals zero for all paths sufficiently upper-bounds the maximum cumulative metric from the end node of the concerned path to a terminal node, and thereby guarantees the finding of the code path with maximum metric. Another example is the well-known Fano path metric [4], which, by its formula, can be equivalently interpreted as the sum of the g-function defined in (2) and a specific h function. The details regarding the Fano metric will be given in the next section.

Notably, the branch metric used to define (2) depends not only on the labels of the concerned branch (i.e., $v_{\ell(n-1)}, \ldots, v_{\ell n-1}$) but also on the respective demodulator outputs (i.e., $r_{\ell(n-1)}, \ldots, r_{\ell n-1}$). Some researchers also view the received vector $\mathbf{r} = (r_0, r_1, \ldots, r_{N-1})$ as labels for another (possibly nonbinary) code tree or trellis; hence the term "received branch," which reflects a branch labeled by the respective portion of the received vector $\mathbf{r}$ was introduced. With such a naming convention, this section concludes by quoting the essential attributes of sequential decoding defined in Ref. 15. According to the authors, the very first attribute of sequential decoding is that "the branches are examined sequentially, so that at any node of the tree the decoder's choice among a set of previously unexplored branches does not depend on received branches deeper in the tree." The second attribute is that "the decoder performs at least one computation for each node of every examined path." The authors then remark at the end that "Algorithms which do not have these two properties are not considered to be sequential decoding algorithms." Thus, an easy way to visualize the defined features of sequential decoding is that the received scalars $r_0, r_1, \ldots, r_{N-1}$ are received *sequentially* in time in order of the subindices during the decoding process. The next path to be examined therefore cannot be in any sense related to the received scalars whose subindices are beyond the deepest level of the paths that are momentarily in the stack, because random usage, rather than sequential usage, of the received scalars (such as the usage of r_j, followed by the usage of r_{j+2} instead of r_{j+1}) implicitly indicates that all the received scalars should be ready in a buffer before the decoding process begins.

Based on the two attributes, the sequential decoding is simply Algorithm A with an evaluation function $f(\cdot)$ equal to the sum of the branch metrics of those branches contained by the examined path (i.e., the path metric), where the branch metric is a function of the branch labels and the respective portion of the received vector. Variants of sequential decoding therefore mostly reside on different path metrics adopted. The subsequent sections show that taking a general view of Algorithm A, rather than a restricted view of sequential decoding, promotes the understanding of various later generalizations of sequential decoding.

The next section introduces the most well-known path metric for sequential decoding, which is named after its discover, R. M. Fano.

5. FANO METRIC AND ITS GENERALIZATION

Since its discovery in 1963 [4], the Fano metric has become a typical path metric in sequential decoding. Originally, the Fano metric was discovered through mass simulations, and was first used by Fano in his sequential decoding algorithm on a code tree [4]. For any path $\mathbf{v}_{(\ell n-1)}$ that ends at level ℓ on a code tree, the *Fano metric* is defined as

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} M(v_j \mid r_j)$$

where $\mathbf{r} = (r_0, r_1, \ldots, r_{N-1})$ is the received vector and

$$M(v_j \mid r_j) = \log_2 \frac{\Pr(r_j \mid v_j)}{\Pr(r_j)} - R$$

is the *bit metric*, and the calculation of $\Pr(r_j)$ follows the convention that the code bits — 0 and 1 — are transmitted with equal probability

$$\begin{aligned} \Pr(r_j) &= \sum_{v_j \in \{0,1\}} \Pr(v_j) \Pr(r_j \mid v_j) \\ &= \tfrac{1}{2} \Pr(r_j \mid v_j = 0) + \tfrac{1}{2} \Pr(r_j \mid v_j = 1) \end{aligned}$$

and $R = k/n$ is the code rate. For example, a hard-decision decoder with $\Pr\{r_j = 0 \mid v_j = 1\} = \Pr\{r_j = 1 \mid v_j = 0\} = p$ for $0 \le j \le N-1$ (i.e., a memoryless BSC channel with crossover probability p), where $0 < p < \frac{1}{2}$, will interpret the Fano metric for path $\mathbf{v}_{(\ell n-1)}$ as

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} \log_2 \Pr(r_j \mid v_j) + \ell n(1-R) \tag{3}$$

where

$$\log_2 \Pr(r_j \mid v_j) = \begin{cases} \log_2(1-p), & \text{for} \quad r_j = v_j \\ \log_2(p), & \text{for} \quad r_j \ne v_j \end{cases}$$

[6] Criteria that guarantee optimal decoding by the Algorithm A are extensively discussed in Refs. 13 and 14.

In terms of the Hamming distance, (3) can be rewritten as

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = -\alpha \cdot d_H(\mathbf{r}_{(\ell n-1)}, \mathbf{v}_{(\ell n-1)}) + \beta \cdot \ell \qquad (4)$$

where $\alpha = -\log_2[p/(1-p)] > 0$, and $\beta = n[1 - R + \log_2(1-p)]$. An immediate observation from (4) is that a larger Hamming distance between the path labels and the respective portion of the received vector corresponds to a smaller path metric. This property guarantees that if no error exists in the received vector (i.e., the bits demodulated are exactly the bits transmitted), and $\beta > 0$ (or equivalently, $R < 1 + \log_2(1-p)$),[7] then the path metric increases along the correct code path, and the path metric along any incorrect path is smaller than that of the equally long correct path. Such a property is essential for a metric to work properly with sequential decoding.

Later, Massey [17] proved that at any decoding stage, extending the path with the largest Fano metric in the stack minimizes the probability that the extending path does not belong to the optimal code path, and the usage of the Fano metric for sequential decoding is thus analytically justified. However, making such a *locally* optimal decision at every decoding stage does not always guarantee the ultimate finding of the *globally* optimal code path in the sense of the MLD rule in (1). Hence, the error performance of sequential decoding with the Fano metric is in general slightly inferior to that of the MLD-rule-based decoding algorithm.

A striking feature of the Fano metric is its dependence on the code rate R. Introducing the code rate into the Fano metric somehow reduces the complexity of the sequential decoding algorithm. Observe from (3) that the first term, $\sum_{j=0}^{\ell n-1} \log_2 \Pr(r_j \mid v_j)$, is the part that reflects the maximum-likelihood decision in (1), and the second term, $\ell n(1-R)$, is introduced as a bias to favor a longer path or specifically a path with larger ℓ, since a longer path is closer to the leaves of a code tree and thus is more likely to be part of the optimal code path. When the code rate increases, the number of incorrect paths for a given output length increases.[8] Hence, the confidence on the currently examined path being part of the correct code path should be weaker. Therefore, the claim that longer paths are part of the optimal code path is weaker at higher code rates. The Fano metric indeed mirrors the above intuitive observation by using a linear bias with respect to the code rate.

[7] The code rate bound below which the Fano-metric-based sequential decoding performs well is the *channel capacity*, which is $1 + p\log_2(p) + (1-p)\log_2(1-p)$ in this case. The alternative larger bound $1 + \log_2(1-p)$, derived from $\beta > 0$, can only justify the subsequent argument, and by no means ensure a good performance for sequential decoding under $1 + p\log_2(p) + (1-p)\log_2(1-p) < R < 1 + \log_2(1-p)$. Channel capacity is beyond the scope of this article. Interested readers can refer to the treatise by Cover and Thomas [16].

[8] Imagine that the number of branches that leave each node is 2^k, and increasing the code rate can be conceptually interpreted as increasing k subject to a fixed n for a (n, k, m) convolutional code.

The effect of taking other bias values has been examined in Refs. 18 and 19. The authors defined a new bit metric for sequential decoding as

$$M_B(r_j \mid v_j) = \log_2 \frac{\Pr(r_j \mid v_j)}{\Pr(r_j)} - B \qquad (5)$$

and found that a tradeoff between computational complexity and error performance of sequential decoding can be observed by varying the bias B.

Researchers began to investigate the effect of a joint bias on $\Pr(r_j)$ and R, providing new generalization of sequential decoding. Han et al. [20] observed that universally adding a constant to the Fano metric of all paths does not change the sorting result at each stage of the sequential decoding algorithm (see step 4 of Algorithm A). They then chose the additive constant $\sum_{j=0}^{N-1} \log_2 \Pr(r_j)$, and found that

$$M(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) + \sum_{j=0}^{N-1} \log_2 \Pr(r_j) = \sum_{j=0}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R] + \sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j) \qquad (6)$$

for which the two terms on the right-hand side of (6) can be immediately reinterpreted as the g function and the h function from the perspective of Algorithm A.[9] As the g function is now defined based on the branch metric $\sum_{j=\ell(n-1)}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R]$, Algorithm A, according to the discussion in the previous section, becomes a search to find the code path $\mathbf{v}^*$ that satisfies

$$\sum_{j=0}^{N-1} [\log_2 \Pr(r_j \mid v_j^*) - R] \geq \sum_{j=0}^{N-1} [\log_2 \Pr(r_j \mid v_j) - R] \quad \text{for all code path } \mathbf{v}$$

This criterion is equivalent to the MLD rule in (1). Consequently, the Fano path metric indeed implicitly uses $\sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j)$ as a heuristic estimate of the upcoming metric from the end node of the current path to a terminal node. A question that naturally follows regards the trustworthiness of this estimate. The question can be directly answered by studying the effect of varying weights

[9] Notably, defining a path metric as $\sum_{j=0}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R] + \sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j)$ does not yield a sequential decoding algorithm according to the definition of sequential decoding in Ref. 15, for such a path metric depends on information of "the received branches" beyond level ℓ, i.e., $r_{\ell n}, \ldots, r_{N-1}$. However, a similarly defined evaluation function surely gives Algorithm A.

on the g function (i.e., the cumulative metric sum that is already known) and h function (i.e., the estimate) using

$$f_\omega(\mathbf{v}_{(\ell n-1)}) = \omega \sum_{j=0}^{\ell n-1} [\log_2 \Pr(r_j \mid v_j) - R] + (1-\omega) \sum_{j=\ell n}^{N-1} \log_2 \Pr(r_j) \tag{7}$$

where $0 \le \omega \le 1$. Subtracting a universal constant $(1-\omega)\sum_{j=0}^{N-1} \Pr(r_j)$ from (7) gives the *generalized Fano metric* for sequential decoding as

$$M_\omega(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} (\log_2 \frac{\Pr(r_j \mid v_j)^\omega}{\Pr(r_j)^{1-\omega}} - \omega R) \tag{8}$$

When $\omega = \frac{1}{2}$, the generalized Fano metric reduces to the Fano metric with a multiplicative constant, $\frac{1}{2}$. As ω is slightly below $\frac{1}{2}$, which can be interpreted from (7) as the sequential search is guided more by the estimate on the upcoming metrics than by the known cumulative metric sum, the number of metric computations reduces but the decoding failure probability grows. When ω is closer to one, the decoding failure probability of sequential decoding tends to be lower; however, the computational complexity increases. In the extreme case, taking $\omega = 1$ makes the generalized Fano metric completely mirror the MLD metric in (1), and the sequential decoding becomes a maximum-likelihood (hence, optimal in decoding failure probability) decoding algorithm. The work in Ref. 20 thereby led to the conclusion that an (implicit) heuristic estimate can be elaborately defined to reduce fairly the complexity of sequential decoding with a slight degradation in error performance. Notably, for discrete symmetric channels, the generalized Fano metric is equivalent to the metric defined in Eq. (5) [21]. However, the generalized Fano metric and the metric of (5) are by no means equal for other types of channels such as the AWGN channel.

6. STACK ALGORITHM AND ITS VARIANTS

The stack algorithm or the ZJ algorithm was discovered by Zigangirov [5] and later independently by Jelinek [6] to search a code tree for the optimal codeword. It is exactly the Algorithm A with g-function equal to the Fano metric and zero h function. Because a stack is involved in searching for the optimal codeword, the algorithm is called the *stack algorithm*. An example is provided blow to clarify the flow of the stack algorithm.

Example 1. For a BSC with crossover probability $p = .045$, the Fano bit metric for a convolutional code with code rate $R = \frac{1}{2}$ can be obtained from (3) as

$$M(v_j \mid r_j) = \begin{cases} \log_2(1-p) + (1-R) = 0.434, & \text{for } r_j = v_j \\ \log_2(p) + (1-R) = -3.974, & \text{for } r_j \ne v_j \end{cases}$$

Consequently, only two Fano bit metric values are possible, 0.434 and −3.974. These two Fano bit metric values can be "scaled" to equivalent "integers" to facilitate the simulation and implementation of the system. Taking the multiplicative scaling factor of 2.30415 yields

$$M_{\text{scaled}}(v_j \mid r_j) = \begin{cases} 0.434 \times 2.30415 \approx 1, & \text{for } r_j = v_j \\ -3.974 \times 2.30415 \approx -9, & \text{for } r_j \ne v_j \end{cases}$$

Now, the convolutional code in Fig. 1 is decoded over its code tree (cf. Fig. 3) using the stack algorithm with the scaled Fano metric. Assume that the received vector is **r**=(11 01 00 01 10 10 11). Figure 5 presents the contents

Loop 1	Loop 2	Loop 3	Loop 4	Loop 5
1 (1 + 1 = 2)	11 (2 + 1 + 1 = 4)	111 (4 – 9 + 1 = –4)	1110 (–4 + 1 + 1 = –2)	110 (–4)
0 (–9 –9 = – 18)	10 (2 –9 –9 = –16)	110 (4 + 1 – 9 = –4)	110 (–4)	11100 (–2 + 1 – 9 = –10)
	0 (–18)	10 (–16)	10 (–16)	11101 (–2 –9 + 1 = –10)
		0 (–18)	0 (–18)	10 (–16)
			1111 (–4 –9 –9 = –22)	0 (–18)
				1111 (–22)

Loop 6	Loop 7	Loop 8	Loop 9
11100 (–10)	11101 (–10)	111010 (–10 + 1 + 1 = –8)	1110100 (–8 + 1 + 1 = –6)
11101 (–10)	1100 (–12)	1100 (–12)	1100 (–12)
1100 (–4 –9 + 1 = –12)	1101 (–12)	1101 (–12)	1101 (–12)
1101 (–4 + 1 – 9 = –12)	10 (–16)	10 (–16)	10 (–16)
10 (–16)	111000 (–10 – 9 + 1 = –18)	111000 (–18)	111000 (–18)
0 (–18)	0 (–18)	0 (–18)	0 (–18)
1111 (–22)	1111 (–22)	1111 (–22)	1111 (–22)

Figure 5. Stack contents after each path metric reordering in Example 1. Here, different from that used in the Fano metric computation, the input bit labels rather than the code bit labels are used for each recorded path. The associated Fano metric follows each path label sequence (inside parentheses).

of the stack after each path metric reordering. Each path in the stack is marked by its corresponding input bit labels rather than by the code bit labels. Notably, while both types of labels can uniquely determine a path, the input bit labels are more frequently recorded in the stack in practical implementation since the input bit labels are the desired estimates of the transmitted information sequences. Code bit labels are used more often in metric computation and in characterizing the code, because the code characteristic, such as error-correcting capability, can only be determined from the code bit labels (codewords).[10] The path metric associated with each path is also stored. The algorithm is terminated at the ninth loop, yielding an ultimate decoding result of $\mathbf{u} = (11101)$.

Maintaining the stack is a significant implementation issue of the stack algorithm. In a straightforward implementation of the stack algorithm, the paths are stored in the stack in order of descending f-function values; hence, a sorting mechanism is required. Without a proper design, the sorting of the paths within the stack may be time-consuming, limiting the speed of the stack algorithm.

Another implementation issue of the stack algorithm is that the stack size in practice is often insufficient to accommodate the potentially large number of paths examined during the search process. The stack can therefore overflow. A common way of compensating for a stack overflow is to simply discard the path with the smallest f-function value [1], since it is least likely to be the optimal code path. However, when the discarded path happens to be an early predecessor of the optimal code path, performance is degraded.

Jelinek proposed the so-called *stack-bucket technique* to reduce the sorting burden of the stack algorithm [6]. In his proposal, the range of possible path metric values is divided into a number of intervals with prespecified, fixed spacing. For each interval, a separate storage space, a *bucket*, is allocated. The buckets are then placed in order of descending interval endpoint values. During decoding, the next path to be extended is always the top path of the first nonempty bucket, and every newly generated path is directly placed on top of the bucket in which interval the respective path metric lies. Some data structure can be used to reduce the maintenance burden and storage demand of stacked buckets, since some buckets corresponding to a certain metric range may occasionally (or even always) be empty during decoding. The sorting burden is therefore removed by introducing the stacked buckets. The time taken to locate the next path no longer depends on the size of the stack, rather on the number of buckets, considerably reducing the time required by decoding. Consequently, the stacked bucket technique was used extensively in the software implementation of the stack algorithm for applications in which the decoding time is precisely restricted [1,22,23]

The drawback of the stacked bucket technique is that the path with the best path metric may not be selected, resulting in degradation in performance. A *metric-first stacked bucket* implementation overcomes the drawback by sorting the top bucket when it is being accessed. However, Anderson and Mohan [24] indicated that the access time of the metric-first stacked buckets will increase at least to the order of $S^{1/3}$, where S is the total number of the paths ever generated.

Another software implementation technique for establishing a sorted stack was discussed by Mohan and Anderson [25], who suggested the adoption of a *balanced binary tree* data structure, such as an AVL tree [26], to implement the stack, offering the benefit that the access time of the stack becomes of order $\log_2(S)$, where S represents the momentary stack size. Briefly, a balanced binary tree is a sorted structure with node insertion and deletion schemes such that its depth is maintained equal to the logarithm of the total number of nodes in the tree, whenever possible. As a result, inserting or deleting a path (which is now a node in the data structure of a balanced binary tree) in a stack of size S requires at most $\log_2(S)$ comparisons (i.e., the number of times the memory is accessed). The balanced binary tree technique is indeed superior to the metric-first stacked bucket implementation, when the stack grows beyond certain size. Detailed comparisons of time and space consumption of various implementation techniques of sequential decoding, including the Fano algorithm to be introduced in the next section, can be found in another article [24].

In 1994, a novel systolic priority queue, called the *parallel entry systolic priority queue* (PESPQ), was proposed to replace the stacked buckets [27]. Although it does not arrange the paths in the queue in strict order, the systolic priority queue technique can identify the path with the largest path metric within a constant time. This constant time was shown to be comparable to the time required to compute the metric of a new path. Experiments revealed that the PESPQ stack algorithm is several times faster than its stacked bucket counterpart. Most importantly, the invention of the PESPQ technique has given a seemingly promising future to hardware implementation of the stack algorithm.

As stated at the end of Section 4, one of the two essential features of sequential decoding is that the next visited path cannot be selected based on the basis of the information deeper in the tree [15]. The above feature is subsequently interpreted as the information that is deeper in the tree is supposed to be received in some future time and hence is perhaps not available at the current decoding stage. This conservative interpretation apparently arises from the aspect of an online decoder. A more general reinterpretation, simply following the wording, is that any codeword search algorithm that decides the next visited path in sequence without using the information deeper in its own search tree is considered to be a sequential decoding algorithm. This new interpretation precisely suits the bidirectional sequential decoding algorithm

[10] For a code tree, a path can be also uniquely determined by its end node in addition to the two types of path labels, so putting the "end node" rather than the path labels of a path in the stack suffices to fulfill the need for the tree-based stack algorithm. Nevertheless, such an approach, while easing the stack maintenance load, introduces an extra conversion load from the path end node to its respective input bit labels (as the latter is the desired estimates of the transmitted information sequence).

proposed by Forney [11], in which each of the two decoders still performs the defined sequential search in its own search tree. Specifically, Forney suggested that the sequential decoding could also start its decoding from the end of the received vector, and proposed a bidirectional stack algorithm in which two decoders simultaneously search the optimal code path from both ends of the code tree. The bidirectional decoding algorithm stops whenever either decoder reaches the end of its search tree. This idea has been greatly improved by Kallel and Li by stopping the algorithm whenever two stack algorithms with two separate stacks meet at a common node in their respective search trees [28]. Forney also claimed that the same idea can be applied to the Fano algorithm introduced in the next section.

7. FANO ALGORITHM

The Fano algorithm is a sequential decoding algorithm that does not require a stack [4]. The Fano algorithm can only operate over a code tree because it cannot examine path merging.

At each decoding stage, the Fano algorithm retains the information regarding three paths: the current path, its immediate predecessor path, and one of its successor paths. On the basis of this information, the Fano algorithm can move from the current path to either its immediate predecessor path or the selected successor path; hence, no stack is required for queuing all examined paths.

The movement of the Fano algorithm is guided by a dynamic threshold T that is an integer multiple of a fixed step size Δ. Only the path whose path metric is no less than T can be next visited. According to the algorithm, the process of codeword search continues to move forward along a code path, as long as the Fano metric along the code path remains nondecreasing. Once all the successor path metrics are smaller than T, the algorithm moves backward to the predecessor path if the predecessor path metric beats T; thereafter, threshold examination will be subsequently performed on another successor path of this revisited predecessor. In case the predecessor path metric is also less than T, the threshold T is one step lowered so that the algorithm is not trapped on the current path. For the Fano algorithm, if a path is revisited, the presently examined dynamic threshold is always lower than the momentary dynamic threshold at the previous visit, guaranteeing that looping in the algorithm does not occur, and that the algorithm can ultimately reach a terminal node of the code tree, and stop.

Figure 6 displays a flowchart of the Fano algorithm, in which $\mathbf{v}_p$, $\mathbf{v}_c$, and $\mathbf{v}_s$ represent the path label sequences of the predecessor path, the current path and the successor path, respectively. Their Fano path metrics are denoted by M_p, M_c, and M_s, respectively. The algorithm begins with the path that contains only the origin node. The label sequence of its predecessor path is initially set to "dummy," and the path metric of such a dummy path is assumed to be negative infinity. The initialization value of the dynamic threshold is zero.

The algorithm then proceeds to find, among the 2^k candidates, the successor path $\mathbf{v}_s$ with the largest path metric M_s. Thereafter, it examines whether $M_s \geq T$. If so, the algorithm moves forward to the successor path and updates the necessary information. Then, whether the new current path is a code path is determined, and a positive result immediately terminates the algorithm. A delicate part of the Fano algorithm is "threshold tightening." Whenever a path is first visited, the dynamic threshold T must be "tightened" such that it is adjusted to the largest possible value below the current path metric, namely, $T \leq M_c < T + \Delta$. Notably, the algorithm can determine whether a path is first visited by simply examining $\min\{M_p, M_c\} < T + \Delta$. If $\min\{M_p, M_c\} < T + \Delta$ holds because of the validity of $M_c < T + \Delta$, the threshold is automatically tightened; hence, only the condition $M_p < T + \Delta$ is required in the tightening test. The above mentioned procedures are repeated until a code path is finally reached.

Along the other route of the flowchart, following a negative answer to the examination of $M_s \geq T$ (which implicitly implies that the path metrics of all the successor paths of the current path are less than T), the algorithm must lower the threshold if M_p is also less than T; otherwise, a deadlock on the current path arises. Using the lowered threshold, the algorithm repeats the finding of the best successor path whose path metric exceeds the new threshold, and proceeds the followup steps.

The rightmost loop of the flowchart considers the case for $M_s < T$ and $M_p \geq T$. In this case, the algorithm can only move backward, since the predecessor path is the only one whose path metric is no smaller than the dynamic threshold. The information regarding the current path and the successor path should be subsequently updated. Yet, the predecessor path, as well as its associated path metric, should be recalculated from the current $\mathbf{v}_p$, because information about the predecessor's predecessor is not recorded. Afterward, the Fano algorithm checks for the existence of a new successor path $\mathbf{v}_t$ that is not the current successor path from which the algorithm has just moved, and whose associated path metric exceeds the current M_s. Restated, this step finds the best successor path other than those that have already been examined. If such a new successor path does not exist, then the algorithm seeks either to reduce the dynamic threshold or to move backward again, depending on whether $M_p \geq T$. In case such a new successor path $\mathbf{v}_t$ with metric M_t is located, the algorithm refocuses on the new successor path by updating $\mathbf{v}_s = \mathbf{v}_t$ and $M_s = M_t$, and repeats the entire process.

A specific example is provided below to help in understanding of the Fano algorithm.

Example 2. Assume the same convolutional code and received vector as in Example 1. Let the step size Δ be four. Figure 7 presents the traces of the Fano algorithm during its decoding.

In this figure, each path is again represented by its input bit labels. S and D denote the paths that contains only the origin node and the dummy path, respectively. According to the Fano algorithm, the possible actions taken include MFTT = "move forward and tighten the threshold," MF = "move forward only," MBS = "move

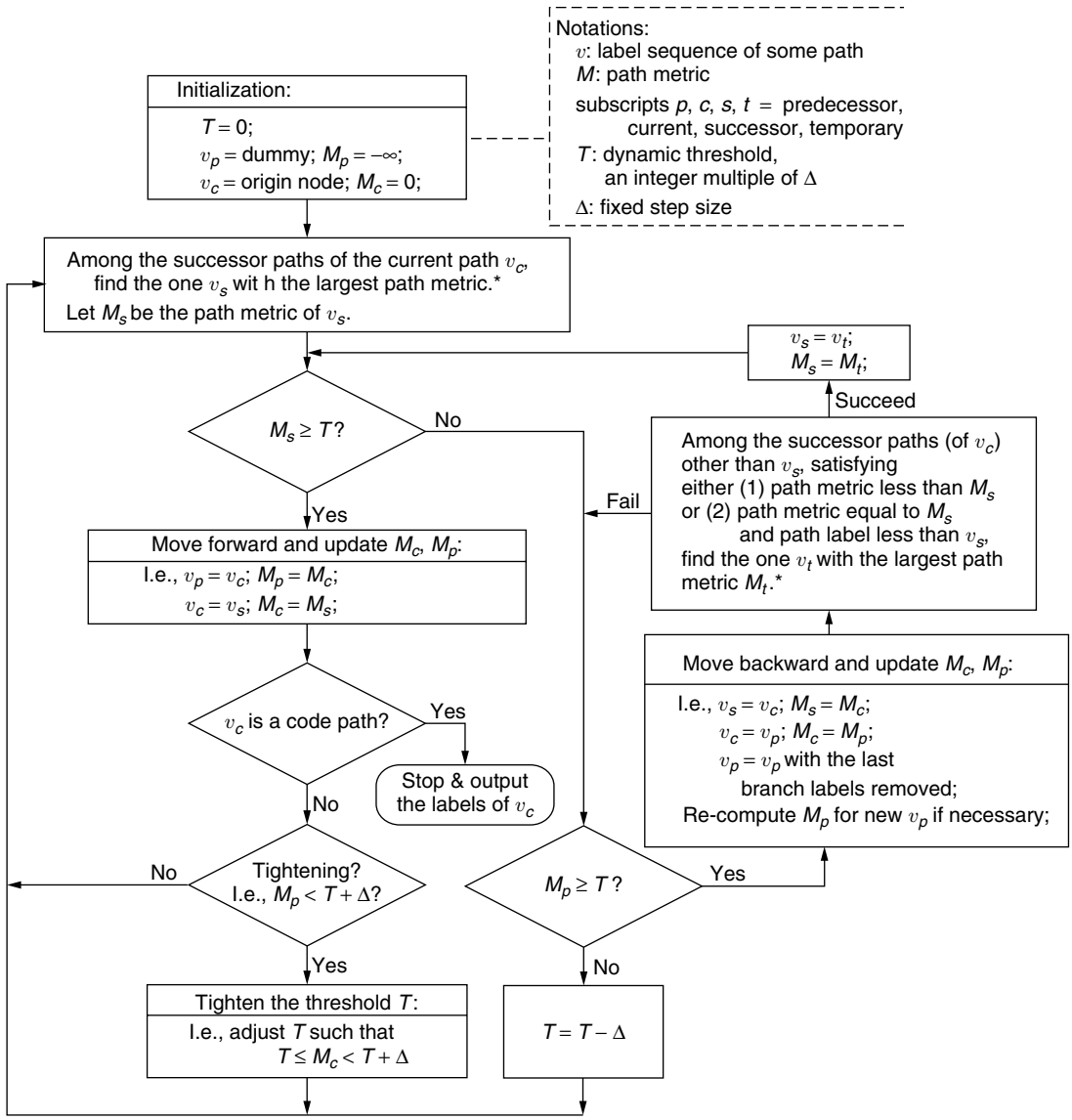


Figure 6. Flowchart of the Fano algorithm.

backward and successfully find the second best successor," MBF = "move backward but fail to find the second best successor," and LT = "lower threshold by one step." The algorithm stops after 36 iterations, and decodes the received vector to the same code path as that obtained in Example 1. This example clearly shows that the Fano algorithm revisits several paths more than once, such as path 11 (eight visits) and path 111 (five visits), and returns to path S three times.

As described in the previous example, the Fano algorithm may move backward to the path that contains only the origin node, and discover all its successors with path metrics less than T. In this case, the only action the algorithm can take is to keep decreasing the dynamic threshold until the algorithm can move forward again because the path metric of the predecessor path of the single-node path is set to $-\infty$. The impact of varying Δ should also be clarified. As stated by Lin and Costello [1], the load of branch metric computations executed during the finding of a qualified successor path becomes heavy when Δ is too small; however, when Δ is too large, the dynamic threshold T might be lowered too much in a single adjustment, forcing an increase in both the decoding error and the computational effort. Experiments suggest [1] that a better performance can be obtained by taking Δ within 2 and 8 for the unscaled Fano metric (Δ should be analogously scaled if a scaled Fano metric is used).

The Fano algorithm, perhaps surprisingly, while quite different in its design from the stack algorithm, exhibits broadly the same searching behavior as the stack algorithm. In fact, with a slight modification to the update procedure of the dynamic threshold (e.g., setting $\Delta = 0$, and substituting the "tightening test" and subsequent "tightening procedure" by "$M_c \neq T$?" and "$T = M_c$," respectively), both algorithms have been proved to visit almost the same set of paths during the decoding process [29]. Their only dissimilarity is that unlike the

Iteration	v_p	v_c	v_s	M_p	M_c	M_s	T	Action
0	*D*	*S*	1	$-\infty$	0	2	0	MFTT
1	*S*	1	11	0	2	4	0	MFTT
2	1	11	111	2	4	−4	4	LT
3	1	11	111	2	4	−4	0	MBS
4	*S*	1	10	0	2	−16	0	MBS
5	*D*	*S*	0	$-\infty$	0	−18	0	LT
6	*D*	*S*	1	$-\infty$	0	2	−4	MF
7	*S*	1	11	0	2	4	−4	MF
8	1	11	111	2	4	−4	−4	MF
9	11	111	1110	4	−4	−2	−4	MFTT
10	111	1110	11100	−4	−2	−10	−4	MBS
11	11	111	1111	4	−4	−22	−4	MBS
12	1	11	110	2	4	−4	−4	MF
13	11	110	1100	4	−4	−12	−4	MBF
14	1	11	110	2	4	−4	−4	MBS
15	*S*	1	10	0	2	−16	−4	MBS
16	*D*	*S*	0	$-\infty$	0	−18	−4	LT
17	*D*	*S*	1	$-\infty$	0	2	−8	MF
18	*S*	1	11	0	2	4	−8	MF
19	1	11	111	2	4	−4	−8	MF
20	11	111	1110	4	−4	−2	−8	MF
21	111	1110	11100	−4	−2	−10	−8	MBS
22	11	111	1111	4	−4	−22	−8	MBS
23	1	11	110	2	4	−4	−8	MF
24	11	110	1100	4	−4	−12	−8	MBF
25	1	11	110	2	4	−4	−8	MBS
26	*S*	1	10	0	2	−16	−8	MBS
27	*D*	*S*	0	$-\infty$	0	−18	−8	LT
28	*D*	*S*	1	$-\infty$	0	2	−12	MF
29	*S*	1	11	0	2	4	−12	MF
30	1	11	111	2	4	−4	−12	MF
31	11	111	1110	4	−4	−2	−12	MF
32	111	1110	11100	−4	−2	−10	−12	MF
33	1110	11100	111000	−2	−10	18	−12	MBS
34	111	1110	11101	−4	−2	−10	−12	MF
35	1110	11101	111010	−2	−10	−8	−12	MFTT
36	11101	111010	1110100	−10	−8	−6	−8	Stop

Figure 7. Decoding traces of the Fano algorithm for Example 2.

stack algorithm which visits each path only once, the Fano algorithm may revisit a path several times, and thus has a higher computational complexity. From the simulations over the binary symmetric channel as illustrated in Fig. 8, the stack algorithm with stacked bucket modification is apparently faster than the Fano algorithm at crossover probability $p = .057$, when their software implementations are concerned. The stack algorithm's superiority in computing time gradually disappears as p becomes smaller (or as the channel becomes less noisy) [22].

In practice, the time taken to decode an input sequence often has an upper limit. If the decoding process is not completed before the time limit, the undecoded part of the input sequence must be aborted or erased; hence, the probability of input erasure is another system requirement for sequential decoding. Figure 9 confirms that the stack algorithm with stacked bucket modification remains faster than the Fano algorithm except when either admitting a high erasure probability (e.g., erasure probability > 0.5 for $p = .045$, and erasure probability > 0.9 for $p = .057$) or experiencing a less noisier channel (e.g., $p = .033$) [22].

An evident drawback of the stack algorithm in comparison with the Fano algorithm is its demand for an extra stack space. However, with recent advances in computer technology, a large memory requirement is no longer a restriction for software implementation.

Hardware implementation is now considered. In hardware implementation, stack maintenance normally

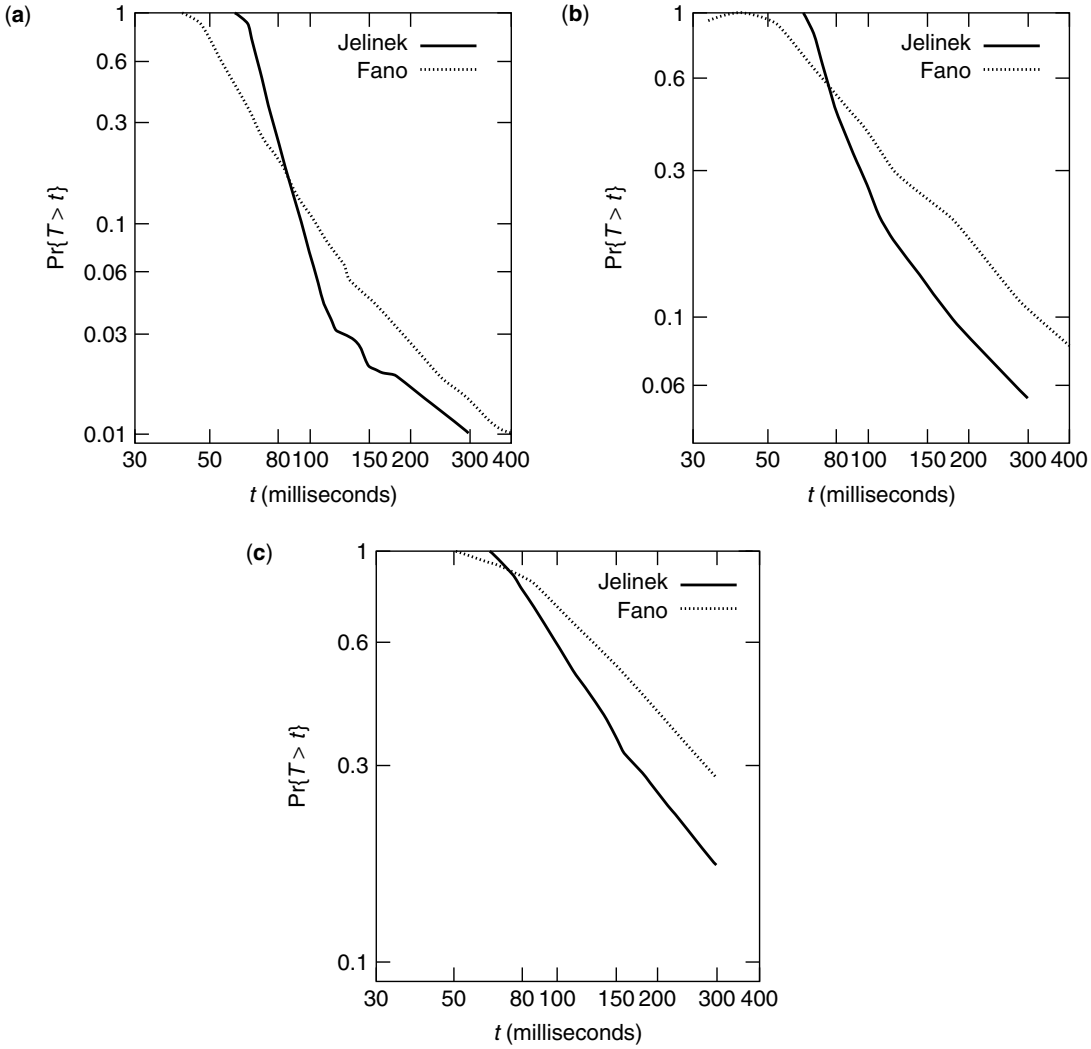


Figure 8. Comparisons of computational complexities of the Fano algorithm and the stack algorithm (with stacked bucket modification) based on the $(2, 1, 35)$ convolutional code with generator polynomials $g_1 = 53, 533, 676, 737$ and $g_2 = 733, 533, 676, 737$ and a single input sequence of length 256; simulations are performed over the binary symmetric channel with crossover probability p: (**a**) $p = .033$; (**b**) $p = .045$; (**c**) $p = .057$. $\Pr\{T \geq t\}$ is the empirical complement cumulative distribution function for the software computation time T. In simulations, $(\log_2[2(1-p)] - \frac{1}{2},\ \log_2(2p) - \frac{1}{2})$, which is derived from the Fano metric formula, is scaled to $(2, -18)$, $(2, -16)$ and $(4, -35)$ for $p = 0.033$, $p = 0.045$ and $p = 0.057$, respectively. In subfigures (a), (b) and (c), the parameters for the Fano algorithm are $\Delta = 16$, $\Delta = 16$ and $\Delta = 32$, and the bucket spacings taken for the stack algorithm are 4, 4, and 8, respectively. (Reproduced from Figs. 1–3 in Ref. 22).

requires accessing external memory a certain number of times, which usually bottlenecks the system performance. Furthermore, the hardware is renowned for its efficient adaptation to a big number of computations. These hardware implementation features apparently favor the no-stack Fano algorithm, even when the number of its computations required is larger than the stack algorithm. In fact, a hard-decision version of the Fano algorithm has been hardware-implemented, and can operate at a data rate of 5 Mbps (megabits per second) [30]. The prototype employs a systematic convolutional code to compensate for the input erasures so that whenever the prespecified decoding time expires, the remaining undecoded binary demodulator outputs that directly correspond to the input sequences are immediately outputted. Other features of this prototype include the following:

- The Fano metric is fixedly scaled to either $(1, -11)$ or $(1, -9)$, rather than adaptable to the channel noise level for convenient hardware implementation.
- The length (or depth) of the backward movement of the Fano algorithm is limited by a technique called *backsearch limiting*, in which the decoder is not allowed to move backward more than some maximum number J levels back from its furthest penetration into the tree. Whenever the backward

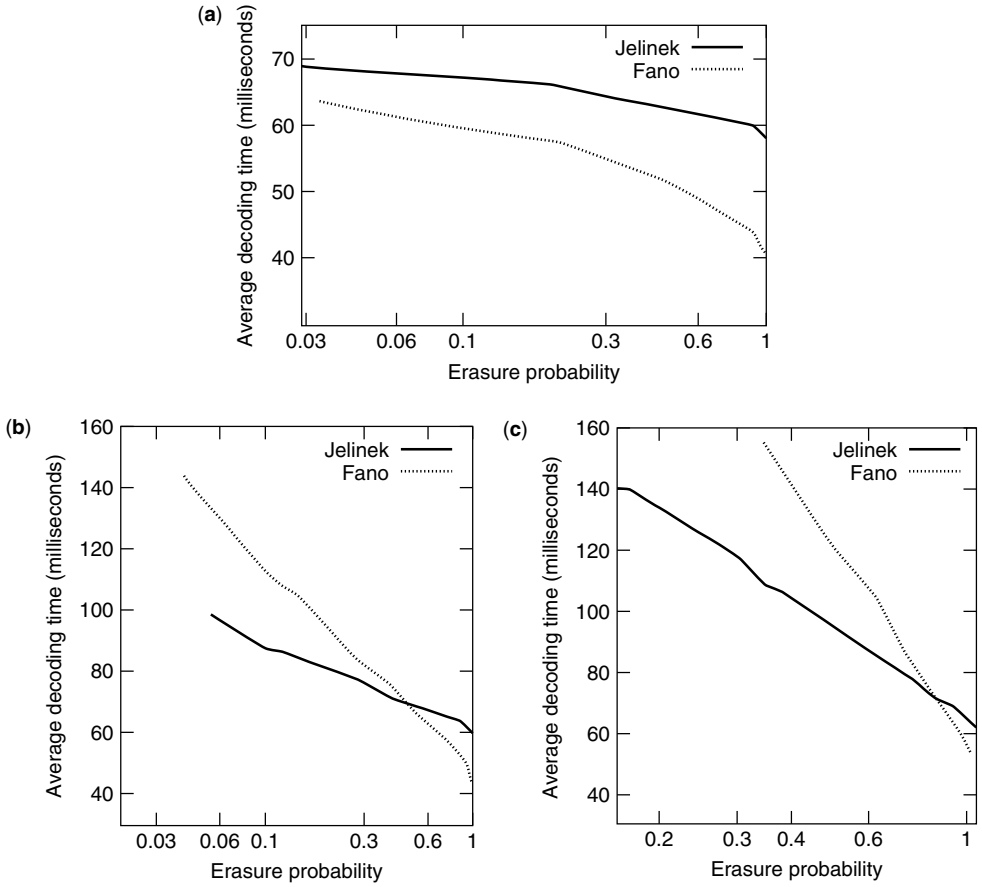


Figure 9. Comparisons of erasure probabilities of the Fano algorithm and the stack algorithm with stacked-bucket modification; all simulation parameters are taken to be the same as those in Fig. 8: (**a**) $p = .033$; (**b**) $p = .045$; (**c**) $p = .057$. (Reproduced from Figs. 5–7 in Ref. 22).

limit is reached, the decoder is forced forward by lowering the threshold until it falls below the metric of the best successor path.

- When the hardware decoder maintains the received bits that correspond to the backsearch J branches for use of forward and backward moves, a separate input buffer must, at the same time, actively receive the upcoming received bits. Once an input buffer overflow is encountered, the decoder must be resynchronized by directly jumping to the most recently received branches in the input buffer, and those information bits corresponding to J levels back are forcefully decoded. The forcefully decoded outputs are exactly the undecoded binary demodulator outputs that directly correspond to the respective input sequences. Again, this design explains why the prototype must use a systematic code.

Figure 10 shows the resultant bit error performances for this hardware decoder for BSCs [30]. An anticipated observation from Fig. 10 is that a larger input buffer, which can be interpreted as a larger decoding time limit, gives a better performance.

A soft-decision version of the hardware Fano algorithm was used for space and military applications in the late 1960s [31,32]. The decoder built by the Jet Propulsion Laboratory [32] operated at a data rate of 1 Mbps, and successfully decoded the telemetry data from the Pioneer Nine spacecraft. Another soft-decision variable-rate hardware implementation of the Fano algorithm was reported in Ref. 33, wherein decoding was accelerated to 1.2 Mbps.

A modified Fano algorithm, named *creeper algorithm*, was proposed in 1999 [34]. This algorithm is indeed a compromise between the stack algorithm and the Fano algorithm. Instead of placing all visited paths in the stack, it selectively stores a fixed number of the paths that are more likely to be part of the optimal code path. As anticipated, the next path to be visited is no longer restricted to the immediate predecessor path and the successor paths but is extended to these selected likely paths. The number of likely paths is usually set less than 2^k times the code tree depth. The simulations given in Ref. 34 indicated that in computational complexity, the creeper algorithm considerably improves the Fano algorithm, and can be made only slightly inferior to the stack algorithm.

8. TRELLIS-BASED SEQUENTIAL DECODING ALGORITHM

Sequential decoding was mostly operated over a code tree in early publications, although some early published

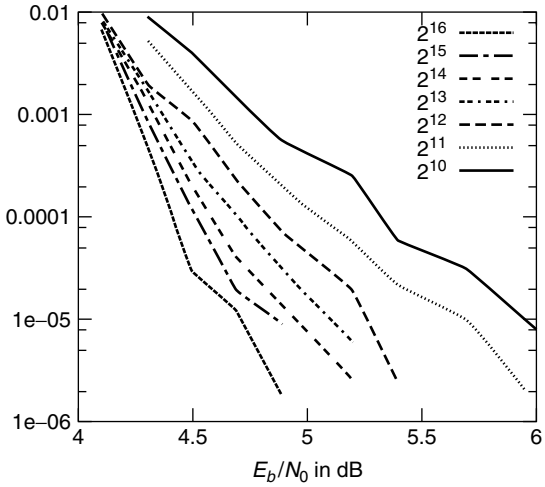


Figure 10. Bit error rate of the hardware Fano algorithm based on systematic $(2, 1, 46)$ convolutional code with generator polynomials $g_1 = 4,000,000,000,000,000$ and $g_2 = 7,154,737,013,174,652$. The legend indicates that the input buffer size tested ranges from $2^{10} = 1024$ to $2^{16} = 65,536$ branches. Experiments are conducted with 1 Mbps data rate over the binary symmetric channel with crossover probability $p = \frac{1}{2}\text{erfc}(\sqrt{E_b/N_0})$, where $\text{erfc}(x) = (2/\sqrt{\pi}) \int_x^\infty \exp\{-x^2\}\, dx$ is the complementary error function. The step size, the backsearch limit, and the Fano metric are set to $\Delta = 6$, $J = 240$, and $(1, -11)$, respectively. (Reproduced from Fig. 18 in Ref. 30).

work already hinted at the possibility of conducting sequential decoding over a trellis [35,36]. The first feasible algorithm that sequentially searches a trellis for the optimal codeword is the *generalized stack algorithm* [23]. The generalized stack algorithm simultaneously extends the top M paths in the stack. It then determines, according to the trellis structure, whether any of the extended paths merge with a path that is already in the stack. If so, the algorithm deletes the newly generated path after ensuring that its path metric is smaller than the cumulative path metric of the merged path up to the merged node. No redirection on the merged path is performed, even if the path metric of the newly generated path exceeds the path metric of the subpath that traverses along the merged path, and ends at the merged node. Thus, the newly generated path and the merged path may coexist in the stack. The generalized stack algorithm, although it generally yields a larger average computational complexity than the stack algorithm, has lower variability in computational complexity and a smaller probability of decoding error [23].

The main obstacle in implementing the generalized stack algorithm by hardware is the maintenance of the stack for the simultaneously extended M paths. One feasible solution is to employ M independent stacks, each of which is separately maintained by a processor [37]. In such a multiprocessor architecture, only one path extraction and two path insertions are sufficient for each stack in a decoding cycle of a $(2, 1, m)$ convolutional code [37]. Simulations have shown that this multiprocessor counterpart not only retained the low variability in computational complexity as the original generalized stack algorithm but also had a smaller average decoding time.

When the trellis-based generalized stack algorithm simultaneously extends 2^K most likely paths in the stack (i.e., $M = 2^K$), where $K = \sum_{j=1}^{k} K_j$ and K_j is the length of the jth shift register in the convolutional code encoder, the algorithm becomes the maximum-likelihood Viterbi decoding algorithm. The optimal codeword is thereby sought by exhausting all possibilities, and no computational complexity gain can be obtained at a lower noise level. Han et al. [38] proposed a true noise-level-adaptable trellis-based maximum-likelihood sequential decoder, called *maximum-likelihood soft-decision decoding algorithm* (MLSDA). The MLSDA adopts a new metric, other than the Fano metric, to guide its sequential search over a trellis for the optimal code path, which is now the code path with the minimum path metric. Derived from a variation of the Wagner rule [39], the new path metric associated with a path $\mathbf{v}_{(\ell n-1)}$ is given by

$$M_{\text{ML}}(\mathbf{v}_{(\ell n-1)} \mid \mathbf{r}_{(\ell n-1)}) = \sum_{j=0}^{\ell n-1} M_{\text{ML}}(v_j \mid r_j) \tag{9}$$

where $M_{\text{ML}}(v_j \mid r_j) = (y_j \oplus v_j) \times |\phi_j|$ is the jth-bit metric, $\mathbf{r}$ is the received vector, $\phi_j = \ln[\Pr(r_j \mid 0)/\Pr(r_j \mid 1)]$ is the jth loglikelihood ratio, and

$$y_j = \begin{cases} 1, & \text{if } \phi_j < 0 \\ 0, & \text{otherwise} \end{cases}$$

is the hard-decision output due to ϕ_j. For AWGN channels, the ML-bit metric can be simplified to $M_{\text{ML}}(v_j \mid r_j) = (y_j \oplus v_j) \times |r_j|$, where

$$y_j = \begin{cases} 1, & \text{if } r_j < 0 \\ 0, & \text{otherwise} \end{cases}$$

As described previously, the generalized stack algorithm, while examining the path merging according to a trellis structure, does not redirect the merged paths. The MLSDA, however, genuinely redirects and merges any two paths that share a common node, resulting in a stack without coexistence of crossed paths. A remarkable feature of the new ML path metric is that when a newly extended path merges with an existing path of longer length, the ML path metric of the newly extended path is always greater than or equal to the cumulative ML metric of the existing path up to the merged node. Therefore, a newly generated path that is shorter than its merged path can be immediately deleted, reducing the redirection overhead of the MLSDA only to the case in which the newly generated path and the merged existing path are equally long.[11] Thus, they merged at their end node. In such case, the

[11] Notably, for the new ML path metric, the path that survives is always the one with smaller path metric, contrary to the sequential decoding algorithm in terms of the Fano metric, in which the path with larger Fano metric survives.

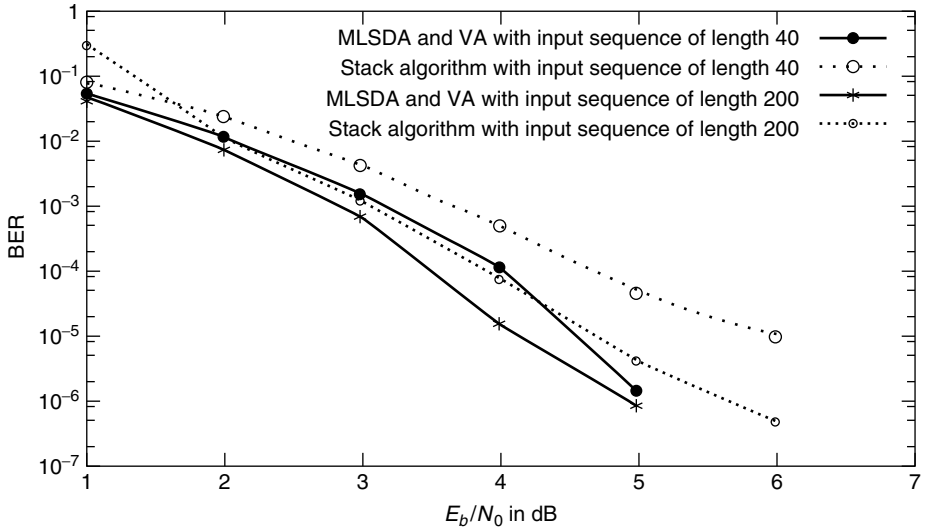


Figure 11. Bit error rates (BERs) of the MLSDA, the Viterbi algorithm (VA), and the stack algorithm for binary $(2, 1, 6)$ convolutional code with generators $g_1 = 634$ (octal), $g_2 = 564$ (octal), and input sequences of lengths 40 and 200. *Source*: Fig. 1 in Ref. 38).

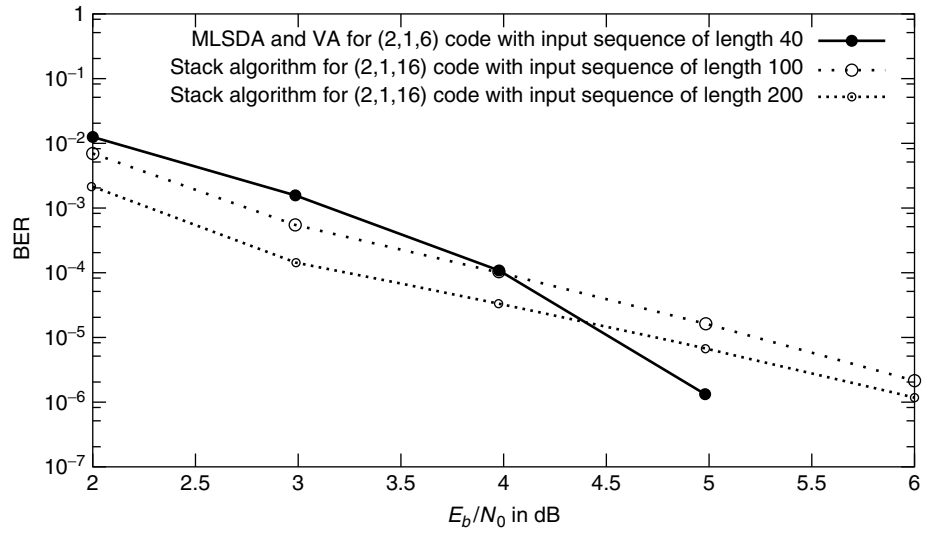


Figure 12. Bit error rates (BERs) of the MLSDA and Viterbi algorithm for binary $(2, 1, 6)$ convolutional code with generators $g_1 = 634$ (octal), $g_2 = 564$ (octal), and an input sequence of length 40. Also, BERs of the stack algorithm for binary $(2, 1, 16)$ convolutional code with generators $g_1 = 1632044, g_2 = 1145734$, and input sequences of lengths 100 and 200. (*Source*: Fig. 2 in Ref. 38).

redirection is merely a deletion of the path with larger path metric.

Figures 11 and 12 show the performances of the MLSDA for $(2, 1, 6)$ and $(2, 1, 16)$ convolutional codes transmitted over the AWGN channel. Specifically, Fig. 11 compares the bit error rate (BER) of the MLSDA with those obtained by the Viterbi and the stack algorithms. Both the MLSDA and the Viterbi algorithm yield the same BER since they are both maximum-likelihood decoders. Figure 11 also shows that the MLSDA provides around 1.5 dB advantage over the stack algorithm at BER $= 10^{-5}$, when both algorithms employ the same input sequence of length 40. Even when the length of the input sequence of the stack algorithm is extended to 200, the MLSDA with input sequence of length 40 still offers an advantage of $\sim$0.5 dB at BER $= 10^{-6}$. Figure 12 collects the empirical results concerning the MLSDA with an input sequence of length 40 for $(2, 1, 6)$ code, and the stack algorithm with input sequences of lengths 100 and 200 for $(2, 1, 16)$ code. The three curves indicate that the MLSDA with an input sequence of smaller length 40 for $(2, 1, 6)$ code provides an advantage of 1.0 dB over the stack algorithm with much longer input sequences and larger constraint length at BER $= 10^{-6}$.

These simulations lead to the conclusion that the stack algorithm normally requires a sufficiently long input sequence to converge to a low BER, which necessarily results in a long decoding delay and high demand for stack space. By adopting a new sequential decoding metric, the MLSDA can achieve the same performance using a much shorter input sequence; hence, the decoding delay and the demand for stack space can be significantly reduced. Furthermore, unlike the Fano metric, the new ML metric adopted in the MLSDA does not depend on the knowledge of the channel, such as SNR, for codes transmitted over the AWGN channel. Consequently, the MLSDA and the Viterbi algorithm share a common nature that their performance is insensitive to the accuracy of the channel SNR estimate for AWGN channels.

9. PERFORMANCE CHARACTERISTICS OF SEQUENTIAL DECODING

An important feature of sequential decoding is that the decoding time varies with the received vector, because the number of paths examined during the decoding process differs for different received vectors. The received vector, in turn, varies according to the noise statistics. The decoding complexity can therefore be legitimately viewed as a random variable whose probability distribution is defined by the statistics of the received vector.

The statistics of sequential decoding complexity have been extensively investigated using the *random coding technique* [15,36,40–44]. Instead of analyzing the complexity distribution with respect to a specific deterministic code, the average complexity distribution for a random code was analyzed. In practical applications, the convolutional codes are always deterministic in their generator polynomials. Nevertheless, taking the aspect of a random convolutional code, in which the coefficients of the generator polynomials are random, facilitates the analysis of sequential decoding complexity. The resultant average decoding complexity (where the decoding complexity is directly averaged over all possible generator polynomials) can nonetheless serve as a quantitative guide to the decoding complexity of a practical deterministic code.

In analyzing the average decoding complexity, a correct code path that corresponds to the transmitted codeword over the code tree or trellis always exists, even if the convolutional encoder is now random. Extra computation time is introduced whenever the search process of the decoder deviates from the correct code path due to a noise-distorted received vector. The incorrect paths that deviate from the correct code path can be classified according to the first node at which the incorrect and the correct code paths depart from each other. Denote by S_j the subtree that contains all incorrect paths that branch from the jth node on the correct path, where $0 \le j \le L-1$ and L is the length of the code input sequences. Then, an upper probability bound[12] on the average computational complexity C_j defined as the number of branch metric computations due to the examination of those incorrect paths in S_j can be established as

$$\Pr\{C_j \ge \mathcal{N}\} \le A\mathcal{N}^{-\rho} \tag{10}$$

for some $0 < \rho < \infty$ and any $0 \le j \le L-1$, where A is a constant that varies for different sequential decoding algorithms. The bound is independent of j because, during its derivation, L is taken to infinity such that all incorrect subtrees become identical in principle. The distribution characterized by the right-hand side of (10) is a *Pareto distribution*, and ρ is therefore named the *Pareto exponent*.

[12] The bound in (10) was first established by Savage for random *tree codes* for some integer value of ρ [41], where random tree codes constitute a super set of convolutional codes. Later, Jelinek [43] extended its validity for random tree codes to real-valued $\rho \ge 1$, satisfying (11). The validity of (10) for random convolutional codes was substantiated by Falconer [42] for $0 < \rho < 1$, and by Hashimoto and Arimoto [44] for $\rho \ge 1$.

Experimental studies indicate that the constant A usually lies between 1 and 10 [1]. The Pareto exponent ρ is uniquely determined by the code rate R using the formula

$$R = \frac{E_0(\rho)}{\rho} \tag{11}$$

for $0 < R < C$, where $E_0(\rho)$ is the *Gallager function* [45 Eq. (5.6.14)] and C is the channel capacity (cf. footnote 7). For example, the Gallager function and the channel capacity for the binary symmetric channel with crossover probability p are respectively given by

$$E_0(\rho) = \rho - (1+\rho)\log_2[p^{1/(1+\rho)} + (1-p)^{1/(1+\rho)}] \tag{12}$$

and

$$C = 1 + p\log_2(p) + (1-p)\log_2(1-p)$$

Equations (11) and (12) together imply that ρ goes to infinity as $R \downarrow 0$, and ρ approaches zero when $R \uparrow C$. Figure 13 gives the Pareto exponents for code rates $R = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}$.

An argument converse to (10), due to Jacobs and Berlekamp [15], states that no sequential decoding algorithm can achieve a computational distribution better than the Pareto distribution of (10), given that no decoding error results for convolutional codes. Specifically, they showed that for a Pareto exponent that satisfies (11)

$$\Pr\{C_j \ge \mathcal{N} \mid \text{correct decoding}\} > [1 - o(\mathcal{N})]\mathcal{N}^{-\rho} \tag{13}$$

where $o(\cdot)$ is the little-o function, satisfying $o(x) \to 0$ as $x \to \infty$. The two bounds in (10) and (13) coincide only when $\mathcal{N}$ is sufficiently large.

On the basis of the multiple branching process technique [46], closed form expressions of the average computational complexity of sequential decoding were derived [47–49]. However, these closed form expressions were suited only for small $\mathcal{N}$. Inequalities (10) and (13)

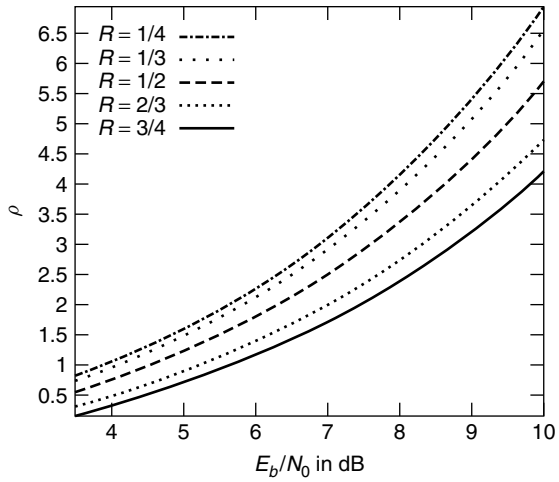


Figure 13. Pareto exponent as a function of E_b/N_0 for a BSC with crossover probability $p = \frac{1}{2}\text{erfc}(\sqrt{E_b/N_0})$, where $\text{erfc}(x) = (2/\sqrt{\pi})\int_x^\infty \exp\{-x^2\}\,dx$ is the complementary error function.

also show that the two bounds are independent of the code constraint length. This observation confirms the claim made in Section 1 that the average computational effort for sequential decoding is in principle independent of the code constraint length.

Inequalities (10) and (13) and the observation that $C_j \geq 1$ jointly yield

$$E[C_j] = \int_1^\infty \Pr\{C_j \geq \mathcal{N}\}\, d\mathcal{N} \leq \int_1^\infty A\mathcal{N}^{-\rho}\, d\mathcal{N}$$

and

$$\begin{aligned} &E[C_j \mid \text{correct decoding}] \\ &\quad = \int_1^\infty \Pr\{C_j \geq \mathcal{N} \mid \text{correct decoding}\}\, d\mathcal{N} \\ &\quad \geq \int_1^\infty [1 - o(\mathcal{N})]\mathcal{N}^{-\rho}\, d\mathcal{N} \end{aligned}$$

Therefore, if the Pareto exponent ρ is greater than unity, $E[C_j]$ is bounded from above. Conversely, if $E[C_j \mid \text{correct decoding}] < \infty$, then $\rho > 1$. Since the probability of correct decoding is very close to unity in most cases of interest, $\rho > 1$ is widely accepted as a sufficient and necessary condition for $E[C_j]$ to be bounded. This result gives rise to the term *computational cutoff rate* $R_0 = E_0(1)$, for sequential decoding.

From (11), $\rho > 1$ if, and only if, $R < R_0 = E_0(1)$, meaning that the cutoff rate R_0 is the largest code rate under which the average complexity of sequential decoding is finite. Thus, sequential decoding becomes computationally implausible once the code rate exceeds the cutoff rate R_0. This theoretical conclusion can be similarly observed from simulations.

Can the computational cutoff rate be improved? The question was answered by a proposal made by Flaconer [42] concerning the use of a hybrid coding scheme. The proposed communication system consists of an $(n_{\text{out}}, k_{\text{out}})$ outer Reed–Solomon encoder, n_{out} parallel $(n_{\text{in}}, 1, m)$ inner convolutional encoders, n_{out} parallel noisy channels, n_{out} sequential decoders for inner convolutional codes, and an algebraic decoder for the outer Reed–Solomon code. These five modules work together in the following fashion. The outer Reed–Solomon encoder encodes k_{out} input symbols into n_{out} output symbols, each of which is b bits long with the last m bits equal to zero. Then, each of n_{out} output symbol is fed into its respective binary $(n_{\text{in}}, 1, m)$ convolutional encoder in parallel, and induces $n_{\text{in}} \times b$ output code bits. Thereafter, these n_{out} $(n_{\text{in}}b)$-bit streams are simultaneously transmitted over n_{out} independent channels, where the n_{out} independent channels may be created by time-multiplexing over a single channel. On receiving n_{out} noise-distorted received vectors of dimension $n_{\text{in}}b$, the n_{out} sequential decoders reproduce the n_{out} output symbols through a sequential codeword search. If any sequential codeword search is not finished before a prespecified time, its output will be treated as an *erasure*. The final step is to use an algebraic Reed–Solomon decoder to regenerate the k_{out} input symbols, based on the n_{out} output symbols obtained from the n_{out} parallel sequential decoders.

The effective code rate of this hybrid system is

$$R_{\text{effective}} = \frac{k_{\text{out}}(b-m)}{n_{\text{out}}n_{\text{in}}b}$$

The largest effective code rate under which the hybrid system is computationally practical has been proved to improve over $E_0(1)$ [42]. Further improvement along the line of code concatenation can be found in Refs. 50 and 51.

The basis of the performance analysis for the aforementioned sequential decoding is random coding, and has nothing to do with any specific properties of the applied code, except the code rate R. Zigangirov [34] proposed to analyze the statistics of C_j for deterministic convolutional codes with an infinite memory order (i.e., $m = \infty$) in terms of recursive equations, and determined that for codes transmitted over BSCs and decoded by the tree-based stack algorithm

$$E[C_j] \leq \frac{\rho}{\rho - 1} 2^{-(-n\alpha+\beta)/(1+\rho)-k} \tag{14}$$

for $R < R_0 = E_0(1)$, where ρ is the Pareto exponent that satisfies (11) and α and β are as defined in the sentence following Eq. (4). Zigangirov's result again suggested that $E[C_j]$ is bounded for $R < R_0$, even when the deterministic convolutional codes are considered. A similar result was also established for the Fano algorithm [34].

Another code-specific estimate of sequential decoding complexity was due to Chevillat and Costello [52,53]. From simulations, they ingeniously deduced that the computational complexity of a sequential decoder is indeed related to the column distance function (CDF) of the applied code. They then established that for a convolutional code transmitted over a BSC with crossover probability p

$$\Pr\{C_j \geq \mathcal{N}\} < A N_d \exp\{-\lambda_1 d_c(\ell) + \lambda_2 \ell\} \tag{15}$$

for $R < 1 + 2p\log_2(p) + (1-2p)\log_2(1-p)$, where A, λ_1, and λ_2 are factors determined by p and code rate R; N_d is the number of length-$[n(\ell+1)]$ paths with Hamming weight equal to $d_c(\ell)$; ℓ is the integer part of $\log_{2^k}\mathcal{N}$; and $d_c(r)$ is the CDF of the applied code. They concluded that a convolutional code with a rapidly increasing CDF can yield a markedly smaller sequential decoding complexity. The outstanding issue is thus how to construct similar convolutional codes for use in sequential decoding. The issue is further explored in Section 11.

Next, the upper bounds on the bit error rate of sequential decoding are introduced. Let P_{S_j} be the probability that a path belonging to the incorrect subtree S_j is decoded as the ultimate output. Then, Chevillat and Costello [53] show that for a specific convolutional code transmitted over a BSC with crossover probability p

$$P_{S_j} < B N_f \exp\{-\gamma d_{\text{free}}\} \tag{16}$$

where B and γ are factors determined by p and code rate R, N_f is the number of code paths in S_j with Hamming weight equal to d_{free}, and d_{free} is the free distance of the applied code. The parameter γ is positive for all convolutional

codes whose free distance exceeds a lower limit determined by p and R. This result indicates that a small error probability for sequential decoding can be obtained by selecting a convolutional code with a large free distance and a small number of codewords with Hamming weight d_{free}. The free distance of a convolutional code generally grows with its constraint length. The bit error rate can therefore be made desirably low when a convolutional code with a sufficiently large constraint length is employed, as the computational complexity of sequential decoding is independent of the code constraint length. However, when a code with a large constraint length is used, the length of the input sequences must also be extended such that the effective code rate $kL/[n(L+m)]$ is closely approximated by code rate $R = k/n$. More discussion of the bit error rates of sequential decoding can be found in the literature [36,40,54].

10. BUFFER OVERFLOW AND SYSTEM CONSIDERATIONS

As already demonstrated by the hardware implementation of the Fano algorithm in Section 7, the *input buffer* at the decoder end for the temporary storage of the received vector is finite. The online decoder must therefore catch up to the input rate of the received vector such that the storage space for obsolete components of the received vector can be freed to store upcoming received components. Whenever an input buffer overflow is encountered, some of the still-in-use content in the input buffer must be forcefully written over by the new input, and the decoder must resynchronize to the new contents of the input buffer; hence input erasure occurs. The overall codeword error P_s of a sequential decoder thus becomes

$$P_s \simeq P_e + P_{\text{erasure}}$$

where P_e is the *undetected word error* under the infinite input buffer assumption and P_{erasure} is the *erasure probability*. For a code with a long constraint length and a practically sized input buffer, P_e is markedly smaller than P_{erasure}, so the overall word error is dominated by the probability of input buffer overflow. In this case, effort is reasonably focused on reducing P_{erasure}. When the code constraint length is only moderately large, a tradeoff between P_e and P_{erasure} must be made. For example, reducing the bucket spacing for the stacked bucket-enabled stack algorithm or lowering the step size for the Fano algorithm results in a smaller P_e, but increases $E[C_j]$ and hence P_{erasure}. The choice of path metrics, as indicated in Section 5, also yields a similar tradeoff between P_e and P_{erasure}. Accordingly, a balance between these two error probabilities must be maintained in practical system design.

The probability of buffer overflow can be analyzed as follows. Let B be the size of the input buffer measured in units of branches; hence the input buffer can store nB bits for an (n, k, m) convolutional code. Also let $1/T$ (bits per second) be the input rate of the received vector. Suppose that the decoder can perform μ branch metric computations in $n \times T$ seconds. Then if over μB branch computations are performed for paths in the jth incorrect subtree, the jth received branch in the input buffer must be written over by the new received branch. To simplify the analysis, assume that the entire buffer is simply reset when a buffer overflow occurs. In other words, the decoder aborts the present codeword search, and immediately begins a new search according to the new received vector. From Eq. (10), the probability of performing more than μB branch computations for paths in S_j is upper-bounded by $A(\mu B)^{-\rho}$. Hence, the erasure probability [41,55] for input sequences of length L is upper-bounded by

$$P_{\text{erasure}} \le L A (\mu B)^{-\rho} \tag{17}$$

Taking $L = 1000, A = 5, \mu = 10, B = 10^5$, and $R = \frac{1}{2}$ yields $\rho = 1.00457$ and $P_{\text{erasure}} \le 4.694 \times 10^{-3}$.

Three actions can be taken to avoid eliminating the entire received vector when the input buffer overflow occurs: (1) just inform the outer mechanism that an input erasure occurs, and let the outer mechanism take care of the decoding of the respective input sequences; (2) estimate the respective input sequences by a function mapping from the received vector to the input sequence; and (3) output the tentative decoding results obtained thus far. The previous section already demonstrated an example of the first action using the hybrid coding system. The second action can be taken whenever an input sequence to a convolutional encoder can be recovered from its codewords through a function mapping. Two typical convolutional codes whose input sequence can be directly mapped from the codewords are the systematic code and the quick-lookin code (to be introduced in the next section). A decoder can also choose to output a tentative decoded input sequence if the third action is taken. A specific example for the third action, named the *multiple stack algorithm*, is introduced in the next paragraph.

In 1977, Chevillat and Costello [56] proposed a *multiple stack algorithm* (MSA), which eliminated entirely the possibility of erasure in the sense that the decoded output is always based on the codeword search. The MSA, as its name implies, acts exactly like the stack algorithm except that it accommodates multiple stacks. During decoding, the MSA first behaves as the stack algorithm by using only its main stack of size z_{main}. When the main stack reaches its limit, the best t paths in the main stack are transferred to a smaller second stack with size $z \ll z_{\text{main}}$. Then the decoding process proceeds just like the stack algorithm, but now using the second stack. If a path reaches the end of the search tree before the second stack is filled, then the path is stored as a tentative decision, and the second stack is eliminated. The MSA then returns to the main stack that has t vacancy spaces for new paths because the top t paths have been removed. If another path reaches the end of the search tree before the main stack is filled again, the decoder compares its path metric with that of the current tentative decision, and outputs the one with larger metric, and stops. Now in case the second stack is filled before a code path is located, a third stack with the same size z is created such that the top t paths in the second stack are transferred to it. The codeword search process then proceeds over the

third stack until either a tentative decision can be made or a new stack needs to be created. Additional stacks of size z are formed whenever necessary. The decoder always compares the newly located code path with the previous tentative decision, and retains the better one. With some properly selected system parameters including z_{main}, z, t, and input buffer size, the MSA guarantees that whenever an input erasure occurs, a tentative decision is always ready for output [56]. Simulation results show that even though the stack maintenance of the MSA is more complex than the stack algorithm, the bit error rate of the former is much lower than that of the latter (see. Fig. 14). Further improvements of the MSA can be found in Refs. 57 and 58.

11. CODE CONSTRUCTION FOR SEQUENTIAL DECODING

A rapid column distance growth in CDF has been conjectured to help the early rejection of incorrect paths for sequential decoding algorithms [59]; this conjecture was later substantiated by Chevillat and Costello both empirically [52] and analytically [53]. In an effort to construct a good convolutional code for sequential decoding, Johannesson proposed [60] that the *code distance profile*, defined as $\{d_c(1), d_c(2), \ldots, d_c(m+1)\}$, can be used, instead of the entire code distance function $d_c(\cdot)$, as a "criterion" for good code construction. His suggestion greatly reduced the number of possible code designs that must be investigated.

A code $\mathcal{C}$ is said to have a better distance profile than another code $\mathcal{C}'$ with the same code rate and memory order, if there exists ℓ with $1 \le \ell \le m+1$ such that $d_c(j) = d'_c(j)$ for $1 \le j \le \ell - 1$ and $d_c(\ell) > d'_c(\ell)$, where $d_c(\cdot)$ and $d'_c(\cdot)$ are the CDFs of codes $\mathcal{C}$ and $\mathcal{C}'$, respectively. In other words, a code with a better distance profile exhibits a faster initial column distance growth in its CDF. A code is said to have an *optimal distance profile*, and is called an ODP code, if its distance profile is superior to that of any other code with the same code rate and memory order.

The searching of ODP codes was extensively studied by Johannesson and Passke [60–63]. They found that the ODP condition could be imposed on a half-rate ($R = \frac{1}{2}$) short constraint length code without penalty in the code free distance [60]; That is, the half-rate ODP code with a short constraint length can also be the code with the largest free distance of all codes with the same code rate and memory order. Tables 1 and 2 list the half-rate ODP codes for systematic codes and nonsystematic codes, respectively. Comprehensive information on ODP codes can be found in Ref. 34. These tables notably reveal that the free distance of a systematic ODP code is always inferior to that of a nonsystematic ODP code with the same memory order.

Employing ODP codes, while notably reduces the number of metric computations for sequential decoding, does not ensure erasure-free performance in practical implementation. If an erasure-free sequential decoding algorithm such as the MSA cannot be adopted due to certain practical considerations, the decoder must still force an immediate decision by just taking a *quick look* at the received vector, once input erasure occurs.

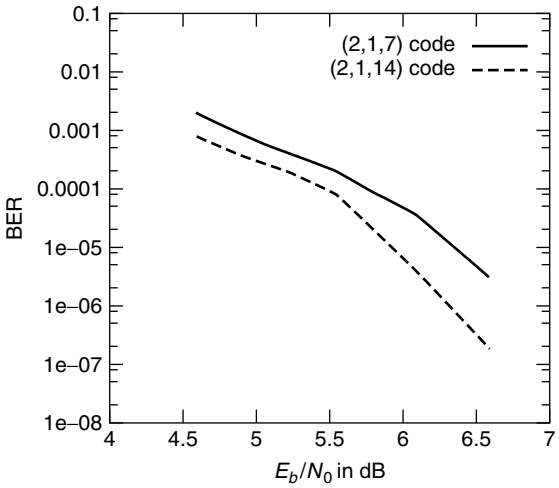


Figure 14. The MSA bit error rates (BER) for $(2, 1, 7)$ and $(2, 1, 14)$ convolutional codes with an input sequence of length 64. The system parameters for $(2, 1, 7)$ and $(2, 1, 14)$ convolutional codes are $(z_{\text{main}}, z, t) = (1024, 11, 3)$ and $(z_{\text{main}}, z, t) = (2900, 11, 3)$, respectively. The input erasure is emulated by an upper limit on branch metric computations C_{lim}, which is 1700 and 3300 for $(2, 1, 7)$ and $(2, 1, 14)$ convolutional codes, respectively. (Reproduced in part from Fig. 3 in Ref. 57).

Table 1. List of Code Rate $R = \frac{1}{2}$ Systematic Codes with Optimal Distance Profile

m	g_2	d_{free}
1	6	3
2	7	4
3	64	4
4	66	5
5	73	6
6	674	6
7	714	6
8	671	7
9	7,154	8
10	7,152	8
11	7,153	9
12	67,114	9
13	67,116	10
14	71,447	10
15	671,174	10
16	671,166	12
17	671,166	12
18	6,711,454	12
19	7,144,616	12
20	7,144,761	12
21	71,447,614	12
22	71,446,166	14
23	67,115,143	14
24	714,461,654	15
25	671,145,536	15
26	714,476,053	16
27	7,144,760,524	16
28	7,144,616,566	16
29	7,144,760,535	18
30	67,114,543,064	16
31	67,114,543,066	18

Source: Ref. 34.

Table 2. List of Code Rate $R = \frac{1}{2}$ Nonsystematic Codes with Optimal Distance Profile

m	g_1	g_2	d_{free}
1	6	4	3
2	7	5	5
3	74	54	6
4	62	56	7
5	77	45	8
6	634	564	10
7	626	572	10
8	751	557	12
9	7,664	5,714	12
10	7,512	5,562	14
11	6,643	5,175	14
12	63,374	47,244	15
13	45,332	77,136	16
14	65,231	43,677	17
15	727,144	424,374	18
16	717,066	522,702	19
17	745,705	546,153	20
18	6,302,164	5,634,554	21
19	5,122,642	7,315,626	22
20	7,375,407	4,313,045	22
21	67,520,654	50,371,444	24
22	64,553,062	42,533,736	24
23	55,076,157	75,501,351	26
24	744,537,344	472,606,614	26
25	665,041,116	516,260,772	27

Source: Ref. 34.

This seems to suggest that a systematic ODP code is preferred, even if it has a smaller free distance than its nonsystematic ODP counterpart. In such case, the deficiency on the free distance of the systematic ODP codes can be compensated for by selecting a larger memory order m. However, when a convolutional code with large m is used, the length of the input sequences must be proportionally extended; otherwise the effective code rate cannot be well approximated by the convolutional code rate, and the performance to some extent degrades. This effect motivates the attempt to construct a class of nonsystematic ODP codes with the "quick look" property and a free distance superior to that of systematic codes.

Such a class of nonsystematic codes has been developed by Massey and Costello, called the *quick-lookin* (QLI) convolutional codes [59]. The generator polynomials of these half-rate QLI convolutional codes differ only in the second coefficient. Specifically, their generator polynomials satisfy $g_1(x) = g_2(x) + x$, where addition of coefficients is based on modulo-2 operation, allowing the decoder to recover the input sequence $\mathbf{u}(x)$ by summing the two output sequences $\mathbf{v}_1(x)$ and $\mathbf{v}_2(x)$ as

$$x \cdot \mathbf{u}(x) = \mathbf{v}_1(x) + \mathbf{v}_2(x) \tag{18}$$

If p is the individual bit error probability in the codeword $\mathbf{v}$, then the bit error probability due to recovering information sequence $\mathbf{u}$ from $\mathbf{v}$ through (18) is shown to be approximately $2p$ [59]. Table 3 lists the QLI ODP convolutional codes [34].

Table 3. List of Code Rate $R = \frac{1}{2}$ QLI Codes with Optimal Distance Profile

m	g_1	d_{free}
2	7	5
3	74	6
4	76	6
5	75	8
6	714	8
7	742	9
8	743	9
9	7,434	10
10	7,422	11
11	7,435	12
12	74,044	11
13	74,046	13
14	74,047	14
15	740,464	14
16	740,462	15
17	740,463	16
18	7,404,634	16
19	7,404,242	15
20	7,404,155	18
21	74,041,544	18
22	74,042,436	19
23	74,041,567	19
24	740,415,664	20
25	740,424,366	20
26	740,424,175	22
27	7,404,155,634	22
28	7,404,241,726	23
29	7,404,154,035	24
30	74,041,567,514	23
31	74,041,567,512	25

Source: Ref. 34.

12. CONCLUSIONS

Although sequential decoding has a longer history than maximum-likelihood decoding based on the Viterbi algorithm, its practical applications are not as popular, because the highly repetitive "pipeline" nature of the Viterbi decoder makes it very suitable for hardware implementation. Furthermore, a sequential decoder usually requires a longer decoding delay (defined as the time between the receipt of a received branch and the output of its respective decoding decision) than a Viterbi decoder. Generally, the decoding delay of a sequential decoder for an (n, k, m) convolutional code is around $n \times B$, where B is the number of received branches that an input buffer can accommodate. Yet, the decoding delay of a Viterbi decoder can be made a small multiple, often ranging from 5 to 10, of $n \times m$. On the other hand, Refs. 64 and 65 showed that sequential decoding is highly sensitive to the channel parameters such as an inaccurate estimate of channel SNR and an incomplete compensation of phase noise. The Viterbi algorithm, however, proved to be robust for imperfect channel identification, again securing the superiority of the Viterbi decoder in practical applications.

Nevertheless, there are certain situations that the sequential decoding fits well, especially in decoding convolutional codes having a large constraint length.

In addition, the sequential decoder can send a timely retransmission request by detecting the occurrence of an input buffer overflow [66]. Sequential decoding has attracted some attention in the field of mobile communications [67] in which a demand of low bit error rate is required. Such applications are beyond the scope of this article, and interested readers can refer to the literature [1,68,69].

Acknowledgments

Professor Marc P. C. Fossorier of the University of Hawaii and Professor John G. Proakis are appreciated for their careful reviews and valuable comments. Mr. Tsung-Ju Wu and Mr. Tsung-Chi Lin are also appreciated for preparing the figures and checking the examples in this manuscript.

BIOGRAPHIES

Yunghsiang S. Han was born in Taipei, Taiwan, on April 24, 1962. He received the B.S. and M.S. degrees in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 1984 and 1986, respectively, and the Ph.D. degree from the school of Computer and Information Science, Syracuse University, Syracuse, New York, in 1993. From 1986 to 1988 he was a Lecture at Ming-Hsin Engineering College, Hsinchu, Taiwan. He was a Teaching Assistant from 1989 to 1992 and from 1992 to 1993 a Research Assistant in the School of Computer and Information Science, Syracuse University. He is a recipient of the 1994 Syracuse University Doctoral Prize. From 1993 to 1997 he was an Associate Professor in the Department of Electronic Engineering at Hua Fan College of Humanities and Technology, Taipei Hsien, Taiwan. He is now with the Department of Computer Science and Information Engineering at National Chi Nan University, Nantou, Taiwan. He was promoted to full Professor in 1998. His research interests are in error-control coding and fault-tolerant computing.

Po-Ning Chen received the B.S. and M.S. degrees in electrical engineering from the National Tsing-Hua University in Taiwan in 1985 and 1987, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland at College Park (USA) in 1994. From 1985 to 1987, he was with Image Processing Laboratory in the National Tsing-Hua University, where he worked on the recognition of Chinese characters. During 1989, he was with StarTech Engineering Inc., where he focused on the development of fingerprint recognition systems. After receiving the Ph.D. degree in 1994, he joined Wan Ta Technology Inc. as a Vice General Manager, conducting several projects on Point-of-Sale systems. In 1995, he joined the research staff at the Advanced Technology Center, Computer and Communication Laboratory, Industrial Technology Research Institute in Taiwan, where he led a project on Java-based Network Managements. Since 1996, he has been an Associate Professor in the Department of Communications Engineering at the National Chiao-Tung University, Taiwan, and became a full Professor in 2001. Dr. Chen received the 2000 Young Scholar Paper Award from Academia Sinica, Taiwan. His areas of interests are information and coding theory, large deviation theory, and distributed detection.

BIBLIOGRAPHY

1. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
2. J. M. Wozencraft, Sequential decoding for reliable communications, *IRE Nat. Conv. Rec.* **5**(Pt. 2): 11–25 (1957).
3. J. M. Wozencraft and B. Reiffen, *Sequential Decoding*, MIT Press, Cambridge, MA, 1961.
4. R. M. Fano, A heuristic discussion of probabilistic decoding, *IEEE Trans. Inform. Theory* **IT-9**(2): 64–73 (April 1963).
5. K. Sh. Zigangirov, Some sequential decoding procedures, *Probl. Peredachi Inform.* **2**: 13–25 (1966).
6. F. Jelinek, A fast sequential decoding algorithm using a stack, *IBM J. Res. Dev.* **13**: 675–685 (Nov. 1969).
7. N. J. Nilsson, *Principle of Artificial Intelligence*, Tioga, Palo Alto, CA, 1980.
8. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
9. A. J. Viterbi, Error bound for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**(2): 260–269 (April 1967).
10. J. L. Massey, *Threshold Decoding*, MIT Press, Cambridge, MA, 1963.
11. G. D. Forney, Jr., Review of random tree codes, in *Appendix A. Study of Coding Systems Design for Advanced Solar Missions*, NASA Contract NAS2-3637, Codex Corp., Dec. 1967.
12. J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, Addison-Wesley, Reading, MA, 1984.
13. Y. S. Han, C. R. P. Hartmann, and C.-C. Chen, Efficient priority-first search maximum-likelihood soft-decision decoding of linear block codes, *IEEE Trans. Inform. Theory* **39**(5): 1514–1523 (Sept. 1993).
14. Y. S. Han, A new treatment of priority-first search maximum-likelihood soft-decision decoding of linear block codes, *IEEE Trans. Inform. Theory* **44**(7): 3091–3096 (Nov. 1998).
15. I. M. Jacobs and E. R. Berlekamp, A lower bound to the distribution of computation for sequential decoding, *IEEE Trans. Inform. Theory* **IT-13**(2): 167–174 (April 1967).
16. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
17. J. L. Massey, Variable-length codes and the fano metric, *IEEE Trans. Inform. Theory* **IT-18**(1): 196–198 (Jan. 1972).
18. E. A. Bucher, Sequential decoding of systematic and nonsystematic convolutional codes with arbitrary decoder bias, *IEEE Trans. Inform. Theory* **IT-16**(5): 611–624 (Sept. 1970).
19. F. Jelinek, Upper bound on sequential decoding performance parameters, *IEEE Trans. Inform. Theory* **IT-20**(2): 227–239 (March 1974).
20. Y. S. Han, P.-N. Chen, and M. P. C. Fossorier, A generalization of the fano metric and its effect on sequential decoding using a stack, *IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, 2002.
21. K. Sh. Zigangirov, *private communication*, Feb. 2002.

22. J. M. Geist, An empirical comparison of two sequential decoding algorithms, *IEEE Trans. Commun. Technol.* **COM-19**(4): 415–419 (Aug. 1971).

23. D. Haccoun and M. J. Ferguson, Generalized stack algorithms for decoding convolutional codes, *IEEE Trans. Inform. Theory* **IT-21**(6): 638–651 (Nov. 1975).

24. J. B. Anderson and S. Mohan, Sequential coding algorithms: A survey and cost analysis, *IEEE Trans. Commun.* **COM-32**(2): 169–176 (Feb. 1984).

25. S. Mohan and J. B. Anderson, Computationally optimal metric-first code tree search algorithms, *IEEE Trans. Commun.* **COM-32**(6): 710–717 (June 1984).

26. D. E. Knuth, *The Art of Computer Programming*, Vol. III: Sorting and Searching, Addison-Wesley, Reading, MA, 1973.

27. P. Lavoie, D. Haccoun, and Y. Savaria, A systolic architecture for fast stack sequential decoders, *IEEE Trans. Commun.* **42**(5): 324–335 (May 1994).

28. S. Kallel and K. Li, Bidirectional sequential decoding, *IEEE Trans. Inform. Theory* **43**(4): 1319–1326 (July 1997).

29. J. M. Geist, Search properties of some sequential decoding algorithms, *IEEE Trans. Inform. Theory* **IT-19**(4): 519–526 (July 1973).

30. G. D. Forney, Jr. and E. K. Bower, A high-speed sequential decoder: prototype design and test, *IEEE Trans. Commun. Technol.* **COM-19**(5): 821–835 (Oct. 1971).

31. I. M. Jacobs, Sequential decoding for efficient communication from deep space, *IEEE Trans. Commun. Technol.* **COM-15**(4): 492–501 (August 1967).

32. J. W. Layland and W. A. Lushbaugh, A flexible high-speed sequential decoder for deep space channels, *IEEE Trans. Commun. Technol.* **COM-19**(5): 813–820 (Oct. 1978).

33. M. Shimada, T. Todoroki, and K. Nakamura, Development of variable-rate sequential decoder LSI, *IEEE Int. Conf. Communications*, 1989, pp. 1241–1245.

34. R. Johannesson and K. Sh. Zigangirov, *Fundamentals of Convolutional Coding*, IEEE Press, Piscataway, NJ, 1999.

35. J. Geist, *Algorithmic Aspects of Sequential Decoding*, Ph.D. thesis, Dept. Electrical Engineering, Univ. Notre Dame, Notre Dame, IN, 1970.

36. G. D. Forney, Jr., Convolutional codes III: Sequential decoding, *Inform. Control* **25**: 267–269 (July 1974).

37. N. Bélanger, D. Haccoun, and Y. Savaria, A multiprocessor architecture for multiple path stack sequential decoders, *IEEE Trans. Commun.* **42**(2–4): 951–957 (Feb.–April 1994).

38. Y. S. Han, P.-N. Chen, and H.-B. Wu, A maximum-likelihood soft-decision sequential decoding algorithm for binary convolutional codes, *IEEE Trans. Commun.* **50**(2): 173–178 (Feb. 2002).

39. J. Snyders and Y. Be'ery, Maximum likelihood soft decoding of binary block codes and decoders for the golay codes, *IEEE Trans. Inform. Theory* 963–975 (Sept. 1989).

40. H. L. Yudkin, *Channel State Testing in Information Decoding*, Ph.D. thesis, MIT, Cambridge, Mass, 1964.

41. J. E. Savage, Sequential decoding-the computation problem, *Bell Syst. Tech. J.* **45**: 149–175 (Jan. 1966).

42. D. D. Falconer, A hybrid decoding scheme for discrete memoryless channels, *Bell Syst. Tech. J.* **48**: 691–728 (March 1969).

43. F. Jelinek, An upper bound on moments of sequential decoding effort, *IEEE Trans. Inform. Theory* **IT-15**(1): 140–149 (Jan. 1969).

44. T. Hashimoto and S. Arimoto, Computational moments for sequential decoding of convolutional codes, *IEEE Trans. Inform. Theory* **IT-25**(5): 584–591 (1979).

45. R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

46. W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley, New York, 1970.

47. R. Johannesson, On the distribution of computation for sequential decoding using the stack algorithm, *IEEE Trans. Inform. Theory* **IT-25**(3): 323–332 (May 1979).

48. D. Haccoun, A branching process analysis of the average number of computations of the stack algorithm, *IEEE Trans. Inform. Theory* **IT-30**(3): 497–508 (May 1984).

49. R. Johannesson and K. Sh. Zigangirov, On the distribution of the number of computations in any finite number of subtrees for the stack algorithm, *IEEE Trans. Inform. Theory* **IT-31**(1): 100–102 (Jan. 1985).

50. F. Jelinek and J. Cocke, Bootstrap hybrid decoding for symmetrical binary input channel, *Inform. Control* **18**: 261–298 (April 1971).

51. O. R. Jensen and E. Paaske, Forced sequence sequential decoding: A concatenated coding system with iterated sequential inner decoding, *IEEE Trans. Commun.* **46**(10): 1280–1291 (Oct. 1998).

52. P. R. Chevillat and D. J. Costello, Jr., Distance and computation in sequential decoding, *IEEE Trans. Commun.* **COM-24**(4): 440–447 (April 1978).

53. P. R. Chevillat and D. J. Costello, Jr., An analysis of sequential decoding for specific time-invariant convolutional codes, *IEEE Trans. Inform. Theory* **IT-24**(4): 443–451 (July 1978).

54. A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill, New York, 1979.

55. J. E. Savage, The distribution of the sequential decoding computation time, *IEEE Trans. Inform. Theory* **IT-12**(2): 143–147 (April 1966).

56. P. R. Chevillat and D. J. Costello, Jr., A multiple stack algorithm for erasurefree decoding of convolutional codes, *IEEE Trans. Commun.* **COM-25**(12): 1460–1470 (Dec. 1977).

57. H. H. Ma, The multiple stack algorithm implemented on a zilog z-80 microcomputer, *IEEE Trans. Commun.* **COM-28**(11): 1876–1882 (Nov. 1980).

58. K. Li and S. Kallel, A bidirectional multiple stack algorithm, *IEEE Trans. Commun.* **47**(1): 6–9 (Jan. 1999).

59. J. L. Massey and D. J. Costello, Jr. Nonsystematic convolutional codes for sequential decoding in space applications, *IEEE Trans. Commun. Technol.* **COM-19**(5): 806–813 (Oct. 1971).

60. R. Johannesson, Robustly optimal rate one-half binary convolutional codes, *IEEE Trans. Inform. Theory* **IT-21**(4): 464–468 (July 1975).

61. R. Johannesson, Some long rate one-half binary convolutional codes with an optimal distance profile, *IEEE Trans. Inform. Theory* **IT-22**(5): 629–631 (Sept. 1976).

62. R. Johannesson, Some rate 1/3 and 1/4 binary convolutional codes with an optimal distance profile, *IEEE Trans. Inform. Theory* **IT-23**(2): 281–283 (March 1977).

63. R. Johannesson and E. Paaske, Further results on binary convolutional codes with an optimal distance profile, *IEEE Trans. Inform. Theory* **IT-24**(2): 264–268 (March 1978).

64. J. A. Heller and I. W. Jacobs, Viterbi decoding for satellite and space communication, *IEEE Trans. Commun. Technol.* **COM-19**(5): 835–848 (Oct. 1971).

65. I. M. Jacobs, Practical applications of coding, *IEEE Trans. Inform. Theory* **IT-20**(3): 305–310 (May 1974).

66. A. Drukarev and Jr. D. J. Costello, Hybrid ARQ error control using sequential decoding, *IEEE Trans. Inform. Theory* **IT-29**(4): 521–535 (July 1983).

67. P. Orten and A. Svensson, Sequential decoding in future mobile communications, *Proc. PIMRC '97*, 1997 Vol. 3, pp. 1186–1190.

68. S. Kallel, Sequential decoding with an efficient incremental redundancy ARQ strategy, *IEEE Trans. Commun.* **40**(10): 1588–1593 (Oct. 1992).

69. P. Orten, Sequential decoding of tailbiting convolutional codes for hybrid ARQ on wireless channels, *Proc. IEEE Vehicular Technology Conf.*, 1999, Vol. 1, 279–284.

SERIALLY CONCATENATED CODES AND ITERATIVE ALGORITHMS

S. Benedetto
G. Montorsi
Politecnico di Torino
Torino (Turin), Italy

1. INTRODUCTION

In his goal to find a class of codes whose probability of error decreased exponentially at rates less than capacity, while decoding complexity increased only algebraically, Dave Forney [1] arrived at a solution consisting of the coding structure known as *concatenated code*. It consists of the cascade of an *inner* code and an *outer* code, which, in Forney's approach, would be a relatively short inner code (typically, a convolutional code) admitting simple maximum-likelihood soft decoding, and a long high-rate algebraic outer code (for most applications a nonbinary Reed–Solomon code) equipped with a powerful algebraic error correction decoding algorithm, possibly using reliability information from the inner decoder.

Initially motivated only by theoretical research interests, concatenated codes have since then evolved as a standard for those applications where very high coding gains are needed, such as (deep-)space applications, digital television broadcasting, compact disk players, and many others. Alternative solutions for concatenation have also been studied, such as using a trellis-coded modulation scheme as inner code [2], or concatenating two convolutional codes [3]. In the latter case, the inner Viterbi decoder employs a soft-output decoding algorithm to provide soft-input decisions to the outer Viterbi decoder. An interleaver was also proposed between the two encoders to separate bursts of errors produced by the inner decoder.

We find then, in a "classical" concatenated coding scheme, the main ingredients that formed the basis for the invention of "Turbo codes" [4], namely two, or more, *constituent* codes (CCs) and an *interleaver*. The novelty of Turbo codes, however, consists of the way they use the interleaver, which is embedded into the code structure to form an overall concatenated code with very large block length, and in the proposal of a parallel concatenation to achieve a higher rate for given rates of CCs. The latter advantage is obtained using systematic CCs and not transmitting the information bits entering the second encoder. In the following, we will refer to turbo codes as *parallel concatenated codes* (PCCs). The so-obtained codes have been shown to yield very high coding gains at bit error probabilities in the range 10^{-5}–10^{-7}; in particular, low bit error probabilities can be obtained at rates well beyond the channel cutoff rate, which had been regarded for long time as the "practical" capacity. Quite remarkably, this performance can be achieved by a relatively simple iterative decoding technique whose computational complexity is comparable to that needed to decode the two CCs.

After the invention of PCCs, other forms of code concatenations with interleaver have been proposed, like the serial concatenation [5] and hybrid concatenations [6,7]. This article deals with serial concatenations with interleaver in a wider sense; that is, we apply the iterative algorithm introduced in 1998 [5] for decoding serially concatenated codes (SCCs) also to systems where the two concatenated blocks represent different functions, such as intersymbol interference channel, a modulator, or a multiuser combiner.

In the first part we consider the serial concatenation of interleaved codes or *serially concatenated codes* (SCCs), called SCBC or SCCC according to the nature of CCs, that can be block (SCBC) or convolutional codes (SCCC). For this class of codes, we give analytical upper bounds to the performance of a maximum-likelihood (ML) decoder, present design guidelines leading to the optimal choice of CCs that maximize the so-called *interleaver gain* and the asymptotic code performance, and present the iterative decoding algorithm yielding results close to capacity limits with limited decoding complexity.

In the second part of the paper we extend the concept of serial concatenation, beyond the case of two codes, to systems in which two functional blocks with memory are cascaded and can be separated by an interleaver. This extension will prove that the "Turbo" principle applied to serial concatenation can have a wide variety of applications in the field of digital transmission.

Throughout the paper, a semi-tutorial approach has been adopted, and all but strictly necessary algebra and mathematical subtleties avoided. In this, we tried to stick to Einstein motto: *Everything should be made as simple as possible, but not simpler*, so as to highlight the meaning of the main properties/results while addressing the interested reader to the appropriate references.

2. SERIAL CONCATENATIONS

Figure 1 is a block diagram of a fairly general serial concatenation of three modules, M1, M2, and M3, with two interleavers I1 and I2 separating each pair of modules.

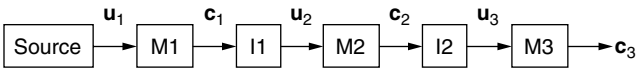


Figure 1. A general serially concatenated structure with three modules and two interleavers.

The symbols **u** and **c** may denote semiinfinite sequences, when the transmitted information flow does not possess a frame structure, or vectors containing a finite number of symbols when the information flow is delivered in a frame-by-frame basis. In the framed case, the interleavers must be *block* interleavers [8]. As we will see in Section 6, a structure like the one depicted in Fig. 1 can represent numerous systems and applications, beyond the basic one in which all modules represent encoders. In the next two sections, however, we will show how to analyze and design a serial concatenation in the particular, yet important, case of two encoders separated by one interleaver. This restriction will simplify the presentation, yet it will lead to conclusions that are applicable to a broader set of systems.

3. ANALYSIS OF SERIALLY CONCATENATED CODES WITH INTERLEAVERS

Consider the serial concatenation shown in Fig. 2. It is formed by the *outer* encoder C_o with rate $R_c^o = k/p$, and the *inner* encoder C_i with rate $R_c^i = p/n$, joined by an interleaver of size N bits, generating a serially concatenated code (SCC) C_S with rate $R_c^S = R_c^o \times R_c^i = k/n$. N will be assumed to be an integer multiple[1] of p. The outer and inner encoders are the constituent codes (CCs) of the SCC. We consider here block constituent codes, in which each code word is formed in a memoryless fashion by the encoder. This is the case when both CCs are block encoders, or when they are *terminated* [9, Chap. 11] convolutional encoders.

For large interleaver sizes, an SCC cannot be decoded as a single code using maximum-likelihood (ML) decoding [8]. Instead, as we will see later, the decoder will use a simpler, suboptimum iterative decoding algorithm based on two soft CC decoders. On the other hand, the analysis seems only be possible for ML decoding, using upper bounds that yield a great insight into the code behavior and provide valuable tools for the code design.

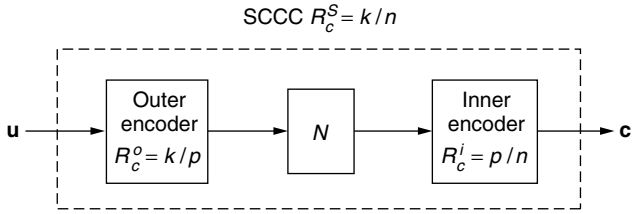


Figure 2. The serial concatenation of two encoders and one interleaver.

[1] Actually, this constraint is not necessary. We can choose in fact inner and outer codes of any rates $R_c^i = k_i/n_i$ and $R_c^o = k_o/n_o$, constraining the interleaver to be an integer multiple of the minimum common multiple of n_o and k_i, that is, $N = K \cdot mcm(n_o, k_i)$. This generalization, though, leads to more complicated expressions and is not considered in the following.

3.1. A Union Bound to the Code Error Probabilities

Upper bounds to the *word*[2] P_w and *bit*[3] P_b error probabilities for an (n, k) linear block code based on the union bound [9] under maximum-likelihood soft decoding for binary PSK (or binary PAM) transmission over an additive white Gaussian noise channel with two-sided noise power spectral density $N_0/2$ are given by

$$P_w \le \frac{1}{2} \sum_{d=d_{\min}}^{n} \sum_{w=1}^{k} A_{w,d} \operatorname{erfc}\left(\sqrt{\frac{dR_cE_b}{N_0}}\right)$$
$$P_b \le \frac{1}{2} \sum_{d=d_{\min}}^{n} \sum_{w=1}^{k} \frac{w}{k} A_{w,d} \operatorname{erfc}\left(\sqrt{\frac{dR_cE_b}{N_0}}\right) \quad (1)$$

where R_c is the code rate, E_b is the energy per *information* bit, and $A_{w,d}$ are the *input–output* coefficients of the encoder, representing the number of codewords with weight d generated by information words of weight w.

Knowledge of the coefficients $A_{w,d}$ is sufficient to evaluate the upper bound to both word and bit error probabilities, and thus we need to evaluate them for the SCC C_S, assuming that we know those of the CCs.

If N in Fig. 2 is low, we can compute the coefficients $A_{w,d}$ by letting each individual information word with weight w be first encoded by the outer encoder C_o and then, after the p bits of the outer codeword have been permuted by the interleaver, be encoded by the inner encoder C_i originating an inner codeword with a certain weight. After repeating this procedure for all the information words with weight w, we should count the inner codewords with weight d, and their number would be the value of $A_{w,d}$.

When N is large the previous operation becomes too complex, and we must resort to a different approach. As thoroughly described elsewhere [5,10], a crucial step in the analysis consists in replacing the actual interleaver that performs a permutation of the N input bits with an abstract interleaver called *uniform interleaver*, defined as a probabilistic device that maps a given input word of weight l into all distinct $\binom{N}{l}$ permutations of it with equal probability $P = 1/\binom{N}{l}$. Use of the uniform interleaver permits the computation of the "average" performance of the SCC, intended as the expectation of the performance of SCCs using the same CCs, taken over the set of all interleavers of a given size. A theorem proved in Ref. 10 guarantees the meaning fullness of the average performance, in the sense that there will always be, for each value of the signal-to-noise ratio, at least one particular interleaver yielding performance better than or equal to that of the uniform interleaver.

[2] The word error probability is defined as the probability that the decoder chooses an incorrect code word, *i.e.*, a code word different from the transmitted one.

[3] The bit error probability is defined as the probability that the decoder delivers an incorrect information bit to the user.

Define as $A_{w,d}^{C_S}$, $A_{w,l}^{C_o}$ and $A_{w,l}^{C_i}$ the input–output coefficients of the SCC C_S, outer and inner encoders, respectively; exploiting the properties of the uniform interleaver, which transforms a codeword of weight l at the output of the outer encoder into all its distinct $\binom{N}{l}$ permutations, we obtain [5]

$$A_{w,d}^{C_s} = \sum_{l=0}^{N} \frac{A_{w,l}^{C_o} \times A_{l,h}^{C_i}}{\binom{N}{l}} \tag{2}$$

This equation (2) permits an easy computation of the input-output coefficients of the SCC, and, together with (1), yields the upper bounds to word and bit error probabilities.

Example 1. Consider the rate-$\frac{1}{3}$ (nominal rate) SCC of Fig. 3 obtained concatenating a terminated rate-$\frac{1}{2}$, 4-state, recursive systematic encoders with generator matrix

$$G_o(Z) = \left[1, \frac{1+Z^2}{1+Z+Z^2}\right] \tag{3}$$

and as inner encoder a terminated 4-state, rate-$\frac{2}{3}$, recursive convolutional encoder with generator matrix

$$G_i(Z) = \begin{bmatrix} 1, 0, \dfrac{1+Z^2}{1+Z+Z^2} \\ 0, 1, \dfrac{1+Z}{1+Z+Z^2} \end{bmatrix} \tag{4}$$

Using the uniform interleaver concept, we have obtained the union bounds to the word and bit error probabilities shown in Figs. 4 and 5. The curves refer to five values of the interleaver size, corresponding to information block sizes $k = 10, 20, 40, 80, 160$. The curves show that increasing the interleaver size yields an increasing coding gain, which is more pronounced for the bit error probability. This phenomenon is known as *interleaving gain*.

Previous analysis can be generalized to n cascaded encoders separated by $n-1$ interleavers. Indeed, the ML performance of the overall encoder improves uniformly when increasing the number of CCs. On the other hand, the suboptimum iterative decoding algorithm that must be used to decode this kind of codes for complexity reasons suffers from a *bootstrap* effect that becomes more and more relevant as we increase the number of CCs. The case of three encoders, which has been proposed and analyzed [11] under the name of double serially concatenated code, seems to be the farthest one can go in practice.

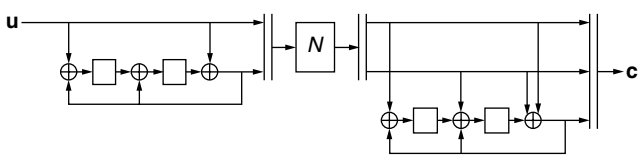


Figure 3. SCCC encoder of Example 1.

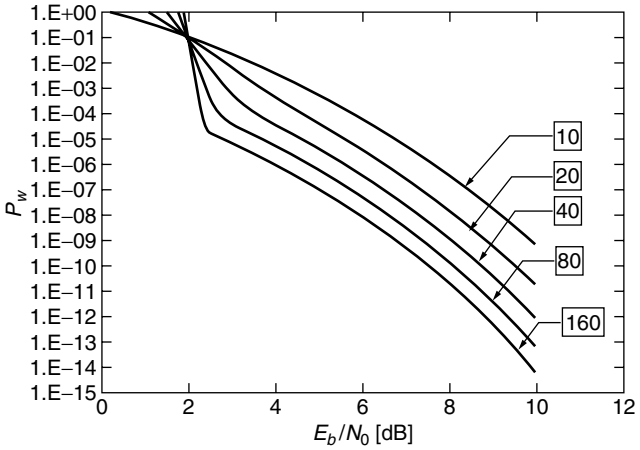


Figure 4. Word error probability bounds for the SCC of Example 1 employing as constituent encoders block codes obtained from 4-state convolutional encoders through trellis termination. The interleaver is uniform and corresponds to information words with size $k = 10, 20, 40, 80, 160$.

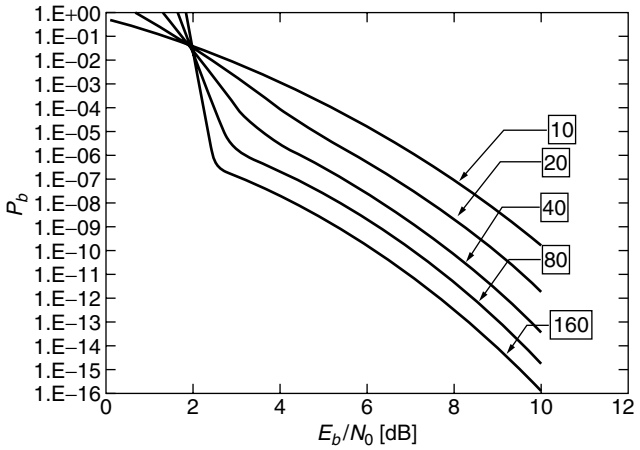


Figure 5. Bit error probability bounds for the SCC of Example 1 employing as constituent encoders block codes obtained from 4-state convolutional encoders through trellis termination. The interleaver is uniform and corresponds to information words with size $K = 10, 20, 40, 80, 160$.

4. DESIGN OF SERIALLY CONCATENATED CODES WITH INTERLEAVER

The error probability performance of concatenated codes with interleaver (and, in particular, of serially concatenated codes) under iterative decoding are invariably represented by curves like the ones depicted in Fig. 6. We can identify three different regions, for increasing values of the signal-to-noise ratio. The first one is the *non-convergence* region, where the error probability keeps high, nearly constant values. At a certain point, the *convergence abscissa*, the curves start a rather steep descent down to medium–low values of the error probability (the *waterfall region*). Finally, in the third region (the *error floor* region), the slope of the curves decreases significantly, and performance improvement are paid with

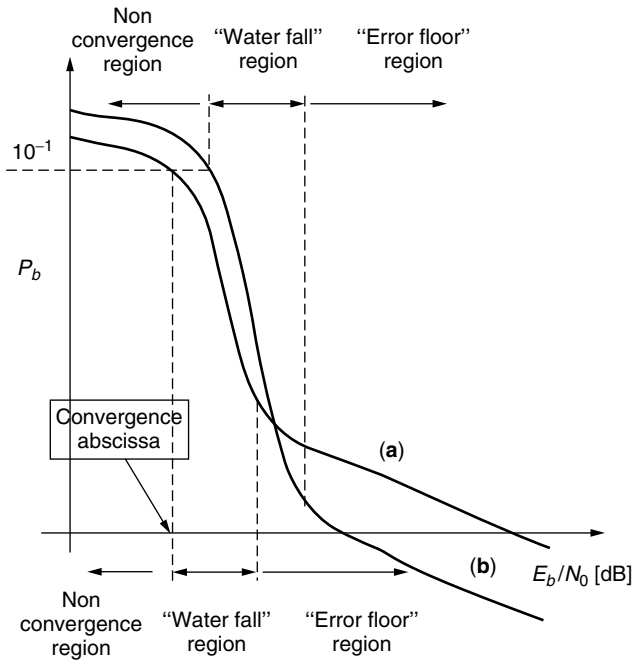


Figure 6. Qualitative behavior of P_b versus E_b/N_0 for concatenated codes with interleavers under iterative decoding.

significant energy expenses. The waterfall region is dominated by the interleaver gain, whereas the error floor region is dictated by the minimum distance of the code. In the first region, the interleaver acts through its size, whereas in the second the kind of interleaver plays a dominant role.

Any attempt to design codes like those in Fig. 2 as a whole leads to discouraging results because of its complexity. So far, only two successful design techniques have been proposed in the literature: the first relies on the previously introduced concept of uniform interleaver, and leads to design rules for the CCs that work nicely for medium–low error probabilities. After designing the CCs, one can improve the code performance by a cut-and-try approach to the interleaver design.

While the first approach is a purely analytical one, the second requires the separate simulation of the behavior of each CC, and is based on the *probability density evolution* technique ([12–14]) or on the *EXIT charts* [15]. This technique leads to codes that exhibit good performance in terms of convergence abscissa, and are suited for medium-high values of the error probability. Unfortunately, the two techniques lead to different code designs; typically, the codes designed (or, better, found) through the second technique show faster convergence of the iterative decoding algorithm, but reach sooner the error floor [see curve (a) in Fig. 6]. The opposite is true for codes designed using the first technique [see curve (b) in Fig. 6]. As a consequence, the choice depends on the quality of service requirements of the system at hand. In the following, we report the main results stemming from the first design technique, addressing the readers to the appropriate references for the second.

4.1. The ML Design

A lengthy analysis [5] shows that, for large interleavers, the union upper bound to the ML error probability of a general concatenated code with uniform interleaver can be written as

$$P \leq \frac{1}{2} \sum_{d=d_{\min}} K_d N^{\alpha(d)} \operatorname{erfc}\left(\sqrt{\frac{dR_c E_b}{N_0}}\right) \tag{5}$$

where K_d does not depend on the interleaver size N. Asymptotically, for large interleavers where $N \to \infty$, the dominant term in the summation of (5) is the one for which the exponent of N is maximum. Denoting that exponent by $\alpha_M \triangleq \max_d \alpha(d)$, and neglecting the other summation terms, yields

$$P \overset{N}{\asymp} \frac{1}{2} K_{d_M} N^{\alpha_M} \operatorname{erfc}\left(\sqrt{\frac{d_M R_c E_b}{N_0}}\right) \tag{6}$$

where $d_M = \arg\max_d [\alpha(d)]$. Negative values of α_M lead to interleaving gains, which are more pronounced for larger magnitudes of α_M. The value of α_M is different in the cases of word and bit error probabilities, and depends on the kind of code concatenation. In the following, we show its values for the bit error probability. Adding one to those values yields the α'_M values pertaining to the word error probability.

For parallel concatenated codes[4] (PCCs) [10] a necessary and sufficient condition to have $\alpha_M < 0$ is that both constituent encoders be *recursive*. In that case, we obtain

$$\begin{aligned} \alpha_M &= -1 \\ d_M &= d_{\text{free,eff}} \end{aligned} \tag{7}$$

where $d_{\text{free,eff}}$ is the *effective* free distance of the PCCC, *i.e.*, the minimum weight of codewords associated to weight 2 information words. The effective free distance of the PCCC is equal to the sum of the effective free distances of the constituent encoders, and thus the optimization of them consists in searching for recursive convolutional encoders with the largest effective free distance, and, possibly, the lowest number of codewords with weight equal to the effective free distance (number of "nearest" neighbors).

For SCCCs, a necessary and sufficient condition to have $\alpha_M < 0$ is that the inner constituent encoder be *recursive* [5]. If that is the case, we obtain

$$\alpha_M = -\left\lfloor \frac{d^o_{\text{free}} + 1}{2} \right\rfloor \tag{8}$$

$$d_M = \begin{cases} \dfrac{d^o_{\text{free}} d^i_{\text{free,eff}}}{2}, & \text{for } d^o_{\text{free}} \text{ odd} \\ \dfrac{(d^o_{\text{free}} - 3) d^i_{\text{free,eff}}}{2} + d_3, & \text{for } d^o_{\text{free}} \text{ even} \end{cases}$$

[4] For comparison purposes, we recall here the main results for the case of parallel concatenation.

where d_3 is the minimum weight of inner code words associated to input words with weight 3.

Example 2. Consider two rate-$\frac{1}{3}$ concatenated encoders. The first is a parallel concatenated convolutional encoder made up with two 4-state systematic recursive convolutional encoders derived from the generator matrix G_0 of (3). The second, instead, is the rate-$\frac{1}{3}$ SCCC of Example 1. The two concatenations use interleaver sizes such that the corresponding information block sizes coincide. Applying the union bound leads to the results reported in Fig. 7, where we plot the bit error probability versus E_b/N_0 for information block sizes equal to 100 and 1000.

The curves make apparent the difference in the interleaver gain. In particular, the parallel concatenation shows an interleaver gain going as $\alpha_M = -1$, whereas the interleaver gain of the SCCC goes as $\alpha_M = -\dfrac{d_{of}+1}{2} = -3$, being the free distance of the outer code equal to 5. This means, for $P_b = 10^{-11}$, a gain of more than 2 dB in favor of the SCCC.

For SCCs, then, the design criteria require the inner code to be recursive, and chosen so as to maximize its effective free distance. The outer encoder can be either recursive or not, and its optimization is based on the usual criterion of maximizing the free distance.

The design criterion based on the *effective* free distance (rather than simply the free distance), and the requirement for the inner constituent encoder to be recursive, is a complete novelty in the panorama of known optimum convolutional encoders, whose search had been based on feedforward encoders with the greatest free distance. The effective free distance of a convolutional encoder can be significantly greater than the free distance, and this, with the interleaver gain, explains the exceptionally good performance of SCCs in the waterfall region. In Refs. 16 and 17 upper bounds to the effective free distance and tables of good recursive convolutional encoders are reported. The findings of the more complete search for good codes matching the design criteria have also been described [18].

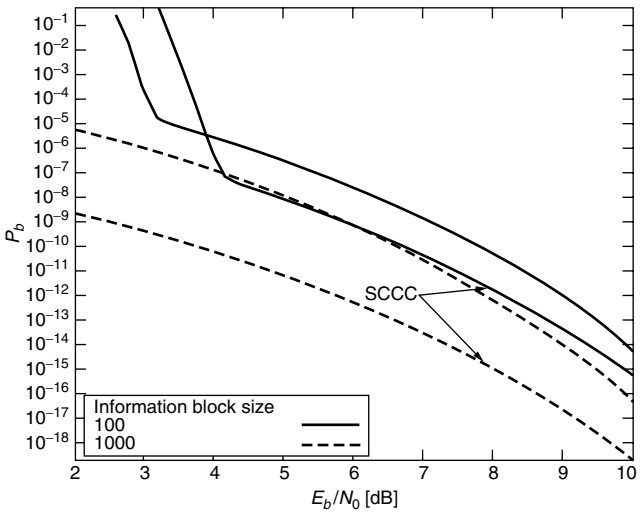


Figure 7. Comparison of serially and parallel rate-$\frac{1}{3}$ concatenated convolutional codes based on 4-state CCs with two interleaver sizes, yielding the same information block size for the two concatenations.

5. ITERATIVE DECODING OF SERIALLY CONCATENATED CODES

In this section, we present the iterative algorithm for decoding serially concatenated codes, with complexity not significantly higher than that needed to separately decode the two CCs. We assume, without loss of generality, that the constituent encoders admit a trellis representation [9].

The core of the decoding procedure consists of a block called *soft-input–soft-output* (SISO). It is a four-port device, which accepts as inputs the likelihood functions (or the corresponding likelihood ratios) of the information and code symbols labeling the edges of the code trellis, and forms as outputs an update of those likelihood functions based on the code constraints. The block SISO is used within the iterative decoding algorithm as shown in Fig. 8, where we also show the block diagram of the encoder to clarify the notations.

We will first explain in words how the algorithm works, according to the blocks of Fig. 8. Subsequently we will give the input-output relationships of the block SISO.

The symbols $\lambda(\cdot;I)$ and $\lambda(\cdot;O)$ at the input and output ports of SISO refer to the logarithmic likelihood ratios (LLRs),[5] unconstrained when the second argument is I, and modified according to the code constraints when it is O. The first argument u refers to the information symbols of the encoder, whereas c refers to code symbols. Finally, the superscript o refers to the outer encoder, and i to the inner encoder. The LLRs are defined as

$$\lambda(x;\cdot) \triangleq \log\left[\frac{P(x;\cdot)}{P(x_{\text{ref}};\cdot)}\right] \tag{9}$$

When x is a binary symbol, "0" or "1," x_{ref} is generally assumed to be the "0." When x belongs to an L-ary alphabet, we can choose as x_{ref} each one of the L symbols.

In contrast to the iterative decoding algorithm employed for PCCC decoding [19], in which only the LLRs

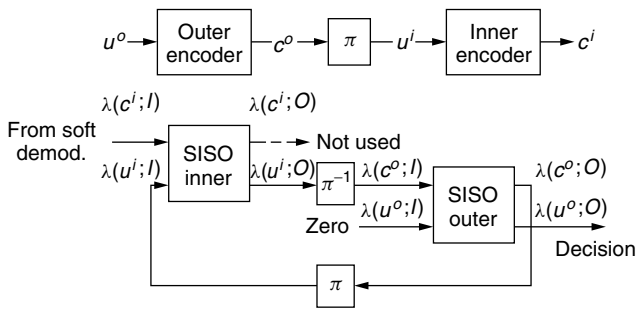


Figure 8. Block diagrams of the encoder and iterative decoder for serially concatenated convolutional codes.

[5] When the symbols are binary, only one LLR is needed; when the symbols belong to an L-ary alphabet, $L-1$ LLRs are required.

of information symbols are updated, here we must update the LLRs of both information and code symbols based on the code constraints.

During the first iteration of the SCC algorithm,[6] the block "SISO inner" is fed with the demodulator soft outputs, consisting of the LLRs of symbols received from the channels, namely, of the code symbols of the inner encoder. The second input $\lambda(u^i;I)$ of the SISO inner is set to zero during the first iteration, since no a- priori information is available on the input symbols u^i of the inner encoder.

The LLRs $\lambda(c^i;I)$ are processed by the SISO algorithm, which computes the *extrinsic* [19] LLRs of the information symbols of the inner encoder $\lambda(u^i;O)$ conditioned on the inner code constraints. The extrinsic LLRs are passed through the inverse interleaver (block labeled "π^{-1}"), whose outputs correspond to the LLRs of the code symbols of the outer code:

$$\lambda(c^o;I) = \pi^{-1}[\lambda(u^i;O)]$$

These LLRs are then fed to the block "SISO outer" in its upper entry, which corresponds to code symbols. The SISO outer, in turn, processes the LLRs $\lambda(c^o;I)$ of its unconstrained code symbols, and computes the LLRs of both code and information symbols based on the code constraints. The input $\lambda(u^o;I)$ of the SISO Outer is always set to zero, which implies assuming equally likely transmitted source information symbols. The output LLRs of information symbols (which yield the a posteriori LLRs of the SCCC information symbols) will be used in the final iteration to recover the information bits. On the other hand, the LLRs of outer code symbols, after interleaving, are fed back to the lower entry (corresponding to information symbols of the inner code) of the block SISO inner to start the second iteration. In fact we have

$$\lambda(u^i;I) = \pi[\lambda(c^o;O)]$$

5.1. Input–Output Relationships for the Block SISO

The block SISO has been described [19]. It represents a slight generalization of the BCJR algorithm [20–22]. Here, we will only recall for completeness its input–output relationships. We will refer, for notations, to the trellis section of the trellis encoder, assumed to be time invariant as we deal with convolutional codes, shown in Fig. 9, where the symbol e denotes the trellis edge, and where we have identified the information and code symbols associated to the edge e as $u(e)$, $c(e)$, and the starting and ending states of the edge e as $s^S(e)$, $s^E(e)$, respectively.

The block SISO works at *symbol* level; thus, for an (n,p) convolutional code, it operates on information symbols u belonging to an alphabet with size 2^p and on code symbols belonging to an alphabet with size 2^n. We will give the general input–output relationships, valid for both outer and inner SISOs, assuming that the information and code symbols are defined over a finite time index set $[1,\ldots,K]$.

[6] To simplify the description, we assume for now that the interleaver acts on symbols instead of bits. In practice, one often uses bit LLRs and bit interleaver, as it will be seen later.

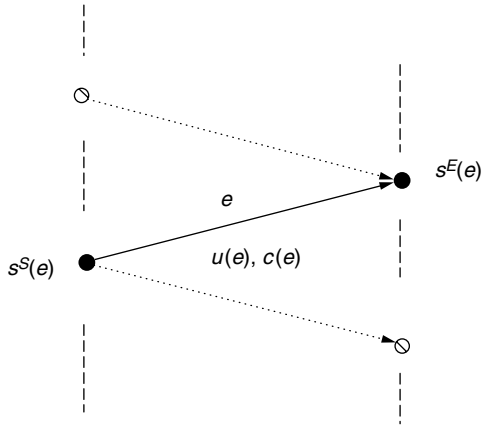


Figure 9. Trellis section defining the notations used for the description of the SISO algorithm.

At time k, $k=1,\ldots,K$, the output extrinsic LLRs are computed as

$$\lambda_k(c;O) = \max_{e:c(e)=c}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[u(e);I] + \beta_k[s^E(e)]\} - \max_{e:c(e)=c_{\text{ref}}}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[u(e);I] + \beta_k[s^E(e)]\} \tag{10}$$

$$\lambda_k(u;O) = \max_{e:u(e)=u}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[c(e);I] + \beta_k[s^E(e)]\} - \max_{e:u(e)=u_{\text{ref}}}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[c(e);I] + \beta_k[s^E(e)]\} \tag{11}$$

The name *extrinsic* given to the LLRs computed according to (10) and (11) derives from the fact that the evaluation of $\lambda_k(c;O)$ [and of $\lambda_k(u;O)$] does not depend on the corresponding simultaneous input $\lambda_k(c;I)$ [and $\lambda_k(u;I)$], so that it can be considered as an update of the input LLR based on informations coming from all homologous symbols in the sequence, except the one corresponding to the same symbol interval.

The quantities $\alpha_k(\cdot)$ and $\beta_k(\cdot)$ in (10) and (11) are obtained through the *forward* and *backward* recursions, respectively, as

$$\alpha_k(s) = \max_{c:s^E(e)=s}{}^*\{\alpha_{k-1}[s^S(e)] + \lambda_k[u(e);I] + \lambda_k[c(e);I]\}, \quad k=1,\ldots,K-1 \tag{12}$$

$$\beta_k(s) = \max_{e:s^S(e)=s}{}^*\{\beta_{k+1}[s^E(e)] + \lambda_{k+1}[u;I] + \lambda_{k+1}[c(e);I]\}, \quad k=K-1,\ldots,1 \tag{13}$$

with initial values

$$\alpha_0(s) = \begin{cases} 0 & s=S_0 \\ -\infty & \text{otherwise} \end{cases}$$

$$\beta_K(S_i) = \begin{cases} 0 & s=S_K \\ -\infty & \text{otherwise} \end{cases}$$

The operator max* performs the following operation:

$$\max_j{}^*(a_j) \triangleq \log\left[\sum_{j=1}^{J} e^{a_j}\right] \tag{14}$$

This operation, a crucial one in affecting the computational complexity of the SISO algorithm, can be performed in practice as

$$\max_j{}^*(a_j) = \max_j(a_j) + \delta(a_1, a_2, \ldots, a_J) \tag{15}$$

where $\delta(a_1, a_2, \ldots, a_J)$ is a correction term that can be computed recursively using a single-entry lookup table [19,23].

The previous description of the iterative decoder assumed that all operations were performed at *symbol* level. Quite often, however, the interleaver operates at *bit* level to be more effective. This is the case, for example, of all results presented in Sections 3–5.

Thus, to perform bit interleaving, we need to transform the symbol extrinsic LLRs obtained at the output of the first SISO into extrinsic bit LLRs, before they enter the deinterleaver. After deinterleaving, the bit LLRs need to be compacted into symbol LLRs before entering the second SISO block, and so on. These operations are performed under the assumption that the LLRs of bits forming a symbol are independent.

Assuming an (n, p) code, and denoting with $\mathbf{u} = [u_1, \ldots, u_p]$ the information symbol formed by p information bits, then the extrinsic LLR λ_i of the ith bit u_i within the symbol $\mathbf{u}$ is obtained as

$$\lambda_{kp+i}(O) = \max_{\mathbf{u}:u_i=1}{}^*[\lambda_k(\mathbf{u}; O) + \lambda_k(\mathbf{u}; I)] - \max_{\mathbf{u}:u_i=0}{}^*[\lambda_k(\mathbf{u}; O) + \lambda_k(\mathbf{u}; I)] - \lambda_{kp+i}(I) \tag{16}$$

Conversely, the extrinsic LLR of the symbol $\mathbf{u}$ is obtained from the extrinsic LLRs of its component bits u_i as

$$\lambda(\mathbf{u}) = \sum_{i=1}^{p} u_i \lambda_i \tag{17}$$

As previous description should have made clear, the SISO algorithm requires that the whole sequence had been received before starting. The reason is due to the backward recursion that starts from the (supposed known) final trellis state. As a consequence, its practical application is limited to the case where the duration of the transmission is short (K small), or, for K long, when the received sequence can be segmented into independent consecutive blocks, like for block codes or convolutional codes with trellis termination [24]. It cannot be used for continuous decoding. This constraint leads to a frame rigidity imposed to the system, and also reduces the overall code rate, because of trellis termination.

A more flexible decoding strategy is offered by modifying the algorithm in such a way that the SISO module operates on a fixed memory span, and outputs the updated LLRs after a given delay D. This algorithm, which we have called the *sliding-window soft-input/soft-output* (SW-SISO) algorithm, is fully described in Ref. 19. In all simulation results the SW-SISO algorithm has been applied.

5.2. Applications of the Decoding Algorithm

We will now use the decoding algorithm to confirm the design rules presented before, and to show the behavior of SCCC in the region of low signal-to-noise ratios (below cutoff rate). Since analytic bounds fail to give significant results in this region, no meaningful quantitative comparisons can be performed between simulated and analytic performance. However, we will show that the hierarchy of the simulation results agrees with the design considerations that had been based on the analysis.

The following aspects will be considered:

- The behavior of the decoding algorithm versus the number of decoding iterations (Section 5.2.1)
- The behavior of the decoding algorithm versus the interleaver length (Subsection 5.2.2)
- The effect of choosing a nonrecursive inner code (Section 5.2.2)
- The SCCC behavior for very low signal-to-noise ratios, to see how close serial concatenation can get to theoretical Shannon bound (Section 5.2.3)
- The comparison between SCCCs and PCCCs (Turbo codes) for the same value of the decoding delay imposed by the two schemes on the input bits (Section 5.2.4).

For all simulated SCCCs, we have used purely random interleavers.

5.2.1. Simulated Coding Gain Versus Number of Iterations. Consider the rate-$\frac{1}{3}$ SCCC1 of Example 1 employing an interleaver of length $N = 2048$. Since the interleaver operates on coded sequences produced by the outer rate-$\frac{1}{2}$ encoder, its length of 2048 bits corresponds to a delay of 1024 information bits. The simulation results are shown in Fig. 10 in terms of bit error probability versus E_b/N_0 for a number of iterations N_I ranging from 1 to 7. The nice convergence of the decoding algorithm is manifest.

5.2.2. The Effect of a Nonrecursive Inner Encoder. In Section 4 we concluded that a nonrecursive inner encoder should yield little interleaver gains. To confirm this theoretical prediction by simulation results, in Fig. 11 we plot the bit error probability versus the input decoding delay obtained by simulating a rate-$\frac{1}{3}$ SCCC that concatenates the outer encoder of Example 1 with the 4-state, nonrecursive inner encoder with generator matrix

$$G_i(Z) = \begin{bmatrix} 1+Z, & Z, & 1 \\ 1+Z, & 1, & 1+Z \end{bmatrix}$$

The curves refer to a signal-to-noise ratio $E_b/N_0 = 1.5$ dB, and to a number of iterations N_I ranging from 1 to 10. It is evident that the bit error probability reaches the floor of

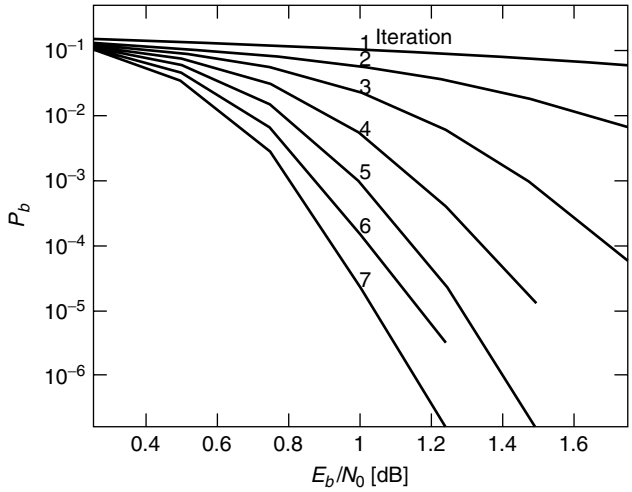


Figure 10. Simulated bit error probability versus the number of iterations for the rate-$\frac{1}{3}$ SCCC of Example 1. The decoding delay in terms of input bits due to the interleaver is 1024.

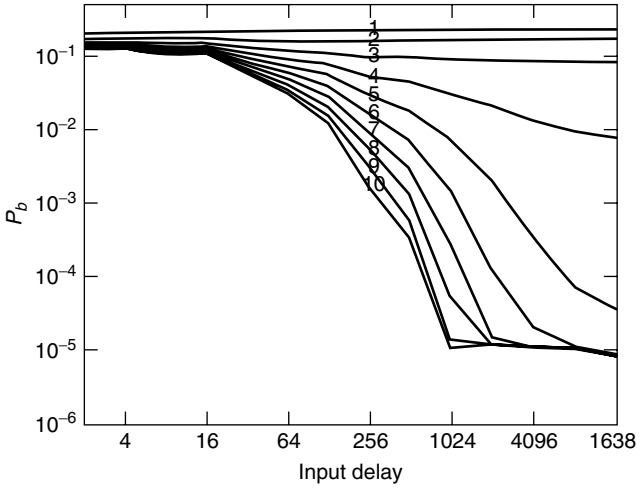


Figure 11. Simulated performance of the SCCC employing a nonrecursive inner encoder described in Section 5.2.3. The bit error probability is plotted versus input decoding delay for different number of iterations. The signal-to-noise ratio is 1.5 dB.

10^{-5} for a decoding delay greater than or equal to 1024, so that no interleaver gain takes place beyond this point. For comparison, in Fig. 12 we show the results obtained for the SCCC concatenating the 4-state, rate-$\frac{1}{2}$ nonrecursive outer encoder with generator matrix

$$G_o(Z) = [1 + Z + Z^2, 1 + Z^2]$$

with the rate-$\frac{2}{3}$ inner encoder of Example 1. The curves refer to a signal-to-noise ratio of 0.75 dB, and show the interleaver gain predicted by the analysis.

5.2.3. Approaching the Theoretical Shannon Limit. We show here the capabilities of SCCCs of yielding results close to the Shannon capacity limit. To this purpose, we have chosen a rate-$\frac{1}{4}$ concatenated scheme with very long interleaver, corresponding to an input decoding delay

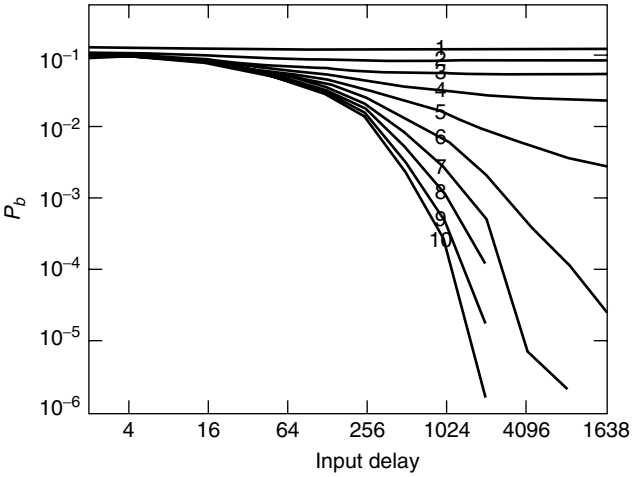


Figure 12. Simulated performance of the SCCC employing a nonrecursive outer encoder described in Section 5.2.2. The bit error probability is plotted versus input decoding delay for different number of iterations. The signal-to-noise ratio is 0.75 dB.

of 16,384. The constituent codes are 8-state codes: the outer encoder is nonrecursive, and the inner encoder is a recursive encoder. Their generating matrices are

$$G_o(Z) = [1 + Z, 1 + Z + Z^3]$$

$$G_i(Z) = \left[1, \frac{1 + Z + Z^3}{1 + Z}\right]$$

respectively. Note the feedback polynomial $(1 + Z)$ in the generator matrix of the inner encoder, which eliminates error events with odd input weights. The results in terms of bit error probability versus signal-to-noise ratio for different number of iterations are presented in Fig. 13 showing that the decoding algorithm works at $E_b/N_0 = -0.05$ dB, at 0.76 dB from the Shannon capacity limit (which is in this case equal to -0.817 dB), with very limited complexity (remember that we are using two rate-$\frac{1}{2}$ codes with 8 states).

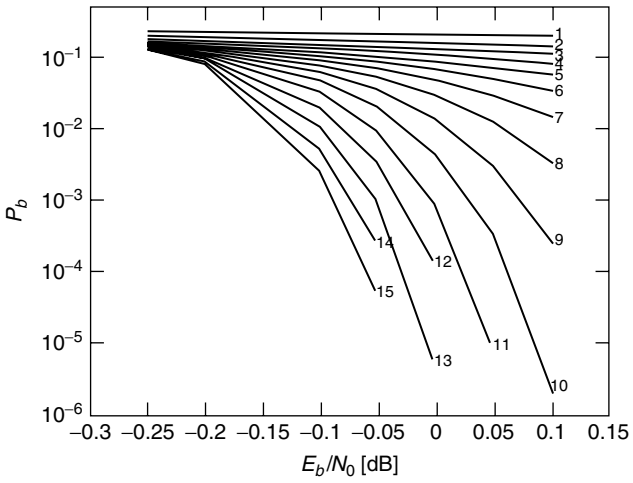


Figure 13. Simulated performance of the rate-$\frac{1}{4}$ SCCC obtained with two 8-state constituent codes and an interleaver yielding an input decoding delay equal to 16,384.

5.2.4. Comparison Between Serially and Parallel Concatenated Codes. Previous analytic results showed that serial concatenation can yield significantly higher interleaver gains and steeper asymptotic slope of the error probability curves. To check whether these advantages are retained when the codes are iteratively decoded at very low signal-to-noise ratios, we have simulated the behavior of two SCCCs and PCCCs in equal system conditions: the concatenated code rate is $\frac{1}{3}$, the CCs (same as Example 2) are 4-state recursive encoders (rates $\frac{1}{2} + \frac{1}{1}$ for PCCCs, and rates $\frac{1}{2} + \frac{2}{3}$ for the SCCCs), and the decoding delays in terms of input bits are 1024 and 16,384, respectively. In Fig. 14 we report the results, in terms of bit error probability versus signal-to-noise ratio, for the case of a decoding delay equal to 1024, after three and seven decoding iterations. As it can be seen from the curves, the PCCC outperforms the SCCC for high values of the bit error probabilities. For bit error probabilities lower than $3 \cdot 10^{-3}$ (for seven iterations), the SCCC outperform PCCC, and does not present the error floor. At 10^{-4}, SCCC has an advantage of 0.7 dB with seven iterations. Finally, in Fig. 15, we report the results for an input decoding delay of 16,384 and six and nine decoding iterations. In this case, the crossover between PCCC and SCCC happens around 10^{-5}. The advantage of SCCC at 10^{-6} is 0.5 dB with nine iterations.

As a conclusion, we can say that the advantages obtained for signal-to-noise ratios above the cutoff rate, where the union bounds can be safely applied, are retained also in the region between channel capacity and cutoff rate. Only when the system quality of service focuses on high values of bit error probability (the threshold depending on the interleaver size) the PCCC are to be preferred. PCCCs, however, present a floor to the bit error probability, which, in the most favorable case seen above, lies around 10^{-6}. This floor is much lower in the case of SCCC.

Finally, it must be recognized that the constituent codes design rules presented in Sec. 4 are based on union bound considerations, and thus yield optimum SCCCs above the cutoff rate. For system applications aiming at very low signal-to-noise ratios, close to the channel capacity (as, for example, in deep-space communications), a general statement is that complex CCs should be avoided, and CCs with low number of states (4–16) should be used. The code design in this signal-to-noise ratio region, a different design approach based on the separate simulation of the behavior of each CC, and based on the probability density evolution technique [12–14] or on the *EXIT charts* [15] must be adopted.

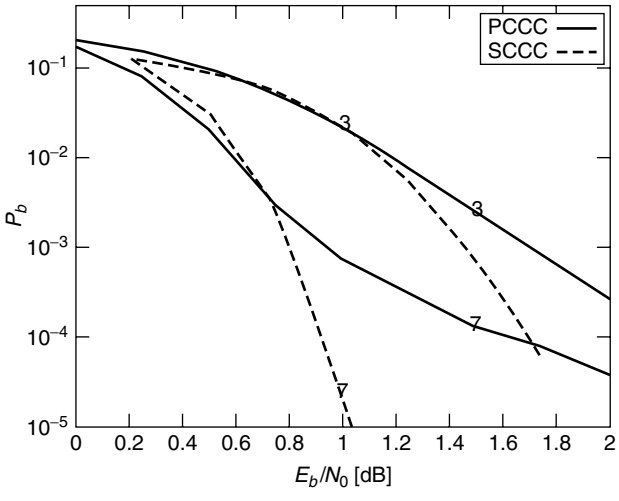


Figure 14. Comparison of the two rate-$\frac{1}{3}$ PCCCs and SCCCs of Example 2. The curves refer to three and seven iterations of the decoding algorithm and to an equal input decoding delay of 1024.

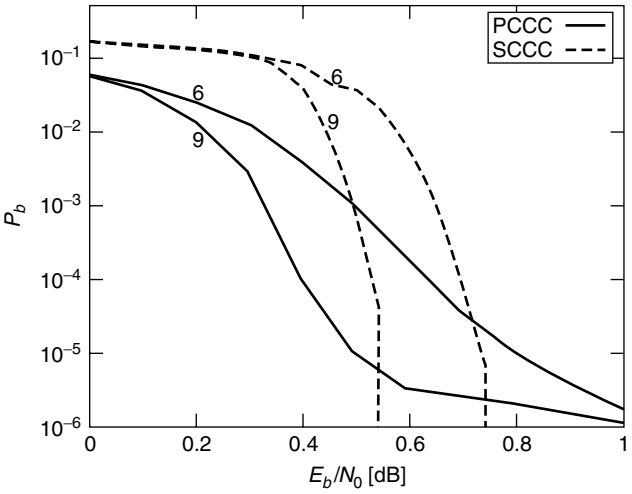


Figure 15. Comparison of the two rate-$\frac{1}{3}$ PCCCs and SCCCs of Example 2. The curves refer to three and seven iterations of the decoding algorithm and to an equal input decoding delay of 16384.

6. EXAMPLES OF SERIAL CONCATENATIONS

In this section we present a few significant examples of digital transmission systems that can be interpreted, and thus be utilized, as serial concatenations of individual modules. We will show that the serial structure fits to a set of systems well beyond the classical one constituted by two concatenated encoders with an interleaver in between.

6.1. Serial Concatenation of an Outer Encoder and an Inner Modulator

The serial concatenation of an outer encoder with rate R_c^o with an inner modulator characterized by an alphabet of $M = 2^m$ channel symbols through an interleaver is shown in Fig. 16 [25].

The binary symbols $\mathbf{u}$ emitted by the source are encoded by the outer binary encoder. The coded bits $\mathbf{c}$ are interleaved first, and then serial-to-parallel converted to m parallel bit streams $\mathbf{c_i}, \mathbf{i} = \mathbf{1}, \ldots, \mathbf{m}$. A mapper associates

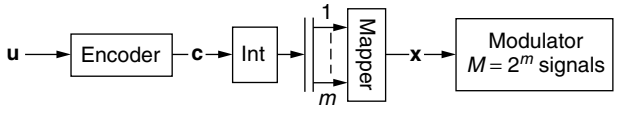


Figure 16. Block diagram of the serial concatenation of an outer encoder with an inner modulator through a bit interleaver.

each bit m-tuple to a modulator symbol x, which is then sent to the channel. The *spectral* efficiency of the scheme is related to the number of information bits per channel symbol $n_m \triangleq mR_c^o$.

A scheme like the one depicted in Fig. 16 can be detected or decoded iteratively, by considering it as a serial concatenation with interleaver. The condition to achieve improved performance through the iterative algorithm is that the mapping of m bits to a modulated signal be an operation with memory

$$\mathrm{P}[u_1, u_2, \ldots, u_m \mid r] \neq \mathrm{P}[u_1 \mid r]\mathrm{P}[u_2 \mid r] \cdots \mathrm{P}[u_m \mid r]$$

where r is the received signal.

Example 3. Consider an outer 8-state, rate-$\frac{1}{2}$ systematic recursive convolutional encoder punctured to rate $\frac{2}{3}$, followed by an interleaver with length 12,003, a mapper performing three different kind of mappings — natural, Gray mapping, or "anti-Gray" mapping — and an 8-PSK modulator. The three mappings are shown in Fig. 17. Note that the anti-Gray mapping has been chosen so as to maximize the sequence d_1, d_2, d_3, d_4, where d_w is the minimum Euclidean distance between pair of signals whose binary labels have a Hamming distance equal to w. This mapping choice is in perfect agreement with the design rule for inner encoders found in Section 4, replacing Hamming with Euclidean distances, and is the opposite with respect to Gray mapping.

The whole system is shown in Fig. 18; its receiver consists of the LLR computation on 8-PSK symbols, followed by a soft demapper that projects the symbol LLRs onto bit LLRs, the interleaver/deinterleaver and the puncturer/depuncturer pairs, and, finally, the outer SISO working on the rate-$\frac{1}{2}$ trellis. In Fig. 19 we report the bit error probability of such a scheme as a function of E_b/N_0, obtained by simulation with one to five iterations of the detection/decoding algorithm. The solid black curves refer to natural mapping, and show a 2.5-dB gain at 10^{-5}. The dashed line refers to Gray mapping, and shows that the iterations do not bring any performance improvement, although yielding better results than the first iteration with natural mapping. Finally, the solid gray curves refer to the anti-Gray mapping. With 1 iteration, this is the worst mapping; as the iterations evolve, however, the gain is very significant, and a gain of almost 4 dB over the Gray mapping is obtained at 10^{-5}, which further increases for higher E_b/N_0 because of the largest slope of the curves.

To further improve the performance, a rate-1, 2-state differential encoder could be added before the modulator, leading to the scheme of Fig. 20, in which the inner "encoder" is now recursive, as prescribed by the design rules of Section 4. Its performance has been reported elsewhere [26], and an extension to the case of intersymbol interference channels and Turbo equalization has been proposed [27].

6.2. The Encoded Continuous-Phase Modulation

Continuous-phase modulation (CPM) is a class of *constant-envelope*, *continuous-phase* modulation schemes obtained through a modulator with finite memory that can be described as a finite-state machine, or, equivalently, a trellis. Because of its attractive properties of constant envelope and bandwidth compactness, also maintained through nonlinear devices, CPM has been a subject of extensive studies and publications in the late 1970s and 1980s, and has then been applied in various radio systems, such as for example in GSM. Reference 28 contains an in-depth treatment of this subject.

A general CPM modulator can be split [29] into the cascade of a finite-state machine, the continuous-phase encoder (CPE), and a memoryless modulator (MM). Adding an outer convolutional encoder and an interleaver leads then to the system depicted in Fig. 21, where the information bits are first encoded by a convolutional encoder with rate k/n, followed by a bit interleaver with length N, and by an M-ary CPM modulator. Such a structure is capable of transmitting $(k/n)\log_2 M$

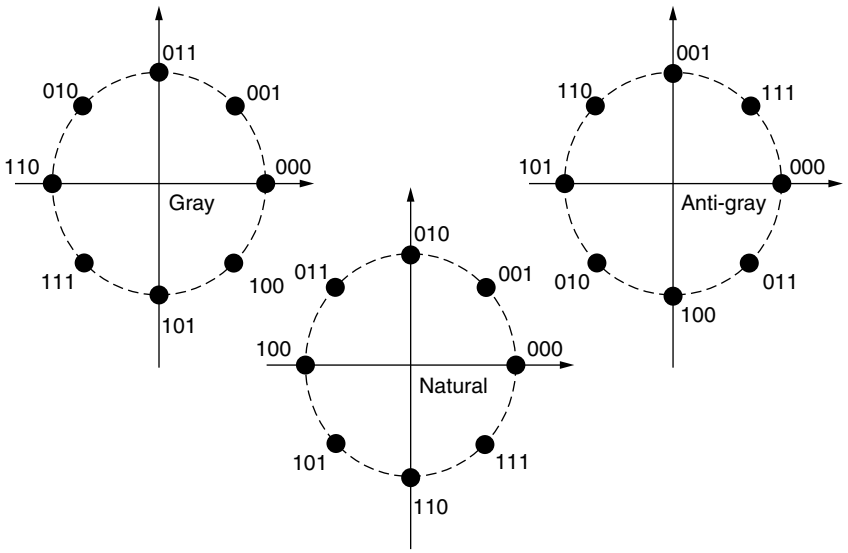


Figure 17. Three different mappings from triplets of bits to 8-PSK signals: Gray, natural, and anti-Gray.

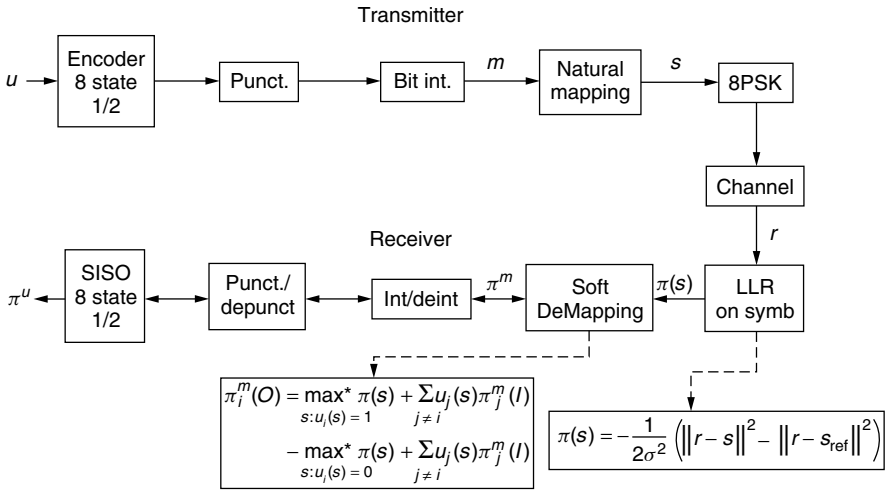


Figure 18. Block diagram of the transmitter and receiver of the serial concatenation of an outer encoder with an inner modulator through a bit interleaver.

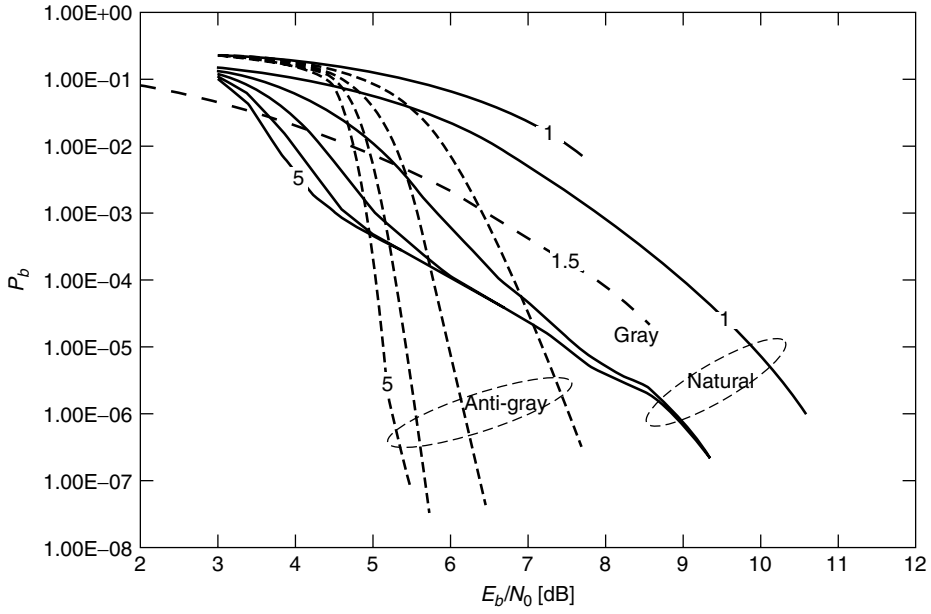


Figure 19. Bit error probability versus E_b/N_0 for the serial concatenation of Example 3. The solid black curves refer to natural mapping, the dashed one to Gray mapping, and the solid gray curves to anti-Gray mapping, with one to five iterations of the decoding algorithm.

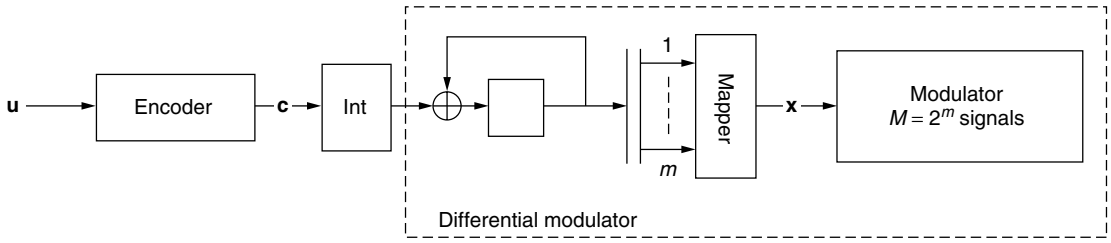


Figure 20. Block diagram of the serial concatenation of an outer encoder with an inner differential encoder and a modulator through a bit interleaver.

information bits per CPM signal, achieving a bandwidth efficiency that depends on the frequency pulse and modulation index of the CPM scheme.

According to Fig. 21, the cascade of a convolutional encoder, an interleaver, and a CPM modulator, a configuration that we will call *serially concatenated convolutional CPM* (SCC-CPM), can be seen as a particular form of serially concatenated convolutional codes with interleaver, and thus demodulated-decoded with an iterative scheme yielding high coding gains.

Example 4. Consider a coded CPM scheme (like the one depicted in Fig. 21), in which the outer encoder is a 4-state, rate-$\frac{2}{3}$ recursive systematic convolutional encoder,

and the inner module is a quarternary partial response CPM scheme [28] employing a rectangular frequency pulse with duration equal to 2 symbol intervals (the so-called 2-REC pulse) and modulation index $h = \frac{1}{3}$. Using the iterative receiver of Fig. 21, one obtains the performance shown in Fig. 22. The simulated scheme, owing to the low modulation index, has a high bandwidth efficiency, and achieves very good performance. Also, the performance improve very significantly with iterations.

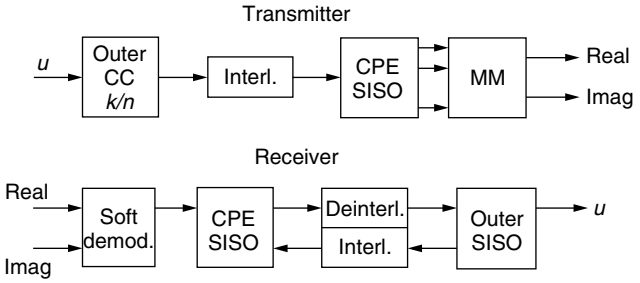


Figure 21. Block diagram of the encoded CPM transmitter and receiver.

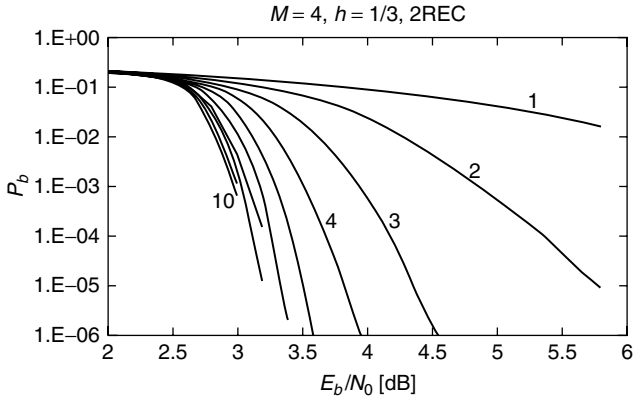


Figure 22. Bit error probability curves versus E_b/N_0 for the quaternary CPM scheme employing the 2-REC pulse with modulation index $h = \frac{1}{3}$. The outer code rate is $\frac{2}{3}$, and the interleaver size corresponds to 8190 information bits.

6.3. Serial Concatenation of an Outer Encoder with the Magnetic Recording Channel

In Fig. 23 we show the block diagram of a coded magnetic recording system based on the serial concatenation paradigm. It uses as outer code a high-rate (like $\frac{8}{9}$, $\frac{15}{16}$; this is a mandatory constraint of this applications that requires very high recording densities) terminated convolutional encoder, and as inner module the magnetic recording channel preceded by a precoder that makes it recursive. An interleaver separates the two blocks. In the same figure, the lower scheme represents the precoder/channel model. Using the Lorentzian model with an EPR4 partial response target [30], the overall magnetic channel transfer function in the transform domain is

$$G_m(Z) = 1 + Z - Z^2 - Z^3$$

It can be seen as an intersymbol interference channel with non binary (five-level) outputs, so this example becomes similar to the case of Turbo equalization [31]. The precoder is characterized by a $1/(1 \oplus Z^2)$ transfer function.

The iterative decoder (see Fig. 23, upper portion) is made by two SISOs. The inner SISO (SISO channel in the figure) is matched to the precoder/channel response, and works on non binary symbols, whereas the outer SISO works for the very high-rate $(n, n-1)$ convolutional encoder. Two solutions are possible: the first makes use of a heavily punctured rate-$\frac{1}{2}$ "mother" encoder, and the second employs unpunctured encoders. In the latter case, the complexity of the SISO, which depends on the number of edges per trellis section per decoded bit is too high, and thus a different approach must be followed, like designing a SISO that works on the $(n, 1)$ *dual* code [32]. This can be made coincident with a standard SISO, provided that its inputs be converted from LLRs to the so-called *log-reflection coefficients* (LRC in the figure).

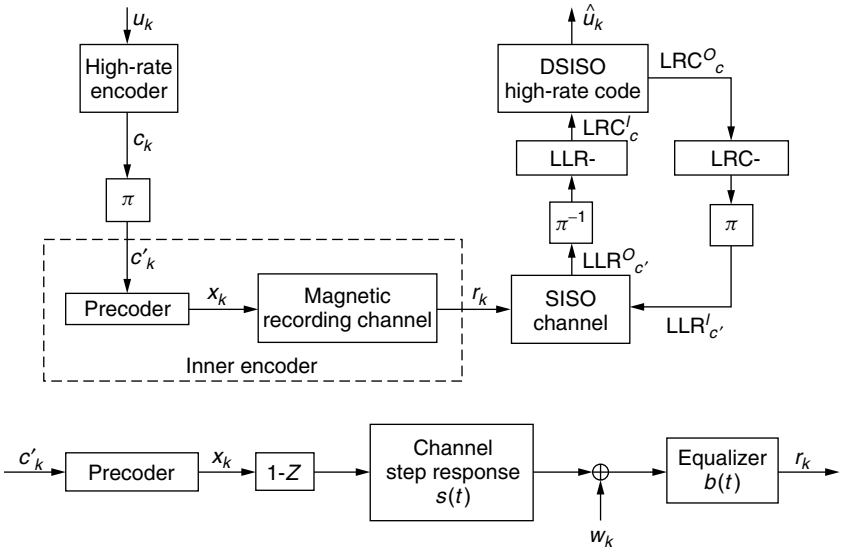


Figure 23. Block diagram of a coded magnetic recording system (upper diagram). Lorentzian model of precoder/magnetic channel (lower diagram).

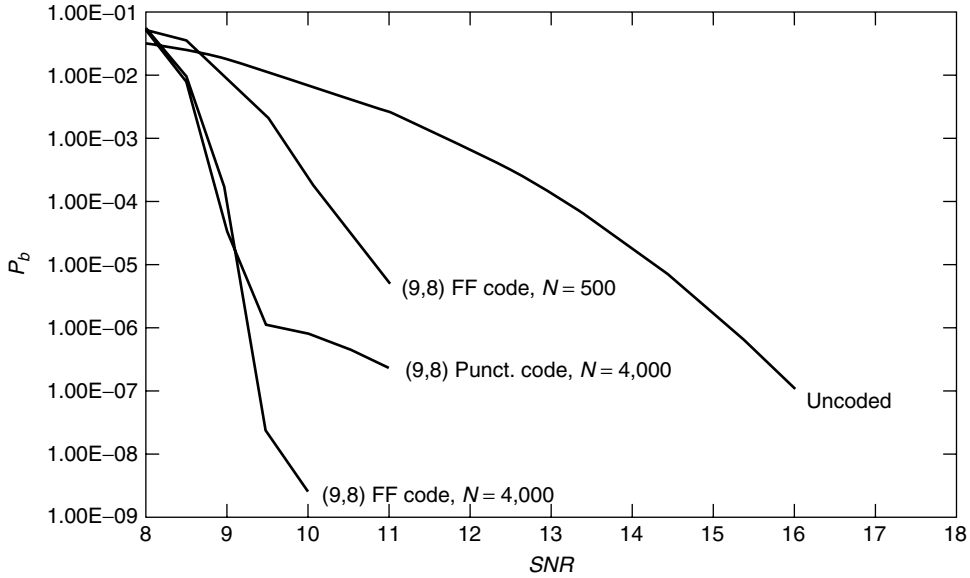


Figure 24. Bit error probability versus signal-to-noise ratio for the magnetic recording system of Fig. 23. The curves refer to the uncoded case, to the coded case employing a punctured rate-$\frac{8}{9}$ outer encoder with input block size 4000, and to the coded case using an unpunctured, optimized rate-$\frac{8}{9}$ outer encoder with block sizes 500 and 4000.

Using the iterative decoding algorithm, we have obtained the results shown in Fig. 24. The curves refer to the uncoded case, to the coded case employing a punctured rate-$\frac{8}{9}$ outer encoder with input block size 4000, and to the coded case using an unpunctured, optimized [33] rate $\frac{8}{9}$ outer encoder with block sizes 500 and 4000. A very large coding gain (6 dB for block size 4000 at 10^{-6}) can be obtained, without any significant error floor down to 10^{-9} for the unpunctured case. The punctured code, on the other hand, behaves similarly down to 10^{-6}, but shows the error floor just below. Notice that the different behavior of the two codes in the error floor region is due to the different free distance of the outer encoder, which is 2 for the punctured code (interleaver gain N^{-1}), and 3 for the unpunctured one (interleaver gain N^{-2}), in perfect agreement with the design rules of Section 4.

6.4. The Multiuser Interfered Channel

In coded code-division multiple access (CDMA) systems, K independent users first encode their information bit streams, then, after interleaving, they transmit them onto the same channel, as shown in Fig. 25. The cascade

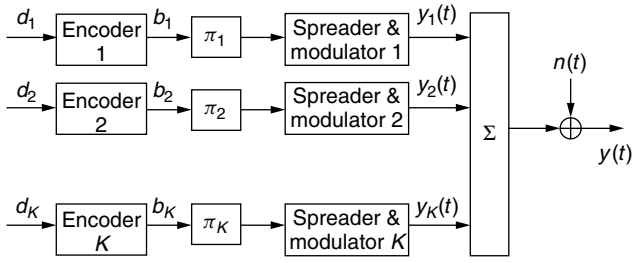


Figure 25. The CDMA transmitter with K users.

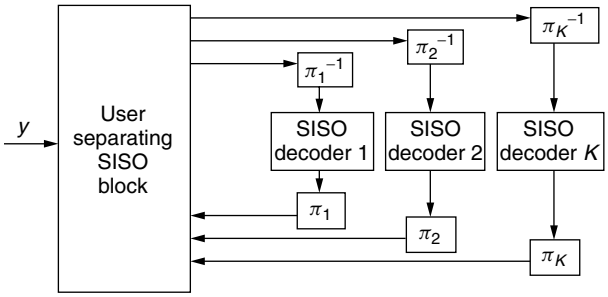


Figure 26. Turbo receiver. At every iteration, the user separator exchanges information with the decoders.

of the channel encoder, interleaver, and multiaccess channel can be viewed as a serial concatenation. As a consequence, at the receiver side, an iterative receiver can be employed, as shown in Fig. 26. A user separator, which is a soft-output version of a multiuser detector, attenuates for each user the multiaccess interference and sends K streams of soft output values to K channel decoders after deinterleaving. The decoders, in their turn, feed back with updated extrinsic LLRs the user separator, which will use the feedback information in the following iteration, to improve the user separation. In the last iteration, the decoders provide the final estimate of the K information bit streams. Figure 27 shows an example of performance along the iterations. It refers to a system with four equipower synchronous users, with correlation among any pairs equal to 0.7, and outer rate $\frac{1}{2}$, equal 16-state convolutional encoders with generator matrix

$$G_o(Z) = [1 + Z + Z^4, 1 + Z^2 + Z^3 + Z^4]$$

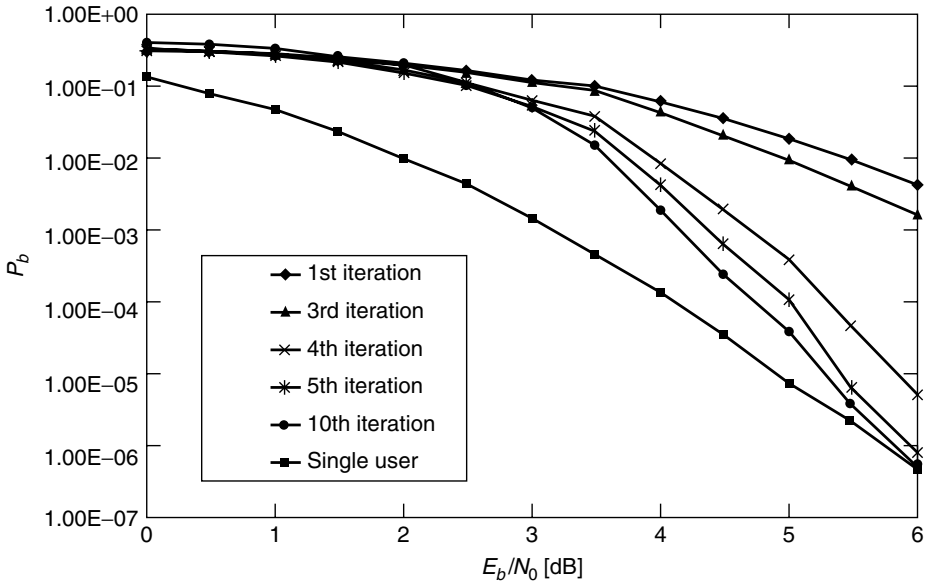


Figure 27. Results of simulation. $K = 4$ synchronous users, all with the same power. The correlation between any pair of users is 0.7. Interleaver size is 256, and the outer encoders are rate-$\frac{1}{2}$, 16-state feedforward convolutional encoders.

The interleavers are purely random interleavers with size 256. Finally, the user separator employs the algorithm described elsewhere [34].

As the curves show, the receiver tends to the single-user performance for sufficiently high signal-to-noise ratios.

7. CONCLUSIONS

In a semitutorial way avoiding all except strictly necessary mathematical developments, in this article we have described the main characteristics of serially concatenated codes with interleavers. They consist of the cascade of an outer encoder, an interleaver permuting the outer codeword bits, and an inner encoder whose input words are the permuted outer code words. Decoding is performed by an iterative technique based on two soft-input/soft-output algorithms tailored to the trellis of the outer and inner encoders. The iterative algorithm can be easily extended to other forms of concatenation, where one or both encoders are replaced with system modules with some form of memory performing different functions, like a modulator, a channel with intersymbol interference, a multiuser combiner. First, upper bounds to the *average* maximum-likelihood error probability of serially concatenated block and convolutional coding schemes have been described. Then, design guidelines for the outer and inner encoders that maximize the *interleaver* gain and the asymptotic slope of the error probability curves have been derived, together with the iterative decoding algorithm. Finally, extensions to different systems that can be seen as serial concatenations and detected or decoded accordingly have been proposed.

Acknowledgments

The authors gratefully acknowledge the contributions of Alex Graell i Amat and Alberto Tarable for providing the magnetic recording and multiuser detection results in Section VII.

BIOGRAPHIES

Sergio Benedetto is a Full Professor of Digital Communications at Politecnico di Torino, Italy since 1981. He has been a Visiting Professor at University of California, Los Angeles (UCLA), at University of Canterbury, New Zealand, and is an Adjoint Professor at Ecole Nationale Superieure de Telecommunications in Paris. In 1998 he received the Italgas Prize for Scientific Research and Innovation. He has also been awarded the title of Distinguished Lecturer by the IEEE Communications Society. He has co-authored two books on probability and signal theory (in Italian), the books *Digital Transmission Theory* (Prentice-Hall, 1987), *Optical Fiber Communications* (Artech House, 1996), and *Principles of Digital Communications with Wireless Applications* (Plenum-Kluwer, 1999), and over 250 papers in leading journals and conferences. He has taught several continuing-education courses on the subject of channel coding for the UCLA Extension Program and for the CEI organization. He was the Chairman of the Communications Theory Symposium of ICC 2001, and is the Area Editor for the *IEEE Transactions on Communications for Modulation and Signal Design*. Professor Benedetto is the Chairman of the Communication Theory Committee of IEEE and a Fellow of the IEEE.

Guido Montorsi was born in Turin, Italy, on January 1, 1965. He received the Laurea in Ingegneria Elettronica in 1990 from Politecnico di Torino, Turin, Italy, with a master thesis, concerning the study and design of coding schemes for high-definition television (HDTV), developed at the RAI Research Center, Turin. In 1992 he spent the year as visiting scholar in the Department of Electrical Engineering at the Rensselaer Polytechnic Institute, Troy, NY. In 1994 he received the Ph.D. degree in telecommunications from the Dipartimento di Elettronica of Politecnico di Torino.

In December 1997 he became assistant professor at the Politecnico di Torino. In July 2001 he became Associate Professor. In 2001–2002 he spent one year in the startup company Sequoia Communications for the innovative design and implementation of a third-generation WCDMA receiver.

He is author of more than 100 papers published in international journals and conference proceedings. His interests are in the area of channel coding and wireless communications, particularly the analysis and design of concatenated coding schemes and study of iterative decoding strategies.

BIBLIOGRAPHY

1. G. D. Forney, Jr., *Concatenated Codes*, MIT Press, Cambridge, MA, 1966.
2. R. H. Deng and D. J. Costello, High rate concatenated coding systems using bandwidth efficient trellis inner codes, *IEEE Trans. Commun.* **COM-37**(5): 420–427 (May 1989).
3. J. Hagenauer and P. Hoeher, Concatenated Viterbi decoding, *Proc. 4th Joint Swedish-Soviet Int. Workshop on Information Theory*, Gotland, Sweden, Studenlitteratur, Lund, Aug. 1989, 29–33.
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. ICC'93*, Geneve, Switzerland, May 1993.
5. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* **44**: 909–926 (May 1998).
6. D. Divsalar and F. Pollara, Serial and hybrid concatenated codes with applications, *Proc. Int. Symp. Turbo Codes and Related Topics*, Brest, France, Sept. 1997.
7. S. Benedetto and G. Montorsi, Generalized concatenated codes with interleavers, *Proc. Int. Symp. Turbo Codes and Related Topics*, Brest, France, Sept. 1997.
8. R. Garello, G. Montorsi, S. Benedetto, and G. Cancellieri, Interleaver properties and their applications to the trellis complexity analysis of turbo codes, *IEEE Trans. Commun.* **49**: 793–807 (May 2001).
9. S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, Kluwer Academic/Plenum, New York, 1999.
10. S. Benedetto and G. Montorsi, Unveiling turbo-codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* **42**(2): 409–429 (March 1996).
11. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Analysis, design and iterative decoding of double serially concatenated codes with interleavers, *IEEE J. Select. Areas Commun.* **16**: 231–244 (Feb. 1998).
12. S. Y. Chung, T. J. Richardson, and R. L. Urbanke, Analysis of sum-product decoding of low-density parity-check codes using a Gaussian approximation, *IEEE Trans. Inform. Theory* **47**: 657–670 (Feb. 2001).
13. H. El Gamal and A. R. Hammons, Jr., Analyzing the turbo decoder using the Gaussian approximation, *IEEE Trans. Inform. Theory* **47**: 671–686 (Feb. 2001).
14. D. Divsalar, S. Dolinar, and F. Pollara, Iterative turbo decoder analysis based on density evolution, *IEEE J. Select. Areas Commun.* **19**: 891–907 (May 2001).
15. S. ten Brink, Convergence behavior of iteratively decoded parallel concatenated codes, *IEEE Trans. Commun.* **49**: 1727–1737 (Oct. 2001).
16. S. Benedetto and G. Montorsi, Design of parallel concatenated convolutional codes, *IEEE Trans. Commun.* **44**: 591–600 (May 1996).
17. D. Divsalar and R. J. McEliece, Effective free distance of turbo codes, *Electron. Lett.* **32**: 445–446 (Feb. 1996).
18. S. Benedetto, R. Garello, and G. Montorsi, A search for good convolutional codes to be used in the construction of turbo codes, *IEEE Trans. Commun.* **46**: 1101–1105 (Sept. 1998).
19. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Soft-input soft-output modules for the construction and distributed iterative decoding of code networks, *Eur. Trans. Telecommun.* **9**: 155–172 (March 1998).
20. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**: 284–287 (March 1974).
21. R. J. McEliece, On the BCJR trellis for linear block codes, *IEEE Trans. Inform. Theory* **42**: 1072–1091 (July 1996).
22. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**(2): 429–445 (March 1996).
23. P. Robertson, E. Villebrun, and P. Hoeher, A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain, *Proc. ICC'95*, Seattle, June 1995.
24. D. Divsalar and F. Pollara, Turbo codes for PCS applications, *Proc. ICC'95*, Seattle, June 1995.
25. X. Li and J. A. Ritcey, Trellis-coded modulation with bit interleaving and iterative decoding, *IEEE J. Select. Areas Commun.* **17**: 715–724 (April 1999).
26. P. Hoeher and J. Lodge, Turbo DPSK: Iterative differential PSK demodulation and channel decoding, *IEEE Trans. Commun.* **47**: 837–843 (June 1999).
27. A. Dejonghe and L. Vandendorpe, Low-complexity turbo-equalization for coded multilevel modulations, *Proc. SCVT2001*, Delft (The Nederlands), Oct. 2001.
28. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
29. B. E. Rimoldi, A decomposition approach to CPM, *IEEE Trans. Inform. Theory* **34**: 260–270 (March 1988).
30. H. N. Bertram, *Theory of Magnetic Recording*, Cambridge Univ. Press, 1994.
31. C. Laot, A. Glavieux, and J. Labat, Turbo equalization: adaptive equalization and channel decoding jointly optimized, *IEEE J. Select. Areas Commun.* **19**: 1744–1752 (Sept. 2001).
32. G. Montorsi and S. Benedetto, An additive version of the SISO algorithm for the dual code, *Proc. IEEE Int. Symp. Inform. Theory*, ISIT'2001, 2001.
33. A. Graell i Amat, G. Montorsi, and S. Benedetto, New high-rate convolutional codes for concatenated schemes, *Proc. IEEE Int. Conf. Communications*, ICC'2002, New York, May 2002.
34. A. Tarable, G. Montorsi, and S. Benedetto, A linear front end for iterative soft interference cancellation and decoding in coded CDMA, *Proc. IEEE Int. Conf. Communications*, ICC'2001, Vol. 1, 2001.

SERIALLY CONCATENATED CONTINUOUS-PHASE MODULATION WITH ITERATIVE DECODING

PÄR MOQVIST
TOR AULIN
Chalmers University of Technology
Göteborg, Sweden

1. INTRODUCTION

Digital radio communications has evolved from a tiny research area to mass-market products used by millions of people worldwide. Especially the 1990s saw a tremendous development of mobile telephony. Compared to earlier analog radio systems, digital communications enables cheaper, smaller, and more complex devices through the progress in integrated digital circuit technology. Naturally, this draws increasing attention to research in digital communications.

In this article, we focus on the transmission of digital messages of any kind; thus, we do not cover how these messages are produced. They may stem from an analog source that is digitized (such as speech, audio, or video), or they may be digital in their original nature (such as text messages and computer files). It is, however, essential that the message does not contain any redundancy, that is, that it consists of digital symbols that can be considered independent and identically distributed from a statistical point of view. This normally requires some pre-processing such as speech or video coding, or data compression.

Even in digital communications, messages are converted to analog signals before transmission on a physical channel. This process, known as digital modulation, is adapted to the specific transmission medium in question, such as coaxial cables, optical fibers, or radio channels. Common digital modulation formats include binary and quadrature phase shift keying (BPSK/QPSK), quadrature amplitude modulation (QAM), frequency shift keying (FSK), and digital phase modulation (PM).

A radio transmitter normally consists of a modulator producing radiofrequency (RF) signals, a signal amplifier and an antenna. In many applications, it is both easier and cheaper to use a nonlinear amplifier than a linear one, meaning that only phase-modulated signals can be amplified without considerable distortion. Thus, modulation formats which also vary the signal amplitude (e.g., BPSK, QPSK, QAM) are not suitable in conjunction with nonlinear transmitter amplifiers. It is here that the broad class of continuous-phase modulation (CPM) schemes find their major applications [1,2]. Examples include the GSM mobile telephony standard, microwave radio links, and ground-to-satellite communications.

Reliable digital transmission can be obtained by channel coding, as discussed by Shannon [3] in his 1948 classic paper. Block codes and convolutional codes are basic channel codes that are able to lower the bit error rate (BER) in the receiver at the cost of increased complexity. More advanced codes can be constructed from these basic elements by concatenation. One such technique is Turbo coding or parallel concatenated codes (PCCs), introduced by Berrou et al. [4–7], in which two convolutional codes separated by a random interleaver (permuter) are concatenated in parallel. The two codes are decoded separately in an iterative manner (see Fig. 1). Turbo codes have been shown to perform very close to Shannon's limit, while maintaining a reasonable decoder complexity. The same is true for serially concatenated codes (SCCs) with random interleaving, as shown in Fig. 2 - as examined by Benedetto et al. [8,9]. Turbo codes are currently being introduced in the third-generation mobile system standards UMTS (Universal Mobile Telecommunications System) [10] and cdma2000 [11].

Channel coding and modulation can also be combined, as in convolutionally coded CPM. By a joint design of the code and the CPM system, improved BER performance can be obtained. This leads us to the technique of combining CPM and SCCs with random interleaving, namely, serially concatenated CPM (SCCPM) investigated by several authors [12–14]. Here, the inner (second) code in an SCC is substituted with a CPM modulator, producing a coded and interleaved CPM scheme which can be decoded iteratively (see Fig. 3). As for Turbo codes and SCCs, the use of a random interleaver leads to substantial performance gains. In this article, we describe SCCPM in detail and show some interesting performance examples.

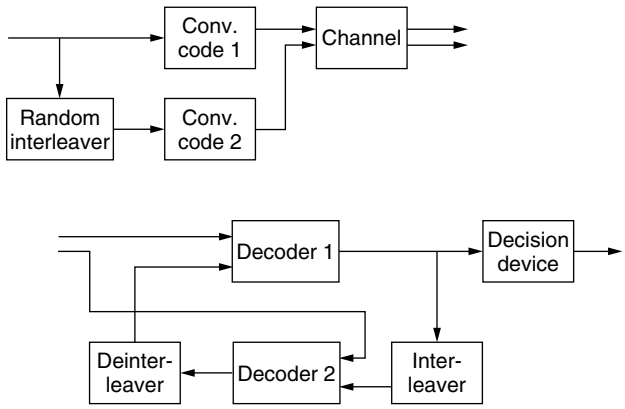


Figure 1. A typical Turbo code (PCC) with encoder (upper) and iterative decoder (lower).

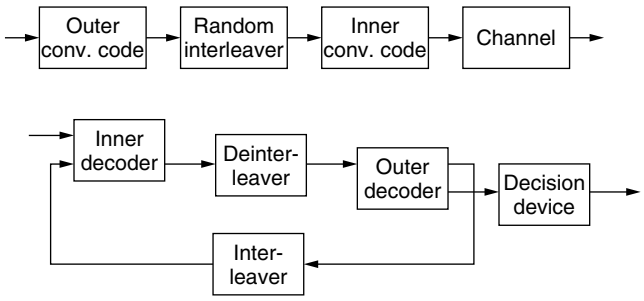


Figure 2. Serially concatenated code (SCC) with encoder (upper) and iterative decoder (lower).

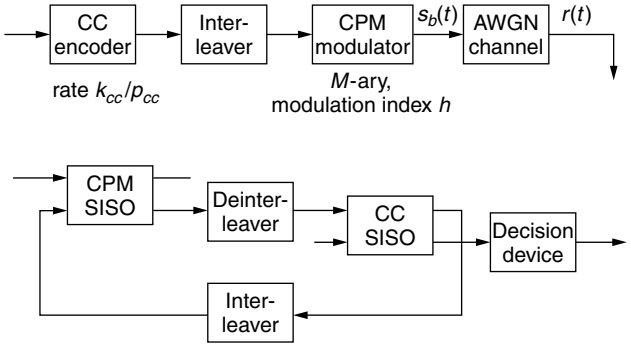


Figure 3. Serially concatenated continuous-phase modulation (SCCPM) with iterative decoding.

2. BACKGROUND

The introduction of the Turbo codes (PCCs) in 1993 [4] led to a vast research interest in concatenated convolutional codes separated by a pseudorandom bit interleaver and decoded iteratively. This technique of achieving large and nearly random codes while maintaining a simple decoder structure has found many applications, primarily where a large coding gain is desired (such as in deep-space applications or terrestrial links). The Turbo codes have shown to yield BERs around 10^{-5} at rates well beyond the channel cutoff rate, [7].

Shortly after the introduction of the turbo codes, it turned out that serially concatenated codes (SCCs) with a pseudo-random bit interleaver were equally suitable for iterative decoding [8]. It is generally understood that SCCs based on convolutional codes can give even better performance than turbo codes for low BERs (typically $<10^{-6}$) [9], while PCCs are better suited for medium-to-high BERs. Still, many concatenated coding schemes assume a simple carrier frequency modulation such as BPSK, meaning that a rate-k/n SCC transmits k/n bits per channel symbol. This bandwidth efficiency can be improved by more advanced modulation techniques such as trellis-coded modulation (TCM) [15] and continuous-phase modulation (CPM) [1,2]. Both TCM and CPM introduce memory into the transmitted signal, in a way that allows them to be described by finite-state machines similar to a convolutional code. Benedetto et al. employed TCM in an SCC with iterative decoding, such that the inner code is a recursive trellis code designed for the chosen signal set [16]. This way, it is possible to use, for example, a rate-$\frac{2}{3}$ SCC combined with an 8-PSK signal set yielding a bandwidth efficiency of 2 bits per symbol.

On the other hand, CPM is the natural choice of modulation when a constant envelope of the transmitted signal is required, such as when the transmitter amplifier is not perfectly linear. CPM was subject to vast research interest in the late 1970s and early 1980s [1], and the reader is referred to Ref. 2 for a comprehensive summary of much of that work. Later, convolutionally coded CPM was investigated as a way of improving performance while maintaining the constant envelope [17–20]. Matched codes [21] were introduced as yielding the minimum number of states in the optimum receiver when combined with CPM. The inherent modulo-2π property of the information-carrying phase in CPM has been exploited [22], such that convolutional codes over rings were combined with CPM. Since these codes showed better performance than did previous ones, much effort was put into code searches for various CPM systems [22–24].

In this article we study SCCPM, *i.e.* coded and interleaved CPM with iterative decoding. First, a brief description of CPM and how it is employed in a serially concatenated system is given (Section 3). Then, we study a transfer function bound to the BER performance, which is based on known concepts from SCCs (Section 4). This gives theoretical insights to the behavior of SCCPM in general. In Section 5, these observations are compared with numerous computer simulation results, demonstrating that the excellent performance of Turbo codes and SCCs indeed can be generalized to a trellis-coded modulation like CPM. Finally, in Section 6, bandwidth considerations are discussed and a bandwidth–performance comparison of some selected systems is presented.

3. SYSTEM DESCRIPTION

3.1. Block Diagram

Figure 3 shows a block diagram of SCCPM on an additive white Gaussian noise (AWGN) channel with iterative decoding. Independent and equiprobable information bits are encoded by a convolutional code (CC) with rate $R_{cc} = k_{cc}/p_{cc}$. For convenience, let p_{cc} be an integer multiple of $\log_2 M$, where M is the symbol alphabet size (cardinality) of the CPM system. As in Turbo codes and SCCs, bitwise block interleaving with block size N bits is employed, yielding groups of $\log_2 M$ bits that are mapped to CPM symbols using some mapping rule, such as natural mapping or gray mapping. The CPM modulator produces a constant envelope signal $s_b(t)$ that is transmitted on the AWGN channel. In this theoretical treatment, continuous transmission is assumed, thus the encoders are not reset to the zero state at the beginning of an interleaver block. The resulting concatenated code is a block code with information words of length $N \cdot R_{cc}$. (In a more practical system, blockwise decoding would be facilitated by trellis termination of both the outer CC and the inner CPM system [25], such that the encoders are enforced to the zero state at both the beginning and the end of a block. This normally leads to a small performance degradation.) In complex baseband representation, the received continuous-time signal is given by $r(t) = s_b(t) + n(t)$, where $n(t)$ is complex white Gaussian noise with double-sided power spectral density $N_0/2$.

3.2. Iterative Decoder

As in Turbo codes and SCCs, the iterative decoder consists of two *a posteriori* probability (APP) algorithms, one for the inner CPM system and one for the outer CC; more specifically, we use the bitwise sliding-window SISO algorithm proposed by Benedetto et al. [26] and formally justified by the present authors [27]. The SISO

algorithm is a nice generalization of the Bahl et al. (BCJR) algorithm [28], taking a priori probabilities of both information symbols and code symbols, and computing extrinsic APPs of both information symbols and code symbols. An extrinsic APP is an APP from which the a priori probability has been divided.

For the inner SISO, channel observations are used as a priori probabilities of the code symbols of the CPM system, as described in Section 3.4. The inner SISO then computes extrinsic APPs of the information bits of the CPM system; these are deinterleaved and used as a priori information on the code bits of the CC in the outer SISO. Applying the constraints of the CC, the outer SISO updates this information to extrinsic APPs that are interleaved and used as a priori information on the information bits of the CPM system in the next iteration. At the same time, the outer SISO computes APPs of the information bits of the CC; these are used by the decision device to select the bit with maximum APP in the last iteration. As can be seen from Fig. 3, those inputs/outputs not needed are left unconnected. A uniform distribution of the bits is assumed for an unconnected input. Note that at the first iteration, no a priori information on the CPM information bits is available in the inner SISO.

3.3. Description of CPM

Rimoldi [29] gave a concise description of CPM as the concatenation of two separate devices: a continuous-phase encoder (CPE) and a memoryless modulator (MM). He used the concept of a tilted-phase representation of CPM, which we repeat here for convenience.

Consider a CPM system with M-ary information symbols $u \in \{0, 1, \ldots, M-1\}$ transmitted every symbol interval T with energy E. To match it to the CC, let $\log_2 M$ be an integer number. The symbols are phase-modulated using a positive normalized frequency pulse $g(t)$ containing no impulses and being nonzero for L symbol intervals; thus the system is full response ($L = 1$) or partial response ($L > 1$). The phase response $q(t)$ is the integral of the frequency pulse, and it is normalized to $q(LT) = \frac{1}{2}$. The LREC and LRC families [2] are examples of frequency pulses.

The modulation index h is assumed to be rational and irreducible, $h = K/P$, since then the system can be described by a trellis [1]. The tilted-phase $\psi(t)$ during symbol interval n ($t = \tau + nT$) is given by [29]

$$\psi(\tau + nT) = \left[2\pi h \left[\sum_{i=0}^{n-L} u_i \mod P \right] + 4\pi h \sum_{i=0}^{L-1} u_{n-i} q(\tau + iT) + W(\tau) \right] \mod 2\pi \qquad 0 \le \tau < T \quad (1)$$

where $\{u_i\}$ are information symbols and the data-independent function $W(\tau)$ is given by

$$W(\tau) = \pi h (M-1) \frac{\tau}{T} - 2\pi h (M-1) \sum_{i=0}^{L-1} 9(\tau + iT) + \pi h (M-1)(L-1) \quad (2)$$

The transmitted signal during the same interval is

$$s(\tau + nT) = \mathrm{Re}\left\{ s_b(\tau + nT) \cdot \exp\left[j(2\pi f_1(\tau + nT) + \varphi_0) \right] \right\} \quad (3)$$

where $s_b(\tau + nT) = (2E/T)^{1/2} \cdot \exp[j\psi(\tau + nT)]$ is the complex baseband equivalent, $f_1 = f_0 - h(M-1)/2T$ is a shift of the center frequency f_0, and φ_0 is the initial phase of the carrier. Note from these relations that the transmitted signal during symbol interval n is completely specified by the current symbol u_n, the $L-1$ previous data symbols $u_{n-L+1}, \ldots, u_{n-1}$, and the accumulated value

$$v_n = \sum_{i=0}^{n-1} u_i \mod P \quad (4)$$

which can take only P values. Hence, in each symbol interval n, the CPE produces the code symbol (vector) $\boldsymbol{c}_n = [v_n, u_{n-L+1}, \ldots, u_n]$ to the MM, which transmits one of $P \cdot M^L$ signals $s_b(t, \boldsymbol{c}_n)$.

3.4. Channel Observations in CPM SISO Algorithm

As mentioned above, the SISO algorithm takes *a priori* probabilities of both information symbols, $\Pr(u_n = U)$, and code symbols, $\Pr(\boldsymbol{c}_n = \boldsymbol{C})$. For the inner CPM SISO, the latter are replaced by channel observations $p(\mathbf{r}_n | \mathbf{c}_n = \mathbf{C})$, where $\mathbf{r}_n$ is a sufficient statistic [30] obtained from $r(t)$. It can be shown [12] that the channel observations are proportional to $\exp[\mathrm{Re}\{r_k\}/N_0]$, where r_k is the sampled output of a filter matched to $s_b(t, \mathbf{c}_n)$,

$$r_k = \int_{nT}^{(n+1)T} r(t) s_b^*(t, \mathbf{c}_n)\, dt \quad (5)$$

and where the superscript * denotes complex conjugate. This sufficient statistic is exactly the same as when performing maximum-likelihood sequence detection (MLSD) in uncoded CPM [2], for example, using the Viterbi algorithm.

4. ANALYSIS OF SCCPM

In this section, an SCCPM system like the one in Fig. 3 is analyzed with the overall goal to demonstrate its ability to provide performance gains similar to Turbo codes and SCCs. These latter codes were analyzed through average transfer function bounds to the BER by Benedetto et al. in Refs. 6 and 9, respectively. As demonstrated in Ref. 12, this type of bound can be applied also to SCCPM. However, for all these codes, several arguments can be raised against the use of the transfer function bound:

1. Since this bound is based on a union bound, it diverges below a signal-to-noise ratio (SNR) threshold of approximately 3 dB, well above the operating point of many Turbolike systems.
2. The bound is a theoretical one in the sense that it is a bound on the average BER over all possible interleavers, each appearing with equal probability (the so-called uniform interleaver [6]). A specific

interleaver may yield substantially lower or higher BER. In fact, more recent interleaver design methods have shown improvements by orders of magnitude compared to the uniform interleaver [e.g., [31]].

3. The transfer function bound is based on optimal MLSD, which has a tremendous complexity for Turbolike systems because of their large and nearly random interleavers. The performance of the suboptimal iterative decoder can deviate substantially from MLSD, especially for low SNRs. Note that MLSD is equivalent to maximum APP sequence detection (MAPSD) because of the independent and equiprobable information bits.

Still, in this article we adhere to the transfer function bound because (1) performance above 3 dB is also of interest, and the transfer function bound is easy to obtain; (2) our primary goal is to demonstrate the Turbolike performance of SCCPM in general, as well as to compare different combinations of CCs and CPM systems; and (3) MLSD/MAPSD represents the ultimate performance limit of the system with any decoder, iterative or not. We do not address the convergence of the iterative decoder in this article, but refer the reader to other articles [e.g., 32–34,38]. The points presented above should anyway be kept in mind when bounds are compared with simulation results.

4.1. Transfer Function Bound for SCCPM

Similar to an SCC [9], the combination of a rate-R_{cc} CC, a bitwise length-N interleaver, and M-ary CPM is a block code with information words of length $k = N \cdot R_{cc}$. However, it does not possess the uniform error property, because CPM is not a linear code. Hence we cannot assume the all-zero information word to be transmitted; instead all pairs of transmitted and hypothesized (candidate) information words in the MLSD receiver must be considered. This is simplified by the fact that error events in CPM do not in general depend on the specific transmitted sequence, but only on the difference sequence between transmitted and hypothesized CPM symbols. Thus we need only enumerate the set of difference sequences in the inner CPM system, if we keep track of the fraction of transmitted sequences they correspond to. (Details of CPM error events can be found in Ref. 2.)

A union bound on the BER $P_b(e)$ in SCCPM with MLSD is

$$P_b(e) \le \frac{1}{2} \sum_{\mathcal{D}} B_{\mathcal{D}} \exp\left(-\frac{\mathcal{D} R_{cc} E_b}{2N_0}\right) \tag{6}$$

where E_b is the energy per information bit entering the CC and

$$B_{\mathcal{D}} = \sum_{w} \frac{w}{N \cdot R_{cc}} \overline{A}_{w,\mathcal{D}} \tag{7}$$

is the bit error multiplicity for error events with normalized squared Euclidean distance (NSED) $\mathcal{D}$ on the channel. The latter is assumed to be normalized to the energy per bit entering the CPM system; thus we use the same NSED as in uncoded CPM (2). $\overline{A}_{w,\mathcal{D}}$ is the number of concatenated error events with input weight w (Hamming distance $d_H(\cdot) = w$ between information words) and output weight (NSED) $\mathcal{D}$, averaged with respect to the number of transmitted codewords they correspond to. Thus $\overline{A}_{w,\mathcal{D}}$ is the input/output weight enumerating function (IOWEF), or input/output weight spectrum of the system. Assuming a uniform interleaver of length N bits, it can be shown that $\overline{A}_{w,\mathcal{D}}$ is obtained — the same as for SCCs — as

$$\overline{A}_{w,\mathcal{D}} = \sum_{l=0}^{N} \frac{A_{w,l}^{out} \cdot \overline{A}_{l,\mathcal{D}}^{in}}{\binom{N}{l}} \tag{8}$$

from the IOWEFs of the outer convolutional code and the inner CPM system. However, due to the linearity of the CC, its IOWEF need not be averaged. The denominator counts the number of distinct permutations that the uniform interleaver can produce from an output weight l codeword from the outer encoder, entering the inner CPM system as an input weight l codeword [9].

4.2. Error Events in CPM

Clearly, in concatenated and interleaved systems the input weight of error events is of interest. In the outer CC, the input weight counts the number of bit errors made during an error event, as given by the factor w in Eq. (7). But even more important, the input weight of error events in the inner CPM system determines not only which concatenated error events will be possible, but also their multiplicity, as can be seen from Eq. (8).

As an example, consider minimum shift keying (MSK) (binary 1REC, $h = \frac{1}{2}$). Error events in MSK start when the transmitted and hypothesized CPM symbols differ from each other for the first time in the block (see Fig. 4). When this occurs, there will be a Hamming distance of 1 between the bit labels of the transmitted and hypothesized symbols, that is, the input weight will increase by 1. At the same time, the NSED of the error event also will increase by 1. If the symbols differ again, the input weight and NSED will increase by the same amount, and the error event will end. However, if they do not differ, the Hamming distance will be 0 and the input weight will not increase, but the error event will continue. In this case, the NSED will increase by 2. Eventually, the symbols will differ once again, and a total input weight of 2 will be obtained, while the NSED will be any multiple of 2 depending on the length of the error event. Concluding, isolated error events in MSK will always have input weight 2, and the minimum NSED will also be 2.

For SCCs, the inner code should be a recursive systematic convolutional (RSC) code to avoid input weight 1 error events. This is because these events nullify

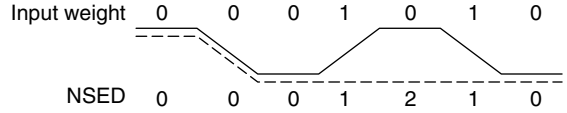


Figure 4. Input weight of an error event in MSK. Transmitted (solid) and hypothesized (dashed) symbol sequences.

the performance gain of an increased interleaver size N, the *interleaver gain* [9]. For CPM with $M > 2$, the input weight depends on the bit labeling (mapping) of the CPM symbols, including natural mapping or gray mapping. In addition, certain combinations of mapping and modulation index $h = K/P$ yield input weight 1 error events, despite the inherent recursive nature of CPM. More specifically, the following conditions apply:

- Binary ($M = 2$) CPM: integer modulation indices $h = K$. These systems are seldom used in practice.
- Quaternary ($M = 4$) CPM with natural mapping: $h = K/2$. With gray mapping: $h = K/3$.
- Octal ($M = 8$) CPM with natural mapping: $h = K/4$. With gray mapping: $h = K/3$, $h = K/5$, and $h = K/7$.
- Hexadecimal ($M = 16$) CPM with natural mapping: $h = K/8$. With gray mapping: $h = K/5$, $h = K/7$, $h = K/9$, $h = K/11$, $h = K/13$, and $h = K/15$.

Thus it is always possible to avoid input weight 1 error events, and hence to obtain interleaver gain, by a careful selection of the mapping.

4.3. Transfer Function Bound for Coded and Interleaved MSK

To find the IOWEF of the CC for all combinations of w and l, the methods described by Benedetto and Montorsi [6] or Divsalar et al. [35] can be used. The former is a recursive algorithm suitable when the output weights are integers. Regarding the inner CPM system, the IOWEF generally contains a manifold of real-valued NSEDs $\mathcal{D}$. However, for the important special case of minimum shift keying (MSK) (binary 1REC, $h = \frac{1}{2}$), all NSEDs are integers and thus the algorithm in Divsalar et al. [35] can be applied.

As an example, consider coded MSK with a uniform interleaver of length $N = 128$, 512, 2048, or 8192. The outer code is the 4-state, nonrecursive, nonsystematic rate-$\frac{1}{2}$ CC specified by the octal connection polynomial (7,5) (generating matrix $G(D) = [1 + D + D^2, 1 + D^2]$) and having free distance 5. Using the IOWEFs discussed above, Eqs. (7) and (8) give the total bit error multiplicity for each output weight $\mathcal{D}$ in the concatenated system, and Eq. (6) gives the upper bound on the average BER, shown in Fig. 5. Note the so-called divergence of the bound for E_b/N_0 (SNR) values below approximately 3 dB. This phenomenon has been observed also for Turbo codes [6] and SCCs [9], and is an inherent effect of the union bound. Above 3 dB, the bound exhibits a rather steep slope similar to those of SCCs. This promises for good practical performance, even with a suboptimal iterative decoder and a real, pseudorandom interleaver. Comparing the bounds for different interleaver sizes, it can be noted that when N increases by a factor of four, the BER bound decreases with roughly a factor 64; thus the interleaver gain appears to be approximately N^{-3}.

4.4. Detailed Analysis of Error Events

With a careful examination of the BER bound in Eqs. (6)–(8), several observations can be made regarding the behavior of SCCPM systems, when a uniform interleaver and MLSD is employed. For large SNRs, the minimum NSED $\mathcal{D}_{\min}$ dominates the bound. Thus SCCPM (as well as Turbo codes and SCCs) is no different from other coded modulation systems. However, the true strength of turbo-like systems is the spectral thinning of the distance (weight) distribution that occurs for large interleavers. This means that the lowest weights (distances) have small bit error multiplicities $B_{\mathcal{D}}$, effectively causing a displacement of the distance distribution toward larger weights. A "thin" weight spectrum gives improved BER performance for small-to-medium SNRs provided that the dominance of the exponential function in Eq. (6) is not too strong (i.e., that E_b/N_0 is not too large). As discussed

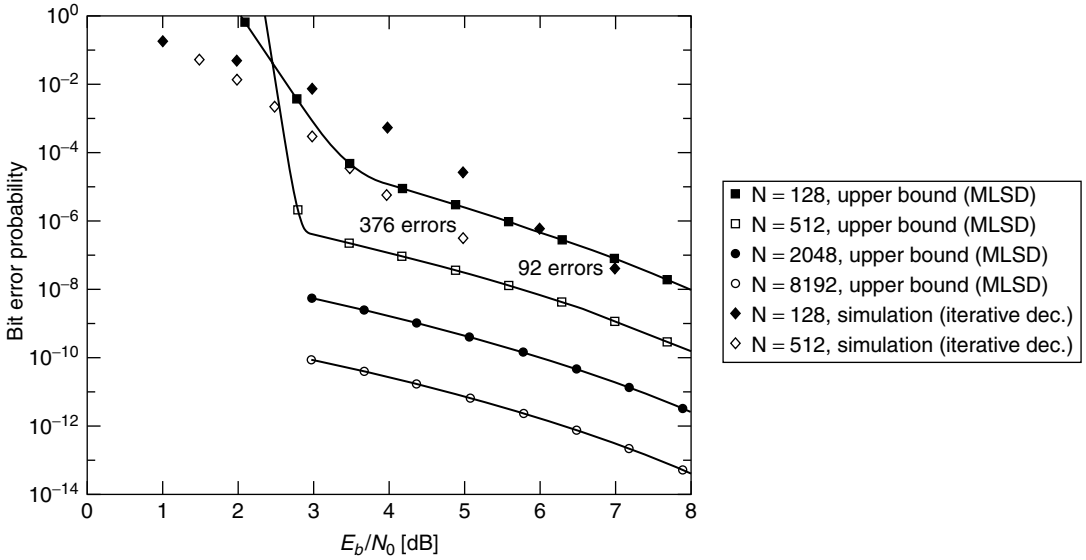


Figure 5. BER bounds and simulation results for (7,5)-coded MSK for various interleaver sizes N. The bounds for $N = 2048$ and $N = 8192$ were calculated from truncated weight spectra.

earlier, Turbolike systems have interleaver gain such that a larger interleaver decreases the bit error multiplicity [cf. Eq. (8)], and thus achieves spectral thinning.

As demonstrated for SCCs by Benedetto et al. [9], the dependency of $B_\mathcal{D}$ on the interleaver size N can be investigated with the aid of the BER bound. Specifically, $B_\mathcal{D}$ can be bounded from above, for each output weight $\mathcal{D}$, by a polynomial in N whose maximum exponent is $\alpha(\mathcal{D})$. When $\alpha(\mathcal{D}) < 0$, there is interleaver gain for the specific output weight $\mathcal{D}$. It can be shown that in general, $\alpha(\mathcal{D}_{\min}) < 0$ such that there is asymptotic (large E_b/N_0) interleaver gain.

However, the overall maximum value of the $\alpha(\mathcal{D})$, $\alpha_{\max}$, is more interesting because it determines whether there is interleaver gain over the whole SNR range. For SCCs, Benedetto et al. distinguish two separate cases [9]: (1) block and nonrecursive convolutional inner codes and (2) recursive convolutional inner codes. In case (1), input weight 1 events exist, yielding $\alpha_{\max} \geq 0$; thus, interleaver gain cannot be assured for all SNRs. In case (2), $\alpha_{\max} = -\lfloor (d_{\text{free}} + 1)/2 \rfloor < 0$ for an outer code with free distance d_{free}, because there are no input weight 1 events in the inner code. The crucial point is thus the existence of these events. We have already seen in Section 4.2 that certain CPM systems in fact do have such events — and we will also see that this affects the BER performance negatively. MSK does not have input weight 1 error events, so $\alpha_{\max} = -3$ and $B_\mathcal{D} \leq N^{-3}$ when it is combined with the (7,5) CC, for which $d_{\text{free}} = 5$. This is consistent with the bounds for different interleaver sizes in Fig. 5.

We can also study the NSED associated with $\alpha_{\max}$, $\mathcal{D}(\alpha_{\max})$, as an alternative to $\mathcal{D}_{\min}$ for small-to-medium SNRs. (For large SNRs, $\mathcal{D}_{\min}$ dominates the bound.) Let $\mathcal{D}^{\text{in}}(i)$ denote the minimum NSED for input weight i error events in the inner CPM system, and if no such event exists, let $\mathcal{D}^{\text{in}}(i) = \infty$. Thus, the classic minimum distance in CPM is the minimum of $\mathcal{D}^{\text{in}}(i)$ over all i. $\mathcal{D}^{\text{in}}(i)$ is given directly by the IOWEF of the CPM system, but it can also be obtained from a search in the difference trellis [12]. For inner CCs in SCCs, $\mathcal{D}^{\text{in}}(2)$ is called the *effective free distance* [9]. It is important since it can be shown [9] that $\mathcal{D}(\alpha_{\max})$ is given by a concatenation of input weight 2 events in the inner code, such that

$$\mathcal{D}(\alpha_{\max}) = \left\lfloor \frac{d_{\text{free}} + 1}{2} \right\rfloor \cdot \mathcal{D}^{\text{in}}(2) \tag{9}$$

(This equation is not valid when d_{free} is odd and $\mathcal{D}^{\text{in}}(3)$ is finite [12].) Clearly, performance can be improved not only by increasing the interleaver size, but also by a larger free distance in the outer CC, or a larger effective free distance in the inner CPM system. However, one should be careful with conclusions from $\mathcal{D}(\alpha_{\max})$ since it does not necessarily dominate the bound. As shown in another study [12], there are more combinations of error events that may play a role.

Example 1. Consider once again (7,5)-coded MSK. For the inner MSK system, $\mathcal{D}^{\text{in}}(2) = 2$ and $\mathcal{D}^{\text{in}}(i) = \infty$ for all other i. Hence $\mathcal{D}(\alpha_{\max}) = 6$ which coincides with $\mathcal{D}_{\min}$ for this specific system. It is interesting to note that this NSED is of the same order as that in many traditional coded CPM systems. However, the associated bit error multiplicity B_6 is overbounded by N^{-3}, which can be made as small as desired by enlarging the interleaver. This spectral thinning is not so easily obtained in traditional coded CPM.

Example 2. (7,5)-coded binary 2RC, $h = \frac{3}{4}$. Here, there are no odd input weight events in the CPM system, while $\mathcal{D}^{\text{in}}(2) = 2.66$. Therefore, $\mathcal{D}(\alpha_{\max}) = 7.97$ which is larger than $\mathcal{D}_{\min} = 5.21$. There is also a combination with $\mathcal{D} = 6.59$. The bit error multiplicities for these events are $B_{7.97} \leq N^{-3}$, $B_{6.59} \leq N^{-4}$, and $B_{5.21} \leq N^{-5}$. Thus the resulting performance is a mixture of the contributions to the bound of these events. Still, we can compare this system with Example 1 by observing that the event with lowest NSED (5.21) has a factor N^2 lower multiplicity, while the other events have larger NSEDs. Thus we can expect it to perform better than Example 1, at least above the divergence threshold around 3 dB. From the simulations in section V, this appears to be the case also for smaller SNRs.

4.5. Summary

Our observations can be summarized as follows. Note that they are generally valid only above the divergence threshold around 3 dB for systems employing a uniform interleaver and MLSD, although we will see in Section 5 that this appears to be the case also for more practical scenarios where a pseudo-random interleaver and iterative decoding is used in the low-SNR region.

- Like SCCs, SCCPM systems are capable of providing interleaver gain, if input weight 1 error events are avoided. This is always possible by a change in the mapping, such as from natural to gray, or vice versa. The order of the interleaver gain is determined by the free distance of the outer code, d_{free}.
- For a general inner CPM system, several combinations of the minimum NSEDs per input weight contribute to the BER bound. In order to compare different CPM systems, these combinations must be examined in detail for a given outer CC.
- It is always possible to avoid input weight 1 events by selecting a different mapping, specifically, natural binary instead of gray, or vice versa.

5. EXAMPLE SYSTEMS

The analysis in Section 4 is based on a union bound on the BER with MLSD, assuming a uniform interleaver. In practice, a suboptimal iterative decoder and a pseudorandom interleaver are used instead. Still, above the divergence threshold at E_b/N_0 around 3 dB, we shall see that the iterative decoder indeed performs quite close to the upper bound for MLSD. Below this value, the bound diverges and thus observations based on it may not be accurate; furthermore, the iterative decoder may not even converge toward the MLSD decision. All these points should be taken into account when performance of the iterative decoder is studied.

In this section, some examples of SCCPM systems are presented together with computer simulation results. We use (7,5)-coded MSK as our reference system, and then study the influence of other outer codes and inner CPM systems. Finally, a system with inner input weight 1 events is evaluated. In all simulations, the number of iterations[1] (8–12) as well as the delay of the sliding-window APP algorithms (10–15 symbols) are chosen such that no practical deterioration of performance could be noticed. The interleavers are chosen at random.

5.1. Reference System: (7,5)-Coded MSK

For this simple SCCPM scheme (Example 1, above), the MLSD transfer function bound is compared with simulation results for the iterative decoder in Fig. 5 for $N = 128$ and 512. For large SNRs, the actual interleaver and iterative decoding used in the simulations give slightly better performance than does the uniform interleaver in the MLSD bound, although the BER is not very accurate in the region below 10^{-6}. In fact, experience with Turbo codes tells us that most interleavers perform better than the average represented by the uniform interleaver. However, the discrepancy around the divergence threshold only be explained can by the fact that a suboptimal iterative decoder is used instead of MLSD.

Figure 6 shows more simulation results for different interleaver sizes, and also without any interleaver (i.e., a traditional coded CPM system, which was treated by Lindell [17] assuming MLSD detection). The interleaved system is evaluated using the iterative decoder (which has $4 + 2$ states) for $N = 128$, 512, 2048, 8192, and 32,768. For the noninterleaved system, both the iterative decoder and MLSD (8 states) is employed. Clearly, without interleaver, iterative decoding is inferior to MLSD by approximately 1.5 dB. At a BER of 10^{-3}, this transforms to a gain of more than 3 dB when a large interleaver is inserted.

[1] Here, and throughout the article, one decoder iteration corresponds to a pass through both the inner and the outer SISO module.

5.2. Influence of the Outer CC

According to the analysis above, the main influence of the outer code is the impact of its free distance d_{free} on the interleaver gain. Here, four different rate-$\frac{1}{2}$ outer CCs concatenated with MSK are investigated, whereby one is the reference system above. The codes are the 2-state (2,3) code, the 4-state (7,5) code, the 16-state (23,35) code, and the 64-state (133,171) code, and all codes are nonrecursive, nonsystematic convolutional codes. They have $d_{\text{free}} = 3$, 5, 7, and 10, yielding interleaver gains between N^{-2} and N^{-5}, and implying $\mathcal{D}(\alpha_{\max}) = 4$, 6, 8, and 10.

In Fig. 7, the codes are compared at an input delay of 4096 information bits ($N = 8192$). As can be seen, the required E_b/N_0 for a given BER $\geq 10^{-4}$ actually *increases* with d_{free}. This is in contrast to the analysis, where a larger d_{free} would give a larger interleaver gain. However, as pointed out earlier, this is true only above the divergence threshold, and for sufficiently large interleavers. In the

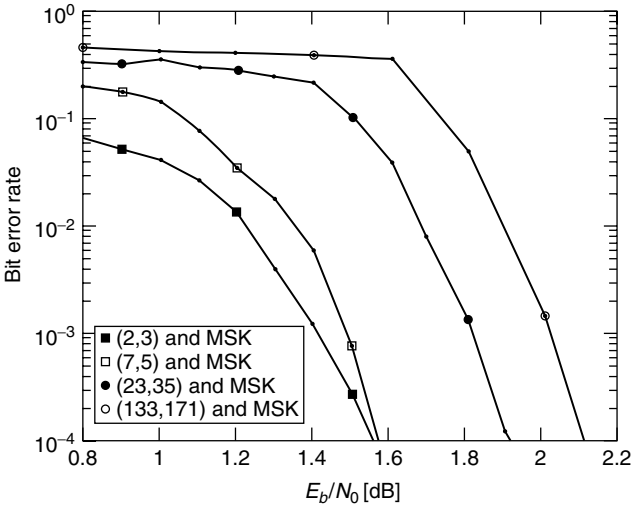


Figure 7. Simulation results for four different outer convolutional codes (CCs) with MSK, at a delay of 4096 information bits ($N = 8192$). All systems utilize the same bandwidth.

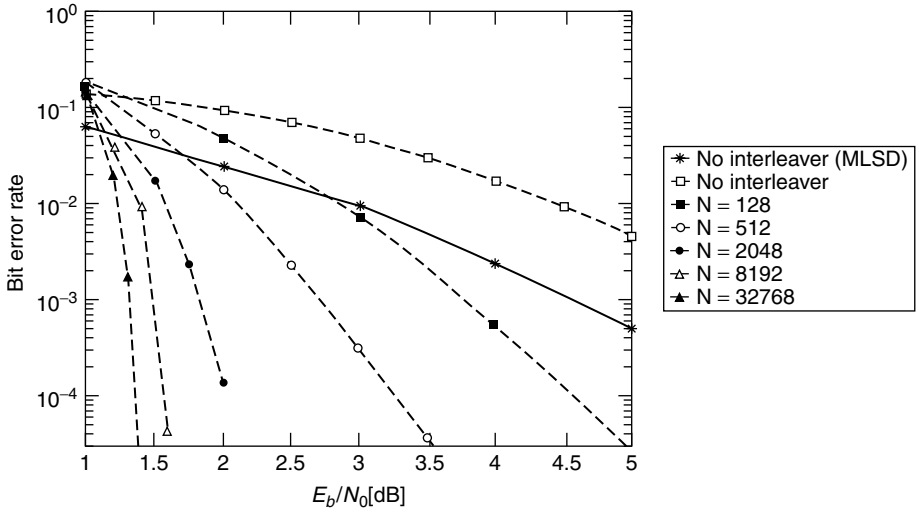


Figure 6. Simulation results for (7,5)-coded MSK ($4 + 2$ states in the iterative decoder).

low-SNR region, outer codes with more states suffer from convergence problems, as has been reported for SCCs [33]. The gain of a larger d_{free} can thus not be exploited in this SNR region. Still, it can be noticed that the 4-state code has a steeper BER slope than the 2-state code above 1.4 dB. Thus we can expect it to perform better asymptotically.

5.3. Influence of the Inner CPM System

We saw in section IV that the BER bound for SCCPM is made up of several combinations of error events in the inner CPM system, which depend on the outer CC. Therefore, in this section, we compare four different CPM systems, including MSK, concatenated through an $N = 8192$ interleaver with the same outer (7,5) CC. The systems certainly have different bandwidths, but we defer that comparison to Section 6. Their dominant error event combinations, as calculated in an earlier study [12], are shown in Table 1.

System 1 is the reference system (MSK, 2 states). System 2 is Example 2 (in Section 4) (binary 2RC, $h = \frac{3}{4}$, 8 states), which is supposed to perform better than system 1. System 3 is binary 3RC, $h = \frac{2}{3}$ (12 states), and system 4 is binary 3RC, $h = \frac{4}{5}$ (20 states). Studying the error events in Table 1 we find that system 3 should be slightly better than system 1, but worse than system 2, and that system 4 should perform about as well as system 2. These conclusions generally hold only above the divergence threshold of the BER bound, but the simulation results in Fig. 8 indicate that the relations stay the same in the low-SNR region as well, although the difference between systems 1 and 3 is larger than expected. Note that system 1 has the largest $\mathcal{D}_{\min}$, but it performs worst of these four systems.

Figure 9 shows simulation results from Ref. 36 for some SCCPM systems with $M = 4$, 8, and 16, together with system 1. The rate of the CC is adapted accordingly, such that $R_{cc} = \frac{1}{2}$, $\frac{2}{3}$, and $\frac{3}{4}$, respectively. In order to compare the systems at equal encoder delays (4096 information bits), interleaver sizes of $N = 8192$, 6144, and 5460 bits were chosen. As can be seen, increasing h and M gives improved performance, although this need not be true for arbitrary modulation indices. As demonstrated earlier [36], 3RC systems generally give worse BER performance — keeping everything else constant — than do their 2RC counterparts, but they consume less bandwidth. We will return to bandwidth in Section 6. In Fig. 9, the best BER performance is obtained by rate-$\frac{3}{4}$ coded $M = 16$ 2RC, $h = \frac{1}{2}$, which reaches BER $= 10^{-3}$ at $E_b/N_0 = 0.35$ dB. This is even better than the rate-$\frac{1}{2}$ Turbo codes with BPSK presented by Berrou et al. [4], but the bandwidth is slightly larger.

Concluding, for SCCPM systems utilizing the same outer code, performance below the divergence threshold is roughly in line with the error event analysis, and the convergence problem is not very pronounced. Still, methods of examining the convergence behavior [32–34] could give insights beyond the traditional union bound approach.

Table 1. Dominant Error Events for Some SCCPM Systems

System	$B_{\mathcal{D}} \sim N^{-5}$	$B_{\mathcal{D}} \sim N^{-4}$	$B_{\mathcal{D}} \sim N^{-3}$
1. (7,5) and binary 1REC, $h = \frac{1}{2}$ (MSK)	—	—	$B_6 \sim 480N^{-3}$
2. (7,5) and binary 2RC, $h = \frac{3}{4}$	$B_{5.21} \sim 2880N^{-5}$	$B_{6.59} \sim 1440N^{-4}$	$B_{7.97} \sim 480N^{-3}$
3. (7,5) and binary 3RC, $h = \frac{2}{3}$	—	$B_{3.55} \sim 120N^{-4}$	$B_{6.06} \sim 60N^{-3}$
4. (7,5) and binary 3RC, $h = \frac{4}{5}$	—	$B_{5.54} \sim 120N^{-4}$	—

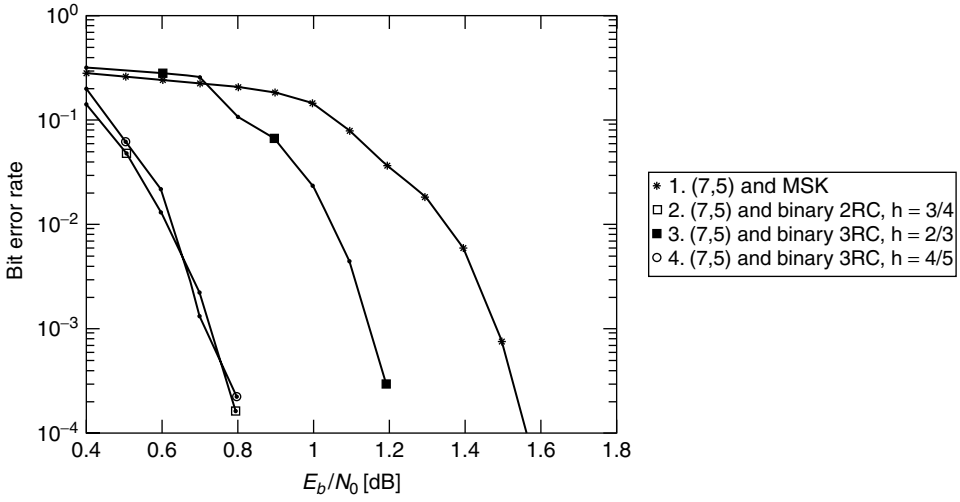


Figure 8. Simulation results for four different inner CPM systems with the (7,5) CC, at a delay of 4096 information bits ($N = 8192$). See Fig. 11 for a bandwidth comparison of these systems.

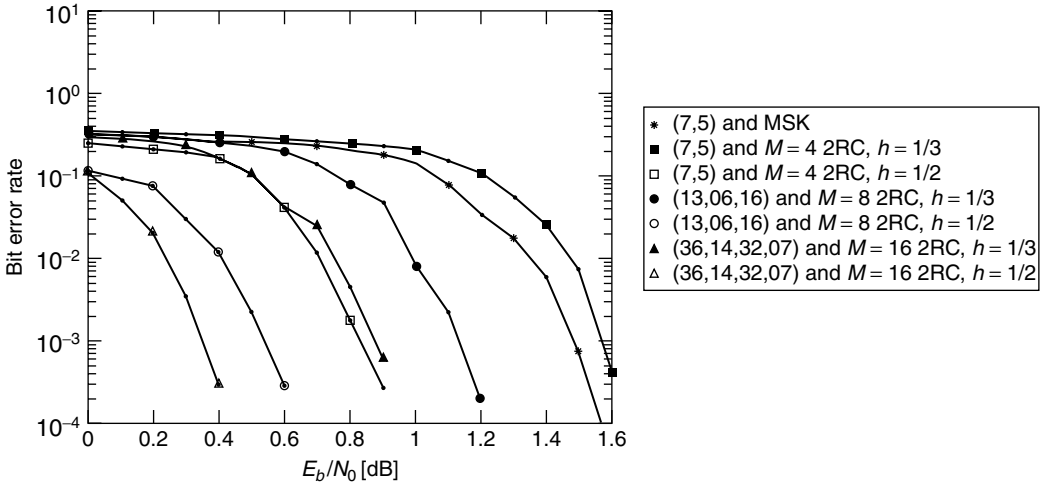


Figure 9. Simulation results for some $M = 4$, 8, and 16 SCCPM systems with rates $\frac{1}{2}$, $\frac{2}{3}$, and $\frac{3}{4}$ outer CCs, respectively, and interleaver sizes $N = 8192$, 6144, and 5460. See Fig. 11 for a bandwidth comparison of some of these systems.

5.4. Influence of Inner Input Weight 1 Error Events

It was stated in Section 4 that if the inner CPM system has input weight 1 error events, the interleaver gain would not be assured for all SNRs. We investigate this phenomenon by considering quaternary 1REC, $h = \frac{1}{3}$ with natural and gray mapping (3 states) together with the (2,3) and (7,5) outer CCs (2 and 4 states, respectively). Figure 10 shows BER simulation results with an $N = 8192$ interleaver. Clearly, the systems with Gray mapping — and thus with input weight 1 error events — do not provide any spectral thinning caused by the random interleaver; hence their BER curves have a modest slope that is due only to the distance profile in Eq. (6). Notice that for the 4-state code, gray mapping still gives better performance down to a BER of 10^{-5}.

6. POWER SPECTRAL DENSITY AND BANDWIDTH

Bandwidth is at least as important as BER performance in many communication systems. The demand for higher data rates in wireless data communications and increased capacity in digital mobile telephony networks is growing constantly. Spectrum reuse in the form of small-cell networks can meet these requirements to a certain extent, but bandwidth-efficient coding and modulation is nevertheless essential. In this section, we evaluate the bandwidth efficiency of SCCPM in general, and compare it with power efficiency (BER) for some selected examples.

For a coded CPM system without interleaving, Ho and McLane [20] have calculated the true power spectral density. However, since random bit interleaving is used in SCCPM, an accurate estimate (see also Ref. 17 and 20) of the baseband power spectral density $S(f)$ is

$$S_{\text{coded}}(f) \approx R_{cc} S_{\text{uncoded}}(R_{cc} f) \tag{10}$$

The power spectral density of the carrier-modulated SCCPM signal is given by

$$P_c(f) = \frac{P}{2}\left[S_{\text{coded}}(f - f_0) + S_{\text{coded}}(-f - f_0)\right] \tag{11}$$

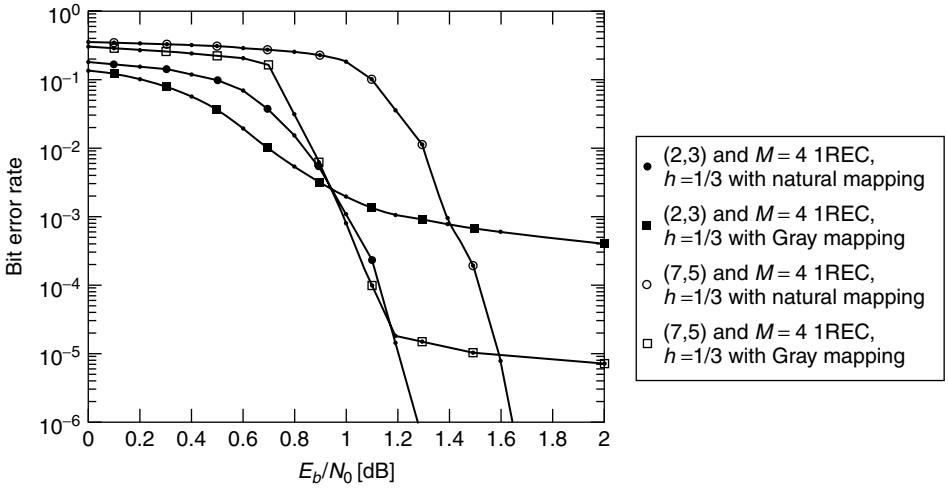


Figure 10. The effect of input weight 1 error events in the inner CPM system, $N = 8192$. Gray mapping yields such events; natural mapping does not.

where $P = E/T$ is the average transmitted signal power. Similarly, the normalized double-sided bandwidth of the coded CPM system $2B_{\text{coded}}T_b$ can be expressed in terms of the corresponding quantity of the uncoded CPM system, $2B_{\text{uncoded}}T_b$:

$$2B_{\text{coded}}T_b \approx \frac{2B_{\text{uncoded}}T_b}{R_{cc}} \tag{12}$$

Note that the rate of the CC is included in the normalized bandwidth in Eq. (12). The bandwidths are defined as 99% in-band power, equivalent with the −20 dB level in the fractional out-of-band power function [2,17]. The uncoded bandwidth of a CPM system can be calculated using the general method described in a previous publication [37].

The normalized bandwidths (bandwidth efficiencies) of some SCCPM systems are compared with their error performance in Fig. 11. Interleaver sizes corresponding to an input delay of 4096 information bits were used, namely, $N \approx 4096/R_{cc}$. Error performance is measured in terms of the required E_b/N_0 for a simulated BER of 10^{-3}, and plotted against the normalized double-sided bandwidth $2B_{\text{coded}}T_b$. Lines are used to interconnect points for similar systems but with different modulation indices. Notice that these lines do not provide any information on modulation indices in between the points. As can be seen, over the whole bandwidth range shown in Fig. 11, the lowest E_b/N_0 among these systems is obtained for the rate-$\frac{2}{3}$, (13,06,16) CC, a 6144-bit interleaver, and $M = 8$ 2RC. At $2B_{\text{coded}}T_b \approx 2.0$, it requires only 0.5 dB, and at $2B_{\text{coded}}T_b \approx 1.15$, it requires roughly 1.6 dB. The latter can be compared to uncoded MSK ($2B_{\text{coded}}T_b \approx 1.20$), which requires 7.3 dB for a BER of 10^{-3}. Thus the gain over uncoded MSK is around 5.7 dB. For smaller BERs, the gain increases because the SCCPM system has a much steeper BER slope. However, if a low-complexity system is the goal, the quaternary 1REC systems may provide a good tradeoff. The rate-$\frac{3}{4}$-coded $M = 16$ 2RC systems are not shown here because they are less power/bandwidth-efficient than are the rate-$\frac{2}{3}$-coded $M = 8$ 2RC systems, as demonstrated earlier [36].

How does SCCPM compare with a typical Turbo code with BPSK? The answer to this question depends on the bandwidth required to transmit one BPSK symbol. While theoretically, a normalized bandwidth of 1 is sufficient if signals extend over infinite time, we here assume a 99% bandwidth of 1.2 as being more practical. Then, a rate-$\frac{1}{2}$ Turbo code will have $2B_{\text{coded}}T_b = 2.4$, while typically yielding a BER of 10^{-3} at roughly 0.8 dB for an input delay of 4096 bits (Berrou et al. [4] obtained 0.6 dB with 65536 bits). This is slightly inferior to the rate-$\frac{2}{3}$-coded $M = 8$ 2RC, $h = \frac{1}{2}$ SCCPM system, which requires around 0.55 dB with $2B_{\text{coded}}T_b \approx 2$.

Finally, we compare some of the coded, noninterleaved CPM systems with MLSD, investigated by Lindell [17]. These systems have large asymptotic gains (upto dB over uncoded MSK), but here we still use simulated BER $= 10^{-3}$ as the performance measure. As can be seen from Fig. 11, SCCPM is able to provide gains of 1.5–2.0 dB at this BER level, although with increased receiver complexity. At a lower BER, the gain would increase even further because the SCCPM systems generally have a much steeper slope than do the noninterleaved systems. For example, rate-$\frac{3}{4}$ and $M = 16$ 1REC, $h = \frac{2}{13}$ with MLSD requires 5.7 dB for BER $= 10^{-5}$, while rate-$\frac{2}{3}$ and $M = 8$ 2RC, $h = \frac{1}{4}$ with a 6144-bit interleaver and iterative decoding requires 1.9 dB, yielding a gain of 3.8 dB.

7. CONCLUSIONS

SCCPM with iterative decoding is an exciting new coding and modulation technique, merging the principles of both coded CPM and Turbo codes. With the simple insertion of a random interleaver between the outer code and the CPM system, dramatic performance gains

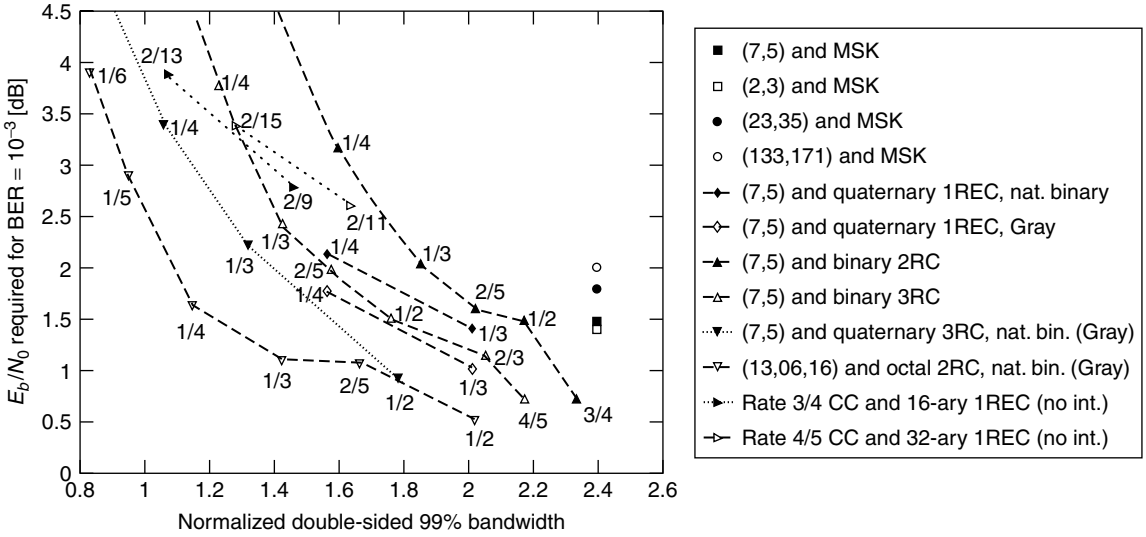


Figure 11. Bandwidth/performance comparison for an input delay of 4096 bits and BER $= 10^{-3}$. The numbers indicate modulation indices. "Nat. bin. (Gray)" refers to gray mapping only for modulation indices that yield input weight 1 error events with natural mapping.

become feasible. Since MLSD (Viterbi decoding) is practically unreasonable, this requires a rather different iterative decoder that consists of SISO algorithms for the constituent code and modulation, passing APPs between each other. This decoder is very similar to that used in SCCs with memoryless modulation.

SCCPM can be analyzed through a transfer function bound to the BER, using a number of assumptions. This includes the averaging over all possible interleavers (the uniform interleaver assumption), the use of an MLSD receiver instead of the iterative decoder, and the assumption of an SNR above 3 dB, which is the divergence threshold of the bound. Still, the transfer function bound gives theoretical insights such as the following:

1. Similar to Turbo codes and SCCs, SCCPM is capable of providing interleaver gain, the order of which is determined by the free distance of the outer CC.
2. Several error events in the inner CPM system contribute to the overall performance; hence it is difficult to give general design guidelines for the inner CPM system.
3. However, CPM systems with input weight 1 error events should be avoided since they do not give interleaver gain. This is always possible by a different choice of mapping.

Simulation results for a variety of SCCPM systems indicate that, in addition to these three conclusions, the convergence of the iterative decoder also affects performance in the low-SNR (waterfall) region. For unequal outer CCs with similar CPM systems, performance in the waterfall region actually deteriorates with increasing free distance of the outer code. This is consistent with conclusions from SCCs. On the other hand, for similar outer CCs with different CPM systems, the transfer function bound analysis appears to give satisfactory conclusions. Also, we demonstrated that CPM systems with input weight 1 error events indeed fail to provide Turbolike performance. We also demonstrated a number of higher-order SCCPM systems with $M = 4$, 8, and 16. These systems both have attractive BER performance and bandwidth efficiency. Comparing power and bandwidth efficiency, it is seen that the $M = 8$ 2RC systems are the most efficient as known today. These systems can provide gains of almost 4 dB at BER $= 10^{-5}$, at the same bandwidth, compared to the best known coded CPM systems without interleaving.

In conclusion, wherever a constant envelope of the transmitted signal is desired, SCCPM should be able to provide appealing power and bandwidth efficiencies. As of today, the decoding complexity may seem high, but we are strongly convinced that in a near future, this will not pose any major obstacles. Of course, much work remains to be done on implementation issues such as timing and carrier phase recovery, reduced sampling rates, and quantization to fix-point representation.

BIOGRAPHIES

Pär Moqvist received his M.S. and Lic.Eng. degrees in electrical engineering from Chalmers University of Technology, Göteborg, Sweden, in 1993 and 1999, respectively. In 1994, he was awarded the John Ericsson Medal for outstanding scholarship. He joined Ericsson AB in 1993 as a radio system engineer, working on the design of digital radio modems for land-mobile applications. Since 1997 he has been a Ph.D. student with the Telecommunication Theory Group at Chalmers University. His research is in the general area of signaling and detection for digital communication systems, with a current focus on iterative decoding methods.

Tor M. Aulin received his M.S. degree in electrical engineering from the University of Lund, Lund, Sweden, in 1974 and the Dr. Techn. (Ph.D.) from the Institute of Telecommunication Theory, University of Lund, Göteborg, Sweden in November 1979. He became a docent there in 1981 and he was also a visiting scientist at the ECSE Department at Rensselaer Polytechnic Institute, Troy, New York. One year was spent at the European Space Agency (ESA, ESTEC) as an ESA research fellow. In 1983, he became a research professor (docent) in information theory at Chalmers University of Technology, Göteborg, Sweden. In 1991, he formed the Telecommunication Theory group there and became a docent in computer engineering in 1995. During 1995 he was a visiting fellow at the telecommunications engineering department, Australian National University, Canberra, ACT, Australia. Some of Dr. Aulin's research interests are communication theory, combined modulation/coding strategies (such as CPM and TCM), analysis of general sequence detection strategies and digital radio channel characterization. Dr. Aulin has authored the book *Digital Phase Modulation*, Plenum Press 1986. He is a fellow of the IEEE and an editor for *IEEE Transactions on Communications* in the area of communication theory and coding. In 1997, he was awarded the prestigious Senior Individual Grant, handed over by the Prime Minister of Sweden.

BIBLIOGRAPHY

1. T. Aulin, *CPM—a Power and Bandwidth Efficient Digital Constant Envelope Modulation Scheme*, Ph.D. dissertation, Univ. Lund, Lund, Sweden, 1979.
2. J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.
3. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 623–656 (1948).
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo codes, *Proc. Int. Conf. Communications* (*ICC'93*), 1993, pp. 1064–1070.
5. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* **44**: 1261–1271 (1996).
6. S. Benedetto and G. Montorsi, Unveiling turbo codes: Some results on parallel concatenated coding schemes, *IEEE Trans. Inform. Theory* **42**: 409–428 (1996).
7. S. Benedetto and G. Montorsi, Design of parallel concatenated convolutional codes, *IEEE Trans. Commun.* **44**: 591–600 (1996).
8. S. Benedetto and G. Montorsi, Serial concatenation of block and convolutional codes, *Electron. Lett.* **32**: 887–888 (1996).

9. S. Benedetto et al., Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding, *IEEE Trans. Inform. Theory* **44**: 909–926 (1998).
10. *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Multiplexing and Channel Coding* (FDD) (Release 1999), (3GPP TS 25.212 V3.5.0), 2000.
11. *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, TIA/EIA/IS-2000.2-A, 2000.
12. P. Moqvist, *Serially Concatenated Systems: An Iterative Decoding Approach with Application to Continuous Phase Modulation*, thesis, Chalmers Univ. Technology, Göteborg, Sweden, 1999; *http://www.ce.chalmers.se/TCT/*.
13. C. Brutel and J. Boutros, Serial concatenation of interleaved convolutional codes and M-ary continuous phase modulations, *Ann. Telecommun.* **54**: 235–242 (1999).
14. K. R. Narayanan and G. L. Stüber, Performance of trellis coded CPM with iterative demodulation and decoding, *Proc. Global Telecommunications Conf.* (*GLOBECOM'99*), 1999, pp. 2346–2351.
15. E. Biglieri et al., *Introduction to Trellis-Coded Modulation with Applications*, Macmillan, New York, 1991.
16. S. Benedetto et al., Serial concatenated trellis coded modulation with iterative decoding, *Proc. Int. Symp. Information Theory* (*ISIT'97*), 1997, p. 8.
17. G. Lindell, *On Coded Continuous Phase Modulation*, Ph.D. dissertation, Univ. Lund, Lund, Sweden, 1985.
18. S. V. Pizzi and S. G. Wilson, Convolutional coding combined with continuous phase modulation, *IEEE Trans. Commun.* **33**: 20–29 (1985).
19. F. Morales-Moreno and S. Pasupathy, Structure, optimization and realization of FFSK trellis codes, *IEEE Trans. Inform. Theory* **34**: 730–741 (1988).
20. P. Ho and P. J. McLane, The power spectral density of digital continuous phase modulation with correlated data symbols, *IEE Proc.* **133**(Pt. F): 95–114 (1986).
21. F. Morales-Moreno, W. Holubowicz, and S. Pasupathy, Optimization of trellis coded TFM via matched codes, *IEEE Trans. Commun.* **42**: 1586–1594 (1994).
22. R. H.-H. Yang and D. P. Taylor, Trellis-coded continuous-phase frequency-shift keying with ring convolutional codes, *IEEE Trans. Inform. Theory* **40**: 1057–1067 (1994).
23. B. Rimoldi and Q. Li, Coded continuous phase modulation using ring convolutional codes, *IEEE Trans. Commun.* **43**: 2714–2720 (1995).
24. G. Karam, I. Fernandez, and V. Paxal, New coded 8-ary CPFSK schemes, *Proc. Int. Conf. Communications,* (*ICC'93*), 1993, pp. 1059–1063.
25. P. Moqvist and T. Aulin, Trellis termination in CPM, *Electron. Lett.* **36**: 1940–1941 (2000).
26. S. Benedetto et al., A soft-input soft-output APP module for iterative decoding of concatenated codes, *IEEE Commun. Lett.* **1**: 22–24 (1997).
27. P. Moqvist and T. Aulin, Certain aspects on MAP algorithms for turbo codes, *Proc. 17th Swedish Conf. Radio Science and Communication* (*RVK'99*), 1999, pp. 623–625.
28. L. R. Bahl et al., Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**: 284–287 (1974).
29. B. Rimoldi, A decomposition approach to CPM, *IEEE Trans. Inform. Theory* **34**: 260–270 (1988).
30. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.
31. M. Breiling, S. Peeters, and J. Huber, Interleaver design using backtracking and spreading methods, *Proc. Int. Symp. Information Theory* (*ISIT'2000*), 2000, pp. 451.
32. H. El Gamal, *On the Theory and Application of Space-Time and Graph Based Codes*, Univ. Maryland, College Park, MD, 1999.
33. S. ten Brink, Convergence of iterative decoding, *Electron. Lett.* **35**: 806–808 (1999).
34. D. Divsalar, S. Dolinar, and F. Pollara, Iterative turbo decoder analysis based on Gaussian density evolution, *Proc. IEEE Military Communications Conf.* (*MILCOM 2000*), 2000, pp. 202–208.
35. D. Divsalar et al., Transfer Function Bounds on the Performance of Turbo Codes, TDA Progress Report, Jet Propulsion Lab., Pasadena, CA, 42-122, 1995, pp. 44–55.
36. P. Moqvist and T. Aulin, Power and bandwidth efficient serially concatenated CPM with iterative decoding, *Proc. IEEE Global Telecomm. Conf.* (*GLOBECOM'00*), 2000, pp. 790–794.
37. T. Aulin and C.-E. Sundberg, An easy way to calculate power spectra for digital FM, *IEE Proc.* **130**(Pt. F): 519–526 (1983).
38. P. Moqvist and T. Aulin, Convergence Analysis of SCCPM with iterative Decoding, *Proc. IEEE Global Telecomm. Conf.* (*GLOBECOM'01*), 2001, pp. 1048–1052.

SERVICES VIA MOBILITY PORTALS

DANIEL RALPH
CHRIS SHEPHARD
BTexact Technologies
Ipswich, Suffolk, United Kingdom

1. INTRODUCTION

Mobility portals look set to become the window through which the user will access a whole range of innovative services [4]. Wherever and whenever, the mobility portal will warn you, inform you or just entertain you.

There is a distinction between the mobility portal and the mobile portal, the concept of mobility extends to include terminal independence through user profiles, additional value add services such as "find me, follow me" call routing and integration with existing systems in the fixed network. The mobile portal however is being expressed as a cut-down version of existing Web-based applications such as news, weather, and email.

The explosion in use of Short Message Service (SMS) for sending text messages demonstrates a natural evolution path to WAP services on smartphone and personal digital assistant (PDA) devices. Support from many vendors has created an environment where the applications delivered through the mobile portal will increasingly substitute access through the fixed network.

However, although the Internet and mobile phones have independently been successful (see Figs. 1 and 2), this

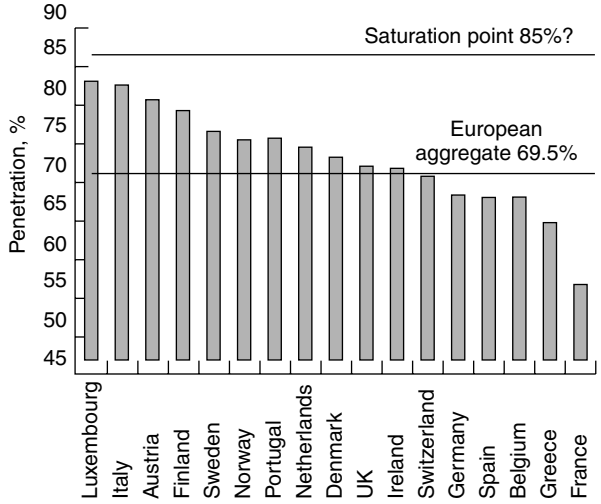


Figure 1. European mobile penetration by country, June 2001 (%) (*Source*: Morgan Stanley Research).

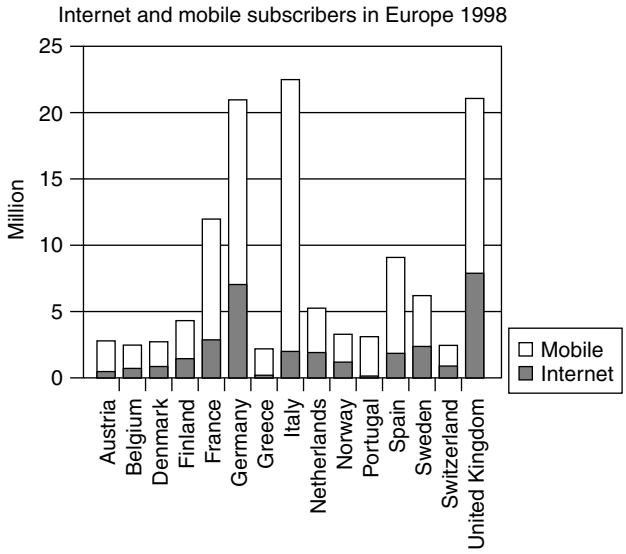


Figure 2. Internet and mobile subscribers in Europe, 1998. (*Source*: Dataquest, Mobile Communications International, Computer Industry Almanac [10]).

does not always mean that combining these technologies will also be successful. Evidence of this is clear in the combination of TV and phone technologies, both successful in their own right but the uptake of personal videophones has not happened. Mobile videophone is considered a killer application. However, it will become clear from this article that to achieve a high takeup is unlikely in the foreseeable future due to bandwidth limitations and device battery life. A more likely scenario is that a short videoclip downloaded in non–real time could be played back to the user to provide visual information.

In addition, users may subscribe to specialized information feeds for stock trading, weather, ski and snow conditions, and horoscopes. Finally, the services will offer specialized alerting through partners such as auction, travel sites, banking, and messaging providers.

2. THE RISE OF THE MOBILITY PORTAL

Without doubt the explosive growth rate of Internet users and the significance of accessing content through the use of Web-based portal services will continue. While this highlights the demand for content from the fixed network, it cannot be assumed that the same type or level of demand will be present in the mobile environment. The differences in network and device capability will require a different solution in providing content through the mobile portal.

In assessing why the mobility portal is important, it can be seen from looking across the value chain that a number of key elements are available to be exploited in delivering new mobility services:

- The existing infrastructure will support these services, in terms of both core IP networks and content availability. Whether this is from the customers Internet service provider or other content providers, it will enable the rise of the mobility portal.
- The prospect of "always on" packet-based services delivered over 2.5G and 3G networks such as GPRS and UMTS will enhance the user experience in the way content is delivered over the next-generation infrastructure provided by the mobile operator.
- The first-generation consumer equipment is readily available and supports basic web access and WAP services. This is in the form of smartphone and PDA devices.
- In the short term, existing content can be redeveloped for delivery to the smartphone or PDA device, this will enable the rapid deployment of mobile services. Although this will only be through provision of existing generic Web portal services on the mobile device, this will include search engines, personalization, snap-shot text information and basic messaging.

It is the very issue of content and what services the user will require that will determine the successful uptake of service access through the mobility portal. Figure 3 illustrates some of the anticipated benefits from a wireless portal.

Questions have to be answered to determine the customer's needs when mobile: What do they want? When? Why? Providers of mobility portals will need to assess customer requirements and actual usage to develop context awareness so as to present the most relevant content. If the network knows the weather conditions, user location, time, and whom the user is accompanied by, then this gives the network an opportunity to provide information ahead of the user requirement for this information, further developing the mobile device as a lifestyle tool [12].

Future applications need to develop the unique attributes of the mobile device. The opportunity for the network to provide location-based services, and the personal nature of the mobile device lend it to acting as a payment device.

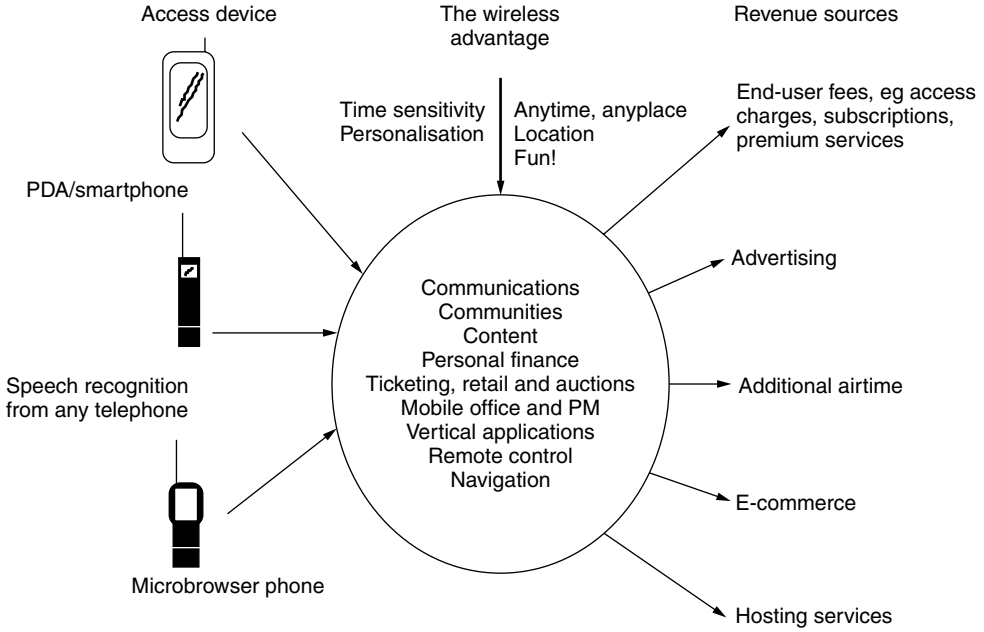


Figure 3. Benefits of a wireless portal.

3. INSIDE THE MACHINE

The current UK digital mobile network has evolved from the first-generation analog network and is based on the Global System for Mobile communications (GSM).

Network technologies are evolving toward the 3G mobile vision, this will require an evolution through extensions to the existing 2G mobile network technologies. In the near future this will involve

- GSM (2G) component technologies:
- SMS and cell broadcast
- GPRS (general packet Radio Service)
- EDGE (enhanced data rates for GSM Evolution)
- UMTS [3G] (Universal Mobile Telecommunications System)
- Bluetooth
- GPS and other location-based techniques (Mobile Positioning Protocol)
- Mobile agents and intelligent networks

The next-generation mobile system will rely heavily on emerging application technologies such as

- Wireless Application Protocol (WAP)
- XML/XSL (eXtensible Markup Language)
- JAVA Technology (Java 2 Micro Edition)
- SIM Application Toolkit
- Lightweight Efficient Application Protocol (LEAP) [1]
- Compact HTML
- XHTML

It is clear that the Internet will play a pivotal role, requiring increasing quality of service (QoS) and allowing better integration of applications across different terminals and bearers, probably using APIs to provide access to functionality.

The importance of voice must not be underestimated as a mechanism for controlling services by natural language commands, such as adding an appointment to your calendar or more simply as an alternative to typing a response to an email. Voice browsing of a mobility portal will be standardized through the use of VoiceXML and provide a development environment to deliver powerful niche applications [2].

Before service providers can offer content or applications using mobility portals, a number of issues require further consideration.

3.1. Wireless Application Protocol

Wireless Application Protocol (WAP) is a specification for a set of communication protocols to standardize the way that wireless devices, such as cellular telephones and radio transceivers, can be used for Internet access, including email, the World Wide Web, newsgroups, and Internet Relay Chat (IRC) [3].

While Internet access has been possible in the past, different manufacturers have used different technologies. In the future, devices and services that use WAP will be able to interoperate (see Fig. 4).

The WAP layers are

- Wireless application environment (WAE)
- Wireless session layer (WSP)
- Wireless transport layer (WTP)
- Wireless transport layer security (WTLS)
- Wireless datagram protocol (WDP)

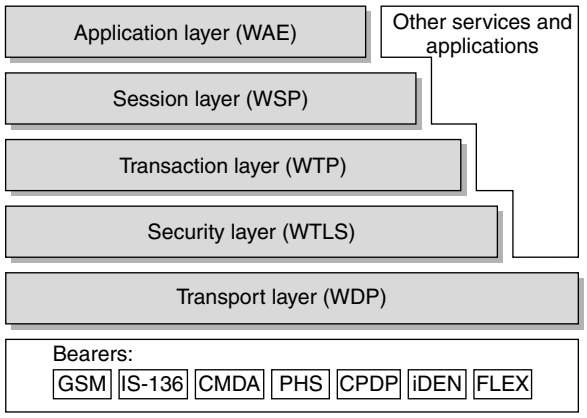


Figure 4. The WAP protocol stack.

Following independent development, four companies proposed the WAP standard: Ericsson, Motorola, Nokia, and Openwave (previously known as *phone.com*).

The use of transcoding engines (described as the HTML filter in Fig. 5) is being proposed as a "quick win" in providing services from existing Web content to the first generation of wireless terminals. This has met with limited success due to the complexity present in much of the multimedia-rich Web content. The use of JavaScript, Frames, and imagemaps causes in many cases a failure to transcode the html into a suitable WML page that can be displayed on a smartphone.

3.2. i-mode

The deployment of WAP in Europe must be contrasted with the developments in Japan, where the i-mode service has significantly better functionality. The service uses a Compact Hypertext Markup Language (CHTML) and is provided over the Personal HandyPhone System (PHS), which is a packet-based "always on" service, operating at 9.6 kbps (kilobits per second). The use of CHTML permits color graphics to be displayed or 10 lines of text, it still has limitations similar to the Wireless Markup Language as used in WAP [11].

There are many reasons for the success of i-mode. It relies on open technology, allowing any Internet site to join in; content is written in a simplified version of HTML (CHTML); and no fees are charged for placement on the i-mode portal. The services are positioned as a unique mobile service and not as "the Internet on your phone," thereby avoiding unfavorable comparisons with service from the fixed Internet. The low penetration of PCs in Japanese homes probably also helps control user expectations. Finally, NTT DoCoMo has concentrated on growing core revenues from airtime usage and has chosen not to support the service through advertising or transaction revenue.

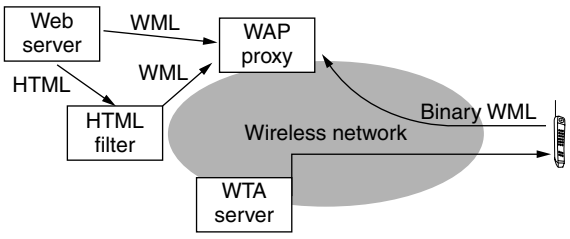


Figure 5. Example WAP network.

Mobile Internet services to compete with i-mode have now been introduced by rival operators. However, they have thus far failed to attract significant numbers of subscribers or content providers, despite offering greater bandwidth.

By being first to market, i-mode may now have an important head start, and competitors offering WAP based services cannot tap into the wealth of available i-mode content because of format incompatibilities.

i-mode is a brand name for NTT DoCoMo's Internet service. The technology and service offerings behind i-mode will continue to evolve to take advantage of improving network capabilities, including the introduction of 3rd generation mobile networks. In addition, DoCoMo has plans to roll out i-mode service in other countries, particularly in mobile networks where NTT owns a share. One example is the cooperation with Hutchinson in Hong Kong. At the same time, DoCoMo is active in the *WAP Forum*, *World Wide Web Consortium* (W3C), and Internet Engineering Task Force (IETF) whose standards, in combination, will map the way forward for a globally compatible mobile Internet.

The deployment of global mobility portals are essential to support the global traveler, they need to support the roaming capabilities of GPRS and GSM; and remove the added complexity for the user to change profiles to connect to the same services in different countries. The geographic distribution of network elements and the seamless access to services across different countries will strengthen the mobile operator's position as a service provider.

3.3. Terminals

The look and feel of the mobile device will have an impact on the user's perception of services, the usability of the device will have an impact on the usability of any service delivered over that device. Regardless of the content the portal provides, the terminal usability will prove significant in the promotion of services through mobility portals.

The types of mobile device range from the smartphone complete with NaviRoller to a pen based PDA device (see Figs. 6 and 7) and upto a subnotebook with keyboard and 1024×480 pixel display. Each of these devices has different attributes related to

- Screen size (resolution)
- Input method, such as through keyboard, touch screen, handwriting recognition
- Presence or absence of a speaker and microphone

The multiple methods for input and display are compounded by the variable and often low bit rate of mobile networks. In order to provide suitable content, device characteristics will need to be available to the service provider to allow appropriate formatting and display to

Figure 6. D503i mobile phone.

Figure 7. Next-generation PDA device with pen and speech based input, integral GPS and GPRS connectivity [9].

the range of devices outlined. The W3C have a working group on device capability and profile called Composite Capabilities/Preference Profiles (CC/PP) that is driving standards efforts to implement device profiles to enable content providers to determine device capability.

Further proposals from the Salutation consortium [6], Jini and Universal PnP (UpnP) [5] are intended to deliver service discovery and device recognition capabilities to enable the network to seek the services that the user requires and that are suitable for the terminal capability. For instance, the user may require a print service, but have available only an IrDa port with which to communicate with the printer.

The limitations of device capabilities in terms of storage and access to information will increase the need for additional support from the network. This may be provided through network based backup and synchronization services or server processing for voice recognition to authenticate to a service.

3.4. Addressing

It has been a concern for some time that the lack of IPv4 address space would prevent connections to the Internet. This problem will only become worsen with an explosion of mobile devices that have an "always on" connection to the Internet. Ultimately the proposed migration to IPv6 with its 128-bit address scheme will remove this problem with available address space for devices connected to fixed or mobile networks. The size of the IPv6 address range will allow 10^{20} addresses (or thousands of IPv6 address for every square inch of the earth.

Each mobile device that a user carries or interacts with will be uniquely identified by an IPv6 address and may well communicate to the user using Bluetooth devices connected within one's "personal bubble" or to an external network using UMTS. It is estimated that 1 billion Bluetooth addresses will be in use by 2005 [7].

The home environment is where a significant portion of this addressing will be required; in particular, the ability for devices around the home to communicate, to share information on your whereabouts, and to anticipate the family's needs. This could be extended to the fridge ordering fresh milk from the supermarket on your arrival home from holiday.

3.5. Billing

Getting information to the handset of a user is the simple part. Devising a mechanism that meets the commercial aspirations of all the players involved in making this happen is much harder. Many network operators will partner with a publisher to manage the day-to-day operations of their portal. Service providers will want recompense for the information provided.

Even in the simple case of portal provision, a number of revenue models have been developed to describe the different ways in which the value can be expressed. For example, a simple revenue share might be appropriate for a traffic-congestion report, the operator charging the user "per click" for information and passing on a share to the information provider.

Initially, WAP services are expected to be expensive to use since the tendency is to be online for a long circuit-switched data (CSD) call since features such as interactivity and selection of more information are used by the end user. With the introduction of GPRS handset the billing model is based on the number of packets sent or received by the user.

In contrast, i-mode uses an efficient micro billing system via the mobile phone bill, which makes it easy for subscribers to pay for value added services and premium sites and is also attractive for site owners, to sell information to users.

3.6. Security

There are several areas of WAP security that will need additional development to ensure the end-to-end security requirements for transactions involving payments or

instructions requiring nonrepudiation. These include what is being called "premature encryption endpoint." The WAP gateway acts as a proxy and decrypts the incoming WTLS session and encrypts as SSL3.0 for the outgoing connection to the Origin server.

The WAP specification 1.1 supports the use of certificates in a way similar to X.509 and encryption is supported at 56 and 128 bits. In the future the WAP specification will be extended to include the wireless identity module (WIM), which may be a replacement for your SIM card containing a personal digital signature to identify and authenticate the User. The addition of a crypto library will enable the developer of WAP services access to a range of functions that allow the User to digitally sign and encrypt a message before it is sent from the mobile device.

3.7. Interoperability

The provision of mobile data services in the past has always been using proprietary technologies [8]. The standardization of WAP allows developers and services to be less restricted in their range of mobile networks or mobile devices. The first release of gateway and device products have the inevitable problems associated with new technology and significantly more testing will be required between different vendors products. This is an area where the service provider can add value in providing an integrated service to an agreed level of quality and compatibility with other services.

3.8. Quality of Service

The issue of quality of service (QoS) has plagued IP networks since it became commercialized. Several IETF standards have been proposed to satisfy the user requirements, these include Resource Reservation Protocol (RSVP) with IETF Integrated Services (IntServ), Differentiated Services (Diffserv), and multiprotocol label switching (MPLS).

The mobile network suffers from QoS issues such as dropped connections, lack of resources, and reduced bandwidth due to too many users sharing the available capacity. These factors will significantly restrict the availability of mobile services that compete directly with fixed network services, such as videoconferencing or high-speed file transfer.

As users are demanding higher levels of QoS from IP networks for fixed services delivered on these networks, so, too, will the mobile user. The demand for real-time services may be sufficiently low to cause less difficulty, videostreaming can use buffering to delay the transmission. Videoconferencing cannot accept high network latency, but it may be that this is not the "killer application."

Improved quality of service will be available with the advent of 3G networks. The rollout plans for many 3G operators are progressing at a rapid pace, with statements from several outlining available data rates between 64 and 384 kbps on initial launch. With these achieving commercial service in the near future, NTT DoCoMo began trials over their installed 3G network in July 2001.

4. CONCLUSION

Future developments of the mobility portal must embrace the entire range of existing web based content if it is to succeed as the default method of access to the network. This will include all applications from gaming to mobile commerce.

Services should be designed so the relevant elements of a service are available over to the particular client device, the requirement for the network to recognize the device capability and network connection characteristics.

There will be no single "killer application" for the mobile device, although its unique attributes, including location positioning and personal nature, provide an opportunity for the mobile device to become the portal through which the user interacts whilst "on the move."

It is highly likely that WAP will be superseded by evolution of existing Internet standards, the requirement for legacy support will remain for WAP enabled mobile phone devices.

The reasons for the limited lifetime of WAP are due to

Low bandwidth
High latency
Low-resolution monochrome displays
Dropped calls and other quality of service issues
Low device processing power

All of these factors are not limitations of WAP; however, once they have been improved or overcome, the tendency will be to use existing Internet standards end to end, as opposed to the WAP architecture of creating a wireless version of these standards. Even if successful, once all devices and networks support WAP, it will cease to differentiate and therefore will not reduce subscriber churn. However, it will provide the user with convenient access to services whenever and wherever required.

With so many heavyweight carriers, equipment manufacturers and software and applications developers backing the Wireless Application Forum, a momentum is being generated that will drive WAP-based equipment and services forward into the future.

Despite the drawbacks discussed in this article, WAP certainly seems to be shaping up to play a major role in facilitating the brave new world of the personalized mobility portal. WAP is a powerful tool. It enables "anytime, anywhere" connectivity to a wealth of information, whether for leisure or business applications.

It is impossible to predict who will win between WAP and i-mode; clearly, DoCoMo has a significantly larger market share at the time of writing. However, if XML becomes the dominant content standard, there may be situations where both standards are absorbed into existing Internet protocols.

This article has demonstrated through the discussion of mobility portals some of the problems involved in providing content, applications, and filtering information to a new generation of wireless devices.

Acknowledgments
The author would like to thank the Institute of Electrical Engineers (IEE) for reproduction of this article.

BIOGRAPHIES

Daniel Ralph joined BTexact Technologies in 1996 following a period working in the biotechnology industry. Since joining BT he has worked on projects as diverse as the deployment of the trial global VoIP network for Concert and the design and implementation of the call processing engine "powering" the network intelligence platform. He is now responsible for a number of technical developments within the arena of mobile Internet technologies. He is a graduate of the Open University, and has recently completed the BT MSc in telecommunications. He is a corporate member of the BCS.

Chris Shephard read theoretical physics at UEA and then University of Birmingham, United Kingdom. He joined BTexact Technologie in 1980 and helped to develop signaling, call control, and maintenance software for System X PSTN and ISDN switches. After two years with Siemens in Florida, where he was part of a large project designed to adapt their range of digital public switching systems (EWSD) to the USA regulatory and market requirements. He returned to BTexact Technologies where he worked initially on the development and delivery of local network switches (LA-30) and subsequently on the design and management of two European broadband ISDN projects. He then switched to the study of network centric computing and the use of thin client devices. He now works in the terminals and applications unit within Multimedia Applications where he leads a team developing XML applications.

BIBLIOGRAPHY

1. The Lightweight & Efficient Application Protocol (LEAP) Manifesto, *http://www.freeprotocols.org/leap*.
2. Voice portals: Ready for Prime Time? (12/7/00), *http://www.zdnet.com/anchordesk/stories/story/0,10738,2601898,00.html*.
3. WAP Forum, WAP Architecture Specification (WAPARCH), April 30, 1998; URL: *http://www.wapforum.org*.
4. European Wireless Portal Use to Boom (7/7/00), *http://www.allnetdevices.com/industry/market/2000/07/07/europeanwireless.html*.
5. UPnP, Jini, Salutation, *http://www.cswl.com/whiteppr/tech/upnp.html*.
6. Service discovery and management, *http://www.salutation.org*.
7. Bluetooth SIG Adds Protocol (10/7/00), *http://www.allnetdevices.com/developer/news/2000/07/10/bluetoothsig.html*.
8. B. Johnston, C. Fenton, and D. Gilliand, Mobile data services, *BT Technol. J.* **14**(3): 92–108 (July 1996) (*http://www.bt.com/bttj*).
9. Future phones, *http://www.futurefonezone.com*.
10. UK demographics from CIA publications, *http://www.odci.gov/cia/publications/factbook/geos/uk.html*.
11. i-mode, *http://imodelinks.com/desktop/faq.html*.
12. W. N. Schilit, N. I. Adams, and R. Want, Context-aware computing applications, *Proc. Workshop on Mobile Computing Systems and Applications*, IEEE Computer Society Press, Santa Cruz, CA, 1994, pp. 85–90.

FURTHER READING

Official Wireless Application Protocol 2.0, *The Complete Standard with Searchable CD-ROM*, Wireless Application Protocol Forum, Ltd.

Holma H. and A. Toskala, eds., *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, rev. ed., New York, Wiley, 2000.

SESSION INITIATION PROTOCOL (SIP)

HENNING SCHULZRINNE
Columbia University
New York, New York

1. OVERVIEW

The Session Initiation Protocol (SIP) is a *signaling* protocol for setting up, tearing down, and modifying multimedia sessions. These sessions can consist of any number of media, such as audio, video, and shared applications, and can be either unicast (point-to-point) or multicast. SIP does not transport media itself, but rather allows endpoints (Internet hosts) to coordinate the exchange of media.

This coordination function becomes necessary since endpoints, called *user agents* by SIP, need to discover one another even if they change their Internet address. Endpoints also need to agree on the characteristics of the session, such as the number and type of media streams, the codecs to be used for audio and video, the properties of shared applications, or the availability of quality of service (QoS) mechanisms. SIP is based on the notion that an endpoint identifier, a SIP URI, is associated with a person or telephone number. A person keeps the same SIP URI even if they change their IP address or use different devices, such as a office PC, wireless 3G phone, or traditional telephone. The ability to reach a person with different devices under the same name is sometimes referred to as *personal mobility* [1].

SIP supports five aspects of managing multimedia sessions:

User location — given a SIP URI, SIP protocol entities find all relevant end systems that are registered for that URI.

User availability — based on information in the SIP request, the called end systems indicate their willingness to participate in the session or may indicate alternate locations.

User capabilities — through protocol exchanges, both end systems learn about the media capabilities of the other side.

Session setup — if there is an intersecting set of capabilities, the called party alerts the user ("rings") and, if a human controls the endpoint, the session may get established.

Session management — SIP requests can transfer sessions to others, terminate sessions, and modify session parameters.

SIP itself does not describe the media content; rather, it relies on external protocols for this task. Currently, SIP applications almost exclusively use the Session Description Protocol (SDP) [2] for this purpose, although other mechanisms can be negotiated between session participants. A newer version of the Session Description Protocol, called SDPng, is under development [3], but it is unclear how fast it can replace SDP.

SIP messages typically, but not necessarily, travel through a sequence of intermediaries, called *SIP proxies*. These intermediaries serve many functions, from user location and providing other services to policing reachability to opening firewalls for media streams. Often, the initial message traverses all such proxies, while subsequent messages within a session are exchanged directly between the two user agents.

SIP can use a variety of network-layer and transport-layer protocols. SIP works equally well over IPv4 [4] and IPv6 [5]; the latter is particularly important for its application to next-generation wireless networks. Currently, operation has been specified for UDP [6] and TCP [7], as well as for SCTP [8,9]. For reliable transport protocols such as TCP and SCTP, SIP can also benefit from transport-layer hop-by-hop security between end systems and proxies as well as between proxies. A single SIP request may traverse a path consisting of a number of different protocols.

Unlike traditional signaling protocol approaches, SIP does not try to model services directly, but rather provide a set of interoperable building blocks that can be used to build common telephony services [10] as well as new services that are beyond the capabilities of the existing circuit-switched network.

While useful for conference establishment, the core SIP specification does not address conference control features such as floor control, specifically, the coordination of access to a shared resource. However, SIP events (Section 5.6) and possibly common remote procedure call mechanisms may be used for this purpose.

SIP is maintained by two working groups within the Internet Engineering Task Force (IETF) and is currently an Internet Proposed Standard [11]. It was published in March 1999 and reissued with corrections and enhancements in March 2002.

2. SIP USAGE

SIP was originally designed to be used with multicast multimedia conferences in the Internet, allowing participants to invite others to join the multicast group. This usage is still supported, but has been overtaken in practical importance by its use in voice over IP (VoIP) applications. Voice over IP applications using SIP can be roughly divided into four areas: enterprise, cable, landline carrier, and third-generation (3G) wireless networks. In enterprise networks, traditional analog or digital handsets are replaced by a mixture of Ethernet-connected IP phones, such as the ones pictured in Fig. 1, and software running on desktop PCs [12]. These devices then use SIP to set up internal calls or to reach a gateway operated by the enterprise, a PBX with an IP interface, or a third party.[1] This arrangement makes it easy to add devices to a local dialing plan regardless of their physical location.

Traditional landline carriers can use SIP as a rough replacement for their current backbone signaling protocols, such as ISUP. In such an arrangement, local or gateway switches offer the same analog or digital circuit-switched service to the carrier's customers, but instead of a dedicated signaling network, the carrier uses SIP to interconnect these switches. SIP has been extended to transport ISUP messages across SIP networks, so that midcall information that is relevant only to ISUP networks is not lost in the protocol translation across an ISUP-SIP-ISUP signaling path [13]. The set of specifications for facilitating the interworking of SIP and ISUP is called SIP-T [14]. This is not a different protocol, just a set of guidelines for interworking.

Cable carriers can also use SIP as the signaling protocol in the Distributed Call Signaling (DCS) version of the PacketCable specification [15]. A number of

[1] The latter arrangement is often called *IP centrex*.

Figure 1. SIP phones.

extensions [16–18] have been proposed to make DCS mimic traditional telephone behavior more closely.

Third-generation wireless networks, using both GSM-evolved and CDM2000-based wireless networks, are being standardized by the 3GGP and 3GPP2 consortia. Both have chosen SIP as the signaling protocol for multimedia sessions. For 3GPP, SIP is slated to appear in release 5 of their UMTS framework.

In addition to VoIP and multimedia session setup, SIP has also been proposed for signaling events, that is, asynchronous notifications of state changes. One of the first applications is instant messaging and presence (Section 5.6). SIP events are also useful to unify many of the features found in traditional telephone systems, such as voicemail notification, users joining and leaving conference calls, or the transmission of DTMF digits [19]. Even the use of SIP events to control home appliances have also been proposed [20].

3. RELATED PROTOCOLS

SIP is related to other signaling protocols such as H.323 [21,22] and ISUP [23]. Like these, it operates out-of-band, that is, using a two separate associations for the signaling information and the actual media data, such as voice. Unlike ISUP, SIP is designed for IP networks and for multimedia sessions. A full comparison with H.323 is beyond the scope of this article, but a number of papers offer perspectives [24–27]. Interworking between the two protocols is also possible [28].

As noted above, SIP uses the Session Description Protocol (SDP) to describe the characteristics of the multimedia streams. Typically, these multimedia sessions use RTP [29,30] to carry audio and video information across IP networks.

RTSP [31] is a related control protocol that sets up streaming media sessions. It shares some characteristics with SIP, such as the protocol format and the use of SDP, but supports the control of stored media, through requests such as pause and play, rather than setting up interactive sessions. It does not emphasize finding locations, as it is assumed that the location of multimedia objects is much more stable than those of people. RTSP can be used in a SIP-based telecommunication system for playing announcements, recording voicemail [32] and other so-called media server functionalities.

SIP can interact directly with Internet telephony gateways that translate IP voice packets into circuit-switched calls. Alternatively, these gateways, called *media gateways*, can be controlled by a media gateway controller (MGC), using a master–slave protocol such as MGCP [33] or MEGACO/H.248 [34]. The media gateway controller translates SIP requests into MGCP or H.248 requests, and vice versa. MGCs may also terminate ISUP or other PSTN signaling. MGCs are sometimes called *soft switches*, although the definition is not very precise. Unlike these media control protocols, SIP is a peer-to-peer protocol, where both endpoints are equal.

As indicated above, SIP can use telephone numbers [35] to reach destinations. Proxies can translate these numbers into SIP URIs, possibly via multiple translations, using the ENUM mechanism [36]. ENUM uses NAPTR DNS records [37] to translate a telephone number such as +1-917-555-1234 to a DNS name, here *4.3.2.1.5.5.5.7.1.9.1.enum.arpa*. This record then contains a SIP URI, for example, allowing subscribers to keep their existing telephone number as they migrate to a SIP-based telephone system.

4. PROTOCOL DESCRIPTION

4.1. Protocol Layers

SIP is an application-layer protocol, with several logical sublayers. The lowest layer describes how SIP requests and responses are encoded as messages (Section 4.4). The second layer is the transport layer, defining how clients send requests and receive responses (Section 4.3). Above this transport layer, the transaction layer deals with the notion of group of requests and responses that form a single SIP *transaction*, namely, a request, its retransmissions, and all provisional and final responses (Section 4.4). While all SIP elements (Section 4.2) share similar encoding, transport, and transaction layers, they are distinguished by their role-specific *core*.

4.2. SIP Elements

There are three principal SIP elements: user agents (UAs), proxies, and registrars. User agents can act as clients, initiating requests and receiving responses, and/or servers, receiving requests, while proxies always combine a client and a server functionality. Registrars receive only REGISTER requests and update bindings between stable names and shorter-term contact addresses (see Section 4.3). These roles are logical only, so that a single software implementation can act as a user agent and a proxy server for different requests.

Proxies can be *stateful* or *stateless*. A stateless proxy simply forwards requests and responses one by one, but does not have to concern itself with transactions or associating responses with requests. Stateful proxies are transaction-aware. Stateful proxies are needed if a single inbound request can generate multiple outbound requests going to different destinations ("forking"). Unlike a normal telephone switch, both types of proxies do not have to keep state for the duration of the call. Some proxies do keep call state, for example, for accounting purposes or to control a firewall, and are referred to as *call-stateful*. Naturally, user agents are always call-stateful. Reducing state in SIP network elements helps with scaling and improves robustness, as any host in a server farm, for example, can easily take over should one host fail (Section 4.3).

While not a different protocol element as such, it has become common to refer to back-to-back user agents (B2BUA) in creating services. B2BUAs can be thought of as two user agents where one receives a request and the other issues a related request. This arrangement can be useful for certain types of media-handling firewalls or to create services where two participants have SIP dialog with the call controller, but media are exchanged between them directly. This arrangement is called *third-party call control* [38].

4.3. Locating Users and Servers

SIP endpoints are identified by SIP URIs that are similar to email addresses. (Users may often be able to use their email address as a SIP URI.) SIP URIs can vary in specificity; they may identify a particular user at a host located at an IP address, or, more commonly, identify a user generically by name and domain. An example of the latter is *sip:alice@example.com*. Such a generic URI is then translated via one or more steps to zero or more host locations, possibly identified by SIP URIs. For example, a call to *sip:alice@example.com* may reach her at *sip:alice@128.59.16.1* and *sip:asmith592@aol.com*. SIP URIs with the "sips" scheme indicate that the destination insists on being contacted via transport-layer security (TLS) [39].

The externally visible address of a person or other endpoint is known as the *address-of-record* (AOR). A single person or SIP telephone may have several such AORs. The AOR is bound to any number of contact addresses using the SIP REGISTER method. These contact addresses are typically SIP URIs, but can be any URI, commonly including mail to URIs and http URIs. Each of the user's hosts sends periodic REGISTER request to the registrar identified by the AOR, refreshing the address bindings. For example, for the AOR *sip:alice@example.com*, all of Alice's SIP-enabled communications devices would send updates to the registrar at *example.com*. SIP registrars are one example of a SIP location service that is then used by proxies for the domain to route calls for the AOR. While less common, implementors can choose any other mechanism to populate the location service. For example, a Webpage might allow manual updates of telephone numbers or other non-SIP URIs.

SIP requests may also carry telephone URLs [40] identifying telephone numbers. An example of such a URL describing a telephone terminal in New Jersey is tel:+1-201-555-1234. However, since this URL does not identify an Internet host, a SIP entity needs to translate it to a SIP URI, either a user name or a telephone number at a particular destination domain. In the example, a SIP proxy may translate the number to *12015551234@bigcarrier.com* if it wants the call to be routed to a gateway operated by *bigcarrier.com*. The actual routing may be determined by routing protocols such as TRIP [41,42] that distribute information about the reachability of telephone numbers and their associated gateways.

When receiving an initial SIP request, a SIP proxy looks at the request URI contained in the first line of the SIP request. (In Fig. 2, this is *bob@biloxi.com*) If the domain name is managed by the proxy, the proxy maps the user name to one or more contact addresses and sends replicas of the request to those addresses, creating *branches*. The requests may be sent sequentially, after a previous branch has failed, or in parallel. This procedure is referred to as *forking* and greatly simplifies reaching one individual that may be using multiple devices to communicate. It roughly mimics, on a global scale, the operation of multiple phones in a household, all ringing at once. Each branch may also return a redirection response, indicating one or more alternate addresses to be tried. As soon as somebody picks up at one of the addresses, that SIP user agent returns a success response and the proxy sends a CANCEL request to all other active branches, terminating ringing there.

If the request URI identifies a domain not handled by the proxy, the proxy is acting as an *outbound proxy*. Outbound proxies are typically used by the initiator of the request to handle all outbound requests, regardless of destination. For example, Alice might use an outbound proxy in her home domain, *atlanta.com*. Outbound proxies may be necessary if firewall policies restrict inbound requests or might be useful to offer additional outbound services, such as custom abbreviated dialing.

Unlike HTTP, SIP URIs attempt to avoid identifying a single physical server, but rather make it possible to direct a request to one of the servers handling SIP requests for a particular domain. This approach was first taken for email, using DNS MX [43] records to list a set of mail transport agents for a domain. SIP employs a two-step process [44], where first DNS NAPTR records [37] indicate the set of transport protocols (UDP, TCP, TCP with TLS, etc.) available for the domain. For each suitable protocol, DNS SRV [45] resource records list a set of candidate servers, qualified by preference and weight. The client looking for a server picks a random server among the highest-preference group of servers, performing a simple form of client-based load balancing, albeit without load feedback.

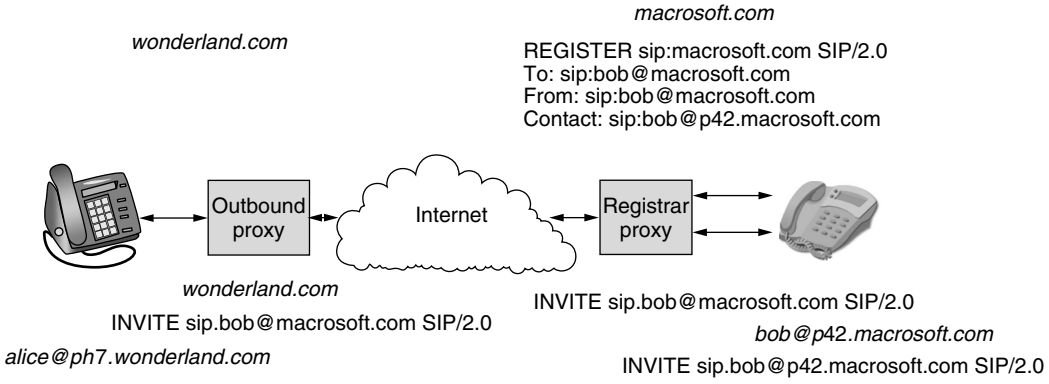


Figure 2. SIP architecture.

4.4. SIP Requests and Responses

SIP is a request-response protocol similar to HTTP. Requests and responses are expressed as plain text, consisting of a header or status line, a set of header fields and a message body. The message body can contain any information, including binary data. Just like HTTP and email, SIP uses MIME [46] to carry multiple distinct message bodies in a single request.

Each request is identified by a *method* that describes its basic functionality. The basic protocol defines six methods — INVITE, ACK, BYE for call setup and termination, OPTIONS for discovering capabilities, CANCEL for terminating pending calls, and REGISTER for establishing address bindings, but other specifications add additional methods. For example, the INFO method [47] can carry midcall ISUP PSTN call signaling information [48]. A request may trigger zero or more provisional responses that indicate call progress and one or more final responses. Each SIP method is qualified by a set of parameters expressed as header fields, similar to email or HTTP headers. Indeed, SIP inherits a large number of HTTP and email header fields.

The details of the header field syntax can be found in the protocol specification [49]. Header fields can be roughly divided into header fields that describe the message body, identify the request or response, help in routing requests, assist in authentication, negotiate protocol capabilities, and provide auxiliary information to end systems and proxies. The content-describing header fields identify the media type, language, handling, and encoding and measure the length of the body. The request is identified by components of the From, To, Call-ID and CSeq header fields. As described in Section 4.5, the Via, Record-Route, and Route header fields trace request paths and allow requests to revisit the same set of servers as a previous request. A set of header fields, mostly borrowed from HTTP [50], provide a challenge-response authentication framework, as described in Section 4.7. Headers also negotiate protocol capabilities, such as the support for media types in message bodies, human languages understood for error messages, and which protocol features are supported or required. A large set of optional header fields identify the caller or call, for example, the purpose of the call, the organization the caller or callee are affiliated with, or additional hints for the call, such as customized alerting or a Webpage related to the caller.

Responses are classified by their three-digit response code. The first digit is sufficient to identify the type of response; the first digit of "1" indicates a provisional response such as "trying" or "ringing", a first digit of "2," success; "3," a redirection; "4," a client error; "5," a server error; and "6," a global failure. (A global failure indicates that further searches for an instance of the callee are fruitless, for example, because the callee refuses to talk to the caller.)

A sample INVITE request and the corresponding final response is shown in Fig. 3. In the example, Alice, using her SIP device located at *pc33.atlanta.com*, calls Bob, at *biloxi.com*. Alice tells Bob that she can receive PCMU (that is, μ-law) audio over RTP using the AVP profile [30] at port 49172 and IP address 100.101.102.103; Bob responds to Alice that he is willing to talk and waiting to receive audio at port 33280 on address 110.111.112.113. The Call-ID field is a randomly chosen identifier for the call; the randomly chosen tags in the From and To fields identify the participants, allowing multiple terminals of Bob and Alice to be distinguished. The CSeq identifier allows Alice to unambiguously associate Bob's response with her request. The example does not show the third and final request in the call setup, with method ACK, that Alice sends to Bob to confirm that she has received Bob's "200 OK" response. The final ACK request establishes a SIP *dialog* between Alice and Bob. Only INVITE requests use an ACK request for confirmation.

As noted, SIP can operate directly on top of unreliable transport protocols such as UDP. To ensure that the call is set up, SIP implements its own retransmission mechanism. Due to forking at proxies, requests are retransmitted and acknowledged on each branch, not end to end. Requests are retransmitted at exponentially increasing intervals starting with the estimate for the round-trip time, if known, and 0.5 s if the round-trip time is unknown. The user agent client retransmits until it receives a provisional or final response or after seven retransmissions.

INVITE transactions differ somewhat in their behavior from all other requests. Non-INVITE requests are assumed to take only a small amount of time to process, so that the client simply retransmits until it gets a final response. It slows down retransmission to once every few seconds after it has received a provisional response, to deal with lost final responses. For INVITE transactions, the final answer can take tens of seconds or minutes since a human may need to answer the ringing phone. Here, the server retransmits the final response until the client confirms its receipt with an ACK request.

During the call, either Alice or Bob can send additional INVITE requests to update the media session, for example, to change the set of media or to adjust the media destination IP address. If either Alice or Bob hang up, their SIP terminal sends a BYE request to the other side. Typically, both midcall INVITE requests and the BYE request bypass any intermediate proxies since Alice and Bob now know each other's IP address.

4.5. Routing Requests and Responses

The routing of requests depends on the request URI and, for all but the first request in a dialog, on the Route header fields that may be present (see below). The response to the first request will contain a Contact header field that describes the location to send subsequent requests to. Thus, subsequent requests will contain that address as the request URI.

Route header fields lists proxies that indicated a desire to stay in the path of subsequent requests within a dialog, for example, any reINVITE requests to change session parameters or the final BYE request terminating the dialog. Requests in both directions, from caller to callee and vice versa, traverse this set of proxies. Route headers offer a functionality somewhat similar to IP source routing. A proxy that wants to see subsequent requests for a dialog

Request:

```
INVITE sip:bob@biloxi.com SIP/2.0
Via: SIP/2.0/UDP pc33.atlanta.com;branch=z9hG4bK776asdhds
To: Bob <sip:bob@biloxi.com>
From: Alice <sip:alice@atlanta.com>;tag=1928301774
Call-ID: a84b4c76e66710@pc33.atlanta.com
CSeq: 314159 INVITE
Contact: <sip:alice@pc33.atlanta.com>
Content-Type: application/sdp
Content-Length: 145

v=0
o=alice 2890844526 2890844526 IN IP4 atlanta.com
s=Session SDP
c=IN IP4 100.101.102.103
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

Response:

```
SIP/2.0 200 OK
To: Bob <sip:bob@biloxi.com>;tag=8321234356
From: Alice <sip:alice@atlanta.com>;tag=1928301774
Via: SIP/2.0/UDP pc33.atlanta.com;branch=z9hG4bK776asdhds
Call-ID: a84b4c76e66710@pc33.atlanta.com
CSeq: 314159 INVITE
Contact: <sip:alice@pc33.atlanta.com>
Content-Type: application/sdp
Content-Length: 140

v=0
o=bob 2890844526 2890844526 IN IP4 biloxi.com
s=Session SDP
c=IN IP4 110.111.112.113
t=0 0
m=audio 33280 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

Figure 3. Sample INVITE request and response.

adds itself to the *route set* for a dialog by adding Record-Route header fields to the initial request. The callee user agent then copies those Record-Route header fields into the response. The callee also keeps the list, in the same order, as its route set. The caller reverses the list and makes that its route set. The route set is then inserted as a Route header into subsequent requests by both caller and callee in the same session. Each proxy inspects the top-most route header and uses it to find the next hop toward the destination. If the Route header names the proxy itself, the proxy removes the header field.

Responses always traverse the same set of proxy servers as the corresponding request, guided by the Via header fields inserted into the request by the proxy servers.

4.6. Extending SIP

Since VoIP and multimedia conferencing are still in their infancy, it is likely that SIP will evolve. Also, some areas where SIP is being applied have specific requirements whose fulfillment should not burden other applications. SIP was designed to be extensible while maintaining a maximum of backward compatibility. Allowing for extensions requires defining suitable behavior when an implementation discovers an unknown protocol element and the ability to ascertain the capabilities of SIP elements.

Since SIP proxies forward requests independent of their method, additional methods can be added without upgrading all proxies. SIP user agents can indicate the methods that they support in an Allow header.

User agents and proxies can add new header field types without coordination, as proxies simply copy them into outgoing requests and receiving user agents ignore unknown header fields. If a user agent client wants to ensure that a particular feature is supported by a proxy or user agent server, it adds a Require header field indicating the name of the feature. In turn, the client or server can summarize its own capabilities with the Supported header field.

4.7. SIP Security

SIP poses a number of security challenges. It carries potentially sensitive information, such as subject information about calls, media encryption keys, and reachability information. The existence of communication relationships themselves may be considered private. Registrations need to be protected against malicious modifications, as such modifications would allow the attacker to redirect requests to any location. Also, attackers must be prevented from injecting requests into existing calls, as those fake requests could be used to terminate or redirect the call. The security challenges are increased by the use of proxies that may be operated by either the caller's network provider, the callee's network provider, or possibly the operator of the network being visited by either caller or callee.

Since it must be possible to make calls to parties with whom the caller does not have a preexisting security relationship, standard shared secret security is of only limited use. A public key infrastructure for large user populations does not currently exist, but organizations can easily obtain public key certificates, for example, for Web servers.

Given those constraints, SIP uses three security mechanisms, namely transport-layer security (TLS) [39,51], digest authentication [52], and secure/multipurpose Internet mail extensions (S/MIME) [53]. Naturally, IPsec can be deployed as well, but it is effectively invisible to SIP. TLS protects exchanges "hop by hop," that is, between user agent and proxy or registrar, and between proxies. With standard server certificates, the user agent can ascertain the identity of the registrar or proxy. Digest authentication is a challenge–response authentication protocol based on a shared secret. It is commonly used to protect registrations, although it offers header integrity only in combination with TLS. Finally, S/MIME is used in a fashion similar to email, but encapsulating the SIP message parts that are to be encrypted or signed in an outer SIP wrapper that remains visible to proxies. For encryption, S/MIME requires that the sender knows the recipients public key or has access to a shared public key infrastructure.

5. SIP EXTENSIONS

The IETF is extending SIP in a variety of ways to accommodate special requirements. Many of the extensions are motivated by the need to interoperate with the PSTN, but a major extension, SIP events (Section 5.6), offers new services and extends SIP to provide event notification. SIP extensions are designed to be backward-compatible, so that peers can establish sessions, with possibly reduced functionality, even if one or both of them do not support a particular extension. Many extensions are specific to particular deployment scenarios or applications and thus are likely to be supported by only a subset of implementations.

5.1. Session Keep-Alive

Once a SIP dialog has been established, the two peers have no direct way to discover if the other one has crashed or has been disconnected from the network. Usually, the lack of media data may provide a clue, but, unlike the situation in traditional phone calls, SIP sessions can persist even if no media are being exchanged. In circumstances where endpoints want to remove state if the other side is no longer reachable, the SIP session timer mechanism [54] has been proposed. The initial INVITE request indicates how often the caller would like the session to be refreshed. The callee can claim the role of refresher or leave this to the caller. The party responsible for refreshes periodically issues another INVITE request. If it receives no response, it terminates the call. Similarly, the party expecting periodic reINVITE requests will terminate the call if they are not forthcoming. The session interval can be lowered by proxies en route, but proxies can reject session intervals that are too short.

As with TCP liveness detection, SIP session timers must be used with caution, as they may unnecessarily terminate a long-lived session during brief network outages.

5.2. Conferencing

SIP supports a variety of multiparty conferencing architectures, including Internet multicast, end-system mixing, dialin conference servers, ad hoc centralized conferences, dialout conferences, or centralized signaling with peer-to-peer or multicast media [55,56]. For multicast and end-system mixing, there are no special servers. However, multicast is not yet widely available in the Internet; end-system mixing is likely to scale only to modest-size conferences, since one of the participants needs to send a mixed audio- or videostream to all other participants. The most common architectures are likely to be dialin and dialout conferences with a centralized conference server, or possibly a hierarchy of such servers.

Conferences are treated just like normal users; they are addressed using SIP URLs such as *sip:staff-meeting@example.com*. For ad hoc conferences, the initiator sends an INVITE request to the conference bridge, with a randomly chosen conference identifier.

The manner of adding users depends on the conference model. For multicast and end-system mixed conferences, regular INVITE requests suffice. For dialin conferences, conference members can ask others to join the conference by issuing a REFER request to the candidate member.

5.3. Multiparty Calls

A number of common telephony features, such as call waiting, blind and attended transfer, conference calls, call parking, call pickup, music on hold, call monitoring, barge-in, whispered call waiting, and single-line extensions, or Internet-oriented features such as presence-enabled conferencing all generally involve multiple participants, both human and machines (media servers). Often, media of these participants are mixed.

Many of these features can be implemented using third-party call control [38], where a central controller maintains and terminates dialogs, using INVITE, reINVITE, and BYE, with the participants and mixes media as appropriate. This approach has the advantage that it requires only basic SIP capabilities in the end systems, but it means that features have to implemented in a single point for each call.

Alternatively, a peer-to-peer approach can be used [57], where all functionality is implemented "at the edges," that is, by terminals. This approach requires a set of primitives to replace an existing dialog, join an existing dialog with another one, perform media replication by the media origin, and allow a user agent to ask another user agent to send a request on its behalf. The latter functionality is accomplished with the REFER request [58] that asks the recipient to construct another SIP request. The SIP request and its parameters and contained in the Refer-To header as a SIP URI.

5.4. Caller Preferences

In SIP, terminal addresses often identify a generic destination, such as a person, and not a particular terminal or phonejack. Thus, it is helpful to provide the caller with the ability to indicate which of the devices reachable by a particular SIP URI the caller would prefer to reach. This preference is expressed through properties that the destination should have, not its address [59]. For example, the caller can indicate a desire to reach voicemail instead of a human being, might prefer not to talk to a mobile phone or might like to be connected to a Chinese-speaking operator. SIP proxies then use this information in making call routing decisions.

5.5. Quality of Service

Since SIP requests do not necessarily traverse the same route as the data packets for the session they establish, SIP cannot directly control quality of service on the data path. Instead, this is left to mechanisms such as the Resource Reservation Protocol (RSVP) [60] or differentiated services. However, SIP can help negotiate the use of such mechanisms and determine when the resource reservation has succeeded. This is needed to prevent scenarios where the called phone rings, but then the resource reservation fails because of lack of network bandwidth, leading to a call without media or with only best-effort media streams. An additional SIP method, COMET (condition met), signals the successful completion of the resource reservation [61].

5.6. SIP for Events, Presence, and Instant Messaging

Events, that is, changes in state, are a useful abstraction in many telecommunication services. Examples include automatic callback, message waiting, and multi party conference management. Also, they underlie the notion of presence (usually combined with instance messaging) that has become a popular Internet service. Event notifications are often approximated by email messages, but since email is picked up only by the recipient, there can be a large and unpredictable latency between the event notification and its receipt. SIP can readily be extended to signal events. Here, SIP follows the common subscribe–publish model. A SIP entity sends a SIP SUBSCRIBE request to the source of the events. The subscriber is then notified, using NOTIFY, each time the event occurs. The subscription has a limited lifetime and needs to be refreshed before it expires.

SIP is suitable since many of the same properties needed for setting up calls apply. For example, the destination of the event subscription and event notification should be identified by a long-term address, while the actual destination may change network addresses.

Beyond signaling events related to SIP sessions, SIP events have also been defined for describing the presence state of a user [62], that is, whether the user is available to communicate by phone, SIP call, text chat, or other means. Since this information needs to be available even if the user terminal is not connected to the network, a presence agent acts as a standin for the user. It also merges presence information from multiple devices. The presence agent can use information from SIP binding updates (REGISTER) to detect changes in presence state.

5.7. Programming SIP Services

While not part of the protocol specification itself, there have been a number of proposals on how to program SIP-based proxies and end systems. Among these are JAIN SIP, SIP servlets [63], sip-cgi [64], and the Call Processing Language (CPL) [65,66]. JAIN SIP is a Java API that allows to construct SIP messages and extract information about SIP headers and responses. SIP servlets are similar to Java servlets for Web servers, offering a model where the servlet handles a full transaction. Sip-cgi attempts to transfer the cgi programming model found in Web servers to SIP proxies. SIP requests are passed as environment variables to scripts or programs, with each request causing a new process to be invoked. The scripts can be written in any language, including such common Web services scripting languages such as Perl, Python, or Tcl. Unlike HTTP requests, which are completely stateless, sip-cgi offers some support for associating multiple requests and responses within the same transaction. The script is responsible for parsing SIP headers and generating responses, but can simplify its task by invoking default behaviors.

CPL is an XML-based tree for handling a SIP transaction in proxies. It is useful primarily for call routing, admission, rejection, and logging. Each transaction is logically handled by a single CPL script, acting as a form of decision tree. The traversal of the tree can depend on the status responses returned by branches. Extensions of CPL to presence services and to end-system services are in progress. For voice-oriented services, voice menu systems such as VoiceXML are available.

5.8. SIP Performance

There currently is no generally accepted, standard way to measure the performance of SIP clients and servers. However, SIPstone [67] has been proposed as a simple metric for common operations, such as calls and registrations.

5.9. Emergency Services

One of the principal functions of the existing PSTN is to summon emergency help. In many countries, a system has emerged where callers can call a single, nationwide emergency number, such as 911 in the United States or 112 in many parts of Europe. The call is then directed to the nearest emergency call center, where the dispatcher

can see the caller's geographic location on her console. A similar architecture has been proposed for SIP [68], with a global emergency identifier, "sos," and a special proxy that maps user location to the nearest emergency call center. Unlike in the PSTN, the user location cannot be directly determined from the device address, as IP subnets can span large geographic regions, particularly with virtual private networks and dialup connections.

5.10. Configuring SIP Networks

SIP attempts to make it possible to set up a SIP network with minimal manual configuration. It is sufficient for a SIP device to know the address-of-record of its owner. The domain part of that address identifies the registrar for the owner. If desired, an outbound proxy can be discovered using DHCP [69]. Configuring other parameters, such as dial plans or SIP protocol timing parameters, is currently under discussion [70].

6. SUMMARY AND CONCLUSION

SIP is a signaling protocol that allows peers to establish multimedia sessions across the Internet. Since some of the requirements, such as mapping from long-term stable user identities to addresses, are similar, SIP can also be used to convey events to users, including changes of presence information. SIP is currently used primarily by voice over IP applications. A number of extensions have been standardized, with many more under discussion.

BIOGRAPHY

Henning Schulzrinne received his undergraduate degree in economics and electrical engineering from the Darmstadt University of Technology, Germany, his M.S.E.E. degree as a Fulbright scholar from the University of Cincinnati, Ohio, and his Ph.D. degree from the University of Massachusetts in Amherst, Massachusetts. He was a member of the technical staff at AT&T Bell Laboratories, Murray Hill and an associate department head at GMD-Fokus (Berlin), before joining the Computer Science and Electrical Engineering Departments at Columbia University, New York. His research interests encompass real-time, multimedia network services in the Internet and modeling and performance evaluation.

He is a division editor of the *Journal of Communications and Networks* and an editor of the *IEEE/ACM Transactions on Networking* and former editor of the *IEEE Internet Computing Magazine* and *IEEE Transactions on Image Processing*. He is member of the Board of Governors of the IEEE Communications Society and the ACM SIGCOMM Executive Committee, former chair of the IEEE Communications Society Technical Committees on Computer Communications and the Internet, and has been technical program chair of Global Internet, Infocom, NOSSDAV, and IPtel. He also was a member of the Internet Architecture Board (IAB).

Protocols codeveloped by him are now Internet standards, used by almost all Internet telephony and multimedia applications. His research interests include Internet multimedia systems, quality of service, and performance evaluation.

BIBLIOGRAPHY

1. H. Schulzrinne, Personal mobility for multimedia services in the Internet, in *European Workshop on Interactive Distributed Multimedia Systems and Services* (*IDMS*), Berlin, Germany, March 1996.
2. M. Handley and V. Jacobson, *SDP: Session Description Protocol*, RFC 2327, Internet Engineering Task Force, April 1998.
3. D. Kutscher, J. Ott, and C. Bormann, *Session Description and Capability Negotiation*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).
4. J. Postel, *Internet Protocol*, RFC 791, Internet Engineering Task Force, Sept. 1981.
5. S. Deering and R. Hinden, *Internet Protocol, Version 6* (*ipv6*) *Specification*, RFC 2460, Internet Engineering Task Force, Dec. 1998.
6. J. Postel, *User Datagram Protocol*, RFC 768, Internet Engineering Task Force, Aug. 1980.
7. J. Postel, *Transmission Control Protocol*, RFC 793, Internet Engineering Task Force, Sept. 1981.
8. R. Stewart et al., *Stream Control Transmission Protocol*, RFC 2960, Internet Engineering Task Force, Oct. 2000.
9. J. Rosenberg, H. Schulzrinne, and G. Camarillo, *SCTP as a Transport for SIP*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).
10. J. Lennox, H. Schulzrinne, and T. F. L. Porta, *Implementing Intelligent Network Services with the Session Initiation Protocol*, Technical Report CUCS-002-99, Columbia Univ., New York, NY, Jan. 1999.
11. S. Bradner, *The Internet Standards Process—Revision 3*, RFC 2026, Internet Engineering Task Force, Oct. 1996.
12. W. Jiang, J. Lennox, H. Schulzrinne, and K. Singh, Towards junking the PBX: deploying IP telephony, *Proc. Int. Workshop on Network and Operating System Support for Digital Audio and Video* (*NOSSDAV*), Port Jefferson, NY, June 2001.
13. International Telecommunication Union, *Introduction to CCITT Signalling System No. 7*, Recommendation Q.700, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, March 1993.
14. A. Vemuri and J. Peterson, *SIP for telephones* (*SIP-t*): *Context and Architectures*, Internet Draft, Internet Engineering Task Force, April 2002 (work in progress).
15. E. Miller, F. Andreasen, and G. Russell, The PacketCable architecture, *IEEE Commun. Mag.* **39**: (June 2001).
16. W. Marshall et al., *SIP Extensions for Supporting Distributed Call State*, Internet Draft, Internet Engineering Task Force, Aug. 2001 (work in progress).
17. W. Marshall, F. Andreasen, and D. Evans, *SIP Extensions for Media Authorization*, Internet Draft, Internet Engineering Task Force, May 2002 (work in progress).
18. W. Marshall et al., *SIP Extensions for Network-Asserted Caller Identity and Privacy within Trusted Networks*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).

19. B. Culpepper, R. Fairlie-Cuninghame, and J. Mule, *SIP Event Package for Keys*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).
20. A. Roychowdhury and S. Moyer, Instant messaging and presence for network appliances using SIP, *Internet Telephony Workshop 2001*, New York, April 2001.
21. J. Toga and J. Ott, ITU-T standardization activities for interactive multimedia communications on packet-based networks: H.323 and related recommendations, *Comput. Networks ISDN Syst.* **31**: 205–223 (Feb. 1999).
22. International Telecommunication Union, *Visual Telephone Systems and Equipment for Local Area Networks Which Provide a Non-Guaranteed Quality of Service*, Recommendation H.323, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1996.
23. International Telecommunication Union, *Functional Description of the ISDN User Part of Signalling System No. 7*, Recommendation Q.761, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 1994.
24. H. Schulzrinne and J. Rosenberg, A comparison of SIP and H.323 for Internet telephony, *Proc. Int. Workshop on Network and Operating System Support for Digital Audio and Video* (*NOSSDAV*), Cambridge, UK, July 1998, pp. 83–86.
25. I. Dalgic and H. Fang, Comparison of H.323 and SIP for IP telephony signaling, *Proc. Photonics East*, Boston, MA, SPIE, Sept. 1999.
26. Nortel Networks, *A Comparison of H.323v4 and SIP*. 3GPP contribution, Jan. 2000.
27. T. Eyers and H. Schulzrinne, Predicting internet telephony call setup delay, *Proc. 1st IP-Telephony Workshop* (*IPtel 2000*), Berlin, Germany, April 2000.
28. K. Singh and H. Schulzrinne, Interworking between SIP/SDP and H.323, *Proc. 1st IP-Telephony Workshop* (*IPtel 2000*), Berlin, Germany, April 2000.
29. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, Internet Engineering Task Force, Jan. 1996.
30. H. Schulzrinne, *RTP Profile for Audio and Video Conferences with Minimal Control*, RFC 1890, Internet Engineering Task Force, Jan. 1996.
31. H. Schulzrinne, A. Rao, and R. Lanphier, *Real Time Streaming Protocol* (*RTSP*), RFC 2326, Internet Engineering Task Force, April 1998.
32. K. Singh and H. Schulzrinne, Unified messaging using SIP and RTSP, *Proc. IP Telecom Services Workshop*, Atlanta, GA, Sept. 2000 pp. 31–37.
33. M. Arango et al., *Media Gateway Control Protocol* (*MGCP*) *Version 1.0*, RFC 2705, Internet Engineering Task Force, Oct. 1999.
34. F. Cuervo et al., *Megaco Protocol Version 1.0*, RFC 3015, Internet Engineering Task Force, Nov. 2000.
35. International Telecommunication Union, *The International Public Telecommunication Numbering Plan*, Recommendation E.164, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1997.
36. P. Faltstrom, *E.164 Number and DNS*, RFC 2916, Internet Engineering Task Force, Sept. 2000.
37. R. Daniel and M. Mealling, *Resolution of Uniform Resource Identifiers Using the Domain Name System*, RFC 2168, Internet Engineering Task Force, June 1997.
38. J. Rosenberg, J. Peterson, H. Schulzrinne, and G. Camarillo, *Third Party Call Control in SIP*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).
39. T. Dierks and C. Allen, *The TLS Protocol Version 1.0*, RFC 2246, Internet Engineering Task Force, Jan. 1999.
40. A. Vaha-Sipila, *URLs for Telephone Calls*, RFC 2806, Internet Engineering Task Force, April 2000.
41. J. Rosenberg and H. Schulzrinne, *A Framework for Telephony Routing over IP*, RFC 2871, Internet Engineering Task Force, June 2000.
42. J. Rosenberg, H. Salama, and M. Squire, *Telephony Routing over IP* (*TRIP*), RFC 3219, Internet Engineering Task Force, Jan. 2002.
43. C. Partridge, *Mail Routing and the Domain System*, RFC 974, Internet Engineering Task Force, Jan. 1986.
44. J. Rosenberg and H. Schulzrinne, *SIP: Locating SIP Servers*, RFC 3263, Internet Engineering Task Force, May 2002.
45. A. Gulbrandsen, P. Vixie, and L. Esibov, *A DNS RR for Specifying the Location of Services* (*DNS SRV*), RFC 2782, Internet Engineering Task Force, Feb. 2000.
46. N. Borenstein and N. Freed, *MIME* (*multipurpose Internet Mail Extensions*) *Part One*: *Mechanisms for Specifying and Describing the Format of Internet Message Bodies*, RFC 1521, Internet Engineering Task Force, Sept. 1993.
47. S. Donovan, *The SIP INFO Method*, RFC 2976, Internet Engineering Task Force, Oct. 2000.
48. E. Zimmerer et al., *MIME Media Types for ISUP and QSIG Objects*, RFC 3204, Internet Engineering Task Force, Dec. 2001.
49. J. Rosenberg et al., *SIP: Session Initiation Protocol*, RFC 3261, Internet Engineering Task Force, May 2002.
50. R. Fielding et al., *Hypertext Transfer Protocol — HTTP/1.1*, RFC 2616, Internet Engineering Task Force, June 1999.
51. E. Rescorla, *HTTP over TLS*, RFC 2818, Internet Engineering Task Force, May 2000.
52. J. Franks et al., *HTTP Authentication: Basic and Digest Access Authentication*, RFC 2617, Internet Engineering Task Force, June 1999.
53. B. Ramsdell, *S/MIME Version 3 Message Specification*, RFC 2633, Internet Engineering Task Force, June 1999.
54. S. Donovan and J. Rosenberg, *SIP Session Timer*, Internet Draft, Internet Engineering Task Force, Oct. 2001 (work in progress).
55. J. Rosenberg and H. Schulzrinne, *Models for Multi Party Conferencing in SIP*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).
56. K. Singh, G. Nair, and H. Schulzrinne, Centralized conferencing using SIP, *Proc. Internet Telephony Workshop 2001*, New York, April 2001.
57. R. Mahy et al., *A Multi-Party Application Framework for SIP*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).
58. R. Sparks, *The Refer Method*, Internet Draft, Internet Engineering Task Force, Oct. 2001 (work in progress).
59. H. Schulzrinne and J. Rosenberg, *SIP Caller Preferences and Callee Capabilities*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).

60. R. Braden et al., *Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification*, RFC 2205, Internet Engineering Task Force, Sept. 1997.

61. W. Marshall, G. Camarillo, and J. Rosenberg, *Integration of Resource Management and SIP*, Internet Draft, Internet Engineering Task Force, April 2002 (work in progress).

62. J. Rosenberg et al., *Session Initiation Protocol (SIP) Extensions for Presence*, Internet Draft, Internet Engineering Task Force, April 2002 (work in progress).

63. A. Kristensen and A. Byttner, *The SIP Servlet API*, Internet Draft, Internet Engineering Task Force, Sept. 1999 (work in progress).

64. J. Lennox, H. Schulzrinne, and J. Rosenberg, *Common Gateway Interface for SIP*, RFC 3050, Internet Engineering Task Force, Jan. 2001.

65. J. Lennox and H. Schulzrinne, *CPL: A Language for User Control of Internet Telephony Services*, Internet Draft, Internet Engineering Task Force, Nov. 2001 (work in progress).

66. J. Lennox and H. Schulzrinne, The call processing language: User control of internet telephony services, *Proc. Lucent Technologies XML Day*, Murray Hill, NJ, Feb. 2000.

67. H. Schulzrinne, S. Narayanan, J. Lennox, and M. Doyle, *SIPstone—Benchmarking SIP Server Performance*, Technical Report CUCS-005-02, Dept. Computer Science, Columbia Univ., New York, March 2002.

68. H. Schulzrinne and K. Arabshian, Providing emergency services in internet telephony, *IEEE Internet Comput.* **6**: 39–47 (May 2002).

69. H. Schulzrinne, *DHCP Option for SIP Servers*, Internet Draft, Internet Engineering Task Force, March 2002 (work in progress).

70. C. Stredicke and I. Butcher, *SIP End Point Configuration Data Format*, Internet Draft, Internet Engineering Task Force, Feb. 2002 (work in progress).

SHALLOW-WATER ACOUSTIC NETWORKS*

JOHN G. PROAKIS
JOSEPH A. RICE
ETHEM M. SOZER
MILICA STOJANOVIC
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Since the early 1980s, the underwater acoustic (UWA) communications technology has progressed significantly. Communication systems with increased bit rate and reliability now enable real-time point-to-point links between underwater nodes such as ocean-bottom sensors and autonomous underwater vehicles (AUVs). Current research is focused on combining various point-to-point links within a network structure to meet the emerging demand for applications such as environmental data collection, offshore exploration, pollution monitoring, and military surveillance [1].

The traditional approach for ocean-bottom or ocean-column monitoring is to deploy oceanographic sensors, record the data, and recover the instruments. This approach has several disadvantages:

- The recorded data cannot be recovered until the end of the mission, which can be several months.
- There is no interactive communication between the underwater instruments and the onshore user. Therefore, it is not possible to reconfigure the system as interesting events occur.
- If a failure occurs before recovery, data acquisition may stop, or all the data may be lost.

The long delay between data acquisition and recovery can be reduced by using expendable or reusable communication probes as in the EMMA [2] and GEOSTAR [3] systems. Both systems have ocean-bottom sensors and a number of probes that have radiocommunication equipment. The data collected by sensors are carried to the surface with probes at preprogrammed intervals or as soon as some interesting events occur. After surfacing, the probe sends the data to an onshore user via satellite. The release of the probes can also be forced by sending acoustic signals from a nearby ship. These systems provide quasi-real-time data collection. However, lack of bidirectional communication links and the high cost of probes limit their usage.

The ideal solution for real-time monitoring of selected ocean areas for long periods of time is to connect various instruments through wireless links within a network structure. Basic underwater acoustic networks are formed by establishing bidirectional acoustic communication between nodes such as autonomous underwater vehicles (AUVs) and fixed sensors. An RF link connects the network to a surface station that can be further connected to terrestrial networks, such as the Internet, through an RF link. Onshore users can extract real-time data from multiple distant underwater instruments. After evaluating the obtained data, they can send control messages to individual instruments. Since data is not stored in the underwater instruments, data loss is prevented as long as isolated node failures can be circumvented by reconfiguring the network.

A major constraint of UWA networks is the limited energy supply. Whereas the batteries of a wireless modem can be easily replaced on land-based systems, the replacement of an underwater modem battery involves ship time and retrieval of the modem from the ocean bottom, which is costly and time-consuming. Therefore, transmission energy is precious in underwater applications. Network protocols should conserve energy by reducing the number of retransmissions, powering down between transactions, and minimizing the energy required per transmission.

* This work was supported by the Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research Contract N00014-00-1-0564, by Small-Business Innovative Research (SBIR) Program, and by ONR 321.

Some underwater applications require that the network be deployed quickly without substantial planning, such as in rescue and recovery missions. Therefore, the network should be able to determine the node locations and configure itself automatically to provide an efficient data communication environment. Also, if the channel conditions change or some of the nodes fail during the mission, the network should be capable of reconfiguring itself dynamically to continue its operation.

2. UNDERWATER ACOUSTIC COMMUNICATIONS

Unlike digital communications through radio channels where data are transmitted by means of electromagnetic waves, acoustic waves are used primarily in underwater channels. The propagation speed of acoustic waves in UWA channels is five orders of magnitude less than that of radiowaves. This low propagation speed increases the latency of a packet network.

The available bandwidth of an UWA channel depends critically on transmission loss, which increases with both range and frequency, and severely limits the available bandwidth [4,5]. For example, long-range systems that operate over several tens of kilometers may have a bandwidth of only a few kilohertz, while a short-range system operating over several tens of meters may have more than a 100 kHz bandwidth [6]. Within this limited bandwidth, the acoustic signals are subject to time-varying multipath [4], which may result in severe intersymbol interference (ISI) and large Doppler shifts and spreads, relative to radio channels, especially in shallow-water channels. Multipath propagation and Doppler effects degrade of acoustic signals and limit the data throughput. Special processing techniques are needed to combat these channel impairments.

Until the year 1990, as a result of the challenging characteristics of UWA channels, modem development was focused on employing noncoherent frequency shift keying (FSK) signals for achieving reliable communication. Since FSK demodulation is based on energy detection, it does not require phase tracking, which is a very difficult task in high Doppler spread environments. The multipath effects are eliminated by inserting guard periods between successive pulses to ensure that all the reverberation vanishes before each subsequent pulse is to be received. In addition, to avoid Doppler effects, some guard bands are employed between frequency tones. By varying the values of the guard bands, the communication signals can be matched to the channel characteristics, providing an adaptive modem structure. Table 1 presents some data on the noncoherent FSK modems described in the literature.

Although noncoherent FSK systems are effective in UWA channels, their low bandwidth efficiency makes them inappropriate for high-data-rate applications such as multiuser networks. The need for high-throughput, long-range systems has resulted in a focus toward coherent modulation techniques.

Today, with the availability of powerful digital signal processing devices, we are able to employ fully coherent PSK modulation in underwater communications. Equalizers are used to undo the effects of ISI, instead of trying to avoid or suppress it. When combined with explicit phase tracking loops, such as phase-locked loops (PLLs), decision feedback equalizers can provide high data throughput [7]. Other similar structures that use transversal filters and various adaptation algorithms are also reported in the literature. Coherent systems are summarized in Table 1.

Table 1. Summary of Performance Metrics for Some UWA Modems Presented in the Literature [5]

Type	Year	Data Rate (bps)	Bandwidth (kHz)	Range (km)[a]
FSK	1984	1200	5	3.0_S
FSK	1991	1250	10	2.0_D
FSK	1997	2400–600	5	10.0_D–5.0_S
Coherent	1989	500,000	125	0.06_D
Coherent	1993	600–300	0.3–1	89_S–203_D
Coherent	1994	20	20	0.9_S
Coherent	1998	1670–6700	2–10	4.0_S–2.0_S

[a] Subscript S indicates a shallow-water result; D indicates a deep-water result, generally a vertical channel.

Current research is focused on DSP algorithms with decreased complexity and multiuser modems that can operate in a network environment.

3. UNDERWATER ACOUSTIC NETWORKS

Information networks are designed in the form of a layered architecture [8]. The first three layers of this hierarchical structure are the physical layer, the data-link control layer, and the network layer.

The function of physical layer is to create a virtual link for transmitting a sequence of logical information (bits 0 and 1) between pairs of nodes. The information bits are converted into acoustic signals (in case of UWA networks), which are transmitted through the acoustic channel. At the receiving node, the physical layer converts the channel corrupted signals back into logical bits. The modem structures that can be used in the physical layer of an acoustic network were discussed in the previous section.

The second layer in the hierarchical structure is the *data-link control* (DLC) layer, which is responsible for converting the unreliable bit pipe of the physical layer into a higher-level error-free link. For this purpose DLC employs two mechanisms: framing and error correction control. Framing is accomplished by adding header information, which consists of a synchronization preamble, with source and destination addresses at the beginning of the information sequence, and cyclic redundancy check (CRC) bits at the end. The CRC bits are formed from the bits in the packet and are used for error correction control.

At the receiver side, the DLC performs a check using the CRC field to detect errors in a packet. If the CRC fails, it may ask for a retransmission depending on the automatic repeat request (also known as the Automatic Repeat reQuest) (ARQ) protocol. Some widely used ARQ schemes are stop and wait, go-back-N, and selective repeat. These protocols control the logical sequence of transmitting packets between two nodes

and acknowledging the correctly received packets. ARQ procedures form the logical link control (LLC) sublayer of the DLC. If the network is based on multiaccess links, rather than point-to-point links, additional measures must be taken to orchestrate the access of multiple sources to the same medium. These measures are called *media access control* (MAC). Commonly used MAC protocols are the ALOHA protocol, the carrier sense media access (CSMA) protocols, and token protocols. These protocols form the MAC sublayer of DLC.

The layer above the DLC layer is the network layer. The main function of the network layer is to transfer information packets to their final destination, which is called *routing*. Routing involves finding a path through the network and forwarding the packets from the source to the destination along this path. If a route is established at the beginning of a transaction and all the packets follow this path, the network is called a *virtual circuit-switching network*. If a new path is determined for each packet of the same transaction, the network employs datagram switching. In case of datagram switching, packets may arrive to the destination out of order. Therefore, the network layer should reorder the packets before passing them to a higher level.

The network layer selects optimal routes by minimizing the end-to-end path distance. The distance metric can be the delay, the number of hops, required energy, or some other "distance" measure. Some well-known static routing algorithms are the Dijkstra algorithm and the Bellman–Ford algorithm. In dynamic environments, the routes provided by these static algorithms are modified as the "distance" metrics change.

In the following paragraphs, we review the methods and protocols used in the DLC layer and network layer, together with their applicability to UWA networks. We also discuss possible network topologies, which is an important constraint in designing a network protocol.

3.1. Network Topologies

Three basic topologies can be used to interconnect network nodes: centralized, distributed, and multihop [9]. In a *centralized* network, the communication between nodes takes place through a central station, which is sometimes called the "hub" of the network. The network is connected to a backbone at this central station. This configuration is suitable for deep-water networks, where a surface buoy with both an acoustic and an RF modem acts as the hub and controls the communication to and from ocean-bottom instruments. A major disadvantage of this configuration is the presence of a single failure point [9]. If the hub fails, the entire network shuts down. Also, because of the limited range of a single modem, the network cannot cover large areas.

The next two topologies support peer-to-peer links. A fully connected peer-to-peer topology provides point-to-point links between every node of the network. Such a topology eliminates the need for routing. However, the output power needed for communicating with widely separated nodes is excessive. Also, a node that is trying to send packets to a far-end node can overpower and interfere with communication between neighboring nodes, which is called the *near–far problem* [9].

Multihop peer-to-peer networks are formed by establishing communication links only between neighboring nodes. Messages are transferred from source to destination by hopping packets along a multinode route. Routing of the messages is handled by intelligent algorithms that can adapt to changing conditions. Multihop networks can cover relatively larger areas since the range of the network is determined by the number of nodes rather than the modem range.

One of the UWA network design goals is to minimize the energy consumption while providing reliable connectivity between the nodes in the network and the backbone. Network topology is an important parameter that determines the energy consumption [10]. A simplified scenario in which a number of nodes and a master node arranged linearly along a line is considered. The nodes are uniformly distributed, and each node tries to send its packets to the master node. Two extreme communication strategies are possible in this scenario. In the first strategy, each node has direct access to the master node (fully connected topology). In the second strategy, each node transmits only to its nearest neighbor, who then relays the information toward the master node (multihop peer-to-peer topology). The energy consumption curves for both of these cases is plotted as a function of the total distance spanned by the network shown in Fig. 1. Dashed curves represent the case of direct access, which obviously requires more energy. For direct access, inclusion of each additional node results in an increase in total energy. For relaying, the situation is reversed; inclusion of each additional node decreases the total energy consumption because the additional node serves as an additional relay along the same distance.

Hence, the strategy that minimizes energy consumption is multihop peer-to-peer topology. The price paid for the decrease in energy consumption is the need for a sophisticated communication protocol and an increase in packet delay. Therefore, special attention should be given to applications that are sensitive to delays.

3.2. Multiple-Access Methods

In many information networks, communication is bursty, and the amount of time a user spends transmitting over the channel is usually smaller than the amount of time it remains idle. Thus, network users should share the available frequency and time in an efficient manner by means of a multiple access method. Frequency-division multiple access (FDMA) divides the available frequency band into subbands and assigns each subband to an individual user. Because of the severe bandwidth limitations and vulnerability of narrowband systems to fading, FDMA systems do not provide an efficient solution for UWA applications. Instead of dividing the frequency band, time-division multiple access (TDMA) divides a time interval, called a "frame," into time slots. Collision of packets from adjacent time slots is prevented by including guard times that are proportional to the propagation delay present in the channel. TDMA systems require very precise synchronization for proper utilization of the time

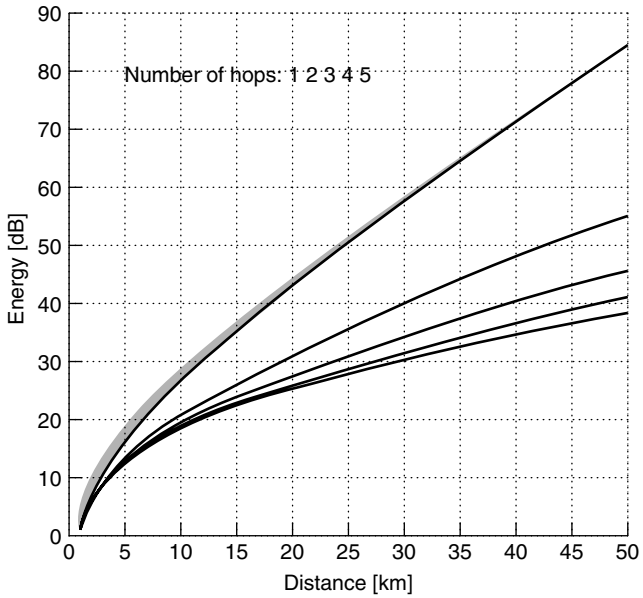


Figure 1. Total (normalized) energy needed to transmit a packet from each network node to the master node. Solid curves represent relaying; dotted curves represent direct access. The parameter on the curves is the number of hops. For relaying, the number of hops increases fro the top curve downward; although more packets are sent when there are more hops, total energy consumption is lower. For direct access, there is little difference between the curves, and the situation is reversed: the number of hops increases from 1 for the lowest energy consumption to 5 for the highest.

slots. High latency present in UWA channels requires long guard times that limits the efficiency of TDMA. Also, establishing a common timing reference is a difficult task. Code-division multiple access (CDMA) allows multiple users to transmit simultaneously over the entire frequency band. Signals from different users are distinguished by means of pseudonoise (PN)codes that are used for spreading the user messages. The large bandwidth of CDMA channels provides resistance to frequency-selective fading and exploits the time diversity present in the UWA channel by employing RAKE filters at the receiver in the case of DS-CDMA (direct-sequence CDMA) [11]. Spread-spectrum signals can be used for resolving collisions at the receiver by using multiuser detectors [12]. In this way, the number of retransmissions and energy requirements of the system can be reduced. This property both reduces battery consumption and increases the throughput of the network. Also, the power requirement of CDMA systems may be less than one-tenth that of TDMA [13]. In conclusion, CDMA appears to be a promising multiple-access technique for shallow-water acoustic networks.

3.3. Media Access Control (MAC) Protocols

Since the number of channels (time frames, frequency bands, or spreading codes) offered by a multiple access method can be much less than the total number of users in a network environment, the same channel is assigned to more than one user. If these users access the channel at the same time, their signals overlap and may be lost (packet collision). Likewise, most underwater acoustic modems are half-duplex in nature, and signals arriving during a transmission are lost, and must also be treated as packet collisions. Media access control (MAC) protocols are used to avoid information loss as a result of packet collisions.

A group of MAC protocols, such as the ALOHA protocol, do not try to prevent collisions but detect collisions and retransmit lost packets. The original ALOHA protocol is based on random access of users to the medium [14]. Whenever a user has information to send, it transmits it immediately. An acknowledgment (ACK) is sent back by the receiver if the packet is received without errors. Because of the arbitrary transmission times, collisions occur and packets are lost. Slotted ALOHA is an enhanced version of the ALOHA protocol, where the time frame is divided into time slots. When a node wants to send a packet, it waits until the next time slot and then begins transmission. Restricting packet transmission to predetermined time slots decreases the probability of collisions [9]. As in the case of TDMA, the ALOHA protocol is inefficient for UWA environment due to slow propagation. Also, the need for retransmissions increases the power consumption of the network nodes and reduces the lifetime of the network.

The number of retransmissions can be reduced if the MAC protocol uses a priori information about the channel state. The media access methods based on this idea are called *carrier sense multiple access* (CSMA) [9]. Details and various forms of this method can be found in papers by Kleinrock and Tobagi [15–18]. The CSMA method tries to avoid the collisions by listening for a carrier in the vicinity of the transmitter. However, this approach does not avoid collisions at the receiver [19]. Let us consider a network formed by three users as shown in Fig. 2. The circles around each node show the communication range of that node. Assume that node A is sending a packet to node B. At the same time, node C listens to the channel and because it is out of the range of A and does not detect the carrier of A, it begins transmission. This creates a collision at B, which is the receiver node. Node A was hidden from node C. This situation is called the "hidden node" scenario [19]. To enable B to hear both messages, node C should defer its transmission. However, if the destination of the packet of C is not B, there is no reason to defer the transmission, provided that node B has the capability to deal with the interference generated by the signal from C [19]. In the case of B sending a packet to A,

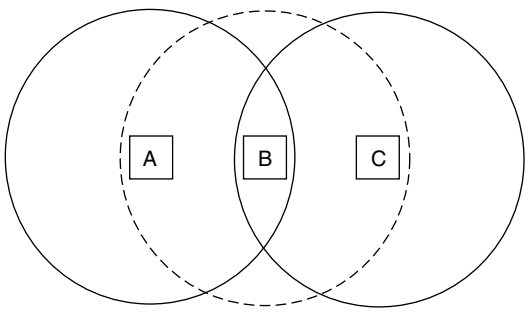


Figure 2. Node A can communicated with node B, but not with node C. Node B can communicated with both A and C.

node C detects a carrier. This creates an "exposed node" situation, where C is exposed to node B.

CSMA cannot solve these problems without adding a guard time between transmissions that is proportional to the maximum propagation time present in the network. The extensive propagation delays in underwater channels can cause this method to become very inefficient. If we consider an underwater acoustic network with a maximum range of 10 km, a data rate of 1 kbps, and a packet size of 1000 bits, the transmission delay and the maximum propagation delay become 1 and 6.7 s, respectively. In this situation, most of the time the channel will be idle, which results in low throughput.

The multiple access with collision avoidance (MACA) protocol was proposed by Karn [20] to detect collisions at the receiver as an alternative to CSMA. This protocol uses two signaling packets: "request to send" (RTS) and "clear to send" (CTS). When node A wants to send a message to B, it first issues an RTS command that contains the length of the message that is to be sent. If B receives the RTS, it sends back a CTS command that also contains the length of the message. As soon as A receives CTS, it begins transmission of the data packet. A node that overhears an RTS (C in this case) defers long enough to let node A receive the corresponding CTS. Also, any node that overhears a CTS defers its transmission for the length of the data packet to avoid collision. If a node overhears an RTS but not a CTS, it decides that it is out of range of the receiver, and transmits its own packet. Therefore, this protocol can solve both the hidden- and the exposed-node problems. The nodes can probe the channel during the RTS–CTS exchange [20]. The channel state information can be used to set the physical-layer parameters, such as output power and modulation type, These properties of the MACA protocol are essential for efficient UWA network design. MACA provides information for reliable communication with minimum energy consumption and can prevent collisions before they occur. The RTS–CTS exchange adds overhead but the reduction of retransmissions can compensate for this increase.

The MACAW protocol was proposed by Bharghavan [19] to improve performance and reliability of the MACA protocol. Instead of creating error-free, reliable point-to-point links with the DLC layer by use of acknowledgments, the MACA protocol ensures the reliability of the end-to-end link with the network layer. If some packets of a message are lost because of errors, the final destination node will ask the originating source to retransmit the lost packets. On highly reliable links, this approach increases throughput, since it eliminates the need to send individual acknowledgments for each hop. In case of poor-quality communication channels, a message will most probably contain erroneous packets. Recovering the errors in the data packet at the network layer will require excessive delay. Generally, error correction is better performed at the data-link layer for channels of low reliability, such as radio or shallow-water acoustic channels. For this purpose, an ACK packet is transmitted after each successful transaction. Including an extra packet in the transaction increases the overhead, which decreases the throughput. However, it has been shown [19] that, for radio channels, the gain in throughput exceeds the increase in overhead. This result may also apply to UWA channels. The MACAW protocol ignores power control and asymmetries that can occur. Its performance under power control needs to be investigated. Also, the effect of adding more overhead to the protocol in an environment where propagation delays are excessive needs to be assessed.

Deng and Haas [21] show that the performance of protocols based on RTS–CTS exchange can degrade as a result of the collision of control packets (RTS–CTS), especially in the case of high propagation and transmission delay. The dual-busy-tone multiple access (DBTMA) protocol is proposed to reduce the packet collision probability. In this protocol, the single shared channel is divided into two subchannels: a data channel and a control channel. Control packets are transmitted on the control channel, while data are carried on the data channel. In addition, two out-of-band tones are introduced to indicate that a node is transmitting on the data channel (BT_t) and a node is receiving on the data channel (BT_r). Researchers face two important limitations for employment of this protocol in an UWA network where energy and bandwidth are scarce: (1) the use of additional tones increases the energy consumption of network nodes, and (2) the divided bandwidth decreases the data throughput of nodes. However, because of the reduced number of collisions, the total energy consumption may be reduced, while network throughput can be increased.

3.4. Automatic Repeat Request (ARQ) Methods

Automatic repeat request (ARQ) is used in the data-link control layer to request the retransmission of erroneous packets. The simplest ARQ scheme that can be directly employed in a half-duplex UWA channel is the stop-and-wait ARQ, where the source of the packet waits for an ACK from the destination node for the confirmation of error-free packet transmission. Since the channel is not utilized during the round-trip propagation time, this ARQ scheme has low throughput. In go-back-N and selective-repeat ARQ schemes, nodes transmit packets and receive ACKs at the same time, and therefore require full-duplex links. Dividing the limited bandwidth of the UWA channels into two channels for full-duplex operation can significantly reduce the data rate of the physical layer. However, the effect on the overall network throughput needs to be investigated.

The selective-repeat ARQ scheme can be modified to work on half-duplex UWA channels. Instead of acknowledging each packet individually at reception time, the receiver will wait for N packet durations and send an ACK packet with the identification number of packets received without errors. Accordingly, the source of the packets will send N packets and wait for the ACK. Then, the source will send another group of N packets that contains the unacknowledged packets and new packets.

Acknowledgments can be handled in two possible ways. In the first approach, which is called *positive acknowledgment*, on reception of an error-free packet, the destination node will send an ACK packet to the source node. If the source does not receive an ACK packet before

a preset timeout duration, it will retransmit the data packet. In the case of a *negative acknowledgment*, the destination sends a packet if it receives a corrupted packet or does not receive a scheduled data packet. A negative acknowledgment may help conserve energy by eliminating the need to send explicit ACK packets and retransmission of data packets in case of a lost ACK packet. When combined with a MACA-type MAC protocol, the negative acknowledgment scheme may provide highly reliable point-to-point links due to the information obtained during RTS–CTS exchange as discussed in Section 3.3.

3.5. Routing Algorithms

As previously indicated, there are two basic methods used for routing packets through an information network: *virtual circuit* routing, where all the packets of a transaction follow the same path through the network, and *datagram* routing, where packets are allowed to pass through different paths. Networks using virtual circuits decide on the path of the communication at the beginning of the transaction. In datagram switching, each node that is involved in the transaction makes a routing decision, which is to determine the next hop of the packet.

Many of the routing methods are based on the *shortest-path* algorithm. In this method, each link in the network is assigned a cost that is a function of the physical distance and the level of congestion. The routing algorithm tries to find the shortest path, that is, the path with lowest cost, from a source node to a destination node. In a distributed implementation each node determines the cost of sending a data packet to its neighbors and shares this information with the other nodes of the network. In this way, every node maintains a database that reflects the cost of possible routes.

For routing, let us consider the most general problem where network nodes are allowed to move. This situation can be viewed as an underwater network with both fixed ocean-bottom sensors and AUVs. The instruments temporarily form a network without the aid of any preexisting infrastructure. These types of networks are called *ad hoc networks* [22].

In ad hoc networks the main problem is to obtain the most recent state of each individual link in the network, so as to decide on the best route for a packet. However, if the communication medium is highly variable as in the shallow-water acoustic channel, the number of routing updates can be very high. Current research on routing focuses on reducing the overhead added by routing messages while at the same time finding the best path, which are two conflicting requirements. Broch et al. [23] compared four ad hoc network routing protocols presented in the literature:

- Destination sequence distance vector (DSDV) [24]
- Temporally ordered routing algorithm (TORA) [25]
- Dynamic source routing (DSR) [26]
- Ad hoc on-demand distance vector (AODV) [27]

DSDV maintains a list of *next hops* for each destination node that belongs to the shortest-distance route. The protocol requires each node to periodically broadcast routing updates to maintain routing tables.

TORA is a distributed routing algorithm. The routes are discovered on demand. This protocol can provide multiple routes to a destination very quickly. The route optimality is considered as a second priority, and the routing overhead is reduced.

DSR employs source routing; that is, the route of each packet is included in its header. Each intermediate node that receives the packet checks the header for the next hop and forwards the packet. This eliminates the need for intermediate nodes to maintain best routing information to route the packets.

AODV uses the on-demand route discovery and maintenance characteristic of DSR and employs them in a hop-by-hop routing scheme instead of source routing. Also, periodic updates are used in this protocol.

In a mobile radio environment DSR provides the best performance in terms of reliability, routing overhead, and path optimality [23]. The effect of long propagation delays and channel asymmetries caused by power control are issues that need to be addressed when considering application of these network routing protocols to UWA channels.

4. EVOLUTION OF UWA NETWORKS

Two types of applications have guided the evolution of underwater networks so far: gathering of environmental data and surveillance of an underwater area. In the first case, the network consists of several types of sensors, some mounted on fixed moorings and others mounted on moving vehicles. This type of network is called an *autonomous ocean sampling network* (AOSN), where the word "sampling" implies collecting the samples of oceanographic parameters such as temperature, salinity, and underwater currents. For surveillance applications, the network consists of a larger number of sensors, typically bottom-mounted or on slowly crawling robots, that can be quickly deployed, and whose task is to map a shallow-water area. In particular, mapping may focus on detection of warfare objects. An example of such a network, called *Seaweb*, will be described in more detail in Section 5.

An AOSN is formed by a number of autonomous underwater vehicles (AUVs), moorings, and surface buoys. The AUVs traverse an ocean area spanned by the network nodes (moorings and surface buoys) collecting scientific data. The coordination of the AUVs is handled from a central location, which can be either at one of the network nodes and/or on shore. AUVs relay key observations and their status to the central location. After evaluating the incoming data, the central location sends control signals to the AUVs through the network nodes. The acoustic communication between the AUVs and the network nodes is designed so that it does not require high data throughput. More complete data sets are transferred to the onshore control center through a radio channel when AUVs dock to a mooring. The control center is connected to a backbone such as the Internet, so that a scientist can reach the sampling network in real time. An important

limitation observed during the tests of the AOSN was the impossibility for the AUVs to instantly respond to commands due to the high-latency environment [28]. As a result of the highly variable acoustic channel, network connections to AUVs were occasionally lost. Therefore, some level of automation is needed in the AUVs to avoid disastrous events, which may occur if the last command sent to an AUV directs it toward an obstacle and the connection is lost.

A deep-water acoustic local-area network (ALAN) was deployed in Monterey Canyon, California, for long-term data acquisition and ocean monitoring from multiple ocean-bottom sources [1]. A centralized network topology was employed with a hub on the surface. The MAC protocol was based on TDMA, where time slots were determined adaptively on the basis of estimated latency. Because this protocol relies on correct estimation of the round-trip propagation times, any error in the estimation process decreases the throughput of the system by causing retransmissions.

The evolution of research on underwater acoustic networks has followed the usual layered architecture. Most work to date has been performed on the physical layer and multiple-access techniques. The data-link-layer protocols have been addressed to a lesser extent, and the work has only begun on the network layer and routing algorithms [5]. In all of these areas, the focus of research has been on adapting the well-known theoretical concepts to the requirements and constraints of the underwater acoustic channels.

Typically, packet transmission in a store-and-forward network is considered in most of the underwater acoustic networks. The design of automatic repeat request (ARQ) protocols is influenced by the long propagation times in underwater channels. Talavage et al. [29] proposed a shallow-water acoustic local-area network (S-ALAN) protocol, which is a modified version of ARPA-supported packet radio network (PRN) protocol. The S-ALAN differs from PRN in the routing algorithm and the data transmission medium. In contrast to PRN, which uses datagram switching over a single channel, S-ALAN employs virtual circuit switching using three separate channels (frequency bands) and selective-repeat ARQ. When a network node gathers enough information to send to the control center, it issues a request to set up a virtual circuit. When the setup request reaches the destination node, the destination node assigns transmit data, receive data, and acknowledgment channels for all the nodes in the virtual circuit and reports the final configuration back to the source node. The use of three separate channels enables the network to fully utilize the ARQ scheme.

A peer-to-peer communication protocol has been developed to control AUVs [30] based on carrier sense multiple access with collision avoidance (CSMA/CA). Since the CSMA/CA protocol relies on acknowledgments, the channel remains idle for an amount proportional to the round-trip propagation time. Because of the long propagation delays in UWA channels, this protocol has a low throughput. On the other hand, it is highly reliable.

A media access protocol proposed for shallow-water acoustic networks is presented in Ref. 10. The protocol is based on the MACA protocol and employs a stop-and-wait ARQ scheme. The RTS–CTS exchange is used to determine the channel conditions, and this information is used to set the acoustic modem parameters such as output power level. The details of this network are given in the following section.

The routing optimization problem for a shallow-water acoustic network is addressed in Ref. 31. The genetic algorithm based routing protocol tries to maximize the lifetime of the battery-powered network by minimizing the total energy consumption of the network. The minimum energy required to establish reliable communication between two nodes is used as the link distance metric. A master node collects the link cost information from the network nodes, determines optimum routes, and sends the routing information back to the nodes. The authors showed that the optimization algorithm favors multihop links at the expense of increased delay.

5. SEAWEB

Seaweb is an acoustic network for communications and navigation of deployable autonomous undersea systems [32]. The U.S. Navy incorporated Seaweb networking in the June 2001 Fleet Battle Experiment India (FBE-I). The Seaweb installation charted in Fig. 3 was part of the overall FBE-I joint forces architecture for command, control, communications, computers, intelligence, surveillance, and reconnaissance (C4ISR) providing wide-area connectivity, enhanced bandwidth, and reachback capability. Seaweb reliably supported asynchronous networking for an improved 688-class (688I) fast-attack nuclear submarine mobile node and two deployable autonomous distributed system (DADS) nodes. Two moored radio-acoustic communications (racom) buoy gateway nodes provided line-of-sight radio links to a shore station having terrestrial Internet connection to an antisubmarine warfare (ASW) command center (ASWCC) located ashore. In addition, 10 undersea repeater nodes highlighted the flexible architecture, indicating expandability and potential area coverage. The network performed reliably with no hardware failures and no lost transmissions. Seaweb supported Internet Protocol (IP) delivery of automated ASW contact reports from the DADS sensors to shore, and command and control (C2) on the IP backlink. Naval messages to and from the submarine via Seaweb protocols permitted assured access at tactical depth. The 15-node FBE-I Seaweb system extended Naval network-centric operations into the undersea battlespace. Analysts executed numerous communication and navigation tests, and proved that the FBE-I network design was overly conservative and could have supported even greater area coverage and traffic load.

5.1. Concept of Operations

Telesonar wireless acoustic links interconnect distributed undersea instruments, potentially integrating them as a unified resource. Seaweb is the realization of an undersea telesonar network [10] of fixed and mobile nodes, with various interfaces to manned command centers.

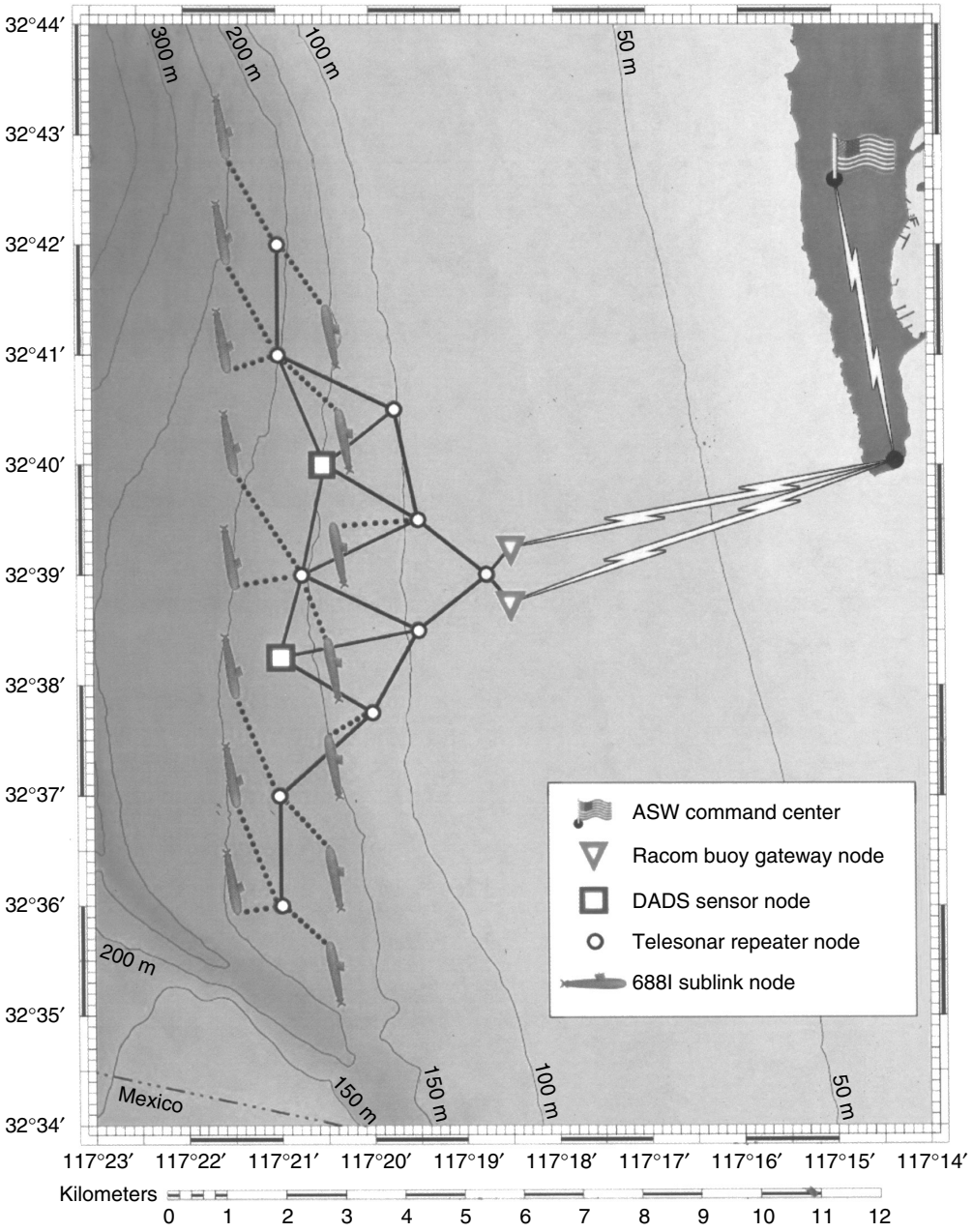


Figure 3. The FBE-I Seaweb network was a 14-node undersea grid. Two nodes were prototype deployable autonomous distributed system (DADS) sensors for littoral ASW, and two nodes were moored radio/acoustic communications (racom) gateway nodes. A mobile submarine node with sublink capacity had full interoperability with the Seaweb network. An ASW command center served as the ashore site. U.S. Navy personnel exercised the complete Seaweb installation for 4 days with high reliability and no component failures.

The architectural flexibility afforded by Seaweb wireless connections permits the network designer to allocate an arbitrary mix of node types with a node density and area coverage appropriate for the given telesonar propagation conditions and for the mission at hand. The concept of operations assumes that the majority of network nodes are inexpensive, autonomous, battery-limited devices deployed from submarine, ship, and aircraft, or from unmanned undersea vehicles (UUVs) and unmanned aerial vehicles (UAVs).

Seaweb networks support asynchronous data communications from autonomous nodes to command centers. On the backlink, Seaweb allows remote command and control of instruments associated with the autonomous nodes. Additionally, network activity supports acoustic navigation and geolocalization of undersea nodes as a natural byproduct of telesonar ranging signals. More generally, Seaweb networking permits wireless transmissions between member nodes in the network using established routes or via an intervening cellular node.

Seaweb enables future Naval capabilities in littoral ASW and undersea autonomous operations. A significant dual use of Seaweb is communication and navigation for oceanographic surveys and environmental assessment. Certainly, a major potential benefit of the technology is cross-system, cross-platform, cross-mission interoperability, providing enormous added value to otherwise solitary systems. For example, a UUV mobile node operating within a grid of fixed sensor nodes benefits from the established network topology for situational awareness, navigation, and communications via gateway nodes to distant command centers. Conversely, UUVs add value to the fixed grid for sensor deployment, search, survey, water-column sampling, popup racom gateway communications, and other functions.

The initial motivation for Seaweb is a requirement for wide-area surveillance in littoral waters. These sensors operate in 50- to 300-m waters with node spacing of 2–5 km. Sensor nodes generate concise ASW contact reports that Seaweb routes to a master node for field-level data fusion [33]. Primary network packets are contact reports with about 1000 information bits [34]. Sensor nodes asynchronously produce these packets at a variable rate dependent on the receiver operating characteristics (ROCs) for a particular sensor suite and mission. The master node communicates with manned command centers via encrypted gateway nodes such as a racom sea-surface buoy linked with space satellite networks. Following ad hoc deployments, the Seaweb network self-organizes, including node identification, clock synchronization on the order of 0.1–1.0 s, node geolocalization on the order of 100 m, assimilation of new nodes, and self-healing following node failures.

As a fixed grid of inexpensive interoperable sensor nodes and repeater nodes, this is the most fundamental Seaweb operating mode based on a stable topology that periodically adjusts itself to optimize overall network endurance and quality of service (QoS). The fixed Seaweb topology provides an underlying cellular network suited for supporting an AOSN [35], including communication and navigation for UUV mobile nodes. The cellular architecture likewise provides seamless connectivity for submarine operations at speed and depth in a manner not unlike those for terrestrial cellular telephone service for automobiles.

5.2. Developmental Approach

The concept of operations emphasizes simplicity, efficiency, reliability, and security, and these attributes therefore govern the design philosophy for Seaweb development. Research is advancing telesonar modem technology for reliable underwater signaling by addressing the issues of (1) adverse transmission channels, (2) asynchronous networking, (3) battery energy efficiency, (4) transmission security, and (5) cost.

Despite a concept of operations emphasizing simplicity, Seaweb is a multifaceted system, and its development is a grand challenge. The high cost of sea testing and the need for many prototype nodes motivate extensive engineering system analysis. Simulations using an optimized network engineering tool (OPNET) with simplified ocean acoustic propagation assumptions permit laboratory refinement of networking protocols [34] and initialization methods [36]. Meanwhile, controlled experimentation in actual ocean conditions incrementally advances telesonar signaling technology [37].

Seaweb development balances the desire for rapid increases in capability and the need for stable operation in support of applications that are themselves developmental. This balance is achieved by an annual cycle culminating with the late-summer Seaweb experiment (Seaweb '98, Seaweb '99, Seaweb 2000, etc.).

The objective of the annual Seaweb experiments is to exercise telesonar modems in networked configurations where various modulation and networking algorithms can be assessed. In the long term, the goal is to provide for a self-configuring network of distributed assets, with network links automatically adapting to the prevailing environment through selection of the optimum transmit parameters. A full year of hardware improvements and in-air network testing helps ensure that the incremental developments tested at sea will provide tractable progress and mitigate overall developmental risk. In preparation for Seaweb experimentation, multiple contributing projects conduct relevant research during the first three-quarters of the fiscal year. The fourth-quarter Seaweb experiment then implements and tests the results from these research activities with a concentration of resources in prolonged ocean experiments. The products of the annual Seaweb experiment are major capability upgrades for integrated Seaweb server software, telesonar modem Seaweb firmware, and telesonar modem hardware. The annual Seaweb experiment also transitions these upgrades into participating application systems. After the annual Seaweb experiment yields a stable level of functionality, the firmware product can be further exercised and refinements instituted during system testing and by spinoff applications throughout the year. For example, in year 2001, Seaweb 2000 technology enabled the March–June FRONT-3 ocean observatory on the continental shelf east of Long Island, New York [38]. These applications afford valuable long-term performance data that are not obtainable during Seaweb experiments when algorithms are in flux and deployed modems are receiving frequent firmware upgrades. At the conclusion of the Seaweb experiment, the upgraded Seaweb capability reaches stability suitable for use with the continuing development of the various application systems during the following year. Meanwhile, the annual cycle repeats, beginning with research and preparations for the next Seaweb experiment. And so, Seaweb capability increases in an incremental manner.

The Seaweb architecture of interest includes the physical layer, the media access control (MAC) layer, and the network layer. These most fundamental layers of communications functionality support higher layers, collectively identified here as the "client" layer. The client layer tends to be application-specific and is not the direct responsibility of telesonar modems or the Seaweb network.

5.3. Telesonar Modems and the Physical Layer

The U.S. Navy has been developing *telesonar* modems designed to function at low bit rates with high reliability and modest processing [39,40]. The basis for this approach is the need for low-cost, energy-efficient workhorse modems suitable for the development of networking technologies [41]. From an interoperability perspective, the low-bit-rate modem offers the lowest common denominator for cross-system networks that may include low-cost, expendable nodes [42,43]. As a pair of modems establishes a low-bit-rate link, they may adaptively negotiate higher-bit-rate modulations if warranted by favorable propagation and available processing resources.

The present telesonar modem [44] normally uses 5 kHz of acoustic bandwidth encompassing 120 discrete MFSK bins and 8 tracking bins [45]. A basic 1-of-4 MFSK modulation carries 2400 bps but lacks data protection and error correction coding (ECC). A constraint-length-9, rate-$\frac{1}{2}$ convolutional code very effectively corrects bit errors by representing binary information across multiple symbols; the rate-$\frac{1}{2}$ reduces throughput by a factor of 2. For protection against multipath-induced inter-symbol interference (ISI), the MFSK chip duration may be lengthened; the modem allows for a doubling from 25 to 50 ms, resulting in another factor of 2 reduction in bit rate. A "Doppler-tolerant" mode skips alternate MFSK bins to allow greater latitude for tracking Doppler shifts caused by node-to-node range rate. The Doppler-tolerant mode also increases robustness by doubling the acoustic energy per chip, but it also causes another halving in bit rate. Finally, a Hadamard MFSK modulation carries 6 Hadamard codewords of 20 tones each. Interleaving the codewords across the band increases immunity to frequency-selective fading, and Hadamard coding yields a frequency diversity factor of 5 for adverse channels having low or modest spectral coherence. Hadamard signaling is effective in channels having frequency-selective fading or narrowband noise. Any combination of the above mentioned options is possible. For FBE-I, a conservative modulation choice combined Hadamard MFSK, rate-$\frac{1}{2}$ convolutional coding, and 50-ms chip lengths to yield a net 300 bps information rate.

For all operational modes, receiving modems process the data noncoherently using a fixed-point TMS320C5410 DSP. Directional transducers can further enhance the performance of these devices [46,47]. The present telesonar modem includes provision for a watchdog function hosted aboard a microchip independent of the DSP. The watchdog resets the DSP on detection of supply voltage drops or on cessation of DSP activity pulses. The watchdog provides a high level of fault tolerance and permits experimental modems to continue functioning in spite of system errors. A watchdog reset triggers the logging of additional diagnostics for thorough troubleshooting after modem recovery.

Low-bandwidth, half-duplex, high-latency telesonar links limit Seaweb quality of service (QoS). Occasional outages from poor propagation or elevated noise levels can disrupt telesonar links [48]. Ultimately, the available energy supply dictates service life, and battery-limited nodes must be energy-conserving [49]. Moreover, Seaweb must ensure transmission security by operating with low bit-energy per noise-spectral-density (E_b/N_0) and by otherwise limiting interception by unauthorized receivers.

Spread-spectrum modulation is consistent with the desire for asynchronous multiple access to the physical channel using CDMA networking [50]. Nevertheless, the Seaweb concept does not exclude TDMA or FDMA methods and is in fact pursuing hybrid schemes suited to the physical-layer constraints. In a data transfer, for example, a concise asynchronous CDMA dialog could queue data packets for transmission during a time slot or within a frequency band such that multiaccess interference (MAI) collisions are avoided altogether.

At the physical layer, an understanding of the transmission channel is obtained through at sea measurements [51] and numerical propagation models [52]. Knowledge of the fundamental constraints on telesonar signaling translates into increasingly sophisticated modems [53]. DSP-based modulators and demodulators permit the application of modern digital communications techniques to exploit the unique aspects of the underwater channel. To aid understanding of telesonar performance, modems automatically log physical-layer diagnostics, including signal-to-noise ratio (SNR), automatic gain control (AGC), bit error rate (BER), and the number of corrected and uncorrected errors.

5.4. Handshake Protocols and the MAC Layer

Developmental Seaweb modem firmware implements the core features of a compact, structured protocol for secure, low-power, point-to-point, connectivity. The protocol efficiently maps MAC-layer functionality onto a physical layer based on channel-tolerant, 64-bit utility packets and channel-adaptive, arbitrary-length data packets. Seaweb firmware implements utility packet types using the basic Hadamard MFSK physical layer. These utility packet formats permit data transfers and node-to-node ranging. A richer set of available utility packets is being investigated with OPNET simulations prior to modem implementation, but seven core utility packets provided substantial networking capability for FBE-I.

The telesonar handshake protocol is suited to wireless half-duplex networking with slow propagation. Handshaking [20] asynchronously establishes adaptive telesonar links [54]. The initial handshake consists of the transmitter sending a request-to-send (RTS) packet and the receiver replying with a clear-to-send (CTS) packet. A busy signal (BSY) packet may be issued in response to an RTS when the receiver node decides to defer data reception in favor of other traffic. Following a successful RTS–CTS handshake, the data packet(s) are sent. This RTS–CTS round trip establishes the communications link and probes the channel to gauge optimal transmit power. Future firmware enhancements will support power control and the adaptive choice of data modulation method, with selection based on channel estimates derived from the RTS role as a probe signal. Telesonar links eventually will be environmentally adaptive [55], with provision for bidirectional asymmetry. Handshaking permits addressing, ranging, channel estimation, adaptive modulation, and power control.

The Seaweb 2000 core protocol implemented stop-and-wait ARQ scheme by providing either positive or negative acknowledgment of a data message. The choice of acknowledgment type depends on the traffic patterns associated with a particular network mission. Handshaking provided the means for resolving packet collisions automatically using retries from the transmitter or automatic repeat request (ARQ) packets from the receiver. For FBE-I, a purely negative acknowledgment was supported by the modem, implemented as an ARQ utility packet. At the client layer, the DADS client system supports positive acknowledgment through its IP implementation. Figure 4 illustrates the MAC layer protocol.

If two nodes send an RTS to each other, unnecessary retries may occur because both nodes will ignore the received RTS command. Each node will then wait for the other node to send a CTS for a timeout duration, and retransmit their RTS packet. This problem is solved by assigning priority to the packets that are directed towards the master node, as explained below (Fig. 5).

Assume that node A is a lower level node than node B; that is, Node A is the parent of B. Node A and node B both send RTS to each other. As a result of transmission delays, packets arrive to their destinations while both nodes are waiting for a CTS packet. When node B receives the RTS, it notices that the packet is from its own destination, node A. Node B checks whether node A is its parent or child. Since node A is its parent, node B has the priority and sends a CTS packet immediately. By that time, node A receives the RTS of node B, does the same check and decides that it should wait for node B to complete its data transmission, since node B is its child. Therefore, node A puts its own data packet into a queue and waits for the CTS packet of node B.

Future implementations of Seaweb firmware will retain the purely negative acknowledgment approach, as analysis has shown this to be the appropriate MAC layer implementation for long-latency, half-duplex links. For communications requiring positive acknowledgments, the Seaweb 2001 firmware includes provision for efficient delivery of a receipt (RCPT) utility packet from the destination node to the source node.

The RTS–CTS approach anticipates eventual implementation of adaptive modulation and secure addressing. The initiating node transmits a RTS waveform with a frequency-hopped, spread-spectrum (FHSS) [44] pattern or direct-sequence spread-spectrum (DSSS) [11] pseudo-random carrier uniquely addressing the intended receiver.

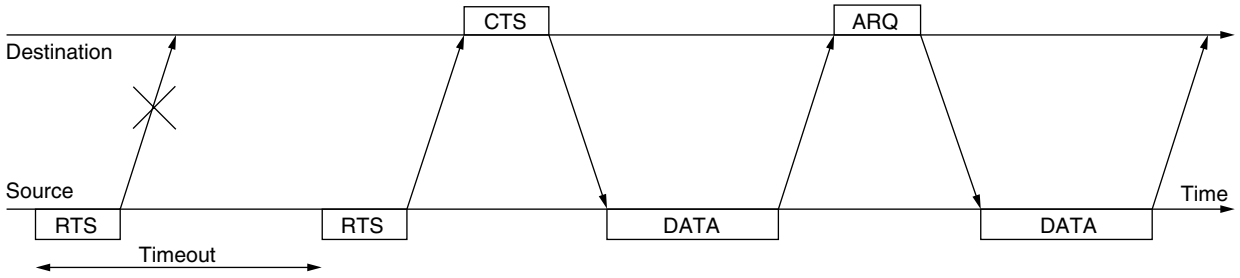


Figure 4. The source node starts the MAC layer handshake protocol by sending an RTS packet to the destination node. If the RTS packet is lost in the channel, the source node retransmits the RTS packet after a timeout duration equal to the round-trip time of an header-only packet (e.g., RTS, CTS, or ARQ), and calculated using the range information in the neighboring tables. When the destination node receives the RTS, it replies with a CTS packet. On receipt of the CTS packet by the source, the DATA packet, which contains a header and the information, is sent to the destination. The destination node issues an ARQ packet if it does not receive the DATA packet before a timeout occurs.

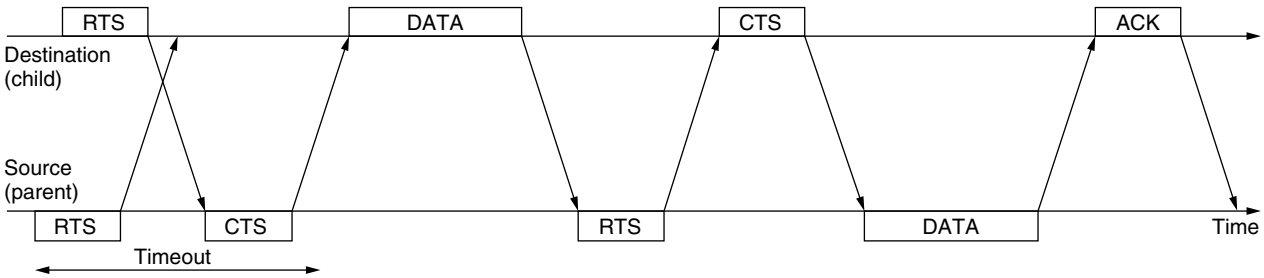


Figure 5. If two nodes send RTS packets to each other with some delay, both nodes receive the packets. When they recognize that the RTS is received from their destination, they check for the priority of the nodes. Higer-level nodes, or children, have higher priority. Therefore, the parent node sends CTS telling the child to send its data. When the parent receives the data, it replies with an RTS, which also act as an acknowledgments. Then the parent sends its data to the child. This example uses positive acknowledgments (ACK).

(Alternatively, the initiating node may transmit a universal code for broadcasting or when establishing links with unknown nodes.) The addressed node detects the request and awakens from an energy-conserving sleep state to demodulate. Further processing of the RTS signal can provide an estimate of the channel scattering function and signal excess. An adaptive power-control technique determines the source level that will deliver sufficient but not excessive SNR. The addressed node then acknowledges receipt with a FHSS or DSSS acoustic response. This CTS reply specifies appropriate modulation parameters for the ensuing message packets based upon the measured channel conditions. Following this RTS–CTS handshake, the initiating node transmits the data packet(s) with nearly optimal bit rate, modulation, coding, and source level.

5.5. Seaweb Server and the Network Layer

The Seaweb backbone is a set of autonomous, stationary nodes (e.g., deployable surveillance sensors, repeaters). Seaweb peripherals are mobile nodes (e.g., UUVs, including swimmers, gliders, and crawlers) and specialized nodes (e.g., bistatic sonar projectors). Seaweb gateways provide connections to command centers submerged, afloat, aloft, and ashore. Telesonar-equipped gateway nodes interface Seaweb to terrestrial, airborne, and space-based networks. For example, a telesonobuoy serves as a racom interface permitting satellites and maritime patrol aircraft to access submerged, autonomous systems. Similarly, submarines can access Seaweb with telesonar signaling through the underwater telephone band or other high-frequency sonar [56]. Seaweb provides the submarine commander digital connectivity at speed and depth with bidirectional access to all Seaweb-linked resources and distant gateways.

At the physical and MAC layers, adaptive modulation and power control are keys to maximizing both channel capacity (bps) and channel efficiency (bit-kilometer/joule). At the network layer, careful selection of routing is required to minimize transmit energy, latency, and net energy consumption, and to maximize reliability and security. Seaweb experimentation underscores the differences between acoustic networks and conventional networks. Limited power, small bandwidth, and propagation latencies dictate that the Seaweb network layer be simple and efficient. For compatibility with Seaweb networks, the higher client layers must utilize lookup tables, data compression, forward error correction, and data filtering to minimize packet sizes and retransmissions, and to avoid congestion at the network layer.

A very significant development was the introduction of the Seaweb server [57]. A Seaweb server resides at manned command centers and is the graphical user interface to the undersea network. It interprets, formats, and routes downlink traffic destined for undersea nodes. On the uplink, it archives incoming data packets produced by the network, retrieves the information for an operator, and provides Web-based read-only database access for client users. The server manages Seaweb gateways and member nodes. It monitors, displays, and logs the network status. The server manages the network routing tables and neighbor tables and ensures network interoperability. Seaweb '99 modem firmware permitted the server to remotely reconfigure routing topologies, a foreshadowing of future self-configuration and dynamic network control. The Seaweb server is a suite of software programs implemented under Linux on a laptop PC with a LabVIEW graphical user interface. A single designated "super" server controls and reconfigures the network.

Network supervisory algorithms can execute either at an autonomous master node or at the Seaweb server. Seaweb provides for graceful failure of network nodes, addition of new nodes, and assimilation of mobile nodes. Essential byproducts of the telesonar link are range measurement, range rate measurement, and clock synchronization. Collectively, these features will support network initialization, node localization, route configuration, resource optimization, and maintenance.

Node-to-node ranging employed a new implementation of a round-trip travel time measurement algorithm with 0.1-ms resolution linked to the DSP clock rate.

As a network analysis aid, all modems now include a data-logging feature. All output generated by the telesonar modem and normally available via direct serial connection is logged to an internal buffer. Thus, the behavior of autonomous nodes can be studied in great detail after recovery from sea. Seaweb 2000 firmware logs diagnostics related to channel estimation (SNR, multipath duration, range rate, etc.), demodulation statistics (BER, AGC, intermediate decoding results, power level, etc.), and networking (data packet source, data packet sink, routing path, etc.). For Seaweb applications, the data-logging feature can also support the archiving of data until such time that an adjacent node is able to download the data. For example, a designated sink node operating without access to a gateway node can collect all packets forwarded from the network and telemeter them to a command center when interrogated by a gateway (such as a ship arriving on station for just such a data download).

Since the network in consideration is an ad hoc network, an initialization algorithm is needed to establish preliminary connections autonomously. This algorithm is based on polling, and as such it guarantees connectivity to all the nodes that are acoustically reachable by at least one of their nearest neighbors. During initialization, the nodes create *neighbor tables*. These tables contain a list of each node's neighbors and a quality measure of their link, which can be the received SNR from the corresponding neighbor. The neighbor tables are then collected by the master node and a routing tree is formed. Table 2 is an example node table for node 3 of the network given in Fig. 6.

The master node decides on the primary (and secondary) routes to each destination, with routing optimization based on the genetic algorithm. Initialization ends when the master node sends primary routes to the nodes. The initialization algorithm provides either a single set of connections, or multiple connections between the nodes. Multiple connections are desirable to provide greater robustness to failures. A possible routing tree with backup routes for the network of Fig. 6 is given in Fig. 7.

Optimum routes are determined with the help of a genetic algorithm-based routing protocol [31]. The routing protocol tries to maximize the lifetime of the

Table 2. Node Table[a] of Node 3 for Network Given in Fig. 6

Neighbor ID	Range (km)	Output Power (dB)
1	6	−9 dB
2	7	−9 dB
4	9	−3 dB

[a] This table contains the ID of the nodes with which node 3 can communicate with a direct link. For each neighbor, the range of the node and the minimum output power required to communicate with that node are entered. The output power is in decibels with respect to the maximum output power of the modem. Because of the channel characteristics, nodes at different ranges may require the same amount of output power.

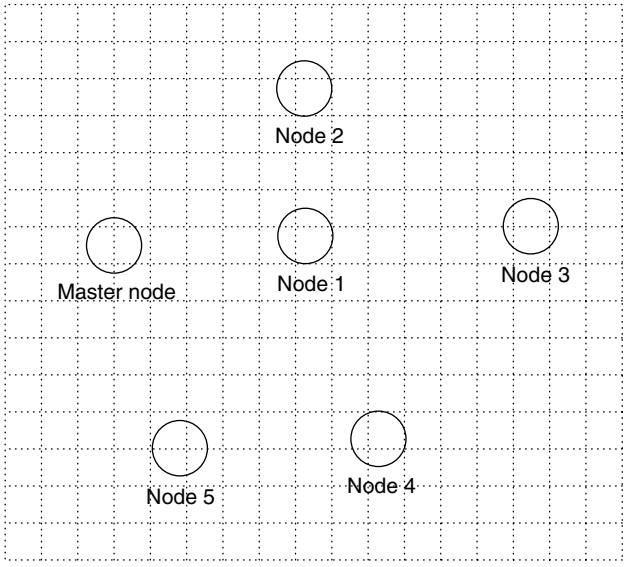


Figure 6. The network consists of a master node and five sensor nodes. The sensor nodes send information packets to the master node, which is the connection point of the network to a backbone. The master node can also send control packets to the sensor nodes.

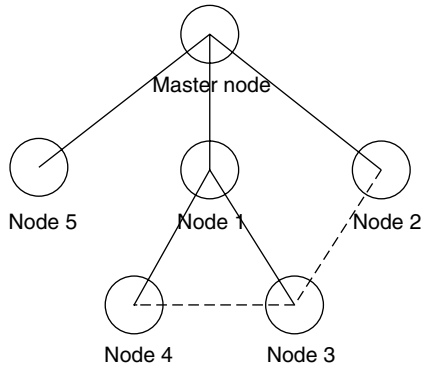


Figure 7. The routing tree is created by the master node. The master node is at the top of the tree. Sensor nodes are the leafs. Nodes 1, 2, and 5 are the children of the master node and form the first layer of the network. Node 1 is the parent of nodes 3 and 4, which form the second layer of the network. Dashed lines represent backup links that can be used in case of a failed link.

battery-powered network by minimizing the total energy consumption of the network. The minimum energy required to establish reliable communication between two nodes is used as the link distance metric. A master node collects the link cost information from the network nodes, determines optimum routes, and sends the routing information back to the nodes. The optimization algorithm favors multihop links at the expense of increased delay.

The performance of acoustic links between nodes can degrade, and even a link can be permanently lost as a result of node failure. In such cases, the network should be able to adapt itself to the changing conditions without interrupting the packet transfer. This robustness can be obtained by updating the routes periodically.

In the current implementation, the network tables are created manually at the Seaweb server with the help of the ranging packets. Also, the initial routes and updates are determined and reported to the network nodes manually through the Seaweb server.

6. CONCLUDING REMARKS

In this article, we presented an overview of basic principles and constraints in the design of reliable shallow-water acoustic networks that may be used for transmitting data from a variety of undersea sensors to onshore facilities. Major impediments in the design of such networks were considered, including (1) severe power limitations imposed by battery power; (2) severe bandwidth limitations; and (3) channel characteristics, including long propagation times, multipath, and signal fading. Multiple-access methods, network protocols, and routing algorithms were also considered.

Of the multiple-access methods considered, it appears that CDMA, achieved either by frequency hopping or by direct sequence, provides the most robust method for the underwater network environment. Currently under development are modems that utilize these types of spread-spectrum signals to provide the multiple-access capability to the various nodes in the network. Simultaneously with current modem development, there are several investigations on the design of routing algorithms and network protocols.

The design example of the shallow-water network employed in Seaweb embodies the power and bandwidth constraints that are so important in digital communication through underwater acoustic channels. As an information system compatible with low bandwidth, high latency, and variable quality of service, Seaweb offers a blueprint for the development of future shallow-water acoustic networks. Experimental data that will be collected in the near future will be used to assess the performance of the network and possibly validate a number of assumptions and tradeoffs included in the design. Over the next decade (at the time of writing), significant improvements are anticipated in the design and implementation of shallow water acoustic networks as more experience is gained through at-sea experiments and network simulations.

BIOGRAPHIES

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering, adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

Joseph A. Rice joined the Naval Postgraduate School's Physics Department in 2001 as the SPAWAR Systems Center San Diego Engineering Acoustics Chair. He is Principal Investigator for numerous ONR projects and tasks collectively performed as the Seaweb Initiative. He is the Technical POC for three SBIR contracts producing undersea acoustic modems, networks, and directional transducers. He has been a U.S. Navy research engineer since 1981, developing digital signal processing and numerical modeling concepts for solving undersea acoustics problems. He received the U.S. Navy Meritorious Civilian Service Award for experimental demonstration of undersea acoustic matched-field processing (MFP) during the Swellex acoustic propagation studies. He is interested in acoustic localization and navigation, having developed the prototype array element localization (PAEL) sonobuoy and associated processing for the Sonobuoy Thinned Random Array Program (STRAP). He holds a B.A. in Mathematics and a B.S. in Engineering Science, both from the University of Cincinnati. He holds an M.S.E.E. in Applied Ocean Science from UCSD. He is a member of IEEE and ASA.

Ethem M. Sozer received his B.S. and M.S. degrees in Electrical Engineering from the Middle East Technical University, Ankara, Turkey, in 1994 and 1997, respectively. He is working toward his Ph.D. in Electrical Engineering at Northeastern University, Boston. His research interests include iterative equalization techniques, underwater acoustic communications, and underwater acoustic networks. He is currently working for Delphi Communication Systems, Inc, Maynard, Maryland.

Milica Stojanovic graduated from the University of Belgrade, Belgrade, Yugoslavia, in 1988, and received the M.S. and Ph.D. degrees in Electrical Engineering from Northeastern University, Boston, in 1991 and 1993, respectively. She is a Principal Research Scientist at the Massachusetts Institute of Technology, and a Guest Investigator at the Woods Hole Oceanographic Institution. Her research interests include digital communications theory and statistical signal processing and their applications to wireless communication systems.

BIBLIOGRAPHY

1. J. Catipovic, D. Brady, and S. Etchemendy, Development of underwater acoustic modems and networks, *Oceanography* **6**: 112–119 (March 1993).
2. R. Conogan and J. P. Guinard, Observing operationally in situ ocean water parameters: The EMMA system, *Proc. OCEANS'98*, Nice, France, Sept. 1998, pp. 37–41.
3. J. Marvaldi et al., GEOSTAR — development and test of a communication system for deep-sea benthic stations, *Proc. OCEANS'98*, Nice, France, Sept. 1998, pp. 1102–1107.
4. M. Stojanovic, Recent advances in high-speed underwater acoustic communications, *IEEE J. Ocean. Eng.* **21**: 125–136 (April 1996).
5. D. B. Kilfoyle and A. B. Baggeroer, The state of the art in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **25**: 4–27 (2000).
6. J. Catipovic, Performance limitations in underwater acoustic telemetry, *IEEE J. Ocean. Eng.* **15**: 205–216 (July 1990).
7. M. Stojanovic, J. Catipovic, and J. Proakis, Phase-coherent digital communications for underwater acoustic channels, *IEEE J. Ocean. Eng.* **19**: 100–111 (1994).
8. A. Tannenbaum, *Computer Networks*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.

9. K. Pahlavan and A. H. Levesque, *Wireless Information Networks*, Wiley, New York, 1995.
10. E. M. Sozer, M. Stojanovic, and J. G. Proakis, Underwater acoustic networks, *IEEE J. Ocean. Eng.* **25**: 72–83 (Jan. 2000).
11. E. M. Sozer et al., Direct sequence spread spectrum based modem for under water acoustic communication and channel measurements, *Proc. OCEANS'99*, Nov. 1999.
12. Z. Zvonar, D. Brady, and J. A. Catipovic, Adaptive decentralized linear multiuser receiver for deep water acoustic telemetry, *J. Acoust. Soc. Am.* 2384–2387 (April 1997).
13. A. C. Chen, Overview of code division multiple access technology for wireless communications, *Proc. IECON'98*, 1998, Issue 24, pp. T15–T24.
14. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
15. L. Kleinrock and F. A. Tobagi, Carrier sense multiple access for packet switched radio channels, *Proc. ICC'74*, June 1974, pp. 21B-1–21B-7.
16. L. Kleinrock and F. A. Tobagi, Packet switching in radio channels, Part I: Carrier sense multiple access modes and their throughput-delay characteristics, *IEEE Trans. Commun.* 1400–1416 (1975).
17. F. A. Tobagi and L. Kleinrock, Packet switching in radio channels, Part II: The hidden terminal problem in carrier sense multiple access and busy tone solution, *IEEE Trans. Commun.* 1417–1433 (1975).
18. H. Takagi and L. Kleinrock, Correction to "Throughput analysis for persistent CSMA systems," *IEEE Trans. Commun.* 243–245 (1987).
19. V. Bharghavan, A. Deers, S. Shenker, and L. Zhang, MACAW: A media access protocol for wireless LAN's, *Proc. ACMSIGCOMM*, Aug. 1994, pp. 212–225.
20. P. Karn, MACA — a new channel access method for packet radio, *Proc. ARRL/CRRL Amateur Radio 9th Computer Network Conf.*, Sept. 1990.
21. J. Deng and Z. J. Haas, Dual busy tone multiple access (DBTMA): A new medium access control for packet radio networks, *Proc. IEEE 49th Vehicular Technology Conf.*, Houston, TX, May 1998, pp. 973–977.
22. D. B. Johnson, Routing in ad hoc networks of mobile hosts, *Proc. Workshop on Mobile Computing and Applications*, Dec. 1994, pp. 159–163.
23. J. Broch et al., A performance comparison of multi-hop wireless ad hoc network routing protocols, *Proc. ACM/IEEE Int. Conf. Mobile Computing and Networking*, Oct. 1998.
24. C. E. Perkins and P. Bhagwat, Highly dynamic destination sequence distance vector routing (DSDV) for mobile computers, *Proc. SIGCOMM'94*, Aug. 1994, pp. 234–244.
25. V. D. Park and M. S. Corson, A highly adaptive distributed routing algorithm for mobile wireless networks, *Proc. INFOCOM'97*, April 1997, pp. 1405–1413.
26. D. B. Johnson and D. A. Maltz, Protocols for adaptive wireless and mobile networking, *IEEE Pers. Commun.* (Feb. 1996).
27. C. E. Perkins, *Ad Hoc on Demand Distance Vector* (*AODV*) *Routing*, Internet-Draft, *draft-ietf-manet-aodv-00.txt*, Nov. 1997.
28. J. H. Kim et al., Experiments in remote monitoring and control of autonomous underwater vehicles, *Proc. OCEANS'96*, Fort Lauderdale, FL, Sept. 1996, pp. 411–416.
29. J. L. Talavage, T. E. Thiel, and D. Brady, An efficient store-and-forward protocol for a shallow-water acoustic local area network, *Proc. OCEANS'94*, Brest, France, Sept. 1994, pp. I883–I888.
30. S. M. Smith and J. C. Park, A peer-to-peer communication protocol for underwater acoustic communication, *Proc. OCEANS'97*, Oct. 97, pp. 268–272.
31. E. M. Sozer, M. Stojanovic, and J. G. Proakis, Initialization and routing optimization for ad hoc underwater acoustic networks, *Proc. Opnetwork'00*, Washington, DC, Aug. 2000.
32. J. A. Rice et al., Seaweb underwater acoustic nets, *SSC San Diego Biennial Review*, Aug. 2001.
33. E. Jahn, M. Hatch, and J. Kaina, Fusion of multi-sensor information from an autonomous undersea distributed field of sensors, *Proc. FUSION'99 Conf.*, Sunnyvale, CA, July 1999.
34. S. McGirr, K. Raysin, C. Ivancic, and C. Alspaugh, Simulation of underwater sensor networks, *Proc. IEEE OCEANS'99 Conf.*, Seattle, WA, Sept. 1999.
35. T. B. Curtin, J. G. Bellingham, J. Catipovic, and D. Webb, Autonomous oceanographic sampling networks, *Oceanography* **6**: 86–94 (1993).
36. J. G. Proakis, M. Stojanovic, and J. A. Rice, Design of a communication network for shallow-water acoustic modems, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1150–1159.
37. V. K. McDonald, J. A. Rice, M. B. Porter, and P. A. Baxley, Performance measurements of a diverse collection of undersea acoustic communication signals, *Proc. IEEE OCEANS'99 Conf.*, Seattle, WA, Sept. 1999.
38. D. L. Codiga, J. A. Rice, and P. S. Bogden, Real-time delivery of subsurface coastal circulation measurements from distributed instruments using networked acoustic modems, *Proc. IEEE OCEANS 2000 Conf.*, Providence, RI, Sept. 2000.
39. S. Merriam and D. Porta, DSP-based acoustic telemetry modems, *Sea Technol.* (May 1993).
40. D. Porta, DSP-based acoustic data telemetry, *Sea Technol.* (Feb. 1996).
41. J. A. Rice and K. E. Rogers, Directions in littoral undersea wireless telemetry, *Proc. TTCP Sympo. Shallow-Water Undersea Warfare*, Halifax, NS, Canada, Oct. 1996, Vol. 1, pp. 161–172.
42. M. D. Green, J. A. Rice, and S. Merriam, Underwater acoustic modem configured for use in a local area network, *Proc. IEEE OCEANS'98 Conf.*, Nice, France, Sept. 1998, Vol. 2, pp. 634–638.
43. M. D. Green, J. A. Rice, and S. Merriam, Implementing an undersea wireless network using COTS acoustic modems, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1027–1031.
44. M. D. Green, New innovations in underwater acoustic communications, *Proc. Oceanology International*, Brighton, UK, March 2000.
45. K. F. Scussel, J. A. Rice, and S. Merriam, A new MFSK acoustic modem for operation in adverse underwater channels, *paper presented at the OCEANS'97*, Halifax, NS, Canada, 1997.
46. N. Fruehauf and J. A. Rice, System design aspects of a steerable directional acoustic communications transducer for autonomous undersea systems, *Proc. OCEANS 2000 Conf.*, Providence, RI, Sept. 2000.

47. A. L. Butler, J. L. Butler, W. L. Dalton, and J. A. Rice, Multimode directional telesonar transducer, *Proc. IEEE OCEANS 2000 Conf.*, Providence, RI, Sept. 2000.

48. J. A. Rice, Acoustic signal dispersion and distortion by shallow undersea transmission channels, *Proc. NATO SACLANT Undersea Research Centre Conf. High-Frequency Acoustics in Shallow Water*, Lerici, Italy, July 1997, pp. 435–442.

49. J. A. Rice and R. C. Shockley, Battery-energy estimates for telesonar modems in a notional undersea network, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1007–1015.

50. M. Stojanovic, J. G. Proakis, J. A. Rice, and M. D. Green, Spread-spectrum methods for underwater acoustic communications, *Proc. IEEE OCEANS'98 Conf.*, Nice, France, Sept. 1998, Vol. 2, pp. 650–654.

51. V. K. McDonald and J. A. Rice, Telesonar testbed advances in undersea wireless communications, *Sea Technol.* **40**(2): 17–23 (Feb. 1999).

52. P. A. Baxley, H. P. Bucker, and J. A. Rice, Shallow-water acoustic communications channel modeling using three-dimensional Gaussian beams, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 1022–1026.

53. M. B. Porter, V. K. McDonald, J. A. Rice, and P. A. Baxley, Relating the channel to acoustic modem performance, *Proc. European Conf. Underwater Acoustics*, Lyons, France, July 2000.

54. J. A. Rice and M. D. Green, Adaptive modulation for undersea acoustic modems, *Proc. MTS Ocean Community Conf.*, Baltimore, MD, Nov. 1998, Vol. 2, pp. 850–855.

55. J. A. Rice, V. K. McDonald, M. D. Green, and D. Porta, Adaptive modulation for undersea acoustic telemetry, *Sea Technol.* **40**(5): 29–36 (May 1999).

56. J. A. Rice, Telesonar signaling and seaweb underwater wireless networks, *Proc. NATO Sympo. New Information Processing Techniques for Military Systems*, Istanbul, Turkey, Oct. 9–11 2000.

57. C. L. Fletcher, J. A. Rice, R. K. Creber, and D. L. Codiga, Undersea acoustic network operations through a database-oriented server client interface, *Proc. IEEE OCEANS 2001 Conf.*, Waikiki, HI, Nov. 2001.

SHELL MAPPING

HENRY K. KWOK
Cisco Systems, Inc.
San Jose, California

DOUGLAS L. JONES
University of Illinois at Urbana — Champaign
Urbana, Illinois

1. CONSTELLATION SHAPING

Constellation shaping [1,2] is a technique that improves the efficiency of high-rate digital communications by reducing the average power without compromising the data rate or bit error rate. It can provide moderate gain (called *shaping gain*) of up to 1.53 dB on top of any coding gain (such as by using convolutional codes) with modest complexity. For instance, it is used in the V.34 high-rate analog voiceband modem standard to reduce the average power of the QAM constellation.

The concept and purpose of shaping is easily illustrated by the following simple example. Consider two 256-QAM constellations (see Figs. 1 and 2). Both constellations have the same number of distinct points, allowing 8 bits to be transmitted with each symbol. The error rate is primarily a function of the distance between constellation points, so it is virtually identical for both constellations. However, the average energy is reduced from 170 to 162.75 by choosing

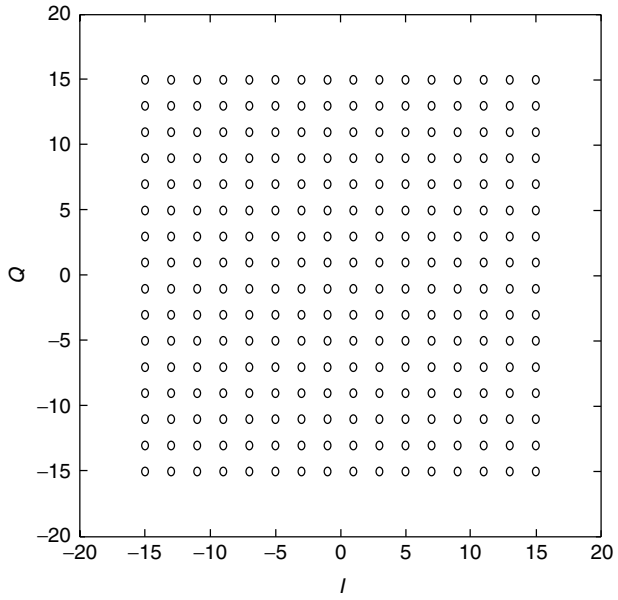


Figure 1. The original (unshaped) 256-QAM constellation has an average energy of 170.

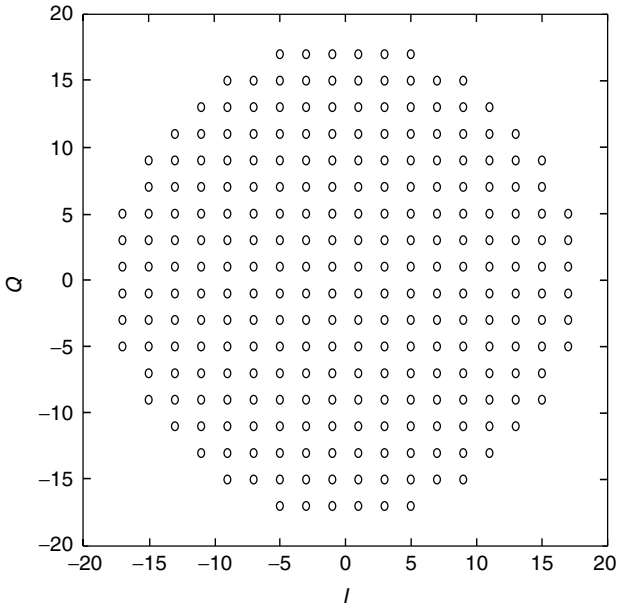


Figure 2. The shaped 256-QAM constellation has an average energy of 162.75 — a reduction of 0.379 dB.

a circular boundary instead of a square one. The reduction in average power is 0.379 dB without compromising the data rate or bit error rate.

Higher-dimensional constellations can provide even more average power reduction (up to 1.53 dB) [1–3]. For example, V.34 modems use a 16-dimensional constellation that provides about 1 dB of average power reduction. The idea is essentially the same as for two dimensions; conceptually, a block of N successive complex-valued symbols are grouped into a single $2N$-dimensional "hypersymbol," and a "hyperconstellation" in this space encodes a large number of bits simultaneously. Use of an approximately hyperspherical constellation with the same distance between constellation points reduces the average energy while preserving the same noise margin.

The primary practical challenge of shaping is to efficiently index the very large number of points in a high-dimensional, shaped constellation. Beyond two dimensions, simple tabulation becomes infeasible. Shell mapping is an efficient algebraic technique for performing this mapping that allows shaping to be used in practical communication systems. Shell mapping groups constellation points of roughly equal energy from a quadrature (QAM) constellation into concentric rings, or shells. A block of successive QAM symbols jointly code a large number of bits, which are divided into two groups: those mapped to the particular constellation points within the chosen shell for each QAM symbol, and another set of bits that indicates the combination of shells used for that block. For example, a system encoding an average rate of 6 bits per QAM symbol might use a block of eight symbols with 16 constellation points per shell; in this case, $8 \times 4 = 32$ bits would be encoded in the specific constellation point used within the shells selected for the eight successive QAM symbols, and the remaining 16 bits map to the 2^{16}th lowest-energy combinations of shells for the block. The total rate per eight symbols is thus $32 + 16 = 48$, for an average rate of 6 bits per QAM symbol. Shell mapping is a mathematical algorithm that efficiently identifies and indexes the lowest-energy sets of shells in a block of several successive QAM symbols using a technique based on combinatorics, generating functions, and finite-field algebra.

2. SHELL MAPPING

Shell mapping provides an efficient way to index lattice points inside a regular solid (such as a hypersphere or a hyperdiamond). It was first used in reducing the average power of QAM-based modems, in which it divides a QAM constellation into concentric rings with equal areas. A certain number of input bits are used to select the rings from which the constellation points will be chosen. A method due to Laroia [3] is used in the V.34 modems. We present a general version of this algorithm from a combinatorial viewpoint.

We use superscript $^{(N)}$ to indicate the dimension of a particular quantity. (The superscript may be suppressed if the dimension is 1.) Bold face represents a vector or a collection of lower-dimensional quantities. A subscript is used to differentiate different quantities of the same dimensionality. For example, the ith ring index is denoted as s_i [or $s_i^{(1)}$], and the N-element combination is denoted as $\mathbf{s}^{(N)} = (s_1, s_2, \ldots, s_N)$.

2.1. Generating Function

In combinatorics [4,5], a generating function is defined as a formal series

$$g(x) = a_0 + a_1 x + \cdots + a_N x^N \tag{1}$$

Generating functions are often used to solve problems involving a collection of objects with costs. We illustrate the method of generating functions with an example.

Example 1. How many ways are there to add three integers from the set of $\{0, 1, 2, 3\}$ such that the sum equals C? To solve this via the generating function, we first define

$$g^{(1)}(x) = 1 + x + x^2 + x^3 \tag{2}$$

The power of each coefficient represents the value of the first number. Then, $g^{(3)}(x) = [g^{(1)}(x)]^3 = 1 + 3x + 6x^2 + 10x^3 + 12x^4 + 12x^5 + 10x^6 + 6x^7 + 3x^8 + x^9$ represents the generating function of the three-integer sum. There is one three-integer combination that adds up to 0 (or 9), three combinations that add up to 1 (or 8), and so on. Table 1 lists the combinations that result in those costs and the corresponding terms in the generating function.

The exponent of the terms indicates the *cost* of selecting a certain object, and the coefficients of the terms denote the *number* of combinations that achieve a particular cost. To compute the N-element generating function, one simply multiplies the generating polynomial N times and locates the coefficient for the term representing the desired cost.

Table 1. Each Term in the Generating Function and Its Associated Inputs and Outputs

Terms in $g^{(3)}(x)$	Combinations
1	(0, 0, 0)
$3x$	(1, 0, 0), (0, 1, 0), (0, 0, 1)
$6x^2$	(1, 1, 0), (0, 1, 1), (1, 0, 1), (2, 0, 0), (0, 2, 0), (0, 0, 2)
$10x^3$	(1, 1, 1), (2, 1, 0), (2, 0, 1), (0, 2, 1), (1, 2, 0), (0, 1, 2), (1, 0, 2), (3, 0, 0), (0, 3, 0), (0, 0, 3)
$12x^4$	(0, 1, 3), (0, 3, 1), (3, 0, 1), (1, 1, 2), (1, 2, 1), (2, 1, 1), (1, 0, 3), (1, 3, 0), (3, 1, 0), (2, 2, 0), (2, 0, 2), (0, 2, 2)
$12x^5$	(3, 2, 0), (3, 0, 2), (0, 3, 2), (2, 2, 1), (2, 1, 2), (1, 2, 2), (2, 3, 0), (2, 0, 3), (0, 2, 3), (1, 1, 3), (1, 3, 1), (3, 1, 1)
$10x^6$	(2, 2, 2), (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1), (3, 3, 0), (3, 0, 3), (0, 3, 3)
$6x^7$	(3, 3, 1), (3, 1, 3), (1, 3, 3), (3, 2, 2), (2, 3, 2), (2, 2, 3)
$3x^8$	(3, 3, 2), (3, 2, 3), (2, 3, 3)
x^9	(3, 3, 3)

2.2. Indexing Algorithm for Sets with Integral Summable Cost

The method of generating functions has been used to solve numerous counting problems. However, for this article, we focus on a specific class of problems.

Let S be a set of objects. Let each element $s \in S$ have an integral cost of $|s|$. Denote the cost of an N-element combination, $\mathbf{s}^{(N)} = (s_1 \cdots s_N) \in S^N$, as

$$|\mathbf{s}^{(N)}| = \sum_{i=1}^{N} |s_i| \tag{3}$$

It is important that the cost of an element take on integral values and that the cost of an N-element combination be the sum of the cost of individual elements. Equation (3) has the same form as the l_1-norm and often leads to geometrical interpretation.

Example 2. In the previous example, $S = \{0, 1, 2, 3\}$ and $|s| = s$. Therefore

$$|\mathbf{s}^{(N)}| = \sum_{i=1}^{N} |s_i| = \|\mathbf{s}^{(N)}\|_1 \tag{4}$$

Graphically, $\|\mathbf{s}^{(N)}\|_1 = c$ depicts a hyperplane in the all-positive quadrant.[1] We denote

$$S_c^{(N)} = \{\mathbf{s}^{(N)} \in S^N : |\mathbf{s}^{(N)}| = c\} \tag{5}$$

We illustrate the first four hyperplanes, $S_0^{(3)}$, $S_1^{(3)}$, $S_2^{(3)}$, and $S_3^{(3)}$, with the lattice points that belong to each hyperplane shown in Fig. 3. The lattice points on $S_c^{(3)}$ represent three-integer combinations that sum up to c. Each layer corresponds to a term in $g^{(3)}(x)$.

We assume that it is desirable to select N-element combinations within a certain range of costs. From this point on, we assume (without loss of generality) that the cost is nonnegative. This assumption implies that the generating functions will consist only of terms x^n where $n \in \{0\} \cup \mathbb{Z}^+$.

[1] Strictly speaking, the set defined is not a hyperplane, but rather a set of lattice points that lie on a common hyperplane. Throughout this article; however, we loosely use the term "hyperplane" to denote a set of lattice points on a common hyperplane.

2.3. Counting the Sets

First, we count the number of combinations that satisfy the cost constraint. To that end, we define the one-element generating function

$$g^{(1)}(x) = \sum_{i=\min_{s\in S}|s|}^{\max_{s\in S}|s|} a_i x^i \tag{6}$$

where a_i represents the number of elements in S that have the cost of i. Next, we compute the N-element generating function

$$g^{(N)}(x) = [g^{(1)}(x)]^N = \sum_i a_i^{(N)} x^i \tag{7}$$

where $a_i^{(N)}$ represents the number of N-element combinations in S^N that have the cost of i. To seek the number of combinations within a cost range, we can sum the coefficients over the range $C_l \le |\mathbf{s}^{(N)}| \le C_h$ as

$$K = \sum_{i=C_l}^{C_h} a_i^{(N)} \tag{8}$$

2.4. Mapping Algorithm

A mapping procedure from an input value to an N-element combination is needed in an actual communication system. Suppose that there are K N-element combinations that satisfy the cost constraint. Let R be an input integer from the set $I^{(N)} = \{i \in \mathbb{Z} : 0 \le i < K\}$, and let the set of the K N-element combinations be $S^{(N)} = \{\mathbf{s} \in S^N : C_l \le |\mathbf{s}| \le C_h\}$. The goal is to find a bijective mapping, $\mathbf{f} : I^{(N)} \to S^{(N)}$. The basic approach is to decompose the N-dimensional problem into multiple lower-dimensional problems. Specifically, Laroia [3] provides a way to decompose the N-dimensional problem into two $\frac{N}{2}$-dimensional problems and another way to decompose the problem into a 1D (one-dimensional) problem and an $(N-1)$-dimensional problem. Arbitrary decompositions

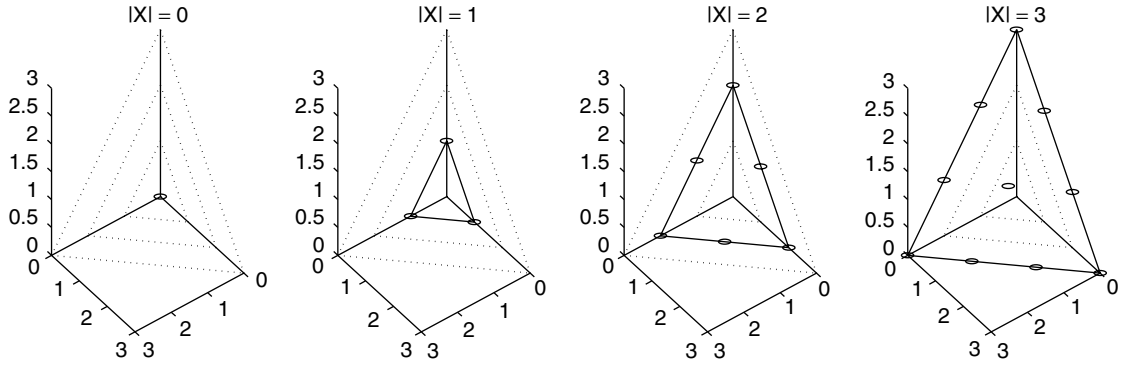


Figure 3. The layering of the generating function.

offering different trade-offs between computational cost and memory requirements can be found in Ref. 6.

2.5. Identifying the Cost of the Combination

First, we partition $S^{(N)}$ into a set of hyperplanes, $\{S_c^{(N)}\}_{c=C_l}^{C_h}$. Since these hyperplanes are parallel, they are also disjoint.[2] Thus, the $S^{(N)}$ set is partitioned into

$$S^{(N)} = S_{C_l}^{(N)} \cup S_{C_l+1}^{(N)} \cup \cdots \cup S_{C_h}^{(N)} \tag{9}$$

Next, we associate ranges of integers from $I^{(N)}$ to these N-dimensional hyperplanes. This is done by successively assigning the set of $a_c^{(N)}$ lowest available input integers in $I^{(N)}$ to the hyperplane of cost c, $S_c^{(N)}$. At the end of this process, we associate each input $R \in I^{(N)}$ to one of the $S_c^{(N)}$. If the input R is associated with $S_c^{(N)}$, we define the N-D *cost* of the input R as

$$C^{(N)} \triangleq |\mathbf{f}(R)| = c \tag{10}$$

This is called the N-dimensional cost because all N-element combinations $\mathbf{s}^{(N)}$ from this hyperplane have the cost $|\mathbf{s}^{(N)}|$ of $C^{(N)}$. Next, we partition the input set into

$$I_c^{(N)} = \{r \in I^{(N)} : \mathbf{f}(r) \in S_i^{(N)}\} \tag{11}$$

In other words, $I_c^{(N)}$ is the subset $I^{(N)}$ that is associated with $S_c^{(N)}$. Since each input is associated to only one hyperplane, the collection of $I_c^{(N)}$'s is disjoint, and we can write

$$I^{(N)} = \cup_c I_c^{(N)} \tag{12}$$

From the example above, it is easy to see that each $I_c^{(N)}$ consists of a range of consecutive integers. We define the N-dimensional *residue* as

$$R^{(N)} = R - \min I_{C^{(N)}}^{(N)} \tag{13}$$

Physically, $R^{(N)}$ uniquely indices a combination in $S_{C^{(N)}}^{(N)}$. (The uniqueness can be shown from the fact that $0 \le R^{(N)} < a_{C^{(N)}}^{(N)}$.) So far, we have created a mapping $(g_c, g_s) : I^{(N)} \to \mathbb{Z} \times \mathbb{Z}$ that maps an input to an N-dimensional cost and an N-dimensional residue. The cost selects the hyperplane $S_{C^{(N)}}^{(N)}$, and the residue points to an N-element combination in $S_{C^{(N)}}^{(N)}$.

Example 3. From Example 2, we have $N = 3$, $S = \{0, 1, 2, 3\}$, $|s| = s$, $C_l = 0$, and $C_h = 3$. Also, let the input $R = 14$. We have

$$K = \sum_{i=C_l}^{C_h} a_i^{(N)} = \sum_{i=0}^{3} a_i^{(3)} \tag{14}$$

$$= \begin{array}{ccccccc} 1 & + & 3 & + & 6 & + & 10 \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ x^0 & & x^1 & & x^2 & & x^3 \end{array} \tag{15}$$

$$= 20 \tag{16}$$

[2] Again, what we mean here is that since the two sets lie on two parallel hyperplanes, there is no element that belongs to both sets.

We associate the range of input to the subset of outputs as

$$\begin{aligned} I_0^{(3)} &= \{0\} \to S_0^{(3)} \\ I_1^{(3)} &= \{1, 2, 3\} \to S_1^{(3)} \\ I_2^{(3)} &= \{4, 5, \ldots, 9\} \to S_2^{(3)} \\ I_3^{(3)} &= \{10, 11, \ldots, 19\} \to S_3^{(3)} \end{aligned} \tag{17}$$

From (17), we find that $R = 14$ belongs to the range of 10–19. This means $C^{(3)} = 3$ and $R^{(3)} = 14 - 10 = 4$ [or we are now considering the fifth combination in the hyperplane of $|\mathbf{s}^{(3)}| = C^{(3)} = 3$; see Fig. 3].

2.6. Decomposition of the Problem

We decompose the N-dimensional problem into a P-dimensional problem and a Q-dimensional problem. The idea is to split $(C^{(N)}, R^{(N)})$ into $(C^{(P)}, R^{(P)})$, and $(C^{(Q)}, R^{(Q)})$. The obvious constraint is that $P + Q = N$.

Now, we perform the actual decomposition of the problem. We break up the N-dimensional hyperplane into the union of a collection of Cartesian product sets of P-dimensional and Q-dimensional hyperplanes, where $Q = N - P$.[3] For $\forall c \in \mathbb{Z}$, we can write

$$\begin{aligned} S_c^{(N)} &= (S_0^{(P)} \times S_c^{(Q)}) \cup (S_1^{(P)} \times S_{c-1}^{(Q)}) \cup \cdots \cup (S_c^{(P)} \times S_0^{(Q)}) \\ &= \bigcup_{i=0}^{c} (S_i^{(P)} \times S_{c-i}^{(Q)}) \end{aligned} \tag{18}$$

Since all product sets on the right side are disjoint, we can relate the size of these sets as follows:

$$\begin{aligned} |S_c^{(N)}| &= |S_0^{(P)}| \cdot |S_c^{(Q)}| + |S_1^{(P)}| \cdot |S_{c-1}^{(Q)}| + \cdots + |S_c^{(P)}| \cdot |S_0^{(Q)}| \\ &= \sum_{i=0}^{c} |S_i^{(P)}| \cdot |S_{c-i}^{(Q)}| \end{aligned} \tag{19}$$

We may also reach the following conclusion by noting that $a_i^{(n)} = |S_i^{(n)}|$ and

$$a_c^{(N)} = \sum_{i=0}^{c} a_i^{(P)} a_{c-i}^{(Q)} \tag{20}$$

In other words, (19) simply states the rule for evaluating the coefficients of $g^{(N)}(x)$ from the coefficients of $g^{(P)}(x)$ and $g^{(Q)}(x)$. Now, we partition the hyperplane $S_c^{(N)}$ into a collection of Cartesian product sets of lower-dimensional hyperplanes. Meanwhile, we partition the corresponding input range, $I_c^{(N)}$, similar to the way described above. The $|S_0^{(P)} \times S_c^{(Q)}| = |S_i^{(P)}| \cdot |S_{c-i}^{(Q)}|$ lowest available integers in $I_c^{(N)}$ are assigned to the Cartesian-product set, $S_i^{(P)} \times S_{c-i}^{(Q)}$. This range of input is denoted as $I_{i,c-i}^{(N)}$. Once again, we have

$$I_c^{(N)} = \bigcup_{i=0}^{c} I_{i,c-i}^{(N)} \tag{21}$$

[3] As mentioned earlier, usually $P = N - P = N/2$. However, for the sake of visual illustration, we choose N to be 3, which is not a power of 2. The shell mapper still operates properly; however, it is less efficient.

Once we decide that an input R is within $I^{(N)}_{i,C^{(N)}-i}$ and is associated with $S^{(P)}_i \times S^{(Q)}_{C^{(N)}-i}$, we have split the cost into

$$C^{(P)} = |\mathbf{s}^{(P)}| = i \tag{22}$$

$$C^{(Q)} = |s^{(Q)}| = C^{(N)} - i \tag{23}$$

In addition, we need to adjust the residue so that it now indexes a combination in $S^{(P)}_i \times S^{(Q)}_{C^{(N)}-i}$ instead of $S^{(N)}_{C^{(N)}}$. This is done by subtracting from $R^{(N)}$ the smallest value in $Z^{(N)}_{i,C^{(N)}-i}$. Now a residue of 0 indicates the first combination in $S^{(P)}_i \times S^{(Q)}_{c-i}$. We denote the new residue as

$$R^{(N)}_{i,c-i} = R^{(N)} - (\min I^{(N)}_{i,c-i} - \min I^{(N)}_c) = R - \min I^{(N)}_{i,c-i} \tag{24}$$

where $c = C^{(N)}$ in our specific case.

Example 4. Continuing with our previous example of $S = \{0, 1, 2, 3\}$, $N = 3$, $C_l = 0$, $C_h = 3$, $K = 20$, and $R = 14$, recall that $(C^{(3)}, R^{(3)}) = (3, 4)$. We now search for the Cartesian product set with which the input is associated. Let $P = 2$, and $Q = N - P = 1$. We associate the input range $I^{(3)}$ to the Cartesian product sets as

$$\begin{aligned} I^{(3)}_{0,3} &= \{10\} \rightarrow S^{(2)}_0 \times S^{(1)}_3 \\ I^{(3)}_{1,2} &= \{11, 12\} \rightarrow S^{(2)}_1 \times S^{(1)}_2 \\ I^{(3)}_{2,1} &= \{13, 14, 15\} \rightarrow S^{(2)}_2 \times S^{(1)}_1 \\ I^{(3)}_{3,0} &= \{16, 17, 18, 19\} \rightarrow S^{(2)}_3 \times S^{(1)}_0 \end{aligned} \tag{25}$$

We search in the following order: $S^{(2)}_0 \times S^{(1)}_3$, $S^{(2)}_1 \times S^{(1)}_2$, and then $S^{(2)}_2 \times S^{(1)}_1$. The search sequence is graphically depicted in Fig. 4. We find that $R = 14$ belongs to $I^{(3)}_{2,1}$. Thus, the two costs are $C^{(2)} = 2$, $C^{(1)}_3 = 1$.

In order to complete the decomposition into two lower-dimensional problems, we need two residues, $R^{(P)}$ and $R^{(Q)}$, for the two costs, $C^{(P)}$ and $C^{(Q)}$. Observe that there are several constraints. First, $R^{(P)}$ (or $R^{(Q)}$) is an index to a P-element (or Q-element) combination in $S^{(P)}_{C^{(P)}}$ (or $S^{(Q)}_{C^{(Q)}}$). Thus, $R^{(P)}$ and $R^{(Q)}$ must satisfy the following inequalities:

$$0 \le R^{(P)} < |S^{(P)}_{C^{(P)}}| \tag{26}$$

$$0 \le R^{(Q)} < |S^{(Q)}_{C^{(Q)}}| \tag{27}$$

In addition, the mapping of $|S^{(P)}_{C^{(P)}}| \cdot |S^{(Q)}_{C^{(Q)}}|$ values of $R^{(N)}_{C^{(P)},C^{(Q)}}$ to $(R^{(P)}, R^{(Q)})$ must be bijective. One such mapping is

$$\mathbf{g}_{|S^{(P)}_{C^{(P)}}|}(R^{(N)}) = \left(R^{(N)}_{C^{(P)},C^{(Q)}} \bmod |S^{(P)}_{C^{(P)}}|, \left\lfloor \frac{R^{(N)}_{C^{(P)},C^{(Q)}}}{|S^{(P)}_{C^{(P)}}|} \right\rfloor \right) \tag{28}$$

This is the same as arranging the $|S^{(P)}_{C^{(P)}}| \cdot |S^{(Q)}_{C^{(Q)}}|$ values of $R^{(N)}$ into an $|S^{(P)}_{C^{(P)}}|$-column $\times$ $|S^{(Q)}_{C^{(Q)}}|$-row array. Then, the column and row index of each array element are taken as $R^{(P)}_1$ and $R^{(Q)}$, respectively. Now, we have transformed $(C^{(N)}, R^{(N)})$ into $(C^{(P)}, R^{(P)})$ and $(C^{(Q)}, C^{(Q)})$. We may recursively apply the procedure to each cost–residue pair. Eventually, the dimension of the cost (and residue) will become 1. If $a^{(1)}_{C^{(1)}} = 1$, we are guaranteed that $R^{(1)} = 0$, signifying the first (and only) element in S with the cost $C^{(1)}$. If $a^{(1)}_{C^{(1)}} > 1$, it is possible that $R^{(1)} > 0$. In that case, we must resort to a lookup table to find the specific element in S.

Example 5. To conclude this section, we finish the shell mapping of the input $R = 14$. Note that $R = 14$ is the second combination ($R^{(3)}_{C^{(2)},C^{(1)}} = 1$) in $S^{(2)}_{C^{(2)}} \times S^{(1)}_{C^{(1)}}$, so the lower-dimensional residues should be

$$R^{(2)} = 1 \bmod 3 = 1 \tag{29}$$

$$R^{(1)} = \lfloor \tfrac{1}{3} \rfloor = 0 \tag{30}$$

We perform the same decomposition procedure on the cost-residue pair of $(C^{(2)}, R^{(2)}) = (2, 2)$. We know that $S^{(2)}_{C^{(2)}} = S^{(2)}_2 = (S^{(1)}_0 \times S^{(1)}_2) \cup (S^{(1)}_1 \times S^{(1)}_1) \cup (S^{(1)}_2 \times S^{(1)}_0)$. Since $S^{(1)}_i$ is a simple point in one dimension, $S^{(1)}_i \times S^{(1)}_j$ is a 2D point. We can also derive this fact from $|S^{(1)}_i \times S^{(1)}_j| = |S^{(1)}_i| \cdot |S^{(1)}_j| = 1 \cdot 1 = 1$. In addition, we know from the generating function $g^{(2)}(x)$ that $|S^{(2)}_{C^{(2)}}| = 3$. On the other hand, from Eq. (19), we have

$$\begin{aligned} |S^{(2)}_{C^{(2)}}| &= |S^{(1)}_0 \times S^{(1)}_2| + |S^{(1)}_1 \times S^{(1)}_1| + |S^{(1)}_2 \times S^{(1)}_0| \\ &= |S^{(1)}_0| \cdot |S^{(1)}_2| + |S^{(1)}_1| \cdot |S^{(1)}_1| + |S^{(1)}_2| \cdot |S^{(1)}_0| \\ &= 1 + 1 + 1 = 3 \end{aligned} \tag{31}$$

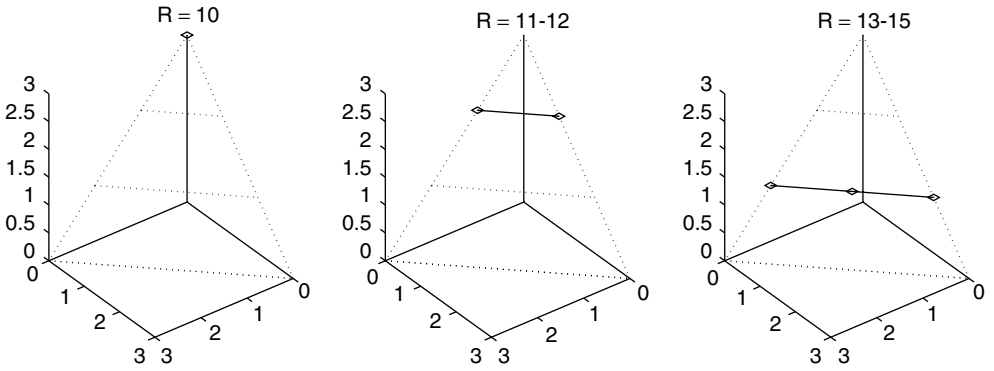


Figure 4. Decomposition of $S^{(3)}_3$ into a collection of Cartesian-product sets.

The corresponding input partitioning is

$$\begin{aligned} I_{0,2}^{(2)} &= \{13\} \rightarrow S_0^{(1)} \times S_2^{(1)} \\ I_{1,1}^{(2)} &= \{14\} \rightarrow S_1^{(1)} \times S_1^{(1)} \\ I_{2,0}^{(2)} &= \{15\} \rightarrow S_2^{(1)} \times S_0^{(1)} \end{aligned} \tag{32}$$

Thus, we have three Cartesian product sets, each with only one element. Since we are seeking the second combination ($R^{(2)} = 1$), we split the 2D cost $C^{(2)}$ into

$$C_1^{(1)} = 1 \tag{33}$$

$$C_2^{(1)} = 2 - 1 = 1 \tag{34}$$

The residue $R^{(2)}$ is reduced to

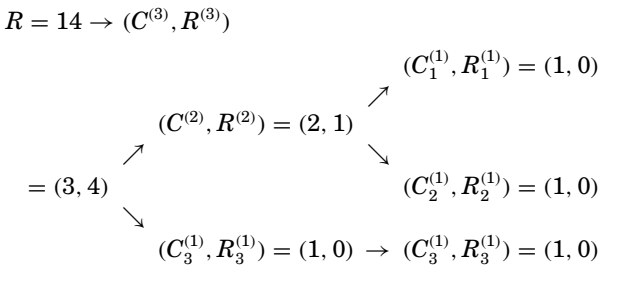
$$\begin{aligned} R_{1,1}^{(2)} &= R^{(2)} - (\min I_{1,1}^{(2)} - \min I_2^{(2)}) \\ &= 1 - (14 - 13) = 0 \end{aligned} \tag{35}$$

$$R_1^{(1)} = R_{1,1}^{(2)} \bmod |S_1^{(1)}| = 0 \bmod 1 = 0 \tag{36}$$

$$R_2^{(1)} = \left\lfloor \frac{R_{1,1}^{(2)}}{|S_1^{(1)}|} \right\rfloor = \left\lfloor \frac{0}{1} \right\rfloor = 0 \tag{37}$$

The final cost-residue pairs are $(C_1^{(1)}, R_1^{(1)}) = (1, 0)$, $(C_2^{(1)}, R_2^{(1)}) = (1, 0)$, and $(C_3^{(1)}, R_3^{(3)}) = (1, 0)$. We may now find the elements that correspond to each cost-residue pair. For example, we need to find the first combination with the cost of 1 for $(C_1^{(1)}, R_1^{(1)})$. Coincidentally, the elements have the same value as the costs. Now, the three 1D costs form the final combination $\mathbf{s}^{(3)} = (C_1^{(1)}, C_2^{(1)}, C_3^{(1)}) = (1, 1, 1)$. The answer is verified graphically in Fig. 5.

The decomposition results in a tree diagram as



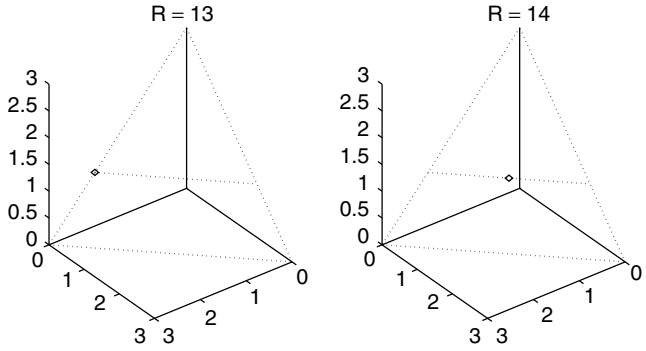


Figure 5. Decomposition of $S_2^{(2)}$ into individual lattice points. The final answer is $s^{(3)} = (1, 1, 1)$.

Note that at each column all costs sum up to 3 ($3 = 2 + 1 = 1 + 1 + 1$). Algebraically, this is equivalent to the decomposition of $g^{(3)}(x) \rightarrow g^{(2)}(x) \cdot g^{(1)}(x) \rightarrow g^{(1)}(x) \cdot g^{(1)}(x) \cdot g^{(1)}(x)$.

2.7. Demapping Algorithm

The demapping operation is the inverse operation of the mapping process. We solve for the inverse of each step and reverse the order of the steps. It is simpler than the mapping process in the sense that there is no need to search for the proper-cost decomposition.

First, given a vector $\mathbf{s}^{(N)} = (s_1 \cdots s_N)$, we find (usually via lookup tables) $\{(C_i^{(1)}, R_i^{(1)})\}_{i=1}^{N}$ such that s_i is the $(R_i^{(1)} + 1)$-th element in S with a cost of $C_i^{(1)}$. Next, for each pair of cost–residue pairs, $(C^{(P)}, R^{(P)})$ and $(C^{(Q)}, R^{(Q)})$, we combine the costs by

$$C^{(P+Q)} = C^{(P)} + C^{(Q)} \tag{38}$$

and the residues by

$$R_{C^{(P)}, C^{(Q)}}^{(P+Q)} = |S_{C^{(P)}}^{(P)}| R^{(Q)} + R^{(P)} \tag{39}$$

$$R^{(P+Q)} = R_{C^{(P)}, C^{(Q)}}^{(P+Q)} + (\min I_{C^{(P)}, C^{(Q)}}^{(P+Q)} - \min I_{C^{(P+Q)}}^{(P+Q)}) \tag{40}$$

We repeatedly apply this procedure until we combine all cost–residue pairs to one pair: $(C^{(N)}, R^{(N)})$. Then, the original index can be recovered by

$$R = \sum_{i=0}^{C^{(N)}} g^{(N)}(i) + R^{(N)} \tag{41}$$

3. APPLICATION AND PRACTICAL CONSIDERATION

Shell mapping can be applied in situations in which the cost (i.e., energy in the case of constellation shaping) per ring is proportional to the ring index and the total cost is the sum of the individual ring costs. This is nearly exact for large-QAM constellations subdivided into many rings. The shaping gain from shell mapping tends to improve as the number of rings increases, thereby reducing the number of symbols in each ring. However, the complexity and cost of shell mapping also increases with the number of rings, so practical systems make a judicious trade-off between the accuracy of the approximation and the number of rings. For example, the V.34 modem standard uses 8 rings of QAM symbols to create a 16-dimensional-shaped constellation. Each QAM symbol is divided into as many as 16 rings from a constellation as large as 1664-QAM.

The maximal shaping gain is obtained asymptotically as the number of dimensions grows arbitrarily large. However, the complexity of the shell mapping algorithm also grows with the dimension, so practical systems select a moderate block size. For example, the V.34 modem uses 16-dimensional blocks; larger block sizes provide incremental performance benefits at the cost of rapidly increasing complexity.

Shell mapping reduces the average energy of a block of QAM symbols by using only the 2^b combinations of

rings with the lowest total energies over the block. For instance, it would not use the high-energy constellation points for every QAM symbol in a single block. To provide the freedom to reject high-energy combinations of symbols, at least a modest overdeterminacy in the individual QAM constellations is required. For example, if a data rate of 8 bits per QAM symbol is desired, a shell-mapped QAM constellation of more than 256 points is required to support shaping. The larger constellation increases both complexity and the peak-to-average power ratio, however, so some compromise between expansion and reduced shaping gain is generally made. Factors of 1.2–1.5 times the base size are usually sufficient to realize most of the available shaping gain.

The ITU V.34 modem standard includes shell mapping. It uses eight successive QAM symbols to create a 16-dimensional-shaped constellation. Each QAM symbol is divided into as many as 16 rings from a base constellation as large as 1664-QAM. The incoming bit stream is segmented into blocks of bits. Within each block, a portion are used for the trellis coder. The output selects one of four QAM constellations. Some bits are sent to the shell mapper, which yields eight ring indices. The remaining bits are used to select the actual constellation points within each ring.

BIOGRAPHY

Douglas L. Jones received th B.S.E.E., M.S.E.E., and Ph.D. degrees from Rice University in 1983, 1985, and 1987, respectively. During the 1987–1988 academic year, he was at the University of Erlangen — Nuremberg in Germany on a Fulbright postdoctoral fellowship. Since 1988, he has been with the University of Illinois at Urbana — Champaign, where he is currently a Professor in Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute. In the Spring semester of 1999 he served as the Texas Instruments Visiting Professor at Rice University. He is an author of the laboratory textbook *A Digital Signal Processing Laboratory Using the TMS32010*, over 150 conference and journal publications, and several patents. His research interests are in digital signal processing and communications, including nonstationary signal analysis, adaptive processing, multisensor data processing, OFDM, and various applications such as advanced hearing aids.

BIBLIOGRAPHY

1. G. D. Forney and L. Wei, Multidimensional constellations — Part I: Introduction, figures of merit, and generalized cross constellation, *IEEE J. Select. Areas Commun.* **7**: 877–892 (1989).
2. G. D. Forney and L. Wei, Multidimensional constellations — Part II: Voronoi constellations, *IEEE J. Select. Areas Commun.* **7**: 941–958 (1989).
3. R. Laroia, On optimal shaping of multidimensional constellations, *IEEE Trans. Inform. Theory* **40**: 1044–1056 (1994).
4. P. J. Cameron, *Combinatorics: Topics, Techniques, Algorithms*, Cambridge Univ. Press, Cambridge, UK, 1994.
5. R. A. Brualdi, *Introductory Combinatorics*, Prentice-Hall, Englewood Cliffs, NJ, 1994.
6. H. K. Kwok, *Shape Up: Peak-Power Reduction via Constellation Shaping*. Ph.D. thesis, Univ. Illinois at Urbana — Champaign, 2000.

SIGMA-DELTA CONVERTERS IN COMMUNICATION SYSTEMS

FRED HARRIS
San Diego State University
San Diego, California

1. THE WHY OF SIGMA-DELTA CONVERTERS

Correlation between a sequence of sample values can be used to quantize, to a specified fidelity, a signal with significantly fewer bits per sample than that required by an instantaneous quantizer. The sigma-delta (Σ-Δ) converter uses feedback to shape the quantizing noise spectrum of an oversampled low-resolution quantizer to obtain low levels of in-band noise in exchange for higher levels of out-of-band-noise. In effect, the quantizer arranges for the quantizing noise spectrum and input signal spectrum to occupy nearly distinct spectral regions. Filtering that rejects the out-of-band-shaped quantizing noise converts the signal correlation to additional bits by the ratio of coherent gain to incoherent gain of the filtering process.

A major application of the sigma-delta process is in the area of analog-to-digital conversion, particularly in the conversion of audio signals, of instrumentation signals, and modem input signals. High-performance converters operating with 1-bit quantizers at an input rate 64 times the desired output rate of 100 kHz can supply 24-bit samples with 120 dB SNR (signal-to-noise ratio). The second major application of the Σ-Δ process has been in the area of digital-to-analog conversion for audio signals, control signals, and modem output signals. It is standard practice to use multirate digital filters to raise the sample rate of a 16-bit sampled data signal by a factor of 64, say, from 48 kHz to 3.072 MHz, then convert the 16-bit data samples to 1-bit data samples with a digital Σ-Δ converter, which is then presented to a 1-bit DAC for conversion to an analog signal.

The enhanced analog-to-digital conversion application uses analog sampled data components in the Σ-Δ modulator. These are implemented with switched capacitor or continuous-time integrators, a pair of low-resolution A-to-D (A/D) and D-to-A (D/A) converters in the feedforward–feedback path, and digital filters to reduce noise bandwidth and signal sample rate. By way of comparison, the enhanced D/A conversion application uses standard digital signal processing blocks in its Σ-Δ modulator. It also uses standard DSP blocks to raise the sample rate, a single low-resolution D/A converter to form an analog signal, and analog filters to reduce the noise bandwidth.

The third major application of the Σ-Δ process is in the area of digital-to-digital (D/D) conversion. In this

application, the input signal, collected and quantized with a conventional A/D, is pre-processed with a digital Σ-Δ to reduce the number of bits required to represent the signal in the bandwidth to be extracted by subsequent processing. The Σ-Δ modulator shapes the quantizing noise spectrum to preserve low noise levels over the frequency span of the passband while permitting significant increase of noise level in the frequency span of the stopband of the subsequent digital filter. Processing resources, such as multiplier widths required to implement digital filters after the digital Σ-Δ modulator, are reduced considerably. Reduction in data bit width from 16 to 4 bits can convert filter multipliers to lookup tables for FPGA implementations, while reductions from 16 bits to 1 bit permits arbitrary FIR filtering with no multipliers at all.

Processing flow diagrams matching the three applications we have described are shown in Fig. 1. The material covered in the remainder of this presentation will be

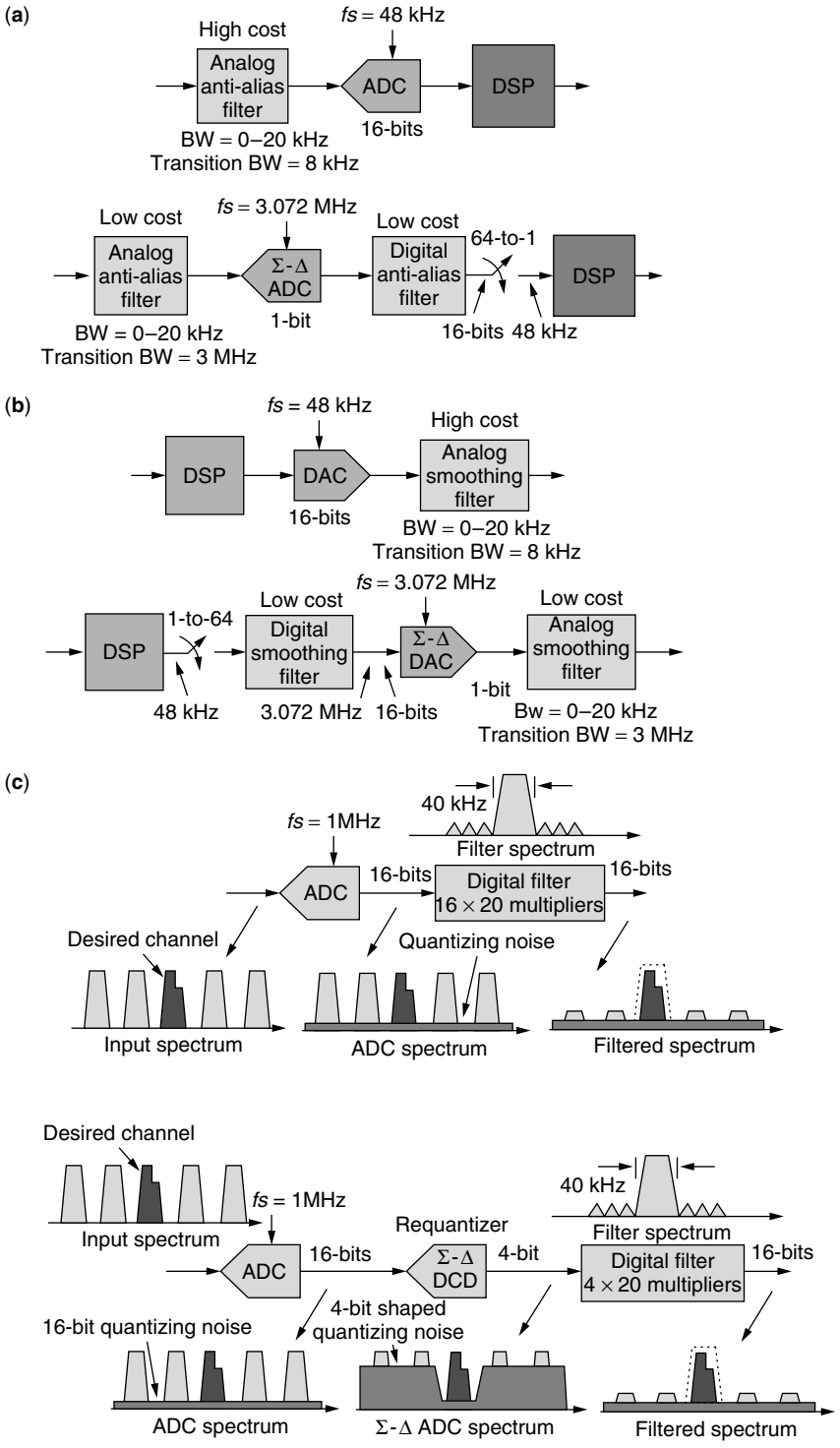


Figure 1. (**a**) A/D conversion with memoryless converter compared to oversampled sigma-delta A/D modulator with low-cost digital filter; (**b**) D/A conversion with memoryless converter compared to low-cost digital filter with oversampled sigma-delta D/A modulator; (**c**) digital signal processing of sampled data with full-bandwidth A/D conversion compared to processing with sigma-delta preprocessed bandwidth-limited data.

generic descriptions of the Σ-Δ process without regard to implementation. Applications presented near the end of this article will emphasize the third application, with use of the all-DSP digital-to-digital process in communication systems.

2. BACKGROUND

We start by reviewing the model of the ideal amplitude quantizer that performs instantaneous mapping from input amplitude x to output amplitude x_q in accord with a specified input–output profile $x_q = Q(x)$. As shown in Eq. (1) and in Fig. 2, the difference between the input and output of the quantizer represents an error, and to control the size of this error, the profile should be close to the errorless profile $x_q = x$, the unit slope line through the origin.

$$x_q = Q(x)$$
$$e_q = Q(x) - x = x_q - x \tag{1}$$

The quantization profile, reminiscent of a staircase, is defined by the locations of its treads and risers. Figure 2 shows the nonlinear mapping profile and the error profile of a uniform quantizer, one exhibiting equally spaced treads and risers. Such a quantizer is described as being a uniform or, paradoxically, a linear quantizer.

Note from the error profile that the quantization error is a deterministic and nonlinear function of the input signal level with a known error for a known input. A linear model of the quantizer is that of a zero-mean, independent, uniformly distributed, white-noise source added to each input sample. Both models are shown in Fig. 3. The linear model is often substituted for the nonlinear model for ease of analysis. In our subsequent discussions, we make this substitution with the awareness that this may a poor representation of the quantizing noise when the number of output levels is less than 16, or when the input samples are highly correlated. These are precisely the conditions we encounter in the Σ-Δ process. It is for this reason that simple linear analysis of a Σ-Δ system does not predict nonlinear modes of behavior, which we discuss in a later section.

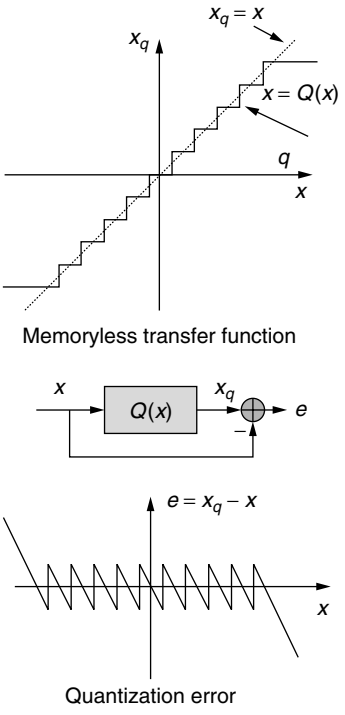


Figure 2. Quantization profile and error profile for uniform quantizer.

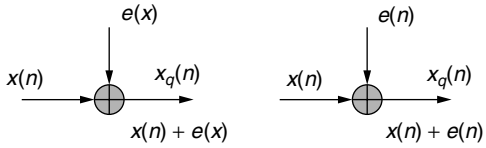


Figure 3. Nonlinear model and linear additive model of quantizer.

The noise in the linear additive noise model of the quantizer is uniformly distributed over a quantile interval of width q, from which we can compute the mean-square quantizing noise as shown in Eq. (2) to be $q^2/12$. We can similarly identify the mean-square level of the signal component at the quantizer output for any input amplitude distribution. The simplest such distribution is uniform over the range of $-Mq$ to $+Mq$, where M is the maximum number of positive and negative quantile increments. The mean-square signal level for this distribution is shown in Eq. (3) to be $(2Mq)^2/12$:

$$\sigma_q^2 = \int_{-q/2}^{+q/2} e^2 f_q(e)\, de = \frac{1}{q}\int_{-q/2}^{+q/2} e^2\, de = \frac{q^2}{12} \tag{2}$$

$$\sigma_s^2 = \int_{-Mq}^{+Mq} s^2 f_s(s)\, ds = \frac{1}{2Mq}\int_{-Mq}^{+Mq} s^2\, ds = \frac{(2M)^2 q^2}{12} \tag{3}$$

The signal : quantizing noise ratio of the quantizer is shown in Eq. (4) to be $(2M)^2$. Equation (4) also shows that when we replace $2M$, the number of levels in a quantizer with a power of 2 of the form 2^b, the SNR is 2^{2b}. Finally we see that the SNR in decibels is seen to be $6b$ dB, from which we obtain the standard quantizer rule of 6 dB per bit. The SNR of a quantizer, for any input amplitude distribution, is of the form shown in Eq. (5), where the offset factor $K_{\text{density}}(\sigma^2)$, is a parameter that varies with amplitude density and signal variance. K_{density} is negative for most densities and becomes more so as we decrease the input signal variance relative to quantizer dynamic range.

The primary message presented by Eq. (5) is that quantizer fidelity or SNR can be purchased with a linear quantizer by increasing the number of bits involved in the quantization process. The first rule of quantizing is: "If you want a higher fidelity representation of the signal, get more bits." In the next section we derive a corollary to this rule: "If you can't get more bits, get correlated samples and convert them to more bits."

$$\text{SNR} = \frac{\sigma_s^2}{\sigma_q^2} = \frac{(2M)^2 q^2/12}{q^2/12} = (2M)^2 = (2^b)^2 = 2^{2b}$$

$$\begin{aligned}\mathrm{SNR}_{\mathrm{dB}} &= 10\log_{10}(\mathrm{SNR}) = 10\log_{10}(2^{2b}) \qquad (4)\\ &= 20b\log_{10}(2) = 6b \text{ dB}\end{aligned}$$

$$\mathrm{SNR}_{\mathrm{dB}} = 6b + K_{\mathrm{density}}(\sigma^2) \qquad (5)$$

3. SIGMA-DELTA MODEL

The model of a Σ-Δ converter can be derived from a number of equivalent perspectives; the two most common are that of an error feedback modulator and that of a standard two-input one-output feedback loop. We first examine the error feedback model and then convert this model to other feedback models. Figure 4 presents the structure of a noise feedback modulator or coder. The coder consists of a prediction filter that computes, from previous quantization errors, an estimate $\hat{q}(n)$ of the next quantization error $q(n)$. This estimate is subtracted from the input and presented to the internal quantizer that adds the actual quantization error, $q(n)$, an error modeled as an additive noise source. The size of this added error is computed as the difference between the output and input of the quantizer and is delivered to the predictor filter for use in the next prediction cycle. From Fig. 4, we can determine the input and output of the quantizer that is shown in Eq. (6).

$$\begin{aligned}&\text{Quantizer input:} && x(n) - \hat{q}(n)\\ &\text{Quantizer output:} && y(n) = [x(n) - \hat{q}(n)] + q(n) \qquad (6)\\ & && = x(n) + [q(n) - \hat{q}(n)]\end{aligned}$$

The transfer function of the noise feedback coder is presented in Eq. (7). Here we see that the output of the system contains the signal input to the system, $X(Z)$, plus a filtered version of the noise input $Q(Z)[1 - P(Z)]$. The term $[1 - P(Z)]$ is denoted the noise transfer function (NTF):

$$\begin{aligned}Y(Z) &= X(Z) + [Q(Z) - \hat{Q}(Z)]\\ &= X(Z) + Q(Z)[1 - P(Z)] \qquad (7)\end{aligned}$$

Rather than continuing with this general model with arbitrary $P(Z)$, we choose to examine a specific example, in particular, the simplest prediction filter from which we will derive insight and guidance to other filter structures. We reason as follows. We want the output $\hat{q}(n)$ of the prediction filter to be a good approximation of $q(n)$, and further we want the computational burden required to perform the prediction to be small. We can realize these conditions if the successive samples of the error are highly correlated and, of course, we obtain high correlation when the input signal is significantly oversampled. This is the justification for the requirement to deliver over sampled data to the Σ-Δ converter. It is common, for instance, for the input data to be sampled at 64 times the signal's Nyquist rate. With significant oversampling, we can easily argue that the correlation between successive input samples is high and consequently the correlation between successive quantization errors is also high. For this condition, a good approximation of the next quantization error $\hat{q}(n)$ is the previous error $q(n-1)$, and the filter that supplies the delayed noise sample is $P(Z) = Z^{-1}$. Figure 5 shows the noise feedback coder with the prediction filter replaced with a single delay element Z^{-1}.

We note that there are two feedback loops in Fig. 5. The first loop starts at the input summing junction, negotiates the lower summing junction, and passes through the delay line back to the input junction. The second loop starts at the input summing junction, passes through the quantizer summing junction, the sign reversal of the lower summing junction, through the delay line and back to the input summing junction. Figure 6 presents the noise feedback quantizer redrawn to explicitly show the two loops just traversed, and then recognizing that the minor loop is a digital integrator, replacing it with the transfer function $Z/(Z-1)$. This figure is the conventional model of a Σ-Δ converter comprising an integrator and quantizer in a feedback loop.

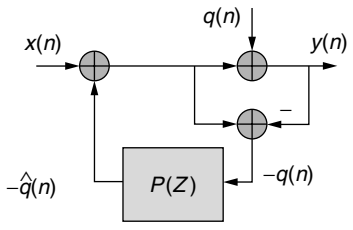


Figure 4. Simple model of noise feedback quantizer.

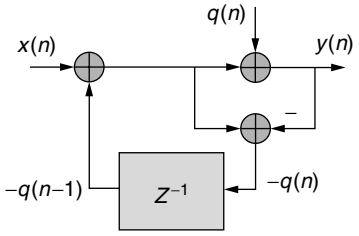


Figure 5. Noise feedback quantizer with delay-line predictor.

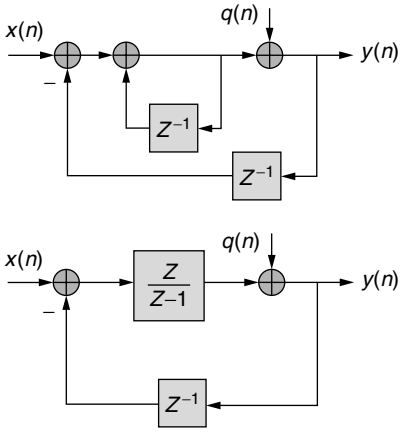


Figure 6. Noise feedback quantizer drawn with inner and outer feedback loops and with inner loop, a digital integrator represented by its transfer function.

Replacing $P(Z)$ with Z^{-1} in Eq. (7) leads to Eq. (8), which describes the input–output relationship of the Σ-Δ converter. Note that the NTF for the quantizing noise input is that of a simple differentiator $[1 - Z^{-1}]$ and that this filter has a transmission zero at $Z = 1$ or at DC. The pole-zero diagram of this NTF and its power spectral response along with the spectra of a typical input signal is shown in Fig. 7. The power spectral response of the NTF is derived in Eq. (9), where sampled data frequency is denoted by the parameter θ, where $\theta = \omega T = 2\pi(f/fs)$ and has units of radians per sample.

$$\begin{aligned} Y(Z) &= X(Z) + Q(Z)[1 - Z^{-1}] \\ &= X(Z) + Q(Z)\left[\frac{Z-1}{Z}\right] \end{aligned} \tag{8}$$

$$\begin{aligned} |\mathrm{NTF}(\theta)|^2 &= \left[\frac{e^{j\theta}-1}{e^{j\theta}}\right]\left[\frac{e^{-j\theta}-1}{e^{-j\theta}}\right] \\ &= 2[1-\cos(\theta)] = 4\sin^2(\theta/2) \end{aligned} \tag{9}$$

In Fig. 7, note that the zero of the NTF is located at DC, which suppressed the quantization noise in the neighborhood of DC. The signal spectral is restricted by the significant oversampling to reside in a small neighborhood of DC with two-sided width on the order of 1.5% of the sample rate. The combination of the over sampling and noise spectral shaping has arranged, to first order, for the signal and the noise spectra to occupy distinct spectral intervals. The signal spectra can be retrieved from the composite signal by a lowpass filter that passes the signal but rejects the noise residing beyond the signal bandwidth. Increasing the number of NTF zeros in the signal bandwidth can further improve the in-band noise suppression. We will examine Σ-Δ converters with multiple NTF zeros. The zeros of the NTF significantly suppress the quantizer noise levels of a quantizer embedded in Σ-Δ, consequently, low levels of output quantizing noise can be obtained from quantizers with relatively large levels of input quantizing noise. Most Σ-Δ converters are designed to operate with 1-bit quantizers with a number of high-performance converters operating with 4-bit quantizers.

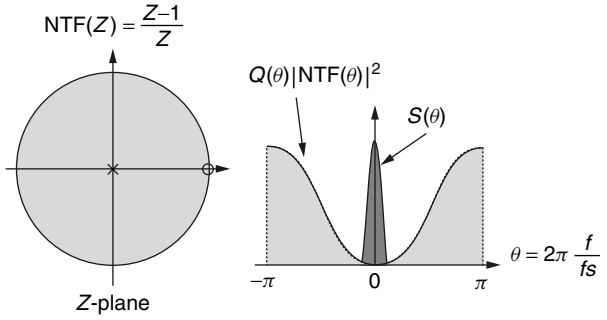


Figure 7. Pole-zero diagram of noise transfer function and power spectrum of input signal and of quantizing noise shaped by NTF.

4. NOISE PERFORMANCE OF SIGMA-DELTA CONVERTERS

We now review how filtering an oversampled and quantized data signal can improve that signal's quantizing SNR. The spectrum of nonshaped quantizing noise is uniformly distributed over a spectral width equal to the sample rate. If the signal is oversampled, say, by a factor of 2, half the quantizing noise power is in-band and half the noise power is out-of-band. We can pass the oversampled signal through a half-band filter without affecting the signal but effectively rejecting half the noise power. Rejecting half the noise bandwidth improves the signal to quantizing noise ratio by 3 dB. Quantizing noise is measured at 6 dB per bit, so reducing the noise bandwidth of uniformly distributed noise by a factor 2 improves the SNR by half a bit. To realize a 1-bit improvement in SNR, a signal would have to be oversampled by a factor of 4 and have its quantizing noise bandwidth reduced by the same factor of 4.

We now address the relationship between the SNR of a Σ-Δ converter with a shaped noise spectrum and its oversample ratio. In many modulators, the NTF of the Σ-Δ contains one or more zeros located at DC. We expand the power spectral response of a single-zero NTF in a Taylor series about DC and truncate the series after the first nonzero term to obtain an approximation to the filter response valid in the neighborhood of the signal spectrum:

$$\begin{aligned} |\mathrm{NTF}(\theta)|^2 &= 4\sin^2\frac{\theta}{2} = 2[1-\cos(\theta)] \\ &= 2\left\{1-\left[1-\frac{\theta^2}{2!}+\cdots+\right]\right\} = \theta^2 \end{aligned} \tag{10}$$

The fraction of the shaped noise spectrum that contributes to the final output of the Σ-Δ converter is that part that survives the filtering operation of the lowpass filter following the initial conversion process. The noise contained in the filter bandwidth is

$$\sigma_q^2 = \frac{N_0}{2}\int_{-\theta_{\mathrm{BW}}}^{\theta_{\mathrm{BW}}}\theta^2 d\theta = \frac{N_0}{2}\frac{1}{3}\theta^3\Big|_{\theta=-\theta_{\mathrm{BW}}}^{\theta=\theta_{\mathrm{BW}}} = \frac{N_0}{3}(\theta_{\mathrm{BW}})^3 \tag{11}$$

where we see that the noise contained in the filtered Σ-Δ output is proportional to the cube of the ratio of bandwidth-to-sample rate. If we reduce this ratio by a factor of 2, the noise contribution is reduced by a factor of 8, or by 9 dB, or equivalently by an improvement of 1.5 bits. Figure 8 illustrates how reduction in bandwidth (relative to sample rate) effects the reduction in output noise. Table 1 lists the rate at which the output SNR increases as a function of oversample ratio for a quantizer with 0, 1, 2, and 3 zeros in the NTF. The next section describes architectures for Σ-Δ converters that have NTFs with multiple zeros.

Figure 9 presents the power spectrum of a Σ-Δ's two-zero noise transfer function along with the spectrum of its 1-bit time series formed when processing an input signal containing two in-band sinusoids. Figure 10 show a segment of the input time series overlaid with the corresponding one-bit output series as well as the spectrum obtained by filtering the one-bit output series of the Σ-Δ

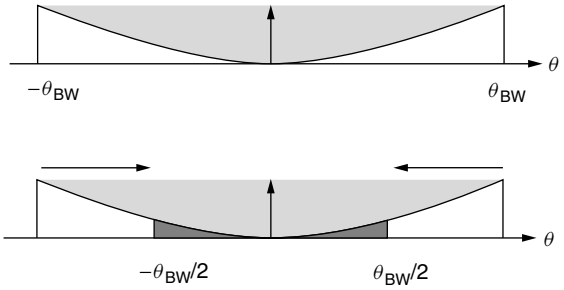


Figure 8. Area under spectrally shaped noise for two different ratios of output bandwidth-to-sample rate ratio.

converter. Note that the normalized bandwidth of the modulator and filter is $\frac{1}{100}$th of the input sample rate, a ratio equal to 6.64 octaves. From Table 1 we see that a 2-zero NTF exhibits a noise processing gain of 15 dB per oversampling octave, from which we expect a 6.64 × 15-dB or 99.6-dB improvement in SNR. With the spectrum normalized to the peak response of the windowed FFT (fast Fourier transform) we observe that the noise level of the filtered output is on the order of 98–105 dB with 103.6 dB as the actual SNR determined by integrating over the filter bandwidth.

5. SIGMA-DELTA ARCHITECTURES

A small number of common structures describe the architecture of the majority of Σ-Δ converters. We now examine the design philosophy common to many of these architectures. Most Σ-Δ architectures are formed about a feedback loop containing a loop filter and low-resolution quantizer that forms an oversampled, spectrally shaped data sequence that is processed by an external band-limiting filter to reject out-of-band noise. The feedback loop of the Σ-Δ quantizer is called the "Σ-Δ modulator," which, when combined with the filter, forms the sigma-delta converter.

Figure 11 is an extension of the redrawn noise feedback structure of Fig. 6 that cast the Σ-Δ as a two-input one-output feedback system. That first system employed an integrator and quantizer in a unity feedback control loop. What we have done here is replace the feedback integrator with an arbitrary filter $H(Z)$ and have also placed filter $G(Z)$ in the input path to enable private zeros for the input signal. We now consider appropriate constraints for the two filters to obtain the desired Σ-Δ performance.

As shown in Eq. (12), the transfer functions from the two inputs, $X(Z)$ and $Q(Z)$, to the common output $Y(Z)$ is computed as their distinct forward gains divide by one minus the loop gain:

$$Y(Z) = \frac{G(Z)}{1+H(Z)}X(Z) + \frac{1}{1+H(Z)}Q(Z) \tag{12}$$

The transfer functions of $H(Z)$ and $G(Z)$ are ratios of numerators to a common denominator polynomial defined by $N_1(Z)/D_1(Z)$ and $N_2(Z)/D_1(Z)$, respectively. Substituting these ratios in Eq. (12) leads to Eq. (13), and clearing the resulting denominators results in Eq. (14):

$$Y(Z) = \frac{N_2(Z)/D_1(Z)}{1+N_1(Z)/D_1(Z)}X(Z) + \frac{1}{1+N_1(Z)/D_1(Z)}Q(Z) \tag{13}$$

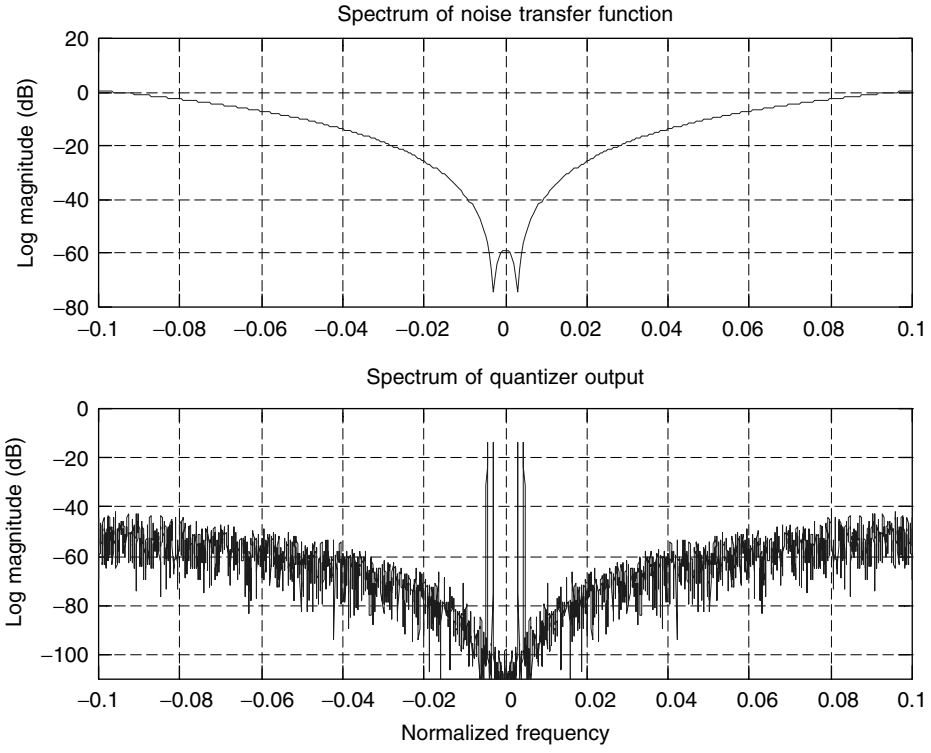


Figure 9. Noise transfer function of a 2-zero sigma-delta modulator and power spectrum of the output series obtained from modulator.

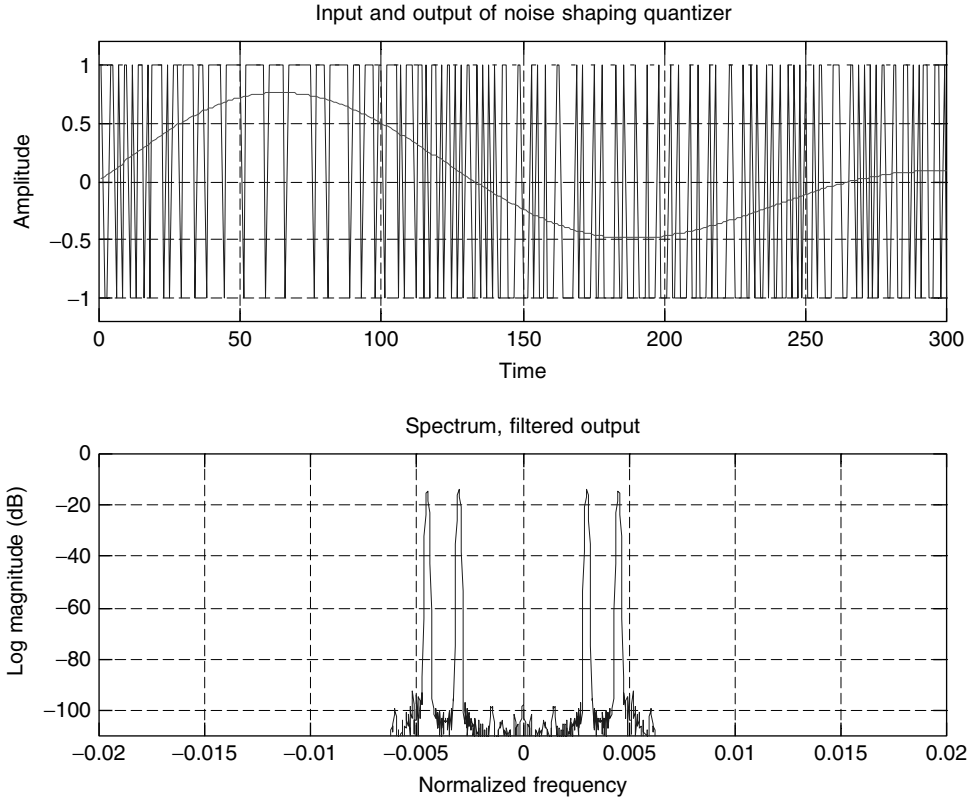


Figure 10. Segment of input and 1-bit output time series and power spectrum of filtered 1-bit output series from modulator.

$$Y(Z) = \frac{N_2(Z)}{D_1(Z) + N_1(Z)} X(Z) + \frac{D_1(Z)}{D_1(Z) + N_1(Z)} Q(Z) \tag{14}$$

5.1. MULTIPLE FEEDBACK LOOPS

We identify the two transfer functions shown in Eq. (14) as the signal transfer function (STF) and as the noise transfer function (NTF), respectively. As expected, the poles, $D(Z)$, of the loop filter transfer function become the zeros of the NTF. To assure good stopband performance of the NTF, $D(Z)$ must have its roots on the unit circle in the signal passband. Poles located at $Z = 1$, obtained by local feedback, form integrators that lead to the desired NTF zeros. Local feedback between the integrators can redistribute the zeros along the unit circle to realize an equal-ripple stopband to exchange excess attenuation for wider stopband bandwidth. Typical NTFs that can be realized with the structure of Eq. (14) are illustrated in Eq. (15); by way of example, these NTFs are third-order NTFs implementing Chebyshev (also transliterated as Tchebyschev), Butterworth, and derivative stopbands. These NTFs have distributed zeros and active poles, repeated zeros and active poles, and repeated zeros with inactive poles; "active" poles are finite poles that affect the spectral magnitude response, and hence are finite poles not at the origin. The frequency responses of these sample NTFs are shown in Fig. 12, where we see that the active poles have the desirable effect of reducing the

Table 1. Improvement in Quantizer SNR in dB and Effective Number of Bits for Each Doubling of Sample Rate Relative to Signal Bandwidth for Sigma-Delta Converters with 0, 1, 2, and 3 Noise Transfer Function Zeros

Number of NTF Zeros	SNR Improvement	
0	3 dB/double	0.5 bit/double
1	9 dB/double	1.5 bit/double
2	15 dB/double	2.5 bit/double
4	21 dB/double	3.5 bit/double

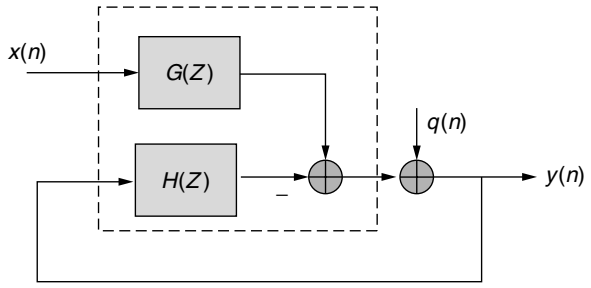


Figure 11. Two-input one-output feedback model of sigma-delta modulator.

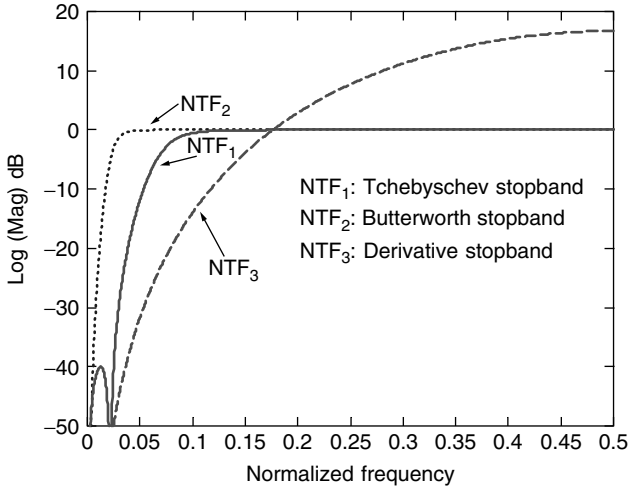


Figure 12. Spectral responses of typical noise transfer functions.

high-frequency gain of the NTF and that the distributed zeros widens the stopband bandwidth of the NTF.

$$\begin{aligned} \mathrm{NTF}_1 &= \frac{(Z-1)(Z^2+c_1Z+1)}{(Z-1)^3+(Z^2+b_1Z+b_2)} \\ \mathrm{NTF}_2 &= \frac{(Z-1)^3}{(Z-1)^3+(Z^2+b_1Z+b_2)} \\ \mathrm{NTF}_2 &= \frac{(Z-1)^3}{Z^3} \end{aligned} \tag{15}$$

The zeros $N_1(Z)$ of the feedback filter $G(Z)$ are selected so that the system poles match a selected prototype transfer function such as an elliptic, Chebyshev, or Butterworth stopband filter.

Figure 13a shows a three-stage version of the general multiple feedback–feedforward filter structure with local feedback forming the discrete integrators. The integrator poles become, via the major feedback loops, the desired NTF zeros. In this configuration, the input data $x(n)$ enter the filter through the feedforward path while the output data $y(n)$ leave the filter at the feedback path. Figure 13b presents the dual three-stage version of the pole structure presented in Fig. 13a. Here the input data $x(n)$ enter the filter at the feedback path while the output data $y(n)$ also leave the filter at the feedback path in which the quantizer must reside. This form of the filter does not offer a feedforward path to form private zeros for the signal transfer function.

5.2. Cascade Converters

The second major architecture for Σ-Δ modulators is that of cascade low-order Σ-Δ modulators. The cascade form is called a MASH converter, for *multiple sample and holds*, a description of the analog implementation. The low-order modulators can be formed by any structure but is usually implemented with one or two integrator loops and a 1-bit quantizer as originally shown in Fig. 6. We derived the expression for the output of the first stage of a single-loop modulator in Eq. (8). The output contains the loop's quantizer noise differentiated by the loop NTF. We obtain an improved NTF by the use of a second Σ-Δ modulator to measure and cancel the noise of the first modulator. This structure is shown in Fig. 14.

The Z transform of the first loop's output is shown in Eq. (16), and that of the second loop's output is shown in Eq. (17). Note that the first output contains $Q_1(Z)[1-Z^{-1}]$, the first loop's noise filtered by the loop NTF, a derivative, while the second output contains $Q_1(Z)$, the first loop's noise without the derivative. Applying a derivative as a $(1-Z^{-1})$ operator to the output of the second loop leads to the terms shown in Eq. (18), and forming the sum of $y_1(n)$ and $[y_2(n)-y_2(n-1)]$ leads to the terms shown in Eq. (19):

$$Y_1(Z) = X(Z) + [1-Z^{-1}]Q_1(Z) \tag{16}$$

$$Y_2(Z) = -Q_1(Z) + [1-Z^{-1}]Q_2(Z) \tag{17}$$

$$[1-Z^{-1}]Y_2(Z) = -[1-Z^{-1}]Q_1(Z) + [1-Z^{-1}]^2Q_2(Z) \tag{18}$$

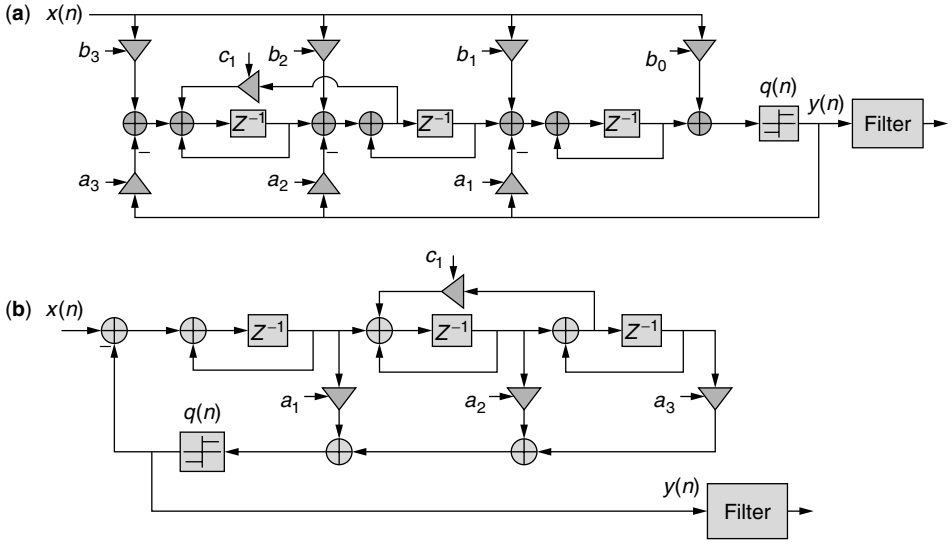


Figure 13. (**a**) Three-stage example of feedback–feedforward filter using cascade discrete integrators: zeros formed at input, poles formed at output; (**b**) dual three-stage example of feedback using cascade discrete integrators: poles formed at input.

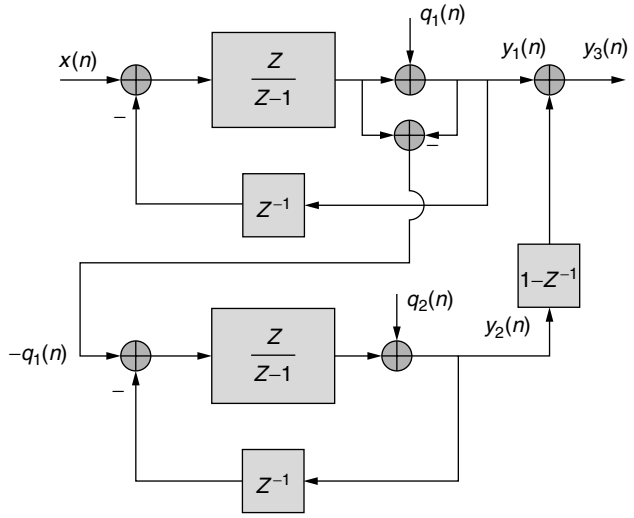


Figure 14. Cascade single-loop, single-bit sigma-delta modulators.

$$\begin{aligned} Y_1(Z) + [1 - Z^{-1}]Y_2(Z) &= X(Z) + [1 - Z^{-1}]Q_1(Z) \\ &\quad + \{-[1 - Z^{-1}]Q_1(Z) \\ &\quad + [1 - Z^{-1}]^2 Q_2(Z)\} \\ &= X(Z) + [1 - Z^{-1}]^2 Q_2(Z) \end{aligned} \tag{19}$$

The output of the cascade modulators contains the input $X(Z)$ and the doubly differentiated noise $Q_2(Z)[1 - Z^{-1}]^2$ of the second modulator loop. The double zero in the NTF has the effect of further suppressing the in-band quantization noise, which improves the output SNR for a given over sample rate. The process of canceling the first loop's noise with differentiated output of the second loop results in bit growth due to the two summations. If each of loop uses a 1-bit quantizer, the modulator outputs are ± 1. The output of the discrete derivative contains the three levels, 0 and ± 2, and the output of the sum of the two paths contains the four levels, ± 1 and ± 3. The result of combining the output of the cascade 1-bit modulators results in an equivalent 2-bit modulator. The cascade can be extended to include three stages by using a third Σ-Δ modulator to measure and cancel the noise of the second Σ-Δ modulator. The output of the third stage is double differentiated and added to the output of the two-stage cascade, which replaces the doubly differentiated second stage noise with the triply differentiated third-stage noise. Here again, the double derivative of a 1-bit sequence results in an increased output bit width. The output levels of the three cascade modulators are ± 1, ± 3, ± 5, and ± 7, the equivalent of a 3-bit modulator. Figure 15a–c presents the time series and the spectrum obtained from successive stages of a cascade of three single-loop modulators. Note the bit growth at each successive output and the enhanced depth of spectral noise suppression with the increased number of NTF zeros of the cascade modulator. For comparison, Fig. 15d presents the time series and spectrum from a single 3-loop, 3-bit sigma delta modulator. The spectral responses of the two 3-loop systems are essentially the same even though the time series from the two systems are very different.

5.3. Noise Prediction Loops

The third and the last major architecture for a Σ-Δ modulators is an enhanced version of the prediction filter introduced as our first model of the Σ-Δ modulator and presented in Fig. 4. We initially replaced the prediction filter with a delay element Z^{-1}, and then elected to operate the loop at rates far in excess of the signal Nyquist rate to assure high correlation between successive errors. We now return to this structure and describe one technique of designing efficient prediction filters, which permit significant reduction in the system over sample rate.

The design of a prediction filter that forms successive estimates of the next input sample is a standard task in signal estimation. The optimum prediction filter processes a sequence of N successive input samples with weights that minimize the mean-squared error between the prediction of the next sample and the ensuing measurement of that sample. This structure is expressed in Eq. (20) and shown in Fig. 16:

$$\begin{aligned} \hat{q}(n) &= \sum_{k=1}^{N} b(k) q(n-k) \\ e(n) &= q(n) - \hat{q}(n) \end{aligned} \tag{20}$$

The process of minimizing the mean-squared error leads to the standard normal equations shown in Eq. (21a) along with the augmented power of the prediction error shown in Eq. (21b). The set of N normal equations can be represented in matrix form as shown in Eq. (22), where the column vector $\overline{b}$ of dimension $(N+1)$ is the augmented coefficient vector $\{1 - b_1 - b_2 \cdots - b_N\}^T$, and the column vector $\overline{r}_e$ is the error correlation vector:

$$r_{qq}(m) = \sum_{k=1}^{N} b(k) r_{qq}(m-k) \quad m = 1, 2, 3, \ldots, N \tag{21a}$$

$$r_{ee}(0) = r_{qq}(0) - \sum_{k=1}^{N} b(k) r_{qq}(k) \tag{21b}$$

$$\begin{bmatrix} r_{qq}(0) & r_{qq}(1) & \cdots & \cdots & r_{qq}(N-1) \\ r_{qq}(1) & r_{qq}(0) & \cdots & \cdots & r_{qq}(N-21) \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ r_{qq}(N-1) & r_{qq}(N-2) & \cdots & \cdots & r_{qq}(0) \end{bmatrix} \times \begin{bmatrix} 1 \\ -b(1) \\ \vdots \\ \vdots \\ -b(N) \end{bmatrix} = \begin{bmatrix} r_{ee}(1) \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \tag{22}$$

Equation (22) can be written concisely in vector–matrix form as shown in Eq. (23) and then solved for the optimum weights as shown in Eq. (24):

$$\mathbf{R}_{qq}\overline{\boldsymbol{b}} = \overline{\boldsymbol{r}}_e \tag{23}$$

$$\overline{\boldsymbol{b}} = \mathbf{R}_{qq}^{-1}\overline{\boldsymbol{r}}_e \tag{24}$$

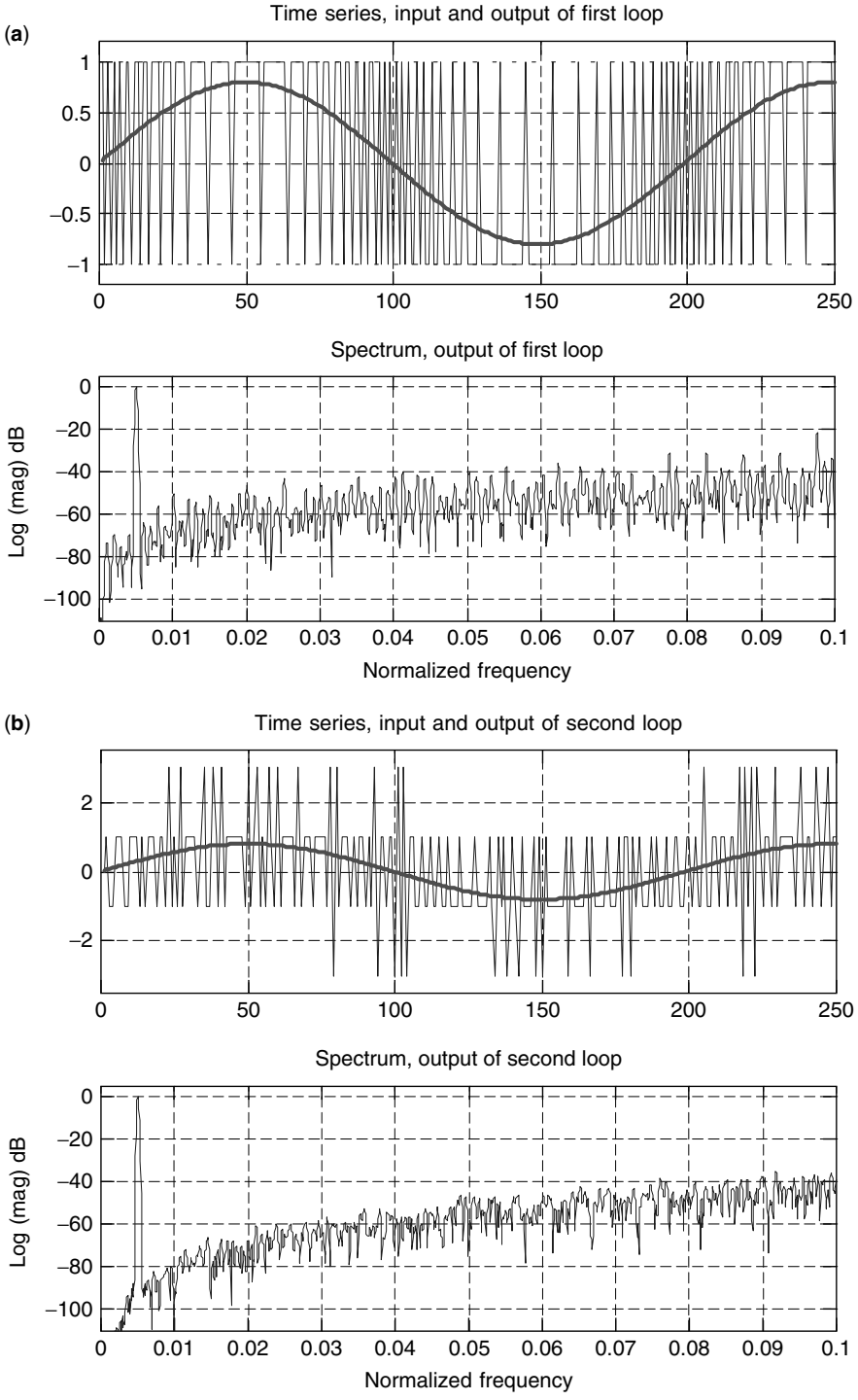


Figure 15. (**a**) Time series and spectral response of (**a**) first 1-bit modulator; (**b**) two-cascade 1-bit modulators; (**c**) three-cascade 1-bit modulators; and (**d**) 3-zero 3-bit modulator.

We have no problem with the existence of, and the process for, computing the weight set that minimizes the mean squared error for a given signal from which we can extract the second-order statistics. The problem is that we do not know the signal a priori hence do not have the required signal statistics. We fall back to a minimax solution, that of finding the solution for the signal with the worst statistics and apply that solution to the arbitrary signal. This signal is band limited white noise, and of course it is the band limiting that permits the successful prediction. The power spectrum of the bandlimited noise is shown in Fig. 17 and expressed as follows:

$$P_{qq}(\theta) = \begin{cases} 1 & -\theta_0 \le \theta \le \theta_0 \\ 0 & \text{elsewhere} \end{cases} \tag{25}$$

The correlation function of the modeled band-limited white noise, found as the inverse DTFT of Eq. (25), is

$$r_{qq}(n) = \frac{\theta_0}{\pi} \frac{\sin(n\theta_0)}{(n\theta_0)} \tag{26}$$

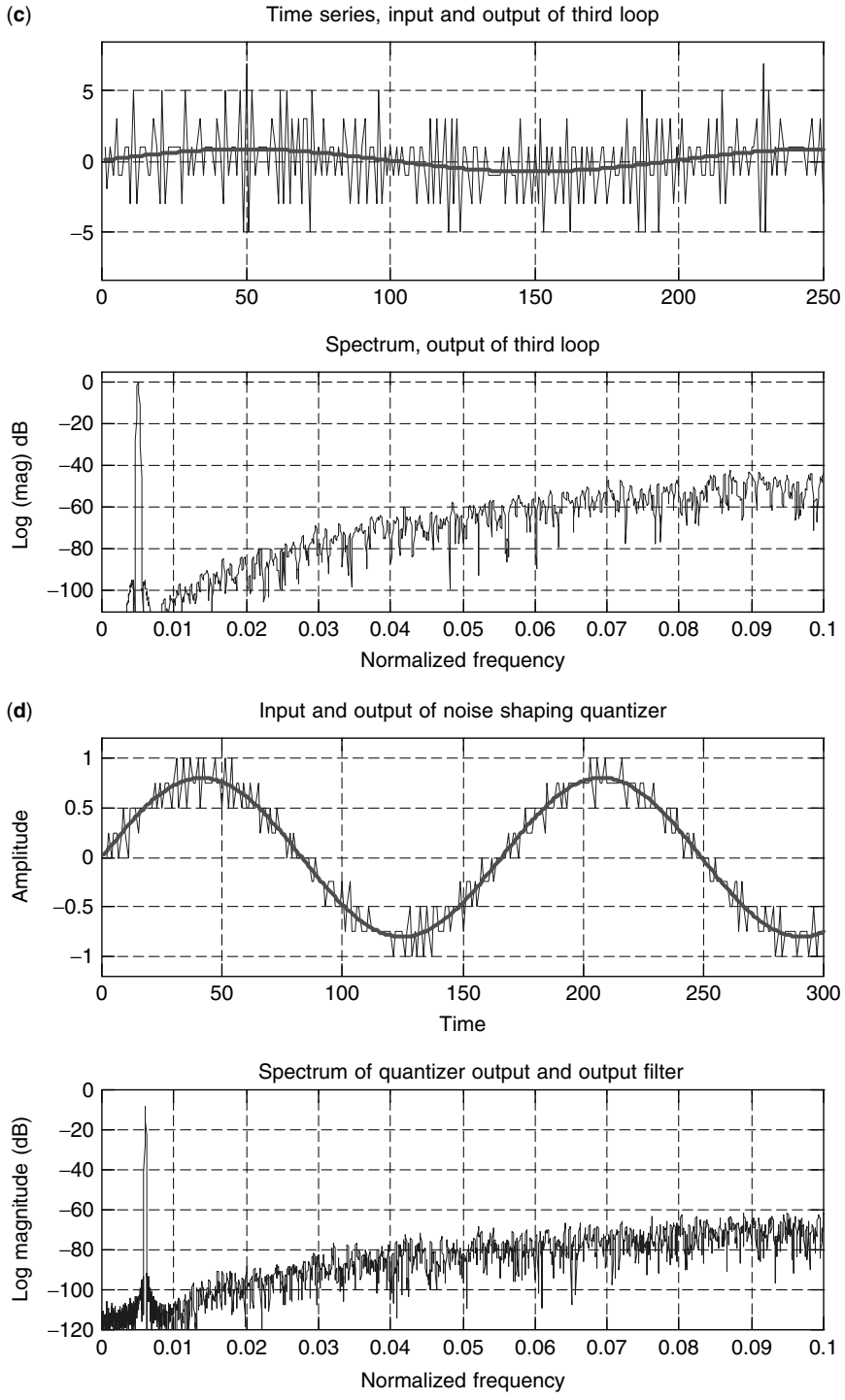


Figure 15. (*Continued*)

Using sample values of the correlation function presented in Eq. (26) to solve for the optimum weights of Eq. (24) leads to an ill-conditioned correlation matrix and to a set of weights with a large spread of coefficient values. This large spread is undesirable, and is due to the fact the optimum predictor is also the whitening filter. This filter tries to whiten the spectrum, and since there is no energy in the out-of-band region of Eq. (25), the filter can and does set arbitrarily large gains in this region. To control the out-of-band spectral gain of the predictor, we overlay the entire frequency band with a low-level white-noise spectrum of amplitude ε, modifying Eq. (25) to take the following form:

$$P_{qq}(\theta) = \begin{cases} 1+\varepsilon & -\theta_0 \le \theta \le \theta_0 \\ \varepsilon & \text{elsewhere} \end{cases} \tag{27}$$

The modified power spectrum has the following correlation function:

$$r_{qq}(n) = \frac{\theta_0}{\pi}\frac{\sin(n\theta_0)}{(n\theta_0)} + \varepsilon\delta(n) \tag{28}$$

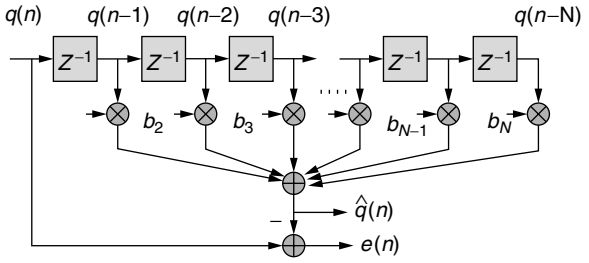


Figure 16. Tapped delay-line prediction filter.

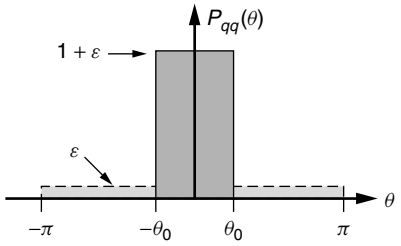


Figure 17. Power spectrum of band-limited white noise with ε white-noise overlay.

The effect of the added low-level white noise is to improve the condition number of the correlation matrix by adding the small value ε to the diagonal terms of the matrix. Adjustment of this parameter is an effective way to control out-of-band gain to improve the modulator stability margin.

Figure 18a presents the spectrum of a 20% bandwidth 10-tap prediction filter designed with the conditioning parameter $\varepsilon = 10^{-5}$. This is an unusually wide bandwidth Σ-Δ modulator with sample rate only 5 times the signal's Nyquist rate. Also shown is the spectrum formed from the Σ-Δ modulator 4-bit output sequence obtained with this prediction filter. Figure 18b shows a segment of the input series and the 4-bit output series from this modulator as well as a zoom to the passband of the output signal spectrum.

5.4. Bandpass Sigma-Delta Modulators

The Σ-Δ modulators we have examined thus far have been baseband since they were originally designed around the spectral characteristics of the pole of a digital integrator. The standard method for quantizing a passband signal involves a downconversion of the center frequency to baseband with a pair of quadrature mixers, a pair of filters to remove the sum frequencies formed by the mixing operation, and finally quantization with a pair of matched converters. It is logical to perform this translation when the Σ-Δ modulator is the conversion device, since the NTF zeros reside at base band. As expected, two converters are required to service the complex base band process. Another option is to move the NTF zeros from baseband to the center frequency of the narrowband input and perform the desired conversions with the signal residing at a low intermediate frequency.

When we move the sigma-delta integrator poles from the neighborhood of zero frequency, the resultant poles are called *resonators*. A resonator formed by structures with real coefficients exhibits both positive- and negative-frequency images, and hence requires twice the number of integrators to build the positive- and the negative-frequency NTF zeros. Doubling the number of integrators is equivalent to building and using two filters in a pair of modulators, so the increase in implementation complexity is not a concern. Feedback around resonator pole pairs results in NTF spectral zeros at these locations; hence a Σ-Δ modulator can be designed to operate at any frequency within the spectral span defined by the sample rate. Rather than pursue the straightforward synthesis problem, we consider techniques that can be applied directly to the structure of the prototype baseband modulator. A number of such techniques can be used to translate the poles of a baseband prototype modulator to an arbitrary center frequency.

We can effect a frequency transformation of an existing filter structure by replacing all-pass networks in the structure with another all-pass network. The simplest such transformation, shown in Eq. (29), replaces a delay line, represented by Z^{-1}, with a phase rotated delay line, represented by $Z^{-1}e^{j\theta}$. When applied to the delays in a filter, the spectral response of the filter is translated to the center frequency θ radians per sample. This relationship, shown in Eq. (30), is known as the *modulation property* of the Z transform:

$$Z^{-1} \Rightarrow Z^{-1}e^{j\theta} \tag{29}$$

$$\text{If} \quad h(n) \Leftrightarrow H(Z)$$

$$\text{Then} \quad h(n)e^{jn\theta} \Leftrightarrow H(Ze^{-j\theta}) \tag{30}$$

Figure 19 is a third-order multiple feedback Σ-Δ modulator with the tuning substitution described in Eq. (29) converting the prototype integrators into complex resonators. This substitution results in complex data due to the complex scalars, requiring the use of complex registers in place of the delay lines of the prototype, and similarly, the need for two real quantizers in place of the original quantizer. Figure 20 shows the frequency response of time series obtained from a baseband prototype, from a complex resonated version, and from a real resonated version of the 3-loop, 1-bit modulator presented in Fig. 19. The real resonator version is described next.

One might object to the computational burden due to the use of a complex phase rotator in each complex resonator, but remember that the scalar phase rotators replace the pair of input mixers and the DDS required for the downconversion. Judicious choice of center frequency can also lead to trivial operators that replace the complex phase rotators. In particular, when the center frequency is the quarter-sample rate, the phase rotator defaults to $\exp(j\,\pi/2)$ or j, a simple data transfer and possible sign reversal between registers.

A related all-pass transformation that can be applied directly to the baseband prototype modulator is the lowpass to bandpass transformation shown in Eq. (31). This transformation appends to each delay element in the filter a sign change, and an all-pass tuning filter with a parameter $c = \cos(\theta c)$, where θc is the center frequency.

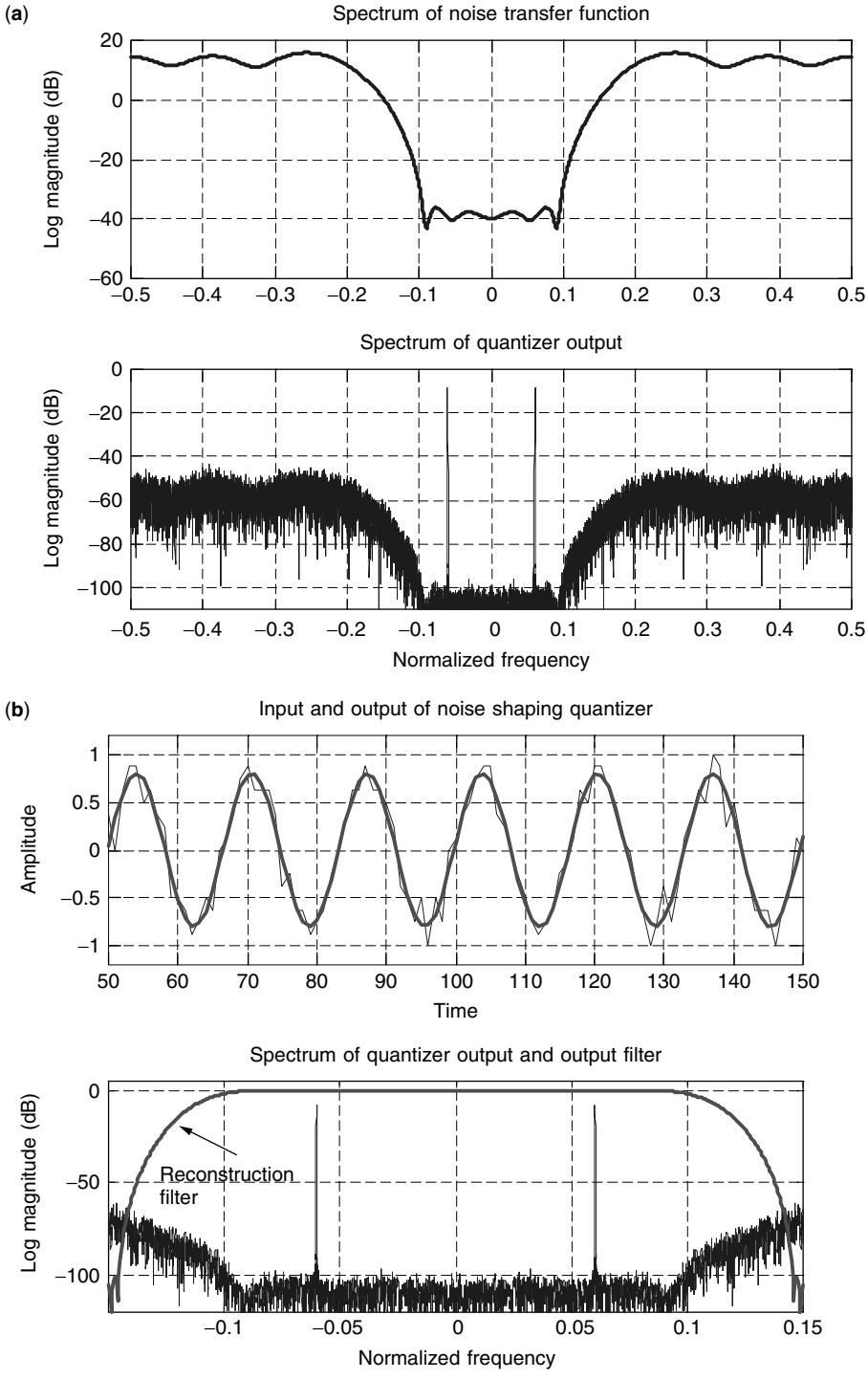


Figure 18. (**a**) Spectrum of noise transfer function obtained with 10-tap prediction filter and spectrum of sigma-delta output 4-bit series; (**b**) segment of input and output time series obtained with 10-tap prediction filter with 4-bit sigma-delta modulator and detail of output series spectrum.

The most common form of this frequency transformation is the default case $\theta c = \pi/2$, which sets c to zero, and consequently replaces Z^{-1} with $-Z^{-2}$. This substitution results in two delays with negative feedback in place of the baseband integrators:

$$\frac{1}{Z} \Rightarrow -\frac{1}{Z}\frac{1-cZ}{Z-c}, \qquad \text{where} \quad c = \cos(\theta_c) \tag{31}$$

Figure 21 illustrates the progression from the base band integrator, through the quarter-sample rate resonator, to the arbitrary all-pass tuned resonator. The lowpass to bandpass generalized delay in the tuning resonator requires two delays, two additions, and a single multiplier to implement both numerator and denominator as well as the cascade delay shown in Eq. (31). The third subplot of Fig. 19 presents the spectrum of a time series formed by the all-pass tuned variant of the modulator shown in Fig. 18.

As a final note on resonated Σ-Δ modulators, we comment that the predictive noise filters of the previous

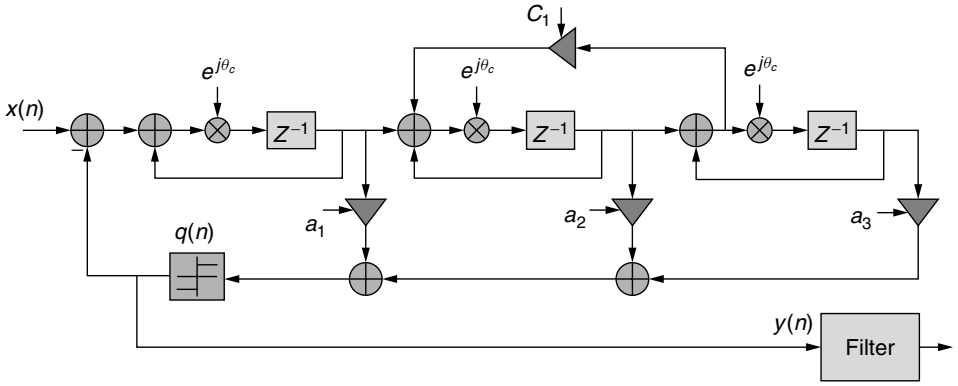


Figure 19. Bandpass sigma-delta converter with complex, phase-rotated resonators.

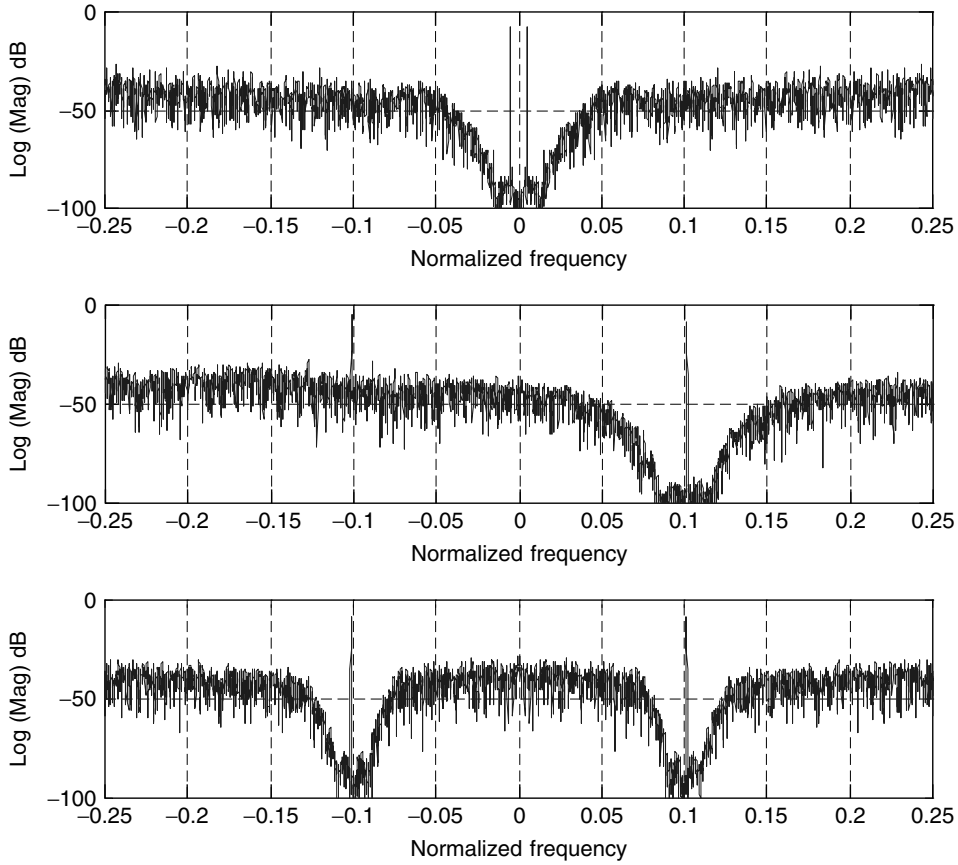


Figure 20. Spectrum of time series from 3-loop, 1-bit baseband prototype, from complex resonator, and from real resonator versions of prototype modulator.

section can also be trivially tuned to become bandpass Σ-Δ modulators. We manage this by spectrally shifting, as a symmetric or asymmetric translation, the spectra of the band-limited noise power spectrum in Eq. (27) used to define the correlation sequence of Eq. (28). As a consequence of this translation, the filter designed by the normal equation will be have a bandstop NTF rather than a baseband NTF.

6. STABILITY CONSIDERATIONS

We now address models of the quantizer in the Σ-Δ modulator. Our first order model of the quantizer is that

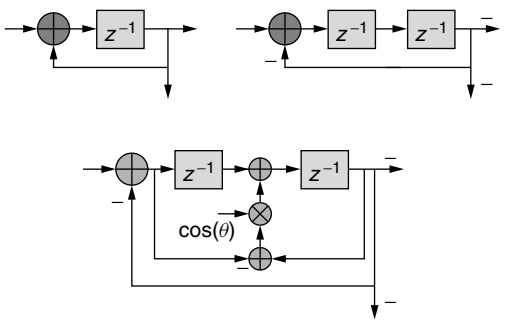


Figure 21. Prototype integrator, quarter-sample rate resonator, and all-pass tuned arbitrary frequency resonator.

of an independent additive noise source, the standard approximation used in a memoryless quantizer. When we have a small number of quantization levels, this model is poor because of the high correlation between the quantized signal and the quantization error. The quality of the model is improved by adding a random dither signal to the data samples prior to the quantization process. When embedded in a feedback loop, the model must also include a gain that may be amplitude-dependent. One linear model of the nonlinear quantizer $Q[u(n)]$ is shown in Fig. 22, where the input signal $u(n)$ is partitioned into two components, $u_{DC}(n)$ and $u_{AC}(n)$, subjected to their separate gains of K_0 and K_1 and then added to the additive noise source. Reasonable questions to ask are (1) whether this is a valid model and (2) over what range of input signal amplitudes is this a valid model.

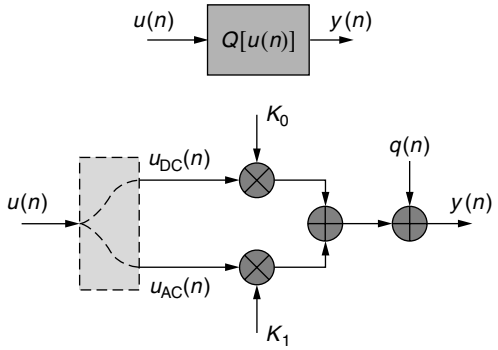


Figure 22. Linear model of quantizer in sigma-delta modulator feedback loop.

We answer the first question, concerning model validity by operating two 3-loop feedback models, one containing a standard 1-bit quantizer, and one an additive noise source in place of the quantizer. Figure 23a presents the

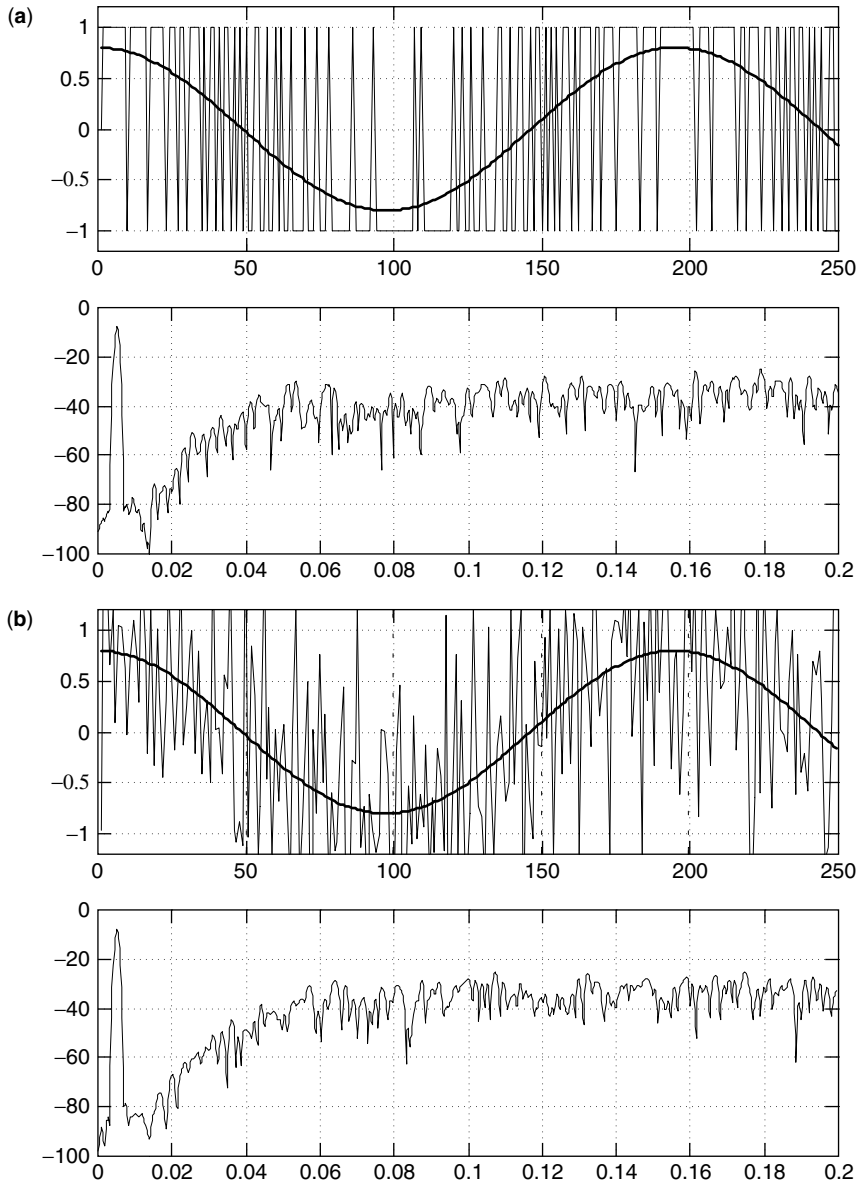


Figure 23. Input and output time series, and output spectra of a (**a**) 3-loop, 1-bit sigma-delta modulator and (**b**) 3-loop, linear additive noise model of sigma-delta modulator.

input and output time sequences with associated output spectrum of the 3-loop, 1-bit modulator. Figure 23b shows the corresponding figures for the same system containing the additive noise source with comparable variance. As projected, the spectra of the two loops is identical; the shaped inserted noise has the same spectrum as the shaped quantization noise. It is interesting to see how well the feedback loop suppresses the nonlinear quantizer behavior, enabling the linearized model to approximate the performance of the nonlinear system.

The next question we address is the range of input amplitudes for which the linearized model is a valid description of the nonlinear system. The standard approach to this inquiry is to collect statistics on the maximum level of the modulator's internal registers as a function of the input signal level. Figure 24a presents curves showing the maximum register levels observed in a 3-loop, 1-bit Σ-Δ modulator for test runs of length 16,384 samples with fixed (DC) levels over the range of 0–1, where 1 is the full-scale 1-bit quantizer output level. Also shown are the maximum register levels as a function of fixed input levels for the linear model of the loop with the inserted noise rather than quantizing noise. We note that the quantized loop and the linear model exhibit similar responses for fixed DC input levels spanning the range 0 to 0.6, and that the quantized loop exhibits poor stability for input signal levels beyond 0.6 and in fact becomes unstable at approximately 0.74.

Figure 24b presents a set of similar curves for the maximum register levels observed in the same 3-loop,

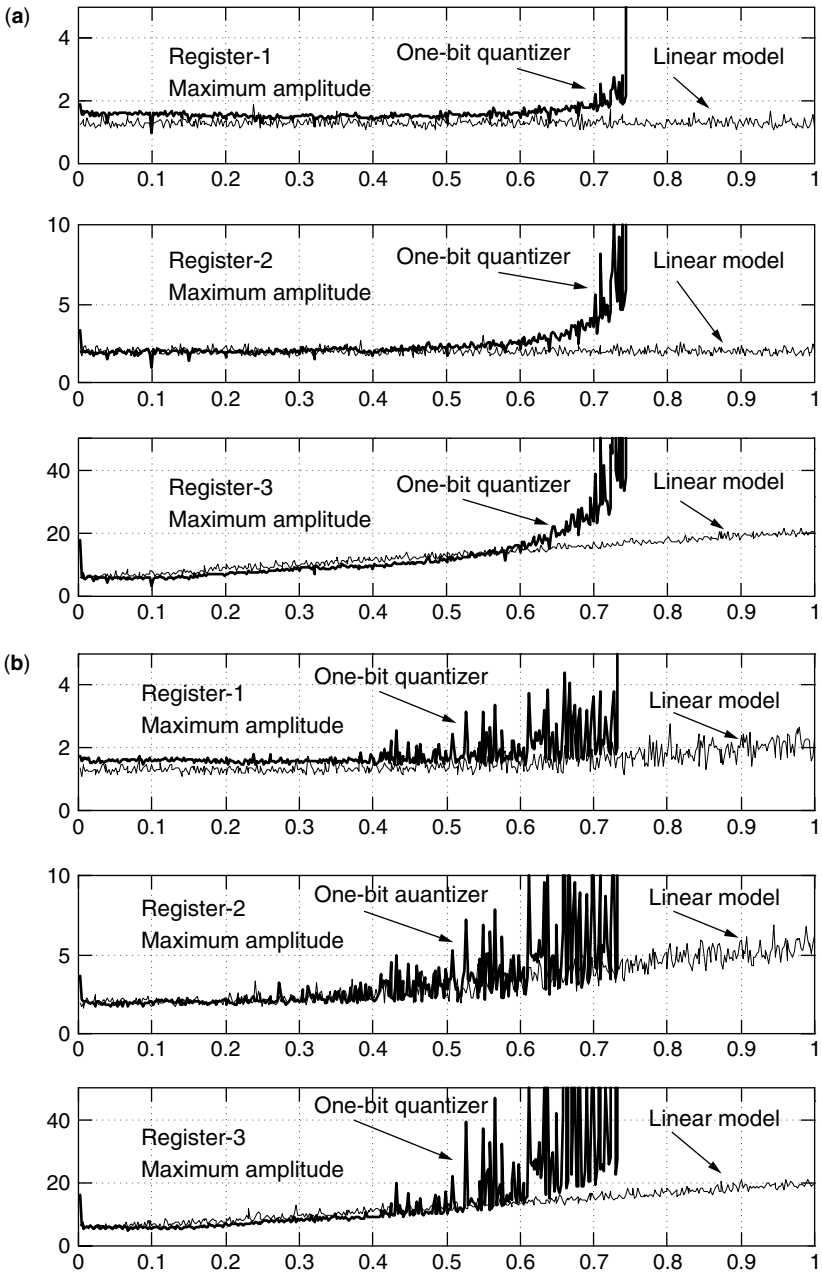


Figure 24. Amplitudes of internal states of 3-loop sigma-delta modulator and linearized noise model as function of (**a**) DC input level and (**b**) in-band AC input level.

1-bit Σ-Δ modulator for the same measurement conditions except that the input signal is an in-band sinusoid with fixed amplitude levels varied over the range of 0–1. As in the companion figure, also shown are the maximum register levels as a function of input levels for the linear model of the loop. We note that the quantized loop and the linear model exhibit similar responses for fixed AC input levels spanning the reduced range 0–0.4, with the quantized loop exhibiting increasingly poorer stability for input signal levels in the range 0.4–0.74 and in fact becoming unstable at 0.74. Note that the range of input AC levels for which the internal states avoid instability is smaller than the range of input DC levels. It is standard practice to restrict the range of input levels to half the input range to avoid the operating at the edge of unstable behavior.

The relationship that couples the instability of the sigma-delta modulator to the amplitude of the input resides in the fact that the linear model gain terms, K_0 and K_1, introduced in Fig. 22, decrease as the input amplitude increases. We can include the gain K_1 in the closed-loop expression for the noise transfer function, as shown in Eq. (32) to illustrate how the denominator of the transfer function changes with K_1, and hence varies with the input amplitude. From conventional feedback analysis we see that the closed-loop zeros are the open-loop poles, and that as K_1 varies from 0 to 1, the closed root poles migrate from the closed root zeros to the unity-gain closed root poles. The locus of the root migration as a function of K_1 is shown in Fig. 25. Note that for values of K_1 less than 0.276, the system poles are outside of the unit circle and the system is unstable:

$$\mathrm{NTF}(Z) = \frac{1}{1 + K_1 N(Z)/D(Z)} = \frac{D(Z)}{D(Z) + K_1 N(Z)} \qquad (32)$$

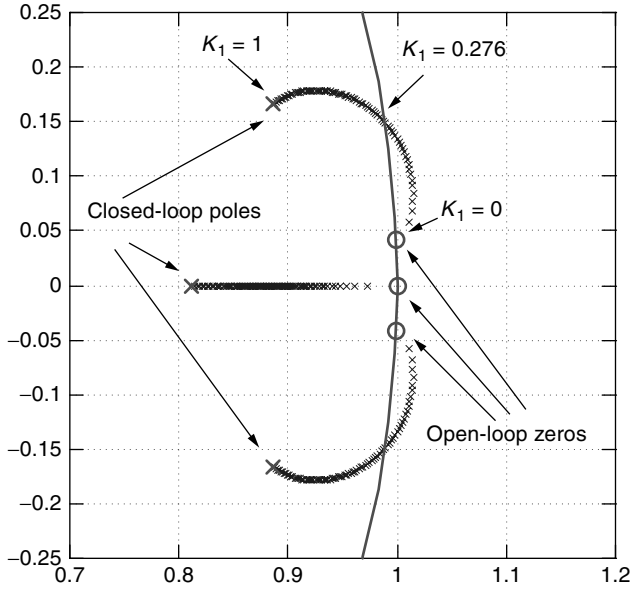


Figure 25. Root locus of closed-loop poles for third-order sigma-delta modulator demonstrating cause of instability as quantizer gain K_1 is reduced.

7. DIGITAL SIGNAL PROCESSING APPLICATIONS OF SIGMA-DELTA CONVERTERS

7.1. Σ-Δ Modulator in Data Preprocessor

The Σ-Δ converter can be used to improve the fidelity of any quantization process in a DSP system, many of which occur naturally and many by virtue of enlightened design. In the spirit of the latter option, the Σ-Δ modulator can be used as a preprocessor in any filtering task for which the filter bandwidth is a small fraction of its sample rate. Under this condition, the bandwidth of interest is already oversampled and can be requantized to a smaller number of bits to reduce the arithmetic resources required for the ensuing filtering process. One example we cite is that of a digital FIR filter to extract a downconverted 0.6-MHz-wide color subcarrier of a composite 6-MHz-wide NTSC signal sampled at 12.0 MHz. We note that the signal bandwidth processed by the filter is already oversampled by a factor of 20 occupying only 5% of the sample rate. The Σ-Δ can requantize the data from 10 bits to 1 bit with a shaped noise spectrum. The shaped noise spectrum preserves the quantizing noise level of the input signal in the signal bandwidth while permitting increased quantizing noise levels in the band to be rejected by the filtering process. Block diagrams of the two approaches to this processing task, conventional and sigma-delta preprocessing, are shown in Fig. 26. Here, a 4-tap preprocessor converts the 10-bit data to 1-bit so that the following 80-tap filter can be implemented without multiplies. Figure 27 shows the input and output time series as well as the output spectrum of the preprocessor. As expected, the 1-bit process successfully preserves the signal fidelity in the important band, the band processed by the filter.

7.2. Σ-Δ Modulator in DC Canceler

A second application of the Σ-Δ modulator as a DSP preprocessor is its insertion in the common signal task of canceling DC. DC components are generated and inserted in the signal a number of mechanisms, these include analog insertion due to untrimmed offsets in A/D converters, and digital insertion due to arithmetic truncation of two's-complement products and summations in various signal processing operations. The bias, or DC offset, on the order of a fraction of a bit per sample, does not appears to be a problem at first glance, but in

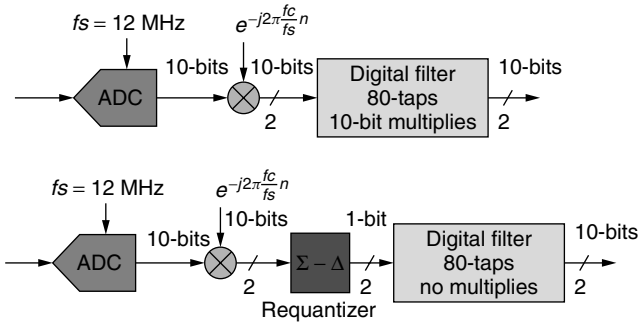


Figure 26. Narrowband processing with conventional and unconventional sigma-delta preprocess filtering.

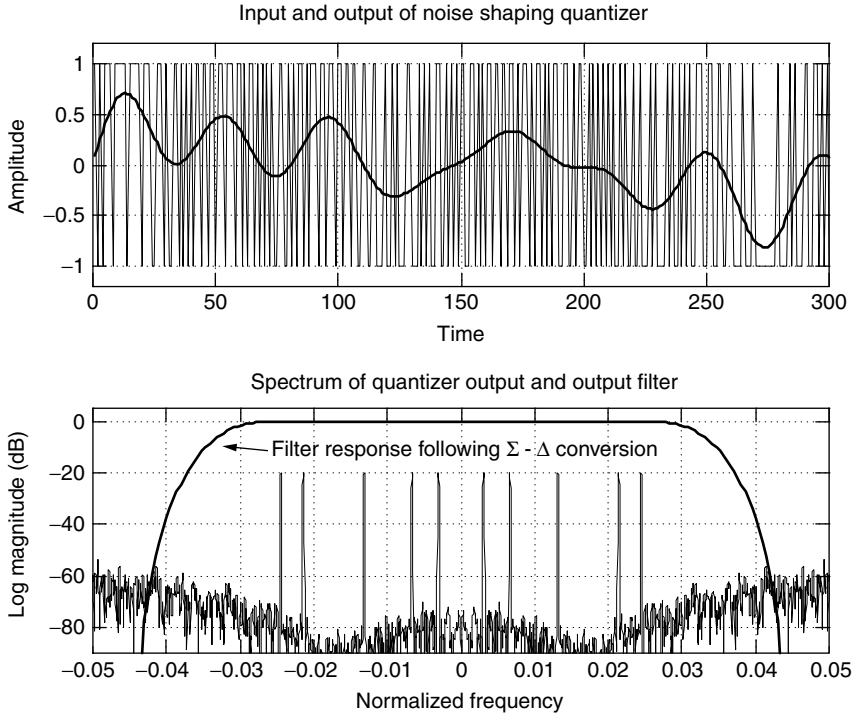


Figure 27. Time series and spectrum of signal preprocess quantized to 1 bit for reduced workload in subsequent filtering process.

fact becomes a problem when the samples are coherently processed. The fractional bit offset can grow to a many-bit offset as the signal experiences the coherent gain of subsequent filters as they reduce signal bandwidth. In some applications, the DC is of no consequence and can be ignored, but in others, the DC offset must be removed early in the signal processing path. We do this to preserve the dynamic range of a fixed-point data set and in digital receivers to avoid decision biases in the detection process following matched-filter processing.

Removal of the DC is performed by a DC notch filter usually implemented as a DC canceler of the form shown in Fig. 28. The integrator in the feedback loop of the filter becomes the transmission zero of this filter that also exhibits a nearby pole at $Z = (1 - \mu)$. The notch filter has the transfer function shown in Eq. (33), and we note that the distance, μ, between the zero and pole defines the bandwidth of the notch. The parameter μ is a small binary number, on the order of 2^{-10}, implemented by a right data shift. The integrator estimates the DC in the series, and the filter subtracts the DC from the input sequence. In order to cancel a DC term whose value is a fraction of a bit, the output of the canceler filter must grow additional bits to the right of the input data's binary point. On leaving the canceler, the lower-order data bits are discarded by the output quantizer that returns the number of output bits to the number of input bits for the benefit of subsequent arithmetic processing. This quantization discards the fractional part of the correction inserted by the canceler, so that the combined canceler and quantizer is capable of rejecting only the integer number of bits of the DC offset:

$$H(Z) = \frac{Z - 1}{Z - (1 - \mu)} \tag{33}$$

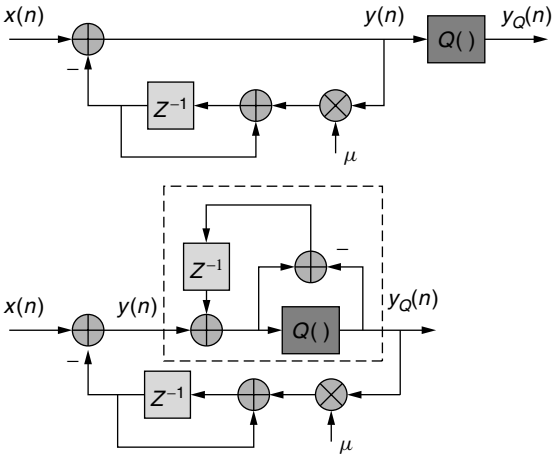


Figure 28. DC canceler with integrator in feedback path and external quantizer and same DC canceler with internal quantizer with noise feedback loop.

To preserve the fractional part of the corrected DC cancellation, we move the quantizer into the feedback loop and wrap a noise feedback loop around the quantizer. This modified structure is shown in the lower section of Fig. 28. The noise feedback quantizer can be placed in either the feedforward or in the feedback portion of the noise canceler. Figure 29 shows the spectrum of the input and output of the DC canceler before and after the external quantizer. We see the DC term present in the input spectrum and its absence in the output spectrum. A DC term is reinserted at the output of the external quantizer as the data samples are returned to 8-bit datawords. In the last figure we note that the internal quantizer with its

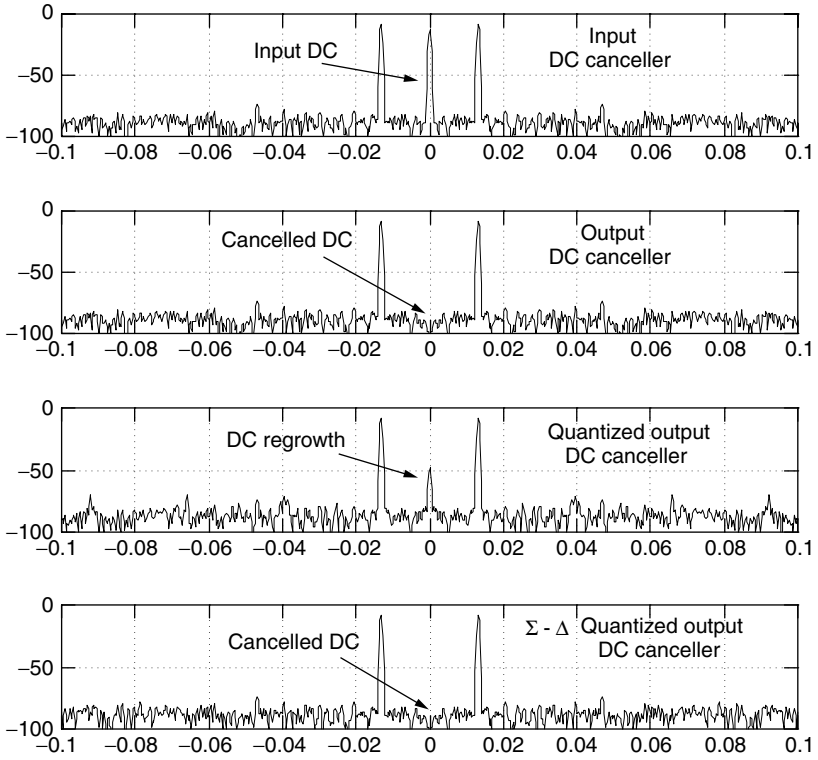


Figure 29. Spectra of input and output of DC canceler, and then output of canceler with 8-bit external quantizer and 8-bit internal noise feedback quantizer.

noise feedback loop does not reinsert the DC as the data are truncated to 8 bits.

7.3. Σ-Δ Regulator in DDS

The last example we cite for the use of a Σ-Δ requantizer is in a direct digital synthesizer (DDS). The DDS uses a fine-resolution overflowing phase accumulator to synthesize a specified phase–time profile for an output complex sinusoid. In a typical system, the output phase word, drawn from a 32-bit accumulator, is quantized to an 8-bit word used as an address to access the sine–cosine values stored in its lookup table. This structure is shown as the first of the three block diagrams in Fig. 30. The quantization forms a correlated error sequence, in fact a sawtooth-shaped periodic phase error of peak amplitude equal to the least significant output bit. The phase error sequence, the difference between the input and output of the quantizer, is shown as the top segment of Fig. 31. This error sequence phase-modulates the output sinusoid, generating an undesired set of line spectra centered about the output center frequency. The amplitude of the maximum phase modulation spurious line is 6 dB per address bit below the desired spectral line. The first curve of Fig. 32 presents the spectrum formed from an 8-bit lookup table containing 16-bit values of sine and cosine of the table address. We can clearly see the line structure and the −48-dB phase modulation spurious tone.

The spectral line structure related to the periodic phase error is undesirable, and the standard remedy to suppress this line structure is the use of additive dither to break up the regularity of the error sequence. The second block diagram of Fig. 30 illustrates the location for the dither

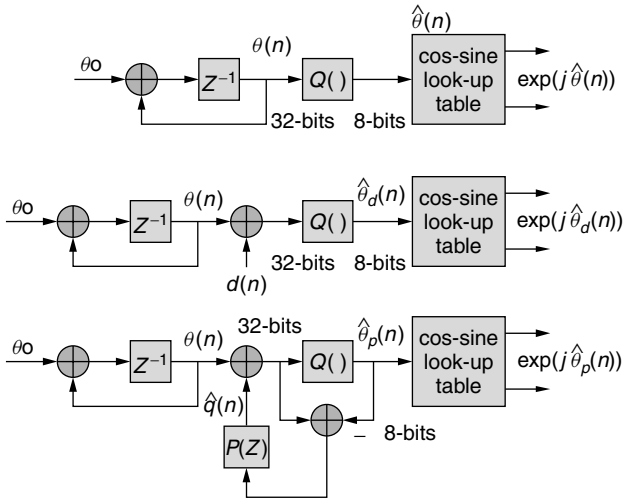


Figure 30. DDS quantization options for table addresses: option 1 — truncation; option 2 — dithered truncation; option 3 — noise feedback truncation.

insertion, the corresponding curve in Fig. 31 shows the dithered phase error structure, and the related curve in Fig. 32 shows the spectrum of the time series generated by the dithered address process. A proper dither suppresses the line structure and pulls the average phase noise level down by 12 dB or 2 bits, in this example from −48 to −60 dB.

It is an easy transition to replace the addition of random dither prior to the quantization process by the addition of correlated dither formed in a noise feedback

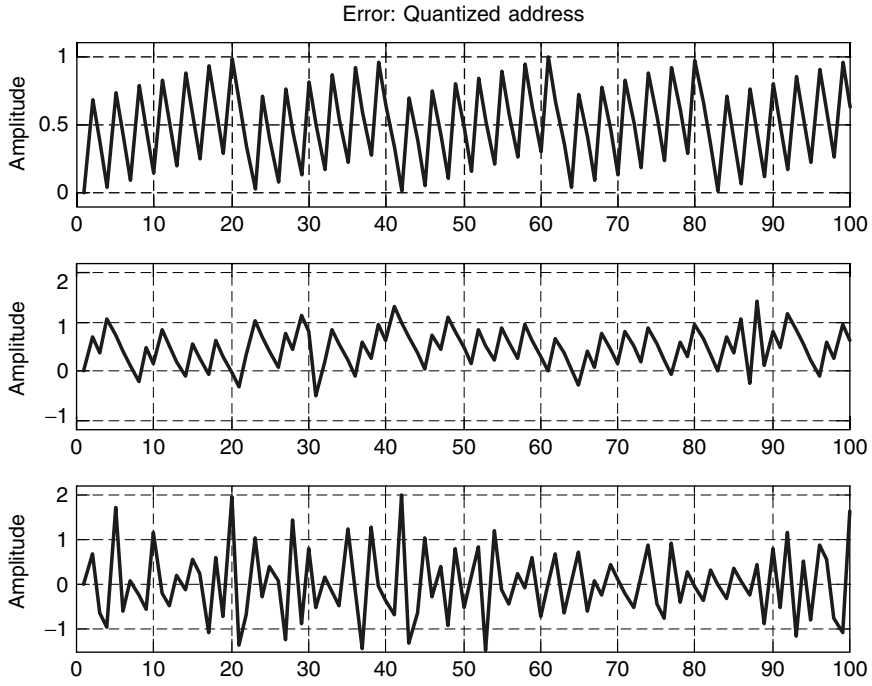


Figure 31. DDS phase error sequences obtained by different quantization options: option 1—truncation; option 2—dithered truncation; option 3—noise feedback truncation.

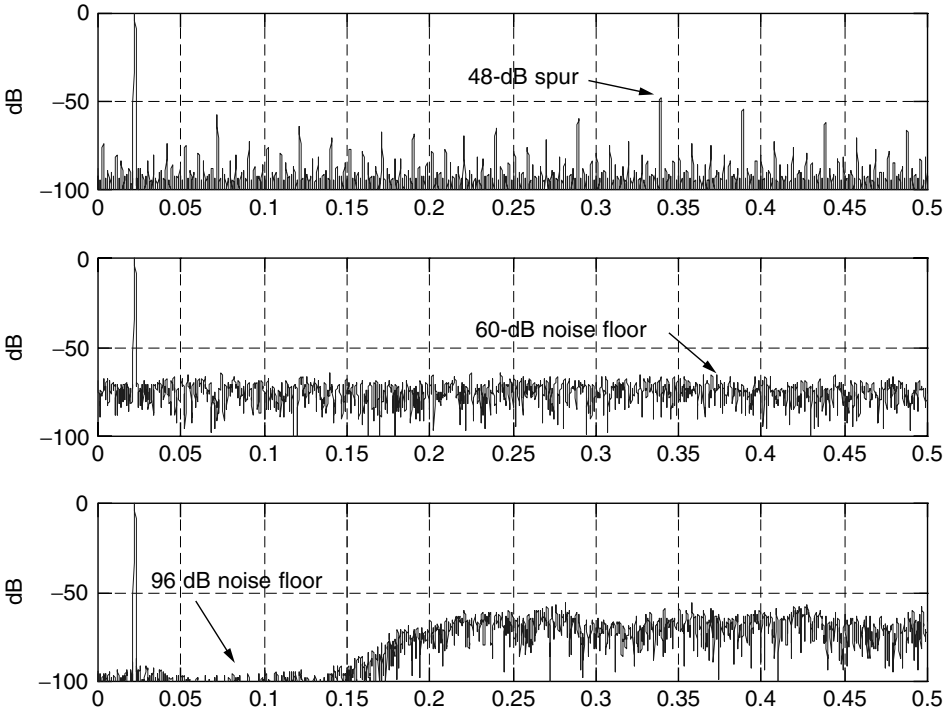


Figure 32. Spectrum of sinusoids obtained from 8-bit table with different quantization options: option 1—truncation; option 2—dithered truncation; option 3—noise feedback truncation.

loop. The third block diagram of Fig. 30 illustrates this option. The corresponding segment of Fig. 31 presents the dithered phase error generated by a 10-tap prediction filter designed using the normal equations presented in Eqs. (22)–(28) to obtain a noise suppression bandwidth equal to 25% of the sample rate. The shaped noise spectrum shown in the third segment of Fig. 32 illustrates a dramatic improvement in the level of phase noise modulation. The in-band level has been pulled down from −60 to −96 dB with rather small increases in out-of-band spectral level. This option of using the sigma-delta loop to manipulate addresses rather than signal amplitude leads to SNR improvements comparable to that available from a dithered 14-bit table with data samples extracted from an 8-bit table. Shorter tables, and predictors with smaller fractional bandwidth can form sinusoids at arbitrary center frequencies exhibiting remarkable fidelity over restricted close-in bandwidths.

8. CLOSING COMMENTS

We have reviewed the theory of operation of noise feedback, hence noise shaping, quantizers commonly called *sigma-delta modulators*. A number of architecture variants were described and their performance illustrated by examining spectra of their output series. Sigma-delta modulator architectures were described that reflect three primary design perspectives: cascade integrators, predictive feedback, and combinations of multiple simple modulators. The cascade integrator model uses minor-loop feedback to place and distribute open-loop poles on the unit circle and then major-loop feedback through a low-resolution quantizer to convert these poles to closed-loop noise transfer function zeros. The noise feedback model measures the quantizing error and uses band-limited prediction filters to predict and precancel the next value of quantizing noise contained within a selected segment of input bandwidth. The cascade models use a succession of low-order modulators to synthesize a high-order modulator. This is accomplished by measuring and canceling the shaped quantizing noise from previous stages with enhanced-shaped noise from later stages.

The tradeoff between oversample ratios, quantization SNR improvement, and order of feedback filter was described for multiple zero NTF. We examined narrow-band resonator-based models of the Σ-Δ process. In particular, we limited our discussion to transformations of baseband prototype systems that support tunable variants of existing systems. Feedback stability considerations were examined to explain the divergence of behavior between the linear and nonlinear models of the Σ-Δ modulator. Finally, a number of Σ-Δ applications were presented and illustrated in which the requantization process was enhanced or invoked to improve the performance of standard DSP-based signal processing tasks.

Material not addressed in this presentation included design and implementation of digital resampling filters that normally accompany the modulator when it is used in A/D and D/A applications. Here the emphasis was the embedding of the Σ-Δ modulator in a DSP system environment. Consistent with this emphasis, analog implementations of the noise feedback process were also intentionally not addressed in this review article.

BIOGRAPHY

Fred J. Harris received the B.S. degree in Electrical Engineering in 1961 from the Polytechnic Institute of Brooklyn, the M.S. degree in Electrical Engineering in 1967 from San Diego State University, and pursued Ph.D. work in electrical engineering from 1967 to 1971 at the University of California at San Diego. Since 1967 he has taught at San Diego State University, were he occupies the CUBIC Corp. Signal Processing Chair. Teaching and research areas include digital signal processing, multirate signal processing, communication systems, source coding, and modem design. He holds a number of patents involving multirate signal processing for satellite and cable modems as well as for sigma-delta implementations. He has contributed to a number of texts and handbooks on various aspects of signal processing. He is the traditional absentminded professor and drives secretaries and editors to distraction by requesting lowercase letters when spelling his name. He roams the world collecting old toys and sliderules and riding old railways.

FURTHER READING

Aziz P. M., H. V. Sorenson, and J. Van Der Spiegel, An overview of sigma delta converters: How a 1-bit ADC achieves more than 16-bit resolution, *IEEE Signal Process. Mag.* **13**(1): 61–84 (Jan. 1996).

Candy J. C. and G. C. Temes, *Oversampling Delta-Sigma Data Converters: Theory, Design, and Simulation*, IEEE Press, 1992.

Dick C. and F. Harris, Narrow-band FIR filtering with FPGAs using sigma-delta modulation encoding, *J. VLSI Signal Process. Signal Image Video Technol.* **14**(3): 265–282 (Dec. 1996).

Dick C. and F. Harris, FPGA signal processing using sigma-delta modulation, *IEEE Signal Process. Mag.* **17**(1): 20–35 (Jan. 2000).

Harris F. and B. McKnight, Error feedback loop linearizes direct digital synthesizers, *28th Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA. Oct. 30–Nov. 1, 1995.

Jayant N. S. and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, 1984, Chaps. 7 and 8.

Norsworthy S. R., R. Schreier, and G. C. Temes, *Delta-Sigma Data Converters: Theory, Design, and Simulation*, IEEE Press, 1997.

Uchimura K., T. Hayashi, T. Kimura, and A. Iwata, Oversampling A-to-D and D-to-A converters with multistage noise shaping modulators, *IEEE Trans. Acoust. Speech Signal Process.* **AASP-36**: 1899–1905 (Dec. 1988).

SIGNAL PROCESSING FOR MAGNETIC-RECORDING CHANNELS

EVANGELOS ELEFTHERIOU
IBM Research
Zurich Research Laboratory
Rüschlikon, Switzerland

1. INTRODUCTION

The main driving force of progress in magnetic-recording technology has been the need for vast and reliable storage. In the past four decades, the areal density of disk drives has increased ten million fold, leading to dramatic reductions in storage cost. In particular, the storage densities of high-end disk drives have been growing at a compound growth rate of 60% annually, starting in 1991. This is due to the introduction of the magnetoresistive (MR) recording head, and the advances in high-performance data and servo channels, as well as in VLSI technology [1]. Today's commercial disk drives use longitudinal recording and can store information at a density of approximately 50 Gbits per square inch. This unprecedented areal density is achieved while maintaining the stringent on-track error rate requirements of 10^{-8} to 10^{-9} before outer error-correction coding.

It is expected that this phenomenal growth in areal density of longitudinal recording will slow down because of the superparamagnetic effect [1]. This has increased research and development efforts in perpendicular recording, which since its inception [2,3] has promised to achieve much higher areal densities than longitudinal recording can. Although indications exist that perpendicular recording may be able to achieve ultra-high areal densities, the most recent laboratory demonstrations indicate that longitudinal recording is still marginally ahead of its perpendicular counterpart. Ultimately, perpendicular recording promises areal densities that are about four to five times higher than those of longitudinal recording [4,5]. However, there are considerable engineering challenges associated with the realization of this promise [6]. A transition from longitudinal to perpendicular recording would involve changes in various disk-drive subsystems, including the head, disk, head/disk interface, and servo. It is expected, however, that from a signal-processing and coding architecture point of view, the read electronics will not undergo substantial changes.

Although advances in head and media technologies have historically been the driving force behind areal density growth, digital signal processing and coding are increasingly recognized as a cost-efficient approach in achieving substantial areal density gains while preserving the high reliability of disk drives. The general similarity of the write/read process in a hard-disk drive to transmission and reception in communication systems has led to the adoption of adaptive equalization and coding techniques to the magnetic-recording channel. The classical communication channel perspective can be applied not only to study equalization, detection, and both inner and outer coding strategies but also to estimate the ultimate information-theoretic limits of longitudinal and perpendicular recording [7–9].

In the past decade, several digital signal-processing and coding techniques have been introduced into hard-disk drives to improve the error-rate performance at ever increasing normalized linear densities as well as to reduce manufacturing and servicing costs. In the early 1990s, partial-response class-4 (PR4) shaping in conjunction with maximum-likelihood sequence detection [10] replaced the peak detection systems employing run-length-limited (RLL) (d, k)-constrained codes, and paved the way for future applications of advanced coding and signal-processing techniques. For example, at moderate storage densities, the introduction of partial-response maximum-likelihood detection (PRML) [10] requires a new class of inner constrained codes. This class of codes, collectively known as PRML(G, I) codes, facilitates timing recovery and gain control, and limits the path memory length of the sequence detector. At higher normalized linear recording densities, generalized partial-response (PR) polynomials with real coefficients reduce noise enhancement at the equalizer output. In particular, shaping polynomials of the form $(1 - D)(1 + p(D))$ and $(1 - D^2)(1 + p(D))$, where D represents the unit delay and $p(D) = \sum_{\ell=1}^{L} p_\ell D^\ell$ is a finite impulse response predictor filter with real coefficients $\{p_\ell\}$, are significant in practice. Generalized PR channels in conjunction with sequence detection give rise to noise-predictive maximum-likelihood (NPML) systems [11–15]. The extension of the NPML detection scheme to handle data-dependent medium noise was proposed in [16–20].

Parallel to the developments on NPML detection for magnetic recording, maximum transition run (MTR) (j, k) codes were introduced as a means to provide coding gains for extended partial response (E^2PR) channels [21]. The theoretical underpinning of this new class of codes and practical constructions of codes that deal with the problem of quasi-catastrophic error propagation are presented in Ref. [22]. The unifying theory presented in Ref. [22] led to an exhaustive characterization of quasi-catastrophic error-propagation-free MTR codes, called MTR(j, k, t) codes, and revealed a connection between the conventional PRML(G, I)-constrained codes used in disk drives and the recently discovered MTR codes.

More recently, constrained codes, such as the PRML(G, I) and MTR(j, k, t) codes, have been combined with multiparity block codes to improve the bit error rate performance of the inner channel even further — at the expense, sometimes, of a slight decrease in code rate. These multiparity linear inner codes deliver substantial gains in performance when decoded by a so-called *soft* post-processor that follows the NPML detector and utilizes some form of reliability information [23–28]. Currently, a 16-state NPML detector for a generalized PR channel with a first-order null at dc followed by a post-processor for soft decoding of a combined multiparity/constrained code represents the de facto industry standard.

In this article the state of the art in signal processing and constrained coding for hard-disk drives is reviewed, with emphasis on the techniques that have been used in commercial systems. In Section 2 the magnetic recording system as a communications channel is introduced, and the various sources of noise are presented. In Section 3 the three families of modulation codes, namely, RLL(d, k), PRML(G, I), and MTR(j, k, t), that have predominantly been used in storage systems are discussed, and the latest developments on combined modulation/parity codes are presented. In Section 4 equalization and detection techniques are discussed, with emphasis on the PRML and NPML detection, and performance results are given. The soft-decoding methods for the combined modulation/parity codes via post-processors are presented in Section 5, and the NPML detector for data-dependent noise is reviewed in Section 6. Finally, Section 7 contains a brief discussion of future trends in signal processing and coding for magnetic storage.

2. MAGNETIC-RECORDING SYSTEM

2.1. Saturation Recording

In hard-disk drives, data is stored by longitudinal magnetization of a layer of magnetic media that has been deposited on a rigid disk. The data is recorded in concentric circles, called *tracks*, by applying an external field using a write head flying over the spinning disk. In general, the recording process is nonlinear, primarily because of the nonlinear nature of the magnetic medium. However, if the applied field exceeds a critical value the magnetization is

completely polarized in one direction. Therefore, at any point along a track, the magnetization can, in principle, be uniformly polarized in one of two possible directions, reflecting a recorded "0" or "1." This approach to storing information is referred to as saturation recording.

Writing is most commonly performed by inductive heads. When a current is applied, a magnetic field is generated through the head and across its gap. Depending on the amplitude of the applied current, the fringe field that escapes the gap can be strong enough to saturate the magnetic medium. By reversing the polarity of the current flowing through the coil, which is wound around the head, the direction of the magnetic field and consequently the direction of the magnetization of the medium are reversed. The size of the magnetization pattern depends on the rate by which the polarity of the current through the write head changes (data rate) and on the spinning velocity of the disk.

Magnetoresistive (MR) heads have become the prevailing magnetic-recording sensor for reading back the stored information because of the higher sensitivity and lower noise than the inductive heads. The term *magnetoresistive* refers to the physical phenomenon in which the resistivity of a metal changes in the presence of a magnetic field. Therefore, as the MR head flies over a spinning disk, the changes in the magnetic field emanating from the disk translate into changes of the resistivity of the head and consequently into changes of the voltage across the head. Because the MR head directly measures the flux from the medium, the head signal is independent of the velocity of the disk. As a consequence, the MR-head readback signal is as strong for high-RPM (revolutions per minute) server-class drives as it is for slowly rotating, mobile hard-disk drives.

2.2. A Communications Channel

One may consider a magnetic-recording system as a communications channel in which information, rather than being transmitted from one point in space to another, is stored at one point in time and retrieved at another. The objective in magnetic recording is to maximize the rate at which data is stored and retrieved as well as to maximize the density of the information stored per square inch of disk space. Figure 1 shows the general architecture of the recording system in commercial magnetic hard-disk drives.

User data are organized in sectors, with each sector typically containing 512 8-bit bytes. Upon a write request, the data are organized into sectors and fed into a Reed–Solomon (RS) encoder. The RS codewords are first byte-interleaved and then encoded by a constrained or modulation code. In storage systems, the latter code imposes constraints on the data stream being stored in the medium. In magnetic storage, modulation or constrained codes are used, for example, to facilitate the operation of the conventional peak detector, to provide timing information and eliminate quasi-catastrophic error propagation in sequence detection, or to eliminate certain predominant error events and increase the Euclidean distance between output sequences as seen by certain types of sequence detectors. Modulation codes usually employ a precoder, either of the form $1/(1 \oplus D)$ (in the RLL and MTR cases) or $1/(1 \oplus D^2)$ (in the (G, I) case), where $\oplus$ denotes modulo-2 addition. The main function of the precoder is to facilitate modulation code design. In today's recording systems the modulation encoder is followed by a parity encoder that usually appends one, two, three, or four parity bits to the modulation codeword. The output of

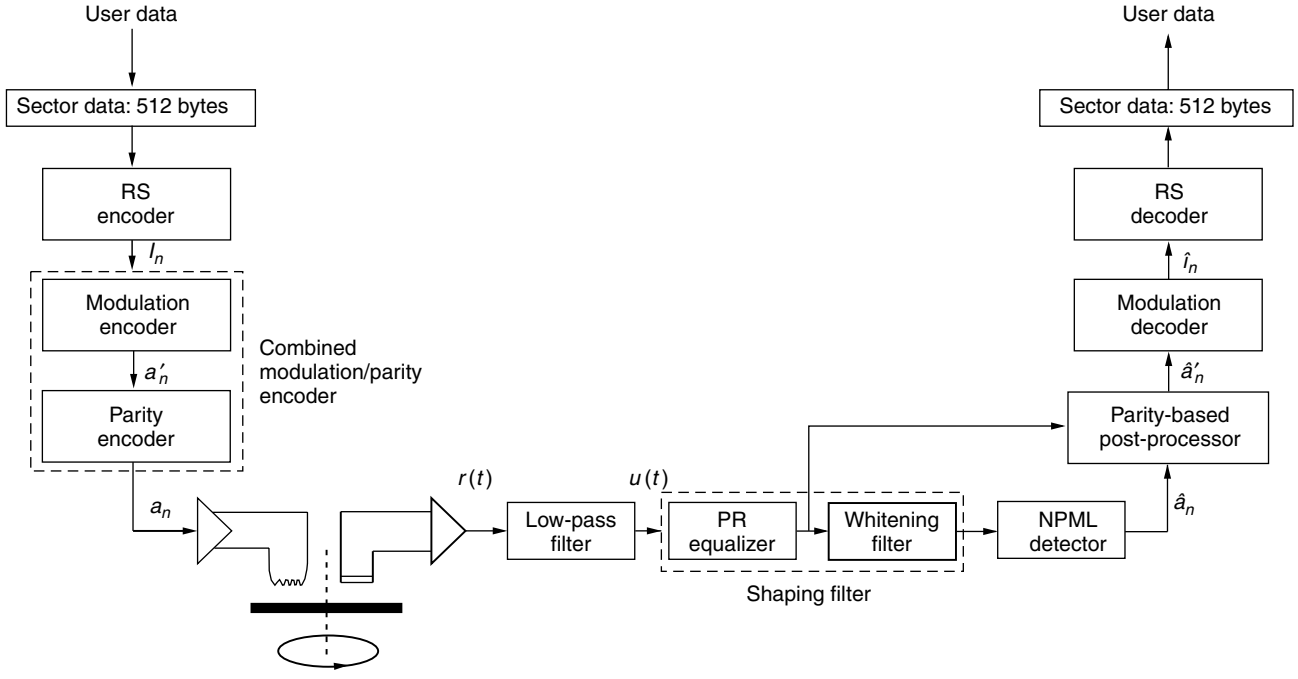


Figure 1. Recording system architecture.

the parity encoder is finally mapped into two-level channel input symbols, $a_i \in \{+1, -1\}$. These are the symbols that are actually being stored in the magnetic medium. The symbol sequence is first converted into current levels by the write head driver, and then the head stores the resulting sequence of rectangular pulses on disk tracks in the form of a series of transitions or magnetization waveforms.

For completeness, it is worthwhile to mention that there are two different formats for mapping the binary data sequence at the output of the modulation encoder to write current waveforms as shown in Fig. 2. In non-return-to-zero inverted (NRZI) recording, a binary 1 corresponds to a change in polarity of the write current, whereas a binary 0 corresponds to no change in polarity of the write current. In non-return-to-zero (NRZ) recording, the binary information sequence is mapped directly to the amplitude level of the write current waveform. From the signaling point of view, these two recording formats are related via the simple $1/(1 \oplus D)$ precoding function as shown in Fig. 3. As can be seen, the current waveform corresponding to NRZI recording of a particular data-input sequence is identical to the NRZ current waveform of the $1/(1 \oplus D)$ precoded data-input sequence. The NRZI recording format played an important role for peak-detection systems, whereas in today's sequence-detection systems, NRZ is the preferred recording format.

To read the data back from the hard disk, a read head is used that senses the changes of the magnetic flux that reflect the changes in polarity of the data sequence stored.

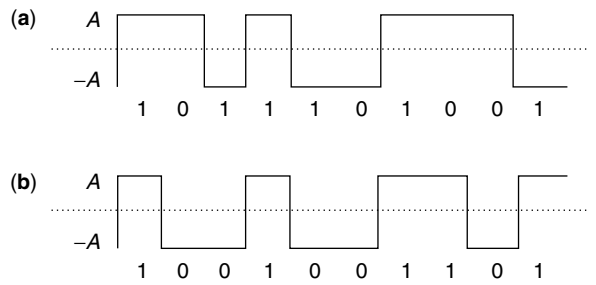


Figure 2. (**a**) NRZI and (**b**) NRZ signals.

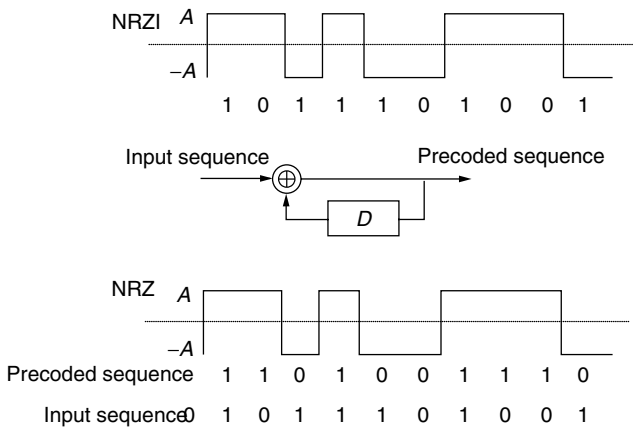


Figure 3. Relationship of NRZI and NRZ.

For an isolated positive transition, the response of the read head is a pulse $s(t)$, which is referred to as the isolated transition response. Analytically this isolated transition or step response is very well approximated by the Lorentzian pulse

$$s(t) = \frac{A}{1 + (2t/\mathrm{PW50})^2} \tag{1}$$

where A is the peak amplitude and PW50 denotes the width of the pulse at its 50% amplitude level. The value of PW50 depends on the physical width of the written transition, the characteristics of the magnetic medium, and the flying height of the head. Figure 4 shows the Lorenztian pulse response as well as the response to an isolated transition known as the Potter pulse [29] for the same value of PW50 and $A = 1$. The Lorentzian pulse is a single-parameter model, whereas the Potter pulse depends on the geometry of the head, the head-to-medium spacing, and the transition parameter [30, Chapter 6]. The Potter analytical pulse appears to be more appropriate for modeling the response of a MR head [30].

The data signal is read back via a low-pass filter (LPF) and a variable gain amplifier (VGA) as an analog signal, $r(t)$. The signal $r(t)$ is sampled periodically at times $t = iT$ to obtain a sequence of samples. The functions of the sampling device and VGA unit are controlled by the timing recovery and gain control loops, respectively. The sequence of samples is first shaped into a suitable PR signal format by the equalizer. The whitening filter then whitens the total distortion at the output of the equalizer, and the NPML sequence detector provides an initial estimate, $\hat{a}_i$, of the channel input symbols. Note that the functions of the PR equalizer and the whitening filter can be combined into a single filter. The initial estimate of the encoded data $\hat{a}_i$ coupled with the equalizer output, that is, hard and soft information, is fed to a noise-predictive parity-based post-processor. The post-processor is a suboptimum soft decision decoder for the parity code that corrects a specified number of the most likely error events at the output of the NPML detector by exploiting the parity information in the incoming sequence. The post-processor produces

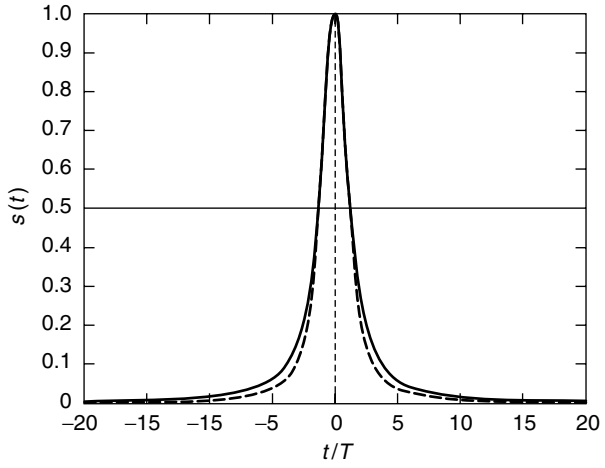


Figure 4. Isolated transition response in a magnetic-recording system. Solid line: Lorentzian pulse, dashed line: Potter pulse.

the final estimate $\hat{a}_i'$ of the modulation-encoded symbols. This sequence is fed to the modulation decoder, which delivers the final decisions. The overall performance of the recording system is very sensitive to error propagation at the modulation decoder. Thus, in practice, the design of modulation decoders with limited error propagation is crucial. Finally, after byte-deinterleaving, the RS decoder corrects any residual errors to obtain a highly reliable estimate of the original user data stored.

2.3. Linear Intersymbol-Interference Channel Model

As mentioned above, the two-level channel-input sequence, $\{a_i\}$, modulates a current source to produce a rectangular current waveform with amplitude -1 or $+1$. This rectangular current waveform can be viewed as the input to the write-head/medium/read-head assembly that produces an output voltage waveform that in essence is the differentiated and low-pass-filtered version of the input waveform corrupted by noise [31,32]. Experimental data have demonstrated that with write precompensation of magnetization transition shifts, the readback waveform is very well approximated by the sum of the appropriate series of isolated transition responses. Adopting therefore a linear model for the write/read process on a magnetic disk, the readback waveform $r(t)$ can be expressed as [31,32]

$$r(t) = \frac{d}{dt}\left[\sum_i a_i \Pi(t - iT)\right] \circledast s(t) + \eta_e(t) \tag{2}$$

where $\Pi(t)$ is a unit-amplitude rectangular pulse of duration T, $s(t)$ is the isolated transition or step response, $\eta_e(t)$ represents the additive white Gaussian electronics noise arising from read head and preamplifier, and $\circledast$ denotes convolution. The effect of medium noise in the case of well-dispersed particulate media can be taken into account by considering a second additive white Gaussian source, $\eta_m(t)$, at the channel input [33]. In this case the model in Eq. (2) becomes

$$r(t) = \frac{d}{dt}\left[\sum_i a_i \Pi(t - iT) + \eta_m(t)\right] \circledast s(t) + \eta_e(t) \tag{3}$$

It can readily be seen that the readback signal can equivalently be expressed as

$$r(t) = \sum_i a_i [s(t - iT) - s(t - T - iT)] + \eta(t) \tag{4}$$

where

$$\eta(t) = \frac{d\eta_m(t)}{dt} \circledast s(t) + \eta_e(t) \tag{5}$$

and $\dfrac{d\eta_m(t)}{dt} \circledast s(t)$ represents the medium-noise contribution to the total noise. Equation (4) describes a pulse-amplitude-modulated waveform of a binary data sequence $\{a_i\}$ transmitted at a rate of $1/T$ through a dispersive linear channel with effective impulse response $h(t) = s(t) - s(t - T)$ and received in the presence of additive noise. In a magnetic-recording system, the impulse response $h(t)$, which represents the effective impulse response of the overall magnetic-recording channel, is called pulse response because it corresponds to the response of the head/medium to a rectangular pulse. Furthermore, the quantity $D_c = \mathrm{PW50}/T$ is called normalized linear density. For a given PW50, the smaller the symbol interval T, that is, the closer the transitions, the higher the value of D_c, which implies the larger the linear density of the recording system. Conversely, the smaller the symbol interval T, the higher the interaction and overlap between the pulses, which gives rise to intersymbol interference (ISI) in a sequence of data symbols. Today's high-performance digital magnetic-recording systems operate at normalized linear densities in the range of $2.5 \le D_c \le 3.5$, where severe ISI is present in the readback signal. In such a case, the recovery of the data sequence from the readback signal requires advanced equalization and detection techniques that compensate for the presence of ISI.

Alternatively, the readback signal may be expressed in the equivalent form:

$$r(t) = \sum_i b_i s(t - iT) + \eta(t) \tag{6}$$

where $b_i = a_i - a_{i-1}$. In this case the symbol sequence $\{b_i\}$ takes values from the ternary alphabet $\{+2, 0, -2\}$. Figure 5 shows a linear ISI model of the magnetic-recording channel, where $s'(t)$ denotes the first derivative of the isolated transition response $s(t)$. The encoded NRZ data symbols a_n are passed through a $1 - D$ filter. A nonzero output of this filter corresponds to a positive or negative transition of the write current and, consequently, a transition in the magnetization pattern on the disk. The output of the $1 - D$ filter is then fed to a linear filter with impulse response $s(t)$, representing a Lorentzian, Potter, or any other analytic or experimental read-head response to an isolated transition. Finally, the readback signal is generated by adding white and colored Gaussian noise.

In the frequency domain, the overall response of the linear model may be expressed as $H(f) = S(f)[1 - e^{-j2\pi fT}]$, where $S(f)$ indicates the Fourier transform of the isolated transition response. The frequency response characteristics $H(f)$, plotted as a function of the normalized frequency fT and with D_c as a parameter, are illustrated in Fig. 6. As expected, the frequency response exhibits a spectral null at $f = 0$ because of the factor $[1 - e^{-j2\pi fT}]$. Furthermore, there is substantial high-frequency attenuation, which increases as the linear normalized density D_c is increased.

So far a linear channel model has been assumed, that is, the noise-free readback signal can be derived by linear

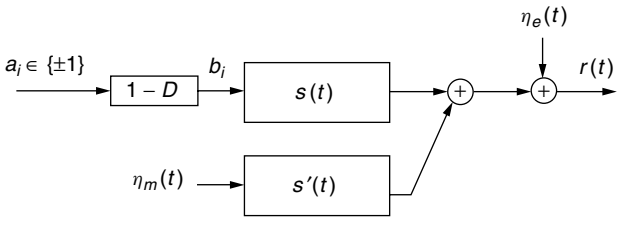


Figure 5. Linear ISI model for a magnetic-recording system.

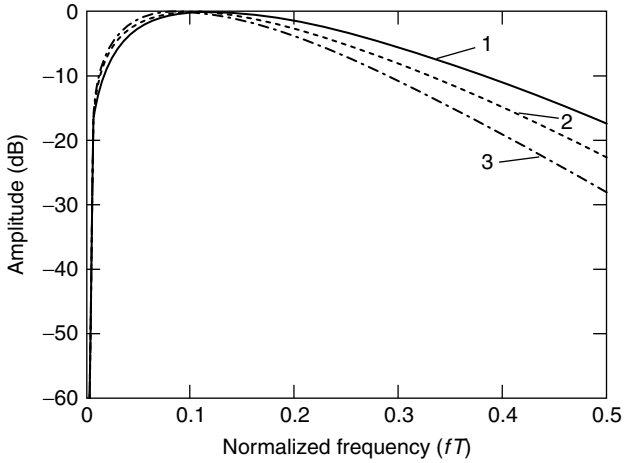


Figure 6. Frequency response to a square pulse of duration T. Curve 1: PW50/$T = 2.5$; curve 2: PW50/$T = 3.0$, and curve 3: PW50/$T = 3.5$.

superposition of isolated step responses, and the noise sources are stationary and additive. At high recording densities, the noise tends to be colored and enhanced by the linear equalizer, the nonstationary data-dependent noise becomes more dominant, and the signal becomes inherently nonlinear. Finally, today's recording systems employ thin-film media, which are dominated, as we will see below, by nonadditive noise sources.

2.4. The Microtrack Model

Recording systems using thin-film media may also exhibit a nonadditive noise known as transition noise. In general, the transition noise is due to fluctuations concentrated close to the recorded transition centers and is attributed to the random microstructural properties of the grains in thin-film recording media [34]. The effect of transition noise on equalization and data detection is more difficult to analyze than that of Gaussian noise owing to its nonstationary and data-dependent characteristics. A simple model for transition noise can be obtained by modeling the width and the position of the isolated transition response as random variables. Taking a Taylor series expansion with respect to small random deviations from the nominal value in the position and width, we can arrive at a channel model with multiplicative position jitter and pulse-widening noise sources [35].

The microtrack model developed in [18] is more general and allows a rather accurate modeling of the noise that occurs when a transition is written in thin-film magnetic-recording media. This model imitates the random zigzag effects when a transition is written. The random zigzag form of a transition is captured by dividing the recording track into N equally-sized microtracks. Figure 7 shows a track modeled by $N = 4$ microtracks, where the vertical dashed lines indicate the ideal positions of two consecutive transitions. The arrows show the direction of magnetization, whereas the short vertical lines on each microtrack indicate the corresponding positions of an instantaneous magnetization reversal. The positions of these instantaneous reversals or flips follow a specified probability density function and are chosen independently for each microtrack. Therefore, if $s(t)$ is the response to an ideal isolated transition across the entire track, then the response of the ℓ-th microtrack to a magnetization reversal at position τ_ℓ is $s(t - \tau_\ell)/N$ [18]. The noiseless output of the magnetic-recording channel to a single transition is then given by

$$\hat{s}(t) = \frac{1}{N} \sum_{\ell=1}^{N} s(t - \tau_\ell) \tag{7}$$

where τ_ℓ is the random shift associated with the ℓ-th microtrack. This random shift or jitter, τ_ℓ, is modeled as an independent and identically distributed (i.i.d.) process according to the derivative of the average cross-track magnetization profile. If the average cross-track magnetization profile has a *tanh*-shape, the jitter probability density function (pdf) is given by [34]

$$p_\tau(\tau) = \frac{1}{\pi a} \operatorname{sech}^2\left(\frac{2\tau}{\pi a}\right) = \frac{1}{\pi a} / \cosh^2\left(\frac{2\tau}{\pi a}\right) \tag{8}$$

In the case of an *erf*-shaped cross-track magnetization profile, the pdf is given by [18]

$$p_\tau(\tau) = \frac{1}{\pi a} \exp\left\{-\left(\frac{\tau}{a\sqrt{\pi}}\right)^2\right\} \tag{9}$$

In both cases a is known as the transition-width parameter, a quantity usually determined experimentally.

The microtrack model is specified by two parameters: the number of microtracks N and the transition-width parameter a. In its more general form, a third parameter, L_e, can be introduced that characterizes partial erasure and its effects. This parameter specifies the threshold below which two adjacent transitions on the same microtrack erase each other [18], see Fig. 7. The microtrack model allows a separate analysis of the write process and the transition noise. The measure of goodness of the write process is the steepness of the cross-magnetization profile. This steepness depends on the transition-width parameter a. An ideal transition

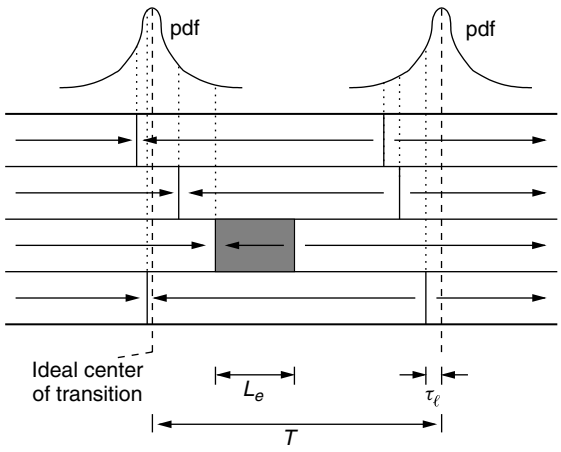


Figure 7. Microtrack model with $N = 4$ microtracks and erasure on the second microtrack.

has width zero, that is, all magnetic particles change their polarization at the same location. However, real transitions exhibit a *tanh*-shape, an *erf*-shape, or another experimentally determined cross-polarization profile. The transition- or data-dependent medium noise produced by the microtrack model is the difference between the actual output and the expected or average output with respect to the jitter distribution, that is, $\hat{s}(t) - E[\hat{s}(t)]$. As the number of microtracks increases, the transition noise decreases. By allowing $N \to \infty$ then $\hat{s}(t) \to E[\hat{s}(t)] = s(t) * p_\tau(t)$, and the transition noise is zero. In this case only electronics noise and the nonideal write process affect the performance of the magnetic-recording system via the transition-width parameter a. Another way to eliminate the effects of transition noise is to set $a = 0$. In this case $E[\hat{s}(t)] = s(t)$, and the magnetic-recording system behaves according to the linear ISI model described above.

Using Taylor's series expansion and keeping only the first two predominant terms, the transition noise for a single transition can be approximated by [18]

$$\begin{aligned}\hat{s}(t) - E[\hat{s}(t)] &= -s'(t)\left[\frac{1}{N}\sum_{\ell=1}^{N}\tau_\ell\right] \\ &\quad + s''(t)\left[\frac{1}{2N}\sum_{\ell=1}^{N}\tau_\ell^2 - \frac{1}{2}E[\tau_\ell^2]\right] \\ &\quad + \cdots \simeq s'(t)n_1 + s''(t)n_2\end{aligned} \tag{10}$$

where $n_1 = \dfrac{1}{N}\sum_{\ell=1}^{N}\tau_\ell$ and $n_2 = \dfrac{1}{2N}\sum_{\ell=1}^{N}\tau_\ell^2 - \dfrac{1}{2}E[\tau_\ell^2]$ are zero-mean random variables [18]. The random variable n_1 controls the amount of position jitter noise and is referred to as the position jitter variable, whereas the random variable n_2 controls the amount of pulse-width variation noise and is referred to as the width variable.

In general, the readback signal according to the microtrack model is expressed as

$$r(t) = \frac{1}{N}\sum_i b_i \sum_{\ell=1}^{N} s(t - iT - \tau_{\ell,i}) + \eta_e(t) \tag{11}$$

where $\tau_{\ell,i}$ is the random shift or jitter associated with the ℓ-th microtrack at the i-th symbol interval, and $\eta_e(t)$ represents additive white Gaussian noise (AWGN) characterized by its one-sided power spectral density N_0. Thus, the behavior of the magnetic-recording system can be described by the five-parameter (PW50/T, N, L_e, a, and N_0) model as shown in Fig. 8. Although for the magnetic-recording channel this model is quite versatile and encompasses the effects of both medium noise in thin-film media and electronics noise, the linear ISI model with additive Gaussian noise is still commonly used to estimate the performance of detection and coding schemes, even at ultra-high densities, and yields very accurate results.

3. MODULATION CODES

The basic principles of modulation coding, also known as coding for input-constrained channels, was established in the classic study of discrete noiseless channels [36]. With modulation coding a desired constraint is imposed on the data-input sequence so that the encoded data stream satisfies certain properties in the time or frequency domain. These codes are very important in digital-recording devices and have become ubiquitous in all data-storage applications [37].

A constrained system is represented by a labeled, directed graph, or a finite-state transition diagram (FSTD). An FSTD consists of states (or vertices) and labeled, directed transitions between states such that the allowable constrained sequences are precisely the sequences obtained by traversing paths of the diagram. The FSTD is called deterministic if at each state all outgoing transitions have distinct labels. The capacity of a constrained set of sequences represents the maximum achievable code rate of an encoder generating sequences satisfying the underlying constraint. It is given by $\log_2 \lambda_{\max}(A)$, where $\lambda_{\max}(A)$ is the largest real eigenvalue of the adjacency matrix associated with a deterministic FSTD that represents the constrained set of sequences. Among the various methods for constructing modulation codes, the state-splitting algorithm in [38] provides a systematic approach to designing finite-state encoders and sliding-block decoders for finite-type constrained systems. In practice, however, the right choices must be made during the code-construction procedure, irrespective of the approach used for code design. For example, quite often look-ahead coding techniques yield more efficient designs than the systematic state-splitting approach does. A more detailed treatment of finite-state modulation codes and their application to digital data recording can be found in [37,39].

In this section the most important classes of modulation codes will be reviewed, with emphasis on the application to magnetic recording, in particular, to hard-disk drives. Finally, the combination of the PRML(G, I) and MTR(j, k, t) codes with parity block codes in order to improve the bit error rate performance will also be discussed.

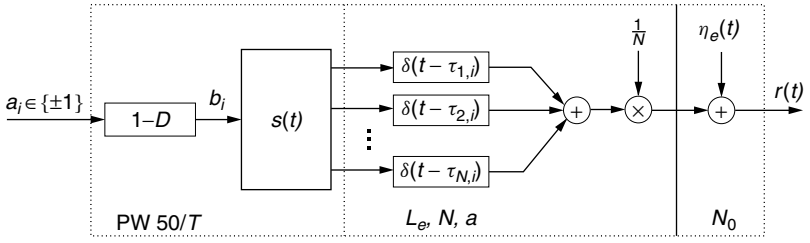


Figure 8. Data-dependent noise channel model for a magnetic-recording system. From [9], ©; IEEE 2002.

3.1. RLL(d, k) Codes

Modulation codes that impose a restriction on the number of consecutive 1s and 0s in the encoded data sequence are generally called RLL(d, k) codes. The code parameters d and k are nonnegative integers with k always larger than d, where d indicates the minimum number of 0s between two 1s and k indicates the maximum number of zeros between two 1s. At low-linear recording densities, peak-detection systems employing RLL(d, k)-constrained codes have been predominant in digital magnetic storage. RLL(d, k) codes reduce the effect of pulse interference and prevent the loss of clock synchronization. When used with the NRZ recording format, the d-constraint has the effect of spreading the transitions by at least $d + 1$ symbols further apart, thereby minimizing intersymbol interference and nonlinear distortion. The k-constraint sets an upper limit on the run of identical symbols to $k + 1$ so that useful timing information can always be extracted from the readback signal.

The set of sequences that satisfy the (d, k) constraints can be generated by the FSTD shown in Fig. 9. For (d, k)-constrained sequences the capacity can be computed as the base-2 logarithm of the largest real solution of one of the following equations [40]:

$$x^{k+2} - x^{k+1} - x^{k+1-d} = 1, \qquad k < \infty \tag{12}$$
$$x^{d+1} - x^{d} = 1, \qquad k = \infty,$$

Table 1 lists the capacity Cap(RLL(d, k)) for various values of d and k. The use of (d, k)-constrained sequences, with $d \geq 1$, allows the information density along a track to be increased while keeping the separation between adjacent recorded transitions fixed. The quantity $(d + 1)$Cap(RLL(d, k)), called the density ratio or packing density [37,41], is a direct measure of the increase in linear recording density as a function of d and of the capacity of the (d, k)-constrained sequences. Clearly, the packing density can be made arbitrarily large by increasing d. Conversely, large values of d lead to codes with very low rate, which implies high-recording symbol rates, and thus renders these codes impractical for storage systems that are limited by the clock speed. Moreover, large values of d lead to a more sensitive timing and detection window. In practical recording systems, codes with $d \leq 2$ have been used. In particular, the rate-2/3 RLL$(1, 7)$ (packing density of 4/3) and rate-1/2 RLL $(2, 7)$ (packing density of 3/2) codes have been widely used in the digital-recording industry at low-normalized densities around $D_c \simeq 1$ [37].

There are several approaches for constructing a rate-2/3 RLL$(1, 7)$ code. The simple construction in Ref. [42] is based on the code assignment of Table 2. Freely concatenating any of the four codewords from this list would violate the $d = 1$ constraint at codeword boundaries. The substitutions given in Table 3 avoid the violations at code boundaries and give rise to a *look-ahead encoder* and *sliding-block decoder*. The encoding table of the other popular rate-1/2 RLL$(2, 7)$ code is illustrated in Table 4.

3.2. PRML(G, I) Codes

The PRML scheme introduced in the early 1990s to advanced hard-disk drives operating at moderate linear recording densities, that is, approximately in the range $1.5 \leq D_c \leq 2.3$, requires a different type of constraint called (G, I) constraint [10,41]. The need for this new constraint is twofold. First, long runs of zero samples (noise-free samples) at the PR4 equalizer output can degrade the tracking performance of the timing recovery

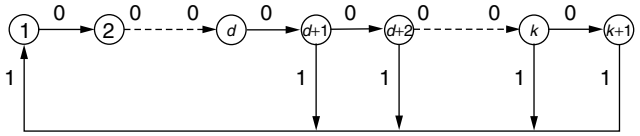


Figure 9. FSDT for RLL(d, k)-constrained sequences.

Table 1. Capacity of RLL(d, k) Constraints

k	$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d = 4$
1	0.6942				
2	0.8791	0.4057			
3	0.9468	0.5515	0.2878		
4	0.9752	0.6174	0.4057	0.2232	
5	0.9881	0.6509	0.4650	0.3218	0.1823
6	0.9942	0.6690	0.4979	0.3746	0.2669
7	0.9971	0.6793	0.5174	0.4057	0.3142
8	0.9986	0.6853	0.5293	0.4251	0.3432
9	0.9993	0.6888	0.5369	0.4376	0.3620
∞	1.00	0.6942	0.5515	0.4650	0.4057

Table 2. Encoder for the Rate-2/3 RLL(1, 7) Code

Information bits	Encoded bits
00	101
01	100
10	001
11	010

Table 3. Substitution Table for the Rate-2/3 RLL(1, 7) Code

Prohibited Pattern	Substitute Pattern
101,101	101,000
101,100	100,000
001,101	001,000
001,100	010,000

Table 4. Encoder/Decoder for the Rate-1/2 RLL(2, 7) Code

Information Bits	Encoded Bits
10	0100
11	1000
000	000100
010	100100
011	001000
0010	00100100
0011	00001000

and gain control loops. Thus, as in peak detection systems and the k-constraint described above, it is necessary to introduce a global constraint G to limit the number of zero samples at the equalizer output. Second, another constraint is necessary to limit the path memory requirements and hence force a finite decision delay in PRML detection, without incurring any significant performance degradation in the sequence of estimates produced by the detector. This constraint, referred to as the I-constraint, is related to the so-called quasi-catastrophic error propagation of PR systems [43], and imposes a limitation on the length of runs of zero samples in each of the odd and even interleaved subsequences at the PR4 equalizer output. In general, a trellis is called quasi-catastrophic if there are distinct states for which some of the output sequences starting from these states are identical and the total probability of all such indistinguishable output sequences is zero [43].

The issue of undesired sequences and the application of PRML(G, I) codes to eliminate them are more general and not restricted to PR4 shaping only. The above discussion has indicated that long strings of zeros at the PR4 channel output as well as in each of the subsequences of even and odd bit positions at the PR4 channel output constitute a set of undesired sequences that need to be eliminated. It can readily be shown that the same holds for all types of generalized PR shaping polynomials of the form $(1-D)^m(1+D)^n(1+p(D)), n, m \geq 1$, where $p(D)$ has no roots on the unit circle [22].

To facilitate the timing and gain control algorithms, in general only those channel input sequences need to be eliminated that have spectral energy at the frequencies where the generalized PR polynomial used for shaping has spectral nulls. For example, for shaping polynomials of the form $(1-D)^m(1+p(D)), m \geq 1$, exhibiting and m-th order spectral null at dc, the channel input sequences $(+1)$ and (-1) with a spectral null at dc should be eliminated. The notation (S) indicates the sequence obtained by periodically repeating the string S infinitely many times, e.g., $(ab) = ababab\cdots$. The k-constraint presented above limits the maximum length of 0s at the input of the $1/(1 \oplus D)$ precoder to k or, equivalently, the length of channel input patterns (after precoding) of type $(+1)$, (-1) to $k+1$. Similarly, it can readily be seen that for an arbitrary channel-shaping polynomial with spectral nulls at both the dc and the Nyquist frequency, the channel input sequences $(+1-1)$ and $(-1+1)$ with a spectral null at the Nyquist frequency should also be eliminated. For example, for channel-shaping polynomials of the form $(1-D^2)(1+p(D))$, the channel input sequences $(+1)$, (-1), with a spectral line at dc, as well as the sequences $(+1-1)$, $(-1+1)$, with a spectral line at the Nyquist frequency, should be eliminated. The G-constraint mentioned above limits the maximum length of 0s at the input of the $1/(1 \oplus D^2)$ precoder to G or, equivalently, the maximum length of channel input patterns of all four types $(+1)$, (-1), $(+1-1)$ and $(-1+1)$ to $G+2$ [22].

In connection with the PR4 shaping polynomial above, we have seen that another desirable code property is the elimination of quasi-catastrophic error propagation, which is inherent in maximum likelihood sequence detection of any generalized PR channel with spectral nulls [43]. This property allows the path memory size of the sequence detector to be reduced without degrading its bit error rate performance. Quasi-catastrophic error propagation is prevented by eliminating all channel-input error sequences $\{\varepsilon_i\} = \{a_i - \hat{a}_i\}$ that have spectral energy at those frequencies where the channel has spectral nulls. For generalized PR polynominals of the form $(1-D)^m(1+p(D)), m \geq 1$, the k-constraint at the input of a $1/(1 \oplus D)$ precoder is sufficient to eliminate quasi-catastrophic error propagation, because it limits the maximum length of channel-input error patterns of type $(+2)$, (-2) to $k+1x$. Conversely, for shaping polynominals that also exhibit spectral nulls at the Nyquist frequency, such as for example the shaping polynominal $(1-D^2)(1+p(D))$, which is very often used in practical systems, more channel-input error patterns need to be eliminated. Specifically, it can readily be seen that in this case it is necessary, and sufficient, to limit the maximum length of channel-input error patterns of the type $(+2)$, (-2), $(+2\,-2)$, $(-2\,+2)$, $(+2\,0)$, $(0\,+2)$, $(-2\,0)$, $(0\,-2)$. In general, these channel-input patterns exhaustively characterize the undesired quasi-catastrophic sequences for arbitrary shaping polynomials of the form $(1-D)^m(1+D)^n(1+p(D))$, where $n, m \geq 1$ (see also [44]). The I-constraint discussed above at the input of a $1/(1 \oplus D^2)$ precoder limits the maximum length of channel-input error patterns of type $(+2)$, (-2), $(+2\,-2)$, $(-2\,+2)$ to $2I+2$, and of type $(+2\,0)$, $(0\,+2)$, $(-2\,0)$, $(0\,-2)$ to $2I+3$. Note that an additional G-constraint, $G \leq 2I$, further reduces the maximum length of error patterns of type $(+2)$, (-2), $(+2\,-2)$, $(-2\,+2)$ [22].

The most widely used code rates for PRML(G, I) codes in the industry have been 8/9 and 16/17. The optimal block lists of length 9 for constructing block-encodable/decodable rate-8/9 PRML$(4, 4)$ and rate-8/9 PRML$(3, 6)$ codes can be found in [45]. By slightly relaxing the G and I parameters, it was also possible to construct a rate-16/17 PRML$(6, 6)$ code [46].

3.3. MTR Codes

The MTR(j, k) codes introduced in [21] result in a direct coding gain or improved performance by eliminating some of the predominant error events in conjunction with sequence detection. In addition to the benefits of an enhanced coding gain, MTR codes are useful in controlling nonlinear phenomena associated with the fast switching of the write head. More specifically, the MTR constraints are characterized by the parameters j, the maximum number of consecutive 1s that can occur, and k, the maximum number of zeros that can occur. Figure 10 shows the FSTDs generating sequences according to the MTR$(j = 2, k = \infty)$ and MTR$(j = 3, k = \infty)$ constraints. The FSTD of the MTR$(j = 2, k = 5)$ constraint is illustrated in Fig. 11. It can readily be verified that in all sequences obtained by traversing the paths of the diagram in Fig. 11, the consecutive runs of 1s are limited to two. Also, the runs of consecutive 0s are limited to five for timing-recovery purposes. The capacities of these FSTDs are Cap(MTR$(j = 2, k = 5)) = 0.8578$, Cap(MTR$(j = 2, k = \infty)) = 0.8791$, and Cap(MTR$(j = 3, k = \infty)) = 0.9468$.

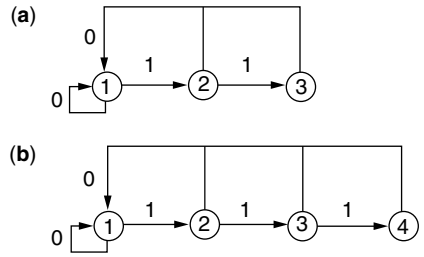


Figure 10. FSTDs for (**a**) MTR($j = 2, k = \infty$) and (**b**) MTR($j = 3$, $k = \infty$) constrained sequences.

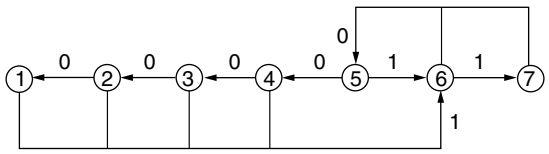


Figure 11. FSTD for MTR($j = 2, k = 5$) constrained sequences.

In general, MTR(j, k) codes do not avoid quasi-catastrophic error propagation in sequence detectors for PR polynomials with spectral nulls at both dc and the Nyquist frequency. The k-constraint avoids channel-input error sequences that have spectral energy only at dc, whereas the j-constraint avoids channel-input error sequences that have spectral energy only at the Nyquist frequency. Therefore, an additional constraint is needed to limit the maximum length of channel-input error sequences of type $(+2\,0)$, $(0\,+2)$, $(-2\,0)$, $(0\,-2)$ that have spectral energy at both dc and the Nyquist frequency. This new constraint for MTR codes, known as the *twins* or t-constraint, has been introduced in [22]. In simple terms the MTR encoder satisfies a t-constraint if it does not allow $t + 1$ consecutive pairs of 0s or 1s (twins). For example if $t = 8$ then nine consecutive twins are not allowed, that is, the string 00 00 11 00 00 00 11 00 11 is not allowed. Sequences that satisfy the t-constraint and at the same time are j- and k-constrained are denoted by MTR(j, k, t) [22]. The derivation of the deterministic FSTD that describes the MTR(j, k, t) constraint is presented in [22]. Figure 12 illustrates a 30-state labeled directed graph for the MTR($j = 2, k = 7, t = 4$) constraint. The capacity of this FSTD is Cap(MTR($j = 2, k = 7, t = 4$) = 0.8591. Tables 5 and 6 list the capacities Cap(MTR($j = 2, k, t$)) and Cap(MTR($j = 3, k, t$)) for various values of the parameters k and t. Codes based on the MTR(j, k, t) constraints eliminate the problem of error propagation in sequence detection for all PR shaping polynomials of the form $(1 - D)^m(1 + D)^n(1 + p(D))$, $m, n \geq 1$. Note that this class of PR polynomials includes PR4, EPR4 (extended PR4), and E^{2}PR4, (extended-square PR4) corresponding to $(1 - D^2)$, $(1 - D^2)(1 + D)$, and $(1 - D^2)(1 + D)^2$ shaping polynomials, respectively, which have been very important in practical systems. For shaping polynomials with memory greater than $j + 1$, the j-constraint can readily be incorporated into the detector to reduce the number of states or branches in the trellis, and to increase the capacity by allowing new potential sequences which were forbidden in MTR(j, k, t) [22].

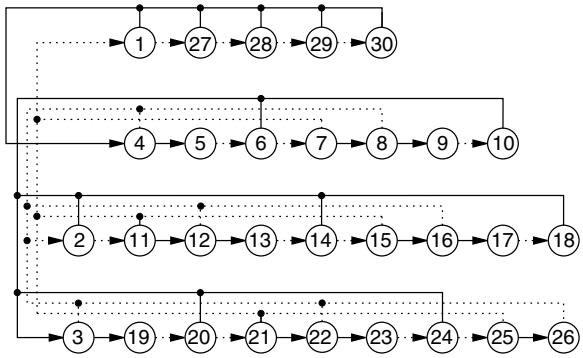


Figure 12. FSTD for MTR($j = 2, k = 7, t = 4$) constrained sequences. Solid line: 1, dotted line: 0.

Table 5. Capacity of MTR($j = 2, k, t$) Constraints

t	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
4	0.85514	0.85915	0.86052	0.86127	
5	0.86259	0.86732	0.86920	0.87026	0.87063
6	0.86567	0.87069	0.87296	0.87424	0.87476
7	0.86699	0.87213	0.87461	0.87599	0.87663
8	0.86756	0.87275	0.87536	0.87678	0.87748
9	0.86782	0.87302	0.87569	0.87714	0.87788
10	0.86792	0.87313	0.87584	0.87730	0.87806
11	0.86797	0.87319	0.87591	0.87737	0.87814
12	0.86799	0.87321	0.87594	0.87740	0.87818

Table 6. Capacity of MTR($j = 3, k, t$) Constraints

t	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
4	0.92830	0.93151	0.93223	0.93260	
5	0.93476	0.93834	0.93969	0.94047	0.94065
6	0.93720	0.94100	0.94265	0.94355	0.94390
7	0.93816	0.94203	0.94385	0.94481	0.94524
8	0.93853	0.94243	0.94434	0.94533	0.94581
9	0.93868	0.94259	0.94454	0.94554	0.94604
10	0.93874	0.94266	0.94462	0.94563	0.94614
11	0.93877	0.94268	0.94466	0.94567	0.94618
12	0.93878	0.94269	0.94467	0.94568	0.94620

In a NRZI format the j-constraint imposes a limit on the maximum number of consecutive transitions in the write current. In particular, the original MTR$(j = 2, k)$ codes that do not allow three consecutive transitions to appear in any encoded sequence have the interesting property of eliminating bit patterns that cause the most common error events in sequence detection. Figure 13 shows typical error patterns that are eliminated by these codes. These error patterns correspond to the NRZ error events of the form $\{\pm 2, \mp 2, \pm 2\}$. It can easily be seen that error bursts of the form $\{\pm 2, \mp 2, \pm 2, \ldots\}$ and length greater than three are also eliminated by the $j = 2$ constraint. These error events correspond to mistaking the polarity of an alternating write current for three or more channel symbol intervals. However, the distance gain realized by eliminating these error events is offset by a significant rate loss penalty. In particular, the maximum possible code rates for the original MTR$(j = 2, k)$ constraint [21]

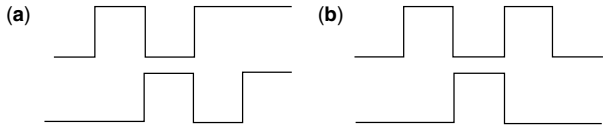


Figure 13. Error patterns eliminated by the $j = 2$ constraint. (**a**) Tribit shift, (**b**) quadbit to dibit.

and the MTR$(j = 2, k, t)$ constraint [22] are less than 8/9, leading to an unacceptable loss of performance due to the low code rate. Time-varying constraints, in which $j = 2$ is observed only at even bit positions (referred to as MTR $j = 2, 3$ constraints), permit the design of higher-rate MTR codes while maintaining their distance gain. The rates of the codes that have been designed to satisfy time-varying MTR$(j = 2, 3, k)$ constraints are 8/9 [47], and slightly higher [48,49]. These rates are still not adequately high for magnetic-recording applications.

An alternative coding strategy is to allow the error event $\{\pm 2, \mp 2, \pm 2\}$ to occur partially and to eliminate all other error events that correspond to mistaking the polarity of an alternating write current for four or more channel symbol intervals. In this way higher-rate codes can be constructed, thereby reducing the signal-to-noise ratio (SNR) penalty due to rate loss. For example, by using a $j = 3$ constraint in conjunction with a j-constrained trellis, all error events of the type $\{\pm 2, \mp 2, \pm 2, \mp 2, \}$, $\{\pm 2, \mp 2, \pm 2, \mp 2, \pm 2\}$, and so on are eliminated. Moreover, also the error event $\{\pm 2, \mp 2, \pm 2\}$ is partially eliminated because the quadbit-to-dibit error cannot occur (see Fig. 13). The rate-16/17 MTR$(j = 3, k = 12, t = 16)$ look-ahead code and the time-varying MTR$(j = 3, 4, k = 18, t = 14)$ block code, described in [22], have this interesting performance-enhancing property and have been implemented in various magnetic-recording systems. The construction of the rate-16/17 block-encodable/decodable MTR$(j = 3, 4, k = 18, t = 14)$ code, as described in [22], will be outlined briefly because of its practical significance. It can be verified that there are in total $65753 > 2^{16}$ potential codewords that can be generated by starting in state two in Fig. 10b, making 17 transitions and terminating in states one, two, or three. Among these codewords there are 199 codewords that begin or end with 10 zeros. After discarding these codewords and 17 more codewords that start with the first 15 bits of one of the strings (1001), (0110), (0011), (1100), or end with the first 16 bits of one of the strings (1001), (0110), (0011), or (1100), a set of 65537 codewords is obtained. These 17-bit codewords can be freely concatenated without violating the $j = 3$ constraint except at the border of two codewords, where the constraint is relaxed to $j = 4$. Furthermore, $k = 18$ and $t = 14$ is obtained.

In general, the performance-enhancing features of the class of high-rate MTR codes render them very attractive compared with conventional PRML(G, I) codes. Of course, ultimately, the coding gain of an MTR code is determined by the tradeoff among various factors such as rate loss, error propagation at the MTR decoder, and performance-improving properties of the MTR code. Finally, it is worthwhile mentioning that for a long time the conventional (G, I) and MTR constraints were discussed separately in the literature, and no connection was made between them. Very recently a theoretical result was presented in [22] showing that the precoded (G, I) constraints are a subclass of the precoded MTR constraints. This very interesting property allows an alternative code-construction methodology for (G, I) codes that is based on employing the $1/(1 \oplus D)$ precoder used by MTR codes.

3.4. Combined Modulation/Parity Codes

As discussed above, PR shaping for the magnetic-recording channel was normally used in conjunction with byte-oriented PRML(G, I) or MTR(j, k, t) modulation codes to aid timing recovery and gain control, to limit the path memory length, and to enhance the performance of the sequence detector. Until recently, these codes have been widely used in the disk-drive industry since the introduction of PR4 and maximum likelihood sequence detection more than ten years ago [10].

Because of the channel memory and the slight noise coloration, sequence detection produces some error patterns more frequently than others. These predominant error patterns or error events at the sequence detector output depend on the generalized PR shaping polynomial, the noise blend, and the normalized linear recording density D_c. For example, Table 7 shows the error events at the 16-state NPML detector output in the case of electronics noise only and $D_c = 3.4$. A shorthand notation to represent ternary error events has been used. For example, $\{+, -, +\}$ is used to denote the events $\{+2, -2, +2\}$ and $\{-2, +2, -2\}$. At high-recording linear densities, the predominant error event is of the type $\{+, -, +\}$. Simulation and experimental data have shown that at low and moderate linear densities the error event $\{+\}$ predominates. Note that the relative percentage of the various error events also depends on the noise conditions, that is, the noise blend between electronics, transition, and colored stationary media noise. These observations suggested the use of modulation codes combined with multiparity block codes to improve performance even further, but at the expense of a slight decrease in code rate. The basic concept is to use parity to detect the presence of an error event from a list of predominant error events. Decoding is achieved by a technique that combines syndrome and soft-decision decoding and is known in the industry as parity-based post-processing [23–28]. A combined modulation/parity code is constructed from a

Table 7. Error Events at NPML Detector Output (PW50/T = 3.4)

Error Events	Relative Frequency
$+-+$	62%
$+$	15%
$+-$	10%
$+-+-+$	4%
$+-+-$	2%
$+\,0\,0+$	1%
$+-+-+-$	1%
other	5%

(G, I) or MTR code by adding parity bits to the (G, I) or MTR codewords. Specifically, the design aims at single or double error-event detection from a prespecified list of error events so that the probability of miscorrection is minimum. Practical reasons, such as complexity and decoding delay, dictate the use of a short list of error events and short codes. Therefore, to keep the code-rate as high as possible, only a small number of parity bits may be used. Typically, the (G, I) codewords are extended by one to four parity bits. In [27] an elaborate approach to designing error-event detection codes is proposed. The code construction methodology is based on a recursive approach for building the parity-check matrix by adding one column at a time. A new column is added to the parity-check matrix if and only if the prespecified error-event detection capability is satisfied [27].

The approach adopted in [23,25,26,28] is based on simple polynomial codes characterized by their generator polynomial $g_c(D)$, whose degree specifies the total number of parity bits. The single-parity code $g_c(D) = 1 \oplus D$ ensures that there is an even number of 1s in the codeword. It can therefore detect error events of odd length. The double-parity code with $g_c(D) = 1 \oplus D^2$ adds two parity bits such that both odd and even interleaves of the codeword have an even number of 1s. The polynomial codes with primitive generator polynomials $g_c(D) = 1 \oplus D \oplus D^3$ and $g_c(D) = 1 \oplus D \oplus D^4$ add three and four parity bits, respectively. In all cases the combined modulation/parity code satisfies the prespecified (G, I) constraint. Examples of combined modulation/parity codes proposed in the literature include rate-32/35 PRML$(G = 7, I = 7)$ single-parity bit code [23], and rate-64/66, -64/67, -64/68, and -64/69 PRML(G, I) single-parity, dual-parity, triple-parity, and quadruple-parity codes, respectively, [25,26,28]. Moreover, longer block-length rate-96/101 and -96/102 PRML(G, I) triple-parity and quadruple-parity codes, respectively, have been discussed in [24–26].

An alternative design approach is to use performance-enhancing MTR modulation codes instead of the simple (G, I) codes. For example, the time-varying MTR$(j = 3, 4, k = 18, t = 14)$ block code presented above eliminates all error events of the type $\{\pm 2, \mp 2, \pm 2, \mp 2\}$, $\{\pm 2, \mp 2, \pm 2, \mp 2, \pm 2\}$, and so on. More important, this code also eliminates almost 50% of the error events of the type $\{\pm 2, \mp 2, \pm 2\}$. Therefore, two parity bits are adequate to detect the remaining predominant error events $\{\pm 2, \mp 2, \pm 2\}$, $\{\pm 2\}$, $\{\pm 2, \mp 2\}$, and $\{\pm 2, 0, 0, \pm 2\}$. By concatenating six 17-bit MTR$(j = 3, 4, k = 18, t = 14)$ codewords and inserting two parity bits in appropriate locations a rate-96/104 MTR$(j = 3, 4, k = 18, t = 14)$ dual-parity code is obtained.

4. EQUALIZATION AND DETECTION TECHNIQUES

The role of advanced signal processing has been crucial in increasing the areal density in magnetic storage devices. Until the beginning of the past decade, all commercially available disk-drive systems used analog peak detection. A peak detector operates on the analog readback signal to detect the presence of a pulse within a sliding observation window. At low linear densities, the location of peaks in the readback waveform will closely correspond to the location of the transitions in the input current waveform. Therefore, with an accurate clock signal, the recorded data pattern can in principle be reconstructed by correctly determining the pulse positions in the readback waveform. For a long time, peak detectors combined with RLL(d, k) modulation codes, which improve missing-bit errors by spacing transitions further apart, provided a low-cost detection strategy.

Recent advances in VLSI technologies and the need for ever higher areal densities paved the way for advanced signal-processing techniques, including maximum-likelihood sequence detection (MLSD) or near-MLSD schemes. The noise in hard-disk drive systems is a combination of electronics noise, colored stationary media noise, and transition noise. Although the first two noise sources can be treated as additive and Gaussian sources, the third is data-dependent, non-Gaussian and nonadditive.

For the purpose of introducing the optimum detection scheme for the readback signal, we will first consider the linear ISI model for the magnetic-recording system shown in Fig. 5. In this case, it can readily be shown [50,51] that the optimum signal processing and detection method consists of a filter matched to the pulse response, $h(t)$, a symbol-rate sampler, a whitening filter, and a MLSD as shown in Fig. 14. The output of the symbol-rate sampler will then be

$$u_i = \sum_{\ell} a_\ell R_h(i - \ell) + \nu_i \tag{13}$$

where $R_h(i)$ is the sampled autocorrelation function of $h(t)$ defined by

$$R_h(i) = \int_{-\infty}^{+\infty} h(t)h(t + iT)\,dt \tag{14}$$

$1/T$ is the rate at which the encoded date are written in the magnetic medium, and $\{\nu_i\}$ is the additive noise sequence given by

$$\nu_i = \int_{-\infty}^{+\infty} \eta(t)h(t - iT)\,dt \tag{15}$$

In Fig. 14, the binary modulation-encoded data sequence $\{a_i\}$ is represented by the D-transform $a(D) = \sum_i a_i D^i$ and

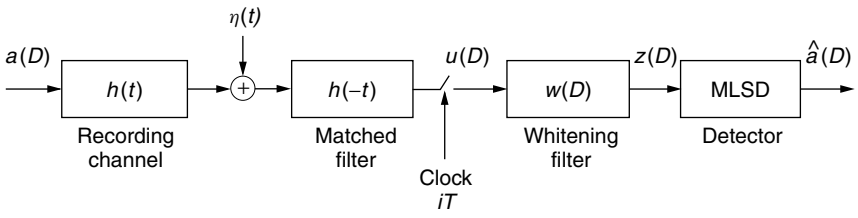


Figure 14. Magnetic-recording system with optimum signal processing and MLSD detection.

has power spectral density $S_a(D)$. Also the D-transform of the symbol-rate sampler output sequence $\{u_i\}$ is denoted by

$$u(D) = a(D)S_h(D) + \nu(D) \tag{16}$$

where $S_h(D)$ indicates the D-transform of the sampled autocorrelation function $R_h(i)$. By observing Eq. (15) it can readily be seen that the power spectral density of the noise sequence $\{\nu_i\}$ is $S_\nu(D) = S_\eta(D)S_h(D)$, where $S_\eta(D)$ denotes the D-transform of the sampled autocorrelation function $R_\eta(i)$ of the additive Gaussian process $\eta(t)$, that is, $R_\eta(i) = \int_{-\infty}^{+\infty} \eta(t)\eta(t+iT)\,dt$. Note that if the only noise source is the electronics noise due to the read head and preamplifier, then $S_\nu(D) = N_0\, S_h(D)$.

After symbol-rate sampling, the sequence at the output of the matched filter enters the whitening filter with transfer function $w(D)$. The output of the whitening filter is given by

$$z(D) = a(D)S_h(D)w(D) + \nu(D)w(D) = a(D)g(D) + e(D) \tag{17}$$

Observe that the desired signal component is affected by ISI through the overall transfer function $g(D) = S_h(D)w(D)$. Although it appears that the ISI may affect an infinite number of previously recorded symbols, in practice only a finite number of terms of $g(D)$ play a role in the sequence-detection process. This allows modeling the digital magnetic-recording system by an equivalent discrete-time finite-impulse-response (FIR) model with a minimum-phase transfer function $g(D)$ [52], where

$$g(D) = \sum_{\ell=0}^{M} g_\ell D^\ell \tag{18}$$

The MLSD that follows the whitening filter determines the most likely recorded sequence $\hat{a}(D)$ based on the observed sequence $z(D)$. It is well known that the MLSD is efficiently implemented with the Viterbi algorithm, whose complexity increases exponentially with the memory length. For an ISI span of M symbols, the state complexity of the optimum MLSD is 2^M. At low-linear recording densities, M may extend to only a few symbols. However, for $D_c > 3$, the ISI may extend to eight or even more symbols.

The optimum detector in the presence of ISI and additive Gaussian noise could therefore be prohibitively complex to implement because of its excessive state complexity. To circumvent this problem, a variety of suboptimal lower-complexity schemes have been developed aiming at reducing the span of ISI and consequently the complexity of the resulting MLSD. PR shaping to prescribed ISI targets has been very early recognized as an affective way to reduce the channel memory and also match the overall channel frequency response. In the following sections the most important suboptimum detection schemes will be reviewed, with emphasis on the ones that have extensively been used in high-performance direct-access storage devices.

4.1. PRML Detection

Partial response techniques for the magnetic-recording channel are similar to those that have been used in digital communication systems [50]. According to these techniques, the overall channel is shaped to some prescribed ISI pattern for which simple detection methods are known. The similarity between the pulse response in a magnetic-recording system and certain PR shapes was first noted in [53]. In particular, it was observed that at recording densities corresponding to $D_c \approx 2$, the frequency-domain representation of the Lorentzian pulse response closely resembles that of a linear filter with impulse response given by

$$f(t) = \operatorname{sinc}\left(\frac{t}{T}\right) - \operatorname{sinc}\left(\frac{t-2T}{T}\right) \tag{19}$$

where $\operatorname{sinc}(t) = \sin(\pi t)/\pi t$. It can readily be seen that at all times that are multiples of T, the value of the function $f(t)$ is zero except at times $t = 0$ and $t = 2T$, where the function takes the values $+1$ and -1, respectively. The discrete-time representation of this recording channel model corresponds to an overall noiseless input–output relationship given by

$$x_i = a_i - a_{i-2} \tag{20}$$

In D-transform notation, the input–output relationship becomes

$$x(D) = a(D)f(D) = a(D)(1 - D^2) \tag{21}$$

where the overall transfer function $f(D) = 1 - D^2$ is known in the literature as partial-response class-4 or PR4 shape. For higher linear recording densities, PR polynomials of the form

$$f_N(D) = (1 - D)(1 + D)^N \qquad N \geq 1 \tag{22}$$

are more suitable because their spectral characteristics match the spectral characteristics of the magnetic-recording channel better [54]. Clearly, the frequency response corresponding to $f_N(D)$ exhibits a spectral null at zero frequency and an N-th order null at the Nyquist frequency. Moreover, the PR4 polynomial corresponds to $N = 1$, whereas the PR polynomials with $N > 1$ are referred to as *extended* PR4 polynomials and are denoted as $\mathrm{E^{N-1}PR4}$. The PR4 and EPR4 shaping polynomials have been used extensively in hard-disk drives.

Simple linear equalization techniques may be employed to yield a prescribed PR shape. After symbol-rate sampling at the output of the matched filter, the samples enter a PR equalizer with transfer function $c(D)$, as shown in Fig. 15. The coefficients of the PR equalizer $\{c_\ell\}$ are optimized such that the overall transfer function, including channel and matched filter, closely matches a desired PR polynomial, that is,

$$c(D) = \frac{f_N(D)}{S_h(D)} \tag{23}$$

This equalizer is called zero-forcing linear equalizer, and shapes the incoming sequence $\{u_\ell\}$ according to any of the PR polynomials given by Eq. (22). If the magnetic-recording channel with additive Gaussian noise

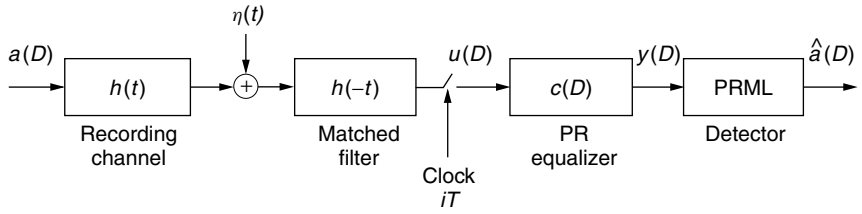


Figure 15. Magnetic-recording system with PRML detection.

is equalized to a PR target $f_N(D)$, the signal at the input of the detector becomes

$$\begin{aligned} y(D) &= a(D)S_h(D)c(D) + \nu(D)c(D) \\ &= a(D)f_N(D) + n(D) = x(D) + n(D) \end{aligned} \tag{24}$$

Thus, the zero-forcing linear equalizer limits the span of ISI to a small number of symbols, allowing an implementation of the MLSD for the desired PR target $f_N(D)$ with a small number of states.

After zero-forcing linear equalization, the noise sequence $n(D)$ is a discrete-time filtered version of the additive Gaussian noise $\eta(t)$. The power spectral density of $n(D)$ is

$$S_n(D) = \frac{f_N(D)f_N(D^{-1})}{S_h(D)} S_\eta(D) \tag{25}$$

Therefore, the variance of the noise sequence $\{n_i\}$ is

$$\sigma_n^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_n(e^{jw})\,dw = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|f_N(e^{jw})|^2}{S_h(e^{jw})} S_\eta(e^{jw})\,dw \tag{26}$$

This variance is a measure of the noise power at the input to the detector. Given that the noise sequence $n(D)$ is Gaussian and assuming that it also is an i.i.d. sequence of random variables (a reasonable assumption provided the selected PR target closely matches the magnetic-recording channel characteristics at the specified linear recording density), the most likely recorded sequence $\hat{a}(D)$ is the one that minimizes the squared Euclidean distance between the equalized sequence $y(D)$ and the noiseless sequence $x(D) = a(D)f_N(D)$. Equivalently,

$$\begin{aligned} \hat{a}(D) &= \arg\min_{a(D)} \|y(D) - x(D)\|^2 \\ &= \arg\min_{a(D)} \|y(D) - a(D)f_N(D)\|^2 \end{aligned} \tag{27}$$

The minimization in Eq. (27) can be efficiently achieved using the Viterbi algorithm. The combination of PR equalization techniques with MLSD is known in the industry as PRML detection [10]. The Viterbi algorithm is usually described via a finite-state transition diagram that evolves in time, known as the trellis diagram. The implementation complexity of the Viterbi algorithm depends on the number of states of the corresponding trellis. A $E^{N-1}PR4$ shaping polynomial gives rise to a trellis with 2^{N+1} states. In the case of PR4 and EPR4, the number of states is four and eight, respectively.

The application of the Viterbi algorithm for a PR4 shaped magnetic-recording channel was first proposed in [55], where it was also shown that a potential gain of 3 dB could be achieved using this dynamic programming procedure. The state complexity of a PR4-based detector can be further reduced by noting that the Euclidean metric in Eq. (27) can be split into two terms: one involving the odd indices and the other the even indices. Thus, in this special case, the Viterbi algorithm can be applied to two independent 2-state trellises operating on the even and odd PR4-equalized subsequences $\{y_{2i}\}$ and $\{y_{2i+1}\}$, respectively [10,56]. The Viterbi algorithm on each of the two 2-state trellises shown in Fig. 16 can be further simplified substantially by considering only the difference between the two survivor metrics and thus transforming the algorithm into a dynamic-threshold computation scheme ([10,56], and references therein). All these simplifications inherent in PR4 shaping led to the incorporation of the PR4-based PRML detection technology into magnetic hard-disk drives as well as magnetic tape systems. It was first introduced in the early 1990s in a commercial 5.25-inch IBM disk drive. The PR4-based PRML technology had a tremendous impact in the hard-disk drive industry and became the de-facto standard. Analytical studies, simulation results, and experimental data have shown that PR4-based PRML systems offer a 30–50% increase in linear recording density over RLL(2, 7) or RLL(1, 7) peak detection systems.

At higher linear recording densities, that is, $D_c > 2$, the linear PR4 equalizer leads to substantial noise enhancement, which increasingly affects the performance of the PRML system. EPR4 shaping alleviates this problem, to a certain extent, but increases the memory of the shaping polynomial to three and hence the number of states in the corresponding trellis to eight. Like PR4 shaping, the EPR4 polynomial exhibits spectral nulls at zero and the Nyquist frequencies; therefore, the magnetic-recording system would require similar PRML(G, I) codes

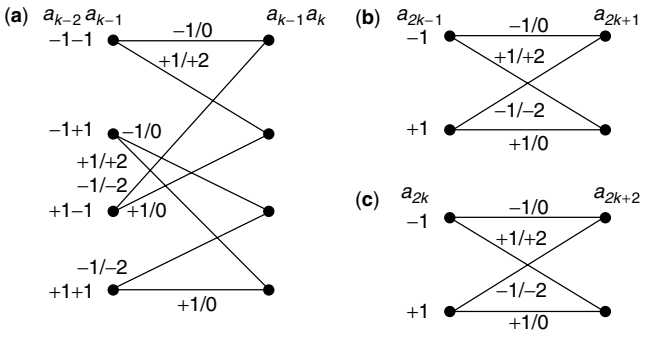


Figure 16. (**a**) 4-state PR4 trellis diagram. (**b**) 2-state odd and (**c**) 2-state even interleaved $1 - D$ trellis.

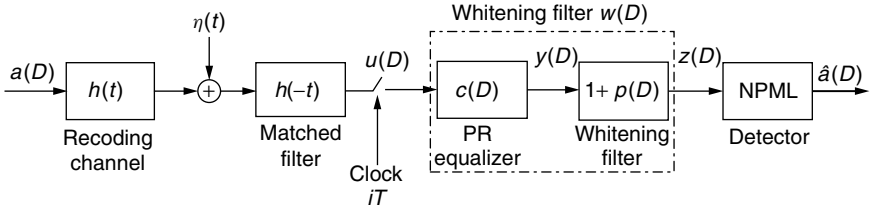


Figure 17. Magnetic-recording system with NPML detection.

to eliminate quasi-catastrophic error propagation, limit the path memory length, and facilitate timing and gain control. EPR4-based PRML systems achieve an additional 15% increase in linear recording density over PR4-based PRML systems, and the scheme has been used by some disk-drive manufactures for a limited period of time. Higher-order PR polynomials, that is, $N > 2$, although suitable for magnetic recording from the shaping characteristics point of view, do not achieve the matched filter bound and thus are of less interest. Hence, generalized PR polynomials in which the coefficients of the desired target response are nonintegers became significant in practical systems.

4.2. NPML Detection

In the absence of noise enhancement and noise correlation, the PRML sequence detector performs maximum-likelihood sequence detection. But there is an obvious loss of optimality associated with linear PR equalization as the operating point moves to higher linear recording densities. Clearly, a very close match between the desired target polynomial and the physical channel will guarantee that this loss will be minimal. An effective way to achieve near optimal performance independent of the operating point — in terms of linear recording density — and the noise conditions is via noise prediction. In particular, the power of the noise sequence $n(D)$ at the output of the PR linear equalizer can be minimized by using an infinitely long predictor. A linear predictor with coefficients $\{p_\ell\}$ operating on the noise sequence $n(D)$ will produce the estimate $\hat{n}(D)$, where the prediction-error sequence is given by

$$e(D) = n(D) + \hat{n}(D) = n(D)(1 + p(D)) \tag{28}$$

The optimum predictor $p(D) = p_1 D + p_2 D^2 + \cdots$, which minimizes the mean-square error $E\{|e_i|^2\}$ is given by $p(D) = q(D)/q_0 - 1$, where $q(D)$ is the minimum phase causal factor of $1/S_n(D)$ in Eq. (25). Using results from prediction theory, one readily obtains the minimum achievable mean-square error, which is also the variance of the white noise sequence at the output of the whitening filter $1 + p(D)$ shown in Fig. 17 and is given by [11]

$$\sigma_e^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \frac{S_\eta(e^{j\omega})}{S_h(e^{j\omega})}\, d\omega\right\} \tag{29}$$

Assuming that the additive noise process before matched-filtering is white and Gaussian with power spectral density $S_\eta(D) = N_0$, that is, the case of electronics noise only, Eq. (29) reduces to an expression that is identical to the minimum mean-square error (MMSE) of an infinitely long zero-forcing decision-feedback equalizer (DFE). Thus, the combination of the PR linear equalizer with the infinitely long noise predictor, shown in Fig. 17, is equivalent to the forward filter of a zero-forcing DFE, which in turn is the optimum whitening filter $w(D)$ of the MLSD [57]. Hence,

$$w(D) = c(D)(1 + p(D)) \tag{30}$$

and the sequence at the input of the detector becomes

$$z(D) = a(D)f(D)(1 + p(D)) + e(D) \tag{31}$$

where $e(D)$ is an AWGN sequence with variance given by Eq. (29). Note that in general the linear PR equalizer $\{c_\ell\}$ can be optimized so that the coefficients of the desired target $\{f_k\}$ can take any arbitrary value, that is,

$$f(D) = \sum_{\ell=0}^{K} f_\ell D^\ell \tag{32}$$

If $f_0 = 1$ and $f(D)$ is minimum phase, then the variance of the noise at the output of the whitening filter is still given by Eq. (29) [11].

So far, the zero-forcing criterion for obtaining the equalizer coefficients $\{c_\ell\}$ has been considered for the development of the NPML detection technique. Similar results can be obtained by using the MMSE criterion for optimizing the linear PR equalizer. For more details on this topic the reader is referred to [11].

The important result of Eq. (31) is that with this method the desired generalized PR target can be factored into two terms. The first factor contains spectral nulls or near nulls at selected frequencies, reflecting the nulls or near nulls of the physical magnetic-recording channel. The second factor, without roots on the unit circle, is optimized depending on the linear recording density and noise conditions. Note that spectral nulls at frequencies $f = 0$ and $f = 1/2T$ play an important role in practical systems because they render the sequence detector insensitive to dc offsets and disturbances around the Nyquist frequency.

An infinitely long predictor filter would lead to a sequence detector structure that requires an unbounded number of states. Finite-length predictors and, in particular, shaping polynomials of the form $g(D) = (1 - D)(1 + p(D))$ with a spectral null at dc and $g(D) = (1 - D^2)(1 + p(D))$ with spectral nulls at both dc and the Nyquist frequency, where $p(D) = \sum_{\ell=1}^{L} p_\ell D^\ell$ is the transfer function of a predictor of finite order L, render the noise at the input of the sequence detector approximately white. This class of generalized PR polynomials, which

is significant in practical applications, as well as any generalized PR shaping polynominal of the form $g(D) = f(D)(1 + p(D))$, when combined with sequence detection, give rise to NPML systems [11,12,14,15]. For a generalized PR target characterized by the polynomial $g(D) = (1 - D)(1 + p(D))$, the effective ISI memory of the system is limited to $M = L + 1$ symbols, and the NPML detector performs maximum-likelihood sequence detection using the 2^{L+1}-state trellis corresponding to $g(D)$. The same holds for generalized PR targets of the form $g(D) = (1 - D^2)(1 + p(D))$. In this case the effective ISI memory of the system is limited to $M = L + 2$ symbols, and the NPML detector performs maximum-likelihood sequence detection using the 2^{L+2}-state trellis corresponding to $g(D)$. Finally, if $g(D) = f(D)(1 + p(D))$, then the effective memory of the system is limited to $M = L + K$, giving rise to a 2^{L+K}-state NPML detector. In any case, the NPML detector is efficiently implemented by using the Viterbi algorithm, which recursively computes

$$\hat{a}(D) = \arg \min_{a(D)} \|z(D) - a(D)g(D)\|^2 \tag{33}$$

For large values of L, which implies white Gaussian noise at the output of the predictor filter, the performance of NPML may be determined by the technique described in [52]. At high SNR, the probability of error for NPML detection can be approximated by

$$P_b \approx K_0 Q\left(\frac{d_{\min}}{2\sigma_e}\right) \tag{34}$$

where $d^2_{\min}$ is the minimum Euclidean distance of an error event, σ_e^2 is the variance of the AWGN at the output of the predictor given in Eq. (29), and K_0 is a constant.

The variance σ_e^2 of the noise is a measure of performance because it represents the effective SNR at the input to the NPML detector. It is plotted in Fig. 18 as a function of the normalized linear density D_c for selected generalized PR shaping polynomials. The system is only affected by electronics noise, and the SNR is assumed constant at 25 dB. Curve 1 corresponds to the case of an infinitely long whitening filter $w(D)$. Curves 2 and 3 correspond to the memory-four shaping polynomials $g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2)$ and $g(D) = (1 - D)(1 + 0.75D)(1 + p_1 D + p_2 D^2)$, respectively. Finally, curve 4 corresponds to a memory-six shaping polynomial assuming a PR4 equalizer followed by a four-coefficient predictor. In all cases a simple 5-pole Butterworth filter instead of a matched filter has been assumed. Furthermore, the results for curves 2 and 4 have been obtained by using a linear PR4 equalizer with 10 coefficients, whereas those for curve 3 have been obtained by using a $(1 - D)(1 + 0.75D)$ 10-coefficient PR linear equalizer. As can be seen a memory 4-target giving rise to a 16-state NPML detector is uniformly good for $2 \le D_c \le 3.6$. In fact, the variance of the noise at the input of the 16-state NPML detector is at most 1.2 dB worse than the variance of an infinitely long whitening filter. The class of 16-state NPML detectors associated with order $L = 3$ predictor filters $(g(D) = (1 - D)(1 + p_1 D + p_2 D^2 + p_3 D^3))$ or order $L = 2$ predictor filters $(g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2))$ provide significant performance gains over the E^2PR4 shaping polynomial and is currently the state-of-the-art detection scheme in the hard-disk drive industry.

In practical applications a simple low-pass filter is used instead of a matched filter. Furthermore, the PR equalizer can be implemented as a FIR digital filter or a continuous-time filter before sampling. There are many commercial products that have used the latter approach, but today most of the hard-disk drives use discrete-time FIR equalization techniques. For known channel characteristics, the coefficients of the finite-length PR equalizer and predictor can be obtained by solving two sets of equations separately. In a first step the optimum coefficients of a finite-length PR equalizer according to the MMSE or zero-forcing criterion are obtained [50]. The predictor coefficients are then the solution of the well-known normal equations. An alternative approach is to combine the PR equalizer and the predictor into a single whitening filter and obtain its coefficients together with the NPML detector target by using a joint optimization procedure. Such an approach is reminiscent of the computation of the coefficients of a PR DFE [11,58]. To cope with the slow variations of the magnetic-recording channel as well as with the need for different sets of coefficients depending on whether the outer or the inner tracks are read back, standard adaptation algorithms can be applied. In some commercial hard-disk drives, the whitening-filter

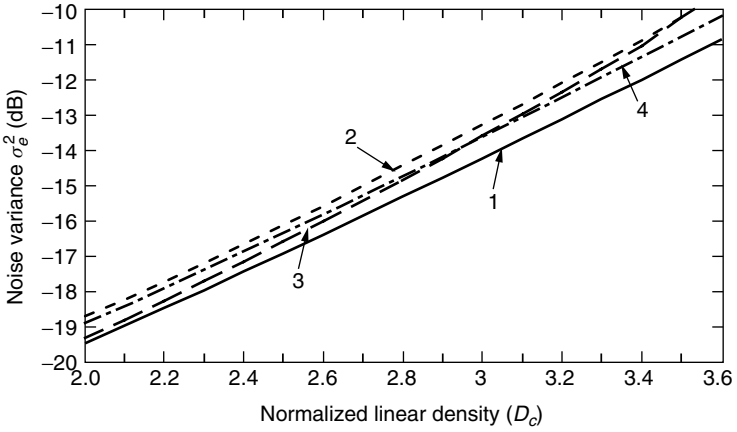


Figure 18. Noise variance at NPML detector input as a function of D_c for a Lorentzian channel with AWGN and an SNR of 25 dB. Curve 1: infinitely long whitening filter $w(D)$; curve 2: $g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2)$; curve 3: $g(D) = (1 - D)(1 + 0.75D)(1 + p_1 D + p_2 D^2)$, and curve 4: $g(D) = (1 - D^2)(1 + p_1 D + p_2 D^2 + p_3 D^3 + p_4 D^4)$.

(PR equalizer/predictor) coefficients are trained during manufacturing, and large sets of coefficients are obtained reflecting the various operating conditions. These sets of coefficients are stored and frozen before shipping the products. Other commercial products have the capability to retrain the whitening filter and NPML detector target adaptively in real time.

4.3. Other Detection Techniques

Several other techniques to increase the areal recording density have been proposed. One of these, which is well known in the field of digital communications, is DFE [50]. The DFE does not force the overall transfer function to a prescribed target, and thus suffers from less noise enhancement than PRML techniques. In fact, the response at the output of the forward section of a DFE could be viewed as a particular form of generalized PR polynomial that matches the magnetic-recording channel characteristics very well. It has already been pointed out above that the forward section of an infinitely long zero-forcing DFE is equivalent to the whitening filter of the optimum MLSD. The feedback section of the DFE using past decisions cancels the causal ISI and a simple threshold symbol-by-symbol detector provides the final decision. Application of DFE to storage channels has been studied in [59]. An adaptive DFE with look-up table implementation of the feedback section (RAM-DFE) has been proposed in [60,61]. By using a dynamic procedure to update the contents of the RAM, a RAM-DFE can also effectively counter nonlinear distortion. Prototype DFE chips for the magnetic-recording channel have been developed in the past [62] but this approach did not enjoy commercial success primarily because of the problem of error propagation.

Another technique that has attracted considerable attention is the fixed-delay tree-search approach with decision feedback (FDTS/DF) [31]. The FDTS/DF employs the forward section of a DFE to create a minimum phase causal response with most of the energy of the causal ISI concentrated in the first few terms. If M is the total span of ISI, the last $M-\tau$ ISI terms are canceled by a feedback mechanism similar to a DFE arrangement, whereas the first τ ISI terms are being processed by a detector structure that searches into a tree with $2^{\tau+1}$ branches. Clearly, the feedback cancellation scheme works properly provided that the FDTS/DF detector releases decisions with a delay of τ symbols. Thus, the detector looks ahead τ steps into the tree, and computes the $2^{\tau+1}$ Euclidean metrics associated with all the look-ahead paths in the tree. These metrics are then used to decide whether the symbol at the root of the tree should be $+1$ or -1. This scheme is a derivative of the delay-constrained optimal detector approach presented in [63]. The application of noise-predictive PR equalization schemes in conjunction with FDTS has been studied in [12].

Finally, reduced-state sequence-detection schemes [64–66] have also extensively been studied for application in the magnetic-recording channel ([12,25] and references therein). For example, it can readily be seen that the NPML detectors can be viewed as a family of reduce-state detectors with imbedded feedback. They also exist in a form in which the decision-feedback path can be realized by simple table look-up operations, whereby the contents of these tables can be updated as a function of the operating conditions [12]. Analytical and experimental studies have shown that a judicious tradeoff between performance and state complexity leads to practical schemes with considerable performance gains. Thus, reduced-state approaches appear promising for increasing the linear density even further.

5. PARITY-BASED POST-PROCESSING

In generalized PR systems combined with sequence detection a short list of error events dominates. In general, post-processors are suboptimum reduced-complexity receiver structures that improve error-rate performance by correcting the most likely error events at the output of the sequence detector. The use of a post-processor requires a modest increase in implementation complexity. Based on the short list of preselected error events the post-processor computes the log-likelihood ratio (LLR) of each of these selected error events at each point in time. These LLRs are then used to decide the type and the location of the most likely error event to be corrected. The LLR corresponding to the i-th error event in the list, $\varepsilon_i(D)$, is computed as the difference of the following Euclidean metrics

$$\begin{aligned}\mathrm{LRR}(\varepsilon_i(D)) = \|(z(D) - \hat{a}(D)g(D)\|^2 - \|z(D) \\ - (\hat{a}(D) + \varepsilon_i(D))g(D)\|^2 \end{aligned} \tag{35}$$

where $\|z(D) - \hat{a}(D)g(D)\|^2$ is the Euclidean distance between the noisy sequence at the input of the sequence detector, $z(D)$ and the sequence $\hat{a}(D)g(D)$ generated by using final decisions produced by the sequence detector, whereas $\|z(D) - (\hat{a}(D) + \varepsilon_i(D))g(D)\|^2$ is the Euclidean distance between the noisy sequence at the input of the sequence detector, $z(D)$ and the sequence $(\hat{a}(D) + \varepsilon_i(D))g(D)$ generated by an alternative data sequence that includes the specific error pattern.

As can be seen, the LLRs computed according to Eq. (35) use the same metric as the NPML detector does. A post-processor using such a soft metric is also referred to as noise-predictive post-processor. Noise-predictive post-processors are reminiscent of the *one-shot* optimum receiver structures in communications theory [67], and can have threshold-based or parity-based triggering mechanisms [23–28,68,69]. In general, threshold-based post-processing schemes do not suffer from rate loss, whereas parity-based post-processing schemes do. However the parity-based triggering mechanism for initiating error-event correction may be more robust in certain situations.

In a parity-based post-processing scheme typically a single or double error event within a codeword is corrected. After NPML detection and during the symbol-by-symbol processing of each codeword, a short list of the most likely error events is maintained and continuously updated together with the associated error type, polarity, and location. Once the entire codeword has been received and the short list finalized, the LLRs and the syndromes of all combinations of error events are computed. Finally,

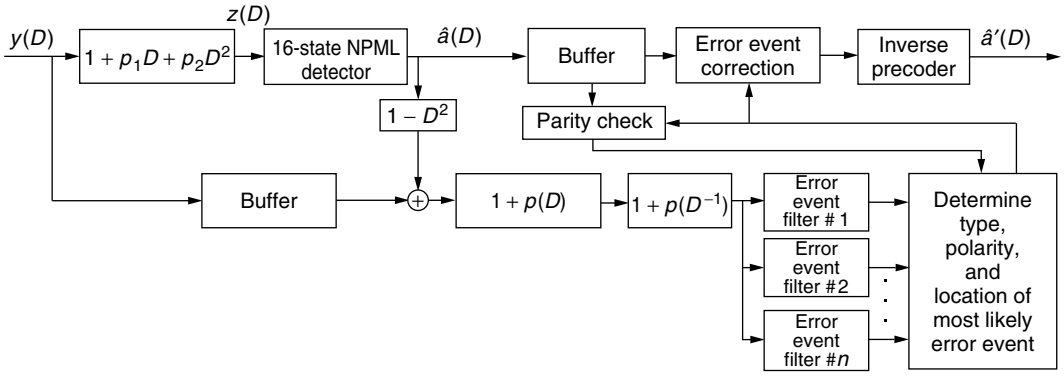


Figure 19. 16-state NPML detector and noise-predictive parity-based post-processor. From [26], © IEEE 2001.

after rejecting error-event candidates that violate certain conditions, such as for example those that do not produce valid syndromes, the single (double) error event with the minimum LLRs is (are) corrected. Figure 19 shows the block diagram of the 16-state NPML detector in tandem with a parity-based post-processor.

Figure 20 illustrates the performance of various modulation/parity codes after 16-state NPML detection and parity-based post-processing. These results have been obtained via computer simulations assuming the Lorentzian channel model and electronics noise. The user linear density is set to $\mathrm{PW50}/Tu = 3.2$. The channel density D_c depends on the rate of the code. The front-end filter is a five-pole low-pass Butterworth filter with 3-dB cutoff frequency at 55% of the channel symbol rate. The equalizer is a 10-coefficient PR4-shaping zero-forcing equalizer. Curve 1 shows the performance of a 16-state NPML detector in conjunction with a rate-32/34 PRML(G, I) single-parity code and a post-processor that detects and corrects the three types of error events $\{\pm 2\}$, $\{\pm 2, \mp 2, \pm 2\}$, and $\{\pm 2, \mp 2, \pm 2, \mp 2, \pm 2\}$. Curve 2 indicates the performance of the 16-state NPML detector in conjunction with a rate-96/102 PRML(G, I) code with quadruple-parity and a post-processor that detects and corrects seven types of error events. Finally, curve 3 corresponds to the performance of the 16-state NPML detector in conjunction with rate-96/104 MTR$(j = 3, 4, k = 18, t = 14)$ code with dual-parity and a post-processor that detects and corrects the four types of error events $\{\pm 2\}$, $\{\pm 2, \mp 2, \pm 2\}$, and $\{\pm 2, 0, 0, \pm 2\}$, and $\{\pm 2, \mp 2, \pm 2, \mp 2\}$. As can be seen, the rate-96/104 MTR-based code outperforms the rate-32/34 and -96/102 (G, I)-based codes by 1.3 and 0.7 dB, respectively, although it is approximately 2% less efficient. The difference in performance between the rate-96/104 MTR-based and rate-96/102 (G, I)-based codes is less pronounced under data-dependent medium noise conditions. This is attributed to the fact that in such a case, even at very high linear densities, the error $\{\pm 2\}$ is the predominant error event and the MTR constraints are not very effective.

Figure 21 shows the evolution of the various architectures that have been used over the past decade by various disk-drive manufactures. In particular, it shows the SNR requirements for achieving a symbol-error rate of 10^{-6} after the parity-based post-processor as a function of the normalized linear density D_c. Curve 1 corresponds to the conventional PRML architecture. It shows the performance of the 2-state interleaved PR4-based PRML

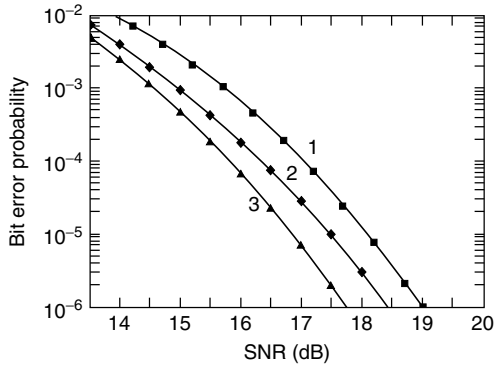


Figure 20. Bit-error rate performance after post-processing for a Lorentzian channel with $\mathrm{PW50}/Tu = 3.2$, AWGN, and a 16-state NPML detector. Curve 1: rate-32/34 PRML(G, I), 1 parity bit; curve 2: rate-96/102 PRML(G, I), 4 parity bits, and curve 3: rate-96/104 MTR(j, k, t), 2 parity bits.

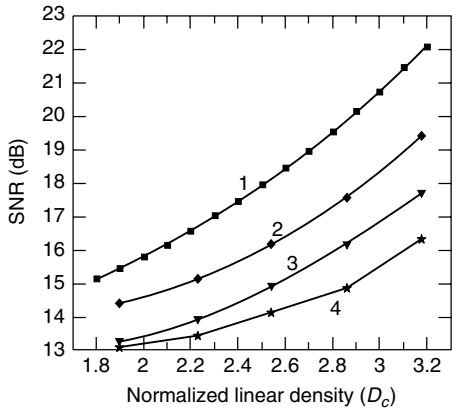


Figure 21. SNR requirements as a function of D_c for a Lorentzian channel with AWGN. Error probability $P_b = 10^{-6}$. Curve 1: PRML architecture, rate-16/17 PRML(G, I); curve 2: EPRML architecture, rate-32/34 PRML(G, I), 1 parity bit; curve 3: NPML architecture, rate-32/34 PRML(G, I), 1 parity bit, and curve 4: NPML architecture, rate-96/104 MTR(j, k, t), 2 parity bits.

detector in conjunction with a rate-16/17 PRML(G, I) constrained code. Curve 2 corresponds to the EPRML architecture. In this architecture the detector is based on the EPR4 shaping polynomial giving rise to an 8-state trellis. The inner code is a rate-32/34 PRML(G, I) constrained code and includes a single parity bit. Curve 2 demonstrates the benefit of the higher-order shaping polynomial as well as of the combined constrained/parity inner code, in particular at high linear recording densities, where a gain of approx. 2.5 dB over the conventional PRML scheme is obtained. Finally, curves 3 and 4 correspond to the NPML architecture, which has only recently been introduced in hard-disk drive products. The shaping polynomial is based on a 2-coefficient predictor and has spectral nulls at both the dc and the Nyquist frequency. Curve 3 shows the performance of the 16-state NPML detector in conjunction with the single-parity rate-32/34 PRML(G, I) constrained code, whereas curve 4 illustrates the performance of the 16-state NPML detector in conjunction with the dual-parity rate-96/104 MTR$(j = 3, 4, k = 18, t = 14)$ code. As can be seen, the 16-state NPML detector in conjunction with the rate-96/104 MTR-based dual-parity code, provides a 5.5 dB gain over the conventional PR4-based PRML detection with the rate-16/17 PRML(G, I) code at $D_c = 3.2$. Alternatively, the current NPML system architecture provides a 55% linear recording density increase over the conventional PR4-based PRML architecture.

6. DATA-DEPENDENT NPML DETECTION

Today's hard-disk drive devices employ thin-film media that appear primarily to exhibit nonstationary data-dependent transition or medium noise as opposed to colored stationary medium noise. Improvements on the quality of the readback head as well as the incorporation of low-noise preamplifiers may render the data-dependent medium noise a significant component of the total noise affecting the performance of the magnetic-recording system. Because the medium noise is correlated and data-dependent, information about the noise and data patterns in past samples can provide information about the noise in the current sample. Thus, the concept of noise prediction for stationary Gaussian noise sources developed in [11] can be naturally extended to the case where the noise characteristics depend highly on the local data patterns [16,18].

By modeling the data-dependent noise as a finite-order Markov process, the optimum MLSD for channels with ISI has been derived in [17,19]. In particular, it has been shown that when the data-dependent noise is conditionally Gauss–Markov, the branch metrics can be computed from the conditional second-order statistics of the noise process. In other words, the optimum MLSD can be implemented efficiently by using the Viterbi algorithm, where the branch-metric computation involves data-dependent noise prediction. Because both predictor coefficients and prediction error depend on the local data pattern, the resulting structure has been called data-dependent NPML detector [20]. In real systems, the data-dependent medium noise is not a strictly Markov noise process, and clearly the data-dependent NPML detector is only a near-MLSD structure. Nevertheless, physical models for data-dependent medium noise such as the one developed in [35] or the more accurate microtrack model [18] can be used to investigate the impact of the model parameters on the data-dependent Markov assumption and the performance of the NPML detector.

Let $\{y_i\}$ be the sequence of samples at the output of the PR linear equalizer with target $f(D)$. Then

$$y_i = x_i + n_i(a) = a_i + \sum_{\ell=1}^{K} f_\ell a_{i-\ell} + n_i(a) \tag{36}$$

where the noise sample $n_i(a)$ at time instance iT is assumed to be a zero-mean Gaussian random variable with statistics depending on the data sequence $a \triangleq \{a_i\}$. Figure 22 shows an example of a channel with ISI that spans K symbols and data-dependent Gauss–Markov noise. This model corresponds to Eq. (36), where the additive noise is generated by a L-order autoregressive filter whose coefficients depend on the last $K + 1$ recorded data symbols $a_i^{i-K} = \{a_i, a_{i-1}, \ldots, a_{i-K}\}$. In this case the correlated and data-dependent noise sample is given by

$$n_i = \sigma_e(a_i^{i-K})\upsilon_i - \sum_{\ell=1}^{L} p_\ell(a_i^{i-K}) n_{i-\ell} \tag{37}$$

where $\{\upsilon_i\}$ denotes a zero-mean unit-variance white Gaussian noise sequence, $\sigma_e(a_i^{i-K})$ indicates the data-dependent standard deviation, and $\{p_\ell(a_i^{i-K})\}$ represents the data-dependent coefficients of the autoregressive filter. Clearly, there are 2^{K+1} sets of filter coefficients, and the total memory of the model is $K + L$, giving rise to a trellis with 2^{K+L} states. A Viterbi algorithm, whose branch metrics have been modified to account for the data-dependent correlated noise, can operate on this trellis to estimate recursively the most likely recorded data sequence $\{a_i\}$. The implementation of this algorithm is very similar to the standard Viterbi algorithm for NPML detection described above. The main difference is that in the computation of the branch metrics, a window of observed values $y_i^{i-L} = \{y_i, y_{i-1}, \ldots, y_{i-L}\}$ is used, instead of just one sample z_i, which is the output of a "global"

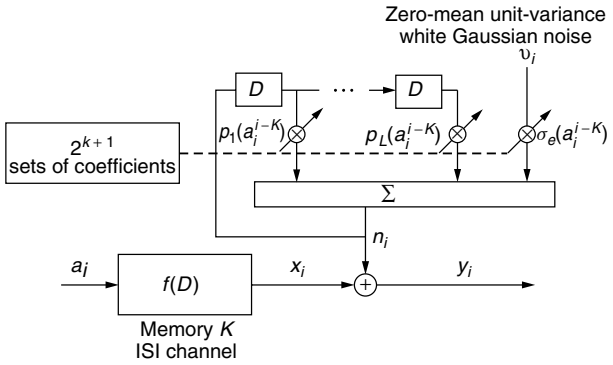


Figure 22. Block diagram of ISI channel with Gauss–Markov data-dependent noise.

whitening/prediction filter. The optimality of this approach for the Gauss–Markov noise case was shown in [19].

Let $\xi_i \triangleq a_i^{i-K-L} = \{a_i, a_{i-1}, \ldots, a_{i-K-L}\}$ denote a transition on the 2^{K+L}-state trellis. Then, assuming Gaussian data-dependent noise, the branch metric associated with transition ξ_i is given by

$$\gamma_i(\xi_i) = \ln \sigma_e^2(a_i^{i-K}) + \frac{[y_i - x_i + \hat{n}_i(\xi_i)]^2}{2\sigma_e^2(a_i^{i-K})} \tag{38}$$

where $\hat{n}_i(\xi_i)$ is a data-dependent predicted value of the current noise sample n_i, based on the past L noise samples $\{n_{i-1}, n_{i-2}, \ldots, n_{i-L}\}$, that is,

$$\hat{n}_i(\xi_i) = \sum_{\ell=1}^{L} p_\ell(a_i^{i-K})(y_{i-\ell} - x_{i-\ell}) = \sum_{\ell=1}^{L} p_\ell(a_i^{i-K}) n_{i-\ell} \tag{39}$$

and $\sigma_e^2(a_i^{i-K})$ represents the data-dependent variance of the prediction error. If the additive noise is correlated and Gaussian and does not depend on the data pattern, then $\{p_l(a_i^{i-K})\}$ and $\sigma_e^2(a_i^{i-K}) = \sigma_e^2$. In such a case, the data-dependent NPML detection structure reduces to the NPML detection technique presented above.

A finite-length predictor filter with coefficients that depend on a particular data pattern can be obtained by applying the normal equations separately and conditioned on each of the 2^{K+1} possible data patterns a_i^{i-K} that affect the noise process. For this purpose, the noise process can be characterized by 2^{K+1} autocorrelation functions of the form

$$R_n(l \mid a_i^{i-K}) = E\{n_i n_{i-\ell} \mid a_i^{i-K}\} \tag{40}$$

where the expectation is taken not only with respect to the noise statistics but also with respect to the data symbols that are not included in the conditioning. Using the autocorrelation functions described by Eq. (40), a set of 2^{K+1} L-th order prediction filters $\{p_l(a_i^{i-K})\}$ can be obtained together with their corresponding prediction errors $\sigma_e^2(a_i^{i-K})$, which can then be used in the branch metric computation of Eq. (38).

In practical systems the noise autocorrelation function can be estimated based on training patterns and statistical averaging, which can be performed during the manufacturing process of the disk drive. To avoid matrix inversions and the numerical problems associated with ill-conditioned matrices, an alternative approach is to use standard adaptation procedures to learn the set of noise predictor coefficients and noise prediction error-variances conditioned on the local data patterns [20]. Finally, note that the reduced-state sequence detection schemes discussed in connection with NPML detection can also be applied to data-dependent NPML, providing a significant reduction of implementation complexity.

Figure 23 illustrates the performance of two 16-state NPML detection schemes in the presence of nonstationary data-dependent medium noise. These results have been obtained via computer simulations assuming the microtrack channel model with a transition-width parameter of $a = 0.15$, and no electronics or any other source of stationary noise. In this case the system is affected by 100% data-dependent medium noise. The user linear recording density is set to PW50/$Tu = 3.6$. In both cases a rate-96/102 PRML(G, I) has been used, resulting in a normalized channel linear density of $D_c = 3.825$. The front-end filter is a five-pole low-pass Butterworth filter with 3-dB cutoff frequency at 55% of the channel symbol rate. The equalizer is a 10-coefficient PR4-shaping zero-forcing equalizer. Curve 1 shows the performance of a 16-state NPML detector whose branch metric computation uses the same set of predictor coefficients optimized as if the system were affected by stationary Gaussian noise only. Curve 2 indicates the performance of the data-dependent 16-state NPML detector whose branch metrics have been modified according to Eq. (38) to account for the presence of 100% data-dependent noise. As can be seen, the data-dependent NPML detector yields a gain of 1.3 dB at a symbol error rate of 10^{-5}. Note however that 100% data-dependent medium noise is not a realistic scenario in today's hard-disk drives, and therefore the gains expected by using a more complex data-dependent NPML detection scheme are less pronounced.

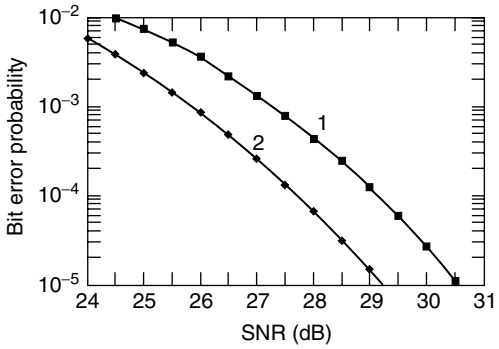


Figure 23. Bit-error rate performance for microtrack channel with PW50/$Tu = 3.6$, $a = 0.15$, and 100% data-dependent medium noise. Curve 1: 16-state NPML detector, and curve 2: 16-state data-dependent NPML detector.

7. FUTURE TRENDS

Currently, MTR or (G, I) codes combined with multiparity block codes, and 16-state NPML detection for generalized PR shaping channels, followed by a post-processor for soft decoding of the combined multiparity/constrained code, represent the state of the art in the industry. Outer error-correction coding has also played an important role in achieving high data integrity in magnetic-recording systems. In hard-disk drives, interleaved byte-oriented RS coding is currently the standard outer coding scheme. In the future, RS symbols with more than eight bits and sector sizes having a length of more than 512 8-bit bytes will have a significant impact on efforts to improve performance and push areal density even further.

The recent advances in coding theory and in particular the introduction of Turbo codes in the mid-1990s and the rediscovery of the powerful low-density parity check (LDPC), hold the promise to push the areal density to the ultimate limit for given magnetic-recording components. In spite of the current limit of the sector-size in a hard-disk drive to 512 bytes, which constrains the block size of an

outer code, and the high code-rate requirements, it has been shown that LDPC or Turbo codes with rather simple iterative decoding schemes can achieve a gain of more than 2 dB over existing systems and bring performance to within 1.5 dB of the ultimate information-theoretic limit: the capacity.

However, despite progress in the area of reduced-complexity detection and decoding algorithms, Turbo equalization structures with iterative detectors/decoders have not yet found their way into digital recording systems because of the still unfavorable tradeoff between performance, implementation complexity, and latency. The design of high-rate, short-block-length Turbo-like codes for recording systems remains an area of active research.

BIOGRAPHY

Evangelos S. Eleftheriou received as B.S. degree in electrical engineering from the University of Patras, Greece, in 1979, and M.Eng. and Ph.D. degrees in electrical engineering from Carleton University, Ottawa, Canada, in 1981 and 1985, respectively. He joined the IBM Zurich Research Laboratory in Rüschlikon, Switzerland, in 1986, where he has been working in the areas of high-speed voice-band data modems, wireless communications, and coding and signal processing for the magnetic recording channel. Since 1998, he has managed the magnetic recording and wired transmission activities at the IBM Zurich Research Laboratory.

His primary research interests lie in the areas of communications and information theory, particularly signal processing and coding for recording and transmission systems. He holds more than 30 patents (granted and pending applications) in the areas of coding and detection for transmission and digital recording systems. He was editor of the IEEE Transactions on Communications from 1994 to 1999 in the area of Equalization and Coding. He was guest editor of the *IEEE Journal on Selected Areas of Communications* special issue, "The Turbo Principle: From Theory to Practice."

BIBLIOGRAPHY

1. D. A. Thompson and J. S. Best, The future of magnetic data storage technology, *IBM J. Res. Develop.* **44**: 311–322 (2000).
2. S. Iwasaki and Y. Nakamura, An analysis for the magnetization mode for high density magnetic recording, *IEEE Trans. Magn.* **MAG-13**: 1272–1277 (1977).
3. S. Iwasaki and K. Ouchi, Co-Cr recording films with perpendicular magnetic anisotropy, *IEEE Trans. Magn.* **MAG-14**: 849–851 (1978).
4. H. Takano et al., Realization of 52.5 Gb/in.^{2} perpendicular recording, *J. Magn. Magn. Mater.* **235**: 241–244 (2001).
5. H. N. Bertram and M. Williams, SNR and density limit estimates: A comparison of longitudinal and perpendicular recording, *IEEE Trans. Magn.* **36**: 4–9 (2000).
6. R. Wood, The feasibility of magnetic recording at 1 terabit per square inch, *IEEE Trans. Magn.* **36**: 36–42 (2000).
7. R. Cideciyan, E. Eleftheriou, and T. Mittelholzer, Perpendicular and longitudinal recording: A signal-processing and coding perspective, *IEEE Trans. Magn.* **38**: 1698–1704 (2002).
8. A. Dholakia, E. Eleftheriou, and T. Mittelholzer, On iterative decoding for magnetic recording channels, in *Proc. 2nd Intl. Symp. on Turbo Codes & Related Topics*, Brest, France, 219–226, 2000.
9. D. Arnold and E. Eleftheriou, On the information-theoretic capacity of magnetic recording systems in the presence of medium noise, *IEEE Trans. Magn.* **38**: (Sept. 2002) (in press).
10. R. D. Cideciyan et al., A PRML system for digital magnetic recording, *IEEE J. Sel. Areas Commun.* **10**: 38–56 (1992).
11. P. R. Chevillat, E. Eleftheriou, and D. Maiwald, Noise predictive partial-response equalizers and applications, *Proc. IEEE Intl. Conf. Commun.* 942–947 (1992).
12. E. Eleftheriou and W. Hirt, Noise-predictive maximum-likelihood (NPML) detection for the magnetic recording channel, *Proc. IEEE Intl. Conf. Commun.* 556–560 (1996).
13. E. Eleftheriou and W. Hirt, Improving performance of PRML/EPRML through noise prediction, *IEEE Trans. Magn.* **32**(part1): 3968–3970 (1996).
14. R. Karabed and N. Nazari, Trellis-coded noise predictive Viterbi detection for magnetic recording channels, in *Dig. The Magnetic Recording Conf.* (*TMRC*), Minneapolis, MN, Aug. 1997.
15. J. D. Coker, E. Eleftheriou, R. L. Galbraith, and W. Hirt, Noise-predictive maximum likelihood (NPML) detection, *IEEE Trans. Magn.* **34**(part1): 110–117 (1998).
16. J. Caroselli, S. A. Altekar, P. McEwen, and J. K. Wolf, Improved detection for magnetic recording systems with media noise, *IEEE Trans. Magn.* **33**: 2779–2781 (1997).
17. A. Kavcic and J. M. F. Moura, Correlation-sensitive adaptive sequence detection, *IEEE Trans. Magn.* **34**: 763–771 (1998).
18. J. P. Caroselli, Modeling, analysis, and mitigation of medium noise in thin film magnetic recording channels, Ph.D. dissertation, University of California, San Diego, 1998.
19. A. Kavcic and J. M. F. Moura, The Viterbi algorithm and Markov noise memory, *IEEE Trans. Inform. Theory* **46**: 291–301 (2000).
20. J. Moon and J. Park, Pattern-dependent noise prediction in signal-dependent noise, *IEEE J. Sel. Areas Commun.* **19**: 730–743 (2001).
21. B. Brickner and J. Moon, Design of a rate 6/7 maximum transition run code, *IEEE Trans. Magn.* **33**(part1): 2749–2751 (1997).
22. R. D. Cideciyan, E. Eleftheriou, B. H. Marcus and D. S. Modha, Maximum transition run codes for generalized partial response channels, *IEEE J. Sel. Areas Commun.* **19**: 619–634 (2001).
23. T. Conway, A new target response with parity coding for high density magnetic recording channels, *IEEE Trans. Magn.* **34**: 2382–2486 (1998).
24. J. L. Sonntag and B. Vasic, Implementation and bench characterization of a read channel with parity check post-processor, in *Dig. The Magnetic Recording Conf.* (*TMRC*), Santa Clara, CA, Aug. 2000.
25. R. D. Cideciyan, J. D. Coker, E. Eleftheriou, and R. L. Galbraith, NPML detection combined with parity-based post-processing, in *Dig. The Magnetic Recording Conf.* (*TMRC 2000*), Santa Clara, CA, Aug. 2000.

26. R. D. Cideciyan, J. D. Coker, E. Eleftheriou, and R. L. Galbraith, NPML detection combined with parity-based postprocessing, *IEEE Trans. Magn.* **37**: 714–720 (2001).
27. B. Vasic, A graph based construction of high-rate soft decodable codes for partial response channels, in *Proc. IEEE ICC'01*, Helsinki, Finland, 2716–2720, 2001.
28. W. Feng, A. Vityaev, G. Burd, and N. Nazari, On the performance of parity codes in magnetic recording systems, in *Proc. IEEE Global Telecommun. Conf.*, 1877–1881, 2000.
29. R. I. Potter, Digital magnetic recording theory, *IEEE Trans. Magn.* **10**(part1): 502–508 (1974).
30. K. G. Ashar, *Magnetic Disk Technology*, IEEE Press, New York, 1997.
31. J. Moon and L. R. Carley, Performance comparison of detection methods in magnetic recording, *IEEE Trans. Magn.* **26**: 3155–3172 (1990).
32. J. Moon, The role of signal processing in data-storage systems, *IEEE Signal Proc. Mag.* 54–72 (July 1998).
33. L. L. Nunnelley, D. E. Heim, and T. C. Arnoldussen, Flux noise in particulate media: Measurement and interpretation, *IEEE Trans. Magn.* **23**: 1767–1775 (1987).
34. H. N. Bertram, *Theory of Magnetic Recording*, Cambridge University Press, Cambridge, UK, 1994.
35. S. K. Nair, H. Shafiee, and J. Moon, Modeling and simulation of advanced read channels, *IEEE Trans. Magn.* **29**: 4056–4058, (1993).
36. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423 (1948).
37. K. A. S. Immink, P. H. Siegel, and J. K. Wolf, Codes for digital recorders, *IEEE Trans. Inform. Theory* **44**: 2260–2299 (1998).
38. R. L. Adler, D. Coppersmith, and M. Hassner, Algorithms for sliding-block codes: An application of symbolic dynamics to information theory, *IEEE Trans. Inform. Theory* **29**: 5–22 (1983).
39. B. H. Marcus, P. H. Siegel, and J. K. Wolf, Finite-state modulation codes for data storage, *IEEE J. Sel. Areas Commun.* **10**: 5–37 (1992).
40. K. A. S. Immink, *Coding Techniques for Digital Recorders*, Prentice-Hall International, UK, 1991.
41. P. H. Siegel and J. Wolf, Modulation and coding for information storage, *IEEE Commun. Mag.* **29**: 68–86.
42. G. Jacoby, A new look-ahead code for increased data density, *IEEE Trans. Magn.* **13**: 1202–1204 (1977).
43. G. D. Forney and A. R. Calderbank, Coset codes for partial-response channels; or, Coset codes with spectral nulls, *IEEE Trans. Inform. Theory* **35**: 925–943 (1989).
44. R. Karabed, P. H. Siegel, and E. Soljanin, Constrained coding for binary channels with high intersymbol interference, *IEEE Trans. Inform. Theory* **45**: 1777–1797 (1999).
45. Method and Apparatus for Implementing Optimum PRML Codes, U. S. Patent 4, 707, 681 (1987) J. Eggenberger and A. M. Patel.
46. A. M. Patel, Rate 16/17 (0,6/6) Code, *IBM Tech. Discl. Bull.* **31**(8): 21–23 (1989).
47. W. G. Bliss, An 8/9 rate time-varying trellis code for high density magnetic recording, *IEEE Trans. Magn.* **33**: 2746–2748 (1997).
48. K. K. Fitzpatrick and C. S. Modlin, Time-varying MTR codes for high density magnetic recording, in *Proc. IEEE Global Telecommun. Conf.*, 1250–1253, 1997.
49. B. E. Moision and P. H. Siegel, Distance enhancing constraints for noise predictive maximum likelihood detectors, in *Proc. IEEE Global Telecommun. Conf.*, 2730–2735, 1998.
50. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
51. J. G. Proakis, Equalization techniques for high-density magnetic recording, *IEEE Signal Proc. Mag.* 73–82 (July 1998).
52. G. D. Forney, Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 363–378 (1972).
53. H. Kobayashi and D. T. Tang, Application of partial-response channel coding to magnetic recording systems, *IBM J. Res. Develop.* **14**: 368–375 (1970).
54. H. K. Thapar and A. M. Patel, A class of partial response systems for increasing storage density in magnetic recording, *IEEE Trans. Magn.* **MAG-23**: 3666–3668 (1987).
55. H. Kobayashi, Application of probabilistic decoding to digital magnetic recording, *IBM J. Res. Develop.* **15**: 65–74 (1971).
56. R. W. Wood and D. A. Peterson, Viterbi detection of class IV partial response on a magnetic recording channel, *IEEE Trans. Commun.* **COM-34**: 454–461 (1986).
57. R. Price, Nonlinearly feedback equalized PAM vs. capacity for noisy filter channels, in *Proc. IEEE Intl. Conf. on Commun.*, 22.12–22.17, 1972.
58. J. W. M. Bergmans, Partial response equalization, *Philips J. Res.* **42**: 209–245 (1987).
59. J. W. Bergmans, Density improvements in digital magnetic recording by decision feedback equalization, *IEEE Trans. Magn.* **22**: 157–162 (1986).
60. K. D. Fisher et al., An adaptive RAM-DFE for storage channels, *IEEE Trans. Commun.* **39**: 1559–1568, (1991).
61. J. Cioffi et al., Adaptive equalization in magnetic-disk storage channels, *IEEE Commun. Mag.* **28**: 14–29 (1990).
62. J. W. M. Bergmans et al. Dual-DFE read/write channel IC for hard-disk drives, *IEEE Trans. Magn.* **34**: 172–177 (1998).
63. K. Abend and B. D. Fritchman, Statistical detection for communication channels with intersymbol interference, *Proc. IEEE* **58**: 779–785 (1970).
64. V. M. Eyuboglu and S. U. Qureshi, Reduced-state sequence estimation with set partitioning and decision feedback, *IEEE Trans. Commun.* **COM-36**: 13–20 (1988).
65. A. Duell-Hallen and C. Heegard, Delayed decision-feedback sequence estimation, *IEEE Trans. Commun.* **COM-37**: 428–436 (1989).
66. P. R. Chevillat and E. Eleftheriou, Decoding of trellis-encoded signals in the presence of intersymbol interference and noise, *IEEE Trans. Commun.* **COM-37**: 669–676 (1989).
67. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons Inc., New York, 1965.
68. R. W. Wood, Turbo-PRML: A compromise EPRML detector, *IEEE Trans. Magn.* **29**: 4018–4020 (1993).
69. H. Sawaguchi and S. Mita, Soft-output decoding for concatenated error correction in high-order PRML channels, in *Proc. IEEE Intl. Conf. on Commun.*, 632–1637, 1992.

SIGNAL QUALITY MONITORING IN OPTICAL NETWORKS

IPPEI SHAKE
NTT Corporation
Kanagawa, Japan

1. INTRODUCTION

Signal quality monitoring is an important issue in relation to the design, operation, and maintenance of optical transport networks. From a network operator's point of view, monitoring techniques are required to provide connections, undertake protection and/or restoration, perform maintenance, and establish service level agreements. To realize these functions, the monitoring techniques should be able to offer the following: in-service (nonintrusive) measurement, signal deterioration detection [both signal-to-noise ratio (SNR) degradation and waveform distortion], fault isolation (locate faulty sections or nodes), transparency and scalability (irrespective of the signal bit rate and signal format), and simplicity (small size and low cost).

There are several approaches, including both digital and analog techniques, that make it possible to detect various types of impairment. Bit error rate (BER) measurement is a fundamental method for evaluating end-to-end signal quality, but it is not a practical solution for optical networks because it requires clock and data synchronization between transmitters and receivers. Error block detection or error counting utilizing a SONET/SDH frame or other frame is currently a practical solution for optical networks. However, these techniques fail to satisfy the aforementioned requirements for performance monitoring, for instance, transparency in an optical transport network (OTN). Several approaches have recently been proposed to overcome this problem. These techniques include optical SNR or power evaluation with optical/electrical spectrum measurement, pilot tone detection, pseudo-BER estimation or error monitoring using variable decision circuits, and a statistical method using histogram evaluation accompanied by synchronous or asynchronous eye diagram measurement. A fundamental performance monitoring parameter of any digital transmission system is its end-to-end BER. However, the BER can be correctly evaluated only with an out-of-service BER measurement by using a known test bit pattern in place of the real signal. By contrast, in-service measurement can provide only rough estimates through the measurement of digital parameters [e.g., BER estimation, error block detection, and error count in forward error correction (FEC) or analog parameters (e.g., optical SNR, optical/electrical spectrum, and Q-factor)].

2. PERFORMANCE MONITORING UTILIZING DIGITAL PARAMETERS

2.1. Digital Frame Based Technique

2.1.1. Bit Interleaved Parity (SONET/SDH). Error block detection is a digital technique for the end-to-end performance monitoring of optical channels. The error block is detected by means of an error detection code, for example, bit interleaved parity (BIP) in a SONET/SDH frame [1]. A BIP code that consists of X bits is called a BIP-X code. The BIP-X code is written as part of the overhead of the following payload. The data sequence to be monitored is divided into sequences, each of which is X bits long, the even parity of the nth bit (n is an integer between 1 and X) of all the X bits sequences is written in the nth bit of the BIP-X code. These procedures are undertaken at the transmitter end first. Then, at the receiver end, the BIP code is calculated again over each frame using the same procedure and compared with the transmitted code. This technique can only be used to monitor odd numbers of bit errors because it uses a BIP code in which an even parity is written over long sequences of data. This means that BIP-X is valid when the BER is low enough to cause only 1 bit error in the nth bit of all the X bit sequences.

2.1.2. Digital Wrapper. The digital wrapper is a SONET/SDH-based frame format technology in OTN [1]. This frame includes not only data and a channel header but also the forward error correction (FEC) code. The FEC makes it possible to correct bit errors that occur in the transmission line at the receiver by sending an additional bit sequence with data and channel header sequences. FEC is a powerful technology for long-haul high-speed optical transmission because of this real-time error correction capability. From the viewpoint of performance monitoring, the FEC procedure can provide us with the number of errors. This frame typically is used on a link per link basis, so it should not be used for end-to-end performance monitoring.

2.1.3. Others. Other frame format technologies have been developed for wide area networks (WAN), although most are also SONET-based. For example, 10 Gbit/s Ethernet (IEEE802.3ae) for WAN (WAN-PHY) has an operation, administration, management, and provisioning function based on SONET technology. Another extension of the 10 Gbit/s Ethernet frame for WAN has also been proposed.

2.2. Pseudo BER Estimation Using Variable Decision Circuit

When the system BER is too low to be measured within a reasonable amount of time, it is necessary to estimate the BER or other parameters representing signal quality. This subsection introduces one BER estimation approach that uses a variable decision circuit [2]. The decision level is changed step by step from a low amplitude to a high amplitude through the whole eye opening. The BER is measured only when the decision levels are relatively low or high. This is because the decision level is directly related to the BER, as shown in the following equation (with a Gaussian assumption),

$$\mathrm{BER} = (1/2)\mathrm{erfc}(Q/\sqrt{2}),\ Q = (\mu_1 - V_t)/\sigma_1 = (V_t - \mu_0)/\sigma_0 \quad (1)$$

where μ_i and σ_i are the mean and the standard deviation of mark ($i = 1$) and space ($i = 0$) noise levels, respectively, and V_t is the decision level. The BER becomes relatively high when the decision level is lower or higher than the

optimum level, and the measurement can be finished in a sufficiently short time. BER data are then plotted on a graph, where the vertical axis is the BER and the horizontal axis is the decision level. There are two linear fitting curves in the figure. Finally, the lowest BER is estimated by extracting the intersection of these two fitting curves.

This method is useful because a very low BER can be easily estimated within a reasonable amount of time. However, its use of BER measurement means that it requires data synchronization between the transmitter and receiver.

Recently, a new approach has been proposed as an application of this method, which resolves the problem of data synchronization by using two decision circuits [3,4]. The receiver equipment uses these two decision circuits. The first is used for a working channel with a fixed decision level and the second is used for measurement with a variable decision level. The received signal is divided into two, launched into these two decision circuits, and then the two logical outputs from the two circuits are combined with an exclusive OR gate (EXOR). The EXOR output is interpreted as the BER (Fig. 1). This provides a pseudo-BER estimation because the result from the fixed decision level is used as the reference data instead of the bit sequence of a real data code. Using these estimated BER values, we can estimate the lowest BER value by employing the same procedure as [2] (extracting the intersection of the fitting curve). The merit of this method is that it requires no knowledge of the transmitted bit sequences, and the performance of the data regeneration process in the master decision circuit is not degraded by the monitoring function.

By counting the bit error number, we also can estimate the amplitude histograms [3]. When the decision level of the second decision circuit is changed step by step through the whole eye opening, the number of "1" events counted at EXOR is recognized as the pseudo-error probability, which is the number of events whose amplitude is between the decision levels of the first and second decision circuits. The amplitude histograms are given by the absolute value of the derivation of this distribution, which is the same as that obtained by the synchronous sampling measurement. The Q-factor is evaluated using the amplitude histograms.

EXOR can be eliminated to realize the same function [3]. By counting "1" events at both the first (fixed decision level) and second (variable decision level) decision circuits while counting the clock period, and subtracting the number of "1" events, we can recognize the difference between a "1" event of the first and second decision circuits. Then the deviation of the probability curve of the difference also provides the amplitude histograms.

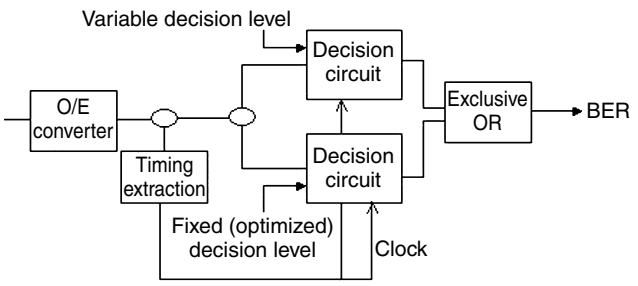


Figure 1. Typical configuration for dual decision circuits technique.

If the profile of the amplitude histograms does not change in the measurement time, the amplitude histograms can be estimated by using one decision circuit and a clock counter [5]. As a single decision circuit with a variable decision level does not have a reference, the error counter cannot reflect the influence of the amplitude fluctuation of each bit. This method is particularly advantageous with regard to BER measurement or error block detection, although it still requires timing extraction, which might depend on the signal format, bit rate, and/or modulation format.

3. ANALOG PARAMETERS

3.1. Statistical Method Using Histogram Evaluation

A method for signal quality monitoring that will provide a good measure of signal quality without the complexity of termination has long been desired and studied. Amplitude histogram estimation using a variable decision level (Section 2) can be recognized as a statistical method, but in this section, I focus on amplitude histogram evaluation using a sampling technique. Q-factor evaluation using synchronous sampling is more of a statistical method than a variable decision circuit technique because the sampling rate typically is low compared with the signal bit rate. However, Q-factor evaluation using the sampling technique is also a useful method when the system BER is too low to be measured within a reasonable amount of time, because it can be used for high bit rate signals, for which no decision circuit is used.

The amplitude histogram of an optical binary signal is obtained from an eye-diagram measured by the synchronous sampling technique. An amplitude histogram exhibits amplitude distributions at both mark and space levels at a fixed timing phase where the eye opening is at its maximum (Fig. 2a). The Q-factor at a fixed timing phase t ($Q(t_0)$) is estimated from the amplitude histogram, which is generally generated at timing phase t in the pattern as opposed to the data eye. It is defined as

$$Q(t_0) = |\mu_1(t) - \mu_0(t)|/(\sigma_1(t) + \sigma_0(t)) \tag{2}$$

where $\mu_i(t)$ and $\sigma_i(t)$ are the mean and standard deviation of the mark ($i = 1$) and space ($i = 0$) levels at t, respectively. The Q-factor (Q) is analytically related to the BER on a Gaussian assumption due to the equation defined by (1).

When the BER is low (e.g., $< 10^{-7}$), the theoretical value of Q is almost the same as the measured $Q(t_0)$ value:

$$Q(t_0) \sim Q \tag{3}$$

Eye monitoring using synchronous electrical sampling is conventionally undertaken with a digital sampling oscilloscope. However, the signal bit rate is limited by the O/E conversion bandwidth. Recently, an optical sampling

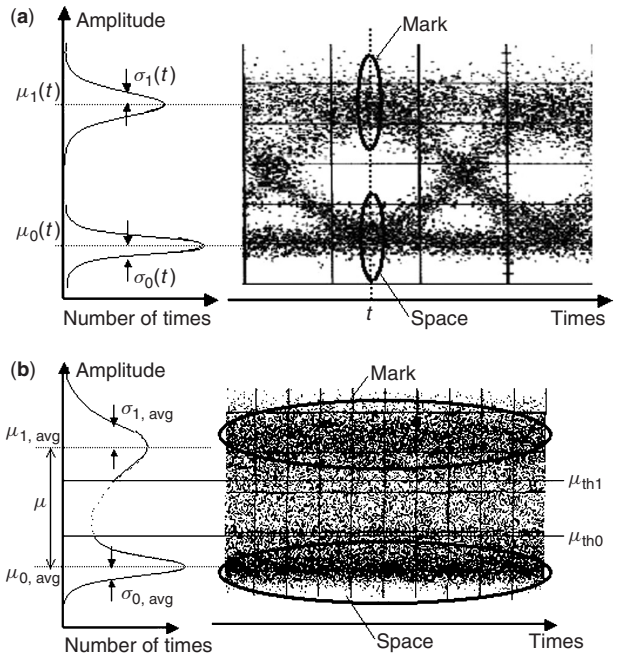


Figure 2. (**a**) Amplitude histograms at a fixed timing phase t of eye diagrams obtained by synchronous sampling; (**b**) amplitude histograms obtained by asynchronous sampling.

technique has been developed that has a temporal resolution of less than 1 ps and which overcomes the bit rate limitation [6]. Signal quality evaluation using eye diagrams obtained by means of optical sampling has also been reported [7,8]. The noise characteristics of optical amplifier systems do not have a Gaussian distribution. The analysis of non-Gaussian distributions is discussed in Ref. 9. Evaluations of crosstalk due to chromatic dispersion using histograms obtained by electrical sampling are presented in Ref. 10.

As mentioned previously, a statistical method using synchronous sampling is a useful technique for Q-factor evaluation. However, all sampling-based methods require synchronization and then some analysis, which makes them similar to protocol-aware termination in terms of cost and complexity. In fact, synchronous sampling also requires timing extraction using complex equipment that is specific to each BER and each format. Recently, the situation has begun to change. A simple, asynchronous histogram method was developed for Q-factor measurement [11]. Performance can be monitored at different monitoring points such as optical line repeaters, regenerators, or optical switching nodes (requires premeasurement). In other words, this method is expected to be applied to monitoring points where electrical termination is impossible. If we consider the all-optical network of the future, an optical switching node that has an all-optical switch will require performance monitoring without electrical termination.

Here, the averaged Q-factor (Q_{avg}) measurement obtained through asynchronous sampling is presented [11]. The asynchronous amplitude histogram is obtained from an eye-diagram measured by asynchronous optical sampling (Fig. 2b). Of the sampling points that constituted the histogram, it is determined that a set of points whose level is higher than a predetermined threshold level, μ_{th1}, belongs to level "mark" (i.e., "1"), while a set of points whose level is lower than a predetermined threshold level, μ_{th0} belongs to level "space" (i.e., "0"). Q_{avg} is defined by

$$Q_{avg} = |\mu_{1,avg} - \mu_{0,avg}|/(\sigma_{1,avg} + \sigma_{0,avg}) \tag{4}$$

where $\mu_{i,avg}$ and $\sigma_{i,avg}$ are the mean and standard deviation of the mark ($i = 1$) and space ($i = 0$) level distributions, respectively. The data obtained by asynchronous sampling include unwanted cross-point data in the eye-diagram, which reduces the measured value of the averaged Q-factor. Thus, it is necessary to remove the cross-point data. In this way, the two threshold levels were set at $\mu_{th1} = \mu_{1,avg} - \alpha\mu$ and $\mu_{th0} = \mu_{0,avg} + \alpha\mu$, and the coefficient α was defined as falling between 0 and 0.5.

In another approach, the peak levels of both mark and space level distribution (μ_1 and μ_0) are extracted, the data between μ_1 and μ_0 are eliminated and residual data at the mark level (larger than μ_1) and space level (smaller than μ_0) are symmetrically the reverse of the μ_1 and μ_0, respectively [12], and then the same evaluation is performed.

The essence of this method is that it does not use timing extraction or evaluate asynchronous eye diagrams. That is why this method provides signal format, modulation format, and bit rate flexibility. However, this technique is not timing jitter-sensitive despite the fact that jitter impairs the BER. This is the tradeoff with this method. Some analysis of bit rate flexibility and chromatic dispersion dependence has been provided [8].

3.2. Optical Power, Wavelength, and Optical SNR Monitor with Spectrum Measurement

Many methods using spectrum measurement have been published with a view to optical power, wavelength, and/or optical SNR (OSNR) monitoring. A simple approach for monitoring such kinds of analog parameter involves measuring the optical power spectrum of a tapped optical signal. However, this technique had two main problems. One is that the experimental equipment is large and expensive. The progress on DWDM networks makes it more difficult to monitor the optical power/ OSNR and wavelength of all channels. The second problem is that an optical spectrum monitor can measure the optical signal to out-of-band noise ratio, but it cannot monitor the optical SNR including in-band noise.

Recently, compact and stable equipment has been proposed by several groups to deal with the first problem. One approach employs an optical power and frequency monitor with a grating and a photo detector (PD) array [13], while another uses an arrayed waveguide grating (AWG) filter. With the latter technique, the AWG filtering wavelengths are controlled with the wavelength monitor and its feedback, using a reference light, thus realizing a precise wavelength and power monitoring. With yet another technique, a tunable Fabry-Perot etalon filter is used to separate the different spectral components in the spatial domain using temperature control [14]. An OSNR and optical wavelength monitor using an acousto-optic tunable filter have also been reported. The monolithic

integration of PD modules into an AWG filter is another approach for realizing an optical power and wavelength monitor in WDM networks. A point of interest as regards the abovementioned optical spectrum monitoring is its applicability to DWDM networks where the channel spacing is less than 50 GHz [15].

Further approaches have also been developed with a view to achieving precise OSNR monitoring. One technique monitors the OSNR by using the polarization-nulling technique, which employs the different polarization properties of optical signals and amplified spontaneous emission (ASE) noise [16]. The most recent approach to use this technique is reported in Ref. 16, where the polarization of an optical signal with ASE noise is controlled so that it is linear. The optical signal is split into a signal +ASE component and an ASE only component with polarization beam splitter (PBS). Then the ASE only component is divided into two with a 3 dB coupler. One of these ASE only components is optically filtered with a band-pass filter (BPF), and the powers of these three components are measured. As a result, OSNR is written as an equation of these three powers.

Another technique monitors OSNR by analyzing the low-frequency noise characteristics at the receiver [17]. With this technique, the total received power and received noise power are measured and compared. The noise power density is measured using an analog to digital converter and an FFT unit operating in the 40–50 kHz range. This technique is ineffective when the pattern length of the optical signal becomes longer than PRBS15. However, this limitation has been relaxed recently. In the latest result [18], the optical signal was split into orthogonally polarized lights by a polarization beam splitter and recombined after an optical delay $\Delta\tau$ in one of the paths. Then the measured electrical power P after the orthogonal delayed-homodyne module is written as

$$P = \{1 - 4\gamma(1-\gamma)\sin 2(\pi f \Delta\tau)\}S + N \qquad (5)$$

where γ is the power ratio at the polarization beam splitter, S and N are the electrical powers of the signal and noise, respectively, and f is the measured frequency. By setting $\Delta\tau$, f, and γ at adequate values, the electrical power component of the signal becomes zero, and the receiver noise component can be measured.

3.3. Pilot Tone Technique, Sub-Carrier Multiplexing

The pilot tone technique has been discussed for several years, and various approaches have been adopted. Sub-carrier multiplexing is used for the pilot tone and there are two main kinds of sub-carrier usage. One approach is to use the sub-carrier for monitoring the signal power, wavelength crosstalk, or OSNR. The other approach involves using the sub-carrier for an optical signal header. This means that the sub-carrier signal, which is modulated by header information data, is multiplexed into an optical signal (payload data) as a packet header or supervisory channel. Moreover, a variety of sub-carrier frequencies are used in these approaches (Fig. 3). The keys to these techniques relate to the methods used for combining and

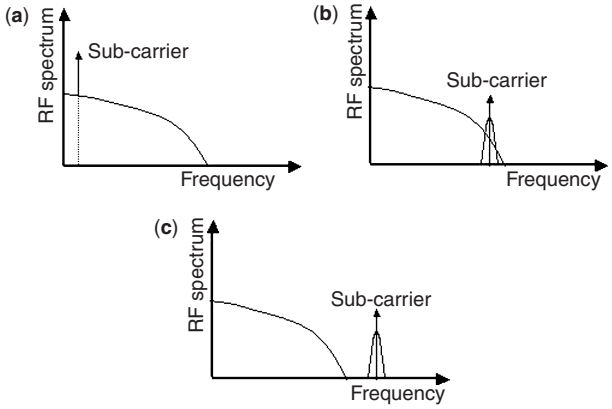


Figure 3. Typical RF spectrum of a data signal and a sub-carrier (pilot tone), when the sub-carrier frequency is (**a**) low (~ MHz), (**b**) near the signal bit rate (modulated by header information), and (**c**) higher than the signal bit rate (modulated by header information).

detecting the sub-carrier, the monitoring accuracy, and the influence of sub-carrier multiplexing on the signal.

The most effective approach for the pilot tone technique is to add a specific sinusoidal tone with a small amplitude to each optical WDM channel. In Ref. 19, the pilot tone is simply added to the laser bias current at the transmitter and is used to supervise individual wavelength channels along the optical path. At the transmitter an electrical pilot tone in the kHz regime is added to the signal, each wavelength channel being coded with a different tone frequency. The pilot tones are extracted at nodes between the transmitter and receiver by tapping a small portion of the signal power into a monitor module. The tapped optical power is detected, the pilot tones are filtered out electronically, and their levels are registered. The tones will provide signal identification and power level information for fault management. The tone amplitude is not more than 10% of the data level, and the tone frequency is below 100 kHz. This ensures that interference with the low-frequency components of the data results in negligible sensitivity degradation. Another approach uses a 10% peak-to-peak pilot tone modulation with a 50-kHz frequency and heterodyne detection with 20-Hz IF electrical filtering [20]. In Ref. 20, OSNR values measured by an optical spectrum analyzer and the pilot tone technique are compared, and good agreement in the OSNR region below 30 dB is reported. The OSNR sensitivity is limited to 30 dB with this technique because the electrical noise in the detector becomes nonnegligible.

Another approach designed simply for accurate frequency monitoring uses an AWG. With this approach, pilot tones with different RF frequencies are used for different wavelength channels. The pilot tones are split by the AWG where each center wavelength of each filter channel is set at the ITU grids using a temperature controller, so one AWG filter channel can split two pilot tones. The ratio of these two pilot tones with the same frequency split by adjacent AWG filters indicates the frequency of the optical channel [21].

Deciding on an adequate frequency range is also an important issue with respect to the accuracy of this

technique. Values ranging from several tens of kHz to the sub MHz level typically are used to minimize the penalty on a payload optical signal (Fig. 3a). However, a recent report states that when pilot tone–based monitoring techniques are used in amplified networks, their performance is deteriorated by the cross-gain modulation (XGM) of erbium-doped fiber amplifiers (EDFAs). The XGM problem is solved by using high-frequency tones in the 1-MHz range, but even when the pilot tone frequencies are in the few MHz range, the performance is limited by the Raman effect. As a result, tone frequencies higher than 100 MHz are recommended [22]. Another analysis suggests that, with regard to the pilot tone carrier to noise ratio, the tone detection sensitivity is limited by the power spectral density of the payload signal, and a pilot tone higher than 1 GHz is recommended for a 2.5-Gbit/s payload signal [23]. Some other approaches use pilot tone frequencies around the bit rate frequency, some of which is modulated by header information (Fig. 3b,c). The subject of pilot tone frequency remains an open issue.

In relation to the influence of the sub-carrier on the signal, the modulation intensity of a sub-carrier influences the payload signal, but detection sensitivity is reduced when the modulation intensity is low. This is a trade off problem and, for example, the number of payload signal wavelengths is limited because of this problem [23].

Wavelength conversion at intermediate nodes will become an important function in future DWDM networks. Some pilot tone–based approaches focusing on this function have already been reported [24]. These papers discuss the influence of the pilot tone in an interferometric wavelength converter using SOA, pilot tone frequency conversion with wavelength conversion using an semiconductor optical amplifier (SOA) + DFB laser, and the pilot tone technique when three SOA-interferometers are cascaded for wavelength conversion.

Some approaches use the sub-carrier technique for an optical signal header. Sub-carrier frequencies higher than the bit rate are sometimes used. The sub-carrier is encoded as a packet header using amplitude shift keying (ASK) or phase shift keying (PSK). The use of frequencies higher than the signal bit rate is advantageous in terms of the crosstalk between the sub-carrier and the payload signal. However, it is necessary to install high-speed, high-bandwidth electrical equipment at the transmitter and receiver. A sub-carrier frequency lower than the signal bit rate is used in [25] as a 50-Mbit/s NRZ channel overhead. A frequency of 9.73 GHz is used when the signal bit rate is 10 Gbit/s, using a differentially driven Mach-Zehnder (MZ) interferometer modulator. The sub-carrier is directly detected and filtered by an LPF. Another sub-carrier detection technique is proposed in Ref. 26. Here, the sub-carrier is encoded on the optical carrier by means of a dual-arm MZ $LiNbO_3$ (LN)-modulator, one arm is used for 10-Gbit/s data, the other is used for a 16.7-GHz RF tone modulated with a 100-Mbit/s ASK. A fiber loop mirror using polarization maintaining (PM) fiber birefringence is used to separate the baseband signal from the sub-carrier.

Other techniques focus on chromatic dispersion monitoring [27]. These techniques employ a chromatic dispersion monitor using the sub-carrier ratio method between higher and lower frequencies. The sub-carrier power measured at the receiver decreases due to the phase delay that the sub-carrier experiences in the dispersive fiber. A chromatic dispersion monitor using the optical side-band suppression method, which uses a sub-carrier frequency higher than the bit rate, measures the relative phase (time) delay between sub-carrier sidebands.

BIOGRAPHY

Ippei Shake was born in Kobe, Japan, in 1970. He received the B.S. and M.S. degrees in physics from Kyoto University, Kyoto, in 1994 and 1996, respectively. In 1996, he joined NTT Optical Network System Laboratories, NTT Corporation, Yokosuka, Japan. Since then he has been engaged in research and development of high-speed optical signal processing and high-speed optical transmission systems. He is now with NTT Network Innovation Laboratories, Yokosuka, Japan. His research interests also include optical networks, optical performance monitoring, and optical time-division multiplexing/demultiplexing circuits. He is a member of the Institute of Electrical and Electronics Engineers and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

BIBLIOGRAPHY

1. ITU-T Recommendation G.709.
2. N. S. Bergano, F. W. Kerfoot, and C. R. Davidson, Margin measurements in optical amplifier systems, *IEEE Photonics Tech. Lett.* **3**: 304–306 (1993).
3. R. Weismann, O. Bleck, and H. Heppner, Cost effective performance monitoring in WDM systems, *Optical Fiber Communication Conference 2000* (*OFC2000*), WK2, 2000.
4. M. Fregolent, S. Herbst, H. Soehnle, and B. Wedding, Adaptive optical receiver for performance monitoring and electronic mitigation of transmission impairments, *26th European Conference on Optical Communication* (*ECOC2000*), S2.1, 2000.
5. S. Ohteru and N. Takachio, Optical signal quality monitor using direct Q-factor measurement, *IEEE Photonics Tech. Lett.* **11**(10): 1307–1309 (1999).
6. H. Takara et al., 100 Gbit/s optical signal eye-diagram measurement with optical sampling using organic nonlinear optical crystal, *Electron. Lett.* **24**: 2256–2258 (1996).
7. C. Schmidt et al., Optical Q-factor monitoring at 160 Gb/s using an optical sampling system in an 80km transmission experiment, *Conference on Lasers and Electro-Optics 2002* (*CLEO 2002*), 579–580, 2002.
8. I. Shake and H. Takara, Transparent and flexible performance monitoring using amplitude histogram method, *OFC2002*, TuE1.
9. S. Norimatsu and M. Maruoka, Accurate Q-factor estimation of optically amplified systems in the presence of waveform distortion, *J. Lightwave, Tech.* **20**(1): 19–29 (2002).

10. C. M. Weinert, C. Caspar, M. Konitzer, and M. Rohde, Histogram method for identification and evaluation of crosstalk, *Electron. Lett.* **36**(6): 2000.

11. I. Shake, H. Takara, S. Kawanishi, and Y. Yamabayashi, Optical signal quality monitoring method based on optical sampling, *Electron. Lett.* **34**(22): 2152–2154 (1998).

12. M. Rasztovits-Wiech, K. Studer, and W. R. Leeb, Bit error probability estimation algorithm for signal supervision in all-optical networks, *Electron. Lett.* **35**(20): 1754–1755 (1999).

13. K. Otsuka et al., A high-performance optical spectrum monitor with high-speed measuring time for WDM optical networks, *23rd European Conference on Optical Communication* (*ECOC'97*), 147–150, 1997.

14. S. K. Shin, C. H. Lee, and Y. C. Chung, A novel frequency and power monitoring method for WDM network, *Optical Fiber Communication Conference '98* (*OFC'98*), WJ7, 1998.

15. H. Suzuki and N. Takachio, Optical signal quality monitor built into WDM linear repeaters using semiconductor arrayed waveguide grating filter monolithically integrated with eight photo diode, *Electron. Lett.* **35**(10): 836–837 (1999).

16. J. H. Lee and Y. C. Chung, Improved OSNR monitoring technique based on polarization-nulling method, *Electron. Lett.* **37**(15): (2001).

17. S. K. Shin, K. J. Park, and Y. C. Chung, A novel optical signal-to-noise ratio monitoring technique for WDM networks, *Optical Fiber Communication Conference 2000* (*OFC2000*), WK6, 2000.

18. C. J. Youn, K. J. Park, J. H. Lee, and Y. C. Chung, OSNR monitoring technique based on orthogonal delayed-homodyne method, *Optical Fiber Communication Conference 2002* (*OFC2002*), TuE3, 2002.

19. G. R. Hill et al., A transport network layer based on optical network elements, *J. Lightwave, Tech.* **11**(5): 667–679 (1993).

20. G. Bendelli, C. Cavazzoni, R. Girardi, and R. Lano, Optical performance monitoring technique, *26th European Conference on Optical Communication* (*ECOC2000*) **4**: 113–116 (2000).

21. C. J. Youn, S. K. Shin, K. J. Park, and Y. C. Chung, Optical frequency monitoring technique using arrayed-waveguide grating and pilot tones, *Electron. Lett.* **37**(16): 2001.

22. H. S. Chung et al., Effects of stimulated Raman scattering on pilot-tone-based WDM supervisory technique, *IEEE Photonics Tech. Lett.* **12**(6): 731–733 (2000).

23. Y. Hamazumi and M. Koga, Transmission capacity of optical path overhead transfer scheme using pilot tone for optical path network, *J. Lightwave, Tech.* **15**(12): 2197–2205 (1997).

24. A. Bissons et al., Analysis of evolution of over-modulated supervisory data in a cascade of all-optical wavelength converters, *Optical Fiber Communication Conference 2000* (*OFC2000*), ThD4, 2000.

25. M. Rohde et al., Control modulation technique for client independent optical performance monitoring and transport of channel overhead, *Optical Fiber Communication Conference 2002* (*OFC2002*), TuE2, 21–22, 2002.

26. G. Rossi, O. Jerphagnon, B.-E. Olsson, and D. J. Blumenthal, Optical SCM data extraction using a fiber-loop mirror for WDM network system, *IEEE Photonics Tech. Lett.* **12**(7): 897–899 (2000).

27. M. N. Petersen et al., Online chromatic dispersion monitoring and compensation using a single inband subcarrier tone, *IEEE Photonics Tech. Lett.* **14**(4): 570–572 (2002).

SIGNATURE SEQUENCES FOR CDMA COMMUNICATIONS

Tony Ottosson
Erik Ström
Arne Svensson
Chalmers University of Technology
Göteborg, Sweden

1. INTRODUCTION

A multiple access method is a method for allowing several users to share a common physical channel, such as, a coaxial cable, an optical fiber, or a band of radio frequencies. Common multiple access methods are frequency-division multiple access (FDMA) and time-division multiple access (TDMA) [1–3]. Strictly speaking, the term "multiple access" is applicable for the case when the users are not at the same geographic location and the term "multiplexing" is applicable when they are; however, we use "multiple access" for both cases. In FDMA, the users' signals are selected such that they do not overlap in the frequency domain (an example is FM or AM broadcast radio), and in TDMA, the users' signals do not overlap in time (i.e., the users take turns using the channel). Ideally, this means that the users do not disturb each other, and the signals are said to be orthogonal. (In practice, there will be some interuser interference due to imperfections in the implementation and nonideal channels, but we will ignore these complications here.)

A third multiple access method that has become increasingly popular is code-division multiple access (CDMA) [4,5]. In CDMA, the users signals are overlapping in both time and frequency. This does not, however, imply that CDMA signals are necessarily nonorthogonal. As we will see shortly it is indeed possible, under certain strong restrictions, to find orthogonal signals that overlap in time and frequency.

Let us consider a simple example to illustrate the main idea behind the most common type of CDMA: direct-sequence CDMA (DS-CDMA). The name DS-CDMA stems from the fact that the users' signals are direct-sequence spread spectrum signals (DS-SS signals) [6–8]. Suppose that we have two users who each want to transmit data at a rate of $1/T$ bits/second. Let $b_1(t)$ and $b_2(t)$ be the data waveform of user 1 and 2, respectively, where bits are coded as ± 1 amplitudes of the waveforms. The kth user is assigned a signature waveform, $c_k(t)$, and the transmitted signal, $s_k(t)$, is formed as $s_k(t) = b_k(t)c_k(t)$, see Fig. 1. The data waveform is defined by the user's information bit sequence: in this case $b_1[n] = \{1, -1, \ldots\}$ for user 1 and $b_2[n] = \{-1, 1, \ldots\}$ for user 2. Similarly, the signature waveforms are defined by the users' signature sequences: $c_1[n] = \{1, -1, 1, 1, 1, -1, 1, 1, \ldots\}$ for user 1 and $c_2[n] = \{1, 1, -1, 1, 1, 1, -1, 1, \ldots\}$ for user 2. We note that the data waveforms can change polarity every T seconds and the signature waveforms can change polarity every T_c seconds. The ratio $N = T/T_c$ is called the spreading factor and also processing gain.

It is the signature waveform (also known as the code waveform or spreading waveform) that enables the

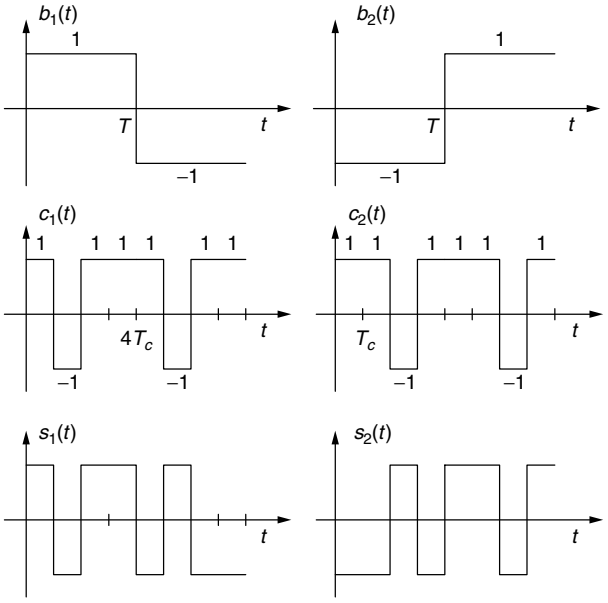


Figure 1. Data waveforms, signature waveforms, and transmitted waveforms for two users in a direct-sequence CDMA system.

receiver to separate users. Hence, the waveforms must be unique among the users. The process of multiplying the data waveform with the signature waveform is called spreading, since the bandwidth of the transmitted signal is significantly larger than the bandwidth of the data waveform. As a matter of fact, the bandwidth of the transmitted signal is approximately N times larger than the data waveform bandwidth (which justifies the terminology "spreading factor" for N).

The transmitted signals are added by the channel[1] and the received waveform is $r(t) = s_1(t) + s_2(t)$ (we ignore channel noise and other nonideal channel effects here). Now suppose that the receiver is interested in detecting the data from user 1. A block diagram of the system under consideration is found in Fig. 2. The receiver multiplies the received signal with the signature waveform of user 1, $c_1(t)$, and this results in the signal

$$\begin{aligned} z_1(t) &= r(t)c_1(t) \\ &= [b_1(t)c_1(t) + b_2(t)c_2(t)]c_1(t) \\ &= b_1(t)\underbrace{c_1^2(t)}_{=1} + b_2(t)c_2(t)c_1(t) \\ &= b_1(t) + b_2(t)c_2(t)c_1(t) \end{aligned}$$

We see that the signal $z_1(t)$ is the sum of two terms: the desired data waveform of user 1 and an interference term that is due to user 2. Since we have reduced the bandwidth of the desired part of the signal, the process of multiplying the received signal with the signature waveform is called despreading.

The standard method for recovering the data from a signal disturbed by additive noise or interference is to process the noisy signal with a matched filter. In this case, the filter should be matched to a rectangular pulse of length T seconds. Hence, the matched filter can be implemented as an integrator over T seconds. That is, the output of the matched filter is

$$y_1(t) = \int_{t-T}^{t} z_1(u)\,du$$

To decided on the nth bit, we sample the matched filter at time $(n+1)T$. Hence, to recover $b_1[0]$, we sample the filter at time T, which results in

$$\begin{aligned} y_1(T) &= \int_0^T z_1(t)\,dt \\ &= \int_0^T b_1(t)\,dt + \int_0^T b_2(t)c_1(t)c_2(t)\,dt \\ &= \int_0^T b_1[0]\,dt + b_2[0]\underbrace{\int_0^T c_1(t)c_2(t)\,dt}_{=0} \\ &= b_1[0]T \end{aligned}$$

where $b_1[0] = 1$ and $b_2[0] = -1$ are the transmitted bits. We see that the data bit from user 1 can be recovered as $\text{sgn}[y_1(T)]$, regardless of the value of $b_2[0]$. The condition

[1] There are situations when the transmitted signals are added already in the transmitter. This is the case when several users' signals are transmitted from the same geographical position. One example of this is a downlink in a cellular CDMA system, where all the users' signals in one cell are transmitted from the same base station [1–3].

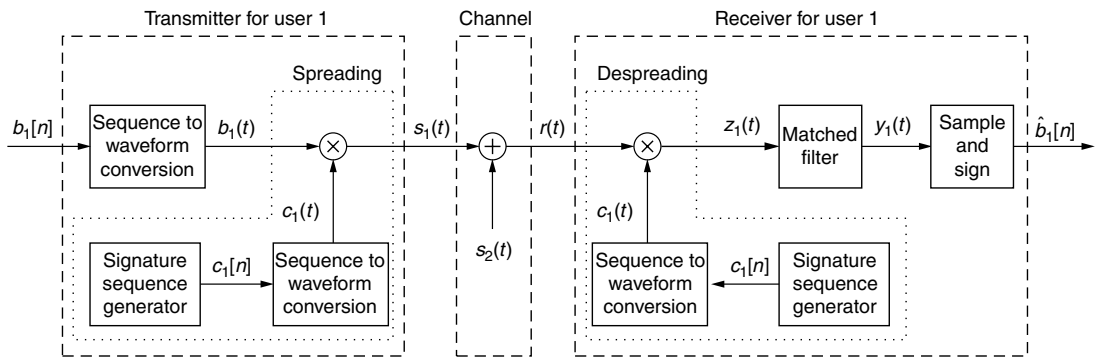


Figure 2. A block diagram for the transmitter and receiver for user 1. The channel is assumed to be noise-free and the only interference added is the signal from user 2.

that the cross-correlation between $c_1(t)$ and $c_2(t)$ is zero, that is,

$$\int_0^T c_1(t)c_2(t)\,dt = 0$$

proves that the signature waveforms $c_1(t)$ and $c_2(t)$ are orthogonal (over the interval $0 \le t \le T$). As long as the signature waveforms are orthogonal, users will not interfere with each other after despreading and matched filtering, and the data can be recovered without error (if we ignore channel noise). However, it is not necessary for the signature sequences to be completely orthogonal for a CDMA system to work. As a matter of fact, as long as the cross-correlation between the signature waveforms is small, the interference alone cannot cause errors (although the resistance against noise can be reduced).

It should now be clear that we need to choose signature sequences carefully to avoid excessive interference between users. In short, we want to select sequences that have low cross-correlations. As is discussed in more detail later, it is also important to select sequences that have good partial cross-correlation and autocorrelation properties. The autocorrelation of a sequence is the cross-correlation between the sequence and a time-shifted version of the same sequence. Ideally, we want the autocorrelation to be small for all nonzero time-shifts. The partial cross-correlation is the correlation between partial segments of the sequences. Ideally, the partial cross-correlation should be small for all choices of sequence segments.

This definition of "good correlation properties" is applicable for DS-CDMA, but not necessarily for other types of CDMA. Although many forms of CDMA exist, the most common type apart from DS-CDMA is frequency-hopping CDMA (FH-CDMA). FH-CDMA is similar to FDMA in that the users are assigned different frequency channels that are meant to be for exclusive use by the users. In FH-CDMA, the frequency channel allocation for a user is constantly changing. That is, the user hops among frequencies, and the hopping is done according to the user's signature sequence. Every now and then, two or more users will be assigned the same frequency. This is known as a collision and is a highly undesirable event, and FH-CDMA signature sequences are therefore primarily designed to avoid collisions. Hence, the signature sequences used for DS-CDMA and FH-CDMA are different. We will not cover FH-CDMA sequences further in this article; the interested reader is referred to Ref. [9] for an overview. More details on FH sequences (and DS sequences) can also be found in Refs. [6–8].

In addition to good cross-correlation properties (or good collision properties), it is desirable that the sequences are easily generated, that is, with as little hardware and software complexity as possible. Furthermore, in some applications, we also are interested in making the signature sequence a secret for all but the intended receiver. It is then important that it will be difficult for an unauthorized receiver to predict future values of the signature sequence by observing its past values. It should be noted that in most current CDMA systems, the signature sequences have not been designed with this last requirement in mind. Hence, security in these systems usually is obtained by other cryptographic methods.

We know that completely random sequences have, on average, good correlation properties and are impossible to predict. However, we cannot use true random sequences since both the transmitter and the receiver must be able to produce the same sequences. Instead, most sequences used in practice are derived from pseudorandom sequences (PN-sequences). The theory of PN sequences involves finite-field algebra; however, we will leave out all details here and rather give a brief description of how PN-sequences can be generated with digital hardware (or software). Most sequences in practical use are binary sequences (also known as bi-phase sequences) or 4-ary sequences (quadriphase sequences). The latter type is useful for quadrature modulated systems (e.g., phase-shift keying or quadrature amplitude modulation).

2. SPREAD SPECTRUM AND CDMA

In the introduction, it is noted by a simple example that, the signature sequences should be unique and have low partial cross-correlations for all time-shifts. To quantify these statements, let us go into more details.

As before, assume that a user k transmits the signal

$$s_k(t) = \sqrt{E_k/T}\,b_k(t)c_k(t) \tag{1}$$

where

$$b_k(t) = \sum_{i=0}^{\infty} b_k[i]h(t - iT) \tag{2}$$

is the data waveform

$$c_k(t) = \sum_{i=0}^{\infty} c_k[i]g(t - iT_c) \tag{3}$$

the signature waveform, and $h(t)$ a rectangular pulse shape given by

$$h(t) = \begin{cases} 1 & 0 \le t \le T \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The chip-pulse-shape $g(t)$ can in general take any form. To simplify the presentation here, only the rectangular chip-pulse-shape is considered, such that

$$g(t) = \begin{cases} 1 & 0 \le t \le T_c \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In practice, more spectrally efficient waveforms are used, but for the purpose of this article this only makes the mathematics untrackable. The properties and results that are discussed here hold with small changes for all pulse shapes used in practice. The data sequence and the chip sequence are denoted $b_k[i]$ and $c_k[i]$, respectively. In general, both of these sequences are complex-valued and take values from limited sets. Again for simplicity, we will concentrate here on real-valued data symbols and binary chips. The only difference in this model compared to the model presented in the introduction is that we have normalized the signal so its energy is

$$\int_0^T s_k^2(t)\,dt = \frac{E_k}{T}\int_0^T b_k^2(t)c_k^2(t)\,dt = \frac{E_k}{T}T = E_k \tag{6}$$

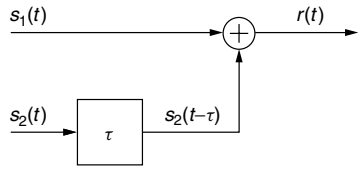


Figure 3. Block diagram of a two-user asynchronous noise-free channel with relative delay τ.

Let us consider a two-user case where the users are asynchronous (i.e., not time-aligned); see Fig. 3. This is the typical situation in a true multiple access scenario (such as the uplink in a cellular system). The received waveform can be expressed as

$$r(t) = s_1(t) + s_2(t-\tau) \tag{7}$$

where τ is the relative time-offset between the signals of users 1 and 2 depending on the propagation times of the user signals, and hence their distances to the receiver. Without loss of generality, we assume that $0 \le \tau < T$. Here we have, for simplicity, neglected the receiver noise that is always present, since it will not have any influence on the properties of the signature waveforms.

Assuming that we are interested in detecting user 1, we first despread the signal by multiplying with the waveform of user 1 as

$$\begin{aligned} z_1(t) &= r(t)c_1(t) \\ &= \left[\sqrt{E_1/T}b_1(t)c_1(t) + \sqrt{E_2/T}b_2(t-\tau)c_2(t-\tau)\right]c_1(t) \\ &= \sqrt{E_1/T}b_1(t) + \sqrt{E_2/T}b_2(t-\tau)c_2(t-\tau)c_1(t) \end{aligned} \tag{8}$$

The despread signal consists of the desired signal part and an interfering part from user 2. To recover $b_1[0]$ we calculate the output of the normalized filter matched to $h(t)$ at time T, which becomes

$$\begin{aligned} y_1(T) &= \frac{1}{\sqrt{T}}\int_0^T z_1(t)\,dt \\ &= \sqrt{E_1}\frac{1}{T}\int_0^T b_1(t)\,dt \\ &\quad + \sqrt{E_2}\frac{1}{T}\int_0^T b_2(t-\tau)c_2(t-\tau)c_1(t)\,dt \\ &= \sqrt{E_1}b_1[0]\frac{T}{T} + \sqrt{E_2}b_2[-1]\frac{1}{T}\int_0^\tau c_2(t-\tau)c_1(t)\,dt \\ &\quad + \sqrt{E_2}b_2[0]\frac{1}{T}\int_\tau^T c_2(t-\tau)c_1(t)\,dt \end{aligned} \tag{9}$$

The integrals in this expression are partial cross-correlations between the waveforms $c_1(t)$ and a time-shifted version of $c_2(t)$. Assuming that τ is a multiple of the chip time given as $\tau = pT_c$ $(0 \le p < N)$, we can calculate the individual partial cross-correlations as function of the discrete signature sequences as

$$\frac{1}{T}\int_0^\tau c_2(t-\tau)c_1(t)\,dt = \begin{cases} \dfrac{T_c}{T}\displaystyle\sum_{n=0}^{p-1} c_2[n-p]c_1[n] & 0 < p < N \\ 0 & p = 0 \end{cases} \tag{10}$$

$$\frac{1}{T}\int_\tau^T c_2(t-\tau)c_1(t)\,dt = \frac{T_c}{T}\sum_{n=p}^{N-1} c_2[n-p]c_1[n] \tag{11}$$

These derivations can easily be extended to more than two users with user i as the desired user. We now introduce the discrete partial cross-correlations as

$$X_{k,i}(p) = \sum_{n=p}^{N-1} c_k[n-p]c_i[n] \quad 0 \le p < N \tag{12}$$

and

$$\overline{X}_{k,i}(p) = \begin{cases} \displaystyle\sum_{n=0}^{p-1} c_k[n-p]c_i[n] & 0 < p < N \\ 0 & p = 0 \end{cases} \tag{13}$$

With these discrete partial cross-correlations the matched filter output in Eq. (9) can be rewritten as

$$\begin{aligned} y_1(T) &= \sqrt{E_1}b_1[0] + \sqrt{E_2}b_2[-1]\overline{X}_{2,1}(p)/N \\ &\quad + \sqrt{E_2}b_2[0]X_{2,1}(p)/N \end{aligned} \tag{14}$$

We clearly see that we get the desired signal $\sqrt{E_1}b_1[0]$ but also two multiple-access interference (MAI) terms that depend on the partial cross-correlations between the signature sequences of the two users. For ideal output (only the desired part), the sum of the two MAI terms should be zero for any combination of data symbols $b_2[-1]$ and $b_2[0]$, and for all values of the time-shift p, since p may vary. These equations can easily be extended to more users. The MAI term from user k on user i is obtained by replacing index 2 by index k and index 1 by index i in Eq. (14). Ideally, the sum of all these MAI terms should be zero for all time-shifts p. This is guaranteed when $\overline{X}_{k,i}(p) = 0$ and $X_{k,i}(p) = 0$ for all $p \in [0, N-1]$ and for all $k \ne i$, although in some cases this is an unnecessarily strong requirement.

For signature sequences with period N the cross-correlations can be expressed as

$$\begin{aligned} X_{k,i}(p) &= C_{k,i}(p) \\ \overline{X}_{k,i}(p) &= C_{k,i}(p-N) \end{aligned} \tag{15}$$

where we have introduced the aperiodic cross-correlation parameter $C_{k,i}(m)$ defined as

$$C_{k,i}(m) = \begin{cases} \displaystyle\sum_{n=0}^{N-1-m} c_k[n]c_i[n+m] & \text{if } 0 \le m \le N-1, \\ \displaystyle\sum_{n=0}^{N-1+m} c_k[n-m]c_i[n] & \text{if } 1-N \le m < 0, \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

which is commonly used in the literature [10].

In many systems, the channel for the kth user can be modeled as a channel with L distinct propagation paths, where the lth path has complex gain $g_k(l)$ and delay $\tau_k(l)$.

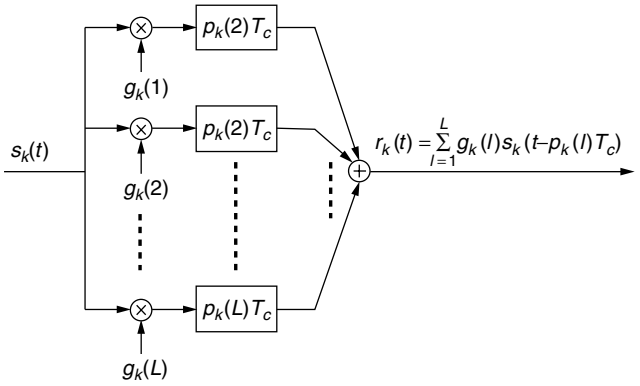


Figure 4. Noise-free L-path channel for the kth user. The lth path has complex gain $g_k(l)$ and delay $p_k(l)T_c$.

For simplicity, we assume that the delays can be written as $\tau_k(l) = p_k(l)T_c$, where $p_k(l)$ is an integer in $[0, N-1]$ and $p_1(1) = 0$. The multipath channel for the kth user is shown in Fig. 4.

Generalizing the given expressions for the case of a uplink CDMA system with K users experiencing delay spread, it can be shown that the matched filter output for user 1 at time T (path 1) can be written as

$$\begin{aligned} y_1(T) = {} & \sqrt{E_1}g_1(1)b_1[0] + \\ & \underbrace{\sum_{l=2}^{L}\sqrt{E_1}g_1(l)\Big(b_1[-1]\overline{X}_{1,1}(p_1(l))/N + b_1[0]X_{1,1}(p_1(l))/N\Big)}_{\text{ISI(+ICI)}} \\ & + \underbrace{\sum_{k=2}^{K}\sum_{l=1}^{L}\sqrt{E_k}g_k(l)\Big(b_k[-1]\overline{X}_{k,1}(p_k(l))/N + b_k[0]X_{k,1}(p_k(l))/N\Big)}_{\text{MAI}} \end{aligned} \tag{17}$$

The desired part is $\sqrt{E_1}g_1(1)b_1[0]$ in Eq. 17. The second part is the intersymbol interference (ISI), which is a weighted sum of aperiodic autocorrelations (partial cross-correlations between two equal sequences with different time-shifts) $\overline{X}_{1,1}(p_1(l))$ and $X_{1,1}(p_1(l))$. The part of the ISI that depends on $b_1[0]$ is sometimes referred to as interchip interference (ICI), but we prefer to refer to all of it as ISI. The third part is the MAI part, which is a weighted sum of the partial cross-correlations $\overline{X}_{k,1}(p_k(l))$ and $X_{k,1}(p_k(l))$. The ideal signature sequences should make the matched filter output as close as possible to the desired component. Again, this is obtained when all partial cross-correlations are zero for all time-shifts and the partial autocorrelation is zero for all time-shifts except zero. In practice, it may be enough to require that these partial correlations are close to zero.

From Eq. (17) it is seen that some information about the desired data symbol $b_1[0]$ is not used to form the desired part of the decision variable but instead becomes interference. The performance of a spread spectrum and CDMA receiver can be improved by using all the information about $b_1[0]$, which is available in the received signal (also that in the multipath components of the desired signal) [6–8]. Such a receiver is not matched to the transmitted waveform of the desired user only, but to that waveform convolved with the channel impulse response, and is commonly referred to as a RAKE receiver [6,8,11]. Now the decision variable becomes the weighted sum

$$y_{\text{rake}} = \sum_{l=1}^{L} g_1^*(l)y_1(T + p_1(l)T_c) \tag{18}$$

This expression can be worked out in the same way as Eq. (17). Again, it consists of three terms: the desired term, an ISI term, and a MAI term. The ISI and MAI terms are guaranteed to becomes zero under the same requirements as above, that is, the partial autocorrelation should be zero at all time-shifts except 0 and the partial cross-correlations should be zero at all time-shifts.

So far we have considered only the decision variable for data detection. The normalized matched filter output

$$y_1(t) = \frac{1}{\sqrt{T}}\int_{t-T}^{t} z_1(u)\,du \tag{19}$$

may also be used for acquisition and synchronization. Note that we have assumed that the sampling point T used to obtain a decision variable for $b_1[0]$ was known in the preceding derivation. In practice, this sampling point must be estimated. One common way of doing this is to find the maximum over a symbol interval of the matched filter output, since this maximum most likely corresponds to the time-shift where the timing of the received signal and the regenerated signal in the receiver are aligned. This normally is implemented as a two-stage procedure, where the matched filter output $y_1(t)$ is compared to a threshold in the first stage. The time where the threshold is exceeded is taken as a rough estimate of the timing (aquisition). In the next stage, the maximum of $y_1(t)$ in the vicinity of the point found in the first stage is found (synchronization) [6,8].

The contribution from the desired signal in $y_1(t)$ becomes very similar to the MAI term in Eq. (9), except that the user index is the same on both signature waveforms. After some derivations,

$$y_1(pT_c) = \sqrt{E_1}b_1[-1]\overline{X}_{1,1}(p)/N + \sqrt{E_1}b_1[0]X_{1,1}(p)/N \tag{20}$$

when $r(t) = s_1(t)$. On a multipath channel, there will be more terms of the type $\overline{X}_{1,1}(p + p_1(l))$ and $X_{1,1}(p + p_1(l))$. To have a distinct peak to use for timing estimation, it is clear that the requirements on the signature sequence used for synchronization is the same as the requirement to obtain low ISI in data detection.

3. COMMON SIGNATURE SEQUENCES

In this section, we describe some commonly used signature sequences and briefly discuss their properties.

The signature waveform of user k will, as before, be denoted as $c_k(t)$. This waveform is obtained as given in Eq. (3) where $c_k[i]$ takes values from $\{\pm 1\}$. The signature sequences are commonly defined by arithmetics in the binary field GF(2) where the elements are denoted as 0 and 1. In the following, we use uppercase letters to denote variables in GF(2) and lowercase letters to denote the corresponding antipodal variables. This means that the antipodal signature sequence is obtained from the corresponding signature sequence in GF(2) from

$$c_k[i] = 1 - 2\,C_k[i] \tag{21}$$

We do not use different symbols for addition and multiplication, because we believe it is clear from the expression whether it is operations in GF(2) or in the field of real numbers. The user index k will be suppressed when there is no reason to distinguish between several users.

Periodic sequences, like the ones discussed in this section, can be used in several ways to form signature waveforms for spread spectrum and CDMA. One way is to map the complete binary sequence to an antipodal sequence using Eq. (21), and then form the signature waveform as described in Eq. (3). When the spreading factor N is identical to the period P, this is referred to as short codes, while the case of $P > N$ corresponds to long codes.[2] With short codes, the signature waveform becomes identical for each transmitted symbol, while it changes over time for long codes. However, sometimes instead the signature waveform is obtained from a periodic repetition of part of the full period of the original sequence. Mathematically, this means that the signature sequence used to form the signature waveform is given by

$$\{C[0], C[1], \ldots, C[Q], C[0], C[1], \ldots, C[Q], \ldots\} \tag{22}$$

where $Q < P$. Different users may use different phase shifts of the same periodic sequence, or they may use different periodic sequences.

3.1. Maximal Length Sequences

Maximal length sequences (m-sequences) are generated by a shift register with feedback and have the maximum period that can be obtained with the given shift register [12]. The signature sequence $\{C[0], C[1], \ldots\}$ is generated by the recursive formula

$$\begin{aligned} C[i] &= G[1]C[i-1] + G[2]C[i-2] + \cdots + G[m]C[i-m] \\ &= \sum_{k=1}^{m} G[k]C[i-k] \end{aligned} \tag{23}$$

where $i \geq m$ and m is the length (memory) of the shift register and is commonly also referred to as the degree of the sequence. The coefficients $\{G[1], G[2], \ldots, G[m]\}$ and the initial state $\{C[0], C[1], \ldots, C[m-1]\}$ specify the sequence. The coefficient $G[m]$ is always 1 for binary m-sequences. The maximum period of this signature sequence is $P = 2^m - 1$ and is obtained when the characteristic polynomial $G(D) = G[m]D^m + G[m-1]D^{m-1} + \cdots + D + 1$ is an irreducible and primitive polynomial that divides $D^P + 1$ [7,8,12,13]. An irreducible polynomial cannot be factored and a primitive polynomial $G(D)$ of degree m is one for which the period of the coefficients of $1/G(D)$ is P.

M-sequences exist for all $m > 1$ and the number of characteristic polynomials is

$$N_p(m) = \frac{2^m - 1}{m} \prod_{i=1}^{k} \frac{P_i - 1}{P_i}$$

where $\{P_1, P_2, \ldots, P_k\}$ are prime numbers such that

$$2^m - 1 = \prod_{i=1}^{k} P_i^{m_i}$$

where $\{m_1, m_2, \ldots, m_k\}$ are integers.[3] It should be noted that a characteristic polynomial $G(D)$ can be reversed as $G'(D) = D^m + G[1]D^{m-1} + \cdots + G[m-1]D + G[m]$ to give a reversed m-sequence. These reversed polynomials are included in $N_p(m)$. For a given characteristic polynomial, different initial states give a different phase shift of the same sequence. In Table 1, the number of characteristic polynomials and the maximum period are given for degree m m-sequences. It is clear that the number of characteristic polynomials and thus sequences increases very fast when m increases. Characteristic polynomials for many periods can be found in Refs. [7,8,11,14,15] and we summarize some of them in Table 2 (the reversed polynomials are not included in this table).

Table 1. The Maximum Period and the Number of Characteristic Polynomials for m-Sequences of Degree *m*

m	$P = 2^m - 1$	$N_p(m)$
2	3	1
3	7	2
4	15	2
5	31	6
6	63	6
7	127	18
8	255	16
9	511	48
10	1023	60
11	2047	176
12	4095	144
13	8191	630
14	16383	756
15	32767	1800
16	65535	2048
17	131071	7710
18	262143	8064
19	524287	27594
20	1048575	24000

[2] The case $N > P$ should be avoided and is not discussed further here.

[3] This is a so-called prime decomposition.

Table 2. Characteristic Polynomials of m-Sequences. The Polynomials Are Given in Octal Notation. After Converting the Octal Numbers to Binary Numbers, the Left-Most Binary 1 Corresponds to $G[m]$. As an Example, $23_{\text{octal}} = 010011_{\text{binary}}$, Which Corresponds to $G(D) = D^4 + D + 1$

m	Characteristic Polynomials in Octal Form: $\{G[m], G[m-1], \ldots, G[1], 1\}$
2	7
3	13
4	23
5	45, 75, 67
6	103, 147 155
7	211, 217, 235, 367, 277, 325, 203, 313, 345
8	435, 551, 747, 453, 545, 537, 703, 543
9	1021, 1131, 1461, 1423, 1055, 1167, 1541, 1333, 1605, 1751, 1743, 1617, 1553, 1157
10	2011, 2415, 3771, 2157, 3515, 2773, 2033, 2443, 2461, 3023, 3543, 2745, 2431, 3177

In every period of length P of a m-sequence, the number of zeros is $2^{m-1} - 1$ and the number of ones is 2^{m-1}. Moreover, half of the number of runs of ones and zeros have length 1, 1/4 have length 2, 1/8 have length 3, and in general $1/2^k$ have length k with $k < m$.

With short codes using the full period of the m-sequences, that is, $Q = P = N$, it is well known that the periodic autocorrelation function is given by Refs. [7,8]

$$\phi(p) = \overline{X}_{k,k}(p) + X_{k,k}(p) = \begin{cases} N & \text{when } p = 0 \\ -1 & 1 \le p < N \end{cases} \tag{24}$$

For large N, $\phi(p)/\phi(0) = -1/N$, where $p \neq 0$, becomes insignificant, and the autocorrelation function approaches the autocorrelation function of an uncorrelated sequence. Therefore, m-sequences have excellent properties for synchronization in direct sequence spread spectrum when the data waveform is constant and the channel flat, since it contains well-defined autocorrelation peaks. They also lead to very small ISI under the same assumption. However, it is also well known that the periodic cross-correlation given by

$$\Phi_{k,i}(p) = \overline{X}_{k,i}(p) + X_{k,i}(p) \tag{25}$$

may have significant values [10,11]. The maximum periodic cross-correlation

$$\Phi_{\max} = \max_{p,k,i \neq k} |\Phi_{k,i}(p)| \tag{26}$$

between any pair of m-sequences is given in Table 3 for some short m-sequences. From table 3 it is clear that the maximum cross-correlation peak may be more than one-third of the length of the sequence. The maximum MAI is proportional to the maximum cross-correlation values when $b_k[-1] = b_k[0]$. Much less appears to be known about ISI and MAI in the general case when $b_k[-1] \neq b_k[0]$ and also about the synchronization properties when $b_i[-1] \neq b_i[0]$.

As an example of short codes based on m-sequences, we choose the length 31 sequence generated by using $G(D) = D^5 + D^4 + D^3 + D^2 + 1$ (75 in octal as seen in Table 2). When the initial values of the signature sequence are all ones, the first period becomes {1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0}. In Fig. 5 the aperiodic and periodic discrete autocorrelation sequences ($\phi(p)$) are shown as well as $X_{1,1}(p) - \overline{X}_{1,1}(p)$, which is proportional to the MAI when $b_k[-1] = -b_k[0]$. With binary antipodal modulation, the absolute value of the maximum MAI between two users using different shift of the same m-sequence is proportional to the maximum of $|X_{1,1}(p) - \overline{X}_{1,1}(p)|$ and

Table 3. The Maximum Periodic Cross-Correlation $\Phi_{\max}$ for m-Sequences of Degree m

m	$\Phi_{\max}$	$\Phi_{\max}/P$
3	5	0.71
4	9	0.60
5	11	0.35
6	23	0.36
7	41	0.32
8	95	0.37
9	113	0.22
10	383	0.37
11	287	0.14
12	1407	0.34

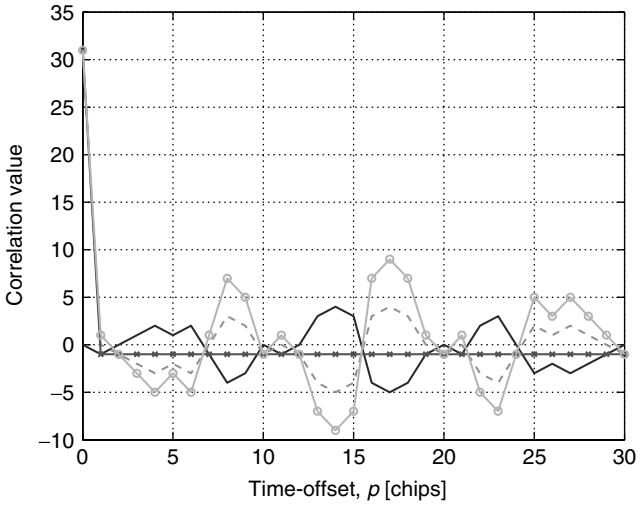


Figure 5. $\overline{X}_{1,1}(p)$ (solid line), $X_{1,1}(p)$ (dashed line), $\phi(p)$ (cross), and $X_{1,1}(p) - \overline{X}_{1,1}(p)$ (circle) for the m-sequence based on characteristic polynomial $G(D) = D^5 + D^4 + D^3 + D^2 + 1$.

$|\phi(p)|$ for $p \neq 0$. From Fig. 5 we see that this maximum MAI in this example is not larger than the maximum periodic cross-correlation $\Phi_{\max}$ as shown in Table 3. However, the maximum cross-correlation $|\Phi_{k,i}(p)|$ between the m-sequences generated by $G(D) = D^5 + D^4 + D^3 + D^2 + 1$ and $G(D) = D^5 + D^2 + 1$ (45 in octal) is 13, so the maximum correlation is not guaranteed to be at most the values given in Table 3. The conclusion is that $\Phi_{\max}$ should not be taken as a guarantee of the maximum MAI.

With long codes, much less is known about the autocorrelation and cross-correlation properties of signature sequences based on m-sequences. In this case, it is the partial correlations over $N < P$ chips that are important. The same is true for both long and short codes on multipath fading channels. In this case, the correlation properties also depend on whether a simple matched filter receiver is used or a RAKE receiver is used (or any other receiver). In the first case, it is the partial correlation between the transmitted signature waveform of user k convolved with the channel impulse response of user k and the signature waveform of the desired used that should have small values to reduce MAI. In the second case, it is the partial correlation between the transmitted signature waveform of user k convolved with the channel impulse response of user k and the signature waveform of the desired used convolved with the channel impulse response of the desired user that influence MAI. It seems that very little is known about these correlation properties on multipath channels in general for m-sequences.

M-sequences are used in the IS-95 CDMA system developed by Qualcomm [16]. In the uplink, a long m-sequence with period $2^{42} - 1$ is used to distinguish different channels (channelization). In both the uplink and the downlink, m-sequences with period $2^{15} - 1$ are used to separate mobiles (uplink) and base stations (downlink). Separate m-sequences are used on the I and Q channels in both directions. In the downlink, the data are also scrambled by a decimated long m-sequence.

3.2. Gold Sequences

M-sequences lead to the excessive amount of MAI in CDMA systems as seen previously. Another family of periodic signature sequences with somewhat better properties are Gold sequences [7,8,11,17,18]. A Gold sequence is obtained as the sum of two so-called preferred pairs of m-sequences, that is, $C[i] = C'[i] + C''[i]$, where $C'[i]$ and $C''[i]$ are the ith chips of two different m-sequences. In fact, each preferred pair of m-sequences generates a whole family of Gold sequences, namely, each of the two m-sequences alone and the sum of one of them with any shift of the other. Therefore, every preferred pair generates $P + 2$ Gold sequences, where P as before is the period of the m-sequence. A limited set of characteristic polynomials for preferred pairs of m-sequences is given in Table 4. A more complete table can be found in Ref. [7, page 502].

When $N = P$, the periodic autocorrelation $\phi(p)$ with $p \neq 0$ and the periodic cross-correlation $\Phi_{k,i}(p)$ can be shown to take at most three values, which are $\{-1, -t(m), t(m) - 2\}$, where

$$t(m) = \begin{cases} 2^{(m+1)/2} + 1 & m \text{ odd} \\ 2^{(m+2)/2} + 1 & m \text{ even} \end{cases} \tag{27}$$

Table 4. Characteristic Polynomial for Preferred Pairs of m-Sequences That Are Used to Form Gold Sequences. A More Complete Table can be found in Ref. [7, page 502]. The Octal Notation is Explained in Table 2

m	$P = 2^m - 1$	Preferred Pairs of Generator Polynomials in Octal Form
5	31	[45,75]
6	63	None
7	127	[211,217], [211,277]
8	255	[747,703]
9	511	[1021,1131], [1131,1423]

Table 5. The Maximum Periodic Autocorrelation and Cross-Correlation $\Phi_{\max}$ for Gold Sequences of Degree m

m	$t(m) = \Phi_{\max}$	$\Phi_{\max}/P$
5	9	0.29
6	17	0.27
7	17	0.13
8	33	0.13
9	33	0.06
10	65	0.06
11	65	0.03
12	129	0.03

These values are given in Table 5 together with the corresponding normalized values. Here it is clearly seen that these values are much improved compared with the corresponding values for m-sequences (see Table 3).

Thus, the MAI on a flat channel is reduced when $b_k[-1] = b_k[0]$. However, for other combinations of data, the MAI may be larger. The properties for synchronization on a flat channel have become poorer as compared with m-sequences, since the periodic autocorrelation function is larger for $0 < p < N$. This is illustrated in Figs. 6 and 7 for

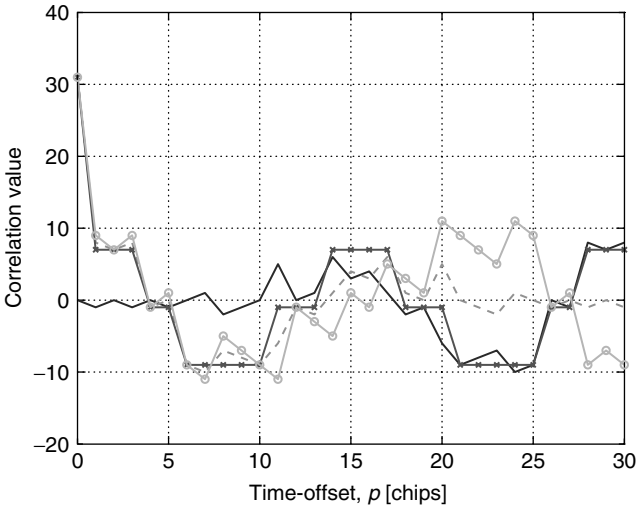


Figure 6. $\overline{X}_{1,1}(p)$ (solid line), $X_{1,1}(p)$ (dashed line), $\phi(p)$ (cross), and $X_{1,1}(p) - \overline{X}_{1,1}(p)$ (circle) for Gold sequence.

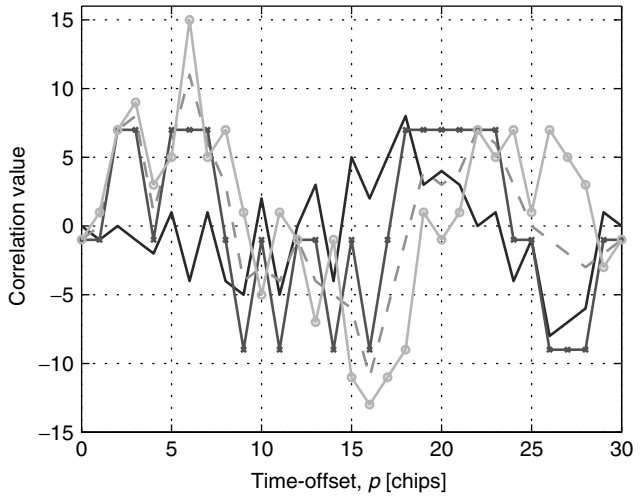


Figure 7. $\overline{X}_{2,1}(p)$ (solid line), $X_{2,1}(p)$ (dashed line), $\Phi_{2,1}(p)$ (cross), and $X_{2,1}(p) - \overline{X}_{2,1}(p)$ (circle) for two Gold sequences.

Gold sequences with period 31. The preferred pairs used are those given in Table 4 for $m = 5$. Both m-sequences have been generated starting from all ones. Figure 6 shows $\overline{X}_{1,1}(p)$ (solid line), $X_{1,1}(p)$ (dashed line), $\phi(p)$ (cross), and $X_{1,1}(p) - \overline{X}_{1,1}(p)$ (circle) for the Gold sequence obtained by adding the two sequences with a shift of 2 chips on the one generated from characteristic polynomial 45. Figure 7 shows $\overline{X}_{2,1}(p)$ (solid line), $X_{2,1}(p)$ (dashed line), $\Phi_{2,1}(p)$ (cross), and $X_{2,1}(p) - \overline{X}_{2,1}(p)$ (circle) where user two uses the sequence above and user 1 the sequence obtained by shifting 15 chips instead of 2. Both figures verify the three valued behavior of $\phi(p)$ and $\Phi_{k,i}(p)$ as discussed previously (curves shown with cross). However, in both cases, it is also clear that larger MAI may occur when $b_2[-1] = -b_2[0]$ (curves with circle).

Again, much less is known about the correlation properties when long codes based on Gold sequences are used and in general on multipath channels. In the Wideband CDMA (WCDMA) system used for third-generation mobile communications, Gold sequences are used to distinguish cells in the downlink and to distinguish mobiles in the uplink when simple receivers are used in the base station [4,5]. In both cases, long codes are used. The Gold sequence used in the downlink has period $2^{18} - 1$, while the Gold sequence used in the uplink has period $2^{41} - 1$. In both cases, only 38,400 chips are used to form a periodic sequence with period 38,400. The spreading factor is variable between 4 and 512.

3.3. Kasami Sequences

The *small set* of Kasami sequences can be obtained in a way similar to the Gold sequences [11,13,19]. Again, two m-sequences are added, but in this case, these two m-sequences have different periods. To obtain a period of $P = 2^m - 1$, with m even, of the Kasami sequence, one starts with an m-sequence with the same period. This m-sequence is then decimated by $2^{m/2} + 1$, which results in another m-sequence of period $2^{m/2} - 1$. By adding these two sequences, a Kasami sequence is obtained. Yet more Kasami sequences can be obtained by adding the original m-sequence with the other $2^{m/2} - 2$ shifts of the decimated sequence. By also including the original m-sequence in the set, $2^{m/2}$ Kasami sequences with period $2^m - 1$ have been obtained. It turns out that another way to generate all these sequences is by using the characteristic polynomial $G(D)G'(D)$ in (23), where $G(D)$ is the characteristic polynomial of the original m-sequence and $G'(D)$ is the characteristic polynomial of the decimated m-sequence.

The periodic discrete autocorrelation and cross-correlation is also three-valued for the *small set* of Kasami sequences. The values are from the set $\{-1, -(2^{m/2} + 1), 2^{m/2} - 1\}$. The *small set* of Kasami sequences therefore satisfies the Welch lower bound [11,20], which states that the maximum cross-correlation $\Phi_{\max}$ between any two sequences in a set of M sequences is bounded as

$$\Phi_{\max} \geq P\sqrt{\frac{M-1}{MP-1}} \approx \sqrt{P} \tag{28}$$

where the approximation is valid for large values of P and M. This set of sequences is therefore considered as optimal.

The *large set* of Kasami sequences with period $P = 2^m - 1$, with m even, contains both the Gold sequences and the *small set* of Kasami sequences and is obtained in the following way. Three different sequences are added. One is an m-sequence of period P. The other two are obtained by decimating the original m-sequence by $2^{m/2} + 1$ and $2^{(m+2)/2} + 1$. The two can be used in any possible shift. The number of sequences obtained are $2^{3m/2}$ when $m = 0 \bmod 2$ and $2^{3m/2} + 2^{m/2}$ when $m = 2 \bmod 2$. All the values of the periodic autocorrelation and cross-correlation are from the set $\{-1, -1 \pm 2^{m/2}, -1 \pm 2^{m/2+1}\}$. The Welch bound is not asymptotically approached with this larger set, but the packing of signal space is more efficient than for the Gold sequences. This set can be generated directly by using the characteristic polynomial $G(D)G'(D)G''(D)$, where $G(D)$ is the characteristic polynomial of the original m-sequence, $G'(D)$ is the characteristic polynomial of the first decimated m-sequence, and $G''(D)$ is the characteristic polynomial of the second decimated m-sequence.

However, still little is known about the performance of the sets of Kasami sequences when they are used as short or long signature sequences on multipath channels in CDMA. The reason, as before, is that it is not only the periodic autocorrelation and periodic cross-correlation between the full period of the signature sequences that matters.

3.4. Walsh–Hadamard Sequences

Two orthogonal signature sequences (on a flat channel) over the full period cannot be obtained by any of the sequence families discussed previously since $\Phi_{k,i}(p) \neq 0$. Complete orthogonality would make MAI to disappear and is therefore interesting. In the introduction, we indicated that under certain restrictions, it is in fact possible to obtain complete orthogonality. The family of orthogonal sequences is referred to as Walsh–Hadamard sequences

(sometimes Hadamard codes or Walsh functions). A Hadamard matrix of length $P = 2^m$ is defined as

$$\mathbf{H}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\mathbf{H}_P = \begin{bmatrix} \mathbf{H}_{P/2} & \mathbf{H}_{P/2} \\ \mathbf{H}_{P/2} & \overline{\mathbf{H}}_{P/2} \end{bmatrix} \tag{29}$$

where $\overline{\boldsymbol{H}}_{P/2}$ denotes the complement of $\mathbf{H}_{P/2}$. The rows of such a matrix are orthogonal. Thus, a user can form its signature sequence as a periodic repetition of one row in the Hadamard matrix of length $P = N$. If another user forms its signature sequence in the same way, but based on another row in the same matrix, and these two users are synchronized such that the periods of the signature sequences completely overlap, the periodic cross-correlation $\Phi_{k,i}(0)$ becomes zero. These two users will therefore not interfere with each other on a channel without multipath propagation. However, $\Phi_{k,i}(p)$ and $\phi(p)$ for $0 < p < N$ are in general not zero and are in many cases quite large (can in fact be as large as $N - 1$). This means that the system must be synchronized and that Walsh-Hadamard sequences are not good for synchronization purposes. The orthogonality is also lost on multipath channels.

Signature sequences based on rows or parts of rows of a Hadamard matrix can also be used when different users have different spreading factors as in WCDMA [4,5]. From Eq. (29), it is seen that each Hadamard matrix can in fact be decomposed into four Hadamard matrices of half the size. Since the rows in Hadamard matrices of all sizes have orthogonal rows, it is possible to allow a user with spreading factor $P/2$ to form its signature sequence as a repetition of one of the rows in $\mathbf{H}_{P/2}$. To keep orthogonality between all signature sequences also over the period $P/2$, users with spreading factor P are now restricted to use one of the remaining rows not starting with the sequence of length $P/2$ already used. This can be generalized to any spreading factor between 2 and P, when the Hadamard matrix of length P is used.

In WCDMA, the Hadamard matrix of length 256 is used to allocate signature waveforms to different channels in a base station or in a mobile (referred to as channelization codes in WCDMA). This matrix can be used for spreading factors from 4 to 256 (only powers of two are used) and the orthogonality is always preserved as long as the channel is flat. A user with spreading factor 4, is allocated the first 4 bits on one row in the matrix. It forms its signature sequence by periodically repeating these four bits. Since 1/4 of the rows in the Hadamard matrix start with the same 4 bits, these cannot be used for other users, because this would not make them orthogonal over a 4 chip period. A user with spreading factor 8 can be allocated the 8 first bits on the one of the remaining (allowed) rows for its signature sequence. This means that another 1/8 of all the rows are not allowed for the rest of the users. This scheme can now be continued as long as there are rows available to allocate. These sequences are referred to as Orthogonal Variable Spreading Factor (OVSF) sequences in WCDMA.

Walsh–Hadamard sequences are also use in IS-95 CDMA [16]. In the downlink, different channels are separated by different Walsh–Hadamard sequences of length 64. Walsh–Hadamard sequences are also used in the uplink, not as signature sequences but to obtain orthogonal 64-level modulation.

4. SOME CONCLUDING REMARKS ON SIGNATURE SEQUENCES

Although we have not covered all known signature sequences, it is clear that families of signature sequences with $\overline{X}_{k,i}(p) = 0$ except for $p = 0$ and $k = i$ do not exist. This means that there will be interference (the sum of ISI, MAI, and receiver noise) in the decision variable of each user. The choice of signature waveforms can to a certain extent reduce the interference, but in practice channel coding is also used in the system to reduce the effects of the interference. Since most known channel codes are designed for independent errors, it is important that interference be white (or almost white), such that the interference in the decision variable is independent from one symbol interval to the next. Furthermore, the power of the total interference should be small. The average interference power (AIP) for user 1 in the output at time T of a filter matched to $h(t)$ is given by

$$\begin{aligned} \mathrm{AIP}_1(T) &= \mathrm{E}\,[y_1(T) - \sqrt{E_1}g_1(1)b_1[0]]^2 \\ &= \sum_{l=2}^{L} E_1 \mathrm{E}\,|g_1(l)|^2 \left(\overline{X}_{1,1}^2(p_1(l))/N^2 + X_{1,1}^2(p_1(l))/N^2\right) \\ &\quad + \sum_{k=2}^{K}\sum_{l=1}^{L} E_k\,\mathrm{E}\,|g_k(l)|^2 \left(\overline{X}_{k,1}^2(p_k(l))/N^2 \right. \\ &\quad \left. + X_{k,1}^2(p_k(l))/N^2\right) \end{aligned} \tag{30}$$

where the same assumptions as in Eq. (17) about the channels have been used. Here, we have used the fact that bits from different users and different times are independent and hence the expected value of the cross-terms are zero. This average depends on the channel (both the complex coefficients and the relative delays) so in order for signature sequences to be good, they must lead to low AIP for all reasonable channels. The average also depends on the received energy from the different users through E_k and $|g_k(l)|^2$. These can to a certain extent be reduced by power control, such that a nearby user is not received at much higher power than a distant user. Results exist that seem to show that all the different families of signature sequences discussed lead to approximately the same AIP with practical channels in asynchronous DS-CDMA [21]. There also is some evidence that Gold sequences lead to reasonably good performance on land mobile radio channels [22].

With short signature sequences, the correlation time of the interference only depends on the correlation time of the channel coefficients. Thus, a large interference value due to high cross-correlation between two users, may remain for a significant time, and this makes channel coding less efficient. Long signature sequences are a means to reduce the correlation between the interference in neighboring symbol intervals, since different segments of the signature

sequences are used in neighboring symbol intervals, leading to different cross-correlations. This significantly improves the performance of channel coding.

There also exist many different receivers that may be used with DS-CDMA [23–25]. Many of these attempt to remove some or all interference before the decision variable is formed. For such receivers, the actual choice of signature sequences may be somewhat less important. The processes of mapping data symbols $b_k(t)$ to a transmitted waveform in DS-CDMA can also be described as repetition coding followed by scrambling, where the scrambling sequences plays the role of the signature sequence above [26]. In such a system, the outer channel coding and the repetition coding can be combined into one encoding process and interleaving can be done either on symbols or on chips. For all these different kind of systems, signature sequences should be designed jointly with several other things like channel coding, interleaving scheme, detection algorithm, and so on. For such more elaborate schemes, it remains an open issue how to design signature sequences and what the actual performance of the system will be on different channels.

BIOGRAPHIES

Tony Ottosson received the M.Sc. in electrical engineering from Chalmers University of Technology, Göteborg, Sweden, in 1993, and the Lic. Eng. and Ph.D. degrees from the Department of Information Theory, Chalmers University of Technology, in 1995 and 1997, respectively. Currently, he is an associate professor at the Communication Systems Group, Department of Signals and Systems, Chalmers University of Technology. During 1999 he was also working as a Research Consultant at Ericsson Inc., Research Triangle Park, North Carolina.

Professor Ottosson's research interests are in communication systems and information theory and are targeted mainly to CDMA systems. Specific topics are modulation, coding, multirate schemes, multiuser detection, combined source-channel coding, joint decoding techniques, and synchronization.

Erik G. Ström received his M.Sc. in electrical engineering in 1990 from the Royal Institute of Technology (KTH), Stockholm, Sweden, and Ph.D. degree in electrical engineering from the University of Florida, Gainesville, in 1994. He joined the Department of Signals, Sensors, and Systems at KTH as a postdoc in 1995 and was appointed assistant professor (forskarassistent) in 1996. Later that year, Ström joined Chalmers University of Technology, Göteborg, Sweden, where he is now an associate professor (högskolelektor/docent). He received the Chalmers Teacher's Prize in 1998. Since 1990 Dr. Ström acted as a consultant for the Educational Group for Individual Development, Stockholm, Sweden. He is a contributing author and associate editor for the Royal Admiralty Publishers' FesGas-series. Ström is a member of the board of the IEEE VT/COM chapter of the Swedish section and was a coguest editor for the special issue IEEE *Journal on Selected Areas in Communications* on "Signal Synchronization in Digital Transmission Systems." His research interests include code-division multiple access, synchronization, and wireless communications. He has published approximately 40 journal and conference papers.

Arne Svensson received the M.Sc. degree in electrical engineering in 1979, and the Dr. Ing. (Tekniska Licentiat) and the Dr. Techn. (Ph.D.) in telecommunication theory in 1982 and 1984, respectively, from University of Lund, Lund, Sweden. He joined Ericsson in 1987 as a research engineer and became later a specialist in communications. At Ericsson he worked on the design and analysis of a communication systems for the Swedish air force and the personal digital cellular system for mobile communications in Japan. Since 1993, he has been professor in communication systems at Chalmers University of Technology, Göteborg, Sweden, where he has been working on design and analysis of air interfaces for mobile communications. Dr. Svensson has published more than 150 papers in international journals and conference proceedings, and he is a fellow of IEEE. In 1986, he received the paper of the year award from the IEEE Vehicular Technology Society. His areas of interest include channel coding/decoding, digital modulation methods, channel estimation, data detection, multiuser detection, digital satellite systems, wireless IP-based systems, CDMA and spread spectrum systems, personal communication networks, and ultra wideband systems.

BIBLIOGRAPHY

1. T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 2002.
2. M. D. Yacoub, *Foundations of Mobile Radio Engineering*, CRC Press, Boca Raton, Fl, 1993.
3. G. L. Stüber, *Principles of Mobile Communication*, 2nd ed., Kluwer, Boston, 2001.
4. H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, Wiley, New York, 2000.
5. T. Ojanperä and R. Prasad, *Wideband CDMA for Third Generation Mobile Communications*, Artech House, Boston, 1998.
6. M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook*, revised edition, McGraw-Hill, New York, 1994.
7. R. C. Dixon, *Spread Spectrum Systems with Commercial Applications*, 3rd ed., Wiley, New York, 1994.
8. R. L. Peterson, R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
9. D. V. Sarwate, Optimum PN sequences for CDMA systems. In *Proc. IEEE Third International Symposium on Spread Spectrum Techniques and Applications*, 27–35, Oulu, Finland, July 1994.
10. D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudorandom and related sequences, *IEEE Proceedings* **68**(5): 593–619 (May 1980).
11. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.

12. S. W. Golomb, *Shift Register Sequences*, revised edition, Aegean Park Press, Laguna Hills, CA, 1982.
13. E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* **36**(9): 48–54 (Sept. 1998).
14. P. Fan and M. Darnell, *Sequence Design for Communication Applications*, UK Research Studies Press, 1996.
15. W. Stahnke, Primitive binary polynomial, *Math. Comp.* **27**: 977–980 (Oct. 1976).
16. S. C. Yang, *CDMA RF System Engineering*, Artech House, Boston, 1998.
17. R. Gold, Optimal binary sequences for spread spectrum multiplexing, *IEEE Trans. Inform. Theory* **IT-13**: 619–621 (Oct. 1967).
18. R. Gold, Maximal recursive sequences with 3-valued recursive cross correlation functions, *IEEE Trans. Inform. Theory*, **IT-14**: 154–156 (Jan. 1968).
19. T. Kasami, Weight distribution formula for some class of cyclic codes, Technical Report R-285, Coordinated Science Laboratory, University of Illinois, Urbana, IL, April 1966.
20. L. R. Welch, Lower bounds on the maximum cross correlation of signals, *IEEE Trans. Inform. Theory* **IT-20**: 397–399 (1974).
21. K. H.A Kärkkäinen and P. A. Leppänen, Comparison of the performance of some linear spreading code families for asynchronous DS/SSMA systems, *Proc. IEEE Military Communications Conference*, 784–790, November 1991.
22. H. Elders-Boll, The optimization of spreading sequences for CDMA system in the presence of frequency-selective fading, *Proc. IEEE 6th Symp. on Spread Spectrum Techniques and Applications*, 414–418, New Jersey Institute of Technology, New Jersey, USA, September 2000.
23. S. Verdú, *Multiuser Detection*, Cambridge University Press, UK, 1998.
24. S. Moshavi, Multi-user detection for DS-CDMA communications, *IEEE Commun. Mag.* **34**(10): 124–136 (Oct. 1996).
25. A. Duel-Hallen, J. Holtzman, and Z. Zvonar, Multiuser detection for CDMA systems, *IEEE Personal Commun. Mag.* **2**(2): 46–58 (April 1995).
26. P. Frenger, P. Orten, and T. Ottosson, Code-spread CDMA using maximum free distance low-rate convolutional codes, *IEEE Trans. Commun.* **48**(1): 135–144 (Jan. 2000).

SIMULATION OF COMMUNICATION SYSTEMS

K. SAM SHANMUGAN
University of Kansas
Lawrence, Kansas

1. INTRODUCTION

Modeling and simulation of communication systems can be viewed in a hierarchical manner starting at the network layer and progressing down to transmission systems level and then on to implementation details. The performance issues and tradeoffs addressed in each layer, and the modeling and simulation methods and the tools used at the various layers differ significantly. At the network layer, the simulation model will consist of processors, routers, traffic sources, buffers, transmission links, network topology, and protocols. The flow of packets and messages over the network will be simulated using an event-driven simulation framework, and system performance metrics such as network throughput, latency, resource utilization, and quality of service will be estimated from simulations. On the other hand, at the bottom level in the hierarchy dealing with implementation details, simulation of digital hardware, for example, will be done using hardware description language (HDL) simulators at the gate level. Performance metrics and design tradeoffs at this level may include power, speed, and chip area.

The focus of this article is on waveform level simulation of transmission systems or communication links, an example of which is shown in Fig. 1.

The primary simulation technique used at the link level is time-driven Monte Carlo simulation of the flow of waveforms or signals over the transmission link (Refs. 1–3 cover this topic in great detail). Waveform-level simulation of communication systems involves the following steps:

1. Modeling the communication system in a block diagram form in which each functional block performs a specific signal processing operation such as modulation, and filtering
2. Generating sampled values of the input signals, noise, and interference
3. Letting the functional blocks in the simulation model operate on the input samples
4. Gathering the samples generated during the simulation and estimating performance measures such as bit error rates as a function of E_b/N_0 and other design parameters

Details of these steps are presented in the following sections.

2. DISCRETE TIME REPRESENTATION OF SIGNALS AND SYSTEMS

In the simulation domain, all systems and signals are represented by their discrete-time equivalents. The simulation models and simulation algorithms draw heavily from DSP (digital signal processing) concepts [4,5]. The fundamental concept that permits us to go

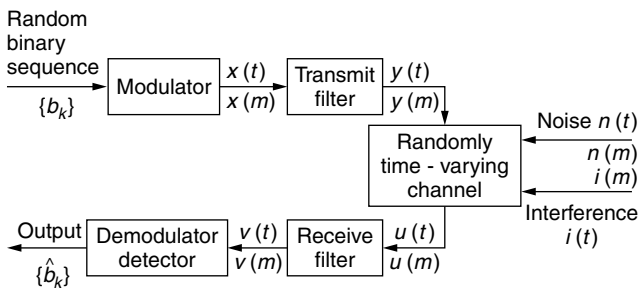


Figure 1. Waveform level simulation model of a communication system.

back and forth between discrete- and continuous-time representation of signals and systems is the sampling theorem. Implications of the sampling theorem as it applies to waveform-level simulation of communication systems are presented below.

2.1. Discrete-Time Representation of Lowpass Signals and Systems

The uniform sampling theorem in its simplest form states that a deterministic signal that is lowpass and band-limited to B Hz in the continuous-time domain can be represented exactly, *without loss of information*, in terms of its sampled values as long as the sampling rate f_s is greater than the *Nyquist rate* of $2B$ samples per second [4,5]. The relationship between the continuous time signal $x(t)$ and its sampled values $x(kT_s)$, where T_s is the time between samples, is given by the following equations:

$$x_s(t) = \sum_{k=-\infty}^{\infty} x(kT_s)\delta(t - kT_s) \tag{1}$$

$$X_s(f) = f_s \sum_{k=-\infty}^{\infty} X(f - kf_s) \tag{2}$$

$$x(t) = x_s(t)^* \frac{\sin(\pi f_s t)}{\pi f_s t} = \sum_{k=-\infty}^{\infty} x(kT_s) \frac{\sin(\pi f_s(t - kT_s))}{(\pi f_s(t - kT_s))} \tag{3}$$

Equation (3) represents an interpolation formula in the time domain that yields the values of $x(t)$ for *all* values of t in terms of the sampled values $x(kT_s)$. This interpolation operation in the time domain is equivalent to filtering of $x_s(t)$ in the frequency domain using an ideal lowpass filter with a bandwidth of $f_s/2$ Hz. If the signal is not strictly band-limited and/or the sampling rate is less than $2B$, then it will not be possible to reconstruct $x(t)$ exactly from $x_s(t)$ due to aliasing, which is a result of spectral overlap that occurs in the frequency domain [see Eq. (2)].

The lowpass sampling theorem applies to the discrete-time representation of both lowpass *signals* and lowpass *systems*. The impulse response $h(t)$ of a lowpass system in the continuous-time domain can also be represented in the discrete-time domain using the sampling theorem.

2.2. Sampling of Bandpass (Deterministic) Signals and Systems

Both lowpass and bandpass signals and components are usually present in communication systems. Information-bearing signals are usually lowpass in nature prior to modulation, after which they become bandpass in nature. Components such as filters can be lowpass or bandpass. Hence we need to represent both lowpass and bandpass signals and systems in the discrete time domain for simulation.

Bandpass signals and systems can be sampled directly using the bandpass version of the sampling theorem, or they can be sampled using the principle of the lowpass sampling theorem and the concept of the lowpass equivalent representation [1]. In the time domain, modulated signals can be expressed in the form

$$x(t) = x_c(t)\cos(2\pi f_c t) - x_s(t)\sin(2\pi f_c t) \tag{4}$$

$$= \text{Re } al\{\tilde{x}(t)\exp(j2\pi f_c t)\} \tag{5}$$

$$\tilde{x}(t) = x_c(t) + jx_s(t) \tag{6}$$

where $x_c(t)$ and $x_s(t)$ are the in-phase and quadrature phase modulating signals, which are usually lowpass; f_c is the carrier frequency; and $\tilde{x}(t)$ is the *complex envelope* of the bandpass signal. $\tilde{x}(t)$ will be lowpass since the modulating signals $x_c(t)$ and $x_s(t)$ are lowpass. Since the bandwidth of the modulating signals will be small compared to f_c, fewer samples will be needed to represent the lowpass complex envelope $\tilde{x}(t)$, which contains all the information in the modulated bandpass signal. Equation (5) clearly shows that the bandpass signal can be reconstructed from $\tilde{x}(t)$ by simply multiplying it with the complex carrier and taking the real part.

The lowpass equivalent representation of a deterministic bandpass signal can also be derived in the frequency domain using the Hilbert transform. When the bandwidth of the signal is small compared to the carrier frequency, the lowpass equivalent representation in the frequency domain is given by [1]

$$\begin{aligned} X_{\text{LP}}(f) &= \{2X_{\text{BP}}(f + f_c)\}_{\text{lowpass}} \\ &= 2X_{\text{BP}}(f + f_c)U(f + f_c) \end{aligned} \tag{7}$$

where $U(f)$ is the unit step function, $U(f) = 1$ for $f > 0$ and zero for $f < 0$.

Note that the lowpass equivalent can be asymmetric around $f = 0$ if the bandpass spectrum is not symmetric around the carrier frequency. This leads to a complex time-domain function as shown in Eq. (6). The minimum sampling rate for the lowpass equivalent representation is B *complex* samples per second or $2B$ real samples per second.

It should be noted that the complex envelope representation is with respect to a single-carrier frequency f_c. This representation can also be used for simulating multicarrier FDM (frequency-division multiplex) systems by using one of the mid-band carriers as the reference and representing the total composite complex envelope of the sum of the FDM carries with respect to the reference carrier according to

$$\sum_{i=1}^{n} A_i e^{j2\pi f_i t + j\phi_i} = e^{j2\pi f_c t} \sum_{i=1}^{n} A_i e^{j2\pi (f_i - f_c)t + j\phi_i} \tag{8}$$

where f_c is the carrier frequency chosen as the reference.

If multiple signals with widely differing bandwidths are present in a system, then a multirate sampling strategy will be useful for simulating such systems. Each signal is sampled at a rate consistent with its bandwidth, and interpolation and decimation techniques are used to upconvert or downconvert the sampling rates as necessary.

2.3. Sampling of Lowpass and Bandpass Random Processes

Signals, noise, and interference in communication systems are usually modeled as stationary random processes

characterized in the frequency domain by power spectral density functions. Frequency domain parameters such as bandwidth, and properties such as lowpass and bandpass are based on power spectral densities (PSDs). The sampling principle and the concept of lowpass equivalent representation also apply to random processes in terms of their PSDs [6].

A lowpass random process in the continuous-time domain can be represented in the discrete-time domain in terms of its sampled values, and it is possible to recover the continuous-time random signal (with a mean-squared error approaching zero) from its sampled values as long as the process being sampled is band-limited to B Hz and the sampling rate is greater than $2B$. The concept of lowpass equivalent representation applies for bandpass random processes also. For example, a bandpass Gaussian process $n(t)$ can be represented in terms of its lowpass equivalent as

$$\begin{aligned} n(t) &= n_c(t)\cos(2\pi f_c t) - n_s(t)\sin(2\pi f_c t) \\ &= \text{Real}\,\{\tilde{n}(t)\exp(j2\pi f_c t)\}, \tilde{n}(t) = n_c(t) + jn_s(t) \end{aligned} \tag{9}$$

where $n_c(t)$ and $n_s(t)$ are real-valued lowpass Gaussian random processes with the power spectral densities

$$\begin{aligned} S_{n_c n_c}(f) = S_{n_s n_s}(f) &= S_{NN}(f+f_c)U(f+f_c) \\ &\quad + S_{NN}(-f+f_c)U(-f+f_c) \\ jS_{n_s n_c}(f) &= S_{NN}(f+f_c)U(f+f_c) \\ &\quad - S_{NN}(-f+f_c)U(-f+f_c) \end{aligned} \tag{10}$$

If the bandpass process is nonsymmetric around the carrier frequency, then $n_c(t)$ and $n_s(t)$ will be correlated with the cross-PSD given above. For simulation purposes, bandpass random processes are sampled using their lowpass equivalent representations.

2.4. Simulation of Bandpass Systems with Bandpass Inputs

In order to minimize the computational burden, bandpass systems with bandpass inputs are simulated using sampled values of lowpass equivalent representations. It should be noted that while the lowpass equivalent representation of bandpass *signals* in the frequency domain contains a factor of 2 in Eq. (7), the lowpass equivalent representation of bandpass *systems* does not include the factor of 2.

The input–output relationship for the bandpass and lowpass equivalent representations are given by

$$\tilde{y}_{\text{LP}}(t) = \tilde{h}_{\text{LP}}(t)^* \tilde{x}_{\text{LP}}(t); \quad y_{\text{BP}}(t) = \text{Real}\,\{\tilde{y}_{\text{LP}}(t)e^{j2\pi f_c t}\} \tag{11}$$

2.5. Factors Influencing the Sampling Rate

The most important factor in determining an appropriate sampling rate for simulations is the amount of aliasing that can be tolerated. Other factors that have to be considered in setting the sampling rate for simulations include the effect of sampling on modeling functional blocks such as filters (frequency warping due to bilinear z transform), nonlinearities (bandwidth expansion), and feedback loops (delay in feedback loops). All the deleterious effects of sampling on simulation accuracy can be minimized by increasing the sampling rate. However, increasing the sampling rate will increase the computational burden.

This tradeoff between sampling rate and the accuracy of simulations is a very important one in simulations. A practical value for sampling rate that offers a good tradeoff between simulation accuracy and computational burden is 16–32 samples per hertz or symbol [1]. It is most convenient to use an integer number of samples per symbol for simulating digital transmission systems, and it is computationally most efficient to choose 16 or 32 samples per hertz so that the fastest version of the discrete Fourier transform (DFT) algorithm can be used during simulations.

3. MODELING AND SIMULATION OF FUNCTIONAL BLOCKS IN COMMUNICATION SYSTEMS

At the waveform level simulation of a communication system, each functional block in the simulation model performs a specific signal processing operation. If the signal processing operation performed by the functional block is discrete time, algorithmic, and at the symbol level (e.g., convolutional encoder/decoder), then there is very little modeling per se for such functional blocks: the simulation model of the functional block is the algorithm itself (e.g., a Viterbi decoder). On the other hand, there are a number of other functional blocks that perform (analog) waveform processing operations such as filtering and amplification. The signal processing operations performed by these blocks as well as the communication channel have to be *modeled* for simulation purposes. Examples of such *models*, which are described below, include infinite and finite impulse response filters and AM-to-AM and AM-to-PM models for memoryless nonlinearities.

3.1. Modeling and Simulation of Linear Time-Invariant (LTIV) Components

Many components in communication systems such as filters, optical fibers, cables, and other guided channels are linear and time-invariant. Such components are described in the time domain in terms of the impulse responses $h(t)$ or transfer functions $H(f)$. If these components are bandpass in nature, then for simulation purposes we will use the lowpass equivalent representations described in Section 2. The DSP literature contains a wide array of algorithms for implementing and virgule or simulating filters [4,5]. (We will use the generic term *filters* to represent LTIV components). The choice of simulation algorithm will depend on the nature of the filter specifications, the duration of the impulse response, and the context in which the filter is used in the overall simulation model.

3.1.1. Finite-Impulse Response (FIR) Model — Time-Domain Convolution. If the filter specification is given empirically in terms of the sampled values of the lowpass equivalent impulse response $\tilde{h}(kT_s)$, $k = 0, 1, 2, \ldots, M-1$, then the simplest simulation model is an FIR structure

that implements the input–output relationship as a finite convolution sum of the form

$$\tilde{y}(nT_s) = T_s \sum_{k=0}^{M-1} \tilde{h}(kT_s)\tilde{x}((n-k)T_s) \tag{12}$$

Simulation of this equation requires M complex multiplications and additions for each output sample. This may impose a considerable computational burden if the impulse response is very long. In order to reduce the processing load, impulses responses are truncated to the shortest possible duration without losing a significant amount of energy outside the truncation window and the truncated impulse response is used in Eq. (12). If the filter is specified in the frequency domain in terms of sampled values of the frequency response $H(kf_c)$, the inverse Fourier transform is used to compute the impulse response, which is then truncated and used in Equation (12) to perform time-domain convolutions.

3.1.2. FIR Model — DFT Implementation. The computational burden of time-domain convolution can be reduced by using DFT operators. In this implementation, the input samples and the impulse response are *padded* with enough zeroes and the padded input vector and the impulse response vectors are convolved using the DFTs operators according to

$$Y(kf_0) = \text{DFT}\ (\tilde{y}(kT_s)) = \text{DFT}\ (\tilde{h}(kT_s))\text{DFT}\ (\tilde{x}(kT_s)) \tag{13}$$

$$\tilde{y}(kT_s) = \text{Inv.DFT}\ (Y(kf_0)) \tag{14}$$

The minimum DFT size n is usually chosen to be a power of 2 nearest to the sum of the length of the unpadded impulse response and the input vectors. This will permit the use of fast FFT operators to perform the convolution [4,5].

DFT/FFT implementation will be computationally very efficient (several orders of magnitude faster) compared to direct-time-domain convolution when the length of the impulse response exceeds about 100 samples. One drawback of the DFT/FFT filters is that they introduce a processing delay of n samples (e.g., the output lags the input by n samples), which might cause a problem if the filter being simulated is part of a feedback loop, for example. The one block processing delay, which is strictly an artifact of the DFT/FFT filter, might render the feedback loop unstable and lead to totally incorrect simulation results. If all the blocks in the simulation model are serially cascaded, then this is not a problem.

If the input sequence is very long, then it can be broken up into smaller blocks and the response of the filter to each input block can be computed separately and the results can be added using the superposition principle. Either the input blocks or the output blocks have to be overlapped in order to produce the correct output [1,5,6].

3.1.3. Infinite Impulse Response (IIR) Models. If the filter specifications are given in the form of poles and zeroes or as a ratio of polynomial in the s domain, then the filter can be simulated using recursive computational structures that are very efficient. These recursive structures are derived using either *the impulse-invariant method or the bilinear z transform* [1,4,5]. In the first method the impulse response of the filter is obtained from the inverse transform of the transfer function and the z transform of the infinite-length impulse response (untruncated) is used to derive a recursive computational structure. In the second method, the transfer function in the Laplace transform domain is mapped into the z-transform domain directly using the bilinear z transform defined by

$$H_d(z) = H(s)@s = \frac{2}{T_s}\left[\frac{1-z^{-1}}{1+z^{-1}}\right] \tag{15}$$

The resulting transfer function in both cases can be factored into a product of quadratic factors:

$$H_d(z) = \alpha_0 \prod_{i=1}^{K} H_d^{(i)}(z); \qquad K = \frac{1}{2}(N+1);$$

$$H_d^{(i)}(z) = \frac{1+\alpha_{1i}z^{-1}+\alpha_{2i}z^{-2}}{1+\beta_{1i}z^{-1}+\beta_{2i}z^{-2}} \tag{16}$$

This leads to a simulation structure shown in Fig. 2.

The main source of error in the impulse-invariant IIR simulation model is aliasing, whereas the bilinear z transform suffers from frequency warping introduced by the transform. Both of these effects can be minimized by choosing a sufficiently high sampling rate. While the IIR structures are computationally faster compared to the FIR models, the IIR structures can be easily derived only for filters whose transfer functions are specified in analytical form by a transfer function in the Laplace or Fourier transform domains. The FIR filters can model and simulate filters with any arbitrary frequency or impulse responses specified empirically. Many important filters, such as the square-root raised-cosine filter, cannot be specified in the Laplace transform domain by poles and zeros and hence are not easy to simulate via the IIR methods.

3.2. Modeling and Simulation of Linear Time-Varying (LTV) Components

There are many components in a communication system that exhibit linear time-varying behavior. An important example of a LTV component is the mobile communication channel in which channel characteristics such as attenuation and delay change as a function of time due to relative motion between the transmit and receive antennas. The input–output relationship for LTV components can be expressed in the time domain by the convolution integral

$$\tilde{y}(t) = \int_{-\infty}^{\infty} \tilde{h}(\tau, t)\tilde{x}(t-\tau)d\tau \tag{17}$$

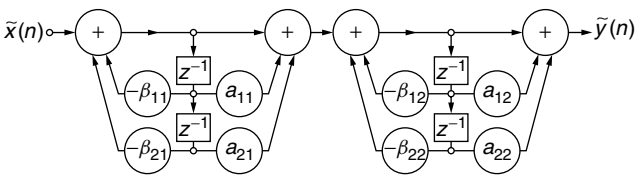


Figure 2. Recursive simulation structure for IIR models [1].

where $\tilde{h}(\tau, t)$ is the time-varying impulse response of the component measured at time t when the impulse is applied to the system input at $t - \tau$. If the input to the system is band-limited to W Hz, then we can derive an FIR model for the time-varying system via the sampling theorem as [7]

$$\begin{aligned}\tilde{y}(kT_s) &\approx \frac{1}{2W}\sum_{n=0}^{\infty}\tilde{h}\left(\frac{n}{2W}, kT_s\right)\tilde{x}\left(kT_s - \frac{n}{2W}\right)\\ &\approx \frac{1}{2W}\sum_{n=0}^{M}\tilde{h}\left(\frac{n}{2W}, kT_s\right)\tilde{x}\left(kT_s - \frac{n}{2W}\right)\\ &= \frac{1}{2W}\sum_{n=0}^{M}\tilde{g}_n(kT_s)\, x\left(kT_s - \frac{n}{2W}\right)\end{aligned} \tag{18}$$

where $\tilde{g}_n(kT_s)$ are called the *tap gain functions* that represent the time-varying impulse response of the system. For a time-*invariant* system, the tap gain functions will be constant and $\tilde{g}_n(kT_s) = \tilde{h}(nT_s)$. The FIR model given above can be implemented as a tapped delay line (TDL) model of the form shown in Fig. 3.

3.3. Modeling and Simulation of Nonlinear Components

Communication systems often contain nonlinear components. Sometimes a nonlinear component is placed in the system to improve performance. For example, a limiter is placed at the front end of a receiver that is subjected to impulsive noise. Also, devices such as high-power amplifiers exhibit an undesirable nonlinear behavior when the input power is high. This introduces nonlinear signal distortion, which might degrade the performance of the communication system significantly. The effects of nonlinearities on system performance are usually difficult to characterize by analytical means but are easy to simulate.

Nonlinearities in communication systems fall into two categories: (1) instantaneous nonlinearities such as limiters and (2) nonlinearities with memory, such as frequency-dependent wideband RF amplifiers. Nonlinearities are also sometimes classified according to whether the input/output signals are baseband or bandpass. Since bandpass nonlinearities are the most common type of nonlinear elements encountered in communication systems, we will concentrate on techniques for modeling and simulation of bandpass nonlinearities in this section.

3.3.1. Lowpass Equivalent Models for Memoryless Nonlinearities. Devices such as bandpass limiters and logarithmic amplifiers can be modeled using the complex envelope representation of the input and output signals. In the bandpass case the input–output relationship of a memoryless nonlinearity can be represented by

$$\begin{aligned}x(t) &= A(t)\cos(\omega_c t + \phi(t))\\ y(t) &= G[x(t)] = G[A\cos(\alpha)];\alpha = \omega_c t + \phi(t)\end{aligned} \tag{19}$$

where $G(\,)$ is a memoryless nonlinearity. If the bandwidth of the input signal is much smaller than f_c, then we can expand $y(t)$ as a Fourier series of the form

$$z = a_0 + \sum_{k=1}^{\infty}(a_k\cos k\alpha + b_k\sin k\alpha) \tag{20}$$

The output of the nonlinearity will contain spectral components in the vicinity of f_c as well as harmonic terms located in the vicinity of kf_c, $k > 1$. If $f_c \gg B$, then the harmonic terms will be far removed from the in-band terms and hence they can be ignored (these components can be easily removed by filtering in a real system). The in-band spectral components or the so-called *first-zone output components* correspond to $k = 1$ in the Fourier series expansion and the first-zone output of the nonlinearity can be expressed in the form

$$y(t) = a_1\cos[\omega_c t + \phi(t)] + b_1\sin[\omega_c t + \phi(t)] \tag{21}$$

where a_1 and b_1 are the Fourier series coefficients defined by

$$\begin{aligned}a_1 &= f_1(A) = \frac{1}{\pi}\int_0^{2\pi} G(A\cos\alpha)\cos\alpha\, d\alpha;\\ b_1 &= f_2(A) = \frac{1}{\pi}\int_0^{2\pi} G(A\cos\alpha)\sin\alpha\, d\alpha\end{aligned} \tag{22}$$

In terms $f_1(A)$ and $f_2(A)$ (which are called the first-order *Chebyshev transforms* of the nonlinearity), the complex lowpass equivalent simulation model for a memoryless nonlinearity is [1]

$$\tilde{y}(t) = f(A)\exp(j\phi + jg(A)) \tag{23}$$

$$f(A)e^{jg(A)} = f_1(A) - jf_2(A) \tag{24}$$

For nonlinearities such as soft and hard limiters, this model, in the form of $f(A)$ and $g(A)$ (which are also called the *AM-to-AM and AM-to-PM* transfer characteristics of

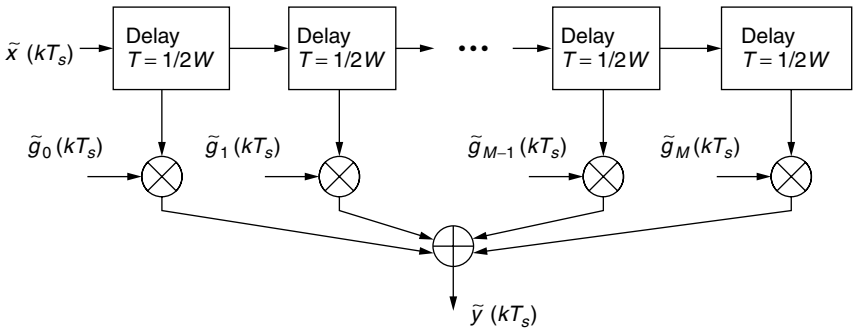


Figure 3. Tapped delay line model for a time-varying component.

the nonlinearity), can be derived in the closed form from the Fourier integrals given in Eq. (22) [1].

The AM–AM and AM–PM characteristics of devices such as high-power amplifiers are usually obtained from *swept-power measurements*, which are made with an input tone of the form $A\cos(\omega_c t + \phi(t))$. The input amplitude and hence the input power $A^2/2$ is varied in steps of 1 dB or so. The AM–AM characteristic is obtained from the input power–output power relationship and the AM–PM characteristic is obtained by measuring the phase offset between the input and output as a function of the input power level. Typical AM–AM and AM–PM characteristics are shown in Fig. 4. The AM–AM and AM–PM model can be simulated using either the empirical AM–AM and AM–PM data or the analytical approximation [8] such as the one shown in Fig. 4.

3.3.2. Lowpass Equivalent Models for Nonlinearities with Memory. When a nonlinear component operates over a wide bandwidth, it might exhibit a frequency selective nonlinear behavior. Models for nonlinearities with memory (or frequency-selective behavior) are difficult to derive analytically, but some useful models can be derived from swept-power/swept-frequency measurements. These measurements are made by probing the device with an unmodulated tone of the form $A\cos(2\pi(f_c t + f_i t))$ and changing both the input power levels and the frequency offset f_i and recording the AM–AM and AM–PM transfer characteristics at different frequencies. If these curves are significantly different at different frequencies, then a model of the form shown in Fig. 5 can be synthesized from the swept-frequency/swept-power measurements to account for the frequency selective behavior of the device. Details of the model synthesis may be found in papers by Saleh and Poza et al. [8,9]. Other simulation models for nonlinearities with memory in the form of Volterra series or nonlinear differential equations may be found in Ref. 1. (The best single source of reference for all the topics addressed in this article is Ref. 1.)

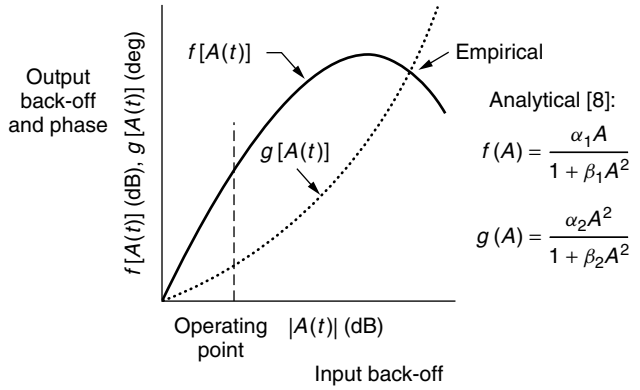


Figure 4. AM–AM and AM–PM characteristics.

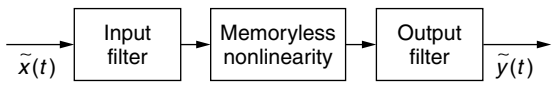


Figure 5. Simulation model for a frequency-selective nonlinearity.

3.4. Modeling and Simulation of Communication Channels

Communication channels introduce a variety of transmission impairments, including attenuation, linear distortion, noise, and interference. Simulation models for communication channels take one of two forms: (1) a transfer function model for time-invariant channels such as optical fibers and cables and (2) a TDL model for time-varying channels such as the mobile radio channel. The transfer function model of a time-invariant channel can be simulated using an FIR or IIR algorithm. Time-varying channels are more difficult to model and simulate. Some of the models and approaches that are used for simulating mobile radio communication channels and other time-varying channels are described below.

3.4.1. Simulation Models for Mobile Communication Channels. In a typical mobile communication environment there will be multiple propagation paths between the transmitter and the receiver due to reflection, refraction, and scattering [10,11]. Also, the path characteristics will be time-varying due to relative motion between the transmit and receive antennas (Fig. 6).

The input–output relationship for the two-ray multipath channel shown in Fig. 6 can be expressed as

$$\tilde{y}(t) = \tilde{a}_1(t)\tilde{x}(t - \tau_1(t)) + \tilde{a}_2(t)\tilde{x}(t - \tau_2(t)) \qquad (25)$$

where $\tau_1(t)$ and $\tau_2(t)$ are the path delays and $\tilde{a}_1(t)$ and $\tilde{a}_2(t)$ are the randomly time-varying complex attenuations of the two multipath components, which causes fluctuations in the received signal power. Changes in the received signal power as a function of time is called *fading*. The terms $\tilde{a}_1(t)$ and $\tilde{a}_2(t)$ are usually modeled as *uncorrelated and stationary* random processes.

Movement of the mobile unit over larger distances ($d \gg \lambda$, where λ is the wavelength), and changes in terrain features affect attenuation and received signal power slowly, and this phenomenon is called *large-scale (or slow) fading*. The received signal in each path in Fig. 6 is made up of a large number of scattered components, and hence the central-limit theorem leads to complex Gaussian process models for the complex attenuation and the complex envelope of the received signal for each path.

Movement over small distances on the order of $\lambda/2$ causes significant phase changes of the scattered components, resulting in rapid fluctuations in signal amplitude and power. This phenomenon is called *small-scale (fast) fading*. Large-scale fading impacts the link

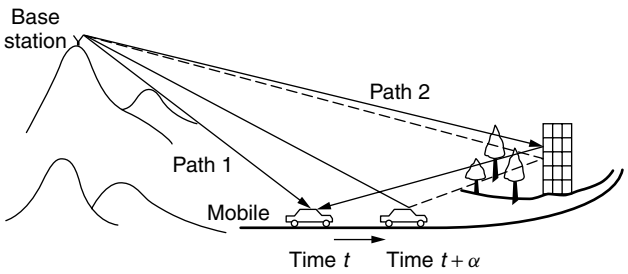


Figure 6. Mobile communication environment.

$S/N + I$ and coverage within a given area. Small-scale fading, on the other hand, impacts all the signal processing operations in the transmitter and receiver, and the effects of small-scale fading are simulated at the waveform level. The basic simulation model for small-scale fading is a TDL model of the form given in Fig. 3.

The tap gains are random processes representing the random variations in the channel characteristics such as the complex attenuation of each multipath component. These are usually modeled as complex Gaussian processes with zero mean and Rayleigh envelope distribution for dense urban environments, and nonzero mean and Rice envelope statistics for rural and suburban environments. The simulation model is specified in terms of the number of multipath components, relative delays, average power received in each path, and the power spectral density of the random process models that describe the random variations of the path characteristics (i.e., the tap gain functions). An example of the multipath model that is used for simulating the mobile radio environments for the design of the third-generation cellular systems is given in Table 1 [12].

In the simulation model, the tap gain functions are generated by filtering uncorrelated Gaussian random processes. The filter transfer function is carefully synthesized to produce the desired Doppler power spectral density. Details of the algorithms used for generating Gaussian sequences with a given power spectral density are discussed in Section 4.

3.4.2. Discrete-Channel Model. Whereas the TDL model is used to simulate the waveform-level distortions introduced by a multipath fading communication channel, a discrete Markov model is often used to characterize the burst errors introduced by fading channels [13–15]. These discrete channels models are computationally more efficient than the waveform-level simulation models for evaluating the performance of error control encoders/decoders, interleaves, and other devices. An example of a simple two-state Markov model for a fading channel is shown in Fig. 7.

Table 1. Example of a TDL Model for an Outdoor Mobile Radio Channel at 2 GHz

Tap Delay (ns)	Average Power (db)
0	0.0
244	−2.4
488	−6.5
732	−9.4
936	−12.7
1220	−13.3
1708	−15.4
1953	−25.4

In the Makov model, the channel is in one of the two states at the beginning of a symbol interval. If the channel is in good state, the transmitted symbol is error-free whereas the probability of transmission error is 0.5 when the channel is in the bad state. While the channel remains in the bad state, it produces bursts of transmission errors. The channel can transition from state i to state j at the end of a symbol interval with a probability of $p_{ij}, i, j = 1, 2$. The rate or the probability of transition between the good and bad states and the error generation probabilities can be derived analytically from the underlying fading channel model or estimated from error sequences obtained using waveform level simulations.

Markov models are also applicable to hard input and soft output channels. Details of the Baum–Welch algorithm used for estimating the structure and parameters of the Markov models and examples of discrete channel models may be found in the literature [16–18].

Simulation of the Markov model is carried out at the symbol rate. For simulating the flow of a symbol through the discrete-channel model, two uniform random numbers are drawn, one to determine the state of the channel (i.e., the transition) and a second random number to decide whether the transmitted symbol suffers a transmission error. Thus, the simulation of the discrete model is very efficient compared to having to generate sampled values of the transmitted waveform and process them through all the functional blocks in the waveform-level simulation model.

4. MONTE CARLO SIMULATION AND RANDOM-NUMBER GENERATION

With respect to the simulation model shown in Fig. 1, the inputs or stimuli that drive the simulation model are sampled values of the input signals, noise, and interference, which are modeled as random processes. Sampled values of random processes are random variables, and hence

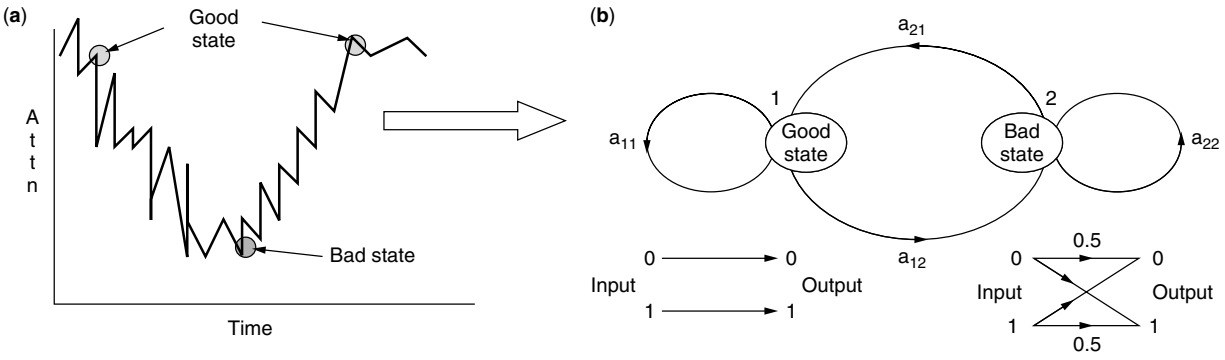


Figure 7. (**a**) Fading channel; (**b**) Markov model for a fading channel.

the input sequences that drive the simulation models are sequences of random numbers from arbitrary distributions and with arbitrary power spectral densities or autocorrelation functions. Hence Monte Carlo simulation involves the generation and processing of random numbers.

4.1. Uniform Random-Number Generator

A typical Monte Carlo simulation might entail the generation and processing of many thousand samples of random variables, and hence we need computationally efficient algorithms for generating random numbers. The starting point of random-number generation is the uniform random-number generator. An independent sequence of uniform random integers in the interval $[0, M-1]$ can be generated using the *linear congruential algorithm* [19–21]

$$X_{j+1} = (aX_j + c) \bmod M \quad \text{(integer arithmetic)} \tag{26}$$

This recursive algorithm is started using an initial random *seed* that is provided by the user and it produces a set of integers uniformly distributed in the interval $[0, M-1]$. A sequence of uniform random numbers in the interval [0,1) can be obtained according to $U_i = float(X_i/M)$.

The output sequence produced by the recursive algorithm given in Eq. (26) will be periodic with a maximum period of M. The values of a, c and M are carefully chosen such that the sequence is independent, is uniformly distributed, and has the maximum period. Two popular algorithms that produce uniform sequences with long periods are the Marsaglia–Zamann algorithm and the Wichmann–Hill algorithm [1].

Since the uniform random-number generators play such a central role in Monte Carlo simulation and random-number generation, they should be thoroughly tested for temporal and distributional properties. Statistical tests for these can be found in the literature [1,6].

4.2. Random-Number Generators for Arbitrary Distributions

An independent sequence of random numbers from arbitrary probability distributions can be generated by applying appropriate memoryless nonlinear transformations to an independent uniform sequence U_i [1]. It can be shown that U_i can be mapped to a sequence of random numbers X_i from an arbitrary distribution $F_X(x)$ according to

$$X_i = F_X^{-1}(U_i) \tag{27}$$

where F_X^{-1} is the inverse cumulative distribution function (CDF) of X. This method is called the *inverse transform method* of generating random numbers, and it can be easily applied to distributions that are analytically tractable. For example, if we want to produce a sequence of random numbers from an exponential probability density function (PDF), the inverse transform method yields the formula $X_i = (-1/\lambda)\ln(1-U_i)$.

If the CDF and/or the inverse of the CDF of X cannot be expressed in closed form, then the inverse transform method can be implemented in empirical form by quantizing the underlying PDF and creating a piecewise linear CDF and applying the inverse transform method empirically. The details of this approach are shown in Fig. 8.

4.3. Gaussian Random-Number Generators

Gaussian random processes are used to model signals, noise, and interference as well as fading in communication channels, and hence it is important to have computationally efficient algorithms for generating Gaussian random numbers. Two algorithms for generating Gaussian random numbers are [20,21]

1. The sum of a large number of uniform random numbers (usually 12), which leads to an approximately Gaussian distribution by virtue of the central-limit theorem
2. The Box–Mueller method, which uses the following algorithm

$$\begin{aligned} X_1 &= [-2\ln(U_1)]^{1/2}\cos 2\pi U_2 \\ X_2 &= [-2\ln(U_1)]^{1/2}\sin 2\pi U_2 \end{aligned} \tag{28}$$

where X_1, X_2 are two independent Gaussian samples derived from two independent uniform random numbers U_1, U_2.

The Box–Mueller method produces a better distribution than does the sum-of-12 method.

4.4. Correlated Gaussian Sequences

Correlated Gaussian sequences with a given power spectral density (PSD) or autocorrelation function can be generated by filtering an uncorrelated Gaussian sequence. The filter transfer function can be synthesized using a number of different approaches. For FIR implementation of the filter, the transfer function of the filter can be chosen according to

$$H(f) = \sqrt{(S_{YY}(f))} \tag{29}$$

where $S_{YY}(f)$ is the desired power spectral density. This method can be used for generating temporally correlated

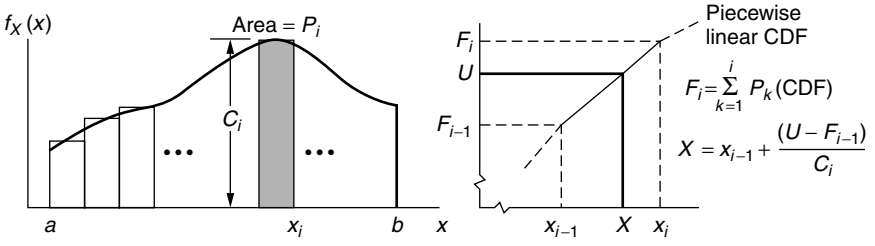


Figure 8. Empirical version of the inverse transform method.

Gaussian processes employed in modeling multipath fading communication channels. An IIR filter transfer function can be synthesized using auotregressive moving-average (ARMA) model. Details of this procedure may be found in the literature [1,6,22].

4.5. Binary and Nonbinary Sequences

The input symbol sequences used in the simulation of digital transmission systems can be generated by mapping the output of a uniform random-number generator into binary and M-ary sequences. Discrete-symbol sequences can also be generated using a linear feedback shift register arrangement in which the feedback tap weights are chosen to be the coefficients of primitive polynomials in Galois field (GF) (2^k), $2^k = M$. An example of this for the binary case is shown in Fig. 9.

These shift register sequences, which are also called *pseudonoise* (PN) *sequences*, have many desirable properties that are useful in the context of simulations. Two of the most important properties of the shift register sequences are that they have the maximum period ($2^m - 1$ in the binary case), and they produce all possible m symbol combinations within one period where m is the number of stages in the shift register structure. This property is very useful for simulating intersymbol interference (ISI) and other forms of linear distortion in digital transmission systems. Although it is possible to do this with random sequences derived from a uniform random-number sequence, the sequence length required to produce all possible m symbol combinations will be much longer than the PN sequences.

Details on PN sequence generation and a list of primitive polynomials can be found in Ref. 1.

5. PERFORMANCE ESTIMATION VIA MONTE CARLO SIMULATION

The primary use of simulation is performance evaluation and tradeoff studies. A number of performance metrics such as power spectral densities, S/N at the output of a receiver and bit error rates (BER) in digital systems can be estimated using Monte Carlo simulations. Just as in measurements, the estimated quantities are subjected to variations that are inherent in estimation of parameters from random observations (or simulation results). The metrics used to judge the quality of the estimators are the bias and variance. While it is easy to construct estimators that are unbiased, the variance of the estimator is not very easy to control since it depends on the type of estimator used as well as the number of observations used to obtain an estimated value. In general, the variance will be inversely proportional to the sample size, and this often leads to long simulation runs, especially in the case of low BER estimation. Since BER estimation is very important in the simulation of digital transmission systems, we present some of the approaches for estimating BERs in the following sections.

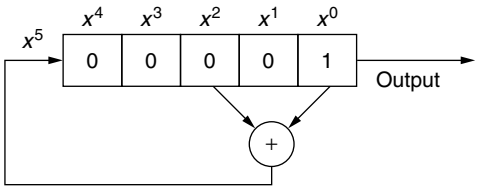


Figure 9. Feedback shift register for generating binary PN sequences; $m = 5$; $g(x) = 1 + x + x^5$.

5.1. MC Techniques for Estimating of BERs in Digital Communication Systems

The Monte Carlo technique is the most general method of estimating BERs and can be applied to any type of communication system with arbitrary distributions for noise and interference. The technique is simple to apply; perform a waveform-level simulation with a long input sequence of length N symbols, count the number of errors between the input symbol stream and the simulated output, and then form a counting estimate for the BER as the ratio of the number of errors counted and the number of symbols simulated (see Fig. 1). If the symbol errors in the system are occurring independently, then the normalized estimation error is given by [1]

$$\text{Normalized error} = \frac{\text{standard deviation of the estimator}}{\text{BER being estimated}} \approx \frac{1}{\sqrt{NP_e}} \tag{30}$$

For estimating a BER on the order of 10^{-6} with a normalized error of, say, 20%, the ordinary MC method requires a sample size on the order of $25(10^6)$ bits to be simulated. Hence the ordinary MC technique is not suitable for estimating very low BERs. A number of alternate methods have been developed in order to reduce sample size requirements for low-BER estimation. An overview of these methods and details may be found in the literature [1,23–25]. Two of these methods are described briefly here.

5.2. Semianalytical (SA) MC Techniques for Low-BER Estimation

This method is applicable to systems in which the effects of noise and interference can be assumed to be additive and Gaussian (of some other known distribution) at the output of the system [1]. In this case the BER in the system can be estimated by running a noiseless simulation to characterize the waveform distortion introduced by all the functional blocks in the system and analytically computing the probability of error due to noise superimposed on the distorted output. A simple example for the binary case with additive noise at the output is shown in Fig. 10. The basic form of the estimator becomes

$$\tilde{P}_e = \frac{1}{N}\sum_{k=1}^{N} P_k; \quad P_k = Q\left(\frac{d_k}{\sigma}\right) \tag{31}$$

where $Q(d_k/\sigma)$ is the analytically computed probability of error for the kth simulated value of the distorted output sample d_k at decision time, N is the number of symbols simulated, and σ is the standard deviation of

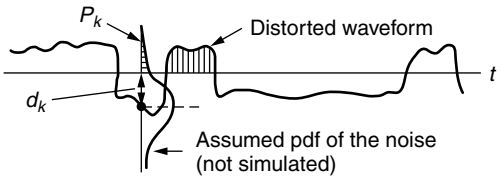


Figure 10. Semianalytical BER (bit error rate) estimator for a binary communication system.

the noise at the output of the system, which is assumed to be Gaussian in this example. With this approach, the simulation length is determined by the number of symbols N needed to simulate the distribution of the waveform distortion accurately. Typically this length will be much smaller than what would be necessary to simulate the effects of noise explicitly, particularly at low BERs. If the primary source of distortion in the system is linear (ISI) and lasts over m symbols, then a PN sequence of length 2^m is all that will be needed in a binary system to simulate all possible distortion values exactly. The variance of the estimator is zero in this case. The SA Monte Carlo methods lead to significant reduction in sample size for low-BER estimation. Indeed, with these methods the sample size requirements are independent of the BER being estimated. The SA methods are applicable to M-PSK and QAM (multiphase shift keying and quadrature amplitude modulation) schemes also [1].

5.3. An Important Sampling Method

This method, which is shown in Fig. 11, is based on biasing the input PDFs such that the important regions of the input PDFs are enhanced to produce a larger number of symbol errors during simulation. The higher error rates can be estimated with smaller sample sizes, and the bias in the estimator, which results from biasing the input PFDs, can be corrected easily at the output of the system where errors are counted. The important sampling method when applied properly has the potential to reduce sample size requirements significantly. Details of the important sampling method can be found in the literature [1,24,26].

6. SUMMARY

Simulation is a very useful tool for the design and analysis of communication systems. In this article we presented the basic principles behind waveform-level simulation of communication systems. The first step is developing a simulation model of the system under study in a block diagram form containing parameterized representations of all the functional blocks in the system that might have a bearing on the design and analysis issues being addressed. The second step is the selection of models for the signal processing operations performed by the functional blocks and representing them appropriately using the lowpass equivalent representation and the sampling theorem. After a simulation model is completely specified in terms of the topology, functional blocks, parameters, and input signals, the next step is to execute the simulations by generating sampled values of all the input signals using appropriate random-number generators and letting the functional blocks in the simulation model operate on the input sequences and produce output sequences. The final step is the estimation of performance metrics of interest such as BERs. This estimation can be done online while the simulation is being executed or at the end of the simulation as an offline postprocessing operation.

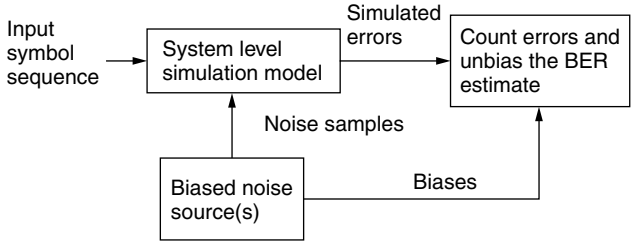


Figure 11. Important sampling method.

The overall simulation accuracy will depend on the modeling assumptions and approximations, accurate representation of signals, and accuracy of the algorithms used for random-number generation, estimation techniques used, and the length of the simulations. All of these issues were addressed in this article.

Simulation of communications is an interdisciplinary activity that requires skills in a broad range of areas, including communication systems, random signal theory, statistics, digital signal processing, and software engineering. A sound understanding of the fundamental principles in these areas as they apply to simulations is essential in order to produce valid and accurate answers to important design and analysis problems that face the communication engineers of today.

BIOGRAPHY

K. Sam Shanmugan received a Ph.D. degree from Oklahoma State University, Stillwater, Oklahoma, in electrical engineering in 1970. He is currently the SW Bell Distinguished Professor of Telecommunication in the Electrical Engineering and Computer Science Departments at the University of Kansas. He has also worked for AT&T Bell Laboratories, TRW, Hughes and Cadence Design Systems. Dr. Shanmugan is the author of over 100 publications in the above areas and is the author/coauthor of three books, *Digital and Analog Communication Systems* (Wiley, 1979), *Random Signals: Detection Estimation and Data Analysis* (Wiley, 1988), and *Simulation of Communication Systems* (Plenum Press, 1992). Dr. Shanmugan is a fellow of the IEEE and is the recipient of many teaching and research awards at the University of Kansas.

BIBLIOGRAPHY

1. M. C. Jeruchim, P. B. Balaban, and K. S. Shanmugan, *Simulation of Communication Systems*, 2nd ed., Kluwer-Plenum Press, New York, 2000 (provides in-depth coverage of all the major topics and also contains several hundred additional references).

2. *IEEE Journal on Selected Areas in Communications*, special issues devoted to computer-aided modeling, analysis and design of communications systems: **SAC-2**(1) (1984); **SAC-6**(1) (1988); **SAC-10**(1) (1992).

3. F. M. Gardner and J. D. Baker, *Simulation Techniques*, Wiley, New York, 1997.

4. A. V. Oppenheim and R. W. Schaffer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

5. A. V. Oppenheim, A. S. Willskey, and L. T. Young, *Signals and Systems*, Prentice-Hall, Englewood-Cliffs, NJ, 1983.

6. K. Sam Shanmugan and A. M. Breipohl, *Random Signals: Detection, Estimation and Data Analysis*, Wiley, New York, 1988.

7. T. Kailath, Channel characterization: Time varying dispersive channels, in *Lecturers in Communication Theory*, McGraw-Hill, New York, 1961.

8. A. A. M. Saleh, Frequency independent and frequency-dependent nonlinear models for TWT amplifiers, *IEEE Trans. Commun.* **Com-29**(11): 1715–1720 (1981).

9. H. M. Poza, Z. A. Sarkozy, and H. L. Berger, A wideband data link computer simulation model, *Proc. NAECON Conf.*, 1975.

10. T. S. Rappaport, *Wireless Communications*, Prentice-Hall, Upper Saddle River, NJ, 1996.

11. B. Sklar, Rayleigh fading channels in mobile digital communications, Parts I and II, *IEEE Commun. Mag.* **35**: 90–110 (July 1997).

12. Modified ITU propagation models for indoor, indoor to pedestrian and vehicular environments, 3GPP Ts.25.101 v.2.1.0 UE Radio Transmission and Reception (FFD), *www.3gpp.org*.

13. B. D. Fritchman, A binary channel characterization using partitioned Markov chains, *IEEE Trans. Inform. Theory* **IT-13**: 221–227 (April 1967).

14. S. Sivaprakasam and K. Sam Shanmugan, An equivalent Markov model for burst errors in digital channels, *IEEE Trans. Commun.* 1347–1356 (April 1995).

15. L. R. Rabiner and B. H. Huang, An introduction to hidden Markov models, *IEEE ASSP Mag.* 4–16 (Jan. 1986).

16. W. Turin, *Performance Analysis of Digital Transmission Systems*, Computer Science Press, Rockville, MD, 1990.

17. W. Turin and P. Balaban, Markov model for burst errors in narrowband CDMA system operating over a fading channel, *Proc. Globecom'98*, Sydney, 1998.

18. A. Beverly and K. Sam Shanmugan, Hidden Markov models for burst errors in GSM and DECT channels, *Proc. Globecom'98*, Sydney, 1998.

19. R. F. W. Coates, G. J. Janacek, and K. V. Lever, Monte Carlo simulation and random number generation, *IEEE J. Select. Areas Commun.* **6**(1): 58–66 (Jan. 1988).

20. D. E. Knuth, *The Art of Computer Programming*, Vol. 2, *Seminumerical Algorithms*, 2nd ed., Addison-Wesley, Reading, MA, 1981.

21. R. Y. Rubinstein, *Simulation and the Monte Carlo Method*, Wiley, New York, 1981.

22. S. L. Marple, Jr., *Digital Spectral Analysis, with Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

23. M. C. Jeruchim, Techniques for estimating the bit error rate in the simulation of digital communication systems, *IEEE J. Select. Areas Commun.* **SAC-2**(1): 153–170 (Jan. 1984).

24. K. Sam Shanmugan and P. Balaban, A modified Monte Carlo simulation technique for evaluation of error rate in digital communication systems, *IEEE Trans. Commun.* **COM-28**(11): 1916–1928 (1980).

25. P. M. Hahn and M. C. Jeruchim, Developments in the theory and application of importance sampling, *IEEE Trans. Commun.* **COM-35**(7): 706–714 (July 1987).

SOFT OUTPUT DECODING ALGORITHMS

LANCE C. PÉREZ
University of Nebraska
Lincoln, Nebraska

1. INTRODUCTION

The success of turbo codes and iterative decoding in the field of channel coding for the additive white Gaussian noise (AWGN) and other channels has led to the investigation of iterative techniques in a wide range of disciplines. Indeed, iterative processing is being considered for virtually every component in single and multiuser digital communication systems. Iterative processing typically involves iterative information exchange between system components, such as an equalizer and a channel decoder or an interference canceler and channel decoder, that traditionally operated independently.

The essential element of all iterative processing techniques is some form of a soft-input, soft-output (SISO) module. In this article, a detailed description is given of the most common SISO modules, namely, the soft-output Viterbi algorithm (SOVA) [1] and several versions of the Bahl, Cocke, Jelinek, and Raviv (BCJR) [2] or maximum a posteriori (MAP) algorithm. To make the descriptions concrete, these algorithms are described in the context of their application to the decoding of binary convolutional codes. The extension to other applications is generally straightforward and may be found in the appropriate literature.

The outline of this article is as follows. Section 2 provides a basic system description and introduces the basic concepts and notations of convolutional codes and their trellis representations necessary for the subsequent development of the SISO algorithms. A detailed description of the SOVA is given in Section 3. In Section 4, the MAP algorithm and its max-log variant are described. Section 5 contains some comparisons between these algorithms in the application of iterative decoding of turbo codes. Finally, some concluding remarks and pointers to areas for further reading are given in Section 6.

2. SYSTEM MODEL

For the purposes of this article, a digital communication system with forward error correction (FEC) channel coding can be represented by the block diagram shown in Fig. 1. A detailed derivation of this model and the required assumptions may be found in Ref. [3]. The source is a binary memoryless source that produces a sequence

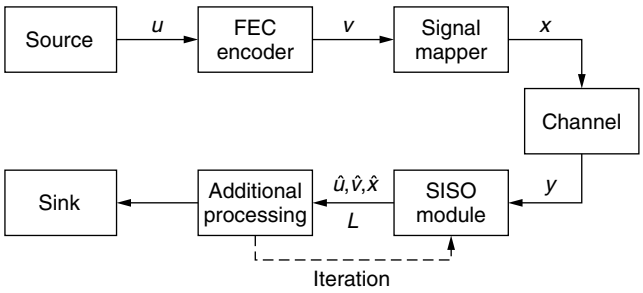


Figure 1. Block diagram of a digital communication system with forward error correction.

$\boldsymbol{u} = [u_0, \ldots, u_r, \ldots, u_{N-1}]$ of independent identically distributed 0's and 1's with equal a priori probabilities of $p_0 = p_1 = 1/2$. The FEC encoder maps the information sequence $\boldsymbol{u}$ to the code sequence $\boldsymbol{v} = [v_0, \ldots, v_r, \ldots, v_{N-1}]$ according to some encoding rule. The signal mapper maps each $n-$tuple v_r to one of 2^n symbols in the signal set. For convenience, the signal mapper is frequently chosen to be a binary phase shift keying (BPSK) or an antipodal modulator that maps each 0 and 1 of the code sequence $\boldsymbol{v}$ to a -1 or $+1$, respectively. The output sequence of the signal mapper, $\boldsymbol{x} = [x_0, \ldots, x_r, \ldots, x_{N-1}]$ is then transmitted across the memoryless AWGN channel, which adds to each transmitted symbol an independent Gaussian random variable with probability density function

$$p_N(n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{n^2}{2\sigma^2}}$$

where $\sigma^2 = \dfrac{N_o}{2}$.

The FEC encoder explicitly considered here is a convolutional encoder of rate $R_c = k/n$ with total encoder memory ν. During each encoding epoch r, the convolutional encoder maps the current $k-$tuple of information bits $u_r = [u_r^{(1)}, \ldots, u_r^{(k)}]$ to an output $n-$tuple of coded bits $v_r = [v_r^{(1)}, \ldots, v_r^{(n)}]$ based on the current input and the current encoder state s_r. The $n-$tuple of coded bits v_r and its corresponding signal mapper output x_r are referred to as a symbol. Although not required, the subsequent decoder descriptions will assume finite length information sequences and thus $r = 0, \ldots, N-1$. The decoding algorithms considered in this article all require the notion of the trellis representation of a convolutional code, which is simply the time expansion of the state transition diagram. In this case, the trellis has a total of 2^ν distinct states $s_r = j$, $j = 0, \ldots, 2^\nu - 1$, at epoch r with 2^k state transitions, or branches, leaving and entering each state.

For example, Fig. 2 depicts a rate $R_c = 1/2$ convolutional encoder with total encoder memory $\nu = 2$ realized in nonsystematic feedforward form [4]. The code is specified by its two generator polynomials, $g_0 = 1 + D + D^2 = 111$ and $g_1 = 1 + D^2 = 101$, which are frequently represented in right justified octal format as $g_0 = 7$ and $g_1 = 5$. The equivalent recursive systematic encoder realization [5] is shown in Fig. 3. Assuming that the encoder is initialized to the all zero state, the trellis diagram with $N = 7$ sections for this encoder is shown in Fig. 4. Each full section of

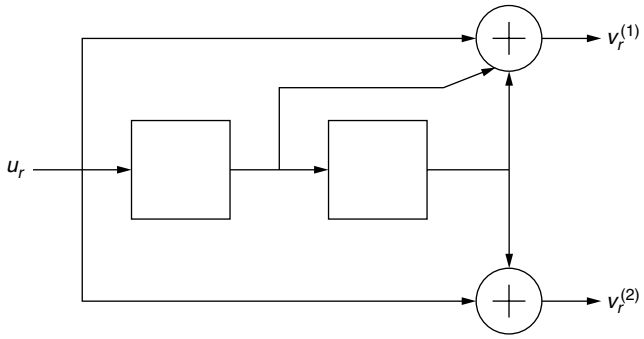


Figure 2. A rate $R_c = 1/2$ convolutional encoder with $\nu = 2$ realized in nonsystematic feedforward form with generator polynomials $g_0 = 1 + D + D^2$ and $g_1 = 1 + D^2$.

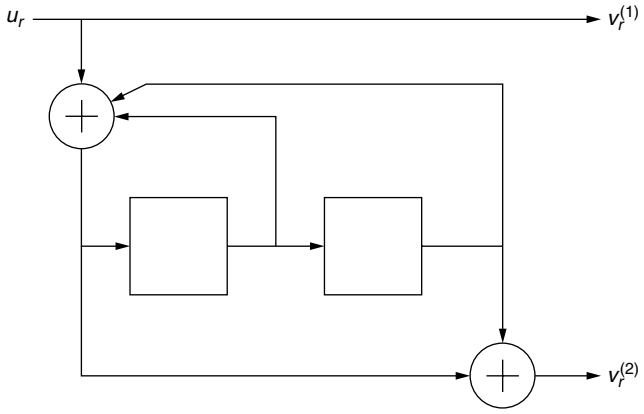


Figure 3. A rate $R_c = 1/2$ convolutional encoder with $\nu = 2$ realized in recursive systematic form with generator polynomials $g_0 = 1 + D + D^2$ and $g_1 = 1 + D^2$.

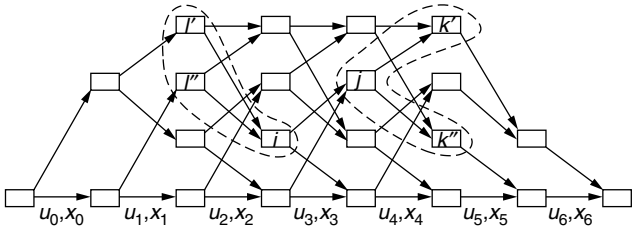


Figure 4. Trellis diagram corresponding to the encoders of Figs. 2 and 3 with information sequences of length 5 and two tail bits.

the trellis has $2^2 = 4$ states and there are $2^1 = 2$ branches leaving and entering each state. In this example, the trellis ends in the all zero state as well. This is referred to as *terminating* the trellis and is accomplished for rate one-half feedforward encoder realizations by appending a *tail* of ν zeroes to the end of the information sequence. Thus, this trellis represents information sequences of length 5 with a tail of 2 zeroes. For recursive systematic and systematic feedback encoder realizations, trellis termination is accomplished by a nonzero, state-dependent tail of ν bits [5]. Although not strictly required by the decoding algorithms described in this article, the subsequent development assumes that the code trellis is terminated.

The decoder operates on the noisy received sequence y. In traditional decoding algorithms for convolutional codes, such as the Viterbi algorithm [6] or sequential decoding, the primary goal of the decoder is to produce an estimate of the transmitted symbol sequence $\hat{\mathbf{x}}$, or equivalently an estimate of the information sequence $\hat{\mathbf{u}}$, consistent with minimizing, or nearly minimizing, an appropriate cost function, such as the probability of an information bit error P_b or the probability of a frame or sequence error P_f. The purpose of the SISO algorithms is to produce an estimate of the transmitted symbol sequence or information sequence along with a sequence $L = [L_0, \ldots, L_r, \ldots, L_{N-1}]$ of soft reliability information indicating the confidence of each component of the sequence estimate. In the context of this article, the reliability information is based on the calculation, or approximation, of the a posteriori probabilities of either the information bits $\Pr[u_r \mid \boldsymbol{y}]$, or the a posteriori probabilities of the transmitted symbols $\Pr[x_r \mid \boldsymbol{y}]$.

In traditional decoding, maximizing the a posteriori probabilities of the information bits, although optimum in terms of P_b, leads to only minor improvements compared to the Viterbi algorithm, which minimizes the sequence error rate $\Pr[\hat{\mathbf{v}} \neq \mathbf{v}] \mid \boldsymbol{y}]$. Thus, even though the MAP algorithm was originally formulated for convolutional codes by Bahl, Cocke, Jelinek, and Raviv [2] in 1972, it was not widely used because it provided no significant improvement over maximum-likelihood decoding and is significantly more complex. Interest in decoding algorithms with soft outputs was rekindled by the work of Hagenauer and Hoeher [1] on the soft output Viterbi algorithm (SOVA) in 1989 and pushed to its current ubiquity with the discovery of turbo codes and iterative decoding in 1993 [7]. For this reason, the discussion of the SISO algorithms begins with the SOVA and then progresses through the MAP algorithm and its variants.

The SISO algorithms described in this chapter all attempt to provide soft reliability information about each bit, either $u_r^{(i)}$ or $v_r^{(i)}$, or each symbol x_r. The basic approach to this is to partition the set of code sequences into two sets Ω_r, which contains sequences where the r^{th} bit or symbol takes on the desired value, and Ω_r^c, which contains sequences where the r^{th} bit or symbol differs from the desired value. The metrics associated with paths in each set may then be used to generate a reliability value for the current bit or symbol. The algorithms described here differ in two fundamental ways: (1) the number of paths used in each set Ω_r and Ω_r^c, and (2) the method for finding these paths.

3. SOFT OUTPUT VITERBI ALGORITHM (SOVA)

As its name suggests, the SOVA is a version of the classical hard output Viterbi algorithm modified to provide soft output reliability information. The Viterbi algorithm minimizes the probability of a sequence error by decoding that sequence $\boldsymbol{v}$ with the largest likelihood $\Pr[\boldsymbol{x} \mid \boldsymbol{y}]$ given the received sequence $\boldsymbol{y}$. For equally likely inputs, this is equivalent to maximizing $\Pr[\boldsymbol{y} \mid \boldsymbol{x}]$. (Thus, the Viterbi algorithm is a *sequence* MAP decoder.) The basic idea of the SOVA is to derive bit or symbol level reliability information from the sequence a posteriori probabilities. The SOVA described here is the algorithm discovered by Hagenauer and Hoeher [1] as modified by Fossorier et al. [8].

To see how this is done, we begin with a brief description of the Viterbi algorithm. For the AWGN channel with BPSK modulation, it is straightforward to show that

$$\Pr[\boldsymbol{y} \mid \boldsymbol{x}] = \prod_{r=0}^{N-1} \prod_{i=1}^{n} \frac{1}{\sqrt{\pi N_o}} \exp\left\{-\frac{(y_r^{(i)} - x_r^{(i)})^2}{N_o}\right\} \sim \prod_{r=0}^{N-1} \prod_{i=1}^{n} \exp\left\{-\frac{(y_r^{(i)} - x_r^{(i)})^2}{N_o}\right\}$$

Using logarithms, this becomes

$$\log \Pr[\boldsymbol{y} \mid \boldsymbol{x}] \sim -\sum_{r=0}^{N-1} \sum_{i=1}^{n} (y_r^{(i)} - x_r^{(i)})^2 \tag{1}$$

and maximizing $\Pr[\boldsymbol{y} \mid \boldsymbol{x}]$ is equivalent to finding the symbol sequence $\boldsymbol{x}$ that is closest to the received sequence in terms of squared Euclidean distance. The Viterbi algorithm is an efficient technique based on the code trellis for finding the closest sequence.

To formulate the Viterbi algorithm, notice that the inner summation in Eq. (1) corresponds to the squared Euclidean distance between the r^{th} symbol in the transmitted sequence $\boldsymbol{x}$ and the r^{th} received symbol in $\boldsymbol{y}$. Since each code sequence is represented by a path through the code trellis, the r^{th} transmitted symbol corresponds to a state transition or branch from a state $s_r = i$ to state $s_{r+1} = j$ and the inner summation in Eq. (1) is referred to as a branch metric. Formally, the branch metric associated with transmitted symbol x_r is defined to be

$$\beta_r(x_r = x) = \sum_{i=1}^{n} (y_r^{(i)} - x^{(i)})^2 \tag{2}$$

Finally, define the *partial path metric* for path l at epoch R and state j as

$$M_{R,l}(j) = \sum_{r=0}^{R-1} \beta_r(x_{r,l} = x) \tag{3}$$

where $j = 0, \ldots, 2^v - 1$ and $x_{r,l}$ is the r^{th} symbol on the l^{th} path.

With these definitions the Viterbi algorithm may now be stated as

Step 1: Initialize $M_0(0) = 0$ and $M_0(i) = \infty$ for $i \neq 0$. Set $R = 0$.

Step 2: Increment $R = R + 1$. For each state $s_R = j$, compute the branch metrics for the 2^k branches entering that state and compute the 2^k partial path metrics

$$M_{R,l}(j) = M_{R-1}(i) + \beta_{R-1}(x_{R-1,l} = x), \tag{4}$$

where $s_{R-1} = i$ is a state at epoch $R-1$ with a branch leading to state $s_R = j$ with branch label $x_{R-1,l}$.

Step 3: For each state $s_R = j$, compare the 2^k partial path metrics $M_{R,l}(j)$ and choose the minimum, that is,

$$M_R(j) = \min_l M_{R,l}(j).$$

Store the corresponding partial sequence $\tilde{\boldsymbol{u}}^j = [\hat{u}_0, \ldots, \hat{u}_{R-1}]$. This partial sequence is known as the survivor for state j at epoch R.

Step 4: If $R < N$, then return to Step 2. If $R = N$, then the survivor to state $s_N = 0$ with path metric $M_N(0)$ is the decoded sequence.

Note that this statement assumes that the encoder started in the all zero state and that the trellis is terminated and ends in the all zero state. Steps 2 and 3 are commonly referred to as the add-compare-select (ACS) operation of the Viterbi algorithm.

The output of the Viterbi algorithm is simply a decoded sequence with no explicit reliability information. The goal of a SISO algorithm is to provide some reliability information, usually in the form of an a posteriori probability, in addition to the decoded sequence. The basic observation behind the SOVA is that the real valued path metric, $M_N(0)$, of the decoded maximum likelihood sequence, $\hat{\mathbf{u}}$ or $\hat{\mathbf{x}}$, provides some reliability information about each decoded information bit $\hat{u}_r$. Assume for the moment that $\hat{u}_r = 1$. If the Viterbi algorithm could be modified to provide the path metric $M_N^c(0)$ of a path with $\hat{u}_r = 0$, then the path metric difference

$$\Delta = |M_N(0) - M_N^c(0)| \tag{5}$$

could be used as a reliability value for the r^{th} bit.

To provide the best reliability measure, the metric $M_N^c(0)$ should correspond to the *best* path with $\hat{u}_r = 0$. That is, $M_N^c(0)$ should correspond to the most likely path in the complementary set of sequences Ω_r^c. The question remains as to how this path can easily be found for each $\hat{u}_r$ of the decoded sequence. One solution is a modified Viterbi algorithm, called the SOVA, that computes reliability information for each bit via a traceback operation during the normal forward recursion of the Viterbi algorithm. These reliability values are stored for every bit in the partial sequence $\tilde{\mathbf{u}}^j$ for every state and the SOVA requires additional memory compared to the standard Viterbi algorithm.

For clarity of exposition and notation, the SOVA will be described for rate $R_c = 1/n$ codes realized in recursive systematic form. The latter assumption incurs no loss of generality, but simplifies the notation by ensuring that the information bit on the last branch of the two competing paths entering each state is different. That is, path 0 entering state $s_R = j$ has $\hat{u}_{R-1,0} = 0$ and path 1 entering $s_R = j$ has $\hat{u}_{R-1,1} = 1$. Consequently, each ACS operation at epoch R immediately generates a reliability value $L_{R-1}^j = \Delta = |M_{R,0} - M_{R,1}|$ for the current bit $\hat{u}_{R-1}$ at the current state $s_R = j$. Since this is the first time that the bit u_{R-1} is decoded, this is the best possible reliability value available at the moment. The situation for the remaining bits in the partial path is more complicated.

As the SOVA progresses through the trellis, it must update these reliability values such that when $R = N$ the reliability values for each bit $\hat{u}_r$ in the decoded sequence is generated from the best path with the complementary bit value in the r^{th} position. This is accomplished through careful updating of the reliability values L_r^j for $r = 0, \ldots, R-2$ following the ACS operation. To understand how this is done, two distinct cases must be considered for each ACS operation. A graphical representation of the ACS operation at state $s_R = 0$ is shown in Fig. 5. Without loss of generality, assume that path 0 is chosen as the survivor.

The first case occurs when the r^{th} bit, for some $0 \le r < R-1$, on path 0, denoted by $\hat{u}_{r,0}$, differs from the corresponding bit $\hat{u}_{r,1}$ on path 1. There are three reliability values to be considered when updating L_r^0:

1. $\Delta = |M_{R,0}(0) - M_{R,1}(0)|$, the difference in the partial path metrics of path 0 and path 1.
2. $L_r^0 = L_{r,0}$, the current reliability value of $u_r^0 = \hat{u}_{r,0}$ on path 0.
3. $L_{r,1}$, the current reliability value of $\hat{u}_{r,1}$ on path 1.

$L_{r,0}$ comes from an ACS operation prior to epoch R between path 0 and, say, path m in which $\hat{u}_{r,0} \neq \hat{u}_{r,m}$. That is, path m belongs to the complementary set Ω_r^c for $\hat{u}_{r,0}$. $L_{r,1}$ comes from an ACS operation prior to epoch R between path 1 and, say, path l in which $\hat{u}_{r,1} \neq \hat{u}_{r,l}$. It follows that $\hat{u}_{r,l} = \hat{u}_{r,0}$ and path l does not belong to the complementary set Ω_r^c for $\hat{u}_{r,0}$. Thus, the update equation for this first case is

$$L_r^j = \min\{\Delta, L_{r,0}\} \tag{6}$$

The second case occurs when the r^{th} bit, for some $0 \le r < R-1$, on path 0, denoted by $\hat{u}_{r,0}$, agrees with the corresponding bit $\hat{u}_{r,1}$ on path 1. Arguing as above, we find that in this case path l is now in the complementary set Ω_r^c with a reliability value of $\Delta + L_{r,1}$ and the update equation is

$$L_r^j = \min\{\Delta + L_{r,1}, L_{r,0}\} \tag{7}$$

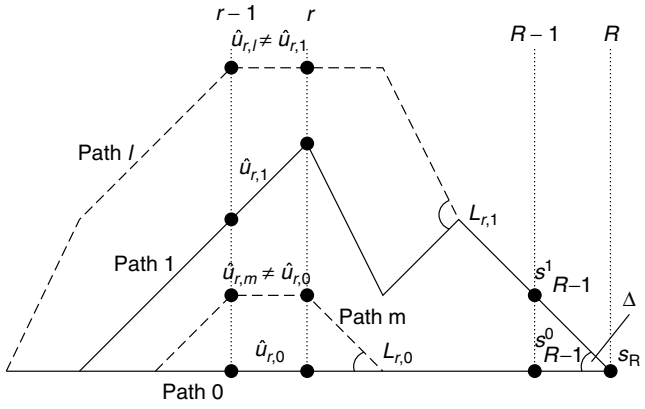


Figure 5. Figure illustrating the paths used in updating the reliability information at state 0 in the SOVA.

With these two update equations, the SOVA may now be stated as follows for rate $R_c = 1/n$ codes.

Step 1: Initialize $M_0(0) = 0$ and $M_0(i) = \infty$ for $i \neq 0$. Set $R = 0$.

Step 2: Increment $R = R + 1$. For each state $s_R = j$, compute the branch metrics for the 2^k branches entering that state and compute the 2^k partial path metrics

$$M_{R,l}(j) = M_{R-1}(i) + \beta_{R-1}(x_{R-1,l} = x) \tag{8}$$

where $s_{R-1} = i$ is a state at epoch $R-1$ with a branch leading to state $s_R = j$ with branch label $x_{R-1,l}$.

Step 3: For each state $s_R = j$, compare the 2^k partial path metrics $M_{R,l}(j)$ and choose the minimum, that is,

$$M_R(j) = \min_l M_{R,l}(j)$$

Store the corresponding partial sequence $\tilde{\boldsymbol{u}}^j = [\hat{u}_0, \ldots, \hat{u}_{R-1}]$. This partial sequence is known as the survivor for state j at epoch R.

Step 3a: For each state $s_R = j$, compute Δ, trace back the survivor and update the reliability values L_R^j of each bit according to Eqs. (6) and (7)

Step 4: If $R < N$, then return to Step 2. If $R = N$, then the survivor to state $s_N = 0$ with path metric $M_N(0)$ is the decoded sequence and the reliability values are L_N^0.

Note that this statement assumes that the encoder started in the all zero state and that the trellis is terminated and ends in the all zero state.

This version of the SOVA computes reliability values based on a single path from each of the sets Ω_r and Ω_r^c. The path used in each set is optimum in the maximum likelihood sequence sense of the Viterbi algorithm. The traceback operation required in this version of the SOVA for updating the reliability values introduces considerable complexity and may not be suitable for some applications. Several alternative algorithms may be found in the literature [9,10]. In the sequel, algorithms that attempt to compute or approximate the a posteriori probabilities directly are discussed.

4. A POSTERIORI PROBABILITY ALGORITHMS

4.1. BCJR or MAP Decoding

The ultimate purpose of this algorithm is the calculation of a posteriori probabilities, such as $\Pr[u_r \mid \boldsymbol{y}]$, or $\Pr[x_r \mid \boldsymbol{y}]$, where $\boldsymbol{y}$ is the received sequence observed at the output of a channel, whose input is the transmitted sequence $\boldsymbol{x}$. Following Ref. 2, it is convenient to calculate the probability that the encoder traversed a specific branch in the trellis, that is, $\Pr[s_r = i, s_{r+1} = j \mid \boldsymbol{y}]$, where s_r is the state at epoch r, and s_{r+1} is the state at epoch $r+1$. The BCJR or MAP algorithm computes this probability as

$$\begin{aligned}\Pr[s_r = i, s_{r+1} = j \mid \boldsymbol{y}] &= \frac{1}{\Pr(\boldsymbol{y})}\Pr[s_r = i, s_{r+1} = j, \boldsymbol{y}] \\ &= \frac{1}{\Pr(\boldsymbol{y})}\alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j) \end{aligned} \tag{9}$$

The α-values are internal variables of the algorithm and are computed by the *forward recursion*

$$\alpha_{r-1}(i) = \sum_{\text{states } l} \alpha_{r-2}(l)\gamma_{r-1}(i,l) \tag{10}$$

This forward recursion evaluates α-values at time $r-1$ from previously calculated α-values at time $r-2$, and the sum is over all states l at time $r-2$ that connect with state i at time $r-1$. The forward recursion for the trellis of Fig. 4 is illustrated in Fig. 6. To enforce the boundary condition that the encoder begins in state 0, the α-values are initialized as $\alpha_0(0) = 1, \alpha_0(1) = \alpha_0(2) = \alpha_0(3) = 0$.

The β-values are calculated in a similar manner, called the *backward recursion*

$$\beta_r(j) = \sum_{\text{states } k} \beta_{r+1}(k)\gamma_{r+1}(k,j) \tag{11}$$

with initial values of $\beta_N(0) = 1, \beta_N(1) = \beta_N(2) = \beta_N(3) = 0$ to enforce the terminating condition of the trellis code. The sum is over all states k at time $r+1$ to which state j at time r connects. The backward recursion is illustrated in Fig. 7.

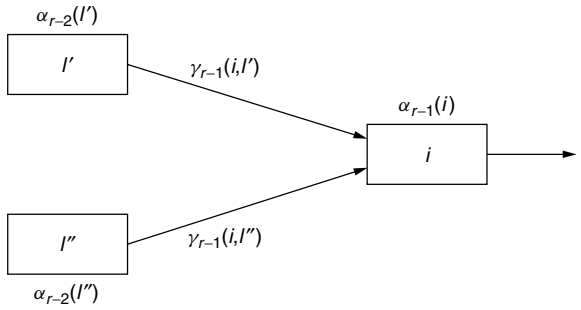


Figure 6. Illustration of the forward recursion of the MAP algorithm.

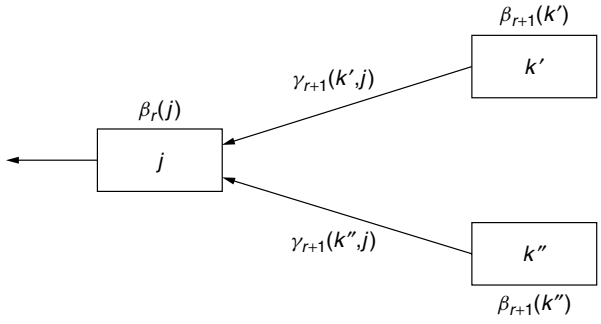


Figure 7. Illustration of the backward recursion of the MAP algorithm.

The γ-values are conditional transition probabilities and are the inputs to the algorithm based on the received sequence. Specifically, $\gamma_r(j,i)$ is the joint probability that the state at time $r+1$ is $s_{r+1}=j$ and that y_r is received, given $s_r=i$. It is calculated using the expression

$$\begin{aligned}\gamma_r(j,i) &= \Pr(s_{r+1}=j, y_r \mid s_r=i)\\ &= \Pr[s_{r+1}=j \mid s_r=i]\Pr(y_r \mid x_r)\end{aligned} \tag{12}$$

where $\Pr[s_{r+1}=j \mid s_r=i]$ is the a priori transition probability and is related to the probability of u_r. For feedforward encoders, the top transition in the trellis diagram of Fig. 4 is associated with $u_r=1$ and the bottom transition with $u_r=0$. This term is used to account for a priori probability information on the bits u_r. To simplify notation, this transition probability will be denoted as

$$p_{ij} = \Pr[s_{r+1}=j \mid s_r=i] = \Pr[u_r] \tag{13}$$

The second term, $\Pr(y_r \mid x_r)$, is simply the conditional channel transition probability, given that symbol x_r is transmitted. Note that x_r is the symbol associated with the transition from state $i \to j$.

The a posteriori symbol probabilities $\Pr[u_r \mid \boldsymbol{y}]$ can now be calculated from the a posteriori transition probabilities (9) by summing over all transitions corresponding to $u_r=1$, and, separately, by summing over all transitions corresponding to $u_r=0$, to obtain

$$p[u_r=1 \mid \boldsymbol{y}] = \frac{1}{\Pr(\boldsymbol{y})}\sum_{u_r=1}\Pr[s_r=i, s_{r+1}=j, \boldsymbol{y}] \tag{14}$$

$$p[u_r=0 \mid \boldsymbol{y}] = \frac{1}{\Pr(\boldsymbol{y})}\sum_{u_r=0}\Pr[s_r=i, s_{r+1}=j, \boldsymbol{y}] \tag{15}$$

From these equations it is clear that the MAP algorithm computes the a posteriori probabilities using all the sequences in the sets Ω_r and Ω_r^c.

The derivation of the MAP algorithm requires the probability

$$q_{ij}(x) = \Pr(\tau(u_r, s_r)=x \mid s_r=i, s_{r+1}=j) \tag{16}$$

that is, the a priori probability that the output x_r at time r assumes the value x on the transition from state i to state j. For convolutional codes, as opposed to coded modulation schemes, this probability is a deterministic function of i and j.

To begin, define the internal variables α and β by their probabilistic meaning. These are

$$\alpha_r(j) = \Pr(s_{r+1}=j, \tilde{\boldsymbol{y}}) \tag{17}$$

the joint probability of the partial sequence $\tilde{\boldsymbol{y}}=(y_{-l},\ldots,y_r)$ up to and including time epoch r and state $s_{r+1}=j$; and

$$\beta_r(j) = \Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j) \tag{18}$$

the conditional probability of the remainder of the received sequence $\boldsymbol{y}$ given that the state at time $r+1$ is j.

It is now possible to calculate

$$\begin{aligned}\Pr(s_{r+1}=j, \boldsymbol{y}) &= \Pr(s_{r+1}=j, \tilde{\boldsymbol{y}}, (y_{r+1},\ldots,y_l))\\ &= \Pr(s_{r+1}=j, \tilde{\boldsymbol{y}})\Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j, \tilde{\boldsymbol{y}})\\ &= \alpha_r(j)\beta_r(j)\end{aligned} \tag{19}$$

where we have used the fact that $\Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j, \tilde{\boldsymbol{y}}) = \Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j)$, that is, if $s_{r+1}=j$ is known, events after time r are independent of the history $\tilde{\boldsymbol{y}}$ up to s_{r+1}.

In the same way we calculate via Bayes' expansion

$$\begin{aligned}\Pr(s_r=i, s_{r+1}=j, \boldsymbol{y}) &= \Pr(s_r=i, s_{r+1}=j, (y_{-l},\ldots,y_{r-1}),\\ &\quad \times y_r, (y_{r+1},\ldots,y_l))\\ &= \Pr(s_r=i, (y_{-l},\ldots,y_{r-1}))\\ &\quad \times \Pr(s_{r+1}=j, y_r \mid s_r=i)\\ &\quad \times \Pr((y_{r+1},\ldots,y_l) \mid s_{r+1}=j)\\ &= \alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j)\end{aligned} \tag{20}$$

Now, again applying Bayes' rule and $\sum_b p(a,b) = p(a)$, we obtain

$$\begin{aligned}\alpha_r(j) &= \sum_{\text{states } i}\Pr(s_r=i, s_{r+1}=j, \tilde{\boldsymbol{y}})\\ &= \sum_{\text{states } i}\Pr(s_r=i, (y_{-l},\ldots,y_{r-1}))\Pr(s_{r+1}=j, y_r \mid s_r=i)\\ &= \sum_{\text{states } i}\alpha_{r-1}(i)\gamma_r(j,i)\end{aligned} \tag{21}$$

For a trellis code started in the zero state at time $r=-l$ we have the starting conditions

$$\alpha_{-l-1}(0)=1, \alpha_{-l-1}(j)=0;\ j\neq 0 \tag{22}$$

As above, we similarly develop an expression for $\beta_r(j)$, that is,

$$\begin{aligned}\beta_r(j) &= \sum_{\text{states } i}\Pr(s_{r+2}=i, (y_{r+1},\ldots,y_l) \mid s_{r+1}=j)\\ &= \sum_{\text{states } i}\Pr(s_{r+2}=i, y_{r+1} \mid s_{r+1}=j)\\ &\quad \times \Pr((y_{r+2},\ldots,y_l) \mid s_{r+2}=i)\\ &= \sum_{\text{states } i}\beta_{r+1}(i)\gamma_{r+1}(i,j)\end{aligned} \tag{23}$$

The boundary condition for $\beta_r(j)$ is

$$\beta_l(0)=1, \beta_l(j)=0; \quad j\neq 0 \tag{24}$$

for a trellis code which is terminated in the zero state.

Furthermore, the general form of the γ values is given by

$$\begin{aligned}\gamma_r(j,i) &= \sum_{x_r} \Pr(s_{r+1}=j \mid s_r=i) \\ &\quad \times \Pr(x_r \mid s_r=i, s_{r+1}=j)\Pr(y_r \mid x_r) \\ &= \sum_{x_r} p_{ij} q_{ij}(x_r) p_N(y_r - x_r) \end{aligned} \tag{25}$$

Equations (21) and (23) are iterative and the a posteriori state and transition probabilities can now be calculated via the following algorithm.

Step 1: Initialize $\alpha_{-l-1}(0)=1, \alpha_{-l-1}(j)=0$ for all non-zero states ($j \neq 0$) of the encoder, and $\beta_l(0)=1, \beta_l(j)=0, j \neq 0$. Let $r=-l$.

Step 2: For all states j calculate $\gamma_r(j,i)$ and $\alpha_r(j)$ via Eqs. (25) and (21).

Step 3: If $r<l$, let $r=r+1$ and go to Step 2, else $r=l-1$ and go to Step 4.

Step 4: Calculate $\beta_r(j)$ using Eq. (23). Calculate $\Pr(s_{r+1}=j, \boldsymbol{y})$ from Eq. (19), and $\Pr(s_r=i, s_{r+1}=j; \boldsymbol{y})$ from Eq. (9).

Step 5: If $r>-l$, let $r=r-1$ and go to Step 4.

Step 6: Terminate the algorithm and output all the values $\Pr(s_{r+1}=j, \boldsymbol{y})$ and $\Pr(s_r=i, s_{r+1}=j, \boldsymbol{y})$.

The a posteriori state and transition probabilities produced by this algorithm can now be used to calculate a posteriori information bit probabilities, that is, the probability that the information k-tuple $u_r=u$, where u can vary over all possible binary k-tuples. Starting from the transition probabilities $\Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y})$, we simply sum over all transitions $i \to j$ that are caused by $u_r=u$. Denoting these transitions by $A(u)$, we obtain

$$\Pr(u_r=u) = \sum_{(i,j)\in A(u)} \Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \tag{26}$$

As mentioned previously, another most interesting product of the APP decoder is the a posteriori probability of the transmitted output symbol x_r. Arguing analogously as above, and letting $B(x)$ be the set of transitions on which the output signal x can occur, we obtain

$$\begin{aligned}\Pr(x_r=x) &= \sum_{(i,j)\in B(x)} \Pr(x \mid y_r)\Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \\ &= \sum_{(i,j)\in B(x)} \frac{p_N(y_r-x_r)}{p(y_r)} q_{ij}(x) \\ &\quad \times \Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \end{aligned} \tag{27}$$

where the a priori probability of y_r can be calculated via

$$p(y_r) = \sum_{\substack{x' \\ ((i,j)\in B(x))}} p(y_r \mid x') q_{ij}(x') \tag{28}$$

and the sum extends over all transitions $i \to j$.

Equation (27) can be much simplified if there is only one output symbol on the transition $i \to j$. In this case, the transition automatically determines the output symbol, and

$$\Pr(x_r=x) = \sum_{(i,j)\in B(x)} \Pr(s_r=i, s_{r+1}=j \mid \boldsymbol{y}) \tag{29}$$

5. LOG-MAP AND THE MAX-LOG-MAP

5.1. The MAP Algorithm in the Logarithm Domain (Log-MAP)

Although the MAP algorithm is concise and consists only of multiplications and additions, current direct digital hardware implementations of the algorithm lead to complex circuits due to many real-number multiplications involved in the algorithm. To avoid these multiplications, we transform the algorithm into the logarithm-domain. This results in the so-called *log-MAP* algorithm.

First, we transform the forward recursion (10), (21) into the logarithm-domain using the definitions

$$A_r(i) = \log(\alpha_r(i)); \quad \Gamma_r(i,l) = \log(\gamma_r(i,l)) \tag{30}$$

to obtain the *log-domain* forward recursion

$$A_{r-1}(i) = \log\left(\sum_{\text{states } l} \exp\left(A_{r-2}(l) + \Gamma_{r-1}(i,l)\right)\right) \tag{31}$$

Likewise, the backward recursion can be transformed into the logarithm-domain using the analogous definition $B_r(j)=\log(\beta_r(j))$, and we obtain

$$B_r(j) = \log\left(\sum_{\text{states } k} \exp\left(B_{r+1}(k) + \Gamma_{r+1}(k,j)\right)\right) \tag{32}$$

The product in Eqs. (9) and (20) now turns into the simple sum

$$\alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j) \to A_{r-1}(i) + \Gamma_r(j,i) + B_r(j) \tag{33}$$

Unfortunately, Eqs. (31) and (32) contain log() and exp() functions, which are more complex than the original multiplications. However, in most cases of current practical interest, the MAP algorithm is used to decode binary codes with $R_c=1/n$ where there are only two branches involved at each state, and therefore only sums of two terms in Eqs. (31) and (32). The logarithm of such a binary sum can be expanded as

$$\begin{aligned}\log(\exp(a)+\exp(b)) &= \log\left(\exp\left(\max(a,b)\right)\right. \\ &\quad \left. \times\ (1+\exp(-|a-b|))\right) \\ &= \max(a,b) + \log\left((1+\exp(-|a-b|))\right) \end{aligned}$$

The second term is now the only complex operation left and there are a number of methods to approach this including a look-up table.

Finally, for binary codes the algorithm computes the log-likelihood ratio (LLR) $\lambda(u_r)$ of the information bits u_r using the a posteriori probabilities (26) as

$$\begin{aligned}\lambda(u_r) &= \log\left(\frac{\Pr(u_r=1)}{\Pr(u_r=0)}\right)\\ &= \log\left(\frac{\displaystyle\sum_{(i,j)\in A(u=1)} \alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j)}{\displaystyle\sum_{(i,j)\in A(u=0)} \alpha_{r-1}(i)\gamma_r(j,i)\beta_r(j)}\right)\end{aligned}$$

$$\lambda(u_r) = \log\left(\frac{\displaystyle\sum_{(i,j)\in A(u=1)} \exp(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j))}{\displaystyle\sum_{(i,j)\in A(u=0)} \exp(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j))}\right) \quad (34)$$

The range of the LLR is $[-\infty,\infty]$, where a large value signifies a high probability that $u_r = 1$.

5.2. Max-Log-MAP

The complexity of the log-MAP algorithm may be further reduced by approximating the forward and backward recursions by

$$\begin{aligned}A_{r-1}(i) &= \log\left(\sum_{\text{states } l} \exp\left(A_{r-2}(l)+\Gamma_{r-1}(i,l)\right)\right)\\ &\approx \max_{\text{states } l}\left(A_{r-2}(l)+\Gamma_{r-1}(i,l)\right)\end{aligned} \quad (35)$$

and

$$\begin{aligned}B_r(j) &= \log\left(\sum_{\text{states } k} \exp\left(B_{r+1}(l)+\Gamma_{r+1}(k,j)\right)\right)\\ &\approx \max_{\text{states } k}\left(B_{r+1}(l)+\Gamma_{r+1}(k,j)\right)\end{aligned} \quad (36)$$

which results in the max-log-MAP algorithm. The final LLR calculation in Eq. (34) is approximated by

$$\begin{aligned}\lambda(u_r) \approx &\max_{(i,j)\in A(u=1)}\left(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j)\right)\\ &-\max_{(i,j)\in A(u=0)}\left(A_{r-1}(i)+\Gamma_r(j,i)+B_r(j)\right)\end{aligned} \quad (37)$$

The advantage of the max-log-MAP algorithm is that it uses only additions and maximization operations to approximate the LLR of u_r. It is very interesting to note that Eq. (35) is the Viterbi algorithm for maximum-likelihood sequence decoding. Furthermore, Eq. (36) is also a Viterbi algorithm, but it is operated in the reverse direction.

Further insight into the relationship between the log-MAP and its approximation can be gained by expressing the LLR of u_r in the form

$$\lambda(u_r) = \log\left(\frac{\displaystyle\sum_{\boldsymbol{x};(u_r=1)} \exp\left(-\frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0}\right)}{\displaystyle\sum_{\boldsymbol{x};(u_r=0)} \exp\left(-\frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0}\right)}\right) \quad (38)$$

where the sum in the numerator extends over all coded sequences $\boldsymbol{x}$ that correspond to information bit $u_r = 1$, and the denominator sum extends over all $\boldsymbol{x}$ corresponding to $u_r = 0$.

It is quite straightforward to see that the max-log-MAP retains only the path in each sum that has the best metrics, and therefore the max-log-MAP calculates an approximation to the true LLR, given by

$$\lambda(u_r) \approx \min_{\boldsymbol{x};(u_r=0)} \frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0} - \min_{\boldsymbol{x};(u_r=1)} \frac{|\boldsymbol{y}-\boldsymbol{x}|^2}{N_0} \quad (39)$$

that is, the metric difference between the nearest path to $\boldsymbol{y}$ with $u_r = 0$ and the nearest path with $u_r = 1$. That is, like the SOVA described earlier, the max-log-MAP approximates the LLR by using the best sequence in each set Ω_r and Ω_r^c. Fossorier et al. [8] have shown that the max-log-MAP and the SOVA presented earlier are identical.

6. PERFORMANCE IN ITERATIVE DECODING

The SOVA, MAP, and max-log-MAP algorithms offer a myriad of complexity and implementation tradeoffs that are application- and technology-dependent and will be left to the literature. The performance tradeoff is clearer, although still somewhat dependent on the application. From the development presented here, one would expect the MAP and log-MAP algorithms to perform identically (ignoring any issues of numerical stability) and that both would slightly outperform the SOVA and max-log-MAP. The latter two algorithms are again essentially identical. To illustrate this, we consider the use of these SISO algorithms in the iterative decoding of a turbo code. It is worth noting that some modifications are required to enable the SOVA to easily use the extrinsic information of the iterative turbo decoder. Details of these modifications may be found in Refs. 10 and 11.

The performance of a turbo code with iterative decoding is shown in Fig. 8. The turbo code is a rate $R_c = 1/3$ code with identical memory $\nu = 4$ constituent encoders and an interleaver of length 4096 bits. The constituent encoders have feedforward polynomial $g_1(D) = 1 + D^4$ and feedback polynomial $g_0(D) = 1 + D + D^2 + D^3 + D^4$. All of the performance curves shown are for 18 complete decoder iterations. As expected, the performance with the MAP and log-MAP component decoders are identical, as is performance with the SOVA and max-log-MAP component decoders. There is a gap of approximately 0.5 dB between the two curves. This loss is attributable to the poorer quality soft information provided by the SOVA and max-log-MAP algorithms due to the use of only a single sequence in complementary set Ω_r^c.

7. CONCLUSION

This article discussed several SISO algorithms suitable, with minor adaptation, to the majority of applications considering soft-information exchange or iterative processing. They were presented in the context of the decoding of binary convolutional codes with finite block lengths. The extensions to sliding windows, block codes, equalization,

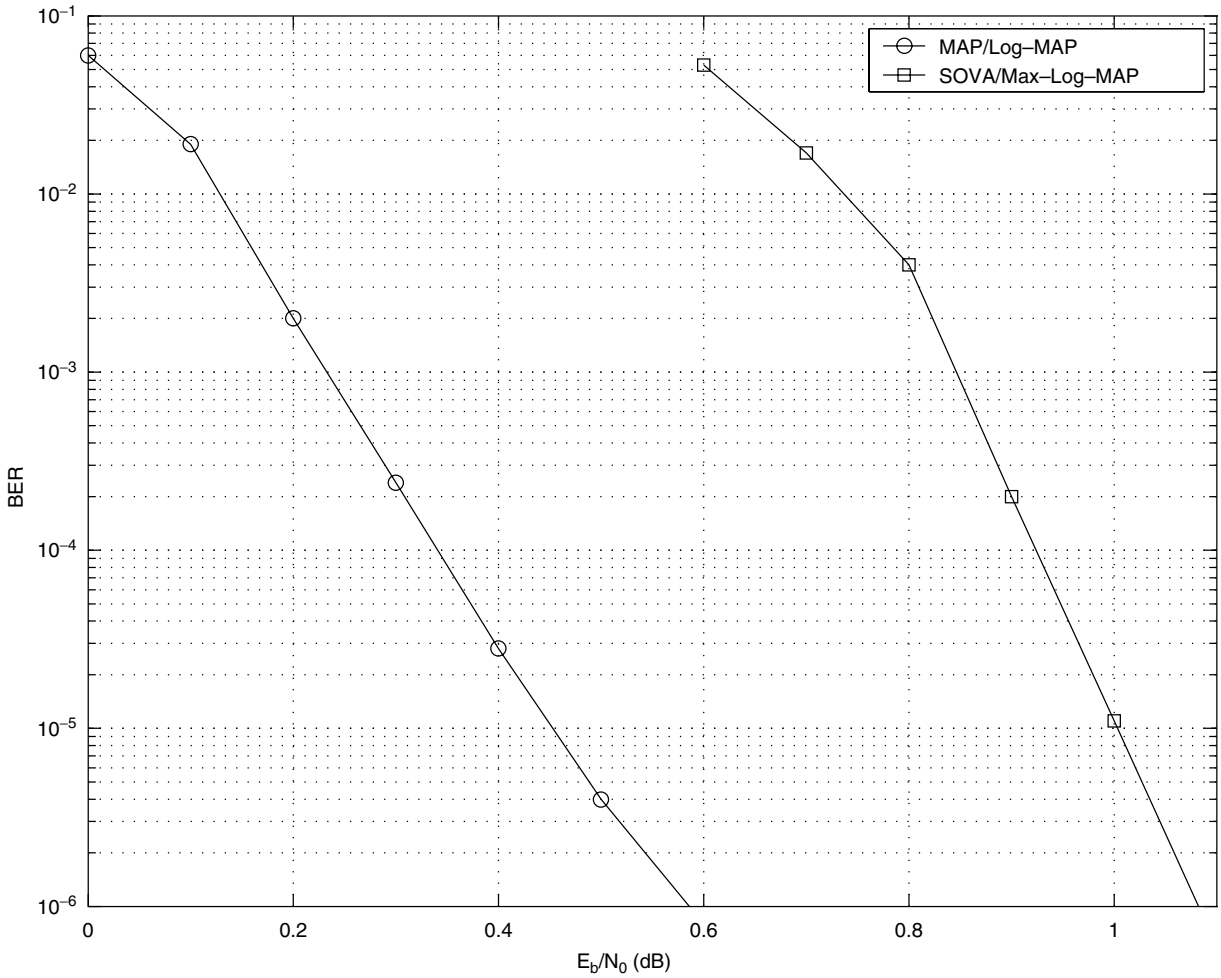


Figure 8. Performance comparison of rate 1/3 turbo code, with interleaver size 4096, and 18 iterations of decoding.

and other applications are relatively straightforward and may be found in the literature.

The virtually universal adoption of SISO algorithms has led to renewed interest in the general theory of decoding algorithms. This has led to several interesting results that unify ideas in fields as diverse as decoding algorithms, graph theory and belief propagation. These results require a degree of mathematical sophistication, but they provide a more general framework in which to understand a broad class of SISO algorithms. The interested reader may find some of these results in Refs. 12–14 and the references therein.

Acknowledgments

The author acknowledge the significant contributions of Christian B. Schlegel to the development of the MAP material in this work and the assistance of Christopher G. Hruby in the preparation of the figures and his valuable comments.

BIOGRAPHY

Lance C. Pérez received the B.S. degree in electrical engineering in 1987 from the University of Virginia, Charlottesville, Virginia, and the M.S. and Ph.D. degrees in electrical engineering from the University of Notre Dame, Notre Dame, Indiana in 1989 and 1995, respectively. From February 1, 1995, to July 31, 1995, Dr. Pérez was also a postdoctoral research associate with a joint appointment from University of Notre Dame and the Institute for Information and Signal Processing at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland. He joined the faculty of the Department of Electrical Engineering at the University of Nebraska, Lincoln, in August 1996 and he is currently an associate professor there. Dr. Pérez is the recipient of a National Science Foundation Career Award and is coauthor with Christian B. Schlegel of the book *Trellis and Turbo Coding* published the IEEE Press/Wiley. His areas of interest are channel coding, digital communications, and engineering education.

BIBLIOGRAPHY

1. J. Hagenauer and P. Hoeher, A Viterbi algorithm with soft-decision outputs and its applications, *Proceedings of GLOBECOM '89*, 1680–1686, Nov. 1989.
2. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).

3. J. G. Proakis, *Digital Communications*, McGraw-Hill, Boston, MA, 1989.
4. S. Lin and D. J. Costello, Jr., *Error Control Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
5. R. Johannesson and K. S. Zigangirov, *Fundamentals of Convolutional Coding*, IEEE Press, Piscataway, NJ, 1999.
6. G. D. Forney, Jr., The Viterbi algorithm, *Proceedings of the IEEE*, PROC-61, 268–278, 1973.
7. Claude Berrou, Alain Glavieux, and Punya Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proceedings of ICC'93*, 1064–1070, May 1993.
8. M. P. C. Fossorier, F. Burkert, S. Lin, and J. Hagenauer, On the equivalence between SOVA and max-log-MAP decodings, *IEEE Comm. Lett.* **COML-2**: 137–139 (May 1998).
9. J. Chen, M. P. C. Fossorier, S. Lin, and C. Xu, Bi-Directional SOVA decoding for turbo-codes, *IEEE Comm. Lett.* **COML-4**: 405–407 (Dec. 2000).
10. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **IT-42**: 429–445 (Mar. 1996).
11. M. Bossert, *Channel Coding for Telecommunications*, Wiley, New York, 1999.
12. R. J. McEliece, On the BCJR trellis for linear block codes, *IEEE Trans. Inform. Theory* **IT-41**: 1072–1092 (July 1996).
13. C. Heegard and S. B. Wicker, *Turbo Codes*, Kluwer, Norwell, 1999.
14. B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.

SOFTWARE RADIO

JOSEPH MITOLA III
Consulting Scientist
Tampa, Florida

1. INTRODUCTION

Software radio (SWR), briefly, is about increasingly wider-bandwidth radiofrequency (RF)-capable digital hardware that is given much or all of its function–personality by software. This chapter provides an overview of the mathematical, engineering, and economics principles of SWR from the theory to practice of radio systems engineering. In the space available, there is little opportunity to address either the larger network architectures, or the software-defined radio (SDR) implementation details. However, the treatment differentiates SWR from SDR. It describes the end-to-end partitioning of SDR requirements. Further, it describes the allocation of critical parameters including dynamic range and processing capacity. Allocation tradeoffs among hardware platforms, firmware and software depend on cost–benefit in the marketplace. In addition, as implementations migrate to software, one must assure that the software is structured well and performs robustly — even when many tasks are competing for processing resources. There are also pointers to relevant industry forums [1] and standards bodies [2].

Some try to "sell" the software radio, but that is not the purpose here. On the contrary, an ideal SWR approach sometimes yields an ineffective product. Pagers, for example, maximize display area and battery life, while minimizing the size and weight of a fixed-function product. Hardware-intensive application-specific integrated circuits (ASICs) are best for that market niche at present. On the other hand, nearly ideal SWR base stations have been deployed since early 2000 because of lower cost of ownership than the baseband–digital signal processor (DSP) base stations that they replace. One therefore must appreciate how analog, digital, and software-intensive approaches complement each other and drive alternative cost–benefit profiles. One may then select the right mix of hardware-intensive and software-defined implementation aspects of a design. This introduction should help the reader appreciate the potential contributions and pitfalls of SWR technology.

SWR is an interdisciplinary technology. It is helpful for software people to understand the RF hardware and air interface standards concepts of an interdisciplinary team. The software-oriented discussion is for people with strong background in RF, analog radio, or DSP but little background in large-scale software. SDRs typically have over one million lines of code (LoC), which is a complex, large-scale software systems. Thus large-scale software tools such as the Unified Modeling Language (UML), the Common Object Request Broker Architecture (CORBA), and the eXtensible Markup Language (XML), typically unfamiliar to RF engineers, loom large in software radio architecture.

The appropriate host platforms for SDR functions change over time. Commercial digital filter ASICs become obsolete as DSP capacity increases, changing the systems level tradeoffs from ASIC to DSP. As needs, technology, and team expertise evolve, the effective choice will also change. Platforms also change with the top–down design constraints, such as market economics, consumer values, and network architecture. For quick time to market, one may procure functions in off-the-shelf code or intellectual property. A sound systems-level architecture facilitates this process, while an inferior architecture inhibits it. The in-depth understanding of SDR therefore includes markets and economics.

One revolutionary aspect of software radio is that knowing how to code a radio algorithm in the programming language C on a DSP chip no longer gives a software engineer the core skills needed to contribute effectively to software radio systems development. In fact, that experience becomes a liability if it causes one to minimize the importance of the new large-scale software engineering methods such as UML, XML, and CORBA.

In addition, European readers will recognize SDL, the ITU-standard Specification and Description Language. In teaching the software radio course on which this article is based, the author finds that Asian and U.S. engineers are less practiced in formal methods for specifying radio functions then their European counterparts. The European Telecommunications Standards Institute (ETSI) emphasis on formal methods and the widespread use of SDL in support of the European standards-setting

process has not permeated U.S. practice, particularly in military, civil, and other non-ITU marketplaces. As a result, U.S. practitioners of radio engineering are doing with generic computer-aided software engineering (CASE) tools what their European counterparts are doing with communications-oriented SDL — defining new radio air interface standards and implementations. This article introduces the formal techniques and software tools needed to effectively develop radios of the next-generation level of complexity.

In addition, software radio has become an industry focus area. Several texts now provide further background reading [3], industry perspectives [4], and in-depth treatment of architecture [5]. This article therefore highlights the numerous approaches with references for the interested reader.

This section introduced SWR and SDR. The next section summarizes the expanding role of SWR in contemporary telecommunications. The subsequent section reviews the fundamental precepts of the ideal SWR: the placement of the analog-to-digital converter (ADC) near the antenna, and the use of software to replace formerly analog or digital hardware. Section 4 introduces implementation-dependent SDR architecture, based on defining functions, components, and design rules that guide SDR product migration. Section 5 examines practical SDR designs, emphasizing implementation constraints. Isochronous performance of the real-time software, for example, is a critical design constraint in the signal processing streams. Section 6 reviews the development parameters and risks associated with choosing designs from the conservative baseband DSP through a variety of SDR alternatives to the ideal SWR. Many SDR projects of the 1990s failed or fell substantially short of requirements because of unanticipated software complexity. Section 7 surveys the broader implications of SWR. Importantly, SWR brings the substantial mitigation of the so-called shortage of radio spectrum, which is an issue more of the economics of spectrum reuse infrastructure than of the laws of physics. Section 8 underscores a few conclusions. A list of acronyms is provided for the reader's convenience, while references are appended in the Bibliography.

2. THE TRANSFORMATION OF RADIO ENGINEERING

We are in the midst of a transformation of radio systems engineering. Throughout the 1970s and 1980s, radio systems migrated from analog to digital in system control, source and channel coding, and baseband signal processing. In the early 1990s, the SWR transformation began to extend these horizons by soft-coding traditionally hardware-defined characteristics of wireless devices, including

- Radiofrequency (RF) band
- RF channel broadband channel coding and bandwidth
- Diversity, intermediate-frequency (IF) combining, beamforming, and antenna characteristics [6]

Today the evolution toward practical SDR is accelerating through a combination of techniques. These include multiband antennas and wideband RF devices. Wideband ADCs and digital-to-analog converters (DACs) now affordably access GHz of spectrum instantaneously. Multiband radios for military, commercial, teleinformatic, intelligent transportation, aircraft, and other applications are increasingly affordable. Processing of IF, baseband, and bit streams is implemented using increasingly general-purpose programmable processors. The complexity of the physical and link-layer protocol software for the many diverse RF bands and physical-layer modes of third- and fourth-generation wireless (3G/4G) now extends to millions of lines of code.

The ideal SWR consists of wideband antenna, wideband ADC/DAC, and general-purpose processor(s). Although SWR configurations are of research interest, they do not meet market constraints. The SDR therefore embraces the evolution of programmable hardware, increasing flexibility via increased programmability within market constraints. The ideal SWR represents and end state of maximum flexibility in this evolution. SDR "futureproofs" practical infrastructure against continually evolving standards and hardware. Strong SWR architecture permits one to insert SDR technology gracefully and affordably. For a clear path for product evolution, one must adapt SWR to specific applications or market niches. The attempts of researchers to build the ideal SWR yield lessons learned from technology pathfinders such as SPEAK*easy* [7,8], FIRST [9,10], and TRUST [11]. Continuing transformation is evident in the creation of the Wireless World Radio Foundation (WWRF) and SDR Forum work in SDR, with global emphasis on SDR as a critical enabler of 3G and 4G wireless. Related work continues to expand in DARPA, the Americas, the ITU, the EC (e.g., GSM MoU and ETSI), and Asia (e.g., ARIB and IEICE).

3. THE IDEAL SOFTWARE RADIO

The top-level components of an ideal SWR handset consist of a power supply, an antenna, a multiband RF converter, a power amplifier, and a single-chip ADC and DAC. These components plus on-chip general-purpose processor and memory perform the radio functions, as illustrated in Fig. 1.

The ideal SWR mobile wireless terminal interfaces directly to the user (e.g., via voice, data, fax, and/or multimedia) and to multiple RF "air" interfaces. Driven by user demands, the mobile unit minimizes dissipated power for long battery life and minimizes manufacturing parts count by maximizing hardware integration for lower unit cost. The generic wireless base station accesses multiple radio air interfaces and the Public Switched Telephone Network (PSTN). With access to the power grid, base stations may employ computationally intensive software designs using modular, open-architecture processing hardware that facilitates technology insertion. Technology insertion futureproofs base-station infrastructure against the continuing evolution of air interfaces. Military base stations ("nodes") need to support multiple networks on

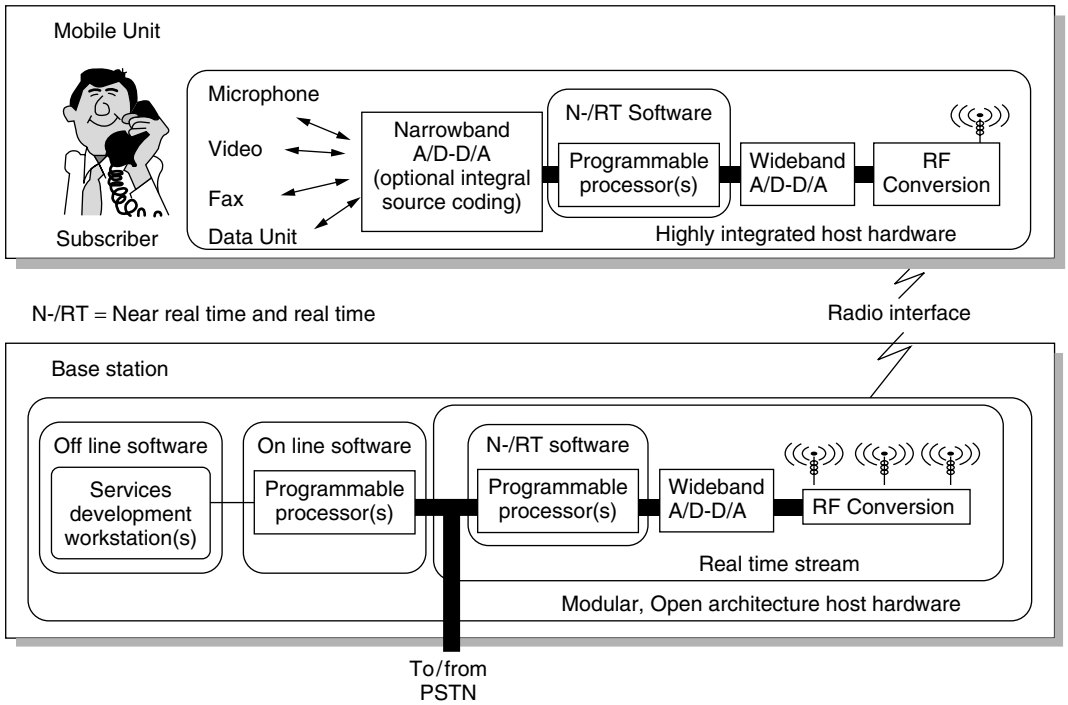


Figure 1. The ideal software radio handset and base station.

multiple RF bands with multiple air interfaces ("modes"). Command-and-control communications (C3) nodes may be formed by the collocation of several radios on mobile vehicles. These configurations often interfere with each other. The military calls this "cosite interference." The SWR basestation attempting to support push-to-talk (PTT) traffic on multiple channels in the same band can also generate self-interference. With software downloads to evolve radio personalities, the management of self-generated interference that was a design issue for the radio engineering laboratory of the 1980s is a deployment-time configuration management issue for SWR.

The placement of the ADC and DAC as close to the antenna as possible and the definition of radio functions in software are the hallmarks of the SWR. Although SWRs use digital techniques, software-controlled digital radios are not ideal SWRs. The essential difference is the total programmability of the SWR, including software-defined RF bands, channel access filters, beamforming, and physical-layer channel modulation. SDR designs, on the other hand, liberally mix analog hardware, digital hardware, and software technologies. SDR has become practical as DSP costs per millions of instructions per second (Mips) have dropped below $10 and continue to plummet. The economics of software radios become increasingly compelling as demands for flexibility increase while these costs continue to drop by a factor of 2 every 1.5–3 years.

By April 2002, SDR technology cost-effectively implemented commercial 1G analog and 2G digital mobile cellular radio air interfaces. 3G SDR base stations were being developed. Over time, wideband 4G air interfaces will also yield to software techniques on wideband RF platforms. In parallel, multiband multimode military radios were being deployed, such as the digital modular radio (DMR) [12], and the Joint Tactical Radio System (JTRS) [13] "clusters."[1] Such SDR implementations require a mix of increasingly sophisticated software technology along with hardware-intensive techniques such as ASICs.[2]

3.1. The Ideal Functional Components

Technology advances have ushered in new radio capabilities that require an expansion of the canonical communications functional model: source coding, the channel, and channel coding. The new aspects are captured in the software radio functional model. First, multiband technology [14], accesses more than one RF band of communications channels at once. The RF channel, then, is generalized to the channel set of Fig. 2. This set includes RF channels, but radio nodes such as personal communications system (PCS) base stations and portable military radios also interconnect to fiber and cable; therefore these are also included in the channel set. The channel encoder

[1] In this article, the conventional notion of cellular radio is extended to embrace the idea that the propagation of RF from any SDR transmitter defines an implicit RF cell. Its size and shape is determined by the physical placement of antenna(s) and the environment. Antenna height, directivity, path loss, diffraction, and multipath loss shape the cell. A multiband, multimode SDR is uniquely suited to turn such implicit cells into explicitly managed ad hoc cellular networks.

[2] In fact, the continuing interplay among military and commercial software radios plays an important role in the evolution of SDR technology. The merger of these market segments around common interest in open architecture SDR platforms is an on-going process, complete with the common interests and discontinuities.

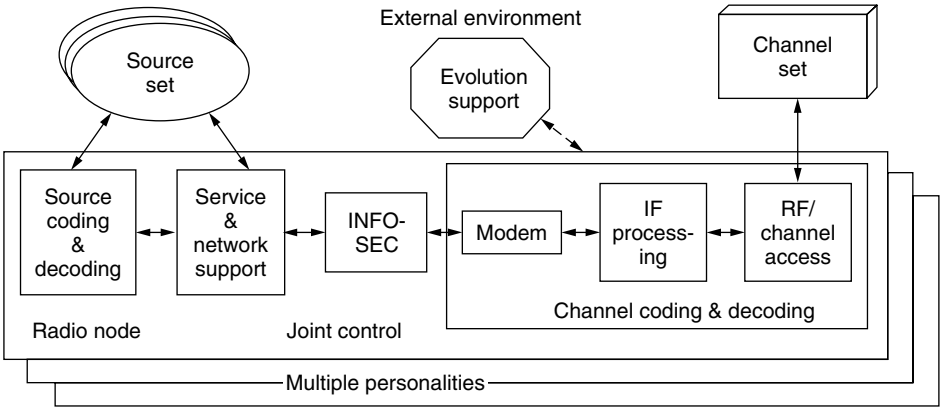


Figure 2. Functional model of a software radiocommunications system.

of a multiband radio includes RF/channel access, IF processing, and modem. The RF/channel access includes wideband antennas, and the multielement arrays of smart antennas [15]. This segment also provides multiple signal paths and RF conversion that span multiple RF bands. IF Processing may include filtering, further frequency translation, space/time diversity processing, beamforming, and related functions. Multimode radios [7] generate multiple air interfaces (also called "waveforms") defined principally in the modem, which is the RF channel modulator–demodulator. These waveforms may operate only in specific RF bands or may span multiple bands. A software-defined personality includes RF band, channel set (e.g., control and traffic channels), air interface waveforms, INFOSEC, network interfaces, and related user interface functions. In this abstraction of information source, the user is just another source (/sink) set.

Although few applications require Information Security (INFOSEC), there are incentives for its use. Authentication reduces fraud. Stream encipherment ensures privacy. Both help assure data integrity. Transmission security (TRANSEC) hides the fact of a communications event (e.g., by spread-spectrum techniques [16]). INFOSEC is therefore included in the functional model, although this function may be null for some applications.

In addition, the source coding/decoding pair of the expanded model includes the data, facsimile, video, and multimedia sources essential for new services. Some sources will be physically remote from the radio node, connected via the synchronous digital hierarchy (SDH), a local-area network (LAN) [17], and other systems, through the Service & Network Support shown in Fig. 2.

These functions may be implemented in multithreaded multiprocessor software orchestrated by a joint control function. Joint control assures system stability, error recovery, timely dataflow, and isochronous streaming of voice and video. As radios become more advanced, joint control becomes more complex, evolving toward autonomous selection of band, mode, and data format in response to implicit user needs. An autonomous software radio capable of machine learning is called a "cognitive radio." Cognitive radio requires a knowledge processing architecture in the joint control function, overlaid on the SWR architecture discussed in this article [18].

Any of the radio node functions may be singleton (e.g., single band vs. multiple bands) or null, further complicating joint control. Agile beamforming supports additional users and enhances quality of service (QoS) [19]. Beamforming today requires dedicated processors, but in the future, these algorithms may timeshare a DSP pool along with, for instance, a rake receiver [20] and other modem functions. Joint source and channel coding [21] also yields computationally intensive waveforms. Dynamic selection of band, mode, and diversity as a function of QoS [22] introduces large variations into demand, potentially causing conflicts for processing resources. These resources may include ASICs, field-programmable gate arrays (FPGAs), DSPs, and general-purpose computers. Channel strapping, adaptive waveform selection, and other forms of data-rate agility [23] further complicate the statistical structure of the computational demand. In addition, processing resources may be unavailable because of equipment failure [24]. Joint control therefore integrates fault modes, personalities, and support functions, mapping the highest priority radio functions onto the available processing resources. to yield a reliable telecommunications object [25].

In a software radio, the user can upload a variety of new air interface personalities [26]. These may modify any aspect of the air interface, including how the carrier is hopped, the spectrum is spread, and beams are formed. The required radio resources are RF access, digitized bandwidth, dynamic range, memory, and processing capacity. Resources used must not exceed those available on the radio platform. Some mechanism for evolution support is therefore needed to define the waveform personalities, to download them (e.g., over the air) and to ensure that each personality is safe before being activated. Evolution support therefore must include a software factory. In addition, the evolution of the radio platform — the analog and digital hardware of the radio node — must also be supported. This may be accomplished via the development of customized hardware/firmware modules, or by the acquisition of commercial off-the-shelf (COTS) modules, or both.

The block diagram of the radio functional model is a partitioning of the blackbox functions of the ideal SWR node into the functional components shown in Fig. 2, characterized further in Table 1.

Table 1. Function Allocation of the Software Radio Functional Model

Functional Component	Allocated Functions	Remarks
Source coding and decoding	Audio, data, video, and fax interfaces	Ubiquitous algorithms (e.g., ITU [27], ETSI [28])
Service and network support	Multiplexing; setup and control; data services; internetworking	Wireline and Internet standards, including mobility [29]
Information security[a]	Transmission security, authentication, nonrepudiation, privacy, data integrity	May be null, but is increasingly essential in wireless applications [30]
Channel coding/decoding: modem[a]	Baseband modem, timing recovery, equalization, channel waveforms, predistortion, black data processing	INFOSEC, modem, and IF interfaces are not yet well standardized
IF processing[a]	Beamforming, diversity combining, characterization of all IF channels	Innovative channel decoding for signal and QoS enhancement
RF access	Antenna, diversity, RF conversion	IF interfaces are not standardized
Channel set(s)	Simultaneity, multiband propagation, wireline interoperability	Automatically employ multiple channels or modes for managed QoS
Multiple personalities[a]	Multiband, multimode, agile services, interoperable with legacy[b] modes	Multiple *simultaneous* personalities may cause considerable RFI[c]
Evolution support[a]	Define & manage personalities	Local or network support
Joint control[a]	Joint source/channel coding, dynamic QoS vs. load control, processing resource management	Integrates user and network interfaces; multiuser; multiband; and multimode capabilities

[a]Interfaces to these functions have historically been internal to the radio, not plug-and-play.
[b]"Legacy" refers to modes that are deployed but may be deprecated.
[c]Radiofrequency interference.

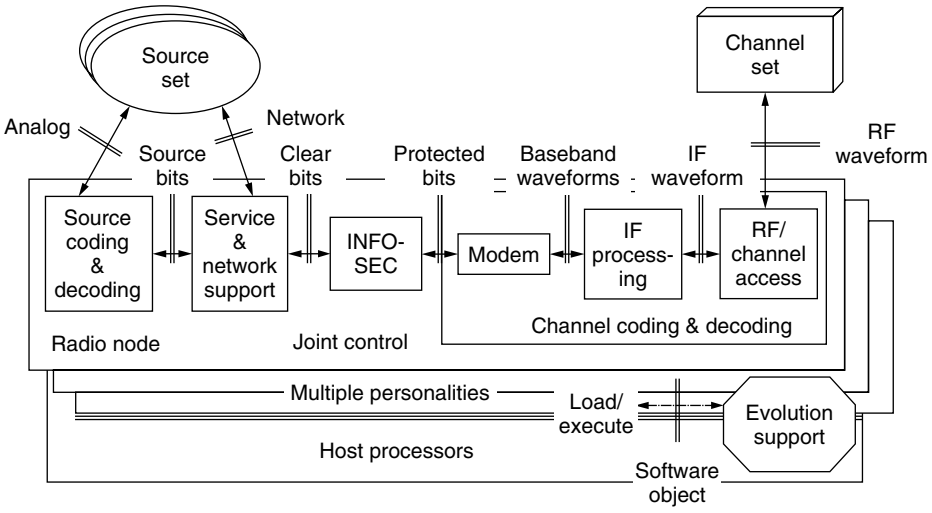


Figure 3. Standard interface points facilitate development, deployment, and evolution.

Not every implementation needs all sub-functions of this functional model. Thus, one may consider the functional model to be a point of departure for the tailoring of SDR implementations.

3.2. The Ideal Functional Interfaces

After identifying the functions to be accomplished in a software radio, one must define the interface points among the functional components. Figure 3 identifies these interfaces. The notation "RF waveform" includes spatial beamforming and the air interface. The IF waveform consists of signals filtered and converted to an intermediate carrier frequency that facilitates analog filtering, signal conditioning, and related analog processing.

In addition, IF processing may include A/D and D/A conversion. If so, some IF processing may be implemented digitally. Baseband waveforms are usually digital streams (e.g., of message packets or coded voice). These digital streams may also be sampled replicas of analog signals, such as digitized FM broadcast waveforms. The modem delivers what may be called *decoded channel bits* that may be encrypted ("black" bits in INFOSEC jargon) to the INFOSEC function if one is present. The modem transforms IF signals to channel bits. INFOSEC then transforms these protected bits into unencrypted ("clear")

bits (also called "red" bits). These bits may be manipulated through a protocol stack in order to yield source or network bit streams. Network bit streams conform to a network protocol, while source bits are appropriate for a source decoder. The interface to local sources of voice, music, video, and other media includes an analog transducer. Access to remote sources is accomplished via the network interface. In addition to these signal processing interfaces, there are control interfaces mediated by the user or network (both of which are source sets).

Personalities are downloaded to the radio via the software object interface. The simplest mechanism for maintaining radio software after deployment is the downloading of a complete binary image of the radio. A more flexible approach allows one to download a specific new function such as a specialized voice coder (vocoder). Such incremental downloads conserve network bandwidth at the expense of increased risk of configuration errors in the download process.

The interfaces thus defined may be thought of as the "horizontal" interfaces of the software radio, since they are concatenated to create the signal and control flows through functional components between sources and channels. These interfaces are further characterized in Table 2. Design-level interfaces ("design to") specify the data to be exchanged, while code-to interfaces directly or indirectly specify the exact format and meaning of each hardware signal, bit, word, byte, and message sequence of the interface.

In traditional radio engineering, these interfaces were addressed primarily in the design and development of the radio. For plug-and-play in software radio, they must be open architecture standards that facilitate the insertion of third-party components in deployment and operations. This is the business model that made the IBM PC a commercial success. How can such interfaces be standardized for industrywide third-party plug-and-play business to grow?

4. SOFTWARE-DEFINED RADIO (SDR) ARCHITECTURE

The *Random House Unabridged Dictionary* defines architecture as "a fundamental underlying design of computer hardware, software, or both" [38]. While this is an agreeable definition, it provides no prescription of what "underlying design" entails. The IEEE prescribes that architecture consists of components and interfaces. This leaves undefined what the components and interfaces are supposed to accomplish. The ideal software radio functional model and interfaces of the previous section begin to specify the functions of the architecture. The Defense Information Systems Agency (DISA) is the U.S. Department of Defense (DoD) agency charged with

Table 2. Top Level Component Interfaces

Interface	Characteristics	Properties
Analog stream	Audio, video, facsimile streams	Continuous, infinite dimensional; filtering constraints are imposed here
Source bit stream	Coded bit streams and packets; ADC, vocoder, text data compression [31]	Includes framing and data structures; finite arithmetic precision defines a coded, Nyquist [32] or oversampled dynamic range[a]
Clear bit stream	Framed, multiplexed, forward-error-controlled (FEC) bit streams and packets	FEC imparts algebraic properties over the Galois fields defined by these bit streams [33]
Protected bit stream	Random challenge, authentication responses; public key; enciphered bit streams [34] and packets	Finite-dimensional; randomized streams, complex message passing for downloads; if null, this interface reverts to clear bits
Baseband waveform	Discrete-time synchronous quantized sample streams (one per carrier)	Digital waveform properties determine fidelity of analytic representation of the signal
IF waveform	Composite, digitally preemphasized waveform ready for upconversion	Analog IF is continuous with infinite dimensions; digital IF may be oversampled
RF waveform	Power level, shape, adjacent channel interference, etc. are controlled	Analog RF: channel impulse response, spatial distributions via beams and smart antennas [35]
Network interface	Packaged bit streams may require asynchronous transfer mode (ATM), SS7, or ISO protocol stack processing	Synchronous digital hierarchy (SDH), ATM, and/or signaling system 7 (SS7)
Joint control	Control interfaces to all hardware and software; initialization; fault recovery	Loads binary images, instantiates waveforms, manipulates control parameters; cognitive joint control learns user needs [36]
Software objects	Download from evolution support systems	Represents binary images, applets; includes self-description of system capabilities [e.g., 37]
Load/execute	Software object encapsulation	Downloads require authentication and integrity

[a]A coded dynamic range is defined by the vocoder. Nyquist dynamic range results when an analog signal is sampled so as to meet the Nyquist criteria for bandwidth recovery of the sampled signal and has been quantized with sufficient bits of sufficient accuracy to represent the two-tone spurious-signal-free dynamic range of the application. Oversampling above the Nyquist rate can yield additional dynamic range through processing gain.

defining architecture. DISA defines architecture in terms of profiles for communications standards [39], defining architecture by analogy to "zoning laws and building codes" that constrain the construction of residential and industrial buildings [40].

4.1. Functions, Components, and Design Rules

None of the many possible definitions of architecture completely suits the architecture needs of the SDR community. One is needed that relates services, systems, technology, and economics. "Architecture" for SDR is therefore defined as a comprehensive, consistent set of *functions, components* and *design rules* according to which radio (and/or "wireless") communications systems may be cost-effectively organized, designed, constructed, deployed, operated, and *evolved over time*. This definition of architecture is consistent with the other definitions, but addresses more clearly the needs for plug-and-play and component reuse. By including functions and design rules, architecture supports component reuse, spanning component migration among hardware and software implementations.

The *design rules* must assure that when SDR hardware and software components are mated, the resulting composite entity accomplishes the intended functions within the performance bounds established by regulatory bodies, service providers, and users. The abstract functions and interfaces of the ideal SWR above constitute a horizontal architecture for software radio, an abstract functional flow from user to antenna and back. The ideal SWR does not specify the physical arrangement of physical components of an actual radio, of an SDR. Vertical levels are also needed to manage SDR hardware platforms and to achieve platform-independence of increasingly complex SDR software.

4.2. Plug-and-Play

In order for SDR architecture to support plug-and-play, design rules must be published that permit hardware and software from different suppliers to work together when plugged into an existing system. Hardware modules will plug-and-play if the physical interfaces and logical structure of the functions supplied by that module are compatible with the physical interfaces, allocation of functions, and related design rules of the host hardware platform. Software modules will plug-and-play if the individual modules and the SDR configuration of modules are computationally stable. For this, there must be a comprehensive interface to the host environment, and the module must describe itself to the host environment so the component can be managed as a radio resource. SDR architecture, then, defines the partitioning of functions at appropriate levels of abstraction so that software functions may be allocated to software components at appropriate levels of abstraction. SDR architecture defines the design rules, including design patterns [41,42] and interface standards. Rules defining logical levels of abstraction hide irrelevant details in the lower layers. These rules comprise a vertical architecture for SDR.

4.3. Vertical Architecture Design Rules

While horizontal architecture applies to any software radio, vertical architecture applies to radio implementations, SDRs. SDR components do not all share the same logical level of abstraction. A DSP module, for example, is part of the SDR platform. CORBA facilities are part of the software infrastructure on which ideal SWR functions are built using practical SDR software modules. In an advanced SDR, a modem may be a software module, a radio application. Bridging from one air interface to another is a service built on air interface radio applications. For example, in a military disaster-relief scenario, a SDR may bridge the Global System for Mobile communications (GSM) [43] to a military air interface like SINCGARS or HAVE QUICK [13]. One must therefore identify the SDR levels of abstraction that naturally partition the hardware and software into radio platforms, middleware,[3] radio applications, and communications services, as illustrated in Fig. 4.

In digital radios of the 1980s, the radio hardware platform ("radio platform") accomplished most of the RF and IF radio functions in hardware. The RF and IF parameters could be set through a microprocessor from a simple

[3] *Middleware* is software that insulates applications from the details of the operating environment (e.g., the hardware).

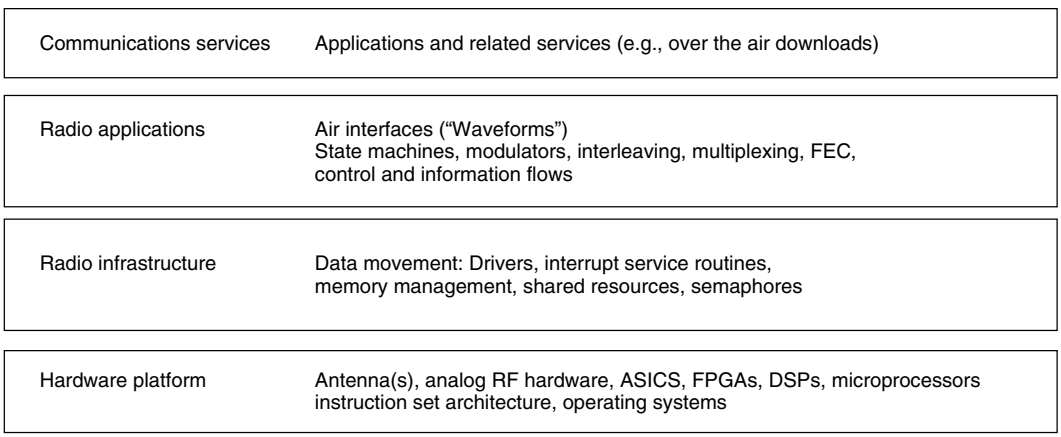


Figure 4. Logical levels of abstraction of the SDR implementations.

user interface or a low-speed data bus. Today's SDR platforms embody GFLOPS of processing capacity that host hundreds of thousands of LoC. This software performs the three top-layer functions of Fig. 4. At the infrastructure level, the code moves data among the distributed multiprocessors of the radio platform. At the applications level, software processes thus distributed cooperate to form radio applications, such as a 3G cellphone (cellular telephone) standard or a military waveform like HAVE QUICK. At the highest level of abstraction, applications software delivers communications services to users. Radio applications may incorporate specialized air interface protocols, and also may employ standard wireline data exchange protocols like TCP/IP.

One must define interfaces among these levels of abstraction, such as using an applications programming interface (API). APIs may map from one horizontal layer to the next. The API calls may be thought of as the vertical interfaces among horizontal layers. This approach has been used with reported success on SWR technology pathfinders [8]. The four layers of abstraction defined above are useful for defining software–software and software–hardware interfaces. Not all API's conform to these four layers. However, they are architecture anchors that help organize the process of evolving SDR implementations.

4.4. Mathematical Structure

Some mathematical principles illuminate the path toward SDR architecture. Some key principles are based on computability and point-set topology [44]. Consider transceiver state, consisting of a set of labels such as "Idle," "Synchronizing," "Receive," "Transmit," and "CarrierFault," asserted by other algorithms such as SquelchDetection (e.g., no squelch means the channel is Idle). SquelchDetection has possible wait states that have topological structure [45]. There is a set (e.g., of state labels, process names) with a family of subsets (e.g., the ones over which the software operations are valid), which has [or fails to have] topological properties. In addition, SDR architecture regarded as a collection of SDR implementations ("instances") has topological structure [46]. Maps over topological spaces define critical mathematical properties such as SDR module composability per the gluing lemma [47]. If the SDR has strong topological structure, then the insertion of plug-and-play components preserves the composability of software modules, including isochronism and controllability [48]. If not, then those critical properties cannot be guaranteed. In the absence of mathematical properties, one must test all possible configurations of SDR modules, a computationally intractable task for merely a few dozen downloadable SDR modules. Industry standards facilitate the application of such mathematical principles to broad classes of SDR.

4.5. Industry-Standard SDR Architectures

An important evolutionary step in the definition of vertical SDR interfaces is use of middleware (e.g., CORBA [49]) in SDR architecture. The Object Management Group (OMG) has defined an Interface Definition Language (IDL) in their Common Object Request Broker Architecture (CORBA). CORBA was developed primarily to define interfaces among software modules that were not originally designed to work together. IDL provides facilities for defining interfaces among software components through the mediation of an Object Request Broker (ORB). Since each new component implements only one interface to the ORB rather than N interfaces to the existing components, the process of integrating a new software component is greatly simplified.

The JTRS JPO began the development of its CORBA-based Software Communications Architecture (SCA) [50] in 1997. Version 2.2 (Fig. 5) includes the architecture specification with supplements on military security,

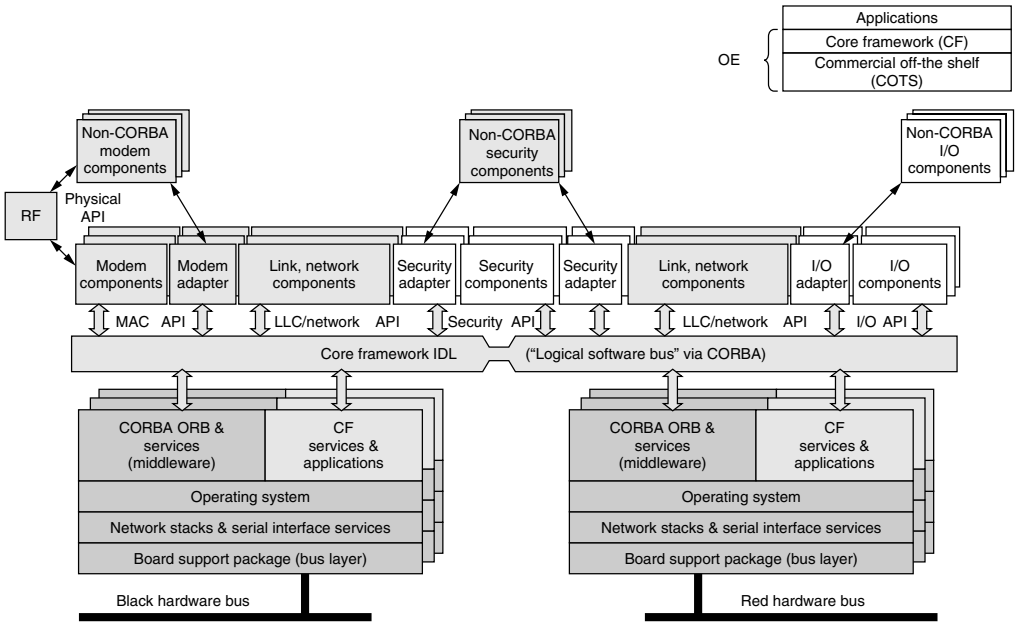


Figure 5. SCA Version 2.2.

APIs, and rationale. The framework applies UML to define hardware devices, software objects, and related interface rules. Its operating environment includes a core framework (CF) consisting of

1. *Base application interfaces* used by all software applications (Port, LifeCycle, TestableObject, PropertySet, PortSupplier, ResourceFactory, and Resource)
2. *Framework control interfaces*, used to control the system (Application, ApplicationFactory, Domain Manager, Device, LoadableDevice, ExecutableDevice, AggregateDevice and DeviceManager)
3. *Framework services interfaces* that support both core-compatible and non-core-compatible (vendor-unique and hardware-defined) applications (File, FileSystem, FileManager, and Timer)
4. A *domain profile* that describes the properties of hardware devices (DeviceProfile) and software components (SoftwareProfile) of the radio system using XML

Since this architecture does not yet map radio communications functions (e.g., 3G or military standards like HAVE QUICK) to its functional model, it provides a framework, an important first step toward plug-and-play architecture. In addition, its code-to level of specification is limited to the XML descriptions of interfaces. Therefore, different "fully compliant" implementations of a given air interface from two different vendors do not necessarily interoperate. For example, one implementation might specify RF in kilohertz while the other specifies it in Hz. With no units-consistency checking or remapping, the different software components would not use the facilities of the RF platform consistently. The government, academic, and industry bodies developing this standard plan to continue to evolve it towards an open-architecture plug-and-play standard.

5. SDR DESIGNS

Design-to and code-to architecture design rules assure that the critical properties of software radios are met as plug-and-play components are configured. Among these is the computational stability of the integrated software, a mathematical property emphasized above. Next is isochronism, the sufficiently precise timing of the real-time signal processing streams. Consider first the signal streams of an ideal SDR, illustrated in Fig. 6. These include a real-time isochronous channel-processing stream, a near-real-time environment management stream, an online control stream to manage the radio's configuration and modes of operation, and radio personalities from offline evolutionary development.

5.1. Real-Time Channel Processing Streams

In practical SDR designs, the real-time channel processing stream is a signal structure within the channel coding/decoding function. In an aggressive SDR design, each broadband channel (e.g., a cellular band) is accessed via a wideband ADC and DAC. This aspect of SDR design merits particular attention. The real-time stream generates subscriber channels in the transmit-path and isolates them in the receive path, such as by

1. Filtering of frequency-division multiple-access (FDMA) [51] waveforms
2. Timing recovery of time-division multiple-access (TDMA) [52] waveforms
3. Despreading military spread-spectrum [53] or commercial code-division multiple-access (CDMA) [54] waveforms

Historically, subscriber channel isolation was allocated to analog IF processing (e.g., in first-generation FDMA cellphones) or ASIC hardware (e.g., a CDMA cellphone).

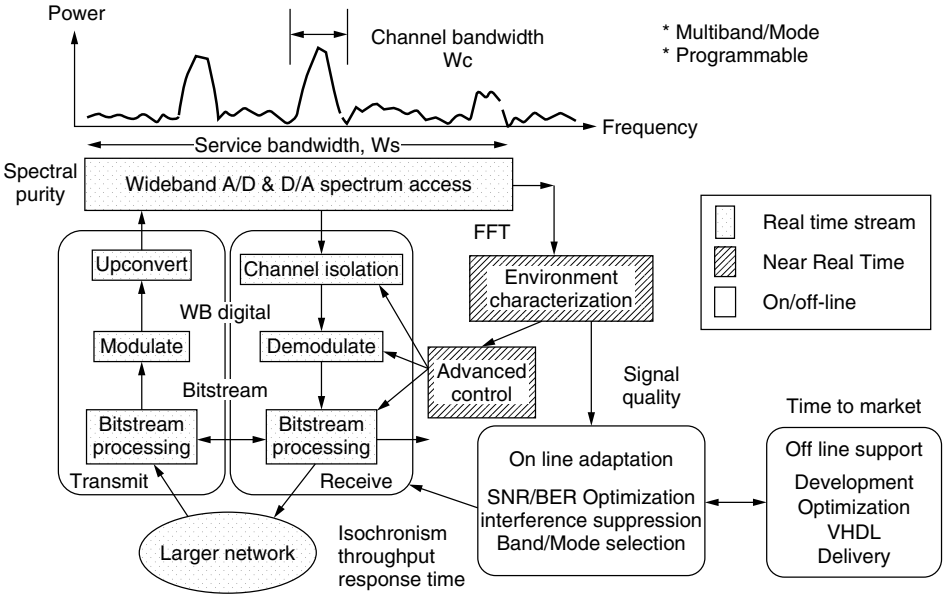


Figure 6. Signal processing streams of software radio.

When implemented in software, the isochronous processing windows for these functions are on the order of the time between DAC/ADC samples: microseconds to tens of nanoseconds. Modulation and demodulation of the channel waveform are also accomplished in the real-time channel-processing stream. Isochronous windows for these functions are on the order of tens to hundreds of microseconds. Once the narrowband subscriber bit streams are recovered, the timing of the isochronous windows increases to milliseconds. INFOSEC encryption and decryption, if applicable, are performed on the subscriber bit streams, with appropriately long isochronous windows — tens to hundreds of milliseconds. For baseband DSP, the time between digital samples (e.g., for baseband voice) is on the order of milliseconds to hundreds of microseconds. This allows plenty of time for processing between samples. In the software radio's IF stream, however, the time between samples is from tens of microseconds to hundreds of nanoseconds. Such point operations require Mflops to Gflops for isochronous performance. For isochronous performance, sampled data values must be computationally produced and consumed within short-buffer timing windows in order to maintain the integrity of the digital signal representation. Subscriber channels may be organized in parallel, resulting in a multiple-instruction multiple-datastream (MIMD) multiprocessing architecture [55]. Input/output (I/O) data rates of this stream approach 200 MB/s (megabytes per second) per IF ADC or DAC. Although these data rates are decimated through processing, to sustain isochronism blocks and events must be timed through I/O interfaces, FPGAS/ASICS and hard real-time embedded software in these streams.

To implement the approaches described above in a practical SDR design requires the integration of the real-time channel processing streams with related radio functions such as local oscillator (LO) signal generation. Figure 7 shows these radio signal flows structured into RF, IF, baseband, bit stream, and source segments, each with order-of-magnitude differences in isochronous windows. This view clarifies the sharing of the power management and low-noise amplifier (LNA) elements with RF conversion and with the RF frequency standard. These RF elements typically need physical proximity to the antenna. The LNA is placed near the antenna in order to set the system sensitivity. The power amplifier is near the antenna for power efficiency. The RF section may be remote from IF processing, such as in diversity architectures.

Digital IF processing in an SDR filters the wideband signal structure from the RF segment to yield the narrower baseband bandwidth. SDR ADCs appear at the IF–RF or RF–antenna interface. The baseband segment performs the modem functions, converting information between channel code and source code. The bit stream segment performs operations on bit streams, including multiplexing, demultiplexing, interleaving, framing, bit stuffing, protocol stack operations, and FEC. SDR system control is included in the bit stream segment because of the digital nature of control messages. The source segment includes the user, the local source and sink of information, and control. Source coding is the transformation of communications signals into bit streams if the design conforms to the ideal SWR functional partitioning described above. The organization of the design of SDR nodes into RF, IF, baseband, bit stream, and source segments promotes the application of a given talent pool and isochronous design discipline within a segment and minimizes the interdependencies between segments.

5.2. The Environment Management Stream

The other shaded boxes in Fig. 6 constitute the near-real-time environment management stream. In an ideal SWR, this stream continuously manages radio environment usage in frequency, time, and space. In a practical SDR design, the message exchanges with the host network are typically defined in specific signaling and multiple-access protocols. These traditionally include the assignment of traffic to clear channels, and the handoff of a mobile subscriber from one cell to the next. This may further include channel identification, equalization, and the estimation of parameters such as multipath time delays and cochannel interference levels. For example, the HF Automatic Link Establishment (ALE) protocol

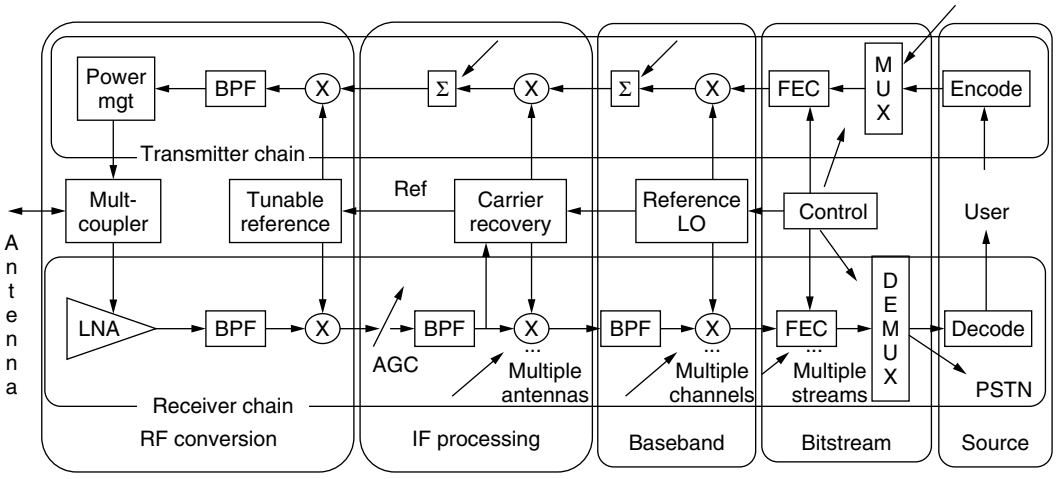


Figure 7. SDR isochronous and interdependent signal flows.

includes probes and responses that characterize several assigned channels [56]. The ALE data are then sent on the channel that is best for the specific subscriber location. In orthogonal FDM (OFDM), narrowband channels with excessive interference may be dynamically deallocated from use [57]. The environment management stream may employ block operations such as matrix multiplications for smart-antenna beamforming. Real-time adaptation by smart antennas may respond to signal parameters computed on every TDMA burst or CDMA symbol. GSM, for example, requires channel identification within 540 μs–2 ms. This establishes top-down limits on the execution time of environment management software. Forgiving operations, such as power-level updates may be refreshed every few frames. Location-aware services (e.g., emergency 911 cellphone location) typically define timing requirements for subscriber emitter location.

5.3. Online Adaptation: Mode Selection and Download Management

Online adaptation includes mode selection, as suggested in Fig. 6, as well as download management. A practical SDR is a multiband/multimode radio. The graceful transition from one RF band or air interface mode to another is called "mode handover." Military radios may change modes as a function of information priority, RF propagation, and other military criteria. An air interface mode typically defines the QoS provided by that mode. 3G air interfaces offer a wide range of data rates. Generally, high data rates require high signal-to-noise ratio (SNR[4]) for a required bit error rate (BER). Online adaptation selects the appropriate air interface mode to satisfy the competing goals of the user and/or of the network. Because of the number and complexity of 3G modes, ITU standards define mode handover in detail.

As modes become more elaborate, users are confronted with an increasing array of choices of QoS versus price. The burden of choosing RF band and mode in the future will be shared among the user, the network, and intelligent wireless appliances [e.g., personal digital assistant (PDA)], SDRs that employ natural-language processing and machine learning to assist the user with mode selection are called "cognitive radios" [18]. Because of Moore's law, cognitive wirelesses PDAs are likely to emerge soon, along with cognitive networks [58]. Cognitive PDAs and networks provide interesting cross-discipline research opportunities, fostering collaboration across natural-language processing, cognitive science, and radio engineering. In the past, mode control was primarily up to the network. As wireless LANs, home wireless networks, and intelligent transportation systems converge with cellular technology, cognitive PDAs will shape offered demand, for example, by delaying a large email attachment until the connection is free.

SDR personality management by over-the-network download also adapts the behavior of practical SDRs. Prior to SDR, the flexibility of RF access of a handheld wireless device was limited to merely choosing one of several predefined air interface modes. With SDR, parameters of the predefined modes may be modified along with higher-level software functions such as the user interface, network applications and air interface protocols.

[4] The SNR may be expressed in terms of unmodulated carrier and interference (CIR), or signal to interference plus noise (SINR).

5.4. Offline Adaptation: The Software Factory

Offline SDR development environments define SWR personalities. Offline functions include radio systems analysis, enhanced algorithms, hardware platform creation, and the rehosting of existing software to new hardware/software platforms. These functions assist in defining incremental service enhancements. For example, an enhanced beamformer, equalizer, and trellis decoder may be developed to increase subscriber density. These enhancements may be prototyped and linked into the channel processing stream in a research, testbed, or evaluation facility [59]. Such an arrangement allows one to debug the algorithm(s) and to experiment with parameter settings prior to deploying new personalities. One may determine the value of the new feature (e.g., in terms of improved subscriber density), as well as its cost. Network traffic to download such features constitutes overhead. Offline adaptation thus includes the definition of personalities [60], download protocols [61], and download traffic [62].

5.5. Software Tools

An advanced SDR does not just transmit a waveform. It characterizes the available transmission channels, probes the available propagation paths, and constructs an appropriate channel waveform. It may also electronically steer its transmit beam in the right direction, select the appropriate power level and pick an appropriate data rate before transmitting. Again, an advanced SDR does not just receive an incoming signal. It characterizes the energy distribution in the channel and in adjacent channels, it recognizes the mode of the incoming transmission, and it creates an appropriate processing stream. With a smart antenna, it also adaptively nulls interfering signals, estimates the dynamics of the multipath, coherently combines desired-signal multipath, and adaptively equalizes this ensemble. It may also trellis decode the channel modulation and then correct residual errors via FEC decoding to receive the signal with the lowest possible bit error rate (BER). Such operations require a family of software components and related tools, including those illustrated in Fig. 8.

Figure 8 organizes software tools according to the time-criticality of the supported software functions. In an SDR, hard real-time software may be delivered as the personality of an ASIC or FPGA. Reduced time-criticality means the function is more compatible with software implementation. The tradeoff among ASIC, FPGA, and software changes with each 18-month Moore's law cycle. The columns labeled C (criticality) and A (availability) identify SDR challenge areas. Bit interleaving, for example, is not challenging in terms of either its criticality to SDR architecture or its availability as a software component. Interference suppression, on the other hand,

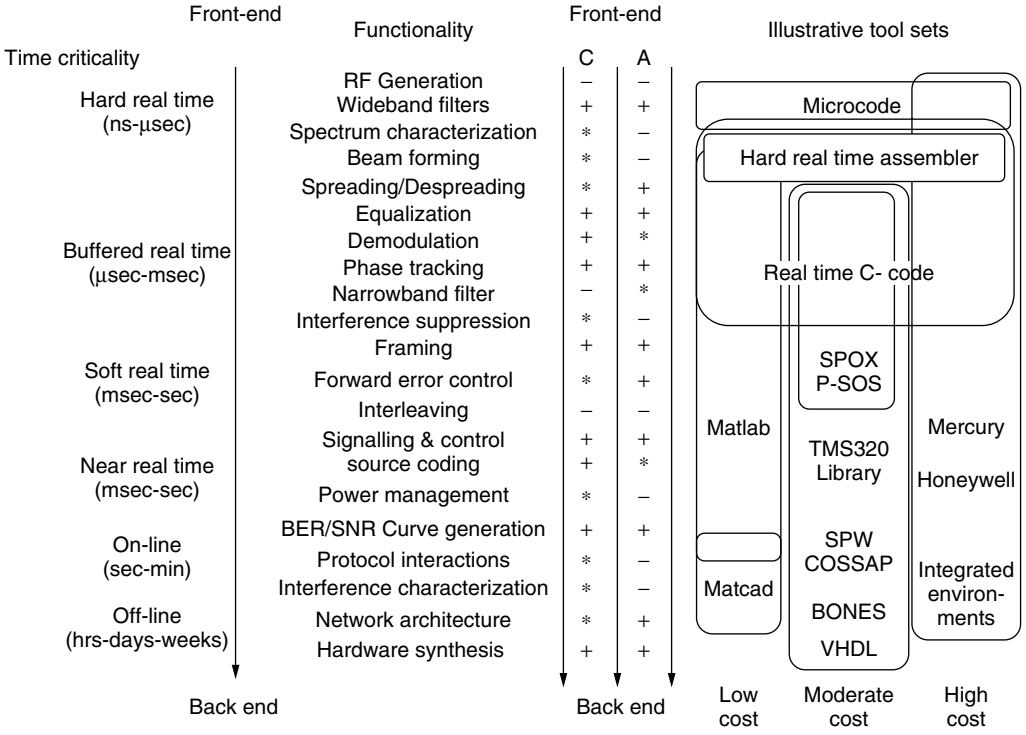


Figure 8. Toolsets of a software factory. (C = criticality; A = availability; $*$ = key performance driver; $+$ = important issue).

is a critical issue differentiating SDR from conventional radios; it also can be a performance driver. To the right are three columns of toolsets that represent the sophistication of the software factory. One may develop software radio products of limited scope [e.g., <40 kloc 40,000 lines of code] using the low-cost tools in the first column. As team size grows, or the mix of ASICs, FPGAs, and DSP hardware in the delivery environment becomes more complex, the investment of tens of thousands of dollars (per design seat) pays off. The largest, most complex systems benefit from the high-cost tool suites costing millions of dollars per system.

5.6. SDR Technology Alternatives

Technology alternatives for digital radios, SDR, and software radios are characterized in the software radio parameter space of Fig. 9. The parameter space compares two critical SDR technology parameters: digital access bandwidth and the flexibility of the processing platform. Digital access bandwidth is approximately half of the sampling rate of the widest bandwidth ADC in the isochronous signal-processing path. Thus, for example, an ideal SDR with 5 GHz conversion rate supports nominally a 2.5-GHz analog bandwidth, based on the Nyquist criterion [32]. Similarly, wideband digital-signal synthesis, digital upconversion, and wideband DAC yield an ideal software radio transmitter.

ADCs with continuous conversion bandwidths of >6 GHz have been built [63], although they are expensive. If all the processing after the ADC were accomplished on a single general-purpose computer, one would have an ideal software radio receiver (the point marked X in the figure). Using a rule of thumb of 100 operations per sample, the digital filtering of a 5-GS/s (gigasamples per second) stream to access a 25-MHz band of RF spectrum requires 500 gigamultiplications (5×10^{11}) per second. This processing capacity is about two orders of magnitude beyond 2002-generation DSPs [64,65] and three or four beyond general-purpose computers. This translates to about 6–10 Moore's law cycles or only 10–15 years of continued exponential development of DSP technology and an additional 5 years beyond that of general purpose computing technology.

Another limitation of the ideal SWR is that no single antenna nor RF stage can sustain the analog bandwidth from 2 MHz to 2.5 GHz RF with reasonable losses or power efficiency. The single wideband RF required for the 5 GHz ADC (and for the transmitter/DAC) is therefore not feasible. Antenna and RF stages depend on properties of materials that have stubbornly resisted pushing bandwidths beyond one RF decade, a 10 : 1 ratio of high to low RF. Thus, the ideal SWR is not possible with today's technology. The ideal properties of such a radio are a useful reference point for measuring progress towards generality and flexibility.

Practical SDR implementations limit RF coverage to medium or narrowband antennas, RF conversion, and IF processing technology. They also use a mix of digital technologies including ASICs, FPGAs, DSP, and general-purpose processors. Examples are illustrated in the figure. The STR-2000 (point *A* in Fig. 9) was an early baseband HF DSP radio developed by Standard Marine AB. This radio digitized its HF IF signal at a 24 kHz sampling

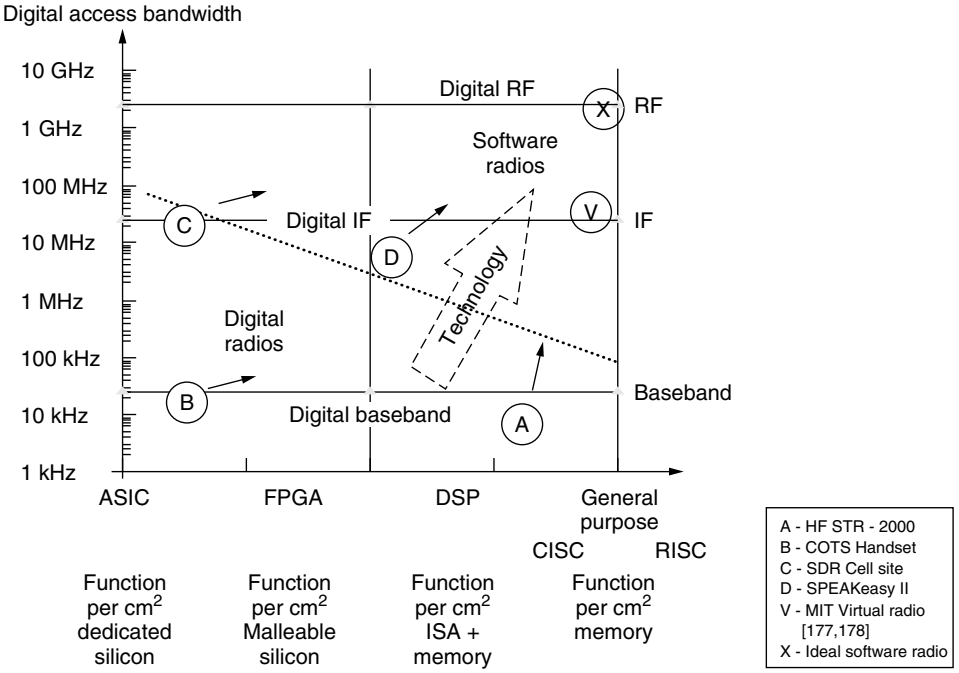


Figure 9. Software radio parameter space.

rate. It used twin Texas Instruments (TI) TMS320C30 DSPs to provide a half-dozen standard HF signal formats digitally. This is readily accomplished by amateur radio operators using general-purpose processors today [66,67]. Second-generation (e.g., GSM) COTS SDR handsets (point *B* in Fig. 9) minimize size, weight, and power using a direct-conversion receiver [68] RF-ASIC [69], and baseband DSP. Combining such ASICs in a handset enclosure for a dual-mode cellphone creates a "Velcro radio" [70]. Aggressive implementations incorporate high-density FPGAs to provide software-driven configurability in a delivery platform that maximizes throughput for a given technology clock rate [71]. 3G CDMA researchers employ FPGAs to replace despreader ASICs [72]. These FPGAs have greater speed and power efficiency for a given clock rate than do DSPs. In addition, the personality of an FPGA can be upgraded in the field by software download. On the other hand, FPGAs lack the silicon area and computational flexibility of the DSP instruction set architecture (ISA). With increasing chip density and area, system-on-chip (SoC) architectures of the near future are likely to include both fixed ISA and variable-personality FPGA coprocessors.

Contemporary software radio cell-site designs (point *C* in Fig. 9) access the allocated unlink[5] RF using a single ADC, such as with 25 MHz of analog bandwidth (viz., 70 MHz conversion rate). These designs employ a bank of digital filter ASICs [73] or parallel digital filters [74] to access a hundred or more subscriber channels in parallel. Research radios like the European ACTS Flexible Interoperable Radio System Technology (FIRST) used Pentek boards [75] with Harris parallel digital filter ASICs. The larger commercial cellular infrastructure suppliers include Alcatel, Ericsson, Fujitsu, Lucent, Motorola, Nippon Electric Corp (NEC), Nokia, NorTel, Siemens, and Toshiba. Although all contribute to open research, few publish the details of their commercial SDR handset or infrastructure products. Research supported by one or more of these industry leaders, explores ASIC [76], DSP [77], and FPGA [78] cell sites [79,80], some with smart antennas [81–82,83]. In addition, emerging products now include ADCs with 200 MHz of bandwidth with >80 dB dynamic range with integrated digital IF processing [84].

Technologically aggressive designs include SPEAK*easy*, the military technology pathfinder. SPEAK*easy* II (point *D* in Fig. 9), which became the baseline for Motorola's WITS 6000 software radio product line [85], incorporated over a Gflops of processing capacity for enhanced flexibility, substantial DSP in 1996–1998. The virtual radio (point *V* in Fig. 9) is the most flexible software radio research implementation reported in the literature [86]. A general purpose DEC Alpha processor running UNIX accesses a wideband IF digitally. Narrowband AM and FM broadcast receivers and an RF LAN were implemented purely in software on this platform. The related SpectrumWare software technology is being commercialized for military and commercial applications [87].

None of the designs *A*–*X* in Fig. 9 is a panacea: the architecture question is the degree of digital RF access and programmability required for the intended market. Contemporary radio designs therefore vary across the dotted line in the phase space. Advancing microelectronics technology moves all implementations inexorably upward and to the right over time. The three fundamental waveform limitations of any SDR implementation, then, are RF access, digital access bandwidth, and digital processing flexibility and capacity.

[5] The *uplink* is the link from mobile to base station. The *downlink* is the reverse link.

5.7. SDR Radio Reference Platform

The definition and use of a radio platform facilitates the evolution of SDR implementations through generations of hardware and software releases. It also enhances the use of UML, CASE tools, and middleware. A radio reference platform is a high-level characterization of the capabilities of the hardware environment of the software radio. Table 3 identifies the critical radio platform parameters that determine the performance of a software radio.

The parameters of Table 3 should be specified with precision. If the platforms in the family are tested for conformance to a well-specified reference platform, then software developed for one member of the family should port readily to another member of the family. The software will not port well (and may not port at all) if special features of the platform beyond the reference set are used. The specification of a minimum level of capability for each parameter defines a reference platform for a family of software radio implementations. Illustrative platforms are suggested in Table 4. The PDAs will have replaced conventional cell phones in this vision of the future. Given the reference platforms, they will have broadband RF, multiple parallel data channels, and wide digital processing bandwidth (BMW).

Devices now in development, mostly in proprietary settings promise to bring such platforms to market in the 2002–2007 timeframe. Such reference platforms closer to the ideal software radio are beginning to make economic and technical sense in infrastructure applications. A reference platform need not have an associated block diagram, but it is often convenient to use such a diagram in the analysis of the feasibility of a reference model.

The reference design of Fig. 10, illustrates the value of a reference platform, but has the following drawbacks. First, it implies that ADCs and DACs are the interface between the digital processing and analog RF sections of the radio. That is often the case, but an ultrawideband (UWB) [88] communications system, for example, uses subnanosecond pulses to spread the communications over 2 GHz or more of bandwidth. These pulses are both transmitted and received with analog circuits, not with DACs and ADCs. The primary value is to associate critical parameters with physical devices in such a way that one may outline an evolutionary path for software radio architecture.

6. DEVELOPMENT PARAMETERS AND RISKS

In 1992 when Mitola [89] introduced the term, almost nobody knew what a "software radio" was. By 1996, 6 months after the publication of the special issue of the *IEEE Communications Magazine* on the software radio, almost every radio vendor claimed to have one. The term had become an industry "buzzword." By 1999, it had become widely understood that nobody even *wanted* to offer an ideal software radio product because one would be unaffordable or inefficient or both. Thus, in 1996, the acronym SDR was introduced as the family affordable, practical implementations of software radio [90]. Today,

Table 3. Software Radio Reference Platform Parameters

Critical Parameter	Remarks
Number of channels	Number of parallel RF, IF, and/or baseband channels
RF access	Continuous coverage from a minimum to a maximum RF
Digital bandwidth	Bandwidth of the maximum ADC for each RF/IF channel
Dynamic range	End to end, including RF, IF, ADC, and processing gain
Interconnect bandwidth	Bandwidth of critical buses, serial ports, backplanes, etc.
Timing accuracy	The precision and stability of system clock(s)
Frequency performance	RF, IF, and local oscillator (LO) accuracy and stability
Processing capacity	Mips, Mflops using standard benchmarks, arithmetic precision (per processor class if appropriate)
Memory capacity	RAM, ROM per processor; mass storage capacity
Hardware acceleration	Parameterize capabilities encapsulated in hardware such as despreader ASICS, FPGAs, and related hybrids.
Operating environment	Operating system and related facilities (including CORBA middleware), interfaces (e.g., APIs), and measured determinism

Table 4. Illustrative Mobile SDR Reference Platforms

Notional Platform	RF Access (MHz)	Channels	Digital Bandwidth (MHz)
Lowband PDA	450–1200	3 (traffic, control, rental)	5
Midband PDA	850–2500	3 (traffic, control, rental)	20
Lowband military	30–500	4 (voice, 2 data, 1 scan)	10
Midband military	88–1200	4 (voice, 2 data, 1 scan)	20
Wideband military	800–4000	6 (4 JTIDS, 1 voice, 1 scan)	250

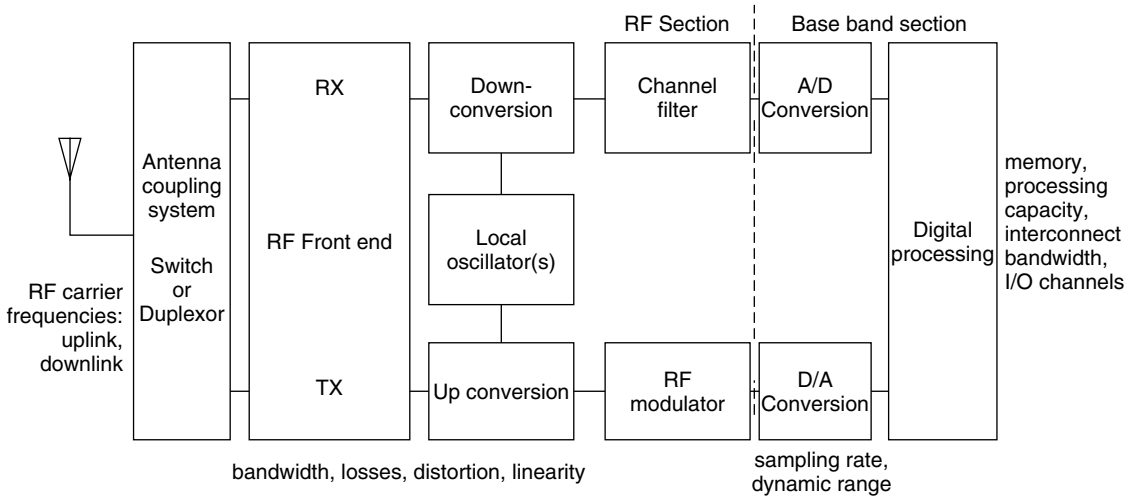


Figure 10. Reference design for an SDR implementation.

SDR is regarded as a ubiquitous technology that is key to the affordable evolution of military and commercial wireless markets. The four key characteristics of SDR from a development or acquisition perspective are

1. The number of air interface channels simultaneously supported (N)
2. The level of programmable digital access (PDA)
3. The degree of hardware modularity (HM)
4. The scope of software flexibility and affordability (SFA)

The number N defines hardware risk. There is low risk with single channel SDRs. With multiple channel (i.e., $N<6$), and full channel access (i.e., N is the full number of subscribers in an allocated RF band), risks increase. Two to four channel nodes are typical of military, civil aviation, and law enforcement hub applications. The full access class is typical of cellular base-station infrastructure. Single-channel SDR provides a baseline of minimum development risk and complexity. Multiple channel nodes require distributed multiprocessing. With small numbers of channels, hardware efficiency per channel is not a major challenge. It becomes a market-discriminator for the full access class, however. This class also carries maximum risk of mismatch between the processing demand offered by the software and the processing capacity deliverable by the hardware. Matching demand to capacity is therefore an essential design issue for practical SDR implementations.

The level of PDA is the point in the software radio functional model at which the conversion to digital occurs. This defines the scope over which the radio's functions are programmable with substantial flexibility. The types of PDA include baseband programmability, IF programmability, and RF programmability.

HM identifies the economic impact of the differences in hardware upgrade paths. Architecture may be based on capability-oriented coarse-grain (possibly programmable) modules such as receivers and exciters that are specific to an air interface. Alternatively, architecture may be based on technology-oriented coarse-grain modules such as COTS ADC and DSP boards. Finer grain modules such as FPGA, ADC and DSP chips are also candidate modules. Finally, the system-on-a-chip approach defines module as a chunk of intellectual property (IP). The granularity of hardware modularity is not prejudicial, but should be determined by the needs of the market segment.

SFA characterizes the service provider's ability to acquire plug-and-play software modules that are driven by a vital, multisupplier marketplace. Software that runs on just one radio platform and is available from only the original manufacturer tends to box the service provider into single-source (sometimes very expensive) maintenance and upgrade paths. If the functionality of the unit will not change over its lifecycle, then this may be a perfectly acceptable path. This would be a rare occurrence in today's fast-moving marketplaces, however. Software that runs on many platforms (e.g., Java) and is available from multiple vendors generally gives the service provider a better software product with more flexibility and at a lower cost over the lifecycle than the alternatives.

7. BROADER IMPLICATIONS

The prospect of a new technology of multiband, multimode software radios — handsets and infrastructure — has social and political implications. Type certification authorities, for example, are charged with administering the equitable use of radio spectrum. Among other things, they certify that radio equipment meets legally imposed constraints. In addition, software radios may operate on any RF band that is within the capabilities of the underlying radio platform, and with any mode for which a software load-image is available. This raises the possibility of truly novel approaches to spectrum management. One of the more interesting is the possibility that software radios could use a spectrum rental protocol to autonomously share spectrum. Another is that by incorporating advanced agent technology, they could evolve their own protocols. As mentioned previously, radios capable of such behavior are called "cognitive radios" [18].

7.1. Type Certification

The prospect of an evolving radio platform raises substantial questions among regulatory bodies about type certification. In remarks before the SDR Forum, the U.S. FCC [91] described type certification of software radios as presenting "regulatory issues." These include the following:

1. To which service(s) is an SDR approved?
2. Is a new approval needed for each "change" to an approved SDR unit?
3. How does the FCC enforce the equipment authorization rules for SDRs?
4. How can an unauthorized use of an SDR be prevented?

Regulators rely on a mix of tactics to achieve their goals. Industry is required to obtain licenses for some uses of spectrum, while others are available without a license, provided the manufacturer complies with the regulations. The FCC relies on legal remedies to motivate manufacturers to comply with the rules. They generally specify license requirements in terms of RF power output, modulation, occupied bandwidth, spurious emissions, and frequency stability (over temperature and voltage supply variations). Analog radios embody these parameters in hardware, so the type certification process has historically focused on the certification of devices. Digital radios, similarly, embody these parameters in a mix of analog and digital hardware, so the process remains valid. Current-generation SDRs with baseband programmable digital access embody these parameters in relatively fixed core images that are tightly coupled to the hardware, and this is compatible with the current process as well.

However, SDRs with at IF programmability embody these parameters in software that is loosely coupled to the hardware. Each combination of band and mode has to be certified separately, according to today's process. Over the air downloads to the SDR complicate the certification process substantially. At present, regulators in the United States are in the process of obtaining the advice of industry through an expected request for comments on proposed rule-making. Industry has the challenge of assisting regulators in defining a certification process that is responsive to the broader social and legal issues, but that does not seriously impede the benefits of SDR technology. Open architecture in some ways exacerbates the certification challenges. A proliferation of software packages enabled by open architecture drives the combinatorial complexity of type certification. Must a service provider certify every possible combination of software modules from every possible vendor? A helpful architecture might have properties that simplify and expedite type certification.

7.2. Incremental Download Stability and Type Certification

In addition to defining a partitioning, an architecture may define principles that assure that plug-and-play with desired properties of controllability and reliability. For example, to type-certify an open-architecture SDR, one must guarantee that the properties specified by the regulatory bodies will be preserved *in spite of the software radio's high degree of flexibility*. The need for such guarantees motivates the study of the mathematical properties of the software radio [92]. For example, one may model the statistical demand for computational resources versus processing capacity using queuing theory [93,94]. Real-time performance can be assured in a fixed architecture using this approach.

The plug-and-play SDR, however, has a *variable architecture* as modules are introduced into the environment and removed. This raises the complexity of the statistics, particularly in complex nodes. In a future 3G cell site, for example, hundreds of users can invoke dozens of variable-bandwidth services via a pool of shared DSP resources. To make this tractable, there should be a predictable relationship of computational demand between plug-and-play software modules and the host processor environment. This calls for a theory of plug-and-play resource bounds for the software radio within which such predictable relationships will exist. The fact that radio software must run to complete in a short, finite time period that can be specified in advance leads to a proof that radio software need not be Turing-computable [92]. The theory translates into a prohibition on unconstrained "while" and "until" loops. These have to be replaced by bounded-while and bounded-until loops that are allowed to run at most n times before generating a protection fault. The related theory of bounded recursion shows how a compiler can calculate n for the programmer so there is no additional programming burden to obtain this protection. Without such protection, while loops may run forever, consuming unacceptable amounts of time and processing power.

This theoretical advance makes it possible for one to provide a software engineering environment that can place tight upper bounds on the computational resources of an arbitrary radio software module. One may therefore prove by induction that a bounded recursive downloaded module will consume resources that are within tightly specified a-priori limits when loaded into a bounded recursive system. This can reduce the combinatorial complexity of the type certification of incremental software downloads. Given, for example, M vocoders and N air interfaces, a bounded recursive software system need test only $M + N$ software configurations, proving the other $MN - (M + N)$ configurations by induction. This supports the incremental download of the M vocoders, reducing download bandwidth on the network. Conventional software has to test all MN integrated load images. Furthermore, a change of vocoder requires the download of a complete load image, with increased network overhead. This article therefore sets forth the technical issues that underlie this tradeoff between network overhead and download complexity.

7.3. Spectrum Management Implications

Given that SDRs will continue to become more capable, one can ask whether they might have some fundamental impact on our approach to the use of the radio spectrum. A new research area, cognitive radio, suggests that this might indeed be the case [18]. Wireless multimedia applications require significant bandwidth, some of which

will be provided by third-generation 3G services. Even with substantial investment in 3G infrastructure, the radio spectrum allocated to 3G will be limited. Cognitive radio is a particular extension of software radio that employs model-based reasoning about users, multimedia content, and communications context. Cognitive radio offers a mechanism for the flexible pooling of radio spectrum using a new class of protocols called "formal radio etiquette." This approach could expand the bandwidth available for conventional uses (e.g., police, fire, and rescue) and extend the spatial coverage of 3G in a novel way. This section characterizes the potential contributions of cognitive radio to spectrum pooling and outlines an initial framework for formal radio etiquette protocols.

Figure 11 illustrates important aspects of spectrum allocation.

Bandwidth that could be made available for the sharing of spectrum, based on current allocations to mobile users, is summarized in Table 5.

The literature describes a protocol for spectrum rental among cognitive radios and infrastructure [95]. The effective use of this new protocol requires software radios that always know where they are (e.g., in latitude, longitude, and altitude above mean sea level), and which embed propagation models that include terrain and buildings. In addition, they must know what their users are doing (e.g., shopping, which is a low precedence use, or in need of emergency assistance, which is a high precedence use). Cognitive radios accomplish this by parsing all incoming and outgoing messages and voice traffic, and analyzing this information to establish the user's priority for use of spectrum. In addition, cognitive infrastructure can offer unused radio spectrum for rent for as little as one second in a microcell. Alternatively, rentals may allow use for minutes to hours in macrocells. The cognitive protocol includes listening for legacy radios to attempt to use the spectrum so that the cognitive radios may politely defer to legacy users. Police, for example, may require the renters to immediately yield the spectrum back to the renting authority. The protocol supports the return of spectrum within 30 ms. Throughput is enhanced if the legacy users can wait for $\geq$0.5 s before being guaranteed clear spectrum.

Table 5. Mobile Spectrum Pools

Band	RF_{min}	RF_{max}	W_c	Remarks
Very low	26.9	399.9	315.21	Long-range vehicular traffic
Low	404	960	533.5	Cellular
Mid	1390	2483	930	PCS
High	2483	5900	1068.5	Indoor and RF LANs

Although cognitive radios may not be practical for years to come, the research points in an interesting direction for spectrum managers. Instead of hard-allocations with primary and secondary users, the spectrum managers at some point in the not-too-distant future should be able to delegate the details of spectrum management to the radios themselves. The spectrum managers would then assume the higher-level task of specifying the rules the radios have to follow to ensure equitable access that conforms to social, political, and legal norms.

This article has provided an overview of software radios. It began top-down by introducing the functional model of the software radio. Next, it introduced the important aspects of software, especially the need for isochronism in multiband multimode radios that share a pool of processing resources among multiple users. A range of hardware implementations were introduced, to differentiated among digital radios, PDRs, SDRs, and ideal software radios. This led to the characterization of acquisition parameters that divide software radios into broad classes. Finally, broader implications were

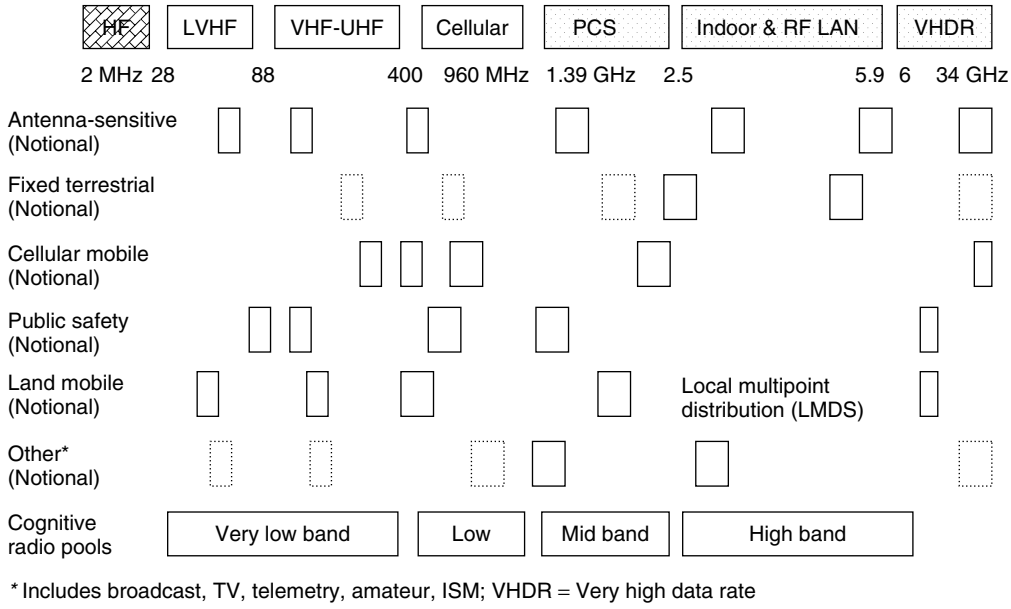


Figure 11. Potential spectrum pools.

presented, including the apparent challenge of type-certifying software radios. The chapter concluded with a view towards the future of the software radio — the evolution toward the cognitive radio.

8. CONCLUSION

Software radio has proved to be a valuable abstraction for the radio science, engineering, and user communities. During 2001, over 65 technical papers published in refereed journals of the IEEE and ACM included software radio or SDR as a theme. Although not a perfect metric, it indicates the degree to which the SWR/SDR abstraction is shaping research and technology development on a global scale. Practical SDR implementations continue to emerge with increasingly wider bandwidths and more tailored air interface capabilities. The challenges to the ideal software radio pose important research challenges, notably in the physics of antennas and RF conversion. The opportunities for innovative SDR implementations continue to attract substantial investments as the potential of this technology to reshape the physical layer of RF communications remains fertile ground.

ACRONYMS

3G/4G	Third- and fourth-generation wireless
ADC	Analog-to-digital-converter
ALE	Automatic Link Establishment, an HF air interface protocol
API	Applications programmer interface
ASICs	Application-specific integrated circuits
BER	Bit error rate
C3	Command-and-control communications
CASE	Computer-aided software engineering
CDMA	Code-division multiple access
CF	Core framework (of the JTRS/SDRF/OMG SCA)
CORBA	The Common Object Request Broker Architecture
DAC	Digital-to-analog converters
DARPA	The Defense Advanced Research Projects Agency
DISA	Defense Information Systems Agency
DMR	Digital modular radio
DoD	Department of Defense (U.S.)
DSP	Digital signal processor (or processing)
EC	The European Community, Brussels, Netherlands
ETSI	European Telecommunications Standards Institute
FDMA	Frequency-division multiple access
FEC	Forward error control
FPGA	Field programmable gate arrays
Gflops	Giga-floating-point operations per second
GHz	Gigahertz, 10^9 Hz
HF	High frequency, nominally 3–30 MHz
Hz	Hertz, cycles per second (e.g., of RF carrier frequency)
I/O	Input/output
IDL	Interface Definition Language
IEEE	Institute of Electrical and Electronics Engineers
IF	Intermediate frequency
INFOSEC	Information Security
IP	Internet Protocol, as in TCP/IP
ISA	Instruction set architecture
ITU	International Telecommunications Union
JTRS	Joint Tactical Radio System
kHz	Kilohertz, 10^3 Hz
LNA	Low-noise amplifier
LO	Local oscillator
loc	Lines of code
LVHF	Low VHF, typically 28–88 MHz
MHz	Megahertz, 10^6 Hz
Mips	Millions of instructions per second
OMG	Object Management Group
ORB	Object Request Broker
PCS	Personal communications system
PDA	Personal digital assistant
PSTN	Public Switched Telephone Network
PTT	Push to talk
QoS	Quality of service
RF	Radiofrequency
SCA	Software Communications Architecture
SDL	Specification and Description Language (ITU Standard Z.100)
SDR	Software-defined radio
SDRF	SDR Forum
SDR Forum	See *www.sdrforum.org*
SNR	Signal-to-noise ratio
SWR	Software radio
TCP	Transmission Control Protocol, usually used with IP, as in TCP/IP
TDMA	Time-division multiple access
UHF	Ultrahigh frequency, typically 300–3000 MHz
UML	The Unified Modeling Language
VHF	Very high frequency, typically 30–300 MHz
XML	The eXtensible Markup Language

BIOGRAPHY

Dr. Joseph Mitola III is an internationally recognized expert on software radio systems and technologies. In addition to having published the first paper on software radio architecture in 1992, he teaches in the United States, Asia, and Europe. He was Founding Chair of the SDR Forum in 1996 and co-chairs the Forum's Technical Symposium, November 2002. He published the first interdisciplinary graduate text on SWR, *Software Radio Architecture* [Wiley Interscience]. His doctoral dissertation, *Cognitive Radio* (KTH, June 2000), created the first teleinformatics framework for autonomous software radios, integrating machine learning and language processing into software radio. He edited the IEEE text *Software Radio Technologies*. With The MITRE Corporation, Dr. Mitola applies his expertise in telecommunications and information processing to the national and tactical needs of DoD. More recently he served as Senior Program Manager at the Defense Advanced Research Projects Agency and General Systems Engineer of the Defense Airborne Reconnaissance Office

(DARO). Prior to MITRE, Dr. Mitola was the Chief Scientist of Electronic Systems, E-Systems Melpar Division, culminating a career at E-Systems that began in 1976. He has also held positions of technical leadership with Harris Corporation, Advanced Decision Systems, and ITT Corporation. He began his career with the U.S. DoD in 1967. Dr. Mitola holds the B.S. in EE (Northeastern University '72); M.S.E. (The Johns Hopkins University, 1974); Licentiate in Engineering (May 1999), and Doctorate in Teleinformatics (The Royal Institute of Technology, KTH, Stockholm, June 2000).

BIBLIOGRAPHY

1. Software-Defined Radio (SDR) Forum (*www.sdrforum.org*).
2. Object Management Group (OMG) (*www.omg.org*).
3. J. Mitola and Z. Zvonar, *Software Radio Technology: Selected Readings*, IEEE Press, New York, 2001.
4. E. Del Re, ed., *Software Radio*, Springer-Verlag, London, 2001.
5. J. Mitola, *Software Radio Architecture*, Wiley, New York, 2000.
6. Kohno, *Software Radio and Software Antenna: Spatial and Temporal Communication Theory Using Software Antenna*, Yokohama National Univ., Yokohama, Japan, 1998.
7. Upmal and Lackey, SPEAKeasy, the military software radio, *IEEE Commun. Mag.* (1995).
8. P. Cook, An architectural overview of the speakeasy system, *IEEE J. Select. Areas Commun.* (April 1999).
9. *ACTS Mobile Communications Summit '98*, European Commission, Rhodes, Greece, June 98.
10. *4th ACTS Mobile Communications Summit '99* (CD-ROM) European Commission, Sorrento, Italy, June 1999.
11. M. Mehta et al., Reconfigurable terminals: An overview of architectural solutions, *IEEE Commun. Mag.* (Aug. 2001).
12. *<http://www.motorola.com/GSS/SSTG/ISSPD/WITS/DMR.html>*.
13. Joint Tactical Radio System homepage, *www.jtrs.saalt.army.mil* (2002).
14. McGarth et al., *RFIC Technology for Wireless Consumer Products — Trends in GaAs*, M/A-COM LOUD & Clear, M/A-COM, Inc., Lowell, MA, 1995.
15. Kennedy and Sullivan, Direction finding and smart antennas using software radio architectures, *IEEE Commun. Mag.* (May 1995).
16. D. Nicholson, *Spread Spectrum Signal Design LPE and AJ Systems*, Computer Science Press, Rockville, MD, 1988.
17. Stallings, *Handbook of Computer-Communications Standards*, Vol. 1, *The Open Systems Interconnection* (*OSI*) *Model and OSI-Related Standards*, Macmillan, New York, 1987.
18. J. Mitola, *Cognitive Radio: Model Based Competence for Software Radios*, Licentiate thesis, KTH (The Royal Institute of Technology), Stockholm, Sweden, Aug. 1999.
19. Pickholtz and Hill, *Adaptive Beamforming for Interference Reduction*, George Washington Univ. PW3312A, Dec. 31, 1990.
20. Zoltowski et al., Blind 2-D rake receivers based on space-time adaptive MVDR processing for IS-95 CDMA system, *Proc. MILCOM 96*, IEEE, New York, Oct. 1996.
21. Belzer et al., *Joint Source Channel Coding of Images with Trellis Coded Quantization and Convolutional Codes*, UCLA, Los Angeles, 1998.
22. Ferguson and Huston, *Quality of Service*, Wiley, New York, 1998.
23. Paradells et al., *DECT Multibearer Channels*, IEEE Press, New York, 1994.
24. Strom and Shaula, Optimistic recovery in distributed systems, *ACM Trans. Comput. Sys.* (1985).
25. Pesonen, *Object-Based Design of Embedded Software Using Real-Time Operating Systems*, IEEE Press, New York, 1994.
26. M. Cummings and S. Heath, Mode switching and software download for software defined radio: The SDR Forum approach, *IEEE Commun. Mag.* (Aug. 1999).
27. ITU Recommendation H.320, *Narrow-band Visual Telephone Systems and Terminal Equipment* (*www.itu.int/publications/itu-t/ituth13.htm*), International Telecommunications Union, 1998.
28. ITU, *Coding of analogue signals by pulse code modulation* (*G.711–G.712*) *and by methods other than PCM* (*G.720–G.729*), International Telecommunications Union, Geneva, Switzerland, 1998.
29. IETF references to internetworking.
30. Mouly and Pautet, Evolution of the GSM system, *IEEE PCS Mag.* (Oct. 1995).
31. J. Storer, *Data Compression*, The Computer Science Press, Rockville, MD, 1988.
32. Ziemer and Peterson, *Digital Communications and Spread Spectrum Systems*, Macmillan, New York, 1985.
33. W. Peterson and E. Weldon, *Error-Correcting Codes*, MIT Press, Cambridge, MA, 1972.
34. G. Simmons, ed., *Contemporary Cryptography*, IEEE Press, New York, 1992.
35. J. Razavilar et al., Software radio architecture with smart antennas: A tutorial on algorithms and complexity, *IEEE J. Select. Areas Commun.* (April 1999).
36. J. Mitola, *Cognitive Radio: An Integrated Agent Architecture for Software-Defined Radio*, doctoral dissertation, KTH (The Royal Institute of Technology), Stockholm, Sweden, June 2000.
37. ITU Recommendation H.320, ITU-T, Geneva, 1998.
38. *Random House Unabridged Webster's Dictionary*, Random House, New York, 1999.
39. *DII Strategic Enterprise Architecture*, DISA, Washington, DC, 1994.
40. *Technical Architecture for Information Management* (*TAFIM*), U.S. DoD, Washington, DC, 1996.
41. E. Gamma et al., *Design Patterns: Elements of Reusable Object-Oriented Software*, Adisson-Wesley, Reading, MA, 1994.
42. K. Gardner et al., *Cognitive Patterns: Problem Solving Frameworks for Object Technology*, Cambridge Univ. Press, Cambridge, UK, 1998.
43. Mouly and Pautet, *The GSM System for Mobile Communications*, (published by the authors), Plaiseau, France, 1992.
44. J. Mitola, Software radio architecture: A mathematical perspective, *IEEE J. Select. Areas Commun.* (April 1999).

45. Hoest and Shavit, Towards a topological characterization of asynchronous complexity, *Proc. PODC'97*, ACM, Santa Barbara, CA, 1997.
46. J. Mitola, Software radios: Technology and prognosis, *Proc. Nat. Telesystems Conf.*, May 1992, IEEE, New York, 1992.
47. Ono, *Introduction to Point Set Topology*, Johns Hopkins Univ. Press, Baltimore, MD, 1974.
48. J. Mitola, Software radio architecture: A mathematical perspective, *IEEE J. Select. Areas Commun.* (April 1999).
49. T. Mowbray and R. Zahavi, *The Essential CORBA*, Wiley, New York, 1995.
50. *Software Communications Architecture Specification*, MSRC-5000SCA V2.2, JTRS Joint Program Office, Rosslyn, VA (online) *www.jtrs.saalt.army.mil/docs/documents/sca.html* (Nov. 17, 2001).
51. W. C. Y. Lee, *Mobile Communications Design Fundamentals*, Sams, Indianapolis, IN, 1986.
52. Mouly and Pautet, Evolution of the GSM system, *IEEE PCS Mag.* (Oct. 1995).
53. D. Nicholson, *Spread Spectrum Signal Design LPE and AJ Systems*, Computer Science Press, Rockville, MD, 1988.
54. Qualcomm, *The Technical Case for Convergence of Third Generation Wireless Systems Based on CDMA* (*www.qualcomm.com*), March 1999.
55. Bensley et al., *Introduction to Parallel Supercomputing*, The MITRE Corp., Bedford, MA, 1988.
56. *Jane's Military Communications 1992–93* Jane's Information Group, Surrey, UK, 1992.
57. K.-C. Chen and S.-T. Wu, A programmable architecture for OFDM-CDMA, *IEEE Commun. Mag.* (Nov. 2000).
58. T. Kanter, *Adaptive Personal Mobile Communication: Service Architecture and Protocols*, doctoral dissertation, KTH (The Royal Institute of Technology), Stockholm, Sweden, Nov. 2001.
59. J. L. Dixon and J. Wilkes, A 'low-cost' software radio testbed, *Proc. IEEE VTS 53rd Vehicular Technology Conf.*, Spring 2001, IEEE Press, New York, 2001.
60. H. Shiba et al., Design and evaluation of software radio prototype with over-the-air download function, *Proc. Vehicular Technology Conf.*, Fall 2001, IEEE Press, New York, 2001.
61. M. Cummings and S. Heath, Mode switching and software download for software defined radio — the SDR Forum approach, *IEEE Commun. Mag.* (Aug 1999).
62. R. Rummler et al., Traffic modeling of software download for reconfigurable terminals, *Proc. 12th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Sept. 2001, IEEE Press, New York, 2001.
63. R. Walden, Analog to digital converter súrvey and analysis, *IEEE J. Select. Areas Commun.* (April 1999).
64. Texas Instruments Corp. homepage, *www.ti.com* (Jan. 2002).
65. Analog Devices homepage, *www.analogdevices.com* (Jan. 2002).
66. R. Dean Straw N6BV, *The ARRL Handbook for Radio Amateurs*, ARRL (National Association for Amateur Radio), Newington, CT, 2000.
67. WinRadio, *www.advdig.com*, Advanced Digital Systems of St. Louis, Saint Louis, MO, Nov. 1999.
68. U. Rhode et al., *Communications Receivers*, McGraw-Hill, New York, 1997.
69. *RF IC Design for Wireless Communication Systems*, Mead Microelectronics, Inc., Corvallis, OR, 1996.
70. *The Software Defined Radio*, BellSouth, Athens, GA, Dec. 1995.
71. C. Dick, Configurable logic for digital communications: Some signal processing perspectives, *IEEE Commun. Mag.* (Aug. 1999).
72. A. Shankiti and M. Leeser, Implementing a RAKE receiver for wireless communications on an FPGA-based computer system, *Proc. FPGA 2000*, Monterey CA, ACM, New York, 2000.
73. *Harris Digital Channelizer Application Note: Channelized Receiver*, Harris Corp. Melbourne, FL, 1990.
74. K. Zangi and R. Koilpillai, Software radio issues in cellular base stations, *IEEE J. Select. Areas Commun.* (April 1999).
75. Pentek homepage, *www.pentek.com* (1999).
76. E. Farag et al., *A Programmable Power-Efficient Decimation Filter for Software Radios*, ACM O-S97913O3-31971O8, ACM, New York, 1997.
77. T. Yokoi et al., Software receiver technology and its applications, *IEICE Trans. Commun.* (Tokyo) (June 2000).
78. X. Reves et al., Software radio implementation of a DS-CDMA indoor subsystem based on FPGA devices, *Proc. 12th IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Sept. 2001, IEEE Press, New York, 2001.
79. Y. Suzuki, Software radio base and personal station prototypes, *IEICE Trans. Commun.* (Tokyo) (June 2000).
80. *Proc. Mobile Communications Summit*, Barcelona, Spain, Sept. 2001, European Commission, Brussels, 2001.
81. R. Kohno, Structures and theories of software antennas for software defined radio, *IEICE Trans. Commun.* (Tokyo) (June 2000).
82. *Proc. ACTS Mobile Communications Summit '98*, June 1998, European Commission, Rhodes, Greece, 1998.
83. *Proc. 4th ACTS Mobile Communications Summit '99*, June 1998 (CD-ROM), European Commission, Sorrento, Italy, 1999.
84. J. Rosa, *RF/IF Subsystem of a Commercial SDR Base Station*, SDR Forum Document SDRF-01-I-0012-V0.00, Hypres Corp., White Plains, NY (*www.hypres.com*) Jan. 2001.
85. Motorola homepage, *www.motorola.com*.
86. V. Bose et al., Virtual radios, *IEEE J. Select. Areas Commun.* (April 1999).
87. Vanu, Inc. homepage, *www.vanu.com*.
88. P. Withington, *Impulse Radio Overview* (*www.timedomain.com*), Time Domain, Inc., 1999.
89. J. Mitola, The software radio architecture, *IEEE Commun. Mag.* (May 1995).
90. *The Software Defined Radio Request for Information*, BellSouth, Atlanta, GA, Dec. 1995.
91. Van Tuyl et al., FCC transmitter certification requirements: Issues related to software defined radio, *Proc. SDR Forum*, SDR Forum, Rome, NY, June 1999.
92. J. Mitola, Software radio architecture: A mathematical perspective, *IEEE J. Selec. Areas Commun.* (April 1999).

93. Tebbs & Garfield, *Real Time Systems*, McGraw-Hill, Berkshire, UK, 1977.

94. K. Ellison, *Developing Real-Time Embedded Software in a Market-Driven Company*, Wiley, New York, 1994.

95. J. Mitola, Cognitive radio for mobile multimedia communications, *Proc. IEEE Mobile Multimedia Communications* (MOMUC) Workshop, San Diego, CA, Nov. 1999, IEEE Press, New York, 1999.

SPACE–TIME CODES FOR WIRELESS COMMUNICATIONS

NAOFAL AL-DHAHIR
A. R. CALDERBANK
AT&T Shannon Laboratory
Florham Park, New Jersey

AYMAN F. NAGUIB
Morphics Technology Inc.
Campbell, California

Space–time coding is a communications technique for wireless systems that employ multiple transmit antennas and single or multiple receive antennas. Information theory has been used to demonstrate that multiple antennas have the potential to dramatically increase achievable data rates. Space–time codes realize these gains by introducing temporal and spatial correlation into the signals transmitted from different antennas. There is, in fact, a diversity gain that results from multiple paths between base station and mobile terminal, and a coding gain that results from how symbols are correlated across transmit antennas. Significant increases in throughput are possible with only two antennas at the base station and one or two antennas at the mobile, and with simple receiver structures. The second antenna at the mobile terminal can be used to further increase system capacity through interference suppression. This article provides an overview of space–time coding techniques and the associated signal processing framework for narrowband and broadband wireless communications.

Current cellular standards such as IS136 support circuit data and fax (facsimile) services at a rate of 9.6 kbps (kilobits per second), and a packet data mode is now being standardized. Rapid growth in mobile computing and other wireless data services is inspiring many proposals for high-speed data services in the range of 64–384 kbps for microcellular wide-area and high-mobility applications, and up to 2 Mbps for indoor applications [1].

However, data rates on band–limited wireless channels are limited by multipath fading and interference from other users [2–8]. Deploying multiple antennas at the both the transmitter and the receiver increases the capacity of wireless channels, and information theory provides measures of this increase [9–11]. The standard approach to increasing capacity is to use linear processing at the receiver with optimum linear combining to combat multipath fading and suppress interference [3,4]. Here, the received signals are weighted and combined to maximize the signal-to-interference-plus-noise ratio (SINR) at the receiver. By contrast, transmit diversity schemes use processing at the transmitter to spread the information across multiple transmit antennas. The earliest forms of transmit diversity were proposed by Uddenfeldt and Raith [12] and Wittneben [13]. The latter is a delay diversity scheme where a signal is transmitted from one antenna, then delayed one symbol, and transmitted from a second antenna. Wittneben's work includes the delay diversity scheme of Seshadri and Winters [14] as a special case (see also Refs. 15 and 16). Winters [17] showed that delay diversity is optimal in the sense that the diversity order experienced by an optimal receiver is equal to the number of transmit antennas. Diversity is the link-level advantage (over a single path) obtained from spreading information over multiple independent paths from base station to mobile unit. Note that superposition of fading statistics at the receiver also reduces variation in signal strength, and allows smoother and more efficient power control. This means that the base station can support significantly more users for a given constraint on radiated signal power. Space–time coding [18–25] combines correlation techniques designed for multiple transmit antennas with the appropriate signal processing at the receiver to provide significant gain over the delay diversity schemes in Refs. 13 and 14.

This article is organized as follows. Section 1 describes fundamentals of space–time coding for flat-fading channels. Equalization schemes necessary for implementation of space–time codes over frequency-selective channels are discussed in Section 2, and channel estimation issues are discussed in Section 3. An extensive reference list is provided to help the reader undertake a more detailed study of any of the topics discussed.

1. SPACE–TIME CODING

This section presents a mathematical model of a *narrowband* communications system with N_t transmit antennas and N_r receive antennas and it assumes a flat-fading channel (see Section 2 for frequency-selective channels). As shown in Fig. 1, the space–time encoder transforms the input data at time l into N_t code symbols

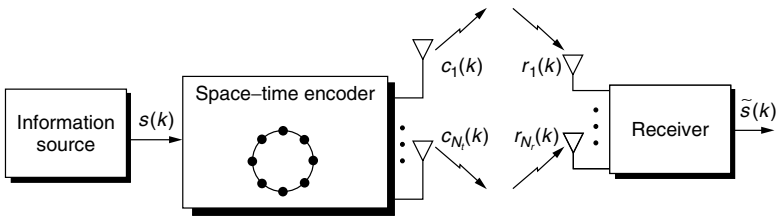


Figure 1. Space–time coding.

$c_1(l), c_2(l), \ldots, c_{N_t}(l)$ that are transmitted *simultaneously* from the different transmit antennas.

Signals arriving at different receive antennas undergo independent fading. The signal at each receive antenna is a noisy superposition of the faded versions of the N_t transmitted signals. Let E_s be the average energy of the signal constellation. The constellation points are scaled by a factor $\sqrt{E_s}$ so that the average energy of the constellation points is 1. Let $r_j(l), j = 1 \cdots N_r$ be the received signal at antenna j after matched filtering. Assuming ideal timing and frequency information, we have

$$r_j(l) = \sqrt{E_s} \cdot \sum_{i=1}^{N_t} \alpha_{ij}(l) c_i(l) + \eta_j(l), \qquad j = 1, \ldots, N_r \tag{1}$$

where $\eta_j(l)$ are independent samples of a zero-mean complex white Gaussian process with two-sided power spectral density $N_0/2$ per dimension. It is also assumed that $\eta_j(l)$ and $\eta_k(l)$ are independent for $j \neq k$, $1 \leq j, k \leq N_r$. The gain $\alpha_{ij}(l)$ models the complex fading channel gain from transmit antenna i to receive antenna j. The channel gain α_{ij} is modeled as a lowpass-filtered complex Gaussian random process with zero mean, variance 1, and autocorrelation function $R_\alpha(\tau) = J_0(2\pi f_d \tau)$, where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind and f_d is the maximum Doppler frequency [26]. It is assumed that signals transmitted from different antennas are subject to independent fades. This can be achieved by separating transmit antennas by more than half the underlying wavelength or by using antennas with different polarizations.

Let $\mathbf{c}_l = [c_1(l), \ldots, c_{N_t}(l)]^T$ be the $N_t \times 1$ *codeword* transmitted from the N_t antennas at time l, $\alpha_j(l) = [\alpha_{1j}(l), \ldots, \alpha_{N_t j}(l)]^T$ be the corresponding $N_t \times 1$ channel vector from the N_t transmit antennas to the jth receive antenna, and $\mathbf{r}(l) = [r_1(l), \ldots, r_{N_r}(l)]^T$ be the $N_r \times 1$ received signal vector. Also, let $\eta(l) = [\eta_1(l), \ldots, \eta_{N_r}(l)]^T$ be the $N_r \times 1$ noise vector at the receive antennas. Furthermore, let us define the $N_r \times N_t$ channel matrix $\mathcal{H}_l$ from the N_t transmit to the N_r receive antennas as $\mathcal{H}(l) = [\alpha_1(l), \ldots, \alpha_{N_r}(l)]^T$. Equation (1) can be rewritten in a matrix form as

$$\mathbf{r}(l) = \sqrt{E_s} \cdot \mathcal{H}(l) \cdot \mathbf{c}_l + \eta(l) \tag{2}$$

We can easily see that the *signal-to-noise ratio* (SNR) *per receive antenna* is given by

$$\text{SNR} = \frac{N_t \cdot E_s}{N_0} \tag{3}$$

1.1. Space–Time Trellis Codes (STTCs)

Suppose that the *codeword* sequence

$$\mathcal{C} = \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L$$

were transmitted and consider the probability that the decoder decides erroneously in favor of another legitimate codeword sequence

$$\tilde{\mathcal{C}} = \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \ldots, \tilde{\mathbf{c}}_L$$

Assume a data frame of length L and define the $N_t \times N_t$ error matrix $\mathcal{A}$ as

$$\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}}) = \sum_{l=1}^{L} (\mathbf{c}_l - \tilde{\mathbf{c}}_l)(\mathbf{c}_l - \tilde{\mathbf{c}}_l)^* \tag{4}$$

The squared distance between $\mathcal{C}$ and $\tilde{\mathcal{C}}$ at the output of the wireless channel turns out to be proportional to $\sum_{j=1}^{N_r} \mathcal{H}_j^* \mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}}) \mathcal{H}_j$, where $\mathcal{H}_j$ is the column vector of path gains from the different transmit antennas to the jth receive antenna. The vector $\mathcal{H}_j$ varies with time, and when it finds the null space of $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$, the jth receive antenna experiences a deep fade. Diversity gain is just the minimum rank of $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$, where the minimization is over all pairs of codewords. Coding gain depends on the product of the nonzero eigenvalues, and again there is a minimization over all pairs of codewords.

If ideal channel state information (CSI) $\mathcal{H}(l), l = 1, \ldots, L$ is available at the receiver, it is straightforward to show that the probability of transmitting $\mathcal{C}$ and deciding in favor of $\tilde{\mathcal{C}}$ is upper-bounded by

$$P(\mathcal{C} \to \tilde{\mathcal{C}}) \leq \left(\prod_{i=1}^{p} \lambda_i \right)^{-N_r} \cdot \left(\frac{E_s}{4N_0} \right)^{-pN_r} \tag{5}$$

where p is the rank of the error matrix $\mathcal{A}$ and $\lambda_i, i = 1, \ldots, p$ are the nonzero eigenvalues of the error matrix $\mathcal{A}$ (see Ref. 27 for details). The bound on probability of error given in Eq. (5) is similar to the probability of error bound for trellis-coded modulation for flat-fading channels. The first term $g_p = (\lambda_1 \lambda_2 \cdots \lambda_p)$ represents the coding gain achieved by the space–time code, and the second term $(E_s/4N_0)^{-pN_r}$ represents a diversity gain of pN_r. This analysis leads to two design criteria for space–time codes. The first is to maximize the rank p of $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$, thereby maximizing diversity gain. The second, for a given diversity gain p, is to maximize the coding gain g_p.

Now consider the problem of decoding space–time codes. Under the assumption that ideal CSI $\mathcal{H}(l), l = 1, \ldots, L$ is available at the receiver, we can derive the maximum-likelihood (ML) decoding rule for the space–time code as follows. Suppose that all codewords are equiprobable, a codeword

$$\mathcal{C} = \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L$$

has been transmitted, and

$$\mathcal{R} = \mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_L$$

has been received, where $\mathbf{r}_l$ is given by Eq. (2). At the receiver, optimum decoding amounts to choosing a codeword sequence

$$\tilde{\mathcal{C}} = \tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \ldots, \tilde{\mathbf{c}}_L$$

for which the a posteriori probability

$$\Pr(\tilde{\mathcal{C}} | \mathcal{R}, \mathcal{H}(l), l = 1, \ldots, L)$$

is maximized. Since the noise vector is assumed to be a multivariate AWGN (additive white Gaussian noise), it can be easily shown [27] that the optimum decoder is

$$\tilde{\mathcal{C}} = \arg \min_{\tilde{\mathcal{C}} = \tilde{\mathbf{c}}_1, \cdots, \tilde{\mathbf{c}}_L} \sum_{l=1}^{L} \| \mathbf{r}(l) - \sqrt{E_s} \cdot \mathcal{H}(l) \cdot \tilde{\mathbf{c}}_l \|^2 \tag{6}$$

It is obvious that the optimum decoder in (6) can be implemented using the Viterbi algorithm (VA) when the space–time code has a trellis representation.

Figure 2 shows an 8-PSK 8-state space–time code designed for two transmit antennas, where the edge label xy means that symbol x is transmitted from the first antenna and symbol y, from the second antenna. The different symbol pairs in a given row label the transitions (edges) out of a given state, in order, from top to bottom. Observe that labels on edges leaving a given state disagree in the first position. It follows that the rank of the matrix $\mathcal{A}(\mathcal{C}, \tilde{\mathcal{C}})$ corresponding to codewords $\mathcal{C}$ and $\tilde{\mathcal{C}}$ (that diverge and then remerge) is equal to 2. The reader may verify that, for odd-numbered states, if the symbol transmitted from the first antenna is negated, the result is the delay diversity scheme proposed by Wittneben. Both schemes provide a diversity gain of 2, but with the space–time code there is an additional coding gain of 2.5 dB.

1.2. Space–Time Block Codes (STBCs)

When the number of antennas is fixed, the decoding complexity of space–time trellis coding (measured by the number of trellis states at the decoder) increases exponentially as a function of the diversity level and transmission rate [22]. In addressing the issue of decoding complexity, Alamouti [18] discovered a remarkable space–time block coding scheme for transmission with two antennas. This scheme supports maximum-likelihood detection based only on linear processing at the receiver. It was later generalized [19] to an arbitrary number of antennas and is able to achieve the full diversity promised by the number of transmit and receive antennas. Here, we will briefly review the basics of space–time block codes [18]. Figure 3 shows the baseband representation for transmit diversity employing space–time block coding with two transmit antennas. The input symbols to the space–time block encoder are divided into pairs. At a given symbol period, the two symbols in each group $\{c_1, c_2\}$ are transmitted simultaneously from the two antennas. The signal transmitted from antenna 1 is c_1, and the signal transmitted from antenna 2 is c_2. In the next symbol period, the signal $-c_2^*$ is transmitted from antenna 1 and the signal c_1^* is transmitted from antenna 2. Let h_1 and h_2 be the channel gains from the first and second transmit antennas to the receive antenna, respectively. The major assumption here is that h_1 and h_2 are constant over two consecutive symbol periods:[1]

$$h_i(nT) = h_i((n+1)T), \qquad i = 1, 2$$

Let r_1 and r_2 be the received signals over two consecutive symbol periods. Then

$$r_1 = h_1 c_1 + h_2 c_2 + \eta_1 \tag{7}$$

$$r_2 = -h_1 c_2^* + h_2 c_1^* + \eta_2 \tag{8}$$

where η_1 and η_2 represent the AWGN and are modeled as i.i.d. complex Gaussian random variables with zero mean and power spectral density $N_0/2$ per dimension. Define the received signal vector $\mathbf{r} = [r_1\ r_2^*]^T$, the codeword vector $\mathbf{c} = [c_1\ c_2]^T$, and the noise vector $\eta = [\eta_1\ \eta_2^*]^{\mathbf{T}}$. Equations (7) and (8) can be rewritten in a matrix form as

$$\mathbf{r} = \mathbf{H} \cdot \mathbf{c} + \eta \tag{9}$$

where the channel matrix $\mathbf{H}$ is defined as

$$H = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \tag{10}$$

The vector η is a complex Gaussian random vector with zero mean and covariance $N_0 \cdot \mathbf{I}$. Define $\mathcal{C}$ as the set of

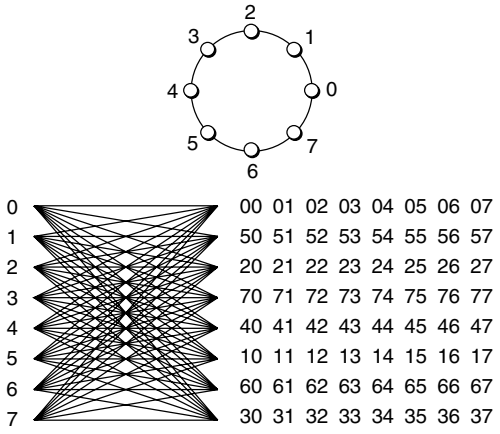


Figure 2. 8-PSK 8-state space–time code with two transmit antennas and a bandwidth efficiency of 3 bits/channel use.

[1] For GSM (Global System for Mobile Communication) mobiles traveling at 60 mph (mi/h) and a carrier frequency of 1 GHz, the channel coherence time is around 11 ms, which is about 3000 GSM symbol durations. Hence, the channel can be safely assumed constant over several hundred symbol durations even at highway speeds. For IS136 mobiles, the symbol duration is around 41 μs; hence, the channel can be assumed constant over few tens of consecutive symbols. Therefore, the assumption of a fixed channel over two consecutive symbols is satisfied for both systems.

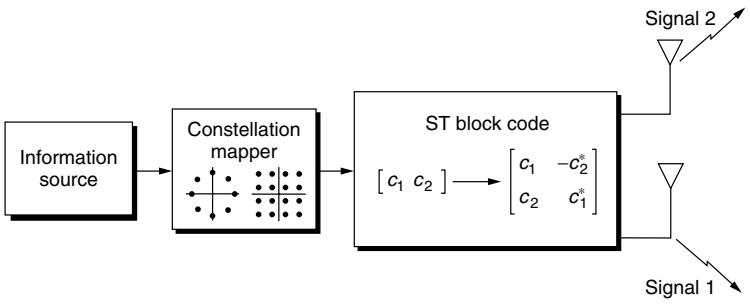


Figure 3. Transmit diversity with space–time block coding.

all possible symbol pairs $\mathbf{c} = \{c_1, c_2\}$ and assume that all symbol pairs are equiprobable. Since the noise vector η is assumed to be a multivariate AWGN, it follows that the optimum ML decoder is

$$\hat{\mathbf{c}} = \arg\min_{\hat{\mathbf{c}} \in \mathcal{C}} \|\mathbf{r} - \mathbf{H} \cdot \hat{\mathbf{c}}\|^2 \tag{11}$$

The ML decoding rule in this equation can be further simplified by realizing that the channel matrix $\mathbf{H}$ is orthogonal ($\mathbf{H}^*\mathbf{H} = (|h_1|^2 + |h_2|^2) \cdot \mathbf{I}$). Consider the modified received signal vector $\tilde{\mathbf{r}}$ given by

$$\tilde{\mathbf{r}} = \mathbf{H}^* \cdot \mathbf{r} = (|h_1|^2 + |h_2|^2) \cdot \mathbf{c} + \tilde{\eta} \tag{12}$$

where $\tilde{\eta} = \mathbf{H}^* \cdot \eta$. In this case the ML decoding rule becomes

$$\hat{\mathbf{c}} = \arg\min_{\hat{\mathbf{c}} \in \mathcal{C}} \|\tilde{\mathbf{r}} - (|h_1|^2 + |h_2|^2) \cdot \hat{\mathbf{c}}\|^2 \tag{13}$$

Since $\mathbf{H}$ is orthogonal, it follows that the noise vector $\tilde{\eta}$ will have a zero mean and covariance $(|h_1|^2 + |h_2|^2) \cdot \mathbf{I}$, that is, the elements of $\tilde{\eta}$ are independent and identically distributed. Hence, it follows immediately that simple linear combining reduces the ML decoding rule in Eq. (13) to two separate, and much simpler, ML decoding rules for c_1 and c_2, as established elsewhere [18]. Assuming that we are using a signaling constellation with 2^b constellation points, this linear combining reduces the number of decoding metrics that have to be computed for ML decoding from 2^{2b} to 2×2^b. It is also straightforward to verify that the SNR for c_1 and c_2 will be

$$\text{SNR} = \frac{(|h_1|^2 + |h_2|^2) \cdot E_s}{N_0} \tag{14}$$

and hence a two-branch diversity performance is obtained at the receiver. When the receiver uses N_r receive antennas, we can write the received signal vector $\mathbf{r}_m$ at receive antenna m and 2, respectively as

$$\mathbf{r}_m = \mathbf{H}_m \cdot \mathbf{c} + \eta_{\mathbf{m}} \tag{15}$$

where $\eta_{\mathbf{m}}$ is the noise vector and $\mathbf{H}_m$ is the channel matrix from the two transmit antennas to the mth receive antenna. In this case the optimum ML decoding rule is

$$\hat{\mathbf{c}} = \arg\min_{\hat{\mathbf{c}} \in \mathcal{C}} \sum_{m=1}^{N_r} \|\mathbf{r}_m - \mathbf{H}_m \cdot \hat{\mathbf{c}}\|^2 \tag{16}$$

As before, in the case of N_r receive antennas, the decoding rule can be further simplified by premultiplying the received signal vector $\mathbf{r}_m$ by $\mathbf{H}_m^*$. In this case, the diversity order provided by this scheme is $2N_r$. Figure 4 shows a simplified block diagram for the receiver with two receive antennas. The properties of the space–time block coding scheme in Fig. 4 and its extension in Ref. 19 can be further exploited to improve wireless capacity and/or throughput. The reader is referred to Ref. 28 for further discussion on this point.

2. EQUALIZATION OF SPACE–TIME CODES ON FREQUENCY-SELECTIVE CHANNELS

As the transmission bandwidth increases beyond the *coherence bandwidth* [29] of the channel, equalization becomes indispensable. Equalization complexity increases with the channel memory (also referred to as the *channel delay spread*), signal constellation size, and the use of multiple transmit and/or receive antennas. The objective of this section is to give an overview of candidate equalization schemes for space–time-coded transmission over broadband wireless channels. In this article, we use the terms *frequency-selective channel*, *broadband channel*, and *intersymbol interference* (ISI) *channel* interchangeably.

We start in Section 2.1 by describing the frequency-selective channel model and assumptions. Effective equalization schemes for STTC and STBC are discussed in Sections 2.2 and 2.3, respectively.

2.1. Channel Model and Assumption

The channel impulse response (CIR) from transmit antenna i to receive antenna j is denoted by the vector $\mathbf{h}_{ij}$. The multiple-input/multiple-output (MIMO) channel memory, denoted by ν, is the maximum memory of all constituent single-input/single-output (SISO) channels. For simplicity, we focus on the case $N_t = 2$ and $N_r = 1$, hence, $\mathbf{h}_{ij}$ will be simply denoted by the vector $\mathbf{h}_i$ or its corresponding D-transform $h_i(D) \stackrel{def}{=} \sum_{k=0}^{\nu} \mathbf{h}_i(k) D^k$. Extension to the general case is straightforward. The CIRs are assumed constant over the transmission block (quasistatic fading) and vary independently from block

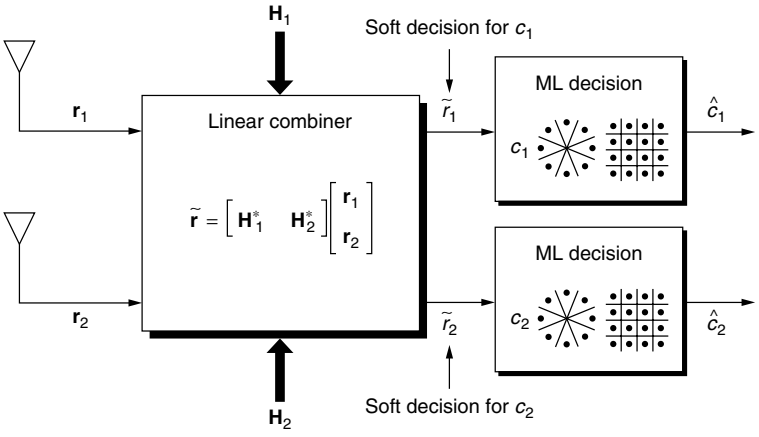


Figure 4. Receiver for space–time block coding.

to block. The input symbols are assumed complex zero-mean and belong to a 2^b signal constellation. The noise is additive white Gaussian and independent of the input.

2.2. Equalization Schemes for Space–Time Trellis Codes

Our focus will be on the 8-state 8-PSK STTC shown in Fig. 2. This code has a rich and transparent structure that can be exploited to simplify equalization.

1. *Turbo Equalization*. While it is possible in theory to model the STTC and the ISI channel, separated by an interleaver,[2] by a single trellis and perform maximum a posteriori (MAP) decoding on this trellis using, for instance, the BCJR (Bahl–Cocke–Jelinek–Raviv) algorithm [30], the complexity would be prohibitive. An alternative lower-complexity decoding scheme views the space–time encoder and the ISI channel, separated by an interleaver, as a serial concatenation of two finite-state machines that can be decoded iteratively using the *Turbo principle* [31]. Using this Turbo equalization scheme, joint space–time equalization and decoding is performed by iteratively exchanging *soft* extrinsic information between the separate BCJR-MAP equalizer and decoder modules. Hard decisions are generated only after the last iteration. The BCJR algorithm consists of a forward and a backward recursion and is usually implemented in the log domain to reduce computational complexity and improve numerical accuracy. Turbo equalization achieves remarkable performance very close to theoretical performance limits [32]. The number of states in the BCJR equalizer module is *exponential* in the channel memory, the number of transmit antennas, and the spectral efficiency (in bps/Hz). The use of MIMO FIR shortening prefilters [33] to reduce the complexity of the BJRC equalizer module and its application to STTC have been studied in [34]. However, for spectrally efficient modulation schemes (such as 8-PSK modulation used in EDGE[3]), the complexity of turbo equalization is still too high [36]. In addition, the long decoding delay might not be acceptable for speech and real-time data applications. An attractive alternative in this case is the M-BCJR equalizer described next.

2. *Prefiltered M-BCJR Equalizer*. The M-BCJR algorithm [37], is a reduced-complexity version of the BCJR algorithm [30] where at each trellis step, only the M active states associated with the highest metrics are retained. An improved version of the M-BCJR algorithm was proposed [38] and applied to the equalization of STTC. Moreover, it was shown [38] that preceding the M–BCJR equalizer with a channel-shortening prefilter improves its performance, especially for small values of M. Even better performance is achieved when a different prefilter is used for the forward and backward recursions of the M-BCJR algorithm. The value of M and the number of prefilter taps can be jointly optimized to achieve the best performance–complexity tradeoffs.

3. *Prefiltered MLSE/DDFSE Equalizer*. Unlike the BJCR-MAP equalizer, which minimizes the *symbol* error probability, maximum-likelihood sequence estimation (MLSE) minimizes the *sequence* error probability assuming equally likely inputs and can be implemented efficiently using the VA. A major advantage of the BCJR-MAP algorithm over the conventional VA[4] is the generation of *soft* information on the decisions. Generalization of the MLSE equalizer to the MIMO case was first reported by Van Etten [40]. For a 2^b signal constellation, N_t transmit antennas, and MIMO channel memory of ν, the MIMO MLSE equalizer has $2^{b \cdot N_t \cdot \nu}$ states in general. The number of equalizer states can be reduced to $2^{b \cdot \nu}$ by using the STTC trellis structure as shown in Ref. 41. However, this complexity is still too high for large signal constellations and long MIMO channel memory. Delayed decision feedback sequence estimation (DDFSE) was introduced [42] as a hybrid scheme between MLSE and decision feedback equalization (DFE) [43] for channels with long memory. Basically, the CIR is divided into a leading part and a tail. Then, an MLSE equalizer is constructed on the basis of the leading part, and the interfering effect of the CIR tail is canceled by feedback using previous (hard) decisions (assumed correct). Like all feedback schemes, DDFSE suffers from error propagation effects. These effects are minimized if most of the channel energy is concentrated in its leading part (as in minimum-phase channels). This is the task of the FIR prefilter designed [44] to improve DDFSE performance in equalizing the 8-state 8-PSK STTC [41].

4. *Orthogonal Frequency Division Multiplexing* (*OFDM*). In OFDM, the high-rate input stream is demultiplexed and transmitted over N low-rate independent frequency subcarriers. This multicarrier transmission scheme is implemented digitally using the efficient fast Fourier transform (FFT) method [45]. OFDM is a block transmission scheme; therefore, a guard sequence (of length at least equal to channel memory) is needed to eliminate interblock interference (IBI). The most popular choice for guard sequence is a *cyclic prefix*, which makes the channel matrix *circulant*; hence, diagonalizable by the FFT. If the FFT size is made large enough such that the width of each frequency bin is less than the *coherence bandwidth* of the channel, then no equalization is needed.[5] A large FFT size (compared to channel memory) also reduces the guard sequence overhead at the expense of increased storage and processing requirements and increased delay which might not be acceptable for delay-sensitive applications.

The OFDM scheme has been extended to the MIMO case [46]. OFDM was successfully applied to equalization of STTC [47].

[2] The randomizing effect of the interleaver is critical to the remarkable performance exhibited by Turbo schemes.

[3] EDGE stands for *enhanced data rates for GSM evolution* and is the proposed third-generation TDMA cellular standard [35].

[4] A modified soft-output Viterbi algorithm (SOVA) was presented by Hagenauer and Hoeher [39]. However, its performance is suboptimal compared to that of BCJR-MAP.

[5] Except for a simple gain and phase adjustment using a single complex tap for each subchannel, assuming negligible intercarrier interference due to Doppler effects or frequency offset errors.

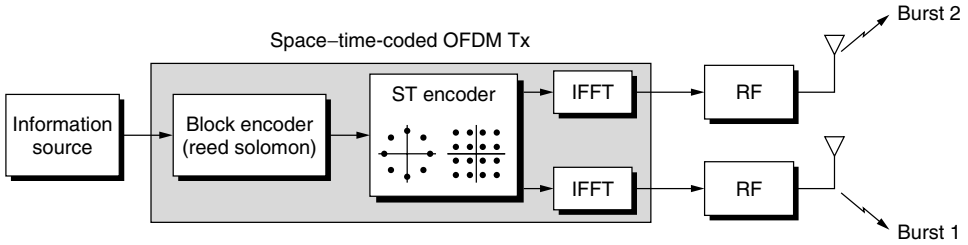


Figure 5. Transmitter for space–time-coded OFDM for broadband applications.

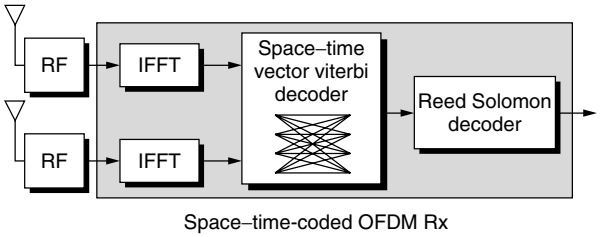


Figure 6. Receiver for space–time-coded OFDM for broadband applications.

Figures 5 and 6 show simplified block diagrams for the transmitter and receiver, respectively, for an OFDM modem with a concatenated space–time coding scheme. The input information symbols are first encoded by an outer conventional channel code. The output of the outer code is then space–time-encoded. Each of the space–time code output streams is then OFDM-modulated and sent over the corresponding antenna. At the receiver, the signal at each receive antenna is OFDM-demodulated. The demodulated signals from the antennas are then fed into the space–time decoder followed by the outer decoder. Figure 7 shows the simulation results for the abovementioned OFDM space–time-coded modem. In this simulation, the available bandwidth is 1 MHz, and the maximum Doppler frequency is 200 Hz. The number of OFDM tones used for modulation is 256. These correspond to a subcarrier separation of 3.9 kHz and OFDM frame duration of 256 μs. To each frame, a cyclic prefix of 40 μs duration is added. Each tone modulates a 4-PSK constellation, although higher-order constellations may be used. We used 16-state 4-PSK space–time code with two transmit and two receive antennas. In addition, an outer (72,64,9) RS code over Galois Field GF(2^7) is used. We plot the frame error probability as function of SNR for different channel delay spreads. From this plot, we can see that an E_b/N_0 between 2.7-4 dB (depending on the delay spread) is needed to achieve a data rate of 1.5 Mbps.

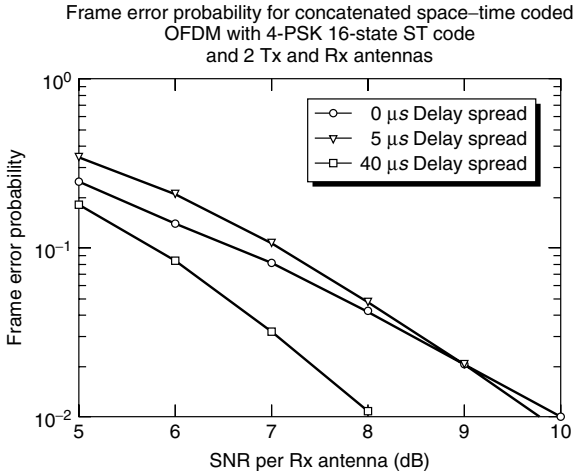


Figure 7. FER of concatenated space–time-coded OFDM with 4-PSK 16-state STC with 2Tx and 2Rx antennas.

2.3. Equalization Schemes for Space–Time Block Codes

Our focus will be on the case of two transmit antennas described in Section 1.2, where a full-rate STBC can be constructed for any signal constellation.

1. *Time-Reversal Space–Time Block Coding* (*TRSTBC*). TRSTBC was introduced [48] as an extension of the Alamouti STBC scheme [18] to frequency-selective channels by imposing the Alamouti orthogonal structure at a *block*, *not symbol*, level as in the flat-fading channel case. At the receiver, TRSTBC employs clever time-domain processing to eliminate the mutual interference effects between the two inputs *while still achieving the maximum diversity gain* of $|\mathbf{h}_1|^2 + |\mathbf{h}_2|^2$. Effectively, TRSTBC converts the two-input/single-output channel to two SISO channels, each with equivalent impulse response $h_{\text{eqv}}(D) = h_1(D)h_1^*(D^{-1}) + h_2(D)h_2^*(D^{-1})$ to which standard SISO equalization schemes such as MLSE [49] or DFE [50] can be applied. TRSTBC assumes that the two channels $h_1(D)$ and $h_2(D)$ are fixed over two consecutive transmission blocks and perfectly known at the receiver and that guard symbols (of length at least equal to channel memory) are inserted between data blocks to eliminate IBI.

2. *Orthogonal Frequency-Division Multiplexing* (*OFDM*). An elegant scheme for combining OFDM and STBC by implementing the Alamouti orthogonal structure at a block level was first reported by Liu et al. [51]. This OFDM–STBC scheme achieves the *full* diversity gain of $|\mathbf{h}_1|^2 + |\mathbf{h}_2|^2$ without bandwidth expansion for two transmit antennas, assuming that the channel is fixed over two consecutive OFDM blocks and known at the receiver and a cyclic prefix is used to eliminate IBI.

3. *Single-Carrier Frequency-Domain Equalization* (*SCFDE*). OFDM has two main drawbacks with respect to single-carrier transmission, namely, a higher peak : average ratio (PAR), which results in larger back-off with nonlinear amplifiers and increased sensitivity to frequency errors and phase noise [52]. An alternative equalization scheme that overcomes these two drawbacks

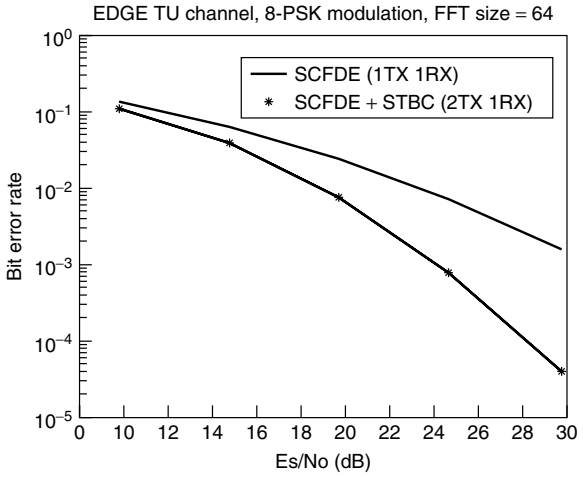


Figure 8. Bit error rate of SCFDE with 1TX and 2TX (STBC) for EDGE TU environment, 8-PSK modulation, and size 64 FFT (1 RX is assumed).

of OFDM while retaining its reduced implementation complexity advantage (due to use of FFT) is single-carrier frequency-domain equalization (SCFDE) [53]. An effective transmit diversity scheme for combining STBC and SCFDE over frequency-selective channels is described in Ref. 54. Figure 8 shows the significant transmit diversity gain achieved as exhibited by the increased slope of the BER curve at high SNR. This simulation assumes a typical urban (TU) EDGE channel with a linearized GMSK transmit pulseshape, 8-PSK modulation, and an FFT size of 64.

3. CHANNEL ESTIMATION ISSUES

Channel estimation for space–time-coded transmissions over flat-fading channels can be performed effectively using orthogonal pilot tones and interpolation as discussed in detail in Ref. 25. Here, we discuss the more challenging frequency-selective channel case.

In single-carrier block transmission systems, a *training sequence* is typically inserted in each block and used to estimate the CIR at the receiver (e.g., using a least-squares algorithm [55]). If the CIR varies within the block, this initial CIR estimate can be tracked using one of various adaptive algorithms.

For single-transmit-antenna scenarios, the training sequence is only required to have a "good" (i.e., impulselike) autocorrelation sequence. However, for the N_t transmit antenna scenarios, the N_t training sequences should, in addition, have "low" (ideally zero) cross-correlation, at least over time lags less than or equal to those of the MIMO channel memory. It can be shown that *perfect root of unity sequences* (PRUS) [56] have these ideal correlation properties. However, PRUS do not belong to standard signal constellations such as PSK. Additional challenges in channel estimation for multiple-transmit-antenna systems over the single-transmit-antenna case are the increased number of channel parameters to be estimated and the reduced transmit power (by a factor of N_t) for each transmit antenna. An obvious transmission scheme that forces the cross-correlation between the N_t training sequences to zero consists of dividing the training interval into N_t subintervals where only one antenna is allowed to transmit its training sequence in each subinterval. This scheme has two major drawbacks: (1) the peak : average ratio (PAR) is increased, which in turn increases amplifier nonlinear distortion; and (2) the effective training period for each transmit antenna is reduced by a factor of N_t.

The rich structure of space–time codes can be used to reduce the number of channel parameters to be estimated. For example, for TRSTBC, decoupling of the two inputs due to the Alamouti orthogonal structure removes the requirement of low cross-correlation between the two training sequences. Similarly, for the 8-state 8-PSK STTC, the special code structure can be exploited to simplify the training sequence design problem while restricting the training symbols to belong to standard signal constellation and incurring negligible performance loss from PRUS [57].

4. CONCLUSION

Space–time coding is a new coding/signal processing framework for wireless communications systems with multiple transmit/receive antennas. This new framework offers the best tradeoff between spectral efficiency and power consumption by optimum combination of modulation, coding, and diversity gains over flat-fading channels.

Space–time trellis codes offer the maximum possible diversity and coding gains without any sacrifice in transmission bandwidth. Their decoding requires a vector Viterbi algorithm. Alamouti-type space–time block codes offer maximum diversity gain, much lower decoding complexity, and full rate transmission but sacrifice coding gain.

For frequency-selective fading channels, we described several competitive equalization schemes for space–time codes that further exploit the temporal diversity of the channel. For space–time trellis codes, prefiltered M-BCJR and OFDM achieve the best performance-complexity tradeoff. For Alamouti's space–time block code, TRSTBC, OFDM–STBC, and FDE–STBC are the most promising candidates. OFDM-based schemes are less attractive when issues of high PAR, frequency errors, and delay become critical. Exploiting the rich structure of space–time codes is critical in simplifying the channel estimation and equalization schemes.

Acknowledgments

We would like to thank the following colleagues (in alphabetical order) for many technical discussions and contributions to this work: G. Bauch, S. N. Diggavi, C. Fragouli, N. Seshadri, A. Stamoulis, V. Tarokh, and W. Younis.

BIOGRAPHIES.

Naofal Al-Dhahir received his M.S. and Ph.D. degrees from Standford University in 1990 and 1994, respectively, in electrical engineering. He was as instructor at Stanford University during Winter 1993. From August 1994 to 1999, he was a member of the technical staff at the Communications Program at GE Corporate R&D Center

in Schenectady, New York, where he worked on various aspects of satellite communication systems design and anti-jam GPS receivers. Since August 1999, he has been a principal member of technical staff at AT&T Shannon Laboratory in Florham Park, New Jersey. His current research intersts include equalization schemes, space-time coding and signal processing, OFDM, and digital subscriber line technology. He has authored more than 40 journal papers and holds 7 U.S. patents in the areas of satellite communications, digital television, and space-time processing. He is a senior member of the IEEE and a member of the IEEE SP4COM technical committee. He is editor for *IEEE Transaction on Signal Processing*, *IEEE Communications Letters*, and *IEEE Transactions on Communications*. He is co-author of the book *Dopler Applications for LEO Satellite System* (Kluwer 2001).

Robert Calderbank is vice president for research at AT&T. He also is responsible for directing the research program in Internet and network systems. This program provides AT&T with technical and industry leadership in all areas of networking technology. These areas include network security, content distribution, operations support, network measurement and management, and end-to-end optical systems.

Dr. Calderbank is an IEEE and AT&T Fellow, and a recipient of the IEEE Third Millennium Medal for his contributions to digital communications. These include the design of high-speed voiceband modems, the development of advanced read channels for magnetic disk storage, and the invention of space-time codes, a breakthrough wireless technology that uses a small number of antennas to significantly improve throughput and reliability.

Ayman Naguib received the B.Sc. Degree (with honors) and the M.S. degree in electrical engineering from Cairo University, Cario, Egypt, in 1987 abd 1990, respectively, and the M.S. degreee in statistics and the Ph.D. degree in electrical engieering from Stanford University, Stanford, California, in 1993 and 1996, respectively.

From 1987 to 1989, he spent his military service at the Signal Processing Laboratory, The Military Technical College, Cairo, Egypt. From 1989 to 1990, he was employed at Cairo University as a research and teaching assistant in the Communication Theory Group, Department of Electrical Engineering. From 1990 to 1995, he was a research and teaching assistant in the Information Systems Laboratories, Stanford University, Stanford California. In 1996, he joined AT&T Labs, Florham Park, New Jersey, as a principal member of technical staff. In September 2000, he joined Morphics Technology Inc. as a technical leader. His current research interests include in general space-time signal processing and coding for high data rate wireless communications (W-CDMA, OFDM, etc.).

BIBLIOGRAPHY

1. Special Issue on the European Path Towards UMTS, *IEEE Pers. Commun. Mag.* **2**: (Feb. 1995).
2. D. J. Goodman, Trends in cellular and cordless communications, *IEEE Commun. Mag.* **29**: 31–40 (June 1991).
3. J. H. Winters, Optimum combining in digital mobile radio with cochannel interference, *IEEE J. Select. Areas Commun.* **JSAC-2**(4): 528–539 (July 1984).
4. J. H. Winters, Optimum combining for indoor radio systems with multiple users, *IEEE Trans. Commun.* **COM-35**(11): 1222–1230 (Nov. 1987).
5. J. H. Winters, On the capacity of radio communication systems with diversity in a Rayleigh fading environment, *IEEE J. Select. Areas Commun.* **JSAC-5**(5): 871–878 (June 1987).
6. P. Balaban and J. Salz, Optimum diversity combining and equalization in digital data transmission with application to cellular mobile radio, *IEEE Trans. Vehic. Technol.* **VT-40**(2): 342–354 (May 1991).
7. P. Balaban and J. Salz, Optimum diversity combining and equalization in data transmission with application to cellular mobile radio — Part I: Theoretical considerations, *IEEE Trans. Commun.* **COM-40**(5): 885–894 (May 1992).
8. P. Balaban and J. Salz, Optimum diversity combining and equalization in data transmission with application to cellular mobile radio — Part II: Numerical results, *IEEE Trans. Commun.* **COM-40**(5): 895–907 (May 1992).
9. G. J. Foschini and M. J. Gans, On limits of wireless communications in a fading environment when using multiple antennas, *Wireless Commun. Mag.* **6**: 311–335 (March 1998).
10. E. Telatar, *Capacity of Multi-Antenna Gaussian Channels*, technical memorandum, AT&T Bell Laboratories, June 1995.
11. G. Foschini, Layered space–time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Tech. J.* **1**: 41–59 (1996).
12. U.S. Patent 5,088,108 (Feb., 1992), J. Uddenfeldt and A. Raith, Cellular digital mobile radio system and method of transmitting information in a digital cellular mobile radio system.
13. A. Wittneben, Base station modulation diversity for digital SIMULCAST, *Proc. IEEE VTC'91*, St. Louis, MO, 1991, Vol. 1, pp. 848–853.
14. N. Seshadri and J. H. Winters, Two schemes for improving the performance of frequency-division duplex (FDD) transmission systems using transmitter antenna diversity, *Int. J. Wireless Inform. Networks* **1**: 49–60 (Jan 1994).
15. A. Wittneben, A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation, *Proc. IEEE ICC'93*, Geneva, Switzerland, 1993, Vol. 3, pp. 1630–1634.
16. J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels, *Proc. IEEE VTC'96*, Atlanta, GA, 1996, Vol. 1, pp. 136–140.
17. J. H. Winters, Diversity gain of transmit diversity in wireless systems with Rayleigh fading, *Proc. IEEE ICC'94*, New Orleans, LA, 1994, Vol. 2, pp. 1121–1125.
18. S. Alamouti, Space block coding: A simple transmitter diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1451–1458 (Oct. 1998).
19. V. Tarokh, H. Jafarkhani, and R. A. Calderbank, Space–time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* **45**: 1456–1467 (July 1999).
20. N. Seshadri, V. Tarokh, and A. R. Calderbank, Space–time codes for high data rate wireless communications: Code

construction, *Proc. IEEE VTC'97*, Phoenix, AZ, 1997, Vol. 2, pp. 637–641.

21. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space–time codes for high data rate wireless communications: performance criterion and code construction, *Proc. IEEE ICC'97*, Montreal, Canada, 1997, Vol. 1, pp. 299–303.
22. V. Tarokh, N. Seshadri, and A. R. Calderbank, Space–time codes for high data rate wireless communications: Performance criterion and code construction, *IEEE Trans. Inform. Theory* **44**: 744–765 (March 1998).
23. V. Tarokh, A. F. Naguib, N. Seshadri, and A. R. Calderbank, Space–time codes for high data rate wireless communications: Mismatch analysis, *Proc. IEEE ICC'97*, Montreal, Canada, 1997, Vol. 1, pp. 309–313.
24. V. Tarokh, A. F. Naguib, N. Seshadri, and A. Calderbank, Space–time codes for high data rate wireless communications: Performance criteria in the presence of channel estimation errors, mobility, and multiple paths, *IEEE Trans. Commun.* **47**: 199–207 (Feb. 1999).
25. A. F. Naguib, V. Tarokh, N. Seshadri, and A. R. Calderbank, A space–time coding based modem for high data rate wireless communications, *IEEE J. Select. Areas Commun.* **16**: 1459–1478 (Oct 1998).
26. W. C. Jakes, *Microwave Mobile Communications*, IEEE Press, 1974.
27. J. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
28. A. F. Naguib and N. Seshadri, Combined interference cancellation and ML decoding of space–time block codes, *IEEE J. Select. Areas Commun.* (in press).
29. T. Rappaport, *Wireless Communications*, IEEE Press, 1996.
30. L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**: 284–287 (March 1974).
31. J. Hagenauer, The Turbo principle: Tutorial introduction and state of the art, *Proc. Int. Symp. Turbo Codes*, Sept. 1997, pp. 1–11.
32. C. Douillard et al., Iterative correction of intersymbol interference: Turbo equalization, *Eur. Trans. Telecommun.* 507–511 (Sept.–Oct. 1995).
33. N. Al-Dhahir, FIR channel-shortening equalizers for MIMO ISI channels, *IEEE Trans. Commun.* **50**: 213–218 (Feb. 2001).
34. G. Bauch and N. Al-Dhahir, Iterative equalization and decoding with channel shortening filters for space–time-coded modulation, in *Vehicular Technology Conference Fall*, pp. 1575–1582, 2000.
35. A. Furuskar, S. Mazur, F. Muller, and H. Olofsson, EDGE: Enhanced data rates for GSM and TDMA/136 evolution, *IEEE Pers. Commun. Mag.* 56–66 (June 1999).
36. G. Bauch and A. Naguib, MAP equalization of space–time-coded signals over frequency-selective channels, in *Wireless Communications and Networking Conference*, pp. 261–265, Sept. 1999.
37. V. Franz and J. Anderson, Concatenated decoding with a reduced-search BCJR algorithm, *IEEE J. Select. Areas Commun.* 186–195 (Feb. 1998).
38. C. Fragouli, N. Al-Dhahir, S. Diggavi, and W. Turin, Prefiltered M-BCJR equalizer for frequency-selective channels, in *Conference on Information Sciences and Systems*, March 2001.
39. J. Hagenauer and P. Hoeher, A Viterbi algorithm with soft-decision outputs and its applications, *Global Telecommunications Conf.*, Nov. 1989, pp. 47.1.1–47.1.7.
40. W. V. Etten, Maximum likelihood receiver for multiple channel transmission systems, *IEEE Trans. Commun.* 276–283 (Feb. 1976).
41. A. Naguib and N. Seshadri, MLSE and equalization of space–time-coded signals, in *Vehicular Technology Conference Spring*, pp. 1688–1693, May 2000.
42. A. Duel-Hallen and C. Heegard, Delayed decision-feedback sequence estimation, *IEEE Trans. Commun.* 428–436 (May 1989).
43. N. Al-Dhahir and A. H. Sayed, The finite-length MIMO MMSE-DFE, *IEEE Trans. Signal Process.* 2921–2936 (Oct. 2000).
44. W. Younis and N. Al-Dhahir, FIR prefilter design for MLSE equalization of space–time-coded transmission over multipath fading channels, in *International Symposium on Circuits and Systems*, May 2001, 362–365.
45. S. Weinstein and P. Ebert, Data transmission by frequency-division multiplexing using the discrete fourier transform, *IEEE Trans. Commun.* **19**: 628–634 (Oct. 1971).
46. G. Raleigh and J. Cioffi, Spatio-temporal coding for wireless communication, *IEEE Trans. Commun.* 357–366 (March 1998).
47. D. Agrawal, V. Tarokh, A. Naguib, and N. Seshadri, Space–time-coded OFDM for high data-rate wireless communication over wideband channels, in *Vehicular Technology Conference*, pp. 2232–2236, May 1998.
48. E. Lindskog and A. Paulraj, A transmit diversity scheme for delay spread channels, in *International Conference on Communications*, pp. 307–311, June 2000.
49. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **18**: 363–378 (May 1972).
50. J. Salz, Optimum mean-square decision feedback equalization, *Bell Syst. Tech. J.* **52**: 1341–1373 (Oct. 1973).
51. Z. Liu, G. Giannakis, A. Scaglione, and S. Barbarossa, Decoding and equalization of unknown multipath channels based on block precoding and transmit-antenna diversity, *Asilomar Conf. Signals, Systems, and Computers*, 1999, pp. 1557–1561.
52. T. Pollet, M. V. Bladel, and M. Moeneclaey, BER sensitivity of OFDM systems to carrier frequency offset and Wiener phase noise, *IEEE Trans. Commun.* 191–193 (Feb.–April 1995).
53. H. Sari, G. Karam, and I. Jeanclaude, Transmission techniques for digital terrestrial TV broadcasting, *IEEE Commun. Mag.* 100–109 (Feb. 1995).
54. N. Al-Dhahir, *Single-Carrier Frequency-Domain Equalization for Space–Time Block-Coded Transmissions over Frequency-Selective Fading Channels*, IEEE Communications Letters, Vol. 5, July 2001, pp. 304–306.
55. S. Crozier, D. Falconer, and S. Mahmoud, Least sum of squared errors (LSSE) channel estimation, *IEE Proc. Part F*, Aug. 1991, pp. 371–378.
56. D. Chu, Polyphase codes with good periodic correlation properties, *IEEE Trans. Inform. Theory* **IT-18**: 531–532 (July 1972).
57. C. Fragouli, N. Al-Dhahir, and W. Turin, *Channel Estimation for Space–Time-Coded Systems*, AT&T Technical Document 4TKQBS, Feb. 2001.

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 5

WILEY ENCYCLOPEDIA OF TELECOMMUNICATIONS

WILEY ENCYCLOPEDIA OF

TELECOMMUNICATIONS

VOLUME 5

John G. Proakis
Editor

WILEY-INTERSCIENCE

A John Wiley & Sons Publication

The *Wiley Encyclopedia of Telecommunications* is available online at
http://www.mrw.interscience.wiley.com/eot

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Wiley encyclopedia of telecommunications / John G. Proakis, editor.
p. cm.
includes index.
ISBN 0-471-36972-1
1. Telecommunication — Encyclopedias. I. Title: Encyclopedia of telecommunications. II. Proakis, John G.
TK5102 .W55 2002
621.382′03 — dc21

2002014432

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

SPATIOTEMPORAL SIGNAL PROCESSING IN WIRELESS COMMUNICATIONS

DIMITRIOS HATZINAKOS
RYAN A. PACHECO
University of Toronto
Toronto, Ontario, Canada

1. INTRODUCTION

Wireless channels are severely limited by fading due to multipath propagation. When fading is "flat," the receiver experiences fast or slow variations of the received signal amplitude, which makes detection of the information symbols difficult. Fading can also be frequency-selective, in which case in addition to signal variations we experience intersymbol interference. An effective strategy to deal with multipath propagation and fading has been the utilization of multiple antennas at the receiver and or at the transmitter. Beamforming with multiple antennas closely spaced together has been traditionally used for directive transmission and rejection of multipath components. However, this approach requires line-of-sight communications and the formation of steerable multiple beams in the case of mobile multiuser communications. Another approach is diversity, which is based on the weighted combination of multiple signal copies that are subject to independent fading. Therefore, two adjacent receive or transmit antennas must be apart by several wavelengths so that signals transmitted or received from different antennas are sufficiently uncorrelated. The latter approach is considered in this article.

The information-theoretic capacity of multiple antenna systems and data communication techniques that exploit the potential of spatial diversity have been studied [3,4], and dramatic increases in supportable data rates are expected. In such systems, the capacity can theoretically increase by a factor up to the number of transmit and receive antennas in the array. Spacetime diversity and coding techniques, such as trellis spacetime codes, block spacetime codes, and variations of the very popular BLAST (Bell Laboratories layered spacetime) architecture, among others [5,7–9] are currently under investigation. However, there is a need for more thorough investigation of the potential and limits of such technologies and the development of efficient low complexity systems. In the first half of this article, we provide a general treatment of the spacetime signal processing scenario and spacetime diversity and coding frameworks based on block spacetime codes and discuss potential benefits, requirements, and limitations.

A common assumption in existing spacetime coding literature is the knowledge or availability of a good estimate for the multipath channel [3,7]. In the absence of channel knowledge, the capacity gains to be achieved depend on the particular characteristics (e.g., coherence time) of the channel. Various schemes, consisting of an antenna array and a spatio-temporal processor at the base station (or at the mobile unit), have been proposed to mitigate the effects of intersymbol interference (ISI), and to reject cochannel interference (CCI) or multiple-access interference (MAI) induced by simultaneous intra- or intercell users [5,17]. To cope with the time-varying and fading nature of wireless channels, data are transmitted in bursts, and attached to each burst is a training sequence of short duration that is used to help the receiver recover the unknown parameters of the communication channel [10]. In many cases unsupervised (i.e., blind) techniques [12,21], which require no training data, as well as semiblind approaches, which provide a synergistic treatment of training and blind principles, are employed to address the channel estimation and/or the direct channel equalization problems [1,11,14]. In the second half of the article, a semiblind channel equalization approach will be presented and integrated with spacetime diversity and coding to form a realistic scenario for detection and estimation.

The aim of the article is by no means to provide an exhaustive coverage but rather to illustrate the principle and potential and to suggest a realistic approach for implementation of spacetime wireless transceivers, and also to provide a motivation for exploring the fast-growing literature and technological developments in the area.

2. THE DISCRETE-CHANNEL MODEL

The basic structure of a wireless transceiver employing spacetime diversity and coding (multiple-input multiple-output system) is depicted in Fig. 1.

The received equivalent discrete-time signal at the receiver antenna a, $a = 1, \ldots, A$, after demodulation, filtering, and oversampling (i.e., taking N samples per symbol period) can be written as [1]

$$r_a[n] = \sum_{m=1}^{M} \sum_{k=0}^{N_b-1} b_m[k] \cdot g_{ma}[n - kN] + v_a[n] \qquad (1)$$

where N_b is the length of the transmitted burst of data and the $\{b_m[k]\}_{k=0}^{N_b-1}$ denote the data symbols transmitted from the antenna $m, m = 1, \ldots, M$. When spacetime coding is employed in the transmitter, we may assume that M

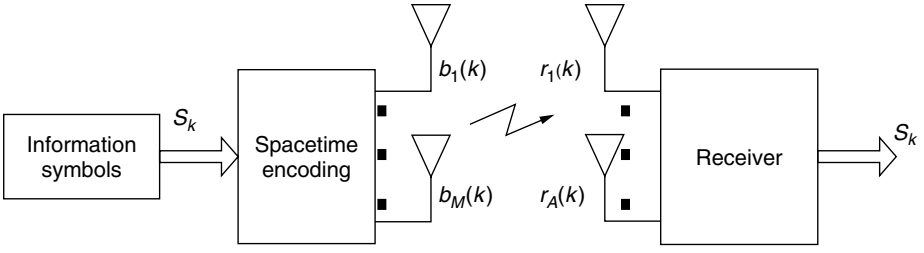


Figure 1. Single-user communication using multiple transmit and receive antennas.

complex symbols are transmitted simultaneously from the M transmit antennas. In this case, we may write [3]

$$\mathbf{b}[k] = \sum_{l=1}^{M} \mathbf{q}_l^r[k] \cdot s_l^r + \sum_{l=1}^{M} j\mathbf{q}_l^j[k] \cdot s_l^j \tag{2}$$

where $\mathbf{b}(k) = [b_1(k), \ldots, b_M(k)]^T$ is the transmitted data vector at time instant k, $s_l = s_l^r + js_l^j, l = 1, \ldots, M$ are the transmitted symbols and $\mathbf{q}_l^r(k)$ and $\mathbf{q}_l^j(k)$ are the $M \times 1$ transmit antenna weight vectors for the real and imaginary part of the symbol l, respectively, at time instant k. The $g_{ma}[n], n = 0, \ldots, LN-1$ accounts for the transmit and receive filters and the multipath channel between transmit antenna m and receive antenna a. Without loss of generality we will assume that all the channels between transmit and receive antennas are of the same length $LN-1$ samples and are time-invariant during the transmission of a burst of N_b data symbols. Finally $v_a[n]$ is circular complex white Gaussian noise, uncorrelated with $b_m[n] \forall m$, with zero mean and variance σ^2.

Assume that we are interested in recovering the mth transmitted symbol at time instant k. To do this we first collect N samples from all antenna elements to form the vector $\mathbf{r}(k)$, which takes the following form:

$$\mathbf{r}(k) = [\mathbf{G}(L-1), \ldots, \mathbf{G}(0)] \cdot [\mathbf{b}(k-L+1)^T, \ldots, \mathbf{b}(k)^T]^T + \mathbf{v}(k) \tag{3}$$

where

$$\mathbf{r}(k) = \begin{bmatrix} r_1[kN] \\ \vdots \\ r_1[(k+1)N-1] \\ \vdots \\ r_A[kN] \\ \vdots \\ r_A[(k+1)N-1] \end{bmatrix}_{AN\times 1}$$

$$\mathbf{G}(l) = \begin{bmatrix} g_{11}[lN] & \cdots & g_{M1}[lN] \\ \vdots & & \vdots \\ g_{11}[(l+1)N-1] & \cdots & g_{M1}[(l+1)N-1] \\ \vdots & & \vdots \\ g_{1A}[lN] & \cdots & g_{MA}[lN] \\ \vdots & & \vdots \\ g_{1A}[(l+1)N-1] & \cdots & g_{MA}[(l+1)N-1] \end{bmatrix}_{AN\times M}$$

$$\mathbf{b}(k) = \begin{bmatrix} b_1[k] \\ \vdots \\ b_M[k] \end{bmatrix}_{M\times 1} \qquad \mathbf{v}(k) = \begin{bmatrix} v_1[kN] \\ \vdots \\ v_1[(k+1)N-1] \\ \vdots \\ v_A[kN] \\ \vdots \\ v_A[(k+1)N-1] \end{bmatrix}_{AN\times 1}$$

In general, the simultaneous transmission of M symbols will span μ received vectors. Thus, it may be necessary to process more than one received vector at a time in order to estimate the mth symbol. Then, assuming a linear receiver structure and in the absence of noise, the recovery of the real information symbol s_m^q, where $q = r$ or $q = j$, at time instant k, requires that the $A \times 1$ weight vectors $\mathbf{W}_{i,m}^q$ of the spacetime equalizer at $i = k, \ldots, k+\mu-1$ satisfy the relation

$$\sum_{i=k}^{k+\mu-1} \mathbf{W}_{i,m}^{qH} \cdot \mathbf{r}(i) = s_m^q[k-d] \tag{4}$$

where d is an arbitrary delay. Throughout the article the symbols $*$, T, and H denote conjugate, transpose, and conjugate transpose operations, respectively.

3. SPACETIME DIVERSITY AND FLAT-FADING CHANNELS

Here we are going to examine the benefits that can be expected by employing both transmit and receive diversities. We follow a procedure based on an excellent analysis [3] of this problem. To simplify the mathematical formulas, we assume a flat-fading channel described by the constant matrix $\mathbf{G}$ and that $N = 1$, with no oversampling. Then, by placing $L = 1, N = 1$, and $\mathbf{G}(0) = \mathbf{G}$ into the relation of Section 2, we obtain the following simplified relation between the $A \times 1$ receive signal vector $\mathbf{r}(k)$ and the $M \times 1$ transmit data symbol vector $\mathbf{b}(k)$:

$$\mathbf{r}(k) = \mathbf{G} \cdot \mathbf{b}(k) + \mathbf{v}(k) \tag{5}$$

where

$$\mathbf{r}(k) = \begin{bmatrix} r_1[k] \\ \vdots \\ r_A[k] \end{bmatrix}_{A\times 1} \qquad \mathbf{G} = \begin{bmatrix} g_{11} & \cdots & g_{M1} \\ \vdots & & \vdots \\ g_{1A} & \cdots & g_{MA} \end{bmatrix}_{A\times M}$$

$$\mathbf{b}(k) = \begin{bmatrix} b_1[k] \\ \vdots \\ b_M[k] \end{bmatrix}_{M\times 1} \qquad \mathbf{v}(k) = \begin{bmatrix} v_1[k] \\ \vdots \\ v_A[k] \end{bmatrix}_{AN\times 1}$$

The transmitted data vector at time instant k is given by Eq. (2). Then, according to Eq. (4), the output of the linear spatiotemporal receiver filter (detector) for the real symbol $s_m^q, q = r$ or $q = j, m = 1, \ldots, M$ is

$$\begin{aligned} \hat{s}_m^q &= Real\left\{ \sum_{i=k}^{k+\mu-1} \mathbf{W}_{i,m}^{qH} \cdot \mathbf{r}(i) \right\} \\ &= Real\left\{ \sum_{i=k}^{\mu+k-1} \mathbf{W}_{i,m}^{qH} \mathbf{G} \sum_{l=1}^{M} \mathbf{q}_l^r(i) s_l^r \right. \\ &\quad \left. + \sum_{i=k}^{\mu+k-1} \mathbf{W}_{i,m}^{qH} \mathbf{G} \sum_{l=1}^{M} j\mathbf{q}_l^j(i) s_l^j + \sum_{i=k}^{\mu+k-1} \mathbf{W}_{i,m}^{qH} \mathbf{v}(i) \right\} \end{aligned} \tag{6}$$

It can be shown [3,6] that the maximum signal to noise ratio (SNR) for this linear spatiotemporal receiver is achieved when (within a scale term)

$$\mathbf{W}_{i,m} = \mathbf{G}\mathbf{q}_m^r(i) \qquad \text{or} \qquad \mathbf{W}_{i,m} = \mathbf{G}_j\mathbf{q}_m^j(i) \tag{7}$$

for s_m^r and s_m^j, respectively. By substituting these values in Eq. (6), we determine the detector output $\hat{s}_m^r$ for s_m^r, where $m = 1, \ldots, M$:

$$\begin{aligned}\hat{s}_m^r &= Real\left\{\sum_{i=k}^{\mu+k-1} \mathbf{q}_m^{rH}(i)\mathbf{G}^H\mathbf{G}\sum_{l=1}^{M}\mathbf{q}_l^r(i)s_l^r\right\} \\ &\quad + Real\left\{\sum_{i=k}^{\mu+k-1} \mathbf{q}_m^{rH}(i)\mathbf{G}^H\mathbf{G}\sum_{l=1}^{M}j\mathbf{q}_l^j(i)s_l^j\right\} \\ &\quad + Real\left\{\sum_{i=k}^{\mu+k-1} \mathbf{q}_m^{rH}(i)\mathbf{G}^H\mathbf{v}(i)\right\} \\ &= Real\left\{\sum_{l=1}^{M} trace(\mathbf{G}\mathbf{Q}_l^r\mathbf{Q}_m^{rH}\mathbf{G}^H)s_l^r\right\} \\ &\quad + Real\left\{\sum_{l=1}^{M} j\,trace(\mathbf{G}\mathbf{Q}_l^j\mathbf{Q}_m^{rH}\mathbf{G}^H)s_l^j\right\} \\ &\quad + Real\left\{\sum_{i=k}^{\mu+k-1} \mathbf{q}_m^{rH}(i)\mathbf{G}^H\mathbf{v}(i)\right\} \qquad (8)\end{aligned}$$

where $\mathbf{Q}_m^r = [\mathbf{q}_m^r(k), \ldots, \mathbf{q}_m^r(k+\mu-1)]_{M\times\mu}$ and $\mathbf{Q}_m^j = [\mathbf{q}_m^j(k), \ldots, \mathbf{q}_m^j(k+\mu-1)]_{M\times\mu}$ are the transmitter weight matrices for symbols s_m^r and s_m^j, respectively. Similarly, we find the detector output $\hat{s}_m^j$ for s_m^j, where $m = 1, \ldots, M$:

$$\begin{aligned}\hat{s}_m^j &= -Real\left\{\sum_{l=1}^{M} trace(\mathbf{G}\mathbf{Q}_l^j\mathbf{Q}_m^{jH}\mathbf{G}^H)s_l^j\right\} \\ &\quad + Real\left\{\sum_{l=1}^{M} j\,trace(\mathbf{G}\mathbf{Q}_l^r\mathbf{Q}_m^{jH}\mathbf{G}^H)s_l^r\right\} \\ &\quad + Real\left\{\sum_{i=k}^{\mu+k-1} \mathbf{q}_m^{jH}(i)\mathbf{G}^H\mathbf{v}(i)\right\} \qquad (9)\end{aligned}$$

By properly choosing the weight matrices $\mathbf{Q}_m^r, \mathbf{Q}_m^j, m = 1, \ldots, M$, the interference from other symbols can be significantly reduced or completely eliminated and the output SNR maximized. In this case, a transmit diversity gain up to the order M and a transmit–receive diversity gain of order up to AM is possible. Another way of looking at the above procedure is that of mapping the set of symbols $\{s_l\}$ to the set of vectors $\{\mathbf{b}(k)\}$, that is

$$[s_1 s_2 \cdots s_M]_{1\times M} \rightarrow [\mathbf{b}(k)\mathbf{b}(k+1)\cdots\mathbf{b}(k+\mu-1)]_{M\times\mu} \quad (10)$$

which is a form of spacetime block code. Thus, the objective is to design efficient such codes. It can be shown that in general this is possible only if $\mu \geq M$ [3–5]. Next we provide a few examples to demonstrate some important points.

Example 1. Let us consider the transmission of two complex symbols s_1, s_2 from a transmitter with two antennas by using the above mentioned process ($M = 2$) and let $\mu = 2$. First let us consider the following choice of weight matrices

$$\mathbf{Q}_1^r = [\mathbf{q}\ \mathbf{0}]\mathbf{Q}_2^r = [\mathbf{0}\ \mathbf{q}]\mathbf{Q}_1^j = [\mathbf{q}\ \mathbf{0}]\mathbf{Q}_2^j = [\mathbf{0}\ \mathbf{q}]$$

where $\mathbf{q}$ is a 2×1 weight vector. The transmitted complex vectors are

$$\mathbf{b}(k) = [\mathbf{q}s_1], \quad \mathbf{b}(k+1) = [\mathbf{q}s_2]$$

This is clearly the situation of *Space-only* transmission. Then, since

$$\mathbf{Q}_l^x\mathbf{Q}_m^{xH} = \mathbf{0}, \quad x = r, j, \quad m \neq l$$

interference from different symbols is eliminated. Furthermore since

$$\mathbf{Q}_m^x\mathbf{Q}_m^{yH}, \quad m = 1, 2$$

are Hermitian, the terms

$$Trace(\mathbf{G}\mathbf{Q}_m^x\mathbf{Q}_m^{yH}\mathbf{G}^H), \quad m = 1, 2$$

are real. Thus, equations (8) and (9) are reduced to

$$\begin{aligned}D_m^x &= \pm trace(\mathbf{G}\mathbf{q}\mathbf{q}_m^H\mathbf{G}^H)s_m^x + \mathbf{q}^H\mathbf{G}^H\mathbf{v} \\ &\quad x = r, j \quad m = 1, 2 \qquad (11)\end{aligned}$$

and the corresponding SNR per real symbol is easily found to be

$$\text{SNR} = \frac{E_s}{\sigma^2} trace(\mathbf{G}\mathbf{q}\mathbf{q}^H\mathbf{G}^H)$$

where E_s is the energy of the real symbol and σ^2 the variance of the white noise.

Example 2. Let again $M = 2$ and $\mu = 2$ but this time choose the weight matrices as follows [3–5]:

$$\mathbf{Q}_1^r = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \quad \mathbf{Q}_2^r = \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}, \quad \mathbf{Q}_1^j = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix},$$
$$\mathbf{Q}_2^j = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$$

and the transmitted complex vectors are

$$\mathbf{b}(k) = \begin{bmatrix}s_1\\ -s_2^*\end{bmatrix}, \quad \mathbf{b}(k+1) = \begin{bmatrix}s_2\\ s_1^*\end{bmatrix}$$

In this case of *spacetime* processing it is easy to show by following arguments similar to those in example 1 that all interference is eliminated and since

$$\mathbf{Q}_m^x\mathbf{Q}_m^{xH} = \mathbf{I}_{2\times 2}, \quad m = 1, 2$$

are identity matrices

$$\begin{aligned}D_m^x &= \pm trace(\mathbf{G}\mathbf{G}^H)s_m^x + \sum_{i=k}^{k+1}\mathbf{q}_m^{xH}(i)\mathbf{G}^H(0)\mathbf{v}(i) \\ &\quad x = r, j \quad m = 1, 2 \qquad (12)\end{aligned}$$

and the corresponding SNR per real symbol is found to be

$$\text{SNR} = \frac{E_s}{\sigma^2} trace(\mathbf{G}\mathbf{G}^H) = \frac{E_s}{\sigma^2}\sum_{i=1}^{2}\sum_{j=1}^{2}|g_j^i|^2$$

Example 3. Consider now the case of $M = 1, A > 1$ corresponding to pure receive diversity. In this case the matrix G reduces to a vector $A \times 1$. Similarly, the transmit weight matrices reduce to scalars and after some calculations we find that the corresponding maximum SNR per real symbol is

$$\text{SNR} = \frac{E_s}{\sigma^2} \sum_{i=1}^{A} |g_1^i|^2$$

which corresponds to the case of *Maximal Ratio Combining*.

From these examples, we conclude the following:

1. In the case of space-only processing (Example 1), it is important that both the transmitter and the receiver know the channel, that is, the matrix **G**, to maximize the output SNR. This is achieved when the transmit weight vector **q** is equal to the eigenvector of **G** corresponding to the largest eigenvalue.
2. On the other hand, in the case of spacetime processing of Example 2, the transmitter does not need to know the channel. Obviously the receiver still requires knowledge of the channel for decoding.
3. Examples 2 and 3 indicate that the spacetime diversity gain with M transmit and A receive antennas is of the order $M \cdot A$. Thus, in Example 3, in order to achieve similar gains as in Example 2, we must choose $A = 4$. Similar arguments can be made for the case of transmit only diversity where $M > 1, A = 1$.
4. Spacetime diversity exploits the propagation environment itself to improve the performance. The *richer* the multipath channel (matrix **G**), the higher the diversity gain is.

4. SPACETIME DIVERSITY AND SEMIBLIND CHANNEL EQUALIZATION

In multiaccess communications, channel resources are allocated in ways that require either strict cooperation among users, such as in time- or frequency-division multiple access (TDMA or FDMA), or not, as in code- or space-division multiple access (DS-CDMA or SDMA). DS-CDMA and SDMA systems generally require more complex receivers than do TDMA or FDMA since the received data may contain interference from other users (multiuser interference, also called *multiple-access interference* (MAI)). The amount of interference that one user contributes to another is dependent on the orthogonality between their received signals. In DS-CDMA systems, orthogonality may exist on the transmitting end by assigning orthogonal spreading codes to users but will not exist on the receiver end because of propagation effects (multipath propagation in the case of wireless channels). The end result is that correlation-type receivers are no longer viable and some method of multiuser detection is needed in order to deal effectively with MAI. Many different multiuser detection strategies have been proposed over the past decade from prohibitively complex [20] to extremely simple [13].

In addition to MAI there is also intersymbol interference (ISI), or *self-interference*, that is also caused by multipath propagation. ISI is negligible in low-rate DS-CDMA applications but is becoming more of an issue in third-generation systems (where the symbol duration is on the order of the multipath delay spread [10]). Consequently, high-rate DS-CDMA systems suffer from MAI and ISI. Thus, receivers for such systems need to perform equalization and multiuser detection.

When MAI and ISI are present, an analytic treatment of the spacetime diversity and coding scenario is not as straightforward as with flat-fading single-user channels. In any case, the unknown wireless channel must be estimated at the receiver or directly equalized for cancellation of the interference and detection of the desired data symbols. In this section, we complement the spacetime diversity and coding principles described in the previous section with a spacetime semiblind channel equalization technique and propose a realistic and effective wireless communications transceiver.

4.1. Revised Signal Model

A K-user DS-CDMA system is considered. Each user is assigned a unique spreading code with a processing gain of N. In addition, each user transmits from an M-element antenna array to an antenna array with A elements. The block diagram of such a transceiver is depicted in Fig. 2. Let $b_m^k[n]$ be the nth symbol for the kth user from the mth transmit antenna. We assume a 4-QAM modulation scheme $b_m^k[n] \in \{e^{\pm j(\pi/4)}, e^{\pm j(3\pi/4)}\}$. Summing over all transmit antennas for all users, we can rewrite the received signal at the ath receive antenna

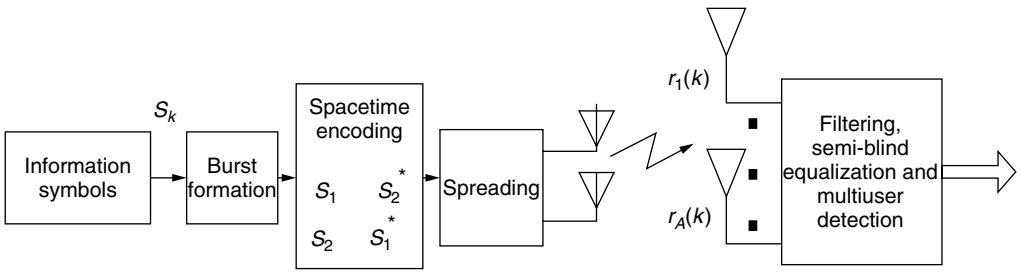


Figure 2. kth-user transceiver using 2 transmit and A receive antennas including spacetime coding and channel equalization.

is as

$$r_a[n] = \sum_{k=1}^{K} \sum_{m1}^{M} \sum_{k=0}^{N_b-1} b_m^k[n_b] g_{ma}^k[n - Nn_b] + v_a[n] \tag{13}$$

where now $g_{ma}^k[n]$ represents the combined impulse response of channel plus spreading code for the kth user from the mth transmit antenna to the ath receive antenna:

$$g_{ma}^k[n] = \sum_{l=0}^{L_k-1} \beta_{ma} k[l] c_k[n-l] \tag{14}$$

In this equation $\{\beta_{ma}^k[l]\}_{l=0}^{L_k-1}$ represents the frequency-selective multipath channel for the given user. The $c_k[n]$ is the normalized spreading code for the kth user: $c_k[n] \in \{\pm 1\}/\sqrt{N}$. Once again, it is assumed that the fading coefficients do not change during the transmission of N_b data symbols and that $\max(L_k - 1) \le (L-1)N$ for some integer L.

Collecting the N chips corresponding to the n_bth transmitted symbol, the received vector is written as

$$\mathbf{r}_a(n_b) = \mathbf{G}_a \cdot \mathbf{b}_{n_b} + \mathbf{v}_a(n_b) \tag{15}$$

where

$$\begin{aligned}
\mathbf{r}_a(l) &= [r_a[lN], \ldots, r_a[(l+1)N-1]]^T \\
\mathbf{G}_a &= [\mathbf{G}_a(L-1), \ldots, \mathbf{G}_a(0)] \\
\mathbf{G}_a(l) &= [\mathbf{g}_{1a}^1(l), \mathbf{g}_{2a}^1(l), \ldots, \mathbf{g}_{Ma}^K(l)] \\
\mathbf{g}_{ia}^k(l) &= [g_{ia}^k[lN], \ldots, g_{ia}^k[(l+1)N-1]]^T \\
\mathbf{b}_{n_b} &= [\mathbf{b}(n_b - L + 1)^T, \ldots, \mathbf{b}(n_b)^T]^T \\
\mathbf{b}(l) &= [b_1^1[l], b_2^1[l], \ldots, b_M^K[l]]^T \\
\mathbf{v}_a(l) &= [v_a[lN], \ldots, v_a[(l+1)N-1]]^T
\end{aligned}$$

4.2. Interference Suppression Strategies

We now fix the number of transmit antennas per user (M) at two and apply the block code of Example 2 (this is similar to the Alamouti scheme [7]). In this case, symbols for the kth user, $\{b_k[n]\}_{n=0}^{N_b-1}$, are mapped as follows for each pair of symbols $(\{b_k[0], b_k[1]\}, \{b_k[2], b_k[3]\}, \ldots)$:

$$\begin{aligned}
b_1^k[0] &= \frac{b_k[0]}{\sqrt{2}}, \quad b_2^k[0] = \frac{-b_k^*[1]}{\sqrt{2}} \\
b_1^k[1] &= \frac{b_k[1]}{\sqrt{2}}, \quad b_2^k[1] = \frac{b_k^*[0]}{\sqrt{2}}
\end{aligned}$$

The additional normalization by $\sqrt{2}$ has been introduced so that the total transmitted energy remains constant and independent of M. In the receiver we employ two linear filters. The first, $\mathbf{W}_1$, is trained to extract the even-numbered symbols $(b_1[0], b_1[2], \ldots)$ and the other, $\mathbf{W}_2$, is trained to extract the odd-numbered symbols (in this case user number one is the desired user).

4.2.1. Least-Squares Optimization.

Using least squares, the solution then becomes

$$\mathbf{W}_1^{\mathrm{LS}} = \arg\min_{\mathbf{W}} \frac{2}{N_t} \sum_{i=0}^{N_t/2-1} |\mathbf{W}^H \mathbf{r}(i) - b_1[2i]|^2 \tag{16}$$

$$\mathbf{W}_2^{\mathrm{LS}} = \arg\min_{\mathbf{W}} \frac{2}{N_t} \sum_{i=0}^{N_t/2-1} |\mathbf{W}^H \mathbf{r}(i) - b_1^*[2i+1]|^2 \tag{17}$$

where N_t is the number of training symbols per burst (see Fig. 3) and

$$\mathbf{r}(i) = [\mathbf{r}_1(2i)^T, \mathbf{r}_1^*(2i+1)^T, \ldots, r_A(2i)^T, r_A^*(2i+1)^T]^T$$

The second vector for each antenna element is conjugated to account for the conjugate introduced by the space-time code.

If $N_t/2 \ge 2AN$, then the solution to Eqs. (16) and (17) is well known:

$$\mathbf{W}_1^{\mathrm{LS}} = \overbrace{\left(\frac{2}{N_t} \sum_{i=0}^{N_t/2-1} \mathbf{r}(i)\mathbf{r}(i)^H\right)^{-1}}^{\mathbf{R}_{N_t}} \overbrace{\left(\frac{2}{N_t} \sum_{i=0}^{N_t/2-1} \mathbf{r}(i) b_1^*[2i]\right)}^{\mathbf{P}_{N_t^1}} \tag{18}$$

$$\mathbf{W}_2^{\mathrm{LS}} = \left(\frac{2}{N_t} \sum_{i=0}^{N_t/2-1} \mathbf{r}(i)\mathbf{r}(i)^H\right)^{-1} \underbrace{\left(\frac{2}{N_t} \sum_{i=0}^{N_t/2-1} \mathbf{r}(i) b_1[2i+1]\right)}_{\mathbf{P}_{N_t^2}} \tag{19}$$

In cases where this is not true, the time-averaged autocorrelation matrix, $\mathbf{R}_{N_t}$, becomes singular, so we use diagonal loading for regularization

$$\mathbf{W}_{1,2}^{\mathrm{LS}} = (\mathbf{R}_{N_t} + \delta \mathbf{I}_{2AN})^{-1} \mathbf{P}_{N_t}^{1,2} \tag{20}$$

where δ is some small positive constant (σ^2 ideally). As $N_t \to \infty$, $\mathbf{R}_{N_t} \to \mathbf{R}$, where $\mathbf{R}$ is the autocorrelation matrix

$$\mathbf{R} = \mathbf{G}\mathbf{B}\mathbf{G}^H + \sigma^2 \mathbf{I}_{2AN} \tag{21}$$

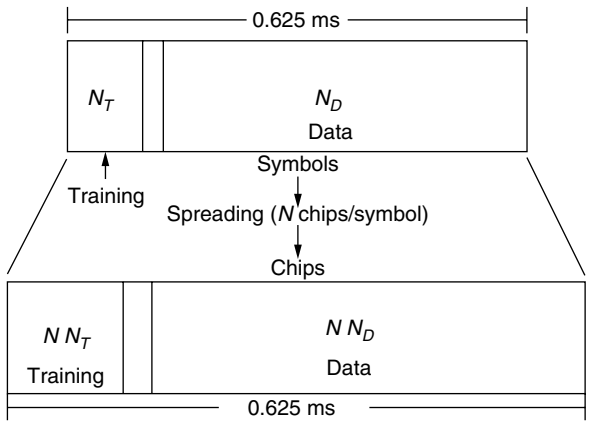


Figure 3. Burst structure.

and

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_1^* \\ \vdots & \vdots \\ \mathbf{G}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_A^* \end{bmatrix}, \quad \mathbf{B} = E\left(\begin{bmatrix} \mathbf{b}_{2i} \\ \mathbf{b}_{2i+1}^* \end{bmatrix} [\mathbf{b}_{2i}^H \quad \mathbf{b}_{2i+1}^T] \right)$$

Assuming that $2AN \geq 4KL$, and that $\mathbf{G}$ has full rank, it can be shown that the dimensionality of the signal space, k, (rank($\mathbf{GBG}^H$)) is given by

$$\kappa = \begin{cases} 2 \cdot K & \text{if } L = 1 \text{ (flat-fading)} \\ 2 \cdot K \cdot (L+1) & \text{if } L > 1 \end{cases} \tag{22}$$

Thus, in the noiseless case ($\sigma^2 = 0$) the minimum number of training symbols required to completely suppress the interference (multiple access and intersymbol) is simply $2 \cdot \kappa$. However, in the presence of noise and in a time-varying propagation environment, longer training sequences are needed. For a given length of training, the performance degrades with decreasing SNR and is highly dependent on the time-varying channel characteristics. This has serious implications on both user detectability and system spectral efficiency, which is measured as a percentage of training data in a burst. To alleviate this problem semiblind algorithms that utilize *blind* signal estimation principles to effectively increase the equivalent length of the training sequence have been proposed [1,14,16]. The semiblind algorithm described next can be effectively incorporated in a spatiotemporal framework and provide significant performance gains as it will be demonstrated by means of computer simulations.

4.2.2. Semiblind Processing. The objective is to find the weight vectors that minimize the cost functions

$$\mathbf{W}_1 = \arg\min_{\mathbf{W}} \frac{2}{N_b} \sum_{i=0}^{N_b/2-1} |\mathbf{W}^H \mathbf{r}(i) - b_1[2i]|^2 \tag{23}$$

$$\mathbf{W}_2 = \arg\min_{\mathbf{W}} \frac{2}{N_b} \sum_{i=0}^{N_b/2-1} |\mathbf{W}^H \mathbf{r}(i) - b_1^*[2i+1]|^2 \tag{24}$$

given the following information about the source symbols:

- $b_1[i]$ is known for $i = 0, \ldots, N_t - 1$ (training symbols),
- $|b_1[i]| = 1, i = 0, \ldots, N_b - 1$.

Our strategy is to use an alternating projection approach such as that discussed by van der Veen [2]. For the ith iteration of the weight vector, let

$$\tilde{b}_1^{1(i)} \triangleq \left\{ \mathbf{b}_{N_t}^1, \frac{\mathbf{W}_1^{H(i)} \mathbf{r}(N_t/2)}{|\mathbf{W}_1^{H(i)} \mathbf{r}(N_t/2)|}, \ldots, \frac{\mathbf{W}_1^{H(i)} \mathbf{r}(N_b/2-1)}{|\mathbf{W}_1^{H(i)} \mathbf{r}(N_b/2-1)|} \right\}$$

$$\tilde{b}_1^{2(i)} \triangleq \left\{ \mathbf{b}_{N_t}^2, \frac{\mathbf{W}_2^{H(i)} \mathbf{r}(N_t/2)}{|\mathbf{W}_2^{H(i)} \mathbf{r}(N_t/2)|}, \ldots, \frac{\mathbf{W}_2^{H(i)} \mathbf{r}(N_b/2-1)}{|\mathbf{W}_2^{H(i)} \mathbf{r}(N_b/2-1)|} \right\}$$

where $\mathbf{b}_{N_t}^1 = [b_1[0], b_1[2], \ldots, b_1[N_t - 2]]$ and $\mathbf{b}_{N_t}^2 = [b_1[1], b_1[3], \ldots, b_1[N_t - 1]]$) are a sequence that contains the known N_t source symbols, and $(N_b - N_t)/2$ estimated source symbols (via the Godard-type nonlinearity $\frac{\mathbf{W}_1^H \mathbf{r}(k)}{|\mathbf{W}_1^H \mathbf{r}(k)|}$). The function $\frac{\mathbf{W}_1^H \mathbf{r}(k)}{|\mathbf{W}_1^H \mathbf{r}(k)|}$ is chosen based on a priori knowledge of the source constellation (it returns a complex number with unit magnitude). If the initialization is accurate, then it will return the correct phase of the source symbol. We refer to the corresponding $(N_b - N_t)/2$ symbols in $\tilde{b}_1^{1(i)}(\tilde{b}_1^{2(i)})$ as *pseudotraining symbols*.

The sequence $\tilde{b}_1^{1(i)}(\tilde{b}_1^{2(i)})$ is then used to estimate a new cross-correlation vector, $\tilde{\mathbf{P}}_{N_b}^{1(i)}(\tilde{\mathbf{P}}_{N_b}^{2(i)})$, and $\mathbf{W}_1^{(i)}(\mathbf{W}_2^{(i)})$ is updated using this new vector and signal space parameters computed from eigen-decomposition of the time-averaged autocorrelation matrix $\mathbf{R}_{N_b}$:

$$\tilde{\mathbf{P}}_{N_b}^{1(i)} = \frac{2}{N_b} \sum_{k=0}^{N_b/2-1} \tilde{b}_1^{*1(i)}(k) \mathbf{r}(k) \tag{25}$$

$$\tilde{\mathbf{P}}_{N_b}^{2(i)} = \frac{2}{N_b} \sum_{k=0}^{N_b/2-1} \tilde{b}_1^{2(i)}(k) \mathbf{r}(k) \tag{26}$$

$$\mathbf{W}_1^{(i+1)} = \hat{\mathbf{U}}_S \hat{\mathbf{\Lambda}}_S^{-1} \hat{\mathbf{U}}_S^H \tilde{\mathbf{P}}_{N_b}^{1(i)} \tag{27}$$

$$\mathbf{W}_2^{(i+1)} = \hat{\mathbf{U}}_S \hat{\mathbf{\Lambda}}_S^{-1} \hat{\mathbf{U}}_S^H \tilde{\mathbf{P}}_{N_b}^{2(i)} \tag{28}$$

where $\mathbf{U}_S$ are the signal space eigenvectors and $\mathbf{\Lambda}_S$ are the signal space eigenvalues. The complete algorithm is given in Table 1. The algorithm is called the (*semiblind constant-modulus algorithm*) (SBCMA).

5. SIMULATION EXAMPLES

This section compares the MSE and probability of error performance of the SBCMA algorithm against LS. Channel coefficients, $\{\beta_{ia}^k[l]\}_{l=0}^{L_k-1}$, are randomly generated for each user $(k = 1, \ldots, K)$ and for each transmitter/receiver pair $(i = 1, 2, a = 1, \ldots, A)$ from a complex Gaussian distribution of unit variance and zero mean. The channel coefficients do not change during the transmission of N_b data symbols. SNR is defined as $10 \log_{10}(1/\sigma^2)$. In all cases

Table 1. SBCMA

In: $\hat{\mathbf{U}}_s$, $\hat{\mathbf{\Lambda}}_s$, $\{\mathbf{r}(k)\}_{k=0}^{N_b/2-1}$, $\{b_1[k]\}_{k=0}^{N_t-1}$, ζ *Out*: $\mathbf{W}_{1,2}$

$\mathbf{H} \triangleq \hat{\mathbf{U}}_s \hat{\mathbf{\Lambda}}_s^{-1} \hat{\mathbf{U}}_s^H$

$\mathbf{X} \triangleq [\mathbf{r}(0), \ldots, \mathbf{r}(N_b/2-1)]$

Choose $\mathbf{W}_1^{(0)} = \mathbf{W}_1^{LS}, \mathbf{W}_2^{(0)} = \mathbf{W}_2^{LS}$

for $i = 0, 1, \ldots$

a. $Y_1 = (\mathbf{W}_1^{(i)H} \mathbf{X}) / |\mathbf{W}_1^{(i)H} \mathbf{X}|$
$Y_2 = (\mathbf{W}_2^{(i)H} \mathbf{X}) / |\mathbf{W}_2^{(i)H} \mathbf{X}|$
$\tilde{\mathbf{b}}_1 = [\mathbf{b}_{Nt}^1, Y_1[N_t/2], \ldots, Y_1[N_b/2-1]]$
$\tilde{\mathbf{b}}_2 = [\mathbf{b}_{Nt}^2, Y_2[N_t/2], \ldots, Y_2[N_b/2-1]]$

b. $\tilde{\mathbf{P}}_{N_b}^{1(i)} = (\mathbf{X} \cdot \tilde{\mathbf{b}}_1^H)/(N_b/2)$
$\tilde{\mathbf{P}}_{N_b}^{2(i)} = (\mathbf{X} \cdot \tilde{\mathbf{b}}_2^T)/(N_b/2)$

c. $\mathbf{W}_1^{(i+1)} = \mathbf{H} \cdot \tilde{\mathbf{P}}_{N_b}^{1(i)}, \mathbf{W}_2^{(i+1)} = \mathbf{H} \cdot \tilde{\mathbf{P}}_{N_b}^{2(i)}$

until $\|\mathbf{W}_1^{(i+1)} - \mathbf{W}_1^{(i)}\|^2 / \|\mathbf{W}_1^{(i)}\|^2 < \zeta$

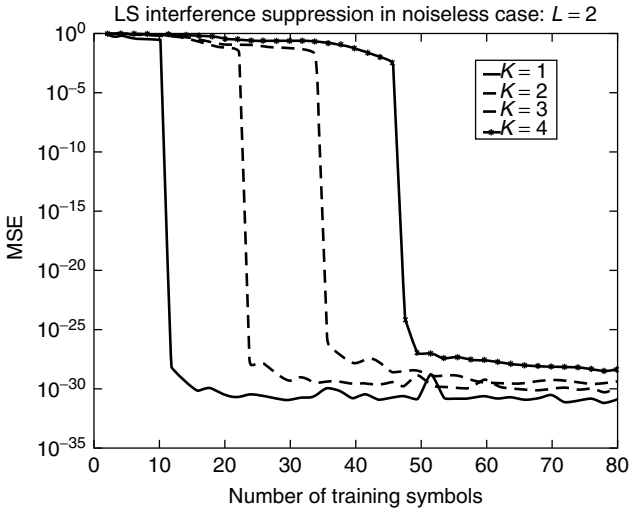


Figure 4. MSE comparison for noiseless case ($\sigma^2 = 0$): $L = 2$; $K = 1, 2, 3, 4$; $A = 2$; $N = 7$.

$\zeta = 10^{-5}$, and $N = 7$. The signal space rank was estimated using the MDL criterion.

Figure 4 demonstrates the zero-forcing ability of the LS filter when the number of training symbols is greater than the signal space dimensionality [as predicted by Eq. (22)]. In this case $L = 2$ and K is increased from 1 to 4. It is seen that once $N_t \geq 2 \cdot \kappa$ there is complete suppression of multiple access and intersymbol interference.

Figure 5 compares the MSE performance of the SBCMA with LS under the conditions that $K = 3, L = 2$ and $A = 2$ or $A = 4$. It is observed that the SBCMA uses significantly fewer training symbols than the LS and that the SBCMA is able to converge using fewer training symbols than required by LS in the noiseless case (36 in this case). When $A = 4$, the MSE is significantly lower but the required number of training symbols for convergence is roughly the same. This can be understood from the fact that increasing A does not increase the column space of $\mathbf{G}$.

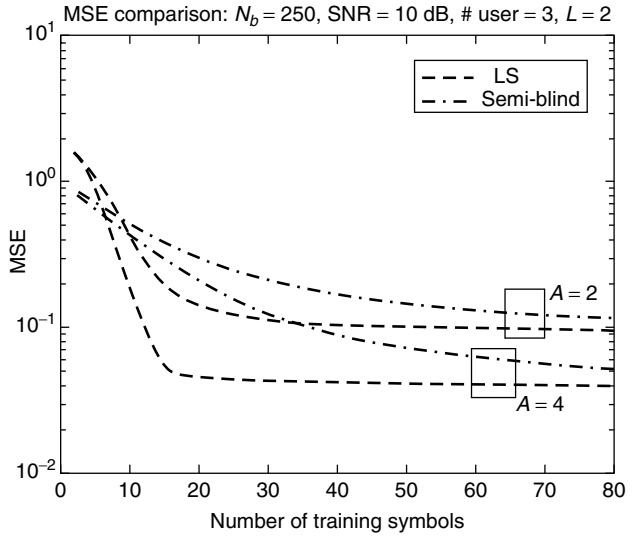


Figure 5. MSE comparison for $A = 2$ and $A = 4$: SNR = 10 dB, $L = 2, K = 3, N = 7$.

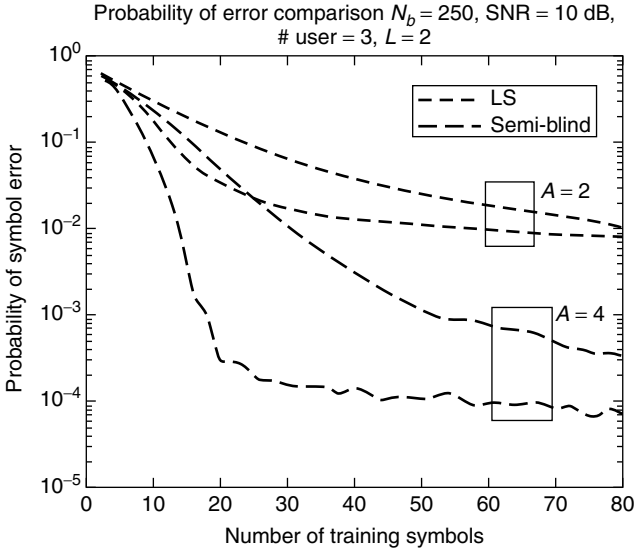


Figure 6. Probability of error comparison for $A = 2$ and $A = 4$: SNR = 10 dB, $L = 2, K = 3, N = 7$.

Finally, Fig. 6 compares the probability of error performance of the SBCMA with LS under the same conditions as above. Again, we observe the superior convergence speed of the SBCMA over LS.

6. CONCLUSION

Some theoretical results and a practical approach on spatio-temporal signal processing for wireless communications have been presented. The investigation for efficient utilization of multiple antennas at both ends of a wireless transceiver combined has raised the expectations for significant increases of the system capacity but at the same time created new challenges in developing the required signal processing technologies. The research and developments in this area are ongoing, and the corresponding literature is rapidly increasing as we now try to define the systems that will succeed the third generation of wireless mobile systems.

BIOGRAPHIES

Dimitrios Hatzinakos, Ph.D., is a Professor at the Department of Electrical and Computer Engineering, University of Toronto. He has also served as Chair of the Communications Group of the Department since July 1, 1999. His research interests are in the area of digital signal processing with applications to wireless communications, image processing, and multimedia. He is author or co-author of more than 120 papers in technical journals and conference proceedings, and he has contributed to five books in his areas of interest.

He served as an Associate Editor for the *IEEE Transactions on Signal Processing* from July 1998 until 2001; Guest Editor for the special issue of *Signal Processing*, Elsevier, on *Signal Processing Technologies*

for Short Burst Wireless Communications, which appeared in October 2000. He was a member of the Conference board of the IEEE Statistical Signal and Array Processing Technical Committee (SSAP) from 1992 to 1995 and was Technical Program co-Chair of the 5th Workshop on Higher-Order Statistics in July 1997. He is a senior member of the IEEE and member of EURASIP, the Professional Engineers of Ontario (PEO), and the Technical Chamber of Greece.

Ryan A. Pacheco is currently working towards a Ph.D. in Electrical Engineering at the University of Toronto. His research is concerned with blind/semiblind interference suppression, tracking, and signal acquisition for wireless communications.

BIBLIOGRAPHY

1. R. A. Pacheco and D. Hatzinakos, Semi-blind strategies for interference suppression in ds-cdma systems, *IEEE Trans. Signal Process.* (in press).
2. A.-J. van der Veen, Algebraic constant modulus algorithms, in *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 2, Prentice-Hall PTR, Englewood Cliffs, NJ, 2001, pp. 89–130.
3. G. Ganesan and P. Stoica, Space-time diversity, in G. B. Giannakis et al., eds., *Signal Processing Advances in Wireless & Mobile Communications*, Vol. 2, Prentice-Hall PTR, Englewood Cliffs, NJ, 2001, pp. 59–87.
4. V. Tarokh, H. Jafarkhani, and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Trans. Inform. Theory* **45**(5): 1456–1467 (July 1999).
5. A. F. Naguib, N. Seshadri, and A. R. Calderbank, Increasing data rate over wireless channels, *IEEE Signal Process. Mag.* 76–92 (May 2000).
6. S. Haykin, *Adaptive Filter Theory* Prentice-Hall, Englewood Cliffs, NJ, 1996.
7. S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Select. Areas Commun.* **16**(8): 1451–1458 (Oct. 1998).
8. G. J. Foschini, Layered space-time architecture for wireless communication in a fading environment when using multiple antennas, *Bell Labs Tech. J.* **1**(2): 41–59 (1996).
9. P. W. Wolniansky, G. J. Foschini, G. B. Golden, and R. A. Valenzuela, V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel, *IEEE Proc. ISSSE* 295–300 (1998).
10. M. Adachi, F. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communication systems, *IEEE Commun. Mag.* **36**: 56–69 (Sept. 1998).
11. A. Gorokhov and P. Loubaton, Semi-blind second order identification of convolutive channels, *Proc. ICASSP*, 1997, pp. 3905–3908.
12. D. Hatzinakos, Blind deconvolution channel identification and equalization, *Control Dynam. Syst.* **68**: 279–331 (1995).
13. M. Honig, U. Madhow, and S. Verdu, Blind multiuser detection, *IEEE Trans. Inform. Theory* **41**: 944–960 (July 1995).
14. A. M. Kuzminskiy, L. Féty, P. Forster, and S. Mayrargue, Regularized semi-blind estimation of spatio-temporal filter coefficients for mobile radio communications, *Proc. GRETSI*, 1997, pp. 127–130.
15. A. M. Kuzminskiy and D. Hatzinakos, Semi-blind estimation of spatio-temporal filter coefficients based on a training-like approach, *IEEE Signal Process. Lett.* **5**: 231–233 (Sept. 1998).
16. A. M. Kuzminskiy and D. Hatzinakos, Multistage semi-blind spatio-temporal processing for short burst multiuser SDMA systems, *Proc. 32nd Asilomar Conf. Signals, Systems, and Computers*, Oct. 1998, pp. 1887–1891.
17. A. Paulraj and C. Papadias, Space-time processing for wireless communications, *IEEE Signal Process. Mag.* **14**: 49–83 (Nov. 1997).
18. J. G. Proakis, *Digital Communications*, 3rd. ed., McGraw-Hill, New York, 1995.
19. A. van der Veen, S. Talwar, and A. Paulraj, A subspace approach to blind space-time signal processing for wireless communication systems, *IEEE Trans. Signal Process.* **45**: 173–190 (Jan. 1997).
20. S. Verdu, Minimum probability of error for asynchronous gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **32**: 85–96 (Jan. 1986).
21. X. Wang and H. V. Poor, Blind equalization and multiuser detection in dispersive CDMA channels, *IEEE Trans. Commun.* **46**: 91–103 (Jan. 1998).

SPEECH CODING: FUNDAMENTALS AND APPLICATIONS

MARK HASEGAWA-JOHNSON
University of Illinois at Urbana–Champaign
Urbana, Illinois

ABEER ALWAN
University of California at Los Angeles
Los Angeles, California

1. INTRODUCTION

Speech coding is the process of obtaining a compact representation of voice signals for efficient transmission over band-limited wired and wireless channels and/or storage. Today, speech coders have become essential components in telecommunications and in the multimedia infrastructure. Commercial systems that rely on efficient speech coding include cellular communication, voice over internet protocol (VOIP), videoconferencing, electronic toys, archiving, and digital simultaneous voice and data (DSVD), as well as numerous PC-based games and multimedia applications.

Speech coding is the art of creating a minimally redundant representation of the speech signal that can be efficiently transmitted or stored in digital media, and decoding the signal with the best possible perceptual quality. Like any other continuous-time signal, speech may be represented digitally through the processes of sampling and quantization; speech is typically quantized using either 16-bit uniform or 8-bit companded quantization. Like many other signals, however, a sampled speech

signal contains a great deal of information that is either redundant (nonzero mutual information between successive samples in the signal) or perceptually irrelevant (information that is not perceived by human listeners). Most telecommunications coders are *lossy*, meaning that the synthesized speech is perceptually similar to the original but may be physically dissimilar.

A speech coder converts a digitized speech signal into a coded representation, which is usually transmitted in frames. A speech decoder receives coded frames and synthesizes reconstructed speech. Standards typically dictate the input–output relationships of both coder and decoder. The input–output relationship is specified using a reference implementation, but novel implementations are allowed, provided that input–output equivalence is maintained. Speech coders differ primarily in bit rate (measured in bits per sample or bits per second), complexity (measured in operations per second), delay (measured in milliseconds between recording and playback), and perceptual quality of the synthesized speech. *Narrowband* (NB) coding refers to coding of speech signals whose bandwidth is less than 4 kHz (8 kHz sampling rate), while *wideband* (WB) coding refers to coding of 7-kHz-bandwidth signals (14–16 kHz sampling rate). NB coding is more common than WB coding mainly because of the narrowband nature of the wireline telephone channel (300–3600 Hz). More recently, however, there has been an increased effort in wideband speech coding because of several applications such as videoconferencing.

There are different types of speech coders. Table 1 summarizes the bit rates, algorithmic complexity, and standardized applications of the four general classes of coders described in this article; Table 2 lists a selection of specific speech coding standards. Waveform coders attempt to code the exact shape of the speech signal waveform, without considering the nature of human speech production and speech perception. These coders are high-bit-rate coders (typically above 16 kbps). Linear prediction coders (LPCs), on the other hand, assume that the speech signal is the output of a linear time-invariant (LTI) model of speech production. The transfer function of that model is assumed to be all-pole (autoregressive model). The excitation function is a quasiperiodic signal constructed from discrete pulses (1–8 per pitch period), pseudorandom noise, or some combination of the two. If the excitation is generated only at the receiver, based on a transmitted pitch period and voicing information, then the system is designated as an LPC vocoder. LPC vocoders that provide extra information about the spectral shape of the excitation have been adopted as coder standards between 2.0 and 4.8 kbps. LPC-based analysis-by-synthesis coders (LPC-AS), on the other hand, choose an excitation function by explicitly testing a large set of candidate excitations and choosing the best. LPC-AS coders are used in most standards between 4.8 and 16 kbps. Subband coders are frequency-domain coders that attempt to parameterize the speech signal in terms of spectral properties in different frequency bands. These coders are less widely used than LPC-based coders but have the advantage of being scalable and do not model the incoming signal as speech. Subband coders are widely used for high-quality audio coding.

This article is organized as follows. Sections 2, 3, 4 and 5 present the basic principles behind waveform coders, subband coders, LPC-based analysis-by-synthesis coders, and LPC-based vocoders, respectively. Section 6 describes the different quality metrics that are used to evaluate speech coders, while Section 7 discusses a variety of issues that arise when a coder is implemented in a communications network, including voice over IP, multirate coding, and channel coding. Section 8 presents an overview of standardization activities involving speech coding, and we conclude in Section 9 with some final remarks.

2. WAVEFORM CODING

Waveform coders attempt to code the exact shape of the speech signal waveform, without considering in detail the nature of human speech production and speech perception. Waveform coders are most useful in applications that require the successful coding of both speech and nonspeech signals. In the public switched telephone network (PSTN), for example, successful transmission of modem and fax signaling tones, and switching signals is nearly as important as the successful transmission of speech. The most commonly used waveform coding algorithms are uniform 16-bit PCM, companded 8-bit PCM [48], and ADPCM [46].

2.1. Pulse Code Modulation (PCM)

Pulse code modulation (PCM) is the name given to memoryless coding algorithms that quantize each sample of $s(n)$ using the same reconstruction levels $\hat{s}_k$, $k = 0, \ldots, m, \ldots, K$, regardless of the values of previous samples. The reconstructed signal $\hat{s}(n)$ is given by

$$\hat{s}(n) = \hat{s}_m \quad \text{for} \quad s(n) \text{ s.t. } (s(n) - \hat{s}_m)^2 = \min_{k=0,\ldots,K} (s(n) - \hat{s}_k)^2 \tag{1}$$

Many speech and audio applications use an odd number of reconstruction levels, so that background noise signals with a very low level can be quantized exactly to

Table 1. Characteristics of Standardized Speech Coding Algorithms in Each of Four Broad Categories

Speech Coder Class	Rates (kbps)	Complexity	Standardized Applications	Section
Waveform coders	16–64	Low	Landline telephone	2
Subband coders	12–256	Medium	Teleconferencing, audio	3
LPC-AS	4.8–16	High	Digital cellular	4
LPC vocoder	2.0–4.8	High	Satellite telephony, military	5

Table 2. A Representative Sample of Speech Coding Standards

Application	Rate (kbps)	BW (kHz)	Standards Organization	Standard Number	Algorithm	Year
Landline telephone	64	3.4	ITU	G.711	μ-law or A-law PCM	1988
	16–40	3.4	ITU	G.726	ADPCM	1990
	16–40	3.4	ITU	G.727	ADPCM	1990
Tele conferencing	48–64	7	ITU	G.722	Split-band ADPCM	1988
	16	3.4	ITU	G.728	Low-delay CELP	1992
Digital	13	3.4	ETSI	Full-rate	RPE-LTP	1992
cellular	12.2	3.4	ETSI	EFR	ACELP	1997
	7.9	3.4	TIA	IS-54	VSELP	1990
	6.5	3.4	ETSI	Half-rate	VSELP	1995
	8.0	3.4	ITU	G.729	ACELP	1996
	4.75–12.2	3.4	ETSI	AMR	ACELP	1998
	1–8	3.4	CDMA-TIA	IS-96	QCELP	1993
Multimedia	5.3–6.3	3.4	ITU	G.723.1	MPLPC, CELP	1996
	2.0–18.2	3.4–7.5	ISO	MPEG-4	HVXC, CELP	1998
Satellite telephony	4.15	3.4	INMARSAT	M	IMBE	1991
	3.6	3.4	INMARSAT	Mini-M	AMBE	1995
Secure communications	2.4	3.4	DDVPC	FS1015	LPC-10e	1984
	2.4	3.4	DDVPC	MELP	MELP	1996
	4.8	3.4	DDVPC	FS1016	CELP	1989
	16–32	3.4	DDVPC	CVSD	CVSD	

$\hat{s}_{K/2} = 0$. One important exception is the A-law companded PCM standard [48], which uses an even number of reconstruction levels.

2.1.1. Uniform PCM. Uniform PCM is the name given to quantization algorithms in which the reconstruction levels are uniformly distributed between $S_{\max}$ and $S_{\min}$. The advantage of uniform PCM is that the quantization error power is independent of signal power; high-power signals are quantized with the same resolution as low-power signals. Invariant error power is considered desirable in many digital audio applications, so 16-bit uniform PCM is a standard coding scheme in digital audio.

The error power and SNR of a uniform PCM coder vary with bit rate in a simple fashion. Suppose that a signal is quantized using B bits per sample. If zero is a reconstruction level, then the quantization step size Δ is

$$\Delta = \frac{S_{\max} - S_{\min}}{2^B - 1} \tag{2}$$

Assuming that quantization errors are uniformly distributed between $\Delta/2$ and $-\Delta/2$, the quantization error power is

$$\begin{aligned} 10\log_{10} E[e^2(n)] &= 10\log_{10}\frac{\Delta^2}{12} \\ &\approx \text{constant} + 20\log_{10}(S_{\max} - S_{\min}) - 6B \end{aligned} \tag{3}$$

2.1.2. Companded PCM. Companded PCM is the name given to coders in which the reconstruction levels $\hat{s}_k$ are not uniformly distributed. Such coders may be modeled using a compressive nonlinearity, followed by uniform PCM, followed by an expansive nonlinearity:

$$\begin{aligned} s(n) \rightarrow \boxed{\text{compress}} \rightarrow t(n) \rightarrow \boxed{\text{uniform PCM}} \\ \rightarrow \hat{t}(n) \rightarrow \boxed{\text{expand}} \rightarrow \hat{s}(n) \end{aligned} \tag{4}$$

It can be shown that, if small values of $s(n)$ are more likely than large values, expected error power is minimized by a companding function that results in a higher density of reconstruction levels $\hat{x}_k$ at low signal levels than at high signal levels [78]. A typical example is the μ-law companding function [48] (Fig. 1), which is given by

$$t(n) = S_{\max}\frac{\log(1 + \mu|s(n)/S_{\max}|)}{\log(1+\mu)}\text{sign}(s(n)) \tag{5}$$

where μ is typically between 0 and 256 and determines the amount of nonlinear compression applied.

2.2. Differential PCM (DPCM)

Successive speech samples are highly correlated. The long-term average spectrum of voiced speech is reasonably

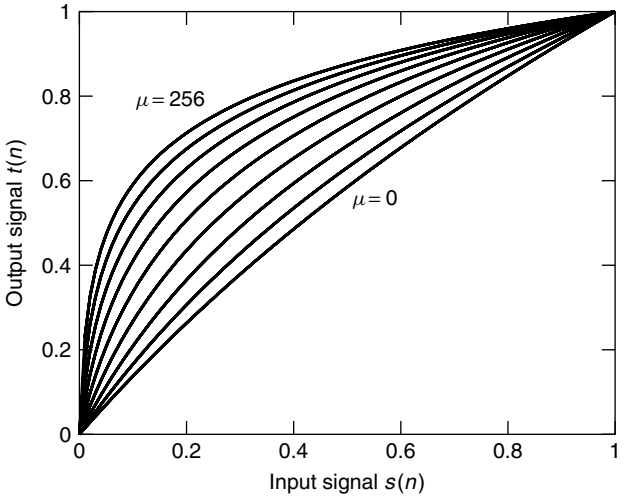


Figure 1. μ-law companding function, $\mu = 0, 1, 2, 4, 8, \ldots, 256$.

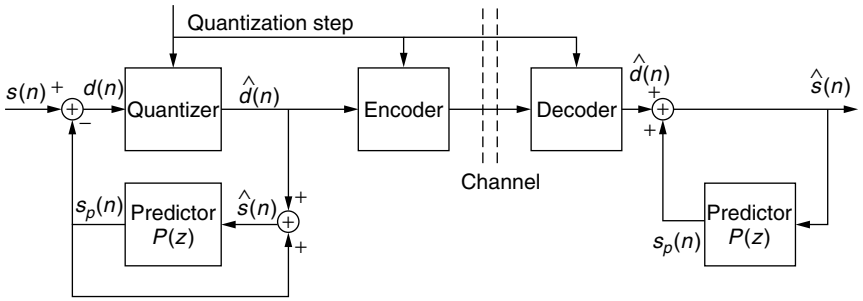


Figure 2. Schematic of a DPCM coder.

well approximated by the function $S(f) = 1/f$ above about 500 Hz; the first-order intersample correlation coefficient is approximately 0.9. In differential PCM, each sample $s(n)$ is compared to a prediction $s_p(n)$, and the difference is called the prediction residual $d(n)$ (Fig. 2). $d(n)$ has a smaller dynamic range than $s(n)$, so for a given error power, fewer bits are required to quantize $d(n)$.

Accurate quantization of $d(n)$ is useless unless it leads to accurate quantization of $s(n)$. In order to avoid amplifying the error, DPCM coders use a technique copied by many later speech coders; the encoder includes an embedded decoder, so that the reconstructed signal $\hat{s}(n)$ is known at the encoder. By using $\hat{s}(n)$ to create $s_p(n)$, DPCM coders avoid amplifying the quantization error:

$$d(n) = s(n) - s_p(n) \tag{6}$$

$$\hat{s}(n) = \hat{d}(n) + s_p(n) \tag{7}$$

$$e(n) = s(n) - \hat{s}(n) = d(n) - \hat{d}(n) \tag{8}$$

Two existing standards are based on DPCM. In the first type of coder, continuously varying slope delta modulation (CVSD), the input speech signal is upsampled to either 16 or 32 kHz. Values of the upsampled signal are predicted using a one-tap predictor, and the difference signal is quantized at one bit per sample, with an adaptively varying Δ. CVSD performs badly in quiet environments, but in extremely noisy environments (e.g., helicopter cockpit), CVSD performs better than any LPC-based algorithm, and for this reason it remains the U.S. Department of Defense recommendation for extremely noisy environments [64,96].

DPCM systems with adaptive prediction and quantization are referred to as adaptive differential PCM systems (ADPCM). A commonly used ADPCM standard is G.726, which can operate at 16, 24, 32, or 40 kbps (2–5 bits/sample) [45]. G.726 ADPCM is frequently used at 32 kbps in landline telephony. The predictor in G.726 consists of an adaptive second-order IIR predictor in series with an adaptive sixth-order FIR predictor. Filter coefficients are adapted using a computationally simplified gradient descent algorithm. The prediction residual is quantized using a semilogarithmic companded PCM quantizer at a rate of 2–5 bits per sample. The quantization step size adapts to the amplitude of previous samples of the quantized prediction error signal; the speed of adaptation is controlled by an estimate of the type of signal, with adaptation to speech signals being faster than adaptation to signaling tones.

3. SUBBAND CODING

In subband coding, an analysis filterbank is first used to filter the signal into a number of frequency bands and then bits are allocated to each band by a certain criterion. Because of the difficulty in obtaining high-quality speech at low bit rates using subband coding schemes, these techniques have been used mostly for wideband medium to high bit rate speech coders and for audio coding.

For example, G.722 is a standard in which ADPCM speech coding occurs within two subbands, and bit allocation is set to achieve 7-kHz audio coding at rates of 64 kbps or less.

In Refs. 12,13, and 30 subband coding is proposed as a flexible scheme for robust speech coding. A speech production model is not used, ensuring robustness to speech in the presence of background noise, and to nonspeech sources. High-quality compression can be achieved by incorporating masking properties of the human auditory system [54,93]. In particular, Tang et al. [93] present a scheme for robust, high-quality, scalable, and embedded speech coding. Figure 3 illustrates the basic structure of the coder. Dynamic bit allocation and prioritization and embedded quantization are used to optimize the perceptual quality of the embedded bitstream, resulting in little performance degradation relative to a nonembedded implementation. A subband spectral analysis technique was developed that substantially reduces the complexity of computing the perceptual model.

The encoded bitstream is embedded, allowing the coder output to be scalable from high quality at higher

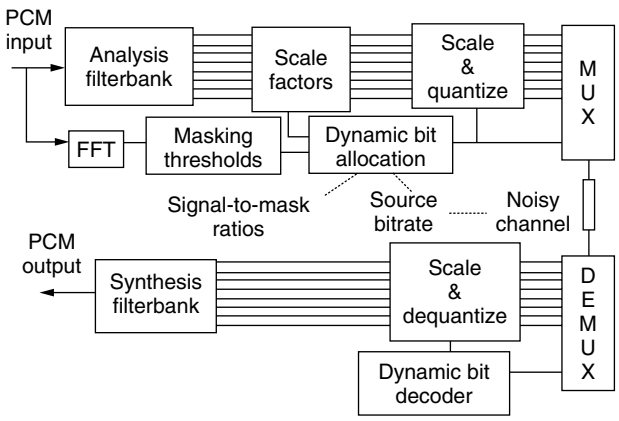


Figure 3. Structure of a perceptual subband speech coder [93].

bit rates, to lower quality at lower rates, supporting a wide range of service and resource utilization. The lower bit rate representation is obtained simply through truncation of the higher bit rate representation. Since source rate adaptation is performed through truncation of the encoded stream, interaction with the source coder is not required, making the coder ideally suited for rate adaptive communication systems.

Even though subband coding is not widely used for speech coding today, it is expected that new standards for wideband coding and rate-adaptive schemes will be based on subband coding or a hybrid technique that includes subband coding. This is because subband coders are more easily scalable in bit rate than standard CELP techniques, an issue which will become more critical for high-quality speech and audio transmission over wireless communication channels and the Internet, allowing the system to seamlessly adapt to changes in both the transmission environment and network congestion.

4. LPC-BASED ANALYSIS BY SYNTHESIS

An analysis-by-synthesis speech coder consists of the following components:

- A model of speech production that depends on certain parameters θ:
$$\hat{s}(n) = f(\theta) \tag{9}$$
- A list of K possible parameter sets for the model
$$\theta_1, \ldots, \theta_k, \ldots, \theta_K \tag{10}$$
- An error metric $|E_k|^2$ that compares the original speech signal $s(n)$ and the coded speech signal $\hat{s}(n)$. In LPC-AS coders, $|E_k|^2$ is typically a perceptually weighted mean-squared error measure.

A general analysis-by-synthesis coder finds the optimum set of parameters by synthesizing all of the K different speech waveforms $\hat{s}_k(n)$ corresponding to the K possible parameter sets θ_k, computing $|E_k|^2$ for each synthesized waveform, and then transmitting the index of the parameter set which minimizes $|E_k|^2$. Choosing a set of transmitted parameters by explicitly computing $\hat{s}_k(n)$ is called "closed loop" optimization, and may be contrasted with "open-loop" optimization, in which coder parameters are chosen on the basis of an analytical formula without explicit computation of $\hat{s}_k(n)$. Closed-loop optimization of all parameters is prohibitively expensive, so LPC-based analysis-by-synthesis coders typically adopt the following compromise. The gross spectral shape is modeled using an all-pole filter $1/A(z)$ whose parameters are estimated in open-loop fashion, while spectral fine structure is modeled using an excitation function $U(z)$ whose parameters are optimized in closed-loop fashion (Fig. 4).

4.1. The Basic LPC Model

In LPC-based coders, the speech signal $S(z)$ is viewed as the output of a linear time-invariant (LTI) system whose input is the excitation signal $U(z)$, and whose transfer function is represented by the following:

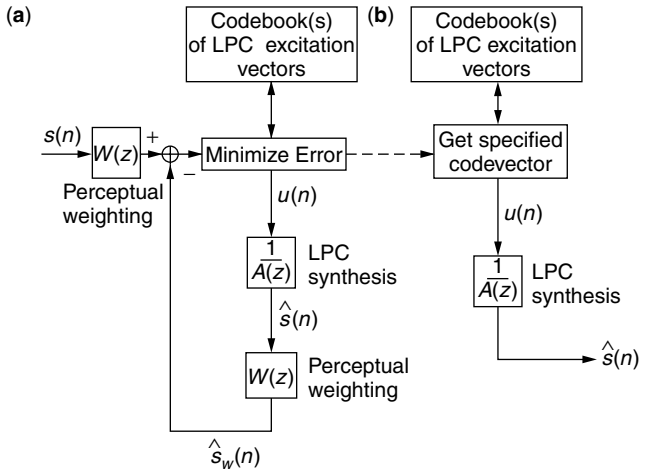


Figure 4. General structure of an LPC-AS coder (**a**) and decoder (**b**). LPC filter $A(z)$ and perceptual weighting filter $W(z)$ are chosen open-loop, then the excitation vector $u(n)$ is chosen in a closed-loop fashion in order to minimize the error metric $|E|^2$.

$$S(z) = \frac{U(z)}{A(z)} = \frac{U(z)}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{11}$$

Most of the zeros of $A(z)$ correspond to resonant frequencies of the vocal tract or *formant frequencies*. Formant frequencies depend on the geometry of the vocal tract; this is why men and women, who have different vocal-tract shapes and lengths, have different formant frequencies for the same sounds.

The number of LPC coefficients (p) depends on the signal bandwidth. Since each pair of complex-conjugate poles represents one formant frequency and since there is, on average, one formant frequency per 1 kHz, p is typically equal to $2BW$ (in kHz) + (2 to 4). Thus, for a 4 kHz speech signal, a 10th–12th-order LPC model would be used.

This system is excited by a signal $u(n)$ that is uncorrelated with itself over lags of less than $p + 1$. If the underlying speech sound is unvoiced (the vocal folds do not vibrate), then $u(n)$ is uncorrelated with itself even at larger time lags, and may be modeled using a pseudo-random-noise signal. If the underlying speech is voiced (the vocal folds vibrate), then $u(n)$ is quasiperiodic with a fundamental period called the "pitch period."

4.2. Pitch Prediction Filtering

In an LPC-AS coder, the LPC excitation is allowed to vary smoothly between fully voiced conditions (as in a vowel) and fully unvoiced conditions (as in /s/). Intermediate levels of voicing are often useful to model partially voiced phonemes such as /z/.

The partially voiced excitation in an LPC-AS coder is constructed by passing an uncorrelated noise signal $c(n)$ through a pitch prediction filter [2,79]. A typical pitch prediction filter is

$$u(n) = gc(n) + bu(n - T_0) \tag{12}$$

where T_0 is the pitch period. If $c(n)$ is unit variance white noise, then according to Eq. (12) the spectrum of $u(n)$ is

$$|U(e^{j\omega})|^2 = \frac{g^2}{1 + b^2 - 2b\cos\omega T_0} \tag{13}$$

Figure 5 shows the normalized magnitude spectrum $(1-b)|U(e^{j\omega})|$ for several values of b between 0.25 and 1. As shown, the spectrum varies smoothly from a uniform spectrum, which is heard as unvoiced, to a harmonic spectrum that is heard as voiced, without the need for a binary voiced/unvoiced decision.

In LPC-AS coders, the noise signal $c(n)$ is chosen from a "stochastic codebook" of candidate noise signals. The stochastic codebook index, the pitch period, and the gains b and g are chosen in a closed-loop fashion in order to minimize a perceptually weighted error metric. The search for an optimum T_0 typically uses the same algorithm as the search for an optimum $c(n)$. For this reason, the list of excitation samples delayed by different candidate values of T_0 is typically called an "adaptive codebook" [87].

4.3. Perceptual Error Weighting

Not all types of distortion are equally audible. Many types of speech coders, including LPC-AS coders, use simple models of human perception in order to minimize the audibility of different types of distortion. In LPC-AS coding, two types of perceptual weighting are commonly used. The first type, perceptual weighting of the residual quantization error, is used during the LPC excitation search in order to choose the excitation vector with the least audible quantization error. The second type, adaptive postfiltering, is used to reduce the perceptual importance of any remaining quantization error.

4.3.1. Perceptual Weighting of the Residual Quantization Error. The excitation in an LPC-AS coder is chosen to minimize a perceptually weighted error metric. Usually, the error metric is a function of the time domain waveform error signal

$$e(n) = s(n) - \hat{s}(n) \tag{14}$$

Early LPC-AS coders minimized the mean-squared error

$$\sum_n e^2(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(e^{j\omega})|^2\, d\omega \tag{15}$$

It turns out that the MSE is minimized if the error spectrum, $E(e^{j\omega})$, is white — that is, if the error signal $e(n)$ is an uncorrelated random noise signal, as shown in Fig. 6.

Not all noises are equally audible. In particular, noise components near peaks of the speech spectrum are hidden by a "masking spectrum" $M(e^{j\omega})$, so that a shaped noise spectrum at lower SNR may be less audible than a white-noise spectrum at higher SNR (Fig. 7). The audibility of noise may be estimated using a noise-to-masker ratio $|E_w|^2$:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|E(e^{j\omega})|^2}{|M(e^{j\omega})|^2}\, d\omega \tag{16}$$

The masking spectrum $M(e^{j\omega})$ has peaks and valleys at the same frequencies as the speech spectrum, but the difference in amplitude between peaks and valleys is somewhat smaller than that of the speech spectrum. A variety of algorithms exist for estimating the masking spectrum, ranging from extremely simple to extremely complex [51]. One of the simplest model masking spectra that has the properties just described is as follows [2]:

$$M(z) = \frac{|A(z/\gamma_2)|}{|A(z/\gamma_1)|}, \quad 0 < \gamma_2 < \gamma_1 \le 1 \tag{17}$$

where $1/A(z)$ is an LPC model of the speech spectrum. The poles and zeros of $M(z)$ are at the same frequencies as the poles of $1/A(z)$, but have broader bandwidths. Since the zeros of $M(z)$ have broader bandwidth than its poles, $M(z)$ has peaks where $1/A(z)$ has peaks, but the difference between peak and valley amplitudes is somewhat reduced.

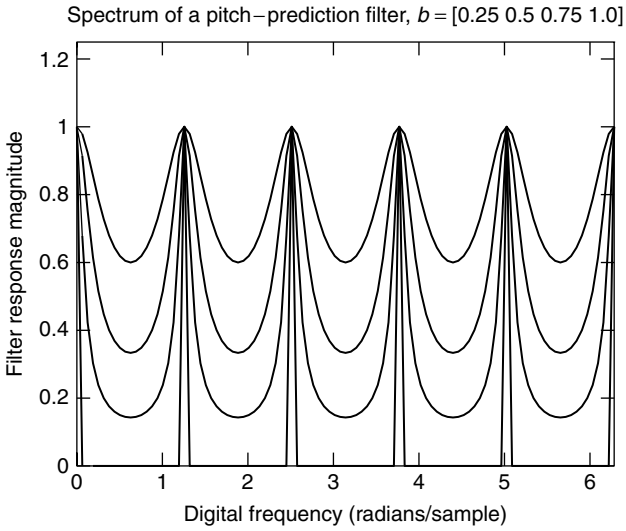


Figure 5. Normalized magnitude spectrum of the pitch prediction filter for several values of the prediction coefficient.

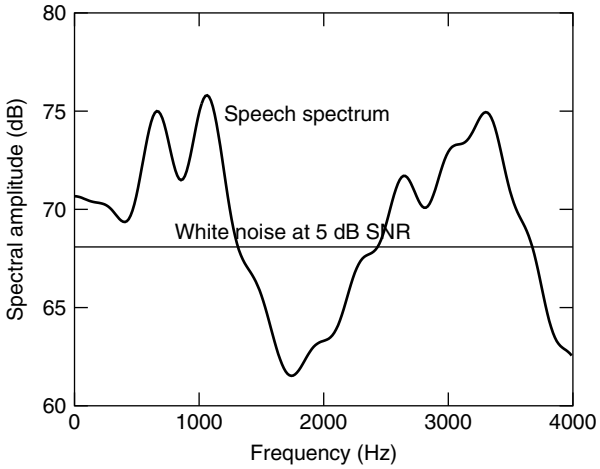


Figure 6. The minimum-energy quantization noise is usually characterized as white noise.

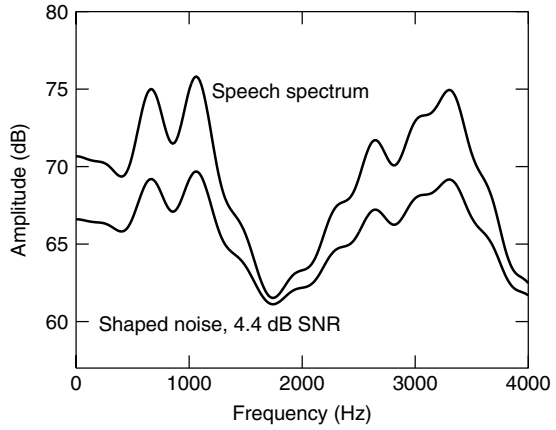


Figure 7. Shaped quantization noise may be less audible than white quantization noise, even at slightly lower SNR.

The noise-to-masker ratio may be efficiently computed by filtering the speech signal using a perceptual weighting filter $W(z) = 1/M(z)$. The perceptually weighted input speech signal is

$$S_w(z) = W(z)S(z) \tag{18}$$

Likewise, for any particular candidate excitation signal, the perceptually weighted output speech signal is

$$\hat{S}_w(z) = W(z)\hat{S}(z) \tag{19}$$

Given $s_w(n)$ and $\hat{s}_w(n)$, the noise-to-masker ratio may be computed as follows:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(e^{j\omega}) - \hat{S}_w(e^{j\omega})|^2\, d\omega = \sum_n (s_w(n) - \hat{s}_w(n))^2 \tag{20}$$

4.3.2. Adaptive Postfiltering. Despite the use of perceptually weighted error minimization, the synthesized speech coming from an LPC-AS coder may contain audible quantization noise. In order to minimize the perceptual effects of this noise, the last step in the decoding process is often a set of adaptive postfilters [11,80]. Adaptive postfiltering improves the perceptual quality of noisy speech by giving a small extra emphasis to features of the spectrum that are important for human-to-human communication, including the pitch periodicity (if any) and the peaks in the spectral envelope.

A pitch postfilter (or long-term predictive postfilter) enhances the periodicity of voiced speech by applying either an FIR or IIR comb filter to the output. The time delay and gain of the comb filter may be set equal to the transmitted pitch lag and gain, or they may be recalculated at the decoder using the reconstructed signal $\hat{s}(n)$. The pitch postfilter is applied only if the proposed comb filter gain is above a threshold; if the comb filter gain is below threshold, the speech is considered unvoiced, and no pitch postfilter is used. For improved perceptual quality, the LPC excitation signal may be interpolated to a higher sampling rate in order to allow the use of fractional pitch periods; for example, the postfilter in the ITU G.729 coder uses pitch periods quantized to $\frac{1}{8}$ sample.

A short-term predictive postfilter enhances peaks in the spectral envelope. The form of the short-term postfilter is similar to that of the masking function $M(z)$ introduced in the previous section; the filter has peaks at the same frequencies as $1/A(z)$, but the peak-to-valley ratio is less than that of $A(z)$.

Postfiltering may change the gain and the average spectral tilt of $\hat{s}(n)$. In order to correct these problems, systems that employ postfiltering may pass the final signal through a one-tap FIR preemphasis filter, and then modify its gain, prior to sending the reconstructed signal to a D/A converter.

4.4. Frame-Based Analysis

The characteristics of the LPC excitation signal $u(n)$ change quite rapidly. The energy of the signal may change from zero to nearly full amplitude within one millisecond at the release of a plosive sound, and a mistake of more than about 5 ms in the placement of such a sound is clearly audible. The LPC coefficients, on the other hand, change relatively slowly. In order to take advantage of the slow rate of change of LPC coefficients without sacrificing the quality of the coded residual, most LPC-AS coders encode speech using a frame–subframe structure, as depicted in Fig. 8. A frame of speech is approximately 20 ms in length, and is composed of typically three to four subframes. The LPC excitation is transmitted only once per subframe, while the LPC coefficients are transmitted only once per frame. The LPC coefficients are computed by analyzing a window of speech that is usually longer than the speech frame (typically 30–60 ms). In order to minimize the number of future samples required to compute LPC coefficients, many LPC-AS coders use an asymmetric window that may include several hundred milliseconds of past context, but that emphasizes the samples of the current frame [21,84].

The perceptually weighted original signal $s_w(n)$ and weighted reconstructed signal $\hat{s}_w(n)$ in a given subframe

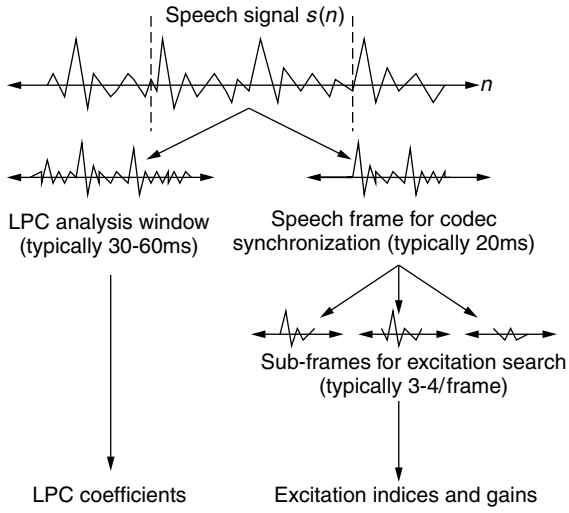


Figure 8. The frame/subframe structure of most LPC analysis by synthesis coders.

are often written as L-dimensional row vectors S and $\hat{S}$, where the dimension L is the length of a subframe:

$$S_w = [s_w(0), \ldots, s_w(L-1)], \quad \hat{S}_w = [\hat{s}_w(0), \ldots, \hat{s}_w(L-1)] \tag{21}$$

The core of an LPC-AS coder is the closed-loop search for an optimum coded excitation vector U, where U is typically composed of an "adaptive codebook" component representing the periodicity, and a "stochastic codebook" component representing the noiselike part of the excitation. In general, U may be represented as the weighted sum of several "shape vectors" X_m, $m = 1, \ldots, M$, which may be drawn from several codebooks, including possibly multiple adaptive codebooks and multiple stochastic codebooks:

$$U = GX, \quad G = [g_1, g_2, \ldots], \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \end{bmatrix} \tag{22}$$

The choice of shape vectors and the values of the gains g_m are jointly optimized in a closed-loop search, in order to minimize the perceptually weighted error metric $|S_w - \hat{S}_w|^2$.

The value of S_w may be computed prior to any codebook search by perceptually weighting the input speech vector. The value of $\hat{S}_w$ must be computed separately for each candidate excitation, by synthesizing the speech signal $\hat{s}(n)$, and then perceptually weighting to obtain $\hat{s}_w(n)$. These operations may be efficiently computed, as described below.

4.4.1. Zero State Response and Zero Input Response. Let the filter $H(z)$ be defined as the composition of the LPC synthesis filter and the perceptual weighting filter, thus $H(z) = W(z)/A(z)$. The computational complexity of the excitation parameter search may be greatly simplified if $\hat{S}_w$ is decomposed into the zero input response (ZIR) and zero state response (ZSR) of $H(z)$ [97]. Note that the weighted reconstructed speech signal is

$$\hat{S}_w = [\hat{s}_w(0), \ldots, \hat{s}_w(L-1)], \quad \hat{s}_w(n) = \sum_{i=0}^{\infty} h(i)u(n-i) \tag{23}$$

where $h(n)$ is the infinite-length impulse response of $H(z)$. Suppose that $\hat{s}_w(n)$ has already been computed for $n < 0$, and the coder is now in the process of choosing the optimal $u(n)$ for the subframe $0 \le n \le L-1$. The sum above can be divided into two parts: a part that depends on the current subframe input, and a part that does not:

$$\hat{S}_w = \hat{S}_{\text{ZIR}} + UH \tag{24}$$

where $\hat{S}_{\text{ZIR}}$ contains samples of the zero input response of $H(z)$, and the vector UH contains the zero state response. The zero input response is usually computed by implementing the recursive filter $H(z) = W(z)/A(z)$ as the sequence of two IIR filters, and allowing the two filters to run for L samples with zero input. The zero state response is usually computed as the matrix product UH, where

$$H = \begin{bmatrix} h(0) & h(1) & \ldots & h(L-1) \\ 0 & h(0) & \ldots & h(L-2) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & h(0) \end{bmatrix},$$
$$U = [u(0), \ldots, u(L-1)] \tag{25}$$

Given a candidate excitation vector U, the perceptually weighted error vector E may be defined as

$$E_w = S_w - \hat{S}_w = \tilde{S} - UH \tag{26}$$

where the target vector $\tilde{S}$ is

$$\tilde{S} = S_w - \hat{S}_{\text{ZIR}} \tag{27}$$

The target vector needs to be computed only once per subframe, prior to the codebook search. The objective of the codebook search, therefore, is to find an excitation vector U that minimizes $|\tilde{S} - UH|^2$.

4.4.2. Optimum Gain and Optimum Excitation. Recall that the excitation vector U is modeled as the weighted sum of a number of codevectors X_m, $m = 1, \ldots, M$. The perceptually weighted error is therefore:

$$|E|^2 = |\tilde{S} - GXH|^2 = \tilde{S}\tilde{S}' - 2GXH\tilde{S}' + GXH(GXH)' \tag{28}$$

where prime denotes transpose. Minimizing $|E|^2$ requires optimum choice of the shape vectors X and of the gains G. It turns out that the optimum gain for each excitation vector can be computed in closed form. Since the optimum gain can be computed in closed form, it need not be computed during the closed-loop search; instead, one can simply assume that each candidate excitation, if selected, would be scaled by its optimum gain. Assuming an optimum gain results in an extremely efficient criterion for choosing the optimum excitation vector [3].

Suppose we define the following additional bits of notation:

$$R_X = XH\tilde{S}', \quad \Sigma = XH(XH)' \tag{29}$$

Then the mean-squared error is

$$|E|^2 = \tilde{S}\tilde{S}' - 2GR_X + G\Sigma G' \tag{30}$$

For any given set of shape vectors X, G is chosen so that $|E|^2$ is minimized, which yields

$$G = R_X'\Sigma^{-1} \tag{31}$$

If we substitute the minimum MSE value of G into Eq. (30), we get

$$|E|^2 = \tilde{S}\tilde{S}' - R_X'\Sigma^{-1}R_X \tag{32}$$

Hence, in order to minimize the perceptually weighted MSE, we choose the shape vectors X in order to maximize the covariance-weighted sum of correlations:

$$X_{\text{opt}} = \arg\max(R'_X \Sigma^{-1} R_X) \tag{33}$$

When the shape matrix X contains more than one row, the matrix inversion in Eq. (33) is often computed using approximate algorithms [4]. In the VSELP coder [25], X is transformed using a modified Gram–Schmidt orthogonalization so that Σ has a diagonal structure, thus simplifying the computation of Eq. (33).

4.5. Types of LPC-AS Coder

4.5.1. Multipulse LPC (MPLPC). In the multipulse LPC algorithm [4,50], the shape vectors are impulses. U is typically formed as the weighted sum of 4–8 impulses per subframe.

The number of possible combinations of impulses grows exponentially in the number of impulses, so joint optimization of the positions of all impulses is usually impossible. Instead, most MPLPC coders optimize the pulse positions one at a time, using something like the following strategy. First, the weighted zero state response of $H(z)$ corresponding to each impulse location is computed. If C_k is an impulse located at $n = k$, the corresponding weighted zero state response is

$$C_k H = [0, \ldots, 0, h(0), h(1), \ldots, h(L-k-1)] \tag{34}$$

The location of the first impulse is chosen in order to optimally approximate the target vector $\tilde{S}_1 = \tilde{S}$, using the methods described in the previous section. After selecting the first impulse location k_1, the target vector is updated according to

$$\tilde{S}_m = \tilde{S}_{m-1} - C_{k_{m-1}} H \tag{35}$$

Additional impulses are chosen until the desired number of impulses is reached. The gains of all pulses may be reoptimized after the selection of each new pulse [87].

Variations are possible. The multipulse coder described in ITU standard G.723.1 transmits a single gain for all the impulses, plus sign bits for each individual impulse. The G.723.1 coder restricts all impulse locations to be either odd or even; the choice of odd or even locations is coded using one bit per subframe [50]. The regular pulse excited LPC algorithm, which was the first GSM full-rate speech coder, synthesized speech using a train of impulses spaced one per 4 samples, all scaled by a single gain term [65]. The alignment of the pulse train was restricted to one of four possible locations, chosen in a closed-loop fashion together with a gain, an adaptive codebook delay, and an adaptive codebook gain.

Singhal and Atal demonstrated that the quality of MPLPC may be improved at low bit rates by modeling the periodic component of an LPC excitation vector using a pitch prediction filter [87]. Using a pitch prediction filter, the LPC excitation signal becomes

$$u(n) = bu(n-D) + \sum_{m=1}^{M} c_{k_m}(n) \tag{36}$$

where the signal $c_k(n)$ is an impulse located at $n = k$ and b is the pitch prediction filter gain. Singhal and Atal proposed choosing D before the locations of any impulses are known, by minimizing the following perceptually weighted error:

$$|E_D|^2 = |\tilde{S} - bX_D H|^2, X_D = [u(-D), \ldots, u((L-1)-D)] \tag{37}$$

The G.723.1 multipulse LPC coder and the GSM (Global System for Mobile Communication) full-rate RPE-LTP (regular-pulse excitation with long-term prediction) coder both use a closed-loop pitch predictor, as do all standardized variations of the CELP coder (see Sections 4.5.2 and 4.5.3). Typically, the pitch delay and gain are optimized first, and then the gains of any additional excitation vectors (e.g., impulses in an MPLPC algorithm) are selected to minimize the remaining error.

4.5.2. Code-Excited LPC (CELP). LPC analysis finds a filter $1/A(z)$ whose excitation is uncorrelated for correlation distances smaller than the order of the filter. Pitch prediction, especially closed-loop pitch prediction, removes much of the remaining intersample correlation. The spectrum of the pitch prediction residual looks like the spectrum of uncorrelated Gaussian noise, but replacing the residual with real noise (noise that is independent of the original signal) yields poor speech quality. Apparently, some of the temporal details of the pitch prediction residual are perceptually important. Schroeder and Atal proposed modeling the pitch prediction residual using a stochastic excitation vector $c_k(n)$ chosen from a list of stochastic excitation vectors, $k = 1, \ldots, K$, known to both the transmitter and receiver [85]:

$$u(n) = bu(n-D) + gc_k(n) \tag{38}$$

The list of stochastic excitation vectors is called a *stochastic codebook*, and the index of the stochastic codevector is chosen in order to minimize the perceptually weighted error metric $|E_k|^2$. Rose and Barnwell discussed the similarity between the search for an optimum stochastic codevector index k and the search for an optimum predictor delay D [82], and Kleijn et al. coined the term "adaptive codebook" to refer to the list of delayed excitation signals $u(n-D)$ which the coder considers during closed-loop pitch delay optimization (Fig. 9).

The CELP algorithm was originally not considered efficient enough to be used in real-time speech coding, but a number of computational simplifications were proposed that resulted in real-time CELP-like algorithms. Trancoso and Atal proposed efficient search methods based on the truncated impulse response of the filter $W(z)/A(z)$, as discussed in Section 4.4 [3,97]. Davidson and Lin separately proposed center clipping the stochastic codevectors, so that most of the samples in each codevector are zero [15,67]. Lin also proposed structuring the stochastic codebook so that each codevector is a slightly-shifted version of the previous codevector; such a codebook is called an *overlapped codebook* [67]. Overlapped stochastic codebooks are rarely used in practice today, but overlapped-codebook search methods are often used to reduce the computational complexity of an adaptive codebook search. In the search of

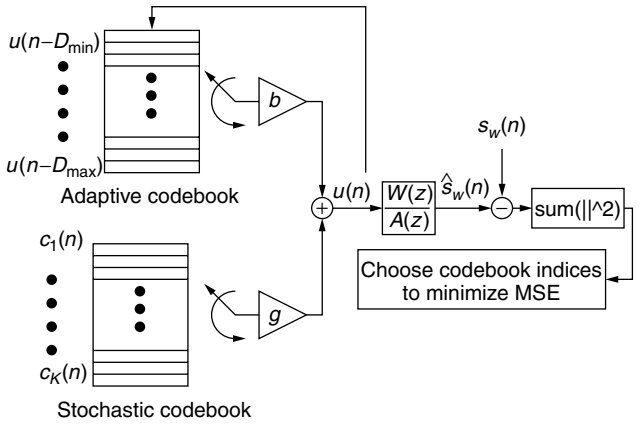


Figure 9. The code-excited LPC algorithm (CELP) constructs an LPC excitation signal by optimally choosing input vectors from two codebooks: an "adaptive" codebook, which represents the pitch periodicity; and a "stochastic" codebook, which represents the unpredictable innovations in each speech frame.

an overlapped codebook, the correlation R_X and autocorrelation Σ introduced in Section 4.4 may be recursively computed, thus greatly reducing the complexity of the codebook search [63].

Most CELP coders optimize the adaptive codebook index and gain first, and then choose a stochastic codevector and gain in order to minimize the remaining perceptually weighted error. If all the possible pitch periods are longer than one subframe, then the entire content of the adaptive codebook is known before the beginning of the codebook search, and the efficient overlapped codebook search methods proposed by Lin may be applied [67]. In practice, the pitch period of a female speaker is often shorter than one subframe. In order to guarantee that the entire adaptive codebook is known before beginning a codebook search, two methods are commonly used: (1) the adaptive codebook search may simply be constrained to only consider pitch periods longer than L samples — in this case, the adaptive codebook will lock onto values of D that are an integer multiple of the actual pitch period (if the same integer multiple is not chosen for each subframe, the reconstructed speech quality is usually good); and (2) adaptive codevectors with delays of $D < L$ may be constructed by simply repeating the most recent D samples as necessary to fill the subframe.

4.5.3. SELP, VSELP, ACELP, and LD-CELP. Rose and Barnwell demonstrated that reasonable speech quality is achieved if the LPC excitation vector is computed completely recursively, using two closed-loop pitch predictors in series, with no additional information [82]. In their "self-excited LPC" algorithm (SELP), the LPC excitation is initialized during the first subframe using a vector of samples known at both the transmitter and receiver. For all frames after the first, the excitation is the sum of an arbitrary number of adaptive codevectors:

$$u(n) = \sum_{m=1}^{M} b_m u(n - D_m) \tag{39}$$

Kleijn et al. developed efficient recursive algorithms for searching the adaptive codebook in SELP coder and other LPC-AS coders [63].

Just as there may be more than one adaptive codebook, it is also possible to use more than one stochastic codebook. The vector-sum excited LPC algorithm (VSELP) models the LPC excitation vector as the sum of one adaptive and two stochastic codevectors [25]:

$$u(n) = bu(n - D) + \sum_{m=1}^{2} g_m c_{k_m}(n) \tag{40}$$

The two stochastic codebooks are each relatively small (typically 32 vectors), so that each of the codebooks may be searched efficiently. The adaptive codevector and the two stochastic codevectors are chosen sequentially. After selection of the adaptive codevector, the stochastic codebooks are transformed using a modified Gram–Schmidt orthogonalization, so that the perceptually weighted speech vectors generated during the first stochastic codebook search are all orthogonal to the perceptually weighted adaptive codevector. Because of this orthogonalization, the stochastic codebook search results in the choice of a stochastic codevector that is jointly optimal with the adaptive codevector, rather than merely sequentially optimal. VSELP is the basis of the Telecommunications Industry Associations digital cellular standard IS-54.

The algebraic CELP (ACELP) algorithm creates an LPC excitation by choosing just one vector from an adaptive codebook and one vector from a fixed codebook. In the ACELP algorithm, however, the fixed codebook is composed of binary-valued or trinary-valued algebraic codes, rather than the usual samples of a Gaussian noise process [1]. Because of the simplicity of the codevectors, it is possible to search a very large fixed codebook very quickly using methods that are a hybrid of standard CELP and MPLPC search algorithms. ACELP is the basis of the ITU standard G.729 coder at 8 kbps. ACELP codebooks may be somewhat larger than the codebooks in a standard CELP coder; the codebook in G.729, for example, contains 8096 codevectors per subframe.

Most LPC-AS coders operate at very low bit rates, but require relatively large buffering delays. The low-delay CELP coder (LD-CELP) operates at 16 kbps [10,47] and is designed to obtain the best possible speech quality, with the constraint that the total algorithmic delay of a tandem coder and decoder must be no more than 2 ms. LPC analysis and codevector search are computed once per 2 ms (16 samples). Transmission of LPC coefficients once per two milliseconds would require too many bits, so LPC coefficients are computed in a recursive backward-adaptive fashion. Before coding or decoding each frame, samples of $\hat{s}(n)$ from the previous frame are windowed, and used to update a recursive estimate of the autocorrelation function. The resulting autocorrelation coefficients are similar to those that would be obtained using a relatively long asymmetric analysis window. LPC coefficients are then computed from the autocorrelation function using the Levinson–Durbin algorithm.

4.6. Line Spectral Frequencies (LSFs) or Line Spectral Pairs (LSPs)

Linear prediction can be viewed as an inverse filtering procedure in which the speech signal is passed through an all-zero filter $A(z)$. The filter coefficients of $A(z)$ are chosen such that the energy in the output, that is, the residual or error signal, is minimized. Alternatively, the inverse filter $A(z)$ can be transformed into two other filters $P(z)$ and $Q(z)$. These new filters turn out to have some interesting properties, and the representation based on them, called the *line spectrum pairs* [89,91], has been used in speech coding and synthesis applications.

Let $A(z)$ be the frequency response of an LPC inverse filter of order p:

$$A(z) = -\sum_{i=0}^{p} a_i z^{-i}$$

with $a_0 = -1$. The a_i values are real, and all the zeros of $A(z)$ are inside the unit circle.

If we use the lattice formulation of LPC, we arrive at a recursive relation between the mth stage $[A_m(z)]$ and the one before it $[A_{m-1}(z)]$ For the pth-order inverse filter, we have

$$A_p(z) = A_{p-1}(z) - k_p z^{-p} A_{p-1}(z^{-1})$$

By allowing the recursion to go one more iteration, we obtain

$$A_{p+1}(z) = A_p(z) - k_{p+1} z^{-(p+1)} A_p(z^{-1}) \tag{41}$$

If we choose $k_{p+1} = \pm 1$ in Eq. (41), we can define two new polynomials as follows:

$$P(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{42}$$

$$Q(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{43}$$

Physically, $P(z)$ and $Q(z)$ can be interpreted as the inverse transfer function of the vocal tract for the *open-glottis* and *closed-glottis* boundary conditions, respectively [22], and $P(z)/Q(z)$ is the driving-point impedance of the vocal tract as seen from the glottis [36].

If p is odd, the formulae for p_n and q_n are as follows:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ &= \prod_{n=1}^{(p+1)/2} (1 - e^{jp_n} z^{-1})(1 - e^{-jp_n} z^{-1}) \end{aligned} \tag{44}$$

$$\begin{aligned} Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \\ &= (1 - z^{-2}) \prod_{n=1}^{(p-1)/2} (1 - e^{jq_n} z^{-1})(1 - e^{-jq_n} z^{-1}) \end{aligned} \tag{45}$$

The LSFs have some interesting characteristics: the frequencies $\{p_n\}$ and $\{q_n\}$ are related to the formant frequencies; the dynamic range of $\{p_n\}$ and $\{q_n\}$ is limited and the two alternate around the unit circle $(0 \le p_1 \le q_1 \le p_2 \ldots)$; $\{p_n\}$ and $\{q_n\}$ are correlated so that intraframe prediction is possible; and they change slowly from one frame to another, hence, interframe prediction is also possible. The interleaving nature of the $\{p_n\}$ and $\{q_n\}$ allow for efficient iterative solutions [58].

Almost all LPC-based coders today use the LSFs to represent the LP parameters. Considerable recent research has been devoted to methods for efficiently quantizing the LSFs, especially using vector quantization (VQ) techniques. Typical algorithms include predictive VQ, split VQ [76], and multistage VQ [66,74]. All of these methods are used in the ITU standard ACELP coder G.729: the moving-average vector prediction residual is quantized using a 7-bit first-stage codebook, followed by second-stage quantization of two subvectors using independent 5-bit codebooks, for a total of 17 bits per frame [49,84].

5. LPC VOCODERS

5.1. The LPC-10e Vocoder

The 2.4-kbps LPC-10e vocoder (Fig. 10) is one of the earliest and one of the longest-lasting standards for low-bit-rate digital speech coding [8,16]. This standard was originally proposed in the 1970s, and was not officially replaced until the selection of the MELP 2.4-kbps coding standard in 1996 [64]. Speech coded using LPC-10e sounds metallic and synthetic, but it is intelligible.

In the LPC-10e algorithm, speech is first windowed using a Hamming window of length 22.5ms. The gain (G) and coefficients (a_i) of a linear prediction filter are calculated for the entire frame using the Levinson–Durbin recursion. Once G and a_i have been computed, the LPC residual signal $d(n)$ is computed:

$$d(n) = \frac{1}{G}\left(s(n) - \sum_{i=1}^{p} a_i s(n-i)\right) \tag{46}$$

The residual signal $d(n)$ is modeled using either a periodic train of impulses (if the speech frame is voiced) or an uncorrelated Gaussian random noise signal (if the frame is unvoiced). The voiced/unvoiced decision is based on the average magnitude difference function (AMDF),

$$\Phi_d(m) = \frac{1}{N - |m|} \sum_{n=|m|}^{N-1} |d(n) - d(n - |m|)| \tag{47}$$

The frame is labeled as voiced if there is a trough in $\Phi_d(m)$ that is large enough to be caused by voiced excitation. Only values of m between 20 and 160 are examined, corresponding to pitch frequencies between 50 and 400 Hz. If the minimum value of $\Phi_d(m)$ in this range is less than

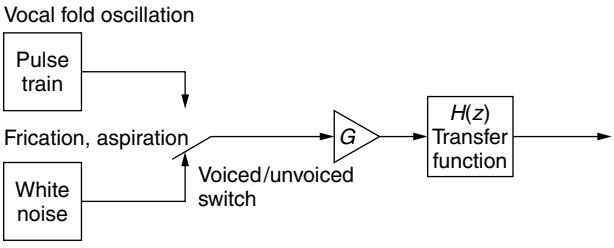


Figure 10. A simplified model of speech production whose parameters can be transmitted efficiently across a digital channel.

a threshold, the frame is declared voiced, and otherwise it is declared unvoiced [8].

If the frame is voiced, then the LPC residual is represented using an impulse train of period T_0, where

$$T_0 = \arg \min_{m=20}^{160} \Phi_d(m) \tag{48}$$

If the frame is unvoiced, a pitch period of $T_0 = 0$ is transmitted, indicating that an uncorrelated Gaussian random noise signal should be used as the excitation of the LPC synthesis filter.

5.2. Mixed-Excitation Linear Prediction (MELP)

The mixed-excitation linear prediction (MELP) coder [69] was selected in 1996 by the United States Department of Defense Voice Processing Consortium (DDVPC) to be the U.S. Federal Standard at 2.4 kbps, replacing LPC-10e. The MELP coder is based on the LPC model with additional features that include mixed excitation, aperiodic pulses, adaptive spectral enhancement, pulse dispersion filtering, and Fourier magnitude modeling [70]. The synthesis model for the MELP coder is illustrated in Fig. 11. LP coefficients are converted to LSFs and a multistage vector quantizer (MSVQ) is used to quantize the LSF vectors. For voiced segments a total of 54 bits that represent: LSF parameters (25), Fourier magnitudes of the prediction residual signal (8), gain (8), pitch (7), bandpass voicing (4), aperiodic flag (1), and a sync bit are sent. The Fourier magnitudes are coded with an 8-bit VQ and the associated codebook is searched with a perceptually-weighted Euclidean distance. For unvoiced segments, the Fourier magnitudes, bandpass voicing, and the aperiodic flag bit are not sent. Instead, 13 bits that implement forward error correction (FEC) are sent. The performance of MELP at 2.4 kbps is similar to or better than that of the federal standard at 4.8 kbps (FS 1016) [92]. Versions of MELP coders operating at 1.7 kbps [68] and 4.0 kbps [90] have been reported.

5.3. Multiband Excitation (MBE)

In multiband excitation (MBE) coding the voiced/unvoiced decision is not a binary one; instead, a series of voicing decisions are made for independent harmonic intervals [31]. Since voicing decisions can be made in different frequency bands individually, synthesized speech may be partially voiced and partially unvoiced. An improved version of the MBE was introduced in the late 1980s [7,35] and referred to as the IMBE coder. The IMBE

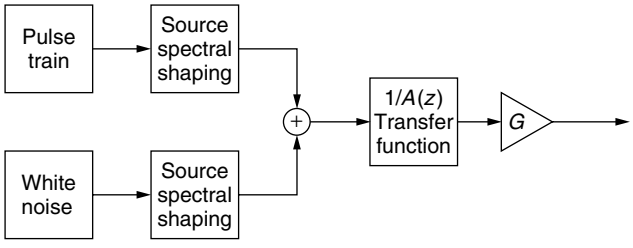


Figure 11. The MELP speech synthesis model.

at 2.4 kbps produces better sound quality than does the LPC-10e. The IMBE was adopted as the Inmarsat-M coding standard for satellite voice communication at a total rate of 6.4 kbps, including 4.15 kbps of source coding and 2.25 kbps of channel coding [104]. The Advanced MBE (AMBE) coder was adopted as the Inmarsat Mini-M standard at a 4.8 kbps total data rate, including 3.6 kbps of speech and 1.2 kbps of channel coding [18,27]. In [14] an enhanced multiband excitation (EMBE) coder was presented. The distinguishing features of the EMBE coder include signal-adaptive multimode spectral modeling and parameter quantization, a two-band signal-adaptive frequency-domain voicing decision, a novel VQ scheme for the efficient encoding of the variable-dimension spectral magnitude vectors at low rates, and multiclass selective protection of spectral parameters from channel errors. The 4-kbps EMBE coder accounts for both source (2.9 kbps) and channel (1.1 kbps) coding and was designed for satellite-based communication systems.

5.4. Prototype Waveform Interpolative (PWI) Coding

A different kind of coding technique that has properties of both waveform and LPC-based coders has been proposed [59,60] and is called *prototype waveform interpolation* (PWI). PWI uses both interpolation in the frequency domain and forward–backward prediction in the time domain. The technique is based on the assumption that, for voiced speech, a perceptually accurate speech signal can be reconstructed from a description of the waveform of a single, representative pitch cycle per interval of 20–30 ms. The assumption exploits the fact that voiced speech can be interpreted as a concentration of slowly evolving pitch cycle waveforms. The prototype waveform is described by a set of linear prediction (LP) filter coefficients describing the formant structure and a prototype excitation waveform, quantized with analysis-by-synthesis procedures. The speech signal is reconstructed by filtering an excitation signal consisting of the concatenation of (infinitesimal) sections of the instantaneous excitation waveforms. By coding the voiced and unvoiced components separately, a 2.4-kbps version of the coder performed similarly to the 4.8-kbps FS1016 standard [61].

Recent work has aimed at reducing the computational complexity of the coder for rates between 1.2 and 2.4 kbps by including a time-varying waveform sampling rate and a cubic B-spline waveform representation [62,86].

6. MEASURES OF SPEECH QUALITY

Deciding on an appropriate measurement of quality is one of the most difficult aspects of speech coder design, and is an area of current research and standardization. Early military speech coders were judged according to only one criterion: intelligibility. With the advent of consumer-grade speech coders, intelligibility is no longer a sufficient condition for speech coder acceptability. Consumers want speech that sounds "natural." A large number of subjective and objective measures have been developed to quantify "naturalness," but it must be stressed that any scalar measurement of "naturalness" is an oversimplification.

"Naturalness" is a multivariate quantity, including such factors as the metallic versus breathy quality of speech, the presence of noise, the color of the noise (narrowband noise tends to be more annoying than wideband noise, but the parameters that predict "annoyance" are not well understood), the presence of unnatural spectral envelope modulations (e.g., flutter noise), and the absence of natural spectral envelope modulations.

6.1. Psychophysical Measures of Speech Quality (Subjective Tests)

The final judgment of speech coder quality is the judgment made by human listeners. If consumers (and reviewers) like the way the product sounds, then the speech coder is a success. The reaction of consumers can often be predicted to a certain extent by evaluating the reactions of experimental listeners in a controlled psychophysical testing paradigm. Psychophysical tests (often called "subjective tests") vary depending on the quantity being evaluated, and the structure of the test.

6.1.1. Intelligibility. Speech coder intelligibility is evaluated by coding a number of prepared words, asking listeners to write down the words they hear, and calculating the percentage of correct transcriptions (an adjustment for guessing may be subtracted from the score). The diagnostic rhyme test (DRT) and diagnostic alliteration test (DALT) are intelligibility tests which use a controlled vocabulary to test for specific types of intelligibility loss [101,102]. Each test consists of 96 pairs of confusable words spoken in isolation. The words in a pair differ in only one distinctive feature, where the distinctive feature dimensions proposed by Voiers are voicing, nasality, sustention, sibilation, graveness, and compactness. In the DRT, the words in a pair differ in only one distinctive feature of the initial consonant; for instance, "jest" and "guest" differ in the sibilation of the initial consonant. In the DALT, words differ in the final consonant; for instance, "oaf" and "oath" differ in the graveness of the final consonant. Listeners hear one of the words in each pair, and are asked to select the word from two written alternatives. Professional testing firms employ trained listeners who are familiar with the speakers and speech tokens in the database, in order to minimize test-retest variability.

Intelligibility scores quoted in the speech coding literature often refer to the composite results of a DRT. In a comparison of two federal standard coders, the LPC 10e algorithm resulted in 90% intelligibility, while the FS-1016 CELP algorithm had 91% intelligibility [64]. An evaluation of waveform interpolative (WI) coding published DRT scores of 87.2% for the WI algorithm, and 87.7% for FS-1016 [61].

6.1.2. Numerical Measures of Perceptual Quality. Perhaps the most commonly used speech quality measure is the mean opinion score (MOS). A mean opinion score is computed by coding a set of spoken phrases using a variety of coders, presenting all of the coded speech together with undegraded speech in random order, asking listeners to rate the quality of each phrase on a numerical scale, and then averaging the numerical ratings of all phrases coded by a particular coder. The five-point numerical scale is associated with a standard set of descriptive terms: 5 = excellent, 4 = good, 3 = fair, 2 = poor, and 1 = bad. A rating of 4 is supposed to correspond to standard toll-quality speech, quantized at 64 kbps using ITU standard G.711 [48].

Mean opinion scores vary considerably depending on background noise conditions; for example, CVSD performs significantly worse than LPC-based methods in quiet recording conditions, but significantly better under extreme noise conditions [96]. Gender of the speaker may also affect the relative ranking of coders [96]. Expert listeners tend to give higher rankings to speech coders with which they are familiar, even when they are not consciously aware of the order in which coders are presented [96]. Factors such as language and location of the testing laboratory may shift the scores of all coders up or down, but tend not to change the rank order of individual coders [39]. For all of these reasons, a serious MOS test must evaluate several reference coders in parallel with the coder of interest, and under identical test conditions. If an MOS test is performed carefully, intercoder differences of approximately 0.15 opinion points may be considered significant. Figure 12 is a plot of MOS as a function of bit rate for coders evaluated under quiet listening conditions in five published studies (one study included separately tabulated data from two different testing sites [96]).

The diagnostic acceptability measure (DAM) is an attempt to control some of the factors that lead to variability in published MOS scores [100]. The DAM employs trained listeners, who rate the quality of standardized test phrases on 10 independent perceptual scales, including six scales that rate the speech itself (fluttering, thin, rasping, muffled, interrupted, nasal), and four scales that rate the background noise (hissing, buzzing, babbling, rumbling). Each of these is a 100-point scale, with a range of approximately 30 points between the LPC-10e algorithm (50 points) and clean speech (80 points) [96]. Scores on the various perceptual scales are combined into a composite quality rating. DAM scores are useful for pointing out specific defects in a speech coding algorithm. If the only desired test outcome is a relative quality ranking of multiple coders, a carefully controlled MOS test in which all coders of interest are tested under the same conditions may be as reliable as DAM testing [96].

6.1.3. Comparative Measures of Perceptual Quality. It is sometimes difficult to evaluate the statistical significance of a reported MOS difference between two coders. A more powerful statistical test can be applied if coders are evaluated in explicit A/B comparisons. In a comparative test, a listener hears the same phrase coded by two different coders, and chooses the one that sounds better. The result of a comparative test is an apparent preference score, and an estimate of the significance of the observed preference; for example, in a 1999 study, WI coding at 4.0 kbps was preferred to 4 kbps HVXC 63.7% of the time, to 5.3 kbps G.723.1 57.5% of the time (statistically significant differences), and to 6.3 kbps G.723.1 53.9% of the time (not statistically significant) [29]. It should be

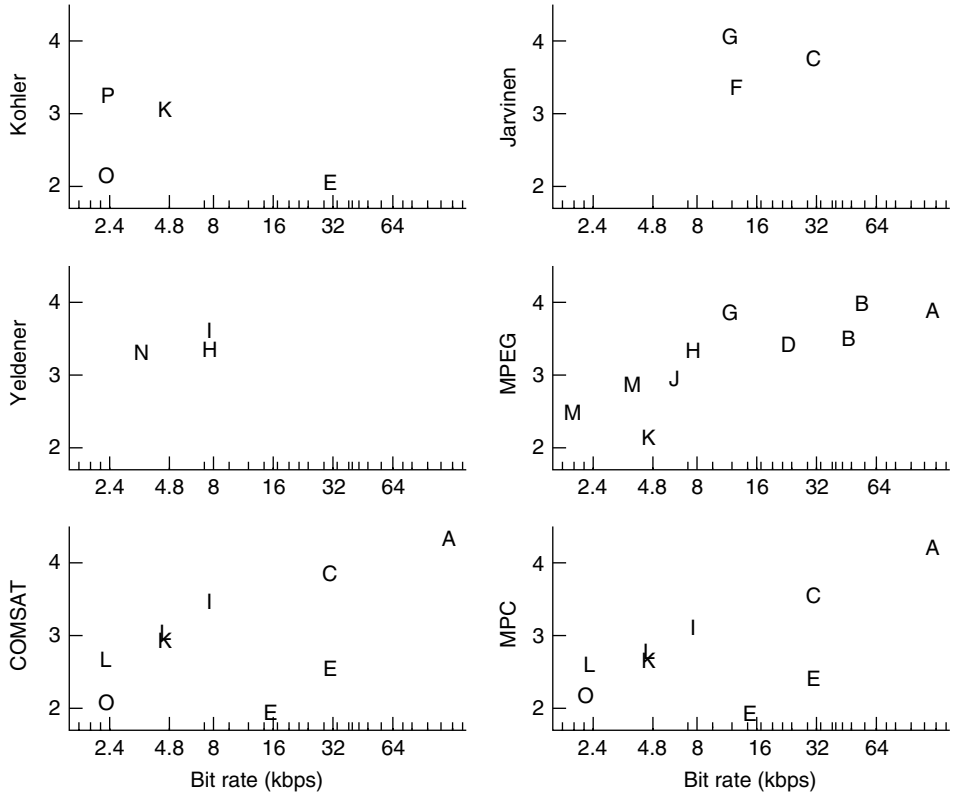


Figure 12. Mean opinion scores from five published studies in quiet recording conditions — Jarvinen [53], Kohler [64], MPEG [39], Yeldener [107], and the COMSAT and MPC sites from Tardelli et al. [96]: (A) unmodified speech, (B) ITU G.722 subband ADPCM, (C) ITU G.726 ADPCM, (D) ISO MPEG-II layer 3 subband audio coder, (E) DDVPC CVSD, (F) GSM full-rate RPE-LTP, (G) GSM EFR ACELP, (H) ITU G.729 ACELP, (I) TIA IS54 VSELP, (J) ITU G.723.1 MPLPC, (K) DDVPC FS-1016 CELP, (L) sinusoidal transform coding, (M) ISO MPEG-IV HVXC, (N) Inmarsat mini-M AMBE, (O) DDVPC FS-1015 LPC-10e, (P) DDVPC MELP.

noted that "statistical significance" in such a test refers only to the probability that the same listeners listening to the same waveforms will show the same preference in a future test.

6.2. Algorithmic Measures of Speech Quality (Objective Measures)

Psychophysical testing is often inconvenient; it is not possible to run psychophysical tests to evaluate every proposed adjustment to a speech coder. For this reason, a number of algorithms have been proposed that approximate, to a greater or lesser extent, the results of psychophysical testing.

The signal-to-noise ratio of a frame of N speech samples starting at sample number n may be defined as

$$\mathrm{SNR}(n) = \frac{\sum_{m=n}^{n+N-1} s^2(m)}{\sum_{m=n}^{n+N-1} e^2(m)} \tag{49}$$

High-energy signal components can mask quantization error, which is synchronous with the signal component, or separated by at most a few tens of milliseconds. Over longer periods of time, listeners accumulate a general perception of quantization noise, which can be modeled as the average log segmental SNR:

$$\mathrm{SEGSNR} = \frac{1}{K}\sum_{k=0}^{K-1} 10\log_{10}\mathrm{SNR}(kN) \tag{50}$$

High-amplitude signal components tend to mask quantization error components at nearby frequencies and times. A high-amplitude spectral peak in the speech signal is able to mask quantization error components at the same frequency, at higher frequencies, and to a much lesser extent, at lower frequencies. Given a short-time speech spectrum $S(e^{j\omega})$, it is possible to compute a short-time "masking spectrum" $M(e^{j\omega})$ which describes the threshold energy at frequency ω below which noise components are inaudible. The perceptual salience of a noise signal $e(n)$ may be estimated by filtering the noise signal into K different subband signals $e_k(n)$, and computing the ratio between the noise energy and the masking threshold in each subband:

$$\mathrm{NMR}(n,k) = \frac{\sum_{m=n}^{n+N-1} e_k^2(m)}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2\, d\omega} \tag{51}$$

where ω_k is the lower edge of band k, and ω_{k+1} is the upper band edge. The band edges must be close enough together that all of the signal components in band k are effective in masking the signal $e_k(n)$. The requirement of effective masking is met if each band is exactly one Bark in width, where the Bark frequency scale is described in many references [71,77].

Fletcher has shown that the perceived loudness of a signal may be approximated by adding the cube roots of the signal power in each one-bark subband, after properly accounting for masking effects [20]. The total

loudness of a quantization noise signal may therefore be approximated as

$$\mathrm{NMR}(n) = \sum_{k=0}^{K-1} \left(\frac{\sum_{m=n}^{n+N-1} e_k^2[m]}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2 \, d\omega} \right)^{1/3} \tag{52}$$

The ITU perceptual speech quality measure (PSQM) computes the perceptual quality of a speech signal by filtering the input and quantized signals using a Bark-scale filterbank, nonlinearly compressing the amplitudes in each band, and then computing an average subband signal to noise ratio [51]. The development of algorithms that accurately predict the results of MOS or comparative testing is an area of active current research, and a number of improvements, alternatives, and/or extensions to the PSQM measure have been proposed. An algorithm that has been the focus of considerable research activity is the Bark spectral distortion measure [73,103,105,106]. The ITU has also proposed an extension of the PSQM standard called perceptual evaluation of speech quality (PESQ) [81], which will be released as ITU standard P.862.

7. NETWORK ISSUES

7.1. Voice over IP

Speech coding for the voice over Internet Protocol (VOIP) application is becoming important with the increasing dependency on the Internet. The first VoIP standard was published in 1998 as recommendation H.323 [52] by the International Telecommunications Union (ITU-T). It is a protocol for multimedia communications over local area networks using packet switching, and the voice-only subset of it provides a platform for IP-based telephony. At high bit rates, H.323 recommends the coders G.711 (3.4 kHz at 48, 56, and 64 kbps) and G.722 (wideband speech and music at 7 kHz operating at 48, 56, and 64 kbps) while at the lower bit rates G.728 (3.4 kHz at 16 kbps), G.723 (5.3 and 6.5 kbps), and G.729 (8 kbps) are recommended [52].

In 1999, a competing and simpler protocol named the Session Initiation Protocol (SIP) was developed by the Internet Engineering Task Force (IETF) Multiparty Multimedia Session Control working group and published as RFC 2543 [19]. SIP is a signaling protocol for Internet conferencing and telephony, is independent of the packet layer, and runs over UDP or TCP although it supports more protocols and handles the associations between Internet end systems. For now, both systems will coexist but it is predicted that the H.323 and SIP architectures will evolve such that two systems will become more similar.

Speech transmission over the Internet relies on sending "packets" of the speech signal. Because of network congestion, packet loss can occur, resulting in audible artifacts. High-quality VOIP, hence, would benefit from variable-rate source and channel coding, packet loss concealment, and jitter buffer/delay management. These are challenging issues and research efforts continue to generate high-quality speech for VOIP applications [38].

7.2. Embedded and Multimode Coding

When channel quality varies, it is often desirable to adjust the bit rate of a speech coder in order to match the channel capacity. Varying bit rates are achieved in one of two ways. In multimode speech coding, the transmitter and the receiver must agree on a bit rate prior to transmission of the coded bits. In embedded source coding, on the other hand, the bitstream of the coder operating at low bit rates is embedded in the bitstream of the coder operating at higher rates. Each increment in bit rate provides marginal improvement in speech quality. Lower bit rate coding is obtained by puncturing bits from the higher rate coder and typically exhibits graceful degradation in quality with decreasing bit rates.

ITU Standard G.727 describes an embedded ADPCM coder, which may be run at rates of 40, 32, 24, or 16 kbps (5, 4, 3, or 2 bits per sample) [46]. Embedded ADPCM algorithms are a family of variable bit rate coding algorithms operating on a sample per sample basis (as opposed to, e.g., a subband coder that operates on a frame-by-frame basis) that allows for bit dropping after encoding. The decision levels of the lower-rate quantizers are subsets of those of the quantizers at higher rates. This allows for bit reduction at any point in the network without the need of coordination between the transmitter and the receiver.

The prediction in the encoder is computed using a more coarse quantization of $\hat{d}(n)$ than the quantization actually transmitted. For example, 5 bits per sample may be transmitted, but as few as 2 bits may be used to reconstruct $\hat{d}(n)$ in the prediction loop. Any bits not used in the prediction loop are marked as "optional" by the signaling channel mode flag. If network congestion disrupts traffic at a router between sender and receiver, the router is allowed to drop optional bits from the coded speech packets.

Embedded ADPCM algorithms produce codewords that contain enhancement and core bits. The feedforward (FF) path of the codec utilizes both enhancement bits and core bits, while the feedback (FB) path uses core bits only. With this structure, enhancement bits can be discarded or dropped during network congestion.

An important example of a multimode coder is QCELP, the speech coder standard that was adopted by the TIA North American digital cellular standard based on code-division multiple access (CDMA) technology [9]. The coder selects one of four data rates every 20 ms depending on the speech activity; for example, background noise is coded at a lower rate than speech. The four rates are approximately 1 kbps (eighth rate), 2 kbps (quarter rate), 4 kbps (half rate), and 8 kbps (full rate). QCELP is based on the CELP structure but integrates implementation of the different rates, thus reducing the average bit rate. For example, at the higher rates, the LSP parameters are more finely quantized and the pitch and codebook parameters are updated more frequently [23]. The coder provides good quality speech at average rates of 4 kbps.

Another example of a multimode coder is ITU standard G.723.1, which is an LPC-AS coder that can operate at

2 rates: 5.3 or 6.3 kbps [50]. At 6.3 kbps, the coder is a multipulse LPC (MPLPC) coder while the 5.3-kbps coder is an algebraic CELP (ACELP) coder. The frame size is 30 ms with an additional lookahead of 7.5 ms, resulting in a total algorithmic delay of 67.5 ms. The ACELP and MPLPC coders share the same LPC analysis algorithm and frame/subframe structure, so that most of the program code is used by both coders. As mentioned earlier, in ACELP, an algebraic transformation of the transmitted index produces the excitation signal for the synthesizer. In MPLPC, on the other hand, minimizing the perceptual-error weighting is achieved by choosing the amplitude and position of a number of pulses in the excitation signal. Voice activity detection (VAD) is used to reduce the bit rate during silent periods, and switching from one bit rate to another is done on a frame-by-frame basis.

Multimode coders have been proposed over a wide variety of bandwidths. Taniguchi et al. proposed a multimode ADPCM coder at bit rates between 10 and 35 kbps [94]. Johnson and Taniguchi proposed a multimode CELP algorithm at data rates of 4.0–5.3 kbps in which additional stochastic codevectors are added to the LPC excitation vector when channel conditions are sufficiently good to allow high-quality transmission [55]. The European Telecommunications Standards Institute (ETSI) has recently proposed a standard for adaptive multirate coding at rates between 4.75 and 12.2 kbps.

7.3. Joint Source-Channel Coding

In speech communication systems, a major challenge is to design a system that provides the best possible speech quality throughout a wide range of channel conditions. One solution consists of allowing the transceivers to monitor the state of the communication channel and to dynamically allocate the bitstream between source and channel coding accordingly. For low-SNR channels, the source coder operates at low bit rates, thus allowing powerful forward error control. For high-SNR channels, the source coder uses its highest rate, resulting in high speech quality, but with little error control. An adaptive algorithm selects a source coder and channel coder based on estimates of channel quality in order to maintain a constant total data rate [95]. This technique is called *adaptive multirate* (AMR) coding, and requires the simultaneous implementation of an AMR source coder [24], an AMR channel coder [26,28], and a channel quality estimation algorithm capable of acquiring information about channel conditions with a relatively small tracking delay.

The notion of determining the relative importance of bits for further unequal error protection (UEP) was pioneered by Rydbeck and Sundberg [83]. Rate-compatible channel codes, such as Hagenauer's rate compatible punctured convolutional codes (RCPC) [34], are a collection of codes providing a family of channel coding rates. By puncturing bits in the bitstream, the channel coding rate of RCPC codes can be varied instantaneously, providing UEP by imparting on different segments different degrees of protection. Cox et al. [13] address the issue of channel coding and illustrate how RCPC codes can be used to build a speech transmission scheme for mobile radio channels. Their approach is based on a subband coder with dynamic bit allocation proportional to the average energy of the bands. RCPC codes are then used to provide UEP.

Relatively few AMR systems describing source and channel coding have been presented. The AMR systems [99,98,75,44] combine different types of variable rate CELP coders for source coding with RCPC and cyclic redundancy check (CRC) codes for channel coding and were presented as candidates for the European Telecommunications Standards Institute (ETSI) GSM AMR codec standard. In [88], UEP is applied to perceptually based audio coders (PAC). The bitstream of the PAC is divided into two classes and punctured convolutional codes are used to provide different levels of protection, assuming a BPSK constellation.

In [5,6], a novel UEP channel encoding scheme is introduced by analyzing how symbol-wise puncturing of symbols in a trellis code and the rate-compatibility constraint (progressive puncturing pattern) can be used to derive rate-compatible punctured trellis codes (RCPT). While conceptually similar to RCPC codes, RCPT codes are specifically designed to operate efficiently on large constellations (for which Euclidean and Hamming distances are no longer equivalent) by maximizing the residual Euclidean distance after symbol puncturing. Large constellation sizes, in turn, lead to higher throughput and spectral efficiency on high SNR channels. An AMR system is then designed based on a perceptually-based embedded subband encoder. Since perceptually based dynamic bit allocations lead to a wide range of bit error sensitivities (the perceptually least important bits being almost insensitive to channel transmission errors), the channel protection requirements are determined accordingly. The AMR systems utilize the new rate-compatible channel coding technique (RCPT) for UEP and operate on an 8-PSK constellation. The AMR-UEP system is bandwidth efficient, operates over a wide range of channel conditions and degrades gracefully with decreasing channel quality.

Systems using AMR source and channel coding are likely to be integrated in future communication systems since they have the capability for providing graceful speech degradation over a wide range of channel conditions.

8. STANDARDS

Standards for landline public switched telephone service (PSTN) networks are established by the International Telecommunication Union (ITU) (*http://www.itu.int*). The ITU has promulgated a number of important speech and waveform coding standards at high bit rates and with very low delay, including G.711 (PCM), G.727 and G.726 (ADPCM), and G.728 (LDCELP). The ITU is also involved in the development of internetworking standards, including the voice over IP standard H.323. The ITU has developed one widely used low-bit-rate coding standard (G.729), and a number of embedded and multimode speech coding standards operating at rates between 5.3 kbps (G.723.1) and 40 kbps (G.727). Standard G.729 is a speech coder operating at 8 kbps, based on algebraic code-excited LPC (ACELP) [49,84]. G.723.1 is a multimode coder, capable of operating at either 5.3 or 6.3 kbps [50]. G.722

is a standard for wideband speech coding, and the ITU will announce an additional wideband standard within the near future. The ITU has also published standards for the objective estimation of perceptual speech quality (P.861 and P.862).

The ITU is a branch of the International Standards Organization (ISO) (*http://www.iso.ch*). In addition to ITU activities, the ISO develops standards for the Moving Picture Experts Group (MPEG). The MPEG-2 standard included digital audiocoding at three levels of complexity, including the layer 3 codec commonly known as MP3 [72]. The MPEG-4 motion picture standard includes a structured audio standard [40], in which speech and audio "objects" are encoded with header information specifying the coding algorithm. Low-bit-rate speech coding is performed using harmonic vector excited coding (HVXC) [43] or code-excited LPC (CELP) [41], and audiocoding is performed using time–frequency coding [42]. The MPEG homepage is at *drogo.cselt.stet.it/mpeg*.

Standards for cellular telephony in Europe are established by the European Telecommunications Standards Institute (ETSI) (*http://www.etsi.org*). ETSI speech coding standards are published by the Global System for Mobile Telecommunications (GSM) subcommittee. All speech coding standards for digital cellular telephone use are based on LPC-AS algorithms. The first GSM standard coder was based on a precursor of CELP called *regular-pulse excitation with long-term prediction* (RPE-LTP) [37,65]. Current GSM standards include the enhanced full-rate codec GSM 06.60 [32,53] and the adaptive multirate codec [33]; both standards use algebraic code-excited LPC (ACELP). At the time of writing, both ITU and ETSI are expected to announce new standards for wideband speech coding in the near future. ETSI's standard will be based on GSM AMR.

The Telecommunications Industry Association (*http://www.tiaonline.org*) published some of the first U.S. digital cellular standards, including the vector-sum-excited LPC (VSELP) standard IS54 [25]. In fact, both the initial U.S. and Japanese digital cellular standards were based on the VSELP algorithm. The TIA has been active in the development of standard TR41 for voice over IP.

The U.S. Department of Defense Voice Processing Consortium (DDVPC) publishes speech coding standards for U.S. government applications. As mentioned earlier, the original FS-1015 LPC-10e standard at 2.4 kbps [8,16], originally developed in the 1970s, was replaced in 1996 by the newer MELP standard at 2.4 kbps [92]. Transmission at slightly higher bit rates uses the FS-1016 CELP (CELP) standard at 4.8 kbps [17,56,57]. Waveform applications use the continuously variable slope delta modulator (CVSD) at 16 kbps. Descriptions of all DDVPC standards and code for most are available at *http://www.plh.af.mil/ddvpc/index.html*.

9. FINAL REMARKS

In this article, we presented an overview of coders that compress speech by attempting to match the time waveform as closely as possible (waveform coders), and coders that attempt to preserve perceptually relevant spectral properties of the speech signal (LPC-based and subband coders). LPC-based coders use a speech production model to parameterize the speech signal, while subband coders filter the signal into frequency bands and assign bits by either an energy or perceptual criterion. Issues pertaining to networking, such as voice over IP and joint source–channel coding, were also touched on. There are several other coding techniques that we have not discussed in this article because of space limitations. We hope to have provided the reader with an overview of the fundamental techniques of speech compression.

Acknowledgments

This research was supported in part by the NSF and HRL. We thank Alexis Bernard and Tomohiko Taniguchi for their suggestions on earlier drafts of the article.

BIOGRAPHIES

Mark A. Hasegawa-Johnson received his S.B., S.M., and Ph.D. degrees in electrical engineering and computer science from MIT in 1989, 1989, and 1996, respectively. From 1989 to 1990 he worked as a research engineer at Fujitsu Laboratories Ltd., Kawasaki, Japan, where he developed and patented a multimodal CELP speech coder with an efficient algebraic fixed codebook. From 1996–1999 he was a postdoctoral fellow in the Electrical Engineering Department at UCLA. Since 1999, he has been on the faculty of the University of Illinois at Urbana-Champaign. Dr. Hasegawa-Johnson holds four U.S. patents and is the author of four journal articles and twenty conference papers. His areas of interest include speech coding, automatic speech understanding, acoustics, and the physiology of speech production.

Abeer Alwan received her Ph.D. in electrical engineering from MIT in 1992. Since then, she has been with the Electrical Engineering Department at UCLA, California, as an assistant professor (1992–1996), associate professor (1996–2000), and professor (2000–present). Professor Alwan established and directs the Speech Processing and Auditory Perception Laboratory at UCLA (http://www.icsl.ucla.edu/~spapl). Her research interests include modeling human speech production and perception mechanisms and applying these models to speech-processing applications such as automatic recognition, compression, and synthesis. She is the recipient of the NSF Research Initiation Award (1993), the NIH FIRST Career Development Award (1994), the UCLA-TRW Excellence in Teaching Award (1994), the NSF Career Development Award (1995), and the Okawa Foundation Award in Telecommunications (1997). Dr. Alwan is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served as an elected member on the Acoustical Society of America Technical Committee on Speech Communication (1993–1999), on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics (1996–2000) and Speech Processing (1996–2001). She is an editor in chief of the journal *Speech Communication*.

BIBLIOGRAPHY

1. J.-P. Adoul, P. Mabilleau, M. Delprat, and S. Morisette, Fast CELP coding based on algebraic codes, *Proc. ICASSP*, 1987, pp. 1957–1960.
2. B. S. Atal, Predictive coding of speech at low bit rates, *IEEE Trans. Commun.* **30**: 600–614 (1982).
3. B. S. Atal, High-quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders, *Proc. ICASSP*, 1986, pp. 1681–1684.
4. B. S. Atal and J. R. Remde, A new model of LPC excitation for producing natural-sounding speech at low bit rates, *Proc. ICASSP*, 1982, pp. 614–617.
5. A. Bernard, X. Liu, R. Wesel, and A. Alwan, Channel adaptive joint-source channel coding of speech, *Proc. 32nd Asilomar Conf. Signals, Systems, and Computers*, 1998, Vol. 1, pp. 357–361.
6. A. Bernard, X. Liu, R. Wesel, and A. Alwan, Embedded joint-source channel coding of speech using symbol puncturing of trellis codes, *Proc. IEEE ICASSP*, 1999, Vol. 5, pp. 2427–2430.
7. M. S. Brandstein, P. A. Monta, J. C. Hardwick, and J. S. Lim, A real-time implementation of the improved MBE speech coder, *Proc. ICASSP*, 1990, Vol. 1: pp. 5–8.
8. J. P. Campbell and T. E. Tremain, Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm, *Proc. ICASSP*, 1986, pp. 473–476.
9. *CDMA, Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard*, Technical Report Proposed EIA/TIA Interim Standard, Telecommunications Industry Association TR45.5 Subcommittee, 1992.
10. J.-H. Chen et al., A low delay CELP coder for the CCITT 16 kb/s speech coding standard, *IEEE J. Select. Areas Commun.* **10**: 830–849 (1992).
11. J.-H. Chen and A. Gersho, Adaptive postfiltering for quality enhancement of coded speech, *IEEE Trans. Speech Audio Process.* **3**(1): 59–71 (1995).
12. R. Cox et al., New directions in subband coding, *IEEE JSAC* **6**(2): 391–409 (Feb. 1988).
13. R. Cox, J. Hagenauer, N. Seshadri, and C. Sundberg, Subband speech coding and matched convolutional coding for mobile radio channels, *IEEE Trans. Signal Process.* **39**(8): 1717–1731 (Aug. 1991).
14. A. Das and A. Gersho, Low-rate multimode multiband spectral coding of speech, *Int. J. Speech Tech.* **2**(4): 317–327 (1999).
15. G. Davidson and A. Gersho, Complexity reduction methods for vector excitation coding, *Proc. ICASSP*, 1986, pp. 2055–2058.
16. DDVPC, *LPC-10e Speech Coding Standard*, Technical Report FS-1015, U.S. Dept. of Defense Voice Processing Consortium, Nov. 1984.
17. DDVPC, *CELP Speech Coding Standard*, Technical Report FS-1016, U.S. Dept. of Defense Voice Processing Consortium, 1989.
18. S. Dimolitsas, Evaluation of voice coded performance for the Inmarsat Mini-M system, *Proc. 10th Int. Conf. Digital Satellite Communications*, 1995.
19. M. Handley et al., *SIP: Session Initiation Protocol*, IETF RFC, March 1999, *http://www.cs.columbia.edu/hgs/sip/sip.html*.
20. H. Fletcher, *Speech and Hearing in Communication*, Van Nostrand, Princeton, NJ, 1953.
21. D. Florencio, Investigating the use of asymmetric windows in CELP vocoders, *Proc. ICASSP*, 1993, Vol. II, pp. 427–430.
22. S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, New York, 1989.
23. W. Gardner, P. Jacobs, and C. Lee, QCELP: A variable rate speech coder for CDMA digital cellular, in B. Atal, V. Cuperman, and A. Gersho, eds., *Speech and Audio Coding for Wireless and Network Applications*, Kluwer, Dordrecht, The Netherlands, 1993, pp. 85–93.
24. A. Gersho and E. Paksoy, An overview of variable rate speech coding for cellular networks, *IEEE Int. Conf. Selected Topics in Wireless Communications Proc.*, June 1999, pp. 172–175.
25. I. Gerson and M. Jasiuk, Vector sum excited linear prediction (VSELP), in B. S. Atal, V. S. Cuperman, and A. Gersho, eds., *Advances in Speech Coding*, Kluwer, Dordrecht, The Netherlands, 1991, pp. 69–80.
26. D. Goeckel, Adaptive coding for time-varying channels using outdated fading estimates, *IEEE Trans. Commun.* **47**(6): 844–855 (1999).
27. R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*, CRC Press, Boca Raton, FL, 2000.
28. A. Goldsmith and S. G. Chua, Variable-rate variable power MQAM for fading channels, *IEEE Trans. Commun.* **45**(10): 1218–1230 (1997).
29. O. Gottesman and A. Gersho, Enhanced waveform interpolative coding at 4 kbps, *IEEE Workshop on Speech Coding*, Piscataway, NY, 1999, pp. 90–92.
30. K. Gould, R. Cox, N. Jayant, and M. Melchner, Robust speech coding for the indoor wireless channel, *ATT Tech. J.* **72**(4): 64–73 (1993).
31. D. W. Griffn and J. S. Lim, Multi-band excitation vocoder, *IEEE Trans. Acoust. Speech Signal Process.* **36**(8): 1223–1235 (1988).
32. Special Mobile Group (GSM), *Digital Cellular Telecommunications System: Enhanced Full Rate (EFR) Speech Transcoding*, Technical Report GSM 06.60, European Telecommunications Standards Institute (ETSI), 1997.
33. Special Mobile Group (GSM), *Digital Cellular Telecommunications System (Phase 2+): Adaptive Multi-rate (AMR) Speech Transcoding*, Technical Report GSM 06.90, European Telecommunications Standards Institute (ETSI), 1998.
34. J. Hagenauer, Rate-compatible punctured convolutional codes and their applications, *IEEE Trans. Commun.* **36**(4): 389–400 (1988).
35. J. C. Hardwick and J. S. Lim, A 4.8 kbps multi-band excitation speech coder, *Proc. ICASSP*, 1988, Vol. 1, pp. 374–377.
36. M. Hasegawa-Johnson, Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model, *J. Acoust. Soc. Am.* **108**(1): 457–460 (2000).
37. K. Hellwig et al., Speech codec for the european mobile radio system, *Proc. IEEE Global Telecomm. Conf.*, 1989.
38. O. Hersent, D. Gurle, and J.-P. Petit, *IP Telephony*, Addison-Wesley, Reading, MA, 2000.

39. ISO, *Report on the MPEG-4 Speech Codec Verification Tests*, Technical Report JTC1/SC29/WG11, ISO/IEC, Oct. 1998.
40. ISO/IEC, *Information Technology — Coding of Audiovisual Objects, Part 3: Audio, Subpart 1: Overview*, Technical Report ISO/JTC 1/SC 29/N2203, ISO/IEC, 1998.
41. ISO/IEC, *Information Technology — Coding of Audiovisual Objects, Part 3: Audio, Subpart 3: CELP*, Technical Report ISO/JTC 1/SC 29/N2203CELP, ISO/IEC, 1998.
42. ISO/IEC, *Information Technology — Coding of Audiovisual Objects, Part 3: Audio, Subpart 4: Time/Frequency Coding*, Technical Report ISO/JTC 1/SC 29/N2203TF, ISO/IEC, 1998.
43. ISO/IEC, *Information Technology — Very Low Bitrate Audio-Visual Coding, Part 3: Audio, Subpart 2: Parametric Coding*, Technical Report ISO/JTC 1/SC 29/N2203PAR, ISO/IEC, 1998.
44. H. Ito, M. Serizawa, K. Ozawa, and T. Nomura, An adaptive multi-rate speech codec based on mp-celp coding algorithm for etsi amr standard, *Proc. ICASSP*, 1998, Vol. 1, pp. 137–140.
45. ITU-T, *40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation* (*ADPCM*), Technical Report G.726, International Telecommunications Union, Geneva, 1990.
46. ITU-T, *5-, 4-, 3- and 2-bits per Sample Embedded Adaptive Differential Pulse Code Modulation* (*ADPCM*), Technical Report G.727, International Telecommunications Union, Geneva, 1990.
47. ITU-T, *Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction*, Technical Report G.728, International Telecommunications Union, Geneva, 1992.
48. ITU-T, *Pulse Code Modulation* (*PCM*) *of Voice Frequencies*, Technical Report G.711, International Telecommunications Union, Geneva, 1993.
49. ITU-T, *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction* (*CS-ACELP*), Technical Report G.729, International Telecommunications Union, Geneva, 1996.
50. ITU-T, *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, Technical Report G.723.1, International Telecommunications Union, Geneva, 1996.
51. ITU-T, *Objective Quality Measurement of Telephone-Band* (*300–3400 Hz*) *speech codecs*, Technical Report P.861, International Telecommunications Union, Geneva, 1998.
52. ITU-T, *Packet Based Multimedia Communications Systems*, Technical Report H.323, International Telecommunications Union, Geneva, 1998.
53. K. Jarvinen et al., GSM enhanced full rate speech codec, *Proc. ICASSP*, 1997, pp. 771–774.
54. N. Jayant, J. Johnston, and R. Safranek, Signal compression based on models of human perception, *Proc. IEEE* **81**(10): 1385–1421 (1993).
55. M. Johnson and T. Taniguchi, Low-complexity multi-mode VXC using multi-stage optimization and mode selection, *Proc. ICASSP*, 1991, pp. 221–224.
56. J. P. Campbell Jr., T. E. Tremain, and V. C. Welch, The DOD 4.8 KBPS standard (proposed federal standard 1016), in B. S. Atal, V. C. Cuperman, and A. Gersho, ed., *Advances in Speech Coding*, Kluwer, Dordrecht, The Netherlands, 1991, pp. 121–133.
57. J. P. Campbell, Jr., V. C. Welch, and T. E. Tremain, An expandable error-protected 4800 BPS CELP coder (U.S. federal standard 4800 BPS voice coder), *Proc. ICASSP*, 1989, 735–738.
58. P. Kabal and R. Ramachandran, The computation of line spectral frequencies using chebyshev polynomials, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**: 1419–1426 (1986).
59. W. Kleijn, Speech coding below 4 kb/s using waveform interpolation, *Proc. GLOBECOM* 1991, Vol. 3, pp. 1879–1883.
60. W. Kleijn and W. Granzow, Methods for waveform interpolation in speech coding, *Digital Signal Process.* **1**(4): 215–230 (1991).
61. W. Kleijn and J. Haagen, A speech coder based on decomposition of characteristic waveforms, *Proc. ICASSP*, 1995, pp. 508–511.
62. W. Kleijn, Y. Shoham, D. Sen, and R. Hagen, A low-complexity waveform interpolation coder, *Proc. ICASSP*, 1996, pp. 212–215.
63. W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, Improved speech quality and efficient vector quantization in SELP, *Proc. ICASSP*, 1988, pp. 155–158.
64. M. Kohler, A comparison of the new 2400bps MELP federal standard with other standard coders, *Proc. ICASSP*, 1997, pp. 1587–1590.
65. P. Kroon, E. F. Deprettere, and R. J. Sluyter, Regular-pulse excitation: A novel approach to effective and efficient multi-pulse coding of speech, *IEEE Trans. ASSP* **34**: 1054–1063 (1986).
66. W. LeBlanc, B. Bhattacharya, S. Mahmoud, and V. Cuperman, Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4kb/s speech coding, *IEEE Trans. Speech Audio Process.* **1**: 373–385 (1993).
67. D. Lin, New approaches to stochastic coding of speech sources at very low bit rates, in I. T. Young et al., ed., *Signal Processing III: Theories and Applications*, Elsevier, Amsterdam, 1986, pp. 445–447.
68. A. McCree and J. C. De Martin, A 1.7 kb/s MELP coder with improved analysis and quantization, *Proc. ICASSP*, 1998, Vol. 2, pp. 593–596.
69. A. McCree et al., A 2.4 kbps MELP coder candidate for the new U.S. Federal standard, *Proc. ICASSP*, 1996, Vol. 1, pp. 200–203.
70. A. V. McCree and T. P. Barnwell, III, A mixed excitation LPC vocoder model for low bit rate speech coding, *IEEE Trans. Speech Audio Process.* **3**(4): 242–250 (1995).
71. B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, San Diego, (1997).
72. P. Noll, MPEG digital audio coding, *IEEE Signal Process. Mag.* **14**(5): 59–81 (1997).
73. B. Novorita, Incorporation of temporal masking effects into bark spectral distortion measure, *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 665–668.
74. E. Paksoy, W-Y. Chan, and A. Gersho, Vector quantization of speech LSF parameters with generalized product codes, *Proc. ICASSP*, 1992, pp. 33–36.
75. E. Paksoy et al., An adaptive multi-rate speech coder for digital cellular telephony, *Proc. of ICASSP*, 1999, Vol. 1, pp. 193–196.

76. K. K. Paliwal and B. S. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, *IEEE Trans. Speech Audio Process.* **1**: 3–14 (1993).

77. L. Rabiner and B.-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffis, NJ, 1993.

78. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

79. R. P. Ramachandran and P. Kabal, Stability and performance analysis of pitch filters in speech coders, *IEEE Trans. ASSP* **35**(7): 937–946 (1987).

80. V. Ramamoorthy and N. S. Jayant, Enhancement of ADPCM speech by adaptive post-filtering, *AT&T Bell Labs. Tech. J.* **63**(8): 1465–1475 (1984).

81. A. Rix, J. Beerends, M. Hollier, and A. Hekstra, PESQ — the new ITU standard for end-to-end speech quality assessment, *AES 109th Convention*, Los Angeles, CA, Sept. 2000.

82. R. C. Rose and T. P. Barnwell, III, The self-excited vocoder — an alternate approach to toll quality at 4800 bps, *Proc. ICASSP*, 1986, pp. 453–456.

83. N. Rydbeck and C. E. Sundberg, Analysis of digital errors in non-linear PCM systems, *IEEE Trans. Commun.* **COM-24**: 59–65 (1976).

84. R. Salami et al., Design and description of CS-ACELP: A toll quality 8 kb/s speech coder, *IEEE Trans. Speech Audio Process.* **6**(2): 116–130 (1998).

85. M. R. Schroeder and B. S. Atal, Code-excited linear prediction (CELP): High-quality speech at very low bit rates, *Proc. ICASSP*, 1985, pp. 937–940.

86. Y. Shoham, Very low complexity interpolative speech coding at 1.2 to 2.4 kbp, *Proc. ICASSP*, 1997, pp. 1599–1602.

87. S. Singhal and B. S. Atal, Improving performance of multi-pulse LPC coders at low bit rates, *Proc. ICASSP*, 1984, pp. 1.3.1–1.3.4.

88. D. Sinha and C.-E. Sundberg, Unequal error protection methods for perceptual audio coders, *Proc. ICASSP*, 1999, Vol. 5, pp. 2423–2426.

89. F. Soong and B.-H. Juang, Line spectral pair (LSP) and speech data compression, *Proc. ICASSP*, 1984, pp. 1.10.1–1.10.4.

90. J. Stachurski, A. McCree, and V. Viswanathan, High quality MELP coding at bit rates around 4 kb/s, *Proc. ICASSP*, 1999, Vol. 1, pp. 485–488.

91. N. Sugamura and F. Itakura, Speech data compression by LSP speech analysis-synthesis technique, *Trans. IECE* **J64-A**(8): 599–606 (1981) (in Japanese).

92. L. Supplee, R. Cohn, and J. Collura, MELP: The new federal standard at 2400 bps, *Proc. ICASSP*, 1997, pp. 1591–1594.

93. B. Tang, A. Shen, A. Alwan, and G. Pottie, A perceptually-based embedded subband speech coder, *IEEE Trans. Speech Audio Process.* **5**(2): 131–140 (March 1997).

94. T. Taniguchi, ADPCM with a multiquantizer for speech coding, *IEEE J. Select. Areas Commun.* **6**(2): 410–424 (1988).

95. T. Taniguchi, F. Amano, and S. Unagami, Combined source and channel coding based on multimode coding, *Proc. ICASSP*, 1990, pp. 477–480.

96. J. Tardelli and E. Kreamer, Vocoder intelligibility and quality test methods, *Proc. ICASSP*, 1996, pp. 1145–1148.

97. I. M. Trancoso and B. S. Atal, Efficient procedures for finding the optimum innovation in stochastic coders, *Proc. ICASSP*, 1986, pp. 2379–2382.

98. A. Uvliden, S. Bruhn, and R. Hagen, Adaptive multi-rate. A speech service adapted to cellular radio network quality, *Proc. 32nd Asilomar Conf.*, 1998, Vol. 1, pp. 343–347.

99. J. Vainio, H. Mikkola, K. Jarvinen, and P. Haavisto, GSM EFR based multi-rate codec family, *Proc. ICASSP*, 1998, Vol. 1, pp. 141–144.

100. W. D. Voiers, Diagnostic acceptability measure for speech communication systems, *Proc. ICASSP*, 1977, pp. 204–207.

101. W. D. Voiers, Evaluating processed speech using the diagnostic rhyme test, *Speech Technol.* **1**(4): 30–39 (1983).

102. W. D. Voiers, Effects of noise on the discriminability of distinctive features in normal and whispered speech, *J. Acoust. Soc. Am.* **90**: 2327 (1991).

103. S. Wang, A. Sekey, and A. Gersho, An objective measure for predicting subjective quality of speech coders, *IEEE J. Select. Areas Commun.* **10**(5): 819–829 (1992).

104. S. W. Wong, An evaluation of 6.4 kbit/s speech codecs for Inmarsat-M system, *Proc. ICASSP*, 1991, pp. 629–632.

105. W. Yang, M. Benbouchta, and R. Yantorno, Performance of the modified bark spectral distortion measure as an objective speech quality measure, *Proc. ICASSP*, 1998, pp. 541–544.

106. W. Yang and R. Yantorno, Improvement of MBSD by scaling noise masking threshold and correlation analysis with MOS difference instead of MOS, *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 673–676.

107. S. Yeldener, A 4 kbps toll quality harmonic excitation linear predictive speech coder, *Proc. ICASSP*, 1999, pp. 481–484.

SPEECH PERCEPTION

HANNES MÜSCH
GN ReSound Corporation
Redwood City, California

SØREN BUUS
Northeastern University
Boston, Massachusetts

Speech is probably one of the oldest methods of communication. It is produced by an articulatory system constituting the respiratory tract, the vocal cords, the mouth, nasal passages, tongue, and lips. Precisely coordinated actions of all these parts produce a highly complex signal that encodes messages in a very robust manner, which enables the receiver to understand the message even if it is severely degraded by noise or distortion. What exactly makes speech intelligible? A definite answer to this question has yet to come, but a large amount of research has revealed a multitude of factors that are important for speech intelligibility. This article seeks to elucidate some of these factors, especially those that are important for the design of communication systems. A brief description of the speech signal is followed by a brief discussion of some basic properties of speech perception and an overview of some theories that have been proposed to account for them. These theories may be thought to provide a microscopic account of speech perception in

that they typically seek to explain how listeners map the acoustic properties of the speech signal onto an internal linguistic representation of the message. Other theories take a macroscopic approach. They are not concerned with the specific errors that a listener may make. Rather, they attempt to predict the overall percentage of the speech that will be understood when the speech signal is degraded by noise, distortion, reverberation, and other artifacts that typically are introduced by communication systems and room acoustics. Although these theories reveal relatively little about the processes involved in speech perception, a large portion of this article is devoted to them because they are useful for predicting how the intelligibility of speech is affected by the properties of a communication system.

1. THE SPEECH SIGNAL

Speech is usually considered to consist of a string of individual speech sounds — phonemes — that combine into words, phrases, and sentences. The phonemes correspond to the individual consonants and vowels in a word. At a finer level of analysis, the phonemes are composed of a combination of phonetic features, each of which corresponds to specifics of its production and/or the acoustic properties that signal it. The phonetic features fall into three classes: voicing, place of articulation, and manner of articulation. *Voicing* refers to whether the vocal cords vibrate when the phoneme is produced. Voiced phonemes include all vowels and consonants such as /b/ (as in bat), /z/ (as in zip); in contrast, consonants such as /p/ (as in pat) and /s/ (as in sat) are voiceless. *Place of articulation* refers to the position of the primary constriction of the vocal tract. Examples are *bilabial* (e.g., /m/ as in mat), *alveolar* (e.g., /n/ as in net), and *velar* (e.g., /g/ as in get). *Manner of articulation* includes features such as *nasal* (e.g., /m/ and /n/, which are produced by lowering the soft palate to open the passage to the nasal cavity), *stop* (e.g., /b/, /p/, and /d/, which are produced by momentarily blocking the vocal tract such that pressure builds up to produce a brief burst of noise when the closure is released), and *fricative* (e.g., /s/ and /f/, which are produced by turbulent air flow through a constriction of the vocal tract). Whether phonemes, phonetic features, or other units of speech constitute basic units in speech perception remains an open question.

1.1. Temporal Representation

Figure 1 shows the acoustic waveform of a male talker saying the phrase "the happy prince." On the timescale of Fig. 1a, the envelope of the signal is the most prominent feature. Bursts of acoustic energy can be seen clearly. These bursts typically are 100–200 ms apart and often coincide with the syllables of the speech. Note that word boundaries are not clearly marked. There is no pause between "the" and "happy," yet there is a pause-like gap within "happy." The complex envelope of the speech signal encodes considerable information about the message.

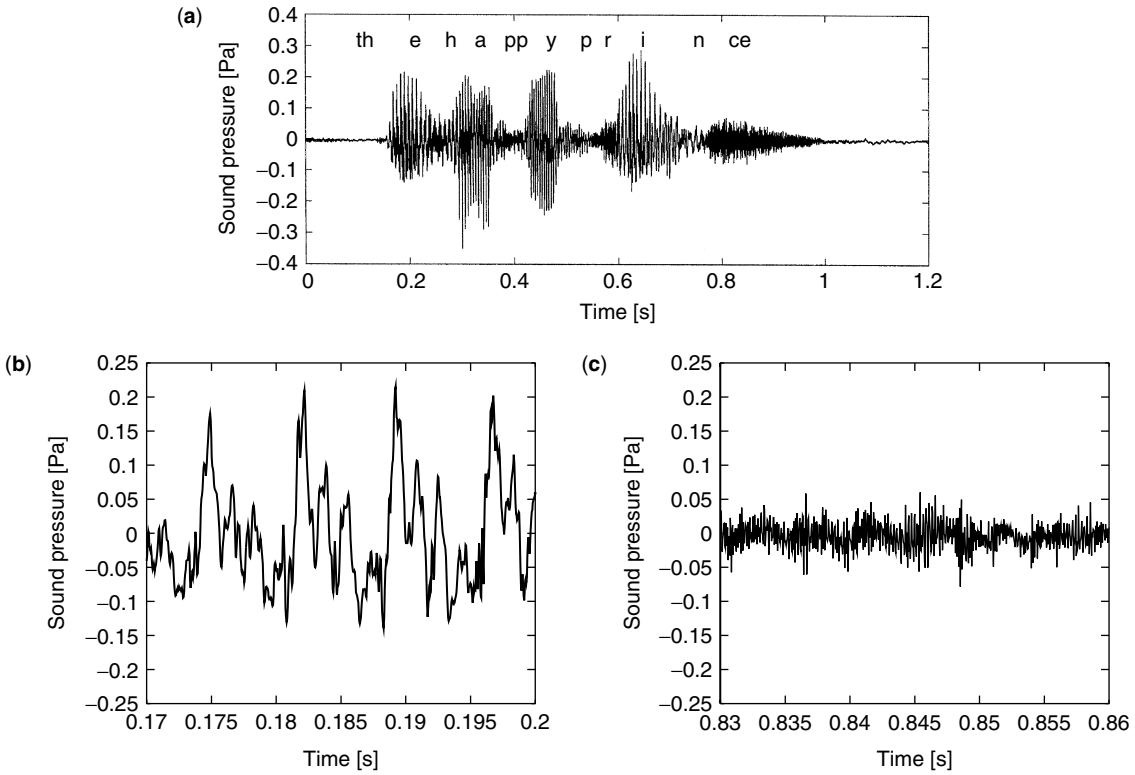


Figure 1. Sound pressure–time function of the phrase "the happy prince," spoken by a male talker. The average speech level is 65 dB SPL. (**a**) The envelope fluctuations are the most prominent feature. (**b**) Fine structure of the vowel /ə/ in "the." (**c**) Fine structure of the fricative /s/ in "prince."

Enlarging part of the graph reveals the fine structure of the acoustic wave. The fine structure also contains much information about the message. The panel in Fig. 1b shows the trace of the sound pressure during the vowel /ə/ in "the." The sound is not strictly periodic, but there is a high degree of regularity in the signal, which gives rise to a well-defined pitch. The repetition rate of the pattern is called the *pitch period*. The pitch helps convey meaning by prosody (e.g., the pitch rises at the end of a question) and specifying the identity of the speaker, but it is not important for the identity of the vowel.

In contrast, the hissing sound /s/ at the end of "prince" has a noise-like fine structure. Because it is a voiceless consonant, the vocal cords do not vibrate during its production. Rather, the sound is generated by the turbulence that results when air is pressed through a constriction in the vocal tract. The acoustic energy of voiceless consonants typically is much lower than that of vowels.

The speech envelope, which is so obvious in Fig. 1, is obliterated when a masking noise is added to the speech. At negative signal-to-noise ratios (SNRs), the envelope of the speech-and-noise composite is almost flat and the time course of the signal looks almost like that of a noise. Nevertheless, most people have no problems communicating at SNRs somewhat below 0 dB. (As will become evident later, the SNR necessary to just understand speech depends on the spectral shape and temporal characteristics of the noise.)

1.2. Spectral Representation

Many aspects of the speech signal are best described in the frequency domain. Spectra of vowels, for example, have prominent peaks, which represent the resonances of the acoustic filter formed by the vocal tract. They are called *formants*, and their interrelationships are important in defining the identity of vowels and some features of consonants. Thus, short-term spectra yield a highly useful characterization of speech. They, too, encode much information about the message contained in the speech and are used as a foundation for most automatic speech-recognition schemes. The changing spectral content of the speech signal is often displayed in a *spectrogram*, which is a power spectrum–time plot. Figure 2 shows the spectrogram of the utterance whose time trace is shown in Fig. 1. The timescale on the abscissa is the same in both figures. Frequency is shown on the ordinate. The signal power at any frequency and point in time is represented by the darkness of the plot. Dark areas correspond to high power and light areas correspond to low power.

The energy fluctuations that were seen in Fig. 1 are also apparent in this presentation. The high-energy periods produced by vocal-cord vibration are easily seen as dark patches, which are separated by light low-energy periods that occur during periods of consonantal noise and during speech pauses. For most speech sounds, the energy is concentrated in the frequency region below 5000 Hz. An exception is the /s/ at the end of "prince," which has most of its energy at high frequencies.

The formants are identified by the dark traces below 3000 Hz. They are visible primarily during the voiced

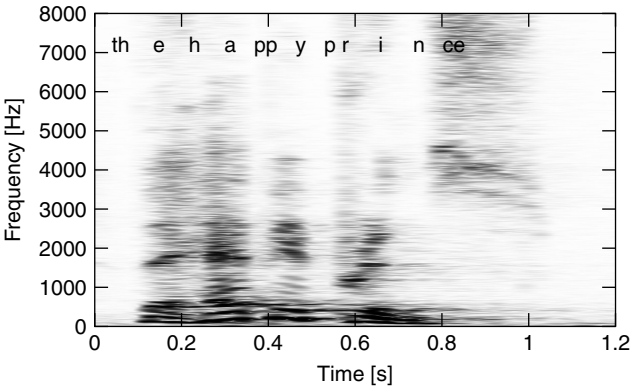


Figure 2. Spectrogram of the signal shown in Fig. 1. Time is plotted along the abscissa and frequency along the ordinate. The instantaneous signal power is shown by the darkness of the plot. Dark areas represent high power and light areas represent low power.

excitation of the vowels. The formants of a steady-state vowel have constant frequencies, which produce horizontal formant trajectories. The formant frequencies reflect the shape of the vocal tract, which the talker adjusts to produce the desired vowel. Ideally, it should attain a specific shape to produce a given vowel. During natural speech, however, the articulators must move quite rapidly to produce the desired sequence of phonemes and they never reach the ideal positions. Rather, they constantly move from one (unreached) target to the next. As a result, the formants of natural speech show a pattern of rising and falling trajectories that reflect the articulator movement. Because they reflect the movement from one vocal tract shape to the next, the formant trajectories encode not only the vowel identity but also the identity of the preceding and succeeding consonant — at least to some degree. This interaction is called *coarticulation*. When a masking noise masks the low-energy consonants but not the high-energy vowels, a large number of the consonants can still be identified correctly. This is possible because a part of the consonant is encoded in the format transitions that result from coarticulation. Coarticulation also is at the heart of one of the most vexing problems in speech recognition by humans or machines: the lack of a one-to-one correspondence between the acoustic properties of the speech signal and the speech units [1]; (see, however, Ref. 2). For example, the acoustic properties of a speech segment that encodes the phoneme /b/ vary greatly depending on the context (i.e., other phonemes) that surrounds it, as well as on the speaker and speaking rate.

2. SPEECH WITH REDUCED ACOUSTIC INFORMATION

The speech signal encodes the message in the envelope, in the fine structure, and in the short-term spectrum to varying degrees. This produces a redundancy that allows speech to retain a high degree of intelligibility, even if some of these aspects are altered or removed by noise, distortion, or signal processing. For example, Shannon et al. [3] modulated bands of noise with the envelope of the speech signal in these bands. As a result of this

operation, the temporal fine structure of the signal was lost completely and the spectral resolution was greatly reduced, but intelligibility remained high. Even when all spectral information was removed — that is, when the broadband envelope of speech was imprinted on a broadband noise — the speech was still recognizable. In this latter condition, the temporal envelope was the only cue available to the listener.

The fact that speech remains intelligible after the entire fine structure has been removed does not mean that only a small fraction of the message is encoded in the fine structure. On the contrary, speech can be highly intelligible when the listeners' only cue is the fine structure. To demonstrate this, Licklider and Pollack [4] passed speech through an infinite peak clipper, so that the processed signal had a fixed positive value whenever the input signal was positive and a fixed negative value whenever the input was negative. This binary code flattens the envelope completely and broadens the spectrum. The processed signal contains information only about the zero crossings of the original signal. Nevertheless, speech processed in this way is intelligible enough to let conversation take place with little difficulty. Even if it sounds very distorted, its quality is high enough to reveal readily the identity of the speaker.

Yet another minimalist representation of speech is sine-wave speech. In this modification, the speech is synthesized by tracing the dominant formants with variable-frequency tone generators. The speech thus synthesized is also highly intelligible, even when only a small number of generators are used [5].

These examples make it clear that speech is very robust. Both spectral and temporal cues carry the message. Speech cannot be defined completely in the spectral domain, nor can it be described only in the temporal domain. The large redundancy allows communication in very unfavorable listening conditions. At the same time, it makes speech a difficult subject to study.

3. SOME BASICS OF SPEECH PERCEPTION

One hallmark of speech perception is that it is categorical. If one constructs a set of speech sounds that gradually changes the acoustic properties from those corresponding to one phoneme (e.g., /d/) to those corresponding to another (e.g., /g/), listeners will report hearing /d/ until the stimuli reach a boundary at which point the perception changes rapidly to become /g/. Discrimination between two members of the same phonemic category usually is difficult, whereas discrimination between members of different phonemic categories is easy [6]. However, listeners can discriminate among stimuli within a single category under ideal conditions [7] and some members of a stimulus set may be shown to be better exemplars than others [8]. Thus, the categorization of speech does not eliminate information about differences between tokens of the same categories. In fact, more recent research indicates that the phonetic categories have a complex internal structure [9], and much current research aims to discover the internal organization of phonetic categories [10,11].

Another hallmark of speech perception is its sensitivity to context. As discussed above, coarticulation makes the acoustic representation of a particular phoneme depend on the context in which it occurs. However, contextual effects are not limited to coarticulation. Variables such as speaking rate and lexical context also may change the perception produced by a given set of acoustical properties. For example, a particular speech sound may be heard as /ba/ if it occurs in a context of slow speech, but as /wa/ if it occurs in a context of rapid speech [12]. Complete theories of speech perception ultimately must explain such context effects, as well as how listeners deal with the lack of invariance between phonemes (or features) and their acoustic representations.

4. THEORIES OF SPEECH PERCEPTION

Various theories have been proposed to account for speech perception at differing levels of detail. Some take a macroscopic approach and seek to predict how the overall percentage of recognition depends on the long-term average properties of speech, distortion, and background noise. This class of theories encompasses the classic articulation index (AI) and derivative theories such as the speech intelligibility index (SII) and speech transmission index (STI). These theories do not attempt to predict how recognition errors depend on the details of the speech signal, but aim to provide a tool that can predict the intelligibility of speech transmitted through a communication system. Because these theories have proved to be very useful engineering tools, they are discussed at some length below. Other theories take a microscopic approach and seek to explain how the detailed properties of the speech signal determine its mapping into linguistic units. This class of theories encompasses a number of very different approaches to various parts of this difficult problem. Some seek to explain speech perception as a result of the general properties of signal processing by the auditory system, sometimes combined with cognitive processing. Others hold that speech recognition is mediated by mapping the acoustic properties of the speech to the processes necessary to produce it. They imply that the recognition of speech is specific to humans. Whether extraction of the phonetic units occurs directly by auditory processes or through reference to articulatory processes, subsequent processing also is important. Thus, other models seek to incorporate the interplay between lexical processes and lower-level phonetic processes.

Auditory theories of speech recognition hold that the auditory processes responsible for perception of any sound generally define the relation between acoustic properties and the categorical perception of phonemes [13], although subsequent lexical, syntactical, and grammatical processing may modify the perception. Specific theories within this class range from auditory processes being responsible for mapping the acoustic properties into features that feed into linguistic processes [14] to ones that claim that a detailed analysis of auditory processing can explain how the categorical boundaries between phonemes depend on context and provide the invariance that appears to be lacking in the acoustical properties of the speech [2]. Holt and

co-workers have suggested that many context effects can be explained as a spectral contrast enhancement mediated, in part, by neural adaptation that occurs throughout the auditory pathway [15,16].

Theories that rely on listeners' knowledge of speech production to explain speech perception hold that listeners achieve a mapping between the context-dependent acoustic properties of speech and the corresponding phonemes because they infer how the speech was produced. The most prominent of these theories is the motor theory of speech perception [17]. This theory holds that speech perception is accomplished, at least in part, by a specialized phonetic processor that maps the acoustic signal into the articulatory gestures that underlie the production of it.

Theories on the processing that occurs subsequent to the extraction of phonetic units often hold that speech perception reflects an intimate interaction between low-level auditory and phonetic processes and higher-level lexical processes (for a review, see the article by Protopapas [18]). Some employ *top–down processing*, in which excitatory and inhibitory connections let nodes corresponding to words modify the response properties of nodes corresponding to individual phonemes [19]. Others employ *bottom–up processing*, in which lower-level nodes of a network activate higher-level nodes in a manner that allows sequences of short-term spectra to map into words [20].

5. THE ARTICULATION INDEX

Although speech intelligibility is determined by many factors, two parameters appear to be of paramount importance: the *bandwidth* and the *audibility* of the signal. Both parameters are affected by the transfer function of a transmission system, and it is important for the communication engineer to understand how they affect intelligibility. A group of models, collectively known as *articulation index* (AI) *models*, attempt to predict intelligibility from the transfer function of a transmission system. A team led by Harvey Fletcher at Bell Labs laid the groundwork for the development of the Articulation Index in the early 1900s (see Allen [21] for a historical perspective). The complete model was published in 1950 by Fletcher and Galt [22], but is rarely used because it is very complex [23]. A simpler version was published earlier [24]. On the basis of French and Steinberg's model, Kryter [25] devised a model that was easy to use. With minor modifications, Kryter's model was later accepted as ANSI standard S3.5 [26]. The newest member of the group of Articulation-Index models is the *speech intelligibility index* (SII) [27].

Articulation index models are macroscopic models of speech recognition. They predict the *average* intelligibility achievable with a linear or nearly linear transmission system, where the average is taken across many talkers and speech materials. The model predictions are based on the statistics of the speech signal. Therefore, AI models do not predict the intelligibility of a short speech segment, nor do they account for the intelligibility of particular phonemes and the patterns of confusions among phonemes that occur in the perception of partly intelligible speech [28].

The basic assumption underlying AI theory is that speech intelligibility is determined by the audibility of the speech signal and that different frequency ranges make unequal contributions to the intelligibility. The AI model provides a method to calculate an articulation index. The AI is highly correlated with speech intelligibility and is a descriptor of the intelligibility that can be achieved with the transmission system for which it was calculated. If two people, one meter apart and facing each other, converse in a quiet room that is free of reverberation, intelligibility can reasonably be expected to be optimal. For such a condition, the AI is unity. In this optimal listening condition, all relevant speech cues are completely audible. If, on the other hand, the listener cannot hear the talker at all, there will be no intelligibility. For such a condition, the AI is zero. In general, the AI is a scalar between zero and unity.

Every listening condition has associated with it a single value of AI. Different speech materials yield the same AI — as long as they are equally audible. The AI does not reflect the fact that the intelligibility differs among syllables, words, and various types of sentences presented through a given transmission system. Such differences occur mainly because these materials differ in terms of how much information they provide by context. To account for the effects of context on the speech test score, AI models use different transformation functions between the AI and the predicted test score. These transformations are independent of the speech transmission system and specific to the test material and the scoring procedure. Typical transformations are shown in Fig. 8 (later in this article). Although equivalent to the AI in many respects, the SII (the result of the speech intelligibility index model [27]) differs from the AI by assuming that the SII is not exclusively determined by audibility but to a small degree also by the type of the speech material used [29].

The AI is calculated as an importance-weighted sum of a quantity W_i that is related to the audibility of the speech signal in a set of spectral bands that span the range of audiofrequencies:

$$\mathrm{AI} = \sum_i I_i W_i \tag{1a}$$

where I_i is the importance weight of the ith band. The product of I_i and W_i is the AI in the ith band, AI_i. Accordingly, Eq. (1a) can be rewritten as

$$\mathrm{AI} = \sum_i \mathrm{AI}_i \tag{1b}$$

5.1. Audibility

The quantity W_i is a fraction less than or equal to unity that indicates how many of the cues carried in the ith band are actually available to the listener. It is related to the audibility of the speech signal. Audibility is expressed in terms of the *sensation level*, which is the number of decibels that a sound is above its threshold. The sensation level of a band of speech can be calculated as the difference between the critical-band level of the speech and the level of a just-audible tone centered in that band. The critical band is a measure of the bandwidth of frequency

analysis in the auditory system. It can be defined as "that bandwidth at which subjective responses [to sound] rather abruptly change" [30] and represents a frequency range across which the acoustic energy is integrated by the auditory system. The critical band is about 100 Hz wide for center frequencies below 500 Hz and 15–20% of the center frequency at higher frequencies [30–32]. [It should be noted that modern measurements of auditory filter characteristics yield an *effective rectangular bandwidth* that is somewhat narrower than the critical bandwidth, especially at low frequencies [33].]

In quiet, the level of a just-audible tone depends only on the listener's sensitivity to sound. The lowest level at which the tone can be heard is called the *absolute threshold*. Threshold levels vary with frequency and among listeners, but established norms exist for normal threshold levels [34,35]. In individuals with hearing loss, thresholds are elevated — that is, the tone levels at threshold are higher than in normal-hearing persons. Therefore, audibility is decreased in listeners with hearing loss. In agreement with the reduced intelligibility of unamplified speech for listeners with hearing loss, the elevated thresholds produce lower AI values.

Interfering sounds (i.e., maskers) also can reduce the audibility of speech. If the masker is above threshold, it may produce masking and make a tone at absolute threshold inaudible. The level of the test tone that can just be detected in the presence of the masker is called the *masked threshold*. The *amount of masking*, which is the difference between the masked threshold and the absolute threshold, depends on the level and spectral shape of the masker and the frequency of the tone (for review, see Buus' article [33]). For typical maskers with spectra that change relatively slowly with frequency, the amount of masking is determined primarily by the sound level of the masker measured within a critical band centered on the tone. For such maskers, masking is linear. Whenever the masker level is increased by, for example, 10 dB, the masked threshold increases by 10 dB. For listening situations with masking sounds, the sensation level of the speech is determined as the difference between the critical-band level of the speech and the masked or absolute threshold for tones at the center of the band, whichever is higher.

The relation between band sensation level and W_i differs somewhat across AI models [36]. In their simplest form, AI models assume a linear relation between sensation level and W_i [25–27] over a 30-dB range of sensation levels (see Fig. 3). Within this range W_i increases as the sensation level increases. In the models of French and Steinberg [24] and Fletcher and Galt [22], the relation is nonlinear and the range of sensation levels over which W_i changes is larger than 30 dB.

Because W_i and band AI_i are proportional [see Eq. (1)], the AI_i also increases with sensation level. The level dependence of the AI can be understood by noting that speech is an amplitude-modulated signal. French and Steinberg [24] analyzed data of Dunn and White [37], who measured the level distribution of 125-ms speech segments in a number of frequency bands. The duration of these segments corresponds roughly to the average duration of syllables. When plotting the percentage of

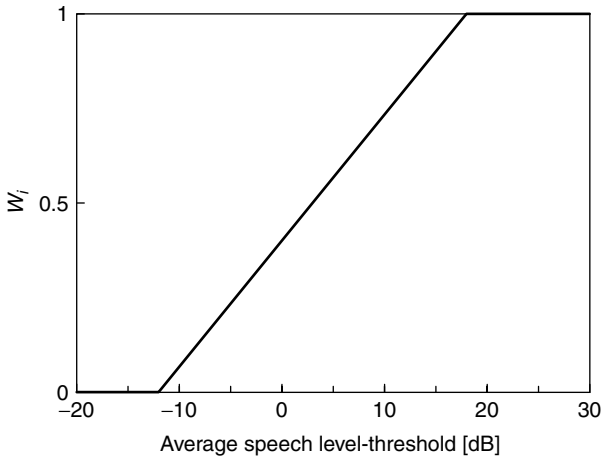


Figure 3. Relation between the audibility-related quantity W_i and the number of decibels that the average speech level is above threshold [25,26]. This function reflects the distribution of the short-term speech level in 125-ms intervals. If the long-term average speech level is 12 dB below the listener's threshold, speech peaks will exceed the listener's threshold and W_i becomes larger than zero.

intervals that exceed a certain level as a function of the difference between that level and the long-term average speech level, French and Steinberg found a relationship that is well described by a linear function. This relationship, together with the masked or absolute threshold, allows one to derive the percentage of syllables that are intense enough to be heard by the listener at any given speech level. The difference between speech level and threshold level is a measure of the proportion of syllables that the listener hears. When this proportion increases, W_i — and the AI — in the band also increase. According to Kryter [25], approximately 1% of the 125-ms intervals of speech is audible when the long-term average speech level is 12 dB below detection threshold. At this point W_i is just above zero. From there, W_i increases linearly with level as a larger proportion of the syllables becomes audible. Once the long-term average speech level is 18 dB above threshold, even the weakest percentile of 125-ms intervals exceeds threshold. At this point, W_i is unity and does not increase further with additional level increases.

5.2. Independence Assumption

As already stated, the AI and the intelligibility score are related by a nonlinear transformation function that is different for different test materials. With one exception, these transformations are derived empirically. Only the transformation between AI and the proportion of phonemes correctly identified in a nonsense-syllable recognition task, also known as the *articulation*, s, is determined by the structure of the AI model. Articulation and AI are related by

$$\mathrm{AI} = -k \cdot \log_{10}(1 - s) \tag{2}$$

where k is a scale factor that is chosen such that AI is unity in the most favorable listening condition. Equation (2) results from the fundamental assumption that speech

bands contribute independently to articulation. This assumption underlies most AI models (see, however, Ref. 38). Combining Eqs. (1) and (2) yields

$$\log_{10}(1-s) = \sum_i \log_{10}(1-s_i) \tag{3}$$

which is equivalent to

$$(1-s) = \prod_i (1-s_i) \tag{4}$$

where s is the articulation in the broadband condition and s_i is the articulation achieved when listening to the individual bands. The terms $(1-s)$ and $(1-s_i)$ are the associated error probabilities. Equation (4) states that the probability of identifying a phoneme in a nonsense syllable incorrectly is equal to the product of the error probabilities when decisions are based on the information available in the individual bands. Whenever a joint probability equals the product of the conditional probabilities, the events are said to be independent. Therefore, Eqs. (3) and (4) reflect the AI model assumption that speech bands contribute independently to the intelligibility of phonemes in nonsense syllables.

5.3. Band Importance

The importance of various frequency bands for speech understanding varies with frequency. This can be seen in Fig. 4, which shows the proportion of correctly understood phonemes in highpass- and lowpass-filtered nonsense syllables as a function of the filters' cutoff frequencies. The speech is at a level that ensures that audibility in the passband is always unity. For the purpose of speech intelligibility predictions, a lowpass filter with a cutoff frequency of 8000 Hz is equivalent to an allpass, because the speech signal carries very little energy in the bands above 8000 Hz. By definition, the AI of a system with a flat transfer function that provides optimal gain is unity. The data in Fig. 4 show that articulation is almost perfect ($s = 0.985$) for this ideal transmission system. Likewise, when the cutoff frequency of a highpass filter is 100 Hz, the same high performance is observed because the acoustic information in bands below 100 Hz is irrelevant for speech intelligibility.

Performance decreases when the low-frequency components of the speech are gradually removed by increasing the cutoff frequency of the highpass filter. Initially, the performance decrease is only slight. Raising the cutoff frequency from 100 to 1000 Hz affects articulation only marginally. This suggests that the speech bands between 100 and 1000 Hz make only small contributions to intelligibility. When the cutoff frequency is raised further, articulation decreases sharply. Removing the frequency range between 1000 and 4000 Hz causes articulation to drop from almost perfect to just above 0.5. Apparently, this frequency band carries a significant portion of the speech cues.

Similarly, performance decreases when the high-frequency speech components are removed by decreasing the cutoff frequency of the lowpass filter. Again, performance declines very gradually at first, but the decrease becomes quite rapid once the cutoff frequency is below 5000 Hz. The frequency at which the two curves intersect divides the speech spectrum into two equally important bands. This frequency, called the *crossover frequency*, is approximately 1700 Hz. Because the AI is unity for a broadband condition in which the speech is clearly audible across the entire frequency range, it follows that the AI for each of the two equally important bands must be 0.5.

Equation (1a) states that the AI is an importance-weighted sum of W_i, which is related to the audibility in the bands. The *importance density function* describes how the inherent potential for intelligibility is distributed across frequency. It can be derived from the relation between the filter cutoff frequency and the associated

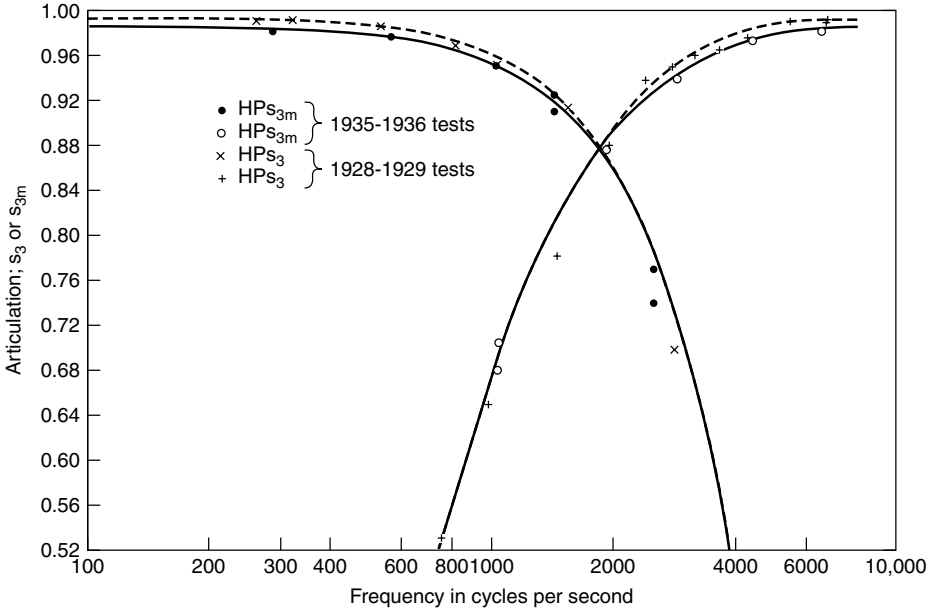


Figure 4. Crossover functions. The proportion of correctly identified phonemes in a nonsense-syllable context is plotted as a function of the cutoff frequencies of ideal highpass and lowpass filters. Filled dots and crosses represent highpass filters; open dots and plusses, lowpass filters. (Reprinted from Fletcher and Galt [22], with permission of the publisher.)

articulation. Using Eq. (2), the articulation scores are transformed into the corresponding AI values. This results in a relation between filter cutoff frequency and AI. Thus with the help of Eq. (2), the data in Fig. 4 can be transformed into a cumulative importance function. Simple differentiation yields the importance-density function. The band-importance weights, I_i, are derived from the importance-density function by integrating over the width of the band. Figure 5 shows the importance weights for critical-band wide bands of nonsense syllables [27].

The various AI models differ in the details of the calculation [39]. For example, different models use different importance functions. The exact shape of the importance function seems to depend on the speech material used [29,40]. Accordingly, the SII specifies several speech-material-specific importance functions. Differences also exist in the degree of sophistication with which the sensation level of the speech is predicted from parameters of the transmission system and the spectrum of the masking noise. The representation of the speech spectrum and the masker spectrum in the inner ear is "smeared" in both the spectral and temporal domains. Not only can the masker mask the speech, but the speech signal also acts as a masker upon itself. The amounts of noise-masking and self-masking depend on the speech level and on the transfer function of the transmission system and can be predicted quite accurately. The model by Fletcher and Galt [22] and the SII [27] exhibit the most sophistication in modeling both forms of masking. The simple AI of ANSI [26] does not model self-masking of speech and its prediction accuracy is compromised if the transmission system filters speech so that high-level portions of the spectrum are adjacent to low-level portions.

The AI model discussed so far predicts that when speech exceeds an optimal level, further level increases do not result in increased intelligibility. However, intelligibility may decrease at excessive sound levels [22,41]. This effect is known as *rollover*. The AI by Fletcher and Galt [22] and the SII [27] model this effect.

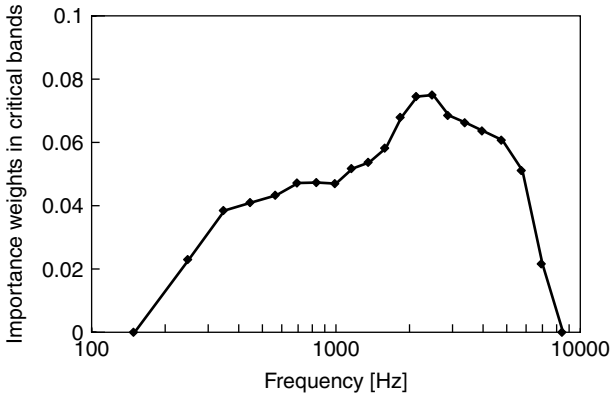


Figure 5. The importance function for nonsense syllables as defined by the speech intelligibility index (SII). Every data point represents the integral of the importance density function over a frequency range equal to one critical bandwidth centered at the frequency of the symbol. This importance function is in good agreement with the data in Fig. 4.

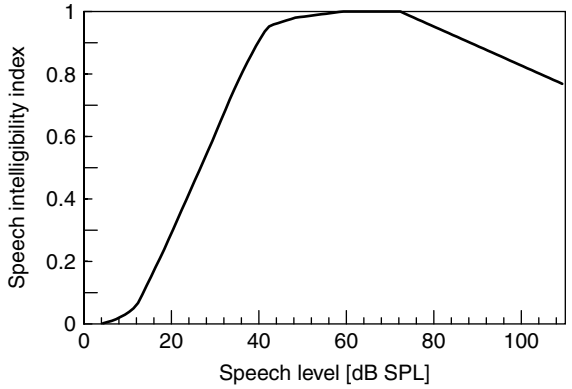


Figure 6. Speech intelligibility index (SII) of undistorted speech as a function of speech level. The predictions are for speech reception in quiet by normal-hearing listeners.

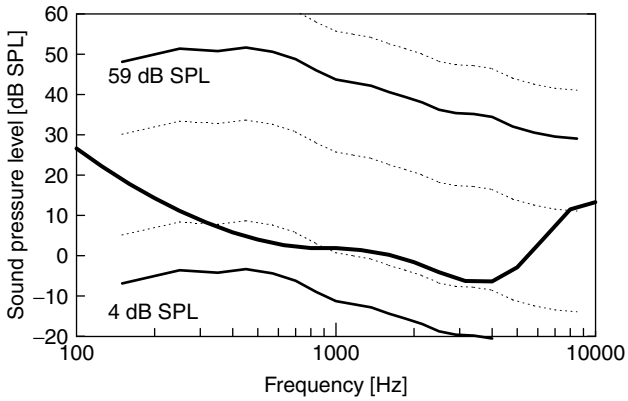


Figure 7. The sound pressure level (SPL) at the detection threshold (bold line, ISO 226) is plotted together with the long-term average level of speech in critical bands (solid lines labeled with the overall speech level). The dashed lines indicate the dynamic range of the speech (from 12 dB above to 18 dB below the long-term average).

Figure 6 shows the SII of unfiltered speech as a function of the overall speech level. At very low speech levels, the signal is inaudible and the SII is zero. As the speech level increases, spectral components in the critical band centered at 500 Hz are the first to exceed threshold. This can be seen from Fig. 7, which shows the long-term average critical-band level of speech with an overall level of 4 dB SPL (lower solid line) in relation to the threshold of normally hearing listeners (bold line [35]). The dashed line 12 dB above the long-term average speech level marks Kryter's [25] estimate of the top of the dynamic range of the speech signal.

The rate at which the SII increases with speech level depends on the sum of the importance weights in the bands whose audibility is affected by the level change. At low levels, the SII increases only slowly with level because only the bands near 500 Hz contribute. At higher levels, all bands are audible and a 3-dB change in speech level changes the SII by 0.1. This rate is equal to the rate at which W_i increases with sensation level (see Fig. 3) and is the steepest slope possible.

Once the speech level reaches 59 dB SPL, all speech bands are completely audible. Figure 6 shows that at this level the SII is unity, indicating that intelligibility is optimal. This is also seen in Fig. 7, where the critical-band level of 59-dB-SPL speech is shown by the upper solid line. The dashed line 18 dB below represents Kryter's [25] estimate of the lower bound of the dynamic range. In frequency bands that carry speech cues, this lower bound is above threshold, indicating that all speech cues are available to the listener. Further increasing the speech level does not result in increased intelligibility. On the contrary, when the speech level increases above 68 dB SPL, the SII decreases.

5.4. Transformation Between Articulation Index and Intelligibility

The articulation index describes the extent to which the sensation generated by the stimulus itself affects intelligibility. However, speech recognition depends not only on the sensory input but also on the context in which it occurs. Semantic and syntactic context, lexical constraints, and the size of the test set strongly affect intelligibility. None of these factors are reflected in the AI (some are reflected in the SII, however). Instead, they are accounted for by a set of transformation functions between AI and the predicted intelligibility score. Because these transformations depend strongly on the contextual constraints in the speech material and on the scoring procedure, their form varies significantly. One set of transformation functions that can serve as a reference is reproduced in Fig. 8. However, many users of AI models prefer to derive transformation functions specifically for their particular speech materials. In fact, the SII [27] states that users must define the appropriate transformations for the particular speech material and scoring method being used. Whatever transformation is used, those shown in Fig. 8 are reasonably representative. They indicate that for highly redundant materials, such as everyday speech, an AI of about 0.2 is needed to obtain a marginal intelligibility of 75%.

It has been shown that words in meaningful sentences are more easily understood than words in isolation or words preceded by a carrier phrase. Lexical constraints cause the intelligibility of phonemes in meaningful words to be higher than the intelligibility of phonemes in nonsense syllables. In general, intelligibility increases as the predictability of the speech sounds increases.

By applying optimal decision theory to speech recognition, Pollack [42] has shown that the increased recognition performance for items with high *a priori* probability of occurrence cannot be explained by more successful guessing. When he calculated the receiver operating characteristics (ROCs) that describe the speech-recognition performance for speech tokens with different *a priori* probabilities of occurrence, he found that the observers were more sensitive to speech tokens with a high *a priori* probability of occurrence than to those with a low *a priori* probability. The better performance for items with high priori probability was not the result of a response

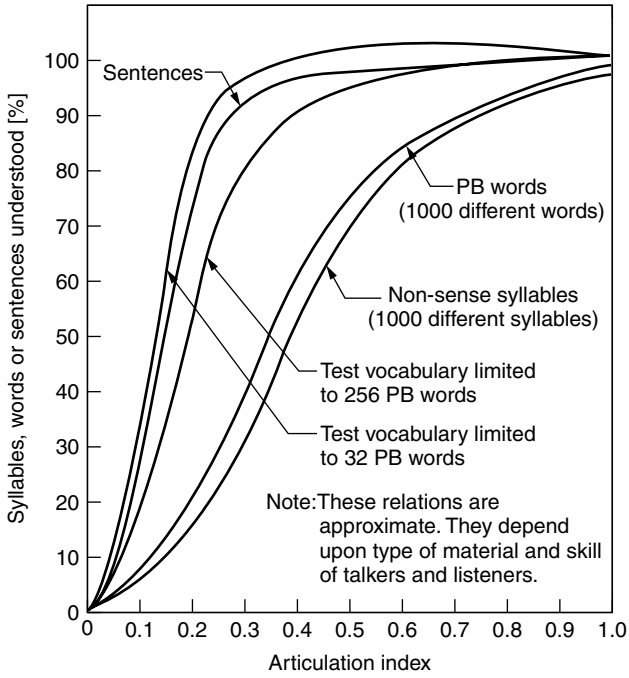


Figure 8. Transformation functions between the AI and the predicted intelligibility of various speech materials. The more contextual constraints are placed on the signal, the higher the intelligibility. (Reprinted from Kryter [25], with permission of the publisher.)

bias. Rather, it appears as if context provides an additional source of information that increases the listener's sensitivity to speech recognition. It has been shown that a signal-detection model may account for such effects by changing the amount of variance contributed to the decision variable by a single, central noise source [43].

An entirely different method of accounting for context effects assumes that contextual information is statistically independent from the sensory information used to recognize the target without context [44]. Expanding on the idea of multiplicative error probabilities and assuming that the context is extracted from speech presented in the same listening condition as the sensory input, this model derives a simple relation between the probability, p_c, of correctly recognizing items in the presence of context and the probability, p_i, of recognizing the same items without context:

$$p_c = 1 - (1 - p_i)^a \tag{7}$$

When this equation was applied to several sets of data, a was nearly constant over a wide range of values of p_i, lending support to the assumption of an independent contribution of the context [44].

5.5. The Speech Transmission Index

The AI model definition of audibility implicitly assumes that the temporal structure of the speech signal is undisturbed and that the short-term level of the masking noise is unrelated to the short-term level of the speech signal. AI models break down when these requirements are not met. To circumvent this

problem, the speech transmission index (STI) [45] converts temporal distortions of the speech signal into an equivalent reduction in the audibility of the speech. This approach extends the AI concept to include temporal distortions such as reverberation and nonlinearities (e.g., a level-compression circuit), which can reduce the modulation depth of an amplitude-modulated signal such as speech. The reduction of modulation depth caused by temporal distortion usually varies across modulation frequencies. For example, reverberation tends to reduce fast modulations more than slow modulations. Therefore, the STI model determines the effective SNR in every frequency band for a set of modulation frequencies ranging from 0.63 to 12.5 Hz, which are typical for speech. Any reduction in modulation depth is transformed into an equivalent noise level, which yields a transmission index for each modulation frequency. The average of the transmission indices at a given spectral frequency band yields an overall modulation transmission index for the band, which is largely equivalent to the AI contribution by the band. Thus, the STI is very similar to the AI models, but its reliance on modulation transfer allows it to account for situations that the AI models do not handle readily.

6. SUMMARY

Speech is a signal in which a message is encoded. The coding is not yet fully understood, but it is known that it is very robust, because speech remains intelligible even after severe signal distortion. Speech perception is categorical. Theories of speech perception attempt to explain the speech-perception process. Examples of such theories are *auditory theories* [e.g., 2] and the *motor theory of speech perception* [17]. Although the process of speech recognition and comprehension is not yet fully understood, models exist that predict human speech-recognition performance without attempting to explain the speech perception process. These models are statistical in nature and predict only the average intelligibility. Examples of such models are the speech intelligibility index (SII), the speech transmission index (STI), and the articulation index (AI).

BIOGRAPHIES

Hannes Müsch received the Diplom Ingenieur degree in acoustics and electrical engineering in 1993 from the Technische Universität Dresden, Germany, and the MSEE and Ph.D. degrees in electrical engineering from Northeastern University, Boston, Massachusetts, in 1997 and 2000, respectively. Dr. Müsch was a consultant in industrial and building acoustics at Müller-BBM, Germany, and a research scientist at GN ReSound's Core Technology Center in Redwood City, California, where he worked on the development of signal processing algorithms for hearing aids. In 2001, he joined Sound ID, Palo Alto, California, where he is the director of signal processing. Dr. Müsch's research interests are psychoacoustics, models for the prediction of speech recognition performance, and the effects of hearing loss on audition. His work is focused on developing signal processing algorithms to improve the listening experience for the hearing impaired.

Søren Buus is the director of the Communications and Digital Signal Processing Center and professor of electrical and computer engineering at Northeastern University Boston, Massachusetts. He graduated with an M.S. degree in electrical engineering and acoustics from the Technical University of Denmark in 1976 and a Ph.D. in experimental psychology from Northeastern University in 1980. His primary research interests are in basic psychoacoustics. In particular, he has published extensively on intensity coding, detection, loudness, and temporal processes in the auditory system. Before joining the faculty at Northeastern University in 1986, Professor. Buus held research appointments at Northeastern University and Harvard University, Massachusetts. He also has been a guest researcher at the Acoustics and Mechanics Laboratory, CNRS, Marseille, France; Institute of Electroacoustics at the Technical University of Munich, Germany; Department of Environmental Psychology, Faculty of Human Sciences, Osaka University, Japan; and the Acoustics Laboratory, Technical University of Denmark. He is a fellow of the Acoustical Society of America.

BIBLIOGRAPHY

1. J. L. Miller, Speech perception, in M. C. Crocker, ed., *Encyclopedia of Acoustics*, Vol. 4, Wiley, New York, 1997, pp. 1579–1588.
2. K. N. Stevens and S. E. Blumstein, The search for invariant acoustic correlates of phonetic features, in P. D. Eimas and J. L. Miller, eds., *Perspectives on the Study of Speech*, Erlbaum, Hillsdale, NJ, 1981, pp. 1–38.
3. R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, Speech recognition with primarily temporal cues, *Science* **270**: 303–304 (1995).
4. J. C. R. Licklider and I. Pollack, Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech, *J. Acoust. Soc. Am.* **20**: 42–51 (1948).
5. R. E. Remez, P. E Rubin, D. B. Pisoni, and T. D. Carrell, Speech perception without traditional speech cues, *Science* **212**: 947–950 (1981).
6. A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, The discrimination of speech sounds within and across phoneme boundaries, *J. Exp. Psychol.* **54**: 358–368 (1957).
7. A. E. Carney, G. P. Widin, and N. F. Viemeister, Noncategorical perception of stop consonants differing in VOT, *J. Acoust. Soc. Am.* **62**: 961–970 (1977).
8. A. G. Samuel, Phonemic restoration: Insights from a new methodology, *J. Exp. Psychol.: Gen.* **110**: 474–494 (1982).
9. P. K. Kuhl, Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not, *Percept. Psychophysiol.* **50**: 93–107 (1991).
10. J. L. Miller, On the internal structure of phonetic categories: A progress report, *Cognition* **50**: 271–285 (1994).
11. P. Iverson and P. K. Kuhl, Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from

a common mechanism? *Percept. Psychophysiol.* **62**: 874–876 (2000).

12. J. L. Miller and A. M. Liberman, Some effects of later-occurring information on the perception of stop consonant and semivowel, *Percept. Psychophysiol.* **25**: 457–465 (1979).
13. J. Coleman, Cognitive reality and the phonological lexicon: A review, *J. Neuroling.* **11**: 295–320 (1998).
14. D. B. Pisoni, Auditory and phonetic memory codes in the discrimination of consonants and vowels, *Percept. Psychophysiol.* **13**: 253–260 (1973).
15. L. L. Holt, A. J. Lotto, and K. R. Kluender, Neighboring spectral content influences vowel identification, *J. Acoust. Soc. Am.* **108**: 710–722 (2000).
16. L. L. Holt and K. R. Kluender, General auditory processes contribute to perceptual accommodation of coarticulation, *Phonetica* **57**: 170–180 (2000).
17. A. M. Liberman and I. G. Mattingly, The motor theory of speech perception, revised, *Cognition* **21**: 1–36 (1985).
18. A. Protopapas, Connectionist modeling of speech perception, *Psychol. Bull.* **125**: 410–436 (1999).
19. J. L. McCelland and J. L. Elman, The TRACE model of speech perception, *Cogn. Psychol.* **18**: 1–86 (1986).
20. D. H. Klatt, Speech perception: A model of acoustic-phonetic analysis and lexical access, *J. Phonet.* **7**: 279–312 (1979).
21. J. B. Allen, How do humans process and recognize speech? *IEEE Trans. Speech Audio Process.* **2**: 567–577 (1994).
22. H. Fletcher and R. Galt, Perception of speech and its relation to telephony, *J. Acoust. Soc. Am.* **22**: 89–151 (1950).
23. H. Müsch, Review and Computer implementation of Fletcher and Galt's method of calculating the Articulation Index, *Acoust. Res. Lett. Online* **2**: 25–30 (2001).
24. N. R. French and J. C. Steinberg, Factors governing the intelligibility of speech, *J. Acoust. Soc. Am.* **19**: 90–119 (1947).
25. K. Kryter, Method for the calculation and use of the Articulation Index, *J. Acoust. Soc. Am.* **34**: 1689–1697 (1962).
26. ANSI, S3.5-1969, *American National Standard Methods for the Calculation of the Articulation Index*, American National Standards Institute, New York, 1969.
27. ANSI, S3.5-1997, *American National Standard Methods for Calculation of the Speech Intelligibility Index*, American National Standards Institute, New York, 1997.
28. G. A. Miller and P. E. Nicely, An analysis of perceptual confusion among some English consonants, *J. Acoust. Soc. Am.* **27**: 338–352 (1954).
29. C. V. Pavlovic, Factors affecting performance on psychoacoustic speech recognition tasks in the presence of hearing loss, in G. A. Studebaker and I. Hochberg, eds., *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, Needham Heights, MA, 1993.
30. B. Scharf, Critical bands, in J. V. Tobias, ed., *Foundations of Modern Auditory Theory*, Vol. I, Academic Press, New York, 1970, pp. 157–202.
31. E. Zwicker, Subdivision of the audible frequency range into critical bands (Frequenzgruppen), *J. Acoust. Soc. Am.* **33**: 248 (1961).
32. E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichten-empfänger*, Hirzel-Verlag, Stuttgart, Germany, 1967 (available in Engl. transl. by H. Müsch, S. Buus, and M. Florentine as *The Ear as a Communication Receiver*, Acoust. Soc. Am., Woodbury, NY, 1999).
33. S. Buus, Auditory masking, in M. J. Crocker, ed., *Encyclopedia of Acoustics*, Vol. 3, Wiley, New York, 1997, pp. 1427–1445.
34. ANSI, *Specifications for Audiometers*, American National Standards Institute, New York, 1989.
35. ISO 389-7, *Acoustics — Reference Zero for the Calibration of Audiometric Equipment* — Part 7*: Reference Threshold of Hearing under Free-Field and Diffuse-Field Listening Conditions*, International Organization for Standardization, Geneva, 1996.
36. C. V. Pavlovic, Derivation of primary parameters and procedures for use in speech intelligibility predictions, *J. Acoust. Soc. Am.* **82**: 413–422 (1987) [erratum: *J. Acoust. Soc. Am.* **83**: 827 (1987)].
37. H. K. Dunn and S. D. White, Statistical measurements on conversational speech, *J. Acoust. Soc. Am.* **11**: 278–288 (1940).
38. H. Steeneken and T. Houtgast, Mutual dependence of the octave-band weights in predicting speech intelligibility, *Speech Commun.* **28**: 109–123 (1999).
39. C. V. Pavlovic and G. A. Studebaker, An evaluation of some assumptions underlying the articulation index, *J. Acoust. Soc. Am.* **74**: 1606–1612 (1984).
40. G. A. Studebaker and R. L. Sherbecoe, Frequency-importance functions for speech recognition, in G. A. Studebaker and I. Hochberg, eds., *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, Needham Heights, MA, 1993.
41. G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney, Monosyllabic word recognition at higher-than-normal speech and noise levels, *J. Acoust. Soc. Am.* **105**: 2431–2444 (1999) [erratum: *J. Acoust. Soc. Am.* **106**: 2111 (1999)].
42. I. Pollack, Message probability and message reception, *J. Acoust. Soc. Am.* **36**: 937–945 (1964).
43. H. Müsch and S. Buus, Using statistical decision theory to predict speech intelligibility, I Model structure, *J. Acoust. Soc. Am.* **109**: 2896–2909.
44. A. Boothroyd and S. Nittrouer, Mathematical treatment of context effects in phoneme and word recognition, *Acoust. Soc. Am.* **84**: 101–114 (1988).
45. H. J. M. Steeneken and T. Houtgast, A physical method for measuring speech-transmission quality, *J. Acoust. Soc. Am.* **67**: 318–326 (1980).

SPEECH PROCESSING

DOUGLAS O'SHAUGHNESSY
INRS-Telecommunications
Montreal, Quebec, Canada

1. INTRODUCTION

When a computer is used to handle the acoustic signal we call speech, sound is captured by a microphone, digitized for storage in the computer, and then manipulated or transformed to serve some useful purpose, such as efficient coding for transmission or analysis for automatic

conversion to text. Such processing of speech signals by computer converts the speech pressure wave variations (which actually constitute speech) to other forms of information that are more useful for practical computer applications.

Speech is normally produced by a human speaker, whose brain does the processing needed to generate the speech signal. Similarly, human listeners receive speech through their ears and process it with their brains to understand the spoken message. In this chapter, we examine ways that computers use to analyze speech and to simulate production or perception of speech signals. There are many applications for this speech processing: text-to-speech synthesis (e.g., a machine capable of reading aloud to the visually handicapped; a system able to respond vocally from a textual database), speech recognition (e.g., controlling machines via voice; fast entry into databases), speaker verification (e.g., using one's voice as an identifier instead of fingerprints or passwords), and speech coding (i.e., compressing speech into a compact form for storage or transmission, and subsequent reconstruction of the speech when and where needed). All these uses require converting the original speech into a compact digital form, or vice versa, and related digital signal processing.

Such processing usually involves either analysis or synthesis of speech. Recognition of speech (where the desired result is the text that one usually associates with the speech) or of speakers (where the output is the identity of the source of the speech, or a verification of a claimed identity) requires analysis of the speech, to extract relevant (and usually compact) features that characterize how the speech was produced (e.g., features related to the shape of the vocal tract used to utter the speech). Text-to-speech synthesis, on the other hand, tries to convert normal text into an understandable synthetic voice. For speech coding, both the input and output of the process are in the form of speech; thus text is not directly involved, and the objectives are a compression of the data rate (for efficient and secure transmission) and a retention of the naturalness and intelligibility of the original speech, during the analysis and synthesis stages.

2. SPEECH ANALYSIS

For speech analysis, we first convert the speech waves (which, like all analog signals in nature, vary continuously in both time and intensity) into a bitstream, that is, analog-to-digital (A/D) conversion, so that a computer can process the speech signal. The resulting information or "bit rate" (which is of prime interest for efficient operations) is controlled by two factors: the sampling rate (number of evaluations or samples per second) and the sampling precision (number of bits per sample). The sampling rate is directly proportional to the allowed bandwidth of the speech; the Nyquist theorem specifies that the rate be at least twice the highest frequency in the processed signal (otherwise, inevitable "aliasing" distortion appears in the later D/A conversion) [1]. The minimum rate used in practical speech applications is usually 8000 samples/s, thus retaining a range from very low frequencies (theoretically, 0 Hz, although no useful audio data occur below 50 Hz) up to 4 kHz (in practice, an analog lowpass filter usually precedes the A/D converter, and there is a gradual cutoff of frequencies below 4 kHz). Such a low rate is standard for telephone applications, because the switched network only preserves approximately the 300–3200-Hz range. The maximal rate typically employed is 44,100/second (used for audio on compact disks [2–5]), which preserves frequencies beyond the normal hearing range.

Unlike the sampling rate, which follows immediately from the selection of bandwidth range, the choice of bits/sample corresponds to how much distortion is tolerable in a given speech application. To minimize distortion to inaudible levels after A/D conversion, we typically need 12 bits/sample. This gives about 60 dB (decibels) in signal-to-noise ratio, which allows for a typical 30-dB range between strong vowels and weak fricatives in speech. Hence, 12-bit A/D converters are acceptable, although 16-bit systems are more common. In the telephone network, where bandwidth is at a high premium, 8-bit logarithmic (nonuniform) quantization is used on most digital links.

The objective of speech analysis is to transform speech samples into a more useful information representation, for specific objectives other than simply furnishing speech to human ears. The bit rate of basic digital speech (via simple or logarithmic A/D conversion) is typically 64 kbit/s or higher. If we compare that rate to the amount of fundamental information in the signal, we can see a large discrepancy. An average speaking rate is about 12 sounds or phonemes/s, and most languages have an inventory of about 32 phonemes. Thus the phonemic sequence of speech can be sent in approximately 60 bps (bits per second) ($12 \log_2 32$). This does not include intonational or speaker-specific aspects of the speech, but ignores the fact that phoneme sequences are not random (e.g., Huffman coding can reduce the rate). The overall information rate in speech is perhaps 100 bps. Current speech coders are far from providing transparent coding at such rates, but some complex systems can reduce the rate to about 8 kpbs (at 8000 samples/s) without losing quality (other than limiting bandwidth to 4 kHz).

Typical analysis methods try to extract features that correspond to some well-known aspects of speech production or perception. For example, experiments have shown that speakers can easily control the intensity of speech sounds (and that listeners can easily detect small changes in intensity) [6]. Thus intensity (or a related parameter, energy) is often determined in speech analysis, by simply summing a sequence of (squared) speech samples. Similarly, the positions of resonance peaks in the amplitude spectrum of speech (known as formants, abbreviated F1, F2, ..., in order of increasing frequency) have been correlated with shapes of the vocal tract in speech production; they are also well discerned perceptually (peaks are much more salient than spectral valleys). Analysis techniques typically try to extract compact parameterizations of these spectral peaks, either modeling the underlying resonances of the vocal tract (including resonance bandwidths) or some form of the detail (both coarse and fine) in the amplitude spectrum.

Another feature that is often examined in speech analysis is that of the fundamental frequency of the vocal cords, which vibrate during "voiced" (periodic) sounds such as vowels. The rate (abbreviated F0) is directly controlled in speech production, and the resulting periodicity is easily discriminated by listeners. It is used in tone languages directly for semantic concepts, and cues aspects of stress and syntactic structure in many languages. While F0 (or its inverse, the pitch period — the time between successive closures of the vocal cords) is rarely extracted for speech recognizers, it is often used in low-bit-rate speech coders and must be properly modeled for speech synthesis.

3. SPEECH CODING

The objective of speech coding is to represent speech as compactly as possible (i.e., few bps), while retaining high intelligibility and naturalness, with minimal time delay in inexpensive hardware. There are many practical compromises possible here, ranging from zero-delay, cheap, transparent, high-rate PCM (pulse-code modulation) to systems that trade off naturalness for decreased bit rates and increased complexity [5]. We distinguish higher-rate waveform-based coders (usually yielding high-quality speech) from lower-rate parametric coders. Coders in the first class reconstruct speech signals sample-by-sample, representing each speech sample directly with at least a few bits, while the latter class typically discards spectral phase information and processes blocks ("frames") of many samples at a time, which allows the average number of bits/sample to be one or below (e.g., speech coding at 8 kbps).

3.1. Structure in Speech That Coders Exploit

Coders typically try to identify structure in signals, extract it for efficient representation, and leave the remaining (less predictable) signal components either to be ignored (in low-rate systems) or coded using simple techniques (in high-rate systems). There are many sources of structure in speech signals (which, indeed, account for the large difference between typical rates of 64 kbps and an approximately 100-bps theoretical limit). Phonemes typically last 80 ms (e.g., approximately 10 pitch periods), largely due to the increasing difficulty of articulating speech more rapidly (but also to avoid losses in human perception at higher speaking rates). However, identification of each sound is possible from a fraction of each pitch period (at least for speech in noise-free environments). The effective repetition of information in multiple periods helps increase redundancy in human speech communication, which allows reliable communication even in difficult conditions.

The spectra of most speech show a regular structure, which is the product of a periodic excitation (at a rate of F0) and the set of resonances of the vocal tract (which appears as a series of peaks and valleys, averaging about one peak every 1 kHz). Rather than code the speech waveform sample by sample or the spectrum point by point, efficient coders extract from the speech some parameters directly related to the overall amplitude, F0, and the spectral peaks, and use these for transmission.

A simple spectral measure for speech analysis is the zero-crossing rate (ZCR), which provides a basic frequency estimate for the major energy concentration. The ZCR is just the number of times the speech signal crosses the time axis (i.e., changes algebraic sign) in a given time period (e.g., taking an overly simple nonspeech case, a sinusoid of 100 Hz has a ZCR of 200/s). Background acoustic noise often has a steady, broad lowpass spectrum and thus has a ZCR corresponding roughly to a stable frequency in the low range of the signal bandwidth. On the other hand, for weak speech obstruents that are difficult to detect against background noise, the ZCR is either high (corresponding to a high-frequency concentration of energy in fricatives and stop bursts) or very low (if a "voicebar," corresponding to radiation of F0 through the throat, dominates). The ZCR is easy to compute, and can be used for speech endpoint detection (see below).

3.2. Exploiting Simple Structure

Advanced coders examine blocks of data, consisting of sequences of speech samples (which necessarily increases the response time, but facilitates temporal exploitation). The simplest coders (PCM) represent each sample independently. Using a uniform quantifier, however, is inefficient because most speech samples tend to be small (i.e., speech is more often weak than strong), and thus the quantifier levels assigned to large samples are rarely used. Optimal encoding occurs when, on average, all levels are equally used. Since the distribution of speech samples typically resembles a Gamma probability density function (their likelihood decaying roughly exponentially as amplitude increases), a logarithmic compression prior to uniform quantization is useful (e.g., the μ-law or A-law log PCM, common in telephone networks).

Many coders adapt in time, exploiting slowly changing characteristics of the speech and/or transmission channel, such as following the shape or movements of the vocal tract. For example, the step size of the quantifier can be adjusted to follow excursions of the speech signal, integrated over time periods ranging from 1 to 100 ms. When the signal has large energy, using larger step sizes can avoid clipping (which causes nonlinear and severe distortion). When the step size is reduced for low-energy samples, the quantization noise is proportionally reduced without clipping.

3.3. Exploiting Detailed Spectral Structure

Many speech coders use linear prediction methods to estimate a current speech sample based on a linear combination of previous samples [8]. This is useful when the speech spectrum is nonuniform (pure white noise, with independent samples and a flat spectrum, would allow no gain through such prediction). The simplest prediction occurs in delta modulation, where one previous speech sample directly provides the estimate of each current sample; the assumption is that a typical sample changes little from its immediately prior neighbor. This becomes truer if we sample well above the Nyquist rate, as is done in delta modulation, which in turn allows the use of a one-bit quantifier (trading off sampling rate for bit precision) [7].

The power of linear prediction becomes more apparent when the order of the predictor (i.e., the window or number of prior samples examined) is about 10 or so, and when the predictor adapts to movements of the vocal tract (e.g., is updated every 10–30 ms). This occurs in two very popular speech coders, ADPCM (adaptive differential PCM) and LPC (linear predictive coding). They exploit the fact that the major excitation of the vocal tract for voiced speech occurs at the time that the vocal cords close (once per pitch period), which causes a sudden increase in speech amplitude, after which the signal decays exponentially (with a rate inversely proportional to the bandwidth of the strongest resonance — usually F1). Each pitch period approximately corresponds to the impulse response of the vocal tract, and is directly related to its resonances. Since there are about four such formants (in a typical 0–4-kHz bandwidth) and since each formant can be directly characterized by a center frequency and a bandwidth, a predictor order of ten or so is adequate (higher orders give more precision, but offer diminishing returns as computation and bit rate increase).

The 10–16 LPC parameters derived from such an analysis form the basis of many coders, as well as for speech recognizers. These parameters are the multiplier coefficients of a direct-form digital filter, which, when used in a feedforward fashion, converts a speech signal being modeled into a "residual" or "error" signal, which is then suitable for either simple coding (with fewer bits, typically 3–4/sample, in ADPCM) or further parameterization (as in LPC systems).

Using the coefficients in a feedback filter, the system acts as a speech synthesizer, when excited by an impulse train (impulses spaced every 1/F0 samples) or a noise signal (simulating the frication noise generated at a narrow constriction in the vocal tract). This latter, dual excitation is a very simple, but powerful, model of the LPC residual signal, which allows very low bit rates in basic LPC systems [9]. The result is intelligible but synthetic-quality speech (less natural than the "toll-quality" speech of the telephone network with log PCM, or with other medium-rate, waveform-coding methods). The LPC model, with its dozen spectral coefficients parameterizing the vocal tract shape (filter) and its three excitation parameters (F0, amplitude, and a single "voicing" bit noting a decision whether the speech is periodic) modeling the residual, often operates around 2.4 kbps (using updates every 20 ms). The LPC coefficients are usually transformed into a more efficient set for coding, such as the reflection coefficients (which are the multipliers in a lattice-form vocal tract filter, and can actually correspond to reflected energy in simple three-dimensional vocal tract models, modeling the two traveling waves of pressure, one going up the vocal tract, the other down). Another popular set is the line spectral frequencies (LSF), which displace the resonance poles in the digital spectral z plane onto the unit circle, which allows more efficient differential coding in one dimension around the circle, rather than the implicit two-dimensional coding of the resonances inherent in other LPC forms.

3.4. Exploiting Structure Across Parameters

More efficient coders go beyond simple temporal and spectral structure in speech, to exploit correlations across parameters, both within and across successive frames of speech. Even relatively efficient spectral representations such as the LSFs do not produce orthogonal parameter sets; that is, there remain significant correlations among sets of parameters describing adjacent frames of speech. Shannon's theorem states that it is always more efficient to code a signal in vector form, rather than code the samples or parameters as a succession of independent numbers. Thus, we can group related parameters as a block and represent the set with a single index. For example, to code a speech frame every 20 ms, a set of 10 reflection coefficients might need 50 bits as scalar parameters (about 5 bits each), but need only 10 bits in vector quantization (VQ). In the latter, we chose 1024 ($= 2^{10}$) representative points in 10-dimensional space (where each dimension corresponds to a parameter). The points are usually chosen in a training phase (i.e., coder development) by examining many minutes of typical speech (ranging over a wide variety of different speakers and phonemes). Each frame provides one point in this space, and the chosen points for the VQ are the centroids of the most densely populated clusters. Since there are only about 1000 different spectral patterns easily discriminable in speech perception (ignoring the effects of F0 and overall amplitude), a 10-bit system is often adequate. Ten-bit VQ represents effectively a practical limit computationally as well, since a search for the optimal point among much more than 1024 possibilities every 20 ms may exceed hardware capacity. VQ basically trades off increased computation during the analysis stage for lower bit rates in later transmission or storage. The most common use today for speech VQ is in CELP (code-excited linear prediction) speech, where short sequences of LPC residual signals are stored in a codebook, to excite an LPC filter (whose vocal tract spectra are themselves represented by another codebook). CELP provides toll-quality speech at low rates, and has become very popular in recent years.

4. FUNDAMENTAL FREQUENCY (F0) ESTIMATION (PITCH DETECTORS)

Both low-rate speech coders and text-to-speech synthesizers require estimation of the F0 of speech signals, as well as the related (and simpler) estimation of the presence or absence of periodicity (i.e., voicing). For many speech signals, it is a relatively simple task to detect periodicity and measure the period. However, despite hundreds of algorithms in the literature [10], no one pitch detector (so called because of the close correlation of F0 and perceived pitch) is fully accurate. Environmental noise often obscures speech periodicity, and the interaction of phase, harmonics, and spectral peaks often creates ambiguous cases where pitch detectors can make small or even large mistakes (especially in weaker sections of speech).

The basic approach to F0 estimation simply looks for peaks in the speech waveform, spaced at intervals roughly corresponding to typical pitch periods. Periods can range

from 2 ms (sounds from small infants) to 20 ms (for large men), but each individual speaker usually employs about an octave range (e.g., 6–12 ms for a typical adult male). Many estimators use heuristics, such as the fact that F0 rarely changes abruptly (except when voicing starts or ceases). Since F0 is roughly independent of which phoneme is being uttered, and since the structure of formants (and related phase effects) in a voiced spectrum can obscure F0 estimation, we often eliminate from analysis the frequencies above 900 Hz (via a lowpass filter), thus retaining one strong formant containing several harmonics to supply the periodicity information. (This also simplifies processing by use of a decimated signal, which allows a lower sampling rate after lowpass filtering.) At extra cost, spectral flattening can be provided by autocorrelation methods or by LPC inverse filtering, to further reduce F0 estimation errors. Other pitch detectors do peak-picking directly on the harmonics after a Fourier transform of the speech.

5. AUTOMATIC SPEECH RECOGNITION (ASR)

Translating a speech signal into its underlying message (i.e., ASR) is a pattern recognition problem. In principle, one could store all possible speech signals, each transcribed with its corresponding text. Given today's faster computers and the decreasing cost of computer memory, one might wonder if this radically simple approach could eventually solve the ASR problem. To see that this is not true, consider the immense number of possible utterances. Typical utterances last a few seconds; at 10,000 samples/s, we potentially have, say, $2^{30,000}$ signals. Even with very efficient coding (e.g., 100 bps), we would still have an immense 2^{300} signals. Most of these would not be recognizable as speech, but it is impossible *a priori* to just consider ones that might eventually occur as an input to an ASR system (even enlisting millions of speakers talking for days would be quite insufficient).

Thus speech signals to be recognized must be processed to reduce the amount of information from perhaps 64 kbps to a much lower figure. Obvious candidates are low-rate coders, since they preserve enough information to reconstruct intelligible speech. (Higher-rate waveform coders retain too much information, which is useful for naturalness but is not needed for intelligibility; only the latter is important for ASR.)

Unlike speech coding, where the objective of speech analysis is to reproduce speech from a compact representation, ASR instead transforms speech into its corresponding textual equivalent. The direct relationship between the spectral envelope of speech and vocal tract shape (and hence to the phoneme being uttered) has led to intense use of efficient representations of spectral envelope for ASR. (The lack of a simple, direct correlation between F0 and phonemes, on the other hand, has led to F0 being largely ignored in ASR, despite its use to cue semantic and syntactic information in human speech recognition.) One difficulty has been how to extract compact yet relevant information about the envelope. Simple energy is useful, but is often subject to variations (e.g., automatic gain control, variable mouth-to-microphone distance, varying channel gain) irrelevant for phonemic distinctions; as a result, a simple energy measure is often not used for ASR, but change in energy between frames is.

LPC parameters were once popular for ASR, but they have been largely replaced by the mel-scale frequency cepstral coefficients (MFCCs) [11,12]. The term *mel-scale* refers to a frequency-axis deformation, to weight the lower frequencies more than higher ones (which is quite difficult to do in LPC analysis). This follows critical-band spacing in audition, where perceptual resolution is fairly linear below 1 kHz, but becomes logarithmic above that. The cepstrum is the inverse transform of the log amplitude of the Fourier transform of speech. The amplitude spectrum of speech in decibels is weighted via triangular filters spaced at critical bands, and then coefficients are produced via weightings from increasingly higher-frequency sinusoids (in the inverse Fourier transform step). The first 10 or so MFCCs provide a good spectral envelope representation for ASR. C0 (the first MFCC) is actually just the overall speech energy (and is often omitted from use); C1 provides a simple measure of the balance between low- and high-frequency energy (the one-period sinusoid weights low frequencies positively and high ones negatively). Higher coefficients provide the increasingly finer spectral details needed to distinguish, say, the vowels /i/ and /e/.

5.1. Timing Problems in ASR

The major difficulty for ASR is the large amount of variability in speech production. In text-to-speech synthesis, one synthetic voice may suffice, and all listeners must adjust to its accent. ASR systems, however, must accommodate different speaking styles, by storing many different speakers' patterns or by integrating knowledge about different styles. Variations occur at several levels: timing, spectral envelope, and intonation. There is much freedom in how a speaker times articulations and how exactly the vocal tract moves. In 1986, ASR commonly stored templates consisting of successive frames of spectral parameters (e.g., LPC coefficients), and compared them with those of an unknown utterance. Since utterances of the same text could easily have different numbers of frames, the alignment of frames could not simply be one-to-one. Nonlinear "dynamic time warping" (DTW) was popular because it compensated for small speaking rate variations. However, it was still computationally expensive and extended awkwardly to longer utterances; it was also difficult to improve models with additional training speech.

5.2. General Stochastic Approach

During the 1980s, the hidden Markov model (HMM) approach became the dominant method for ASR. It accepted large amounts of computation during an initial training phase in order to get a more flexible and faster model at recognition time. Furthermore, HMMs provided a mathematically elegant and computationally practical solution to some of the serious problems of variability in speech production. DTW had partially accommodated the timing problem, via its allowance of nonlinear temporal paths in the recognition search space,

but was unsatisfactory for variations in pronunciation (e.g., spectral differences due to perturbations in vocal tract shape, different speakers, phonetic contexts). DTW was essentially a deterministic approach to a stochastic problem.

With HMMs, speech variability is treated in terms of probabilities. The likelihood of a speaker uttering a certain sound in a certain context is modeled via probability distributions, which are estimated from large amounts of "training data" speech. Given sufficient computer power and memory, ASR systems tend to have improved recognition accuracy as more data are implicated in the stochastic models, to make them more reliable. While computer resources are never infinite, this approach is feasible to improve systems for "speaker-independent ASR," where training speech is obtained from hundreds of different speakers, and the system then accepts input speech from all users.

For alternative "speaker-dependent" recognizers, which need training from individual users and only employ models trained on each speaker's voice at recognition time, the large amounts of training data are less feasible, given most users' reluctance to provide more than a few minutes of speech. The latter systems have better accuracy, since the HMM models are directly relevant for each user's speech, whereas speaker-independent HMMs must model much more broadly across the diversity of many speakers (such models, as a result, are less discriminative).

Most ASR is done using the ML (maximum-likelihood) approach, in which statistical models for both speech and language are estimated based on prior training data (of speech and texts, respectively). After training establishes the models, at recognition time an input speech signal is analyzed and the text with the corresponding highest likelihood is chosen as the recognition output. Thus, given a signal S, we choose text T, which maximizes the a posteriori probability $P(T|S) = P(S|T)P(T)/P(S)$, using Bayes' rule. It is impossible to directly get good estimates for $P(T|S)$ because of the extremely large number of possible speech signals S. Instead, we develop estimates for $P(S|T)$ (the acoustic model) and for $P(T)$ (the language or text model). When maximizing across possible texts T, we ignore the denominator $P(S)$ term as irrelevant for the best choice of T. Even for large vocabularies, the number of text possibilities is much smaller than the number of speech signals; thus it is more practical to estimate $P(S|T)$ than $P(T|S)$. For each possible text, the acoustical statistics are obtained from speakers repeatedly uttering that text (in practice, such estimates can be obtained for small text units, such as words and phonemes, while still using speech conveniently consisting of sentences). The a priori likelihood of a text T being spoken is $P(T)$, which is obtained by examining computerized textual databases.

There are efficient methods to develop such statistics and to evaluate the large number of possibilities when searching for the maximum likelihood (the search space for a vocabulary with thousands of words and for an utterance of several seconds, at typically 100 frames/s, is quite large). The "forward–backward" method examines all possible paths, summing many small joint likelihoods, while the more efficient Viterbi method looks for the single best path [11]. The latter is much faster, and is commonly used because in practice it tends to sacrifice little in recognition accuracy. When trying to discriminate among similar words, however, the ML approach sometimes fails, because it does not examine how close alternative possible texts are. In addition, HMMs treat all speech frames as equally important, which is not the case in human speech perception. Alternative methods such as maximum mutual information estimation or linear discriminant analysis are considerably more expensive and complex, but are more selective in examining the data, focusing on the differences between competing similar texts, when examining a speech input.

5.3. Details of the Hidden Markov Model (HMM) Approach

The purpose of this method is to model random dynamic behavior via a mathematical method where different "states" represent some aspects of the behavior. The basic, first-order Markov chain has several states, connected by transitions among states. The likelihood of leaving each state is modeled by a probability distribution (usually a PDF — probability density function), and each state itself is also so modeled. For speech, these two sets of PDFs attempt to model the variable timing and articulation of utterances, respectively. Each state roughly corresponds to a vocal tract shape (or more precisely to a speech spectral envelope with formants resulting from the vocal tract shape). Each transition models the likelihood that the vocal tract moves from one position to another.

The PDF for transitions from a given state usually has a simple form: a relatively high likelihood of remaining in that state (e.g., 0.8), a smaller chance of moving to the next state in the time chain (e.g., 0.15), and a yet smaller chance of skipping to the state after the next one. Since time always moves forward, we do not allow backward transitions (e.g., we go through an HMM modeling a word, starting with its first sound and proceeding to its last sound). Since each state models roughly a vocal tract shape (or more often an average of shapes), we stay in each state for several 10-ms frames (hence the high self-loop likelihood). Allowing an occasional state to be skipped accounts for some variability in speech production, especially for rapid, unstressed speech. For example, the training speech may be clear and slow, thus creating states that need not always be visited in later (perhaps fast) test speech.

The transition PDF thus described leads to an exponential PDF for the duration of state visits, which is an inaccurate model for actual phoneme durations. There is too much bias toward short sounds. For instance, very few phonemes last only 1–2 frames. Some more complicated HMM approaches allow direct durational modeling, but at the cost of increased complexity.

In practice, every incoming speech signal is divided into successive 10-ms frames of data for analysis (as in speech coding). During the training phase, many frames are assigned to each given HMM state, and the state's PDF is simply the average across all such frames. Similarly, the PDF describing which state B follows any given state A in modeling the dynamics of the speech simply follows the likelihood of moving to a nearby vocal tract shape (B)

in one frame, given that the previous speech frame was assigned to state A. This simple approach in which the model takes no direct account of the history of the speech (beyond one state in the past) is called a *first-order Markov model*. It is clear, when dealing with typical 10-ms frames, that there is certainly significant correlation across many successive frames (due to coarticulation in vocal tract movements). Thus the first-order assumption is an approximation, to simplify the model and minimize computer memory. Attempts to use higher-order models have not been successful because of the immense increase in complexity needed to accommodate reasonable amounts of coarticulation. The HMM is called "hidden" because the vocal tract behavior being modeled is not directly observable from the speech signal (if the input to an ASR system were X rays of the actual moving vocal tract during speech, we could use direct Markov models, but this is totally impractical).

Gaussian PDFs are normally used to specify each state in an HMM. Such a PDF for an N-dimensional feature vector $\mathbf{x}$ for a spoken word assigned index i is

$$P_i(x) = (2\pi)^{-N/2}|\mathbf{W}_i|^{-1/2}\exp\left[\frac{-(\mathbf{x}-\mu_i)^{\mathrm{T}}\mathbf{W}_i^{-1}(\mathbf{x}-\mu_i)}{2}\right] \quad (1)$$

where $\mathbf{W}_i$ is the covariance matrix (noting the individual correlations between parameters), $|\mathbf{W}_i|$ is the determinant of $\mathbf{W}_i$, and μ_i is the mean vector for word i. Most systems use a fixed $\mathbf{W}$ matrix (instead of individual $\mathbf{W}_i$ for each word) because (1) it is difficult to obtain accurate estimates for $\mathbf{W}_i$ from limited training data, (2) using one $\mathbf{W}$ matrix saves memory and computation, and (3) $\mathbf{W}_i$ matrices are often similar for different words. If one chooses parameters that are independent (i.e., no relationship among the numbers in the vector $\mathbf{x}$), the $\mathbf{W}$ matrix simplifies to a diagonal matrix, which greatly simplifies calculation of the PDF (which must be repeatedly done for each speech frame and for each HMM). Thus many ASR systems assume (often without adequate justification, other than reducing cost) such a diagonal $\mathbf{W}$. Unless an orthogonalization procedure is performed (e.g., a Karhunen–Loeve transformation—itself quite costly), most commonly used feature sets (LPC coefficients, MFCCs) have significant correlation.

The elements along the main diagonal of $\mathbf{W}$ indicate the individual variances of the speech analysis parameters. The use of $\mathbf{W}_i^{-1}$ in the Gaussian PDF notes that those features with the smallest variances are the most useful for discriminating sounds in ASR. Parameter sets that lead to small variances and widely spaced means $\boldsymbol{\mu}_i$ are best, so that similar sounds can be consistently discriminated.

The Gaussian form for the state PDF is often appropriate for modeling many physical phenomena; the idea follows from basic stochastic theory: the sum of a large number of independent, identically distributed random variables resembles a Gaussian. Natural speech, coming from human vocal tracts, can be so treated, but only for individual sounds. If we try to model too many different vocal tract shapes with one HMM state (as occurs in multispeaker, or context-independent ASR), the Gaussian assumption is much less reasonable (see text below for the discussion of mixtures of Gaussian PDFs).

We create HMMs to model different units of speech. The most popular approaches are to model either phonemes or words. Phonemic HMMs require fewer states than word-based HMMs, simply because phonemes are shorter and less phonetically varied than words. If we ignore coarticulation with adjacent sounds, many phonemes could be modeled with just one state each. Inherently dynamic phonemes, such as stops and diphthongs, would require more states (e.g., a stop such as /t/ would at least need a state for the silence portion, a state for the explosive release, and probably another state to model ensuing aspiration). There is a very practical issue of how many states are appropriate for each HMM. There is no simple answer; too few states lead to diffuse PDFs and poor discriminability (especially for similar sounds such as /t/ and /k/), due to averaging over diverse spectral patterns. Too many states lead to excessive memory and computation, as well as to "undertraining," where there are not enough speech data in the available training set to provide reliable model parameters.

Consider the (very practical) task of simply distinguishing "yes" versus "no" (or perhaps just the 10 digits: 0, 1, 2,...9); it is efficient to create an HMM for every word in the allowed vocabulary. There will be several states in each HMM to model the sequence of phonemes in each word. HMMs work best when there is a rough correspondence between the number of states and the number of distinguishable sounds in the speech unit being modeled. If the model creation during the training phase is done well, this allows each state to have a PDF with minimal variance. At recognition time, such tight PDFs will more easily discriminate words with different phoneme sequences.

As the size of the allowed vocabulary increases, however, it becomes much less practical to have one or more models for every word. Even in speaker-independent ASR, where we could ask thousands of different people to furnish speech data to model the many tens of thousands of words in any given language such as English, the memory and search time needed for word-based HMMs goes beyond current computer resources. Thus, for vocabularies larger than 1000 words, most systems employ HMMs that model phonemes (or sometimes "diphones," which are truncated sequences of phoneme pairs, used to model coarticulation during transitions between two phonemes). Diphones are obtained by dividing a speech waveform into phoneme-sized units, with the cuts in the middle of each phone (thus preserving in each diphone the transition between adjacent phonemes). In text-to-speech applications, concatenating diphones in a proper sequence (so that spectra on either side of a boundary match) usually yields smooth speech because the adjacent sounds at the boundaries are spectrally similar; e.g., to synthesize *straight*, the six-diphone sequence /#s-st-tr-re-et-t#/ would be used (# denoting silence).

Using phonemic HMMs reduces memory and search time, as well as allowing a fixed set of models, which does not have to be updated every time a word is

added to the vocabulary. If the models are "context-independent," however, their discriminability is small. When averaging over a wide range of phonetic contexts, the states modeling the initial and final parts of each phoneme have varying spectral patterns and thus broad PDFs of poorer discriminability.

Nowadays, more advanced ASR systems employ "context-dependent" phonemic HMMs (e.g., diphone models), where each phoneme is assigned many HMMs, depending on its immediate neighbors. A simple and common (but expensive) technique uses triphone models, where for a language with N phonemes we have N^3 models; each phoneme has N^2 models, one for each context (looking before and after by one phoneme). There is much evidence from the speech production literature about the effects of coarticulation, which are significant even beyond a triphone window. (A diphone approach is a compromise between a simple context-independent scheme and a complex triphone method; it has N^2 models.) For English, with about 32 phonemes, the triphone method leads to tens of thousands of HMMs (which need to be adequately analyzed and stored in the training phase, and all must be examined at recognition time). Many systems adopt a compromise by grouping or clustering similar models according to some set of phonetic rules (e.g., the decision-tree approach). If the clustering is done well, one could even expand the analysis window beyond triphones, to accommodate further coarticulation (e.g., in the word "strew," lip rounding for the vowel /u/ affects the spectrum of the initial /s/; thus a triphone model for /u/ looking leftward only at the neighbor /r/ is inadequate).

5.4. Language Models for ASR

Going back only a decade or so, we find much less successful ASR methods that concentrated only on the acoustic input to make decisions. It was thought that all the required information to translate speech to text could be found in the speech signal itself; thus no cognitive modeling of the listener seemed necessary. Some surprisingly simple experiments that modeled some aspects of language changed that naive approach, with positive results in recognition accuracy. Indeed, it can be easily observed that words in speech (as in text) rarely occur in random order. Both syntactically and semantically, there is much redundancy in word order. For example, when talking about a dog, one may well find the semantically related words "small" or "black" immediately prior to "dog." As for syntax, there are many restrictions on word sequences, such as the common structure of article+adjective+noun for English noun phrases, and strict order in verb phrases such as "may not have been eaten."

Thus researchers developed models for language, in which the likelihood of a given word occurring after a prior word sequence is evaluated and used in the decision process of speech recognition. The most common approach is that of a trigram language model, where we estimate the probability that any given word in text (or speech) will follow its preceding two words (written or spoken). Textual redundancy in English (and indeed many languages) goes well beyond a trigram window of analysis, but practical issues of computer memory and the availability of training data have so far limited most models to a three-word range. Training for language models has been almost exclusively done using written texts (and rarely transcriptions of speech), despite the inappropriateness for ASR of written compositions (i.e., most speech is spontaneous, rarely from written texts), owing to the cost (and often limited accuracy) of transcribing large amounts of speech. Researchers find it much easier to use the many databases of text available today, despite their shortcomings.

The major advantage of models such as trigrams is that they incorporate diverse aspects of practical text redundancies automatically (no complex semantic or syntactic analysis is needed). This reflects the current premature state of natural language processing (e.g., the continuing difficulties in automatic language translation). This lack of intelligent analysis in trigram models, however, leads to serious inefficiencies. As the allowed vocabulary increases in size, for applications not limited to specific topics of conversation, the availability of sufficient text to obtain reliable statistics is a major problem. Depending on what is counted as a word (e.g., do "eat, eats, eating, eaten" count as four individual words?), there are easily hundreds of thousands of words in English; taking a very conservative estimate of 10^5 words, there are 10^{15} trigrams, which strains current computer capacities, or at least significantly increases costs. As a result, many acceptable trigrams (and even many bigrams and unigrams) are not observed in any given training text (even large ones containing millions of words). The usual estimation of probabilities simply divides the number of occurrences of each specific trigram (or bigram) by the total number of occurrences of each word being modeled in the training text. For sequences occurring frequently in training texts, the estimates are reasonable, at least for the purposes of modeling (and recognizing speech from) additional text from the same source. One may readily create different language models for different types of text, such as those for physicians, lawyers, and other professionals, corresponding to specific applications.

A serious problem with the approach above is how to handle unobserved sequences (e.g., the very large number of three-word sequences not found in a given training text, but nonetheless permitted in the language). Assigning them all a probability of zero is quite inappropriate, since such word sequences would then never appear in a speech recognizer's output, even if they were actually spoken. One solution simply assigns all unseen sequences the same tiny probability, and reduces the probabilities of all observed sequences by a compensatory amount (to ensure that the total probability remains equal to one). This is somewhat better than arbitrarily assigning zero likelihoods, but treats all unseen sequences as equally likely, which is a very rough and poor estimate. A more popular way is the "backoff" approach, in which the overall assigned likelihood for a word is a weighted combination of trigram, bigram, and unigram probabilities. The weightings are appropriately adjusted for unseen sequences. If no trigram (or bigram) estimate is available, one simply assigns its corresponding weighting a value of zero, and uses existing

(bigram and) unigram statistics. Of course, words that never appear at all in a training set have no statistics to which to back off; their likelihoods must be estimated another way.

While the trigram method has considerable power as a language model for ASR, there is also considerable waste. The memory for trigram statistics for large-vocabulary applications is very large, and many of the probabilities are poorly estimated due to insufficient training data. It is more efficient if one clusters similar words together, thus reducing memory and making the fewer statistics more reliable. One extreme case is the tri-PoS approach, where all words are classified by their syntactic category (e.g., part of speech, such as noun, verb, or preposition). Depending on how many categories are used, the statistics can be very reduced in size, while still retaining some powerful syntactic information to predict common word-class sequences. Unfortunately, the semantic links among adjacent words are largely lost with this purely syntactic language model. Owing to the reduced model size, one can easily extend the tri-PoS method to windows of more than three words, or one can extend the word categories to include semantic labels as well. Word-class language models using perhaps hundreds of hybrid syntactic-semantic classes may eventually be a good compromise approach between the too simple tri-PoS method and the huge trigram approach.

5.5. Practical Issues in ASR — Segmentation

In addition to the serious ASR issues of computer resources and of speech variability, there is another aspect of recognizing speech that makes the task more difficult than synthesizing speech: segmentation. In the case of automatic speech synthesis, the computer's task is easier for two reasons: (1) the input text is clearly divided into separate words, which simplifies processing; and (2) the burden is placed on the human listener to adapt to the weaknesses of the computer's synthetic speech. In ASR, on the other hand, the machine must adapt to each user's different style of speech, and there are few clear indications to boundaries in a speech signal. In particular, acoustic cues to word boundaries in speech are very few. Automatic segmentation of a long utterance into smaller units (ideally into words) is very difficult, but is effectively necessary to reduce computation and lower error rates. One is tempted to exploit periods of silence for segmentation, but they are unreliable cues to linguistic boundaries. Long pauses are usually associated with sentence boundaries, but they often occur within a sentence and even within words (e.g., in hesitations). Short pauses are easily confused with phonemic silences (i.e., the vocal tract closures during unvoiced stops).

Segmenting speech into syllabic units is easier, because of the typical rise and fall of speech energy between vowels and consonants. However, many languages (e.g., English) allow many different types of syllables (ranging from ones with only a vowel to ones with several preceding and ensuing consonants). Many languages (e.g., Japanese) are much easier to segment because of their consistent alternation of consonants and vowels. Segmenting speech into phonemes is also difficult if the language allows sequences of similar phonemes, such as vowels, or large variations in phonemic durations (e.g., English allows severe reduction of unstressed phonemes, such that recognizers often completely miss brief sounds; at the other extreme, long vowels or diphthongs can often be misinterpreted as containing multiple phonemes).

To facilitate segmentation, many commercial recognizers still require speakers to adopt an artificial style of talking, pausing briefly (at least a significant fraction of a second) after each word. Silences of more than, say, 100 ms should not normally be confused with (briefer) stop closures, and such sufficiently long pauses allow simple and reliable segmentation of speech into words.

In order of increasing recognition difficulty, four styles of speech are often distinguished: *isolated-word* or *discrete-utterance* speech, *connected-word* speech, *continuous*-read speech, and normal conversational speech. The last two categories concern continuous-speech recognition (CSR), which requires little or no adaptation of speaking style by system users. CSR allows the most rapid input (e.g., 150–250 words/min), but is the most difficult to recognize. For isolated-word recognition, requiring the speaker to pause after each word is very unnatural for speakers and significantly slows the rate at which speech can be processed (e.g., to about 20–100 words/min), but it clearly alleviates the problem of isolating words in the input speech signal.

Using word units, the search space (in both memory and computation) is much smaller than for longer utterances, and this leads to faster and more accurate recognition. However, few speakers like to talk in "isolated word" style. There remains also the serious issue of "endpoint detection," where the beginning and end times for speech must be determined; against a background of noise, it is often difficult to discriminate weak sounds from the noise. Endpoint detection is more difficult in noise: speaker-produced (lip smacks, heavy breaths, mouth clicks), environmental (stationary: fans, machines, traffic, wind, rain; nonstationary: music, shuffling paper, door slams), and transmission (channel noise, crosstalk). The large variability of durations and amplitudes for different sounds makes reliable speech versus silence detection difficult; strong vowels are easy to locate, but the boundaries between weak obstruents and background noise are often poorly estimated. Typically, endpoint location uses energy as the primary measure to cue the presence of speech, but also employs some spectral parameters. Clear endpoint decisions are required very often in isolated-word speech. For these various reasons, most current research focuses on more normal "continuous" speech.

Stochastic modeling has an important place in speech processing, given the large amount of variability in speech production. Humans are indeed incapable of producing exactly the same utterance twice. With effort, they can likely make utterances sound virtually the same to listeners, but at the sampling and precision levels of automatic speech analyzers, there are always differences, which result in parameter variations. If one selects ASR parameters well and such variation is minimal, system performance can be high with proper stochastic models.

However, all too often, variability is large and goes beyond the modeling capability of current ASR systems. Future ASR must find a way to combine deterministic knowledge modeling (e.g., the "expert system" approach to artificial intelligence tasks) with well-controlled stochastic models to accommodate the inevitable variability. Currently, the research pendulum has swung far away from the expert-system approach common in the mid-1970s.

5.6. Practical Issues in ASR: Noise

In many applications, the speech signal is subject to various distortions before it is received by a recognizer [13]. Noise may occur at the microphone (e.g., in a telephone booth on the street) or in transmission (e.g., fading over a portable telephone). Noise leads to decreased accuracy in the parameters during speech analysis, and hence to poorer recognition, especially (as is often the case) if there is a mismatch in conditions between the training and testing speech. It is very difficult (and/or expensive) to anticipate all distortion possibilities during training, and thus ASR accuracy certainly decreases in such cases.

As an attempt to normalize across varying conditions, ASR systems often calculate an average parameter vector over several seconds (e.g., over the entire utterance), and then subtract this from each frame's parameters, applying the net results to the recognizer. This approach takes account of slow changes in average energy as well as the filtering effects of the transmission microphone and channel, which usually change slowly over time. Such "mean subtraction" is simple and useful in noisy conditions, but can delay recognition results if we must determine the mean for several seconds of speech and only then start to recognize the speech. A similar method called RASTA (RelAtive SpecTrAl) uses a highpass filter (with a very low-frequency cutoff) to eliminate very slowly varying aspects of a noisy speech signal, as being specific to the transmission channel and irrelevant to the speech message.

More traditional speech enhancement techniques can also be used in ASR [14]. These include spectral subtraction (or related Wiener filtering), where the average amplitude spectrum observed during signal periods that are estimated to be devoid of actual speech is subtracted from the spectrum of each speech frame. In this way, stationary noise can be somewhat suppressed. Frequency ranges dominated by noise are thus largely eliminated from consideration in the analysis to determine relevant parameters. The output of speech enhancers sounds more pleasant, but intelligibility is not enhanced by this removal of noise.

Sometimes used for speech enhancement, another approach called "comb filtering" estimates F0 in each frame and suppresses portions of the speech spectrum between the estimated harmonics. Again, the enhanced output speech sounds more pleasant, but this method requires a reliable F0 detector, and the resulting comb filter must adapt dynamically to the many changes in pitch found in normal speech.

Much more powerful speech enhancement is possible if several microphones are allowed to record in a noisy environment. For example, in a noisy plane cockpit, one microphone close to the pilot's lips would capture a signal containing speech plus some noise, while another microphone outside the helmet would capture a version of the noise corrupting the initial signal. Using adaptive filtering techniques very similar to those for echo cancellation at 2/4-wire junctions in the switched telephone network, one can improve very significantly both the quality and the intelligibility of such speech.

5.7. Artificial Neural Networks (ANNs)

Since 1990, in another attempt to solve many problems of pattern recognition, significant research has been done in the field of artificial neural networks. These ANNs simulate very roughly the behavior of neurons in the human central nervous system. Given a signal (e.g., speech) to recognize, each node of an ANN accepts a set of signal-based inputs, computes a linear combination of the these values, compares the result against a threshold, and provides a binary output (1 if the linear combination exceeds the threshold, 0 otherwise). By combining such nodes in a network, quite complicated decision spaces can be constructed, which have proven useful in numerous pattern recognition applications, including speech recognition [15]. For ASR, the ANN typically accepts a long vector of L parameters (e.g., for 10 MFCCs per frame and a 100-frame utterance, one has a vector of $L = 1000$ dimensions), as emitted by the usual speech analysis methods discussed above, and these L numbers provide the input to a set of M nodes (each individual input value may be input to several nodes). Each node thus makes a linear combination of a selected set of values from the input vector. The conceptually simplest ANN would assign one output node to each possible output of the recognition process (e.g., if the utterance must be one of the 10 spoken digits, $M = 10$, and one tries to specify the weights in the ANN such that, on any given trial, only one of the M outputs will be 1; the label on that output node provides the textual output. Thus, with proper training of the ANN to choose the appropriate weights, when one says "six" and the L resulting analysis parameters are fed through the ANN, ideally only one of the 10 output nodes (i.e., that corresponding to "six") will show a 1 (the others showing 0). This assumes, however, that each of the allowed vocabulary words corresponds to a simple distribution in L-dimensional acoustic space, such that basic hyperplanes can partition the space into 10 separable clusters (and furthermore that a training algorithm looking at many utterances of the 10 digits can determine the correct weights). It is very rare that these assumptions are exactly true, however. To adapt to practical speech cases, a partial solution has been to add extra layers of nodes to the ANN. We allow the number M_2 of nodes in the second layer to be larger than M, and allow the binary outputs of those M_2 nodes to pass through another layer of weights and N nodes. Such a double-layer ANN can partition the speech space into the much more complex shapes that correspond to practical systems. For speech, in general, it appears that a three-tier system is needed; the original input speech-parameter vector provides the numbers for the lowest layer, the information propagates through three levels of weightings

(two "hidden" layers of nodes) to finally appear at the output layer (with one node for each textual label).

For static patterns (e.g., simple steady vowels), ANNs have provided excellent classification when properly trained. However, general ASR must handle dynamic speech, and thus the practical success of ANNs has been more limited. So-called time-delay ANNs allow complicated feedback within ANN layers to try to account for the fact that small delays (variability in timing) in speech signals occur often in human speech communication and do not hinder perception. In a basic ANN, however, shifting the set of inputs can cause large changes in the ANN outputs. As a result, for ASR, ANNs have typically found application, not as a replacement for HMMs, but as an additional aid to the basic HMM scheme (e.g., ANNs can be used in the training phase for better estimation of HMM parameters).

6. SPEAKER VERIFICATION

We have shown speech analysis to be useful in coding, where speech is regenerated after efficient compression, and in ASR, where speech is converted to text. Another application for speech analysis is speaker verification, where one determines whether speakers are who they claim to be. Fingerprints or retinal scans are often more accurate in identifying people than speech-based methods, but speech is much less intrusive and is feasible over the telephone. Many financial institutions, as well as companies furnishing access to computer databases, would like to provide automatic customer service by telephone. Since personal number codes (keyed on a telephone pad) can easily be lost, stolen, or forgotten, speaker verification may provide a viable alternative. The required decision process can be much simpler than for ASR, since a given speech signal is converted into a simple binary digit ("yes," to accept the claim, or "no," to refuse it). However, the accuracy of such verification has only relatively recently reached the point of commercialization.

In ASR, for speech signals corresponding to the same spoken text, any variation due to different speakers is viewed as "noise" to be either (1) (ideally) eliminated by speaker normalization or (2) (more commonly) accommodated through models based on a large number of speakers. When the task is to verify the person talking, rather than what that person is saying, the speech signal must be processed to extract measures of speaker variability instead of segmental acoustic features. What to look for in a speech signal for the purpose of distinguishing speakers is much more complex than for ASR. In ASR we look for spectral features to distinguish different fundamental vocal tract positions, and we can exploit language models to raise accuracy.

One common speaker verification approach indeed employs ASR techniques, with the exception of labeling models with speaker names instead of word or phoneme labels. When two speakers utter the same word or phoneme, spectral patterns are compared to see which speaker provides the more precise match. In ASR, competing phoneme or word candidates can often be distant in spectral space, making the choice relatively easy. In speaker verification, however, many people have quite similar vocal tract shapes and speaking styles; using analysis over many frames, however, renders the task feasible.

For speech recognition, much is known about the human speech production process, which links a text and its phonemes to the spectra and prosodics of a corresponding speech signal. Each phoneme has specific articulatory targets, and the corresponding acoustic events have been well studied (but still are not fully understood). For speaker verification, on the other hand, the acoustic aspects of what characterizes the principal differences between voices are difficult to separate from signal aspects that reflect ASR. There are three major sources of variation among speakers: differences in vocal cords and vocal tract shape, differences in speaking style (including variations in both target positions for phonemes and dynamic aspects of coarticulation such as speaking rate), and differences in what speakers choose to say. Automatic speaker verification exploits only the first two of the variation sources, examining low-level acoustic features of speech, since a speaker's tendency to use certain words and syntactic structures (the third source) is difficult to exploit (e.g., impostors could easily mimic a speaker's choice of words, but would find simulating a specific vocal tract shape or speaking style much more difficult).

Unlike the relatively clear correlation between phonemes and spectral resonances, there are no evident acoustic cues specifically or exclusively dealing with speaker identity. Most of the parameters and features used in speech analysis contain information useful for the identification of both the speaker and the spoken message. The two types of information, however, are coded quite differently in the speech signal. Unlike ASR, where decisions are made for every phoneme or word, speaker verification requires only one decision, based on parts or all of a test utterance, and there is no clear set of acoustic cues that reliably distinguishes speakers. Speaker verification typically utilizes long-term statistics averaged over whole utterances or exploits analysis of specific sounds. The latter approach is common in "text-dependent" applications where utterances of the same text are used for both training and testing.

There are two classes of errors in speaker verification: *false rejections* and *false acceptances*. A false rejection occurs when the system incorrectly rejects a true speaker. A false acceptance occurs when the system incorrectly accepts an impostor. The decision to accept or reject usually depends on a threshold: if the distance between a test speech pattern and a reference pattern exceeds the threshold (or equivalently, using a PDF for the claimed speaker, the evaluated probability value is too low), the system rejects a match. Depending on the costs of each type of error, verification systems can be designed to minimize an overall penalty by biasing the decisions in favor of less costly errors. Low thresholds are generally preferred because false acceptances are usually more expensive (e.g., admitting an impostor to a secure facility might be disastrous, while excluding some authorized customers is often merely annoying). System parameters are often

adjusted so that the two types of error occur equally often (*equal error rate condition*).

One popular method is the *Gaussian mixture model* [16], which follows the popular modeling approach for ASR, but eliminates separate HMMs for each phoneme as unnecessary. It is essentially a one-state HMM, but allows the PDF to be complicated. In ASR, we can often justify that each state's PDF may be approximated as approaching a Gaussian distribution (especially for triphone HMMs, where context is well controlled). When merging all speech from a speaker into one PDF, however, that PDF is far from Gaussian. To retain the simple statistics of Gaussians (complete specification with only the mean and variance), however, both ASR and speaker verification often allow a state's PDF to be a weighted combination of many Gaussian PDFs. Using the same set of Gaussians for all speakers, each speaker is characterized simply by the corresponding set of weights. Evaluation is thus quite rapid, even when all the frames of an utterance contribute to the overall speaker decision.

Such an approach ignores dynamic speech behavior (as well as ignoring intonation, as do most ASR systems), and further assumes that both training and testing utterances will be roughly similar in phonetic composition. This latter assumption is ensured when doing "text-dependent" verification, where the speaker is asked to utter words already made known to the system during the training phase; for instance, a commercial system may train on two-digit numbers (e.g., "74") and then ask each candidate speaker to utter a few such numbers at test time. False rejections are much lower for these systems than for "text-independent" verifiers, although the latter are sometimes needed for forensic work (e.g., there is no text control in wiretapped conversations). The former, however, run the risk of an impostor playing a recording of the desired speaker (over the telephone, or if there is no camera surveillance at a secure facility). Such impostors, of course, elevate false-acceptance rates with "text-independent" verifiers (which have lower accuracy) as well, but impostors would have difficulty anticipating the requested text that customers are asked to speak, in the case of the much larger vocabulary in the latter systems.

In any pattern recognition task (ASR included), training and test data must be kept separate (i.e., testing on data already used for training artificially raises performance to levels that will usually not be replicated in practical testing situations), but this is even more important in the case of speaker verification. Speakers tend to vary their style of speech over time (e.g., morning versus evening, Monday–Saturday, healthy–ill). Training a speaker verifier in only one session, and then testing a few months later will usually not give good results. Ideally, speaker verification needs months of training (or at least an adaptive system, which allows for periodic updates while being used). If the same utterances are used to train and test a system, a high degree of accuracy is due to the model parameters often being heavily tuned toward the training data. For good results, the training data must be sufficiently diverse to represent many different possible future input speech signals. With shared training and test data, it is difficult to know whether the system has been designed to take advantage of specific speech or speaker characteristics that may not be repeated in new data. Given K utterances per speaker as data, one common procedure trains its system using $K-1$ as data and one as test, but repeats the process K times treating each utterance as test once. Technically, this "leave one out" method designs K different systems, but this method verifies whether the system design is good using a limited amount of data, while avoiding the problems of common train and test data.

7. TEXT-TO-SPEECH SYNTHESIS (TTS)

Our final speech processing application concerns the automatic generation of speech from text [17–19]. Here, speech processing occurs both in the development stage of the system, when spoken units from one or more speakers are recorded, and again at synthesis time, when selected stored units are concatenated to produce the synthetic output voice. The units involved can range from full sentences (e.g., as in talking cars and ovens) to shorter phrases and words (e.g., telephone directory assistance) to phonemes (e.g., for unlimited text applications).

We distinguish true TTS systems, which accept any input text in a chosen language (this could include new words and typographical errors), from "voice response systems," which accept a very limited vocabulary and are essentially voice coders of much simpler complexity (but suffer from inflexibility). Commercial synthesizers have now become much more widespread owing to both advances in computer technology and improvements in the methodology of speech synthesis.

The simplest applications requiring only small vocabularies are speech coders that merely play back the speech when needed. This is impossible for general TTS applications, because no one training speaker can provide all possible utterances. For TTS, small units (typically, phonemes or diphones) are concatenated to form the synthetic speech, and significant adjustments must be done at unit boundaries to avoid unacceptable disjointed (jumpy) speech. The quantity and complexity of such adjustments vary in indirect proportion to the unit size, which in turn also varies inversely with quality (e.g., general text-to-speech using smaller units sounds less natural).

The critical issues for synthesizers concern tradeoffs among the conflicting demands of simultaneously maximizing speech quality, while minimizing memory space, algorithmic complexity, and computation speed. While simple TTS is possible in real time with low-cost hardware, there is a trend toward using more complex programs (tens of thousands of lines of code, and megabytes of storage). Real-time TTS systems produce speech that is generally intelligible, but lacks naturalness.

Most synthesizers reproduce speech of bandwidth ranging from 300–3000 Hz (e.g., for telephone applications) to 100–5000 Hz (for higher quality). A spectral range up to 3 kHz is sufficient for vowel perception because vowels are adequately specified by the lowest three formants. The perception of some consonants, however, is slightly impaired when energy in the 3–5-kHz range is omitted. Frequencies above 5 kHz help improve speech naturalness but seldom aid speech intelligibility. If we assume

that a synthesizer reproduces speech up to 4 kHz, a rate of 8000 samples/s is needed. Since linear PCM requires 12 bits/sample for toll-quality speech, storage rates near 100 kbps result, which are acceptable only for synthesizers with very small vocabularies.

The memory requirement for a simple synthesizer is often proportional to its vocabulary size. Continued decreases in memory costs have allowed the use of very simple TTS systems. Nonetheless, storing all possible speech waveforms for synthesis purposes (even with efficient coding) is impractical for TTS. The sacrifices usually made for large-vocabulary synthesizers involve simplistic modeling of spectral dynamics, vocal tract excitation, and intonation. Such modeling usually causes quality limitations that are the primary problems for current TTS research.

7.1. The Steps in Producing Speech from Text

TTS requires several steps to convert the linguistic textual message into an acoustic signal. The linguistic processing is, in a sense, an inverse of the procedures for ASR. First, a "preprocessing" stage "normalizes" the input text so that it is a series of spelled-out words (retaining any punctuation marks as well). All abbreviations and digits are converted to words, typically via a lookup table or simple programs (sophisticated systems however might distinguish ways of saying "$19.99" and "1999"); word context may be necessary to handle cases such as "St. Peter St." The words are then converted into a string of phonemes (and basic intonation parameters), usually via a combination of a dictionary and pronunciation rules. Reduced memory costs have led to use of large dictionaries containing all the words in a given language and their phonemic pronunciations. With a dictionary, additional relevant information can be readily available, including lexical stress (which syllable in each word is stressed), part of speech, and even semantics.

The alternative approach of letter-to-phoneme rules is useful to handle new, foreign, and mistyped words (cases where simple dictionaries fail). Complex languages such as English (derived from both Romance and Germanic languages) require many hundreds of these rules to pronounce letter sequences correctly (e.g., consider "-ough-:" rough, cough, though, through, thought, drought). Languages are often much simpler phonetically; for instance, Spanish employs just one rule per letter. In all cases, developing TTS capability for a language requires establishing a dictionary and/or pronunciation rules. Hence commercial TTS exists only for about 20 of the world's languages, whereas many commercial ASR systems (those based on word units and without a language model) can handle virtually all languages, since most languages employ similar versions of stops, fricatives, vowels, and nasals.

The next TTS step is intonation specification, determining a duration and amplitude for each phoneme, as well as a complete F0 pattern for the utterance. This is much more difficult than the prior two TTS steps, and requires a syntactic and semantic analysis of the input text. Most languages use intonation in complex ways to cue many aspects of speech communication. Many languages (including English) stress only a small number of syllables in any utterance. This involves not just simple lexical stress from a dictionary but also a judgment about the semantic importance of the words in a sentential context; for synthetic speech comparable to that of humans, this would require a natural-language processor typically beyond current capabilities. People often cue syntactic structure via intonation; for example, in English, long word phrases often start with an F0 rise and end with a fall. Questions in many languages are cued with a large final F0 rise if they request a yes/no answer (but not questions with the "wh-words": what, when, why, who, how). Tone languages (e.g., Chinese, Thai) employ four or five different F0 patterns on syllables to distinguish different words with the same phoneme sequence. Finally, speech uttered with emotion often changes intonation significantly.

In the last TTS step, speech units are concatenated using the specified intonation, and adjustments are made to the model parameters at unit boundaries. Few such manipulations are needed for phrasal concatenation, but smaller units require at least smoothing of all parameters across the boundary for several frames. This is relatively straightforward when the units contain spectral parameters (e.g., LPC coefficients or formants), although improved quality occurs with more complicated smoothing rules. Storing small units with waveform coding (e.g., log PCM or ADPCM) is often not suitable (despite the higher general quality of such speech) here because smoothing the available parameters does not approximate well the actual coarticulation and F0 manipulation found in human speech production.

Smoothing of parameters at the boundaries between concatenated units is most important for short units (e.g., phones) and decreases in importance for larger units owing to fewer boundaries. Smoothing is much simpler when the joined units approximately correspond at the boundaries. Since diphone boundaries link spectra from similar sections of two realizations of the same phoneme, smoothing rules for their concatenation are simple. Systems that link phones, however, must use complex smoothing rules to represent coarticulation in the vocal tract. Not enough is understood about coarticulation to establish a complete set of rules to describe how the spectral parameters for each phone are modified by its neighbors. Diphone synthesizers circumvent this problem by storing parameter transitions from one phone to the next, since coarticulation influences primarily the immediately adjacent phones. However, since coarticulation often extends over several phones, using only average diphones or those from a neutral context leads to lower-quality synthetic speech. Improved quality is possible by using multiple diphones dependent on context, effectively storing "triphones" of longer duration (which may substantially increase memory requirements). Some coarticulation effects can be approximated by simple rules, such as lowering all resonant frequencies during lip rounding, but others such as the undershoot of phoneme target positions (which occurs in virtually all speech) are much harder to model accurately.

It is in this last TTS step that the system yields poorer quality than many speech coders, because TTS is often

forced to employ synthetic-quality coding techniques. The traditional excitation for vocal tract filters in TTS (either LPC or formant-based approaches, which constitute most commercial methods) is a simple train of periodic pulses (for voiced speech) or white noise (a random-number generator) for unvoiced speech. The combination of oversimplified excitation and the limited modeling accuracy of the LPC or simple formant models leads to intelligible, but slightly unnatural, synthetic speech.

Since the late 1980s, some commercial systems have been successful in concatenating waveform-coded small units, with limited amounts of perceptually annoying spectral jumps at unit boundaries; for instance, the PSOLA (pitch-synchronous overlap and add) method outputs successive smoothed pitch periods [20]. While such speech can sound more natural, it is inflexible in producing alternative voices. It is typically based on one speaker uttering a large inventory of diphones; for a language such as English with about 32 phonemes, about 1000 diphones must be uttered and in a uniform fashion. One cannot simply adjust some synthesizer parameters here to get other synthetic voices (as is possible with formant synthesizers).

8. CURRENT TECHNOLOGY

Today's speech coders deliver toll quality (i.e., equivalent to the analog telephone network) at 8 kbps with minimal amounts of delay (and even at 4 kbps, if delay is less of an issue). The favored approach is CELP, for which there are several standards accepted internationally (e.g., in digital cellular telephony). The digital links in most telephone networks still employ simpler 64 kbps log PCM coding; 24–32 kbps ADPCM and delta modulation are also still popular because of their relative simplicity. We are still far from an ultimate limit of perhaps 100 bps to code speech. Despite the reducing cost of computer memory and speed, further research in coding is needed because of the increasing use of wireless telephony, where limited bandwidth is very much an issue.

The lack of formal standards for human–machine applications of speech (i.e., synthesis and recognition) hinders comparison of commercial systems. Several companies offer unlimited text-to-speech for several languages (typically the major European languages, plus Japanese and Chinese). Such synthetic speech is largely intelligible, but is easily discerned as synthetic and lacking naturalness. All synthesizers suffer from our lack of understanding of the complex relationships between text and intonation. Synthesizers are clearly increasing in use, but wider public acceptance awaits further improvements in quality.

Speech recognizers are also increasing in use, but their severe limitations (compared to human speech perception) have also hindered wider acceptance. The need to pause between words, restrict the choice of words, and/or do prior training, as well as frequent recognition errors, have significantly limited the use of ASR. Systems eliminating all these restrictions (i.e., continuous, speaker-independent, large-vocabulary ASR) still suffer from high cost and frequent errors, especially if they are used in noisy or telephone environments, and such conditions occur often in practical applications. Progress in ASR has been attributed more to general improvements in computers and to the wider availability of training data, than to algorithmic breakthroughs. The basic HMM approach using MFCCs was developed largely before 1985. Even more recent additions, such as delta coefficients, mean subtraction, Gaussian mixtures, and language models, have been in wide use since 1990.

Future systems will likely integrate more structure into the stochastic approach. It is clear that the expert-system approach to ASR common in the early 1970s will never replace stochastic methods, for the simple reason that individual human phoneticians can never assimilate enough information from hundreds of hours of speech, in ways that probabilistic computer models can improve with larger amounts of training data. The extremely simple stochastic models in current widespread ASR use, however, are too unstructured and allow too much freedom (similar complaints hold for recent neural network approaches to ASR). For example, the MFCCs, while appropriately scaling the frequency axis to account for perceptual resolution, do not take account of the wide perceptual difference between resonances and spectral valleys. First-order HMMs ignore the high degree of correlation across many frames of speech data (compensating by using delta coefficients is only a very rough use of speech dynamics). Intonation is widely ignored (e.g., F0) or treated as noise (e.g., durational factors), despite evidence of its use in human speech perception. Of course, in a practical world, you use whatever works, and current systems, despite their flaws, provide sufficiently high accuracy for small vocabularies (e.g., recognizing the digits in spoken telephone or credit card numbers, or controlling computer menu selections via voice). More widespread use of speech in telephone dialogs will await advances in both the quality of synthetic speech and the recognition accuracy of spontaneous conversations.

BIOGRAPHY

Douglas O'Shaughnessy has been a professor at INRS-Telecommunications at the University of Quebec and adjunct professor at McGill University in Montreal, Québec, Canada, since 1977. His interests include automatic speech synthesis, analysis, coding, and recognition. He has been an associate editor for the Journal of the Acoustical Society of America since 1998, and will be the general chair of the 2004 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in Montreal, Canada. Dr. O'Shaughnessy received all his degrees from the Massachusetts Institute of Technology (MIT), and is a fellow of the Acoustical Society of America.

He is the author of the textbook *Speech Communications: Human and Machine* (IEEE Press, 2000).

BIBLIOGRAPHY

1. J. Deller, J. G. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
2. P. Noll, Digital audio coding for visual communications, *Proc. IEEE* **83**: 925–943 (1995).

3. J.-P. Adoul and R. Lefebvre, Wideband speech coding, in W. Kleijn and K. Paliwal, eds., *Speech Coding and Synthesis*, Elsevier, New York, 1995, Chap. 8.
4. J. Johnston and K. Brandenburg, Wideband coding — perceptual considerations for speech and music, in S. Furui and M. Sondhi, eds., *Advances in Speech Signal Processing*, Marcel Dekker, New York, 1992, pp. 109–140.
5. A. Gersho, Advances in speech and audio compression, *Proc. IEEE* **82**: 900–918 (1994).
6. D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley, Reading, MA, 1987.
7. A. Spanias, Speech coding: A tutorial review, *Proc. IEEE* **82**: 1541–1582 (1994).
8. P. Kroon and W. Kleijn, Linear predictive analysis by synthesis coding, in R. Ramachandran and R. Mammone, eds., *Modern Methods of Speech Processing*, Kluwer, Norwell, MA, 1995, pp. 51–74.
9. R. Cox and P. Kroon, Low bit-rate speech coders for multimedia communication, *IEEE Commun. Mag.* **34**(12): 34–41 (1996).
10. W. Hess, Pitch and voicing determination, in S. Furui and M. Sondhi, eds., *Advances in Speech Signal Processing*, Marcel Dekker, New York, 1992, pp. 3–48.
11. L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
12. J. Piconi, Signal modeling techniques in speech recognition, *Proc. IEEE* **81**: 1215–1247 (1993).
13. J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer, Norwell, MA, 1996.
14. Y. Ephraim, Statistical-model-based speech enhancement systems, *Proc. IEEE* **80**(10): 1526–1555 (1992).
15. N. Morgan and H. Bourlard, Continuous speech recognition, *IEEE Signal Process. Mag.* **12**(3): 25–42 (1995).
16. D. Reynolds and R. Rose, Robust text-independent speaker indentification using Gaussian mixture speaker models, *IEEE Trans. SAP* **3**: 72–83 (1995).
17. D. Klatt, Review of text-to-speech conversion for English, *J. Acoust. Soc. Am.* **82**: 737–793 (1987).
18. J. van Santen, Using statistics in text-to-speech system construction, *Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis*, 1994, pp. 240–243.
19. T. Dutoit, *From Text to Speech: A Concatenative Approach*, Kluwer, Norwell, MA, 1997.
20. E. Moulines and F. Charpentier, Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Commun.* **9**: 453–467 (1990).

SPEECH RECOGNITION

JAYADEV BILLA
BBN Technologies
Cambridge, Massachusetts

1. INTRODUCTION

Speech is a natural and preferred form of communication among humans. Automatic speech recognition attempts to extend this mode of communication to human–machine interaction. As a field of study, automatic speech recognition has a history extending back to the late 1960s. Over these years our understanding of the process of speech communication has increased substantially. This greater understanding, coupled with an even more developed ability to harness powerful computational models to encapsulate our knowledge of speech communication, has allowed for the development of increasingly capable speech recognition systems. Today, in some specialized applications, state-of-the-art speech recognition systems are capable of recognition and transcription of speech at performance levels that rival human ability.

In this article, we introduce the problem of automatic speech recognition and describe the reasons that make speech recognition by machine difficult. We also describe the science behind automatic speech recognition and discuss the principles and architecture of a typical large vocabulary speech recognition system. We then provide an overview of current state-of-the-art speech recognition research and applications, and conclude with a description of current trends in speech recognition system research and development.

2. WHY SPEECH RECOGNITION IS DIFFICULT

The ease with which humans use speech understates the complexity of this task for machines. The fundamental difficulty with speech recognition is the overwhelming variability in the production of speech. Indeed, if everyone always spoke in exactly the same way so that each word, whether spoken by the same person or different people, always sounded exactly the same, then we could simply store all the words and recognize speech by a simple comparison with the words we had heard and stored earlier. However, this is not the case; there is a considerable amount of variation in spoken speech even when the same speaker repeats the same sentence. Some forms of variability include those arising in the speaker due to mood, stress, and state of health, and those arising from the environment such as background noise or room acoustics. Other causes of variability in spoken speech include those due to gender, age, accent, or speaking rate among speakers and due to language differences, such as regional dialects, or formal versus conversational speaking style.

Figure 1 shows the speech signal and spectrogram for the utterance "gray whales". Spectrograms are pictures of the distribution of energy in frequency over time. The dark horizontal bars correspond to resonances of the vocal tract, and change their vertical frequency positions over time, depending on the sound being produced as well as on neighboring sounds. The dependency of each sound on its neighboring sounds further increases the number of ways in which a particular word can be pronounced. If this were not enough, normal human speech is a continuous stream of words without any interword separation, unlike the clearly separated words of written text. As an example of the difficulty in word boundary identification, consider spoken instances of the phrases "recognize speech" and "wreck a nice beach."

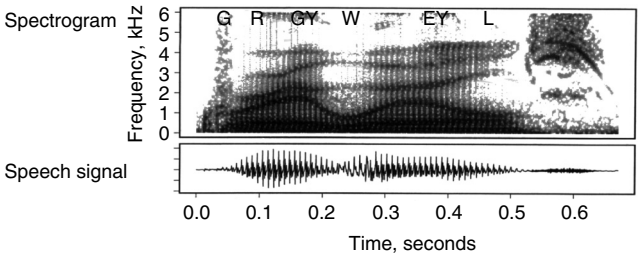


Figure 1. A spoken instance of the utterance "gray whales." The speech signal (bottom) is displayed as a simple amplitude plot over time. The spectrogram (top) shows the energy distribution of the utterance over frequency and time. The horizontal dark areas correspond to the resonances of the vocal tract.

An automatic speech recognition system must successfully address these variabilities in speech before any useful output can be provided to the user. Current understanding of the speech production process and the semantic and grammatic structure of language provides a means to model some of the variability. We are, however, limited by an incomplete knowledge of the process of speech communication. We do not really know how humans organize acoustic information and store knowledge of language; not do we know how cognitive processes are formed and used to perform recognition tasks. Current automatic speech recognition systems reflect these shortcomings and attempt to fit mathematical models to account for our incomplete understanding of the speech recognition problem. Tremendous advances in speech recognition technology have come from our ability to build sophisticated models that compensate for this lack of knowledge. Nevertheless, the lack of complete understanding of the speech communication process is an additional obstacle for automatic speech recognition systems to overcome, in addition to the inherent difficulty of the speech recognition task.

Moreover, automatic speech recognition systems must work effectively over the majority of acoustic environments that the systems may encounter. The difficulty in producing robust speech recognition systems that can work effectively over different acoustic environments is the primary reason for the limited use of such systems. Advances since 1980 have enabled us to build and deploy speech recognition systems for tasks ranging from responses to interactive voice response (IVR) systems with prompts of the type, "please say or press one to select choice one for ..., say or press two" to highly accurate transcription of dictation in certain specialized areas. The corresponding research recognition systems are close to achieving real-time transcription of general English such as in news broadcasts. Figure 2 charts commercial speech recognition system complexity, measured by speaking mode and vocabulary size, and its corresponding application area over time, against a projection for the near future. Early commercial speech recognition systems possessed vocabularies of several words that were spoken in isolation. More recently deployed speech recognition systems have had vocabularies in the tens of thousands of words and are capable of recognizing fluent speech. In the future, commercial systems are likely to recognize normal conversational speech with vocabulary sizes approaching that of humans.

3. THE SCIENCE BEHIND AUTOMATIC SPEECH RECOGNITION

The term *recognition* in the context of speech recognition is typically framed as a classification problem. Classification is when one has a finite set of possible outcomes, such as in a multiple-choice question, and attempts to classify an object or event of interest as one of these possibilities based on some observations. In automatic speech recognition systems, classification is usually considered a statistical pattern recognition problem. In the context of speech recognition, this involves building mathematical models of spoken speech and using them to identify the most likely sequence of words in the vocabulary as the recognized sentence or phrase.

A typical speech recognition system, illustrated in Fig. 3, consists of three main components: a *feature extraction stage*, where one extracts a set of speech features that can minimize some of the variability in speech without loss of information; a *training stage*, where one builds a set of mathematical models of speech; and a *recognition stage*, where one uses the trained models to make the classification of the spoken speech into a word or sentence.

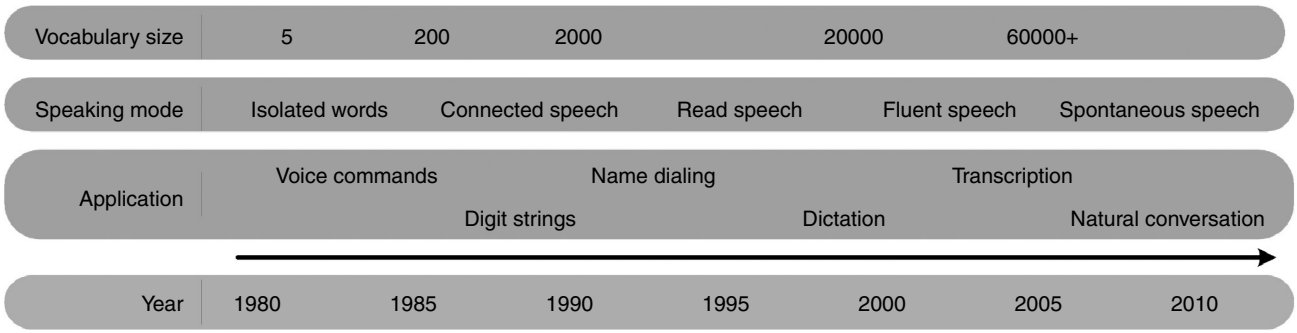


Figure 2. Timeline of the increase in speech recognition system complexity and its corresponding area of application. Early systems were capable of recognition of a few words spoken in isolation. Future systems are likely to have vocabularies sizes similar to human vocabularies and capable of recognition of natural spoken language.

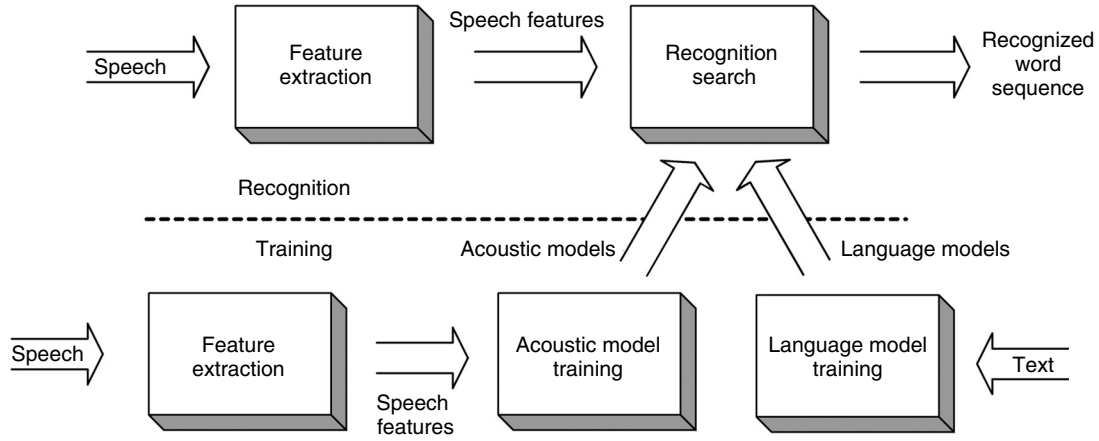


Figure 3. Components of a typical speech recognition system. A speech recognition system must first be trained (bottom) to generate models of speech. Recognition proceeds by comparing the input speech to the speech models to determine which speech model is closest to the input speech.

The catalog of mathematical models with which we compare features of a spoken word or sentence must be created beforehand, and often involves *levels of knowledge* akin to the many ways in which humans use prior knowledge to recognize speech. Typically, automatic speech recognition systems use at least two levels of knowledge: an *acoustic model* describing the actual acoustics of the speech signal, and a *language model* describing what word follows or is most likely to follow the current word or set of words.

The acoustics of a spoken sentence can be decomposed into a sequence of progressively finer acoustic events such as words, syllables, allophones, and phonemes. Fig. 4 shows the decomposition of the utterance "gray whales" into allophones and phonemes. The term *phoneme* has its origins in linguistics and is defined as the smallest unit of speech that can distinguish one word from another, such as *bit* versus *pit*. An allophone, on the other hand, is one of two or more variants of the same phoneme such as the phoneme \p\ in *pin* and *spin*. Each of these acoustic events can be modeled individually, and recognition can be achieved by identifying the appropriate sequence of acoustic events. The fundamental acoustic event is often determined by the application and by the amount of training data. For example, if we intend to build a recognizer that recognizes the 10 digits, 0–9, we can simply build a model for each of these digits. It is likely, in this case, that we will have many instances of each digit spoken by many people with which to build a representative model. On the other hand, if we intend to build a speech recognition system for general dictation with a vocabulary of 60,000 words, it is unlikely that we will have many spoken instances for most of the words in the vocabulary. In these cases, it is advisable to build phonetic models as the number of phonemes is much smaller than the number of words and therefore far more likely to have sufficient training instances. These models of acoustic events encapsulate the characteristics of the actual spoken speech signal and form the acoustic model.

Words	Gray			Whales			
Phonemes	-g	r	ey	w	ey	l	z-
Allophones	- [g] r	g [r] ey	r [ey] w	ey [w] ey	w [ey] l	ey [l] z	l [z] -

Figure 4. Decomposition of the utterance "gray whales" into words, phonemes, and allophones. Depending on the application and availability of training data, any of these acoustic events can be modeled individually. Recognition is then achieved by identifying the appropriate sequence of acoustic events.

Acoustic models are created by organizing labeled samples of speech, specifically, spoken speech and its corresponding text, using a training paradigm. Most often, acoustic modeling is attempted via statistical models that are capable of automatically modeling the key distinguishing features of their training data. This ability to "learn" the key features from highly variable input is crucial to the application of statistical models to the speech recognition problem. Of these statistical models, hidden Markov models (HMMs) are the most widely used methodology for acoustic modeling. HMMs possess many interesting properties and are supported by elegant algorithms that allow for their application to the speech recognition problem.

A language model, in contrast, determines valid linguistic constructs and encapsulates the ways in which words are connected to form phrases and sentences. An alternate description of the language model is that it specifies the grammar that the speech recognizer must adhere to when producing a recognition result. Depending on the intended area of application of the speech recognition system, language modeling can be one of two types: rule-based or statistical. Rule-based language models are used when there is a rigid structure to the spoken sentence, such as a credit card number spoken by a person, and are created by writing explicit rules that specify allowable sentences. Statistical language models are used when there is no rigid structure to the spoken utterances, such as in normal conversational speech. Unlike rule-based language models, statistical language models do not explicitly specify all allowable sentences, but assign scores to each sequence or group of words,

proportional to the frequency that a particular sequence or group of words occurs in the training data.

3.1. Feature Extraction

The goal of feature extraction is to translate the raw speech waveform into a set of features that effectively capture the salient aspects of the speech signal and at the same time reduce the effect of variability due to speaker or environment in the final representation. There is a considerable amount of information in the speech signal. Some information, such as the resonances of the vocal tract that correspond to vowel-like sounds in spoken words, is crucial to speech recognition. Other information, such as gender and mood of the speaker or the acoustic environment, provides no useful information for speech recognition.

Information in a speech signal is encapsulated mostly in the energy and in the frequency distribution of that energy. The spectral characteristics of speech change rapidly over time as different words are spoken. To capture these rapid changes in speech over the course of a word or sentence, features must be generated sequentially over the duration of the utterance. Furthermore, features must be generated on sufficiently short time segments so that the spectral characteristics are relatively invariant. Typically, small Fourier analysis algorithms such as the fast Fourier transform (FFT) are used to perform the spectral analysis of the speech segment. Since the analysis is performed on relatively short segments of speech, this mode of analyzing speech is often referred to as *short-time Fourier analysis*.

The source-filter model of speech production provides a convenient parametric model for feature extraction. Parametric modeling reduces the representation of a signal to the estimation of model parameters that, in turn, are likely to form a much more compact representation of the signal. Figure 5 provides a graphical view of a model of human speech production and its source-filter model equivalent. The idea behind the source-filter model is that speech is a result of airflow (source) being shaped by the vocal tract (filter). The type of source excitation controls the realization of different types of speech. For example, voiced speech, which includes sounds with strong periodic structure such as vowels, can be generated with periodic source excitation. Unvoiced speech, which encompasses sounds with little or no periodic structure such as the \s\ in *set*, is generated with white-noise excitation. Further, it is assumed that information is carried primarily by the shape of the vocal tract rather than in the airflow, and that variability in airflow is the primary source of variability among speakers. These simplifying assumptions render the feature extraction problem mathematically tractable and allow the immediate applicability of many techniques from signal processing theory.

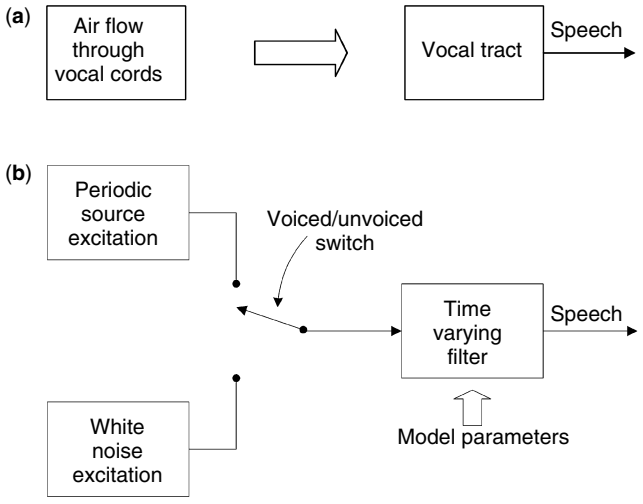


Figure 5. A model of human speech production and its corresponding source-filter model equivalent: (**a**) human speech production; (**b**) speech production model.

Figure 6 illustrates a typical speech feature extraction paradigm. From basic Fourier analysis, given a source-filter model, one can consider the speech spectrum to be the product of the excitation spectrum and the filter spectrum. Furthermore, if one takes the logarithm of the speech spectrum and transforms it back to the time domain (by way of the inverse Fourier transform) one can represent the excitation and vocal tract separately. The result of this inverse transform of the log spectrum is the *cepstrum*. The term *cepstrum* is actually derived from "spectrum" spelled with the first four letters in reverse order as *c-e-p-s*-trum to signify the inversion of the spectrum. Excitation and vocal tract have very different spectral characteristics; excitation is usually located in the higher order terms and the vocal tract in the lower-order terms of the cepstrum. The noninformative excitation can be easily discarded by considering only the lower-order cepstral terms.

This basic paradigm can be further enhanced by incorporating information from psychoacoustic studies. For example, the human auditory system is more sensitive to lower-frequency components than higher-frequency components. The differential emphasis of frequency content can be incorporated into the feature extraction paradigm by warping the frequency scale to compress the higher frequencies relative to the lower frequencies. This allows us to place more emphasis on low-frequency components over high-frequency components. Common

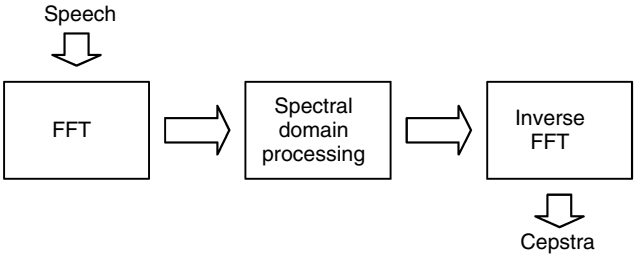


Figure 6. Block diagram of a typical speech feature extraction paradigm. Speech is first transformed into the spectral domain with an FFT algorithm. Various techniques, such as log compression and frequency warping, are used to emphasize or deemphasize characteristics of the speech signal depending on their importance for speech recognition. Finally, the modified spectrum is inverted using an inverse FFT algorithm to yield cepstra. The lower order ceptral terms typically form the speech features used during speech recognition.

warping strategies include mel-scale warping and bark-scale warping, both of which are approximately linear to about 1000 Hz and logarithmic thereafter.

Excellent overviews of the various techniques for speech feature extraction can be found in the literature [1,2].

3.2. Model Training

Speech model training allows us to encapsulate the acoustics of the speech signal and the manner in which words are connected to form phrases and sentences. As mentioned earlier, the acoustics of the speech signal are captured in the acoustic model, and the language and word structure in the language model.

3.2.1. Acoustic Modeling. Acoustic modeling in current speech recognition systems is most commonly based on *hidden Markov models* (HMMs), statistical models that are capable of automatically extracting statistically significant information from available speech data. In HMM theory, the distance or closeness to incoming speech is measured in terms of the probability that the input speech could be generated by that model. This probabilistic approach allows us to absorb the variation in speech features and provides a powerful paradigm for speech recognition.

Often in physical processes, the current state of the process has a significant influence on subsequent events. This concept is embodied in the Markov property, which states that given the current state of the process or system, the future evolution of the system is independent of its past. Models based on the Markov property are called *Markov models*. Figure 7a shows a simple four-state Markov model of weekly weather. Each state in this Markov model is associated with an output symbol indicative of a possible weather observation: hot, cold, windy, or cloudy. These weather observations are called *output symbols* because the Markov model can be regarded as a generative model; it outputs symbols as a transition is made from one state to the other. The arrows connecting the states indicate allowable transitions, and each number on the arcs corresponds to the *transition probability*: the probability of an arc being taken. For example, the self-arc on the "cloudy" state has a probability of 0.6, indicating that if one makes the observation that it was "cloudy" this week, then there will be a 60% chance of making the observation that the weather is still "cloudy" the following week. Note that in the Markov model, the transition from one state to another is probabilistic, but the production of the output symbol is deterministic and known. Also note that at any given instance the current state completely determines the set of all possible states to transition to without regard to how one transitioned into the current state. For example, in the Markov model of Fig. 7a, if the current weekly weather corresponds to the state "hot," next week's weather can only be hot, windy, or cold and this holds true irrespective of whether last week was observed as cloudy, hot, or cold.

As it turns out, this definition of Markov models is too restrictive to be useful in many interesting problems. In this example, it is too restraining to confine the output symbols from any state to correspond to a particular weather observation. Indeed, it is far more meaningful to describe the weekly weather as some combination of the four weather patterns: hot, cloudy, cold, or windy. In *hidden* Markov models (HMMs) the output of each state is associated with an output probability distribution rather than a deterministic and known outcome. In an HMM all output symbols are possible at each state, albeit with differing probabilities. The probabilities associated with each state are known as output probabilities.

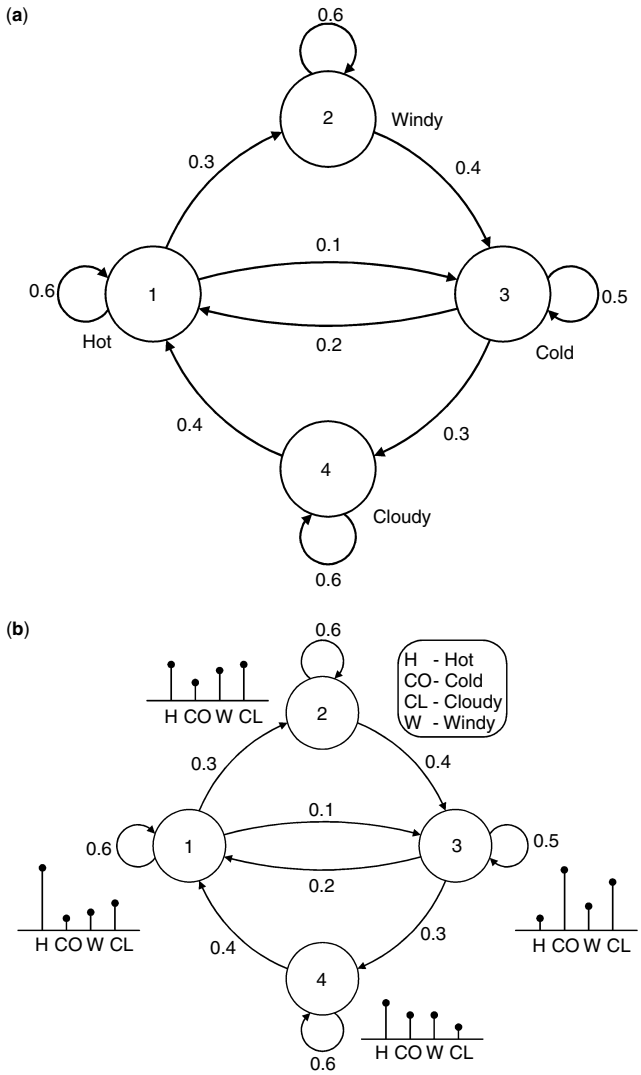


Figure 7. A model of weekly weather patterns. A Markov model of weather is presented in (**a**). A corresponding hidden Markov model (HMM) is presented in (**b**). Note the output probability distributions in the HMM compared to the deterministic single output in the Markov model.

Figure 7b shows the Markov model of Fig. 7a extended to the HMM formulation. The difference is that we now have a probability distribution associated with each state, and when a transition is made into a state, the output symbol is chosen according this probability distribution. Since weather of any type is possible with varying likelihoods in every state, given a sequence of weekly weather readings, the state sequence is unobservable and *hidden*.

The HMM methodology is very well suited for speech recognition and provides a natural and highly reliable way of modeling and recognizing speech. In the case of speech, we can visualize the states as corresponding to functional (or largely unchanging) portions of the word or subword, such as the beginning or end, and the speech features as corresponding to the outputs of the state. HMMs simultaneously model time and feature variability. Time variability is modeled by state-transition probabilities, and feature variability by state output probability distributions. Training the HMM involves estimation of the output state probability distributions and the state-transition probabilities. Figure 8 shows the structure of a typical three-state speech HMM. The structure of a speech HMM has a natural flow of transitions from left to right to indicate the forward time flow of speech. In a typical speech recognition system there is one HMM per phonetic context, to reflect the differing pronunciation of phonemes in different contexts. Although different phonetic contexts could have different structures, the HMM structure is usually kept the same, with HMMs being distinguished from each other by their transition and output probabilities.

Thus far we have referred to neither the language or type of acoustic event that the HMM is modeling. This brings out an important quality of statistical modeling in general, and HMMs in particular: that the HMM is independent of language and type of acoustic event. An HMM does not require any specific changes for language or type of acoustic event.

An HMM can be regarded as a generative model. For example, in Fig. 8, as we enter into state 1, a speech feature is emitted. Based on the transition probabilities out of state 1, a transition can be made back to state 1, 2, or 3, and another speech feature emitted based on the probability distribution corresponding to the state to which the transition was made. This continues until a transition is made out of state 3. At that point, the sequence of emitted speech features corresponds to the phoneme or phonetic context for which the HMM had been constructed.

In a recognition paradigm, the same HMM can be used to solve the reverse problem — given a sequence of speech features, what is the probability that the HMM could have generated this speech feature sequence? In this mode, given a starting speech feature, one can estimate the probability of the feature being generated by the probability distribution of state 1. One can now assume that a transition is made from state 1 to state 2 and estimate the probability that the next speech feature was generated by the probability distribution of state 2. This is continued until the last state is exited. The product of all the encountered output probabilities and transition probabilities gives the probability that this specific state sequence generated the speech feature sequence. For every possible sequence of states, one obtains a different probability value. During recognition, the probability computation is performed for all speech HMMs and all possible state sequences. The state sequence and speech HMM that gives the highest probability is declared to be the recognized phoneme or phonetic context.

Mathematically tractable techniques such as the Baum–Welch reestimation and Viterbi algorithms provide elegant automatic solutions to these problems and have allowed HMMs to be successful in a wide variety of difficult speech recognition applications. In-depth presentations of HMMs and their application to speech recognition can be found in Rabiner and Juang [2] and Huang et al. [3].

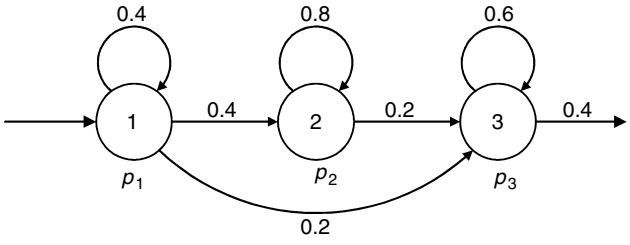


Figure 8. Typical structure of a three-state speech HMM. Numbers on the arcs indicate transition probabilities and p_1, p_2, and p_3 refer to state output probability distributions. HMM training involves the estimation of both the transition and state output probabilities.

3.2.2. Language Modeling. A language model acts as a grammar dictating which word can or cannot follow another word or group of words. Without a language model, every word in the vocabulary of the speech recognition system is equally likely and must be considered at the end of every other word. Language modeling provides a means to limit the vocabulary at each decision point either by eliminating unlikely words or by increasing the probability of more likely words.

There are two approaches to language modeling for speech recognition: rule-based and statistical. *Rule-based language modeling* involves the construction of a set of rules that encompass all allowable sentences. Since the set of rules is explicit and must specify *all* allowable sentences, rule-based language models are primarily used in very restrictive tasks such as IVR systems. A sample rule-based language model for a speech recognition system, used to recognize names for automatic telephone dialing by name, is shown in Fig. 9. Allowable names are John Doe, Jane Smith, Simon Smith, and Mike Phillips with an additional *garbage* name. Since a rule-based grammar completely specifies all the allowable output of the speech recognizer, a valid name is always recognized even when invalid names are spoken. The catchall *garbage* word is often included to prevent the recognition of one of the allowable names when an invalid input is presented to

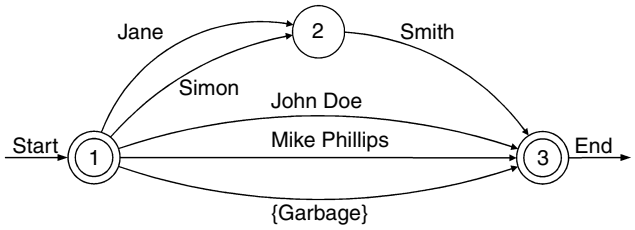


Figure 9. A rule-based language model for recognizing names for an automatic dialing task. The "garbage" name is used to prevent incorrect results with invalid input.

the system. Since rule-based language modeling imposes a rigid structure on the recognized utterance, it provides an excellent means for task automation. In the name-dialer example, the person using the system would be required to say the first and last name in that order. Any valid name can be dialed out simply by looking up the name in the directory. The advantage of rigid structure is also the greatest shortcoming of rule-based language modeling. To describe a speech recognition task of any complexity, it becomes increasingly harder to describe all valid sentences. Consider, for example, designing a rule-based grammar to recognize an interview on television. It would be virtually impossible to describe the myriad ways in which the conversation can even start, let alone proceed.

Statistical language models, on the other hand, are based on probability estimates of the likelihood that one word will follow one or more other words. Because they are estimated directly from training data, statistical models have the advantage of be able to encapsulate the colloquial characteristics of speech in addition to its semantic and syntactic elements. However, statistical language modeling requires a large pool of training data. Without a large training data set, many word sequences are either not observed or are observed far too infrequently to reliably estimate probabilities. Despite this constraint, statistical language models have contributed significantly to improvement in speech recognition system performance, especially in large vocabulary speech recognition tasks, where it is far too complex and virtually impossible to build rule-based language models. A detailed discourse on statistical language modeling can be found Jelinek's treatise [4].

4. CURRENT STATE OF SPEECH RECOGNITION AND FUTURE DIRECTIONS

Automatic speech recognition technology has now advanced to the point where speech recognition systems with vocabularies of a hundred to several thousand words can be robustly deployed in a variety of tasks, ranging from service automation to general consumer applications to the management of spoken information. Table 1 shows the error rates of current high-performance speech recognition systems on various tasks. Observe the increase in error rates as the vocabulary of the speech recognition system and the complexity of the speech increases.

A logical area of deployment has been in the telephony market, where speech is the primary modality for information exchange. Telephony applications include service automation, an early example of which is the automation of operator services with *yes*/*no* responses in the collect call prompt: "You have a collect call from . . ., to accept this call press one or say yes". A telephony application rapidly now gaining popularity is voice dialing. Today, many vendors offer the ability to dial numbers via simple voice commands such as "call home."

Consumer applications of speech recognition have been primarily in the dictation and computer command and control arenas, allowing users to dictate for a variety of word processing tasks or to perform basic computer commands such as "open file" or "close file." Some popular products aimed for the general consumer include Dragon Systems' Dragon NaturallySpeaking, IBM's ViaVoice, and L&H's Voice Xpress.

A relatively new area of application of speech recognition technology has been in the management of spoken information. The conversion of spoken information into textual form allows rapid access to information which would otherwise require the arduous task of listening to large archives of spoken material. An example of an application providing this ability is BBN's Rough 'n' Ready (RnR) audio indexer system [5]. The RnR system is capable of real-time speech recognition of broadcast news, and provides information extraction and indexing coupled with a Web-enabled interface allowing for rapid retrieval of spoken information in a computer accessible form.

Although speech recognition by machine is increasingly used in popular applications, significant issues remain. The translation of speech recognition technology from the laboratory into the "real world" is still fraught with problems. Fielded systems typically show higher error rates than systems evaluated in the laboratory, due mostly to the lack of control over the working environment. To address this vulnerability, current speech recognition systems are normally trained on large amounts of real-world training data. This ensures that, when fielded outside the laboratory, the speech recognition system can handle the unconstrained environment of the real world. Typical real-world training data include recordings of broadcast news such as from CNN or NPR (National Public Radio), and recordings, taken with participants' consent, of telephone conversations between friends and family. Broadcast news recordings include the accompanying commercials, background music, and noise. On the other hand, telephone conversation between family and friends is typical of normal conversational speech with the attendant disfluencies. State-of-the-art speech recognition systems have an average word error rate of close to 10% on broadcast speech, and close to 30% on conversational speech, with vocabularies of up to 65,000 words [6,7].

Table 1. Error Rates of High-Performance Speech Recognition Systems in 2001

Task	Type of Speech (complexity)	Vocabulary Size (Words)	Word Error Rate (%)
Connected digits	Read	10	< 0.3
Automated travel agent	Spontaneous	2,500	< 2
Broadcast news	Read and spontaneous (mixed)	64,000	15
Telephone conversations	Spontaneous and conversational	45,000	35

While the use of real-world training data has enhanced the usefulness of speech technology in everyday applications, significant barriers remain to the seamless integration of speech technology in day-to-day environments. Some issues currently being addressed in speech recognition research laboratories to improve the practical use of speech technology are presented below.

4.1. Speaker and Environmental Adaptation and Robustness

Current speech recognition is often handicapped by the fact that it typically does not run "out of the box." Noise in the environment is a common reason for the failure of speech recognition systems. In some cases, even using a different microphone to capture the speech signal can significantly change the properties of the resultant speech features, with consequent loss in system performance [8]. Speaker and environment adaptation allow an existing speech recognition system to handle new speakers and environment without substantial change in performance. Adaptation can be performed in two different paradigms: supervised, where the system is adapted based on enrollment data, or unsupervised, where the system adapts automatically without any specific input to the system beyond that spoken for the purpose of recognition.

4.2. Conversational Speech

In normal spontaneous spoken conversation, there is a significant amount of disfluency such as hesitations and false starts. In addition, conversational speech commonly suffers from the lack of clear articulation as well as other speaker maladies such as highly variable speaking rate or increased emotional emphasis. These factors combine to cause significant problems for accurate speech recognition. It would not be unusual to see a doubling in error rates by moving from broadcast news to conversational speech recognition task with all other factors remaining the same.

4.3. Speed Versus Accuracy

Another issue with state-of-art speech recognition systems is the large amount of time needed to provide the most accurate recognition result or transcript. Speech recognition systems often need to be tuned to run in real time by sacrificing performance. This is slowly changing due to the ever-increasing speed of computers as well as due to algorithmic improvements. Nevertheless, significant work still needs to be done to allow for real-time speech recognition without losing accuracy.

4.4. Language and Task Independence

Most speech recognition systems are built for specific languages and, more often than not, specific tasks within those languages. This artifact of the specificity of the training that goes into the system invariably requires retraining the system if the language or task changes. The idea behind language and task independence is to lessen the effect of specificity by either generalization of the system or by developing the system's capability to automatically acquire new language or task skills.

4.5. Understanding

After speech recognition, understanding speech is the next frontier in human–machine interaction. Give a system capable of consistently accurate speech recognition, the next logical step would be to use this ability to extract the meaning of the spoken words and provide the appropriate response.

5. CONCLUSION

Speech recognition is perhaps one of a handful of technologies that have the potential to fundamentally change the way we interact with machines. The science of speech recognition continues to develop, and as it does, so will its impact on our lives. Even now, simple applications such as voice dialing for mobile phones or voice controllable car radios are making everyday tasks both safe and convenient. In the future, these and other emerging applications will surely change the nature of human–machine interaction and be responsible for many new and exciting changes to the ways we work and live.

BIOGRAPHY

Jayadev Billa received the B.E. degree in electronics and communication engineering in 1991 from Osmania University, Hyderabad, India, and M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pennsylvania, in 1993 and 1997, respectively. He joined the BBN Technologies, Cambridge, Massachusetts, in 1996 where he is now a senior scientist. At BBN he works on the design and development of large vocabulary speech recognition system in a variety of languages. His areas of interest are speech feature extraction and acoustic modeling for speech recognition.

BIBLIOGRAPHY

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.
2. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
3. X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh Univ. Press, Edinburgh, Scotland, 1990.
4. F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1998.
5. J. Makhoul et al., Speech and language technologies for audio indexing and retrieval, *Proc. IEEE* **88**(8): 1338–1354 (Aug. 2000).
6. S. Matsoukas et al., The 1998 BBN Byblos primary system applied to English and Spanish broadcast news transcription, *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Herndon, UK, March 1999.
7. J. Billa et al., Recent experiments in large vocabulary conversational speech recognition, in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Phoenix, AZ, IEEE, April 1999.
8. A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, 1993.

SPREAD SPECTRUM SIGNALS FOR DIGITAL COMMUNICATIONS

JOHN G. PROAKIS
Northeastern University
Boston, Massachusetts

1. INTRODUCTION

Spread spectrum signals is a class of signals that were designed primarily for use in digital communication systems to overcome either intentional or unintentional interference. Such signals were originally used in military communication systems either to provide resistance to jamming or to hide the signal by transmitting it at low power and, thus, making it difficult for an unintended listener to detect its presence in noise (low probability of intercept). However, today, spread spectrum signals are used to provide reliable communications in a variety of commercial applications. For example, spread spectrum signals are used in the so-called unlicensed frequency bands, such as the Industry, Scientific, and Medical (ISM) band at 2.4 GHz. Typical applications in the ISM band are cordless telephones, wireless LANs, and cable replacement systems such as Bluetooth. Since the band is unlicensed, there is no central control over the radio resources, and the systems have to function even in the presence of severe interference from other communication systems and other electrical and electronic equipment (e.g., microwave ovens, radars, etc.). Here the interference is not intentional, but the interference may nevertheless be enough to disrupt the communication for non–spread spectrum systems.

Code-division multiple access systems (CDMA systems) use spread spectrum techniques to provide communication to several concurrent users. CDMA is used in one second generation (IS-95) and several third generation wireless cellular systems (e.g., cdma2000 and WCDMA). One advantage of using interference-resistant signals in these applications is that the radio resource management (primarily the channel allocation to the active users) is significantly reduced.

The name spread spectrum stems from the fact that the transmitted signals occupy a much wider frequency band than is required to transmit the information. There are many different ways to spread the bandwidth of the information-bearing signal. The most common ones are called direct-sequence (DS) and frequency-hopping (FH) spread spectrum (SS). In DS-SS, the information-bearing signal is spread over the entire channel bandwidth in a manner that appears random. In a FH-SS, the transmitter changes the carrier frequency of the relatively narrowband information-bearing signal in a fashion that appears random. At any given time, only a small fraction of the available channel bandwidth is used and exactly which fraction used is known only to the intended receiver. A hybrid of DS and FH spread spectrum is also possible. Several other methods are available for spreading the bandwidth of the information-bearing signals; however, the clear majority of the implemented systems are either DS or FH or a hybrid of DS/FH.

The literature in the field of spread spectrum communications is quite voluminous and ranges from text books to specialized conference and journal papers. For an interesting review of the history of the development of spread spectrum, we recommend the papers [1–3]. Among the available books, we would like to especially mention the text by Simon, Omura, Scholtz, and Levitt [4] which covers quite a lot of the classical spread spectrum techniques. The books by Ziemer, Peterson, and Borth [5] and by Dixon [6], which cover more of the current commercial applications, are also recommended. The reference lists of the above mentioned books contain several thousand entries. Among the tutorial-style papers that are available in the literature, we would like to especially mention the 1982 paper by Pickholtz, Schilling, and Milstein [7].

2. MODEL OF A SPREAD SPECTRUM DIGITAL COMMUNICATION SYSTEM

The basic elements of a spread spectrum digital communication system are illustrated in Fig. 1. We observe that the channel encoder and decoder and the modulator and demodulator are the basic elements of a conventional digital communication system. In addition to these elements, a spread-spectrum system employs two identical pseudorandom sequence generators, one which interfaces with the modulator at the transmitting end and the second which interfaces with the demodulator at the receiving end. These two generators produce a pseudorandom or pseudonoise (PN) binary-valued sequence, which is used to spread the transmitted signal at the modulator and to despread the received signal at the demodulator.

Time synchronization of the PN sequence generated at the receiver with the PN sequence contained in the received signal is required to properly despread the received spread-spectrum signal. In a practical system, synchronization is established prior to the transmission

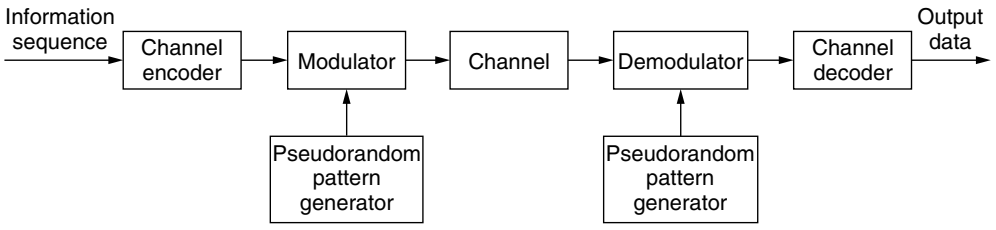


Figure 1. Model of spread-spectrum digital communications system.

of information by transmitting a fixed PN bit pattern which is designed so that the receiver will detect it with high probability in the presence of interference. After time synchronization of the PN sequence generators is established, the transmission of information commences. In the data mode, the communication system usually tracks the timing of the incoming received signal and keeps the PN sequence generator in synchronism. The synchronization of spread spectrum signals is treated in Refs. 4 and 7 and in the article by Luise et al. [8] in this encyclopedia.

Interference is introduced in the transmission of the spread-spectrum signal through the channel. The characteristics of the interference depend to a large extend on its origin. The interference may be generally categorized as being either broadband or narrowband (partial band) relative to the bandwidth of the information-bearing signal, and either continuous in time or pulsed (discontinuous) in time. For example, an interfering signal may consist of a high-power sinusoid in the bandwidth occupied by the information-bearing signal. Such a signal is narrowband. As a second example, the interference generated by other users in a multiple-access channel depends on the type of spread-spectrum signals that are employed by the various users to transmit their information. If all users employ broadband signals, the interference may be characterized as an equivalent broadband noise. If the users employ frequency hopping to generate spread-spectrum signals, the interference from other users may be characterized as narrowband or partial band.

Our discussion will focus on the performance of spread-spectrum signals for digital communication in the presence of narrowband and broadband interference. Two types of digital modulation are considered, namely, PSK and FSK. PSK modulation is appropriate for applications where phase coherence between the transmitted signal and the received signal can be maintained over a time interval that spans several symbol (or bit) intervals. This is usually the case in DS-SS, where a single carrier frequency is modulated by the spread spectrum signal that covers the entire channel bandwidth. On the other hand, FSK modulation with noncoherent detection is appropriate in applications where phase coherence of the carrier cannot be accurately estimated. This is usually the case in FH-SS, where the carrier frequency is hopped rapidly, typically at the transmitted symbol (or bit) rate. In such a case, the relatively short time interval spanned by a single symbol (or bit) is not sufficient to obtain an accurate estimate of the carrier phase.

The PN sequence generated at the modulator is used in conjunction with the PSK modulation to shift the phase of the PSK signal pseudorandomly, as described below at a rate that is an integer multiple of the bit rate. The resulting modulated signal is called a *direct-sequence* (*DS*) *spread-spectrum signal*. When used in conjunction with binary or M-ary ($M > 2$) FSK, the PN sequence is used to select the carrier frequency of the transmitted signal pseudorandomly. The resulting signal is called a *frequency-hopped* (*FH*) *spread-spectrum signal*.

3. DIRECT-SEQUENCE SPREAD-SPECTRUM SYSTEMS

Let us consider the transmission of a binary information sequence by means of binary PSK. The information rate is R bits/sec and the bit interval is $T_b = 1/R$ sec. The available channel bandwidth is W Hz, where $W \gg R$. At the modulator, the bandwidth of the information signal is expanded to W Hz by shifting the phase of the carrier pseudorandomly at a rate of W times/sec according to the pattern of the PN generator. The basic method for accomplishing the spreading is shown in Fig. 2.

The information-bearing baseband signal is denoted as $v(t)$ and is expressed as

$$v(t) = \sum_{n=-\infty}^{\infty} a_n g_T(t - nT_b) \tag{1}$$

where $\{a_n = \pm 1, -\infty < n < \infty\}$ and $g_T(t)$ is a rectangular pulse of duration T_b. This signal is multiplied by the signal from the PN sequence generator, which may be expressed as

$$c(t) = \sum_{n=-\infty}^{\infty} c_n p(t - nT_c) \tag{2}$$

where $\{c_n\}$ represents the binary PN code sequence of ± 1's and $p(t)$ is a rectangular pulse of duration T_c, as illustrated in Fig. 2. This multiplication operation serves to spread the bandwidth of the information-bearing signal (whose bandwidth is R hz, approximately) into the wider bandwidth occupied by PN generator signal $c(t)$

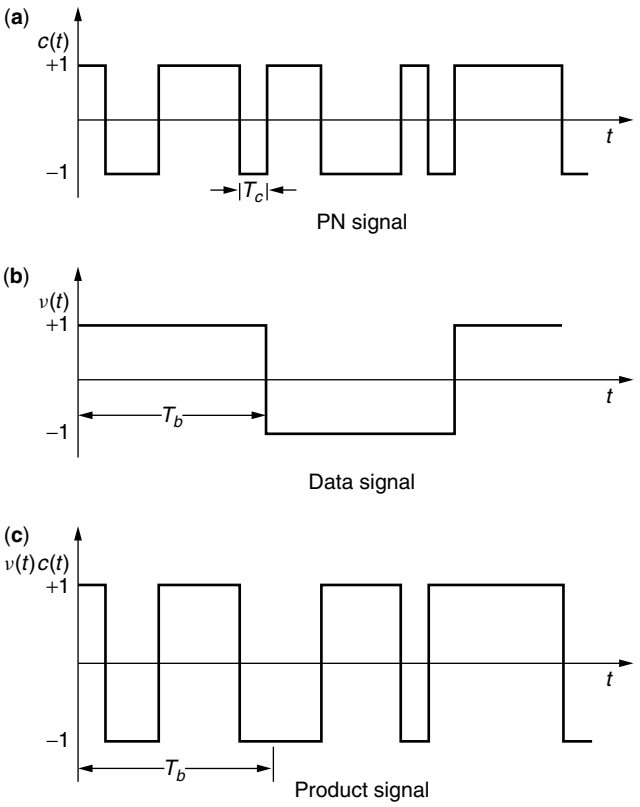


Figure 2. Generation of a DS spread-spectrum signal.

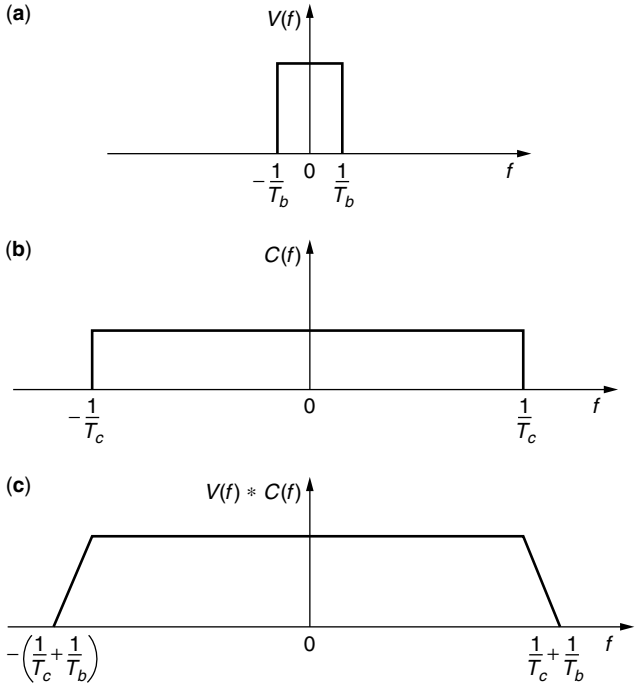


Figure 3. Convolution of spectra of the (**a**) data signal with the (**b**) PN code signal.

(whose bandwidth is $1/T_c$, approximately). The spectrum spreading is illustrated in Fig. 3, which shows, in simple terms, using rectangular spectra, the convolution of the two spectra, the narrow spectrum corresponding to the information-bearing signal and the wide spectrum corresponding to the signal from the PN generator.

The product signal $\upsilon(t)c(t)$, also illustrated in Fig. 2, is used to amplitude modulate the carrier $A_c \cos 2\pi f_c t$ and, thus, to generate the double-sideband, suppressed carrier (DSB-SC) signal

$$u(t) = A_c \upsilon(t) c(t) \cos 2\pi f_c t \tag{3}$$

Since $\upsilon(t)c(t) = \pm 1$ for any t, it follows that the carrier-modulated transmitted signal may also be expressed as

$$u(t) = A_c \cos[2\pi f_c t + \theta(t)] \tag{4}$$

where $\theta(t) = 0$ when $\upsilon(t)c(t) = 1$ and $\theta(t) = \pi$ when $\upsilon(t)c(t) = -1$. Therefore, the transmitted signal is a binary PSK signal.

The rectangular pulse $p(t)$ is usually called a *chip* and its time duration T_c is called the *chip interval*. The reciprocal $1/T_c$ is called the *chip rate* and corresponds (approximately) to the bandwidth W of the transmitted signal. The ratio of the bit interval T_b to the chip interval T_c usually is selected to be an integer in practical spread spectrum systems. We denote this ratio as

$$L_c = \frac{T_b}{T_c} \tag{5}$$

Hence, L_c is the number of chips of the PN code sequence/information bit. Another interpretation is that L_c represents the number of possible 180° phase transitions in the transmitted signal during the bit interval T_b.

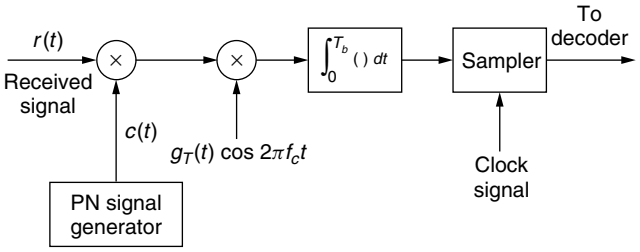


Figure 4. Demodulation of DS spread spectrum signal.

The demodulation of the signal is performed as illustrated in Fig. 4. The received signal is first multiplied by a replica of the waveform $c(t)$ generated by the PN code sequence generator at the receiver, which is synchronized to the PN code in the received signal. This operation is called (spectrum) despreading, since the effect of multiplication by $c(t)$ at the receiver is to undo the spreading operation at the transmitter. Thus, we have

$$A_c \upsilon(t) c^2(t) \cos 2\pi f_c t = A_c \upsilon(t) \cos 2\pi f_c t \tag{6}$$

since $c^2(t) = 1$ for all t. The resulting signal $A_c \upsilon(t) \cos 2\pi f_c t$ occupies a bandwidth (approximately) of R hz, which is the bandwidth of the information-bearing signal. Therefore, the demodulator for the despread signal is simply the conventional cross correlator or matched filter. Since the demodulator has a bandwidth that is identical to the bandwidth of the despread signal, the only additive noise that corrupts the signal at the demodulator is the noise that falls within the information-bandwidth of the received signal.

3.1. Effect of Despreading on a Narrowband Interference

It is interesting to investigate the effect of an interfering signal on the demodulation of the desired information-bearing signal. Suppose that the received signal is

$$r(t) = A_c \upsilon(t) c(t) \cos 2\pi f_c t + i(t) \tag{7}$$

where $i(t)$ denotes the interference. The despreading operation at the receiver yields

$$r(t)c(t) = A_c \upsilon(t) \cos 2\pi f_c t + i(t)c(t) \tag{8}$$

The effect of multiplying the interference $i(t)$ with $c(t)$, is to spread the bandwidth of $i(t)$ to W HZ.

As an example, let us consider a sinusoidal interfering signal of the form

$$i(t) = A_I \cos 2\pi f_I t \tag{9}$$

where f_I is a frequency within the bandwidth of the transmitted signal. Its multiplication with $c(t)$ results in a wideband interference with power-spectral density $I_0 = P_I/W$, where $P_I = A_I^2/2$ is the average power of the interference. Since the desired signal is demodulated by a matched filter (or correlator) that has a bandwidth R,

the total power in the interference at the output of the demodulator is

$$I_0 R_b = P_I R_b / W = \frac{P_I}{W/R_b} = \frac{P_I}{T_b/T_c} = \frac{P_I}{L_c} \tag{10}$$

Therefore, the power in the interfering signal is reduced by an amount equal to the bandwidth expansion factor W/R. The factor $W/R = T_b/T_c = L_c$ is called the *processing gain* of the spread-spectrum system. The reduction in interference power is the basic reason for using spread-spectrum signals to transmit digital information over channels with interference.

In summary, the PN code sequence is used at the transmitter to spread the information-bearing signal into a wide bandwidth for transmission over the channel. By multiplying the received signal with a synchronized replica of the PN code signal, the desired signal is despread back to a narrow bandwidth while any interference signals are spread over a wide bandwidth. The net effect is a reduction in the interference power by the factor W/R, which is the processing gain of the spread-spectrum system.

The PN code sequence $\{c_n\}$ is assumed to be known only to the intended receiver. Any other receiver that does not have knowledge of the PN code sequence cannot demodulate the signal. Consequently, the use of a PN code sequence provides a degree of privacy (or security) that is not possible to achieve with conventional modulation. The primary cost for this security and performance gain against interference is an increase in channel bandwidth utilization and in the complexity of the communication system.

3.2. Probability of Error

In this section we derive the probability of error for a DS spread spectrum system assuming that the information is transmitted via binary PSK. Within the bit interval $0 \le t \le T_b$, the transmitted signal is

$$s(t) = a_0 g_T(t) c(t) \cos 2\pi f_c t, \quad 0 \le t \le T_b \tag{11}$$

where $a_0 = \pm 1$ is the information symbol, the pulse $g_T(t)$ is defined as

$$g_T(t) = \begin{cases} \sqrt{\dfrac{2\mathcal{E}_b}{T_b}}, & 0 \le 1 \le T_b \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

and $c(t)$ is the output of the PN code generator which, over a bit interval, is expressed as

$$c(t) = \sum_{n=0}^{L_c - 1} c_n p(t - nT_c) \tag{13}$$

where L_c is the number of chips per bit, T_c is the chip interval, and $\{c_n\}$ denotes the PN code sequence. The PN code chip sequence $\{c_n\}$ is designed to be uncorrelated (white); that is,

$$E(c_n c_m) = E(c_n) E(c_m) \text{ for } n \ne m \tag{14}$$

and each chip is $+1$ or -1 with equal probability. These conditions imply that $E(c_n) = 0$ and $E\left(c_n^2\right) = 1$.

The received signal is assumed to be corrupted by an additive interfering signal $i(t)$. Hence,

$$r(t) = a_0 g_T(t) c(t) \cos(2\pi f_c t + \phi) + i(t) \tag{15}$$

where ϕ represents the carrier phase shift. Since the received signal $r(t)$ is typically the output of an ideal bandpass filter in the front end of the receiver, the interference $i(t)$ is also a bandpass signal, and may be represented as

$$i(t) = i_c(t) \cos 2\pi f_c t - i_s(t) \sin 2\pi f_c t \tag{16}$$

where $i_c(t)$ and $i_s(t)$ are the two quadrature components of $i(t)$.

We assume that the receiver is perfectly synchronized to the received signal and the carrier phase is perfectly estimated by a PLL. Then, the signal $r(t)$ is demodulated by first despreading through multiplication by $c(t)$ and then crosscorrelation with $g_T(t) \cos(2\pi f_c t + \phi)$, as shown in Fig. 5. At the sampling instant $t = T_b$, the output of the correlator is

$$y(T_b) = \mathcal{E}_b + y_i(T_b) \tag{17}$$

where $y_i(T_b)$ represents the interference component, which has the form

$$\begin{aligned} y_i(T_b) &= \int_0^{Tb} c(t) i(t) g_T(t) \cos(2\pi f_c t + \phi)\, dt \\ &= \sum_{n=0}^{L_c - 1} c_n \int_0^{Tb} p(t - nT_c) i(t) g_T(t) \cos(2\pi f_c t + \phi)\, dt \\ &= \sqrt{\frac{2\mathcal{E}_b}{T_b}} \sum_{n=0}^{L_c - 1} c_n v_n \end{aligned} \tag{18}$$

where, by definition,

$$v_n = \int_{nT_c}^{(n+1)T_c} i(t) \cos(2\pi f_c t + \phi)\, dt \tag{19}$$

The probability of error depends on the statistical characteristics of the interference component. Its mean value is

$$E[y_i(T_b)] = 0 \tag{20}$$

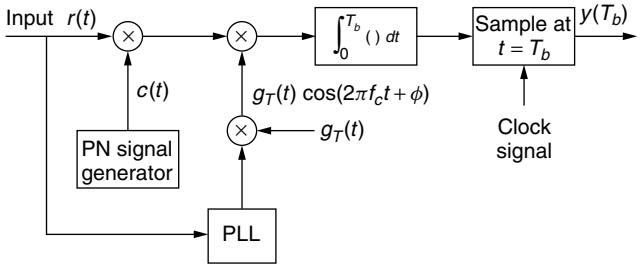


Figure 5. DS spread-spectrum signal demodulator.

Its variance is

$$E[y_i^2(T_b)] = \frac{2\mathcal{E}_b}{T_b} \sum_{n=0}^{Lc-1} \sum_{m=0}^{Lc-1} E(c_n c_m) E(v_n v_m)$$

But $E(c_n c_m) = \delta_{mn}$. Therefore,

$$E[y_i^2(Tb)] = \frac{2\mathcal{E}_b}{T_b} \sum_{n=0}^{L_c-1} E\left(v_n^2\right) = \frac{2\mathcal{E}_b}{T_b} L_c E\left(v^2\right) \tag{21}$$

where $v = v_n$, as given by Eq. (9).

Consider for sinusoidal interfering signal at the carrier frequency, that is,

$$i(t) = \sqrt{2P_I}\cos(2\pi f_c t + \Theta_I) \tag{22}$$

where P_I is the average power and Θ_I is the phase of the interference, which is random and uniformly distributed over the interval $(0, 2\pi)$. If we substitute for $i(t)$ in Eq. (19), it is easy to show that

$$E(v^2) = \frac{T_c^2 P_I}{4} \tag{23}$$

and, therefore,

$$E[y_i^2(T_b)] = \frac{\mathcal{E}_b P_I T_c}{2} \tag{24}$$

The ratio of $\{E[y(T_b)]\}^2$ to $E[y_i^2(T_b)]$ is the SNR at the detector. In this case we have

$$(\text{SNR})_D = \frac{\mathcal{E}_b^2}{\mathcal{E}_b P_I T_c/2} = \frac{2\mathcal{E}_b}{P_I T_c} \tag{25}$$

To see the effect of the spread-spectrum signal, we express the transmitted energy $\mathcal{E}_b$ as

$$\mathcal{E}_b = P_S T_b \tag{26}$$

where P_S is the average signal power. Then, if we substitute for $\mathcal{E}_b$ in Eq. (25) we obtain

$$(\text{SNR})_D = \frac{2P_S T_b}{P_I T_c} = \frac{2P_S}{P_I/L_c} \tag{27}$$

where $L_c = T_b/T_c$ is the processing gain. Therefore, the spread-spectrum signal has reduced the power of the interference by the factor L_c.

Another interpretation of the effect of the spread-spectrum signal on the sinusoidal interference is obtained if we express $P_I T_c$ in Eq. (27) as follows. Since $T_c \simeq 1/W$, we have

$$P_I T_c = P_I/W = I_0 \tag{28}$$

where I_0 is the power-spectral density of an equivalent interference in a bandwidth W. Therefore, in effect, the spread-spectrum signal has spread the sinusoidal interference over the wide bandwidth W, creating an equivalent spectrally flat noise with power-spectral density I_0. Hence,

$$(\text{SNR})_D = \frac{2\mathcal{E}_b}{I_0} \tag{29}$$

The probability of error for a DS spread-spectrum system with binary PSK modulation is easily obtained from the SNR at the detector, if we make an assumption on the probability distribution of the sample $y_i(T_b)$. From Eq. (18) we note that $y_i(T_b)$ consists of a sum of L_c uncorrelated random variables $\{c_n v_n, 0 \le n \le L_c - 1\}$, all of which are identically distributed. Since the processing gain L_c is usually large in any practical system, we may use the Central Limit Theorem to justify a Gaussian probability distribution for $y_i(T)$. Under this assumption, the probability of error is

$$P_b = Q\left(\sqrt{\frac{2\mathcal{E}_b}{I_0}}\right) \tag{30}$$

where I_0 is the power-spectral density of an equivalent broadband interference and $Q(x)$ is defined as

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt \tag{31}$$

Hence, the effect of the interference on the error probability is equivalent to that of broadband white Gaussian noise with spectral density I_0.

3.3. The Interference Margin

We may express $\dfrac{\mathcal{E}_b}{I_0}$ in the Q-function in Eq. (30) as

$$\frac{\mathcal{E}_b}{I_0} = \frac{P_S T_b}{P_I/W} = \frac{P_S/R}{P_I/W} = \frac{W/R}{P_I/P_S} \tag{32}$$

Also, suppose we specify a required E_b/I_0 to achieve a desired level of performance. Then, using a logarithmic scale, we may express Eq. (32) as

$$10\log\frac{P_I}{P_S} = 10\log\frac{W}{R} - 10\log\frac{\mathcal{E}_b}{I_0}$$
$$\left(\frac{P_I}{P_S}\right)_{\text{dB}} = \left(\frac{W}{R}\right)_{\text{dB}} - \left(\frac{\mathcal{E}_b}{I_0}\right)_{\text{dB}} \tag{33}$$

The ratio $(P_I/P_S)_{\text{dB}}$ is called the *interference margin*. This is the relative power advantage than an interference may have without disrupting the communication system.

For example, suppose we require an $(\mathcal{E}_b/I_0))_{\text{dB}} = 10$ dB to achieve reliable communication. What is the processing gain that is necessary to provide an interference margin of 20 dB? Clearly, if $W/R = 1000$, then $(W/R)_{\text{dB}} = 30$ dB and the interference margin is $(P_I/P_S)_{\text{dB}} = 20$ dB. This means that the average interference power at the receiver may be 100 times the power P_S of the desired signal and we can still maintain reliable communication.

3.4. Performance of Coded Spread-Spectrum Signals

It is shown in many textbooks on digital communications (for reference, see [9], Chapter 8), that when the transmitted information is coded by a binary linear (block or convolutional) code, the SNR at the output of a soft-decision decoder in the presence of spectrally flat Gaussian interference is increased by the coding gain, defined as

$$\text{coding gain} = R_c d_{\min}^H \tag{34}$$

where R_c is the code rate and $d_{\min}^H$ is the minimum Hamming distance of the code. Therefore, the effect of coding is to increase the interference margin by the coding gain. Thus, Eq. (33) may be modified as

$$\left(\frac{P_I}{P_S}\right)_{\text{dB}} = \left(\frac{W}{R}\right)_{\text{dB}} + (CG)_{\text{dB}} - \left(\frac{\mathcal{E}_b}{I_0}\right)_{\text{dB}} \tag{35}$$

where $(CG)_{\text{dB}}$ denotes the coding gain. Typical coding gains obtained by use of binary block or convolutional codes are in the range of 4 to 7 dB.

3.5. Pulsed Interference

A very damaging type of interference for DS-SS is broadband pulsed noise, whose power is spread over the entire system bandwidth W. The pulsed interference is transmitted for a fraction ρ of the time, that is, ρ is the duty cycle of the transmitted interference, where $0 < \rho \leq 1$. If this signal is being transmitted by a jammer, this allows the jammer to transmit pulses with a power level P_I/ρ for ρ percent of the time, with an equivalent spectral density of $I_0 = P_I/W$, where P_I is the average transmitted power. For simplicity, let us assume that the interference pulse spans an integer number of symbols (or bits) and that the pulsed noise is Gaussian distributed. When the interferer is not transmitting, the received information bits are assumed to be error-free, and when the interferer is transmitting, the probability of error for an uncoded DS-SS system is

$$P(\rho) = \rho Q\left(\sqrt{\frac{2\mathcal{E}_b}{I_0}\rho}\right) \tag{36}$$

The worst case duty cycle that maximizes the probability of error for the communication system can be found by differentiating $P(\rho)$ with respect to ρ. Thus, we find that the worst-case pulsed noise occurs when

$$\rho^* = \begin{cases} \dfrac{0.71}{\mathcal{E}_b/I_0}, & \dfrac{\mathcal{E}_b}{I_0} \geq 0.71 \\ 1, & \dfrac{\mathcal{E}_b}{I_0} < 0.71 \end{cases} \tag{37}$$

and the corresponding probability of error is

$$P(\rho^*) = \begin{cases} \dfrac{0.082}{\mathcal{E}_b/I_0} = \dfrac{0.082 P_I/P_S}{W/R}, & \dfrac{\mathcal{E}_b}{I_0} \geq 0.71 \\ Q\left(\sqrt{\dfrac{2\mathcal{E}_b}{I_0}}\right) = Q\left(\sqrt{\dfrac{2W/R}{P_I/P_S}}\right), & \dfrac{\mathcal{E}_b}{I_0} < 0.71 \end{cases} \tag{38}$$

The error rate performance given by Eq. (38) for $\rho = 1.0, 0.1$, 0.01, and 0.001 along with the worst-case performance based on ρ^* is plotted in Fig. 6. When we compare the error rate for continuous wideband Gaussian noise interference ($\rho = 1$) with worst-case pulse interference, we find a large difference in performance; for example, approximately 40 dB at an error rate of 10^{-6}. This is, indeed, a large penalty.

If we simply add coding to the DS spread-spectrum system, the performance in SNR is improved by an amount equal to the coding gain, which in most cases is limited to less than 10 dB. The reason that the addition of coding does not improve the performance significantly is that the interfering signal pulse duration (duty cycle) may be selected to affect many consecutive coded bits. Consequently, the code word error probability is high due to the burst characteristics of the interference.

In order to improve the performance of the coded DS spread-spectrum system, we should interleave the coded bits prior to transmission over the channel. The effect of interleaving is to make the coded bits that are affected by the interferer statistically independent. Figure 7 illustrates a block diagram of a DS spread-spectrum system that employs coding and interleaving. By selecting a sufficiently long interleaver so that the burst characteristics of the interference are eliminated, the penalty in performance due to pulse interference is significantly reduced; for example, to the range of 3–5 dB for conventional binary block or convolutional codes (for reference, see [9], Chapter 13).

4. FREQUENCY-HOPPED SPREAD SPECTRUM

In frequency-hopped (FH) spread spectrum, the available channel bandwidth W is subdivided into a large number of nonoverlapping frequency slots. In any signaling interval the transmitted signal occupies one or more of the available frequency slots. The selection of the frequency slot (s) in each signal interval is made pseudorandomly according to the output from a PN generator.

A block diagram of the transmitter and receiver for an FH spread-spectrum system is shown in Fig. 8. The modulation is either binary or M-ary FSK (MFSK). For example, if binary FSK is employed, the modulator selects one of two frequencies say f_0 or f_1, corresponding to the transmission of a 0 for a 1. The resulting binary FSK signal is translated in frequency by an amount that is determined by the output sequence from a PN generator, which is used to select a frequency f_c that is synthesized by the frequency synthesizer. This frequency is mixed with the output of the FSK modulator and the resultant frequency-translated signal is transmitted over the channel. For example, by taking m bits from the PN generator, we may specify $2^m - 1$ possible carrier frequencies. Figure 9 illustrates an FH signal pattern.

At the receiver, there is an identical PN sequences generator synchronized with the received signal, which is used to control the output of the frequency synthesizer. Thus, the pseudorandom frequency translation introduced at the transmitter is removed at the demodulator by mixing the synthesizer output with the received signal.

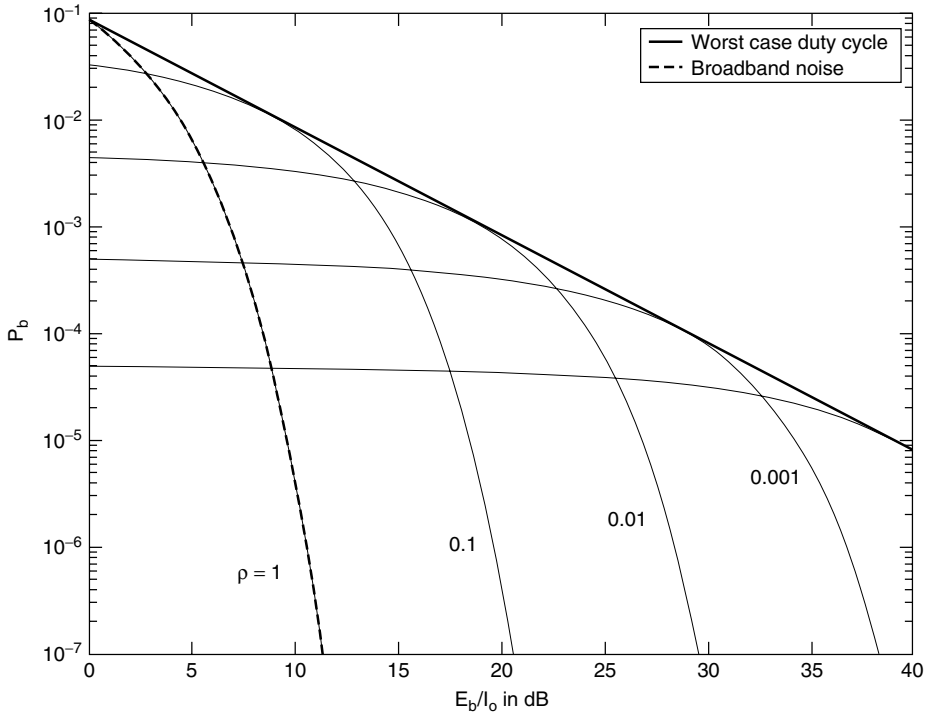


Figure 6. Bit error probability for BPSK modulated DS spread spectrum with pulsed interference having a duty cycle P.

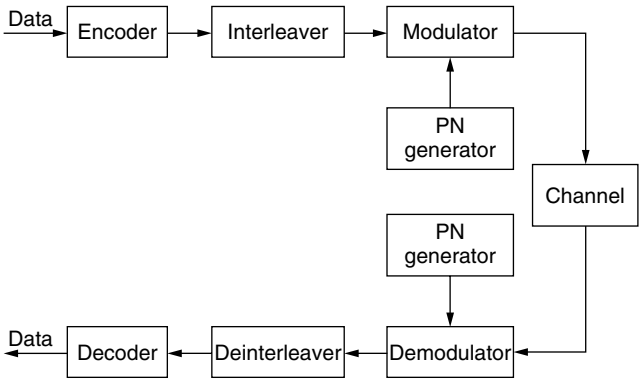


Figure 7. Black diagram of communication system with coding and interleaving.

The resultant signal is then demodulated by means of an FSK demodulator. A signal for maintaining synchronism of the PN sequence generator with the FH received signal is usually extracted from the received signal.

Although binary PSK modulation generally yields better performance than binary FSK, it is difficult to maintain phase coherence in the synthesis of the frequencies used in the hopping pattern and, also, in the propagation of the signal over the channel as the signal is hopped from one frequency to the another over a wide bandwidth. Consequently, FSK modulation with noncoherent demodulation is usually employed in FH spread-spectrum systems.

The frequency-hopping rate, denoted as R_h, may be selected to be either equal to the symbol rate, or lower than the symbol rate, or higher than the symbol rate. If R_h is equal to or lower than the symbol rate, the FH system is called a slow-hopping system. If R_h is higher than the symbol rate; that is, there are multiple hops/symbols, the FH system is called a fast-hopping system. However, there is a penalty incurred in subdividing an information symbol into several frequency-hopped elements, because the energy from these separate elements is combined noncoherently (for reference, see [9], Chapter 12).

FH spread-spectrum signals may be used in CDMA where many users share a common bandwidth. In some cases, an FH signal is preferred over a DS spread-spectrum signal because of the stringent synchronization requirements inherent in DS spread-spectrum signals. Specifically, in a DS system, timing and synchronization must be established to within a fraction of a chip interval $T_c = 1/W$. Conversely, in an FH system, the chip interval T_c is the time spend in transmitting a signal in a particular frequency slot of bandwidth $B \ll W$. But this interval is approximately $1/B$, which is much larger than $1/W$. Hence, the timing requirements in an FH system are not as stringent as in a DS system.

4.1. Slow Frequency-Hopping Systems

Let us consider a slow frequency-hopping system in which the hop rate $R_h = 1$ hop/bit. If the interference on the channel is broadband and is characterized as AWGN with power-spectral density I_0, the probability of error for the detection of noncoherently demodulated binary FSK is

$$P_b = \tfrac{1}{2} e^{-\mathcal{E}_b/2I_0} \tag{39}$$

where $\mathcal{E}_b/I_0$ is the SNR/bit.

As in the case of a DS spread-spectrum system, we observe that $\mathcal{E}_b$, the energy/bit, can be expressed as $\mathcal{E}_b = P_S T_b = P_S/R$, where P_S is the average transmitted power and R is the bit rate. Similarly, $I_0 = P_I/W$, where

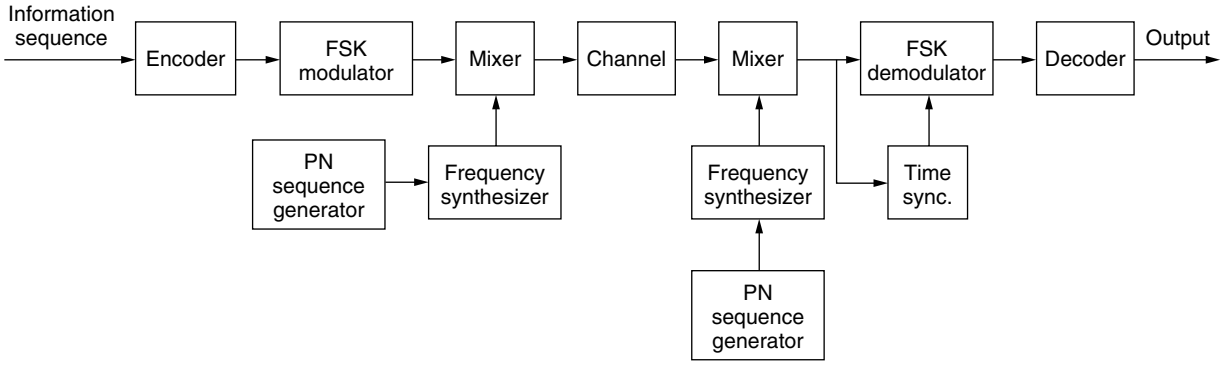


Figure 8. Black diagram of an FH spread-spectrum system.

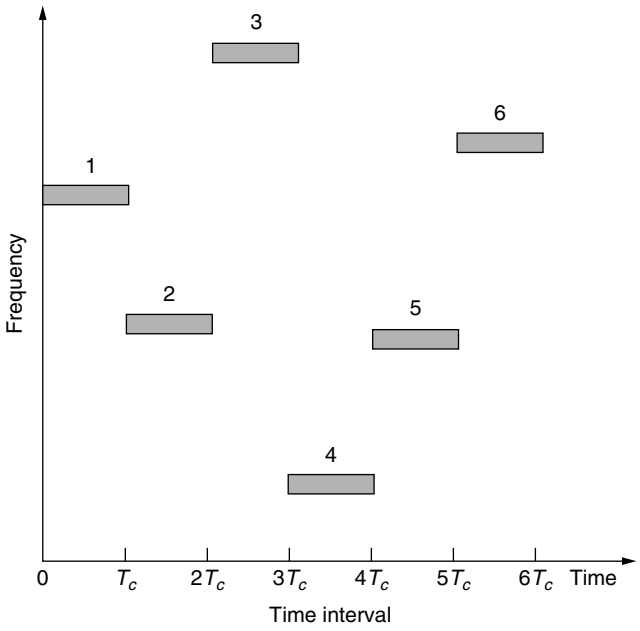


Figure 9. An example of an FH pattern.

P_I is the average power of the broadband interference and W is the available channel bandwidth. Therefore, the SNR/bit can be expressed as

$$\frac{\mathcal{E}_b}{I_0} = \frac{W/R}{P_I/P_S} \tag{40}$$

where W/R is the processing gain and P_I/P_S is the interference margin for the FH spread-spectrum signal.

Slow FH spread-spectrum systems are particularly vulnerable to partial-band interference that may result in FH CDMA systems. To be specific, suppose that the partial-band interference is modeled as a zero-mean Gaussian random process with a flat power-spectral density over a fraction of the total bandwidth W and zero in the remainder of the frequency band. In the region or regions where the power-spectral density is nonzero, its value is I_0/ρ, where $0 < \rho < 1$. In other words, the interference average power P_I is assumed to be constant.

Let us consider the worst-case partial-band interference by selecting the value of ρ that maximizes the error probability. In an uncoded slow-hopping system with binary FSK modulation and noncoherent detection, the transmitted frequencies are selected with uniform probability in the frequency band W. Consequently, the received signal will be corrupted by interference with probability ρ. When the interference is present, the probability of error is $\frac{1}{2}\exp(-\rho\mathcal{E}_b/2I_0)$ and when it is not, the detection of the signal is assumed to be error free. Therefore, the average probability of error is

$$\begin{aligned} P_b(\rho) &= \frac{\rho}{2} e^{-\rho\mathcal{E}_b/2I_0} \\ &= \frac{\rho}{2}\exp\left(\frac{\rho W/R}{2P_I/P_S}\right) \end{aligned} \tag{41}$$

Figure 10 illustrates the error rate as a function of $\mathcal{E}_b/I_0$ for several values of ρ. By differentiating $P_b(\rho)$, and solving for the value of ρ that maximizes $P_b(\rho)$, we find

$$\rho^* = \begin{cases} 2I_0/\mathcal{E}_b, & \rho \geq 2 \\ 1, & \mathcal{E}_b/I_0 < 2 \end{cases} \tag{42}$$

The corresponding error probability for the worst case partial-band interference is

$$P_b = \begin{cases} e^{-1}/\mathcal{E}_b/I_0, & \mathcal{E}_b/I_0 \geq 2 \\ \frac{1}{2}e^{-\mathcal{E}_b/2I_0}, & \mathcal{E}_b/I_0 < 2 \end{cases} \tag{43}$$

which is also shown in Fig. 10. Whereas the error probability decreases exponentially for full-band interference as given by Eq. (39), the error probability for worst-case partial band interference decreases only inversely with $\mathcal{E}_b/I_0$. This result is similar to the error probability for DS spread-spectrum signals in the presence of pulse interference. It is also similar to the error probability for binary PSK in a Rayleigh fading channel.

An effective method for combatting partial band interference in a FH system is signal diversity, which can be obtained by simple repetition of the transmitted information bit on different frequencies (or by means of block or convolutional coding). Signal diversity obtained through coding provides a significant improvement in performance relative to uncoded signal transmission. In fact, it has been shown by Viterbi and Jacobs [10] that by optimizing the code design for the partial-band

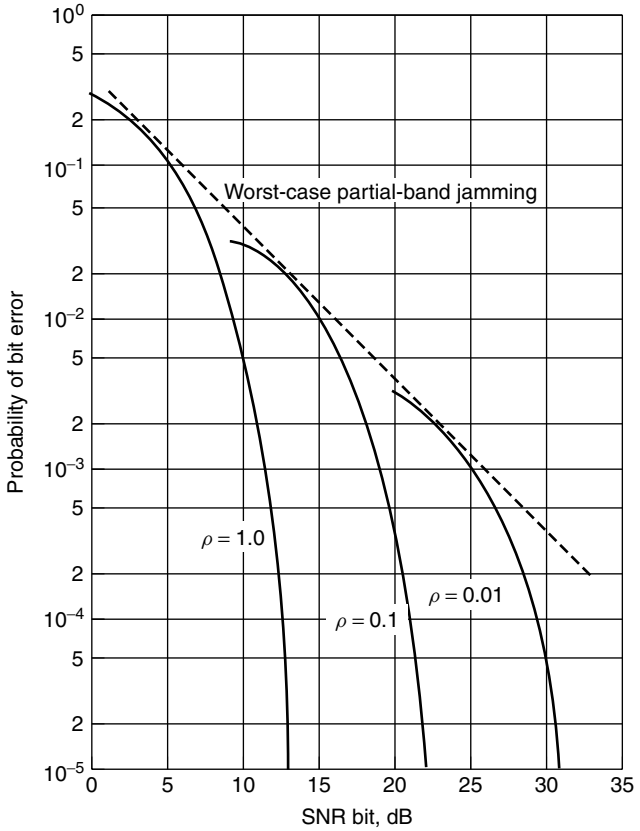


Figure 10. Performance of binary FSK with Partial-band interference.

interference, the communication system can achieve an average bit-error probability of

$$P_b = e^{-\mathcal{E}_b/4I_0} \tag{44}$$

Therefore, the probability of error achieved with the optimum code design decreases exponentially with an increase in SNR and is within 3 dB of the performance obtained in an AWGN channel. Thus, the penalty due to partial-band interference is reduced significantly.

4.2. Fast Frequency-Hopping Systems

In fast FH systems, the frequency-hop rate R_h is some multiple of the symbol rate. Basically, each (M-ary) symbol interval is subdivided into N subintervals, which are called *chips* and one of M frequencies is transmitted in each subinterval. Fast frequency-hopping systems are particularly attractive for military communications. In such systems, the hop rate R_h may be selected sufficiently high so that a potential intentional interferer does not have sufficient time to detect the presence of the transmitted frequency and to synthesize a jamming signal that occupies the same bandwidth.

To recover the information at the receiver, the received signal is first dehopped by mixing it with the hopped carrier frequency. This operation removes the hopping pattern and brings the received signal in all subintervals (chips) to a common frequency band that encompasses the M possible transmitted frequencies. The signal in each subinterval is then passed through the M matched filters (or correlators) tuned to the M possible transmitted frequencies which are sampled at the end of each subinterval and passed to the detector. The detection of the FSK signals is noncoherent. Hence, decisions are based on the magnitude of the matched filter (or correlator) outputs.

Since each symbol is transmitted over N chips, the decoding may be performed simply on the basis of hard decisions for the chips on each hop.

To determine the probability of error for the detector, we recall that the probability of error for noncoherent detection of binary FSK for each hop is

$$p = \tfrac{1}{2}e^{-\mathcal{E}_b/2NI_0} \tag{45}$$

where N is the number of frequency hops (chips) per bit, and $\mathcal{E}_b/N$ is the energy of the signal per hop. Assuming that N is odd, the decoder decides in favor of the transmitted binary FSK bit that is larger in at least $(N+1)/2$ chips. Thus, the decision is made on the basis of a majority vote given the decisions on the N chips. Consequently, the probability of a bit error is

$$P_b = \sum_{m=(N+1)/2}^{N} \binom{N}{m} p^m (1-p)^{N-m} \tag{46}$$

where p is given by Eq. (45). We should note that the error probability P_b for hard-decision decoding of the N chips will be higher than the error probability for a single hop/bit PSK system, which is given by Eq. (39), when the SNR/bit $\mathcal{E}_b/I_0$ is the same in the two systems.

5. COMMERCIAL SPREAD SPECTRUM SYSTEMS

As mentioned earlier, the number of nonmilitary spread spectrum systems have increased rapidly the last decades. The applications are quite diverse: underwater communications [11], wireless local loop systems [12], wireless local area networks, cellular systems, satellite communications [13], and ultra wideband systems [14]. Spread spectrum is also used in wired application in, for example, power-line communication [15] and have been proposed for communication over cable-TV networks [12] and optical fiber systems [16,17]. Finally, spread spectrum techniques have been found to be useful in ranging, such as, radar and navigation and the Global Positioning System (GPS) [18]. Other applications are watermarking of multimedia [19] and (mentioned here as a curiosity) in clocking of high-speed electronics [20,21]. Due to space constraint, we will only briefly mention some of the hot wireless applications here.

The wireless local area network (WLAN) standard IEEE 802.11 was originally designed to operate in the ISM band at approximately 2.4 GHz. The standard supports several different coding and modulation formats and several data rates. The first version of the standard was released in 1997 and supports both FH and DS spread spectrum formats with data rates of 1 or 2 Mbit/s [22]. The FH modes are slow hopping and use so-called Gaussian

FSK (GFSK) modulation (binary for 1 Mbit/s and 4-ary for 2Mbit/s). The system hops over 79 subcarriers with 1-MHz spacing. The DS-SS modes use a 11-chip long Barker sequence which is periodically repeated for each symbol. The chip rate is 11 Mchips/s, and the symbol rate is 1 Msymbols/s. The modulation is differentially encoded BPSK or QPSK (for 1 and 2 Mbit/s, respectively). We note that the processing gain is rather low, especially for the DS-SS modes.

The 802.11 standard has since 1997 been extended in several directions (new bands, higher data rates, etc). In 1999, the standard was updated to IEEE 802.11b (also known as Wi-Fi, if the equipment, also passes an interoperability test). In addition to the original 1 and 2 Mbit/s modes, IEEE 802.11b also supports 5.5 and 11 Mbit/s DS-SS modes [23] and several other optional modes with varying rates. The higher rate DS-SS modes uses so-called complementary code keying (CCK). The chip rate is still 11 Mchips/s and each symbol is represented by 8 complex chips. Hence, for the 5.5-Mbit/s mode, each symbol carries 4 bits, and for the 11-Mbit/s mode, each symbol carries 8 bits. Hence, the processing gains is reduced compared to the 1- and 2-Mbit/s modes. As a matter of fact, the 11 Mbit/s is perhaps not even a spread-spectrum system. The CCK modulation is a little bit complicated to describe, but in essence it forms the complex chips by combining a block code and differential QPSK [22], Section 18.4.6.5].

Bluetooth is primarily a cable replacement system, that is, a system for short range communication with relatively low-data rate. It is designed for the ISM band and uses slow hopping FH-SS with GFSK modulation ($BT = 0.5$ and modulation index between 0.28 and 0.35). The system hops over 79 subcarriers with a rate of 1600 hop/s. The subcarrier spacing is 1 MHz, and in most countries the subcarriers are placed at $f_k = 2402 + k$ MHz for $k = 0, 1, \ldots, 78$. Bluetooth supports both synchronous and asynchronous links and several different coding and packet schemes. The user data rates varies from 64 kbits/s (symmetrical and synchronous) to 723 kbits/s (asymmetrical and asynchronous). The maximum symmetrical rate is 434 kbits/s. The range of the system is quite short, probably less than 10 m in most environments. It is likely that future versions of Bluetooth will support higher data rates and longer ranges.

The first cellular system with a distinct spread spectrum component was IS-95 (also known as cdmaOne or somewhat pretentiously as CDMA). Although the Global System for Mobile Communications (GSM), has a provision for frequency hopping, it is not usually considered to be a spread spectrum system. Often, spread spectrum and code-division multiple access (CDMA) are used as synonyms, although they really are not. A multiple access method is a method for allowing several links (that are not at the same geographical location) to share a common communication resource. CDMA is a multiple access method where the links are spread spectrum links. A receiver that is tuned to a certain user relies on the anti-jamming properties of the spread spectrum format to suppress the other users' signals.

IS-95 uses DS-SS links with a chip rate of 1.2288 Mchips/s and a bandwidth of (approximately) 1.25 MHz. In the downlink (forward link or base-station-to-terminal link) the chips are formed by a combination of convolutional encoding, repetition encoding, and scrambling. The chips are transmitted both in inphase and quadrature (but scrambled by different PN sequences). In the uplink (reverse link), the transmitted chips are formed by a combination of convolutional coding, orthogonal block coding, repetition coding, and scrambling. In the original IS-95 (IS-95A), the uplink was designed such that the detection could be done noncoherently. In the third generation evolution of IS-95, known as cdma2000, the modulation and coding has changed and the transmitted bandwidth tripled to allow for peak data rates exceeding 2 Mbit/s [24,–26].

Another third generation system is Wideband CDMA (WCDMA) [27]. WCDMA is a rather complex system with many options and modes. We will here only briefly describe the frequency-division duplex (FDD) mode. The FDD mode uses direct-sequence spreading with a chip rate of 3.84 Mchips/s. The chip waveform is a root-raised cosine pulse with roll-off factor 0.22, and the bandwidth of the transmitted signal is approximately 5 MHz. WCDMA supports many different information bit rates by changing the spreading factor (from 4 to 256 in the uplink and from 4 to 512 in the downlink) and the error control scheme (no coding, convolutional coding, or turbo coding); however, the modulation is QPSK with coherent detection in all cases. Today, the maximum information data rate is roughly 2 Mbits/s, but it is likely that future revisions of the standard will support higher rates through new combinations of spreading, coding, and modulation.

Acknowledgment

The author wishes to thank Professors Erik G. Strom, Tony Ottosson, and Arne Svensson for contributing Section 5 of this article.

BIOGRAPHY

Dr. John G. Proakis received the B.S.E.E. from the University of Cincinnati in 1959, the M.S.E.E. from MIT in 1961, and the Ph.D. from Harvard University in 1967. He is an Adjunct Professor at the University of California at San Diego and a Professor Emeritus at Northeastern University. He was a faculty member at Northeastern University from 1969 through 1998 and held the following academic positions: Associate Professor of Electrical Engineering, 1969–1976; Professor of Electrical Engineering, 1976–1998; Associate Dean of the College of Engineering and Director of the Graduate School of Engineering, 1982–1984; Interim Dean of the College of Engineering, 1992–1993; Chairman of the Department of Electrical and Computer Engineering, 1984–1997. Prior to joining Northeastern University, he worked at GTE Laboratories and the MIT Lincoln Laboratory.

His professional experience and interests are in the general areas of digital communications and digital signal processing and more specifically, in adaptive filtering,

adaptive communication systems and adaptive equalization techniques, communication through fading multipath channels, radar detection, signal parameter estimation, communication systems modeling and simulation, optimization techniques, and statistical analysis. He is active in research in the areas of digital communications and digital signal processing and has taught undergraduate and graduate courses in communications, circuit analysis, control systems, probability, stochastic processes, discrete systems, and digital signal processing. He is the author of the book *Digital Communications* (McGraw-Hill, New York: 1983, first edition; 1989, second edition; 1995, third edition; 2001, fourth edition), and co-author of the books *Introduction to Digital Signal Processing* (Macmillan, New York: 1988, first edition; 1992, second edition; 1996, third edition), *Digital Signal Processing Laboratory* (Prentice-Hall, Englewood Cliffs, NJ, 1991); *Advanced Digital Signal Processing* (Macmillan, New York, 1992), *Algorithms for Statistical Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2002), *Discrete-Time Processing of Speech Signals* (Macmillan, New York, 1992, IEEE Press, New York, 2000), *Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ: 1994, first edition; 2002, second edition), *Digital Signal Processing Using MATLAB V.4* (Brooks/Cole-Thomson Learning, Boston, 1997, 2000), and *Contemporary Communication Systems Using MATLAB* (Brooks/Cole-Thomson Learning, Boston, 1998, 2000). Dr. Proakis is a Fellow of the IEEE. He holds five patents and has published over 150 papers.

BIBLIOGRAPHY

1. Robert A. Scholtz, The origins of spread-spectrum communications, *IEEE Trans. Commun.* **30**(5): 822–854 (May 1982). (Part 1).
2. Robert A. Scholtz, Notes on spread-spectrum history, *IEEE Trans. Commun.* **31**(1): 82–84 (Jan. 1983).
3. R. Price, Further notes and anecdotes on spread-spectrum origins, *IEEE Trans. Commun.* **31**(1): 85–97 (Jan. 1983).
4. Marvin K. Simon, Jim K. Omura, Robert A. Scholtz, and Barry K. Levitt, *Spread Spectrum Communications Handbook*, revised edition McGraw-Hill, 1994.
5. Rodger E. Ziemer, Roger L. Peterson, and David E. Borth, *Introduction to Spread Spectrum Communications*, Prentice-Hall, 1995.
6. Robert C. Dixon, *Spread Spectrum Systems with Commercial Applications*, 3rd ed., John Wiley and Sons, 1994.
7. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications — A tutorial. *IEEE Trans. Commun.* **30**(5): 855–884 (May 1982).
8. M. Luise, U. Mengali, and M. Morelli, Synchronization in Digital Communications, *The Wiley Encyclopedia of Telecommunications*.
9. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
10. A. J. Viterbi and I. M. Jacobs, Advances in Coding and Modulation for Noncoherent Channels Affected by Fading, Partial Band, and Multiple-Access Interference, in A. J. Viterbi ed., *Advances in Communication Systems*, Vol. 4, Academic, New York, 1975.
11. Charalampos C. Tsimenidis, Oliver R. Hinton, Alan E. Adams, and Bayan S. Sharif, Underwater acoustic receiver employing direct-sequence spread spectrum and spatial diversity combining for shallow-water multiaccess networking, *IEEE J. Oceanic Engineering* **26**(4): 594–603 (Oct. 2001).
12. D. Thomas Magill, Spread spectrum techniques and applications in America, *Proceedings of International Symposium on Spread Spectrum Techniques and Applications*, 1–4, 1995.
13. D. Thomas Magill, Francis D. Natali, and Gwyn P. Edwards, Spread-spectrum technology for commercial applications, *Proceedings of the IEEE* **82**(4): 572–584 (April 1994).
14. Moe Z. Win and Robert A. Scholtz, Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communications, *IEEE Trans. Commun.* **48**(4): 679–691 (April 2000).
15. Denny Radford, Spread spectrum data leap through AC power wiring, *IEEE Spectrum* **33**(11): 48–53 (Nov. 1996).
16. Jawad A. Salehi, Code division multiple-access techniques in optical fiber networks — Part I: fundamental principles, *IEEE Trans. Commun.* **37**(8): 824–833 (Aug. 1989).
17. Jawad A. Salehi and Charles A. Brackett, Code division multiple-access techniques in optical fiber networks — Part II: Systems performance analysis. *IEEE Trans. Commun.* **37**(8): 834–842 (Aug. 1989).
18. Michael S. Braasch and A. J. van Dierendonck, GPS receiver architectures and measurements, *Proceedings of the IEEE* **87**(1): 48–64 (Jan. 1999).
19. Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon, Secure spread spectrum watermarking for multimedia, *IEEE Trans. Image Processing* **6**(12): 1673–1687 (Dec. 1997).
20. Harry G. Skinner and Kevin P. Slattery, Why spread spectrum clocking of computing devices is not cheating, *Proceedings 2001 International Symposium on Electromagnetic Compatibility* 537–540 (Aug. 2001).
21. S. Gardiner, K. Hardin, J. Fessler, and K. Hall, An introduction to spread-spectrum clock generation for EMI reduction, *Electronic Engineering* **71**(867): 75, 77, 79, 81 (April 1999).
22. IEEE standard for information technology–telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements-part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. IEEE Std 802.11–1997, November 1997. ISBN: 1-55937-935-9.
23. Supplement to IEEE standard for information technology–telecommunications and information exchange between systems–local and metropolitan area networks-specific requirements- part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Higher-speed physical layer extension in the 2.4 GHz band. IEEE Std 802.11b- 1999, January 2000. ISBN: 0-7381-1811-7.
24. Daisuke Terasawa and Jr. Edward G. Tiedemann, cdmaOne (IS-95) technology overview and evolution. In *Proceedings IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, 213–216, June 1999.
25. Douglas N. Knisely, Sarath Kumar, Subhasis Laha, and Sanjiv Nanda, Evolution of wireless data services: IS-95 to cdma2000, *IEEE Commun. Mag.* **36**(10): 140–149 (Oct. 1998).
26. Theodore S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ 2001.

27. 3rd generation partnership project; technical specification group radio access network; physical layer - general description (release 4), 3GPPTS 25.201 V4.1.0 (2001-12). Online: *http://www.3gpp.org. Accessed March 20, 2002*.

STATISTICAL CHARACTERIZATION OF IMPULSIVE NOISE*

ARTHUR A. GIORDANO
AG Consulting Inc., LLC
Burlington, Massachusetts

THOMAS A. SCHONHOFF
Titan Systems Corporation
Northboro, Massachusetts

1. INTRODUCTION

A statistical characterization of impulsive noise is a multifaceted task. For the characterization to be meaningful and useful in predicting communications performance, knowledge of the underlying physical mechanisms is required. In order to develop a statistical noise model that accurately represents the communications channel under investigation, physical measurements of the impulsive noise characteristics are generally needed. With this understanding, a statistical noise model can be developed and used to construct optimum or near-optimum detectors, estimate physical system and noise parameters, and determine communications performance. As a consequence, there is no single noise model that can be used as a representation of all impulsive noise channels. This article describes a number of important noise models where the reader is cautioned to first confirm that the appropriate noise model is selected for the communications channel under investigation.

Gaussian noise, which is the predominant noise source associated with thermal noise in receiver front-ends and numerous communications channels such as microwave and satellite channels, is analytically tractable and used to predict communications performance for a broad class of communications systems. Noise statistics are characterized by a Gaussian probability density function (pdf) where only knowledge of the mean and variance of the noise is needed to completely characterize the noise statistics. The spectral characteristics of the noise may be white or nonwhite, with both cases having been extensively investigated [23]. Perhaps the most common and simple case is associated with additive white Gaussian noise (AWGN) where a number of analytical performance results have been obtained [33]. In fact, Gaussian noise is so prevalent that all other cases are categorized as non-Gaussian. This categorization is perhaps unfortunate in that it leads to the assumption that non-Gaussian noise has a single representation.

Impulse noise is typically associated with noise pulses that have large peak amplitudes and bandwidths that generally exceed the receiver bandwidth. The tails of the pdf of impulse noise are greater in extent than that of Gaussian noise and often lead to moments of the distribution that are ill defined. In contradistinction to the Gaussian case where only the pdf is needed, statistical characterization of an impulsive noise process requires multiple statistics for a complete characterization. The most extensively studied statistic is acknowledged to be the amplitude probability distribution (APD) and is defined as the probability that the noise envelope exceeds a specified value. This statistic is referred to as a first-order statistic. A complete characterization of the noise process requires higher-order statistics generally associated with the time statistics of the noise. Examples of these statistics include the autocorrelation, the pulse spacing distribution (also termed pulse interarrival times), the pulse duration distribution, and the average envelope crossing rate.

When a Gaussian noise model is accepted as a meaningful model for the communications channel under investigation, it is well known [48] that the optimum detector is linear and is implemented as either a matched filter or a correlation receiver.[1] A consequence of this result is that bit error rate (BER) computations are often analytically tractable for a large class of modulations. In contrast, for an impulsive noise process the optimum detector is nonlinear with a structure that is dependent on the noise model utilized. Thus, many different receiver structures can be derived where each structure is optimized for the specified noise model.

To determine the detector structure for an impulsive noise model, two approaches, referred to as analytical and ad hoc, are followed. The analytical approach is based on an accurate model of the noise statistics developed from the physical processes that generate the noise. The ad hoc approach is based on suboptimal detector structures that utilize nonlinearities that are easily implemented and reduce the tails in the noise distribution resulting in improved performance over linear detectors operating in the same impulsive noise environment. This latter approach has the advantage that the receiver performance is likely to be less sensitive to the noise model and therefore may offer more robust performance in time-varying or unknown noise statistics.

2. CHARACTERIZATION OF IMPULSIVE NOISE

Common examples of communication channels with impulsive noise include atmospheric radio noise, man-made noise, telephone communications, underwater acoustic noise, and magnetic recording noise. Atmospheric radio noise, arising from lightning discharges in the atmosphere, is an electromagnetic interference that can seriously degrade receiver performance [43,44]. Man-made noise occurs in automotive ignitions [22], electrical machinery such as welders, power transmission and distribution lines [32], medical and scientific apparatus, and so on. In telephone communications impulsive noise is generated

* This article is adapted, with permission, from a chapter in a textbook to be published by Prentice-Hall.

[1] In the non–white noise case, the detector uses a whitening filter prior to detection.

from telephone switches and is particularly important in characterizing the performance of high-speed digital subscriber loops [20,21,34,46]. These examples represent a diverse subset of impulsive noise environments where knowledge of the physical characteristics dictates the noise model selection.

To maintain generality but provide a basis for relating the noise model to a physical case, the important case of atmospheric radio noise will be emphasized. Although much of the treatment that follows is specific to this channel, the process is likely to be extendable to other physical channels.

A summary of several principal noise models is provided.

2.1. Statistical-Physical Model

The noise $\mathbf{n}(t)$ is represented over an interval T by a summation of impulses filtered by the channel and expressed as

$$\mathbf{n}(t) = \sum_{i=1}^{\mathbf{N}} \mathbf{a}_i \delta(t - \mathbf{t}_i) \tag{1}$$

where $\mathbf{N}$ is the number of impulses occurring in the interval T, $\mathbf{a}_i$ is a random variable (rv) representing the strength of the ith impulse and $\mathbf{t}_i$ is the occurrence time of the ith impulse.

$\mathbf{N}$ is typically assumed to be Poisson distributed with parameter μT where μ is the average rate of arrival of the impulses. Other investigators postulate a Poisson-Poisson distribution [8] or a Pareto [24] distribution leading to different pulse clustering behavior than that of a Poisson. For example, in the Poisson-Poisson case clusters of noise pulses occur where the pulses within a cluster occur at a Poisson rate μ_p and the clusters themselves are Poisson at a slower rate than μ_p.

The rv $\mathbf{a}_i$ can be often assumed to be Gaussian [34]. However, for the atmospheric radio noise case, the statistical physical model is well founded based on measurements and analytical modeling of the physical behavior of the communications environment. Giordano has shown that the rv $\mathbf{a}_i$ is proportional to the received field strength which depends on the source to receiver distance $\mathbf{r}_i$ in accordance with a generalized propagation law g, where $\mathbf{a}_i = g(\mathbf{r}_i)$ [9]. By assuming a spatial distribution of noise sources and a specific propagation law, the pdf of the rv $\mathbf{a}_i$ can be determined.

Middleton has extended the statistical-physical model developed in Giordano by introducing more general assumptions on the noise spatial conditions and propagation assumptions [25,26,28]. These models are termed "canonical" in that their form is invariant of the physical source mechanisms. Two important cases, referred to as Class A and Class B, were developed. Class A models are used for the narrowband case where the noise bandwidth is comparable or less than the receiver front-end bandwidth and Class B models are used in the wideband case.[2]

[2] That is, the noise bandwidth is wider than the receiver front-end bandwidth.

A Class C model has also been defined consisting of a combination of Class A and B models.

2.2. Generalized-t or Hall Model

This model, developed by Hall [13], has the form represented by a the product of a zero mean, narrowband Gaussian process $\mathbf{n}_g(t)$ with variance σ_1^2 and a slowly varying modulating stationary process $\mathbf{a}(t)$ that is independent of $\mathbf{n}_g(t)$, that is

$$\mathbf{n}(t) = \mathbf{a}(t)\mathbf{n}_g(t) \tag{2}$$

This model is premised on the fact that the noise sources vary with time over a large dynamic range and that, unlike a Gaussian noise source where energy is delivered at a constant rate, the impulses tend to occur in bursts. To model atmospheric radio noise, the slowly varying modulating process $\mathbf{a}(t)$ is selected to behave in a way that is similar to that of an empirical model obtained from measurements.

A variant of this model referred to as a truncated Hall model [35] removes some of the undesirable attributes of the Hall model where moments associated with the APD are undefined. This model is obtained by truncating the pdf of the noise envelope thereby producing finite moments and a noise process that is physically realizable. It has a further advantage in that the impulsiveness of the noise can be specified by means of the parameter Vd, which is the rms to average envelope ratio.

2.3. Mixture Model

The mixture model [29] is obtained by judiciously selecting two noise models which can be added together to produce the desired noise statistics. Thus, the term mixture model actually represents a collection of models since the two subsidiary noise processes can be chosen from many possibilities. One form of the mixture model is represented as a zero mean noise process $\mathbf{n}_0(t)$ that is,

$$\mathbf{n}_0(t) = (1 - \varepsilon_r)\mathbf{n}_g(t) + \varepsilon_r \mathbf{n}_i(t) \tag{3}$$

where $\mathbf{n}_g(t)$ is a Gaussian noise process, $\mathbf{n}_i(t)$ is an impulsive noise process and ε_r is a constant with $0 < \varepsilon_r < 1$. (An alternate formulation can be obtained by adding a combination of noise envelopes.) This model is appealing in that for small amplitudes approaching zero, $\mathbf{n}_0(t)$ approaches a Gaussian distribution ($\varepsilon_r \to 0$); for large noise amplitudes which create "spikes" of noise, the heavy tails of the impulsive noise distribution predominate ($\varepsilon_r \to 1$).

2.4. Empirical Model

The empirical model is based on graphical fits of the APD [7]. This model assumes that the small amplitude part of the APD is represented by a Rayleigh distribution and that the large amplitude region can be represented by a log-normal or power Rayleigh distribution. The parameters for the distributions are based on measurements of three statistical moments which are the average noise envelope, the rms noise

envelope, and the average logarithm of the noise envelope. Denoting the noise envelope as $v(t)$, the average noise envelope V_{ave} over an interval T is given by

$$V_{\text{ave}} = 1/T \int_0^T v(t)\,dt \tag{4}$$

Similarly, the rms noise envelope V_{rms} and average logarithm of the noise envelope $V_{\log}$ are given respectively by

$$V_{\text{rms}} = \left[1/T \int_0^T v^2(t)\,dt\right]^{1/2} \tag{5}$$

and

$$V_{\log} = \text{antilog}\left(1/T \int_0^T \log v(t)\,dt\right) \tag{6}$$

These moments are ordinarily measured in one bandwidth and are converted to other bandwidths as needed [39].

3. APD COMPUTATION USING HANKEL TRANSFORMS

When the statistical-physical model given in Eq. (1) is adopted, a very powerful method based on Hankel transforms is available to compute the APD and relate it to the underlying physical environment [9]. This method has been utilized in the atmospheric noise channel but is not restricted to that case. The method in fact applies to any narrowband receiver output representation when the input is a time sequence of independent impulses. An outline of the approach is presented here with a more complete derivation provided in Ref. 9.

Let us assume that a narrowband receiver with a carrier frequency f_c has an impulse response given by

$$h(t) = \beta(t)\cos(\omega_c t - \psi)u(t) \tag{7}$$

where $\beta(t)$ is the amplitude of $h(t)$, ψ is the phase, $\omega_c = 2\pi f_c$, and $u(t)$ is the unit step function. If the noise process given by Eq. (1) is applied to the input of a receiver with an impulse response given by Eq. (7), the output noise process $\mathbf{n}_0(t)$ can be expressed as

$$\mathbf{n}_0(t) = \sum_{i=1}^{\mathbf{N}} \mathbf{v}_i(t)\cos(\omega_c t - \omega_c \mathbf{t}_i - \psi)u(t - \mathbf{t}_i) \tag{8}$$

where $\mathbf{v}_i(t) = \mathbf{a}_i\beta(t - \mathbf{t}_i)$. The output noise can also be expressed in terms of its envelope $\mathbf{v}(t)$ and phase $\psi_s(t)$ as

$$\mathbf{n}_0(t) = \mathbf{v}(t)\cos(\omega_c t + \psi_s(t)) \tag{9}$$

The $\mathbf{N}$ terms in Eq. (8) can now be viewed as $\mathbf{N}$ vectors (or phasors) where the ith vector amplitude is $\mathbf{v}_i(t) = \mathbf{a}_i\beta(t - \mathbf{t}_i)$ with a phase $\psi_i = -\omega_c \mathbf{t}_i - \psi$ as shown in Fig. 1. The vectors then sum to the total vector with amplitude $\mathbf{v}(t)$ and phase $\psi_s(t)$.

The Hankel transform is the tool required to relate $\mathbf{v}(t)$ to $\mathbf{v}_i(t)$ for a fixed value of $\mathbf{N} = k$. The method is based on the characteristics function (cf) theorem for determining the pdf of a sum of independent random vectors where the cf of the sum vector is the product of the cf's of the individual vectors. It is now assumed

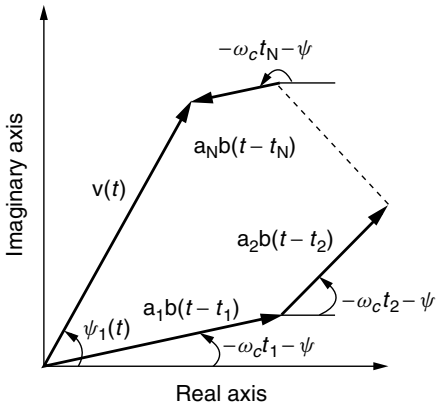


Figure 1. Phasor representation of output noise process.

that the phase of each vector ψ_i is uniformly distributed. Then for a sum of independent random vectors with uniformly distributed phase, the Hankel transform of the magnitude of the sum vector can be found by forming the product of the Hankel transforms of the magnitudes of the individual vectors.

Let $H_{\mathbf{v}_i}(z)$ denote the Hankel transform of the ith vector and $H_{\mathbf{v}_{N=k}}(z)$ denote the Hankel transform of the sum of $\mathbf{N} = k$ vectors. Then, the above description leads to

$$H_{\mathbf{v}_{N=k}}(z) = \prod_{i=1}^{k} H_{\mathbf{v}_i}(z) \tag{10}$$

Reference 9 shows that the Hankel transform of the ith vector can be defined as the expected value of the $rvJ_0(z\mathbf{v}_i)$ that is,

$$H_{\mathbf{v}_i}(z) = E[J_0(z\mathbf{v}_i)] \tag{11}$$

or equivalently

$$H_{\mathbf{v}_i}(z) = E[J_0(z\mathbf{a}_i\beta(t - \mathbf{t}_i))] \tag{12}$$

Because the individual vectors are identically distributed and the number of vectors $\mathbf{N}$ is a Poisson distributed rv, it can be shown that $H_{\mathbf{v}}(z)$, the Hankel transform of the envelope has a convenient representation in terms of the envelope of the receiver impulse response $\beta(t)$, the rate of arrival of the impulses μ, and the pdf of $\mathbf{a}_i$ given by $f_{\mathbf{a}}(a_i)$. The actual form derived in Ref. 9 is

$$H_{\mathbf{v}}(z) = \exp\left\{-\mu \int_0^\infty f_{\mathbf{a}}(a_i) \int_0^T [1 - J_0(za_i\beta(s))]\,ds\,da_i\right\} \tag{13}$$

Because $\mathbf{a}_i = g(\mathbf{r}_i)$, an alternative form of Eq. (13) can be obtained in terms of the pdf of the source to receiver distance $f_{\mathbf{r}}(r_i)$ that is,

$$H_{\mathbf{v}}(z) = \exp\left\{-\mu \int_0^\infty f_{\mathbf{r}}(r_i) \int_0^T [1 - J_0(zg(r_i)\beta(s))]\,ds\,dr_i\right\} \tag{14}$$

The pdf of the envelope can now be obtained by computing the inverse Hankel transform, that is,

$$p_{\mathbf{v}}(v) = v \int_0^\infty z H_{\mathbf{v}}(z) J_0(zv)\, dz \tag{15}$$

for $v \geq 0$ and 0 otherwise.

Reference 9 also shows that the cumulative distribution function (cdf) of the envelope can be expressed as

$$F_{\mathbf{v}}(v) = v \int_0^\infty H_{\mathbf{v}}(z) J_1(zv)\, dz \tag{16}$$

for $v \geq 0$ and 0 otherwise.

The APD is defined as the probability that the envelope $\mathbf{v}$ exceeds a value v_0 and is represented as $P[\mathbf{v} > v_0]$. It is given in terms of the cdf

$$P[\mathbf{v} > v_0] = 1 - F_{\mathbf{v}}(v_0) \tag{17}$$

or equivalently for $v_0 > 0$

$$P[\mathbf{v} > v_0] = 1 - v_0 \int_0^\infty H_{\mathbf{v}}(z) J_1(zv_0)\, dz \tag{18}$$

for $v_0 > 0$.

Example. In the case of atmospheric noise we now assume that lightning discharge occur in a uniform spatial distribution about the receiver out to a specified maximum range r_m resulting in a pdf given by

$$f_{\mathbf{r}}(r_i) = 1/r_m \tag{19}$$

where $0 < r_i < r_m$.

It is further assumed that the individual impulses arrive at the receiver in accordance with an inverse propagation law $\mathbf{a}_i = k_0/\mathbf{r}_i$, where k_0 is a constant. Then, it can be shown using Eqs. (14) and (18) that the APD for $v > 0$ can be expressed as [9]

$$P[\mathbf{v} > v_0] = K_u/[K_u^2 + {v_0}^2]^{1/2} \tag{20}$$

where

$$K_u = \mu k_0/r_m \int_0^T \beta(s)\, ds \tag{21}$$

This APD form has been shown to be a reasonable fit to measured atmospheric noise APDs [9,13]. Note that other assumptions on spatial distributions and propagation laws will produce other APD forms.

4. ATMOSPHERIC NOISE CHANNEL MODELS

One of the most extensively investigated impulsive noise channels is the atmospheric radio noise channel. Communications systems with center frequencies below 100 MHz must operate in the presence of atmospheric noise that arise from lightning discharges as a result of storms occurring throughout the world. Experimental data on some statistical properties of atmospheric noise can be found in several publications [5,14,15].

Receivers operating in atmospheric noise may experience two types of noise behavior, that is, 1) highly impulsive noise from local storms associated with distances that are within 1000 Km and 2) continuous noise from distant storms. With local storms radiated energy travels along the ground with low attenuation so that the receiver is subjected to strong electrical fields over a short duration. Storms that occur at ranges greater than 1000 Km propagate in modes within the earth-ionosphere cavity and arrive at the receiver with only a small portion of their original energy. In this latter case large numbers of lightning discharges occur simultaneously at various locations throughout the world causing bursts from distant storms to be numerous and overlap in time thereby producing the continuous background noise. As a result, APDs tend to have two dominant regions where the low-amplitude region is Rayleigh and arises via the central limit theorem from many independent, weak components whereas the high-amplitude region is dominated by strong distinguishable impulses that follow another distribution such as a power Rayleigh, log-normal, and so on.

The receiver bandwidth and operating frequency are also significant in characterizing impulsive noise. Very low frequency (VLF) (3–30 KHz) receivers experience significant interference as a result of the large radiated energy from the short (100 μsec) main strokes (also referred to as the return stroke) of lightning discharges which tend to be centered around 10 KHz [45]. Inan [16] provides recent progress and results in this band. The radiated spectrum from a lightning discharge decays with frequency with the result that the predischarge consisting of several stepped discrete leaders prior to the stronger main stroke produces interference in higher frequency bands. With regard to the receiver bandwidth a receiver that uses a wide bandwidth will be subjected to individual pulses from lightning discharges that are more easily distinguished so that the noise appears impulsive. Narrow bandwidths produce overlapping pulses resulting in noise that appears more continuous.

Another mechanism that strongly impacts the time statistics of atmospheric noise is due to multiple stroking. Multiple stroking from "long" (200 μsec) discharges usually consists of three or four return strokes spaced at about 40 msec apart. This phenomenon accounts for the departure from a random distribution in time and introduces the dependencies evident in the pulse spacing distributions.

Researchers have expended considerably less effort in modeling higher-order statistics that are important in characterizing receiver performance. One study that has investigated the effects of time dependencies in the noise pulses, involves noise measurements in the medium frequency band (300KHz–3MHz) [11]. In this case BER performance for linear and nonlinear receivers was obtained and compared by simulation using channels that included either a truncated Hall model or measured noise having pulse dependencies. The results show that when the bursty nature of the noise is neglected by assuming independent noises samples, the performance of nonlinear receivers over linear receivers is significantly greater than if pulse dependencies are incorporated in

the analysis. Typically, in measured noise it is noted that nonlinearities can improve the performance of a linear receiver by an amount that is on the order of the V_d value.

With the above limited explanation of the underlying physical mechanisms, we now return to more detailed descriptions of atmospheric noise models. The models introduced in Section 2 will be extensively described and subsequently used to estimate bit error rate performance in selected cases. Other performance results on coherent and noncoherent signaling in impulsive noise can be found in Refs. 1,4,6,30,40,41.

4.1. Hall Model

The Hall model presented above is repeated here and used to compute the envelope, that is

$$\mathbf{n}(t) = \mathbf{a}(t)\mathbf{n}_g(t) \tag{22}$$

The envelope $\mathbf{v}(t)$ can now be expressed as

$$\mathbf{v}(t) = |\mathbf{a}(t)\mathbf{n}_g(t)| \tag{23}$$

By selecting a "two-sided" Chi distribution for $\mathbf{b}(t) = 1/\mathbf{a}(t)$, the envelope distribution fits empirical data for large values of the envelope. The pdf of $\mathbf{b}(t)$ with parameters m and σ^2 is then given by

$$p(b) = \frac{(\frac{m}{2})^{m/2}}{\sigma^m \Gamma(\frac{m}{2})} |b|^{m-1} \exp\left(-\frac{m}{2\sigma^2} b^2\right) \tag{24}$$

Note that for $m = 1$ the pdf of $\mathbf{b}(t)$ is Gaussian so that the pdf of $\mathbf{n}_0(t)$ is the ratio of two Gaussian processes. Hall then shows that the pdf of $\mathbf{n}_0(t)$ is given by

$$p(n_0) = \frac{\Gamma(\frac{\theta}{2})}{\Gamma(\frac{\theta-1}{2})} \frac{\gamma^{\theta-1}}{\sqrt{\pi}(n_0^2 + \gamma^2)^{\theta/2}} \tag{25}$$

where $\gamma = \sqrt{m}\frac{\sigma_1}{\sigma}$ and $\theta = m + 1 > 1$. For $\sigma_1 = \sigma$ the above equation is a Student t distribution with parameter m so that in the general case Hall named Eq. (25) as a generalized t distribution with parameters θ and γ. Letting $\mathbf{x}(t)$ and $\mathbf{y}(t)$ denote the inphase and quadrature components of $\mathbf{n}_0(t)$ respectively, Hall derives the joint pdf $p_{\mathbf{xy}}(x, y)$ from a transformation of the product of random variables in Eq. (2) resulting in

$$p_{\mathbf{xy}}(x, y) = \frac{(\theta - 1)\gamma^{\theta-1}}{2\pi} \frac{1}{[x^2 + y^2 + \gamma^2]^{(\theta+1)/2}} \tag{26}$$

If we let $\mathbf{x} = \mathbf{v}\cos\psi_s$ and $\mathbf{y} = \mathbf{v}\sin\psi_s$ where $\mathbf{v}$ and ψ_s denote the envelope and phase respectively of the noise, we can write the joint pdf of the envelope and phase as

$$p_{\mathbf{v}\psi_s}(v, \psi_s) = \frac{v(\theta - 1)\gamma^{\theta-1}}{2\pi[v^2 + \gamma^2]^{(\theta+1)/2}} \tag{27}$$

From this equation, it is seen that ψ_s is independent of $\mathbf{v}$ and has a uniform pdf in the interval $(0, 2\pi)$ so that

$$\begin{aligned} p(v) &= \frac{1}{2\pi} \int_0^{2\pi} p_{\mathbf{v}\psi_s}(v, \psi_s) d\psi_s \\ &= \frac{(\theta - 1)\gamma^{\theta-1} v}{[v^2 + \gamma^2]^{(\theta+1)/2}}, \quad 0 \le v \le \infty \end{aligned} \tag{28}$$

Asymptotic forms of the pdf are

$$p(v) \approx \begin{cases} (\theta - 1)\gamma^{\theta-1}/v^\theta, & \text{for } v \text{ large} \\ (\theta - 1)v/\gamma^2, & \text{for } v \text{ small} \end{cases} \tag{29}$$

The form of the pdf for large v is consistent with empirical data and for small v follows a limiting form of a Rayleigh distributed envelope.

The APD or exceedance distribution can be shown to be

$$P(\mathbf{v} > v_0) = \int_{v_0}^{\infty} p(v)\, dv = \frac{\gamma^{\theta-1}}{(v_0^2 + \gamma^2)^{(\theta-1)/2}} \tag{30}$$

To fit measured data θ is typically taken to be an integer between 2 and 5 and γ is related to the average or rms value of v for the specified value of θ. Note that if $\theta = 2$ the above equation takes the same form as Eq. (20).

The APD $P(\mathbf{v} > v_0)$ is plotted as a function of v_0/γ for $\theta = 2, 3, 4$, and 5. The Rayleigh exceedance distribution is

$$P(\mathbf{v} > v_0) = \exp\left(-\frac{v_0^2}{2\gamma^2}\right)$$

and is shown in Fig. 2 for reference purposes.[3]

Examination of Fig. 2 shows that longer tails are exhibited with $\theta = 2$ rather than with $\theta = 3, 4$, or 5. This behavior is consistent with impulsive noise that has a large dynamic range. A measure of the impulsiveness of the noise is V_d, the ratio of the rms to average envelope value defined as

$$V_d = 20 \log\left(\frac{\sqrt{\mu_2}}{\mu_1}\right) \tag{31}$$

where

$$\mu_j = \int_0^{\infty} v^j p(v)\, dv \quad j = 1, 2 \tag{32}$$

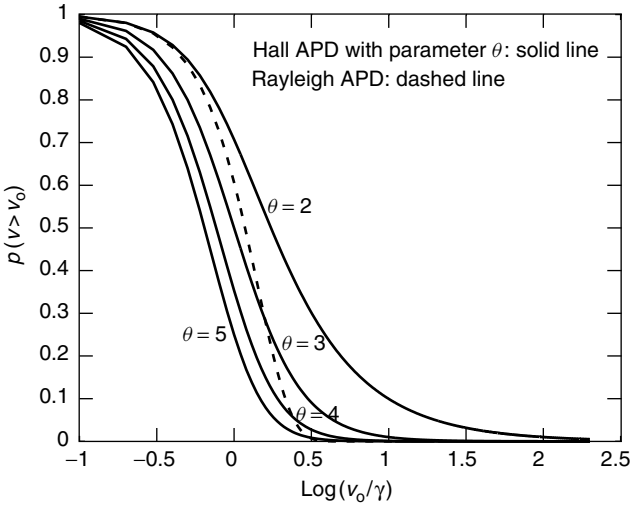


Figure 2. APD for Hall model and Rayleigh random variables.

[3] These results are obtained with MATLAB.

For $\theta = 4$ and 5 $V_d = 3$ dB and 2 dB respectively and the noise is considered to be moderately impulsive. For $\theta = 3$ the second moment does not exist and for $\theta = 2$ neither the first or second moments exist. One explanation of the infinite moments can be attributed to the propagation model where the received field strength follows an inverse distance relation. Near the origin the field strength is arbitrarily large which does not correspond to a physical condition. The actual noise moments must be finite and can be forced to this condition by truncating and normalizing the envelope distribution. Thus, a Hall model envelope pdf can be developed for the truncated case and is

$$p_E(v) = \begin{cases} \dfrac{c(\theta-1)\gamma^{\theta-1}v}{[v^2+\gamma^2]^{(\theta+1)/2)}} & 0 \le v \le v_m \\ 0, & v > v_m \end{cases} \tag{33}$$

where v_m is the maximum allowed envelope level and c is selected to ensure that

$$\int_0^\infty p_E(v)\,dv = 1 \tag{34}$$

Applying the above equation, we can show that

$$c = \frac{D^{\theta-1}}{D^{\theta-1}-1} \tag{35}$$

where $D = \sqrt{1+(v_m/\gamma)^2}$. For $\theta = 2$ the envelope pdf of the truncated Hall model is

$$p_E(v) = \begin{cases} \dfrac{D\gamma}{D-1}\dfrac{v}{[v^2+\gamma^2]^{3/2}}, & 0 \le v \le \gamma\sqrt{D^2-1} \\ 0, & v > \gamma\sqrt{D^2-1} \end{cases} \tag{36}$$

and

$$V_d = 20\log\left(\frac{(D-1)^{3/2}}{-\sqrt{D^2-1}+D\ln(D+\sqrt{D^2-1})}\right) \tag{37}$$

The truncated Hall model APD is given by

$$P(\mathbf{v} > v_0) = \frac{D}{D-1}\left[\frac{1}{\left(1+\dfrac{v_0^2}{\gamma^2}\right)^{1/2}} - \frac{1}{D}\right] \tag{38}$$

With this formulation more impulsive noise conditions can be realized and larger values of V_d can be obtained.

An advantage of the truncated Hall model is that computations can be performed for a specified V_d. For example, a highly impulsive case with $V_d = 10$ dB can be obtained by use of a normalized truncation point $v_m/\gamma = 290$. Other values for the normalized truncation point can be obtained from the plot given in Fig. 3.

4.2. Mixture Model

As described previously, mixture models are obtained by a "mixing" two simpler noise models. (See, for example, [35], [31], or [29]), that is

$$\mathbf{n}_0(t) = (1-\varepsilon_r)\mathbf{n}_g(t) + \varepsilon_r\mathbf{n}_i(t) \tag{39}$$

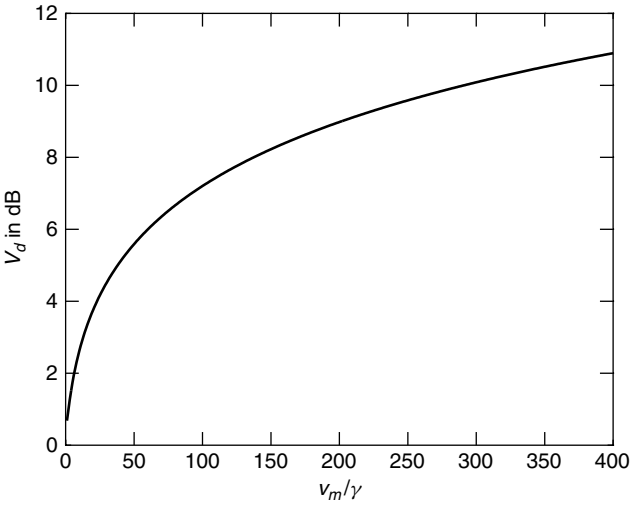


Figure 3. V_d as a function of truncation level for Hall model.

An alternate formulation of the mixture model can be obtained in terms of the noise envelope

$$\mathbf{v}(t) = \left(\frac{\mathbf{v}_i(t)+\mathbf{v}_R(t)}{2}\right) + \left(\frac{\mathbf{v}_i(t)-\mathbf{v}_R(t)}{2}\right)\mathbf{u}(t) \tag{40}$$

where $\mathbf{v}_R(t)$ is a Rayleigh envelope corresponding to the Gaussian noise component, $\mathbf{v}_i(t)$ is the envelope of the impulsive noise component, and $\mathbf{u}(t)$ is $+1$ with probability ε and -1 with probability $1-\varepsilon$. Thus, when $\mathbf{u}(t) = 1$, the envelope is Rayleigh distributed and given by

$$p(v_R) = \begin{cases} \dfrac{v_R}{\sigma^2}\exp\left(-\dfrac{v_R^2}{2\sigma^2}\right), & v_R \ge 0 \\ 0, & \text{otherwise} \end{cases} \tag{41}$$

and when $\mathbf{u}(t) = -1$, the envelope is distributed in accordance with an assumed distribution. One specific impulsive envelope pdf which has been used is referred to as a generalized Laplacian [35] and is given by

$$p(v_i) = \frac{v_i^v}{2^{v-1}(P_r\sigma)^{v+1}\Gamma(v)}K_{1-v}\left(\frac{v_i}{P_r\sigma}\right), \quad v_i \ge 0 \tag{42}$$

and zero when $v_i < 0$ where $K_{1-v}()$ is a modified Bessel function of order $1-v$ with $v > 0$ and P_r controls the ratio of the power between the impulsive and Rayleigh portions of the noise. The pdfs of the inphase and quadrature components can be determined from the joint pdf of the envelope v and phase ψ_s similar to that of Eq. (27) resulting in

$$p_{xy}(x,y) = \frac{(x^2+y^2)^{(v-1)/2}}{2\pi(P_r\sigma)^{v+1}\Gamma(v)2^{v-1}}K_{1-v}\left(\frac{\sqrt{x^2+y^2}}{P_r\sigma}\right) \tag{43}$$

The pdf of the inphase (or quadrature) components is then, from item 6.596-3 in [12],

$$p(x) = \frac{|x|^{v-1/2}}{\sqrt{\pi}(P_r\sigma)^{v+1/2}\Gamma(v)2^{v-1/2}}K_{1/2-v}\left(\frac{|x|}{P_r\sigma}\right) \tag{44}$$

This equation becomes a Laplace pdf when $v = 1$ and $P_r = 1/\sqrt{2}$ by using the relationships on page 444 in Ref. 2, that is,

$$K_{1/2}(z) = \left(\frac{\pi}{2z}\right)^{1/2} e^{-z}$$

and

$$K_{1/2-v}(z) = K_{v-1/2}(z)$$

The Laplace pdf then becomes, from Ref. 18

$$p(x) = \frac{1}{\sqrt{2\sigma^2}} \exp\left(-|x|\sqrt{\frac{2}{\sigma^2}}\right), -\infty < x < \infty \tag{45}$$

Note that the Laplace pdf has zero mean and variance σ^2 so that the Gaussian and Laplacian random variables have the same mean and variance.

By computing the first and second moments of the envelope of the mixture model, the V_d ratio for the mixture model can then be determined, that is,

$$V_d = 20 \log \left[\frac{\sqrt{4vP_r^2\varepsilon + 2(1-\varepsilon)}}{\dfrac{\varepsilon P_r \sqrt{\pi}\,\Gamma(v+1/2)}{\Gamma(v)} + (1-\varepsilon)\sqrt{\dfrac{\pi}{2}}} \right] \tag{46}$$

For $v = 1$ this reduces to

$$V_d\,|_{v=1} = 20 \log \left[\frac{\sqrt{4P_r^2\varepsilon + 2(1-\varepsilon)}}{\varepsilon P_r \dfrac{\pi}{2} + (1-\varepsilon)\sqrt{\dfrac{\pi}{2}}} \right] \tag{47}$$

With $P_r = 1$ and $\varepsilon = 0$, corresponding to the Gaussian only case, the V_d ratio is 1.05 dB. For $\varepsilon = 1$ and $v = 1$ the noise is impulsive according to a Laplacian distribution and $V_d = 20 \log 4/\pi = 2.1$ dB so that the noise is only mildly impulsive. More impulsive noise cases can be obtained by allowing $v < 1$ to be small in Eq. (46) Plots of V_d for several values of P_r and v are computed and shown in Figs. 4, 5, and 6 as a function of ε.

The APD for the mixture model can be obtained from

$$\begin{aligned} P\{\mathbf{v} > v_0\} &= P\{\mathbf{v} > v_0 \mid \mathbf{u} = 1\}P\{\mathbf{u} = 1\} \\ &\quad + P\{\mathbf{v} > v_0 \mid \mathbf{u} = -1\}P\{\mathbf{u} = -1\} \end{aligned} \tag{48}$$

Using item 6.561–12 in Ref. 12, this can be shown to be

$$\begin{aligned} P\{\mathbf{v} > v_0\} &= P\{\mathbf{v_i} > v_0\}\varepsilon + P\{\mathbf{v}_R > v_0\}(1-\varepsilon) \\ &= \varepsilon \int_{v_0}^{\infty} \frac{v_i^v}{2^{v-1}(P_r\sigma)^{v+1}\Gamma(v)} K_{1-v}\left(\frac{v_i}{P_r\sigma}\right) dv_i \\ &\quad + (1-\varepsilon)\int_{v_0}^{\infty} \frac{v_R}{\sigma^2} \exp\left(-\frac{v_R^2}{2\sigma^2}\right) dv_R = \varepsilon \left(\frac{v_0}{P_r\sigma}\right)^v \\ &\quad \times \frac{K_v\left(\dfrac{v_0}{P_r\sigma}\right)}{2^{v-1}\Gamma(v)} + (1-\varepsilon)\exp\left(-\frac{v_0^2}{2\sigma^2}\right) \end{aligned} \tag{49}$$

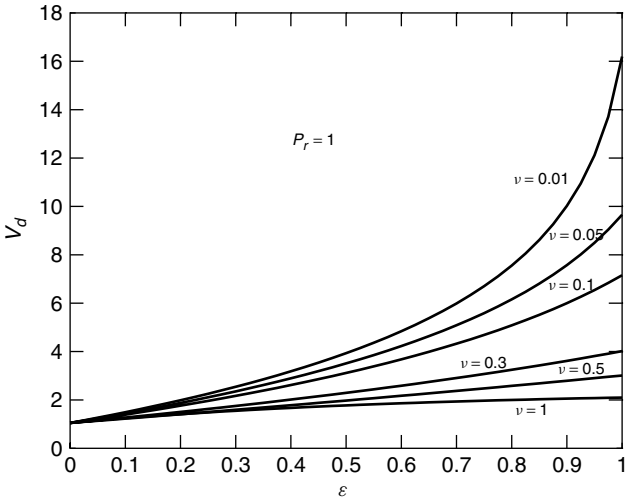


Figure 4. V_d as a function of v and ε with $P_r = 1$.

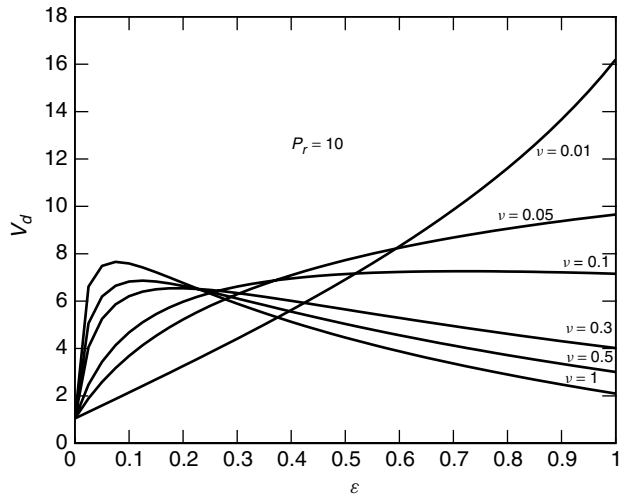


Figure 5. V_d as a function of v and ε with $P_r = 10$.

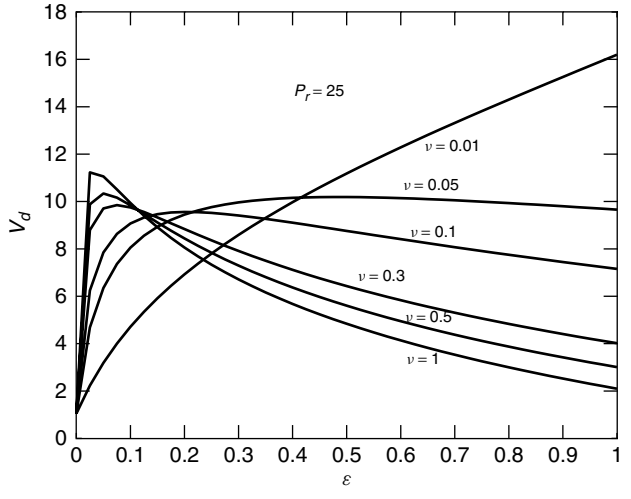


Figure 6. V_d as a function of v and ε with $P_r = 25$.

4.3. Middleton Class A and B Models

Middleton's models are statistical physical models like that given in Eq. (1) but with more general assumptions regarding noise source spatial distributions and propagation conditions. See Refs. 25–27,38. Thus, these models are regarded as canonical because noise source spatial distributions and propagation formulas need not be explicitly specified. Both Class A and B models assume that the noise sources are Poisson distributed in space with received waveforms produced by interfering sources that are independent and Poisson distributed in time. The broadband Class B models are useful in representing atmospheric noise statistics.

For Class A noise the instantaneous amplitude has a pdf described in terms of a parameter A referred to as the impulsive index and a parameter $P_r = \sigma_g^2/\sigma_i^2$ representing the ratio of the Gaussian noise power σ_g^2 to the impulsive noise power σ_i^2. The Class A instantaneous amplitude pdf is given by Ref. 49 as

$$p(w) = e^{-A} \sum_{m=0}^{\infty} \frac{A^m e^{-w^2/2\sigma^2\sigma_m^2}}{m!\sqrt{2\pi\sigma_m^2\sigma^2}} \tag{50}$$

where $\sigma^2 = \sigma_g^2 + \sigma_i^2$ and

$$\sigma_m^2 = \frac{\dfrac{m}{A} + P_r}{1 + P_r} \tag{51}$$

Middleton uses normalized instantaneous amplitudes defined as $x = w/\sigma$ so that

$$p(x) = e^{-A} \sum_{m=0}^{\infty} \frac{A^m}{m!\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{x^2}{2\sigma_m^2}\right) \tag{52}$$

Small values of the parameter A produce large impulsive tails whereas large values of $A >\approx 10$ yield the limiting case of Gaussian interference. Thus, the reciprocal of A behaves as V_d in the truncated Hall and mixture models. Middleton also derives the pdf of the normalized envelope defined as $\mathrm{v}_N = \mathrm{v}/\sqrt{2\sigma^2}$ resulting in

$$p(v_N) = 2e^{-A} \sum_{m=0}^{\infty} \frac{A^m}{m!\sigma_m^2} v_N \exp\left(-\frac{v_N^2}{\sigma_m^2}\right), \quad v_N \geq 0 \tag{53}$$

and zero for $v_N < 0$.

A special case of a mixture model can be obtained by splitting the above expression into two terms as

$$p(v_N) = \frac{2e^{-A}}{\sigma_0^2} v_N e^{-v_N^2/\sigma_0^2} + 2e^{-A} \sum_{m=1}^{\infty} \frac{A^m}{m!\sigma_m^2} v_N \times \exp\left(-\frac{v_N^2}{\sigma_m^2}\right), \quad v_N \geq 0 \tag{54}$$

where

$$\sigma_0^2 = \frac{P_r}{1+P_r} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_i^2} \tag{55}$$

The first term in Eq. (54) can be seen to be Rayleigh distributed whereas the second term represents an impulsive distribution. The normalized form of the APD is obtained from

$$P\{\mathbf{v}_N > v_{N_0}\} = \int_{v_{N_0}}^{\infty} p(v_N)\, dv_N \tag{56}$$

where $v_{N_0} = N_0/\sqrt{2\sigma^2}$ resulting in

$$P\{\mathbf{v}_N > v_{N_0}\} = e^{-A} \sum_{m=0}^{\infty} \frac{A^m}{m!} \exp\left(-\frac{v_{N_0}^2}{\sigma_m^2}\right) \tag{57}$$

For Class B noise the instantaneous noise amplitude can be written in normalized form as

$$p(x) = \frac{e^{-x^2/W}}{\pi\sqrt{W}} \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} A_\alpha^m \Gamma\left(\frac{m+1}{2}\right) {}_1F_1\left(-\frac{m\alpha}{2}; \frac{1}{2}; \frac{x^2}{W}\right) \tag{58}$$

where ${}_1F_1$ is the confluent hypergeometric function, A_α is a parameter that includes the impulsive index A and other parameters that depend on the physical mechanism, α is a constant between 0 and 2 related to the noise source density and propagation law, and W is a parameter that normalizes the noise process to the energy contained in the Gaussian portion of the noise. As indicated in Ref. 42, the normalization cannot be associated with the total energy because the moments of Eq. (58) are not finite.

The pdf of the normalized envelope for Class B noise is

$$p(v_N) = \frac{2v_N}{W} \exp\left(-\frac{v_N^2}{W}\right) \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \times A_\alpha^m \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_1F_1\left(-\frac{m\alpha}{2}; 1; \frac{v_N^2}{W}\right), \quad v_N \geq 0 \tag{59}$$

and zero for $v_N < 0$. The normalized APD for Class B noise is

$$P\{\mathbf{v}_N > v_{N_0}\} = \exp\left(-\frac{v_{N_0}^2}{W}\right) \left[1 - \frac{v_{N_0}^2}{W} \sum_{m=1}^{\infty} \frac{(-1)^m}{m!} \times A_\alpha^m \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_1F_1\left(1 - \frac{m\alpha}{2}; 2; \frac{v_{N_0}^2}{W}\right)\right] \tag{60}$$

An alternate form for the APD of Class B noise is given by Ref. 26 as

$$P\{\mathbf{v}_N > v_{N_0}\} = 1 - \exp\left(-\frac{v_{N_0}^2}{W}\right) \frac{v_{N_0}^2}{W} \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \times A_\alpha^m \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_1F_1\left(1 - \frac{m\alpha}{2}; 2; \frac{v_{N_0}^2}{W}\right) \tag{61}$$

To see that Eqs. (60) and (61) are equivalent, we can rewrite the last equation as

$$P\{\mathbf{v}_N > v_{N_0}\} = 1 - \exp\left(-\frac{v_{N_0}^2}{W}\right) {}_1F_1\left(1; 2; \frac{v_{N_0}^2}{W}\right) - e^{-v_{N_0}^2/W} \frac{v_{N_0}^2}{W} \times \sum_{m=1}^{\infty} \frac{(-1)^m}{m!} A_\alpha^m \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_1F_1\left(1 - \frac{m\alpha}{2}; 2; \frac{v_{N_0}^2}{W}\right) \tag{62}$$

However, from Ref. 2,

$$ {}_1F_1(1;2;z) = \frac{e^z - 1}{z} \tag{63} $$

which allows the above equation to be written as

$$ P\{\mathbf{v}_N > v_{N_0}\} = \exp\left(-\frac{v_{N_0}^2}{W}\right) - \exp\left(-\frac{v_{N_0}^2}{W}\right)\frac{v_{N_0}^2}{W} \times \sum_{m=1}^{\infty} \frac{(-1)^m}{m!} A_\alpha^m \Gamma\left(1 + \frac{m\alpha}{2}\right) {}_1F_1\left(1 - \frac{m\alpha}{2}; 2; \frac{v_{N_0}^2}{W}\right) \tag{64} $$

For large values of the argument $z = v_{N_0}^2/W$ an approximate form of the APD, which is useful for numerical evaluation, can be obtained by replacing the confluent hypergeometric function with the expression, using Ref. 2,

$$ {}_1F_1(a;b;z) = \frac{\Gamma(b)}{\Gamma(a)} z^{a-b} e^z, \quad \text{large } z \tag{65} $$

Using Eq. (65) in Eq. (61) leads to

$$ \begin{aligned} P\{\mathbf{v}_N > v_{N_0}\} &= 1 - ze^{-z} \sum_{m=0}^{\infty} \frac{(-1)^m A_\alpha^m}{m!} \\ &\quad \times \Gamma\left(1 + \frac{m\alpha}{2}\right) \frac{\Gamma(2)}{\Gamma(1 - \frac{m\alpha}{2})} z^{-\frac{m\alpha}{2}-1} e^z \\ &= -\sum_{m=1}^{\infty} (-1)^m \frac{A_\alpha^m}{m!} \frac{\Gamma(1 + \frac{m\alpha}{2})}{\Gamma(1 - \frac{m\alpha}{2})} z^{-m\alpha/2}, \quad \text{large } z \end{aligned} \tag{66} $$

From Ref. 2, we can use the identities

$$ \Gamma\left(1 + \frac{m\alpha}{2}\right) = \frac{m\alpha}{2} \Gamma\left(\frac{m\alpha}{2}\right) \tag{67} $$

and

$$ \Gamma\left(\frac{m\alpha}{2}\right) \Gamma\left(1 - \frac{m\alpha}{2}\right) = \frac{\pi}{\sin \pi\alpha \frac{m}{2}} \tag{68} $$

in Eq. (66) resulting in

$$ P\{\mathbf{v}_N > v_{N_0}\} = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{A_\alpha^m}{(m-1)!} \frac{\alpha}{2\pi} \times \Gamma^2\left(\frac{m\alpha}{2}\right) \sin\left(\pi\alpha\frac{m}{2}\right) z^{-m\alpha/2}, \quad \text{large } z \tag{69} $$

An example of the APD of Middleton's Class B noise for $V_d = 10$ dB is given below in Section 4.5.

4.4. Empirical First-Order Model

The empirical first-order (simulation) noise model is based upon the Crichlow graphical model for the APD of the noise envelope [7]. This model uses a special type of probability graph paper; one on which the power Rayleigh functions plot as straight lines.[4] The coordinate transformations for the probability paper is found by considering the APD of a normalized Rayleigh function

$$ \frac{n}{\sqrt{\mu_2}} = \text{APD}(w) = e^{-w^2} \tag{70} $$

We use the transformation

$$ x' = -20 \log(-\ln w) \tag{71} $$

$$ y' = 20 \log \frac{n}{\sqrt{\mu_2}} \tag{72} $$

On this probability paper, the APD of atmospheric noise can be represented by a three-section curve as shown in Fig. 7. The lower region of the curve, representing random low-amplitude envelopes of high probability, approaches a Rayleigh distribution. Hence, it can be approximated by a straight line (R). The higher region of the curve, representing high-impulsive envelopes of low probability, approaches a power Rayleigh distribution. It can also be approximated by a straight line (PR). The center region of the curve corresponds to a circular arc tangent to the two straight lines. The circular arc is also tangent to the line (T) which is parallel to line (BI), which bisects the acute angle formed by the intersection of the Rayleigh and power Rayleigh lines. Four parameters are necessary to specify a unique pair of lines and an arc. They are:

1. Slope of the power Rayleigh line;
2. Point through which the power Rayleigh line passes;
3. Point through which the Rayleigh line passes (the slope is known to be $\frac{-1}{2}$);
4. Parameter determining the radius of the circular arc.

Crichlow defined the four parameters as follows:

1. $X = -2s$, where s is the slope of the power Rayleigh line;
2. C (dB) = the dB difference between the power Rayleigh line and the Rayleigh line at $p = 0.01$;
3. A (dB) = the dB value of the Rayleigh line at $p = 0.5$;
4. B (dB) = the dB difference between the y' -axis intercepts of lines (BI) and (T).

Experimentally measured APDs indicate that the parameter B is linearly related to first order to the parameter X by

$$ B = 1.5(X - 1) \tag{73} $$

Thus, once the values of parameters X, C, and A are known, a unique APD can be constructed. The X, C, and A parameters vary according to the V_d value of the APD. Wilson [47] has calculated the values of these parameters for V_d values of 4.0 through 30.0. Parameter values for $V_d = 2.0$ and 3.0 were determined by extrapolating upon Wilson's values along with experimental curve fitting. The above transformations were thoroughly tested with a total of 20,000 samples, for each of ten different V_d values. The APDs were then determined by experimental

[4] A Rayleigh function plots as a straight line with slope = −.5.

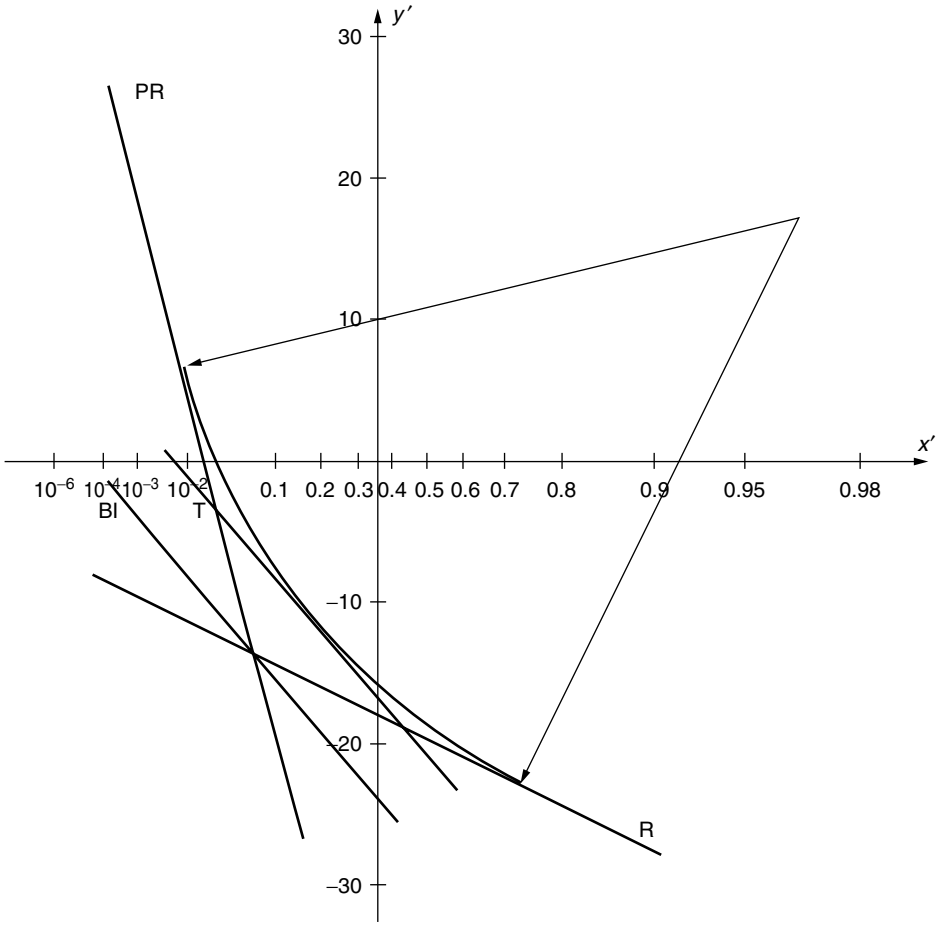


Figure 7. Crichlow APD model.

measurements. This can be seen in Fig. 8 for V_d ratios of 6, 10, and 14 dB. It is clear from the figure that more than 20,000 samples should be used to test the APD curves for values of Δ greater than 20.

4.5. Comparison of Selected APD Results for Different Models

In this section the Middleton Class B model, the truncated Hall model, and the empirical model are compared with CCIR 322 data [5]. The main point of this section is to show that numerous models can be used to obtain good fits to measured APD data and that no matter what model is selected, error rate performance estimates will be the same as long as the noise samples are independent.[5]

Figure 9 shows a comparison of the APD curves for the truncated Hall model with $V_d = 10$ dB, the Middleton Class B model with $A_\alpha = 1, \alpha = 1$, and $W = 0.007$, and CCIR 322 data with $V_d = 10$ dB. For reference purposes, the Rayleigh APD is also shown; it is known that the Rayleigh distribution corresponds to $V_d = 1.05$ dB. The

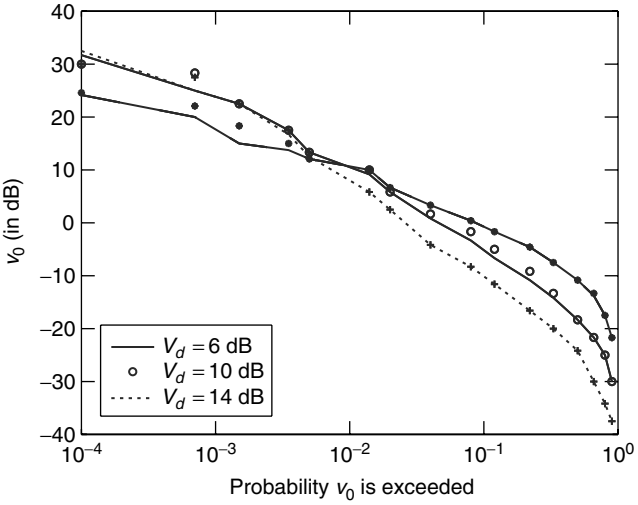


Figure 8. Verification of Crichlow APD using Wilson's parameters.

curves are plotted on semilog paper where the vertical scale is in dB above the rms level. For the trancated Hall case, the rms level is $\gamma\sqrt{D-1}$ and for the Rayleigh case the rms level is $\sqrt{2}\gamma$. The parameter W provides the normalization for the Class B Middleton model. A comparison of the empirical model for $V_d = 10$ dB

[5] Atmospheric noise measured data reveals that noise samples are correlated because of multiple lightning discharges. Nevertheless, it can be shown that the APD is essentially the same; higher order statistics will, however, be affected, although this behavior is rarely modeled. See Ref. 9.

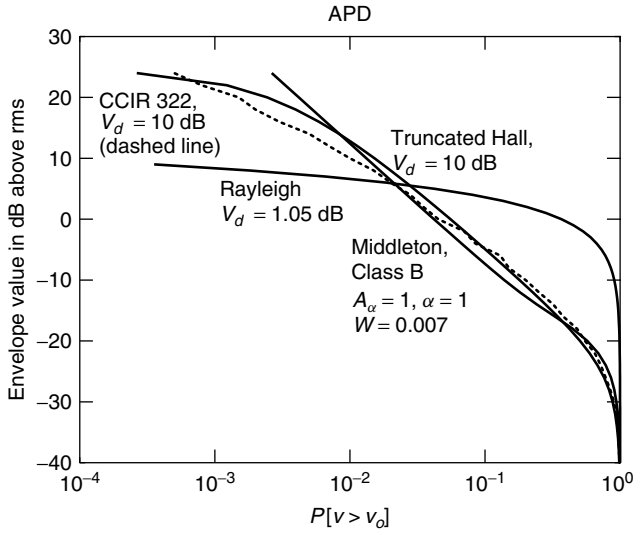


Figure 9. Comparison of some atmospheric noise models.

is presented in Fig. 8. Good fits with CCIR 322 data are obtained in these particular examples, especially for envelope values in dB above rms that are below about 20 dB. Computational problems arise with both the truncated Hall model and the Class B model for large envelope values. In the trancated Hall case the curve stops when the truncation point is exceeded. In the Class B case the exact formulation in Eq. (61) converges poorly so that the approximation of Eq. (66) is involved for $z > 100$. The Class B parameters were selected to provide good but not necessarily an optimum fit to the CCIR 322 data.

5. DETECTOR STRUCTURES IN NON-GAUSSIAN NOISE

At this point it is apparent that numerous models exist to represent non-Gaussian noise. The development of optimal structures would, in principle, require computation of the likelihood ratio or log-likelihood ratio. Because of the non-Gaussian noise, however, each model can in general lead to a unique optimum detector structure. This section describes a few of the common cases which have been derived and built. As indicated below in Section 5.1, an optimal detector structure under the conditions of small snr leads to the general form of a nonlinearity followed by a linear correlator or matched filter demodulation for making a decision. Figure 10 shows the general structure. A wide variety of nonlinearities have been used, including hard limiters, soft clippers, hole punchers, and logarithmic devices [10].

Section 5.2 shows one detector structure that has been developed based upon Hall's model. Finally, in Section 5.3, some commonly-used *ad-hoc* structures are presented.

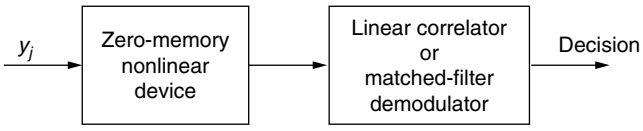


Figure 10. Common detector structure for impulsive noise.

5.1. Weak Signal Detection

A commonly derived receiver structure for the case of a weak signal is referred to as the locally optimum Bayes' detector (LOBD).(The LOBD is also called a low snr detector or a threshold receiver.) References 17, 19, and 40 give up-to-date derivations of this detector, although Ref. 3 gives one of the earliest presentations. Following the presentations in Refs. 3 and 40, we define two hypotheses as

$$\vec{\mathbf{y}} = \vec{\mathbf{u}}_i + \vec{\mathbf{n}}, \quad i = 0, 1 \tag{74}$$

where each vector contains k samples. The likelihood ratio is given by

$$L(\vec{y}) = \frac{p_1(\vec{y})}{p_0(\vec{y})} = \frac{p_{\mathbf{n}}(\vec{y} - \vec{u}_1)}{p_{\mathbf{n}}(\vec{y} - \vec{u}_0)} \tag{75}$$

We now expand the pdf $p_{\mathbf{n}}(\vec{y} - \vec{u})$ in a vector Taylor series and ignore, for small signals, terms of degree two and higher. This results in

$$p_{\mathbf{n}}(\vec{y} - \vec{u}_i) \simeq p_{\mathbf{n}}(\vec{y}) - \sum_{j=1}^{k} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_i} u_{ij} \tag{76}$$

Substituting this expression into the likelihood ratio above yields

$$L(\vec{y}) = \frac{p_{\mathbf{n}}(\vec{y}) - \sum_{j=1}^{k} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_i} u_{1j}}{p_{\mathbf{n}}(\vec{y}) - \sum_{j=1}^{k} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_i} u_{0j}} \tag{77}$$

Dividing the numerator and denominator by $p_{\mathbf{n}}(\vec{y})$ results in

$$\begin{aligned} L(\vec{y}) &= \frac{1 - \sum_{j=1}^{k} \frac{1}{p_{\mathbf{n}}(\vec{y})} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_i} u_{1j}}{1 - \sum_{j=1}^{k} \frac{1}{p_{\mathbf{n}}(\vec{y})} \frac{\partial p_{\mathbf{n}}(\vec{y})}{\partial y_i} u_{0j}} \\ &= \frac{1 - \sum_{j=1}^{k} \frac{d}{dy_j} \ln p_{\mathbf{n}}(y_j) u_{1j}}{1 - \sum_{j=1}^{k} \frac{d}{dy_j} \ln p_{\mathbf{n}}(y_j) u_{0j}} \end{aligned} \tag{78}$$

The unity constant does not affect the decision and can be ignored leading to the decision to choose H_1 if[6]

$$-\sum_{j=1}^{k} \frac{d}{dy_j} \ln p_{\mathbf{n}}(y_j) u_{1j} > -\sum_{j=1}^{k} \frac{d}{dy_j} \ln p_{\mathbf{n}}(y_j) u_{0j} \tag{79}$$

The receiver structure for this case is depicted in Fig. 11 and is often referred to as a threshold receiver, that is,

[6] These results assumes equal a priori probabilities and uniform costs.

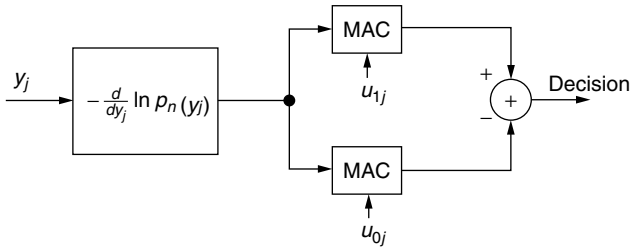


Figure 11. LOBD detector.

LOBD, because the signal is assumed to be small. It is apparent that the structure of Fig. 11 is an example of that of Fig. 10.

It is instructive to examine the form of Fig. 11 when the noise is white and Gaussian. In this case, the density is

$$p_{\mathbf{n}}(\vec{y}) = \frac{1}{(2\pi\sigma^2)^{k/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{k} y_j^2\right), \tag{80}$$

$$\ln p_{\mathbf{n}}(\vec{y}) = -\frac{k}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{k} y_j^2, \tag{81}$$

and

$$-\frac{d}{dy_j}\ln p_{\mathbf{n}}(\vec{y}) = \frac{y_j}{\sigma^2} \tag{82}$$

In other words, the nonlinearity, in this case, is linear. This is comforting for we know that the optimum detector in white Gaussian noise is linear.

An example of a non-Gaussian noise distribution for which the detector can be solved in its entirety is the Laplace distributed noise. In this case, k independent noise samples comprise the vector $\vec{\mathbf{n}}$ so that the received signal vector can be represented by

$$\vec{\mathbf{y}} = \vec{\mathbf{u}}_i + \vec{\mathbf{n}} \tag{83}$$

where $\vec{\mathbf{u}}_i,\ i = 0, 1$ is the transmitted signal vector. The pdf of $\vec{\mathbf{n}}$ is given by

$$p(\vec{n}) = \frac{1}{\sqrt{2\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}}\sum_{j=1}^{k}|n_j|\right) \tag{84}$$

Under hypothese $H_i, i = 0, 1$ the pdf of $\vec{y}$ becomes

$$p(\vec{y}_j) = \frac{1}{\sqrt{2\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}}\sum_{j=1}^{k}|y_j - u_{ij}|\right) \tag{85}$$

The log-likelihood ratio can now be computed as

$$\ell(\vec{y}) = \ln\frac{p_1(\vec{y})}{p_0(\vec{y})} = \sqrt{\frac{2}{\sigma^2}}\sum_{j=1}^{k}[|y_j - u_{0j}| - |y_j - u_{1j}|] \tag{86}$$

and the decision rule is to choose H_1 when[7]

$$\sum_{j=1}^{k}|y_j - u_{0j}| > \sum_{j=1}^{k}|y_j - u_{1j}| \tag{87}$$

Continuing with the example, if we let $u_{1j} = c$ and $u_{0j} = -c$ for all j and define the function $g(y_j)$ as

$$g(y_j) = |y_j + c| - |y_j - c|, \tag{88}$$

then the receiver structure can be obtained and is shown in Fig. 12 where the Laplace noise nonlinearity is shown in Fig. 13. This detector structure is of the general form of Fig. 11 but in this example it is not necessary to impose the weak signal restriction.

5.2. Hall's Log-Correlator

Hall derived the optimum receiver for Hall model noise by forming a likelihood ratio computed from the pdf of the complex envelope of the noise as shown in [13]. A few definitions are required before the pdf of the complex noise envelope can be obtained. Assume samples of the received noise are taken at a spacing Δt and let $\mathbf{v}_g(t)$ be the complex envelope of the noise $\mathbf{n}_g(t)$ in Eq. (2). We further assume that $\mathbf{v}_g(t)$ has the covariance

$$E\{\mathbf{v}_g(t_i)\mathbf{v}_g^*(t_j)\} = N_0\delta_{ij}, \quad i, j = 1, \ldots, k \tag{89}$$

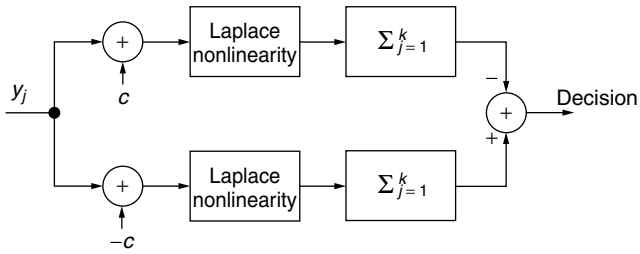


Figure 12. Detector structure for laplace noise.

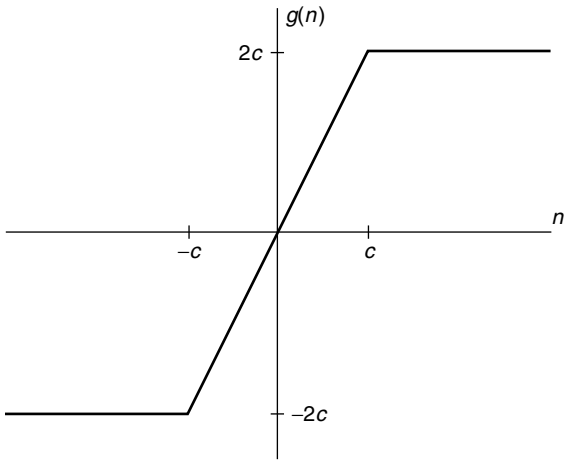


Figure 13. Laplace noise nonlinearity.

[7] This decision rule assumes equal a priori probabilities and uniform costs.

where k represents the number of received complex samples. If the noise signal has a duration T and $2B$ is the RF bandwidth of the receiver front-end, $k \approx 2BT$, and a good approximation of the complex noise envelope results.

The process $\mathbf{b}(t)$, which is the reciprocal of the slowly varying modulating process $\mathbf{a}(t)$ in Eq. (2) has a zero mean with a pdf given by Eq. (24) and a covariance that is assumed to be

$$E\{\mathbf{b}(t_i)\mathbf{b}(t_j)\} = \frac{B_0}{2}\delta_{ij}, \quad i,j = 1, \cdots, k \tag{90}$$

Let $\vec{\mathbf{z}}$ be a vector of k complex envelope samples for the Hall model in Eq. (2) for the special case $\theta = 2$. Hall then shows that the pdf of the noise in the independent sample case can be expressed as

$$p(\vec{z}) = \left(\frac{\gamma_2}{2\pi}\right)^k \prod_{j=1}^{k} \frac{1}{[|z_j|^2 + \gamma_2^2]^{3/2}} \tag{91}$$

where $\gamma_2^2 = (N_0\Delta t)/B_0$. Under hypothesis H_i, $i = 0, 1$, the received signal vector $\vec{\mathbf{y}}$ can be written as a sum of the complex signal vector $\vec{u}_i$ and the noise vector $\vec{\mathbf{z}}$ as

$$\vec{\mathbf{y}} = \vec{u}_i + \vec{\mathbf{z}}, \quad i = 0, 1 \tag{92}$$

and from Eq. (91) the pdf of $\vec{\mathbf{y}}$ is given by

$$p(\vec{y}) = \left(\frac{\gamma_2}{2\pi}\right)^k \prod_{j=1}^{k} \frac{1}{[|y_j - u_{ij}|^2 + \gamma_2^2]^{3/2}} \tag{93}$$

The log-likelihood ratio can then be written as

$$\ell(\vec{y}) = \ln \frac{p_1(\vec{y})}{p_0(\vec{y})} = \sum_{j=1}^{k} \ln[|y_j - u_{0j}|^2 + \gamma_2^2]^{3/2} - \sum_{j=1}^{k} \ln[|y_j - u_{1j}|^2 + \gamma_2^2]^{3/2} \tag{94}$$

Note that the power 3/2 has no effect on the decision so that an equivalent rule is to choose H_1 when

$$\sum_{j=1}^{k} \ln[= |y_j - u_{0j}|^2 + \gamma_2^2] > \sum_{j=1}^{k} \ln[|y_j - u_{1j}|^2 + \gamma_2^2] \tag{95}$$

Hall's log-correlator is depicted at baseband in Fig. 14. Note that the term γ_2^2 is referred to as the bias and must in general be estimated. Although the log-correlator receiver was developed for $\theta = 2$, Hall shows that the decision rule of Eq. (95) is the same when γ_2^2 is replaced by $\gamma^2 = m\gamma_2^2$.

Exact error rate computations for this receiver are intractable even with a central limit theorem argument which requires only the first and second moments of the log of the decision variable. Instead Hall computes bounds on the probability of error. Below, in Section 6.2, we present error rate results for the log-correlator and other receiver structures obtained by simulation.

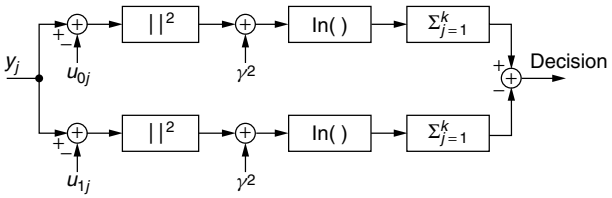


Figure 14. Log-correlator receiver.

5.3. Ad Hoc Receiver Structures

Optimal Zero Memory Nonlinear (ZMNL) devices are often difficult to implement in practice. Thus, more practical, yet suboptimal, nonlinearities are used. Figure 15 shows the three most common ad hoc nonlinearities. In Fig. 15a the transfer characteristics of a hole puncher are shown. In this case, the input signal is passed undistorted as long as its envelope is smaller than a threshold t_h. If any samples of the envelope are larger than t_h, these samples are totally suppressed.

In Fig. 15b, the transfer characteristics for a clipper ZMNL device is shown. Its characteristics are similar to a hole puncher except that if the envelope exceeds t_h, a constant output proportional to t_h is available rather than a total suppression of the signal. By allowing t_h to get very small relative to the signal plus noise envelope, the clipper approximates a hard limiter, whose characteristics are shown in Fig. 15c.

6. SELECTED EXAMPLES OF NOISE MODELS, RECEIVER STRUCTURES, AND ERROR RATE PERFORMANCE

Sample results are presented in this section for several noise models and corresponding receiver structures. No attempt is made to be exhaustive because in general, every noise model yields a different receiver structure and its associated error rate performance. Conversely, as mentioned previously, if the APD of the noise models is essentially the same, then the resulting error rate performance for any *specific* receiver structure will be similar for all the noise models, assuming independence of the noise samples.

The results provided here illustrate the available improvement in error rate performance when a nonlinear receiver is utilized in the presence of impulsive noise. Sample cases presented below include:

- Bandpass limiter and linear correlator in Gaussian noise, outlined in Section 6.1.

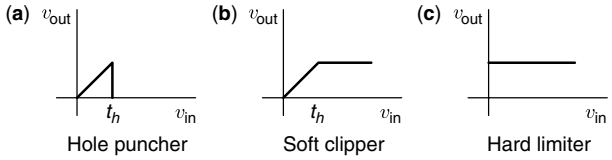


Figure 15. Ad Hoc zero memory nonlinearities used on atmospheric noise channel receivers.

- Bandpass limiter, linear correlator, and log-correlator performance in truncated Hall noise with a V_d of 10 dB, presented in Section 6.2.
- Weak signal receiver with Middleton noise, presented in Section 6.3, and
- Hole puncher and soft clipper in empirically generated noise with a V_d of 6 dB, detailed in Section 6.4.

In the impulsive noise cases multiple samples per symbol are assumed to allow the matched filter or linear correlator to average the residual noise following the nonlinearity. It can be shown that the case of a single sample per symbol provides no advantage from use a nonlinearity. Error rate performance is obtained by simulation for coherent antipodal signaling with the number of samples per symbol denoted by NSAM and the number of symbols used denoted by NSYM. For an RF receiver filter bandwidth denoted by $2B$ and a symbol duration of T, then NSAM $= 2BT$ so that the signal to noise ratio SNR is related to the energy contrast ratio E_b/N_0 by SNR*NSAM $= E_b/N_0$.

6.1. Bandpass Limiter and Linear Correlator in Gaussian Noise

Because the optimum receiver in independent Gaussian noise is a linear receiver, use of a bandpass limiter (or any other nonlinearity) in this case prior to the linear correlator will therefore degrade the error rate performance. Figures 16 and 17 show the BPL/linear correlator and linear correlator receivers respectively. These receivers structures were used in a simulation to estimate error rate performance. The simulated transmitted sequence was assumed to be symbols with values that are ± 1. White Gaussian noise was added to the transmitted sequence and then input to each receiver. Using MATLAB, error rate performance shown in Fig. 18 was obtained for NSAM $= 10$ and NSYM $= 5000$. The theoretical error rate is expressed as

$$P_e = \tfrac{1}{2}\text{erfc}\left(\sqrt{\frac{E_b/N_0}{\text{NSAM}}}\right) \tag{96}$$

The simulation shows good agreement between the linear correlator simulation and the theoretical performance and

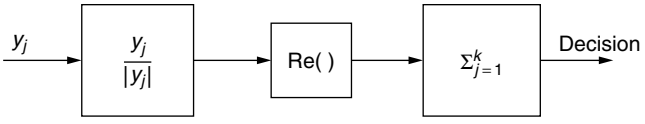


Figure 16. Bandpass limiter/linear correlator receiver.

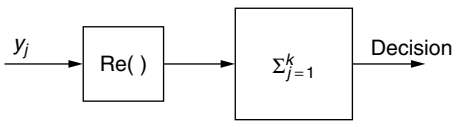


Figure 17. Linear correlator receiver.

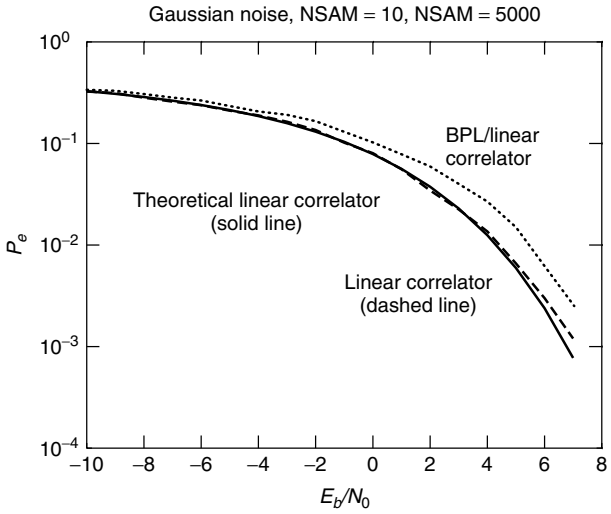


Figure 18. Error rate performance for a linear and BPL/linear receiver in Gaussian noise.

that the BPL performance is approximately 1-dB poorer than the linear receiver for larger values of E_b/N_0.

6.2. Bandpass Limiter, Linear Correlator, and Log-Correlator in Truncated Hall Noise

To generate the noise samples, in-phase and quadrature samples of the Hall model noise are computed from

$$\mathbf{n}_{0r} = |\mathcal{T}| \cos \mathbf{u} \tag{97}$$

and

$$\mathbf{n}_{0i} = |\mathcal{T}| \sin \mathbf{u} \tag{98}$$

where $\mathcal{T}$ is a random variable from a generalized t distribution and $\mathbf{u}$ is uniformly distributed in the interval $(-\pi, \pi)$. The random variable $\mathcal{T}$ is defined by

$$\mathcal{T} = \frac{\mathcal{G}}{\sqrt{\frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i^2}} \tag{99}$$

where $\mathcal{G}$ is a zero mean Gaussian random variable with variance σ_1^2 and the $\mathbf{x}_i$ are independent zero mean Gaussian distributed random variables with variance σ^2. Note that $\sum_{i=1}^{m} \mathbf{x}_i^2$ is chi-squared with m degrees of freedom and is independent of $\mathcal{G}$. Therefore $|\mathcal{T}|$, can be generated from the ratio of a Rayleigh distributed random variable $\mathbf{r}$ to the square root of a chi-squared random variable with m degrees of freedom. $\mathbf{r}$ in turn can be obtained from a transformation of a uniformly distributed random variable $\mathbf{u}$ on the interval (0, 1) by

$$\mathbf{r} = [-2\sigma_1^2 \ln \mathbf{u}]^{1/2} \tag{100}$$

The Hall bias term γ is computed from $\gamma^2 = m\sigma_1^2/\sigma^2$. The above procedure can also be used to compute noise samples from a truncated Hall distribution. In such a case, the samples obtained from the Hall model noise with $\theta = 2$

are generated and samples whose amplitude exceed the truncation value v_m are discarded.

Before presenting the simulation results for the truncated Hall model, a result for one sample per symbol in a linear receiver is derived. For symbols $\mathbf{u}$ that assume the values of $\pm a$ with equal probability, the error rate can be expressed as

$$P_e = P\{\mathbf{u} > 0 \mid \mathbf{u} = -a + \mathbf{n}\} \tag{101}$$

where $\mathbf{n}$ is the additive truncated Hall model noise with parameter γ, v_m, and $\theta = 2$. Because the pdf of $\mathbf{n}$ is

$$p(n) = \frac{1}{\pi\gamma[(\frac{n}{\gamma})^2 + 1]} \tag{102}$$

it follows that

$$P_e = \int_0^{n_m} \frac{du}{\pi\gamma[(\frac{u+a}{\gamma})^2 + 1]} \tag{103}$$

where n_m is the maximum value of the noise corresponding to the truncation point. For high V_d, the truncation point can be approximated as infinite so that the resulting error probability becomes

$$P_e \approx \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left(\frac{a}{\gamma}\right) \tag{104}$$

For $a = 1$, the average transmitted power is unity and the noise power is $\gamma^2(D-1)$ so that the parameter γ can be replaced by

$$\gamma = \sqrt{\frac{1}{(D-1)E_b/N_0}} \tag{105}$$

Figure 19 shows[8] the error rate results for the linear correlator, BPL/linear correlator, and log-correlator receivers for NSAM = 10 and NSYM = 10000; for reference the theoretical error rate for one sample per symbol and $a = 1$ is also shown. The curves are obtained by simulation using MATLAB. It can be seen that the log-correlator receiver is the best and the BPL/linear correlator is slightly degraded. Conversely, both nonlinear receivers are significantly better than the linear correlator. These results neglect pulse dependencies from multiple stroking where the bursty nature of measured atmospheric noise tends to limit the performance improvement of nonlinear receivers to an amount that is on the order of the V_d value [15,11].

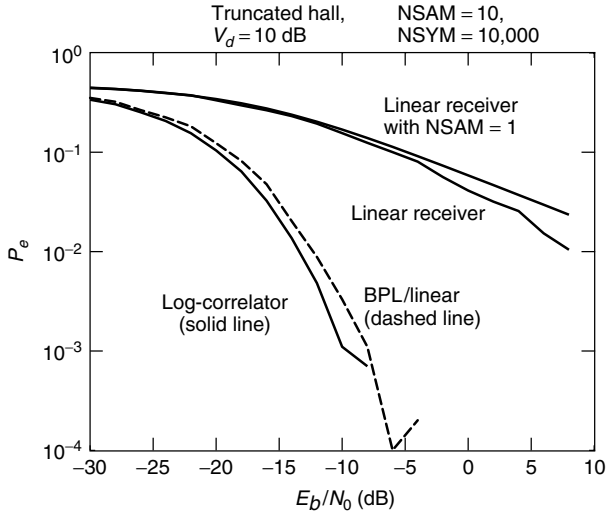


Figure 19. Error rate performance for various receivers in truncated hall noise.

[8] Note that some scatter occurs at low estimated error probabilities as a result of an insufficient number of samples for the estimate.

6.3. WEAK SIGNAL DETECTOR IN MIDDLETON NOISE

Returning to the result of Fig. 11, Spaulding and Middleton, in Refs. 37 and 40, provide a general result for the error rate performance of antipodal signals in Middleton's Class A noise as

$$P_e = \tfrac{1}{2}\text{erfc}\,(\sqrt{k\Gamma_o f/2}),\ \Gamma_o f \ll 1 \tag{106}$$

where Γ_o is the signal to noise ratio and f is defined as

$$f = \int_{-\infty}^{\infty}\left[\frac{d}{dy}\ln p_n(y)\right]^2 p_n(y)\,dy \tag{107}$$

Performance results for noncoherent signaling in Middleton's Class A noise and weak signals are provided by Spaulding and Middleton in Refs. 37 and 41, leading to an approximate error rate for the LOBD as

$$P_e = \tfrac{1}{2}\ \text{erp}\left(-\frac{k\Gamma_o f}{4}\right),\ \Gamma_o f \ll 1 \tag{108}$$

6.4. Hole Puncher and Clipper in Empirically Generated Noise

In this section the noise is generated using the empirical method described in Section 4.4 using $V_d = 6$ dB. The error rate performance, displayed in Figs. 20 and 21, for the soft clipper and hole puncher receivers respectively, are obtained parametric in the threshold t_h in dB above the average envelope. Although these figures are displayed using MATLAB, the original data is derived from a simulation reported on in Ref. 36. The simulated transmitted signal is Minimum Shift Keying (MSK) with a two-sided bandwidth of 900 Hz and a signaling rate of 100 bps. The best threshold for the soft clipper is 0 dB with little degradation for a setting of 3 dB. The best threshold for the hole puncher is 3 dB for high signal to noise ratios. It should be noted that the performance for both nonlinearities obtained with their optimum thresholds is about the same, as evident in Fig. 22. For reference, a Gaussian error rate curve for coherent detection of MSK is displayed in all three figures.

7. CONCLUSIONS

This article has provided a brief introduction on impulsive noise physical phenomena, impulsive noise statistics,

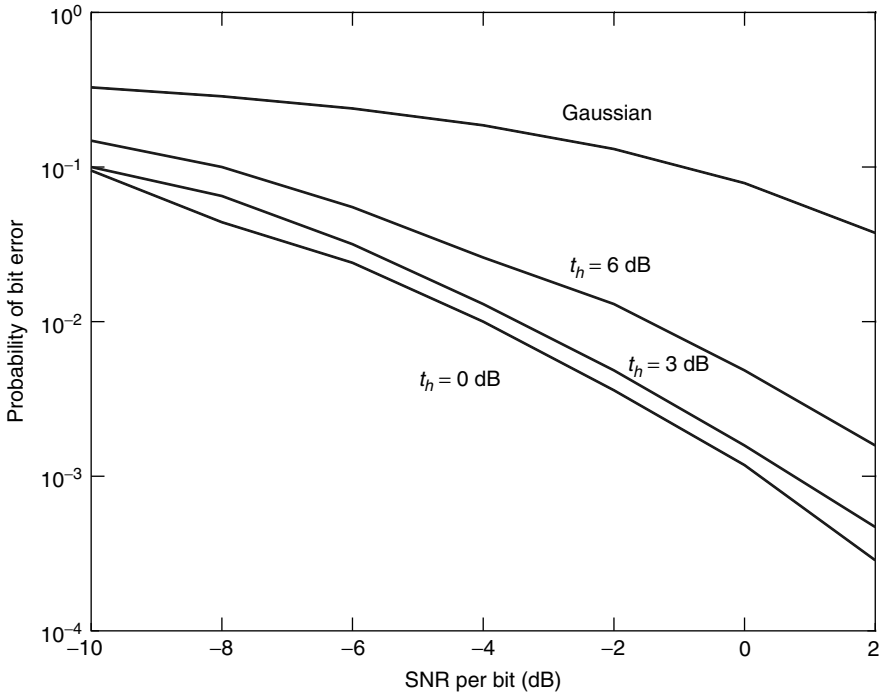


Figure 20. Probability of error as a function of SNR per bit with clipper threshold as a parameter.

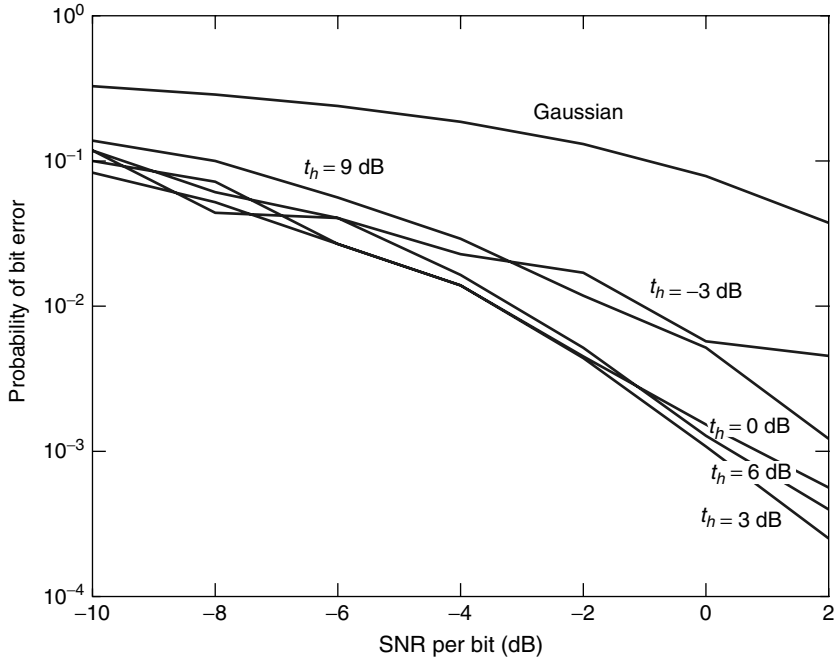


Figure 21. Probability of error as a function of SNR per bit with hole-puncher threshold as a parameter.

appropriate utilized receiver structures and some selected BER performance results. A few significant points that have been presented are now emphasized:

1. There is no single noise model that can be used as a representation of all impulsive noise channels.
2. Physical measurements of the impulsive noise characteristics are generally needed to understand the behavior of the noise allowing noise models, receiver structures, and performance estimates to be obtained.
3. Optimum receiver structures are nonlinear but are "optimum" only in terms of the noise model assumed.
4. Specific channels require knowledge of higher order statistics to completely characterize the noise process and produce reliable performance estimates.

It is anticipated that future work on the statistical characterization of impulsive noise for various communications channels will involve new measurements and continued emphasis on first order statistical models that

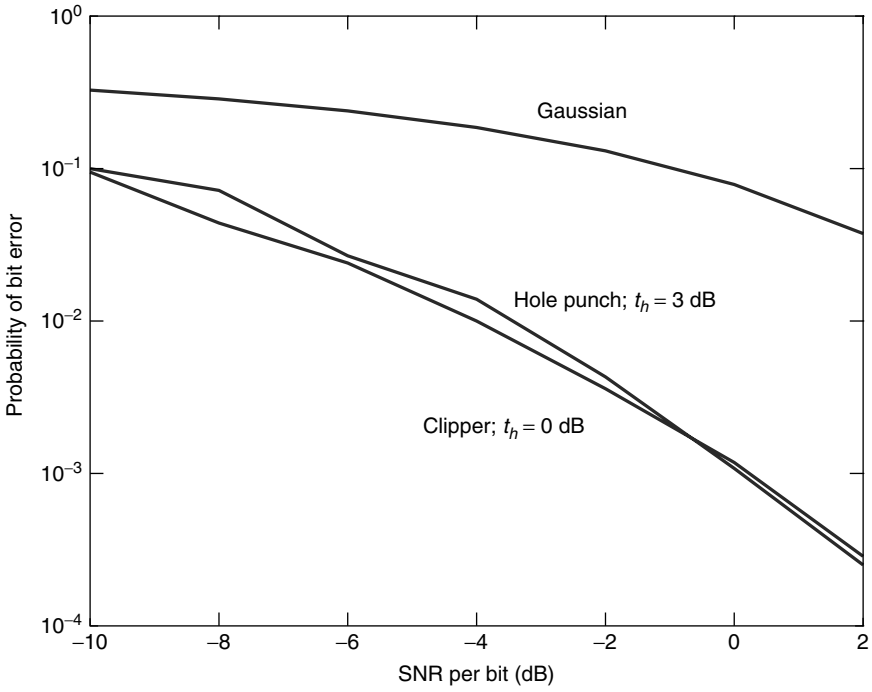


Figure 22. Comparison of bit error rates for two nonlinearities.

are analytically tractable. It is also expected that significant research will focus on the time statistics of the noise for the channel under investigation. Gratifying results may very well be attained by continued emphasis on the underlying physical process such as that developed for statistical physical models.

BIOGRAPHIES

Arthur A. Giordano formed AG Consulting, LLC in June 2001 to consult in the field of military and commercial communications. Specific consulting activities have included: 1) investigation of secure net broadcast for voice and data to evaluate the potential interoperability of commercial cellular systems including CDMA, GSM, TDMA, and next generation (3G) cellular systems with planned military systems; 2) investigation and documentation on a comparison of SMR, CDMA, 1xRTT, and GPRS 3) expert witness on cases involving RF coverage and next generation CDMA; and 4) development of a satellite channel simulator. From 1993 to June 2001, he worked as a wireless manager for GTE and later Verizon Laboratories and was primarily involved in the analysis, design, and deployment of cellular and fixed wireless networks. From 1985 to 1993, he was a vice president at CNR responsible for a critical government program to provide secure, formatted message traffic for a user network employing multimedia communications assets consisting of landlines, satellites, and HF radios. Prior to 1985, he held several engineering positions encompassing communications network architecture development, spread spectrum communications, modulation and coding designs, adaptive signal processing, and measurement and modeling of atmospheric noise. He has a Doctorate from the University of Pennsylvania in EE and MS and BS degrees in EE from Northeastern University. He has published numerous technical articles, holds two patents, has coauthored a book entitled Least Square Estimation with Applications to Digital Signal Processing and is currently coauthoring a text on Detection and Estimation Theory.

Thomas A. Schonhoff received his Bachelor's degree from M.I.T., his Master's degree from Johns Hopkins University, Baltimore, Maryland, and his Ph.D. from Northeastern University, Boston, Massachusetts. He has worked at six different corporations, although he has been with LinCom Corporation (now Titan System Corporation, Communication and Software Solutions Division) since 1985. For the last 21 years, Dr. Schonhoff has also taught graduate courses as an adjunct at Worcester Polytechnic Institute, Massachusetts.

BIBLIOGRAPHY

1. B. Aazhang and H. V. Poor, performance of DS/SSMA communications in impulsive channels. II hard-limiting correlation receivers, *IEEE Trans. Comm.* **36**(1): 88 an 1988.
2. M. Abramowitz and I. A. Stegun, ed., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, U.S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series 55, 1968.
3. O. Y. Antonov, Optimum Detection of Signals in Non-Gaussian Noise, *Radio Eng. Elec. Physics* **12**: 541–548 April 1967.
4. S. Buzzi, E. Conte, and M. Lops, Optimum Detection over Rayleigh Fading Dispersive Channels, with Non-Gaussian Noise, *IEEE Trans. Comm.* **45**(9): 1061–1069 (Sept. 1997).
5. CCIR Report 322, World distribution and characteristics of atmospheric radio noise, International Radio Consultative Committee, Genevsa, Feb. 7, 1964.

6. E. Conte, M. DiBisceglie, and M. Lops, Optimum detection of fading signals in impulsive noise, *IEEE Trans. Comm.* **43**(2): Part 3 869–876 (Feb.-March-April 1995).
7. W. Q. Crichlow et al., Determination of the amplitude-probability distribution of atmospheric radio noise from statistical moments, *J. Res. Natl. Bureau Stand.—D: Radio Propag.* **64D**: 49–56 (1960).
8. K. Furutsu and T. Ishida, On the theory of amplitude distribution of impulsive noise, *J. Appl. Phys.* **32**: 1206–1221 (July 1961).
9. A. Giordano and F. Haber, Modeling of atmospheric noise, *Radio Sci.* **7**(11): 1011–1023 (1972).
10. A. A. Giordano and H. E. Nichols, Simulated error rate performance of nonlinear receivers in atmospheric noise, *National Telecommunications Conference*, 1977.
11. A. A. Giordano et al., Measurement and statistical analysis of wide-band MF atmospheric radio noise, 2, impact of data on bandwidth and system performance, *Radio Sci.* **21**(2): 203–222 (Mar.–Apr. 1986).
12. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, corrected and enlarged edition, prepared by A. Jeffrey, Academic Press, 1980.
13. H. M. Hall, *A New Model for 'Impulsive' Phenomena: Application to Atmospheric-Noise Communication Channels*, Ph.D. dissertation, Stanford University, CA, Aug. 1966.
14. J. R. Herman et al., Measurement and statistical analysis of wide-band MF atmospheric radio noise, 1 structure and distribution and time variation of noise power, *Radio Sci.* **21**(1): 25–46 (Jan.–Feb. 1986).
15. J. R. Herman et al., considerations of atmospheric noise effects on wideband MF communications, *IEEE Commun. Mag.* 24–29 (Nov. 1983).
16. U Inan, Holographic Array for Ionospheric Lightning (HAIL), *http://www-star.stanford.edu/~vlf/*
17. S. A. Kassam, *Signal Detection in Non-Gaussian Noise*, Springer-Verlag, New York, 1988.
18. S. M. Kay *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
19. S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1998.
20. K. J. Kerpez and K. Sistanizadeh, High bit rate asymmetric digital communications over telephone loops, *IEEE Trans. Comm.* **43**(6): 2038–2049 (June 1995).
21. K. Kumozaki, Error correction performance in digital subscriber loop transmission systems, *IEEE Trans. Comm.* **39**(8): 1170–1174 (Aug. 1991).
22. W. R. Lauber and J. M. Bertrand, Statistics of motor vehicle ignition noise at VHF/UHF, *IEEE Trans. Electromag. Compat.* **41**(3): 257–259 (Aug. 1999).
23. R. N. McDonough and A. D. Whalen, *Detection of Signals in Noise*, 2nd ed., Academic Press, New York, 1995.
24. S. McLaughlin, The Non-Gaussian Nature of Impulsive Noise in Digital Subscriber Lines, *http://www.ee.ed.ac.uk/bin/pubsearch?McLaughlin*, Jan. 2000.
25. D. Middleton, Statistical-physical models of electromagnetic interference, *IEEE Trans. Electromagn. Compat.* **EMC-19**: 106–127 (Aug. 1977).
26. D. Middleton, Non-Gaussian noise models in signal processing for telecommunications: New methods and results for class a and b noise models, *IEEE Trans. Inform. Theory* **45**(4): 1129–1149 (May 1999).
27. D. Middleton and A. D. Spaulding, Elements of Weak Signal Detection in Non-Gaussian Noise Environments, *Advances in Statistical Signal Processing*, Vol. 2, Signal Detection, H. V. Poor and J. B. Thomas editors, JAI Press, Inc., Greenwich, CT, 1993.
28. D. Middleton, Procedures for determining the parameters of the first-order canonical models of class a and class b electromagnetic interference, *IEEE Trans. Electromagn. Compat.* **EMC-21**(3): 190–208 (Aug. 1979).
29. J. H. Miller and J. B. Thomas, Detection of signals in impulsive noise modeled as a mixture Process, *IEEE Trans. Comm.* **Com-24**: 559–563 (May 1976).
30. S. Miyamoto, M Katayama, and N Morinaga, Performance analysis of QAM systems under class a impulsive noise environment, *IEEE Trans. Electromagn. Compat.* **37**(2): 260–267 (May 1995).
31. J. W. Modestino, Locally optimum receiver structures for known signals in non-Gaussian narrowband Noise, 13th *Annual Allerton Conference on Circuit and System Theory*, Oct. 1975.
32. J. B. O'Neal Jr., Substation noise at distribution-line communications frequencies, *IEEE Trans. Electromagn. Compat.* **30**(1): 71–77 (Feb. 1988).
33. John G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
34. D. H. Sargrad and J. W. Modestino, Errors-and-erasure coding to combat impulse noise on digital subscriber loops. *IEEE Trans. Comm.* **38**(8): 1145–1155 (Aug. 1990).
35. T. A. Schonhoff, A. A. Giordano, and Z. McC, Huntoon, Analytical representation of atmospheric noise distributions constrained in V_d, 1977 International Conference on Communications.
36. T. A. Schonhoff, A differential GPS receiver system using atmospheric noise mitigation techniques, *Proceedings of the ION GPS—91*, September 1991.
37. A. D. Spaulding and D. Middleton, *Optimum Reception in an Impulsive Interference Environment*, U.S. Dept. of Commerce OT Report 75–67, June 1975.
38. A. D. Spaulding, Stochastic modeling of the electromagnetic interference environment, *1977 International Conference on Communications*, pp. 42.2–114, 42.2–123.
39. A. D. Spaulding, C. J. Roubique, and W. Q. Crichlow, Conversion of the amplitude probability distribution function for atmospheric radio noise from one bandwidth to another, *J. Research NBS* **66D**(6): 713–720 (Nov.-Dec.-Feb. 1962).
40. A. D. Spaulding and D. Middleton, Optimum reception in an impulsive interference environment—part I: Coherent detection, *IEEE Trans. Commun.* **COM-25**: 910–923 (Sept. 1977).
41. A. D. Spaulding and D. Middleton, Optimum reception in an impulsive interference environment—part II: Incoherent detection, *IEEE Trans. Commun.* **COM-25**: 924–934 (Sept. 1977).
42. A. D. Spaulding, Optimum threshold signal detection in broad-band impulsive noise employing both time and spatial sampling, *IEEE Trans. Commun.* **COM-29**: (Feb. 1981).
43. URSI, Review of Radio Science 1981–1983.
44. URSI Commission E.

45. A. D. Watt, *VLF Radio Engineering*, Pergamon, London, 1967.

46. J. Werner, The HDSL environment (high bit rate digital subscriber line), *IEEE J. Select. Areas Commun.* **9**(6): 785–800 (Aug. 1991).

47. K. E. Wilson, Analysis of the Crichlow Graphical Model of Atmospheric Radio Noise at Very Low Frequencies M.S. Thesis, Air Force Institute of Technology, Nov. 1974.

48. J. W. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons, New York, 1965.

49. S. M. Zabin and H. V. Poor, Recursive algorithm for identification of impulsive noise channels, *IEEE Trans. Inform. Theory* **36**(3): (May 1990).

STATISTICAL MULTIPLEXING

KAVITHA CHANDRA
Center for Advanced
Computation and Telecommunications
University of Massachusetts at Lowell
Lowell, Massachusetts

1. INTRODUCTION

Twenty-first century communications will be dominated by intelligent high-speed information networks. Ubiquitous access to the network through wired and wireless technology, adaptable network systems, and broadband transmission capacities will enable networks to integrate and transmit media-rich services within specified standards of delivery. This vision has driven the standardization and evolution of broadband integrated services network (B-ISDN) technology since the early 1990s. The narrowband ISDN proposal for integrating voice, data, and video on the telephone line was the precursor to broadband services and asynchronous transport mode (ATM) technology. The ISDN and B-ISDN recommendations are put forth in a series of documents published by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T), which was formerly CCITT [1,2] and the ATM Forum study groups [3,4]. The concept of multiplexing integrated services traffic on a common channel for efficient utilization of the transmission link capacity is central in the design of ISDN and ATM networks. ATM in particular advocates an asynchronous allocation of time slots in a time-division multiplexed frame for servicing the variable-bit-rate (VBR) traffic generated from video and data services. ATM multiplexing relies on the transport of information using fixed-size cells of 53 bytes in length and the application of fast cell-switching architectures made possible by advances in digital technology. It utilizes the concepts of both circuit and packet switching by creating virtual circuits that carry VBR streams generated by multiplexing ATM cells from voice, video, and data sources.

The ATM architecture is designed to efficiently transport traffic sources that alternate between bursts of transmission activity and periods of no activity. It also supports traffic sources with continuously changing transmission rates. One measure of traffic burstiness is the ratio of peak to average rate of the source. A circuit-switched network would conservatively allocate to each source a capacity equal to its peak rate. In this case, full resource utilization takes place only when all of the sources transmit at their peak rates. This is typically a low-probability event when the sources are statistically independent of each other. A statistical multiplexer, however, allocates a capacity that lies between the average and peak rates and buffers the traffic during periods when demand exceeds channel capacity. The process of buffering the multiplexed stream smooths the relatively high variations in the traffic rate of individual sources. The multiplexed traffic is expected to have a smaller variance about the mean rate in the limit as the number of sources multiplexed increase to a large value. As a result, there is a diminishing magnitude in the probability of occurrence of source rates that are greater than the available capacity. This feature leads to the economies of scale paradigm of statistical multiplexing that is at the core of B-ISDN and ATM transmission technologies.

Statistical multiplexers have been integral components in packet switches and routers on data networks since the 1960s. They have gained increased prominence since 1990 with the availability of broadband transmission speeds exceeding 155 Mbps and ranging upto 10 Gbps in the core of the network. In conjunction with gigabit switching speeds, the new-generation of internetworks have the hardware infrastructure for delivering broadband services. However, since broadband traffic features are highly unpredictable, the control of service quality such as packet delay and loss probabilities must be managed by a suite of intelligent and adaptable protocols. The development of these techniques is the present focus of standards bodies, researchers, and developers in industry. New Internet protocols and services are currently being proposed by the Internet Engineering Task Force (IETF) to enable integrated access and controlled delivery of multimedia services on the existing Internet packet-switching architecture [5,6]. These include integrated (Intserv) and differentiated (Diffserv) services [7], multiprotocol label switching (MPLS) [8], and resource reservation protocols (RSVPs) [9]. It is expected that ATM and the Internet will coexist with ATM infrastructure deployed at the corporate, enterprise, and private network levels. The Internet will continue to serve connectivity on the wide-area network scale. The design and performance of these new protocols and services will depend on the traffic patterns of voice, video, and data sources and their influence on queues in statistical multiplexers. These problems have been the focus of numerous studies since 1990. This article is organized as follows. Section 2 describes stochastic traffic descriptors and models that have been applied to characterize voice, video, and data traffic. In Section 3 the methods applied for performance analysis of queues driven by the aforementioned models are discussed. Section 4 concludes with a discussion on the open problems in this area.

2. TRAFFIC DESCRIPTORS AND MODELS

The characterization of network traffic with parametric models is a basic requirement for engineering communications networks. Statistical multiplexers in particular

are modeled as queueing systems with finite buffer space, served by one or more transmission links of fixed or varying capacity. The service structure typically admits packets of multiple sources on a first-come first-serve (FCFS) basis. Priority-based service may also be implemented in ATM networks and more recent invocations of the Internet protocol. The statistical multiplexing gain (SMG) is an important performance metric that quantifies the multiplexing efficiency. The SMG may be calculated as the ratio of the number of VBR sources that can be multiplexed on a fixed capacity link under a specified delay or loss constraint and the number of sources that can be supported on the basis of peak rate allocation. To determine and maximize the SMG, admission control rules are formulated that can relate traffic characteristics to performance constraints and system parameters.

In this section, the analytic, computational, and empirical approaches for modeling traffic are discussed. A more detailed taxonomy of traffic models is presented by Frost and Melamed [10] and by Jagerman et al. [11]. The traffic is assumed to be composed of discrete units referred to as packets. The packets arriving at the multiplexer input are characterized using the sequence of random arrival times $T_1, T_2, \ldots, T_n$ measured from an origin assumed to be zero. The packets are associated with workloads $W_1, W_2, \ldots, W_n$ that may also be random variables. These workloads can represent variable Internet packet sizes fixed ATM cell sizes, or in case of batch arrivals, where more than one packet may arrive at a time instant, the workload represents the batch size. The packet interarrival times $\tau_n = T_n - T_{n-1}$ or the counting process $N(t)$, which represents the number of packets arriving in the interval $(0, t]$ are representative and equivalent descriptors of the traffic.

The most tractable traffic models result when interarrival times and workload sequences are independent random variables and independent of each other. A renewal process model is readily applicable as a traffic model in such a case. Telephone traffic on circuit-switched networks has been shown to be adequately modeled by independent negative exponential distributions for the interarrival times and call holding times. As shown by A. K. Erlang in his seminal study [12] of circuit-switched telephone traffic, the Poisson characteristics of teletraffic greatly simplify the analysis of queueing performance. Packet traffic measurement studies since 1970 have, however, shown that the arrival process of data, voice, and video applications rarely exhibit temporal independence. Traffic studies [13–15] conducted during the Arpanet days examined data traffic generated by user dialogs with distributed computer systems and showed that computer terminals transmitted information in bursts that occurred at random time intervals. Pawlita [16] presented a study of four different user applications in data networks and identified bursty traffic patterns, clustered dialog sequences and hyperexponential distributions for the user dialog times. Traffic measurement studies conducted on local-area networks [17–19] and wide-area networks [20] have found similar statistics in the packet interarrival times.

A measurement and modeling study of traffic on a token ring network by Jain and Routhier [17] showed that the packet arrivals occurred in clusters, for which they proposed a packet train model. The time between packet clusters was found to be a function of user access times, whereas the intracluster statistics were a function of the network hardware and software. More recent analyses of Internet traffic by Paxson and Floyd [21] and Caceres et al. [22] have shown that packet interarrival times generated by protocol-based applications such as file transfer, network news protocol, simple mail transfer, or remote logins are neither independent nor are they exponentially distributed.

Meier-Hellstern et al. [23], in their study of ISDN data traffic, have shown that the interarrival times for a user's terminal generated packet traffic can be modeled by superposing a gamma and power-law type probability density functions. The traffic generated in an Ethernet local area network of workstations has been shown by Gusella [24] to be nonstationary and characterized by a long-tailed interarrival time distribution. Leland et al. [25] analyzed aggregated Ethernet traffic on several timescales. A self-similar process was proposed as a model based on scale invariant features in the traffic. This model implies that the traffic variations are statistically similar over many, theoretically infinite ranges of timescales. As a result, one observes temporal dependence in the traffic structures over large time intervals. Erramilli et al. [26] propose a deterministic model based on chaotic maps for modeling these long-range dependence features. A compilation of references to work done on self-similar traffic modeling can be found in the study by Willinger et al. [27]. The aforementioned data traffic studies indicate that temporal dependence features found in measurements must be described accurately by traffic models. In this regard, static traffic descriptors such as the first- and second-order moments and marginal distributions of the traffic have been proposed. Dynamic models that capture some of the temporal features of the arrival process have also been proposed.

Traffic bursts are structures characterized by a successive occurrence of several short interarrival times followed by a relatively long interarrival time. This feature has been characterized using simple first-order descriptors such as the ratio of peak to average rate. In terms of the random interarrival times τ, the coefficient of variation c_τ captures the dispersion in the traffic through the ratio of the standard deviation and the expectation of the interarrival times

$$c_\tau = \frac{\sigma[\tau]}{E[\tau]} \tag{1}$$

Alternately, the index of dispersion $I_N(t)$ of the counting process $N(t)$ can be calculated for increasing time intervals of length t. This is a second-order characterization that captures the burstiness as a function of the variance of the process and is given by

$$I_N(t) = \frac{\text{Var}\,[N(t)]}{E[N(t)]} \tag{2}$$

An index of dispersion for intervals $I_\tau(n)$ may be similarly defined by replacing the numerator and denominator of Eq. (2) by the variance and expectation of the sum

of n successive interarrival times. The correlation in the workloads may also be characterized using the aforementioned indices of dispersion. The magnitude and rate of increase in these traffic descriptors can capture succintly, the degree of correlation in the arrival process. An increasing magnitude of the index of dispersion with the observation time indicates highly correlated streams that are in turn linked to large packet delays and packet losses. For example, the expected number in a single server queue driven by Poisson arrivals and a general service time distribution (M/G/1) is given by the Pollaczek–Khintchine mean value formula [28], which shows that the average queue size increases in direct proportion to the square coefficient of variation of the service times. For stationary arrival processes, the limiting values of the indices of dispersion as n and t tend to infinity are shown [24] to be related to the normalized autocorrelation coefficients $\rho_\tau(j)\ \ j = 0, 1, 2, \ldots$, as

$$I_N = I_\tau = {c_\tau}^2 \left[1 + 2 \sum_{j=1}^{\infty} \rho_\tau(j) \right] \tag{3}$$

Sriram and Whitt [29] and others [30] apply the index of dispersion of counts (IDC) and intervals for examining the burstiness effects of superposed packet voice traffic on queues. The IDC of a single packet voice source approached a limiting value of 18 in comparison to a value of unity for a Poisson process. It has been shown [29] that under superposition, the magnitudes of the IDC of the multiplexed process approached Poisson characteristics for short time intervals. As the time interval increased the positive autocorrelations of the individual sources interact, leading to increased values of the IDC parameter. The larger the number of sources superposed, the larger is the time interval at which the superposed process deviates from Poisson-like statistics. These concepts showed the importance of identifying a relevant timescale for the superposed traffic that allows the sizing of the buffers in a queue. Although the index of dispersion descriptors have proved useful for evaluating the burstiness property in a qualitative way, they have limited application for deriving explicit measures of queue performance.

A method for estimating an index of dispersion of the queue size using the peakedness functional is presented by Eckberg [31,32]. The "peakedness" of the queue represents the ratio of the variance and expectation of the number of busy servers in an infinite server system driven by a stationary traffic process. This approach incorporates second-order traffic descriptors. The traffic workloads represented by random service times S are modeled by the service time distribution $F(t)$, its complement $Q(t) = 1 - F(t) = \Pr[S > t]$, and the autocorrelation of $Q(t)$ denoted by $R_Q(t) = \int_0^\infty Q(x)Q(t+x)\,dx$. In addition, the arrival process may be characterized by a time-varying, possibly random, arrival rate $\lambda(t)$ and its covariance density $k(\tau)$. An arrival process characterized by a random arrival rate belongs to the category of doubly stochastic processes. Cox and Lewis [33] derive the covariance density of a doubly stochastic arrival process as $k(\tau) = \sigma_\lambda^2 \rho_\lambda(\tau)$, where σ_λ^2 and $\rho_\lambda(\tau)$ are the variance and normalized autocovariance functions of $\lambda(t)$, respectively. With the traffic specified by these functions, the expected value and variance of the number of busy servers $L(t)$ at a time t may be obtained as follows:

$$E[L(t)] = \int Q(t-\tau)\lambda(\tau)\,d\tau \tag{4}$$

$$\text{Var}\,[L(t)] = \int [Q(t-\tau)\{1 - Q(t-\tau)\}\lambda(\tau) + k(\tau)R_Q(\tau)]\,d\tau \tag{5}$$

The presence of correlations in the arrival rate for lags greater than zero cause the arrival process to be overdispersed relative to Poisson processes with constant arrival rate. The degree of dispersion is proportional to the magnitude of $k(\tau), \tau > 0$ and the decay rate of $Q(t)$. This increased traffic variability has an impact on the problem of resource allocation and engineering. The peakedness functional $Z[F] = \frac{E[L(t)]}{\text{Var}\ [L(t)]}$ provides a measure of the influence of traffic variance and correlations on queue performance. $Z[F]$ has a magnitude that is greater than one for processes with nonzero $k(s), s > 0$. The peakedness of the process influences the traffic engineering rules used for sizing system resources. The application of peakedness to estimate the blocking probability of finite server systems is presented by Fredericks [34]. Here a knowledge of $Z[F]$ is used in Hayward's approximation, an extension of Erlang's blocking formula to estimate the additional servers needed for $Z[F] > 1$. The utility of the peakedness characterization for analysis of delay systems has been discussed by Eckberg [32].

A more descriptive representation of the multiplexer queues is provided by the steady-state probability distribution of the buffer occupancy. If the random variable X represents the buffer occupancy in the steady state, the shape of the complementary queue distribution $G(x) = \Pr[X > x]$ provides information on the timescales at which traffic burstiness and correlations impact the queue. Livny et al. [35] have shown that the positive autocorrelations in traffic have significant impact in generating increased queue sizes and blocking probabilities relative to independent identically distributed processes. In this context it is useful to differentiate between two types of queue phenomena: queues arising from packet or cell level congestion and those arising from burst level congestion [36]. Packet level queues occur due to an instantaneous arrival of packets from different sources in the same time-slot resulting in a cumulative rate that is greater than the service capacity. It may be due to the chance occurrence of a set of interarrival times of different sources that cause individual packets to collide in time at the multiplexer input. This phenomenon can also occur for deterministic traffic, such as periodic sources, when the starting epochs are randomly displaced from one another [37,38]. Queues arising from packet-level congestion are typically of small to moderate size and can be accommodated using small buffers.

Larger queue sizes result when multiple traffic sources start transmitting in the burst state. Here, sustained transmission of a number of sources at the peak rate leads to a buildup in the queue size for time durations that

are functions of the burst state statistics. Since individual times of packet arrivals are not important in this case, burst-level queues have been analyzed using the fluid flow model [39]. In the fluid approximation, the discrete packet arrival process and the buffer occupancy variables are replaced by real-valued random processes. Although burst level congestion leads to lower probability events than does packet level congestion, the decay rate of these probabilities are functions of both the service rate and traffic source burst statistics. Figure 1 shows a depiction of a typical structure of the packet- and burst-level queue components in $G(x)$. In the design of a statistical multiplexer the size of buffer is typically set to absorb packet-level queues. Burst-level queues estimated from infinite buffer queue analysis can be used to approximate the losses that take place in a finite buffer system. Finite buffer systems typically require more complex analysis than infinite buffer systems.

The fluid approximation requires that the time variation and correlation of the arrival rate process be prescribed. Finite-state Markov chain models of traffic have been applied extensively in fluid buffer analysis. Discrete- and continuous-time Markov chains (DTMC, CTMC) with finite-state space [40] are among the simplest extensions to the renewal process model for incorporating temporal dependence. The traffic correlation structure exhibits geometric or exponential rate of decay for the discrete and continuous-time Markov chains, respectively. A K-state discrete time Markov chain $Y[n], n = 0, 1, 2, \ldots$ resides in one of K states $S_1, S_2, \ldots, S_K$ at any given time n. By the Markov property, the probability of transitioning to a particular state at time n is a function of the state of the process at $n-1$ only. These one-step transition probabilities are specified in a K-dimensional matrix P_Y as elements $p_{ij} = \Pr[Y[n] \in S_j | Y[n-1] \in S_i]$. The elements in each row of P_Y sum to unity. In a continuous time Markov chain, the transition rates are captured by an infinitesimal generating matrix Q_Y containing elements q_{ij} that represent the transition rate from state S_i to S_j for $i \neq j$. In this case, the sum of the rates in each row is equal to zero. The probability transition matrix and the generator matrix uniquely determine the rate of decay in the autocorrelations of the Markov chain. In the context of modeling traffic arrivals the transition matrix is supported by a K-state rate vector that describes the arrival rate when the traffic is in a particular state. This feature allows the variable rate features of network traffic to be represented. The rate vector in the simplest case is a set of constants that may represent the average traffic rate in each state. More general models based on a stochastic representation for the rate selection have also been considered. The Markov modulated Poisson process (MMPP) [30,41,42] is one example where the Markov process is characterized by a state-dependent Poisson process. These Markovian models of traffic can capture time variations in the arrival rate and associate these variations with a temporal correlation envelope that is determined by the magnitude of the transition probabilities. These models, however, cannot address nonexponential trends in the correlation function. To accommodate more general shapes of the correlation functions, Li and Hwang [43,44] propose the application of linear systems analysis using power spectral representation of traffic.

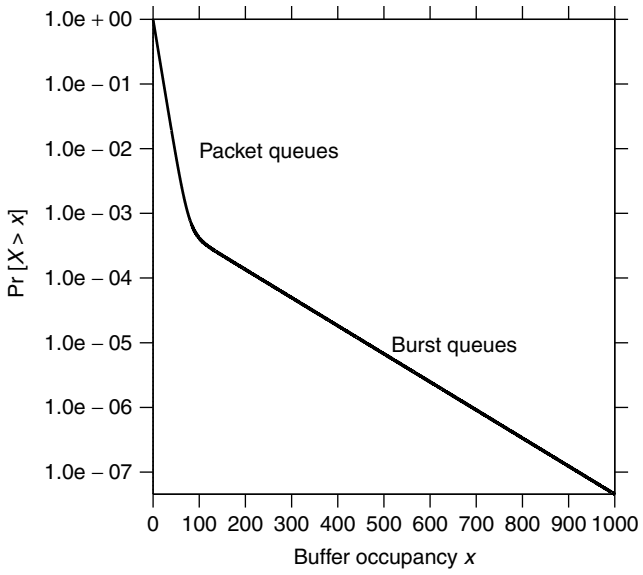


Figure 1. Packet- and burst-level components of queue size distribution.

3. PERFORMANCE OF STATISTICAL MULTIPLEXERS

Statistical multiplexing is designed to increase utilization of a resource that is subject to random usage patterns. In this work, the resource is considered to be a transmission link of finite capacity. Multiple sources access the channel on a first-come first-serve or priority basis and are allowed to queue in a buffer when the channel is busy. Statistical multiplexing gains come at the expense of a packet loss or delay probability that is considered tolerable for the applications being transported. The performance constraint is typically specified by the acceptable probability of loss for a given buffer size B. In some cases an infinite buffer queue is analyzed for tractability and the loss probability is approximated by the tail probability of queue lengths $P(X > B)$. For the limiting case of zero buffer size, the probability of loss may be calculated by determining the probability of the aggregate input rate exceeding the capacity.

The approaches to performance analysis in the literature may be classified by applications. The multiplexing of packet voice with data using two state Markov chains to model the ON and OFF states and the application of analytic and simulation-based performance analysis was the subject of numerous studies since the 1970s. With the availability of larger transmission capacities and standardization of encoding schemes for digital video, the transport and multiplexing of packet video became an active area of research in the early 1990s. Models for variable-bit-rate (VBR) video were found to be more complex and of higher dimensionality. The Markov representation for packet video typically required a large state space to capture the temporal variations and amplitude distributions. As a result, several approximation methodologies such as effective bandwidth formalisms and large

deviations analyses were proposed to relate the traffic characteristics, performance constraints, and statistical multiplexer parameters. These approximations have been particularly useful in formulating admission control decisions for multiservice networks.

In the following section a review of voice/data multiplexing schemes is provided first. This is followed by a presentation of video traffic models and the analysis of their multiplexing performance using fluid buffer approximations. This leads to a discussion of admission control algorithms with particular focus on the effective bandwidth approximations.

3.1. Voice and Data Multiplexers

In the early 1980s integrated services digital networks (ISDNs) [45] were envisoned to support multiplexed transport of voice-, data-, and image-based applications on a common transport infrastructure that included both telephone and data networks. The digital telephone network with a basic transmission rate of 64 kbps (kilobits per second) was considered to be the dominant transport network. Different local user interfaces were standardized to connect end systems such as telephones, data terminals, or local area networks to a common ISDN channel. At any given time, the traffic generated on this link could be a mix of data, voice, and associated signaling and control information. The performance requirements of voice and data traffic [46] govern the design of the multiplexing system. The buffer overflow probability is a chief concern for data transmission, whereas bounding transmission delay is critical for speech signals [47]. Initial studies on the performance of voice/data multiplexing systems assumed fixed duration time-division multiplexing frames in which time slots were distributed between voice and data packets. In this context, the multiplexing efficiency of voice and data has been analyzed using various approaches that involve moving frame boundaries between voice and data slots [48–50], separate queueing buffers for voice and data [51], encoder control for voice [52], application of circuit-switching concepts for both voice and data [53], and hybrid models of circuit-switched voice/packet-switched data [54]. Maglaris and Schwartz [55] describe a variable-frame multiplexer that admits long messages of variable length and single packets that arrive as a Poisson process. The Poisson model is assumed to impose a degree of traffic burstiness on the otherwise continuous rate process. The ability to adapt frame sizes in response to traffic variations showed improved performance in terms of bandwidth utilization and delays relative to fixed-frame movable-boundary schemes. The system requirements for multiplexing data during silence periods of speech is presented by Roberge and Adoul [56]. In this work the accurate discrimination of speech and data signals is proposed using statistical pattern classification algorithms based on zero-crossing statistics of the quadrature-amplitude-modulated speech signal. A speech-data transition detector is proposed for detecting switching points in time with accuracy.

With the evolution of fast packet switching devices, the more recent approaches to voice/data integration have examined the performance of asynchronous multiplexing on a single high-speed channel. Voice/data integration concepts were motivated in part, by the traffic characteristics of data applications such as Telnet, File Transfer Protocol (FTP), and Simple Mail Transfer Protocol (SMTP). These applications generate bursts of activity separated by random durations of inactive periods. Speech patterns in telephone conversations are also characterized by random durations of talk spurts that are followed by silence periods. Brady [57] presented experimental measurements of the average durations of ON and OFF periods and transition rates between these states from a study of telephone conversations. Typically the average speech activity is found to range from 28 to 40% of the total connection time and is a function of users, language, and other such factors. Average length of talk and silence spurts are in the 0.4–1.2-s and 0.6–1.8-s ranges.

A two-state Markov process [58] representing the ON and OFF states has been the canonical model for characterizing speech-based applications, although the silence durations are seldom exponentially distributed. The characteristic transition rates between ON and OFF states can be significantly different for voice and data sources. The alternating talk spurts and silence durations of speech applications exhibit relatively slow transition rates, allowing the data to be multiplexed in the OFF periods. Data sources exhibit faster transitions between active and inactive states. A problematic feature in packet voice traffic is its temporal correlation which is induced by speech encoders and voice activity detectors [29,30]. As a result of multiplexing with voice in a queue, the departing data flow takes on the characteristics of the superposed voicestream. This feature influences the performance of other multiplexers in the transmission path.

Heffes and Lucantoni [30] model the dependence features of multiplexed voice and data traffic using a two-state Markov modulated Poisson process (MMPP). Asynchronous voice data multiplexing of MMPP sources is examined by evaluating the delay distributions of a single-server queue with first-in first-out (FIFO) service and general service time distribution. The application of this model for evaluation of overload control algorithms is discussed. Sriram and Whitt [29] extract the dependence features of aggregate voice packet arrival process from a highly variable renewal process model of a single voice source. The aggregation of multiple independent voice sources is examined using the index of dispersion of intervals (IDI). The motivation behind this approach is that the limiting value of IDI as number of sources tend to infinity completely characterizes the effect of the arrival process on the congestion characteristics of a FIFO queue in heavy traffic. This work also shows that the positive dependence in the packet arrival process is a major cause of congestion in the multiplexer queue at heavy loads. Buffer sizes larger than a critical value as determined by the characteristic correlation time scale will allow a sequence of dependent interarrival times to build up the queue, causing congestion. Limiting the size of the buffer, at the cost of increased packet loss is proposed as an approach for controlling congestion. To control packet loss that occurs from dependence in arrival process, Sriram and Lucantoni [59] propose

dropping the least significant bits in the queue when the queue length reaches a given threshold. They show that under this approach the queue performance is comparable to that of a Poisson traffic source. These pioneering studies provided a comprehensive understanding on the efficiency of synchronous and asynchronous approaches for multiplexing voice and data traffic on a common channel. A quantitative characterization of the dependence features in traffic was shown to be one of the most important requirements for performance evaluation. In this regard, finite-state Markov processes were found to be amenable in both capturing some of the dependence features and allowing tractable analysis of the multiplexer queues. These studies also had limitations in that traffic measurements and measurement based models did not play a prominent role in analysis of multiplexers. However, with transition from ISDN to B-ISDN and the recognition that simple two-state Markov models are inadequate for broadband sources, more emphasis has been placed on measurement based analysis of video and data traffic. The developments in video models and multiplexers are discussed next.

3.2. Video Models and Multiplexers

Video communication services are important bandwidth consuming applications for B-ISDN. In an early study, Haskell [60] showed that multiplexing outputs of picture-phone video encoders into a common buffer could achieve significant multiplexing gains. Although current compression techniques for digital video can achieve video bit rates of acceptable quality in the range of 1–5 Mbps, when hundreds of such flows are to be transported, efficient multiplexing schemes are still required. Figure 2 depicts a comparison of the temporal variation of measured video frame rates for a low-activity videoconference encoded with H.261 standard and high-activity MPEG-2 encoded entertainment video. The signals represent the number of bits in each encoded video frame as a function of the frame index. The large dispersions about a mean rate are evident, as are the sudden transitions in frame rate amplitudes when encoding changes from predictive to refresh mode.

The transport of variable bit rate (VBR) video using statistical multiplexing has been examined in numerous studies. VBR video is preferred over traditional constant-bit-rate (CBR) video due to the improved image quality and shorter delays at the encoder. Statistical multiplexing invariably results in buffering delays and losses, which can significantly degrade video quality. To minimize the amount of delay and loss, the networking community has focused on the development of effective and implementable congestion control schemes, including connection admission control and usage parameter control. To minimize the impact of delay and loss, the video community has focused on developing good error concealment algorithms and designing efficient two-layer coding algorithms [61,62] for use in combination with the dual-priority transport provided by ATM networks. For example, while one-layer MPEG-2 produces generally unacceptable video quality with a cell loss ratio of 10^{-3}, losses at this rate with SNR scalability (one of the four standardized layered coding algorithms of MPEG-2) are generally invisible, even to experienced viewers [63].

Various application- and coding-specific models of one-layer VBR video have been proposed in the literature. Maglaris et al. [64] was among the first to analyze short (10-s) segments of low-activity videophone signals. A first

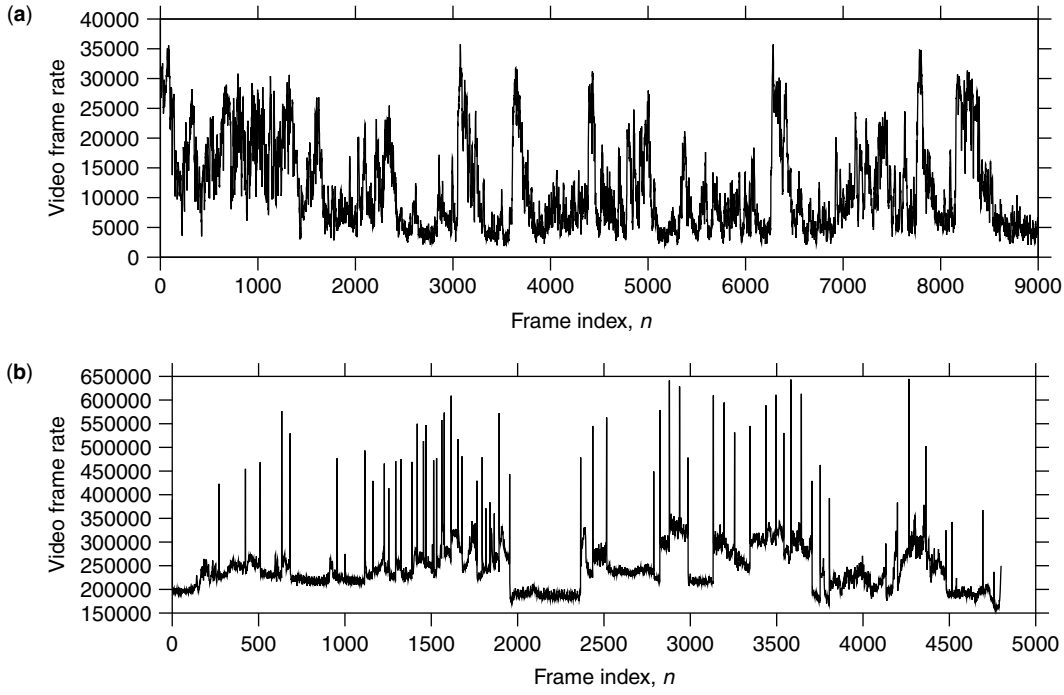


Figure 2. Sample paths of VBR video in videoconferencing and entertainment applications.

order autoregressive (AR) process was proposed for the number of bits in successive video frames of a single source. The multiplexed video was modeled by a birth–death Markov process. In this model, transitions are limited to neighboring states. The model parameters were selected to match the mean and short-term covariance structure in the measurements. The states of the Markov chain were derived by quantizing the aggregate source rate histogram into a fixed number of levels. A choice of 20 levels per source was found adequate for the low-activity videoconferencing source. Sen et al. [65] extended this model to accommodate moderate activity sources, using additional states to model low and high activity levels. The resulting source model is equivalent to that obtained by superposition of independent ON–OFF processes. Grunenfelder et al. [66] used a conditional replenishment encoder that exhibits strong correlations effects. The superposed video process was assumed to be wide sense stationary. The multiplexer was modeled using a general arrival process with independent arrivals and deterministic service times. A source model for full motion video was presented by Yegenoglu et al. [67] using an autoregressive process with time-varying coefficients. The selection of the coefficients was based on the state of a discrete-time Markov chain. The transition and rate matrices were constructed by matching the rate probability density function with that obtained from measurements. Moderate-activity videoconference data were also modeled by Heyman et al. [68]. The rate evolution was modeled by a Markov chain with identical transition probabilities in each row. These probabilities were modeled as a negative binomial distribution. The number of states in the model was of the order of the peak rate scaled by a factor of 10. This ranged from 400 to 500 states. The within state correlations were modeled by a discrete AR (DAR) process that resulted in a diagonally dominant matrix structure. This structure did not model single-source statistics very well since there were no selective transitions between states based on source characteristics. The DAR model failed to capture the short-term correlations in the traffic of a single source. Lucantoni et al. [69] proposed a Markov renewal process (MRP) model for VBR source traffic. Results of this model show that burstiness in video data can be captured more accurately than in the DAR model. Although the MRP was shown to perform better than the DAR model in capturing the burstiness, the MRP still did not match the cell loss probabilities for large buffer sizes. Heyman and Lakshman [70] studied high-activity video sources and concluded that their DAR model proposed for videoconference sources could not be applied as a general model to all sources. Skelly et al. [71] also used Markov chains to verify a histogram-based queueing model for multiplexing. They determined, on the basis of simulation, that a fixed number of eight states were sufficient to model the video source.

The video encoding system effects play a significant role in shaping the temporal and amplitude variation of compressed video. Traffic shaping algorithms at policing systems that enforce constraints on the output rate of the encoding system also play an important role in shaping digital video traffic [72–74]. In general, the coder that produces a bit stream conforming to constraints will not have the same statistical characteristics of an unconstrained coder. The idea that encoders could be constrained to generate traffic described by Markov chains was explored by Heeke [75] for designing better traffic policing and control algorithms. Pancha and Zarki [76] have examined the traffic characteristics resulting from various combinations of the quantization parameter, the inter-to-intraframe ratio and the priority breakpoint in MPEG one-layer and two-layer encoding, respectively. Data generated for each parameter set were modeled by a Markov chain by selecting the number of states based on the ratio of peak rate to the standard deviation of the frame rates. Frater et al. [77] verify the performance of a non-Markovian model for full motion video based on scene characterization by matching the cell loss probabilities at different buffer sizes. Krunz and Hughes [78] modeled MPEG, with distinct models for each different frame type. The selection of an adequate number of states in the Markov chain model of video such as to adequately model the spectral content is discussed by Chandra and Reibman [79]. Adequate spectral content in single sources is found necessary to understand the scaling aspects of Markov models under multiplexing.

Non-Markovian models exhibiting long-range dependence have been proposed by Garrett and Willinger [80] and others [77,81]. Ryu and Elwalid [82] show that long-term correlations do not significantly affect network performance over a reasonable range of cell losses, buffer sizes, and network operating parameters. Grossglauser and Bolot [83] also propose that correlation timescales to be considered in the traffic depend on the operating parameters and that full long-range dependence characterization of traffic is unnecessary. The impact of temporal correlation in the output rate of a VBR video source on the queue response has been examined [43], and it has been shown that macrolevel correlations can be modeled by Markov-chain-based models. Long-range dependence seen in VBR video has also been examined and the queueing results have been compared to those obtained using the DAR model [84]. It was concluded that for moderate buffer sizes, the short-range correlations obtained using Markov chain models are sufficient to estimate the buffer characteristics.

The aforementioned studies have determined that correlations on many timescales are an inherent feature in video sources and that Markov modulated source models are appropriate for capturing these dynamics. The ubiquitous use of multistate Markov models has led to work on their performance in queues. The application of finite-state Markov chain video traffic models for H.261 and MPEG coders in simulation studies [79,85] has shown that with an increase in the number of multiplexed video sources and corresponding increase in the channel capacity, the loss probability can be significantly reduced and reasonable multiplexing gains achieved. It was shown that typically 15–20 states are required to faithfully model the queue behavior of moderate to high-activity video sources. When using too few states, the tail probabilities of the rate histogram will not be captured, thereby yielding

an underestimate of the packet delay or loss probability. This situation has been observed by Hasslinger [86] in modeling VBR sources using semi-Markov models.

The performance analysis of queues driven by large-dimensional Markovian traffic sources may be approached using exact queueing analysis in discrete time and discrete state space. This approach becomes quickly intractable as the number of sources increase due to exponential increase in state dimension. In the limit of a large number of sources operating in the heavy-traffic regime, the discrete arrival and departure times may be replaced by a fluid approximation. This analysis technique is discussed in the next section.

3.3. Fluid Buffer Models

Fluid flow models assume that the packet arrival process at a multiplexer occurs continuously in time and may be characterized by continuous random fluctuations in the arrival rate [87]. This approach is applicable when the packet sizes are small relative to the link capacity. The computational model presented by Anick et al. [39] affords the estimation of the delay and loss distributions in multiplexers fed by Markov modulated fluid sources and served at constant rate. In this method, the buffer occupancy X is assumed to be a continuous valued random variable. The arrival process of each source is represented by a finite-state continuous-time Markov generator Q and associated diagonal rate matrix R. If K is the number of states required to represent a single source, and N is the number of sources (assumed identical) being multiplexed, the superposition can be modeled by the Markov generator Q_N and diagonal rate matrix R_N, which are computed as the N-fold Krönecker sums $Q \oplus Q \oplus \cdots \oplus Q$ and $R \oplus R \cdots \oplus R$, respectively. The Krönecker sum operation increases the dimension of the multiplexed source generator matrix to $M = K^N$.

The aggregated traffic stream enters a queue with finite or infinite waiting room. Packets in the buffer are serviced on a first-in first-out basis at a constant service rate. The cumulative probability distribution of the buffer occupancy x in steady state is specified by the row vector $\vec{p}$: $[p_0(x), p_1(x), \ldots, p_{M-1}(x)]$, where element $p_i(x) = \text{Prob}\,[X \le x; \textit{source in state } i]$. For a service rate of C packets per second, the probabilities satisfy the equation

$$\frac{\partial \vec{p}}{\partial x} D = \vec{p} Q_N \tag{6}$$

where the matrix $D = [R_N - IC]$ captures the drift from the service rate in each state. Here I is the identity matrix. The solution of Eq. (6) follows that of an eigenvalue problem and may be represented in terms of the tail probability distribution $G(x)$ as

$$G(x) = \Pr[X > x] = \sum_{i=0}^{M-1} a_i(x) e^{-z_i x} \tag{7}$$

where z_i, $i = 0, \ldots, M-1$ are the eigenvalues of the matrix $Q_N D^{-1}$. The coefficients $a_i(x)$ are functions of the eigenvalues and eigenvectors [88,89] of $Q_N D^{-1}$. For an infinite buffer, subject to consideration that the solution is bounded at x equal to infinity, the coefficients of the exponentially growing modes are set equal to zero. The amplitudes of the remaining modes are determined by applying the appropriate boundary conditions for overload and underload states. Underload states represented by states of the drift matrix with negative elements are subject to the condition $p_i(x = 0) = 0$. The coefficients for overload states are solved by equating $p_i(\infty)$ to the steady-state probability of the multiplexed source being in state i. For a buffer of finite size, all of the eigenvalues are retained and boundary condition at infinity replaced by the corresponding value at the buffer size.

The aforementioned approach requires the solution of an eigenvalue problem for a matrix whose upper bound on dimension scales as $O(K^N)$. To counter this dimensionality problem, reduced order traffic models have been used as approximations. For two-state ON–OFF Markov processes the superposition yields a generator of $O(N)$ states. Multi-state Markov sources have been approximated by the superposition of multiple two-state ON–OFF sources by matching first and second moments of the two processes [64,65,90]. The number of two-state sources selected for this model is often an arbitrarily choice. For moderate to high-activity video sources, this approximation can be shown to underestimate the packet delays.

Correlation effects afforded by the generator matrix play a dominant role in structuring the features of burst-level queueing delays. A traffic source represented by a finite-state Markov chain exhibits temporal autocorrelations that decay exponentially in time. The rate of decay is governed by the dominant eigenvalues of the generator matrix Q. It can be shown that the characteristic correlation time-scale of a single source is retained in the superposed traffic. The selection of an adequate single source model order K that captures all of the dominant modes is therefore an important consideration in building the traffic model. High-activity video sources often require K to be in the range of 15–20 states. Figure 3 shows the effect of choosing an inadequate number of states K for modeling the H.261 encoded videoconference source shown in Fig. 2a. As K is increased from 5 to 16, the asymptotic decay rate of the complementary delay distribution approaches that exhibited by the measurements.

For sources with large-dimensional K, even for moderate values of N, the estimation of buffer occupancy distributions for the multiplexed system becomes computationally intensive. A method for reducing the state-space dimension of multiplexed source generator is given by Thompson et al. [91]. The reduction process involved the quantization of the rates and the fundamental rate was chosen to yield the best match to the mean and variance of the rate. States having equal rates were aggregated, thereby reducing the number of states in the generator matrix. The resulting model allowed for scalable analysis as the number of sources was increased.

For large-dimensional systems, asymptotic approximations to the model given in Eq. (7) may be obtained for large buffer sizes and small delay or loss probabilities [92]. In this approximation

$$G(x) \sim \alpha e^{-\beta x} \quad x \to \infty \tag{8}$$

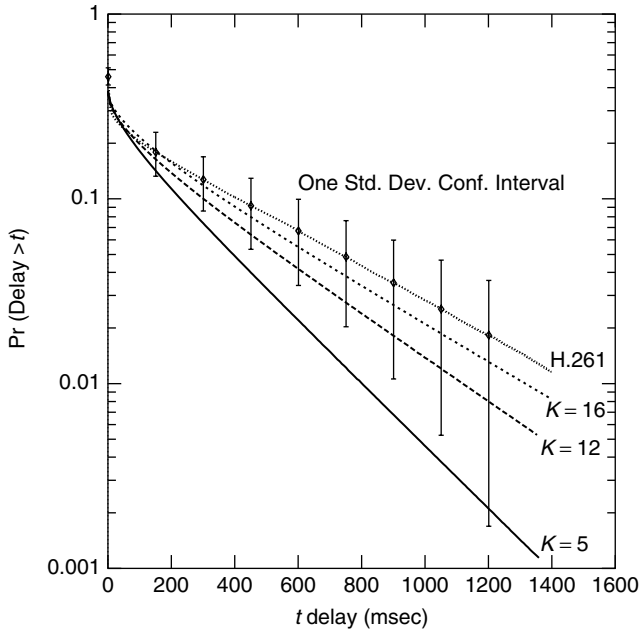


Figure 3. Influence of number of states K chosen to model video source.

where α is referred to as the *asymptotic constant* and β is the largest negative eigenvalue of the matrix $Q_N D^{-1}$. The asymptotic decay rate $-\beta$ is a function of the service rate C and may be determined with relative ease. The asymptotic constant, however, requires knowledge of all the eigenvectors and eigenvalues of the system. Approximate methods for estimating α for Markovian systems have been derived [93,94]. Although most of the asymptotic representations have considered Markovian sources, there have been some results for traffic modeled as stationary Gaussian processes [95] and fractional Brownian motion [96].

Very often, $\alpha = G(0)$ is assumed to be unity in the heavy-traffic regime. This allows a very usable descriptor of multiplexer performance that is referred to as the *effective bandwidth* of a source, which assumes the limiting form of the tail probabilities be structured as

$$G(x) \approx e^{-\beta x} \tag{9}$$

To achieve a specified value of β that satisfies given performance constraints, the required capacity may be shown [97] to be obtained as the maximal eigenvalue of the matrix $[R_N + \frac{Q_N}{\beta}]$. This capacity is referred to as the effective bandwidth (EB) of the multiplexed source. In the limit as β approaches zero and ∞, EB approaches the source average and peak rate, respectively. However, as noted by Choudhury et al. [98], the EB approximation can lead to conservative estimates for bursty traffic sources that undergo significant smoothing under multiplexing. It was shown that the asymptotic constant was itself asymptotically exponential in the number of multiplexed sources N. For traffic sources with indices of dispersion greater than Poisson, this parameter decreased exponentially in N, reflecting the multiplexing gain of the system. The application of these approximations in designing efficient multiplexing systems through admission control is discussed next.

3.4. Effective Bandwidths and Admission Control

Network traffic measurements have identified application and system dependent features that influence the traffic characteristics. Characterization in terms of the mean traffic rate is inadequate because of the large variability in its value over time. As traffic mixes and their performance requirements change, network mechanisms that adapt to these variations are of critical importance in broadband networks. Admission and usage control policies are two proposals in place in ATM networks and the next-generation Internet An admission control process determines if a source requesting connection can be admitted into the system without perturbing the performance of existing connections. This control algorithm should therefore make its decision taking into account a specific set of traffic parameters, available link capacity, and performance specifications. If a flow is admitted, the usage control policy monitors the flow characteristics to ensure that its bandwidth usage is within the admitted values.

To facilitate admission control, the concept of service classes has been introduced to categorize traffic with disparate traffic and service characteristics. Multiplexing among the same and across different service classes have been analyzed. The performance of five classes of admission control algorithms is reviewed by Knightly and Shroff [99]. These classes include scheduling based on average and peak rate information [100], effective bandwidth calculations [97,101,102], their refinements from large deviation principles [103], and maximum variance approaches based on estimating the upper tails of Gaussian process models of traffic [95]. A overview of the EB and its refinements is presented, since it appears to be the most generally applicable formalism.

It is assumed that K classes of traffic are to be admitted into a node served by a link capacity C packets per second. If N_i sources of type i exist, each characterized by an effective bandwidth E_i, then the simplest admission policy is given by the linear control law:

$$\sum_{i=1}^{K} N_i E_i \leq C \tag{10}$$

The effective bandwidths are derived taking into consideration the traffic characteristics and performance requirements of each class and available capacity C. Defining the traffic generated by type i source on a timescale t by a random variable $A_i[0, t]$, the effective bandwidth derived from large-deviation principles is given by [103]

$$E_i(s, t) = \frac{1}{st} \log E[e^{sA_i[0,t]}] \tag{11}$$

where the parameter s is related to the decay rate of $G(x)$ and captures the multiplexing efficiency of the system. It is calculated from the specified probability of loss or delay bounds [89]. The term on the right is the log moment generating function of the arrival process. The

workload can be described over a time t that represents the typical time taken for the buffer to overflow starting from an empty state. For a fixed value of t, the EB is an increasing function of s and lies between the mean and peak values of $A_i[0,t]$. This may be shown by a Taylor series approximation of E_i as $s \to 0$ and $s \to \infty$ respectively. Methods for deriving E_i for different traffic classes are discussed by Chang [104]. The aforementioned model assumes that all the multiplexed sources have the same quality of service requirements. If not, all sources achieve the performance of the most stringent source. Kulkarni et al. [105] consider an extension of this approach for addressing traffic of multiple classes. For the superposition, since the total workload is given by $A[0,t] = \sum_{i=1}^{K} A_i[0,t]$, the effective capacity C_e of the multiplexed system is

$$C_e = \sum_{i=1}^{K} E_i \tag{12}$$

The admission control algorithm simply compares C_e with available capacity C and if $C_e < C$ allows the new source to be admitted into the system.

4. CONCLUSIONS AND OPEN PROBLEMS

The current status on statistical multiplexing in broadband telecommunication networks has been presented. The multiplexing issues in the early 1980s were concerned with voice and data integration on 63-kbps telephone channels. Variations of synchronous time-division multiplexing using moving boundaries between voice and data slots, silence detection for insertion of data packets, and development of adaptive speech encoders were the primary concerns in designing efficient multiplexers. The transition to broadband era characterized by capacities exceeding 155 Mbps evolved with the design and standardization of asynchronous transfer mode networks. ATM networks were envisioned to integrate and optimize features of circuit- and packet-switched networks. The increase in switching speeds and network capacities in the 1990s and the invention of the World Wide Web concept, accelerated the development of many new applications and services that involved networked voice, video, and data. With increased accessibility of the Internet, many of the ATM related paradigms such as traffic characterization, admission control, and statistical multiplexing efficiency are now more relevant for the public Internet.

The important open issues at present are robust characterization of traffic and the derivation of traffic models that can be tractably analyzed in a queueing system. With the automation of end systems and application of complex encoders and detectors, the traffic patterns seen on networks today may not readily map to a pure stochastic model framework. On the contrary, traffic measurements indicate that deterministic patterns and nonlinearities in traffic amplitudes are new prevailing features in broadband networks. Large-dimensional stochastic models are seen to be required to capture these features. The computational complexity associated in analyzing the multiplexing problem for such models has led to some innovative approximation techniques. The characterization of a traffic source by an effective bandwidth is one such result that is derived by application of large-deviation principles. Derived in the asymptotic limit of large number of multiplexed sources, large buffer sizes and link capacities and small probabilities of delay and loss, effective bandwidths offer a conservative, but computationally feasible model for evaluating multiplexing efficiency in theory. The design of real-time algorithms for applying these concepts on a network and discovery of their effectiveness is expected to be the next step in the statistical multiplexing analysis.

BIOGRAPHY

Kavitha Chandra received her B.S. degree in electrical engineering in 1985 from Bangalore University, India, her M.S. and D.Eng. degrees in computer and electrical engineering from the University of Massachusetts at Lowell in 1987 and 1992, respectively. She joined AT&T Bell Laboratories in 1994 as a member of technical staff in the Teletraffic and Performance Analysis Department. From 1996 to 1998 she was a senior member of technical staff in the Network Design and Performance Analysis department of AT&T Laboratories. She is currently an associate professor in the Department of Electrical and Computer Engineering and principal faculty in the Center for Advanced Computation and Telecommunications at the University of Massachusetts at Lowell. Dr. Chandra received the Eta Kappa Nu Outstanding Electrical Engineer Award (honorable mention) in 1996 and the National Science Foundation Career Award in 1998. Her research interests are in the areas of network traffic and performance analysis, wireless networks, acoustic and electromagnetic wave propagation, adaptive estimation and control.

BIBLIOGRAPHY

1. CCITT Red Book, *Integrated Services Digital Network (ISDN), Series I Recommendation*, Vol. III, Fascicle III.5, 1985.
2. ITU-T, *Recommendation i.113. Vocabulary of Terms for Broadband Aspects of ISDN*, Vol. Rev. 1, Geneva, 1991.
3. ATM Forum, *User-Network Interface (UNI) Specification Version 3.1*, 1994.
4. ATM Forum, *Traffic Management Specification Version 4.0*, 1996.
5. D. D. Clark, S. Shenker, and L. Zhang, Supporting real-time applications in an integrated services packet network: Architecture and mechanism, *Proc. SIGCOMM 92*, 1992, pp. 14–26.
6. T. Chen, Evolution to the programmable Internet, *IEEE Commun. Mag.* **38**: 124–128 (2000).
7. G. Eichler, Implementing integrated and differentiated services for the Internet with ATM networks: A practical approach, *IEEE Commun. Mag.* **38**: 132–141 (2000).

8. D. Awduche, MPLS and traffic engineering in IP networks, *IEEE Commun. Mag.* **37**: 42–47 (1999).
9. L. Zhang et al., RSVP: A new resource reservation protocol, *IEEE Network* **7**(5): 8–18 (1993).
10. V. S. Frost and B. Melamed, Traffic modeling for telecommunications networks, *IEEE Commun. Mag.* **32**(4): 70–81 (1994).
11. D. Jagerman, B. Melamed, and W. Willinger, Stochastic modeling of traffic processes, in J. Dshalalow, ed., *Frontiers in Queuing: Models, Methods and Problems*, CRC Press, 1996.
12. A. K. Erlang, The theory of probabilities and telephone conversations, *Nyt Tidsskrift Matematik B* **20**: 33 (1909). (English translation in E. Brockmeyer, H. L. Halstrom and A. Jensen (1948), The life and works of A. K. Erlang, The Copenhagen Telephone Company, Copenhagen.
13. E. G. Coffman and R. C. Wood, Interarrival statistics for time sharing systems, *Commun. ACM* **9**: 5000–5003 (1966).
14. P. E. Jackson and C. D. Stubbs, A study of multi-access computer communications, *AFIPS Conf. Proc.* **34**: 491–504 (1969).
15. E. Fuchs and P. E. Jackson, Estimates of distributions of random variables for certain computer communications traffic models, *Commun. ACM* **13**: 752–757 (1970).
16. P. F. Pawlita, Traffic measurements in data networks, recent measurement results, and some implications, *IEEE Trans. Commun.* **29**(4): 525–535 (1981).
17. R. Jain and S. A. Routhier, Packet trains- measurements and a new model for computer network traffic, *IEEE J. Select. Areas Commun.* **4**(6): 986–995 (1986).
18. J. F. Schoch and J. A. Hupp, Performance of the Ethernet local network, *Commun. ACM* **23**(12): 711–721 (1980).
19. D. N. Murray and P. H. Enslow, An experimental study of the performance of a local area network, *IEEE Commun. Mag.* **22**: 48–53 (1984).
20. F. A. Tobagi, Modeling and measurement techniques in packet communication networks, *Proc. IEEE* **66**: 1423–1447 (1978).
21. V. Paxson and S. Floyd, Wide-area traffic: The failure of Poisson modeling, *IEEE/ACM Trans. Network.* **3**(3): 226–244 (1996).
22. R. Caceres, P. Danzig, S. Jamin, and D. Mitzel, Characteristics of wide-area Tcp/Ip conversations, *Proc. ACM/SIGCOMM*, 1991, pp. 101–112.
23. K. S. Meier-Hellstern, P. E. Wirth, Y. Yan, and D. A. Hoeflin, Traffic models for ISDN data users: Office automation application, in A. Jensen and V. B. Iversen, eds., *Teletraffic and Data Traffic, a Period of Change*, Elsevier Science Publishers, 1991, pp. 167–172.
24. R. Gusella, Characterizing the variability of arrival processes with indexes of dispersion, *IEEE J. Select. Areas Commun.* **9**(2): 203–211 (1991).
25. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Network.* **2**(1): 1–15 (1994).
26. A. Erramilli, R. P. Singh, and P. Pruthi, Chaotic maps as models of packet traffic, *Proc. 14th Int. Teletraffic Congress*, 1994, Vol. 1, pp. 329–338.
27. W. Willinger, M. S. Taqqu, and A. Erramilli, A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks, in F. P. Kelly, S. Zachary, and I. Ziedins, eds., *Stochastic Networks: Theory and Applications*, Oxford Univ. Press, 1996, pp. 339–366.
28. R. B. Cooper, *Introduction to Queueing Theory*, 3rd ed., CEEPress Books, 1990.
29. K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, *IEEE J. Select. Areas Commun.* **4**(6): 833–846 (1986).
30. H. Heffes and D. M. Lucantoni, A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE J. Select. Areas Commun.* **4**: 856–868 (1986).
31. A. E. Eckberg, Jr., Generalized peakedness of teletraffic processes, *10th Int. Teletraffic Congress*, 1983.
32. A. E. Eckberg, Jr., Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness, in *11th Int. Teletraffic Congress*, 1985.
33. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Chapman & Hall, 1966.
34. A. A. Fredericks, Congestion in blocking systems — a simple approximation technique, *Bell Syst. Tech. J.* **59**: 805–827 (1980).
35. M. Livny, B. Melamed, and A. K. Tsiolis, The impact of autocorrelation on queueing systems, *Manage. Sci.* **39**(3): 322–339 (1993).
36. W. Roberts, ed., *Performance Evaluation and Design of Multiservice Networks*, COST 224 Final Report, Commission of the European Communities, 1992.
37. I. Norros, J. W. Roberts, A. Simonian, and J. T. Virtamo, The superposition of variable bit rate sources in an ATM multiplexer, *IEEE J. Select. Areas Commun.* **9**(3): 378–387 (1991).
38. J. W. Roberts and J. T. Virtamo, The superposition of periodic cell arrival streams in an ATM multiplexer, *IEEE Trans. Commun.* **39**: 298–303 (1991).
39. D. Anick, D. Mitra, and M. M. Sondhi, Stochastic theory of a data-handling system with multiple sources, *Bell Syst. Tech. J.* **8**: 1871–1894 (1982).
40. E. Cinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
41. H. Heffes, A class of data traffic processes-covariance function characterization and related queueing results, *Bell Syst. Tech. J.* **59**: 897–929 (1980).
42. W. Fischer and K. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook, *Perform. Eval.* **18**: 149–171 (1992).
43. S. Q. Li and C. L. Hwang, Queue response to input correlation functions: Discrete spectral analysis, *IEEE/ACM Trans. Network.* **1**(5): 522–533 (1993).
44. S. Q. Li and C. L. Hwang, Queue response to input correlation functions: Continuous spectral analysis, *IEEE/ACM Trans. Network.* **1**(6): 678–692 (1993).
45. M. Decina, W. S. Gifford, R. Potter, and A. A. Robrock (guest eds.), Special issue on Integrated Services Digital Network: Recommendations and Field Trials — I, *IEEE J. Select. Areas Commun.* **4**(3): (1986).
46. J. G. Gruber and N. H. Le, Performance requirements for integrated voice/data networks, *IEEE J. Select. Areas Commun.* **1**(6): 981–1005 (1983).

47. J. G. Gruber, Delay related issues in integrated voice and data networks, *IEEE Trans. Commun.* **29**(6): 786–800 (1980).

48. N. Janakiraman, B. Pagurek, and J. E. Neilson, Performance analysis of an integrated switch with fixed or variable frame rate and movable voice/data boundary, *IEEE Trans. Commun.* **32**: 34–39 (1984).

49. A. G. Konheim and R. L. Pickholtz, Analysis of integrated voice/data multiplexing, *IEEE Trans. Commun.* **32**(2): 140–147 (1984).

50. K. Sriram, P. K. Varshney, and J. G. Shantikumar, Discrete-time analysis of integrated voice-data multiplexers with and without speech activity detectors, *IEEE J. Select. Areas Commun.* **1**(6): 1124–1132 (1983).

51. H. H. Lee and C. K. Un, Performance analysis of statistical voice/data multiplexing systems with voice storage, *IEEE Trans. Commun.* **33**(8): 809–819 (1985).

52. T. Bially, B. Gold, and S. Seneff, A technique for adaptive voice flow control in integrated packet networks, *IEEE Trans. Commun.* **28**: 325–333 (1980).

53. E. A. Harrington, Voice/data integration using circuit switched networks, *IEEE Trans. Commun.* **28**: 781–793 (1980).

54. C. J. Weinstein, M. K. Malpass, and M. J. Fisher, Data traffic performance of an integrated circuit and packet-switched multiplex structure, *IEEE Trans. Commun.* **28**: 873–877 (1980).

55. B. Maglaris and M. Schwartz, Performance evaluation of a variable frame multiplexer for integrated switched networks, *IEEE Trans. Commun.* **29**(6): 801–807 (1981).

56. C. Roberge and J. Adoul, Fast on-line speech/voiceband-data discrimination for statistical multiplexing of data with telephone conversations, *IEEE Trans. Commun.* **34**(8): 744–751 (1986).

57. P. T. Brady, A statistical analysis of on-off patterns in 16 conversations, *Bell Syst. Tech. J.* **47**: 73–91 (1968).

58. P. T. Brady, A model for generating on-off speech patterns in two-way conversations, *Bell Syst. Tech. J.* **48**: 2445–2472 (1969).

59. K. Sriram and D. M. Lucantoni, Traffic smoothing effects of bit dropping in a packet voice multiplexer, *IEEE Trans. Commun.* **37**(7): 703–712 (1989).

60. B. G. Haskell, Buffer and channel sharing by several interframe picturephone coders, *Bell Syst. Tech. J.* **51**(1): 261–289 (1972).

61. M. Ghanbari, Two-layer coding of video signals for VBR networks, *IEEE J. Select. Areas Commun.* **7**: 771–781 (1989).

62. S. Tubaro, A two layers video coding scheme for ATM networks, *Signal Process. Image Commun.* **3**: 129–141 (1991).

63. R. Aravind, M. R. Civanlar, and A. R. Reibman, Packet loss resilience of mpeg-2 scalable coding algorithms, *IEEE Trans. Circuits Syst. Video Technol.* **6**: 426–435 (1996).

64. B. Maglaris et al., Performance models of statistical multiplexing in packet video communications, *IEEE Trans. Commun.* **36**(7): 834–844 (1988).

65. P. Sen, B. Maglaris, N. Rikli, and D. Anastassiou, Models for packet switching of variable bit-rate video sources, *IEEE J. Select. Areas Commun.* **7**(5): 865–869 (1989).

66. R. Grunenfelder, J. P. Cosmos, S. Manthorpe, and A. Odinma-Okafor, Characterization of video codecs as autoregressive moving average processes and related queueing system performance, *IEEE J. Select. Areas Commun.* **9**: 284–293 (1991).

67. F. Yegenoglu, B. Jabbari, and Y. Zhang, Motion classified autoregressive modeling of variable bit rate video, *IEEE Trans. Circuits Syst. Video Technol.* **3**: 42–53 (1993).

68. D. Heyman, A. Tabatbai, and T. V. Lakshman, Statistical analysis and simulation study of video teletraffic in atm networks, *IEEE Trans. Circuits Syst. Video Technol.* **2**: 49–59 (1992).

69. D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, Methods for performance evaluation of VBR video traffic models, *IEEE/ACM Trans. Network.* **2**: 176–180 (1994).

70. D. P. Heyman and T. V. Lakshman, Source models for VBR broadcast-video traffic, *IEEE/ACM Trans. Network.* **4**: 40–48 (1996).

71. P. Skelly, M. Schwartz, and S. Dixit, A histogram based model for video traffic behavior in an ATM multiplexer, *IEEE/ACM Trans. Network.* **1**(4): 447–459 (1993).

72. P. Pancha and M. El Zarki, Prioritized transmission of VBR MPEG video, *Proc. GLOBECOM'92*, 1992, pp. 1135–1139.

73. M. R. Ismail, I. E. Lambadaris, M. Devetsikiotis, and A. R. Raye, Modelling prioritized MPEG video using tes and a frame spreading strategy for transmission in ATM networks, *Proc. INFOCOM'9*, 1995, Vol. 5, pp. 762–769.

74. A. R. Reibman and A. W. Berger, Traffic descriptors for VBR video teleconferencing, *IEEE/ACM Trans. Network.* **3**: 329–339 (1995).

75. H. Heeke, A traffic control algorithm for ATM networks, *IEEE Trans. Circuits Syst. Video Technol.* **3**(3): 183–189 (1993).

76. P. Pancha and M. El. Zarki, Bandwidth allocation schemes for variable bit rate MPEG sources in ATM networks, *IEEE Trans. Circuits Syst. Video Technol.* **3**(3): 192–198 (1993).

77. M. R. Frater, J. F. Arnold, and P. Tan, A new statistical model for traffic generated by VBR coders for television on the broadband isdn, *IEEE Trans. Circuits Syst. Video Technol.* **4**(6): 521–526 (1994).

78. M. Krunz and H. Hughes, A traffic model for MPEG-coded VBR streams, *Perform. Eval. Rev.* (*Proc. ACM SIGMETRICS'95*) **23**: 47–55 (1995).

79. K. Chandra and A. R. Reibman, Modeling one- and two-layer variable bit rate video, *IEEE/ACM Trans. Networks* **7**(3): 398–413 (1999).

80. M. W. Garrett and W. Willinger, Analysis, modeling and generation of self-similar VBR video traffic, *Proc. ACM SIGCOMM'94*, 1994, pp. 269–280.

81. J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, Long range dependence in variable bit-rate video traffic, *IEEE Trans. Commun.* **43**: 1566–1579 (1995).

82. B. K. Ryu and A. Elwalid, The importance of long-range dependence of VBR video traffic in atm traffic engineering: Myths and realities, *Proc. ACM SIGCOMM'96*, 1996, pp. 3–14.

83. M. Grossglauser and J. D. Bolot, On the relevance of long-range dependence in network traffic, *Proc. ACM SIGCOMM'96*, 1996, pp. 15–24.

84. D. Heyman and T. V. Lakshman, What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Trans. Networking* **4**: 301–317 (1996).

85. K. Chandra and A. R. Reibman, Modeling two-layer SNR scalable MPEG-2 video traffic, *Proc. 7th Int. Workshop Packet Video*, 1996, pp. 7–12.

86. G. Hasslinger, Semi-markovian modelling and performance analysis of variable rate traffic in ATM networks, *Telecommun. Syst.* **7**: 281–298 (1997).

87. L. Kleinrock, *Queueing Systems*, Vol. 2, Wiley, New York, 1976.

88. J. Walrand and P. Varaiya, *High-Performance Communication Networks*, Morgan Kaufmann, 1996.

89. M. Schwartz, *Broadband Integrated Networks*, Prentice-Hall, 1996.

90. J. W. Mark and S.-Q. Li, Traffic characterization for integrated services networks, *IEEE Trans. Commun.* **38**: 1231–1242 (1990).

91. C. Thompson, K. Chandra, S. Mulpur, and J. Davis, Packet delay in multiplexed video streams, *Telecommun. Syst.* **16**: 335–345 (2001).

92. J. Abate, G. L. Choudhury, and W. Whitt, Asymptotics for steady-state tail probabilities in structured Markov queueing models, *Stochastic Models* **10**: 99–143 (1994).

93. R. G. Addie and M. Zukerman, An approximation for performance evaluation of stationary single server queues, *IEEE Trans. Commun.* **42**: 3150–3160 (1994).

94. A. Elwalid et al., Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing, *IEEE J. Select. Areas Commun.* **13**: 1004–1016 (1995).

95. J. Choe and N. B. Shroff, A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks, *IEEE/ACM Trans. Network.* **6**(5): 659–671 (1998).

96. I. Norros, On the use of Fractal Brownian motion in the theory of connectionless networks, *IEEE J. Select. Areas Commun.* **13**: 953–962 (1995).

97. A. I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, *IEEE/ACM Trans. Network.* **1**(3): 329–343 (1993).

98. G. Choudhury, D. M. Lucantoni, and W. Whitt, Squeezing the most of ATM, *IEEE Trans. Commun.* **44**(2): 203–217 (1996).

99. E. W. Knightly and N. B. Schroff, Admission control for statistical qos: Theory and practice, *IEEE Network* **13**(2): 20–29 (1999).

100. D. Ferrari and D. Verma, A scheme for real-time channel establishment in wide-area networks, *IEEE J. Select. Areas Commun.* **8**: 368–379 (1990).

101. G. Kesidis, J. Walrand, and C. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, *IEEE/ACM Trans. Network.* **1**(4): 424–428 (1993).

102. C. S. Chang and J. A. Thomas, Effective bandwidth in high-speed digital networks, *IEEE J. Select. Areas Commun.* **13**: 1019–1114 (1995).

103. F. Kelly, *Stochastic Networks: Theory and Applications*, Oxford Univ. Press, 1996.

104. C. S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, *IEEE Trans. Automatic Control* **39**: 913–931 (1994).

105. V. G. Kulkarni, L. Gun, and P. F. Chimento, Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer, *IEEE J. Select. Areas Commun.* **6**(13): 1039–1047 (1995).

STREAMING VIDEO

ROBERT A. COHEN
Rensselaer Polytechnic Institute
Troy, New York

HAYDER RADHA
Michigan State University
East Lansing, Michigan

1. INTRODUCTION

In the early 1990s, as the Internet and the World Wide Web were becoming ubiquitous in the office and home, users would download compressed digital video files onto their computers. Once downloaded, these clips would be viewed by opening the video files with a compatible player program on the computer. These video files often required several minutes or hours to download, especially through slow modem connections over telephone lines. Depending on which compression scheme was used, an incomplete video file may not have been viewable at all. If a connection were not reliable, the download process would have to be repeated, which forced the user to wait even longer to view the clip. This dilemma was solved through the use of videostreaming.

Videostreaming is the transmission, at a regulated rate, of digital video over a network from a server to a client (player) in such a way that the client can display the video while the video is still being transmitted. In other words, the client can start playing the video without having to wait for the entire clip to be downloaded. Using an assortment of compression techniques, packetization, and transport protocols, video can be manipulated to be streamed over a wide variety of networks. Publicly available methods for streaming video over the Internet first came about in 1994 using the Multicast Backbone (MBone) [1], in which many users could simultaneously receive multimedia in real time. Several commercial videostreaming solutions first became available between 1995 and 1997. Between 1997 and today, these streaming systems have evolved to address a variety of streaming needs, ranging from small personal streaming applications all the way to large distributed streaming systems designed for thousands of viewers. Several articles describing the many technologies and current trends in videostreaming can be found in Civanlar et al. [2]. Information regarding the histories and uses of specific commercial streaming systems can also be found [3–5].

The purpose of this article is to give an overview of how videostreaming works, the various technical aspects of

videostreaming, video and network-related issues involved in streaming, technical solutions to these issues, and future directions of this field.

2. VIDEOSTREAMING SYSTEM DESCRIPTION AND OPERATION

The two main components of a videostreaming system are the streaming server and the client. The *server*, which can range in size from a small PC-like system with a camera to a large collection of networked computers, streams video into the network. The *client* is the device or computer on which the video is displayed. In bidirectional systems such as those used with videoconferencing, the server and client are combined into one device. The focus of this chapter will be on one-way videostreaming. For more information on videoconferencing, see Schaporst [6]. For an in-depth look at streaming video over the Internet, see [7].

2.1. Video Server Architecture

The components of a typical videostreaming server are shown in Fig. 1. Digitized video typically has a very high bit rate. For example, uncompressed 320×240-pixel RGB (red-green-blue) video at 30 frames per second (FPS) requires a transmission bandwidth of over 55 Mbps, which is well over the bandwidth available on most of the networks in use today. A video encoder therefore is used to compress the program into a lower-rate bitstream. Details on video coders that are used for streaming are given in Section 3. Once the video is compressed, it can be streamed live over the network, or it can be stored for streaming later on demand.

The session controller chooses a program to be streamed. It typically receives commands from the client, which is shown in Fig. 2. Issues related to session control and management are discussed in Section 5. Once a program is selected, the rate of the compressed video data will be chosen on the basis of data from the session controller and from the network and client limitations. On the basis of the performance of the network, the congestion controller can also control the rate at which a video program is being streamed. Since video is usually streamed over packet-based networks, it must be packetized in a way that is appropriate given the type of network and video encoding method. Section 4 gives more information on packetization schemes and network-related issues. If the video is going to be streamed over a lossy network, error control data can be added to the video packets to reduce the effects of packet loss or delay. These application-layer packets are then fed to the transport layer, which uses protocols such as UDP or ATM to send data over the network.

2.2. Video Client Architecture

A client system is shown in Fig. 2. The user typically first chooses a program to be viewed. The program list is usually sent over the network via a separate application, such as a Web browser. Once the program is chosen, the client's session controller tells the server's controller to start streaming the video. The client receives and demultiplexes the packets to reconstruct the encoded videostream. If errors are detected, the client will, if possible, correct the errors before sending the reconstructed stream to the video decoder. As described later, error detection information can be fed back to the server for congestion and rate control. The video decoder then decompresses the stream to produce video that can be displayed. In some cases, the decoder can try to conceal uncorrectable errors so the user does not see too many distracting artifacts in the displayed video.

3. VIDEO CODING FOR STREAMING

3.1. Basics of Video Coding

As discussed above, video coding is needed to reduce the amount of information transmitted over a streaming session. The basic objective of any video coding method is to reduce or virtually eliminate the amount of redundant data contained within a series of pictures. Usually, each picture contains a large number of pixels that are very similar. This is particularly true for pixels located within a small neighborhood of the picture. Therefore, the total number of pixels within a picture can be represented by a smaller number of bytes. Moreover, consecutive pictures within a video sequence are very similar. Consequently,

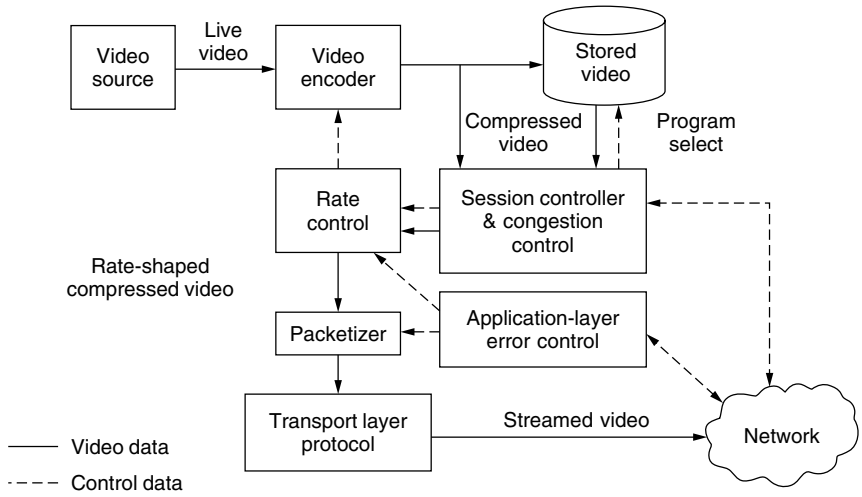


Figure 1. Videostreaming server.

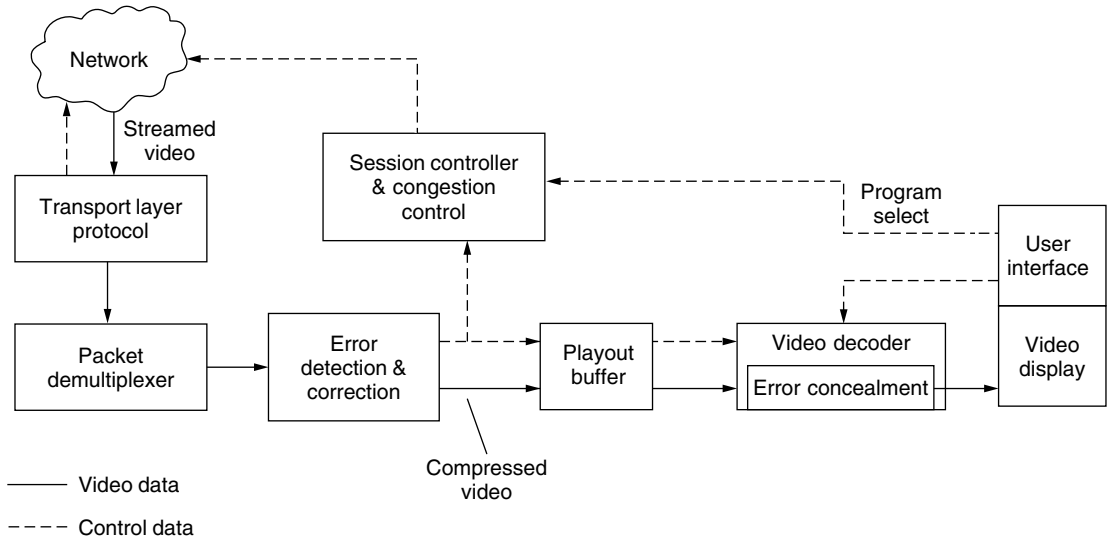


Figure 2. Videostreaming client.

there are two types of redundancies within a video sequence that a video encoder exploits to reduce the amount of transmitted data: (1) redundancies within each picture and (2) redundancies among consecutive pictures. Redundancies among consecutive pictures are known as *temporal redundancies* or *interframe* (or *interpicture*) *redundancies*. When coding a new picture, temporal redundancies are reduced through prediction from one or more previously coded pictures. The most popular picture prediction approach is based on motion estimation/motion compensation (ME/MC) using block matching (BM) among adjacent pictures [8].

International video coding standards such as those from MPEG [9–11], H.261 [12], and H.263 [13] employ *intracoded pictures* (I frames), which do not depend on prediction from other pictures. Any video sequence requires a minimum of one I frame, which is usually the first picture of the sequence. The abovementioned standards also use *prediction frames* (P frames), which depend on a previously coded I or P frame. This results in the picture coding structure shown in Fig. 3. MPEG-2 [10], MPEG-4 [11], and other more recent compression methods employ *bidirectional prediction frames* (B frames), which use prediction from a previously coded I or P frame *and* a future P frame as shown in Fig. 3.

The pixels of an I frame or the *residual pixels* of P and B frames are usually coded using transform-domain methods. The *discrete cosine transform* (DCT) is the most popular transform coding method used for video compression [14]. MPEG-1 [9], MPEG-2, H.261, and H.263 are all based on a DCT coding method. Another popular transform that has received a great deal of attention is the wavelet transform [15]. For example, JPEG-2000 [16] is based on a wavelet transform coding method.

3.2. Scalable Video Coding for Streaming Applications

Scalable video coding is a desirable tool for many multimedia applications and services. Video scalability, for example, can be used in systems employing decoders with a wide range of processing power. In this case, processors with low computational power decode only a subset of the scalable videostream. Another use of scalable video is in environments with a variable and unpredictable transmission bandwidth (e.g., the Internet or wireless networks). In this case, receivers with low access bandwidth receive, and consequently decode, only a subset of the scalable videostream, where the amount of that subset is proportional to the available bandwidth. Without scalability, a videostream could be rendered useless if the network bandwidth drops below the coding rate.

Several video scalability approaches have been adopted by leading video compression standards such as MPEG-2, MPEG-4, and H.263. Temporal, spatial, and quality [signal-to-noise ratio (SNR)] scalability types have been defined in these standards. Temporal scalability applies to the frame rate (in frames per second) of the video. With spatial scalability, the size or resolution of each video frame can vary. In SNR-scalable systems, the quality of the video can be increased or decreased. All of these standardized types of scalable video consist of a *base layer*

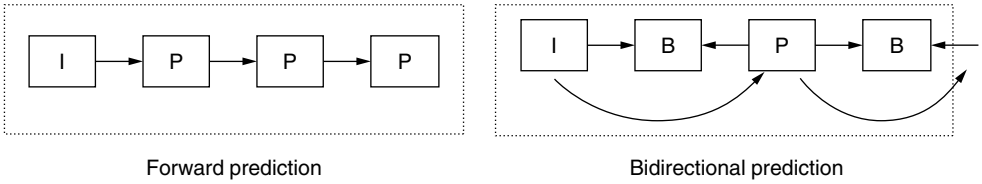


Figure 3. Examples of prediction structures used for video coding.

(BL) and one or multiple *enhancement layers* (ELs). The BL part of the scalable videostream represents, in general, the minimum amount of data needed for decoding a viewable video sequence from the stream. The EL part of the stream represents additional information, and therefore enhances the video signal representation when decoded by the receiver.

For each type of video scalability, a certain *scalability structure* is used. The scalability structure defines the relationship among the pictures of the BL and the pictures of the enhancement layer. Figure 4 illustrates examples of video scalability structures. MPEG-4 also supports object-based scalability structures for arbitrarily shaped video objects.

Another type of scalability, which has been primarily used for coding still images, is *fine-granular scalability* [17]. Images coded with this type of scalability can be decoded progressively. In other words, the decoder can start decoding and displaying the image after receiving a very small amount of data. As more data are received, the quality of the decoded image is progressively enhanced until the complete information is received, decoded, and displayed. Among leading international standards, progressive image coding is one of the modes supported in JPEG-2000 and the still-image coding tool in MPEG-4 video.

When compared with nonscalable methods, a disadvantage of scalable video compression is coding efficiency. In order to increase coding efficiency, video scalability methods normally rely on relatively complex structures (such as the spatial and temporal scalability examples shown in Fig. 4). By using information from as many pictures as possible from both the BL and EL, coding efficiency can be improved when compressing an enhancement-layer picture. However, using prediction among pictures within the enhancement layer either eliminates or significantly reduces the fine-granular scalability feature, which is desirable for environments with a wide range of available bandwidth (e.g., the Internet).

3.2.1. MPEG-4 Fine-Granular Scalability (FGS) Video Coding. In order to strike a balance between coding efficiency and fine-granularity requirements, a more recent activity in MPEG-4 adopted a hybrid scalability structure characterized by a DCT motion-compensated base layer and a fine-granular scalable enhancement layer. This scalability structure is illustrated in Fig. 5.

The base layer carries a minimally acceptable quality of video to be reliably delivered using a, packet-loss recovery method such as retransmission. The enhancement layer improves the base layer video by fully utilizing the bandwidth available to individual clients. By employing a motion-compensated base layer, coding efficiency from temporal redundancy exploitation is partially retained. The base and a single-enhancement layer streams can be either stored for later transmission, or can be directly streamed by the server in real time. The encoder generates a compressed bitstream that can be transmitted over any bit rate available over the range of bandwidth $[R_{\min}, R_{\max}]$. The base-layer bit rate has to meet the following constraint: $R_{\mathrm{BL}} \leq R_{\min}$. The enhancement layer is overcoded using a bit rate $(R_{\max} - R_{\mathrm{BL}})$. It is important to note that the range $[R_{\min}, R_{\max}]$ can be determined off-line for a particular set of Internet access technologies.

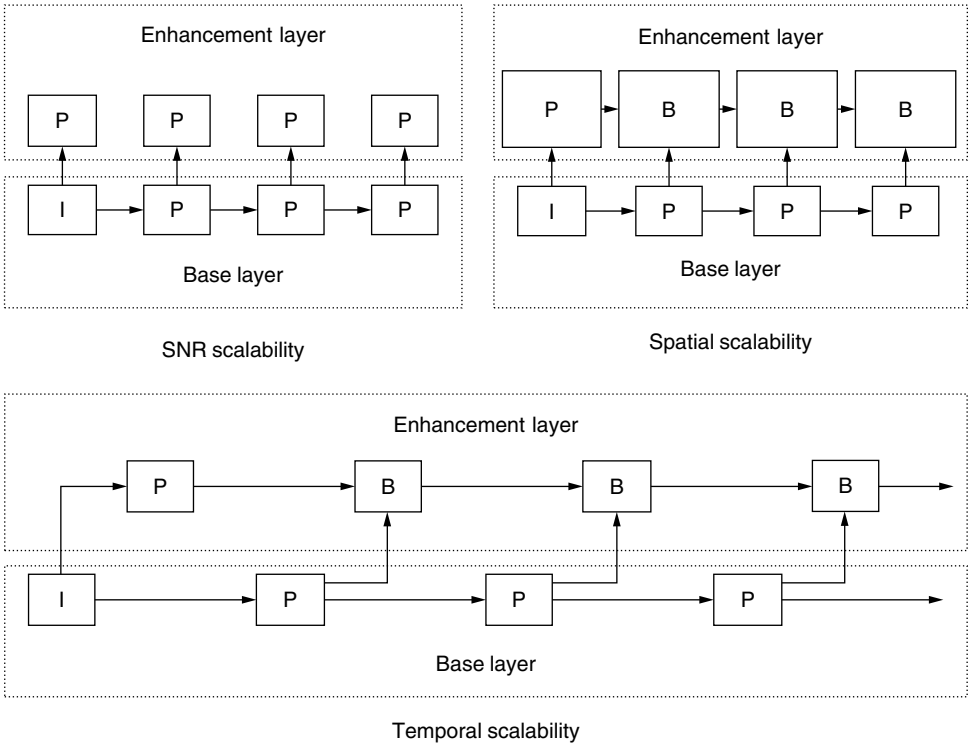


Figure 4. Examples of video scalability structures.

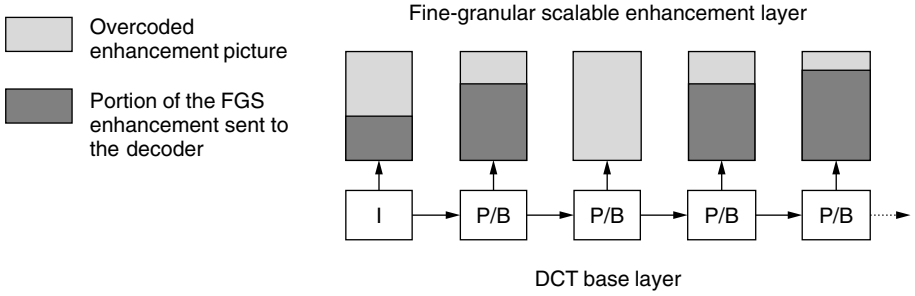


Figure 5. Video scalability structure with fine granularity.

For example, $R_{\min} = 20$ kbps and $R_{\max} = 100$ kbps can be used for analog modem/ISDN access technologies. More sophisticated techniques can also be employed in real-time to estimate the range $[R_{\min}, R_{\max}]$. For unicast streaming, the system estimates, in real time, the available bandwidth R for a particular session. On the basis of this estimate, the server transmits the enhancement layer using a bit rate R_{EL}:

$$R_{\text{EL}} = \min(R_{\max} - R_{\text{BL}}, R - R_{\text{BL}})$$

Because of the fine granularity of the enhancement layer, its real-time rate control aspect can be implemented with minimal processing. For multicast streaming, a set of intermediate bit rates $R_1, R_2, \ldots, R_N$ can be used to partition the enhancement layer into substreams. In this case, N fine-granular streams are multicasted using the bit rates:

$$R_{e1} = R_1 - R_{\text{BL}}, R_{e2} = R_2 - R_1, \ldots, R_{eN} = R_N - R_{N-1}$$

where $R_{\text{BL}} < R_1 < R_2 \cdots < R_{N-1} < R_N \leq R_{\max}$.

One can choose from many alternative compression methods when coding the BL and EL layers of the FGS structure shown in Fig. 5. For the FGS MPEG-4 standard, the base layer is coded using a DCT-based set of video compression tools. The FGS MPEG-4 enhancement layer is coded using an embedded DCT coding scheme. MPEG-4 FGS also support temporal scalability and hybrid SNR/temporal scalabilities. For more details regarding the MPEG-4 FGS scalable video method and many of its coding tools, the reader is referred to the paper by Radha et al. [18]. FGS is also very resilient to packet losses [19], which are common in streaming applications over the Internet.

4. PACKETIZATION AND TRANSPORT-LAYER ISSUES

Once coded, the next step in streaming is to transmit the compressed video over a network. Underlying network transport protocols such as TCP, UDP, or ATM [20] work well for transferring data over networks, but they are not very effective alone as the only packetization and fragmentation methods for streaming time-dependent media such as video or audio. When ignored, factors such as end-to-end delay, packet loss, delay jitter, network bandwidth bottlenecks, congestion, buffering, and decoder complexity/capability all can render a videostream useless.

4.1. Application-Layer Packetization

Many of the abovementioned issues can be addressed by intelligently packetizing video data at the application layer prior to sending it through the network transport layer using a transport protocol. Forming packets by breaking the data at their natural separation points is known as application-layer framing [21]. Frame, slice [10], or DCT block boundaries are good natural separation points for packetizing videostreams. In most cases, if properly framed packets are lost during streaming, the client will still be able to make use of the other packets to decode a viewable program. These application-layer packets can be created and stored prior to streaming, or they may be generated in real time as the video is being streamed.

At the application layer, framing alone, however, will not solve timing and synchronization issues in videostreaming. Transport protocols such as TCP and UDP can determine sequential relations between packets, but they have nothing to resolve real-time temporal relationships. Adding time indicators, or *timestamps*, to application-layer packets allows the client to know the temporal relationships between video packets. Timestamps allow clients to do things such as synchronize media streams, throw out packets that arrive too late to be decoded and displayed, and manage buffer control issues. One popular method of application-layer framing that has the above-mentioned features is the Real-Time Transport Protocol (RTP). Another standard that addresses properties of multimedia streams is Quicktime [22].

4.2. The Real-Time Transport Protocol (RTP)

The Real-Time Transport Protocol (RTP) [23] is a packetization protocol that is most commonly used at the application layer to packetize time-dependent data such as video or audio. RTP packetization is often done prior to transport-layer packetization such as UDP, so that RTP handles media-dependent issues such as timing, synchronization, and multiplexing, and UDP handles network-dependent issues such as data packet framing and multicasting. An example of this kind of multilevel packetization is shown in Fig. 6. Like other network protocols, RTP consists of a header and a payload. The RTP header contains bit fields to represent information such as sequencing, source identification, and timing. Detailed descriptions of the RTP header may be found elsewhere [23,24]. A few of the fields that are of particular relevance to videostreaming are:

Timestamp. This 32-bit field represents a sampling of a clock consistent with the type of data being

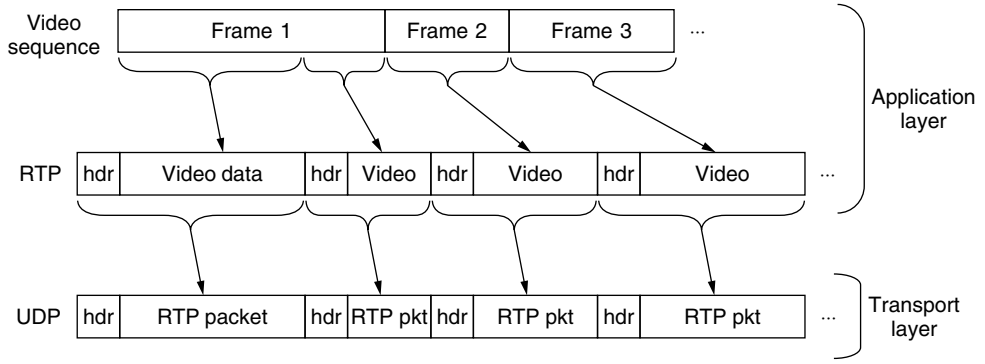


Figure 6. Multilayer packetization.

streamed. For frame-based video, every RTP packet containing data from the same video frame will have the same timestamp. For MPEG video, the timestamp typically has a resolution of 90 kHz.

Marker bit. A complete frame of compressed video may still be too large to put into one packet. If a video frame is split into multiple RTP packets, the marker bit typically is set for the last packet in the video frame. By setting the marker bit in the last packet, the client does not have to wait for the packet containing the start of the next frame to tell us that we can decode the current frame.

Sequence number. Since all packets of a single video frame have the same RTP timestamp, the 16-bit RTP sequence number, which is incremented for each RTP packet sent, can be used to determine the proper order of packets in a video frame. This way, the client does not have to look at the video data in the payload to determine the proper RTP packet sequence.

4.3. Packet Size Considerations

If the length of a packet set through the transport layer is larger than the network's maximum transmission unit (MTU) [20], the packet will be fragmented. It is therefore desirable to ensure that the RTP packet size is less than the network MTU so that the RTP packet will not be split into multiple packets. In the example shown in Fig. 6, the first video frame is too large to fit into one MTU, so it is split into two RTP packets. The obvious splitting method would be to simply fill one packet so that the resulting network packet size is just below the MTU value, and then put whatever is left over into the next packet. Since a transmitted packet may be delayed or lost, however, it is better to use application-layer framing, and choose a splitting point (or points) that take the video structure into account. For example, with MPEG-based video, the split could be done at a slice boundary. If, for example, the second packet of a two-packet frame is lost, the client will not have to throw away data in the first packet since they can be decoded into a viewable partial picture (assuming that a partial picture is something that the client would want to display). As another example, for multilayer scalable video, large frames could be split at layer boundaries so that a missing packet will have a lower probability of making the received packets for a frame unusable.

Large variations in packet size can also cause problems for videostreaming. In MPEG-2 and MPEG-4, I frames are typically large, while P and B frames are relatively small. Suppose that a 10-FPS video program is being streamed. The average time needed to transmit one frame would therefore be 0.1 s. The I frame, however, could require more than 0.1 s to stream, and the P and B frames would use less than 0.1 s. If, for example, the Group of Pictures (GOP) size (i.e., I frame period) is 2 s, we would have one I frame and 19 P or B frames in every GOP. If this video is coded so that the I frame is very large, the server could take over one second to transmit the I frame. This amount of time has the potential to cause problems if the decode buffer in the client is close to being empty, since a large I frame won't be available for display until all the packets in the frame have been received. It is therefore prudent for the content creator to use an encoding rate control method that is amenable toward streaming. The rate control method that results in the best compression ratio may not necessarily be the best method to use for streaming.

5. END-TO-END SESSION CONTROL AND MANAGEMENT

As described earlier, continuous decoding and playback of video at the receiver characterizes videostreaming and differentiates it from traditional applications such as email, FTP, and Webpage download. Consequently, additional challenges of videostreaming are to establish and maintain a regulated and continuous transmission of video data between the server and the client. This makes end-to-end session and rate control crucial aspects of any videostreaming solution.

5.1. Session Control

Some of the basic steps in multimedia session control are as follows:

- Choose a program to be streamed.
- Choose a rate at which the program will be streamed.
- Identify the network architecture over which the streaming will be done (e.g., single vs. multiple clients, unicast vs. multicast, UDP vs. http).

- Address other factors such as error control, media property rights, and encryption.
- Control the positioning, starting, stopping, and playing of the media stream.
- Adapt the streaming properties on the basis of various conditions related to the server, client, or network.

Some of the session control mechanisms used by today's commercially available streaming solutions are proprietary. An example of a publicly specified session control scheme is RTSP [25], which was designed for use over the Internet. A fundamental overview on how multimedia streams can be controlled over the Internet can be found in a paper by Schulzrinne [26].

The bidirectional nature of session control must also be taken into consideration when designing a streaming system. If the round-trip time of a network is not too large, proprietary methods or public protocols such as the RTP Control Protocol (RTCP) [23] can be used to help the server to adapt to varying network conditions and client configurations. In RTCP, sender and receiver reports are used to feedback timing, quality, and other related quantities to allow the server to adapt accordingly.

An aspect of session control that relates to video coding is the choice of rates used for streaming. If a viewer chooses to stream a nonscalable 128-kbps video program over a 56-kbps connection, a long startup delay may be incurred, as explained in the next section. One way commonly used to solve this problem is to encode the video at different rates, and store the multiple streams in different files. This method would allow the client (either automatically or user-driven) to choose the stream best suited for the given network connection. Encoding a video program at several different rates in separate files can be time-consuming, and it can create disk-space and file management problems for the content creator. To improve the server/client system, switched-rate streaming can be used. In switched-rate streaming, the video is encoded into a file that can be streamed at different rates. This way, the content creator can generate a file that can be streamed, for example, at 28, 56, and 128 kbps (and perhaps a few rates in between), and then the server can switch rates during streaming on the basis of feedback from the client. The problem of choosing rates during streaming can also be solved by using scalable video, as described in Section 3. If the video is continuously scalable, as in FGS, a variety of methods can be used to choose the appropriate packet size and streaming rate for the given connection [27].

5.2. Playout Buffer and Delay Considerations

If video were being transmitted or streamed over a guaranteed constant-rate network, the buffer internal to the video decoder (e.g., the decoder buffer in MPEG) would be sufficient to ensure a consistent viewing experience for the end user. Networks such as the Internet, however, are affected by loss and delays, which necessitate the use of a larger buffer at the client to store enough video so that the program will continue to be played on the client during these periods of rate variation, delay, or packet loss. A playout buffer is therefore used to store the videostream as it is being received. This buffer isolates the decoder from the network.

There is a tradeoff between the size of the playout buffer and the amount of delay incurred from when the server starts streaming to when the video is actually displayed on the client. This delay is also encountered when the playout buffer empties due to network congestion, and therefore must be refilled with an appropriate amount of video. One extreme way to abate network effects is to make the playout buffer large enough to contain the entire videostream, and then wait for this buffer to fill with the entire program before playing. This solution, of course, is not acceptable since the whole idea of streaming is to allow us to view the program without waiting for all of it to be sent to the client. If we initially fill the playout buffer with only a second or two of video, the viewer will be less likely to be annoyed by long waits. Having a short playout delay, however, increases the likelihood of having the buffer empty, resulting in more frequent occurrences of video pausing due to buffer refills.

If video is being streamed in real time (i.e., at the rate of the encoded video), then it would take, for example, 10 s to fill the buffer with 10 s of video. If the streaming session is not rate-limited, and if the network allows, the initial buffer-filling data can be sent faster than video rates. This way, the viewer may have to wait only a few seconds to have 10 s of video in the buffer. This solution is used by some of today's commercially available streaming systems to reduce waiting times at the client. It is important to note that if this high-speed buffer filling is done too frequently, it greatly increases the data rate used over the network, so a server should be designed not to overwhelm the network in cases when the client asks for buffer refills too frequently.

5.3. Congestion Control

In particular, a streaming video session needs to estimate the effective available bandwidth between the server and the client. This bandwidth estimate can be used to regulate the rate at which video is transmitted over the end-to-end session. Bandwidth estimation over a shared network, such as the Internet, is a very challenging problem. More importantly, bandwidth estimation is crucial for eliminating "gridlock" or *congestion* over the shared network. In other words, the large numbers of sessions (streaming or other applications) that use the Internet simultaneously need to employ some mechanism that provides a fair usage of the shared resources of the Internet. Without such a mechanism, the different sessions would be transmitting data at rates higher than the available bandwidth. This would lead to congestion over the shared network and eventually may lead to some form of gridlock. Consequently, this aspect of videostreaming, or available bandwidth estimation, is commonly known by the networking and Internet community as *congestion control*.

It is important to note that, prior to the emergence of streaming applications, congestion control represented one of the cornerstones of TCP. In fact, many experts

attribute the continuous success of the Internet and its growth to the ability of the TCP/IP protocol stack to provide a robust and scalable congestion control algorithm. The scalability of a congestion control algorithm is its ability to support a large number of sessions while (1) maintaining a robust level of fairness among these sessions and (2) eliminating the possibility of congestion (or gridlock) over the shared network.

The TCP congestion control algorithm is based on the following simple strategy. The server increases its sending rate linearly until a packet-loss event occurs. This event is detected by the receiver and communicated back to the server. Once a packet loss is detected, the server reduces its sending rate exponentially. Therefore, the TCP congestion control can be expressed as a linear increase–exponential decrease congestion control. Another popular characterization for TCP congestion control is additive increase/multiplicative decrease (AIMD) [28]. The original work in TCP congestion control is attributed to Jacobson [29].

Since the emergence of streaming applications, one of the key concerns has been the impact of these applications on the congestion of the Internet. In particular, streaming applications are supported by the UDP protocol rather than by TCP. UDP, however, does not provide any standardized congestion control mechanism. Consequently, for streaming applications, congestion control is the responsibility of the application. Therefore, real-time streaming applications that do not support some form of congestion control may become *misbehaving* in the sense that they may monopolize the available shared bandwidth over the Internet, and hence, they may become *unfair* to the mainstream (well-behaving) TCP applications. So far, this concern has been addressed by popular streaming solutions, such as those from RealNetworks [30] and Microsoft [31], through a relatively straightforward method. Multiple streams are generated at different bit rates and stored at the server. For example, streams for 56-kbps modems as well as for ISDN and higher rates (e.g., cable-modem or DSL rates) are compressed and made available for users. The application relies on the user to select the appropriate stream at the beginning of the streaming session. This simple approach is not sustainable for the long term, especially if we anticipate that streaming applications will become increasingly popular. Consequently, researchers have proposed several approaches for congestion control of media streaming, in general, and videostreaming in particular.

One of the most popular congestion control strategies that have been proposed and studied thoroughly for streaming applications is what is known as *TCP-friendly congestion control* [32]. The basic premise of this approach is that a streaming application follows a congestion control mechanism that is similar to the congestion control mechanism employed by TCP. This naturally makes streaming applications *fair* to the more popular TCP applications in terms of sharing the available bandwidth over Internet sessions. This fairness explains the reason behind the label "TCP-friendly."

6. FUTURE DIRECTIONS IN VIDEOSTREAMING

Videostreaming is expected to continue its growth toward a mainstream Web application. This growth has to be accompanied with successful efforts in addressing some of the key challenges in video coding and congestion control strategies. Regarding video coding, the key challenge is to provide new solutions that address the need for (1) scalability (i.e., to address the bandwidth variation, devices and network heterogeneity, and related QoS issues over the Internet); (2) new functionality (e.g., interactivity); and (3) providing high-quality video. New trends in video coding, such as 3D motion-compensated wavelet compression [33], could provide some answers to these challenges. Multiple description coding [34] is being looked at to stream video over channels with different characteristics. In addition to streaming frames of video, the object-based coding capabilities of MPEG-4 could be used to stream objects from various sources, which are composited at the client for new kinds of interactive streaming experiences. Moreover, proxy-based services that provide some form of *transcoding* of video may be a viable approach for addressing the need for scalable and high quality video. In particular, a new framework known as *TranScaling* has been proposed [35] to address the scalability and video quality issues of videostreaming over the wireless Internet. Under TranScaling, a scalable stream is mapped at a gateway server to one or more scalable streams with higher-quality video.

Regarding congestion control, further studies are needed in the area of TCP-friendly algorithms and related approaches. It is important to note that TCP-friendly congestion control represents a class of (or an umbrella of) mechanisms that are friendly to TCP sessions. This class of mechanisms includes, for example, the AIMD congestion control strategy mentioned above. Therefore, different types of TCP-friendly algorithms provide different levels of performance in terms of fairness and scalability. For example, it has been shown that AIMD TCP-friendly congestion control has many optimal and desirable attributes when compared with other TCP-friendly algorithms [36]. Moreover, other and more general congestion control frameworks have been proposed for streaming applications. This includes *equation-based* congestion control [37], *binomial* [38], and *ideally scalable* [36] congestion control algorithms.

Error resilience is also becoming increasingly important, especially when video is streamed over specialized channels. As the use of wireless networks increases, adding error detection, correction, and concealment to videostreams becomes just as important as the method used to compress the video. MPEG-4, for example, has error resilience built into the standard [39]. Finally, protecting digital property rights has become a topic of interest for streaming copyrighted material. Property rights management is being added to the latest versions of some of the more popular commercially available streaming systems described earlier in this article. Another way to protect the video is through the use of watermarking, in which a nonvisible (or barely visible) signal is added to the video [40]. This watermark would be present in

subsequent copies of the video, even if it is transcoded or recompressed. Solutions to all these challenges, combined with the increasing speed of network connectivity for end users, could soon make videostreaming a popular alternative to standard broadcast television.

BIOGRAPHIES

Robert Cohen received a B.S., summa cum laude, and a M.S., both in computer and systems engineering from Rensselaer Polytechnic Institute in Troy, New York. In 1990, he joined Philips Research in Briarcliff Manor, New York as a senior member of the research staff. At Philips Research, he was a project leader or team member doing research, development, patenting, and publishing in areas including video coding and signal processing, rapid prototyping for VLSI video systems, the Grand Alliance HDTV decoder, statistical multiplexing for MPEG video encoders, scalable MPEG-4 video streaming, and next-generation video surveillance systems. He has recently returned to Rensselaer Polytechnic Institute to complete his Ph.D. in electrical engineering. His current research interests include video coding and transmission, multimedia streaming, and image and video processing algorithms and architectures.

Hayder Radha received his Ph.D. ('93) and Ph.M. ('91) degrees from Columbia University, his M.S. degree from Purdue University in 1986, and his B.S. with honors degree from Michigan State University in 1984, all in electrical engineering. He joined Bell Laboratories in 1986 as a Member of Technical Staff. He worked at Bell Labs between 1986 and 1996 in the areas of digital communications, signal and image processing, and broadband multimedia communications. In 1996, he joined Philips Research as a principal member of research staff, and worked in the areas of video communications, networking, and high definition television. He initiated an Internet video research program at Philips Research and led a team of researchers working on scalable video coding, networking, and streaming algorithms. In 2000, he joined Michigan State University as an associate professor in the Department of Electrical and Computer Engineering.

He served as a cochair and an editor of the ATM and LAN Video Coding Experts Group of the ITU-T between 1994 and 1996. His research interests include image and video coding, multimedia communications and networking, and the transmission of multimedia data over wireless and packet networks. He has 25 patents in these areas (granted and pending). Dr. Radha received the Bell Laboratories Distinguished Member of Technical Staff Award and Appointment in 1993 and the Research Fellow Appointment at Philips Research in 2000. He is also a senior member of the IEEE.

BIBLIOGRAPHY

1. H. Eriksson, MBONE: The multicast backbone, *Commun. ACM* **36**: 68–77 (Jan. 1993).
2. M. R. Civanlar et al., guest eds., *IEEE Trans. Circuits Syst. Video Technol.* (Special Issue on Streaming Video) **11**(3) (March 2001).
3. H. P. Alesso, *e-Video: Producing Internet Video as Broadband Technologies Converge*, Addison-Wesley Professional, 2000.
4. G. J. Conklin et al., Video coding for streaming media delivery on the Internet, *IEEE Trans. Circuits Syst. Video Technol.* **11**(3): 269–281 (March 2001).
5. J. Alvear, *Web Developer.com Guide to Streaming Multimedia*, Wiley Computer Publishing, New York, 1998.
6. R. Schaporst, *Videoconferencing and Videotelephony: Technology and Standards*, 2nd ed., Artech House, Boston, 1999.
7. D. Wu et al., Streaming video over the Internet: approaches and directions, *IEEE Trans. Circuits Syst. Video Technol.* **11**(3): 282–299 (March 2001).
8. A. M. Tekalp, *Digital Video Processing*, Prentice-Hall, Upper Saddle River, NJ, 1995.
9. D. J. LeGall, MPEG: A video compression standard for multimedia applications, *Commun. ACM* **34**: 46–58 (1991).
10. J. L. Mitchell et al., eds., *MPEG Video Compression Standard* Kluwer, 1996.
11. A. Puri and T. Chen, eds., *Multimedia Systems, Standards, and Networks*, Marcel Dekker, New York, 2000.
12. *Video Codec for Audiovisual Services at 64 kBit/s*, CCITT Recommendation H.261, 1990.
13. *Video Coding for Low Bitrate Communications*, ITU-T Recommendation H.263, Nov. 1995.
14. A. N. Netravali and B. G. Haskell, *Digital Pictures — Representation and Compression*, 2nd ed., Plenum Press, New York, 1995.
15. S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, San Diego, 1999.
16. D. S. Taubman and M. W. Marcellin, JPEG 2000: *Image Compression Fundamentals, Standards, and Practices*, Kluwer, 2001.
17. H. Radha et al., Scalable Internet video using MPEG-4, *Signal Process. Image Commun.* **15**: 95–126 (Sept. 1999).
18. H. Radha, M. van der Schaar, and Y. Chen, The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP, *IEEE Trans. Multimedia* **3**(1): 53–68 (March 2001).
19. M. van der Schaar and H. Radha, Unequal packet loss protection for fine-granular-scalability video, *IEEE Trans. Multimedia* **3**(4): 381–394 (Dec. 2001).
20. W. R. Stevens, *UNIX Network Programming*, Vol. 1, 2nd ed., *Networking APIs: Sockets and XTI*, Prentice-Hall, Engelwood Cliffs, NJ, 1998.
21. D. D. Clark and D. L. Tennehouse, Architecture considerations for a new generation of protocols, *Proc. ACM SIGCOMM '90*: 201–208, Sept. 1990.
22. E. Hoffert et al., QuickTime: An extensible standard for digital multimedia, *Proc. IEEE Compcon Spring '92*, 1992, pp. 15–20.
23. H. Schulzrinne et al., *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, Internet Engineering Task Force, Jan. 1996.
24. M.-T. Sun and A. R. Reibman, eds., *Compressed Video over Networks*, Marcel Dekker, New York, 2001.
25. H. Schulzrinne, A. Rao, and R. Lanphier, *Real Time Streaming Protocol* (*RTSP*), RFC 2326, Internet Engineering Task Force, Apr. 1998.

26. H. Schulzrinne, A comprehensive multimedia control architecture for the Internet, *Proc. IEEE Workshop on Network and Operating System Support for Digital Audio and Video*, 1997, pp. 65–76.

27. R. Cohen and H. Radha, Streaming fine-grained scalable video over packet-based networks, *Proc. IEEE GlobeCom '00* **1**: 288–292 (Nov.–Dec. 2000).

28. D.-M. Chiu and R. Jain, Analysis of the increase and decrease algorithms for congestion avoidance in computer networks, *Comput. Networks ISDN Syst.* **17**: 1–14 (1989).

29. V. Jacobson, Congestion avoidance and control, *ACM SIGCOMM Comput. Commun. Rev.* **18**(4): 314–329 (Aug. 1988).

30. RealNetworks home page (online), *http://www.realnetworks.com*.

31. Microsoft Windows Media Technologies Player home page (online), *http://www.microsoft.com/windows/windowsmedia*.

32. M. Allman, V. Paxson, and W. Stevens, *TCP Congestion Control*, RFC 2581, Internet Engineering Task Force, April 1999.

33. S.-J. Choi and J. W. Woods, Motion-compensated 3-D subband coding of video, *IEEE Trans. Circuits Syst. Video Technol.* **8**(2): 155–167 (Feb. 1999).

34. J. K. Wolf, A. D. Wyner, and J. Ziv, Source coding for multiple descriptions, *Bell Syst. Tech. J.* **59**(10): 1909–1921 (Dec. 1980).

35. H. Radha, TranScaling: A video coding and multicasting framework for wireless IP multimedia services, *Proc. ACM SIGMOBILE Workshop on Wireless Mobile Multimedia*, July 2001, pp. 13–23.

36. D. Loguinov and H. Radha, Increase-decrease congestion control for real-time streaming: Scalability, *Proc. IEEE INFOCOM*, June 2002.

37. S. Floyd et al., Equation-based congestion control for unicast applications, *ACM SIGCOMM* **30**(4): 43–56 (Aug. 2000).

38. D. Bansal and H. Balakrishnan, Binomial congestion control algorithms, *Proc. IEEE INFOCOM 2001* **2**: 631–640 (April 2001).

39. R. Talluri, Error-resilient video coding in the ISO MPEG-4 Standard, *IEEE Commun. Mag.* **2**(6): 112–119 (June 1999).

40. S.-J. Lee and S.-H. Jung, A survey of watermarking techniques applied to multimedia, *Proc. IEEE Int. Symp. Industrial Electronics 2001* **1**: 272–277 (June 2001).

SURFACE ACOUSTIC WAVE FILTERS

PANKAJ K. DAS
ROBERT J. FILKINS
University of California at San Diego
La Jolla, California

1. FUNDAMENTALS OF SAW DEVICES

Electronic systems have made use of acoustic waves for many years. Early examples of acoustoelectric devices include delay lines that exploit slow acoustic velocities to provide long delays in a small package, and high-Q filters that use quartz resonators. The use of surface acoustic wave (SAW) devices came about with the development of the interdigital transducer (IDT). The interdigital transducer allows SAW devices to be mass produced using IC fabrication techniques and thus reduces manufacturing costs tremendously. Filters in color television were the first consumer application of SAW devices. Today SAW filters play a very significant role in the wireless and cellular phone industries. A typical mobile phone uses half a dozen SAW filters, which might constitute one-fifth of the total fabrication costs. Because of the demands of the wireless industry, the center frequencies of SAW filters have been pushed from 2.5 to 5 GHz and beyond. The major objective of this article is to introduce the reader to the fundamentals of surface acoustic wave devices. A second objective is to discuss how these devices apply to communication systems, such as wireless transceivers, optical receivers, cellular phones, spread-spectrum processors, and RF filters in general.

Surface acoustic waves have been known to scientists since 1885, when Lord Rayleigh presented his paper to the Royal Society. For this reason, surface waves are often referred to as *Rayleigh waves*. The Rayleigh wave propagates along the surface of a solid material with particle motion in the plane defined by the surface normal and the propagation direction. The wave amplitude decreases with distance from the surface. The surface wave is quite strong when compared with a bulk wave as the energy spreads in only two dimensions instead of three.

SAW devices developed through the 1980s continued to utilize the Rayleigh wave in their operation. However, as the need for higher-frequency and lower-loss devices grew (with the onset of wireless communication needs), researchers began developing devices based on other types of surface waves. For example, so-called leaky surface waves (LSAW) have been used in the development of 900-MHz filters for wireless radio transceivers. Other quasisurface acoustic waves include surface skimming bulk waves (SSBW), surface transverse waves (STW), and Bluestein–Gulyaev (BG) waves. Devices employing these different surface or quasisurface waves look very similar. They all use IDTs for generation and detection. They differ in the crystal orientation of the substrate materials. Each type of device will be cut to a different orientation, leading to acoustic wave propagation along different crystal axes. The resulting pseudosurface waves have desirable properties, which we'll discuss later. We will generally refer to all these types of waves as *surface acoustic waves*.

Before describing surface acoustic waves in greater detail we shall make a digression into the use of bulk acoustic waves in the electronics industry. This provides a basis for understanding the later development of SAW devices.

1.1. Bulk Ultrasound Devices

Two basic types of ultrasonic waves can exist within a solid material, namely, longitudinal and transverse. In the longitudinal (or compressional) wave type, the displacement of particles within the material is in the

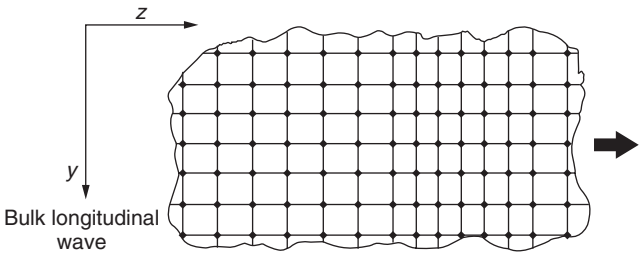


Figure 1. Bulk longitudinal ultrasound propagation in a isotropic solid [1].

direction of the propagating wave. As shown in Fig. 1, if a solid is divided into minute grid points, the longitudinal wave propagation causes some of the grid points to approach one another while other points are separated. The longitudinal wave moving in the z direction is composed of a series of compressions and rarefactions within the material. (Note that the lack of variation along the depth is the reason for calling a wave of this type a "bulk" wave.) Figure 2 shows the case of the transverse (or shear) wave type. The particle motion in this case is in the direction perpendicular to the direction of wave propagation, and the grid points now are translated either up or down. In three dimensions, there are two possible directions that are mutually perpendicular to the direction of motion. Therefore, in real solids two transverse wave modes are possible. These two shear waves are often referred to as vertically polarized (as in Fig. 2) and horizontally polarized (consider the particles in Fig. 2 were moving in and out of the page). The longitudinal and transverse bulk nondispersive waves can theoretically exist only within a solid of infinite dimension. For practical purposes, however, as long as the solid medium is considerably larger (e.g., 100 times greater) than the acoustic wavelength, the deviation from the ideal case is negligible.

The velocity of bulk transverse waves is roughly 3000 m/s, which is five orders of magnitude smaller than the velocity of electromagnetic waves in a solid — a fact that has been used to make ultrasonic delay lines. Consider the example of trying to obtain a 3-μs delay of a radar signal. To delay the electromagnetic wave, one requires 900 ft of coaxial cable. The equivalent delay for an ultrasonic wave traveling in solid substrate requires only 9 mm of material. The price to be paid for this compactness is the need for two transducers, one to convert the electrical energy to ultrasound, and a second to convert the ultrasound back to electrical energy. A scheme for a simple acoustic delay

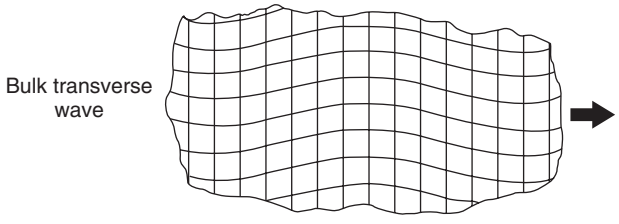


Figure 2. Bulk transverse wave traveling in a solid.

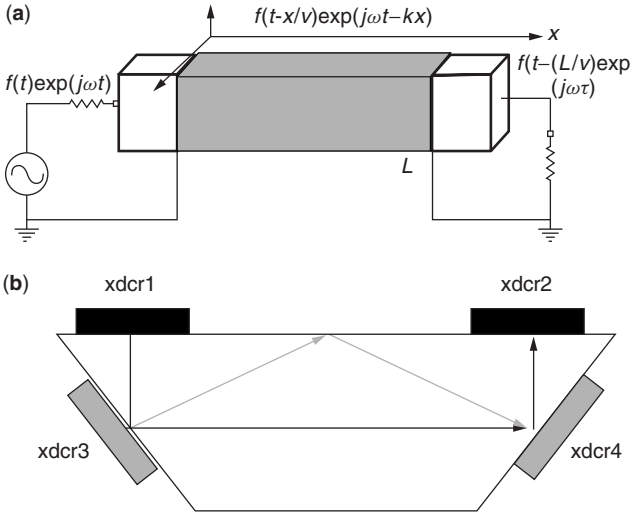


Figure 3. (**a**) Simplified bulk longitudinal delay line of length L — delay time τ is equal to L/v, where v is the longitudinal acoustic velocity; (**b**) compact delay line using low loss, bulk shear waves [3].

line is shown in Fig. 3a. The transducers are generally made from a parallel slab of piezoelectric material such as quartz or lithium niobate. The parallel sides of the transducer are metallized so that voltages can be applied. The excitation voltage is an amplitude modulated signal of the form $h(t)e^{j\omega t}$, where $\omega = 2\pi f$ is the carrier frequency. The ultrasonic signal generated within the solid can be represented by

$$h\left(t - \frac{x}{v}\right) e^{j\omega(t-x/v)} = h\left(t - \frac{x}{v}\right) e^{j(\omega t - kx)} \tag{1}$$

where the wavenumber

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{v} \tag{2}$$

and v is the velocity of wave propagation. The output signal at the second transducer for a delay line of length l is given by

$$h\left(t - \frac{l}{v}\right) e^{j\omega t} \tag{3}$$

which excludes a constant factor that accounts for propagation or transduction losses. Here the delay time τ is given by

$$\tau = \frac{l}{v} \tag{4}$$

A small device of this type can provide long delays; such devices have been used in electronic systems since the 1940s. In fact, early devices also utilized shear waves and internal reflections to generate compact delay lines of the type shown in Fig. 3b. Two of the possible delay paths are shown in this figure. Many other examples can be found in early literature [3].

Of course, the device structure can be extended to provide multiple delays. The so-called tapped delay line is useful in many electronic applications such as radar

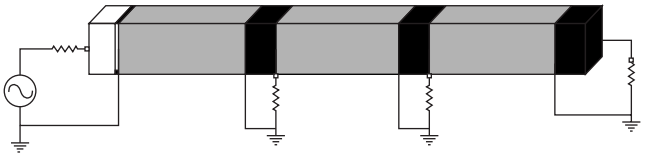


Figure 4. Simple representation of a tapped bulk wave ultrasound delay line.

processing. To obtain a tapped delay line with N taps using the abovementioned approach requires $N + 1$ transducers as shown in Fig. 4. In order to operate at high frequencies, each transducer must be carefully glued to the substrate, making the device rather cumbersome to fabricate.

A bulk wave in an unbounded medium is nondispersive. Often signal processing applications, such as pulse compression, require a dispersive delay, where each frequency travels with a different velocity. In this case a dispersive ultrasonic wave such as a plate or Lamb wave is used. A pulse compression filter, like those used in the 1950s to process radar chirp signals, is illustrated in Fig. 5. The dispersive property is shown in Fig. 6 as frequency versus delay time, which is

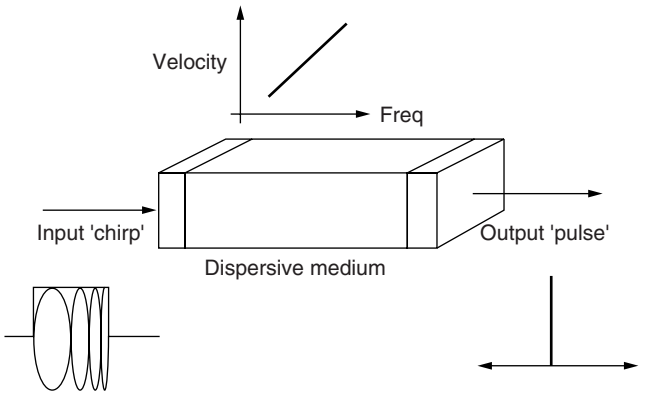


Figure 5. The use of a dispersive medium for pulse compression. The ultrasonic velocity is a function of the propagating frequency [1].

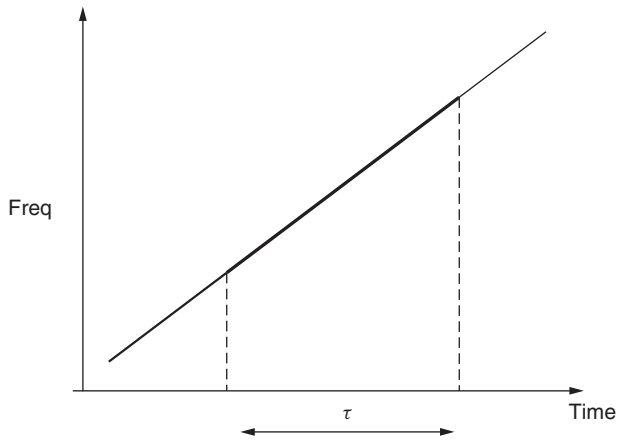


Figure 6. Dispersion represented by frequency as a function of time delay.

related to the frequency–velocity property of the wave. In other words, higher frequencies travel faster than do lower ones through the device. The desired input for this dispersive delay line is a "chirped" pulse whose instantaneous frequency is low at the beginning and increases progressively as function of time. As such a chirped pulse propagates through the delay line, the beginning of the wave is delayed more than the end of the wave. If everything is well matched, all the pulse energy will appear at the output at the same time, producing a spike [5]. Of course, the dispersion and the signal must be precisely matched in order to obtain good pulse compression.

1.2. Surface Acoustic Waves

Surface acoustic waves (SAWs) are more complicated than their bulk wave counterparts. The pure SAW or Rayleigh wave is a combination of the longitudinal and transverse wave components, which are elliptically polarized [1]; that is, the individual particles of the medium move in elliptical paths around their rest positions. This elliptical path is confined to the plane defined by the surface normal and the direction of wave propagation. A gridline representation of the particle behavior is shown in Fig. 7.

The exact derivation of the surface acoustic wave and how it comes to be confined to the free surface is quite complex [2]. The Rayleigh wave is a combination of the more general Lamb wave modes for a free plate. As the thickness of a plate becomes "infinite," the lowest-order antisymmetric (flexural) mode and the lowest-order symmetric (dilational) mode become degenerate, and combine to form the Rayleigh wave. The wave becomes tightly bound to the surface, leaving the interior of the plate undisturbed. The existence of such a wave was predicted from seismological research, where the earth's crust was considered to be a plate of infinite thickness.

The maximum SAW amplitude occurs right at the surface of the device. The wave amplitude decays rapidly in the direction perpendicular to the surface. The decay constant is approximately one wavelength long. For example, the acoustic wavelength in Y-cut $LiNbO_3$ at 100 MHz is roughly 36 μm for a wave propagating in

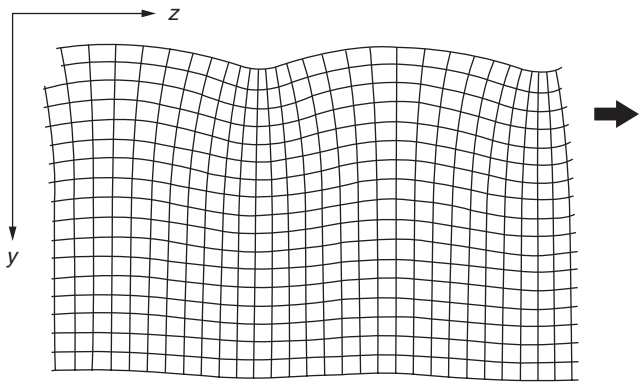


Figure 7. The displacements of a rectangular grid of material points as a result of SAW propagation in an isotropic material.

the z direction. (Note that "Y-cut" means that the normal to the surface is the y-direction of the crystal.) Therefore the substrate of a 100-MHz SAW delay line need only be a few hundred micrometers thick in order to work. Figure 8a summarizes the properties of SAW, showing the compressional and shear component motions of the grid points as well as the stress depth in wavelengths. That the wave is confined to the surface means that energy will spread in only two dimensions, and attenuate as $1/r^2$, instead of $1/r^3$, as for the bulk wave case. This feature gives Rayleigh waves another advantage for delay-line devices.

The SAW velocity is approximately 90% of the shear wave velocity. This is an important point to consider. Since the SAW phase velocity is slower than lowest bulk wave velocity, the Rayleigh wavelength measured along the surface is larger than any projected bulk waves. The Rayleigh wave cannot, in general, phase-match to any bulk wave components. Only in the presence of certain types of anisotropy or surface discontinuities can SAW combine with bulk waves.

1.3. Pseudo-SAW or Shallow Bulk Waves

In addition to Rayleigh waves, SAW devices may exploit other modes of surface wave propagation. Leaky SAW (LSAW), surface skimming bulk wave (SSBW), and surface transverse wave (STW) modes can be produced using interdigital transducers on various engineered substrates. The main advantages of these devices are

1. Wave propagation velocities are much higher than pure Rayleigh waves, allowing devices to operate at higher frequencies for the same lithographic tolerances. That is, the spacing between IDT fingers will represent a quarter wavelength at a higher acoustic frequency.
2. The electromechanical coupling coefficient, K, can be larger, leading to an increase in operational bandwidth and lower obtainable insertion loss.
3. Pseudosurface waves penetrate deeper into the substrate, allowing for larger acoustic power handling before nonlinear piezoelectric and acoustoelastic effects occur.
4. The temperature coefficient of the acoustic velocity can be carefully tuned for pseudosurface waves to enhance stability.

A pictorial summary of the various surface acoustic wave modes is given in Fig. 8b.

1.4. Piezoelectricity

An ultrasonic wave moving through a solid is a manifestation of stress applied to a crystal lattice. The

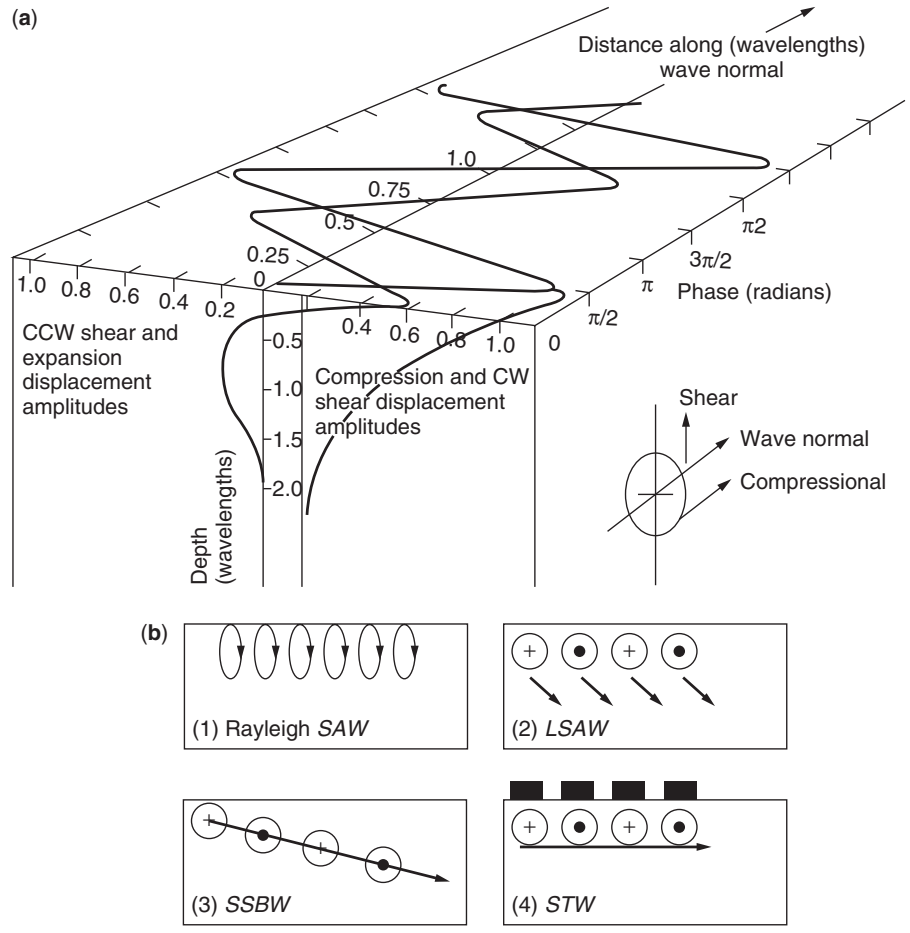


Figure 8. (**a**) Representation of surface acoustic wave. The grid points go through both shear and compressional motions [6]; (**b**) summary of surface acoustic varieties: (1) Rayleigh or pure SAW mode; (2) leaky SAW propagation from surface into substrate; (3) surface skimming bulk wave (SSBW); (4) surface transverse wave (STW) [4].

resulting strain alters the ionic equilibrium positions for the lattice, and in some cases a net electric polarization occurs. This stress-induced polarization is known as the *piezoelectric effect*. Piezoelectric materials also exhibit the reciprocal behavior; when an electric field is applied to the crystal, the ionic positions move, the lattice changes size or shape, and strain is produced. This behavior allows piezoelectric materials to behave as acoustoelectric transducers.

The topic of piezoelectricity is treated in great depth in a number of other texts [3], but we shall briefly describe some of the key aspects of piezoelectric materials. First, what is it that makes a material piezoelectric? To answer that question, one should first consider that crystalline materials are comprised of a regularly repeating pattern of atoms. As a result of chemical bonding, the atoms in the lattice share electrons. When atoms of different sizes share bonding electrons, it is possible that the electrons will "spend slightly more time" with one of the elements. Therefore, on average, one atom can appear negatively charged, and another can appear positively charged. A net electric dipole is formed by the atoms. The polarization effect is accentuated by different interatomic distances, atomic sizes, and crystal orientation. This built-in dipole will repeat over the entire crystal, compensating itself so that the crystal has no net potential. At the ends of the crystal, impurities or other matter will eventually appear to compensate the built-in electric field. The application of mechanical stress will disturb the compensated dipole and create an electric potential. Alternatively, if an electric field is applied, the built-in dipoles will respond. The strength of the response depends on the orientation of the dipole to the applied electric field or mechanical stress. For example, the torque applied to a dipole is greatest when the electric field is initially oriented perpendicular to it.

What we have just described is a coupling between the dielectric and elastic properties of an anisotropic crystal. The precise coupling is described mathematically in terms of the elastic, dielectric, and piezoelectric constants. Their derivation is based on thermodynamic potentials and is beyond the scope of this article, but the key points are summarized. Each set of constants comprises a vector or tensor matrix. The elastic constants relate two second-order tensors (stress and strain) and are therefore a fourth-order tensor. It can be shown that in general there are at most 21 (twenty-one) elastic constants for a crystal. The dielectric constants relate two vectors (electric and polarization fields) and is therefore a second-order tensor. There can be at most six (6) permittivities. The piezoelectric constants relate a second-order symmetric tensor (stress or strain) to a vector (electric field) and is therefore a third-order tensor. There can be as many as 18 (eighteen) piezoelectric constants. The electromechanical coupling efficiency, K (or sometimes K^2), is defined as the ratio of the mutual elastic and dielectric energy density (U_m) to the geometric mean of the dielectric (U_d) and elastic (U_e) self-energy densities.

$$K = \frac{U_m}{(U_d \times U_e)^{1/2}} \tag{5}$$

The value of K can also be described in terms of the individual material constants: s (elastic), ε (dielectric), and d (piezoelectric). A simple expression is

$$K = \frac{d}{(s\varepsilon)^{1/2}} \tag{6}$$

One final point to make about the effect of piezoelectricity on substrates is that it alters the mechanical stiffness of the material. The additional reaction force provided by the piezoelectric field can be combined with the normal elastic coefficients to create a set of stiffened elastic coefficients for the material [2]. The importance of this fact will be seen when we consider waveguides, transducers, and gratings.

In summary, we see that piezoelectric properties of a material are based on the absence of crystal symmetry, and that the exact piezoelectric behavior depends very much on the orientation of the crystal lattice to the applied external forces. These facts form the basis of substrate engineering for SAW devices.

1.5. Substrate Materials

A good piezoelectric substrate is a vital ingredient for a SAW device. Quartz is the only naturally occurring piezoelectric material used to manufacture devices. The largest piezoelectric constant for quartz is $d_{11} = 2.31(\times 10^{-12} C/N)$, which is smaller than other synthetic (human-made) SAW substrates [3]. Quartz does have the advantage of low acoustic loss, and low dielectric constant (which leads to low capacitance per unit length) when compared to synthetic piezoelectrics. Most of the crystals used in manufacturing devices are grown in a laboratory environment. Computer modeling of the crystal parameters was used to develop ST-cut quartz, the most readily used quartz cut for SAW devices. The object of ST-cut quartz is to provide very stable acoustic velocity over temperature; the temperature coefficient is nearly zero ppm/°C. The SAW propagation direction for ST quartz is the x direction with a velocity of 3.158 mm/μs.

Lithium niobate is probably the most important synthetic material for SAW substrates. It was discovered in the late 1950s, a time of great interest in synthetic piezoelectric materials. Typically, lithium niobate ($LiNbO_3$) used for devices is a congruent crystalline mixture of Li_2O and Nb_2O_5, composed of 48.6% Li_2O [22]. Most SAW devices are built of Y-cut, z-propagating $LiNbO_3$ substrates. The Rayleigh velocity is 3.487 mm/μs for this cut.

Lithium tantalate ($LiTaO_3$) is another popular substrate material. A rotated Y-cut, z-propagating crystal is used for pure SAW devices. The velocity is 3.254 mm/μs. The mechanical coupling factor is lower than that of lithium niobate, owing in part to a lower dielectric constant. However, lithium tantalate has the advantages of reduced degree of dielectric anisotropy, smaller temperature coefficient, and lower acoustic diffraction.

As described earlier, pseudo-SAW devices are obtained by using different crystal orientations and in some cases different substrate materials. Examples of useful orientations for propagation of LSAW on lithium niobate include 64° YX-cut, 41° YX-cut, and 36° YX-cut. Lithium tantalate is used as a substrate material for LSAW-

and SSBW-type devices. For example, impedance filter elements based on SAW gratings or resonator structures employ a 36° YX-cut $LiTaO_3$ substrate. Each substrate and crystal cut has a different acoustic velocity, temperature coefficient, electromechanical coupling, and attenuation factor. The designer must choose what is best for a given filter application. The interested reader should consult Ref. 23 for a complete list of SAW substrates and their properties.

1.6. Thin Films

Thin films are often used in the design of surface acoustic wave devices [16]. As we shall see in subsequent sections, thin metal films are essential for the operation of SAW devices. They provide a means of generation, detection, and directional control of surface acoustic waves. The need for higher-performance devices pushes research toward better metal films. For example, to increase operating power and bandwidth requires metal films that resist electromigration and acoustic migration breakdown, while maintaining high conductivity at high frequencies.

Dielectric films can be added to SAW substrates to provide a variety of performance enhancements. A thin amorphous film, such as glass, deposited on a piezoelectric substrate provides the following advantages: surface passivation, reduction of pyroelectric effects (the buildup of a surface potential due to temperature changes), and smoothing to reduce propagation loss. Dielectric films are also used to modify the electromechanical coupling factor, reduce the frequency dependent temperature coefficient of velocity, and tune the acoustic velocity. Velocity tuning is an essential part of designing acoustic waveguides.

Piezoelectric films are used in several ways in advanced SAW devices. The first application is to provide substrate materials with high acoustic velocity. The higher velocity allows fabrication of higher-frequency devices without changing the lithographic feature sizes. Piezoelectric films are also useful for increasing coupling factors, and increasing the acoustic nonlinearity of the substrate. The nonlinearity is used in acoustoelectric convolver devices. A final advantage of piezoelectric thin films is to allow the integration of SAW devices with other microelectronic circuits. For example, depositing ZnO on silicon or gallium arsenide substrates to create integrated amplifier and SAW filter circuits.

2. SAW BUILDING BLOCKS AND DEVICES

2.1. Acoustic Impedance, Waveguides, and Gratings [2,5]

In order to understand how surface acoustic wave devices operate, we should say a little bit about wave reflection and transmission at material discontinuities. The Rayleigh phase velocity is the simplest parameter to describe the dispersion relation for the propagation of acoustic waves across the piezoelectric substrate. But in order to describe the behavior at material interfaces, a second parameter is often introduced: the *acoustic impedance*. The surface acoustic impedance is defined as

$$Z = \rho V_R \tag{7}$$

where ρ is the material density in kg/m^3 and V_R is the Rayleigh wave phase velocity in m/s. This impedance definition allows one to adapt a transmission-line model for the wave propagation. At the interface between materials of different acoustic impedance, some of the acoustic energy is transmitted, some reflected. A reflection coefficient R and a transmission coefficient T can be defined as

$$R = \frac{Z_1 - Z_2}{Z_1 + Z_2} = \frac{\rho V_{R1} - \rho_2 V_{R2}}{\rho V_{R1} + \rho_2 V_{R2}} \tag{8a}$$

$$T = 1 - R \tag{8b}$$

This description is oversimplified, as it neglects the effects of wave polarization, incidence angle, and propagating modes, but it gives the reader a feel for what happens in a SAW substrate when the acoustic velocity changes abruptly.

As mentioned in the preceding section, the presence of thin films on the surface of a SAW substrate affects wave velocity. The first, and maybe most important, example is the presence of a metal film on the substrate. The conductivity of the metal film has the effect of shorting (short-circuiting) the piezoelectric field in the material. This, in turn, alters the stiffness and therefore the velocity of the material under the metal film. The change in SAW velocity due to the shorting of the piezoelectric field allows for direct measurement of the electromechanical coupling coefficient

$$K^2 = \frac{-2\Delta V_R}{V_R} \tag{9}$$

where ΔV_R is the fractional change in SAW velocity and V_R is the free surface SAW velocity. An impedance mismatch also occurs for the propagating surface wave. This action forms the basis of acoustic waveguides and gratings, which we will discuss next.

The basic function of an *acoustic waveguide* is to confine the acoustic wave propagation to an area of the substrate, much like a fiberoptic cable for light, or a coaxial transmission line for RF energy. The waveguide confinement is used in SAW devices to overcome beamspreading losses, to redirect signals, to correct phase fronts, to increase the acoustic energy density (important in nonlinear devices like convolvers), and to generally create the notion of an acoustic "circuit." All acoustic waveguides operate on the basis of the $\Delta V_R/V_R$ action, but different structures are possible to achieve this result. Waveguides can generally be divided into three main classes: flat overlay waveguides, topographic waveguides, and other engineered waveguides. Several possible structures exists within each class. Examples of each type are shown in Fig. 9.

The first type of waveguide we shall discuss is the flat overlay waveguide. The simplest type of overlay waveguide is called a strip waveguide (see Fig. 9a). In this waveguide, a thin layer of dielectric material is deposited on the SAW substrate. The dielectric strip will typically have a slower acoustic velocity when compared to the SAW substrate. The $\Delta V_R/V_R$ action confines the acoustic wave to within the slower material. Note that in the most general case,

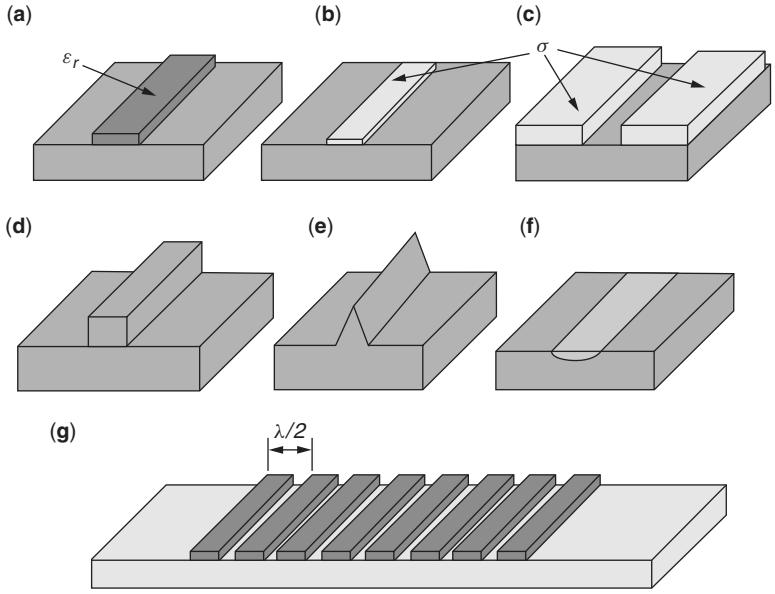


Figure 9. Various types of acoustic waveguides structures. Flat overlay types: (**a**) mass loading strip of dielectric film; (**b**) shorting type with thin metal layer; (**c**) slot waveguide using shorting type metal films. Topographic waveguides; (**d**) rectangular ridge waveguide; (**e**) wedge waveguide, Engineered type; (**f**) in-diffused waveguide structure [5]; (**g**) acoustic grating structure.

neither the substrate nor the overlay strip needs to be piezoelectric.

A second type of overlay waveguide is the shorting (short-circuiting) strip guide. In this structure (Fig. 9b) a thin conducting film is placed on a piezoelectric SAW substrate. The conducting film short-circuits the piezoelectric field and reduces the acoustic velocity under the strip. The acoustic wave is weakly confined to the area under the strip. By changing the thickness of the conducting film, it is possible to introduce a mass loading effect as well. The combination of mass loading and piezoelectric shorting can be used to carefully tune the dispersion behavior of the waveguide.

A slot waveguide is produced by depositing a film with a faster acoustic velocity on the SAW substrate as shown in Fig. 9c. In this case the acoustic wave is confined to the "slot" area between the film overlays.

The topographic waveguide is produced by selectively removing an area of the substrate to create a ridge or wedge confinement region (see Fig. 9d,e). In the ridge or rectangular waveguide, the acoustic wave propagates as a function of the bending modes of the ridge section. The bandwidth of such a structure is therefore limited by the geometry of the ridge. The wedge is a means of creating a broader band waveguide. The tapered geometry of the wedge creates a wide range of frequencies that can propagate [5]. A key advantage of the topographic waveguide structure is its low loss.

One final type of waveguide structure is shown in Fig. 9f. This is the diffused waveguide structure. The acoustic velocity is altered within a region of the substrate by diffusing another material into the substrate. One example is to diffuse titanium into lithium niobate to create a confining region. The advantage of diffused waveguides over strip types is also lower loss.

Another SAW device structure based on the $\Delta V_R/V_R$ effect is the acoustic *grating*. The object of the grating is to create an acoustic reflector. The reflecting element can then be used to create a variety of wave control devices including filters. The device is analogous to the Bragg grating in optics. The grating structure, as shown in Fig. 9g, is a series of strips. The strips may be created using any of the structures described for waveguides. Two of the more common structures are shorting metallic strips and ridge guides. The metallic strips are deposited and etched using photolithography, whereas a ridge structure can be created by e-beam (electron-beam) lithography. The strips are separated by a half the acoustic wavelength of interest. Recall that at each impedance interface, an amount of the energy is reflected. Separating the strips by one-half wavelength ($\lambda/2$) allows each reflection to add in phase. A large number of strips in sequence are needed to reflect all the energy. The optimum number of strips is a function of the structure used and the substrate material. For example, to obtain near-100% reflection using shorting strips on lithium niobate requires on the order of 100 strips [4].

This brief introduction into acoustic waveguides and gratings lays the foundation for discussing the structure and design SAW devices, the devices that, in turn, are used to create SAW filters. The $\Delta V_R/V_R$ effects in some cases are the basis of device operation, and in other cases (such as IDTs) create nonideal behaviors that must be compensated for in some way. Now let's discuss interdigital transducers.

2.2. Interdigital Transducers

The interdigital transducer (IDT) is shown in Fig. 10b. The IDT is simply a set of metal strips placed on the piezoelectric substrate. Alternate strips, or "fingers", are interconnected to form two interdigitated electrical contacts. The width of each finger and the distance between fingers is usually one-quarter of the acoustic wavelength to be generated. When a RF voltage is applied to the two contacts, an electric field is simultaneously set up between all adjacent fingers. This has the effect of alternately compressing and elongating the piezoelectric substrate between the fingers, producing a distributed

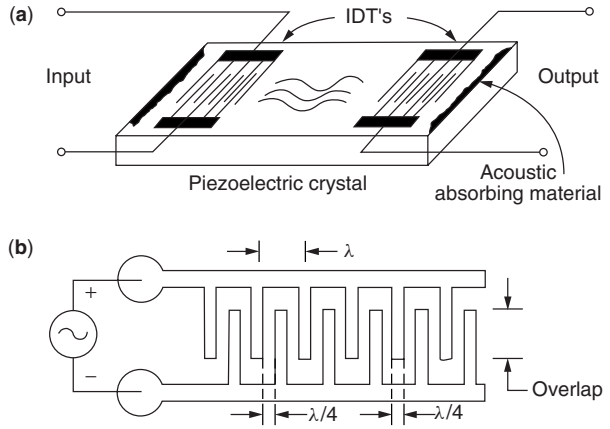


Figure 10. Generation of a surface wave using a bulk transducer — this is improved using IDTs: (**a**) SAW delay line made from two IDTs; (**b**) diagram of an IDT composed of a set of metal strips placed on a piezoelectric substrate [7].

source of strain. The resulting acoustic waves propagate to the right and left, as the superposition of all of the waveguide modes of the substrate plate structure. Because the electric and strain fields are confined close to the substrate surface, the generation of surface waves is strongly favored. The type of surface waves excited are a function of the applied electric field, the piezoelectric matrix for the substrate, and the relative orientation of the crystal axes. In some cases, for example Y-cut lithium niobate, Rayleigh waves are generated. In other cases, for example, 36° YX-cut lithium tantalate, LSAW waves are favored.

To efficiently generate a surface acoustic wave, all the sources must add coherently. Consider the RF voltage with period T; to propagate a distance $\lambda/2$ takes $T/2$ seconds. This is also the time it takes for the electric field between two adjacent fingers to reverse polarity. Therefore, all the generated waves will add together constructively. The reverse process occurs for detection; that is, the elastic deformation passes under the fingers and induces in-phase voltages between each pair of electrodes. This type of operation makes the IDT a simple and efficient SAW generator.

Next, we shall describe the operation of SAW delay lines, and in the process we'll uncover some of the limitations of and improvements for the simple IDT structure.

3. SAW DEVICES FOR FILTERS

3.1. Delay Lines

A delay line is constructed using two IDTs on a piezoelectric substrate as shown in Fig. 10a. If lithium niobate is used as the substrate the surface velocity is about 3600 m/s, which translates into an acoustic wavelength of 36 μm at 100-MHz operation. Using the guidelines previously mentioned, the IDT finger width and spacing would be 9 μm. Of course, the delay is a function of the spacing between the two IDTs; For example, if the IDTs are separated by 1 cm, the delay for the SAW device will be 2.3 μs.

Why use many finger pairs instead of only one pair for IDTs? Actually, an IDT can be (and sometimes is) made using only a single pair. However, greater transducer efficiency is achieved for the same applied voltage, since for N finger pairs the resulting N sources will add coherently. As if often the case, there is an inherent gain–bandwidth trade-off. There is a reduction in operating bandwidth as more and more finger pairs are added. In general, the optimum number of finger pairs (N_{opt}) is proportional to the reciprocal of the electromechanical coupling coefficient [5]. That is

$$N_{\text{opt}}^2 = \frac{\pi}{4}\frac{1}{K^2} \tag{10}$$

where K^2 is the squared coupling coefficient of Eq. (9).

The two characteristics of SAW delay line filters that limit the performance are delay-line losses and triple-transit echo. Delay line losses can be divided into four: IDT loss, propagation loss, misalignment or steering loss, and diffraction loss.

An inherent loss of 6 dB occurs in a SAW delay line because an ordinary IDT generates waves in the forward and backward directions. This division of energy gives 3 dB of loss per IDT, transmitter, and receiver. The loss can be reduced or eliminated through the use of unidirectional transducers. The design of unidirectional transducers is reviewed in the next section.

Now let's move to the second loss mechanism in SAW delay lines — propagation loss. Propagation loss is characterized by a frequency-dependent loss per unit length. The following functional form can be used to describe this behavior:

$$A(x) = A_0 e^{-\alpha(f)x} \tag{11}$$

As SAW propagates along the surface of the device, the amplitude starts out with amplitude A_0, and decays as function of x and $\alpha(f)$. As the SAW frequency increases, the material deformation is unable to accurately follow the RF field and produces a phase term. This, in turn, creates a propagation loss. In general, the value of $\alpha(f)$ is proportional to the square of the frequency. An additional loss term is created by air loading of the surface. If the delay line operated in a vacuum, the loading loss would be zero. In most cases, a SAW device package is hermetically sealed with an inert gas such as nitrogen. In this case the loading loss is negligible, becoming more dominant at low frequencies.

Beam steering loss is ideally zero if the substrate is cut to the proper direction. The proper direction is the one where the power flow angle is zero. If the IDTs become misaligned with the crystal axis, or the cut is incorrect, the power flow angle will be nonzero, and loss will develop. As with propagation loss, the steering loss is worse for long delay lines.

Diffraction loss is the result of the finite-sized aperture of the IDT. The acoustic wavefront can be described in terms of a near-field, or Fresnel region, and far-field or Fraunhofer region. This is just as in the case of optics. In the far-field region the wavefront spreads out to a width larger than the aperture of either the transmitting or

receiving IDT, placing a limit on the length of a delay line for a given IDT aperture length. The calculation of this loss is complicated by the anisotropic nature of the substrate crystals. Models are available that predict this loss to within a fraction of a decibel. In principle, acoustic waveguides can be used to overcome this loss in long-delay devices.

3.1.1. Triple-Transit Echo. We expect that when an RF signal is applied to a SAW delay line, the output will be an exact copy of the input, delayed only by some time τ and perhaps lower in amplitude. It is possible for other outputs to occur. The most important and troublesome extraneous output is the triple-transit echo. As the name suggests, the extra output appears at time $t = 3\tau$ and is caused by reflections of the SAW off the IDTs. For a delay line with two matched IDTs, this echo is 12 dB lower than the main signal. It is worth noting how so much SAW energy can be reflected off the IDTs. As mentioned previously, the acoustic velocity is different in the regions of the substrate covered by metal electrodes. The net $\Delta v/v$ change in the acoustic impedance along the SAW transmission line generates small reflections at each IDT finger. The reflections add coherently due to phase matching, and the result is a significant extra echo. A split-finger IDT is one means of reducing the phase-match condition from occurring (see Fig. 11). Each quarter-wave finger pair is separated into two lambda (2λ) by eight fingers. A multistrip coupler can also be used to trap the triple transit echo (this will be discussed in a subsequent section).

If we pause here for a moment and compare a SAW delay line to a bulk ultrasound equivalent device, we immediately see the advantage of the SAW device. The SAW device can be fabricated using a one-step photolithographic process, whereas the bulk ultrasound device required gluing at least two transducers to the substrate. In fact, for a delay line with N taps, the bulk ultrasound device would require gluing $N + 1$ transducers. The SAW device is fabricated in one step regardless of the number of taps, owing to the lithographic processing. The cost advantages in both time saved and device yield are clear. Also, the two-dimensional nature of the SAW device makes it consistent with other integrated circuit techniques for packaging and testing.

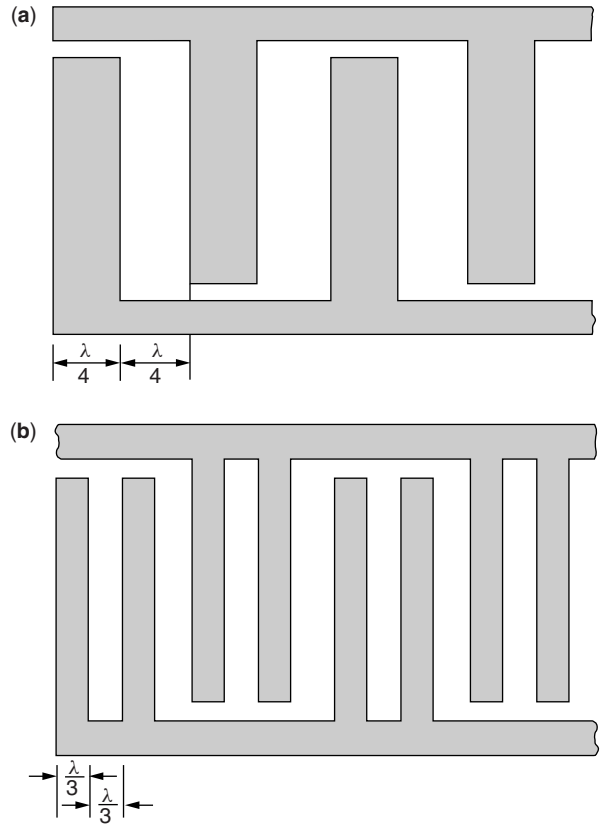


Figure 11. Diagram of transducers with (**a**) single electrode and (**b**) double or split electrodes per half wavelength. The split electrodes provide a means of canceling the triple-transit echo in delay lines.

3.2. Unidirectional Transducers

Unidirectional transducers correct the 3-dB loss of the bidirectional IDT, and overcome the problem of triple-transit echo. There are four techniques for making low-loss unidirectional transducers:

1. Two IDTs separated by a quarter wavelength
2. Three-phase excitation
3. Single-phase unidirectional transducer (SPUDT)
4. Multistrip couplers

The first type of *unidirectional transducer* is shown in Fig. 12. The transducer consists of two IDTs separated by a quarter-wavelength gap, which produces a phase difference of 90°. The second IDT is also terminated such that the load will completely reflect the SAW. In operation, the excited backward wave is reflected by the second

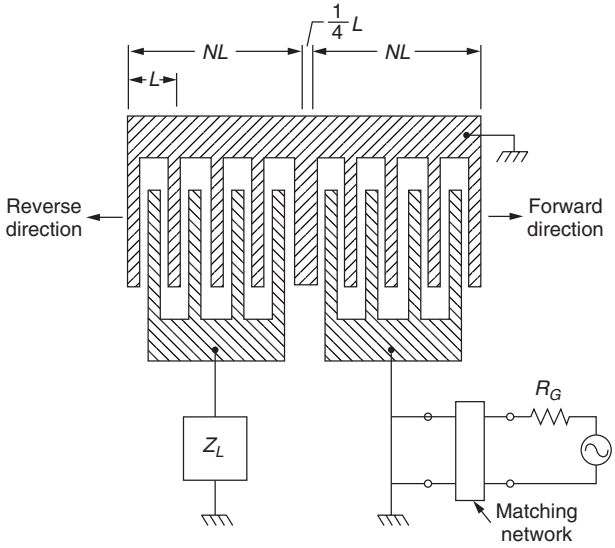


Figure 12. A unidirectional transducer using a reflector IDT spaced a quarter wavelength from the driven IDT [9].

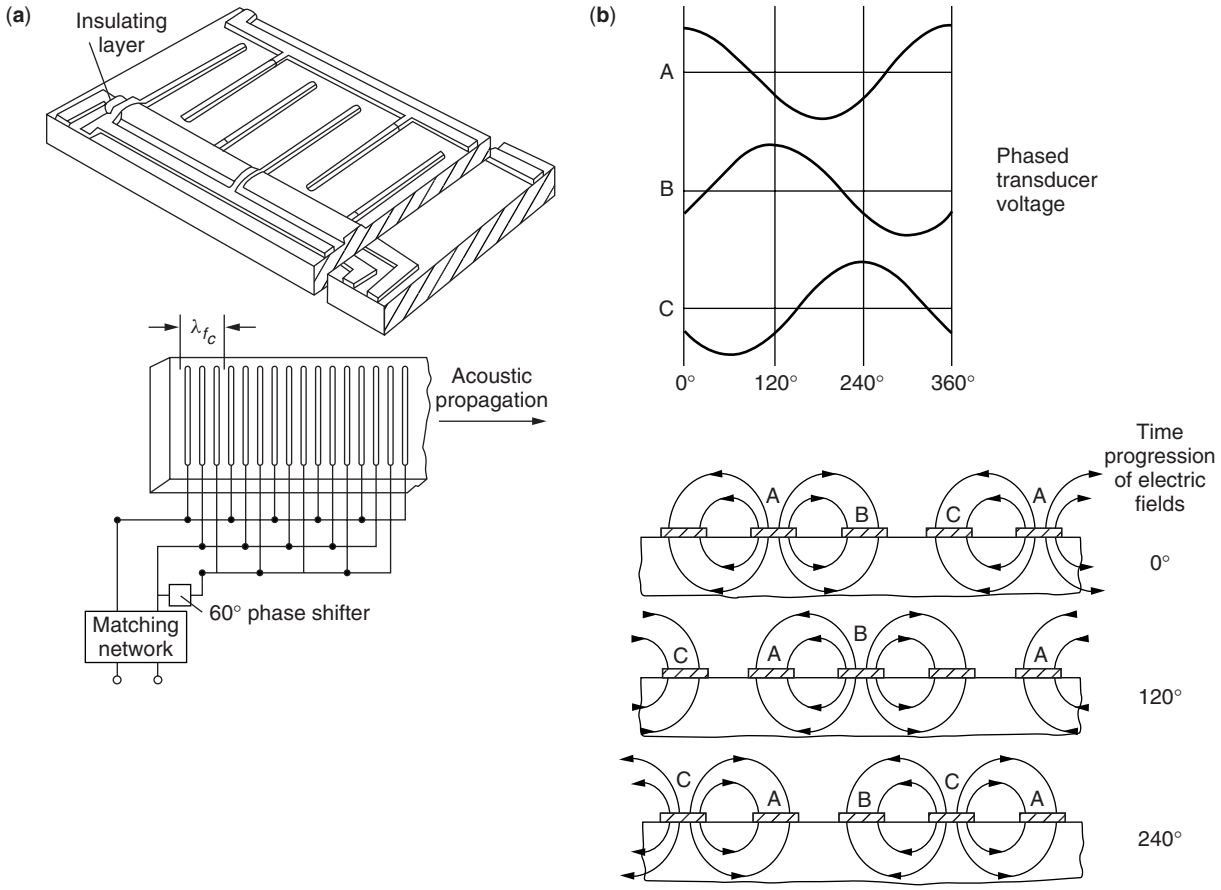


Figure 13. (**a**) Three-phase excitation of an IDT [10]; (**b**) the resulting interelectrode electric field evolution for three-phase excitation.

transducer with a phase shift, allowing it to reinforce the forward moving wave. This cancels the 3-dB loss of a simple IDT. One drawback, however, is a limitation on achievable frequency response.

The *three-phase excitation IDT* is more complex, but effective. As shown in Fig. 13, the IDT now consists of three electrode fingers. The RF drive signal is converted into a three-phase signal using a 60° phase shifter. Each drive signal is thus separated by 120°, as illustrated in Fig. 13a. The resulting interelectrode electric field evolution is shown in Fig. 13b. Only the forward traveling wave propagates; the backward wave is canceled out. Of course, the direction can be reversed by adding an additional 120° phase shift to the drive signals. Fabrication of the three-phase IDT is further complicated by the need for an extra insulator to isolate one of the IDT fingers.

The *single-phase unidirectional transducer* (SPUDT) is shown in several forms in Fig. 14. The SPUDT uses floating electrodes arranged to form an acoustic grating, designed to reflect the traveling SAW. When properly designed, the reflector reinforces the forward moving SAW, and cancels the backward moving SAW. The basic SPUDT device has three configurations [4]: single floating electrode (Fig. 14a), double floating electrode (Fig. 14b), and a comb type (Fig. 14c).

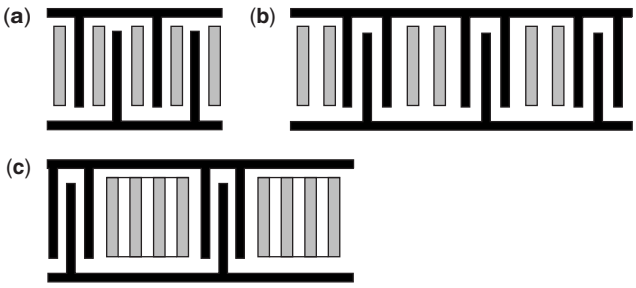


Figure 14. Finger layout for a single-phase unidirectional transducer (SPUDT). The basic device has three configurations: (**a**) single floating electrode; (**b**) double floating electrode; (**c**) comb-type SPUDT using longer grating elements [2].

3.3. Surface Acoustic Wave Filters Based on Delay Lines and Apodization

We shall now begin our discussion of SAW filters by showing how a tapped-delay structure can be used to create a transversal, or finite-impulse-response (FIR), filter. The concept of transversal filters has been developed fully in other texts. We start here by noting that the transversal filter has a transfer function given by

$$H(\omega) = \sum_{n=0}^{N-1} C_n e^{-j\omega T_d} \tag{12}$$

This transfer function corresponds to a delay line where each of the N output taps is multiplied electronically by the appropriate coefficient C_n. Each of the N products is then summed electronically. A programmable filter can be achieved by simply switching different values for $\{C_n\}$. If one is merely interested in a fixed transversal filter, the value of C_n can be designed into the transducer itself. The transducer gain adjustment can be achieved many different ways, but we shall discuss a method called *apodization*.

Consider the IDT shown in Fig. 15 and assume that the same SAW is incident on both finger pairs A and B. One sees that the finger pair A overlaps for the entire width of the SAW wavefront as drawn. The finger pair B, however, overlaps for only a 10% fraction of the wavefront region. It is therefore expected that the voltage detected for each transducer will correspond to coefficients $C_A = 1.0$ and $C_B = 0.1$. SAW devices become an easy platform for making inexpensive filters. The calculated coefficient values are incorporated directly into the mask used to fabricate the IDT pattern. For example, to implement some specific transfer function $H_1(\omega)$ [assuming that $H_1(\omega)$ is suitably band-limited], one can expand $H_1(\omega)$ in a Fourier series to generate a series as in Eq. (12). In particular, if the impulse response of the filter is $h_1(t)$ then the coefficients $\{C_n\}$ are given by

$$C_n = h_1(nT_d) \tag{13}$$

where T_d is now chosen to provide the correct sampling interval for reconstruction of $h_1(t)$. When we look at the SAW tapped delay line under a microscope, we actually see a spatially sampled replica of the impulse response in the IDT finger pair overlap. Any of the well-known techniques for digital filter design can be applied to designing SAW filter masks using apodization. The production of finite impulse response or transversal filters is readily achievable.

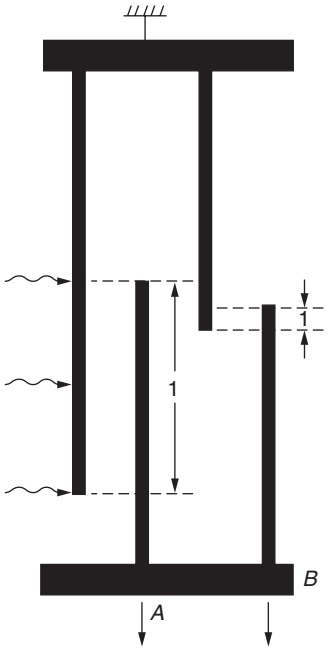


Figure 15. Overlap of the IDT finger pairs determines the transversal filter coefficients [7].

Now let's examine the frequency response of the basic IDT arrangement using the idea of a tapped delay line with N fingers. In the simple case all the IDT finger pairs are the same, which corresponds to all $C_n = 1$ using the notation given above. The spatially sampled impulse response is therefore a rectangular pulse. The frequency-domain transfer function for this arrangement is

$$H(\omega) = \sum_{n=0}^{N-1} e^{-j\omega T_d} \qquad \text{where} \quad T_d = \frac{1}{f_0} \tag{14}$$

and f_0 is the center frequency of the filter design. In the vicinity of f_0, the magnitude of the transfer function (14) is approximately

$$|H(\omega)| = N\frac{\sin(N\pi\,\Delta f/f_0)}{N\pi\,\Delta f/f_0} \qquad \text{where} \quad \Delta f = f - f0$$

$$\text{and } f = \frac{\omega}{2\pi} \tag{15}$$

This is the familiar sinc function response centered around f_0, which has bandwidth that is inversely proportional to N. N in this case is the spatial representation of time; a larger N means a longer impulse or time response for the filter. Thus the bandwidth of the filter shrinks as the number of IDT pairs is increased. That the frequency response is the Fourier transform of the aperture or apodization is a very satisfying and useful result for understanding SAW filters. Figure 16 summarizes the Fourier transform behavior of the SAW filter.

When selecting SAW filter bandwidth, one must also consider the device insertion loss. Figure 17 shows the fractional bandwidth versus minimum theoretical insertion loss of a SAW delay line using different materials. The minimum insertion loss is 6 dB and increases as the fractional bandwidth of the filter is increased. (The plot assumes that the IDTs are bidirectional.) Using unidirectional transducers, it is possible to build a SAW filter with 0 dB insertion loss. The optimum fractional bandwidth is represented by the expression

$$\left(\frac{\Delta f}{f}\right)_{\text{opt}} = \frac{1}{N_{\text{opt}}} = \frac{2K}{\pi^{1/2}} \tag{16}$$

where N_{opt} is the optimum number of finger pairs and K is electromechanical coupling coefficient.

The apodized IDT (see Figs. 18–20) in its simplest form has the problem of increased diffraction loss. This is due to variations in electrode overlap. When the electrodes overlap, only a small amount a large nonmetallized area results. As the SAW passes under the apodized electrodes, it sees a varying, parasitic acoustic grating. The result is wavefront distortion and added diffraction losses. An improved IDT, shown in Fig. 18, corrects this problem by adding floating electrodes. The floating electrodes maintain a consistent grating structure.

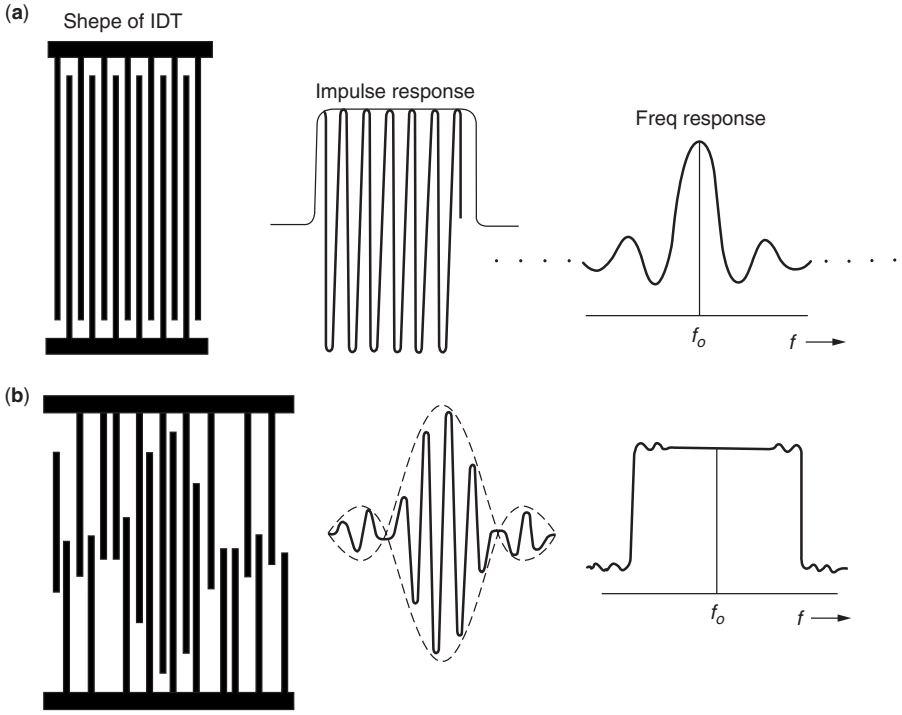


Figure 16. Two IDT apodizations and there corresponding filter characteristics [7].

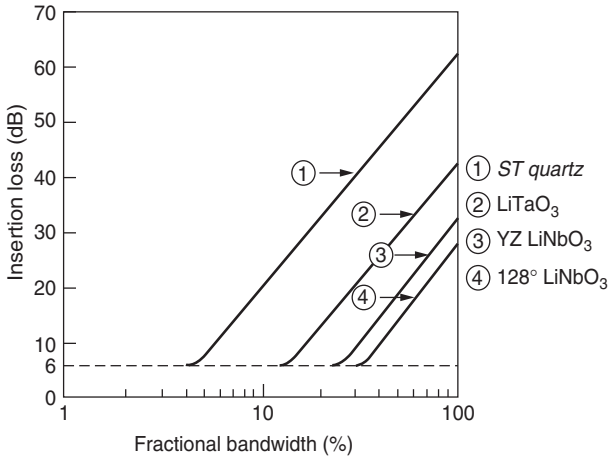


Figure 17. Minimum theoretical insertion loss versus fractional bandwidth for different piezoelectric substrates used for SAW devices.

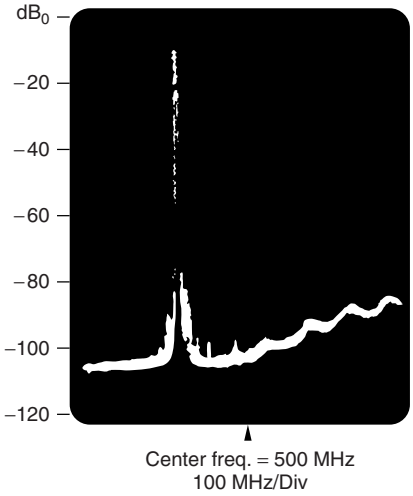


Figure 19. Typical response of a SAW filter using apodization [11].

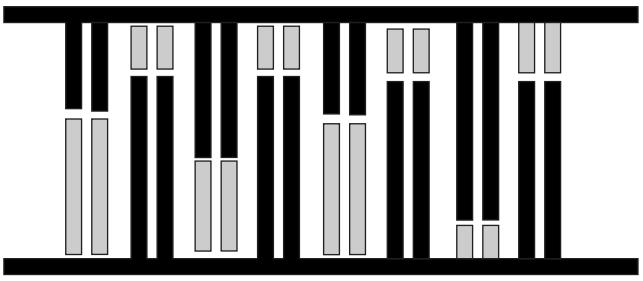


Figure 18. Improved IDT for transversal filter applications. This structure uses split-finger IDT to reduce triple-transit echo, and floating electrodes (gray stripes) to reduce the diffraction effects of apodized electrodes.

3.4. Multistrip Couplers

The *multistrip coupler* (MSC) is a useful structure for creating unidirectional IDTs and for suppressing undesirable bulk modes in SAW filters. In this section, we shall develop a simple understanding of the multistrip coupler and discuss some of its applications. The multistrip coupler is a set of parallel metal fingers that are not electrically connected to each other. To appreciate the operation of these fingers, we must first consider the two-dimensional wavefront of the propagating surface acoustic wave.

3.4.1. Simple Theory. A typical multistrip coupler is shown in Fig. 21. The device is comprised of two pairs of

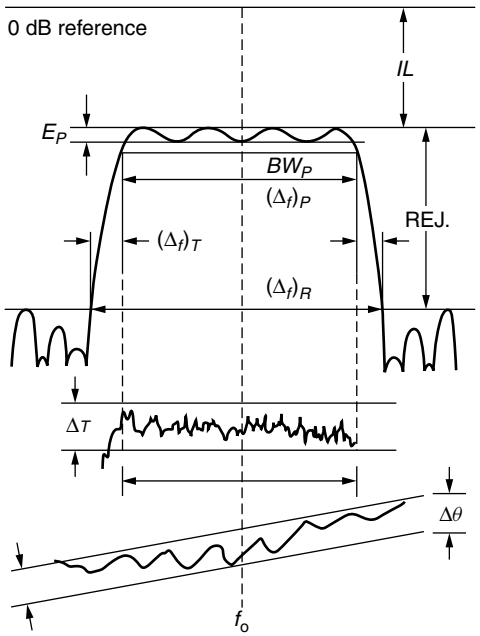


Figure 20. SAW frequency response parameters [7].

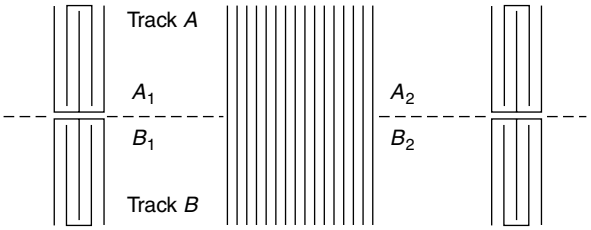


Figure 21. Multistrip directional coupler with IDTs at the input and output ports [8].

IDTs, labeled A_1, A_2, and B_1, B_2. The normal path for SAW would be from A_1 to A_2, and from B_1 to B_2, thus the device has two acoustic paths or tracks. We shall call these tracks A and B. The multistrip coupler is formed by the parallel rows of fingers that cross the device, extending over both track A and track B. With the addition of the multistrip coupler, the acoustic path is altered such that if the IDT A_1 is excited, then all the acoustic energy is transferred to track B and appears at IDT B_2. The output at A_2 will be zero if the coupler is properly designed. If only IDT B_1 is excited, then the SAW energy is transferred to track A and appears at IDT A_2. At first glance, it might seem impossible for the mechanical energy of the SAW to be transferred from one track to another simply by virtue of the parallel metal fingers. However, it is important to recognize that the piezoelectric substrate produces an electric field and thereby voltages on the metal fingers as the acoustic wave passes under them. This provides a means of coupling a common electric field between the two tracks. The electric field, in turn, produces a surface acoustic wave in the previously unexcited acoustic track. By using a multitude of parallel fingers, this passively generated SAW can be made quite strong.

The presence of metal fingers in both of the tracks on the piezoelectric substrate leads to coupling between the two tracks. This provides a means for energy transfer for a long multistrip coupler (MSC). The operation of the MSC and the coupling effect can be understood by considering Fig. 22. In this figure the SAW wavefront generated by the transducer in track A has been separated into symmetric and antisymmetric parts, labeled s and a, respectively. The wavefront of the symmetric part extends uniformly over both tracks and has a magnitude of $A/2$, where A is the amplitude of the total SAW wavefront. The antisymmetric part has the same $A/2$ magnitude and extends over both tracks; however, in track B the amplitude is 180° out of phase with respect to the amplitude in track A. The amplitude in track A has the same phase as the symmetric component. Without the metal fingers, if the symmetric and antisymmetric parts are combined, then the SAW cancels in track B and the full amplitude appears in track A.

While both the symmetric and antisymmetric parts propagate under the MSC, we note that there is an important difference between the two. The symmetric part of the SAW induces voltages on the fingers that are the same in both tracks. In contrast, the antisymmetric part of the SAW induces voltages that are of opposite phase on either ends of the metal fingers. Because the couplers are unable to support current flow parallel to the fingers, the antisymmetric field is essentially unaffected by the presence of the metal contact. In other words, the antisymmetric field does not "see" the metal contacts and propagates with the unstiffened velocity. The symmetric part of the wave will travel with the stiffened velocity. Well, actually because of the gaps between the metal fingers the symmetric and antisymmetric wavefronts will not be exactly equal to the stiffened and unstiffened velocities. The purely stiffened and unstiffened velocities correspond to the completely metallized and bare substrates respectively.

Using these arguments it can be shown that complete power transfer from track A to track B can take place for an MSC of length L_T

$$L_T = \frac{\lambda}{K^2} \tag{17}$$

where the coupling coefficient $K^2 \sim 2\Delta v/v$. If d is the MSC repeat distance, then the number of fingers N_T needed for

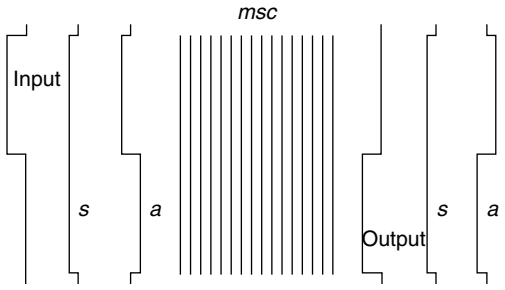


Figure 22. The input and output field distributions of a MSC separated into a symmetric (s) and antisymmetric (a) modes; the diagram shows 100% coupling from track A to track B [8].

a complete transfer is

$$N_T = \frac{\lambda}{K^2 d} \tag{18}$$

For a length x of the MSC, the power in each of the tracks is given by

$$P1 \sim \cos^2\left(\frac{\pi x}{L_T}\right) \tag{19a}$$

$$P2 \sim \sin^2\left(\frac{\pi x}{L_T}\right) \tag{19b}$$

The following is a brief list of possible applications for multistrip couplers:

- Coupler
- Bulk wave suppressor
- Coupling between different substrates
- Aperture transformations
- Precise attenuator — phase correction by offset
- Delay-line tap
- Beamwidth compressor
- Magic tee
- Beam redirection
- Reflector
- Unidirectional transducer
- Reflecting track changer
- Echo trap
- Better filter design
- Multiplexing
- Strip coupled amplifier and convolver
- Compressed convolver

Details of the applications can be found in the literature [7,8].

3.5. SAW Oscillators and Resonators

There are two distinct ways in which a SAW device can be used as the high-Q resonator circuit in an oscillator. The first case is shown in Fig. 23, where a SAW delay line is used in the feedback path around an amplifier. The second case is a SAW resonator structure, a planar cavity formed by two grating reflectors, which results in the high-Q element. The SAW-based oscillator has many advantages over bulk wave quartz oscillators in the frequency range beyond 100 MHz up to ~5 GHz. The operating frequency of bulk wave devices is based on crystal thickness, and for these high-frequency applications the required thickness are too thin to fabricate reliably. This means that for higher-frequency devices, overtones of the bulk resonator's fundamental frequency must be generated, which, in turn, requires multipliers and associated filters. The SAW device does not require these multiplier/filter combinations, thereby providing great advantage in size, weight, power, and cost.

The delay-line oscillator is a phase shift oscillator that operates at a frequency of

$$f_n \approx n\frac{v}{L} \tag{20}$$

where n is an integer. Of course, in order to oscillate, the amplifier gain must compensate for the delay-line losses, and the phase shift around the feedback path must be multiples of 2π. A particular frequency can be selected by designing some filter properties into the IDTs used to constitute the delay line.

In SAW resonators, the acoustic wave is trapped in a planar cavity formed by two acoustic grating reflectors. (In a conventional bulk wave resonator, the acoustic wave is confined by the two parallel surfaces of the crystal. The acoustic impedance mismatch between the air and crystal form a perfect reflector.) As mentioned in Section 2.1 with respect to acoustic gratings, the SAW decomposes into reflected longitudinal and shear waves when incident on

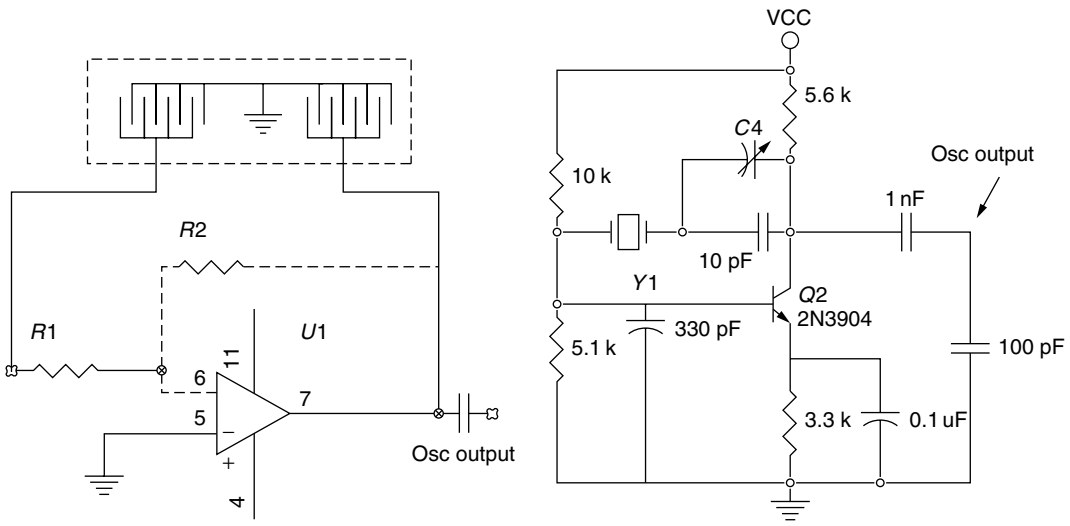


Figure 23. (**a**) An oscillator using a SAW delay line to produce the required phase shift; (**b**) discrete Pierce oscillator using a SAW-based impedance element Y1.

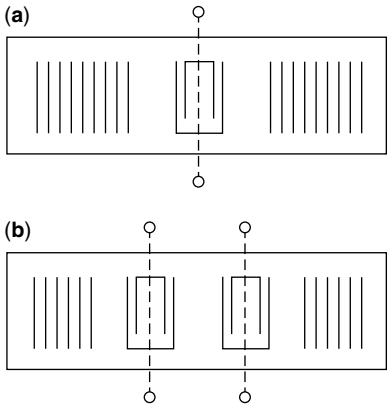


Figure 24. (**a**) Typical one-port SAW resonator; (**b**) two-port SAW resonator.

an abrupt surface discontinuity. The acoustic grating is made up of a large number of small, periodic surface perturbations.

The typical schematic for one-port and two-port SAW resonators is shown in Fig. 24. In a two-port resonator separate IDTs are used for generation and detection of the surface waves. The number of reflectors in the grating strips is typically in the hundreds. The performance of a resonator can be characterized almost entirely in terms of its reflectors. The Q factor of the resonant cavity with no propagation loss can be written as

$$Q = \frac{2\pi l}{\lambda(l - |r_f|^2)} \tag{21}$$

where l is the effective cavity length and $|r_f|$ is the amplitude reflection factor. The energy lost in the reflector due to mode conversion or absorption reduces the reflection factor and therefore the Q. The cavity length, l, is the sum of the separation between the gratings and the penetration of the SAW into the grating regions. The penetration of the SAW into the grating regions again depends on the design of the gratings. Oscillators with a Q of 30,000 have been reported using SAW resonators, compared to a Q of only a few thousand for oscillators based on delay lines. Using resonators, an oscillator can be made with stability and noise performance as good as in devices made from quartz overtone crystals, and with higher operating frequencies. For example, Rayleigh wave resonator oscillators on ST-cut quartz have a typical noise floor of -176 dBc/Hz with long-term stability on the order of 1 ppm per year [2]. For this reason, SAW resonators find use in UHF and VHF oscillators and wireless filters.

SAW resonators are a building block for filters operating in the 0.9–5-GHz range.

3.6. Convolvers

The nonlinear interaction of surface acoustic waves in materials can be used to make convolvers for signal processing. The nonlinearity of the material can be produced by either large-amplitude acoustic signals or by acoustoelectric interactions. Acousto-electric SAW convolvers can be configured in three different ways: separate medium, combined medium, and strip-coupled. What this classifications mean and how each operates will be discussed in the following section.

Let's begin with convolvers based on acoustic nonlinearity. The large amplitude acoustic nonlinearity can be considered the breakdown of Hooke's law within the material. One generally uses a piezoelectric substrate as the nonlinear acoustic material. This has two advantages: (1) piezoelectric nonlinearity is less than elastic stress–strain nonlinearity and (2) more importantly, the second harmonic of the strain wave has an associated electric field that is easy to detect and integrate by measuring the total voltage across (or current flowing through) the interaction region.

Convolvers based on acoustic nonlinearity have been demonstrated at different frequencies. They tend to have large inherent loss because the nonlinear coupling coefficient, K, is small. The dynamic range of these convolvers is also limited by failure of the elastic media. Higher power levels must be achieved in order for these devices to work well. Multistrip couplers, acoustic waveguides, or curved transducers are often used to boost power levels. Examples of each type of device structure are shown in Figs. 25–27.

Acoustoelectric phenomena produce a different type of nonlinearity. This nonlinearity is produced by the interaction of a SAW-generated piezoelectric field with the space charge arising from the displacement of free carriers in a semiconductor. The current density within the semiconductor is proportional to the product of the free-carrier density and the piezoelectric field. The carrier

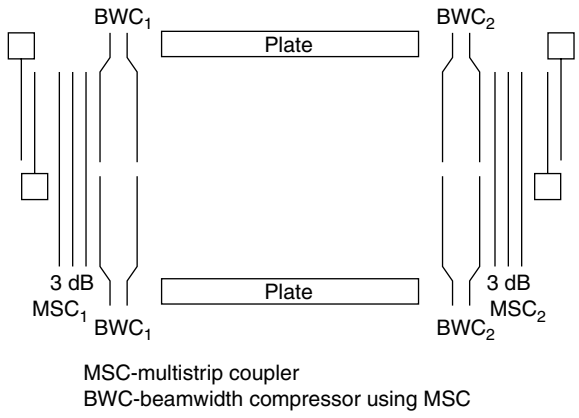


Figure 25. Diagram of a convolver using multistrip coupler compression, acoustic waveguides.

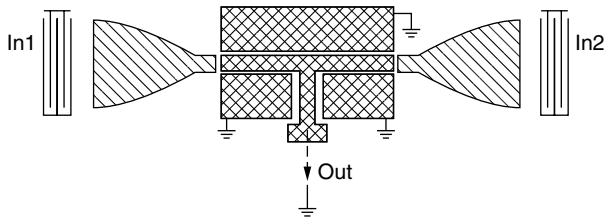


Figure 26. Diagram of a convolver using horn waveguide to increase power density.

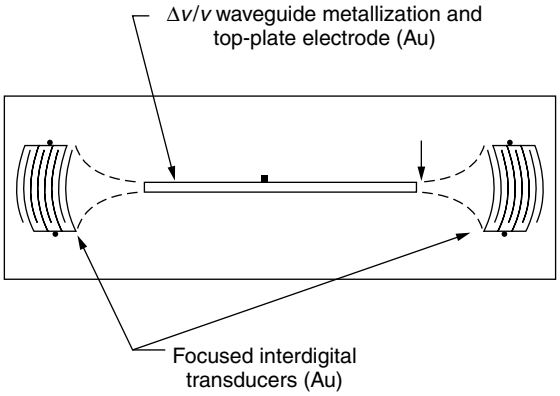


Figure 27. A convolver using curved transducers and a waveguide.

density is itself a function of the applied electric field, since the carriers are rearranged to produce a space charge. Thus, the normal component of the electric field at the semiconductor surface creates a voltage that is proportional to the square of the field. Two possibilities exist for creating an acoustoelectric convolver. One example is to use a piezoelectric semiconductor, such as GaAs, to generate the SAW — this is known as a *combined medium structure*. A second possibility is referred to as the *separate medium structure*, where a semiconductor is placed in close proximity to a piezoelectric substrate such as lithium niobate (Fig. 28). One can also evaporate the semiconductor directly over the piezoelectric substrate. A final variation is to evaporate a piezoelectric film such as ZnO over the semiconductor substrate.

3.7. Summary of SAW Devices

As late as the early 1980s SAW devices were a scientific curiosity; now they have matured to the point where they are routinely used in communication and signal processing applications. SAW filters can be divided into five categories:

1. Tapped delay line with fixed taps
2. Tapped delay line with programmable taps
3. Resonators
4. Convolvers
5. Transform-domain processors

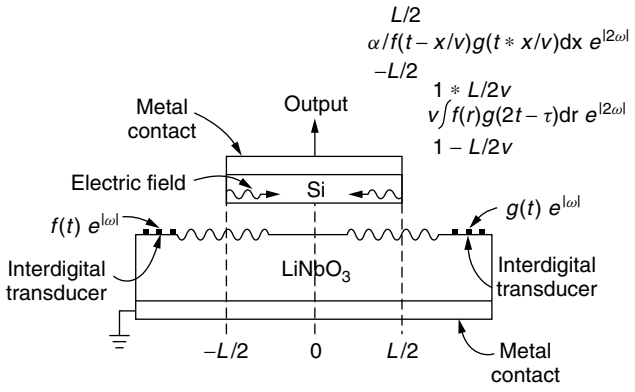


Figure 28. Diagram of a convolver made using a separated medium structure [12].

Let's take a moment to summarize the key feature of each device category. A tapped delay line with fixed-weight taps produces a filter with a very steep filter transition band; that is, when compared to an LC filter, the SAW delay-line filter produces larger stopband rejection closer to the corner frequency. A tapped delay line with programmable taps is used as programmable, matched, or with some added complexity as Widrow–Hoff LMS processor. Resonators can be used in oscillator circuits or cascaded to build narrow passband filters. A convolver is a three-terminal device, which utilizes its nonlinear response to perform programmable convolution of signals. Finally, transform-domain processors are subsystems used in real-time signal processing. For example, a real-time Fourier transform processor can be created using a SAW chirp filter.

4. SAW FILTER DESIGN

4.1. Finite-Impulse-Response Filters

In the previous section, we introduced the notion of a finite-impulse-response (FIR) filter based on apodized IDTs. In reviewing the loss mechanism in delay line devices, improvements were made to the basic IDT to arrive at the structure shown at Fig. 18. A high-performance IDT uses split electrodes and floating electrodes to compensate for nonideal behavior. There are, in addition to insertion loss and fractional bandwidth, other parameters that one is concerned with in FIR filter design. These other parameters are listed in the table in Fig. 19, along with typical values achieved using SAW filters. The overall design parameters from the table in Fig. 19 for a filter are shown in Fig. 20 [7]. These include insertion loss, pass bandwidth, rejection bandwidth, transition bandwidth, fractional, bandwidth, rejection, shape factor, amplitude ripple, phase ripple, and group delay. Here again, the overall design of a filter becomes quite complicated and beyond the scope of this short article. We do wish, however, to highlight the performance that one can achieve by using SAW for FIR filters. One such high-performance filter is shown in the top trace in Fig. 19. The overall rejection of roughly 90 dB demonstrates the advantages of SAW devices over conventional *LC* technology.

4.2. SAW Resonator Filters

For many designs operating beyond 100 MHz, filters based on SAW resonators have been more advantageous. Indeed, for operating frequencies beyond 1 GHz, resonator structures have proved to be the best path for SAW designers. These designs provide low insertion loss, flat passband, steep transition, and excellent close-in rejection. Filters operating at 2 and 5 GHz based on LSAW impedance elements have been demonstrated [15]. The resonator structures for these filters have been etched into lithium tantalate substrates using electron-beam

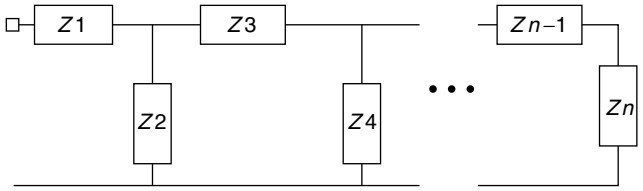


Figure 29. General filter form for a Cauer-type ladder [14].

lithography. Feature sizes on the order of 200 nm are necessary to create 5-GHz acoustic gratings. The SAW impedance elements are designed into ladder networks, as would be used for conventional LC filter design. Let's review ladder networks.

The general Cauer form of a ladder network is shown in Fig. 29 [14]. The ladder is composed of series and shunt impedances Z_1 through Z_n. The process of synthesizing a filter response involves partial fraction expansion of the desired transfer function. Each term in the expanding transfer function corresponds to the impedance at each ladder rung. In practice, predetermined filter prototypes are used to create ladder filters. These prototypes include the maximally flat Butterworth, the equiripple Chebyshev, and the linear phase Bessel filter responses. The designer can look up the required filter coefficients (L and C values) for the type and order of the filter. Once a lowpass prototype design is established, the filter can be translated to the bandpass of interest by replacing the Ls and Cs with LC combinations. As shown if Fig. 30, standard equations are available to scale a design to the desired operating point.

The equivalent circuit model for a periodic SAW grating shown in Fig. 33 shows how it, too, forms a ladder network. All the SAW ladder elements, Z_1, Z_2, Z_{1m}, Z_{2m}, are functions of the electrode width, spacing, and acoustic velocity. With the aid of CAD, one can use the relations shown in Fig. 33 to synthesize impedances. Each SAW resonator can, in turn, be generalized to create a ladder network as shown in Fig. 31. The SAW impedance element filters (IEFs) are combined into ladder networks. The results are summarized in the literature [15,17].

A high-performance 900-MHz filter design, based on LSAW resonators, is shown in Fig. 32. The design is based on a two-port SAW resonator structure on 36° $LiTaO_3$. Two outer IDTs drive an interior and two exterior grating structures. The entire resonator is then duplicated in an image connection design technique [17] (Note how the filter image has horizontal and vertical symmetry). The

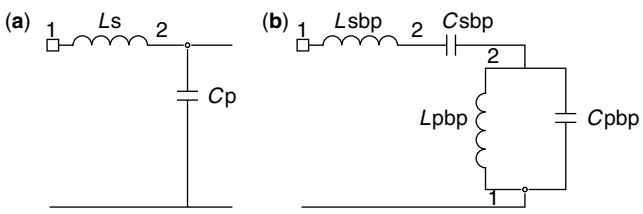


Figure 30. Designing ladder networks is often performed by first synthesizing a lowpass filter using LC pair (**a**), and then scaling the response using a bandpass transformation (**b**) [14].

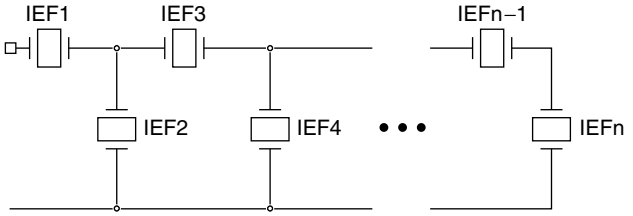


Figure 31. Ladder network composed of SAW impedance element filters (IEF), for synthesizing filters in the gigahertz range [15].

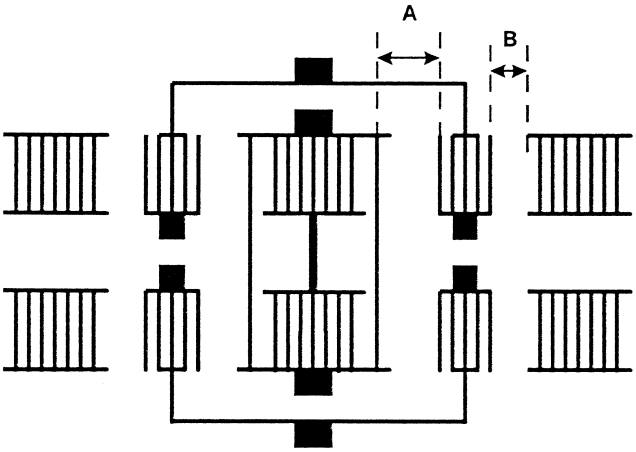


Figure 32. Low-loss, sharp-cutoff 900-MHz ISM band SAW filter design [17]. Filter structure utilizes image connection design technique, combined with resonator structures to produce excellent performance. Critical design spacings, A and B, highlighted in diagram.

design procedure involves: selecting the grating period, then adjusting the IDT to grating spacings (labeled A and B in Fig. 32). Each spacing is adjusted to minimize ripple, and maximize the transition band. The overall frequency response is the combination of the IDTs by themselves, and the grating response. So the response of each is tuned to create the best overall response. The final filter specifications are 903 MHz center frequency, 5.6 MHz transition bandwidth, 2 MHz passband, 2.4 dB insertion loss, 65 dB sidelobe suppression, and a die size of 3 mm^2.

Other designs are shown in Figs. 33–35.

4.3. SAW CAD Design

Several computer-aided design tools exist for the development of SAW filters. One such example is SAWCAD-PC available online from the University of Central Florida (*http://www.ucf.edu*).

5. SAW FILTER APPLICATIONS IN TELECOMMUNICATIONS

The following is brief list of telecommunication applications where SAW filters are often employed:

1. Nyquist filters for digital radio
2. IF filters for mobile and wireless transceivers

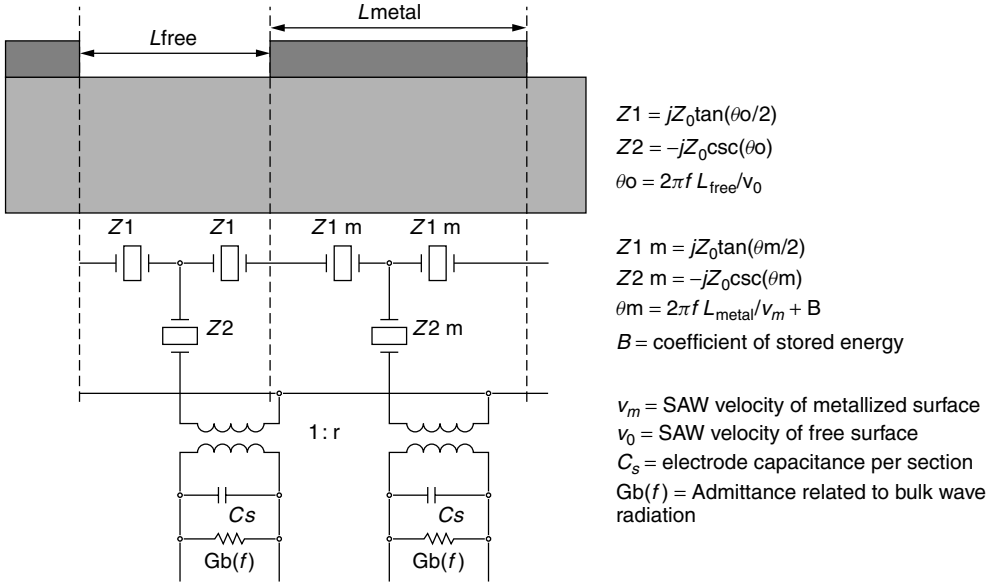


Figure 33. Improved equivalent-circuit model for IDT design that considers secondary effects of stored energy and bulk waves [17].

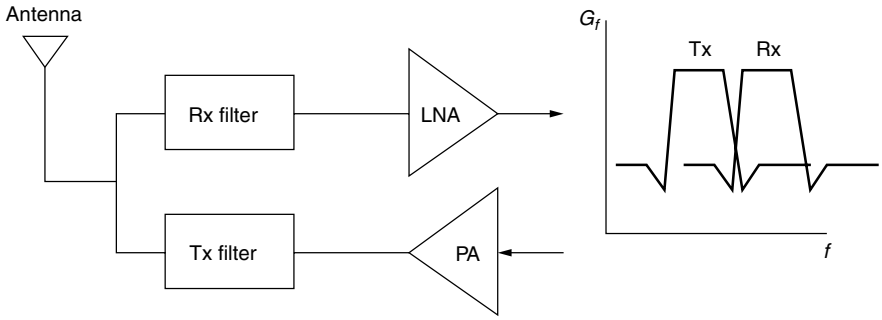


Figure 34. General block diagram for a antenna duplexer in a wireless system. Two SAW- or LSAW-based filters are used to allow transmit and receive frequencies to share a single antenna.

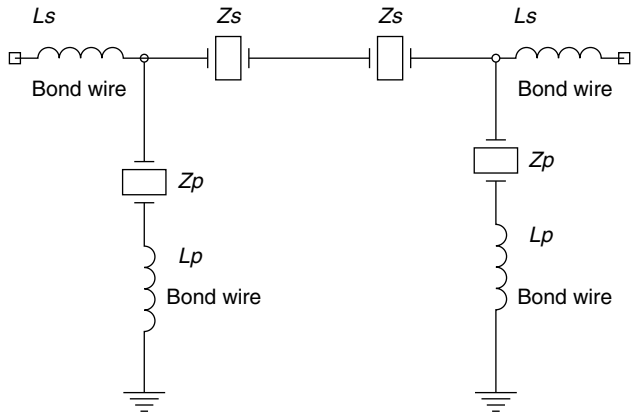


Figure 35. A simple ladder-type filter structure used for designed Tx and Rx filters. Each element Z_s and Z_p is made from a SAW resonator structure [16].

3. Antenna duplexers
4. Clock recovery circuit for optical fiber communication
5. Delay lines for path length equalizers
6. RF front-end filters and channelization in mobile communications
7. Precision fixed frequency and tunable oscillators
8. Resonant filters for automotive keyless entry, garage door transmitter circuit, and medical alert transmitter circuit.
9. Pseudonoise (PN)-coded tapped delay line
10. Convolvers and correlators for spread-spectrum communications
11. Programmable and fixed matched filters.

A partial list of commercially available SAW filters is shown in Table 1. The table includes applications, center frequency, bandwidth, and typical insertion loss for each device. In the following sections, some of these applications will described in greater detail.

5.1. SAW Antenna Duplexers

An antenna duplexer is an example of the use of SAW filters in modern wireless applications. A general diagram of an antenna duplexer system is shown in Fig. 29. The transmit channel (Tx) outputs signals from a power amplifier (PA) and conveys them to the antenna via a transmit filter. The receive channel (Rx) decodes incoming signals by first processing them through a receive filter (Rx) and then amplifying with

Table 1. Commercially Available SAW Filters for Telecommunication

Application	Center Frequency (MHz)	BW (MHz)	Fractional BW (%)	Insertion Loss (dB)
900-MHz cordless phone Tx	909.3	3	0.33	5
900-MHz cordless phone Rx	920	3	0.33	5
AMPS/CDMA Rx	881.5	25	2.84	3.5
AMPS/CDMA Tx	836.5	35	4.18	1.8
CDMA IF	85.38	1.26	1.48	12.00
CDMA IF	130.38	1.23	0.94	7.50
CDMA IF	183.6	1.26	0.69	10.50
CDMA IF	210.38	1.26	0.60	9.00
CDMA IF	220.38	1.26	0.57	8.00
CDMA Base Tcvr Subsys(BTS)	70	9.4	13.43	28.00
CDMA BTS	150	1.18	0.79	25.00
Broadband access	1086	10	0.92	5.00
Broadband access	499.25	1	0.20	9.00
Broadband access	479.75	23	4.79	13.00
Broadband access	333	0.654	0.20	9.00
Cable TV tuner	1220	8	0.66	4.2
DCS	1842.5	75	4.07	3.6
EGSM	942.5	35	3.71	2.6
GSM	947.5	25	2.64	3
GSM	902.5	25	2.77	3
GSM IF	400	0.49	0.12	6.5
GSM BTS	71	0.16	0.23	9
GSM BTS	87	0.4	0.46	7
GSM BTS	170.6	0.3	0.18	9
PCS	1960	60	3.06	2.4
PCS	1880	60	3.19	2
W-CDMA	1960	60	3.06	2.1
W-CDMA IF	190	5.5	2.89	8
W-CDMA IF	380	5	1.32	9
W-LAN	900	25	2.78	6
Satellite IF	160	2.5	1.56	25.1
Satellite IF	160	4	2.50	24.8
GPS	1575.42	50	3.17	1.4
Wireless data	2441.75	83.5	3.42	5
Wireless data	770	17	2.21	7
Wireless data	570	17	2.98	17.2
Wireless data	374	17	4.55	10.5
Wireless data	240	1.3	0.54	11

an LNA. Thus the two channels are multiplexed by frequency division.

The filters used in these applications require low insertion loss and large stopband rejection, as well as small transition band. For these reasons, SAW filters are excellent candidate devices. Figure 30 shows a general schematic for a ladder-type filter. The parallel and series impedances, Z_p and Z_s, are created using SAW resonator (Fig. 24) structures. Additional inductance due to bond wires are also shown as part of the ladder network as they contribute to the performance at gigahertz-range operating frequencies.

5.2. Cellular Phone Transceiver Modules

Both analog (e.g., AMPS) and digital (e.g., GSM) cellular phone technologies employ several SAW filters and oscillators. As an example, Fig. 36 depicts an AMPS cellular phone handset. AMPS employs frequency-division multiple access, transmitting at 824–859 MHz and receiving at 869–894 MHz. The transceiver system employs six SAW devices. The antenna duplexer for the 800-MHz transmit and receive carriers is built from two SAW filters labeled Rx1 and Tx1. Additional carrier filtering is provided by SAW Rx2 and Tx2. The IF stage and VCO and PLL synthesizer also utilize SAW devices.

The SAW devices used in the receiver, Rx1 and Rx2, are typically of the LSAW resonator type. They are required to suppress harmonics, reject image frequency noise, and suppress switching noise from the power supplies and power amplifiers. The transmit filter, Tx1 and Tx2, is also typically an LSAW device, especially as it is required to handle up to 1 W of transmit power. The IF stage SAW filter is required to be very selective (30-kHz spacing) and very stable. Therefore this filter is often a waveguide-coupled resonator built on ST-cut quartz. The VCO and

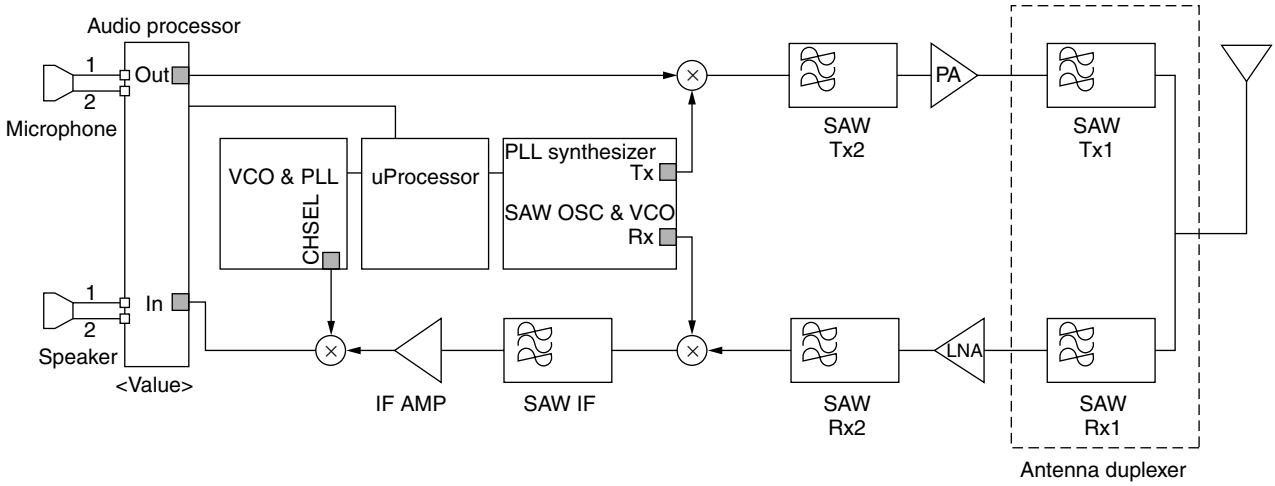


Figure 36. Simplified block diagram of AMPS cellular phone handset depicting the use of up to six SAW devices in a 800-MHz dual heterodyne transceiver circuit [2].

PLL SAW devices are often dual-mode resonator filters or wideband delay lines. These same filter attributes are required in digital phone systems as well, the operating frequencies are simply altered accordingly.

5.3. Nyquist Filters for Digital Radio Link

Digital communication systems require a Nyquist filter for bandlimiting and to prevent intersymbol interference (ISI) [24]. The details of Nyquist filters are developed in the many texts on the subject of digital communications. Stated briefly, the Nyquist criterion for a zero ISI filter requires that frequency-domain sum of the entire filter spectrum be a constant. Or to put it another way, that the digital pulse spectrum is band-limited. One such filter that meets this criterion is the raised-cosine (RC) filter. Digital microwave radios often employ raised cosine filters in both the transmit and receive sides. In this way, the received matched filter response is raise-cosine-squared, which also can be shown to satisfy the Nyquist criterion. SAW filters are used to implement raised cosine and similar pulseshaping Nyquist filters for the North American common microwave carrier bands (4, 6, 8, and 11 GHz) [2].

5.4. Convolvers as Correlators for Spread-Spectrum Receivers

SAW convolvers are used in spread-spectrum communication links because of their small size, large processing gain, and broad bandwidth [2]. Typically a SAW convolver is implemented at the IF rather than RF stage. Actually, the convolver is used to perform autocorrelation of pseudonoise spreading codes. One port of the convolver (see Figs. 25–28) receives the IF coded sequences, while the second port receives the time-reversed reference code. The autocorrelated output is obtained at the third port. The interaction area, such as the semiconductor strip in a separated media structure, must be long enough to hold an entire bit sequence. One of the inputs, most likely the reference sequence, must be strong enough to generate the nonlinear action within the SAW substrate, or separated strip.

SAW convolver operating frequencies include the 900-MHz band, the 2-GHz spread-spectrum band, and the Japanese license-free spread-spectrum band below 322 MHz.

5.5. Clock Recovery Circuit for Optical Communications Link

SAW oscillators and filters are used in clock recovery circuits for optical communication links. The key advantages of SAW again lie in their small size, stability, and low jitter performance. SAW filters are used at center frequencies that correspond to the ATM/SONET/SDH clock frequencies. Example operating frequencies include 155.52, 622.08, and 2488.32 MHz. SAW oscillators are often used in PLL circuits as local oscillator references and VCO references.

These are some of the applications of SAW devices. There are others, including Global Positioning System (GPS) receivers, pagers, and ID tags — almost every telecommunication application. Additional applications are available in the literature [4,5,18,19].

BIOGRAPHY

Robert J. Filkins received the B.S. and M.S. degrees in electrical engineering in 1990 and 1997, respectively, from Rensselaer Polytechnic Institute, Troy, New York. He joined General Electric in 1993 as an Electronics Engineer, developing low-noise electronics systems for ultrasonic inspection of turbine components. Since 1995, he has been with the General Electric Global Research Laboratory, working on low-noise amplifier systems for ultrasonic, eddy-current, and laser ultrasonic inspection of components, as well as wireless and photonic communication components. Mr. Filkins currently holds six patents in the areas of electronics and nondestructive inspection systems. His areas of interest include optoelectronic devices

and materials, laser generation of SAW, and microwave design. He is currently working toward a Ph.D. degree at Rensselaer Polytechnic Institute.

BIBLIOGRAPHY

1. P. K. Das, *Optical Signal Processing Fundamentals*, Springer-Verlag, New York, 1991.
2. B. A. Auld, *Acoustic Fields and Waves in Solids*, Vol. II, 2nd ed., Krieger Publishing, 1990.
3. W. P. Mason, *Physical Acoustics*, Vol. I, Part A, Academic Press, New York, 1964.
4. C. K. Campbell, *Surface Acoustic Wave Devices for Mobile and Wireless Communications*, Academic Press, New York, 1998.
5. E. A. Ash et al., in A. A. Oliner, ed., *Acoustic Surface Waves*, Topics in Applied Physics Vol. 24, Springer-Verlag, New York, 1978.
6. J. deKlerk, *Physics Today* **25**: 32–39 (Nov. 1972).
7. L. B. Milstein and P. K. Das, *IEEE Commun. Mag.* **17**: 25–33 (1979).
8. F. G. Marshall, C. O. Newton, and E. G. S. Paige, *IEEE Trans.* **SU-20**: 124–134 (1973).
9. W. R. Smith et al., *IEEE Trans.* **MTT-17**: 865–873 (1969).
10. C. S. Hartmann, W. J. Jones, and H. Vollers, *IEEE Trans.* **SU-19**: 378–384 (1972).
11. R. M. Hays and C. S. Hartmann, *Proc. IEEE* **64**: 652–671 (1976).
12. A. Chatterjee, P. K. Das, and L. B. Milstein, *IEEE Trans.* **SU-32**: 745–759 (1985).
13. T. Martin, Low sidelobe IMCON pulse compression, *Ultrasonic Symp. Proc.* (J. deKlerk, ed.), IEEE, New York, 1976.
14. A. Budak, *Passive and Active Network Analysis*, Houghton-Miflin, Boston, 1973, Chap. 4.
15. S. Lehtonen et al., SAW Impedance Element Filters towards 5 GHz, *IEEE Ultrasonics Symp.*, 1998, pp. 369–372.
16. T. A. Fjeldy and C. Ruppel, eds., *Advances in Surface Acoustic Wave Technology, Systems, and Applications*, Vol. 1, World Scientific, 2000.
17. Y. Fujita et al., Low loss and high rejection RF SAW filter using improved design techniques, *IEEE Ultrasonics Symp.*, 1998, pp. 399–402.
18. G. Kino, *Acoustic Waves: Devices, Imaging, and Analog Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
19. D. P. Morgan, *Surface Wave Devices for Signal Processing*, Elsevier, New York, 1985.
20. R. C. Williamson and H. I. Smith, *IEEE Trans.* **SU-20**: 113–123 (1973).
21. P. K. Das and W. C. Wang, *Ultrasonics Symp. Proc.*, IEEE, New York, 1972, p. 316.
22. E. D. Palik, *Handbook of Optical Constants of Solids*, Academic Press, New York, 1997.
23. A. J. Slobodnik, E. D. Conway, and R. T. Delmonico, *Microwave Acoustics Handbook,* Vol. IA*, Surface Wave Velocities*, AFCRL-TR-73-0597, Air Force Cambridge Research Labs, Cambridge, MA.
24. J. G. Proakis, *Digital Communications*, 2nd ed., McGraw Hill, New York, 1989, pp. 532–536.

SURVIVABLE OPTICAL INTERNET

Hussein T. Mouftah
Pin-Han Ho
Queen's University at Kingston
Kingston, Ontario, Canada

1. INTRODUCTION

The Internet has revolutionized the computer and communications world like nothing before, which has changed the lifestyles of human beings by providing at once a worldwide broadcasting capability, a mechanism for information dissemination, and a medium for collaboration and interaction between individuals and their computers without regard for geographic location. As the importance of the Internet is overwhelming, the strategies of constructing the infrastructure are becoming more critical. The design of the Internet architecture with a suite of control and management strategies, by which the most efficient, scalable, and robust deployment can be achieved, has been a focus of researches in industry and academia since the late 1980s.

The Internet is a core network, or a wide-area network (WAN), which is usually across countries, or continents, for the purpose of interconnecting hundreds or even thousands of small-sized networks, such as metropolitan-area networks (MANs) and local area networks (LANs). The small-sized networks are also called *access networks*, which upload or download data flows to or from the core network through a traffic grooming mechanism.

The proposal of constructing networks on fiberoptics has come up with the progress in the photonic electronics since the 1960s or so, however, most of which were focused on the access networks before 1990, such as the fiber distributed data interface (FDDI) and synchronous optical network (SONET) ring networks. With the advances in the optical technologies, most notably *dense wavelength-division-multiplexed* (DWDM) transmission, the amount of raw bandwidth available on fiberoptic links has increased by several orders of magnitude, in which a single fiber cut may influence a huge amount of data traffic in transmission. As a result, the idea of using optical fibers to build the infrastructure of the Intern along with *survivability* issues emerged in the early 1990s. The adoption of a survivable *optical Internet* is for the purpose of reducing the network cost and enabling versatile multimedia applications by a simplification of administration efforts and a guarantee of service continuity (or system reliability) during any network failure.

This article presents the state-of-the-art progress of the survivable *optical Internet*, and provides thorough overviews on a variety of aspects of technologies and proposals for the protection and restoration mechanisms. To provide enough background knowledge, the following subsections describe the evolution of data communication networks, as well as the definition of *survivability* as the introduction of this article.

1.1. Evolution of Data Communication Networks

Since the late 1980s, router-based cores have been largely adopted, where traffic engineering (TE) (i.e., the manipulation of traffic flow to improve network performance) was achieved by simply manipulating routing metrics and link states of interior gateway protocol (IGP), such as open shortest path first (OSPF) and intermediate system–intermediate system (IS-IS). As the increase in Internet traffic of new applications such as multimedia services, a number of limitations in terms of bandwidth provisioning and TE manipulation came up:

1. Software-based routers had the potential of becoming traffic bottlenecks under heavy load. Hardware-based IP switching apparatus (which is defined below), however, are vendor-specific and expensive.
2. Static metric manipulation was not scalable and usually took a trial-and-error approach rather than a scientific solution to an increasingly complex problem. Even after the addition of TE extensions with dynamic metrics such as maximum reservable bandwidth of links to the IGPs, the improvements remain very limited [1,2].

To overcome the problems of bandwidth limitation and TE requirements, application-specific integrated circuit (ASIC)-based switching devices had been developed since the early 1990s. As prices of such devices were getting cheaper, IP over asynchronous transfer mode (ATM) was largely adopted and provided benefits for the Internet service providers (ISPs) in the aspects of the high-speed interfaces, deterministic performance, and TE capability by manipulating permanent virtual circuitry (PVC) in the ATM core networks.

While the IP over ATM core networks mentioned above solved the problems in the IP routing cores for the ISPs, however, the expense paid for the complexity of the overlay model of the IP over ATM networks has motivated the development of various multilayer switching technologies, such as Toshiba's IP switching [3] and Cisco's tag switching [4].

Multiprotocol label switching (MPLS) is the latest step in the evolution of multilayer switching in the Internet. It is an Internet Engineering Task Force (IETF) standards-based approach built on the efforts of the various proprietary multilayer switching solutions. MPLS is composed of two distinct functional components: a control component and a forwarding component. By completely separating the control component from the forwarding component, each component can be independently developed and modified. The only requirement is that the control component continues to communicate with the forwarding component by managing the packet forwarding table. When packets arrive, the forwarding component searches the forwarding table maintained by the control component to make a routing decision for each packet. Specifically, the forwarding component examines information contained in the packet's header, searches the forwarding table for a match, and directs the packet from the input interface to the output interface across the system's switching fabric.

The forwarding component of MPLS is based on a label-swapping forwarding algorithm. This is the same algorithm used to forward data in ATM and frame relay switches [1,2,5]. An MPLS packet has a 32-bit header carried in front of the packet to identify a forwarding equivalence class (FEC) [5]. A label is analogous to a connection identifier, which encodes information from the network layer header, and maps traffic to a specific FEC. An FEC is a set of packets that are forwarded over the same path through a network even if their ultimate destinations are different.

1.2. Survivability Issues

A network is considered to be survivable if it can maintain service continuity to the end users during the occurrence of any failure by preplanned or real-time mechanisms of protection and restoration. Network *survivability* has become a critical issue as the prevalence of DWDM, by which a single fiber cut may interrupt huge amount of bandwidth in transmission. In general, Internet backbone networks are overbuilt in comparison to the average traffic volumes, in order to support fluctuations in traffic levels, and to stay ahead of traffic growth rates. With the underutilized capacity in networks, the most widely recognized strategy to perform protection service is to find protection resources that are physically disjoint (or diversely routed) from the working paths, over which the data flow could be switched to the protection paths during any failure of network elements along the working paths.

This article is organized as follows. Section 2 presents a framework of interest in this article for performing protection and restoration services in the *optical Internet*, and includes some background knowledge and important terminology for the subsequent discussions. Section 3 describes several classic approaches of achieving *survivability* with implementation issues. Section 4 investigates into the strategies newly reported in the literature. Section 5 summarizes this article.

2. FRAMEWORK FOR ACHIEVING SURVIVABILITY

This section focuses on the framework of protection and restoration for the *optical Internet*. Some of the most important background knowledge and concepts in this area are presented.

2.1. Background

Although the use of DWDM technology enables a fiber to accommodate a tremendous amount of data; it may also risk a serious data loss when a fault occurs (e.g., a fiber cut or a node failure), which could downgrade the service to the customers to the worst extent. To improve the *survivability*, the ISPs are required to equip the networks with protection and restoration schemes that can provide end-to-end guaranteed services to their customers according to the service-level agreements (SLAs).

Faults can be divided into four categories: path failure (PF), path degraded (PD), link failure (LF), and link degraded (LD) [6]. PD and LD are cases of loss of signal (LoS) in which the quality of the optical flow is

unacceptable by the terminating nodes of the *lightpath*. To cope with this kind of failure, Hahm et al. [7] suggested that a predetermined end-to-end path that is physically disjoint from the working path is desired, since a fault localization cannot be conducted with respect to LoS along the intermediate optical network elements in an all-optical network. Because the probability of occurrence of an LoS fault in the optical networks is reported to be as low as 1×10^{-7} [8], it rarely needs to be considered. Furthermore, an LoS fault is generally caused by a failed transmitter or wavelength converter, which could be overcome by switching the traffic flowing on the impaired channel to the spare wavelength(s) along the same physical conduit. Since this article focuses on *shared protection* schemes in which protection paths are physically disjoint from working paths, the protection and restoration of an LoS fault on a channel is not included in this paper. Gadiraju and Mouftah [9] have further discussed and analyzed this topic.

In the PF and LF cases, the continuity of a link or a path is damaged (e.g., a fiber cut). This kind of failure can be detected by a loss of light (LoL) detection performed at each network element so that fault localization [7] can be easily performed. The restoration mechanisms may have the protection path traversing the healthy NEs along the original working path as much as possible while circumventing from the failed NEs to save the restoration time and improve network resources. In this article, all nodes are assumed to be capable of detecting an LoL fault in the optical layer. Optical detectors residing in the optical amplifiers at the output ports of a node monitor the power levels in all outgoing fibers. An alarm mechanism is performed at the underlying optical layer to inform the upper control layer of a failure once a power-level abnormality is detected.

In this article, protection and restoration schemes are discussed and evaluated with the following criteria: *scalability*, *dynamicity*, *class of service*, *capacity efficiency*, and *restoration speed*. The *scalability* of a scheme is important in a sense that networks are required to be scalable to any expansion in the number of nodes or edges with the same computing power in the control center or in each node. The *scalability* ensures that the scheme can be applied to large networks in size and capacity. The *dynamicity* of a scheme determines whether the networks can deal with dynamic traffic that arrives at the networks one after the other without any prior knowledge, which makes the optimization of spare capacity a challenge. The networks with *class of service* can provide the end users with protection services with wider spectrum and finer granularity, and are the basis on which the ISPs charge their customers. The *capacity efficiency* is concerned with the cost of networks. To be capacity-efficient, the schemes should make the most use of every piece of spare capacity through optimization or heuristics. *Restoration speed* describes how fast a failure can be recovered after its occurrence. However, *restoration speed*, *capacity efficiency*, and *dynamicity* are tradeoffs in the design spectrum of network protection and restoration schemes most of the time.

Protection can be defined as the efforts made before any failure occurs, including the failure detection and localization. *Restoration*, on the other hand, is the reaction of the protocol and network apparatus toward the failure, including any signaling mechanisms and network reconfiguration used to recover from the failure. A *lightpath* is defined as a data path in an optical network that may traverse one or several nodes. A *working path* (or primary path) is defined as a *lightpath* that is selected for transmitting data during normal operation. A *protection path* (or spare path, secondary path) is the path used to protect a specific segment of working path(s).

Various types of protection and restoration mechanisms have been reported, and can be categorized as *dedicated* and *shared protection* in terms of whether the spare capacity can be used by more than one working paths. The *dedicated protection* can be either $1+1$ or $1:1$. The $1+1$ protection is characterized by having two disjoint paths transmitting the data flow simultaneously between sender and destination nodes, thus an ultrafast *restoration speed* is achieved. As a failure occurs, the receiver node does not have to switch over the traffic flow along the working path for keeping the service continuity. The $1:1$ protection, on the other hand, has a *dedicated protection* path for the working path without passing data traffic along the protection path. In other words, the network resources along the protection path is not configured, and can be used for the other traffic with a lower priority during normal operation (this kind of low-priority traffic is also called "best-effort" traffic). Once a failure occurs, the best-effort traffic has to yield the right of way, so that the traffic in the affected working path is switched over to the corresponding protection path.

The *shared protection* has versatile types of design originality, which can be categorized as path-based, link-based, and *short leap shared protection* (SLSP) according to the location where fault localization is performed. In the case of a number of working paths sharing a protection path, shared protection can be either $1:\mathrm{N}$ or $\mathrm{M}:\mathrm{N}$. In the $1:\mathrm{N}$ case, a protection path is shared by N physically disjoint working paths. In the $\mathrm{M}:\mathrm{N}$ case, M protection paths are shared by N working paths. The working paths sharing the same protection path are required to be physically disjoint because a protection path can afford a switchover of a single working path only at the same moment. This is also called a *shared risk link group* (SRLG) constraint, which will be discussed in later sections.

Opposite to the investigation into the relationship between paths in the networks, proposals were made to preplan or preconfigure spare capacity at the network planning stage according to the working capacity along each span (or edge). The planning-type schemes are used in networks with static traffic that is rarely changed as time goes by, and usually requires a time-consuming optimization process (e.g., integer linear programming or some flow theories).

2.2. Shared Risk Link Group (SRLG) Constraint

The *shared risk link group* (SRLG) [10] is defined as a group of working *lightpath*s that run the same risk of

service interruption by a single failure, which has the following characteristics:

$$\mathrm{SRLG}(P^m_{M,u}) = \bigcup_{M\cap Q\neq\varnothing} \bigcup_{k\in K} \{P^m_{Q,k}\},$$

$$P_1 \in \mathrm{SRLG}(P_2) \Leftrightarrow P_2 \in \mathrm{SRLG}(P_1)$$

where u and k are wavelength planes (which could be the same in the case of a multifiber system), and M and Q are two sets of *optical network element*s (ONEs) traversed by the two *lightpath*s $P^m_{M,u}$ and $P^m_{Q,k}$. The SRLG of the *lightpath* $P^m_{M,u}$, SRLG ($P^m_{M,u}$), is the union of all *lightpath*s (which belong to the set of all wavelength planes, K) that run the same risk of a single failure with $P^m_{M,u}$. In other words, they traverse at least one common ONE.

SRLG is a dynamic link state that needs to be updated whenever a *lightpath* is modified (e.g., a teardown or a buildup). SRLG is hierarchical [10], and is not limited to physical components (but also protocols, etc). The SRLGs existing in the network can be derived by the following pseudocode:

```
For n = 1 to N do
   Derive all lightpaths traversing the nth ONE
     and put them into temp;
   covered = false;
   If temp ≠ ∅
      For l = 1 to N do
         If temp ⊃ S_l
            S_l ← ∅    //S_l can not be a root element
         End if
         If temp ⊂ S_l and S_l ≠ ∅
            covered = true; // the nth ONE is not a root element
             Break;
         End if
      End for
      If covered = false
          S_n ← temp
       End if
   End if
End for
```

Here, S_l and *temp* are lightpath sets. S_l represents the SRLG with all its lightpaths overlapping at the lth ONE. A root element is where the corresponding SRLG is defined, in which all the lightpaths traverse through the ONE. To update the SRLG information on-line, the processing of the algorithm should be finished before the next event arrives.

The SRLG constraint is a stipulation on the selection of protection resources with which a ONE cannot be reserved for protection by two or more working lightpaths if they belong to the same SRLG. The purpose of following the SRLG constraint is to guarantee the 100% survivability for a single failure occurring to any ONE in the network. When *shared protection* is adopted, in the case that a working and protection paths can be on different wavelength planes, the SRLG constraint impairs the performance more than the case where working and protection paths have to be on the same wavelength plane. It is obvious that the former situation outperforms the latter because of a more flexible utilization of wavelength channels; however, the expense paid for the better performance is that more tunable transceivers and optical amplifiers are required in a node. Note that in the latter, the SLRG constraint exists only if the system is a multifiber system, in which two or more wavelength channels on the same wavelength plane could be in the same SRLG. The lightpaths that take the same wavelength plane may suffer from the SRLG constraint since they cannot share the same protection resources (e.g., a wavelength channel) if they belong to the same SRLG.

As shown in Fig. 1a, working paths P1 and P2 are in the same SRLG with each other since the two working paths have a physically overlapped span F–G. To restore P1 and P2 at the same time once a failure occurred on span F–G, the number of spare links prepared for P1 and P2 should be the summation of bandwidth of P1 and P2 along the spans A–K–H–I–J–D–G. On the other hand, for the two path segments P1 (S–E–F) and P2 (A–B–F–G) in which there is no overlapped span, as shown in Fig. 1b, they are not in the same SRLG. The spare capacity along the spans A–K–H–I–J–D for P1 and P2 can be the maximum of the two paths.

3. CONVENTIONAL TECHNIQUES

This section introduces some classical techniques for protection and restoration, which are frameworks proposed by both industry and academia. The topics included are *SONET self-healing ring*, path-based, link-based *shared protection* schemes, and *short leap shared protection* (SLSP).

The first scheme, *SONET self-healing ring*, is to set up a network in a ring architecture, or in a concatenation of rings, with the facilitation of standard signaling protocols. *SONET self-healing ring*s are characterized by stringent recovery time scales from a failure and a recovery service time of 50 ms. The restoration process performed within the recovery time includes detection, switching time, ring propagation delays, and resynchronization, which are derived from a frame synchronization at the lowest frame speed, (DS1, 1.5 Mb/ps) [8]. On the other hand, the expense paid for the *restoration speed* is the *capacity efficiency*, in which more than 100% (and up to 300%) of capacity redundancy is required.

The second and third schemes are designed for a mesh network, which is characterized by high-*capacity efficiency*, expensive optical switching components, and

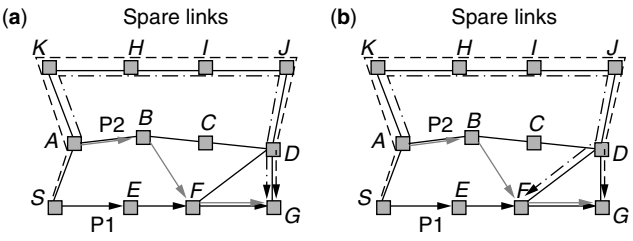


Figure 1. Since there is an overlapped span F–G for P1 and P2 in (**a**), the spare capacity along A–K–H–I–J–D–G has to be the summation of the two working paths. In (**b**), since there is no overlapped span and node, the spare capacity along A–K–H–I–J–D can be shared by P1 and P2.

relatively slow restoration services compared with the SONET self-healing ring.

3.1. SONET Self-Healing

This subsection introduces the *SONET self-healing ring*, which has been an industry standard. The content in this subsection is adopted mainly from the *Internet Draft* [11].

Two types of SONET self-healing rings are widely used: (1) a *unidirectional path-switched ring* (UPSR), which consists of two unidirectional counter-propagating fiber rings, referred to as *basic rings*; and (2) a two-fiber (or four-fiber) bidirectional line-switched ring (BLSR/2 or BLSR/4), which consists of two unidirectional counter-propagating basic rings as well.

In UPSR, the basic concept is to design for channel level protection in two-fiber rings. UPSR rings dedicate one fiber for working time-division multiplexing channels (TDM time slots) and the other for corresponding protection channels (counterpropagating directions). The Traffic is permanently sent along both fibers to exert a $1+1$ protection. Different rings are connected via bridges. To satisfy a bidirection connection, all the resources along working and protection fibers are consumed, as a result, the throughput is restricted to that of a single fiber.

Clearly, UPSR rings represent simpler designs and do not require any notification or switchover signaling mechanisms between ring nodes, namely, receiver nodes perform channel switchovers. As such, they are resource-inefficient since they do not reuse fiber capacity (both spatially, and between working and protection paths). Moreover, span (i.e., fiber) protection is undefined for UPSR rings, and such rings are typically most efficient in access rings where traffic patterns are concentrated around collector hubs.

As for the BLSR, it is designed to protect at the line (i.e., fiber) level, and there are two possible variants, namely two-fiber (BLSR/2) and four-fiber (BLSR/4) rings. The BLSR/2 concept is designed to overcome the spatial reuse limitations associated with two-fiber UPSR rings and provides only path (i.e., line) protection. Specifically, the BLSR/2 scheme divides the capacity timeslots within each fiber evenly between working and protection channels with the same direction (and having working channels on a given fiber protected by protection channels on the other fiber). Therefore bidirectional connections between nodes will now traverse the same intermediate nodes but on differing fibers. This allows for sharing loads away from saturated spans and increases the level of spatial reuse (sharing), a major advantage over two-fiber UPSR rings. Protection slots for working channels are preassigned on the basis of a fixed odd/even numbering scheme, and in case of a fiber cut, all affected time slots are looped back in the opposite direction of the ring. This is commonly termed "loopback" line/span protection and avoids any per-channel processing. However, loopback protection increases the distance and transmission delay of the restored channels (nearly doubling pathlengths in the worst case). More importantly, since BLSR rings perform line switching at the switching nodes (i.e., adjacent to the fault), more complex active signaling functionality is required. Further bandwidth utilization improvements can also be made here by allowing lower-priority traffic to traverse on idle protection spans. Four-fiber BLSR rings extend on the BLSR/2 concepts by providing added span switching capabilities. In BLSR/4 rings, two fibers are used for working traffic and two for protection traffic (counterpropagating pairs, one in each direction). Again, working traffic can be carried in both directions (clockwise, counterclockwise), and this minimizes spatial resource utilization for bi-directional connection setups. Line protection is used when both working and protection fibers are cut, looping traffic around the long-side path. If, however, only the working fiber is cut, less disruptive switching can be performed at the fiber level. Here, all failed channels are switched to the corresponding protection fiber going in the same direction (and lower-priority channels preempted).

Obviously, the BLSR/4 ring capacity is twice that of the BLSR/2 ring, and the four-fiber variant can handle more failures. In addition, it should be noted that both two- and four-fiber rings provide node failure recovery for passthrough traffic. Essentially, all channels on all fibers traversing the failed node are line-switched away from the failed node. BLSR rings, unlike UPSR rings, require a protection signaling mechanism. Since protection channels can be shared, each node must have a global state, and this requires state signaling over both spans (directions) of the ring. This is achieved by an automatic protection switching (APS) protocol, or also commonly termed *SONET APS*. This protocol uses a 4-bit node identifier and hence allows only up to 16 nodes per ring. Additional bits are designated to identify the type of function requested (e.g., bidirectional or unidirectional switching) and the fault condition (i.e., channel state). Control nodes performing the switchover functions utilize framepersistency checks to avoid premature actions and discard any invalid message codes.

3.2. Path-Based Shared Protection

For a path-based *shared protection*, as shown in Fig. 2, the first hop node [6] of the working path w1 computes the protection path p1, which has to be diversely routed from the working path according to the SRLG information. If a fault occurs on the working path, whether it is an LoL or an LoS, the terminating node (N1) in its control plane realizes the fault and sends a notification indicator

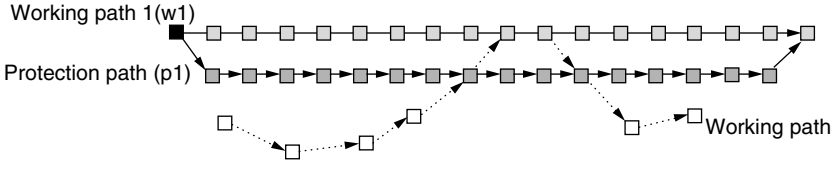


Figure 2. Ordinary path-based 1 : N protection.

signal (NIS) [6] to the first hop node of the path to activate a switchover. Then, the first hop node immediately sends a wakeup packet to activate the configuration of the nodes along the protection path and then switch over the whole traffic on the working path to the protection path.

One of the most important merits of the path-based protection scheme is that it handles LoS and LoL in a single move with relatively lower expense of network resources that need to be reserved. In addition, restoration becomes simpler in terms of signaling algorithm complexity since only the terminating node of the path needs to respond to the fault. On the other hand, it incurs the following difficulties and problems: (1) the complexity of calculation for the diverse protection route grows fast with the increasing number of nodes in the domain, and (2) the protection resources cannot be shared by any other working path that violates the SRLG constraint with the protected working path. For example, in Fig. 2, p1, the protection path for w1, cannot share any of its resources to protect w2 because w2 shares the same link group with w1 only in 1 out of the 17 links.

3.3. Link-Based Shared Protection

Link-based protection was originally devised for the networks with a ring-based architecture such as SONET, in which the effects of network planning play an important role in performance. The *scalability* and reconfigurability have been criticized [12,13] in that a global reconstruction may be realized as a result of a small localized change, which limits the network's scale. The migration of link-based protection from ring-based networks to mesh networks needs more network planning efforts and heuristics, which has been explored intensely [14–16]. In general, the link-based protection in mesh networks is defined as a protection mechanism that performs a fault localization during the occurrence of a failure, and restores the interrupted services by circumventing the traffic from a failed link or node at the upstream neighbor node and merging back to the original working path at the downstream neighbor node. In other words, to protect both the downstream neighbor link failure and node failure, two merge nodes have to be arranged for every node along a working path: one for the upstream neighbor link and one for the upstream neighbor node. As shown in Fig. 3, node

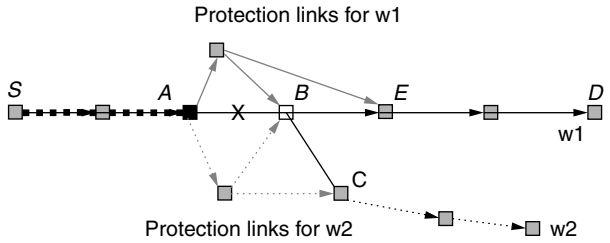


Figure 3. Link-based protection in a mesh WDM network. The node localizing an upstream fault behaves as a PML *path merge LSR (label switch router)* (PML) [6], which only needs to notify its upstream neighbor that behaves as a *path switch LSR* (PSL) [6] before traffic can be switched over to the protection links. The time for transmitting NIS is totally saved in this local restoration mechanism.

A has B and E as merge nodes for w1 and has B and C as merge nodes for w2.

With the associated signaling mechanisms, link-based protection provides a faster restoration due to the fault localization and a better throughput due, in turn, to the relaxation of the SRLG constraint. However, the large amount of protection resources consumed by this scheme may impair the performance and leave room for improvement. Because of its design for the ring-based architecture and its goal for protecting links between nodes, an end-to-end protection service is hard to perform by a pure link-based protection. In addition, it is debatable whether the downstream neighbor node and link are required to have separate protection circles. In Section 3.4, a new definition of link-based protection for achieving an end-to-end protection service in mesh networks will be introduced.

3.4. Short Leap Shared Protection (SLSP)

SLSP [18,19] is an end-to-end service-guaranteed *shared protection* scheme, which is an enhancement of the link- and path-based shared protection, for providing finer service granularities and more network throughput. The main idea of SLSP is to subdivide a working path into several fixed-size and overlapped segments, each of which is assigned by the first hop node a protection domain ID (PDID) after the working path is selected, as shown in Fig. 4. The diameter of a protection domain (or *P domain*) is defined as the hop count of the shortest path between the PSL and PML in the *P domain*. The definition of SLSP generalizes the *shared protection* schemes, in which the link- and path-based shared protection can be categorized as two extreme cases of SLSP with domain diameters of 2 and H, respectively, where H is the hop count of the working path.

The protection path can be calculated for each *P domain* either by the first hop node alone, or distributed to the PSL in each *P domain*, depending on how the SRLG information is configured in each node and how heavy a workload the first hop node can afford at that time. Figure 4 illustrates how a path under SLSP is configured and recovered when a fault occurs. Node A is the first hop node, and node N is the last hop node, which could respectively be the source node and the destination node of this path. The first *P domain* (PDID = 1) starts at node A and ends at node F. The second *P domain* (PDID = 2) is from node E to node J, and the third is from node I to node N. In this case, (A,F), (E,J), and (I,N) are the corresponding PSL–PML pairs for each *P domain*.

Since each *P domain* is overlapped with its neighboring *P domains* by a link and two nodes, a single failure on any link or node along the path can be handled by at least one *P domain*. If a fault occurs on the working path, the Link Management Protocol (LMP) helps localize the fault, and the PSL of the *P domain* in which the fault is located will be notified to activate a traffic switchover. For example, a fault on link 4 or node E is localized by node D. A fault on link 5 or node F is localized by node E. In the former case, node D sends a *notification indicator signal* (NIS) to notify node A that a fault occurred in their *P domains*. In the later case, node E is itself a PSL. In each

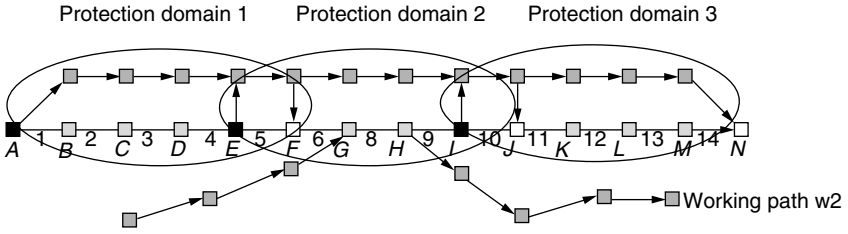


Figure 4. SLSP protection scheme divides the working path into several overlapped *P domains*. Nodes A, E, I are the PSLs, and nodes F, J, N are the PMLs.

case, the PSL (i.e., A or E) immediately sends a wake-up packet to activate the configuration of each node along the corresponding protection path, and then the traffic can be switched over to the protection path. A *tell-and-go* (TAG) [17] strategy can be adopted at this moment so that the PSL (i.e., node A or E) may switch the traffic to the protection path before an acknowledgment packet is received from the PML (i.e., node F or I). At completion of the switchover, the information associated with this rearrangement has to be disseminated to all the other nodes so that the best-effort traffic will not be arranged to use these resources. Under the single-failure assumption, it is impossible that more than one working path will switch their traffic to protection paths at the same time. However, for the environment where multiple failures are considered, a working path has to possess two or more sets of partially or totally disjoint protection paths to prevent the possibility that its protection resources are busy while it needs them. When the fault on the working path is fixed and a switchback to the original working path is required, a notification for releasing the protection resources has to be sent by the PSL to the first hop node right after the traffic is switched. With this, the protection resources can be reported as "free" again to all the other nodes at the next OSPF dissemination.

To implement the protection information dissemination, the association of the protection resources in each *P domain* with corresponding working path segments (PDID) has to be included in its forwarding adjacency. The other working paths must know this association before they can reserve any piece of protection resources for a protection purpose. An example is shown in Fig. 4, where w1 and w2 possess the same SRLG on link 8. However, w2 can share all the protection resources of w1 except those in the second *P domain* (PDID2). In addition, the signaling protocol, such as the resource reservation protocol (RSVP) or label distribution protocol (LDP), needs further extensions for the "path message" and "label request message" to carry object to assign PSLs and PMLs in the SLSP paradigm.

The advantages of the SLSP framework over the ordinary path protection schemes are as follows:

1. The complexity of calculating a diverse route under the constraint of whole domain's SRLG information can be segmented and largely diminished to several *P domains*, in which the provisioning latency for dynamic path selection can be reduced.
2. Both the notification and the wakeup message may be performed only within a *P domain*; therefore, the restoration time is reduced according to the size of the *P domain*.
3. The protection service can be guaranteed more readily since the average/longest restoration time does not vary with the length of the whole path; instead, the average or largest size of the *P domains* will be the dominant factor, which can be an item with which the service providers bill their customers.
4. The computation complexity of protection paths is simplified by the segmentation of working paths. Section 4 introduces signaling issues in which the distributed allocation scheme can reduce the computation efforts of the first hop node that is usually a heavy-loaded border router.
5. The SRLG constraint is relaxed. SLSP divides a working lightpath into several segments, which results in two effects: an increase in the number of SRLGs in the network and a decrease in the size (or the number of lightpaths) of SRLGs. The former has little influence on the network performance while the later can improve the relaxation of SRLG constraint.

An example of advantage 5 (above) is illustrated in Fig. 5. In Fig. 5a, the two working lightpaths (A, G, H, I, F) and (A, G, J, K, L, M) are in the same SRLG if a path-based protection scheme is adopted, so that they cannot share any protection resources. On the other hand, with the adoption of SLSP, as shown in Fig. 5b, the same protection resources (G, R, S, T) can be shared by both of the working paths due

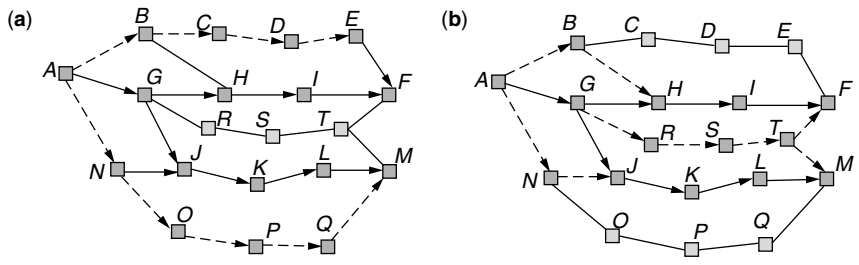


Figure 5. (**a**) Path-based protection; (**b**) SLSP with diameters of 2 and 3.

to the segmentation into two *P domains*. Note that the two lightpaths in Fig. 5 are in the same wavelength plane so that they are stipulated by the SRLG constraint.

Compared with the pure link protection approach, SLSP provides adaptability in compromising restoration time and protection resources required, with which the class of service can be achieved with more granularity. The Isps can put proper constraints on the path selection according to the service-level agreement with each of their customers. The constraining parameters for the selection of the protection path can be those related to the restoration time along the working path, such as the diameter of each *P domain* and the physical distance between each PSL–PML pair.

4. SPARE CAPACITY ALLOCATION

This section introduces the protection and restoration schemes, which are based on the efforts of optimization on the *spare capacity allocation* performed either at the network planning stage, or during the time interval between two network events (i.e., a connection setup or teardown) if each connection request arrives one after the other with a specific holding time. Since the optimization needs to know all the working paths in the network, it has to be processed in a static manner. In other words, any network changes, including a change of traffic, or an extension to the topology, or a change of network administrative policies and routing constraint, will require the optimization to be performed again. The static restoration schemes that will be presented in this section are *ring cover/node cover* [14,15], preconfigured cycle (or p-cycle) [20–22], *protection cycle* [23–25], *survivable routing* [26,27,29], and static SLSP [30].

4.1. Ring Cover/Node Cover

Algorithms were developed to cover all the link/node with least spare capacity in a shape of cycle, tree, or a mixture of both, which have been proved to be NP-hard. Once a fault occurs on any edge or node, a restoration process can be activated to switch over the traffic along the impaired edge onto the spare route. Since every edge/node is covered by the preplanned spare resource, a fault can be recovered any time and anywhere in the network [14,15].

The fatal drawback of *ring / node cover* is that the spare capacity along an edge has to be the maximum capacity of all edges in the network, so that all the possible failure events can be dealt with. This characteristic of *ring / node cover* has motivated the development of more capacity-efficient schemes based on the similar design originality. *Preconfigured cycle* (or *p-cycle*) is one of the successful proposals as far as we can see, and will be presented in the next paragraphs.

4.2. Preconfigured Cycle

The *preconfigured cycle* (or *p-cycle*) is one of the most successful and well-developed strategies for performing spare capacity management at the network planning stage, which is an extension of ring cover. The structure of spare capacity deployed into networks is in a shape of a ring, which is where the "cycle" comes from. Different from ring cover, *p-cycle* addresses networks in which working and spare capacity can vary from link to link, so as to reduce the waste of segmentation of capacity by a ring-based architecture [20–22]. The spare capacity along each ring (or a "pattern" in the terminology of *p-cycle*) is preconfigured instead of being only preplanned, which means that the best-effort traffic can hardly utilize these resources. With knowledge of all the working capacity in networks, *p-cycle* formulates the optimization of the summation of spare capacity in each span into an integer linear programming (ILP) problem, in which different patterns are chosen from the prepared candidate cycles. The result of the optimization is the number of copies for each different cycle pattern, which can be 0 or any positive integer.

Inevitably, the deployment of ring-shaped patterns into mesh networks may result in excess spare links due to a uniform capacity along a ring. On the other hand, since *p-cycle* provides restoration only for the on-cycle and straddling failure, as shown in Fig. 6, intercycle resource sharing is not supported. This above observation motivates us to solve this problem from another point of view; more mesh-based characteristics are taken into account along the design spectrum, which has two ends on the pure ring-based and pure mesh-based design originality.

Because the *p-cycle* does not allow intercycle sharing, the relationship between working paths is not considered. The optimization problem is formulated as follows [20]:

Target:

Minimize

$$\sum_{j=1}^{S} c_j \cdot s_j$$

Subject to the following constraints:

$$s_j = \sum_{i=1}^{N_p} (x_{i,j}) \cdot n_i$$

$$\sum_{i=1}^{N_p} (y_{i,j}) \cdot n_i = w_j + r_j,$$

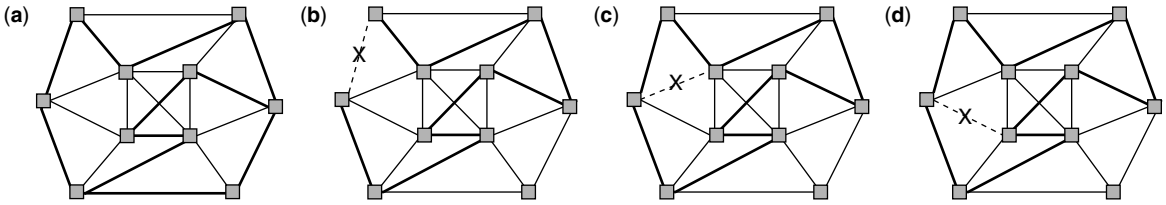


Figure 6. (**a**) A pattern of *p-cycle* in the network. (**b**) An on-cycle failure is restored by using the rest of the pattern. (**c**, **d**) a straddling failure can be restored by either side of the pattern.

where $j = 1, 2, \ldots, S$, and

$$x_{i,j} = \begin{cases} 1 & \text{span } j \text{ has a link on pattern } i \\ 0 & \text{otherwise} \end{cases}$$

$$y_{i,j} = \begin{cases} 0 & \text{both nodes of span } j \text{ not on pattern } i \\ 1 & \text{span } j \text{ is on-cycle with the pattern } i \\ 2 & \text{span } j \text{ is straddling with the pattern } i \end{cases}$$

For each of the parameters, n_j — number of copies of the pattern i; w_j, s_j — number of working and spare links on span j; r_j — spare links excess of those required for span j.

The process of solving the ILP for *p-cycle* is notorious with its NP completeness and a long computation latency, which can prevent the scheme from being used in networks with larger size or with a dynamic traffic pattern. Therefore, the *dynamicity* is traded for the optimized *capacity efficiency* and the swift restoration. Heuristics was provided to improve the computation complexity [22], in which different cost functions were adopted to address either *capacity efficiency*, traffic capture efficiency, or a mixture of both, at the expense of performance. After a ring is allocated, the covered working links are never considered in the subsequent iterations. The iteration ends when all the working links are covered. In considering both intracycle and intercycle protection, the cost function developed takes both *capacity efficiency* η_{capacity} and traffic capture efficiency η_{capture} into account.

4.3. Protection Cycles

Another important proposal for ring-based protection is *protection cycle*s, which is inherent from node and ring cover, and is intended to overcome their drawbacks [23–25]. Algorithms were developed to cover each link with double cycles for both planar and nonplanar networks, so that *automatic protection switching* (APS) can be performed in an optical network with arbitrary mesh topologies. *Protection cycles* also suffer from the problem of *scalability* and requiring a global reconfiguration in response to a local network variation. In addition, the granularity of infrastructure is limited to four-fiber instead of two-fiber in case a wavelength partition is not considered. As a result, the system flexibility is reduced in the implementation of networks with smaller service granularity, such as middle-sized or metropolitan-area networks.

4.4. Survivable Routing

Survivable routing is defined as an approach to derive a better backup path by using the shortest path algorithm with a well-designed cost function and link metrics. Two types of *survivable routing* schemes are introduced in this section; (1) *successive survivable routing* (SSR) [26] and *asymmetrically weighted survivable routing* (AWSR).

In SSR, each traffic flow routes its working path first, then its backup path in the source node next so that the difference in importance between primary and secondary paths can be emphasized. This sequential derivation of the working and secondary paths is where the "successive" comes from. With SSR, protection paths are optimized according to each working path in the network instead of the working bandwidth along each edge, in order to facilitate the intercycle sharing of spare resource and to improve the *capacity efficiency*. For dealing with the relationship between working paths, the SRLG constraint has to be considered and included into the optimization process. Therefore, the optimization process can only be formulated into an integer programming (IP) problem with longer computation time consumed instead of an Integer Linear Programming problem, since the derivation of the SRLG relationship in networks can only be through a nonlinear operation.

In AWSR, on the other hand, working and protection paths are derived at the same stage with a weighting on the cost of a working path, namely, a cost function $\alpha \cdot C + CP$, where C and CP are costs of the working path and its corresponding protection path, respectively; and the parameter α is a weighting on the working path. The meaning of using the weighting parameter α is to distinguish the importance between the resources utilized by the working and protection paths, in which the derivation of the two paths is based on the same link metrics. A special case of $\alpha = 1$ can be solved by the SUURBALLE's algorithm [28] within a polynomial computation time. In case a *dedicated protection* (e.g., $1 + 1$) is adopted, the best value of α could be as small as unity, since there is no difference to the consumption of network resources between a working path and a protection path. On the other hand, for a *shared protection* scheme (e.g., 1:N or M:N), the best value of α could be set large since the protection resources are shared by several working paths (or, in other words, the protection paths are less important than the working paths by α times).

In addition to the SUURBALLE's algorithm, the node-disjoint diverse routing problem can be solved intuitively by the two-step algorithm [28], which first finds the shortest path, and then finds the shortest path in the same graph with the edges and nodes of the first path being erased. Although this method is straightforward and simple, it may fail to find any disjoint path pair after erasing the first path that isolates the source node from the destination node in the network. Besides, in some cases the algorithm cannot find the optimal path pair if there is another better disjoint path pair in which the working path is not the shortest one. To avoid these drawbacks, an enhancement for the two-step algorithm is necessary. The authors of Ref. 27 have conducted a study in which not only the shortest path is examined but also the loopless k shortest paths, where $k = 1, 2, \ldots, K$. The approach proposed in that paper [27] can also deal with the situation in which the working and protection paths are required to take different (or heterogeneous) link metrics, in order to further distinguish the characteristics of the two paths. On the other hand, another study [29] focused on the optimization or approximate optimization for the derivation of working and protection path pairs with uniform link metrics.

4.5. Static SLSP (S-SLSP)

S-SLSP is aimed at the task of an efficient spare capacity reallocation, which takes the *scalability* (or computation complexity) and network *dynamicity* into

consideration [30]. In terms of resources sharing and *capacity efficiency*, the fundamental difference between S-SLSP and *p-cycle* or any other ring-based *spare capacity allocation* schemes is that the former investigates the relationship between lightpaths, so that all types of resource sharing are supported, which is potentially more capacity-efficient. As mentioned in Section 3.4, *P domains* are allocated in a cascaded manner along a working path, which facilitates not only link, but also node protection, so that an all-aspect restoration service can be provided. Because spare resources are preplanned instead of being preconfigured, better throughput can be achieved by launching best-effort traffic onto the spare capacity in normal operation.

There are two phases for the implementation of the S-SLSP algorithm:

1. Prepare all the cycles in the network up to a limited size.
2. Optimize the deployment of each cycle according to the existing working traffic in the network.

In step 1, the cycle listing algorithm can be seen in Grover et al. [22] and Mateti and Deo [31]. Step 2 can be simply formulated as an integer programming (IP) problem shown as follows.

Objective:
Minimizing

$$\delta \cdot \sum_{j=0}^{S} c_j \cdot s_j$$

subject to the constraints

$$s_j = \max_i \{x_{i,j} \cdot ns_i\} + SR_j$$

$$w_j = \sum_{i=1}^{N_p} (x_{i,j}) \cdot nw_i$$

$$n_i = nw_i + ns_i$$

$$W_j \geq w_j + s_j$$

where c_j and s_j are the cost and the spare capacity on the span j, respectively, and S is the number of spans in the network; n_i is the number of copies for the *P domain* i; SR_j is the extra spare links needed on span j for meeting the SRLG constraint; ns_i and nw_i are the numbers of spare and working links on span i; and W_i is the number of links on span i. The binary parameter $x_{i,j}$ is 1 if the span j has a link on domain i, and 0 otherwise. The binary parameter δ is 1 if the plan meets the requirement of SLSP, 0 otherwise.

Although the formulation above is feasible, the brutal-attack-type optimization by examining the combinations of all cycles for each working path may yield terribly high computation complexity. S-SLSP reduces the computation complexity by interleaving the optimization process into several sequential sub-processes, in which an improvement from $O(2^N)$ to $O(m \cdot 2^n)$ can be made, where $N = n \cdot m$ is the number of candidate *P domains* for all working paths, m is the number of interleaved processes, and n is the number of candidate *P domains* in each subset of working paths. In other words, several subsets of working paths are optimized one after the other. It can be easily seen that the larger the number of working paths in a subset, the closer is the optimality that can be achieved, but at an expense of computation time.

Grouping of the existing working paths into several subsets is implemented according to the following two principles: (1) The working paths in a subset should be overlapped with each other as much as possible and (2) Since the candidate cycles can be up to a limited size according to the SLA of each working path, working paths with a similar requirement of restoration time are also grouped into the same subset. Working paths in a subset are processed with the integer programming solver together to determine the corresponding spare capacity. A framework of optimization on the performance can be developed with the size of each subset of working paths varied. Since every network event (i.e., a connection setup or teardown) will change the network capacity distribution, the spare capacity reallocation has to restart to follow the new link state. Therefore, a "successful" reallocation process is completed before the next network event arrives. Note that the computation time is exponentially increased with the volume of each subset. Since the increase in computation time will decrease the chance of completing the Integer programming task, as a consequence, the performance is impaired. On the other hand, the increase of the volume of each subset of working paths can improve the quality of optimization. Because of the two abovementional criteria, an optimized number of working paths in a subset can be derived to yield the best performance.

5. SUMMARY

The *optical internet* has changed the lifestyle of human beings by providing various applications, such as e-commerce, e-conferencing, Internet TV, Internet phone, and video on demand (VoD). This article introduced the survivable *optical internet* from several aspects, which includes a number of selected proposals and implementations for performing protection and restoration mechanisms in order to deal with single failure in network components.

This article first described the evolution of data networks and the importance of *survivability*, then some background knowledge and terminology were presented. Sections 3 and 4 described the latest progress in the protection and restoration strategies proposed by both industry and academia, which includes a comparison between all the approaches mentioned in terms of *scalability*, *dynamicity*, *class of service*, *capacity efficiency*, and *restoration speed*. Table 1 lists the results of the discussion, and compares the strategies mentioned in this article.

BIOGRAPHIES

Hussein Mouftah joined the Department of Electrical and Computer Engineering at Queen's University,

Table 1. Protection versus Restoration Strategies

	Scalability	Dynamicity	Class of Service	Capacity Efficiency	Restoration Speed
Ring/node cover	Low	Low	Low	Low	Low
p-cycle	Low	Low	Low	High	High[b]
Protection cycles	Low	Low	N/A[a]	Low	High[a]
Successive Survivable Routing (SSR)	Medium	High	Low	Medium	Medium
Asymmetrically Weighted Survivable Routing (AWSR)	Medium	High	Low	Medium	Medium
Static-SLSP	High[c]	High[c]	High[c]	Medium	Medium

[a] Protection Cycles is a pure link-based protection, in which every link has a *dedicated protection* cycle.
[b] The p-cycle has the highest *restoration speed* because of its preconfigured spare resources instead of being only preplanned.
[c] S-SLSP divides a working path into several protection domains, and support dynamic allocation of working protection path pairs, which can achieve *scalability*, *dynamicity*, and *class of service*.

Kingston, Canada, in 1979, where he is now a full professor and the department associate head, after three years of industrial experience mainly at Bell Northern Research of Ottawa (now Nortel Networks). He obtained his Ph.D. in EE from Laval University, Quebec, Canada, in 1975 and his B.Sc. in EE and M.Sc. in CS from the University of Alexandria, Egypt, in 1969 and 1972, respectively. He served as editor in chief of the IEEE Communications Magazine (1995–97) and IEEE Communications Society Director of Magazines (1998–99). Dr. Mouftah is the author or coauthor of two books and more than 600 technical papers and eight patents in this area. He is the recipient of the 1989 Engineering Medal for Research and Development of the Association of Professional Engineers of Ontario (PEO). He is the joint holder of a honorable mention for the Frederick W. Ellersick Price Paper Award for Best Paper in Communications Magazine in 1993. He is also the joint holder of two Outstanding Paper Awards; for papers presented at the IEEE Workshop on High Performance Switching and Routing (2002) and at the IEEE 14th International Symposium on Multiple-Valued Logic (1984). He is the recipient of the IEEE Canada (Region 7) Outstanding Service Award (1995). Dr. Mouftah is a fellow of the IEEE (1990).

Pin-Han Ho received his B.S. and M.S. degrees from the Department of Electrical Engineering, National Taiwan University, Taiwan, in 1993 and 1995, respectively. He also received an M.Sc. (Eng) degree from the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada, in 2000. He is currently finishing his Ph.D. degree in the same Department of Electrical and Computer Engineering at Queen's University. He is the joint holder of the Outstanding Paper Award for a paper presented at the IEEE Workshop on High Performance Switching and Routing (2002). His area of interest is optical networking and routing and wavelength assignment in survivable WDM-routed networks.

BIBLIOGRAPHY

1. A. S. Tanenbaum, *Computer Networks*, 2nd ed., Prentice-Hall, 1996.
2. W. Stallings, *Data & Computer Communications*, 6th ed., Prentice-Hall, 2000.
3. D. Awduche and G.-S. Kuo, MPLS in broadband IP networks, *Tutorial T11, IEEE Int. Conf. Communications* (ICC2000), New Orleans, June 22, 2000; *http://www.icc00.org/technic/tutor/thursday.htm*.
4. Y. Rekhter et al., *Cisco systems tag switching architecture overview*, RFC 2105.
5. E. Rosen, A. Viswanathan, and R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, Jan. 2001.
6. S. Makam et al., Framework for MPLS-based recovery, *Internet Draft, <draft-makam-mpls-recovery-frmwrk-03.txt>, work in progress*, July 2001.
7. J. H. Hahm et al., Restoration mechanisms and signaling in optical networks, *Internet Draft, <draft-many-optical-restoration-00.txt>, work in progress*, Feb. 2001.
8. R. Ramaswami and K. N. Sivarajan, *Optical Networks—A Practical Perspective*, Morgan Kaufmann Publishers, 1998.
9. P. Gadiraju and H. T. Mouftah, Channel protection in WDM mesh networks, *IEEE Workshop on High Performance Switching and Routing, Dallas*, May 2001, pp. 26–30.
10. D. Papadimitriou et al., Inference of shared risk link groups, *Internet Draft, <draft-many-inference-srlg-00.txt>, work in progress*, Feb. 2001.
11. N. Ghani et al., Architectural framework for automatic protection provisioning in dynamic optical rings, *Internet Draft, <draft-ghani-optical-rings-01.txt>, work in progress*, Sept. 2001.
12. M. Kodialam and T. V. Lakshman, Dynamic routing of locally restoration bandwidth guaranteed tunnels using aggregated link usage information, *Proc. IEEE, Infocom'01*, 2001, pp. 376–385.
13. D. Zhou and S. Subramaniam, Survivability in optical networks, *IEEE Network* 16–23 (Nov./Dec. 2000).

14. O. J. Asem, Optimal topologies for survivable fiber optic networks using SONET self-healing rings, *Proc. IEEE GLOBECOM*, 1991, Vol. 3, pp. 2032–2038.

15. W. D. Grover, Case studies of survivable ring, mesh and mesh-arc hybrid networks, *Proc. IEEE GLOBECOM*, 1992, Vol. 1, pp. 633–638.

16. D. Stamatelakis and W. D. Grover, IP layer restoration and network planning based on virtual protection cycles, *IEEE Journal of Selected Areas in Communications* **18**(10): 1938–1949 (Oct. 2000).

17. C. Qiao, A high speed protocol for bursty traffic in optical networks, *SPIE All-Opt. Commun. Syst.* **3230**: 79–90 (Nov. 1997).

18. P.-H. Ho and H. T. Mouftah, A framework for service-guaranteed path protection of the optical internet, *optical Network Design and Modeling* (*ONDM*), Vienna, Austria, Feb. 2001, pp. MO3.2.1–MO3.2.13.

19. P.-H. Ho and H. T. Mouftah, SLSP: A new path protection scheme for the optical Internet, *Optical Fiber Communications* (*OFC*) *2001*, Anaheim, CA, March 2001, pp. TuO1.1–TuO1.3.

20. D. Stamatelakis and W. D. Grover, *Network Restorability Design Using Pre-configured Trees, Cycles, and Mixtures of Pattern Types*, TRLabs Technical Report TR-1999-05, Issue 1.0, Oct. 2000.

21. W. D. Grover and D. Stamatelakis, Cycle-oriented distributed preconfiguration: Ring-like speed with mesh-like capacity for self-planning network restoration, *Proc. IEEE Int. Conf. Communications*, 1998, Vol. 1, pp. 537–543.

22. W. D. Grover, J. B. Slevinsky, and M. H. MacGregor, Optimized design of ring-based survivable networks, *Can. J. Electric. Comput. Eng.* **20**(3): 138–149 (Aug. 1995).

23. C. Thomassen, On the complexity of finding a minimum cycle cover of a graph, *SIAM J. Comput.* **26**(3): 675–677 (1997).

24. G. Ellinas and T. E. Stern, Automatic protection switching for link failures in optical networks with bidirectional links, *Proc. IEEE GLOBECOM*, 1996, Vol. 1, pp. 152–156.

25. G. Ellinas, A. G. Hailemariam, and T. E. Stern, Protection cycles in mesh WDM networks, *IEEE J. Select. Areas Commun.* **18**(10): 1924–1937 (Oct. 2000).

26. Y. Lie, D. Tipper, and P. Siripongwutikorn, Approximating optimal spare capacity allocation by successive survivable routing, *Proc. IEEE Infocom'01*, 2001, Vol. 2, pp. 699–708.

27. P.-H. Ho and H. T. Mouftah, Issues on diverse routing for WDM mesh networks with survivability, *10th IEEE Int. Conf. Computer Communications and Networks* (*ICCCN'01*) (in press).

28. R. Bhandari, *Survivable Networks: Algorithms for Diverse Routing, The Kluwer International Series in Engineering and Computer Science*, Kluwer, Boston, 1999.

29. J. Tapolcai et al., Algorithms for asymmetrically weighted pair of Disjoint Paths in survivable networks, *Proc. IEEE 3rd Int. Workshop on Design of Reliable Communication Networks* (*DRCN 2001*), Budapest, Oct. 2001.

30. P.-H. Ho and H. T. Mouftah, *Spare Capacity Re-allocation with S-SLSP for the Optical Next Generation Internet*, Research Report-01-011, Optical Networking Lab., ECE Dept., Queen's University, Kingston, Ontario, Sept. 2001.

31. P. Mateti and N. Deo, On algorithms for enumerating all circuits of a graph, *SIAM J. Comput.* **5**(1): 90–99 (March 1976).

SYNCHRONIZATION IN DIGITAL COMMUNICATION SYSTEMS

M. Luise
U. Mengali
M. Morelli
University of Pisa
Pisa, Italy

The word *synchronization* (often abbreviated *sync*) refers to signal processing functions accomplished in digital communication receivers to achieve correct alignment of the incoming waveform with certain locally generated references. For instance, in a baseband pulse amplitude modulation system the signal samples must be taken with a proper phase so as to minimize the intersymbol interference. To this purpose the receiver generates clock ticks indicating the location of the optimum sampling times. As a second example, in bandpass transmissions a coherent receiver needs *carrier* synchronization (or recovery), which means that the demodulation sinusoid must be locked in phase and frequency to the incoming carrier. Clock and carrier recovery are instances of *signal* synchronization, which is carried out within the physical layer of the system. By contrast, *network* synchronization is usually performed on digital data streams at the higher layers of the ISO/OSI stack. This happens for example with the time alignment of data frames within the format of the digital stream. As network synchronization is usually easier to implement, we will concentrate to a larger extent on signal synchronization, with a short overview of the former at the end of the article.

Signal synchronization is a crucial issue in digital communication receivers. The advent of low-cost and high-power VLSI circuits for digital signal processing (DSP) has dramatically affected the receiver design rules. At the moment of writing this article the vast majority of newly designed transmission equipment is heavily based on DSP components and techniques.[1] For this reason, we concentrate in the following on DSP-based synchronization, with just occasional references to analog methods. After a brief introduction to the general topic of clock and carrier recovery (Section 1), we consider synchronization for narrowband signals in Section 2 and wideband spread-spectrum signals in Section 3. The case of multicarrier transmission is dealt with in Section 4, and is followed by a short review of frame synchronization in Section 5.

1. AN INTRODUCTION TO SIGNAL SYNCHRONIZATION

The ultimate task of a digital communication receiver is to produce an accurate replica of the transmitted data sequence. With Gaussian channels, the received signal component is completely known except for the data

[1] Exceptions are very-high-speed systems for fiberoptic communications [beyond 1 Gbps (gigabits per second)], wherein the signal processing functions are implemented with analog components.

symbols and a group of variables, denoted *synchronization parameters*, which are ultimately related to the signal time displacement with respect to a local reference in the receiver. Reliable data detection requires good estimates of these parameters. Their measurement is referred to as *synchronization* and represents a crucial part of the receiver. The following examples help illustrate the point.

In baseband synchronous transmission the information is conveyed by uniformly spaced pulses representing bit values. The received signal is first passed through a filter (usually matched to the incoming pulses) and then is sampled at the symbol rate. To achieve the best detection performance, the arrival times of the pulses must be accurately located so that the filtered waveform is sampled at "optimum" times. A circuit implementing this function is called *timing* or *clock synchronizer*.

As a second example, consider a bandpass communication system with coherent demodulation. Optimum detection requires that the received signal be converted to baseband making use of a local reference with the same frequency and phase as the incoming carrier. This calls for accurate frequency and phase measurements, since phase errors may severely degrade the detection process. Circuits performing such measurements, as well as appropriate corrections to the local carrier, are called *carrier synchronizers*.

In the following, we will specifically deal with carrier and clock synchronization of bandpass-modulated signals. The function of clock synchronization for (carrierless) baseband transmission will be seen as a special case that can be derived from bandpass techniques with minor modifications only.

From the observations above it is clear that synchronization plays a central role in the design of digital communication equipment since it greatly affects overall system performance. Also, synchronization circuits represent such a large portion of a receiver's hardware and/or software that their implementation has a considerable impact on the overall cost [1,2].

2. SIGNAL SYNCHRONIZATION IN NARROWBAND TRANSMISSION SYSTEMS

By *narrowband* we mean signals whose bandwidth around their carrier frequency is comparable to the information bit rate. Most popular narrowband-modulated signals are phase shift keying (PSK), quadrature amplitude modulation (QAM), and a few nonlinear modulations such as Gaussian minimum shift keying and the broader class of continuous-phase modulations (CPM). We will restrict our attention to PSK and QAM for the sake of simplicity.

2.1. Signal Model and Synchronization Functions

The mathematical model of a modulated signal is

$$s_{\mathrm{mod}}(t) = s_R(t)\cos(2\pi f_0 t) - s_I(t)\sin(2\pi f_0 t) \tag{1}$$

where f_0 is the carrier frequency. The quantity $s(t) = s_R(t) + js_I(t)$ is the *complex envelope* of $s_{\mathrm{mod}}(t)$ (where $j = \sqrt{-1}$), and consists of uniformly spaced pulses

$$s(t) = \sum_i c_i g(t - iT) \tag{2}$$

where c_i is the ith transmitted symbol, T is the signaling interval, and $g(t)$ is the basic pulseshape. Symbols $\{c_i\}$ are taken from a *constellation* of M points in the complex plane. With M-PSK modulation, c_i may be written as $c_i = e^{j\alpha_i}$, where $\alpha_i \in \{0, 2\pi/M, \ldots, 2\pi(M-1)/M\}$. With M-QAM signaling we have $c_i = a_i + jb_i$, where a_i and b_i belong to the set $\{\pm 1, \pm 3, \ldots, \pm(\sqrt{M}-1)\}$.[2]

The physical channel corrupts the information-bearing signal with different impairments such as distortion, interference, and noise. When the only impairment is additive noise, the received waveform takes the form

$$r_{\mathrm{mod}}(t) = s_{\mathrm{mod}}(t - \tau) + w_{\mathrm{mod}}(t) \tag{3}$$

where τ is the propagation delay and $w_{\mathrm{mod}}(t)$ is background noise. For the additive white Gaussian noise (AWGN) channel, $w_{\mathrm{mod}}(t)$ is a Gaussian random process with zero mean and two-sided power spectral density $N_0/2$. The signal-to-noise ratio (SNR) is defined as the ratio of the power of the signal component to the noise power in a bandwidth equal to the inverse of the signaling interval T (the so-called Nyquist bandwidth).

As is shown in Fig. 1, the demodulation is performed by multiplying (downconverting) $r_{\mathrm{mod}}(t)$ by two local quadrature references: $2\cos(2\pi f_{\mathrm{LO}} t + \varphi)$ and $-2\sin(2\pi f_{\mathrm{LO}} t + \varphi)$. The products are lowpass (LP)-filtered to eliminate the frequency components around $f_{\mathrm{LO}} + f_0$. In general, the local carrier frequency f_{LO} is not exactly equal to f_0, and the difference $\nu = f_{\mathrm{LO}} - f_0$ is referred to as *carrier frequency offset*. Assuming that the filters' bandwidth is large enough to pass the signal components undistorted, and collecting the filter outputs into a single complex-valued waveform $r(t) = r_R(t) + jr_I(t)$, after some mathematics it is found that

$$r(t) = e^{j(2\pi\nu t + \theta)} \sum_i c_i g(t - iT - \tau) + w(t) \tag{4}$$

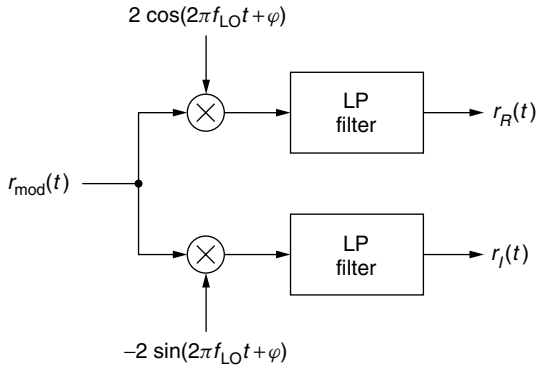


Figure 1. I/Q signal demodulation.

[2] Here, the number of points in the constellation is an *even* power of 2. Different constellations with M equal to an *odd* power of 2 or with an arbitrary number of points are commonly used as well.

where $\theta = -(2\pi f_0 \tau + \varphi)$ and $w(t) = w_R(t) + jw_I(t)$ is the noise contribution. Inspection of this equation reveals that the demodulated signal contains the following unknown parameters: the *frequency offset* ν, the *phase offset* θ, and the *timing offset* τ. As is now explained, reliable data detection cannot be achieved without knowledge of these parameters.

Assume that the frequency offset is much smaller than the signal bandwidth, as is often the case. Then, passing $r(t)$ through a filter matched to $g(t)$ produces

$$x(t) = e^{j(2\pi \nu t + \theta)} \sum_i c_i h(t - iT - \tau) + n(t) \tag{5}$$

where $n(t)$ is filtered noise and $h(t)$ is the convolution of $g(t)$ with $g(-t)$. To single out the effects of frequency and phase errors, we assume for the moment that τ is perfectly known and that $h(t)$ satisfies the first Nyquist criterion for the absence of intersymbol interference (ISI):

$$h(kT) = \begin{cases} 1 & \text{for} \quad k = 0 \\ 0 & \text{for} \quad k \neq 0 \end{cases} \tag{6}$$

Then, sampling $x(t)$ at the ideal clock instants $t_k = kT + \tau$, we get

$$x[k] = c_k e^{j[2\pi \nu (kT + \tau) + \theta]} + n[k] \tag{7}$$

where $n[k]$ is the noise sample. From (7) it is seen that the useful component of $x[k]$ is rotated by a time-varying angle $\psi[k] = 2\pi \nu (kT + \tau) + \theta$ with respect to its correct position, and this may have a catastrophic effect on system performance. For instance, consider a simple binary PSK signal with $c_k \in \{-1, +1\}$. Regeneration of the digital datastream is carried out making the decision $\hat{c}_k = \pm 1$ according to whether $\Re e\{x[k]\} = \pm 1$. Neglecting the noise for simplicity, it is easily found that $\Re e\{x[k]\} = c_k \cos(\psi[k])$, which means that the receiver makes decision errors whenever $\pi/2 < |\psi[k]| \leq \pi$. This condition is certainly met (sooner or later) in the presence of a frequency offset, but it may also hold true due to a *phase* offset even when the frequency offset is negligible.

Now consider a timing error. In doing so we assume that frequency and phase offsets have already been compensated for, and that the receiver has elaborated an *estimate* $\hat{\tau}$ of the channel delay (in general, $\hat{\tau} \neq \tau$). Then, sampling the matched-filter output at the clock instants $\hat{t}_k = kT + \hat{\tau}$ yields

$$x[k] = c_k h(-\varepsilon) + \sum_{m \neq 0} c_{k-m} h(mT - \varepsilon) + n[k] \tag{8}$$

where $\varepsilon = \tau - \hat{\tau}$ is the *timing error*. Note that, as ε is usually small compared to the sampling period T, we have $h(-\varepsilon) \approx 1$. The first term on the right-hand side (RHS) of (8) is the useful component, the second is a disturbance referred to as *intersymbol interference* (ISI), and the third is Gaussian noise. Clearly, the ISI tends to "mask" the useful component and can produce decision errors even in the absence of noise. For this reason $h(t)$ is usually given a shape satisfying the first Nyquist criterion. This would make the ISI vanish if $x(t)$ were sampled exactly at the nominal clock instants $t_k = kT + \tau$ (i.e., $\varepsilon = 0$). In practice some residual ISI is inevitable, but its amount can be limited by keeping ε small.

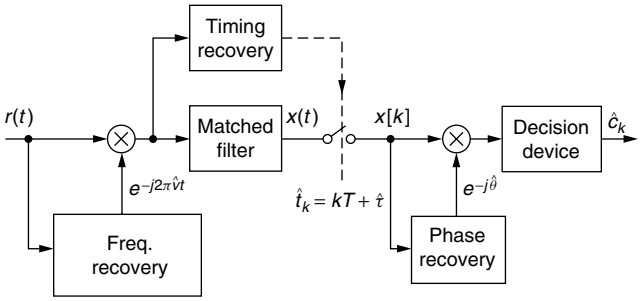


Figure 2. Coherent receiver with sync functions.

To sum up, Fig. 2 illustrates the signal synchronization functions of a coherent receiver for PSK or QAM signals. The blocks indicated as frequency, phase, and timing recovery have the task of providing reliable estimates $\hat{\nu}$, $\hat{\theta}$, and $\hat{\tau}$ of the corresponding sync parameters ν, θ, and τ, respectively. With such estimates, the receiver *compensates* for the sync offsets. To do so, it first multiplies (downconverts) the received signal by $e^{-j2\pi \hat{\nu} t}$; next it takes samples at the correct clock instants $\hat{t}_k = kT + \hat{\tau}$, and finally it multiplies the signal samples by $e^{-j\hat{\theta}}$ to cancel out the residual phase offset. It is worth noting that Fig. 2 has only illustration purposes; indeed, the real receiver architecture may be somewhat different. For instance, phase recovery may be accomplished before matched filtering and without exploiting timing information. Similarly, timing recovery may be derived directly from the demodulated waveform $r(t)$, prior to frequency compensation.

So far our discussion has concentrated on the simple case of transmission on the AWGN channel, which is typical of satellite communications. Unfortunately this simplified scenario does not fit many modern communication systems, as, for example, mobile cellular networks or high-speed wireline systems for the access network. In such systems the channel frequency response is not flat across the signal bandwidth and the signal undergoes unpredictable linear distortions. This implies that, in addition to the sync parameters mentioned above, the receiver also has to estimate the *channel impulse response* (CIR). Indeed, the received signal takes the form

$$r(t) = e^{j2\pi \nu t} \sum_i c_i p(t - iT) + w(t) \tag{9}$$

where $p(t)$, the (complex-valued) CIR to be estimated, may be modeled as the convolution of three impulse responses: those of the transmit filter, the physical channel, and the receive filter. Comparing (9) to (4), it is seen that the parameters θ and τ are no longer visible as they have been incorporated into $p(t)$. Thus, CIR estimation has become an augmented synchronization problem. Although CIR estimation and synchronization functions tend to overlap and possibly merge in modern communication equipment, in the following we stick to synchronization issues only. The interested reader is referred to articles on estimation and equalization chapters elsewhere in this encyclopedia.

A further warning about Fig. 2 is that the diagram seems to suggest that synchronization is always derived from the information-bearing signal (*self-synchronization*). Actually, the transmitted signal often contains known sequences periodically inserted into the datastream. They are referred to as *preambles* or *training sequences* and serve to ease the estimation of the sync parameters. Notwithstanding, in other cases synchronization is carried out by exploiting a dedicated channel, the so-called *pilot channel* (*pilot-aided synchronization*). Even though the method requires extra bandwidth, it is often adopted whenever channel distortion and multiple access interference make the detection process critical. This occurs with third-generation cellular systems (UMTS in Europe and Japan, and cdma2000 in the Americas) where both the downlink (from the radio base station to the mobile phones) and the uplink do employ pilot channels.

2.2. Performance of Synchronization Functions

From the discussion in the previous section it is recognized that synchronization functions may be split into two parts: *estimation* of the sync parameters and *compensation* of the corresponding estimated offsets. Strategies to accomplish these tasks are now pinpointed, and the major performance indicators are introduced.

The estimation of a sync parameter is accomplished by operating on the digitized samples of the received signal. Methods to do so fall into two categories: either *feedforward* (open loop) or *feedback* (closed loop). The former provide estimates based on the observation of the received signal over a finite time interval (*observation window*), say, $0 \le t \le T_0$. At the end of the interval the estimate is released and is used for compensation of the relevant offset. Compensation is applied either back on the stored signal samples within the observation window (if adequate digital memory is available), or on the samples of the subsequent window (assuming that the sync parameters are almost constant from one window to the next). An attractive feature of feedforward schemes is that they have short estimation times and, as such, are particularly suited for burst-mode transmissions where fast synchronization is mandatory. Feedback schemes, on the other hand, have much longer estimation times as they operate in a recursive fashion. For instance, in a clock recovery circuit the timing estimate is periodically updated (usually at symbol rate) according to an equation of the type

$$\hat{\tau}[k+1] = \hat{\tau}[k] - \gamma e_\tau[k] \tag{10}$$

where $\hat{\tau}[k]$ is the estimate at the kth step, $e_\tau[k]$ is an *error signal*, and γ is a design parameter (*step size*). The error signal $e_\tau[k]$ is provided by the synchronizer and (to a first approximation) is proportional to the kth estimation error $\hat{\tau}[k] - \tau$. Equation (10) is reminiscent of a feedback control system; since the correction term $-\gamma e_\tau[k]$ has a sign opposite the error $\hat{\tau}[k] - \tau$, the latter is steadily forced toward zero. Feedback synchronizers are akin to the analog phase-locked loops (PLLs) used for carrier acquisition and tracking. Feedforward synchronizers have no counterparts in analog hardware.

The estimate of a sync parameter λ is a random variable depending on samples of the received signal. Accordingly, it is customary to qualify the accuracy of an estimate by specifying its *mean value* and *mean-square error* (MSE). The estimate $\hat{\lambda}$ is said to be *unbiased* if its mean value equals the true value λ. The bias is defined as

$$b(\hat{\lambda}) = E\{\hat{\lambda}\} - \lambda \tag{11}$$

where $E\{\cdot\}$ means statistical expectation. In general we would like $b(\hat{\lambda})$ to be zero since this means that the estimates would coincide with λ at least "on average." The MSE of the estimator is given by

$$\mathrm{MSE}(\hat{\lambda}) = E\{(\hat{\lambda} - \lambda)^2\} \tag{12}$$

Clearly, the smaller $\mathrm{MSE}(\hat{\lambda})$ is, the more accurate the estimate $\hat{\lambda}$ is, in that the latter will be "close" to λ with high probability.

In searching for good estimators, one wonders what is the ultimate accuracy that can be achieved. An answer is given by the Cramér–Rao bound (CRB) [3], which represents a lower limit to the MSE of *any unbiased estimator*:

$$\mathrm{MSE}(\hat{\lambda}) \ge \mathrm{CRB}(\hat{\lambda}) \tag{13}$$

Knowledge of the CRB is very useful as it establishes a benchmark against which the performance of practical estimators can be compared.

Another fundamental performance parameter of a sync system is the *acquisition time*. With feedforward schemes this is just the time needed to compute the estimate $\hat{\lambda}$ and coincides with the length of the observation window (plus possibly some extra time to carry out the calculations). With feedback systems, however, the acquisition time represents the number of iterations needed to achieve a steady-state estimation condition. As is intuitively clear from Ref. 10, the acquisition time depends on the step size γ. In general, the larger γ is, the quicker the convergence. However, increasing γ degrades the estimation accuracy in the steady state. In practice, some tradeoff between these conflicting requirements must be sought.

A possible drawback of closed-loop schemes is the *cycle slipping* phenomenon [4]. Briefly, imagine a steady-state condition in which $\hat{\lambda}[k]$ is fluctuating around the true value λ. Fluctuations are usually small but, occasionally, they grow large as a consequence of the combined effects of noise and the random nature of the data symbols. If a large deviation occurs, it may happen that $\hat{\lambda}[k]$ is "attracted" toward an equilibrium point other than the initial steady-state value [1]. When this happens, a synchronization failure (*cycle slip*) occurs, and we say that the loop has *lost lock*. Cycle slips must be rare events in a well-designed loop because they reflect the presence of large errors. Their rate of occurrence is usually measured by the *mean time to lose lock*, specifically, the average time between two consecutive sync loss events.

2.3. Frequency Synchronization

Feedforward frequency synchronizers are typically used in burst-mode transmissions (as in satellite time-division

multiple access), with frequency offsets much smaller than the signaling rate. In these circumstances timing recovery is performed first and the frequency estimator operates on symbol rate samples. Most frame formats are endowed with a *preamble* that is used to cancel the modulation in the samples (7) from the matched filter by dividing $x[k]$ by c_k. This *Data-Aided* (DA) operation produces

$$z[k] = e^{j[2\pi\nu(kT+\tau)+\theta]} + n'[k] \quad 0 \le k \le N-1 \tag{14}$$

where $n'[k] = n[k]/c_k$ and N is the observation length in symbol intervals. When no preamble is available, the data modulation is usually wiped out by processing $x[k]$ through a nonlinear function. With M-ary PSK, for example, $x[k]$ is raised to the Mth power. This is called *non-data-aided* (NDA) or "blind" processing.

Several algorithms have been devised to estimate ν from $z[k]$. An efficient method, proposed by Rife and Boorstyn (RB) [5], consists of computing the Fourier transform (FT) of the sequence $z[k]$

$$Z(f) = \sum_{k=0}^{N-1} z(k) e^{-j2\pi fkT} \tag{15}$$

and taking $\hat{\nu}$ as the value of f where $|Z(f)|$ achieves a maximum. It turns out that the RB algorithm is unbiased and attains the Cramér–Rao bound

$$\mathrm{CRB}_\nu = \frac{3}{2\pi^2 T^2 N(N^2-1)} \times (\mathrm{SNR})^{-1}(\mathrm{Hz})^2 \tag{16}$$

at intermediate to high signal-to-noise ratios. At low SNRs, however, the estimates are plagued by occasional large estimation errors (*outliers*) and the variance of the estimator exhibits a *threshold effect*, typical of nonlinear estimation schemes. The threshold effect manifests as an abrupt increase of the estimation MSE as the SNR decreases below a certain value (the estimator's threshold).

Although the RB algorithm can be implemented through efficient fast Fourier transform (FFT) techniques, its complexity is too high with large data records. Simpler methods based on linear regression of $\arg\{z[k]\}$ have been proposed in [6,7]. Their main drawback is that their threshold may be as high as 7–8 dB, and cannot be lowered by increasing the observation length (as occurs with the RB scheme). Other estimators are discussed in [8,9] and are based on the sample correlation function of $z[k]$

$$R[m] = \frac{1}{N-m} \sum_{k=m}^{N-1} z[k] z^*[k-m] \quad 1 \le m \le Q \tag{17}$$

where Q is a design parameter. In general, increasing Q improves the accuracy of the estimates but reduces the maximum offset that can be estimated (*estimation range*). A method to make the estimation range independent of Q is given elsewhere [10].

The frequency estimators described in other studies [5–10] were conceived specifically for unmodulated signals (or for DA operation), but they can also be used in NDA recovery [11]. The price to pay is an increase of the threshold with respect to DA operation, especially with large-signal constellations. This may be a serious drawback with efficient forward error-correcting codes, like Turbo codes, where the operating SNR is as low as 1–2 dB.

All of the above frequency synchronizers produce biased estimates when the transmitted signal undergoes linear distortion, as happens in wideband transmissions. The issue of frequency recovery for selective channels is difficult to deal with. Suffice it to say that, even theoretically, it is not clear whether unbiased frequency estimates can be obtained without knowledge of the CIR. The problem of joint estimating CIR and frequency offset is addressed in Ref. 12 assuming that a preamble is available. The resulting scheme performs well but is complex to implement. A different solution [13] exploits knowledge of the channel statistics.

A mixed analog/digital feedback frequency synchronizer is sketched in Fig. 3. As is seen, the frequency-corrected signal $r'(t) = r(t)e^{-j\phi(t)}$ is fed to a *frequency error detector* (FED) whose purpose is to generate the error signal $e_\nu[k]$. The latter gives an indication of the difference between the current estimate $\hat{\nu}(k)$ from the digitally controlled oscillator (DCO) and the true value ν. Similar to (10), the error signal is then used in a feedback loop to update the estimates in a recursive fashion:

$$\hat{\nu}[k+1] = \hat{\nu}[k] - \gamma e_\nu[k] \tag{18}$$

The frequency estimate is then used in the DCO to generate the exponential $e^{-j\phi(t)}$ such that

$$\frac{d\phi(t)}{dt} = 2\pi\hat{\nu}[k], \quad kT \le t < (k+1)T \tag{19}$$

Compared with feedforward schemes, feedback synchronizers typically have larger acquisition times but can track large slow time-varying frequency offsets. For this reason they are well suited for continuous-mode transmissions. Their counterpart in analog hardware is the traditional automatic frequency control (AFC) loop, commonly used in radio receivers.

The heart of closed-loop schemes is the FED. The maximum-likelihood-based FED [14], the quadricorrelator [15] and the dual-filter detector [16], all share the same kind of loop error signal:

$$e_\nu[k] = \int_{kT}^{(k+1)T} \Im m\{y^*(t)z(t)\}dt \tag{20}$$

where $y(t)$ and $z(t)$ are obtained by passing $r'(t)$ through two suitable lowpass filters, $\Im m\{\cdot\}$ means *imaginary part*

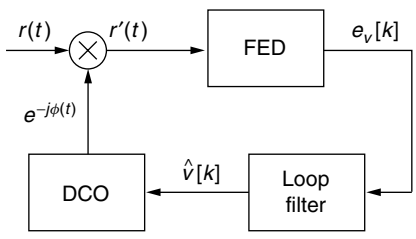


Figure 3. Feedback frequency recovery.

of the enclosed quantity, and the asterisk denotes complex conjugation. In its simplest version, $y(t)$ is just the matched filter output and $z(t)$ is a replica of $y(t)$ delayed by a symbol interval, namely, $z(t) = y(t - T)$. The preceding schemes do exploit neither knowledge of data symbols nor timing information and can recover frequency offsets as large as the symbol rate $1/T$. Unfortunately, their accuracy is far from the CRB and their estimates are biased in the presence of frequency-selective fading. The reason of the bias is that they take the center of gravity of the received spectrum as an estimate of the carrier frequency. With multipath propagation, however, the signal spectrum is distorted by the channel frequency response, and its center of gravity is moved off its original position. A closed-loop frequency synchronizer for transmissions over multipath channels has been proposed [17]. It gives unbiased estimates provided that the channel is flat over the rolloff regions of the received spectrum.

2.4. Phase Synchronization

When using phase-coherent detection the phase offset must be properly compensated for before data detection takes place. In a DSP-based receiver, phase recovery is usually performed after timing correction using the symbol rate samples $x[k]$ from the matched filter. Information symbols may be known (training sequence) or not. Assuming that the frequency offset is zero or has been perfectly compensated for [i.e., setting $\nu = 0$ in (7)], we have

$$x[k] = e^{j\theta} c_k + n[k] \tag{21}$$

When the symbols c_k are known, a feedforward phase synchronizer yields the maximum-likelihood (ML) estimate of θ as follows:

$$\hat{\theta} = \arg \left\{ \sum_{k=0}^{N-1} c_k^* x[k] \right\} \quad -\pi \leq \theta < \pi \tag{22}$$

where N is the observation length in symbol intervals and $\arg\{\cdot\}$ denotes the argument of the enclosed complex value. The estimate (22) is unbiased and achieves the CRB

$$\mathrm{CRB}_\theta = \frac{1}{2N} (\mathrm{SNR})^{-1} (\mathrm{rad})^2 \tag{23}$$

When no preamble is available, data modulation can be removed from $x[k]$ by means of suitable nonlinear processing. For example, with M-ary PSK, $x[k]$ is raised to the Mth power to yield the following open-loop NDA phase estimate:

$$\hat{\theta} = \frac{1}{M} \arg \left\{ \sum_{k=0}^{N-1} x^M[k] \right\} \quad -\frac{\pi}{M} \leq \theta < \frac{\pi}{M} \tag{24}$$

A variant of this estimator is proposed by Viterbi and Viterbi (VV) in [18]. The VV algorithm converts $x[k]$ to polar coordinates $x[k] = \rho[k] e^{j\varphi[k]}$ and then replaces $x^M[k]$ in (24) with either $\rho^2[k] e^{jM\varphi[k]}$ or $e^{jM\varphi[k]}$ (depending on the SNR and the number M of constellation points). Although this nonlinear processing degrades the estimation accuracy, the VV algorithm achieves the CRB at high SNR (i.e., asymptotically).

As indicated in (24) the NDA schemes are forced to operate on a smaller phase interval than the whole 2π angle, namely, $-\pi/M \leq \theta < \pi/M$. This implies that they give estimates that are ambiguous by multiples of $2\pi/M$. Distinct phase offsets differing by $2\pi/M$ are "mapped" onto the same estimated value within the base interval $-\pi/M \leq \theta < \pi/M$. The ambiguity is usually resolved by means of a *unique word* [1] or by differential encoding and decoding [19].

NDA phase estimators for QAM constellations are obtained by letting $M = 4$ in (24) (fourth power estimators) or in a VV-like algorithm. The estimation accuracy of such schemes tends to fall short of the CRB (even at high SNRs) as the number of constellation points increases [20].

The phase recovery algorithms illustrated so far have all a feedforward structure that makes them suited to burst-mode transmission. However, feedback loops are preferred with continuous transmissions for the sake of simplicity. Probably the most popular feedback scheme is illustrated in Fig. 4. Here the phase-compensated signal samples $y[k] = x[k] e^{-j\hat{\theta}[k]}$ and the detected symbols $\hat{c}_k$ are fed to the phase error detector (PED) to generate the error signal

$$e_\theta[k] = \Im m\{y^*[k] \cdot \hat{c}_k\} \tag{25}$$

The latter is then passed through an IIR filter which updates the phase estimate according to

$$\hat{\theta}[k+1] = \hat{\theta}[k] - \gamma e_\theta[k] \tag{26}$$

where γ is the step size. Finally, the lookup table produces the map $\hat{\theta}[k] \rightarrow e^{-j\hat{\theta}[k]}$.

As the PED uses data decisions to build up the error signal, the loop is said to operate in a *decision-directed* (DD) mode. When a preamble is available it may also operate in a data-aided (DA) "training mode" to ease initial acquisition. In this case data decisions $\hat{c}_k$ in (25) are replaced by the preamble symbols.

In the presence of an uncompensated residual frequency error the phase estimates $\hat{\theta}[k]$ are biased if the loop filter is first-order as indicated in (26). To cope with moderate-frequency errors, second-order loop filters are adopted [1]. The phase recovery algorithm (25)–(26) is often referred to as "digital Costas loop" since it resembles

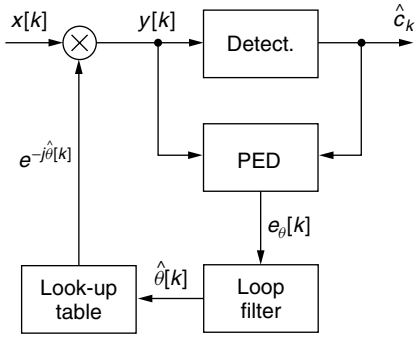


Figure 4. Digital Costas loop for phase recovery.

the analog PLL invented in the 1950s by Costas to track the chrominance subcarrier of color television signals [21].

The steady-state performance of Costas loops is good but, as happens with the NDA feedforward estimators, the estimates are ambiguous by multiples of $2\pi/M$ with M-ary PSK and by multiples of $\pi/4$ with QAM modulation. Again, the ambiguity can be resolved either with unique words or with differential encoding and decoding.

In the previous discussion we have concentrated on uncoded transmissions. With trellis-coded modulations the tracking performance of a conventional Costas loop may fail as a result of the decision delay inherent in the detection process. In these circumstances per-survivor-processing (PSP) techniques [22] are preferable. In practice, each surviving path in the trellis diagram generates a phase estimate based on its own tentative decisions. The estimate is then used to extend that survivor one step further in the trellis. The issue of phase recovery with large decoding delays (like those experienced in Turbo decoding [23]) is still an open problem.

2.5. Timing Synchronization

As with the other sync functions, timing recovery consists of two distinct operations: (1) estimation of the timing offset τ (*timing estimation*) and (2) application of the estimate to the sampling process (*timing compensation*).

Two different timing recovery architectures are possible. Figure 5 shows a feedforward scheme with *asynchronous signal sampling*. The received signal is passed through an antialias filter (AAF) and it is then fed to the A/D converter (represented by the sampler in Fig. 5). The latter is controlled by a free-running oscillator, having no reference whatsoever with the clock of the data signal. The sampling rate $1/T_s$ is usually higher than the symbol rate by an oversampling factor >2. Timing correction is achieved by *interpolating* the samples $x[m] = x(mT_s)$ from the matched filter according to the estimates of τ. This "resynthesizes" signal samples at the correct timing instants, which are seldom present in the digitized stream $x[m]$. Simple piecewise polynomial interpolators are described by Erup et al. [24].

The alternative architecture (see Fig. 6) involves *synchronous signal sampling* and is particularly useful with high-data-rate modems where oversampling is too expensive or not feasible. Here, timing correction is accomplished through a feedback loop wherein a timing error detector (TED) drives a numerically controlled oscillator (NCO). The latter updates the timing estimates

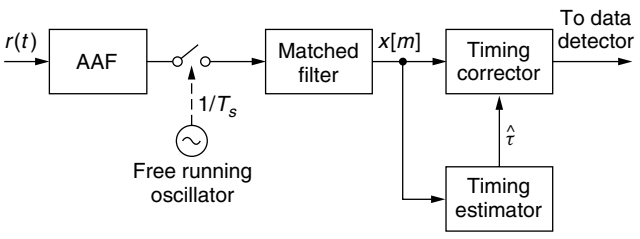


Figure 5. Feedforward timing recovery with asynchronous sampling.

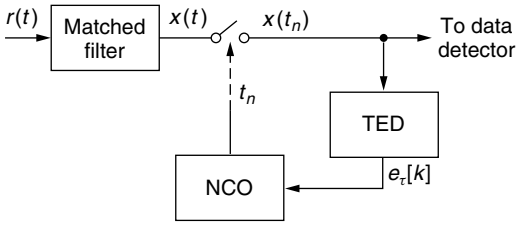


Figure 6. Feedback timing recovery with synchronous sampling.

according to

$$\hat{\tau}[k+1] = \hat{\tau}[k] - \gamma e_\tau[k] \tag{27}$$

where $e_\tau[k]$ is the TED output and γ is the step size. In so doing, the NCO produces a sequence of clock pulses that are *synchronous* with the data symbol clock.

Feedback timing estimation may be performed jointly with carrier phase recovery in a DD mode. To this end the zero-crossing detector [15] or the Muller–Mueller detector [25] may be employed. The main drawback of these methods is that spurious locks may occur with large QAM constellations due to complex interactions between phase and timing loops. A simpler approach is to recover timing independently of the carrier phase. The Gardner detector (GD) [26] and the NDA early–late detector (ELD) [27] are widely used for timing recovery in the absence of phase information. They operate with an oversampling factor of 2 and generate the following error signals

$$e_{\mathrm{GD}}[k] = \Re e\{(x(kT + \hat{\tau}[k]) - x(kT - T + \hat{\tau}[k-1])) \times x^*(kT - T/2 + \hat{\tau}[k-1])\} \tag{28}$$

$$e_{\mathrm{ELD}}[k] = \Re e\{(x(kT - T/2 + \hat{\tau}[k-1]) - x(kT + T/2 + \hat{\tau}[k])) \times x^*(kT + \hat{\tau}[k])\} \tag{29}$$

where $x(t)$ is the output of the matched filter. Note that both $e_{\mathrm{GD}}[k]$ and $e_{\mathrm{ELD}}[k]$ are insensitive to phase errors, since any possible phase offset term $e^{j\theta}$ on the received signal vanishes in the product between the samples of signal $x(t)$ and its own complex conjugate.

Returning to feedforward schemes, the most popular algorithm in this class is that proposed by Oerder and Meyr (OM) [28]. It needs an oversampling factor M greater than 2, and has the form

$$\hat{\tau} = -\frac{T}{2\pi} \arg\left\{ \sum_{k=0}^{MN-1} |x(kT_s)|^2 e^{-j2\pi k/M} \right\} \tag{30}$$

where N is the observation length (in symbol intervals). Simulations indicate that both GD and the OM algorithms have good performance with raised-cosine-shaped pulses, provided that the rolloff factor is sufficiently large. The estimation accuracy becomes poor as the signal bandwidth decreases.

The above algorithms are tailored for transmissions over the AWGN channel. With multipath channels the situation is more complex. For example, the MSE at the output of a symbol-spaced equalizer is very sensitive

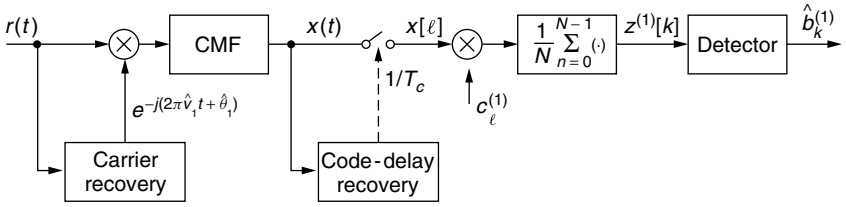


Figure 7. Coherent CDMA receiver.

to the timing phase and, for a given CIR, an optimal phase exists that minimizes the bit error rate (BER). The key point is how this phase can be found and properly tracked as the channel varies in time. A reasonable solution is proposed by Godard [29] under some restrictive assumptions. Timing recovery may be avoided by resorting to fractionally spaced equalizers whose performance is insensitive to the sampling phase [30].

2.6. Timing Synchronization in Baseband Transmission

The timing algorithms for bandpass signals described in Section 2.5 can be used with minor changes for baseband transmission as well. The GD and ELD are directly translated to baseband by just interpreting $x(t)$ as a real-valued signal and consequently by dropping the complex-conjugate operation. The same is true with the OM open-loop estimator. It is worth noting that the ELD in (29) is just the digital counterpart of the popular analog early–late synchronizer for rectangular data pulses [31].

3. SYNCHRONIZATION FOR SPREAD-SPECTRUM AND CDMA SIGNALS

3.1. Signal Model and Synchronization Functions

With the term *spread-spectrum* we address any modulated signal whose bandwidth around the carrier frequency is much larger than its information bit rate. The most popular spread-spectrum format for commercial applications is *direct-sequence* spread-spectrum (DS/SS) wherein spectral spreading is directly obtained through the use of a high-rate sequence (the *code*) of binary values called *chips*. DS/SS is the basis for code-division multiple access (CDMA) techniques adopted in mobile cellular systems such as IS95, cdma2000, and UMTS. In a CDMA network, each user is assigned a specific code (or *signature sequence*). It consists of a pseudorandom sequence with a repetition period of L chips, and with a chip rate $1/T_c$ equal to N times the symbol rate $1/T$. The parameter N is called the *spreading factor* since the signal bandwidth is about $1/T_c$ and is N times larger than the bandwidth of the modulating signal, $1/T$.

Consider the ith user in a CDMA network and call $b_k^{(i)}$ his/her binary datastream, with index k running at the symbol rate $1/T$. Denoting by $\lfloor n/N \rfloor$ the integer part of n/N and by $|n|_N$ the remainder of the division, the baseband equivalent signal transmitted by the ith user takes the form

$$s^{(i)}(t) = \sum_{\ell} c^{(i)}_{|\ell|_L} b^{(i)}_{\lfloor \ell/N \rfloor} g(t - lT_c) \tag{31}$$

where $g(t)$ is the chip pulseshape and $\{c_\ell^{(i)}; 0 \le \ell \le L-1\}$ is the signature sequence (with ℓ ticking at the chip rate).

Figure 7 is a block diagram of a conventional coherent receiver for user 1. Waveform $r(t)$ is the sum of signals from all the active users (in number of I). Assuming an AWGN channel, it takes the form

$$r(t) = e^{j(2\pi \nu_1 t + \theta_1)} s^{(1)}(t - \tau_1) + \sum_{i=2}^{I} A_i e^{j(2\pi \nu_i t + \theta_i)} s^{(i)}(t - \tau_i) + w(t) \tag{32}$$

where $w(t)$ is the noise; ν_i and θ_i are carrier frequency and phase offsets, respectively; while A_i and τ_i account for the attenuation and delay experienced by the ith signal. We say that the CDMA system is *synchronous* when all signals share the same chip and symbol framework: $\tau_i = \tau$ for $i = 1, 2, \ldots, I$. This happens for example in the downlink of a mobile communication network. Otherwise, when the signals are not bound by synchronicity constraints (as in the case of the uplink from mobiles to base station), we speak of *asynchronous* multiple access.

Returning to Fig. 7, after frequency and phase correction, the received waveform is fed to the chip matched filter (CMF) and then is sampled at chip rate. Spectral despreading is performed by multiplying the samples $x[\ell]$ by a locally generated replica of the user's code and, finally, the resulting sequence is accumulated over a symbol period.

Correct receiver operation requires accurate recovery of the signal time offset τ_1 to ensure that the local code replica is properly aligned with the signature sequence in the received signal. This goal is usually achieved in two steps — a coarse alignment is obtained first and is then used as a starting point for code tracking.

Assuming perfect carrier recovery and code alignment, it turns out that the decision variable $z^{(1)}[k]$ at the accumulator output is given by

$$z^{(1)}[k] = b_k^{(1)} + \eta^{(1)}_{\mathrm{MAI}}[k] + n[k] \tag{33}$$

where $n[k]$ is the noise contribution while $\eta^{(1)}_{\mathrm{MAI}}[k]$, the *multiple-access interference* (MAI), accounts for the presence of the other users. MAI can be reduced to zero in synchronous CDMA by using *orthogonal* code sequences (Walsh–Hadamard). With asynchronous CDMA, however, orthogonality cannot be enforced even with Walsh–Hadamard sequences because of the different time offsets $\tau_1, \ldots, \tau_I$. In most cases MAI is the major limiting factor to achieve accurate synchronization and reliable data detection in CDMA systems. Approximating MAI as Gaussian noise, the last two terms on the RHS of (33) can be lumped together to give

$$z^{(1)}[k] = b_k^{(1)} + n'[k] \tag{34}$$

where $n'[k]$ is zero-mean Gaussian with a variance equal to the sum of the variance of $n[k]$ and of the MAI term $\eta_{\mathrm{MAI}}^{(1)}[k]$.

From the discussion above it appears that the synchronization problem in CDMA systems is similar to that encountered with narrowband signals, with two main differences: (1) the presence of MAI and (2) the broader signal bandwidth that makes the time offset compensation (code synchronization) much more complex, as discussed in Section 3.2.

3.2. Code Synchronization

As mentioned earlier, signal demultiplexing relies on the availability at the receiver of a time-aligned version of the spreading code. The delay τ_1 must be estimated and tracked to ensure such an alignment. The main difference with respect to narrowband modulations is the estimation accuracy. In narrowband systems, timing errors must be small compared to the *symbol time* whereas in spread-spectrum systems they must be small compared with the *chip* time, which is N times smaller.

Assume again an AWGN channel and, for simplicity, that the spreading factor equals the code repetition length (the so-called *short-code* DS/SS format). Then, code acquisition amounts to finding the start of the symbol interval and is usually performed by correlating the input signal with the local replica of the code sequence (*sliding correlator*). The magnitude of the correlation is used in two ways: it is compared to a threshold to decide whether the intended user is actually transmitting and, if this is the case, to locate the position of the maximum. This location is taken as a coarse estimate of the delay τ_1.

The search for the maximum of the correlation may be performed with either serial or parallel schemes [32]. Serial schemes test all the possible code delays in sequence until the threshold is crossed. They are simple to implement but are inherently time-consuming and their acquisition time cannot be established a priori (it can only be predicted statistically). Parallel schemes look for all the possible code epochs in parallel and choose the one corresponding to the maximum correlation. They guarantee short acquisitions but are computationally intensive.

A simplified scheme of a serial detector is sketched in Fig. 8. The PN-code generator provides a spreading code with a tentative initial epoch δ, which is correlated with the (digitized) received signal on a symbol period. The result is squared to cancel out any possible phase offset (noncoherent processing); then it is smoothed on a W-symbol dwell time, and finally it is compared to a threshold. If the threshold is crossed, the code epoch is frozen and the fine timing recovery process is started. Otherwise, the PN-code generator is advanced by one chip and a new trial acquisition is performed. The value of the threshold λ is a design parameter. Low values of λ correspond to high probabilities of false acquisition events caused by large noise peaks. On the other hand, high threshold values produce occasional acquisition failures. The false-alarm probability and missed-detection probability depend on λ and on the SNR. The threshold value is chosen on the basis of the operating conditions.

Conventional correlation-based methods have satisfactory performance in a power-controlled system, when signals from different users arrive at the receiver with comparable amplitudes. However, they fail in a *near–far* situation where strong-powered users interfere with weaker ones. In these cases MAI becomes a serious impairment to achieve accurate code synchronization. Improvements are obtained by taking into account the statistical properties of the MAI. For example, near–far resistant code acquisition is achieved by modeling MAI as colored Gaussian noise [33].

Initial code acquisition provides the receiver with a coarse estimate of the delay τ_1. Fine timing recovery (also called *code tracking*) is then needed to locate the optimum chip-rate sampling instants. Code tracking is typically performed by means of feedback loops, where a suitable timing error signal is used to update the timing estimate at symbol rate according to an equation of the type (27). Figure 9 depicts the architecture of a digital

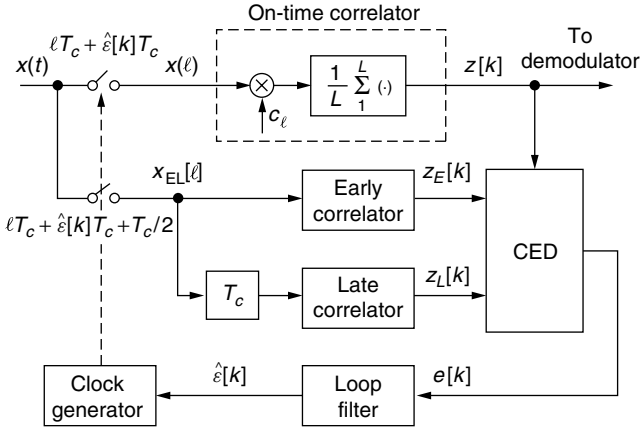


Figure 9. Digital delay-lock loop for fine timing recovery of a DS/SS or CDMA signal.

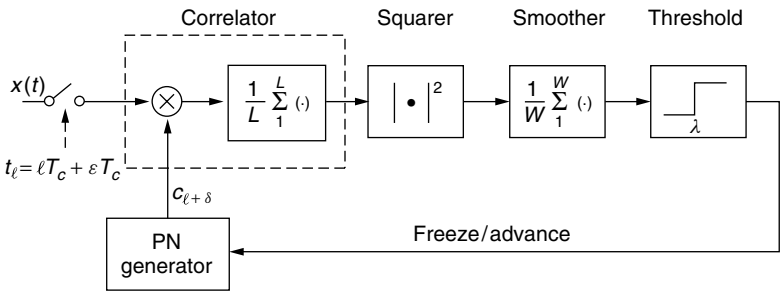


Figure 8. Serial code acquisition of a DS/SS signal.

delay-lock loop (DLL) [34] performing such a function. It is seen that data demodulation relies on the "on-time samples," those taken at the optimum sampling instants. The DLL also needs the "early/late samples," those taken midway between two consecutive on-time samples. The estimate $\hat{\varepsilon}[k]$ of the normalized chip delay ε is then used to drive an NCO or an interpolator, depending on whether synchronous or asynchronous sampling is employed. Moeneclaey and De Jonghe [35] give some examples of low-complexity chip-timing error detectors (CEDs). For instance, detectors insensitive to the phase offset (not requiring prior phase recovery) are expressed by

$$e_1[k] = \Re e\{(z_E[k] - z_L[k])z^*[k]\} \tag{35}$$

$$e_2[k] = |z_E[k]|^2 - |z_L[k]|^2 \tag{36}$$

The first is reminiscent of the early–late timing detector for narrowband transmissions; the second is typical of spread-spectrum signals. Their performance is satisfactory in the absence of multipath and near–far effects.

The foregoing discussion has been concerned with DS-CDMA transmissions over the AWGN channel. On the other hand, third generation CDMA-based systems are expected to support multimedia services with data rates up to 2 Mbps. In these circumstances the transmission channel is characterized by multipath propagation. A nice feature of CDMA systems is that they can resolve multipath components and optimally combine them by means of a RAKE receiver [19] or other more sophisticated receiver structures based on multiuser detection [36]. In all cases, accurate estimates of the relative delays, amplitudes, and phases of the propagation paths are needed to achieve reliable data detection [37–39]. In particular, the problem of measuring the propagation delays seems crucial since performance degrades rapidly with timing misalignments in excess of a small fraction of the chip interval.

3.3. Carrier Frequency and Phase Synchronization

Code timing recovery is typical of CDMA signals. Carrier recovery, on the other hand, does not differ much with respect to narrowband modulations. In general, carrier recovery is performed after code acquisition, exploiting the despread signal $z[k]$ from the correlator in Fig. 9. A notable exception occurs when the frequency offset is comparable with the symbol rate. In this case code acquisition becomes unreliable because the frequency offset *decorrelates* the signal within the integration window. Indeed, things go as if the signal power were reduced by a factor

$$L = \left[\frac{\pi \nu T_i}{\sin(\pi \nu T_i)}\right]^2 \tag{37}$$

where T_i is the window length. For example, a frequency offset of $1/(2T_i)$ causes a loss of more than 6 dB. On the other hand, frequency offset cannot be reliably estimated unless the code sequence is coarsely acquired. A chicken–egg problem arises that can be approached with *joint* estimation of the frequency offset and the code phase. This leads to a *bidimensional* grid search in which the frequency uncertainty range is partitioned into a number of "bins" and the code acquisition test is repeated for all the bins [32]. In the end, coarse estimates of code phase and frequency offset are available. The latter is then used to correct the local oscillator so as to reduce the residual frequency error to a small fraction of the symbol rate. At that stage frequency tracking is performed by means of conventional feedback schemes such as a quadricorrelator [15] or a dual-filter detector [16].

Frequency recovery for CDMA transmissions on frequency-selective channels is still an open problem. Current research investigates methods to alleviate the combined effects of MAI and multipath on the acquisition process.

4. SYNCHRONIZATION IN MULTICARRIER TRANSMISSION

4.1. Signal Model and Synchronization Functions

In multicarrier transmission, the output of a high-rate data source is split into many low-rate streams modulating adjacent subcarriers within the available bandwidth. If N is the number of the subcarriers, the symbol rate on each of them is reduced by a factor N with respect to the source rate, and this squeezes the signal bandwidth around the subcarrier to a point that the transmission channel appears to be locally flat. Correspondingly, the channel distortion on each subcarrier is reduced to a multiplicative factor that can be compensated for by a simple one-tap equalizer. The possibility of easing the equalization function has motivated the adoption of multicarrier transmission as a standard in a number of current applications, for example, European *digital audiobroadcasting* (DAB) and terrestrial *digital videobroadcasting* (DVB), IEEE 802.11 wireless local area networks, and asymmetric digital subscriber line (ADSL) and its high-speed variant VDSL, to mention only a few.

Figure 10 shows the transmitter of the most popular multicarrier transmission technique, namely, orthogonal frequency-division multiplexing (OFDM), as adopted in DAB or DVB. The input data symbols c_i at rate $1/T$ are serial-to-parallel (S/P)-converted and partitioned into blocks of length N. The mth block $\boldsymbol{c}^{(m)} = [c_0^{(m)}, c_1^{(m)}, \ldots, c_{N-1}^{(m)}]$ is fed to an N-point inverse discrete Fourier transform (IDFT) unit to produce the N-dimensional vector $\boldsymbol{b}^{(m)}$ (an OFDM block). The Fourier transform is efficiently computed via fast Fourier transform (FFT) techniques. Next, $\boldsymbol{b}^{(m)}$ is extended by appending to its end a copy of the first part of the vector, say, from $b_0^{(m)}$ to $b_{N_g-1}^{(m)}$, denoted by the *cyclic prefix*. The resulting extended vector drives a linear modulator with a rectangular impulse response $g(t)$ and a signaling interval

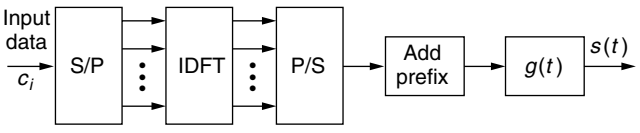


Figure 10. OFDM transmitter.

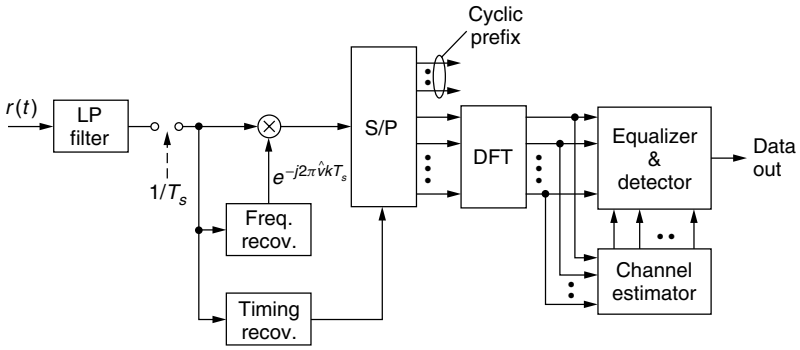


Figure 11. OFDM receiver.

$T_s = NT/(N + N_g)$. The cyclic prefix makes the signal insensitive to intersymbol interference, provided N_g is greater than the duration of the channel impulse response expressed in symbol intervals.

The OFDM receiver is sketched in Fig. 11. After lowpass (LP) filtering, the signal is sampled at rate $1/T_s$ and frequency and timing synchronization is performed. Next, the samples are serial-to-parallel-converted and, after removal of the cyclic prefix, they are passed to an N-point DFT unit whose output drives the decision device. Note that the timing circuit does not control the sampling operations. Its only purpose is to locate an appropriate window containing the samples to feed into the DFT.

Assuming perfect timing and frequency correction, the output of the DFT corresponding to the mth OFDM block is found to be

$$X^{(m)}[n] = c_n^{(m)} H[n] + w^{(m)}[n] \quad 0 \le n \le N-1 \tag{38}$$

where $w^{(m)}[n]$ is channel noise and $H[n]$ is the channel response at the frequency $f_n = n/(NT)$ affecting the nth subcarrier. From (38) it is seen that the frequency selectivity of the channel only appears as a multiplicative term (amplitude/phase factor) on each data symbol. Accordingly, channel equalization can be performed in the frequency domain through a bank of complex multipliers [40]. This requires estimation of the channel response, which is usually performed by means of suited pilot symbols inserted in the transmitted data frame.

4.2. Frequency and Timing Estimation

Timing recovery in OFDM is significantly different from that in single-carrier systems. It just amounts to estimating where the OFDM block starts. Because of the cyclic prefix, timing offsets need not be particularly small. As long as they are shorter than the difference Δ between the length of the cyclic prefix and the CIR duration, their only effect is to generate a linear phase term across the DFT output, which can be compensated for by the channel equalizer. This is seen as follows. Denoting by εT_s the timing error (less than Δ), the DFT output is found to be

$$\begin{aligned} X^{(m)}[n] &= c_n^{(m)} H[n] e^{j2\pi n\varepsilon/N} + w^{(m)}[n] \\ &= c_n^{(m)} H'[n] + w^{(m)}[n] \quad 0 \le n \le N-1 \end{aligned} \tag{39}$$

In the second equality the factor $e^{j2\pi n\varepsilon/N}$ and the channel gain $H[n]$ have been lumped together into a single term $H'[n]$. This implies on one hand that the channel estimator cannot distinguish between timing errors and channel distortions and, on the other hand, that the equalizer itself can perform fine-timing synchronization.

The main problem with OFDM systems is their sensitivity to frequency errors. It can be easily argued that the accuracy of a frequency synchronizer for OFDM has to be N times larger than with a conventional equivalent monocarrier modulation. Depending on the application, the initial frequency offset can be as large as many times the subcarrier spacing $1/(NT)$. If not properly compensated for, it gives rise to intercarrier interference (ICI), meaning that the nth DFT output $X^{(m)}[n]$ depends not only on $c_n^{(m)}$ [as indicated in (39)] but also on all the other symbols within the mth block.

Frequency estimation in OFDM systems can be performed jointly with timing recovery by exploiting either the redundancy introduced by the cyclic prefix [41] or pilot symbols inserted at the start of the frame [42]. The timing estimate is given by the location of the maximum of a correlation computed from the time-domain samples. The phase of the correlation is then used to estimate the frequency offset. Some feedback schemes for frequency tracking are discussed in the literature [43,44].

Only single-user OFDM systems have been considered so far. Orthogonal frequency-division multiple access (OFDMA) is a multiuser and multicarrier application that has been proposed for the uplink of wireless systems and cable TV [45] because of its robustness against multipath distortion and multiuser interference. Timing and frequency synchronization in the uplink of an OFDMA system is still an open problem that is currently under investigation.

5. FRAME SYNCHRONIZATION

In any digital stream the transmitted data have always some kind of "framing" depending on the communication system characteristics. For instance, computer data are organized into 8-bit packets (bytes), which in turn may be grouped into 512-byte or 1024-byte blocks. As a second example, some data-link layer functions (e.g., those for error control) require a specific segmentation of data. In all these cases correct detection and interpretation of data requires that the receiver be synchronized with such framing. This is the task of *frame synchronization*, which is the only instance of network synchronization that we will deal with.

The most common technique to achieve frame synchronization makes use of framing markers. In essence, a *sync word or unique word* (UW) is periodically inserted into the data pattern to mark the start of a frame. The receiver knows the UW in advance and performs a search for its location in the received stream. The operation is called *frame acquisition* and is usually performed on the regenerated data. Figure 12 depicts a digital correlator for frame synchronization. The receiver continuously *correlates* the incoming bits b_k with the sync word bits u_k until a match is found. Note that b_k and u_k take on the logical values 0 or 1 and the inverted XOR gates act as comparators, meaning that their output is 1 only if the two inputs have the same value. An "in-sync" condition is declared whenever a threshold λ is crossed:

$$\sum_{i=k-N+1}^{k} b_i \circ u_{k-i} \geq \lambda \tag{40}$$

where $\circ$ denotes the comparator operation. To avoid false locks, the transmitter is prevented from sending a data pattern equal to the UW. Stuffing additional bits in the "forbidden sequence" does this.

To reduce the probability of declaring a false lock, the UW bit pattern is designed in such a way as to exhibit a low level of "off-sync" correlation. To see this point, define the cross-correlation function of the UW as

$$R_k = \sum_{i=k}^{N-1} u_i \circ u_{i-k} \geq 0 \quad k = 0, \ldots, N-1 \tag{41}$$

Clearly R_k equals N when $k = 0$ (in-sync condition) and must be much smaller than N when $k \neq 0$ (off-sync condition), otherwise a false threshold crossing might occur when searching for the correct frame alignment. Barker sequences [46] have off-sync correlation less than unity. Other good sequences are found in the textbook by Wu [47].

Transmission errors in the regenerated bits may impair frame synchronization. The performance metric in this context is the *probability of missed detection* P_{MD}, which is defined as the probability that the digital correlator in Fig. 12 fails to detect the UW because of bit errors.

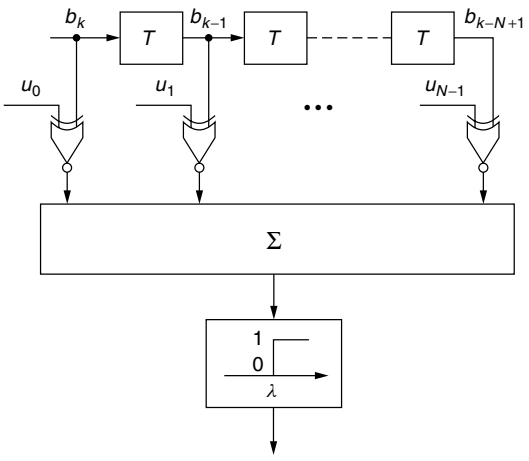


Figure 12. Digital frame synchronizer.

Denoting with P_e the probability of a bit error, the following relationship between P_{MD} and P_e is found:

$$P_{\mathrm{MD}} = \sum_{i=N-\lambda+1}^{N} \binom{N}{i} P_e^i (1-P_e)^{N-i} \tag{42}$$

For low values of P_e, this equation becomes

$$P_{\mathrm{MD}} \cong \binom{N}{N-\lambda+1} P_e^{N-\lambda+1} \tag{43}$$

and indicates that P_{MD} decreases exponentially as λ decreases. At first sight this would suggest taking a small threshold value. Doing so, however, might result in *false alarms*, meaning that random bit patterns might cause threshold crossings. Assuming independent and equiprobable bits, the *probability of false alarm* P_{FA} is found to be

$$P_{\mathrm{FA}} = \frac{1}{2^N} \sum_{i=0}^{N-\lambda} \binom{N}{i} \tag{44}$$

Since P_{FA} *increases* as λ decreases, the value of λ must be chosen as a tradeoff between contrasting requirements about P_{MD} and P_{FA}. Note that P_{MD} and P_{FA} both decrease exponentially with the UW length N. Practical values of N are on the order of a few tens.

BIOGRAPHIES

Marco Luise is a full professor of telecommunications at the University of Pisa, Italy. He received M.S. and Ph.D. degrees in electronic engineering from the University of Pisa. In the past, he was a research fellow of the European Space Agency (ESA) at the European Space Research and Technology Centre (ESTEC), Noordwijk, the Netherlands, and a research scientist of the Italian National Research Council (CNR), at the Centro Studio Metodi Dispositivi Radiotrasmissioni (CSMDR), Pisa, Italy. Professor Luise cochaired four editions of the Tyrrhenian International Workshop on Digital Communications, and in 1998 was the General Chairman of the URSI Symposium ISSSE '98. He has been the Technical Chairman of the 7th International Workshop on Digital Signal Processing Techniques for Space Communications and of the Conference European Wireless 2002. As a Senior Member of the IEEE, he served as editor for *Synchronization of the IEEE Transactions on Communications*, and is currently editor for *Communications Theory of the European Transactions on Telecommunications*. His main research interests lie in the broad area of wireless communications, with particular emphasis on CDMA systems and satellite communications.

Umberto Mengali received his training in electrical engineering from the University of Pisa, Italy, where he received his degree in 1961. In 1971 he got the Libera Docenza in Telecommunications from the Italian Education Ministry and in 1975 was made a professor of electrical engineering in the Department of Information Engineering of the University of Pisa. In 1994 he was a Visiting Professor at the University of Canterbury, New Zealand, as an Erskine fellow. His research interests are in the

area of digital communications and communication theory, with emphasis on synchronization methods and modulation techniques. He has published over 80 journal papers and has coauthored the book *Synchronization Techniques for Digital Receivers* (Plenum Press, 1997). Professor Mengali has been an editor of the *IEEE Transactions on Communications* and of the *European Transactions on Telecommunications*. He is an IEEE fellow and is listed in *American Men and Women in Science*.

Michele Morelli was born in Pisa, Italy, in 1965. He received the Laurea (cum laude) in electrical engineering and the "Premio di Laurea SIP" from the University of Pisa, Italy, in 1991 and 1992, respectively. From 1992 to 1995 he was with the Department of Information Engineering of the University of Pisa, where he received a Ph.D. degree in electrical engineering. He is currently a research fellow at the Centro Studi Metodi e Dispositivi per Radiotrasmissioni of the Italian National Research Council (CNR) in Pisa. His interests are in digital communication theory, with emphasis on synchronization algorithms for CDMA and multicarrier systems.

BIBLIOGRAPHY

1. U. Mengali and A. N. D'Andrea, *Synchronization Techniques for Digital Receivers*, Plenum Press, New York, 1997.
2. H. Meyr, M. Moeneclaey, and S. A. Fechtel, *Digital Communication Receivers*, Wiley, New York, 1998.
3. S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
4. G. Asheid and H. Meyr, Cycle slips in phase-locked loops: A tutorial survey, *IEEE Trans. Commun.* 2228–2241 (Oct. 1982).
5. D. C. Rife and R. R. Boorstyn, Single-tone parameter estimation from discrete-time observation, *IEEE Trans. Inform. Theory* 591–598 (Sept. 1974).
6. S. A. Tretter, Estimating the frequency of a noisy sinusoid by linear regression, *IEEE Trans. Inform. Theory* 832–835 (Nov. 1985).
7. S. M. Kay, A fast and accurate single frequency estimator, *IEEE Trans. Acoust. Speech Signal Process.* 1987–1990 (Dec. 1989).
8. M. P. Fitz, Further results in the fast estimation of a single frequency, *IEEE Trans. Commun.* 862–864 (March 1994).
9. M. Luise and R. Reggiannini, Carrier frequency recovery in all-digital modems for burst-mode transmissions, *IEEE Trans. Commun.* 1169–1178 (March 1995).
10. U. Mengali and M. Morelli, Data-aided frequency estimation for burst digital transmission, *IEEE Trans. Commun.* 23–25 (Jan. 1997).
11. M. Morelli and U. Mengali, Feedforward frequency estimation for PSK: A tutorial review, *Eur. Trans. Telecommun.* 103–116 (March/April 1998).
12. M. Morelli and U. Mengali, Carrier frequency estimation for transmission over selective channels, *IEEE Trans. Commun.* 1580–1589 (Sept. 2000).
13. M. G. Hebley and D. P. Taylor, The effect of diversity on a burst-mode carrier-frequency estimator in the frequency-selective multipath channel, *IEEE Trans. Commun.* **46**: 553–560 (April 1998).
14. A. N. D'Andrea and U. Mengali, Noise performance of two frequency-error detectors derived from maximum likelihood estimation methods, *IEEE Trans. Commun.* 793–802 (Feb./March/April 1994).
15. F. M. Gardner, *Demodulator Reference Recovery Techniques Suited for Digital Implementation*, European Space Agency, Final Report, ESTEC Contract 6847/86/NL/DG, Aug. 1988.
16. T. Alberty and V. Hespelt, A new pattern jitter-free frequency error detector, *IEEE Trans. Commun.* 159–163 (Feb. 1989).
17. K. E. Scott and E. B. Olasz, Simultaneous clock phase and frequency offset estimation, *IEEE Trans. Commun.* 2263–2270 (July 1995).
18. A. J. Viterbi and A. M. Viterbi, Nonlinear estimation of PSK-modulated carrier phase with application to burst digital transmission, *IEEE Trans. Inform. Theory* 543–551 (July 1983).
19. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1989.
20. M. Moeneclaey and G. de Jonghe, ML-oriented NDA carrier synchronization for general rotationally symmetric signal constellations, *IEEE Trans. Commun.* 2531–2533 (Aug. 1994).
21. F. M. Gardner, *Phaselock Techniques*, 2nd ed., Wiley, New York, 1979.
22. A. Polydoros, R. Raheli, and C-K. Tzou, Per survivor processing: A general approach to MLSE in uncertain environments, *IEEE Trans. Commun.* 354–364 (Feb./March/April 1995).
23. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding, *IEEE Trans. Commun.* 1261–1271 (Oct. 1996).
24. L. Erup, F. M. Gardner, and R. A. Harris, Interpolation in digital modems — Part II: Implementation and performance, *IEEE Trans. Commun.* 998–1008 (June 1993).
25. K. H. Mueller and M. Mueller, Timing recovery in digital synchronous data receivers, *IEEE Trans. Commun.* 516–531 (May 1976).
26. F. M. Gardner, A BPSK/QPSK timing-error detector for sampled receivers, *IEEE Trans. Commun.* 423–429 (May 1986).
27. W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
28. M. Oerder and H. Meyr, Digital filter and square timing recovery, *IEEE Trans. Commun.* 605–611 (May 1988).
29. P. Godard, Passband timing recovery in all-digital modem receiver, *IEEE Trans. Commun.* 517–523 (May 1978).
30. G. Ungerboeck, Fractional tap-spacing equalizer and its consequences for clock recovery in data modems, *IEEE Trans. Commun.* 856–864 (Aug. 1976).
31. M. K. Simon, Nonlinear analysis of an absolute value type of early-late-gate bit synchronizer, *IEEE Trans. Commun. Technol.* 589–596 (Oct. 1970).
32. R. De Gaudenzi, F. Giannetti, and M. Luise, Signal synchronization for direct-sequence code-division multiple-access radio modems, *Eur. Trans. Telecommun.* 73–89 (Jan./Feb. 1998).
33. S. E. Bensley and B. Aazhang, Maximum likelihood synchronization of a single-user for code-division multiple-access communication systems, *IEEE Trans. Commun.* 392–399 (March 1998).

34. R. de Gaudenzi, M. Luise, and R. Viola, A digital chip timing recovery loop for band-limited direct-sequence spread-spectrum signals, *IEEE Trans. Commun.* **COM-41**(11): (Nov. 1993).

35. M. Moeneclaey and G. De Jonghe, Tracking performance of digital chip synchronization algorithms for bandlimited direct-sequence spread-spectrum communications, *IEE Electron. Lett.* 1147–1149 (June 1991).

36. S. Verdú, *Multiuser Detection*, Cambridge Univ. Press, 1998.

37. R. A. Iltis and L. Mailaender, An adaptive multiuser detector with joint amplitude and delay estimation, *IEEE J. Select. Areas Commun.* **12**: 774–785 (June 1994).

38. E. G. Strom and F. Malmsten, A maximum likelihood approach for estimating DS-CDMA multipath fading channels, *IEEE J. Select. Areas Commun.* **18**: 132–140 (Jan. 2000).

39. V. Tripathi, A. Montravadi, and V. V. Veeravalli, Channel acquisition for wideband CDMA signals, *IEEE J. Select. Areas Commun.* **18**: 1483–1494 (Aug. 2000).

40. H. Sari, G. Karam, and J. Janclaude, Transmission techniques for digital terrestrial TV broadcasting, *IEEE Commun. Mag.* **36**: 100–109 (Feb. 1995).

41. J.-J. van de Beek, M. Sandell, and P. O. Borjesson, ML estimation of time and frequency offset in OFDM systems, *IEEE Trans. Signal Process.* 1800–1805 (July 1997).

42. T. M. Schmidl, and D. C. Cox, Robust frequency and timing synchronization for OFDM, *IEEE Trans. Commun.* 1613–1621 (Dec. 1997).

43. F. Daffara and O. Adami, A novel carrier recovery technique for orthogonal multicarrier systems, *Eur. Trans. Telecommun.* 323–334 (July–Aug. 1996).

44. M. Morelli and U. Mengali, Feedback frequency synchronization for OFDM applications, *IEEE Commun. Lett.* 28–30 (Jan. 2001).

45. H. Sari and G. Karam, Orthogonal frequency-division multiple access and its application to CATV networks, *Eur. Trans. Telecommun.* 507–516 (Dec. 1998).

46. S. W. Golomb and R. A. Scholtz, Generalized barker sequences, *IEEE Trans. Inform. Theory* **IT-11**: 533–537 (Oct. 1965).

47. W. W. Wu, *Elements of Digital Satellite Communications*, Comp. Science Press, Rockville, MD, 1984.

SYNCHRONOUS OPTICAL NETWORK (SONET) AND SYNCHRONOUS DIGITAL HIERARCHY (SDH)

ROGER FREEMAN*
Independent Consultant
Scottsdale, Arizona

1. BACKGROUND AND INTRODUCTION

SONET and SDH are similar digital transport formats that were developed for the specific purpose of providing a reliable and versatile digital structure to take advantage of the higher bit rate capacity of optical fiber. SONET is an acronym standing for *synchronous optical network*. In a similar vein, SDH stands for *synchronous digital hierarchy*. We could say that SONET has a North American flavor, and SDH has a European flavor. This may be stretching the point, because the two systems are very similar.

The original concept in developing a digital format for high-bit-rate capacity optical systems was to have just one singular standard for worldwide application. This did not work out. The United States wanted the basic bit rate to accommodate DS3 [around 50 Mbps (megabits per second)]. The Europeans had no bit rates near this value and were opting for a starting rate around 150 Mbps. Another difference surfaced for framing alternatives. The United States perspective was based on a frame of 13-row $\times$180-byte columns for the 150 Mbps rate reflecting what is now called STS-3 structure. Europe advocated a 9-row $\times$270-byte column STS-3 frame to efficiently transport the E1 signal (2.048 Mbps) using 4 columns of 9 bytes, based on 32 bytes/125 μs.

The ANSI T1X1 committee approved a final standard in August 1988, with CCITT following suit, and a global SONET/SDH standard was established. This global standard was based on a 9-row frame, wherein SONET became a subset of SDH [1].

Both SONET and SDH use basic building-block techniques. As we mentioned above, SONET starts at a lower bit rate, at 51.84 Mbps. This basic rate is called STS-1 (synchronous transport signal layer 1). Lower-rate payloads are mapped into the STS-1 format, while higher rate signals are obtained by byte interleaving N frame-aligned STS-1s to create an STS-N signal. Such a simple multiplexing approach results in no additional overhead; as a consequence the transmission rate of an STS-N signal is exactly $N \times 51.84$ Mbps where N is currently defined for the values: 1, 3, 12, 24, 48, and 192 [2].

The basic building block of SDH is the synchronous transport module level 1 (STM-1) with a bit rate of 155.52 Mbps. Lower-rate payloads are mapped into STM-1, and higher-rate signals are generated by synchronously multiplexing N STM-1 signals to form the STM-N signal. Transport overhead of an STM-N signal is N times the transport overhead of an STM-1, and the transmission rate is $N \times 155.52$ Mbps. Presently, only STM-1, STM-4, STM-16, and STM-64 have been defined by the ITU-T organization [5].

Both with SONET and SDH, the frame rate is 8000 per second, resulting in a 125-μs frame period. There is high compatibility between SONET and SDH. Because of the different basic building-block size, they differ in structure: 51.84 Mbps for SONET and 155.52 Mbps for SONET. However, if we multiply the SONET rate by

* Roger Freeman took an early retirement from the Raytheon Company, Equipment Division, in 1991 where he was Principal Engineer to establish *Roger Freeman Associates*, Independent Consultants in Telecommunications. He has been writing books on various telecommunication disciplines for John Wiley & sons, Inc. since 1973. Roger has seven titles which he keeps current including *Reference Manual for Telecommunication Engineers* now in 3rd edition. His website is *www.rogerfreeman.com* and his email address is rogerf67@cox.net.

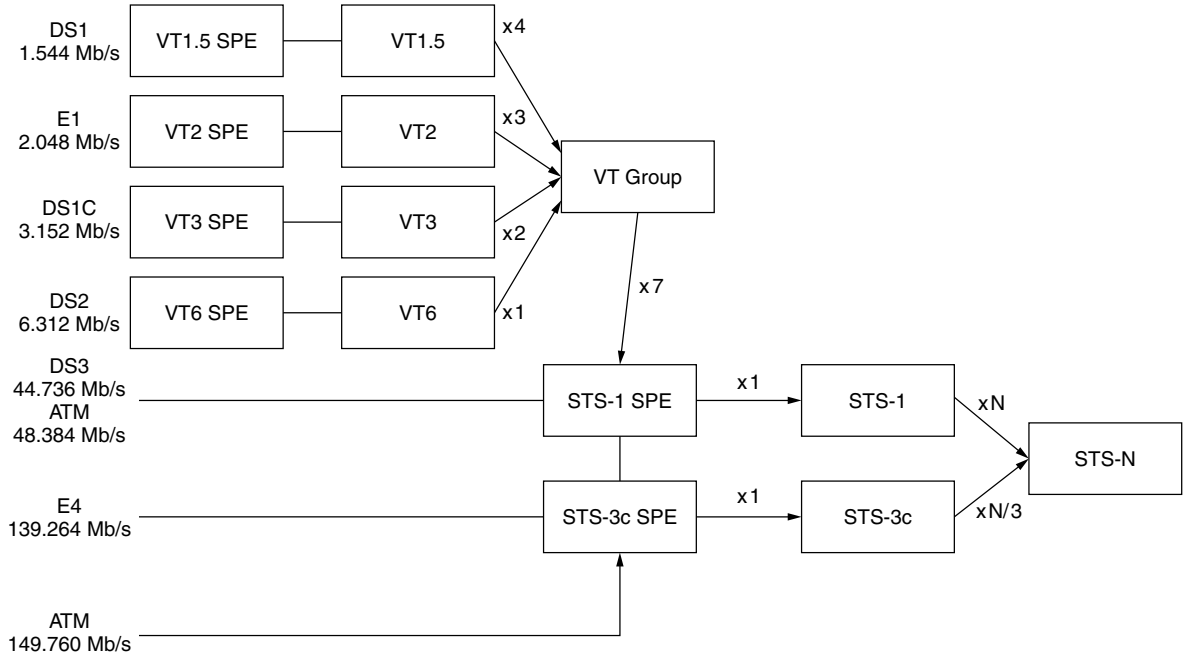


Figure 1.1. SONET multiplexing structure. (From Fig. 3, p. 4, C. A. Siller and M. Shafi, eds., *Synchronous Optical Network, Synchronous Digital Hierarchy: An Overview of Synchronous Network*, IEEE Press, New York, 1996 [3].)

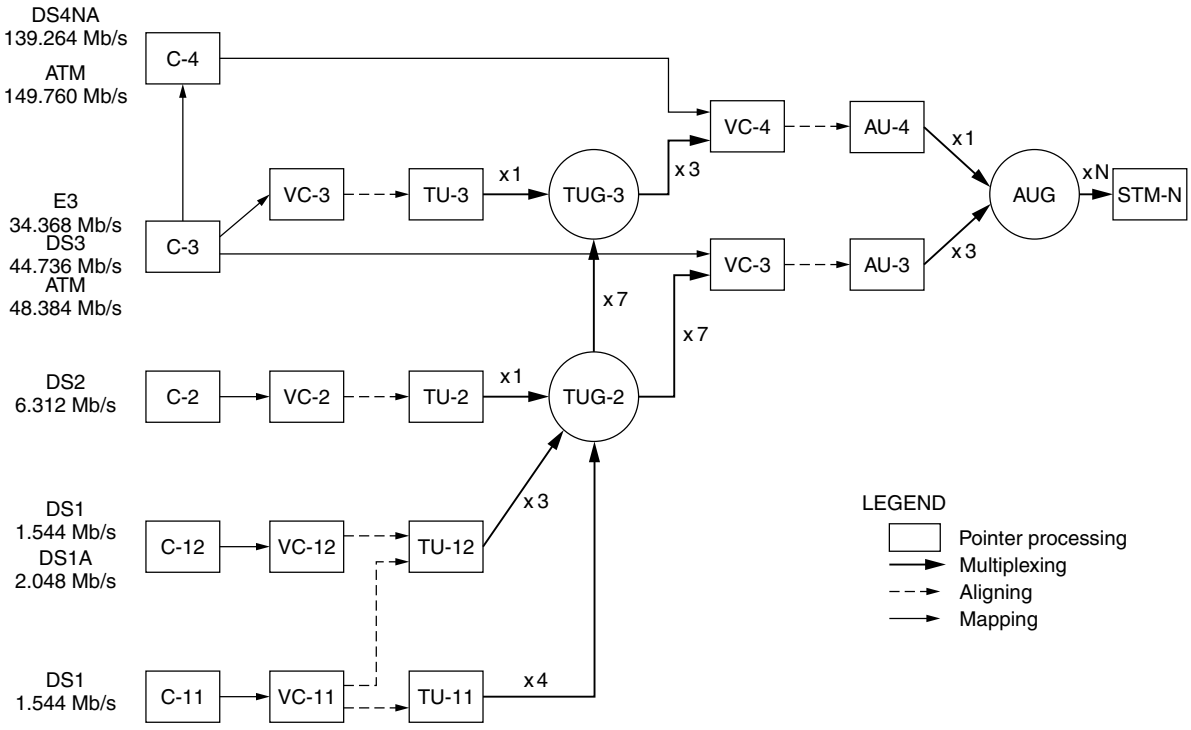


Figure 1.2. SDH multiplexing structure. (From Fig. 4, p. 4, Siller and Shafi [3] as Fig. 8-1.)

3, developing the STS-3 signal, we do indeed have the SDH initial bit rate of 155.52 Mbps. Figures 1.1 and 1.2 compare the multiplexing structures of each. Table 1.1 lists and compares the bit rates of each standard.

Besides differing in basic building-block bit rates, SONET and SDH differ with respect to overhead usage. These overhead differences have been grouped into two broad categories: format definitions and usage interpretation. As a result, we have opted to segregate our descriptions of each.

We must dispel a concept that has developed by the unfortunate use of words and terminology. Some believe

Table 1.1. SONET and SDH Transmission Rates

SONET Optical Carrier Level OC-N	SONET Electrical Level STS-N	Equivalent SDH STM-N	Line Rate (Mbps)
OC-1	STS-1	—	51.84
OC-3	STS-3	STM-1	155.52
OC-12	STS-12	STM-4	622.08
OC-24	STS-24	—	1244.16
OC-48	STS-48	STM-16	2488.32
OC-192	STS-192	STM-64	9953.28
OC-768	STS-768	STM-256	39,813,120[a]

[a] On the drawing boards as of this writing [2–5].

because SONET stands for *synchronous optical network*, it will operate only on optical fiber lightguide. This is patently incorrect. Any transport medium that can provide the necessary bandwidth (measured in hertz, as one would expect), will transport the requisite SONET or SDH line rates. For example, digital loss-of-signal (LoS) microwave, using heavy bit packing modulation schemes, readily transports 622 Mbps (STS-12 or STM-4) per carrier at the higher frequencies using a 40-MHz assigned bandwidth.

The objective of this article is to provide an overview of these two standards and some of their challenging innovations that make them interesting, such as the payload pointer. Section 2 of this article deals with SONET, the North American standard, and Section 3 covers SDH. Section 4 presents a summary of the two standards in tabular form.

It should be noted that we keep custom set in this work that ITU recommendations will be identified by the CCITT or CCIR logo if that document was issued prior to January 1, 1993. If the document was issued after that date, where it originated from the Telecommunications Standardization Sector of the ITU, it will be titled ITU-T Recommendation XYZ or from the Radiocommunications Sector, ITU-R XYZ.

2. SYNCHRONOUS OPTICAL NETWORK (SONET)

2.1. Synchronous Signal Structure

SONET is based on a synchronous digital signal comprised of 8-bit octets, which are organized into a frame structure. The frame can be represented by a two-dimensional map comprising N rows and M columns, where each box so derived contains one octet (or byte). The upper left-hand corner of the rectangular map representing a frame contains an identifiable marker to tell the receiver if is the start of frame.

SONET consists of a basic, first-level structure known as STS-1, which is discussed in the following paragraphs. The definition of the first level also defines the entire hierarchy of SONET signals because higher-level SONET signals are obtained by synchronously multiplexing the lower-level modules. When lower-level modules are multiplexed together, the result is denoted STS-N (STS stands for synchronous transport signal), where N is an integer. The resulting format can be converted to an OC-N (OC stands for optical carrier) or STS-N electrical signal. There is an integer multiple relationship between the rate of the basic module STS-1 and the OC-N electrical equivalent signals (i.e., the rate of an OC-N is equal to N times the rate of an STS-1). Only OC-1, OC-3, OC-12, OC-24, OC-48, and OC-192 are supported by today's SONET.

2.1.1. The Basic Building Block. The STS-1 frame is shown in Fig. 2.1. STS-1 is the basic module and building block of SONET. It is a specific sequence of 810 octets (6480 bits) that includes various overhead octets and an envelope capacity for transporting payloads.[1] STS-1 is depicted as a 90-column, 9-row structure. With a frame period of 125 μs (i.e., 8000 frames per second). STS-1 has a bit rate of 51.840 Mbps. Consider Fig. 2.1, where the order of transmission is row-by-row, from left to right. In each octet of STS-1 the most significant bit (MSB) is transmitted first.

As illustrated in Fig. 2.1, the first three columns of the STS-1 frame contain the transport overhead. These three columns have 27 octets (i.e., 9×3), 9 of which are used for the *section overhead*, with 18 octets containing the *line*

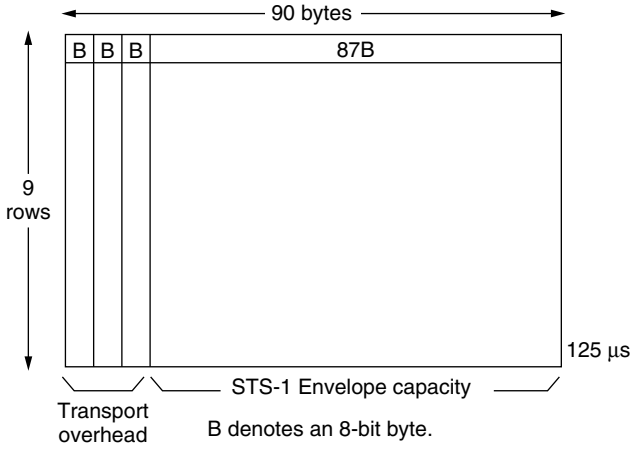


Figure 2.1. The STS-1 frame.

[1] The several reference publications use the term *byte*, meaning, in this context, an 8-bit sequence. We prefer the term *octet*. The reason is that some argue that byte is ambiguous, having conflicting definitions.

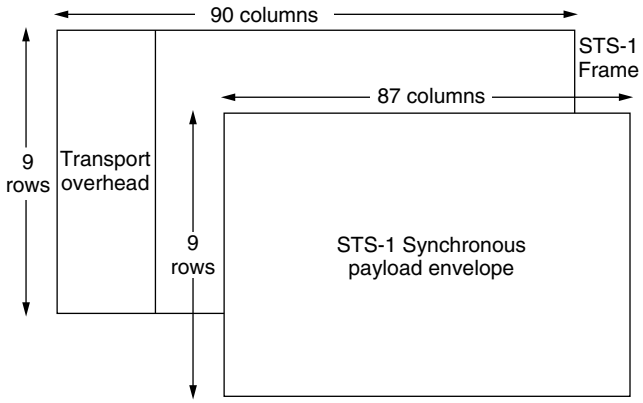


Figure 2.2. STS-1 synchronous payload envelope (SPE).

overhead. The remaining 87 columns make up the STS-1 envelope capacity, as shown in Fig. 2.2.

The STS-1 synchronous payload envelope (SPE) occupies the STS-1 envelope capacity. The STS-1 SPE consists of 783 octets and is depicted as an 87-column × 9-row structure. In that structure, column 1 contains 9 octets and is designated as the *STS path overhead* (POH). In the SPE, columns 30 and 59 are not used for payload but are designated *fixed-stuff* columns and are undefined. However, the values used as "stuff" in columns 30 and 59 of each STS-1 SPE will produce even parity when calculating BIP-8 of the STS-1 path BIP (bit-interleaved parity) value. The POH column and fixed stuff columns are shown in Fig. 2.3. The 756 octets in the remaining 84 columns are used for the actual STS-1 payload capacity.

The STS-1 SPE may begin anywhere in the STS-1 envelope capacity. Typically, the SPE begins in one STS-1 frame and ends in the next. This is illustrated in Fig. 2.4. However, on occasion, the SPE may be wholly contained in one frame. The *STS payload pointer* resides in the transport overhead. It designates the location of the next octet where the SPE begins. Payload pointers are described in the following paragraphs.

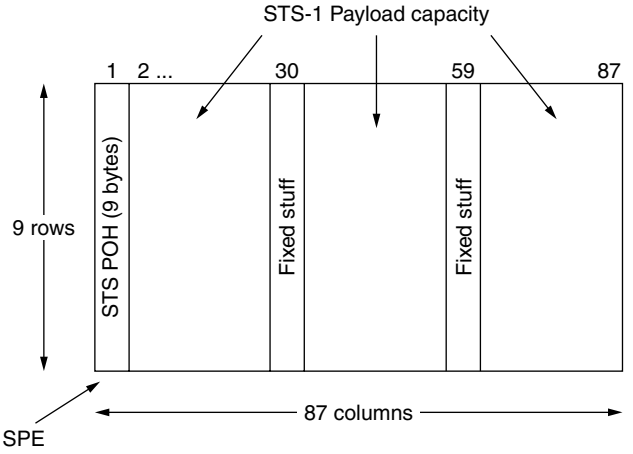


Figure 2.3. Path overhead (POH) and the STS-1 payload capacity within the STS-1 SPE. Note that the net payload capacity of the STS-1 is only 84 columns.

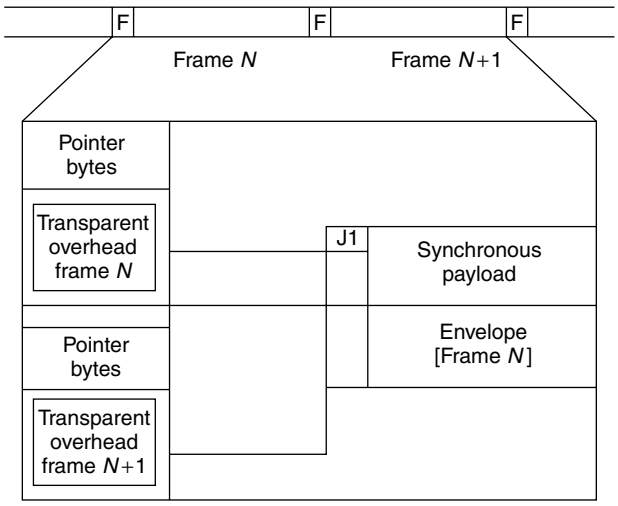


Figure 2.4. STS-1 SPE typically located in STS-1 frames. (Figure courtesy of Agilent Technologies [7].)

The STS POH is associated with each payload and is used to communicate various pieces of information from the point where the payload is mapped into the STS-1 SPE to the point where it is delivered. Among the pieces of information carried in the POH are alarm and performance data [6].

2.1.2. STS-N Frames. Figure 2.5 illustrates the structure of an STS-N frame. The frame consists of a specific sequence of $N \times 810$ octets. The STS-N frame formed by octet-interleaved STS-1 and STS-M modules ($< N$). The transport overhead of the associated STS SPEs are not required to be aligned because each STS-1 has a payload pointer to indicate the location of the SPE or to indicate concatenation.

2.1.3. STS Concatenation. Superrate payloads require multiple STS-1 SPEs. FDDI and some B-ISDN payloads fall into this category. Concatentation means the linking together. An STS-Nc module is formed by linking N constituent STS-1s together in a fixed-phase alignment. The superrate payload is then mapped into the resulting STS-Nc SPE for transport. Such STS-Nc SPE requires an OC-N or STS-N electrical signal. Concatenation indicators contained in the second through the Nth STS payload pointer are used to show that the STS-1s of an STS-Nc are linked together.

The are $N \times 783$ octets in an STS-Nc. Such an NTS-Nc arrangement is illustrated in Fig. 2.6 and is depicted as a $N \times 87$column × 9-row structure. Because of the linkage, only one set of STS POHs is required in the STS-Nc SPE. Here the STS POH always appears in the first of the N STS-1s that make up the STS-Nc [10].

Figure 2.7 shows the assignment of transport overhead of an OC-3 carrying an STS-3c SPE.

2.1.4. Structure of Virtual Tributaries (VTs). The SONET STS-1 SPE with a channel capacity of 50.11 Mbps has been designed specifically to transport a DS3 tributary signal. To accommodate sub-STS-1 rate payloads such as

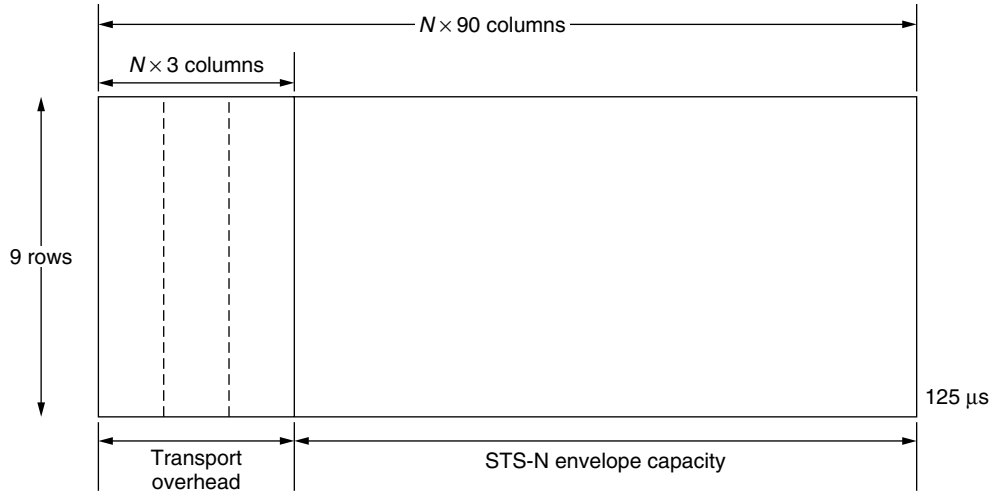


Figure 2.5. STS-N frame.

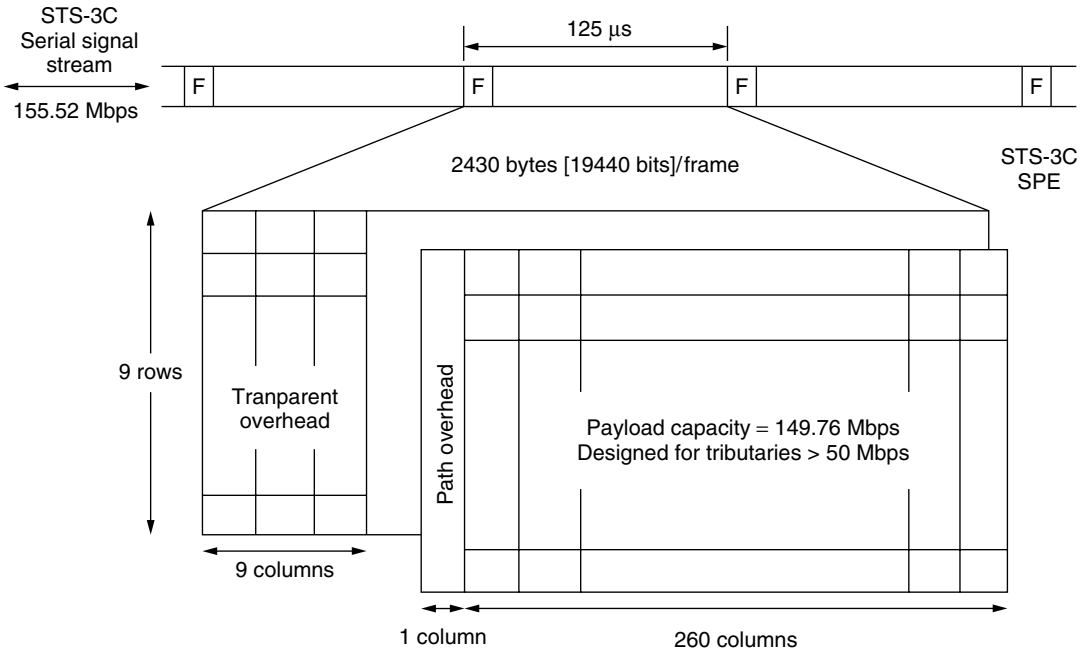


Figure 2.6. STS-3c concatenated SPE. (Courtesy of Agilent Technologies [7].)

DS1, the VT structure is used. It consists of four sizes: VT1.5 (1.728 Mbps) for DS1 transport, VT2 (2.304 Mbps) for E1 transport, VT3 (3.456 Mbps) for DS1C transport, and VT6 (6.912 Mbps) for DS2 transport. The virtual tributary concept is illustrated in Fig. 2.8. The four VT configurations are illustrated in Fig. 2.9. In the 87-column × 9-row structure of the STS-1 SPE, the VTs occupy 3, 4, 6, and 12 columns respectively [6,7].

2.2. Payload Pointer

The STS payload pointer provides a method for allowing flexible and dynamic alignment of the STS SPE within the STS envelope capacity, independent of the actual contents of the SPE. SONET, by definition, is intended to be synchronous. It derives its timing from the master network clock.

Modern digital networks must make provision for more than one master clock. Examples in the United States are the several interexchange carriers which interface with local exchange carriers (LECs), each with its own master clock. Each master clock (stratum 1) operates independently. And each of these master clocks has excellent stability (i.e., better than 1×10^{-11} per month), yet there may be some small variance in time among the clocks. Assuredly they will not be phase-aligned. Likewise, SONET must take into account loss of master clock or a segment of its timing delivery system. In this case, network switches fall back on lower-stability internal

Section overhead	A1 framing	A1 framing	A1 framing	A2 framing	A2 framing	A2 framing	J0 Section trace	Z0 Section growth	Z0 Section growth
	B1 BIP-8 parity (1)	unspecified	unspecified	E1 orderwire	unspecified	unspecified	F1 user option	unspecified	unspecified
	D1 section data com channel	unspecified	unspecified	D2 data channel	unspecified	unspecified	D3 data com channel	unspecified	unspecified
	H1 pointer	H1* pointer	H1* pointer	H2 pointer	H2* pointer	H2* pointer	H3 pointer action byte	H3 pointer action byte	H3 pointer action byte
	B2 BIP-8 parity (1)	B2 BIP-8 parity (1)	B2 BIP-8 parity (1)	K1 automatic protection switching	unspecied	unspecied	K2 automatic protection switching	unspecied	unspecied
Line overhead	D4 data channel	unspecified	unspecified	D5 data channel	unspecified	unspecified	D6 data channel	unspecified	unspecified
	D7 data channel	unspecified	unspecified	D8 data channel	unspecified	unspecified	D9 data channel	unspecified	unspecified
	D10 data channel	unspecified	unspecified	D11 data channel	unspecified	unspecified	D12 data channel	unspecified	unspecified
	S1 synchroni-zation	Z1 line growth	Z1 line growth	Z2 line growth	Z2 line growth	M1 REI (2)	E2 express orderwire	unspecified	unspecified

(1) BIP = Bit Interleaved Parity
(2) REI = Remote Error Indication

Each box represents 1 byte (8 bits)

Figure 2-7. STS-3 Transport overhead assignment. *Asterisk indicates (Based on Section 8.2, ANSI T.1.105-1995, Ref. 1).

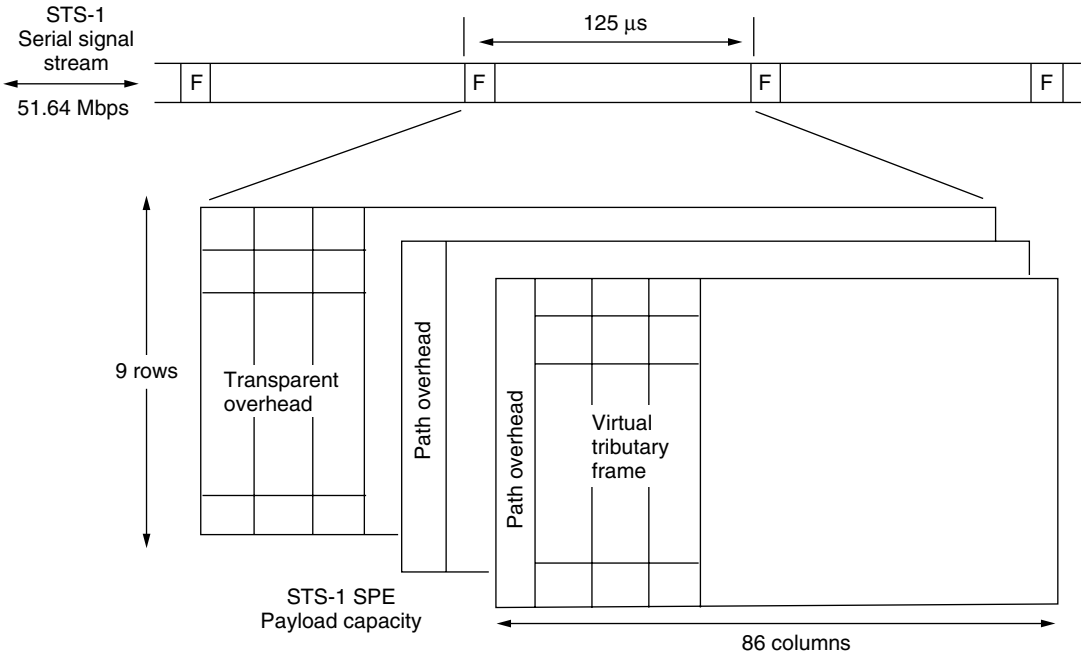


Figure 2.8. The virtual tributary (VT) concept. (From Ref. 7, Courtesy of Agilent Technologies.)

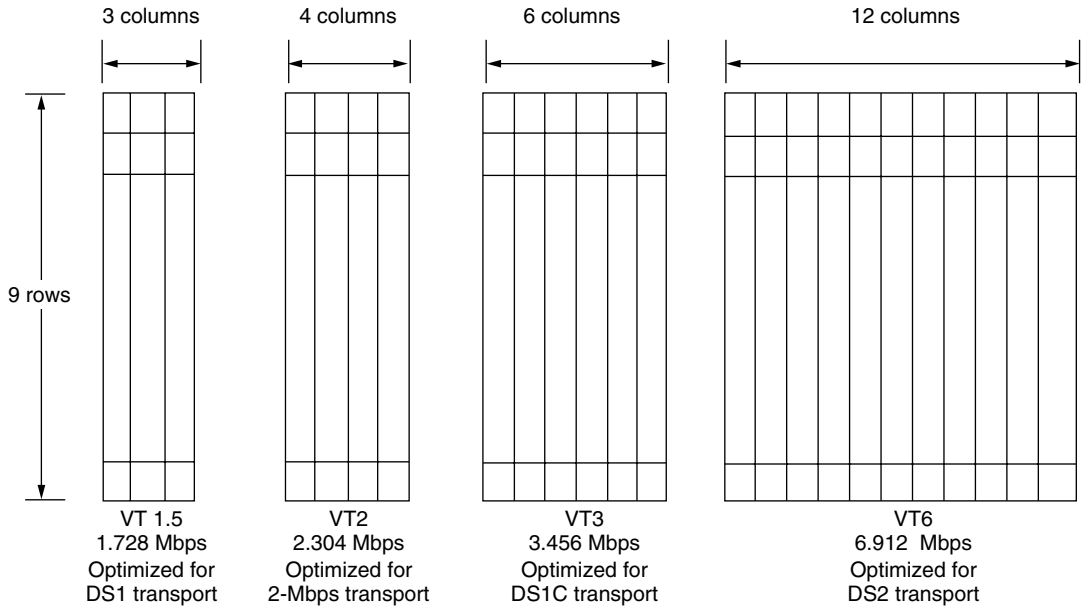


Figure 2.9. The four sizes of virtual tributary frames. (From Ref. 7, Courtesy of Agilent Technologies.)

clocks.[2] The situation must be handled by SONET. Therefore, synchronous transport is required to operate effectively under these conditions, where network nodes are operating at slightly difference rates [4].

To accommodate these clock offsets, the SPE can be moved (justified) in the positive or negative direction one octet at a time with respect to the transport frame. This is accomplished by recalculating or updating the payload pointer at each SONET network node. In addition to clock offsets, updating the payload pointer also accommodates any other timing-phase adjustments required between the input SONET signals and the timing reference at the SONET node. This is what is meant by *dynamic alignment*, where the STS SPE is allowed to float within the STS envelope capacity.

The payload pointer is contained in the H1 and H2 octets in the line overhead (LOH) and designates the location of the octet where the STS SPE begins. These

[2] It is general practice in digital networks that switches provide timing supply for transmission facilities.

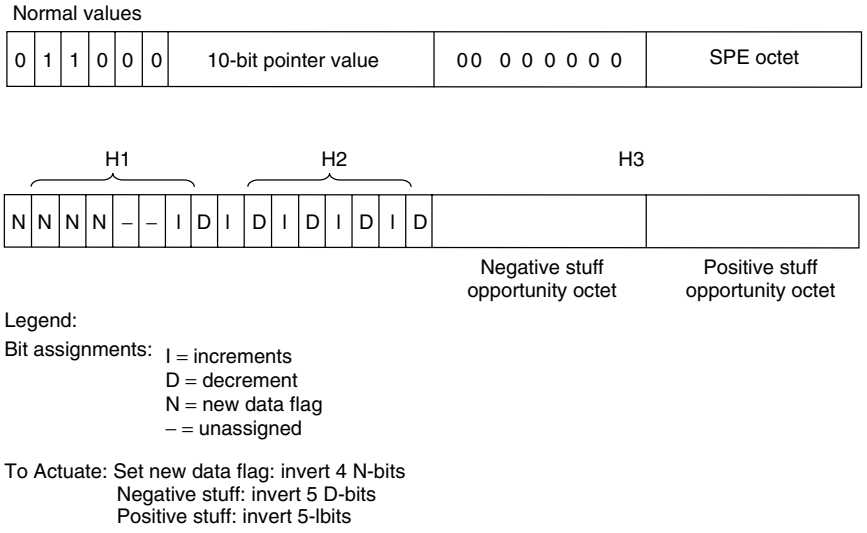


Figure 2.10. STS payload pointer (H1, H2) coding.

two octets are illustrated in Fig. 2.10. Bits 1 through 4 of the pointer word carry the *new data flag* (NDF), and bits 7 through 16 carry the pointer value. Bits 5 and 6 are undefined.

Let us discuss bits 7 through 16, the pointer value. This is a binary number with a range of 0 to 782. It indicates the offset of the pointer word and the first octet of the STS SPE (i.e., the J1 octet). The transport overhead octets are not counted in the offset. For example, a pointer value of 0 indicates that the STS SPE starts in the octet location that immediately follows the H3 octet, whereas an offset of 87 indicates that it starts immediately after the K2 octet location. These overhead octets are shown in Fig. 2.7.

Payload pointer processing introduces a signal impairment known as *payload adjustment jitter*. This impairment appears on a received tributary signal after recovery from a SPE that has been subjected to payload pointer changes. The operation of the network equipment processing the tributary signal immediately downstream is influenced by this excessive jitter. By careful design of the timing distribution for the synchronous network, payload jitter adjustments can be minimized, thus reducing the level of tributary jitter that can be accumulated through synchronous transport.

2.3. The Three Overhead Levels of SONET

The three embedded overhead levels of SONET are

1. Path (POH)
2. Line (LOH)
3. Section (SOH)

These overhead levels, represented as spans, are illustrated in Fig. 2.11. One important function carried out by this overhead is the support of network operation, administration, and maintenance (OA&M),

The path overhead (POH) consists of 9 octets and occupies the first column of the SPE, as pointed out previously. It is created by and included in the SPE as part of the SPE assembly process. The POH provides the facilities to support and maintain the transport of the SPE between path terminations, where the SPE is assembled and disassembled. Among the POH specific functions are

- An 8-bit wide (octet B3) BIP (bit-interleaved parity) check calculated over all bits of the previous SPE. The computed value is placed in the POH of the following frame.

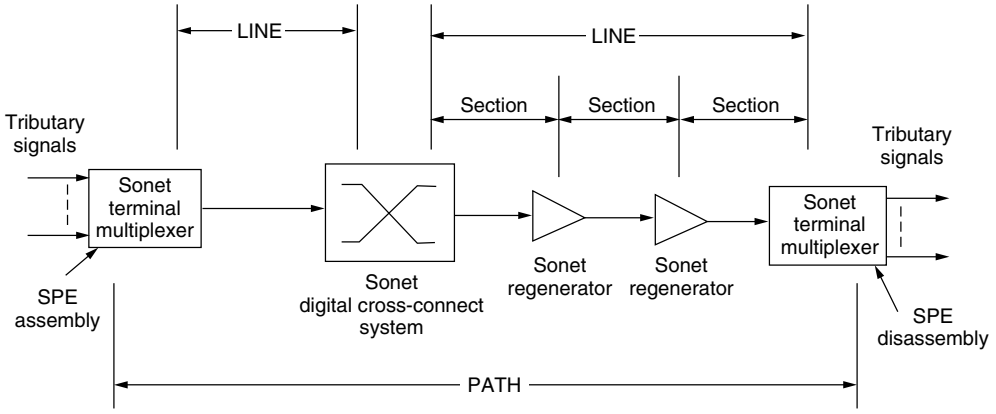


Figure 2.11. SONET section, line, and path definitions.

- Alarm and performance information (octet G1).
- A path signal label (octet C2); gives details of SPE structure. It is 8 bits wide, which can identify up to 256 structures (2^8).
- One octet (J1) repeated through 64 frames can develop an alphanumeric message associated with the path. This allows verification of continuity of connection to the source of the path signal at any receiving terminal along the path by monitoring the message string.
- An orderwire for network operator communications between path equipment (octet F2).

Facilities to support and maintain the transport of the SPE between adjacent nodes are provided by the line and section overhead. These two overhead groups share the first three columns of the STS-1 frame. The SOH occupies the top three rows (total of 9 octets, and the LOH occupies the bottom 6 rows (18 octets).

The line overhead functions include:

- Payload pointer (octets H1, H2, and H3) (each STS-1 in an STS-N frame has its own payload pointer)
- Automatic protection switching control (octets K1 and K2)
- BIP parity check (octet B2)
- 576-kbps data channel (octets D4–D12)
- Express orderwire (octet E2)

A *section* is defined in Fig. 2.11. Section overhead functions include [6,7]

- Frame alignment pattern (octets A1 and A2)
- STS-1 identification (octet C1): a binary number corresponding to the order of appearance in the STS-N frame, which can be used in the framing and deinterleaving process to determine the position of other signals
- BIP-8 parity check (octet B1): section error monitoring
- Data communications channel (octets D1, D2, and D3)
- Local orderwire channel (octet E1)
- User channel (octet F1)

2.4. SPE Assembly–Disassembly Process

Payload mapping is the process of assembling a tributary signal into an SPE. It is fundamental to SONET operation. The payload capacity provided for each individual tributary signal is always slightly greater than that required by that tributary signal. The mapping process, in essence, is to synchronize the tributary signal with the payload capacity. This is achieved by adding stuffing bits to the bitstream as part of the mapping process.

An example might be a DS3 tributary signal at a nominal bit rate of 44.736 of the 49.54 Mbps provided by an STS-1 SPE. The addition of path overhead completes the assembly process of the STS-1 SPE and increases the bit rate of the composite signal to 50.11 Mbps. The SPE assembly process is shown graphically in Fig. 2.12. At the terminus or drop point of the network, the original DS3 payload must be recovered, as in our example. The process of SPE disassembly is shown in Fig. 2.13. The term used in this case is *payload demapping*.

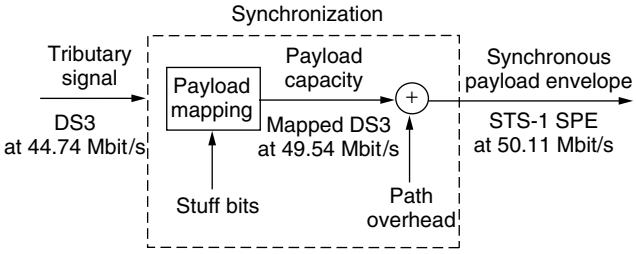


Figure 2.12. The SPE assembly process. (Courtesy of Agilent Technologies [7].)

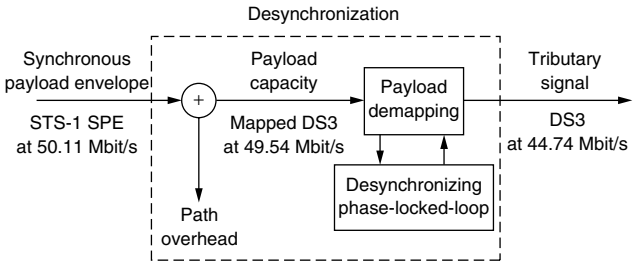


Figure 2.13. The SPE disassembly process. (Courtesy of Agilent Technologies [7].)

The demapping process desynchronizes the tributary signal from the composite SPE signal by stripping off the path overhead and the added stuff bits. In the example, an STS-1 SPE with a mapped DS3 payload arrives at the tributary disassembly location with a signal rate of 50.11 Mbps. The stripping process results in a discontinuous signal representing the transported DS3 signal with an average signal rate of 44.74 Mbps. The timing discontinuities are reduced by means of a desynchronizing phase-locked loop, which then produces a continuous DS3 signal at the required average transmission rate of 44.736 Mbps [1,6,7].

2.5. Add/Drop Multiplex (ADM)

A SONET ADM multiplexes one or more DS-n signals into a SONET OC-N channel. In its converse function, a SONET ADM demultiplexes a SONET STS-n configuration into its component DS-n components to be passed to a user or to be forwarded on a tributary bitstream. An ADM can be configured for either the add/drop or terminal mode. In the ADM mode, it can operate when the low-speed DS1 signals terminating at the SONET derive timing from the same or equivalent source (i.e., synchronous) as the SONET system it interfaces with, but do not derive timing from asynchronous sources.

Figure 2.14 is an example of an ADM configured in the add/drop mode with DS1 and OC-N interfaces. A SONET ADM interfaces with two full-duplex OC-N signals and one or more full-duplex DS1 signals. It may optionally provide low-speed DS1C, DS2, DS3, or OC-M (where $M \leq N$).

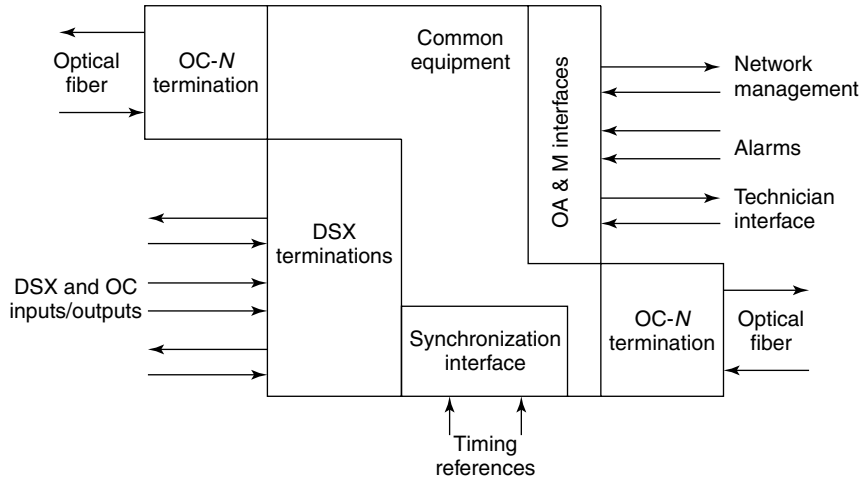


Figure 2.14. SONET ADM add/drop configuration example [2,8].

There are non-path-terminating information payloads from each incoming OC-N signal, which are passed to the SONET ADM and transmitted by the OC-N interface on the other side.

Timing for transmitted OC-N is derived from either an external synchronization source, an incoming OC-N signal, from each incoming OC-N signals in each direction (called *through timing*), or from its local clock, depending on the network application. Each DS1 interface reads data from an incoming OC-N and inserts data into an outgoing OC-N bit stream as required. Figure 2.14 also shows a synchronization interface for local switch application with external timing and an operations interface module (OIM) that provides local technician orderwire,[3] local alarm and an interface to remote operations systems. A controller is part of each SONET ADM, which maintains and controls ADM functions, to connect to local or remote technician interfaces, and to connect to required and optional operations links that permit maintenance, provisioning, and testing.

Figure 2.15 shows an example of an ADM in the terminal mode of operation with DS1 interfaces. In this case, the ADM multiplexes up to NX(28DS1) or equivalent signals into an OC-N bitstream.[4] Timing for this terminal configuration is taken from either an external synchronization source, the received OC-N signal (called *loop timing*) or its own local clock, depending on the network application [8].

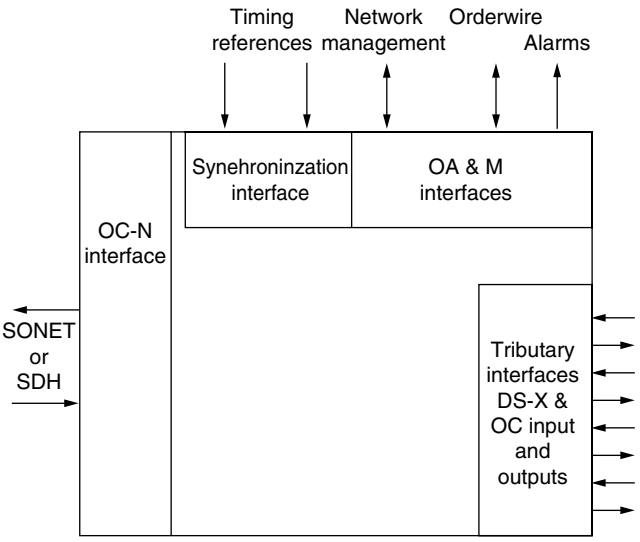


Figure 2.15. A SONET or SDH ADM in a terminal configuration Refs. 2,8, and 12.

2.6. Automatic Protection Switching (APS)

First, we distinguish $1+1$ protection from $N+1$ protection. These two SONET linear APS options are shown in Fig. 2.16. APS can be provided in a linear or ring architecture. SONET NEs (network elements) that have *line termination equipment* (LTE) and terminate optical lines may provide *linear* APS. Support of linear APS at STS-N electrical interfaces is not provided in the relevant Telcordia and ANSI standards.

Linear APS, and in particular, the protocol for the APS channel, is standardized to allow interworking between SONET LTEs from different vendors. Therefore, all the STS SPEs carried in an OC-N signal are protected together. The ANSI and Telcordia standards define two linear APS architectures:

$1+1$
$N+1$ (also called $1:n/1:1$)

The $1+1$ is an architecture in which the headend signal is continuously bridged to working and protection equipment so the same payloads are transmitted identically to the tailend working and protection equipment (see top of Fig. 2.16). At the tailend, the working and protection OC-N signals are monitored independently and identically for failures. The receiving equipment chooses either the working or protection signals as the one from which to select

[3] An *orderwire* is a voice or keyboard–printer–display circuit for coordinating setup and maintenance activities among technicians and supervisors (related word: *service channel*).
[4] This implies a DS3 configuration as it contains 28 DS1s.

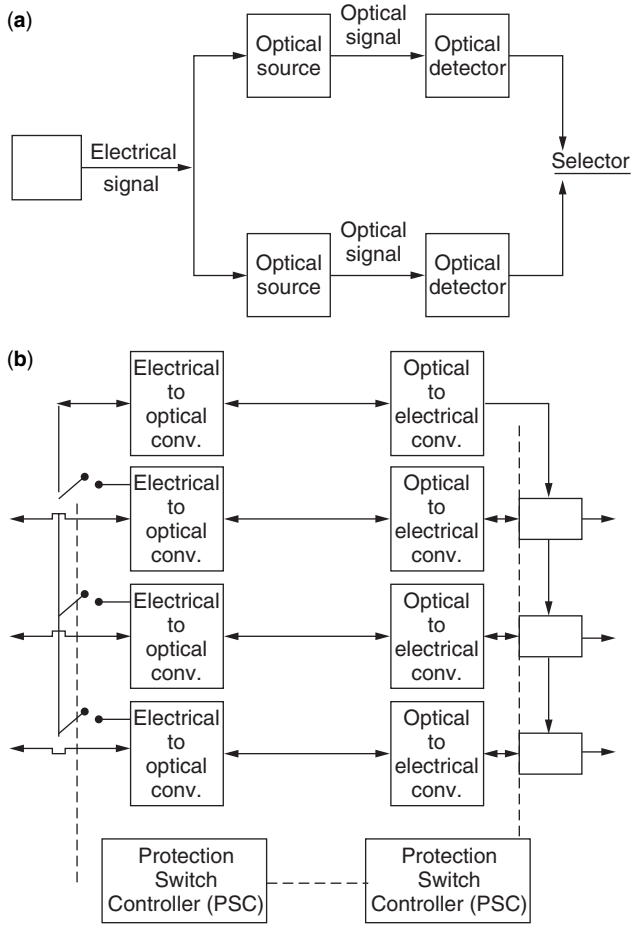


Figure 2.16. (**a**) Linear SONET APS, $1+1$ protection Refs. 1,9, and 12; (**b**) Linear SONET APS, $N+1$ protection Refs. 1,9, and 12.

traffic, based on the switch initiation criteria [e.g., loss of signal (LoS), signal degraded]. Because of the continuous headend bridging, the $1+1$ architecture does not allow an unprotected extra traffic channel to be provided.

To achieve full redundancy, $1+1$ protection is very effective. This type of configuration is widely employed usually with a ring architecture. In the basic configuration of a ring, the traffic from the source is transmitted simultaneously over both bearers and the decision to switch between main and standby is made at the receiving location. In this situation only *loss of signal* or similar indications are required to initiate changeover and no command and control information needs to be passed between the two sites. It is assumed that after the failure in the mainline, a repair crew will restore it to service. Rather than have the repaired line placed back in service as the "mainline," it is designated the new "standby." Thus only one line interruption takes place and the process of repair does not require a second break in service.

The best method of configuring $1+1$ service is to have the standby line geographically distant from the mainline. This minimizes common-mode failures. Because of its simplicity, this approach assures the fastest restoration of service with the least requirement for sophisticated monitoring and control equipment. However, it is more costly and involves less efficient equipment usage than does a $N+1$ approach. It is inefficient in that the standby equipment sits idle nearly all the time, not bringing in any revenue.

The $N+1$ (also denoted as $1:n$ or one for n) protection is an architecture in which any of n working lines can be bridged to a single protection line (see bottom of Fig. 2.16). Permissible values for n are 1–14. The APS channel refers to the K1 and K2 bytes in the line overhead (LOH), which are used to accomplish headend–tailend signaling. Because the headend is switchable, the protection line can be used to carry an extra traffic channel. Some texts call a subset of $N+1$ architecture the $1:1$ architecture.

The $N+1$ link protection method makes more efficient use of standby equipment. It is merely an extension of the $1+1$ technique described above. With the excellent reliability of present-day equipment, we can be fairly well assured that there will not be two simultaneous failures on a route. This makes it possible to share the standby line among N working lines.

Although $N+1$ link protection makes more cost-effective use of equipment, it requires more sophisticated control and cannot offer the same level of availability as $1+1$ protection. Diverse routing of main and standby lines is also much harder to achieve.

2.7. SONET Ring Configurations

A ring network consists of network elements connected in a point-to-point arrangement that forms an unbroken circular configuration as shown in Fig. 2.17. As we must realize, the main reason for implementing path-switched rings is to improve network survivability. The ring provides protection against fiber cuts and equipment failures.

Various terms are used to describe path switched ring functionality, for example, *unidirectional path-protection-switched* (UPPS) *rings, unidirectional path-switched rings* (UPSRs), *uniring, and counterrotating rings*.

Ring architectures may be considered a class of their own, but we analyze a ring conceptually in terms of $1+1$ protection. Usually, when we think of a ring architecture, we think of route diversity; and there are two separate directions of communications. The ring topology is most popular in the long-haul fiberoptic community. It offers what is called *geographic diversity*. Here we mean that there is sufficient ring diameter (e.g., >10 mi) that there is an excellent statistical probability that at least one side of the ring will survive forest fires, large floods, hurricanes, earthquakes, and other force majeure events. It also means that only one side of the ring will suffer an ordinary nominal equipment failure or "backhoe fade." Some forms of ring topology are used in CATV HFC systems, but more for achieving efficient and cost-effective connectivity than as a means of survivability. Rings are not used in building and campus fiber routings.

There are two basic SONET self-healing ring (SHR) architectures: the unidirectional and the bidirectional. Depending on the traffic demand pattern and some other factors, some ring types may be better suited to an application than others.

In a unidirectional SHR, shown in Fig. 2.17a, working traffic is carried around the ring in one direction only. For example, traffic going from node A to node D would

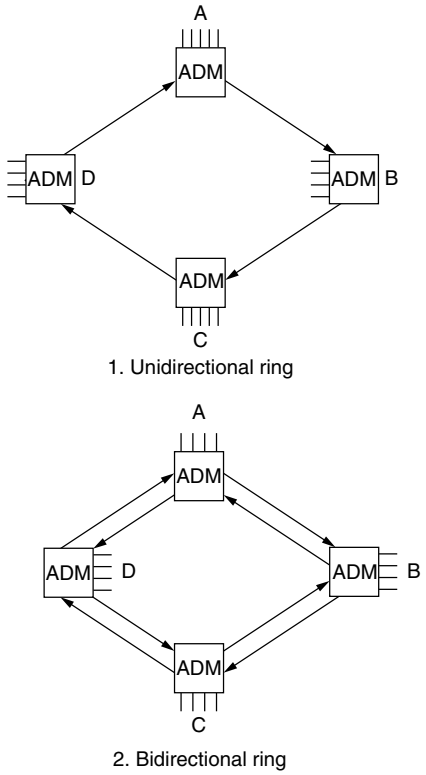


Figure 2.17. Ring definitions — working path direction(s) Refs. 1, 8, 9, 10, 11, and 12.

traverse the ring in a clockwise direction, and traffic going from node *D* to node *A* would also traverse the ring in a clockwise direction. The capacity of a unidirectional ring is determined by the total traffic demands between any two node pairs on the ring.

In a bidirectional SHR (Fig. 2.17b), working traffic is carried around the ring in both directions, using two parallel paths between nodes (e.g., sharing the same fiber cable sheath). Using an example similar to that described above, traffic going from node *A* to node *D* would traverse the ring in a clockwise direction through intermediate nodes *B* and *C*, and traffic going from node *D* to node *A* would return along the same path also going through intermediate nodes *B* and *C*.

In a bidirectional ring, traffic in both directions of transmission between two nodes traverse the same set of nodes. Thus, unlike a unidirectional ring time slot, a bidirectional ring time slot can be reused several times on the same ring, allowing better utilization of capacity. All the nodes on the ring share the protection bit rate capacity, regardless of the number of times the time slot has been reused. Bidirectional routing is also useful on large rings, where propagation delay can be a consideration, because it provides a mechanism for ensuring that the shortest path is used (under normal conditions) against failures affecting both working and protection paths as well as node failures [6,9].

2.7.1. UPSR–BLSR Comparison. As was discussed previously, the UPSR and the BLSR have different characteristics. The UPSR offers simplicity and efficiency in hubbed traffic environments. The BLSR, although more complex than the UPSR, offers protection from some failures against which the UPSR cannot provide protection. In addition, in traffic environments that are *not* hubbed, the BLSR can potentially offer more efficiency than a UPSR. Table 2.1 compares these two techniques.

Table 2.2 contains ring capacities for different types of rings. The span capacity is the capacity of an individual span between two nodes. The ring capacity is the total bit rate capacity of all connections through a ring. Capacities are best expressed as equivalent STS-1s.

3. SYNCHRONOUS DIGITAL HIERARCHY (SDH)

3.1. Introduction

SDH resembles SONET in most respects. It uses different terminology, often for the same function as SONET. It is behind SONET in maturity by several years. History tells us that SDH will be more pervasive worldwide than SONET and is or will be employed in all countries using an E1-based PDH (plesiochronous digital hierarchy).

3.2. SDH Standard Bit Rates

SDH bit rates are built on the basic rate of STM-1 (synchronous transport module 1) of 155.520 Mbps.

Table 2.1. Comparison of SONET UPSR with SONET BLSR

UPSR (Unidirectional Path-Switched Ring)	BLSR (Bidirectional Line-Switched Ring)
Path uses bit rate capacity around entire ring	More efficient bit rate capacity utilization
Less complex, no switching protocol	Switching protocol complicated
More economic	Less economic
Used in access networks	Basic use in inter-switch (trunk) applications
Interoperability: Different node on a UPSR can be from different manufacturers	Such interoperability yet to be estabilished

Refs. 2, 10, 11, and 12.

Table 2.2. Ring Bit Rate Capacities for an OC-N Ring

Ring Type	Maximum Span Capacity	Maximum Ring Capacity
UPSR 2-fiber	OC-N	OC-N
BLSR 2-fiber	OC-N/2	≥ OC-N ≤ OC-XN/2 (see Note)
BLSR 4-fiber	OC-N	≥ OC-2N ≤ OC-XN (see Note)

Note: Depending on the traffic pattern, X = Number of ring nodes.

Table 3.1. SDH Bit Rates

SDH Level	Equivalent SONET Level	Bit Rate (kbps)
1	STS-3/OC-3	155,520
4	STS-12/OC-12	622,080
16	STS-48/OC-48	2,488,320
64	STS-192/OC-192	9,953,280

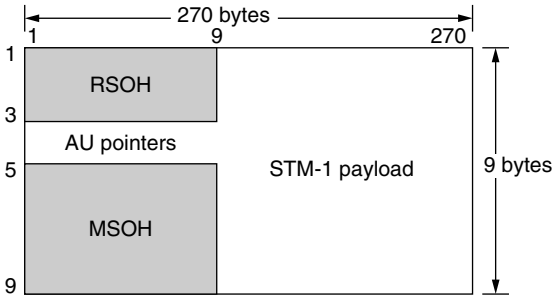


Figure 3.1. STM-1 frame structure (RSOH = regenerator section overhead; MSOH = multiplex section overhead).

Higher-capacity STMs are formed at rates equivalent to N times this basic rate. STM capacities for $N = 4, N = 16$, and $N = 64$ are defined. Table 3.1 shows the bit rates presently available for SDH (Ref. G.707 3/96) with its SONET equivalents.

The basic SDH multiplexing structure is shown in Figure 1.2 [4].

3.3. Definitions

3.3.1. Synchronous Transport Module (STM). An STM is the information structure used to support section-layer connections in the SDH. It consists of information payload and section overhead (SOH) information fields organized in a block frame structure that repeats every 125 μs. The information is suitably conditioned for serial transmission on the selected medium at a rate that is synchronized to the network. As mentioned above, a basic STM is defined at 155.520 Mbps; this is termed STM-1. Higher-capacity STMs are formed at rates equivalent to N times the basic rate. STM capacities are currently defined for $N = 4, N = 16$, and $N = 64$. Figure 3.1 shows the frame structure for STM-1.

The STM-1 comprises a single Administrative Unit Group (AUG) together with the section overhead (SOH). The STM-N contains N AUGs together with SOH. Figure 3.2 illustrates an STM-N.

3.3.2. Virtual Container n (VC-n). A *virtual container* is the information structure used to support path-layer connections in the SDH. It consists of information payload and path overhead (POH) information fields organized in a block frame structure that repeats every 125 or 500 μs. Alignment information to identify VC-n frame start is provided by the server network.

Two types of virtual containers have been identified:

- *Lower-Order Virtual Container n*: *VC-n* ($n = 1, 2, 3$). This element contains a single container n ($n = 1, 2, 3$) plus the lower-order virtual container POH appropriate to that level.
- *Higher-Order Virtual Container n*: *VC-n* ($n = 3, 4$). This element comprises either a single container n ($n = 3$) or an assembly of tributary groups (TUG-2s or TUG-3s), together with virtual container POH appropriate to that level.

3.3.3. Administrative unit n (AU-n). An *administrative unit* is the information structure that provides adaptation between the higher-order path layer and the multiplex section layer. It consists of an information payload (the higher-order virtual container) and an administrative unit pointer that indicates the offset of the payload frame start relative to the multiplex section frame start.

Two administrative units are defined. The AU-4 consists of a VC-4 plus an administrative unit pointer that indicates the phase alignment of the VC-4 with respect to the STM-N frame. The AU-3 consists of a VC-3 plus an administrative unit pointer that indicates the phase alignment of the VC-3 with respect to the STM-N frame. In each case the administrative unit pointer location is fixed with respect to the STM-N frame.

One or more administrative units occupying fixed, defined positions in an STM payload are termed an *administrative unit group* (AUG). An AUG consists of an homogeneous assembly of AU-3s or an AU-4.

3.3.4. Tributary Unit N (TU-n). A *tributary unit* is an information structure that provides adaptation between

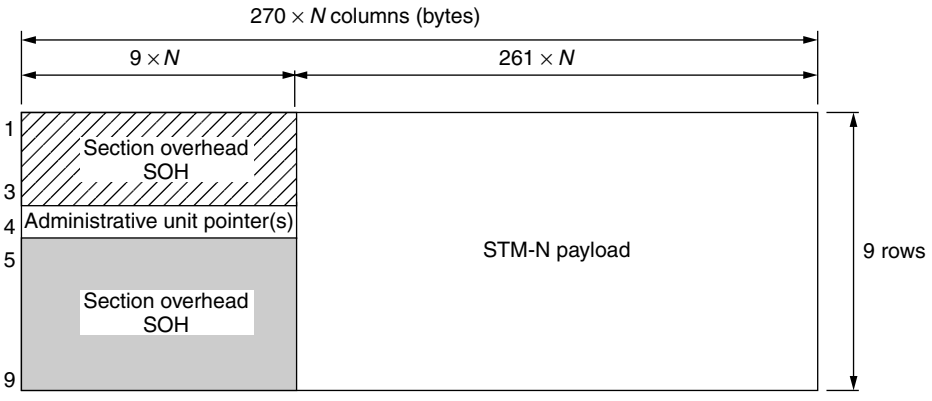


Figure 3.2. STM-N frame structure.

the lower-order path layer and the higher-order path layer. It consists of an information payload (the lower-order virtual container) and a tributary unit pointer that indicates the offset of the payload frame start relative to the higher-order VC frame start. The TU-n ($n = 2, 2, 3$) consists of a VC-n together with a tributary unit pointer.

One or more tributary units, occupying fixed defined positions in a higher order VC-n payload is termed a *tributary unit group* (TUG). TUGs are defined in such a way that mixed-capacity payloads made up of different-size tributary units can be constructed to increase flexibility of the transport network. A TUG-2 consists of a homogeneous assembly of identical TU-1s or a TU-2. A TUG-3 consists of a homogeneous assembly of TUG-2s or a TU-3.

3.3.5. Container n ($n = 1–4$). A *container* is the information structure that forms the network synchronous information payload for a virtual container. For each defined virtual container, there is a corresponding container. Adaptation functions have been defined for many common network rates into a limited number of standard containers. These include all of the rates defined in CCITT Rec. G.702.

3.3.6. Pointer. A *pointer* is an indicator whose value defines the frame offset of a virtual container with respect to the frame reference of the transport entity that is supported.

3.4. Conventions

The order of transmission of information in all diagrams and figures in this section is first from left to right and then top to bottom. Within each byte (octet), the most significant bit is transmitted first. The most significant bit (bit 1) is illustrated at the left in all the diagrams, figures, and tables in this section [5].

3.5. Basic SDH Multiplexing

Figure 3.3 illustrates the relationship between various multiplexing elements, which are defined in the text below. It also shows common multiplexing structures.

Figures 3.4–3.8 illustrate examples of various signals that are multiplexed using the multiplexing elements shown in Fig. 3.3.

3.6. Administrative Units in the STM-N

The STM-N payload can support N AUGs, where each AUG may consist of one AU-4 or three AU-3s.

The VC-n associated with each AU-n does not have a fixed phase with respect to the STM-N frame. The location of the first byte of the VC-n is indicated by the AU-n pointer. The AU-n pointer is in a fixed location in the STM-N frame. Examples are illustrated in Figs. 3.2 and 3.4–3.9.

The AU-4 may be used to carry, via the VC-4, a number of TU-n ($n = 1, 2, 3$) units forming a two-stage multiplex. An example of this arrangement is illustrated in Figs. 3.8a and 3.9a. The VC-n associated with each TU-n does not have a fixed-phase relationship with respect to the start of the VC-4. The TU-n pointer is in a fixed location in the VC-4, and the location of the first byte of the VC-n is indicated by the TU-n pointer.

The AU-3 may be used to carry, via the VC-3, a number of TU-n ($n = 1, 2$) units forming a two-stage multiplex. An example of this arrangement is illustrated in Figs. 3.8b and 3.9b. The VC-n associated with each TU-n does not have a fixed-phase relationship with respect to the start of the VC-3. The TU-n pointer is in a fixed location in the VC-3, and the location of the first byte of the VC-n is indicated by the TU-n pointer.

3.6.1. Interconnection of STM-Ns. SDH is designed to be universal, allowing the transport of a large variety of signals including those specified in CCITT Rec. G.702. However, different structures can be used for the transport of virtual containers. The following interconnection rules will be used (Ref. G.707):

1. The rule for interconnecting two AUGs based on two different types of administrative unit — namely, AU-4 and AU-3 — will be to use the AU-4 structure.

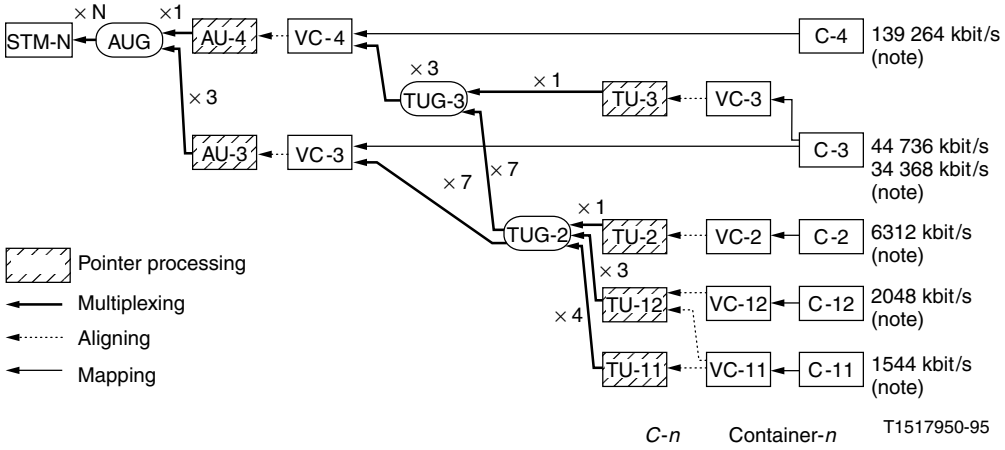


Figure 3.3. Multiplexing structure overview. *Note*: G.702 tributaries associated with containers C-x are shown; other signals, e.g., ATM, can also be accommodated.) (From Ref. 5, Fig. 6-1/G.707, p. 6.)

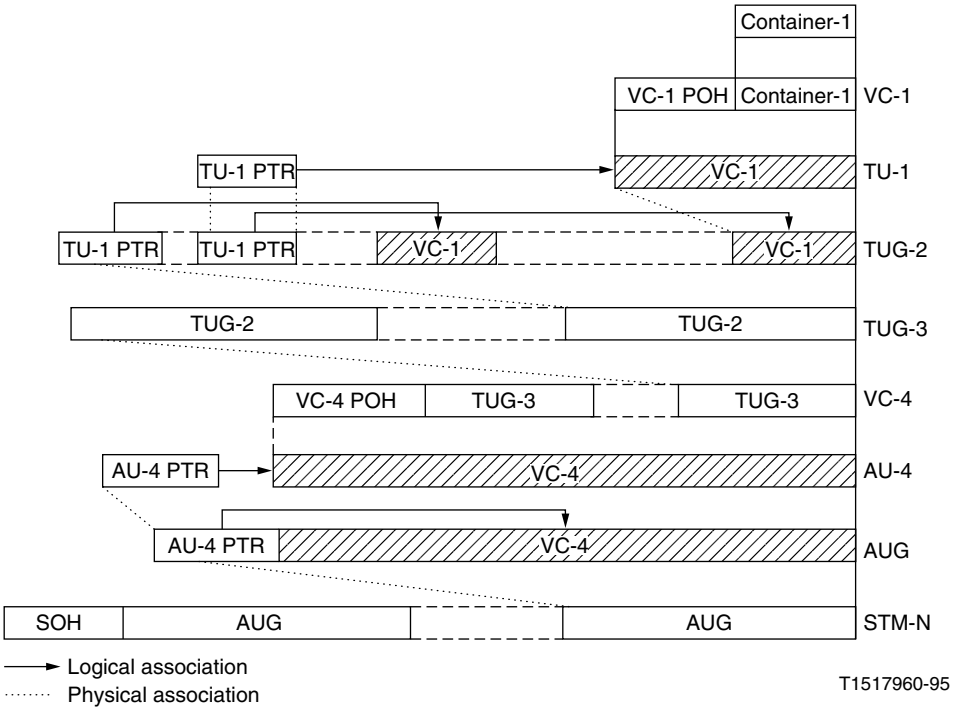


Figure 3.4. Multiplexing method directly from container 1 using AU-4. [*Note*: Unshaded areas are phase-aligned. Phase alignment between the unshaded and shaded areas is defined by the pointer (PTR) and is indicated by the arrow.] (From Ref. 5, Fig. 6-2/G.707, p. 7.)

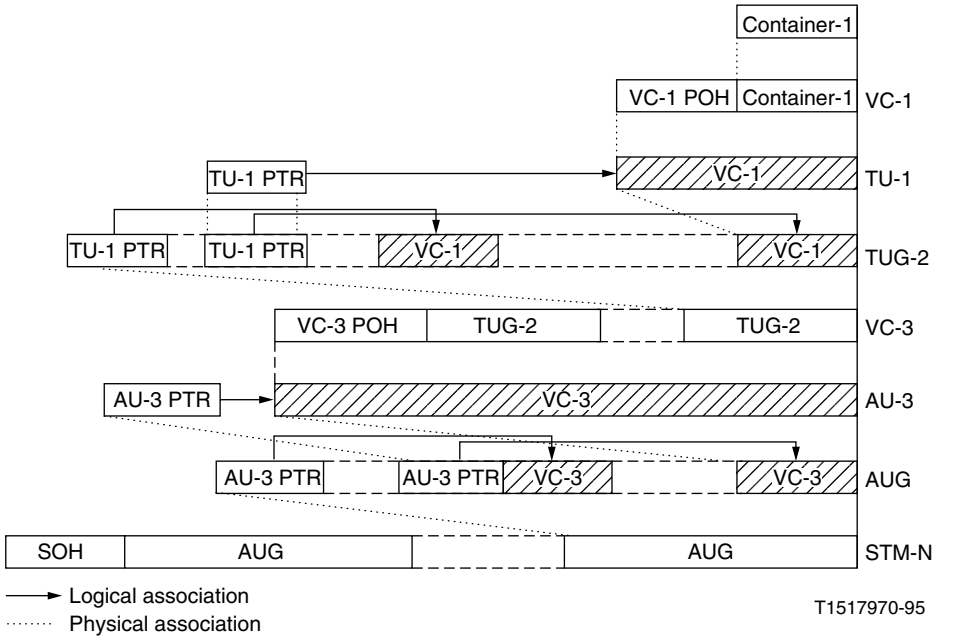


Figure 3.5. Multiplexing method directly from container 1 using AU-3. [*Note*: Unshaded areas are phase-aligned. Phase alignment between the unshaded and shaded areas is defined by the pointer (PTR) and is indicated by the arrow.] (From Ref. 5, Fig. 6-3/G.707, p. 7.)

Therefore, the AUG based on AU-3 will be demultiplexed to the VC-3 or TUG-2 level according to the type of payload, and remultiplexed within an AUG via the TUG-3/VC-4/AU-4 route. This is illustrated in Fig. 3.9a,b.

2. The rule for interconnecting VC-11s transported via different types of tributary unit — namely, TU-11 and TU-12 — will be to use the TU-11 structure. This is illustrated in Fig. 3.10c. VC-11, TU-11, and TU-12 are described below.

This SDH interconnection rule does not modify the interworking rules defined in ITU-T Rec. G.802 for networks based on different PDHs and speech encoding laws.

3.6.2. Scrambling. Scrambling assures sufficient bit timing content (transitions) at the NNI to maintain synchronization and alignment. Figure 3.11 is a functional block diagram of the frame synchronous scrambler. The generating polynomial for the scrambler is $1 + X^6 + X^7$.

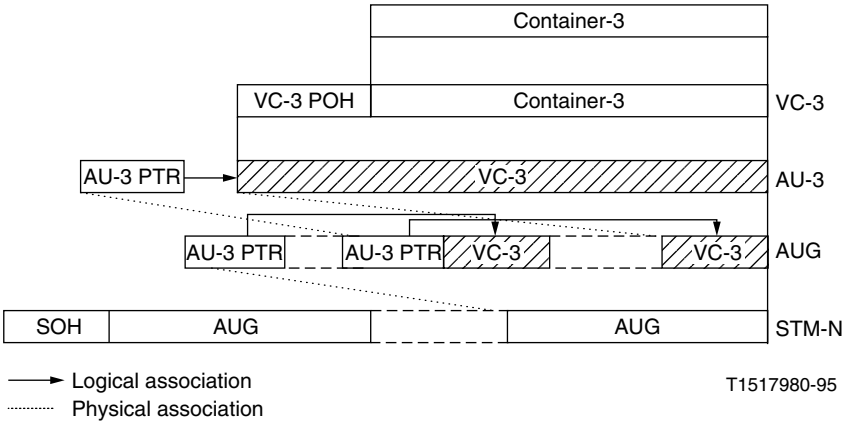


Figure 3.6. Multiplexing method directly from container 3 using AU-3. [*Note*: Unshaded areas are phase-aligned. Phase alignment between the unshaded and shaded areas is defined by the pointer (PTR) and is indicated by the arrow.] (From Ref. 5, Fig. 6-4/G.707, p. 9.)

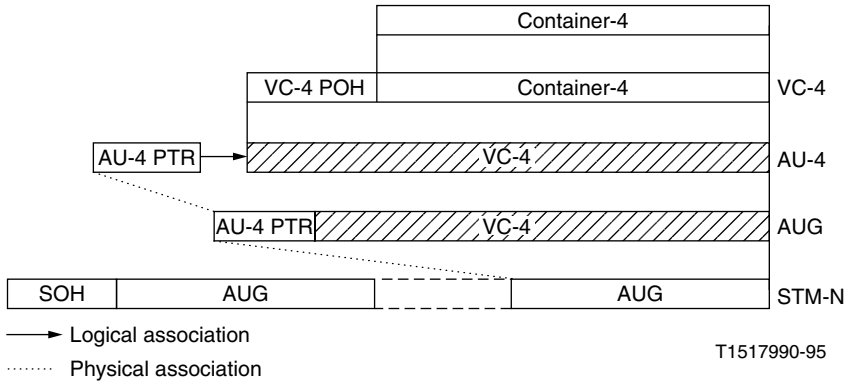


Figure 3.7. Multiplexing method directly from container 4 using AU-4. [*Note*: Unshaded areas are phase-aligned. Phase alignment between the unshaded and shaded areas is defined by the pointer (PTR) and is indicated by the arrow.] (From Ref. 5, Fig. 6-5/G.707, p. 10.)

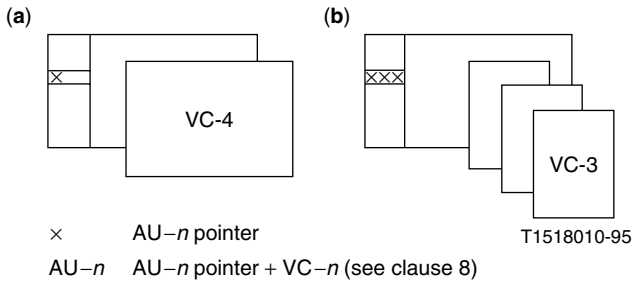


Figure 3.8. Administrative units in an STM-1 frame: (**a**) with one AU-4; (**b**) with three AU-3s.

3.7. Frame Structure for 51.840-Mbps Interface

Low/medium-capacity SDH transmission systems based on radio and satellite technologies that are not designed for the transmission of STM-1 signals may operate at a bit rate of 51.840 Mbps across digital sections. However, this bit rate does not represent a level of the SDH or a NNI bit rate (ITU-R Rec. G.707).

The recommended frame structure for a 51.840-Mbps signal for satellite (ITU-R Rec. S.1149) and LoS microwave application (ITU-R Rec. F.750) is shown in Fig. 3.12.

3.8. Multiplexing Methods

3.8.1. Multiplexing of Administrative Units into STM-N

3.8.1.1. Multiplexing of Administrative Unit Groups (AUGs) into an STM-N. The arrangement of N AUGs multiplexed into an STM-N is illustrated in Fig. 3.13. The AUG is a structure of 9 rows × 261 columns plus 9 bytes in row 4 (for the AU-n pointers). The STM-N consists of an SOH described below and a structure of 9 rows × (N × 261) columns with N × 9 bytes in row 4 (for the AU-n pointers). The N AUGs are one single-byte-interleaved into this structure and have a fixed-phase relationship with respect to the STM-N.

3.8.1.2. Multiplexing of an AU-4 via AUG. The multiplexing arrangement of a single AU-4 via the AUG is illustrated in Fig. 3.14. The 9 bytes at the beginning of row 4 are assigned to the AU-4 pointer. The remaining 9 rows × 261 columns are allocated to virtual container 4 (VC-4). The phase of the VC-4 is not fixed with respect to the AU-4. The location of the first byte of the VC-4 with respect to the AU-4 pointer is given by the pointer value. The AU-4 is placed directly in the AUG.

3.8.1.3. Multiplexing of AU-3s via AUG. The multiplexing arrangement of three AU-3s via the AUG is shown in Fig. 3.15. The 3 bytes at the beginning of row 4 are assigned to the AU-3 pointer. The remaining 9 rows × 87 columns are allocated to the VC-3 and two columns of fixed stuff. The byte in each row of the two columns of fixed stuff of each AU-3 shall be the same. The phase of the VC-3 and the two columns of fixed stuff is not fixed with respect to the AU-3. The location of the first byte of the VC-3 with respect to the AU-3 pointer is given by the pointer value. The three AU-3s are single-byte-interleaved in the AUG.

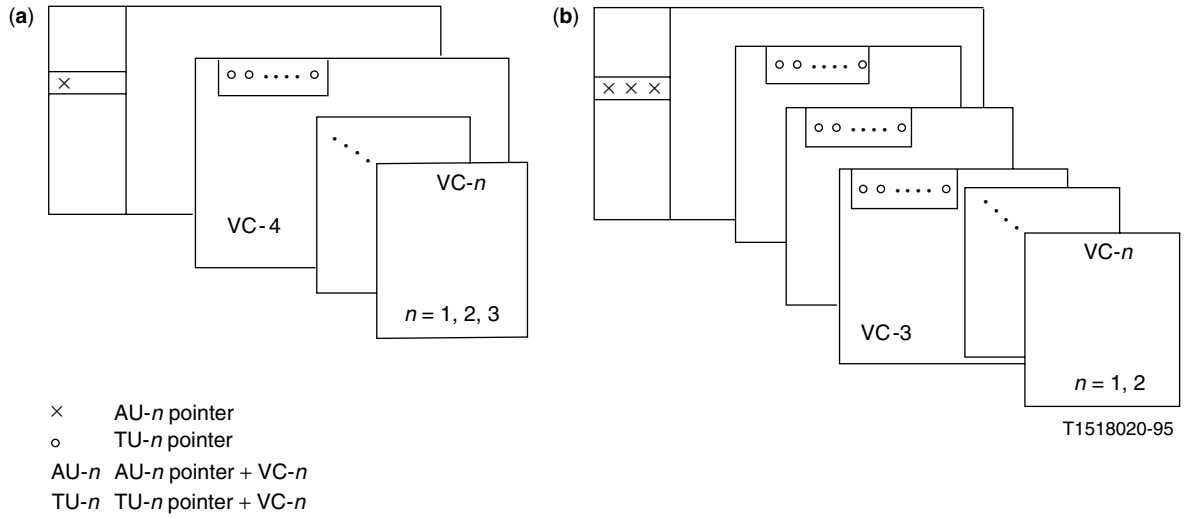


Figure 3.9. Two-stage multiplex: STM-1 with (**a**) one AU-4 containing TUs and (**b**) three AU-3s containing TUs. (From Ref. 5, Fig. 6-8/G.707, p. 12.)

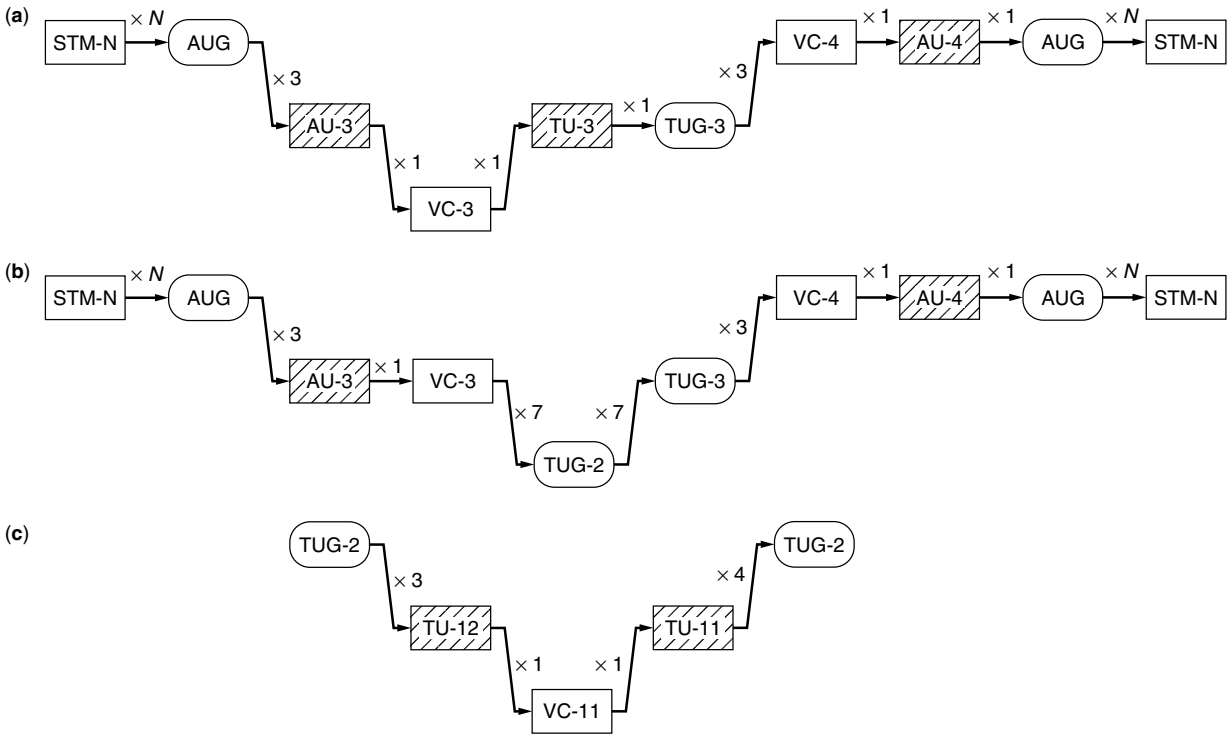


Figure 3.10. Interconnection of STM-Ns: (**a**) of VC-3 with C-3 payload; (**b**) of TUG-2; (**c**) of VC-11. (From Ref. 5, Fig. 6-9/G.707, p. 16.)

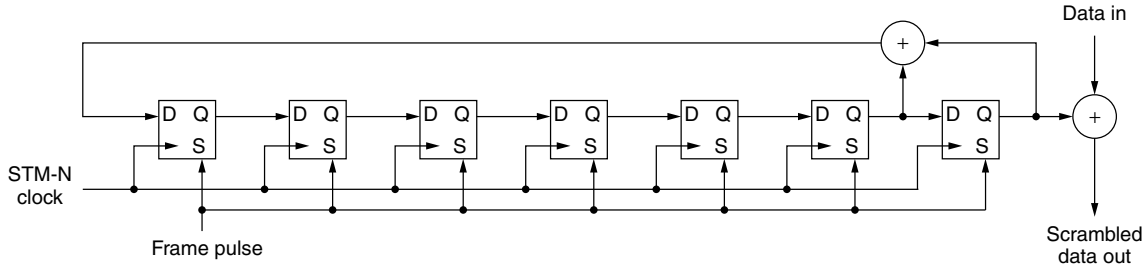
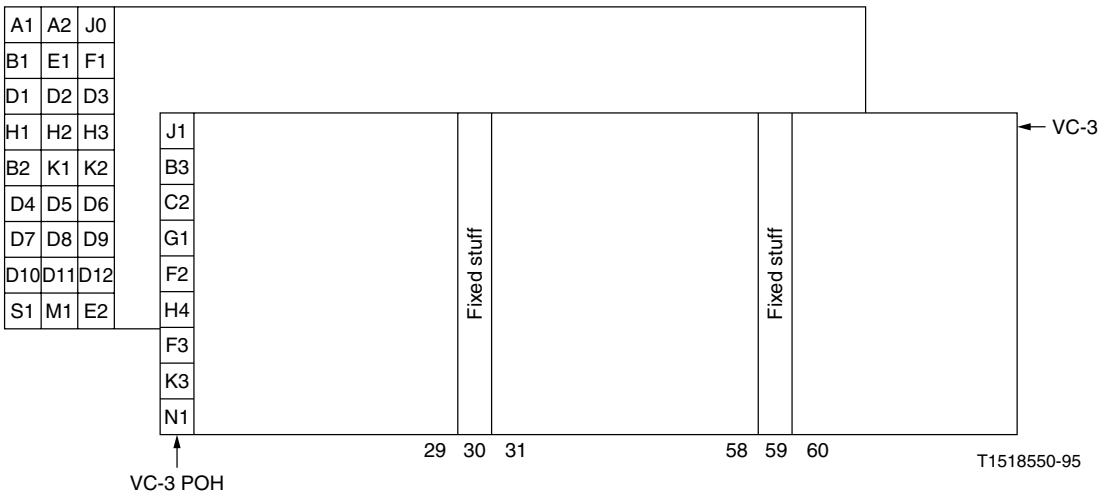


Figure 3.11. Functional block diagram of a frame-synchronous scrambler. (From Ref. 5, Fig. 6-10/G.707, p. 17, ITU-T.)



NOTES

1 M1 position is not the same position [9, 3N + 3] as in a STM-N frame.
2 Fixed stuff columns are not part of the VC-3.

Figure 3.12. Frame structure for 51.840-Mbps (SDH) operation. (From Ref. 5, Fig. A.1/G.707, p. 89.)

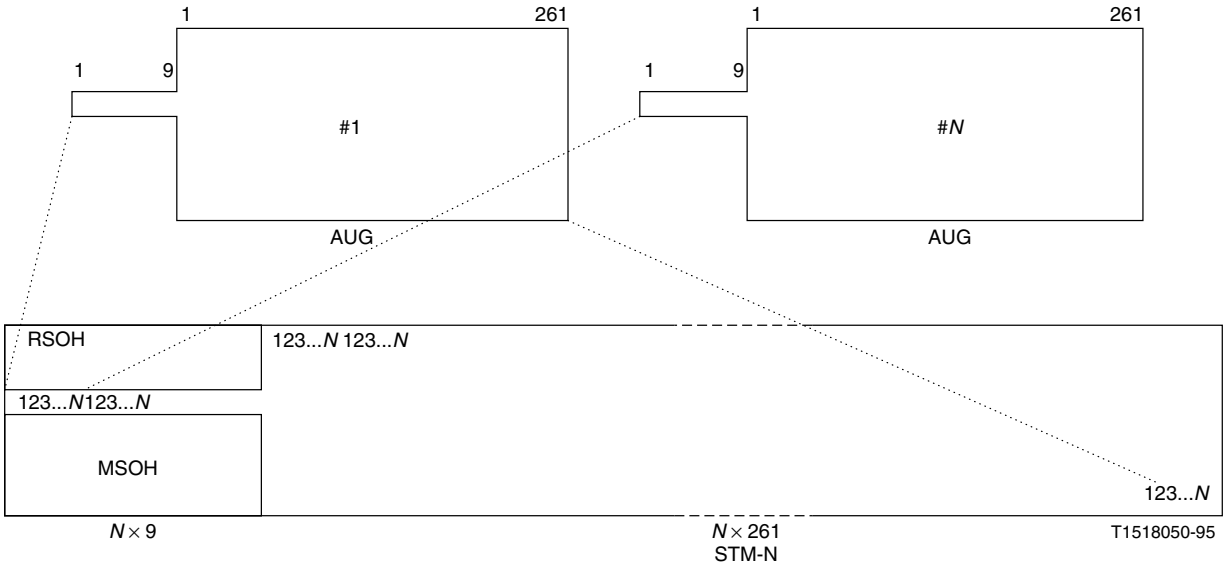


Figure 3.13. Multiplexing of *N* AUGs into an STM-N. (From Ref. 5, Fig. 7-1/G.707, p. 18.)

3.8.2. Multiplexing of Tributary Units into VC-4 and VC-3

3.8.2.1. Multiplexing of Tributary Unit Group 3S (TUG-3s) into a VC-4

The arrangement of three TUGs multiplexed in the VC-4 is illustrated in Fig. 3.16 The TUG-3 is a 9-row × 86-column structure. The VC-4 consists of one column of VC-4 POH, two columns of fixed stuff and a 258-column payload structure. The three TUG-3s are single-byte-interleaved into the 9-row × 258-column VC-4 payload structure and have a fixed phase with respect to the VC-4. As described in 3.9.1.1, the phase of the VC-4 with respect to the AU-4 is given by the AU-4 pointer.

3.8.2.2. Multiplexing of a TU-3 via a TUG-3. The multiplexing of a single TU-3 via the TUG-3 is shown in Fig. 3.17. The TU-3 consists of the VC-3 with a 9-byte VC-3 POH and the TU-3 pointer. The first column of the 9-row × 86-column TUG-3 is assigned to the TU-3 pointer (bytes H1, H2, H3) and fixed stuff. The phase of the VC-3 with respect to the TUG-3 is indicated by the TU-3 pointer.

3.8.2.3. Multiplexing of TUG-2s via a TUG-3. The multiplexing format for the TUG-2 via the TUG-3 is illustrated in Fig. 3.18. The TUG-3 is a 9-row × 86-column structure with the first two columns of fixed stuff.

3.8.2.4. Multiplexing of TUG-2s into a VC-3. The multiplexing structure for TUG-2s into a VC-3 is shown

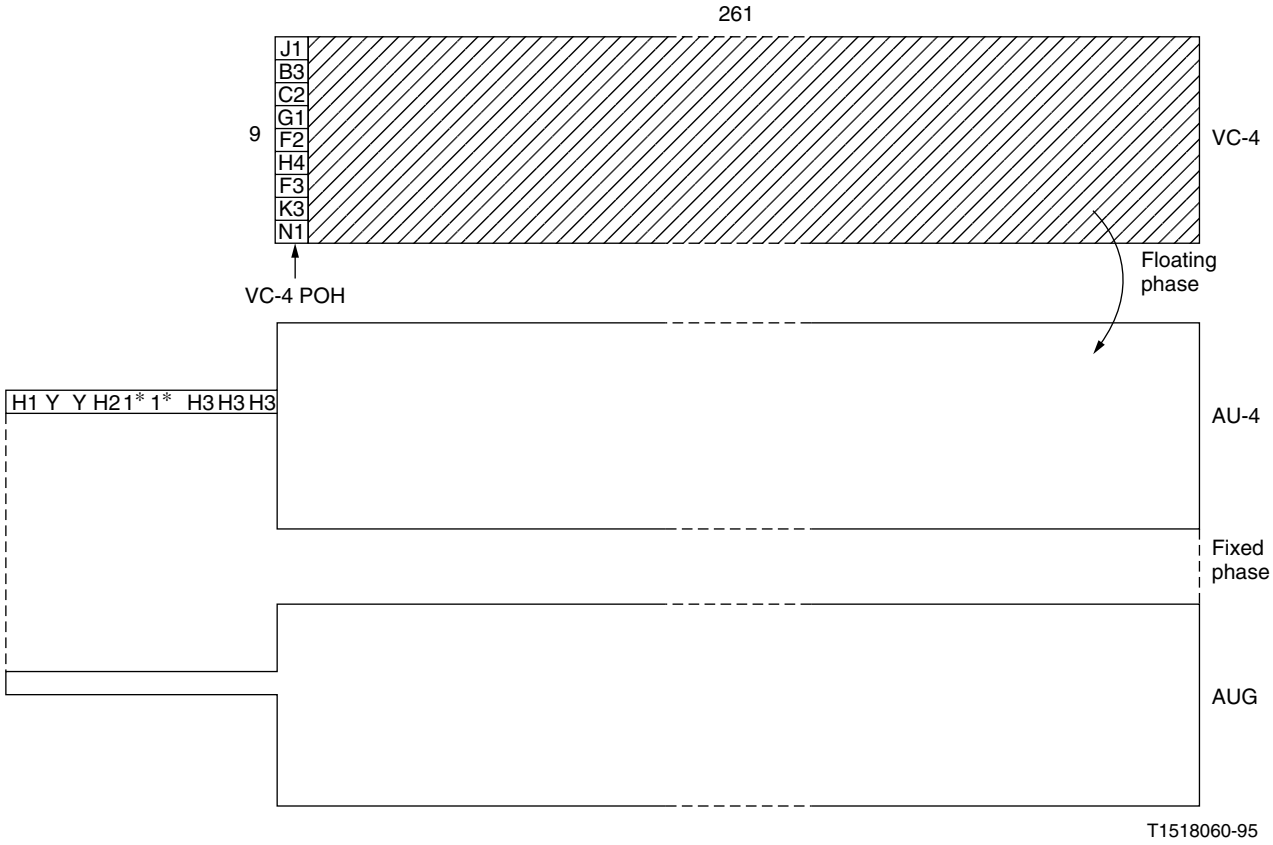


Figure 3.14. Multiplexing of AU-4 via AUG. (From Ref. 5, Fig. 7-2/G.707, p. 19.)

in Fig. 3.19. The VC-3 consists of VC-3 POH and a 9-row × 84-column payload structure. A group of seven TUG-2s can be multiplexed into the VC-3.

3.9. Pointers

3.9.1. AU-*n* Pointer. The AU-*n* pointer provides a method of allowing flexible and dynamic alignment of the VC-*n* within the AU-*n* frame. Dynamic alignment means that the VC-*n* is allowed to "float" within the AU-*n* frame. Thus, the pointer is able to accommodate differences, not only in the phases of the VC-*n* and the SOH, but also in the frame rates.

3.9.1.1. AU-n Pointer Location. The AU-4 pointer is contained in bytes H1, H2, and H3 as shown in Fig. 3.20. The three individual AU-3 pointers are contained in three separate H1, H2, and H3 bytes as shown in Fig. 3.21.

3.9.1.2. AU-n Pointer Value. The pointer contained in H1 and H2 designates the location of the byte where the VC-*n* begins. The 2 bytes allocated to the pointer function should be viewed as one word, as illustrated in Fig. 3.22. The last ten bits (bits 7–16) of the pointer word carry the pointer value.

As illustrated in Fig. 3.22, the AU-4 pointer value is a binary number with a range of 0–782 that indicates the offset in 3-byte increments, between the pointer and the first byte of the VC-4 (see Fig. 3.20). Figure 3.22 also indicates one additional valid pointer, the Concatenation Indication. The Concatenation Indication is indicated by binary 1001 in bits 1–4, bits 5–6 are unspecified, and 10 binary 1s in bit positions 7–16. The AU-4 pointer is set to "concatenation indication" for AU-4 concatenation.

As shown in Fig. 3.22, the AU-3 pointer value is also a binary number with a range of 0–782. Since there are three AU-3s in the AUG, each AU-3 has its own associated H1, H2, and H3 bytes.

Note that the H bytes are shown in sequence in Fig. 3.20. The first H1,H2,H3 set refers to the first AU-3, and the second set to the second AU-3, and so on. For the AU-3s, each pointer operates independently.

In all cases, the AU-*n* pointer bytes are not counted in the offset. For example, in an AU-4, the pointer value of 0 indicates that the VC-4 starts in the byte location that immediately follows the last H3 byte, whereas an offset of 87 indicates that the VC-4 starts 3 bytes after the K2 byte.

3.9.1.3. Frequency Justification. If there is a frequency offset between the frame rate of the AUG and that of the VC-*n*, then the pointer value will be incremented or decremented as needed, accompanied by a corresponding positive or negative justification byte or bytes. Consecutive pointer operations must be separated by at least three

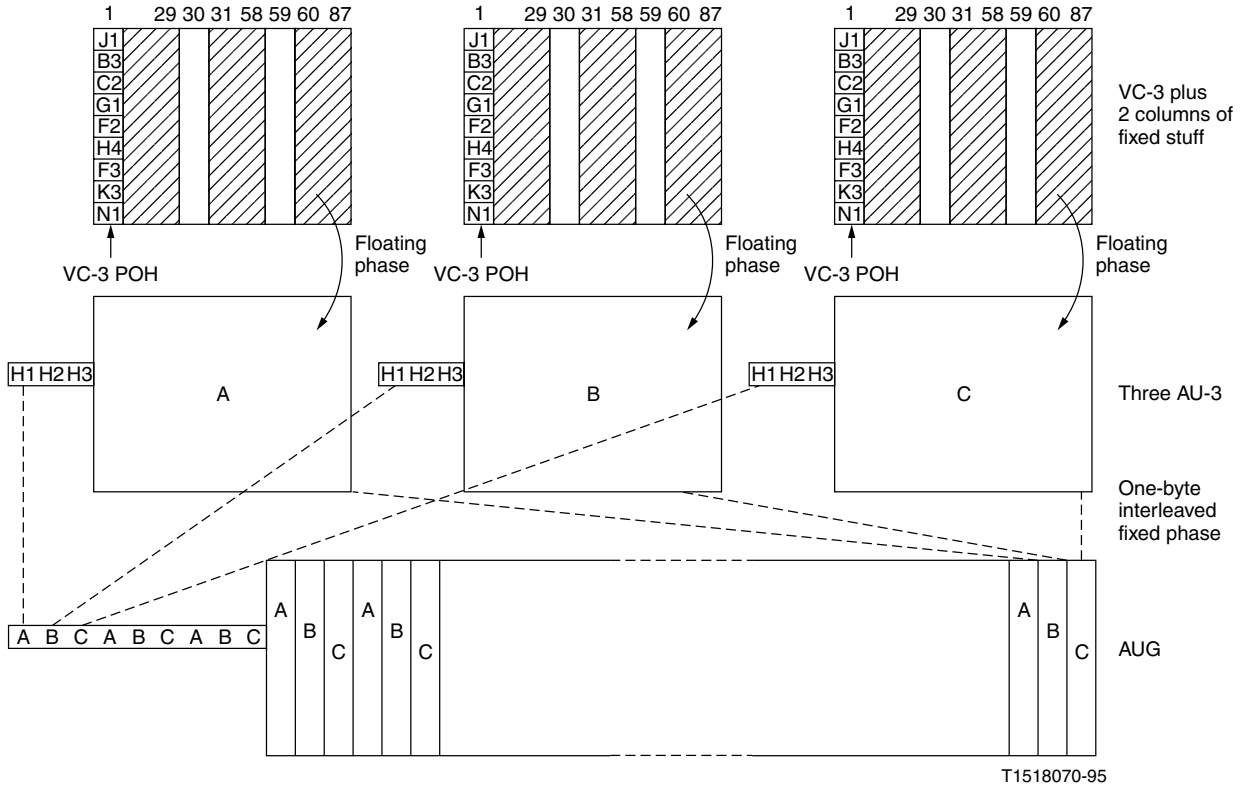


Figure 3.15. Multiplexing of AU-3s via AUG. (*Note*: The byte in each row of the two columns of fixed stuff of each AU-3 shall be the same.) (From Ref. 5, Fig. 7-3/G.707, p. 20.)

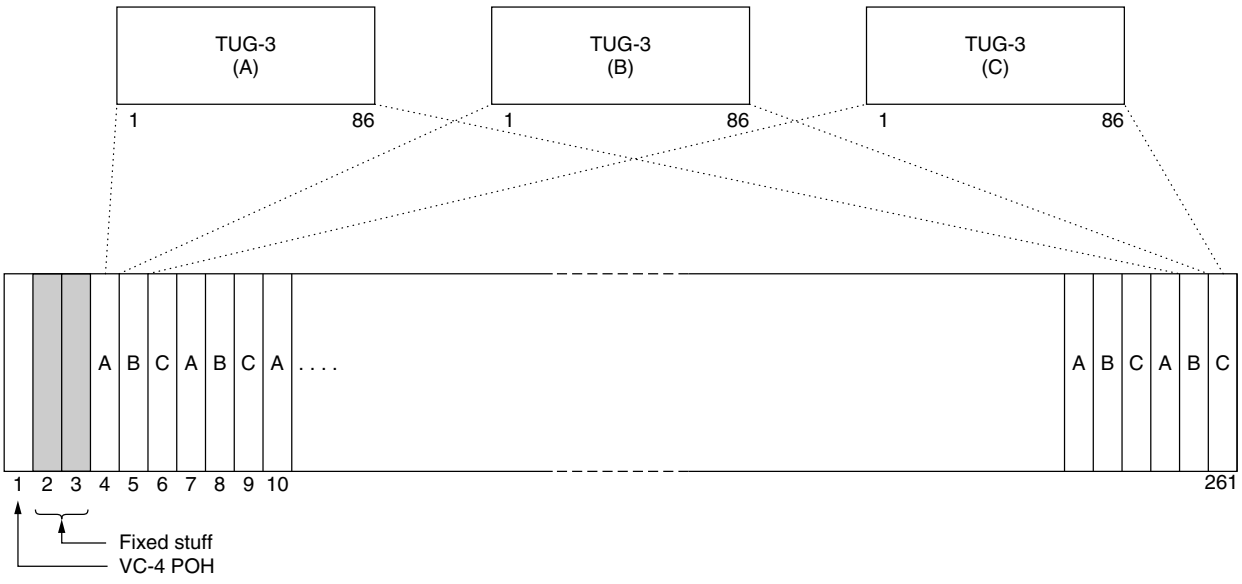


Figure 3.16. Multiplexing of three TUG-3s into a VC-4. (From Ref. 5, Fig. 7/4/G.707, p. 21.)

frames (i.e., every fourth frame) in which the pointer value remains constant.

If the frame rate of the VC-*n* is too slow with respect to that of the AUG, then the alignment of the VC-*n* must periodically slip back in time and the pointer value must be incremented by one. This operation is indicated by inverting bits 7, 9, 11, 13, and 15 (*I* bits) of the pointer word to allow 5-bit majority voting at the receiver. Three positive justification bytes appear immediately after the last H3 byte in the AU-4 frame containing inverted *I* bits. Subsequent pointers will contain the new offset. This is depicted in Fig. 3.23.

For AU-3 frames, a positive justification byte appears immediately after the individual H3 byte of the AU-3

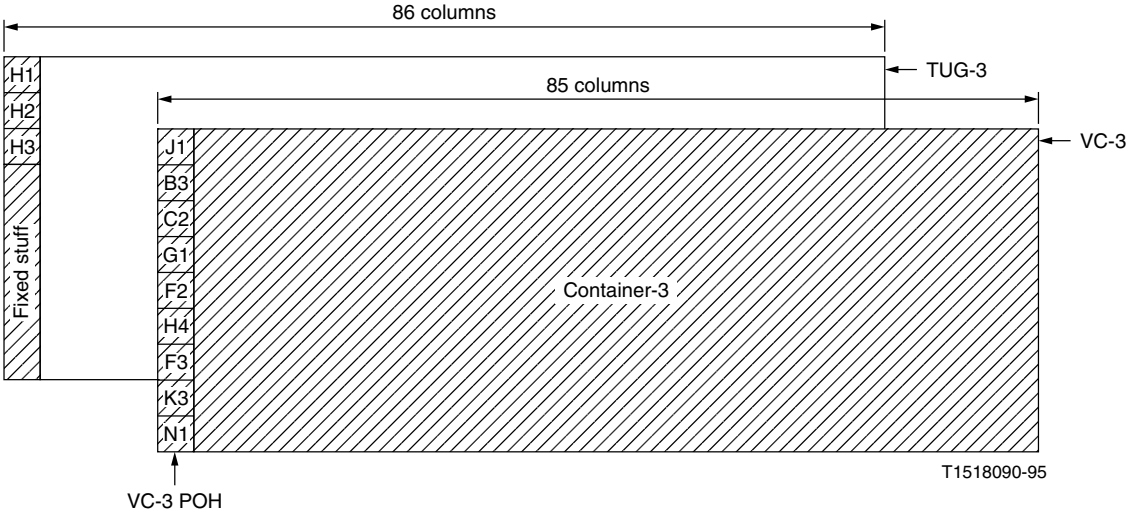


Figure 3.17. Multiplexing a TU-3 via a TUG-3. (From Ref. 5, Fig. 7-5/G.707, p. 21.)

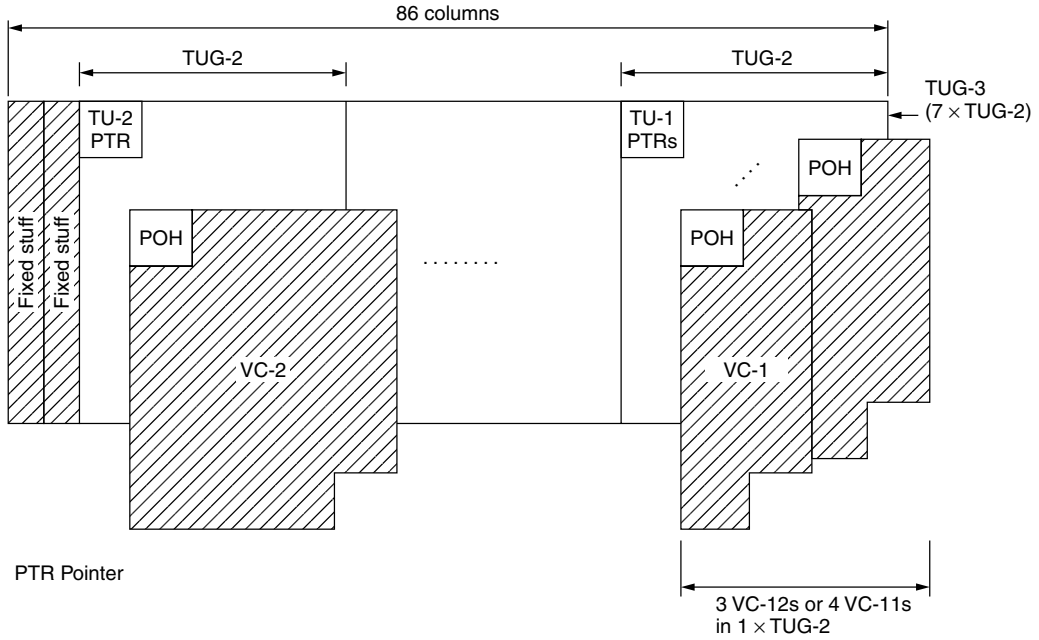


Figure 3.18. Multiplexing of TUG-2s via a TUG-3. (From Ref. 5, Fig. 7-6/G.707, p. 22.)

frame containing inverted *I* bits. Subsequent pointers will contain the new offset.

If the frame rate of the VC-*n* is too fast with respect to that of the AUG, then the alignment of the VC-*n* must periodically be advanced in time and the pointer value must be decremented by one. This operation is indicated by inverting bits 8, 10, 12, 14, and 16 (*D* bits) of the pointer word to allow 5-bit majority voting at the receiver. Three negative justification bytes appear in the H3 bytes in the AU-4 frame containing inverted *D* bits. Subsequent pointers will contain the new offset.

For AU-3 frames, a negative justification byte appears in the individual H3 byte of the AU-3 frame containing inverted *D* bits. Subsequent pointers will contain the new offset.

3.9.1.4. Pointer Generation. The following summarizes the rules for generating the AU-*n* pointers:

1. During normal operation, the pointer locates the start of the VC-*n* within the AU-*n* frame. The NDF (new data flag) is set to binary 0110. The NDF consists of the *N* bits, bits 1–4 of the pointer word.
2. The pointer value can be changed only by operation 3, 4, or 5 (see items below).
3. If a positive justification is required, the current pointer value is sent with the *I* bits inverted and the subsequent positive justification opportunity is filled with dummy information. Subsequent pointers contain the previous pointer value incremented by one. If the previous pointer is at its maximum value,

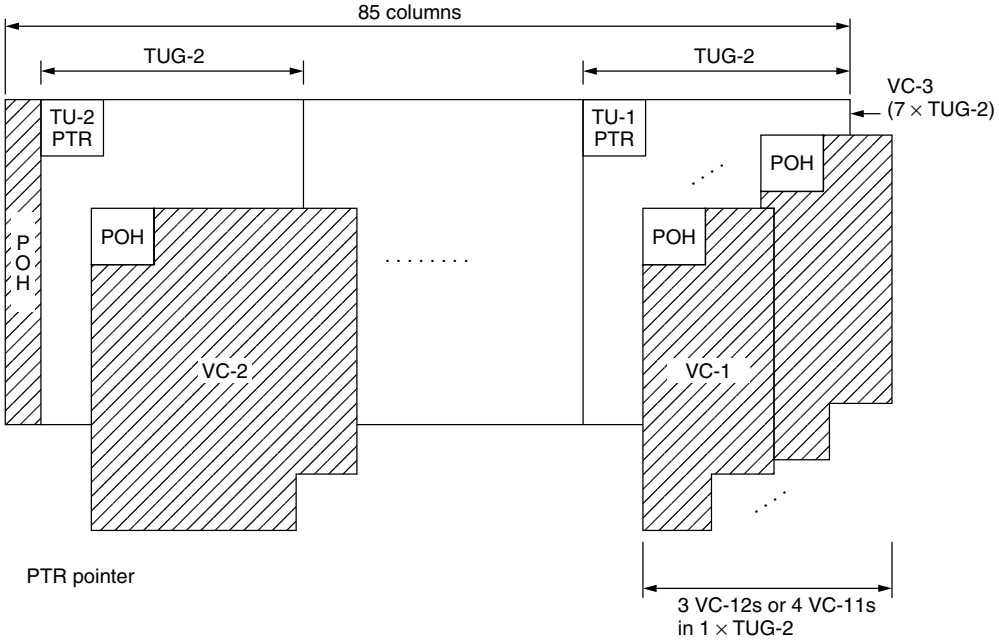


Figure 3.19. Multiplexing seven TUG-2s into a VC-3. (From Ref. 5, Fig. 7-8/G.707, p. 24.)

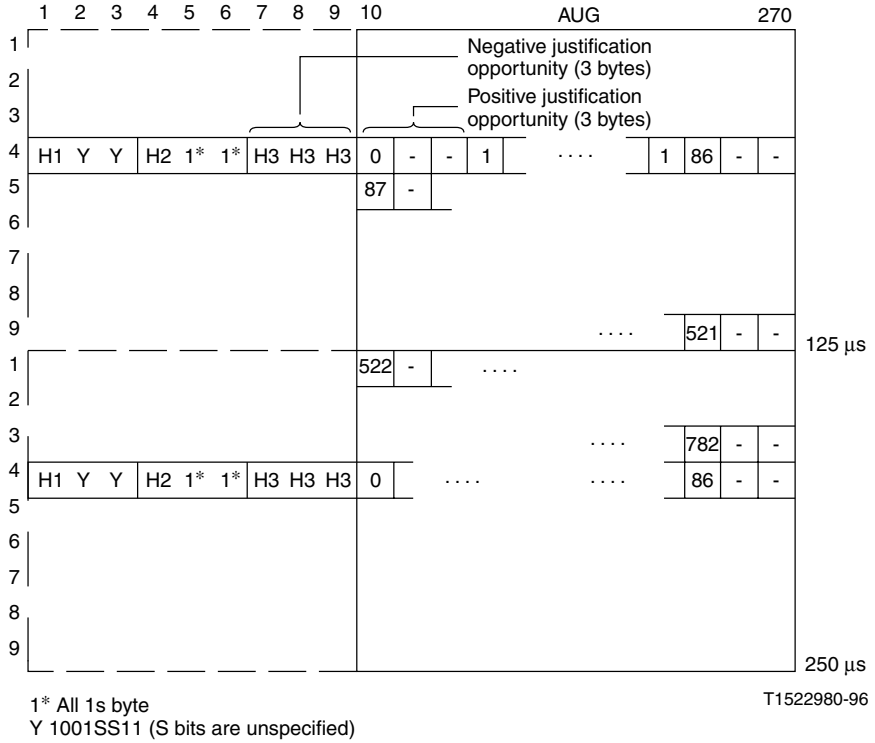


Figure 3.20. AU-4 pointer offset numbering. (From Ref. 5, Fig. 8-1/G.707, p. 34.)

the subsequent pointer is set to zero. No subsequent increment or decrement operation is allowed for at least three frames following this operation.

4. If a negative justification is required, the current pointer value is sent with the *D* bits inverted and the subsequent negative justification opportunity is overwritten with actual data. Subsequent pointers contain the previous pointer value decremented by one. If the previous value is zero, the subsequent pointer is set to its maximum value. No subsequent increment or decrement operation is allowed for at least three frames following this operation.
5. If the alignment of the VC-*n* changes for any reason other than rules 3 or 4, the new pointer value should be sent accompanied by the NDF set to 1001. The NDF only appears in the first frame that contains the new values. The new location of the VC-n begins at the first occurrence of the offset

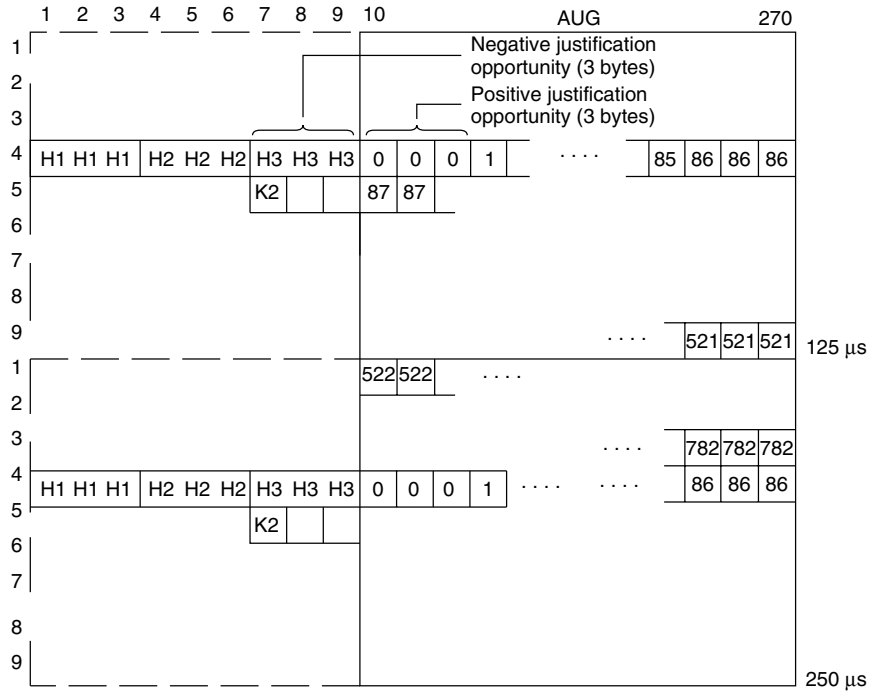
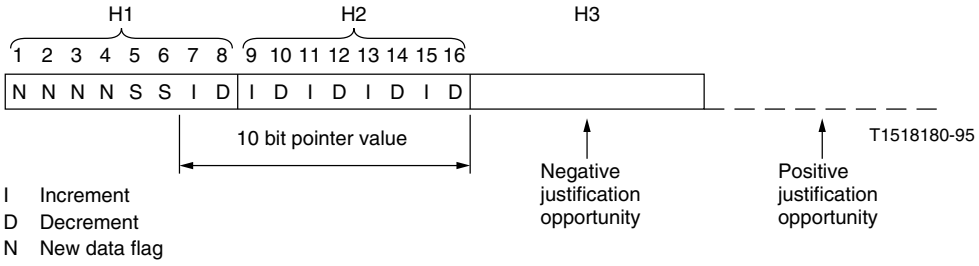


Figure 3.21. AU-3 pointer offset numbering. (From Ref. 5, Fig. 8-2/G.707, p. 325.)



I Increment
D Decrement
N New data flag

New data flag
- Enabled when at least 3 out of 4 bits match "1001"
- Disabled when at least 3 out of 4 bits match "0110"
- Invalid with other codes

Negative justification
- Invert 5 D-bits
- Accept majority vote

Pointer value (b7-b16)
- Normal range is:
  for AU-4, AU-3: 0-782 decimal
  for TU-3: 0-764 decimal

Positive justification
- Invert 5 I-bits
- Accept majority vote

Concatenation indication
- 1001SS1111111111 (SS bits are unspecified)

SS bits	AU-*n*/TU-*n* type
10	AU-4, AU-3, TU-3

Figure 3.22. AU-*n*/TU-3 pointer (H1, H2, H3) coding. (From Ref. 5, Fig. 8-3/G.707, p. 36.)

indicated by new pointer. No subsequent increment or decrement operation is allowed for at least three frames following this operation.

3.9.1.5. Pointer Interpretation. The following list summarizes the rules for interpreting the AU-*n* pointers:

1. During normal operation, the pointer locates the start of the VC-*n* within the AU-*n* frame.
2. Any variation from the current pointer value is ignored unless a consistent new value is received 3 times consecutively or is preceded by either rule 3, 4, or 5. Any consistent new value received three times consecutively overrides (i.e., takes priority over) rules 3 and 4.
3. If the majority of the *I* bits of the pointer word is inverted, a positive justification operation

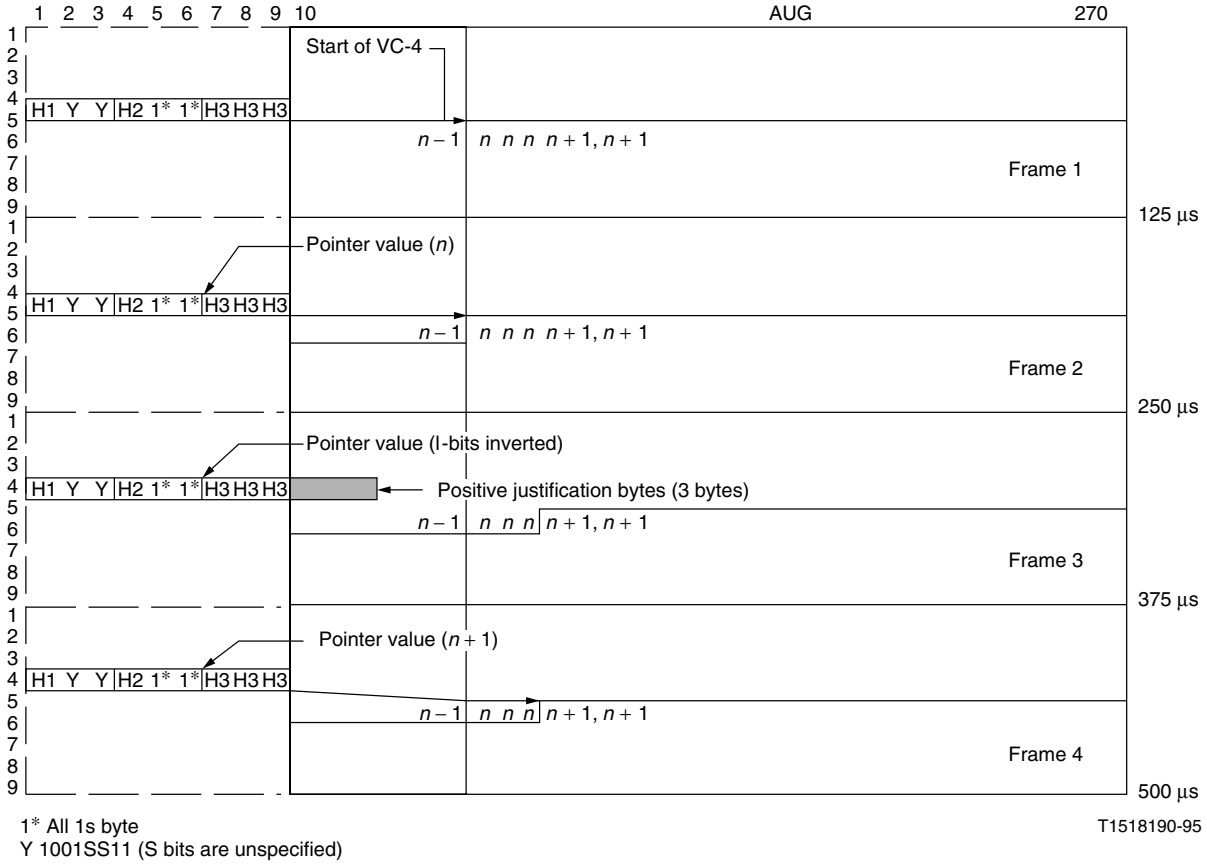


Figure 3.23. AU-4 pointer adjustment operation, positive justification. (From Ref. 5, Fig. 8-4/G.707, p. 37.)

Table 4.1. Summary of SDH/SONET Payloads and Mappings[a]

	SDH				SONET			
Payload	Container	Actual Payload Capacity	Payload and POH	Mapping AU-3/AU-4 Based	Container SPE	Actual Payload Capacity	Payload and POH	Mapping
DS1 (1.544)	VC-11	1.648	1.728	(AU-3),AU-4	VT 1.5	1.648	1.664	STS-1
	VC-12	2.224	2.304	AU-3,AU-4				
E1 (2.048)	VC-12	2.224	2.304	(AU-3),AU-4	VT2	2.224	2.240	STS-1
DS1C (3.152)					VT3	3.376	3.392	STS-1
DS2 (6.312)	VC-2	6.832	6.912	(AU-3),AU-4	VT6	6.832	6.848	STS-1
E3 (34.368)	VC-3	48.384	48.960	AU-3,AU-4				
DS3 (44.736)	VC-3	48.384	48.960	(AU-3),AU-4	STS-1	49.536	50.112	STS-1
E4 (139.264)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	STS-3c
ATM (149.760)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	STS-3c
ATM (599.040)	VC-4-4c	599.040	601.344	(AU-4)	STS-12c	599.040	601.344	STS-12c
FDDI (125.000)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	STS-3c
DQDB (149.760)	VC-4	149.760	150.336	(AU-4)	STS-3c	149.760	150.336	STS-3c

[a](AU-*n*) indicates compatible mapping to SONET Note 2: Numbers are in Mbit/s unit.
Source: Shafi and Siller, Ref. 3 Table 2, p. 64, 3-, reprinted with permission.

is indicated. Subsequent pointer values shall be incremented by one.

4. If the majority of the *D* bits of the pointer word is inverted, a negative justification operation is indicated. Subsequent pointer values shall be decremented by one.
5. If the NDF is interpreted as enabled, the coincident pointer value shall replace the current one at the offset indicated by the new pointer value unless the receiver is in a state that corresponds to a loss of pointer.

4. SONET/SDH SUMMARY

Table 4.1 summarizes SDH/SONET payloads and mappings. Table 4.2 reviews the various overheads for SDH STM-1 and SONET STS-3c.

Table 4.2. Overhead Summary, STM-1, STS-3c

Framing A1	Framing A1	Framing A1	Framing A2	Framing A2	Framing A2	STS-1 ID C1	STS-1 ID C1	STS-1 ID C1		Trace J1	
BIP-8 B1			Order-wire E1			User F1			For STS-3c only, not included in STM-1	BIP-8 B3	
Data Communication D1			Data Communication D2			Data Communication D3				Sig Label C2	
Pointer H1	1001 ss11	1001 ss11	Pointer H2	1111 1111	1111 1111	Ptr Action H3	Ptr Action H3	Ptr Action H3		Path status G1	
BIP-8 B2	BIP-8 B2	BIP-8 B2	APS K1			APS K2				User F2	Path Overhead
Data Communication D4			Data Communication D5			Data Communication D6				Multiframe H4	
Data Communication D7			Data Communication D8			Data Communication D9				Growth Z3	
Data Communication D10			Data Communication D11			Data Communication D12				Growth Z4	
Growth Z1	Growth Z1	Growth Z1	Growth Z2	Growth Z2	Growth Z2	Orderwire E2				Growth Z5	

Transport overhead | Payload capacity

Source: Siller and Shafi, Ref. 3, Table 3, p. 64, reprinted with permission.

BIOGRAPHY

Roger Freeman has over 50 years experience in telecommunications including a stint in the US Navy and radio officer on merchant vessels. He attended Middlebury College and has two degrees from New York University. He has had assignments with the Bendix Corporation in Spain and North Africa which was followed by five years as a member technical staff for ITT Communications Systems. Roger then became manager of microwave systems for CATV extension at Jerrold Electronics Corporation followed by assignments at Page Communications Engineers in Washington, DC where he was a project engineer on earth stations and on various data communication programs. During this period he was assigned by the ITU as Regional Planning Expert for northern Sourth America based in Quito, Ecuador. From Quito he took a position with ITT at their subsidiary in Madrid, Spain where he did consulting in telecommunication planning. In 1978 he joined the Raytheon Company as principal engineer in their Communication Systems Directorate where he held design positions on military communications. At the same time he taught various telecommunication courses in the evenings at Northeastern University and 4-day seminars at the University of Wisconsin. These seminars were based on his several textbooks on telecommunications published by John Wiley & Sons, New York. He also gives telecommunication seminars (in Spanish) in Monterrey, Mexico City and Caracas. Roger is a contributor and guest editor (Desert Storm edition) of the IEEE Communications magazine and was advanced by the IEEE to senior life member in 1994. He served on the board of directors of the Spain Section of the IEEE and was its secretary for four years. In 1991 Roger took early retirement from the Raytheon Company and organized Roger Freeman Associates, Independent Consultants in Telecommunications. The group has undertaken over 50 assignments from Alaska to South America.

Roger may be reached at *rogerf67@cox.net*; his website is *www.rogerfreeman.com*. Also of interest would be *www.telecommunicationbooks.com* where the reader may subscribe to the on-line Reference Manual for Telecommunication Engineering, 3rd ed, updated quarterly.

BIBLIOGRAPHY

1. *Synchronous Optical Network* (*SONET*)—*Basic Description Including Multiplex Structure, Rates and Formats*, ANSI T1.105-1995, ANSI New York, 1995.
2. R. L. Freeman, *Reference Manual for Telecommunications Engineering*, 3rd ed., Wiley, New York, 2001.
3. C. A. Siller and M. Shafi, eds., *SONET/SDH*, IEEE Press, New York, 1996.
4. R. L. Freeman, *Telecommunication Transmission Handbook*, 4th ed., Wiley, New York, 1998.
5. *Network-Node Interface for the Synchronous Digital Hierarchy* (*SDH*), ITU-T Rec. G.707, ITU Geneva, March, 1996.
6. *Synchronous Optical Network* (*SONET*), *Transport Systems, Common Generic Criteria*, Telcordia GR-253-CORE, Issue 2, Rev. 2, Piscataway, NJ, Jan. 1999.
7. *Introduction to SONET*, Seminar, Hewlett-Packard, Burlington, MA, Nov. 1993.
8. *SONET Add-Drop Multiplex Equipment* (*SONET ADM*) *Generic Criteria*, Bellcore, TR-TSY-000496, Issue 2, Bellcore, Piscataway, NJ, 1989.
9. *Automatic Protection Switching for SONET*, Telcordia Special Report SR-NWT-001756, Issue 1, Telcordia, Piscataway, NJ, Oct. 1990.
10. *SONET Dual-Fed Unidirectional Path Switched Ring* (*UPSR*) *Equipment Generic Criteria*, Telcordia GR-1400-CORE, Issue 2, Telcordia, Piscataway, NJ, Jan. 1999.
11. *SONET Bidirectional Line-Switched Ring Equipment Generic Criteria*, Telcordia GR-1230-CORE, Issue 4, Telcordia, Piscataway, NJ, Dec. 1998.
12. *Telcordia Notes on the Synchronous Optical Network* (*SONET*), Special Report, SR-NOTES-Series-01, Issue 1, Piscataway, NJ, Dec. 1999.

T

TAILBITING CONVOLUTIONAL CODES

MARC HANDLERY
ROLF JOHANNESSON
PER STÅHL
Lund University
Lund, Sweden

1. INTRODUCTION

Error-correcting codes should protect digital data against errors that occur on noisy communication channels. There are two basic types of error-correcting codes: block codes and convolutional codes, which are similar in some ways but differ in many other ways. For simplicity, we consider only binary codes.

We first consider *block codes* and divide the entire sequence of data or information digits into *blocks* of length K called *information words*. Then the *block encoder* maps the set of information words one to one to the set of *codewords* of *blocklength* N, where $N \geq K$. The block code $\mathcal{B}$ is the set of $M = 2^K$ codewords and the block code *rate* $R = K/N$ is the fraction of binary digits in the codeword that is needed to represent the information word; the remaining fraction, $1 - R$, represents the *redundancy* that can be used to combat channel noise. The theory of block codes is very rich, and many algebraic concepts have been used to design block codes with mathematical structures that can be exploited in efficient decoding algorithms. Block codes are used in many applications; for example, all compact-disk (CD) players use a very powerful block code called the *Reed–Solomon code*.

The encoder for a convolutional code maps the sets of information words (in this context also called an *information sequences*) one to one to the set of codewords. An information word is regarded as a continuous sequence of b-tuples, and a codeword is regarded as a continuous sequence of c-tuples. The convolutional encoder introduces dependencies between the codeword c-tuples; the present codeword c-tuple depends not only on the corresponding information word b-tuple but also on the m (memory) previous b-tuples. The *convolutional code* is the (infinite) set of infinitely long codewords and the convolutional code rate is $R = b/c$. For block codes, K and N are usually large, whereas for convolutional codes, b and c are usually small. Convolutional codes are routinely used in many applications, for example, mobile telephony and modems.

A major drawback of many block codes is that it is difficult to make use of soft-decision information provided by a channel (or demodulator) that, instead of outputting a hard decision of a channel symbol, outputs likelihood information. Such channels are called *soft-output channels* and are often better models of the situation encountered in practice. For convolutional codes there exist decoding algorithms that can easily exploit the soft-decision information provided by a soft-output channel. This is one reason why convolutional codes are often used in practice.

Tailbiting convolutional codes, in the sequel simply called *tailbiting codes*, are block codes that are obtained from convolutional codes and can be regarded as a link between these two types of codes. They inherit many properties, such as distance properties and the error-correcting capability, from convolutional codes. Many decoding algorithms for convolutional codes that make use of soft-decision information can be extended to tailbiting codes.

The basic idea of obtaining good block codes from convolutional codes by using the tailbiting method was first presented by Solomon and van Tilborg [1]. The term *tailbiting convolutional code* was, however, not used until Ma and Wolf presented their results in 1986 [2]. It has since been shown that many good block codes can be considered as tailbiting codes. Tailbiting codes are also an interesting option in practical applications where information is to be transmitted in rather large blocks and where the advantages of convolutional codes can be exploited.

2. TERMINATING CONVOLUTIONAL CODES

In order to explain the tailbiting termination method, a short introduction to convolutional codes and convolutional encoders is first given. A rate $R = b/c$, binary convolutional encoder has b inputs and c outputs and encodes a semiinfinite information sequence $\mathbf{u} = \mathbf{u}_k\mathbf{u}_{k+1}\mathbf{u}_{k+2}\ldots$, where $\mathbf{u}_i = (u_i^{(1)}\ u_i^{(2)}\cdots u_i^{(b)})$ is a binary b-tuple and k is an integer (positive or negative), into a code sequence $\mathbf{v} = \mathbf{v}_k\mathbf{v}_{k+1}\mathbf{v}_{k+2}\cdots$, where $\mathbf{v}_i = (v_i^{(1)}\ v_i^{(2)}\cdots v_i^{(c)})$ is a binary c-tuple. For simplicity we will first only consider convolutional encoders without feedback. In this case the code symbol $\mathbf{v}_t$ at time t depends on both the information symbol $\mathbf{u}_t$ and the m previous information symbols. The encoding rule can be written as

$$\mathbf{v}_t = \mathbf{u}_t G_0 + \mathbf{u}_{t-1} G_1 + \cdots + \mathbf{u}_{t-m} G_m \tag{1}$$

where $t \geq k$ is an integer and $\mathbf{u}_j = 0$ when $j < k$. The parameter m is called the *memory* of the encoder and G_i, $0 \leq i \leq m$, is a binary $b \times c$ matrix. The arithmetic in (1) is in the binary field $\mathbb{F}_2$. A convolutional code encoded by a convolutional encoder is the set of all code sequences $\mathbf{v}$ obtained from the encoder resulting from all possible different information sequences $\mathbf{u}$. For simplicity we will assume here that $k = 0$.

The convolutional encoder can be regarded as a linear finite-state machine. Figure 1 shows an example of a code rate $R = \frac{1}{2}$ convolutional encoder. It has the following encoding rule

$$\mathbf{v}_t = \mathbf{u}_t(1\ 1) + \mathbf{u}_{t-1}(1\ 0) + \mathbf{u}_{t-2}(1\ 1) \tag{2}$$

where $\mathbf{u}_j$, $j = t, t-1, t-2$, is a binary 1-tuple, and its memory is $m = 2$. The encoder has two delay elements;

hence, at each time t the encoder can be in four different *states* depending on the two previous information symbols. We denote the state at time t by $\mathbf{s}_t$, the set of all possible encoder states by $\mathcal{S}$, and the number of possible encoder states by n_s. A convolutional encoder is assumed to start in the all-zero state: $\mathbf{s}_0 = \mathbf{0}$. A trellis [3] describes all possible state sequences $\mathbf{s}_0\mathbf{s}_1\mathbf{s}_2\cdots$ of the convolutional encoder. The trellis for the rate $R = \frac{1}{2}$ memory $m = 2$ encoder in Fig. 1 is shown in Fig. 2. From each state in the trellis there are at each time instant two possible state transitions. The state transitions are represented by branches. The upper branch corresponds to the encoder input 0 and the lower one, to the encoder input 1. The symbols on each branch denote the encoder output.

A convolutional code is a set of (semi)infinitely long code sequences resulting from (semi)infinite strings of data. However, since data usually are transmitted in packets of a certain size and not as a stream of symbols, it is in a practical situation necessary to split the infinite datastream into blocks, and let each block of data be separately encoded into a block of code symbols by the convolutional encoder. We say that the convolutional code is *terminated* into a block code. An encoder for an (N, K) block code maps blocks of K information bits $\mathbf{u} = (u_0\ u_1 \cdots u_{K-1})$ one to one to codewords $\mathbf{v} = (v_0\ v_1 \cdots v_{N-1})$ of N bits.

The straightforward method for terminating a convolutional code into a block code is to use the termination method called *direct truncation*. Assume that a rate $R = b/c$ convolutional encoder with memory m is used to encode a block of $K = Lb$ information bits $\mathbf{u} = (\mathbf{u}_0\ \mathbf{u}_1 \cdots \mathbf{u}_{L-1})$, where $\mathbf{u}_i = (u_i^{(1)}\ u_i^{(2)} \cdots u_i^{(b)})$. Using direct truncation, the convolutional encoder is started in a fixed state, and is simply fed with the L information b-tuples. The resulting output is a codeword $\mathbf{v} = (\mathbf{v}_0\ \mathbf{v}_1\ \ldots\ \mathbf{v}_{L-1})$, where $\mathbf{v}_i = (v_i^{(1)}\ v_i^{(2)} \cdots v_i^{(c)})$, consisting of $N = Lc$ bits and the set

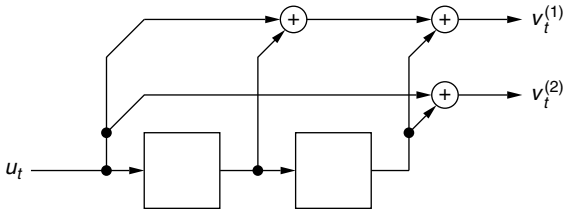


Figure 1. A rate $R = \frac{1}{2}$ memory $m = 2$ convolutional encoder.

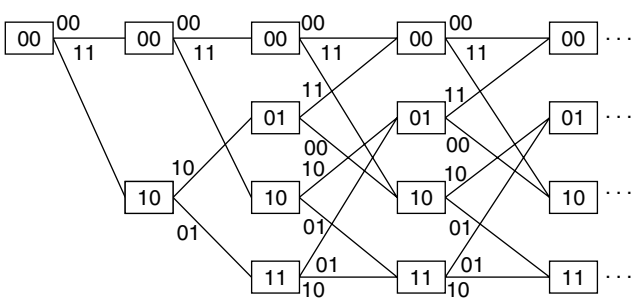


Figure 2. The trellis for the encoder in Fig. 1.

of codewords resulting from all the $M = 2^K$ possible different information blocks makes up a linear block code. The blocklength of this code is $N = Lc$, and the code's rate is $R_{\mathrm{dt}} = K/N = b/c$.

The disadvantage with the direct-truncation method is that the last information b-tuples are less protected than the other data bits. A remedy for this is to feed the encoder with m additional b-tuples, after the information symbols, which forces the encoder in to an advanced determined encoder state (known to the decoder). These mb symbols carry no information and are called "dummy symbols." Usually the predetermined ending state is chosen to be the all-zero state, and we reach this state by feeding the (feedback-free) encoder by m dummy zero b-tuples. Hence, this second termination method is called the *zero-tail method*. With this method also the last information symbols are protected, and this is the most common method used to terminate a convolutional code. The block code obtained with the zero-tail method has blocklength $N = (L + m)c$ and rate

$$R_{\mathrm{zt}} = \frac{K}{N} = \frac{L}{L+m}\frac{b}{c} = \frac{L}{L+m}R \tag{3}$$

which is less than the rate $R = b/c$ of the convolutional encoder. The rate loss $L/(L + m)$, caused by the m dummy b-tuples, is negligible if $L \gg m$, but if L is small, that is, if the data arrives in short packets, this rate loss might not be acceptable.

In the third method, called *tailbiting*, this rate loss is removed. In order to protect all data bits equally, we impose the restriction that the rate $R = b/c$ convolutional encoder should start and, after feeding it with the $K = Lb$ information bits $\mathbf{u}$, end in the same encoder state. To avoid the use of dummy symbols, this starting–ending state is *not* fixed, but depends on the actual information sequence to be encoded. Since no dummy symbols are used, the codewords have length $N = Lc$, and the rate of the block code obtained is the same as for the encoder:

$$R_{\mathrm{tb}} = \frac{K}{N} = \frac{b}{c} = R \tag{4}$$

We have no rate loss. The following example illustrates our three different termination methods.

Example 1. Assume that the rate $R = \frac{1}{2}$ encoder with memory $m = 2$ in Fig. 1 is used to encode a block of 4 information bits. Figure 3a–c shows the trellis when direct truncation, zero-tail, and tailbiting termination types are used, respectively. With direct truncation the encoder starts in the all-zero state and can end in any of the four encoder states, whereas with zero-tail termination the encoder can start and end in one predetermined state, that is, in the all-zero state only. With tailbiting the encoder, as with zero-tail termination, also ends in the same state as it started in, however, all four encoder states are possible starting-ending states.

Consider the encoding of the information block $\mathbf{u} =$ (1 0 0 1) using the tailbiting technique. If the encoder is loaded with (starts in) state 10, then after feeding it with the information sequence $\mathbf{u}$, the encoder again ends in

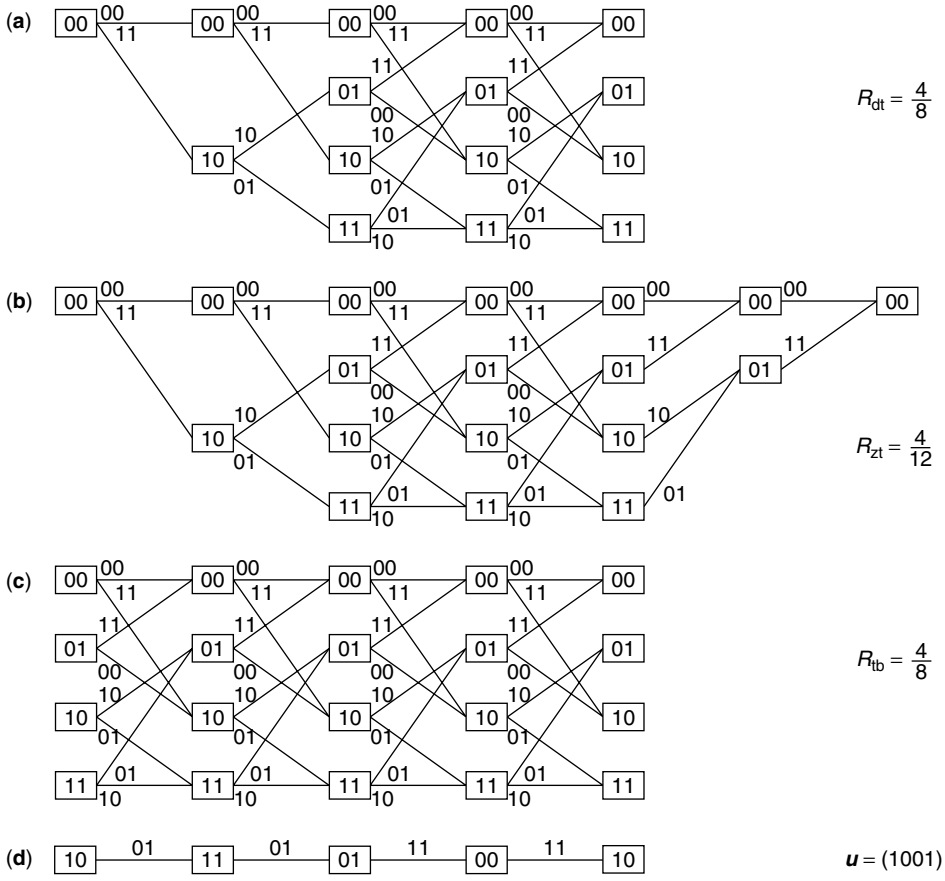


Figure 3. Trellises for encoding a block of 4 information bits using the encoder in Fig. 1 and (**a**) direct truncation, (**b**) zero-tail termination, (**c**) tailbiting termination. In (**d**) the encoder state sequence when encoding $\mathbf{u} = (1\ 0\ 0\ 1)$ using tailbiting is shown.

state 10; that is, the tailbiting restriction is fulfilled. The corresponding codeword is $\mathbf{v} = (01\ 01\ 11\ 11)$. The encoder state sequence is shown in Fig. 3d.

We call a linear block code obtained by terminating a convolutional code using the tailbiting termination method a *tailbiting code*, and the number of encoded b-tuples L is called the *tailbiting length*. The corresponding trellis is called a *tailbiting trellis*. Since the convolutional encoder starts and ends in the same state, we can view the tailbiting trellis as a *circular* trellis. Figure 4 shows a circular trellis of length $L = 6$ for a tailbiting code encoded by a rate $R = 1/c$ encoder with four states. Every valid codeword in a tailbiting code corresponds to a circular path in the circular trellis. From the regular structure of the circular trellis obtained from a time-invariant convolutional encoder, it follows that if a codeword is cyclically shifted c symbols (corresponding to one cyclic step in the circular trellis), the result is also a codeword in the tailbiting code. Block codes with this property are called *quasicyclic* codes. Hence, all $R = b/c$ tailbiting codes obtained from time-invariant convolutional encoders are quasicyclic under a cyclic shift of c symbols.

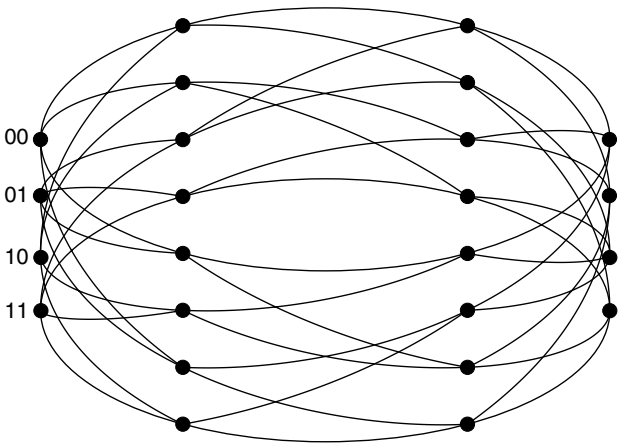


Figure 4. A circular trellis for a tailbiting code of length $L = 6$ encoded by a rate $R = 1/c$ encoder with four states.

3. ENCODING TAILBITING CODES

Using the tailbiting termination method we must find the correct starting state for each information word to be encoded such that the encoder starts and ends in the same state. For feedback-free encoders realized in controller canonical form, like the encoder in Fig. 1, this is very easy. Consider the encoding of the information word $\mathbf{u} = (1\ 0\ 0\ 1)$ in Example 1. After feeding the convolutional encoder with the 4 information bits, it ends in state (1 0), which is simply the 2 last information bits in reversed order, and, hence, if this state is chosen as starting state, then the tailbiting criterion will be fulfilled. Similarly, for feedback-free encoders realized in controller canonical form the encoder starting state is the reverse of the m last

information b-tuples $\mathbf{u}_{L-1} \cdots \mathbf{u}_{L-m}$. If the convolutional encoder has feedback, the starting state does not depends not only on the m last information b-tuples but, in general, also on all information bits. The starting state can be obtained by performing a simple matrix multiplication as shown in Ref. 4, where finding the starting state of a convolutional encoder realized in observer canonical form is also investigated. Another method to find the starting state of a feedback convolutional encoder when tailbiting is used is given by Weiss et al. [5]. This method requires that the encoding is performed twice: once in order to find the correct starting state and a second time for the actual encoding.

From Eq. (1) and the fact that the starting state is given by the reverse of the last m information b-tuples to be encoded, it follows that for feedback-free encoders

$$\mathbf{v}_t = (v_t^{(1)}\ v_t^{(2)} \cdots v_t^{(c)}) = \sum_{k=0}^{m} \mathbf{u}_{((t-k))} G_k \tag{5}$$

where the double parentheses denote modulo L arithmetic on the indices; that is, $((t-k)) \equiv t-k \pmod{L}$. Then the encoding of a tailbiting code using a feedback-free encoder can be compactly written as

$$\mathbf{v} = \mathbf{u}\mathbf{G}^{\text{tb}} \tag{6}$$

where

$$\mathbf{G}^{\text{tb}} = \begin{pmatrix} G_0 & G_1 & G_2 & \dots & G_m & & & \\ & G_0 & G_1 & G_2 & \dots & G_m & & \\ & & \ddots & \ddots & \ddots & & \ddots & \\ & & & G_0 & G_1 & G_2 & \dots & G_m \\ G_m & & & & G_0 & G_1 & \dots & G_{m-1} \\ G_{m-1} & G_m & & & & G_0 & \dots & G_{m-2} \\ \vdots & & \ddots & & & & \ddots & \vdots \\ G_1 & G_2 & \dots & G_m & & & & G_0 \end{pmatrix} \tag{7}$$

is an $L \times L$ matrix and where each entry is a $b \times c$ matrix.

Example 2. The generator matrix that corresponds to the tailbiting encoding in Example 1 is

$$\mathbf{G}^{\text{tb}} = \begin{pmatrix} 11 & 10 & 11 & 00 \\ 00 & 11 & 10 & 11 \\ 11 & 00 & 11 & 10 \\ 10 & 11 & 00 & 11 \end{pmatrix}. \tag{8}$$

We verify that when encoding $\mathbf{u} = (1\ 0\ 0\ 1)$ the corresponding codeword is $\mathbf{v} = (1\,0\,0\,1)\mathbf{G}^{\text{tb}} = (01\,01\,11\,11)$.

Often it is convenient to express the information word and codeword in terms of the delay operator D:

$$\mathbf{u}(D) = \mathbf{u}_0 + \mathbf{u}_1 D + \cdots + \mathbf{u}_{L-1} D^{L-1} \tag{9}$$

$$\mathbf{v}(D) = \mathbf{v}_0 + \mathbf{v}_1 D + \cdots + \mathbf{v}_{L-1} D^{L-1} \tag{10}$$

Each convolutional encoder can also be described by a generator matrix $G(D)$ using the delay operator. For example, the encoder in Fig. 1 has generator matrix $G(D) = (1 + D + D^2\ 1 + D^2)$. If the convolutional encoder has feedback, the entries of the generator matrix are rational functions. Using the delay operator, we can write the tailbiting encoding procedure in compact form as

$$\mathbf{v}(D) \equiv \mathbf{u}(D)G(D) \pmod{1 + D^L} \tag{11}$$

where $G(D)$ is a rational generator matrix.

Example 3. We repeat the calculations in Example 2 using (11). The generator matrix is $G(D) = (1 + D + D^2\ 1 + D^2)$, and the information sequence $u(D) = 1 + D^3$ gives the codeword

$$\begin{aligned} \mathbf{v}(D) &\equiv (1 + D^3)(1 + D + D^2\ 1 + D^2) \\ &\equiv (D^2 + D^3\ 1 + D + D^2 + D^3) \\ &\equiv (01) + (01)D + (11)D^2 + (11)D^3 \pmod{1 + D^4} \end{aligned} \tag{12}$$

Example 4. Consider tailbiting of length $L = 4$ using the (systematic) feedback encoder shown in Fig. 5. It is equivalent to the encoder in Fig. 1 and its generator matrix is

$$G(D) = \left(1 \frac{1 + D^2}{1 + D + D^2}\right) \tag{13}$$

The information sequence $u(D) = 1 + D + D^3$ gives the codeword

$$\begin{aligned} \mathbf{v}(D) &\equiv (1 + D + D^3)\left(1\ \frac{1 + D^2}{1 + D + D^2}\right) \\ &\equiv (1 + D + D^3)(1\ (1 + D^2)(1 + D^2 + D^3)) \\ &\equiv (1 + D + D^3)(1\ D + D^3) \equiv (1 + D + D^3\ D + D^3) \\ &\equiv (10) + (11)D + (00)D^2 + (11)D^3 \pmod{1 + D^4} \end{aligned} \tag{14}$$

where we have assumed that $(1 + D + D^2)^{-1} \equiv 1 + D^2 + D^3 \pmod{1 + D^4}$.

For certain encoders at certain tailbiting lengths the tailbiting technique will fail to work; we do not have a one-to-one mapping between the information sequences and the codewords.

Example 5. Assume that the feedback encoder in Fig. 5 is used for tailbiting. We start the encoder in state (0 1)

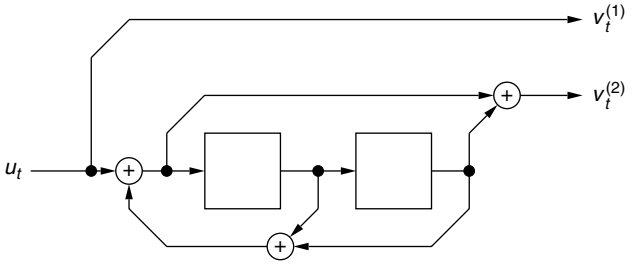


Figure 5. A rate $R = \frac{1}{2}$ (systematic) feedback convolutional encoder.

and feed it with zeros. The encoder passes the states (1 0) and (1 1) and after three steps the encoder will be back again in state (0 1). The corresponding encoder output is 00 01 01. If the tailbiting length L is a multiple of 3, the repetition of this cycle $L/3$ times will be a tailbiting path corresponding to a nonzero codeword. Since the all-zero codeword always corresponds to the all-zero input, we will have more than one codeword corresponding to an all-zero input; that is, one information sequence corresponds to at least two codewords; hence, tailbiting fails when L is a multiple of 3.

The generator matrix $G(D)$ used to generate a tailbiting code of length L can be decomposed in its invariant factor decomposition $G(D) = A(D)\Gamma(D)B(D)$ (see, for example, [6]), where $A(D)$ and $B(D)$ are $b \times b$ and $c \times c$ binary polynomial matrices with unit determinants, respectively, and where $\Gamma(D)$ is the $b \times c$ matrix

$$\Gamma(D) = \begin{pmatrix} \dfrac{\gamma_1(D)}{q(D)} & & & \\ & \ddots & & \\ & & \dfrac{\gamma_b(D)}{q(D)} & 0 \quad \dots \quad 0 \end{pmatrix} \tag{15}$$

where $q(D)$ and $\gamma_i(D)$, $i = 1, 2, \dots, b$, are binary polynomials. It was shown [4] that if and only if both $q(D)$ and $\gamma_b(D)$ are relatively prime to $1 + D^L$, then (11) describes a one-to-one mapping between $\mathbf{u}(D)$ and $\mathbf{v}(D)$ for all $\mathbf{u}(D)$.

Example 6. Consider the feedback encoder in Fig. 5 with generator matrix $G(D)$ given in (13). Using the invariant factor decomposition, we can write

$$G(D) = (1) \begin{pmatrix} \dfrac{1}{1+D+D^2} & 0 \end{pmatrix} \begin{pmatrix} 1+D+D^2 & 1+D^2 \\ 1+D & D \end{pmatrix} \tag{16}$$

and, hence, $q(D) = 1 + D + D^2$ and $\gamma_b(D) = 1$. Since $q(D)$ divides $1 + D^{3k}$, where $k = 1, 2, \dots$, it follows in agreement with Example 5 that the tailbiting technique fails for $L = 3k$.

4. DECODING TAILBITING CODES

Assume that the information block $\mathbf{u}$ is encoded as the codeword $\mathbf{v}$ and then transmitted over a noisy channel. At the receiver side, the decoder is provided with a noise-corrupted version of $\mathbf{v}$ that we denote by $\mathbf{r}$. The task of the decoder is to make an estimate $\hat{\mathbf{u}}$ of the information block $\mathbf{u}$ based on the received $\mathbf{r}$. A decoder that chooses $\hat{\mathbf{u}}$ in such a manner that

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{x}} \mathrm{P}(\mathbf{r} \mid \mathbf{u} = \mathbf{x}) \tag{17}$$

where the maximum is taken over all possible information blocks, is called a (sequence) *maximum-likelihood* (*ML*) *decoder*. If all possible information blocks are equally likely to occur, then an ML decoder minimizes the probability that $\hat{\mathbf{u}} \neq \mathbf{u}$; thus, an ML decoder minimizes the decoding block error probability. A drawback of the ML decoder is that it does not provide any information on how reliable the estimation $\hat{\mathbf{u}}$ is. A decoder that computes

$$\mathrm{P}(u_t^{(i)} = x \mid \mathbf{r}), \; x \in \{0, 1\}$$

for all t and i, is called a (symbol-by-symbol) *a posteriori probability* (*APP*) *decoder*. The estimated information bit $\hat{u}_t^{(i)}$ is then chosen to be the binary digit x maximizing the probability $\mathrm{P}(u_t^{(i)} = x \mid \mathbf{r})$; the probability $\mathrm{P}(u_t^{(i)} = x|\mathbf{r})$ also provides valuable information on the reliability of the estimation $\hat{u}_t^{(i)}$, which is crucial in many modern coding schemes, including Turbo codes.

A code described via a trellis with only one possible starting state (as the example in Fig. 3b) may be decoded by trellis-based algorithms that can easily make use of the received soft channel output. Examples of such algorithms are the Viterbi decoder [7], which is an ML decoder, and the two-way [(BCJR) (Bahl–Cocke–Jelinek–Raviv)] decoder [8], which is an APP decoder. The main problem when decoding tailbiting codes is that the decoder does not know the encoder starting–ending state (since the starting–ending state depends on the information sequence to be transmitted), and, hence, the Viterbi decoder and two-way decoder cannot be used in their original forms. These algorithms, however, can be applied to tailbiting codes described via their tailbiting trellises with n_s possible encoder starting-ending states as follows. Let us divide the set of tailbiting codewords into n_s disjoint subsets or subcodes $\mathcal{C}_{\mathbf{s}}^{tb}$, where $\mathbf{s} \in \mathcal{S}$ and $\mathcal{C}_{\mathbf{s}}^{tb}$ is the set of codewords that correspond to trellis paths starting and ending in the state $\mathbf{s}$. Hence, each subcode is characterized by its encoder starting-ending state and corresponds to a trellis with one starting and one ending state. Each of these subcodes may be decoded by the Viterbi algorithm producing n_s candidate estimates of $\mathbf{u}$. An ML decoder for tailbiting codes then chooses the best one, namely, the one with highest probability, among these n_s candidates as its output. Using a similar approach, we can also implement an a posteriori *probability* (APP) decoder [6].

The complexities of the described ML and APP decoders for tailbiting codes are of the order of n_s^2. For a rate $R_{\mathrm{tb}} = b/c$ tailbiting code with memory m the number of possible encoder states n_s can be as large as 2^{mb}, and hence the complexity of the ML and APP decoders are then of the order of 2^{2mb}. Because of this high computational complexity, suboptimal decoding algorithms, that despite their lower complexities achieve performances close to those of the ML and APP decoders, have been widely explored.

Many suboptimal decoding methods make use of the circular trellis representations of tailbiting codes. The basic idea to circumvent the problem of the unknown encoder starting–ending state is to go around and around the circular trellis while decoding until a decoding decision is reached (see Fig. 6). The longer the decoder goes around the circular trellis, the less influence the unknown encoder starting state has on the performance of the decoder. There are many near-ML decoding algorithms using the Viterbi algorithm when circling around the tailbiting trellis; a nonexhaustive list of such decoders

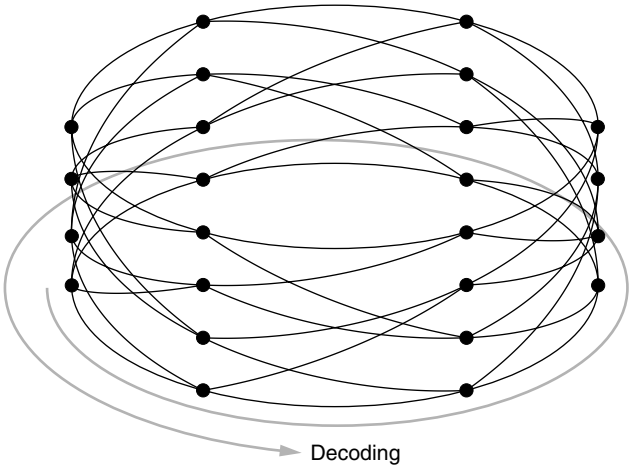


Figure 6. Illustration of a suboptimal decoding algorithm going around and around a circular trellis.

is given by Calderbank et al. [9]. An approximate APP decoding algorithm using the two-way (BCJR) algorithm when circling around the tailbiting trellis was introduced by Anderson and Hladik [10]. Experiments have shown that often very few decoding cycles suffice to achieve satisfactory results. The decoding performance may, however, be significantly affected by *pseudocodewords* corresponding to trellis paths of more than one cycle that do not pass through the same trellis state at any integer multiple of the cycle length other than the pseudocodeword length. The described suboptimal tailbiting decoders going around and around the circular trellis have computational complexities of the order of 2^{mb}, which is the square root of the decoding complexities of the optimal algorithms.

5. APPLICATIONS OF TAILBITING CODES

The class of tailbiting codes includes many powerful block codes, as, for example, the binary (24, 12, 8) Golay code [9]. Furthermore, it was shown [11] that many of the best tailbiting codes have minimum distances as large as the best linear codes. Since for tailbiting codes regular trellis structures are directly obtained from their generators, these codes are often used when communicating over soft-output channels.

Using tailbiting codes as component codes in concatenated coding schemes is often an attractive option. For example, concatenated codes with outer Reed–Solomon codes and inner tailbiting codes have been adopted in the IEEE Standard 802.16-2001 for local and metropolitan area networks and in the ETSI EN 301 958 standard for digital videobroadcasting (DVB). Since tailbiting codes are easily decoded by (approximate) APP decoders, tailbiting codes are attractive component codes in concatenated coding schemes using iterative decoders [5].

BIOGRAPHIES

Marc Handlery received the Diploma degree in electrical engineering from the Swiss Federal Institute of Technology (ETH) Zúrich, Switzerland, in 1999. He is currently working at the Department of Information Technology at Lund University, Lund, Sweden, where he also received the Ph.D. degree in 2002. His research interests include convolutional codes, concatenated codes, tailbiting codes, and their trellis representations.

Rolf Johannesson received the M.S. and Ph.D. degrees in 1970 and 1975, respectively, both from Lund University, Lund, Sweden. He was awarded the degree of Professor, honoris causa, from the Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, in 2000. Since 1976, he has been with Lund University, where he is now Professor of Information Theory. His scientific interests include information theory, error-correcting codes, and cryptography. In addition to papers and book chapters in the area of convolutional codes and cryptography, he has authored two textbooks on switching theory and digital design and one on information theory and coauthored *Fundamentals of Convolutional Coding*.

Per Ståhl received his M.S. degree in electrical engineering in 1997 and his Ph.D. degree in information theory in 2001 from Lund University, Sweden. His research interests include tailbiting trellises and structure problems in error correcting codes. Since January 2002 he has been working with data security at Ericsson Mobile Platforms AB in Lund.

BIBLIOGRAPHY

1. G. Solomon and H. C. A. van Tilborg, A connection between block and convolutional codes, *SIAM J. Appl. Math.* **37**: 358–369 (Oct. 1979).
2. H. H. Ma and J. K. Wolf, On tail biting convolutional codes, *IEEE Trans. Commun.* **COM-34**: 104–111 (Feb. 1986).
3. G. D. Forney, Jr., *Review of Random Tree Codes*, NASA Ames Research Center, Contract NAS2-3637, NASA CR 73176, Final Report, Appendix A, Dec. 1967.
4. P. Ståhl, J. B. Anderson, and R. Johannesson, A note on tailbiting codes and their feedback encoders, *IEEE Trans. Inform. Theory* 529–534 (Feb. 2002).
5. C. Weiss, C. Bettstetter, and S. Riedel, Code construction and decoding of parallel concatenated tail-biting codes, *IEEE Trans. Inform. Theory* **IT-47**: 366–386 (Jan. 2001).
6. R. Johannesson and K. Sh. Zigangirov, *Fundamentals of Convolutional Coding*, IEEE Press, Piscataway, NJ, 1999.
7. A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**: 260–269 (April 1967).
8. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).
9. A. R. Calderbank, G. D. Forney, Jr., and A. Vardy, Minimal tail-biting trellises: The Golay code and more, *IEEE Trans. Inform. Theory* **IT-45**: 1435–1455 (July 1999).
10. J. B. Anderson and S. M. Hladik, Tailbiting MAP decoders, *IEEE J. Select. Areas Commun.* **16**: 297–302 (Feb. 1998).
11. I. E. Bocharova, R. Johannesson, B. D. Kudryashov, and P. Ståhl, Tailbiting codes: Bounds and search results, *IEEE Trans. Inform. Theory* **48**: 137–148 (Jan. 2002).

TELEVISION AND FM BROADCASTING ANTENNAS

HARUO KAWAKAMI
YASUSHI OJIRO
Antenna Giken Corp. Laboratory
Saitama City, Japan

1. INTRODUCTION

At the present time, the superturnstile antenna (of which the batwing antenna is a radiating element) [1], the supergain antenna (dipole antennas with a reflector plate), invented in the United States; and the Vierergruppe antenna, invented in Germany and sometimes called the *two-dipole antenna* in Japan, are widely used for very high-frequency television (VHF-TV) broadcasting around the world. The superturnstile antenna (2) is used in Japan, as well as in the United States.

Figure 1 shows the appearance of the original batwing antenna, when it was first made public by Masters [1]. The secret lies in its complex shape. It is called a *batwing antenna* in the United States and *Schmetterlings antenne* (butterfly antenna) in Germany, in view of this shape.

The characteristics of the batwing antenna were calculated using the moment method proposed by Harrington [3], who conducted experiments on this antenna. The model antenna was approximately about two-fifths the size of a full-scale batwing antenna, with its design center frequency at 500 MHz.

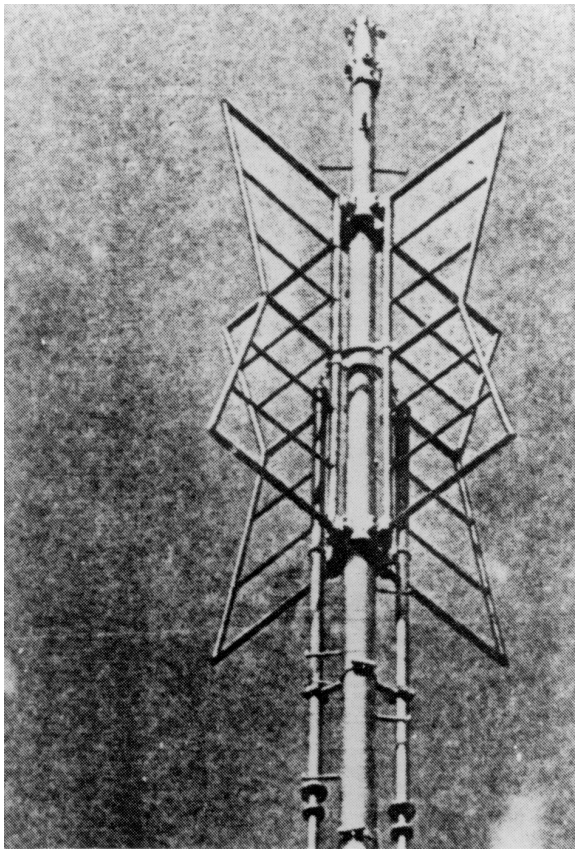


Figure 1. Historical shape of the batwing radiator.

It has been reported that for thick cylindrical antennas, a substantial effect on the current distribution appears to be due to nonzero current on the flat end faces. The assumption of zero current at the flat end face is appropriate for a thin cylindrical antenna; however, in the case of a thick cylindrical antenna, this assumption is not valid.

Specifically, this article applies the moment method to a full-wave dipole antenna, with a reflector plate supported by a metal bar, such as those widely used for TV and frequency-modulated (FM) broadcasting [4]. In the present study, the analysis is made by including the flat end-face currents. As a result, it is found that the calculated and measured values agree well, and satisfactory wideband characteristics are obtained.

Next, the twin-loop antennas, most widely used for ultra-high-frequency (UHF)- TV broadcasting, are considered. Previous researchers analyzed them by assuming a sinusoidal current distribution. Others adopted the higher-order expansions (Fourier series) of the current distribution; however, their analyses did not sufficiently explain the wideband characteristics of this antenna.

This article applies the moment method to a twin-loop antenna with a reflector plate or a wire-screen-type reflector plate. As for the input impedance, 2L-type twin-loop antennas have reactance near zero [in the case where $l_1 = 0.15\lambda_0$, i.e., where the voltage standing-wave ratio (VSWR) is nearly equal to unity]. Also, satisfactory wideband characteristics are obtained. The agreement between the measurement and the theory is quite good. Thus it may become possible in the future to improve practical antenna characteristics, based on the results obtained.

The digital terrestrial broadcasting station will use multiple channels in common. Therefore, antennas of the transmitting station will have the required properties of broad bandwidth and high gain. We have investigated the two-element modified batwing antenna with the reflector (UMBA) for UHF digital terrestrial broadcasting. It is calculated with the balanced feedshape. As a result, the broadband input impedance of 100 is obtained. However, in practical use, this antenna requires the balun and the impedance transformer as another circuit. So it has a complex feed system and high cost. In this section, we propose the unbalance-fed modified batwing antenna (UMBA), which does not use a balun and an impedance matching circuit. The same broad bandwidth and high gain compared with the previous feedshape are obtained. Next, the parallel coupling UMBA (PCUMBA) to obtain high gain is investigated. The gain of about 14 dBi is obtained for coupling four elements.

2. MOMENT METHOD

2.1. Thin Wire

This section discusses the moment method proposed by Harrington [3] and deals with the Garlerkin method where antenna current is developed using a triangle function and the same weight function as the current development function is used. The Garlerkin method can save calculation time because the coefficient matrix

is symmetric, and the triangle function is widely used because it presents a better performance in calculation time and accuracy and in other parameters for an antenna element without sudden change in antenna current.

Scattering electric field E^s by antenna current and charge is given by

$$E^s = -j\omega A - \nabla\phi \tag{1}$$

Vector potential A and scalar potential ϕ are given by

$$A = \frac{\mu}{4\pi}\iint_S J\frac{e^{-jkr_0}}{r_0}\,dS \tag{2a}$$

$$\phi = \frac{1}{4\pi\varepsilon}\iint_S \sigma\frac{e^{-jkr_0}}{r_0}\,dS \tag{2b}$$

The following relation exists between charge density σ and current J:

$$\sigma = \frac{-1}{j\omega}\nabla\cdot J \tag{2c}$$

Now, assuming that the surface of each conductor is a perfect conductor and letting E^i be the incoming electric field, the following equation must hold true:

$$n \times E^S = -n \times E^i \tag{3}$$

The following simultaneous equation holds true from the boundary condition of the antenna surface of this antenna system:

$$\sum_{i=1}^{N} I_i\langle W_j, LF_i\rangle = \langle W_j, E^i_{\tan}\rangle,$$
$$\times (i = 1,\ldots,N, \quad j = 1,\ldots,M) \tag{4}$$

where L is the operator for integration and differentiation. The current is given, from Eq. (4), by

$$[I_i] = [\langle W_j, LF_i\rangle]^{-1}[\langle W_j, E^i_{\tan}\rangle] \tag{5}$$

and the matrix representation of (5) gives:

$$[I] = [Z]^{-1}[V] \tag{6}$$

Now, let the current at point t on each element be represented by

$$I(t) = \hat{t}\sum_{i=1}^{N} I_i T_i(t) \tag{7}$$

where $\hat{t}$ is the unit vector in the direction of the antenna axis, and coefficient I_i is the complex coefficient determined by boundary condition. Letting $T_i(t)$ be the triangle development function, $T_i(t)$ is given by

$$T_i(t) = \begin{cases} 1 - \dfrac{|t - t_i|}{\Delta l_i} & t_{i-1}\langle t\langle t_{i+1} \\ 0 \end{cases} \tag{8}$$

where $\Delta l_i = t_i - t_{i-1}$ and $\Delta l_i = t_{i+1} - t_i$.

The impedance matrix Z in (6) is given by,

$$Z_{ji} = \int_{\text{axis}} dl \int_C dl' \left[j\omega\mu W_{jm}F_{\text{in}} + \frac{1}{j\omega\varepsilon}\frac{dW_{jm}}{dl}\cdot\frac{dF_{\text{in}}}{dl'}\right]\frac{e^{-jkr_0}}{4\pi r_0} \tag{9}$$

where C represents an antenna surface l' parallel to the antenna axis l. F_{in} and W_{jm} are divided into four in the triangle development function shown in Fig. 2, where the triangle is configured so that the value is one at the center and the divided antenna elements are obtained approximately by the four pulse functions. The expansion equation of (5) is used for the Green function.

2.2. Thick Wire

As shown in Fig. 3, a monopole antenna excited by a coaxial cable line consists of a perfectly conducting body of revolution being coaxial with the z axis. Conventionally, the integral equation is derived on the presumption that the tangential component of the scattered field cancels the corresponding impressed field component on the conductor surface.

Alternately, according to the boundary condition proposed by Dr. P. C. Waterman [6], an integral equation can

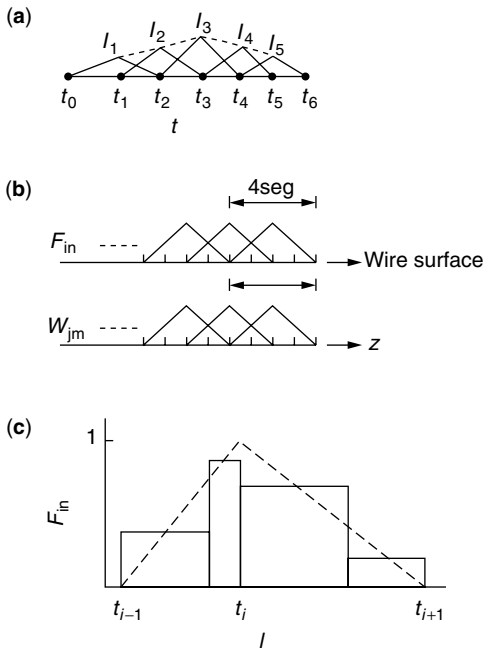


Figure 2. Approximation to the expansion and weighting function: (**a**) triangle function; (**b**) expansion and weighting function; (**c**) approximation to the expansion function F_{in}.

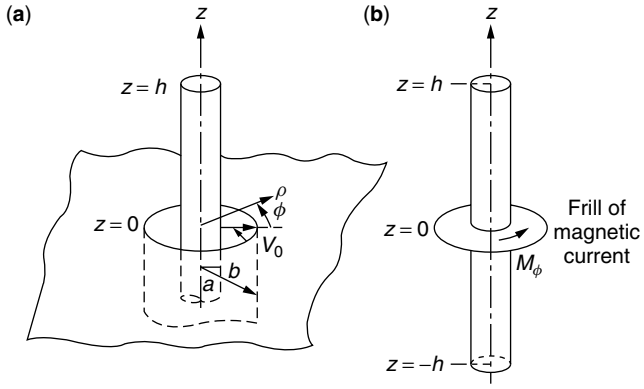


Figure 3. Coaxial cable line feeding a monopole through a ground plane and mathematical model of the antenna.

be derived by utilizing the field behavior within the conductor. Also, the conventional integral equation has a singular point when the source and the observation point are the same. On the other hand, the integral equation, after Waterman, is well behaved, and so it is more convenient for numerical calculation. When applying the extended boundary condition to an antenna having an axial symmetry as shown in Fig. 3, the axial component of the electric field is required to vanish along the axis of the conductor.

More explicitly, on the axis inside the conductor, the axial component of the total field is the sum of the scattered field (the field from current on the conductor) and the impressed field (the field from excitation). The corresponding integral equation is written as

$$\frac{j\eta}{4\pi k}\int_{-h}^{h} I_z(z')\left(k^2+\frac{\partial^2}{\partial z^2}\right)G(z,z')\,dz = E_z^{\text{inc}}$$
$$I_z(\pm h)=0 \qquad (10)$$

with

$$G(z,z') = \frac{e^{-jk\sqrt{(z-z')^2+a^2}}}{\sqrt{(z-z')^2+a^2}}, \quad k=\frac{2\pi}{\lambda},$$
$$\eta = 120\pi, \qquad I_z = 2\pi a J_z$$

This integral equation reduces to the well-known equation (11) when the current is assumed to be zero on the antenna end faces:

$$\frac{j\eta}{4\pi k}\left\{k^2\int_S J_z(z')G(z,z')\,ds' + \frac{\partial}{\partial z}\int_S \nabla'\cdot J G(z,z')\,ds'\right\} = E_S^{\text{inc}} \qquad (11)$$

Taking the current at the end faces into account, (11) becomes

$$\frac{j\eta}{4\pi k}\left\{k^2\int_{-h}^{h} I_z(z')G(z,z')\,dz' + \frac{\partial}{\partial z}\int_{-h}^{h}\frac{\partial I_z(z')}{\partial z'}G(z,z')\,dz' + \frac{\partial}{\partial z}\int_{S'}\frac{1}{\rho'}\frac{\partial J_\rho(\rho')}{\partial \rho'}G(z,z')\,ds'\right\} = E_S^{\text{inc}},$$
$$ds' = \rho' d\phi'\,d\rho' \qquad (12)$$

where, on the end surface S', a in $G(z,z')$ becomes ρ.

Dr. C. D. Taylor and Dr. D. R. Wilton [7] analyzed the current distribution on the flat end by a quasi-static-type approximation method; the resulting theoretical and experimental values were in good agreement. This analysis makes the same assumption that, as shown in Fig. 4, the current flowing axially to the center of the end surface without modification.

In applying the moment method, sinusoidal functions were used as expansion and weight functions. Accordingly, the Galerkin method was used to generate the integral equation.

Notice (Fig. 4) that the expansion and weight functions have been changed only on the dipole end faces. This is done so that the impedance matrix becomes symmetric for the end current and for the current flowing on the antenna surface (excluding the antenna end face).

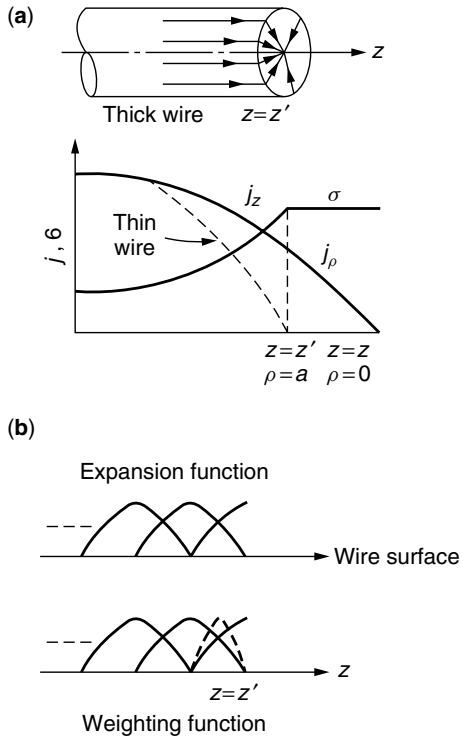


Figure 4. Current and charge for flat end faces, expansion, and weighting function.

Expanding the unknown current in terms of sine functions, we have

$$I_z(z') = \sum_{n=1}^{N} I_n F_n \qquad (13)$$

If the expansion functions are overlapped

$$F_n = \begin{cases} \dfrac{\sin k(\Delta - |z'-z_n|)}{\sin k\Delta} & z' \in (z_{n-1}, z_{n+1}) \\ 0 & z' \notin (z_{n-1}, z_{n+1}) \end{cases}$$
$$\Delta = |z_{n+1} - z_n| \qquad (14)$$

where z' indicates an axial coordinate taken along the conductor surface. Equations (13) and (14) are substituted into (12) to obtain

$$\frac{j\eta}{4\pi k \sin k\Delta}\sum_{n=1}^{N} I_n\left[\int_{Z_{n-1}}^{Z_{n+1}} G(z,z')\left(k^2+\frac{d}{dz'^2}\right)F_n\,dz' + k\{G(z,z_{n+1}) + G(z,z_{n-1}) - 2\cos k\Delta G(z,z_n)\}\right]$$
$$+\frac{j\eta}{4\pi k}\frac{\partial}{\partial z}\int_{S'}\frac{1}{\rho'}\frac{\partial J_\rho(\rho')}{\partial \rho'}G(z,z')ds' = E_z^{\text{inc}} \qquad (15)$$

The first term in the integral is zero because F_n consists of sine functions. Assuming that

$$E_z^{\text{end}} = \frac{j\eta}{4\pi k}\frac{\partial}{\partial z}\int_{S'}\frac{1}{\rho'}\frac{\partial J_\rho(\rho')}{\partial \rho'}G(z,z')\,ds' \qquad (16)$$

then (15) becomes

$$\frac{j\eta}{4\pi \sin k\Delta}\sum_{n=1}^{N} +I_n\{G(z,z_{n+1})+G(z,z_{n-1}) - 2\cos k\Delta G(z,z_n)\}+E_z^{\text{end}}=E_z^{\text{inc}} \tag{17}$$

Applying a weight function of the same form as (13) and taking the inner product of both sides of (17) and the weight function, we see that the equation becomes

$$\frac{j\eta}{4\pi \sin k\Delta}\sum_{n=1}^{N} I_n\langle G(z,z_{n+1})+G(z,z_{n-1}) - 2\cos k\Delta G(z,z_n), W_m\rangle + \langle E_z^{\text{end}}, W_m\rangle = \langle E_z^{\text{inc}}, W_m\rangle \quad (m=1,2,\cdots\cdots N) \tag{18}$$

where

$$W_m=\begin{cases}\dfrac{\sin k(\Delta-|z-z_m|)}{\sin k\Delta} & z\in(z_{m-1},z_{m+1})\\ 0 & z\notin(z_{m-1},z_{m+1})\end{cases}$$
$$\Delta=|z_{m+1}-z_m| \tag{19}$$

Here z indicates a coordinate taken on the axis. The above results can be expressed in matrix form as

$$[Z][I]=[V] \tag{20}$$

The impressed field E_z^{inc} of the $[V]$ matrix is considered to be excited by a frill of magnetic current (8) across the aperture of the coaxial cable line feeding a monopole as shown in Fig. 3. In other words, assuming the field on the aperture of a coaxial cable line at $z=0$ to be identical to that of a transverse electromagnetic (TEM) mode on the coaxial transmission line, the equivalent frill of magnetic current can be determined. Also, by considering the image of an excitation voltage as V_0, and the inner diameter and outer diameter of the coaxial cable line by a and b, respectively, it follows that

$$E_z^{\text{inc}}=\frac{V_0}{2l_n(b/a)}\left\{\frac{e^{-jk\sqrt{z^2+a^2}}}{\sqrt{z^2+a^2}}-\frac{e^{-jk\sqrt{z^2+b^2}}}{\sqrt{z^2+b^2}}\right\} \tag{21}$$

2.3. Charge Density and Current on End Faces

Now, we proceed with the study of the current over the antenna end surface. We presume that the current is zero at the center to the edge (in the case of a thin cylindrical antenna, the end current is assumed to be zero as shown by the dotted line in Fig. 3). If we express the total charge on the end surface by Q and its radius by a, then the charge density, σ, and the current density, J_ρ, are given by

$$\sigma=\frac{Q}{\pi a^2},\quad J_\rho=\frac{-j\omega Q}{2\pi a^2}\rho \tag{22}$$

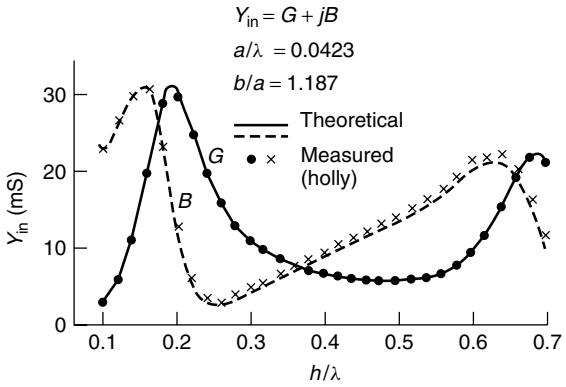
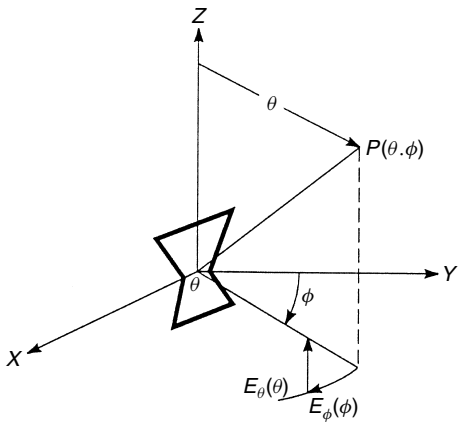


Figure 5. Input admittance of thick cylindrical antenna.



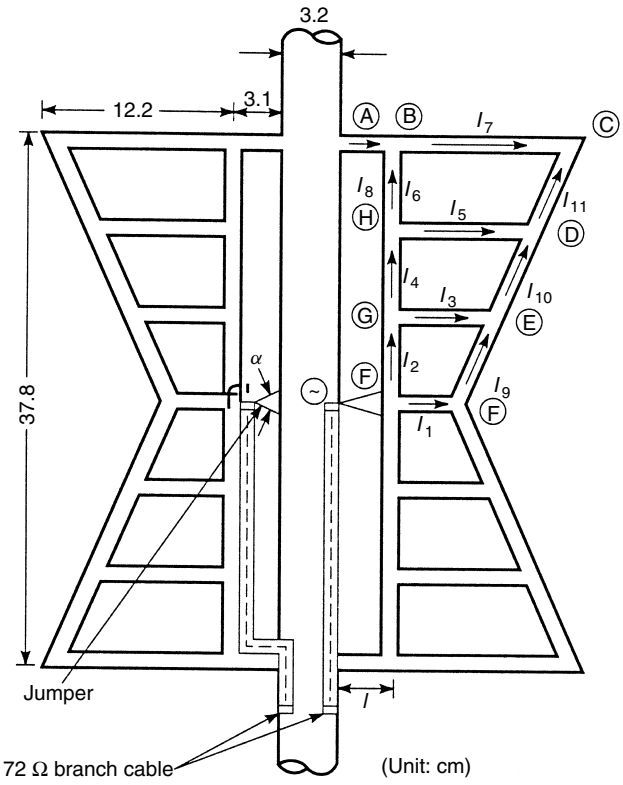


Figure 6. Construction of the model batwing antenna and its coordinate system.

Also, by the current continuity condition on the edge, we have

$$Q = \pm\frac{I_z(z')}{j\omega} \tag{23}$$

Furthermore, the axial component of the field strength on the z axis, produced by sources on the end surfaces, is given by

$$E_z = \pm\frac{j\eta I_z(z')}{2\pi ka^2}(z-z')\left\{\frac{e^{-jk\sqrt{(z-z')^2+a^2}}}{\sqrt{(z-z')^2+a^2}} - \frac{e^{-jk|z-z'|}}{|z-z'|}\right\} \tag{24}$$

We impose the boundary condition that the total axial field is zero in the range of $-h\langle z\langle h$, not including $z = \pm h$. Now, if we select the end face weight functions to be the same as the expansion functions, it follows that the same weight function will also be applied to the end faces $z = \pm h$, as shown in Fig. 4.

Referring to the difference between the analysis of Taylor and Wilton and that presented here, the former uses the point-matching method and is applied only to the body of revolution. On the other hand, the latter employs the Galerkin method and is applied to antennas that are asymmetric and that also contain discontinuities in the conductors. By using the moment method and taking the end surfaces into consideration, a relatively thick antenna can be treated.

For example, the calculated input admittance for $a = 0.0423\lambda$ and $b/a = 1.187$ is given in Fig. 5, as a function of h. For comparison, the values measured by Holly [9] are also shown. In the calculation, the number of subsections, N, was chosen as 50–60 per wavelength. The theoretical values agree well with the measured values. Therefore, it is concluded that the analytic technique is adequate for thick cylindrical antennas.

3. THE BATWING ANTENNA ELEMENT

The antenna is installed around a support mast, as shown in Fig. 6, and fed from points f and f', through a jumper from a branch cable with a characteristic impedance of 72 Ω. The conducting support mast is idealized by an infinite, thin mast. The batwing antenna element is divided into 397 segments for the original type, with triangular functions as the weighting and expansion functions, and the analysis of the batwing antenna elements is carried out using the Galerkin method. The batwing antenna is fed with unit voltage. The currents flowing in each antenna conductor are calculated over a frequency range of 300–700 MHz.

Figure 7a,b illustrates these current distributions $I_i (i = 1 \sim 12)$ on the conductors at frequencies of 300, 500, and 700 MHz. Since the distribution of currents along each conductor is calculated, this allows calculation of the radiation characteristics. Figure 8 illustrates the amplitude and phase characteristics of radiation patterns in the horizontal and vertical planes. It is seen from this figure that the theoretical values agree well with the measurements.

Figure 9 illustrates the theoretical and measured input impedance of a batwing antenna mounted on an aluminum plate, $3m \times 3m$. Both curves coincide closely with each other, with the input impedance having a value close to 72 Ω, which is the proper match to the characteristic impedance of the branch cable. Vernier impedance matching is carried out in practice by connecting a metal jumper between the end of the branch cable and the feed point of the antenna element or the support mast. The feed strap's length, width, or form is varied to derive VSWR values below 1.10. In the case of the distance between the support mast and antenna elements being varied, Fig. 10 shows that the real part of the input impedance changes as the distance l between the support mast and antenna elements is varied.

Figure 11 shows that the real part of the input impedance is not largely affected by the angle of the jumper and that the reactance is likely to shift as a whole to the left side of the Smith chart because of the capacitance.

Figures 11 and 12 show that matching can be obtained by means of adjusting appropriately the distance between the support mast and antenna elements l, and the angle of the jumper in order to reduce the input VSWR.

The power gain of the antenna at 500 MHz is calculated to be 3.3 dB. Figure 13 shows the gain of the antenna in the $\phi = 0^\circ$ direction as a function of the frequency, referenced to a half-wavelength dipole. Figure 14 illustrates three-dimensional amplitude characteristics of radiation patterns in the horizontal and vertical planes at each frequency. For example, Fig. 15 indicates the polarization pattern of the calculated performance characteristics.

4. MUTUAL RADIATION IMPEDANCE CHARACTERISTICS

Two units of the batwing antenna illustrated in Fig. 6 were stacked one above the other in the same plane to constitute a broadside array with a center-to-center separation distance d.

Measurements were made as follows:

1. Frequency was fixed at 500 MHz, with distance d as the parameter.
2. Distance was fixed at 60 cm (full wavelength), with frequency as the parameter.

In both 1 and 2, measurements of mutual impedance were made under both open- and short-circuit conditions. Figure 16 shows a schematic of the equipment used in the measurement setup. The data obtained are shown in Fig. 17a,b. The theoretical values agree well with the measurements as shown in Fig. 17a,b. The mutual impedance was below a few ohms when the distance between elements exceeded 0.8 λ.

5. THEORETICAL ANALYSIS OF METAL-BAR-SUPPORTED WIDEBAND FULL-WAVE DIPOLE ANTENNAS WITH A REFLECTOR PLATE

Wideband full-wave dipole antennas with a reflector plate supported by metal bars were invented in Germany. The construction is shown in Fig. 18. A full-wave dipole antenna is located in front of a reflector, and supported

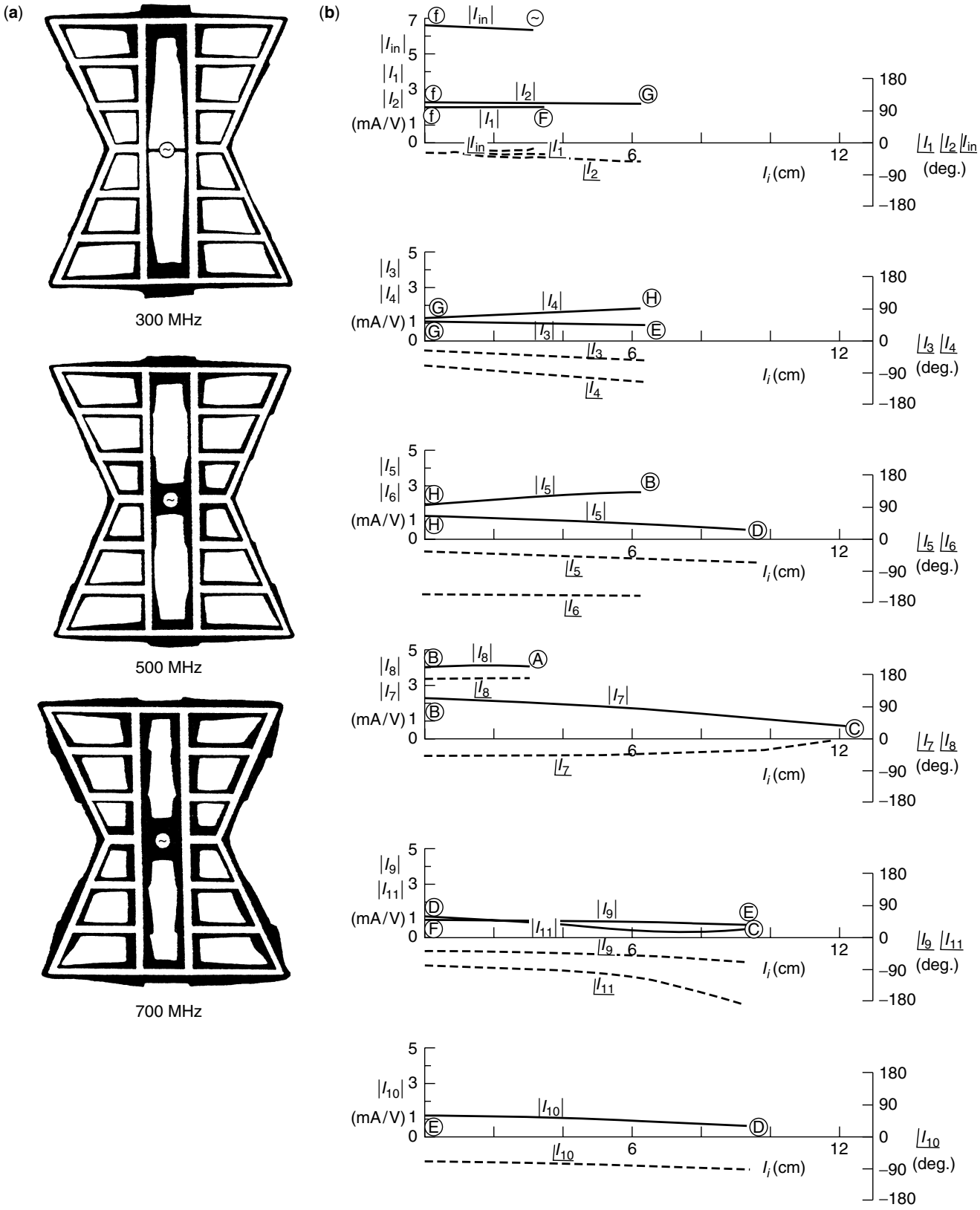


Figure 7. (**a**) Amplitude characteristics of current distribution for frequency range from 300, 500, 700 MHz of shaded areas; (**b**) current distribution at 500 MHz.

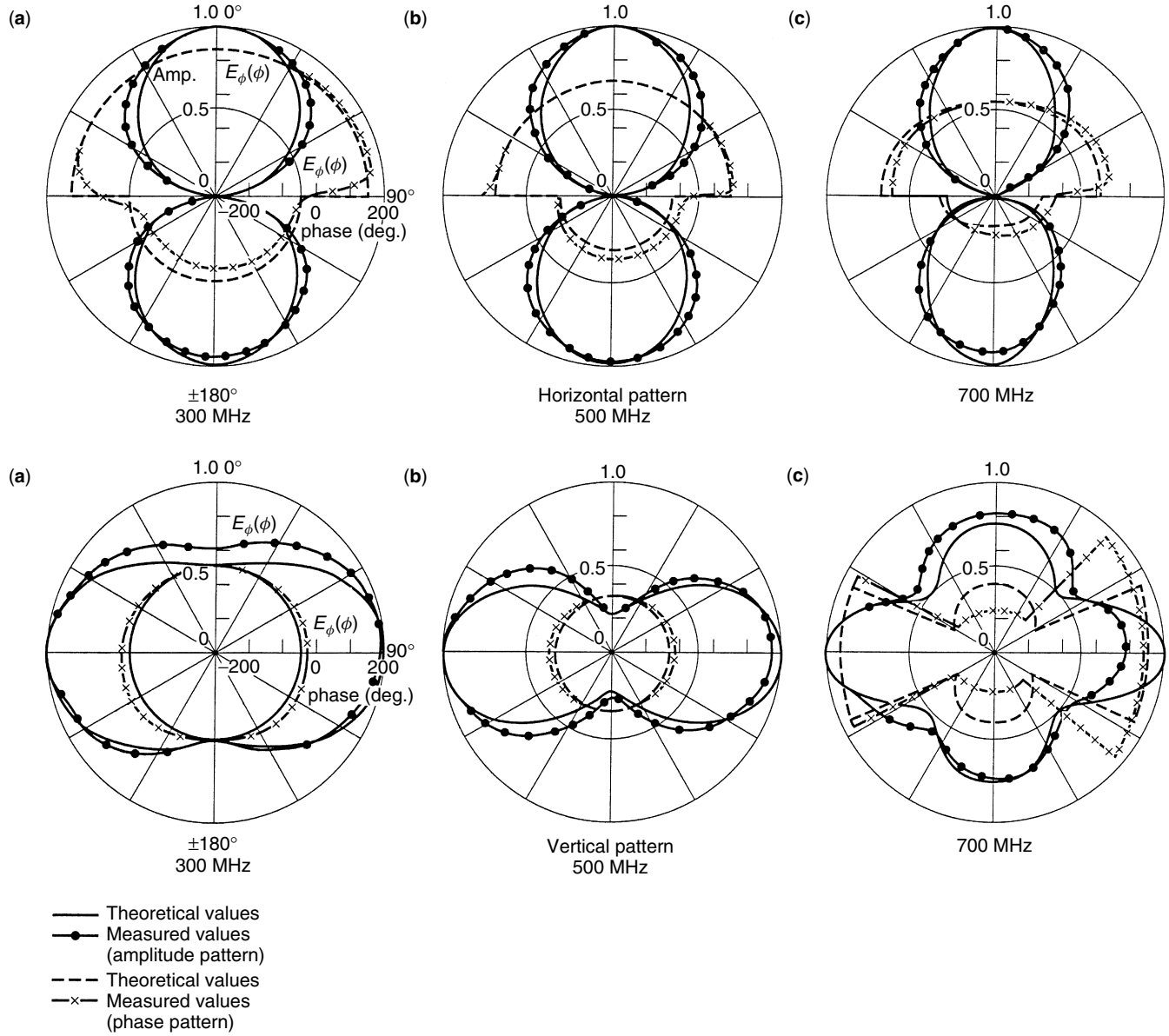


Figure 8. Amplitude and phase characteristics of radiation patterns at 300, 500, 700 MHz (with support mast).

directly by a metal bar attached to a reflector. This antenna was also analyzed by the moment method described previously.

Because the supporting bar (see Fig. 18) is metallic, leakage currents may cause degradation of the radiation characteristics. To calculate these effects, the radial component of field E_ρ must be taken into consideration. In other words, E_ρ is needed for the calculation of $Z_{m,n}$, as defined by inner products of the expansion functions on the supporting bar and weighting functions on the antenna element or on the parallel conductors. We assume the supporting bar to be separated from the feed point by a distance l_1. Also, the radius of the supporting bar is fixed at a fourth of the radius of the antenna element (i.e., at $\lambda_0/100$), and then is varied to be $0.2\lambda_0$, $0.25\lambda_0$, and $0.3\lambda_0$. Figures 19–22 indicate various calculated performance

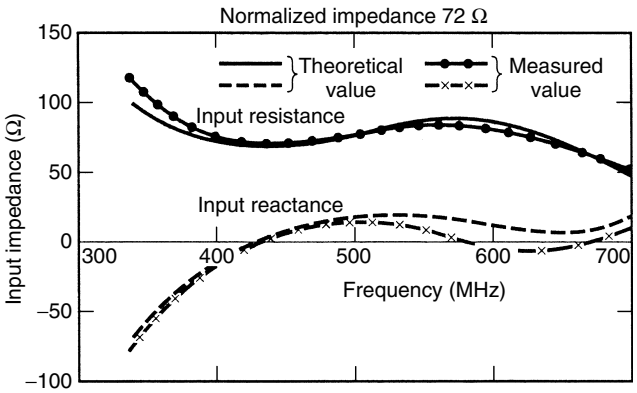


Figure 9. Theoretical and measured values of the input impedance as function of frequency (mast is infinite thin, $\alpha = 0°$).

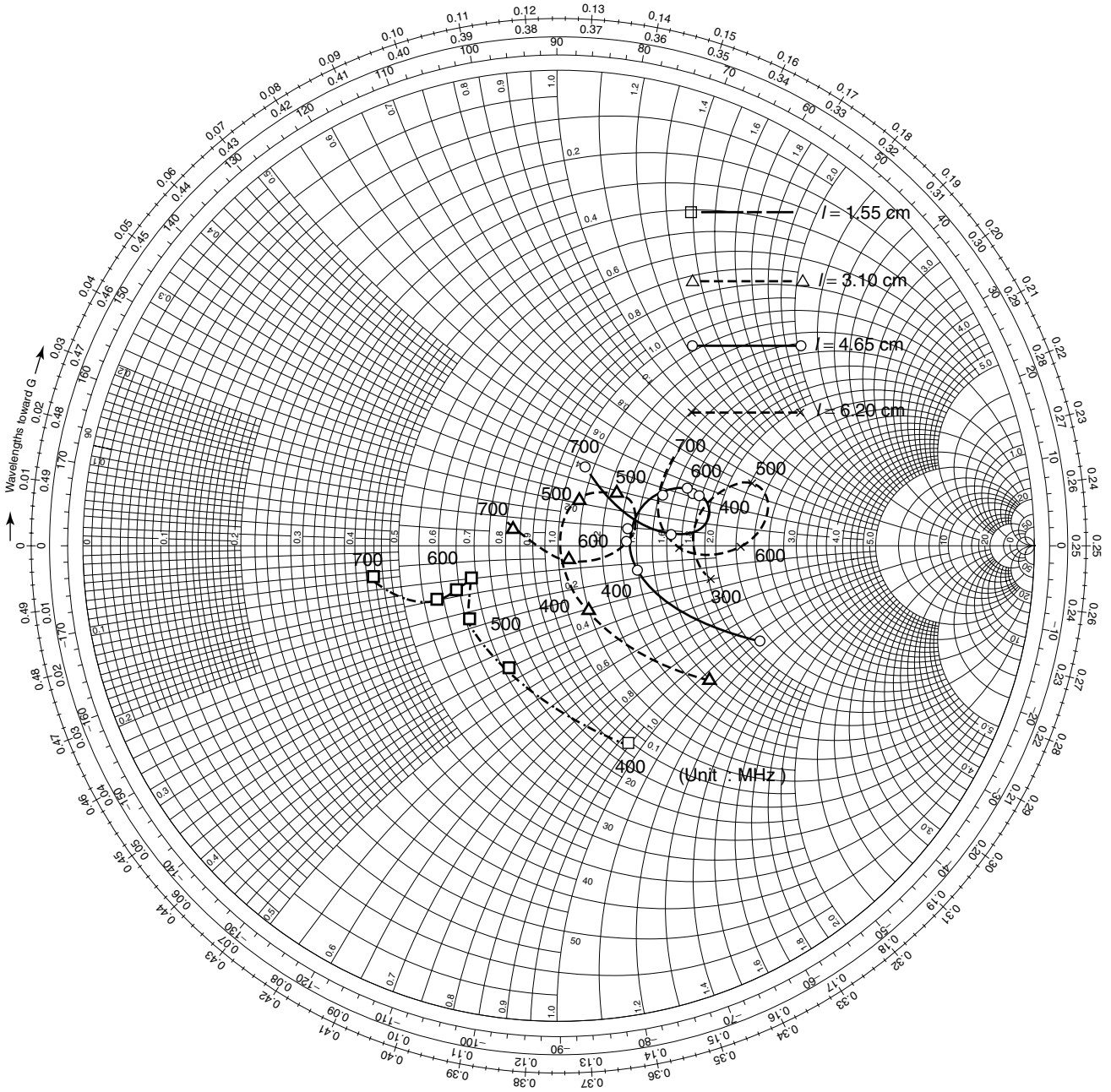


Figure 10. Changing of distance between the support mast (infinite thin, $\alpha = 0^\circ$) antenna element.

characteristics. Note that the leakage current due to the supporting bar is minimized for $f/f_0 = 0.7$, and that this current is substantial at other frequencies. The current distribution is shown only for the case of $l_1 = 0.25\lambda_0$.

6. CHARACTERISTICS OF 2L TWIN-LOOP ANTENNAS WITH INFINITE REFLECTOR

As shown in Fig. 23, a twin-loop antenna has the loops connected by a parallel line: The 2L, 4L, and 6L types are used, according to the number of loops. For actual use, a reactive load is provided by the trap at the top end, which also serves as the antenna support. The dimensions used for this article are as follows: center frequency $f_0 =$ 750 MHz (wavelength $\lambda_0 = 40$ cm), length of the parallel line part $2l_1 = \lambda_0/2$ ($l_1 = 10$ cm), $l_2 = \lambda_0/2 (l_2 = 20$ cm), interval of the parallel line part, $d = \lambda_0/20$ ($d = 2$ cm), loop radius $b = \lambda_0/2\pi (b = 6.366$ cm), distance from the reflector to the antenna $l_3 = \lambda_0/4 (l_3 = 10$ cm), conductor diameter $\phi = 10$ mm, and top end trap $l_t = 0$ to $\lambda_0/4$, changed in intervals of $\lambda_0/16$. 2L twin-loop antennas were arranged in front of an *infinite* reflector, and calculations were executed in regard to the frequency characteristics of the trap length.

The radiation pattern of the 2L-type antenna is shown in Fig. 24. Up to $l_t = \lambda_0/8$, the main beam gradually

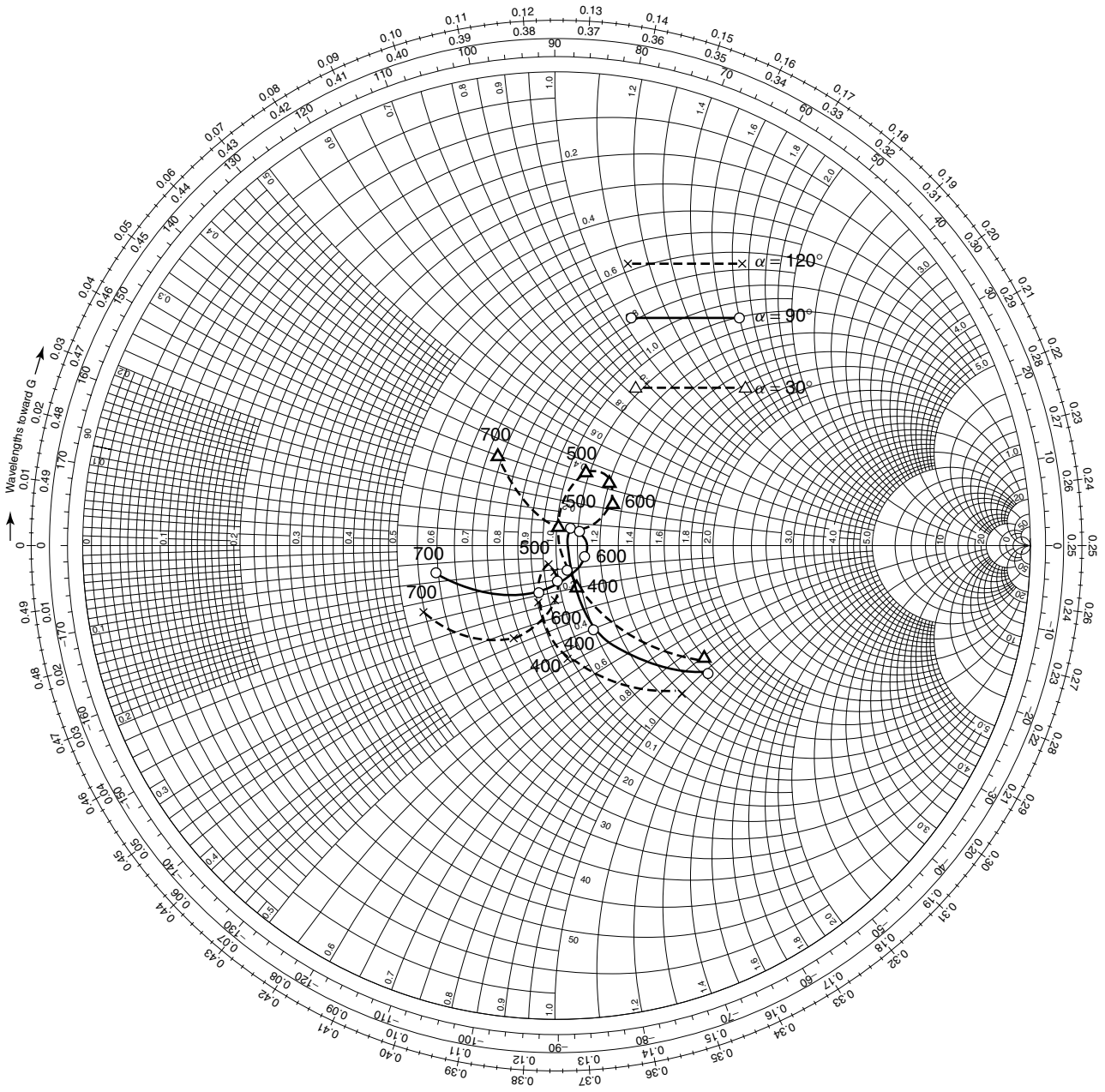


Figure 11. Changing of the shape of the jumper ($\ell = 3.10$ cm).

becomes sharper with increasing frequency, and it can be seen that the sidelobes increase. When l_t increases in this way to $\lambda_0/8$ and $\lambda_0/4$, the directivity becomes disturbed.

Figure 24 shows $l_1 = 0.15\lambda_0$ and $l_1 = 0.25\lambda_0$ characteristics of the radiation pattern in a polar display. The antenna gain for both lengths ($l_1 = 0.15\lambda_0$ and $l_1 = 0.25\lambda_0$) shows a small change of approximately $9.5 - 8.5$ dB. The input impedance has a value very close to 50 Ω, essentially the same as the characteristic impedance of the feed cable. As for the input impedance, the 2L twin-loop antenna has reactance nearest zero (for the case where $l_1 = 0.15\lambda_0$); that is, the VSWR is nearly equal to unity.

In the calculation above, the reflector was considered to be an infinite reflector, and the effect of the reflector on the antenna elements was treated by the image method. In the case of practical antennas, however, it is the usual practice to make the reflector finite, or consisting of several parallel conductors. Therefore, a calculation was executed for a reflector in which 21 linear conductors replaced the infinite reflector, as shown in Fig. 25.

The results are shown along with those for the infinite-reflector case. On the basis of these results, it was concluded that no significant difference was observed in input impedance and gain between the infinite reflector case and the case where the reflector consisted of parallel conductors.

The wire-screen-type reflector plate had a height of $3\lambda_0$(120 cm), a width of λ_0(40 cm), and a wire interval of

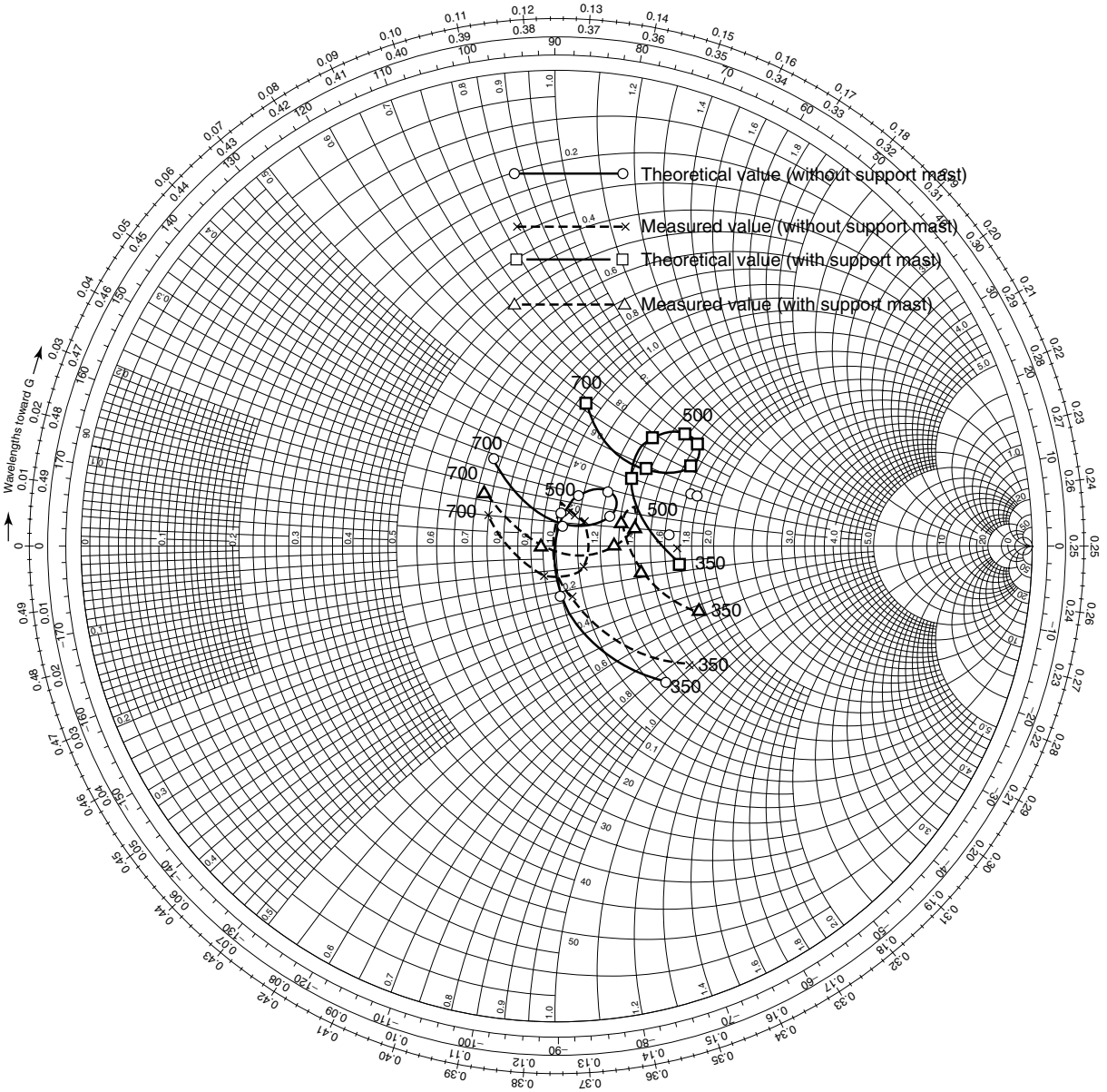


Figure 12. Theoretical and measured values of the input impedance of batwing antenna with support $\ell = 3.10$ cm $\alpha = 0^\circ$).

$0.15\lambda_0$ (6 cm). The radiation pattern is shown in Fig. 26. With regard to the pattern in the horizontal plane, no difference was found in comparison with an infinite reflector, but a backlobe of approximately -16 dB exists to the rear of the reflector. The same figure also shows the phase characteristics. With regard to the pattern in the vertical plane, the phase shows a large change where the pattern shows a cut.

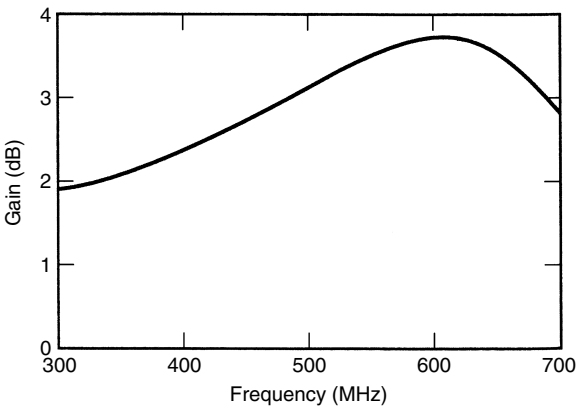


Figure 13. Batwing antenna gain with $\lambda/2$ dipole.

7. THE DIGITAL TERRESTRIAL BROADCASTING ANTENNAS

7.1. Unbalance-Fed Modified Batwing Antenna

The configuration of the unbalance-fed modified batwing antenna (UMBA) is shown in Fig. 27. The reflector is

assumed an infinite perfect conducting plane for the calculation.

Parallel lines are connected perpendicularly to the center of batwing radiators. One side of the parallel line is connected perpendicularly to the reflector; the other side is bent above the reflector and connected to the other parallel line. UMBA is excited by the unbalanced generator, which is assumed to be a coaxial cable, between the center of the line and the reflector. The center frequency is 500 MHz ($\lambda_0 = 600$ mm) and the conductor diameter is $0.013\lambda_0$. The center frame of the batwing radiator, that is the length of LW0 $= 0.25\lambda_0$ and the spacing of d, acts as a quarter wavelength stab. When the total length of the radiating elements LW1, LW2, and LW3 in Fig. 27 is about $0.5\lambda_0$, a broad-bandwidth input impedance is obtained. Therefore, LW1 $= 0.167\lambda_0$, LW2 $= 0.063\lambda_0$, and LW3 $= 0.271\lambda_0$ are chosen. Two batwing radiators are set on the height of $H = 0.25\lambda_0$ above the reflector. The spacing of d and Hf are adjusted in the range from $0.016\lambda_0$ to $0.025\lambda_0$ to obtain the input impedance of 50 Ω. The element spacing of D is $0.63\lambda_0$. NEC-Win Pro is used in the calculation.

In Fig. 28 the measured and the calculated VSWR (50 Ω) for a two-element UMBA are shown. The bandwidth ratio of 30% (VSWR≤1.15) is obtained for both results. In Fig. 29 the power gain is shown. The gain of about 11 dBi and the broadband property are obtained, and measured results agree well with calculated results. In Fig. 30, the normalized radiation patterns in the vertical plane (Z–Y plane: $E\phi$) and the horizontal plane (Z–X plane: E_θ) at frequencies of 450, 500, and 550 MHz are shown. Good symmetric patterns on the horizontal plane are obtained. The half-power beamwidth for five frequencies is shown in Table 1.

7.2. Parallel Coupling UMBA

The configuration of the parallel coupling UMBA (PCUMBA) is shown in Fig. 31. Four batwing elements are coupled without a space. The outside elements (1 and

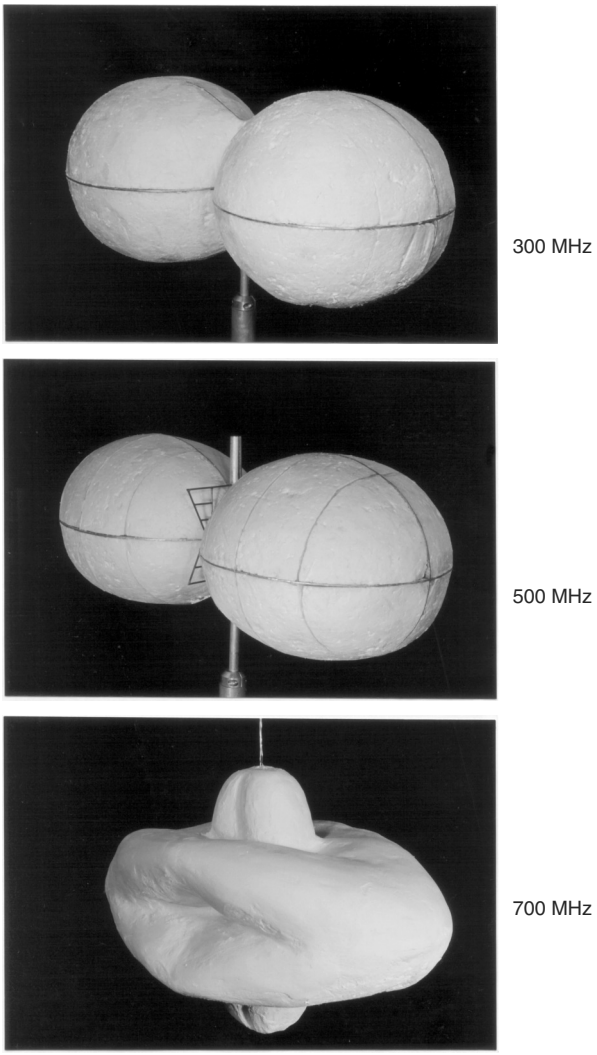


Figure 14. Three-dimensional amplitude characteristics of radiation patterns at 300, 500, 700 MHz.

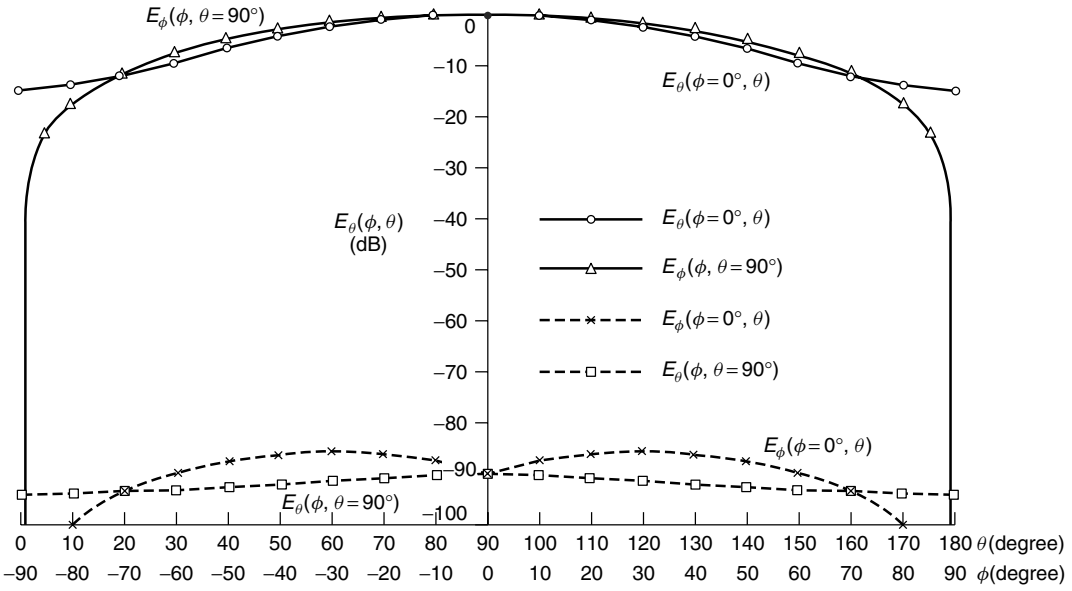


Figure 15. Polarization pattern of the model batwing antenna at 500 MHz (without support mast).

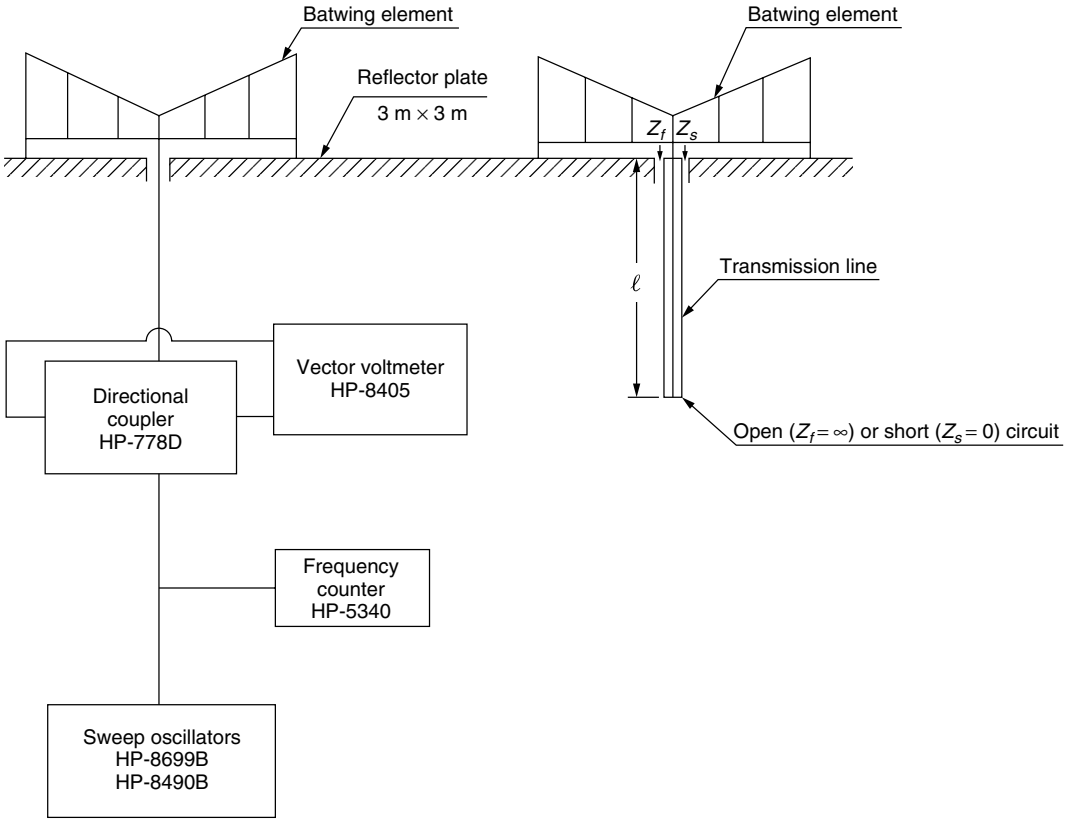


Figure 16. Block diagram of experimental measurement setup.

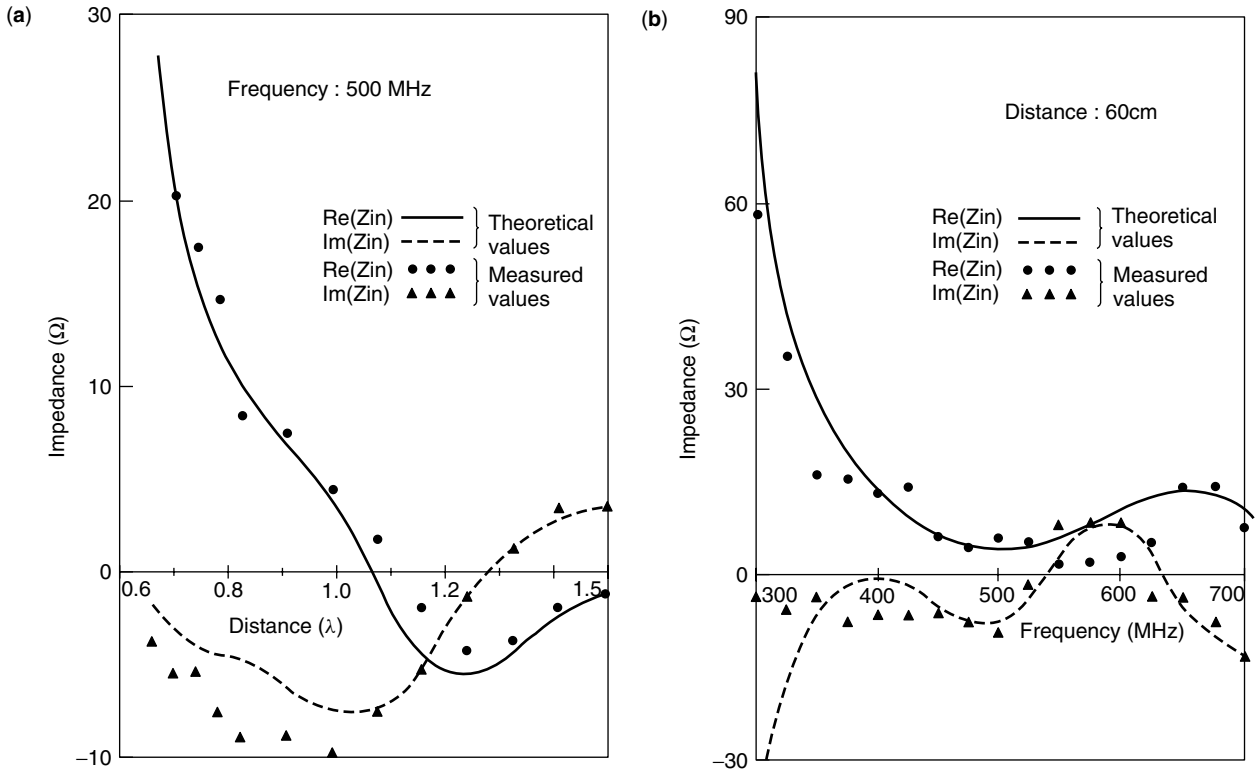


Figure 17. (**a**) Mutual radiation impedance at 500 MHz; (**b**) mutual radiation impedance changing of distance $d = 1\lambda$.

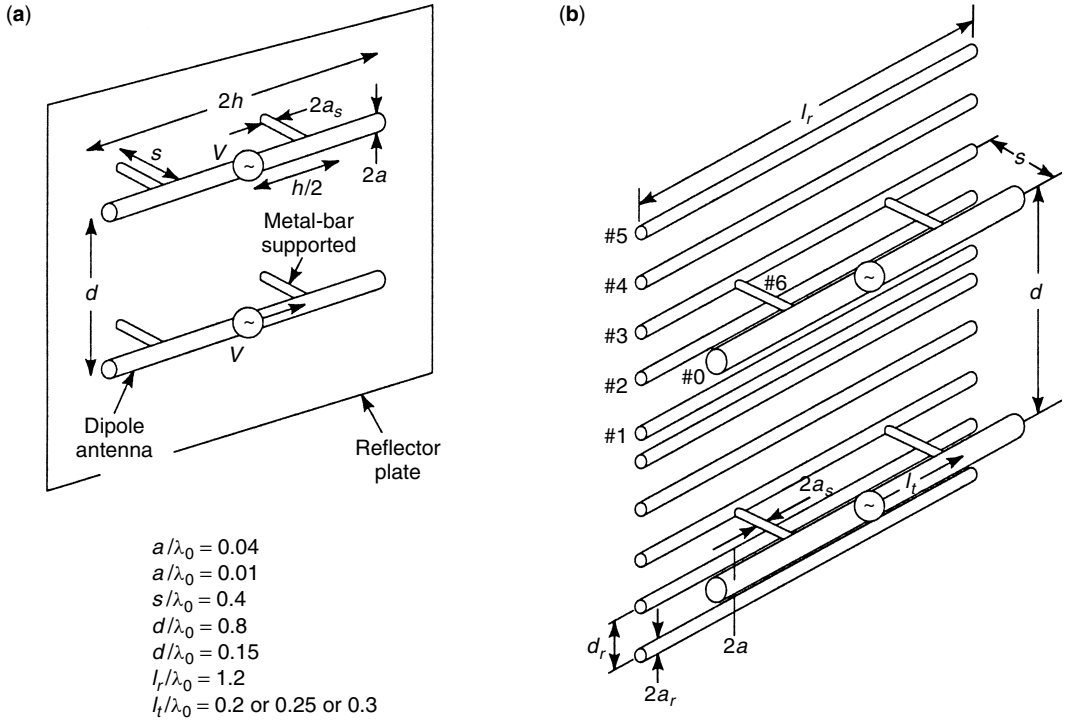


Figure 18. Metal-bar-supported wideband full-wave dipole antennas with a reflector plate.

4) are cross-connected with the inside elements (2 and 3) by the strait wire. Then the outside element is fed by in-phase against the inside one. Because the input impedance of UMBA is about 50 Ω, an impedance transformer is required for PCUMBA matched with the coaxial cable. The length from the center of PCUMBA to the center of 2(3) is a quarter-wavelength. Therefore, by changing the height and the radius of the center feed line, the impedance is transformed as a quarter-wave transformer. In Fig. 32 the calculated and measured VSWR normalized by 50 Ω for PCUMBA are shown. For the calculation, the bandwidth ratio (less than 1.15) is 30%. For the measurement, after adjusting stabs, the bandwidth ratio (less than 1.05) is 30%. Thus a broadband antenna is realized with the simple feed structure. In Fig. 33 the calculated power gain is shown. The gain of about 14 dBi equivalent for the five-element dipole array is obtained at the center frequency. In Fig. 34 radiation patterns with the infinite ground plane are shown. The half-power beamwidth for five frequencies is shown in Table 2.

8. CONCLUSION

Previous researchers [1,2] have analyzed batwing antennas by approximating the current distribution as a sinusoidal distribution. Wideband characteristics are not obtained with a sinusoidal current distribution. In this article, various types of modified batwing antennas, as the central form of the superturnstile antenna system, were analyzed theoretically with the aid of the moment method. The results were compared to measurements in order to examine the performance of the antenna elements in detail.

It is also evident from this research that the shape of the jumper has a remarkable effect on the reactance of the input impedance, and that the distance between the support mast and the antenna element also markedly influences the resistance of this impedance. Thus, a satisfactory explanation is given with regard to the matching conditions. As a result, it was found that the calculated and the measured values agree well, and satisfactory wideband characteristics are obtained. Also, it was found that a broadside array spaced over a full wavelength is to be of the order of a few ohms for the designed frequency at 500 MHz, and there would be small mutual coupling effect. It has been made clear that a superturnstile antenna made by arranging these antennas in multiple bays has an arrangement with small mutual coupling effect.

Next, an analytic method and calculated results for the performance characteristics of a thick cylindrical antenna were presented. The analysis used the moment method and took the end face currents into account. The calculated results were compared with measured values, demonstrating the accuracy of the analytic method.

Using this method, a full-wave dipole antenna with a reflector supported by a metal bar was analyzed. The input impedance was measured for particular cases, thus obtaining the antenna dimensions for which the antenna input impedance permits broadband operation. In conclusion, wideband characteristics are not obtained with a one-bay antenna. The wideband characteristic is

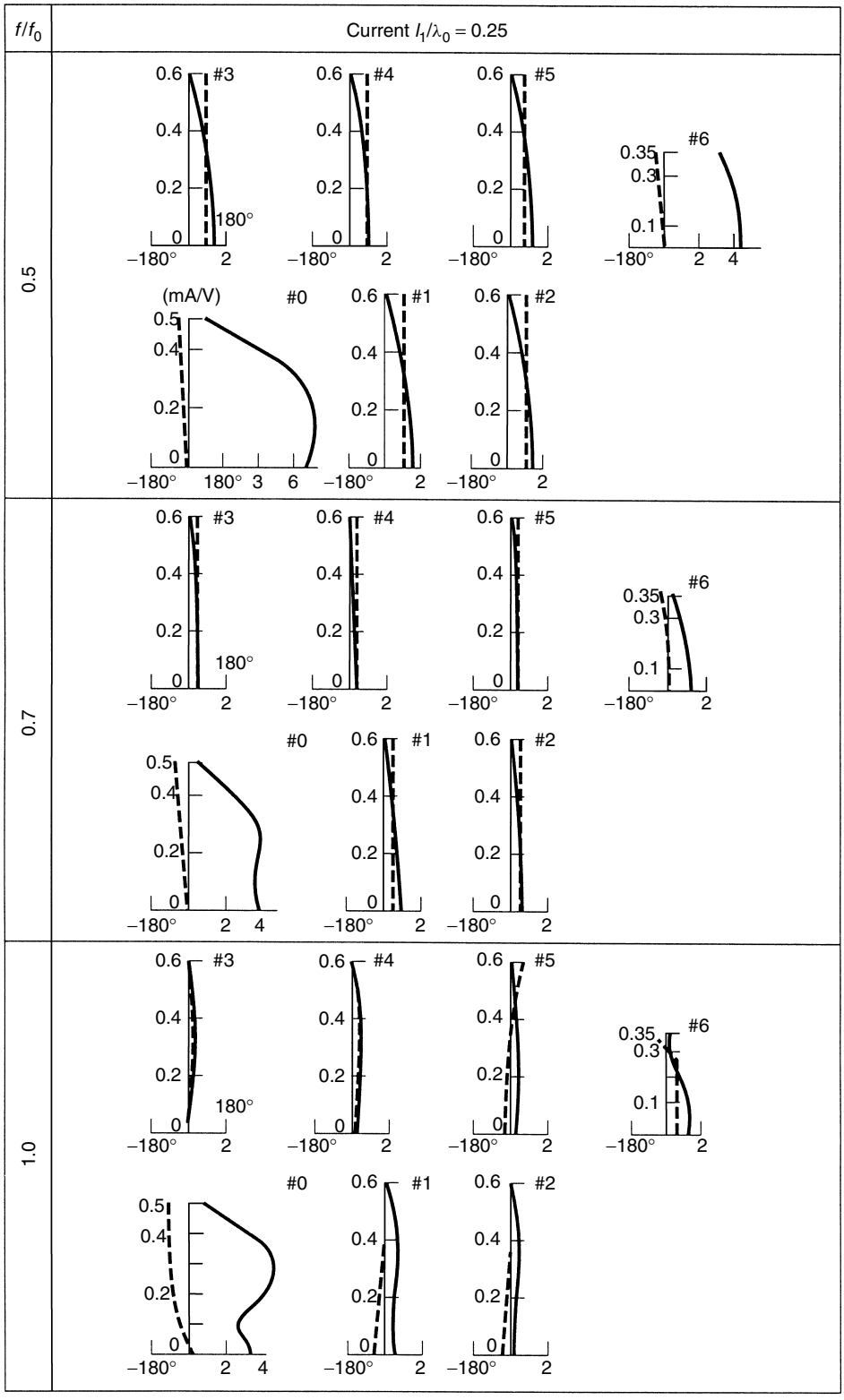


Figure 19. Current distribution of metal-bar-supported full-wave dipole antennas (two-bay) with a wire-screen-type reflector plate.

obtained by means of the mutual impedance of the two-bay arrangement. In the frequency region of $f/f_0 = 0.7$, the resistance of the input impedance is considered to be constant. In this case, the leakage current to the support bar is small. With regard to the radiation pattern, it was seen that a degradation of characteristics was caused by the metal-support bar.

It is noted that the present method should be similarly useful for analyzing antennas of other forms where the end face effect is not negligible.

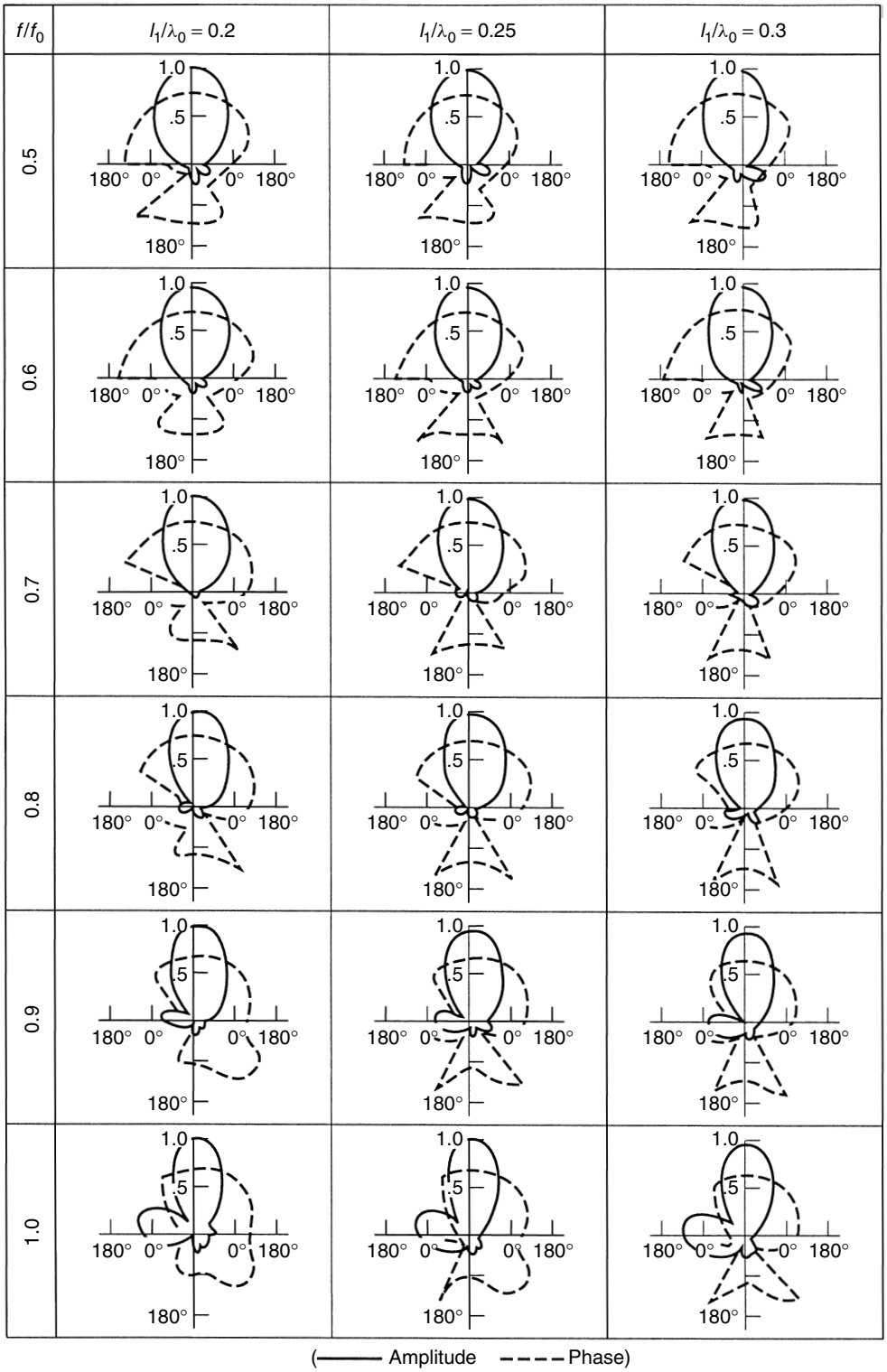


Figure 20. Radiation pattern of metal-bar-supported full-wave dipole antennas (two-bay) with a wire-screen-type reflector plate.

Next, the twin-loop antennas were considered for use as wideband antennas. The analysis results for the 2L type showed that the change in the characteristics with a change in frequency becomes more severe with increasing trap length l_t, and the bandwidth becomes small, while a short trap length l_t shows a small change and a tendency for the bandwidth to become wide. For $l_t = 0$, a wide bandwidth for pattern and gain was obtained for the 2L type. The input impedance has a value very close to 50 Ω, essentially the same as the characteristic impedance of

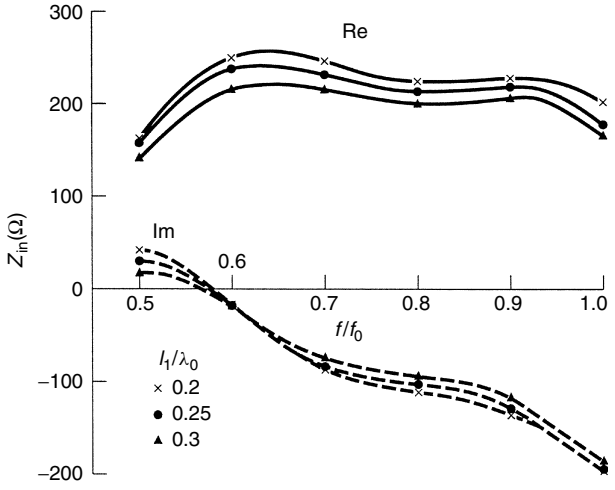


Figure 21. Input impedance characteristics of metal-bar-supported full-wave dipole antennas (two-bay) with a wire-screen-type reflector plate.

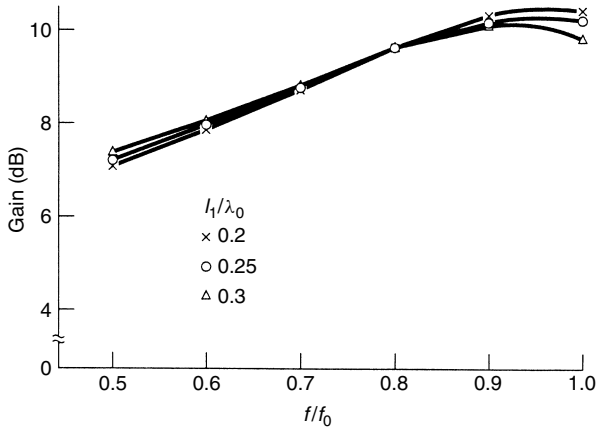


Figure 22. Gain of metal-bar-supported full-wave dipole antennas (two-bay) with a wire-screen-type reflector plate.

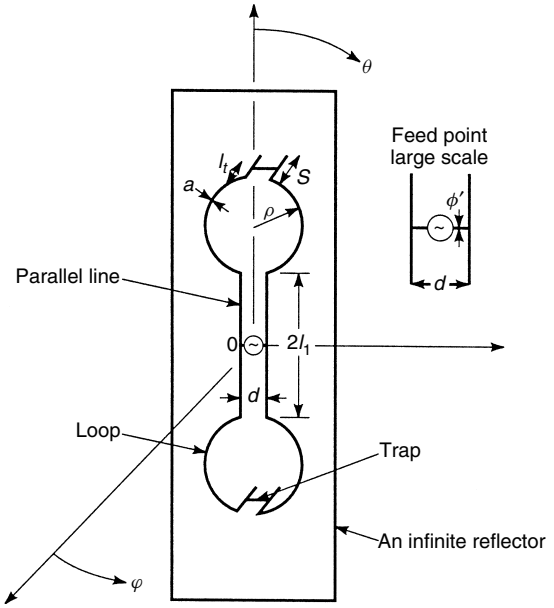


Figure 23. Structure of 2L-type twin-loop antenna and its coordinate system for analysis.

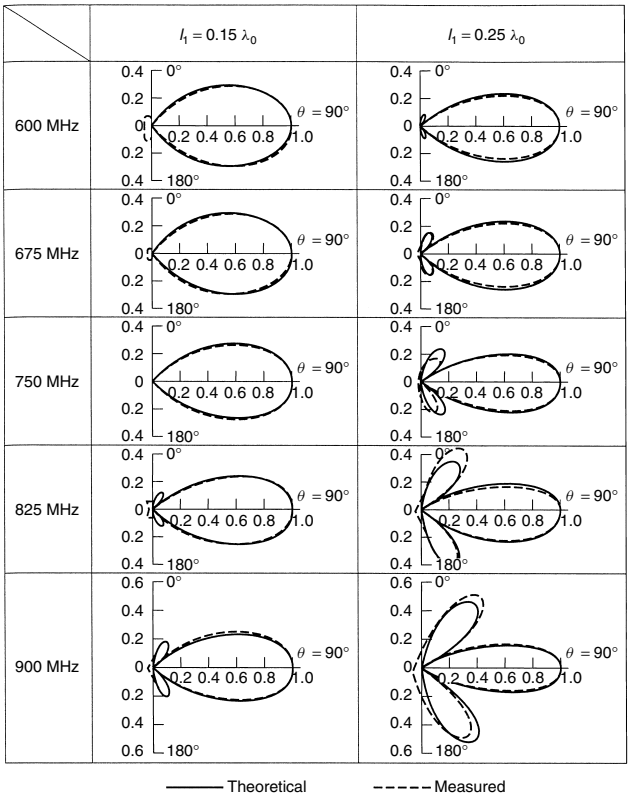


Figure 24. Vertical radiation pattern of 2L-type twin-loop antenna.

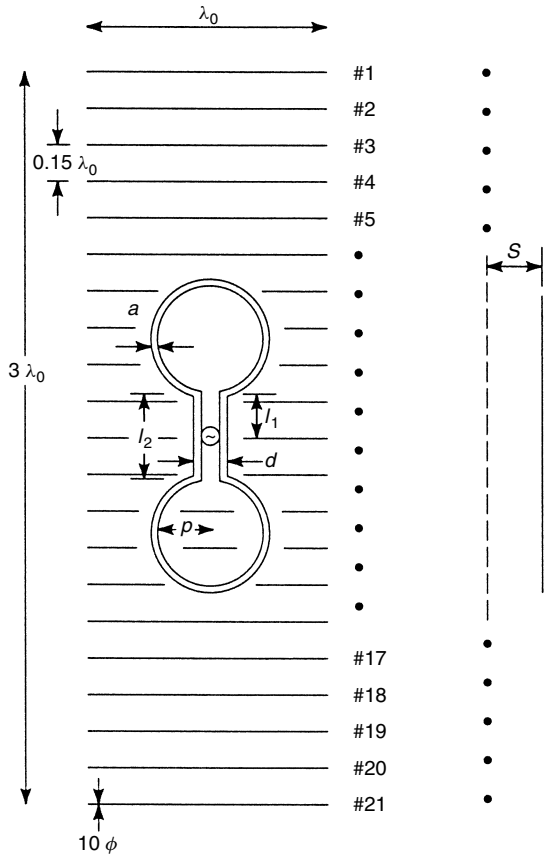


Figure 25. Structure of 2L-type twin-loop antenna with a wire-screen-type reflector plate.

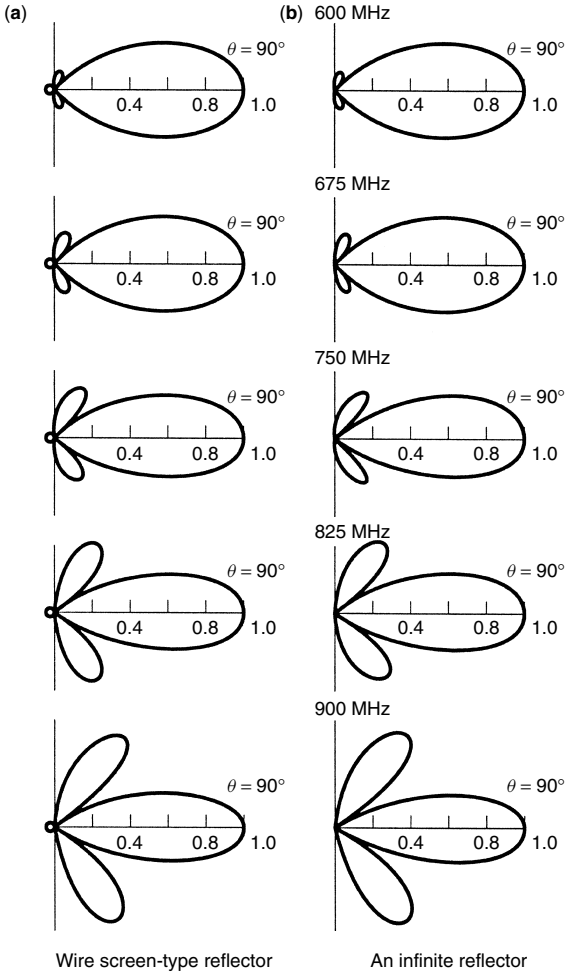


Figure 26. Comparison between characteristics of 2L-type twin-loop antenna with a wire-screen-type reflector and with an infinite reflector.

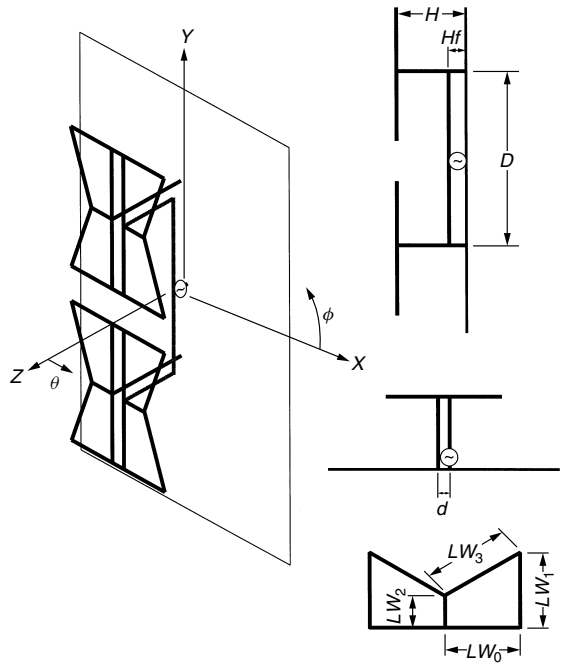


Figure 27. Configuration of UMBA.

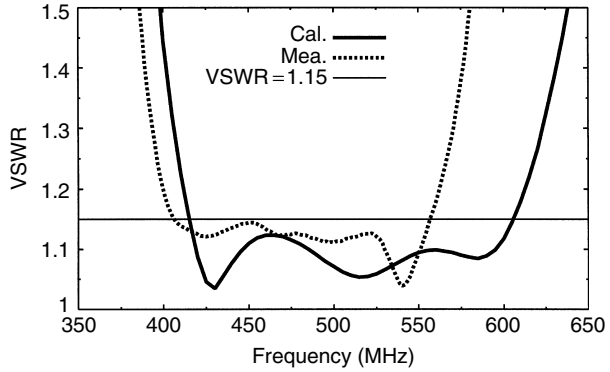


Figure 28. VSWR characteristics of UMBA (normalized impedance $Z_0 = 50\Omega$).

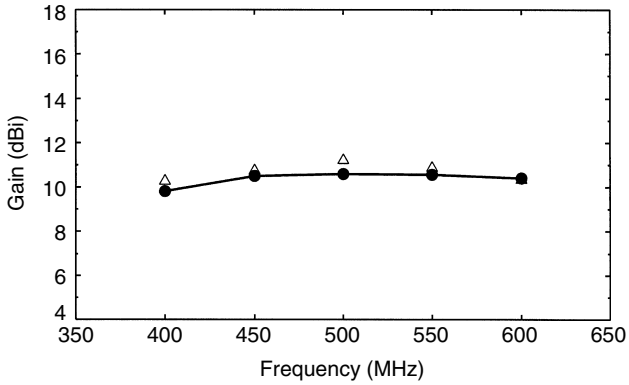


Figure 29. Power gain of UMBA.

the feed cable over a very wide frequency range. Thus, a satisfactory explanation was given with regard to the matching conditions.

The unbalance fed modified batwing antenna and the parallel coupling UMBA were also investigated. Broadband properties the same as the balance-fed type is obtained and the input impedance is matched ell with a coaxial cable of 50 Ω. The unbalanced current are decreased, and good symmetric patterns are obtained by the contribution of the quarter-wavelength stab on the batwing element. A power gain for the parallel coupling four-element PCUMBA of about 14 dBi is obtained. The unbalance-fed shape and the parallel coupling shape have the effect that the digital terrestrial broadcasting antenna is simpler, smaller, and of wider bandwidth than another transmitting antennas for UHF-TV broadcasting.

It has been reported here that a rigorous theoretical analysis has been achieved about 45 years after the invention of these VHF-UHF antennas.

BIOGRAPHIES

Haruo Kawakami received the B.E. degree from the Department of Electrical Engineering, Meiji University, Tokyo, Japan, in 1962, and the Ph.D. degree from Tohoku University, Sendai, Japan, in 1983.

In 1962, he joined YAGI Antenna Co., Ltd. In 1964, he become a research associate of at the Department

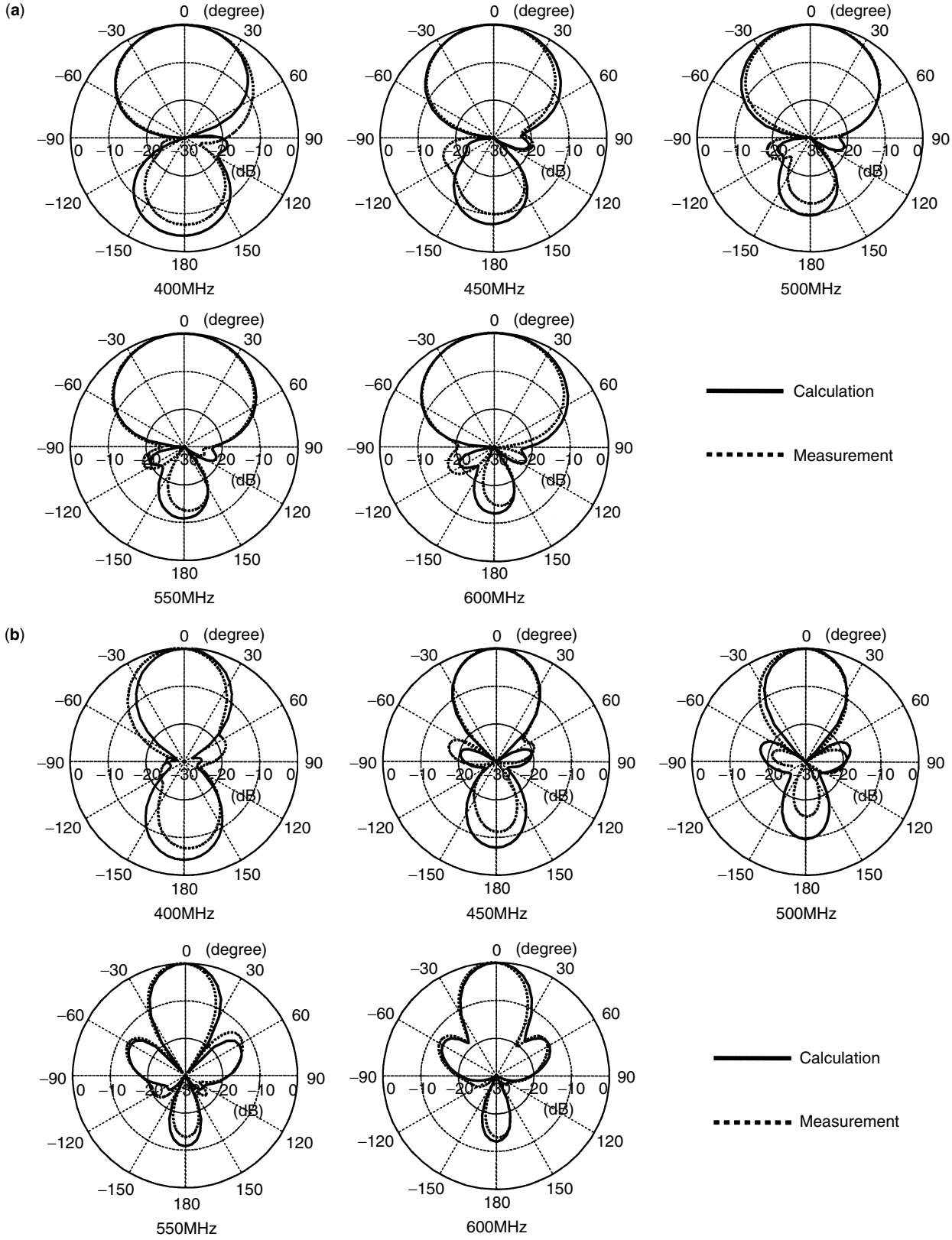


Figure 30. Radiation pattern of UMBA: (**a**) horizontal plane ($E\theta$); (**b**) vertical plane ($E\phi$).

Table 1. Half-Power Beamwidth of UMBA

	400 MHz	450 MHz	500 MHz	550 MHz	600 MHz
Horizontal plane ($E\theta$)	71°	64°	69°	73°	73°
Vertical plane ($E\phi$)	45°	39°	40°	32°	32°

Table 2. Half-Power Beamwidth of PCUMBA

	400 MHz	450 MHz	500 MHz	550 MHz	600 MHz
Vertical plane ($E\phi$)	34°	30°	24°	22°	22°

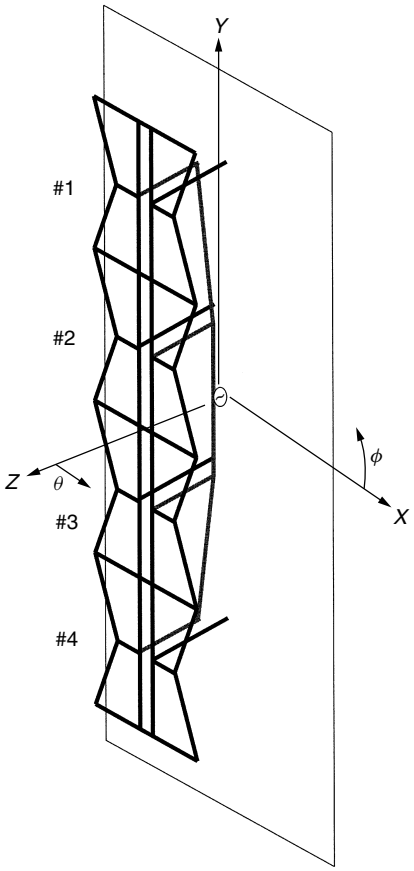


Figure 31. Configuration of PCUMBA.

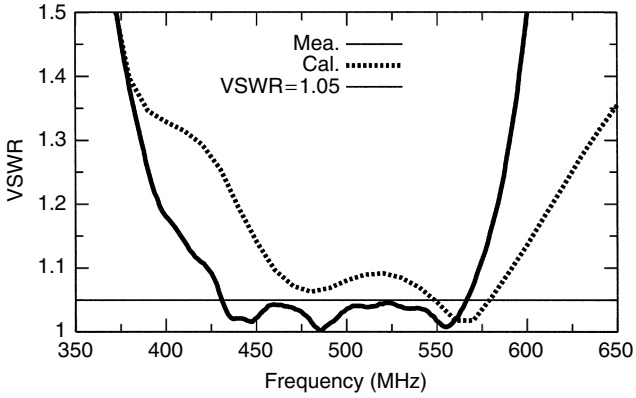


Figure 32. VSWR characteristics of PCUMBA (normalized impedance $Z_0 = 50\Omega$).

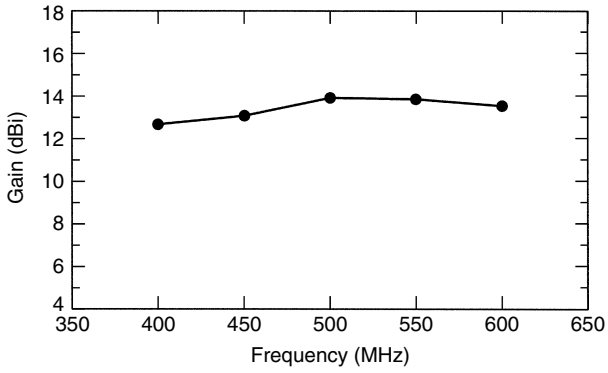


Figure 33. Power gain of PCUMBA.

of Electric and Electronic Engineering, Sophia University, and was promoted to lecturer and associate professor there in 1985 and 1992, respectively. He is currently a director of Antenna Giken Co., Ltd. In 1989, he was a part-time lecturer at the Department of Electronics Engineering, Tokyo Metropolitan College of Aeronautical Engineering. In 1998, he was a visiting professor at Utsunomiya University.

Dr. Kawakami is the author of *Antenna Theory and Application*, *New Electrical Circuit Practice*, and *New Alternate Current Circuit Practice*, Mimatsu Data, Tokyo, 1991, Kogaku-Tosho, Tokyo, 1990 and 1992, respectively.

He has been engaged in research and development of mobile systems, automobile-borne antennas, and electromagnetic compatibility. Dr. Kawakami received a Best Defense Technology Paper Award in 1987.

He is a member of the Applied Computational Electromagnetic Society, the Institute of Electronics, Information and Communication Engineers of Japan, and the Institute of Image Information and Television Engineers of Japan. Kawakami's name has been listed in Marquis "Who's who in the World."

Yasushi Ojiro received the B.E. degree from the College of Engineering Sciences, Tsukuba University, Ibaraki, Japan, in 1993, and the Ph.D. degree from Tsukuba University in 1998. In 1998, he jointed Antenna Giken Co., Ltd. He has been engaged in research and development of HF antenna, TV broadcasting antenna, and adaptive array. He is a member of the Institute of Electonics, Information and Communication Engineers of Japan.

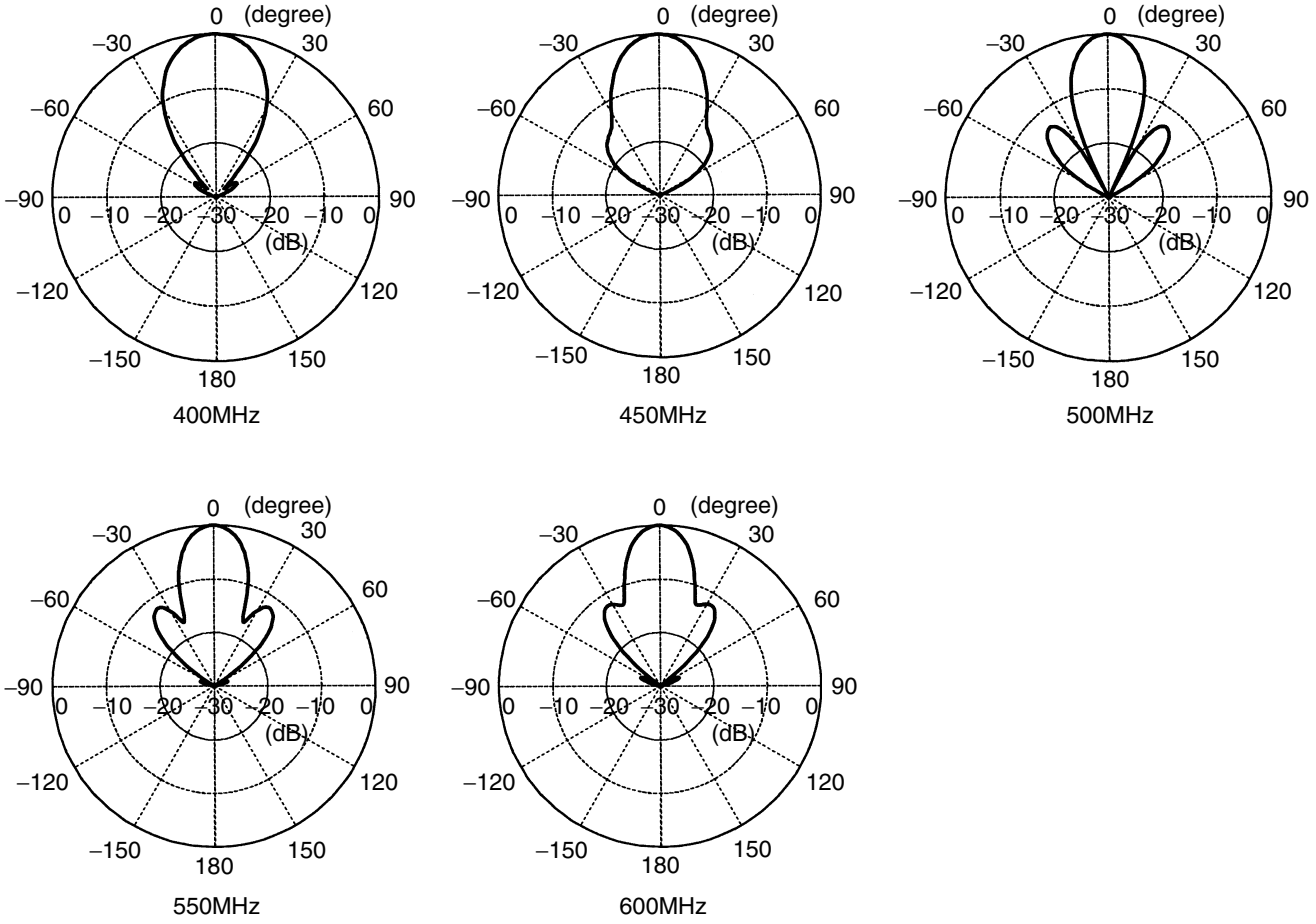


Figure 34. Radiation pattern of PCUMBA [vertical plane (E_ϕ)].

BIBLIOGRAPHY

1. R. W. Masters, The super turnstile, *Broadcast News* **42**: (Jan. 1946).
2. Y. Mushiake, ed., *Antenna Engineering Handbook*, The OHM-Sha, Ltd., Oct. 1980 (in Japanese).
3. R. F. Harrington, *Field Computation by Moment Method*, Macmillan, New York, 1968.
4. W. Berndt, Kombinierte sendeantennen fur fernseh-und UKW-runfunk (teil II), *Telefunken-Zeitung, Jahrgang* **101**: (Aug. 1953).
5. R. F. Harrington, Matrix method for field problems, *Proc. IEEE* **55**(2): 136 (Feb. 1967).
6. P. C. Waterman, Matrix formulation of electromagnetic scattering, *Proc. IEEE* **53**(8): 785 (Aug. 1965).
7. C. D. Taylar and D. R. Wilton, The extended boundary condition solution of the dipole antenna of revolution, *IEEE Trans. Antennas Propag.* **AP-20**(6): 772 (Nov. 1972).
8. L. L. Tsai, Numerical solution for the near and far field of an annular ring of magnetic current, *IEEE Trans. Antennas Propag.* **AP-20**(5): 569 (Sept. 1972).
9. R. W. P. King, *Tables of Antenna Characteristics*, IFI/Plenum, New York, 1971.

TERNARY SEQUENCES

TOR HELLESETH
University of Bergen
Bergen, Norway

1. INTRODUCTION

Sequences with good correlation properties have many applications in modern communication systems. Applications of sequences include signal synchronization, navigation, radar ranging, random-number generation, spread-spectrum communications, multipath resolution, cryptography, and signal identification in multiple-access communication systems.

In *code-division multiple-access* (CDMA) systems several users share a common communication channel by assigning a distinct signature sequence to each user, enabling the user to distinguish his/her signal from those of other users. It is important to select the distinct signature sequences in a way that minimizes the interference from the other users. It is also important to select the sequences such that synchronization is easily achieved.

Phase shift keying (PSK) is a commonly used modulation scheme. In this case the code symbols are

pth roots of unity. Even though binary sequences ($p = 2$) are used in most applications, one can sometimes arrive at better results by going to a larger alphabet. In this article we will consider constructions of ternary sequences ($p = 3$). Ternary sequence families can be constructed with better correlation properties than is possible to obtain by using comparable binary sequence families.

In many cases the best ternary sequences belong to a family of sequences that can be constructed for any p. In other cases the ternary sequences do not have a known nonternary generalization. We will therefore often consider sequences over an alphabet of size being a prime p, when the best ternary sequences can be obtained in this way by letting $p = 3$.

Many sequence families with good correlation properties that are used in many applications employ maximum-length sequences, in short denoted as m sequences, as one of the main ingredients. We will therefore give a detailed definition and presentation of the properties of these sequences. This requires some background material on finite fields (or Galois fields), which will be provided for the sake of completeness.

The m sequences are perhaps the most well known sequences because of their many applications in communication and cryptographic systems. They are deterministic and easy to generate in hardware using linear shift registers and still they possess many randomlike properties. One important property is that an m sequence has a two-level autocorrelation function, a fact that is very useful for synchronization purposes.

From m sequences one can construct ternary sequences with even better autocorrelation properties that are not achievable using binary sequences. Further, we present some more recent constructions of sequences with two-level autocorrelation obtained from ternary m-sequences.

The important aspects for applications are the correlation properties of a sequence or a family of sequences. We define the auto- and cross-correlations of sequences and discuss the important design parameters required by a family of sequences that are to be applied in CDMA systems.

In many practical applications the correlations that occur are rather more aperiodic than periodic in nature. The problem of designing families of sequences with good aperiodic correlation values is very hard in general. Therefore a common approach has been to design sequences with good periodic correlations and see if they satisfy the requirements needed by the applications at hand. The focus in this article will therefore be on periodic correlation.

The main part will be to construct large families of ternary sequences with good correlation properties. Since many of the best sequence families are based on the properties of the cross-correlation between two m sequences, this will be studied in detail.

In order to compare the different constructions of sequence families, we will describe the best known bounds, due to Sidelnikov [1] and Welch [2], on the correlation values of the families. These bounds indicate that nonbinary sequence families can have better correlation parameters than the best binary families.

In later sections we will construct several families of sequences with the best known correlation properties among ternary sequences. These constructions will include sequences by Trachtenberg [3] that represent generalizations of the important binary sequences due to Gold [4]. Some of the constructed sequence families due to Sidelnikov [5], Kumar and Moreno [6], and Helleseth and Sandberg [7] can be shown to be asymptotically optimal. Finally, we will also include tables containing the parameters of some of the best ternary sequence families.

2. CORRELATION OF SEQUENCES

The correlation between two sequences $\{u(t)\}$ and $\{v(t)\}$ of length n is the complex inner product of one of the sequences with a shifted version of the other. The correlation is *periodic* if the shift is a cyclic shift, *aperiodic* if the shift is noncyclic, and a *partial-period* correlation if the inner product involves only a partial segment of the two sequences. We will concentrate on the periodic correlation.

Definition 1. Let $\{u(t)\}$ and $\{v(t)\}$ be two complex-valued sequences of period n, not necessarily distinct. The periodic correlation of the sequences $\{u(t)\}$ and $\{v(t)\}$ at shift τ is defined by

$$\theta_{u,v}(\tau) = \sum_{t=0}^{n-1} u(t+\tau)v^*(t)$$

where the sum over t is computed modulo n and * denotes complex conjugation.

When the two sequences are the same, the correlation is called the *autocorrelation* of the sequence $\{u(t)\}$, whereas when they are distinct, it is common to refer to the *crosscorrelation*.

Example 1. The sequence of period $n = 13$,

$$u(t) = 0\ 0{-}1\ 0{-}1{-}1{-}1{+}1{+}1\ 0{-}1{+}1{-}1$$

has autocorrelation $\theta_{u,u}(\tau) = 0$ for all $\tau \neq 0 \pmod{n}$.

Example 2. Let ω be a complex third root of unity: $\omega^3 = 1$. The sequence of period $n = 26$

$$\{u(t)\} = 11\omega^2 1\omega^2\omega^1\omega^2\omega^2\omega^1 1\omega^2\omega^2\omega^2 11\omega^1 1\omega^1\omega^2\omega^1\omega^1\omega^2 \times 1\omega^1\omega^1\omega^1$$

has autocorrelation $\theta_{u,u}(\tau) = -1$ for all $\tau \neq 0 \pmod{n}$.

Frequently, as $\{u(t)\}$ in example 2, the sequences $\{u(t)\}$ and $\{v(t)\}$ are of the form $u(t) = \omega^{a(t)}$ or $v(t) = \omega^{b(t)}$, where ω is a complex pth root of unity and where the sequences $\{a(t)\}$ and $\{b(t)\}$ take values in the set of integers mod p. In this case we will also sometimes use the notation $\theta_{a,b}$ instead of $\theta_{u,v}$.

For synchronization purposes one prefers sequences with low absolute values of the maximum out-of-phase autocorrelation; that is, $|\theta_{u,u}(\tau)|$ should be small for all

values of $\tau \neq 0 \pmod{n}$. For most applications one needs families of sequences with good simultaneously auto- and cross-correlation properties.

Let $\mathcal{F}$ be a family consisting of M sequences

$$\mathcal{F} = \{\{s_i(t)\}: i = 1, 2, \ldots, M\}$$

where each sequence $\{s_i(t)\}$ has period n.

The cross-correlation between two sequences $\{s_i(t)\}$ and $\{s_j(t)\}$ at shift τ is denoted by $C_{i,j}(\tau)$. In CDMA applications it is desirable to have a family of sequences with certain properties. To facilitate synchronization, it is desirable that all the out-of-phase autocorrelation values ($i = j, \tau \neq 0$) are small. To minimize the interference due to the other users in a multiple-access situation, the cross-correlation values ($i \neq j$) must also be kept small. For this reason the family of sequences should be designed to minimize

$$C_{\max} = \max\{|C_{i,j}|: 1 \leq i, j \leq M, \text{ and either } i \neq j \text{ or } \tau \neq 0\}$$

For practical applications one needs a family $\mathcal{F}$ of sequences of period n, such that the number of users $M = |\mathcal{F}|$ is large and simultaneously $C_{\max}$ is small.

3. GALOIS FIELDS

Many sequence families are most easily described using the theory of finite fields. This section gives a brief introduction to finite fields. In particular the finite field GF(3^3) will be studied in order to provide detailed examples of good sequence families.

There exists finite fields, also known as *Galois fields*, with p^m elements for any prime p and any positive integer m. A Galois field with p^m elements is unique (up to an isomorphism) and is denoted GF(p^m).

For a prime p, let GF(p) = $\{0, 1, \ldots, p-1\}$ denote the integers modulo p with the two operations addition and multiplication modulo p.

To construct a Galois field with p^m elements, select a polynomial $f(x)$ of degree m, with coefficients in GF(p) that is irreducible over GF(p); thus, $f(x)$ can not be written as a product of two polynomials, of degree ≥ 1, with coefficients from GF(p) [irreducible polynomials of any degree m over GF(p) exist].

Let

$$\text{GF}(p^m) = \{a_{m-1}x^{m-1} + a_{m-2}x^{m-2} + \cdots + a_0: a_0, \ldots, a_{m-1} \in \text{GF}(p)\}$$

Then GF(p^m) is a finite field when addition and multiplication of the elements (polynomials) are done modulo $f(x)$ and modulo p. To simplify the notations, let α denote a zero of $f(x)$, that is, $f(\alpha) = 0$. Such an α exists, it can formally be defined as the equivalence class of x modulo $f(x)$.

Example 3. The Galois field GF(3^3) can be constructed as follows. Let $f(x) = x^3 + 2x + 1$, which is easily seen to be an irreducible polynomial over GF(3). Then $\alpha^3 = \alpha + 2$ and

$$\text{GF}(3^3) = \{a_2\alpha^2 + a_1\alpha + a_0: a_0, a_1, a_2 \in \text{GF}(3)\}$$

Computing the powers of α, we obtain

$$\alpha^4 = \alpha \cdot \alpha^3 = \alpha(\alpha + 2) = \alpha^2 + 2\alpha$$

$$\alpha^5 = \alpha \cdot \alpha^4 = \alpha(\alpha^2 + 2\alpha) = \alpha^3 + 2\alpha^2 = 2\alpha^2 + \alpha + 2$$

$$\alpha^6 = \alpha \cdot \alpha^5 = \alpha(2\alpha^2 + \alpha + 2) = 2\alpha^3 + \alpha^2 + 2\alpha = \alpha^2 + \alpha + 1$$

and, similarly, all higher powers of α can be expressed as a linear combination of α^2, α, and 1. In particular, the calculations give $\alpha^{26} = 1$. In Table 1 we give all the powers of α as a linear combination of 1, α, and α^2. The polynomial $a_2\alpha^2 + a_1\alpha + a_0$ is represented as $a_2a_1a_0$.

Hence, the elements $1, \alpha, \alpha^2, \ldots, \alpha^{25}$ are all the nonzero elements in GF(3^3). In general, such an element α, which generates the nonzero elements of GF(p^m), is called a *primitive element* in GF(p^m). Every finite field has a primitive element. An irreducible polynomial $g(x)$ of degree m with coefficients in GF(p) with a primitive element as a zero is called a *primitive polynomial*.

All elements in GF(p^m) are roots of the equation $x^{p^m} - x = 0$. Let β be an element of GF(p^m). It is important to study the polynomials of smallest degree with coefficients in GF(p^m) that has β as a zero. This polynomial is called the *minimum polynomial* of β over GF(p). Note that since all the binomial coefficients $\binom{p}{i}$ are divisible by p, for $1 < i < p$, it follows for any $a, b \in \text{GF}(p^m)$ that

$$(a + b)^p = a^p + b^p$$

Observe that if $m(x) = \sum_{i=0}^{\kappa} m_i x^i$ has coefficients in GF(p) and β as a zero, then

$$m(\beta^p) = \sum_{i=0}^{\kappa} m_i \beta^{pi} = \sum_{i=0}^{\kappa} m_i^p \beta^{pi} = \left(\sum_{i=0}^{\kappa} m_i \beta^i\right)^p = (m(\beta))^p = 0$$

Hence, $m(x)$ has $\beta, \beta^p, \ldots, \beta^{p^{\kappa-1}}$, as zeros where κ is the smallest integer such that $\beta^{p^\kappa} = \beta$. Conversely, the polynomial with exactly these zeros can be shown to be an irreducible polynomial.

Example 4. We will find the minimal polynomial of all the elements in GF(3^3). Let α be a root of $x^3 + 2x + 1 = 0$,

Table 1. The Galois Field GF(3^3)

i	α^i	i	α^i
0	001	13	002
1	010	14	020
2	100	15	200
3	012	16	021
4	120	17	210
5	212	18	121
6	111	19	222
7	122	20	211
8	202	21	101
9	011	22	022
10	110	23	220
11	112	24	221
12	102	25	201

that is, $\alpha^3 = \alpha + 2$. The minimal polynomials over GF(3) of α^i for $0 \le i \le 25$ are denoted $m_i(x)$. Observe by the argument above that $m_{pi}(x) = m_i(x)$, where the indices are taken modulo 26. It follows that

$$\begin{aligned}
m_0(x) &= (x - \alpha^0) = x + 2,\\
m_1(x) &= (x - \alpha)(x - \alpha^3)(x - \alpha^9) = x^3 + 2x + 1,\\
m_2(x) &= (x - \alpha^2)(x - \alpha^6)(x - \alpha^{18}) = x^3 + x^2 + x + 2,\\
m_4(x) &= (x - \alpha^4)(x - \alpha^{12})(x - \alpha^{10}) = x^3 + x^2 + 2,\\
m_5(x) &= (x - \alpha^5)(x - \alpha^{15})(x - \alpha^{19}) = x^3 + 2x^2 + x + 1,\\
m_7(x) &= (x - \alpha^7)(x - \alpha^{21})(x - \alpha^{11}) = x^3 + x^2 + 2x + 1,\\
m_8(x) &= (x - \alpha^8)(x - \alpha^{24})(x - \alpha^{20}) = x^3 + 2x^2 + 2x + 2,\\
m_{13}(x) &= (x - \alpha^{13}) = x + 1,\\
m_{14}(x) &= (x - \alpha^{14})(x - \alpha^{16})(x - \alpha^{22}) = x^3 + 2x + 2,\\
m_{17}(x) &= (x - \alpha^{17})(x - \alpha^{25})(x - \alpha^{23}) = x^3 + 2x^2 + 1.
\end{aligned}$$

This also leads to a factorization into irreducible polynomials of $x^{27} - x$:

$$\begin{aligned}
x^{27} - x &= x\prod_{j=0}^{25}(x - \alpha^j)\\
&= x m_0(x) m_1(x) m_2(x) m_4(x) m_5(x) m_7(x)\\
&\quad \times m_8(x) m_{13}(x) m_{14}(x) m_{17}(x)
\end{aligned}$$

Since a polynomial $m_i(x)$ is primitive if it is irreducible and its zeros are primitive elements, it follows that the polynomial $m_i(x)$ is primitive whenever $\gcd(i, p^m - 1) = 1$. In general the number of primitive polynomials of degree m over GF(p) is $\phi(p^m - 1)/m$, where ϕ is Euler's ϕ function, that is, $\phi(n)$ is the number of integers i such that $1 \le i < n$ and $\gcd(i, n) = 1$. Thus, in our example, $m_1(x), m_5(x), m_7(x), m_{17}(x)$ are the $\phi(26)/3 = 4$ primitive polynomials of degree 3 over GF(3).

4. THE TRACE FUNCTION

In order to describe sequences generated by a linear recursion, it is convenient to introduce the *trace function* from GF(p^m) to GF(p). The *trace* function from GF(3^3) to GF(3) is illustrated in Example 5.

Example 5. The trace function from GF(3^3) to GF(3) is defined by

$$Tr(x) = x + x^3 + x^9$$

Since $x^{27} = x$ and $(x + y)^3 = x^3 + y^3$ holds for all $x, y \in$ GF(3^3), it is easy to see that $\mathrm{Tr}(x) \in$ GF(3) for all $x \in$ GF(3^3), because

$$\begin{aligned}
(\mathrm{Tr}(x))^3 &= (x + x^3 + x^9)^3 = x^3 + x^9 + x^{27}\\
&= x^3 + x^9 + x = \mathrm{Tr}(x)
\end{aligned}$$

Further, it follows that

$$\begin{aligned}
\mathrm{Tr}(x + y) &= \mathrm{Tr}(x) + \mathrm{Tr}(y), \mathrm{Tr}(ax)\\
&= a\mathrm{Tr}(x), \text{ and } \mathrm{Tr}(x^3) = \mathrm{Tr}(x)
\end{aligned}$$

for any $x, y \in$ GF(3^3) and $a \in$ GF(3).

It is straightforward to calculate the trace of the elements in GF(3^3):

$$\begin{aligned}
\mathrm{Tr}(1) &= 1 + 1 + 1 = 0\\
\mathrm{Tr}(\alpha) &= \alpha + \alpha^3 + \alpha^9 = 0\\
\mathrm{Tr}(\alpha^2) &= \alpha^2 + \alpha^6 + \alpha^{18} = 2\\
\mathrm{Tr}(\alpha^4) &= \alpha^4 + \alpha^{12} + \alpha^{10} = 2\\
\mathrm{Tr}(\alpha^5) &= \alpha^5 + \alpha^{15} + \alpha^{19} = 1\\
\mathrm{Tr}(\alpha^7) &= \alpha^7 + \alpha^{21} + \alpha^{11} = 2\\
\mathrm{Tr}(\alpha^8) &= \alpha^8 + \alpha^{24} + \alpha^{20} = 1\\
\mathrm{Tr}(\alpha^{13}) &= \alpha^{13} + \alpha^{13} + \alpha^{13} = 0\\
\mathrm{Tr}(\alpha^{14}) &= \alpha^{14} + \alpha^{16} + \alpha^{22} = 0\\
\mathrm{Tr}(\alpha^{17}) &= \alpha^{17} + \alpha^{25} + \alpha^{23} = 1
\end{aligned}$$

Since $\mathrm{Tr}(x) = \mathrm{Tr}(x^3)$ for all $x \in$ GF(3^3), the trace of all the elements $\mathrm{Tr}(\alpha^t)$ for $t = 0, 1, \ldots, 25$ are given by

$$00202122102220010121120111$$

The sequence $\{\mathrm{Tr}(\alpha^t)\}$ is a ternary m sequence of period $n = 3^3 - 1$, to be studied further in the next section. Note also that $\mathrm{Tr}(\alpha^{t+13}) = \alpha^{13}\mathrm{Tr}(\alpha^t) = 2\mathrm{Tr}(\alpha^t)$.

A more general description of the trace function is given below. It is well known that GF(q^k) $\subset$ GF(q^m) if and only if k divides m (denoted by $k \mid m$). Let q be a power of a prime p, and let $m = ke, k, e \ge 1$. Then the *trace function* Tr_k^m is a mapping from the finite field GF(q^m) to the subfield GF(q^k) given by

$$\mathrm{Tr}_k^m(x) = \sum_{i=0}^{e-1} x^{q^{ki}}$$

Lemma 1. The trace function satisfies the following:

(a) $\mathrm{Tr}_k^m(ax + by) = a\mathrm{Tr}_k^m(x) + b\mathrm{Tr}_k^m(y)$, for all $a, b \in$ GF(q^k), $x, y \in$ GF(q^m).

(b) $\mathrm{Tr}_k^m(x^{q^k}) = \mathrm{Tr}_k^m(x)$, for all $x \in$ GF(q^m).

(c) Let l be an integer such that $k|l|m$. Then

$$\mathrm{Tr}_k^m(x) = \mathrm{Tr}_k^l(\mathrm{Tr}_l^m(x)), \text{ for all } x \in \mathrm{GF}(q^m).$$

(d) For any $b \in$ GF(q^k), it holds that

$$|\{x \in \mathrm{GF}(q^m) \mid \mathrm{Tr}_k^m(x) = b\}| = q^{m-k}$$

(e) Let $a \in$ GF(q^m). If $\mathrm{Tr}_k^m(ax) = 0$ for all $x \in$ GF(q^m), then $a = 0$.

In particular, it is useful to observe that the trace mapping $\mathrm{Tr}_k^m(x)$ takes on all values in the subfield $\mathrm{GF}(q^k)$ equally often when x runs through $\mathrm{GF}(q^m)$. In the cases when the fields involved are clear from the context, we will normally omit the subscripts.

5. MAXIMAL-LENGTH SEQUENCES (*m* SEQUENCES)

Maximal-length linear feedback shift register sequences (or m-sequences) are important in many applications and are building blocks for many important sequence families. An m-sequence has period $n = q^m - 1$ and symbols from a Galois field $\mathrm{GF}(q)$. During a period of an m sequence, each m-tuple of m consecutive symbols, except for the all zero m-tuple, occurs exactly once. Any m sequence can be generated from a linear recursion using a primitive polynomial. The following is an example of a ternary m sequence of period $n = 3^3 - 1 = 26$ having symbols from GF(3).

Example 6. Let $f(x)$ be the ternary primitive polynomial defined by $f(x) = x^3 + 2x + 1$. Define the sequence $\{s(t)\}$ by $s(t+3) + 2s(t+1) + s(t) = 0 \pmod 3$ i.e., $s(t+3) = s(t+1) + 2s(t) \pmod 3$. Therefore, starting with the initial state $(s(0), s(1), s(2)) = (002)$ one generates the m sequence

$$00202122102220010121120111\ldots$$

of period $n = 3^3 - 1 = 26$. Three consecutive positions $(s(t), s(t+1), s(t+2))$ take on all possible nonzero values in $\mathrm{GF}(3)^3$ during a period of the sequence. All other nonzero initial states will generate a cyclic shift of this sequence.

Using the trace function, we can define a sequence $\{s(t)\}$ with symbols in GF(3) such that $s(t) = \mathrm{Tr}(\alpha^t)$, where α is a zero of $f(x) = x^3 + 2x + 1$. This m sequence obeys the recursion $s(t+3) + 2s(t+1) + s(t) = 0 \pmod 3$, since

$$\begin{aligned} s(t+3) + 2s(t+1) + s(t) &= \mathrm{Tr}(\alpha^{t+3}) + 2\mathrm{Tr}(\alpha^{t+1}) + \mathrm{Tr}(\alpha^t) \\ &= \mathrm{Tr}(\alpha^{t+3} + 2\alpha^{t+1} + \alpha^t) \\ &= \mathrm{Tr}(\alpha^t(\alpha^3 + 2\alpha + 1)) \\ &= \mathrm{Tr}(0) \\ &= 0 \end{aligned}$$

All cyclic shifts of $\{s(t)\}$ are described by $\{s(t+\tau)\} = \{\mathrm{Tr}(c\alpha^t)\}$, where $c = \alpha^\tau$.

We next describe m sequences in general over the alphabet $\mathrm{GF}(q)$. A common method to generate sequences is via linear recursions. A linear recursion over $\mathrm{GF}(q)$ is given by

$$\sum_{i=0}^{m} f_i s(t+i) = 0, \text{ for all } t \tag{1}$$

where $f_i \in \mathrm{GF}(q)$ for $i = 0, 1, \ldots, m$. The sequence $\{s(t)\}$ is completely determined by the recursion above and the initial values $s(0), s(1), \ldots, s(m-1)$. The *characteristic polynomial* of the recursion is defined to be

$$f(x) = \sum_{i=0}^{m} f_i x^i$$

The maximum possible period of a sequence generated by a recursion of degree m is $q^m - 1$. This follows since m consecutive symbols uniquely determine the sequence and there are only $q^m - 1$ possible nonzero possibilities for m successive symbols. The maximum period $n = q^m - 1$ is obtained in the case when $f(x)$ is a primitive polynomial.

Some important properties of m sequences are listed next. These properties are easily verified by the sequence in Example 6, but hold for all m sequences in general.

Lemma 2. Let $\{s(t)\}$ be an m sequence of period $q^m - 1$ over $\mathrm{GF}(q)$.

(a) (Balance property) All nonzero elements occur equally often q^{m-1} times and the zero element occurs $q^{m-1} - 1$ times during a period of the m sequence.

(b) (Run property) As t varies over $0 \le t \le q^m - 2$, the m-tuple

$$(s(t), s(t+1), \ldots, s(t+m-1))$$

runs through all the elements in $\mathrm{GF}(q)^m$ exactly once, with the exception of the all-zero m-tuple, which does not occur.

(c) (Shift and add property) For any τ, $0 < \tau \le q^m - 2$, there exists a δ for which

$$s(t) - s(t+\tau) = s(t+\delta), \text{ for all } t$$

(d) (Constancy on cyclotomic cosets) There exists a cyclic shift τ of $\{s(t)\}$, such that $s(p^i t + \tau) = s(t + \tau)$ for all t.

The m sequences generated by the different primitive polynomials $m_i(x)$ are described in Table 2. The m sequences generated by $m_i(x)$ are all cyclic shifts of the sequence $\{\mathrm{Tr}(\alpha^{it})\}$, which is the one listed in Table 2. The sequence $\{s(dt)\}$ is said to be a decimation (by d) of the sequence $\{s(t)\}$, where indices are computed modulo the period n of $\{s(t)\}$. Further, from the trace representation, one observes that an m sequence generated by $m_i(x)$ is obtained by decimating an m-sequence generated by $m_1(x)$ by i.

Table 2. *m* Sequences of Period $n = 3^3 - 1 = 26$

i	$m_i(x)$	m Sequence
1	$x^3 + 2x + 1$	00202122102220010121120111
5	$x^3 + 2x^2 + x + 1$	01211120011010212221002202
7	$x^3 + x^2 + 2x + 1$	02022001222120101100211121
17	$x^3 + 2x^2 + 1$	01110211210100222012212020

Definition 2. We define two sequences $\{s_1(t)\}$ and $\{s_2(t)\}$ to be cyclically equivalent if there exists an integer τ such that

$$s_1(t+\tau) = s_2(t)$$

for all t, otherwise they are said to be *cyclically distinct*.

The sequences generated by different $m_i(x)$ can be shown to be cyclically distinct. Therefore, there are $\phi(3^3 - 1)/3 = 4$ cyclically distinct m sequences of period 26.

6. SEQUENCES WITH LOW AUTOCORRELATION

One attractive property of m sequences of period $n = p^m - 1, p$ a prime, is their *two-level autocorrelation* property.

Theorem 1. The autocorrelation function for an m sequence $\{s(t)\}$ of period $n = p^m - 1$, where p is a prime, is given by

$$\theta_{s,s}(\tau) = \begin{cases} -1 & \text{if } \tau \neq 0 \pmod{p^m - 1} \\ p^m - 1 & \text{if } \tau = 0 \pmod{p^m - 1} \end{cases}$$

Proof In the case $\tau = 0 \pmod{p^m - 1}$, the result follows directly from the definition. In the case $\tau \neq 0 \pmod{p^m - 1}$, define $u(t) = s(t+\tau) - s(t)$ and observe that $\{u(t)\}$ is a nonzero sequence. Further, $\{u(t)\}$ is an m sequence by the shift–add property. The balance property of an m sequence implies that

$$\begin{aligned} \theta_{s,s}(\tau) &= \sum_{i=0}^{p^m-2} \omega^{s(t+\tau)-s(t)} \\ &= \sum_{i=0}^{p^m-2} \omega^{u(t)} \\ &= (p^{m-1} - 1)\omega^0 + p^{m-1} \sum_{i=1}^{p-1} \omega^i = -1 \end{aligned}$$

where ω is a complex primitive pth root of unity.

Other well-known sequences with two-level autocorrelation properties are the *GMW sequences*, due to Gordon, Welch, and Mills. These sequences can most easily be described in terms of the trace function.

Theorem 2. Let k, m be integers with $k \mid m, k \geq 1$. Let $r, 1 \leq r \leq q^k - 2$ satisfy $\gcd(r, q^k - 1) = 1$. Let a be a nonzero element in $\mathrm{GF}(q^m)$, and let $\{s(t)\}$ be the GMW sequence of period $q^m - 1$ defined by

$$s(t) = \mathrm{Tr}_1^k\{[\mathrm{Tr}_k^m(a\alpha^t)]^r\}$$

where α is a primitive element in $\mathrm{GF}(q^m)$. Then

(a) $\{s(t)\}$ is balanced.

(b) If q is a prime, then

$$\theta_{s,s}(\tau) = \begin{cases} -1 & \text{if } \tau \neq 0 \pmod{q^m - 1} \\ q^m - 1 & \text{if } \tau = 0 \pmod{q^m - 1} \end{cases}$$

The low values of the out-of-phase autocorrelation of m sequences and GMW sequences make them attractive in many applications that require synchronization. The *linear span* of a sequence is the smallest degree of a characteristic polynomial that generates the sequence. One important advantage with GMW sequences is that they in general have a much larger linear span than do m sequences. This is important in some applications.

Definition 3. A sequence $\{s(t)\}$ of period n is said to have a "perfect" autocorrelation function if

$$\theta_{s,s}(\tau) = 0 \quad \text{for all } \tau \neq 0 \pmod{n}$$

Sequences with perfect autocorrelation functions do not always exist. For example, binary $\{-1, +1\}$ sequences with odd period cannot have this property since all the autocorrelation values necessarily have to be odd. For any period $n > 4$, no binary $\{-1, +1\}$ sequence is known to have perfect autocorrelation.

One family of *ternary* sequences $\{u(t)\}$ with perfect autocorrelation are the Ipatov sequences. These sequences are based on m-sequences over $\mathrm{GF}(q)$, q odd, of period $n = q^m - 1$, combined with the quadratic character of $\mathrm{GF}(q)$, which is a mapping from $\mathrm{GF}(q)$ to the set $\{0, -1, +1\}$.

A nonzero element x is said to be a square in $\mathrm{GF}(q)$ if $x = y^2$ has a solution $y \in \mathrm{GF}(q)$. In order to define the ternary Ipatov sequences, first define the quadratic character χ:

$$\chi(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \text{ is a square in } \mathrm{GF}(q) \\ -1 & \text{if } x \text{ is a nonsquare in } \mathrm{GF}(q) \end{cases}$$

The quadratic character χ has the property that $\chi(xy) = \chi(x)\chi(y)$ for all $x, y \in \mathrm{GF}(q)$. In particular, if α is a primitive element in $\mathrm{GF}(q)$, then the squares in $\mathrm{GF}(q)$ are exactly the even powers of α; thus $\chi(\alpha^i) = (-1)^i$.

Combining the properties of m sequences and the quadratic character χ leads to the following ternary sequences where the out-of-phase autocorrelation is always zero. Note, however, that these sequences contain q^{m-1} zero elements.

Theorem 3. Let $\{s(t)\}$ be an m sequence over $\mathrm{GF}(q)$ of period $n = q^m - 1$ where $q = p^r, p$ an odd prime, and m odd. Let $f(x)$ be the characteristic polynomial of $\{s(t)\}$, and let α be a zero of $f(x)$. The ternary sequence $\{u(t)\}$ with symbols from $\{0, -1, +1\}$ defined by

$$u(t) = (-1)^t \chi(s(t))$$

has period $N = (q^m - 1)/(q - 1)$ and perfect autocorrelation.

Proof The proof follows nicely from the properties of m-sequences and the quadratic character χ. First note that from the trace representation of an m-sequence it follows without loss of generality that

$$s(t+N) = \mathrm{Tr}(\alpha^{t+N}) = \alpha^N \mathrm{Tr}(\alpha^t)$$

where $\alpha^N \in \mathrm{GF}(q)$. Hence

$$u(t+N) = (-1)^{t+N}\chi(s(t+N)) = (-1)^{t+N}\chi(\alpha^N)\chi(s(t)) = (-1)^t\chi(s(t)) = u(t)$$

which implies that $\{u(t)\}$ has period N.

To show that $\{u(t)\}$ has perfect autocorrelation, let $\tau \neq 0 \pmod N$ and use the property that $\{u(t)\}$ has period N. Then

$$\theta_{u,u}(\tau) = \sum_{t=0}^{N-1} u(t+\tau)u^*(t) = \frac{1}{q-1}\sum_{t=0}^{q^m-2}(-1)^{\tau}\chi(s(t+\tau)s(t))$$

Since each pair $(s(t+\tau), s(t)) \neq (0,0)$ occurs q^{m-2} times in an m sequence when t runs through $t = 0, 1, \ldots, q^m - 2$, it follows that each nonzero element in $\mathrm{GF}(q)$ appears $q^{m-1}(q-1)$ times as a product $s(t+\tau)s(t)$. Hence

$$\theta_{u,u}(\tau) = (-1)^{\tau}q^{m-1}\sum_{x\in\mathrm{GF}(q)}\chi(x) = 0,$$

since the numbers of squares equals the number of nonsquares in $\mathrm{GF}(q)$.

Example 7. This example constructs an Ipatov sequence of period 13 from an m sequence of period $n = 3^3 - 1 = 26$. Note that $\chi(0) = 0, \chi(1) = 1, \chi(2) = -1$. Let $+$ and $-$ denote $+1$ and -1 respectively. The m sequence

$$00202122102220010121120111$$

leads to the following ternary Ipatov sequence of period 13

$$00-0----++0-+-$$

which possesses perfect autocorrelation.

Ipatov sequences can be constructed for all length $n = (q^m - 1)/(q-1)$ whenever p and m are odd. Høholdt and Justesen have constructed ternary sequences with perfect autocorrelation also in the case when $q = 2^r$ is even.

7. THE CROSS-CORRELATION OF *m* SEQUENCES

In this section we study the cross-correlation between two different m sequences of period $n = p^m - 1$ with symbols from $\mathrm{GF}(p)$, where p is a prime. Many families of sequences with good correlation properties can be derived from these cross-correlation properties.

Any m sequence of length $n = p^m - 1$ can (after a cyclic shift) be obtained by decimating $\{s(t)\}$ by a d relatively prime to n. We use $C_d(\tau)$ to denote the cross-correlation function between the m sequence $\{s(t)\}$ and its decimation $\{s(dt)\}$. By definition, we have

$$C_d(\tau) = \sum_{t=0}^{p^m-2}\omega^{s(t+\tau)-s(dt)}$$

We can assume after suitable shifting that $\{s(t)\}$ is in the form $s(t) = \mathrm{Tr}(\alpha^t)$, where $\mathrm{Tr}(x)$ denotes the trace function from $\mathrm{GF}(p^m)$ to $\mathrm{GF}(p)$. Using the properties of the trace function, this can be reformulated as

$$C_d(\tau) = \sum_{t=0}^{p^m-2}\omega^{\mathrm{Tr}(\alpha^{t+\tau})-\mathrm{Tr}(\alpha^{dt})} = \sum_{x\in\mathrm{GF}(p^m)\setminus\{0\}}\omega^{\mathrm{Tr}(cx-x^d)}$$

where $c = \alpha^{\tau}$.

Hence, finding the cross-correlation function between m sequences is equivalent to evaluating the sum above. Such sums, called *exponential sums*, have been extensively studied in the literature. Several of the best sequence families applied in CDMA systems are based on estimates of such exponential sums.

In the case when $\{s(t)\} = \{s(dt)\}$, that is, when $d \equiv p^i \pmod{p^m - 1}$, then we have the two-valued autocorrelation function of the m sequence $\{s(t)\}$. When the two sequences are cyclically distinct, that is, when $d \not\equiv p^i \pmod{p^m - 1}$, then the cross-correlation function $C_d(\tau)$ is known to take on at least three different values when $\tau = 0, 1, \ldots, p^m - 2$.

It is therefore of special interest to study the cases when exactly three values occur. In addition to being a long studied and challenging mathematical problem, it turns out that several of these cases often lead to a low maximum absolute value of the cross-correlation. In the binary case when $p = 2$, the following six decimations give three-valued cross-correlation; these decimations cover all the known binary cases, but there are no proofs that these are the only ones:

1. $d = 2^k + 1, m/\gcd(m,k)$ odd.
2. $d = 2^{2k} - 2^k + 1, m/\gcd(m,k)$ odd.
3. $d = 2^{m/2} + 2^{(m+2)/4} + 1, m \equiv 2 \pmod 4$.
4. $d = 2^{(m+2)/2} + 3, m \equiv 2 \pmod 4$.
5. $d = 2^{(m-1)/2} + 3, m$ odd.
6. $d = \begin{cases} 2^{(m-1)/2} + 2^{(m-1)/4} - 1 & \text{if } m \equiv 1 \pmod 4 \\ 2^{(m-1)/2} + 2^{(3m-1)/4} - 1 & \text{if } m \equiv 3 \pmod 4 \end{cases}$

Case 1 is the oldest result due to Gold [4] in 1968, and forms the basis for the binary Gold sequences found in numerous applications. Case 2 produces sequences with properties similar to those of the Gold sequences and was first proved by Kasami and Welch in the late 1960s ago.

In the nonbinary case when $p > 2$, fewer cases give three-valued cross-correlation. The following two cases have a three-valued cross-correlation; these were proved by Trachtenberg [3] for m odd and extended to the case $m/\gcd(m,k)$ odd in Helleseth [8]:

1. $d = (p^{2k} + 1)/2, m/\gcd(m,k)$ odd.
2. $d = p^{2k} - p^k + 1, m/\gcd(m,k)$ odd.

These decimations are analogs to the Gold and Kasami–Welch decimations, respectively. The three values of the cross-correlation that occur in these two cases are $\{-1, -1 \pm p^{(m+e)/2}\}$, where $e = \gcd(k,m)$. The maximum

absolute values of the cross-correlation function are smallest in the case when m is odd and $\gcd(k, m) = 1$, when the cross-correlation values are $\{-1, -1 \pm p^{(m+1)/2}\}$. For $p > 3$, there are no other decimations known that have a three-valued cross-correlation.

Dobbertin et al. [9] found new decimations for ternary m sequences that give three-valued cross-correlation. The family of ternary sequences presented below do not have a known analogue when $p > 3$.

Theorem 4. Let $d = 2 \cdot 3^{(m-1)/2} + 1$, where m is odd, then the cross-correlation function $C_d(\tau)$ takes on the following three values:

$-1 + 3^{(m+1)/2}$	occurs	$\frac{1}{2}(3^{m-1} + 3^{(m-1)/2})$	times
-1	occurs	$3^m - 3^{m-1} - 1$	times
$-1 - 3^{(m+1)/2}$	occurs	$\frac{1}{2}(3^{m-1} - 3^{(m-1)/2})$	times

Numerical results have also revealed some other decimations with the same cross-correlation properties. At present this is the only known open case of three-valued cross-correlation of m sequences.

Conjecture 1. Let $d = 2 \cdot 3^r + 1$, where m is odd, and

$$r = \begin{cases} \dfrac{m-1}{4} & \text{if } m \equiv 1 (\text{mod } 4) \\ \dfrac{3m-1}{4} & \text{if } m \equiv 3 (\text{mod } 4) \end{cases}$$

then the cross-correlation function $C_d(\tau)$ is as in Theorem 4.

New ternary sequences of period $n = 3^m - 1$, with the same autocorrelation values as m sequences, have been constructed by adding two m sequences of this period. Two simple examples are given below. Let $d = 2 \cdot 3^{(m-1)/2} + 1$, where m is odd or $d = 2^{2k} - 2^k + 1$, where $m = 3k$ [10]. Let $s(t) = \text{Tr}(\alpha^t)$; then the ternary sequence $\{u(t)\}$, where

$$u(t) = s(t) + s(dt)$$

has the same two-level autocorrelation function as the m sequence of the same period. More recently, these results have been generalized to nonternary sequences.

8. BOUNDS ON SEQUENCE CORRELATIONS

Practical applications require a family $\mathcal{F}$ of sequences of period n, such that the number of users $M = |\mathcal{F}|$ is large and simultaneously the maximum value of the autocorrelation and crosscorrelation, $C_{\max}$, is as small as possible. Welch and Sidelnikov provide important lower bounds on the minimum value of $C_{\max}$ for a family of sequences of size M and period n. These bounds are important in comparing different sequence designs with the optimal possible achievable parameters.

The following bound is due to Welch [2].

Theorem 5. Let $\mathcal{F}$ be a family of M complex-valued sequences of period n

$$\mathcal{F} = \{\{b_i(t)\}: i = 1, 2, \ldots, M\}$$

where each sequence has norm (or energy) n:

$$\sum_{t=0}^{n-1} |b_i(t)|^2 = n, \quad \text{for all } i = 1, 2, \ldots, M$$

Let $\rho_{\max}$ be the maximum nontrivial correlation value of the family $\mathcal{F}$

$$\rho_{\max} = \max \left\{ \begin{array}{l} \displaystyle\sum_{t=0}^{n-1} b_i(t+\tau) b_j^*(t): 1 \le i, \\ j \le M \text{ either } i \ne j \text{ or } \tau \ne 0 \end{array} \right\}$$

where * denotes complex conjugation. Then for all $k \ge 0$, it holds that

$$\rho_{\max}^{2k} \ge \frac{1}{nM - 1} \left\{ \frac{Mn^{2k+1}}{\binom{k+n-1}{n-1}} - n^{2k} \right\}$$

To obtain a lower bound on the maximum nontrivial correlation for a family of sequences, $\{b_i(t)\}$, over an alphabet of size p, one applies the Welch bound to the family

$$\mathcal{F} = \{\{\omega^{b_i(t)}\}: i = 1, 2, \ldots, M\}$$

The following bound is due to Sidelnikov [1].

Theorem 6. Let $\mathcal{F}$ be a family of M sequences of period n over an alphabet of size p. In the case $p = 2$, then

$$C_{\max}^2 > (2k+1)(n-k) + \frac{k(k+1)}{2} - \frac{2^k n^{2k+1}}{m(2k)! \binom{n}{k}}, \quad 0 \le k < \frac{2n}{5}$$

In the case $p > 2$, then

$$C_{\max}^2 > \left(\frac{k+1}{2}\right)(2n-k) - \frac{2^k n^{2k+1}}{m(k!)^2 \binom{2n}{k}}, \quad k \ge 0$$

The Welch bound applies to complex-valued sequences, while the Sidelnikov bound applies to sequences over any finite alphabet. To find the best result one normally has to examine both bounds. Some improvements of these bounds have been obtained by Levenshtein [11]. Applying the Sidelnikov bound for a family of size $M = n^u$ for some $u \ge 1$, one observes that the best bound is usually obtained by setting $k = \lfloor u \rfloor$. The Sidelnikov bound is well approximated, when $n \gg u$, by

$$C_{\max}^2 > n \left\{ (2u+1) - \frac{1}{(2u-1)!!} \right\} \quad \text{for} \quad p = 2$$

and

$$C_{\max}^2 > n \left\{ (u+1) - \frac{1}{(u)!} \right\} \quad \text{for} \quad p > 2$$

where $(2u-1)!!$ denotes $1 \cdot 3 \cdot 5 \cdots (2u-1)$. From these approximations, one sees that the bounds indicate that one may improve the performance in terms of $C_{\max}$ using

a nonbinary alphabet instead of a binary alphabet. The improvement can be obtained for a small alphabet, even for ternary sequences.

Using the tightest lower bound on the maximum correlation parameters obtained from the Welch or Sidelnikov bounds one can derive asymptotic bounds. For example, in the case when the size of the sequence family $M = |\mathcal{F}|$ is approximately equal to the length n, the lower bounds are $C_{\max} \approx (2n)^{1/2}$ for $p = 2$ and $C_{\max} \approx n^{1/2}$ for $p > 2$.

9. SEQUENCE FAMILIES WITH LOW CORRELATIONS

This section surveys and compares some of the best-known ternary sequence designs for CDMA applications. Several of the sequence designs are asymptotically optimal and give good parameters also when $p \neq 3$. First we describe sequences based on the crosscorrelation of m sequences.

Theorem 7. Let $\{s(t)\}$ and $\{s(dt)\}$ be two m sequences of period $n = p^m - 1$, p a prime. Let $\mathcal{S}$ be the family of sequences given by

$$\mathcal{S} = \{\{s(t+\tau) + s(dt)\}: \tau = 0, 1, \ldots, p^m - 2\} \cup \{s(t)\} \cup \{s(dt)\}$$

Then $|\mathcal{S}| = n + 2 = p^m + 1$ and the maximum nontrivial auto- or cross-correlation value in the family is $C_{\max} = \max\{|C_d(\tau)|: \tau = 0, 1, \ldots, n-1\}$.

Proof The idea behind this construction is that the cross-correlation between any two sequences in $\mathcal{S}$ equals the cross-correlation between two m sequences that differ by a decimation d. Let $\{u_1(t)\}$ and $\{u_2(t)\}$ be two sequences in $\mathcal{S}$. Consider the typical case when $u_1(t) = s(t+\tau_1) + s(dt)$ and $u_2(t) = s(t+\tau_2) + s(dt)$.

Using the shift–add property of m sequences, the difference $u_1(t+\tau) - u_2(t)$ can be calculated by

$$\begin{aligned} u_1(t+\tau) - u_2(t) &= (s(t+\tau_1+\tau) + s(d(t+\tau))) \\ &\quad - (s(t+\tau_2) + s(d(t))) \\ &= (s(t+\tau_1+\tau) - s(t+\tau_2)) \\ &\quad - (s(dt) - s(dt+d\tau)) \\ &= s(t+\gamma_1) - s(dt+\gamma_2) \\ &= s(t+\gamma) - s(dt) \end{aligned}$$

Therefore, the correlation between two sequences in $\mathcal{S}$ equals the cross-correlation between $\{s(t)\}$ and $\{s(dt)\}$, which has a maximal absolute value $C_{\max} = \max\{|C_d(\tau)|: \tau = 0, 1, \ldots, n-1\}$.

It follows from this result and the known results on the cross-correlation function of m sequences that the best sequence designs of this form are obtained for the values of d for which the maximum absolute value of the cross-correlation is as small as possible. This happens in the following cases:

1. $d = (p^{2k}+1)/2$, m odd, $\gcd(k, m) = 1$.
2. $d = p^{2k} - p^k + 1$, m odd, $\gcd(k, m) = 1$.
3. $d = 2 \cdot 3^{(m-1)/2} + 1$, m odd and $p = 3$.

These cases give $C_{\max} = p^{(m+1)/2} + 1 = (p(n+1))^{1/2} + 1$, and thus the parameters of the sequence families above are

$$M = |\mathcal{S}| = n + 1, n = p^m - 1, \text{ and } C_{\max} = (p(n+1))^{1/2} + 1$$

For m even, there are better constructions using the cross-correlation of other pairs of m sequences. The best sequence family of this type has parameters

$$M = |\mathcal{S}| = n + 1, n = p^m - 1, \text{ and } C_{\max} = 2(n+1)^{1/2} - 1$$

This family is based on m sequences with a four-valued cross-correlation function with values $\{-1 - p^{m/2}, -1, -1 + p^{m/2}, -1 + 2p^{m/2}\}$; see Helleseth [8], where

$$d = 2p^{m/2} - 1 \text{ and } p^{m/2} \neq 2 (\text{mod } 3)$$

It is natural to compare the parameters of these designs with the best possible that may be achieved according to the Welch or Sidelnikov bound. For a sequence family of size approximately equal to the length (i.e., $M \approx n$), both bounds give an asymptotic lower bound on $C_{\max} \approx n^{1/2}$ (for $p > 2$). The best sequence families obtained from Theorem 7 are a factor $p^{1/2}$ off the optimal value for m odd and a factor of 2 off the optimal value for m even.

It is interesting to observe that the popular binary Gold sequence family has also size $M \approx n$, and $C_{\max} \approx (2n)^{1/2}$. According to the Sidelnikov bound for $p = 2$, this is optimal for binary sequences. The Welch and Sidelnikov bounds indicate that this can be improved with a factor of $2^{1/2}$ for nonbinary sequences since asymptotically $C_{\max} \approx n^{1/2}$ may be achievable.

We next construct sequence families better than the ones obtained from the cross-correlation of m sequences. These sequence families are asymptotically optimal with respect to the Welch and Sidelnikov bounds. These sequences can be described using *perfect nonlinear* (*PN*) *mappings*.

Let $f(x)$ be a function $f: \mathrm{GF}(p^m) \to \mathrm{GF}(p^m)$. Then f is said to be a perfect nonlinear mapping, if $f(x+a) - f(x)$ is a permutation of $\mathrm{GF}(p^m)$ for any nonzero $a \in \mathrm{GF}(p^m)$. From PN mappings with $f(x) = x^d$, one can construct families of sequences with good correlation properties when p is an odd prime.

Example 8. There are three known families of PN mapping of the form $f(x) = x^d$; the first two work for any odd prime p, while the third works only for $p = 3$:

(a) $d = 2$.
(b) $d = p^k + 1$ where $m/\gcd(k, m)$ is odd.
(c) $d = (3^k + 1)/2$ where k is odd and $\gcd(k, m) = 1$.

Let ω be a complex pth root of unity, and let $\mathrm{Tr}(x)$ denote the trace mapping from $\mathrm{GF}(p^m)$ to $\mathrm{GF}(p)$. For any PN mapping, it holds for any nonzero $a \in \mathrm{GF}(p^m)$, that

$$\sum_{x \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(f(x+a)-f(x))} = \sum_{b \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(b)} = 0$$

since the trace function takes on all values in $\mathrm{GF}(p)$ equally often.

Let $c, \lambda \in \mathrm{GF}(p^m)$ and $c \neq 0$; then the following exponential sum is the key to the asymptotically optimal sequence families constructed from PN mappings:

$$S(c,\lambda) = \sum_{x \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(cf(x)+\lambda x)}$$

It is important to determine $|S(c,\lambda)|$, which can be done as follows

$$\begin{aligned} |S(c,\lambda)|^2 &= \sum_{x,y \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(c(f(x)-f(y))+\lambda(x-y))} \\ &= \sum_{y,z \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(c(f(y+z)-f(y))+\lambda z)} \\ &= p^m \end{aligned}$$

since the PN property only gives a contribution when $z = 0$. Hence

$$|S(c,\lambda)| = |\sum_{x \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(cf(x)+\lambda x)}| = p^{m/2}$$

This leads to the following asymptotically optimal sequence family.

Theorem 8. Let $f(x) = x^d$ be a PN mapping of GF (p^m). Let α be a primitive element in GF (p^m). Let $\{s_c(t)\}$ be the sequence of period $n = p^m - 1$ defined by

$$s_c(t) = \mathrm{Tr}(\alpha^t + cf(\alpha^t))$$

Let $\mathcal{F}$ be the family of sequences

$$\mathcal{F} = \{\{s_c(t)\} : c \in \mathrm{GF}(p^m)\}$$

Then $\mathcal{F}$ is a family of $M = p^m$ sequences of period $n = p^m - 1$ and $C_{\max} \leq 1 + p^{m/2}$.

Proof The cross-correlation between two sequences in the family is

$$\begin{aligned} \theta s_{c_1}, s_{c_2}(\tau) &= \sum_{t=0}^{p^m-2} \omega^{s_{c_1}(t+\tau)-s_{c_2}(t)} \\ &= \sum_{t=0}^{p^m-2} \omega^{\mathrm{Tr}((c_1\alpha^{d\tau}-c_2)\alpha^{dt}+(\alpha^\tau-1)\alpha^t)} \\ &= -1 + \sum_{x \in GF(p^m)} \omega^{\mathrm{Tr}(cf(x)+\lambda x)} \end{aligned}$$

where $c = c_1\alpha^{d\tau} - c_2$ and $\lambda = \alpha^\tau - 1$. Hence, the sum above gives $|\theta s_{c_1}, s_{c_2}(\tau)| \leq 1 + p^{m/2}$, except when $c_1 = c_2$ and $\tau = 0$.

Thus, the family $\mathcal{F}$ is asymptotically optimal with respect to the Welch–Sidelnikov bound. The family corresponding to the PN mapping with $d = 2$ is due to Sidelnikov, while $d = p^k + 1, m/\gcd(k,m)$ odd is due to Kumar and Moreno [6]. The connection between PN functions and these optimal families were pointed out by Helleseth and Sandberg [7], who also described the ternary sequence family for $d = (3^k + 1)/2$, where k is odd and $\gcd(k.m) = 1$.

Example 9. In the following example we describe a family of ternary Kumar–Moreno sequences of period $n = 3^3 - 1 = 26$ for $d = 3^k + 1$, where $k = 1$, namely, $d = 3^1 + 1 = 4$. The ith sequence in the family is defined to be

$$s_i(t) = \mathrm{Tr}(\alpha^t + \beta_i \alpha^{dt}), \text{ where } \beta_i \in \mathrm{GF}(3^3)$$

Let $u(t) = \mathrm{Tr}(\alpha^t), v_0(t) = \mathrm{Tr}(\alpha^{4t}), v_1(t) = \mathrm{Tr}(\alpha^{4t+1})$, then

$$\begin{aligned} u(t) &= 00202122102220010121120111 \\ v_0(t) &= 02120112220200212011222020 \\ v_1(t) &= 01001210221110100121022111 \end{aligned}$$

All sequences in the family $\mathcal{F}$ can be obtained by adding different cyclic shifts of the sequences $\{v_0(t)\}$ or $\{v_1(t)\}$ (both of period 13) to the sequence $\{u(t)\}$. Figure 1 shows the shiftregister that generates all the sequences in $\mathcal{F}$. The maximal absolute value of the correlation for this family is $C_{\max} \leq 1 + (27)^{1/2}$.

Most of the families discussed above have family size M approximately equal to the length n. Because of the large increase in the number of users in modern-day communication systems, there is a demand for larger sequence families. A general construction due to Sidelnikov [1] based on general bounds on exponential sums is presented below. These families have a flexible and larger family size.

The main idea is that if $f(x)$ is a polynomial of degree $d \geq 1$ with coefficients from GF (p^m) that cannot be expressed in the form $f(x) = g(x)^p - g(x) + c$ for any $g(x)$ with coefficients from GF (p^m) and any $c \in GF(p^m)$, then

$$\left| \sum_{x \in \mathrm{GF}(p^m)} \omega^{\mathrm{Tr}(f(x))} \right| \leq (d-1)\sqrt{p^m}$$

Theorem 9. Let p be a prime, $q = p^m$ and α an element of order $n \mid q - 1$ in GF (q). Consider the family of all sequences $\{s(t)\}$ over GF (p) of the form

$$s(t) = \mathrm{Tr}\left\{\sum_{k=1}^{d} a_k \alpha^{kt}\right\}$$

where $(a_1, a_2, \ldots, a_d) \in GF(q)^d$ and d is an integer satisfying $1 \leq d \leq p^{\lfloor (m+1)/2 \rfloor}$. Let $\mathcal{F}$ be the subset of this family consisting of those sequences having period equal to n. If M denotes the number of cyclically distinct sequences in this family and $C_{\max}$ the maximum correlation value, then

$$M \geq \frac{q-1}{n} q^{d - \lfloor d/p \rfloor - 1}$$

and

$$C_{\max} \leq \left(\frac{d-n}{(q-1)}\right) q^{1/2} + \frac{n}{q-1}$$

The main idea behind this construction is that if $\{s_1(t)\}$ and $\{s_2(t)\}$ correspond to polynomials $f_1(x)$ and $f_2(x)$ of degree $\leq d$, respectively, then the difference

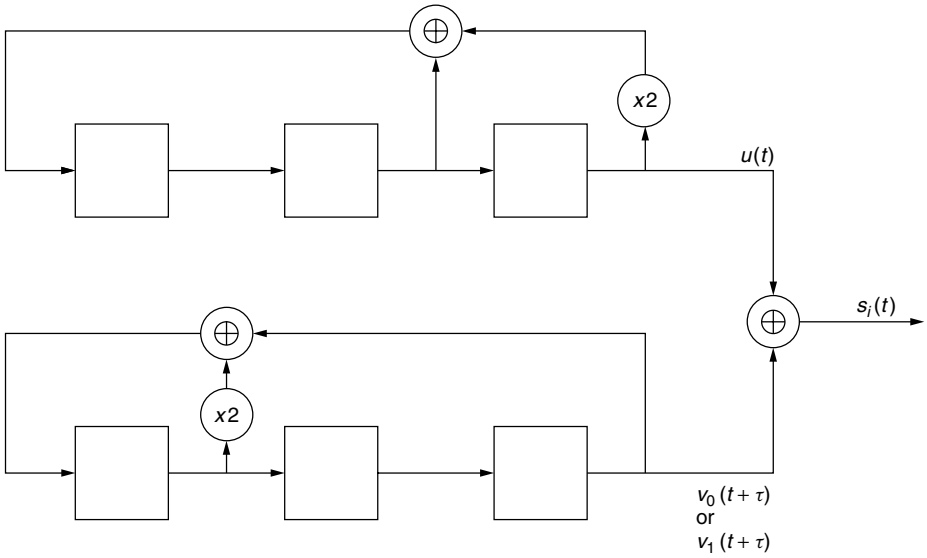


Figure 1. A ternary Kumar–Moreno sequence family.

$s_1(t+\tau) - s_2(t)$ involved in the correlation calculations corresponds to the polynomial $f_1(cx) - f_2(x)$, where $c = \alpha^\tau$. Since $f_1(cx) - f_2(x)$ has degree $\leq d$, the exponential sum bound above gives the result.

The entry attributed to Sidelnikov in Table 3 corresponds to the sequences above with $n = q - 1$ and $d = 2$. These sequences are the same as the ones obtained from the PN mapping with $d = 2$ above.

Bent sequences is another family of sequences that in addition to low cross-correlation have large linear span. This is a family of sequences that exists for all primes p. The sequences in the family have period $n = p^m - 1$ for even values of m. The sequences in the family are balanced and have maximum correlation $C_{\max} = p^{m/2} + 1$. For further information on nonbinary bent sequences, see Kumar et al. [12].

Table 3 compares some of the families of sequences constructed in this section. The families have size M approximately equal to the period n.

For further information on sequences and their correlation properties, the reader is referred to surveys on sequences that can be found in Refs. 13–15.

BIOGRAPHY

Tor Helleseth received the Cand.Real. and Dr.Philos. degrees in mathematics from the University of Bergen, Bergen, Norway, in 1971 and 1979, respectively.

From 1973 to 1980 he was a Research Assistant at the Department of Mathematics, University of Bergen. From 1981 to 1984 he was a Researcher at the Chief Headquarters of Defense in Norway. Since 1984 he has been a Professor at the Department of Informatics at the University of Bergen.

From 1991 to 1993 he served as an Associate Editor for Coding Theory for *IEEE Transactions on Information Theory*. Since 1996 he has been on the editorial board of *Designs, Codes and Cryptography*. He was Program Chairman for Eurocrypt'93 and for the 1997 Information Theory Workshop. He was one of the organizers of the conferences on Sequences and Their Applications SETA'98 and SETA'01. He has published more than 100 scientific papers in international journals.

In 1997 he was elected an IEEE Fellow for his contributions to coding theory and cryptography. His

Table 3. Families of Sequence Designs

Family	Length	Alphabet	Family Size	$C_{\max}$
Trachtenberg [3]	$n = p^m - 1$			
	p prime, m odd	p	$n+2$	$1 + ((n+1)p)^{1/2}$
Dobbertin et al. [9]	$n = 3^m - 1$			
	m odd	3	$n+2$	$1 + ((n+1)3)^{1/2}$
Helleseth [8]	$n = p^m - 1$			
	p prime, m even	p	$n+2$	$-1 + 2(n+1)^{1/2}$
	$p^{m/2} \neq 2 \pmod 3$			
Sidelnikov [1,5]	$p^m - 1$			
	p prime	p	$n+1$	$1 + (n+1)^{1/2}$
Kumar–Moreno [6]	$n = p^m - 1$			
	p prime	p	$n+1$	$1 + (n+1)^{1/2}$
Helleseth–Sandberg [7]	$n = 3^m - 1$	3	$n+1$	$1 + (n+1)^{1/2}$
Bent sequences	$p^m - 1$			
	p prime, m even	p	$(n+1)^{1/2}$	$1 + (n+1)^{1/2}$

research interests include coding theory, sequence designs, and cryptography.

BIBLIOGRAPHY

1. V. M. Sidelnikov, On mutual correlation of sequences, *Soviet Math. Doklad.* **12**: 197–201 (1971).
2. L. R. Welch, Lower bounds on the maximum cross correlation of signals, *IEEE Trans. Inform. Theory* **IT-20**: 397–399 (1974).
3. H. M. Trachtenberg, *On the Cross-Correlation Functions of Maximal Linear Recurring Sequences*, Ph.D. thesis, Univ. Southern California, 1970.
4. R. Gold, Maximal recursive sequences with 3-valued recursive cross-correlation functions, *IEEE Trans. Inform. Theory* **IT-14**: 154–156 (1968).
5. V. M. Sidelnikov, Cross correlation of sequences, *Probl. Kybern.* **24**: 15–42 (1971) (in Russian).
6. P. Kumar and O. Moreno, Prime-phase sequences with periodic correlation properties better than binary sequences, *IEEE Trans. Inform. Theory* **IT-37**: 603–616 (1991).
7. T. Helleseth and D. Sandberg, Some power mappings with low differential uniformity, *Appl. Algebra Eng. Commun. Comput.* **8**: 363–370 (1997).
8. T. Helleseth, Some results about the cross-correlation function between two maximal linear sequences, *Discrete Math.* **16**: 209–232 (1976).
9. H. Dobbertin, T. Helleseth, V. Kumar, and H. Martinsen, Ternary m-sequences with three-valued crosscorrelation function: Two new decimations, *IEEE Trans. Inform. Theory* **IT-47**: 1473–1481 (2001).
10. T. Helleseth, P. V. Kumar, and H. Martinsen, A new family of sequences with ideal two-level autocorrelation function, *Designs, Codes Cryptogr.* **23**: 157–166 (2001).
11. V. I. Levenshtein, Bounds on the maximal cardinality of a code with bounded modules of the inner product, *Soviet Math. Doklad.* **25**: 526–531 (1982).
12. P. V. Kumar, R. A. Scholtz, and L. R. Welch, Bent function sequences, *IEEE Trans. Inform. Theory* **IT-28**: 858–864 (1982).
13. J. Gibson, ed., *The Mobile Communications Handbook*, CRC Press and IEEE Press, New York, 1996.
14. V. S. Pless and W. C. Huffman, eds., *The Handbook of Coding Theory*, North-Holland, Amsterdam, 1998.
15. M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread-Spectrum Communications Handbook*, McGraw-Hill, New York, 1994.

TERRESTRIAL DIGITAL TELEVISION

YIYAN WU
Communications Research Centre Canada
Ottawa, Ontario, Canada

1. INTRODUCTION

A *terrestrial digital television* (TDT) system broadcasts video, audio, and ancillary data services over the VHF and UHF bands (50–800 MHz). It is a wideband digital point-to-multipoint transmission system — a high-speed digital pipe for multimedia services to the general public. The payload data rate for a TDT system is between 19 and 25 Mbps (megabits per second). The TDT will eventually replace existing analog television services, such as NTSC, PAL, and SECAM.

The first commercial TDT services were introduced in the United States and in the United Kingdom in November 1998. Since then, Australia, Sweden, Spain, Singapore, and Korea have also started TDT broadcast services. Many countries are having field trials and pilot projects of TDT systems and are planning to launch commercial service in the near future. Existing analog TV services are expected to be phased out by 2010–2020.

The bandwidth of a TDT system can either be 6, 7, or 8 MHz, depending on the countries and regions. Generally, the TDT system bandwidth is identical to the bandwidth of the analog television system that it will replace in any particular country.

The advantages of a TDT system in comparison to an analog television system are typically

1. *Better Picture Quality*. A TDT system can deliver high-definition television (HDTV) pictures at about twice the vertical resolution of an analog television system with the same bandwidth. It can also provide wide-screen formatted pictures with an aspect ratio of 16 : 9 rather than the traditional near-square, 4 : 3 television format.
2. *Better Audio Quality*. A TDT system can deliver CD-quality stereo or multichannel audio (5.1 channels). It can also carry multiple audio channels for multilanguage implementation.
3. *Ghost-Free and Distortion-Free Pictures*. There are no ghosts or any other forms of transmission distortions (e.g., co- and adjacent-channel interference, tone interference, impulse noise) visible on the screen.
4. *High Spectrum and Power Efficiency*. Each TDT channel can carry up to eight analog TV quality programs. In addition, the TDT system requires much less transmission power for the same coverage.
5. *User-Friendly Interface*. The TDT system provides an electronics program guide (EPG) for easy program and channel selection. Time-shifted viewing can easily be implemented by preselecting programs for recording.
6. *Easy Interface with Other Media*. Since a TDT system is a high-speed "digital pipe," it can have seamless interfaces with other communication systems and computer networks. Nonlinear editing, lossless storage, and access to video databases can easily be implemented.
7. *Multimedia Services and Data Broadcasting*. As a fully digital system, the TDT system can be used to deliver multimedia services and to provide data broadcasting service. Interactive services will also be available.

8. *Conditional Access*. In a TDT system, it is possible to address each receiver for content control or for value-added services.
9. *Providing Fixed, Mobile, or Portable Services*. Based on reception conditions, the TDT system can provide high-data-rate service for fixed reception, or intermediate-data-rate service for portable reception, or lower-data-rate service to automobile-mounted high-speed mobile reception.
10. *Single-Frequency Network* (SFN) *or Diversified Transmission*. Because of the strong multipath immunity of the TDT system, it is possible to operate a series of TDT transmitters on the same RF frequency to provide better coverage and service quality, and to achieve better spectrum efficiency.

The disadvantages of a TDT system in comparison to an analog television system are

1. *Sudden Service Dropout*. This is the "cliff effect." As a digital transmission system, the TDT system video and audio signals fail abruptly when the signal-to-noise ratio (SNR) drops below a critical operating threshold. Analog TV systems, on the other hand, have graceful degradation of signal power versus picture quality and the audio is extremely robust.
2. *Longer Acquisition Time*. For an analog TV system, signal acquisition is almost instantaneous. A TDT system typically takes about half a second (0.5 s) to acquire a signal. This can be bothersome when changing channels or channel surfing. The delay is mostly caused by the video decoding system. Synchronization, channel estimation, equalization, channel decoding, and conditional access also contribute to the delay.
3. *Frame Rate*. The TDT system uses the same frame rate as the analog TV system: 25 or 30 frames per second. For high-intensity pictures, the screen can show an annoying flicking, especially for large-screen-display devices.

One of the most challenging issues of introducing the TDT is that the TDT service has to coexist for some time in the same spectrum band with the existing analog TV service in a frequency-interlaced approach. Spectrum is a valuable and limited resource. In most countries and regions, there is no additional spectrum in the VHF/UHF bands available for TDT implementation. The TDT system has to use "taboo channels" [1], those channels that are unusable for broadcasting analog TV services, because of the existence of intermodulation products and other forms of interference, such as co-channel and adjacent channel interference. In other words, a TDT system has to be robust to various types of interference. The analog TV system is quite sensitive to interference (even if the interference level is 50 dB below the analog TV signal, viewers can still notice it). The service quality of the analog TV should not suffer from much degradation due to the introduction of the TDT service. Therefore, the TDT system has to transmit at low power, or, in other words, to operate at low SNR, which means that strong channel coding and forward error correction (FEC) techniques have to be implemented.

Video compression, or video source coding, is another challenge. The raw data rate of a HDTV program (without data compression) is about 1 Gbps. The TDT system data throughput is only 18–22 Mbps. Therefore, the video source coding system has to achieve at least a 50 : 1 data compression ratio, in real time, and still provide a high video quality.

2. SYSTEM DESCRIPTION

As shown in Fig. 1, a TDT system typically consists of source coding, multiplexing, and transmission systems. The video and audio source coding are implemented to compress the video and audio source data to reduce the data rate. The compressed data are multiplexed with the program-related data, such as the electronics program guide (EPG), the Program and System Information Protocol (PSIP) or services information (SI) data, program rating data, closed captioning data, and conditional access data, as well as other data that are not related to the television programs, such as opportunistic data and bandwidth-guaranteed data. The multiplexed output, or transport stream (TS), is channel coded and modulated for transmission over a terrestrial RF channel.

At the receiving end, the signal undergoes demodulation and FEC to correct the transmission errors. The resulting data are demultiplexed to sort out video output data, audio output data, and other program-related and non–program-related data.

2.1. Transmission System

2.1.1. Transmission System Description

2.1.1.1. Channel Coding. Figure 2 is a diagram for a TDT transmission system. As mentioned earlier, the combination of low power emission and robustness against interference requires that powerful channel coding must be implemented to reduce the TDT system SNR threshold. This is achieved in practice by the use of concatenated channel coding. In such a coding system, two levels of FEC codes are employed: an "inner" modulation or convolution code and an "outer" symbol error correction code. Bandwidth-efficient trellis-coded modulation (TCM) or convolutional coding [2] is implemented as the inner code to achieve high coding gain over the additive white Gaussian noise (AWGN) channel and to correct short bursts of interference, such as those created by analog TV synchronization pulses. This code has good noise-error performance at low SNR. The required output BER for the inner coder is on the order of 10^{-3} to 10^{-4}. The Reed–Solomon (RS) code [3] is used as the outer code to handle the burst of errors generated by the inner code and to provide a system BER of 10^{-11} or lower. The merit of a concatenated coding system is that it can achieve a very low SNR threshold. The drawback is that it has a "cliff" threshold, meaning that the signal dropout is within 1 dB decrease of SNR near the threshold.

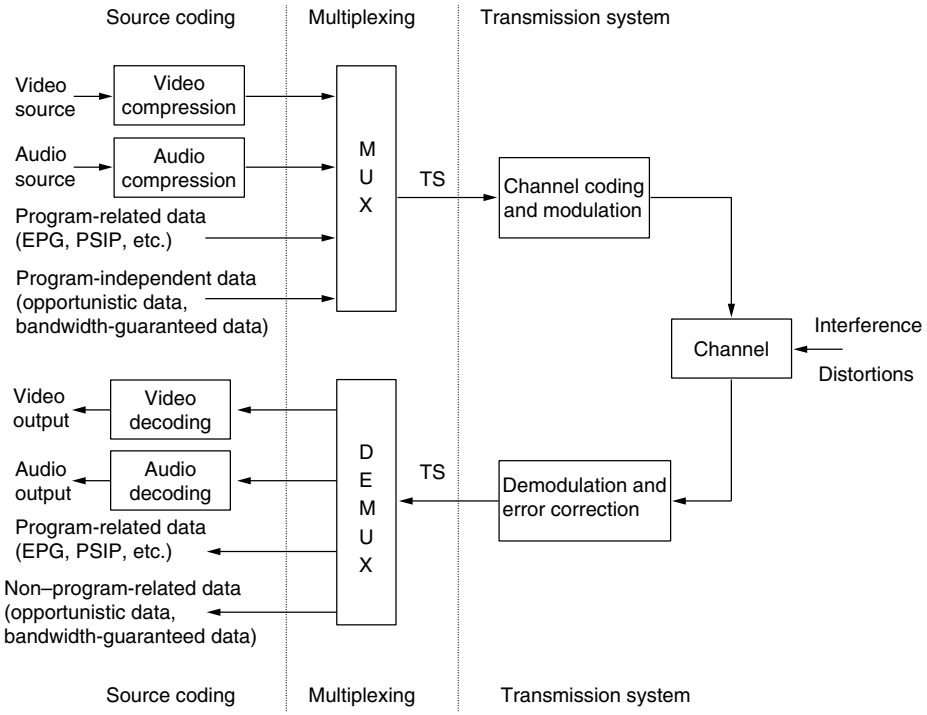


Figure 1. Terrestrial digital television system diagram.

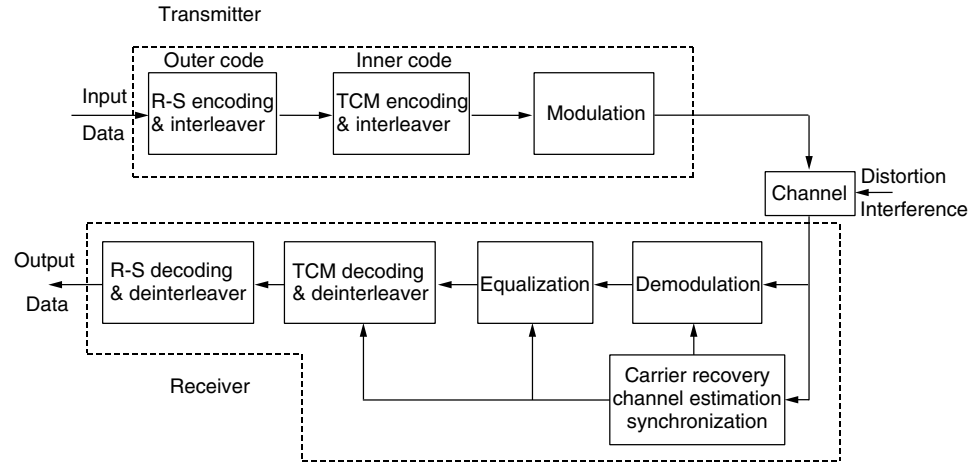


Figure 2. Terrestrial digital television transmission system diagram.

In Fig. 2, an interleaver and deinterleaver are used to fully exploit the error correction ability of the FEC code [3]. Since most errors occur in bursts (this is especially true for the outer code), they often exceed the error correction capability of the FEC code. Interleaving is used to spread out, or decorrelate, the burst of errors into shorter error sequences or isolated errors that are within the capability of the FEC code.

2.1.1.2. Modulation. From Fig. 2, the input data, or transport stream data, are coded and interleaved by the outer and inner error correction codes and, then, modulated and frequency shifted to a RF channel for transmission.

There are two types of modulation techniques used in the TDT systems; single-carrier modulation (SCM), such as vestigial sideband (VSB) [4] modulation, and multicarrier modulation (MCM), such as orthogonal frequency-division multiplexing (OFDM) [5,6] modulation.

In a SCM system, the information-bearing data are used to modulate one carrier, which occupies the entire RF channel. VSB is a one-dimensional modulation scheme. Its symbol rate is about twice the usable bandwidth. The information bits are transmitted in a vestigial sideband, occupying a bandwidth that is slightly larger than half the symbol rate. VSB modulation is mathematically equivalent to the offset quadrature amplitude modulation (OQAM).

Adaptive equalization is implemented in SCM systems to combat the intersymbol interference (ISI) that is usually caused by multipath distortion [4]. A raised-cosine filter is generally used for spectrum shaping.

In a MCM system, quadrature amplitude modulation (QAM) [4] is used to modulate multiple low-data-rate carriers, which are transmitted concurrently using frequency-division multiplexing (FDM) [4]. Since each QAM carrier spectrum envelope is of the form $\sin(x)/x$, that is, with periodic spectrum zero-crossing points, the QAM carrier spacing can be carefully selected so that each carrier spectrum peak is located on all the other carriers spectrum zero-crossing points. Although the carrier spectra overlap, they do not interfere with each other and the information bits they carry can be demodulated independently without intercarrier interference. In other words, all carriers maintain orthogonality in a FDM fashion. Hence the name OFDM. Spectrum shaping is not needed for OFDM modulation.

The OFDM can be efficiently implemented via the digital fast Fourier transform (FFT), where an inverse FFT is used as the OFDM modulator and a FFT as the demodulator [5,6]. Each inverse FFT output block is called an OFDM symbol, which contains multiple data samples. The FFT sizes used in the TDT systems are 2048, 4096, and 8192. One important feature of the OFDM is the use of a guard interval (GI). By inserting a GI between the OFDM symbols, or FFT blocks, using "cyclic extension" [5,6], the intersymbol interference can be eliminated. However, the amplitude and phase distortion, or fading, within an OFDM symbol still exists. Channel coding and equalization are used to mitigate the fading channel. Channel coding combined with the OFDM is called coded OFDM or COFDM.

2.1.2. Terrestrial Digital Television Transmission Standard. Currently, there are three TDT transmission standards [7]:

1. The Advanced Television Systems Committee (ATSC) system [8] standardized by the ATSC, a United States–based digital television standards body
2. The Digital Video Broadcasting — Terrestrial (DVB-T) standard [9] developed by the DVB project, a European-based digital broadcasting standard body, and standardized by the European Telecommunication Standard Institution (ETSI)
3. The Integrated Service Digital Broadcasting — Terrestrial (ISDB-T) standard [10] developed and standardized by the Association of Radio Industries and Businesses (ARIB) in Japan.

The main characteristics of the three TDT systems are summarized in Table 1. All three systems employ concatenated channel coding. All three standards can be

Table 1. Main Characteristics of Three Terrestrial Digital Television Systems

Systems	ATSC 8-VSB	DVB-T COFDM	ISDB-T BST-OFDM
Source coding			
Video	Main profile syntax of ISO/IEC 13818-2 (MPEG-2 — video)		
Audio	ATSC Standard A/52 (Dolby AC-3)	ISO/IEC 13818-3 (MPEG-2 — layer II audio) and Dolby AC-3	ISO/IEC 13818-7 (MPEG-2 — AAC audio)
Transport stream	ISO/IEC 13818-1 (MPEG-2 TS) transport stream		
Transmission system			
Channel coding			
Outer coding	RS (207, 187, $t = 10$)	RS (204, 188, $t = 8$)	
Outer interleaver	52 RS block interleaver	12 RS block interleaver	
Inner coding	Rate $\frac{2}{3}$ trellis code	Punctured convolutional code — rate: $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, $\frac{5}{6}$, $\frac{7}{8}$ Constraint length = 7, Polynomials (octal) = 171, 133	
Inner interleaver	12 : 1 trellis-code interleaving	Bitwise interleaving and frequency interleaving	Bitwise interleaving, frequency interleaving, and selectable time interleaving
Data randomization	16-bit PRBS	16-bit PRBS	16-bit PRBS
Modulation	8-VSB	COFDM Subcarrier modulation: QPSK, 16 QAM and 64 QAM hierarchical modulation: multiresolution constellation (16 QAM and 64 QAM) Guard intervals: $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$ of OFDM symbol 2 modes: 2K, 8K FFT	BST-OFDM with 13 frequency segments Subcarrier modulation: DQPSK, QPSK, 16 QAM, 64 QAM Hierarchical transmission: choice of three different subcarrier modulations on each segment Guard intervals: $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$ of OFDM symbol 3 modes: 2K, 4K, 8K FFT

scaled to any RF channel bandwidth (6, 7, or 8 MHz) with corresponding scaling in the data capacity.

It is believed that China is developing another TDT transmission system, which will be finalized in the 2002/2003 timeframe.

2.1.2.1. The ATSC 8-VSB System. The ATSC system [8] uses trellis-coded 8-level vestigial sideband (8-VSB) modulation with the RS (Reed–Solomon) (207, 187) [3] as the outer code and Ungerboeck rate-$\frac{2}{3}$ TCM [2] as the inner code. A raised-cosine filter with a roll-off factor of 11.5% is used for spectrum shaping.

The ATSC system was designed to transmit high-quality video and audio (HDTV) and ancillary data over a 6-MHz channel in the VHF/UHF band. The data throughput is 19.4 Mbps. The system was designed to eventually replace the existing analog TV service in the same frequency band. Therefore, one of the system requirements was to allow the allocation of an additional digital transmitter with equivalent coverage for each existing analog TV transmitter. Another requirement was to cause minimum disturbance to the existing analog TV service in terms of service area and population.

Various picture qualities can be achieved using one of 18 possible video formats (standard definition or high-definition pictures, progressive or interlaced scan format, as well as different frame rates and aspect ratios). The system can accommodate fixed and possibly portable reception.

The system is designed to withstand different types of interference: existing analog TV services, white noise, impulse noise, phase noise, continuous-wave, and multipath distortions. The system is also designed to offer spectrum efficiency and ease of frequency planning.

The main characteristics of the ATSC 8-VSB system are listed in Table 1. It should be mentioned that currently there is an ongoing 8-VSB enhancement project that will enable the ATSC system to accommodate different transmission modes (2-VSB, 4-VSB, and 8-VSB modulation) in order to offer a tradeoff between data rates and system robustness. It could also transmit dual datastreams (a robust datastream and a high-speed datastream time-division multiplexed, i.e., mixed-mode operation with different mixed ratios) within one RF channel. The project is to be completed by mid 2002.

2.1.2.2. DVB-T COFDM System. The DVB-T system [9] uses the coded orthogonal frequency-division multiplexing (COFDM) modulation system with the RS (204, 188) [3] as the outer code and a punctured convolutional code (rates: $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, $\frac{5}{6}$, $\frac{7}{8}$; constraint length = 7; polynomials (octal) = 171, 133) as the inner code. Different QAM modulations (QPSK, 16 QAM, and 64 QAM) can be implemented on the OFDM carriers. The system was designed to operate within the existing UHF spectrum allocated to analog television transmission. The payload data rates range between 4 and 32 Mbps, depending on the choice of channel coding parameters, modulation type, guard interval duration, and channel bandwidth. DVB-T can also accommodate a large range of SNR and different types of channels. It allows fixed, portable, or mobile reception, with a consequential trade-off in the usable bit rate.

The OFDM system has two operational modes: a "2K mode," which uses a 2048-point FFT, and an "8K mode," which requires an 8192-point FFT. The system makes provisions for selection between different levels of QAM modulation and different inner code rates and also allows two-level hierarchical channel coding and modulation. Moreover, a guard interval with selectable length separates the transmitted symbols, which makes the system robust to multipath distortion and allows the system to support different network configurations, such as large area SFNs and single transmitter operation. The "2K mode" is suitable for single transmitter operation and for small-scale SFN networks. The "8K mode" can be used both for single transmitter operation and for small and large SFN networks.

The main characteristics of the DVB-T COFDM system are listed in Table 1.

2.1.2.3. ISDB-T BST-OFDM. The ISDB-T system [10] uses the band-segmented transmission (BST)-OFDM modulation system. It uses the same channel coding as the DVB-T system, namely, the RS (204, 188) [3] as the outer code and a punctured convolutional code [rate $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, $\frac{5}{6}$, $\frac{7}{8}$; constraint length 7; polynomials (octal) = 171, 133] as the inner code. It also uses a large time-interleaver ($\leq$0.5 s) to deal with signal fading in mobile reception and to mitigate impulse noise interference.

The ISDB-T system is intended to deliver digital television, sound programs, and offer multimedia services to fixed, portable, and mobile terminals in the VHF and UHF bands. To meet different service requirements, the ISDB-T system provides a range of modulation and error protection schemes. The payload data rate ranges between 3.7 and 31 Mbps; again, with a consequential tradeoff in the robustness of the transmission.

The system uses a modulation method referred to as *band-segmented transmission* (BST) OFDM, which consists of a set of common basic frequency blocks called BST segments. Each segment has a bandwidth corresponding to $\frac{1}{14}$th of the terrestrial television channel bandwidth. Thirteen segments are used for data transmission within one terrestrial television channel and one segment is utilised as guard band.

The BST-OFDM modulation provides hierarchical transmission capabilities by using different carrier modulation schemes and coding rates of the inner code on different BST segments. Each data segment can have its own error protection scheme (coding rates of inner code, depth of the time interleaving) and type of modulation (QPSK, DQPSK, 16 QAM, or 64 QAM). Each segment can then meet different service requirements. A number of segments may be combined flexibly to provide a wideband service (e.g., HDTV). By transmitting OFDM segment groups with different transmission parameters, hierarchical transmission is achieved. Up to three service layers (three different segment groups) can be provided in one terrestrial channel. Partial reception of services contained in the transmission channel can be obtained using a narrowband receiver that has a bandwidth as low as one OFDM segment.

Table 2. Video Formats

Vertical Scanlines	Horizontal Samples per Line	Picture Aspect Ratio	Frame Rate	Scan Format
		a. ATSC System		
1080p	1920	16:9	24, 29.97	Progressive
1080i	1920	16:9	29.97	Interlaced
720p	1280	16:9	24, 29.97, 59.94	Progressive
480p	704	4:3, 16:9	24, 29.97, 59.94	Progressive
480i	704	4:3, 16:9	29.97	Interlaced
480p	640	4:3	24, 29.97, 59.94	Progressive
480i	640	4:3	29.97	Interlaced
		b. ISDB-T System		
1080i	1920	16:9	29.97	Interlaced
1080i	1440	16:9	29.97	Interlaced
720p	1280	16:9	59.94	Progressive
480p	720	16:9	59.94	Progressive
480i	720	16:9, 4:3	29.97	Interlaced
480i	544	16:9, 4:3	29.97	Interlaced
480i	480	16:9, 4:3	29.97	Interlaced

The main characteristics of the ISDB-T BST-OFDM system are listed in Table 1.

2.2. Source Coding

From Fig. 1, two types of source coding are implemented in a TDT system: video coding and audio coding.

2.2.1. Video Coding and Video Formats. All TDT systems adopted the MPEG-2, or ISO/IEC 13818-2, video compression standard [11]. ISO/IEC 13818-2 is a discrete cosine transform (DCT) based, motion-compensated interframe coding. It was developed by the Motion Picture Experts Group (MPEG). The standard supports a wide range of picture qualities, data rates, and video formats for broadcast and multimedia applications (see Table 2).

Different video formats are used by the TDT systems from different parts of the world. In Europe, TDT is implemented to broadcast multiple Standard Definition Television (SDTV) signals over a TDT RF channel, which is called a "multiplexer." The video format is the same as the analog TV system in Europe, specifically, 720 pixels per scanline, 576 interlaced active scanlines per frame, and 50 frames per second, or a 50-Hz $720 \times 576i$ format. The aspect ratios are 4:3 and 16:9. The letter "i" indicates interlaced scanning, a scheme which displays every second scanline and then fills in the gaps in the next pass. "Interlacing" is a scanning format used in the analog TV system to reduce the TV signal bandwidth by one-half [1]. The impact of using interlaced scanning is a slight reduction of vertical resolution.

In North America, the TDT system was designed to accommodate both HDTV and SDTV formats. Table 2, part (a) lists the 18 video formats supported by the ATSC standard [8]. Although this table is not mandated by the U.S. Federal Communications Commission (FCC), it is supported by all consumer receiver manufacturers as a voluntary industry standard.

In Japan, the TDT system was also planned to provide the HDTV and SDTV services. Table 2, part (b) shows the available video formats.

Currently, two HDTV formats are used in TDT broadcasting: 30 Hz — 1920×1080i and 60 Hz — $1280 \times$ 720p, where "p" stands for the progressive (noninterlaced) scan format, which is widely used on computer terminals. For comparable picture quality, the progressive scanning format has twice the frame rate but requires fewer scanlines. Both HDTV formats use a 16:9 aspect ratio.

2.2.2. Audio Coding. Three audio coding standards are implemented in the TDT systems:

1. MPEG layer II audio coding
2. Dolby AC-3 multichannel audio system
3. MPEG-2 advanced audio coding (AAC)

The MPEG audio coding was developed by the Motion Picture Experts Group (MPEG). It can further be classified into MPEG-1 audio coding, or ISO/IEC 11172-3 [12]; and MPEG-2 audio coding, or ISO/IEC 13818-3 and ISO/IEC 13818-7 [11].

The MPEG-1 audio coding has three layers: I, II, and III. Layers I and II are based on subband coding using a 32-subband filterbank. The coded information includes scale factors per band, the bit allocation for each band, and the quantized values for each band. The MPEG-1 layer III is based on the same subband filterbank followed by a modified discrete cosine transform (MDCT) stage that yields a filterbank with 576 bands. The quantized values are Huffman-coded to obtain greater efficiency. The MPEG-1 layer III is often called MP3 for downloading compressed music files via the Internet.

The MPEG-2 audio (ISO/IEC 13818-3 [11]) specifications extend the MPEG-1 layer II coders to lower sample rates and to multichannel audio.

Dolby AC-3 [13,14] is based on an MDCT filterbank with 256 bands. A differentially coded representation of the spectral envelope is transmitted along with the quantized values. The bit allocation is derived from the spectral envelope, identically by the encoder and decoder, with the encoder having the ability to adjust to the psychoacoustic model. The Dolby AC-3 multichannel audio system can provide the 5.1 channels (front right, front left, front center, rear right, rear left, and subwoofer) that have been widely used in the film industry.

The AAC is also part of the MPEG-2 audio, i.e., ISO/IEC 13818-7 [11]. It is based on an MDCT filterbank with 1024 bands, and uses Huffman coding for the quantized values. The AAC can also provide multichannel audio (5.1 channels) service.

The coding complexity increases from the MPEG-1 layers I, II, and III, to AC-3 and to AAC. For the same audio quality, more complicated algorithms provide higher compression ratios, or lower bit rates [16]. But they are more vulnerable to transmission errors.

Dolby AC-3 is part of the ATSC DTV standard [13] and has also been adopted by the DVB-T standard [15]. The DVB-T system deployed in Europe implemented MPEG-1 audio layer II to deliver high-quality stereo service. The ISDB-T system adopted AAC as its audio standard.

2.3. Transport Layer

2.3.1. Transport Stream. All TDT systems implemented the MPEG-2 transport system specified in the ISO/IEC 13818-1 standard [11] for data multiplexing. The MPEG-2 transport layer is designed for broadcast and multimedia applications. It has a large packet size and a small amount of overhead to achieve high spectrum efficiency. The MPEG-2 transport packet stream can carry multiplexed compressed video, compressed audio, and data packets.

The MPEG-2 transport stream (TS) packet size is 188 bytes, as shown in Fig. 3. The first byte is the synchronization byte. The next 3 bytes are fixed header bytes for error handling, encryption control, picture identification (PID), priority, and so on. The other 184 bytes are for the adaptation header and payload. The adaptation header is of variable length up to 184 bytes. It is for time synchronization, random-access flagging splicing indication, and so forth.

2.3.2. Program and System Information Protocol and Service Information. The Program and System Information Protocol (PSIP) and service information (SI) are data that are transmitted along with a station's TDT signal to provide TDT receivers important information about the television station and what is being broadcast [17]. In the ATSC system, these data are called PSIP [18]. In the DVB-T and the ISDB-T systems, it is called SI [19,20]. The most important function of PSIP and SI is to provide a method for TDT receivers to identify a TDT station and to determine how a receiver can tune to it. The PSIP and SI also tell the receiver whether multiple program channels are being broadcast and how to find each channel; identify whether the program is closed captioned; and convey program rating information and other data associated with the program. If the TV station inserted the wrong PSIP or SI, the receivers might not be able to find and decode the TDT signal.

4. SERVICES AND COVERAGE

For analog TV coverage planning, the reference receiver setup assumed the use of a directional antenna at a 10-m height at the edge of the coverage area. This means very high antenna gain and directivity (a narrow beamwidth). The 10-m antenna height likely provides a line-of-sight (LoS) path to the transmitter, with resulting high signal strength. This receiver setup is still used for TDT system coverage prediction for fixed-antenna outdoor reception of full data rate transmission. Table 3 [7] provides TDT protection ratios for frequency planning used by different countries and regions.

Since the early 1950s, major population centers have expanded substantially and there have been many high-rise buildings and other human-made structures erected throughout urban and suburban areas. In North America, some community bylaws even prohibited the use of outdoor antennas. The widespread use of electrical appliances and high-voltage power lines has substantially increased the level of impulsive noise (especially in the VHF band). All these result in much worse reception conditions, namely, the loss of LoS to the transmitter and the increase of the interference and noise levels. During the analog TV–DTV transition period, the TDT transmission power is further

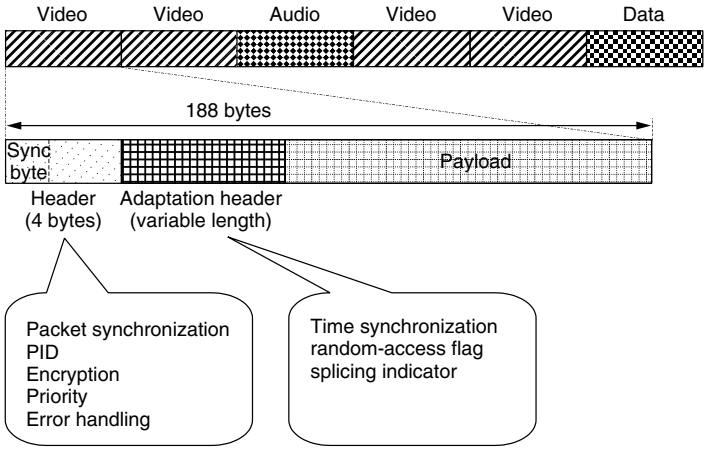


Figure 3. MPEG-2 transport stream.

Table 3. Digital Television Protection Ratios (in dB) for Frequency Planning

System Parameters (Protection Ratios)	Canada	USA	EBU[a]	Japan[a]
SNR for AWGN channel	+19.5	+15.19	+19.3	+20.1
Cochannel DTV into analog TV	+33.8	+34.44	+34∼37	+39
Cochannel analog TV into DTV	+7.2	+1.81	+4	+5
Cochannel DTV into DTV	+19.5	+15.27	+19	+21
Lower adjacent channel DTV into analog TV	−16	−17.43	$-5 \sim -11^b$	−6.0
Upper adjacent channel DTV into analog TV	−12	−11.95	$-1 \sim -10^b$	−6.0
Lower adjacent channel analog TV into DTV	−48	−47.33	$-34 \sim -37^b$	−31
Upper adjacent channel analog TV into DTV	−49	−48.71	$-38 \sim -36^b$	−33
Lower adjacent channel DTV into DTV	−27	−28	−30	−26
Upper adjacent channel DTV into DTV	−27	−26	−30	−27

[a] DVB-T (8 MHz, 64 QAM, $R = \frac{2}{3}$). ISDB-T (6 MHz, 64 QAM, $R = \frac{3}{4}$), Analog TV (M/NTSC).
[b] Depending on analog TV systems used.

limited to prevent interference into the existing analog TV services. At the same time, DTV has to withstand interference from the analog TV.

For fixed services, TDT has to compete with cable, satellite, MMDS/LMDS, and other communication systems. On the other hand, there is an increasing demand for indoor and mobile television and data services.

4.1. Indoor Reception

Indoor reception presents particularly difficult conditions, because of the lower signal strength due to building penetration losses, which can be as high as 25 dB for the VHF/UHF signals. Indoor signals also suffer from strong static and dynamic multipath distortion, due to reflections from indoor walls, as well as from outdoor structures. The lack of LoS path often results in strong pre-echoes. Nearby traffic and the movement of human bodies or even pets can significantly alter the distribution of indoor signals, causing time-varying echoes and field strength variations.

The indoor signal strength and its distribution are related to many factors, such as building structure (concrete, brick, wood), siding material (aluminum, plastic, wood), insulation material (with or without metal coating), and window material (tinted and metal-coated glasses, multilayer glass).

The indoor settop antenna gain and directivity depend very much on frequency and location. For "rabbit ear" antennas, the measured gain varied from about −10 to −4 dBi. For 5-element logarithmic antennas, the gains are between −15 and +3 dBi. The low height of indoor antennas also results in lower signal strength. Meanwhile, indoor environments sometime experience high levels of impulse noise from power lines and home appliances.

The low receiver antenna height, low antenna gain, and poor building penetration mean that an additional 30–40 dB of signal power is needed for reliable indoor reception compared to outdoor reception. However, the use of the single-frequency networks (SFNs) and receiving antenna diversity can considerably improve the indoor reception. Reducing the data rate and consequently lowering the required SNR can also improve the location availability.

4.2. Mobile Reception

Mobile reception also suffers from low antenna gain and low antenna height. The difference in field strength is larger than 10 dB between a receiving antenna at 1.5 m and one at a height of 10 m. In urban areas, building blockage and shadowing also reduce the signal strength significantly.

Mobile reception requires that the receivers withstand strong and dynamic multipath distortion, Doppler effects, and signal fading. A lower data rate is used, on the order of 50–25% of the rate used for the fixed reception. The system SNR over AWGN is generally under 10 dB.

To guarantee a satisfactory service quality, the transmission system must provide the required field strength within the service area, or a high level of location availability. Single-frequency networks (SFNs) and receiving antenna diversity can definitely improve the service quality of mobile reception.

4.3. Single-Frequency Networks (SFNs) and Diversified Transmission

The SFN or a diversified transmission approach can provide stronger field strength throughout the core coverage area and, therefore, can significantly improve the service availability. The receivers have more than one transmitter from which they can receive the signal (diversity gain). They have better chances of having a strong signal path to a transmitter, which makes it more likely to achieve reliable service.

Optimizing the transmitter network (transmitter density, tower height and location, as well as the transmission power at each transmitter) can result in better coverage with lower total transmit power and better spectrum efficiency. Special measures must be taken to minimize the frequency offset among the repeaters and to flexibly address each transmitter with respect to its exact site, power, antenna height, and the insertion of specific local signal delays.

The key difference between a TDT and an analog TV system is that the TDT can withstand at least 20 dB of DTV–DTV cochannel interference, as shown in Table 3, while the analog TV co-channel threshold of visibility is around 50 dB (30–35 dB for CCIR grade 3). In other words, TDT is 10–30 dB more robust than analog TV, which provides more flexibility for the repeater design and siting.

One drawback of using multiple transmitters is that "active" multipath distortion can occur when coverage from transmitters overlaps. The receiver must be able

to deal with these strong echoes. Achieving frequency and time synchronization of multiple transmitters and feeding the same program source to multiple transmitter sites will increase the complexity of the transmission facilities. It could also double the Doppler effects, if the mobile terminal is leaving a reception area from one transmitter and driving toward another transmitter in the overlapping area.

5. CONCLUSION

In comparison to the analog TV services, TDT can provide much better picture and audio qualities. It is robust to multipath distortion and various forms of interference. It is spectrum-efficient and can provide better service quality. It can easily interface with computer systems, and can provide multimedia and data broadcasting services. The TDT will eventually replace all the existing analog TV services.

BIOGRAPHY

Dr. Yiyan Wu received his B.Eng. degree in 1982 from the Beijing University of Posts and Telecommunications, Beijing, China, and M.Eng. and Ph.D. degrees in electrical engineering from Carleton University, Ottawa, Canada, in 1986 and 1990, respectively. He joined Telesat Canada in 1990 as a senior satellite communication systems engineer. Since 1992, he has been a senior research scientist at the Communications Research Centre Canada, where he has been working on digital television and broadband wireless multimedia communication research and standards development. Dr. Wu is a fellow of the IEEE, a member of the editorial board of the proceedings of the IEEE, and an associate editor of the IEEE *Transactions on Broadcasting*. He has served on many international committees (ATSC, IEEE, ITU) and has been a consultant to many industry and government institutions. He was the recipient of the 1999 IEEE Consumer Electronics Society Chester Sally Paper Award and 2002 Canadian Government Federal Partners in Technology Transfer Innovator Awards for scientific achievement. He is an adjunct professor of Carleton University and the Beijing University of Posts and Telecommunications. Dr. Wu has published more than 150 scientific papers and book chapters.

BIBLIOGRAPHY

1. K. B. Benson, ed., *Television Engineering HandBook*, McGraw-Hill, New York, 2000.
2. G. Ungerboeck, Trellis coded modulation with redundant signal sets, *IEEE Commun. Mag.* **27**: 5–21 (Feb. 1987).
3. G. C. Clark and J. B. Cain, *Error-Correction Coding for Digital Communications*, Plenum Press, New York, 1981.
4. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1995.
5. Y. Wu and W. Y. Zou, Orthogonal frequency division multiplexing: A multiple-carrier modulation scheme, *IEEE Trans. Consumer Electron.* **41**(3): 392–399 (Aug. 1995).
6. S. B. Winstein and P. M. Ebert, Data transmission by frequency division multiplexing using the discrete Fourier transform, *IEEE Trans. Commun.* **19**: 628–634 (Oct. 1971).
7. Y. Wu et al., Comparison of terrestrial DTV transmission systems: The ATSC 8-VSB, the DVB-T COFDM and the ISDB-T BST-OFDM, *IEEE Trans. Broadcast.* **46**(2): 101–113 (June 2000).
8. ATSC, *ATSC Digital Television Standard*, ATSC Standard A/53, April 1, 2001, *http://www.atsc.org/*.
9. ETS 300 744, *Digital Video Broadcasting (DVB); Digital Broadcasting Systems for Television, Sound and Data Services; Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television*, ETSI Draft EN 300 744 V1.2.1, (1999-1). *http://www.dvb.org/ and http://www.etsi.org/*.
10. ARIB, *Terrestrial Integrated Services Digital Broadcasting (ISDB-T)—Specifications of Channel Coding, Framing Structure, and Modulation*, Sept. 28, 1998, *http://www.arib.or.jp/arib/english/*.
11. ISO/IEC 13818, *Information Technology—Generic Coding of Moving Pictures and Associated Audio Information, Part 1 (Systems), 2 (Video), 3 (Audio), 7 (AAC Audio)*, 1994.
12. ISO/IEC 11172, *Information Technology—Coding for Moving Pictures and Associated Audio for Digital Storage at about 1.5 Mbits, Part 1 (Systems), 2 (Video), 3 (Audio)*, 1993.
13. ATSC, *ATSC Digital Audio Compression Standard (AC/3)*, ATSC Standard A/52, December 20, 1995.
14. ITU-R, *Audio Coding for Digital Terrestrial Television Broadcasting*, ITU-R Rec. BS.1196, 1995.
15. ETSI TR 101 154, *Implementation Guidelines for the Use of MPEG-2 Systems, Video and Audio in Satellite, cable and Terrestrial broadcasting Applications*, ETSI TR 101 154 V1.4.1 (2000-07).
16. G. A. Soulodre et al., Subjective evaluation of state-of-the-art two-channel audio codecs, *J. Audio Eng. Soc.* **46**(3): 164–177 (March 1998).
17. A. Allison and K. T. Williams, *Channel Branding and Navigation for DTV: Understanding PSIP*, NAB, Washington DC, 2000.
18. ATSC, *Program and System Information Protocol*, ATSC Standard A/65A, May 2000.
19. ETS 300 468, *Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB Systems*, ETSI 300 468 ed. 2, Oct. 1996.
20. ARIB, *Service Information for Digital Broadcasting System*, ARIB STD B-10, V3.0, May 2001.

TERRESTRIAL MICROWAVE COMMUNICATIONS

DAVID R. SMITH
George Washington University
Ashburn, Virgina

1. INTRODUCTION

The basic components required for operating a radio over a microwave link are the transmitter, towers, antennas, and receiver. Transmitter functions typically include

multiplexing, encoding, modulation, upconversion from baseband or intermediate frequency (IF) to radiofrequency (RF), power amplification, and filtering for spectrum control. Antennas are placed on a tower or other tall structure at sufficient height to provide a direct, unobstructed line-of-sight (LoS) path between the transmitter and receiver sites. Receiver functions include RF filtering, downconversion from RF to IF amplification at IF equalization, demodulation, decoding, and demultiplexing.

In describing terrestrial microwave, the focus is on digital, line-of-sight, point-to-point microwave systems (~1–30 GHz). The design of millimeter-wave (30–300-GHz) radio links is also considered, however. Further, much of the material on line-of-sight propagation, including multipath and interference effects, and link design methodology also applies to the design of mobile and analog radio systems. Many of the design and performance characteristics considered here apply to satellite communications as well.

2. LINE-OF-SIGHT PROPAGATION

The modes of propagation between two radio antennas may include a direct, line-of-sight (LoS) path but also a ground or surface wave that parallels the earth's surface, a sky wave from signal components reflected off the troposphere or ionosphere, a ground reflected path, and a path diffracted from an obstacle in the terrain. The presence and utility of these modes depend on the link geometry, both distance and terrain between the two antennas, and the operating frequency. For frequencies in the microwave band, the LoS propagation mode is the predominant mode available for use; the other modes may cause interference with the stronger LoS path. Line-of-sight links are limited in distance by the curvature of the earth, obstacles along the path, and free-space loss. Average distances for conservatively designed LoS links are 25–30 mi, although distances of $\leq$100 mi have been used. The performance of the LoS path is affected by several phenomena addressed in this section, including free-space loss, terrain, atmosphere, and precipitation. The problem of fading due to multiple paths is addressed in the following section.

2.1. Free-Space Loss

Consider a radio path consisting of isotropic antennas at both transmitter and receiver. An isotropic transmitting antenna radiates its power P_t equally in all directions. In the absence of terrain or atmospheric effects (i.e., free space), the radiated power density is equal at points equidistant from the transmitter. The effective area A of a receiving antenna determines how much of the incident signal from the transmitting antenna is collected by the receiving antenna is given by

$$A = \frac{\lambda^2 G_r}{4\pi} \tag{1}$$

where λ = wavelength in meters and G_r is the ratio of the receiving antenna power gain to the corresponding gain of an isotropic radiator. The power received by the receiving antenna at a distance r in watts, P_r, is then given by

$$P_r = \frac{P_t G_t}{4\pi r^2} \cdot \frac{\lambda^2 G_r}{4\pi} \tag{2}$$

If we assume that both the transmitting and receiving antennas are isotropic radiators, then $G_t = G_r = 1$, and the ratio of received to transmitted power becomes the free-space path loss

$$l_{\text{fs}} = \frac{P_r}{P_t} = \frac{\lambda^2}{(4\pi r)^2} \tag{3}$$

The free-space path loss is then determined as the ratio of the received power to the transmitted power and is given in decibels by

$$\begin{aligned} L_{\text{fs}} &= 10\log_{10}(l_{\text{fs}}) = 10\log_{10}\left(\frac{P_r}{P_t}\right) \\ &= 32.44 + 20\log_{10}(D_{\text{km}}) + 20\log_{10}(f_{\text{MHz}})\ \text{dB} \end{aligned} \tag{4}$$

where $r = D_{\text{km}}$ is the distance in kilometers and $f_{\text{MHz}} = 300/\lambda_{\text{meters}}$ is the signal frequency in megahertz (MHz). Note that the doubling of either frequency or distance causes a 6-dB increase in path loss.

2.2. Terrain Effects

Obstacles along a LoS radio path can cause the propagated signal to be reflected or diffracted, resulting in path losses that deviate from the free space value. This effect stems from electromagnetic wave theory, which postulates that a wavefront diverges as it advances through space. A radio beam that just grazes the obstacle is diffracted, with a resulting obstruction loss whose magnitude depends on the type of surface over which the diffraction occurs. A smooth surface, such as water or flat terrain, produces the maximum obstruction loss at grazing. A sharp projection, such as a mountain peak or even trees, produces a knife-edge effect with minimum obstruction loss at grazing. Most obstacles in the radio path produce an obstruction loss somewhere between the limits of smooth earth and knife edge.

2.2.1. Reflections. When the obstacle is below the optical LoS path, the radio beam can be reflected to create a second signal at the receiving antenna. Reflected signals can be particularly strong when the reflection surface is smooth terrain or water. Since the reflected signal travels a longer path than does the direct signal, the reflected signal may arrive out of phase with the direct signal. The degree of interference at the receiving antenna from the reflected signal depends on the relative signal levels and phases of the direct and reflected signals.

At the point of reflection, the indirect signal undergoes attenuation and phase shift, which is described by the reflection coefficient R, where

$$R = \rho \exp(-j\phi) \tag{5}$$

The magnitude ρ represents the change in amplitude, and ϕ is the phase shift on reflection. The values of

ρ and ϕ depend on the wave polarization (horizontal or vertical), angle of incidence ψ dielectric constant of the reflection surface, and wavelength λ of the radio signal. The mathematical relationship has been developed elsewhere [1] and will not be covered here. For microwave frequencies, however, two general cases should be mentioned:

1. For horizontally polarized waves with small angle of incidence, $R = -1$ for all terrain, such that the reflected signal suffers no change in amplitude but has a phase change of 180°.
2. If the polarization is vertical with grazing incidence, $R = -1$ for all terrain. With increasing angle of incidence, the reflection coefficient magnitude decreases, reaching zero in the vicinity of $\psi = 10°$.

To examine the problem of interference from reflection, we first simplify the analysis by neglecting the effects of the curvature of the earth's surface. Then, when the reflection surface is flat earth, the geometry is as illustrated in Fig. 1a, with transmitter (Tx) at height h_1 and receiver (Rx) at height h_2 separated by a distance D and the angle of reflection equal to the angle of incidence ψ. Using plane geometry and algebra, the path difference δ between the reflected and direct signals can be given by

$$\delta = (r_2 + r_1) - r = \frac{2h_1h_2}{D} \tag{6}$$

The overall phase change experienced by the reflected signal relative to the direct signal is the sum of the phase difference due to the pathlength difference δ and the phase ϕ due to the reflection. The total phase shift is therefore

$$\gamma = \frac{2\pi}{\lambda}\frac{2h_1h_2}{D} + \phi \tag{7}$$

At the receiver, the direct and reflected signals combine to form a composite signal with field strength E_C. By the simple geometry of Fig. 1b and using the law of cosines, we obtain

$$E_C = E_D^2 + E_R^2 + 2E_DE_R\cos\gamma \tag{8}$$

where E_D = field strength of direct signal
E_R = field strength of reflected signal
ρ = magnitude of reflection coefficient = E_R/E_D
γ = phase difference between direct and reflected signal as given by Eg. (7)

The composite signal is at a minimum, from (8), when $\gamma = (2n+1)\pi$, where n is an integer. Similarly, the composite signal is at a maximum when $\gamma = 2n\pi$. As noted earlier, the phase shift ϕ due to reflection is usually around 180° for microwave paths since the angle of incidence on the reflection surface is typically quite small. For this case, the received signal minima, or nulls, occur when the path difference is an even multiple of a half wavelength, or

$$\delta = 2n\left(\frac{\lambda}{2}\right) \quad \text{for minima} \tag{9}$$

The maxima, or peaks, for this case occur when the path difference is an odd multiple of a half-wavelength, or

$$\delta = \frac{(2n+1)\lambda}{2} \quad \text{for maxima} \tag{10}$$

2.2.2. Fresnel Zones. The effects of reflection and diffraction on radiowaves can be more easily seen by using the model developed by Fresnel for optics. Fresnel accounted for the diffraction of light by postulating that the cross section of an optical wavefront is divided into zones of concentric circles separated by half-wavelengths. These zones alternate between constructive and destructive interference, resulting in a sequence of dark and light bands when diffracted light is viewed on a screen. When viewed in three dimensions, as necessary for determining path clearances in LoS radio systems, the Fresnel zones become concentric ellipsoids. The first Fresnel zone is that locus of points for which the sum of the distances between the transmitter and receiver and a point on the ellipsoid is exactly one half-wavelength longer than the direct path between the transmitter and receiver. The nth Fresnel zone consists of that set of points for which the difference is n half-wavelengths. The radius of the nth Fresnel zone at a given distance along the path is given by

$$F_n = 17.3\left(\frac{nd_1d_2}{fD}\right)^{1/2} \text{ meters} \tag{11}$$

where d_1 = distance from transmitter to a given point along the path (km)
d_2 = distance from receiver to the same point along the path (km)
f = frequency (GHz)
D = pathlength (km) ($D = d_1 + d_2$)

As an example, Fig. 2 shows the first three Fresnel zones for an LoS path of length (D) 40 km and frequency (f) 8 GHz. The distance h represents the clearance between the LoS path and the highest obstacle along the terrain.

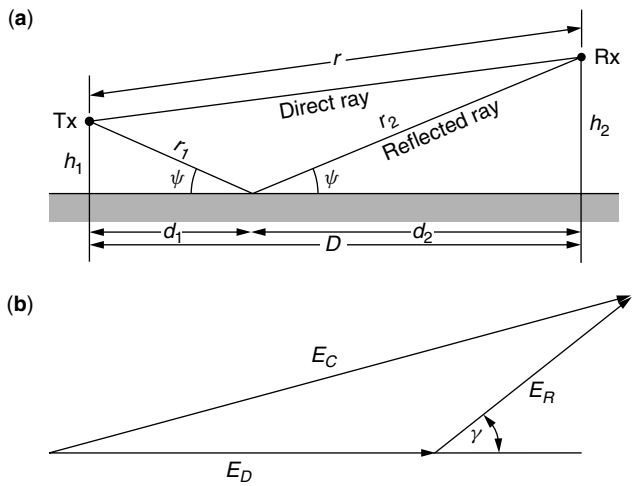


Figure 1. Geometry of two-path propagation: (**a**) flat-earth geometry; (**b**) two-ray geometry.

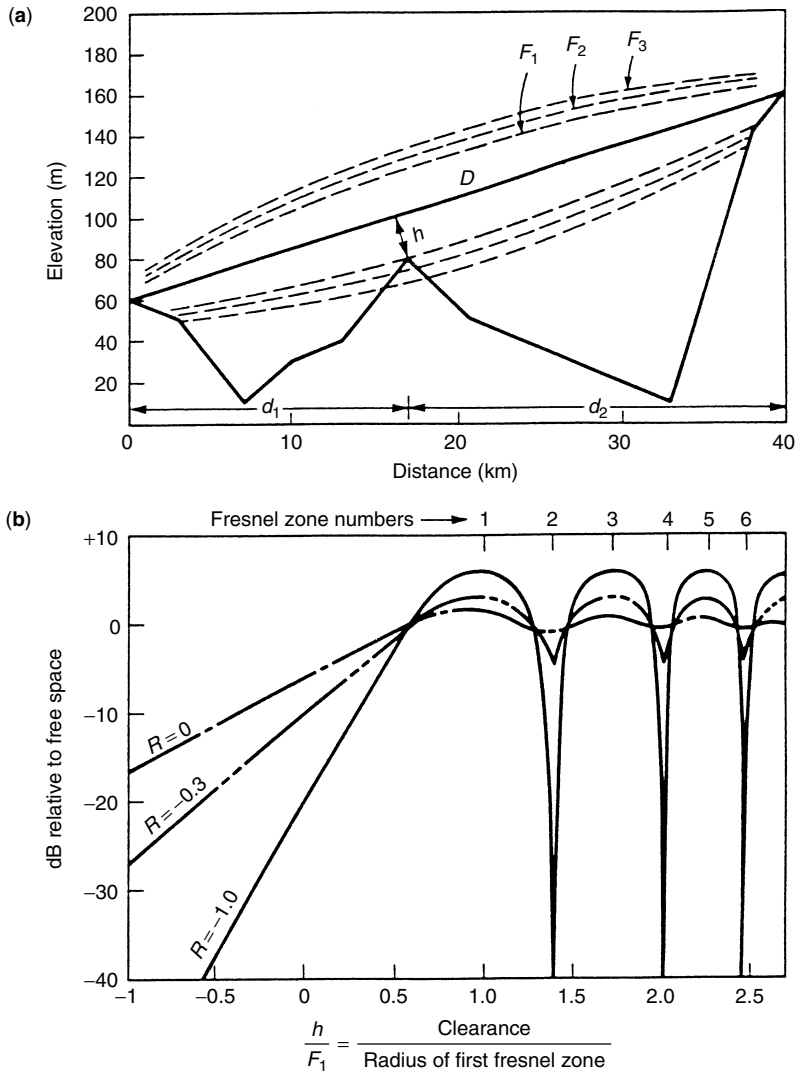


Figure 2. Fresnel zones: (**a**) Fresnel zones for an 8-GHz, 40-km LoS path; (**b**) attenuation versus path clearance.

Using Fresnel diffraction theory, we can calculate the effects of path clearance on transmission loss (see Fig. 2b). The three cases shown in Fig. 2b correspond to different reflection coefficient values as determined by differences in terrain roughness. The curve marked $R = 0$ represents the case of knife-edge diffraction, where the loss at a grazing angle (zero clearance) is equal to 6 dB. The curve marked $R = -1.0$ illustrates diffraction from a smooth surface, which produces a maximum loss equal to 20 dB at grazing. In practice, most microwave paths have been found to have a reflection coefficient magnitude of 0.2–0.4; thus the curve marked $R = -0.3$ represents the ordinary path [2]. For most paths, the signal attenuation becomes small with a clearance of 0.6 times the first Fresnel zone radius. Thus microwave paths are typically sited with a clearance of at least $0.6F_1$.

The fluctuation in signal attenuation observed in Fig. 2b is due to alternating constructive and destructive interference with increasing clearance. Clearance at odd-numbered Fresnel zones produces constructive interference since the delayed signal is in phase with the direct signal; with a reflection coefficient of −1.0, the direct and delayed signals sum to a value 6 dB higher than free-space loss. Clearance at even-numbered Fresnel zones produces destructive interference since the delayed signal is out of phase with the direct signal by a multiple of $\lambda/2$; for a reflection coefficient of −1.0, the two signals cancel each other. As indicated in Fig. 2b, the separation between adjacent peaks or nulls decreases with increasing clearance, but the difference in signal strength decreases with increasing Fresnel zone numbers.

2.3. Atmospheric Effects

Radiowaves travel in straight lines in free space, but they are bent, or refracted, when traveling through the atmosphere. Bending of radiowaves is caused by changes with altitude in the index of refraction, defined as the ratio of propagation velocity in free space to that in the medium of interest. Normally the refractive index decreases with altitude, meaning that the velocity of propagation increases with altitude, causing radiowaves to bend downward. In this case, the radio horizon is extended beyond the optical horizon.

The index of refraction n varies from a value of 1.0 for free space to approximately 1.0003 at the surface of the earth. Since this refractive index varies over such a small range, it is more convenient to use a scaled unit, N, which is called *radio refractivity* and defined as

$$N = (n - 1)10^6 \tag{12}$$

Thus N indicates the excess over unity of the refractive index, expressed in millionths. When $n = 1.0003$, for example, N has a value of 300. Owing to the rapid decrease of pressure and humidity with altitude and the slow decrease of temperature with altitude, N normally decreases with altitude and tends to zero.

To account for atmospheric refraction in path clearance calculations, it is convenient to replace the true earth radius a by an effective earth radius a_e and to replace the actual atmosphere with a uniform atmosphere in which radiowaves travel in straight lines. The ratio of effective to true earth radius is known as the k factor:

$$k = \frac{a_e}{a} \tag{13}$$

By application of Snell's law in spherical geometry, it may be shown that as long as the change in refractive index is linear with altitude, the k factor is given by

$$k = \frac{1}{1 + a(dn/dh)} \tag{14}$$

where dn/dh is the rate of change of refractive index with height. It is usually more convenient to consider the gradient of N instead of the gradient of n. Making the substitution of dN/dh for dn/dh and also entering the value of 6370 km for a into (14) yields the following:

$$k = \frac{157}{157 + (dN/dh)} \tag{15}$$

where dN/dh is the N gradient per kilometer. Under most atmospheric conditions, the gradient of N is negative and constant and has a value of approximately

$$\frac{dN}{dh} = -40 \text{ units/km} \tag{16}$$

Substituting (14) into (13) yields a value of $k = \frac{4}{3}$, which is commonly used in propagation analysis. An index of refraction that decreases uniformly with altitude resulting in $k = \frac{4}{3}$ is referred to as *standard refraction*.

2.3.1. Anomalous Propagation. Weather conditions may lead to a refractive index variation with height that differs significantly from the average value. In fact, atmospheric refraction and corresponding k factors may be negative, zero, or positive. The various forms of refraction are illustrated in Fig. 3 by presenting radio paths over both true earth and effective earth. Note that radiowaves become straight lines when drawn over the effective earth radius. Standard refraction is the average condition observed and results from a well-mixed atmosphere. The other refractive conditions illustrated in Fig. 3 — including subrefraction, superrefraction, and ducting — are observed a small percentage of the time and are collectively referred to as *anomalous propagation*.

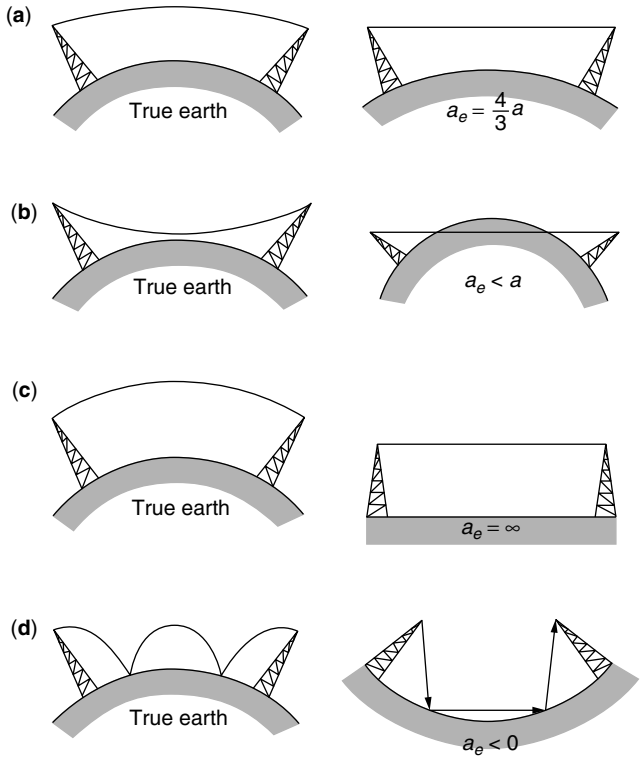


Figure 3. Various forms of atmospheric refraction; (**a**) standard refraction ($k = \frac{4}{3}$); (**b**) subrefraction ($0 < k < 1$); (**c**) superrefraction ($2 < k < \infty$); (**d**) ducting ($k < 0$).

Subrefraction ($k < 1$) leads to the phenomenon known as *inverse bending* or "earth bulge", illustrated in Fig. 3b. This condition arises because of an increase in refractive index with altitude and results in an upward bending of radiowaves. Substandard atmospheric refraction may occur with the formation of fog, as cold air passes over a warm earth, or with atmospheric stratification, as occurs at night. The effect produced is likened to the bulging of the earth into the microwave path that reduces the path clearance or obstructs the LoS path.

Superrefraction ($k > 2$) causes radiowaves to refract downward with a curvature greater than normal. The result is an increased flattening of the effective earth. For the case illustrated in Fig. 3c the effective earth radius is infinity — that is, the earth reduces to a plane. From Eq. (15) it can be seen that an N gradient of -157 units per kilometer yields a k equal to infinity. Under these conditions radiowaves are propagated at a fixed height above the earth's surface, creating unusually long propagation distances and the potential for overreach interference with other signals occupying the same frequency allocation. Superrefractive conditions arise when the index of refraction decreases more rapidly than normal with increasing altitude, which is produced by a rise in temperature with altitude, a decrease in humidity, or both. An increase in temperature with altitude, called a *temperature inversion*, occurs when the temperature of

the earth's surface is significantly less than that of the air, which is most commonly caused by cooling of the earth's surface through radiation on clear nights or by movement of warm dry air over a cooler body of water.

A more rapid decrease in refractive index gives rise to more pronounced bending of radiowaves, in which the radius of curvature of the radiowave is smaller than the earth's radius. As indicated in Fig. 3d, the rays are bent to the earth's surface and then reflected upward from it. With multiple reflections, the radiowaves can cover large ranges far beyond the normal horizon. In order for the radiowave's bending radius to be smaller than the earth's radius, the N gradient must be less than -157 units per kilometer. Then, according to (15), the k factor and effective earth radius both become negative quantities. As illustrated in Fig. 3d, the effective earth is approximated by a concave surface. This form of anomalous propagation is called *ducting* because the radio signal appears to be propagated through a waveguide, or duct. A duct may be located along or it elevated above the earth's surface. The meteorological conditions responsible for either surface or elevated ducts are similar to conditions causing superrefractivity. With ducting, however, a transition region between two differing air masses creates a trapping layer. In ducting conditions, refractivity N decreases with increasing height in an approximately linear fashion above and below the transition region, where the gradient departs from the average. In this transition region, the gradient of N becomes steep.

2.3.2. Atmospheric Absorption. For frequencies above 10 GHz, attenuation due to atmospheric absorption becomes an important factor in radio-link design. The two major atmospheric gases contributing to attenuation are water vapor and oxygen. Studies have shown that absorption peaks occur in the vicinity of 22.3 and 187 GHz due to water vapor and in the vicinity of 60 and 120 GHz for oxygen [3]. The calculation of specific attenuation produced by either oxygen or water vapor is complex, requiring computer evaluation for each value of temperature, pressure, and humidity. Formulas that approximate specific attenuation may be found in ITU-R Rep. 721-2 [4]. At millimeter wavelengths (30–300 GHz), atmospheric absorption becomes a significant problem. To obtain maximum propagation range, frequencies around the absorption peaks are to be avoided. On the other hand, certain frequency bands have relatively low attenuation. In the millimeter-wave range, the first two such bands, or windows, are centered at approximately 36 and 85 GHz.

2.3.3. Rain Attenuation. Attenuation due to rain and suspended water droplets (fog) can be a major cause of signal loss, particularly for frequencies above 10 GHz. Rain and fog cause a scattering of radiowaves that results in attenuation. Moreover, for the case of millimeter wavelengths where the raindrop size is comparable to the wavelength, absorption occurs and increases attenuation. The degree of attenuation on an LoS link is a function of (1) the point rainfall rate distribution; (2) the specific attenuation, which relates rainfall rates to point attenuations; and (3) the effective pathlength, which is multiplied by the specific attenuation to account for the length of the path. Heavy rain, as found in thunderstorms, produces significant attenuation, particularly for frequencies above 10 GHz. The point rainfall rate distribution gives the percentage of a year that the rainfall rate exceeds a specified value. Rainfall rate distributions depend on climatologic conditions and vary from one location to another. To relate rainfall rates to a particular path, measurements must be made by use of rain gauges placed along the propagation path. In the absence of specific rainfall data along a specific path, it becomes necessary to use maps of rain climate regions, such as those provided by ITU-R Rep. 563.3 [4].

To counter the effects of rain attenuation, it should first be noted that neither space nor frequency diversity is effective as protection against rainfall effects. Measures that are effective, however, include increasing the fade margin, shortening the pathlength, and using a lower-frequency band. A mathematical model for rain attenuation is found in ITU-R Rep. 721-3 [4] and summarized here. The specific attenuation is the loss per unit distance that would be observed at a given rain rate, or

$$\gamma_R = kR^{\beta} \tag{17}$$

where γ_R is the specific attenuation in dB/km and R is the point rainfall rate in mm/h. The values of k and β depend on the frequency and polarization, and may be determined from tabulated values in ITU-R Rep. 721-2. Because of the nonuniformity of rainfall rates within the cell of a storm, the attenuation on a path is not proportional to the pathlength; instead it is determined from an effective pathlength given in [5] as

$$L_{\text{eff}} = \frac{L}{1 + (R - 6.2)L/2636} \tag{18}$$

Now, to determine the attenuation for a particular probability of outage, the ITU-R method calculates the attenuation $A_{0.01}$ that occurs 0.01% of a year and uses a scaling law to calculate the attenuation A, at other probabilities

$$A_{0.01} = \gamma_R L_{\text{eff}} \tag{19}$$

where γ_R and L_{eff} are determined for the 0.01% rainfall rate. The value of the 0.01% rainfall rate is obtained from world contour maps of 0.01% rainfall rates found in ITU-R Rep. 563-4.

2.4. Path Profiles

In order to determine tower heights for suitable path clearance, a profile of the path must be plotted. The path profile is obtained from topographic maps that should have a scale of 1 : 50,000 or less. For LoS links under 70 km in length, a straight line may be drawn connecting the two endpoints. For longer links, the great circle path must be calculated and plotted on the map. The elevation contours are then read from the map and plotted on suitable graph paper, taking special note of any obstacles along the path. The path profiling process may be fully automated by use of CD-ROM technology and an appropriate computer

program. CD-ROM data storage disks are available that contain a global terrain elevation database from which profile points can be automatically retrieved, given the latitudes and longitudes of the link endpoints [12].

The path profile may be plotted on special graph paper that depicts the earth as curved and the transmitted ray as a straight line or on rectilinear graph paper that depicts the earth as flat and the transmitted ray as a curved line. The use of linear paper is preferred because it eliminates the need for special graph paper, permits the plotting of rays for different effective earth radius, and simplifies the plotting of the profile. Figure 4 is an example of a profile plotted on linear paper for a 40-km, 8-GHz radio link.

The use of rectilinear paper, as suggested, requires the calculation of the earth bulge at a number of points along the path, especially at obstacles. This calculation then accounts for the added elevation to obstacles due to curvature of the earth. Earth bulge in meters may be calculated as

$$h = \frac{d_1 d_2}{12.76} \tag{20}$$

where d_1 = distance from one end of path to point being calculated (km)
d_2 = distance from same point to other end of the path (km)

As indicated earlier, atmospheric refraction causes ray bending, which can be expressed as an effective change in earth radius by using the k factor. The effect of refraction on earth bulge can be handled by adding the k factor to the denominator in (20) or, for d_1 and d_2 in miles and h in feet

$$h = \frac{d_1 d_2}{1.5\,k} \tag{21}$$

To facilitate path profiling, Eq. (21) may he used to plot a curved ray template for a particular value of k and for use with a flat-earth profile. Alternatively, the earth bulge can be calculated and plotted at selected points that represent the clearance required below a straight line drawn between antennas; when connected together, these points form a smooth parabola whose curvature is determined by the choice of k.

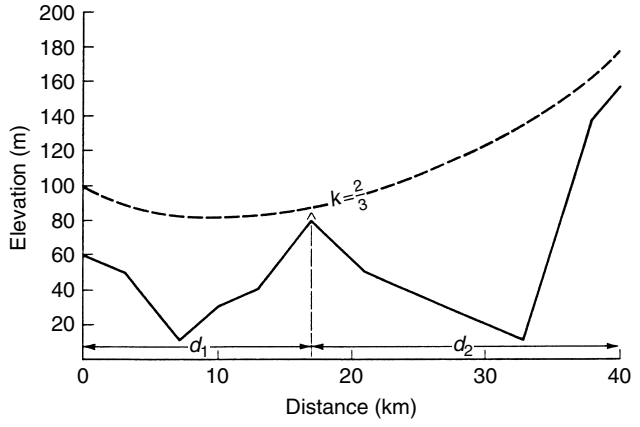


Figure 4. Example of a LoS path profile plotted on linear paper.

In path profiling, the choice of k factor is influenced by its minimum value expected over the path and the path availability requirement. With lower values of k, earth bulging becomes pronounced and antenna height must be increased to provide clearance. To determine the clearance requirements, the distribution of k values is required; it can be found by meteorological measurements [4]. This distribution of k values can be related to path availability by selecting a k whose value is exceeded for a percentage of the time equal to the availability requirement.

Apart from the k factor, the Fresnel zone clearance must be added. Desired clearance of any obstacle is expressed as a fraction, typically 0.3 or 0.6, of the first Fresnel zone radius. This additional clearance is then plotted on the path profile, shown as a small tickmark on Fig. 4, for each point being profiled. Finally, clearance should be provided for trees (nominally 15 m) and additional tree growth (nominally 3 m) or, in the absence of trees, for smaller vegetation (nominally 3 m).

The clearance criteria can thus be expressed by specific choices of k and fraction of first Fresnel zone. Here is one set of clearance criteria that is commonly used for highly reliable paths [7]:

1. Full first Fresnel zone clearance for $k = \frac{4}{3}$
2. 0.3 first Fresnel zone clearance for $k = \frac{2}{3}$

whichever is greater. Over the majority of paths, the clearance requirements of criterion 2 will be controlling. Even so, the clearance should be evaluated by using both criteria along the entire path.

Path profiles are easily obtained using digital terrain data and standard computer software. The United States Geological Survey (USGS) database called *Digital Elevation Models* (DEM) contains digitized elevation data versus latitude and longitude throughout the United States. A similar database for the world has been developed by the National Imagery and Mapping Agency (NIMA) called *Digital Terrain Elevation Data* (DTED). The USGS DEM data and DEM viewer software are available from the USGS web site.

The smallest spacing between elevation points in the USGS is 100 ft (~30 m), while larger spacings of 3 arcseconds (300 ft north–south and 230 east–west) are also available. The USGS 3-arcsecond data are provided in $1^\circ \times 1^\circ$ blocks for the United States. The 1° DEM is also referred to as DEM 250 because these data were collected from 1:250,000-scale maps. DTED level 1 data are identical to DEM 250 except for format differences. Even at the 100-ft spacing, there is a significant number and diversity of terrain features excluded from the database that could significantly impact a propagating wave at microwave radiofrequencies. These terrain databases also contain some information regarding land use/land clutter (LULC). The LULC data indicate geographic areas covered by foliage, buildings, and other surface details. The USGS data contain LULC information that could be used to estimate foliage losses in rural areas and manmade noise levels near built-up areas. In addition, they also contain other terrain features such as roadways, bodies of water,

and other information directly relevant to the wireless environment.

3. MULTIPATH FADING

Fading is defined as variation of received signal level with time due to changes in atmospheric conditions. The propagation mechanisms that cause fading include refraction, reflection, and diffraction associated with both the atmosphere and terrain along the path. The two general types of fading, referred to as *multipath* and *power fading*, are illustrated by the recordings of RF received signal levels shown in Fig. 5.

Power fading, sometimes called *attenuation fading*, results mainly from anomalous propagation conditions, such as (see Fig. 3) *subrefraction* $k < 1$), which causes blockage of the path due to the effective increase in earth bulge; *superrefraction* ($k > 2$), which causes pronounced ray bending and decoupling of the signal from the receiving antenna; and *ducting* ($k < 0$), in which the radio beam is trapped by atmospheric layering and directed away from the receiving antenna. Rainfall also contributes to power fading, particularly for frequencies above 10 GHz. Power fading is characterized as slowly varying in time, usually independent of frequency, and causing long periods of outages. Remedies include greater antenna heights for subrefractive conditions, antenna realignment for superrefractive conditions, and added link margin for rainfall attenuation.

Multipath fading arises from destructive interference between the direct ray and one or more reflected or refracted rays. These multiple paths are of different lengths and have varied phase angles on arrival at the receiving antenna. These various components sum to produce a rapidly varying, frequency-selective form of fading. Deep fades occur when the primary and secondary rays are equal in amplitude but opposite in phase, resulting in signal cancellation and a deep amplitude null. Between deep fades, small amplitude fluctuations are observed that are known as *scintillation*; these fluctuations are due to weak secondary rays interfering with a strong direct ray.

Multipath fading is observed during periods of atmospheric stratification, where layers exist with different refractive gradients. The most common meteorological cause of layering is a temperature inversion, which commonly occurs in hot, humid, still, windless conditions, especially in late evening, at night, and in early morning. Since these conditions arise during the summer, multipath fading is worst during the summer season. Multipath fading can also be caused by reflections from flat terrain or a body of water. Hence multipath fading conditions are most likely to occur during periods of stable atmosphere and for highly reflective paths. Multipath fading is thus a function of pathlength, frequency, climate, and terrain. Techniques used to deal with multipath fading include the use of diversity, increased fade margin, and adaptive equalization.

3.1. Statistical Properties of Fading

The random nature of multipath fading suggests a statistical approach to its characterization. The statistical parameters commonly used in describing fading are

- Probability (or percentage of time) that the LoS link is experiencing a fade below threshold
- Average fade duration and probability of fade duration greater than a given time
- Expected number of fades per unit time

The terms to be used are defined in graph form in Fig. 6. The threshold L is the signal level corresponding to the minimum acceptable signal-to-noise ratio or, for digital transmission, the maximum acceptable probability of error. The difference between the normal received signal level and threshold is the fade margin. A *fade* is defined as the downward crossing of the received signal through

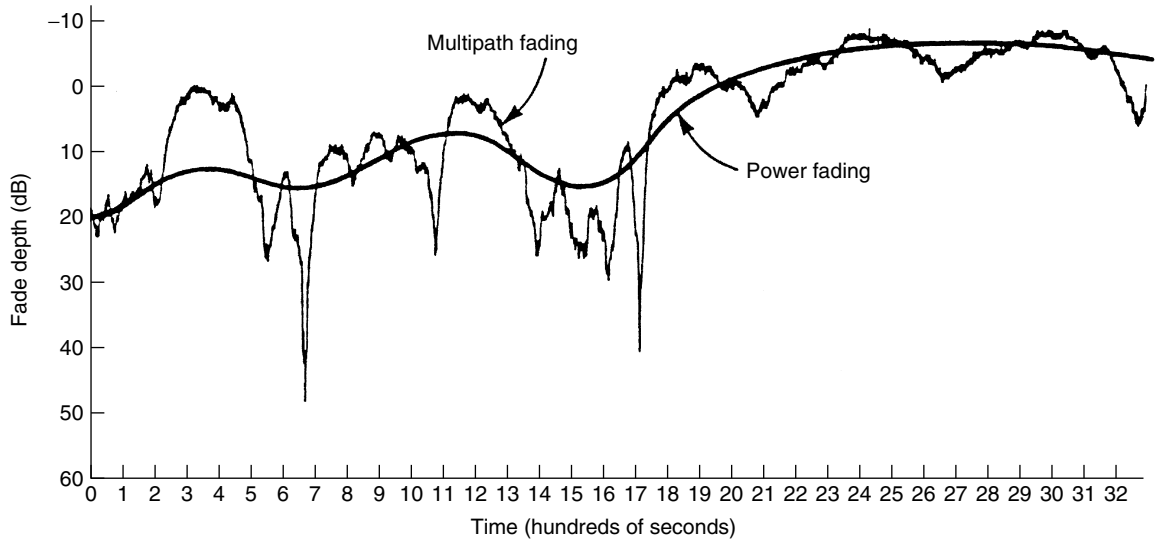


Figure 5. Example of multipath and power fading for LoS link.

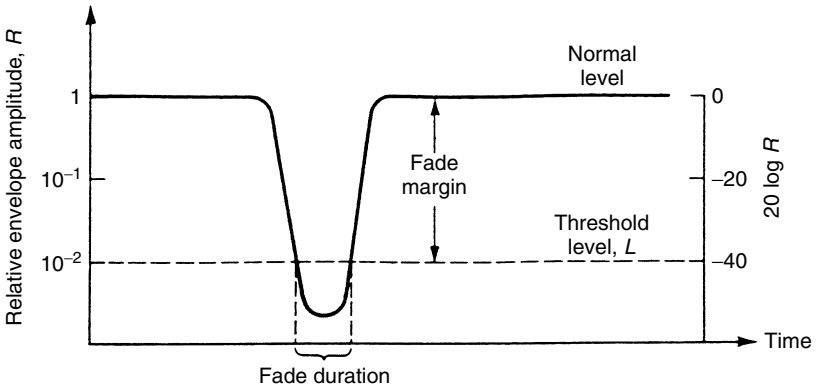


Figure 6. Definition of fading terms.

the threshold. The time spent below threshold for a given fade is then the fade duration.

For LoS links, the probability distribution of fading signals is known to be related to and limited by the Rayleigh distribution, which is well known and is found by integrating the curve shown in Fig. 7. The Rayleigh probability density function is given by

$$p(r) = \begin{cases} (r/\sigma^2)e^{-r^2/2\sigma^2} & 0 \le r < \infty \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

for envelope amplitude r and mean square amplitude σ^2. The Rayleigh distribution function has the form

$$P(r_0) = \Pr(r \le r_0) = \int_0^{r_0} p(r)\,dr = 1 - \exp -\frac{r_0^2}{2\sigma^2} \tag{23}$$

For relative envelope amplitude $R = r/\sigma\sqrt{2}$ and relative threshold amplitude $L = r_0/\sigma\sqrt{2}$, the distribution function becomes

$$P(R < L) = 1 - \exp(-L^2) \tag{24}$$

An approximation to (24) valid for small values of L (representing deep fades) is

$$P(R < L) \approx L^2 \quad \text{for} \quad L < 0.1 \tag{25}$$

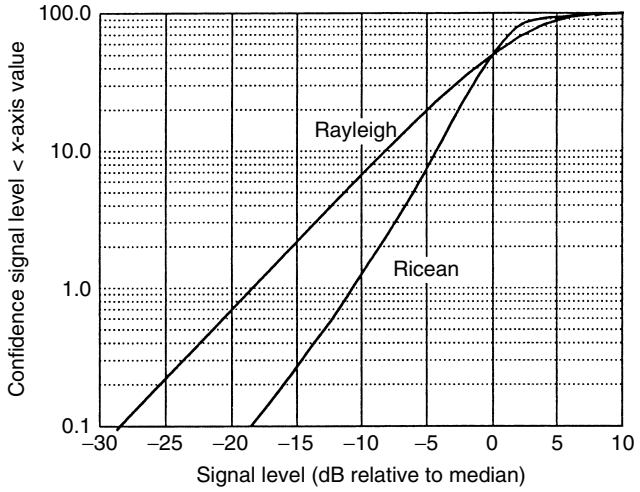


Figure 7. Rayleigh and Ricean distributions.

Fading probabilities are more conveniently expressed in terms of the fade margin F in decibels by letting $F = -20 \log L$. Then

$$P(R < L) = 10^{-F/10} \tag{26}$$

Actual observations of multipath fading indicate that in the region of deep fades, amplitude distributions have the same slope as the Rayleigh distribution but displaced. This characteristic corresponds to the special case of the Ricean distribution where a direct (or specular) component exists that is equal to or less than the Rayleigh fading component [8]. Thus, for deep fading, the distribution function becomes

$$P(R < L) = d(1 - \exp(-L^2)) \tag{27}$$

The parameter d that modifies the Rayleigh distribution has been termed a *multipath occurrence factor*. Experimental results of Barnett [9] show that

$$d = \frac{abD^3f}{4} \times 10^{-5} \tag{28}$$

where D = pathlength (mi)
f = frequency (GHz)
a = terrain factor: = 4 for overwater or flat terrain;
= 1 for average terrain;
= $\frac{1}{4}$ for mountainous terrain
b = climate factor: = $\frac{1}{2}$ for hot, humid climate;
= $\frac{1}{4}$ for average; temperate climate;
= $\frac{1}{8}$ for cool, dry climate

Combining this factor with the basic Rayleigh probability of (14) results in the following overall expression for probability of outage due to fading deeper than the fade margin:

$$P(o) = d10^{-F/10} = \frac{abD^3f \times 10^{-5}}{4}(10^{-F/10}) \tag{29}$$

The Rayleigh distribution given by (14) is the limiting value for multipath fading. Note that the distributions all have a slope of 10 dB per decade of probability.

3.2. Diversity Improvement

Diversity is used in LoS radio links to protect against either equipment failure or multipath fading. Here we

consider the improvement in multipath fading afforded by the two most commonly used diversity techniques:

- *Space diversity*, which provides two signal paths by use of vertically separated receiving antennas
- *Frequency diversity*, which provides two signal frequencies by use of separate transmitter–receiver pairs

The degree of improvement provided by diversity depends on the degree of correlation between the two fading signals. In practice, because of limitations in allowable antenna separation or frequency spacing, the fading correlation tends to be high. Fortunately, improvement in link availability remains quite significant even for high correlation. To derive the diversity improvement, we begin with the joint Rayleigh probability distribution function to describe the fading correlation between diversity signals, given by

$$P(R_1 < L, R_2 < L) = \frac{L^4}{1 - k^2} \quad \text{(for small } L\text{)} \tag{30}$$

where R_1 and R_2 are signal levels for diversity channels 1 and 2 and k^2 is the correlation coefficient. By experimental results, empirical expressions for k^2 have been established that are a function of antenna separation or frequency spacing, wavelength, and pathlength.

3.2.1. Space Diversity Improvement Factor. Vigants [10] has developed the following expression for k^2 in space diversity links:

$$k^2 = 1 - \frac{S^2}{2.75 D\lambda} \tag{31}$$

where S = antenna separation; D = pathlength; λ = wavelength; and S, D, and λ are in the same units. A more convenient expression is the space diversity improvement factor, given by

$$I_{\text{SD}} = \frac{P(R_1 < L)}{P(R_1 < L, R_2 < L)} \cong \frac{L^2}{L^4/(1 - k^2)} = \frac{1 - k^2}{L^2} \tag{32}$$

Using Vigants' expression for k^2 [Eq. (31)], we obtain

$$I_{\text{SD}} = \frac{S^2}{2.75\, D\lambda L^2} \tag{33}$$

When D is given in miles, S in feet, and λ in terms of carrier frequency f in gigahertz, I_{SD} can be expressed as

$$I_{\text{SD}} = \frac{(7.0 \times 10^{-5}) f S^2}{D} (10^{F/10}) \tag{34}$$

where F is the fade margin associated with the second antenna.

3.2.2. Frequency Diversity Improvement Factor. Using experimental data and a mathematical model, Vigants and Pursley [11] have developed an improvement factor for frequency diversity, given by

$$I_{\text{FD}} = \frac{(50 \Delta f)}{fD} (10^{F/10}) \tag{35}$$

where f = frequency (GHz)
Δf = frequency separation (GHz)
F = fade margin (dB)
D = pathlength (mi)

3.2.3. Effect of Diversity on Fading Statistics. The effect of diversity improvement, I_d, on probability of outage due to fading can be expressed as

$$P(o) = P_d \frac{(o)}{I_d} \tag{36}$$

where $P_d(o)$ is the outage probability of simultaneous fading in the two diversity signals. For space diversity, substituting Eqs. (29) and (34) into (36), we obtain

$$P_{\text{sd}}(o) = \frac{abD^4}{28S^2} 10^{-F/5} \tag{37}$$

where the fade margins on the two antennas are assumed equal. Likewise, for frequency diversity we obtain

$$P_{\text{fd}}(o) = (5 \times 10^{-8}) \frac{abf^3 D^4}{\Delta f} 10^{-\text{F}/5} \tag{38}$$

3.3. Frequency-Selective Fading

The first experiences with wideband digital radios revealed that measured error performance fell far short of the performance predicted by the fiat fading model assumed in our discussions so far. This result is due to the presence of frequency-selective fading during which the amplitude and group delay characteristics become distorted. For digital signals, this distortion leads to intersymbol interference that, in turn, degrades the system error rate. This degradation is directly proportional to system bit rate, since higher bit rates mean smaller pulsewidths and greater susceptibility to intersymbol interference. Previously, in analog radio transmission, frequency-selective fading caused intermodulation distortion, but this effect was always secondary when compared to the received signal power. For digital radio systems, however, the traditional fade depth is found to be a poor indicator of error rate.

When the amplitude of the received signal is plotted versus frequency, as shown in Fig. 8, deep amplitude notches appear when the direct ray is out of phase with the indirect rays. These notches are separated in frequency by $1/\tau$, where τ is the time delay between the direct and indirect rays. The notch depth is determined by the relative amplitude of the direct and indirect rays. When an amplitude notch or slope appears in the band of a radio channel, degradation in error rate can be expected. This variation of amplitude with frequency, known as *amplitude dispersion*, is often the main source of degradation in digital radio systems.

3.3.1. Channel Models. Both low-order power series [12] and multipath transfer functions [13] have been used to model the effects of frequency-selective fading. Several multipath transfer function models have been developed, usually based on the presence of two [14] or three [13]

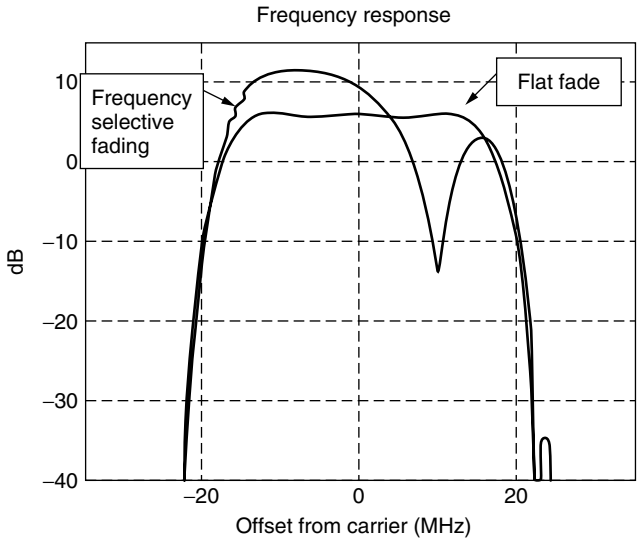


Figure 8. Multipath fading effect.

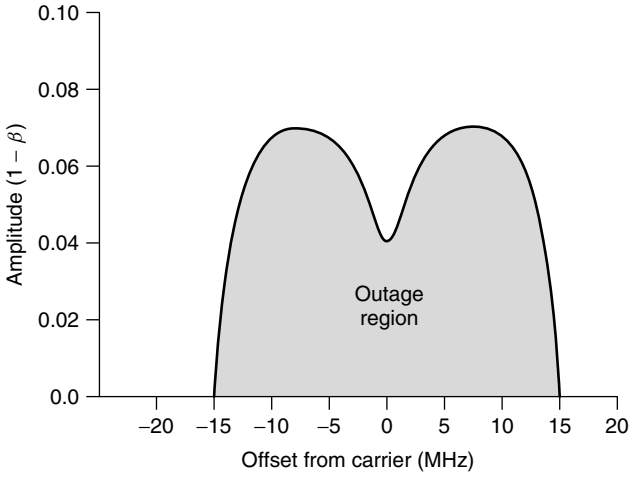


Figure 9. M curve.

rays. In general, the multipath channel transfer function can be written as

$$H(\omega) = 1 + \sum_{i=1}^{n} \beta_i \exp(j\omega\tau_i) \tag{39}$$

where the direct ray has been normalized to unity and the β_i and τ_i are amplitude and delay of the interfering rays relative to the direct ray. The two-ray model can thus be characterized by two parameters, β and τ. In this case, the amplitude of the resultant signal is

$$R = (1 + \beta^2 + 2\beta \cos \omega\tau)^{l/2} \tag{40}$$

and the phase of the resultant is

$$\phi = \arctan \frac{\beta \sin \omega\tau}{1 + \beta \cos \omega\tau} \tag{41}$$

Although the two-ray model is easy to understand and apply, most multipath propagation research points toward the presence of three (or more) rays during fading conditions. Out of this research, Rummler's three-ray model [13] is the most widely accepted.

3.3.2. Dispersive Fade Margin. The effects of dispersion due to frequency-selective fading are conveniently characterized by the dispersive fade margin (DFM), defined as in Fig. 6 but where the source of degradation is distortion rather than additive noise. The DFM can be estimated from M-curve signatures. This signature method of evaluation has been developed to determine radio sensitivity to multipath dispersion using analysis [15], computer simulation, or laboratory simulation [16]. The signature approach is based on an assumed fading model (e.g., two-ray or three-ray). The parameters of the fading model are varied over their expected ranges, and the radio performance is analyzed or recorded for each setting. To develop the signature, the parameters are adjusted to provide a threshold error rate (say, 1×10^{-7}) and a plot of the parameter settings is made to delineate the outage area. A typical signature developed by using the two-ray model is shown in Fig. 9. The area under the signature curve corresponds to conditions for which the bit error rate exceeds the threshold error rate; the area outside the signature corresponds to a BER that is less than the threshold value. The M shape of the curve (hence the term M curve) indicates that the radio is less susceptible to fades in the center of the band than to off-center fades. As the notch moves toward the band edges, greater notch depth is required to produce the threshold error rate. If the notch lies well outside the radioband, no amount of fading will cause an outage.

3.3.3. Improvements Due to Diversity and Equalization. Both diversity and adaptive equalization can be used, separately or together, to improve digital radio performance in the presence of frequency-selective fading. Diversity reduces the probability of in-band dispersion. Adaptive equalization reduces the in-band difference between the minimum- and maximum-amplitude values and, depending on the type of equalizer, reduces the in-band difference between group delay values also. In many instances, both diversity and equalization have been necessary to meet performance objectives. Interestingly, the combined improvement obtained by simultaneous use of diversity and equalization has been found to be larger than the product of the individual improvements. This synergistic effect has been reported in several experiments [18,19], where the added improvement has resulted from the diversity combiner's ability to replace in-band notches with slopes that are easier to equalize. Field measurements of EER distributions for a 6-GHz, 90-Mbps, 8-PSK radio on a 37.3-mi link are shown in Fig. 10 [20]. This system was tested in four configurations: unprotected, with adaptive equalization, with space diversity, and with equalization plus diversity. These results indicate a synergistic effect, with large improvement observed in the combination of equalization and diversity.

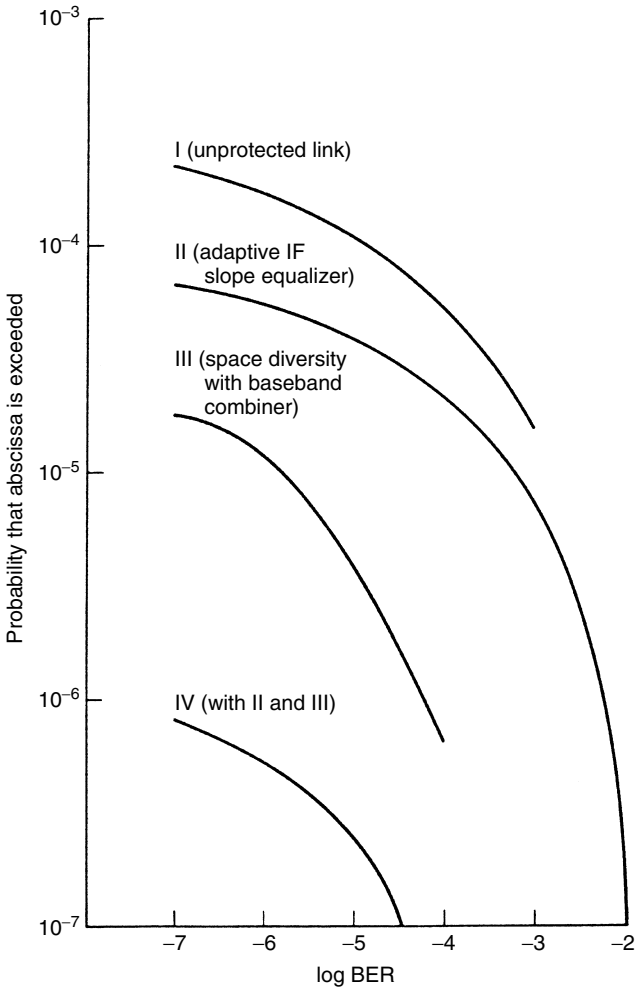


Figure 10. BER Distributions for 6-GHz, 90-Mbps, 8-PSK, 37.3-mi link showing effects of equalization and diversity [20].

4. FREQUENCY ALLOCATIONS AND INTERFERENCE EFFECTS

The design of a radio system must include a frequency allocation plan, which is subject to approval by the local frequency regulatory authority. In the United States, radio channel assignments are controlled by the Federal Communications Commission (FCC) for commercial carriers and by the National Telecommunications and Information Administration (NTIA) for government systems. Figure 11 shows the spectrum allocation process in the United States. The FCC's regulations for use of microwave spectrum establish eligibility rules, permissible-use rules, and technical specifications. There are four principal users who either share or exclusively use a particular spectrum allocation: common carriers, broadcasters, cable TV operators, and private companies. FCC regulatory specifications are intended to protect against interference and to promote spectral efficiency. Equipment-type acceptance regulations include transmitter power limits, frequency stability, out-of-channel emission limits, and antenna directivity. For digital microwave, specifications are added

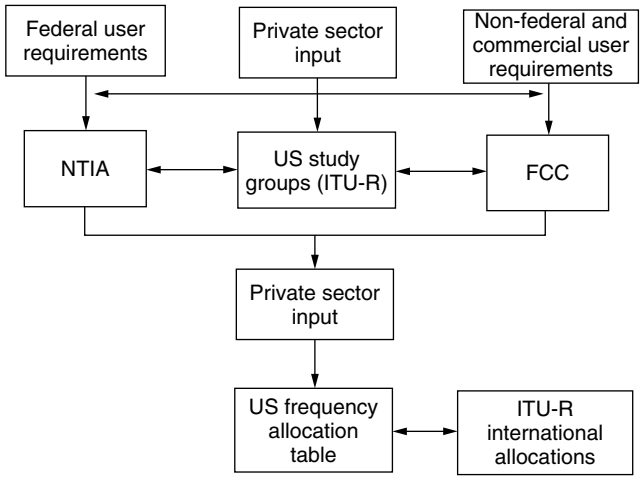


Figure 11. Spectrum allocation process in the United States.

for digital modulation, spectral efficiency in bps/Hz, and voice channel capacity [21].

The International Telecommunications Union Radio Committee (ITU-R) issues recommendations on radio channel assignments for use by national frequency allocation agencies. Although the ITU-R itself has no regulatory power, it is important to realize that ITU-R recommendations are usually adopted on a worldwide basis. With regard to digital microwave systems, the ITU-R has issued recommendations beginning with the 1982 Plenary Assembly.

The usual practice in frequency channel assignments for a particular frequency band is to separate transmit (GO) and receive (RETURN) frequencies by placing all GO channels in one half of the band and all RETURN channels in the other half. With this approach, all transmitters on a given station are in either the upper or lower half of the band, with receivers in the remaining half. Within each half-band, adjacent channels must be spaced far enough apart to avoid energy spillover between channels. A common scheme used to increase adjacent channel discrimination is to alternate between vertical and horizontal polarization. The isolation provided by cross-polarizing adjacent channels is on the order of $\geq$20 dB. At the edges of the band, a guard spacing is necessary to protect against interference into and from adjacent bands.

Another important consideration in radio-link design is RF interference, which may occur from sources internal or external to the radio system. The system designer should be aware of these interference sources in the area of each radio link, including their frequency, power, and directivity. Certain steps can be taken to minimize the effects of interference: good site selection, use of properly designed antennas and radios to reject interfering signals, and use of a properly designed frequency plan. A comprehensive frequency study is required in order to select frequencies that will not create or experience interference with existing systems operating in the same frequency bands. Access to a frequency database is required for interference analyses. The FCC prescribes the

use of EIA/TIA Bulletin TSB10-F for interference analysis of microwave systems [22].

The effect of RF interference on a radio system depends on the level of the interfering signal and whether the interference is in an adjacent channel or is cochannel. A cochannel interferer has the same nominal radiofrequency as that of the desired channel. Cochannel interference arises from multiple use of the same frequency without proper isolation between links. Adjacent-channel interference results from the overlapping components of the transmitted spectrum in adjacent channels. Protection against this type of interference requires control of the transmitted spectrum, proper filtering within the receiver, and orthogonal polarization of adjacent channels.

The performance criteria for digital radio systems in the presence of interference are usually expressed in one of two ways: allowed degradation of the S/N (SNR) threshold or allowed BER. Both criteria are stated for a given signal-to-interference ratio (S/I).

5. DIGITAL RADIO DESIGN

A block diagram of a digital radio transmitter and receiver is shown in Fig. 12. The traffic data streams at the input to the transmitter are usually in coded form, for example bipolar, and therefore require conversion to an NRZ (non-return-to-zero) signal with an associated timing signal. The multiplexer combines the traffic NRZ streams and any auxiliary channels used for orderwires into an aggregate data stream. This step is accomplished either by using pulse stuffing, which allows the radio clock rate to be independent of the traffic data, or by using a synchronous interface, which requires the radio and traffic data to be controlled by the same clock. The aggregate signal is scrambled to obtain a smooth radio spectrum and ensure recovery of the timing signal at the receiver. For phase modulation, some form of differential encoding is often employed to map the data into a change of phase from one signaling interval to the next. The modulator converts the digital baseband signals into a modulated intermediate frequency (IF), which is typically at 70 MHz when the final frequency is in the microwave band. The RF carrier is generated by a local oscillator, which is mixed with the IF-modulated signal to produce the microwave signal. The RF power amplification is accomplished by a traveling-wave tube (TWT) or a solid-state amplifier such as the gallium arsenide field-effect transistor (GaAs FET) amplifier. The final component of the transmitter is the RF filter, which shapes the transmitted spectrum and helps control the signal bandwidth.

At the receiver, the RF signal is filtered and then mixed with the local oscillator to produce an IF signal. The IF signal is filtered and amplified to provide a constant output level to the demodulator. Automatic gain control (AGC) in the IF amplifier provides variable gain to compensate for signal fading. Because the AGC voltage is a convenient indicator of received signal level, it is often used for performance monitoring or diversity combining. Fixed equalization is required to compensate for static amplitude or delay distortion from radio components, such as a TWT or filter, or to build out differential delay between RF channels in diversity operation. Adaptive equalization may also be required to deal with frequency-selective fading on the transmission path. Using the amplified and equalized IF signal, the demodulator recovers data and corresponding timing signals. Some type of performance monitoring is also commonly found in the demodulator, often based on eye pattern opening or pseudoerror techniques. The recovered baseband signal is next decoded and descrambled to reconstruct the aggregate datastream. The demultiplexer recovers the traffic datastreams and auxiliary channels. Finally, in the baseband encoder the standard data interface is generated.

5.1. Antennas

Microwave antennas used in line-of-sight transmission (terrestrial or satellite) provide high gain because radiated energy is focused within a small angular region. The power gain of a parabolic antenna is directly related to both the aperture area of the reflector and the frequency of operation. Relative to an isotropic antenna, the gain can be expressed as

$$G = \frac{4\pi\eta A}{\lambda^2} \tag{42}$$

where λ is the wavelength of the frequency, A is the actual area of the antenna in the same units as λ, and η is the efficiency of the antenna. The antenna efficiency is a number between 0 and 1 that reduces the theoretical gain because of physical limitations such as nonoptimum illumination by the feed system, reflector spillover, feed system blockage, and imperfections in the reflector surface. Converting (42) to more convenient units, the gain may be expressed in decibels as

$$G_{\mathrm{dB}} = 20\log f + 20\log d + 10\log\eta - 49.92 \tag{43}$$

where f is the frequency in megahertz and d is the antenna diameter in feet.

Waveguide or coaxial cable is required to connect the top of the radio rack to the antenna. Coaxial cable is limited to frequencies up to the 2-GHz band. Three types of waveguide are used — rectangular, elliptical, and circular — which vary in cost, ease of installation, attenuation, and polarization. The rectangular waveguide has a rigid construction, is available in standard lengths, and is typically used with short runs requiring only a single polarization. Standard bends, twists, and flexible sections are available, but multiple joints can cause reflections. The elliptical waveguide is semirigid, is available in any length, and is therefore easy to install. It accommodates only a single polarization, but has lower attenuation than does the rectangular waveguide. The circular waveguide has the lowest attenuation, about half that of the rectangular one, and also provides dual polarization and multiple band operation in a single waveguide. Its disadvantages are higher cost and more difficult installation.

5.2. Diversity Design

Diversity in LoS links is used to increase link availability by reducing the effects of multipath fading, improving the

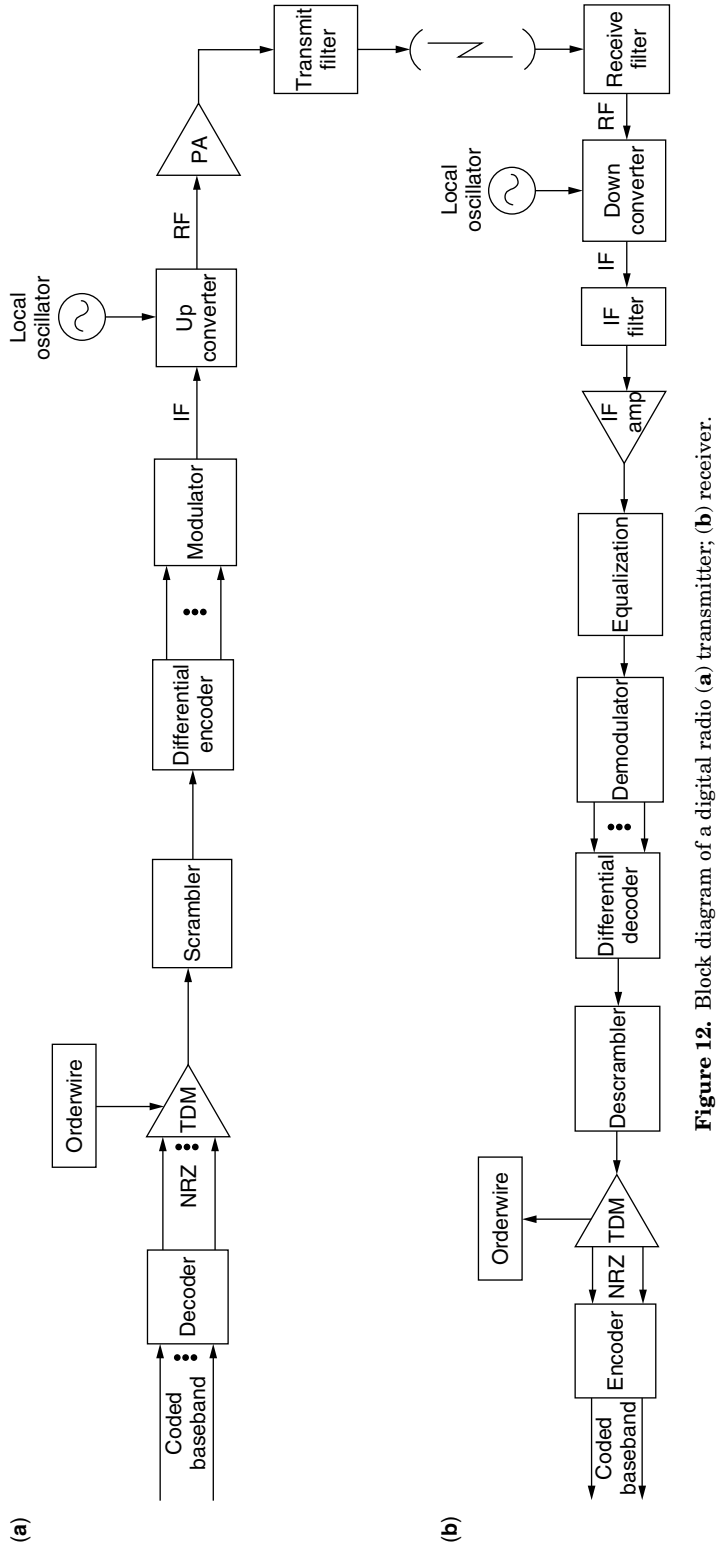


Figure 12. Block diagram of a digital radio (**a**) transmitter; (**b**) receiver.

combined output SNR ratio, and protecting against equipment failure. The most common forms of diversity use two parallel paths, separated in frequency or space, to provide one-for-one (1 : 1) protection on each link. The improvement afforded each link depends on the degree of correlation in fading between the two paths and the ability of a combiner to recognize and mitigate the effects of fading or equipment failure.

A typical arrangement for space diversity has two transmitters that operate on the same frequency and can be switched for output to a common antenna. One transmitter can operate in a hot standby mode, while the other is online; but as an alternative configuration, they can be combined to provide a 6-dB increase over the power available from a single transmitter. In this latter case, the failure of one transmitter power amplifier causes a 6-dB drop in output power but the link remains operational. The two receivers are connected to different antennas that are physically separated to provide the desired space diversity effect. The receiver outputs are fed to the combiner, which combines the two received signals. Space diversity, unlike frequency diversity, does not require an additional frequency assignment and is therefore more efficient in the use of spectrum. Its disadvantage is that additional antennas and waveguides are required, making it more expensive than frequency diversity arrangements.

In a typical arrangement for frequency diversity, two transmitters operate continuously on different frequencies but carry identical traffic. The receivers are connected to the same antenna but are tuned to separate frequencies. The combiner function is identical to that of the space diversity configuration. The use of frequency diversity doubles the spectrum amount required — a significant disadvantage in congested frequency bands. Unlike space diversity, however, frequency diversity provides two complete, independent paths, allowing testing of one path without interrupting service, while requiring only a single antenna per link end. Angle diversity is a newer form of diversity that has been shown to have cost and performance advantages over the more conventional diversity techniques already discussed. The basic idea behind angle diversity is that most deep fades, which are caused by two rays interfering with one another, can be mitigated by a small change in the amplitude or phase of the individual rays. Therefore, if a deep fade is observed on one beam of an angle diversity receiving antenna, switching to (or combining with) the second beam will change the amplitude or phase relationship and reduce the fade depth. Two techniques have been used to achieve angle diversity: a single antenna with two feeds that have a small angular displacement between them, or two separate antennas having a small angular difference in vertical elevation.

Variations of these protection arrangements include hot standby, hybrid diversity, and M : N protection. Hot standby arrangements apply to those cases where a second RF path is not deemed necessary, as with short paths where fading is not a significant problem. Here both pairs of transmitters and receivers are operated in a hot standby configuration to provide protection against equipment failure. The transmitter configuration is identical to that of space diversity. The receiving antenna feeds both receivers, tuned to the same frequency, through a power splitter. Hybrid diversity is provided by using frequency diversity but with the receivers connected to separate antennas that are spaced apart. This arrangement, which combines space and frequency diversity, improves the link availability beyond that realized with only one of these schemes.

For more efficient use of equipment and spectrum, diversity techniques are sometimes applied to a section of one or more links. In its simplest form, frequency diversity is used per section, with one protection channel used for N operational channels. This method can be extended to provide M : N protection, where M protection channels are shared by N operational channels. Further protection can be provided by using space diversity on a per-hop basis and frequency diversity on a section basis.

A diversity combiner performs the combining or selection of diversity signals. This function can be performed at RF [24], IF [25], or baseband [23]. Combiner techniques used in analog radio transmission [8] are generally applicable to digital radio. Phase alignment of the diversity signals becomes more important in digital radio systems, however, because of the potential occurrence of error bursts or loss of timing synchronization when combining misaligned diversity signals. Since delay equalization is simpler at baseband than at IF or RF baseband, "hitless" switching is a popular choice in digital radio combiners. This selection combiner uses some form of in-service performance monitor in each receiver to select the output signal after demodulation and data detection.

5.3. Adaptive Equalizer Design

Initial applications of wideband digital radios revealed that dispersion due to frequency-selective fading was the dominant source of multipath outages. These experiences led to the development and use of adaptive equalization so that outage requirements could be met. Since the introduction of the first adaptive equalizers to digital radio, these devices have undergone a rapid evolution. The degree of sophistication required of the equalizer depends primarily on the bit rate and pathlength. The types of equalizer in use today range from simple amplitude slope equalizers to complex transversal filter equalizers.

To obtain best equalizer performance, particularly for group delay distortion, transversal equalization techniques are now commonly applied to digital radio [26]. Figure 13 shows a block diagram of a QAM demodulator equipped with a transversal filter equalizer. The demodulator outputs are equalized by a forward baseband equalizer whose tap weights are controlled by decision feedback. The tap weights are continuously adjusted to correct channel distortion and provide the desired pulse response. The number of taps and the tap spacing in the transversal filters are design parameters that are chosen according to the system bit rate, channel fading characteristics, and radio outage requirements. Field tests of a baseband adaptive transversal filter applied to a 90-Mbps, 16-QAM radio show outage improvement by a factor of >3 compared with the same radio equipped with an IF slope equalizer only [27].

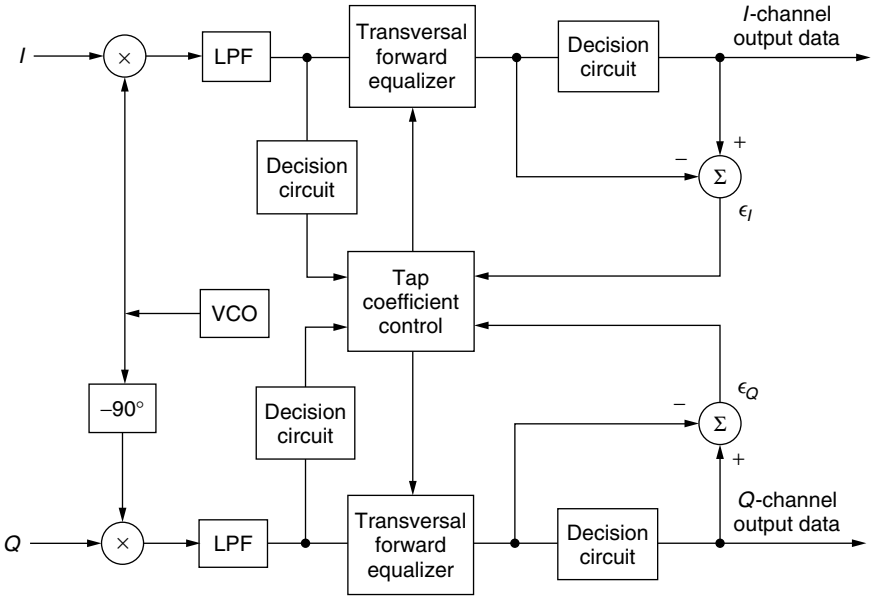


Figure 13. QAM demodulator equipped with decision-directed transversal filter equalizer.

6. RADIO-LINK CALCULATIONS

Procedures for allocating radio-link performance must be based on end-to-end system performance requirements. Outages result from both equipment failure and propagation effects. Allocations for equipment and propagation outage are usually separated because of their different effects on the user and different remedy by the system designer. Here we will consider procedures for calculating values of key radio-link parameters, such as transmitter power, antenna size, and diversity design, based on propagation outage requirements alone. These procedures include the calculation of intermediate parameters such as system gain and fade margin. Automated design of digital LoS links through use of computer programs is now standard in the industry [6,17].

6.1. System Gain

System gain (G_s) is defined as the difference, in decibels, between the transmitter output power (P_t) and the minimum receiver signal level (RSL) required to meet a given bit error rate objective (RSL_m):

$$G_s = P_t - \text{RSL}_m \tag{44}$$

The minimum required RSL_m, also called receiver threshold, is determined by the receiver noise level and the signal-to-noise ratio required to meet the given BER. Noise power in a receiver is determined by the noise power spectral density (N_0), the amplification of the noise introduced by the receiver itself (noise figure N_f), and the receiver bandwidth (B), The total noise power is then given by

$$P_N = N_0 B N_f \tag{45}$$

The source of noise power is thermal noise, which is determined solely by the temperature of the device. The thermal noise density is given by

$$N_0 = kT_0 \tag{46}$$

where k is Boltzmann's constant (1.38×10^{-23} J/K) and T_0 is absolute temperature (K)

The reference for T_0 is normally assumed to be room temperature, 290 K, for which $kT_0 = -174$ dBm/Hz. The minimum required RSL may now be written as

$$\begin{aligned} \text{RSL}_m &= P_N + \text{SNR} \\ &= kT_0 B N_f + \text{SNR} \end{aligned} \tag{47}$$

It is often more convenient to express (47) as a function of data rate R and E_b/N_0. Since $\text{SNR} = (E_b/N_0)(R/B)$, we can rewrite (47) as

$$\text{RSL}_m = kT_0 R N_f + \frac{E_b}{N_0} \tag{48}$$

The system gain may also be stated in terms of the gains and losses of the radio link:

$$G_s = L_p + F + L_t + L_m + L_b - G_t - G_r \tag{49}$$

where G_s = system gain (dB)
L_p = free-space path loss (dB), given by Eq. (4)
F = fade margin (dB)
L_t = transmission line loss from waveguide or coaxials used to connect radio to antenna (dB)
L_m = miscellaneous losses such as minor antenna misalignment, waveguide corrosion, and increase in receiver noise figure due to aging (dB)
L_b = branching loss due to filter and circulator used to combine or split transmitter and receiver signals in a single antenna
G_t = gain of transmitting antenna
G_r = gain of receiving antenna

The system gain is a useful figure of merit in comparing digital radio equipment. High system gain is desirable since it facilitates link design — for example, by easing

the size of antennas required. Conversely, low system gain places constraints on link design — for example, by limiting path length.

6.2. Fade Margin

The traditional definition of fade margin is the difference, in decibels, between the nominal RSL and the threshold RSL as illustrated in Fig. 6. An expression for the fade margin F required to meet allowed outage probability $P(o)$ may be derived from (29) for an unprotected link, from (37) for a space diversity link, and from (38) for a frequency diversity link, with the following results:

$$F = 30 \log D + 10 \log(abf) - 56 - 10 \log P(o) \quad \text{(unprotected link)} \tag{50}$$

$$F = 20 \log D - 10 \log S + 5 \log(ab) - 7.2 - 5 \log P(o) \quad \text{(space diversity link)} \tag{51}$$

$$F = 20 \log D + 15 \log f + 5 \log(ab) - 5 \log(\Delta f) - 36.5 - 5 \log P(o) \quad \text{(frequency diversity link)} \tag{52}$$

This definition of fade margin has traditionally been used to describe the effects of fading at a single frequency for radio systems that are unaffected by frequency-selective fading or to radio links during periods of flat fading. For wideband digital radio without adaptive equalization; however, dispersive fading is a significant contributor to outages so that the flat fade margins do not provide a good estimate of outage and are therefore insufficient for digital link design. Several authors have introduced the concept of effective, net, or composite fade margin [18] to account for dispersive fading. The *effective* (or *net* or *composite*) *fade margin* is defined as that fade depth that has the same probability as the observed probability of outage. The difference between the effective fade margin measured on the radio link and the flat fade margin measured with an attenuator is then an indication of the effects of dispersive fading. Since digital radio outage is usually referenced to a threshold error rate, BER_t, the effective fade margin (EFM) can be obtained from the relationship

$$P(A \geq \text{EFM}) = P(\text{BER} \geq \text{BER}_t) \tag{53}$$

where A is the fade depth of the carrier. The results of Eqs. (50)–(52) can now be interpreted as yielding the effective fade margin for a probability of outage given by the right-hand side of (53). Conversely, note that Eqs. (29), (37), and (38) are now to be interpreted with F equal to the EFM.

The effective fade margin is derived from the addition of up to three individual fade margins that correspond to the effects of flat fading, dispersion, and interference:

$$\text{EFM} = -10 \log(10^{-\text{FFM}/10} + 10^{-\text{DFM}/10} + 10^{-\text{IFM}/10}) \tag{54}$$

Here the flat fade margin (FFM) is given as the difference in decibels between the unfaded signal-to-noise ratio $(\text{SNR})_u$ and the minimum signal to-noise ratio $(\text{SNR})_m$ to meet the error rate objective, or

$$\text{FFM} = (\text{SNR})_u - (\text{SNR})_m \tag{55}$$

Similarly, we define the dispersive fade margin (DFM) and interference fade margin (IFM) as

$$\text{DFM} = (\text{SNR})_d - (\text{SNR})_m \tag{56}$$

and

$$\text{IFM} = (\text{S/I}) - (\text{SNR})_m \tag{57}$$

where $(\text{SNR})_d$ represents the effective noise due to dispersion and (S/I) represents the critical S/I below which the BER is greater than the threshold BER. Each of the individual fade margins can also be calculated as a function of the other fade margins by using (54), as in

$$\text{FFM} = -10 \log(10^{-\text{EFM}/10} + 10^{-\text{DFM}/10} + 10^{-\text{IFM}/10}) \tag{58}$$

BIOGRAPHY

David R. Smith is a professor in the Department of Electrical and Computer Engineering at The George Washington University, Washington, D.C. He received a D.Sc. in electrical engineering and computer science from The George Washington University in 1977; an M.S.E.E. from Georgia Tech, Atlanta, in 1970; and a B.S in physics from Randolph-Macon College, Randolph, Virginia, in 1967. He has worked as an electrical engineer and manager within the Department of Defense, Washington, D.C., for both the Navy Department and the Defense Information Systems Agency. Since 1967 he has held adjunct, visiting, research, and regular faculty positions at Georgia Tech, The George Washington University, and George Mason University, Fairfax, Virginia. He has also consulted with numerous companies in the areas of wireless and digital communications. His publications include over 20 IEEE articles, the books *Digital Transmission Systems* published by Kluwer and *Emerging Public Safety Wireless Communication Systems* published by Artech House, and chapters contributed to two other books. Current interests include research in areas of wireless telecommunications, advanced networking, digital communications, propagation modeling, and computer simulation and modeling of communication systems.

BIBLIOGRAPHY

1. Henry R. Reed and Carl M. Rusell, *Ultra High Frequency Propagation*, Chapman & Hall, London, 1966.
2. K. Bullington, Radio propagation fundamentals, *Bell Syst. Tech. J.* **36**: 593–626 (May 1957).
3. J. H. Van Vleck, *Radiation Laboratory Report* 664, MIT, Cambridge, MA, 1945.
4. Reports of the ITU-R, 1990, Annex to Vol. V, *Propagation in Non-Ionized Media*, ITU, Geneva, 1990.
5. S. H. Lin, Nationwide long-term rain rate statistics and empirical calculation of II-GHt microwave rain attenuation, *Bell Syst. Tech. J.* **56**: 1581–1604 (Nov. 1977).

6. D. R. Smith, A computer-aided design methodology for digital line-of-sight links, *Proc. 1990 Int. Conf. Communications*, 1990, pp. 59–65.

7. *Engineering Considerations for Microwave Communications Systems* GTE Lenkurt Inc., San Carlos, CA, 1981.

8. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill, New York, 1966.

9. W. T. Barnett, Multipath propagation at 4, 6, and 11 GHz, *Bell Syst. Tech. J.* **S1**: 321–361 (Feb. 1972).

10. A. Vigants, Space diversity performance as a function of antenna separation, *IEEE Trans. Commun. Technol.* **COM-16**(6): 831–836 (Dec. 1968).

11. A. Vigants and M. V. Pursley, Transmission unavailability of frequency diversity protected microwave FM radio systems caused by multipath fading, *Bell Syst. Tech. J.* **58**: 1779–1796 (Oct. 1979).

12. L. C. Greenstein and B. A. Czekaj, A polynomial model for multipath fading channel responses, *Bell Syst. Tech. J.* **59**: 1197–1205 (Sept. 1980).

13. W. D. Rummler, A new selective fading model: Application to propagation data, *Bell Syst. Tech. J.* **58**: 1037–1071 (May/June 1979).

14. W. C. Jakes, Jr., An approximate method to estimate an upper bound on the effect of multipath delay distortion on digital transmission, *Proc. 1978 Int. Conf. Communications*, 1978, pp. 47.1.1–47.1.5.

15. M. Emshwiller, Characterization of the performance of PSK digital radio transmission in the presence of multipath fading, *Proc. 1978 Int. Conf. Communications*, 1978, pp. 47.3.1–47.3.6.

16. C. W Lundgren and W. D. Rummler, Digital radio outage due to selective fading — observation vs. prediction from laboratory simulation, *Bell Syst. Tech. J.* **58**: 1073–1100 (May/June 1979).

17. T. C. Lee and S. H. Lin, The DRDIV computer program — a new tool for engineering digital radio routes, *1986 GLOBECOM*, pp. 51.5.1–51.5.8.

18. C. W. Anderson, S. G. Barber, and R. N. Patel, The effect of selective fading on digital radio, *IEEE Trans. Commun.* **COM-27**(12): 1870–1876 (Dec. 1979).

19. T. S. Giuffrida, Measurements of the effects of propagation on digital radio systems equipped with space diversity and adaptive equalization, *Proc. 1979 Int. Conf. Communications*, 1979, pp. 48.1.1–48.1.6.

20. D. R. Smith and J. J. Cormack, Improvement in digital radio due to space diversity and adaptive equalization, *Proc. 1984 Global Telecommunications Conf.*, 1984, pp. 45.6.1–45.6.6.

21. Federal Communications Commission Rules and Regulations, Part 101, *Fixed Microwave Services*, Aug. 1, 1996.

22. TIA/EIA Telecommunications Systems Bulletin, TSB10-F, *Interference Criteria for Microwave Systems*, June 1994.

23. C. M. Thomas, J. E. Alexander, and E. W. Rahneberg, A new generation of digital microwave radios for U.S. military telephone networks, *IEEE Trans. Commun.* **COM-27**(12): 1916–1928 (Dec. 1979).

24. 1. Horikawa, Y. Okamoto, and K. Morita, Characteristics of a high capacity 16 QAM digital radio system on a multipath fading channel, *Proc. 1979 Int. Conf. Communications*, 1979, pp. 48.4.1–48.4.6.

25. G. deWitte, DRS-8: System design of a long haul 91 Mb/s digital radio, *Proc. 1978 Natl. Telecommunications Conf.*, 1978, pp. 38.1.1–38.1.6.

26. C. A. Siller, Jr., Multipath propagation, *IEEE Commun. Mag.* 6–15 (Feb. 1984).

27. G. L. Fenderson, S. R. Shepard, and M. A. Skinner, Adaptive transversal equalizer for 90 Mbls 16-QAM systems in the presence of multipath propagation, *Proc. 1983 Int. Conf. Communications*, 1983, pp. C8.7.1–C8.7.6.

TEST AND MEASUREMENT OF OPTICALLY BASED HIGH-SPEED DIGITAL COMMUNICATIONS SYSTEMS AND COMPONENTS

GREG D. LECHEMINANT
Agilent Technologies
Santa Rosa, California

Gauging the performance of a communications system can be accomplished in a number of ways and with a variety of parameters. The fundamental task of any communications system is to reliably transport information. Thus the one parameter that can describe the overall health of a digital communications system is the bit error-rate (BER). A BER test is a direct measure of the number of bits received incorrectly compared to the total number of bits transmitted. Prior to integration into a working system, it is also important to characterize the performance of the individual components and building blocks that make up the system. Rather than "system level" performance parameters such as BER, tests that describe physical or parametric properties can be used. Examples include signal strength, noise, spectral content, and tolerance to jitter.

1. MEASURING BIT ERROR RATIO

Modern optically based communications systems perform extremely well. BERs on the order of 1 bit received in error per trillion bits transmitted are common. This is often expressed as a BER of 1×10^{-12}. It is not uncommon for this parameter to be incorrectly referred to as the bit error *rate* instead of bit error ratio. However, the 1×10^{-12} number mentioned above does not indicate the rate at which bits are received in error. This would require knowing both the transmission rate in bits per second and the BER (ratio of errored bits to transmitted bits).

Measuring BER is typically achieved using a BER tester or "BERT." A BERT consists of two major building blocks. First there is a pattern generator. The pattern generator performs two major functions. It determines the specific data sequence that the system under test will transmit. This is commonly referred to as the *test pattern*. It is essential that specific, known patterns be used so that it can be determined whether the system under test correctly transported the data. The pattern generators also can provide control over characteristics of the test signal such as waveform amplitude and to some degree the pulseshape. The second block of the BERT is the error

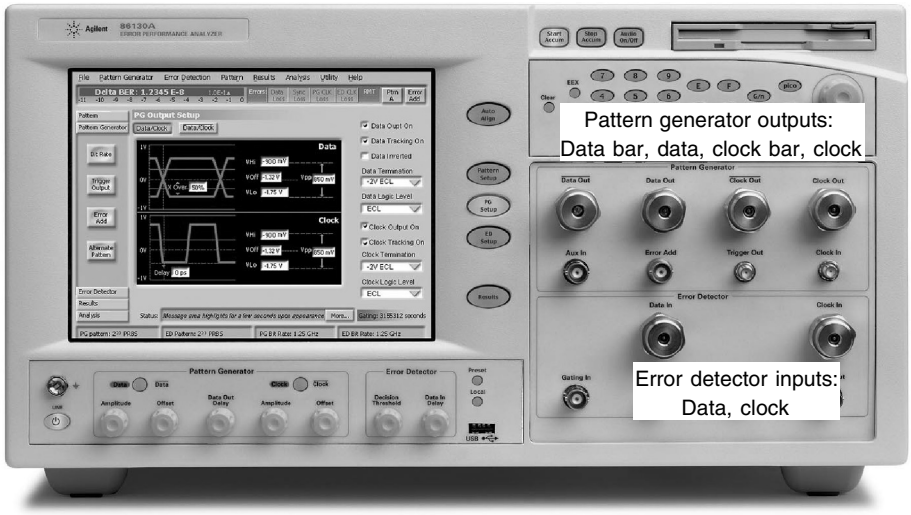


Figure 1. BERT test set.

detector. The error detector is used to verify that each bit sent through the system under test has been correctly received.

The process of performing a BER measurement is as follows (see Fig. 1):

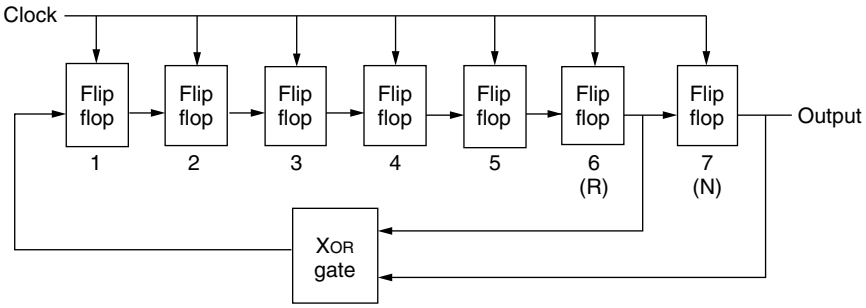
1. The pattern generator is configured to produce a specific data sequence, which is fed to the system under test.
2. The system transmits the pattern to its receiver. The receiver must determine the logic level for each incoming bit. In most cases, the signal at the receiver decision circuit will have been degraded because of attenuation and dispersion in the transmission path. The system receiver effectively produces a regenerated output signal or pattern according to what it thought the incoming bits were. A perfect receiver would never make a mistake in assessing the logic level of the input signal.
3. A separate pattern generator within the error detector section of the BERT produces a pattern identical to that originally fed to the system under test.
4. The error detector pattern and the signal or pattern from the system receiver are aligned in time and compared bit for bit. Whenever there is disagreement, an error is counted.
5. The BER is determined from the number of bits found in error and the total number of bits checked.

An error detector can be viewed as a decision circuit followed by an "exclusive OR" (XOR) logic circuit. Data from the system or circuit under test are fed to the error detector. A clock signal is also fed to the error detector. This clock signal is often a clock signal that the receiver of the system or circuit under test has derived from the data being transmitted. The clock is used to time the decision circuit and XOR gate. The timing of the decision circuit must be optimized within one clock cycle to sample the data signal at the ideal point in time (usually the middle of the bit period). Thus a "clock to data" alignment must be performed. This is usually an automatic adjustment available in the BERT. There is also an optimum amplitude threshold the error detector decision circuit will use to determine whether incoming data are at the 1 or 0 level. A BERT will also have an alignment procedure to determine this optimum level. The third basic step in setting up a BER test is to align the internal pattern generator of the error detector with the pattern from the system or circuit under test so that the XOR function is being performed on identical locations in each pattern. This too is an automatic procedure available in all modern BERTs.

Although a pattern generator can produce virtually any pattern sequence desired, most BER measurements are performed using pseudorandom binary sequences (PRBSs). A PRBS is a data pattern that attempts to replicate truly random data yet is completely deterministic (a requirement for performing a BER measurement). PRBS patterns have lengths of $2^N - 1$. For example, a $2^7 - 1$ pattern is 127 bits in length and will include all sequences 7 bits in length with the exception of 7 sequential 0's. A $2^{10} - 1$ PRBS is 1023 bits in length and has all sequences 10 bits in length with the exception of 10 sequential 0's (zeros). Pattern generators commonly produce pattern lengths of up to $2^{31} - 1$.

PRBS patterns are generated using a series of shift registers with feedback taps (see Fig. 2). One of the unique properties of a PRBS is that when compared to itself, the BER will be 50% except when the patterns from the system under test and the error detector are exactly aligned. This allows the alignment process to be performed quickly and efficiently.

There are several beneficial properties of PRBS patterns. The spectral content of a PRBS is quite broad; thus it can expose resonances in test devices. PRBS patterns can be decimated to produce shorter PRBS patterns. This is useful when testing multiplexing and demultiplexing circuits. (For example, the shorter patterns produced by a 1-to-4 demultiplexer are still PRBSs and can thus be easily tested for BER.) PRBS patterns can also



Sequence length	Shift-register configuration
2^7-1	$D^7 + D^6 + 1 = 0$, inverted
$2^{10}-1$	$D^{10} + D^7 + 1 = 0$, inverted
$2^{15}-1$	$D^{15} + D^{14} + 1 = 0$, inverted
$2^{23}-1$	$D^{23} + D^{18} + 1 = 0$, inverted
$2^{31}-1$	$D^{31} + D^{29} + 1 = 0$, inverted

Figure 2. Generating the PRBS.

be generated at extremely high speeds, limited only by the hardware generating the pattern.

In addition to PRBS patterns, specialized patterns can also be created and generated with a BERT. A special pattern might be created that causes a transmitter to produce more jitter than if a PRBS were used. Communications protocols often require specific bit sequences to "frame" the data payload. Thus a pattern may be built to include the framing bits. These patterns are not produced by a specific circuit topology, but instead are loaded into the memory of the pattern generator. The pattern generator then produces a data sequence according to the loaded pattern. (The internal pattern generator of the error detector can perform similarly.)

While the BER result is indicative of the system performance, it does not provide any indication of underlying reasons for poor or marginal performance. Sometimes the way in which errors occur can provide insight into the causes for errors. Consider the case when 2 or 3 adjacent bits are received in error. If the overall BER is 1×10^{-9}, the probability of 2 bits being received in error as a result of random mechanisms is one in 1×10^{18}. For three to occur in sequence the probability would be one in 1×10^{27}. Because 10^{27} bits at 1 Gbps (gigabits per second) takes millions of years to transmit, 3 sequential bits being in error due to random causes is extremely unlikely in this case. It is far more likely that these errors would be due to something deterministic.

Several techniques can help troubleshoot the root causes for BER. Examples include examining the intervals (time or bits) between errors, or whether errors occur mainly for logic 1s or logic 0s. If errors occur separated by some multiple of 10 bits, and there is a 10-bit-wide parallel bus somewhere in the system, there could be something physically wrong with one of the bus lines.

With the common occurrence of BERs of 1×10^{-12} and even lower in today's high-speed telecommunications systems, what should be the expected time required to verify that a system is operating at a low BER? It is ideal to collect 100 or even 1000 errors to be confident that the error ratio performance is being achieved. However, at an error ratio of 1×10^{-12} and a data rate on the order of 1 Gbps, it will take $11\frac{1}{2}$ days to collect 1000 errors! If the data rate is increased to 10 Gbps and the number of errors collected is reduced to only 100, it will still take several hours to complete the measurement.

Certain techniques can be used to try to determine BER in a reduced timeframe. Both techniques to be discussed rely on making BER measurements in nonideal conditions, thus increasing the BER and extrapolation to determine the ideal BER. One technique is to stress the signal being tested. The other technique is to adjust the sampling threshold of the error detector to nonideal levels. A common technique to stress a signal is to simply attenuate it prior to reaching the system receiver. As the signal level is reduced, the likelihood that a receiver will incorrectly interpret 1s as 0s and 0s as 1s increases. The BER will be artificially increased and then take less time to measure. By measuring the BER at several levels of attenuation, a BER–received power curve can be plotted (see Fig. 3). Extrapolation of the curve can then be used to determine the BER with no external signal attenuation. Extrapolation requires an assumption that the dominant error mechanisms are the same at both high and low signal levels. This usually implies random noise. However, it is possible that a deterministic but extremely low-probability mechanism exists and is not seen except at high power levels where noise is less likely to cause errors. Thus final system qualification may require a lengthy BER measurement at full signal power.

The BER will also be artificially increased if the error detector sampling threshold is increased or decreased from the ideal. However, this type of BER analysis is typically performed on the signal at the test receiver input rather than its output. The BER analysis is not made on the complete system, but instead is an assessment of the quality of the signal being presented to the receiver. This is called a *Q factor* measurement. The *Q* factor is essentially the signal-to-noise ratio (SNR) of the signal. If the dominant mechanism for errors in a system is random noise, then BER can be estimated by the *Q* factor parameter.

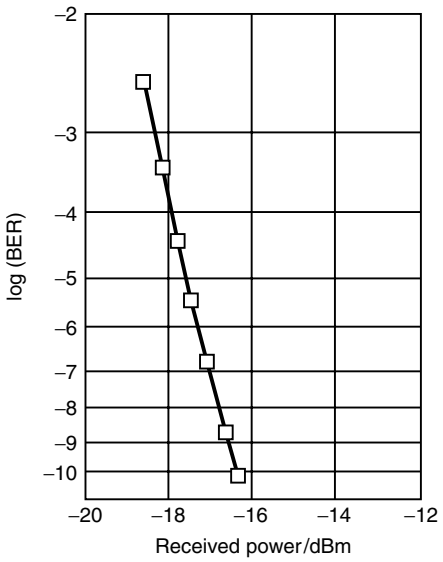


Figure 3. BER versus received power.

The measurement process is as follows. The transmitted signal from the system under test is injected directly into the error detector. (If the system is optical, a linear optical to electrical converter is used to create an electrical input to the error detector.) BER measurements are made with the error detector sampling threshold at several levels above and below the ideal threshold. From each BER result, a signal-to-noise parameter or Q value can be estimated. Again, these measurements are made with sampling thresholds that yield BERs worse than 1×10^{-9} and thus can be performed quickly. With several BER values collected, the BERs are converted to Q values and plotted against sampling threshold. From this plot, the optimum Q factor and optimum sampling threshold can be extrapolated (see Fig. 4).

What does the optimum Q factor tell us? Just as Q factor can be derived from BER, the optimum BER can be derived from the optimum Q factor. The optimum Q factor indicates the highest level of system performance that can be achieved with an ideal decision circuit as a receiver. This is useful for characterizing ultra-low-BER systems that would require extremely long test times for actual BER verification. It is important to recognize the assumptions that are made. The most important one is that the dominant error mechanism is random noise. The mathematical Q/BER relationships are based on this. In the end, final qualification of a system will likely require basic BER verification.

2. WAVEFORM ANALYSIS

Analysis of waveforms can yield a wealth of information about the overall quality of a high-speed digital transmitter or the transmission section of a communication system. In the R&D lab, a simple visual inspection of the signal allows a designer to quickly assess basic performance. In a manufacturing environment, several key parameters can

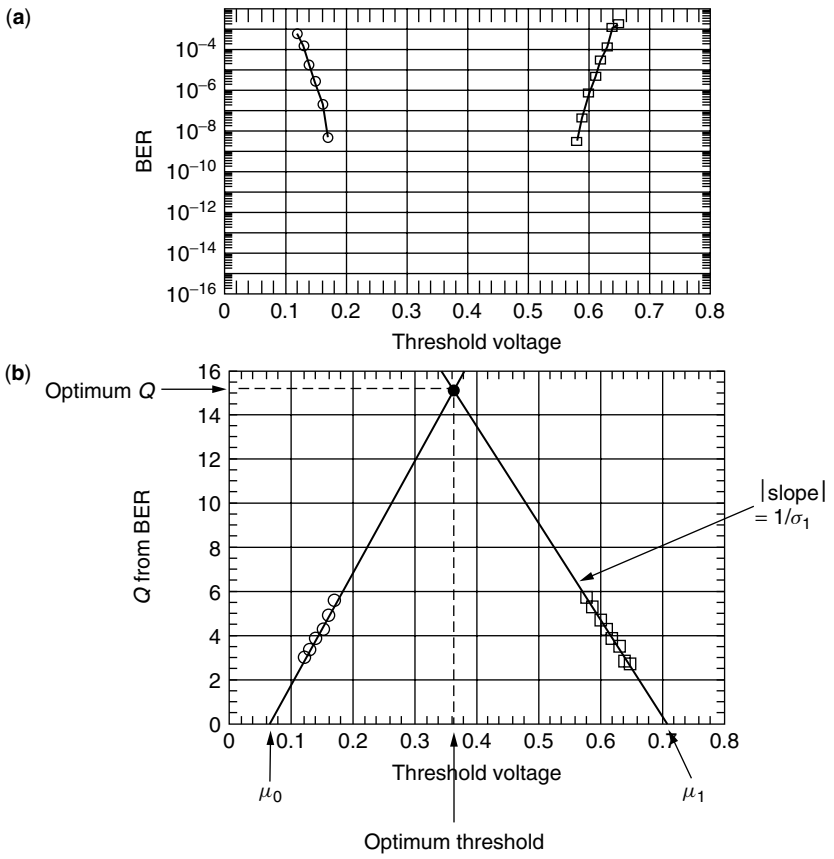


Figure 4. Estimating low BERs through Q-factor measurements.

be derived from a transmitter's waveform. Waveforms are viewed on oscilloscopes that display signal amplitude as a function of time. There are two basic formats to view a digital communications signal, either as a pulsetrain or as an eye diagram.

A *pulsetrain* is a display of some segment or sequence of the communications signal. For example, if a 1-Gbps signal were being examined, each bit is 1 ns (nanosecond) in duration. If the oscilloscope timebase were set to a width of 10 ns, 10 sequential bits could be displayed. Determining which 10 bits from a long data sequence are displayed is a function of how the oscilloscope is triggered. Typically, a pattern generator will produce a "pattern trigger." This is a pulse produced at the beginning of each repetition of the data pattern, such as a PRBS. Triggering the oscilloscope with the pattern trigger and adjusting the oscilloscope time delay can display any specific section of the pattern (see Fig. 5). Very long patterns present some difficulty in displaying pulsetrains. Because the trigger pulse is generated only once per every repetition of the pattern, the time between trigger events can become very large. For wide bandwidth sampling oscilloscopes, one data point is sampled for every trigger event. If a waveform is composed of 1000 sample points, the entire pattern must be transmitted 1000 times to complete acquisition of the waveform. For long pattern lengths it can take several minutes and even hours to produce a single waveform. Thus pulsetrains are usually displayed only when examining relatively short patterns.

Another difficulty when examining pulsetrains is that only a few bits can be examined at a given time. More bits can be displayed by decreasing the resolution of the oscilloscope timebase, but important details are usually lost because of this reduced resolution. Often when examining a high-speed digital communication signal it is desirable to determine the overall performance of the system for all patterns of data. It would be ideal to see this in one simple display. This can be achieved through the eye diagram. The *eye diagram* is a composite display of waveform samples acquired throughout the entire data pattern displayed on a common timebase. Consider the eight waveforms that can be generated from a 3-bit sequence (000, 001, . . . , 110, 111). If these eight waveforms are all placed on a common amplitude–time grid, the eye diagram is displayed (see Fig. 6).

With an eye diagram it is difficult to view the waveform from any individual bit. However, much information is available regarding the overall performance of the

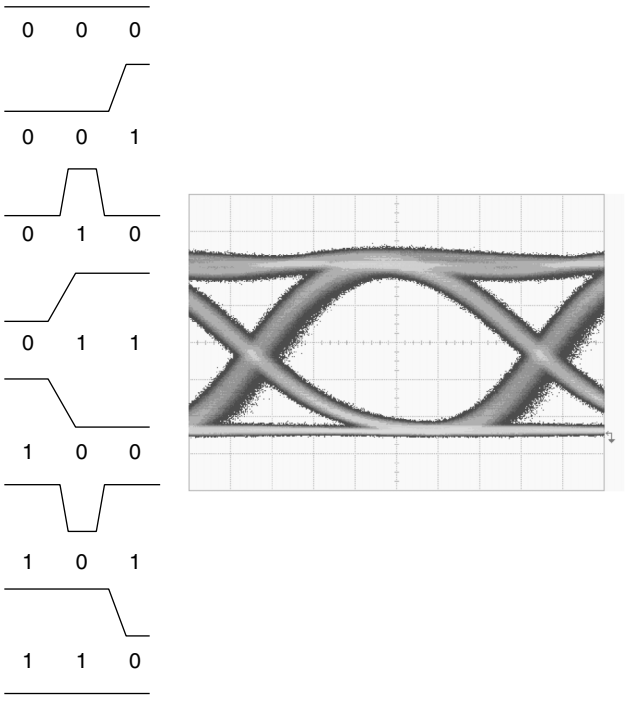


Figure 6. Building the eye diagram.

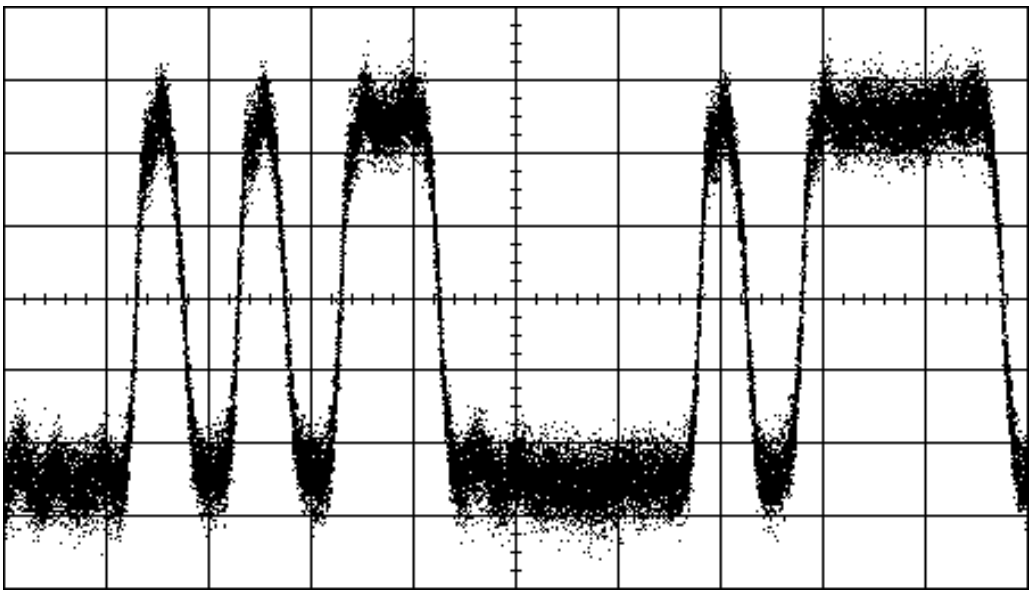


Figure 5. Pulsetrain waveform of a digital communications signal.

transmitted waveform. If the eye diagram begins to close horizontally, this is an indication of excessive waveform timing jitter. Slow rise and fall times cause vertical eye closure. Eye closure due to any mechanism presents a significant system level problem because it makes the decision process more difficult for the receiver at the end of the communication system.

An eye diagram is created when the oscilloscope is triggered with a clock signal that is synchronous to the data. In contrast to triggering with a pattern trigger, triggering with a clock signal allows the oscilloscope to acquire samples throughout the data pattern. Divided clocks, such as a rate $\frac{1}{4}$ or rate $\frac{1}{16}$ are also acceptable. The requirement is that the divided clock be synchronous with the data, and that the divisor be an integer. The oscilloscope can trigger on the data signal itself and create an eye diagram. However, although the displayed eye diagram may appear to be complete, approximately 75% of the data will be missing form the eye diagram. This is because only one of the four combinations of 2 adjacent bits (e.g., the 0−1 combination) will produce a signal transition that the oscilloscope can trigger on. Thus triggering on the data itself should be avoided whenever possible.

In an R&D environment many engineers and scientists have learned how to quickly gauge the quality of a signal through quick visual inspection of the eye diagram. In the most basic sense, an "open" eye diagram is indicative of a quality signal, while a "closed" eye is indicative of signal impairments.

In a manufacturing environment, the eye diagram is used to obtain specific parameters that indicate performance of high-speed digital transmitters. One common example is the measurement of the extinction ratio. The *extinction ratio* is used to describer how efficiently an optical transmitter converts its available signal strength to modulation power. In mathematical terms, it is simply the ratio of the power in a logic level one (1) to the power in a logic level zero (0). Since the eye diagram is composed of a multitude of logic ones and logic zeros, a statistical analysis is performed to determine the aggregate logic one power and the aggregate logic zero. This is achieved through the use of histograms. First a slice of data is acquired for the upper central portion of the eye diagram. A vertical histogram is constructed from these data. The mean of this histogram represents the power level of the aggregate ones making up the eye diagram. A similar process is used in the lower central portion of the eye diagram to determine the power of the aggregate logic zero (see Fig. 7). (A high extinction ratio is typically achieved by using very little power to transmit zeros. When this is achieved, virtually all of the available laser power is being used to transmit information. Thus, a high extinction ratio is indicative of a highly efficient use of laser power.)

The extinction ratio is just one example of histogram based statistical analysis to derive specific parameters of the eye diagram. Other measurements include optical modulation amplitude (OMA), rise time, fall time, jitter, and signal-to-noise-ratio. Again, these measurements are not made on individual bits, but rather on the composite eye diagram, thus yielding the overall performance of the transmitter with a single measurement.

Recall that it is desirable to have an open eye diagram in both the vertical and horizontal sense. It is difficult to measure a simple numerical parameter that can describe the openness of the eye diagram. Instead, a process called "eye-mask testing" is used. An eye-mask is a constellation of solid polygons that represent where the eye diagram waveform may *not* exist. A typical eye-mask consists of a polygon located in the center of the eye diagram, as well as

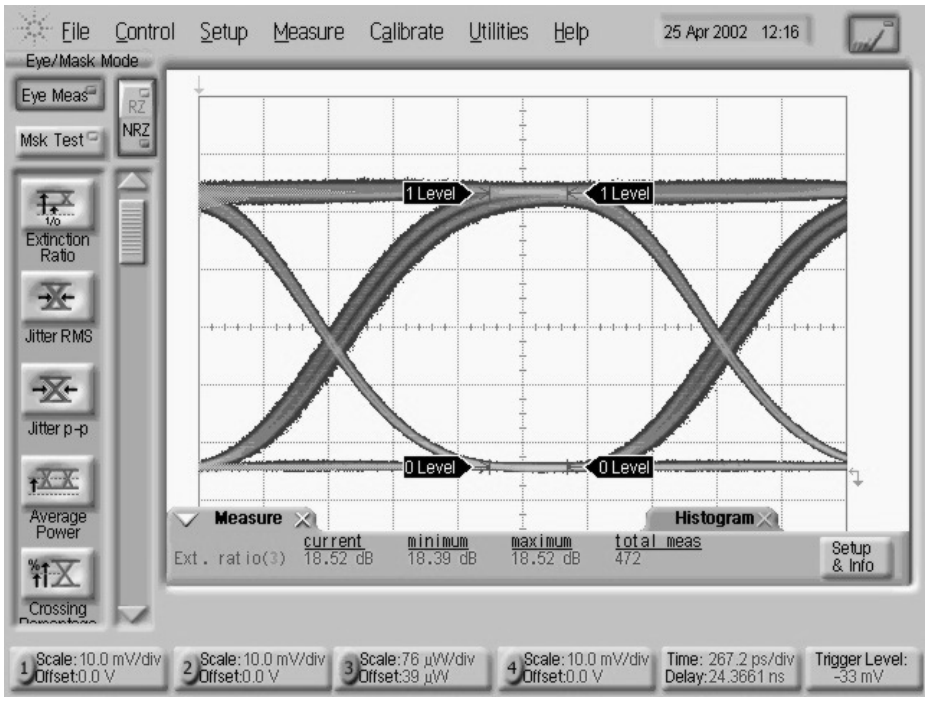


Figure 7. Measuring extinction ratio.

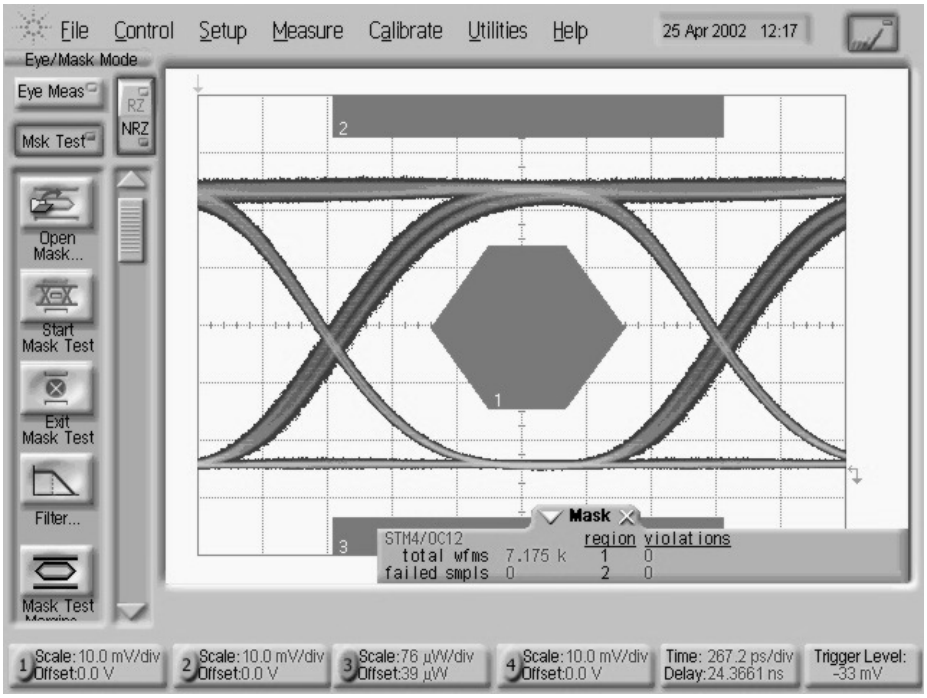


Figure 8. Eye-mask testing.

one polygon above and one polygon below (see Fig. 8). The eye diagram is not allowed to intersect or "violate" any of the mask polygons. The minimum acceptable opening for the eye diagram is then defined by the size and shape of the central polygon.

The shape of the displayed eye diagram can be altered by the frequency response of the oscilloscope measurement channel. It is common for directly modulated high-speed communication lasers to exhibit significant overshoot and ringing during the transition from a low-power logic 0 to a higher-power logic 1. It takes a wide-bandwidth oscilloscope channel to view this phenomenon. For example, if a laser is transmitting 2.5-Gbps data, the oscilloscope bandwidth needs to approach 10 GHz or higher for an accurate representation of the true waveform.

Eye-mask testing is a key element of most industry standards that specify high-speed optically based communication systems. To achieve consistent results across the industry, it is essential to specify the frequency response of the measurement channel. Thus in addition to defining the shape of the eye-mask, the measurement system is also defined through the concept of a reference receiver. A reference receiver usually consists of a photodetector followed by a lowpass filter (see Fig. 9). The combined response of the two elements typically follows a fourth-order Bessel–Thomson frequency response. This response is chosen since it closely approximates the response of a Gaussian filter. A Gaussian response yields minimal distortion of the waveform.

It is interesting to note the bandwidth that is normally specified for a reference receiver. In most communication standards, the −3-dB bandwidth is set to be 75% of the optical bit rate. For example, a reference receiver for a 10-Gbps system would have a bandwidth of 7.5 GHz. Initially

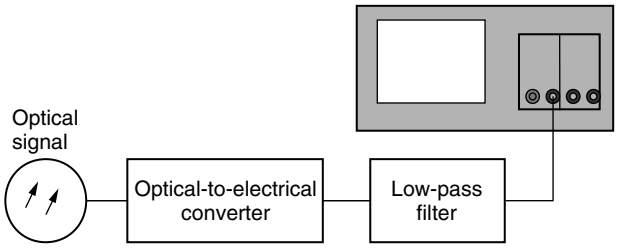


Figure 9. Eye-mask reference receiver.

this seems counterintuitive. Intentionally reducing the measurement bandwidth is likely to change the shape of the eye diagram. Effects such as the overshoot and ringing mentioned above can literally disappear when the bandwidth is reduced. The reduced bandwidth of the reference receiver can actually yield a waveform that will pass the eye-mask test that would otherwise fail with a wider-bandwidth receiver.

To understand the logic behind this, consider the main intent of the test. The usability of the transmitter in a real communications system needs to be verified. This objective is very different from trying to produce the most accurate image of the waveform shape. Most communication systems have receivers with bandwidths just wide enough to allow accurate determination of input signal levels. If the receiver bandwidth is wide, internal noise will increase, and likely degrade system BER. If the receiver bandwidth is too low, a signal making a 0-to-1 transition will be sluggish and not reach full amplitude within the bit period. Somewhere in between these extremes is an ideal receiver bandwidth. Although communication system receivers are not normally specified to have a "75% of bit rate" bandwidth, they do not normally have

relatively wide bandwidths. Thus the reference receiver measurement approach has proved to be an effective way to verify transmitter performance.

BIOGRAPHY

Greg D. LeCheminant received the B.S. degree in 1983 in Electronics Engineering Technology and M.S. degree in 1984 in Electrical Engineering from Brigham Young University in Provo, Utah. He joined the Hewlett-Packard company in 1985 as a manufacturing development engineer working in the production of microwave subassemblies for high-performance signal generators. In 1989 he accepted a marketing position for development of instrumentation used for high-speed optoelectric device and component characterization. He is currently employed by Agilent Technologies (Formerly Hewlett-Packard Test and Measurement) involved in the development of communications industry standards and measurement techniques and applications in high-speed digital communications.

THRESHOLD DECODING

WILLIAM W. WU
JAMES L. MASSEY
Consultare Technology Group
Bethesda, Denmark

1. INTRODUCTION

Threshold decoding was one of the earliest practical techniques introduced for the decoding of linear error-correcting codes (or "parity-check codes"). Before explaining threshold decoding in general, we give two simple examples to illustrate its main features. Throughout this article, we consider only binary codes — partly for simplicity but also because threshold decoding is not well suited to the decoding of nonbinary codes.

Example 1. Consider the encoder for the binary linear $(n = 6,\ k = 3)$ code in which the length $n = 6$ binary code word $\mathbf{v} = [v_1\ v_2\ v_3\ v_4\ v_5\ v_6]$ is determined from the length $k = 3$ binary information sequence $\mathbf{u} = [u_1\ u_2\ u_3]$ by

$$[v_1\ v_2\ v_3\ v_4\ v_5\ v_6] = [u_1\ u_2\ u_3]\begin{bmatrix}1 & 0 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 1 & 0 & 1\\ 0 & 0 & 1 & 1 & 1 & 0\end{bmatrix}$$

where binary arithmetic, that is, arithmetic modulo two in which $1 \oplus 1 = 0$, is understood. The above 3×6 matrix is called a *generator matrix* for the code. This generator matrix $\mathbf{G}$ specifies a *systematic encoder* in the sense that the information bits appear unchanged within the code word, viz. in the first $k = 3$ positions, as follows from the fact that $\mathbf{G} = [\mathbf{I}_3 \vdots \mathbf{P}]$ where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. From this matrix equation, we see that $v_1 = u_1$, that $v_3 \oplus v_5 = u_3 \oplus (u_1 \oplus u_3) = u_1$, and that $v_2 \oplus v_6 = u_2 \oplus (u_1 \oplus u_2) = u_1$. One says that these three sums of encoded bits are *orthogonal on the information bit* u_1 in the sense that each sum is equal to u_1 plus one or more encoded bits, but no encoded bit appears in more than one of the sums. Suppose now that the code word $\mathbf{v}$ is transmitted over a binary symmetric channel (BSC) with crossover probability p where $0 < p < 1/2$. Then the length $n = 6$ binary received word $\mathbf{r}$ can be written as $\mathbf{r} = [r_1\ r_2\ r_3\ r_4\ r_5\ r_6] = \mathbf{v} \oplus \mathbf{e}$ where $\mathbf{e} = [e_1\ e_2\ e_3\ e_4\ e_5\ e_6]$ is the binary error pattern, each of whose bits independently has probability p of being 1. Because $r_i = v_i \oplus e_i \neq v_i$ if and only if $e_i = 1$, the Hamming weight of $\mathbf{e}$, that is, the number of its nonzero components, is the actual number of errors that occurred during the transmission of $\mathbf{v}$ over the BSC. If we now form the above three sums orthogonal on u_1 using the received bits in place of the transmitted bits, we obtain $r_1 = v_1 \oplus e_1 = u_1 \oplus e_1$, $r_3 \oplus r_5 = v_3 \oplus e_3 \oplus v_5 \oplus e_5 = u_1 \oplus e_3 \oplus e_5$, and $r_2 \oplus r_6 = v_2 \oplus e_2 \oplus v_6 \oplus e_6 = u_1 \oplus e_2 \oplus e_6$. These three sums of received bits are orthogonal on the information bit u_1 in the sense that each sum is equal to u_1 plus one or more error bits, but no error bit appears in (or "corrupts") more than one of these sums. It follows that if there is at most one actual error, that is, if at most one of the error bits is a 1, then the information bit u_1 can be correctly found at the receiver as the majority vote of r_1, $r_3 \oplus r_5$, and $r_2 \oplus r_6$. Entirely similar arguments show that, again if at most one of the error bits is a 1, the information bit u_2 can also be correctly found by taking the majority vote of r_2, $r_3 \oplus r_4$, and $r_1 \oplus r_6$ and that the information bit u_3 can be correctly found by the majority vote of r_3, $r_1 \oplus r_5$, and $r_2 \oplus r_4$. This manner of decoding is an example of *majority decoding*, the earliest and simplest form of "threshold decoding." Because majority decoding in this example corrects all single errors, the minimum distance $d_{\min}$ of the code must be at least 3. That the minimum distance is exactly three can be seen from the fact that the information sequence $\mathbf{u} = [1\ 0\ 0]$ gives the code word $\mathbf{v} = [1\ 0\ 0\ 0\ 1\ 1]$ with Hamming weight 3, that is, at distance 3 from the all-zero code word.

From this example, one deduces that if, for each information bit in a binary linear code, a set of δ sums of encoded bits orthogonal on this bit can be formed, then all patterns of $(\delta - 1)/2$ or fewer errors can be corrected by majority decoding. This implies that the minimum distance $d_{\min}$ of the code is at least δ. One says that the code can be *completely orthogonalized* if $d_{\min} = \delta$.

Example 2. Consider the encoder for the (7, 4) binary cyclic code in which the code word $\mathbf{v} = [v_1\ v_2\ v_3\ v_4\ v_5\ v_6\ v_7]$ is determined from the information sequence $\mathbf{u} = [u_1\ u_2\ u_3\ u_4]$ by

$$[v_1\ v_2\ v_3\ v_4\ v_5\ v_6\ v_7] = [u_1\ u_2\ u_3\ u_4]\begin{bmatrix}1 & 0 & 0 & 0 & 1 & 1 & 0\\ 0 & 1 & 0 & 0 & 0 & 1 & 1\\ 0 & 0 & 1 & 0 & 1 & 1 & 1\\ 0 & 0 & 0 & 1 & 1 & 0 & 1\end{bmatrix}$$

This is the (7, 4) Hamming code with $d_{\min} = 3$. It is easy to verify that it is impossible to form three sums of encoded bits that are orthogonal on any one of the four information bits. However, we note that $v_2 \oplus v_3 = u_2 \oplus u_3$, $v_1 \oplus v_6 = u_2 \oplus u_3$ and $v_4 \oplus v_7 = u_2 \oplus u_3$ so that one can form three sums of encoded bits that

are orthogonal on the sum $u_2 \oplus u_3$ of information bits. If the code word $\mathbf{v}$ is transmitted over a BSC, majority decoding of the three corresponding sums of received bits will determine the sum $u_2 \oplus u_3$ correctly when at most one actual error occurs in transmission. Let $(u_2 \oplus u_3)^{\Delta}$ denote the decoding decision for $u_2 \oplus u_3$. Then $v_6 \oplus (u_2 \oplus u_3)^{\Delta} = u_1 \oplus (u_2 \oplus u_3) \oplus (u_2 \oplus u_3)^{\Delta}$ is equal to u_1 when the decoding decision is correct. It follows that $v_1 = u_1$, $v_3 \oplus v_4 \oplus v_5 = u_1$ and $v_6 \oplus (u_2 \oplus u_3)^{\Delta} = u_1$ constitute a set of three *sums of encoded bits and previously decoded bits* that are orthogonal on the information bit u_1 when the previous decoding decision is correct. Hence, u_1 can now be determined correctly by majority decoding when at most one actual error occurs in transmission. Because the code is cyclic, the information bit u_2 can similarly be determined simply by increasing the indices cyclically (i.e., increasing $n = 7$ gives 1) in the expression used to determine u_1, and similarly for u_3 and u_4. Thus, majority decoding of this code corrects all single errors. This is in fact complete minimum-distance decoding because, in a Hamming code, every received word is at distance at most 1 from some code word.

This decoding of a Hamming code exemplifies majority decoding with L-*step orthogonalization* for $\mathrm{L} = 2$. The number of steps is the number of levels of decoding decisions (all of which use at least δ orthogonal sums) required to determine an information bit. We can deduce from this example that if, for each information bit in a binary linear code, a set of δ sums of encoded bits and previously decoded bits orthogonal on this information bit can be formed by L-step orthogonalization, then all patterns of $(\delta - 1)/2$ or fewer errors can be corrected by majority decoding and hence the minimum distance $d_{\min}$ of the code is at least δ. One says that the code can be L-*step orthogonalized* if $d_{\min} = \delta$. Note that 1-step orthogonalization is the same as complete orthogonalization as defined above.

2. EARLY HISTORY

The decoding algorithm given by Reed [1] in 1954 for the binary linear codes found two years earlier by Muller, which are now universally called the Reed-Muller codes, was the first application of majority decoding to multiple-error-correcting codes. The μ^{th} order Reed-Muller code of length $n = 2^m$, where $0 \le \mu < m$, has $k = \sum_{i=0}^{\mu} \binom{m}{i}$ information bits and minimum distance $d_{\min} = 2^{m-1}$. Reed showed that this code can be $(\mu + 1)$-step orthogonalized (although this terminology did not come into use until eight years later).

Yale [2] and Zierler [3] independently showed in 1958 that the binary maximal-length codes could be completely orthogonalized. Mitchell et al. [4] made an extensive study of majority decoding of binary cyclic codes in 1961 that included majority decoding algorithms for the Hamming codes, for the $(n = 73,\ k = 45,\ d_{\min} = 9)$ and the $(n = 21,\ k = 11,\ d_{\min} = 5)$ cyclic codes found by Prange [5], and for the $(n = 15,\ k = 7,\ d_{\min} = 5)$ Bose-Chaudhuri-Hocquenghem (BCH) code.

In his 1962 doctoral thesis at M.I.T., which appeared essentially unchanged as a monograph [6] the following year, Massey formulated threshold decoding as a general technique comprising both majority decoding and also what he called *APP decoding*, where "APP" is short for "a posteriori probability."

In APP decoding, one again uses orthogonal sums on some bit (or sum of bits), but now one takes into account the *probability* that each individual sum will give an erroneous value for that bit (or some of bits). For instance, in Example 1, the "sum" $r_1 = u_1 \oplus e_1$ has probability p of giving an erroneous value of u_1, where p is the crossover probability on the BSC, that is, $\Pr(e_i = 1) = p$ for all i. However, the sum $r_3 \oplus r_5 = u_1 \oplus e_3 \oplus e_5$ has probability $2p(1-p)$ of giving an erroneous value of u_1 because it gives an erroneous value only when exactly one of e_3 and e_5 is an actual error, that is, has value 1.

APP decoding permits the use of soft-decision demodulation as opposed to the hard-decision demodulation that produces the BSC. In soft-decision demodulation, each received bit r_i is tagged by the demodulator with its probability p_i of being erroneous. With soft-decision demodulation the "sum" $r_1 = u_1 \oplus e_1$ in Example 1 has the probability p_1 of giving an erroneous value of u_1, whereas the sum $r_3 \oplus r_5 = u_1 \oplus e_3 \oplus e_5$ has probability $p_3(1-p_5) + p_5(1-p_3)$ of giving an erroneous value of u_1.

It is straightforward to show, cf. Ref. 6, that if $B_1, B_2, \ldots, B_\delta$ are the values of δ sums of received bits orthogonal on the information bit u (or on a sum of information bits), then the decision rule for u (or for the sum of information bits) that minimizes error probability from observation of $B_1, B_2, \ldots, B_\delta$ is: choose $u = 1$ (or choose the sum of information bits equal to 1) if and only if

$$\sum_{i=1}^{\delta} w_i B_i > T$$

where the value of B_i is treated as a real number in this sum, where the weighting factor w_i is given by $w_i = 2 \log \frac{1-P_i}{P_i}$ wherein P_i is the probability that B_i gives an erroneous value for u, where the threshold T is given by $T = \frac{1}{2} \sum_{i=1}^{\delta} w_i$, and where the information bits are assumed to be independent and equally likely to be 0 or 1. Note that the decision rule for majority decoding can be written in this form by taking $w_i \equiv 1$. This motivated Massey [6] to introduce the term *threshold decoding* to describe both majority decoding and APP decoding. Majority decoding can be considered as a nonoptimum but simple approximation to APP decoding.

3. THRESHOLD DECODING WITH PARITY CHECKS

It is often convenient to formulate the orthogonal sums discussed above in terms of the parity-check equations of the binary linear code. A *parity check* can be defined as a sum of error bits whose value can be calculated exactly from the received code word, which is equivalent to saying that the sum of the corresponding encoded bits must be zero.

Example 3. Recall that, for the $(n = 6,\ k = 3)$ code of Example 1, the three sums of received bits orthogonal on the information bit u_1 were $r_1 = u_1 \oplus e_1$, $r_3 \oplus r_5 = u_1 \oplus e_3 \oplus e_5$, and $r_2 \oplus r_6 = u_1 \oplus e_2 \oplus e_6$. If we now add $r_1 = u_1 \oplus e_1$ to each of these sums, we obtain $0 = e_1 \oplus e_1$, $r_1 \oplus r_3 \oplus r_5 = e_1 \oplus e_3 \oplus e_5$, and $r_1 \oplus r_2 \oplus r_6 = e_1 \oplus e_2 \oplus e_6$, which constitute three *parity checks orthogonal on the error bit* e_1 in the sense that each parity check is equal to e_1 plus one or more corrupting error bits, but no error bit corrupts more than one of these parity checks. (Note that e_1 itself corrupts the first of these three parity checks, viz. the trivial parity check 0.) Thus, e_1 will be correctly given by the majority vote of these three parity checks if there is at most one actual error in transmission. Because the first parity check always votes for 0, one can ignore this parity check and say that one decides that e_1 is a 1 if and only if both the second and third parity checks are 1.

One can infer from this example that $\delta - 1$ nontrivial parity checks orthogonal on an error bit (or a sum of error bits) are equivalent to δ sums of received bits orthogonal on an information bit (or a sum of information bits). Moreover, if $A_1, A_2, \ldots, A_{\delta-1}$ are the values of $\delta - 1$ nontrivial parity checks orthogonal on the error bit e (or on a sum of error bits), then the APP decoding rule for e (or for the sum of error bits) from observation of $A_1, A_2, \ldots, A_{\delta-1}$ becomes: decide $e = 1$ (or decide that the sum of error bits is equal to 1) if and only if

$$\sum_{i=1}^{\delta-1} w_i A_i > T$$

where the value of A_i is treated as a real number in this sum, where the weighting factor w_i is given by $w_i = 2 \log \frac{1-P_i}{P_i}$ wherein P_i is the probability that A_i gives an erroneous value for u, where the threshold T is given by $T = \frac{1}{2} \sum_{i=0}^{\delta-1} w_i$, and where $w_0 = 2 \log \frac{\Pr(e=0)}{\Pr(e=1)}$. The decision rule for majority decoding is obtained by taking $w_i \equiv 1$. For $\delta - 1 = 2$ as in Example 3, the majority decoding rule is decide $e = 1$ if and only if $A_1 + A_2 > 3/2$ or, equivalently, if and only if $A_1 = A_2 = 1$.

A *(reduced) parity-check matrix* for an (n, k) linear code is any $(n - k) \times n$ matrix $\mathbf{H}$ such that $\mathbf{v} = [v_1\ v_2\ \cdots\ v_n]$ is a code word if and only if $\mathbf{v}\mathbf{H}^T = \mathbf{0}$ where the superscript "T" denotes transpose. Again let $\mathbf{r} = \mathbf{v} \oplus \mathbf{e}$ where $\mathbf{r}$ is the binary received word and $\mathbf{e}$ is the binary error pattern. Then the *syndrome* $\mathbf{s}$ of $\mathbf{r}$ relative to the parity-check matrix $\mathbf{H}$ is defined as $\mathbf{s} = \mathbf{r}\mathbf{H}^T$. It follows that $\mathbf{s} = (\mathbf{v} \oplus \mathbf{e})\mathbf{H}^T = \mathbf{e}\mathbf{H}^T$ and hence that the syndrome bits are parity checks. In fact, every parity check is either a syndrome bit or a sum of syndrome bits.

If the systematic generator matrix $\mathbf{G} = [\mathbf{I}_k \vdots \mathbf{P}]$ is used for encoding, then the information bits $[u_1\ u_2\ \cdots\ u_k] = [v_1\ v_2\ \cdots\ v_k]$ determine the "parity bits" $[v_{k+1}\ v_{k+2}\ \cdots\ v_n]$ of the code word as $[v_{k+1}\ v_{k+2}\ \cdots\ v_n] = [v_1\ v_2\ \cdots\ v_k]\mathbf{P}$. Thus, $\mathbf{v}$ is a code word if and only if $[v_{k+1}\ v_{k+2}\ \cdots\ v_n] \oplus [v_1\ v_2\ \cdots\ v_k]\mathbf{P} = 0$ or, equivalently, if and only if $\mathbf{v}[\mathbf{P}^T \vdots \mathbf{I}_{n-k}] = \mathbf{0}$. This shows that $\mathbf{H} = [\mathbf{P}^T \vdots \mathbf{I}_{n-k}]$ is a (reduced) parity-check matrix that we will call the *systematic parity-check matrix* of the code. Relative to this parity-check matrix, the syndrome becomes $\mathbf{s} = [r_{k+1}\ r_{k+2}\ \cdots\ r_n] \oplus [r_1\ r_2\ \cdots\ r_k]\mathbf{P}$. This shows the very useful fact that the syndrome relative to the systematic parity-check matrix can be formed by adding the received parity digits to the parity digits computed from the received information bits.

Example 4. For the $(n = 6,\ k = 3)$ code of Example 1, the systematic parity-check matrix is

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The syndrome bits $[s_1\ s_2\ s_3] = \mathbf{r}\mathbf{H}^T$ are given by $s_1 = r_2 \oplus r_3 \oplus r_4 = e_2 \oplus e_3 \oplus e_4$, $s_2 = r_1 \oplus r_3 \oplus r_5 = e_1 \oplus e_3 \oplus e_5$, and $s_3 = r_1 \oplus r_2 \oplus r_6 = e_1 \oplus e_2 \oplus e_6$. We see that s_2 and s_3 are the $\delta - 1 = 2$ nontrivial parity checks orthogonal on e_1 that we exploited in Example 4. We also note that s_1 and s_3 are two nontrivial parity checks orthogonal on e_2, and that s_1 and s_2 are two nontrivial parity checks orthogonal on e_3.

4. THRESHOLD DECODING OF CONVOLUTIONAL CODES

4.1. Preliminaries

Massey [6] formulated the first threshold decoders for multiple-error-correcting convolutional codes. The simplicity of threshold decoders for convolutional codes has led them to dominate practical applications of threshold decoding. We begin here with a brief discussion of those aspects of convolutional codes that are needed to understand threshold decoders for these codes.

In convolutional coding, the information sequences and encoded sequences are semi-infinite sequences that we will represent as *power series* in D. For instance, $U(D) = u_0 \oplus u_1 D \oplus u_2 D^2 \oplus \cdots$ where u_i is the information bit at "time" i. In an (n_o, k_o) convolutional code, there are k_o such information sequences $U_1(D), U_2(D), \ldots, U_{k_o}(D)$, and n_o corresponding encoded sequences $V_1(D), V_2(D), \ldots, V_{n_o}(D)$. The encoded sequences are the result of passing the information sequences through a k_o-input/n_o-output binary finite-state linear system. An example will clarify matters.

Example 5. Consider the $(n_o = 2,\ k_o = 1)$ convolutional code with the systematic encoder $\mathbf{G(D)} = [1 \vdots P(D)]$ where $P(D) = 1 \oplus D \oplus D^4 \oplus D^6$. The (single) information sequence $U_1(D) = U(D)$ yields the code word $[U(D) \vdots U(D)P(D)]$, whose two encoded sequences are multiplexed together for transmission over a single channel. For simplicity of notation, let $V(D) = v_0 \oplus v_1 D \oplus v_2 D^2 \oplus \cdots$ denote the sequence of "parity bits" produced by the systematic encoder. Because $V(D) = U(D)P(D) = U(D)(1 \oplus D \oplus D^4 \oplus D^6)$, we see that $v_i = u_i \oplus u_{i-1} \oplus u_{i-4} \oplus u_{i-6}$ for all $i \geq 0$ where it is understood that $u_j = 0$ if $j < 0$. Because the syndrome relative to the systematic parity-check matrix can be formed by adding the received parity digits to the parity digits computed from the received

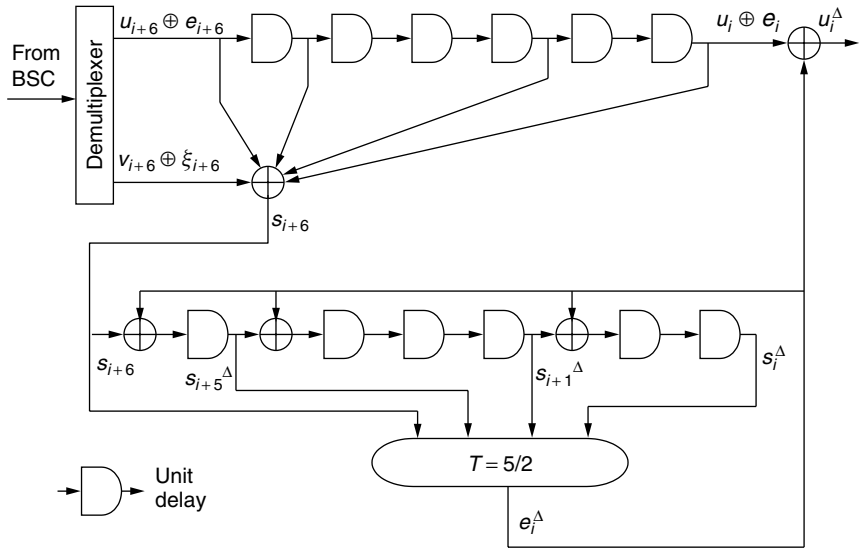


Figure 1. A majority decoder for the (2, 1) convolutional code of Example 5.

information bits, it follows that the syndrome sequence $S(D) = s_0 \oplus s_1 D \oplus s_2 D^2 \oplus \cdots$ can be formed by the simple "linear filter" with a memory of 6 bits shown in Fig. 1. Letting $E(D) = e_0 \oplus e_1 D \oplus e_2 D^2 \oplus \cdots$ and $\Xi(D) = \xi_0 \oplus \xi_1 D \oplus \xi_2 D^2 \oplus \cdots$ denote the error sequences in the information sequence and in the parity-digit sequence, respectively, we further see that the syndrome bits are given by $s_i = e_i \oplus e_{i-1} \oplus e_{i-4} \oplus e_{i-6} \oplus \xi_i$. In particular, we see that $s_6 = e_6 \oplus e_5 \oplus e_2 \oplus e_0 \oplus \xi_6$, $s_4 = e_4 \oplus e_3 \oplus e_0 \oplus \xi_4$, $s_1 = e_1 \oplus e_0 \oplus \xi_1$, and $s_0 = e_0 \oplus \xi_0$ are a set of $\delta - 1 = 4$ parity checks orthogonal on e_0. Thus, if there are two or fewer actual errors among the 11 error bits that enter into these parity checks, e_0 will be correctly given by majority decoding with threshold $T = 5/2$, i.e., ${e_0}^{\Delta} = 1$ if and only if three or more of these parity checks have value 1. One says that the *effective constraint length* of the convolutional code is $n_E = 11$ bits. We can then feed ${e_0}^{\Delta}$ back to remove e_0 from s_6, s_4, and s_1, following which s_7, s_5, s_2 and ${s_1}^{\Delta} = s_1 \oplus {e_0}^{\Delta}$ become a set of 4 parity checks orthogonal on e_1 (on the assumption that ${e_0}^{\Delta} = e_0$). Figure 1 shows the complete double-error correcting majority decoder for the code of this example.

4.2. Early Applications

Codex Corporation was founded in Cambridge, MA, in 1962 as the first organization dedicated solely to the practical application of information-theoretic research. The two innovations that the fledgling company hoped to exploit were Massey's threshold decoders for convolutional codes and Gallager's low-density parity-check (LDPC) codes, the latter a product of a 1960 M.I.T. doctoral thesis and the subject of a 1963 monograph [7]. LDPC codes were "ahead of their day" and defied realization with the discrete logic available in the 1960s. But LDPC codes with Gallager's iterative decoding algorithm have become a hot topic in recent years — reliable transmission at rates extremely close to the capacity of a Gaussian channel has been achieved. Threshold decoders for convolutional codes, however, exhibit a simplicity (cf. Fig. 1) that was virtually ideally suited to realization with the discrete logic available in the 1960s. For this reason, threshold decoders became the first product of Codex Corporation and remained their mainstay product for many years — even though Massey [6] had shown that capacity could not be closely approached with such decoders. Simplicity trumped asymptotics in the 1960s.

It was soon realized at Codex Corporation that most potential applications for threshold decoders were for "bursty channels" in which the actual errors tend to cluster, as opposed to the BSC where the actual errors are scattered. Again convolutional codes were well suited to this problem. Kohlenberg and Forney [8] found by increasing properly the degrees of the nonzero terms in the encoding polynomial $P(D)$ (in the notation of Example 5), a majority decoder acting on the resulting orthogonal parity-checks corrected not only all patterns of $(\delta - 1)/2$ or fewer actual errors among the error bits appearing in these parity checks, but also corrected all bursts of some large length or less. These so-called "diffuse" threshold decoders were in fact the principal coding product of Codex Corporation in the 1960s.

4.3. Self-Orthogonal Codes

A *convolutional self-orthogonal code* (CSOC) was defined by Massey [6] as a convolutional code with the property that, when the systematic parity-check former is used, all the syndrome bits that check a time-0 error in an information bit are orthogonal on that error bit. Macy [9] and Hagelbarger [10] independently developed majority decoding procedures for CSOCs.

Macy [9] and Robinson et al. [11] independently made the important connection between CSOCs and difference sets. Let $\{i_0, i_i, \ldots, i_q\}$ be a set of $q + 1$ nonnegative integers. We will say that $\{i_0, i_i, \ldots, i_q\}$ is a *distinct-differences set* if the $(q + 1)q$ differences of ordered pairs of integers in the set are all distinct. For example, $\{0, 1, 3\}$ is a distinct-differences set because the $(q + 1)q = 3 \cdot 2 = 6$ differences $3 - 0 = 3$, $3 - 1 = 2$, $1 - 0 = 1$, $0 - 3 = -3$, $1 - 3 = -2$, and $0 - 1 = -1$, are all distinct. A distinct-differences set $\{i_0, i_i, \ldots, i_q\}$ is said to be a *perfect difference*

set (or a *planar difference set*) if its $(q+1)q$ differences are all distinct and nonzero when taken modulo $(q+1)q+1$. For example, $\{0,\ 1,\ 3\}$ is a perfect difference set because its 6 differences taken modulo 7 are 3, 2, 1, 4, 5, and 6. The distinct-differences set $\{0,\ 2,\ 6\}$ is also a perfect difference set since its 6 differences are 6, 4, 2, -6, -4 and -2, which taken modulo 7 give 6, 4, 2, 1, 3, and 5. However, the distinct-differences set $\{0,\ 1,\ 4\}$ is not a perfect difference set because $4-1=3$ and $0-4=-4$ are both 3 when taken modulo 7. Perfect difference sets of $q+1$ integers are known to exist whenever q is a power of a prime [12] and perhaps only then.

The convolutional code with encoding polynomial $P(D)=1\oplus D\oplus D^4\oplus D^6$ in Example 5 is a CSOC. The set $\{0,\ 1,\ 4,\ 6\}$ of powers of D appearing in this polynomial is a *perfect difference set*, that is, the $(q+1)q=4\cdot 3=12$ differences between ordered pairs of numbers in this set are all distinct and nonzero modulo $(q+1)q+1=13$.

The general result embodied in Example 5 is that an $(n_o=2,\ k_o=1)$ convolutional code with systematic encoder $\mathbf{G(D)}=[1\ \vdots\ P(D)]$ is a CSOC just when the set of powers of D appearing in $P(D)$ is a distinct-differences set. One usually desires that the degree of $P(D)$, which is the decoding delay of the threshold decoder and determines the number of delay elements therein, be as small as possible. We point out that the set $\{0,\ 2,\ 4,\ 10\}$ is also a perfect difference set but would be a poorer choice for the degrees of the terms in $P(D)$ than $\{0,\ 1,\ 4,\ 6\}$ because it would give a decoding delay of 10 rather than 6 for the same error-correcting capability. The decoding delay of a CSOC is generally minimized when the distinct-differences set is a perfect difference set, but it takes some skill to find the best perfect difference set. We have restricted our discussion here to $(n_o=2,\ k_o=1)$ CSOCs, but distinct-differences sets and perfect difference sets also can be used to describe and construct (n_o, k_o) CSOCs with $k_o>1$ and/or with $n_o>2$, cf. Ref. 11.

The decoder of Fig. 1 is a so-called *feedback decoder* in which decoded information bits are fed back to remove their effect from the syndrome bits used in future decoding decisions. A decoder without such feedback is called a *definite decoder* [11] for the convolutional code. Hagelbarger [10] and Robinson et al. [11] independently observed that the feedback could be removed from the majority decoder for a CSOC without reducing the number of errors guaranteed correctable but at the expense of enlarging the effective constraint length n_E. The reason for this is that all the syndrome bits that check each information error bit in a CSOC remain orthogonal even without the removal of past decoded error bits. Recall that in Example 5 the syndrome bits are given by $s_i=e_i\oplus e_{i-1}\oplus e_{i-4}\oplus e_{i-6}\oplus\xi_i$ for all $i\geq 0$. Thus, the syndrome bits that check e_i are $s_i=e_i\oplus e_{i-1}\oplus e_{i-4}\oplus e_{i-6}\oplus\xi_i$, $s_{i+1}=e_{i+1}\oplus e_i\oplus e_{i-3}\oplus e_{i-5}\oplus\xi_{i+1}$, $s_{i+4}=e_{i+4}\oplus e_{i+3}\oplus e_i\oplus e_{i-2}\oplus\xi_{i+4}$, and $s_{i+6}=e_{i+6}\oplus e_{i+5}\oplus e_{i+2}\oplus e_i\oplus\xi_{i+6}$, which we see are $\delta-1=4$ parity checks orthogonal on e_i with effective constraint length $n_E=17$. The decoder of Fig. 1 is converted to a definite decoder simply by removing the feedback of $e_i{}^{\Delta}$ to the three modulo-two adders in the "syndrome register." The use of a definite decoder eliminates entirely the *error propagation* that results when incorrect decisions are fed back in a feedback decoder. However, the more important fact that this error propagation is very slight for CSOCs was shown in Ref. 11. In virtually all applications of threshold decoding (not only for CSOCs), the performance of the feedback decoder is substantially better than that of the definite decoder.

The (dimensionless) rate R of a (n_o, k_o) convolutional code is defined as $R=k_o/n_o$. Note that $0<R\leq 1$. The bandwidth expansion of the code is $1/R$ so that high-rate codes are preferred in applications where bandwidth is restricted. Wu [13] made an extensive search for good high-rate CSOCs for use in satellite systems. His codes were extensively used in COMSAT and INTELSAT single-channel-per-carrier systems in the 1970s and 1980s in what was perhaps the most significant practical application of threshold decoding yet.

5. THRESHOLD DECODING OF BLOCK CODES

The development of threshold decoding techniques for convolutional codes led to parallel developments for block codes, but without the practical applications for which the convolutional coding systems were so well suited in the 1960s and 1970s. We describe some of these theoretical developments here.

A *quasi-cyclic code* is a $(n=Mk_o,\ k=Mn_o)$ linear code having generator matrices and parity-check matrices that can be partitioned into $M\times M$ blocks, each of which is a *circulant matrix*, that is, a square matrix each of whose rows after the first is the right cyclic shift of the previous row. The (6, 3) binary code of Example 4 is a quasi-cyclic code with $M=3$. A block self-orthogonal code (BSOC) can be defined as a binary linear code such that, when the systematic parity-check matrix is used, all the syndrome bits that check an error in an information bit are orthogonal on that error bit. The (6, 3) binary code of Example 4 is a BSOC. Townsend and Weldon [14] showed that difference sets play essentially the same role in determining cyclic BSOCs as they do in determining CSOCs. In the systematic parity-check matrix $\mathbf{H}=[\mathbf{P}^T\ \vdots\ \mathbf{I}_3]$ of Example 4, the first row [0 1 1] of the circulant matrix $\mathbf{P}^T$ corresponds to the polynomial $D+D^2$ having $\{1,\ 2\}$ as the set of powers of D appearing therein. But $\{1,\ 2\}$ is a distinct-differences set whose 2 differences $2-1=1$ and $1-2=-1$ are distinct modulo $M=3$. This is an illustration of the fact [14] that a binary $(n=2M,\ k=M)$ quasi-cyclic code is a BSOC if and only if, in its systematic parity-check matrix $\mathbf{H}=[\mathbf{P}^T\ \vdots\ \mathbf{I}_M]$, the set of powers of D appearing in the polynomial specified by the first row of the circulant matrix $\mathbf{P}^T$ is a distinct-differences set whose differences are also distinct when taken modulo M. Such a code can be completely orthogonalized. If there are $q+1$ powers of D in the distinct-differences set, then $\delta=d_{\min}=q+2$. Perfect difference sets with $(q+1)q+1=M$ are an obvious source for constructing good BSOCs of this type. For instance, using the perfect difference set $\{0,\ 1,\ 4,\ 6\}$ with $M=13$ gives a (26, 13) BSOC with $\delta=d_{\min}=5$. We have restricted our discussion

to quasi-cyclic BSOCs with $k_o = 1$ and $n_o = 2$, but distinct-differences sets and perfect difference sets also can be used to construct quasi-cyclic BSOCs with $k_o > 1$ and/or with $n_o > 2$, cf. Ref. 14.

Weldon [15] in 1967 showed perfect difference sets can also be used to design majority-decodable cyclic codes as we now explain. First we remark that the so-called *parity-check polynomial* $h(X) = X^k \oplus h_{k-1}X^{k-1} \cdots \oplus h_1 X \oplus 1$ of a (n, k) binary cyclic code, which divides the polynomial $X^n \oplus 1$, determines all the parity checks of the cyclic code in the manner that the binary n-tuple $[b_{n-1}\ b_{n-2}\ \cdots\ b_1\ b_0]$ corresponds to (the coefficients of $e_1,\ e_2,\ \cdots,\ e_n$ in) a parity check if and only if the polynomial $b(X) = b_{n-1}X^{n-1} \oplus b_{n-2}X^{n-2} \oplus \cdots \oplus b_1 X \oplus b_0$ is divisible by $h(X)$. For simplicity, we will refer to $[b_{n-1}\ b_{n-2}\ \cdots\ b_1\ b_0]$ itself as a parity check. Because the code is cyclic, every cyclic shift of a parity check is again a parity check.

Example 6. The set $\{0,\ 2,\ 3\}$ of $q + 1 = 3$ integers is a perfect difference set because the $(q + 1)q = 6$ differences are nonzero and distinct modulo $n = (q + 1)q + 1 = 7$. Set $b(X)$ equal to the polynomial whose powers of X correspond to this perfect difference set, that is, $b(X) = X^3 \oplus X^2 \oplus 1$. Now find the polynomial $h(X)$ of largest degree k that divides both $b(x)$ and $X^n \oplus 1$, that is, find the greatest common divisor of $b(x)$ and $X^n \oplus 1$. In this example, $b(x)$ itself divides $X^n \oplus 1 = X^7 \oplus 1$ so that $h(X) = b(X) = X^3 \oplus X^2 \oplus 1$. This $h(X)$ is the parity check polynomial of a cyclic (7, 3) code in which the 7-tuple [0 0 0 1 1 0 1] corresponding to $b(X)$ is a parity check. This 7-tuple and its left cyclic shifts by 1 and 3 positions, respectively, viz. [0 0 1 1 0 1 0] and [1 1 0 1 0 0 0], constitute a set of $\delta - 1 = 3$ parity checks orthogonal on e_4. Because the code is cyclic, a set of $\delta - 1 = 3$ parity checks orthogonal on every error bit can be formed by cyclic shifting of these 7-tuples.

Cyclic codes constructed as in Example 6 are called *difference-set cyclic codes* [15] and are always completely orthogonalizable. The particular code in Example 6 happens also to be a BSOC because $h(X) = b(x)$. It is also a maximal-length code; indeed, all maximal-length codes are difference-set cyclic codes and BSOCs. The (21,12) and (73, 45) codes mentioned in Section 2 are also difference-set cyclic codes.

The connection between certain threshold-decodable codes and *finite geometries*, both finite Euclidean geometries and finite projective geometries, was first noted by Rudolph in 1967 [16]. The ramifications of this connection have been of great importance in connection with the general theory of block codes, but of less importance in the practical application of threshold decoding because the finite-geometry codes with parameters suitable for practical implementation were mostly already known. There is a close relationship between perfect difference sets and finite geometries. In fact, Singer's proof [12] that perfect difference sets of $q + 1$ integers exist whenever q is a power of a prime is based on properties of finite projective geometries.

The codes obtained from finite geometries all rely on viewing the parity checks of the code as *incidence vectors* of some type, for example, as the incidence vectors for points on the lines of the geometry, in which case error bits are regarded as points and parity checks are regarded as lines of the geometry. Perhaps the key contribution was made by Kasami et al. [17] who showed that the Reed-Muller codes shortened by one bit were cyclic finite-geometry codes and subcodes of BCH codes.

6. RECENT DEVELOPMENTS

There has been a steady trickle of new results on threshold decoding since its heyday in the 1960s and 1970s. Threshold decoders continue to be employed in many applications of transmission or storage of information where simplicity and/or high speed in the decoder is a prerequisite. We close this article by mentioning two recent developments that suggest some promising new directions for threshold decoding.

Riedel and Svirid [18] investigated the use of "parallel concatenated" CSOCs in a "turbo-like" structure. They modified the APP decoding algorithm to obtain a soft-in soft-out APP decoder. Iterative decoding of the CSOCs with this algorithm yielded surprisingly good results for the Gaussian channel and for the Rayleigh fading channel.

Meier and Staffelbach [19] devised a cryptanalytic attack on certain stream ciphers that is equivalent to decoding very long maximal-length codes, which we recall from Section 5 are BSOCs, transmitted over a BSC whose crossover probability p is only slightly less than 1/2. Their attack, that is, their decoding algorithm, uses threshold decoding but applied to only a few of the possible parity checks orthogonal on each bit to be decoded (although their paper, which is written for cryptographers, does not state this). This process is iterated until the decoding becomes stable. This "partial decoding" is necessary because the attack must be made with only a small portion of the received code word available to the decoder. Recall from Section 2 that the code word length is $n = 2^m$ but that there are only $k = m$ information bits so that one needs to decode only m consecutive received bits to obtain the entire code word. Meier and Staffelbach showed that their attack finds the underlying code word with high probability for surprisingly large channel crossover probabilities. This suggests that their novel way of iterating the decoding of BSOCs may be a useful alternative in data transmission applications to Riedel and Svirid's [18] iterative decoding of CSOCs [18].

BIOGRAPHIES

William W. Wu is the founder of Advanced Technology Mechanization Company (ATMco). He initiated the Internet Domino Technologies and the Universal Transmission Techniques; NSF has supported both programs. He was a PI winner of a multiyear contract from NASA/Glenn Research Center for loss cell recovery in ATM. From 1992 to 1998 he headed the Consultare Group, as advisors to RAFAEL, Israel; TELESAT, Canada; Stanford Telecom in California, COMSAT; POLSPACE, Poland; Motorola; and Hyundai Electronic Industries, Korea. From 1989 to 1991, he was a director at Stanford Telecom. From 1978

to 1989 he was with INTELSAT's executive organ as chief scientist; senior MTS is Communication Engineering, R. and D. Departments. He also participated in *INTELSAT IV, IVA, V VI*, AND—*K*. From 1967 to 1978, he was a senior scientist at Advance System Div., COMSAT Corporate Headquarters; and at COMSAT Labs. He authored 2 books on digital satellite communications, and has contributed 2 books, 3 encyclopedia, and 60 technical papers. He was the editor of Satellite Communications and Error Coding, *IEEE Transactions On Communications*. He was also a professional lecture at George Washington University, Washington, D.C., (1977–1989, and 1999), and has given 39 Invited lectures, seminars, and/or short courses worldwide. He is an IEEE Fellow, was chairman for ITU-T SG-XVIII, and a MIT "Distinguished Alumnus" (1998). He has a Ph.D. from the Johns Hopkins University, Maryland MSEE from MIT, Cambridge, Massachusetts, and a BSEE from Purdue University West Lafayette, Indiana.

James L. Massey served on the faculties of the University of Notre Dame, Indiana (1962–1977), the University of California, Los Angeles (1977–1980), and the Swiss Federal Institute of Technology (ETH), Zürich (1980–1998), where he now hold emeritus status. He currently is an adjunct professor at the University of Lund, Sweden.

He has served the *IEEE Transactions on Information Theory* as editor and as associate editor for *Algebraic Coding* and the *Journal of Cryptology* as an associate editor. He is a past president of the IEEE Information Theory Society and of the International Association for Cryptologic Research. He was a founder of Codex Corporation (later a division of Motorola) and of Cylink Corporation, Santa Clara, California.

His awards include the 1988 Shannon Award of the IEEE Information Theory Society, the 1992 IEEE Alexander Graham Bell Medal, and the 1999 Marconi International Fellowship. He is a fellow of the IEEE, a member of the Swiss Academy of Engineering Sciences and the U.S. National Academy of Engineering, an honorary member of the Hungarian Academy of Science, and a foreign member of the Royal Swedish Academy of Sciences.

BIBLIOGRAPHY

1. I. S. Reed, A class of multiple-error-correcting codes and the decoding scheme, *IRE Trans. Inform. Theory* **IT-4**: 38–49 (1954).
2. R. B. Yale, Error correcting codes and linear recurring sequences, Rept. 34–77, M.I.T. Lincon Lab., Lexington, MA, 1958.
3. N. Zierler, On a variation of the first order Reed-Muller codes, Rept. 95, M.I.T. Lincon Lab., Lexington, MA, 1958.
4. M. Mitchell, R. Burton, C. Hackett, and R. Schwartz, Coding and operations research, Rept. on Contract AF19(604)-6183, General Electric, Oklahoma City, OK, 1961.
5. E. Prange, The use of coset equivalence in the analysis and decoding of group codes, AFCRC Rept. 59–164, Bedford, MA, 1959.
6. J. L. Massey, *Threshold Decoding*, M.I.T. Press, Cambridge, MA, 1963.
7. R. G. Gallager, *Low-Density Parity-Check Codes*, M.I.T. Press, Cambridge, MA, 1963.
8. A. Kohlenberg and G. D. Forney, Jr., Convolutional coding for channels with memory, *IEEE Trans. Inform. Theory* **IT-14**: 618–626 (1968).
9. J. R. Macy, *Theory of Serial Codes*, Ph.D. dissertation, Stevens Inst. Tech., Hoboken, NJ, 1963.
10. D. W. Hagelbarger, Recurrent codes for the binary symmetric channel, Lecture Notes, Summer Conf. Theor. Codes, Univ. Michigan, Ann Arbor, MI, 1962.
11. J. P. Robinson and A. J. Bernstein, A class of binary recurrent codes with limited error propagation, *IEEE Trans. Inform. Theory* **IT-13**: 106–113 (1967).
12. J. Singer, A theorem in finite projective geometry and some applications to number theory, *Trans. Amer. Math. Soc.* **43**: 377–385 (1938).
13. W. W. Wu, New convolutional codes, Part I, Part II and Part III, *IEEE Trans. Commun.* **COM-23**: 942–956 (1975), **COM-24**: 19–33 and 946–955 (1976).
14. R. L. Townsend and E. J. Weldon, Jr., Self-Orthogonal Quasi-Cyclic Codes, *IEEE Trans. Inform. Theory* **IT-13**: 183–195 (1967).
15. E. J. Weldon, Jr., Difference set cyclic codes, *Bell Syst. Tech. J.* **45**: 1045–1055 (1966).
16. L. D. Rudolph, A class of majority logic decodable codes, *IEEE Trans. Inform. Theory* **IT-13**: 305–307 (1967).
17. T. Kasami, S. Lin, and W. W. Peterson, New generalizations of the Reed-Muller codes–Part I: Primitive codes, *IEEE Trans. Inform. Theory* **IT-14**: 189–199 (1968).
18. S. Ried and Y. V. Svirid, Iterative ("turbo") decoding of threshold decodable codes, *European Trans. Telecomm.* **6**: 527–534 (1995).
19. W. Meier and O. Staffelbach, Fast correlation attacks on certain stream ciphers, *J. Cryptology* **1**: 159–176 (1989).

TIME DIVISION MULTIPLE ACCESS (TDMA)

PETER JUNG
Gerhard-Mercator-Universität
Duisburg
Duisburg, Germany

1. INTRODUCTION

Communication, stemming from the Latin word for "common," is a most important desire of humankind. The combination of communication with mobility has accelerated the evolution of society worldwide, particularly during the past decade.

The history of mobile radio communication, however, is still young and dates back to the discovery of electromagnetic waves by the German physicist Heinrich Hertz in the nineteenth century. About 100 years ago, Guglielmo Marconi showed that long-haul wireless communication was technically possible using the radio principle based on Hertz's discovery and anticipating what we know as mobile radio today [1].

With the invention of the cellular principle in the early 1970s by engineers of AT&T Bell Labs, the basis for cellular mobile radio systems with high capacity was set [1]. In the early 1980s, the first commercial and civil mobile communication systems like the AMPS (American Mobile Phone Service), the NMT (Nordic Mobile Telecommunication), and the German C450 were introduced, allowing several hundreds of thousands of subscribers [2].

However, technology could not provide digital signal processing at a reasonable degree. Hence, multiple access had to be FDMA (frequency division multiple access). Although FDMA is indispensable for the planning of mobile radio networks, it has some technological drawbacks that led to high-priced base stations and cell phones [2].

However, during the past 20 years, the technological evolution provided us with unprecedented technological possibilities that help to overcome drawbacks of the early mobile radio systems:

- The development of digital signal processing became more and more mature.
- The integration density of microelectronic circuits increased beyond expected limits, providing increasing processing power in small ICs with low power consumption.

In mobile radio, more freedom of choice of the multiple access scheme could be exploited to invent mobile radio systems with more flexibility, higher capacity, and still lower price than the first generation, which relied on FDMA alone.

An increase in system capacity requires base stations that can handle an increased number of traffic channels. With respect to a reduced implementation complexity of the elaborate radiofrequency design of base stations, it is required to support several traffic channels per carrier. Furthermore, to realize radiofrequency front ends with low complexity in cell phones, forward and reverse links should be separated in time. Therefore, TDMA (time division multiple access) provides these assets and hence became the choice for the second generation of mobile radio. Nonetheless, radio network planning remains an important requirement. Hence, TDMA had to be combined with the well-established FDMA, resulting in a hybrid multiple access scheme that is often termed F/TDMA (frequency divided time division multiple access) [3]. These ideas and their further evolutions are considered in what follows.

This article is structured as follows: In Section 2, multiple access principles and hybrid multiple access schemes, which are feasible in mobile radio, are discussed. Section 3 presents signal and system structures used in TDMA systems. The author gives a brief discussion of important TDMA systems for mobile communication in Section 4. An evolution of TDMA toward the third generation of mobile communications, termed UMTS (universal mobile telecommunication system), is presented in Section 5. Section 6 presents concluding remarks.

2. MULTIPLE ACCESS PRINCIPLES AND HYBRID MULTIPLE ACCESS SCHEMES

In Table 1 [3], an overview of the four classic multiple access principles FDMA, TDMA, CDMA (code division multiple access) and SDMA (space division multiple access) is presented. Besides the basic concepts of these multiple access principles, the cellular aspect and an overall evaluation are offered in Table 1.

TDMA, CDMA, and SDMA have become feasible with the introduction of digital technology. However, CDMA was not well understood in civil communications engineering in the 1980s. Only after considerable research effort undertaken in the past two decades, has CDMA been identified as a superb means for mobile multimedia and is therefore deployed in UMTS [4,5]. With the exception of the well-known ANSI/TIA-95, second-generation mobile radio systems did not use CDMA.

In the context of multiple access, SDMA, which requires still expensive smart antennas, has not yet been deployed. However, it is anticipated that SDMA-type technologies will be introduced in upcoming releases of UMTS. SDMA can be regarded as a natural extension of the other three multiple access principles and thus is not considered separately in this article.

According to Table 1, all multiple access principles have specific advantages and drawbacks. To benefit from their advantages and to alleviate the effect of the drawbacks, a combination of multiple access principles, resulting in hybrid multiple access schemes, is recommended.

Considering FDMA, TDMA, and CDMA, four hybrid multiple access schemes are conceivable [3,4], namely, the aforementioned F/TDMA, F/CDMA (frequency divided code division multiple access), T/CDMA (time divided code division multiple access) and F/T/CDMA (frequency and time divided code division multiple access), cf. Fig. 1.

As already discussed, F/TDMA has been chosen for most second-generation mobile radio systems except ANSI/TIA-95, which deploys F/CDMA. Since T/CDMA does not support radio network planning, it has not yet been taken into account for communication systems. F/T/CDMA has been identified as a viable multiple access scheme for the UTRA TDD (UMTS terrestrial radio access time domain duplex) mode. This mode is also known as TD/CDMA and presents an extension of F/TDMA (cf. Section 5).

In what follows, we assume that TDMA is always combined with FDMA, resulting in F/TDMA for the reasons presented in this section. Since TDMA is the most significant part of this hybrid multiple access scheme, the expression TDMA refers to F/TDMA in the sequel.

3. SIGNAL AND SYSTEM STRUCTURES FOR TDMA

3.1. Physical Layer Subscriber Signal Structures

The physical layer subscriber signals carry the data sequences, which shall be transmitted to the receiver. These data sequences consist of encoded subscriber data, which can be any type of information stemming from higher layers (i.e., layers above the physical layer). These subscriber data could for example, be digitally encoded

Table 1. Comparison of Multiple Access Principles [3]

	Multiple Access Principle			
	FDMA	TDMA	CDMA	SDMA
General Features				
Basis	Division of system bandwidth B into N_F directly adjacent, disjoint subscriber frequency bands of width B_u ($B_u \ll B$)	Division of the transmission period into directly adjacent, disjoint TDMA frames of duration T_{fr} comprising of N_Z subscriber time slots of duration T_u ($T_u \ll T_{fr}$)	Spectrum spreading by using K_g subscriber specific CDMA codes	Division of the cell space into K_g sectors
Subscriber activity	N_F subscribers are simultaneously and continuously active; each subscriber uses a single subscriber frequency band	N_Z subscribers are consecutively active for a short period; each subscriber uses a single subscriber time slot per TDMA frame	K_g subscribers are simultaneously and continuously active; each subscriber uses a single subscriber specific CDMA code	K_g subscribers are simultaneously and continuously active; each subscriber has its own sector
Differentiation between subscriber signals	In the frequency domain	In the time domain	Based on CDMA codes	Based on direction of arrival at receiving antennas
Separating the subscriber signals ...	... by filtering	... by deploying synchronization; guard periods between consecutively transmitted subscriber signals are required	... by deploying synchronization, single-user detection (SUD) or multi-user detection (MUD)	... by using antenna arrays
Area of deployment	Analog and digital	Digital	Digital	Digital
Advantages	Simple; robust; supports network planning; simple equalization	Frequency diversity; receiver is insensitive to time variation of the mobile radio channel; time diversity; high spectral capacity owing to missing intra cell interference; reduced complexity in radio frequency design for cell phones and base stations possible; allows time domain duplexing (TDD)	Frequency diversity; receiver is insensitive to time variation of the mobile radio channel; simple equalizers; interference diversity; soft degradation; no network planning required; flexibility; reduced complexity in radio frequency design for base stations possible	Simple; reduction of multiple access interference; supports network planning; softer handover and space diversity; renunciation on equalizer possible
Disadvantages	Low flexibility; little frequency diversity; receiver sensitive to time variation of the mobile radio channel; little interference diversity; space diversity is necessary; considerable complexity in radio frequency design for cell phones and base stations	Low flexibility; latencies; equalizer is required due to intersymbol interference; little interference diversity; global synchronization of all subscribers, at least in a cell	Low spectral capacity without multi-user detection	Low flexibility; little frequency diversity; receiver sensitive to time variation of the mobile radio channel; reduces interference diversity; low spectral capacity; high implementation complexity of radio frequency design
Cellular Aspects				
Typical frequency re-use factor	$r > 1$ due to intercell interference	$r > 1$ due to intercell interference	$r \approx 1$	$r > 1$ due to intercell interference
Evaluation				
	Required for mobile radio; combination with TDMA and/or CDMA is favorable	Applicable in mobile radio; combination with FDMA is strongly suggested and with CDMA is favorable	Applicable in mobile radio; combination with FDMA is strongly suggested and with TDMA is favorable	Applicable in mobile radio; combination with FDMA is strongly suggested and with TDMA and CDMA, respectively, is favorable

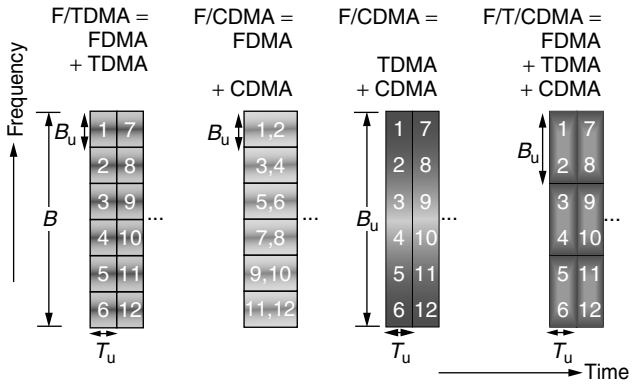


Figure 1. Hybrid multiple access schemes [3,4].

speech. The physical layer subscriber signals must contain signaling information that is required to set up, maintain, and release the connection between transmitter and receiver [3].

For mobile communication, a time-varying multipath channel with an unknown impulse response must be taken into account. To support coherent data detection, channel estimation must be carried out at least once per subscriber time slot. This channel estimation is based on training sequences, which are part of the aforementioned signaling information and which must therefore be embedded in the physical layer subscriber signals. Furthermore, the physical layer subscriber signals are concluded by guard periods of duration T_g in order to guarantee a reasonable separation between consecutive physical layer subscriber signals [3].

As illustrated in Section 2 and Table 1, TDMA allows a subscriber to be active only for a short time before the next period of activity occurs in the next TDMA frame. A typical duration of a subscriber time slot, T_u, is about 0.5 ms, whereas a TDMA frame consists of several subscriber time slots and has a typical duration, T_{fr}, on the order of 5 ms. Hence, the physical layer subscriber signals have a finite duration of T_u. Such signals are usually termed *bursts*.

Figure 2 shows two commonly used burst types [3]. The first burst type [Fig. 2(a)], uses a preamble, which contains the signaling information, including the aforementioned training sequence. When a preamble is used, the aforementioned channel estimation can take place at the beginning of the signal reception. The channel estimate, which is based on noisy samples, is affected by estimation errors due to noise in the received signal. Owing to these estimation errors, the data detection can be only quasi-coherent. The noisy channel estimates are fed into the quasi-coherent data detector, which carries out the data detection based on the sample values obtained after the reception of the preamble. Ideally, this quasi-coherent data detection can be carried out without having to store any sample values.

However, in the case of a low correlation time of the mobile radio channel (i.e., at high mobile velocities), the true channel impulse response varies over the duration T_u of the subscriber signals. The error between the noise channel estimate and the true channel impulse response increases nonlinearly with increasing distance from the preamble. In the case of long bursts, this effect leads to considerable systematic errors resulting in dramatic degradations of the quasi-coherent data detection (i.e., of the bit error ratio at a given E_b/N_0).

In order to alleviate this effect, midambles are used instead of preambles [Fig. 2(b)]. In this case, the data are divided in two parts, usually of equal size and half as long as the data carrying part shown in Fig. 2(a). The signaling information is located between these two parts. Then the effect of the above mentioned systematic errors on the bit error ratio is considerably smaller. However, in order to carry out a quasi-coherent data detection, at least those samples associated with the first part of encoded subscriber data must be stored before the channel estimation can be carried out. Nevertheless, thanks to high integration densities in CMOS technology, memory ICs or

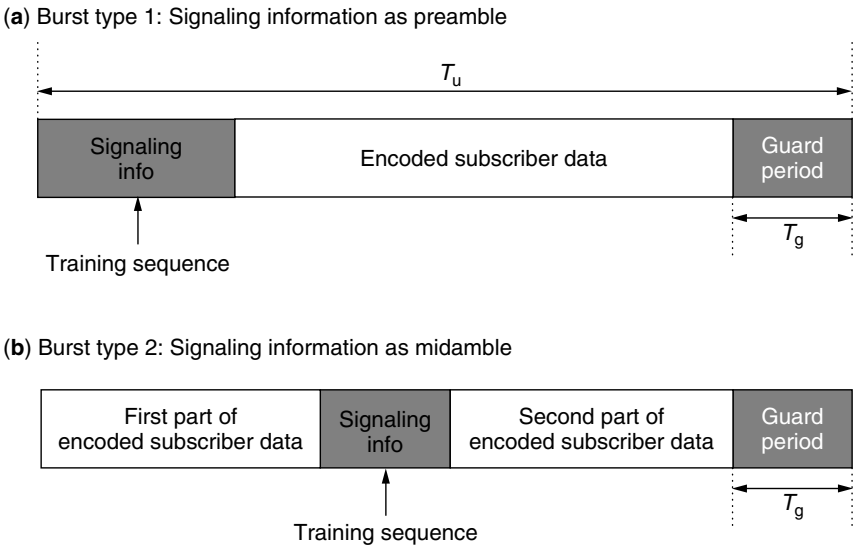


Figure 2. Burst types for TDMA [3]. (**a**) Burst type 1: Signaling information as preamble. (**b**) Burst type 2: Signaling information as midamble.

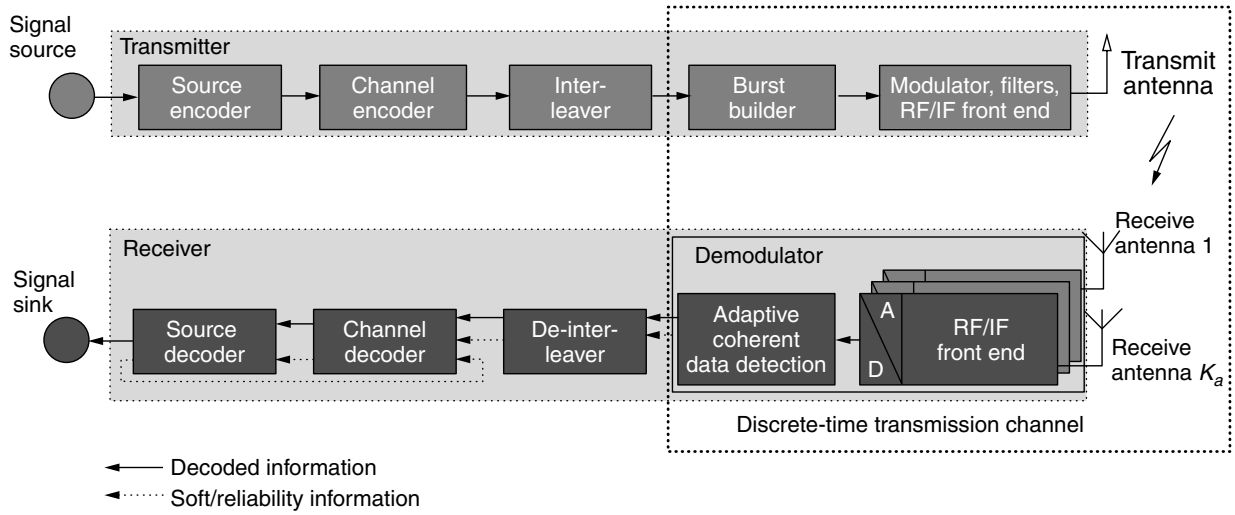


Figure 3. System structure for TDMA [3].

embedded on-chip memories are available at a reasonably low price, thus alleviating this drawback.

A third possibility, using a postamble, suffers from all the drawbacks of the aforementioned two burst types without having further advantages. To the knowledge of the author, this third possibility has not yet been implemented and is not further considered in this article.

3.2. System Structure

Figure 3 shows the corresponding system structure for the physical layer data path between the signal source and the signal sink (cf., e.g., Ref. 3). The system structure consists of a transmitter, a receiver, and the transmission channel.

The transmitter contains source and channel encoders, an interleaver, a burst builder, a modulator, digital and analog filters, the analog RF/IF transmit front end, and at least one transmit antenna. The receiver, particularly the base station receiver, consists of up to K_a receive antennas, K_a RF/IF receive front ends, K_a ADCs (analog-to-digital converters), an adaptive (quasi-) coherent data detector, a de-interleaver, and channel and source decoders.

Usually, hard decided, decoded information combined with the corresponding soft/reliability information are exchanged between the different receiver stages. In this way, a desirably good system performance can be guaranteed.

The system structure shown in Fig. 3 is the basis for the extension to TD/CDMA used in UMTS (see Section 5). There, the corresponding system structure is discussed.

4. TWO IMPORTANT TDMA SYSTEMS FOR MOBILE COMMUNICATION

4.1. Overview

In Fig. 4, the vision generated by the Wireless Strategic Initiative (www.ist-wsi.org) on the further evolution of mobile radio is summarized. The two most important and most successful TDMA systems are the European GSM (global system for mobile communication) [2] and the American UWC-136 (universal wireless communications) [5].

Both TDMA systems started with circuit switched data transmission. In its second phase, GSM was extended to high-speed circuit switched data (HSCSD) with minimal data rates of approximately 14.4 kbit/s and typical data rates between approximately 50 and 60 kbit/s. The corresponding first version of UWC 136 was termed D-AMPS (digital advanced mobile phone service) or IS-54. Both systems were further developed to incorporate packet switching based on GPRS (general packet radio service) and higher data rates using EDGE (enhanced data rates for GSM evolution). Typically, the different EDGE variants provide data rates of approximately 144 kbit/s, with a maximum about 400 kbit/s. GPRS and EDGE evolutions will be part of the family of third-generation mobile communication systems, briefly termed 3G mobile radio systems, with data rates above 400 kbit/s [5–9].

4.2. Global System for Mobile Communication (GSM)

Undoubtedly, the most successful mobile communication system is GSM. Today, GSM provides a multitude of both circuit and packet switched services and applications, including Internet access by using WAP (wireless application protocol) or i-mode. Maximal data rates are currently approximately 50 kbit/s. However, an increase up to approximately 400 kbit/s has already been introduced into the GSM standard.

Table 2 presents important system parameters of GSM, and Fig. 5 gives a comparison between the energy density spectra of well-known digital modulation schemes with those of GMSK (Gaussian minimum shift keying) and the GMSK main impulse [2,3,5,6].

5. TD/CDMA

As mentioned previously, F/TDMA lends itself to further extensions to 3G mobile radio systems. In the early 1990s,

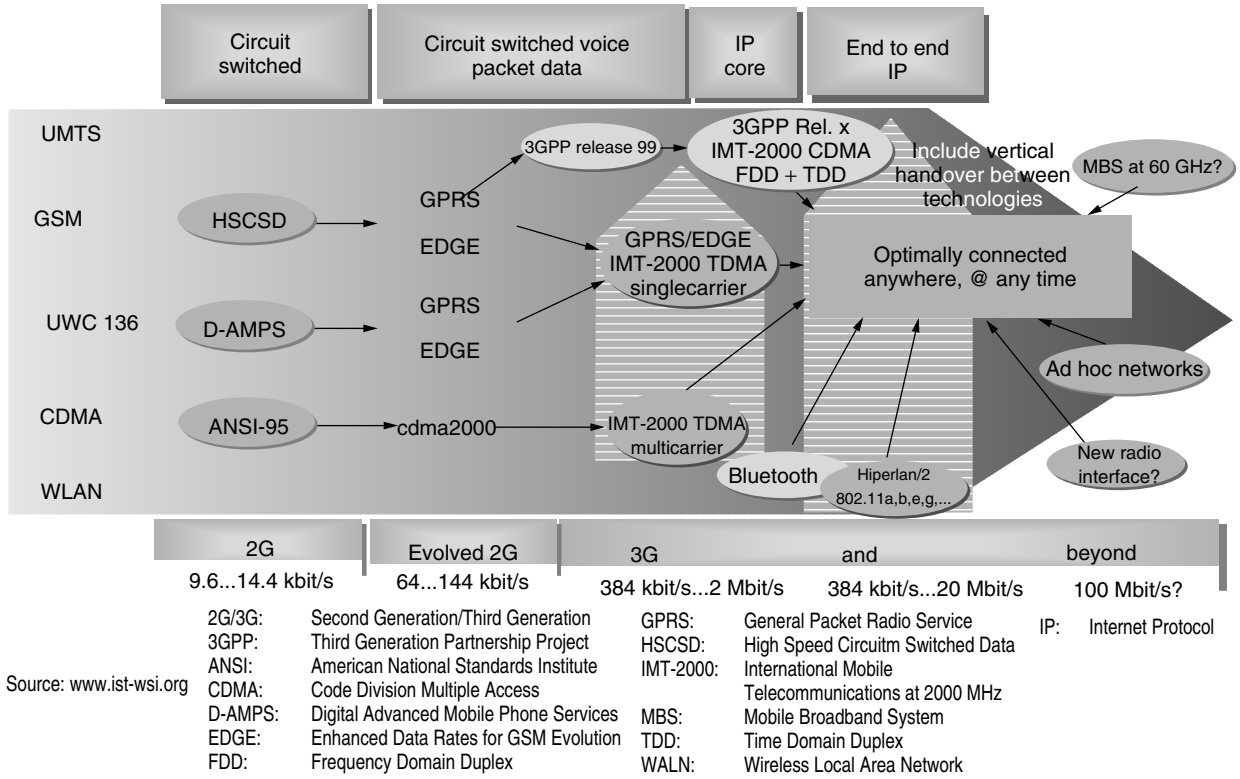


Figure 4. Evolution of wireless communications (source: *www.ist-wsi.org*).

the combination of F/TDMA with CDMA was presented, resulting in the hybrid multiple access scheme F/T/CDMA already considered in Section 2 and Fig. 1. Now, up to K subscribers could operate simultaneously within a TDMA time slot [3,5].

A major problem to be solved in a CDMA-based system is the near-far problem. In order to alleviate the necessity of fast power control, which cannot be provided at high velocities when a TDMA component is used, multiuser detection is a must in F/T/CDMA. It has been shown that suboptimal joint detection (JD) techniques based on block linear equalization and block decision feedback equalization lend themselves as viable means for such F/T/CDMA-based systems.

The channel estimation to be used must be capable of estimating a multitude of simultaneous transmission channels in the uplink. Steiner proposed a means of generating good training sequences for this purpose and proposed a novel channel estimator [10], sometimes termed the Steiner estimator. These milestone developments — the JD techniques and the Steiner estimator — paved the way toward what is known as TD/CDMA or UTRA TDD, today [3,5].

By moderately modifying the system structure shown in Fig. 3, the corresponding system structure for TD/CDMA can be found (Fig. 6). In Fig. 7, the physical layer subscriber signal is shown schematically. Both system and signal structures are consequent evolutions toward more multimedia in mobile communications, and it can be anticipated that TDMA components will remain important in the future.

6. CONCLUSIONS

TDMA as a viable means to solve the multiple access problem in mobile communication was discussed in this article. Besides giving an overview of the four classic multiple access principles, hybrid multiple access schemes were illustrated. Furthermore, we discussed both physical layer signal structures and the corresponding system structures for TDMA systems deployed in mobile communications. Finally, important TDMA systems and the evolution toward 3G mobile radio systems were briefly sketched.

To conclude, it should be mentioned that TDMA also provides excellent features in short-range communication systems. Therefore, wireless local area networks and some future wireless communication system concepts "beyond 3G" rely on TDMA components.

BIOGRAPHY

Peter Jung received the diploma (M.Sc. equiv.) in physics from the University of Kaiserslautern, Germany, in 1990, and the Dr.-Ing. (Ph.D.EE equiv.) and Dr.-Ing. habil. (D.Sc.EE equiv.), both in electrical engineering with a focus on microelectronics and communications technology, from the University of Kaiserslautern in 1993 and 1996, respectively. In 1996, he became private educator (equiv. to reader) at the University of Kaiserslautern and in 1998 also at Technical University of Dresden, Germany. From March 1998 to May 2000, he was with

Table 2. Important System Parameters of GSM [5,6]

Multiple Access Scheme		F/TDMA
Modulation scheme	Phases 1,2,2+:	GMSK (Gaussian minimum shift keying)
	EDGE:	GMSK, $(3/8)\pi$-Offset-8-PSK (8-ary Phase Shift Keying) with spectral forming by GMSK main impulse
Subscriber bandwidth	200 kHz	
Symbol rate	270.833 ksymbols/s	
Duration of a TDMA frame, T_{fr}	4.615 ms	
Number of subscriber time slots per TDMA frame	8	
Uplink (reverse link) frequency bands	880 . . . 915 1720 . . . 1785 1930 . . . 1990	(GSM 900, e.g., German D networks) (DCS 1800, e.g., German E networks) (American GSM 1900)
Downlink (forward link) frequency bands	935 . . . 960 1805 . . . 1880 1850 . . . 1910	(GSM 900, e.g., German D networks) (DCS 1800, e.g., German E networks) (American GSM 1900)
Maximal information rate per subscriber	Speech full rate: half rate: enhanced full rate: Data TCH/9.6 (phase 1): phase 2+ HSCSD: phase 2+ GPRS:	13 kbit/s 6.5 kbit/s 12.2 kbit/s 9.6 kbit/s 115.2 kbit/s 171.2 kbit/s (four coding schemes, today, only coding scheme CS-2 is used; three classes of mobile equipment; 18 multi-slot classes, today, classes 4 and 8 are usually implemented, cf. *www.csdmag.com*)
	EDGE:	packet switching with data rates of minimally 384 kbit/s for velocities below 100 km/h; packet switching with data rates of minimally 144.4 kbit/s for velocities between 100 km/h and 250 km/h
Frequency hopping (optional)	1 hop per TDMA frame, i.e., 217 hops/s	

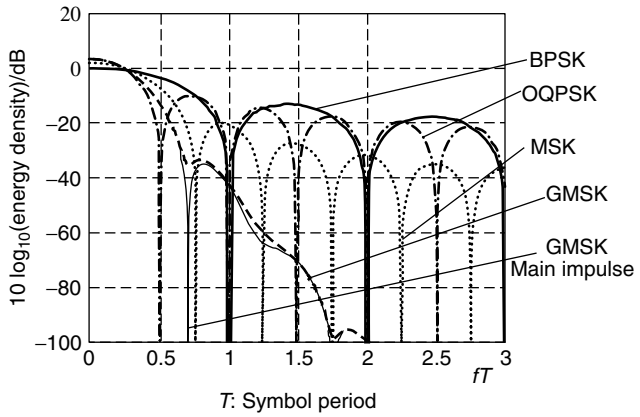


Figure 5. Energy density spectra.

Siemens AG, Bereich Halbleiter, now Infineon Technologies, as Director of Cellular Innovation and later Senior Director of Concept Engineering Wireless Baseband. In June 2000, he became full professor and chair for communication technology at Gerhard-Mercator-University Duisburg and director of the Fraunhofer-Institut für Mikroelekronische Schaltungen und Systeme (IMS), Duisburg. In 1995, he was co-recipient of the best paper award at the ITG-Fachtagung Mobile Kommunikation, Ulm, Germany, and in 1997, he was co-recipient of the Johann-Philipp-Reis-Award for his work on multicarrier CDMA mobile radio systems. His areas of interest include wireless communication technology, software-defined radio, and system-on-a-chip integration of communication systems.

BIBLIOGRAPHY

1. G. Calhoun, *Digital Cellular Radio*, Norwood, Mass., Artech House, 1988.
2. M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications*, 1992.
3. P. Jung, *Analyse und Entwurf digitaler Mobilfunksysteme*, Stuttgart, Teubner, 1997.
4. P. W. Baier, P. Jung, A. Klein, Taking the challenge of multiple access for third generation mobile radio systems — a European view, *IEEE Commun. Mag.* 82–89 (Feb. 1996).
5. T. Ojanperä and R. Prasad, eds., *Wideband CDMA for Third Generation Mobile Communications*, Boston, Artech House, 1998.

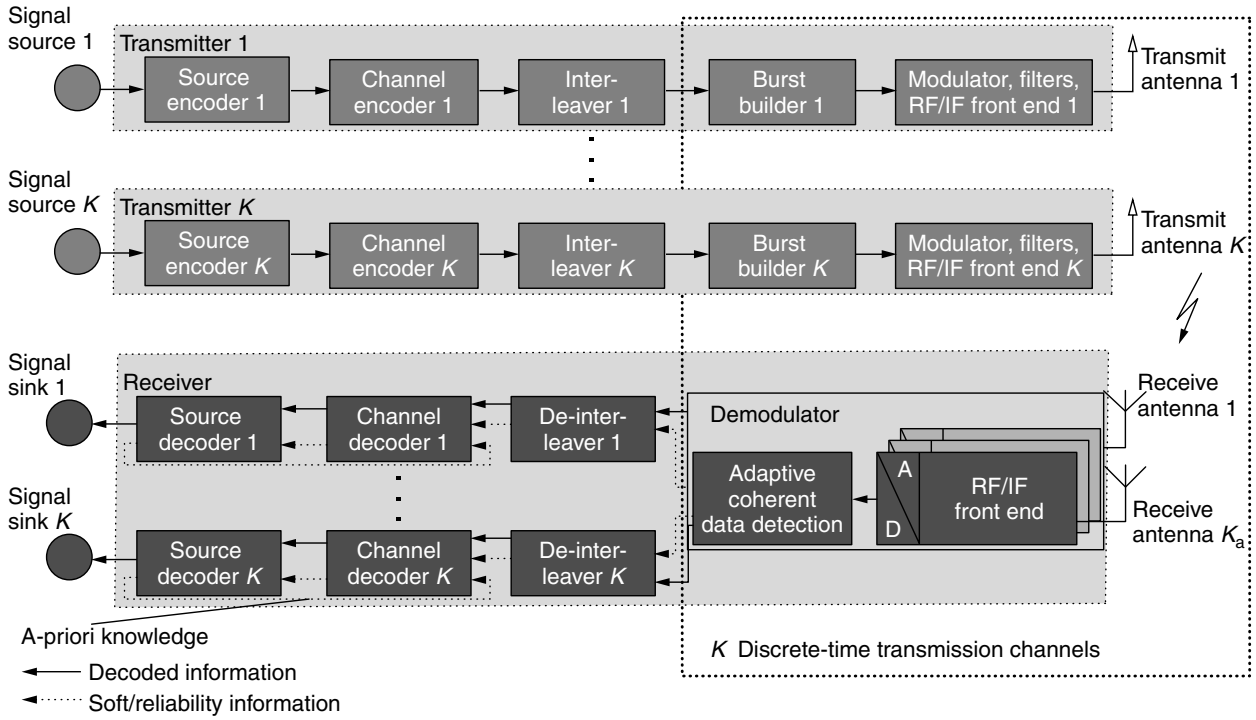


Figure 6. System structure for TD/CDMA [3].

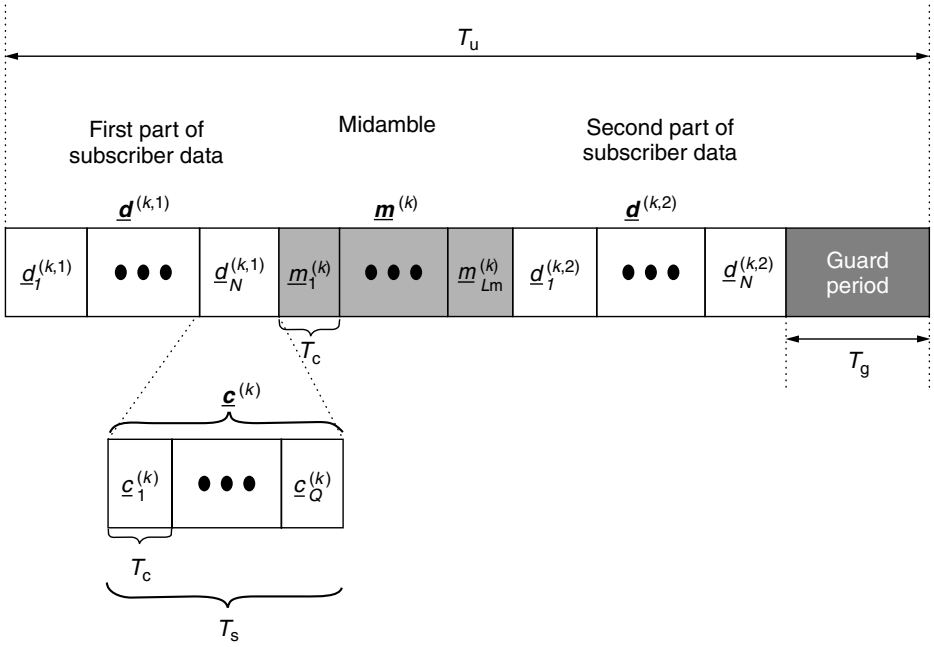


Figure 7. Data structure in a TD/CDMA burst [3].

6. A. Furuskär, S. Mazur, F. Müller, and H. Olofsson, EDGE: Enhanced Data Rates for GSM and TDMA/136 Evolution, *IEEE Personal Commun.* **6**: 56–66 (1999).
7. R. Prasad, W. Mohr, and W. Konhäuser, eds., *Third Generation Mobile Communication Systems*, Boston, Artech House, 2000.
8. B. Walke, M. P. Althoff, and P. Seidenberg, *UMTS—Ein Kurs*, Weil der Stadt, Schlembach, 2001.
9. H. Holma and A. Toskala, eds., *WCDMA for UMTS*, Chichester, Wiley, 2000.
10. B. Steiner and P. Jung, Optimum and suboptimum channel estimation for the uplink of CDMA mobile radio systems with joint detection, *European Transactions on Telecommunications and Related Technologies* (*ETT*) **5**: 39–50 (1994).

TRANSFORM CODING

VIVEK GOYAL
Digital Fountain Inc.
Fremont, California

1. INTRODUCTION

Transform coding is a type of source coding characterized by a modular design that includes a linear transformation of the original data and scalar quantization of the resulting coefficients. It arises from applying the "divide and conquer" principle to lossy source coding. This principle of breaking a major problem into smaller problems that can be more easily understood and solved is central in engineering and computational science. The resulting modular design is advantageous for implementation, testing, and component reuse.

Everyday compression problems are unmanageable without a divide-and-conquer approach. Effective compression of images, for example, depends on the tendencies of pixels to be similar to their neighbors, or to differ in partially predictable ways. These tendencies, arising from the continuity, texturing, and boundaries of objects, the similarity of objects in an image, gradual lighting changes, an artist's technique and color palette, or similar may extend over an entire image with a quarter-million pixels. Yet the most general way to utilize the probable relationships between pixels (later described as *unconstrained source coding*) is infeasible for this many pixels. In fact, 16 pixels is a lot for an unconstrained source code.

To conquer the compression problem — allowing, for example, more than 16 pixels to be encoded simultaneously — state-of-the-art lossy compressors divide the encoding operation into a sequence of three relatively simple steps: the computation of a linear transformation of the data designed primarily to produce uncorrelated coefficients, separate quantization of each scalar coefficient, and entropy coding. This process is called *transform coding*. In image compression, a square image with N pixels is typically processed with simple linear transforms (often discrete wavelet transforms) of size $N^{1/2}$.

This article explains the fundamental principles of transform coding with reference to abstract sources. These principles apply equally well to images, audio, video, and various other types of data.

1.1. Source Coding

Source coding is to represent information in bits, with the natural aim of using a small number of bits. When the information can be exactly recovered from the bits, the source coding or *compression* is called *lossless*; otherwise, it is called *lossy*. The transform codes in this article are lossy. However, lossless entropy codes appear as components of transform codes, so both lossless and lossy compression are of present interest.

In our discussion, the "information" is denoted by a real column vector $x \in \mathbb{R}^N$ or a sequence of such vectors. A vector might be formed from pixel values in an image or by sampling an audio signal; $K \cdot N$ pixels can be arranged as a sequence of K vectors of length N. The vector length N is defined such that each vector in a sequence is encoded independently. For the purpose of building a mathematical theory, the source vectors are assumed to be realizations of a random vector $\mathbf{x}$ with a known distribution. The distribution could be purely empirical.

A source code is composed of two mappings: an encoder and a decoder. The encoder maps any vector $x \in \mathbb{R}^N$ to a finite string of bits, and the decoder maps any of these strings of bits to an approximation $\hat{x} \in \mathbb{R}^N$. The encoder mapping can always be factored as $\gamma \circ \alpha$, where α is a mapping from $\mathbb{R}^N$ to some discrete set $\mathcal{I}$ and γ is an invertible mapping from $\mathcal{I}$ to strings of bits. The former is called a lossy encoder and the latter a lossless code or an entropy code. The decoder inverts γ and then approximates x from the index $\alpha(x) \in \mathcal{I}$. This is shown in the top half of Fig. 1. It is assumed that communication between the encoder and decoder is perfect.

To assess the quality of a lossy source code, we need numerical measures of approximation accuracy and description length. The measure for description length is simply the expected number of bits output by the encoder divided by N; this is called the *rate* in bits per scalar sample and denoted by R. Here we will measure approximation accuracy by squared Euclidean norm divided by the vector length:

$$d(x, \hat{x}) = \frac{1}{N}\|x - \hat{x}\|^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$$

This accuracy measure is conventional and usually leads to the easiest mathematical results, though the theory of source coding has been developed with quite general measures [1]. The expected value of $d(\mathbf{x}, \hat{\mathbf{x}})$ is called the mean-squared error (MSE) *distortion* and is denoted by $D = E[d(\mathbf{x}, \hat{\mathbf{x}})]$. The normalizations by N make it possible to fairly compare source codes with different lengths.

Fixing N, a theoretical concept of optimality is straightforward: A length-N source code is *optimal* if no other length-N source code with at most the same rate has lower distortion. This concept is of dubious value. First, it is very difficult to check the optimality of a source code. Local optimality — being assured that small perturbations of α and β will not improve performance — is often the best that can be attained [14]. Second, and of more practical consequence, a system designer gets to choose the value of N. It can be as large as the total size of the data set — like the number of pixels in an image — but can also be smaller, in which case the data set is interpreted as a sequence of vectors.

There are conflicting motives in choosing N. Compression performance is related to the predictability of one part of x from the rest. Since predictability can only increase from having more data, performance is usually improved by increasing N. (Even if the random variables producing each scalar sample are mutually independent, the optimal performance is improved by increasing N; however, this "packing gain" effect is relatively small [14].) The conflict comes from the fact that the computational complexity of encoding is also increased. This is particularly dramatic if one looks at complexities of optimal source codes. The

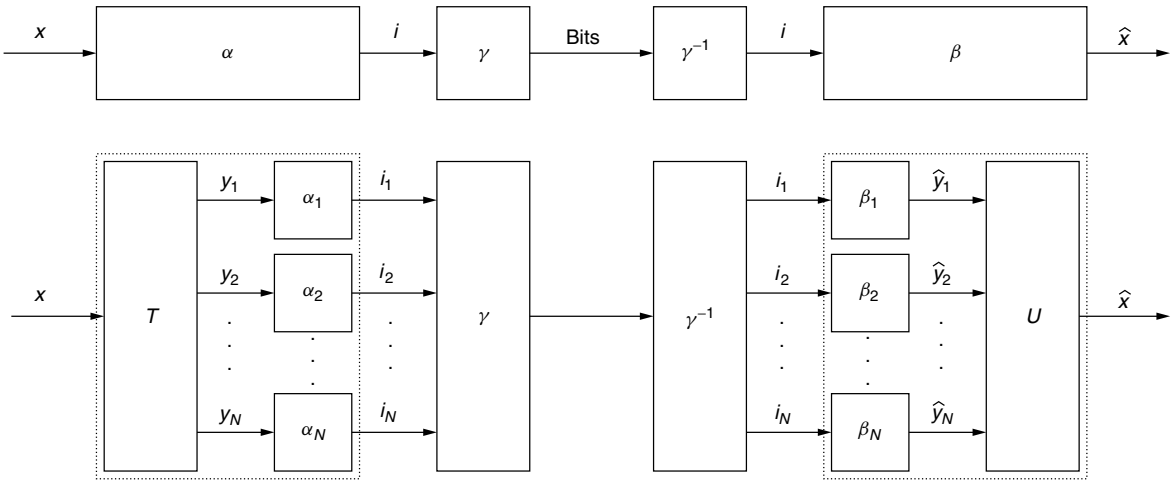


Figure 1. Any source code can be decomposed so that the encoder is $\gamma \circ \alpha$ and the decoder is $\beta \circ \gamma^{-1}$, as shown at top. γ is an entropy code and α and β are the encoder and decoder of an N-dimensional quantizer. In a transform code, α and β each have a particular constrained structure. In the encoder, α is replaced with a linear transform T and a set of N scalar quantizer encoders. The intermediate y_is are called *transform coefficients*. In the decoder, β is replaced with N scalar quantizer decoders and another linear transform U. Usually $U = T^{-1}$.

obvious way to implement an optimal encoder is to search through the entire codebook, giving running time exponential in N. Other implementations reduce running time while increasing memory usage [24].

State-of-the-art source codes result from an intelligent compromise. There is no attempt to realize an optimal code for a given value of N because encoding complexity would force a small value for N. Rather, source codes that are good, but plainly not optimal, are used. Their lower complexities make much larger N values feasible. This has eloquently been called "the power of imperfection" [5]. The paradoxical conclusion is that the best codes to use in practice are *suboptimal*.

1.2. Constrained Source Coding

Transform codes are the most often used source codes because they are easy to apply at any rate and even with very large values of N. The essence of transform coding is the modularization shown in the bottom half of Fig. 1. The mapping α is implemented in two steps. First, an invertible linear transform of the source vector x is computed, producing $y = Tx$. Each component of y is called a *transform coefficient*. The N transform coefficients are then quantized independently of each other by N scalar quantizers. This is called *scalar quantization* since each scalar component of y is treated separately. Finally, the quantizer indices that correspond to the transform coefficients are compressed with an entropy code to produce the sequence of bits that represent the data.

To reconstruct an approximation of x, the decoder essentially reverses the steps of the encoder. The action of the entropy coder can be inverted to recover the quantizer indices. Then the decoders of the scalar quantizers produce a vector $\hat{y}$ of estimates of the transform coefficients. To complete the reconstruction, a linear transform is applied to $\hat{y}$ to produce the approximation $\hat{x}$. This final step usually uses the transform T^{-1}, but for generality the transform is denoted U.

Most source codes cannot be implemented in the two stages of linear transform and scalar quantization. Thus, a transform code is an example of a *constrained source code*. Constrained source codes are, loosely speaking, source codes that are suboptimal but have low complexity. The simplicity of transform coding allows large values of N to be practical. Computing the transform T requires at most N^2 multiplications and $N(N-1)$ additions. Specially structured transforms — such as discrete Fourier, cosine, and wavelet transforms — are often used to reduce the complexity of this step, but this is merely icing on the cake. The great reduction from the exponential complexity of a general source code to the (at most) quadratic complexity of a transform code comes from using linear transforms and scalar quantization.

The difference between constrained and unconstrained source codes is demonstrated by the partition diagrams in Fig. 2. In these diagrams, the cells indicate which source vectors are encoded to the same index and the dots are the reconstructions computed by the decoder. Four locally optimal fixed-rate source codes with 12-element codebooks were constructed. The two-dimensional, jointly Gaussian source is the same as that used later in Fig. 3.

The partition for an unconstrained code shares the symmetries of the source density but is otherwise complicated because the cell shapes are arbitrary. Encoding is difficult because there is no simple way to get around using both components of the source vector simultaneously in computing the index.

Encoding with a transform code is easier because after the linear transform the coefficients are quantized separately. This gives the structured alignment of partition cells and of reconstruction points shown.

It is fair to ask why the transform is linear. In two dimensions, one might imagine quantizing in polar

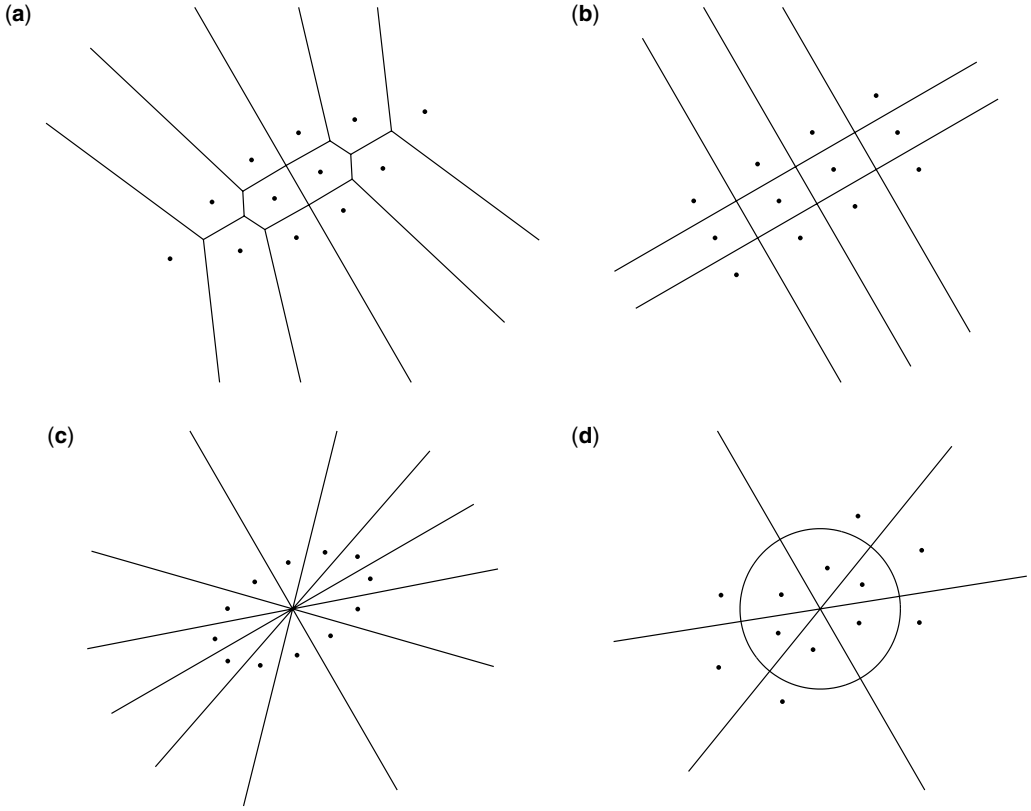


Figure 2. Partition diagrams for (**a**) unconstrained code ($D = 0.055$); (**b**) transform code ($D = 0.066$); (**c**) angular quantization ($D = 0.116$); (**d**) polar-coordinate quantization ($D = 0.069$).

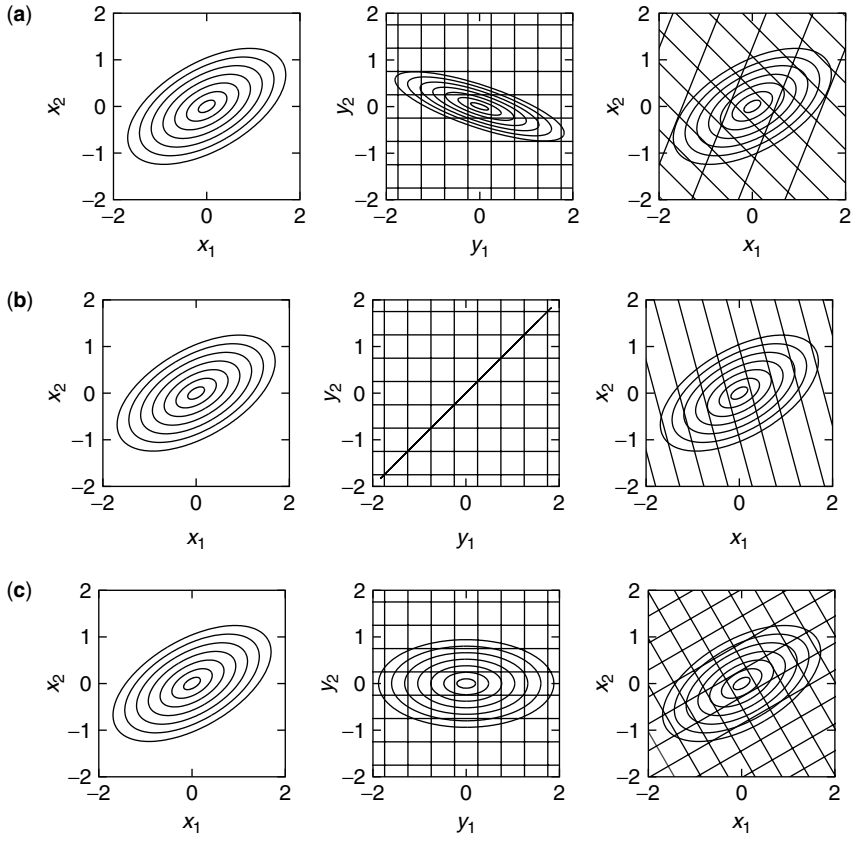


Figure 3. Illustration of various basis changes: (**a**) a basis change generally includes a nonhypercubic partition; (**b**) a singular transformation gives a partition with unbounded cells; (**c**) a Karhunen–Loève transform is an orthogonal transform that aligns the partitioning with the axes of the source PDF. The source is depicted by level curves of the PDF (left). The transform coefficients are separately quantized with uniform quantizers (center). The induced partitioning is then shown in the original coordinates (right).

coordinates. Two examples of partitions obtained with separate quantization of radial and angular components are shown in Figs. 2c and 2d, and these are as elegant as the partition obtained with a linear transform. Yet nonlinear transformations — even transformations to polar coordinates — are rarely used in source coding. With arbitrary transformations, the approximation accuracy of the transform coefficients does not easily relate to the accuracy of the reconstructed vectors. This makes designing quantizers for the transform coefficients more difficult. Also, allowing nonlinear transformations reintroduces the design and encoding complexities of unconstrained source codes.

Constrained source codes need not use transforms to have low complexity. Techniques described in the literature [8,14] include those based on lattices, sorting, and tree-structured searching; none of these techniques is as popular as transform coding.

2. THE STANDARD MODEL AND ITS COMPONENTS

The standard theoretical model for transform coding has the strict modularity shown in the bottom half of Fig. 1, meaning that the transform, quantization, and entropy coding blocks operate independently. In addition, the entropy coder can be decomposed into N parallel entropy coders so that the quantization and entropy coding operate independently on each scalar transform coefficient.

This section briefly describes the fundamentals of entropy coding and quantization to provide background for our later focus on the optimization of the transform. The final part of this section addresses the allocation of bits among the N scalar quantizers. Additional information can be found in the literature [3,8,12,14].

2.1. Entropy Codes

Entropy codes are used for lossless coding of discrete random variables. Consider the discrete random variable $\mathbf{z}$ with alphabet $\mathcal{I}$. An entropy code γ assigns a unique binary string, called a *codeword*, to each $i \in \mathcal{I}$ (see Fig. 1).

Since the codewords are unique, an entropy code is always invertible. However, we will place more restrictive conditions on entropy codes so they can be used on sequences of realizations of $\mathbf{z}$. The *extension* of γ maps the finite sequence $(z_1, z_2, \ldots, z_k)$ to the concatenation of the outputs of γ with each input, $\gamma(z_1)\gamma(z_2)\cdots\gamma(z_k)$. A code is called *uniquely decodable* if its extension is one to one. A uniquely decodable code can be applied to message sequences without adding any "punctuation" to show where one codeword ends and the next begins. In a *prefix code*, no codeword is the prefix of any other codeword. Prefix codes are guaranteed to be uniquely decodable.

A trivial code numbers each element of $\mathcal{I}$ with a distinct index in $\{0, 1, \ldots, |\mathcal{I}| - 1\}$ and maps each element to the binary expansion of its index. Such a code requires $\lceil \log_2 |\mathcal{I}| \rceil$ bits per symbol. This is considered the *lack* of an entropy code. The idea in entropy code design is to minimize the mean number of bits used to represent $\mathbf{z}$ at the expense of making the worst-case performance worse. The expected code length is given by

$$L(\gamma) = E[\ell(\gamma(\mathbf{z}))] = \sum_{i \in \mathcal{I}} p_{\mathbf{z}}(i)\ell(\gamma(i))$$

where $p_{\mathbf{z}}(i)$ is the probability of symbol i and $\ell(\gamma(i))$ is the length of $\gamma(i)$. The expected length can be reduced if short codewords are used for the most probable symbols — even if this means that some symbols will have codewords with more than $\lceil \log_2 |\mathcal{I}| \rceil$ bits.

The entropy code γ is called *optimal* if it is a prefix code that minimizes $L(\gamma)$. Huffman codes are examples of optimal codes. The performance of an optimal code is bounded by

$$H(\mathbf{z}) \le L(\gamma) < H(\mathbf{z}) + 1 \tag{1}$$

where

$$H(\mathbf{z}) = -\sum_{i \in \mathcal{I}} p_{\mathbf{z}}(i) \log_2 p_{\mathbf{z}}(i)$$

is the *entropy* of $\mathbf{z}$.

The up to one bit gap in Eq. (1) is ignored in the remainder of the article. If $H(\mathbf{z})$ is large, this is justified simply because one bit is small compared to the code length. Otherwise note that $L(\gamma) \approx H(\mathbf{z})$ can attained by coding blocks of symbols together; this is detailed in any information theory or data compression textbook.

2.2. Quantizers

A quantizer q is a mapping from a source alphabet $\mathbb{R}^N$ to a *reproduction codebook* $\mathcal{C} = \{\hat{x}_i\}_{i \in \mathcal{I}} \subset \mathbb{R}^N$, where $\mathcal{I}$ is an arbitrary countable index set. Quantization can be decomposed into two operations $q = \beta \circ \alpha$, as shown in Fig. 1. The *lossy encoder* $\alpha : \mathbb{R}^N \to \mathcal{I}$ is specified by a partition of $\mathbb{R}^N$ into *partition cells* $S_i = \{x \in \mathbb{R}^N \mid \alpha(x) = i\}$, $i \in \mathcal{I}$. The *reproduction decoder* $\beta : \mathcal{I} \to \mathbb{R}^N$ is specified by the codebook $\mathcal{C}$. If $N = 1$, the quantizer is called a *scalar quantizer*; for $N > 1$, it is a *vector quantizer*.

The quality of a quantizer is determined by its distortion and rate. The MSE distortion for quantizing random vector $\mathbf{x} \in \mathbb{R}^N$ is

$$D = E[d(\mathbf{x}, q(\mathbf{x}))] = N^{-1}E[\|\mathbf{x} - q(\mathbf{x})\|^2]$$

The rate can be measured in several ways. The lossy encoder output $\alpha(\mathbf{x})$ is a discrete random variable that usually should be entropy coded because the output symbols will have unequal probabilities. Associating an entropy code γ to the quantizer gives a *variable-rate quantizer* specified by (α, β, γ). The rate of the quantizer is the expected code length of γ divided by N. Not specifying an entropy code (or specifying the use of fixed-rate binary expansion) gives a *fixed-rate quantizer* with rate $R = N^{-1} \log_2 |\mathcal{I}|$. Measuring the rate by the idealized performance of an entropy code gives $R = N^{-1}H(\alpha(\mathbf{x}))$; the quantizer in this case is called *entropy-constrained*.

The optimal performance of variable-rate quantization is at least as good as that of fixed-rate quantization, and entropy-constrained quantization is better yet. However, entropy coding adds complexity, and variable-length

output can create difficulties such as buffer overflows. Furthermore, entropy-constrained quantization is only an idealization since an entropy code will generally not meet the lower bound in Eq. (1).

2.2.1. Optimal Quantization. An *optimal quantizer* is one that minimizes the distortion subject to an upper bound on the rate or minimizes the rate subject to an upper bound on the distortion. Because of simple shifting and scaling properties, an optimal quantizer for a scalar $\mathbf{x}$ can be easily deduced from an optimal quantizer for the normalized random variable $(\mathbf{x} - \mu_{\mathbf{x}})/\sigma_{\mathbf{x}}$, where $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ are the mean and standard deviation of $\mathbf{x}$, respectively. One consequence of this is that optimal quantizers have performance

$$D = \sigma^2 g(R) \tag{2}$$

where σ^2 is the variance of the source and $g(R)$ is the performance of optimal quantizers for the normalized source. Equation (2) holds, with a different function g, for any family of quantizers that can be described by its operation on a normalized variable, not just optimal quantizers.

Optimal quantizers are difficult to design, but locally optimal quantizers can be numerically approximated by an iteration in which α, β, and γ are separately optimized, in turn, while keeping the other two fixed. For details on each of these optimizations and the difficulties and properties that arise, see Refs. 5 and 14.

Note that the rate measure affects the optimal encoding rule because $\alpha(x)$ should be the index that minimizes a Lagrangian cost function including both rate and distortion; for example

$$\alpha(x) = \operatorname*{argmin}_{i \in \mathcal{I}} \left[\frac{1}{N} \ell(\gamma(i)) + \lambda \frac{1}{N} \|x - \beta(i)\|^2 \right]$$

is an optimal lossy encoder for variable-rate quantization. (By fixing the relative importance of rate and distortion, the Lagrange multiplier λ determines a rate-distortion operating point among those possible with the given β and γ.) Only for fixed-rate quantization does the optimal encoding rule simplify to finding the index corresponding to the nearest codeword.

In some of the more technical discussions that follow, one property of optimal decoding is relevant: The optimal decoder β computes

$$\beta(i) = E[\mathbf{x} \mid \mathbf{x} \in S_i]$$

which is called *centroid reconstruction*. The conditional mean of the cell, or centroid, is the minimum MSE estimate [26].

2.2.2. High-Resolution Quantization. For most sources, it is impossible to analytically express the performance of optimal quantizers. Thus, aside from using Eq. (2), approximations must suffice. Fortunately, approximations obtained when it is assumed that the quantization is very fine are reasonably accurate even at low to moderate rates. Details on this "high resolution" theory for both scalars and vectors can be found in Refs. 7 and 14 and other sources cited therein.

Let $f_{\mathbf{x}}(x)$ denote the probability density function (PDF) of the scalar random variable $\mathbf{x}$. High-resolution analysis is based on approximating $f_{\mathbf{x}}(x)$ on the interval S_i by its value at the midpoint. Assuming that $f_{\mathbf{x}}(x)$ is smooth, this approximation is accurate when each S_i is short.

Optimization of scalar quantizers turns into finding the optimal lengths for the S_is, depending on the PDF $f_{\mathbf{x}}(x)$. One can show that the performance of optimal fixed-rate quantization is approximately

$$D \approx \frac{1}{12} \left(\int_{\mathbb{R}} f_{\mathbf{x}}^{1/3}(x)\, dx \right)^3 2^{-2R} \tag{3}$$

Evaluating this for a Gaussian source with variance σ^2 gives

$$D \approx \tfrac{1}{2} 3^{1/2} \pi \sigma^2 2^{-2R} \tag{4}$$

For entropy-constrained quantization, high-resolution analysis shows that it is optimal for each S_i to have equal length [9]. A quantizer that partitions with equal-length intervals is called *uniform*. The resulting performance is

$$D \approx \frac{1}{12} 2^{2h(\mathbf{x})} 2^{-2R} \tag{5}$$

where

$$h(\mathbf{x}) = -\int_{\mathbb{R}} f_{\mathbf{x}}(x) \log_2 f_{\mathbf{x}}(x)\, dx$$

is the *differential entropy* of $\mathbf{x}$. For Gaussian random variables, Eq. (5) simplifies to

$$D \approx \frac{\pi e}{6} \sigma^2 2^{-2R} \tag{6}$$

Summarizing Eqs. (3)–(6), the lesson from high resolution quantization theory is that quantizer performance is described by

$$D \approx c \sigma^2 2^{-2R} \tag{6}$$

where σ^2 is the variance of the source and c is a constant that depends on the normalized density of the source and the type of quantization (fixed-rate, variable-rate or entropy-constrained). This is consistent with Eq. (2).

The computations we have made are for scalar quantization. For vector quantization, the best performance in the limit as the dimension N grows is given by the distortion rate function [1]. For a Gaussian source this bound is $D = \sigma^2 2^{-2R}$. The approximate performance given by Eq. (6) is worse by a factor of only $\pi e/6$ ($\approx$1.53 dB). This can be expressed as a redundancy $\frac{1}{2} \log_2(\frac{\pi e}{6}) \approx 0.255$ bits. Furthermore, a numerical study has shown that for a wide range of memoryless sources, the redundancy of entropy-constrained uniform quantization is at most 0.3 bits per sample at all rates [6].

2.3. Bit Allocation

Coding (quantizing and entropy coding) each transform coefficient separately splits the total number of bits among the transform coefficients in some manner. Whether done with conscious effort or implicitly, this is a *bit allocation* among the components.

Bit allocation problems can be stated in a single common form: One is given a set of quantizers described by their distortion–rate performances as

$$D_i = g_i(R_i), \quad R_i \in \mathcal{R}_i, \quad i = 1, 2, \ldots, N$$

Each set of available rates $\mathcal{R}_i$ is a subset of the nonnegative real numbers and may be either discrete or continuous. The problem is to minimize the average distortion $D = N^{-1} \sum_{i=1}^{N} D_i$ given a maximum average rate $R = N^{-1} \sum_{i=1}^{N} R_i$.

As is often the case with optimization problems, bit allocation is easy when the parameters are continuous and the objective functions are smooth. Subject to a few other technical requirements, parametric expressions for the optimal bit allocation can be found elsewhere [22,29]. The techniques used when the $\mathcal{R}_i$s are discrete are quite different and play no role in forthcoming results [30].

If the average distortion can be reduced by taking bits away from one component and giving them to another, the initial bit allocation is not optimal. Applying this reasoning with infinitesimal changes in the component rates, a necessary condition for an optimal allocation is that the slope of each g_i at R_i is equal to a common constant value. A tutorial treatment of this type of optimization has been published [25].

The approximate performance given by Eq. (7) leads to a particularly easy bit allocation problem with

$$g_i = c_i \sigma_i^2 2^{-2R_i}, \quad \mathcal{R}_i = [0, \infty), \quad i = 1, 2, \ldots, N \tag{8}$$

Ignoring the fact that each component rate must be nonnegative, an equal-slope argument shows that the optimal bit allocation is

$$R_i = R + \frac{1}{2} \log_2 \frac{c_i}{\left(\prod_{i=1}^{N} c_i \right)^{1/N}} + \frac{1}{2} \log_2 \frac{\sigma_i^2}{\left(\prod_{i=1}^{N} \sigma_i^2 \right)^{1/N}}$$

With these rates, all the D_is are equal and the average distortion is

$$D = \left(\prod_{i=1}^{N} c_i \right)^{1/N} \left(\prod_{i=1}^{N} \sigma_i^2 \right)^{1/N} 2^{-2R}. \tag{9}$$

This solution is valid when each R_i given above is nonnegative. For lower rates, the components with smallest $c_i \cdot \sigma_i^2$ are allocated no bits and the remaining components have correspondingly higher allocations.

2.3.1. Bit Allocation With Uniform Quantizers. With uniform quantizers, bit allocation is nothing more than choosing a step size for each of the N components. The equal-distortion property of the analytical bit allocation solution gives a simple rule: Make all of the step sizes equal. This will be referred to as "lazy" bit allocation.

Our development indicates that lazy allocation is optimal when the rate is high. In addition, numerical studies have shown that lazy allocation is nearly optimal as long as the minimal allocated rate is at least 1 bit [12,13].

3. OPTIMAL TRANSFORMS

It has taken some time to set the stage, but we are now ready for the main event of designing the analysis transform T and the synthesis transform U. Throughout this section the source $\mathbf{x}$ is assumed to have mean zero, and $R_{\mathbf{x}}$ denotes the covariance matrix $E[\mathbf{x}\mathbf{x}^T]$, where ^{T} denotes the transpose. The source is often — but not always — jointly Gaussian.

A signal given as a vector in $\mathbb{R}^N$ is implicitly represented as a series with respect to the standard basis. An invertible analysis transform T changes the basis. A change of basis does not alter the information in a signal, so how can it affect coding efficiency? Indeed, if arbitrary source coding is allowed after the transform, it does not. The motivating principle of transform coding is that *simple* coding may be more effective in the transform domain than in the original signal space. In the standard model, "simple coding" corresponds to the use of scalar quantization and scalar entropy coding.

3.1. Visualizing Transforms

Beyond two or three dimensions, it is difficult to visualize vectors — let alone the action of a transform on vectors. Fortunately, most people already have an idea of what a linear transform does: it combines rotating, scaling, and shearing such that a hypercube is always mapped to a parallelepiped.

In two dimensions, the level curves of a zero-mean Gaussian density are ellipses centered at the origin with collinear major axes, as shown in the left panels of Fig. 3. The middle panel of Fig. 3a shows the level curves of the joint density of the transform coefficients after a more or less arbitrary invertible linear transformation. A linear transformation of an ellipse is still an ellipse, although its eccentricity and orientation (direction of major axis) may have changed.

The grid in the middle panel indicates the cell boundaries in uniform scalar quantization, with equal step sizes, of the transform coefficients. The effect of inverting the transform is shown in the right panel; the source density is returned to its original form and the quantization partition is linearly deformed. The partition in the original coordinates, as shown in the right panel, is what is truly relevant. It shows which source vectors are mapped to the same symbol, thus giving some indication of the average distortion. Looking at the number of cells with appreciable probability gives some indication of the rate.

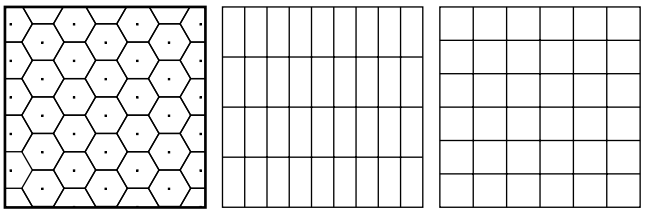
A singular transform is a degenerate case. As shown in the middle panel of Fig. 3b, the transform coefficients have probability mass only along a line. (A *line segment* is an ellipse with unit eccentricity.) Inverting the transform is not possible, but we may still return to the original coordinates to view the partition induced by quantizing the transform coefficients. The cells are unbounded in one

direction, as shown in the right panel. This is undesirable unless variation of the source in the direction in which the cells are unbounded is very small.

3.1.1. Shapes of Partition Cells. Although better than unbounded cells, the parallelogram-shaped partition cells that arise from arbitrary invertible transforms are inherently suboptimal. To understand this better, note that the quality of a source code depends on the shapes of the partition cells $\{\alpha^{-1}(i), i \in \mathcal{I}\}$ and on varying the sizes of the cells according to the source density. When the rate is high, and either the source is uniformly distributed or the rate is measured by entropy $[H(\alpha(\mathbf{x}))]$, the sizes of the cells should essentially not vary. Then, the quality depends on having cell shapes that minimize the average distance to the center of the cell.

For a given volume, a body in Euclidean space that minimizes the average distance to the center is a sphere. But spheres do not work as partition cell shapes because they do not pack together without leaving interstices. Only for a few dimensions N is the best cell shape known [2]. One such dimension is $N = 2$, where the hexagonal packing shown below (left) is best.

The best packings (including the hexagonal case) cannot be achieved with transform codes. Transform codes can only produce partitions into parallelepipeds, as shown for $N = 2$ in Fig. 3. The best parallelepipeds are cubes. We get a hint of this by comparing the two rectangular partitions of a unit-area square shown below. Both partitions have 36 cells, so every cell has the same area. The partition with square cells gives distortion $1/432 \approx 2.31 \times 10^{-3}$, while the other gives $97/31,104 \approx 3.12 \times 10^{-3}$.



This simple example can also be interpreted as a problem of allocating bits between the horizontal and vertical components. The "lazy" bit allocation arising from equal quantization step sizes for each component is optimal. This holds generally for high-rate entropy-constrained quantization of components with the same normalized density.

Returning now to transform choice, to get rectangular partition cells the basis vectors must be orthogonal. For square cells, when quantization step sizes are equal for each transform coefficient, the basis vectors should in addition to being orthogonal have equal lengths. When orthogonal basis vectors have unit length, the resulting transform is called an orthogonal transform. (It is regrettable that of a matrix or transform, "orthogonal" means orthonormal.)

3.1.2. Karhunen–Loève Transforms. A *Karhunen–Loève transform* (KLT) is a particular type of orthogonal transform that depends on the covariance of the source. An orthogonal matrix T represents a KLT of $\mathbf{x}$ if $TR_{\mathbf{x}}T^T$ is a diagonal matrix. The diagonal matrix $TR_{\mathbf{x}}T^T$ is the covariance of $\mathbf{y} = T\mathbf{x}$; thus, a KLT gives uncorrelated transform coefficients. KLT is the most commonly used name for these transforms in signal processing, communication, and information theory, recognizing the works by Karhunen and Loève [19,21]; among the other names are Hotelling transforms [16] and principal component transforms.

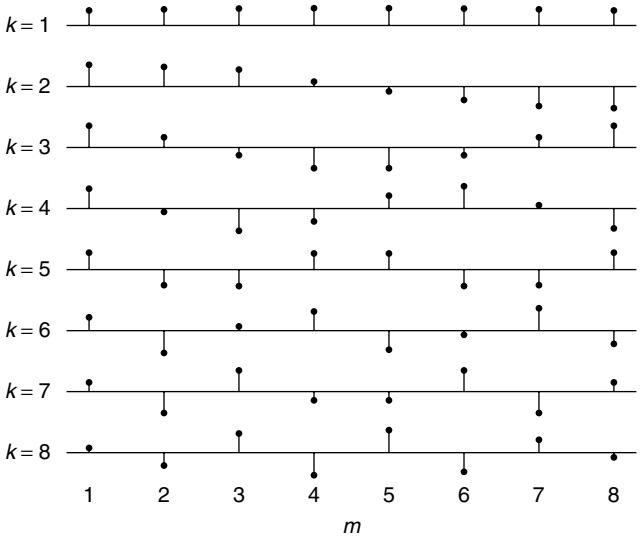
A KLT exists for any source because covariance matrices are symmetric, and symmetric matrices are orthogonally diagonalizable; the diagonal elements of $TR_{\mathbf{x}}T^T$ are the eigenvalues of $R_{\mathbf{x}}$. KLTs are not unique; any row of T can be multiplied by ± 1 without changing $TR_{\mathbf{x}}T^T$, and permuting the rows leaves $TR_{\mathbf{x}}T^T$ diagonal. If the eigenvalues of $R_{\mathbf{x}}$ are not distinct, there is additional freedom in choosing a KLT.

For an example of a KLT, consider the first-order autoregressive signal model that is popular in many branches of signal processing and communication. Under such a model, a sequence is generated as

$$\mathbf{x}[k] = \rho\mathbf{x}[k-1] + \mathbf{z}[k]$$

where k is a time index, $\mathbf{z}[k]$ is a white sequence, and $\rho \in [0, 1)$ is called the *correlation coefficient*. It is a crude but useful model for the samples along any line of a grayscale image, with $\rho \approx 0.9$.

An $\mathbb{R}^N$-valued source $\mathbf{x}$ can be derived from a scalar autoregressive source by forming blocks of N consecutive samples. With normalized power, the covariance matrix is given elementwise by $(R_{\mathbf{x}})_{ij} = \rho^{|i-j|}$. For any particular N and ρ, a numerical eigendecomposition method can be applied to $R_{\mathbf{x}}$ to obtain a KLT. (An analytic solution also happens to be possible [15].) For $N = 8$ and $\rho = 0.9$, the KLT can be depicted as follows:



Each subplot (value of k) gives a row of the transform or, equivalently, a vector in the analysis basis. This basis is superficially sinusoidal and is approximated by

a discrete cosine transform (DCT) basis. There are various asymptotic equivalences between KLTs and DCTs for large N and $\rho \to 1$ [18,28]. These results are often cited in justifying the use of DCTs.

For Gaussian sources, KLTs align the partitioning with the axes of the source PDF, as shown in Fig. 3c. It appears that the forward and inverse transforms are rotations, although actually the symmetry of the source density obscures possible reflections.

3.2. The Easiest Transform Optimization

Consider a jointly Gaussian source, and assume U and T are orthogonal and $U = T^{-1}$. The Gaussian assumption is important because any linear combination of jointly Gaussian random variables is Gaussian. Thus, any analysis transform gives Gaussian transform coefficients. Then, since the transform coefficients have the same normalized density, for any reasonable set of quantizers, Eq. (2) holds with a single function $g(R)$ describing all the transform coefficients. Orthogonality is important because orthogonal transforms preserve Euclidean lengths, which gives $d(x, \hat{x}) = d(y, \hat{y})$.

With these assumptions, for any rate and bit allocation a KLT is an optimal transform:

Theorem 1 [13]**.** Consider a transform coder with orthogonal analysis transform T and synthesis transform $U = T^{-1} = T^T$. Suppose that there is a single function g to describe the quantization of each transform coefficient through

$$E[(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2] = \sigma_i^2 g(R_i), \quad i = 1, 2, \ldots, N$$

where σ_i^2 is the variance of $\mathbf{y}_i$ and R_i is the rate allocated to y_i. Then for any bit allocation $(R_1, R_2, \ldots, R_N)$ there is a KLT that minimizes the distortion. In the typical case where g is nonincreasing, a KLT that gives $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$ sorted in the same order as the bit allocation minimizes the distortion.

Since it holds for any bit allocation and many families of quantizers, Theorem 1 is stronger than several earlier transform optimization results. In particular, it subsumes the low-rate results of Lyons [23] and the high-rate results that are reviewed presently.

Recall that with a high average rate of R bits per component and quantizer performance described by Eq. (8), the average distortion with optimal bit allocation is given by Eq. (9). With Gaussian transform coefficients that are optimally quantized, the distortion simplifies to

$$D = c \left(\prod_{i=1}^{N} \sigma_i^2 \right)^{1/N} 2^{-2R} \tag{10}$$

where $c = \pi e/6$ for entropy-constrained quantization or $c = 3^{1/2}\pi/2$ for fixed-rate quantization. The choice of an orthogonal transform is thus guided by minimizing the geometric mean of the transform coefficient variances.

Theorem 2. The distortion given by Eq. (10) is minimized over all orthogonal transforms by any KLT.

Proof: Applying Hadamard's Inequality to $R_{\mathbf{y}}$ gives

$$(\det T)(\det R_{\mathbf{x}})(\det T^T) = \det R_{\mathbf{y}} \le \prod_{i=1}^{N} \sigma_i^2$$

Since $\det T = 1$, the left-hand side of this inequality is invariant to the choice of T. Equality is achieved when a KLT is used. Thus KLTs minimize the distortion.

Equation (10) can be used to define a figure of merit called the *coding gain*. The *coding gain* of a transform is a function of its variance vector, $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$, and the variance vector without a transform, $\mathrm{diag}(R_x)$:

$$\text{Coding gain} = \frac{\left(\prod_{i=1}^{N} (R_x)_{ii} \right)^{1/N}}{\left(\prod_{i=1}^{N} \sigma_i^2 \right)^{1/N}}.$$

The coding gain is the factor by which the distortion is reduced because of the transform, assuming high rate and optimal bit allocation. The foregoing discussion shows that KLTs maximize coding gain. Related measures are the *variance distribution*, *maximum reducible bits*, and *energy packing efficiency* or *energy compaction*. All of these are optimized by KLTs [28].

3.3. More General Results

The results of the previous section are straightforward and Theorem 2 is well known. However, KLTs are not always optimal. With some sources, there are nonorthogonal transforms that perform better than any orthogonal transform. And, depending on the quantization, $U = T^{-1}$ is not always optimal — even for Gaussian sources. This section provides results that apply without the presumption of Gaussianity or orthogonality.

3.3.1. The Synthesis Transform *U*. Instead of assuming the decoder structure shown in the bottom of Fig. 1, let us consider for a moment the best way to decode given only the encoding structure of a transform code. The analysis transform followed by quantization induces some partition of $\mathbb{R}^N$, and the best decoding is to associate with each partition cell its centroid. Generally, this decoder cannot be realized with a linear transform applied to $\hat{y}$. For one thing, some scalar quantizer decoder β_i could be designed in a plainly wrong way; then it would take an extraordinary (nonlinear) effort to fix the estimates.

The difficulty is actually more dramatic because even if the β_i mappings are optimal, the synthesis transform T^{-1} applied to $\hat{y}$ will seldom give optimal estimates. In fact, unless the transform coefficients are independent, there may be a *linear transform* better suited to the reconstruction than T^{-1}.

Theorem 3 [12]**.** In a transform coder with invertible analysis transform T, suppose that the transform coefficients are independent. If the component quantizers reconstruct to centroids, then $U = T^{-1}$ gives centroid

reconstructions for the partition induced by the encoder. As a further consequence, T^{-1} is the optimal synthesis transform.

Examples where the lack of independence of transform coefficients or the absence of optimal scalar decoding makes T^{-1} a suboptimal synthesis transform are given in Ref. 12.

3.3.2. The Analysis Transform T. Now consider the optimization of T under the assumption that $U = T^{-1}$. The first result is a counterpart to Theorem 2. Instead of requiring orthogonal transforms and finding uncorrelated transform coefficients to be best, it requires independent transform coefficients and finds orthogonal basis vectors to be best. It does not require a Gaussian source; however, it is only for Gaussian sources that $R_{\mathbf{y}}$ being diagonal implies that the transform coefficients are independent.

Theorem 4 [12]**.** Consider a transform coder in which analysis transform T produces independent transform coefficients, the synthesis transform is T^{-1}, and the component quantizers reconstruct to their respective centroids. To minimize the MSE distortion, it is sufficient to consider transforms with orthogonal rows, that is, T such that TT^T is a diagonal matrix.

The scaling of a row of T is generally irrelevant because it can be completely absorbed in the quantizer for the corresponding transform coefficient. Thus, Theorem 4 implies furthermore that it suffices to consider orthogonal transforms that produce independent transform coefficients. Together with Theorem 1, it still falls short of showing that a KLT is necessarily an optimal transform — even for a Gaussian source.

Heuristically, independence of transform coefficients seems desirable because otherwise dependencies that would make it easier to code the source are "wasted." Orthogonality is beneficial for having good partition cell shapes. A firm result along these lines requires high resolution analysis:

Theorem 5 [12]**.** Consider a high-rate transform coding system employing entropy-constrained uniform quantization. A transform with orthogonal rows that produces independent transform coefficients is optimal when such a transform exists. Furthermore, the norm of the ith row divided by the ith quantizer step size is optimally a constant. Thus, normalizing the rows to have an orthogonal transform and using equal quantizer step sizes is optimal.

For Gaussian sources there is always an orthogonal transform that produces independent transform coefficients — the KLT. For some other sources there are only nonorthogonal transforms that give independent transform coefficients, but for most sources there is no linear transform that does so. Is it more important to have orthogonality or independent transform coefficients? Examples in Ref. 12 demonstrate that there is no unequivocal answer. However, in practical situations the point is moot because it is not possible to assure independence of transform coefficients. Thus orthogonal transforms are almost always used.

4. DEPARTURES FROM THE STANDARD MODEL

4.1. Scalar and Vector Entropy Coding

Practical transform coders differ from the standard model in many ways. One particular change has significant implications for the relevance of the conventional analysis and has lead to new theoretical developments: Transform coefficients are often not entropy coded independently.

Allowing transform coefficients to be entropy coded together, as it is drawn in Fig. 1, throws the theory into disarray. Most significantly, it eliminates the incentive to have independent transform coefficients. As for the particulars of the theory, it also destroys the concept of bit allocation because bits are shared among transform coefficients.

The status of the conventional theory is not quite so dire, however, because the complexity of unconstrained joint entropy coding of the transform coefficients is prohibitive. Assuming an alphabet size of K for each scalar component, an entropy code for vectors of length N has K^N codewords. The problems of storing this codebook and searching for desired entries prevent large values of N from being feasible. Entropy coding without explicit storage of the codewords — as, for example, in arithmetic coding — is also difficult because of the number of symbol probabilities that must be known or estimated.

Analogous to constrained lossy source codes, joint entropy codes for transform coefficients are usually constrained in some way. In the original JPEG standard [27], the joint coding is limited to transform coefficients with quantized values equal to zero. This type of joint coding does not eliminate the optimality of the KLT (for a Gaussian source); in fact, it makes it even more important for a transform to give a large fraction of coefficients with small magnitude. The empirical fact that wavelet transforms have this property for natural images (more abstractly, for piecewise smooth functions) is a key to their current popularity.

Returning to the choice of a transform assuming joint entropy coding, the high-rate case gives an interesting result — quantizing in any orthogonal analysis basis and using uniform quantizers with equal step sizes is optimal. All the meaningful work is done by the entropy coder. Given that the transform has no effect on the performance, it can be eliminated. There is still room for improvement, however.

Producing transform coefficients that are independent allows for the use of scalar entropy codes, with the attendant reduction in complexity, without any loss in performance. A transform applied to the quantizer outputs, as shown in Fig. 4, can be used to achieve or approximate this. For Gaussian sources, it is even possible to design the transform so that the quantized transform coefficients have approximately the same distribution, in addition to being approximately independent. Then the same scalar entropy code can be applied to each transform

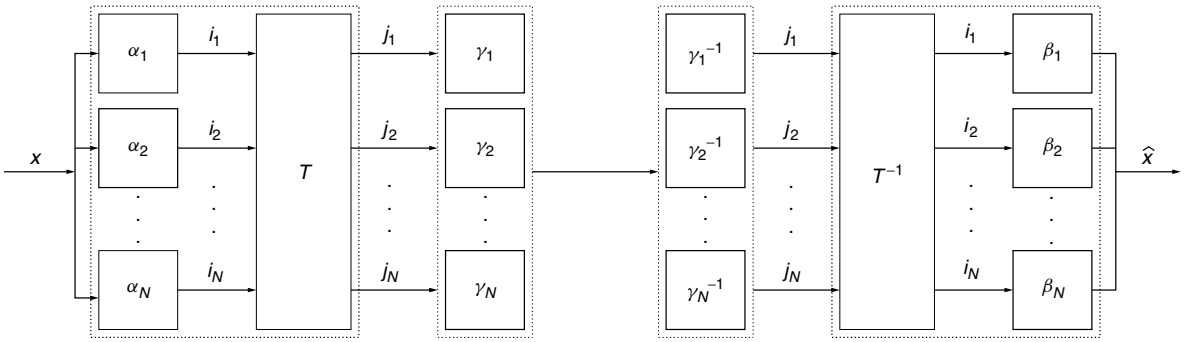


Figure 4. An alternative transform coding structure introduced in Ref. 10. The transform T operates on a vector of discrete quantizer indices instead of on the continuous-valued source vector, as in the standard model. Scalar entropy coding is explicitly indicated. For a Gaussian source, a further simplification with $\gamma_1 = \gamma_2 = \cdots = \gamma_N$ can be made with no loss in performance.

coefficient [10]. Some of the lossless codes used in practice include transforms, but they are not optimized for a single scalar entropy code.

4.2. Transmission with Losses

When source coding is separated completely from the underlying communication problem, it is assumed, as we have done so far, that bits are communicated without loss or error from the encoder to the decoder. Transforms may also be used when some data is lost in transmission though the design objectives for the transform may be changed. Some techniques for these situations have been described elsewhere [11].

5. HISTORICAL NOTES

Transform coding was invented as a method for conserving bandwidth in the transmission of signals output by the analysis unit of a 10-channel vocoder ("*vo*ice *coder*") [4]. These correlated, continuous-time, continuous-amplitude signals represented estimates, local in time, of the power in ten contiguous frequency bands. By adding modulated versions of these power signals, the synthesis unit resynthesized speech. The vocoder was publicized through the demonstration of a related device, called the *Voder*, at the 1939 World's Fair.

Kramer and Mathews [20] showed that the total bandwidth necessary to transmit the signals with a prescribed fidelity can be reduced by transmitting an appropriate set of linear combinations of the signals instead of the signals themselves. Assuming Gaussian signals, KLTs are optimal for this application.

The technique of Kramer and Mathews is not source coding because it does not involve discretization. Thus, one could ascribe a later birth to transform coding. Huang and Schultheiss [17] introduced the structure shown in the bottom of Fig. 1, which we have referred to as the *standard model*. They studied the coding of Gaussian sources while assuming independent transform coefficients and optimal fixed-rate scalar quantization. First they showed that $U = T^{-1}$ is optimal and then that T should have orthogonal rows. These results are subsumed by Theorems 3 and 4. They also obtained high-rate bit allocation results.

6. SUMMARY

The theory of source coding tells us that performance is best when large blocks of data are used. But this same theory suggests codes that are too difficult to use — because of storage, running time, or both — if the block length is large. Transform codes alleviate this dilemma. They perform well, although not optimally, but are simple enough to apply with very large block lengths.

Transform codes are easy to implement because of a divide-and-conquer strategy; the transform exploits dependencies in the data so that the quantization and entropy coding can be simple. As for the choice of the transform, two qualitative intuitions arise from high-resolution transform coding theory: (1) try to have independent transform coefficients; and (2) use orthogonal transforms. Sometimes you can only have one or the other, but for jointly Gaussian sources this works perfectly because you can have both: the transform can produce independent transform coefficients, and then little is lost by using scalar quantization and scalar entropy coding.

As with source coding generally, it is hard to make use of the theory of transform coding with real-world signals. Nevertheless, the principles of transform coding certainly do apply, as evidenced by the dominance of transform codes in audio, image, and video compression.

Acknowledgment

Artical adapted, with permission, from "Theoretical Foundations of Transform Coding," *IEEE Signal Processing Magazine*, Vol. 18, No. 5, pp. 9–21, September 2001. © 2001 IEEE. Condensed from [12].

BIOGRAPHY

Vivek K. Goyal received his B.S. degree in mathematics and his B.S.E. in electrical engineering (both with highest distinction), in 1993, from the University of Iowa, Iowa City. He received his M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1995 and 1998, respectively. In 1998 he received the Eliahu Jury Award of the University of California, Berkeley, awarded to a graduate student or recent alumnus for

outstanding achievement in systems, communications, control, or signal processing.

Dr. Goyal was a research assistant in the Laboratoire de Communications Audiovisuelles at Ecole Polytechnique Federale de Lausanne, Switzerland, in 1996. He worked in the Mathematics of Communications Research Department of Lucent Technologies Bell Laboratories as an intern in 1997 and again as a member of technical staff from 1998 to 2001. He is currently a senior research engineer for Digital Fountain, Inc., Fremont, California. Dr. Goyal is a member of Phi Beta Kappa, Tau Beta Pi, Sigma Xi, Eta Kappa Nu, IEEE, and SIAM. He serves on the program committee of the IEEE Data Compression Conference. His research interests include source coding theory, quantization theory, and practical, robust network content delivery.

BIBLIOGRAPHY

1. T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
2. J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, Vol. 290 of *Grundlehren der mathematischen Wissenschaften*, 3rd ed., Springer-Verlag, New York, 1998.
3. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
4. H. W. Dudley, The vocoder, *Bell Lab. Rec.* **18**: 122–126 (Dec. 1939).
5. M. Effros, Optimal modeling for complex system design, *IEEE Signal Process. Mag.* **15**(6): 51–73 (Nov. 1998).
6. N. Farvardin and J. W. Modestino, Optimum quantizer performance for a class of non-Gaussian memoryless sources, *IEEE Trans. Inform. Theory* **IT-30**(3): 485–497 (May 1984).
7. A. Gersho, Asymptotically optimal block quantization, *IEEE Trans. Inform. Theory* **IT-25**(4): 373–380 (July 1979).
8. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, Boston, 1992.
9. H. Gish and J. P. Pierce, Asymptotically efficient quantizing, *IEEE Trans. Inform. Theory* **IT-14**(5): 676–683 (Sep. 1968).
10. V. K. Goyal, Transform coding with integer-to-integer transforms, *IEEE Trans. Inform. Theory* **46**(2): 465–473 (March 2000).
11. V. K. Goyal, Multiple description coding: Compression meets the network, *IEEE Signal Process. Mag.* **18**(5): 74–93 (Sept. 2001).
12. V. K. Goyal, *Single and Multiple Description Transform Coding with Bases and Frames*, SIAM, 2002.
13. V. K. Goyal, J. Zhuang, and M. Vetterli, Transform coding with backward adaptive updates, *IEEE Trans. Inform. Theory* **46**(4): 1623–1633 (July 2000).
14. R. M. Gray and D. L. Neuhoff, Quantization, *IEEE Trans. Inform. Theory* **44**(6): 2325–2383 (Oct. 1998).
15. U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, Univ. California Press, Berkeley, CA, 1958.
16. H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.* **24**: 417–441, 498–520 (1933).
17. J. J. Y. Huang and P. M. Schultheiss, Block quantization of correlated Gaussian random variables, *IEEE Trans. Commun. Syst.* **11**: 289–296 (Sept. 1963).
18. A. K. Jain, A sinusoidal family of unitary transforms, *IEEE Trans. Pattern Anal. Mach. Int.* **PAMI-1**(4): 356–365 (Oct. 1979).
19. K. Karhunen, Über lineare methoden in der Wahrscheinlichkeitsrechnung, *Ann. Acad. Sci. Fenn., Ser. A.1.: Math.-Phys.* **37**: 3–79 (1947).
20. H. P. Kramer and M. V. Mathews, A linear coding for transmitting a set of correlated signals, *IRE Trans. Inform. Theory* **23**(3): 41–46 (Sept. 1956).
21. M. Loève, Fonctions aleatoires de seconde ordre, in P. Levy, ed., *Processus Stochastiques et Mouvement Brownien*, Gauthier-Villars, Paris, 1948.
22. D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.
23. D. F. Lyons III, *Fundamental Limits of Low-Rate Transform Codes*, Ph.D. thesis, Univ. Michigan, 1992.
24. N. Moayeri and D. L. Neuhoff, Time-memory tradeoffs in vector quantizer codebook searching based on decision trees, *IEEE Trans. Speech Audio Process.* **2**(4): 490–506 (Oct. 1994).
25. A. Ortega and K. Ramchandran, Rate-distortion methods for image and video compression, *IEEE Signal Process. Mag.* **15**(6): 23–50 (Nov. 1998).
26. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York, 1991.
27. W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, 1993.
28. K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press, San Diego, CA, 1990.
29. A. Segall, Bit allocation and encoding for vector sources, *IEEE Trans. Inform. Theory* **IT-22**(2): 162–169 (March 1976).
30. Y. Shoham and A. Gersho, Efficient bit allocation for an arbitrary set of quantizers, *IEEE Trans. Acoust. Speech Signal Process.* **36**(9): 1445–1453 (Sept. 1988).

TRANSMISSION CONTROL PROTOCOL

JAMES AWEYA
Nortel Networks
Ottawa, Ontario, Canada

1. INTRODUCTION

The Internet, which has become a global communication network, uses packet switching techniques to enable its attached devices — personal computers, workstations, servers, and wireless devices — to exchange information. The information is encoded as long strings of bits called *packets*. In order to achieve the transfer of packets between the attached devices, certain rules about packet format and processing must be followed. These rules are called *protocols*, and the suite of protocols used by the Internet is the TCP/IP. The Internet's marked success has made the TCP/IP protocol suite the most

ubiquitous tool for computer networking. Hence, the most widely used transport protocols today are Transmission Control Protocol (TCP) [1] and its companion transport protocol, the User Datagram Protocol (UDP) [2]. Although a number of other transport protocols have been developed or proposed [3], the success of the Internet has made TCP and UDP the dominant transport protocols for networking.

The Internet Protocol (IP) is fundamental to Internet addressing and routing, while the transport protocols provide an end-to-end connection between application processes running in the source and destination end devices. The application processes are identified by *port numbers*. Both TCP and UDP run on top of IP and build on the connectionless datagram services provided by IP. TCP provides a reliable, connection-oriented, bytestream service between applications. UDP, on the other hand, provides service without reliability guarantees. It is a lightweight protocol that allows applications to make direct use of the unreliable datagram service provided by the underlying IP service. UDP is basically an interface to IP, adding little more than multiplexing/demultiplexing (port numbers) and optional data integrity service. By not providing reliability service, UDP's overhead is significantly less than that of TCP. Although UDP includes an optional checksum, incoming datagrams with checksum errors are silently discarded and no error message is generated, while valid datagrams are passed to the application.

The reliable transport service provided by TCP is used by most Internet applications, including interactive Telnet, file transfer, electronic mail, and Webpage access via the Hypertext Transfer Protocol (HTTP). In this article, we look at the basic design and operation of TCP, particularly the main features that make TCP a reliable transport protocol.

2. BASIC TCP FEATURES

TCP provides reliability and ensures end-to-end delivery of data by adding services on top of IP. IP is connectionless and does not guarantee delivery of packets. TCP provides a full-duplex (i.e., can carry data in both directions), virtual circuit connection between two applications communicating with each other. The applications communicate across the TCP connection by exchanging a stream of 8-bit bytes in each direction. TCP groups a set of bytes that need to be sent into a message *segment* that is passed to IP. Message segments can be of arbitrary length, but for reasons of efficiency in managing messages, message segments can be limited by a *maximum segment size* (MSS) that each end has the option of announcing when a connection is established. Applications that use TCP send data in whatever size is convenient for sending. Applications can send data to TCP a few bytes (as little as one byte) or several kilobytes at a time. TCP buffers these data and sends these bytes either as single message segment or as several smaller message segments. Ultimately, the messages are sent in IP datagrams that are limited by the maximum transmission unit (MTU) of a network interface.

TCP treats the actual data it sends as an unstructured stream of bytes. It does not contain any facility to superimpose an application-dependent structure on the data. For example, an application cannot instruct TCP to treat the data as a set of records in a database and to send one record at a time. Any such structuring must be handled by the application that communicates using TCP. Because TCP sends data as a stream of bytes, there is no real end-of-message marker in the datastream.

TCP keeps track of each byte that is sent/received. It has no inherent notion of a block of data, unlike other transport protocols, which typically keep track of the Transport Protocol Data Unit (TPDU) number and not the byte number. TCP numbers each byte that it sends and the number assigned to each byte is called the *sequence number*. This number is necessary to ensure that the bytes are delivered to the application at the receiving end in the order in which they are sent. This process is called *sequencing* of the bytes.

In order to provide a reliable service, TCP must recover from data that is lost, damaged, duplicated, or delivered out of order by IP. TCP achieves this using the positive acknowledgment retransmission (PAR) scheme. TCP implements PAR by assigning a sequence number to each byte that is transmitted and requiring a positive acknowledgment (ACK) from the receiving TCP module. If the ACK is not received within a timeout interval, the data are retransmitted. At the receiver TCP module, the sequence numbers are used to correctly order segments that may have arrived out of order and to eliminate duplicates. Corruption of data is detected by using a checksum field in the TCP header. Data segments that are received with a bad checksum field are discarded.

Since Internet devices can send and receive TCP data segments at different rates because of differences in processor and network bandwidth, it is quite possible for a sending device to send data at a much faster rate than the receiver can handle. TCP implements a flow control mechanism that controls the amount of data sent by the sender. TCP uses a *sliding window* mechanism for implementing flow control. The goal of the sliding window mechanism is to keep the channel full of data and to reduce to a minimum the delays experienced in waiting for acknowledgments.

TCP enables many application processes within a single device to use the TCP services simultaneously; this is termed TCP *multiplexing*. Those processes that may be communicating over the same network interface are identified by the IP address of the network interface. TCP associates a port number value for applications that use TCP. This association enables several connections to exist between application processes on devices because each connection uses a different pair of port numbers. The binding of ports to application processes is handled independently by a device.

TCP applications must establish a connection between them before they can exchange data. A TCP connection identifies the endpoints involved in the connection. An *endpoint* is defined as a pair that includes the unique IP address of the network interface over which the application communicates, and the port number that identifies the application. The TCP connection is identified by the parameters of both endpoints as follows: {IP address1, port

number1, IP address2, port number2}. A connection is fully specified by the pair of endpoints. These parameters make it possible to have several application processes connected to the same endpoint. A local endpoint can participate in many connections to different foreign endpoints.

3. TCP MESSAGE FORMAT

A TCP message segment consists of a header part and an optional data part. The header part, which can be up to 60 bytes long, further comprises a fixed section of 20 bytes to carry 15 fields, and an optional section that can carry up to 40 bytes of TCP options.

3.1. Fixed Header Fields

The TCP header format is shown in Fig. 1. The header has a normal size of 20 bytes, unless TCP options are present.

The pair of 16-bit source and destination *port number* fields is used to identify the endpoint applications of the TCP connection. The source and destination IP addresses, and source and destination port numbers uniquely identify a TCP connection. Some port numbers are well-known port numbers; others have been registered, and still others are dynamically assigned.

The 32-bit *sequence number* identifies the first byte of the data in a message segment and is sent in the TCP header for that segment. If the SYN flag field is set to 1, the sequence number field defines the *initial sequence number* (ISN) to be used for that session, and the first data offset is ISN+1. The ISN does not take a value of 1 for a new TCP connection. The value of the ISN selected is intended to prevent delayed data from an old connection (i.e., old sequence numbers that already may have been assigned to data that are in transit on the network) from being incorrectly interpreted as being valid within the current connection.

Since TCP transmissions are full-duplex, a TCP module is both a sender and receiver of data. When a message is sent by the receiver to the sender, it also carries a 32-bit *acknowledgment number*, which indicates the sequence number of the next byte expected by the receiver. That is, the acknowledgment number is 1 plus the sequence number of the last successfully received byte of data. TCP acknowledgments are cumulative; that is, a single acknowledgment can be used to acknowledge a number of prior TCP message segments. Transmission by TCP is made reliable via the use of sequence numbers and acknowledgment numbers.

The 4-bit *data offset* (or *header length*) field is the number of 32-bit words in the TCP header. This field is needed because the TCP options field could be variable in length. Without TCP options, the data offset field is 20 bytes (5 words). The 4-bit field limits the TCP header to 60 bytes. The 6-bit field marked *reserved* is reserved for future use.

The six 1-bit flag fields in the TCP header serve the following purposes:

- Urgent Pointer Flag (URG) When the URG flag is set, it indicates that the *urgent pointer* is valid. Using these two fields, TCP allows one end of the connection to inform the other end that urgent data have been placed in the normal data stream. This feature requires that when urgent data are found, the receiving TCP should notify whatever application is associated with the connection to go into "urgent mode." After all urgent data have been consumed, the application returns to normal operation.
- Acknowledgment Flag (ACK) When the ACK flag is set, it indicates that the acknowledgment number field is valid.
- Push Flag (PSH) When the PSH flag is set, it tells TCP immediately to deliver data for this message to the upper layer process.
- Reset Flag (RST) The RST flag is used to reset the connection. When an RST is received in a TCP segment, the receiver must respond by immediately terminating the connection. A reset causes the immediate release of the connection and its resources. Sending a RST is not the normal way to close a TCP connection. The normal way where a FIN is sent after all previously queued data has been successfully sent

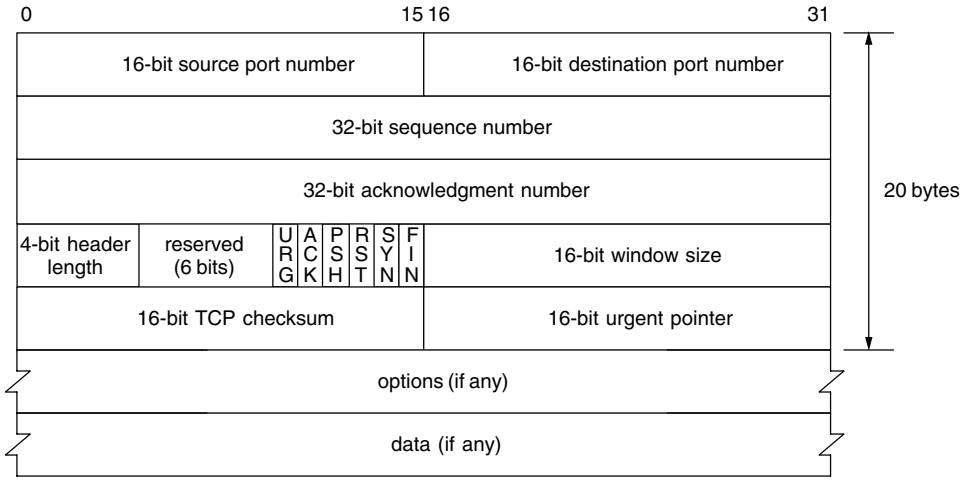


Figure 1. TCP header format.

is sometimes called an *orderly release*. Aborting a connection by sending a RST instead of a FIN is sometimes referred to as an *abortive release*. Other than for an abortive release, one common reason for generating a reset is when a connection request arrives at a destination port that is not in use by an application process.

- Synchronize Sequence Numbers Flag (SYN) The SYN flag is used to indicate the opening of a TCP connection.
- Finish Flag (FIN) The FIN flag is used to terminate the connection.

The 16-bit *window size* field is used to implement flow control and reflects the amount of buffer space available for new data at the receiver. The receiving TCP reports a window to the sending TCP, and this window specifies the number of bytes, starting with the acknowledgment number, that the receiving TCP is currently prepared to receive. The 16-bit window limits the window to 65,535 bytes, but a window scale option allows this value to be scaled to provide larger windows.

The 16-bit *checksum*, which is mandatory, is used to verify the integrity of the TCP header as well as the data. The TCP checksum is an end-to-end checksum. It is calculated by the sender and then verified by the receiver. The checksum is the one's complement of the one's-complement sum of all the 16-bit words in the TCP packet. A 12-byte pseudoheader (see Fig. 2) is prepended to the TCP header for checksum computation. The pseudoheader is used to identify whether the packet has arrived at the correct destination. The pseudoheader gives the TCP protection against misrouted segments. The TCP segment can be an odd number of bytes, while the checksum algorithm adds 16-bit words. So a pad byte of 0 is appended to the end, if necessary, just for checksum computation. The 16-bit TCP length field in Fig. 2 (which is a computed quantity and appears twice in the checksum computation) is the TCP header length plus the data length in bytes. This field does not count the 12 bytes of the pseudoheader.

The URG and *urgent pointer* fields constitute a mechanism by which TCP marks urgent data when transmitting segments. The 16-bit urgent pointer is a positive offset when added to the sequence number field of the segment, indicates the sequence number of the last byte in the sequence of urgent data. This feature enables the sender to send interrupt signals to the receiver and prevents these signals from ending up in the normal data queue at the receiver.

3.2. TCP Options

Many options can be specified in the *TCP options* field up to a maximum of 40 bytes as allowed by the 4-bit data offset field (which limits the TCP header to 60 bytes). A number of TCP options have been defined [1,4,5]; however, the current options relevant to TCP performance are

- *Maximum Segment Size* (*MSS*) *Option*. The MSS option [1] is used only when a connection is being established, and appears only in the initial SYN segment used to open the connection. A TCP sender uses this option to inform the remote end of the largest unit of information or maximum segment size it is willing to receive on the TCP connection. The setting of the MSS option can be up to the MTU of the outgoing interface minus the size of the TCP and IP headers. The MSS option when used together with path MTU discovery [6] allows for the establishment of a segment size that can be sent across the path between two hosts without fragmentation. Fragmentation is undesirable because if one fragment is lost, TCP will timeout and retransmit the entire TCP segment.
- *Window-Scale Option*. This option allows TCP to use window sizes that can operate efficiently over large-bandwidth-delay networks. The 16-bit window size field in the TCP header limits the window to 65,535 bytes. However, most networks, in particular high-speed networks and networks with satellite links, will require a much higher window than this for maximum TCP throughput. The window-scale option effectively increases the size of the window to a 30-bit field, but only the most significant 16 bits of the value are transmitted. This allows larger windows, up to 2^{30} bytes, to be transmitted [4]. This option can only be sent in a SYN segment at the start of the TCP connection.
- *SACK Option*. The selective acknowledgment (SACK) option in TCP is defined in RFC 2018 [5]. This option is used to modify the acknowledgment

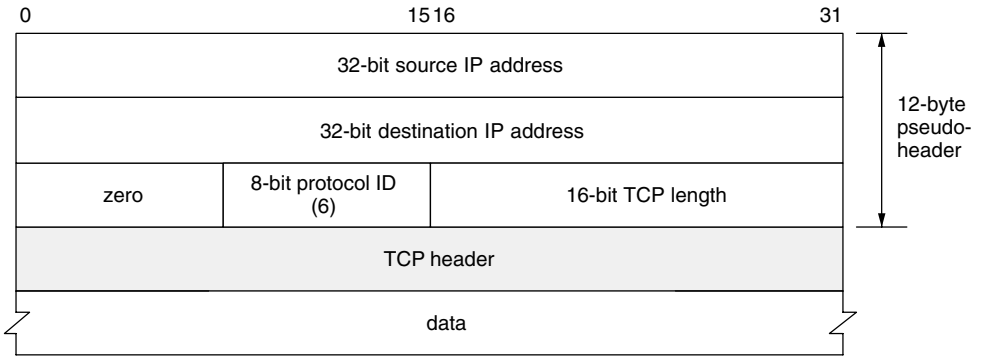


Figure 2. Fields used for TCP checksum computation.

behavior of TCP. The default behavior is to acknowledge the highest sequence number of bytes received in order. The SACK option allows for more robust operation when multiple segments are lost from a single window of data. It enables a TCP receiver to inform the sender of what specific segments were lost (selective acknowledgment) so that the TCP sender can retransmit them. Thus, when faced with multiple segment losses in a single window, TCP SACK enables a sender to continue to transmit segments (retransmissions and new segments) without entering into a time-consuming slow-start phase.

The *data* portion of the TCP segment is optional; a TCP segment does not necessarily need to carry user data. In particular, when a connection is established, and when a connection is terminated, segments with TCP header and possibly with options, are exchanged. The TCP segment used to acknowledge received data when there are no data to be transmitted in that direction consists of a header only.

4. TCP OPERATION

Because IP provides no sequencing or acknowledgment of data and is connectionless, the tasks of connection establishment and termination, reliability in data transfer, data sequencing, and flow control are given to TCP [1]. These operational features of TCP are discussed in this section.

4.1. TCP Connection Establishment

Two applications using TCP must establish a *connection* with each other before they can exchange data. A TCP connection is established using a *three-way handshake*, which ensures that both ends of the connection have a clear understanding of the initial sequence number, and possibly, the TCP options of the remote end. The three steps involved in the three-way handshake are summarized as follows:

Step 1 — endpoint 1 sends a SYN segment (with flag fields SYN = 1, ACK = 0) specifying its initial sequence number, x, and the port number to endpoint 2.

Step 2 — endpoint 2 responds with a SYN segment (with flag fields SYN = 1, ACK = 1) containing its own initial sequence number, y, and an acknowledgment (ACK) of the received initial sequence number (acknowledgment number is 1 greater then x).

Step 3 — endpoint 1 acknowledges the received remote sequence number by sending a segment (with flag fields SYN = 0, ACK = 1) containing the acknowledgment number $y + 1$.

The three steps are shown in Fig. 3. It takes three segments to establish a connection, and 1.5 round-trip times (RTTs) for the two end systems to synchronize state before data exchange. An *active open* is said to be performed by the endpoint that sends the first SYN, while a *passive open* is performed by the other end that receives this first SYN and sends the next SYN.

After a TCP connection has been established, the ACK flag is always set to 1 to indicate that the acknowledgment number field is valid. Data transmission then takes place with messages being exchanged by both sides in a full-duplex fashion until the TCP connection is ready to be closed. When an application process notifies TCP that it has no more data to send, TCP will close the connection in *the sending direction*. Thus, each direction of data flow must be shut down independently since a TCP connection is full-duplex, and can be viewed as containing two independent stream transfers, one going in each direction. *TCP half-close* refers to the ability of one end of a TCP connection to terminate data transfer (except the output of acknowledgment segments) while still receiving data from the other half.

The FIN control flag is issued when a TCP connection is ready to close. One end sends a segment with the FIN flag set to close its half of the connection when it has finished sending data. The receiving TCP acknowledges the FIN segment and notifies the application process at its end that no more data will be delivered to it. However, data can continue to flow in the other direction (i.e., other half of the connection) until that direction of data flow is closed. Acknowledgment segments continue to flow back

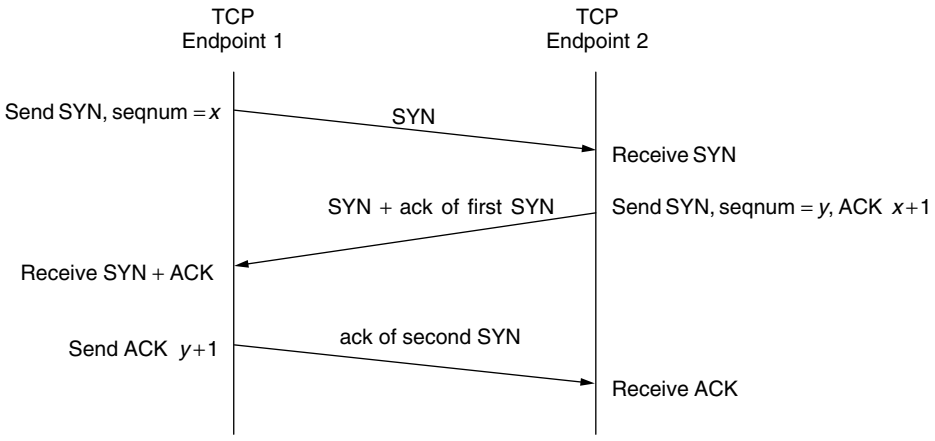


Figure 3. TCP connection establishment.

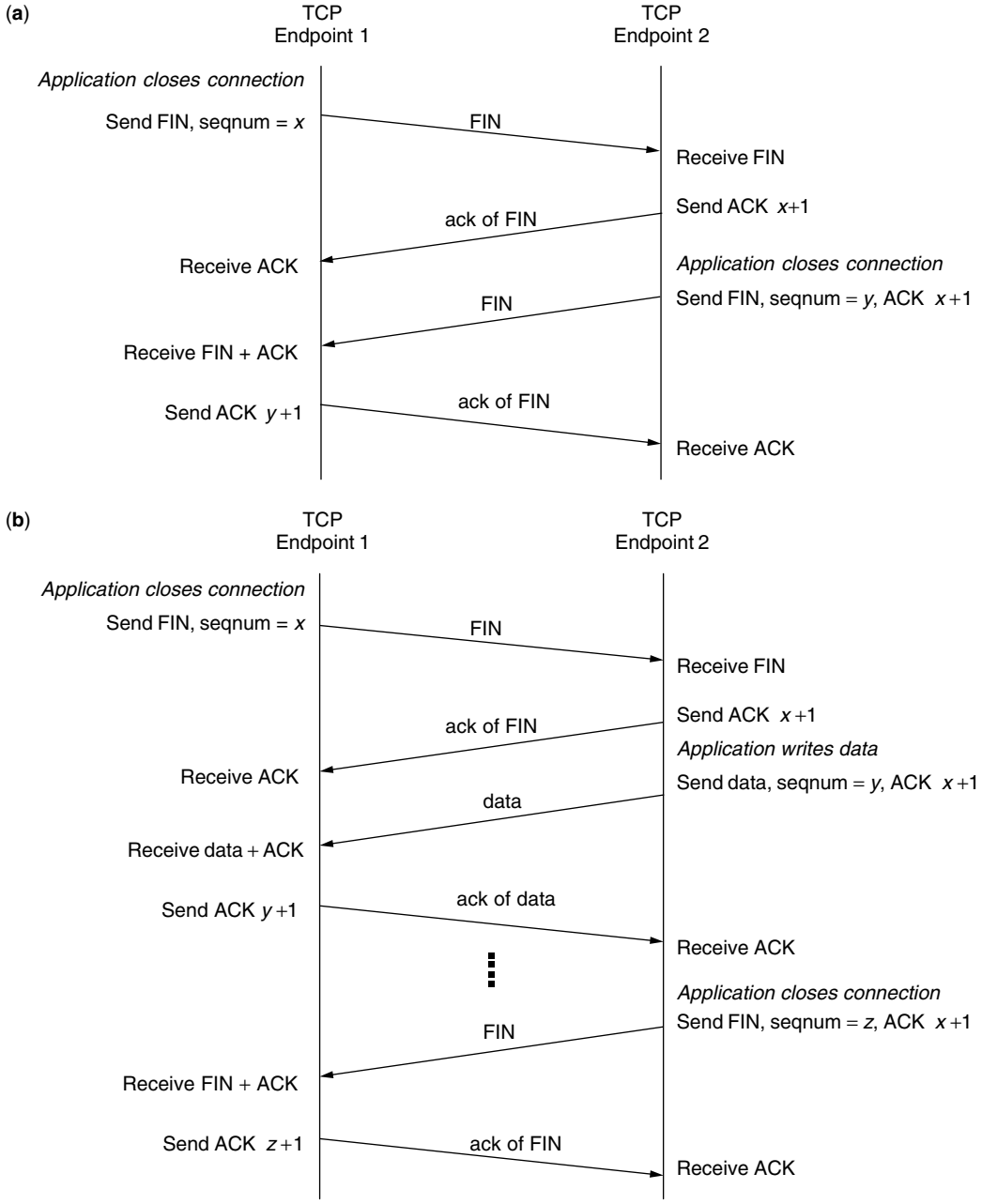


Figure 4. TCP connection termination: (**a**) normal connection termination; (**b**) TCP half-close.

to the sender even after a connection is half-closed. The end that sends the first FIN is said to perform an *active close* and the end that receives this FIN performs a *passive close*. The connection termination procedure is illustrated in Fig. 4. When both directions have been closed, then data completely stop flowing in both directions. In practice, few TCP applications use the half-close feature.

4.2. Data Transfer

TCP can handle both interactive data transfers and bulk data transfers. However, the methods of flow control and data acknowledgment differ between these two application areas.

4.2.1. Interactive Data Transfer. *Telnet* and *rlogin* are examples of applications that generate interactive data. Interactive applications typically send data in very small units. For example, with rlogin, only one byte is sent in a segment. Obviously, this type of operation translates into considerable overhead, since this one byte generates a 41-byte packet: 20 bytes for the TCP header and 20 bytes for IP header. Some performance improvements in data exchange can be obtained through the use of *delayed acknowledgments* [7] and *ACK piggybacking*, where the receiver holds back the ACK for a brief time (*delayed acknowledgment*) and attempts to piggyback the ACK onto data going back to the sender (*ACK piggybacking*). This helps reduce the number of segments sent into the network

since some interactive applications, such as rlogin, require that the receiver echo back the character (byte) that is received. Some operations of interactive data exchange are illustrated in Fig. 5.

For short-delay links, the high-overhead packets (called *tinygrams*) do not pose significant performance problems in the interactive data exchange. However, for slow large-delay links, these high-overhead packets can be a source of congestion traffic. A mechanism commonly called the *Nagle algorithm* was proposed in RFC 896 [8] to reduce the number of these small packets. The Nagle algorithm allows only one small segment to be outstanding at a time without acknowledgment. If more small segments are generated while awaiting the acknowledgment for the first one, then these segments are coalesced into a larger segment and sent when the acknowledgment arrives. On short-delay links where the return of ACKs is faster, the algorithm has negligible impact on data transfer. However, on slow links, where it is desirable to reduce the number of small packets, fewer such packets are transmitted. Although the Nagle algorithm can improve data transfer efficiency by transmitting packets with higher payload-to-header ratios, it may not be appropriate for all interactive applications, especially applications that are jitter sensitive. Jitter sensitive interactive applications typically demand that the small messages they generate be delivered without delay to provide real-time feedback to users. Using the Nagle algorithm can result in an increase in the session jitter by up to a round-trip time (RTT) interval. Interactive applications that are jitter-sensitive typically disable this algorithm.

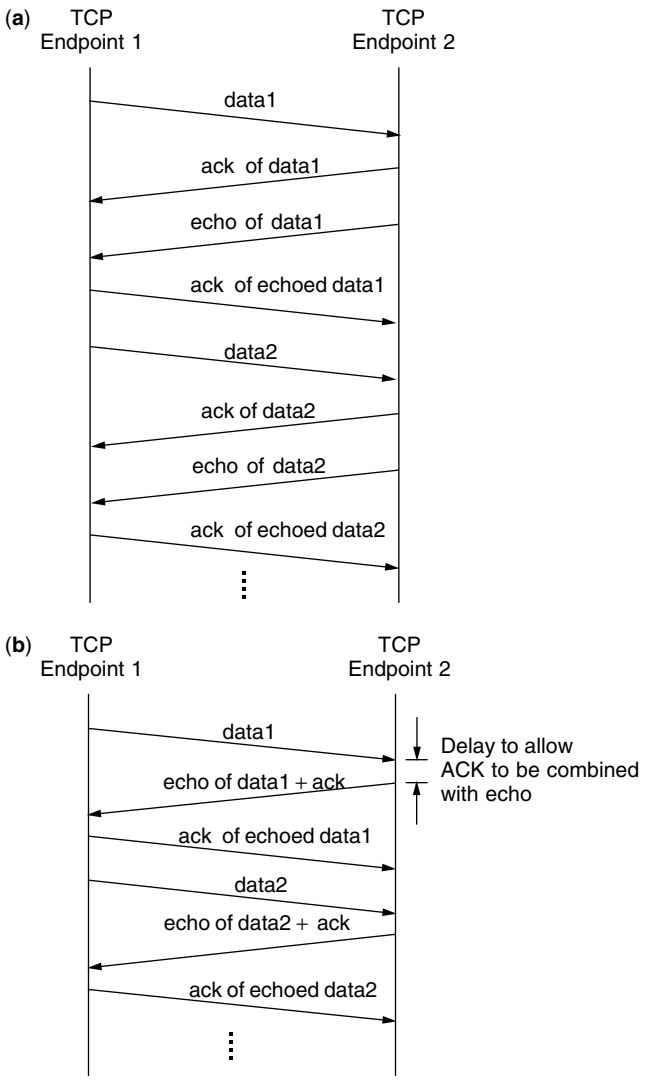


Figure 5. Interactive data exchange: (**a**) without delayed ACK; (**b**) with delayed ACK.

4.2.2. Bulk Data Transfer. This section describes the mechanisms used by TCP for bulk data transfer. These mechanisms allow a sender to transmit multiple data segments before receiving an acknowledgment for those segments. They also allow a TCP sender to respond to congestion and data loss at any point along the network path between the sender and the receiver.

4.2.2.1. Sliding-Window Protocol. TCP uses a *sliding-window* mechanism for bulk data transfer (see Fig. 6). The stream of data has a sequence number assigned at the byte level. The receiver TCP module sends back to the sender an acknowledgment that indicates the range of acceptable sequence numbers beyond the last segment successfully received. This range of acceptable sequence numbers is called a *window*. The window, therefore, indicates the number of bytes that the sender can transmit before receiving further permission.

In Fig. 6, bytes that are to the left of the window range have already been sent and acknowledged. Bytes in the window range can be sent without any delay. Some of the bytes in the window range may already have been sent, but they have not been acknowledged. Other bytes may be waiting to be sent. Bytes that are to the right of the window range have not been sent. These bytes can be sent only when they fall in the window range.

The left edge of the window is the lowest numbered byte that has not been acknowledged. The window can advance; that is, the left edge of the window can move to the right when an acknowledgment is received for the data that have been sent. The window size (called the

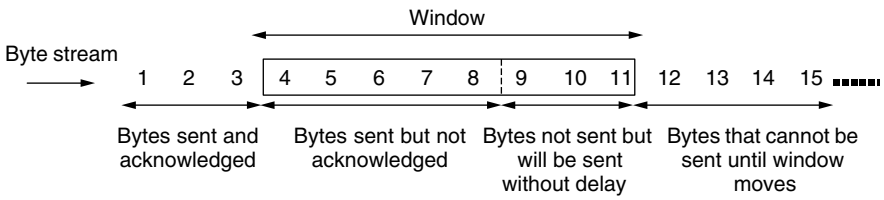


Figure 6. TCP sliding-window flow control.

"receiver's advertised window size") reflects the amount of buffer space available for new data at the receiver. When the receiving TCP process frees up buffer space by reading acknowledged data, the right edge of the window moves to the right. If this buffer space size shrinks because the receiver is being overrun, the receiver will send back a smaller window size. In the extreme case, it is possible for the receiver to send a window size of only one byte (instead of waiting until a larger window could be sent), which means that only one byte can be sent. This situation is referred to as the "silly-window syndrome," and most TCP implementations take special measures to avoid it. The silly-window syndrome can also be caused by the sender when it generates and transmits small amounts of data at a time instead of waiting to generate enough data that can be sent in larger segments.

When a TCP module sends back a window size of zero, it indicates to the sender that its buffers are full and no additional data should be sent. This happens when the left edge of the window reaches the right edge. The sliding window mechanism allows TCP to shrink the window size when the receiver experiences congestion of data and to expand the window size as the congestion problem clears.

It takes about one round-trip time (RTT) between sender and receiver for a sender to receive acknowledgment to data sent. For maximum transfer efficiency, the TCP sender must be able to completely fill the data pipe of the connection with data. Thus, the size of the window offered by the TCP receiver must be large, since it limits how much the sender can transmit. Making the advertised window no smaller than the *bandwidth-delay product* (i.e., capacity in bytes or bits) of the connection path allows maximum data transfer. This results in a window size of

$$\text{Window size} \geq \text{bandwidth (bps)} \times \text{round-trip time (sc)}$$

4.2.2.2. TCP Congestion Control Protocols. We have seen above that either the bandwidth or the RTT of a TCP connection can affect its window size. We have also seen that the larger the window size, the more data can be sent across the connection. However, given a window size, a TCP sender adopts a more controlled behavior in utilizing that window size. This is because a TCP sender initially has no idea of the available network capacity. Also, if a TCP sender commences by injecting a full window size into the network, particularly using a large window size, then there is a strong chance that much of this burst of data would be lost because of transient congestion in the network. Congestion can occur in the network nodes (e.g., routers) when data arrive on a large-bandwidth link and exit on a smaller-bandwidth link. Congestion can also occur when multiple input datastreams converge at a network node whose output bandwidth is less than the sum of the inputs. For these reasons, TCP starts by transmitting small amounts of data noting that these have a better chance of getting through to the receiver, and then probing the network with increasing amounts of data until the network shows signs of congestion. When TCP determines that the network is showing signs of congestion, it reduces its sending rate and then resumes the probing for additional bandwidth.

This seemingly seesaw behavior of TCP is its way of locating the point of equilibrium of maximum network efficiency where its sending rate is maximized just prior to the onset of sustained data loss. The bandwidth probing action of the TCP sender also helps it locate the point at which its data transmission rate is synchronized to the data extraction rate of the receiver. At this point, the rate of return of ACKs to the sender is identical to the rate of transmission of data segments. This is called the *self-clocking* behavior of TCP. This is so called because the receiver can only generate ACKs when data arrive, and the rate of arrival of the ACKs at the sender identifies the arrival rate of the segments at the receiver.

Current TCP implementations contain a number of mechanisms aimed at controlling data transmission and network congestion [9–11]. In addition to the receiver's advertised window (i.e., the buffer size advertised in acknowledgments), a TCP sender maintains a second window called the *congestion window* (*cwnd*). The *cwnd* tracks the maximum amount of data a sender can transmit before an ACK is required. TCP data transfer commences in a phase called *slow start*. The limit of the slow-start process is maintained in the state variable *ssthresh*. The initial *cwnd* used by the sender at the start of this phase is set to a segment size that is equal to any one of the following: the MSS obtained during the three-way handshake, the segment size resulting from the use of the path MTU discovery protocol, the MTU of the sending TCP interface less the TCP and IP headers, or the default segment size of 536 bytes. TCP then sends a segment not exceeding this first window size and waits for the corresponding acknowledgment.

When the acknowledgment is received, *cwnd* is incremented from 1 to 2, which means two segments can be sent. When each of these two segments is acknowledged, the congestion window *cwnd* is increased to four. Each time an ACK is received, *cwnd* is increased by one segment. This essentially leads to an exponential increase in the congestion window in slow start if the receiver sends an ACK for every segment received. The rate of congestion window increase would be slightly lower if the receiver implements delayed ACKs; nevertheless, the rate of increase is rapid. TCP maintains the congestion window *cwnd* in bytes, but always increments it by the segment size in slow start. Each time, the actual window size that is used by the TCP sender is the smaller of the congestion window and the receiver's advertised window.

The slow-start phase is terminated when *cwnd* equals the value of *ssthresh*, the receiver's advertised window, or when congestion (or packet loss) occurs. TCP then goes into what is termed the *congestion avoidance* phase, where it takes a more conservative approach in probing for network bandwidth. In this phase, the congestion window *cwnd* is incremented by 1/*cwnd* each time an ACK is received. Congestion avoidance thus increases the TCP sending rate in a linear fashion. In other words, the *cwnd* is incremented by at most one segment per round-trip time (regardless of how many ACKs are received) until congestion is detected, or the receiver's advertised window is reached. Unlike slow start's exponential increase, where *cwnd* is incremented by the number of ACKs received in a

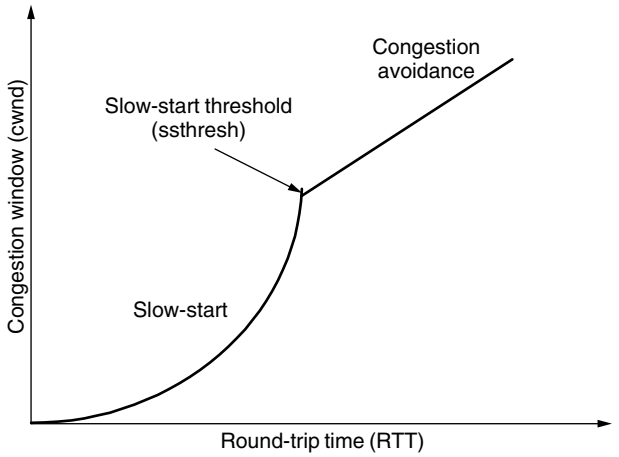


Figure 7. The slow-start and congestion avoidance phases.

RTT, congestion avoidance adopts an additive increase process. The slow-start and the congestion avoidance phases are illustrated in Fig. 7. The *ssthresh* is initially set to the receiver's maximum window size. However, when congestion occurs, the *ssthresh* is set to one-half of the current window size (i.e., the minimum of *cwnd* and the receiver's advertised window, but at least two segments). This provides TCP with the best guess of the point where network congestion could occur in the future.

TCP uses the reception of *duplicate* ACKs or the expiration of a time (*timeout*) to detect when a data segment loss occurs. The behavior of TCP when a segment is lost or received out of order can be explained as follows:

1. When a single segment is lost in a sequence of segments, the successful reception of segments following the lost segment will cause the receiver to issue a duplicate ACK for each successfully received segment.
2. When the TCP receiver gets a segment with an out-of-order sequence number value, the receiver is required to generate an immediate ACK of the highest in-order data byte received (a duplicate ACK of an earlier transmission). The purpose of the duplicate ACK is to let the sender know that a segment was received out of order, and also, the sequence number expected.
3. However, if a segment at the end of a sequence of segments is lost, there are no subsequent segments to generate duplicate ACKs. In this case, the sender's *retransmission time* (discussed below) will expire (a *timeout* will occur) since there are no corresponding ACKs for this segment and the sender cannot wait indefinitely for a delayed ACK.

From cases 1 and 2, it is difficult for a TCP sender to know whether the reception of a duplicate ACK is an indication of a lost segment or just an out-of-order data delivery. Thus, a TCP sender must wait for a small number of duplicate ACKs (typically three) before deciding whether a segment is lost or simply out of order. The assumption here is that if there is just an out-of-order delivery, the receiver will issue only one or two duplicate ACKs before the out-of-order segment is processed, at which time a new ACK will be generated. However, if three or more duplicate ACKs are issued in a row, then there is a strong chance that a segment has been lost.

When the TCP sender receives three duplicate ACKs, it immediately retransmits the lost segment (*fast retransmit*) and enters into a phase called *fast recovery*. Fast retransmit enables a TCP sender to rapidly recover from a single lost segment, or one that is delivered out of sequence, without shutting down the *cwnd*. The receiver responds with a cumulative ACK for all segments received up to that point. *Fast recovery* is based on the notion that, since the subsequent packets generating the duplicate ACKs were successfully transmitted through the network, there is no need to enter slow start and dramatically reduce the data transfer rate; therefore *cwnd* should stay open. Thus, in fast recovery, the *ssthresh* is set to half of the current window size and *cwnd* is set three segments greater than *ssthresh* to allow for three segments already buffered at the receiver. On the arrival of each additional duplicate ACK, *cwnd* is incremented by a segment allowing more data to be sent. When an ACK arrives that acknowledges new data, *cwnd* is set back to *ssthresh*, and TCP enters the congestion avoidance mode.

In the congestion avoidance mode, the TCP sender increments *cwnd* by one segment every RTT until data loss occurs or the receiver's advertised window is reached. If the loss is isolated to a single segment, the resultant duplicate ACKs will cause the sender to halve *cwnd* and then continue a linear growth of *cwnd* from this new point.

The expiration of a retransmission timer may indicate more serious congestion. In this case, the *ssthresh* is set to half of the current window size. But because the sender has not received any useful information for more than one RTT, it adopts a more conservative approach where it closes the congestion window *cwnd* back to one segment, and resumes data transmission in the slow-start mode.

Experimentation has shown that starting off with a larger initial window size of three or four segments in slow start will allow more segments to flow into the network, generating more ACKs, and will decrease the time it takes to complete slow start [12]. Because the *cwnd* can open up faster as a result, better performance is gained, in particular for small files transmitted over links with long RTTs. The short-duration TCP sessions typical of Web fetches can benefit a lot from this feature. However, a starting value of four segments in slow start may be too many for low-speed links with limited buffers, so a starting value of not more than two is a more robust approach [12].

4.2.3. TCP Timeout and Retransmission. To provide reliable data transfer, a TCP receiver is expected to acknowledge the data that it receives from the sender. Every time a segment is sent to a receiver, the sending TCP starts a timer and waits for an acknowledgment. If the timer expires before the receiver acknowledges receipt of the data in the segment, the sending TCP assumes that the segment was lost or corrupted and retransmits it.

TCP monitors the delay performance of each connection to derive reasonable estimates for timeouts. The timeout

estimates are adjusted from time to time to account for changes in the delay performance of the connection. This is because successive TCP segments in a connection may not be sent on the same path and may experience different delays as they traverse different sets of routers and links. To accommodate the varying network delays experienced by segments in a connection, TCP uses an adaptive retransmission algorithm. The measurement of the round-trip time (RTT) experienced on a connection forms the key input to this retransmission algorithm.

TCP measures the elapse time between sending a data byte with a particular sequence number and receiving an acknowledgment that covers that sequence number. This measured elapse time is the sample round-trip time (*Sample_RTT*). On the basis of the *Sample_RTT*, a smoothed round-trip time (*SRTT*) is computed using the exponentially weighted moving average filter

$$SRTT \leftarrow \alpha \times SRTT + (1 - \alpha) \times Sample\_RTT$$

where α is a weighting factor whose value is between 0 and 1 ($0 \leq \alpha < 1$). TCP then computes a retransmission timeout (RTO) value that is a function of the current estimated round-trip time (*SRTT*). RFC 793 [1] recommended the RTO to be computed as

$$RTO = \beta \times SRTT$$

where β is a constant delay weighting factor ($\beta > 1$) with a recommended value of 2. It has been shown [9] that this computation does not adapt to wide fluctuations in the RTT, thus causing unnecessary retransmissions that add to network traffic. Jacobson [9] suggests that both the smoothed RTT estimates (i.e., mean values) and the variance in the RTT measurements should be maintained. These can then be used to compute the RTO instead of just computing the RTO as constant multiple of the smoothed RTT. These changes make TCP more responsive to wide fluctuations in the round-trip times and yield higher throughput. Thus, the following computations are required for each RTT measurement:

$$\begin{gathered} Diff = Sample\_RTT - SRTT \\ SRTT \leftarrow SRTT + \eta \cdot Diff \\ Dev \leftarrow Dev + \theta \cdot (|Diff| - Dev) \\ RTO = SRTT + \phi \cdot Dev \end{gathered}$$

where *Dev* is the smoothed mean deviation, η and θ are filter gains with values in the range $0 < \eta, \theta < 1$, and ϕ is a factor that determines how much influence the deviation has on the RTO. For efficient TCP implementation, η and θ are typically selected to each be an inverse power of 2 so that the operations can be done using shifts instead of multiplies and divides. Jacobson [9] specified $\phi = 2$, but after further research, it was changed to $\phi = 4$ [13].

Although the RTO estimation process described above is relatively straightforward, a problem occurs when a segment is retransmitted. Let us say that TCP transmits a segment, a timer expires and TCP retransmits the segment. When an acknowledgment is received, the sender has no way of knowing whether the acknowledgment corresponds to the original or retransmitted segment. Perhaps the first segment was delayed and not lost or the ACK for the first segment was simply delayed. This situation is sometimes referred to as *acknowledgment ambiguity*. Now, what should TCP do with regard to RTT estimates when the TCP acknowledgments are ambiguous? The answer to this problem is provided by *Karn's algorithm* [14], which adjusts the estimated round-trip time only for unambiguous acknowledgments. When retransmissions take place in Karn's algorithm, the SRTT is not adjusted. Instead, a *timer backoff* strategy is used to adjust the timeout by a factor. The idea is that if a retransmission occurs, something drastic could have happened in the network, and the timeout should be increased sharply to avoid further retransmission. The timer backoff strategy is implemented with a multiplicative factor γ as

$$New\_timeout = \gamma \times timeout$$

Typically, γ is set to 2 so that this algorithm behaves like a binary exponential backoff algorithm. The algorithm uses the SRTT to compute an initial timeout value (*timeout*), then backs off the timeout on each retransmission (to obtain *New_timeout*) until a segment is successfully transmitted. The timeout value that results from backoff is retained when subsequent segments are sent. When an acknowledgment is received for a segment that does not require retransmission, TCP measures the RTT value (*Sample_RTT*) and uses this value to compute the SRTT and the RTO values.

5. TCP FUTURES

The rapid growth of the Internet continues to evolve TCP in several ways. In particular, it is becoming more apparent that enhancements to TCP are required to address the challenges presented by the different transmission media encountered in the Internet. High-speed (optical fiber) links, long and variable delay (satellite) links, lossy (wireless) networks, asymmetric paths (in hybrid satellite networks), and other linkages are becoming widely embedded in the Internet. These different transmission media present a wide range of characteristics that can cause degradation of TCP performance. For example, long propagation delay and losses on a satellite link, handover and fading in a wireless network, bandwidth asymmetry in some media, and other phenomena have been shown to seriously affect the throughput of a TCP connection. Also, TCP assumption that all losses are due to congestion becomes quite problematic over wireless links. TCP considers the loss of packets as a signal of network congestion and reduces its window consequently. This results in severe throughput deterioration when packets are lost for reasons other than congestion. Noncongestion losses are mostly caused by transmission errors.

Some of the techniques currently under study to enhance TCP performance involve modifying TCP to help it cope with these new media types. Others keep the protocol unchanged, but place some intelligence in the intermediate network elements along TCP connections to

provide faster and more accurate feedback to the TCP endpoints about the conditions on the connection.

BIOGRAPHY

James Aweya received his B.Sc. degree in electrical and electronics engineering from the University of Science & Technology, Kumasi, Ghana, the M.Sc. degree in electrical engineering from the University of Saskatchewan, Saskatoon, Canada, and a Ph.D. degree in electrical engineering from the University of Ottawa, Canada. Since 1996, he has been with Nortel Networks where he is currently a systems architect in the Advanced Technology Group. His current activities include the design of resource management and control functions for communication networks, the development and analysis of new network architectures and protocols, and the design and analysis of switch and router architectures. He has published more than 50 journal and conference papers and has a number of patents pending. Dr. Aweya was the recipient of the IEEE Communications Magazine Best Paper Award at OPNETWORK 2001. In addition to communication networks and protocols, he has other interests in neural networks, fuzzy logic control, and application of artificial intelligence to computer networking. Dr. Aweya also collaborates with a number of researchers at the University of Ottawa and Carleton University, Ottawa, Canada, to research issues on communication network design and quality of service control.

BIBLIOGRAPHY

1. J. Postel, *Transmission Control Protocol*, RFC 793, Sept. 1981.
2. J. Postel, *User Datagram Protocol*, RFC 768, Aug. 1980.
3. S. Iren, P. D. Amer, and P. T. Conrad, The transport layer: tutorial and survey, *ACM Comput. Surv.* **31**(4): 360–405 (Dec. 1999).
4. V. Jacobson, R. Braden, and D. Borman, *TCP extensions for High Performance*, RFC 1323, May 1992.
5. M. Mathis, J. Madavi, S. Floyd, and A. Romanov, *TCP Selective Acknowledgment Options*, RFC 2018, Oct. 1996.
6. J. Mogul and S. Deering, *Path MTU Discovery*, RFC 1191, Nov. 1990.
7. R. Braden, ed., *Requirements for Internet Hosts—Communication Layers*, RFC 1122, Oct. 1989.
8. J. Nagle, *Congestion Control in IP/TCP Internetworks*, RFC 896, Jan. 1984.
9. V. Jacobson, Congestion avoidance and control, *ACM Comput. Commun. Rev.* **18**(4): 314–329 (Aug. 1988).
10. W. R. Stevens, *TCP/IP Illustrated*, Vol. 1, Addison-Wesley, Reading, MA, 1994.
11. M. Allman, V. Paxson, and W. Stevens, *TCP Congestion Control*, RFC 2581, April 1999.
12. M. Allman, S. Floyd, and C. Partridge, *Increasing TCP's Initial Window*, RFC 2414, Sept. 1998.
13. V. Jacobson, Berkeley TCP evolution from 4.3-Tahoe to 4.3 Reno, *Proc. 18th Internet Engineering Task Force*, Univ. British Columbia, Vancouver, BC, Sept. 1990.
14. P. Karn and C. Partridge, Improving round-trip time estimates in reliable transport protocols, *Comput. Commun. Rev.* **17**(5): 2–7 (Aug. 1987).

TRANSPORT PROTOCOLS FOR OPTICAL NETWORKS

TIM MOORS
University of New South Wales
Sydney, Australia

1. INTRODUCTION

Transport protocols are used in the end systems that are connected by communication networks to match the characteristics and requirements of the network to those of the users of the communication network in the end systems. In particular, transport protocols are used to match the aspects of reliability, security, transmission unit size, and timing of information transfer. For example, when the network can corrupt, lose, missequence, or duplicate information, a transport protocol may be used to enhance the reliability of the end-to-end path to match the requirements of end users. A transport protocol may also segment information from its users to meet the maximum transmission unit requirements of the network, and may pace the transmission of information supplied by the user so as to prevent or limit congestion in the network. In addition to matching the network path to users, a transport protocol may also be used to match users to each other, such as by controlling the flow of information from source to destination(s) so that its rate does not exceed the capacity of the destination(s).

Optical networks have grown in importance since the introduction of low-attenuation fiber in the 1970s. This article focuses on networks constructed from optical fiber, although many of the principles also apply to networks that employ free-space optics. The first generation of optical fiber networks used optical transmission systems, and converted the optical signal to electronic form for switching. More recently, optical networks have appeared in which the switching is also done optically, creating "all-optical networks," in which a "lightpath" extends between communicating endpoints, with optoelectronic conversion only occurring in the end systems, not along the end-to-end path.

Since the end systems, where transport protocols are implemented, always provide optoelectronic conversion (even when these end systems are connected to all-optical networks), transport protocols are generally implemented in the electronic domain. This leads to a common misconception that existing transport protocols that were designed for electronic networks are also appropriate and optimal for optical networks, and directs the attention for innovation in optical networks toward the transmission and switching systems. In reviewing transport protocols for optical networks, this article will show the fallacy of this misconception: that transport protocols are *strongly* affected by the existence of optical networks.

This article starts by reviewing the features of optical networks that are particularly salient to transport protocols. It then provides an overview of the generic functions that transport protocols are expected to provide, and discusses how these functions are affected by the use of optical networks. Section 4 then describes specific transport protocols that have been used with optical networks. Finally, transport protocols for optical networks need to be implemented in a manner that is appropriate for the high speeds of optical networks. In Section 5, this article concludes by considering such implementation issues.

Before considering transport protocols for optical networks in depth, it is worth making two points about terminology and the scope of transport protocols:

1. The term "transport protocol," as used in this article and in the context of computer communication systems [e.g., 1], refers to an end-to-end protocol. This can be confusing because in the context of optical transmission systems (in particular, in the context of the International Telecommunication Union's G series of recommendations [e.g., 2]) the term "transport" often refers to the transmission of information between adjacent points, such as across one of the multiple hops of an end-to-end communication path.
2. While the most prominent role of transport protocols in optical networks is in end systems, they are also used in routers, switches, and other intermediate systems to ensure the reliable delivery of information between such systems. For example, the popular Border Gateway Protocol for routing sends routing topology update information over the transport protocol TCP.

Despite these different uses of the term "transport" and different applications of transport protocols, this article will focus on the end-to-end transport protocols that are used with optical networks.

2. SALIENT FEATURES OF OPTICAL NETWORKS

Optical fibers can be readily manufactured with loss of a couple of decibels (e.g., 1–4 dB) per kilometer. This allows optical networks to span vast distances, and a transport protocol for an optical network may need to deal with large propagation delays, simply because of the large distances and the limited propagation speed of light. Comparing networks that span the same distance, optical networks may exhibit lower end-to-end delay than their electronic or radio-based counterparts. This is not due to inherently faster propagation of optical signals, since they propagate at a speed of around 2×10^8 m/s in silica fiber — a figure that is comparable to the propagation speed in electrical wiring or radio transmission. Rather, optical networks tend to reduce the delays that the signal will incur as it passes through intermediate systems for two reasons: (1) optical networks tend to minimize buffers because they are expensive — it is difficult to buffer information in the optical domain (e.g., using delay lines built using long threads of fiber) for all-optical networks, and electronic buffers that can match the bandwidth of optical transmission systems are expensive for optoelectronic networks; and (2) the high transmission rates of optical networks mean that large volumes of information can be "in flight" from source to destination at any time. Rather than allow all this information to accumulate in a buffer in a router or switch (which would require inordinately large buffers), optical networks tend to be used with connection-oriented call admission and congestion control schemes that limit the burstiness of traffic entering points of the network, and so reduce the delays incurred by buffering.

Optical fibers exhibit low loss for wavelengths in the range of 1280–1620 nm. This broad range of wavelengths creates a bandwidth of ~30 THz, so with proper modulation, each fiber has the capacity to carry information at rates of terabits per second. At present, line-terminating equipment (e.g., lasers and receivers) can only operate at Gbps (gigabits per second) rates, so the capacity of the fiber can be exploited only by multiplexing multiple signals on a fiber, such as by using wavelength-division multiplexing (WDM). Even so, the transmission rate of optical networks is vastly higher than that of electronic or radio systems, and optical networking products are currently even available for the consumer market with transmission rates of 1 Gbps. The high transmission speed of optical networks shifts the performance bottleneck from transmission links to processing and buffering systems. The high transmission speed also reduces the effect of the transmission time on the end-to-end delay, and, instead, the end-to-end delay becomes dominated by propagation delays that are essentially fixed by the speed of light [3]. The only way to reduce the delay for such transmissions is to reduce the distance that signals must propagate, for example, by reducing the number of round-trip times involved in signaling, or by retrieving information from caches that are near the destination.

Optical transmission systems exhibit extremely low bit error rates compared to electrical (e.g., coaxial cable) or wireless transmission systems. For example, optical transmission systems often exhibit bit error rates of the order of 1×10^{-11} [4], and have been reported with error rates as low as 1×10^{-15}. While the line error rates may be low, random transmission errors may still occur on buses internal to routers and switches. In optical networks, there may also be burst loss or errors caused by buffer overflow, or by fixed-duration events (e.g., protection switching) affecting large numbers of bits. Thus, while optical transmission error rates may be low, transport protocols still need mechanisms to ensure reliable transfer.

The security of optical fiber systems is often considered to be stronger than that of other transmission media because the fiber confines the propagation of the optical signal. While it is still possible to tap fiber in an almost undetectable manner, this tends to be more difficult than intercepting electrical or wireless communications.

Compared to electronic networks, optical networks shift the emphasis from packet switching toward circuit switching. This is often done because either optical

processing within the network is difficult (e.g., WDM for all-optical networks), to reduce buffering within the network, or in order to simplify electronic processing to match it to the rate of optical links [e.g., label swapping technologies such as multiprotocol label swapping (MPLS) and asynchronous transfer mode (ATM)]. In the case of WDM, the circuits may be real lightpaths, while in the cases of MPLS/ATM, the circuits tend to be "virtual," in that they may define a path from source to destination, but may not reserve resources for the exclusive use of traffic that uses the "virtual circuit." Circuits can isolate the traffic of one source from that of others, making transport-layer congestion control redundant. With some optical networks [e.g., those based on WDM or the synchronous digital hierarchy (SDH)], the granularity of bandwidth that can be assigned to a virtual circuit may be coarse (e.g., in units of 51 Mbps for SDH), and this elevates the importance of multiplexing by the transport protocol. Circuits also preserve the sequence of information that they carry, simplifying the transport protocol function of error control. (On the other hand, in some optical networks, the difficulty of buffering in the optical domain leads to switches "deflecting" incoming information that is destined for an outgoing port that is busy. Such deflection routing [5] can lead to significant missequencing.)

When an optical lightpath extends end-to-end between communicating terminals, the optoelectronic conversion is physically collocated near the implementation of the transport protocol. This specialized hardware can include hardware support for transport protocols for optical networks, including calculation of integrity checks.

3. TRANSPORT PROTOCOL FUNCTIONS

The introduction to this article stated that transport protocols "match the characteristics and requirements of the network to those of the users of the communication network in the end systems . . . [and] may also be used to match users to each other." This section describes, in depth, the functions that a transport protocol may implement to provide this matching, and how they are affected by the presence of optical networks. The functions covered are as follows: reliable transfer (Section 3.1); flow and congestion control (Section 3.2); security (Section 3.3); framing, segmentation, and reassembly (Section 3.4); multiplexing (Section 3.5); and state management (Section 3.6).

The scope of this section (and, indeed, this article) is limited to *unicast* transfers from a single source to a single destination. Multicast is another important mode of communication; however, multicast transport protocols that provide functions such as reliability are still in their infancy, and the impact of optical networks on these transport protocols is still uncertain. Furthermore, while this section considers the impact of traffic that has stringent timing requirements (e.g., voice and video) on the function of multiplexing, it does not address how transport protocols (such as the Real-time Time Protocol [6]) may provide functions that reconstruct the timing of such traffic, since this function is independent of the existence of optical networks.

3.1. Reliable Transfer

Many applications seek assurance of the integrity of the information that they exchange. Transport protocols provide this assurance by adding well-defined redundancy (e.g., cyclic redundancy codes or checksums) to the payload information that they send, and destinations check that the information that they receive preserves this redundancy, suggesting that it has retained its integrity. If a discrepancy is found, then errors may be corrected either using the redundancy (forward error correction) or the source retransmitting the information. To limit the volume of information that a source must retransmit, and so improve transmission efficiency, it is common for the source to segment the information that it transmits into smaller parts (segments) that are often carried in separate packets that flow through the network.

There are six aspects to the reliable transfer of these segments, which can be understood in analogy to transferring the sections (segments) of this article from the author to the reader:

1. *Integrity*. Received segments should have the same content as transmitted segments; for instance, no typographic errors should be introduced.
2. *Uniqueness*. Each segment sent should be received only once; for instance, sections of the article should not be duplicated.
3. *Completeness*. Each segment sent should be received; for instance, no sections should be missing from the midst or the end of the article.
4. *Sequence*. Each segment should be received in the proper position relative to other segments.
5. *Relevance*. No extraneous segments (e.g., from other sources) should be inserted in the midst of the segments sent by the source.
6. *Delivery*. In addition to these five aspects of reliability that may interest the recipient, there is another aspect that may interest the source. The source may be interested in whether the destination successfully received all segments.

Dividing the broad concept of "reliability" into six aspects is important because different applications are concerned with different aspects of reliability, and these concerns translate into the functionality that these applications seek of the transport layer. For example, real-time media such as voice and video tend to be less concerned with integrity and completeness than they are concerned with sequence. Transaction systems are often unconcerned with delivery acknowledgments, because the client (who is the source of the request, and destination of the response) will issue another request if a response is not received.

Optical networks have differing effects on how transport protocols address the six aspects of reliability. The high *integrity* of optical networks allows transport protocols to use larger segments than for other networks, while still retaining the same transmission efficiency. Larger segments are important because they help reduce the rate at which segments (and packets) must be processed, helping redress the difference in processing and

transmission rates introduced by optical networks. The integrity of optical transmission systems also suggests that transport protocol integrity checks should emphasize detection of errors introduced in switches and routers, rather than those introduced on the transmission line. The performance of integrity checks is highly sensitive to implementation techniques that are discussed in Section 5 of this article.

While optical networks generally preserve packet *sequence*, transport protocols still tend to use sequence numbers to identify segments that need to be retransmitted. The preservation of sequence also helps destinations detect in *complete* delivery. Consecutive received segments that do not have consecutive sequence numbers suggest that either intermediate segments have been lost, or the latter segment is a retransmission — a case that can be identified if retransmitted segments are explicitly identified as such in their headers. This allows the retransmission process to be initiated by the destination sending a negative acknowledgment to the source when it detects missing information, rather than the source having to time out while waiting for an acknowledgment from the destination. This allows prompt retransmission, rather than waiting for a loosely calibrated timeout, and eliminates the need for complicated timers based on round-trip time measurements [7]. Such negative acknowledgments are used by several transport protocols designed for high-speed optical networks, including XTP [8], NETBLT [9], VMTP [10], and SSCOP [11].

Optical networks can help the transport protocol to estimate the round-trip time, if positive acknowledgments *are* used, or if this is needed for congestion control purposes. This is because optical networks have limited buffering and may provide rate guarantees, thus reducing the variation in round-trip times, and allowing timeouts to be calibrated more accurately. The small buffers also reduce the maximum packet lifetime in the network, which reduces the period for which the transport protocol needs to preserve state after the active data transfer has ended in order to ensure that only *relevant* segments are delivered. In TCP, this period is called the "time wait" period, and reducing it reduces the volume of state information that an end system must maintain, which is important for busy servers.

The high bandwidth of optical transmission systems can lead to a large volume of information being "in flight" between source and destination at any time. Transport protocols must choose between using selective retransmissions and a "go-back-n" policy in which the source retransmits everything from the damaged segment onward, including segments that were originally received intact. If errors are truly rare, then the transmission efficiency of go-back-n may be tolerable; otherwise the transport protocol should use selective retransmissions, which require a resequencing buffer in the destination whose size increases with the transmission rate. Like the resequencing buffer at the destination, sources need to maintain a retransmission buffer that may be large for optical networks since its size increases with the volume of information "in flight." Finally, transport protocols for optical networks need to have large-sequence-number fields so that they can uniquely identify each segment (or byte) of information that is in flight from source to destination. These large sequence number fields are used for both reliable transfer and for flow control.

3.2. Flow and Congestion Control

Flow control ensures that a source does not transmit at either a sustained rate or in bursts that exceed the capacity of the destination to process incoming traffic. Flow control works by the source and destination agreeing on a constraint that will govern the source's transmissions, and the destination sending feedback to update this constraint. Traditional transport protocols have constrained the *volume* of information sent; the protocol allows the source to transmit a certain number of bytes or segments before it must wait to receive feedback from the destination, allowing the source to transmit more. The high bandwidth of optical networks, combined with even modest transmission delays, can give the end-to-end path a large "bandwidth-delay product," meaning that the source must have a large volume of information outstanding at any time in order to transmit at the rate of the network. Unless the source spreads the transmission of this large volume of information over some interval, it will transmit very bursty traffic that is likely to lead to congestion in network elements.

One way to address avoid bursts is to use the self-clocking "slow start" [12] technique, in which the source transmits a small window of information, and then expands this window (up to some limit) with each acknowledgment that it receives from the destination. While this technique avoids an initial burst, it prevents the source from transmitting at the full rate until some interval after the initial transmission. This significantly affects sources that have small amounts of information to transmit, since they may always have their transmission rate limited by this slow-start regime.

An alternative approach to flow control that is popular amongst transport protocols for optical networks [e.g., 8–10] is to use "rate"-based control, where the constraint on source transmissions is defined in terms of a transmission rate rather than a window. Often the "rate" is specified in terms of bits per second, or a mean segment interarrival time, *in addition to* a burst size, which limits the volume of information that the source can send at any instant.

Rate-based controls are complicated by the fact that few current end systems employ real-time operating systems that can guarantee to a process the resources needed to sustain consumption of traffic at a certain rate, independent of competing demands from other processes. Furthermore, few processes can predict how much processing incoming information will require. Thus, while a destination process or transport protocol may propose a rate that it can accept, it will often monitor the level to which its buffers are filled in order to decide whether to adjust the rate up or down. Rate-based controls can also be complicated by operating systems that provide timers that have coarse resolution, since maintaining a high transmission rate with such timers requires that

large bursts also be possible. On the positive side, rate-based controls are becoming increasingly common at the interface between the transport and network layer of modern optical networks. Here, rate-based controls are used to "condition" (also known as "shape" and "police") traffic entering the network so that it doesn't overwhelm the network; either overflowing the limited buffers in an optical network, or interfere with traffic from other sources that has tight delay constraints. Some transport protocols, such as XTP [8], allow sharing of rate control information between the network and transport layer, so that the rate control by the transport layer can also help satisfy network rate requirements.

While flow control prevents a source from overloading the limited capacity of a *destination* endpoint, congestion control is designed to prevent a source from overloading the *network's* capacity. While the transport layer may know how the application would prefer to respond to congestion (e.g., by slowing down, or discarding, traffic), it is the network layer that is ultimately responsible for ensuring that this response occurs, preventing an end system from sending so much traffic that the network congests, to the detriment of other network users. In the past, it was often convenient to implement congestion control in the transport layer (e.g., congestion control was added to the Internet through a small change to TCP [12]). However, networks cannot rely on transport layer congestion control because transport protocols are implemented in end systems where users, who may prefer to place their interests before those of the network, can readily change them. Transport layer congestion control may also be inappropriate for optical networks that provide end-to-end circuits, since the congestion control may only unnecessarily delay source transmissions.

3.3. Security

Although fibers give optical signals some physical security, providing logical security requires encryption of the information transmitted over the optical network. According to the end-to-end arguments [13], the transport protocol is a natural location for encryption, since it is implemented in the end systems, where users can control and trust the encryption and keys that are used. However, encryption of data for optical networks is challenging because of the computational complexity of cryptographic functions, when processing is already a bottleneck compared to the transmission system. This encourages hardware implementations of cryptographic functions [14].

Transport protocols for optical networks need to consider the dependency of many encryption systems on processing data in sequence. For example, traditional encryption often uses ciphers in closed-loop cipher feedback or cipher block chaining modes, in which the output from encrypting initial data is fed back into the cipher to influence the encryption of subsequent data. Such serial processing restricts the possibility of parallelism in the encryption process, which could raise the throughput to match optical transmission rates. Consequently, optical networks may favor ciphers used in an output feedback mode, in which multiple cryptographic systems operate in parallel to generate streams of bits that are unpredictable without the correct key, and the payload is simply (and rapidly) exclusive-ORed with these bitstreams to protect it prior to transmission.

While a transport protocol may be expected to enhance security through encryption of the payload information, it should not itself introduce security vulnerabilities of its own. For example, transport protocols introduce state information, which can lead to denial of service attacks (whereby an attacker attempts to exhaust all state storage space available on a server, preventing service to genuine clients). Such attacks can be countered by the use of cryptographic "cookies" [15]. Similarly, while optical networks may promote the use of negative acknowledgments for error control, these should include information that allows the source to authenticate that they came from the destination. Otherwise, a simple denial of service attack can be effected by a third party repeatedly sending negative acknowledgments to the source, forcing it to retransmit information rather than transmitting new information.

3.4. Framing, Segmentation, and Reassembly

Packet-switched networks impose limitations on the maximum length of each packet for reasons such as controlling the delay that one packet may experience when serialized behind another packet in a multiplexer. A transport protocol for an optical network may segment information from the application prior to transmission to satisfy these maximum transmission unit requirements of the network, and later reassemble it at the destination. It may also use segmentation and reassembly to limit the size of information that needs to be retransmitted when recovering from a bit error, and this is done for both packet- and circuit-switched networks. Whenever segmentation is performed, it may be important for the optical network to be aware of where transport layer segments begin and end so that it can discard a few whole segments, rather than many segment parts, using techniques to be described in Section 4.2.

While the network may provide a conduit that carries information from source to destination, the data units that applications send across this conduit are often discrete, and should not be merged. Several transport protocols provide framing of application data units, so that the destination can determine where, among the information received, application data units start and end. Some transport protocols (notably TCP) provide a bytestream only between communicating applications, and applications can indicate where their data units begin and end only by creating separate connections for each data unit.

3.5. Multiplexing

While the network delivers information to the appropriate *network interface*, often identifying a physical entity such as a computer, the transport layer is responsible for delivering this information to the appropriate *process* operating in the device that has that network interface. Thus, transport protocols provide multiplexing at sources to allow multiple processes to share a single network

interface, and demultiplexing at destinations to direct incoming information to the appropriate interface.

Multiplexing tends to be implemented by including in transport layer segments fields that indicate the destination and source ports of the segment, although transport layers may also use the multiplexing identifiers that are used by connection-oriented network layers (such as ATM) in order to avoid the harmful effects of layered multiplexing [16]. Since some optical networks offer circuits only with coarse granularity of bandwidth, multiplexing is particularly important to allow multiple processes on the end systems to share the large end-to-end bandwidth. When multiplexing together traffic from multiple applications, the traffic from one application (e.g., a delay-insensitive file transfer) may interfere with the delay experienced by other applications (e.g., a delay-sensitive telephone conversation). Some more recent transport protocols designed to support multimedia [e.g., 17] include schedulers that are designed to avoid such interference.

3.6. State Management

Unlike the functions discussed in preceding sections that directly satisfy network or application needs, transport protocols use state information to support their implementation of other functions. In particular, state information is important in the provision of reliable transfer, flow, and congestion control. State information can also be used to record negotiated values of parameters that govern the operation of the protocol during the transfer, including segment size or whether end systems should represent sequence numbers in big- or little-endian format [18]. Such negotiation is often relatively simple for transport protocols since for unicast transfer, they involve only two parties that must negotiate. Such negotiation is important for transport protocols for optical networks since end systems that can agree to simplify the communication may be able to employ simpler protocols, and so keep up with the rate of optical transmission systems. An extreme form of negotiation is where transport protocols are composed [19,20] on demand to include minimal functionality that satisfies source, destination, and network requirements. Such composition has the potential to eliminate unnecessary functionality, and so simplify transport protocols so that they can match the transmission rates of optical networks.

The signaling to establish or release state information can be done either *in band*, on the same channel that is used for data transfer, or *out of band*, on a separate channel. Sending signaling information on the same channel that is used for data transfer avoids the potential problem of completing signaling when no data channel is available, but has the disadvantage of complicating the processing that is required on the frequently used data channel with rarely used and complicated signaling functionality. Most transport protocols to date (e.g., TCP) have emphasized in-band signaling, but out-of-band signaling is used in the Advanced Peer-to-Peer Networking protocol [21] and may become more widespread with all-optical networks in which network signaling must be performed electronically, creating an electronic control plane that is separate from the optical data plane. ATM also emphasizes separation of the control and "user" (data) planes.

Servers that serve multitudes of clients (e.g., server of a popular Website) often require the high transmission rate of optical networks. For such servers, the overhead of retaining state information for each client for long periods can become onerous, and this promotes the conveyance of state information through cookies [15]. Note that it is the number of clients that causes this issue, and leads to the server using a high-speed (e.g., optical) network. That is, this issue is *correlated* to the use of optical networks, but is not *caused* by the use optical networks.

4. SPECIFIC TRANSPORT PROTOCOLS FOR OPTICAL NETWORKS

This section describes specific transport protocols for optical networks, namely, TCP, those relating to ATM, and XTP. It emphasizes the match between the functionality of these protocols and that required for optical networks. While the functionality of a protocol is important, the method by which that functionality is implemented also affects the suitability of a transport protocol to optical networks, and will be addressed in Section 5.

4.1. Transmission Control Protocol

George Santayana wrote, "Those who cannot remember the past are condemned to repeat it" and this has led to a networking proverb: "Those who ignore TCP are doomed to reinvent it." The Transmission Control Protocol (TCP) is currently the most widely used transport protocol on existing electronic and optoelectronic networks. While originally introduced in the 1970s [22], TCP has since been extended with new options [e.g., 23] that improve its suitability to high-speed optical networks. This section considers the application of TCP to optical networks. The key features of TCP are its provision of all six aspects of reliability, and the minimal information that it needs from the network layer, which allows it to operate over varied networks, including optical networks.

TCP provides reliable transfer by including in the header of each segment a 32-bit sequence number field, a 32-bit acknowledgment field, and a 16-bit checksum field. During connection establishment, the source and destination agree on the initial sequence number to use for the transfer, and the source sets the sequence number field in each segment that it sends to indicate the position of the first byte of the segment in the sequence number space. For a TCP destination to correctly locate incoming segments, it must ensure that the sequence number of each segment is unique relative to other segments that are also in transit from the source. For a 1-Gbps transmission system, this means that each segment should take no longer than 17 s to propagate through the network [23]. A timestamp option allows TCP to operate across networks with higher maximum segment lifetimes [23].

A TCP destination uses the checksum field to verify the integrity of incoming segments. Whenever the destination receives two maximum-sized segments worth of data (or

500 ms elapses since receiving a segment that it has not yet acknowledged), it sends a cumulative positive acknowledgment to the source to indicate the sequence number of the next segment that it is expecting to receive, that is, that follows any contiguous set of segments that it has received so far. If a segment is lost, then the destination will send duplicate acknowledgments when it receives subsequent segments, acknowledging receipt of the same contiguous set of segments. The source waits a certain period after transmitting a segment for that segment to be acknowledged, and if an acknowledgment is not forthcoming, then it will retransmit the segment. This timeout may occur after a reasonably long time, and with high-speed optical links, the source may stall before this time, being forced by flow or congestion control mechanisms to defer additional transmissions until it receives a new acknowledgment. The "fast retransmit" extension to TCP [24] can circumvent this stalling on high-speed links, by the source retransmitting a segment when it receives multiple duplicate acknowledgments that suggest loss of that segment, rather than waiting for a timeout to occur. Selective acknowledgments have also been added as a TCP option [25] to improve transmission efficiency on high-speed long-delay paths.

The congestion control of TCP [12] is related to its error control through the interpretation of suggestions of segment loss as indicators of congestion. This interpretation is more valid for optical networks than it is for wireless networks, where appreciable loss can also occur as a result of transmission errors. More recently, TCP has been extended to allow processing of explicit congestion notifications from routers [26]. A TCP source will gradually increase its transmission rate, by one segment per round-trip time, while it does not observe signs of congestion (duplicate acknowledgments, retransmission timeouts, or explicit congestion notifications). When a TCP source observes signs of congestion, it quickly halves its transmission rate. This additive-increase, multiplicative-decrease behavior allows a TCP source to adapt to changing network conditions, and helps multiple sources converge on a "fair" allocation of bandwidth. The fact that a TCP source discovers the permissible transmission rate by itself makes this congestion scheme remarkably general, being able to operate over wireless and optical packet-switched and circuit-switched networks, as well as being flexible and thus able to adapt to changing network conditions. However, for optical networks, a TCP source can take some time to ramp up its transmission rate to that allowed by the network, even when using the slow-start function. Furthermore, a TCP source will continue to probe for additional bandwidth, increasing its transmission rate until segments are lost (e.g., due to source transmission buffers overflowing) and then back off. This behavior is suboptimal for optical networks in which the capacity of a (virtual) circuit may be fixed — increasing the rate only leads to unnecessary loss, and backing off unnecessarily slows the transmission. While there has been research into improving TCP when the network can provide minimum rate guarantees [27], further research is needed to optimize the performance of TCP on optical networks so that the source transmission rate converges on the path capacity, rather than oscillating around it.

The flow control mechanism of TCP originally limited its performance on high-speed optical networks. TCP's flow control works by the destination setting a field in segments returning to the source to indicate the size of its receive window. The 16-bit size of this field, and the fact that it measured bytes, limited a TCP source to having at most 65536 bytes of information unacknowledged at any time — a volume that is insufficient to fill many high-speed transmission links. This has since been corrected by the addition of a TCP window-scale option [23] that effectively increases the size of the receive window to 32 bits.

While TCP has successfully evolved to match the capacity and requirements of optical networks, it is showing signs of its age through the susceptibility of its state management to denial of service attacks, and lack of support for multihoming and partially ordered delivery. While none of these perceived deficiencies particularly limit its applicability to optical networks, it is likely that TCP will be succeeded in the future by newer protocols such as the Stream Control Transmission Protocol [15].

4.2. The Impact of ATM

The asynchronous transfer mode (ATM) was heralded in the early 1990s as a technology that would replace existing telephony and packet-switched technologies (e.g., Ethernet and Frame Relay) with a single broadband integrated services digital network. It was designed with optical networks in mind, by fixing the packet ("cell") size and including in each cell labels ("virtual channel" and "path" identifiers) that were intended to facilitate high-speed switching that could match the rate of optical transmission systems. ATM affects transport protocols in two ways. First, ATM adaptation layers were designed to adapt the native ATM service into a service that better matches the requirements of communicating applications, essentially creating a set of new transport protocols. The Service Specific Connection Oriented Protocol (SSCOP), described below, is an example of one of these transport protocols. The second effect results from the different framing used by ATM and transport protocols, and led to particular discard techniques within the optical network to accommodate transport protocols.

The impetus for SSCOP's design [11] was high-speed user data transfer, but it was first used to carry signaling in ATM networks. SSCOP exploits the fact that ATM networks preserve sequence (as do many optical networks), allowing the destination to detect loss when it receives a segment with a sequence number that does not follow its predecessor. An SSCOP destination will then immediately send an unsolicited status message (negative acknowledgment) to the source that requests retransmission of the missing segment. SSCOP is simple, allowing it to keep up with optical transmission speeds, by virtue of only requiring one timer at the source to trigger the transmission of periodic poll messages to the destination. SSCOP recovers from segment loss by the destination receiving retransmissions that are triggered either by the destination's immediate unsolicited negative acknowledgment or, if that or the corresponding

retransmission is lost, by negative acknowledgments in the destination's status report in response to receiving a periodic poll message. Other aspects of SSCOP that make it suitable for high-speed optical networks include its 32-bit aligned trailer-oriented protocol fields, and separation of control and information flow.

The size of ATM cells (53 bytes, including ATM overheads) is much smaller than that of conventional transport protocol segments [e.g., 1 KB (kilobyte)]. Furthermore, ATM does not recover from cell loss within the network and, in its native form, is oblivious to transport layer framing. This could lead to poor throughput of transport layers over ATM, as ATM could drop an early cell from one transport layer segment, and then waste resources on transmitting latter cells from that segment, and then discard a cell from the next segment, causing multiple segments to be damaged and need retransmission. This led to "ATM" switches being designed to be aware of transport layer framing. Switches that used the *partial packet discard* scheme [28] would continue discarding cells from a segment that had one of its cells discarded, confining loss to segments that would be retransmitted anyway. The *early packet discard* scheme [29] took this further, not only assuming that the transport layer would retransmit the whole segment, but assuming that the transport layer (like TCP) interpreted loss to indicate congestion, so when switch buffers were filling (but not yet overflowing), this scheme would discard all cells belonging to certain segments, preventing any partial packet delivery and prompting TCP's congestion avoidance, improving the aggregate throughput of TCP over ATM.

4.3. Xpress Transport Protocol

The Xpress Transport Protocol (XTP) [8] was developed in conjunction with the Protocol Engine project, with the aim of being a high-speed transport protocol that was suitable for VLSI implementation. The high speed of optical networks often motivates hardware implementations of protocols, and the next section of this article considers this topic in more depth.

Like SSCOP, XTP provides a "fastnak" mode, in which the destination immediately sends a negative acknowledgment in response to receiving a segment with a sequence number that does not follow its predecessor. This mode of operation is well matched to optical networks that tend to preserve sequence. XTP also provides the user with a choice of whether to use go-back-n or selective retransmission. The flow control of XTP contains both rate-based and window-based components.

5. IMPLEMENTATION ISSUES

Previous sections have addressed end-to-end protocol issues, whereas this section addresses how the transport protocol should be implemented in end systems. Proper implementation is important to ensure that transport protocols can handle the high transmission rates of optical networks.

The high transmission rates of optical networks encourage the use of large data units (packets and segments) in transport, and other protocols. Increasing the data unit size reduces the frequency with which per-data-unit operations (such as classification for demultiplexing and state lookup) need to be performed, and only impacts transmission efficiency if the application's payload is smaller than the transmission data unit. Consequently, optical networks often use packets that may be $\geq$8 KB "jumbograms." It is also desirable to avoid interrupting the destination processor at high speed as each data unit arrives because of the overhead in context switching. The interrupt frequency can be reduced by using *interrupt coalescing* [30], in which the processor is not interrupted for every packet received, but rather is interrupted for every n packets received (or after a short timeout).

The processing of packet headers can also be expedited by the format of the header reflecting processing requirements. For example, fields should be aligned on processor word boundaries whenever possible so that the processor does not need to shift them before operating on them. The packet header should also ideally fit within a processor cache line, and be aligned in memory with cache lines, so as to expedite header processing. Integer values in header fields may be represented in either little- or big-endian format. Ideally, the representation should match the native representation of the processors used in the end systems, and end systems may negotiate this representation as part of connection establishment [18]. The common alternative is to use a representation that is statically defined by the transport protocol, and for processors to adapt though translation. Since the headers of consecutive segments often differ little (e.g., perhaps in only a sequence number and checksum), they can be rapidly prepared at the source by copying the header from a template and then adjusting the fields that differ for this segment. Similarly, the destination can rapidly check the header of each incoming segment by comparing it to the header of the preceding segment [31].

Memory technologies have not increased in bandwidth as fast as optical transmission technologies, so high-speed transport protocols need to be implemented with appropriate memory management in order to match the rates of optical networks. This means minimizing the number of times that payload information is shifted in memory, such as using "buffer cutthrough" [32], in which data units remain stationary in memory when they are passed between layers of a protocol stack, with layers exchanging pointers to these data units, and encapsulating (or decapsulating) them in situ in memory.

Caching, as used to match memory speeds to increasing processor speeds, is of limited use for network interfaces, since they exhibit little temporal locality of reference. However, data sent over a network do exhibit spatial locality, and this makes video RAM technology [33] attractive for interfacing end systems to optical networks. Video RAMs consist of a large array of dynamic memory cells, and a fast static memory that can be rapidly stored or loaded with a row of the dynamic memory cells, or sequentially stored or loaded with values from off the chip. The conventional application of video RAMs is for

the static memory to serially feed a videodisplay raster scan, while a processor can concurrently modify values in the dynamic memory. In the networking application [e.g., 8], the static memory feeds (or is fed by) the high-speed network interface, and the application accesses the dynamic memory.

Proper partitioning of the implementation of a transport protocol between hardware and software, and between the user-space and kernel-space software components can help a transport protocol keep up with optical networks. Functions that "touch" each byte that is transmitted (e.g., checksum calculations and encryption) are particularly appropriate for hardware implementation because of the high processor overhead of implementing these functions in software. If these functions must be implemented in software (e.g., so that they are readily changed), then the memory overhead of accessing each byte can be reduced by integrating these functions with functions from other layers of the protocol stack that also touch each byte [34]. The Protocol Engine project [8] attempted to implement complete protocol stacks, including the Xpress Transport Protocol, in hardware.

In addition to the partitioning of an implementation between hardware and software, there is also the issue of partitioning the software components of an implementation between user-space and kernel-space parts of the memory system. A transport protocol implementation needs to access services that an operating system provides, such as buffer management, scheduling, input/output access. Kernel-based transport protocols may reduces the overhead in accessing these services, increasing performance, but they are difficult to develop and deploy. Careful implementation can produce high-performance user-space implementations of transport protocols [e.g., 35].

BIOGRAPHY

Tim Moors is a Senior Lecturer in the School of Electrical Engineering and Telecommunications at the University of New South Wales, in Sydney, Australia. He researches transport protocols for wireless and optical networks, wireless LAN MAC protocols that support bursty voicestreams, communication system modularity, and fundamental principles of networking. Previously, he was with Polytechnic University in Brooklyn, New York, and the Communications Division of the Australian Defence Science and Technology Organisation. He received his Ph.D. and B.Eng.(Hons.) degrees from universities in Western Australia (Curtin and UWA).

BIBLIOGRAPHY

1. W. Doeringer et al., A survey of light-weight transport protocols for high-speed networks, *IEEE Trans. Commun.* **38**(11): 2025–2039 (Nov. 1990) (provides a classic survey of transport protocols for high-speed networks).
2. International Telecommunications Union-T, *Framework of Optical Transport Network Recommendations*, ITU-T, Recommendation G.871, Oct. 2000.
3. L. Kleinrock, The latency/bandwidth tradeoff in gigabit networks, *IEEE Commun. Mag.* **30**(4): 36–40 (April 1992).
4. A. Jalali, P. E. Fleischer, W. E. Stephens, and P. C. Huang, 622 Mbps SONET-like digital coding, multiplexing, and transmission of advanced television signals on single mode optical fiber, *Proc. ICC*, April 1990, pp. 1043–1048.
5. N. F. Maxemchuk, Problems arising from deflection routing: Live-lock, lockout, congestion and message reassembly, *Proc. NATO Advanced Research Workshop on Architecture and Performance Issues of High-Capacity Local and Metropolitan Area Networks*, June 1990, pp. 209–333.
6. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, IETF, RFC 1889, Jan. 1996.
7. L. Zhang, Why TCP timers don't work well, *Proc. SIGCOMM*, Aug. 1986, pp. 397–405.
8. G. Chesson, XTP/PE overview, *Proc. 13th Conf. Local Computer Networks*, Oct. 1988, pp. 292–296.
9. D. D. Clark, M. L. Lambert, and L. Zhang, NETBLT: A high throughput transport protocol, *Proc. SIGCOMM*, Aug. 1987, pp. 353–359.
10. D. R. Cheriton and C. Williamson, VMTP as the transport layer for high-performance distributed systems, *IEEE Commun. Mag.* **27**(6): 37–44 (June 1989).
11. T. R. Henderson, Design principles and performance analysis of SSCOP: A new ATM Adaptation Layer protocol, *Comput. Commun. Rev.* **25**(2): 47–59 (April 1995).
12. V. Jacobson, Congestion avoidance and control, *Proc. SIGCOMM '88*, Aug. 1988, pp. 314–329.
13. J. H. Saltzer, D. P. Reed, and D. D. Clark, End-to-end arguments in system design, *Proc. 2nd Int. Conf. Distributed Computing Systems*, April 1981, pp. 509–512.
14. J. D. Touch, Performance analysis of MD5, *Proc. SIGCOMM*, Aug. 1995, pp. 77–86.
15. R. Stewart and C. Metz, SCTP: A new transport protocol for TCP/IP, *IEEE Internet Comput.* (Nov. 2001).
16. D. L. Tennenhouse, Layered multiplexing considered harmful, *Proc. Protocols for High Speed Networks*, Nov. 1989, pp. 143–148.
17. A. T. Campbell, G. Coulson, F. Garcia, and D. Hutchison, Resource management in multimedia communications stacks, *Proc. IEE Int. Conf. Telecommunications*, April 1993.
18. S. Boecking, *Object-Oriented Network Protocols*, Addison-Wesley, Reading, MA, 2000.
19. S. Murphy and A. Shankar, Service specification and protocol construction for the transport layer, *Proc. SIGCOMM*, Aug. 1988, pp. 88–97.
20. B. Stiller, PROCOM: A manager for an efficient transport system, *Proc. IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems* (*HPCS '92*), Feb. 1992, pp. 0_14–0_17.
21. A. Baratz et al., SNA networks of small systems, *IEEE J. Select. Areas Commun.* **3**(3): 416–426 (May 1985).
22. V. Cerf, Y. Dalal, and C. Sunshine, *Specification of Internet Transmission Control Program*, IETF, RFC 675, 1974.
23. V. Jacobson, R. Braden, and D. Borman, *TCP extensions for High Performance*, IETF, RFC 1323, May 1992.
24. M. Allman, V. Paxson, and W. Stevens, *TCP Congestion Control*, IETF, RFC 2581, April 1999.

25. M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, *TCP Selective Acknowledgement Options*, IETF, RFC 2018, Oct. 1996.

26. K. Ramakrishnan, S. Floyd, and D. Black, *The Addition of Explicit Congestion Notification* (*ECN*) *to IP*, IETF, RFC 3168, Sept. 2001.

27. W.-C. Feng, D. D. Kandlur, D. Saha, and K. G. Shin, Understanding and improving TCP performance over networks with minimum rate guarantees, *IEEE/ACM Trans. Network* **7**(2): 173–187 (April 1999).

28. G. Armitage and K. Adams, Packet reassembly during cell loss, *IEEE Network* **7**(5): 26–34 (Sept. 1993).

29. A. Romanow and S. Floyd, Dynamics of TCP traffic over ATM networks, *Proc. SIGCOMM '94*, 1994.

30. S. Muir and J. Smith, AsyMOS — an asymmetric multiprocessor operating system, *Proc. OPENARCH*, April 1998, pp. 25–34.

31. D. Clark, V. Jacobson, J. Romkey, and H. Salwen, An analysis of TCP processing overhead, *IEEE Commun. Mag.* **27**(6): 23–29 (June 1989).

32. C. Woodside and J. Montealegre, The effect of buffering strategies on protocol execution performance, *IEEE Trans. Commun.* **37**(6): 545–554 (June 1989).

33. J. Nicoud, Video RAMs: Structure and applications, *Proc. IEEE Micro*, Feb. 1988, pp. 8–27.

34. D. Clark and D. Tennenhouse, Architectural considerations for a new generation of protocols, *Proc. SIGCOMM 90*, Sept. 1990, pp. 200–208.

35. A. Edwards, G. Watson, J. Lumley, and C. Calamvokis, User-space protocols deliver high performance to applications on a low-cost Gbps LAN, *Proc. SIGCOMM*, Sept. 1994, pp. 14–23.

36. R. W. Watson and S. A. Mamrak, Gaining efficiency in transport services by appropriate design and implementation choices, *ACM Trans. Comput. Syst.* **5**(2): 97–120 (May 1987) (Focuses on implementation issues).

37. J. P. G. Sterbenz and J. D. Touch, *High-Speed Networking: A Systematic Approach to High-Bandwidth Low-Latency Communication*, Wiley, New York, 2001 (provides in-depth coverage of high-speed networks).

TRELLIS-CODED MODULATION

STEPHEN G. WILSON
University of Virginia
Charlottesville, Virginia

1. INTRODUCTION

Coding for noisy channels has a half-century history of theory and applications, but until the early 1980s the applications usually presumed a binary-to-binary encoding function whose output was transmitted with a binary, say, phase-shift-keyed (PSK), modulation format. The emphasis was on achieving coding gain over an uncoded system, and it was understood that bandwidth expansion was one price to pay, often a tolerable price as in deep-space missions. In this regime, where bandwidth efficiency is less than 1 bps/Hz, finding strong codes amounts to finding codes with optimal binary Hamming distance properties.

On the other hand, most contemporary systems are pressed to communicate increasing throughput in limited bandwidth. Information theorists knew that substantial coding gain was also possible in the "bandwidth-efficient regime", where one is interested in communicating multiple bits per second per hertz. It was not until the seminal work of G. Ungerboeck [1] that this potential came to fruition, and thereafter the technique quickly penetrated applications. The general term for this channel coding technique is *trellis-coded modulation* (TCM). This article will present a tutorial overview of the main ideas behind TCM, with several examples. References to literature are given throughout for deeper investigation. In particular, Biglieri et al. [2] incorporate a wealth of more detailed information.

Coded modulation refers to the intelligent integration of channel coding and modulation to produce efficient digital transmission in the bandwidth-efficient regime. The essential theme is to map sequences of information bits to sequences of modulator symbols, these symbols drawn from a constellation having high spectral efficiency, in such a way as to maximize a relevant performance criterion. For the classic problem of signaling over the Gaussian channel, the objective becomes to maximize the minimum *Euclidean* distance between valid modulated sequences. The mapping can be block-oriented, called *block-coded modulation*.[1] TCM, on the other hand, is best viewed as a stream mapping, associating input bit (or symbol) streams with sequences of signal points from the modulator constellation.

The ideas generally are traced to the late 1970s and the work of Ungerboeck, although, as with most important innovations, earlier roots are evident (see, e.g., the preface to Ref. 2). The 1982 paper by Ungerboeck [1], following a 1976 Information Theory Symposium talk, prompted a great deal of subsequent research in TCM, and the technique quickly became part of the coding culture, appearing in several important voiceband modem and satellite communication standards and, more recently, wireless standards. It is relatively easy to obtain coding gains of several decibels relative to uncoded signaling for modest encoding/decoding complexity, *without* bandwidth expansion. For example, one of the designs discussed below sends 2 bits per modulator interval using 8-PSK modulation, with a 4-state encoder. The design achieves 3-dB asymptotic energy savings over uncoded QPSK, yet occupies the same bandwidth as the latter.

The basic themes common to TCM are rather simple. To achieve bandwidth efficiency in the first place, we need to communicate multiple bits per signal space dimension, since it can be shown that bandwidth occupancy is related to the number of complex signal space dimensions per unit time. To send k message bits per modulation interval, we adopt a constellation $\mathcal{C}$ containing more points than

[1] Block-coded modulation has never attained the foothold enjoyed by TCM in this application, probably because TCM represents a cleaner solution, and maximum-likelihood decoding is readily obtainable in TCM.

2^k. (Typically, as seen shortly, the set size will be twice as large, viz., 2^{k+1}.) This constellation expansion provides the *redundancy* important to coded modulation, as opposed to simply sending extra points per unit time from a constellation of size 2^k. We divide the constellation into regular disjoint subsets, called *cosets*, such that the minimum Euclidean distance between members of these cosets is maximized (and greater than the original minimum distance of the constellation). k input bits are presented per unit time, and of these, $\tilde{k} \leq k$ bits are input to a finite-state encoder, having S states. This encoder produces a label sequence that picks a subset (or coset). The remaining $k - \tilde{k}$ bits select the specific constellation point in the chosen subset. Signal points are produced at the same rate as input vectors are presented. This selection process has *memory* induced by the finite-state encoder, representing the other crucial aspect of coded communication. Figure 1 illustrates this generic encoding operation, sometimes called *coset coding* [3]. One might view this process as that of a time-varying modulation process, whereby the subset selector defines a signal set at each time interval from which the actual transmitted signal is to be selected. Of course, there are important details defining any particular design; in particular it is important to choose the proper value of $\tilde{k}$ as well as the proper finite-state encoder.

In the next section, relevant information theory is summarized that points to the potential of TCM as well as proper design choices. Section 3 provides basic design principles with examples for simple TCM codes. Following that, more detailed information on performance evaluation and tables of best codes are provided. The article closes with a brief discussion of related topics, including design of codes for rotational invariance, multidimensional transmission, and fading channels.

2. RELEVANT INFORMATION THEORY

Before delving into the specifics of TCM, it is helpful to examine the potential efficiency gain that exists, according to information theory, for bandwidth-efficient communication. In addition, this study will provide insight into the appropriate choice of constellation size. This will be addressed for the important case of two-dimensional (in-phase/quadrature) modulation.

Consider the two-dimensional Gaussian noise channel shown in Fig. 2, where it is assumed that a signal vector $\mathbf{x} = (x_1, x_2)$ is presented at each channel use. This pair of

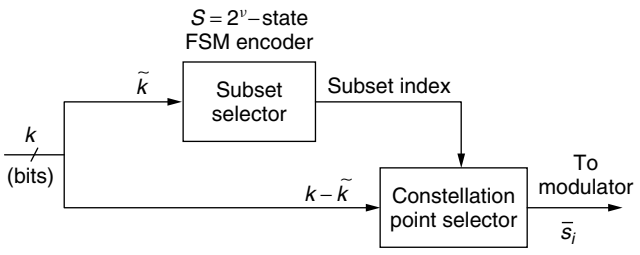


Figure 1. Depiction of generic TCM system.

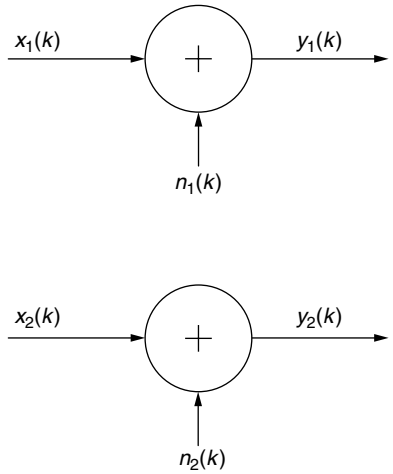


Figure 2. Two-dimensional gaussian channel model.

real numbers corresponds also to a single complex input. The input is constrained only by its expected energy (i.e., $\mathcal{E}[x_1^2 + x_2^2] \leq E_s$), so that E_s represents the average symbol energy at the demodulator input. The noisy channel adds two-dimensional zero-mean Gaussian noise, with independent noise components in each coordinate. In keeping with standard notation, the variance of each component of the additive noise will be denoted $N_0/2$, where $N_0/2$ represents the power spectral density in watts per hertz of the physical white-noise process in the receiver.[2]

The channel capacity of such a channel is defined as the maximum of the mutual information between input and output vectors, the maximum taken over all probability assignments on input symbols satisfying the energy constraint. It is a standard result of information theory [4] that the capacity is attained when the input variables are independent Gaussian, with each subchannel allocated half the available energy. Moreover, the subchannel capacity is

$$C' = \frac{1}{2} \log_2 \left(\frac{1 + E'}{N_0/2} \right) \text{ bits per channel use} \qquad (1)$$

where E' is the energy available to each subchannel. Since the capacity of parallel independent channels equals the sum of the respective subchannel capacities and $E_s = 2E'$, we have

$$C = \log_2 \left(\frac{1 + E_s}{N_0} \right) \text{ bits per channel use} \qquad (2)$$

Figure 3 plots this relation versus E_s/N_0, scaled in dB (see the leftmost curve). Note that the plot, consistent with (2), is linear at high SNR, and every 3 dB increase in SNR buys an additional bit of channel capacity. On this same plot, we mark with circles the values of E_s/N_0

[2] This diagram encapsulates the actual physical process of waveform modulation, physical channel effects, and demodulation of waveforms to real numbers.

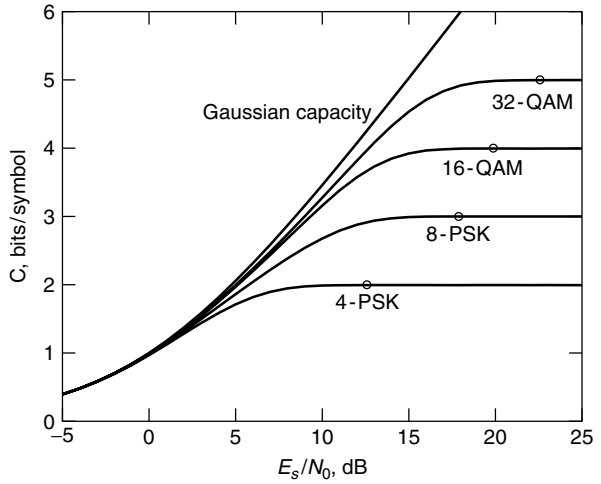


Figure 3. Capacity of two-dimensional signaling, AWGN channel.

needed to achieve bit error probability 10^{-5} when using *uncoded* 4-PSK, 8-PSK, 16-QAM, and 32-QAM (cross) constellations, techniques that are efficient designs for sending 1, 2, 3, or 4 bits per symbol, respectively. (See any text on digital communication, e.g., Ref. 5 or 6.) Noting that the capacity constraint implies the theoretical minimum E_s/N_0 capable of sending k bits per symbol, but that this limit is in principle closely approachable, we conclude that roughly 8–10 dB of potential efficiency improvement exists all across the high-rate regime, indicating significant potential exists for bandwidth-efficient signaling, as well as in the classic coded binary signaling realm. This result was evident prior to the advent of TCM, but it is somewhat surprising that it took researchers so long to tap this potential.

To translate this into implications for bandwidth-efficient signaling and for bottom-line energy efficiency limits, we argue that information rate R bits per channel use can be communicated reliably, provided $R < C$. When communicating R bits per symbol, we have $E_s = RE_b$, where E_b is the energy associated with an information bit, and obtain the relation

$$R \leq \log_2\left(\frac{1 + RE_b}{N_0}\right) \tag{3}$$

which thus requires that

$$\frac{E_b}{N_0} > \frac{2^R - 1}{R} \tag{4}$$

(This critical value is necessary for reliable communication, but is approachable by sufficiently sophisticated coding methods.) For example, to send $R = 1$ bit per 2D (two-dimensional) symbol, no matter what the constellation and no matter how complicated the coding process, it is impossible to achieve *reliable* communication if the bit SNR, E_b/N_0, is less than 1, or 0 dB. More pertinent to our interest, if we wish to send $R = 4$ bits per interval, we must provide at least $E_b/N_0 > \frac{15}{4}$, or 5.7 dB. In terms of corresponding spectral efficiency, we can (optimistically) estimate that the required bandwidth needed to send R_s modulator symbols per second without intersymbol interference is $B = R_s$, approachable with Nyquist pulseshaping in 2D modulation. Thus, since $R_b = 4R_s$, we can communicate with (nearly) 4 bps/Hz spectral efficiency provided E_b/N_0 exceeds 5.7 dB.

In practice, code sequences are not fashioned from independent Gaussian random variables, although this result does illuminate the design of efficient TCM systems. Rather, the inputs to the channel are chosen from some finite, regular arrangement of M points (the constellation). Letting $P(\mathbf{x_i})$ represent the probability of sending constellation point $\mathbf{x_i}$ into the channel of Fig. 2, we can write that the mutual information between input and output is [4]

$$I(\mathbf{X};\mathbf{Y}) = \sum_{i=0}^{M-1} P(\mathbf{x}_i) \int f(\mathbf{y}|\mathbf{x_i}) \log_2 \times \left[\frac{f(\mathbf{y}|\mathbf{x})}{f(\mathbf{y})}\right] d\mathbf{y} \text{ bits/channel use} \tag{5}$$

This may be evaluated numerically by noting that the conditional PDFs are Gaussian 2D forms as described above. Normally, the input symbols are assumed to be equiprobable, which gives the "symmetric" capacity, although one should realize that choosing high-energy constellation points with smaller probability than inner, low-energy points, is a slightly superior choice mimicking the Gaussian distribution above. So-called *shaping* can extract some of this gain, perhaps amounting to 1 dB [7]. Nonetheless, the symmetric capacity for QPSK, 8-PSK, 16-QAM (quadrature amplitude modulation), and 32-QAM (cross-constellation) are shown in Fig. 3 as well. Notice that the capacity curves saturate, as expected, at $\log_2 M$, where M is the constellation size, since even without additive noise, $\log_2 M$, bits is the maximum error-free information per symbol.

An important conclusion from this analysis, drawn by Ungerboeck, is that to reliably communicate k bits per modulator symbol, constellations of size 2^{k+1} can achieve virtually the same energy efficiency as a hypothetical Gaussian code without constellation constraints would achieve. In specific terms, notice that the 8-PSK curve in Fig. 3 closely follows the Gaussian capacity curve up through capacity of 2 bits per symbol. Thus, to send $k = 2$ bits/symbol, 8-PSK represents a sensible choice. Whereas 16-QAM offers a greater selection of signals to build code sequences, the theoretical potential, expressed by channel capacity limits, is only incrementally better. To send $k = 7$ bits per interval, achieving even greater spectral efficiency, a sensible choice for modulation would be 256-QAM. (Note, however, that not any constellation will suffice; it is important that the constellation be an efficient packing of points into signal space.)

This "constellation doubling" is certainly convenient, for it represents the need for the encoder to produce one extra bit beyond the k input bits. The recommendation holds across the range of spectral efficiency, and is appropriate for constellations of any dimension, including 1D and 4D cases, for example.

3. TCM DESIGN PRINCIPLES

The discussion here will concentrate on 2D TCM, the case of primary practical interest. (A brief treatment of higher-dimensional TCM is given at the end.) The underlying objective is to communicate k bits/modulator interval, thereby achieving a spectral efficiency approaching k bps/Hz. The ideas extend readily to higher dimensions though, and more will be said about the benefits of this later. Also, 1D, that is, pulse amplitude modulation (PAM), designs evolve easily from this basic formulation.

It is first helpful to understand what *not* to do in design. We should avoid the separation of the problem into design of a "good" binary encoder and the design of a "good" modulator for sending binary code symbols. In the case of $k = 2$, this strategy might locate tables of optimal Hamming distance codes, appropriate for binary modulation, with 2 input bits and 3 output bits per time step. Then we could adopt a Gray-labeled 8-PSK modulator for transmission of the 3 code bits. While such a design achieves the objective of sending 2 bits per modulator symbol, this decoupled approach cannot in general attain the same energy efficiency that a more integrated approach follows. (An exception is the case of binary codes mapped onto Gray-coded QPSK; there the squared Euclidean distance between signal points is proportional to Hamming distance between their bit labels, so optimal binary codes produce optimal TCM codes, and the bandwidth doubling normally attached to a rate-$\frac{1}{2}$ binary code, say, is avoided when mapping onto the larger QPSK constellation. In some sense rate-$\frac{1}{2}$-coded QPSK represents the earliest exemplar of TCM.)

3.1. Set Partitioning

Returning to the integrated design methodology, we adopt a constellation $\mathcal{C}$ that is a regular arrangement of $M = 2^m$ points in 2D signal space. We assume this constellation contains more than 2^k points, typically larger by a factor of 2 as discussed above. The constellation is first partitioned into disjoint subsets, $\mathcal{A}_i$, whose union is the original constellation. For constellations we will discuss, this partition tower will involve successive steps of splitting-by-2. Subsets are chosen so that the intraset minimum Euclidean distance (within subsets) is maximized. These subsets are often denoted as cosets, for one subset is merely a translation or rotation of another subset.

At the next level of the partition chain, each subset is further subdivided into smaller subsets, denoted $\mathcal{B}_i$, again increasing the intraset Euclidean distance. In principle the process continues until sets of size 1 are produced, although typically one does not need to proceed to this level.

The process is now illustrated for 8-PSK and 16-QAM constellations.

Example 1. Partitioning of 8-PSK. The 8-PSK constellation is comprised of 8 points equally spaced on a circle with radius $E_s^{1/2}$, so that E_s represents both the peak and average energy per modulator symbol. The partition shown in Fig. 4 first divides the constellation into two

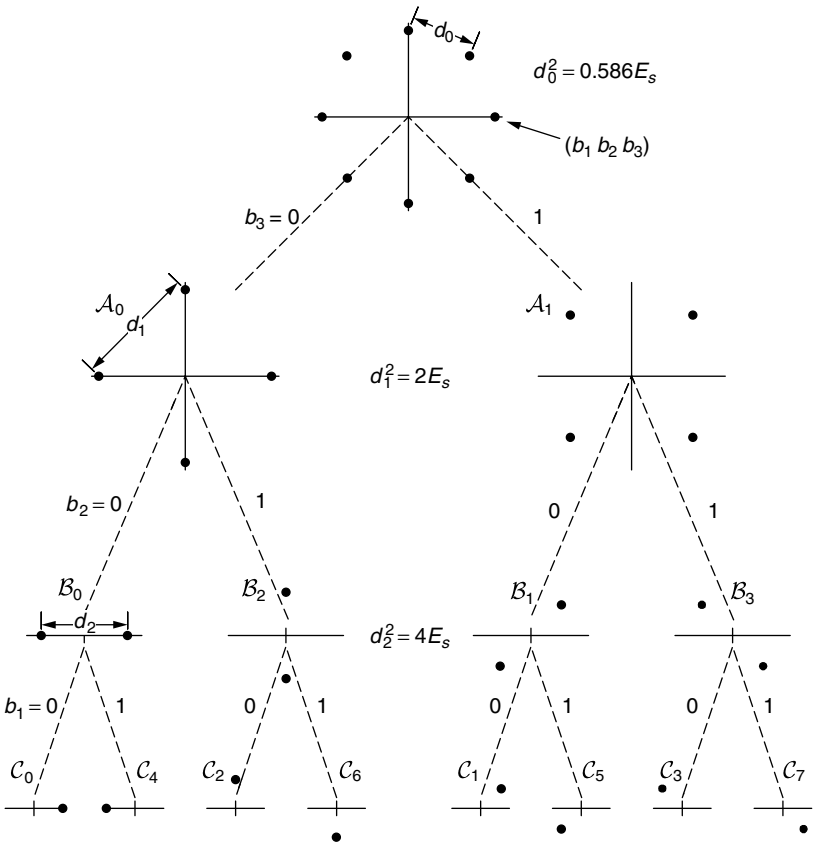


Figure 4. Partition chain for 8-PSK.

QPSK sets, denoted $\mathcal{A}_0, A_1$, one a rotation of the other. Notice that the intraset squared distance is $2E_s$ in normalized terms, whereas the original squared minimum distance is $d_0^2 = 0.586E_s$. Further splitting of each coset produces four antipodal (2-PSK) sets, within which the intraset squared distance is now $4E_s$. We denote these subsets by $\mathcal{B}_0$, $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$. Finally, a third splitting produces singleton sets $\mathcal{C}_i$.

Eventually, each constellation point will need to carry some binary label, 3 bits in this example. The partitioning process above provides a convenient means of doing so, and is called "mapping by set partitioning" by Ungerboeck [1]. In the partition chain, we have arbitrarily attached a 0 bit to a left branch and a 1 bit to a right branch at each stage. Reading the bits from bottom to top gives a 3-bit label to each point, as indicated in Fig. 4. Coincidentally, the bit labeling is the same as that of natural binary progression around the circle.

Example 2. Partitioning of 16-QAM. The partition chain for 16-QAM is depicted in Fig. 5, again showing a twofold splitting of sets at each stage. Here the squared distance growth is more regular than for M-PSK constellations specifically $4a^2$, $8a^2$, $16a^2, \ldots$, where $E_s = 10a^2$ is the average energy per symbol for the constellation. This doubling behavior holds for arbitrary 2D constellations that are subsets of the integer lattice Z^2. Note again the partitioning process supplies bit labels to each constellation point.

Other familiar constellations are easily partitioned in a similar manner, including 1D pulse amplitude modulation (PAM), M-ary phase shift keying, and large M-QAM sets. Identical methods pertain to partitioning of multidimensional lattice-based constellations as well, although the splitting factors are seldom twofold.

3.2. Trellis Construction

Given such a constellation partitioning, it remains to design a trellis code. The parameters of a trellis encoder are (1) the number of bits/symbol, k, as above, and (2) the number of encoder states, S, normally taken to be a power of 2, so $S = 2^v$. A trellis is merely a directed graph with S nodes (states) per timestep, and with 2^k edges joining each of these nodes to nodes at the next time stage. It is these edge label sequences that form the valid sequences of the trellis code. Already there exists some design choice, namely, how these 2^k edges emanating from each state connect with states at the next level. Equivalently, in Fig. 1, of the k input bits, how many will be used to influence the state of the encoder and how many remain to define a point within a subset. Suppose $k = 2$ and $S = 8$, for example. We might construct trellis graphs with each state branching to 4 distinct next states ($\tilde{k} = 2$), or, what is common in TCM designs, we could have two groups of "parallel" edges joining each state with two distinct states at the next level ($\tilde{k} = 1$). We call this "2 sets of 2" branching, versus "4 sets of 1" in the first case. The better policy is unknown at the outset, but the optimal choice becomes

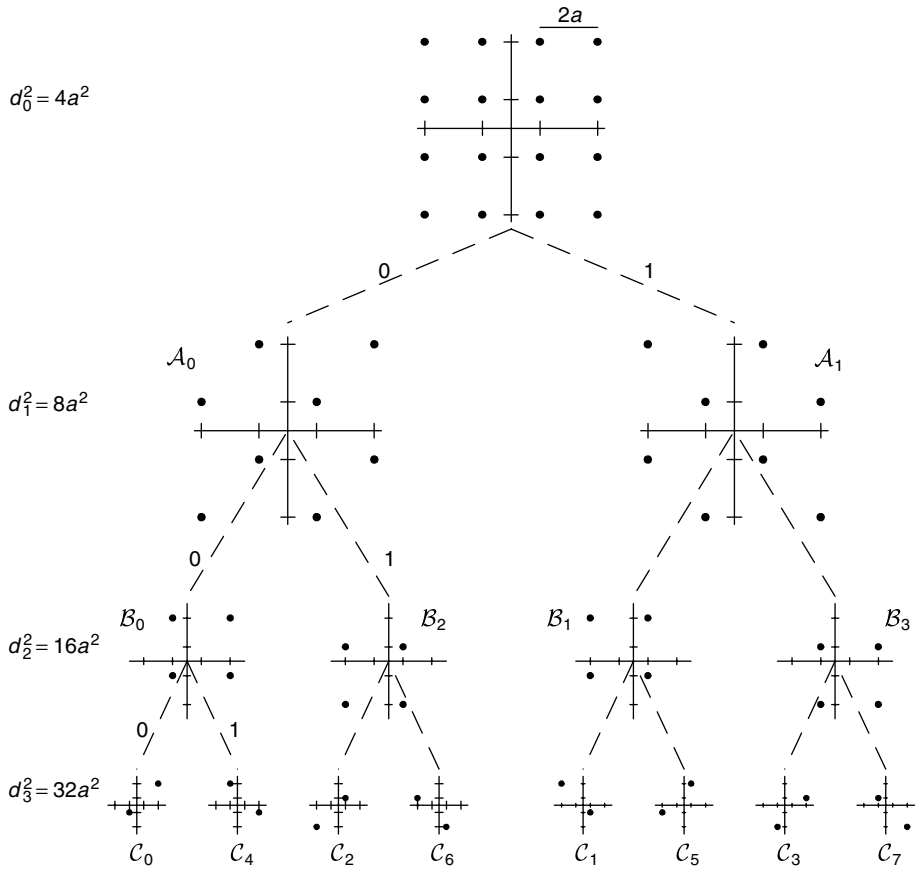


Figure 5. Partition chain for 16-QAM, $E_s = 10a^2$.

evident after study of all cases. In general, the trellis branching will become more diffuse when progressing to larger state complexity.

Given a choice of trellis topology we assign subsets of size $2^{k-\tilde{k}}$ to each state transition. These are commonly called *parallel edges* when $k - \tilde{k} \geq 1$. For small trellises it is not difficult to exhaustively try all possibilities, but for larger trellises there rapidly become many permutations of subset assignments, and their respective performances may differ.

Realizing that maximization of squared distance between distinct channel sequences is the objective, Ungerboeck proposed a "greedy" algorithm as follows:

1. Assign subsets to *diverging* edge sets in the trellis so that the interset minimum distance (between subsets) is maximized. This implies that two sequences that differ in their *state* sequences obtain maximal squared distance on the splitting stage.
2. Assign subsets to *merging* edge sets so that the interset minimum distance is maximized, for similar reason.

Further, all subsets are to be used equally often.

It may not be possible to achieve objectives 1 and 2 in small trellises, 2-state designs, for example. Also, this policy is only a heuristic that seems true of best codes found to date via computer search, but appears to have no provable optimality. Even within the class of "greedy" labelings there will remain many choices in general.

Example 3. 4-State TCM for 8-PSK, $k = 2$. The archetypal example of TCM designs is the 4-state code for 8-PSK, sending 2 bits per symbol. The trellis topology options are 2 sets of 2 ($\tilde{k} = 1$), or 4 sets of 1 ($\tilde{k} = 2$). Adopting the former, along with the greedy policy, and with reference to Fig. 4, we assign subsets of size 2, namely, $\mathcal{B}_i$, to the state transitions as shown in Fig. 6. One should note the symmetry present in the design — each constellation point is used exactly twice throughout the trellis, and each state has its exiting or entering arcs labeled with either $\mathcal{A}_0 = \mathcal{B}_0 \cup \mathcal{B}_2$ or $\mathcal{A}_1 = \mathcal{B}_1 \cup \mathcal{B}_3$. The greedy policy can also be seen — for example, sets $\mathcal{B}_0$ and $\mathcal{B}_2$ have maximal interset distance among the subset choices, and these are assigned to splits and merges at state 00.

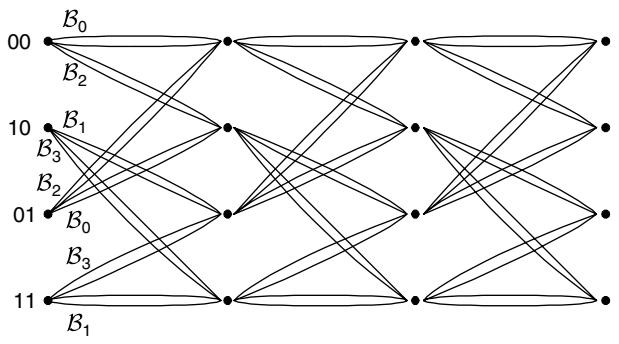


Figure 6. 4-state trellis for $R = 2$, 8-PSK.

It should be noted at this point that information bit sequences are not attached to trellis arcs, but the trellis only specifies valid sequences of modulator symbols. The focus is on maximizing the Euclidean distance between valid sequences, without concern as yet about the underlying message bit sequences.

3.3. Decoding of TCM

Before proceeding further on the design and performance aspects of TCM, we digress to describe optimal decoding, namely, maximum-likelihood sequence decoding.[3] The TCM encoder produces a sequence c_i of 2D signal points for transmission over an additive Gaussian noise channel. Optimal decoding is easily understood in the context of Viterbi's algorithm for decoding the noisy output of any finite-state process [8,9]. The task is to find the most likely message sequence given the noisy observations, and the solution is the sequence whose sum of loglikelihoods for branch symbols is largest. For the Gaussian channel treated here, these branch metrics are merely the negative squared distances between the measurement and the constellation point being evaluated. Further simplification is possible if the constellation points have constant energy, as with M-PSK.

One may implement the decoder in an obvious manner by noting that each survivor state in the trellis such as Fig. 6 has 2^k branches entering it, and the survivor sequence to each state can then be computed by forming the cumulative metric for each of these 2^k sequences, and retaining the largest metric, as well as the route of the best path. In this view, decoding is little different from standard Viterbi processing. An alternate approach that exploits the structure of TCM first views the problem as finding the best-metric choice within each coset (the parallel edges of the graph), then having an add–compare–select routine evaluate the contending coset winners entering a given state. These coset winners can be found once and reused as needed for processing the remaining states at the same time index. Otherwise, aspects of the decoder are identical with standard Viterbi decoding. Issues of metric quantization and range, as well as survivor memory depth are relevant engineering issues, but will not be discussed here.

3.4. Design Assessment

We speak of error events as occurring when the decoder opts for some sequence other than that transmitted. These events generally correspond to short detours from the correct path, during which a small number of message bit errors occur. In contrast with conventional convolutional codes, 1-step error events often exist in TCM decoding. The fundamental parameter of interest in design of TCM codes is the *Euclidean free distance*, d_f, defined as the minimum, over all sequence pairs, of the Euclidean signal space distance between two valid code sequences, without regard to detour length. This is an extension of the notion

[3] This does not exactly minimize the decoded bit error probability, however.

of free distance for convolutional codes, where Hamming distance is the normal distance measure.

The possible error event lengths in the trellis of Fig. 6 are 1-step, along with 3-step, 4-step, and so on. Two-step error events are not possible in this case. For the present, assume that the transmitted sequence follows the top route in the trellis, corresponding to the "all-zeros path". A 1-step error event is the event that the decoder chooses the antipodal mate of the transmitted signal in a one-symbol measurement, and the squared distance for this pair is $4E_s$. (Recall that the radius of the constellation is $E_s^{1/2}$.) This is also the intraset squared distance for $\mathcal{B}_0$.

There are four 3-step error events of the form 2–1–2, 2–1–6, 6–1–2, or 6–1–6, where the numbers represent subscripts of $\mathcal{C}_i$ in Fig. 4. Each sequence has squared distance from the 0–0–0 sequence of $2E_s + 0.586E_s + 2E_s = 4.586E_s$. It is straightforward, though tedious, to demonstrate that all 4-step, and longer, error events have distance at least this large.[4] Thus, the free Euclidean distance of this code is $4E_s$, and we say 1-step error events dominate, since at high SNR, this kind of error event is more likely than any specific 3-step or longer error event. The probability of confusing two signals in Gaussian noise is given by

$$P_2 = Q\left[\left(\frac{d_E^2}{2N_0}\right)^{1/2}\right] \tag{6}$$

where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$ is the Gaussian tail integral function. Given that each transmitted sequence has only one such nearest neighbor, and on each such error event, only 1 of the 2 information bits is in error, we predict that the asymptotic (high SNR) performance of this code is

$$P_b \approx \frac{1}{2} Q\left[\left(\frac{4E_b}{N_0}\right)^{1/2}\right] \tag{7}$$

We have also used the fact that the energy per symbol is $E_s = 2E_b$ (not $3E_b$). In comparison, the bit error probability of uncoded Gray-labeled 4-PSK is

$$P_{b_{\text{QPSK}}} = Q\left[\left(\frac{2E_b}{N_0}\right)^{1/2}\right] \tag{8}$$

The *asymptotic coding gain* (ACG) is defined as the ratio of the arguments of the high-SNR error expressions for the coded and uncoded cases, and ignores multiplier constants. Here the ACG is 2, or 3.01 dB. The spectral occupancy remains unchanged, however, since we are sending one complex symbol for every 2 information bits in either case.

It is interesting to note that the error probability attached to the various message bits is in general not equal. In this example, the message bit defining which of the two parallel transitions is taken from a state to the next state is slightly more error-prone for large SNR. This is because the 1-step error event defines the free Euclidean distance for this code. We thus may find an unequal error protection property, which is sometimes taken as an opportunity if some message bits are deemed more important.

The encoder can be realized in more than one manner, in particular as a feedforward finite-state machine, or as one having feedback.[5] Figure 7 illustrates two "equivalent" realizations of the 4-state encoder for 8-PSK. These encoders produce the same set of coded sequences, although the association between inputs and outputs differs. The 3-step error events attached to the first realization are produced by the lower input bit sequence 0001000…, while the same coded sequence is produced by the input 00010100…. in the second encoder. Corresponding to this, the decoded *bit* error probability will also differ slightly. As with binary convolutional codes [10], we can always realize a TCM encoder as a systematic form with feedback, and this will be the emphasis from here on. Searching over this class of codes is convenient because in some sense this gives a minimal search space, and the encoders are automatically noncatastrophic as well, [1,2,6].

The alternative trellis topology, 4-sets-of-1 branching, cannot be made this efficient. Although 1-step error events are no longer possible, 2-step error events are present with squared distance inferior to the previous design, no matter the subset labeling on the trellis transitions. This latter configuration is, however, the topology associated with choice of an optimal 4-state binary encoder for maximizing Hamming distance — where parallel edges are seldom found.

When moving to an 8-state encoder, note that maintaining 2-sets-of-2 topology retains the 1-step error event, and thus the same free distance. (It is true that the longer error events become less problematic.) To increase the free distance, 1-step error events must be eliminated by switching to 4-sets-of-1 topology. The resulting optimal trellis is shown in Fig. 8, which, the reader may notice,

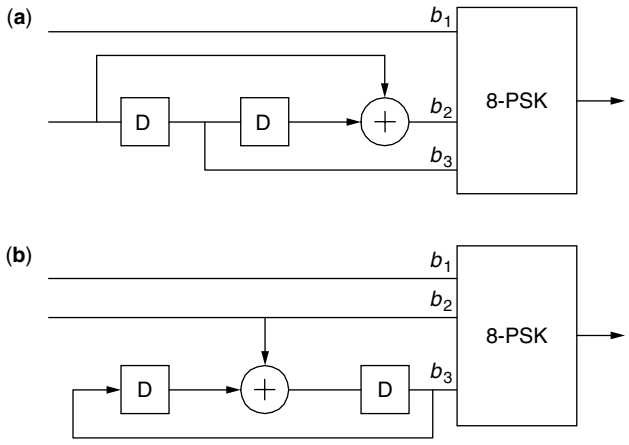


Figure 7. Encoders for 4-state TCM, $R = 2$, 8-PSK: (**a**) feedforward realization; (**b**) systematic, feedback realization.

[4] A computer program can compute the minimum distance in more complicated situations, and would be used to evaluate codes in a code search.

[5] One should also realize that "different" realizations exist when the constellation labeling is changed.

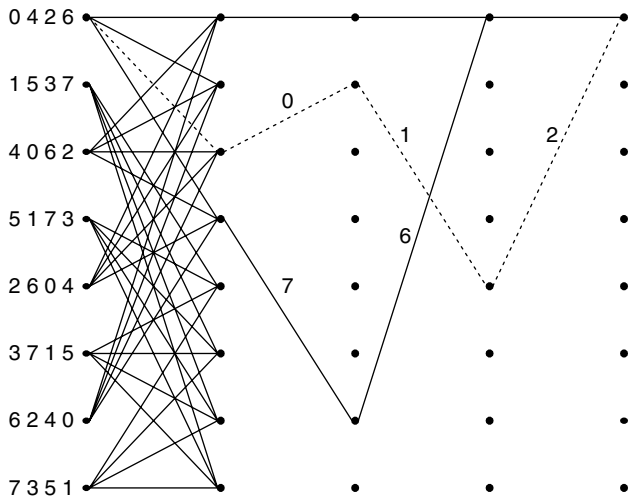


Figure 8. 8-state trellis, $R = 2$, 8-PSK, dominant error events shown relative to upper path.

subscribes still to the greedy policy. The two dominant (3-step and 4-step) error events are shown, both with $d_f^2 = 4.586E_s$, and the asymptotic coding gain over QPSK is now 3.6 dB. Two-step events, although now present, have greater distance.

Example 4. 4-State Code for 16-QAM, $k = 3$. Another example that is feasible to design by hand is the 4-state code for 16-QAM. Each state now has $2^3 = 8$ entering and exiting branches, and the topology options are 2 sets of 4, or four sets of two. The former turns out to be best, again after studying the alternatives, and its trellis labeling is the same as that of Fig. 6, but uses the subset notation of Fig. 5. Since parallel edge sets are now size four, there are three 1-step error events relative to any transmitted sequence, but only two have the smallest squared distance, namely, the intraset minimum distance of $\mathcal{B}_i$ in Fig. 5, $16a^2 = 1.6E_s = 4.8E_b$. Once again, there are no 2-step error events, and 3-step and longer error events have larger squared distance.

Following the methodology of the 8-PSK example, we can determine that for large SNR

$$P_b \approx N_b Q\left[\left(\frac{2.4E_b}{N_0}\right)^{1/2}\right] \tag{9}$$

where $N_b = \frac{1}{3}$ since message labels can be assigned so that at most one of the 3 input bits is incorrect on these two nearest sequences. In general this multiplier constant represents the average frequency of bit errors incurred among all the dominant distance error events, normalized by k, and depends on the code structure as well as the actual encoder realization.

This can be compared with the performance of uncoded 8-PSK, which is bounded by

$$P_b \le \frac{2}{3}Q\left[\left(\frac{0.88E_b}{N_0}\right)^{1/2}\right] \tag{10}$$

The asymptotic coding gain over uncoded 8-PSK (Gray-labeled) is thus $10\log_{10}(2.4/0.88) = 4.3$ dB. Again, this comparison is at equal spectral efficiency (3 bits/symbol); a comparison with uncoded 16-QAM would give slightly larger coding gain.

4. PERFORMANCE

Performance analysis for TCM resorts to upper, and perhaps lower, bounding of events known as the *node error* (or *first error*) *probability* and the *bit error probability*. The latter is generally of most interest, and the former is a useful step along the way. As with analysis of other coding techniques, we normally apply the union bound to obtain an upper bound, and this bound becomes tight at high SNR.

Suppose $\mathbf{c}_i$ is an arbitrary transmitted sequence in a long trellis, and let $\mathbf{c}_j$ be any other valid sequence in the trellis. Define $\mathcal{I}_i$ to be the incorrect subset, namely, the set of valid trellis paths that split from $\mathbf{c}_i$ at specific time k, and remerge at some later time. Then we have that the first-error-event probability is union-bounded by

$$P_e \le \sum_{\mathbf{c}_i} P[\mathbf{c}_i] \sum_{\mathbf{c}_j \in \mathcal{I}_i} P[\Lambda(\mathbf{c}_j) > \Lambda(\mathbf{c}_i)] \tag{11}$$

where $\Lambda(\mathbf{c}_i)$ is the total path metric for the sequence $\mathbf{c}_i$. Note that the bound is a sum of "2-codeword" error probabilities. These 2-codeword probabilities can be written for the AWGN channel as

$$P[\mathbf{c}_i \to \mathbf{c}_j] = Q\left[\left(\frac{d_E^2(\mathbf{c}_i, \mathbf{c}_j)}{2N_0}\right)^{1/2}\right] \tag{12}$$

where $d_E(\mathbf{c}_i, \mathbf{c}_j)$ is the Euclidean distance between the two sequences.

One important distinction for trellis codes is that the usual invariance to transmitted sequence may disappear; that is, the inner sum in (11) depends on the reference sequence $\mathbf{c}_i$, because either the sets of 2-codeword distances in the incorrect subsets vary, or more typically, the multiplicity at each distance varies with reference sequence. For linear binary codes mapped onto 2-PSK or 4-PSK, an invariance property holds that the all-zeros sequence can be taken as the reference path, without loss of generality. A simple counterexample for TCM is provided by coded 16-QAM; transmitted sequences involving inner constellation points have a larger number of nearest-neighbor sequences than do transmitted sequences involving corner points, even though each has the same minimum distance to error sequences. Further details are not included here, but this issue is discussed under the topic of uniformity of the code, and conditions can be found for varying degrees of uniformity [2,11]. It turns out that the 8-PSK codes presented here are strongly uniform; that is, any transmitted sequence can be taken as a reference sequence. Other TCM schemes exhibit a weaker kind of uniformity in which every transmitted sequence has the same nearest-neighbor distance.

Given the general lack of a strong invariance property, the traditional transfer function bounding approach used to calculate (11) for convolutional codes must be generalized to average over sequence pairs. Reference 12 provides a graph-based means of doing this averaging, which provides numerical upper bounds on error event probability, and by differentiating the transfer function expression, bounds on decoded bit error probability (see also Ref. 2). These expressions are series expressions involving increasing effective distance, and as SNR increases, the leading (free distance) term in the expansion dominates.

The resulting expressions will be of the form

$$P_e \approx N_e Q\left[\left(\frac{d_E^2}{2N_0}\right)^{1/2}\right] \tag{13}$$

for error event probability and

$$P_b \approx N_b Q\left[\left(\frac{d_E^2}{2N_0}\right)^{1/2}\right] \tag{14}$$

for decoded bit error probability. The multiplier N_b can be interpreted as the average multiplicity of information bit errors over *all* error events whose distance equals the free distance, divided by the number of information bits released per trellis level, k. These expressions are the aforementioned dominant terms in the series expansions for error probability.

The multipliers either emerge from the transfer function bounding, retaining the dominant term, or, for simple codes, they may be found by counting. The $k = 2$ coded 8-PSK code with 4 states has $N_e = 1$ and $N_b = \frac{1}{2}$ as discussed earlier, but the multipliers can be significantly larger, especially for codes with many states.

One should be cautious in use of (13) or (14), for they represent neither a strict upper or lower bound. Instead, these are asymptotically correct expressions, tightening as SNR increases. Free Euclidean distance is not the entire story at high SNR, but the multiplier factor is also relevant, although a more second-order effect. It can be noted that in the vicinity of $P_b = 10^{-5}$, every factor of 2 increase in the multiplier coefficient costs roughly 0.2 dB in effective SNR for typical TCM codes.

5. TABLES OF CODES

In distinction with algebraic block code constructions, good TCM codes are produced by computer search that optimizes free Euclidean distance. In this section, we list properties of optimal codes taken from Ref. 13. Code information is listed for state complexities ranging from 4 to 64 states, deemed to be the range of most practical interest.

Table 1 summarizes data for the best 1D TCM codes for M-level PAM, where $M = 2^{k+1}$. Data are provided about (1) the systematic feedback encoder connection polynomials, using the notation of Fig. 9 and right-justified octal form, where 13 corresponds to 001011; (2) the squared free distance normalized by $4a^2$, where

Table 1. Encoder Summary for Best PAM TCM Designs

States	$\tilde{k}$	$\mathbf{h}^1$	$\mathbf{h}^0$	$d_f^2/4a^2$	ACG(dB) A $k=1$, (1)	ACG(dB) A $k=2$, (2)	N_e
4	1	2	5	9.0	2.6	3.3	4
8	1	04	13	10.0	3.0	3.8	4
16	1	04	23	11.0	3.4	4.2	8
32	1	10	45	13.0	4.2	4.9	12
64	1	024	103	14.0	4.5	5.2	36

Key: (1) gain relative to 2-PAM; (2) gain relative to 4-PAM.
Source: Ref. 13.

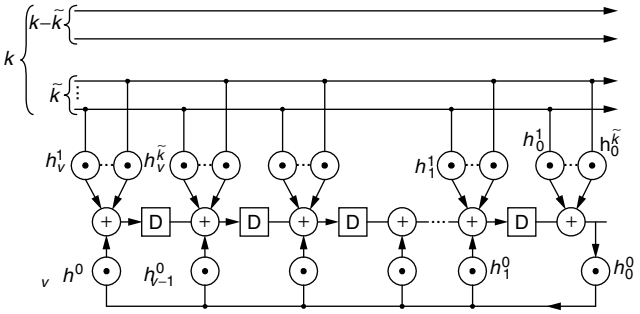


Figure 9. General systematic trellis encoder with feedback.

$2a$ is the PAM signal spacing along the real line; (3) the asymptotic coding gain over uncoded 2^k-ary PAM; and (4) the multiplier N_e in the first term of the union-bound expression for error event probability, as k becomes large. (In this case the same encoder operates on one input bit, regardless of k, at a given S.) The subset labels are 2 bits, meaning that the constellation is divided into four subsets. The remaining $k - \tilde{k}$ bits define a point from the selected subset. The first entry is the encoder used in the digital TV terrestrial broadcast standard in North America (see applications section below.)

Tables 2 and 3 provide similar information for 8-PSK and 16-PSK codes. Whereas 8-PSK is a good packing of 8 points in two dimensions, 16-PSK is not so attractive in an average energy sense, compared to 16-QAM. If peak energy (or amplitude) is of more concern, then 16-PSK improves [14]. The code described in Example 3 (above) is the 4-state code listed in Table 3, and it can be readily checked that the systematic form encoder of Fig. 7b concurs with the connection polynomials listed.

Table 2. Encoder Summary for Best 8-PSK TCM Designs, $k = 2$

States	$\tilde{k}$	$\mathbf{h}^2$	$\mathbf{h}^1$	$\mathbf{h}^0$	d_f^2/E_s	ACG (dB)	N_e
4	1	—	2	5	4.00	3.0	1
8	2	04	02	11	4.59	3.6	2
16	2	16	04	23	5.17	4.1	4
32	2	34	16	45	5.76	4.6	4
64	2	066	030	103	6.34	5.0	≈ 5.3

Key: (1) gain relative to 2-PAM; (2) gain relative to 4-PAM.
Source: Ref. 13.

Table 3. Encoder Summary for Best 16-PSK TCM Designs, $k = 3$

States	$\tilde{k}$	$\mathbf{h}^2$	$\mathbf{h}^1$	$\mathbf{h}^0$	d_f^2/E_s	ACG (dB)	N_e
4	1	—	02	05	1.324	3.5	4
8	1	—	04	13	1.476	4.0	4
16	1	—	04	23	1.628	4.4	8
32	1	—	10	45	1.910	5.1	8
64	1	—	024	103	2.000	5.3	2

Key: (1) gain relative to 2-PAM; (2) gain relative to 4-PAM.
Source: Ref. 13.

Table 4 compiles code and performance data for perhaps the most important case, 2D TCM using QAM constellations. As was the case in Table 1 for M-PAM, optimal codes with fixed state complexity share much in common as we increase k. For example in moving from $R = 3$ bits per symbol to $R = 4$, and so on, we can achieve this by simply growing the constellation by a factor of 2, adding one additional uncoded input bit and retaining the structure of the encoder for choosing cosets. This presents an attractive rate flexibility option.

Observe that for all 1D and 2D constellations, it is relatively easy to attain asymptotic coding gains of 3 dB to more than 5 dB with TCM, relative to an uncoded system with the same spectral efficiency. Notice also that for all these codes at most 2 of the k input bits influence the encoder state, regardless of state complexity. (This holds actually up through 256 states [13].) As k increases, the degree of parallelism in branching increases.

6. POWER SPECTRUM

If symbols are selected equiprobably and independently from a symmetric constellation (e.g., 16-QAM), the power spectrum of the transmitted signal does not contain spectral lines, and the continuum portion of the spectrum is given by the magnitude-squared of the Fourier transform of the modulator pulseshape [5,6]. A raised-cosine or root raised-cosine pulseshape is often selected to achieve band-limited transmission. The bandwidth of this spectrum scales according to the *symbol* rate, which is why high bandwidth efficiency accrues for large QAM signal constellations.

A standard approximation for the power spectrum of channel-coded signals is to adopt the spectrum of the uncoded modulation scheme, and frequency-scale according to the coded symbol rate, $R_{cs} = R_b/k$. This approximation models the coded symbol stream as an independent selection from the constellation, which, of course, is not valid. However, it turns out that for many coding techniques, the coded symbol stream exhibits pairwise independence; that is, the probability of successive pairs of symbols equals the product of the marginal probabilities. (This is not enough for strict independence.) This in turn implies that the coded stream is an uncorrelated one, and this is sufficient to yield a power spectrum identical to that of uncoded transmission, with care for scaling of the bandwidth according to number of information bits per symbol. In particular, codes proposed by Ungerboeck based on set partitioned labeling of the constellation and a linear convolutional encoder behave this way [15].

Example 5. Power Spectrum for Satellite Transmission. High-speed data transmission via satellite at 120 Mbps might utilize $R = 2$ coded 8-PSK transmission, so the symbol rate is 60 Msps (60 million symbols per second). Using root raised-cosine pulseshaping with rolloff factor 0.2, the total signal bandwidth occupies a range of 72 MHz, consistent with certain satellite transponder bandwidths.

7. APPLICATIONS

In this section, two contemporary applications of TCM are described. The actual transmission rates differ markedly, with one intended for dialup modems over the telephone network, while the other supports RF broadcast of digital TV. Nonetheless the motivation and operating principles are similar.

7.1. ATSC Television

The North American standard for terrestrial broadcasting of digital television goes under the name of the American Television Standards Committee (ATSC) standard [16]. There are numerous modes of operation, but one employs 4-state TCM encoding of 8-level PAM (a 1D TCM system). Coded symbols are passed through pulseshaping filters and vestigial sideband modulation is employed to keep the signal bandwidth within a 6-MHz allocation, even though the PAM symbol rate is 10.7 Msps.

The 4-state encoder is shown in Fig. 10. Since only 1 of the 2 input bits influences the encoder state, the $2^2 = 4$ branches leaving each state are organized into 2 sets of

Table 4. Encoder Summary for Best TCM Designs for QAM

						ACG (dB)			
States	$\tilde{k}$	$\mathbf{h}^2$	$\mathbf{h}^1$	$\mathbf{h}^0$	$d_f^2/4a^2$	$(k = 3)$,(1)	$(k = 4)$,(2)	$(k = 5)$,(3)	N_e
4	1	—	02	05	4.0	4.4	3.0	2.8	4
8	2	04	02	11	5.0	5.3	4.0	3.8	16
16	2	16	04	23	6.0	6.1	4.8	4.6	56
32	2	10	06	41	7.0	6.1	4.8	4.6	16
64	2	064	016	101	8.0	6.8	5.4	5.2	56

Key (1) gain relative to 8-PSK; (2) gain relative to 16-QAM; (3) gain relative to 32-QAM.
Source: Ref. 13.

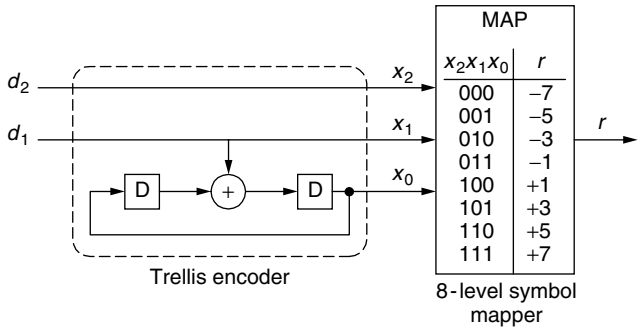


Figure 10. Trellis encoder for ATSC digital TV standard.

2, as for the trellis of Example 3 above. However, here the 1-step error event is not dominant due to the large intraset distance between points in sets of size 2. It can be shown that a 3-step error event dominates the distance, with free Euclidean distance $d_f^2 = \frac{36}{21}E_s$, leading to a 3.3-dB asymptotic coding gain over four-level PAM, (see Table 1).

Coherent detection is performed in the receiver with the aid of a pilot carrier to achieve phase lock without phase ambiguity. The TCM system is actually concatenated with an outer Reed–Solomon code to further improve performance on noisy channels. Channel equalization for multipath effects is also important.

7.2. V.32 Modem

One of the earliest applications of TCM appeared in the V.32 modem standard for 9600-bps transmission over the dialup telephone network. Prior to this time, all modem standards utilized uncoded transmission of modulator symbols. Use of 16-QAM, for example, can achieve 9600 bps with a symbol rate of 2400 sps, consistent with the voiceband channel bandwidth. Of course this data speed is now regarded as slow, but similar ideas have pushed speeds up by a factor of 3 over the same media [17].

To achieve the same throughput without increase in bandwidth, the V.32 standard employs a TCM design due to Wei [18] that maps onto 32-QAM with an 8-state encoder shown in Fig. 11 along with the constellation labeling of Fig. 12. In this trellis, 2 of the 4 bits entering the encoder do not influence the state vector, and are called *uncoded bits*. The remaining 2 are differentially encoded to handle rotational ambiguity (see Section 8), and the differential encoder output influences the state sequence, and hence the sequence of cosets. Each state has 16 branches emanating, organized as 4 sets of 4. In this trellis there are 1-step, 2-step, 3-step, and longer error events. The dominant error event(s) are 3-step events, and the asymptotic coding gain, relative to uncoded 16-QAM, is 4 dB.

Notice that the encoder is nonlinear over the binary field, due to the AND gates, but this does not complicate the decoder relative to a linear encoder. It is also noteworthy that this particular code has the same asymptotic coding gain as the best non–rotationally invariant code with 8 states (Table4). Normally, this extra constraint implies a small distance penalty, however.

A simulation of the performance of the V.32 design was performed with 10^6 bits sent through the system at each SNR in 1-dB steps from 6 to 11 dB. Results for decoded bit error probability, after differential decoding, are shown in Fig. 13, along with a plot of the expression $50Q[(2E_b/N_0)^{1/2}]$. The argument of the Q-function is obtained from the fact that the coding gain over uncoded 16-QAM is 4.0 dB, and that uncoded 16-QAM is 4.0 dB inferior in distance to uncoded QPSK. Thus the asymptotic energy efficiency is equivalent to that of uncoded QPSK, yet the system sends 4, rather than 2, bits per 2D symbol. The factor 50 is an empirically determined value that seems to fit the data well, and is consistent with the multipliers found in Table 4, given that multiple bit errors may occur per error event, and that the differential decoder increases the final bit error probability.

8. FURTHER TOPICS

8.1. Rotational Invariance

To perform coherent (known carrier phase) detection, the receiver must estimate the carrier's phase/frequency from the noisy received signal. This is normally done by some sort of feedback tracking loop with decision-directed operation to remove the influence of data symbols. However, given a symmetric constellation in one or two dimensions, this estimator will exhibit a phase ambiguity of 0/180° or 0/90/180/270°, respectively. Essentially, without additional side information, the demodulator has no way of discriminating phase beyond an ambiguity implied by the symmetry order of the constellation.

To resolve this ambiguity without first making hard decisions on symbols, then performing differential decoding prior to trellis decoding, a *rotationally invariant*

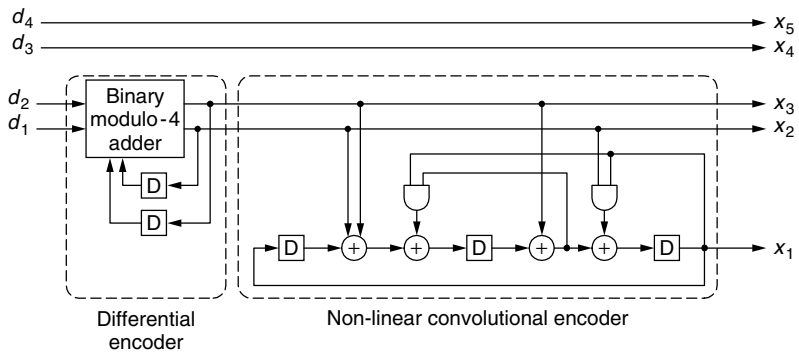


Figure 11. Trellis encoder for V.32 modem standard [18].

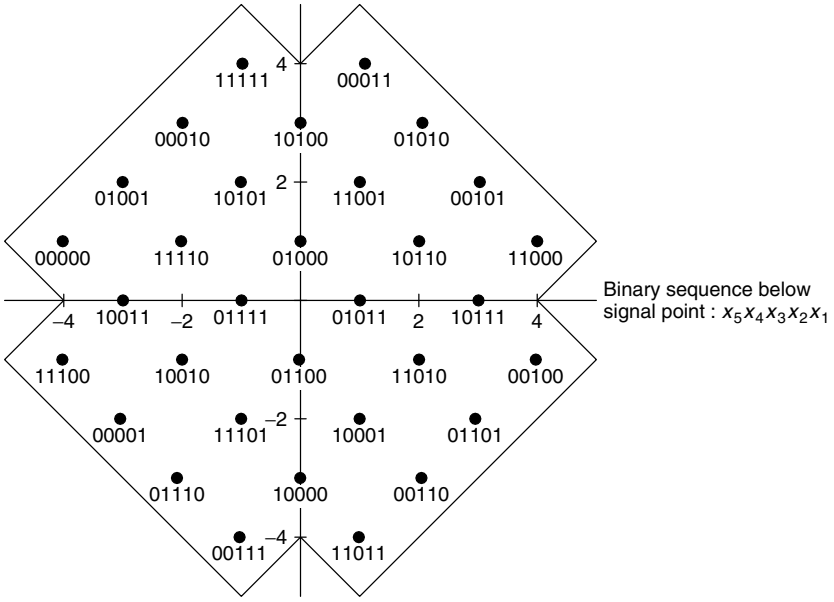


Figure 12. Constellation labeling for V.32 modem standard [18].

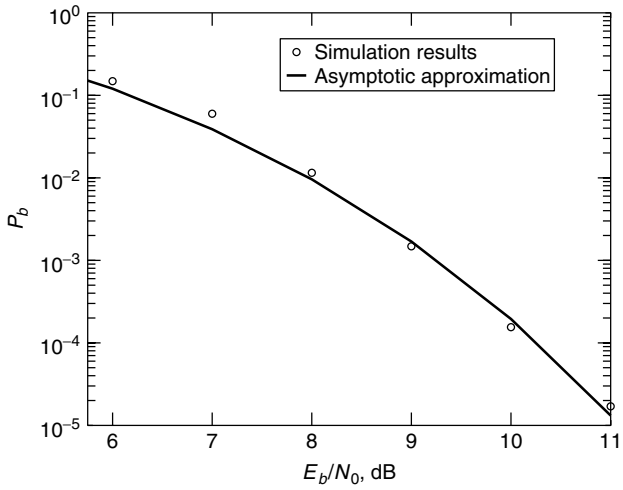


Figure 13. Performance of V.32 modem standard.

design is often preferred. This is a TCM design for which (1) a rotated version of every valid code sequence is also in the code space and (2) all such rotated versions correspond to the same information sequence, when launched from a different initial state. Usually a somewhat weaker condition is imposed — namely, that all rotated versions of the same code sequence correspond to encoder input sequences that, when differentially decoded, correspond to the same information sequence. In the latter case, the encoder is preceded by a precoder on some set of input bits, and the output of the TCM decoder is processed by a differential decoder acting on these same bits, thereby resolving the ambiguity. The penalty for not knowing phase outright is a slight increase in bit error probability.

Conditions for rotational invariant design were studied in detail by Trott et al. [19]. A notable example of a RI design is the 8-state TCM code for 32-QAM modulation due to Wei [18]. The encoder is depicted in Fig. 11 as described above.

8.2. Multidimensional TCM

The codes described thus far produce symbol streams of one- or two-dimensional constellation symbols. Slightly more efficient designs are available when we treat the constellation as a four-dimensional (or larger) object. The typical means of fashioning a four-dimensional constellation is to use two consecutive 2-D symbols. If $\mathcal{C}$ is 2D, then the set product $\mathcal{C} \times \mathcal{C}$ is four-dimensional.

This higher-dimensional constellation can be partitioned in a manner similar to that described earlier for 2D constellation, and then subsets are assigned to trellis arcs so that Euclidean distance is maximized. To keep the throughput fixed, however, one must realize that subsets are much larger in this case. For example, if a 2D TCM system is to send 3 bits per 2D symbol, say, using 16-QAM, then in a 4D TCM design, the encoder sends 6 bits per pair of 2D symbols. If the encoder has, say, 8 states, with arcs to 2 other states, then each arc must be labeled with a subset of size 32 (4D) points.

There are a few advantages offered by higher-dimensional codes, although they should be understood as second-order improvements: (1) it is possible to achieve slightly larger coding gains when throughput and trellis complexity are fixed; (2) rotationally invariant codes are easier to obtain in higher dimensions; and (3) Finally, one may exploit constellation shaping to better advantage, whereby the multi-dimensional constellation is more spherical, and 2D cross sections of these constellations are smaller than the corresponding constellation for 2D coding. Eyuboglu et al. [17] provide an example of a modem standard that employs multidimensional TCM together with shaping. Additional discussion on multidimensional TCM designs can be found in Refs. 13 and 20.

8.3. TCM on Fading Channels

Some propagation channels, notably wireless channels between terminals among buildings or influenced by terrain effects, are beset with the phenomenon of

fading. Essentially, the multiple paths for energy from transmitter to receiver may combine in a constructive or destructive manner, varying with time assuming some motion of the physical process. A common assumption, more accurate in the case of large numbers of scattering objects, is that the aggregate signal amplitude obeys a Rayleigh distribution, leading to the Rayleigh fading assumption. Normally, the fading is "slow"; that is, the amplitude is essentially fixed over many consecutive symbols. Traditional TCM coding does not fare well in such cases, for a single span of faded code symbols easily defeat the power of the decoder to combat noise.

If the fading effect can be made to appear essentially independent from symbol to symbol, on the other hand, then TCM can achieve *diversity* protection against fading, essentially mimicking a time diversity technique without the loss of throughput the latter implies. One way to achieve this independence is via interleaving, either in block or convolutional manner, to sufficient depth. (On slow-fading media, this may be impossible if constraints on delay are tight, as they are in two-way voice communication.) The encoder output is interleaved prior to modulation, and the demodulator output, together with a channel amplitude estimate, is deinterleaved prior to TCM decoding. The amplitude estimate is used in building the branch metric for the various trellis levels according to

$$\lambda = r_n a_n^* x_n^* \tag{15}$$

where a_n is the complex channel gain, assumed known, at time n, r_n is the complex channel measurement at time n, and x_n is the signal being evaluated. Recalling that the union bound on TCM performance involves a sum of 2-codeword probabilities, we are interested in minimizing the 2-codeword probability for all sequence pairs in the face of independent Rayleigh variables. It may be shown that the 2-codeword probability for confusing two sequences in independent Rayleigh fading, assuming perfect CSI (channel state information), is, at high SNR [21,22]

$$P(\mathbf{c}_1 \to \mathbf{c}_2) \le \frac{c}{(E_b/N_0)^D} \tag{16}$$

where D is the *Hamming* symbol distance between the two sequences. Here c is a proportionality factor dependent on the constellation and code, and E_b should be interpreted as the average received energy per bit.

This implies that the design criterion should (1) maximize the minimum Hamming distance between sequences (as opposed to Euclidean distance) and (2) within this class of codes, maximize the product of Euclidean distances between symbols on sequences on the minimum Hamming distance pairs, the so-called product distance [21]. This gives much different codes than the AWGN optimization provides. For example, a 4-state code for 8-PSK, to be used on the interleaved Rayleigh channel, should not have parallel transitions as in Fig. 6, but instead should have a 4-sets-of-1 trellis topology [21,22]. Here the minimum Hamming distance between sequences is 2, and the TCM system provides dual diversity.

Another approach to providing diversity is to use multiple symbols per trellis branch, amounting to multidimensional signaling. Continuing the previous case, instead of sending 2 bits per single constellation symbol, we could agree to send 4 bits per pair of symbols, keeping the rate the same. This so-called *multiple trellis-coded modulation* can achieve higher diversity order in some cases, as well as achieve greater product distance [23].

Another approach is to employ bit interleaving, rather than symbol interleaving. Here the objective is to design the code for maximal binary Hamming distance, rather than symbol distance. Readers are referred to Ref. 24 for further information in this regard.

Acknowledgments

The author gratefully acknowledges the assistance of Griffin Myers in preparing this work and the helpful comments of William Ryan.

BIOGRAPHY

Stephen Wilson is currently Professor of Electrical Engineering at the University of Virginia, Charlottesville, Virginia. His research interests are in applications of information theory and coding to modern communication systems, specifically digital modulation and coding techniques for satellite channels and wireless systems; spread spectrum technology; wireless antenna arrays; transmission on time-dispersive channels; and software radio. Prior to joining the University of Virginia faculty, Dr. Wilson was a staff engineer for The Boeing Company, Seattle, Washington, engaged in system studies for deep-space communication, satellite air-traffic-control systems, and military spread spectrum modem development. Prof. Wilson is presently area editor for Coding Theory and Applications of the IEEE Transactions on Communications, and the author of the graduate-level text Digital Modulation and Coding. He also acts as consultant to several industrial organizations in the area of communication system design and analysis and digital signal processing.

BIBLIOGRAPHY

1. G. Ungerboeck, Channel coding with amplitude/phase modulation, *IEEE Trans. Inform. Theory* **IT-28**: 55–67 (Jan. 1982).
2. E. Biglieri, D. Divsalar, P. McLane, and M. K. Simon, *Introduction to Trellis Coded Modulation with Applications*, Macmillan, New York, 1991.
3. G. D. Forney, Jr., Coset codes — part 1: Introduction and geometrical classification, *IEEE Trans. Inform. Theory* **IT-34**: 1123–1151 (Sept. 1988).
4. T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
5. J. Proakis, *Digital Communications*, McGraw-Hill, New York, 2001.
6. S. Wilson, *Digital Modulation and Coding*, Prentice-Hall, New York, 1996.
7. G. D. Forney, Jr., Trellis shaping, *IEEE Trans. Inform. Theory* **IT-38**: 281–300 (1992).
8. A. J. Viterbi, Error bounds for convolutional codes and asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**: 260–269 (1967).

9. G. D. Forney, Jr., The Viterbi algorithm, *IEEE Proc.* **61**: 268–278 (1973).

10. G. D. Forney, Jr., Convolutional codes I: Algebraic structure, *IEEE Trans. Inform. Theory* **IT-16**: 720–738 (1970).

11. G. D. Forney, Jr., Geometrically uniform codes, *IEEE Trans. Inform. Theory* **IT-37**: 1241–1260 (1991).

12. E. Zehavi and J. K. Wolf, On the performance evaluation of trellis codes, *IEEE Trans. Inform. Theory* **IT-32**: 196–202 (March 1987).

13. G. Ungerboeck, Trellis coded modulation with redundant signal sets, parts 1 and 2, *IEEE Commun. Mag.* **25**(2): 5–21 (Feb. 1987).

14. S. G. Wilson, P. J. Schottler, H. A. Sleeper, and M. T. Lyons, Rate 3/4 trellis coded 16-PSK: Code design and performance evaluation, *IEEE Trans. Commun.* **COM-32**: 1308–1315 (Dec. 1984).

15. E. Biglieri, Ungerboeck codes do not shape the power spectrum, *IEEE Trans. Inform. Theory* **IT-32**: 595–596 (July 1986).

16. J. Whitaker, *DTV Handbook*, McGraw-Hill, New York, 2001.

17. M. V. Eyuboglu, G. D. Forney, Jr., P. Dong, and G. Long, Advanced modulation techniques for V. fast, *Eur. Trans. Telecommun.* **4**: 243–256 (May 1993).

18. L. F. Wei, Rotationally invariant convolutional channel coding with expanded signal space, part II: Nonlinear codes, *IEEE J. Select. Areas Commun.* **SAC-2**: 672–686 (Sept. 1984).

19. M. D. Trott, S. Benedetto, R. Garello, and M. Mondin, Rotational invariance of trellis codes: Encoders and precoders, *IEEE Trans. Inform. Theory* **IT-42**: 751–765 (1996).

20. S. S. Pietrobon et al., Trellis-coded multidimensional phase modulation, *IEEE Trans. Inform. Theory* **IT-36**: 63–89 (Jan. 1990).

21. D. Divsalar and M. K. Simon, The design of trellis-coded MPSK for fading channels: Set partitioning for optimum code design, *IEEE Trans. Commun.* **COM-36**: 1004–1011 (Sept. 1988).

22. Y. S. Leung and S. G. Wilson, Trellis coding for fading channels, in *ICC Conf. Record* (Seattle), 1987.

23. D. Divsalar and M. K. Simon, Multiple trellis-coded modulation, *IEEE Trans. Commun.* **COM-36**: 410–419 (April 1988).

24. G. Caire, G. Taricco, and E. Biglieri, Bit-interleaved coded modulation, *IEEE Trans. Inform. Theory* **IT-44**: 927–946 (1998).

TRELLIS CODING

CHRISTIAN SCHLEGEL
University of Alberta
Edmonton, Alberta, Canada

1. INTRODUCTION

The tremendous growth of high-speed logic circuits and very large-scale integration (VLSI) has ushered in the digital information age, where information is stored, processed, and moved in digital format. Among other advantages, digital signals possess an inherent robustness in noisy communications environments. If distorted, they can be restored, and, through signal processing techniques, errors that occur during transmission can be corrected. This process is called *error control coding*, and is accomplished by introducing dependencies among a large number of digital symbols. *Trellis coding* is a specific, widespread error control methodology.

Figure 1 shows the basic configuration of a point-to-point digital communications link. The digital data to be transmitted over this link typically consists of a string of binary symbols: ones and zeros. These symbols enter the *encoder/modulator* whose function it is to prepare them for transmission over a channel, which may be any one of a variety of physical channels. The encoder accepts the input digital data and introduces controlled redundancy for transmission over the channel. The modulator converts discrete symbols given to it by the encoder into waveforms that are suitable for transmission through the channel. On the receiver side, the demodulator reconverts the waveforms back into a discrete sequence of received symbols, and the decoder reproduces an estimate of the digital input data sequence, which is subsequently used by the data sink. The purpose of the error control functions is to maximize data reliability when transmitted over an unreliable channel.

An important auxillary function of the receiver is synchronization, that is, the process of acquiring carrier frequency and phase, and symbol timing in order for the receiver to be able to operate. Compared to data detection, synchronization is a relatively slow process, and therefore we usually find these two operations separated in receiver implementations.

Another important mechanism in many communication systems is *automatic repeat request* (ARQ). In ARQ the receiver additionally performs error detection, and, through a return channel, requests retransmission of data blocks that cannot be reconstructed with sufficient confidence. ARQ can usually improve the data transmission quality substantially, but a return channel, which is needed for ARQ, is not always available, or may be impractical. For a deep-space probe, for example, ARQ is infeasible since the return path takes too long (several hours!). Equally so, for speech encoded signals ARQ is usually infeasible since only a maximum speech delay of 200 ms is acceptable. In broadcast systems, ARQ is ruled out for obvious reasons.

Error control coding without ARQ is termed *forward error control* (FEC) coding. FEC is more difficult to perform than simple error detection and ARQ, but dispenses with the return channel. Often, FEC and ARQ are combined in hybrid error control systems.

2. IMPORTANCE OF FORWARD ERROR CONTROL (FEC) CODING IN A DIGITAL COMMUNICATIONS SYSTEM

The modern approach to error control coding in digital communications started with the groundbreaking work of Shannon [1], Hamming [2], and Golay [3]. While Shannon advanced a theory to explain the fundamental limits on the efficiency of communications systems, Hamming and Golay were the first to develop practical error control schemes. The new revolutionary paradigm born was

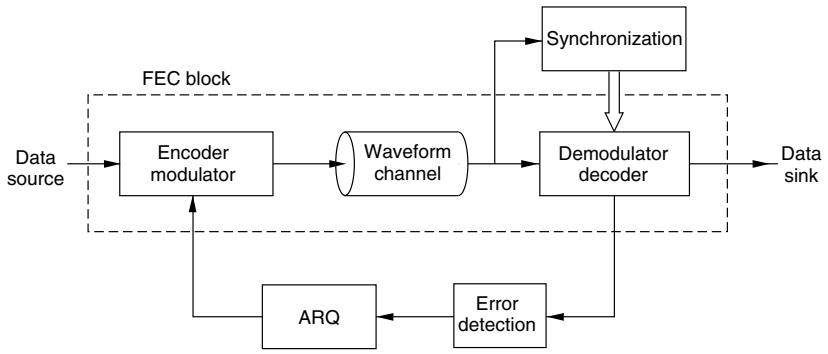


Figure 1. System diagram of a complete point-to-point communication system for digital data.

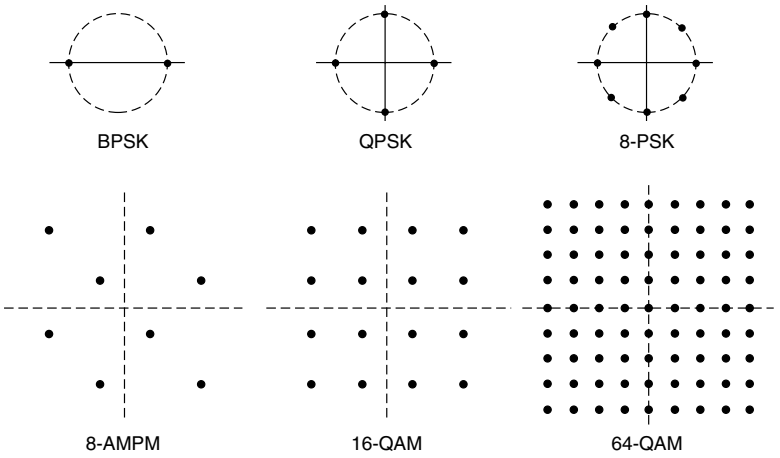


Figure 2. Popular 2D signal constellations used for digital radio systems.

one in which errors are no longer synonymous with data that are irretrievably lost, but by clever design, errors could be corrected, or avoided altogether. Although Shannon's theory promised that large improvements in the performance of communication systems could be achieved, practical improvements had to be excavated by laborious work. In the process, coding theory has evolved into a flourishing branch of applied mathematics [4].

The starting point to coding theory is Shannon's celebrated formula for the capacity of an ideal band-limited Gaussian channel, given by

$$C = W \log_2 \left(1 + \frac{S}{N}\right) \qquad \text{bps (bits per second)} \qquad (1)$$

In this formula C is the channel capacity, which is the maximum rate of information, measured in bits per second, which can be transmitted through this channel; W is the bandwidth of the channel, and S/N is the signal-to-noise power ratio at the receiver. Shannon's main theorem, which accompanies Eq. (1), asserts that error probabilities as small as desired can be achieved as long as the transmission rate R through the channel (in bits per second) is smaller than the channel capacity C. This can be achieved by using an appropriate encoding and decoding operation. On the other hand, the converse of this theorem states that for rates $R > C$ there is a significant error rate which cannot be reduced no matter what processing is invoked.

2.1. Bandwidth and Power

In order to be able to appreciate fully the concepts of trellis coding and trellis-coded modulation, some signal basics are needed. Nyquist showed in 1928 [25] that a channel of bandwidth W (in Hertz) is capable of supporting approximately $2W$ independent signal dimensions per second. If two carriers [$\sin(2\pi f_c)$ and $\cos(2\pi f_c)$] are used in quadrature, as in double-sideband suppressed-carrier (DSBSC) amplitude modulation, we alternatively have W pairs of dimensions (or complex dimensions) per second, leading to the popular QAM (quadrature amplitude modulation) constellations, represented by points in two-dimensional (2D) space. Some popular constellations for digital communications are shown in Fig. 2.

The parameter that characterizes how efficiently a system uses its allotted bandwidth is the *spectral efficiency* η, defined as

$$\eta = \frac{\text{bit rate}}{\text{channel bandwidth } W} \qquad \text{(bps/Hz)} \qquad (2)$$

Using (1) and dividing by W, we obtain the maximum spectral efficiency for an additive white Gaussian noise (WGN) channel, the *Shannon limit*, as

$$\eta_{\max} = \log_2 \left(1 + \frac{S}{N}\right) \qquad \text{(bps/Hz)} \qquad (3)$$

To calculate η, we must suitably define the channel bandwidth W. One commonly used definition is the 99%

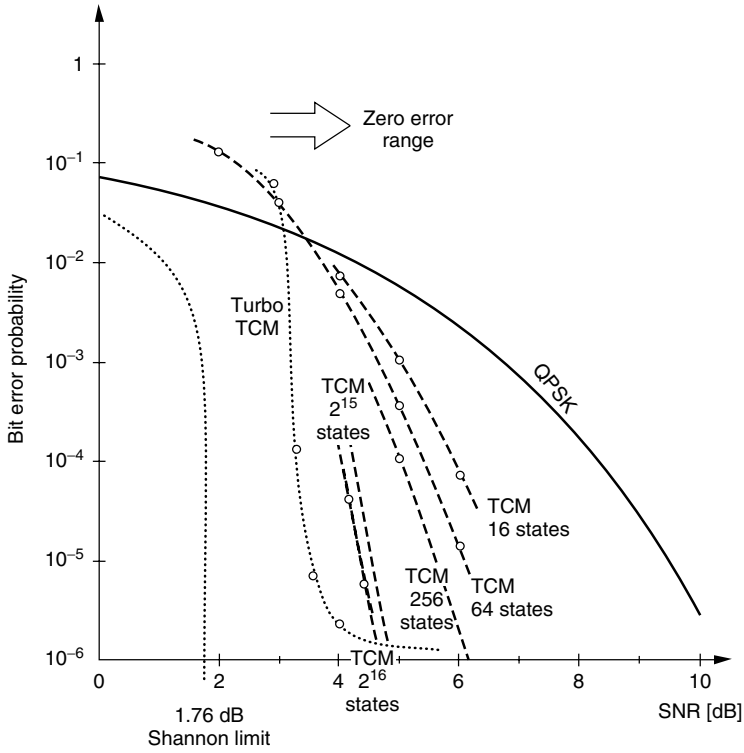


Figure 3. Bit error probability of quadrature phase shift keying (QPSK) and selected 8-PSK trellis-coded modulation (TCM) methods as a function of the normalized signal-to-noise ratio.

bandwidth definition, where W is defined such that 99% of the transmitted signal power falls within the band of width W. This 99% bandwidth corresponds to an out-of-band power of -20 dB.

The average signal power S can be expressed as

$$S = \frac{kE_b}{T} = RE_b \tag{4}$$

where E_b is the energy per bit, k is the number of bits transmitted per symbol, and T is the duration of that symbol. The parameter $R = k/T$ is the transmission rate of the system in bits per second. Rewriting the signal-to-noise power ratio S/N, where $N = WN_0$, where total noise power equals the noise power spectral density N_0 multiplied by the width of the transmission band, we obtain

$$\eta_{\max} = \log_2\left(1 + \frac{RE_b}{WN_0}\right) = \log_2\left(1 + \eta\frac{E_b}{N_0}\right) \tag{5}$$

Since $R/W = \eta_{\max}$ is the limiting spectral efficiency, we obtain a bound from (5) on the minimum bit energy required for reliable transmission, given by

$$\frac{E_b}{N_0} \geq \frac{2^{\eta_{\max}} - 1}{\eta_{\max}} \tag{6}$$

which is also called the *Shannon bound*.

In the limit as we allow the signal to occupy an infinite amount of bandwidth, that is, $\eta_{\max} \to 0$, we obtain

$$\frac{E_b}{N_0} \geq \lim_{\eta_{\max}\to 0} \frac{2^{\eta_{\max}} - 1}{\eta_{\max}} = \ln(2) = -1.59 \text{ dB} \tag{7}$$

the minimum bit energy to noise power spectral density required for reliable transmission.

2.2. Communications System Performance with FEC

In order to compare different communications systems, a second parameter, expressing the power efficiency, has to be considered also. This parameter is the information bit error probability P_b. Figure 3 shows the error performance of QPSK, a popular modulation method for satellite channels that allows data transmission of rates up to 2 bps/Hz (bits per second per Hertz). The bit error probability of QPSK is shown as a function of the signal-to-noise ratio S/N per dimension normalized per bit, henceforth called SNR. It is evident that an increased SNR provides a gradual decrease in error probability. This contrasts markedly with Shannon's theory, which promises zero(!) error probability at a spectral rate of 2 bps/Hz, if SNR > 1.5 (1.76 dB). The dashed line in the figure represents the *Shannon bound* adjusted to the bit error rate.

Also shown in Fig. 3 is the performance of several trellis-coded modulation (TCM) schemes using 8-ary phase-shift keying (8-PSK), and the improvement of coding becomes evident. The difference in SNR for an objective target bit error rate between a coded system and an uncoded system is termed the *coding gain*. It is important to point out here that TCM achieves these coding gains without requiring more bandwidth than the uncoded QPSK system. The figure also shows the performance of a more recently proposed method, Turbo trellis-coded modulation (TTCM). This extremely powerful coding scheme comes very close to the Shannon limit.

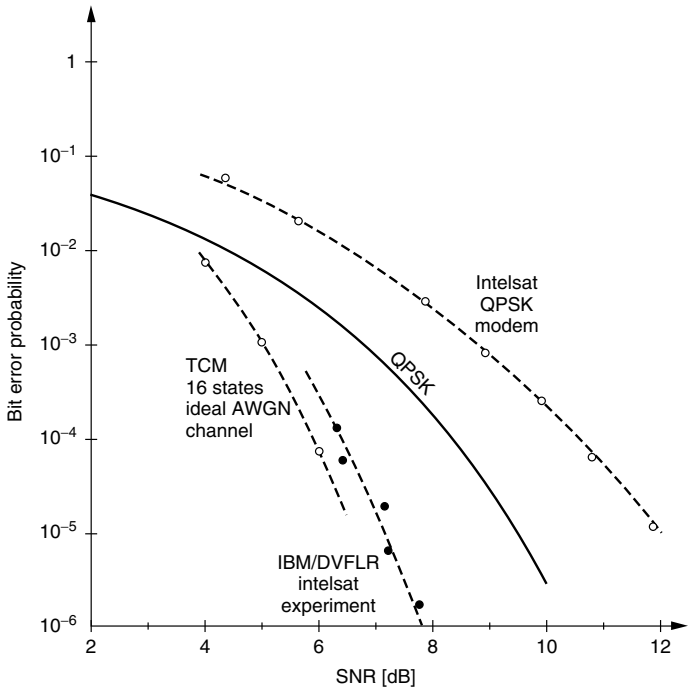


Figure 4. Measured bit error probability of QPSK and a 16-state 8-PSK (TCM) modem over a 64-kps satellite channel [10].

As discussed later, a trellis code is generated by a circuit with a finite number of internal states. The number of these states is a measure of its decoding complexity if optimal decoding, that is, *maximum-likelihood* (ML) decoding, is used. However, the two very large codes shown are not decoding via optimal decoding methods; they are sequentially decoded (see Schlegel's [5] discussion of Wang and Costello's paper [7]). TTCM is the most complex coding scheme, requiring not only two trellis decoders but also soft information decoding and storage of large blocks of data. FEC realizes the promise of Shannon's theory, which states that for a desired error rate of $P_b = 10^{-6}$ we can gain almost 9 dB in expended signal energy with respect to QPSK.

In Fig. 4 we compare the performance of a 16-state 8-PSK TCM code used in an experimental implementation of a single-channel-per-carrier (SCPC) modem operating at 64 kbps (1000 bits per second) [10] against QPSK and the theoretical performance established via simulations. As an interesting observation, the 8-PSK TCM modem comes much closer to its theoretical performance than the original QPSK modem, and a coding gain of 5 dB is achieved.

Figure 5 shows the performance of selected convolutional codes on an additive white Gaussian noise channel. Contrary to TCM, convolutional codes (with BPSK modulation) do not preserve bandwidth and the gains in power efficiency in Fig. 5 are partly obtained by a power bandwidth tradeoff; specifically, the rate $\frac{1}{2}$ convolutional codes plotted here require twice as much bandwidth as does uncoded transmission. This bandwidth expansion may not be an issue in deep-space communications and the application of error control to spread-spectrum systems. As a consequence, for the same complexity, convolutional codes achieve a higher coding gain than does TCM. Turbo coding, the most complex of the coding schemes, achieves a performance within fractions of 1 dB of the Shannon limit.

The field of error control and error-correction coding somewhat naturally breaks into two disciplines, namely, block coding and trellis coding. While block coding, which is approached mostly as applied mathematics, has produced the bulk of publications in error control, trellis coding seems to be favored in most practical applications. One reason for this is the ease with which soft-decision decoding can be implemented for trellis codes. Soft decision is the operation when the demodulator no longer makes any (hard) decisions on the transmitted symbols, but passes the received signal values directly on to the decoder. The decoder, in turn, operates on reliability information obtained by comparing the received signals with the possible set of transmitted signals. This gives soft decision a 2-dB advantage. Also, trellis codes are better matched to high-noise channels, that is, their performance is less sensitive to SNR variations than the performance of block codes. In many applications the trellis decoders act as "SNR transformers"; that is, they lower the signal-to-noise ratio from the input to the output. Such SNR transformers find application in Turbo decoders, coded channel equalizers, and coded multiple-access systems. The irony is, that in many cases, block codes can be decoded more successfully using methods developed originally for trellis codes than using their specialized decoding algorithms.

Figure 6 shows the power and bandwidth efficiencies of some popular uncoded quadrature constellations as well as that of a number of coded transmission schemes. The plot clearly demonstrates the advantages of coding. The trellis-coded modulation schemes used in practice, for example, achieve a power gain of up to 6 dB without loss in spectral efficiency. The convolutionally encoded methods [e.g., the points labeled with (2,1,6) CC (convolutional code), which

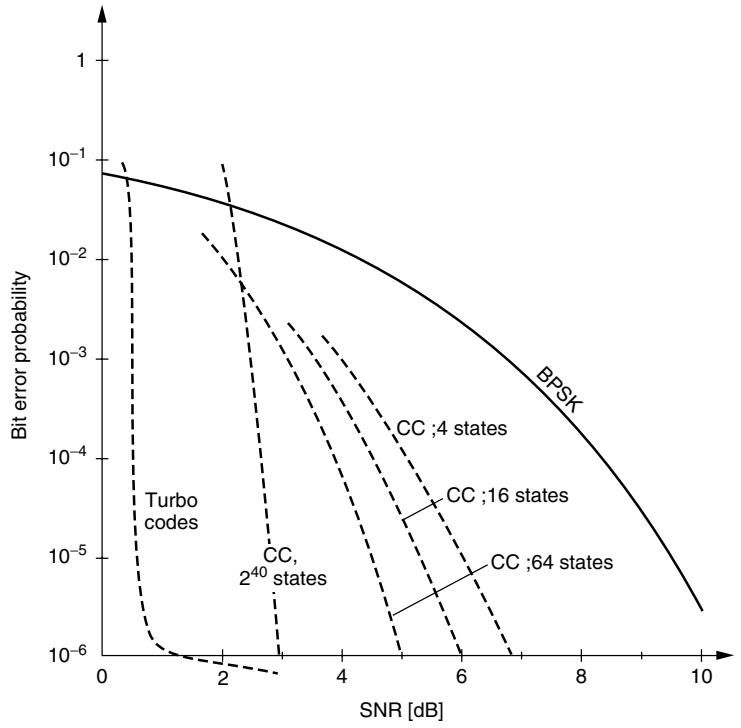


Figure 5. Bit error probability of selected rate $R = \frac{1}{2}$ convolutional codes as a function of the normalized SNR. The very large code is decoded sequentially, while the performance of the other codes, except the Turbo code, is for maximum-likelihood decoding, discussed in Section 5. (*Sources:* Refs. 8 and 9).

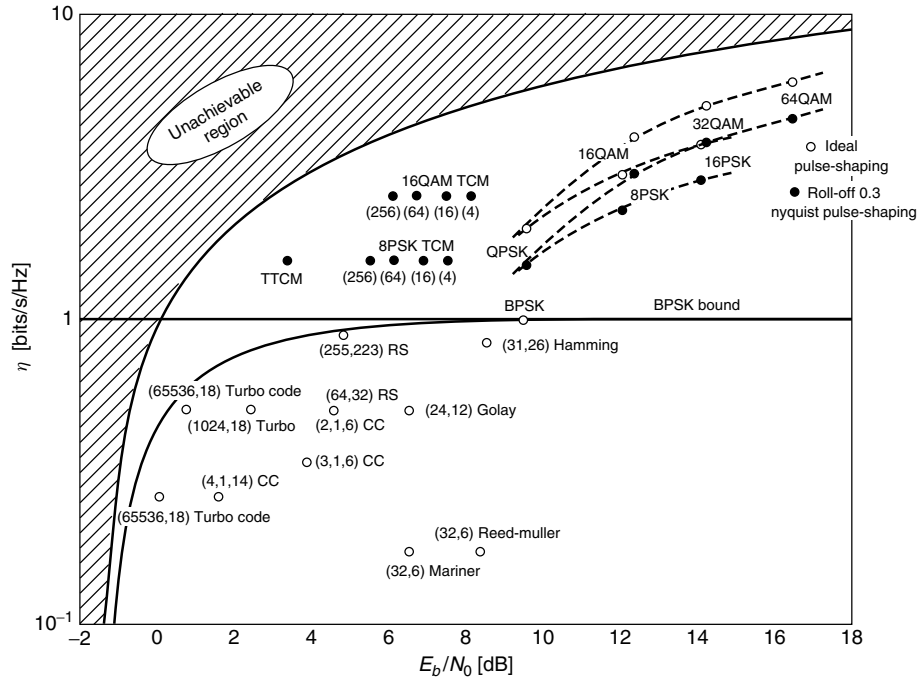


Figure 6. Spectral and power efficiencies achieved by various coded and uncoded transmission methods.

is a rate $R = \frac{1}{2}$ convolutional code with 2^6 states, and (4,1,14) CC, a rate $R = \frac{1}{4}$ convolutional code with 2^{14} states] achieve a gain in power efficiency, at the expense of spectral efficiency.

2.3. A Brief History of Error Control Coding

Trellis coding celebrated its first success in the application of convolutional codes to deep-space probes in the 1960s and 1970s. For a long time afterward, error-control coding was considered a curiosity with deep-space communications as its only viable application.

If we start with uncoded binary phase shift keying (BPSK) as our baseline transmission method, and assume coherent detection, we can achieve a bit error rate of $P_b = 10^{-5}$ at a bit energy : noise power spectral density ratio of $E_b/N_0 = 9.6$ dB, and a spectral efficiency of 1 bit per dimension. From the Shannon limit in Fig. 6 it can be seen that error-free transmission is theoretically

achievable with $E_b/N_0 = -1.59$ dB, indicating that a power savings of over 11 dB is possible by using FEC.

One of the early attempts to close this signal energy gap was the use of a rate $\frac{6}{32}$ biorthogonal (Reed–Muller) block code [4]. This code was used on the *Mariner, Mars* and *Viking* missions. This system had a spectral efficiency of 0.1875 bits/symbol and an optimal soft-decision decoder achieved $P_b = 10^{-5}$ with an $E_b/N_0 = 6.4$ dB. Thus, the (32,6) biorthogonal code required 3.2 dB less power than BPSK at the cost of a fivefold increase in the bandwidth (see Fig. 6).

In 1967, a new algebraic decoding technique was discovered for Bose–Chaudhuri–Hocquenghem (BCH) codes, which enabled the efficient *hard-decision decoding* of an entire class of block codes. For example, the (255,223) BCH code has an $\eta \approx 0.5$ bits/symbol, assuming ideal pulse shaping, and achieves $P_b = 10^{-5}$ with $E_b/N_0 = 5.7$ dB using algebraic decoding.

Sequential decoding allowed the decoding of long-constraint-length convolutional codes, and was first used on the *Pioneer* 9 mission. The *Pioneer* 10 and 11 missions in 1972 and 1973 both used a long-constraint-length (2,1,31) nonsystematic convolutional code [26]. A sequential decoder was used that achieved $P_b = 10^{-5}$ with $E_b/N_0 = 2.5$ dB, and $\eta = 0.5$. This is only 2.5 dB away from the capacity of the channel.

The *Voyager* spacecraft launched in 1977 used a short-constraint-length (2,1,6) convolutional code in conjunction with a soft-decision optimal decoder achieving $P_b = 10^{-5}$ at $E_b/N_0 = 4.5$ dB and a spectral efficiency of $\eta = 0.5$ bits/symbol. The biggest such optimal decoder built to date [27] found application in the *Galileo* mission, where a (4,1,14) convolutional code is used. This code has a spectral efficiency of $\eta = 0.25$ bits/symbol and achieves $P_b = 10^{-5}$ at $E_b/N_0 = 1.75$ dB. Its performance is therefore also 2.5 dB away from the capacity limit. The systems for *Voyager* and *Galileo* are further enhanced by the use of concatenation in addition to the convolutional inner code. An outer (255,223) Reed–Solomon code [4] is used to reduce the required signal-to-noise ratio by 2.0 dB for the *Voyager* system and by 0.8 dB for the *Galileo* system.

More recently, Turbo codes [6] using iterative decoding virtually closed the gap to capacity by achieving $P_b = 10^{-5}$ at a spectacularly low E_b/N_0 of 0.7 dB with $R_d = 0.5$ bits/symbol. It appears that the half-century of effort to reach capacity has been achieved with this latest invention. More recently, low-density parity-check codes have also been demonstrated to obtain performances very close to capacity.

Space applications of error-control coding have met with spectacular success, and were for a long time the major, if not only, area of application for FEC. The belief that coding was useful only in improving power efficiency of digital transmission was prevalent. This attitude was overturned only by the spectacular success of error-control coding on voiceband data transmission modems. Here it was not the power efficiency that was the issue, but rather the spectral efficiency, that is given a standard telephone channel with an essentially fixed bandwidth and SNR, what was the maximum practical rate of reliable transmission.

The first commercially available voice-band modem in 1962 achieved a transmission rate of 2400 bps. Over the next 10–15 years these rates improved to 9600 bps, which was then considered to be the maximum achievable rate, and efforts to push the rate higher were frustrated. Ungerböck's invention of trellis-coded modulation in the late 1970s, however, opened the door to further, unanticipated improvements. The modem rates jumped to 14,400 bps and then to 19,200 bps, using sophisticated TCM schemes [28]. The latest chapter in voiceband data modems is the establishment of the CCITT (Consultative Committee for International Telephony and Telegraphy) V.34 modem standard [29]. The modems specified therein achieve a maximum transmission rate of 28,800 bps, and extensions to V.34 (V.34bis) to cover two new rates at 31,200 bps and 33,600 bps have been established. These rates need to be compared to estimates of the channel capacity for a voiceband telephone channel, which are somewhere around 30,000 bps, essentially achieving capacity on this channel also. Note that the common 56-kbps modems used in the downward direction exploit the higher capacity provided by a digital connection to the switching station.

3. TRELLIS CODING

A trellis encoder consists of two parts, a *finite-state machine* (FSM) which generates the trellis of the code, and a modulator, called the *signal mapper*, which maps state transitions of the FSM into output symbols suitable for transmission. This encoder/modulator pair is shown in Fig. 7 for a specific example. The FSM in this example has a total of eight states, where the state s_r at time r of the FSM is defined by the contents of the delay cells: $s_r = (s_r^{(2)}, s_r^{(1)}, s_r^{(0)})$. The signal mapper performs a memoryless mapping of the 3 bits $v_r = (u_r^{(2)}, u_r^{(1)}, v_r^{(0)})$ into one of the eight symbols of an 8-PSK signal set. The FSM accepts 2 input bits $u_r = (u_r^{(2)}, u_r^{(1)}), u_r^{(i)} = \{0, 1\}$ at each symbol time r, and transitions from a state s_r to one of four possible successor states s_{r+1}. In this fashion the trellis encoder generates the (possibly) infinite sequence of symbols $\underline{x} = (\ldots, x_{-1}, x_0, x_1, x_2, \ldots)$. There are four choices at each time r, which allows us to transmit 2 information bits/symbol, the same as for QPSK, but using a larger constellation. This fact is called *signal set expansion*, and is necessary in order to introduce the redundancy required for error control.

A graphical interpretation of the function of the FSM is given by the state-transition diagram shown in Fig. 8. The nodes in this transition diagram are the possible states of the FSM, and the branches represent the possible transitions between them. Each branch can now be labeled by the pair of input bits $u = (u^{(2)}, u^{(1)})$ which cause the transition, and by either the output triple $v = (u^{(2)}, u^{(1)}, v^{(0)})$, or the output signal $x(v)$. [In Fig. 8 we have used $x(v)$, represented in octal notation, i.e., $x_{\text{oct}}(v) = u^{(2)}2^2 + u^{(1)}2^1 + v^{(0)}2^0$.]

If we index the states by both their content and the time index r, Fig. 8 expands into the *trellis diagram*, or simply the *trellis* of the code, shown in Fig. 9. This trellis is the two-dimensional representation of the state

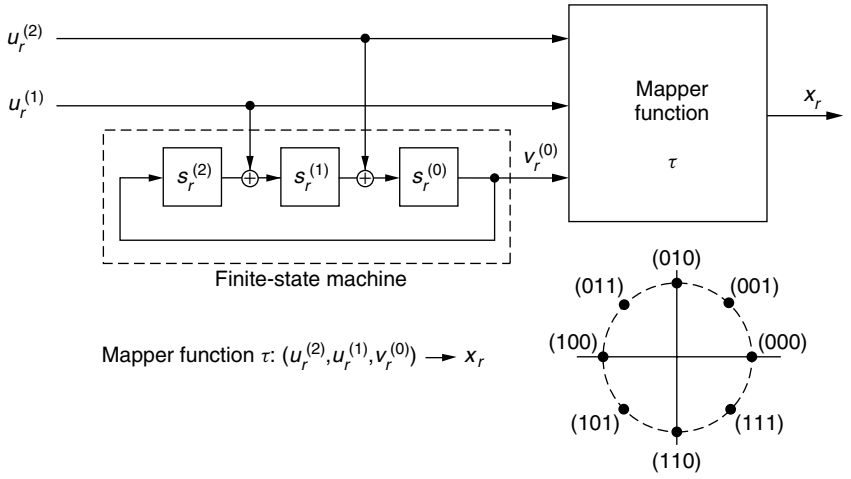


Figure 7. Trellis encoder with an 8-state finite-state machine (FSM) driving a 3-bit to 8-PSK signal mapper. All inputs and outputs of the FSM and all operations are binary modulo-2 operations.

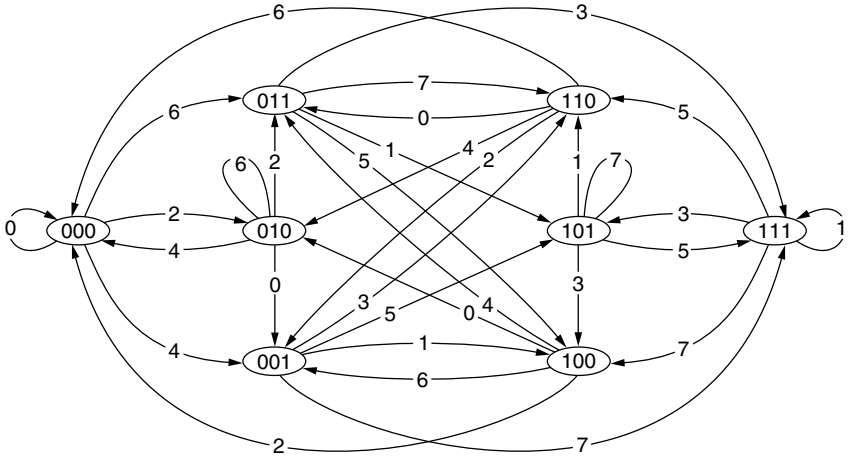


Figure 8. State-transition diagram of the encoder from Fig. 7. The labels on the branches are the encoder output signals x, in decimal notation.

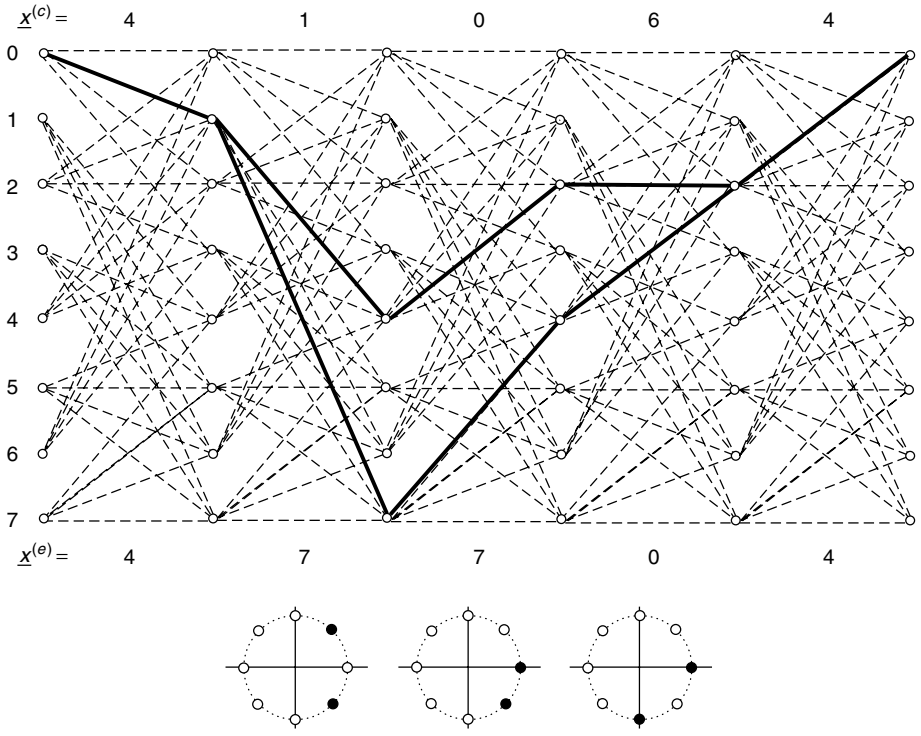


Figure 9. Section of the trellis of the encoder in Fig. 7. The two solid lines depict two possible paths with their associated signal sequences through this trellis. The numbers on top are the signals transmitted if the encoder follows the upper path, and the numbers at the bottom are those on the lower path.

and time–space of the encoder; it captures all achievable states at all time intervals, usually starting from an originating state (commonly state 0), and terminating in a final state (commonly also state 0). In practice the length of the trellis will be several hundred or thousands of time units, possibly even infinite, corresponding to continuous operation. When and where to terminate the trellis is a matter of practical consideration.

Each path through the trellis corresponds to a unique message, as a sequence of symbols, and is associated with a unique sequence of signals. The term *trellis-coded modulation* originates from the fact that these encoded sequences consist of high-level modulated symbols, rather than simple binary symbols.

The FSM puts restrictions on the symbols that can be in a sequence, and these restrictions are exploited by a smart decoder. In fact, what counts is the distance between signal sequences $\underline{x}$, and not the distance between individual signals as in uncoded transmission. Let us then assume that such a decoder can follow all possible sequences through the trellis, and it makes decisions between sequences. This is illustrated in Fig. 9 for two sequences $\underline{x}^{(e)}$ (erroneous) and $\underline{x}^{(c)}$ (correct). These two sequences differ in the three symbols shown. An optimal decoder will make an error between these two sequences with probability $P_s = Q(\sqrt{d_{ec}^2 E_s/2N_0})$, where $d_{ec}^2 = 4.586 = 2 + 0.586 + 2$ is the squared Euclidean distance between $\underline{x}^{(e)}$ and $\underline{x}^{(c)}$, which is much larger than the QPSK distance of $d^2 = 2$, and E_s is the energy per signal. Examining all possible sequence pairs $\underline{x}^{(e)}$ and $\underline{x}^{(c)}$, one finds that those highlighted in Fig. 9 have the smallest squared Euclidean distance, and hence, the probability that the decoder makes an error between those two sequences is the most likely error event.

We now see that by virtue of performing sequence decoding, rather than symbol decoding, the distances between competing candidates can be increased, even though the signal constellation used for sequence coding has a smaller minimum distance between signal points than the uncoded constellation for the same rate. For this code, we may decrease the symbol power by about 3.6 dB; thus, we use less than half the power needed for QPSK to achieve the same performance. This superficial analysis belies the complexity of a more precise error analysis [5] but serves as a crude estimate.

A more precise error analysis is not quite so simple since the possible error paths in the trellis are highly correlated, which makes an exact analysis of the error probability impossible for all except the most simple cases. In fact, much work has gone into analyzing the error behavior of trellis codes [5].

4. CONSTRUCTION OF CODES

From Fig. 9 we see that an error path diverges from the correct path at some state and merges with the correct path again at a (possibly) different state. The task of designing a good trellis code means designing a trellis code for which different symbol sequences are separated by large squared Euclidean distances. Of particular importance is the *minimum squared Euclidean distance*, termed d_{free}^2, namely, $d_{\text{free}}^2 = \min_{\underline{x}^{(i)}, \underline{x}^{(j)}} \|\underline{x}^{(i)} - \underline{x}^{(j)}\|^2$. A code with a large d_{free}^2 is generally expected to perform well, and d_{free}^2 has become the major design criterion for trellis codes.

One heuristic design rule [11], which was used successfully in designing codes with large d_{free}^2, is based on the following observation. If we assign to the branches leaving a state signals from a subset with large distances between points, and likewise assign such signals to the branches merging into a state, we are assured that the total distance is at least the sum of the minimum distances between the signals in these subsets. For our 8-PSK code example, we can choose these subsets to be QPSK signal subsets of the original 8-PSK signal set. This is done by partitioning the 8-PSK signal set into two QPSK sets as illustrated in Fig. 10. The mapper function is now chosen such that the state information bit $v^{(0)}$ selects the subset and the input bits u select a signal within the subset. Since all branches leaving a state have the same state information bit $v^{(0)}$, all the branch signals are in either subset A or subset B, and the difference between two signal sequences picks up an incremental distance of $d^2 = 2$ over the first branch of their difference. These are Ungerböeck's [11] original design rules.

The values of the tap coefficients (see Fig. 13) $h^{(2)} = 0, h_1^{(2)}, \ldots, h_{\nu-1}^{(2)}, 0; h^{(1)} = 0, h_1^{(1)}, \ldots, h_{\nu-1}^{(1)}, 0;$ and $h^{(0)} = 0, h_1^{(0)}, \ldots, h_{\nu-1}^{(0)}, 0$ in the encoder are usually found via computer search programs or heuristic construction algorithms. The parameter ν is called the *constraint length*, and is the length of the shortest error path. Table 1 shows the best 8-PSK trellis codes found to date using 8-PSK with *natural mapping*, which is the bit assignment shown in Fig. 7. The figure gives the connector coefficients, d_{free}^2, the average number $A_{d_{\text{free}}}$ of paths with d_{free}^2, and the average number $B_{d_{\text{free}}}$ of bit differences on these paths. Both $A_{d_{\text{free}}}$ and $B_{d_{\text{free}}}$, as well as the higher-order average path pair number, called *multiplicities*, are important parameters determining the error performance of a trellis code. The tap coefficients are given in octal notation, for instance, $h^{(0)} = 23 = 10111$, where a 1 means connected and a 0 means no connection.

From Table 1 one can see that an asymptotic coding gain (coding gain for SNR $\to \infty$ over the reference constellation that is used for uncoded transmission at the same rate) of about 6 dB can quickly be achieved

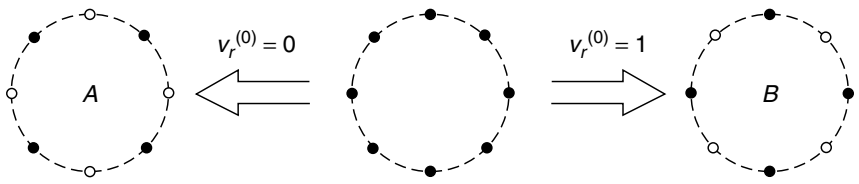


Figure 10. 8-PSK signal set partitioned into constituent QPSK signal sets.

Table 1. Connectors, Free-Squared Euclidean Distance, and Asymptotic Coding Gains of Some Maximum Free-Distance 8-PSK Trellis Codes

Number of States	$h^{(0)}$	$h^{(1)}$	$h^{(2)}$	d^2_{free}	$A_{d_{\text{free}}}$	$B_{d_{\text{free}}}$	Asymptotic Coding Gain (dB)
4	5	2	—	4.00[a]	1	1	3.0
8	11	2	4	4.59[a]	2	7	3.6
16	23	4	16	5.17[a]	2.25	11.5	4.1
32	45	16	34	5.76[a]	4	22.25	4.6
64	103	30	66	6.34[a]	5.25	31.125	5.0
128	277	54	122	6.59[a]	0.5	2.5	5.2
256	435	72	130	7.52[a]	1.5	12.25	5.8
512	1,525	462	360	7.52[a]	0.313	2.75	5.8
1,024	2,701	1,216	574	8.10[a]	1.32	10.563	6.1
2,048	4,041	1,212	330	8.34	3.875	21.25	6.2
4,096	15,201	6,306	4,112	8.68	1.406	11.758	6.4
8,192	20,201	12,746	304	8.68	0.617	2.711	6.4
32,768	143,373	70,002	47,674	9.51	0.25	2.5	6.8
131,072	616,273	340,602	237,374	9.85	—	—	6.9

[a] Codes found by exhaustive computer searches [13,33]; other codes (without the[a] superscript) were found by various heuristic search and construction methods [33,34]. The connector polynomials are in octal notation.

with moderate effort. Since the asymptotic coding gain is a reasonable yardstick at the bit error rates of interest, codes with a maximum of about 1000 states seem to exploit most of what can be gained by this type of coding.

Some other researchers have used different mapper functions in an effort to improve performance, in particular the bit error performance that could be improved by up to 0.5 dB. (For instance, 8-PSK Gray mapping [labeling the 8-PSK symbols successively by (000), (001), (011), (010), (110), (111), (101), (100)] was used by Du and Kasahara [31] and Zhang [32]. Zhang also used another mapper [labeling the symbols successively by (000), (001), (010), (011), (110), (111), (100), (101)] to further improve on the bit error multiplicity. The search criterion employed involved minimizing the bit multiplicities of several spectral lines in the distance spectrum of a code. Table 2 gives the best 8-PSK codes found so far with respect to the bit error probability.

If we go to higher-order signal sets such as 16-QAM, 32-cross, and 64-QAM, there are, at some point, not enough states left such that each diverging branch leads to a different state, and we have parallel transitions, that is, two or more branches connecting two states. Naturally we would want to assign signals with large distances to such parallel branches to avoid a high probability of error, since the probability of these errors cannot be influenced by the code.

The situation of parallel transition is actually the case for the first 8-PSK code in Table 1, whose trellis is given in Fig. 11. Here the parallel transitions are by choice, not by necessity. Note that the minimum-distance path pair through the trellis has $d^2 = 4.586$, but that is not the most likely error to happen. All signals on parallel branches are from a BPSK subset of the original 8-PSK set, and hence their distance is $d^2 = 4$, which gives the 3-dB asymptotic coding gain of the code over QPSK.

In general, we partition a signal set into a partition chain of subsets, such that the minimum distance between signal points in the new subsets is maximized at every level. This is illustrated in Fig. 12 with the 16-QAM signal set and a binary partition chain, which splits each set into two subsets at each level. Note that the partitioning can be continued until there is only one signal left in each subset. In such a way, by following the partition path, a "natural" binary label can be assigned to each signal point. This method of partitioning a signal set is called *set partitioning* with increasing intrasubset distances.

The idea is to use these constellations for codes with parallel transitions by not encoding all the input bits of u_r. Using the encoder in Fig. 13 with a 16-QAM constellation,

Table 2. Table of Improved 8-PSK Codes Using a Different Mapping Function [32]

Number of States	$h^{(0)}$	$h^{(1)}$	$h^{(2)}$	d^2_{free}	$A_{d_{\text{free}}}$	$B_{d_{\text{free}}}$
8	17	2	6	4.59	2	5
16	27	4	12	5.17	2.25	7.5
32	43	4	24	5.76	2.375	7.375
64	147	12	66	6.34	3.25	14.8755
128	277	54	176	6.59	0.5	2
256	435	72	142	7.52	1.5	7.813
512	1377	304	350	7.52	0.0313	0.25
1024	2077	630	1132	8.10	0.2813	1.688

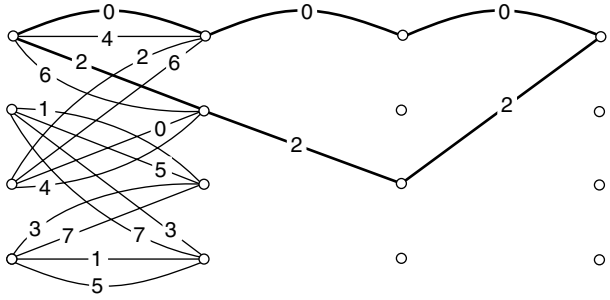


Figure 11. Four-state 8-PSK trellis code with parallel transitions.

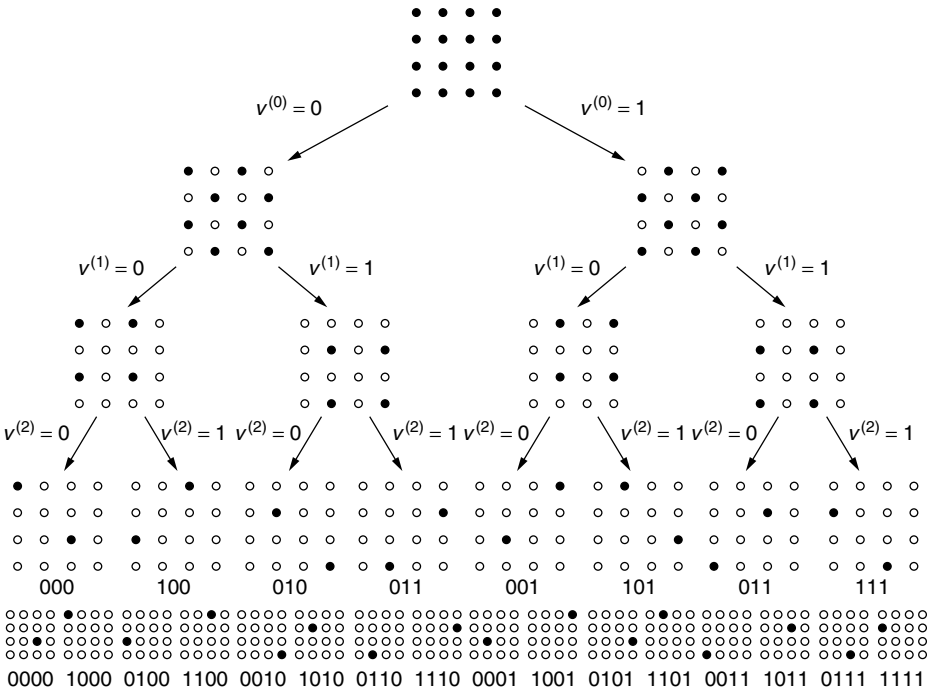


Figure 12. Set partitioning of a 16-QAM signal set into subsets with increasing minimum distance. The final partition level used by the encoder in Fig. 13 is the fourth level, that is the subsets with two signal points each.

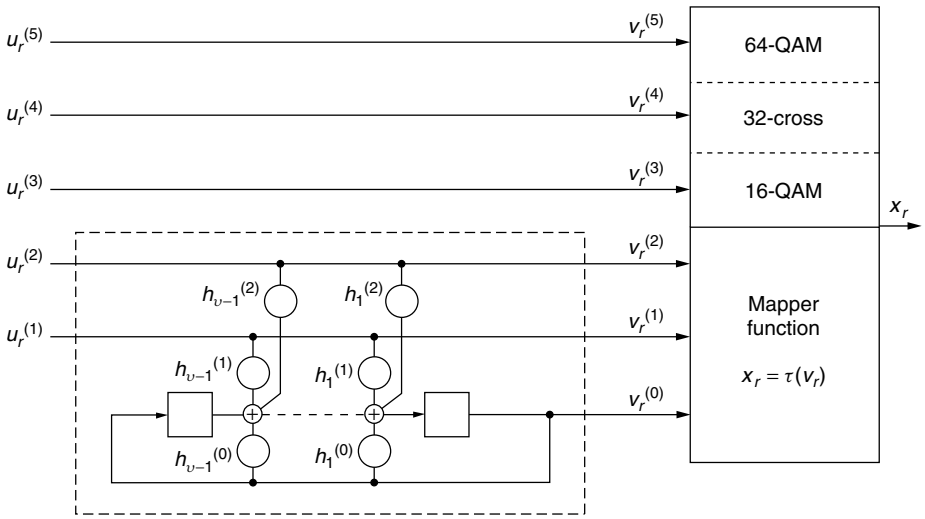


Figure 13. Generic encoder for QAM signal constellations.

for example, the first information bit $u_r^{(3)}$ is not encoded and the output signal of the FSM selects now a subset rather than a signal point. This subset is at the fourth partition level in Fig. 12. The uncoded bit(s) select the actual signal point within the subset. Analogously, then, the encoder now has to be designed to maximize the minimum interset distances of sequences, since it cannot influence the signal point selection within the subsets. The advantage of this strategy is that the same encoder can be used for all signal constellations with the same intraset distances at the final partition level, in particular for all signal constellations that are nested versions of each other, such as 16-QAM, 32-cross, and 64-QAM.

Figure 13 shows such a generic encoder that maximizes the minimum interset distance between sequences, and it can be used with all QAM-based signal constellations. Only the two least significant information bits affect the encoder FSM. All other information bits cause parallel transitions. Table 3 shows the coding gains achievable with such an encoder structure. The gains when going from 8-PSK to 16-QAM are most marked since rectangular constellations have a better power efficiency than do constant-energy circular constellations.

As in the case of 8-PSK codes, efforts have been made to improve the distance spectrum of a code, in particular to minimize the bit error multiplicity of the first few spectral lines. Some of the improved codes using a 16-QAM constellation are listed in Table 4 together with the original codes from Table 3, which are marked by "Ung." The improved codes, taken from Zhang [32], use the signal mapping shown in Fig. 14. Note also that the input line $u_r^{(3)}$ is also fed into the encoder for these codes.

Table 3. Connectors and Gains of Maximum Free Distance QAM Trellis Codes

Number of States	Connectors $h^{(0)}$	$h^{(1)}$	$h^{(2)}$	d^2_{free}	Asymptotic Coding Gain (dB) 16-QAM/ 8-PSK	32-cross/ 16-QAM	64-QAM/ 32-cross
4	5	2	—	4.0	4.4	3.0	2.8
8	11	2	4	5.0	5.3	4.0	3.8
16	23	4	16	6.0	6.1	4.8	4.6
32	41	6	10	6.0	6.1	4.8	4.6
64	101	16	64	7.0	6.8	5.4	5.2
128	203	14	42	8.0	7.4	6.0	5.8
256	401	56	304	8.0	7.4	6.0	5.8
512	1001	346	510	8.0	7.4	6.0	5.8

Source: The codes in this table were presented by Ungerböeck [13].

Table 4. Original 16-QAM Trellis Code and Improved Trellis Codes Using Nonstandard Mapping

2^{ν}	$h^{(0)}$	$h^{(1)}$	$h^{(2)}$	$h^{(3)}$	d^2_{free}	$A_{d_{\text{free}}}$	$B_{d_{\text{free}}}$
Ung 8	11	2	4	0	5.0	3.656	18.313
8	13	4	2	6	5.0	3.656	12.344
Ung 16	23	4	16	0	6.0	9.156	53.5
16	25	12	6	14	6.0	9.156	37.594
Ung 32	41	6	10	0	6.0	2.641	16.063
32	47	22	16	34	6.0	2	6
Ung 64	101	16	64	0	7.0	8.422	55.688
64	117	26	74	52	7.0	5.078	21.688
Ung 128	203	14	42	0	8.0	36.16	277.367
128	313	176	154	22	8.0	20.328	100.031
Ung 256	401	56	304	0	8.0	7.613	51.953
256	417	266	40	226	8.0	3.273	16.391

5. CONVOLUTIONAL CODES

Convolutional codes historically were the first trellis codes. They were introduced in 1955 by Elias [54]. Since then, much theory has evolved to understand convolutional codes. A convolutional code is obtained by using a special modulator. The output binary digits of the encoder, shown again in Fig. 15, are no longer jointly encoded into a modulation symbol, but each bit is encoded into a binary BPSK signal, using the mapping $0 \to -1, 1 \to 1$. It is immediately clear now that the input and output symbol rates are different. For example, in Fig. 15, this rate changes from 2 input bits/time unit to 3 output symbols/time unit. This causes a bandwidth expansion of $\frac{3}{2}$ with respect to uncoded BPSK modulation. It is this bandwidth expansion that is partly responsible for the excellent performance of convolutional codes.

It is interesting to note that the convolutional encoder in Fig. 15 has an alternate "incarnation," which is given in Fig. 16. This form is called the *controller canonical nonsystematic form*; the term stems from the fact that inputs can be used to control the state of the encoder in a very direct way, in which the outputs have no influence. Both encoders generate the same code, but individual input bit sequences map onto different output bit sequences.

The squared Euclidean distance between two signal sequences depends only on the *Hamming distance* $H_d(\underline{v}^{(1)}, \underline{v}^{(2)})$ between the two output symbol sequences, where the Hamming distance between two sequences is defined as the number of bit positions in which the two sequences differ. Consequently, the minimum squared Euclidean distance d^2_{free} depends only on the number of binary differences between the closest code sequences. This number is the *minimum Hamming distance* of a convolutional code, denoted by d_{free}. Convolutional codes are linear in the sense that the modulo-2 addition of

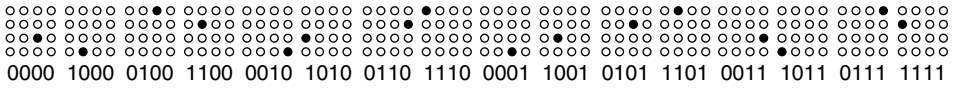


Figure 14. Mapping used for the improved 16-QAM codes.

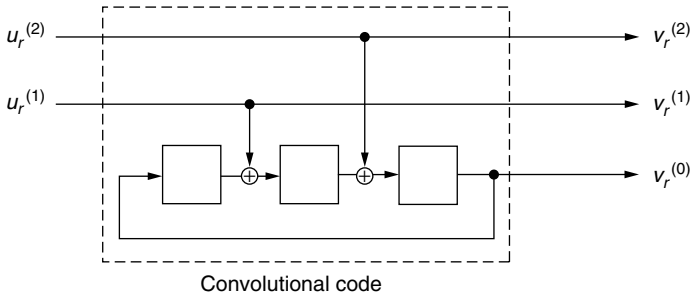


Figure 15. Rate $R = \frac{2}{3}$ convolutional code which was used in Fig. 7 to generate an 8-PSK trellis code.

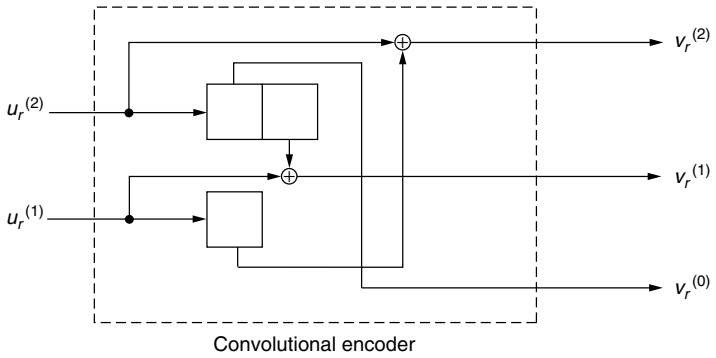


Figure 16. Rate $R = \frac{2}{3}$ convolutional code from above in controller canonical non-systematic form.

two output bit sequences is another valid code sequence, and finding the minimum Hamming distance between two sequences $\underline{v}^{(1)}$ and $\underline{v}^{(2)}$ amounts to finding the minimum Hamming weight of any code sequence $\underline{v}$.

Finding convolutional codes with large minimum Hamming distance is exactly as difficult as finding good general trellis codes with large Euclidean free distance, and, as in the case for trellis codes, computer searches are usually used to find good codes [58–60]. Most often the controller canonical form (Fig. 16) of an encoder is preferred in these searches. The procedure is then to search for a code with the largest minimum Hamming weight by varying the taps either exhaustively or according to heuristic rules. In this fashion, the codes in Tables 5–12 were found [15–17,59]. They are the rate $R = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}$, and $R = \frac{2}{3}$ codes with the greatest minimum Hamming distance d_{free} for a given constraint length.

6. DECODING

6.1. Sequence Decoding

There are basically two major decoding strategies in common use. These are the sequence decoders and the symbol decoders. A sequence decoder looks for the most likely transmitted code sequence. Thus, if $\underline{y}$ is the received sequence from the channel, an optimal sequence decoder calculates the conditional probability

$$\hat{\underline{x}} = \max_{\underline{x}} \Pr(\underline{x}|\underline{y}) = \max_{\underline{x}} \Pr(\underline{y}|\underline{x}), \tag{8}$$

assuming that all $\underline{x}$ are a priori equally likely to have been sent. Such a decoder is referred to as a *maximum-likelihood* (ML) *sequence decoder*. For a trellis code, this amounts to an algorithm which exhaustively

Table 5. Connectors[a] and Free Hamming Distance of the Best $R = \frac{1}{2}$ Convolutional Codes [16]

Constraint Length ν	$g^{(1)}$	$g^{(0)}$	d_{free}
2	5	7	5
3	15	17	6
4	23	35	7
5	65	57	8
6	133	171	10
7	345	237	10
8	561	753	12
9	1161	1545	12
10	2335	3661	14
11	4335	5723	15
12	10533	17661	16
13	21675	27123	16
14	56721	61713	18
15	111653	145665	19
16	347241	246277	20

[a]The connectors are given in octal notation (e.g., $g = 17 = 1111$).

Table 6. Connectors and Free Hamming Distance of the Best $R = \frac{1}{3}$ Convolutional Codes [16]

Constraint Length ν	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	5	7	7	8
3	13	15	17	10
4	25	33	37	12
5	47	53	75	13
6	133	145	175	15
7	225	331	367	16
8	557	663	711	18
9	1,117	1,365	1,633	20
10	2,353	2,671	3,175	22
11	4,767	5,723	6,265	24
12	10,533	10,675	17,661	24
13	21,645	35,661	37,133	26

Table 7. Connectors and Free Hamming Distance of the Best $R = \frac{2}{3}$ Convolutional Codes [16]

Constraint Length ν	$g_2^{(2)}, g_1^{(2)}$	$g_2^{(1)}, g_1^{(1)}$	$g_2^{(0)}, g_1^{(0)}$	d_{free}
2	3,1	1,2	3,2	3
3	2,1	1,4	3,7	4
4	7,2	1,5	4,7	5
5	14,3	6,10	16,17	6
6	15,6	6,15	15,17	7
7	14,3	7,11	13,17	8
8	32,13	5,33	25,22	8
9	25,5	3,70	36,53	9
10	63,32	15,65	46,61	10

Table 8. Connectors and Free Hamming Distance of the Best $R = \frac{1}{4}$ Convolutional Codes [58]

Constraint Length ν	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	5	7	7	7	10
3	13	15	15	17	13
4	25	27	33	37	16
5	53	67	71	75	18
6	135	135	147	163	20
7	235	275	313	357	22
8	463	535	733	745	24
9	1,117	1,365	1,633	1,653	27
10	2,387	2,353	2,671	3,175	29
11	4,767	5,723	6,265	7,455	32
12	11,145	12,477	15,537	16,727	33
13	21,113	23,175	35,527	35,537	36

Table 9. Connectors and Free Hamming Distance of the Best $R = \frac{1}{5}$ Convolutional Codes [15]

Constraint Length ν	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	7	7	7	5	5	13
3	17	17	13	15	15	16
4	37	27	33	25	35	20
5	75	71	73	65	57	22
6	175	131	135	135	147	25
7	257	233	323	271	357	28

Table 10. Connectors and Free Hamming Distance of the Best $R = \frac{1}{6}$ Convolutional Codes [15]

Constraint Length ν	$g^{(5)}$	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	7	7	7	7	5	5	16
3	17	17	13	13	15	15	20
4	37	35	27	33	25	35	24
5	73	75	55	65	47	57	27
6	173	151	135	135	163	137	30
7	253	375	331	235	313	357	34

searches through all the possible paths through the trellis. However, some simplifications are possible.

The algorithms will start in the known state at the beginning of the trellis, and explore all possible paths one step at a time. It will keep a measure of reliability at each state and time, called the *state metric*. This state metric for state $s_n = i$ at time n is calculated as the partial path probability $\Pr(\underline{\tilde{y}}|\underline{\tilde{x}})$ for the partial path $\underline{\tilde{x}}(i) = (\ldots, x_0, x_1, \ldots, x_n)$ that leads to state $s_n = i$, given the partial received sequence $\underline{\tilde{y}} = (\ldots, y_0, y_1, \ldots, y_n)$ up to time n.

This probability can be expressed recursively, and the metric at state $s_n = i$ and time n, denoted by $J_n(i)$, can be calculated from previous state metrics according to

$$J_n(i) = J_{n-1}(j) + \lambda_n(j \to i) \tag{9}$$

In other words, the metric at state i at time n is calculated from the metric at state j at time $n-1$ by adding a measure $\lambda_n(j \to i)$, called the *branch metric*, that depends on the symbol on the transition from $j \to i$ and the received signal y_n. State j must connect to state i in the trellis of the code. However, state j is not usually the only state that connects to state i, and the algorithm will select only the best metric among all merging connections. This is known as the *add–compare–select* step in the algorithm and is formally given by

$$J_n(i) = min_{j \to i}(J_{n-1}(j) + \lambda_n(j \to i)) \tag{10}$$

The algorithm furthermore needs to remember the history of selection decisions, since the winning sequence can be determined only at the end of the received sequence. This method was introduced by Viterbi in 1967 [18,19] in the context of analyzing convolutional codes, and has since become widely known as the *Viterbi algorithm* [20]. The algorithm for a block of length L is as follows:

Step 1. Initialize the S states of the maximum-likelihood decoder with the metric $J_0(i) = -\infty$ and survivors $\underline{\tilde{x}}(i) = \{\ \}$. Initialize the starting state of the decoder, usually state $i = 0$, with the metric $J_0(0) = 0$. Let $n = 1$.

Step 2. Calculate the branch metric

$$\lambda_n = |y_n - x_n(j \to i)|^2 \tag{11}$$

for each state j and each signal $x_n(j \to i)$ that is attached to the transition from state i to state j.

Step 3. For each state i, choose from the 2^k merging paths the survivor $\underline{\tilde{x}}(i)$ for which $J_n(i)$ is maximized.

Table 11. Connectors and Free Hamming Distance of the Best $R = \frac{1}{7}$ Convolutional Codes [15]

Constraint Length ν	$g^{(6)}$	$g^{(5)}$	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	7	7	7	7	5	5	5	18
3	17	17	13	13	13	15	15	23
4	35	27	25	27	33	35	37	28
5	53	75	65	75	47	67	57	32
6	165	145	173	135	135	147	137	36
7	275	253	375	331	235	313	357	40

Table 12. Connectors and Free Hamming Distance of the Best $R = \frac{1}{8}$ Convolutional Codes [15]

Constraint Length ν	$g^{(7)}$	$g^{(6)}$	$g^{(5)}$	$g^{(4)}$	$g^{(3)}$	$g^{(2)}$	$g^{(1)}$	$g^{(0)}$	d_{free}
2	7	7	5	5	5	7	7	7	21
3	17	17	13	13	13	15	15	17	26
4	37	33	25	25	35	33	27	37	32
5	57	73	51	65	75	47	67	57	36
6	153	111	165	173	135	135	147	137	40
7	275	275	253	371	331	235	313	357	45

Step 4. If $n < L$, let $n = n + 1$ and go to step 2, or else go to step 5.

Step 5. Output the survivor $\underline{x}(i)$ that maximizes $J_L(i)$ as the maximum-likelihood estimate of the transmitted sequence.

The Viterbi algorithm has enjoyed tremendous popularity, not only in decoding trellis codes but also in symbol sequence estimation over channels affected by intersymbol interference [17,21] and multiuser optimal detectors [22]. Whenever the underlying generating process can be modeled as a finite-state machine, the Viterbi algorithm finds application.

A rather large body of literature deals with the Viterbi decoder, and there are a number of books dealing with the subject [e.g., 16,17,23,24]. One of the more important results is that it can be shown that there is no need to wait until the entire sequence is decoded before starting to output the estimated symbols $\tilde{x}_n$, or the corresponding data. The probability that the symbols in all survivors $\underline{\tilde{x}}(i)$ are identical for $m \leq n - n_t$ is very close to unity for $n_t \approx 5\nu$. n_t is called the *truncation length* or *decision depth*. We may therefore modify the algorithm to obtain a fixed-delay decoder by modifying steps 4 and 5 of the algorithm outlined above as follows:

Step 4. If $n \geq n_t$, output $x_{n-n_t}(i)$ from the survivor $\underline{\tilde{x}}(i)$ with the largest metric $J_n(i)$ as the estimated symbol at time $n - n_t$. If $n < L$, let $n = n + 1$ and go to step 2.

Step 5. Output the remaining estimated symbols $x_n(i)$; $L - n_t < n \leq L$ from the survivor $\underline{x}(i)$ that maximizes $J_L(i)$.

We recognize that we may now let $L \to \infty$; thus, the complexity of our decoder is not determined by the length of the sequence, and it may be operated in a continuous fashion.

6.2. Symbol Decoding

The other important class of decoders targets symbols, rather than sequences. The goal is to calculate the a posteriori probability (APP)

$$\Pr(u_r|\underline{y}) \tag{12}$$

that is, the probability of u_r after observing $\underline{y}$. This value can be used to find the maximum APP (MAP) decision of u_r as

$$\hat{u}_r = \max_{u_r} \Pr(u_r|\underline{y}) \tag{13}$$

However, as a symbol estimator, the resulting error probability is not significantly better than that achieved with the Viterbi decoder (8). The soft value of (12) is used mostly in iterative decoding such as Turbo decoding, where soft output values from component decoders are required.

The APP of u_r is calculated by the *backward–forward algorithm*, a procedure that sweeps through the trellis in both the forward and the backward directions. The algorithm functions as follows. First the probability of the transition from state j to i, given $\underline{y}$ is calculated. It can be broken into three factors, given by

$$\Pr(s_{r-1} = j, s_r = i, \underline{y}) = \alpha_{r-1}(j)\gamma_r(j \to i)\beta_r(i) \tag{14}$$

where the α values are the result of the forward pass and are calculated recursively according to

$$\alpha_r(j) = \sum_{\text{states } l} \alpha_{r-1}(l)\gamma_r(l \to j) \tag{15}$$

Furthermore, for a trellis code started in the zero state at time $r = 0$ we have the starting conditions

$$\alpha_0(0) = 1, \alpha_0(j) = 0; \quad j \neq 0 \tag{16}$$

Similarly

$$\beta_r(i) = \sum_{\text{states } l} \beta_{r+1}(l)\gamma_{r+1}(i \to l) \tag{17}$$

The boundary condition for $\beta_r(i)$ for a code that is terminated in the zero state at time $r = L$ is

$$\beta_L(0) = 1, \beta_L(i) = 0; \quad i \neq 0 \tag{18}$$

The values $\gamma_r(j \to i)$ associated with the transition from state j to state i are external values and are calculated as

$$\gamma_r(j \to i) = \sum_{x_r} p_{ij} q_{ij}(x_r) p_n(y_r - x_r) \tag{19}$$

where $p_n(\cdot)$ is the probability density function of the AWGN (additive white Gaussian noise) channel given by $\Pr(y_r|x_r) = p_n(y_r - x_r)$. p_{ij} is the a priori probability of the state transitions, usually equal for all transitions, and $q_{ij}(x_r)$ is the probability of choosing x_r if there is more than one signal choice per transition. The calculation of $\gamma_r(j \to i)$ is not very complex and can most easily be implemented by a table lookup procedure.

Equations (15) and (17) are iterative updates of internal variables. The complete algorithm to calculate the a posteriori state-transition probabilities is as follows:

Step 1. Initialize $\alpha_0(0) = 1, \alpha_0(j) = 0$ for all non-zero states ($j \neq 0$), and $\beta_L(0) = 1, \beta_L(j) = 0, j \neq 0$. Let $r = 1$.

Step 2. Calculate $\gamma_r(j \to i)$ using (19), and $\alpha_r(j)$ using (15) for all states j.

Step 3. If $r < L$, let $r = r + 1$ and go to step 2, else $r = L - 1$ and go to step 4.

Step 4. Calculate $\beta_r(i)$ using (17), and $\Pr(s_{r-1} = j, s_r = i, \underline{y})$ from (14).

Step 5. If $r > 1$, let $r = r - 1$ and go to step 4.

Step 6. Terminate the algorithm and output all the values and $\Pr(s_{r-1} = j, s_r = i, \underline{y})$.

Contrary to the ML algorithm, the MAP algorithm needs to go through the trellis twice, once in the forward direction, and once in the reverse direction. What weighs even more is that all the values $\alpha_r(j)$ must be stored from the first pass through the trellis. For a rate k/n convolutional code, for example, this requires $2^{k\nu}L$ storage locations since there are $2^{k\nu}$ states, for each of which we need to store a different value $\alpha_r(j)$ at each time epoch r. The storage requirement grows exponentially in the constraint length ν and linearly in the block length L.

The a posteriori transition probabilities produced by this algorithm can now be used to calculate a posteriori information bit probabilities, that is, the probability that the information k-tuple $u_r = u$. Starting from the transition probabilities $\Pr(s_{r-1} = j, s_r = i|\underline{y})$ we simply sum over all transitions $j \to i$ that are caused by $u_r = u$. Denoting these transitions by $A(u)$, we obtain

$$\Pr(u_r = u) = \sum_{(j \to i) \in A(u)} \Pr(s_{r-1} = j, s_r = i|\underline{y}) \tag{20}$$

Another most interesting product of the MAP decoder is the a posteriori probability of the transmitted output symbol x_r. Arguing analogously as above, and letting $B(x)$ be the set of transitions on which the output signal x can occur, we obtain

$$\Pr(x_r = x) = \sum_{(j \to i) \in B(x)} \Pr(x|y_r)\Pr(s_{r-1} = j, s_r = i|\underline{y})$$

$$= \sum_{(j \to i) \in B(x)} \frac{p_n(y_r - x_r)}{p(y_r)} q_{ij}(x)$$

$$\Pr(s_r = i, s_{r+1} = j|\underline{y}) \tag{21}$$

where the a priori probability of y_r can be calculated via

$$p(y_r) = \sum_{x'} p(y_r|x')q_{ij}(x') \tag{22}$$

Equation (21) can be greatly simplified if there is only one output symbol on the transition $j \to i$. In this case, the transition automatically determines the output symbol, and

$$\Pr(x_r = x) = \sum_{(j \to i) \in B(x)} \Pr(s_{r-1} = j, s_r = i|\underline{y}) \tag{23}$$

7. PARALLEL CONCATENATED TRELLIS CODES (TURBO CODES)

7.1. Code Structure

A Turbo encoder consists of the *parallel concatenation* of two or more, usually identical, rate $\frac{1}{2}$ encoders, realized in systematic feedback form, and a pseudorandom interleaver. This encoder structure is called a parallel concatenation because the two encoders operate on the same *set* of input bits, rather than one encoding the output of the other as in serial concatenation. A block diagram of a Turbo encoder with two constituent convolutional encoders is shown in Fig. 17.

The interleaver is used to permute the input bits such that the two encoders are operating on the same *set* of input bits, but different input *sequences*. Thus, the first encoder receives the input bit u_r and produces the output pair $(u_r, v_r^{(1)})$, while the second encoder receives the input bit u_r' and produces the output pair $(u_r', v_r^{(2)})$. The input bits are grouped into finite-length sequences whose length, N, equals the size of the interleaver. Since both the encoders are systematic and operate on the same set of input bits, it is only necessary to transmit the input bits once and the overall code has rate $\frac{1}{3}$. In order to increase the overall rate of the code to $\frac{1}{2}$, the two parity sequences $\underline{v}^{(1)}$ and $\underline{v}^{(2)}$ can be punctured by alternately deleting $v_r^{(1)}$ and $v_r^{(2)}$. We will refer to a Turbo code whose constituent encoders have parity-check polynomials h_0 and h_1, expressed in

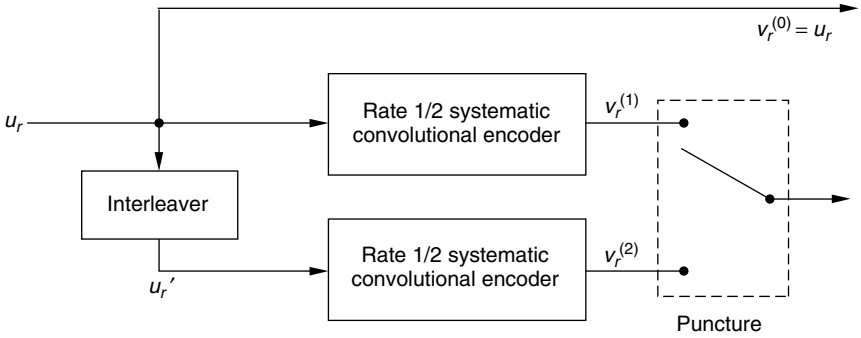


Figure 17. Block diagram of a Turbo encoder with two constituent encoders and an optional puncturer.

octal notation, and whose interleaver is of length N as an (h_0, h_1, N) Turbo code.

Several salient points concerning the structure of the codewords in a Turbo code are

1. Because the pseudorandom interleaver permutes the input bits, the two input sequences $\underline{u}$ and $\underline{u}'$ are almost always different, although of the same weight, and the two encoders will (with high probability) produce parity sequences of different weights.
2. It is easily seen that a codeword may consist of a number of distinct detours in each encoder. Note that since the constituent encoders are realized in systematic feedback form, a nonzero bit is required to return to the all-zero state and thus all detours are associated with information sequences of weight 2 or greater. Finally, with a pseudorandom interleaver it is highly unlikely that both encoders will be returned to the all-zero state at the end of the codeword even when the last ν bits of the input sequence $\underline{u}$ are used to force the first encoder back to the all zero state.

If neither encoder is forced to the all-zero state, that is, no tail is used, then the sequence consisting of $N-1$ zeros followed by a one is a valid input sequence $\underline{u}$ to the first encoder. For some interleavers, this $\underline{u}$ will be permuted to itself and $\underline{u}'$ will be the same sequence. In this case, the maximum weight of the codeword with puncturing, and thus the free distance of the code, will be 2! For this reason, it is common to assume that the first encoder is forced to return to the all zero state. The ambiguity of the final state of the second encoder results in negligible performance degradation for large interleavers.

7.2. Iterative Decoding of Turbo Codes

It is clear from the discussion of the codeword structure of Turbo codes that the state space of these codes is too large to perform optimum decoding. To overcome this, the discoverers of Turbo codes proposed a novel iterative decoder based on the a posteriori (AAP) symbol decoding algorithm.

The AAP decoder for each component code computes the a posteriori probability $\Pr(u_r = u|\underline{y})$ conditioned on the received sequence $\underline{y}$. The iterative Turbo decoder makes use of these a posteriori probabilities in the form of a log-likelihood ratio (LLR) given by

$$L(u_r) = \log \frac{\Pr(u_r = 1|\underline{y})}{\Pr(u_r = 0|\underline{y})} \tag{24}$$

which, from (14) and (20), is given by

$$L(u_r) = \log \frac{\sum_{(j\to i)\in A(u_r=1)} \gamma_r(j \to i)\alpha_{r-1}(j)\beta_r(i)}{\sum_{(j\to i)\in A(u_r=0)} \gamma_r(j \to i)\alpha_{r-1}(j)\beta_r(i)}. \tag{25}$$

Let $y_r^{(0)}$ be the received systematic bit and $y_r^{(m)}$, $m = 1, 2$, the received parity bit corresponding to the mth constituent encoder. With these, $\gamma_r(j \to i)$ may be expressed as

$$\gamma_r(j \to i) = p_{ij}\Pr(y_r^{(0)}, y_r^{(m)}|u_r, v_r^{(m)}) \tag{26}$$

For systematic codes, (26) may be factored as

$$\Pr(y_r^{(0)}, y_r^{(m)}|u_r, v_r^{(m)}) = \Pr(y_r^{(0)}|u_r)\Pr(y_r^{(m)}|v_r^{(m)}) \tag{27}$$

since the received systematic sequence and the received parity sequence are conditionally independent of each other. Finally, substituting (26) and (27) into (25) and factoring yields

$$\begin{aligned} L(u_r) = &\log \frac{\sum_{(j\to i)\in A(u_r=1)} \Pr(y_r^{(m)}|v_r^{(m)})\alpha_{r-1}(j)\beta_r(i)}{\sum_{(j\to i)\in A(u_r=0)} \Pr(y_r^{(m)}|v_r^{(m)})\alpha_{r-1}(j)\beta_r(i)} \\ &+ \log \frac{\Pr(u_r = 1)}{\Pr(u_r = 0)} \\ &+ \log \frac{\Pr(y_r^{(0)}|u_r = 1)}{\Pr(y_r^{(0)}|u_r = 0)} = \Lambda_{e,r}^{(m)} + \Lambda_r + \Lambda_s \end{aligned} \tag{28}$$

where $\Lambda_{e,r}^{(m)}$ is called the *extrinsic information* from the mth decoder, Λ_r is the a priori log-likelihood ratio of the systematic bit u_r and Λ_s is the log-likelihood ratio of the a posteriori probabilities of the systematic bit.

A block diagram of an iterative Turbo decoder is shown in Fig 18, where each APP decoder corresponds to a constituent code. The interleavers are identical to

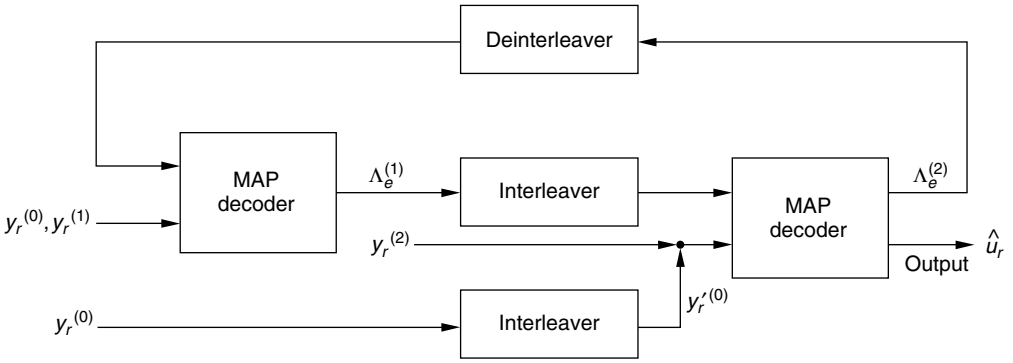


Figure 18. Block diagram of a Turbo decoder with two constituent decoders.

the interleavers in the Turbo encoder and are used to reorder the sequences so that each decoder is properly synchronized. For the first iteration, the first decoder computes the log-likelihood ratio of Eq. (28) with $\Lambda_r = 0$, since u_r is equally likely to be a 0 or a 1, using the received sequences $\underline{y}^{(0)}$ and $\underline{y}^{(1)}$. The second decoder now computes the log-likelihood ratio of Eq. (28) on the basis of the received sequences $\underline{y}^{(0)}$ and $\underline{y}^{(2)}$ (suitably reordered).

The second decoder, however, has available an estimate of the a posteriori probability of u_r from the first decoder, namely, $L^{(1)}(u_r)$. The second decoder may consider this the a priori probability of u_r in (28) and compute

$$\begin{aligned} L^{(2)}(u_r) &= \Lambda_{e,r}^{(2)} + L^{(1)} + \Lambda_s \\ &= \Lambda_{e,r}^{(2)} + \Lambda_{e,r}^{(1)} + \Lambda_r^{(1)} + \Lambda_s + \Lambda_s \end{aligned} \tag{29}$$

as the new LLR. Close examination of (29) reveals that by passing $L^{(1)}$, the second decoder is given $\Lambda_r^{(1)}$, the previous estimate of the a priori probability, which is unnecessary. In addition, it is seen that as the decoder continues to iterate, the LLR accumulates Λ_s and the systematic bit becomes overemphasized. In order to prevent this, the second decoder subtracts $\Lambda_r^{(1)}$ and Λ_s from the information passed from the first decoder and calculates

$$L^{(2)}(u_r) = \Lambda_{e,r}^{(2)} + \Lambda_{e,r}^{(1)} + \Lambda_s \tag{30}$$

What is passed between the two decoders is in fact the extrinsic information only. This process continues until a desired performance is achieved, at which point a final decision is made by comparing the final log-likelihood ratio to the threshold 0.

The extrinsic information is a reliability measure of each component decoder's estimate of the transmitted sequence on the basis of the corresponding received component parity sequence and is essentially independent of the received systematic sequence. Since each component decoder uses the received systematic sequence directly, the extrinsic information allows the decoders to share information without significant error propagation. The efficacy of this technique can be seen in Fig 19, which shows the performance of the original (37,21,65536) Turbo code as a function of the decoder iterations. It is impressive that the performance of the code with iterative decoding continues to improve up to 18 iterations (and beyond).

BIOGRAPHY

Christian Schlegel received the Dipl. El. Ing. ETH degree from the Federal Institute of Technology, Zürich, Switzerland, in 1984, and his M.S. and Ph.D. degrees in electrical engineering from the University of Notre Dame, Indiana, in 1986 and 1989. From 1988 to 1992 he was with Asea Brown Boveri, Ltd., Baden, Switzerland, from 1992–1994 with the Digital Communications Group at the University of South Australia in Adelaide, from 1994–2001 he held faculty positions at the Universities of Texas at San Antonio and Utah in Salt Lake City. In 2001 he was named iCORE Professor for High-Capacity Digital Communications at the University of Alberta, Canada.

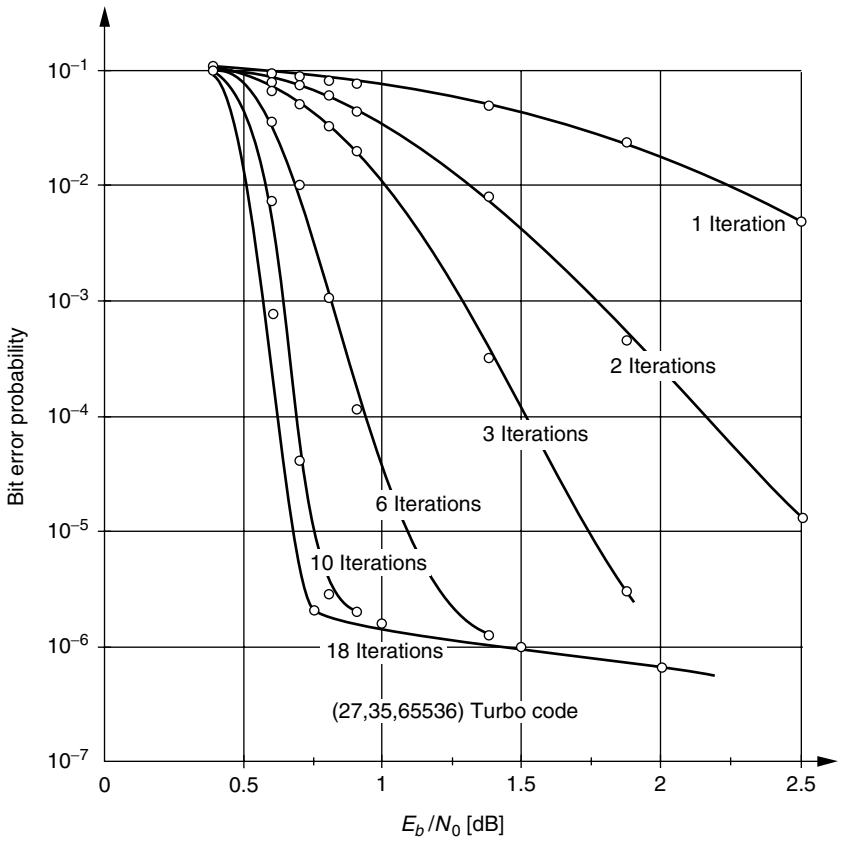


Figure 19. Performance of the (37,21,65536) Turbo code as function of the number of decoder iterations.

His interests are in the area of digital communications and mobile radio systems, error control coding, multiple access communications, and analog and digital system implementations. He is the author of *Trellis Coding* (IEEE 1997) and *Trellis and Turbo Coding* (Wiley/IEEE 2002). Dr. Schlegel received an NSF 1997 Career Award and a Canada Research Chair in 2001.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communications, *Bell Syst. Tech. J.* **27**: 379–423 (July 1948).
2. R. W. Hamming, Error detecting and error correcting codes, *Bell Syst. Tech. J.* **29**: 147–160 (1950).
3. M. J. E. Golay, Notes on digital coding, *Proc. IEEE* **37**: 657 (1949).
4. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, New York, 1988.
5. C. Schlegel, *Trellis Coding*, IEEE Press, Piscataway, NJ, 1997.
6. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. 1993 IEEE Int. Conf. Communication*, Geneva, Switzerland, 1993, pp. 1064–1070.
7. F.-Q. Wang and D. J. Costello, Probabilistic construction of large constraint length trellis codes for sequential decoding, *IEEE Trans. Commun.* **COM-43**: 2439–2448 (September 1995).
8. J. A. Heller and J. M. Jacobs, Viterbi detection for satellite and space communication, *IEEE Trans. Commun. Technol.* **COM-19**: 835–848 (Oct. 1971).
9. J. K. Omura and B. K. Levitt, Coded error probability evaluation for antijam communication systems, *IEEE Trans. Commun.* **COM-30**: 896–903 (May 1982).
10. G. Ungerböeck, J. Hagenauer, and T. Abdel-Nabi, Coded 8-PSK experimental modem for the INTELSAT SCPC system, *Proc. ICDSC, 7th*, 1986, pp. 299–304.
11. G. Ungerböeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**(1): 55–67 (Jan. 1982).
12. G. Ungerböeck, Trellis-coded modulation with redundant signal sets part I: Introduction, *IEEE Commun. Mag.* **25**(2): 5–11 (Feb. 1987).
13. G. Ungerböeck, Trellis-coded modulation with redundant signal sets part II: State of the art, *IEEE Commun. Mag.* **25**(2): 12–21 (Feb. 1987).
14. H. Imai et al., *Essentials of Error-Control Coding Techniques*, Academic Press, New York, 1990.
15. D. G. Daut, J. W. Modestino, and L. D. Wismer, New short constraint length convolutional code construction for selected rational rates, *IEEE Trans. Inform. Theory* **IT-28**: 793–799 (Sept. 1982).
16. S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
17. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
18. A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inform. Theory* **IT-13**: 260–269 (April 1969).
19. J. K. Omura, On the Viterbi decoding algorithm, *IEEE Trans. Inform. Theory* **IT-15**: 177–179 (Jan. 1969).
20. G. D. Forney, Jr., The Viterbi algorithm, *Proc. IEEE* **61**: 268–278 (1973).
21. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363–378 (May 1972).
22. S. Verdú, Minimum probability of error for asynchronous Gaussian multiple-access channels, *IEEE Trans. Inform. Theory* **IT-32**: 85–96 (Jan. 1986).
23. R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, Reading, MA, 1987.
24. A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill, New York, 1979.
25. H. Nyquist, Certain topics in telegraph transmission theory, *AIEE Trans.* 617 ff. (1946).
26. J. L. Massey and D. J. Costello, Jr., Nonsystematic convolutional codes for sequential decoding in space applications, *IEEE Trans. Commun. Technol.* **COM-19**: 806–813 (1971).
27. O. M. Collins, The subtleties and intricacies of building a constraint length 15 convolutional decoder, *IEEE Trans. Commun.* **COM-40**: 1810–1819 (1992).
28. G. D. Forney, Jr., Coded modulation for bandlimited channels, *IEEE Inform. Theory Soc. Newsl.* (Dec. 1990).
29. CCITT Recommendations V.34.
30. U. Black, *The V Series Recommendations, Protocols for Data Communications over the Telephone Network*, McGraw-Hill, New York, 1991.
31. J. Du and M. Kasahara, Improvements of the information-bit error rate of trellis code modulation systems, *IEICE, Japan* **E72**: 609–614 (May 1989).
32. W. Zhang, *Finite-State Machines in Communications*, Ph.D. thesis, Univ. South Australia, Australia, 1995.
33. J. E. Porath and T. Aulin, *Fast Algorithmic Construction of Mostly Optimal Trellis Codes*, Technical Report 5, Division of Information Theory, School of Electrical and Computer Engineering, Chalmers Univ. Technology, Göteborg, Sweden, 1987.
34. J. E. Porath and T. Aulin, Algorithmic construction of trellis codes, *IEEE Trans. Commun.* **COM-41**(5): 649–654 (May 1993).
35. J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, New York, 1988.
36. G. D. Forney, Jr. et al., Efficient modulation for band-limited channels, *IEEE J. Select. Areas Commun.* **SAC-2**(5): 632–647 (1984).
37. L. F. Wei, Trellis-coded modulation with multidimensional constellations, *IEEE Trans. Inform. Theory* **IT-33**: 483–501 (1987).
38. A. R. Calderbank and N. J. A. Sloane, An eight-dimensional trellis code, *Proc. IEEE* **74**: 757–759 (1986).
39. L. F. Wei, Rotationally invariant trellis-coded modulations with multidimensional M-PSK, *IEEE J. Select. Areas Commun.* **SAC-7**(9): 1281–1295 (Dec. 1989).
40. A. R. Calderbank and N. J. A. Sloane, New trellis codes based on lattices and cosets, *IEEE Trans. Inform. Theory* **IT-33**: 177–195 (1987).

41. S. S. Pietrobon, G. Ungerböeck, L. C. Perez, and D. J. Costello, Jr., Rotationally invariant nonlinear trellis codes for two-dimensional modulation, *IEEE Trans. Inform. Theory* **IT-40**(6): 1773–1791 (Nov. 1994).

42. S. S. Pietrobon et al., Trellis-coded multidimensional phase modulation, *IEEE Trans. Inform. Theory* **IT-36**: 63–89 (Jan. 1990).

43. IBM Europe, Trellis-coded modulation schemes for use in data modems transmitting 3–7 bits per modulation interval, CCITT SG XVII Contribution COM XVII, No. D114, April 1983.

44. L. F. Wei, Rotationally invariant convolutional channel coding with expanded signal space — Part I: 180 degrees, *IEEE J. Select. Areas Commun.* **SAC-2**: 659–672 (Sept. 1984).

45. L. F. Wei, Rotationally invariant convolutional channel coding with expanded signal space — Part II: Nonlinear codes, *IEEE J. Select. Areas Commun.* **SAC-2**: 672–686 (Sept. 1984).

46. IBM Europe, Trellis-coded modulation schemes with 8-state systematic encoder and 90° symmetry for use in data modems transmitting 3–7 bits per modulation interval, CCITT SG XVII Contribution COM XVII, No. D180, Oct. 1983.

47. S. S. Pietrobon and D. J. Costello, Jr., Trellis coding with multidimensional QAM signal sets, *IEEE Trans. Inform. Theory* **IT-39**: 325–336 (March 1993).

48. M. V. Eyuboglu, G. D. Forney, P. Dong, and G. Long, Advanced modem techniques for V. Fast, *Eur. Trans. Telecommun.* **ETT-4**(3): 234–256 (May–June 1993).

49. G. D. Forney, Jr., Coset codes — Part I: Introduction and geometrical classification, *IEEE Trans. Inform. Theory* **IT-34**: 1123–1151 (1988).

50. G. D. Forney, Jr., Coset codes — Part II: Binary lattices and related codes, *IEEE Trans. Inform. Theory* **IT-34**: 1152–1187 (1988).

51. G. D. Forney, Jr., Geometrically uniform codes, *IEEE Trans. Inform. Theory* **IT-37**: 1241–1260 (1991).

52. D. Slepian, On neighbor distances and symmetry in group codes, *IEEE Trans. Inform. Theory* **IT-17**: 630–632 (Sept. 1971).

53. J. L. Massey, T. Mittelholzer, T. Riedel, and M. Vollenweider, Ring convolutional codes for phase modulation, *IEEE Int. Symp. Inform. Theory*, San Diego, CA, Jan. 1990.

54. P. Elias, Coding for noisy channels, *IRE Conv. Rec.* (Pt. 4): 37–47 (1955).

55. R. Johannesson and Z.-X. Wan, A linear algebra approach to minimal convolutional encoders, *IEEE Trans. Inform. Theory* **IT-39**(4): 1219–1233 (July 1993).

56. G. D. Forney, Jr., Convolutional codes I: Algebraic structure, *IEEE Trans. Inform. Theory* **IT-16**(6): 720–738 (Nov. 1970).

57. J. E. Porath, Algorithms for converting convolutional codes from feedback to feedforward form and vice versa, *Electron. Lett.* **25**(15): 1008–1009 (July 1989).

58. J. P. Odenwalder, *Optimal Decoding of Convolutional Codes*, Ph.D. thesis, Univ. California, Los Angeles, 1970.

59. K. J. Larsen, Short convolutional codes with maximum free distance for rates 1/2, 1/3, and 1/4, *IEEE Trans. Inform. Theory* **IT-19**: 371–372 (May 1973).

60. E. Paaske, Short binary convolutional codes with maximum free distance for rates 2/3 and 3/4, *IEEE Trans. Inform. Theory* **IT-20**: 683–689 (Sept. 1974).

TRENDS IN BROADBAND COMMUNICATION NETWORKS

SASTRI KOTA
Loral Skynet
Palo Alto, California

1. INTRODUCTION

In the 1970s, the Advanced Research Projects Agency (ARPA) initiated the development of ARPANET, a very successful resource-sharing computer network [1,2]. ARPANET was a wide-area packet switching network, which later evolved into the Internet. This ARPANET resulted in a digital network resolution covering the telecommunication networks, data networks, and multimedia networks. The rapid growth of digital technologies allowed the integration of voice, data, and video to be processed, as a single stream and transported over global networks. The advances in high-speed networking broadband access technologies, the increasing power of the personal computer, the availability of information at the click of a button, and the exponential growth of the Internet, makes digital networks demand grow at a very rapid pace.

Traditionally, telecommunication networks utilized hierarchical controlled circuit switch technologies. Data networks demonstrated the strengths of packet-switched technology. The recent internetwork utilized various broadband services, such as voice, data, videostreams, multimedia, and group working, for schools, universities, hospitals, transported over either fiber, cable, satellite, and/or wavelength-division multiplexing (WDM). For example, the processor in a videogame is 10,000 times faster than Electronic Numerical Integrator and Computer (ENIAC) (1947) computer, and genesis game has more processing power than the 1976 Cray Super computer. Some of the chips used in some videocameras are more powerful than IBM 360. The increasing demand for broadband services, high-speed data transmission over the Internet and video on demand, require significant capacities challenging the fixed and mobile network designers and operations to deploy infrastructures to meet these future demands.

High speed Internet access was previously limited to enterprises, using technologies such as leased T-1, frame relay, or asynchronous transfer mode (ATM). But with the exponential growth of the Internet access for residential users, service providers have recognized the great opportunity in the residential broadband as well as more broadband enterprise markets.

It was reported that high-speed Internet access totaled 1.19 billion hours, 51% of the total 2.3 billion hours spent online, in January 2002. Twenty-two million used broadband to surf the Internet, an increase of 67% while enterprise usage increased by 42% [3]. As a result, the telephone, cable, and satellite companies have been developing xDSL, cable, and satellite access technologies. The objectives of these infrastructures are to provide higher bandwidth and speed to optimize the use of Internet for emerging applications such as content delivery distribution, e-finance, telemedicine, distance learning,

streaming video and audio, and interactive games. This article focuses mainly on the technology options for enterprise and residential users emphasizing the technical challenges through network examples.

This article is organized as follows. Broadband services, future applications and their requirements are discussed in Section 2. Section 3 provides a generic global communication model interconnecting the backbones based on either frame relay, ATM, or IP and even satellite technologies. The different broadband access technologies, their concepts and the enterprise access network solutions are described in Section 4. Section 5 discusses the residential broadband access technologies including Digital Subscriber Line (DSL), cable, wireless, and satellite solutions. Section 6 lists the standard organizations developing the protocols and interface standards, for enhancing the cost-effective interoperability of these infrastructures to meet the requirements of the growing high-speed applications. Section 7 describes the technical challenges for future networking with a multiprotocol label-switching service (MPLS)-based solution.

2. BROADBAND SERVICES AND APPLICATIONS

The Internet evolution has accompanied new opportunities for multimedia applications and services, ranging from simple file transfer to broadband access, IP multicast, media streaming, and content delivery distribution for both enterprise and consumer services. Figure 1 shows a projected growth of worldwide Internet users, where most users have demand for broadband services. In the consumer market the growing awareness of the Internet and activities ranging from shopping to finding local entertainment options to children's homework are driving the steep demand for more bandwidth. Education and entertainment content delivery has become one of the prime applications of Internet. During business globalization an increase of virtual business teams, enterprises, increase in competition for highly skilled workers, service providers and equipment vendors, are driving the demand for higher bandwidths or broadband.

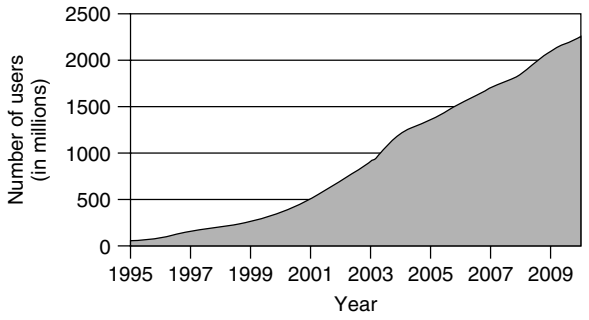


Figure 1. Internet users. (*Source*: Nua Internet Surveys + vgc projections.)

Some of the observations are that 75% of traffic on the Internet is Web-based; there are 3.6 million Web sites with 300–700 million Web pages; the traffic consists of 80% data and 20% voice with a traffic growth of 100–1000% per year [4].

2.1. Broadband Services

Table 1 shows an example of the broadband services and applications. These include entertainment, broadband, and business services. A major challenge for these emerging services supported by Internet, which is an IP-based network, is to provide adequate QoS (quality of service). The normal QoS parameters from a user's perspective for both enterprise and residential users include throughput, packet loss, end-to-end delay, delay jitter, and reliability.

2.2. Network Requirements

The networking infrastructure supporting the enterprise and residential services should meet the following requirements:

- *High Data Rates*. The future applications such as videostreaming, media cast distributions, and high-speed Internet access for telemedicine applications and two-way telephonic education require rates ranging from a few hundred megabits to several gigabits. The three-generation (3G) systems cover up to 2 Mbps (megabits per second) for indoor environments and 144 kbps for vehicular environments. The IEEE 802.11-based broadband systems have approximately 20–30 Mbps transmission speeds. The data

Table 1. Broadband Services

Entertainment	Broadband	Business	Voice and Data Trunking
Broadcasting (DTH)	High-speed Internet access for consumer and enterprise	Telecommuting	IP Transport and ISP
Video on demand (VOD)	Electronic messaging	Videoconferencing	Voice over IP
Network or TV distribution	News on demand	E-finance, B2B	Video, audio, and data file transfer
TV cotransmissions	Multimedia	Home security	
Karaoke on demand	Distance learning	Unified messaging	
Games	MAN and WAN connectivities		
Gambling	Telemedicine		

rates for future generations will range from 2 to 600 Mbps depending on the system. The target speed for 4G cellular will be around 10–20 Mbps.

- *Delay*. Many real-time applications require minimum delay and the packet transfer delays for other classes of services are even stringent, reducing the queuing and processing delays.
- *Mobility*. The 4G cellular systems might be required to provide at least 2 Mbps for moving vehicles.
- *Wide Coverage*. The next-generation systems must provide good coverage area and roaming and handover to other systems.

Next-generation systems should provide at least an order of magnitude higher capacity, but the bit cost should be reduced to make the service more affordable. In addition to these requirements, the network must be scalable and provide security.

Figure 2 shows how different applications vary in their QoS requirements.

The most important QoS parameters are [5]

Throughput. Throughput is the effective data transfer rate in bits per second (bps) of the network. Sharing of network capacity by a number of users reduces the throughput per user; as does the overhead of the packet. A service provider guarantees a minimum bandwidth rate.

Delay. The time taken to transport data from the source to destination is known as *delay* or *latency*. For the public Internet, a voice call may easily exceed 150 ms of delay, because of processing delay and congestion. In GEO satellites one-way propagation delay can reach 250 ms. Such GEO satellite networks cannot service many real-time applications.

Jitter. Jitter is caused by delay variation related to processing and queuing. These delay variations might be due to packet reassembly.

Reliability. The percentage of network availability is an important parameter for the user. In satellite-based networks, the availability depends on the frequency band of operation, power levels, antenna size, and the traffic for the service provided. Advanced error control techniques are used to provide good link availability [6].

Packet Loss. Packet loss occurs as a result of network congestion, error conditions, and link outages. Whenever buffers overflow, packets are dropped. Several buffer management schemes have been proposed to reduce the packet loss rate.

To meet future service requirements, a combination of old and new broadband networking technologies for enterprises include Gigabit Ethernet, frame relay, asynchronous transfer mode (ATM), multiprotocol label switching (MPLS), and broadband satellite networks. There are several options for residential users with respect to broadband access such as (1) dialup, (2) cable, (3) Digital Subscriber Loop (DSL) and its variants, (4) hybrid fiber–coax, (5) wireless including local multipoint distribution services (LMDS) and multichannel multipoint distribution services (MMDS), (6) satellite access, and (7) leased lines.

3. GLOBAL NETWORK MODEL

The tremendous success and exponential growth of the Internet and the new applications such as broadcasting, multicasting, and video distribution have resulted in significantly higher data rates. One of the key requirements for the emerging "global network," which is a "network of networks," is rich connectivity among fixed as well as mobile users. Advances in switching and transport technologies have made increase in transmission bandwidth and switching speeds possible, and still more dramatic increases are possible via optical switching [7]. The future generation of communications networks provides "multimedia services," "wireless (cellular and satellite) access to broadband networks," and "seamless roaming among different systems."

Figure 3 shows a global communication network scenario providing connectivity among corporate networks, Internet, and the ISPs. The networking technologies vary

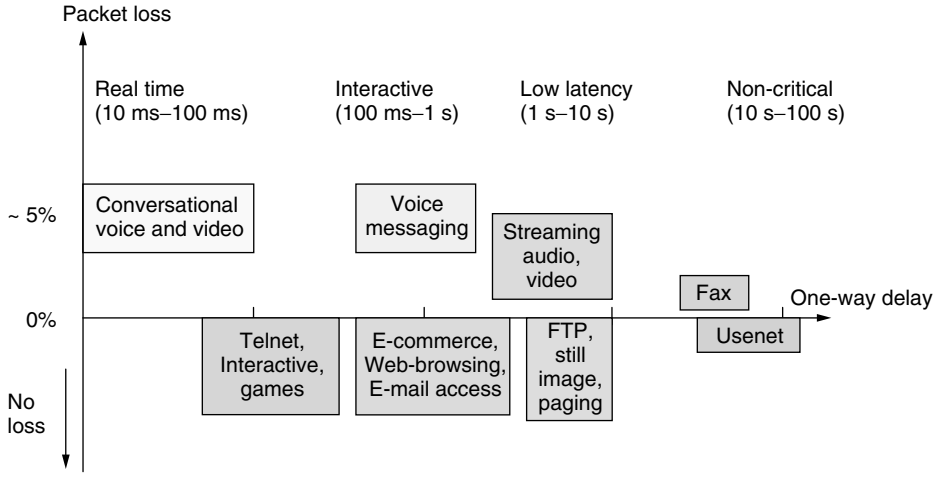


Figure 2. Application-specific QoS requirements example.

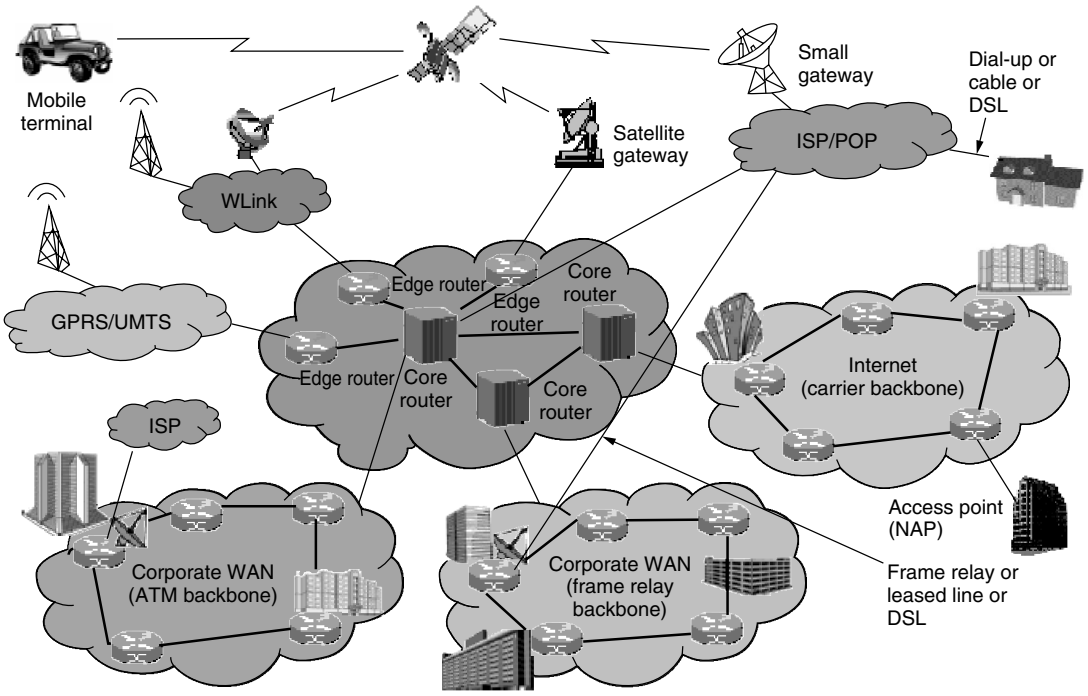


Figure 3. Communication network scenario.

can vary between ATM, frame relay, IP and optical backbones. The access technologies could be dialup, cable, DSL, and satellite.

Mobile communications are supported by second-generation digital cellular (GSM); data service is supported by GPRS. Third-generation systems such as IMT-2000 can provide 2 Mbps and 144 kbps indoors and in vehicular environments. Even 4G and 5G systems are being studied to provide data rates 2–20 Mbps and 20–100 Mbps, respectively. See Section 5.4 for further details.

Several broadband satellite networks at Ka band are planned and being developed to provide such global connectivity for both fixed satellite service (FSS) and mobile satellite service (MSS) using geosynchronous (GSO) and nongeosynchronous (NGSO) satellites as discussed in Section 4.5. Currently GSO satellite networks with very-small-aperture terminals (VSATs) at Ku bands are being used for several credit card verifications, rental cars, banking, and other applications. Satellite networks such as StarBand, Direway, and WildBlue are being developed for high speed Internet access (see Section 5.3).

4. BROADBAND ENTERPRISE NETWORKING

This section describes the networking technologies for enterprise including the use of Gigabit Ethernet, frame relay, ATM, IP, and broadband satellite. The technology concepts and advantages are discussed with some examples.

4.1. Gigabit Ethernet

Gigabit Ethernet is an extension of the IEEE 802.3 Ethernet standard. It builds on the Ethernet protocol but increases speed 10-fold over Fast Ethernet to 10 Gbps. In March 1999, a working group was formed at IEEE 802.3 to develop a standard for 10-gigabit Ethernet. Gigabit Ethernet is basically the faster-speed version of Ethernet. It will support the data rate of 10 Gbps and offers benefits similar to those of the preceding Ethernet standard [8]. The potential applications for 10-gigabit Ethernet enterprise users are universities, telecommunication carriers, and Internet service providers. One of the main benefits of the 10-gigabit standard is that it offers a low-cost solution to solve the demand on bandwidth. In addition to the low cost of installation, the cost of network maintenance and management is minimal. Local network administrators manage and maintain for 10-gigabit Ethernet networks.

In addition to the cost reduction benefit, 10-gigabit Ethernet allows for faster switching and scalability. 10-gigabit Ethernet uses the same Ethernet format, it allows seamless integration of LAN, MAN, and WAN. There is no need for packet fragmentation, reassembling, or address translation, eliminating the need for routers that are much slower than switches. 10-gigabit Ethernet also offers straightforward scalability since the upgrade paths are similar to those of 1-gigabit Ethernet.

In LAN markets, applications typically include in-building computer servers, building-to-building clusters, and data centers. In this case, the distance requirement is relaxed, usually between 100 and 300 m. In the medium-haul market, applications usually include campus backbones, enterprise backbones, and storage area networks. In this case, the distance requirement is moderate, usually between 2 and 20 kms. The cabling infrastructure usually already exists. The technologies must operate over it. The initial cost is not so much of an issue. Normally, users are

willing to pay for the cost of installation but not the cost of network maintenance and management. Ease of technology is preferred. This gives an edge to 10-gigabit Ethernet over the only currently available 10-gigabit SONET technology (OC-192c).

WAN markets typically include Internet service providers and Internet backbone facilities. Most of the access points for long-distance transport networks require the OC-192c data rate. So the key requirement for these markets is the compatibility with the existing OC-192c technologies. Mainly, the data rate should be 9.584640 Gbps. Thus, the 10-Gigabit Ethernet standard specifies a mechanism to accommodate the rate difference.

4.1.1. Protocol. The Ethernet protocol basically implements the bottom two layers of the Open Systems Interconnection (OSI) seven-layer model, that is, the data-link and physical sublayers. Figure 4 depicts the typical Ethernet protocol stack and its relationship to the OSI model.

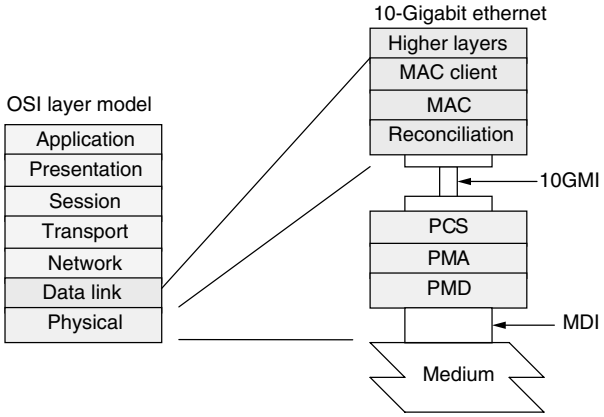


Figure 4. Ethernet protocol layer.

Media Access Control (*MAC*). The media access control sublayer provides a logical connection between the MAC clients of itself and its peer station. Its main responsibility is to initialize, control, and manage the connection with the peer station.

Reconciliation Sublayer. The reconciliation sublayer acts as a command translator. It maps the terminology and commands used in the MAC layer into electrical formats appropriate for the physical-layer entities.

10-Gigabit Media-Independent Interface (*10GMII*). 10GMII provides a standard interface between the MAC layer and the physical layer. It isolates the MAC layer and the physical layer, enabling the MAC layer to be used with various implementations of the physical layer.

Physical Coding Sublayer (*PCS*). The PCS sublayer is responsible for coding and encoding data streams to and from the MAC layer. A default coding technique has not been defined.

Physical Medium Attachment (*PMA*). The PMA sublayer is responsible for serializing code groups into bit streams suitable for serial bit-oriented physical devices and vice versa. Synchronization is also done for proper data decoding in the sublayer.

Physical Medium-Dependent (*PMD*). The PMD sublayer is responsible for signal transmission. The typical PMD functionality includes amplifier, modulation, and waveshaping. Different PMD devices may support different media.

Medium-Dependent Interface (*MDI*). MDI is referred to as a connector. It defines different connector types for different physical media and PMD services.

4.1.2. Fiber Network Architecture. Figure 5 shows Gigabit Ethernet used in an evolving fiber enterprise

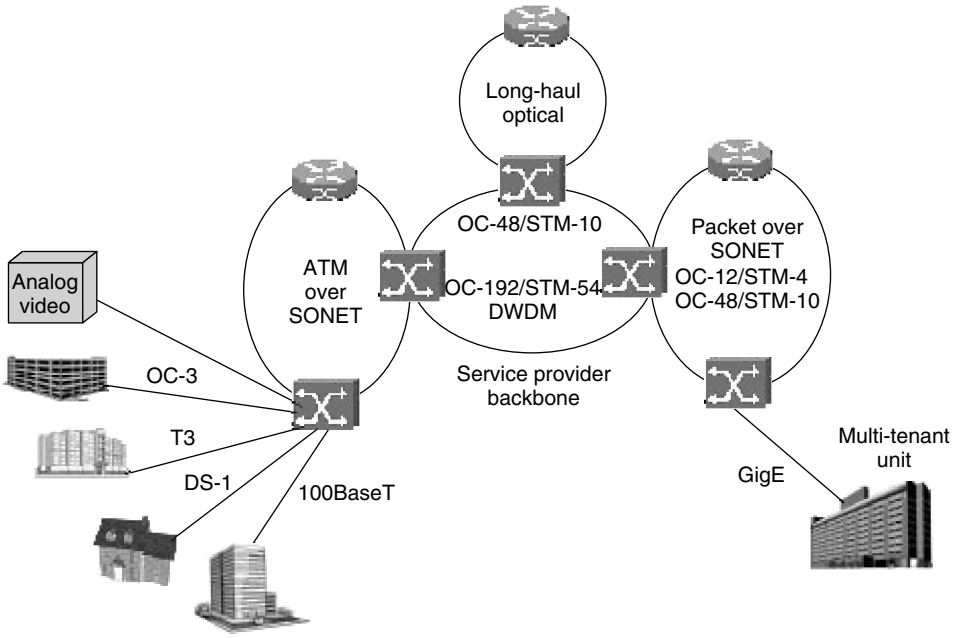


Figure 5. Gigabit Ethernet in a fiber enterprise network.

network configuration. The service provider backbone consists of an OC-48 (2.48 Gbps) or an OC-192 (10 Gbps) using dense wavelength-division multiplexing (DWDM) with connections to ATM over SONET and packet over SONET networks via core routers. The access networks could use OC-3 (155.52 Mbps), or T3 (44.736 Mbps), or DS1 (1.544 Mbps). A typical application using such an enterprise network solution could be distribution of analog and digital video.

4.2. Frame Relay

Frame relay is a standard communication protocol that is specified in ITU-T (formerly CCITT) recommendations I.122 and Q.922, which add relay and routing functions to the data-link layer of the OSI model [9]. Subsequently, the Frame Relay Forum has developed the frame relay specification for wide-area networks. Frame relay services were developed by service providers for enterprise as a cost effective and a better flexible alternative to time-division multiplexing (TDM) and private line services. Enterprises needed dedicated connectivity between offices but could not necessarily afford dedicated circuits. Meanwhile service providers required a reliable means to subscribe their bandwidth-constrained networks.

The frame relay protocol has been particularly effective for data traffic. Carriers generally use frame relay as access technology and ATM as a transport. The network architecture requires carriers to maintain a completely dedicated ATM/frame relay network in addition to their IP and voice networks.

The rapid increase in high bandwidth communication is the main reason for using frame relay technology. Two main factors influence the rapid demand for high-speed networking: (1) rapid increase in use of LANs, and (2) use of fiber optic links. Frame relay is a packet-switching technology, which relies on low error rate digital transmission links and high-performance processors. Frame relay technology was designed to cover (1) low latency and higher throughput, (2) bandwidth on demand, (3) dynamic sharing of bandwidth, and (4) backbone network.

For enterprises, frame relay is a well-understood technology, and by definition, is a layer 2 technology that supports oversubscription. The frame relay technology has some design advantages as well as restrictions including

- Unpredictable bandwidth and maximum speed capacity at DS3
- Hierarchical aggregation schemes that use hub and spoke architectures
- Scaling complexities by having to add additional layer 2 addresses to different sites rather than by IP's inherent self-healing and learning capabilities
- Used for interconnecting LANs and particularly WANs, and more recently, for voice and videoconferencing
- Provides LAN-to-LAN connectivity from 56 kbps to 1.5 Mbps
- Offers congestion control and higher performance

4.2.1. Frame Relay Data Unit. The frame structure in a frame relay network consists of two flags indicating the beginning and the end of the frame, an address field, an information field, and a frame check sequence [10]. In addition to the address, the address field contains functions that warn of overload and indicate which frames should be discarded first. The fields and their purpose are discussed in detail below.

- *Flag*. All frames begin and end with a flag consisting of an octet composed of a known bit pattern: a zero followed by 6 ones and a zero (01111110).
- *Address*. In the two octets in the address field, the first 6 bits of the first octet and the first 4 bits of the second octet are used for addressing. These 10 bits, which form the DLCI, select the next destination to which the frame is to be transported.
- *CR*. Command response is not used by the frame relay protocol. It is sent transparently through the frame relay nodes and can be made available to users as required.
- *EA*. At the end of each address octet there is an extended address bit that can allow extension of the DLCI field to more than 10 bits. If the EA bit is set to "0," another address octet will follow. If it is set to "1," the octet in question is the last one in the address field.
- *FECN*. If overload occurs in the network, forward explicit congestion notification is indicated to alert the receiving end. The network makes this indication, and end users need not take any specific action.
- *BECN*. Similar to FECN, backward explicit congestion notification alerts the sending end to an overload situation in the network.
- *DE*. Discard eligibility indicates that the frame is to be discarded in case of overload. This indication can be regarded as a prioritizing function, although frames without a DE indication can also be discarded.
- *Information Field*. This is where we find user information. The network operator decides how many octets the field is allowed to contain, but the Frame Relay Forum recommends a maximum of 1600. The information passes through the network completely unchanged and is not interpreted by the frame relay protocol.
- *FCS*. The frame check sequence checks the frame for errors. All bits in the frame, except the flags and FCS, are checked.

The frame relay frame format is shown in Fig. 6.

4.2.2. Frame Relay Examples

4.2.2.1. Frame Relay LAN-to-LAN. Figure 7 shows LAN-to-LAN and an ISP connectivity using frame relay at 56 kbps, 384 kbps, T1, and T1/T3, respectively.

4.2.2.2. Frame Relay WAN Architecture. Figure 8 shows a frame relay network with various offices connected via frame relay to four aggregation hubs. Depending on

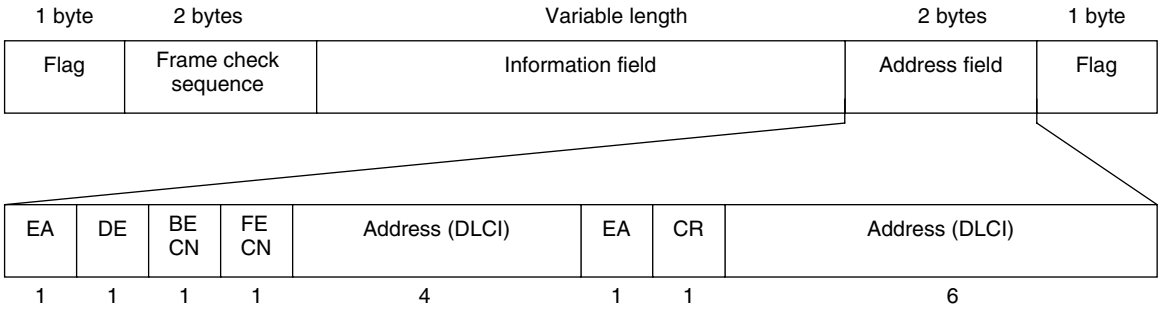


Figure 6. Frame relay frame structure.

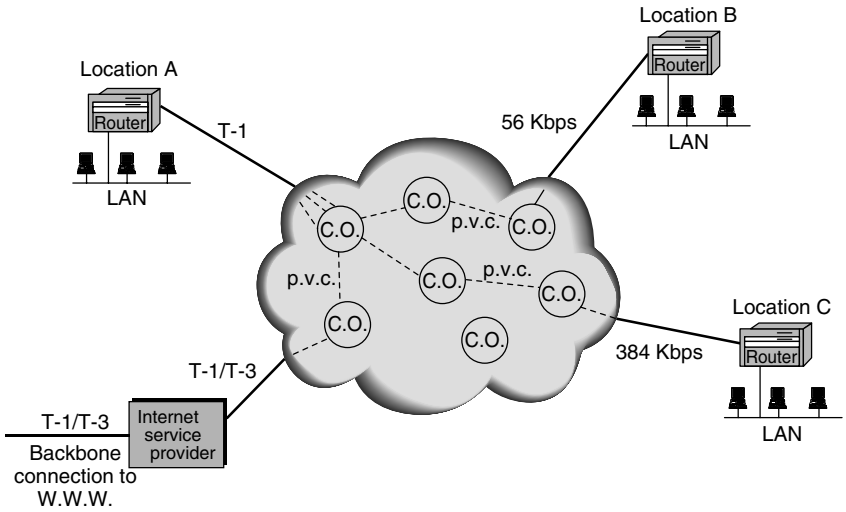


Figure 7. Frame relay overview.

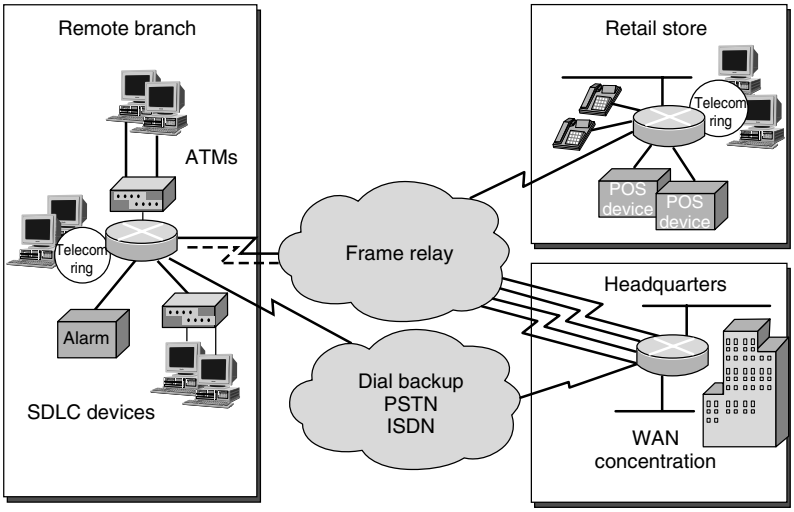


Figure 8. WAN architecture.

the size of the organization and the speeds at which different regional offices connect, the hub will have at least two large WAN routers. For some organizations, hubs may be fed by hundreds of regional locations through frame relay or private line connections. Although some of the traffic may terminate at its locally connected "hub" data center, most of the traffic "hairpins" in and out of the local data center on its way to a remote hub or destination. In this case, the hub sites provide statistical multiplex gains.

4.3. Asynchronous Transfer Mode (ATM)

ATM is an International Telecommunication Union—Telecommunication Standardization Sector (ITU-T) standard for cell relay wherein information for multiple service types, such as voice, video, or data, is conveyed in small, fixed-size cells. ATM networks are connection oriented. ATM is based on the efforts of the ITU-T Broadband Integrated Services Digital Network (BISDN) standard. It was

originally conceived as a high-speed transfer technology for voice, video, and data over public networks. The ATM Forum extended the ITU-T's vision of ATM for use over public and private networks.

ATM is a cell switching and multiplexing technology that combines the benefits of circuit switching (guaranteed capacity and constant transmission delay) with those of packet switching (flexibility and efficiency for intermittent traffic) [11]. It provides scalable bandwidth from a few megabits per second (Mbps) to many gigabits per second (Gbps). Because of its asynchronous nature, ATM is more efficient than synchronous technologies, such as time-division multiplexing (TDM). With TDM, each user is assigned to a time slot, and no other station can send in that time slot. If a station has a lot of data to send, it can send only when its time slot comes up, even if all other time slots are empty. If, however, a station has nothing to transmit when its time slot comes up, the time slot is sent empty and is wasted. Because ATM is asynchronous, time slots are available on demand with information identifying the source of the transmission contained in the header of each ATM cell.

4.3.1. ATM Reference Model. Figure 9 illustrates the B-ISDN Protocol Reference Model, which is the basis for the protocols that operate across the User Network Interface (UNI). The B-ISDN reference model consists of three planes: the user plane, the control plane, and the management plane. Reference ITU-T I.121 describes the ATM reference model and various functions in detail.

4.3.2. ATM Cell Format. An ATM cell consists of 48 bytes of data with a 5-byte header as shown in Fig. 10 [11]. The cell size was determined by ITU-T as a compromise between voice and data requirements. The header fields are as follows:

- *Generic Flow Control* (*GFC*). Provides local functions, such as identifying multiple stations that share a single ATM interface.
- *Virtual Path Identifier* (*VPI*). In conjunction with the VCI, identifies the next destination of a cell as it passes through a series of ATM switches on the way to its destination.
- *Virtual Channel Identifier* (*VCI*). In conjunction with the VPI, identifies the next destination of a cell as it passes through a series of ATM switches on the way to its destination.
- *Payload Type* (*PT*). If the cell contains user data, the second bit indicates congestion, and the third bit indicates whether the cell is the last in a series of cells that represent a single AAL5 frame.
- *Cell Loss Priority* (*CLP*). Indicates whether the cell should be discarded if it encounters extreme congestion as it moves through the network. If the CLP bit equals 1, the cell should be discarded in preference to cells with the CLP bit equal to zero.
- *Header Error Control* (*HEC*). Calculates checksum only on the header.

There are two types of ATM services: permanent virtual circuits (PVCs) and switched virtual circuits (SVCs). A PVC allows direct static connectivity between sites similar to a leased line. It guarantees availability of a connection and does not require a signaling protocol. On the other hand, SVC allows dynamic setup and release of connections. Dynamic call control requires a signaling protocol between the ATM endpoint and the switch. This service provides flexibility. However, it results in a signaling overhead in setting up the connection.

4.3.3. Classes of Service. Table 2 provides different classes of network traffic that need to be treated differently by an ATM network [12].

4.3.4. Traffic Management and QoS. One of the significant advantages of ATM technology is providing QoS guarantees as described in the ATM Forum's *Traffic Management Specification* [13]. The framework supports five service categories, namely, constant bit rate (CBR), real-time variable bit rate (rt-VBR), non-real-time VBR (nrt-VBR), unspecified bit rate (UBR), and available bit

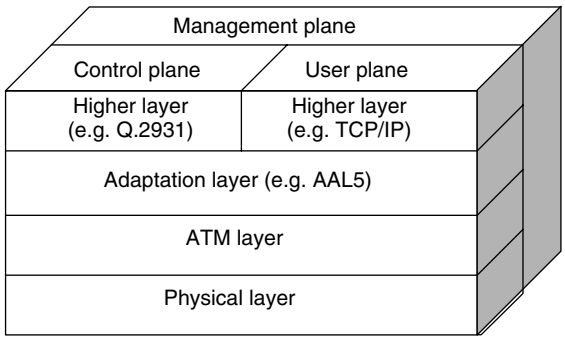


Figure 9. ATM protocol architecture.

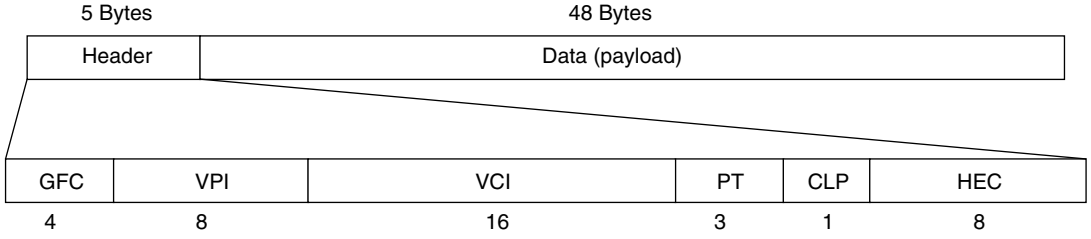


Figure 10. ATM cell structure.

Table 2. Classes of Service

	Constant-Bit-Rate (CBR) Service	Real-Time Service (VBR-rt)	Non-Real-Time Connection-Oriented Data Service (VBR-nrt-COD)	Non-Real-Time Connectionless Data Service (VBR-nrt-CLS)	Unspecified Bit Rate (UBR)	Available Bit Rate (ABR)
Bearer class	Class A	Class B	Class C	Class D	Class X	Class Y
Applications	Voice and clear channel	Packet video and voice	Data	Data	Data	Data
Connection mode	Connection oriented PVC or SVC	Connection oriented PVC or SVC	Connection oriented PVC or SVC	Connection less	Connection oriented	Connection oriented
Bit rate	Constant	Variable	Variable	Variable	Variable	Variable
Timing required	Required	Required	Not required	Not required	Not required	Not required
Services	Private line	None	Frame relay	SMDS	Raw cell	
AAL	1	2	3/4 & 5	3/4	Any	3/4 & 5

rate (ABR), with an additional one, guaranteed frame rate (GFR). With the exception of UBR, all ATM service categories require incoming traffic regulation to control network congestion and ensure QoS guarantees. This function is performed by access policing devices to determine whether the traffic conforms to certain traffic characterizations. The conformant cells are allowed to enter the ATM network and receive QoS guarantees, whereas the nonconformant cells will be either dropped or tagged. Tagged cells may be allowed into the network but will not receive any QoS guarantees. The other traffic management functions include connection admission control (CAC), traffic shaping, usage parameter control (UPC), resource management, priority control, cell discarding, and feedback controls. The ATM Forum [13] provides the details of these functions and traffic management algorithm recommendations for different service classes. Table 3 provides the various traffic parameters and the QoS parameters.

4.4. IP Enterprise Network

Since the late 1990s the Internet has become the major vehicle for most of the broadband applications and for the growth of the telecommunication network. The Internet protocol has become the universal network layer protocol for both wireline and wireless networks. At the transport layer, Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) have become the most popular ones. The UDP has been attracted with the multimedia services. In this section, an overview of TCP and IP protocols with formats and examples is described.

4.4.1. TCP/IP Protocol. The TCP/IP protocol suite is the mostly used Internet protocol for global networks. Many of the Internet applications such as file transfer, email, Web browsing, streaming media, and newsgroups use TCP/IP protocols. Figure 11 shows the TCP/IP protocol stack with respect to ISO/OSI protocol. The overhead introduced at each layer to the application datagram is also described.

4.4.1.1. Transmission Control Protocol (TCP). The TCP provides reliable transmission of data in an IP environment. TCP corresponds to the transport layer (layer 4) of the OSI reference model [14]. Among the services TCP provides are stream data transfer, reliability, efficient flow control, full-duplex operation, and multiplexing. TCP offers reliability by providing connection-oriented, end-to-end reliable packet delivery through an internetwork by using acknowledgments. The reliability mechanism of TCP

Table 3. ATM Service Category Attributes

	Traffic Parameters			QoS Parameters			
Service Categories	PCR, $CDVT_{PCR}$	SCR, MBS, $CDVT_{SCR}$	MCR	Peak-to-Peak CDV	Maximum CTD	CLR	Others
CBR	Yes	No	No	Yes	Yes	Yes	No
rt-VBR	Yes	Yes	No	Yes	Yes	Yes	No
nrt-VBR	Yes	Yes	No	No	No	Yes	No
UBR	Yes	No	No	No	No	No	No
ABR	Yes	No	Yes	No	No	No	Feedback
GFR	Yes	MFS, MBS	Yes	No	No	No	No

Key: PCR — peak cell rate; SCR — sustainable cell rate; MCR — minimum cell rate; MBS — maximum burst size; CDVT — cell delay variation tolerance; CDV — cell delay variation; CTD — cell transfer delay; CLR — cell loss ratio; MFS — maximum frame size.

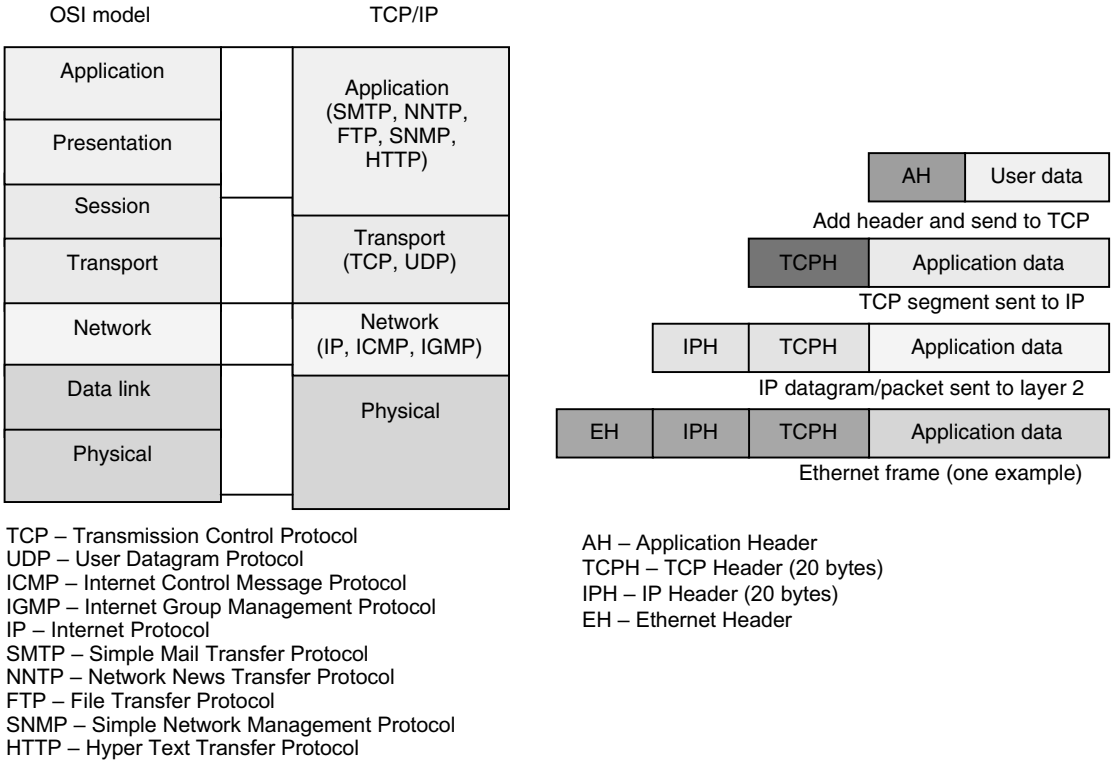


Figure 11. TCP/IP protocol stack.

Source port (16)			Destination port (16)
Sequence number (32)			
Acknowledgement number (32)			
Header size (4)	Reserved (6)	6 Control bits (6)	Window size (16)
Checksum (16)			Urgent pointer (16)
Options			Padding
Data (variable length)			

Figure 12. TCP segment format.

Source port (16)	Destination port (16)
Length (16)	Checksum (16)
Data (variable length)	

Figure 13. UDP segment format.

allows devices to deal with lost, delayed, duplicate, or misread packets. The lost packets are detected by a timeout mechanism. TCP offers a efficient flow control mechanism by using sliding windows to avoid buffer overflow. TCP provides a full-duplex operation and multiplexing of data. Along with the IP Security protocol (IPSec), Internet security can be obtained. Figure 12 shows the format of a TCP segment. The header is 20 bytes long.

4.4.1.2. User Datagram Protocol (UDP). UDP is a connectionless transport-layer protocol that enables best-effort datagrams to be transmitted between host system applications. Unlike TCP, UDP adds no reliability, flow control, or error recovery functions to IP [14]. It uses an 8-byte header. The format of a UDP segment is shown in Fig. 13.

4.4.1.3. Internet Protocol (IP). IP is a connectionless protocol working at the network layer (layer 3). IP has two primary responsibilities: providing connectionless "best effort" delivery of datagrams across and between networks; and providing fragmentation and reassembly of datagrams to support data links with different maximum transmission unit (MTU) sizes.

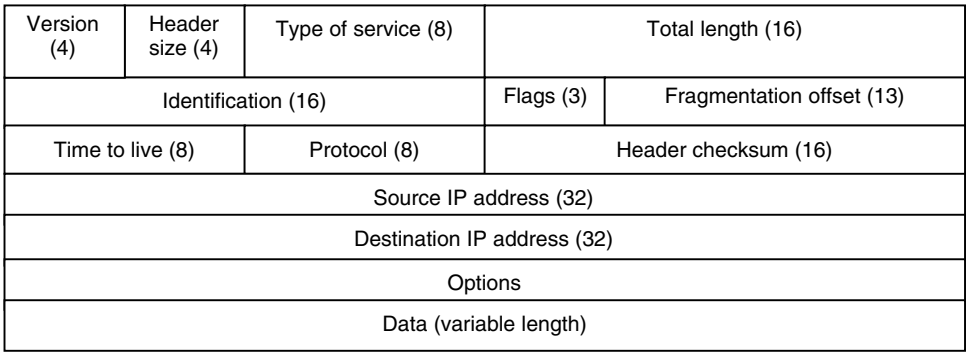


Figure 14. IPv4 segment format.

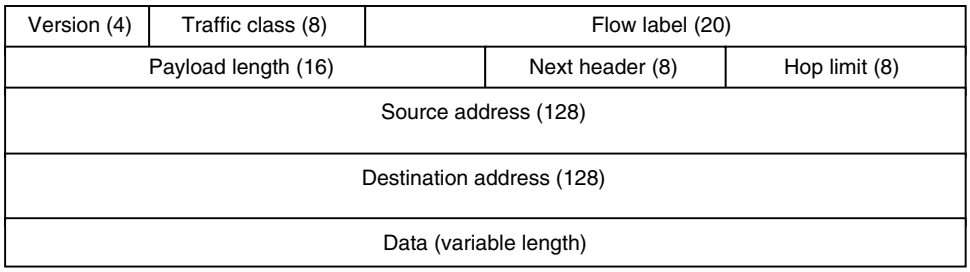


Figure 15. IPv6 segment format.

4.4.2. IPv4 and IPv6 Formats. Currently most of the Internet uses IP version 4 (IPv4) [14]. IP, as originally developed, has no reliability or QoS mechanisms. A new version, IP version 6 (IPv6) has been designed as the successor of IPv4. IPv6 provides more QoS and address space than IPv4 [15]. Figures 14 and 15 show the formats of IPv4 and IPv6 segments. The changes from IPv4 to IPv6 fall into the following categories:

- *Expanded Address Capabilities*. IPv6 increases the IP address size from 32 to 128 bits, to support more levels of addressing hierarchy, and a much greater number of addressable nodes.
- *Header Format Simplification*. Some IPv4 header fields have been dropped or made optional, to reduce the processing cost of packet handling.
- *Improved Support for Extensions and Options*. Changes in the way IP header options are encoded allows for more efficient forwarding and greater flexibility for introducing new options in the future.
- *Flow Labeling Capability*. A new capability is added to enable the labeling of packets belonging to a particular traffic "flow" for which the sender requests special handling.

The fields in the IPv6 segment are

- *Version* — 4-bit Internet Protocol version number = 6.
- *Traffic class* — 8-bit traffic class field. Available for use by originating nodes and/or forwarding routers to identify and distinguish between different classes or priorities of IPv6 packets.
- *Flow label* — 20-bit flow label. Used by a source to label sequences of packets for which it requests special handling by the IPv6 routers.
- *Payload length* — 16-bit unsigned integer. Length of the rest of the packet following the header.
- *Next header* — 8-bit selector. Identifies the type of header immediately following the IPv6 header.
- *Hop limit* — 8-bit unsigned integer. Decremented by 1 by each node that forwards the packet. The packet is discarded if hop limit is decremented to zero.
- *Source address* — 128-bit address of the originator of the packet.
- *Destination address* — 128-bit address of the intended recipient of the packet.

4.4.3. Network Example. As a less expensive WAN technology than frame relay, IP virtual private networks (IPVPNs) are being developed. The premise behind the IPVPNs is to provide a logically private network across a shared infrastructure using tunneling or IPSec protocol to provide security to the enterprise network [16]. Figure 16 shows such an IPVPN network for an enterprise solution consisting of a number of virtual routers from the different sites connected together across the Internet. These virtual routers are used to enable the VPN services connecting all sites in a mesh topology.

The advantage of IPVPNs is that the public Internet provides ubiquitous connectivity. IPSec is used to tunnel the data between the sites across an entirely best-effort public network. The corporations use IPVPNs

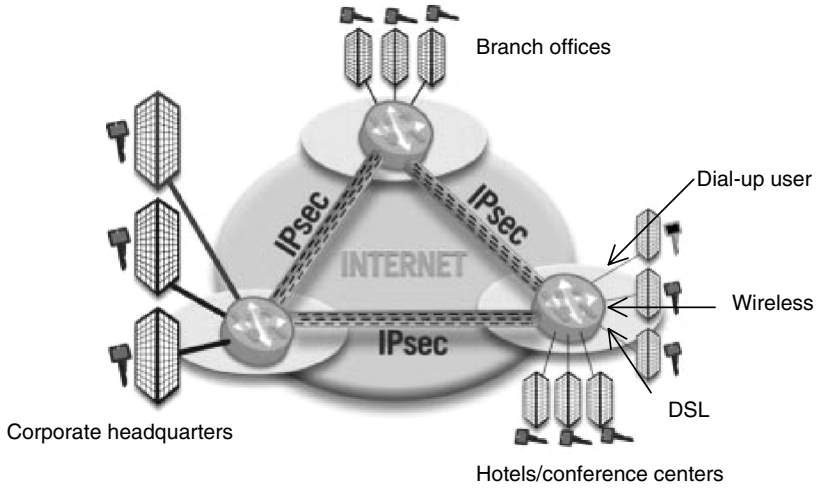


Figure 16. IP enterprise network model.

to connect their data centers, remote offices, mobile employees, telecommuters, customers, suppliers, and business partners. The IPVPN solutions is very attractive because it offers a dedicated private network and is relatively cost-effective.

Many large enterprises have not deployed IPVPNs as expected for the following reasons:

- *Virtual Routers*. The IPVPNs use a virtual router per VPN and the use of virtual router has disadvantages such as (1) no autonomy due to the use of public backbone, (2) enterprise does not have any control on the traffic flows across the network, and (3) the performance of the virtual router is inversely proportional to the number of routers in the network.
- *Service-Level Agreement*. It is very difficult for an enterprise to have a reasonable SLA.
- *Security*. In the enterprise solution, security protocols like IPSec are employed to secure the data transported over a public Internet. However, large corporations normally spend a good portion of their resources monitoring and improving security of their network.
- *Service Guarantees*. There is no possibility of bandwidth or connectivity guarantee when the traffic is carried over a public Internet comprised of many service providers.

Although the IPVPN solution for enterprise network discussed here provides cheap connectivity, it is not well adopted by large enterprises because of the nature of the use of public backbone. It is not suitable for high-bandwidth applications. IPVPNs have been adopted, however, as a cheap solution for network connectivity for small–medium-sized companies.

4.5. Broadband Satellite Networks

Because of the inherent broadcast and multiple access characteristics, satellites have become an attractive solution for enterprise and consumer users. This section describes the future satellite systems operating at the C, Ku, and Ka bands for enterprise applications [17].

4.5.1. Satellite Network Characteristics. The principal attributes of satellites include (1) broadcasting and (2) long-haul communications. For example, Ka-band spot beams generally cover a radius of 200 km (125 mi) or more, while C- and Ku-band beams can cover entire continents. This is in contrast to last-mile-only technologies that generally range from 2 to 50 km (or 1–30 mi).

The main advantages of satellite communications are [18]

- *Ubiquitous Coverage*. A single satellite system can reach every potential user across an entire continent regardless of location, particularly in areas with low subscriber density and/or otherwise impossible or difficult to reach. Current satellites have various antenna types that generate different footprint sizes. The sizes range form coverage of the whole earth as viewed from space (about a third of the surface) down to a spot beam that covers much of Europe or North America. All these coverage options are usually available on the same satellite. Selection between coverage is made on transparent satellites by the signal frequencies. It is spot beam coverage that is most relevant for access since they operate to terminal equipment of least size and cost. Future systems will have very narrow spot beams of a few hundred miles across that have a width of a fraction of a degree.
- *Bandwidth Flexibility*. Satellite bandwidth can be configured easily to provide capacity to customers in virtually any combination or configuration required. This includes simplex and duplex circuits from narrowband to wideband and symmetric and asymmetric configurations. Future satellite networks with narrow spot beams are expected to deliver rates of up to 100 Mbps with 90-cm antennas, and the backplane speed within the satellite switch could be typically in the Gbps range. The uplink rate from a 90-cm user terminal is typically 384 kbps.

- *Cost*. The cost is independent of distance; the wide area coverage from a satellite means that it costs the same to receive the signal from anywhere within the coverage area.
- *Deployment*. Satellites can initiate service to an entire continent immediately after deployment, with short installation times for customer premise equipment. Once the network is in place, more users can be added easily.
- *Reliability and Security*. Satellites are amongst the most reliable of all communication technologies, with the exception of SONET fault-tolerant designs. Satellite links require only that the end stations be maintained, and they are less prone to disabling though accidental or malicious damage.
- *Disaster Recovery*. Satellite communication provides an alternative to damaged fiberoptic networks for disaster recovery options and provide emergency communications.

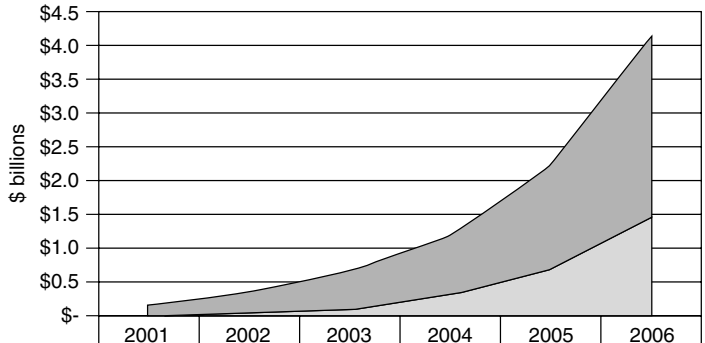
4.5.2. Market Potential. Figure 17 shows the market potential for satellite-based broadband enterprise and residential access users. The enterprise market is expected to grow up to $4 billion by 2006 for content delivery distribution (CDD) [19].

4.5.3. Next-Generation Ka Band. Until the late 1990s, the Ka band was used for experimental satellite programs in the United States, Japan, Italy, and Germany. In the United States, the Advanced Communications Technology Satellite (ACTS) is being used to demonstrate advanced technologies such as onboard processing and scanning spot beams. A number of applications were tested, including distance learning, telemedicine, credit card financial transactions, high-data-rate computer interconnections, videoconferencing, and high-definition television (HDTV). The growing congestion of the C and Ku bands and the success of the ACTS program increased the interest of satellite system developers in the Ka-band satellite communications network for exponentially growing Internet access applications. A rapid convergence of technical, regulatory, and business factors has increased the interest of system developers in Ka-band frequencies.

Several factors influenced the development of multimedia satellite networks at Ka-band frequencies:

- *Adaptive Power Control and Adaptive Coding*. Adaptive power control and adaptive coding technologies have been developed for improved performance, mitigating propagation error impacts on system performance at the Ka band.
- *High Data Rate*. A large bandwidth allocation to geosynchronous fixed satellite services (GSO FSS) and nongeosynchronous fixed satellite services (NGSO FSS) makes high-data-rate services feasible over Ka-band systems.
- *Advanced Technology*. Development of low-noise transistors operating in the 20-GHz band and high-power transistors operating in the 30-GHz band have influenced the development of low-cost earth terminals. Space-qualified higher-efficiency traveling-wave tubes (TWTAs) and ASICs development have improved the processing power. Improved satellite bus designs with efficient solar arrays and higher efficiency electric propulsion methods resulted in cost-effective launch vehicles.
- *Global Connectivity*. Advanced network protocols and interfaces are being developed for seamless connectivity with terrestrial infrastructure.
- *Efficient Routing*. Onboard processing and fast packet or cell switching (e.g., ATM, IP) makes multimedia services possible.
- *Resource Allocation*. Demand assignment multiple access (DAMA) algorithms along with traffic management schemes provide capacity allocation on a demand basis.
- *Small Terminals*. Multimedia systems will use small and high-gain antenna on the ground and on the satellites to overcome path loss and gain fades.
- *Broadband Applications*. Ka-band systems, combining traditional satellite strengths of geographic reach and high bandwidth, provide the operators a large subscriber base with scale of economics to develop consumer products.

Figure 18 illustrates a broadband satellite network architecture represented by a ground segment, a space segment, and a network control segment. The ground segment consists of terminals and gateways (GWs),



	2001	2002	2003	2004	2005	2006
Enterprise	$0.166	$0.297	$0.528	$0.877	$1.542	$2.698
Consumer	$0.009	$0.033	$0.093	$0.292	$0.661	$1.453

Figure 17. Satellite CDD service growth. (*Source*: Pioneer Consulting 2002.)

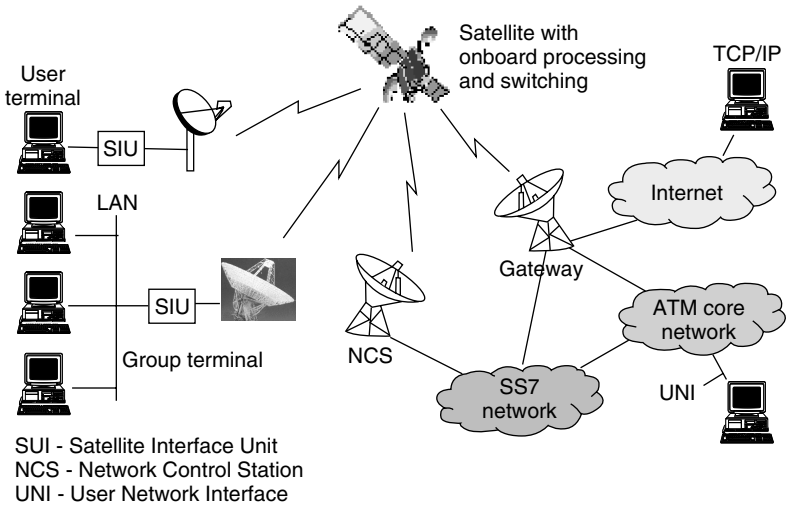


Figure 18. Broadband satellite network configuration.

which may be further, connected to other legacy public and/or private networks. The network control station (NCS) performs various management and resource allocation functions for the satellite media. Intersatellite crosslinks in the space segment to provide seamless global connectivity via the satellite constellation are optional. Hybrid network architecture allows the transmission of packets over satellite, and multiplexes and demultiplexes datagram streams for uplinks, downlinks, and interfaces to interconnect terrestrial networks. The satellite network configuration also illustrates the signaling protocol (e.g., SS7, UNI) and the satellite interface unit. The architectural options could vary from ATM switching, IP transport, to MPLS over satellite.

4.5.4. Future Broadband Satellite Networks. The satellite network architectural options are

- GSO versus NGSO
- No onboard processing or switching
- Onboard processing with ground-based cell or ATM switching or fast packet switching
- Onboard processing and onboard ATM or fast packet switching

The first-generation services that are now in place use existing Ku-band fixed satellite service (FSS) for two-way connections. Using FSS a large geographic area is covered by a single broadcast beam. The new Ka-band systems use spot beams that cover a much smaller area, say, hundreds of miles across. Adjacent cells can use a different frequency range, but a given frequency range can be reused many times over a wide geographic area. The frequency reuse in the spot beam technology increases the capacity. In general, Ka spot beams can provide 30–60 times the system capacity of the first-generation networks.

The next-generation satellite multimedia networks can be divided into two classes: (1) the broadband satellite *connectivity network*, in which full end-to-end user connectivity was established — the proposed global satellite connectivity networks such as Astrolink, Spaceway, and EuroSkyway have onboard processing and switching capabilities; and (2) regional access networks, such as StarBand, IPStar, and WildBlue, which are intended to provide Internet access. These access systems employ nonregenerative payloads.

In a nonregenerative architecture, the satellite receives the uplink and retransmits it on the downlink without onboard demodulation or processing. In a processing architecture with cell switching or layer 3 package, the satellite receives the uplink, demodulates, decodes, switches, and buffers the data to the appropriate beam after encoding and remodulating the data, on the downlink. In a processing architecture, switching and buffering are performed on the satellite; in a nonprocessing architecture, switching/routing and buffering are performed within a gateway. A nonregenerative architecture has physical-layer flexibility but limited, if any, network-layer flexibility. A process architecture has limited physical-layer flexibility, but greatly increased network-layer flexibility, and permits any network topology from point-to-point, hub-and-spoke architecture. The selection of the satellite network architecture is strictly dependent on the target customer applications and performance/cost tradeoffs.

Table 4 compares some of the new-generation C-, Ka-, and Ku-band satellite systems. These systems, which are under development, provide global coverage and high bandwidth.

5. RESIDENTIAL BROADBAND ACCESS

This section describes residential access technologies including DSL, hybrid fiber–coax, fixed wireless, satellite access, and mobile wireless. The technology and examples are described [20].

5.1. Residential Access Market Potential

The different access technologies for broadband applications include cable modem, xDSL, satellite, wireless, and fiber to the home. A report from ARC group concluded that the residential broadband market would be worth $80 billion by 2007 with nearly 300 million residential and

Table 4. Global Broadband Satellite Networks

Services	Spaceway	Astrolink[a]	EuroSky Way	Teledesic	Intelsat	Eutelsat
Data uplink	384 kbps–6 Mbps	384 kbps–2 Mbps	160 kbps–2 Mbps	16 kbps–2 Mbps	—	≤2 Mbps
Data downlink	384 kbps–20 Mbps	384 kbps–155 Mbps	128–640 kbps	16 kbps–64 Mbps	≤45 Mbps	55 Mbps
Number of satellites	8	9 (4 initially)	5	30	—	—
Satellite	GEO	GEO	GEO	MEO	GEO	GEO (Hotbird 3–6)
Frequency band	Ka	Ka	Ka	Ka	C, Ku	Ku, Ka
Onboard processing	Yes	Yes	Yes	Yes	No	—
Operation scheduled	2003	2003	2004	2004/5	—	2001

[a]Program currently on hold.

commercial sites using broadband. Nearly a third of broadband connections will be DSL, with cable close behind. Satellite, wireless, and others will fill the remainder. In fact, for broadband satellite, the high growth rate for CDD applications by 2006 is projected to be $1.45 billion and the revenue growth for wireless, about $3.3 billion in for that same year. It is estimated that 3.8 million subscribers for DSL in 2001 will grow to 20.7 million subscribers in 2006, and to 21 million subscribers to cable modem, according to a Pioneer Consulting report [19].

5.2. Broadband Access Technologies

Figure 19 shows the available broadband access technology options. The Asymmetric DSL (ADSL) provides upstream at 64 kbps, and downstream at 1.8 Mbps whereas very high bit rate (VDSL) supports 26 Mbps symmetric, and in the asymmetric VDSL supports less than 6.4 Mbps upstream and less than 52 Mbps downstream [20]. A digital subscriber line access multiplexer (DSLAM) delivers high-speed data transmission over the existing copper telephone lines. The DSLAM separates the voice frequency signals from high-speed data traffic and controls and routes xDSL traffic between the subscriber's end-user equipment (router, modem, or network interface card) and the network service provider. The cable head end, which includes cable modem termination system, enables high-speed Internet access. Details of the cable modem protocol and access are described in Section 5.2.2.

Table 5 provides a set of advantages and disadvantages of the various broadband access technologies. The technologies compared are satellite, hybrid fiber–coax, DSL, and LMDS/MMDS. The different access technologies are described in the following text.

5.2.1. Digital Subscriber Line (DSL) Access. Digital subscriber line (DSL) is a technology that uses regular telephone lines to transmit a high volume of data at a very high speed. The telephone uses only part of the frequency available on these copper lines; DSL gets more from them by splitting the line — using the higher frequencies for data, the lower for voice and fax [21–26].

DSL offers high-speed broadband connectivity over existing copper telephony infrastructure. Upgrading copper, LECs can offer telephony and data services simultaneously.

5.2.1.1. Key Benefits of DSL

- Compared to a 56K modem, DSL speeds range from twice as fast up to 125 times as fast.

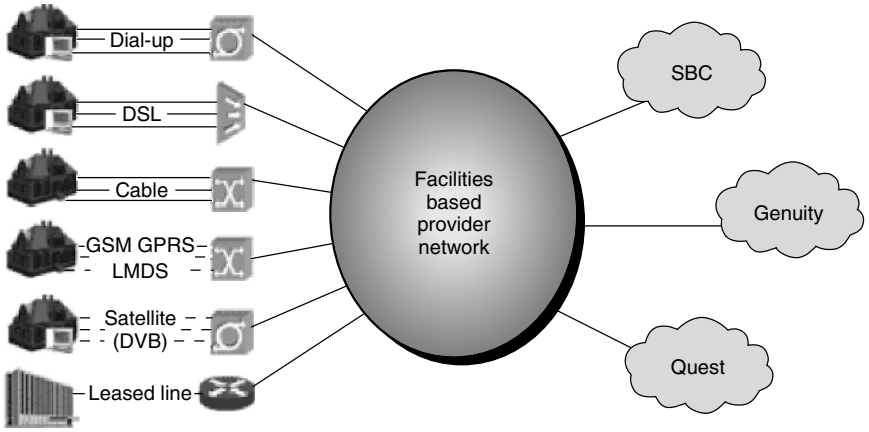


Figure 19. Broadband access options for residential service.

Table 5. Broadband Access Technologies

Access Technology	Advantages	Disadvantages
Satellite	No router hops	High installation cost
	Excellent for content distribution and management through broadcast and multicast	Adequate TCP/IP performance
	Excellent video quality	Signal propagation delay limits real-time applications
	Experiences less packet loss than terrestrial options	
Hybrid fiber–coax	Established presence in most residences — no need to draw new cables	Low security
	Good picture quality	Shared bandwidth may limit performance
	Good customer satisfaction for service	Limited enterprise market coverage
		Prone to external noise degradation
DSL	Reuse of existing copper	Access rate is function of distance from central office
	Dedicated bandwidth per user	Relatively slow rollout
		Poor customer experience during installation and service
LMDS/MMDS	Can offer broadband to areas otherwise not provided "wired" access	Subscribers must be within line of sight
	Offers higher data rates	High weather and radio interference
		Difficult to gain roof rights
		Lack of standards for Interoperability
		Higher Installation cost
Fiber to the home	Very high bandwidths and thus excellent video/TV quality	High investment and deployment costs
	Simple network architecture	Not trivial to move connections
		Hard to reach rural areas

- DSL is typically billed at a flat rate, so you can use it as much as you want without incurring more charges.
- There is no dialup — DSL makes the Internet available 24 × 7 (24 h/day, 7 days/week).
- Phone company networks are among the most reliable, with only minutes of downtime each year.
- Because the connection is made over individual phone lines, each user has a point-to-point connection to the Internet.

5.2.1.2. Types of DSL. Table 6 shows the characteristics of the DSL technologies. DSL is sometimes referred to as xDSL because there are several different variations:

Asymmetric DSL (ADSL) — delivers high-speed data and voice service over the same line. The distance from the CO determines speeds; as the distance increases, the speed available decreases.

G.Lite — variation of ADSL, a DSL that the end user can install and configure. It is not yet fully plug-and-play, and has lower speeds than full-rate ADSL.

Symmetric DSL (SDSL) — downstream speed is the same as upstream. Does not support voice connections on the same line. The distance from the CO determines speeds; as the distance increases, the speed available decreases.

ISDN DSL (IDSL) — a hybrid of ISDN and DSL; it's always an alternative to dialup ISDN. Does not support voice connections on the same line.

High-bit-rate DSL (HDSL) — the DSL that is already widely used for T1 lines. Requires 4 wires instead of the standard single pair.

Very high-bit-rate DSL (VDSL) — still in an experimental phase, this is the fastest DSL, but deliverable over a short distance from the CO.

Voiceover DSL (VoDSL) — an emerging technology that allows multiple phone lines to be transmitted over one phone wire, while still supporting data transmission. VoDSL can be used for small businesses that can balance a need for several phone extensions against their Internet connectivity needs.

Figure 20 shows an example of the ADSL network architecture.

5.2.2. Cable Access. Cable systems were originally designed to deliver broadcast television signals efficiently

Table 6. DSL Technologies

DSL Service	Data Speeds	Information
ADSL (asynchronous DSL)	Downstream: 1.5–1.8 Mbps	Most popular DSL service — based on inherent traffic flow of Internet
	Upstream: 64 kbps	Operating range <18,000 ft from CO
		Speeds are faster when close to CO
		Well suited for high-speed Internet/intranet access and telecommuter applications
ADSL Lite	Downstream: 1 Mbps	Operating range up to 18,000 ft from CO
	Upstream: 384 Kbps	Can travel over longer distances than most DSL services
		Does not require a splitter, thus limiting installation and service costs
		Consumer Internet access
SDSL (synchronous DSL)	160 kbps–2.3 Mbps	Uses one line
		Operating range <10,000 ft from CO
		Suited for videoconferencing applications and/or remote LAN access
		Business Internet access
IDSL (ISDN DSL)	144 kbps downstream and upstream	Operating range up to 18,000 ft from CO (extra equipment can increase distance)
HDSL (high-bit-rate DSL)	1.5 Mbps downstream and upstream	Uses two or four lines
		Operating range <12,000 ft from CO
		Replaces T1/E1 service
		Used primarily for PBX network connections, Internet servers, and private data networks
VDSL (very-high-bit-rate DSL)	26 Mbps symmetric	Symmetric and asymmetric configurations
	<52 Mbps asymmetric downstream	High-capacity service usually served to SOHO/SME users
	<6.4 Mbps asymmetric upstream	Capable of HDTV delivery
		Operating range 1000–4500 ft from CO
		Positioned as service of choice for eventual fiber-based all-optical networks

to subscribers' homes. The coaxial cable systems typically operate with 330 or 450 MHz of capacity, whereas hybrid fiber–coax (HFC) systems are expanded to ≥750 MHz.

Logically, downstream video programming signals begin around 50 MHz, the equivalent of channel 2 for over-the-air television signals. The 5–42 MHz portion of the spectrum is usually reserved for upstream communications from subscribers' homes. Each standard television channel occupies 6 MHz of RF the spectrum. Thus a traditional cable system with 400 MHz of downstream bandwidth can carry the equivalent of 60 analog TV channels and a modern HFC system with 700 MHz of downstream bandwidth has the capacity for some 110 channels.

To support data services over a cable network, a television channel, in the 50–750-MHz range, is typically allocated for downstream traffic to homes and other channels, in the 5–42-MHz band are used to carry upstream signals.

A head-end cable modem termination system (CMTS) communicates through these channels with cable modems located in subscriber homes to create a virtual local-area network (LAN) connection. Most cable modems are external devices that connect to a personal computer through a standard 10base-T Ethernet card or universal serial bus (USB) connection, although internal PCI modem cards are also available. The cable modem access network operates at layer 1 (physical) and layer 2 (media access control/logical link control) of the OSI Reference Model. Thus, layer 3 (network) protocols, such as IP traffic, can be seamlessly delivered over the cable modem, platform to end users.

A single downstream 6-MHz television channel may support up to 27 MHz of downstream data throughput from the cable head end using 64-QAM (quadrature amplitude modulation) transmission technology. Speeds can be boosted to 36 Mbps using 256-QAM. Upstream channels may deliver 500 kbps–10 Mbps from homes

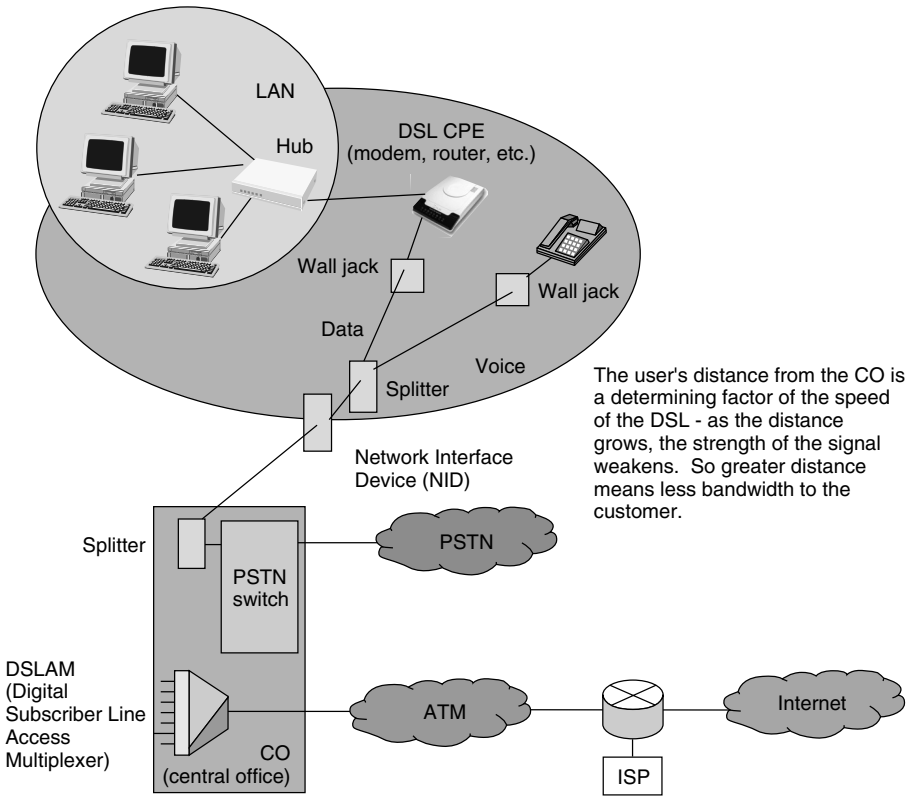


Figure 20. ADSL network architecture (using a splitter).

using 16-QAM or QPSK (quadrature phase shift key) modulation techniques, depending on the amount of spectrum allocated for service. This upstream and downstream bandwidth is shared by the active data subscribers connected to a given cable network segment, typically 500–2000 homes on a modern HFC network.

An individual cable modem subscriber may experience access speeds from 500 kbps to 1.5 Mbps or more — depending on the network architecture and traffic load — blazing performance compared to dialup alternatives. However, when surfing the Web, performance can be affected by Internet backbone congestion.

5.2.2.1. Data over Cable Service Interface Specification (DOCSIS). The DOCSIS was developed by the North American Cable Industry under the auspices of Cable Labs to create a competitive market for cable modem equipment. It was developed as a cheap Web-serving platform. The main specification work for DOCSIS 1.0 was completed in March 1997 [27].

A cable data system consists of multiple cable modems (CMs), in subscriber locations, and a cable modem termination system (CMTS), all connected by a CATV plant. The CMTS can reside in a head-end or a distribution hub. DOCSIS products have been available since 1999. The DOCSIS 1.1 version has enhanced the specification in terms of quality of service (QoS), IP multicast, and security. The DOCSIS 2.0 version has been released in 2002 downstream. The DOCSIS supports an upstream of 320 kbps–10.24 Mbps and downstream rates of 36 Mbps.

5.2.2.2. DOCSIS Cable Modem Protocol. Figure 21 shows the cable modem protocol stack. CM performs the lower four layers. It receives IP over Ethernet, adds encryption, mediates access to the return path, and modulates the data on to the cable network on the forward

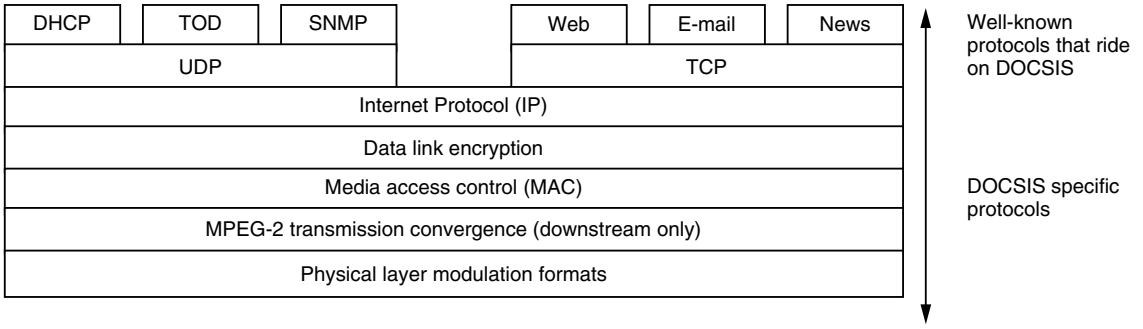


Figure 21. DOCSIS cable modem protocol stack.

path. CMTS also adds an MPEG-2 framing layer. Above the DOCSIS protocol layers is the IP layer and then the Internet applications and protocols.

5.2.3. Fixed Wireless Access. New wireless technologies such as LMDS, MMDS, and 38-GHz platforms provide very high data rates to end users. Many wireless networks utilize hybrids of wireless, satellite, fiber, and copper to create a complete end-to-end platform [20,28].

The traditional version of fixed wireless solution for point-to-point uses microwave at 1.7–40 GHz. The 38-GHz carrier service includes 14 pairs of 50-MHz wide channels available for last-mile connections. Spectrum is licensed primarily to Winstar and Advanced Radio Telecommunications. The technology is mainly used to extend fiber networks.

5.2.3.1. Local Multipoint Distribution Service (LMDS). The broadband LMDS is used to provide point-to-multipoint communications, two-way voice data, Internet access, and video services. It operates within 27.3–31.5 GHz. Cells are no larger than 2 mi in radius but require many cells to cover a large area. LMDS can serve roughly 80,000 customers from a single node. The downstream data rates top out at 38 Mbps.

5.2.3.2. Multipoint Distribution System (MMDS). MMDS is based on 200 MHz of spectrum allocated for TV transmission, increasing the flexibility for two-way communications. It allows transmission up to 1 Gbps per band for send and receive applications within a 35-mi radius. MMDS utilizes cellularization and is well suited for residential and small-business markets.

5.3. Satellite Access Networks

The access systems provide regional coverage and are less complicated than their global connected counterparts. They are more cost-effective, have less associated technical risk, and have less regulatory issues [29]. Table 7 provides a partial list of access satellite systems for regional coverage.

5.3.1. DVB Satellite Access. In this section Internet access via satellite using digital videobroadcasting–return channel by satellite (DVB-RCS) is discussed. The DVB network elements consist of an enterprise model, service-level agreements (SLA), and TCP Protocol Enhancement Proxy (PEP), the hub station, and the satellite interactive terminal (SIT). The target applications of the DVB network could be small and medium enterprises and residential users. One of the major advantages of DVB-RCS is that multicasting is possible at a low cost using the existing Internet standards. The multicast data is tunneled over the Internet via a multicast streaming feeder link from a streaming source to a centralized multicast streaming server and is then broadcasted over the satellite medium to the intended target destination group. The DVB-RCS system supports relatively large streaming bandwidths compared to existing terrestrial solutions (from 64 Kbps to 1 Mbps).

In the DVB network, *a satellite* forward and return links typically use frequency bands in Ku (12–18 GHz) and/or Ka (18–30 GHz). The return links use spot beams and the forward link global beams are used for broadcasting and Ku band. Depending on the frequency bands (Tx/Rx), three popular versions are available: (1) Ku/Ku (14/12 GHz), (2) Ka/Ku (30/12 GHz), and (3) Ka/Ka (30/20 GHz). In business-to-business applications, the SIT is connected to several user PCs via a LAN and a point-of-presence (POP) router. The hub station implements the forward link via a conventional DVB-S chain (similar to Digital TV broadcasting) whereby the IP packet is encapsulated into DVB streams, IP over DVB. The return link is implemented using the DVB-RCS standard *MF-TDMA Burst Demodulator Bank*, IP over ATM. The hub station is connected to the routers of several ISPs via a broadband access server. The hub maps the traffic of all SITs belonging to each ISP in an efficient way over the satellite. The selection of a suitable residential access technology depends on the type of application, site location, required speed, and affordable cost.

5.3.1.1. DVB-RCS. The DVB return channel system via satellite (DVB-RCS) was specified by an ad hoc ETSI

Table 7. Broadband Access Systems

Services	StarBand	WildBlue	iPStar	Astra-BBI	Cyberstar
Data uplink	38–153 kbps	384 kbps–6 Mbps	2 Mbps	2 Mbps	0.5–6 Mbps
Data downlink	40 Mbps	384 kbps–20 Mbps	10 Mbps	38 Mbps	Max. 27 Mbps
Coverage area	USA	Americas	Asia	Europe	Multiregional
Market	Consumer	Business/SME	Consumer and business	Business	ISPs, multicast
Terminal cost (U.S.$)	<$350	<$1000	<$1000	~$1800 <$450 (2001)	—
Monthly Access fee (U.S.$)	$60	$45	—	—	—
Antenna size (M)	1.2	0.8–1.2	0.8–1.2	0.5	
Frequency band	Ku	Ka	Ku, Ka	Ku/Ka	Ku, Ka
Satellite	GEO	GEO	GEO	GEO	GEO
Operation scheduled	Nov. 2000	Mid-2002	Late 2002	Late 2000	1999–2001

technical group founded in 1999. The DVB-RCS system specification in ETSI EN 301 790, v1.2.2 (2000–2012) specifies a satellite terminal (sometimes known as a *satellite interactive terminal* (SIT) or *return channel satellite terminal* (RCST) supporting a two-way DVB satellite system [30,31]. Another CDMA-based spread ALOHA has been proposed for return channel access [32]. This section describes the DVB-RCS protocol. The use of standard system components provides a simple approach and should reduce time to market.

Customer premises equipment (CPE) receives a standard DVB-S transmission generated by a satellite gateway. Packet data may be sent over this forward link in the usual way (e.g., MPE, data streaming) DVB-RCS provides transmit capability from the user site via the same antenna. The transmit capability uses a multifrequency time-division multiple access (MFTDMA) access scheme to share the capacity available for transmission by the user terminal. The return channel is coded using rate $\frac{1}{2}$ convolution FEC and Reed–Solomon coding. The standard is designed to be frequency-independent and does not specify the frequency band(s) to be used — thereby allowing a wide variety of systems to be constructed. Data to be transported may be encapsulated in ATM cells, using ATM adaptation layer 5 (AAL-5), or use a native IP encapsulation over MPEG-2 transport. It also includes a number of security mechanisms.

Figure 22 shows an example of broadband satellite network using the DVB-RCS standard for the return channel protocol.

5.3.1.2. DVB-RCS–CPE Operations. A return channel satellite terminal (RCST), once powered on, will start to receive general network information from the DVB-RCS network control center (NCC). The NCC provides monitoring and control functions, and generates the control and timing messages required for operation of the satellite network. All messages from the NCC are sent using the MPEG-2 TS using private data sections (DVB SI tables). These are transmitted over the forward link. Actually the DVB-RCS specification calls for two forward links — one for interaction control and another for data transmission. Both links can be provided using the same DVB-S transport multiplex. The term "forward link" refers to the link from the gateway that is received by the user terminal. DVB-RCS allows this communication to use the same transmission path as sued for data (i.e., the DVB-S receive path), or an alternative interaction path. Conversely, the return link is the link from the user terminal to the gateway using the DVB interaction channel. The control messages received over the forward link also provide the network clock reference (NCR).

The NCC controls user terminal transmissions. Before a terminal can send data, it must first join the network by communicating (logging on) with the NCC describing the configuration. The logon message is sent using a frequency channel also specified in the control messages. This channel is shared between user terminals wishing to join the network using the slotted ALOHA access protocol. After receiving a logon message from a valid terminal, the NCC returns a series of tables including the terminal burst time plan (TBTP) for the use terminal. The MF-TDMA burst time plan (TBTP) allows the terminal to communicate at specific time intervals using specific assigned carrier frequencies at an assigned transmit power.

The terminal transmits a group of ATM cells (or MPEG-TS packets). This block of information may be encoded in one of several ways using convolutional coding, RS/convolutional coding or Turbo coding. The block is prefixed by a preamble and optional control data and followed by a postamble to flush the convolutional encoder. The complete burst is sent using QPSK modulation. Before each terminal can use its allocated capacity, it must first achieve physical-layer synchronization of time, power, and frequency, a process completed with the assistance of special synchronization messages sent over the satellite channel. A terminal normally logs off the system when it

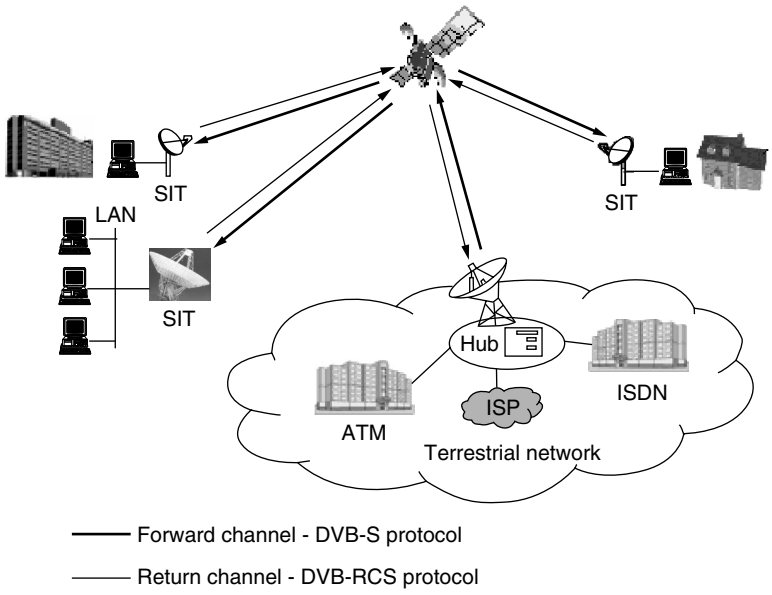


Figure 22. Broadband satellite access — DBV-RCS.

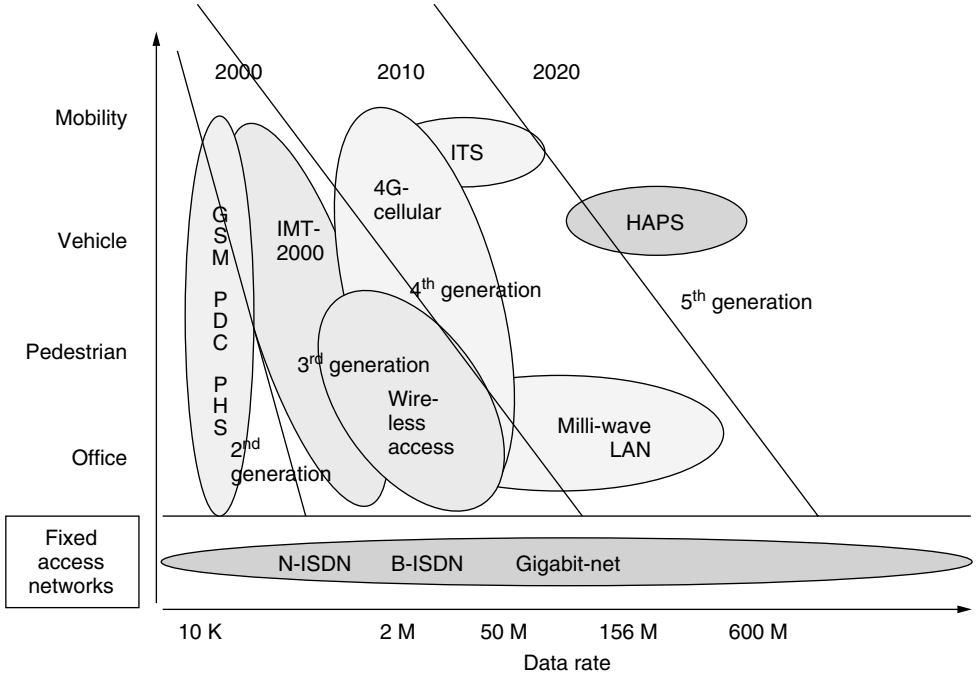


Figure 23. Future mobile wireless technologies.

has completed its communication. Alternately, if there is a need, the NCC may force a terminal to log off.

5.4. Mobile Wireless Access

Figure 23 shows a future trend of mobile communication systems. The major requirements to be supported by these systems are high data rate, high mobility, and seamless coverage. It might be difficult to realize a single system satisfying all these requirements. Some of them can provide high data rates, and others can support high mobility and coverage. However, the future intelligent integrated system solutions could satisfy all the user demands. The first-generation systems are analog cordless and analog cellular. The second-generation systems are digital systems with digital cellular such as GSM, IS54 digital cellular, personal digital cellular (PDC), and IS95. The digital cordless are DECT, PHS. These systems are operated nation wide or internationally and are today's mainstream ones. The data rates for users in air links are limited to less than several tens of kilobits per second. The IMT-2000 is the third-generation (3G) cellular systems, which provide 2 Mbps and 144 kbps indoor vehicular environments. Satellite-based Iridium and GLOBALSTAR belong to this category. As candidates of future mobile systems, 4G cellular support high data rates up to 20 Mbps and high mobility include intelligent transport systems and high-altitude stratospheric platform stations (HAPSs) [33–37].

6. STANDARDS STATUS

The standardization process is in progress for the various broadband access technologies discussed in the previous section. The different standards organizations and the fora along with their Websites are provided in Table 8.

Table 8. Technology Standards and Organizations

Technology	Standard/Organization
IP	IETF (*http://www.ietf.org*)
	ITU-T (*http://www.itu.int/ITU-T*)
	MPLS Forum (*http://www.mplsforum.org*)
FR	Frame Relay Forum (*http://www.frforum.com*)
ATM	ATM Forum (*http://www.atmforum.com*)
Cable	Cable Labs (*http://www.cablelabs.com*)
DSL	DSL Forum (*http://www.adsl.com*)
Wireless	DAVIC (*http://www.davic.org*)
Video	MPEG (*http://www.mpeg.org*)
	DVB (*http://www.dvb.org*)
	ETSI (*http://www.etsi.org*)
Broadband content delivery	BCD (*http://www.bcdforum.org*)
Satellite ATM	ITU-R (*http://www.itu.int/ITU-R*)
Satellite IP	ITU-R (*http://www.itu.int/ITU-R*)
	ITU-T (*http://www.itu.int/ITU-T*)
	IETF (*http://www.ietf.org*)
	TIA (*http://www.tiaonline.org*)
DVB-RCS	ETSI/DVB-RCS (*http://www.etsi.org*)

7. FUTURE NETWORKING: CHALLENGES

The broadband network infrastructure for the enterprise and residential access, for the emerging applications

such as videostreaming, content distribution delivery, telemedicine, two-way interactive learning, and games, must address the following challenges:

- High-speed access
- QoS
- Scalability
- Interworking
- Interoperability
- Security
- Cost-effective solutions (cost per bit)

Many of the standards organizations addressed in Section 6 are developing technical solutions and recommendations for interoperable infrastructure. One of the examples based on MPLS is discussed in the following paragraphs.

7.1. Multiprotocol Label Switching (MPLS)

Multiprotocol label switching (MPLS) is a switching method in which a label field in the incoming packets is used to determine the next hop [38,39]. At each hop, the incoming label is replaced by another label that is used at the next hop. The path thus realized is called a label-switched path (LSP). Devices that base their forwarding decision solely on the basis of the incoming labels (and ports) are called *label-switched routers* (LSRs). In MPLS, the assignment of a particular packet to a particular of forwarding equivalence classes (FECs) is done just once, as the packet enters the network. The FEC to which the packet is assigned is encoded as a short fixed-length value known as a "label." When a packet is forwarded to its next hop, the label is sent along with it; that is, the packets are "labeled" before they are forwarded.

At subsequent hops, there is no further analysis of the packet's network-layer header. Rather, the label is used as an index into a table, which specifies the next hop, and a new label. The old label is replaced with the new label, and the packet is forwarded to its next hop. In the MPLS forwarding paradigm, once a packet is assigned to FEC, subsequent routers do no further header analysis; the labels drive all forwarding. This has a number of advantages over conventional network-layer forwarding and is a great tool for traffic engineering.

Traffic engineering (TE) is concerned with performance optimization of operational networks. In general, it encompasses the application of technology and scientific principles to the measurement, modeling, characterization, and control of Internet traffic, and the application of such knowledge and techniques to achieve specific performance objectives [40,41]. A major goal of Internet traffic engineering is to facilitate efficient and reliable network operations while simultaneously optimizing network resource utilization and traffic performance. Traffic engineering has become an indispensable function in many large autonomous systems because of the high cost of network assets and the commercial and competitive nature of the Internet. All these factors emphasize the need for maximal operational efficiency, which TE can help to achieve. MPLS-based traffic engineering is a good candidate to provide hard QoS guarantees to important services such as voice over IP, video and multimedia over IP, and virtual private networks [42]. MPLS can help to maximize both resource utilization and QoS offered to a given traffic or aggregation of traffics. There are efforts under way to develop satellite over MPLS in addition to MPLS over terrestrial networks to provide QoS guaranteed solutions for both wired and wireless satellite networks.

7.2. Future MPLS Enterprise Network Example

Figure 24 shows a future MPLS-based enterprise network example.

MPLS-based VPNs are widely considered to be a viable next-generation VPN technology MPLS-VPNs were developed to simplify the VPNs without requiring encryption. Instead, MPLS used tunnels to create connectivity between sites. This allows enterprises to nail up bandwidth between sites, providing for bandwidth and connection reliability along with ensuring predictable service for different applications. However, one of the drawbacks is that the processing power required to support each MPLS-VPN becomes a bottleneck on existing monolithic routers because of their centralized router processor architecture that limit processing scalability.

In addition, all the core routers in the network must be configured to run Border Gateway Protocol (BGP). BGP is a major task for MPLS. To receive the full benefits of MPLS-VPNs, vendors must cooperate heavily while implementing the standards without prioritized methods. Some of the benefits of MPLS based VPNs are

- Service guarantees provided with respect to packet prioritization and bandwidth reservation
- Scalable infrastructure
- Seamless ATM interworking
- Small to large group connectivity

But some of the shortcomings of MPLS-VPNs are

- The inability of the enterprise to manage its network independent of the carrier
- Inability of scaling of the MPLS edge control plane

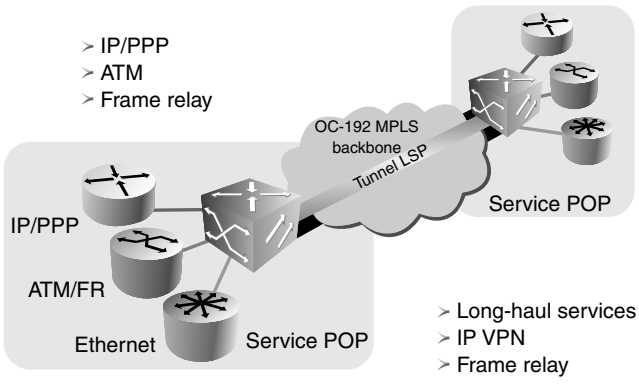


Figure 24. MPLS-based enterprise network.

- Scaling the number of virtual routers within a monolithic system

However, MPLS with traffic engineering developments provide quality of service (QoS) of the network.

8. CONCLUSIONS

Increasing demand for high-speed applications over Internet is driving new broadband network infrastructures. The applications range from simple file transfer and remote login to IP multicast, media streaming, and content delivery distribution. These emerging applications require larger bandwidths than 64 kbps or T-1 rates, and user service guarantees as opposed to "best effort" service over the today's public Internet. To meet such application requirements, different technology options are available. However, research and development in the areas of efficient protocols, QoS architectures, and security mechanisms is urgently required.

In this article, enterprise access technologies such as Gigabit Ethernet, frame relay, ATM, IP and broadband satellite technologies with network examples were provided. The broadband access technologies for residential access, DSL, cable, hybrid coax–fiber, and satellite were also discussed. Current system examples were provided. These discussions are in no way completely exhaustive. The references should provide additional resources for more depth on any single technology topic.

Finally, future networking requirements including speed, QoS, interoperability, security, and cost per bit were described with an illustrative example of an MPLS based enterprise solution.

ACRONYMS

10GMII	10-gigabit media-independent interface
AAL	ATM adaptation layer
ABR	Available bit rate
ACTS	Advanced communication technology satellite
ADSL	Asymmetric DSL
AH	Application header
ARPANET	Advanced Research Projects Agency Network
ATM	Asynchronous transfer mode
BCDF	Broadband Content Delivery Forum
BECN	Backward error congestion notification
BGP	Border Gateway Protocol
B-ISDN	Broadband Integrated Service Data Network
B2B	Business to business
CAC	Connection admission control
CBR	Constant bit rate
CDD	Content delivery distribution
CDV	Cell delay variation
CDVT	Cell delay variation tolerance
CLP	Cell loss priority
CLR	Cell loss ratio
CM	Cable modem
CMTS	Cable modem termination system
CPE	Customer premise equipment
CR	Command response
CTD	Cell transfer delay
DAMA	Demand assignment multiple access
DAVIC	Digital Audio-Visual Council
DE	Discard eligibility
DLCI	Data-link control identifier
DOCSIS	Data over Cable Service Interface Specification
DS-1	Digital signal level one
DSL	Digital subscriber line
DSLAM	Digital subscriber line access multiplexer
DVB	Digital video broadcast
DVB-S	Digital video broadcast-satellite
DVB-RCS	Digital video broadcast–return channel system
DWDM	Dense wave-division multiplexing
EA	Extended address
EH	Ethernet header
ENIAC	Electronic numerical integrator and computer
ETSI	European Telecommunication Standards Institute
FCS	Frame check sequence
FEC	Forward error correction; Forwarding equivalence class
FECN	Forward explicit congestion notification
FR	Frame relay
FSS	Fixed satellite service
FTP	File Transfer Protocol
GEO	Geostationary Earth Orbit
GigE	Gigabit Ethernet
GFC	Generic flow control
GFR	Guaranteed frame rate
GPRS	General packet radio service
GSM	Global System for Mobile Communication
GSO	Geosynchronous orbit
GW	Gateway
HAPS	High-altitude stratospheric platform station
HDTV	High-definition TV
HDSL	High-bit-rate DSL
HEC	Header error control
HFC	Hybrid fiber–coaxial cable (coax)
HTTP	Hyper Text Transfer Protocol
ICMP	Internet Control Message Protocol
IDSL	ISDN DSL
IEEE	Institute of Electronics and Electrical Engineers
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
IMT2000	International Mobile Telecommunications
IP	Internet Protocol
IPv4/IPv6	Internet Protocol version 4/Internet Protocol version 6
IPH	IP header
IPSec	Internet Protocol Security
ISDN	Integrated Services Digital Network
ISP	Internet service provider
ITU-R	International Telecommunications Union — Radio Sector

ITU-T	International Telecommunications Union — Telecommunications
LAN	Local-area Networks
Lec	Local Exchange Carrier
LMDS	Local multipoint distribution services
LSP	Label-switched plan
LSR	Label-switched router
MAC	Media access control
MAN	Metropolitan-area network
MBS	Maximum burst size
MCR	Minimum cell rate
MF-TDMA	Multifrequency time-division multiple access
MMDS	Multichannel multipoint distribution services
MPE	Multiprotocol encapsulation
MPEG	Moving Picture Expert Group
MPLS	Multiprotocol label switching
MSS	Mobile satellite service
MTU	Maximum transmission unit
NCC	Network control center
NCR	Network clock reference
NCS	Network control station
NGSO	Nongeosynchronous orbit
NNTP	Network News Transfer Protocol
nrt-VBR	Non-real-time variable bit rate
OC	Optical carrier
OSI	Open System Interconnect
PBX	Private branch exchange
PCR	Peak cell rate
PDC	Personal digital cellular
PEP	Protocol enhancement proxy
PMA	Physical medium attachment
PMD	Physical medium-dependent
POP	Point of presence
PSTN	Public switched telephone network
PT	Payload type
PVC	Permanent virtual circuit
QAM	Quadrature amplitude modulation
QoS	Quality of service
QPSK	Quadrature phase shift key
RCST	Return channel satellite terminal
RFC	Request for Comments (IETF Document)
rt-VBR	Real-time variable bit rate
SCR	Sustainable cell rate
SDSL	Symmetric DSL
SIT	Satellite interactive terminal
SLA	Service-level agreement
SME	Small- and medium-size enterprises
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SOHO	Small office/home office
SONET	Synchronous optical network
SS7	Signaling System 7
SVC	Switched virtual circuits
TBTP	Terminal burst time plan
TCP	Transmission Control Protocol
TCPH	Transmission Control Protocol Header
TDM	Time-division multiplexing
UBR	Unspecified bit rate
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunication System
UNI	User-network interface
UPC	Usage parameter control
VCI	Virtual channel identifier
VDSL	Very-high-bit-rate DSL
VOD	Video on demand
VoDSL	Voice over DSL
VPI	Virtual path identifier
VPN	Virtual private network
WAN	Wide-area network
WDM	Wavelength-division multiplexing

BIOGRAPHY

Sastri Kota has been a technical consultant with Loral Skynet, Palo Alto, California since 2001. Since the early 1970s he has held various technical and management positions and contributed to the military and commercial satellite systems in the areas of network design, broadband ATM and IP network architectures, and protocol analyses at Lockheed Martin, SRI International, Ford Aerospace, The MITRE and Computer Sciences Corp. Currently he is the U.S. Chair for ITU-R, Working Party 4B. He was the Chair for Wireless ATM Working Group and was the recipient of the ATM Forum Spotlight award. He holds a B.S in Physics, B.S.E.E, M.S.E.E from India, and Electrical Engineer's degree from Northeastern University, Boston. He has published over 90 technical papers in journals, book chapters, and conference proceedings. He was the Guest Editor for *IEEE Communications Magazine*. He has served as Satellite Communications Symposium Chair for IEEE GLOBECOM '00, Assistant Technical Chair for IEEE MILCOM '97, '90, and SPIE '91 conferences. He also served on conference technical committees and as Session Chair for IEEE GLOBECOM, ICC, WCNC, MILCOM, and AIAA communication satellite systems, in the areas of broadband, wireless, and satellite networks. His research interests include QoS for satellite IP, traffic management, wireless and mobile IP networks, and broadband access. He is a senior member of IEEE, Associate Fellow of AIAA, and member of ACM.

BIBLIOGRAPHY

1. L. G. Roberts and B. D. Wessler, Computer network development to achieve resource sharing, *Proc. AFIPS Conf.* 1970, Vol. 36, pp. 543–549.
2. L. Kleinrock, *Queuing Systems,* Vol. 2*; Computer Applications*, Wiley, New York, 1976.
3. Renters Story Network Fusion, April 2002.
4. S. Aidarous, Challenges in evolving to IP-based global networks, *GLOBECOM 2001*, San Antonio, TX, Nov. 25–29, 2002.
5. S. Kota, *Quality of Service* (*QoS*) *Architecture for Satellite IP Networks*, Document 4B/86, ITU-R Working Party 4B, Geneva, Switzerland, 2002.
6. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2000.

7. R. Ramaswami and K. Sivarajan, *Optical Networks: A Practical Perspective*, Academic Press/Morgan Kaufman, 1998.
8. S. Saunders, *Data Communications Gigabit Ethernet Handbook*, McGraw-Hill, New York, 1998.
9. P. Smith, *Frame Relay, Principles and Applications*, Addison-Wesley, Reading, MA, 1993.
10. Frame Relay Forum, *http://www.frforum.com*.
11. H. J. R. Dutton and P. Lenhard, *Asynchronous Transfer Mode (ATM): Technical Overview*, Prentice-Hall, Saddle River, NJ, 1995.
12. D. E. McDysan and D. L. Spohn, *ATM Theory and Application*, McGraw-Hill Series on Computer Communications, McGraw-Hill, New York, 1998.
13. The ATM Forum, *Traffic Management Specification*, version 4.0, 1996.
14. W. R. Stevens, *TCP/IP Illustrated,* Vol. 1*, The Protocols*, Addison-Wesley, Reading, MA, 1994.
15. S. Deering and R. Hinden, *Internet Protocol Version 6* (IPv6) Specification, IETF, RFC 2460, Dec. 1998.
16. S. Kent and R. Atkinson, *Security Architecture for the Internet Protocol*, Internet RFC 2401, 1998.
17. G. Maral and M. Bousquet, *Satellite Communications Systems, Systems, Techniques and Technology*, Wiley, New York, 1993.
18. S. Kota, A. Durresi, and R. Jain, Realizing future broadband satellite network services, in *Modeling and Simulation Environment for Satellite and Terrestrial Communication Networks*, A. Nejat Ince, ed., Kluwer, Boston, 2002.
19. Pioneer Consulting, *Broadband Satellite: Analysis of Global Market Opportunities and Innovation Challenges*, 2002.
20. R. J. Bates, *Broadband Telecommunications Handbook*, McGraw-Hill, New York, 2000.
21. R. W. Smith, *Broadband Internet Connections: A User's Guide to DSL and Cable*, Addison-Wesley, Reading, MA, 2002.
22. W. J. Goralski, *ADSL*, McGraw-Hill, New York, 1998.
23. G. Abe, *Residential Broadband*, CISCO Press, 1997.
24. I. Cooper and M. A. Bramhall, ATM passive optical networks and integrated VDSL, *IEEE Commun. Mag.* **38**(3): 174–179 (2000).
25. A. Azzam, *High-Speed Cable Modems*, McGraw-Hill, New York, 1997.
26. G. Held, *Next-Generation Modems: A Professional Guide to DSL and Cable Modems*, Wiley, New York, 2000.
27. D. Fellows and D. Jones, DOCSIS™ cable modem technology, *IEEE Commun. Mag.* **39**(3): 202–209 (2001).
28. J. R. Vacca, *Wireless Broadband Networks Handbook*, McGraw-Hill, New York, 2001.
29. A. Durresi and S. Kota, Satellite TCP/IP, in *High Performance TCP/IP*, M. Hassan and R. Jain, eds., Prentice-Hall, Englewood Cliffs, NJ, 2002.
30. *DVB, Interactive Channel for Satellite Distribution Systems*, DVBRCS001, rev. 14, ETSI EN 301 790, V1.22 (2000–2012).
31. J. Neale, R. Green, and J. Landovskis, Interactive channel for multimedia satellite networks, *IEEE Commun. Mag.* **39**(3): 192–198 (2001).
32. S. Kota, M. Vazquez-Castro, and J.Carlin, Spread ALOHA multiple access for broadband satellite return channel, *Proc. AIAA 20th Int. Communication Satellite Systems Conf.*, 2002.
33. S. Ohmori, Y. Yamao, and N. Nakayima, The future generations of mobile communications based on broadband access methods, *Int. J. Wireless Pers. Commun.* **17**(2–3): 175–190 (2001).
34. C. Perkins, *IP Mobility Support*, Internet RFC 2002, 1996.
35. J. Rapeli, Future directions for mobile communications—business, technology and research, *Int. J. Wireless Pers. Commun.* **17**(2–3): 155–173 (2001).
36. K. Pahlavan and P. Krishnamurthy, *Principles of Wireless Networks—a Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 2002.
37. G. Wu, M. Mizuno, and P. J. M. Havinga, MIRAI architecture for heterogeneous network, *IEEE Commun. Mag.* **40**(2): 126–134 (2002).
38. A. Ghanwani et al., Traffic engineering standards in IP network using MPLS, *IEEE Commun. Mag.* **37**(12): 49–53 (1999).
39. G. Swallow, MPLS advantages for traffic engineering, *IEEE Commun. Mag.* **37**(12): 54–57 (1999).
40. E. W. Gray, *MPLS: Implementing Technology*, Addison-Wesley, New York, 2001.
41. B. Davie and Y. Rekhter, *MPLS Technology and Applications*, Morgan Kaufman, San Francisco, 2000.
42. G. Armitage, *Quality of Service in IP Networks, Foundations for a Multi-Service Internet*, MTP, Indianapolis, IN, 2000.

TRENDS IN WIRELESS INDOOR NETWORKS

K. Pahlavan
J. Beneat
X. Li
Center for Wireless Information Network Studies
Worcester Polytechnic Institute
Worcester, Massachusetts

1. INTRODUCTION

This article provides an overview of the trends in wireless indoor networks. Since the late 1970s, when the concept of wireless LAN was first introduced, this area has gone through significant ups and downs. After a very exciting development period in the late 1980s, market revenues still remained far below predictions in the mid-1990s. However, during the late 1990s with the introduction of wireless personal area networking, Bluetooth technology, and home networking, a new surge of excitement has emerged in the wireless indoor communication industry, resulting in a number of new startup projects. Furthermore, the emergence of wireless indoor positioning systems to augment wireless indoor telecommunication services has attracted tremendous attention to wireless indoor networks. This chapter starts with an overview of the traditional wireless LAN industry and then moves to explain the new emerging wireless

personal-area network (WPAN), home networking, and indoor geolocation industries.

2. WIRELESS LANs

Since 1980 the perception of a WLAN industry has evolved. It was implemented on a variety of innovative technologies and raised great hopes for developing a sizable market a couple of times. Today, the major differentiation of WLANs from wide area cellular services is the method of delivery to the users, data rate limitations, and frequency band regulations. Cellular data services are delivered by operating companies as services while WLAN users own their network. At a time where the 3G cellular industry is striving for 2-Mbps (megabit/second) packet data services, WLAN standards are focusing on 54-Mbps services. Another differentiation with other radio networks is that, today, almost all WLANs operate in unlicensed bands where frequency regulations are loose and there is no charge or waiting time to obtain the band. To obtain a deeper understanding of all these issues, it is very useful to go over the history of the WLAN industry to see how all these unique issues evolved.

2.1. Early Experiences

The idea of a wireless LAN was first introduced by Gfeller at the IBM Rueschlikon Laboratories in Switzerland in the late 1970s [1]. The number of terminals in manufacturing floors was growing and wiring within the manufacturing floor was difficult. In offices, wires are normally snaked under the suspended ceilings and through the interior partitioning wall. This was not possible in manufacturing floors. In offices, in extreme cases, the wiring could be installed under the floor or simply laid on the floor with some cover. In manufacturing floors, the floors are more rugged, making underfloor wiring more expensive, while throwing wires over the floor is not acceptable because of the danger presented by moving heavy machinery. Diffused IR technology was selected for the implementation of the WLAN. Diffused IR avoided interference problems due to electromagnetic signals radiating from the machinery and avoided dealing with cumbersome administrative procedures from frequency administration agencies. Unfortunately, the principal researcher abandoned the project when the goal of 1 Mbps with reasonable coverage did not materialize.

At about the same time, a second noticeable project on WLANs was performed by Ferert at Hewlett-Packard Pal Alto Research Laboratories in California [2]. In this project, a 100-kbps direct-sequence spread-spectrum (DSSS) WLAN operating at 900 MHz using a carrier sense multiple-access (CSMA) access method technique was developed for office areas. The project was conducted under an experimental license agreement from the FCC. However, when Ferert filed to obtain bands from the FCC, he was discouraged by the administrative complexity of securing a band for his application and he also abandoned the project. A couple of years later, Codex/Motorola attempted to implement a WLAN at 1.73 GHz, but the project was also dropped during negotiations with the FCC.

Although all the pioneering WLAN projects were abandoned, the area continued to attract attention and negotiations with the FCC to secure bands continued [3]. These projects revealed several important challenges facing the WLAN industry that still remain:

1. *Complexity and cost* — WLAN implementation alternatives using IR, spread-spectrum communication, or traditional radios are far more complex and diversified than those in wired LANs.
2. *Bandwidth* — data rate limitations in a wireless medium are far greater than in a wired medium.
3. *Coverage* — point-to-point coverage of a WLAN operating in a building is smaller than cables or even twisted-pair (TP) LAN solutions.
4. *Interference* — WLANs are subject to interference from other overlaying WLANs or other devices operating in the same frequency medium.
5. *Frequency administration* — radiofrequency-based WLANs are subject to expensive and timely frequency regulations.

2.2. Emergence of Unlicensed Bands

Wireless LANs need a significant amount of bandwidth, at least several tens of MHz. Yet, in the 1980s it had not been shown that a strong market existed, comparable to that of the cellular voice industry when it originally started with two 25-MHz bands. In addition, frequency bands of comparable bandwidth for PCS applications were auctioned in the United States for tens of billions of dollars while the WLAN market was under a billion dollars per year. The dilemma for the frequency administration agencies was how to justify this frequency allocation. In the mid-1980s, the FCC found two solutions to this problem. The first and simplest solution was to go beyond the 1–2-GHz band used for cellular telephone and PCS applications to higher frequencies at several tens of GHz where plenty of unused bands were available. This solution was first negotiated between Motorola and the FCC, resulting in Altair, the first wireless LAN product operating in a licensed 18–19-GHz band. Motorola also established a headquarter to facilitate negotiations with the FCC regarding the usage of WLANs in different locations. If the location of operation of a WLAN was substantially changed (e.g., from one town to another), those responsible for the network would contract Motorola to manage the necessary frequency administration issues with the FCC.

The second solution used a more innovative approach to the problem by resorting to the creation of unlicensed bands. In response to the pressing need for suitable bands and motivated by recent studies depicting various implementations of wireless LANs [3], Mike Marcus of the FCC initiated the release of the unlicensed ISM bands (902–928/2400–2483/5725–5875 MHz) in May 1985 [4]. The ISM bands were the first unlicensed bands for consumer product development and played a major role in the development of the WLAN industry. In simple words, licensed and unlicensed bands are compared to backyard and public gardens. Anyone who can afford it

can own a private backyard (licensed band) and arrange a barbeque dinner (a wireless product). If one cannot afford to buy a house with a backyard, he/she simply moves the barbeque party to the public park (unlicensed band) where he/she should observe certain rules or etiquette that allows others to share the public resource as well. The rules enforced on ISM bands restricted the transmission power to 1 W. Modems radiating more than 1 mW had to employ spread-spectrum technology. It was perceived that spread-spectrum communication would limit interference and allow the coexistence of several wireless applications in the same band.

Encouraged by the FCC ruling [4] and some visionary publications in wireless office information networks [3,5,6], a number of WLAN product development projects mushroomed in the North American continent. By late 1980s, the first generation of WLAN products appeared in the market. These products used three different technologies: microwave technology in the licensed 18–19-GHz band, spread spectrum in the ISM bands, and IR. They were shoebox-size access points and receiver boxes. The perception at that time was that a WLAN would be used to connect workstations to the LAN wherever wiring difficulties justified using a more expensive wireless solution. Today, we call this application *LAN extension* [7,8]. At that time, market predictions were estimating a shift of around 15% of the LAN market to WLAN that would generate a few billion dollars of sales per year by the early the 1990s.

Also in the late 1980s, a standardization activity was initiated under the IEEE 802.4L to provide some guidance for development of WLANs. This activity soon turned to the IEEE 802.11 standard, but it would take many years, up until 1997, for the standard to be finalized. In May 1991, to create a scientific forum for the exchange of knowledge on WLANs, the first IEEE sponsored WLAN workshop was organized concurrent to the 802.11 meeting at Worcester Massachusetts [9].

In 1992, following the momentum for WLAN developments, an industrial alliance led by Apple Computers called WINForum was formed aiming at obtaining more unlicensed bands from the FCC for the so-called Data-PCS activities. WINForum finally succeeded in securing 20 MHz of bandwidth in the PCS bands that was divided into two 10-MHz bands, the 1910–1920 MHz band for *isochronous* (voicelike) and the 1920–1930-MHz band for *asynchronous* (data type) applications. The original aim of the WINForum was to secure 40 MHz for Asynchronous applications. WINForum defined a set of rules or etiquettes for these bands to allow coexistence. There are three basic rules: (1) to listen before talk (or transmit) or LBT protocol, (2) to use low transmitter power, and (3) to restrict the duration of the transmissions. The WINForum etiquette is based on CSMA rather than CDMA spread-spectrum communications. This was a better choice since CDMA implementations require careful power control schemes that are not feasible in an uncoordinated multiuser, multivendor WLAN environment. However, spread-spectrum communications without CDMA are less bandwidth-efficient.

Another standardization activity that started in 1992 was HIPERLAN. This ETSI-based standard aimed at high-performance WLANs with data rates up to 20 Mbps, an order of magnitude higher than the original 802.11 data rates of 2 Mbps. To support these data rates, the HIPERLAN community was able to secure two 200-MHz bands at 5.15–5.35 GHz and 17.1–17.3 GHz. This initiation encouraged the FCC to release the so-called *unlicensed national information infrastructure* (U-NII) bands in 1997 at the time the original HIPERLAN now called HIPERLAN-1 was completed. Table 1 summarizes the U-NII bands and their restrictions.

Today, IEEE 802.11a and HIPERLAN-2 projects use the U-NII bands for the implementation of 54-Mbps OFDM-based WLANs. More details of the WLAN standards can be found in Section 2.5 and Table 2.

2.3. Shift in Marketing Strategy

In the first half of the 1990s, a sizable market of around a few billions of dollars per year was expected for the shoebox-type products used as LAN-extensions in indoor areas. This did not materialize. As a consequence, two new directions for product development emerged. The first and simplest approach was to take the existing shoebox-type WLAN products, boost the transmitted power to the maximum authorized limit, and add a high-gain directional antenna for outdoor interbuilding LAN interconnects. This technically simple solution would allow wireless connectivity up to a few tens of kilometers with suitable rooftop antennas. The new inter-LAN wireless bridges could connect corporate LANs within range at a much lower cost than the wired alternative such as T1 carrier lines leased from PSTN service providers. The second approach was to reduce the size of the product to a PCMCIA WLAN card suitable for the laptops that

Table 1. The U-NII-bands

Band of Operation (GHz)	Maximum Transmission Power (mW)	Maximum Power with Antenna Gain of 6 dBi (mW)	Maximum PSD (mW/MHz)	Applications: Suggested and/or Mandated	Other Remarks
5.15–5.25	50	200	2.5	Restricted to indoor applications	Antenna must be an integral part of the device
5.25–5.35	250	1000	12.5	Campus LANs	Compatible with HIPERLAN
5.725–5.825	1000	4000	50	Community networks	Longer range in low-interference (rural) environments

Table 2. Summary of WLAN Standards

Parameters	IEEE 802.11	IEEE 802.11b	IEEE 802.11a	HIPERLAN/2	HIPERLAN/1
Status	Approved, products	Final ballot, Nov. 1999, products	In preparation	In preparation	Approved, no products
Frequency band	2.4 GHz	2.4 GHz	5 GHz		5 GHz
PHY, modulation	DSSS: BPSK, QPSK FHSS: GFSK	DSSS: BPSK, QPSK, CCK	OFDM		GMSK
Delay spread robustness	1, 2 Mbps: 200–400 ns 11 Mbps: 20–60 ns Fallback mechanism		12 Mbps: 350 ns 36 Mbps: 125 ns		Unknown
Data rate	1, 2 Mbps	1, 2, 5.5, 11 Mbps	6, 9, 12, 18, 24, 36, 54 Mbps		23.5 Mbps
Access method	Distributed control, CSMA/CA or RTS/CTS			Central control reservation based-access, scheduled by access point	Active contention resolution, priority signalling

were enjoying a sizable growth. However, this approach was mainly suitable for the spread spectrum products operating at lower frequencies. Figure 1 illustrates all three applications for WLANs.

The early marketing strategy used by companies for LAN-extension aimed at a horizontal market by selling individual WLAN pieces directly to the customers. In the mid-1990s, a few successful companies adopted a major shift in marketing strategy. The new strategy aimed at a vertical market by selling the entire wireless network as a complete solution. The vertical markets approached by the WLAN industry were *"bar code" industries* providing wireless inventory check and tracking in warehouses and manufacturing floors, *financial services* providing wireless financial updates in large stock exchanges, *healthcare* networks providing wireless mobile services inside hospitals, and *wireless campus-area networks* (WCAN) providing wireless classrooms and offices. All these efforts boosted the market for WLANs to over half a billion dollars per year during the late 1990s.

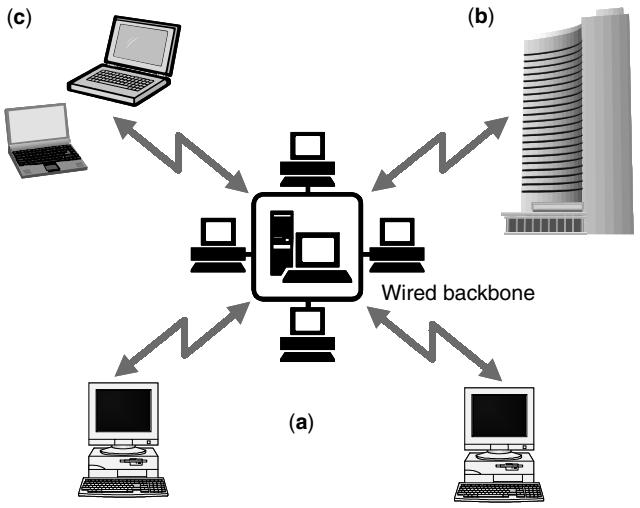


Figure 1. Different forms of WLAN products: (**a**) LAN extension; (**b**) Inter-LAN bridge; (**c**) PCMCIA cards for laptops.

An experimental NSF-sponsored WCAN is represented in Fig. 2. It was designed as a testbed for performance monitoring of WLAN products at Center for Wireless Information Network Studies (CWINS), Worcester Polytechnic Institute (WPI) in 1996. The testbed connects five buildings with inter-LAN bridges using different technologies. Inside each building access points provide coverage to the laptops that are carried by the students. The professor broadcasts his/her image and writings on the electronic board to allow students to participate in the wireless classroom from different buildings on the campus. The entire wireless network is connected to the backbone through a router to isolate the traffic for traffic monitoring experimentations.

Today, horizontal markets for the WLAN industry mainly focus on WLAN as an alternative to LANs wherever the additional cost of the wireless solution is justifiable. This occurs for example in installations with frequent relocations where the additional cost of the WLAN solution is justified by the relocation costs of the wired solution. Temporary networking such as registration sites at conferences or fairs (jobs, food, etc.) is another example where the wireless solution is preferred to the less expensive and more reliable wired alternative. Buildings with difficult or impossible-to-wire situations, such as marble buildings or historical monuments where drilling for wiring is not favored, provides another example where WLAN is justifiable. The most prominent incentive for WLANs in vertical markets is the general use of laptops at home and in the offices.

2.4. New Interest from Military and Service Providers

In the mid-1990s, when the WLAN industry was struggling to find a market, a new wave of interest for WLAN came from the U.S. Department of Defense for

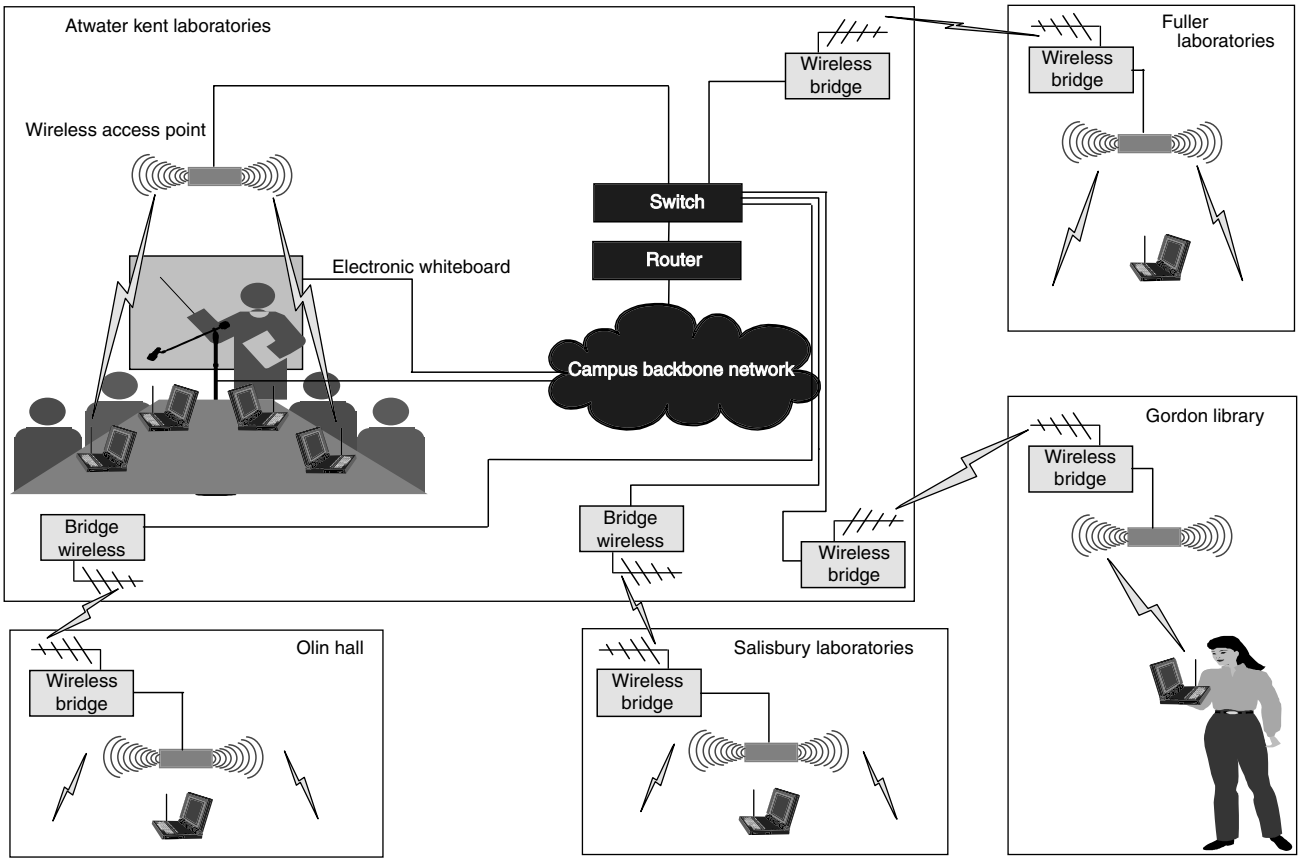


Figure 2. The experimental NSF sponsored WCAN at WPI.

military applications and from the European Community for commercial applications. These projects poured a considerable amount of research investments that further brightened the future of this industry [10].

The incentive for the military was to discover new horizons for implementation of global mobile military networks that support integration of computing and positioning systems. For instance, the InfoPAD project at the University of California, Berkeley [11] was one of the early WLAN DARPA projects. The environment is like a battleship equipped with a number of computing facilities. Soldiers in the environment are carrying InfoPADs that are small asymmetric communication devices carrying user instructions to the computing backbone to initiate computational operations whose results are downloaded to the PAD. A main challenge in this project was the implementation of reasonable size PADs capable of supporting multimedia applications. BodyLAN [12] was another DARPA-sponsored WLAN project initiated at BBN, Cambridge, Massachusetts. This project intended to design a low-power network capable of monitoring vital human body condition information (heartbeat, temperature, etc.) and communicating this information to other soldiers in the proximity. A more recent DARPA project was the Small Unit Operations/Situation Awareness Systems (SUO/SAS). The goal was to design an integrated telecommunication and geolocation network for modern fighting scenarios. The technical challenges included providing accurate indoor position information [13] and communicating situation awareness information to the war fighters. This system was expected to provide a full communication and positioning link to the soldier operating inside a building.

The commercial interest from the European Community (EC) was initiated by the equipment manufacturers seeking solutions for the service providers that were keen on incorporating higher data rate services into the evolving rich cellular industry. In the mid-1990s, both commercial service providers and military network designers believed that the future of backbone networks would be an end-to-end ATM based network. In response to this perception, wideband local networking industries initiated the wireless ATM movement.

From the application point of view, service providers intend to integrate WLAN products into their existing services. A popular scenario used in HIPERLAN-2 to represent the service providers point of view is shown in Fig. 3. In this scenario it is assumed that a WLAN user carries his/her laptop in the office, home, and in public places (airports, train stations, etc.). In the home and office, the laptop connects to a free network whose infrastructure is owned by the user or his/her company. In public places, such as an airport or other transit buildings, WLAN access points belonging to a service provider can provide high-speed access. A wide-area backbone wireless network could also provide the

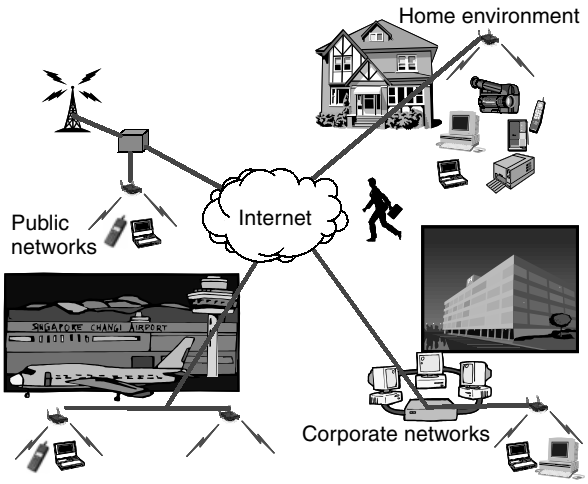


Figure 3. The service provider's view of LANs.

connection but at lower data rates. In all public places the service provider that owns the infrastructure will charge the user. One of the technical challenges for the implementation of this scenario is the vertical roaming among different networks [14]. Another challenge is to incorporate roaming and tariff mechanisms for WLANs. These issues are currently under investigations.

2.5. A New Explosion of Market and Technology

During 1998 and 1999, interest in WLANs exploded. The WLAN industry that relied almost exclusively on a North American market with a market only a fraction of that of the cellular industry, suddenly attracted widespread attention from Japan and the EC and a renewed interest in the United States. There is now growing hope for a WLAN market comparable to that of the cellular industry. In Japan, small office spaces promoted the popular usage of laptops to replace PCs. The natural networking solution for laptops was nothing except WLANs. In the EC, the rich cellular industry started considering WLANs as part of their next generation high-speed packet data services. The interest is twofold. WLANs provide a practical higher-speed solution and operate in unlicensed bands free of charge while the cost of licensed bands is constantly increasing. In the North American continent, the successful growth of broadband Internet access to the homes opened a new window for a sizable market in home networking. This perception trend for a new sizable market was strengthened by the emergence of new low-power personal-area ad hoc wireless networking technologies such as Bluetooth and ultrawideband (UWB) for local distribution, LMDS for home access, and indoor positioning for a variety of applications. The availability of low-power, low-cost wireless chip sets started a new revolution in consumer product development, raising hope for sales in the order of hundreds of millions of chip sets per year. All together these hopes initiated a *Gold Rush* in chip manufacturing for WLAN and WPAN applications that is still ongoing. Technical directions for this industry remain as providing for higher data rates, more comprehensive coverage, less interference, and lower cost.

2.6. WLAN Standards

Wireless LANs provide very high data rates ($\geq$1 Mbps) in local areas ($<$100 m) for access to wired LANs and high speed Internet. Today, all successful wireless LANs operate in unlicensed bands that are free of charge and rigorous regulations. Sine the late 1990s, considering that PCS bands were auctioned at very high prices, wireless LANs have attracted renewed attention.

Table 2 provides a summarizes the IEEE 802.11 and HIPERLAN standards for wireless LANs. IEEE standards include 802.11 and 802.11b operating at 2.4 GHz and 802.11a operating at 5 GHz. Both HIPERLAN-1 and -2, developed under ETSI, operate at 5 GHz. The 2.4-GHz products use spread-spectrum technology to support data rates ranging from 1 to 11 Mbps. HIPERLAN-1 uses GMSK modulation with DFE signal processing at the receiver and supports up to 23.5 Mbps. IEEE 802.11a and HIPERLAN-2 use an OFDM physical layer to support up to 54 Mbps. The access method for all 802.11 standards is the same and includes CSMA/CA, point coordination function (PCF), and request to send (RTS)/clear to send (CTS).

The access method for HIPERLAN-1 is comparable to 802.11, but the access method for HIPERLAN-2 is a voice-oriented access technique suitable for integration of voice and data services. IEEE 802.11, IEEE 802.11b, and HIPERLAN-1 are completed standards, and IEEE 802.11 and 11.b are today's dominant products in the market. IEEE 802.11a and HIPERLAN-2 are still under development. The IEEE 802.11 and HIPERLAN standards can be considered as the second generation of wireless LANs, while OFDM wireless LANs are forming the next generation of these products.

3. WPANs

3.1. Introduction

The first announced personal-area network (PAN) was the BodyLAN that emerged from a DARPA project in the mid-1990s. It was specified as a low-power, small-size, inexpensive, modest-bandwidth solution that could connect personal devices in many collocated systems within a range of $\sim$5 f [12]. Motivated by the BodyLAN project, a WPAN group originally started in June 1997 as a part of the IEEE 802.11 standardization activities. In January 1998, the WPAN group published the original functionality requirements. In May 1998, the study group invited participation from several related groups such as WATM, Bluetooth, HomeRF, BRAN (HIPERLAN), IrDA (IR short-range access), IETF (Internet standardization), and WLANA (a marketing alliance of WLAN companies in the United States). Only the HomeRF and Bluetooth groups responded to the invitation. In March 1998, the Home RF group was formed. In May 1998, the Bluetooth development was announced and a Bluetooth special group was formed within the WPAN group [15]. In March 1999, the IEEE 802.15 was approved as a separate group in the 802 community to handle WPAN standardization. At the time of this writing, IEEE 802.15 WPAN has four subcommittees on Bluetooth, coexistence, high data rate, and low data rate.

3.2. What Is IEEE 802.15 WPAN?

The 802.15 WPAN group is focused on development of short-distance wireless networks used for networking of portable and mobile computing devices such as PCs, personal digital assistants (PDA), cellular phones, printers, speakers, microphones and other consumer electronics. The WPAN group intends to publish standards that allow these devices to coexist and interoperate with one another and other wireless and wired networks in an internationally acceptable frequency band of operation.

The original functional requirement published in January 22, 1998 was based on the BodyLAN project and specified devices with [15]:

- Power management: small current consumption
- Range: 0–10 m
- Speed: 19.2–100 kbps (actual)
- Small size: ~0.5 in.^{3}, no antenna
- Low cost: relative to target device
- Should allow overlap of multiple networks in the same area
- Networking support for a minimum of 16 devices

These specifications well fit the Bluetooth specifications, a technology announced after this premier announcement. The initial activities in the WPAN group included both the HomeRF and Bluetooth groups, but today HomeRF maintains its own Website at *www.homerf.org*. IEEE 802.15 WPAN includes four taskgroups. The first taskgroup is based on Bluetooth and aims to define PHY and MAC specifications for wireless connectivity among fixed, portable, and moving devices within or entering a *personal operating space* (POS), the space about a person or object that typically extends up to 10 m in all directions and envelops the person whether stationary or in motion. This taskgroup will address quality of service to support a variety of traffic classes.

The second taskgroup focuses on coexistence between WPAN and 802.11 WLANs. This group is developing a coexistence model to quantify the mutual interference and a coexistence mechanism to facilitate coexistence between an IEEE 802.11 WLAN and an IEEE 802.15 WPAN device. One goal of the WPAN group is to achieve a sufficient level of interoperability between a WPAN and an 802.11 device to allow transfer of data between the two devices.

The third taskgroup works on PHY and MAC layer specifications for high rate (HR) WPANs with data rates higher than 20 Mbps. This standard will provide for low-power, low-cost solutions that address the needs of portable consumer digital imaging and multimedia applications. This standard aims at providing compatibility with the Bluetooth specification of the taskgroup one. This standard is expected to be completed by early 2002.

The fourth taskgroup is chartered to investigate ultra-low complexity, ultra-low-power consumption, ultra-low-cost PHY/MAC-layer specification to support data rates of up to 200 kbps. Potential applications are sensors, interactive toys, smart badges, remote controls, and home automation. This taskgroup may also address location tracking capabilities required to support the use of smart tags and badges.

3.3. What Is Home RF?

According to the standard committee, the mission of the HomeRF working group is to provide the foundation for a broad range of interoperable consumer devices by establishing an open industry specification for wireless digital communications between PCs and consumer electronic devices anywhere in and around the home.

Figure 4 represents the overall vision of HomeRF. The architecture can support both ad hoc and connected networks. In a popular home setup, the Internet access and PSTN connection arrives at a control HomeRF distribution box that can support HomeRF wireless as well as HPNA networks. The HomeRF wireless network can accommodate isochronous clients interconnecting up to six cordless telephone devices and asynchronous clients interconnecting a number of data devices. The two major competitors for HomeRF are HIPERLAN-2 and Bluetooth. As compared with HIPERLAN-2, the HomeRF solution provides smaller data rates ($\leq$2 Mbps as opposed to 54 Mbps in HIPERLAN-2) that cannot support wireless transmission of video for TV and VCR applications. Compared with Bluetooth, HomeRF provides higher data rates, but Bluetooth was introduced as an inexpensive chip set that early on attracted a large alliance.

The HomeRF workgroup has developed a specification for wireless communications in the home called *shared wireless access protocol* (SWAP). The SWAP specification defines a new common interface to support wireless voice and data networking in the home. The SWAP specification is an extension of DECT (TDMA) for voice, and a relaxed 802.11 (CSMA/CA) for high-speed data applications.

The reader interested in more details on HomeRF is referred to Ref. 16.

3.4. What Is Bluetooth?

Bluetooth is an open specification for short-range wireless voice and data communications that was originally developed for wire replacement in personal area networking to operate all over the world. In 1994, the initial study for development of this technology started at Ericsson, Sweden. In 1998, Ericsson, Nokia, IBM, Toshiba, and Intel formed a special-interest group (SIG) to expand the concept and develop a standard under IEEE 802.15 WPAN. In 1999, the first specification, v1.0b, was released and then accepted as the IEEE 802.15 WPAN standard for a 1-Mbps network. At the time of this writing over 1000 companies participate as members in the Bluetooth SIG, and a number of companies all over the world are developing Bluetooth chip sets. Marketing forecasts indicate penetration of Bluetooth in more than 100 million cellular phones and several million of other communication devices. The IEEE 802.15 is also studying coexistence and interference between Bluetooth and IEEE 802.11 products operating at 2.4 GHz.

The story of the origin of the name Bluetooth is interesting and worth mentioning. "Bluetooth" was the nickname of Harald Blaatand (940–981 A.D.), King of

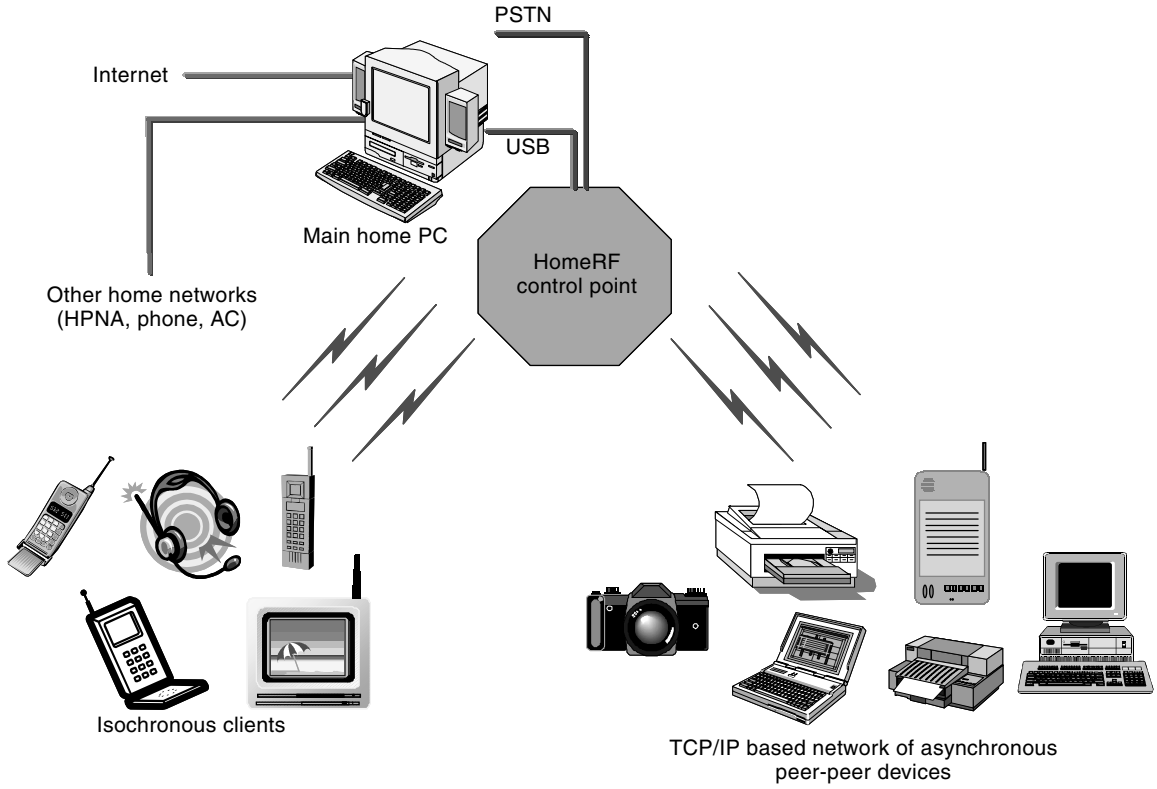


Figure 4. Overview of the HomeRF vision.

Denmark and Norway. When Bluetooth was introduced to the public, a stone carving erected from Harald Blaatand's capital city Jelling was presented [17]. This strange carving was interpreted as Bluetooth connecting a cellular phone and a wireless notepad in his hands. This picture was used to symbolize the vision of using "Bluetooth" to connect personal computing and communication devices. Bluetooth, the king, was also known as a peacemaker and a person who brought Christianity to Scandinavians to harmonize their beliefs with the rest of Europe. This fact is used to symbolize the need for harmony among manufacturers of WPANs around the world to support the growth of the WPAN industry.

Bluetooth is the first popular technology for short-range ad hoc networking designed for integrated voice and data applications. As compared with WLANs, Bluetooth has a lower data rate but has an embedded mechanism to support voice applications. As compared with 3G cellular systems, Bluetooth is an inexpensive personal-area ad hoc network operating in unlicensed bands and owned by the user.

The Bluetooth SIG considers three basic application scenarios [17]. The first scenario is the wire replacement to connect a personal computer or laptop to the keyboard, mouse, microphone, and notepad. As the name indicates this avoids the problem of multiple short-range wirings surrounding today's personal computing devices. The second scenario is an ad hoc network of several different users in a very short range of each other such as in a conference room. WLAN standards and products are also commonly considered for this scenario. The third scenario considers Bluetooth access points to redistribute wide-area voice and data services provided by cellular networks, wired connections or satellite links, in a fashion similar to that of the WLAN scenario in an airport. Contrary to 802.11 however, Bluetooth has provisions for both voice and data and can be used as an integrated voice/data access point to connect to both voice and data backbone infrastructures. The HIPERLAN-2 standard will provide a more expensive version of similar connections that supports a larger number of users and higher data rates.

The topology of a Bluetooth network is referred to as *scattered ad hoc topology*. In a scattered ad hoc environment a number of small networks, each supporting a few terminals, can coexist and possibly interoperate with each another. Bluetooth specifications have been selected to operate in the unlicensed ISM bands at 2.4 GHz. The advantage is the worldwide availability of the bands. The disadvantage is the existence of other users, in particular IEEE 802.11 and 802.11b products in the same band. At the time of this writing, a subcommittee of the IEEE 802.15 is working on the interference issues related to Bluetooth and IEEE 802.11 and 802.11b.

4. HOME NETWORKING

In a house, the number of devices that are connected together or need to communicate with the outside world presents a new challenge. Figure 5 provides an illustration of how diverse and fragmented networking connections can become at home. The house is connected to the

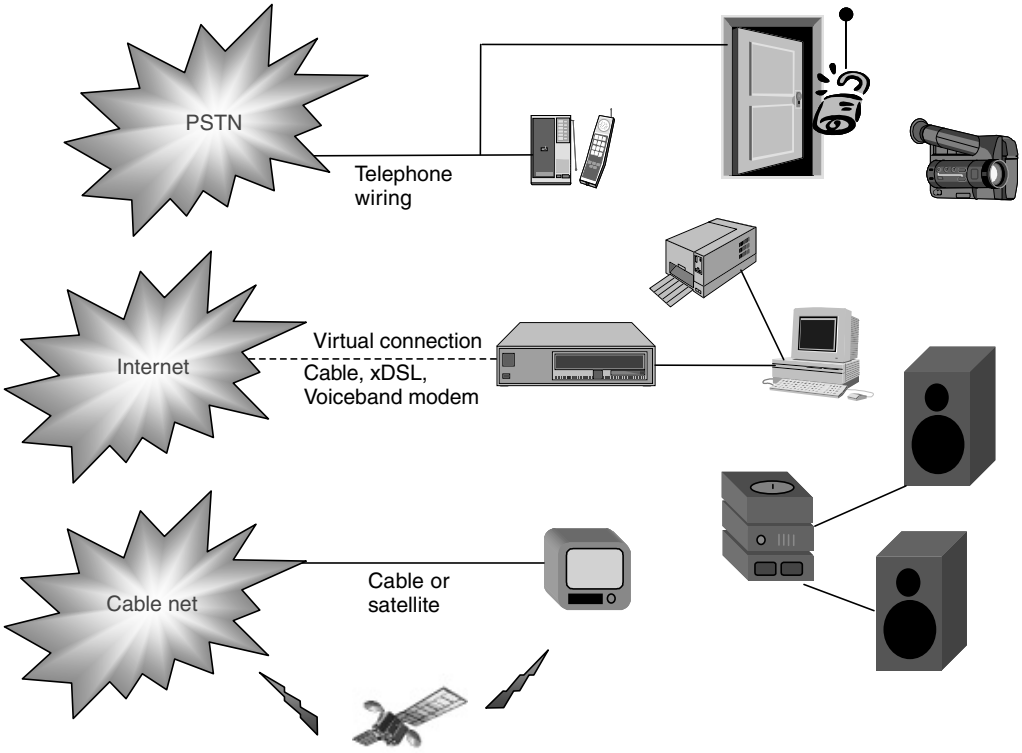


Figure 5. Today's fragmented home access and distribution networks.

PSTN for telephone services, to the Internet for Web access, and to cable network for multichannel TV services. Inside the home, computers and printers are connected to the Internet through voiceband modems, xDSL services, or cable modems. The telephone services and security systems are connected through the phone line. The TV is connected to the multichannel services through HFC cables or satellite dishes. Audio and video entertainment equipment such as a videocamera and a stereo system, and computing systems such as laptops are either isolated or have proprietary wired connections.

This fragmented networking environment has prompted a number of initiatives to create a home network. The home networking industry started in the late 1990s by the design of the so-called home gateways to connect the increasing number of computer appliances, and distribute a single Internet connection. The number of home networks in the United States is expected to almost double each year. This industry has two distinct branches: home access and home distribution segments. The home access technology employs different wireless and wired alternatives to secure a broadband Internet access to the home gateway to be distributed to the users information appliances.

The home distribution or *home-area network* (HAN) interconnects all the home appliances and connects them to the Internet through the home gateway. For the access industry, it is expected that 80% of U.S. households will have a broadband data access by the year 2004. For the distribution industry it is expected that the number of sold "information appliances" will exceed the sold number of PCs by the year 2002. It is also expected that to interconnect PCs and information appliances to the broadband services, 10 million home networks will be installed by the year 2004.

4.1. What Is a HAN?

The home-area network (HAN) provides an infrastructure to interconnect a variety of home appliances and connect them to the Internet through a central home gateway. A number of home appliances are emerging in the market that are in need of a HAN. Figure 6 provides an overview of these appliances classified into logical groups.

Home computing equipment is used for computing and Internet transaction interface access and includes PCs, laptops, printers, scanners, and QuickCAMs. If there is no home distribution network all the equipment is connected together either through the PC or laptop ports. A home computing network allows multiple computers as well as multiple devices to connect with a network protocol. A wireless network allows flexibility in installation and relocation of these devices in different rooms of a home. *Phone appliances* used for two-way conversations are cordless telephones, intercommunication devices, and standard wired telephone sets. All the telephone services have an interface to communicate with the PSTN. Currently they are connected through the home telephone wiring to the PSTN. With a HAN these devices can share the home access medium (cable or TP) allowing one service provider to bring both data and voice services. The *entertainment audiovisual appliances* include TVs, stereo systems, CD players, VCRs, DVD players, tape

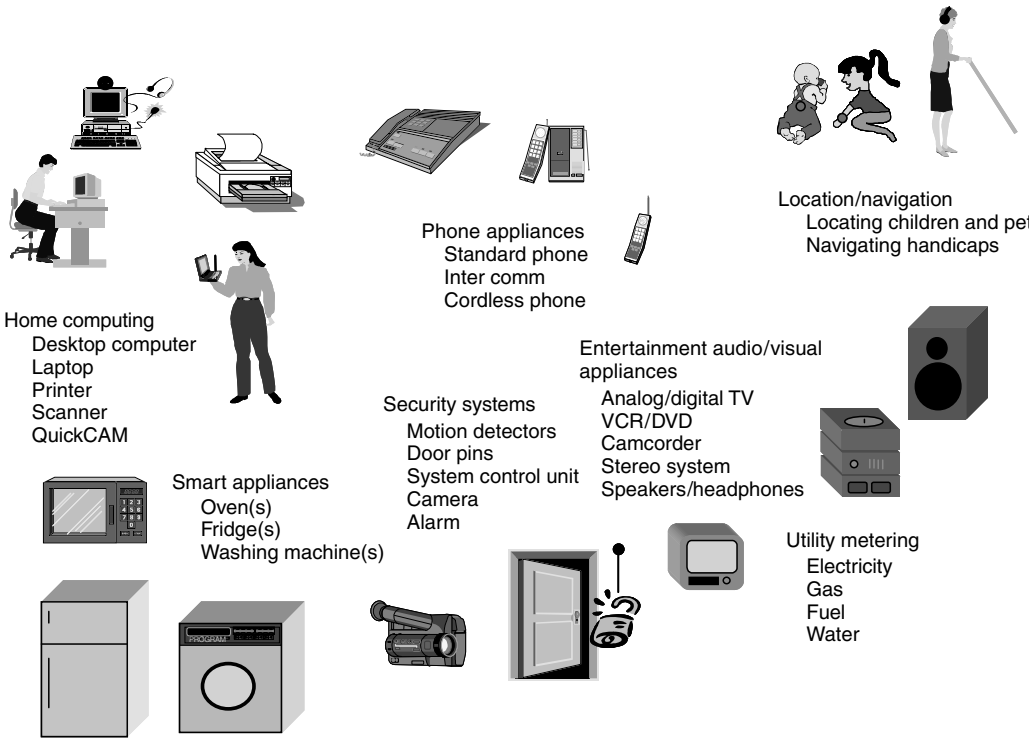


Figure 6. Classification of home equipment demanding networked operation.

recorders, camcorders, speakers, and headphones. These devices communicate through their own protocols such as the more recent IEEE 1394 or HAVi [18]. The *security system* includes motion detectors, door pins, system control panel, camera, and alarm that are networked separately using protocols such as EIA-600 CEBus [19] or the industrial initiative EIA-709 LonWorks [20]. Currently, the access to these systems is through the telephone lines that initiate emergency request alarms to the police stations. Appliance manufacturers are working on "smart" appliances capable of communication. This intelligence allows remote checking test for maintenance (e.g., receiving alarm that refrigerator's filter needs to be replaced) and remote control of operation (e.g., turning a washing machine on from outside). Besides LonWorks, another industrial initiative comes from Merloni's Ariston Digital and is called *Web-ready appliances protocol* (WRAP) [21]. Another wave of interest has been initiated by utility companies for distance *utility metering*. The electricity, gas, water, or fuel companies would like to read the meter at home for billing or other needs (e.g., refilling the gas tank). One of the solutions is to communicate this information through the HAN and its access to the Internet. More recently, a number of startup companies have been designing *indoor locating systems* that can be used for distance monitoring of children, the elderly, and pets or for navigating the blind. These systems are expected to be integrated with the computing networks to provide access through the Internet.

4.2. Why Do We Need a HAN?

The existing LANs designed for office environments do not provide a good solution for home networking. The application diversity, network requirements, building infrastructure, and market size of HANs are distinctly different from that of LANs. At home, the number of users of the network is much smaller than in the offices but the diversity of the device types and their bandwidth requirements are much larger than in the offices. The diversity of bandwidth requirement can range from multichannel video to monthly meter reading. In an office, computing devices are predominant, but the home environment includes new applications that were not needed for LANs such as audiovideo broadcasting and positioning/navigation. Office environments are larger than residential homes and are made up of material more concrete than that found in the home. Therefore physical wiring and wireless coverage in the homes is easier than in offices. Homeowners are more reluctant to allow service workers to enter their homes, and they cannot afford a network manager to operate their network. The number of homes is an order of magnitude larger than that of the offices, so the market for home networking is expected to be much larger than for LANs.

These specific requirements on home applications impose certain constraints on the design of HANs. A HAN needs to be *user-friendly* because it is used and managed by nonprofessionals with limited technical skills and small budget size. A HAN must be low cost, easy to install and relocate, and easy to upgrade. In terms of *performance* a HAN should enable multimedia applications and be capable of accommodating legacy voice and data services. A HAN also needs to be *flexible and scalable* to allow location independent easy-to-reconfigure networks without significant performance degradations.

To avoid eavesdropping a HAN also needs *security and privacy* provisions.

4.3. HAN Technologies

For the offices, a company recognizes the need for networking, decides on installing a network, opens a budget for expensive wiring, and installs the LAN infrastructure. In the homes, users gradually build their networks at their leisure with an investment that is spread over a relatively long period. An average customer of home network does not spend a sizable budget on the wiring to develop an infrastructure. Therefore, the trend in today HANs is to avoid adding any new wiring by either using existing wires or by using a wireless solution. The existing wirings at homes are twisted-pair (TP) telephone wirings, power-line wirings and cable TV wirings. The *phone-line* wirings have a relatively good distribution and in most modern homes at least one telephone outlet can be located in every room. The wiring for phone lines is voice grade TP that is suitable for Ethernet connections. However, this line is used for carrying regular phone and xDSL services that will interfere with an Ethernet signal. *Power-line* wirings are even better distributed because every room has several power outlets. However, the quality of the line is poorer and the level of noise is much larger than in TP phone-line wirings. The characteristics of the lines impose limitations on data rates that can be overcome only by using more complex transmission techniques. Existing *cable TV* wirings have very restricted distribution and only a few outlets are available in each household. This wiring is used for multichannel TV distribution that will interfere with an Ethernet signal. The expensive broadband cable TV modems can be used to overcome this problem. Because of its limited distribution and expensive modem requirements, cable TV wirings are not considered seriously for home distribution. *Wireless* solutions appear ideal for home networking. The ease of installation and relocation provides an excellent solution. Challenges for wireless is reliability, bandwidth, coverage, and interference. Comparing wired and wireless solutions, current wired HANs can be implemented using less expensive network cards and are expected to support higher data rates. Wireless HANs provide ideal ad hoc solutions that support portability.

4.3.1. HPNA. In outdoor areas the PSTN network has two parts, the TP analog access wirings and the backbone digital wiring connecting PSTN switches together. The digital segment of the PSTN fully utilizes the wiring while the traditional access wiring was using only $\sim$4 kHz for analog POTS and the rest was unexplored until xDSL and HPNA were introduced. Further, today computers purchased with network capability as well as PCMCIA network cards for laptops are exclusively Ethernet-based. However, as we mentioned before, the TP phone line wirings at home are also used for analog voice and xDSL access. HPNA is an Ethernet-compatible LAN over the random-tree home phone lines. It uses a standalone adapter to connect directly to the in-home telephone jacks any device having an Ethernet 10base-T interface card, operating at 10 Mbps over the TP lines.

HPNA shares the TP line medium with POTS and xDSL using FDM. POTS uses the 20–3400 Hz band for analog voice transmission, xDSL uses a 25–1100 kHz to provide high-speed Internet access, and HPNA uses a 2–30 MHz band for home distribution networking. The HPNA is based on a patented physical-layer design that is more immune to high noise conditions in the home telephone wirings. The MAC layer for the HPNA is the same as the MAC layer of the IEEE 802.3 Ethernet. From the user's point of view, the HPNA network accommodates the exiting legacy Ethernet software and hardware. Only an adaptor is placed between the Ethernet connection and the phone plugs. The next step for HPNA is to boost the data rate to accommodate video applications.

4.3.2. Power-Line Modems. As compared with telephone wirings, power line wirings have the best wiring distribution in the home due to the superior number of electric outlets that can be found. The entire power line wiring is used for transmission of a 50–60-Hz waveform, and the rest is available for other applications. For many years power lines were used for low-data-rate ($<$100 kbps) control and security networks operating below 500 kHz. These systems were mostly using X-10 and the CEBus/CAL standards [17]. More recently, power lines are being considered for high data rate communications ($>$1 Mbps) and to operate above 1 MHz to provide adequate speeds for computer networking. This area is still in the preliminary stages of development with no clear standard initiative. European regulations prohibit power line signaling above 150 kHz due to potential interference with low-frequency licensed radio services. Current U.S. and Japanese regulations allow the use of a somewhat broader spectrum up to $\sim$525 kHz where AM radios begin. In the power line, a low-frequency band of up to a few kilohertz is used for low-data-rate applications such as security, and a high-frequency band (1–30 MHz) is used for high-speed data communications.

Power lines suffer from tremendous interference from electrical appliances, high attenuation, reflection caused from varying input impedance, and multipath phenomena that makes communication over this medium as challenging as communication over radio channels. As a result, a variety of complex transmission techniques and medium-access protocols has been examined for different power line applications. Traditional FSK and QPSK are used in lower bands and more complex spread spectrum and OFDM modems are used in the higher bands [22]. The difficulty of the medium has complicated the development of low-cost solutions and is a drawback in growth of this market. More recently, smart appliances have been emerging in the market that have some built-in intelligence, can sense other appliances on the power lines, and can be accessed through the Internet. Electric companies are investigating using the outside AC lines to deliver various services such as meter reading, energy management, and even Internet access to the homes. The main current research thrust is to enable access to the control and security systems through the Internet.

4.3.3. Wireless Solution Alternatives. Compared with wired networks, wireless solutions can provide mobility

and coverage to the home as well as the yard. Cordless telephones were very successful from the early days they appeared in the market. Since cordless telephones provide mobility and extended coverage, users have paid higher prices to purchase cordless telephones instead of wired phones. The other advantages are that wireless solutions are easy to install, relocate, scale, and maintain.

The introduction of wireless solutions to home security systems resulted in a sizable growth of that industry. From the user's point of view, wireless security systems are installed very quickly without additional holes in the walls or wires distributed in the home. Furthermore, selecting the location for placing a wireless product is more flexible and allows a better blending with the decoration of the home. Other advantages are that wireless solutions are easier to be expanded, moved to new homes, maintained, and upgraded.

The examples above described are selected applications where the wireless solution is preferred. However, there are numerous home applications and a wireless solution may not be the best one for all of them. Home automation network applications such as switching a light through remote control is done by sending a single bit (ON/OFF) through the power lines to an inexpensive ON/OFF X-10 switch attached to the lamp. Simple infrared remote control can provide the additional mobility by sending the control command to a general control box connected to the power lines. It is not difficult to find other examples involving the power lines or phone lines where wireless might not be beneficial. The important conclusion from this discussion is that in home networking we want to avoid new wiring, not avoid using the existing wires. The home therefore can be seen as a nonhomogeneous environment for networking.

The principal candidates for wireless home networking are WLANs and WPANs discussed in the previous sections. There are a number of home-specific challenges for wireless home networking. The use of noncompatible wireless devices operating in the same unlicensed band can become an issue if they interfere with each other. Handling interference in such cases is an important issue. Another issue is that as we move to higher frequencies of operation in the 5-GHz range to support higher data rates, the coverage of the wireless solution may become a challenge. Wireless home networking needs designing inexpensive reconfigurable devices and internetworking between diverse mediums using different protocols. The network to incorporate cable TV applications requires high transmission rates as well as new delivery techniques to the TV set. Today, the coaxial cable that connects to the converter box on the TV set carries around a hundred analog video channels. This extent of bandwidth is not feasible in a wireless medium, and therefore, the entire system needs to be redesigned.

4.4. Home Access Networks

The early home access technology was based on voice band modems over the phone line. Today, broadband home access with data rates on the order of 10 Mbps is provided through cable modems or xDSL services. The cable distribution in the residential areas has a bus topology that is optimally designed for one-way TV signal distribution. The bus carries all the stations in the neighborhood. Cable modems use a bandpass channel allocated to a TV channel to provide high data rates for transmission using QAM modulation. Broadband cable services use one of the video channels and a reverse channel to establish a two-way communication and access to Internet. The xDSL service uses a 25–1100-kHz band on the phone line and multisymbol QAM modulation to support high data rates to the users. The topology of the telephone line is a star topology that connects every user directly to the end office, where the xDSL data is directed to Internet through a router.

Higher-speed wireless home access uses a local multipoint distribution system (LMDS) or even existing WLAN inter-LAN bridges to provide the service. The advantage of using a fixed wireless solution is that it does not involve wiring under the streets. The wireless solution is certainly attractive when there is no existing wiring in the neighborhood or when obtaining city permission might be troublesome. The HIPER-ACCESS preprogram in the EC and IEEE 802.16 are currently studying the specifications for the next generation of wireless local access. Other wireless alternatives are direct satellite TV broadcasting and 3G wireless networks. Direct broadcast suffers by the lack of a reverse channel and high delays that challenges the implementation of broadband services. The high-speed 3G wireless packet data services are expected to provide up to 2 Mbps, suitable for Internet access. However, besides a lower data rate than the others, these networks will be using licensed bands, and ultimately may be expensive as well. Figure 7 summarizes the existing solutions for the home access technologies.

5. INDOOR GEOLOCATION

5.1. Introduction

An important evolving technology recent has been indoor geolocation technology for both military and commercial applications. There is an increasing need in hospitals to locate patients or expensive equipment and in homes

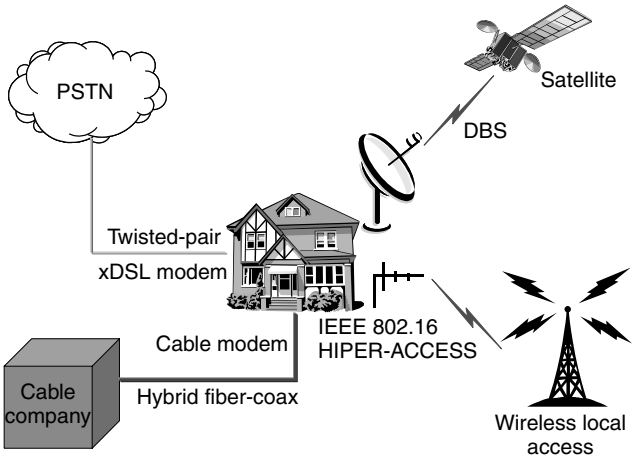


Figure 7. Broadband home access alternatives.

to locate children and equipment. In the Department of Defense, Small Unit Operation/Situation Awareness Systems require a modern warfighter to be able to communicate in an urban environment, and his/her position to be known at all times including when he is inside a building. An indoor geolocation system can also be crucial for the safety of a firefighter entering a blazing building when his/her position is being monitored. Similar scenarios can be found in other public safety agencies. These incentives have lead to research in indoor geolocation systems [13,23,24]. In an indoor environment, traditional GPS or E-911 location systems do not work properly due to severe multipath effects. As a result, dedicated indoor systems have to be developed to provide accurate indoor geolocation services.

5.2. Wireless Geolocation Methods and Metrics

Most geolocation system architectures and methods developed for cellular systems are applicable for indoor geolocation systems although special considerations are needed for indoor radio channels. The most widely used wireless geolocation metrics include angle of arrival (AoA), time of arrival (ToA), time differences of arrival (TDoA), received-signal strength (RSS), and received-signal phase (RSP). In the following, we present an overview of overall system architectures and basic concepts of geolocation metrics as well as corresponding geolocation methods.

5.2.1. Overall System Architecture. Similar to cellular geolocation systems, the architecture of indoor geolocation systems can be roughly grouped into two main categories: mobile-based architecture and network-based architecture. Most of indoor geolocation applications proposed to date have been focused on a network-based system architecture as shown in Fig. 8 [25,26].

The geolocation base stations (GBSs) extract location metrics from the radio signals transmitted by the mobile station and relay the information to a geolocation control station (GCS). The connection between GBS and GCS can be either wired or wireless. Then the position of the mobile station is estimated, displayed and tracked at the GCS. With the mobile-based system architecture, the mobile station estimates self-position by measuring the received radio signals from multiple fixed GBS. Compared to a mobile-based architecture, the network-based system has the advantage that the mobile station can be implemented as a simple-structured transceiver with small size and low power consumption, easily carried by people or attached to valuable equipments as a tag.

5.2.2. Angle of Arrival. The AoA geolocation method uses simple triangulation to locate the transmitter as shown in Fig. 9.

The receiver measures the direction of the received signals (i.e., angle of arrival) from the target transmitter using directional antennas or antenna arrays. If the accuracy of the direction measurement is $\pm\theta_s$, AoA measurement at the receiver will restrict the transmitter position around the line-of-sight (LoS) signal path with an angular spread of $2\theta_s$. AoA measurements at two receivers will provide a position fix as illustrated in Fig. 9. We can clearly observe that given the accuracy of AoA measurements, the accuracy of the position estimation depends on the transmitter position with respect to the receivers. When the transmitter lies between the two receivers, AoA measurements will not be able to provide a position fix. As a result, more than two receivers are normally needed to improve the location accuracy. For a macrocellular environment where the primary scatters are located around the transmitter and far away from the receivers, the AoA method can provide acceptable location accuracy [27]. But dramatically large location errors will occur if the LoS signal path is blocked and the AoA of a reflected or a scattered signal component is used for

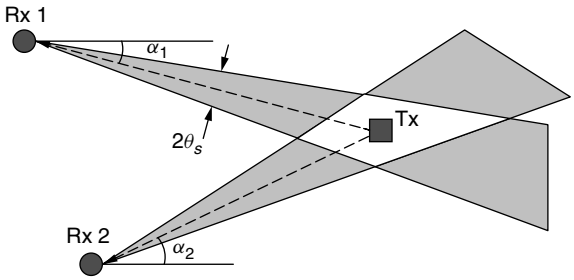


Figure 9. Angle-of-arrival geolocation method.

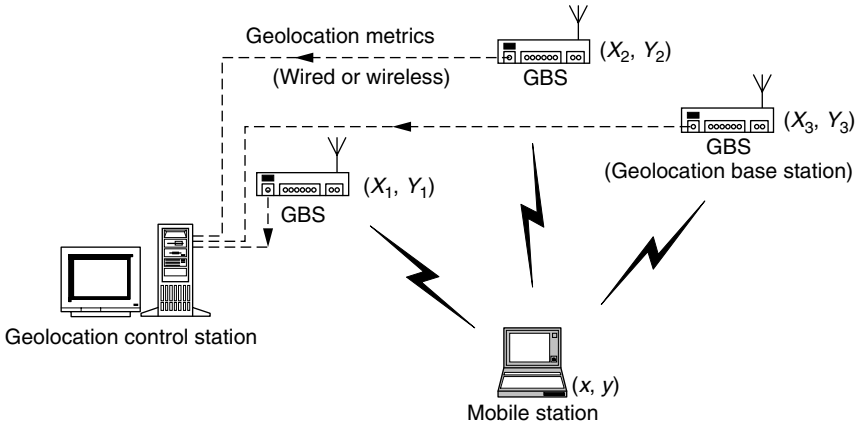


Figure 8. Overall architecture of indoor geolocation system.

estimation. In indoor environments, the LoS signal path is usually blocked by surrounding objects or walls. Thus the AoA method alone cannot provide sufficient accuracy for indoor geolocation systems.

5.2.3. Time of Arrival and Time Difference of Arrival. The ToA method is based on estimating the propagation time of the signals from a transmitter to multiple receivers. Several different methods can be used to obtain ToA or TDoA estimates, including pulse ranging [28,29], phase ranging [28], and spread-spectrum techniques [25,30].

Once the ToA is measured, the distance between the transmitter and receiver can be determined simply since the propagation speed of the radio signal is approximately the speed of light. The estimated distance at the receiver will geometrically define a circle, centered at the receiver, of possible transmitter positions. ToA measurements at three receivers will provide a position fix and given receiver coordinates and distances from the transmitter to receivers, the transmitter coordinates can be easily calculated. In indoor environments, the strongest signal received can come from a longer reflection path and not from the direct line-of-sight path. The ToA-based distance estimates are therefore always larger than the true distance between the transmitter and the receiver as illustrated in Fig. 10, where $\hat{r}_1, \hat{r}_2$, and $\hat{r}_3$ are the estimated distances and r_1, r_2, and r_3 are the true distances. Three ToA measurements determine a region of possible transmitter position as shown in Fig. 10. A nonlinear least-square (NLLS) method is usually used to obtain the best estimation iteratively [25,27]. A constrained NLLS algorithm is also available that makes use of the fact that ToA-based distance estimates are always larger than the true distance [31]. As in the AoA method, more than three ToA measurements are needed to improve the accuracy of position estimation.

Instead of using ToA measurements, time difference measurements can also be employed to locate the receiver position. A constant time difference of arrival (TDoA) for two receivers defines a hyperbola, with foci at the receivers, on which the transmitter must be located. Three or more TDoA measurements provide a position fix at the intersection of hyperbolas. NLLS method can also be used to obtain the best estimation of the transmitter position.

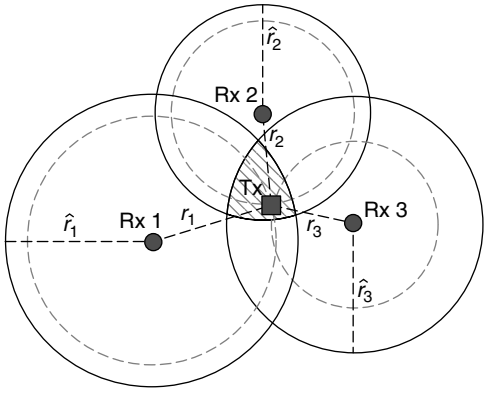


Figure 10. Time of Arrival geolocation method.

Some other methods have been used to solve the hyperbolic position estimation problem [32–35]. Compared to the ToA method, the main advantage of the TDoA method is that it does not require the knowledge of the transmit time from the transmitter while the ToA method requires it. As a result, strict time synchronization between the transmitter and receivers is not required. However, the TDoA method requires time synchronization among all the receivers.

5.2.4. Received-Signal Strength. If the power transmitted by the mobile terminal is known, measuring the received-signal strength (RSS) at the receiver can provide an estimate of the distance when the path-loss characteristics of the environment are known. As in the ToA method, the measured distance will determine a circle, centered at the receiver, on which the mobile transmitter must lie. Three RSS measurements will provide a position fix for the mobile. As a result of shadow fading effects, the RSS method results in large range estimation errors. The accuracy of this method can be improved by utilizing a pre-measured received signal strength contour centered at the receiver [36]. A fuzzy-logic algorithm has been shown [37] to be able to significantly improve the location accuracy.

5.2.5. Received-Signal Phase. Signal phase is another possible geolocation metric. It is well known that with the aid of reference receivers to measure the carrier phase, differential GPS (DGPS) can improve the location accuracy from ~20 m to within 1 m compared to the standard GPS, which only uses pseudorange measurements. One problem associated with the phase measurements lies in the ambiguity resulting from the periodicity of the signal phase while the standard pseudorange measurements are unambiguous. Consequently, in the DGPS, the ambiguous carrier phase measurement is used to fine-tune the pseudorange measurement. A complementary Kalman filter is used to combine the low-noise ambiguous carrier phase measurements with the unambiguous but noisier pseudorange measurements [38]. For indoor geolocation systems, it is possible to use the received-signal phase method together with the ToA/TDoA or the RSS method to fine-tune the location estimate. However, unlike DGPS, where the LoS signal path is always observed, multipath and non-line-of-sight conditions in the indoor environment can cause large errors in phase measurements.

BIOGRAPHIES

Kaveh Pahlavan, is a professor of ECE and CS, and director of the CWINS laboratory at WPI, Worcester, Massachusetts. He is also a visiting professor of TLab and CWC, University of Oulu, Finland. He is the principal author of the *Wireless Information Networks*, John Wiley and Sons, 1995 (with A. Levesque) and *Principles of Wireless Networks—A Unified Approach*, Prentice Hall, 2002 (with P. Krishnamurthy). He has published numerous papers, served as a consultant to a number companies, and sits in the board of a few companies. He is the editor in chief and founder of the *International Journal of Wireless Networks*, the founder of

the IEEE Workshop on Wireless LANs, and a cofounder of the IEEE PIMRC conference. For his contributions to evolution of the wireless networks he has received the Weston Hadden Professor of ECE at WPI, been elected as a fellow of the IEEE in 1996, become a fellow of Nokia in 1999, and received the first Fulbright–Nokia Scholarship Award in 2000. Because of his inspiring visionary publications and his international conference activities for the growth of the wireless LAN industry he is referred to as one of the founding fathers of the wireless LAN industry. Details of his contributions to this field are available at www.cwins.wpi.edu.

Jacques Beneat received his Ph.D. degree in electrical and computer engineering from Worcester Polytechnic Institute (WPI), Massachusetts, in 1993 with focus on the design of data communication circuits and advanced microwave structures for satellite communications. During several years, he served as adjunct faculty in the University of Bordeaux, France, Worcester State College, and WPI. In 1996, he joined the Center for Wireless Information Network Studies at WPI, and during the past five years, he has served as a research scientist with focus on indoor radio propagation measurements and modeling using ray tracing techniques, real-time channel simulators and performance evaluation for emerging wireless networks. He was the general conference secretary for the ninth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) in 1998, for the third IEEE Workshop on Wireless LANs in 2001, and was a guest editor of the *International Journal of Wireless Information Networks* for a special series on implementation issues in wireless communications in 1997–98. In 2002, he will be joining the EE Department at Norwich University in Vermont.

XINRONG LI (xinrong@wpi.edu) received his B.E. and M.E. from the University of Science and Technology of China, Hefei, China, and the National University of Singapore, Singapore, in 1995 and 1999, respectively. He is currently pursuing his Ph.D. degree in wireless communications and networks at Worcester Polytechnic Institute (WPI), Worcester, Massachusetts. He has been working as research assistant in the Center for Wireless Information Network Studies (CWINS), WPI, and his recent research has been focused on indoor geolocation, statistical signal processing, and wireless networks. During the summer of 2002, he was also involved in the development of 3G All-IP CDMA2000 1xEV-DO Wireless Data Networks at Airvana, Inc., Chelmsford, MA.

BIBLIOGRAPHY

1. F. R. Gfeller and U. H. Bapst, Wireless in-house data communication via diffuse infrared radiation, *Proc. IEEE* **67**(11): 1474–1486 (Nov. 1979).
2. P. Ferert, Application of spread spectrum radio to wireless terminal communications, *Proc. NTC*, Houston, TX, Dec. 1980, pp. 244–248.
3. K. Pahlavan, Wireless communications for office information networks, *IEEE Commun. Mag.* **23**(6): 19–27 (June 1985).
4. M. J. Marcus, Regulatory policy considerations for radio local area networks, *IEEE Commun. Mag.* **25**(7): 95–99 (July 1987).
5. K. Pahlavan, *Wireless Intra-office Networks*, ACM Transactions Office Information Systems, July 1988.
6. M. Kavehrad and P. J. McLane, Spread spectrum for indoor digital radio, *IEEE Commun. Mag.* **25**(6): 32–40 (June 1987).
7. K. Pahlavan and A. Levesque, *Wireless Data Communication*, invited paper, IEEE Proc., Sept. 1994.
8. K. Pahlavan, T. Probert, and M. Chase, Trends in local wireless networks, invited paper, *IEEE Commun. Soc. Mag.* (March 1995).
9. First IEEE Workshop on Wireless LANs, Worcester, Mass., May 1991, *http://www.cwins.wpi.edu/wlans91/index.html*.
10. K. Pahlavan, A. Zahedi, and P. Krishnamurthy, Wideband local access: WLAN and WATM, *IEEE Commun. Mag.* (Special Series on Wireless ATM), (Nov. 1997).
11. S. Narayanaswamy et al., Application and network support for infopad, *IEEE Pers. Commun. Mag.* (March 1996).
12. L. Dennison, Body LAN, a wearable personal network, *Proc. 2nd IEEE Workshop on Wireless LANs*, Worcester, MA, Oct. 1996.
13. K. Pahlavan, P. Krishnamurthy, and J. Beneat, Wideband radio propagation modeling for indoor geolocation applications, *IEEE Commun. Mag.* 60–65 (April 1998).
14. K. Pahlavan et al., Handoff in hybrid mobile data networks, *IEEE PCS Mag.* **7**(2): 34–47 (April 2000).
15. T. Siep, I. Gifford, R. Braley, and R. Heile, Paving the way for personal area network standards: An over view of the IEEE P802.15 Working Group for Wireless Personal Area Networks, *IEEE Commun. Soc. Mag.* (Feb. 2000).
16. K. Negus, A. Stephens, and J. Lansfield, HomeRF: Wireless networking for the connected home, *IEEE Commun. Soc. Mag.* (Feb. 2000).
17. Bluetooth SIG, *http://www.bluetooth.com*.
18. HAVi (home audiovideo interoperability), *http://www.havi.org/home.html*.
19. CEBus — Consumer Electronics Bus, *http://www.cebus.org/*.
20. LonWorks Technology Platform, *http://www.echelon.com/*.
21. WRAP — Ariston Digital Web-ready appliance protocol, *http://www.margherita2000.com/wrap/uk/main/index.htm*.
22. Intellon High Speed Power Line Communications, white paper, *http://www.intellon.com/*.
23. P. Krishnamurthy, K. Pahlavan, and J. Beneat, Radio propagation modeling for indoor geolocation applications, *Proc. IEEE PIMRC'98*, Sept. 1998.
24. J. Werb and C. Lanzl, Designing a positioning system for finding things and people indoors, *IEEE Spectrum* **35**(9): (Sept. 1998).
25. PinPoint Local Positioning System, *http://www.pinpointco.com/*.
26. PalTrack Tracking Systems, *http://www.sovtechcorp.com/*.
27. J. Caffery, Jr. and G. L. Stuber, Subscriber location in CDMA cellular networks, *IEEE Trans. Vehic. Technol.* **47**(2): (May 1998).

28. G. Turin, W. Jewell, and T. Johnston, Simulation of urban vehicle-monitoring systems, *IEEE Trans. Vehic. Technol.* **VT-21**: 9–16 (Feb. 1972).

29. H. Hashemi, Pulse ranging radiolocation technique and its application to channel assignment in digital cellular radio, *Proc. IEEE VTC'91*, 1991, pp. 675–680.

30. P. Goud, A. Sesay, and M. Fattouche, A spread spectrum radiolocation technique and its application to cellular radio, *Proc. IEEE Pacific Rim Conf. Communications, Computer and Signal Processing*, 1991, pp. 661–664.

31. G. Morley and W. Grover, Improved location estimation with pulse-ranging in presence of shadowing and multipath excess-delay effects, *Electron. Lett.* **31**: 1609–1610 (Aug. 1995).

32. W. H. Foy, Position-location solutions by Taylor-series estimation, *IEEE Trans. Aerospace Electron. Syst.* **AES-12**: 187–194 (March 1976).

33. D. J. Torrieri, Statistical theory of passive location system, *IEEE Trans. Aerospace Electric Syst.* **AES-20**(2): (March 1992).

34. J. S. Abel and J. O. Smith, A divide-and-conquer approach to least-squares estimation, *IEEE. Trans. Aerospace Electric Syst.* **26**: 423–427 (March 1990).

35. Y. T. Chan and K. C. Ho, A simple and efficient estimator for hyperbolic location, *IEEE Trans. Signal Process.* **42**(8): 1905–1915 (Aug. 1994).

36. W. Figel, N. Shepherd, and W. Trammell, Vehicle location by a signal attenuation method, *IEEE Trans. Vehic. Technol.* **VT-18**: 105–110 (Nov. 1969).

37. H.-L. Song, Automatic vehicle location in cellular communications systems, *IEEE Trans. Vehic. Technol.* **43**(4): 902–908 (Nov. 1994).

38. E. D. Kaplan, *Understanding GPS: Principles and Applications*, Artech House, Boston, 1996.

TROPOSPHERIC SCATTER COMMUNICATION

PETER MONSEN
P.M. Associates
Stowe, Vermont

1. INTRODUCTION

Prior to the launching of the first active repeater satellite TELSTAR on July 10, 1962, long-distance communication was limited to cable systems and radio communication techniques that exploited characteristics of either the ionosphere or the troposphere. Radio systems offered advantages of greater network flexibility, the ability to transverse difficult terrain, and less susceptibility to connectivity loss due to either catastrophic failure or sabotage. In high-frequency (HF) radio systems, ionospheric reflections made communication possible up to thousands of kilometers but with a limited bandwidth that would permit transmission of only one or two telephone channels.

When the transmitting and receiving antennas of a radio link are reciprocally visible, the link is classified as line-of-sight (LoS) and the received signal is reduced as the square of the distance. Beyond the LoS zone which can extend up to about 50 km, the received signal power is more severely reduced with distance because now the radio waves must bend or diffract with the curvature of the earth. In 1937 VanDerPool and Bremmer [1] determined losses due to diffraction over a spherical surface. These losses increase exponentially with distance beyond the LoS zone, so communication dependent on diffraction is limited to short distances beyond the horizon.

The introduction of higher-power radars during World War II, however, led to radio interferences at beyond LoS distances that was considerably larger than predicted by diffraction theory [2]. Some of the contribution to stronger signals was attributed to a refractive index (the ratio of the speed of light in a vacuum relative to the medium of interest) that decreased with height, thus increasing the "effective earth radius." Measurement of interference after World War II from frequency modulation radio stations and television stations confirmed that additional factors were contributing to the enhanced signal levels.

The troposphere extends from the ground to about 10 km and includes virtually all weather phenomena. In the 1950s it was recognized that the turbulence in the troposphere due to variations of meteorological characteristics such as temperature and humidity, would produce inhomogeneities in the refractive index. As a result of these refractive index variations within the scattering common volume corresponding to the intersection of the transmit and receive antenna beams, some fraction of the radio energy would be scattered in the direction of the receive antenna [3,4]. Experimental links implemented between 1950 and 1955 established that although the received signal was continuously variable it was permanently present so that reliable communication would be possible. Because of the scattered radio signals were weak, it was necessary to use very large antennas, sensitive low-noise receivers, and powerful klystron amplifiers with kilowatt outputs. The first "troposcatter" military and commercial links were installed from 1953 onward and provided up to 60 telephone channels in the 400–900-MHz frequency band over distances up to 300 km [5].

2. TROPOSPHERIC SCATTER RADIO LINKS

The provision of multiple voice channel circuits over long distances was the primary purpose of early troposcatter radio links. These links would be connected together to provide telephone conversations, for example, between the United States and Europe via a network of military links that traversed Greenland, Iceland, and Scotland and into Europe as far as Turkey.

In these troposcatter radio links the strongest signal is received when the antennas are aimed approximately at the horizon. The angle between the transmit and receive antenna beam centerlines at their intersection is called the scattering angle θ. In the LoS zone, for example, where the antennas are reciprocally visible this angle is zero. The transmission loss in troposcatter mode is proportional to $\theta^{-\alpha}$, where various theories predict α to be between $\frac{11}{3}$ and 5. Thus the antennas are pointed as low to the horizon as possible without suffering undo blockage. The antennas at

both the transmitter and receiver were relatively large in order to successfully capture the weak scatter signal and provide reliable communication. On some of the longest links these antennas could be 120 ft in diameter.

The standard telephone channel has a 4-kHz bandwidth. In the earlier analog troposcatter systems these channels would be frequency-division multiplexed (FDM) and frequency modulated (FM) to a radiofrequency (RF) carrier between a few hundred megahertz and 5000 MHz. The power amplifier was typically a klystron power tube capable of generating kilowatts of power.

Digital systems used a time-division multiplex (TDM) of digitally converted voice channels. Modulation techniques such as quadrature phase shift keying (QPSK) could be used to convert the composite high-rate digital stream into an RF signal for subsequent amplification and transmission.

At the receiver, a low-noise amplifier was employed for each antenna input in order to increase the margin between the received signal and the ever-present thermal noise. Demodulation in analog systems of the received signal was accomplished in a frequency modulation discriminator that produced a voltage proportional to the instantaneous frequency. In a digital system, adaptive processing was included in the QPSK demodulation process in order to cope with time-delayed scattered components. After demodulation the composite detected signal would be demultiplexed and converted to a series of telephone channels. In an analog system the noise introduced by transmission impairments would be passed on to the next link and would accumulate over a series of tandem links. A digital system offered the advantage of regenerating (although with occasional bit errors) the original transmitted data. In repeater systems the demultiplex operation would be omitted and the detected composite signal could serve as the input for the next troposcatter link.

Because the received signal strength depends on a scattering process that is dependent on fluctuations in the refractive index within the common volume, variations in meteorologic conditions such as temperature and humidity in the common volume produce long-term signal fading. This long-term fading is usually characterized by an hourly median value of received signal power or transmission loss in decibels. Its variation has been found to closely approximate a Gaussian distribution so it is completely described by the hourly median and the standard deviation.

Of course, the scattering process itself results in a fading signal as different scattering paths will add or subtract at the receiver in a random fashion. The individual scattering "blobs" in the common volume tend to be statistically independent so the received signal can be represented in complex notation as the sum of a large number of independent complex random variables. Using the central-limit theorem, the complex received signal is then close to a complex Gaussian process with an envelope, i.e., its magnitude, that follows the Rayleigh probability distribution. If x is the received signal power value and x_m is its median value, the probability that the random variable x is less then a critical value x_c is

$$\rho(x < x_c | x_m) = 1 - e^{-x_c \ln 2 / x_m}$$

which can be approximated for small values as

$$\rho(x < x_c | x_m) \doteq \frac{x_c \ln 2}{x_m} \quad x_c \ll x_m$$

The fading probability distribution above shows that deep fades have a slope of 10 dB per probability decade. This implies that to achieve short-term availabilities of 99.9% there must be a fade margin protection of about 30 dB. One way to reduce this severe power penalty due to fading is to provide redundant signal paths called *diversity paths*. Different diversity configurations are considered in the next section. In general for Mth-order diversity (i.e., M independent transmission paths between transmitter and receiver), the fading slope reduces to $10/M$ dB per probability decade. Thus a quadruple diversity system requires about $\frac{30}{4} = 7.5$ dB of fade margin protection relative to the 30 dB for the no diversity system at a 99.9% availability.

The advantages of troposcatter systems include

- Realization of long paths beyond the LoS zone
- Use over difficult terrain where installing cable systems is unattractive
- Physical security because prevention of sabotage or catastrophic failure is easy to achieve with a small number of stations
- High immunity to interception or to jamming because of the use of narrow beam antennas
- Provision of a communication service that does not depend on satellite services
- Rapid link installation (about one day)

The disadvantage of tropospheric scatter networks include poorer quality signals and lower data rate capacities than alternative satellite and cable systems.

3. DIVERSITY CONFIGURATIONS

In a multipath fading environment it is common to add or utilize redundant transmission paths so that the fading of one or more paths will not necessarily prevent correct reception of the transmitted information. This redundancy, called *diversity*, is a critical element in the successful operation of troposcatter links. For example, in the Rayleigh fading environment for tropospheric circuits, if an outage level of 1% was acceptable, a no diversity system would require about 11.5 dB more transmit power than a dual-diversity system with two separate receive antennas. Since the antennas and power amplifiers in most troposcatter circuits are near technological limits, diversity is an extremely attractive performance improvement approach. Diversity techniques include frequency using multiple radiofrequency carriers, space using multiple transmit or receive antennas, polarization using orthogonal polarization transmissions,

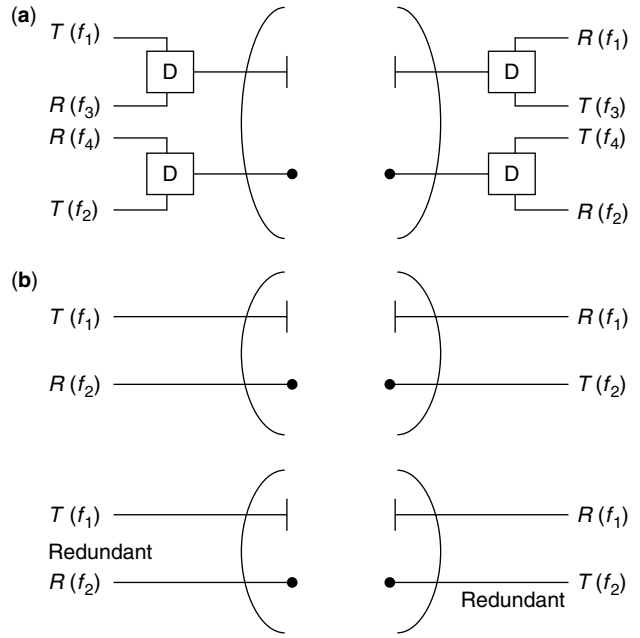


Figure 1. Dual-diversity configurations: (**a**) dual-frequency (2F) diversity; (**b**) dual space (2S) diversity. (D = duplexer).

and angle using an antenna feedhorn that produces multiple beams.

3.1. Frequency and Space Diversity

The most common diversity configurations use combinations of frequency and space as shown in Figs. 1 and 2. A diamond and bar are used in these figures to denote horizontal and vertical polarization, respectively. Troposcatter circuits are always duplex, so frequency diversity requires a duplexer that allows a transmitter and receiver operating in two different frequency bands to be connected to the same antenna. Transmit powers are on the order of 1 kW and receivers may be required to detect signals at levels near — 100 dBm, a dynamic range of 160 dB. In the dual-space diversity, only one transmit antenna needs to be used but equipment redundancy requires a second power amplifier. For equal power amplifier outputs the dual-space (2S) diversity of Fig. 1b is then 3 dB better in total received power than the dual frequency (2F) diversity of Fig. 1a because the total transmit power is the same but the 2S system receives on two antennas.

A 2S/2F quadruple diversity can be achieved with a combination of frequency and space as shown in Fig. 2a. Under certain conditions [6] a savings of two frequency channels per path can be achieved by replacing the frequency channels with orthogonal polarizations. This configuration can be denoted as quadruple space (4S) because there are four separate space diversity paths. The decorrelation of these paths is due to their spatial separation through the scattering common volume. This configuration is also called dual space/dual polarization (2S/2P) although the polarization serves only to separate the signals to each antenna for receiver processing. Generally polarization diversity with

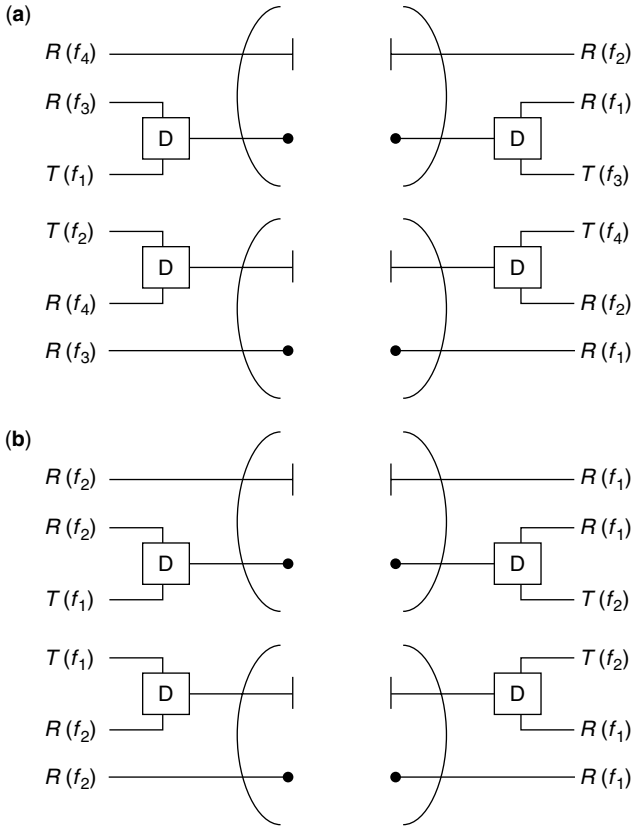


Figure 2. (**a**) Quadruple space/frequency (2S/2F) diversity and (**b**) quadruple space (4S) diversity (D = duplexer).

the same spatial path, e.g., a 2P system with one antenna at each terminal, does not realize sufficient decorrelation to provide a diversity effect.

The antenna spacing in space diversity systems is more conveniently accomplished in the horizontal direction and a separation of 100 wavelengths provides good decorrelation [6] and is widely adopted.

3.2. Angle Diversity

Redundant and decorrelated paths can be realized with multiple beams produced at either the transmitter or receiver antenna. Beams can be separated either horizontally or vertically. The antenna beams can be produced by replacing the feedhorn at the focal point of a parabolic reflector with a multiple feedhorn structure. This method is normally accomplished at the receiver because transmit diversity requires either extra power amplifiers or a division of the available transmitter power.

The selection of a horizontal or vertical displacement of beams depends on the beam correlation for these axes. Beam correlation is inversely related to the common volume power density width on that axis. Horizontally the common volume "size" is limited to approximately the beamwidth. In the vertical dimension the common volume "size" extends beyond the beamwidth limit with a scattering angle dependence of $\theta^{-11/3}$, where θ is generally larger than the beamwidth. Thus the beam correlation for conventional narrowbeam antennas is smaller in

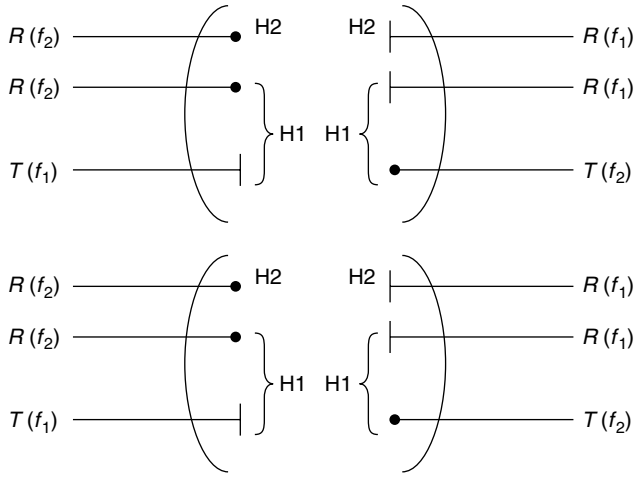


Figure 3. Dual-space/dual-angle (2S/2A) diversity.

the vertical direction. An experimental angle diversity system [7] at 5 GHz used a two-beam vertical splay with a feedhorn design that produced a beam separation or "squint angle" of approximately 1.3 beamwidths compared to a theoretical optimum of about 1 beamwidth. The loss in performance in this experimental study with the slightly larger squint angle was determined to be between 0.1 and 0.4 dB. The configuration of the experimental dual-space/dual-angle (2S/2A) system is shown in Fig. 3.

A major advantage of angle diversity is that it can be used to replace frequency diversity in a 2S/2F system to produce a 2S/2A system that saves two frequency channels per path and results in better performance. In a vertical dual-angle system, there is a performance loss of approximately one-half the decibel value of the loss associated with the larger scattering angle of the elevated beam and a small loss because of beam correlation. These combined losses in angle diversity are generally less than the 3 dB received power advantage relative to frequency diversity. Troposcatter systems require two power amplifiers for equipment redundancy, but in angle diversity they operate in the same frequency band (see Fig. 3), producing a potential received power 3 dB larger than frequency diversity at each diversity receiver. In addition to this short-term advantage, the 2S/2A system is superior to the 2S/2F system because long-term variations are decorrelated in the two antenna beams; thus troposcatter systems should have adaptively compensated takeoff angles. The experimental results reported [7,23] confirm both the short- and long-term advantages of the angle diversity system.

4. TRANSMISSION PARAMETER PREDICTION

Troposcatter propagation is possible in the frequency range from a lower limit determined by antenna size of a few hundred megahertz (MHz) up to about 10,000 MHz where water vapor losses become excessive. Practical link distances range from 50 to 500 km and channel capacities are typically 60–120 telephone channels in analog systems and up to about 12 Mbps in digital systems. These systems must cope with long-term (hourly) and short-term (seconds) variations in received signal strength and multiple paths between the transmitter and receiver that produce self-interference. Predictions of transmission loss and multipath characteristics are essential in the communication link design.

4.1. Transmission Loss

Comprehensive methods for predicting cumulative probability distributions of transmission path loss as a function of radiofrequency and distance over any type of terrain and in different climate regions were produced by the Comite Consultatif International Radio (CCIR) [9] and in the United States by the National Bureau of Standards (NBS) [10]. The detailed point-to-point predictions depend on propagation path geometry, atmospheric refractivity near the center of the earth, and specified characteristics of antenna directivity.

Path loss can be broken into two major components: (1) a basic transmission loss for a system with hypothetical loss-free isotropic antennas and (2) an antenna gain loss that accounts for the reduction in the illuminated portion of the scattering common volume when the beamwidth is reduced. The basic transmission loss depends on the radio frequency, the distance, and the scattering angle between the centerlines of the transit and receive beams at their intersection in the scattering region. The scattering angle, θ, depends on the pathlength d and the transmit and receive elevation angles, θ_{et} and θ_{er}, respectively. For $\theta d < 2$, the scattering angle can be approximated by

$$\theta = \frac{d}{a} + \theta_{et} + \theta_{er}$$

where $a = kR$ is the effective earth radius to account for bending due to a decrease in refractive index with height. Typical troposcatter calculations multiply the earth radius R by a factor of k that is $\frac{4}{3}$ [5].

Except for the small correction factors, the NBS method [10] computes the basic transmission loss in dB as

$$L_{bsr} = 30 \log f - 20 \log d + F(\theta d)$$

where the frequency is in MHz, the path distance is in kilometers, and F is an empirical attenuation function that depends on surface refractivity N_s. The surface refractivity at a height h km above sea level is related to the refractive index n_0 at sea level by

$$N_s = (n_0 - 1) \times 10^6 \times e^{-0.1057h}$$

The attenuation function is provided graphically in Fig. 9.1 in Ref. 10 for various values of N_s. At a nominal value $N_s = 301$, this function can be approximated for $0.017\theta d \leq 10$ by

$$F(\theta d) = 30 \log(\theta d) + 0.332\theta d + 135.82$$

The loss associated with the scattering process can be appreciated by a comparison with free space loss:

$$L_{\mathrm{FS}} = 32.446 + 20 \log f + 20 \log d$$

Thus at 1000 MHz and 200 km the free-space loss is 138.5 dB, whereas the corresponding troposcatter loss with a nominal scattering angle of 0.01 rad is 188.8 dB. This additional 50 dB loss has to be overcome with large antennas and power amplifiers.

The second component of troposcatter transmission loss, antenna gain loss, occurs with high-gain antennas. An increase in gain increases the illumination in the scattering common volume proportionally to the square of the beamwidth, but the number of illuminated scatterers in the common volume of beam intersection decreases as the cube of the beamwidth so there is an overall antenna gain loss proportional to the beamwidth. For antenna gains less than 55 dB and for gains that are not very different for the two antennas, the antenna gain loss in dB can be expressed as [5]

$$\Delta g = 0.07 \exp\{0.055(G_t + G_r)\}$$

where G_t and G_r are the free-space antenna gains in decibels (dB) of the transmitter and receiver, respectively. A more exact calculation of gain loss can be obtained by numerical integration of the scattering common volume defined by the antenna pattern distribution. Parl [11] has developed closed-form results for the path loss and its component antenna gain loss that agree well with the numerical integration results. Antenna gain loss is also called *aperture-to-medium coupling loss* and was the subject of numerous analyses [e.g., 12,13].

4.2. Variability

The transmission loss calculation for a troposcatter communication link provides a prediction of the median value over a measurement period that might be months or a year. Within that measurement period the received signal will undergo variations due to changes in the troposphere meteorologic conditions. Slow changes in temperature and humidity result in corresponding variations in refractive index that can be represented by hourly median values during the measurement period. The distribution of these hourly medians expressed in a dB measure has been found to be approximately Gaussian. The prediction of transmission loss corresponds to the median of this distribution and the standard deviation in dB completes the statistical definition of this long-term variation.

In the NBS method the prediction of the long-term variation is accomplished from a set of empirical curves [14] of variability $V(\rho, \theta)$ in dB as a function of the scatter angle θ and the link availability that the transmission median loss is not exceeded $\rho\%$ of the time. For example, at $\rho = 99.9\%$ the standard deviation of the Gaussian variability distribution is $V/3.1$.

It was recognized by system designers that the prediction of the transmission loss is subject to uncertainty. This uncertainty was expressed as a variance $\sigma_c^2(\rho)$ dependent on the link availability ρ. The system would be designed with a service probability $F(\tau)$, where τ is the standard normal deviate and the predicted transmission loss would be increased by $\tau\sigma_c(\rho)$. Empirical curves for the prediction uncertainty $\sigma_c(\rho)$ are given in Ref. 14.

There is also a variation within the hour due to changes in the pathlengths associated with individual scatterers in the common volume of the intersection of the antenna beams. This short term fading of the received signal envelope is due to multipath and has been found to follow a Rayleigh distribution. The system designer will normally require an hourly median value of transmission loss that includes a fade margin to ensure satisfactory performance under short term Rayleigh fading conditions. The Rayleigh distribution was defined and fade margins are discussed in Section 2.

4.3. Multipath Dispersion

The delay difference between the shortest and longest paths from the transmitter to receiver through the scattering common volume is the delay dispersion. In an analog system, telephone channels are at different frequencies in the composite transmitted signal. A delay difference might produce additive interference for one channel and subtractive interference for another channel. This type of interference and the nonlinear demodulation of frequency modulation produces intermodulation distortion in analog systems.

In a digital system, delay dispersion will cause interference between adjacent symbols, namely, intersymbol interference (ISI). If this ISI is not taken into account by the digital demodulation process, serious performance degradation will result. Adaptive receivers exploit this multipath phenomenon as a form of signal transmission redundancy; that is, the same transmitted information is associated with multiple delay paths.

The basic prediction method for delay dispersion is to calculate the path distances and their corresponding delay through the common volume. Sunde [15] uses an approximate formula to obtain the maximum departure from the mean transmission delay. Sunde also used this result to predict intermodulation distortion in analog systems [16].

In a channel model developed by Bello [17], a one-dimensional approximation to the common volume was used to derive estimates of the root-mean-square (RMS) value of multiple delay spread. Multipath measurements by Sherwood and Suyemoto [18] found two-sided RMS multipath spread, 2σ, to be considerably wider than predicted by the Bello model. They also observed considerable variation in measured 2σ values but no significant correlation with variables such as path loss, surface refractivity, and effective earth radius. Collin [19] measured the cumulative distribution of the 2σ multipath spread and approximated the distribution by a lognormal law. From these empirical data, 99% 2σ values were predicted and compared favorably with data on 0.9- and 4.8-GHz links.

Multipath spread variation has two major causes: changes in the effective earth radius and layering due to turbulence that modify the refractive index height profile within the scattering common volume. In the absence of empirical data on this height profile, a worst-case

maximum delay spread τ can be calculated [20] as

$$\tau = d(\alpha_{t1}\alpha_{t0} - \alpha_{r1}\alpha_{r0})$$

where d is the pathlength, c is the speed of light, and α_{t1}, $\alpha_{t0}(\alpha_{r1}, \alpha_{r0})$ are the maximum and minimum takeoff angles at the transmitter (receiver) measured from the straight line bisecting the transmitter and receiver to the lowest and highest points in the scattering volume, respectively. The worst-case multipath spread can be used to ensure that intermodulation noise does not limit capacity in an analog system and that adaptive systems are sufficiently robust in digital systems. For predicting nominal performance the yearly median 2σ value can be calculated [20] by three-dimensional integration for a fixed refractive index within the scattering common volume.

5. ANALOG SYSTEM PERFORMANCE

In an analog troposcatter link typically 60–120 4-kHz telephone channels are frequency division multiplexed to provide a baseband signal that is subsequently frequency modulated. Link performance can be determined by computing the median value of the voice channel signal-to-noise ratio (SNR) or an outage probability representing the fraction of time the SNR is below a critical threshold. The latter criterion is more meaningful in typical applications where the telephone call includes multiple troposcatter links. For small outage probabilities, the network outage probability is simply the sum of the link probabilities.

An analysis adapted from Ref. 21 is presented here for analog systems that include the effects of both signal-level variations and multipath delay distortion. The measure of performance selected is the signal-to-noise ratio r in a voice channel. The noise is defined to include the effects of thermal noise due to signal level variations and intermodulation (IM) noise due to multipath delay variations. The short-term probability is

$$P(\overline{C}, S) = \text{prob}(r < r_c) \tag{1}$$

where r_c is a critical value of r, the voice channel SNR, $\overline{C}$ is the average received unmodulated carrier power, and S is the 2σ multipath delay spread.

The effects of thermal noise and path intermodulation disturbance add on a noise power basis, so that we have

$$r = (x^{-1} + y^{-1})^{-1} \tag{2}$$

where x is the signal-to-thermal noise ratio and y is the signal-to-path IM ratio.

5.1. Thermal Noise Effects

The signal-to-thermal noise ratio has a mean value $\overline{x}$ from FM theory [22]:

$$\overline{x} = \frac{\overline{C}}{N_0 B}\left(\frac{f_d}{f_1}\right)^2 \frac{B}{N} g_p, \quad \frac{\overline{C}}{N_0 B} > T_{\text{FM}} \tag{3}$$

where B is the IF bandwidth, f_d is the RMS deviation of the voice channel signal, f_1 is the frequency location of the voice channel in the multiplex format, b is the voice channel bandwidth, and g_p is the preemphasis factor for the voice channel location. Equation (3) is valid provided the carrier-to-noise ratio $\text{CNR} = \overline{C}/N_0 B$ is greater than the FM threshold, T_{FM}. With threshold extension T_{FM} occurs at a CNR value of about 7 dB. Below FM threshold the signal-to-noise ratio drops rapidly with CNR. The random variable x can then be approximated by a piece-wise linear function of CNR in dB. For a general demodulator, we define

$$x = g(C) = \begin{cases} g_1(C) & \dfrac{C}{N_0 B} \geq T_{\text{FM}} \\ g_2(C) & \dfrac{C}{N_0 B} \prec T_{\text{FM}} \end{cases} \tag{4}$$

where C is the instantaneous carrier power and $\overline{C}$ is its short-term mean. The linear functions g_1 is given by Eq. (3) and g_2 is obtained from FM demodulator characteristics below threshold.

Since C is a short-term random variable, so is the signal-to-thermal noise ratio x. With optimum combining, C has a gamma PDF of order D, where D is the number of diversities. The probability distribution for C, that is, the probability that the carrier power after diversity combining is less than C, is given by

$$F_C(C) = e^{-DC/\overline{C}} \sum_{I=D}^{\infty} \frac{(DC/\overline{C})^I}{I!} \tag{5}$$

When path IM is negligible, the outage probability can be computed from knowledge of $F_C(A)$ by

$$P(\overline{C}, O) = F_C(g^{-1}(r_c)) \tag{6}$$

where $g^{-1}(x)$ is the inverse relation defined by (4).

5.2. Path Intermodulation Effects

In general the analysis becomes much more difficult when multipath delay effects have to be considered. Multipath delay variations result in an intermodulation noise when the distorted signal passes through the FM discriminator of an analog system. The performance due to this effect is summarized in a signal-to-intermodulation noise ratio (SINR). One commonly used approach for calculating the median SINR ratio is due to Sunde [16]. His results are expressed in terms of the noise power ratio (NPR) [2]. When this ratio is converted to voice channel signal-to-interference ratio, the result for the SINR ratio is

$$y(\Delta) = \left(\frac{f_d}{B}\right)^2 \frac{f_1}{b} \frac{1}{GH(\gamma)} \tag{7}$$

where B is the baseband bandwidth, G is the preemphasis factor, and $H(\gamma)$ is the degradation due to phase distortion γ. The preemphasis factor used by Sunde was $G = 0.192$. The phase distortion γ is a function of the baseband RMS deviation F_r and a delay departure Δ from the mean transmission delay:

$$\gamma = 8F_r^2\Delta^2 \tag{8}$$

In Sunde's original work the function $H(\gamma)$ is proportional to γ^2 (i.e., Δ^4) for small phase distortion values. The saturation effect is due to the use of only the quadratic phase distortion term and omission of higher-order terms. Since for large phase distortion, the multipath effect in an FM system should be proportional to the multipath power, it is more reasonable to have $H(\gamma)$ asymptotically be proportional to γ (i.e., Δ^2). This leads to a modified version for the phase distortion function of the form

$$H(\gamma) = \begin{cases} \gamma^2 & \gamma < \frac{1}{2} \\ \frac{1}{2}\gamma & \gamma > \frac{1}{2} \end{cases} \tag{9}$$

This function is identical to Sunde's below $\gamma = \frac{1}{2}$ and is approximately tangent to the saturation portion around $\gamma = 1$. Equations (7) and (9) thus allow one to compute the SINR value in terms of a multipath delay Δ and the FM system parameters. Unfortunately, little is known about the short-term distribution of Δ for which S is a statistic.

In Ref. 21 a short-term model is proposed that treats Δ as a random variable and the probability distribution F_Δ is derived as an implicit function of the multipath spread S.

The short-term outage probability due to path intermodulation effects alone can then be found from this multipath delay distribution and the relation between signal-to-interference ratio $y(d)$ as a function of delay by Eqs. (7) and (8). The outage probability for path IM alone is

$$P(\infty, S) = F_\Delta(\Delta(r_c)) \tag{10}$$

where $\Delta(y)$ is the inverse relation to Eq. (7) for delay in terms of SINR ratio.

5.3. Combined Thermal Noise and Path IM Effects

Given the probability distributions for the signal-to-thermal noise ratio x and the signal-to-intermodulation noise ratio y, it is a straightforward but tedious task to compute the probability distribution of the total SNR [Eq. (2)] for independent x and y values. The result is

$$\begin{aligned} p(\overline{C}, S) &= \text{prob}(r < r_c) \\ &= \int_0^{r_c} dF_c(g^{-1}(x)) \int_0^{(r_c^{-1} - x^{-1})^{-1}} dF_\Delta(\Delta(y)) \end{aligned} \tag{11}$$

In most cases however the outage probability is dominated by one effect or the other. Thus, a reasonable approximation is to use

$$p(\overline{C}, S) \doteq p(\overline{C}, O) + p(\infty, S) \tag{12}$$

as the short-term outage probability.

Empirical evidence [10,19] suggests that the long-term fluctuations in $\overline{C}$ and S can be described by a lognormal density function. Also the joint process appears to be sufficiently decorrelated that performance estimates can assume independence in terms of the means and standard deviation of $\overline{C}$ and S. Estimates of these parameters for $\overline{C}$ can be obtained in Ref. 10 and for S in Refs. 17–19.

6. DIGITAL SYSTEMS

The multipath delay spread limits the channel capacity that can be achieved in analog systems. Only transmission bandwidths less than the reciprocal of this multipath delay spread can be achieved. Signals of larger bandwidths become distorted due to the multipath dispersion. In FM systems this dispersion causes intermodulation noise after detection.

With digital signal formats, adaptive methods can be used to measure the multipath structure and exploit it as an extra form of diversity to improve performance. Unlike the capacity of analog systems, the capacity of digital systems is not restricted by the multipath delay spread. From a network viewpoint, fades in tandem digital links do not have a cumulative effect because the signal can be regenerated at each node.

Adaptive troposcatter systems have been demonstrated that are efficiently able to detect digital signals perturbed by a fading channel medium while tracking the fading variations. Applicability of adaptive signal processing techniques is critically dependent on whether the rate of fading is slower than the rate of signaling. As discussed below, troposcatter radio links can be considered to be slow-fading multipath channels.

6.1. Slow-Fading Multipath Channels

For digital communication over troposcatter radio links, an attempt is made to maintain transmission linearity; thus the receiver output should be a linear superposition of the transmitter input plus channel noise. This is accomplished by operation of the power amplifier in a linear region or, with saturating power amplifiers, by using constant-envelope modulation techniques. For linear systems, multipath fading can be characterized by a transfer function of the channel $H(f; t)$ that is the frequency domain response at a carrier of 0 Hz as a function of time t. Let t_d and f_d be the decorrelation separations in the time and frequency variables, respectively. If t_d is a measure of the time decorrelation in seconds, then

$$\sigma_t = \frac{1}{2\pi t_d} \text{ Hz}$$

is a measure of the fading rate or bandwidth of the random channel. The quantity σ_t is often referred to as the *Doppler spread* because it is a measure of the width of the received spectrum when a single sine wave is transmitted through the channel. The dual relationship for the frequency decorrelation f_d in hertz suggests that a delay variable

$$\sigma_f = \frac{1}{2\pi f_d} \text{ s}$$

(in seconds) defines the extent of the multipath delay. The quantity σ_f is approximately equal to the RMS multipath delay spread Φ that represents the RMS width of the received process in the time domain when a single impulse function is transmitted through the channel.

Typical values of Doppler and delay spread for troposcatter communication are around 1 Hz and 100 ns, respectively.

The spreads can be defined as moments of spectra in a channel model [24] that assumes wide-sense stationary (WSS) in the time variable and uncorrelated scattering (US) in the multipath delay variable. This WSSUS model and the assumption of Gaussian statistics for $H(f;t)$ provide a statistical description in terms of a single two-dimensional correlation function of the random process $H(f;t)$.

This characterization has been quite useful and accurate for a variety of radio link applications. However the stationary and Gaussian assumptions are not necessary for the utilization of adaptive signal processing techniques on these channels. What is necessary is first that sufficient time exists to "learn" the channel characteristics before they change, and second, that decorrelated portions of the frequency band be excited such that a diversity effect can be realized. These conditions are reflected in the following two relationships in terms of the previously defined channel factors, the data rate R and the bandwidth B.

$$R \text{ (bps)} \gg \sigma_t \text{(Hz) learning requirement}$$

$$B \text{ (Hz)} \geq f_d \text{(Hz) diversity requirement}$$

The learning requirement insures that the channel remains approximately fixed for an interval containing many received bits. The energy in these received bits provides a basis for measurement of the channel characteristics. If $R \sim \sigma_t$, the channel would change before significant energy for measurement purposes could be collected. The signal processing techniques in an adaptive receiver do not necessarily need to measure the channel directly in the optimization of the receiver, but the requirements on learning are approximately the same. If only information symbols are used in the sounding signal, the learning mode is referred to as decision-directed. When digital symbols known to both the transmitter and receiver are employed, the learning mode is called *reference-directed*. Adaptation of troposcatter receivers with no wasted power for sounding signals can be accomplished using the decision-directed mode. This is possible because of the small number of adaptation parameters, the continuous nature of the communication, and the high likelihood that receiver decisions are correct.

As described in Section 2, diversity in fading applications is used to provide redundant communications channels so that when some of the channels fade, communication will still be possible over the others that are not in fade. These diversity techniques are sometimes called explicit diversity because of their externally visible nature. An alternate form of diversity is termed implicit diversity because the channel itself provides redundancy. In order to capitalize on this implicit diversity for added protection, receiver techniques have to be employed to correctly assess and combine the redundant information. The potential for implicit frequency diversity arises because different parts of the frequency band fade independently. Thus, while one section of the band may be in a deep fade, the remainder can be used for reliable communication. However, if the transmitted bandwidth B is small compared to the frequency decorrelation interval f_d, the entire band will fade and no implicit diversity can result. Thus, the second requirement $B \geq f_d$ must be met if an implicit diversity gain is to be realized. In diversity systems a little decorrelation between alternative signal paths can provide significant diversity gain. Thus it is not necessary for $B \gg f_d$ in order to realize implicit frequency diversity gain, although the implicit diversity gain clearly increases with the ratio B/f_d. Note that the condition $R \ll B \geq f_d$ does not preclude the use of implicit diversity because a bandwidth expansion technique can be used in the modulation process to spread the transmitted information over the available bandwidth B. Most digital troposcatter applications, however, are high-rate where R and B are about the same.

The implicit diversity effect described here results from decorrelation in the frequency domain in a slow-fading ($R \gg \sigma_t$) application. This implicit frequency diversity can in some circumstances be supplemented by an implicit time diversity effect, which results from decorrelation in the time domain. In fast-fading applications ($R > \sigma_t$) redundant symbols in an error-correcting coding scheme can be used to provide time diversity provided the codeword spans more than one fade epoch. In our slow-fading application this condition of spanning the fade epoch can be realized by interleaving the codewords to provide large time gaps between successive symbols in a particular codeword. The interleaving process requires the introduction of signal delay longer than the time decorrelation separation t_d. Methods of realizing implicit diversity gain with decoding delays as short as $\frac{1}{4}$ second have been studied [25]. However in many practical applications that require transmission of digitized speech over multiple troposcatter links, the required time delay is unsatisfactorily long for two-way speech communication. For this reason implicit frequency rather than time diversity techniques are used in existing systems. The receiver structures to be discussed next are applicable to situations where implicit frequency diversity can be realized.

6.2. Digital Receivers

When the transmitted symbol rate is on the order of the frequency decorrelation interval of the channel, the frequencies in the transmitted pulse will undergo different gain and phase variations resulting in reception of a distorted pulse.

Although there may have been no intersymbol interference (ISI) at the transmitter, the pulse distortion from the channel medium will cause interference between adjacent samples of the received signal. In the time domain, ISI can be viewed as a smearing of the transmitted pulse by the multipath causing overlap between successive pulses. The condition for ISI can be expressed in the frequency domain as

$$T^{-1}\text{(Hz)} \geq f_d\text{(Hz)}$$

or in terms of RMS multipath delay spread as

$$T\text{(seconds)} \leq 2\pi\sigma\text{(seconds)}$$

Bandwidth limitations and SNR efficiency in the presence of multipath fading make quadrature phase shift keying (QPSK) a common choice for modulation. Since the bandwidth of a QPSK signal is at least on the order of the symbol rate T^{-1} Hz, there is no need for bandwidth expansion under ISI conditions in order to provide signal occupancy of decorrelated portions of the frequency band for implicit diversity. However, it is not obvious whether the presence of the intersymbol interference can wipe out the available implicit diversity gain. It has been established that adaptive receivers can be used that cope with the intersymbol interference and in most cases wind up with a net implicit diversity gain. These receiver structures fall into three general classes: matched filters, equalizers, and maximum likelihood detectors.

6.2.1. Matched Filters. A matched filter has characteristics that "match" the characteristics of the combined transmitter and channel. If the combined frequency response is denoted $H(f)$, the matched filter is $H^*(f)$. In the time domain the combined impulse response can be denoted as $h(t)$. The matched filter impulse response is an inverted and conjugated response: $h^*(-t)$. The matched filter can also be realized by correlating the received signal with the combined impulse response. Since the channel part of the combined response is changing and unknown to the receiver, adaptation is required. Also the process of matched filtering increases the total impulse response thus increasing intersymbol interference (ISI) effects. A reduction in the ISI can be realized by transmitting a shorter pulse and leaving a time gap before the next pulse begins.

Figure 4 illustrates an adaptive matched filter based on these principles. Data modulation is stripped from the received signal by multiplying by previous decisions. The resulting received pulse is then averaged in a recirculating delay line to produce an estimate of $h(t)$. This channel estimate is correlated with the received signal to complete the matched filter operation. The

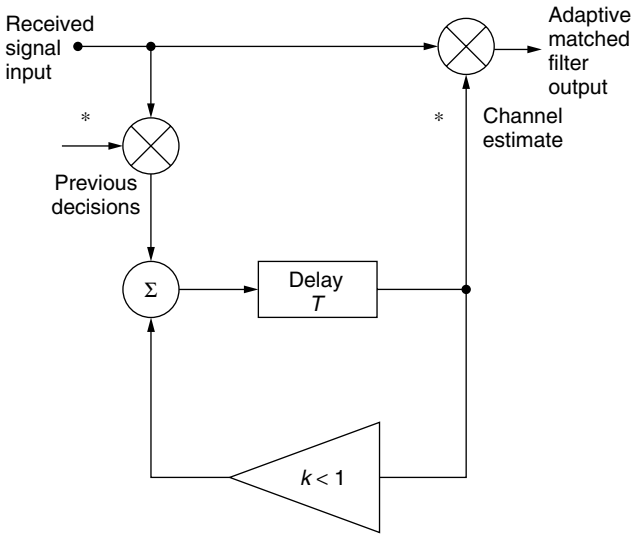


Figure 4. Adaptive matched filter for QPSK system.

adaptive matched filter shown in Fig. 4 will degrade when there is intersymbol interference between received pulses because the averaging process would add overlapped pulses incoherently. When the multipath spread is less than the symbol interval, this condition can be alleviated by transmitting a time-gated pulse whose OFF time is approximately equal to the width of the channel multipath. The multipath causes the gated transmitted pulse to be smeared out over the entire symbol duration but with little or no intersymbol interference. Because the multipath components are adaptively combined an implicit diversity [26] effect is obtained. In a configuration with both explicit and implicit diversity, moderate intersymbol interference can be tolerated because the diversity combining adds signal components coherently and ISI components incoherently.

Because the OFF time of the pulse cannot exceed 100%, this approach is clearly data-rate-limited for fixed multipath conditions. In addition, the time gating at the transmitter results in an increased bandwidth that may be undesirable in a bandwidth-limited application. The power loss in peak power limited transmitters due to time gating can be partially offset by using two carrier frequencies with independent data modulation [27]. This technique was successfully developed by Raytheon for use in the U.S. Air Force tactical troposcatter system, AN/TRC-170. In this system a V2 model provides a 2S/2F diversity configuration with traffic capability of up to 120 voice channels and a range between 60 and 140 mi. A smaller V3 model uses a dual-frequency configuration. The AN/TRC-170 has hundreds of units in the U.S. military inventory, it has been sold to other countries, and it continues to be used in tactical military applications.

6.2.2. Adaptive Equalizers. Adaptive equalizers use linear filter subsystems with electronically adjustable parameters that are controlled in an attempt to combine multipath components and compensate for intersymbol interference. Tapped delay-line filters are a common choice for the equalizer structure as the tap weights provide a convenient adjustable parameter set. Adaptive equalizers have been widely employed in telephone channel applications [28] to reduce ISI effects due to channel filtering. In a fading multipath channel application, the equalizer can provide three functions simultaneously: noise filtering, matched filtering for explicit and implicit diversity, and removal of ISI. These functions are accomplished by adapting a tapped delay-line filter (TDF) to force an error measure to a minimum. By designing the error measure to include the degradation due to correlated noise, ISI, filtering, and improper diversity combining, the TDF will minimize their combined effects.

A linear equalizer (LE) is defined as an equalizer that linearly filters each of the N explicit diversity inputs. An improvement to the LE is realized when an additional filtering is performed on the detected data decisions. Because it uses decisions in a feedback scheme, this equalizer is known as a *decision-feedback equalizer* (DFE).

The operation of a matched filter receiver, an LE, and a DFE can be compared from examination of the received

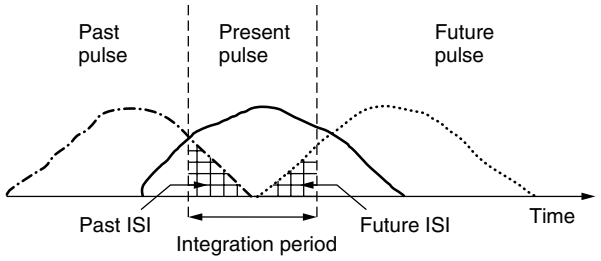


Figure 5. Received pulse sequence.

pulsetrain example of Fig. 5. The binary modulated pulses have been smeared by the channel medium producing pulse distortion and interference from adjacent pulses. Conventional detection without multipath protection would integrate the process over a symbol period and decide whether a $+1$ was transmitted if the integrated voltage is positive and -1 if the voltage is negative. The pulse distortion reduces the margin against noise in that integration process. A matched filter correlates the received waveform with the received pulse replica, thus increasing the noise margin. The intersymbol interference arises from both future and past pulses in these radio systems since the multipath contributors near the mean path delay normally have the greatest strength. This ISI can be compensated for in a linear equalizer by using properly weighted time-shifted versions of the received signal to cancel future and past interferers. The DFE uses time-shifted versions of the received signal only to reduce the future ISI. The past ISI is canceled by filtering past detected symbols to produce the correct ISI voltage from these interferers. The matched filtering property in both the LE and DFE is realized by spacing the taps on the received signal TDFs at intervals smaller than the symbol period.

The DFE is shown in Fig. 6 for the Nth-order explicit diversity system. A forward filter (FF) TDF is used for each diversity branch to reduce correlated noise effects, provide matched filtering and proper weighting for explicit diversity combining, and reduce ISI effects. After diversity combining, demodulation, and detection, the data decisions are filtered by a backward filter TDF to eliminate intersymbol interference from previous pulses. Because the backward filter compensates for this "past" ISI, the forward filter need only compensate for "future" ISI.

A decision-directed error signal for adaptation of the DFE is shown as the difference between the detector input and output. Qualitatively one can see that if the DFE is well adapted, this error signal should be small. Reference-directed adaptation can be accomplished by multiplexing a known bit pattern into the message stream for periodic adaptation.

When error propagation due to detector errors is ignored, the DFE has the same or smaller mean-square error than the LE for all channels [29]. The error propagation mechanism has been examined by a Markov chain analysis [30] and shown to be negligible in practical fading applications. Also in an Nth-order diversity application, the total number of TDF taps is generally less for the DFE than for the LE. This follows because the former uses only one backward filter after combining of the diversity channels in the forward filter.

The performance of a DFE on a fading channel can be predicted [31] using a transformation technique that converts implicit diversity into explicit diversity and treats the ISI effects as a Gaussian interferer. An example of this calculation for a quadruple diversity system with $2\sigma/\mathrm{T} = 0$ and 0.5 is given in Ref. 32. The average probability of error is poorest when there is no multipath spread because there is no implicit diversity. The performance difference between a three-tap (3T) forward filter and an ideal matched filter was also shown to be less than a dB. Thus with moderate $2\sigma/\mathrm{T}$ values the DFE performance is close to optimum.

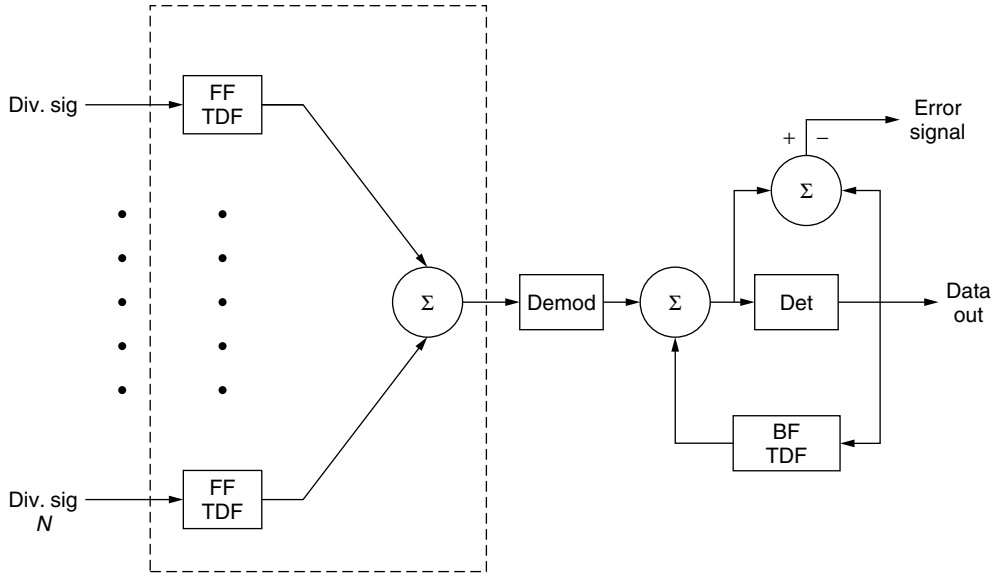


Figure 6. Decision-feedback equalizer, Nth-order diversity.

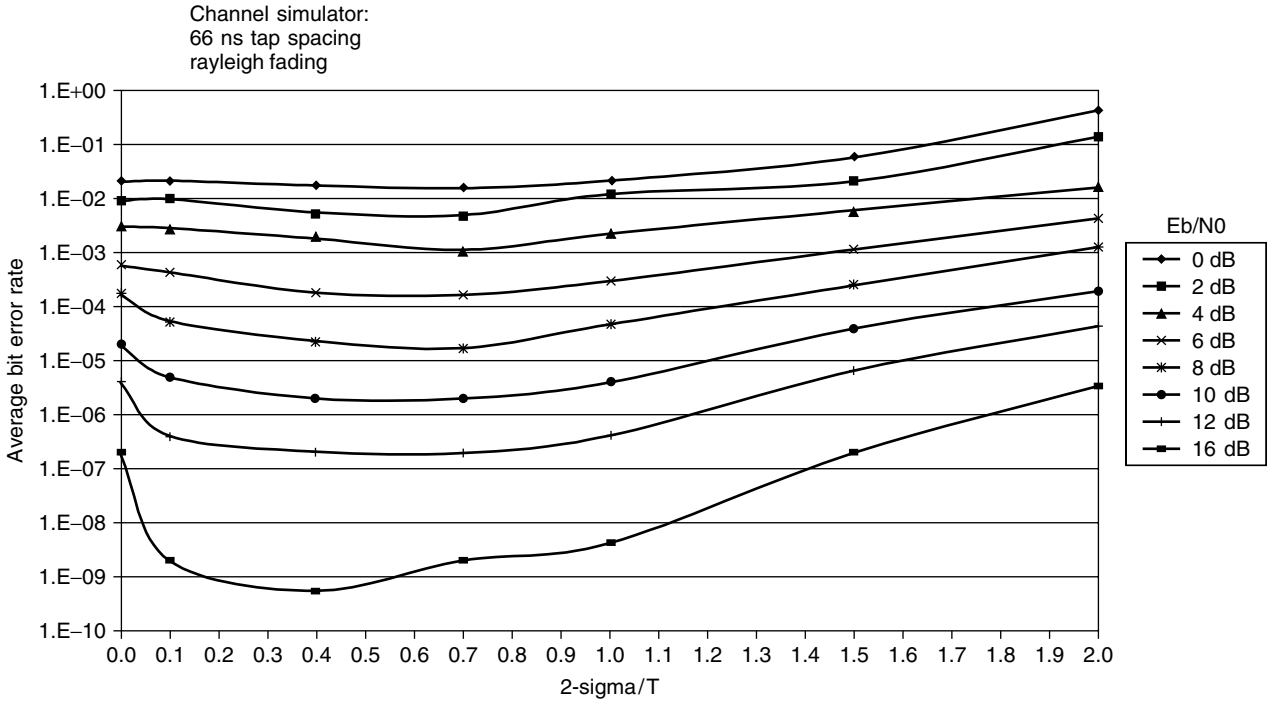


Figure 7. Diversity DFE performance versus. multipath spread QPSK modulation at data rates of 3.2, 6.4, and 9.7 Mbps: three forward filter taps at T/2 spacing, three backward filter taps, Eb/N0 in dB/diversity.

A DFE modem was developed [33] with data rates up to 12.5 Mbps for application on troposcatter channels with up to four orders of diversity. This DFE modem uses only a three-tap forward filter TDF and a three-tap backward filter TDF. Extensive simulator and field tests [31,33,34] have shown that implicit diversity gain is realized over a wide range of actual conditions while ISI effects are mostly eliminated. Figure 7, reproduced from Fig. 6.5 of Ref. 34, summarizes simulated performance as a function of 2σ multipath spread and signal-to-noise ratio. The improvement in bit error rate when 2σ/T increases from zero is due to implicit diversity. For larger values of 2σ/T approaching 2, the ISI begins to dominate the implicit diversity effect. Measured results agree well with the predicted performance from Ref. 31.

An 8-Mbps version of the DFE modem is produced commercially by Comtech Systems Inc., Orlando FL and has been incorporated in both military and civilian tropospheric scatter radio systems. Applications of the latter include ocean oil platforms and a backbone angle diversity system between islands in the Bahamas for cellular telephone service.

6.2.3. Maximum-Likelihood Detectors. The DFE is not optimum for all channels with respect to bit error probability. By considering intersymbol interference as a conventional code defined on the real line (or complex line for bandpass channels), maximum-likelihood sequence estimation algorithms have been derived [35,36] for the linear modulation channel. These algorithms provide a decoding procedure for receiver decisions that minimize the probability of sequence error. A maximum-likelihood sequence estimator (MLSE) receiver requires matched filters for each diversity channel and a combiner. After these filtering and combining operations, a trellis decoding technique is used to find the most likely transmitted sequence.

The MLSE algorithm works by assigning a state for each intersymbol interference combination. Because of the one-to-one correspondence between the states and the ISI, the maximum-likelihood source sequence can be found by determining the trajectory of states.

If some immediate state is known to be on the optimum path, then the maximum-likelihood path originating from that state and ending in the final state will be identical to the optimal path. If at time n, each of the states has associated with it a maximum likelihood path ending in that state, it follows that sufficiently far in the past history will not depend on the specific final state to which it belongs. The common path history is the maximum-likelihood state trajectory [36].

Since the number of ISI combinations and thus the number of states is an exponential function of the multipath spread, the MLSE algorithm has complexity that grows exponentially with multipath spread. The equalizer structure exhibits a linear growth with multipath spread. In return for additional complexity, the MLSE receiver results in a smaller (sometimes zero) intersymbol interference penalty for channels with isolated and deep frequency selective fades. However in many applications where high orders of diversity are employed, these deep selective frequency fades do not occur frequently enough to significantly affect the average error probability [33]. Partly for these reasons the MLSE

technique for troposcatter applications never advanced beyond the experimental model phase.

7. PRESENT APPLICATIONS OF TROPOSCATTER SYSTEMS

Satellite and cable systems can better provide for large traffic capacities or generally high data rates. However, for links with total data rate of about 10 Mbps, the fixed cost of a digital troposcatter system–satellite transponder lease cost compares favorably enough that troposcatter systems continue to be implemented. Ocean oil platforms and backbone systems for cellular telephone service have been cited here as examples. The utility of troposcatter systems in a tactical military application is also still in evidence. One also notes applications in smaller countries that do not have the technology or finances to implement their own satellite system and are reluctant to lease satellite services that could be monitored or terminated for political reasons. Finally there are applications in disaster relief and in data internet services to remote locations.

BIOGRAPHY

Peter Monsen received a B.S. in electrical engineering from Northeastern University in 1962, a M.S. in operations research from Massachusetts Institute of Technology in 1963, and the Doctor of Engineering Science in electrical engineering from Columbia University in 1970.

From 1964 to 1966 Dr. Monsen served as a Lieutenant in the U.S. Army at the Defense Communications Agency. From 1966 to 1972 he was employed by AT&T Bell Laboratories as a supervisor of a Transmission Studies Group working on fading-channel characterization and adaptive equalization. During this period, the optimum decision-feedback equalizer and a method to adapt it on a time-varying radio channel were developed. From 1972 to 1984 he was employed by SIGNATRON, Inc., working on a wide range of signal processing techniques including adaptive equalization, diversity combining, dual polarization utilization, and interference cancellation. This work led to the conception and development of a 12.6 Mb/s adaptive equalizer troposcatter modem for Defense Communication System and NATO applications with a first operational link between Berlin and West Germany.

In 1984, PM Associates was founded through which consulation for a broad range of telecommunication companies has been provided. Independently, Dr. Monsen has recently submitted three patent applications on multiple access systems with unity reuse factor.

BIBLIOGRAPHY

1. Van Der Pool and H. Bremmer, The diffraction of electromagnetic waves from an electrical point source around a finitely conducting sphere, with applications to radiotelegraphy and the theory of the rainbow, *London, Edinburg, Dublin Philos. Maga. J. Sci.* **24**(164): 141 (July 1937) (series 7).
2. P. F. Panter, *Communications System Design-Line of Sight and Troposcatter Systems*, McGraw-Hill, New York, 1972.
3. W. E. Gordon, Radio scattering in the troposphere, *Proc. IRE* **43**: 23 (Jan. 1955).
4. F. duCartel, *Propagation Tropospherique et Faisceaux Hertiziens Transhorizon*, Cheron, Paris, 1961.
5. G. Roda, *Troposcatter Radio Links*, Artech House, Norwood, MA, 1988.
6. R. Larsen, Quadruple space diversity in troposcatter systems, *Marconi Rev.* 28–55 (first quarter 1980).
7. G. Krause and P. Monsen, Results of an angle diversity test experiment, *Natl. Telecommunications Conf. Record*, 1978.
8. P. Monsen and S. Parl, *Adaptive Antenna Control (AAC) Program*, Final Report, ADA091764, Aug. 1980.
9. CCIR, *The Concept of Transmission Loss in Studies of Radio Systems*, documents of the Xth Plenary Assembly, ITU, Geneva, Vol. III, Recommendation 341, 1963.
10. P. L. Rice, A. G. Longley, K. A. Norton, and A. P. Barsis, *Transmission Loss Predictions for Tropospheric Communication Circuits*, Tech. Note 101, Vols. I–II, National Bureau of Standards, Boulder, CO, 1965.
11. S. Parl, New formulas for tropospheric scatter path loss, *Radio Sci.* **14**(1): 49–57 (Jan.–Feb. 1979).
12. L. P. Yeh, Experimental aperture-to-medium coupling loss, *Proc. IRE* **50**: 203 (1962).
13. L. Boithias and J. Battesti, Etude experimental de la baisse du gain d'antenna dans less liaisons transhorizon, *Ann. Telecommun.* **19**(9–10): 221,229 (1964).
14. A. F. Barghausen et al., *Ground Telecommunication Performance Standards*, NBS Report 6767, Boulder, CO, Part 5 of 6, Tropospheric Systems, Chaps. 5 and 6, 15 June 1961.
15. E. D. Sunde, Digital troposcatter transmission and modulation theory, *Bell Syst. Tech. J.* 143–214 (Jan. 1964).
16. E. D. Sunde, Intermodulation distortion in analog FM troposcatter systems, *Bell Syst. Tech. J.* 399–435 (Jan. 1964).
17. P. A. Bello, A troposcatter channel model, *IEEE Trans. Commun. Technol.* **COM-17**: 130–137 (April 1969).
18. A. Sherwood and L. Suyemoto, *Multipath Measurements over Troposcatter Paths*, Mitre Report MTP-170, Bedford, MA, April 1976.
19. C. Collin, *Empirical Evaluation of the Correlation Bandwidth in Troposcatter Paths*, Revue Technique Thomson-CSF, Vol. 11, No. 3, Sept. 1979.
20. P. Monsen et al., *Digital Troposcatter Performance Model: Final Report*, Signatron, Inc., Lexington, MA, par. 2.5.5, Dec. 1983.
21. P. Monsen, Analog and digital transmission performance models for troposcatter links, *AGARD Conf. Electromagnetic Wave Propagation Panel Symp.*, Copenhagen, Denmark, May 1982.
22. E. D. Sunde, *Communications Systems Engineering Theory*, Wiley, New York, 1969, Chap. 4.
23. P. Monsen and J. Eschle, *Test Report Adaptive Antenna Control Program*, ADA04416, U.S. Army, Ft. Monmouth NJ, Sept. 1980.
24. P. A. Bello, Characterization of randomly time-variant linear channels, *IEEE Trans. Commun. Syst.* **CS-11**: 360,393 (Dec. 1963).
25. D. Chase, P. A Bello, L-J. Weng, and R. W. Spencer, *Troposcatter Interleaver Study Report*, RADC TR-75-19, Griffiss AFB NY, Feb. 1975.

26. M. Unkauf and O. A. Tagliaferri, An adaptive matched filter modem for digital troposcatter, *Conf. Rec., Int., Conf. Commun.*, June 1975.

27. M. Unkauf and O. A. Tagliaferri, Tactical digital troposcatter systems, *Conf. Rec. Nat. Telecomm. Conf.*, Vol. 2, Dec. 1978, pp. 17.4.1–17.4.5.

28. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*, McGraw-Hill, New York, 1986, Chap. 6.

29. P. Monsen, Feedback equalization for fading dispersive channels, *IEEE Trans. Inform. Theory* **1T-17**: 56–64 (Jan. 1971).

30. P. Monsen, Adaptive equalization of the slow fading channel, *IEEE Trans. Commun.* **COM-22**: (Aug. 1974).

31. P. Monsen, Theoretical and measured performance of a DFE modem on a fading multipath channel, *IEEE Trans. Commun.* **COM-25**(10): (Oct. 1977).

32. P. Monsen, Fading channel communications, *IEEE Commun. Mag.* 16–25 (Jan. 1980).

33. D. H. Kern and P. Monsen, *MDTS Final Report*, Rep. ECOM-0040-F ECOM, Fort Monmouth, NJ, July 1976.

34. J. B. Gadoury, II, *Error Performance Characterization Study of a Digital Troposcatter Modem* (*MD-918/GRC*), Defense Communications Agency, Washington, DC, Final Report, July 2, 1983.

35. G. D. Forney, Jr., Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference, *IEEE Trans. Inform. Theory* **IT-18**: 363–377 (May 1974).

36. G. Ungerboeck, Adaptive maximum-likelihood receiver for carrier-modulated data transmission systems, *IEEE Trans. Commun.* **COM-22**: 624–636 (May 1974).

TURBO CODES

Claude Berrou
Alain Glavieux
ENST Bretagne
Brest, France

1. INTRODUCTION

Turbo codes are error-correcting codes that are able to perform in conditions close to the theoretical limits predicted by C. E. Shannon [1]. They were presented to the scientific community in 1993 [2] and at first aroused a certain amount of skepticism, because it was believed at that time that reconciling theory and practice in the matter of channel coding and spectral efficiency would be a longer-term task. In addition, the concepts involved in Turbo coding and decoding, although not revolutionary, were not very familiar to most people in the field. The invention of Turbo codes was the result of a pragmatic construction conducted by C. Berrou and A. Glavieux, based on the intuitions of some European researchers, G. Battail, J. Hagenauer, and P. Hoeher, who in the late 1980s aroused the interest of probabilistic processing in digital communication receivers [3–6]. Previously other researchers, mainly R. Gallager [7] and M. Tanner [8], had already imagined coding and decoding techniques whose general principles are closely related to those of Turbo codes. Since 1993, Turbo codes have been widely studied, and adopted in several communication systems, and the inherent concepts of the "Turbo" principle have been applied to topics other than error-correcting coding, such as demodulation, detection and multidetection, and equalization.

This article is organized as follows. The next section gives a practical point of view concerning the search for good codes. Sections 3 and 4 are devoted to the building block in Turbo code structure: the Recursive Systematic Convolutional code. Sections 5 and 6 deal with code construction and the permutation issue, and Section 7 presents the Turbo decoding algorithm. Finally, some applications and examples of performance are given.

2. WHAT IS A GOOD ERROR-CORRECTING CODE?

Figure 1 represents, as well as performance without coding, three possible behaviors for an error-correcting coding scheme on an additive white Gaussian noise (AWGN) channel with BPSK or QPSK modulation and rate-$\frac{1}{2}$ coding. To be concrete, the information block is assumed to be around 188 bytes (MPEG application). The error probability P_e of the "no coding" performance is given by the complementary error function, as a function of E_b/N_0, where E_b is the energy per information bit and N_0 is the monolateral noise density:

$$P_e = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\right); \operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty \exp(-u^2)\,du \quad (1)$$

Behavior 1 (in Fig. 1) corresponds to the ideal Shannon system, with the theoretical limit around 0.8 dB, estimated from the report by Dolinar et al. [9]. It assumes random coding and that the minimum distance $d_{\min}$, deducted from the Gilbert–Varshamov bound [10], would be around 380. The theoretical asymptotic gain, approximated by

$$G_a \approx 10\log(Rd_{\min}) \quad (2)$$

would then be higher than 22 dB.

Behavior 2 has good convergence and low $d_{\min}$. This is, for instance, what is obtained with Turbo coding when the permutation function is not properly designed. Good convergence means that the bit error rate (BER) decreases noticeably, close to the theoretical limit, and low $d_{\min}$ brings about a severe change in the slope, due to an insufficient asymptotic gain. This gain is around 7.5 dB in the example given, and is reached at medium BER ($\approx 10^{-7}$). Below that, the curve remains parallel to the "no coding" one. A possible solution to overcome this *flattening* involves using an outer code, like the Reed–Solomon (RS) code, provided that the statistics of errors is suitable, at the output of the inner decoder, which is not obvious with Turbo codes.

Behavior 3 has poor convergence and high $d_{\min}$. This is representative of a decoding procedure that does not take advantage of all the information available at the receiver side. A typical example is the classical concatenation of

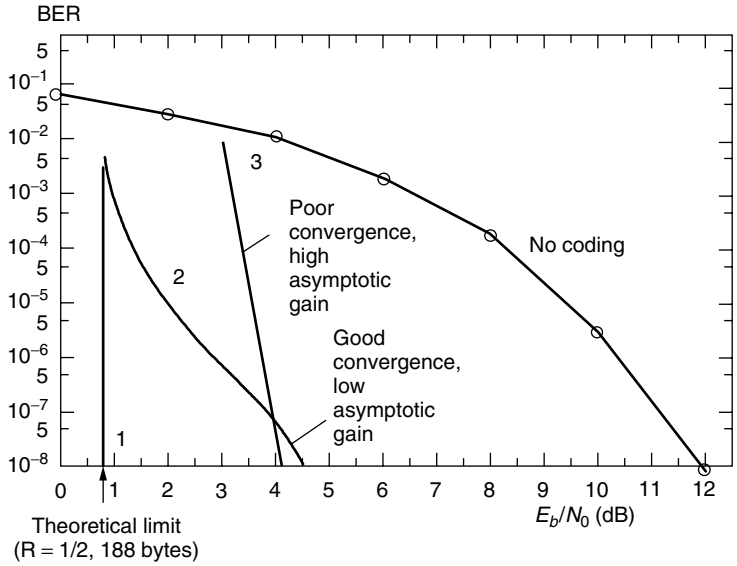


Figure 1. Possible behaviors for a coding/decoding scheme.

an RS code and a simple convolutional code. Whereas the minimum distance may be very large (but depending on the interleaver depth between the outer and the inner codes), the decoder is clearly suboptimal because the convolutional inner decoder does not take advantage of the RS redundant symbols.

The search for the perfect coding/decoding scheme has always faced the "convergence versus $d_{\min}$" dilemma. Usually, improving one of either aspect, in some more or less relevant way, weakens the other.

Turbo codes constituted a real breakthrough with regard to the convergence problem. But it was only several years later that, in addition to good convergence properties, sufficient minimum distances were achieved, which then led to powerful standalone channel coding. As can again be observed from Fig. 1, minimum distances as large as those given by random coding are not necessary. Hence, $d_{\min} = 25$ for a 188-byte block with rate $\frac{1}{2}$ would be sufficient to reach the theoretical limit at BER $= 10^{-8}$, instead of the value 380 given by random coding; $d_{\min} = 32$ would be necessary at BER $= 10^{-10}$.

To conclude, if a good code has to possess some random properties in order to display the same convergence threshold as random coding, its minimum distance may, in practical cases, be much more reasonable.

3. RECURSIVE SYSTEMATIC CONVOLUTIONAL (RSC) CODES

Pseudorandom generators are widely used in digital communications circuits for encrypting, scrambling, randomizing, spreading, encoding, and other functions. Figure 2 represents a pseudorandom generator, a scrambler, and an RSC encoder, all based on the same linear feedback register (LFR) structure. The register length is denoted ν, also called the *code memory* in the context of coding, and the register state is the vector $\mathbf{S} = (s_1, \ldots, s_j, \ldots, s_\nu)$. An RSC code is completely defined by two D polynomials with degree ν, where D is the delay

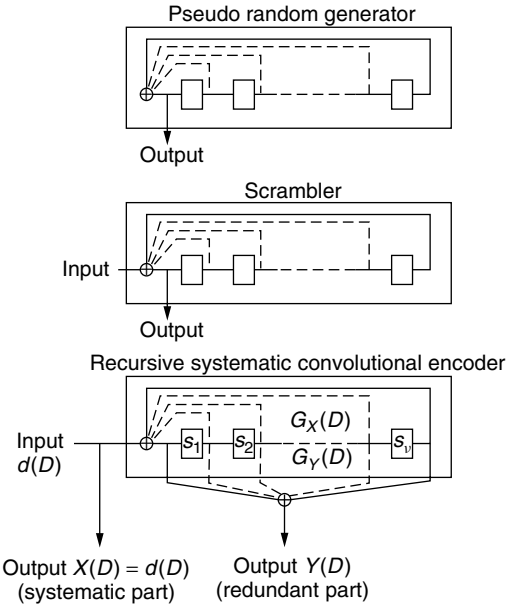


Figure 2. The same linear feedback register (LFR) structure is used for different fundamental operations, such as pseudorandom generation, scrambling, and encoding.

operator:

$$\begin{aligned} G_X(D) &= 1 + \sum_{j=1}^{\nu-1} G_X^{(j)} D + D^\nu \\ G_Y(D) &= 1 + \sum_{j=1}^{\nu-1} G_Y^{(j)} D + D^\nu \end{aligned} \tag{3}$$

where $G_X(D)$ and $G_Y(D)$ are the recursivity and the redundancy polynomials, respectively. $G_X^{(j)}$ (resp. $G_Y^{(j)}$) is equal to 1 if the register tap at level j $(1 \le j \le \nu - 1)$ is used in the construction of recursivity (resp. redundancy), and 0 otherwise. $G_X(D)$ and $G_Y(D)$ are generally defined in octal forms. For instance, $1 + D^3 + D^4$ is referenced as polynomial 23.

In this section, we consider only the encoding of semi-infinite-length information sequences, beginning at discrete time $i = 0$ and never ending. In addition to the systematic part of the output: $X(D) = d(D)$, the encoder yields redundancy $Y(D)$, given by

$$Y(D) = \frac{G_Y(D)}{G_X(D)} d(D) \tag{4}$$

Note that actual redundancy depends not only on the message but also on the LFR initial state, because of the recursivity of the encoder.

Thanks to the linearity property, the code characteristics are expressed with respect to the "all zero" sequence. In this case, any nonzero sequence $d(D)$, accompanied by redundancy $Y(D)$, will represent a possible error pattern for the coding/decoding system. Relation (4) indicates that only a fraction of sequences $d(D)$, which are multiples of $G_X(D)$, lead to finite-length redundancy. We call these particular sequences *return-to-zero* (RTZ) sequences [11], because they force the encoder, if initialized in state 0, to retrieve this state after the encoding of $d(D)$. In what follows, we will be interested only in RTZ patterns, assuming that the decoder will never decide in favor of a sequence whose distance from the "all zero" sequence is infinite. The fraction of sequences $d(D)$ which are RTZ, is exactly

$$p(\mathrm{RTZ}) = 2^{-\nu} \tag{5}$$

because the encoder has 2^ν possible states and an RTZ sequence finishes systematically at state 0. Denoting $p(\mathrm{NRTZ})$ as the proportion of non-RTZ sequences ($p(\mathrm{NRTZ}) = 1 - p(\mathrm{RTZ})$), we have

$$\frac{p(\mathrm{NRTZ})}{p(\mathrm{RTZ})} = 2^\nu - 1 \tag{6}$$

This is also the maximum possible value for the period L of the pseudorandom generator from which the RSC encoder is derived. This maximum value is obtained if $G_X(D)$ is a prime polynomial.

The shortest RTZ sequence is $G_X(D)$ and any RTZ sequence may be expressed as

$$\mathrm{RTZ}(D) = \sum_{i=0}^{\infty} a_i D^i G_X(D) \tag{7}$$

where a_i takes value 0 or 1. The minimum number of "1"s belonging to a RTZ sequence is 2. This is because $G_X(D)$ is a polynomial with at least two nonzero terms, and Eq. (5) then guarantees that $\mathrm{RTZ}(D)$ also has at least two nonzero terms. In general, the number of "1"s in a particular RTZ sequence is called the *input weight*, or simply the *weight*, and is denoted w. We then have $w_{\min} = 2$ for RSC codes, and the RTZ sequences with weight 2 are of the general form

$$\mathrm{RTZ}_2(D) = D^\tau (1 + D^{pL}) \tag{8}$$

where τ is the starting time, p any positive integer, and L the period of the encoder.

RTZ sequences with weight 3 may either exist or not, depending on the expression of $G_X(D)$. For instance, symmetric forms of $G_X(D)$, like polynomial 37, which was used in the first studies on Turbo coding [2], preclude odd values for w. RTZ sequences with even weights always exist, especially of the form

$$\mathrm{RTZ}_{2l}(D) = \sum_{j=1}^{l} D^{\tau_j} (1 + D^{p_j L}) \tag{9}$$

that is, as a combination of l any weight 2 RTZ sequences. This sort of composite RTZ sequence has to be considered closely when trying to design good permutations for Turbo codes.

RSC codes are decoded with the same trellis approach as classical (nonrecursive, nonsystematic) convolutional codes. The decoding relies either on the Viterbi algorithm [12] for hard-output operation, or the *maximum a posteriori* (MAP) algorithm, also called *a posteriori probability* (APP) or BCJR from the names of its inventors [13], or its simplified versions [14], for soft-output computation, as required by Turbo decoding. The Viterbi algorithm may also be adapted for soft-output decoding purposes [4,5].

Compared with classical convolutional codes, RSC codes offer better performance at low signal-to-noise ratio and/or high coding rates. Figure 3 depicts a significant experiment realized with RSC codes [15]. The MAP algorithm was used to decode RSC codes for different code rates and four increasing values of ν: 2, 4, 6 and 8, and the BER obtained by Monte Carlo simulation was plotted for very low signal-to-noise ratios E_b/N_0. What is interesting to observe is that, in each case, the four curves corresponding to the different values of ν seem to cross at one point (or at worst, in a small cloud of points, probably because the polynomials and the puncturing patterns were chosen arbitrarily). In comparison with the crossing point abscissas, the theoretical limits are indicated by arrows. The conformity between crossing points and theoretical limits suggests that increasing the code memory up to large values, say, several dozens, would offer optimum coding. This is quite natural because the theoretical limits were calculated using the random coding model, and RSC codes become more and more random as the register length increases. When ν tends to infinity, the period $L = 2^\nu - 1$ of the maximal-length pseudorandom generator, as well as the ratio between non RTZ and RTZ sequences, as given by (6), also tend to infinity.

We can conclude from this experiment that optimum coding/decoding seems to be achievable by adopting a long RSC code and decoding it using the MAP algorithm. Turbo coding is in fact an artifice for mimicking such a structure while keeping decoding complexity within reasonable boundaries.

In order to achieve coding rates higher than $\frac{1}{2}$, which is the natural rate of the encoder in Fig. 2, "puncturing" may be performed. As many symbols as necessary are discarded before transmission, so as to obtain rates between $\frac{1}{2}$ and 1. In the case of Turbo codes, only redundant symbols from the encoder are punctured. Another possibility for increasing the rate, without (or with less) puncturing, is

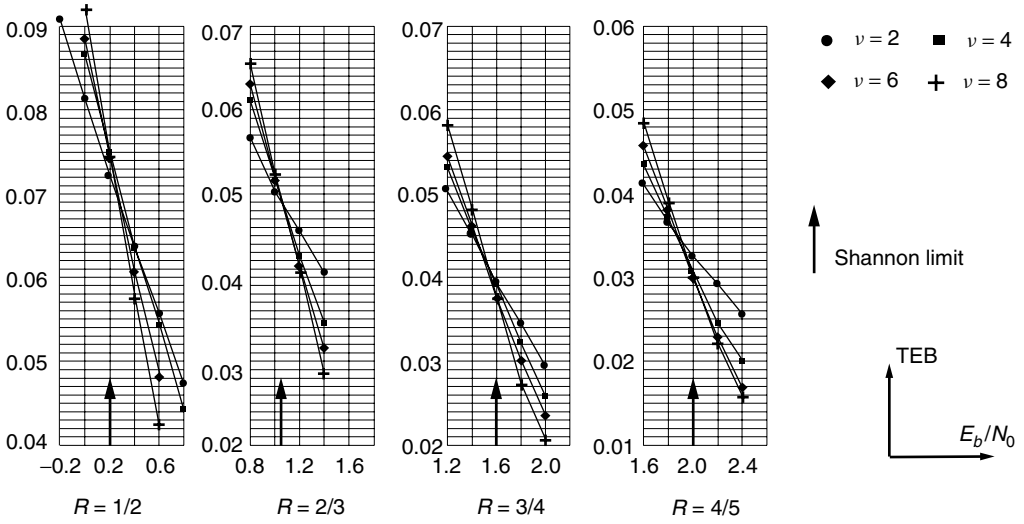


Figure 3. Performance of RSC codes for different code memories ($\nu = 2, 4, 6, 8$) and different rates, at very low signal-to-noise ratios, using the MAP decoding algorithm [13]. The Shannon limits, according to Dolinar et al. [9], are indicated by arrows.

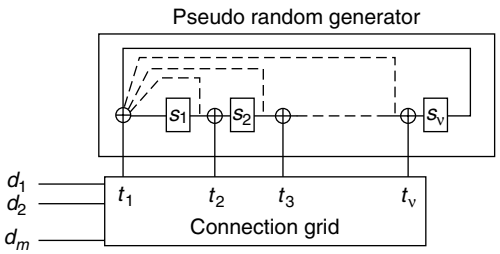


Figure 4. The general structure of an m-binary RSC encoder with code memory ν. The outputs of the encoder are not represented.

by using m-binary RSC codes, which are at the root of powerful high-rate Turbo codes.

Figure 4 depicts the general structure of an m-binary RSC encoder. It uses a pseudorandom generator with code memory ν and generator matrix $\mathbf{G}$ (size $\nu.\nu$), such that

$$\mathbf{S}_{i+1} = \mathbf{G}\mathbf{S}_i + \mathbf{T}_i \tag{10}$$

where $\mathbf{S}_i = (s_{1,i} \cdots s_{\nu,i})$ is the encoder state vector and $\mathbf{T}_i = (t_{1,i} \cdots t_{\nu,i})$ is the input vector to the encoder taps, both considered at time i. The m-component input vector $\mathbf{d}_i = (d_{1,i} \cdots d_{m,i})$ is connected to the ν possible taps via a connection grid whose binary matrix, of size $\nu.m$, is denoted $\mathbf{C}$. The ν-tap vector $\mathbf{T}_i$ is then given by

$$\mathbf{T}_i = \mathbf{C}\mathbf{d}_i \tag{11}$$

In order to avoid parallel transitions in the corresponding trellis, condition $m \leq \nu$ has to be respected. Except in very particular cases, this encoder is not equivalent to a binary encoder fed successively by $d_1, d_2, \ldots, d_m$; that is, the m-binary encoder is not generally decomposable.

The redundant output of the machine (not represented in the figure) is calculated, at time i, as

$$y_i = \sum_{j=1\cdots m} d_{j,i} + \mathbf{R}^T\mathbf{S}_i \tag{12}$$

where $\mathbf{R}^T$ is the transposed redundancy vector. The pth component of $\mathbf{R}$ is "1" if the pth register tap ($1 \leq p \leq \nu$) is used in the construction of y_i, "0" otherwise. It can easily be shown that y_i can also be written as

$$y_i = \sum_{j=1\cdots m} d_{j,i} + \mathbf{R}^T\mathbf{G}^{-1}\mathbf{S}_{i+1} \tag{13}$$

provided that

$$\mathbf{R}^T\mathbf{G}^{-1}\mathbf{C} \equiv \mathbf{0} \tag{14}$$

The set of relations (10)–(14) defines completely an m-binary RSC code.

On one hand, Eq. (12) ensures that the Hamming weight of $(d_{1,i}, d_{2,i}, \ldots, d_{m,i}, y_i)$ is at least 2, when leaving the reference path (null path), because changing one d value also changes the y value. On the other hand, (13) indicates that the Hamming weight of $(d_{1,i}, d_{2,i}, \ldots, d_{m,i}, y_i)$ is also at least 2 when merging the reference path. Hence, relations (12) and (13) together guarantee that the minimum free distance of the code, whose rate is $R = m/(m+1)$, is at least 4, whatever m.

As the minimum distance of a concatenated code is much larger than that of its component codes, we can imagine that very large minimum distances may be obtained for Turbo codes, for low as well as high rates. Of course, choosing large values of m implies high-complexity decoding because ν also has to be large. For this reason, only values of m up to 4 are, for the time being, considered in practice.

4. TERMINATION OF RSC CODES

The optimal decoding of a particular data bit in a convolutionally coded stream requires the knowledge of symbols preceding it and subsequent to it. Therefore, using a convolutional code for encoding a block poses a discontinuity problem at its extremities. With multidimensional codes such as Turbo codes, the question arises each time that the block is encoded. There are three possible solutions, referred to as *trellis termination*, to overcome this problem, which holds for RSC codes as well as for nonrecursive codes:

1. Initialize the encoder in state 0 and do nothing about the final state. Data located at the block end are then less protected than the other data. Depending on the target error rate, this penalty may be acceptable. Note, however, that the frame error rate (FER) is more affected than the BER.
2. Fix both starting and ending states to 0. The encoder is initialized in state 0 and, after the encoding of the block, the register is forced to state 0 by using ν additional bits, called "tail bits." These tail bits, and the redundant symbols associated with them, are transmitted in order to allow the decoder to choose the most likely path merging to state 0. The actual rate of the code is decreased because of the additional symbols, but the loss is quite negligible for medium and long blocks. This method is not absolutely suitable for multidimensional codes, because the tail bits are encoded only once and the composite code may suffer from this imperfection, which, however, is palpable only at low error rates.
3. Use circular (tail-biting) termination. Let us consider an RSC encoder, for instance, the one depicted in Fig. 5 (duobinary code with memory $\nu = 3$). At time $i+1$, register state $\mathbf{S}_{i+1}$ is a function of previous state $\mathbf{S}_i$ and tap vector $\mathbf{T}_i$, as given by (10). For the encoder of Fig. 5, vectors $\mathbf{S}_i$ and $\mathbf{T}_i$, and matrix $\mathbf{G}$ are given by

$$\mathbf{S}_i = \begin{bmatrix} s_{1,i} \\ s_{2,i} \\ s_{3,i} \end{bmatrix}; \qquad \mathbf{T}_i = \begin{bmatrix} d_{1,i} + d_{2,i} \\ d_{2,i} \\ d_{2,i} \end{bmatrix}; \qquad \mathbf{G} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

From Eq. (10) we can infer

$$\mathbf{S}_i = \mathbf{G}\mathbf{S}_{i-1} + \mathbf{T}_{i-1}$$
$$\mathbf{S}_{i-1} = \mathbf{G}\mathbf{S}_{i-2} + \mathbf{T}_{i-2}$$

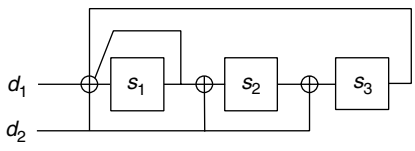


Figure 5. Recursive convolutional (duobinary) encoder with memory $\nu = 3$. The redundancy output, which is not relevant to the operation of the shift register, has been omitted.

$$\vdots$$
$$\mathbf{S}_1 = \mathbf{G}\mathbf{S}_0 + \mathbf{T}_0$$

Hence, $\mathbf{S}_i$ may be expressed as a function of initial state $\mathbf{S}_0$ and of data feeding the encoder between times 0 and i:

$$\mathbf{S}_i = \mathbf{G}^i \mathbf{S}_0 + \sum_{p=1}^{i} \mathbf{G}^{i-p} \mathbf{T}_{p-1} \tag{15}$$

If k is the input sequence length (the number of couples for the encoder of Fig. 5), it is possible to find a state $\mathbf{S}_c$, such that $\mathbf{S}_c = \mathbf{S}_k = \mathbf{S}_0$. Its value is derived from (15)

$$\mathbf{S}_c = \left(\mathbf{I} + \mathbf{G}^k\right)^{-1} \sum_{p=1}^{k} \mathbf{G}^{k-p} \mathbf{T}_{p-1} \tag{16}$$

where $\mathbf{I}$ is the $\nu.\nu$ identity matrix. State $\mathbf{S}_c$ depends on the sequence of data and exists only if $\mathbf{I} + \mathbf{G}^k$ is invertible. In particular, k cannot be a multiple of the period L of the recursive generator, which is such that $\mathbf{G}^L = \mathbf{I}$.

The term $\mathbf{S}_c$ is called the *circulation state*. Thus, if the encoder starts from state $\mathbf{S}_c$, it comes back to the same state when the encoding of the k data (k couples for the encoder of Fig. 5) is completed. Such an encoding process is called *circular* because the associated trellis may be viewed as a circle, without any discontinuity on transitions between states.

Determining $\mathbf{S}_c$ requires a preencoding operation. First, the encoder is initialized in state 0. Then, the data sequence of length k is encoded once, leading to final state $\mathbf{S}_k^0$ (no redundancy is produced during this operation). Then, from Eq. (15), we obtain

$$\mathbf{S}_k^0 = \sum_{p=1}^{k} \mathbf{G}^{k-p} \mathbf{T}_{p-1}$$

Combining this result with (16) gives the value of $\mathbf{S}_c$ as follows:

$$\mathbf{S}_c = \left(\mathbf{I} + \mathbf{G}^k\right)^{-1} \mathbf{S}_k^0 \tag{17}$$

In the second step, data are definitely encoded starting from state $\mathbf{S}_c$.

In practice, the relation between $\mathbf{S}_c$ and $\mathbf{S}_k^0$ is provided by a small combinational operator with ν input and output bits. The disadvantage of this method lies in having to encode the sequence twice: once from state 0 and the second time from state $\mathbf{S}_c$. Nevertheless, in most cases, the double-encoding operation can be performed at a frequency much higher than the data rate, so as to reduce the latency effects. Because $\mathbf{S}_c$ is not a priori known by the decoder, the latter has to estimate it by a preliminary step of processing information available preceding $\mathbf{S}_c$ (Fig. 6). So this operation, which we call *prologue*, falls on some data located at the end of the encoded block, and the prologue starts by assigning equal probabilities (or metrics) to all trellis states. The estimate of $\mathbf{S}_c$ is reliable as long as at

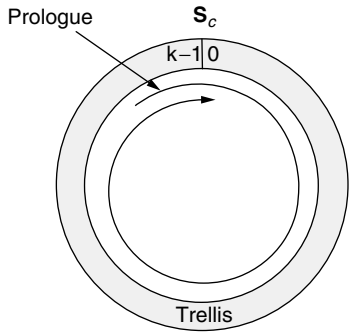


Figure 6. As a preamble to normal decoding, a prologue is carried out by the decoder in order to estimate the circulation state $\mathbf{S}_c$ of the circular (tail biting) trellis.

least a dozen or so redundant symbols, are exploited in the prologue.

The circular trellis termination of convolutional codes is a very powerful technique, enabling block encoding for any size and any rate, without the slightest loss in performance. Because a circle has no discontinuity, circular recursive systematic convolutional (CRSC) codes are not weakened by any side effect and, for this reason, are well suited to multidimensional coding.

5. TURBO CODES

A simple means to construct a quasirandom decodable code is a multiple parallel concatenation of CRSC codes, as depicted in Fig. 7. The block of k bits is encoded N times by N CRSC encoders, in a different order each time. Permutations Π_j are drawn at random, except the first one, which is the permutation identity (no permutation). Each component encoder delivers k/N (where k is a multiple of N) redundant symbols, and the global rate is $\frac{1}{2}$. We have already observed (Section 3) that the proportion of RTZ sequences for a simple RSC code, with code memory ν, is $p_1 = 2^{-\nu}$. The proportion of RTZ sequences for the N-dimensional code is lowered to

$$p_N = 2^{-N\nu} \tag{18}$$

because the sequence must remain RTZ after N different permutations. The other sequences, with proportion $1 - p_N$, yield codewords with a minimum distance at least equal to

$$d_{\min} = \frac{k}{2N} \tag{19}$$

This value assumes that only one sequence is not RTZ (the worst case) and that Y redundancy on the corresponding circle takes value "1" every other time statistically. For instance, with $N = 8$ and $\nu = 3$, we have $p_8 \approx 10^{-7}$, and for sequences of length $k = 1024$, we obtain $d_{\min} = 64$, which is quite a comfortable minimum distance (see Section 2).

Fortunately, from the complexity standpoint, it is not necessary to adopt such a large dimension. In fact, by replacing random permutation Π_2 with a carefully designed permutation, very good performance can be obtained while limiting the composite code to dimension 2. This is the principle of Turbo coding.

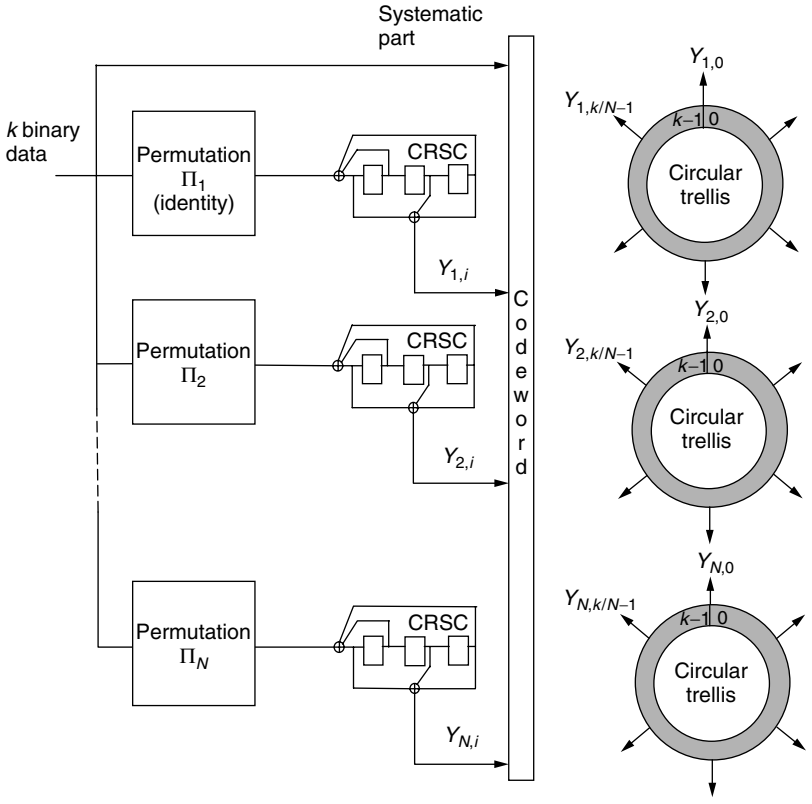


Figure 7. Multiple parallel concatenation of circular recursive systematic convolutional (CRSC) codes. Each encoder delivers k/N redundant symbols, uniformly distributed. Global rate: $\frac{1}{2}$.

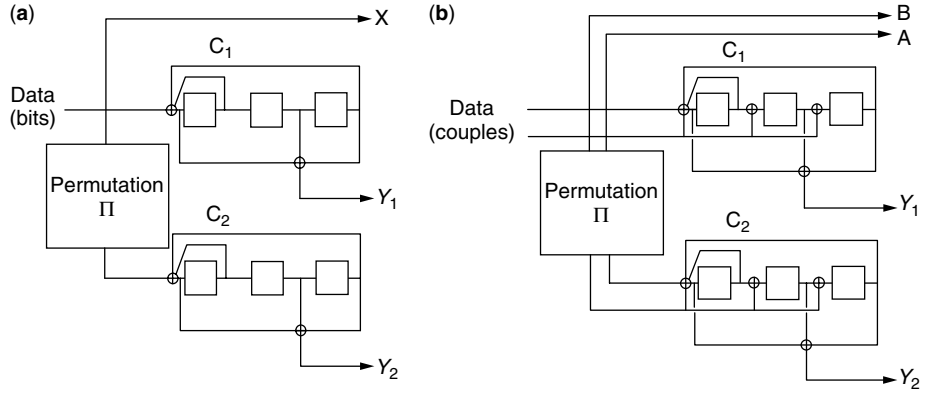


Figure 8. Binary and duobinary 8-state Turbo codes with memory $\nu = 3$, using the same RSC encoders (polynomials 13, 15). Natural rates, without puncturing, are $\frac{1}{3}$ and $\frac{1}{2}$, respectively.

Figure 8 represents two Turbo codes, in the classical binary and the duobinary versions. The original message (length k bits or couples) is encoded twice, in the natural and the permuted orders, by two RSC codes, denoted C_1 and C_2. In both examples, the component encoders are identical (polynomials 13 for recursivity and 15 for parity redundancy), but this is not a necessity. Natural rates, without puncturing, are $\frac{1}{3}$ and $\frac{1}{2}$. The binary Turbo code is suitable for low code rates ($R \leq \frac{1}{2}$), while the duobinary Turbo code is appropriate for high rates ($R \geq \frac{1}{2}$).

As the permutation function falls on finite-length sequences, the Turbo code is a block code by construction. To distinguish them from concatenated algebraic codes, like product codes, which are decoded by the Turbo algorithm and which were later called *block Turbo codes*, these coding schemes are known as *convolutional Turbo codes* or more technically, as *parallel concatenated convolutional codes* (PCCCs).

The arguments in favor of these coding schemes are as follows:

1. The decoding of a convolutionally encoded sequence is very sensitive to errors arriving in packets. Encoding the sequence twice, in different orders, before and after permutation, makes less likely the simultaneous appearance of clustered errors at the decoder inputs of C_1 and C_2. If packets of errors come to the decoder input of C_1, the permutation scatters them and they become isolated errors for the decoder of C_2, and vice versa. Thus, the bidimensional encoding, formed by either a parallel or a serial concatenation, markedly reduces the vulnerability of convolutional encoding toward packets of errors. But which decoder should one rely on to take the final decision? No criterion allows us to trust either one or the other. The answer is supplied by the Turbo algorithm, which spares us from having to make a choice. This algorithm works out exchanges of probabilistic information between both decoders and forces them to converge toward the same decision, as these exchanges take place.
2. Parallel concatenation combines two codes with rates R_1 (code C_1, with possible puncturing) and R_2 (code C_2, also with possible puncturing), and the global rate is

$$R_p = \frac{R_1 R_2}{1 - (1 - R_1)(1 - R_2)} \tag{20}$$

This rate is higher than that of a serially concatenated code ($R_s = R_1 R_2$), for the same values of R_1 and R_2, and the lower these rates, the larger the difference. Thus, with the same performance of component codes, parallel concatenation offers a better global rate, but this advantage is lost when the rates come close to unity.

3. Parallel concatenation relies on systematic codes. At least one of these codes has to be recursive, for a fundamental reason related to the minimum data input weight ($w_{\min}$). Figure 9 depicts two nonrecursive systematic convolutional codes, concatenated in parallel. The input sequence is "all zero," except in one position. This single "1" disrupts the encoder output during a short lapse of time, given by the constraint length. Actually, this sequence with only one "1" is the minimal RTZ sequence of any nonrecursive code. Thus, redundancy Y_1 is very poor with respect to this particular sequence. After the permutation, the sequence remains "all zero" except in one position, and redundancy Y_2 is as poor as the first one. In fact, the minimum distance of this

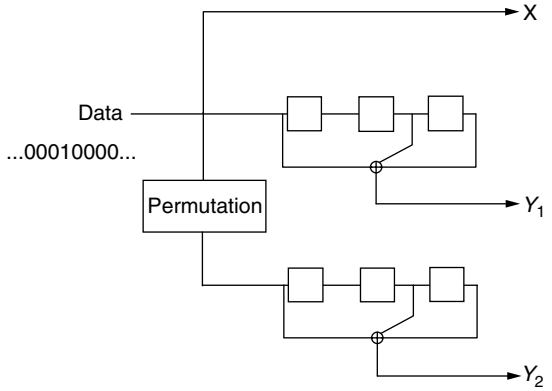


Figure 9. The parallel concatenation of nonrecursive systematic convolutional codes makes up a poor code with regard to weight 1 sequences.

composite code is not higher than that obtained from a single code, with the same rate. If we replace at least one of the nonrecursive encoders by a recursive one, the considered input sequence is no longer an RTZ sequence for this encoder, and the redundancy weight is considerably increased.

4. As seen at the beginning of this section, it is possible to increase the Turbo code dimension N by using more than two component encoders. The result is a noticeable increase in the minimum distance; with $N \geq 5$ and a set of random permutations, the Turbo code is comparable to a quasirandom code. Unfortunately, the convergence threshold of the decoder (the signal-to-noise ratio above which most errors are corrected) deteriorates when the dimension is increased. This is due to the Turbo decoding principle, which is to consider iteratively each dimension, one after the other. As the redundancy rate of each component code decreases when their number increases, the first steps of the decoding are penalized in comparison with the two-dimensional code. This conflict between large minimum distance and low convergence threshold, already mentioned in Section 2, is a permanent feature of error correction coding.

A particular Turbo code is defined by

- m, the number of bits in the input words. Applications known so far consider binary ($m = 1$) and duobinary ($m = 2$) words. More recent decoding techniques using the dual-code approach [16] could allow larger values of m to be considered without increasing the decoding complexity too much.
- The component codes C_1 and C_2 (code memory ν, recursivity and redundancy polynomials). The values of ν are 3 or 4 in practice and the polynomials are generally those that are recognized as the best for simple unidimensional convolutional coding, that is, (15,13) for $\nu = 3$ and (23,35) for $\nu = 4$, or their symmetric forms.
- The permutation function, which plays a decisive role when the target BER is lower than about 10^{-5}. Above this value, the permutation may be any, provided obviously that it respects at least the scattering property (e.g., the permutation may be the regular one).
- The puncturing pattern. This has to be as regular as possible, such as that for simple convolutional codes. In addition to this rule, the puncturing pattern is defined in close relationship with the permutation function when very low errors rates are sought. Puncturing is achieved on the systematic part of the codewords. In some cases, it may be conceivable to puncture the redundant part instead, with the aim of increasing the minimum distance. This is then achieved to the detriment of the convergence threshold, because, from this standpoint, deleting data shared by all the decoders is more penalizing than deleting data that are useful to only one of them.

6. THE PERMUTATION FUNCTION

Either called *permutation* or *interleaving*, the technique involving scattering data over time is of great service in digital communications. It is used, for instance, to reduce the effects of dimming in fading channels, and more generally to combat perturbations that affect consecutive symbols. In the case of Turbo codes, permutation also plays this role for at least one dimension of the composite code. But its importance goes beyond this; permutation also fixes the minimum distance of the concatenated code, in close relationship to the properties of component codes.

Let us consider the binary Turbo code represented in Fig. 8a, with permutation falling on k bits. The worst permutation we can imagine is permutation identity, which minimizes the coding diversity ($Y_1 = Y_2$). On the other hand, the best permutation that could be used, but that probably does not exist [17], could allow the concatenated code to be equivalent to a sequential machine whose irreducible number of states would be 2^{k+6}. There are actually $k + 6$ binary storage elements in the structure: k in the permutation memory and 6 in the encoders. Assimilating this machine to a convolutional code would give a very long code and very large minimum distances, for usual values of k. From the worst to the best of permutations, there is great choice between the $k!$ possible combinations, and we still lack a sound theory about it. Nevertheless, good permutations have already been designed to elaborate normalized Turbo codes, using pragmatic approaches.

6.1. Regular Permutation

The starting point in the design of a permutation is the regular permutation, illustrated in two ways in Fig. 10, for binary codes. The first one assumes that the block of k bits can be organized as a table with M rows and N columns ($k = M.N$). The permutation then involves writing data

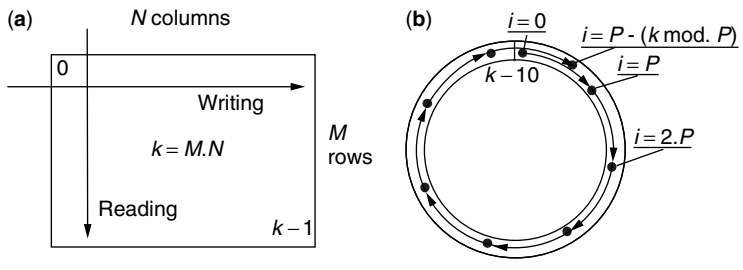


Figure 10. Regular permutation with rectangular (**a**) or circular (**b**) forms.

linewise in an appropriate memory and reading them columnwise, possibly with skips of columns [18]. The second one is used without any hypothesis on the value of k. After writing the data in a linear memory, with address i ($0 \le i \le k-1$), the block is likened to a circle, and both extremities of the block ($i = 0$ and $i = k-1$) then become contiguous. The data are read out such that the jth datum read was written at position i given by

$$i = Pj \mod k \tag{21}$$

where P is an integer, prime with k. In order to maximize the spatial distance after permutation between two consecutive data, whatever they are, and vice versa, P has to be close to $\sqrt{2k}$, with the condition

$$k \approx \frac{P}{2} \mod P \tag{22}$$

6.2. The Statistical Approach

An upper bound of the error probability P_e of a code, assuming optimal decoding, is given by

$$P_e \le \sum_{d=d_{\min}}^{\infty} M_d \operatorname{erfc}\left(\sqrt{Rd\frac{E_b}{N_0}}\right) \tag{23}$$

where $d_{\min}$ is the minimum distance, R the rate considered, and M_d the sum of the weights, called *multiplicity*, of all codewords at distance d. When designing a permutation, the aim is thus to maximize $d_{\min}$ and minimize the multiplicities. Before doing this work, it is interesting to have an idea about the performance that any typical permutation might give. Benedetto and Montorsi [19] proposed to use uniform (or statistical) model of permutation, which is a device that associates with a message of length k and weight w, one of the possible $\binom{k}{w}$ messages obtained by the permutation of w bits among k. The $\binom{k}{w}$ permuted messages have the same probability $\dfrac{1}{\binom{k}{w}}$ of being present at the second encoder input.

This statistical interleaver performs similarly to an interleaver that would be the average of all possible deterministic interleavers with size k. Therefore, there exists at least one interleaver with fixed rules that allows one to reach, and even exceed, the performance given by the uniform interleaver. In fact, it is easy to find deterministic permutations better than the statistical one when the target error rate is low.

Let us assume that A_{wd} denotes the number of codewords at distance d and with weight w, without the permutation. It is then demonstrated [19] that the uniform permutation modifies this value into $w!k^{1-w}A_{wd}$, so the value is reduced if $w!k^{1-w}$ is less than 1. The worst case is given by the minimum value of w, that is, $w_{\min}$, which is 1 for algebraic codes (BCH, etc.) and nonrecursive convolutional codes, but 2 for RSC codes. The term $k^{1-w_{\min}}$, called *interleaving gain*, is then favorable only in the latter case and is equal to $k^{1-2} = 1/k$. It also follows from this result that the longer the encoded block, the lower the multiplicities and the better the performance. This point is in agreement with the theoretical limits obtained on finite-length blocks [9].

In conclusion, the statistical approach of the permutation problem confirms the need to use RSC codes and gives a means of observing and estimating an interleaving gain that depends on the block size. This gain is essentially visible at large error rates (BER $\ge 10^{-5}$). For lower error rates, the main parameter is the minimum distance, which has to be maximized, and the uniform interleaver is not suitable for this.

6.3. Real Permutations

The dilemma in the design of a good permutation lies in the need to obtain a sufficient minimum distance for two distinct classes of codewords, which require conflicting treatment [20]. The first class contains all codewords with input weight $w \le 3$, and a good permutation for this class is as regular as possible. The second class encompasses all codewords with input weight $w > 3$, and nonuniformity (controlled disorder) has to be introduced in the permutation function to obtain a large minimum distance. Figure 11 illustrates the situation, showing the example of a rate-$\frac{1}{3}$ Turbo code, using component binary encoders with code memory $\nu = 3$ and periodicity $L = 2^{\nu} - 1 = 7$.

For the sake of simplicity, the block of k bits is organized as a rectangle with M rows and N columns ($M \approx N \approx \sqrt{k}$). Regular permutation is used; that is, data are written linewise and read columnwise:

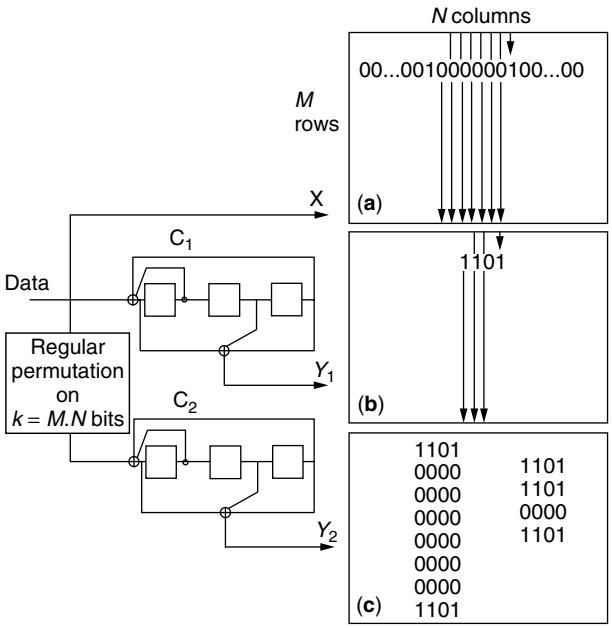


Figure 11. Some possible RTZ (return-to-zero) sequences for both encoders C_1 and C_2, with $G_X(D) = 1 + D + D^3$ (period $L = 7$): (**a**) with input weight $w = 2$; (**b**) with $w = 3$; (**c**) with $w = 6$ or 9.

Case (a) in Fig. 11 depicts a situation where encoder C_1 (the horizontal one) is fed by an RTZ sequence with input weight $w = 2$. Redundancy Y_1 delivered by this encoder is poor, but redundancy Y_2 produced by encoder C_2 (the vertical one) is very informative for this pattern, which is also an RTZ sequence, but whose span is $7.M$ instead of 7. The associated minimum distance would be around $7.M/2$, which is a large minimum distance for typical values of k. With respect to this $w = 2$ case, the code is said to be "good" because $d_{\min}$ tends to infinity when k tends to infinity.

Case (b) in Fig. 11 deals with a weight 3 RTZ sequence. Again, whereas the contribution of redundancy Y_1 is not high for this pattern, redundancy Y_2 gives relevant information over a large span, of length $3.M$. The conclusions are the same as for case (a).

Case (c) in Fig. 11 shows two examples of sequences with weights $w = 6$ and $w = 9$, which are RTZ sequences for both encoders C_1 and C_2. They are obtained by a combination of two or three minimal-length RTZ sequences. The set of redundancies is limited and depends on neither M nor N. These patterns are typical of codewords that limit the minimum distance of a Turbo code, when using a regular permutation.

In order to "break" rectangular patterns, some disorder has to be introduced into the permutation rule, while ensuring that the good properties of regular permutation, with respect to weights 2 and 3, are not lost. This is the crucial problem in the search for good permutation, which has not yet found a definitive answer. Nevertheless, some good permutations have already been devised for several applications (CCSDS [21], IMT-2000 [22,23], DVB [24,25]).

7. TURBO DECODING

Decoding a composite code by a global approach is not possible in practice because of the tremendous number of states to consider. A joint probabilistic process by the decoders of C_1 and C_2 has to be elaborated. Because of latency constraints, this joint process is worked out in an iterative manner in a digital circuit (analog versions of the Turbo decoder are also considered, offering much larger throughputs [26]).

Turbo decoding relies on the following fundamental criterion, which is applicable to all "message passing" or "belief propagation" [27] algorithms:

> When having several probabilistic machines work together on the estimation of a common set of symbols, all the machines have to give the same decision, with the same probability, about each symbol, as a single (global) decoder would.

To make the composite decoder satisfy this criterion, the structure of Fig. 12 is adopted. The double loop enables both component decoders to benefit from the whole redundancy. The term "Turbo" was given to this feedback construction with reference to the principle of the turbo-charged engine.

The components are soft-in/soft-out (SISO) decoders, and permutation (Π) and inverse permutation (Π^{-1}) memories. The node variables of the decoder are logarithms of likelihood ratios (LLRs). An LLR related to a particular binary datum d_i is defined as

$$\mathrm{LLR}(d_i) = \ln\left(\frac{\Pr(d_i = 1)}{\Pr(d_i = 0)}\right) \tag{24a}$$

For a decoder processing m-binary words instead of binary data, the LLRs associated with the 2^m possible values of the word vector $\mathbf{d}_i$, could be written as

$$\mathrm{LLR}_j(\mathbf{d}_i) = \ln\left(\frac{\Pr(\mathbf{d}_i \equiv j)}{\Pr(\mathbf{d}_i \equiv 0)}\right) \tag{24b}$$

where $\Pr(\mathbf{d}_i \equiv j)$ is the probability that vector $\mathbf{d}_i$ takes the jth value, numbered from 0 to $2^m - 1$. Because LLR_0 is always equal to 0, there are only $2^m - 1$ LLRs to calculate in practice.

The role of a SISO decoder is to process an input LLR and, thanks to local redundancy (i.e., y_1 for DEC1, y_2 for

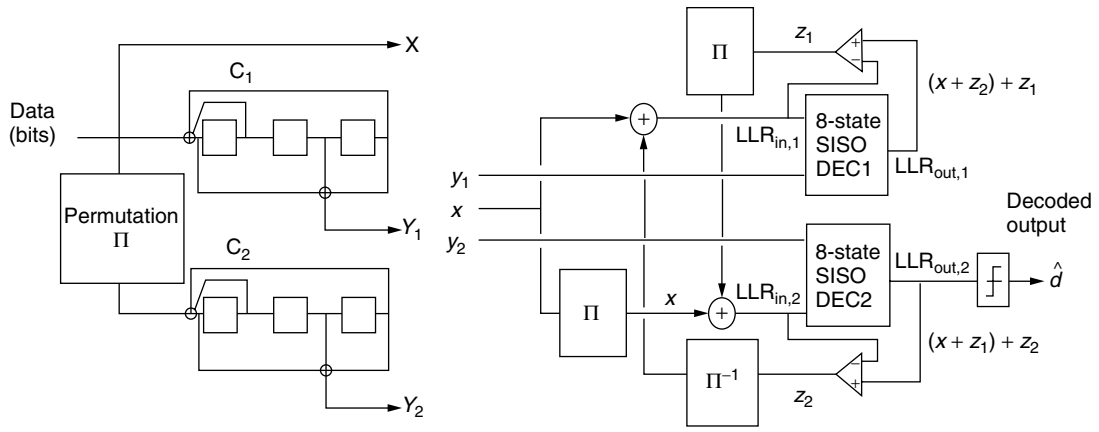


Figure 12. An 8-state Turbo code and its associated decoder (basic structure assuming no delay processing).

DEC2), to try to improve it. The output LLR of a SISO decoder, for a binary datum, may be simply written as

$$\mathrm{LLR}_{\mathrm{out}}(d) = \mathrm{LLR}_{\mathrm{in}}(d) + z(d) \tag{25}$$

where $z(d)$ is the *extrinsic* information about d, provided by the decoder. If this works properly, $z(d)$ is usually negative if $d = 0$, and positive if $d = 1$.

The composite decoder is constructed in such a way that only extrinsic terms are passed by one component decoder to the other. The input LLR to a particular decoder is formed by the sum of two terms: the information symbols (x) stemming from the channel and the extrinsic term (z) provided by the other decoder, which serves as a priori information. The information symbols are common inputs to both decoders, which is why the extrinsic information must not contain them. In addition, the outgoing extrinsic information does not include the incoming extrinsic information, in order to reduce correlation effects in the loop.

The practical course of operation is

Step 1. Process the data peculiar to one code, say, C_2 (x and y_2) by decoder DEC2, and store the extrinsic pieces of information (z_2) resulting from the decoding in a memory. If data are missing because of puncturing, the corresponding values are set to analog 0 (neutral value).

Step 2. Process the data specific to C_1 (x, deinterleaved z_2 and y_1) by decoder DEC1, and store the extrinsic pieces of information (z_1) in a memory. By properly organizing the read/write instructions, the same memory can be used for storing both z_1 and z_2.

Steps 1 and 2 make up the first iteration.

Step 3. Process C_2 again, now taking interleaved z_1 into account, and store the updated values of z_2.

And so on.

The process ends after a preestablished number of iterations, or after the decoded block has been estimated as correct, according to some stop criterion (see the report by Matache et al. [28] for possible methods of stopping rules). The typical number of iterations for the decoding of convolutional Turbo codes is 4–10, depending on the constraints relating to complexity, power consumption, and latency.

According to the structure of the decoder, after p iterations, the output of DEC1 is

$$\mathrm{LLR}_{\mathrm{out1},p}(d) = (x + z_{2,p-1}(d)) + z_{1,p}(d)$$

where $z_{u,p}(d)$ is the extrinsic piece of information about d, yielded by decoder u after iteration p, and the output of DEC2 is

$$\mathrm{LLR}_{\mathrm{out2},p}(d) = (x + z_{1,p-1}(d)) + z_{2,p}(d)$$

If the iterative process converges toward fixed points, $z_{1,p}(d) - z_{1,p-1}(d)$ and $z_{2,p}(d) - z_{2,p-1}(d)$ both tend to zero when p goes to infinity. Therefore, from the equations above, both LLRs become equal, which fulfills the fundamental condition of equal probabilities provided by the component decoders for each datum d. As for the proof of convergence itself, one can refer to various papers dealing with the theoretical aspects of the subject [e.g., 29,30].

Turbo decoding is not optimal. This is because an iterative process obviously must begin, during the first half-iteration, with only a part of the redundant information available (either y_1 or y_2). Fortunately, loss due to suboptimality is small: about 0.5 dB for binary Turbo codes and 0.3 dB for duobinary Turbo codes.

There are two families of SISO algorithms: those based on the Viterbi algorithm [12], which can be used for high throughput continuous stream applications; and others based on the MAP (also known as BCJR or APP) algorithm [13] or its simplified derived versions [14], for block decoding. If the full MAP algorithm is chosen, it is better for extrinsic information to be expressed by probabilities instead of LLRs, which avoids the need to calculate a useless variance for extrinsic terms.

The following practical parameters must be factored into the design of a Turbo decoder that processes LLRs:

- The number of quantization bits for the channel samples: typically 3 or 4 if the code is associated with BPSK or QPSK modulation; 5 or 6 with higher-order modulations that require greater accuracy.
- The number of quantization bits for extrinsic information: typically 1 bit more than those of channel samples.
- The scale factor, which is the ratio between the mean absolute value of the channel data and its maximum absolute value. This factor depends on the coding rate, on the type of channel and also on quantization accuracy.

In practice, depending on the kind of SISO algorithm chosen, some tuning operations (multiplying, limiting) on extrinsic information are added to the basic structure to ensure stability and convergence within a small number of iterations.

8. APPLICATIONS

Table 1 summarizes normalized applications of convolutional Turbo codes known to date. These use either 8-state binary or duobinary RSC component encoders or 16-state binary encoders. Figure 13 shows some representative examples of performance obtained from various Turbo codes associated with QPSK and 8-PSK modulation on Gaussian channels. For the latter case, the so-called pragmatic scheme [31], which is the simplest way to combine channel coding and modulation, was used. The component decoding algorithm was the max–log–MAP (also called subMAP) algorithm [14], which is derived from the exact MAP procedure [13] by doing operations in the logarithmic domain (multiplications become additions), and by replacing additions by maximum (Max) functions.

To simplify, let us say that 8-state Turbo codes are suitable for the medium error rates (FER $\approx 10^{-4}$) that are

Table 1. Applications of Convolutional Turbo Codes

Application	Turbo Code	Termination	Polynomials	Rates
CCSDS [21]	Binary, 16-state	Tail bits	23, 33, 25, 37	$\frac{1}{6}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$
IMT-2000 [22,23]	Binary, 8-state	Tail bits	15, 13, 17	$\frac{1}{4}, \frac{1}{3}, \frac{1}{2}$
DVB-RCS [24]	Duobinary, 8-state	Circular	15, 13	$\frac{1}{3} - \frac{6}{7}$
DVB-RCT [25]	Duobinary, 8-state	Circular	15, 13	$\frac{1}{2}, \frac{3}{4}$
Inmarsat (mini M)	Binary, 16-state	No	23, 35	$\frac{1}{2}$
Eutelsat (skyplex)	Duobinary, 8-state	Circular	15, 13	$\frac{4}{5}, \frac{6}{7}$

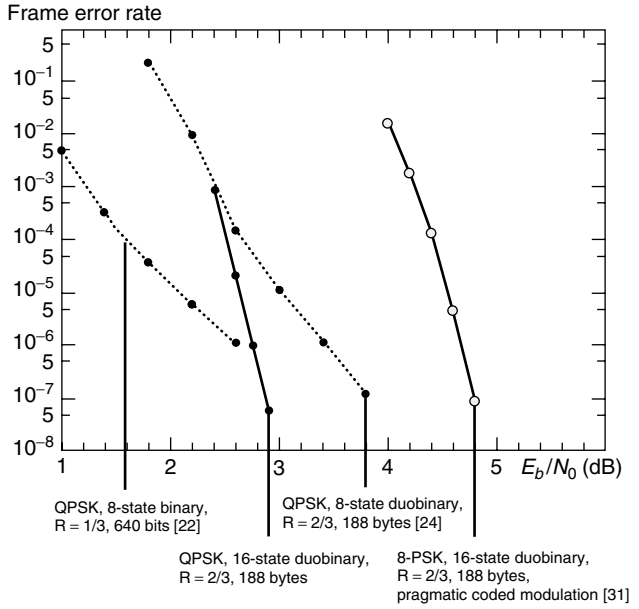


Figure 13. Some examples of performance, expressed in FER, achievable with Turbo codes on Gaussian channels. In all cases, decoding was performed using the max−log−MAP algorithm [14] with eight iterations and 4-bit input quantization.

required, for instance, by ARQ (Automatic Repeat reQuest) systems, whereas 16-state Turbo codes are necessary when the target error rates are lower (FER $\approx 10^{-8}$), for broadcasting in particular.

Acknowledgment
Warm thanks to Janet Ormrod for her contribution to the writing of this article.

BIOGRAPHIES

Claude Berrou was born in Penmarc'h, France, in 1951. He received the Electrical Engineering Degree from the Institut National Polytechique, Grenoble, France, in 1975. In 1978, he joined the Ecole Nationale Supérieure des Télécommunications de Bretagne, Brest, France, where he is currently Professor in the Electronics Department. His research topics include algorithm−silicon interaction, electronics and digital communications, error-correcting codes, Turbo codes that he discovered in 1991, soft-in/soft-out decoders, and genetic coding. He is the author or co-author of eight registered patents and about 40 publications in the field of Turbo coding. He was the co-recipient of the 1997 IEEE Trans. com. Paper Award and the 1998 IEEE Information Theory Society Paper Award. He also received the Médaille Ampère (SEE) and one of the Golden Jubilee Awards for Technological Innovation (IEEE IT Society) in 1998.

Alain Glavieux was born in Paris, France, in 1949. He received the Electrical Engineering Degree from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1978. In 1979, he joined the Ecole Nationale Supérieure des Télécommunications de Bretagne, Brest, France, where he is currently Professor and Director of Corporate Relations. His research interest includes Turbo coding, Turbo equalization, and communications over fading channels. He was the co-recipient of the 1997 IEEE Trans. Com. Paper Award and the 1998 IEEE Information Theory Society Paper Award. He also received one of the Golden Jubilee Awards for Technological Innovation (IEEE IT Society) in 1998.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: (July, Oct. 1948).
2. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: turbo-codes, *Proc. IEEE ICC '93*, Geneva, May 1993, pp. 1064−1070.
3. G. Battail, Coding for the Gaussian channel: The promise of weighted-output decoding, *Int. J. Satellite Commun.* **7**: 183−192 (1989).
4. G. Battail, Pondération des symboles décodés par l'algorithme de Viterbi, *Ann. Télécommun.* **42**(1−2): 31−38 (Jan. 1987).
5. J. Hagenauer and P. Hoeher, A Viterbi algorithm with soft-decision outputs and its applications, *Proc. GLOBECOM '89*, Dallas, TX, Nov. 1989, pp. 47.11−47.17.
6. J. Hagenauer and P. Hoeher, Concatenated Viterbi-decoding, *Proc. Int. Workshop on Information Theory*, Gotland, Sweden, Aug.−Sept. 1989.
7. R. G. Gallager, Low-density parity-check codes, *IRE Trans. Inform. Theory* **IT-8**: 21−28 (Jan. 1962).
8. R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **IT-27**: 533−547 (Sept. 1981).
9. S. Dolinar, D. Divsalar, and F. Pollara, *Code Performance as a Function of Block Size*, TMO Progress Report 42-133, JPL, NASA, May 1998.
10. J. G. Proakis, *Digital Communications*, 2nd ed., McGraw-Hill, New York, 1989, pp. 426−428.

11. R. Podemski, W. Holubowicz, C. Berrou, and G. Battail, Hamming distance spectra of turbo-codes, *Ann. Télécommun.* **50**(9–10): 790–797 (Sept.–Oct. 1995).

12. G. D. Forney, The Viterbi algorithm, *Proc. IEEE* **61**(3): 268–278 (March 1973).

13. L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 248–287 (March 1974).

14. P. Robertson, P. Hoeher, and E. Villebrun, Optimal and suboptimal maximum a posteriori algorithms suitable for turbo decoding, *Eur. Trans. Telecommun.* **8**: 119–125 (March–April 1997).

15. C. Berrou, Some clinical aspects of turbo codes, *Proc. 1st Int. Symp. Turbo Codes & Related Topics*, Brest, France, Sept. 1997, pp. 26–31.

16. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**(2): 429–445 (March 1996).

17. Y. V. Svirid, Weight distributions and bounds for turbo-codes, *Eur. Trans. Telecommun.* **6**(5): 543–555 (Sept.–Oct. 1995).

18. E. Dunscombe and F. C. Piper, Optimal interleaving scheme for convolutional coding, *Electron. Lett.* **25**(22): 1517–1518 (Oct. 1989).

19. S. Benedetto and G. Montorsi, Design of parallel concatenated convolutional codes, *IEEE Trans. Commun.* **44**(5): 591–600 (May 1996).

20. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* **44**(10): 1261–1271 (Oct. 1996).

21. Consultative Committee for Space Data Systems, *Recommendations for Space Data Systems. Telemetry Channel Coding*, Blue Book, May 1998.

22. 3GPP Technical Specification Group, *Multiplexing and Channel Coding* (*FDD*), TS 25.212 v2.0.0, June 1999.

23. TIA/EIA/IS 2000-2, *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, July 1999.

24. DVB, *Interaction Channel for Satellite Distribution Systems*, ETSI EN 301 790, V1.2.2, Dec. 2000, pp. 21–24.

25. DVB, *Interaction Channel for Digital Terrestrial Television*, ETSI EN 301 958, V1.1.1, Aug. 2001, pp. 28–30.

26. H.-A. Loeliger, F. Tarkoy, F. Lustenberger, and M. Helfenstein, Decoding in analog VLSI, *IEEE Commun. Mag.* **37**(4): 99–101 (April 1999).

27. R. J. McEliece and D. J. C. MacKay, Turbo decoding as an instance of Pearl's "belief propagation" algorithm, *IEEE J. Select. Areas Commun.* **16**(2): 140–152 (Feb. 1998).

28. A. Matache, S. Dolinar, and F. Pollara, *Stopping Rules for Turbo Decoders*, TMO Progress Report 42-142, JPL, NASA, Aug. 2000.

29. Y. Weiss and W. T. Freeman, On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs, *IEEE Trans. Inform. Theory* **47**(2): 736–744 (Feb. 2001).

30. L. Duan and B. Rimoldi, The iterative turbo decoding algorithm has fixed points, *IEEE Trans. Inform. Theory* **47**(7): 2993–2995 (Nov. 2001).

31. S. Le Goff, A. Glavieux, and C. Berrou, Turbo-codes and high spectral efficiency modulation, *Proc. IEEE ICC'94*, New Orleans, May 1994, pp. 645–649.

TURBO EQUALIZATION

KRISHNA R. NARAYANAN
Texas A&M University
College Station, Texas

1. INTRODUCTION

Intersymbol interference (ISI) can often be a limiting factor when communicating through band-limited channels and, hence, it is important to use efficient equalization techniques to combat the effect of ISI. A good discussion of equalization techniques for uncoded systems including both sequence estimation type equalizers and symbol-by-symbol equalizers can be found in another study [1]. When an error correction code (ECC) is used in conjunction with an ISI channel, the equalizer should take advantage of the error correction capability of the code. The optimal receiver which performs joint equalization and decoding can be computationally complex and is usually not implementable in practice. Therefore, several suboptimal solutions have been studied. One straightforward technique is to perform equalization without taking the code structure in to account, followed by decoding. Since the operating signal-to-noise ratio (SNR) for coded systems is usually much smaller than that for the uncoded case, the performance of the equalizer can be severely affected. Techniques that use the tentative decision from the decoder in the equalization step, such as in delayed decision feedback sequence estimation [2], have been shown to perform better than separate equalization and decoding.

In 1995, Douillard et al. proposed another sub-optimal joint equalization and decoding technique [3] by extending the idea of iterative decoding that was used to decode Turbo codes [4]. The idea was to exchange soft information between a soft-input soft-output (SISO) equalizer and an SISO decoder in an iterative fashion and they naturally named it *Turbo equalization*. Since then, several researchers have shown that turbo equalization can significantly improve the performance over separate equalization and decoding, and is a practical solution to obtaining close to capacity performance on ISI channels. Currently, Turbo equalization is being considered for use in future-generation digital magnetic recording systems and wireless systems. Here, we will explain the main concepts behind turbo equalization and summarize some of the results in this area.

2. SYSTEM MODEL

A typical system model with an error correction code and an ISI channel is shown in Fig. 1. A block of K bits of the binary data sequence $\boldsymbol{a}$ is first encoded by an ECC (referred to as *outer code* here) into N coded bits. Any code that permits efficient SISO decoding can be used, but we will restrict our attention to a convolutional code, parallel concatenated convolutional code (PCCC or turbo code), or low-density parity-check (LDPC) code as the outer code. The multiplexed output $\boldsymbol{x}$ is interleaved and the interleaved sequence $\boldsymbol{y}$ is modulated into $\boldsymbol{z}$. We will assume

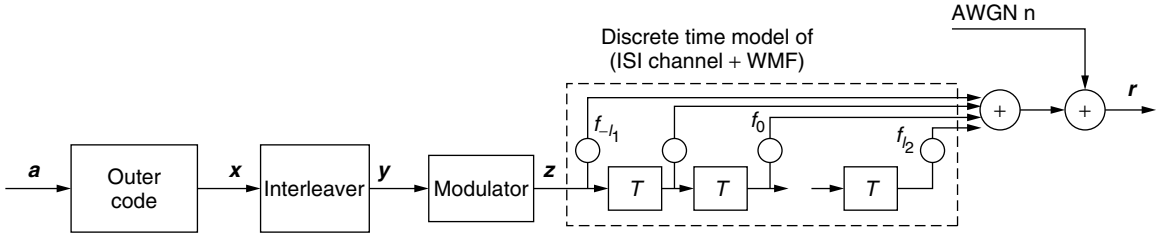


Figure 1. Discrete-time system model.

that the modulation is memoryless and the modulated sequence $\boldsymbol{z}$ is then transmitted over the ISI channel. At the receiver, the received signal is passed through a whitened matched filter (WMF) and the output of the WMF is sampled every T seconds, where T is the symbol duration. The combination of the ISI channel, WMF, and the sampler is equivalent to a discrete-time transversal filter (DTTF). The DTTF can be represented by an L tap-delay line with weights $f_{-l_1}, \ldots, f_{-1}, f_0, f_1, \ldots, f_{l_2}$, as shown in Fig. 1, where l_1 and l_2 are the number of postcursor and precursor taps, respectively, and $L = l_1 + l_2 + 1$. Therefore, the WMF output at time instant k can be expressed as

$$r_k = \sum_{i=-l_1}^{l_2} z_{k-i} \quad f_i + n_k, \tag{1}$$

where n_k are samples of a white Gaussian noise process with zero mean and variance σ_{ch}^2.

3. TURBO EQUALIZATION ALGORITHM

We will start with a few preliminaries and then describe the Turbo equalization algorithm. To keep the discussion simple and clear, let us assume that $\boldsymbol{a}$ is a binary sequence of equiprobable and independent bits, the outer code is a binary convolutional code, and that the modulation is binary phase shift keying (BPSK). Further, let us assume that the receiver has perfect knowledge of the tap coefficients and the variance of the additive noise. For any sequence $\boldsymbol{x}$, let $\boldsymbol{x}_k^N$ denote the sequence $\boldsymbol{x}$ from time k to N. The conditional log-likelihood ratio (LLR) of a binary random variable $x_k \in \{0, 1\}$ given a noisy observation of x, namely r, is defined as

$$\Lambda(x_k) \triangleq \log \frac{P(x_k = 1|r)}{P(x_k = 0|r)} \tag{2}$$

The optimal receiver that minimizes the bit error rate (BER) computes an estimate $\hat{a}_k$ according to

$$\hat{a}_k = \begin{matrix} 1, & \text{if} \quad P(a_k = 1|r_1^N) > P(a_k = 0|r_1^N) \\ 0, & \text{otherwise} \end{matrix} \tag{3}$$

Although conceptually simple, it is quite difficult (almost impossible) to implement the above receiver for most cases of practical interest. The Turbo equalization algorithm computes a suboptimal solution to the problem presented above as explained below. We will first give the steps of the Turbo equalization algorithm without explaining the details of the computation at each step. In the next sections, we will detail the computations to be performed at each step. The overall algorithm is an iterative algorithm that consists of an SISO equalizer and an SISO decoder that exchanges LLR estimates of the bits x_k in the sequence $\boldsymbol{x}$. At the mth stage, the equalizer provides *extrinsic* information in the form of a loglikelihood ratio on each bit x_k, denoted by $L_i^m(x_k)$. The extrinsic information generated by the equalizer $L_i^m(x_k)$ does not include any information on x_k given by the decoder. It represents the contribution of the received signal $\boldsymbol{r}$ and the *a priori* information on x_i given by the outer decoder for all $i \neq k$. In the following section, we will use L to denote extrinsic LLRs and Λ to denote the total LLR. Subscripts i and o denote quantities generated from the inner SISO (equalizer) and outer SISO, respectively. Superscript m refers to the iteration number; and when there is no danger of confusion, we will drop the superscript. The outer decoder uses the extrinsic information $L_i^m(x_k)$ as though they were the output of a hypothetical channel and provides extrinsic information on x_k, denoted by $L_o^m(x_k)$. This extrinsic information $L_o^m(x_k)$ is based purely on the code constraints imposed by the outer code, and is used as a priori information in the $(m+1)$th stage in the equalizer (inner decoder). The outer decoder also produces likelihood ratios for the information bits $\Lambda^m(a_k)$ from which hard-decision estimates $\hat{a}_k$ can be obtained. The iterations proceed until a stopping criterion (e.g., a cyclic redundancy check) is satisfied or a maximum of M iterations have been performed. Note that the extrinsic information generated by the outer decoder should be interleaved before being used in the equalizer and the extrinsic information generated by the equalizer should be deinterleaved at each stage as shown in Fig. 2. The turbo equalization algorithm can be summarized as follows:

1. *Initialization*: $L_o^0(x_k) = 0, \forall k$.
2. *Iterations*: During the mth iteration, for $m = 1, 2, \ldots, M$
   a. *SISO equalization*: compute extrinsic LLR for the coded bits in the sequence $\boldsymbol{x}$ based on the extrinsic information provided by the outer decoder in the previous iteration and the received signal $\boldsymbol{r}$. Formally, we can denote the output of the SISO equalizer as
   For $k = 1, 2, \ldots, N$, compute $L_i^m(x_k)$ based on $\boldsymbol{r}$, and
   $$(L_o^{m-1}(x_1), \ldots, L_o^{m-1}(x_{k-1}),$$
   $$L_o^{m-1}(x_{k+1}), \ldots, L_o^{m-1}(x_N))$$

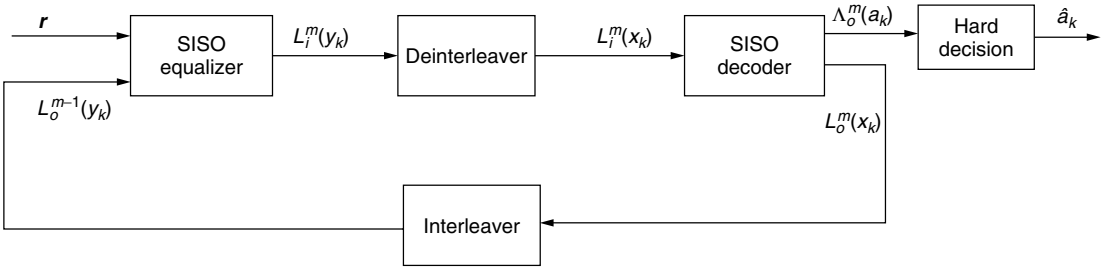


Figure 2. Turbo equalizer structure.

b. *SISO decoding*: compute extrinsic LLRs for the coded bits in the sequence $\boldsymbol{x}$ and the total LLRs for the data bits in the sequence $\boldsymbol{a}$ based on the extrinsic information provided by the SISO equalizer (generated in the previous step):
For $k = 1, 2, \ldots, N$, compute $L_o^m(x_k)$ based on $(L_i^m(x_1), \ldots, L_i^m(x_N))$
For $k = 1, 2, \ldots, K$, compute $\Lambda^m(a_k)$ based on $(L_i^m(x_1), \ldots, L_i^m(x_N))$

c. Hard decision $\hat{a}_k^m = 1$ if $\Lambda(a_k) > 0$ and $\hat{a}_k^m = 0$ otherwise.

d. Check for stopping criterion and if not satisfied, set $m \leftarrow m + 1$ and go to step (a).

4. SOFT-OUTPUT EQUALIZATION

Several algorithms are used to implement the SISO equalizer providing an option to tradeoff complexity for performance. These can be broadly classified into trellis based approaches and filtering based approaches as described below.

4.1. Trellis-Based Approaches

These algorithms exploit the trellis structure of the ISI channel or, equivalently, the Markov nature of the outputs of the ISI channel to efficiently compute the extrinsic LLRs (or APPs). We first note that the ISI channel with L taps can be considered as a convolutional code with constraint length L and a trellis structure can be associated with the channel. At each time instant k, the input to the channel z_k and the state of the ISI channel (which corresponds to the previous $L - 1$ bits) determine the next state. Along with the channel tap coefficients they also determine the output at each time instant. The trellis structure of a 2-tap ISI channel with taps f_0 and f_1 is shown in Fig. 3.

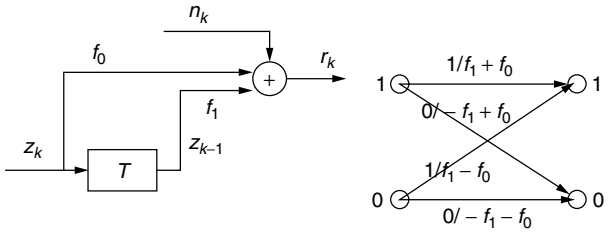


Figure 3. 2-tap ISI channel and trellis.

The optimal equalizer given the channel impulse response and the noise variance at the receiver can be implemented using the Bahl, Cocke, Jelinek, and Raviv (BCJR) algorithm [5]. The BCJR algorithm produces optimal *a posteriori* probabilities (APP's) on x_k given the received signal $\boldsymbol{r}$ by using forward and backward recursions on the trellis of the ISI channel. For an explanation of the algorithm and implementation aspects, the reader is referred to Ref. 6. The BCJR algorithm is optimal; however, its complexity is quite high. Lower-complexity variants of the BCJR algorithm such as the max-log-MAP can also be used [7]. Another alternative is the use of the soft-output Viterbi algorithm (SOVA) as proposed by Hoeher and Hagenauer [8], which modifies the hard-output Viterbi algorithm (VA) to provide reliabilities for the decoded bits in addition to the hard decisions. The SOVA is less complex compared to the BCJR algorithm; however, it is suboptimal. It is also known that the soft-output of the SOVA is optimistic [9] and that the performance can be improved by scaling the extrinsic information produced by the SOVA. Douillard et al. use the SOVA with scaling in their first paper on Turbo equalization [3]. The decoding complexity of the BCJR algorithm, max-log MAP algorithm, and the SOVA increases exponentially with L and even for moderate L, it becomes difficult to implement these algorithms.

One way to reduce the complexity is to use an SISO algorithm based on reduced-state sequence estimation such as the M and T BCJR algorithms [10]. In M-type algorithms, only the best M paths in each stage of the trellis are retained and the rest of the paths are discarded. Hence, the complexity is independent of the channel memory and depends only on M. In T-type algorithms, all paths whose metric is less than a threshold are dropped and, hence, complexity reduction is achieved. However, the reduction in complexity depends on the channel conditions. When used in turbo equalization, during the first few iterations, more number of paths will be retained and as the iterations progress, the effective channel has a higher SNR and, hence, the complexity dynamically decreases. The concepts of M and T algorithms have now been extended to the SOVA and shown to perform well for the case of Turbo equalization [11]. Another technique to reduce the equalization complexity is to truncate the channel memory or use hard-decision feedback from the decoder to cancel the ISI due to some of the taps and use a trellis based equalizer with the shortened channel [12].

4.2. Soft Interference Cancellation with Filtering Approach

A different approach to soft output equalization is to modify conventional hard-output decision feedback equalizers in order to accept and provide soft output. The main advantage of this approach is that the complexity is only $O(L^2)$ compared to the exponential dependence on L for the optimal equalizer. Several different forms of this equalizer have been proposed [13–15]. We will now describe the one due to Wang and Poor [14] in detail. We begin by rewriting (1) in matrix form:

$$\underbrace{\begin{bmatrix} r_{k-\ell_1} \\ \vdots \\ r_k \\ \vdots \\ r_{k+\ell_2} \end{bmatrix}}_{\boldsymbol{r_k}} = \underbrace{\begin{bmatrix} f_{k-\ell_1,\ell_2} \cdots f_{k-\ell_1,0} & \cdots & f_{k-\ell_1,-\ell_1} & 0 & \cdots & 0 \\ \ddots & f_{k,-\ell_2} \cdots & f_{k,0} & \cdots & f_{k,-\ell_1} & 0 & \vdots \\ & \ddots & \cdots & \ddots & \cdots & \ddots & 0 \\ & & f_{k+\ell_2,\ell_2} & \cdots & f_{k+\ell_2,0} & \cdots & f_{k+\ell_2,-\ell_1} \end{bmatrix}}_{\boldsymbol{F_k}} \times \underbrace{\begin{bmatrix} z_{k-\ell_1-\ell_2} \\ \vdots \\ z_k \\ \vdots \\ z_{k+\ell_1+\ell_2} \end{bmatrix}}_{\boldsymbol{z_k}} + \underbrace{\begin{bmatrix} n_{k-\ell_1} \\ \vdots \\ n_k \\ \vdots \\ n_{k+\ell_2} \end{bmatrix}}_{\boldsymbol{n_k}}, \tag{4}$$

or

$$\boldsymbol{r}_k = \boldsymbol{F}_k \boldsymbol{z}_k + \boldsymbol{n}_k, \quad \text{with} \quad \boldsymbol{n}_k \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{\Sigma} = \sigma_{\text{ch}}^2 I). \tag{5}$$

Let us assume that the modulation is BPSK. Hence, there is a simple mapping between the coded bits x_k and the modulated symbols z_k. To keep the notation simple, we will not use interleaved indices, and appropriate interleaving and deinterleaving should be assumed. Based on the extrinsic LLR of the coded bits provided by the channel decoder, $\{L_o^{m-1}(x_k)\}$, we first form soft estimates of the modulated symbols z_k as

$$\begin{aligned} \tilde{z}_k &= 1\,P(z_k = 1) + (-1)P(z_k = -1) \\ &= 1P(x_k = 1) + (-1)P(x_k = 0) \\ &= 1\frac{e^{L_o^{m-1}(x_k)}}{1 + e^{L_o^{m-1}(x_k)}} - 1\frac{1}{1 + e^{L_o^{m-1}(x_k)}} \\ &= \tanh\left(\frac{L_o^{m-1}(x_k)}{2}\right). \end{aligned} \tag{6}$$

Define

$$\tilde{\boldsymbol{z}}_k \triangleq [\tilde{z}_{t-\ell_1-\ell_2}, \ldots, \tilde{z}_{t-1}, 0, \tilde{z}_{t+1}, \ldots, \tilde{z}_{t+\ell_1+\ell_2}]^T \tag{7}$$

Then the soft estimate is used to cancel the intersymbol interference from the received signal r_k to obtain

$$\tilde{\boldsymbol{r}}_k \triangleq \boldsymbol{r}_k - \boldsymbol{F}_k \tilde{\boldsymbol{z}}_k \tag{8}$$

$$= \boldsymbol{F}_k(\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_k) + \boldsymbol{n}_k \tag{9}$$

Next an instantaneous linear MMSE filter is applied to $\tilde{\boldsymbol{r}}_k$, to obtain

$$u_k = \boldsymbol{w}_k^H \tilde{\boldsymbol{r}}_k \tag{10}$$

where the filter $\boldsymbol{w}_k \in \mathbb{C}^{\ell_1+\ell_2+1}$ is chosen to minimize the expected value of mean-square error between the modulated symbol z_k and the filter output u_k:

$$\begin{aligned} \boldsymbol{w}_k &= \arg \min_{\boldsymbol{w} \in \mathbb{C}^{\ell_1+\ell_2+1}} E\{|z_k - \boldsymbol{w}^H \tilde{\boldsymbol{r}}_k|^2\} \\ &= \arg \min_{\boldsymbol{w} \in \mathbb{C}^{\ell_1+\ell_2+1}} \boldsymbol{w}^H E\{\tilde{\boldsymbol{r}}_k \tilde{\boldsymbol{r}}_k^H\}\boldsymbol{w} - 2\Re[\boldsymbol{w}^H E\{z_k \tilde{\boldsymbol{r}}_k\}] \end{aligned} \tag{11}$$

where E denotes expectation and $\Re$ denotes the real part of a complex number. Note that

$$E\{\tilde{\boldsymbol{r}}_k \tilde{\boldsymbol{r}}_k^H\} = \boldsymbol{F}_k \boldsymbol{\Delta}_k \boldsymbol{F}_k^H + \boldsymbol{\Sigma}, \tag{12}$$

$$E\{z_k \tilde{\boldsymbol{r}}_k\} = \boldsymbol{F}_k \boldsymbol{e}, \tag{13}$$

where $\boldsymbol{\Delta}_k$ is defined as

$$\begin{aligned} \boldsymbol{\Delta}_k \triangleq \text{cov}\{\mathbf{z}_k - \tilde{\mathbf{z}}_k\} = \text{diag}\{1 - \tilde{z}_{t-\ell_1-\ell_2}^2, \ldots, 1 - \tilde{z}_{t-1}^2, 1, \\ \times 1 - \tilde{z}_{t+1}^2, \ldots, 1 - \tilde{z}_{t+\ell_1+\ell_2}^2\} \end{aligned} \tag{14}$$

and $\boldsymbol{e}$ denotes a $[2(\ell_1 + \ell_2) + 1]$-vector with all-zero entries, except for the $(\ell_1 + \ell_2 + 1)$th entry, which is 1 (hence $\boldsymbol{F}_k \boldsymbol{e}$ is the $(\ell_1 + \ell_2 + 1)$th column of $\boldsymbol{F}_k$). The solution to (11) is given by

$$\boldsymbol{w}_k = (\boldsymbol{F}_k \boldsymbol{\Delta}_k \boldsymbol{F} + \boldsymbol{\Sigma})^{-1} \boldsymbol{F}_k \boldsymbol{e} \tag{15}$$

In order to form the LLR of the modulated symbol z_k, the instantaneous MMSE filter output u_k in (10), which represents a soft estimate of the bit z_k, is treated as a Gaussian random variable:

$$p(u_k \mid z_k) \sim \mathcal{N}_c(\mu_k z_k, \nu_k^2) \tag{16}$$

Conditioned on the modulated symbol z_k, the mean and variance of u_k are given by

$$\begin{aligned} \mu_k &\triangleq E\{u_k \mid z_k\} \\ &= \boldsymbol{e}^T \boldsymbol{F}_k^H (\boldsymbol{F}_k \boldsymbol{\Delta}_k \boldsymbol{F}_k^H + \boldsymbol{\Sigma})^{-1} \boldsymbol{F}_k \boldsymbol{e} \end{aligned} \tag{17}$$

$$\begin{aligned} \nu_k^2 &\triangleq \text{var}\{u_k \mid z_k\} = E\{|u_k|^2 \mid z_k\} - \mu_k^2 = \boldsymbol{w}_k^H E\{\tilde{\boldsymbol{y}}_k \tilde{\boldsymbol{y}}_k^H\} \boldsymbol{w}_k - \mu_k^2 \\ &= \mu_k - \mu_k^2 \end{aligned} \tag{18}$$

Note that the mean μ_k is real. Therefore the extrinsic information $L_i^m(x_k)$ delivered by the SISO equalizer is given by

$$\begin{aligned} L_i^m(x_k) &= \log \frac{P(u_k \mid x_k = 1)}{P(u_k \mid x_k = 0)} = \log \frac{p(u_k \mid z_k = +1)}{p(u_k \mid z_k = -1)} \\ &= -\frac{|u_k + \mu_k|^2}{\nu_k^2} + \frac{|u_k - \mu_k|^2}{\nu_k^2} \\ &= \frac{4\Re\{\mu_k u_k\}}{\nu_k^2} = \frac{4\Re\{u_k\}}{1 - \mu_k} \end{aligned} \tag{19}$$

For real-valued channels, $L_i^m(x_k) = \dfrac{2u_k}{1 - \mu_k}$.

4.3. SISO Decoding

When convolutional codes are used as outer codes, the BCJR algorithm, any of its variants discussed in Section 4.1, or the SOVA can be used to generate soft output for the coded bits and the information bits. When parallel concatenated or serial concatenated convolutional codes (SCCC) are used, the Turbo decoding algorithm explained in Ref. 6 can be used. For low-density parity-check codes, the belief propagation decoder [16] naturally produces soft output in order to iterate between the SISO equalizer and the decoder. When PCCC, SCCC, or LDPC outer codes are used, the decoding algorithm for soft-output decoding of the outer code itself is an iterative algorithm, which can be combined with the iterations between the decoder and the equalizer such as in Ref. 17.

We show some simulation results to demonstrate the efficiency of the Turbo equalization algorithm. Figure 4 shows the bit error rate as a function of the number of iterations for a 5-tap ISI channel with frequency response $F_1(z) = \sqrt{0.45} + \sqrt{0.25}z^{-1} + \sqrt{0.15}z^{-2} + \sqrt{0.1}z^{-3} + \sqrt{0.05}z^{-4}$ when the outer code is a 16-state convolutional code with generator polynomials $[1 + D^2 + D^4, 1 + D + D^2 + D^4]$ and the block length is $K = 2048$ information bits. The equalizer and the decoder use the SOVA. The performance of the convolutional code in an AWGN channel is also shown. It can be seen that the bit error rate improves with iterations and the performance after 6 iterations is comparable to the performance of the code on an AWGN channel. Figure 5 shows the performance of a parallel concatenated outer code where the component codes are 16-state convolutional codes on the same channel for $K = 5000$. In every iteration of turbo equalization one iteration of Turbo decoding is performed within the outer decoder. The performance can be seen to improve steadily with iterations and is within 0.8 dB from the capacity for this channel. These results show that turbo equalization is quite effective in removing ISI.

5. PERFORMANCE ANALYSIS

It is quite difficult to analytically predict the performance of the Turbo equalizer for a given interleaver. However, it is possible to derive bounds on the performance over the ensemble of all possible interleavers both when E_b/N_o is very high and when the length of the codewords $N \to \infty$. The following two types of analysis—distance spectrum based analysis and analysis of the iterative algorithm provide some insight into the performance of Turbo equalization.

5.1. Distance-Spectrum-Based Analysis

The main idea here is to treat the ISI channel as a convolutional code over complex (or real) field and, hence, to treat the overall system in Fig. 1 as a serial concatenated convolutional code (SCCC) where the error correction code is the outer code and the ISI channel is the inner code. Then, the performance of an ML decoder can be computed over the ensemble of all interleavers, according to the technique pioneered by Benedetto et al. [18]. Our explanation below is based on Ref. 18 but is adapted to the case when the inner code is an ISI channel.

Since data are transmitted in the form of blocks, we can think of the outer code and the ISI channel as equivalent block codes of codeword length N. For a given reference codeword (sequence), let $A^{C_o}(w, h)$ denote the total number of error sequences with input Hamming weight w and output Hamming weight h for the outer

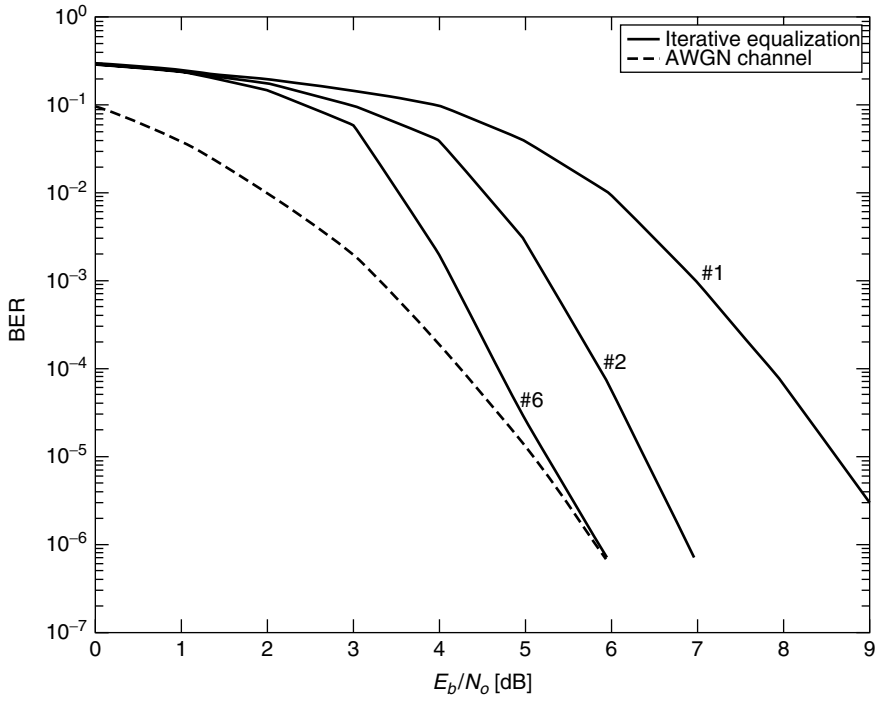


Figure 4. Bit error rate performance with turbo equalization for 5-tap ISI channel and 16-state rate-$\frac{1}{2}$ convolutional outer code.

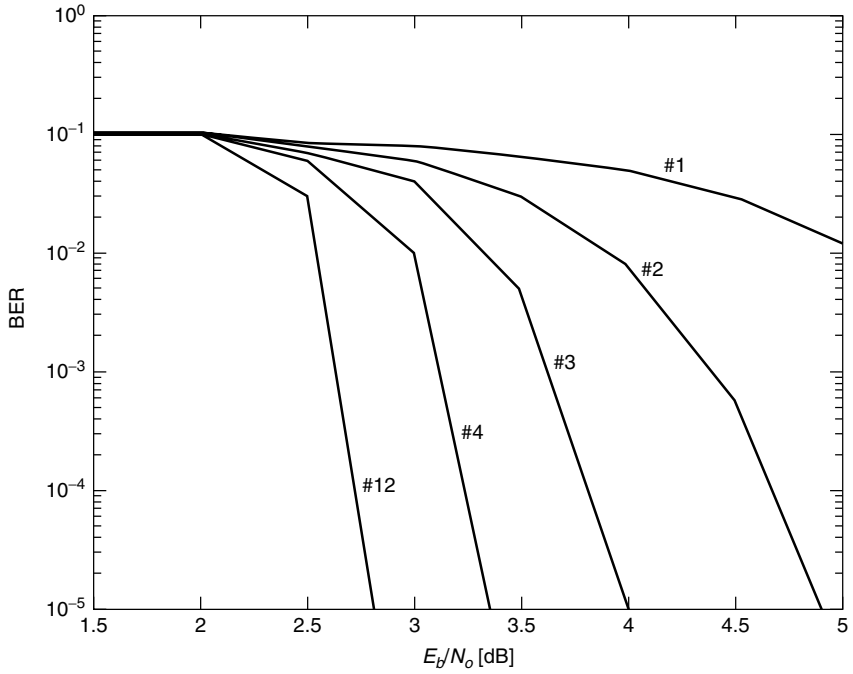


Figure 5. Bit error rate performance with Turbo equalization for 5-tap ISI channel and 16-state rate-$\frac{1}{2}$ Turbo outer code.

code. Each error sequence may result in one or more error events and h is the Hamming weight of the sum of all error events. When the outer code is a linear code, the all-zero sequence can be taken as the reference codeword. Then, the number of error sequences of a given w and h is the same as that of the number of codewords of the outer code with Hamming weight h and information weight w. Similarly, let $A^{C_{\text{ISI}}}(h, \delta^2)$ denote the average number of error sequences with input weight h and squared Euclidean distance (SED) δ^2 for the ISI channel. The average is over all possible reference sequences, since the SED corresponding to an error sequence depends on the reference codeword as well and, hence, the all-zero sequence cannot be assumed as the reference. Techniques used to compute $A^{C_{\text{ISI}}}(h, \delta^2)$ can be found in the literature [19,20]. Under the assumption of a uniform interleaver, that is, over the ensemble of all interleavers, the average number of error sequences of input weight w and SED δ^2, $A^{C_s}(w, \delta^2)$ can be shown to be [18]

$$A^{C_s}(w, \delta^2) = \sum_{h=d_f^o}^{N} \frac{A^{C_o}(w, h) \times A^{C_{ISI}}(h, \delta^2)}{\binom{N}{h}} \tag{20}$$

where N is the length of the interleaver and d_f^o is the free distance of the outer code. The probability of bit error under maximum-likelihood decoding can be upper bounded by using the union bound as

$$P_b(e) \leq \sum_{w=1}^{K} \frac{w}{K} \sum_{h} \sum_{\delta^2 \in \Delta} A^{C_s}(w, \delta^2) Q\left(\sqrt{\frac{\delta^2 \text{RE}_b}{4N_o}}\right) \tag{21}$$

where Δ is the set of all possible squared Euclidean distances δ^2 and R is the rate of the outer code. This result should be interpreted with some caution. On one hand, it is an upper bound computed over the ensemble of all possible interleavers and it is possible to design interleavers that perform better than the average. On the other hand, the Turbo equalization algorithm is not a true ML decoding algorithm and, hence, the performance can be worse than that predicted by (21). Nevertheless, the abovementioned approach can be used to derive some understanding of the effect of the interleaver and different parameters of the outer code. For large E_b/N_o, only the first few terms of the summation corresponding to small values of δ^2 and h dominate the performance. For small h, $A^{C_o}(w, h)$ and $A^{C_{ISI}}(h, \delta^2)$ can be approximated as [18]

$$A^{C_o}(w, h) \approx \sum_{n=1}^{n_{\max}^o(w)} T^{C_o}(w, h, n) \binom{N}{n} \tag{22}$$

$$A^{C_{ISI}}(h, \delta^2) \approx \sum_{n=1}^{n_{\max}^i(h)} T^{C_{ISI}}(h, \delta^2, n) \binom{N}{n} \tag{23}$$

where $T^{C_o}(w, h, n)$ is the number of error events for the outer code of input weight w and output weight h that are the concatenation of exactly n consecutive error events. By consecutive we mean that the error path diverges from the reference path and on merging back immediately diverges again until n such error events occur. Then, the paths remain merged until the end of the block. Similarly, let $T^{C_{\text{ISI}}}(h, \delta^2, n)$ be the number of error events of input weight h and output SED δ^2 that are the concatenation of n consecutive error events for the ISI channel. The quantities $n_{\max}^o(w)$ and $n_{\max}^i(h)$ refer to the maximum number of error events possible due an input error sequence of weight w and h for the outer code and ISI channel, respectively. It is important to note that for *convolutional* outer codes, the quantities $T^{C_o}(w, h, n)$ and $T^{C_{\text{ISI}}}(h, \delta^2, n)$ in (23) are independent of N. The probability

of bit error can then be rewritten as

$$P_b(e) \le \sum_w \frac{w}{K} \sum_h \sum_{\delta^2} \sum_{n^o=1}^{n^o_{\max}(h)} \sum_{n^i=1}^{n^i_{\max}(h)} T^{C_o}(w,h,n^o) \times T^{C_{\mathrm{ISI}}}(h,\delta^2,n^i) \times \frac{\binom{N}{n^o}\binom{N}{n^i}}{\binom{N}{h}} Q\left(\sqrt{\frac{\delta^2 \mathrm{RE}_b}{4N_o}}\right) \quad (24)$$

Using the approximation $\binom{N}{n} \approx \frac{N^n}{n!}$, and the fact that $K = NR$, the probability of bit error can be written as

$$P_b(e) \le \sum_w \frac{w}{R} \sum_h \sum_{\delta^2} \sum_{n^o=1}^{n^o_{\max}(h)} \sum_{n^i=1}^{n^i_{\max}(h)} T^{C_o}(w,h,n^o) \times T^{C_{\mathrm{ISI}}}(h,\delta^2,n^i) N^{n^o+n^i-h-1} Q\left(\sqrt{\frac{\delta^2 \mathrm{RE}_b}{4N_o}}\right) \quad (25)$$

where we can see that the probability of bit error depends on the interleaver length through the term $N^{n^o+n^i-h-1}$. Since the ISI channel is a non-recursive encoder, $n^i_{\max}(h) = h$ [6,18] and, hence, the exponent of N is always greater than or equal to zero. Therefore, the probability of bit error is at best independent of the length N of the codewords, which is the same situation as that for convolutional codes. This means that although, the system in Fig. 1 is a serial concatenated convolutional code, there is no interleaving gain when the outer code is a simple convolutional code, since the ISI channel is a nonrecursive inner code. The interleaver is still useful since the h different error events each correspond to an SED of at least $\delta^2_{\min}$ and, hence, the overall free distance can be up to $d^o_f \times \delta^2_{\min}$. The interleaver is also required to break up the correlation at the output of the equalizer.

This tells us that in order to achieve close to capacity performance, the outer code should be a Turbo code or an LDPC code, in which case, the performance of the code gets better with increasing N and, hence, the overall performance improves with increasing N.

5.1.1. Binary Precoding. It is known from Benedetto et al. [18] that an interleaving gain is possible even with a simple convolutional outer code if the inner encoder is recursive. It is possible to make the ISI channel appear recursive to the outer code by encoding the interleaved output $\boldsymbol{y}$ in Fig. 1 by a rate-1 recursive convolutional encoder as shown in Fig. 6. The precoder is a recursive encoder with generator polynomial $1/g(D)$ with $g(D) = \sum_i^J g_i D^i$, where J is the number of memory elements in the precoder and $g_i \in \{0,1\}$. The outputs of the precoder are transmitted over the ISI channel after modulation. This type of precoding, which we refer to as *binary precoding*, should not be confused with the conventional type of precoding such as Tomlinson–Harashima (TH) precoding, which is intended to cancel ISI at the transmitter. Unlike TH precoding, in binary precoding, channel knowledge is not required at the transmitter and the purpose is only to make the channel trellis appear recursive. When the modulation is memoryless, the trellis of the precoder and that of the ISI channel can be combined into one trellis as shown in Fig. 7. If $J \le L$, then the resulting combined trellis has the same number of states as that of the ISI channel and, hence, there is no increase in the equalization complexity. However, since the precoder is recursive, the combination of the channel and the precoder becomes recursive and, hence, a significant interleaving gain is possible. From the combined transversal filter, the trellis diagram can be easily found. An example of a precoded 2-tap ISI channel with $g(D) = 1 + D$ and the associated trellis diagram is shown in Fig. 8. The ISI channel considered is the same as that in Fig. 3 with transfer function $F_2(z) = f_0 + f_1 z^{-1}$.

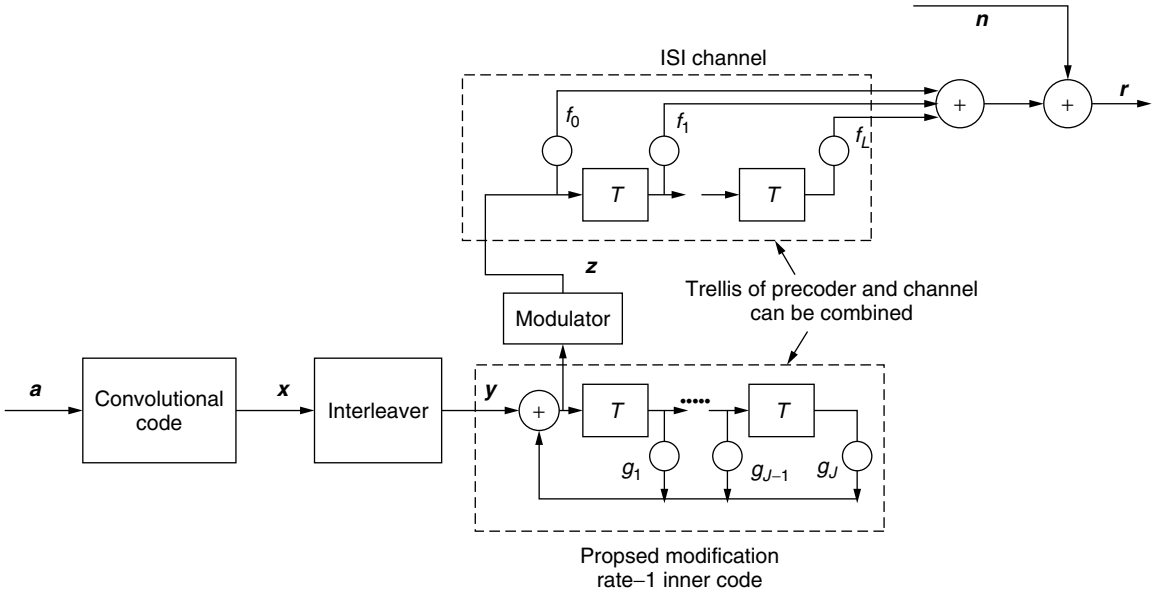


Figure 6. System model with binary precoding.

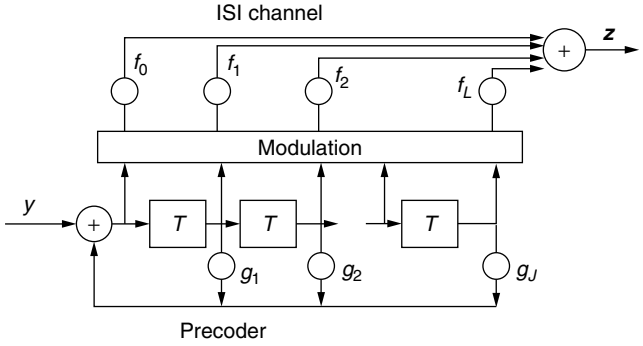


Figure 7. Transversal filter for precoded ISI channel.

Note from the trellis in Fig. 8 that for the precoded channel, a weight-1 error sequence produces an infinite-length error event and that a minimum input weight of 2 is required to produce a finite-length error event. Therefore, an input error sequence with weight h can produce a maximum of $\lfloor h/2 \rfloor$ error events and $n^i_{\max}(h) = \lfloor h/2 \rfloor$ in (25). Therefore, the maximum exponent of N is $\lfloor h/2 \rfloor - h$ and since $h \geq d^o_f$, the exponent of N is always negative if $d^o_f \geq 3$. Thus the BER in (25) decreases with increasing interleaver length; this phenomenon has been termed as interleaving gain [18].

Since the free distance of the outer code needs to be greater than or equal to only 3, even simple codes such as convolutional codes with small constraint length usually perform very well with binary precoding. Thus significant reduction in complexity can be obtained compared to using Turbo outer codes such as in Ref. 21 or, better performance can be obtained compared to using convolutional outer codes without precoding such as in Ref. 3. Figure 9 shows the bit error rate performance

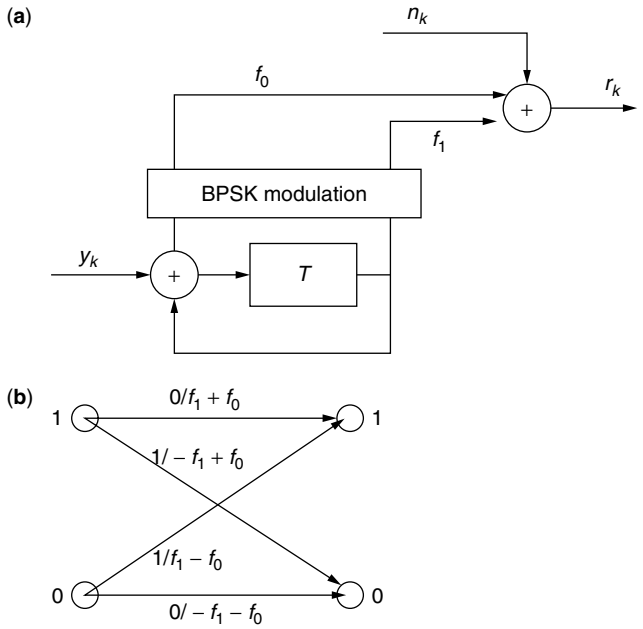


Figure 8. (**a**) Two-tap ISI channel with precoding; (**b**) trellis of precoded channel.

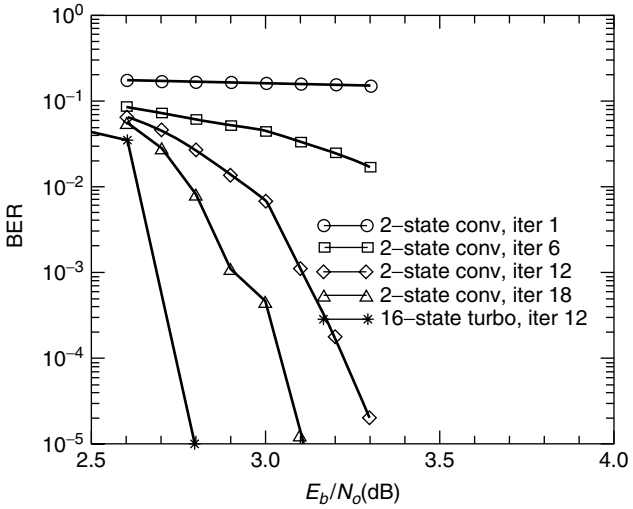


Figure 9. Bit error rate performance for a 5-tap static ISI channel with and without binary precoding.

on a 5-tap ISI channel with transfer function $F_1(z) = \sqrt{0.45} + \sqrt{0.25}z^{-1} + \sqrt{0.15}z^{-2} + \sqrt{0.1}z^{-3} + \sqrt{0.05}z^{-4}$ for a block length of $K = 5000$ bits. The outer code used is a simple rate-$\frac{1}{2}$, 2-state convolutional code with generator polynomials $[1, 1 + D]$. The precoder used is $1/(1 + D)$ and the error performance is shown for different iterations up to 18 iterations. Also shown for comparison is the performance of a 16-state Turbo code (parallel concatenated code) with 12 iterations. It can be seen that even a simple 2-state outer code can provide performance within 0.3 dB as that of a 16-state Turbo code; but the decoding complexity for the former is significantly lesser compared to that of the Turbo code.

The combination of the outer code, interleaver, and the rate-1 recursive precoder can be thought of as a serial concatenated outer code with a recursive inner encoder whose output is transmitted over the ISI channel and that the interleaving gain is really due to the concatenated outer code. However, it should be noted that there is no interleaver between the precoder and the ISI channel and, hence, a separate decoder is not required for the precoder. By combining the trellis of the precoder and the channel, the equalization complexity is reduced and, hence, it is useful to think of this technique as a precoding technique rather than a mere concatenated outer code. Analysis of precoded ISI channels based on the distance spectrum for partial response channels can be found elsewhere [22–24]. The design of precoders based on optimizing the distance spectrum has been considered by Lee [25].

5.2. Analysis of the Iterative Equalization Algorithm

The aforementioned analysis is based on the assumption of an ML decoder whereas the Turbo equalization algorithm is not an ML decoding algorithm and, therefore, the performance of the Turbo equalization can be quite different from that predicted by the analysis above. In order to accurately characterize the iterative process, we need to determine how the extrinsic information evolves from one iteration to another. Since the extrinsic

information vector is an N-dimensional vector, it is quite difficult to characterize the evolution of this vector. When the length $N \to \infty$, we can assume that the extrinsic information $(\ldots, L_i^m(x_k), \ldots,)$ or $(\ldots, L_o^m(x_k), \ldots)$ are sequences of independent identically distributed random variables and, hence, only the PDFs of $L_i^m(x_k)$ and $L_o^m(x_k)$ need to be tracked from one iteration to another. It is still quite difficult to compute the pdf as a function of iterations. Ten Brink [26] and El Gamal and Hammons [27] showed that for Turbo codes, if the PDF is assumed to be Gaussian and only a single parameter related to the PDF is tracked, the approximation is still quite good. This idea was extended to analyze Turbo equalization in Ref. 28. Here, we will explain how to use this technique to analyze the performance of Turbo equalization.

We first introduce the concept of equivalent channels. Consider the interleaved coded sequence $\boldsymbol{y} = \{y_k\}$ and BPSK modulation with $z_k = (2y_k - 1)$. During the mth iteration in the Turbo equalization algorithm, the inner equalizer provides LLRs (or extrinsic information) $L_i^m(x_k)$. This LLR can be thought of as the output of a hypothetical channel with binary inputs as "seen" by the outer decoder. Let us define two quantities for this equivalent channel:

$$\mu_i^m \triangleq E_{y_k}[(2y_k - 1)L_i^m(y_k)] \tag{26}$$

$$(\sigma_i^m)^2 \triangleq E_{y_k}[((2y_k - 1)L_i^m(y_k))^2] = E_{y_k}[(L_i^m(y_k))^2] \tag{27}$$

A measure of reliability of this equivalent channel is the ratio $\mathrm{SNR}_i^m = \left(\dfrac{\mu_i^m}{\sigma_i^m}\right)^2$, which can be thought of as the SNR of the channel as seen by the outer decoder during the mth iteration. The higher SNR_i^m is, the more reliable the effective channel as seen by the outer decoder during the mth iteration is. Similarly, the outer decoder produces extrinsic information $L_o^m(y_k)$ during the mth iteration and a similar SNR, SNR_o^m can be defined as the SNR of the equivalent channel as seen by the equalizer during the mth iteration. During the mth iteration, the inner SISO equalizer is provided with an equivalent channel with SNR SNR_o^{m-1} by the outer code and an AWGN channel with variance σ_{ch}^2. At the output, the equalizer produces an equivalent channel with SNR SNR_i^m. The outer decoder in turn observes this channel and increases the equivalent SNR at the output of the decoder to SNR_o^m. Let us denote the SNR at the output of the equalizer by the transfer function $\mathcal{F}_i(p, q)$, where p is the input SNR from the outer decoder and $q = \sigma_{\mathrm{ch}}^2$ is the variance of the AWGN. Similarly, let $\mathcal{F}_o(p)$ denote the equivalent SNR at the decoder output when the input SNR is p. Then, we have

$$\mathrm{SNR}_i^m = \mathcal{F}_i(\mathrm{SNR}_o^{(m-1)}, \sigma_{\mathrm{ch}}^2) \tag{28}$$

$$\mathrm{SNR}_o^m = \mathcal{F}_o(\mathrm{SNR}_i^m) \tag{29}$$

The evolution of the equivalent SNR with iterations is best illustrated with the help of a diagram such as the one suggested by Ten Brink [26] and shown in Figs. 10 and 11. Figure 10 shows two curves: the function $\mathcal{F}_o(p)$ and the function $\mathcal{F}_i^{-1}(p, \sigma_{\mathrm{ch}}^2)$ for a fixed σ_{ch}^2 [so we refer to this simply as $\mathcal{F}_i^{-1}(p)$]. The function $\mathcal{F}_i^{-1}(p)$ is represented by simply drawing the function $\mathcal{F}_i(p)$ with the x and y axes inverted. Since the output of one of the SISOs is the input to the other, by drawing the curves with

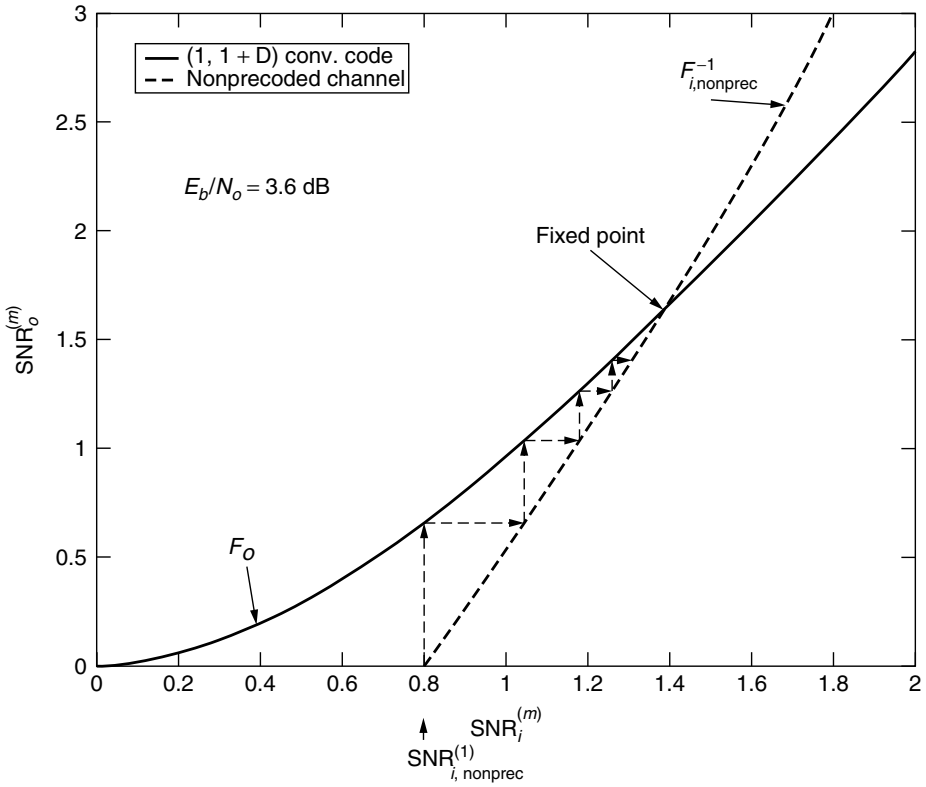


Figure 10. Graphical demonstration of convergence of iterative equalization and decoding: serial concatenation between a 2-state convolutional outer code and a 5-tap ISI channel.

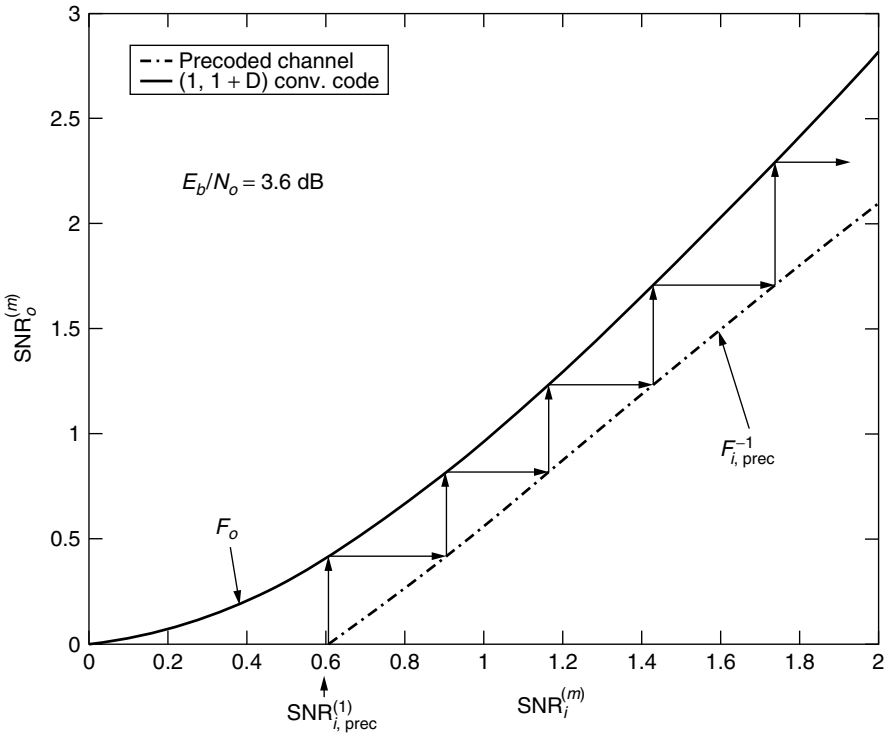


Figure 11. Graphical demonstration of convergence of iterative equalization and decoding: serial concatenation between a 2-state convolutional outer code and a precoded 5-tap ISI channel.

the axis reversed for one of the curves, the evolution of the SNRs can be traced by drawing vertical and horizontal lines between the two curves. That is, the iterations begin at the point $\text{SNR}_i^{(1)}$ and a vertical line from the curve $\mathcal{F}_i^{-1}(p)$ to the curve $\mathcal{F}_o(p)$ followed by a horizontal line to the curve $\mathcal{F}_i^{-1}(p)$ denotes the change of the equivalent SNR during one iteration. The outer code in this case is a 2-state convolutional code and the channel is the 5-tap ISI channel with frequency response $F_1(z) = \sqrt{0.45} + \sqrt{0.25}z^{-1} + \sqrt{0.15}z^{-2} + \sqrt{0.1}z^{-3} + \sqrt{0.05}z^{-4}$ and the $E_b/N_o = 3.6$ dB. It can be seen from Fig. 10 that the two curves intersect, which represents a fixed point for the iterations and the equivalent SNR cannot improve beyond the fixed point. The achievable bit error rate can be computed from the equivalent SNR corresponding to the fixed point. If the two curves $\mathcal{F}_o(p)$ and $\mathcal{F}_i^{-1}(p)$ do not intersect, then the equivalent SNR steadily increases to infinity, which denotes correct decoding or convergence of the turbo equalization algorithm to the correct solution. Such a situation occurs for the same convolutional code and same E_b/N_o, when precoding is used with the channel and is shown in Fig. 11.

Note that in order for the iterative procedure to converge to the correct decision, it is not necessary that both SNR_i^m and SNR_o^m go to infinity; it is sufficient that any one of them, SNR_i^m or SNR_o^m tends to ∞, since the SNRs are defined using the extrinsic LLRs only. The shape of $\mathcal{F}_i^{-1}(p)$ depends on the σ_{ch}^2 (and, hence, on E_b/N_o). The minimum E_b/N_o for which $\mathcal{F}_i^{-1}(p, E_b/N_o)$ does not intersect $\mathcal{F}_o(p)$ is called the *threshold*. In general, for finite impulse response ISI channels, $\mathcal{F}_i^m(\infty, \sigma_{\text{ch}}^2)$ is finite and, hence, in order to get arbitrarily small probability of error for a fixed σ_{ch}^2 capacity approaching codes such as Turbo codes or LDPC codes must be used. With convolutional outer codes, the bit error rate cannot be made arbitrarily small. However, with precoding $\mathcal{F}_i^m(\infty, \sigma_{\text{ch}}^2) \to \infty$ and, hence, even simple convolutional codes can be used to obtain arbitrarily small probability of error.

Finally, it should be noted that parameters other than the equivalent SNR have been used to characterize the iterative process. Ten Brink used mutual information [26], and Narayanan has used the correlation between the equivalent soft output and the actual coded bit [28].

5.2.1. Computing the Transfer Functions. For trellis-based equalizers such as the BCJR algorithm or SOVA, it is quite difficult to analytically compute the functions $\mathcal{F}_i$; hence, Monte Carlo simulations have to be used and ensemble averages in (27) are replaced by a time average. In order to simulate the input extrinsic information, $L_o^m(y_k)$ can be assumed to be a Gaussian random variable whose variance is twice the mean. This assumption has been shown to be reasonably accurate [29]. The following procedure is then used to evaluate $\mathcal{F}_i(p, q)$:

1. Draw a sequence of random i.i.d. bits $y_k \in \{0, 1\}$, set $z_k = 2y_k - 1$ for $k = 1, 2, \ldots, N$
2. Compute $r_k = \sum_{i=-l_1}^{l_2} f_{k-i} z_k + n_k$ where n_k are samples of i.i.d. Gaussian random variables with zero mean and variance q.
3. Draw a sequence of random variables for $L_o^{(m-1)}(y_k)$ from the distribution $\mathcal{N}(2pz_k, 4p)$, for $k = 1, 2, \ldots, N$.
4. Compute the equalizer output $L_i^m(y_k)$, $k = 1, 2, \ldots, N$.

5. Compute $\mathrm{SNR}_i^m = \dfrac{\left(\frac{1}{N}\sum_{k=1}^{N} z_k L_i^m(y_k)\right)^2}{\frac{1}{N}\sum_{k=1}^{N}(z_k L_i^m(y_k))^2}$.

When MMSE equalizers are used such as the one described in Section 4.2, it is possible to obtain the output SNR_i^m semianalytically. That is, we do not have to simulate the equalizer. The following procedure can be used to compute $\mathcal{F}_i(a,b)$ — given the channel $\boldsymbol{F}_t = \boldsymbol{F}$, the noise variance $\sigma_{\mathrm{ch}}^2 = q$:

1. For $i = 1, 2, \ldots, N$ draw i.i.d. $L_i^m(y_k) \sim \mathcal{N}(2az_k, 4a)$
2. Let $\boldsymbol{\Delta}_k \triangleq \mathrm{diag}\left\{1 - \tanh\left(\frac{L_i(y_{k-\ell_1-\ell_2})}{2}\right)^2, \ldots, 1 - \tanh\left(\frac{L_i(y_{k-1})}{2}\right)^2, 1, 1 - \tanh\left(\frac{L_i(y_{k+1})}{2}\right)^2, \ldots, 1 - \tanh\left(\frac{L_i(y_{k+\ell_1+\ell_2})}{2}\right)^2\right\}$.
3. Compute

$$\mu_k \triangleq \boldsymbol{e}^T \boldsymbol{F}^H \left(\boldsymbol{F}\boldsymbol{\Delta}_k \boldsymbol{F}^H + \boldsymbol{\Sigma}\right)^{-1} \boldsymbol{F}\boldsymbol{e} \tag{30}$$

$$\mathrm{SNR}_i^m \triangleq \frac{1}{N}\sum_{k=1}^{N} \frac{1}{2}\frac{2q\mu_k}{1-\mu_k} \tag{31}$$

where $q = 1$ for real channels and $q = 2$ for complex channels. The transfer function of the outer code $\mathcal{F}_o(a)$ cannot be computed analytically and has to be computed via Monte Carlo simulations using the following procedure:

1. Draw a sequence of random i.i.d. bits $y_k \in \{0, 1\}$.
2. Draw a sequence of random variables for $L_i^m(y_k)$ from the distribution $\mathcal{N}(2az_k, 4a)$, for $k = 1, 2, \ldots, N$.
3. Compute the decoder output $L_o^m(y_k),\ k = 1, 2, \ldots, N$.
4. Compute $\mathrm{SNR}_o^m = \dfrac{\left(\frac{1}{N}\sum_{k=1}^{N} z_k L_o^m(y_k)\right)^2}{\frac{1}{N}\sum_{k=1}^{N}(z_k L_o^m(y_k))^2}$.

6. MORE RECENT RESULTS

Turbo equalization is an area of active research and new results on theory and application are emerging. Turbo equalization with carefully designed LDPC outer codes has been shown to provide close to capacity performance. The design of LDPC codes with different kinds of equalizers is an area of current research. Reduced-complexity Turbo equalization, adaptive Turbo equalization, and applications to multiantenna systems and other wireless systems [30,31] are few of the areas under study.

Acknowledgments

The author would like to thank Dũng Ngoc Doan for help with Figs. 10 and 11 and for proofreading the manuscript. The author would also like to thank Prof. Xiaodong Wang for help with the material in section 4.2.

BIOGRAPHY

Krishna R. Narayanan received his Ph.D. degree from Georgia Institute of Technology, Atlanta, in 1988. Since then he has been an assistant professor in the Department of Electrical Engineering at Texas A&M University. His research interests are mainly in the areas of advanced modulation, coding and receiver design for wireless communications, and magnetic recording. Specifically, he has worked in the areas of turbo coding, and iterative signal processing. He currently serves as an associate editor for *IEEE Communication Letters* and an editor for *IEEE Transactions on Wireless Communications*. He is a recipient of the CAREER award from the National Science Foundation in 2001.

BIBLIOGRAPHY

1. J. G. Proakis, *Digital Communications*, McGraw-Hill, 2001.
2. A. Duel-Hallen and C. Heegard, Delayed decision feedback sequence estimation, *IEEE Trans. Commun.* **37**: 428–436 (May 1989).
3. C. Douillard, Iterative correction of intersymbol interference: Turbo-equalisation, *European Trans. Commun.* 507–511 (Sept.–Oct. 1995).
4. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding, *Proc. IEEE Int. Conf. Communication*, (Geneva, Switzerland), June 1993, pp. 1064–1070.
5. L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **20**: 284–287 (March 1974).
6. J. Proakis, ed., *Wiley Encyclopedia in Telecommunications*, "Concatenated coding and iterative decoding," Wiley, 2002.
7. P. Roberston, E. Villebrun, and P. Hoeher, A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain, *Proc. IEEE Int. Conf. Communication*, 1995, pp. 1009–1013.
8. J. Hagenauer and P. Hoher, A Viterbi algorithm with soft-decision outputs and its applications, *Proc. IEEE GLOBECOM*, (Dallas, TX), Nov. 1989, pp. 47.1.1–47.1.7.
9. L. Pake and P. Robertson, Improved decoding with SOVA in a parallel concatenated (turbo-code) scheme, *Proc. IEEE Int. Conf. Communication*, 1996, pp. 102–106.
10. V. Franz and J. B. Anderson, Concatenated decoding with a reduced search BCJR algorithms, *IEEE J. Select. Areas Commun.* **16**: 186–195 (Feb. 1998).
11. K. R. Narayanan, U. D. Gupta, and B. Lu, Low-complexity iterative equalization and decoding with binary precoding, *Proc. IEEE Int. Conf. Communication*, 2000, pp. 1–5.
12. S. H. Muller, W. H. Gerstacker, and J. B. Huber, Reduced state soft output trellis equalization incorporating soft

feedback, *Proc. IEEE Global Communication Conf.*, 1996, pp. 95–100.

13. A. Glavieux, C. Laot, and J. Labat, Turbo equalization over a frequency selective channel, *Proc. Int. Symp. Turbo Codes and Related Topics*, Sept. 1997, pp. 96–102.
14. X. Wang and H. Poor, Iterative (turbo) soft interference cancellation and decoding for coded CDMA, *IEEE Trans. Commun.* **47**: 1046–1061 (July 1999).
15. M. Tüchler, R. Koetter, and A. Singer, Iterative correction of isi via equalization and decoding with priors, *Proc. IEEE Int. Symp. Information Theory*, 2000, p. 100.
16. D. MacKay and R. M. Neal, Near shannon limit performance of low density parity check codes, *Electron. Lett.* 457–458 (March 1996).
17. J. Fan, E. Kurtas, A. Friedmann, and S. W. McLaughlin, Low density parity check codes for digital magnetic recording, *Proc. Allerton Conf. Communication Control and Computing*, 1999.
18. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Serial concatenation of interleaved codes: Design and performance analysis, *IEEE Trans. Inform. Theory* **42**: 409–429 (April 1998).
19. E. Zehavi and J. K. Wolf, On the performance evaluation of trellis codes, *IEEE Trans. Inform. Theory* **33**: 483–501 (July 1987).
20. S. A. Raghavan, J. K. Wolf, and L. B. Milstein, On the performance evaluation of isi channels, *IEEE Trans. Inform. Theory* **39**: 957–965 (1993).
21. D. Raphaeli and Y. Zarai, Combined turbo equalization and turbo decoding, *IEEE Commun. Lett.* **2**: 107–109 (April 1998).
22. T. Souvignier et al., Turbo decoding for PR4: Parallel versus serial concatenation, *Proc. IEEE Int. Conf. Communication*, 1999, pp. 1638–1642.
23. T. Duman and E. Kurtas, Performance bounds for high rate linear codes over partial response channels, *IEEE Trans. Inform. Theory* **47**: 1201–1205 (March 2001).
24. L. McPheters, S. W. McLaughlin, and K. R. Narayanan, Precoded PRML, serial concatenation, and iterative (turbo) decoding for digital magnetic recording, *IEEE Trans. Magn.* 2325–2327 (Sept. 1999).
25. I. Lee, The effect of a precoder on serially concatenated coding system with ISI channel, *IEEE Trans. Commun.* 1168–1175 (July 2001).
26. S. Tenbrink, Iterative decoding for multicode CDMA, *Proc. IEEE Vehicular Technology Conf.*, 1999, pp. 1876–1880.
27. H. El Gamal and R. Hammons, Analyzing the turbo decoder using the gaussian assumption, *Proc. Conf. Information Sciences and Systems*, March 2000.
28. K. R. Narayanan, Effect of precoding on the convergence of turbo equalization for partial response channels, *IEEE J. Select. Areas Commun.* **19**: 686–698 (April 2001).
29. D. Divsalar, S. Dolinar, and F. Pollara, Iterative turbo decoder analysis based on density evolution, *IEEE J. Select. Areas Commun.* 891–907 (Feb. 2001).
30. G. Bauch, H. Khorram, and J. Hagenauer, Iterative equalization and decoding in mobile communications systems, *Proc. Eur. Personal Mobile Communication Conf*, 1997, pp. 307–312.
31. G. Bauch and V. Franz, Iterative equalization and decoding for the GSM system, *Proc. IEEE Vehicular Technology Conf.* (Ottawa), May 1998, pp. 2262–2266.

TURBO PRODUCT CODES FOR OPTICAL CDMA SYSTEMS

STEVEN W. MCLAUGHLIN
CENK ARGON
Georgia Institute of Technology
Atlanta, Georgia

1. INTRODUCTION

There exist various access schemes for participants of an optical fiber communications system [1]:

1. Time-division multiple access (TDMA)
2. Wavelength-division multiple access (WDMA)
3. Code-division multiple access (CDMA)

In TDMA [2] systems, each user (or transmitter) is assigned a specific time slot for data transmission. On the other hand, a specific wavelength is reserved for each transmitter in WDMA [1] systems. CDMA [2] is a different approach from TDMA or WDMA. In CDMA, not a wavelength or a time slot, but a specific pseudorandom address sequence is assigned to each participant. This is advantageous if compared to TDMA and WDMA because pseudorandom address sequences enable data transmission of users at overlapping times and wavelengths. Other advantages of CDMA include [2] (1) security against unauthorized users, (2) protection against jamming, (3) flexibility of adding users, and (4) asynchronous access capability.

In an optical CDMA system [3,4], an optical orthogonal code (OOC) sequence is assigned to each user and data are sent via the destination user's OOC sequence. Different from electrical pseudorandom sequences, which can consist of bipolar pulses, OOC sequences are unipolar, where instead of the -1 and $+1$ levels, only the 0 and $+1$ levels are available in intensity-modulated, direct-detection optical systems. Because of this, true orthogonality cannot be achieved. Therefore, OOC sequences have to be very long to support large numbers of users, and this introduces large bandwidth expansion [2].

The employment of error correction codes (ECCs) is possible in order to improve the performance (and hence to reduce the bandwidth expansion factor) of optical CDMA systems [5–7]. Zhang [5] proposes the use of asymmetric ECCs for optical CDMA with ON/OFF keying (OOK), whereas Kim and Poor [6] and Ohtsuki and Kahn [7] suggest using parallel concatenated convolutional codes (PCCC), namely, Turbo codes [8,9] in optical CDMA with pulse position modulation (PPM). This article presents the application of Turbo product codes (TPC) [17] for performance improvement of optical CDMA systems with either OOK or binary PPM.

TPCs have near-optimum performance [17] like Turbo codes. Furthermore, in terms of storage requirements and number of operations, TPCs have considerably less complexity if compared to Turbo codes. This makes TPCs very attractive for various applications that require high performance at high code and data rates. For example, the application of TPCs is considered in Ref. 26 for optical systems employing dense wavelength-division multiplexing (DWDM).

The organization of this article is as follows: Section 2 summarizes the TPC decoding algorithm for binary memoryless channels. Section 3 describes optical CDMA systems and their channel models. The application of TPCs in optical CDMA and performance curves obtained via simulations are given is Section 4 and Section 5, respectively. Finally, we conclude with some remarks in Section 6.

2. TURBO PRODUCT CODES FOR BINARY MEMORYLESS CHANNELS

In Sections 2.1–2.4 we give some background on product codes and in Section 4 describe their use in optical CDMA systems.

2.1. Product Code Construction

Product codes, or "iterated" codes, are in the class of linear block codes [21], which are widely used for forward error correction (FEC) in many of today's communications systems. Product codes were first introduced by Elias [18] in 1954, and their construction is basically an efficient way for building long block codes from two or more shorter block codes. The construction of a two-dimensional product code can be described as follows: Consider a $k_1 \times k_2$ array of information bits as shown in Fig. 1. First, the columns are encoded using a systematic (n_1, k_1, d_1) linear code C_1; then, the rows are encoded using a systematic (n_2, k_2, d_2) linear code C_2. Here, n_i, k_i, and $d_i (i = 1, 2)$ are the codeword length, the number of information bits, and the minimum Hamming distance, respectively. The linear codes C_1 and C_2 are called constituent or component codes and the resultant $(n_1 n_2, k_1 k_2, d_1 d_2)$ product code has a code rate of $R_c = (k_1 k_2)/(n_1 n_2)$. The idea of a product code

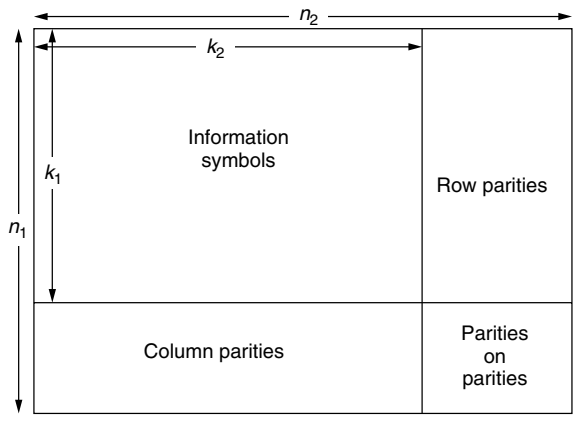


Figure 1. Product code construction.

can be further expanded to dimensions greater than 2. Nevertheless, to reduce encoding and decoding complexity, two-dimensional product codes are usually considered.

Product codes can be decoded by a hard-decision row decoder followed by a hard-decision column decoder (or vice versa). On reception of a noisy data vector, a hard-decision decoder strictly decides on bit 1 or 0 for each component of this vector and operates on these values. This is a low-complexity decoding method; nevertheless, for better performance, an iterative soft-input/soft-output (SISO) (i.e., soft-decision) decoding algorithm has to be used [17,19]. Different from the hard-decision decoder (which uses a threshold device), a soft-decision decoder uses the raw received analog signal to obtain reliability information on each bit position. The block diagram of an iterative SISO decoder for a product code is shown in Fig. 2.

On reception of the noisy data matrix R, the SISO row decoder calculates extrinsic (reliability) information matrix $W^{(1)}$ and passes this as a priori information to the SISO column decoder. The extrinsic information about a bit position is obtained via the help of all other bit positions, as described in the next section. After obtaining the information from the SISO row decoder, the SISO column decoder in turn calculates extrinsic information matrix $W^{(2)}$ and passes this back as a priori information to the row SISO decoder. Decoding continues in an iterative fashion until either estimate $R^{(1)}$ or $R^{(2)}$ is assigned as the decoded matrix. One efficient method of iterative SISO decoding is the Turbo product decoding algorithm proposed by Pyndiah [17] that is described and adapted for the binary memoryless channel model in the following section.

2.2. Extraction of Soft Information from a Binary Memoryless Channel

As we will observe later, optical CDMA systems can be modeled as binary memoryless channels [2]. Hence, there is a need to adapt a method for extracting soft information for these channels. That is, even though the channel is a "hard" channel that uses threshold detection, we can extract certain "soft" information given the error probability of the channel. It will take some time to explain this next. Let us assume that $X = x_0\ x_1\ \cdots\ x_{n-1}$ denotes the transmitted codeword and $R = r_0\ r_1\ \cdots\ r_{n-1}$ denotes the received vector, namely, a possibly noise corrupted version of X. The reliability for the jth component of R is given by the loglikelihood ratio (LLR):

$$\Lambda(r_j) = \log\left[\frac{\Pr(x_j = 1 \mid r_j)}{\Pr(x_j = 0 \mid r_j)}\right] \tag{1}$$

For $\Pr(x_j = 1) = \Pr(x_j = 0) = \frac{1}{2}$, Eq. (1) can be expressed using the transition probabilities of the binary memoryless channel and is found to be

$$\Lambda(r_j) = \log\left[\frac{\Pr(r_j \mid x_j = 1)}{\Pr(r_j \mid x_j = 0)}\right]. \tag{2}$$

The Turbo product decoding algorithm applies the Chase algorithm [20], a suboptimum maximum-likelihood decoding method for a linear (n, k, d) block code C, iteratively to the columns and rows of the product

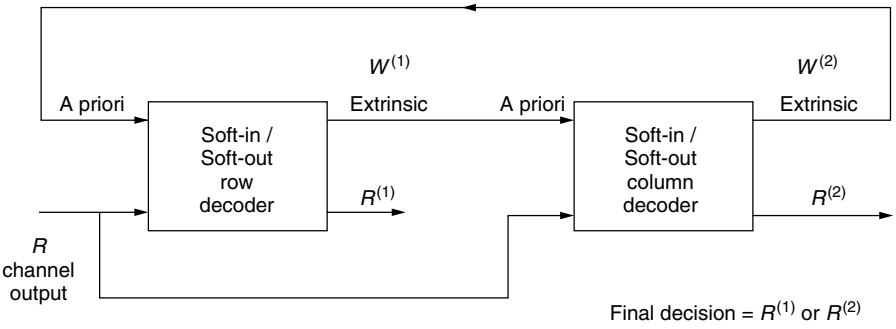


Figure 2. Iterative soft-input/soft-output decoder.

codeword. The Chase algorithm can be described as follows. First, the f least reliable bit positions of R are determined and 2^f test sequences are formed by perturbing these positions. Here, *perturbation* means using all possible combinations of zeros and ones in the least reliable bit positions. The parameter f is chosen as $f \ll k$. After perturbation of the least reliable bit positions, the 2^f test sequences are passed through an algebraic decoder for linear code C. The codewords obtained via the algebraic decoder are called *candidate codewords* and are denoted by $C^i (i = 1, 2, \ldots, 2^f)$. The distance between candidate codeword $C^i = c_0^i \; c_1^i \; \cdots \; c_{n-1}^i (c_j^i \in \{0, 1\})$ and received vector R is found by using the following metric, called "analog weight" [20]:

$$l(R, C^i) = \sum_{j=0}^{n-1} \left| (r_j \oplus c_j^i) \Lambda(r_j) \right| \tag{3}$$

where $\oplus$ denotes modulo-2 addition. This metric is calculated for all candidate codewords. The final output of the Chase decoder is the decoded codeword, which is the candidate codeword closest to received vector R. If C^D denotes the decoded codeword, then the following condition is used to determine C^D:

$$l(R, C^D) \leq l(R, C^i) \text{ for all } i \tag{4}$$

2.3. Computation of Extrinsic Information

Once a decision C^D is made, soft output (or extrinsic information) for each bit position c_j^D is needed to be passed to the next decoding stage. The reliability information of c_j^D is given by the LLR of the transmitted symbol x_j and can be written as

$$\Lambda(c_j^D) = \log \left[\frac{\Pr(x_j = 1 \mid R)}{\Pr(x_j = 0 \mid R)} \right] \tag{5}$$

Here, the computation of the LLR is different than the LLR in Eq. (1), since it takes into account that C^D is one of the 2^k codewords of linear code C. The probabilities in Eq. (5) can be expressed as

$$\Pr(x_j = v \mid R) = \sum_{C^i \in S_j^v} \Pr(X = C^i \mid R) \tag{6}$$

where S_j^v denotes the set of all candidate codewords with $v \in \{0, 1\}$ at their jth bit position. Applying Bayes' rule [21] and assuming that $\Pr(X = C^i)$ to be equal for all i, the LLR in (5) can be written as

$$\Lambda(c_j^D) = \log \left[\frac{\sum_{C^i \in S_j^1} \Pr(R \mid X = C^i)}{\sum_{C^i \in S_j^0} \Pr(R \mid X = C^i)} \right] \tag{7}$$

Assuming independent and identically distributed (i.i.d.) signal components r_j, $\Pr(R \mid X = C^i)$ can be expressed as

$$\Pr(R \mid X = C^i) = \prod_m \Pr(r_m \mid x_m = c_m^i) \tag{8}$$

Using Eq. (8) in Eq. (7), we obtain the following expression for the LLR:

$$\Lambda(c_j^D) = \Lambda(r_j) + w_j \tag{9}$$

where w_j is defined as the extrinsic information given by

$$w_j = \log \left[\frac{\sum_{C^i \in S_j^1} \prod_{m \neq j} \Pr(r_m \mid x_m = c_m^i)}{\sum_{C^i \in S_j^0} \prod_{m \neq j} \Pr(r_m \mid x_m = c_m^i)} \right] \tag{10}$$

As can be observed, the extrinsic information for bit position j is calculated via information obtained from all bit positions except bit position j itself.

The extrinsic information w_j, calculated as in Eq. (10), requires all candidate codewords. To reduce complexity, the following approach is used. Let $C^{\min(1),j} \in S_j^1$ be the codeword closest to R with a 1 in its jth bit position. Similarly, $C^{\min(0),j} \in S_j^0$ is the closest codeword to R with a 0 in its jth bit position. Using these closest (and hence dominant) codewords, the extrinsic information w_j can be approximated by

$$w_j \approx \log \left[\frac{\prod_{m \neq j} \Pr(r_m \mid x_m = c_m^{\min(1),j})}{\prod_{m \neq j} \Pr(r_m \mid x_m = c_m^{\min(0),j})} \right] \tag{11}$$

The calculation in this equation is not as accurate as the expression in Eq. (10). However, it has reduced complexity, since only two candidate codewords need to be considered to obtain the extrinsic information.

2.4. Turbo Decoding of Product Codes

Having defined the computation of the extrinsic information, we are now able to summarize the Turbo decoding algorithm for a binary memoryless channel:

1. The Chase algorithm is applied to the rows of the product codeword and decision $C^D = c_0^D\, c_1^D \cdots c_{n-1}^D$ is obtained for each row.
2. $C^{\min(1),j}$ and $C^{\min(0),j}$ are determined among the candidate codewords. In fact, one of these was already found in the previous step; thus, C^D is equal to either $C^{\min(1),j}$ or $C^{\min(0),j}$.
3. If both $C^{\min(1),j}$ and $C^{\min(0),j}$ exist among the candidate codewords, then the extrinsic information w_j is evaluated using Eq. (11). If one of these codewords cannot be found, then the extrinsic information is approximated as

$$w_j \approx \beta(2c_j^D - 1) \tag{12}$$

   where $\beta \geq 0$ is a predetermined reliability factor that increases with each iteration.
4. The extrinsic information w_j is scaled so that the average of $|w_j|$ is equal to one. The reason for this will be explained later.
5. The reliability information of r_j is updated as

$$\Lambda(r_j) = \Lambda_j + \alpha w_j \tag{13}$$

   where Λ_j is the reliability of the original received jth symbol and $\alpha \geq 0$ is a weight factor to combat high standard deviation in w_j and high BER during the first iterations.
6. The product codeword with the updated reliability information is the input to the next decoding stage. The above-mentioned procedure is applied this time to the columns of the product codeword. Decoding continues in an iterative fashion (i.e., row decoding $\rightarrow$ column decoding $\rightarrow$ row decoding $\rightarrow \ldots$) until user-defined termination.

One iteration of the preceding algorithm means row decoding (a half-iteration) followed by column decoding (another half-iteration); For instance, if we speak of four iterations, we actually mean eight half-iterations. The Turbo product decoder structure for a half-iteration is shown in Fig. 3, where the extrinsic information at the mth half-iteration is stored in matrix $W(m)$, whereas the reliability information of the original received data is stored in matrix Λ_R. For TPC decoding, usually six to eight half-iterations are sufficient to obtain good performance results.

The reason of scaling of the extrinsic information in step (4) in the preceding algorithm is to make the reliability and weight factors independent of the chosen product code. In fact, these factors change at each half-iteration and should be optimized for the employed product code and the channel characteristics of the communications system. If scaling is employed, one suggestion for these factors versus m (number of half-iteration), is as given in Table 1 [22].

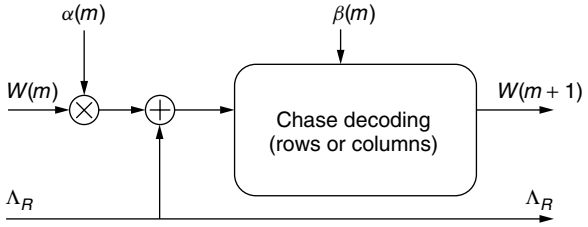


Figure 3. TPC decoder (half-iteration).

Table 1. Weight and Reliability Factors

m	0	1	2	3	4	5	6	7
α	0.0	0.5	0.7	0.9	1.0	1.0	1.0	1.0
β	0.2	0.3	0.5	0.7	0.9	1.0	1.0	1.0

3. OPTICAL CDMA SYSTEMS

In the following sections, we describe optical orthogonal codes and optical CDMA systems using different modulation types.

3.1. Optical Orthogonal Codes (OOCs)

In an optical CDMA system, a unique OOC [3] sequence is assigned to each user; thus, each user has a predetermined address sequence. To ensure orthogonality, OOC sequences have to satisfy the following two properties [3,10]:

1. Each sequence should be distinguished from a time-shifted version of itself (autocorrelation constraint),
2. Each sequence should be distinguished from a possibly time-shifted version of another sequence (cross-correlation constraint).

The parameters of an OOC are denoted by F, K, λ_a, and λ_c, which represent codelength (i.e., number of chips), code weight, autocorrelation constraint, and cross-correlation constraint, respectively. There exist various design techniques for OOCs [10–14]. OOCs with $\lambda_a = \lambda_c = 1$ have the best achievable correlation constraints for a direct-detection optical CDMA system. For example, three OOC sequences are shown in Fig. 4 [15].

Here, the OOC parameters are $F = 28, K = 3$, and $\lambda_a = \lambda_c = 1$, and the number of users is $N = 3$. The sequences are denoted as $S_u = s_{u,0}\; s_{u,1}\; s_{u,2} \cdots s_{u,n_{\text{chip}}-1}$, where $u = 1, 2, \ldots, N$, $s_{u,j} \in \{0, 1\}$ and n_{chip} denotes the number of chips (i.e., F, the number of time slots in one bit interval for OOK and half the number of time slots for binary PPM). It is assumed that the K light pulses of an OOC sequence have a normalized amplitude equal to 1 and since usually $K \ll n_{\text{chip}}$, it is more convenient to indicate the OOC sequences with the chip numbers at which pulses are placed. For the sequences shown in Fig. 4, the notation would be $S_1 = \{0, 2, 8\}$, $S_2 = \{0, 3, 10\}$, and $S_3 = \{0, 4, 13\}$.

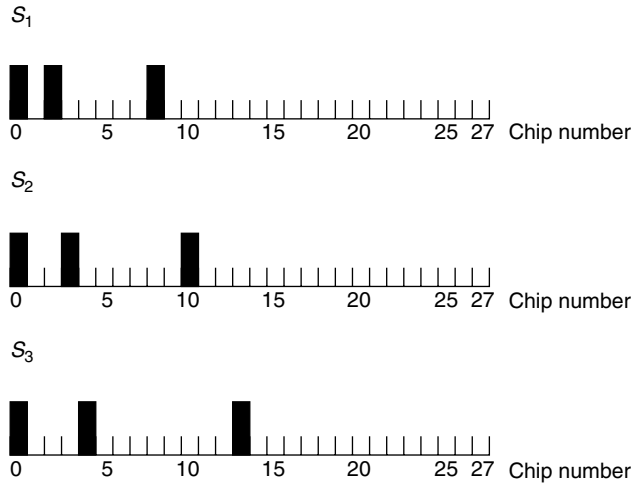


Figure 4. (28,3,1,1) optical orthogonal code sequences.

As stated before, OOCs need to satisfy certain correlation constraints to ensure orthogonality. These constraints can be expressed as

1. Autocorrelation property:

$$\Theta_{uu}(m) = \sum_{i=0}^{n_{\text{chip}}-1} s_{u,i} s_{u,(i+m \bmod n_{\text{chip}})} \le \lambda_a \tag{14}$$

for any sequence S_u and any integer $m \neq 0$. $\Theta_{uu}(0) = K$ since each sequence has K pulses.

2. Cross-correlation property:

$$\Theta_{uw}(m) = \sum_{i=0}^{n_{\text{chip}}-1} s_{u,i} s_{w,(i+m \bmod n_{\text{chip}})} \le \lambda_c \tag{15}$$

for any two sequences S_u and S_w ($S_u \neq S_w$) and any integer m.

For an $(F, K, \lambda_a = 1, \lambda_c = 1)$ OOC, it can be shown that the upper bound on the number of address sequences (i.e., number of users) is equal to [3,10]

$$N_{\max} = \left\lfloor \frac{F-1}{K(K-1)} \right\rfloor \tag{16}$$

where $\lfloor x \rfloor$ denotes the integer part of the real number x. Perfect optimal OOCs are OOCs with length

$$F = N_{\max} K(K-1) + 1 \tag{17}$$

3.2. System Model

Figure 5 shows the block diagram of an intensity modulated, direct-detection fiberoptic CDMA network that employs all-optical signal processing. The configuration shown is a star network with N users; however, other structures like ring networks are also possible.

A user in the system shown in Fig. 5 sends data using the address sequence of the destination user. Let us now briefly describe how this is established. At the transmitter side, the binary source is followed by a modulator. The modulation type is OOK or PPM. In case of OOK, the user transmits bit 1 by sending the address sequence, whereas bit 0 is transmitted by leaving the bit interval

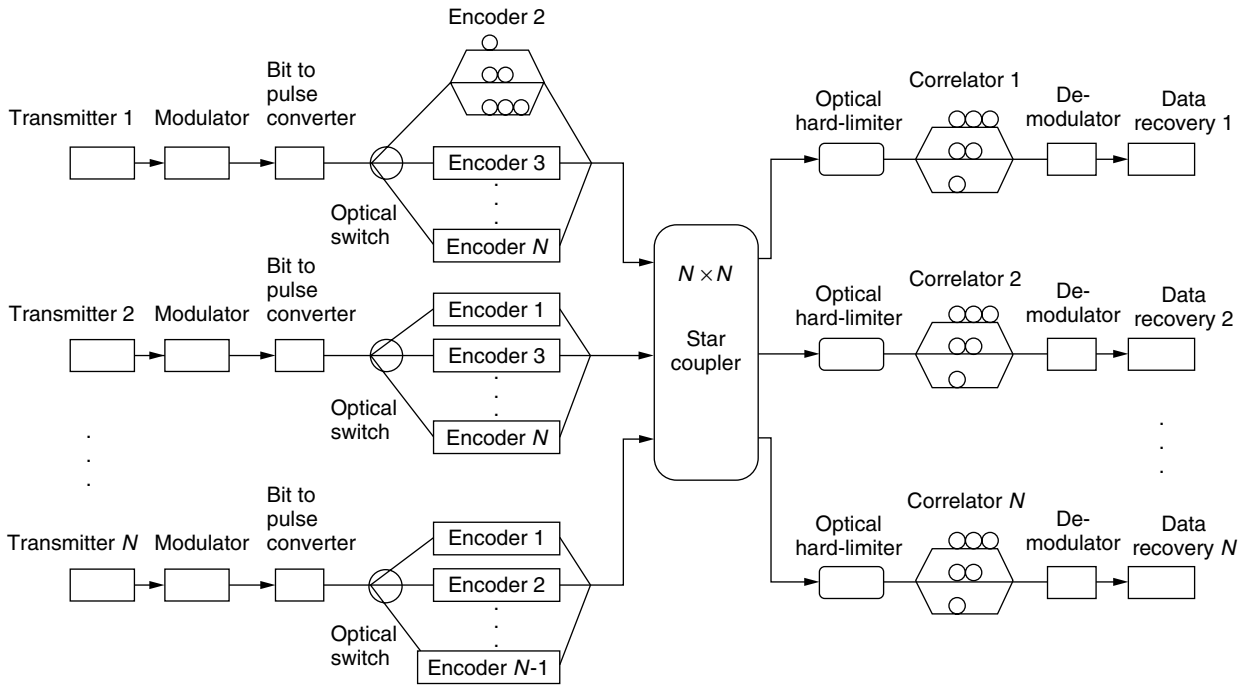


Figure 5. Optical CDMA system.

empty. The major drawbacks of OOK systems are that synchronization and baseline wander problems [1] may occur as a result of long sequences of zeros or ones. To avoid these, scramblers [23] or $m\mathbf{B}n\mathbf{B}$ codes [1] could be used. An alternative to OOK is PPM [6,7]. In case of binary PPM, the bit interval is divided into two time slots and binary 1 and 0 are sent as OOC sequences in their assigned time slots.

The modulator in Fig. 5 is followed by a bit-to-pulse converter to generate a short light pulse of duration T_c, which is the duration of one chip. If T_b denotes the bit duration, then for OOK, $T_c = T_b/F$, whereas for binary PPM, $T_c = T_b/(2F)$. The bandwidth expansion factor is equal to F and $2F$ for OOK and binary PPM, respectively. An optical switch is placed after the bit-to-pulse converter. The purpose of this switch is to select the OOC sequence of the destination user. After selecting the destination address, the OOC sequence is formed by passing the short light pulse through K fiber delay lines, which place the K pulses according to the OOC sequence of the destination user. All users are transmitting their data via an $N \times N$ optical star coupler. Hence, all users are receiving the same signals at overlapping times and wavelengths.

An optical CDMA receiver consists of the following devices: an optical hard-limiter [3], a correlator, and an OOK or PPM demodulator. The output of the hard-limiter is modeled as

$$h(\phi) = \begin{cases} 1 & \text{if} \quad \phi \geq 1 \\ 0 & \text{if} \quad \phi < 1 \end{cases} \tag{18}$$

where ϕ is the normalized input light intensity. Employment of the hard-limiter at the receiver is optional. However, it is shown that placing an optical hard-limiter before the optical correlator enhances the system performance significantly [3,4]. In fact, performance can be improved further by using two hard-limiters, one before and one after the correlator [25]. This article considers the employment of a single hard-limiter or no hard-limiter at all.

Like the encoder, the correlator is also a configuration composed of fiber delay lines that are matched to the destination user's address sequence. Each receiver receives the same sequences; however, because of the orthogonality of the sequences, the data can be decoded only by its intended correlator. The correlator are followed by a demodulator after which the data are finally recovered.

3.3. Channel Model for Optical CDMA with OOK

Without loss of generality, user 1 can be considered as the destination address. Also, the main source of noise can be assumed as optical multiple access interference caused by other users. For an OOK-CDMA system with hard-limiter, the correlator output of user 1 may be modeled as

$$Z_{u_1,\mathrm{HL}} = \begin{cases} K & \text{if} \quad x_j = 1 \\ I_{u_1,\mathrm{HL}} & \text{if} \quad x_j = 0 \end{cases} \tag{19}$$

and for an OOK system without hard-limiter

$$Z_{u_1,\mathrm{NHL}} = x_j K + I_{u_1,\mathrm{NHL}} \tag{20}$$

Here, $x_j \in \{0, 1\}$ represents the data sent to user 1, $I_{u_1,\mathrm{HL}}$ is the interference signal due to other users for a system with hard-limiter, $I_{u_1,\mathrm{NHL}}$ is the interference signal due to other users for a system without hard-limiter, and K is the weight of the OOC. In the correlator output signals, photodetector and amplifier noise may be ignored since these are relatively small if compared to optical multiple access interference.

In an OOK system, the decision criterion at the receiver is

$$r_j = \begin{cases} 0, & \text{if} \quad Z_{u_1} < \mathrm{Th} \\ 1, & \text{if} \quad Z_{u_1} \geq \mathrm{Th} \end{cases} \tag{21}$$

where $\mathrm{Th}(0 < \mathrm{Th} \leq K)$ is a predefined threshold level; thus, if the signal level is above Th, it is assumed that bit 1 was received; otherwise, the assumption is made that the received signal represents bit 0. The probability of bit error for an OOK system may be expressed as

$$\begin{aligned} P_{e,\mathrm{OOK}} = {} & \Pr(r_j = 1 \mid x_j = 0) \Pr(x_j = 0) \\ & + \Pr(r_j = 0 \mid x_j = 1) \Pr(x_j = 1) \end{aligned} \tag{22}$$

Since direct detection is used, we observe that for OOK, $\Pr(r_j = 0 \mid x_j = 1) = 0$. Hence, the OOK system is characterized by an asymmetric channel called Z channel [5] as shown in Fig. 6.

Under the assumption that $\Pr(x_j = 1) = \Pr(x_j = 0) = \frac{1}{2}$, we can write the conditional probability $\Pr(r_j \mid x_j)$ as

$$\Pr(r_j \mid x_j) = \begin{cases} 2P_{e,\mathrm{OOK}} & \text{if} \quad r_j = 1, x_j = 0 \\ 1 - 2P_{e,\mathrm{OOK}} & \text{if} \quad r_j = 0, x_j = 0 \\ 1 & \text{if} \quad r_j = 1, x_j = 1 \\ 0 & \text{if} \quad r_j = 0, x_j = 1 \end{cases} \tag{23}$$

It is shown that for an OOK system with hard-limiter, the probability of bit error can be upper-bounded by [3]

$$P_{e,\mathrm{OOK,HL}} \leq \frac{1}{2} \binom{K}{\mathrm{Th}} \prod_{i=1}^{\mathrm{Th}} (1 - v^{N-i}) \tag{24}$$

where N is the number of active users and

$$v = 1 - \frac{K}{2F} \tag{25}$$

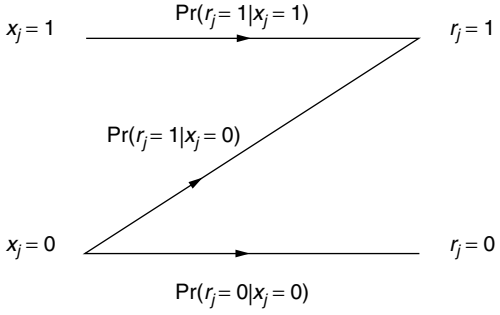


Figure 6. Z-channel model.

For an OOK system without hard-limiter, the probability of bit error has the following upperbound [3]:

$$P_{e,\mathrm{OOK,NHL}} \le \frac{1}{2}\sum_{i=\mathrm{Th}}^{N-1}\binom{N-1}{i}q^i(1-q)^{N-1-i} \tag{26}$$

where

$$q = \frac{K^2}{2F} \tag{27}$$

3.4. Channel Model for Optical CDMA with Binary PPM

As indicated earlier, for an optical CDMA system with binary PPM, one bit duration is split into two time slots. We assume that bit 0 is transmitted as an OOC sequence in slot Z_0 and that bit 1 is transmitted as an OOC sequence in slot Z_1. Let Z_{0,u_1} and Z_{1,u_1} denote the correlator outputs for the two time slots of user 1. At the receiver, the decision as to whether bit 1 or 0 is received is made by

$$r_j = \begin{cases} 0 & \text{if} \quad Z_{0,u_1} > Z_{1,u_1} \\ 1 & \text{if} \quad Z_{1,u_1} > Z_{0,u_1} \end{cases} \tag{28}$$

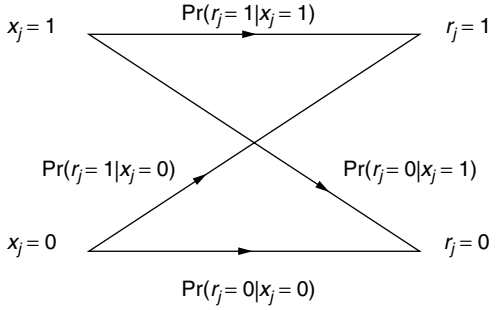
In case of equality, $Z_{1,u_1} = Z_{0,u_1}$, one bit is randomly chosen. PPM systems have the main advantage that there is no need to define a threshold Th as in the case of OOK. The drawback is that PPM requires more system bandwidth; binary PPM occupies twice as much bandwidth as does OOK.

For binary PPM, the probability of bit error, $P_{e,\mathrm{PPM}}$, is equal to

$$\begin{aligned} P_{e,\mathrm{PPM}} = {} & \Pr(Z_{0,u_1} \ge Z_{1,u_1} \mid x_j = 1)\Pr(x_j = 1) \\ & + \Pr(Z_{1,u_1} \ge Z_{0,u_1} \mid x_j = 0)\Pr(x_j = 0) \end{aligned} \tag{29}$$

Assuming $\Pr(x_j = 1) = \Pr(x_j = 0) = \frac{1}{2}$, an optical PPM-CDMA system can be characterized as a binary symmetric channel (BSC) as shown in Fig. 7, where, the crossover probabilities are the same. The transition probabilities can be expressed as

$$\Pr(r_j \mid x_j) = \begin{cases} P_{e,\mathrm{PPM}} & \text{if} \quad r_j = 1, x_j = 0 \\ 1 - P_{e,\mathrm{PPM}} & \text{if} \quad r_j = 0, x_j = 0 \\ 1 - P_{e,\mathrm{PPM}} & \text{if} \quad r_j = 1, x_j = 1 \\ P_{e,\mathrm{PPM}} & \text{if} \quad r_j = 0, x_j = 1 \end{cases} \tag{30}$$



$$\Pr(r_j = 1 | x_j = 0) = \Pr(r_j = 0 | x_j = 1)$$
$$\Pr(r_j = 1 | x_j = 1) = \Pr(r_j = 0 | x_j = 0)$$

Figure 7. Binary symmetric channel.

For an optical PPM system without hard-limiter, the correlator outputs for the two time slots for user 1 can be written as

$$Z_{0,u_1,\mathrm{NHL}} = (1 - x_j)K + I_{0,u_1,\mathrm{NHL}} \tag{31}$$

and

$$Z_{1,u_1,\mathrm{NHL}} = x_j K + I_{1,u_1,\mathrm{NHL}} \tag{32}$$

where $x_j \in \{0, 1\}$ represents the transmitted bit to user 1; $I_{0,u_1,\mathrm{NHL}}$ is the interference due to other users sending bit 0 and similarly, $I_{1,u_1,\mathrm{NHL}}$ is the interference due to other users sending bit 1. Using (29) and following the analysis in Ref. 16, the probability of bit error for an optical PPM-CDMA system without hard-limiter can be obtained as

$$\begin{aligned} P_{e,\mathrm{PPM,NHL}} = {} & \sum_{i=K}^{N-1}\sum_{m=0}^{N-1-i}\binom{N-1}{m}\binom{N-1}{m+i} \\ & \times q^{2m+i}(1-q)^{2(N-1-m)-i} \end{aligned} \tag{33}$$

If we consider an optical PPM-CDMA system with hard-limiter, the correlator outputs for the two time slots for user 1 can be written as

$$Z_{0,u_1,\mathrm{HL}} = \begin{cases} K & \text{if} \quad x_j = 0 \\ I_{0,u_1,\mathrm{HL}} & \text{if} \quad x_j = 1 \end{cases} \tag{34}$$

and

$$Z_{1,u_1,\mathrm{HL}} = \begin{cases} K & \text{if} \quad x_j = 1 \\ I_{1,u_1,\mathrm{HL}} & \text{if} \quad x_j = 0 \end{cases} \tag{35}$$

where $x_j \in \{0, 1\}$ represents the transmitted bit to user 1; $I_{0,u_1,\mathrm{HL}}$ is the interference due to other users sending bit 0 and similarly, $I_{1,u_1,\mathrm{HL}}$ is the interference due to other users sending bit 1. Applying (29) and assuming equally likely symbols x_j, we obtain the probability of bit error for an optical PPM-CDMA system with hard-limiter as [16]

$$P_{e,\mathrm{PPM,HL}} = \prod_{i=1}^{K}(1 - v^{N-i}) \tag{36}$$

Until now, we presented the probability of bit errors and transition probabilities for four possible cases of optical CDMA systems, namely, OOK with and without hard-limiter and binary PPM with and without hard-limiter. We observed that an OOK system is a binary asymmetric memoryless channel, whereas a binary PPM system is a binary symmetric memoryless channel. How to implement turbo product codes in these systems is described next.

4. APPLICATION OF TURBO PRODUCT CODES IN OPTICAL CDMA

To increase the performance (and hence, the number of users) of an optical CDMA system, a TPC encoder may be

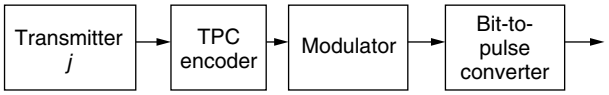


Figure 8. Optical CDMA transmitter with TPC encoder.

employed right before the modulator at the transmitter side in Fig. 8.

As shown, all data is encoded as a product code before it is modulated and passed to the bit-to-pulse converter. In the previous section, upper bounds were presented for the probability of bit errors and transition probabilities for the binary memoryless channels which characterize optical CDMA systems. These results are summarized in Tables 2 and 3. With this information, the implementation of the turbo product decoding algorithm described earlier is possible.

From our discussion on turbo product decoding, we know that initial reliability information is needed on received symbol r_j in order to apply the iterative decoding procedure. For an OOK system, we find the reliability information for r_j as

$$\Lambda(r_j = 1) = \log\left(\frac{1}{2P_{e,\mathrm{OOK}}}\right) \tag{37}$$

and

$$\Lambda(r_j = 0) = -\infty \tag{38}$$

where $P_{e,\mathrm{OOK}}$ is upper-bounded by (24) or (25) depending on whether a hard-limiter is used or not. The interpretation of (37) and (38) is that we are sure about a received 0, whereas a received 1 could be in error and has the reliability given in (37). Hence, when applying the Turbo product decoding algorithm in a direct detection OOK-CDMA system, only those bit positions that have initially (37) as their reliability information need to be considered when evaluating the extrinsic information of the received product codeword.

For the binary PPM system, the initial reliability information of r_j is obtained by

$$\Lambda(r_j) = \begin{cases} \log\left(\dfrac{1-P_{e,\mathrm{PPM}}}{P_{e,\mathrm{PPM}}}\right) & \text{if} \quad r_j = 1 \\ \log\left(\dfrac{P_{e,\mathrm{PPM}}}{1-P_{e,\mathrm{PPM}}}\right) & \text{if} \quad r_j = 0 \end{cases} \tag{39}$$

Since both symbols are equally likely, reliability and extrinsic information must be obtained for all bit positions when applying TPCs to optical PPM-CDMA systems.

We observe that in both OOK and PPM systems, the number of active users (i.e., N) is required to estimate the probability of bit error P_e. This information could be retrieved using an estimator placed at the front end at the receiving side (see Fig. 9).

Depending on the total light power per bit period, this device can estimate the number of active users and pass this information to the TPC decoder, which is then capable of calculating an upper bound of P_e depending on the modulation type and whether an optical hard-limiter is employed. The TPC decoder could also perform a lookup

Table 2. Probability of Bit Error for Optical CDMA Systems

P_e	Upperbound
$P_{e,\mathrm{OOK,HL}}$	$\frac{1}{2}\binom{K}{Th}\prod_{i=1}^{\mathrm{Th}}(1-v^{N-i})$
$P_{e,\mathrm{OOK,NHL}}$	$\frac{1}{2}\sum_{i=Th}^{N-1}\binom{N-1}{i}q^i(1-q)^{N-1-i}$
$P_{e,\mathrm{PPM,HL}}$	$\prod_{i=1}^{K}(1-v^{N-i})$
$P_{e,\mathrm{PPM,NHL}}$	$\sum_{i=K}^{N-1}\sum_{m=0}^{N-1-i}\binom{N-1}{m}\binom{N-1}{m+i}q^{2m+i}(1-q)^{2(N-1-m)-i}$

Table 3. Transition Probabilities

System	$\Pr(r_j=0 \mid x_j=0)$	$\Pr(r_j=1 \mid x_j=0)$	$\Pr(r_j=0 \mid x_j=1)$	$\Pr(r_j=1 \mid x_j=1)$
OOK	$1-2P_{e,\mathrm{OOK}}$	$2P_{e,\mathrm{OOK}}$	0	1
PPM	$1-P_{e,\mathrm{PPM}}$	$P_{e,\mathrm{PPM}}$	$P_{e,\mathrm{PPM}}$	$1-P_{e,\mathrm{PPM}}$

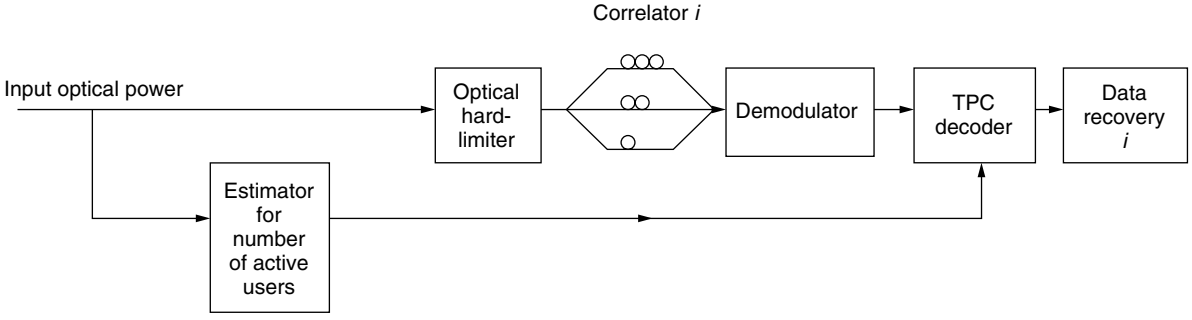


Figure 9. Optical CDMA receiver with TPC decoder.

if a table for P_e versus N is stored beforehand. Having an estimate of P_e enables the TPC decoder to perform iterative decoding via calculation of extrinsic and initial reliability information as described before.

5. PERFORMANCE RESULTS

The following simulations were carried out in another study by the authors [16] to demonstrate the performance improvement due to the employment of TPCs in optical CDMA systems. It is assumed that optical multiple access interference is the main source of noise. Other noise sources, such as photodetector and thermal noise and losses related to optical transmission, are ignored. Transmission of all OOC sequences is modeled to be chip-synchronous, which is a worst-case assumption [3]. Furthermore, it is assumed that perfect optimal OOCs are employed.

The first set of simulations presented here is carried out for an optical OOK-CDMA system with hard-limiter. Assuming that optical multiple access interference is the main source of interference, the threshold at the receiver side is set to $\mathrm{Th} = K - 1$. Figure 10 shows the bit error rate (BER) versus number of active users for various coded and uncoded systems.

The BER curves for the uncoded systems with OOCs of weight $K = 4$, $K = 5$, and $K = 6$ are plotted using the upperbound for $P_{e,\mathrm{OOK,HL}}$. It is observed that a $K = 4$ system with hard decision and a BCH(64,57,4) linear block code or a Reed–Solomon RS(255,239,17) linear block code shows minor performance improvement if compared to the uncoded $K = 4$ system. On the other hand, a $K = 4$ system with a TPC BCH(64,57,4)^2 code (i.e., both component codes of the product code are BCH(64,57,4) codes) outperforms the previously mentioned systems, the uncoded $K = 5$ system, and even the uncoded $K = 6$ system. The number of half-iterations for the TPC is 6, and the f parameter for the Chase algorithm is chosen to be 4 (i.e., 16 test patterns are formed for each row or column of the product codeword). The TPC decoder uses the weight and reliability factors given in Table 1. Similar results as those observed in Fig. 10 are obtained for an optical OOK-CDMA system with no hard-limiter and the same conditions as above.

The performance curves for an optical PPM-CDMA system with no hard-limiter is shown in Fig. 11. Again, the TPC BCH(64,57,4)^2 coded system outperforms all other coded and uncoded systems. It has been shown [16] that similar performance improvements may be achieved for optical PPM-CDMA systems with hard-limiter.

All simulation results show that for a given BER, it is possible to implement optical CDMA systems with low-weight OOCs (i.e., low K) if a TPC is employed. The performance curves show that we might switch from an uncoded $K = 6$ system to a TPC coded $K = 4$ system and obtain an improved BER. The TPC discussed here has an coding overhead of 25%; nevertheless, if we compare the $K = 4$ system with the $K = 6$ systems (the ratio of OOC codelengths vs. number of users as shown in Fig. 12), we observe that the $K = 6$ system requires OOCs with about 2.5 times more codelength.

Considering the coding overhead, the net reduction in bandwidth expansion factor is $2\times$ for the system under consideration. Hence, for a given BER, the employment of a TPC reduces the required OOC codelength significantly and enables an increase in available bandwidth. Furthermore, having a smaller K value means fewer fiber delay lines in the encoders and correlators. A reduction in K provides also a reduction in the amount of power to be transmitted because the

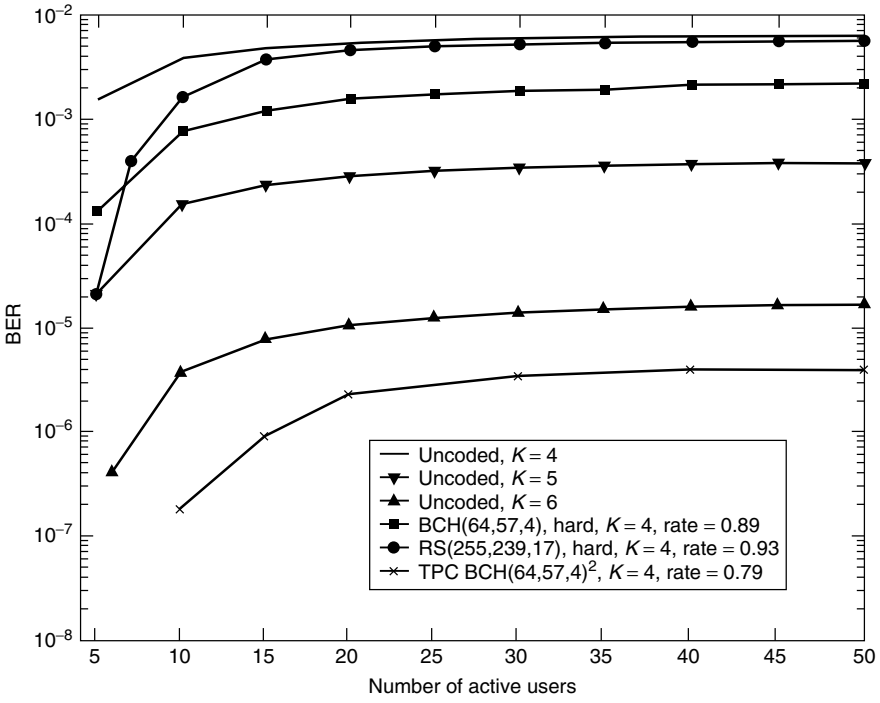


Figure 10. Performance of optical OOK-CDMA system with hard-limiter.

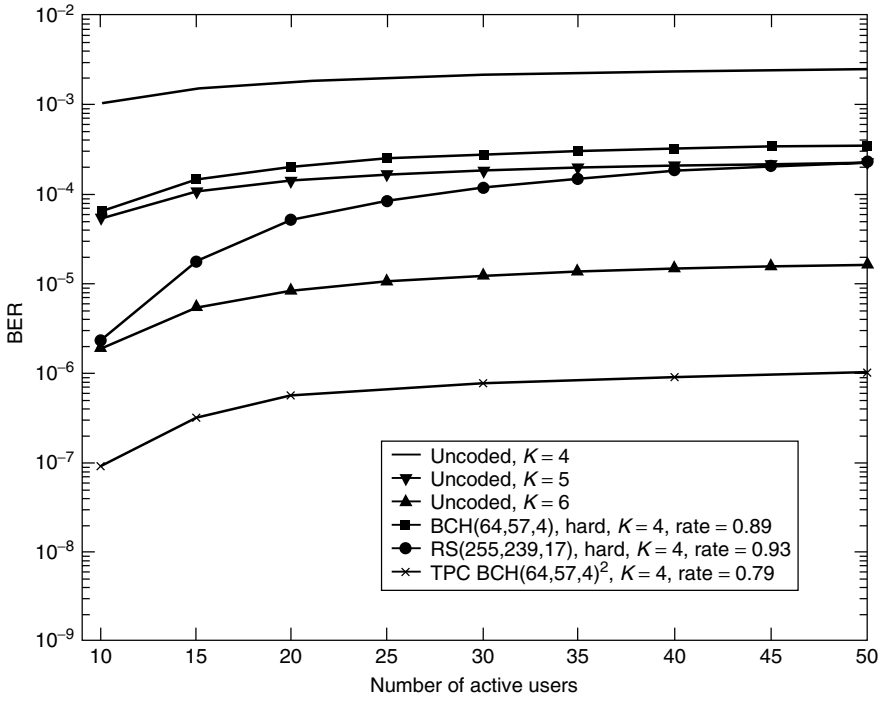


Figure 11. Performance of optical PPM-CDMA system without hard-limiter.

attenuation of the light pulses decreases when they are split into fewer fiber delay lines. Using a TPC BCH(64,57,4)^2 code requires about 1 dB more power to be transmitted per bit. However, switching from $K = 6$ to $K = 4$ provides a gain of 1.76 dB. As a result, for the TPC coded system, the net power gain would be 0.76 dB.

6. CONCLUSION

The wide deployment of optical CDMA systems was not possible in the past because of the huge bandwidth expansion imposed by very long optical orthogonal code (OOC) sequences. However, it appears that using error correction codes might be an efficient way to reduce bandwidth expansion for a given BER target, enabling wider application possibilities of optical CDMA systems.

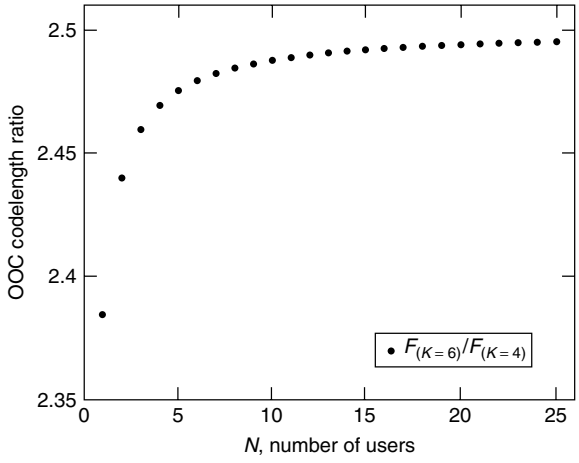


Figure 12. OOC codelength ratio $F_{(K=6)}/F_{(K=4)}$.

In this article, the application of Turbo product codes (TPC) was considered for an intensity modulated, direct-detection optical CDMA system with OOK or binary PPM. It was shown that TPC-coded optical CDMA systems have the potential to outperform their uncoded counterparts or systems with simple one-dimensional error correction codes. Furthermore, it was observed that the usage of OOC sequences with less weight and length is possible via TPCs. TPCs discussed here use only hard (thresholded) decisions and estimates of the number of active users in the system, which make them much easier to fit into existing networks.

Beside optical CDMA systems, other interesting approaches would be the application of TPCs in optical systems with dense wavelength division multiplexing (DWDM) [26]. Also, the application in hybrid WDMA/CDMA or TDMA/CDMA systems [24] might be considered.

BIOGRAPHIES

Steven W. McLaughlin received his B.S.E.E. degree in 1985 from Northeastern University, Boston, Massachusetts, his M.SE. degree in 1986 from Princeton University, New Jersey, and a Ph.D. degree in 1992 from the University of Michigan. He joined the Electrical Engineering Department at the Rochester Institute of Technology, New York, in 1992. In 1996, he joined the School of ECE at the Georgia Institute of Technology, Atlanta, where he is now associate professor. He holds 21 U.S. patents and has more than a dozen pending in the areas of coding and signal processing for data storage and fiber optic transmission systems. In 1997, he received the Presidential Early Career Award for Scientists and Engineers PECASE and was cited by President Clinton

for "for leadership in the development of high-capacity, nonbinary optical recording formats." He also received the National Science Foundation CAREER Award in 1997. His areas of interest are communication and information theory with specific interest in coding for data storage and fiber optic transmission systems.

Cenk Argon received his B.S.E.E. degree in 1992 and his M.S.E.E. degree in 1994 from the Middle East Technical University, Ankara, Turkey, and his M.S. degree in 1999 from the Georgia Institute of Technology, Atlanta. He is currently pursuing his Ph.D. degree and working as a graduate research assistant in the Communications and Information Theory Research Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology. His areas of interest are forward error correction (FEC) and signal processing techniques for optical networks, turbo product codes, and optical CDMA systems.

BIBLIOGRAPHY

1. G. Keiser, *Optical Fiber Communications*, 3rd ed., McGraw-Hill, 2000.
2. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, 1995.
3. J. A. Salehi, Code division multiple access techniques in optical fiber networks — Part I: Fundamental principles, *IEEE Trans. Commun.* **37**: 824–833 (Aug. 1989).
4. J. A. Salehi and C. A. Brackett, Code division multiple access techniques in optical fiber networks — Part II: Systems performance analysis, *IEEE Trans. Commun.* **37**: 834–842 (Aug. 1989).
5. J.-G. Zhang, Use of error-correction codes to improve the performance of optical fiber code division multiple access systems, *Proc. URSI Int. Symp. Signals, Systems, and Electronics*, 1998, pp. 361–366.
6. J. Y. Kim and H. V. Poor, Turbo-coded optical direct-detection CDMA system with PPM modulation, *IEEE J. Lightwave Technol.* **19**(3): 312–323 (March 2001).
7. T. Ohtsuki and J. M. Kahn, BER performance of turbo-coded PPM CDMA systems on optical fiber, *IEEE J. Lightwave Technol.* **18**(12): 1776–1784 (Dec. 2000).
8. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. IEEE ICC*, 1993, pp. 1064–1070.
9. C. Berrou and A. Glavieux, Near optimum error correcting coding and decoding: Turbo-codes, *IEEE Trans. Commun.* **44**(10): 1261–1271 (Oct. 1996).
10. F. R. K. Chung, J. A. Salehi, and V. K. Wei, Optical orthogonal codes: Design, analysis, and applications, *IEEE Trans. Inform. Theory* **35**: 595–604 (May 1989).
11. A. A. Shaar and P. A. Davies, Prime sequences: Quasi-optimal sequences for OR channel code division multiplexing, *Electron. Lett.* **19**(21): 888–890 (Oct. 1983).
12. S. V. Maric, Z. I. Kostic, and E. L. Titlebaum, A new family of optical code sequences for use in spread spectrum fiber-optic local area networks, *IEEE Trans. Commun.* **41**(8): 1217–1221 (Aug. 1993).
13. H. Chung and P. V. Kumar, Optical orthogonal codes-New bounds and an optimal construction, *IEEE Trans. Inform. Theory* **36**(4): 866–873 (July 1990).
14. C. Argon and R. Ergül, Optical CDMA via shortened optical orthogonal codes based on extended sets, *Opt. Commun.* **116**: 326–330 (May 1995).
15. C. Argon and R. Ergül, Detection of shortened OOC codewords in optical CDMA systems with double hard-limiters, *Opt. Commun.* **177**: 277–281 (April 2000).
16. C. Argon and S. W. McLaughlin, Optical OOK-CDMA and PPM-CDMA systems with turbo product codes, *IEEE J. Lightwave Technol.* (in press).
17. R. Pyndiah, Near-optimum decoding of product codes: Block turbo codes, *IEEE Trans. Commun.* **46**(8): 1003–1010 (Aug. 1998).
18. P. Elias, Error-free coding, *IRE Trans. Inform. Theory* **IT-4**: 29–37 (Sept. 1954).
19. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. Inform. Theory* **42**: 429–445 (March 1996).
20. D. Chase, A class of algorithms for decoding block codes with channel measurement information, *IEEE Trans. Inform. Theory* **IT-18**(1): 170–182 (Jan. 1972).
21. S. B. Wicker, *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, 1995.
22. A. Picart and R. Pyndiah, Adapted iterative decoding of product codes, *Proc. IEEE Globecom*, 1999, pp. 2357–2362.
23. B. P. Lathi, *Modern Digital and Analog Communication Systems*, 2nd ed., Holt, Rinehart and Winston, 1989.
24. W. Huang, M. H. M. Nizam, I. Andonovic, and M. Tur, Coherent optical CDMA (OCDMA) systems used for high-capacity optical fiber networks — system description, OTDMA comparison and OCDMA/WDMA networking, *IEEE J. Lightwave Technol.* **18**: 765–778 (June 2000).
25. T. Ohtsuki, Performance analysis of direct-detection optical asynchronous CDMA systems with double optical hard-limiters, *IEEE J. Lightwave Technol.* **15**(3): 452–457 (March 1997).
26. O. Aitsab and V. Lemaire, Block turbo code performances for long-haul DWDM optical transmission systems, *Proc. OFC*, 2000, pp. 280–282.

TURBO TRELLIS-CODED MODULATION (TTCM) EMPLOYING PARITY BIT PUNCTURING AND PARALLEL CONCATENATION

PATRICK ROBERTSON
Institute for Communications Technology
German Aerospace Center (DLR)
Wessling, Germany

THOMAS WÖRZ
Audens ACT Consulting GmbH
Wessling, Germany

1. INTRODUCTION

In 1993, powerful Turbo codes were introduced [1] that achieve good bit error rates ($10^{-3} \cdots 10^{-5}$) at very low SNR close to the channel capacity. They are of interest in a wide

range of digital telecommunications applications such as satellite communications and mobile radio. For instance, they are deployed in the air interface of the UMTS mobile radio system [2]. They comprise two binary component codes and an interleaver and were originally proposed for binary modulation schemes [e.g., BPSK (binary phase shift keying)]. Successful attempts were soon undertaken to combine binary Turbo codes with higher-order modulation [e.g., 8-PSK, 16-QAM (quadrature amplitude modulation)] using Gray mapping [3], and alternatively as component codes within multilevel codes [4]. In contrast, in the approach called Turbo trellis-coded modulation (TTCM) one employs two Ungerboeck type codes [5] in combination with trellis-coded modulation (TCM) in their recursive systematic form as component codes in an overall structure rather similar to binary Turbo codes [6].

Before the combination of Turbo codes and TCM is illustrated, the main idea behind conventional TCM codes is briefly reviewed. Generally, the bit error rate (BER) of an uncoded modulation scheme, at least for medium to large signal-to-noise ratio E_s/N_0, depends on the minimum squared Euclidean distance between all members of the considered signal set (e.g., QPSK: $d^2_{\text{free}} = 4$). Increasing d^2_{free} can be achieved by standard channel coding, which adds coded bits to the information bits. As a result, more bandwidth is required to transmit the information bits including the redundancy with the same signal set. However, in the case of TCM the signal set is basically extended such that the same bandwidth is occupied compared to the uncoded case, for example, from QPSK to 8-PSK. Then, a rate-$\frac{2}{3}$ convolutional code is applied to generate sequences of 8-PSK symbols, whose d^2_{free} is larger than for the uncoded QPSK transmission. The design of the convolutional code is based on the set partitioning of the underlying signal set aiming at the optimization of d^2_{free}. Summarizing, one transmits the same number of information bits per modulation symbol with increased d^2_{free} and without increasing the necessary bandwidth. The principle is applicable in the same way to higher-order modulation schemes such as 16- and 64-QAM. Turbo trellis-coded modulation codes can be decoded with the Viterbi or the Bahl–Jelinek (symbol-by-symbol MAP) algorithm [7]. Multidimensional TCM allows even higher bandwidth efficiency than traditional Ungerboeck TCM by assigning more than one symbol per trellis transition or step [8]. In this case, the set partitioning takes into account the union of more than one two-dimensional signal set.

The basic principle of Turbo codes is applied to TCM by retaining the important properties and advantages of both their structures. Essentially, TCM codes can be seen as systematic feedback convolutional codes followed by one (or more for multidimensional codes [8]) signal mapper(s). Just as binary Turbo codes use a parallel concatenation of two binary recursive convolutional encoders, in TTCM one concatenates two recursive TCM encoders, and adapts the interleaving and puncturing. Naturally, this has consequences at the decoding side, which are explained in depth later.

One can also apply the concept of TTCM to incorporate multidimensional component codes, which allows a higher overall bandwidth efficiency for a given signal constellations than ordinary TTCM. By applying the technique to 8-PSK, 16-QAM, and 64-QAM modulation formats, we will show its viability over a large range of bandwidth efficiency and signal-to-noise ratios. In all cases, low BERs ($10^{-4} \cdots 10^{-5}$) can be achieved within 1 dB or less from Shannon's limit — a finding that in the context of binary Turbo codes was responsible for the interest they generated.

The article begins by describing the generic encoder (beginning with a motivation for its structure); an encoder with 8-PSK signaling will serve as a salient example. We then present the results of a search for component codes, taking into consideration the puncturing at the encoder. This is followed by a section on the iterative decoder using symbol-by-symbol MAP component decoders whose structures are derived for our case of nonbinary trellises and special metric calculation. Finally, we present simulation results of the TTCM scheme with two- and four-dimensional 8-PSK, as well as two-dimensional 16-QAM and 64-QAM. The influence of varying the block size and interleaver type — both of important practical relevance — is also subject of investigation. We conclude by presenting further literature to the subject and current areas of investigation.

2. THE ENCODER

2.1. Motivation for the Structure

Let us recall that two important characteristics of Turbo codes are their simple use of recursive systematic component codes in a parallel concatenation scheme. Pseudorandom bitwise interleaving between encoders ensures a small bit error probability [9]. What is crucial to their practical suitability is the fact that they can be decoded iteratively with good performance [1]. It is well known that Ungerboeck codes combine coding and modulation by optimizing the Euclidean distance between codewords and achieve high spectral efficiency (m bits per 2^{m+1}-ary symbol from the two-dimensional signal space) through signal set expansion. The encoder can be represented as combination of a systematic recursive convolutional encoder and symbol mapper. If $\tilde{m}$ out of m bits are encoded, the resulting trellis diagram consists of $2^{\tilde{m}}$ branches per state, not counting parallel transitions. This results in more than two branches per state for $\tilde{m} > 1$ — we call this a *nonbinary trellis*.

We have employed Ungerboeck codes (and multidimensional TCM codes) as building blocks in a Turbo coding scheme in a similar way as binary codes were used through so-called parallel concatenation [1]. The major differences are (1) the interleaving now operates on short groups of m bits (e.g., pairs for 8-PSK with two-dimensional TCM schemes) instead of single bits; (2) to achieve the desired spectral efficiency, puncturing the parity information is not quite as straightforward as in the binary Turbo coding case; and (3) there are special constraints on both the component encoders as well as the structure of the interleaver.

2.2. Definition of the Encoder

In this section we begin by defining the generic encoder for TTCM and then continue to illustrate a simple

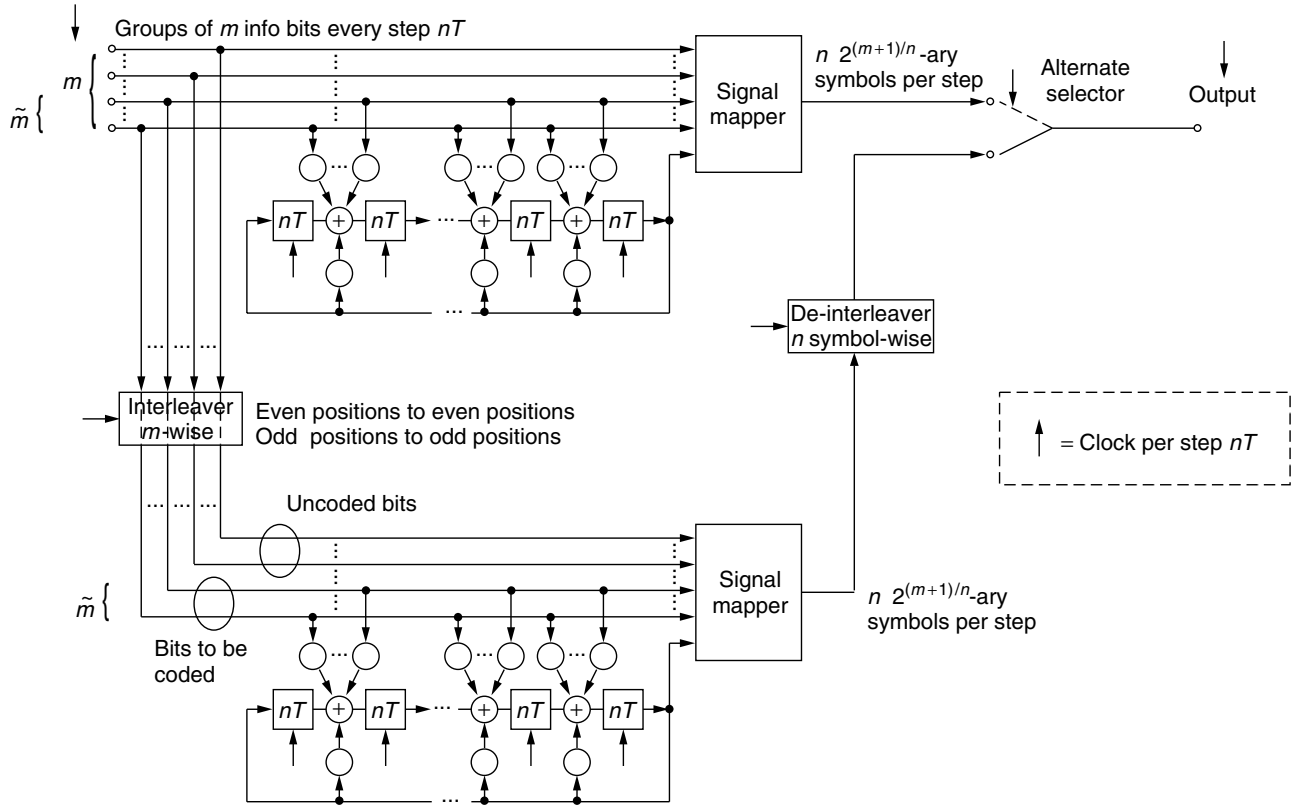


Figure 1. The generic encoder that treats uncoded bits as coded bits from a structural point of view.

example encoder. Figure 1 shows this generic encoder, comprising two TCM encoders linked by the interleaver. It is important to remember that the interleaver operates on small groups of m bits. Let the size of the interleaver — the number of these groups — be N. The number of modulated symbols per block is $N \cdot n$, with $n = D/2$, where D is the signal set dimensionality. The number of information bits transmitted per block is $N \cdot m$. The encoder is clocked in steps of $n \cdot T$ where T is the symbol duration of each transmitted $2^{(m+1)/n}$ -ary symbol. In each step, m information bits are input and n symbols are transmitted, yielding a spectral efficiency of m/n bits per symbol usage. A signal mapper follows each recursive systematic convolutional encoder where the latter each produce one parity bit in addition to retaining the m information bits at their inputs. For clarity we have not depicted any special treatment of the $m - \tilde{m}$ uncoded bits as opposed to the $\tilde{m}$ bits to be encoded: in practice, uncoded bits would not need to be passed through the interleaver but would simply be used to choose the final signal point from a subset of points after the selector. We will return to the problem of parallel transitions shortly.

For the moment, the interleaver is restricted to keeping each group of m bits unchanged within itself (as visualized by the dashed lines passing through the interleaver in Fig. 1). The output of the lower encoder/mapper is deinterleaved according to the inverse operation of the interleaver. This ensures that at the input of the selector, the m information bits partly defining each group of n symbols of both the upper and lower input are identical. Therefore, if the selector is switched such that a group of n symbols is chosen alternately from the upper and lower inputs, then the sequence of $N \cdot n$ symbols at the output has the important property that each of the N groups of m information bits defines part of each group of n output symbols. The remaining bit, which is needed to define each group of n symbols, is the parity bit taken alternatively from the upper and lower encoders.

A simple example will now serve to clarify the operation of the encoder for the case $n = 1$, $m = 2$, $N = 6$ and 8-PSK signaling: it is illustrated in Fig. 2. The set partitioning is shown in Fig. 3. The 6-long sequence $(d_1, d_2, \ldots, d_6) = (\mathbf{00,01,11,10,00,11})$ of information bit pairs ($m = 2$) is encoded in an Ungerboeck style encoder to yield the 8-PSK sequence (**0,2,7,5,1,6**). The information bits are interleaved — on a pairwise basis- and encoded again into the sequence (6,7,0,3,0,4). We deinterleave the second encoder's output symbols to ensure that the ordering of the two information bits partly defining each symbol corresponds to that of the 1st encoder; thus we now have the sequence (0,3,6,4,0,7). Finally, we transmit the 1st symbol of the first encoder, the second symbol of the second encoder, the third of the first encoder, the fourth symbol of the second encoder, and so on: (**0**,3,**7**,4,**1**,7). Thus the parity bit is alternately chosen from encoders 1 and 2 (bold, non-bold, bold ...). Also, the kth information bit pair exactly determines 2 of the 3 bits of the kth symbol x_k. This ensures that each information bit pair defines part of the constellation of an 8-PSK symbol exactly once.

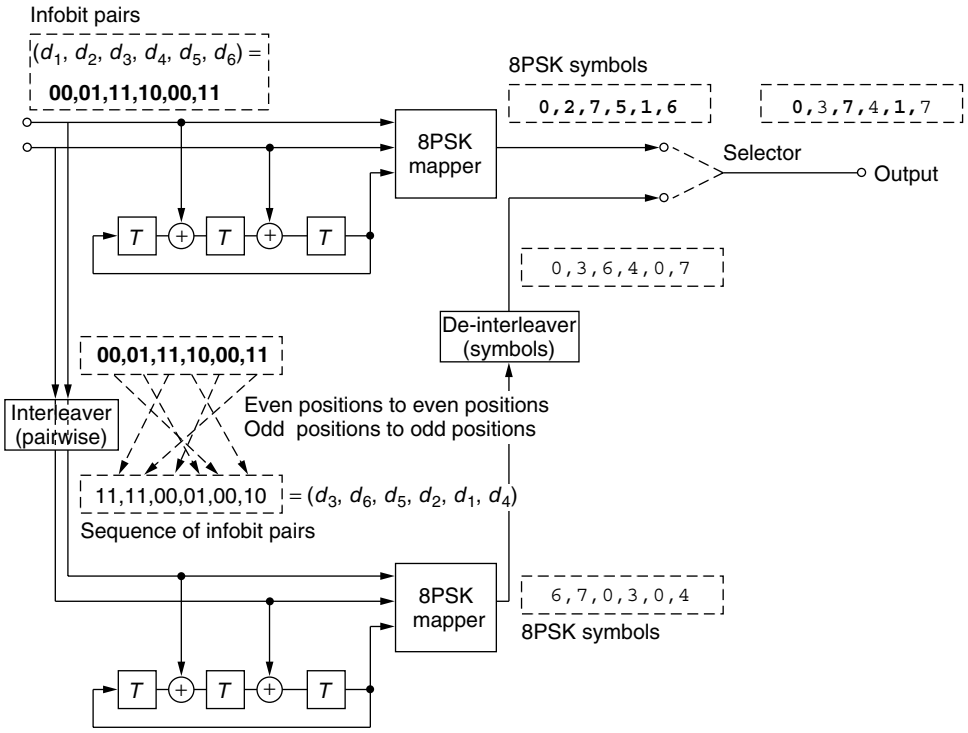


Figure 2. The encoder shown for 8-PSK with two-dimensional component codes memory 3. An example of interleaving with $N = 6$ is shown. Bold letters indicate that symbols or pairs of bits correspond to the upper encoder.

2.3. Interleaver and Code Constraints

2.3.1. Basic Interleaver Types. By deinterleaving the output of the second encoder, each symbol index k before the selector in Fig. 1 has the property of being associated with input information bit group index k, regardless of the actual interleaving rule. However, to ensure that punctured and unpunctured symbols are uniformly spread, that is, occur alternately, at the input of both decoders, the interleaver must map even positions to even positions and odd ones to odd ones (or even–odd, odd–even). Other than this constraint, the interleaver can be chosen to be pseudorandom or modified to avoid low distance error events. It is important to remember that we have so far assumed that the interleaver keeps the input unchanged within each group of information bits and that the corresponding symbol deinterleaver does not modify its symbol inputs (except for the actual reordering of their positions, of course); see the top example in Fig. 4.

We have also forced a constraint on the component code such that the corresponding trellis diagram of the convolutional encoders should have no parallel transitions. This ensures that each information bit benefits from the parallel concatenation and interleaving. This condition can be relaxed in a number of cases. The first [10] applies if the interleaver no longer keeps each group of m bits unchanged during interleaving: This method alters the position of the bits within each group as they are interleaved, see the bottom example in Fig. 4. The argument is that each information bit should influence the state of at least one encoder; bits that lead to only parallel transitions in *one* encoder will thus cause the *other* encoder to change its state and accrue distance henceforth. All bits would thus benefit from the interleaving and parallel concatenation. In one study [10] a slight performance gain was reported for the first few iterations compared to TTCM schemes with no parallel transitions.

The second case in which we allow parallel transitions in the component code is when we desire a very high bandwidth efficiency. Because of the higher operating SNR and the large Euclidean distance that separates the subsets of signal points that define parallel transitions (e.g., the lowest partition step in Fig. 3), corresponding uncoded information bits receive ample protection in cases such as 8-PSK transmitting 2.5 information bits per symbol and 64-QAM with 5 bits per symbol, which we will investigate below. The transmission of uncoded bits has been proposed for the multilevel approach [4], where channel capacity arguments show that when 5 information bits are sent using one 64-QAM symbol, the last 2 partitioning bits theoretically need only minimal (if any) coding protection.

2.3.2. Design Rule for Selecting the Number of Uncoded Bits. In the following, a heuristic rule is given in order to determine the number of uncoded bits per symbol. It is based on the experience that the BER of TTCM schemes (with large block lengths) reaches a value of $P_b \approx 10^{-5}$ at a signal-to-noise ratio E_s/N_0^*, which is approximately 1 dB above the corresponding channel capacity. Let us consider the sequence of increasing inner-set distances Δ_i when following down the partitioning of the corresponding signal set (for an example of partitioning an 8-PSK constellation,

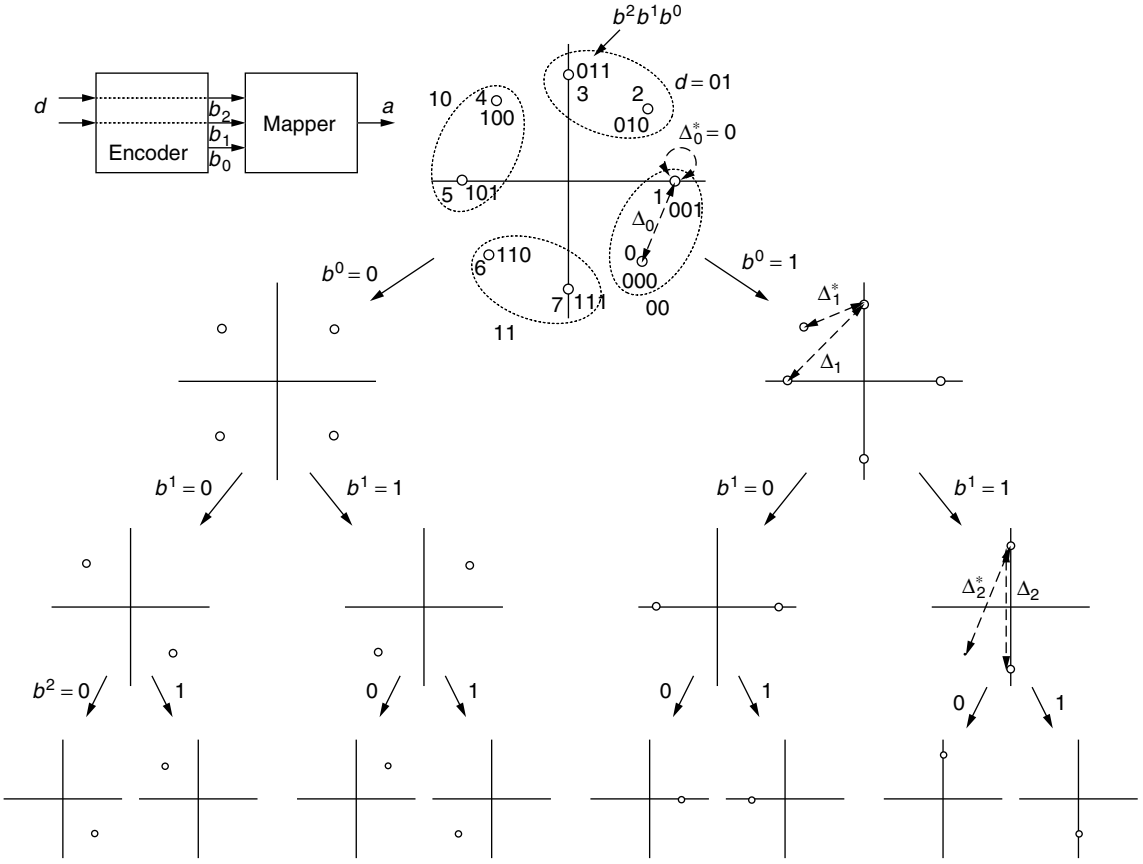


Figure 3. The set partitioning for 8-PSK. Dotted ovals denote subsets corresponding to the different combinations of d. The distances Δ_i are relevant for code design.

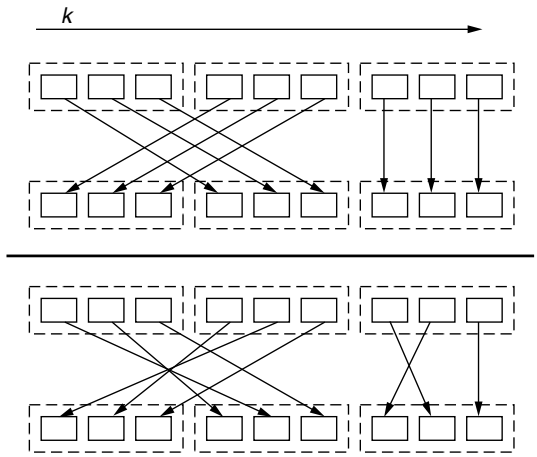


Figure 4. Two kinds of interleaver for Turbo TCM, block length $N = 3$. *Top*: — position invariant group interleaving; *bottom*: — position swapping group interleaving.

refer to Fig. 3). For each distance we can evaluate a rough approximation of the BER in the uncoded case, by applying the well-known formula [11]

$$P_b(\Delta_i) = \frac{1}{2}\,\mathrm{erfc}\sqrt{\frac{E_s\Delta_i}{4N_0}} \tag{1}$$

By using this formula to approximate the BER of the uncoded bits with $P_b(\Delta_i)$, two approximations are included:

- The error propagation from the partition levels that include coded bits into the partition levels with uncoded bits is neglected.
- Moreover, the number of nearest neighbors is not included in the calculation, only the pure distance is used to evaluate (1).

As a result, we can identify at which level of the partition chain the corresponding uncoded bits have enough protection based on the distance Δ_i and the given signal-to-noise ratio E_s/N_0^* to bring the BER below $P_b = 10^{-5}$. Two examples are given in the following:

- Example 1:
  Signal set: four-dimensional 8-PSK.
  Desired information rate: 2.5 bits per symbol.
  The two 8-PSK symbols are generated by the following rule [8]: $\binom{y_1}{y_2} = z^5\binom{4}{4} + z^4\binom{0}{4} + z^3\binom{2}{2} + z^2\binom{0}{2} + z^1\binom{1}{1} + z^0\binom{0}{1}$ modulo 8. The parity bit is z^0; the information bits are z^1 to z^5.
  Corresponding channel capacity: 8.8 dB [5] $\Rightarrow$ $E_s/N_0^* = 9.8$ dB.

Sequence of distances Δ_i for the partition chain of the signal set [8] and corresponding uncoded bit error rates

Partition Level	Δ_i	$P_b(\Delta_i)$
0	0.586	0.05
1	1.172	0.009
2	2	0.001
3	**4**	$6.5 \cdot 10^{-6}$
4	4	$6.5 \cdot 10^{-6}$
5	8	$3.5 \cdot 10^{-10}$
6	∞	—

Conclusion: 3 encoded bits (including the parity bit) are necessary to reach the desired BER for the uncoded bits (hence $\tilde{m} = 2$).

- Example 2:
  Signal set: two-dimensional 64-QAM.
  Desired information rate: 5 bits per symbol.
  Corresponding channel capacity: 16.2 dB [5] $\Rightarrow$ $E_s/N_0^* = 17.2$ dB.
  Sequence of distances Δ_i for the partition chain of the signal set [5] and corresponding uncoded bit error rates

Partition Level	Δ_i	$P_b(\Delta_i)$
0	0.095	0.06
1	0.19	0.013
2	0.38	$8 \cdot 10^{-4}$
3	**0.76**	$4 \cdot 10^{-6}$
4	1.52	$1.3 \cdot 10^{-10}$
5	3.05	$1.8 \cdot 10^{-19}$
6	∞	—

  Conclusion: again 3 encoded bits are necessary to reach the desired BER for the uncoded bits ($\tilde{m} = 2$).

For small block lengths we will operate the coding scheme at even higher signal-to-noise ratios, so we will be on the safe side as far as this design rule is concerned. To target lower BER we will have to adjust P_b accordingly, of course, and yield a higher value for $\tilde{m}$.

2.3.3. Special Interleaver Design for Improved Performance. Several researchers have worked on designing good interleavers for binary Turbo codes. We would like to point out the work by Hokfelt [12–15], where the effect of the interleaver is evaluated in terms of (1) the influence on the iterative decoding algorithm and (2) the avoidance of low-weight error events. Hokfelt proposed the design of interleavers that have positive characteristics with respect to both of these criteria. Specifically, they are designed to improve the decoding performance at lower signal-to-noise ratios and simultaneously improve the BER at high signal-to-noise ratios where Turbo codes often exhibit the "flattening" of the BER curves. We have applied Hokfelt's interleavers to some of the examples for Turbo TCM, also with marked performance gains (see Section 4).

2.4. Component Code Search

In order to find good component codes, one can perform an exhaustive computer search similar to that in [5] that maximizes the minimal distance of each component code under consideration of randomly selecting the parity bits of each second symbol. Furthermore, one should restrict the search to those codes with a primitive feedback polynomial that are widely accepted to yield good performance for Turbo codes (all codes with primitive feedback polynomial thus found have a minimal distance as good as the best candidate codes with nonprimitive feedback polynomial).

A further condition on the code is that the information bits in step k do not affect the value of the parity bits at step k; this condition was also proposed for good TCM codes [5].

Equation (15b) in [5] states that the minimal distance is bounded by

$$d_{\text{free}}^2 \geq \Delta_{\text{free}}^2 = \min \sum_{i=k}^{k+L} \Delta_{q(\mathbf{E}_i)}^2 \equiv \min \Delta^2 \left[\mathbf{E}(D)\right] \qquad (2)$$

minimizing over all nonzero code sequences $\mathbf{E}(D)$. The variable $q(\mathbf{E}_i)$ is the number of trailing zeros in $\mathbf{E}_i$. The values $\Delta_0^2, \Delta_1^2, \Delta_2^2, \ldots$, are the squared minimal Euclidean distances between signals of each subset and must be replaced by $\Delta_0^{*2}, \Delta_1^{*2}, \Delta_2^{*2}, \ldots$, when the corresponding transmitted symbol was "punctured"; the distances are shown in Fig. 3 for 8-PSK. These new distances can be calculated by assuming that the "random" parity bit takes its worst-case value and minimizes the distance between elements of the subsets. In our search we also test both possible states of the puncturing pattern (punctured, unpunctured, punctured, ..., vs. unpunctured, punctured, unpunctured, ...) and retain the lowest distance obtained. After such a search one obtains the results of Table 1, where the parity-check polynomials in octal notation are given as in [5]. Note that in the case of 8-PSK the punctured code has a loss compared to uncoded QPSK ($d_{\text{free}}^2/d_{\text{QPSK}}^2 = d_{\text{free}}^2/2 = 0.878$), but we must not forget that we are able to transmit an *additional* (parity) bit every $2 \cdot n$ 8-PSK symbols, albeit with little protection within the signal constellation.

Table 1. "Punctured" TCM Codes with Best Minimal Distance and Primitive Feedback Polynomial for 8-PSK and QAM (in Octal Notation)

Code	$\tilde{m}$	$H^0(D)$	$H^1(D)$	$H^2(D)$	$H^3(D)$	$d_{\text{free}}^2/\Delta_0^2$
2D-8-PSK, 8 states	2	11	02	04	—	3
4D-8-PSK, 8 states	2	11	06	04	—	3
2D-8-PSK, 16 states	2	23	02	10	—	3
4D-8-PSK, 16 states	2	23	14	06	—	3
2D Z^2, 8 states	3	11	02	04	10	2
2D Z^2, 16 states	3	23	02	16	04	3
2D Z^2, 8 states	2	11	04	02	—	3
2D Z^2, 16 states	2	23	04	10	—	4

3. THE DECODER

3.1. Differences to Binary Turbo Codes

The iterative decoder is similar to that used to decode binary Turbo codes, except that there is a difference in the nature of the information passed from one decoder to the other, and in the treatment of the very first decoding step. Turbo codes received their name from the iterative nature of the decoding algorithm. The main building block of the complete Turbo decoder is the component decoder, which may be a soft-output Viterbi decoder [16], or a symbol-by-symbol (S-b-S) maximum a priori (MAP) decoder [7,17,18]. The component decoders each perform optimal (or close-to-optimal) decoding of each component code, and use the output of the *other* component decoder as if it were an *independent* estimate of the information bits. In the binary Turbo coding scheme, it can be shown that the component decoder's output can be split into three additive parts (when in the logarithmic or loglikelihood ratio domain [17,18]) for each information bit with index k: the *systematic component* (corresponding to the received systematic value for bit k), the *a priori component* (the information given by the other decoder for bit k), and the *extrinsic component* (that part that depends on all other inputs, i.e., those to the "left" and "right" of the bit k in the associated trellis diagram and also the parity information). Only the so-called extrinsic component may be given to the next decoder; otherwise information will be used more that once in the next decoder [1,19]. Furthermore, these three components are disturbed by independent noise.

A major novelty when decoding TTCM is the fact that each decoder alternately sees *its* corresponding encoder's noisy output symbol(s) and then the *other* encoder's noisy output symbol(s). The information bits, that is, systematic bits, that partly resulted in the mapping of each of these symbols are correct — in the sense of being identical to the corresponding encoder output — in both cases. However, this is not so for all the parity bits, since these originate from the other encoder every other group of n symbol — we have indexed these symbols with "$*$," and will call these symbols "punctured" for brevity. Note that in the following, the attribute "$*$" or "punctured" refers to the pertinent component decoder only. The situation is further complicated by the fact that the systematic component cannot be separated from the parity one. This is because the noise that affects the parity component also affects the systematic component since (unlike in the binary case) the systematic information is transmitted together with parity information in the same symbol(s).

Fortunately, these two problems, alternating "punctured" parity bits, and inseparability of systematic and parity components can be solved simultaneously. The trick is to split the output into just two different components: (1) a priori and (2) extrinsic together with systematic. Furthermore, care is taken to avoid using the systematic information more than once in each decoder as will be explained later.

We recommend that the reader briefly review Appendix A, where we have derived the symbol-by-symbol MAP decoder for nonbinary trellises and thus to become familiar with the terms forward and backward variables (α and β), and transitions in the trellis, before continuing with Section 3.2.

3.2. Extrinsic, A Priori, and Systematic Components

Because we will now take a close look at the way the iterative decoder works, we have decided to write logarithms of probabilities, denoted by $L()$, for brevity and clarity. Let us thus define

- $L(d_k = i)$, the logarithm of the decoder output (A.10), written as L in diagrams.
- $L_a(d_k = i) = \log \Pr\{d_k = i\}$, the logarithm of the a-priori term (A.4), written as a in diagrams.
- $L_{(p\&s)}(d_k = i) = \log p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M')$, the logarithm of the inseparable parity and systematic components. Note that we have written $(p\&s)$ in parentheses to stress their inseparability. It is written as $(p\&s)$ in diagrams.
- $L_{(e\&s)}(d_k = i)$, the logarithm of the combined extrinsic and systematic components, written as $(e\&s)$ in diagrams. It will be discussed in the following.

We had stated above that we wish to pass the combined extrinsic and systematic components, $L_{(e\&s)}$, to the next decoder in which it is used as a priori information. This part of the decoder output does not depend on the a priori information $\Pr\{d_k = i\}$. In other words, we must subtract the logarithm of the a priori term $L_a(d_k = i) = \log \Pr\{d_k = i\}$ from the logarithm of (A.10) to obtain a term independent of the a priori information $\Pr\{d_k = i\}$. Thus we compute:

$$L_{(e\&s)}(d_k = i) = \log \Pr\{d_k = i \mid \underline{\mathbf{y}}\} - \log \Pr\{d_k = i\} \qquad (3)$$

$\forall i \in \{0, \ldots, 2^m - 1\}$. This can be done since $\Pr\{d_k = i\}$ is a factor in γ_i that does not depend on M or M' and can be written outside the summations in (A.10). Note that the parity component cannot be separated from the extrinsic once since the former depends on M and M' and cannot be written outside the summations in (A.10). Finally, since the systematic component cannot be split from the parity component, we cannot split $(e\&s)$ at all — hence the parentheses in $L_{(e\&s)}$.

However, the decoder must be formulated in such a way that it correctly uses the channel observation $\mathbf{y}_k$ and the a priori information $\Pr\{d_k = i\}$ at each step k. This is best illustrated in a diagram (see Fig. 5). Shown on the left is the interrelation of both MAP decoders for one information bit in a binary Turbo coding scheme. We have denoted the extrinsic component — omitting the index k — with e, the a priori component with a and the systematic and parity ones with s and p. Bold letters indicate that the variables correspond directly to the upper decoder, nonbold ones correspond directly to the lower decoder. Thus bold (nonbold) extrinsic and L values are *produced* by the upper (lower) decoder and bold (nonbold) a priori values *used* by the upper (lower) decoder. Of course, the decoders have memory (indicated by inputs α and β), so each input will affect many neighboring outputs; we have shown the relationships for only one trellis transition

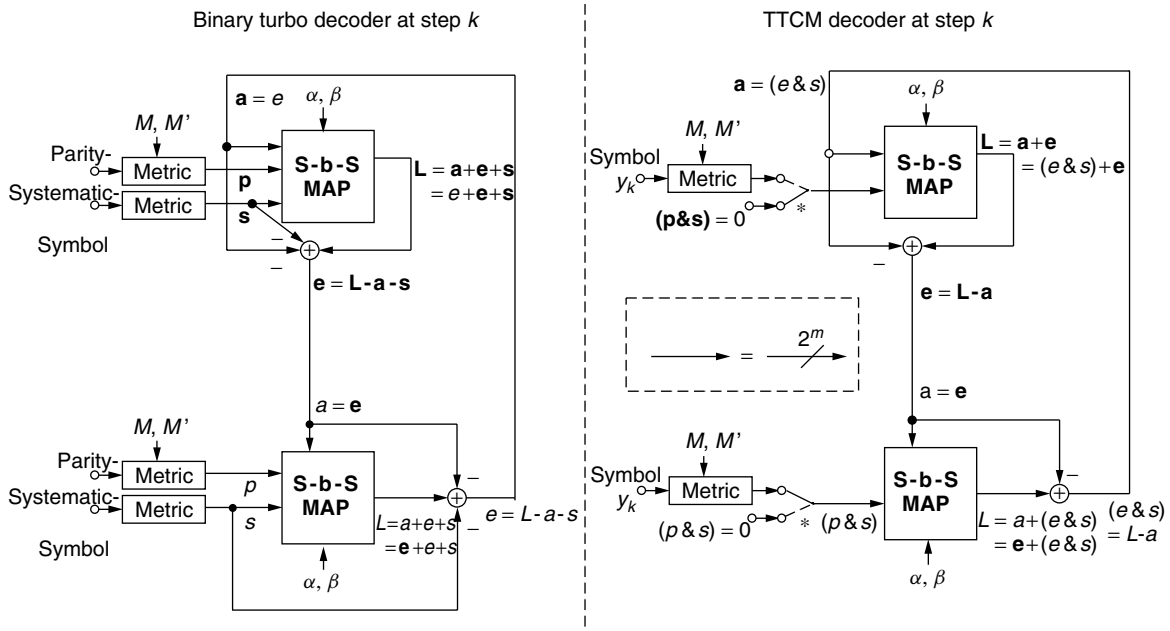


Figure 5. The decoders for binary Turbo codes and TTCM. Note that the labels and arrows apply only to one specific info bit (left) or group of m info bits (right). The interleavers/deinterleavers are not shown.

(step). Both decoders are symmetric as they only pass the newly generated extrinsic information to the next decoder.

The right side of Fig. 5 shows the decoders for TTCM where the upper decoder sees a punctured symbol (which was output by the other decoder: "$*$ mode"), in the example of the encoder in Fig. 2 it might have received a noisy observation of symbol $x_2 = 3$. The corresponding symbol from the upper encoder (**2**) was not transmitted. The upper decoder now ignores this symbol — indicated by the position of the upper switch — as far as the direct channel input is concerned: in Eq. (A.3) we set

$$L_{(p\&s)} = \log p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M') \to 0 \qquad (4)$$

illustrated in Fig. 5 by $(\mathbf{p}\&\mathbf{s}) = \mathbf{0}$. The only input for this step in the trellis is a priori information $L_{\mathbf{a}}$ from the lower decoder, which includes the systematic and (lower) parity information ($p\&s$). The output of the MAP, for this transition, is the sum of this a-priori information $L_{\mathbf{a}}$ and newly computed extrinsic information $L_{\mathbf{e}}$, since we have set $L_{(\mathbf{p\&s})}$ to zero. The a priori information $L_{\mathbf{a}}$ is subtracted, and the extrinsic information $L_{\mathbf{e}}$ is passed to the lower decoder as *its* a-priori information, L_a (see the equations written in Fig. 5). The lower decoder, however, sees a symbol that *was* generated by its encoder; hence it can compute

$$L_{(p\&s)}(d_k = i) = \log p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M') \qquad (5)$$

for each i, and subsequently $L_{(e\&s)}(d_k = i)$ using (3). Then $L_{(p\&s)}(d_k = i)$ is used as the a priori input of the upper decoder in the next iteration. The setting of the switches will alternate from one group of bits (index k) to another. The *symmetry of the decoders* seeing alternately punctured and unpunctured symbols allows decoders to include the systematic information despite of the fact that it cannot be separated from the parity part.

3.2.1. Summary of the Decoder Iteration. We have summarized the steps of one complete iteration in the following algorithm and have numbered the steps in Fig. 6 accordingly. The horizontal line delimits the upper from the lower decoder. We begin with the left hand side of the figure which corresponds to the lower decoder seeing an unpunctured symbol:

1. Compute the logarithm of the branch transition probability [Eq. (A.3)] for each possible value of $d_k = i$. This now takes the systematic and parity components into account for the transition k [Eq. (5)]. Note that the last part of (A.3) denotes the a priori information ($L_a = L_{\mathbf{e}}$) computed by the upper decoder associated with trellis transition k, for each possible value of $d_k = i$.
2. Compute the logarithms of the forward and backward variables, α_{k-1} and β_k, with Eqs. (A.1) and (A.2) for all trellis states M' and M. This now takes into account the code constraint and thus includes all a priori information generated by the upper decoder for the neighboring trellis transitions $\neq k$.
3. Compute the logarithm of the MAP output (A.10) for each possible value of $d_k = i$ using the results of steps 1 and 2. Then subtract the corresponding a priori information ($L_a = L_{\mathbf{e}}$) computed by the upper decoder for each possible value of $d_k = i$ associated with the transition k. The subtraction is a vector operation of length 2^m [see Eq. (3)].
4. Pass the result ($L_{\mathbf{a}} = L_{(e\&s)}$) of step 3 to the upper decoder for it to use in its next decoding iteration. It is the combined extrinsic and systematic component.

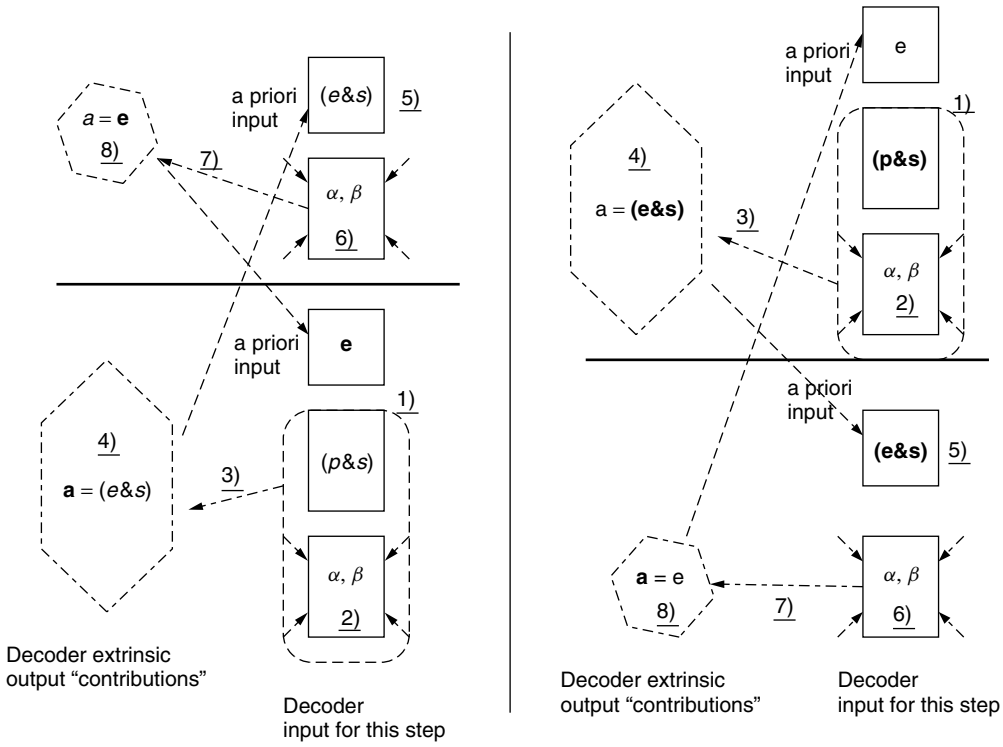


Figure 6. Illustration of the decoder algorithm for TTCM. The symbols used are those used in the right side of Fig. 5 and explained in the text; underlined numbers refer to the steps of the detailed algorithm. The left hand refers to the case where the upper decoder sees a punctured symbol; the right hand, where the lower decoder sees a punctured symbol. The horizontal line delimits the upper decoder from the lower decoder.

In steps 5–8 the index k is the deinterleaved position of index k in steps 1–4.

5. Because this is a punctured symbol seen from the upper decoder's stance, compute the logarithm of the branch transition probability [Eq. (A.3)] for each possible value of $d_k = i$ setting the logarithm of $p(y_k \mid d_k = i, S_k = M, S_{k-1} = M')$ to zero [Eq. (4)]. Note that the last part of (A.3) denotes the a priori information $[L_{\mathbf{a}} = L_{(e\&s)}]$ computed by the lower decoder associated with trellis transition k, for each possible value of $d_k = i$.
6. Same as step 2, exchange "upper decoder" by "lower decoder."
7. Compute the logarithm of the MAP output (A.10) for each possible value of $d_k = i$ using the results of steps 5 and 6. Then subtract the corresponding a priori information $[L_{\mathbf{a}} = L_{(e\&s)}]$ computed by the lower decoder for each possible value of $d_k = i$ associated with the transition k. The subtraction is a vector operation of length 2^m.
8. Pass the result L_a of step 7 to the lower decoder to use in its next decoding iteration. Note that L_a comprises the extrinsic component $L_{\mathbf{e}}$ *without* any systematic or parity component.

For trellis transitions where the lower decoder sees a punctured symbol, the right side of Fig. 6 applies. The same steps (1–8) are performed, except that we swap "upper decoder" and "lower decoder" and swap bold and nonbold notation accordingly.

For one whole iteration:

- Begin with the upper decoder.
- Use the a priori information generated by the lower decoder in its last decoding phase.
- Go through all values of k for $0 \le k < N$ applying steps 1–4 or 5–8 from above as applicable for the puncturing (seen from the upper decoder's stance) of that symbol k.
- Then go to the lower decoder.
- Use the a priori information generated by the upper decoder in its last decoding phase.
- Go through all values of k for $0 \le k < N$ applying steps 1–4 or 5–8 from above as applicable for the puncturing (seen from the lower decoder's stance) of that symbol k.

3.3. Metric Calculation in the First Decoding Stage

The description above assumes that in case a decoder sees a punctured symbol, the systematic and parity information is available from the a priori information received from the other decoder. This is the case in all except the very first decoding stage of the upper decoder. Hence, before the first decoding stage of the upper decoder, we need to set the a priori information to contain the systematic information for the $*$ transitions, where the transmitted symbol was

determined partly by the information group d_k but also by the unknown parity bit $b^{0,*} \in \{0, 1\}$ produced by the *other* encoder. We thus set the a priori information, by applying the mixed Bayes' rule, to

$$\begin{aligned}\Pr\{d_k = i\} \leftarrow \Pr\{d_k = i \mid \mathbf{y}_k\} &= \text{const} \cdot p(\mathbf{y}_k \mid d_k = i) \\ &= \text{const} \cdot \sum_{j\in\{0,1\}} p(\mathbf{y}_k, b_k^{0,*} = j \mid d_k = i) \\ &= \frac{\text{const}}{2} \cdot \sum_{j\in\{0,1\}} p(\mathbf{y}_k \mid d_k = i, b_k^{0,*} = j) \qquad (6)\end{aligned}$$

where it is assumed that $\Pr\{b_k^{0,*} = j \mid d_k\} = \Pr\{b_k^{0,*} = j\} = \frac{1}{2}$, that is, that the parity bit in the symbol x_k is statistically independent of the information bit group d_k and equally likely to be zero or one. Furthermore, the initial a priori probability of d_k — prior to any decoding — is assumed to be constant for all i. Above, it is not necessary to calculate the value of the constant, since the value of $\Pr\{d_k = i \mid \mathbf{y}_k\}$ can be determined by dividing the summation $\sum_{j\in\{0,1\}}$ by its sum over all i (normalization). If the upper decoder is not at a $*$ transition, then we simply set $\Pr\{d_k = i\}$ to $\frac{1}{2^m}$.

3.4. The Complete Decoder

The complete decoder is shown in Fig. 7. By 'metric s' we mean the evaluation of (6). All thin signal paths are channel outputs or values of $\log p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M')$; thick paths represent a group of 2^m values of logarithms of probabilities.

3.4.1. Avoiding Calculation of Logarithms and Exponentials. Since we work with logarithms of probabilities, it is undesirable to switch between probabilities and their logarithms. This, however, becomes necessary at the following four stages in the decoder:

1. In (6) when we sum over probabilities $\left(\sum_{j\in\{0,1\}} p(\mathbf{y}_k \mid d_k = i, b_k^{0,*} = j)\right)$, but the demodulator provides us with $\log(p(\mathbf{y}_k \mid d_k = i, b_k^{0,*} = j))$.
2. When evaluating $\sum_i \sum_{j\in\{0,1\}} p(\mathbf{y}_k \mid d_k = i, b_k^{0,*} = j)$ to normalize (6) to unity.
3. When normalizing the sum of (A.10) to unity.
4. When calculating the hard decision of each individual bit given the values of (A.10).

All of the preceding mandate the calculation of the logarithm of the sum over exponentials (when the decoder otherwise operates in the log domain). By recursively applying the relation [18]

$$\begin{aligned}\ln(e^{\delta_1} + e^{\delta_2}) &= \max(\delta_1, \delta_2) + \ln(1 + e^{-|\delta_2 - \delta_1|}) \\ &= \max(\delta_1, \delta_2) + f_c(|\delta_1 - \delta_2|) \qquad (7)\end{aligned}$$

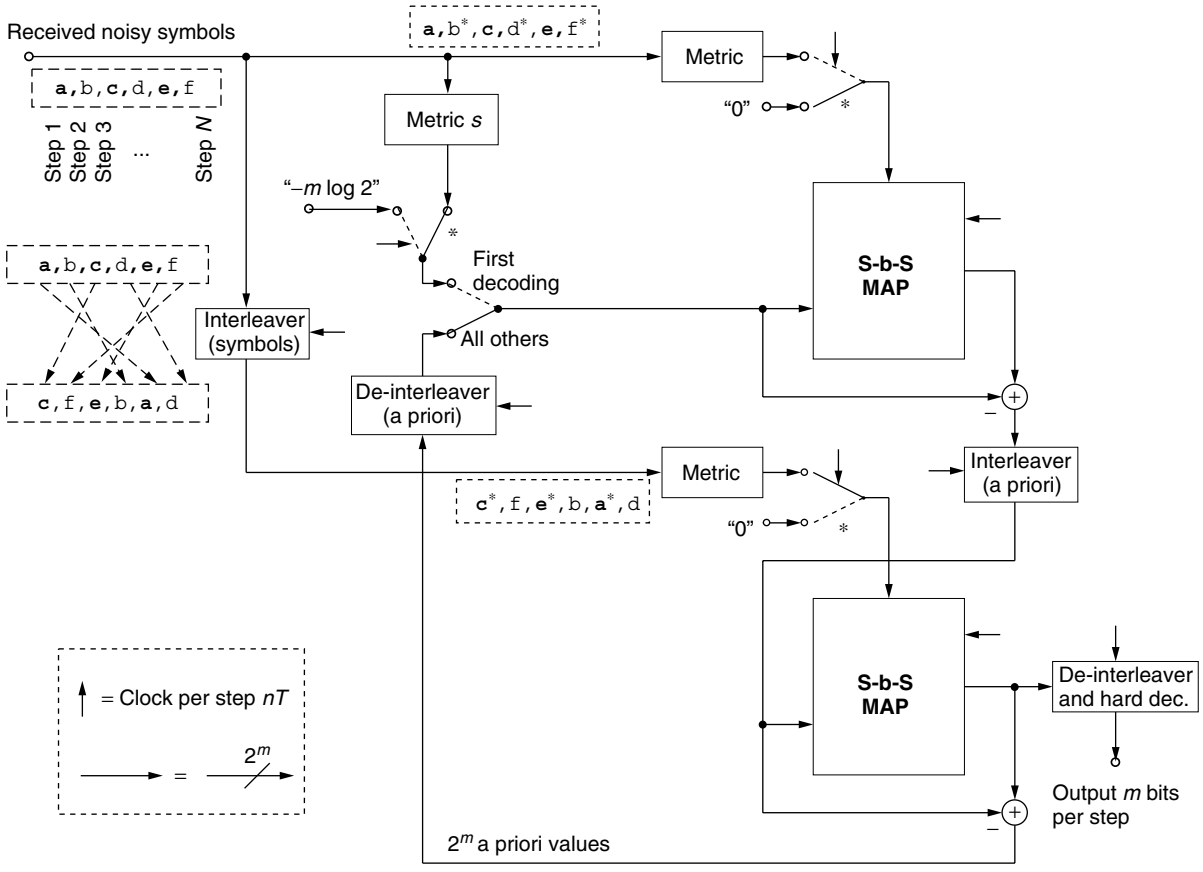


Figure 7. The complete decoder.

the problem can be solved for an arbitrary number of exponentials. The correction function $f_c(.)$ can be realized with a one-dimensional table with as few as eight stored values [18]. When implementing the preceding, we noticed negligible performance degradation.

3.4.2. Subset Decoding. When the component code's trellis contains parallel transitions, this reduces the required decoding complexity: During the iterations, it is not necessary to decide on, or calculate soft outputs for, the uncoded bits that cause these parallel transitions. In the MAP decoders, the parallel transitions can be merged, which mathematically corresponds to adding the path transition probabilities $\gamma_i(\mathbf{y}_k, M', M)$ of the parallel transitions. It is clear that the sum is over just those $2^{(m-\tilde{m})}$ values of i that represent all combinations of the statistically independent uncoded bits. There is one such sum for every particular combination of the remaining $\tilde{m}$ bits that are encoded. The MAP decoder calculates and passes on only the likelihoods of these $\tilde{m}$ bits; hence the (de)interleaver needs to operate only on groups of $\tilde{m}$ bits. During the very last decoding stage, decisions (and, if desired, reliabilities) for the $(m - \tilde{m})$ uncoded bits can be generated by the MAP decoder; either optimally or suboptimally, for example, by taking into account only those transitions between the most likely states along the trellis.

4. EXAMPLES AND SIMULATIONS

As examples we have used 2D-8-PSK (with $N = 1024$), 2D-16-QAM (with $N = 683$), 4D-8-PSK (with 200), and 2D-64-QAM (with $N = 200$ and 1024). Unless stated otherwise, the interleavers were chosen to be pseudorandom, and identical for each transmitted block. In all cases the component decoders were symbol-by-symbol MAP decoders operating in the log domain. The number of trellis states was either 4, 8, or 16. To help the reader compare curves for different values of N, the x axes of the respective curves were chosen to show the same range of SNR. The channel was modeled to be AWGN, where N_0 is the one-sided noise power spectral density. The small block sizes were included to verify that the schemes work well in applications that tolerate only short end-to-end delays. In general, it must be borne in mind that when comparing different approaches to channel coding, the block size (or other measure of fundamental delay) must be kept constant.

The BER curves resulting from Monte Carlo computer simulations are shown in Figs. 8 and 9 for 8-PSK with 2 bits per symbol (b/s); in Fig. 10 for 16-QAM with 3 b/s; in Fig. 11 for 8-PSK with 2.5 b/s and finally in Fig. 12 for 64-QAM with 5 b/s. One iteration is defined as comprising two decoding steps: one in each dimension. The weak asymptotic performance of the component code (evident from the high BER after the very first decoding step) does not seem to affect the performance of the Turbo code after a few iterations, since good BER can be achieved at less than 1 dB from Shannon's limit for large interleaver sizes N. For comparison, Fig. 8 includes the results for a Gray mapping scheme for 2D-8-PSK as presented in [3]; it has the same complexity (when measured as the number of trellis branches per information bit) as the TTCM four-iteration scheme and the same number of information bits per block: 2048. The number of states of the binary trellis for the Gray mapping scheme is 8; hence there are $2048 \times 8 \times 2$ trellis branches per decoding in each dimension; in the TTCM scheme there are $1024 \times 8 \times 4$ branches. Compared to TCM with 64-state Ungerboeck codes and 8-PSK (not included in the figures), we achieve a gain of 1.7 dB at a BER of 10^{-4}. At this BER, the TTCM system has a 0.5-dB advantage over the Gray mapping scheme after four iterations. Rather than comparing all our examples with other coding techniques, we simply point out that good BER can be achieved within 1 dB from Shannon's limit as long as the block size is sufficiently large [20]. The use of designed interleavers for 2D-8-PSK and $N = 1024$ is shown in Fig. 9; we see a marked improvement when using

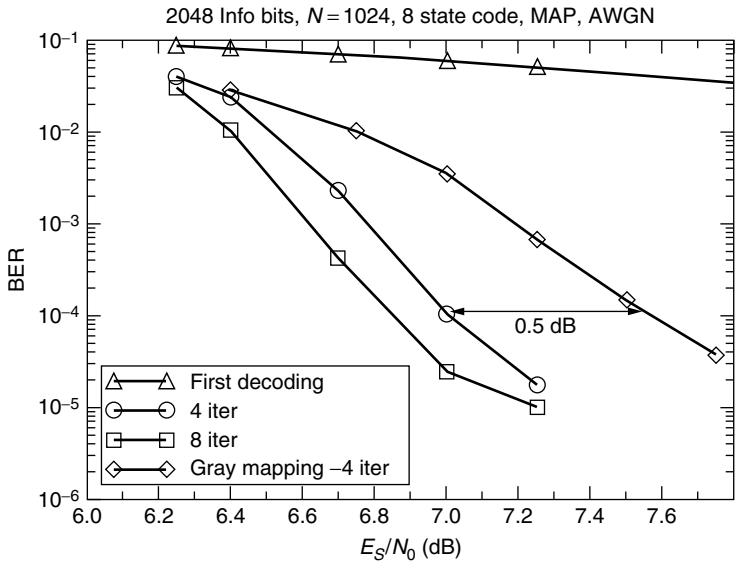


Figure 8. TTCM for 2D-8-PSK, 2 bits per symbol (b/s). Channel capacity: 2 b/s at 5.9 dB. Random interleaver. Code with $\tilde{m} = 2$.

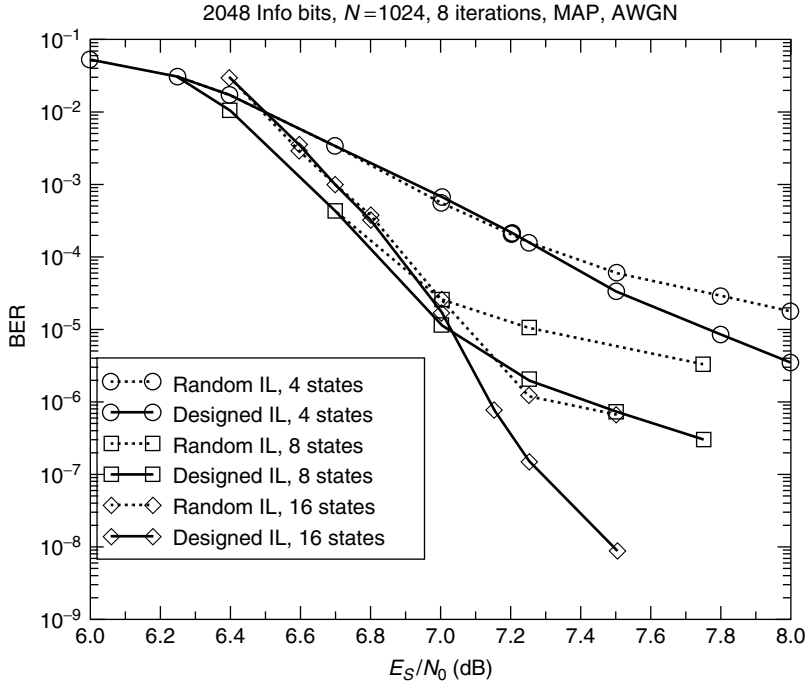


Figure 9. TTCM for 2D-8-PSK and different interleavers and code memory, 2 b/s. Channel capacity: 2 b/s at 5.9 dB. Code with $\tilde{m} = 2$.

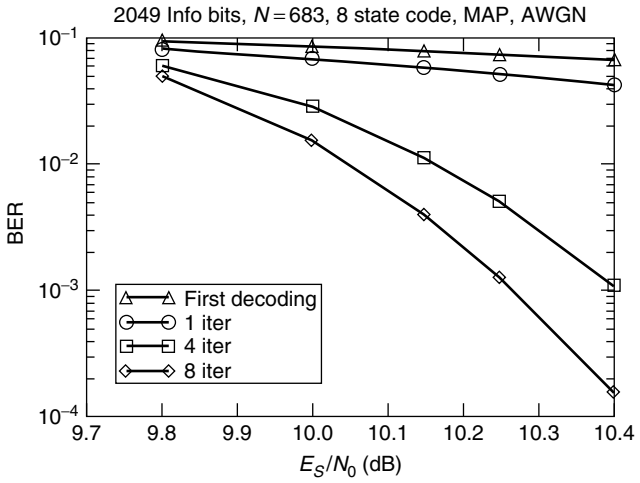


Figure 10. TTCM for 2D-16-QAM, 3 b/s. Channel capacity: 3 b/s at 9.3 dB. Random interleaver. Code with $\tilde{m} = 3$.

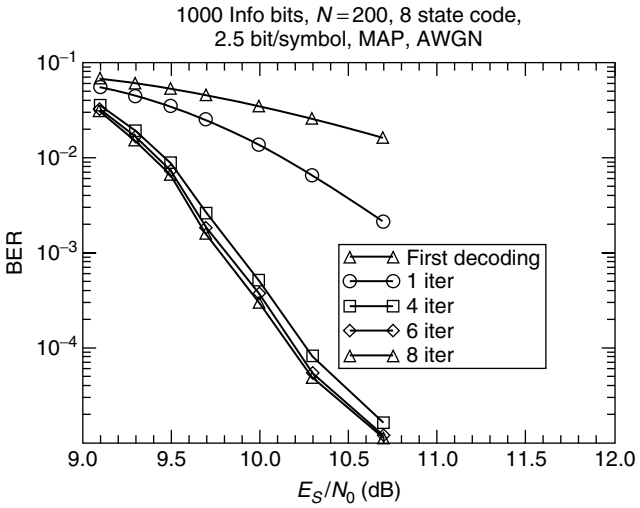


Figure 11. TTCM for 4D-8-PSK, 2.5 b/s. Channel capacity: 2.5 b/s at 8.8 dB. Random interleaver. Code with $\tilde{m} = 2$.

the designed interleaver of Hokfelt [12,13], especially for 16 states.

The results for the higher-bandwidth-efficient examples (2D-64-QAM and 4D-8-PSK) are also encouraging. For most of the simulations we used a random interleaver, unadapted to the component code. This results in the characteristic flattening of the BER curves for higher signal-to-noise ratios and BER lower than 10^{-5}. This BER is consistent with our target BER for the uncoded bits when choosing $\tilde{m}$ as explained in Section 2.3. If we target a lower BER, then we need to

1. Choose $\tilde{m}$ such that the uncoded bits are better protected and suffer from a BER at least as good as our new target BER (e.g., 10^{-7}). For 64-QAM and 4D-8-PSK this means that $\tilde{m}$ should now be 3.
2. Choose component codes with 16 states (as given in Table 1).
3. Choose an interleaver designed for the component codes according to the technique of Hokfelt et al. [12,13].

In Fig. 13 we show results for 64-QAM modulation memory 4 component codes, 5120 information bits in 1024 blocks, using random and constructed interleavers, and after eight decoding iterations.

It is to be noted that Turbo-coded systems will often be employed as an inner coding stage by concatenating an

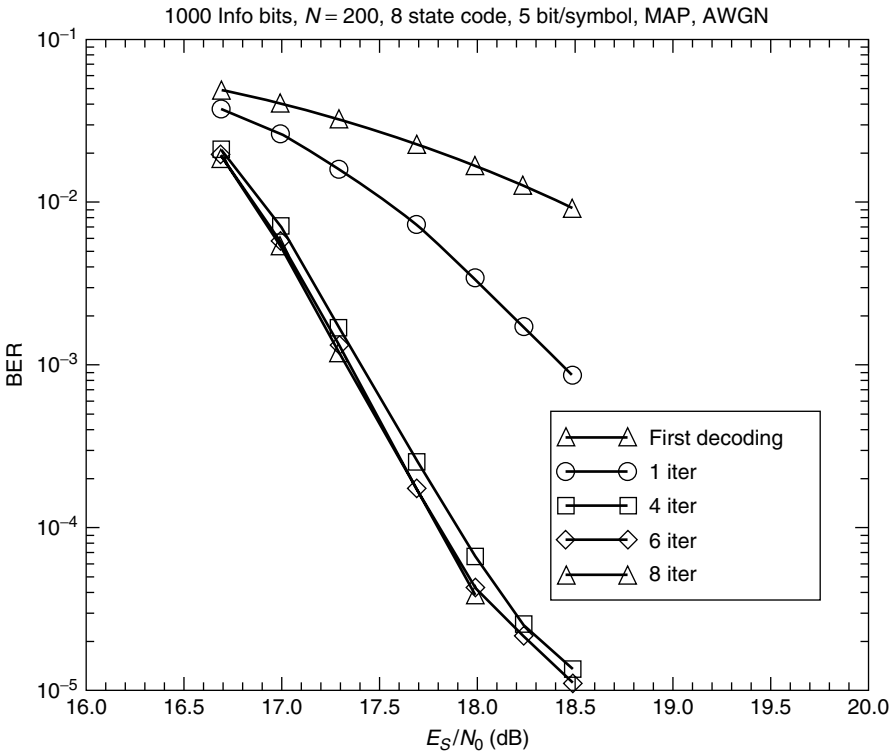


Figure 12. TTCM for 2D-64-QAM, 5 b/s. Channel capacity: 5 b/s at 16.2 dB. Random interleaver. Code with $\tilde{m} = 2$.

outer block code (e.g., RS or BCH code [11]) with a Turbo code, in order to reach very low BER; in these cases, BERs of around 10^{-4} are often sufficient. When we use TTCM with specially designed interleavers and achieve BERs of 10^{-7}, then only very high rate outer block codes will be needed. In the actual computer simulations performed for 2D-8-PSK and memory 4, with the designed interleaver, 2048 information bits per block, and E_S/N_0 7.5 dB, we never encountered a block with more than 11 bit errors (the majority of erroneous blocks had just 4 errors).

5. OTHER WORK AND FURTHER READING

There is a vast literature covering Turbo codes and iterative decoding. Good starting points for decoding principles and an overview of Turbo codes are [17,21], and [22]. Online resources — with further links — related to Turbo codes — can be found in [23–25].

A different approach for bandwidth-efficient coding using recursive parallel concatenation of nonbinary component codes was proposed in [26] and [27], where there is no puncturing of parity bits or symbols. A scheme using serial concatenation and four-dimensional modulation has been proposed [28]. Another serial concatenation scheme has been presented [29] that shows a marked improvement to the parallel concatenated scheme of Benedetto et al. [26,27] at high SNR but a worse performance at lower SNR, especially for larger interleavers.

Work is being carried out to assess the performance of TTCM in fading channels; see [30] and [31] for examples — the former work includes multicarrier transmission. Also, TTCM has been successfully employed in transmit antenna diversity systems in conjunction with spacetime codes in fading channels, including frequency-selective channels [32,33].

6. SUMMARY

We have illustrated the channel coding scheme called Turbo trellis-coded modulation (TTCM), which is bandwidth-efficient and allows iterative "Turbo" decoding of codes built around punctured parallel concatenated trellis codes together with higher-order modulation. The bitwise interleaver known from classic binary Turbo codes is replaced by an interleaver operating on a group of bits. By adhering to a set of constraints for component code and interleaver, the resulting code can be decoded iteratively using, for example, symbol-by-symbol MAP component decoders working in the logarithmic domain to avoid numerical problems and reduce the decoding complexity. We outlined the structure of the iterative decoder and derived the symbol-by-symbol MAP algorithm for nonbinary trellises. Furthermore, we illustrated the differences to the binary case as far as the use of extrinsic, systematic, and parity components of the symbol-by-symbol decoder output are concerned.

The search results for good component codes are shown, taking into account the puncturing at the transmitter. Simulation results are presented for codes with 4, 8, and 16 states, employing random and designed interleavers, and various signal sets such as two- and four-dimensional 8-PSK, 16-QAM, and 64-QAM. With TTCM, error correction close to Shannon's limit is possible for highly bandwidth-efficient schemes that are of relatively low complexity. It remains to be seen whether further

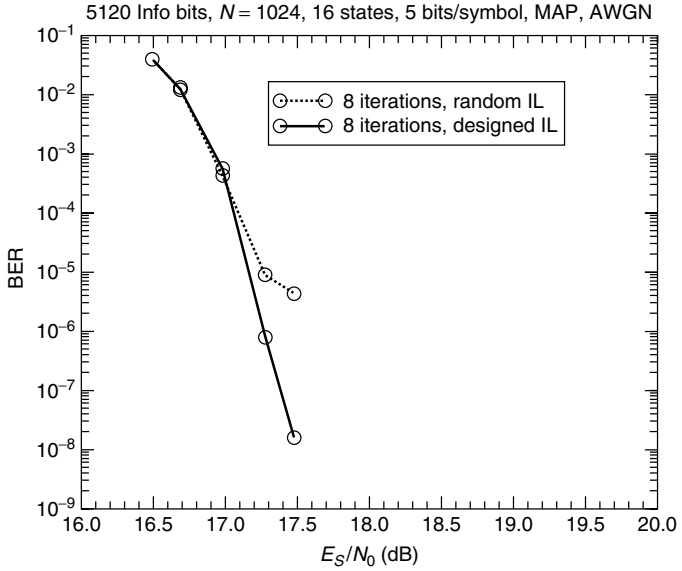


Figure 13. TTCM for 2D-64-QAM, 5 b/s. Channel capacity: 5 b/s at 16.2 dB. Code with $\tilde{m} = 3$.

refinement of the interleaver construction will reduce the flattening of the BER curves below 10^{-8} or 10^{-10}.

Acknowledgments

The authors would like to thank Dr. Joachim Hagenauer for valuable discussions and Dr. Johan Hokfelt for providing designed interleavers.

APPENDIX A. THE SYMBOL-BY-SYMBOL MAP ALGORITHM FOR NONBINARY TRELLISES

We will briefly show the derivation of the symbol-by-symbol MAP algorithm [7] (MAP for short) to be used as the component decoder in the iterative decoding scheme for TTCM built with for nonbinary trellises. For the derivation we consider just a conventional, unpunctured TCM scheme, with a priori information -on each group of information bits d_k- to be used at the input of the decoder. Let the number of states be 2^ν, and the state at step k be denoted by $S_k \in \{0, 1, \ldots, 2^\nu - 1\}$. The group of m information bits d_k can be represented by an integer in the range $(0 \ldots 2^m - 1)$ and is associated with the transition from step $k-1$ to k. The receiver observes N sets of n noisy symbols, where n such symbols are associated with each step in the trellis; specifically, from step $k-1$ to step k the receiver observes $\mathbf{y}_k = [y_k^0, \ldots, y_k^{(n-1)}]$. Let the total received sequence be $\underline{\mathbf{y}} = \underline{\mathbf{y}}_1^N = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. This is the TCM encoder output sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ that has been disturbed by additive white Gaussian noise with one-sided noise-power spectral density N_0. Each $\mathbf{x}_k = [x_k^0, \ldots, x_k^{(n-1)}]$ is the group of n symbols output by the mapper at step k.

The goal of the decoder is to evaluate $\Pr\{d_k \mid \underline{\mathbf{y}}_1^N\}$ for each d_k, and for all k. Let us introduce and define the so-called forward and backward variables:

$$\alpha_{k-1}(M') = \frac{p(S_{k-1} = M', \underline{\mathbf{y}}_1^{k-1})}{p(\underline{\mathbf{y}}_1^{k-1})} \tag{A.1}$$

$$\beta_k(M) = \frac{p(\underline{\mathbf{y}}_{k+1}^N \mid S_k = M)}{p(\underline{\mathbf{y}}_{k+1}^N \mid \underline{\mathbf{y}}_1^k)} \tag{A.2}$$

The branch transition probability for step k, $p(d_k = i, \mathbf{y}_k, S_k = M \mid S_{k-1} = M')$, is denoted by, and calculated as

$$\begin{aligned}\gamma_i(\mathbf{y}_k, M', M) = {} & p(\mathbf{y}_k \mid d_k = i, S_k = M, S_{k-1} = M') \\ & \cdot q(d_k = i \mid S_k = M, S_{k-1} = M') \\ & \cdot \Pr\{S_k = M \mid S_{k-1} = M'\}\end{aligned} \tag{A.3}$$

$q(d_k = i \mid S_k = M, S_{k-1} = M')$ is either zero or one depending on whether encoder input $i \in \{0, 1, \ldots, 2^m - 1\}$ is associated with the transition from state $S_{k-1} = M'$ to $S_k = M$. The first component of (A.3) represents the parity and systematic information available at the output of the transmission channel; its computation depends on the channel noise variance for the case of channels with additive white Gaussian noise (AWGN) [17] and on the actually transmitted symbol associated with the transition from state $S_{k-1} = M'$ to $S_k = M$ for encoder input i. In the last component of (A.3) we use the a priori information. For codes without any parallel transitions:

$$\begin{aligned}&\Pr\{S_k = M \mid S_{k-1} = M'\} = \\ &\begin{cases} \Pr\{d_k = 0\} & \text{if } q(d_k = 0 \mid S_k = M, S_{k-1} = M') = 1 \\ \Pr\{d_k = 1\} & \text{if } q(d_k = 1 \mid S_k = M, S_{k-1} = M') = 1 \\ \quad\vdots & \qquad\qquad \ldots \\ \Pr\{d_k = 2^m - 1\} & \text{if } q(d_k = 2^m - 1 \mid S_k = M, \\ & \qquad S_{k-1} = M') = 1 \end{cases} \\ &= \Pr\{d_k = j\},\end{aligned} \tag{A.4}$$

where $j: q(d_k = j \mid S_k = M, S_{k-1} = M') = 1$. Naturally, when $q(d_k = i \mid S_k = M, S_{k-1} = M') = 1$, then $j = i$; otherwise the value of $\Pr\{S_k = M \mid S_{k-1} = M'\}$ and hence j will be irrelevant anyway. Formally, if there does not exist a j such that $q(d_k = j \mid S_k = M, S_{k-1} = M') = 1$, then $\Pr\{S_k = M \mid S_{k-1} = M'\}$ is set to zero.

We shall now try to combine (A.1), (A.2), and (A.4). We must first bear in mind that the event $(d_k = i, \mathbf{y}_k, S_{k-1} =$

M') has no influence on $\underline{\mathbf{y}}_{k+1}^{N}$ if S_k is given; hence we can write

$$p(\underline{\mathbf{y}}_{k+1}^{N} \mid S_k = M) = p(\underline{\mathbf{y}}_{k+1}^{N} \mid d_k = i, \mathbf{y}_k, S_k = M, S_{k-1} = M') \tag{A.5}$$

Using (A.5) and the fact that

$$p(\underline{\mathbf{y}}_{1}^{k-1}) = \frac{p(\underline{\mathbf{y}}_{1}^{k})}{p(\mathbf{y}_k \mid \underline{\mathbf{y}}_{1}^{k-1})} \tag{A.6}$$

the product of (A.1), (A.2), and (A.4) can now be shown to be:

$$\begin{aligned}
&\alpha_{k-1}(M') \cdot \beta_k(M) \cdot \gamma_i(\mathbf{y}_k, M', M) \\
&\quad = p(S_{k-1} = M', \underline{\mathbf{y}}_{1}^{k-1}) \\
&\qquad \cdot p(\underline{\mathbf{y}}_{k+1}^{N}, d_k = i, \mathbf{y}_k, S_k = M \mid S_{k-1} = M') \\
&\qquad \cdot \frac{p(\mathbf{y}_k \mid \underline{\mathbf{y}}_{1}^{k-1})}{p(\underline{\mathbf{y}}_{1}^{N})}
\end{aligned} \tag{A.7}$$

Obviously

$$\begin{aligned}
&p(\underline{\mathbf{y}}_{k+1}^{N}, d_k = i, \mathbf{y}_k, S_k = M \mid S_{k-1} = M') \\
&\quad = p(\underline{\mathbf{y}}_{k+1}^{N}, d_k = i, \mathbf{y}_k, S_k = M \mid S_{k-1} = M', \underline{\mathbf{y}}_{1}^{k-1})
\end{aligned} \tag{A.8}$$

so we can re-write (A.7) as

$$\begin{aligned}
&\alpha_{k-1}(M') \cdot \beta_k(M) \cdot \gamma_i(\mathbf{y}_k, M', M) \cdot \frac{1}{p(\mathbf{y}_k \mid \underline{\mathbf{y}}_{1}^{k-1})} \\
&\quad = p(S_{k-1} = M', S_k = M, d_k = i, \underline{\mathbf{y}}_{1}^{N}) \frac{1}{p(\underline{\mathbf{y}}_{1}^{N})} \\
&\quad = p(S_{k-1} = M', S_k = M, d_k = i \mid \underline{\mathbf{y}}_{1}^{N})
\end{aligned} \tag{A.9}$$

Therefore, the desired output of the MAP decoder is

$$\begin{aligned}
\Pr\{d_k = i \mid \underline{\mathbf{y}}\} = \text{const} \cdot &\sum_{M} \sum_{M'} \gamma_i(\mathbf{y}_k, M', M) \\
&\cdot \alpha_{k-1}(M') \cdot \beta_k(M)
\end{aligned} \tag{A.10}$$

$\forall i \in \{0, \ldots, 2^m - 1\}$. The constant can be eliminated by normalizing the sum of (A.10) over all i to unity. The probability $\Pr\{d_k = i \mid \underline{\mathbf{y}}\}$ comprises a priori, systematic, parity, and extrinsic components, since it depends on the complete received sequence as well as the a priori likelihoods of d_k.

All that remains now is to recursively define $\alpha_{k-1}(M')$ and $\beta_k(M)$. We begin by writing

$$\begin{aligned}
&\Pr\{S_k = M \mid \underline{\mathbf{y}}_{1}^{k-1}, \mathbf{y}_k\} \cdot p(\mathbf{y}_k \mid \underline{\mathbf{y}}_{1}^{k-1}) \\
&\quad = p(\mathbf{y}_k, S_k = M \mid \underline{\mathbf{y}}_{1}^{k-1})
\end{aligned} \tag{A.11}$$

and dividing both sides by $p(\mathbf{y}_k \mid \underline{\mathbf{y}}_{1}^{k-1})$ and expanding into the form

$$\begin{aligned}
\Pr\{S_k = M \mid \underline{\mathbf{y}}_{1}^{k}\} &= \alpha_k(M) \\
&= \frac{\sum_{M'} p(\mathbf{y}_k, S_k = M, S_{k-1} = M' \mid \underline{\mathbf{y}}_{1}^{k-1})}{\sum_{M} \sum_{M'} p(S_k = M, S_{k-1} = M', \mathbf{y}_k \mid \underline{\mathbf{y}}_{1}^{k-1})}
\end{aligned} \tag{A.12}$$

Because of (A.8), we can write

$$\begin{aligned}
\alpha_k(M) = \frac{\sum_{M'} Pr\{\mathbf{y}_k, S_k = M \mid S_{k-1} = M'\} \cdot p(S_{k-1} = M' \mid \underline{\mathbf{y}}_{1}^{k-1})}{\sum_{M} \sum_{M'} p(\mathbf{y}_k, S_k = M \mid S_{k-1} = M') \cdot \Pr\{S_{k-1} = M' \mid \underline{\mathbf{y}}_{1}^{k-1}\}}
\end{aligned} \tag{A.13}$$

Defining

$$\gamma_T(\mathbf{y}_k, M', M) = \sum_{i=0}^{2^m - 1} \gamma_i(\mathbf{y}_k, M', M) \tag{A.14}$$

yields

$$\alpha_k(M) = \frac{\sum_{M'} \gamma_T(\mathbf{y}_k, M', M) \cdot \alpha_{k-1}(M')}{\sum_{M} \sum_{M'} \gamma_T(\mathbf{y}_k, M', M) \cdot \alpha_{k-1}(M')} \tag{A.15}$$

Similarly

$$\begin{aligned}
\beta_k(M) &= \frac{\sum_{M''} p(S_{k+1} = M'', \underline{\mathbf{y}}_{k+1}^{N} \mid S_k = M)}{p(\underline{\mathbf{y}}_{k+1}^{N} \mid \underline{\mathbf{y}}_{1}^{k})} \\
&= \frac{\sum_{M''} p(S_{k+1} = M'', \mathbf{y}_{k+1} \mid S_k = M) \cdot p(\underline{\mathbf{y}}_{k+2}^{N} \mid S_{k+1} = M'')}{p(\underline{\mathbf{y}}_{k+1}^{N} \mid \underline{\mathbf{y}}_{1}^{k})}
\end{aligned} \tag{A.16}$$

since $p(\underline{\mathbf{y}}_{k+2}^{N} \mid S_{k+1} = M'') = p(\underline{\mathbf{y}}_{k+2}^{N} \mid S_{k+1} = M'', \mathbf{y}_{k+1}, S_k = M)$. Finally, we can calculate $\beta_k(M)$ recursively using

$$\begin{aligned}
&\beta_k(M) \\
&= \frac{\sum_{M''} p(S_{k+1} = M'', \mathbf{y}_{k+1} \mid S_k = M) \cdot \dfrac{p(\underline{\mathbf{y}}_{k+2}^{N} \mid S_{k+1} = M'')}{p(\underline{\mathbf{y}}_{k+2}^{N} \mid \underline{\mathbf{y}}_{1}^{k+1})}}{\sum_{M''} \sum_{M} p(S_{k+1} = M'', S_k = M, \mathbf{y}_{k+1} \mid \underline{\mathbf{y}}_{1}^{k})} \\
&= \frac{\sum_{M''} \gamma_T(\mathbf{y}_{k+1}, M, M'') \cdot \beta_{k+1}(M'')}{\sum_{M''} \sum_{M} \gamma_T(\mathbf{y}_{k+1}, M, M'') \cdot \alpha_k(M)}
\end{aligned} \tag{A.17}$$

In our implementation of the preceding algorithm, we have used logarithms of probabilities and logarithms of $\alpha_{k-1}(M')$, $\beta_k(M)$, and $\gamma_i(\mathbf{y}_k, M', M)$ employing the quasioptimal log-MAP algorithm [18] that uses the max function in conjunction with a table lookup to compute the logarithm of a sum of exponentials. The loss incurred through the use of the log-MAP algorithm is less than 0.1 dB even when using a lookup table with eight stored values.

BIOGRAPHIES

Patrick Robertson received the Dipl.-Ing. degree in electrical engineering from the Technical University of Munich, Munich, Germany, in 1989 and a Ph.D. from the University of the Federal Armed Forces, Munich, Germany, in 1995. Since 1990, he has been working at the Institute for Communications Technology at the German Aerospace Centre (DLR). From 1990 to 1993 he shared this position with a part-time teaching post at the University of the Federal Armed Forces. His contributions within a number of EU and national R&D projects have included work on key technical definition and standardization, such as for the DVB-T system for digital terrestrial television transmission. In 1999 he became leader of the research Group "Broadband Systems and Navigation." Dr. Robertson's current interests include wireless communications and channel coding, navigation systems, navigation channel measurement, as well as mobile multimedia platforms and service architectures. He holds several patents and has published numerous research papers on these areas.

Thomas Worz was born in Stuttgart, Germany, in 1961. He received the Dipl. Ing. degree in electrical engineering from the Technical University of Stuttgart, Germany, in 1988 and his Ph.D. from the Technical University of Munich, Munich, Germany, in 1995. Since 1988, he has been with the Institute of Communications Technology of the German Aerospace Center (DLR), Oberpfaffenhofen. In 1991, he spent a three-month period as a guest scientist at the Communications Research Centre (CRC), Ottawa, Canada. In 1999, he cofounded the AUDENS Advanced Communications Technology Consulting GmbH and works as a technical consultant to industry and agencies. His research interests include channel coding, coded modulation, synchronization, signal processing, and system design. Currently, he is involved in the definition of the Galileo European Navigation system.

BIBLIOGRAPHY

1. C. Berrou, A. Glavieux, and P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *Proc. ICC'93*, May 1993, pp. 1064–1070.
2. 3GPP, *Universal Mobile Telecommunications System* (*UMTS*) *Multiplexing and Channel Coding* (*FDD*), No. 125.212, v.4.1.0 (standard), 2001.
3. S. Le Goff, A. Glavieux, and C. Berrou, Turbo-codes and high spectral efficiency modulation, *Proc. ICC'94*, May 1994, pp. 645–649.
4. U. Wachsmann and J. Huber, Power and bandwidth efficient digital communication using turbo codes in multilevel codes, *Eur. Trans. Telecommun.* **6**(5): (1995).
5. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28**: 55–67 (Jan. 1982).
6. P. Robertson and T. Woerz, Bandwidth efficient turbo trellis coded modulation using punctured component codes, *IEEE J. Select. Areas Telecommun.* **16**: (Feb. 1998).
7. L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* **IT-20**: 284–287 (March 1974).
8. S. Pietrobon et al., Trellis-coded multidimensional phase modulation, *IEEE Trans. Inform. Theory* **IT-36**: 63–89 (Jan. 1990).
9. S. Benedetto and G. Montorsi, Performance evaluation of parallel concatenated codes, *Proc. ICC'95*, June 1995, pp. 663–667.
10. W. Blackert and S. Wilson, Turbo trellis code modulation, *Proc. CISS'96*, 1996.
11. J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001.
12. J. Hokfelt, O. Edfors, and T. Maseng, Interleaver design for turbo codes based on the performance of iterative decoding, *Proc. ICC'99*, June 1999.
13. J. Hokfelt, O. Edfors, and T. Maseng, On the theory and performance of trellis termination methods for turbo codes, *IEEE J. Select. Areas Commun.* **19**: (May 2001).
14. J. Hokfelt, *On the Design of Turbo Codes*, Ph.D. thesis, Lund Univ., Sweden, Aug. 2000 (published by Dept. Advanced Electronics, ISBN 91-7874-061-4).
15. J. Hokfelt, J. Hokfelt's Turbo codes Internet homepage (online), *http://www.tde.lth.se/home/jht/index.html* (Oct. 2001).
16. J. Hagenauer and P. Hoeher, A Viterbi algorithm with soft-decision outputs and its applications, *Proc. GLOBECOM'89*, Nov. 1989, pp. 1680–1686.
17. J. Hagenauer, E. Offer, and L. Papke, Iterative decoding of binary block and convolutional codes, *IEEE Trans. IT* (March 1996).
18. P. Robertson, P. Hoeher, and E. Villebrun, Optimal and sub-optimal maximum a posteriori algorithms suitable for turbo decoding, *Eur. Trans. Telecommun.* **8**(2): 119–125 (1997).
19. J. H. Lodge, R. Young, P. Hoeher, and J. Hagenauer, Separable MAP 'filters' for the decoding of product and concatenated codes, *Proc. ICC'93*, May 1993, pp. 1740–1745.
20. P. Robertson, An overview of bandwidth efficient turbo coding schemes, *Proc. Int. Symp. Turbo Codes and Related Topics*, Sept. 1997, pp. 103–110.
21. S. ten Brink, Convergence behaviour of iteratively decoded parallel concatenated codes, *IEEE Trans. Commun.* **49**: 1727–1737 (Oct. 2001).
22. J. Woodard and L. Hanzo, Comparative study of turbo decoding techniques: An overview, *IEEE Trans. Vehic. Technol.* **49**: 2208–2233 (Nov. 2000).
23. S. Pietrobon, ITR's Turbo coding home page (online), *http://www.itr.unisa.edu.au/steven/turbo/* (Oct. 2001).
24. R. Pyndiah, Block turbo codes (online), *http://www-sc.enst-bretagne.fr/turbo/principale.html* (Oct. 2001).
25. M. Valenti, Turbo codes at West Virginia University (online), *http://www.csee.wvu.edu/mvalenti/turbo.html* (Oct. 2001).
26. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Bandwidth efficient parallel concatenated coding schemes, *Electron. Lett.* **31**(24): 2067–2069 (1995).
27. S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, Parallel concatenated trellis coded modulation, *Proc. ICC'96*, June 1996, pp. 974–978.

28. D. Divsalar and F. Pollara, Serial and hybrid concatenated codes with applications, *Proc. Int. Symp. Turbo Codes and Related Topics*, Sept. 1997, pp. 80–87.

29. H. Ogiwara and V. Bajo, Iterative decoding of serially concatenated punctured trellis-coded modulation, *IEICE Trans. Fund.* (special section on information theory and its applications) **E82-A**: 2089–2095 (Oct. 1999).

30. L. Piazzo and L. Hanzo, TTCM-OFDM over dispersive fading channels, *Proc. VTC'2000*, May 2000.

31. N. Ha and R. Rajatheva, Performance of turbo trellis coded modulation (T-TCM) on frequency-selective Rayleigh fading channels, *Proc. ICC'99*, June 1999.

32. G. Bauch, J. Hagenauer, and N. Seshadri, Turbo processing in transmit antenna diversity systems, *Ann. Telecommun.* (Special Issue: Turbo Codes — a Widespreading Technique) **56**: 455–471 (Aug. 2001).

33. G. Bauch, Concatenation of space-time block codes and Turbo-TCM, *Proc. ICC'99*, June 1999, pp. 1202–1206.

U

ULTRAWIDEBAND RADIO

KAZIMIERZ (KAI) SIWIAK
LAURA L. HUCKABEE
Time Domain Corporation
Huntsville, Alabama

1. INTRODUCTION

Ultrawideband signaling is essentially the art of generating, modulating, emitting, and detecting baseband digital signals that inherently occupy large bandwidths. Impulse transmissions date back to the infancy of wireless technology. They include the experiments of Heinrich Hertz in the 1880s, and the 100-year-old spark-gap "impulse" transmissions of Guglielmo Marconi, who in 1901 sent the first ever over-the-horizon wireless transmission from the Isle of Wight to Cornwall on the British mainland. Early radio circuits consisted solely of passive electrical components, no tubes or transistors, and hence lacked the means to efficiently deal with short transient impulses. Therefore radio subsequently developed along narrowband frequency-selective analog techniques. This led to voice broadcasting and telephony — and more recently to digital telephony and wireless data. Through the years, a small cadre of scientists have worked to develop and refine impulse technologies. Before 1970 the primary focus in impulse radio research was on impulse radar techniques and government-sponsored projects. In late 1970s and early 1980s, however, digital techniques began to mature to the point where the practicality of modern low-power impulse radiocommunications could be demonstrated using the impulse time coding and time modulation approach. Digital impulse radio [1–9], the modern echo of Marconi's century-old transmissions, now emerges under the banner "ultrawideband" radio. Alternate methods of generating signals having UWB characteristics are being developed including the use of continuous streams of pseudonoise (PN)-coded impulses that resemble code-division multiple access (CDMA) signaling that employ a chip rate commensurate with the emission center frequency. The industry is now moving to commercial deployment.

UWB signaling is more nearly characterized by transient circuit responses, whereas conventional radio tends to deal with the steady state. Impulse propagation, especially indoors, also differs significantly from that of narrowband carrier-based systems in that multipath is described as distinct short impulses that sometimes overlap rather than continuous sine waves that form complex interference patterns. Many applications of UWB systems have been enabled by the unique characteristics of this technology.

2. CODED UWB IMPULSES AND IMPULSE STREAMS

UWB radio is the transmission and reception of ultrashort electromagnetic energy impulses. It is the generic term describing radio systems having very large instantaneous bandwidths. The U.S. Federal Communications Commission (FCC), for example, has tentatively defined UWB systems as "having bandwidths greater than 25% of the center frequency measured at the 10 dB down points" or "RF bandwidths greater than 1.5 GHz," whichever is smaller. There are several methods of generating, radiating and receiving such UWB signals, including TM-UWB, DS-UWB, and TRD-UWB. Wide spectra are generated in each method, however, radio techniques, signal characteristics, and application capabilities vary considerably.

Developers of UWB technology have perfected various ways for creating and receiving these signals, and for encoding information in the transmissions. Pulses can be sent individually, in bursts, or in near-continuous streams, and they can encode information in pulse amplitude, polarity, and position. Modulations vary from simple pulse position, to a more energy-efficient pulse polarity [10], and to the very-energy-efficient *M-ary* (multilevel) pulse position modulation. Modern UWB radio is characterized by very low effective radiated power (in the submilliwatt range) and extremely low power spectral densities, by virtue of the wide bandwidths (>1 GHz). The emissions are targeted to be below an effective isotropicaly radiated power (EIRP) of −41.25 dBm/MHz, with restrictions, in bands below 960 MHz, between 1.99 GHz and 10.6 GHz, and at 24 GHz, under U.S. CFR-47 Part 15 Report and Order issued February 2002.

The following three commercially useful UWB communications techniques exemplify the wide range of implementation possibilities: TM-UWB, DS-UWB, and TRD-UWB. All systems use transient switching techniques to generate brief (typically subnanosecond) impulses or "monocycles" having a small number of zero crossings. The impulses are radiated by specialized wideband antennas [11].

TM-UWB impulses are transmitted at high rates, in the millions to tens of millions of impulses per second. However, the pulses are not necessarily evenly spaced in time, but rather they may be spaced at random or pseudorandom time intervals. The process creates a noise-like signal in both the time and frequency domains. Data modulation is applied by further dithering the timing of the pulse transmissions, by signal polarity and perhaps pulse amplitude. A coherent correlation-type receiver and integrator converts the UWB pulses to a baseband digital signal that has a bandwidth commensurate with the data rate. The correlation operation and subsequent integration filtering provide significant processing gain, which is effective against interference and jamming. Time coding of the pulses allows for channelization, while the time dithering, pulse position, and signal polarity provide the modulation. UWB systems built around this technique and operating at very low RF power levels have demonstrated very impressive short- and long-range data links, positioning measurements accurate to within a few

centimeters, and high-performance through-wall motion sensing radars.

DS-UWB uses high duty-cycle phase-coded sequences of wideband impulses transmitted at gigahertz rates. Sequences of tens to thousands of impulse "chips" encode data bits in scalable data rates from a one to hundreds of Mbps (megabits per second). The modulation is by pulse polarity and resembles a baseband binary phase-shift-keyed (BPSK) CDMA system with the chipping rate commensurate with the center frequency. The PN (pseudonoise) encoding per data bit provides a measure of multipath delay spread tolerance, allows for channelization, and provides processing gain against interferers. A direct sequence-type of receiver can be used to correlate with the PN code and convert the integrated impulses to data rate bandwidths.

TRD-UWB employs impulse pairs that are differentially polarity encoded by the data with the transmitted pulse pairs having a precise spacing D. The receiver comprises a correlator with one input fed directly and another input delayed by D. It is similar to a conventional differential phase-shift-keyed (DPSK) system, except that rather than integrating over a bit time, here the integration time is commensurate with multipath decay time. The differentially encoded delayed-reference impulse and its data impulse are affected in the same way by multipath. Hence, the delayed-reference detection and integration operation can be made to behave like a near-perfect RAKE receiver capturing a large percentage of the multipath-induced signal echoes.

3. UWB TECHNOLOGY BASICS

UWB spectra can be generated in several different ways, such as by TM-UWB, which uses low duty-cycle impulses; by DS-UWB, which uses high-duty-cycle waveforms that are direct-sequence phase-modulated; and by coded pulse pairs in TRD-UWB. Ultra-short-impulse waveforms are common to both technologies.

The monocycle waveform applied to the transmitting antenna, and represented in Fig. 1 along with its frequency spectrum, is the most basic element of UWB signaling. A useful analytic representation of the monopulse wave form is given by $p(t)$, the first time derivative of a Gaussian monocycle pulse

$$p(t) = -2\pi f_c t \exp\left[\tfrac{1}{2}\{1 - (2\pi f_c t)^2\}\right] \tag{1}$$

where the center frequency is f_c. The spectrum of $p(t)$ is given by

$$P(f) = \frac{f}{f_c} \exp\left[\frac{1}{2}\left[1 - \left(\frac{f}{f_c}\right)^2\right]\right] \tag{2}$$

Actual radiated waveforms and spectra, like the E field shown in Fig. 1, as well as the received waveform and spectra, are further shaped by the bandpass and transient response characteristics of the transmitting antenna. When the transmitting antenna has a wideband and linear phase response, the radiated waveform approximately resembles the time derivative of the signal supplied to the transmitting antenna. The waveform and its spectrum change again in the receiving antenna load, reflecting the transient impulse response of the entire UWB radio link.

If the pulses had been sent at a regular interval without PN encoding, the resulting spectrum would contain "comb lines" separated by the inverse of the pulse repetition rate. The resulting peak power in the comb lines would limit the total transmit power undesirably, as measured in any 1-MHz bandwidth. To make the spectrum more noiselike and provide for channelization in TM-UWB, the monocycle impulses are pseudorandomly placed within each timeframe.

TM-UWB employs PN encoded time dithering to place pulses to picosecond accuracy within a time window equal to the inverse of the average pulse repetition rate.

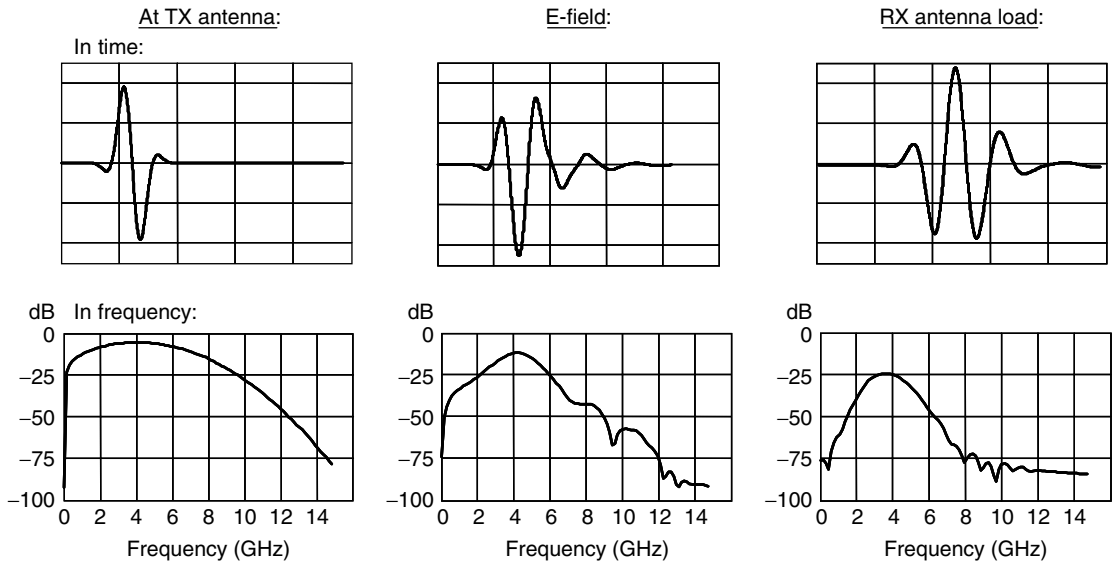


Figure 1. Source, emitted, and received monopulses. (After Ref. 12.)

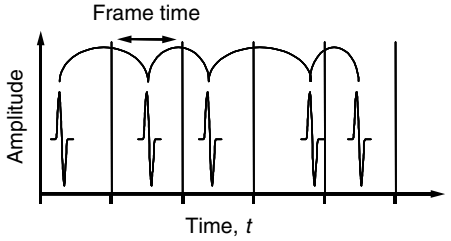


Figure 2. PN-coded UWB waveform sequence in time.

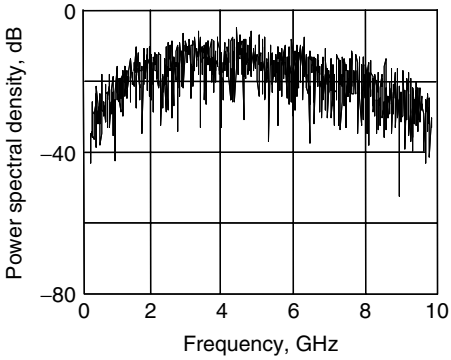


Figure 3. PN coded UWB waveform sequence in frequency.

Figure 2 illustrates a "pulsetrain" that has been PN-time-coded, and Fig. 3 shows the resulting noiselike frequency spectrum. The PN coding uses a pseudorandom timeshift within each time frame. DS-UWB, on the other hand, uses PN codes to polarity-modulate pulse sequences that are closely spaced and at regular intervals. TRD-UWB uses precisely spaced pulse pairs that are polarity-modulated. The resulting spectra of DS-UWB and TRD-UWB are similar to those of TM-UWB.

3.1. TM-UWB Technology

TM-UWB transmitters emit ultrashort monocycle waveforms with tightly controlled pulse-to-pulse intervals. The waveform pulsewidths are typically between 0.2 and 1 ns, corresponding to center frequencies between 5 and 1 GHz, with pulse-to-pulse intervals of 25–1000 ns. The systems typically use pulse position and polarity modulation. The pulse-to-pulse interval is varied on a pulse-by-pulse basis in accordance with two components: an information signal and a channel code. The TM-UWB receiver directly converts the received RF signal into a baseband digital or analog output signal. A front-end correlator coherently converts the electromagnetic pulsetrain to a baseband signal in one stage. There is no intermediate-frequency stage, greatly reducing complexity. A single bit of information may spread over multiple monocycles, providing a way of scaling the energy content of a data bit with the data rate. The receiver coherently sums the proper number of pulses to recover the transmitted information.

TM-UWB systems use a fine pulseshift modulation by positioning the pulse one quarter-cycle (60 ps for a 240-ps pulse) early or late relative to the nominal PN-coded location, or by pulse polarity. Furthermore, multilevel pulse position modulation may be used to provide enhanced bit-energy-to-noise ratio performance. The error probability P_F of the fine-shift modulation in additive white Gaussian noise (AWGN) follows the same behavior as conventional orthogonal or on/off keying (OOK)

$$P_F = \frac{1}{2}\text{erfc}\left(\sqrt{\frac{\gamma_b}{2}}\right) \tag{3}$$

where γ_b is the received signal-to-noise ratio (SNR) per information bit. The error probability P_P of pulse polarity modulation in AWGN follows the same behavior as that of conventional BPSK or antipodal signaling

$$P_p = \tfrac{1}{2}\text{erfc}(\sqrt{\gamma_b}) \tag{4}$$

where γ_b is the SNR per information bit. Pulses may also be transmitted in a "one of many positions" *M-ary* pulse position modulation, which, if the impulse positions do not overlap, resembles the performance of conventional *M-ary* orthogonal signaling in AWGN. The probability of a symbol error is

$$P_m = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[1-\left(1-\frac{1}{2}\text{erfc}\left(\frac{y}{\sqrt{2}}\right)\right)^{M-1}\right] \exp\left[\frac{-(y-\sqrt{2\gamma})^2}{2}\right]dy \tag{5}$$

where $\gamma = \gamma_b \log(M)/\log(2)$ and γ_b is the received SNR per information bit, and average bit error probability P_M is then

$$P_M = P_m\frac{M}{2(M-1)} \tag{6}$$

Modulation further "smooths" the signal spectrum, thus making the signals more noiselike. The probability of a bit error as a function of SNR per bit for the various modulations is depicted in Fig. 4. See also Ref. 13 for derivations of P_F, P_P, and P_M.

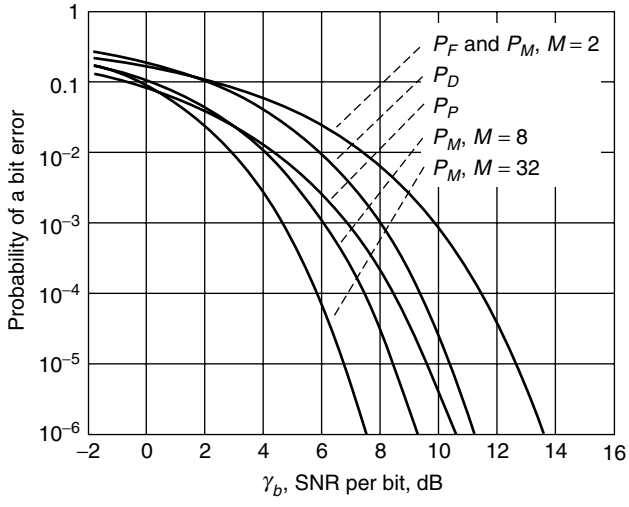


Figure 4. Probability of error for various UWB modulations.

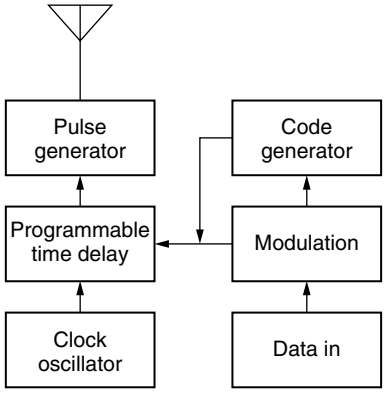


Figure 5. A TM-UWB transmitter.

3.2. A TM-UWB Transmitter

Figure 5 shows a high-level block diagram of a TM-UWB transmitter. The transmitter has no power amplifier, but rather, pulses are generated at the required power. A precision-programmable delay implements the PN time coding and both fine and *M-ary* pulse/time position modulation. Alternatively or in addition, modulation can be encoded in pulse polarity. The precise timing capability of the timer operation (several picoseconds resolution) enables not only precise time modulation and precise PN encoding, but also precision distance determination. The picosecond precision timer, implemented in an integrated circuit, is a key technological component of the TM-UWB system.

3.3. A TM-UWB Receiver

The receiver shown in Fig. 6 resembles the transmitter, except that the pulse generator feeds the multiplier within the correlator. The performance of this type of correlator receiver is described in Ref. 12. Baseband signal processing extracts the modulation and controls signal acquisition and tracking. Baseband signal processing also drives a tracking loop that locks onto the time-coded sequence. Modulation is decoded as either an "early" or "late" pulse in time modulation and/or as a positive or negative pulse in polarity modulation. Different PN time codes are used for channelization. Precise pulse timing inherently enables exceptional positioning and location capabilities in TM-UWB communications systems.

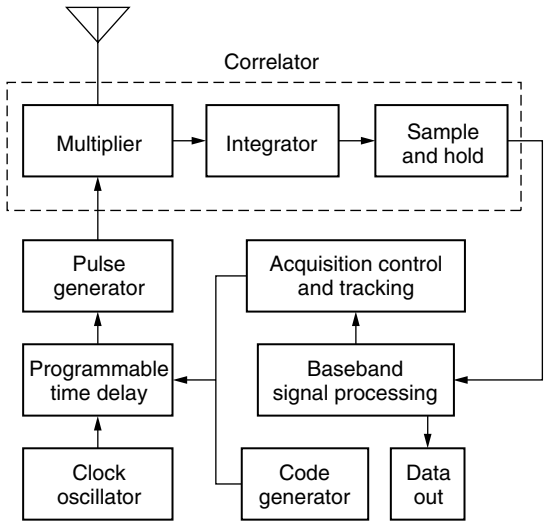


Figure 6. A TM-UWB receiver.

3.4. DS-UWB Technology

A second method of generating useful signals having UWB spectra represents a DS-UWB approach, similar to an RF carrier-based CDMA system. Impulse sequences at duty cycles approaching that of a sine-wave carrier are direct-sequence polarity-modulated (like binary-phase-shift-keying). The PN sequence provides smoothing, channelization and modulation. The chipping rate is some fraction $1/N$ (N need not be an integer) of the "carrier" center frequency. For illustration, Fig. 7 shows the approximate spectral envelope of a 4 GHz impulse sequence that is DS modulated by a zero mean PN code for the cases $N = 1$ and $N = 2$. Actual PN sequences are relatively short and the spectra contain more features, as depicted in Fig. 3. Both signals in Fig. 7 have the same power in a 1-MHz bandwidth at 4 GHz, but the $N = 1$ signal carries the greater total power in the spectrum. The total power and occupied bandwidth can be traded off subject to regulatory emissions limits.

3.5. TRD-UWB Technology

A method of transmitting and receiving impulses that can implement a near-perfect RAKE receiver is exemplified by TRD-UWB and described in Ref. 14. The method employs differentially encoded impulse pairs sent at a precise spacing D. The system is shown in the simplified block diagram of Fig. 8. The transmitter sends a pair of pulses separated by a delay D, and differentially encoded by pulse polarity. The pulses, including propagation induced multipath replicas, are received and detected using a correlator with one input fed directly and another input delayed by D. The receiver resembles a conventional DPSK receiver, which in AWGN exhibits an error probability P_D [13] of

$$P_D = \frac{1}{2}\exp\left(-\gamma_b \frac{N-1}{N}\right) \tag{7}$$

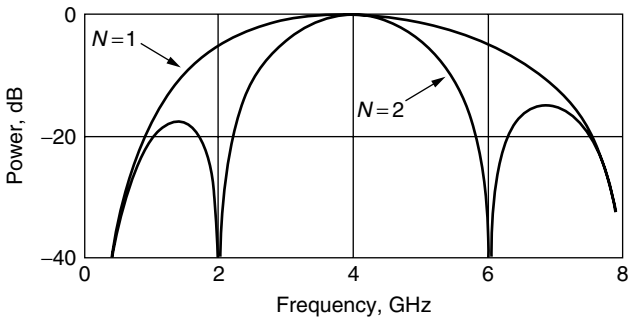


Figure 7. Spectral envelope of DS-UWB signals with $N = 1$ and $N = 2$.

where $N > 1$ is the number of differentially encoded pulses in a sequence and γ_b is the SNR per bit. The integration interval is sufficiently long to RAKE in a significant amount of the multipath energy. The TRD-UWB receiver tends to behave like a near-perfect RAKE receiver capturing a large percentage of the multipath induced signal echoes.

One channelization method employing TRD-UWB [14] has $N = 2$ and employs a family of delays D_i. Impulse pair sequences of these delay combinations constitute the channels. Figure 4 compares the error probability P_D of TRD-UWB with N arbitrarily large to the performance of other modulations.

4. UWB SIGNAL PROPAGATION

Some UWB techniques thrive in multipath, enabling positioning accuracies to better than a few centimeters, and generally follow a free-space propagation law [15]. Further indoor channel characteristics are described elsewhere [16,17]. Multipath fading, characteristic in conventional RF communications, is the result of coherent interaction of sinusoidal signals arriving by many paths. Spread spectrum TIA/EIA-95-B cellular and PCS systems with a 1.228-MHz spreading bandwidth can resolve multipath signals having differential delays of slightly less than one microsecond. Some communications channels, particularly outdoors, can have rms delay spreads measuring many microseconds; therefore, some multipath components can be resolved and received using RAKE techniques. However, in-building communications channels exhibit multipath differential delays and rms delay spreads in the several tens of nanoseconds as seen in Fig. 9 and cannot be resolved in the relatively narrow TIA/EIA-95-B channel. Those systems must therefore contend with significant Rayleigh fading, which may require signals up to tens of decibels above the static signal level for a given measure of performance.

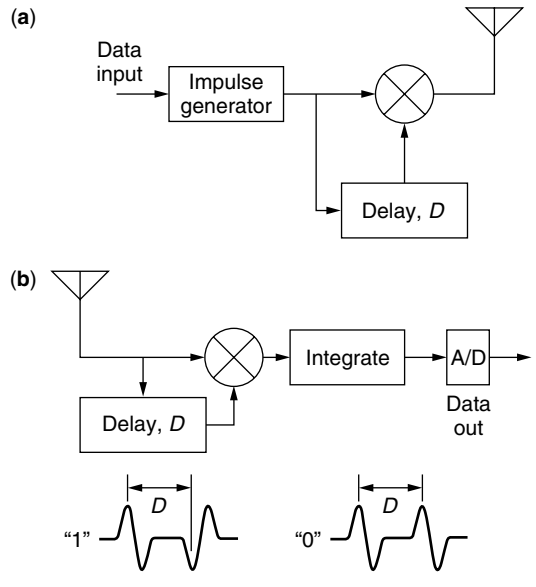


Figure 8. A TRD-UWB transmitter (**a**) and receiver (**b**).

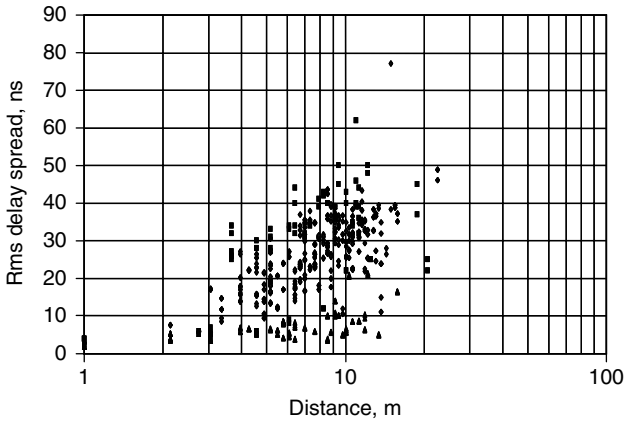


Figure 9. Measured RMS delay spread in small offices and homes.

4.1. In-Building Propagation of Impulses

UWB signal propagation in free space between two unity gain antennas separated by d is very nearly

$$P_L = 20 \log \left(\frac{c}{4\pi d f_c} \right) \tag{8}$$

where c is the velocity of light and f_c is the center frequency of the emitted spectrum. The frequency dependence of propagation comes from the frequency dependence of a unity-gain receive antenna aperture area $A_e = c^2/(4\pi f_c^2)$ evaluated here at f_c. Equation (8) is only approximately correct for the large bandwidth UWB signal [12], because total received power involves an integral in frequency over the product of the received power spectral density and A_e. A typical UWB impulse subjected to multipath is shown in Fig. 10. The multipath is evident as delayed echoes of the first-arriving impulse adding to the signal voltage. This measurement, and subsequent propagation measurements were gathered using Time Domain Corporation's *pulson application demonstrator* (PAD) radios, which have a built-in waveform scanning mode capable of resolving impulses to a fraction of a nanosecond.

Signal measurements in several multipath office and home environments using PADs were processed to determine the strongest impulse in a waveform versus distance and are shown in Fig. 11. The signal attenuation is shown relative to the $d = 1$ meter signal level. The median attenuation with distance is approximately 29 $\log(d)$ with the distance d in meters, and is typical of what can also be expected for conventional narrowband channels.

The measurements reprocessed to "perfectly RAKE" the total power in all of the multipath impulses waveform echoes, are shown in Fig. 12. The total power median attenuation with distance follows a square law, 20 $\log(d)$. A "perfect RAKE" receiver can provide up to the limit of the difference between the strongest impulse and total

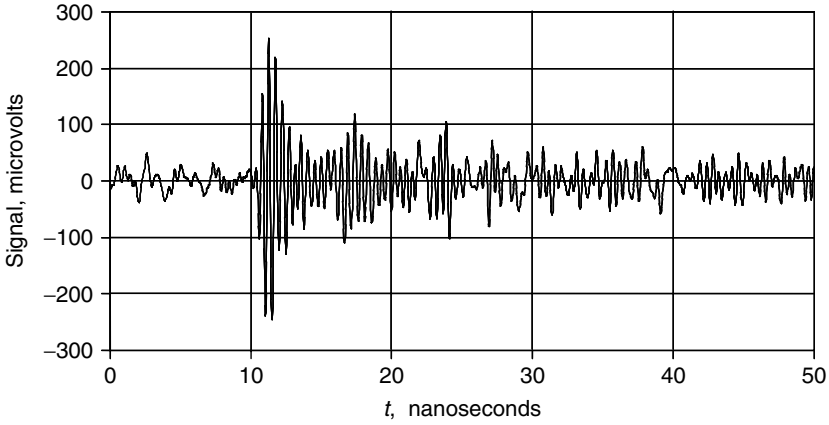


Figure 10. Typical measured UWB received signal in low multipath.

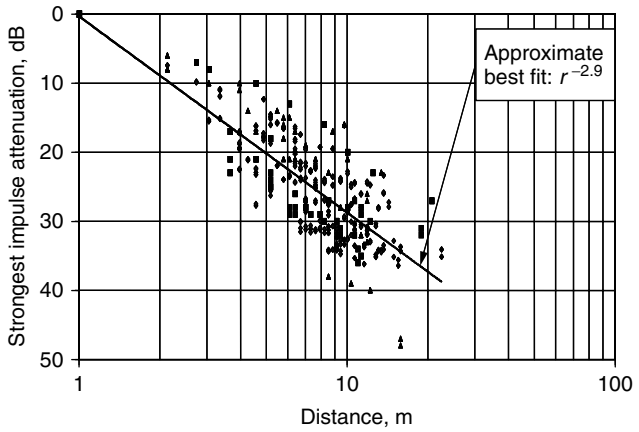


Figure 11. Strongest received UWB impulse versus distance. (After Ref. 12.)

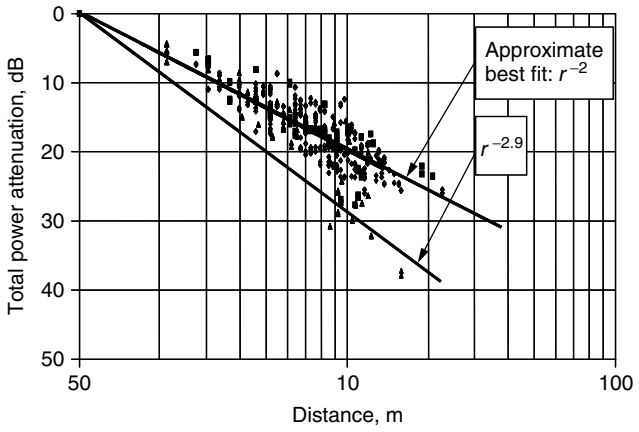


Figure 12. Total UWB impulse power versus distance. (After Ref. 12.)

power, or approximately an average 9 log(d) of RAKE gain for the measured environments considered.

With perfect RAKE gain the signal in multipath follows a near-free-space propagation law, and represents one of the special benefits of UWB impulse technology.

4.2. Impulse Propagation with a Ground Reflection

Propagation over a smooth earth involves a reflection from the ground (Fig. 13). The direct path and reflected pathlengths D and R in terms of the antenna heights H_1 and H_2, and the separation distance d are

$$D = \sqrt{d^2 + (H_1 - H_2)^2} \tag{9}$$

and

$$R = \sqrt{d^2 + (H_1 + H_2)^2} \tag{10}$$

See Ref. 18 for details. The differential delay between the reflected path and the direct path over a plane earth is

$$\Delta t = \frac{R - D}{c} \tag{11}$$

where c is the velocity of light.

The ground reflection coefficient for the cases of interest (shallow incidence angles) is very nearly -1, so the reflected pulse undergoes a polarity inversion. Reflected pulses that arrive by paths having differential delays greater than a half pulselength, as portrayed in Fig. 14, add to the total received energy. Pulses with a differential

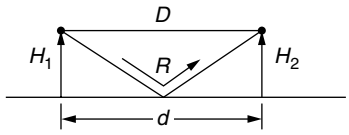


Figure 13. Geometry for two-path propagation.

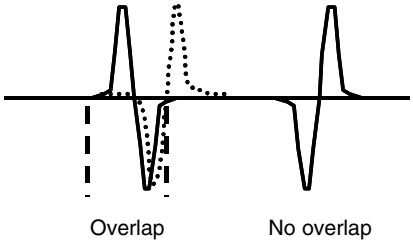


Figure 14. Overlapping and nonoverlapping pulses.

delay of *less* than half a pulselength begin to exhibit destructive interference in the receive window.

The overlapping pulse echo arriving by ground reflection is not delayed enough to be distinct from the directly arriving pulse. The nonoverlapping pulse is distinct, and adds to the received energy if a RAKE receiver is employed. With RAKE gain, the bold line in Fig. 15 in the "no-overlap region" would be raised by 3 dB. Overlapping pulses, as seen in Fig. 15, at first add constructively when the overlap is less than a half pulselength, then when nearly fully overlapping exhibit an inverse 4th power with distance behavior similar to harmonic wave propagation. In contrast, harmonic waves exhibit multiple constructive and destructive interferences as multiple sinusoidal cycle delays interact at close distances.

4.3. Reception of UWB Impulses

UWB signals are detected in a correlation-type receiver, (Figs. 6 and 8). A filter with impulse response $h(t)$ is optionally placed between signal $s(t)$ at the receiver antenna load and the correlator input. The correlation template pulse $p(t)$, locally generated in Fig. 6 and derived from the transmitted reference pulse in Fig. 8, multiplies the received data pulse and is integrated and sampled at the correlator output. The receiver implementation efficiency e_c [12] of this operation is

$$e_c = 10\log\left[\frac{\left|\int\int s(\tau)h(\tau-t)d\tau p(t)dt\right|^2}{\int s(t)^2dt\int\left|\int p(\tau)h(\tau-t)d\tau\right|^2dt}\right] \tag{12}$$

and is maximized when

$$C\int p(\tau)h(\tau-t)d\tau = s(t) \tag{13}$$

provided $h(t)$ is causal and where C is the rms value of $s(t)$. Solutions to Eq. (13) range from the matched template, $h(t)=\delta(t)$ [the Dirac delta function] with $p(t)=s(t)$, to the matched filter, $h(t)=s(-t)$ with $p(t)=\delta(t)$, see (13). The correlator efficiency depends strongly on the shape of the signal $s(t)$ and its relationship to $h(t)$ and $p(t)$. The efficiency e_c can typically range from -6 to -2 dB for simple rectangular templates, see (12), to a RAKE gain of several decibels for a delayed reference receiver.

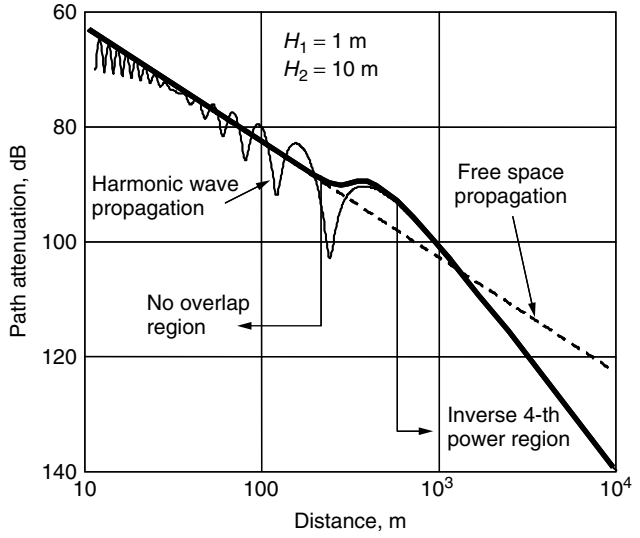


Figure 15. Impulse (bold) and harmonic wave propagation near ground.

4.4. A UWB Link Budget

The UWB link specifies a transmitter and antenna providing an EIRP of P_{Tx}, a receiver with sensitivity S_{Rx}, and because the waveform changes shape in the link, a propagation factor P_L determined at a convenient distance. A transmitter operating with about 2 dB margin to a -41.25-dBm/MHz limit over an equivalent bandwidth of 1.3 GHz emits $P_{\mathrm{Tx}} = -12$ dBm. A companion UWB receiver operating in AWGN (-174 dBm/Hz), referenced to a data bandwidth W Hz, with noise figure, implementation loss and margin totaling L dB, and operating at an SNR dB signal-to-noise ratio per bit, has a sensitivity of

$$S_{\mathrm{Rx}} = -174 + 10\log(W) + L + \mathrm{SNR\ dBm} \tag{14}$$

The propagation term P_L is evaluated conveniently at one meter using Eq. (8) and the resulting system gain SG at a one meter distance including a receiver antenna gain of G_{Rx} dBi is

$$\mathrm{SG} = P_{\mathrm{Tx}} - S_{\mathrm{Rx}} + P_L + G_{\mathrm{Rx}}\mathrm{dB} \tag{15}$$

When $f_c = 4$ GHz, then $P_L = -44$ dB, and using a data bandwidth of $W = 40$ MHz, losses $L = 10$ dB, SNR = 7 dB, and $G_{\mathrm{Rx}} = 5$ dBi, the receiver sensitivity is $S_{\mathrm{Rx}} = -69$ dBm, and the system gain at one meter is SG = 30 dB.

Referring to Fig. 11, a 30-dB system gain permits a median range of approximately 10.8 m without RAKE gain. From Fig. 12, the "perfect RAKE" receiver would permit a median range of more than 31 m. Practical implementations would result in a range performance between 10 and 30 m at a 40-Mbps throughput data rate.

5. APPLICATIONS OF UWB

UWB technology uniquely harnesses an ultrawideband of spectrum to provide high bandwidth communications, but in certain implementations also enables indoor precision tracking and radar sensing on top of communications. The unique capabilities of UWB driving it into applications spaces markets include

High Spatial Capacity. The number of impulses that can be discerned in time over the propagation distance ultimately limits UWB channelization and bandwidth per square meter. For example, with 0.25-ns impulses, the upper limit on pulse rate is 4 billion pulses per second. Because of low emitted power ranges are confined to several tens of meters, thus providing exceptional spatial capacities.

High Channel Capacity and Scalability. Scalability accommodates various channel profiles to harness

the desired data rate given a channel impulse response. Phase coding of impulses simultaneously integrates impulses to improve the energy per data bit, and can RAKE energy from multipath delayed signal echoes.

Robust Multipath Performance. Multipath signal echoes can be RAKE-received for superior performance indoors. In the limit, the total impulse power on average propagates with an inverse square law just as in free space, in contrast to conventional narrowband harmonic wave radios which tend to propagate more nearly like inverse 3rd power indoors.

Very Low Transmit Power. Submilliwatt power levels spread over several gigahertz of bandwidth means that the UWB signals will not cause harmful interference to current users of the spectrum, and also will generally be stealthy and less susceptible to detection.

High System Link Rate. Data rates can be from a high in the hundreds of Mbps, a goal of communications standards developers [19], down to hundreds of kbps. Given a fixed power level, the data rate may be traded off for additional range.

Location Awareness and Tracking. Some implementations of UWB signaling inherently provide 3D sensing and tracking at centimeter accuracies [17].

5.1. The Role of UWB in Wireless Markets

Advancements in UWB radio technology promise the opportunity of creating unique solutions meeting emerging market needs. There are essentially three basic market spaces in which UWB plays a role:

1. *Wireless Communications*. As available bandwidth to users increases, applications will continue to evolve to fill the available bandwidth and demand further increases. On top of this increasing demand for bandwidth, the increase in mobile telephony and travel has spurred demand for bandwidth mobility, implying wireless technology. Initial applications of UWB will evolve from the existing market needs for higher speed data transmission, but demand for multimedia-capable wireless is already driving multiple initiatives in the wireless standards bodies. UWB solutions will emerge that are tailored for these applications because of the available high bandwidth. In particular, high-density multimedia applications, such as multimedia streaming in "hotspots" such as airports or shopping centers or even in multidwelling units, will require bandwidths not currently enabled by continuous-wave "narrowband" technologies. The ability to tightly pack high bandwidth UWB "cells" into these areas without degrading performance will further drive the development of UWB solutions. Full-duplex and simplex radio systems using submilliwatt power levels have already been demonstrated with data rates from 78 kbps to hundreds of Mbps at useful ranges in home and office environments.

2. *Precision Tracking*. As the mobility of people and objects increases, up-to-date and precise information about their location becomes a relevant market need. While GPS and some E911 technologies promise to deliver some level of accuracy outdoors, current indoor tracking technologies remain relatively scarce and have accuracies on the order of 3–10 m. UWB implementations are an adjunct to GPS and E911 that allow the precise determination of location and the tracking of moving objects within an indoor space to an accuracy of a few centimeters. This in turn enables the delivery of location-specific content and information to individuals on the move, and the tracking of high-value assets for security and efficient utilization. While this is an emerging market segment, the accuracy provided by UWB will accelerate market growth and the development of new applications in this area.

3. *Radar*. Finally, UWB signals enable inexpensive short range high-definition radar. With the new radar capability created by the addition of UWB, the radar market will grow dramatically and radar will be used in areas currently unthinkable. Some of the key new radar applications where UWB is likely to have a strong impact include automotive sensors, collision avoidance sensors, smart airbags, intelligent highway initiatives, personal security sensors, precision surveying, and through-the-wall public safety applications. Through-wall radar is already being tested to assist law enforcement and public safety personnel in clearing and securing buildings more quickly and with less risk by providing the capability to detect human presence and movement through walls. Radar enhanced security domes based on precision radar have already demonstrated the capability to detect motion near protected areas, such as high value assets, personnel, or restricted areas. The dome is software configurable to detect movement passing through the edge of the dome, but can disregard movement within or beyond the dome edge.

5.2. Addressing the Wireless Spectrum Squeeze with UWB

UWB operates at ultra-low-power, transmitting impulses over multiple gigahertz of bandwidth. Each pulse, or pulse sequence, is pseudorandomly modulated, thus appearing as "white noise" in the "noise floor" of other radiofrequency devices. UWB operates with emission levels commensurate with levels of unintentional emissions from common digital devices such laptop computers and pocket calculators. Today we have a "spectrum drought" in which there is a finite amount of available spectrum, yet there is a rapidly increasing demand for spectrum to accommodate new commercial wireless services. Even the defense community continues to find itself defending its spectrum allocations from the competing demands of commercial users and other government users. UWB exhibits incredible spectral efficiency that takes advantage of underutilized spectrum, effectively creating a "new" spectrum for existing and future services by making productive use of what appears as the "noise floor" in conventional receiver bandwidths. UWB technology represents a win–win innovation that makes available a critical spectrum to government, public safety, and commercial users.

The best applications for UWB are for indoor use in high-clutter environments. UWB products for the commercial market will make use of the most recent technological advancements in receiver design and will

transmit at very low power (submilliwatts). UWB technology enables not only communications devices but also positioning capabilities of exceptional performance. The fusion of positioning and data capabilities in a single technology opens the door to exciting and new technological developments.

BIOGRAPHIES

Kazimierz "Kai" Siwiak received his B.S.E.E. and M.S.E.E. degrees from the Polytechnic Institute of Brooklyn and his Ph.D. from Florida Atlantic University, Boca Raton, Florida. He designed radomes and phased array antennas at Raytheon before joining Motorola, where he received the Dan Noble Fellow Award for his research in antennas, propagation, and advanced communications systems. In 2000, he joined Time Domain Corporation to lead strategic technology development. He has lectured and published internationally; and holds more than 70 patents worldwide, including 31 issued in the United States. He was awarded Paper of the Year by IEEE–VTS and has authored, *Radiowave Propagation and Antennas for Personal Communications*, (Artech House), now in second edition, and contributed chapters to several other books and encyclopedias.

Laura L. Huckabee received her B.A. degree in physical chemistry in 1986 from Princeton University, her M.A. in comparative culture in 1993 from Sophia University, Tokyo, Japan, and her M.B.A from INSEAD, Fontainebleau, France, in 1994. Her technical work focused on laser spectroscopy and solid state physics. She consulted with U.S. and European firms trying to enter the Japanese market from 1987 through 1993 and joined Procter and Gamble in 1995. At P&G, she developed new businesses in Asia and the Middle East, and moved to the Coca-Cola Company in 1997, developing the Polish soft drink business. In 2000, she returned to the United States to join Time Domain Corporation, the pioneer in UWB radio chipsets, leading international business development and strategic partners.

BIBLIOGRAPHY

1. K. Siwiak and L. L. Huckabee, An introduction to ultra-wide band wireless technology, in B. Bing, ed., *Wireless Local Area Networks — the New Wireless Revolution*, Wiley, New York, 2001.
2. R. A. Scholtz and M. Z. Win, Impulse radio, *Proc. IEEE Personal, Indoor and Mobile Radio Communications, PIMRC 1997*, Helsinki, Finland, 1997.
3. M. Z. Win and R. A. Scholtz, Impulse radio: How it works, *IEEE Commun. Lett.* **2**(1): (Jan. 1998).
4. Time Domain Corporation, Huntsville, AL (online): <*http://www.timedomain.com*> (Nov. 5, 2001).
5. K. Siwiak, Ultra-wide band radio: Introducing a new technology, invited plenary paper, *Proc. IEEE Vehicular Technology Conf. 2001*, Rhodes, Greece, May 2001.
6. H. L. Bertoni, L. Carin, and L. B. Felson, eds., *Ultra-Wideband Short-Pulse Electromagnetics*, Plenum Press, New York, 1993.
7. L. Carin and L. B. Felson, eds., *Ultra-Wideband Short-Pulse Electromagnetics*, Vol. 2, Plenum Press, New York, 1995.
8. C. E. Baum, L. Carin, and A. P. Stone, eds., *Ultra-Wideband Short-Pulse Electromagnetics*, Vol. 3, Plenum Press, New York, 1997.
9. E. Heyman, B. Mandelbaum, and J. Shiloh, eds., *Ultra-Wideband Short-Pulse Electromagnetics*, Vol. 4, Plenum Press, New York, 1999.
10. M. Welborn, System considerations for ultra-wideband wireless networks, *Proc. IEEE Radio and Wireless Conf., RAWCON 2001*, Boston, MA, Aug. 19–22, 2001, pp. 5–8.
11. H. G. Schantz and L. Fullerton, The diamond dipole: A Gaussian impulse antenna, *IEEE APS Conf.*, Boston, July 2001.
12. K. Siwiak, T. M. Babij, and Z. Yang, FDTD simulations of ultra-wideband impulse transmissions, *Proc. IEEE Radio and Wireless Conf., RAWCON 2001*, Boston, Aug. 19–22, 2001.
13. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1983.
14. R. Hoctor, Transmitted-reference, delay-hopped Ultra-Wideband Communications, *Forum on Ultra-Wide Band*, Hillsboro, OR (online): <*http://www.ieee.or.com/IEEEProgram Committee/uwb/uwb.html*> (Oct. 11–12, 2001).
15. K. Siwiak and A. Petroff, A path link model for ultra-wide band pulse transmissions, *Proc. IEEE Vehicular Technology Conf. 2001*, Rhodes, Greece, May 2001.
16. J. Foerster, The effects of multipath interference on UWB performance in an indoor wireless channel, *Proc. IEEE Vehicular Technology Conf. 2001*, Rhodes, Greece, May 2001.
17. K. Siwiak, The future of UWB — a fusion of high capacity wireless with precision tracking, *Forum on Ultra-Wide Band*, Hillsboro, OR (online): <*http://www.ieee.or.com/IEEEProgramCommittee/uwb/uwb.html*> (Oct. 11–12, 2001).
18. K. Siwiak, *Radiowave Propagation and Antennas for Personal Communications*, 2nd ed., Artech House, Norwood, MA, 1998.
19. IEEE P802.15.3, High Rate (HR) Task Group (TG3) for Wireless Personal Area Networks (WPANs), (online): <*http://grouper.ieee.org/ groups/ 802/15/pub/TG3.html*> (Dec. 1, 2001).

UNEQUAL ERROR PROTECTION CODES

ARDA AKSU
North Carolina State University
Raleigh, North Carolina

1. INTRODUCTION AND THEORY

The error correcting capability of the majority of algebraic codes is described in terms of correcting errors in codewords, rather than correcting errors in individual digits. However, in many applications some message positions are more important than others. For instance, in transmitting numerical data, errors in the high-order digits are more serious than errors in the low-order digits. Therefore each block of data can be partitioned into classes of different importance (i.e., of different sensitivity to errors).

It is possible for a linear block code to provide more protection for selected positions in the input message words than is guaranteed by the minimum distance of the code. Then, it is apparent that the best coding strategy aims at achieving lower bit error rate (BER) levels for more important information bits while admitting higher BER levels for the less important ones. This feature is referred to as unequal error protection (UEP) and linear codes having this property are called *linear unequal error protecting* (LUEP) codes.

UEP codes were first studied by Masnick and Wolf in 1967 [1]. Since then, there has been a proliferation of studies in the field of unequal error protection codes, in both the theory and practical applications. Later work [2–7] investigated various approaches to the construction of UEP codes. A separation vector was later [3], introduced as a measure of the error correcting capability of an UEP code at various locations. For a binary linear (n, k) code C, with generator matrix G, the separation vector $\underline{s} = \{s_1, s_2, \ldots, s_k\}$ is defined by

$$s_i = \min\{w(\underline{m}G) \mid \underline{m} \in \{0, 1\}^k, m_i \neq 0\} \quad i = 1, \ldots, k$$

where $w(\cdot)$ denotes the Hamming weight of the argument, namely, the number of nonzero components in the argument.

Another way of looking at it is that the sets $\{\underline{m}G \mid \underline{m} \in \{0, 1\}^k, m_i = 0\}$ and $\{\underline{m}G \mid \underline{m} \in \{0, 1\}^k, m_i = 1\}$ are at a Hamming distance s_i apart. Hence, for a linear binary (n, k) code C that uses a matrix G for its encoding, complete nearest-neighbor decoding guarantees the correct interpretation of the ith digit whenever the error pattern has a Hamming weight less than or equal to $\lfloor (s_i - 1)/2 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer contained in x. From this, it is immediately clear that the minimum distance of the code is $d_{\min} = \min\{s_i \mid i = 1, \ldots, k\}$. Therefore, if a linear code C has a generator matrix G such that the components of the separation vector are not equal, then the code C is called a *linear unequal error protecting* (LUEP) code.

Usually, decoding algorithms for UEP codes are complicated. It has been shown [1,4] that a modified syndrome decoding method using a standard array can be implemented for UEP codes. Essentially, if efficient decoding procedures are available for component codes, then the resulting UEP code can be decoded too. But it is still necessary to design UEP codes, which can be implemented easily.

A binary cyclic code $C(n, k)$ is the direct sum of a number of ideals in the residue class ring $\mathrm{GF}(2)[x]/(x^n - 1)$ (where GF is the Galois field) of polynomials in x. In [4], it is shown that an ordering $M_1, M_2, \ldots, M_v$ of generator matrices of these ideals exists such that

$$G = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_v \end{bmatrix}$$

is an optimal generator matrix. The ith and jth components of the separation vector $(\underline{s})$ are equal if the ith and jth rows of G are in the same ideal of $\mathrm{GF}(2)[x]/(x^n - 1)$.

If the weight of the generator polynomial of a cyclic code C equals $d_{\min}$ of the code, then all components of the separation vector are equal. If this is not the case, the separation vector of a cyclic code can be computed by comparing the weight distributions of its cyclic subcodes. In Van Gils [4], compiled a table for UEP capabilities of binary cyclic codes of odd lengths up to 39. Later, Lin et al. [7], extended the table to lengths up to 65 by using exhaustive computer search.

Many of the best known codes can be constructed as generalized concatenated (GC) codes. The construction of these codes use outer codes with different lengths and inner codes (in the columns of the code matrix) with different lengths and distances. The inner code is multiply partitioned, and this partitioning into subcodes is protected by different outer codes. With this method, a large class of optimal linear UEP codes can be generated ([17]). These codes also contain most of the constructions found in van Gils' 1984 paper [6].

Others have investigated coding and decoding schemes to achieve UEP using several convolutional codes with different error correcting capabilities [18–20]. Lower bounds on the free distance of convolutional codes with unequal information protection have been investigated [21,22]. The asymptotic behaviors of these bounds indicate that more gains can be attained for the important data by enlarging the corresponding constraint length. This comes at the cost of reduced performance for the less significant data.

It is desirable to design UEP codes, which can be implemented easily. UEP, in conjunction with modulation, can be achieved by employing either *time-division coded modulation* (TDCM) or *superposition coded modulation* (SCM) [8–12]. TDCM is a form of resource sharing in which bit streams of differing importance are transmitted in disjoint modulation intervals. In SCM, the different bit streams are transmitted in the same modulation intervals. Let B_1 and B_2 denote, in decreasing order of importance, the two bit streams to be unequally protected against channel noise, and let r_1 and r_2 denote their respective rates. The rates are given in terms of the bit rate normalized by the total number of modulation intervals available for transmission of these bit streams. To specify unequal error protection requirements, let N_1 and N_2 denote the variances of the Gaussian noise that the respective bit streams need to withstand, where $N_1 > N_2$. In other words, as long as the variance of the channel noise is less than N_i, the bit stream B_i can be decoded with a bit error rate below some prescribed value.

TDCM is a scheme in which the bit streams B_1 and B_2 are transmitted on distinct modulation intervals. Bit stream B_i is transmitted over the fraction α_i of the available modulation intervals (where $\alpha_1 + \alpha_2 = 1$) using channel code C_i and transmission energy e_i. The design problem is usually to select the parameters (α_1, α_2) and (e_1, e_2) such that desired levels of UEP are achieved while minimizing the average transmission energy per modulation interval $e_T = \alpha_1 e_1 + \alpha_2 e_2$. Figure 1 shows a generalized transceiver structure for TDCM.

Superposition coded modulation (SCM) consists of transmitting both bit streams on all the available

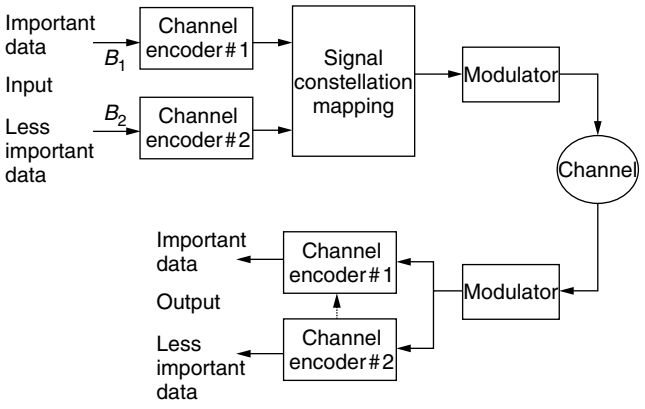


Figure 1. Generalized transceiver structure for TDCM.

modulation intervals using a superposition of channel codes in the modulation space. Let us choose constellations S_1 and S_2 with average energies e_1 and e_2, respectively. Bit stream B_i is encoded with channel code C_i and transmitted using constellation S_i. More specifically, the code C_i generates the codeword $\underline{x}_i$, which is a sequence of signal points $\{x_i\}$, where $x_i \in S_i$. The codewords $\underline{x}_1$ and $\underline{x}_2$ are superimposed and transmitted on the channel as $\underline{x} = \underline{x}_1 + \underline{x}_2$. C_1 and S_1 are respectively referred as *outer code* and *outer constellation* and C_2 and S_2, as inner code and inner constellation. At the decoding end, the received sequence is $\underline{y} = \underline{x} + \underline{n}$, where N is the variance of the Gaussian noise. Then, if $N_2 < N < N_1$, only bit stream B_1 is reliably decoded. If $N < N_2$, then both bit streams are reliably decoded. The SCM design method aims at achieving the prescribed levels of protection for the two bit streams while minimizing the average transmission energy.

For higher rate transmission with multilevel modulation over bandwidth-limited channels, UEP coding can be achieved in the context of combined coding and modulation because of its efficiency compared to time-sharing techniques [13,14]. It has been shown [15] that, asymptotically, the superposition coded modulation technique always outperforms the later one. On the basis of this result, much of the earlier work concentrated on the first technique. Other authors [16] show that this theoretical result doesn't always hold for practical channel codes, which do not achieve capacity. In fact, the opposite happens when a ratio, which measures the degree of inequality in protection, is below a critical threshold.

UEP codes have generated much interest since they are increasingly important in various applications, such as visual communication systems, speech communication, storage and computer systems, and satellite communications. The data in such systems are not equally important, especially after source encoding. Coupled with the proliferation of high-speed wireless networks, which present a very challenging channel for data transfer around the globe, the need for highly efficient UEP coding becomes even more important. The rest of this article focuses on examples of the use of UEP codes in various applications.

2. UEP CODES FOR SPEECH CODING

The trend in current and future cellular mobile radio systems is toward using more sophisticated, digital speech coding techniques. However, mobile radio channels are subject to signal fading and interference, which causes significant transmission errors. The design of speech and channel coding is therefore quite challenging, and UEP codes are considered in the literature for improved performance.

The effects of digital transmission errors on a family of variable-rate embedded subband speech coders have been analyzed [23]. Subband coding of speech is a relatively mature form of waveform coding of speech, in which the signal is first divided into a number of subbands, which are then individually encoded. The underlying principle for the coder is that the bit allocation can be weighed so that those subbands with the most important information get most of the bits. The initial subband coders used fixed bit allocations based on the average spectrum of speech. Later on, the idea of dynamically changing the bit allocation based on the energy of each subband was introduced. An example of the block diagram of the transmitter portion of this coder is shown in Fig. 2. The authors show that there is a difference in error sensitivity of four orders of magnitude between the most and the least sensitive bits of the speech coder. As a result, a family of *rate-compatible punctured convolutional* (RCPC) codes with flexible UEP capabilities, matched to the speech coder, provides significant gains over a wide range of channel signal-to-noise ratios [24].

Perceptual audio coders (PACs) and similar audio compression techniques are inherently packet-oriented. Audio information for a fixed interval (frame) of time is represented by a variable-bit-length packet. Each packet consists of certain control information followed by quantized spectral/subband description of the audio frame. The key idea behind an UEP scheme is that different components of a packet exhibit varying degrees of sensitivity to channel errors. Experimental results on multilevel UEP schemes, which exploit the unequal impact of transmission errors on various audio components, exhibit significant gains and graceful degradation compared to equal error protection [25].

Use of UEP codes have been suggested for speech coding schemes employed in developing (at the time of writing) third-generation (3G) wireless systems [26] and digital audiobroadcasting, mainly because of its superior performance and graceful performance degradation.

Embedded joint source-channel coding schemes of speech also take advantage of UEP by puncturing symbols of the "RCPC Trellis" encoders [27]. The coder is claimed to be robust to acoustic noise and produce good quality speech for a wide range of channel conditions.

3. UEP CODES FOR IMAGE TRANSMISSION

Robust image transmission has received increasing attention with the advances in wireless communication and image coding. Numerous schemes have been developed for transmission of images over noisy channels.

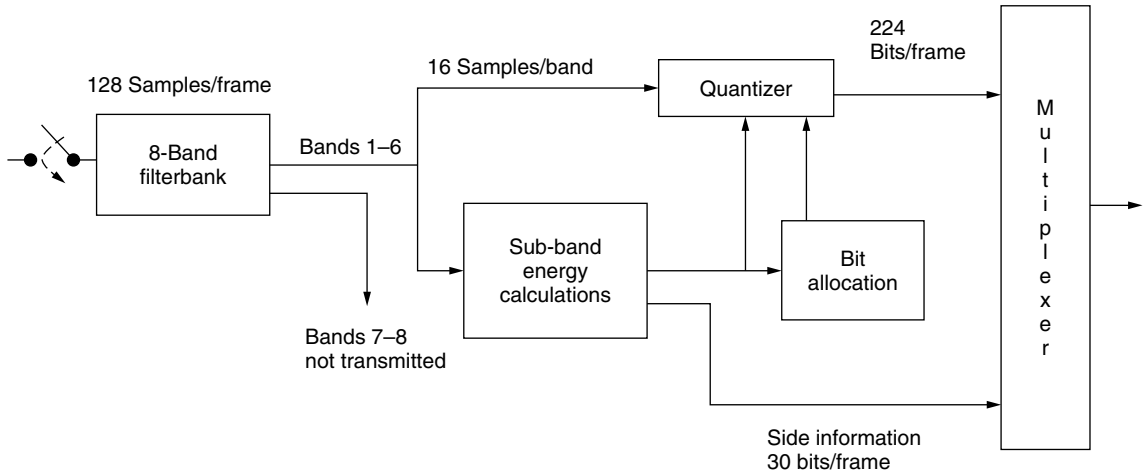


Figure 2. An example of a dynamic bit allocation subband coder.

One such method employs selective channel coding and transmission energy allocation in conjunction with sequence maximum a posteriori soft-decision detection [28]. An UEP scheme is proposed for transmitting discrete cosine transform (DCT) compressed images over an additive white Gaussian noise (AWGN) channel. The system consists of a TCM scheme that uses selective transmission energy allocation to the DCT coefficients. It also employs a sequence maximum a posteriori (MAP) detection scheme that exploits both channel soft-decision information and the statistical image characteristics. Experimental results indicate that this scheme provides substantial objective and subjective improvements over equal-error protection systems as well as graceful performance degradation. Fei and Ko [29] have used UEP codes for JPEG compressed images in a wireless environment in order to provide extra protection for the header syntax of JPEG compressed images.

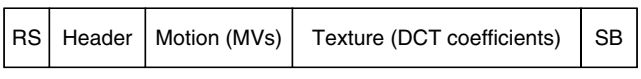
Rate-distortion-based error control schemes are also used widely in wireless image communications. Although error correction codes can effectively protect transmitted bits, the introduced overhead bits decrease the available channel capacity. Among various compression techniques for wireless channel transmission, embedded wavelet coders possess excellent rate distortion performance as well as progressive representation property in the transmitted bit stream. In other words, the quality of the decoded images at the receiver end progressively improves as more bits are received. Furthermore, the same number of bits received at the receiver end inherit different capabilities to increase the quality of decoded images due to the different importance of bits at different locations. Therefore, UEP coding schemes based on a rate-distortion model of embedded wavelet image coders generally provide better performance than do equal-error-protecting schemes [30,31].

Many other schemes investigate the design of the bit allocation between the source coder and channel coder by jointly considering effects of quantization errors and channel errors. Individual bits show different importance for the image reconstruction and different sensitivity to channel errors. Many joint source-channel coding techniques use "subbandwise" UEP, in which the parity bit budget is allocated among subbands proportional to the contained energy [32–34]. This scheme is not optimal since coefficients with large magnitudes at finer scales have more contribution to the distortion reduction than coefficients with small magnitude at coarser scales. Thus "bitplanewise" transmission with UEP is also frequently used in advanced wavelet image codecs [35].

4. UEP CODES FOR VIDEO TRANSMISSION

The standard video compression algorithms (e.g., H.263, MPEG) use predictive coding of frames and variable-length codewords to obtain a large amount of compression. This renders the compressed video bit stream sensitive to channel errors, as predictive coding causes errors in the reconstructed video to propagate in time to future frames of video, and the variable-length codewords cause the decoder to easily lose synchronization with the encoder in the presence of bit errors. To make the compressed bit stream more robust to channel errors, the MPEG-4 video compression standard incorporated several error resilience tools (resynchronization markers, header extension codes, data partitioning, etc.) to enable detection and containment of errors.

However, since wireless channels present quite harsh conditions, powerful error control coding is always needed, and UEP codes present a very good match. The output of an MPEG-4 typical video encoder is a bit stream that contains video packets. These video packets begin with a header, which is followed by the motion information, the texture information, and the stuffing bits (see Fig. 3). The header of each packet begins with a resynchronization



RS: Resynchronization marker
SB: Stuffing bits

Figure 3. An example of a video packet from an MPEG-4 encoder.

marker, which is followed by the important information needed to decode the data bits in the packet. This is the most important information, since the whole packet will be dropped if the header is received in error. The motion information has the next level of information, as motion compensation cannot be performed without it. The texture information is the least important of the four segments of the video packet. Without the texture information, motion compensation concealment can be performed without too much degradation of the reconstructed picture. The stuffing information at the end of the packet has the same priority as the header bits because reversible decoding cannot be performed if this information is corrupted and the following packet may be dropped if the stuffing bits are received in error. When using UEP, the header and stuffing bits normally get the highest amount of protection, the motion bits get the next highest level of protection, and the texture bits receive the lowest level of protection. Using this system, the errors are less likely to occur in the important sections of the video packet. A number of studies are done in order to investigate the performance gains of UEP codes applied to video coding [36,37]. A comparison between using a fixed rate-$\frac{7}{10}$ convolutional code for the whole packet or UEP using rate-$\frac{3}{5}$ convolutional code for the header and stuffing segments, a rate-$\frac{2}{3}$ convolutional code for the motion segment, and a rate-$\frac{3}{4}$ convolutional code for the texture segment, shows gains as much as 1 dB in video quality.

Burlina and Alajaji discuss the UEP capabilities of convolutional codes belonging to the family of RCPC codes applied to image coding [38]. The H.263-compatible datastream is partitioned into classes of different sensitivity, without any modification to the standard. Experimental results show that UEP coding provides very good performance when video is transmitted over channels having high round-trip delay and limited bandwidth [39,40].

UEP coding is a very attractive technique in broadcast systems because it gradually reduces the transmission rate. The use of UEP codes for satellite broadcasting of digital TV signals has been considered in [41,42]. The coding scheme is designed in such a way that the information bits carrying the basic definition TV signal have a lower error rate than the high-definition information bits. Because of the nonlinear nature of the channel, the constellations are also chosen to have constant envelope. Reported results achieve graceful degradation, and no error propagation from the first level decoder to the second-level decoder.

5. UEP CODES FOR WIRELESS COMMUNICATIONS

Rapidly developing standards, as well as the technological advancement in wireless communications has enabled the wireless transmission of not only low-quality voice but also high-fidelity multimedia data.

One of the most important performance measures in wireless communication system design is the bit error rate. While this is a meaningful measure for computer applications such as data transfer, it doesn't necessarily reflect the perceptual quality of multimedia data. In particular, for highly compressed audio or video signals, a rare bit error can lead to a highly annoying or unacceptable perceptual distortion. Instead of demanding a very low bit error rate, which is costly to achieve, the use of UEP and joint design of source and channel coders is desirable. Because of this, employing UEP codes in wireless systems have been studied in detail [43,44].

Jung et al. [45], propose a technique for the H.263-compatible video datastream, based on the data partitioning technique. The proposed algorithm employs bit rearrangement, which provides the UEP against channel errors. The unequal error protection capabilities of convolutional codes belonging to the family of RCPC codes are presented in many studies [46–48]. RCPC codes can be decoded by the maximum likelihood Viterbi algorithm with full channel state information and soft decisions; hence they are very desirable for use in digital mobile radio channels. Moreover, they are adopted in International Telecommunication Union (ITU) standards H.223 and H.324.

Orthogonal frequency-division multiplexing (OFDM), a type of multicarrier modulation, is the standard modulation scheme for digital audiobroadcasting, and is also proposed for digital television and beyond third generation (3G) cellular communication systems. OFDM can be used naturally and effectively to provide UEP by properly allocating power and assigning a constellation to each individual subchannel. The principle and algorithm of using power allocation is studied in Ho's paper [49].

It has been discovered that the use of soft bits in source decoding of audio, image, and video is beneficial. "Soft" means that the channel or the channel decoder supplies not only binary decisions but also reliabilities to the source decoder. This requires the "soft-in/soft-out" decoders that became available in the late 1990s, as they are used in the powerful "Turbo" decoding schemes. Furthermore, Turbo-detection is a suitable method to improve the bit error rate of not only protected bits but also of bits transmitted uncoded in a system with UEP, such as the class 2 bits in the GSM speech channel [50,51].

Figure 4 shows the GSM speech channel at full rate. The RPE-LTP (regular pulse excited–long-term prediction) speech encoder generates a block of 260 bits per 20-ms speech. According to their function and importance for the quality of the speech, the bits of one block are divided into three classes. The most important bits are the class 1a bits (50 bits), which describe the filter coefficients, block amplitudes, and the LTP parameters. Next in importance are the class 1b bits (132 bits), which consist of RPE pointers, RPE pulses, and some other LTP parameters. Least important are the class 2 bits (78 bits), which contain the RPE pulse and filter parameters. First the class 1a bits are encoded by a weak error detecting (53,50,2) CRC (cyclic redundancy check) code. The class 1a and 1b bits are then protected by a convolutional code whereas class 2 bits are transmitted unprotected. The combined sequences of data classes are modulated with the same Euclidean distance properties. As an alternative to the multiplexing of two independent codes, one may employ explicit UEP codes. Then the task of providing unequal protection is divided between the channel encoder and a nonuniform signal set, which discriminates in

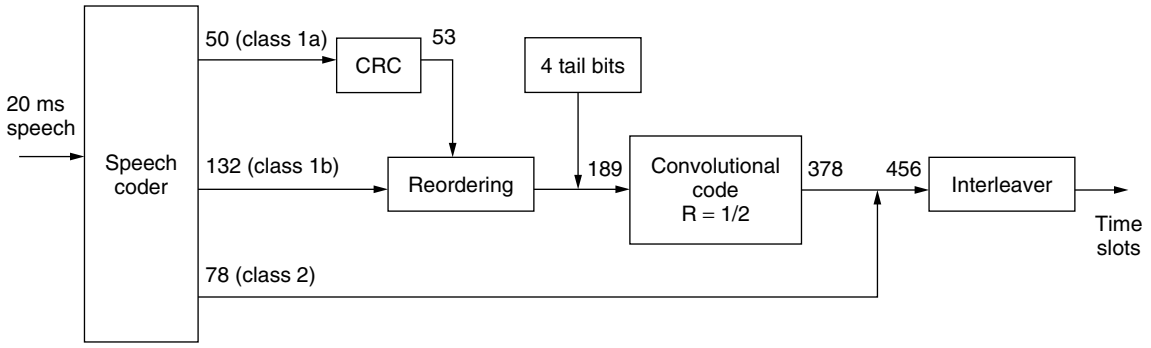


Figure 4. Encoding scheme for the GSM speech channel (full-rate TCH/FS).

favor of the more important bits. Sajadieh et al. [52] show that with the above modifications, better bit-error-rate performance for class 1a bits as well as reduced complexity for the overall decoder can be attained.

The wireless interfaces of asynchronous transfer mode (ATM) networks are gaining importance in order to provide mobility. In a wireless ATM, the ATM cells are transmitted via radioframes between a base station and a mobile station. The wireless interface of ATM networks is a band-limited channel with much higher error rate than the wired one. In multimedia wireless ATM interface, the header and various payloads to be transmitted have different importance or error protection needs. Therefore, it is desirable that channel coding provides UEP, matched to the different sensitivity of source encoded symbols. Several papers have been published for studying UEP block codes [53] and convolutional codes [54] in wireless ATMs.

6. MISCELLANEOUS USE OF UEP AND CONCLUDING REMARKS

Asymmetric digital subscriber lines (ADSLs), a transmission system capable of realizing very high bit-rate services over existing telephone lines, are well suited to multimedia applications, such as videotelephony and videoconferencing. Zheng and Liu have shown [55] that UEP codes can be used to provide extra protection for some subchannels, and proposed a method that achieves significant performance improvement compared to spectrally shaped channels commonly used in ADSL systems.

Another interesting study proposes using a coded modulation technique to increase the storage capacity of multilevel and analog memory cells by one extra bit per cell [56]. The technique is applied to improve readout bit -error probability and provide unequal error protection for the bits to be stored.

As can be seen from the previous examples, UEP codes have been applied to many aspects of modern communication systems when there is a natural need to protect some part of the information content more than the others. It is also not too difficult to see that many data sources fall into this category. Throughout the published literature on UEP codes, some commonalities can be observed: (1) if designed properly, UEP codes generally provide better performance than can equal error protection codes of the same category, and (2), UEP codes achieve gradual and smooth degradation of performance. And because of these properties, UEP codes are employed in various types of communication networks.

BIOGRAPHY

Arda Aksu received his B.Sc. degree from Middle East Technical University, Ankara, Turkey, in 1989, and his M.Sc. and Ph.D. degrees all in electrical engineering from Northeastern University, Boston, Massachusetts, in 1992 and 1995, respectively. From 1994 to 2000 he was with Verizon Technology Organization (formerly known as GTE Laboratories Inc.) in Waltham, Massachusetts. His responsibilities as a principal member of technical staff mainly included design, analysis, and testing of cellular and fixed-access wireless communications systems. Since 2001, he has been a senior technologist with the office of the CTO in Comverse working on the conceptualization and development of new and enhanced services for the next generation wireless networks.

Dr. Aksu's research interests are in the area of technology assessment and strategy development for the provision of wireless and personal communication services and products, as well as the analysis, design, and implementation of wireless networks. He is the author of over 25 scientific and technical publications as well as industry standards contributions in the above disciplines. He has served in the capacity of session chair, paper reviewer, and invited author in IEEE communities. He is the coinventor in five U.S. patent filings.

BIBLIOGRAPHY

1. B. Masnick and J. K. Wolf, On linear unequal error protection codes, *IEEE Trans. Inform. Theory* **IT-3**: 600–607 (Oct. 1967).
2. W. C. Gore and C. C. Kilgus, Cyclic codes with unequal error protection, *IEEE Trans. Inform. Theory* (March 1971).
3. I. M. Boyarinov and G. L. Katsman, Linear unequal error protection codes, *IEEE Trans. Inform. Theory* **IT-27**: 168–175 (March 1981).
4. W. J. van Gils, Two topics on linear unequal error protection codes: Bounds on their length and cyclic code classes, *IEEE Trans. Inform. Theory* **IT-29**: 866–876 (Nov. 1983).

5. L. A. Dunning and W. E. Robbins, Optimal encoding of linear block codes for unequal error protection, *Inform. Control* **37**: 150–177 (1978).
6. W. J. van Gils, Linear unequal error protection codes from shorter codes, *IEEE Trans. Inform. Theory* **IT-30**: 544–546 (May 1984).
7. M. C. Lin, C. C. Lin, and S. Lin, Computer search for binary cyclic UEP codes of odd length up to 65, *IEEE Trans. Inform. Theory* **36**: 924–935 (July 1990).
8. C. Zhi, F. Pingzhi and J. Fan, New results on self-orthogonal unequal error protection codes, *IEEE Trans. Inform. Theory* **36**: 1141–1144 (Sept. 1990).
9. A. R. Calderbank and N. Seshadri, Multilevel codes for unequal error protection, *IEEE Trans. Inform. Theory* **39**: 1234–1248 (July 1993).
10. L.-F. Wei, Coded modulation with unequal error protection, *IEEE Trans. Commun.* **41**: 1439–1449 (Oct. 1993).
11. A. Aksu and M. Salehi, Transmission of analog sources using unequal error protecting codes, *IEEE Trans. Commun.* **43**: 1225–1229 (Feb./March/April 1995).
12. N. Seshadri and C.-E. W. Sundberg, Multilevel block coded modulations for the Rayleigh fading channel, *Proc. IEEE Global Commun. Conf.*, Phoenix, AZ, 1991.
13. M. Isaka and H. Imai, Hierarchical coding based on adaptive multilevel bit-interleaved channels, *Proc. IEEE Vehic. Tech. Conf.*, 2000, pp. 2227–2231.
14. M. Isaka et al., Multilevel codes and multistage decoding for unequal error protection, *Proc. IEEE Int. Personal Wireless Commun. Conf.*, 1999, pp. 249–254.
15. P. P. Bergmans and T. M. Cover, Cooperative broadcasting, *IEEE Trans. Inform. Theory* **20**: 317–324 (May 1974).
16. S. Gadkari and K. Rose, Time-division versus superposition coded modulation schemes for unequal error protection, *IEEE Trans. Commun.* **47**: 370–379 (March 1999).
17. U. Dettmar, G. Yan and U. K. Sorger, Modified generalized concatenated codes and their application to the construction and decoding of LUEP codes, *IEEE Trans. Inform. Theory* **41**: 1499–1503 (Sept. 1995).
18. M. Matsunaga, D. K. Asano, and R. Kohno, Unequal error protection scheme using several convolutional codes, *Proc. IEEE Int. Symp. Inform. Theory*, 1997, p. 101.
19. M.-C. Chiu, C.-C. Chao, and C.-H. Wang, Convolutional codes for unequal error protection, *Proc. IEEE Int. Symp. Inform. Theory*, 1997, p. 290.
20. C.-H. Wang and C.-C. Chao, Further results on unequal error protection of convolutional codes, *Proc. IEEE Int. Symp. Inform. Theory*, 2000, p. 35.
21. T. Danon and S. I. Bross, Free distance lower bounds for unequal error-protection convolutional codes, *Proc. IEEE Int. Symp. Inform. Theory*, 2000, pp. 261.
22. E. K. Englund, Nonlinear unequal error-protection codes are sometimes better than linear ones, *IEEE Trans. Inform. Theory* **IT-37**(5): 1418–1420 (Sept. 1991).
23. R. V. Cox, J. Hagenauer, N. Seshadri, and C.-E. W. Sundberg, Subband speech coding and matched convolutional channel coding for mobile radio channels, *IEEE Trans. Signal Process.* **39**: 1717–1731 (Aug. 1991).
24. H. Shi, P. Ho, and V. Cuperman, Combined speech and channel coding for mobile radio applications, *Proc. IEEE Int. Conf. Select Topics Wireless Commun.*, 1992, pp. 180–183.
25. D. Sinha and C.-E. W. Sundberg, Unequal error protection methods for perceptual audio coders, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 1999, pp. 2423–2426.
26. S. Kim, B. Kang, and K. Oh, An improved scheme in CDMA forward link using unequal error protection, *Proc. IEEE Int. Conf. Universal Personal Commun.*, 1998, pp. 1093–1096.
27. A. Bernard, X. Liu, R. Wesel, and A. Alwan, Embedded joint source-channel coding of speech using symbol puncturing of trellis codes, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.* 1999, pp. 2427–2430.
28. F. I. Alajaji, S. Al-Semari, and P. Burlina, Visual communication via trellis coding and transmission energy allocation, *IEEE Trans. Commun.* **47**: 1722–1728 (Nov. 1999).
29. X. Fei and T. Ko, Turbo-codes used for compressed image transmission over Rayleigh fading channel, *Proc. IEEE Int. Conf. Universal Personal Commun.*, 1997, pp. 505–509.
30. T.-C. Yang and C.-C. J. Kuo, Error correction for wireless image communication with a rate-distortion model, *Proc. Signals, Systems, Computers Conf.* (*Asilomar*), 1998, pp. 963–967.
31. I. Kozintsev and K. Ramchandran, Hybrid compressed-uncompressed framework for wireless image transmission, *Proc. Int. Conf. Image Processing*, 1997, pp. 77–80.
32. P. G. Sherwood and K. Zeger, Error protection for progressive image transmission over memoryless and fading channels, *IEEE Trans. Commun.* **46**: 1555–1559 (Dec. 1998).
33. M. J. Ruf, A high performance fixed rate compression scheme for still image transmission, *Proc. Data Comput. Conf.* (*DCC*), 1994, pp. 294–303.
34. S. Manji and G. Djuknic, Bandwidth efficient and error resilient image coding for Rayleigh fading channels, *Proc. IEEE Vehic. Technol. Conf.* (*VTC*), 1999, pp. 1485–1489.
35. J. Vass and X. Zhuang, Joint source-channel coding for highly efficient error resilient image transmission, *Proc. IEEE Int. Symp. Circuits and Systems* (*ISCAS*), 2000, pp. 311–314.
36. W. R. Heinzelman, M. Budagavi, and R. Talluri, Unequal error protection of MPEG-4 compressed video, *Proc. Int. Conf. Image Processing* (*ICIP*), 1999, pp. 530–534.
37. T. Kawahara and S. Adachi, Video transmission technology with effective error protection and tough synchronization for wireless channels, *Proc. Int. Conf. Image Processing* (*ICIP*), 1996, pp. 101–104.
38. P. Burlina and F. Alajaji, An error resilient scheme for image transmission over noisy channel with memory, *IEEE Trans. Image Process.* 593–600 (April 1998).
39. A. Andreadis, G. Benelli, A. Garzelli, and S. Susini, FEC coding for H.263 compatible video transmission, *Proc. IEEE Int. Cong. Image Processing*, 1997, pp. 579–581.
40. C. W. Yap and K. N. Ngan, Unequal error protection of images over Rayleigh fading channels, *Proc. Int. Symp. Signal Proc. Applications* (*ISSPA*), 1999, 19–22.
41. R. H. Morelos-Zaragoza, O. Y. Takeshita, and H. Imai, Coded modulation for satellite broadcasting, *Proc. IEEE Global Telecommun. Conf.* (*GLOBECOMM'96*), 1996, 31–35.
42. G. Taricco and E. Biglieri, Some simple coded-modulation schemes for unequal error protection in satellite communications, *Proc. IEEE Int. Symp. Spread Spectrum Techniques, Applications*, 1996, 1288–1294.

43. H. Gharavi and S. M. Alamouti, Multi-priority video transmission for third generation wireless communication systems, *Proc. IEEE* **87**: 1751–1763 (Oct. 1999).

44. F. Babich, G. Lombardi, and F. Vatta, Performance evaluation of source-matched channel coding for mobile communications, *Proc. IEEE Vehic. Technol. Conf.*, 1998, 2517–2521.

45. H.-S. Jung, R.-C. Kim, and S.-U. Lee, On the robust transmission technique for H.263 video data stream over wireless networks, *Proc. IEEE Int. Image Proc. Conf.*, 1998, 463–466.

46. J. Hagenauer, N. Seshadri, and C.-E. Sundberg, The performance of rate-compatible punctured convolutional codes for digital mobile radio, *IEEE Trans. Commun.* **38**: 966–980 (July 1990).

47. S. Kang and K.-Y. Yoo, Region and time based unequal error protection for video transmission over mobile links, *Proc. IEEE Int. Symp. Circuits and Systems*, 1999, 511–514.

48. J. Hagenauer and T. Stockhammer, Channel coding and transmission aspects for wireless multimedia, *Proc. IEEE* **87**: 1764–1777 (Oct. 1999).

49. K.-P. Ho, Unequal error protection based on OFDM and its application in digital audio transmission, *Proc. IEEE Global Telecommun. Conf.* (*GLOBECOM'98*), 1998, 1320–1325.

50. G. Bauch and V. Franz, Iterative equalization and decoding for the GSM system, *Proc. IEEE Vehic. Technol. Conf.*, 1998, 2262–2266.

51. F. Burkert et al., Turbo decoding with unequal error protection applied to GSM speech coding, *Proc. IEEE Global Telecommun. Conf.* (*GLOBECOM*), 1996, 2044–2048.

52. M. Sajadieh, F. R. Kschischang, and A. Leon-Garcia, Modulation-assisted unequal error protection over the fading channel, *IEEE Trans. Vehic. Technol.* **47**: 900–908 (Aug. 1998).

53. S. Aikawa, Y. Motoyama, and M. Umehira, Forward error correction schemes for wireless ATM, *Proc. IEEE Int. Commun. Conf.*, June 1996, pp. 454–458.

54. Z. Sun, S. Kimura and Y. Ebihara, Adaptive two-level unequal error protection convolutional code scheme for wireless ATM networks, *Proc. IEEE INFOCOM*, 2000, pp. 1693–1697.

55. H. Zheng and K. J. Ray Liu, Robust image and video transmission over spectrally shaped channels using multicarrier modulation, *IEEE Trans. Multimedia* 88–103 (March 1999).

56. H.-L. Lou and C.-E. Sundberg, Coded modulation to increase storage capacity of multilevel memories, *Proc. IEEE Global Telecommun. Conf.,* (*GLOBECOM'98*), 1998, 3379–3384.

V

V.90 MODEM

BRENT TOWNSHEND
Townshend Computer Tools
Menlo Park, California

1. INTRODUCTION

Since the era of the earliest time-shared computers, data communications methods have followed a path parallel to computer development. Each new generation of faster CPU and denser memory has been accompanied by improvements in communication capabilities. One particularly important mode of communication has been transmission of data over regular telephone lines. However, since the telephone system was not designed with data transmission in mind, creative manipulations of the signal were required to achieve this. One of the earliest examples of this was the teletype machine, which sent data at a speed of 110 bits per second (bps). This was achieved by coding the bits into one of two different tone signals and sending one such symbol 110 times each second. The tones were chosen to be in the audio range that the telephone system was designed to pass, and adequately separated to allow the receiving end to distinguish which tone was present at each time interval. This pattern of modulating a signal, such as alternating between the two tones above, and the demodulation at the receiver to reconstruct the original data signal was the basic model and namesake for all of the subsequent modems (*mod*ulator-*dem*odulators) that followed.

Clearly, the key requirement of a modem was to accurately transmit the data accurately and to do so as fast as possible. Researchers found ways to continually increase the speed by applying more and more complex techniques to the modulation so that more data could be squeezed into the telephone channel designed for voice. The result was a rapid succession of improvements, typically with the transmission rate doubling every 3 years. It would be misleading, however, to simply imagine that these developments were successive improvements of one idea. Each new generation of modem pioneered a qualitatively different technique of modulation or demodulation to achieve the next step in the progression. The 300- and 1200-baud[1] modems based on frequency-shift keying gave way to 2400-bps devices, by exploiting phase shift keying and coherent demodulation. With the introduction of adjustable and then adaptive equalizers, more complex signaling schemes were introduced that could carry more per second in the same bandwidth. Amplitude modulation was combined with phase shift keying to give the quadrature amplitude modulation allowing rates of 9600, 14,400, and 19,200 bps by the late 1980s.

Viterbi decoding and increasingly complex modulation patterns allowed further optimization to bring the data rates to 33,600 bps by 1994. But at this point it was thought that the maximum had been obtained, and there was little optimism that any further improvement could be made regardless of the ingenuity of the modulation designers.

2. SHANNON LIMIT

Throughout and even before the evolution described above, there was a firm theoretical foundation for the envelope of possibilities that had been developed by Claude Shannon several decades earlier. In his seminal paper of 1948 [1], he provided a simple equation to predict that maximum data rate that can be obtained on any channel given the width of the frequency band that is passed by the channel, W, the signal power, S, and the noise power N:

$$C = W \log_2 \frac{S+N}{N}$$

This agrees with intuitive ideas of how data rate should vary. If the bandwidth is doubled (i.e., equivalent to having two identical channels, each with the original bandwidth), then the data rate is doubled. As the background noise level is lowered or the signal power is increased, each symbol is less "fuzzy" and more symbols can be packed into the available space, again resulting in higher data rates.

Applying the Shannon limit to the telephone channel is straightforward. The telephone channel passes audio frequencies in the range of 300–3400 Hertz, giving a bandwidth of 3100 Hz. AT&T and its predecessors [2] have repeatedly surveyed the noise level of the U.S. telephone channels, finding a consensus signal-to-noise ratio of approximately 35 dB. Using these values in the above equation yields a result of 36,000 bps. Thus, it seemed clear at one time that the 33,600-bps rate attained above leaves little further room for improvement.

3. BREAKING THE LIMIT

As we now know, 56-kbps modems are possible over the same channel analyzed above. Does this mean that the Shannon limit is in error or somehow inapplicable? No, the theory is correct, and even the calculation for the telephone channel is accurate. However, what the analysis presented above overlooks is that the definition of the inputs to the equation depends on your point of view. Specifically, the signal-to-noise ratio depends on your definition of what constitutes noise. For example, imagine tuning a radio between two stations such that

[1] The baud rate is the number of symbols that are transmitted per second — if the symbol conveys one bit, then this is equal to the bit rate.

both are heard. If you are interested in listening to one, say, a jazz station, then you may think of the other, a rock station, as a noise source. Conversely, a different listener may swap the classification of which is the signal and which is the noise.

In the case of the telephone channel, the definition of noise is equally important. In the original telephone system design, signals were carried in their analog form from one endpoint to the other through a series of switches and relays. This system picked up electrical noise from a variety of sources including cross-talk from other calls and inductive pickup from motors and other electrical sources. Eventually, the network was replaced by a digital system where the signals are carried between central offices in a digital form that is immune to electrical noise pickup. In the design of the digital system, a choice had to be made as to the sampling rate and precision of these data. To minimize the data requirements and maximize the system capacity in terms of the number of concurrent voice connections, these choices were made to approximate the quality of the older analog system. Thus, during that evolution of the long-distance network from analog to hybrid to all digital, the electrical noise, which gave a signal noise ratio of around 35 dB, was replaced by quantization noise of the same magnitude. As far as telecom designers were concerned, the system did not look very different — the bandwidth and signal-to-noise ratios were similar and could generally be treated identically. But this rough equivalence ignores a critical difference. Specifically, if you are in control of the digital words used by the telephone network, then the quantization is no longer noise — it is a deterministic phenomenon [3,4].

The insight that the quantization is not noise radically changes the parameters of the problem. The pertinent noise in the Shannon calculation is then the noise due to electrical pickup, crosstalk, and other uncontrollable sources. Surveys of the local loops [2] estimate these noise levels at lower than -79 dBm[2] for at least 90% of the cases. Since telephone companies typically limit transmit levels to -12 dBm, the Shannon limit gives a maximum data capacity of

$$C = 3100 \log_2 10^{(79-12)/10} = 69 \text{ kbps}$$

This capacity for the local loop is further demonstrated by the operation of products such as ISDN and DSL, which use the same local loop to achieve data rates of $\geq$1.5 Mbps, although they use a wider bandwidth.

4. OVERVIEW OF V.90

A V.90 communication system is shown in Fig. 1. The digital modem endpoint usually resides at a commercial server such as an Internet service provider (ISP) and may be part of a remote access server that contains many such devices. The digital modem is connected to the telephone network via ISDN, T1, or E1 digital subscriber lines. These types of connections maintain digital signaling from the modem to the digital-to-analog converter at the digital central office (CO) where the end user's line is terminated.

Digital telephone networks carry both voice and data signals as digital symbols or codewords. Each symbol consists of 8 bits and is sent at a rate of 8000 symbols per second, giving a data rate of 64 kbps. When the symbols represent a voice signal, they are interpreted according to either a μ-law or A-law encoding rule[3] [5]. This nonlinear encoding allows the dynamic range and noise characteristics of a typical telephone channel to be preserved when the signal is carried digitally. As can be seen in Fig. 2, these encoding rules allow small signals to be encoded with high precision while still permitting large signals to be transmitted.

Unlike previous telephone modems such as V.34, the V.90 modem uses a baseband signaling scheme without modulation. The data to be transmitted are mapped onto a sequence of telephone codewords and then presented to the telephone network as if they represented an analog signal. At the central office the stream of symbols received from the network must be converted to an analog signal to be sent to the consumer's telephone. The central office does not distinguish between these signals and normal voice calls, so it applies the digital data to a codec, which converts the symbols to an analog signal using the μ-law or A-law rules. The codec also filters the output of the digital-to-analog conversion to smooth the signal and reduce high-frequency components before sending the signal along the two-wire local loop to the consumer.

At the customer premises, an analog modem is installed that is connected to a telephone jack. It receives the audio signal that was sent by the codec after additional filtering by the transmission lines and pickup of noise and

[2] The unit dBm is a measure of absolute power in decibels such that 0 dBm is one milliwatt.

[3] The μ-law nonlinear mapping is used primarily in North America, while A law is used in other parts of the world.

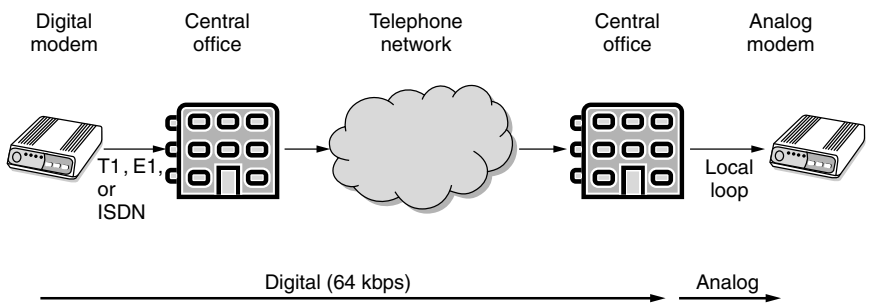


Figure 1. V.90 communications path.

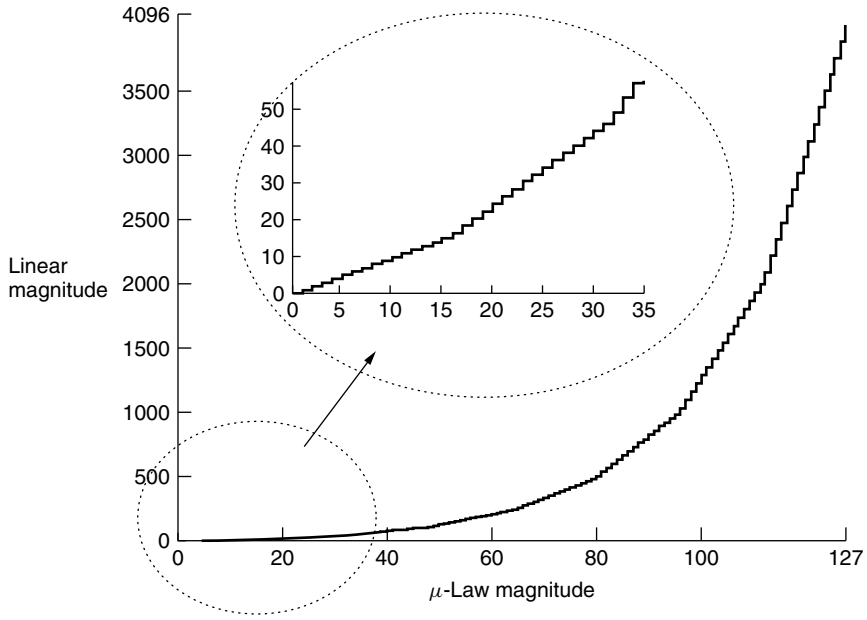


Figure 2. Nonlinear mapping of μ-law codewords.

crosstalk. The analog modem must undo those effects as well as the effects of the codec's smoothing filter. It must then reconstruct the sample timing of the codec and infer the sequence of codewords that was transmitted over the network. Once the symbol stream has been reconstructed, the mapping used by the digital modem can be reversed to recover the original data.

In addition, data must be sent back from the analog modem to the digital end. This upstream transmission is usually done using a lower bandwidth modulation technique such as V.34 at 28.8 or 33.6 kbps. This can be done without significantly impacting the downstream data through use of echo cancelers at both ends, which separate the upstream and downstream signals even though both are carried on the same 2-wire local loop. The resulting asymmetric data rates, 56 kbps downstream and 33.6 kbps upstream, are well suited to most traffic patterns. The typical use of Internet access requires significantly more downstream data such as HTML, images, and video, but the upstream data more often consists of short requests, acknowledgments, and so on.

5. DIGITAL MODEM OPERATION

The internal structure of the digital modem is shown in Fig. 3. It starts with an interface to the host computer or source of data. Since the digital modem usually resides at a central server, such as an Internet service provider (ISP), several, if not hundreds, of digital modems may be implemented by a single hardware device. In any case, the data from the host computer eventually reach the PCM encoder, where the data are is converted into a sequence of 8-bit symbols for presentation to the telephone network.

On the surface, the PCM encoder (Fig. 4) appears to have a simple function; to pack the incoming data onto telephone network codewords. Once encoded, the codewords would be transmitted at a rate of 8000 per second, providing a 64-kbps data path. However, several factors make the operation more complex, including

Codec Filtering. At the receiving central office, the smoothing filter in place for all telephone calls filters low-frequency components out of the reconstructed analog signal. A typical codec (compression/decompression) filter is shown in Fig. 5. As can be seen from the figure, signals below 50 Hz are removed by this filter. Since this filter cannot be bypassed without making changes at the central office, data sequences that result in low-frequency components would not be uniquely identifiable by the analog modem. To avoid this, the digital

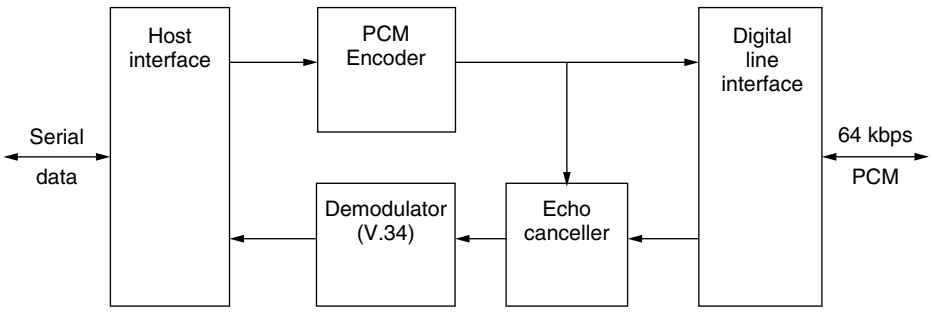


Figure 3. Digital modem block diagram.

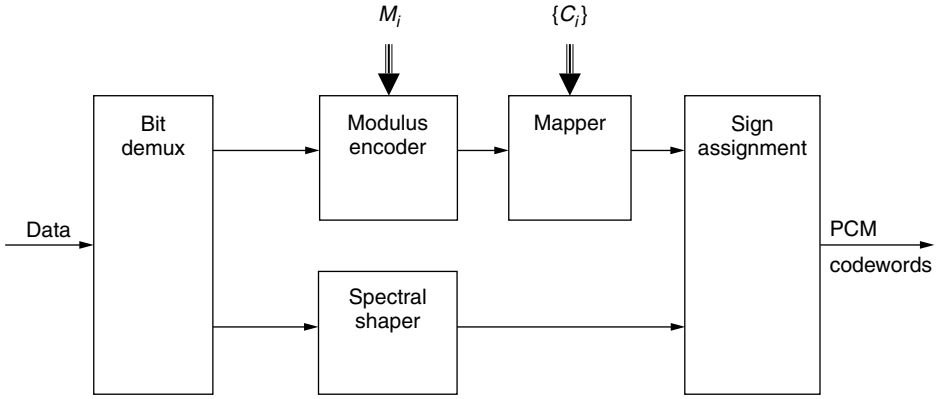


Figure 4. PCM encoder.

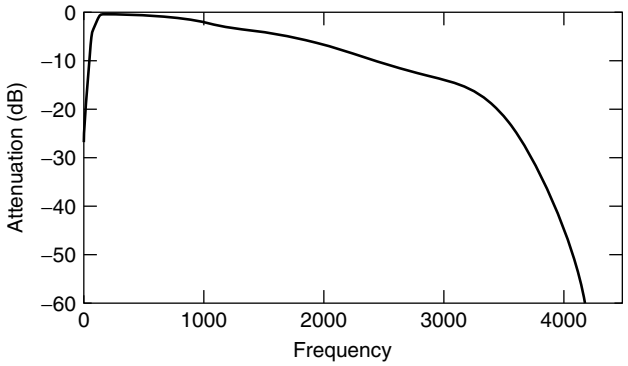


Figure 5. Codec digital-to-analog converter smoothing filter transfer function.

modem sends only a subset of all possible codeword sequences such that the resulting signal has a spectral shape with little energy outside the frequencies passed by the codec.

Codec Nonlinearities. The digital-to-analog converters used in codecs are far from ideal. Although the analog levels corresponding to each codeword value are specified in the G.711 standard [5], actual implementations may introduce DC biases, nonlinearities, or asymmetries in the mapping. These nonlinearities need to be learned by V.90 modems for each call and appropriate compensation must be applied.

Local Loop Transmission. The physical pair of wires between a central office and a consumer's home or business is far from an ideal medium. Wiring anomalies, such as bridge taps, create topologies with unterminated branches that cause signal reflections. Loading coils, which are used to improve the frequency response for voice calls, can wreak havoc for data signals. Modems must identify the existence of these and choose appropriate compensation or fallback to lower speeds.

Power Constraints. Arbitrary sequences of codewords may result in signal levels on the local loop that are higher in power than typical telephone signals. This could give rise to crosstalk to adjacent telephone lines causing noise on other calls. To prevent this, the Federal Communications Commission (FCC) and other regulatory organizations have specified power limits making certain codeword sequences impermissible.

Digital Impairments. Even though telephone codewords are carried digitally from one central office to another, the network does, at times, modify that datastream. Digital or even analog attenuation is sometimes inserted to control echo levels. This may be done using a remapping of codewords or with a tandem of digital-to-analog and analog-to-digital converters. When central offices need to communicate signaling information, one method commonly used is "robbed bit" signaling [12]. The network will use the least-significant bit of one out of every 6 (or sometimes 12) codewords for its own signaling. During a single call, multiple hops may entail use of multiple codewords for signaling within each 6-codeword frame. Since the original value of these bits is lost, fewer bits of each codeword are usable for data transmission. For international calls a remapping between the two G.711 codeword sets (μ-law and A-law) may occur.

Speech Coding. Some transmissions undergo compression, such as ADPCM, designed to pass speech signals at lower bit rates. This allows more circuits to be carried on the same digital trunks. However, such lossy coding techniques limit the data rates and, in many cases, prevent the use of high speed techniques such as V.90.

For these reasons, the encoder must detect these problems and, if possible, provide some preprocessing of the data stream to avoid them and ensure that the data is recoverable by a decoder.

Figure 4, a block diagram of the PCM Encoder, shows the sequence of these transformations given in the ITU V.90 Recommendation [6]. These have been standardized to guarantee inter-operability between different implementations of the modems.

Generation of codewords is done in frames of 6 codewords to allow repeatable handling of digital impairments such as robbed-bit signaling, which usually occurs in fixed positions within each 6-codeword frame. By creating an encoder that handles each of these 6 symbol intervals separately, it is possible to make use of the least significant

bit in intervals where robbed-bit signaling is not occurring and avoid intervals where it does occur.

The first step in packing the incoming data into these 6 codewords is bit selection. As will be discussed below, the sign bit of each symbol is treated specially, so at this point blocks of data bits are separated into S bits that will be encoded into the codewords' sign bits and K bits that will determine the magnitude part of each codeword. For example, at the maximum rate allowed by the standard, groups of 42 bits are read from the interface and separated so that 3 are used for sign bit encoding and the remaining 39 are passed on to the modulus encoder. The data rate using this breakdown is

$$\begin{aligned}\text{Rate} &= \frac{8000}{6} * (K + S) \\ &= \frac{8000}{6} * (39 + 3) \\ &= 56{,}000 \text{ bps}\end{aligned}$$

The standard also allows for other divisions where between 3 and 6 bits are used for sign bit encoding and 15–39 bits are used by the modulus encoder, providing data rates ranging from 28 to 56 kbps.

The next step in the encoding process is using the K bits to choose the magnitude portion of the 6 codewords that will be sent to the telephone network. Each magnitude has 7 bits of precision, which would allow up to 42 bits of encoding. However, only a subset of all possible codewords is used in each symbol interval. These codesets, $\{C_0\}$ to $\{C_5\}$, are determined in the training phase during the initiation of a call (see Section 9 below). Each codeset has M_i elements where $M_i <= 128$ and the modulus encoding step consists of choosing six integers, K_0 to K_5, from the six codesets $\{C_0\}$ to $\{C_5\}$. The algorithm for choosing the K_i has the following steps:

1. Use the K incoming bits to represent an integer, R_0:

$$R_0 = \sum_{j=0}^{k-1} b_j 2^j$$

2. Perform the following recursion to obtain the K_i:

$$\begin{aligned}K_i &= R_i \text{ modulo } M_i \\ R_{i+1} &= R_i - K_i M_i\end{aligned}$$

Assuming that the product of the codeset sizes is greater than 2^K, these steps result in a set of K_i that can be used to reconstruct the R_0 and the original K data bits:

$$R_0 = \prod_{i=0}^{6} k_i$$

For example, if $K = 15$, $S = 4$, and, during training, the M_i were chosen as $\{6, 4, 6, 6, 7, 6\}$, then to encode the data bits {1101010100010010011}, we would use the first 4 bits {1101} for sign encoding. The remaining 15 bits would then be applied to the modulus encoder:

$$\begin{aligned}R_0 &= \sum_{j=0}^{14} b_j 2^j \\ &= 0.2^0 + 1.2^1 + 0.2^2 + 1.2^3 + 0.2^4 + 0.2^5 \\ &\quad + 0.2^6 + 1.2^7 + 0.2^8 + 0.2^9 + 1.2^{10} + 0.2^{11} \\ &\quad + 0.2^{12} + 1.2^{13} + 1.2^{14} = 10{,}387\end{aligned}$$

$$\begin{aligned}K_0 &= (R_0 \text{ modulo } M_0) \\ &= 10{,}387 \text{ modulo } 6 = 1 \\ K_1 &= 1731 \text{ modulo } 4 = 3 \\ K_2 &= 432 \text{ modulo } 6 = 0 \\ K_3 &= 72 \text{ modulo } 6 = 0 \\ K_4 &= 12 \text{ modulo } 7 = 5 \\ K_5 &= 1 \text{ modulo } 6 = 1\end{aligned}$$

$$\begin{aligned}R_1 &= \frac{R_i - K_i}{M_i} = 1731 \\ R_2 &= \frac{1731 - 3}{4} = 432 \\ R_3 &= \frac{432 - 0}{6} = 72 \\ R_4 &= \frac{72 - 0}{6} = 12 \\ R_5 &= \frac{12 - 5}{7} = 1\end{aligned}$$

The resulting output from the modulus encoder would then be {1, 3, 0, 0, 5, 1}. These values are then used to index into the codesets, $\{C_0\}$ to $\{C_5\}$ to choose the magnitude portion of each of the six codewords to be transmitted next to the telephone network.

The S sign bits taken from the data are combined with $S_r = 6 - S$ additional bits that are chosen by a spectral shaper. The function of the spectral shaper is to choose the S_r additional bits to control the spectrum of the analog signal that will be constructed by the codec at the consumer's CO. As discussed above, the codec applies its own filtering to the signal. Performance can be improved by avoiding symbol sequences that have significant energy in parts of the spectrum where the codec attenuates the signal.

During training, the characteristics of the codec filter and transmission lines are probed and a spectral shaping filter is chosen by the analog modem of the form

$$F(z) = \frac{(1 - b_1 z^{-1})(1 - b_2 z^{-1})}{(1 - a_1 z^{-1})(1 - a_2 z^{-1})}$$

The spectral shaper applies this filter to the projected output over the next several frames for each possible choice of the S_r additional sign bits and chooses the values that minimize the energy output from the filter.

In the example above, $S = 4$ and $S_r = 2$, so the energy output from the spectral filter would be measured using the four possible settings of the two additional sign bits combined with the magnitude bits output from the modulus encoder. The chosen bits are the ones that result in the lowest energy over the current and subsequent frames (the standard allows this window to range from 1 to 4 frames). Note that the V.90 standard includes

additional manipulations of the sign bits, but the effect is similar.

6. ANALOG MODEM OPERATION

The analog modem end of a V.90 connection must reverse the effects of the codec at the central office and the distortions caused by the local loop. It must also synchronize itself to the clock used by the telephone company to infer the digital codewords that were sent on the digital telephone network and undo the digital modem transformations to recover the original data signal.

The V.90 standard does not specify how the decoder should operate since its implementation does not impact the interoperability aspects. Given the function of the encoder, any decoder that extracts the original bits is compliant. However, a decoder implementation will typically consist of an equalizer, a clock recovery circuit, decision logic, and bit demapping to reverse the effects of the modulus encoding and spectral shaping. A block diagram of a possible implementation is shown in Fig. 6.

The main elements of the decoder are an equalizer that provides spectral compensation to undo the effects of the codec and local loop filtering; a clock recovery circuit to synchronize the modem with the central office's 8-kHz sampling rate, and decision logic that makes a decision as to which bits were transmitted by the digital modem. In addition, the analog modem will include logic for initialization, training, retraining, compression, error correction, and upstream data transmission as described below.

6.1. Equalizer

The equalizer can be implemented using an adaptive equalizer such as those described by Gitlin et al. [9]. The combination of the codec filter and the local loop has a spectral characteristic, shown in Fig. 5, that attenuates signals below 400 Hz and those above 3400 Hz. It also creates a slight spectral slope and adds some phase distortion to the signal. The equalizer must amplify the low and high frequencies, compensate for the spectral slope, and remove the phase distortion. One possible implementation of an equalizer that achieves these goals is the combination of a fractionally spaced feedforward equalizer with a decision feedback equalizer as shown in Fig. 7. The fractional spacing allows recovery of the high-frequency signals that are near the Nyquist rate of 4 kHz with better performance and lower complexity than sampling at 8 kHz would allow. The decision feedback equalizer permits reconstruction of the low-frequency components where inversion of the nulls of the codec filter using a conventional filter would result in an unstable system. By including the decision logic in the feedback loop, noise is not amplified by the filter as long as the decisions are correct. By choosing the codeword set appropriately, the error rate for the decisions can be made arbitrarily small.

The equalizer will normally be trained and its parameters established during the call initiation. In addition, to track changing characteristics of the channel, adaptation can occur continuously using an internally generated error signal. The tap weights of the equalizer are adjusted to minimize the error estimate using any of a host of update algorithms, such as LMS.

6.2. Clock Recovery

As well as equalization, the analog modem must lock to the codec's 8000-Hz sample clock used by the telephone network without having any direct connection to that clock. During call initiation a known pattern of codewords is transmitted and the analog modem deduces the frequency and phase of the sampling clock from these. In much the same way that the equalizer is adapted to changing channel conditions, the phase-locked clock can also track the codec clock using an error signal.

6.3. Decision Logic

With the equalized signal and a recovered sampling clock, the analog modem can then estimate the analog signal

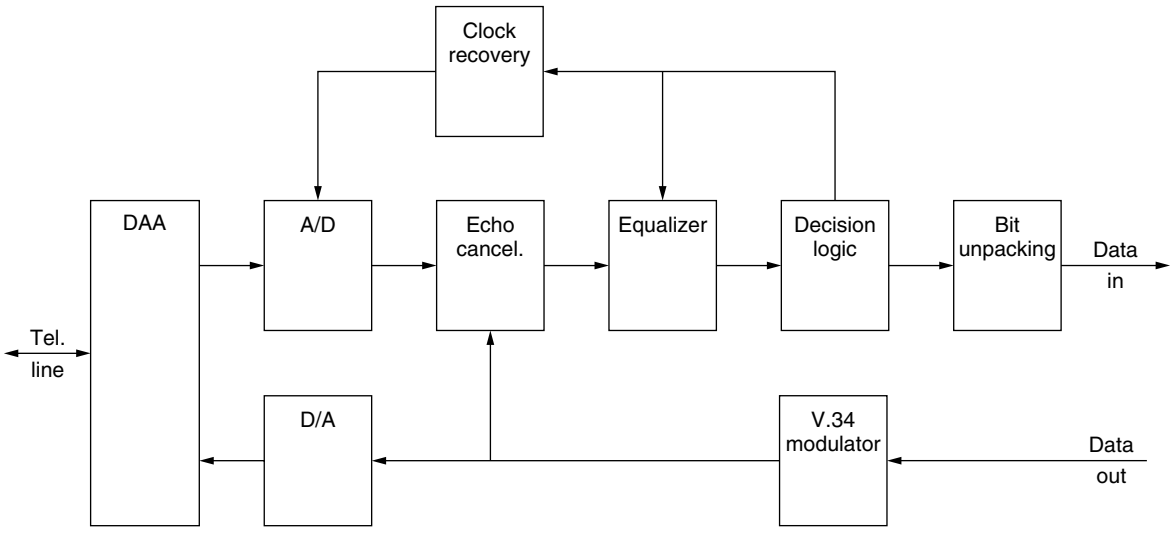


Figure 6. Analog modem block diagram.

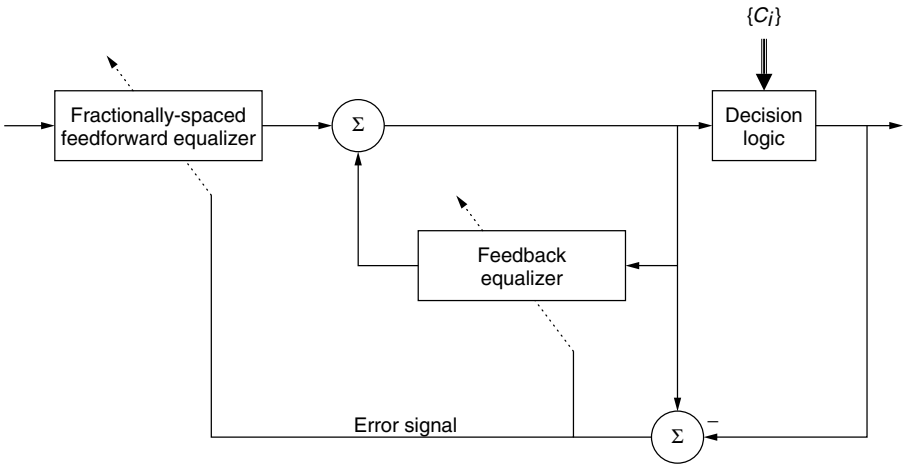


Figure 7. Equalizer structure.

output by the codec at each sampling instant. With the knowledge of the output levels for each of the M_i possible input codewords, the codeword that would give the closed analog signal is chosen to reconstruct K_i. In this way, the codewords that arrived at the codec at each frame are recovered. The operations performed by the digital modem's PCM encoder can then be reversed to reconstruct the $K + S$ data bits for each frame. Assuming the decision of the K_i was correct, this also gives the analog modem information as to the degree of error in the equalizer outputs. By subtracting the theoretical analog value for the received codeword from the equalizer output, an error signal is formed. This signal can be used to control adaptation of the equalizer and clock synchronizer.

7. REVERSE CHANNEL AND ECHO CANCELLATION

The preceding descriptions of the digital and analog modem focus only on the downstream channel from the digital modem to the analog modem. In V.90, the upstream channel is implemented using V.34 modulation techniques, giving a maximum upstream data rate of 33.6 kbps. This asymmetric channel is well suited to most applications where significantly more data flow downstream to the client than upstream.

Introduction of the reverse channel does add significant complexity to the system in that echo cancellation has to be added so that the modems do not receive delayed versions of their own outgoing signals at levels that would corrupt the incoming data.

8. COMPRESSION AND ERROR RECOVERY

Data sent over a modem connection often have some structure that can be exploited by a lossless compressor to improve throughput. HTML, for example, can typically be compressed by a factor of 4 or more. V.90 and prior modem standards often employ a form of Lempel–Ziv–Welch [7] compression to achieve this. Data that are not compressible (such as data that have already been compressed) can be detected using a continuous "compressibility" monitor. The compressor would then be switched off and the data passed verbatim. The details of these compression algorithms can be found in ITU Recommendation V.42bis [8].

Error control is usually implemented using the *Link-access procedure for modems* (LAPM), which has been standardized as V.42 [14]. This procedure blocks data together in packets and adds a checksum to each packet. When packets with correct checksums are received, an acknowledgement is returned to the transmitter. Multiple packets may be outstanding using a sliding-window protocol. If a packet is not acknowledged within a specific time, the transmitter backs up to the last successfully received packet and retransmits the following packets.

9. CALL INITIATION AND TRAINING

After call initiation, the pair of modems must exchange information and analyze line conditions to choose the optimum modulation methods, transmission speeds, codeword sets, spectral shaping parameters, and other features. All of this is done during a startup procedure involving several phases over a period of 10–30 s. Any modem user that has listened in on a data call will be familiar with the pings and chirps exchanged between the modems during these phases [13].

Phase	Description	Operations
1	V.8, V.8bis setup	Identify V.90 support, type of connection (digital or analog)
2	Probing	Exchange of modem capabilities; line probing; ranging
3	Half-duplex training	Equalizer and echo canceler training; digital impairment learning
4	Full-duplex training	Final training and fine-tuning.

The first phase allows the two modems to exchange information about their capabilities. It is during this phase that the modems will identify themselves as V.90-capable modems and whether they are in the role of the digital

or analog modem. The protocols used in this phase are specified in ITU Recommendation V.8, which is also used by most other modem types, such as V.34. This allows full backward and forward compatibility between any combination of modem types.

Phase 2 is the probing phase and is identical to that used in V.34. During this phase signals are exchanged that allow the modems to provide details about their capabilities and to identify some characteristics of the transmission path, such as the round-trip delay, available bandwidth, and signal level. Also some of the parameters to be used in the subsequent training phases are chosen here.

The third phase is where most of the training of the equalizer and echo cancellers occurs and where the codeword set is chosen through a procedure known as *digital impairment learning*. By exchanging test signals the analog modem can determine the characteristics of the analog channel from the codec to the modem, lock its clock to the central office clock, and choose a subset, $\{C_i\}$, of the 256 possible codewords in each timeframe that make error-free decoding possible.

Phase 4, the final phase, allows exchange of the parameters chosen during prior phases such as the spectral shaping filter parameters, the elements of the codeword sets, and the overall transmission rates to be used. The parameters for the V.34 upstream transmissions are also provided during this phase.

In addition to the detailed interactions during each of these phases, the V.90 standard provides procedures for error recovery, retraining of the modems, and cleardown.

10. STANDARDIZATION

Standardization is an essential element of communications protocols. With adoptation of a standard, devices from multiple manufacturers can interoperate and achieve high levels of compatibility. For modems, the International Telecommunications Union (*www.itu.int*) in Geneva, Switzerland takes the lead role in overseeing this process. The name "V.90" refers to a Recommendation (the ITU uses this term rather than "Standard") from the ITU Telecommunication Standardization Sector (ITU-T). The recommendations in the "V" series all relate to data communication over the telephone network. Previous modem recommendations include the following:

- V.21 A 300-bps duplex modem standardized for use in the general switched telephone network
- V.22 A 1200-bps duplex modem standardized for use in the general switched telephone network and on point-to-point 2-wire leased telephone-type circuits
- V.32 A family of 2-wire, duplex modems operating at data signaling rates of up to 9600 bps for use on the general switched telephone network and on leased telephone-type circuits
- V.34 A modem operating at data signaling rates of up to 33,600 bps for use on the general switched telephone network and on leased point-to-point 2-wire telephone-type circuits.

11. COMMERCIALIZATION

Modems have undergone an evolution not only in their algorithms and capabilities but also in the way they are implemented, packaged, and sold. Up until the mid-1990s, most modems were standalone devices that included all the hardware and software needed to connect between a telephone line and a serial port of a computer. In this "hard" configuration, the modem consists of interface circuitry, a controller, and a dedicated signal processor that runs the modem algorithms. Hard modems can be sold as external devices (box modems) that connect to a computer via a serial port, or as an add-on board such as a PCI card that interfaces directly to the computer's internal bus. In either case, hard modems do not require any data processing resources from the host computer.

To reduce costs, the host computer can instead handle some of the modem functions. For example, the controller, which provides the command set (usually the Hayes command set[4]), can be moved from the modem to the host processor. The computer can process the commands and then direct the modem datapump operations, eliminating a microprocessor from the modem. The amount of processing power required to provide these functions is minimal, resulting in very little impact on the host computer. However, since this "controllerless" configuration requires a host computer to operate, it is sold as either an add-on board or as part of the computer itself, by integrating the modem onto the motherboard. The driver software for the modem contains the software needed to implement the controller, making the modem processor-dependent and operating-system-dependent.

The third generation of modems is the software modem. In this form, both the controller and the datapump operations are implemented on the host computer. This eliminates the signal processor from the modem reducing cost even further. The modem hardware then consists of only the DAA (the data access arrangement — an interface to the telephone line), and a bridge to the computer bus. All the signal processing is done on the host computer by executing software included with the driver for the modem. As with the controllerless modem, this type of modem is dependent on the processor type and the operating system. In addition, the datapump operations are much more CPU-intensive than the controller, resulting in reduced performance of the host computer for user operations while the modem is active. However, new protocols and enhancements can be added with a simple software upgrade of the driver.

By 2001, almost half of the 110 million V.90 modems shipped annually were soft modems with controllerless and hard modems sharing the remaining fraction [10].

12. PERFORMANCE

The V.90 standard supports downstream rates of $\leq$56 kbps. In practice, however, these modems operate

[4] The Hayes Command Set, which includes various commands prefixed by "AT," was introduced by Hayes Microcomputer for an early 300 baud modem and became a de facto industry standard.

at lower speeds; 48–52 kbps is more typical. In part, this is due to the condition of the user's local loop. A long loop, or one that has anomalies, introduces distortions and noise that cannot be fully compensated for. The digital path through the telephone network may also introduce digital impairments that reduce the number of bits available for transmission. Another factor is the regulatory limits on the signal energy that reduce the available set of codewords such that speeds greater than 53 kbps are not possible even with very clean lines. During training, the modems choose a set of codewords and other parameters to maximize the data rate.

It is important to note that although data rate is used overwhelmingly as the critical performance measure, latency can have a greater effect on performance. In the typical use of a modem for an Internet connection, many protocols are layered on top of each other that require blocking of data or lookahead. Each of these layers adds delay in transmitting or receiving each byte of data. With TCP/IP, LAPM, spectral shaping, and other operations, round-trip delays between the endpoints can add up to more than 100 ms. For many separate accesses to small amounts of data, such as loading Webpages that contain embedded images, this latency, not the throughput, will dominate performance.

The actual rates and latencies obtained will depend on these factors as well as the details of the manufacturer's implementation. Although the standard fixes the protocol to permit interoperability, there is still great leeway in the implementations possible, creating significant performance differences between modem models.

13. FUTURE DEVELOPMENTS

ITU Recommendation V.92 [11] contains enhancements to V.90, most notably a significantly faster startup time. The negotiation and training time was reduced from as much as 45 seconds in V.90 to less than 15 s in V.92. This feature reduces not only the time a user has to wait for a connection but also the load on Internet providers by making it possible for users to drop and resume connections as needed. The faster reconnect times enabled another useful feature that V.92 adds: "modem-on-hold," which allows the modem call to be put on hold so the telephone can be used for a voice call. The data call can then be resumed where it was left off without dropping open connections.

Data transmission rates were improved in V.92 in two ways. Coding techniques that take advantage of the digital PCM network infrastructure have been applied to the upstream channel. Instead of using V.34 at rates of $\leq$33.6 kbps, V.92 modems can transmit data upstream synchronously at rates of $\leq$48 kbps while still allowing downstream rates of $\leq$56 kbps. Improved compression based on the V.44 standard further improves data rates.

Since telephone calls are carried digitally within the telephone network at 64 kbps, it is not possible to go faster than this over a regular telephone line without modifications to the telephone network. Thus, there is unlikely to be any significant improvement in speed for telephone modems. However, it is possible to use the same physical wire on the local loop for much higher data rates if it is connected to different equipment at the central office. ISDN services provide two 64-kbps data channels and one 16-kbps signaling channel over the same twisted pair. Digital subscriber line (DSL) systems allow rates of $\leq$1.5 Mbps, and several systems that move data at rates exceeding 10 Mbps have been proposed — all using the same local loop. However, in all of these cases the local loop must be terminated at the central office with equipment different from that used currently for regular telephone lines.

Eventually, V.90 and V.92 modems will be replaced by DSL, cable modems, or other broadband access solutions for many data needs, such as home and business Internet access. However, the universality of basic telephone lines will ensure that telephone modems will remain present in laptops, settop boxes, games, and other devices for the foreseeable future.

BIOGRAPHY

Brent Townshend received a B.A.Sc. degree in engineering science from the University of Toronto in 1982. He received his M.S. in computer science, and M.S and Ph.D. degrees in Electrical Engineering all from Stanford University, Stanford, California, in 1983 through 1987. In 1987, he joined AT&T Bell Laboratories in Murray Hill, New Jersy, where he worked in the Acoustics Research Department on speech coding, speech recognition, and psychoacoustics. He also developed object-oriented programming systems and software architectures. Dr. Townshend moved to Montreal, Canada, in 1990 where he started Townshend Computer Tools. This company created new intellectual property through research and development, and then licensed the technology to other companies better suited to market the results. Example projects include the 56k modem, FAX coding systems, cryptography systems, and digital audio products. In 1997, Dr. Townshend cofounded Ordinate Corp in Menlo Park, California, a speech recognition/language testing company. Concurrently, Dr. Townshend has held adjunct or consulting professor positions at McGill University, Quebec, Canada, (1990–1994) and Stanford University (1994–2001) and is actively involved in the Silicon Valley venture capital community. Dr. Townshend has published numerous articles and over 20 patents. His current research interests include speech recognition, language assessment, music analysis, and digital photography.

BIBLIOGRAPHY

1. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**: 379–423, 623–656 (1948).
2. D. V. Batorsky and M. E. Burke, 1980 Bell System noise survey of the loop plant, *AT&T Bell Labs. Tech. J.* **63**(5): 775–818 (May–June 1984).
3. I. Kalet, J. E. Mazo, and B. R. Saltzberg, The capacity of PCM voiceband channels, paper presented at IEEE Int. Conf. Communications, Geneva, 1993.
4. U.S. Patent 5,801,695, Sept., 1998, B. Townshend, High speed communications system for analog subscriber connections.

5. ITU Recommendation G.711, *Pulse Code Modulation (PCM) of Voice Frequencies*, International Telecommunication Union, Nov. 1988.
6. ITU Recommendation V.90, *A Digital Modem and Analogue Modem Pair for Use on the Public Switched Telephone Network (PSTN) at Data Signaling Rates of up to 56,000 bit/s Downstream and up to 33,600 bit/s Upstream*, International Telecommunication Union, Sept. 1998.
7. T. Welch, A technique for high performance data compression, *IEEE Comput.* 8–19 (June 1984).
8. ITU Recommendation V.42bis, *Data Compression Procedures for Data Circuit-Terminating Equipment (DCE) Using Error Correction Procedures*, International Telecommunication Union, Jan. 1990.
9. R. D. Gitlin, J. F. Hayes, and S. B. Weinstein, *Data Communications Principles*, Plenum Press, New York, 1992.
10. *Worldwide Modem Markets—a Semiconductor Perspective Year-to-date 2001 Market Wrap-up with Outlooks through 2005*, VisionQuest, 2001.
11. ITU Recommendation V.92, *Enhancements to Recommendation V.90*, International Telecommunication Union, Nov. 2000.
12. L. Brown and M. Davidson, PCM modem technology: Extending V.34, *Commun. Syst. Design* (Dec. 1997).
13. L. Brown, PCM modem design: V.90 characteristics, *Commun. Syst. Design* (June 1998).
14. ITU Recommendation V.42, *Error-Correcting Procedures for DCEs Using Asynchronous-to-Synchronous Conversion*, International Telecommunication Union, Oct. 1996.

VERY HIGH-SPEED DIGITAL SUBSCRIBER LINES (VDSLs)

J. CIOFFI
Stanford University
Stanford, California

1. INTRODUCTION

Few, if any, other technical inventions have had as much impact on mankind as Bell's invention of the telephone in 1876. Nearly a billion telephone lines exist worldwide, reliably enabling communication from nearly any point on earth with any other point. However, lesser known but potentially of yet greater and more lasting impact was Bell's invention of the less exciting twisted-pair cable in 1881. Since that time, twisted pairs have carried electrical phone signals reliably to those billion phones. But despite their relative age, twisted-pair lines have only begun to realize their potential to carry computer, television, radio, and other digital signals of the information age in addition to the plain old telephone service (POTS) they've quietly and reliably carried for over a century. Digital subscriber line (DSL) technology has emerged to service tens of millions of customers, with hundreds of millions envisioned before 2010. The latest in DSL technology is VDSL, which is overviewed in this article. Treatments of the earlier forms of DSL, such as the most heavily used and deployed ADSL, can be found in several books [1–5].

Very high-speed digital subscriber line (VDSL) service delivers very high bit rates over ordinary phone lines to customers. VDSL provides tens of megabits per second to those customers who desire broadband entertainment or data services while prudently leveraging infrastructure costs of fiber, avoiding wireless equipment placement, and rendering unnecessary coaxial-cable reengineering. VDSL modems can be programmed to carry symmetric (data network or LAN extension) or asymmetric (Internet Web surfing or TV) data rates over a variety of phone line types. Specific VDSL applications also encompass videoconferencing, telecommuting, telemedicine, distance learning, home shopping, and so on.

Figure 1 addresses the often asked question "How can DSLs go so much faster on the same phone lines that supposedly were already close to operating at theoretical limits with 33 kbps and 56 kbps voiceband modems?" Figure 1a illustrates the use of voiceband modems, specifically calling attention to the long transmission path that includes voiceband channels through telephone company switches. It is the bandwidth allocated by the switch to voice calls that limits the bandwidth of voiceband modems (i.e., the switch does not know it is a digital modem and treats the signal like voice). Of particular note, *it is not the phone line that currently prevents transport of broadband data signals to the customer*. Capacities of phone lines depend heavily on length, but a huge percentage are capable of carrying very high data rates if the narrowband switch can be avoided. DSLs (Fig. 1b), avoid the voice switch and instead have a transmission path that includes only the twisted pair. Digital signals are extracted by a second modem in the phone company central office (or a fiber-fed cabinet in the street) before they get to the switch and are then routed as broader bandwidth digital signals through an appropriate broadband network based on technologies such as ATM and IP. VDSL delivers the highest data rates on the shortest of the twisted pairs.

1.1. VDSL History and Basics

The VDSL concept was first published in 1991 [6], and was a result of a joint Bellcore–Stanford research study into the feasibility of 10+ Mbps symmetric and asymmetric data rates on short phone lines. The study specifically searched for the potential successors to the then more prevalent 1.5 Mbps HDSL and the then relatively new (then also only 1.5 Mbps) ADSL.[1] The first serious suggestions that VDSL be standardized first came almost simultaneously in the American T1E1.4 group from Amati Communications Corp. [7] and in the ETSI group from British Telecom [8] as a function of the first ADSL trials in Britain at 2 Mbps and 6 Mbps (supplied by Amati

[1] The author would like to gratefully acknowledge encouragement and support from then-retiring Dr. Joseph Lechleider of Bellcore, who encouraged and financially supported this early "VDSL" study.

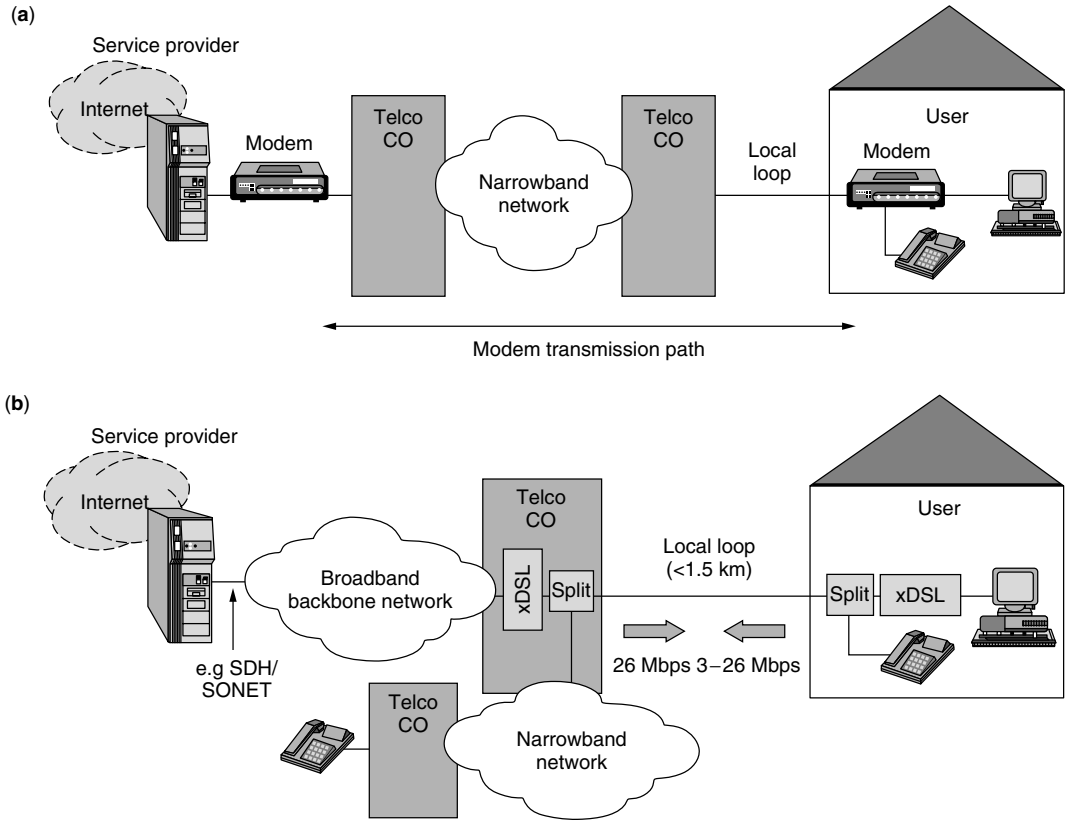


Figure 1. DSL versus voiceband modems: (**a**) voiceband modem communication model; (**b**) DSL modem communication model.

to BT) where discussions on the potential of higher speeds on shorter fiber-fed ONU-based copper loops were active between the two companies. VDSL history also has connection to early high-speed ATM network studies in the ATM forum [9] and DAVIC [10], which attempted to transmit 26 and 52 Mbps symmetrically on one or more twisted pairs over very short distances (<100 ms) for local-area networks. While the latter ATM and DAVIC efforts are somewhat forgotten, in that instead 100base-T and now Gigabit Ethernet became the methods of ubiquitous use for internal computer networks on twisted pair, these early ATM and DAVIC efforts did also provide useful information to the development of present-day VDSL standards.

Asymmetric VDSL is viewed more as a residential service, introduced into the existing twisted-pair loop that carried only POTS or ISDN-BA (Integrated Services Digital Network, Basic rate Access) services. A general case of asymmetric service deployment is illustrated in Fig. 1b. Coexistence of POTS/ISDN and VDSL signals in the same twisted pair is allowed by the separation in frequency of their transmission bands, provided by the service splitter, shown as "split" in Fig. 1b.

Figure 2 presents the spectrum of the POTS/ISDN and VDSL signals running over the customer twisted pair. Fiber loop-carrier systems are being deployed worldwide to bring the high-bandwidth promise of fiber closer to groups of hundreds of phone company customers whom are served by what is called an "ONU" (Optical Network

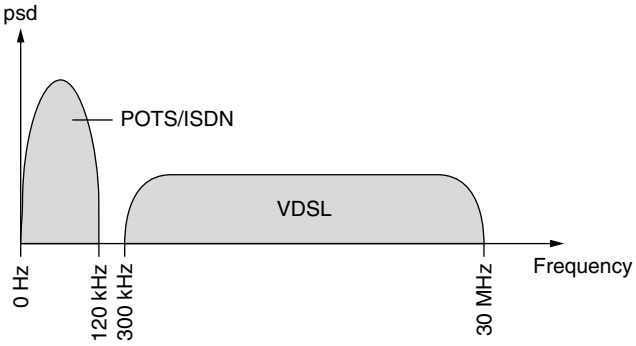


Figure 2. Spectral allocation of VDSL and ordinary phone services.

Unit). Twisted pairs emanate from the ONU and connect to the customers, completing the path started with fiber to the ONU. Because of the service splitter operation, ordinary phone service appears the same to customers who seldom know that they are served by the ONU. (ONU's serve about 15% of the American population and often a smaller percentage in other countries.) The ONU is usually located less than about 1 km (3000 ft) from the customer. Desired asymmetric VDSL data rates over these distances are 13–26 Mbps downstream and 2–3 Mbps upstream, to allow delivery of digital TV (DTV) and high-definition TV (HDTV) services, superfast Web surfing and file transfer, and virtual offices at home. For shorter distances ($\leq$300 ft) the downstream rate can be as high as 52-100 Mbps,

offering simultaneous delivery of several DTV or HDTV channels.

Symmetric VDSL is usually viewed as a business service, allowing 10-Mbps connections over a twisted pair of up to 1 (5000 ft) and 25 Mbps over shorter loops (<3000 ft). A typical application is Ethernet/IP LANs' interconnection with data rates of 100 Mbps (on 4 coordinated phone lines) or 10 Mbps on 1 or more pair, respectively, over a twisted pair between buildings in a corporate campus environment, as shown in Fig. 3 or between a telco central office and a business. Fiber may connect one central building with a network provider while the other buildings within the campus are connected by twisted pairs. Local area networks may service the individual buildings at 25.6 Mbps for ATM or 10 or 100 Mbps for Ethernet/IP over coax, fiber, wireless, twisted-pair, or other media. To expand the network between buildings making use of the existing phone lines, VDSL modems implement the high-speed connection. The traditional method to provide building-to-building connectivity requires use of many phone lines, each at 1.5 Mbps, with inverse multiplexers. This method, shown at the bottom of Fig. 3, is very expensive (inverse multiplexers are much more expensive than VDSL modems) and wastes twisted-pair bandwidth.

There has been an increased interest to use VDSL to transmit multiple T1 (1.5 Mbps), E1 (2 Mbps), and other T3 (45 Mbps) tributary datastreams to serve business customers in the area close to the local exchange (CO — central office). A typical deployment scenario, usually called "CO-based VDSL," is presented in Fig. 4. In this case, connections from the CO to the business are usually established on leased twisted pairs.

1.1.1. ADSL Extension. ADSL is now acknowledged as a successful telecommunications service with tens of millions of lines in deployment, and hundreds of millions to be deployed in the next decade or two. However, in its earliest days of standardization, ADSL faced the severe criticism that even its greatest standardized speed of 8 Mbps was too slow to match the data rates possible on what are called "hybrid fiber coax" (HFC) networks. HFC networks upgrade existing unidirectional cable-TV networks in two ways:

1. The cable TV networks are made two-way in HFC by replacing upstream-blocking filters in TV by more sophisticated two-way non-upstream-blocking filters.
2. The cable TV networks are increased in bandwidth in HFC by replacing coax near the TV head-end by

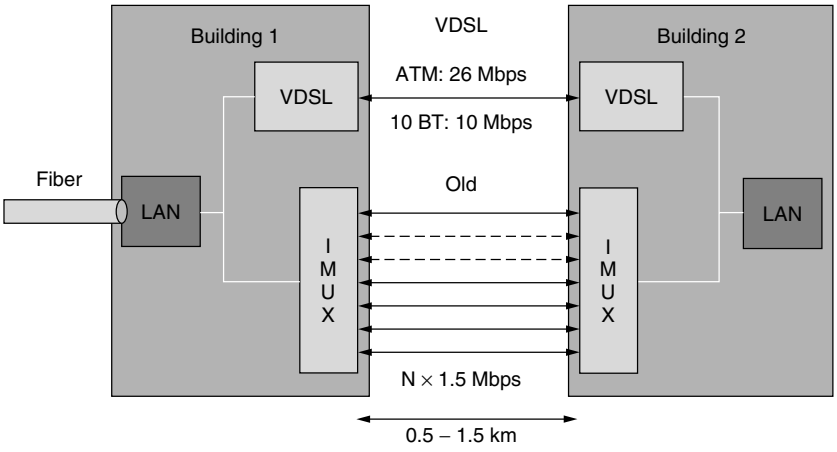


Figure 3. Symmetric VDSL for campus LAN interconnection.

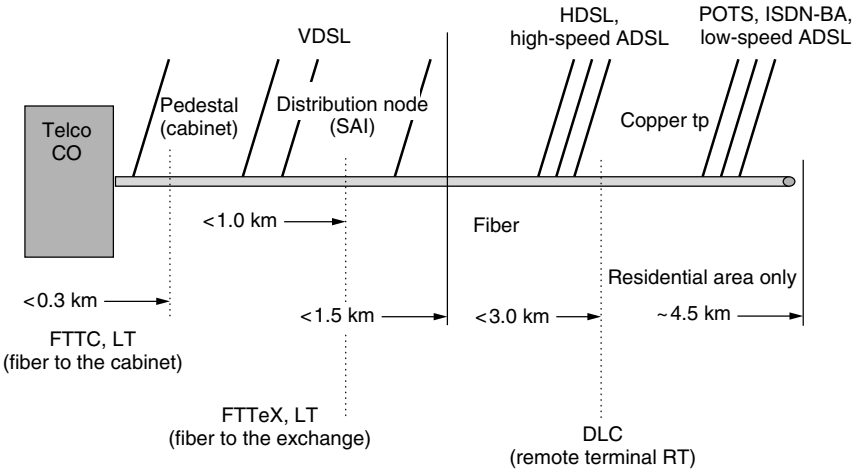


Figure 4. CO-based VDSL for business services.

fiber, multiplexing multiple separate coax signals on the same fiber in different bands, and thus rendering fewer subscribers per coaxial section (thus sharing less bandwidth).

Phone companies believed in 1994 and 1995 that they must replace their existing phone line networks with HFC, and several attempted to do so, only to later find costs prohibitive.

ADSL was already bidirectional, but with limited speeds downstream and yet more limited speeds upstream. (The asymmetry in ADSL allowing a longer line length for reliable transmission of a given data rate [1].) ADSL in 1994 and 1995 was perceived by telcos as too slow with respect to HFC. VDSL emerged from ADSL proponents as a next higher-data-rate step for ADSL — if fiber can be installed in HFC, then why not install it in existing networks when there are customers ready to pay for higher speeds than ADSL and instead use fiber-based loop shortening to increase the speed and symmetry of ADSL?

The initial VDSL architecture, shown in Fig. 5, ensued as the future of DSL deployment when (and if) customers were willing to pay for more and more fiber. The VDSL story is thus far more incremental in nature for telco deployment than HFC, which required an entire network to be replaced on the hope that enough customers would pay for it. In 1995 and beyond, this type of incremental DSL deployment won increasing favor with telephone service providers and is the actual mode of choice today. Cable suppliers continue to upgrade their TV networks to HFC at significant cost, but it becomes increasingly clear that the merits of DSL will prevail for nearly all services other than (unidirectional) analog and newer digital television delivery, for which cable seems to be still well conceived and currently the favorite.[2] The optional splitter of ADSL is preserved in VDSL so that analog voice service can be protected and preserved on the same line as VDSL. The cost of the fiber section is high, but can be divided by the number of customers served. As fiber penetrates closer and closer to the customer, that cost is shared by a smaller number of customers. Thus an important tradeoff in VDSL is the length of the fiber versus the length of the remaining copper. There is no single good answer to this tradeoff as it depends on applications, customer willingness to pay, transmission method, and, of course, the cost of the fiber — however, VDSL allows a wide range of trade-off as this chapter will illustrate.

Figure 6 illustrates data rates for both upstream and downstream VDSL transmission on 24-gauge twisted pair (0.5 mm European) versus loop length for the American

[2] However, it is conceivable that Internet-based approaches to TV may allow an opportunity for DSL.

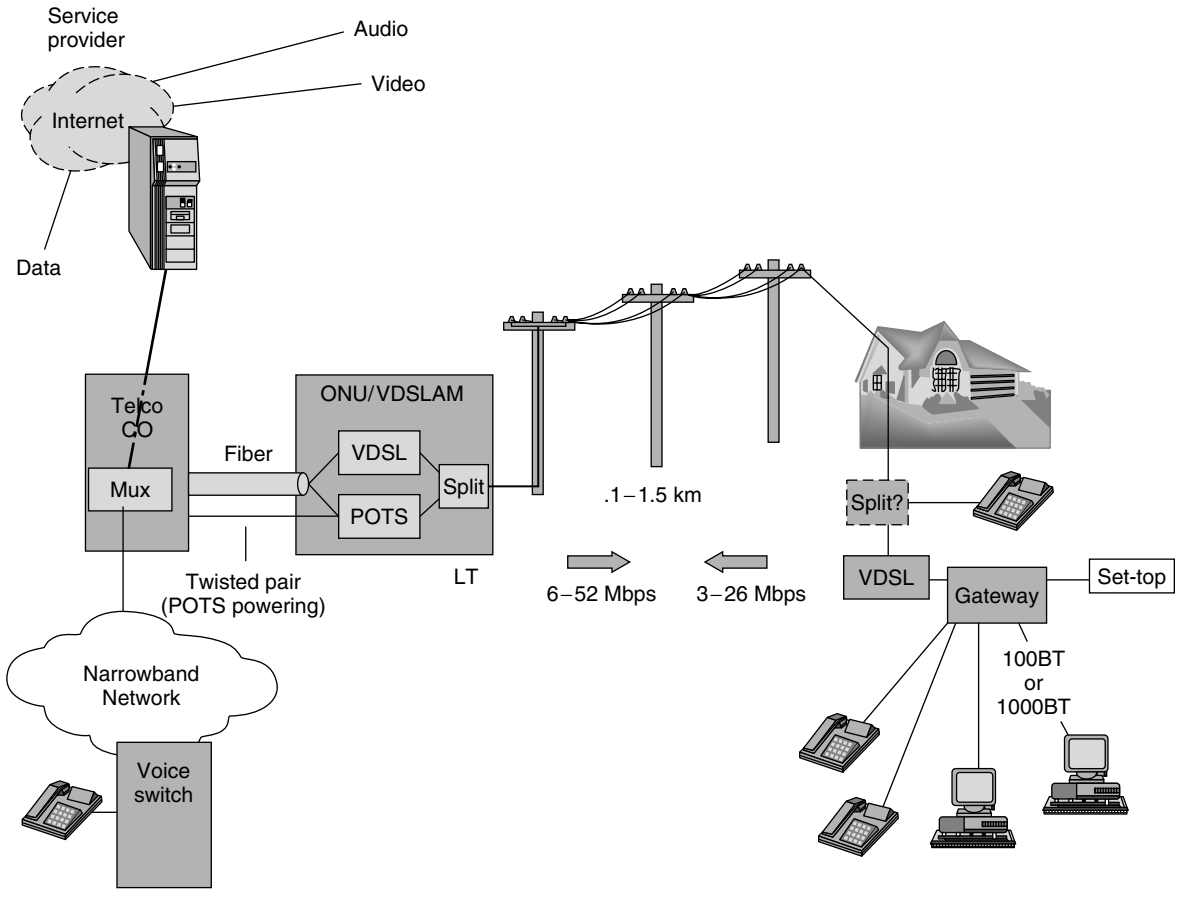


Figure 5. VDSL system architecture with ONU/fiber loop carrier system.

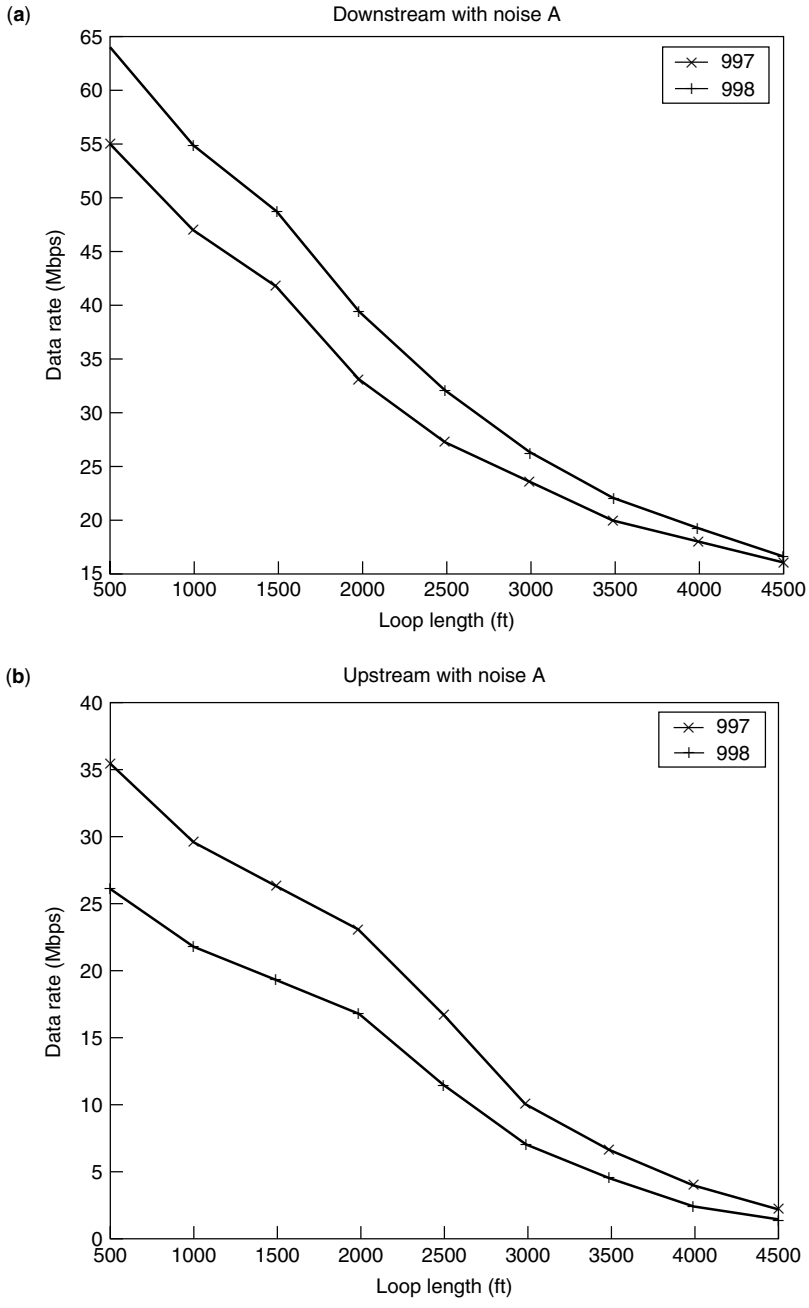


Figure 6. Downstream (**a**) and upstream (**b**) data rates for American and European standard DMT VDSL.

(998 curve) and European (997 curve) DMT VDSL standards.[3] Clearly, the data rates are quite high on short loops, ensuring a greater individual bandwidth per user than cable networks (which customers must share inherently in architecture).

The premium paid in range loss for symmetry of data rate is less as loop lengths get shorter, and so then VDSL also offers a way to offer increasingly symmetric individual service to customers. As the number of small businesses worldwide explodes, most often in urban areas where line lengths are short, the potential for symmetric support of the voice, conferencing, peer-to-peer gaming or working, "home" Web server upstream bandwidths is then evident with VDSL. Today, an increasing number of service providers consider exploring early VDSL deployment for business services, particularly what is often called "fractional T3" support. In 2001, only 12,000 of the nearly 10,000,000 businesses in the United States were connected by fiber (and a smaller percentage in other countries). One large fiber installer and service provider estimates that they can increase this number by 2000 in the near future with cost of $1 billion [17]. Thus, VDSL will play a major role in the future service offerings to small and many large businesses before fiber connection is financially viable

[3] Achievable data rates for the other "single-carrier" standard will be less, with these curves as an upper bound [1].

or completed. Other service providers still believe that support of video and television may be viable also in the future, although the economics of this application may be harder to justify versus cable.

The wealth of ADSL installations also then mandates another practical requirement that VDSL service must be compatible in many respects with existing ADSL. Existing customers with ADSL modems on their premises (perhaps in their portable computers) may move or travel into another area, or may live in an area where VDSL arrives, and will still want their ADSL modem to function as it always has. Thus, the ONU-side modem in Fig. 5, often called the LT (line termination) in VDSL, would need to support ADSL service, but would, of course, also allow higher speed service if/when that customer decides to purchase a higher-speed VDSL modem and the higher-speed VDSL service. Also, a customer who buys a VDSL modem will certainly want that modem to work with an existing ADSL connection at lower speed if that is all that is available. In addition to interoperation with ADSL modems, VDSL modems must also be compatible spectrally with ADSL modems that may share the same binder and with existing home premises networks, as well as with perhaps a plethora of other standard or nonstandard systems that exist in/near the cable (for instance, HDSL, SHDSL, or nearby ham radio). Subsection 7.1.3 deals more directly with this spectrum issue.

Overall, VDSL offers a mechanism for service providers to upgrade their networks incrementally and with continued profitability to include increasing amounts of fiber, approaching an ultimate goal of a network based entirely of fiber. Telephone line service providers have a very powerful story and future with VDSL, now following ADSL. The author suggests that perhaps the term "VDSL" will become synonymous with DSL in the near future.

2. VDSL ARCHITECTURE

With VDSL providing a significant growth opportunity for DSL in the future, the question becomes the details of "where, when, and how" to install fiber. Replacement of copper by fiber first occurs in cables closest to the central office. The cost of trenching for the new fiber in the cables closest to the central office can be shared by all the customers who are served by that cable of wires replaced. Further from the central office, newly installed fiber increasingly services fewer customers, ultimately just one customer when it connects to a specific customer premises. Thus, fiber installation cost increases per customer roughly with distance from the central office. In 2001, 15% of the American network had what is called "fiber feeder" as shown (and marked just "fiber") in Fig. 5. Virtually no fiber went to residential customers in 2001. Only about 1% of the many small businesses (about 10 million) in the United States are connected by fiber directly.[4] Initially with the fiber loop carrier system being deployed today (Fig. 5), usually only the POTS/voice connections to residential and small business customers exist. However, VDSL or any DSL is increasingly added by placing the DSLAM[5] at the end of the fiber, sometimes known as an *optical network unit* (ONU). Many phone companies have massive expenditures and fiber deployments under way to augment their ADSL service deployments so that line lengths are shortened and ensure higher ADSL speeds more reliably. The number of such DSL-enabled ONUs thus will grow rapidly in future years. This position of a DSLAM is often called a "line terminal" (LT), depending on the country and the exact type of DSL deployed (but unfortunately the use of these terms is inconsistent). We will use the term LT in the ensuing part of this article. VDSL-enabled DSLAMs in 2001 were just being introduced on a experimental basis in a few locations, but their number can be expected to increase especially as backward-ADSL-compatibile VDSL DSLAMs enter the market from major suppliers in 2002.

The trend toward more fiber and shorter twisted pairs thus will continue with time and VDSL. Splitter circuits can be used at both DSLAM and customer-premises ends of the line to preserve the existing POTS service in analog. Additionally, the high speeds of VDSL allow multiple digital voice signals to also be carried to the customer. Within the customer premises (home or business, home is shown in Fig. 5), a gateway is used to demultiplex the various VDSL signals and route them to the appropriate applications device, which could be a phone, computer, or television/entertainment system. Within the central office, another demultiplexer/multiplexor can be used to extract application signals and route them appropriately. Figure 5 presumes a heavy use of internet delivery, as well as Ethernet distribution within the home, but other mechanisms for such multiplexing are possible and discussed. In particular, wireless LANs [18] and/or home-phone distribution systems [19] have also been used. The economic tradeoffs, speeds, demand for services in DSL make the actual tradeoff of length of fiber versus length of copper issue difficult to assess in 2001.

However, a point of service potentially is the so-called "distribution point" (or sometimes CSI), which typically is within 3000 ft of the customer, is shown in Fig. 4. This point is typically where larger cables are terminated and smaller cables servicing up to a few hundred customers begin. Usually, the box at the CSI basically serves as a cross-connection point for twisted pairs. However, the entire distribution-point box can be replaced if fiber feeds this CSI point. VDSL modems placed in such an enclosure then energize the subscriber-side twisted pairs that emanate. Power and size constraints are at least as difficult at this point as at the remote terminal, usually leaving a small area (a few square inches) and about 1 W of available power per DSL customer. Another intermediate point is yet closer to individual customers is often called the "cabinet" or "pedestal." Usually only 4–16 customers are served from the pedestal with individual twisted pairs

[4] A large business typically has a small central office or other switching/routing mechanism within its largest campuses, and this switch will often be connected by fiber to the larger telecommunications network.

[5] DSLAM is a *digital subscriber line access multiplexer*, a piece of equipment that houses several DSL modems at the service-provider end of the telephone line.

emanating to these customers. The pedestal again is normally a cross-connection point for telephone lines, but fiber can be deployed to this physically accessible point, and a VDSL modem deployed there. Very high speeds are possible on the resulting phone lines of 100 m or less, potentially hundreds of Mbps or more, higher yet than current VDSL.

Placing fiber to each successive point is increasingly costly because the cost of the fiber per subscriber necessarily increases as the number of customers decreases. Considerable "digging," "wall cracking," or physical labor may be necessary as the fiber proceeds closer to the customer. However, in the future if the customer demands higher bandwidth, then potentially higher revenue is possible also to pay for the fiber deployment costs. Ultimately fiber can be run to the home or even into the home to the desk/TV-top. The key to VDSL is the incremental deployment if and where customers are willing to pay for more fiber. The cost of deploying fiber can be from $250,000 to $1,000,000 per half-mile in areas of reasonable customer density.

2.1. Unbundling Issue

Colocation of VDSL modems is yet more difficult when the VDSL modems are not in a central office. This is because sharing of space by different service providers at the cabinet, carrier service area (CSA), or distribution point is physically difficult (there is not enough space). Today, this is a hotly debated issue in DSL deployment, and a single solution has not yet emanated. Some service providers accuse incumbent local exchange carriers of installing more fiber just to complicate colocation. Potential solutions for VDSL colocation are to

1. Standardize the backplane interface and card size(s) of VDSL so that many service providers may plug into an ONU.
2. Provide separate fibers to the ONU for each service provider and divide the existing small space according to the fibers that enter.
3. Use the HFT concept and colocate at the central office where more space is available.
4. Provide higher-level unbundling at layer 2 or 3 in the protocol stack.

Other solutions may evolve. VDSL standards to date have only encompassed colocation by mandating that a single spectrum type shall be used in all VDSL transmission types to minimize crosstalk between VDSLs. Largely, current VDSL standards are just beginning to address the intricacies of the VDSL colocation issue. In addition, the American ANSI group TIE1.4 has a new standards' effort in Dynamic Spectrum Management (DSM) that offers very attractive alternatives for all co-location alternatives.

2.2. POTS Splitters in VDSL

Splitter circuits for ADSL and VDSL are described in basic detail and design in Ref. 1, Chap. 3. For VDSL, the necessity of a splitter continues to receive attention. The heavy use of splitterless ADSL suggests that perhaps splitterless VDSL is also advisable for compatibility and volume deployment reasons. The first splitterless VDSL proposals appeared in Refs. 16 and 17 — while these proposals in standards met with minority opposition (which is sufficient to block standardization), most advocates of the design in Section 3 are pursuing various forms of splitterless operation as an additional feature and option, albeit proprietary. The VDSL technology in Section 4 will not operate without splitters because of the consequent home wiring effects.

VDSL transmission designs on a splitterless channel will need to be robust to bridged taps, increasing amounts of radio interference, potential crosstalk (on same or other lines) from home services already present on the phone lines, and further signal attenuation. Such modems may also have need for control of power spectral density masks also to avoid excessive emission on the customer's premises that might interfere with local ham operators, emergency radio, or other appliances. The methods in Section 3 address these problems.

2.2.1. Common VDSL Reference Configurations. A common reference model was adopted and illustrated in Fig. 7 [20]. The interfaces and functionality associated with the two γ interfaces are common to both VDSL transmission methods. The PMS-TC (physical medium specific transmission convergence) and PMD (physical-medium-dependent) are specified for each transmission method [21,22], while the spectra of the U interfaces and the splitter are again specified in common for the two transmission methods [20]. Like ADSL, VDSL also specifies two paths, a slow or interleaved path and a fast or noninterleaved path as in Figs. 8 and 9. The former undergoes interleaving as well as forward error correction to allow for maximum impulse noise protection while the latter allows for minimum delay (no more than 1 ms in VDSL). The application specific reference is the DSLAM device that basically makes use of a subset of the functionality for a given PMS-TC interface, converting from the γ interface. The γ can be an ATM or STM interface, for instance at speeds well above those of an individual DSL modem. The application specific layer then extracts the pertinent bits (presumably set up by the normal ATM method for setting permanent or virtual channels) for each of the fast and slow data paths through the VDSL modem, formats those bits into a known and reversible format within each stream and then forwards fast and slow bits to the PMS-TC interface.

Indeed, the greatest challenge of VDSL, presuming a working modem, may be the high-speed extraction and identification of the individual application signals, likely sent through an ATM or IP switching system. This section notes a few characteristics of interest for transmission.

One can envision the application devices as multiplexing and demultiplexing the applications signals, of which some may be simultaneously present in both the fast and slow buffers, and formatting them for/from the modem itself. The high-bandwidth data channel created by VDSL may allow for numerous applications to flow simultaneously.

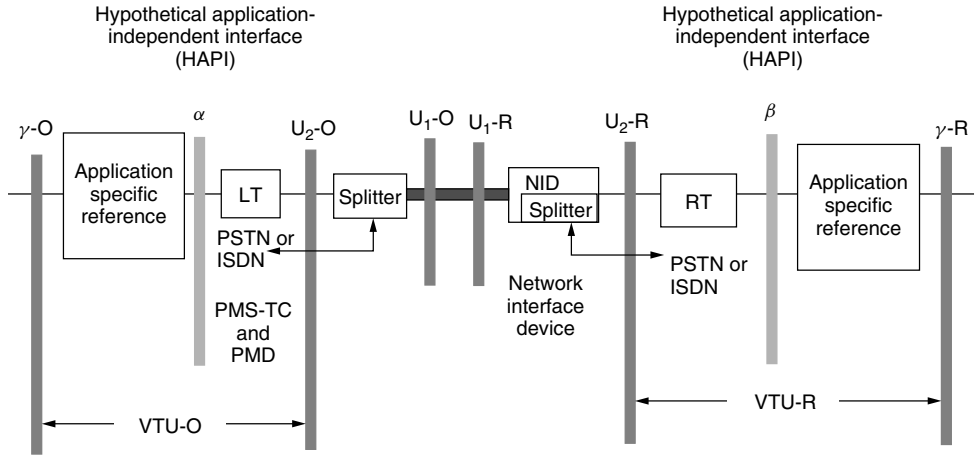


Figure 7. VDSL reference model from wang [20].

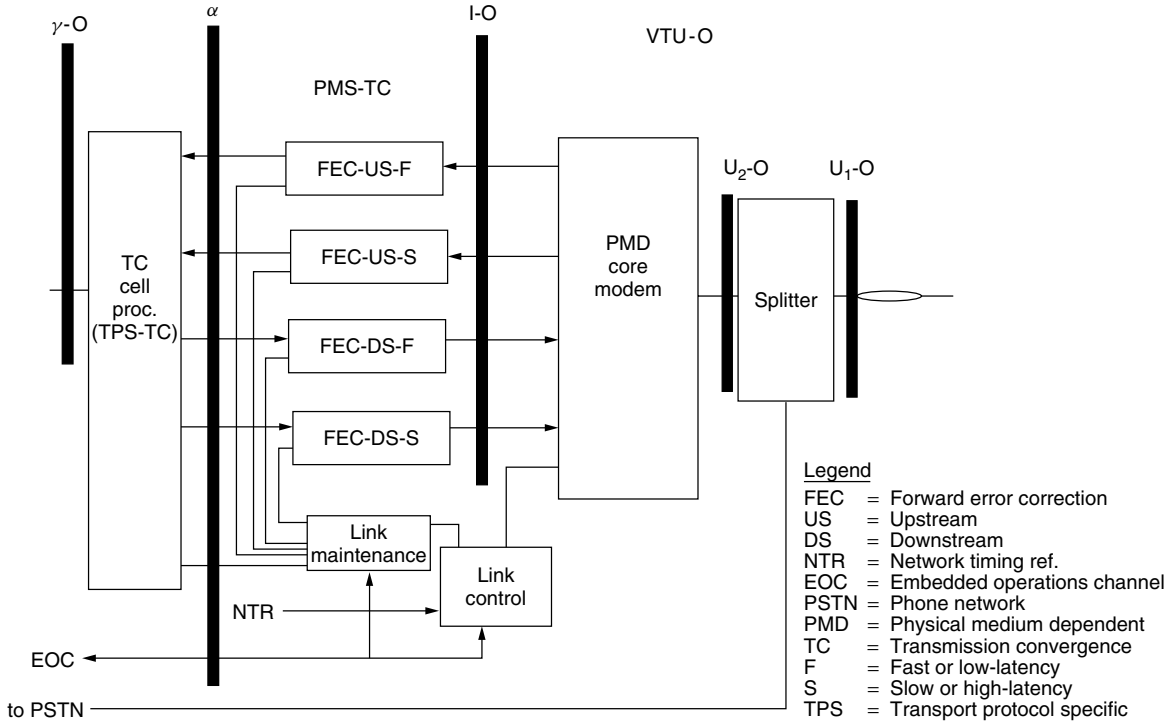


Figure 8. VTU-O reference model.

The VDSL standard Part I [20] has an elaborate list of operational and maintenance capabilities to which we refer the reader. As with previous DSLs, the main parameters of interest are the state of the modems, the likelihood or presence of errors on the link, and the synchronization of network functions. VDSL allows passage of the 8-kHz network timing reference.

2.3. VDSL Spectrum Issues

As the highest speed DSL, VDSL uses the greatest amount of spectrum. Thus, it has the greatest concerns for spectrum compatibility. The issues of crosstalk and emissions from VDSL into surrounding telephone lines and radio receivers is more important and complicated in particular. Also, the crosstalk from existing services also affects VDSL spectrum design and performance. *Near-end crosstalk* (NEXT) is the radiation from one line's signals in one direction into to another line's signals moving in an opposite direction — in effect, a "near end" other-line's transmitter signals into the local receiver. Far-end crosstalk (FEXT) is the same effect but into signals going the same direction, thus a "far end's" other-line's transmitter signals into the local receiver. Furthermore, increasingly popular home LANS on twisted pair within customer premises will also complicate issues and tradeoffs, as there is both spectrum overlap on the same line with VDSL as well as crosstalk issues into other VDSL lines from the home-LANS. Considerable debate occurred in

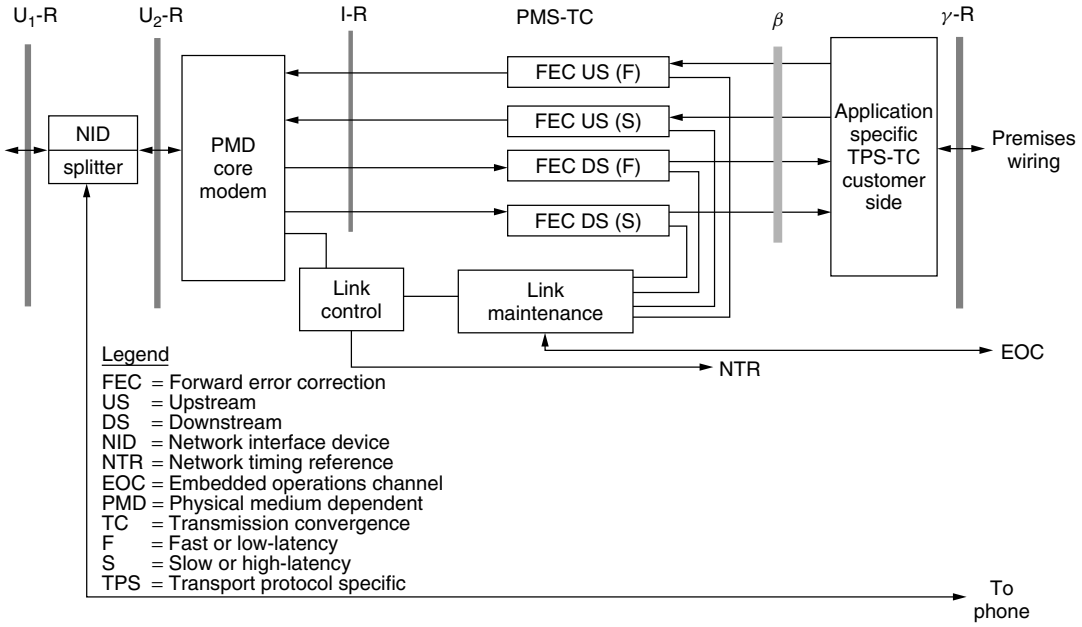


Figure 9. VTU-R reference model.

standards meetings for the design of VDSL spectrum, and there are correspondingly three internationally approved spectrum plans (presumably one selected for any specific geographic region). Unfortunately, the competitive interests between incumbent and competitive service providers and the competitive interests between different transmission techniques did not work to VDSL's best spectrum advantage so far, as significant compromise can occur sometimes for nontechnical reasons or rationale. This area has been reopened several times as VDSL spectrum management standardization continues, and as advanced transmission enhancements alter basic parameters and issues. The three options do provide considerable flexibility for the future, although fortunately, as issues are revisited by technologists, marketing persons, and national regulators. This uncertainty makes this section at this time a bit difficult to write, but this chapter will focus on technical issues and describe the three plans, illustrating the various tradeoffs. In the long-term massive deployment of DSL, the service providers and equipment/chip vendors who best comprehend all the aspects of this area will be able to garner the best business advantages in massive DSL deployment. These spectral options appear in Fig. 10 and are discussed in the next subsection. The previously mentioned and very new DSM effort targets specifically these problems.

2.3.1. Spectral Plans. The need for a fixed spectrum plan is only necessary for compatibility of the "single-carrier" plans, Section 4 of this chapter, whereas the digital duplexing of the DMT spectrum allows arbitrary placement of band edges without excess-bandwidth penalty (although a 7.8% cyclic prefix penalty is necessary, see Section 7.3). It is possible that the plans in Fig. 10a,b will be replaced by those that fully consider all aspects of applications and deregulation in the future, or may be dynamic as in DSM. The 3rd international spectral plan in Fig. 10c encompasses the possibility of spectrum flexibility. This option is implemented only in the DMT VDSL standard [22].

2.3.2. Robustness. VDSL must be able to accommodate frequency-selective disturbances, the best known of which are bridged taps of different lengths. Bridged taps occur in the loop plants of all operators (even though some try to deny it) and in particular are extremely pronounced in occurrence when splitterless designs are used. Immunity to bridged-taps is here called "robustness." Although it is impossible to avoid their effects completely, it is highly desirable for performance to degrade somewhat gracefully with bridged taps; for a symmetric service this can be interpreted as maintaining the ratio of upstream rate to downstream rate close to 1 even if total sum data rate up and down decreases slightly. In this symmetric case, huge rate loss in one of the directions because of bridged taps would be highly undesirable. In asymmetric transmission, it is desirable to maintain the ratio of asymmetry under different bridged-tap configurations.

First, this section illustrates the adverse effects of bridged-taps on transmission performance. Figure 11 shows the transfer function (in dB) of a 4050-ft loop, with bridged taps (66, 56, 46, and 36 ft long) and without bridged taps. The bridged taps cause the transfer function to exhibit notches periodically in frequency. As the bridged taps get shorter, the notches become more deep and move to higher frequencies. The existence of such notches (10–20 dB deep) can seriously harm transmission.

Below the graph, two different 4-band frequency plans are drawn. Note that this specific loop has very large attenuation at frequencies above 7 MHz, so the spectrum above 7 MHz is unsuitable for data transmission. Therefore, only the lower 2 bands would actually be used. Each frequency plan copes differently with this kind of

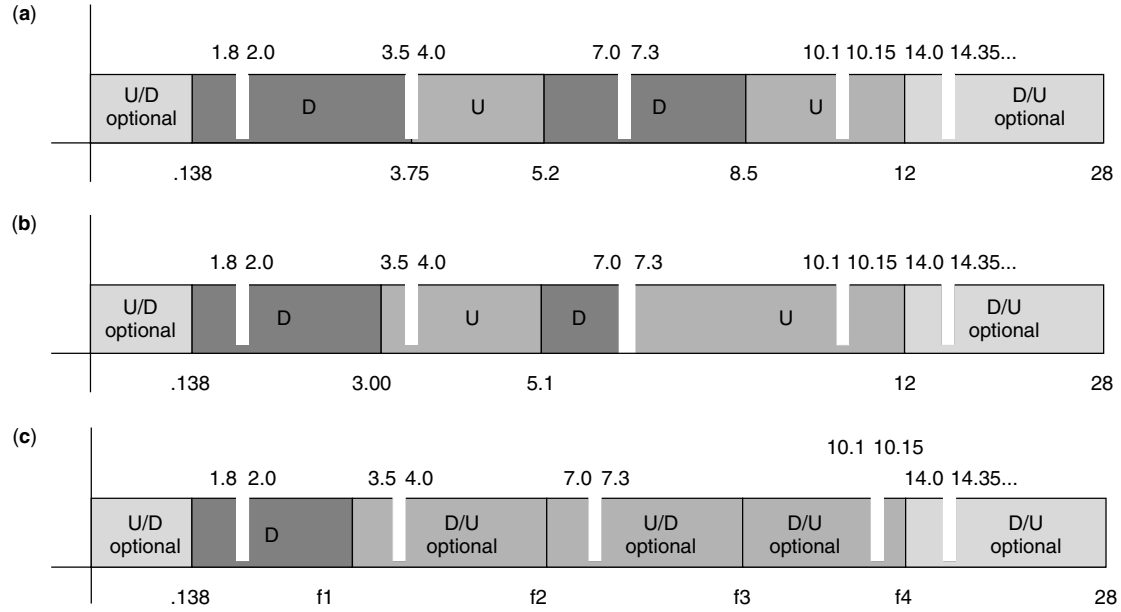


Figure 10. (**a**) plan 998 — North American VDSL spectrum (U = upstream, D = downstream) (additional radio bands notched when used at 18.068-18.168, 21-21.45, 24.89-24.99); (**b**) plan 997 — European VDSL spectrum (same unshown notched bands as 4a); (**c**) international flexible VDSL spectrum plan (f1, f2, f3, f4 determined programmably).

disturbance. If plan A were used in the presence of a bridged tap 36 to 66 ft long, then upstream transmission performance would be degraded significantly, although the downstream direction would not be affected. If plan B were used, then the downstream transmission performance would be degraded. Both plans fail to be robust.

One might argue that some other fixed 4-band plan would actually show more immunity to such situations. We explain why this is false: the bridged-tap length is not determined, so it may vary from 10 to more than 100 ft. This means that the notches may actually occur in almost any frequency of the VDSL spectrum. For any 4-band plan, there will always be a bridged tap with such a length that performance will solely be degraded in one direction. This direction is often the upstream direction using conventional models as are used in this article. However, there is one optimal frequency-division duplexing scheme, which one can prove attains the maximum possible robustness. The solution is to partition the spectrum into infinitesimally small bands and alternatively assign them to upstream and downstream transmission. Then, any frequency-selective disturbance (such as a bridged-tap) will have an equal impact on both directions of transmission. Figure 6 shows such a frequency plan, and illustrates why symmetric service is maintained.

Implementation of this optimal scheme may prove too complex,[6] so suboptimal schemes with adequate robustness may have to be used instead. By interpolating between the 4-band plan and the optimal plan, we deduce that a number of bands as large as possible is highly desirable. As the number of bands increases, the data rate loss caused by a bridged tap will be distributed more evenly between the two directions of transmission. The simulations to follow demonstrate this fact.

2.3.2.1. Simulations. The simulation results that are shown below were obtained using a popular telco simulation tool. The 4 different frequency plans that were evaluated are shown below (numbers refer to MHz):

998
up = (3.75–5.2, 8.5–12)
down = (0.138–3.75, 5.2–8.5)

997
Up = (3.25–5.1, 7.1–12)
Down = (.138–3.25, 5.1–7.1)

Digital Duplexing 5-band plan
up = (0.03–0.138, 3.08–4.78, 10.242–17.66)
down = complement of up

Digital Duplexing 7-band plan
up = (0.03–0.138, 2.5–3.5, 4.5–5.5, 11–17.66)
down = complement of up

Digital Duplexing 15-band plan
up = (0.03–0.138, 2.1–2.5, 2.75–3, 3.25–3.5, 4–4.25, 4.5–4.75, 5–5.5, 10.5–17.66)
down = complement of up

The services evaluated were medium symmetric, long symmetric, extralong symmetric, medium asymmetric, and long asymmetric in a noise A and noise D environment [18]. For each service the reach in meters was computed both with and without bridged taps. Table 1 shows

[6] Demonstrations of full zippering have been able to suggest that at least in some situations, large numbers of alternating up/down bands are indeed feasible with acceptable implementation.

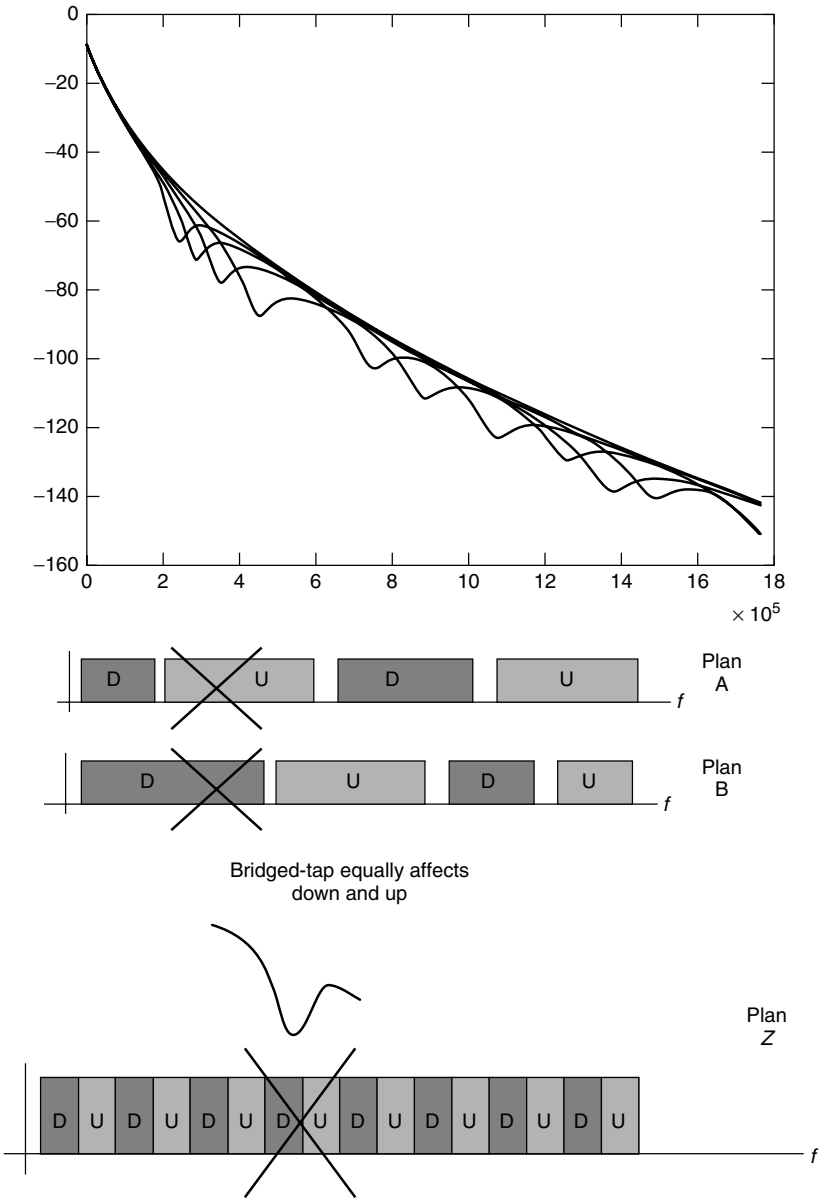


Figure 11. Illustration of robustness with bridged taps. The graph shows the insertion loss (in dB) of a 4050-ft loop with bridged-taps of length 66, 56, 46, and 36 ft (20, 17, 14, and 11 m, respectively). Below the graph, two different frequency plans are shown. When plan A is used, only upstream transmission is affected. When plan B is used, only downstream transmission is affected. In both cases symmetric service is disabled.

the resulting reaches, and the percentage gains achieved by the plans using more than 4 bands in comparison to the 4-band plans.

We immediately see from Table 1 that using a larger number of bands always improves performance. A 4-band plan has upstream data rate annihilated with the 998 or 997 plans, a particularly concerning problem for those desired symmetric service, and for those who have been told that their wishes were accommodated by those plans. It is worth noting that the reach of the extralong symmetric service is improved by more than 35% (300–500 m), when more than 7 bands are used. This service represents an important market segment, and might be the first VDSL service to be deployed. A more detailed set of results appears elsewhere [19].

2.3.2.2. Mobile Radio Noise Robustness. Mobile radio noise robustness is important also to VDSL. The area is sensitive because ingress transmissions can include those used in defense of various countries. Roughly, though, individuals without security clearances in various countries can learn of two types of disturbance:

1. Several narrowband analog voice signals in a band of a few hundred kilohertz that can be anywhere between 1 and 20 MHz and can move (or hop) in their general band location
2. Direct-sequence spread-spectrum signals of 300–500 kHz bandwidth (spread voice or data) with center frequency that hops throughout the 1–20-MHz band

One cannot specify in advance those bands to be annihilated by radio noise, and the joint operation of VDSL and of defense systems is highly preferred, especially in emergency situations.

Table 1. Bridged-Tap Robustness Results for 4500-ft 26-Gauge Loop

ANSI Noise A	998	997	5-band	7-band	15-band
Rate 4500 ft of 26 gauge					
Upstream	40 kbps	260 kbps	1.69 Mbps	2.68 Mbps	3.37 Mbps
Downstream	15.6 Mbps	15.4 Mpbs	15.6 Mbps	14.7 Mbps	14.0 Mbps

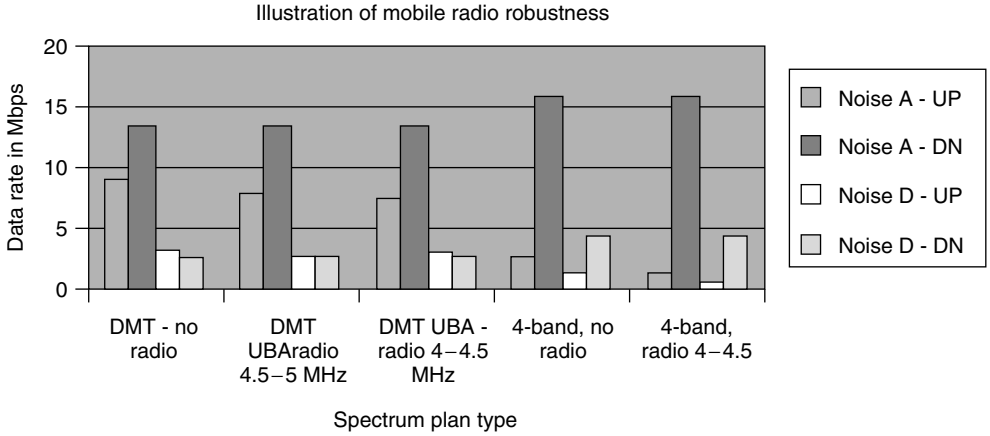


Figure 12. Illustration of mobile radio noise robustness for two plans with 1.1-km loop, 20 VDSL FEXT, noises A and D, 0.4-mm line.

Again the solution is, as shown in Fig. 12, robust with VDSL duplexing. This section models a single data signal of width 500 kHz coupling into a phone line at a level sufficient to cause loss of use of the same band on that line. This level can vary from −80 to −110 dBm/Hz depending on the length of line and other system parameters. In this case, the duplexing plans were again evaluated. The loop simulated is again a 1.1 km, 0.4-mm loop, and noises A and D [20] were used in addition to the radio noise, along with 20 VDSL FEXT. The worst-case position of the radio noise actually was the same for both plans, 4–4.5 MHz. The loss is again less for the plan with more subbands.

The 7-band plan again robustly achieves over 7 Mbps symmetrically in all cases for noise A while the 4-band plan is only 1.4 Mbps. The relative drop in performance for the 4-band plan is also larger (even though absolute data rates are smaller because of analog duplexing). For noise D, the data rates are 2.5 Mbps for the universal band allocation plan and only 575 kbps for 4-band. Again the relative as well as absolute loss is larger for 4-bands because it is not robust to radio noise. A larger number of bands (more than 8) can actually increase the noise A result for the universal band allocation to over 8 Mbps symmetric, which may be of specific interest in Europe.

2.3.2.3. HPNA. Home Phone Network of America (HPNA) or in-progress standard draft G.pnt of the ITU [14] specify a use of the telephone line bandwidth between 5 and 10 MHz for internal home phone line networking. This spectrum of course overlaps HPNA, leading to signal disruption of VDSL on the same line as well as generated large NEXT into neighboring VTU-Rs of other customers than the HPNA user.

Political maneuvering led to this issue being ignored by standards bodies where, for instance, the bands used by G.vdsl and G.pnt, standards documents produced by the same standards body, overlap. One reason for this ignorance was a stipulation by a few that unidirectional lowpass filters could be installed by every user of an HPNA network (presuming that the customer takes the time to locate his/her network interface somewhere outside his/her home and then properly installs the filter), even though that filter is not necessary for his/her internal computers to talk to one another. Remembering that all HPNA users would need to have such a filter self-installed before the very first VDSL could be installed at least begs the question as to the merit of such a solution.

A second, more elegant, solution, which does involve complexity increase for the VTU-R is described in Ref. 28 where G.pnt signals could be cancelled from a VDSL signal when on the same line or on neighboring lines as long as there were only 1 or 2 of significant amplitude. G.pnt systems however do not appear to be gaining true market acceptance, and so this may be less of a problem for VDSL. If they do, interference cancellation of G.pnt by VDSL may become a necessity in practice.

Note that this spectral incompatibility concern does not apply to Ethernet, which is typically installed on category 5 wiring that is separate and isolated by nature and design from the telephone company network. Thus, even though Ethernet uses the same band, there is no actual spectral overlap on physically colocated wires.

2.4. The Grand Debate

Dating to the days of ADSL standardization, there has been a debate over the best transmission technology to use for DSL. While ADSL standards have universally selected the specific multicarrier transmission method known as DMT after considerable deliberation and testing, and a universal consensus, a few CAP (carrierless amplitude and phase) and QAM proponents nevertheless marketed nonstandard ADSL modems for a significant time period, before most switched to and supported standardized DMT. Subsequent debate was heated in the marketplace, and there were several attempts to reverse ADSL standards (from DMT to CAP/QAM) that were abortive. The supposed threat of fundamentally high complexity of DMT leading to high prices eventually was unequivocally refuted in the ADSL marketplace, where low-cost components abound today. Where VDSL was originally intended as an extension to the ADSL DMT standard [7], VDSL then instead became another chance for standardization of the QAM/CAP technologies in DSL. Because the American T1E1.4 group by-laws prohibited standardization in 1994 on line rates above 10 Mbps, that group's charter had to be rewritten for a new limit of 100 Mbps. As the charter was rewritten, CAP proponents insisted that the VDSL area be a new standard and that the line code issue be revised, even though the original VDSL was intended to extend ADSL speed and symmetry. Thus, the opportunity for yet another debate unfortunately emerged and continues to date.

This newer debate has continued in VDSL for many years with two large industrial consortia emerging with two complete transmission specifications for VDSL:

1. *VDSL Alliance* — discrete multitone transmission (DMT) with digital or analog duplexing
2. *VDSL Coalition* — "single"-carrier modulation (SCM) with analog duplexing

("Single" is in quotes here because the VDSL Coalition actually advocates a solution with two carriers in each direction, and an optional third carrier upstream.) Both groups have contributed a "temporary working" standard to T1E1.4, which appear in three documents: a common reference document [20], an SCM document [21], and a DMT document [22]. The DMT specification supports up to 4096 4.3125-kHz ADSL-style tones, and any number of up/down stream transmission bands. The lower 256 tones are exactly the same as ADSL, facilitating backward compatibility. Both groups have about 50 companies in them, with about 10 common members. The companies in both groups represent an enormous cross-section of the telecommunications world.

The VDSL Alliance solution is backward-interoperable with ADSL and has taken greater time to develop into its present converged state [22], but offers some outstanding flexibility and performance features in a large number of possible configurations. Some vendors sell chips that implement up to 32 ADSL modems or up to 4 VDSL modems. The VDSL Coalition specification [16] is slightly simpler to understand (although more pages in reality) and to design to, but not interoperable with ADSL. The Coalition specification had the advantage of earlier availability of transmission components that were partially compliant with it. International standardization in the ITU has steadfastly held to the principle that only one will become an international standard. The T1E1.4 group will revisit the two standards for permanency in 2003, when it is likely only one will survive. The IEEE 802.3 group on Ethernet in the First Mile (see Section 5) currently is also considering the two VDSL standards for that decision. At time of writing, it appears that the considerable advantages of backward interoperability with ADSL (now installed in over 20 million locations) and the spectral flexibility of the DMT approach will again cause it to prevail over the single-carrier method, and the number of single-carrier supporters dwindle to two or three companies. Clearly DMT is the best solution technically and from many other perspectives. However, the politics of standards groups can sometimes lead to tragic decisions, and this issue is not yet formally decided.

3. DMT PHYSICAL-LAYER STANDARD

Discrete multitone (DMT) transmission is the only worldwide standard for ADSL and VDSL transmission. DMT is the most efficient of the high-performance transmission methods that allow a transmission system to perform near the fundamental limit known as capacity [1]. For difficult transmission lines, there is no other cost-effective high-performance alternative presently, and VDSL has the most difficult transmission environment of all DSLs. Variants of DMT for wireless transmission (known as OFDM) have also come into strong use in the area of wireless local area networks (IEEE 802.11(a), [13]) and wireless broadband access (IEEE 802.16 [25]), as well as digital terrestrial television (HDTV and digital TV) broadcast [26] in most of the world, all of which are known to also be particularly difficult transmission problems. The standardized VDSL DMT method is a natural extension of the method used for ADSL and backward compatible with it, as described in Section 3.1. Section 3.2 describes the "zipper" duplexing method that is also called "digital duplexing," which is an enhancement to the original DMT method that allows the upstream and downstream transmissions to be compactly placed in the limited transmission bandwidth of a telephone line. Section 3.4 investigates initialization.

3.1. Basic Multicarrier Concept

Starr et al. explained the basic multicarrier transmission concept of dividing a transmission band into a large number of subcarriers and adaptively allocating fractions of total energy and data rate to each to match an individual line characteristic [1]. It is this adaptive loading feature that sets multicarrier methods in a higher league of performance than other DSL transmission methods. VDSL presents a highly variable transmission environment with bandwidths that can vary from a few MHz to nearly 20 MHz, with intervening radio interference in several narrow bands, with huge spectrum notching effects from

bridged taps, and with a variety of crosstalking situations. VDSL is undoubtedly the most difficult and highly variable transmission problem yet faced by DSL engineers. Any thing less than an excellent design will risk the viability of the DSL industry in the future, and multicarrier methods meet that challenge.

The first challenge for VDSL is interoperability with existing ADSL. One of the applications for VDSL is simply speed extension of ADSL, meaning that it is possible for an existing ADSL customer to have that service provided by a new ONU in his/her neighborhood that is VDSL-ready. A VDSL modem in the ONU that will interoperate with that existing ADSL modem is highly desirable so that no extra labor or purchases are necessary at the customer's premises should that customer elect to rest with their current ADSL service for a period of time before then electing to move to VDSL to increase the speeds of their ADSL service. Similarly, an existing ADSL customer may elect to purchase a VDSL modem (or may have one from a previous residence or business address) and then needs to interoperate with an existing ADSL CO modem. Thus, a requirement for incremental DSL rollout to higher speeds and increasing use of fiber is that VDSL interoperate with existing ADSL, meaning the lower 256-down/32-up DMT tones of an ADSL modem must also be implemented by an interoperating VDSL modem. For this reason, the DMT VDSL standard [22] uses the same tone spacing of 4.3125 kHz that was used in ADSL. The VDSL standard allows for the DMT VDSL modem to symmetrically use numbers of tones of 256, 512, 1024, 2048, and 4096 or 2^{n+8} $n = 0, 1, 2, 3, 4$. The number 2048 is considered a default for full compliance with other VDSL modems, but interoperation with smaller numbers of tones is illustrated in Fig. 13, where it is clear whether upstream or downstream, the modem with the smallest number of DMT tones then dictates the maximum that can be used in that direction. Such elimination of superfluous tones can occur naturally during training, or may be selectively programmed during special initialization exchanges; the latter is usually the preferred implementation.

The following terms apply here:

Extension. The cyclic extension [1] size for DMT VDSL is optionally programmable to sizes $m \times 2^{n+1}$, where m is an integer. The modem must be able to implement at least the default of $20 \times 2^{n+1} = 40 \times 2^n$ for the case of $n = 0$, which corresponds to interoperability with ADSL.[7] Longer cyclic prefixes on shorter channels allow some additional performance-enhancing features to be added for DMT VDSL that are not implemented in ADSL, which are described in more detail in Section 3.2.

Encoder. The constellation encoder for DMT VDSL and tone-ordering procedures are identical to those of the worldwide ADSL standard [1], although implemented perhaps over a larger set of tones for VDSL.

Pilot. The pilot of ADSL has been made optional and generalized in VDSL. In ADSL, the pilot was sent downstream always on tone number 64 (276 kHz). In VDSL, the VTU-R can decide to use a pilot on any (or no) tone in initialization. If a pilot is used, the 00 point in the standardized 4-point QAM constellation on that selected tone is sent in all symbols. The synchronization symbol of ADSL has been eliminated in VDSL, except when interoperating with an older ADSL modem.

Timing Advance. The VTU-R is capable of changing the symbol boundary of the downstream DMT symbol it receives by a programmable amount, which is communicated during initialization to the VTU-O modem. This is also a new feature for VDSL that is used for implementation of digital duplexing as described in Section 3.2.

Power Backoff (PBO). PBO has been studied by standards groups as a way to prevent upstream FEXT from a customer closer to an ONU from acting as a large noise for a VDSL customer further away. The basic problem is that the closer user could operate at a higher data rate or with better performance than is necessary or fair to other customers. Various methods for reducing the disparity among lines vary from introducing a flat power backoff at all frequencies as a function of measured received signal [21] to

[7] Actually, in this case, the cyclic prefix is reduced by 8 samples if the synchronization symbol of ADSL is to be inserted every 69 symbols or 17 ms.

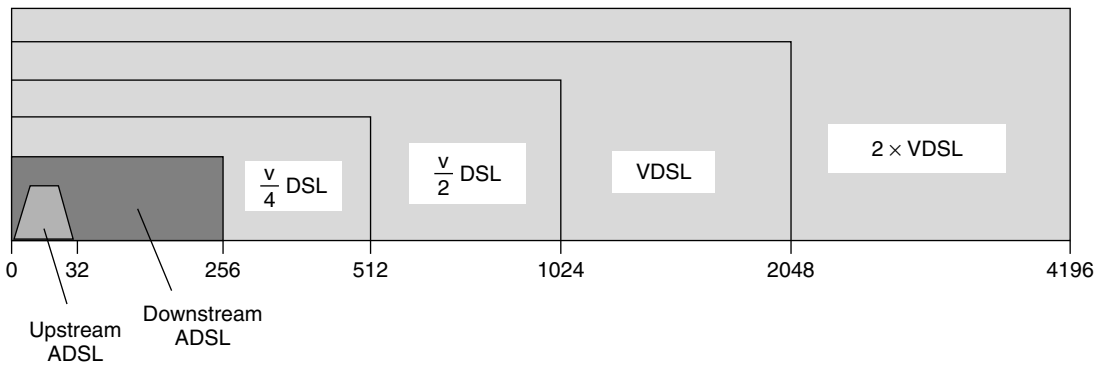


Figure 13. Interoperability diagram for standardized DMT modems of increasing speeds.

spectrally shaped methods that attempt to apply either the (1) *reference noise method* — forcing all upstream transmissions to have a FEXT of common spectral shape (and thus harm), or the (2) *reference length method* — forcing all upstream transmissions to have a FEXT the same as that of a nominal "reference" length VDSL line. The latter two methods, particularly (2), seem to have won favor with standards groups, but the area is still debated at time of writing. A method making use of actual line measurements [40] was introduced for future spectrum management and essentially eliminates the PBO issue, but came after VDSL standards had entered final voting and could not yet then be standardized.

Express Swapping. The standardized bit swapping of ADSL is mandatory also in VDSL [1]. However, VDSL also offers a highly robust and high-speed optional capability of instead altering the bit distribution all at once (instead of one tone at a time as in the older bit swapping). This is known as *express swapping*. The additional commands are described in the VDSL standards documents [22]. However, express swapping allows the new bit table is sent in one command and protected by CRC check — if correctly received, all tones are replaced with the new bit distribution at an immediately succeeding point specified in the commands and protocol. This allows a system to react very quickly to abrupt transients caused by excitation of crosstalkers or RF interferers, or perhaps an off-hook line change in splitterless operation. It also enables advanced spectrum management features that may occur in the future.

3.2. Digital Duplexing

This subsection describes *digital duplexing*, which is a method for minimizing bandwidth loss in separating downstream and upstream DMT VDSL transmissions. Originally, this method was introduced by Isaksson, Sjoberg, Nilsson, Mesdagh, and others in a series of papers that refer to the method as "zipper" [27,34–37]. General principles, as well as some simple examples, illustrate how digital duplexing works and why it saves precious bandwidth in VDSL. This description is intended for readers familiar with basic discrete multitone transmission (DMT). Thus, readers can use their DMT knowledge and the examples and explanations of this section to understand the relationship of the cyclic suffix to the cyclic prefix, and thus consequently to comprehend the benefits of *symbol-rate loop timing* and to appreciate the use of windows without intermodulation loss.

Excess bandwidth is a term used to quantify the additional dimensionality necessary to implement a practical transmission system. The excess-bandwidth concept is well understood in the theory of intersymbol interference where various transmit pulseshapes are indexed by their percentage excess bandwidth. In standardized and implemented DMT designs for ADSL, for instance, the symbol rate is 4000 Hz while the tone width is 4312.5 Hz, rendering the excess bandwidth $(0.3125/4) = 7.8\%$. In ADSL, additional bandwidth is lost in the transition band between upstream and downstream signals when these signals are frequency-division-multiplexed. In VDSL, this additional bandwidth loss is zeroed through an innovation [27] known here as *digital duplexing*, which particularly involves the use of a "cyclic suffix" in addition to the well-known "cyclic prefix" of standardized DMT ADSL. This subsection begins with a review of basic DMT and of its extension with the use of the cyclic suffix, including a numerical example that illustrates symbol-rate loop timing. This discussion illustrates why the time-domain overhead is all that is necessary to allow full use of the entire bandwidth without frequency guard bands in a very attractive and practical implementation.

This section proceeds to investigate crosstalk issues both when other VDSL lines are synchronized and not synchronized. Windowing and its use to mitigate crosstalk into other DSLs or G.pnt are also discussed, as are conversion device requirements.

This section then continues specifically to use a second example that compares a proposed analog duplexing plan for VDSL with the use of digital duplexing in a second proposal. In particular, 4.5 MHz of excess bandwidth is necessary in the analog duplexing while only the equivalent of 1.3 MHz is necessary with the more advanced digital duplexing. The difference in bandwidth loss of 3 MHz accounts for at least a 6–12-Mbps total data rate advantage for digital duplexing in the example, which provides a realistic illustration of the merit of digital duplexing.

3.2.1. Basic DMT. Figure 14 illustrates a basic DMT system for the case of baseband transmission in DSL [7], showing a transmitter, a receiver, and a channel with impulse response characterized by a phase delay Δ and a response length ν in sample periods. $N/2$ tones are modulated by QAM-like two-dimensional input symbols (with appropriate N-tone conjugate symmetry in frequency) so that an N-point inverse fast fourier transform (IFFT) produces a corresponding real baseband time-domain output signal of N real samples.

For basic DMT, the last $L = \nu$ of these samples are repeated at the beginning of the packet of transmitted samples so that $N + L$ samples are transmitted, leading to a time-domain loss of transmission time that is L/N. This ratio is the excess bandwidth. The minimum size for $L = \nu$ is the channel impulse response duration (in sampling periods) for basic DMT (later for digitally duplexed DMT, L will be the total length for all prefix/suffix extensions and thus greater than ν). Sometimes DMT systems use receiver equalizers [1] to reduce the channel impulse response length and thus decrease ν. Such equalizers are common in ADSL. The objective is to have small excess bandwidth by decreasing the ratio ν/N. In both DMT ADSL and VDSL, the excess bandwidth is 7.8%.

The IFFT of the DMT transmitter implements the equation

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n \cdot e^{j(2\pi/N)nk}$$

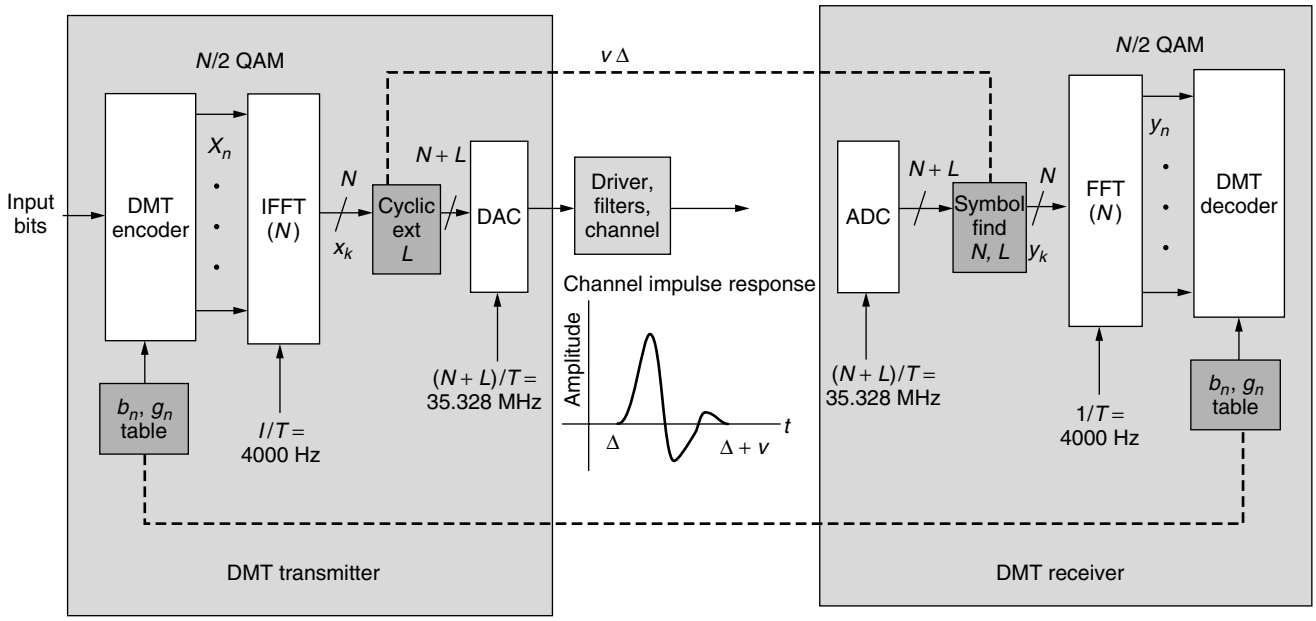


Figure 14. Basic DMT transmission system.

where x_k $k = 0, \ldots, N-1$ are the N successive time-domain transmitter outputs (the prefix is trivial repeat of last ν). The values X_n are the two-dimensional modulated inputs that are derived from standard QAM constellations (with the number of bits carried on each "tone" adaptively determined by loading [6] and stored in the b_n, g_n tables at both ends). To produce a real time-domain output, $X_n = X^*_{N-n}$ in the ubiquitous case that N is an even number. In ADSL, $N = 512$, while in VDSL, $N = 512 \cdot 2^{n+1}$ $n = 0, 1, 2, 3, 4$. One packet of $N + \nu$ samples is transmitted every T seconds for a sampling rate of $N + L/T$. With $1/T = 4000$ Hz in ADSL and VDSL, the sampling rates are 2.208 MHz and up to 35.328 MHz, respectively, leading to a cyclic prefix length of $\nu = L = 40$ samples in ADSL and a cyclic extension length of up to $L = 640$ samples in VDSL. The extension length in VDSL also includes the cyclic suffix to be later addressed.

When the cyclic extension length is equal to or greater than the impulse response length of the channel ($L \geq \nu$), DMT decomposes the transmission path into a maximum of $N/2$ independent simple transmission channels that are free of intersymbol interference and that can be easily decoded. The receiver in Fig. 14 extracts the last N of the $N + L$ samples in each packet at the receiver (when no cyclic suffix is used) and forms the FFT according to the formula

$$Y_n = \sum_{k=0}^{N-1} y_k \cdot e^{-j(2\pi/N)nk}$$

where the reindexing of time is tacit and really means samples corresponding to times $k = L+1, \ldots, L+N$ in the receiver. *The receiver must know the symbol alignment* and cannot execute the FFT at any arbitrary phase of the symbol clock. If the receiver were to be somehow offset in timing phase, then time-domain samples from another adjacent packet would enter the FFT input, displacing some corresponding time-domain samples from the current packet. For instance, suppose the FFT executed m sample times too late, then the output would be

$$\tilde{Y}_n = \sum_{k=0}^{N-m-1} y_{k+m} \cdot e^{-j(2\pi/N)nk} + \sum_{k=N-m}^{N-1} u_{N-m+k} \cdot e^{-j(2\pi/N)nk}$$

where u_k are samples from the next packet that are unwanted and act as a disturbance to this packet. Furthermore, m samples from the current packet were lost (and the rest offset in phase). Thus

$$\tilde{Y}_n = Y_n + E_n$$

where E_n is a distortion term that includes the combined effects of u_k, the missing terms y_k, and the timing offset in the packet boundary. $E_n = 0$ only (in general) when the correct symbol alignment is used by the receiver FFT. DMT systems easily ensure proper phase alignment through the insertion of various training and synchronization patterns that allow extraction of correct symbol boundary.

It is important to note that the FFT of any other signal with the same N might also have such distortion unless the symbol boundaries of that signal and the DMT signal were coincident. In the later case of time coincidence, the FFT output is simply the sum of the two signals' independent FFTs. Indeed the receiver would have no way of distinguishing the two signals and would simply see them as the sum in the time-coincident case.

The second signal could be the opposite-direction signal leaking through the imperfect hybrid in VDSL. If the transmitted and received symbols are aligned in time and frequency-division duplexing is used, tones are zeroed in one direction if used in the other. Then, the sum at the FFT output is simply either the upstream or the downstream signal, depending on the duplexing choice for the set of

indices n. No zeroed tones are necessary between upstream and downstream frequency bands as that is simply a waste of good undistorted DMT bandwidth. No analog filtering is necessary — the IFFT, cyclic prefix, and the FFT do all the work if the system is fortunate enough to have time coincidence of the two DMT signals traveling in opposite directions. The establishment of time coincidence of the symbols at both ends of the loop is the job of the cyclic suffix, which the next subsection addresses. In other words, the FFT works on any DMT signal of packet size N samples, regardless of source or direction as long as the packet is correctly positioned in time with respect to the FFT. This separation is not easily possible unless the DMT signals are aligned — thus VDSL DMT systems ensure this alignment through a cyclic suffix to be subsequently described.

Table 2 provides a comparison and summary of DMT use in ADSL and in VDSL. Note that DMT VDSL uses digital frequency-division multiplexing (FDM) and spans at most 16 times the bandwidth at its highest bandwidth use. This full bandwidth form is actually optional and the default values are shown in parentheses on the right, with the default actually being exactly half full.

3.2.2. Cyclic Suffix. The cyclic suffix occurs at the end of a DMT symbol (the opposite side of the cyclic prefix) and could repeat, for instance, the first 2Δ samples of the DMT symbol (not counting the prefix samples) at the end as in Fig. 16, where Δ is the phase delay in the channel (phase delay or absolute delay, not group delay, which is related to υ). A symbol is then of length $N+L$. The value for L must be sufficiently large that it is possible to align the DMT symbols, transmit and receive, at *both* ends of the transmission line. Clearly alignment at one end of a loop-timed line[8] is relatively easy in that the line terminal (LT) for instance need wait only until an upstream DMT symbol has been received before transmitting its downstream DMT symbols in alignment on subsequent boundaries. Such single-end alignment is often used by some designers to simplify various portions of an ADSL implementation. The alignment is not necessary unless digital duplexing is used. However, alignment at one end almost surely forces misalignment of symbol boundary at the VTU-R as in Fig. 15.

In Fig. 15, let us suppose that $N = 10$, $\nu = 2$, and that the channel phase delay or overall delay is $\Delta = 3$. The reader can pretend they have a master time clock and that the LT begins transmitting a prefixed DMT symbol at time sample 0 of that master clock. The first two samples at times 0 and 1 are the cyclic prefix samples, followed by 10 samples of the DMT symbol, that is, at times 0–11. At the receiver, all samples undergo an absolute delay of 3 samples (in addition to the dispersion of $\nu = 2$ samples about that average delay of 3). Thus, the cyclic prefix' first sample appears in the VTU-R at time 3 of the master clock and the DMT symbol then exists from time 3 to time 14. The samples used by the receiver for the FFT are samples 5–14, while samples 3 and 4 are discarded because they also contain remnants from a previous DMT symbol. The LT has DMT symbol alignment in Fig. 13, so that it also received the upstream prefix first sample at time 0 and continued to receive the corresponding samples of the upstream DMT symbol until time 11. Thus, the DMT symbols are aligned at the LT. However, in order for the upstream DMT symbol to arrive at this time, it had to begin at time -3 in the VTR-R. Thus the upstream symbol transmitted by the VTU-R occurs in the VTU-R at times -3 through 8 of the reader's master clock. Clearly, the DMT symbols are not aligned at the VTU-R.

In Fig. 16, a cyclic suffix of 6 samples is now appended to DMT symbols in both directions, making the total symbol length now 18 samples in duration (thus slowing the symbol rate or using excess bandwidth indirectly). The LT remains aligned in both directions and transmits the downstream DMT symbol from master clock times 0–17. These samples arrive at the NT at times 3–20 of the master clock, and valid times for the receiver FFT are now $5\text{–}14, 6\text{–}15, \ldots, 11\text{–}20$. Each of these windows of 10 successive receiver points carry the same information from the transmitter and differ at the FFT output by a trivial phase rotation on each tone that can easily be removed. The upstream symbol now transmits corresponding valid DMT symbols from time -1 to time 8, also 0–9, 1–10, and the last valid upstream symbol is time 5–14. At the VTU-R, the first downstream valid symbol boundary from samples 5–14 and the last upstream valid symbol boundary, also at samples 5–14, are coincident in time — thus the receiver's FFT can correctly find both LT and VTU-R transmit signals by executing at sample times 5–14 without distortion of or interference between tones.

[8] Loop timing of the sample clock means that the NT (VTU-R) uses the derived sample clock from the downstream signal as a source for the upstream sample clock in DMT.

Table 2. Comparison of DMT for ADSL and VDSL

	ADSL	VDSL (default)
FFT size	512	8192 (4096)
# of tones	256	4096 (2048)
Cyclic extension length	$L = \nu = 40$	$L = 640 \geq \nu + \Delta$
Sampling rate	2.208 MHz	35.328 (17.664) MHz
Bandwidth	1.104 MHz	17.664 (8.832) MHz
Duplexing	Analog FDM	Digital FDM
Tone width	4.3125 kHz	4.3125 kHz
Excess bandwidth	86 kHz (prefix) +40 kHz (filters)	1.3 MHz (650 kHz) (no filters)

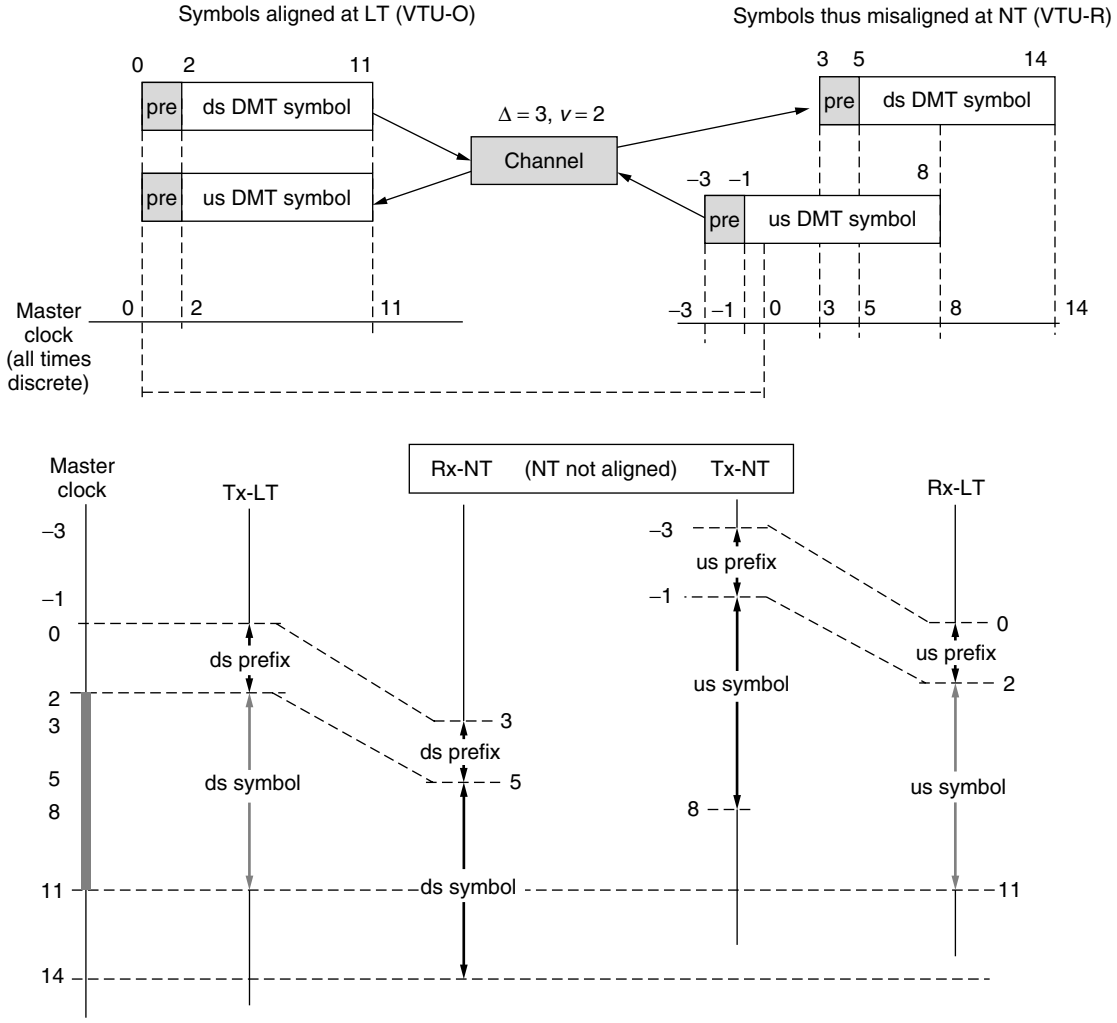


Figure 15. DMT symbol alignment without cyclic suffix.

At this time only (for cyclic suffix length 6), the receiver FFT and IFFT can be executed in perfect alignment, and thus downstream and upstream signals are perfectly separated. This DMT system in Fig. 16 has *symbol-rate loop timing*, and thus a single FFT can be used at each end to extract both upstream and downstream signals without distortion. (There is still an IFFT also present for the opposite-direction transmitter). This loop is now digitally duplexed.

Note that the cyclic suffix has length double the channel delay (or equivalently was equal to the round-trip delay) in the example. More generally, one can see that from Fig. 16 that VTU-R symbol alignment will occur when the equivalent of time 5, which is generally $\nu + \Delta$, is equal to the time $-3 + 2 + L_{\text{suf}}, = -\Delta + \nu + L_{\text{suf}}$. Equivalently $L_{\text{suf}} = 2\Delta$. In fact, any $L_{\text{suf}} \geq 2\Delta$ is sufficient with those cyclic suffix lengths that exceed 2Δ just allowing more valid choices for the FFT boundary in the VTU-R. For instance, the designer had chosen a cyclic suffix length of 7 in our example, then valid receiver FFT intervals would have been both 5–14 and 6–15. This condition can be halved using the timing advance method in the next subsection.

3.2.3. Timing Advance at LT. Figure 17 shows a method to reduce the required cyclic suffix length by one-half to $L_{\text{suf}} \geq \Delta$. This method uses a *timing advance* in the LT modem where downstream DMT symbols are advanced by Δ samples. The two symbols now align at both ends at time samples 2–11, and the cyclic suffix length is reduced to 3 in the specific example of Fig. 17. Thus, for VDSL with timing advance, the total length of channel impulse response length and phase delay must be less than 18 μs, which is easily achieved with significant extra samples for the suffix in practice. In VDSL, the proposal for L is 640 when $N = 8192$. The delays of even severe VDSL channels are almost always such that the phase delay Δ plus the channel impulse response length ν are much less than the 18-μs cyclic extension length (if not, then a time equalizer [6]). One notes that the use of the timing advance causes the transmitters and receivers at both ends to all be operational at the same phase in the absolute time measured by the master clock.

Digital duplexing thus achieves complete isolation of downstream and upstream transmission with no frequency guard band — there is however, the 7.8% cyclic extension penalty (which is equivalent to 1.3 MHz loss

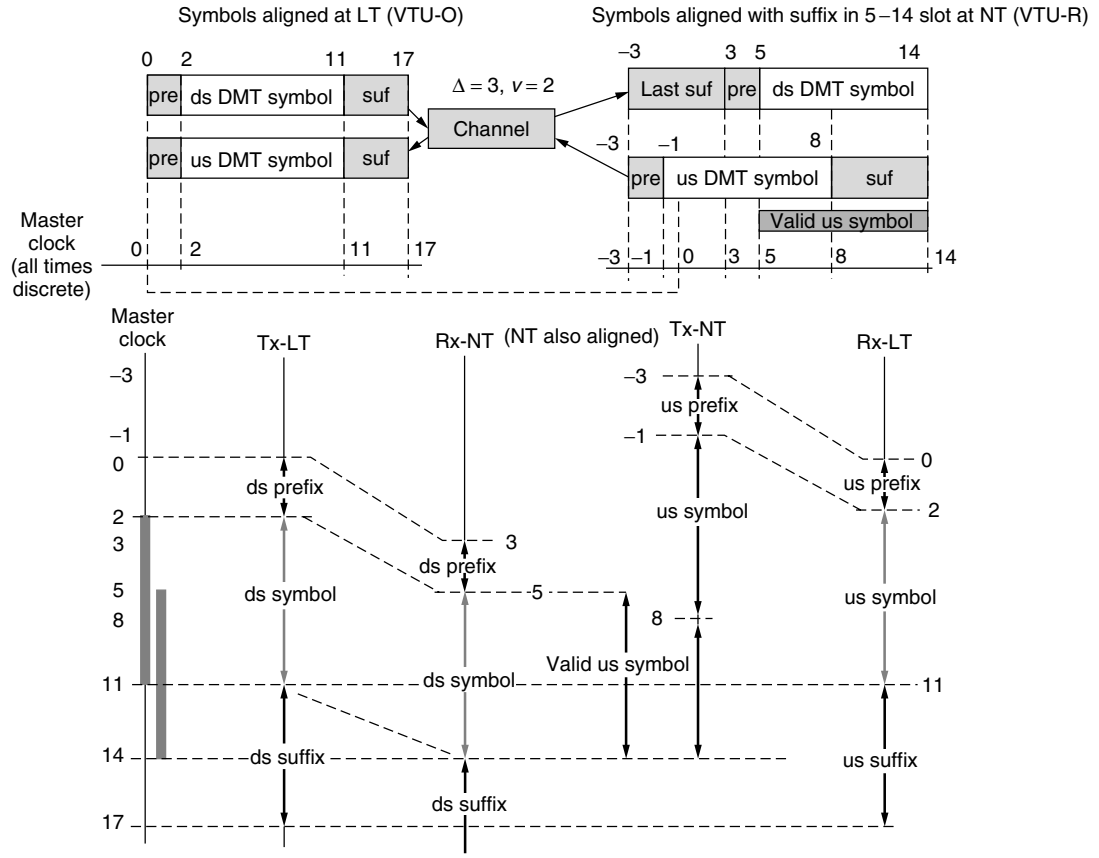


Figure 16. DMT symbol alignment at both LT and NT through use of suffix.

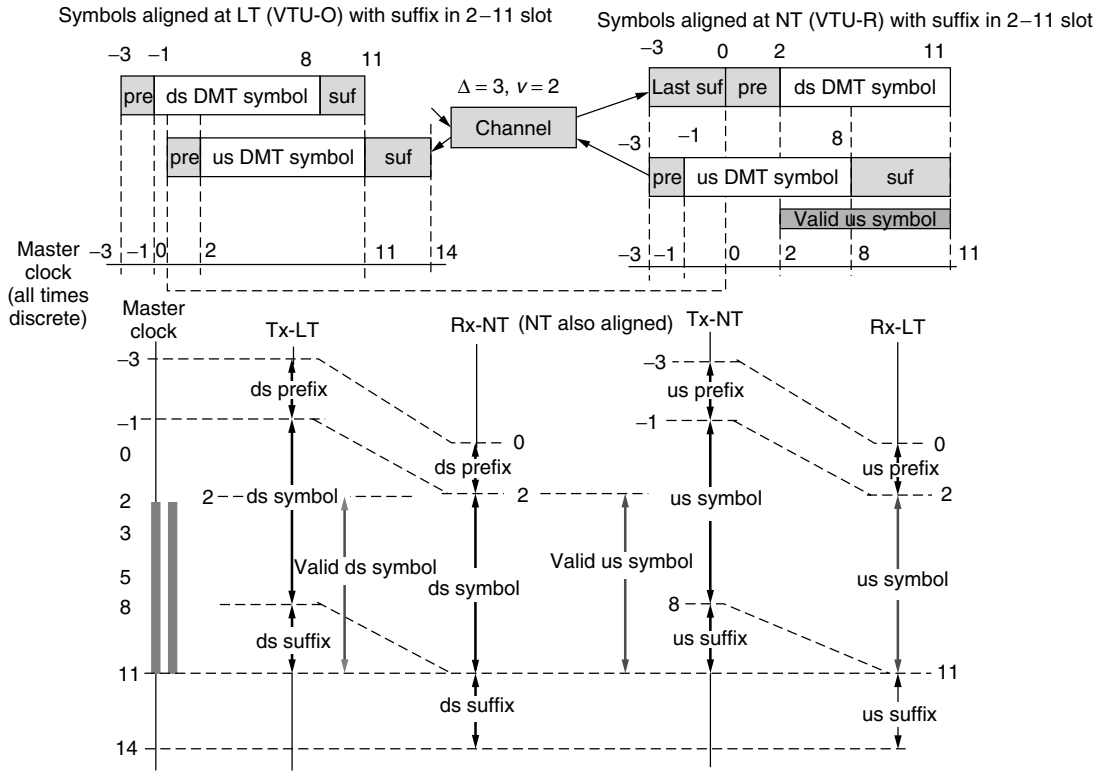


Figure 17. Illustration of cyclic suffix and LT timing advance.

in bandwidth in full VDSL and 650 kHz loss in the default or "lite" VDSL). Thus it is not correct, nor appropriate, to place frequency guard bands in studies of DMT performance in VDSL.

Digital duplexing in concept allows arbitrary assignment of upstream and downstream DMT tones, which with FDM VDSL means that these two sets of upstream and downstream tones are mutually exclusive. On the same line, there is no interference or analog filtering necessary to separate the signals, again because the cyclic suffix (for which the equivalent of 1.3 MHz of bandwidth has been paid or 650 kHz in default VDSL) allows full separation even between adjacent tones in opposite directions via the receiver's FFT. However, while theoretically optimum, one need not "zipper" the spectrum in extremely narrow bands, and instead upstream and downstream bands consisting of many tones may be assigned as described earlier. Some alternation between up and down frequencies is universally agreed as necessary for reasons of spectrum management and robustness, although groups differ on the number of such alternations.

3.2.4. Crosstalk. Analysis of NEXT between neighboring VDSL circuits needs to consider two possibilities:

1. Synchronization of VDSL lines
2. Asynchronous VDSL lines

The first case of synchronous crosstalk is trivial to analyze and implement with FDM. Adjacent lines have exactly the same sampling clock frequency (but not necessarily the same symbol boundaries). There is no NEXT from other synchronized VDSLs with FDM, as will be clear shortly. The second case of asynchronous VDSL NEXT (from other VDSL lines) into VDSL is more interesting. In this case, the sidelobes of the modulation pulseshapes for each tone are of interest.

A single DMT tone consists of the sinusoidal component

$$\begin{aligned} x_n(t) &= \left[X_n \cdot \exp\left(j\frac{2\pi}{N} \cdot \frac{N+\nu}{T} \cdot nt \right) \right. \\ &\quad \left. + X_{-n} \cdot \exp\left(-j\frac{2\pi}{N} \cdot \frac{N+\nu}{T} \cdot nt \right) \right] \cdot w_T(t) \\ &= 2|X_n| \cdot \cos(2\pi f_0 nt + \angle X_n) \cdot w_T(t) \end{aligned}$$

where $f_0 = 2\pi/N \cdot (N+\nu)/T$ or 4.3125 kHz in ADSL and VDSL, and $w_T(t)$ is a windowing function that is a rectangular window in ADSL, but is more sophisticated and exploits digital-duplexing's extra cyclic suffix and extra cyclic prefix in VDSL. Figure 18 shows the relative spectrum of a single tone with respect to an adjacent tone for the rectangular window.

Note the notches in the crosstalker's spectrum at the DMT frequencies, all integer multiples of f_0. Thus, the contribution of other VDSL NEXT will clearly be zero if all systems use the same clock for sampling, regardless of DMT symbol phase with respect to that clock. This is an inherent advantage of DMT systems with respect to themselves since it is entirely feasible that VDSL modems in the same ONU binder group could share the same clock and thus have no NEXT into one another at all. Indeed, this is a recommended option in [22].

When the sampling clocks are different, however, the more that sampling clocks of VDSL systems differ, the greater the deviation in frequency in Fig. 18 from the nulls, allowing for a possibility of some NEXT.

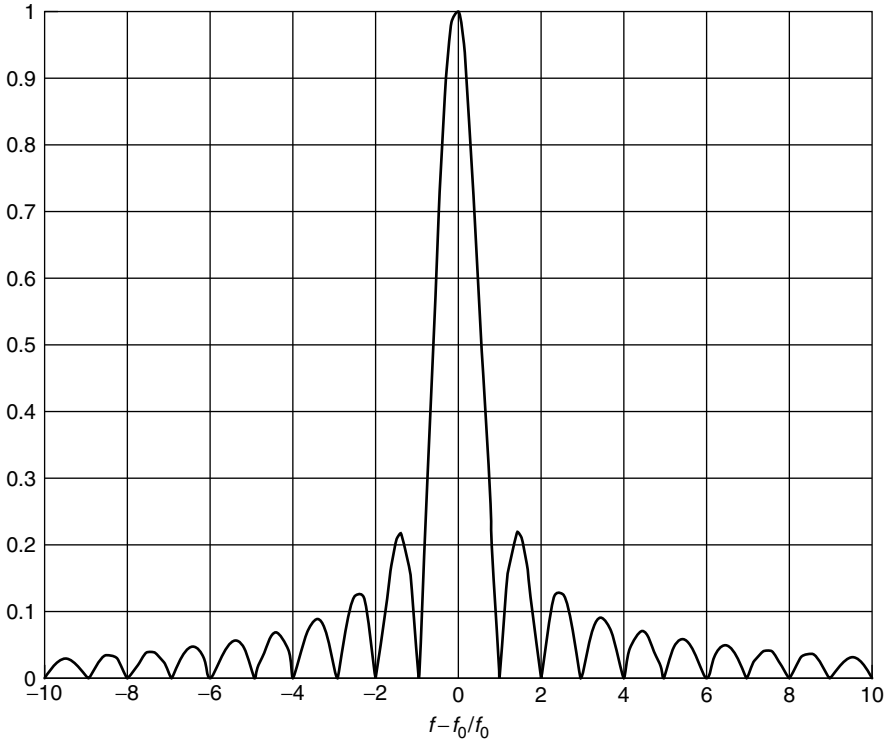


Figure 18. Magnitude of windowed sinusoid versus frequency; note notches at DMT frequencies.

Studies of such NEXT for DMT digital duplexing are highly subjective and depend on assumptions of clock accuracy, number of crosstalkers with worst-case clock deviation, and the individual contribution to NEXT transfer function of each of these corresponding worst-case crosstalkers. Nonetheless, reasonable implementation renders NEXT of little consequence between DMT systems.

If the VDSL PSD transmission level is $S = -60$ dBm/Hz, and the crosstalk coupling is approximated by $(m/49)^{.6} \cdot 10^{-13} \cdot f^{1.5}$ for m crosstalkers the crosstalk PSD level is

$$S_{\text{xtalk}}(f) = S \cdot (1/49)^{.6} \cdot f^{1.5} \cdot \left| \sin c \left(\frac{f}{f_0} \right) \right|^2$$

The peak or sidelobes can be only 12 dB down with such rectangular windowing of the DMT signal, as in Fig. 18. Asynchronous crosstalk may be such that especially with misaligned symbol boundaries that a really worst-case crosstalker could have its peak sidelobe aligned with the null of another tone (this is actually rare, but clearly represents a worst case). To confine this worst case, the methods of the next two subsections are used.

3.2.4.1. Windowing of Extra Suffix and Prefix. This section explains how windowing can be implemented without the consequence of intermodulation distortion when digital duplexing is used. Windowing in digitally duplexed DMT exploits the extra samples in the cyclic suffix and cyclic prefix beyond the minimum necessary. Since the cyclic extension is always fixed at $L = 640$ samples in VDSL, there are always many extra samples. Figure 19 shows the basic idea — the extra suffix samples are windowed as shown with the extra extension samples now being split between a suffix for the current block and a prefix for the next block. The two are smoothly connected by windowing, a simple operation of time-domain multiplication of each real sample by a real amplitude that is the window height. The smooth interconnection of the blocks allows for more rapid decay in the frequency domain of the PSD, which is good for crosstalk and other emissions purposes. A rectangular window will have the per tone (baseband) rolloff function given by

$$W_T(f) = \frac{\sin\left(\frac{\pi f}{f_0}\right)}{\left(\frac{\pi f}{f_0}\right)}$$

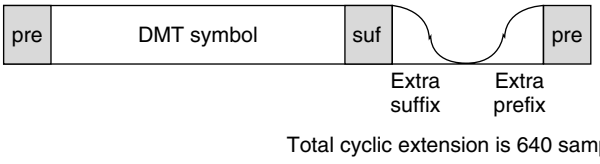


Figure 19. Illustration of windowing in extra suffix/prefix samples — smooth connection of blocks without affecting necessary properties for digital duplexing.

or the so-called sinc function in frequency. Clearly, a smoother window could produce a more rapid decay with frequency. A logical and good choice is the so-called raised-cosine window. Let us suppose that the extra cyclic suffix contains $2L' + 1$ samples in duration, an odd number.[9] Then, the raised cosine function has the following time-domain window (letting the sampling period be T'), where time zero is the first point in the extra cyclic suffix and the last sample being time $2L'T'$ and the centerpoint thus $L'T'$:

$$W_{T,rcr}(t) = \frac{1}{2}\left\{1 + \cos\left(\pi \cdot \left[\frac{t}{L'T'}\right]\right)\right\}$$
$$t = 0, \ldots, L'T', \ldots, 2L'T'$$

One notes that the window achieves values 1 at the boundaries and is zero on the middle sample and follows a sampled sinusoidal curve in between. The points before time $L'T'$ are part of the prefix of the current symbol, while the points after $L'T'$ are part of the suffix of the last symbol.

The overall window (which is fixed at 1 in between) has Fourier transform (let $\alpha = L'/L - L'$), ignoring an insignificant phase term

$$W_T(f) = \frac{L'}{\alpha} \cdot \frac{\sin\left(\frac{\pi f}{f_0}\right)}{\left(\frac{\pi f}{f_0}\right)} \cdot \frac{\cos\left(\frac{\alpha \pi f}{f_0}\right)}{1 - \left(\frac{2\alpha \pi f}{f_0}\right)^2}$$

Larger α means faster rolloff with frequency. This function is improved with respect to the sinc function, especially a few tones away from an up/down boundary. Reasonable values of α corresponding to 100–200 samples will lead to even the peaks of the NEXT sidelobes below −140 dBm/Hz at about 200 kHz spacing below 5 MHz. The reduction becomes particularly pronounced just a few tones away, and so at maximum, a very small loss may occur with asynchronous crosstalk. Thus, signals other than 4.3125-kHz DMT see more crosstalk, but within 200 kHz of a frequency edge, such NEXT is negligible. This observation is most important for studies of interference into home LAN signals like G.pnt, which at present almost certainly will not use 4.3125-kHz-spaced DMT.

3.2.4.1.1. Overlapped Transmitter Windows. Figure 20 shows *overlapped windows* in the suffix region. The smoothing function is still evident and some symmetrical windows (i.e., square-root cosine) have constant average power over the window and the effective length of

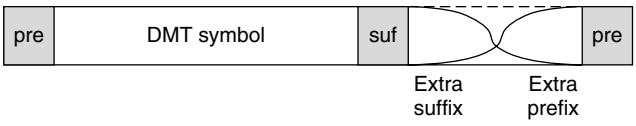


Figure 20. Illustration of overlapped windowing.

[9] If even, just pretend it is one less and allow for two samples to be valid duplexing endpoints.

the window above L' can be doubled, leading to better sidelobe reduction. This overlapping requires an additional $2L' + 1$ additions per symbol, a negligible increase in complexity.

3.2.4.1.2. Receiver Windowing. Receiver windowing can also be used to again filter the extra suffix and extra prefix region in the receiver, resulting in further reduction in sidelobes. Figure 21 shows the effect on VDSL NEXT for both the cases of a transmitter window and both a transmitter and receiver window. Note the combined windows has very low transmit spectrum (well below FEXT in a few tones) and below -140 dBm/Hz AWGN floor by 40 tones. If it is desirable to further reduce VDSL NEXT to zero, an additional small complexity can be introduced as in the next subsection with the adaptive NEXT canceler.

3.2.4.1.3. Adaptive NEXT Canceler for Digital Duplexing. Figure 22 illustrates an adaptive NEXT canceler and its operation near the boundary of up and down frequencies in a digitally duplexed system. Figure 22 is the downstream receiver, but a dual configuration exists for the upstream receiver. Note that any small residual upstream VDSL NEXT left after windowing in the downstream tone n (or in tones less than n in frequency index) must be a function of the upstream signal extracted at frequencies $n + 1, n + 2, \ldots$ at the FFT output. This function is a function of the frequency offsets between all the NEXTs and the VDSL signal. This timing clock offset is usually fixed, but can drift with time slowly. An adaptive filter can eliminate the NEXT as per standard noise cancellation methods [1]. A very small number of tones are required for the canceler per up/down edge if transmit windowing and receiver windowing are used. Adaptive noise cancellation can be used to make VDSL self-NEXT negligible with respect to the -140-dBm/Hz noise level. This allows full benefit of any FEXT reduction methods that may also be also in effect (note that the NEXT is already below the FEXT even without the NEXT canceler, but reducing it below the noise floor anticipates a VDSL system's potential ability to eliminate or dramatically reduce FEXT).

3.3. DMT VDSL Framing

The DMT transmission format supports Reed–Solomon forward error correction [1] and convolutional/triangular interleaving. The Reed–Solomon code is the same as that used for ADSL with up to 16 bytes of overhead allowed per codeword. There is no fixed relationship between symbol boundaries and codeword boundaries, unlike ADSL.

Instead any payload data rate that is an integer multiple of 64 kbps (implying an even integer multiple of payload bytes on average per symbol) can be implemented with dummy byte insertion where necessary and as described by Schelestrate [22]. Triangular interleaving that allows interleaving at a block length that is any integer submultiple of a codeword length (in bytes) is allowed (ADSL forced the block size of the interleaver to be equal to the codeword length). Given the high speeds of VDSL, the loss caused by dummy insertion is small, compared to the implementation advantage of decoupling symbol length from codeword length. Triangular interleaving was described by Starr et al. [1] and described again by Schelestrate [22].

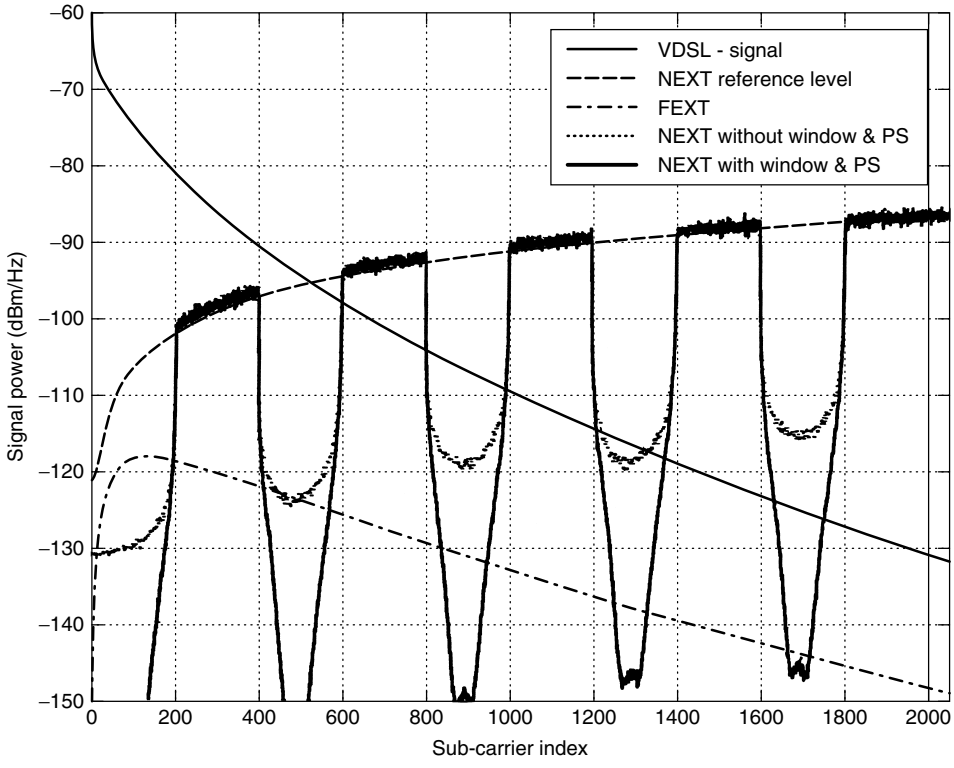


Figure 21. PS = transmitter windowing (pulseshaping) and "window" here means receiver window. This simulation is for a 1000-m loop of 0.5-mm transmission line (24-gauge).

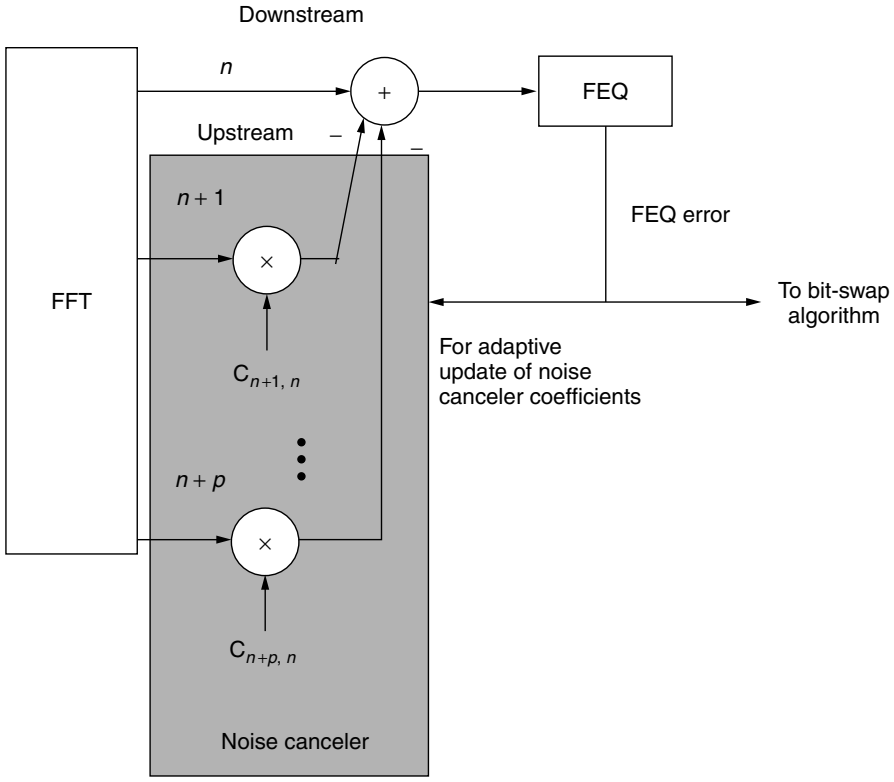


Figure 22. Adaptive noise canceler for elimination of VDSL self-NEXT in asynchronous VDSL operation. Shown for one upstream/downstream boundary tone (which can be replicated for each up/down transition tone that has NEXT distortion with asynchronous VDSL NEXT). No canceler is necessary if NEXT is synchronous.

Latency can take on any value between 1 ms (fast buffer requirement) and 10 ms (slow buffer default) or more. The latency is determined according to codeword, data rate, and interleave depth parameter choices as in [22]. Fast and slow data is combined according to a frame format that no longer includes the synchronization symbol of ADSL, and has updates of the fast and slow control bytes with respect to ADSL. Superframes are no longer restricted to just 69 symbols as in ADSL.

3.4. Initialization

The various aspects of training of a DMT modem [1]. The VDSL training procedure is described in Ref. 22 and compatible with the popular "g.handshake" (g.994) methods of the ITU. The fundamental steps of training are the same as in [1] with the LT now being expected to set a timing advance and measure round trip delay of signals so that the digital-duplexing becomes automatic. The length of cyclic prefix versus suffix and other detailed framing parameters are set through various initialization exchanges.

One feature of digital duplexing is that it does allow very simple echo cancellation if there is band overlap. With synchronized symbols, there is only one tap per tone to do full echo cancellation where that may be appropriate. However, the NEXT generated by overlapping bands at high frequencies might discourage one from trying unless NEXT cancellation (coordinated transmitters and receivers) can also be used, which would also only be one tap per tone per significant crosstalker. The reader is referred to Ref. 22 for more details.

4. MULTIPLE-QAM APPROACHES AND STANDARDS

The VDSL system specified by Oksman [21] uses either CAP or QAM as a modulation scheme [1] and frequency-division duplexing (FDD) to separate the upstream and downstream channels. There are two carriers or equivalently center frequencies, both with 20% excess bandwidth raised cosine transmission in each direction, following frequency plans 997 (Europe) or 998 (North America). The symbol rate of each of the signals is any integer multiple of 67.5 kHz, and the carrier/center frequencies can also be programmed as any integer multiple of 33.75 kHz. This allows for the receiver for each signal to estimate signal quality and request and appropriate center frequency and symbol rate, as well as corresponding signal constellation, which can be any integer QAM constellation from 4 points to 256 points as described in detail by Oksman [21]. Radio-frequency emission control occurs through programmable notch filters in the transmitter, for which a decision feedback equalizer in the receiver can partially compensate.

With overhead included, data rates are certain integer multiples (not all) of 64 kbps up to 51.84 Mbps downstream and 25.92 Mbps upstream.

4.1. Profiling in SCM VDSL

To accommodate short (<1 or 1000 ft), medium (1000–3000 ft), and long (>3000 ft) transmission at both asymmetric and symmetric rates, SCM VDSL can transmit up to 4 QAM signals, two in each direction. For long loops, only one carrier downstream and one carrier upstream (just above the downstream band) is

permitted. For medium range, a second downstream carrier is permitted in addition to the two carriers of long loops, while the short loops can use all four carriers. It was basically this 4-max carrier feature of SCM that forced the number of bands in the 997 and 998 frequency plans.

Figure 23 depicts the concepts of the profiles that are created by altering the center frequencies and symbol rates chosen. Notches at radio bands must be inserted to ensure emissions meet radio requirements, however there are not a sufficient number of carriers to simply achieve this notching by reversing direction. The 998 spectrum plan does leave one radio band as a reversal point near 4 MHz between upstream and downstream transmission. As Fig. 23 illustrates, larger symbol rates will likely be accompanied by larger center/carrier frequencies.

Decision-feedback equalization is presumed in the receiver and no Tomlinson precoding is used, even though FEC is used. The FECs in Ref. 21 and interleaving are basically identical to those used in the DMT standard. The DFE is described in the next subsection.

4.2. Operation of the DFE

The way a DFE handles RF interference (RFI) and notching is briefly discussed with reference to Fig. 24.

The analog front end (AFE) of any VDSL system needs to do some analog processing to reduce very strong RFI to an acceptable level and avoid overloading of the receiver's A/D and other input circuitry. This issue is not discussed any further here. The feedforward filter of the DFE creates a notch around the frequency at which the RF interferer is located, so that very little RFI is present at the output of this adaptive filter. The energy that is removed from the received signal by this notch is then restored by the feedback filter in such a way that the folded spectrum at the input of the slicer is flat. Actually, this is often claimed to be optimum performance by SCM advocates, but is not — optimum performance only can occur when the transmit band is silenced and more carriers are used to have a QAM signal on each side of the notch [25]. DFEs with multiple or single deep notches can require high precision implementation and must execute at the symbol rate, leading to billions of operations per second being required at VDSL speeds. Most QAM designers reduce the number of taps from the levels needed for excellent performance because of this complexity problem. Instead, designers hope that the difficult notching is not required often and then the small number of taps in the equalizer is sufficient.

In a way, SCM VDSL designers reduced system complexity in early chips by ignoring difficult channels and hoping that the low-complexity chips would be consequently attractive. QAM is suited well with the DFE on channels that have continuous transmission bands and mediocre distortion, where it can eliminate generation of multiple carriers. However, as the channel distortion grows, the complexity of the DFE quickly overtakes the complexity of the multiple carrier generation and QAM cannot handle channels with the severest distortion of VDSL well. SCM designers hope that an early low-cost solution can be replaced by increasing complex QAM components of the future that gradually address an increasing number of severe distortion situations in VDSL.

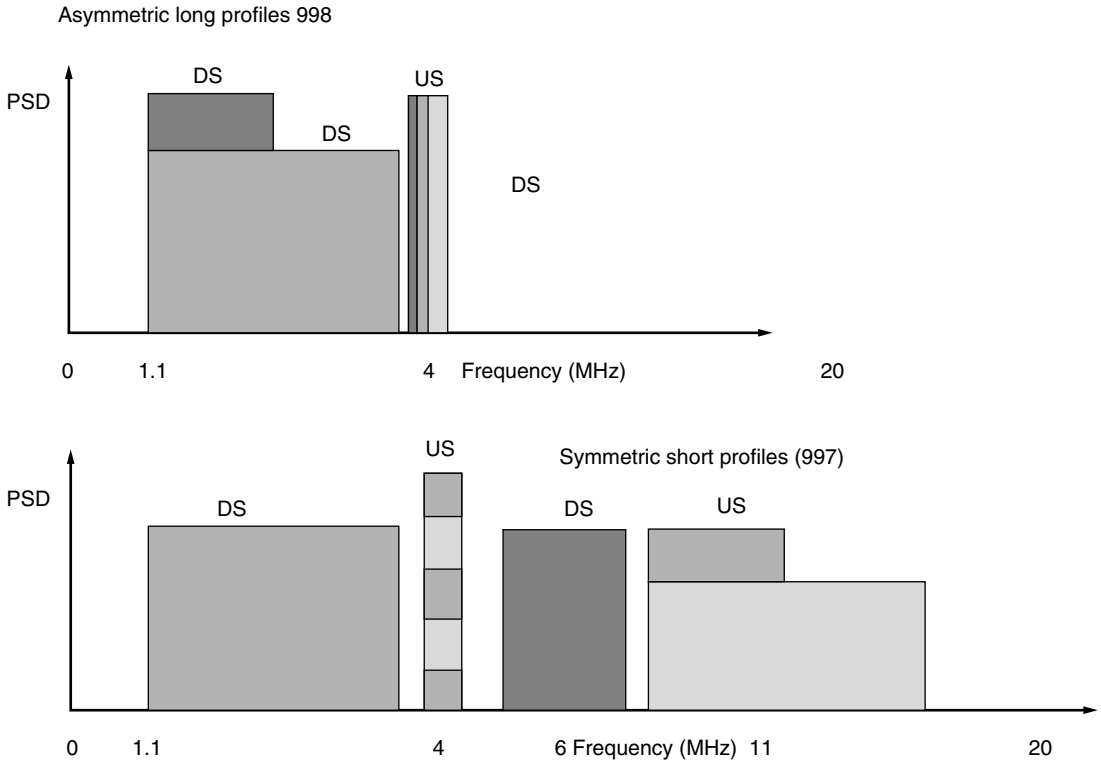


Figure 23. Basic concept of profiling in SCM VDSL.

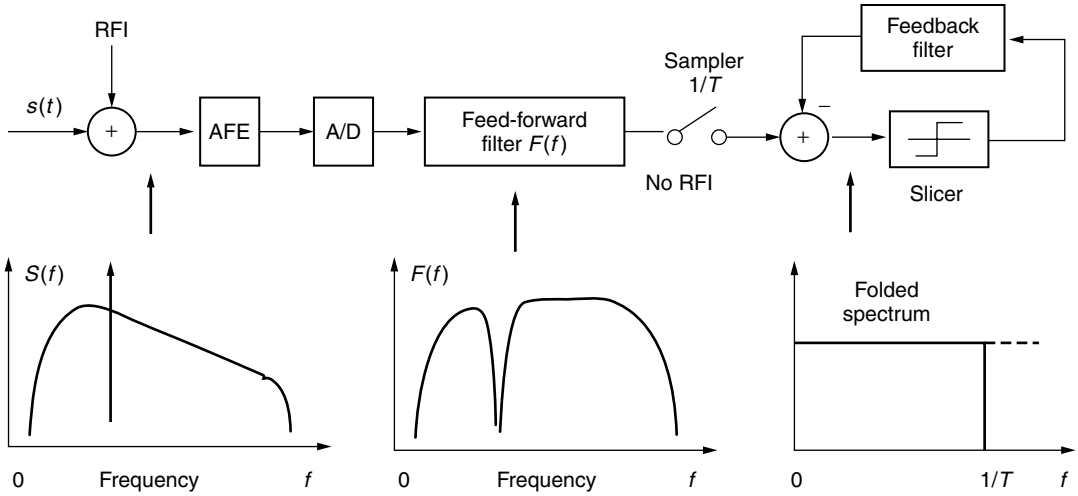


Figure 24. Principle of operation of the DFE in the presence of RFI.

5. ETHERNET IN THE FIRST MILE (EFM)

Figure 26 shows the basic concept of EFM [38]. One, two, or four phone lines may be coordinated to deliver 10 or 100 Mbps symmetric VDSL service in EFM. The 10 mbps version is also recently known as MDSL. While "Ethernet" physical-layer copper twisted-pair standards have long been established, they are restricted to a length of 100 m for 10-Mbps (10base-T), 100-Mbps (100base-T) and 1000-Mbps (1000base-T). Each uses a category 5 set of 4 twisted pairs in synchronous point-to-point transmission. Longer-length transmission fundamentally requires a different physical-layer modulation, while the upper layer "Ethernet/TCP/IP" functionality can be maintained so that the DSL line essentially looks like a "long-range ethernet." Since the current physical-layer Ethernet often uses 3 or 4 twisted pairs in coordination, admitting that same possibility for the EFM versions (clearly longer length for any given speed can be attained by sharing the transmission bandwidth of several lines to achieve the desired rate of 10 or 100 Mbps) as well as the possibility of carrying the entire data rate on just one line also (over a distance shorter than 2 or 4 coordinated lines).

Presently, the IEEE 802.3 standards committee is studying EFM possibilities, and has selected VDSL for the transmission format. EFM appears interested in only symmetric transmission where VDSL under plans 997 and 998 are clearly designed for asymmetric transmission. However, the flexible VDSL spectra under the DMT standard in Section 3 clearly does allow different band use for EFM where appropriate.

Some documents on EFM line modeling have appeared [38,39], but models are not fully accepted nor standardized presently for long-length lines. In particular, FEXT modeling at high frequencies becomes very important, especially with the use of multiple lines at wide coordinated bandwidths.

5.1. Multiline FEXT Modeling

The multiple-input/multiple-output (MIMO) characterization of a cable of twisted pairs merits attention and measurement for studies in Ethernet in First Mile (EFM) efforts. As noted, groups of twisted pair within a cable may be combined for better transmission/duplexing: The interaction between lines within a subgroup or the entire cable can be exploited to improve performance and reduce transceiver complexity, motivating a model. Reference 39 suggests a temporary model for MIMO FEXT that can be used to evaluate/test EFM.

5.1.1. The MIMO FEXT Channel. Figure 25 illustrates the matrix or MIMO FEXT twisted-pair channel. Each of the M inputs to this matrix channel may produce a component of the signal at each of the K outputs. Usually, $M = K$. For instance, a quad (4 twisted pairs tightly packed together) has $M = 4$ inputs, $K = 4$ outputs, and a total of $16 = M \cdot K$ transfer functions of interest. These $M \cdot K$ transfer functions can be summarized in a $K \times M$ matrix

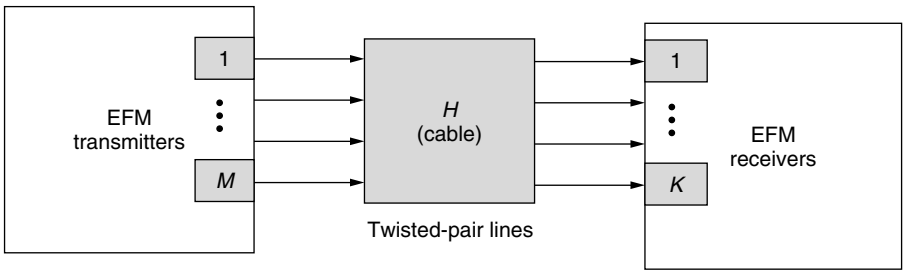


Figure 25. Matrix channel.

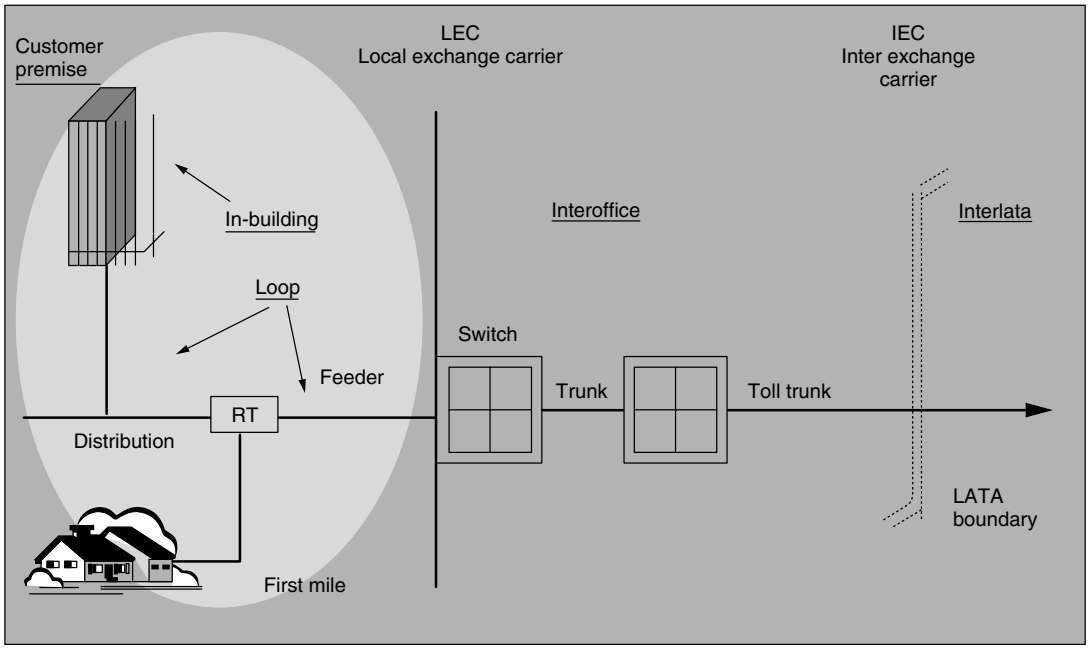


Figure 26. Ethernet in first mile illustration (shaded area is EFM interest) (courtesy of H. Barass).

H. The $K \times 1$ vector of channel outputs **Y** is then related to the $M \times 1$ vector of channel inputs **X** by $\mathbf{Y} = \mathbf{HX}$.

Ultimately, the designer would desire the exact **H** for each binder of wires: Any FEXT information is contained within this matrix. Approximate models are of interest in evaluating the various EFM opportunities in terms of range, rates, and service applications/market. In recognition that such transfer matrices are not well known, Reference 39 suggests an **H** model for temporary use in EFM studies. NEXT matrix models are of less MIMO interest since NEXT is either avoided by duplexing choice or by echo/NEXT cancellation between lines.

The kmth element of $\mathbf{H} = [H_{km}(f)]_{\substack{k=1,\ldots,K \\ m=1,\ldots,M}}$ is the transfer function from input m to output k. When $m = k$, $H(f)$ is simply the transfer function of the kth line, $H_{kk}(f)$, and can be determined from basic transmission line theory, given the length and RLCG parameters of the line [10]. Reference 10 also models FEXT power transfer of the off-diagonal terms as proportional to the line transfer function $|H_{kk}(f)|^2$, the square of frequency f^2, and the length of the line (in meters), d, which is explained on p. 90 of Ref. 11. This corresponds to a crosstalk–insertion loss transfer path of

$$H_{km}(f) = h_{\text{fext}} H_{kk}(f) \cdot (jf) \cdot \sqrt{d} \tag{1}$$

with a worst-case value of $h_{\text{fext}} = \sqrt{7.74 \times 10^{-21} \cdot (.3048 \text{ m/ft})} = 4.8 \times 10^{-11}$ for two adjacent category 3 phone company (telco) plant crosstalking lines. Equation (1) is for one crosstalking line — thus an extra multiplicative factor of $(K-1)_r^{0.6}$ used in models that average several lines is not used because that factor is for more than just two adjacent crosstalking lines. The factor h_{fext} is reduced nominally by a factor of 10 (or 20 dB) for category 5 wiring with tighter twisting. However, quads in the telephone plant that instead sometimes twist all 4 lines in an ensemble, may have a value higher than the one above, as much as an increase by a factor of 20 dB. Thus, a range of h_{fext} may be given by

$$\text{(Category 5 independent twists) } 4.8 \times 10^{-12} \le h_{\text{fext}} \le 4.8 \times 10^{-10} \text{ (category 3 ensemble–twisted quads).}$$

This same FEXT model is often seen in a form that describes only energy transfer, in other words, the squared magnitude of Eq. (1). Here, the model is converted to voltage transfer because EFM studies may desire phase information also. The equation above may sometimes be augmented by a linear phase term $e^{j2\pi f\tau}$, where τ is chosen to make the corresponding crosstalk impulse response causal. The matrix **H** can then be formed by finding the insertion loss function for any twisted pair in the bundle and inserting this insertion loss along the diagonal terms of the matrix **H**. The off-diagonal terms are equal to Eq. (1) with possibly randomly chosen phase offsets and/or linear phase. Some schemes that coordinate the lines may find the details of the off-diagonal terms increasing important.[10] In all cases, simple squaring of Eq. (1) and treating the FEXT like Gaussian noise (which is what current transceiver designs and implementations do) leads to the lowest possible (i.e., uncoordinated) performance.

However, actual individual FEXT insertion losses do not follow such a smooth characteristic with frequency

[10] When the off-diagonal terms are significantly smaller than the diagonal, the best coordinated schemes all converge to make the line appear as if there were no FEXT. While the details of the transfer function are then not of consequence to performance, the implementation still depends on knowing the phase.

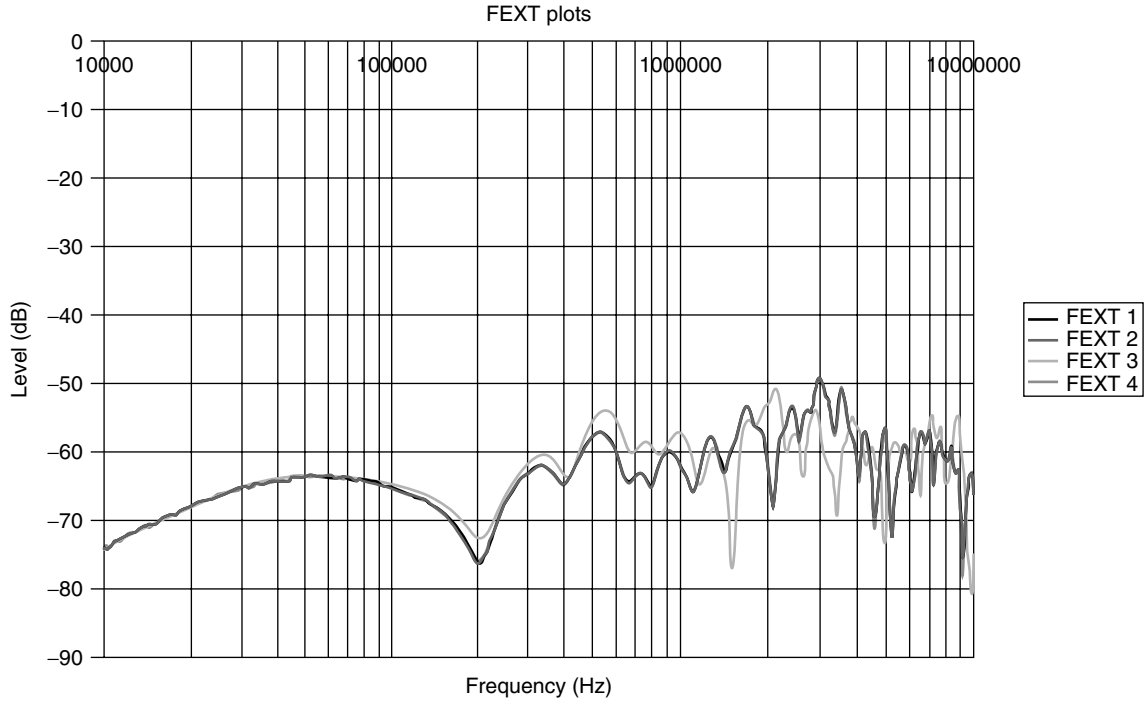


Figure 27. Illustration of 500-m example FEXT insertion loss functions (courtesy of John Cook of BTexact).

and indeed vary up and down with respect to the model in Eq. (1) (see Fig. 27). The variations can vary with the individual pairs as in Fig. 27.

The inaccuracy of the Eq. (*) model is increasingly evident at higher frequencies. There is a raised sinusoidal appearance to the magnitude transfer. This plot shows coupling functions between a pair on a 500 meter .5 mm cable with 50 pairs. Cosine terms can be added to the model to closely approximate the location of the dips and peaks in frequency seen in the measured FEXT insertion loss functions. Thus, the proposed model is

$$H_{km}(f) = \begin{cases} H(f) & m = k \\ n_{\text{lines}} \cdot \sqrt{h_{\text{fext}} d} \cdot (j2\pi f) & \\ \quad \cdot \left[1 + 0.3 \cdot \cos\left(\dfrac{2\pi f d}{c_{\text{line}}} \right) \right. & \\ \quad \left. -0.3 \cdot \cos\left(\dfrac{4\pi f d}{c_{\text{line}}} \right) \right] \cdot H(f) \cdot e^{j\varphi} & m \neq k \end{cases}$$

where $H(f)$ is as derived above from standard transmission theory. c_{line} is the speed of light on the media (often just use 300 Mm/s, and φ is a phase term [i.e., $\varphi = 2\pi f\tau + \phi_{km}$, where τ makes the response causal and ϕ_{km} is chosen from a uniform distribution over $(0, 2\pi)$ independently for each pair of indices m and k]. Normalizing each off-diagonal entry by the factor $n_{\text{lines}} = (K-1)^{.6/2}/\sqrt{(K-1)} = (K-1)^{-0.2}$ on average can account for the fact that distant lines have less crosstalk than close lines within the bundle and produces a slightly more accurate model when $K = 25$ or 50.

The following table suggests values for h_{fext} and n_{lines}:

	Category 5 Quad	Category 3	Telco Quads
h_{fext}	4.8×10^{-12}	4.8×10^{-11}	4.8×10^{-10}
n_{lines}	$3^{-0.2} = 0.803$	$49^{-0.2} = 0.459$	$3^{-0.2} = 0.803$

5.1.2. Summary. This proposed MIMO FEXT model attempts to augment well-known existing models to include

Sets of 4 lines — which may have better or worse coupling depending on the type of quad and associated independent (Cat 5) or ensemble (Cat 3 telco quad) twisting

Phase — phase coupling between lines that can be important for implementation of coordinated transmission schemes

Notches — the length-dependent frequency variation not included in earlier models, but that may become important in EFM studies

This model is in somewhat of a state of improvement at the time of writing and readers may want to pursue dynamic spectrum management and future EFM standards for any updates occurring after the time of writing. Reference 41 enumerates EFM data-rate vs range possibilities.

BIOGRAPHY

John M. Cioffi — BSEE, 1978, Illinois; PhDEE, 1984, Stanford; Bell Laboratories, 1978–1984; IBM Research, 1984–1986; EE Prof., Stanford, 1986–present. Cioffi founded Amati Com. Corp in 1991 (purchased by TI in

1997) and was officer/director from 1991–1997. Currently he is on the boards or advisory boards of BigBand Networks, Coppercom, GoDigital, Ikanos, Ionospan, IteX, Marvell, Kestrel, Teknovus, Charter Ventures, and Portview Ventures, and a member of the U.S. National Research Council's CSTB. Cioffi's specific interests are in the area of high-performance digital transmission. Various Awards: Member, National Academy of Engineering 2001; IEEE Kobayashi Medal (2001), IEEE Millennium Medal (2000), IEEE fellow (1996), IEE JJ Tomson Medal (2000), 1999 University of Illinois Outstanding Alumnus, 1991 *IEEE Comm. Mag.* best paper; 1995 ANSI T1 Outstanding Achievement Award; NSF Presidential Investigator (1987–1992). Cioffi has published over 200 papers and holds over 40 patents, most of which are widely licensed, including basic patents on DMT, VDSL, and vectored transmission.

BIBLIOGRAPHY

1. T. Starr, J. M. Cioffi, and P. Silverman, *Understanding DSL Technology*, Prentice-Hall, Upper Saddle River, NJ, 1999.
2. W. Chen, *DSL: Simulation Techniques and Standards Development for Digital Subscriber Lines*, Macmillan Technical Publishing, March 1998.
3. J. A. C. Bingham, *ADSL, VDSL, and Multicarrier Modulation*, Wiley, New York, 2001.
4. D. A. Rauschmayer, *ADSL/VDSL Principles: A Practical and Precise Study of Asymmetric Digital Subscriber Lines and Very High Speed Digital Subscriber Lines*, Macmillan, 1998.
5. C. K. Summers, *ADSL: Standards, Implementation, and Architecture*, CRC Press, Boca Raton, FL, June 1999.
6. P. S. Chow, J. S. Tu, and J. M. Cioffi, Performance evaluation of a multichannel transceiver for ADSL and VADSL services, *IEEE J. Select. Areas Commun.* **9**(6): 909–919 (Aug. 1991).
7. J. Cioffi and J. A. C. Bingham, A proposal for consideration of a VADSL standards project, *ANSI Contribution T1E1.4/94–183*, Dec. 5, 1994; see also J. Cioffi and K. Jacobsen, 15 Mbps on the Mid-CSA, *ANSI Contribution T1E1.4*/94–088, Palo Alto, CA, April 18, 1994, and Range/Rate Projections for NxDMT, *ANSI Contribution T1E1.4/94–125*, June 6, 1994.
8. K. Foster, A proposal to study the feasibility of high-speed copper drop standardization, *British Telecom Contributions TD 56 and WD 62 to ETSI RG12 subgroup of TM3*, Rome, Italy, Oct. 10, 1994.
9. ATM Forum, ATM User-Network Interface Specification, V.3.0, Physical Layer UNI Interfaces STS-1 for 51.84 Mbps 100 meter transmission on unshielded twisted pair; see *http://www.atmforum.com/atmforum/library/53bytes/backissues/others/53bytes-0795-5.html*, May 1995.
10. DAVIC (Digital Audio Visual Council), Specification 1.4 — Part 8 Lower-Layer Protocols and Physical Interfaces, *http://www.ccett.fr/dam/dvc spec.htm*, Short-range baseband asymmetrical PHY on copper and coax, ca., 1995.
11. T. Starr, M. Sorbara, J. M. Cioffi, and P. Silverman, *DSL Advances*, Prentice-Hall, Upper Saddle River, NJ, 2002.
12. J. Singh, (CEO, OnFiber), Presentation on future of fiber transmission and broadband access, to Computer Science and Telecommunications Board of United States National Research Council, Palo Alto, CA, Jan. 26, 2001.
13. IEEE Standard 802.11a-1999, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification: High-speed Physical Layer in the 5 GHz Band*, issued by IEEE Computer Society, Sept. 16, 1999.
14. International Telecommunications Draft Standard G.pnt, Study Group SG15/Q4, 2001.
15. American National Standard T1.417, *Spectrum Management*, 2001.
16. T. Riley, Proposal to add a new work program to the VDSL project, T1E1.4 *Contribution 99-*, Clearwater, FL, Dec. 5, 1999.
17. J. Cioffi, G. Ginis, and W. Yu, Performance of splitterless DSL at higher speeds, *ANSI Contribution T1E1.4/99-558*, Clearwater, FL, Dec. 5, 1999.
18. ETSI TS 101 270-2 v1.1.5, (2000-12) TM, VDSL Part 2 Transceiver Specification, European Telecommunications Standards Institute Report on VDSL, TM (transmission and multiplexing), *Access Transmission Systems on Metallic Access Cables*, ETSI, Sophia Antipolis, 2000.
19. J. Cioffi and W. Yu, G.vdsl: Robust VDSL in the presence of bridged taps, *ITU Contribution NG-039*, Nashville, TN, Nov. 2, 1999.
20. Q. Wang, ed., Very-high-speed digital subscriber line (VDSL) metallic interface, Part 1: Functional requirements and common specification, *ANSI Temporary Standard, T1.424-2002.*
21. V. Oksman, ed., Very-high-speed digital subscriber line (VDSL) metallic interface, Part 2: Single-carrier modulation specification, *ANSI Temporary Standard, T1.424-2002.*
22. S. Schelestrate, ed., Very-high-speed digital subscriber line (VDSL) metallic interface, Part 3: Multicarrier modulation (MCM) specification, *ANSI Temporary Standard, T1.424-2002.*
23. J. Cioffi et al., North American Prioritization of high-speed asymmetric VDSL service with respect to other VDSL services in response to G.vdsl issue item 2.7, *ANSI Contribution T1E1.4/99-393*, Baltimore, MD, Aug. 23, 1999.
24. FCC Order 00-336, *Second Memorandum Opinion and Order*, adopted Sept. 7, 2000; released Sept. 8, 2000, Ameritech and SBC Corporations.
25. V. Erceg et al., Channels or fixed wireless applications, *IEEE 802.16 Standards Contribution*, Xc-00/NNr0, Feb. 23, 2001, Tampa, FL; see also The IEEE 802.16 Working Group on Broadband Wireless Access Standards, fixed-wireless access standard and activities, Webpage *http://wirelessman.org/*, Jan. 16, 2001.
26. ETSI EN 300 429 V1.1.2 (1997-08), *Digital Video Broadcasting* (DVB)*, Framing structure, Channel Coding, and Modulation for Cable Systems*.
27. M. Isaksson et al., Zipper — a flexible duplex method for VDSL, *Proc. Int. Conf. Copper Wire Access Systems* (CWAS '97), Budapest, Hungary, Oct. 1997, pp. 95–99; see also M. Isaksson et al., Zipper — a duplex scheme for VDSL based on DMT, *ANSI T1E1.4/97-016*, Feb. 3–7, 1997.
28. K. Cheong et al., Soft cancellation via iterative decoding to mitigate the effect of home-LANs on VDSL, *ANSI Contribution T1E1.4/99- 333R2*, Baltimore, MD, Aug. 23, 1999.
29. J. Cioffi, G. Ginis, W. Yu, and C. Zeng, Example improvements of dynamic spectrum management, *ANSI Contribution T1E1.4/2001*-089R1, Costa Mesa, CA, Feb. 23, 2001.

30. J. Cioffi et al., Construction of modulated signals from filter bank elements and equivalence of line codes, *ANSI Contribution T1E1.4/99-395*, Baltimore, MD, Aug. 23, 1999.

31. G. Cherubini, E. Eleftheriou, S. Oelcer, and J. Cioffi, Filtering elements to meet requirements on power spectral density, *ANSI Contribution T1E1.4/99-429*, Baltimore, MD, Aug. 23, 1999.

32. J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney, Jr., MMSE decision feedback equalizers and coding: Parts I and II, *IEEE Trans. Commun.* **43**(10): 2582–2604 (Oct. 1995).

33. J. M. Cioffi and G. D. Forney, Jr., Generalized decision-feedback equalization for packet transmission with ISI and Gaussian noise, in A. Paulraj, V. Roychowdhury, and C. Schaper, eds., *Communication, Computation, Control, and Signal Processing*, Kluwer, Boston, 1997 (a tribute to Thomas Kailath).

34. F. Sjöberg et al., Zipper — a duplex method for VDSL based on DMT, *IEEE Trans. Commun.* **47**(8): 1245–1252 (Aug. 1999).

35. D. J. G. Mestdagh, M. Isaksson, and P. Ödling, Zipper VDSL: A solution for robust duplex communication over telephone lines, *IEEE Commun. Mag.* **38**(5): 90–96 (May 2000).

36. F. Sjöberg et al., Asynchronous zipper, *Proc. IEEE Int. Confer. Communications* (ICC'99), Vancouver, Canada, June 1999, Vol. 1, pp. 231–235.

37. F. Sjöberg et al., Performance evaluation of the zipper duplex method, *Proc. IEEE Int. Confe. Communications* (ICC'98), Atlanta, GA, June 1998, pp. 1035–1039.

38. IEEE 802.3 Draft, Ethernet in the First Mile Copper Standard Test and Data Rate Project, *IEEE 802.3 Standards Contribution*, Los Angeles, Oct. 18, 2001.

39. J. Cioffi and J. Fang, A Temporary model for EFM/MIMO cable characterization, *IEEE 802.3 Standards Contribution*, Los Angeles, Oct. 18, 2001.

40. ANSI Draft Dynamic Spectrum Management Report. *ANSI T1E1.4/2002-R2*, August, 2002, Westminster, Co.

41. J. Cioffi, G. Ginis, and K. B. Song, "Coordinated Level 2 DSM Results: Vectoring of multiple DSLS," *ANSI Contribution TlE1.4/2002-059*, Vancouver, BC, Feb. 18 2002.

42. J. Cioffi, J. Lee, and S. T. Chung, "10 MDSL Beyond all Goals, and Spectrally Compatible with ADSL and VDSL, From Co or RT," *ANSI Contribution TlE1.4/2002-129*, Atlanta, GA, April 8, 2002.

VIRTUAL PRIVATE NETWORKS

Marc-Alain Steinemann
Torsten Braun
Marc Danzeisen
Manuel Günter
University of Bern
Bern, Switzerland

1. INTRODUCTION

A *virtual private network* (VPN) is a private network constructed by public lines or connections using secure methods to transfer information. For example, VPN technology allows organizations to securely extend their network services across shared public networks such as the Internet to remote users, branch offices, and partner companies.

Large corporations used to interconnect local headquarters and branch offices with leased connections provided by telecommunication companies and ran private networks, so-called corporate networks. With the rise of the Internet technology more and more corporate networks switched from various networking protocols such as Novell to the TCP/IP protocol suite. Such private networks based on Internet technology are also referred to as *intranets*. Since leased lines are expensive and the corporations often already have Internet connectivity, there is an economic incentive to replace the expensive leased connections and to use the wide-area interconnectivity of the global Internet instead. However, two basic problems must be emphasized:

1. The intranet may use private addresses that are not unique in the global Internet and thus not routable [1].
2. The Internet protocol version 4 does not assure transmission privacy. While IP packets travel through the public Internet they may be viewed or even altered by third parties.

2. DIFFERENT TYPES OF VPNs

2.1. Subnet-to-Subnet and Access VPNs

Virtual private networks [2–4] encapsulate the packets with private addresses into packets with public addresses. This process is called *tunneling*. If privacy and authenticity of the encapsulated packets are desired, these can be ensured with cryptographic means.

Figure 1 shows the two most prominent VPN types: subnet-to-subnet VPNs and access VPNs. The subnet-to-subnet VPN interconnects geographically distributed

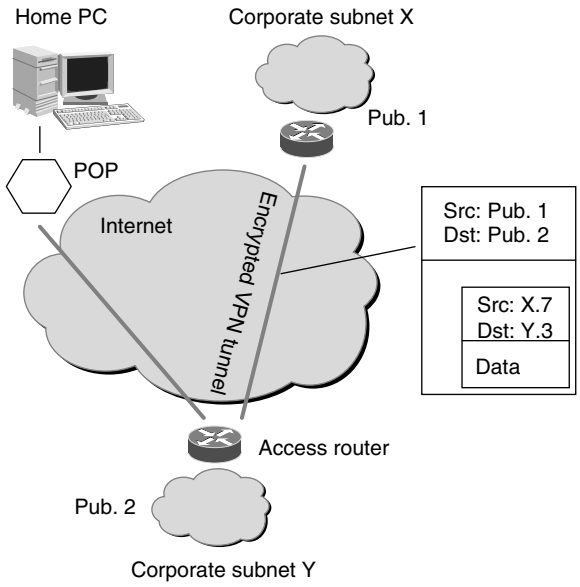


Figure 1. Virtual private network types.

private IP subnets. All traffic leaving one subnet destined for another one is tunneled through the public Internet. The access VPN allows roaming users to dial into the virtual network from their home computers or via an arbitrary Internet point of presence (POP).

Figure 1 also illustrates the tunneling mechanism. It shows the structure of a tunneled IP packet originating from an application that runs within the private subnet X. The packet's destination is a computer in a remotely located part of the VPN (the private subnet Y). The subnets X and Y use private IP addresses that cannot be routed in the public Internet. The address structure of the VPN is invisible from the outside. The access routers of subnets X and Y incorporate VPN functionality. They have an interior network interface with a private IP address and an exterior network interface with a public IP address. The access router at X recognizes that the packet in question must be tunneled. It knows the public interface of the access router of subnet Y and uses that address as the destination address and its own public address as the source address. The access router (also referred to as the *tunnel endpoint*) creates a new IP packet with these new addresses and puts the original packet into the payload of the new packet. The payload is then encrypted. The new packet is sent to the tunnel endpoint at Y. There, the router extracts the payload of the packet and decrypts the content. In this manner, the original packet is restored and can be routed on the private subnet Y toward the originally intended destination.

The access VPN case also uses tunnels. However, there are two distinct possibilities. Either the home PC acts as a tunnel endpoint or the POP of an Internet service provider (ISP) acts as tunnel endpoint.

While a VPN may be useful for a small-to-medium-sized company, the management of the VPN would require additional equipment and personnel. As a consequence, there exists a market for VPN services that lets the customers outsource the management of their VPN. The ISP can deploy VPN capable border routers and use them to introduce a VPN on-demand service [5]. Thereby, several VPNs can be managed on the same infrastructure by the same personnel (ISP staff) so that both the customer and the provider can profit from the economy of scale.

2.2. Encapsulation

Today, many different types of VPN technologies exist such as layer 2 VPNs based on frame relay and asynchronous transfer mode (ATM) networks; remote-access VPNs such as PPTP and L2TP; and IPSec-based VPNs.

2.2.1. Link-Layer VPNs (Layer 2). The Integrated Services Digital Network (ISDN), frame relay and asynchronous transfer mode are connection-oriented networks on link level (layer 2) that support the establishment of link-layer VPNs. Nowadays, most link-layer VPNs are established by frame relay and ATM technology. IP network links over these underlying connection-oriented network technologies are based on overlay models. In this case, meshes of connections have been established to interconnect IP routers of particular VPNs by providing a tunneling infrastructure.

Other but similar types of virtual networks based on link-level mechanisms are virtual local-area networks (VLANs) that can be established using IEEE 802.1Q, ATM LAN emulation (LANE), or multiprotocol over ATM (MPOA) technologies.

A major disadvantage of layer 2 VPNs and also VLANs is the need for a homogeneous topology throughout the entire VPN and the complexity involved in managing two different network technologies, namely, IP and the underlying network technology, for a single VPN. An advantage lies in the connection-oriented structure of those technologies. Links stay established and the tunneled packets follow the link and don't need to be routed as in IP-based VPNs. In addition, quality of service (QoS) is often provided implicitly by the connection-oriented network technologies.

2.2.2. Network-Layer VPNs (Layer 3). In contrast to the link-layer VPNs, where the location-independent IP provides layer 3 addresses and the location-dependent addresses are provided by layer 2 technology, in network layer VPNs, IP provides location-independent as well as the location-dependent addressing. For example, in a link-layer VPN, the location-independent IP addresses can be chosen by the user and the fixed medium-access channel (MAC) addresses are delivered by the network interface. In a network-layer VPN, the location-dependent IP addresses are provided by the intranet and the location-independent IP addresses are provided by the VPN. VPNs based on tunneling mechanisms that use network-layer protocols such as IP or MPLS as outer header are called *network-layer VPNs*.

Tunneling (also called *packet encapsulation*) is a method of wrapping a packet into a new one by prepending a new header. The whole original packet becomes the payload of the new one. At the tunnel endpoints (usually border routers) the header is added (respectively removed) and the result is then forwarded again. Tunneling is often used to transparently transport packets of one network protocol through a network running another protocol.

IP VPN tunneling mechanisms often encapsulate IP packets into IP packets. This tunneling method is called *IP in IP* encapsulation (IPIP). With IPIP encapsulation encryption can be applied to the inner packet by using IPSec protocols.

Generic routing encapsulation (GRE) is another popular tunneling method. GRE is a multiprotocol carrier protocol. With GRE a router at each VPN site encapsulates protocol-specific packets in an IP header, creating a virtual point-to-point link to routers at other ends of an IP cloud, where the IP header is stripped off. By connecting multiprotocol subnetworks in a single-protocol backbone environment, IP tunneling allows network expansion across a single-protocol backbone environment. GRE tunnels do not provide true confidentiality (no encryption functionality) but can carry encrypted traffic. It is possible to encapsulate almost every existing network protocol in GRE.

Standard protocols such as the Point to Point Tunneling Protocol (PPTP) and Layer 2 Forwarding (L2F) are required for supporting remote VPN access by single end systems. The protocols establish virtual point-to-point

links between an end system and a VPN server. The VPN server acts as an interface of a VPN for remote end systems. The protocols mentioned above can carry any other network protocol and are themselves encapsulated in IP. PPTP and L2F have been developed further resulting in a standard called Layer 2 Tunneling Protocol (L2TP).

With multiprotocol label switching (MPLS) routing is independent from the destination address in the encapsulated packet. This independence from the routing decision and the destination address is obtained by establishing a label-switched path (LSP) instead of establishing an IP tunnel between the two routers of a common VPN. MPLS allows setting up tunnels by appending a MPLS header in front of the IP header. This 32-bit MPLS header avoids the large overhead by another IP header as it is required with IP-in-IP tunneling. Multiple MPLS headers are possible; thus labels can be stacked onto each other. Label stacking supports hierarchical tunnels and is in particular being used when building MPLS-based VPNs.

In a typical MPLS VPN scenario as shown in Fig. 2, a packet is classified at an ingress router of an ISP based on the incoming port number as belonging to a particular VPN. The ingress router has learned via Boarder Gateway Protocol (BGP) to which VPN it belongs, to which egress router the packet must be sent, and via which egress interface the destination is reachable. The ingress router appends two labels to a packet belonging to a VPN; the inner label specifies the egress port at the ISPs egress router, namely, the link toward the destination subnetwork of the VPN. The outer label is being used to forward the packet toward the egress router and can be learned by MPLS signaling protocols such as Constraint-based Routing (CR) using Label Distribution Protocol CR-LDP or Traffic Engineering Resource Reservation Setup Protocol (TE-RSVP). Both labels are popped by the egress router (edge router). Figure 2 shows an example VPN/MPLS scenario with a label-switched path (LSP) set up between ingress and egress of an ISP. This LSP is set up along the path and carries the traffic between the VPN subnets. Note that MPLS makes the private VPN addresses of a customer transparent to the routers of the ISP and that MPLS does not provide security mechanisms as IPSec does.

2.2.3. Firewalls and VPNs. VPN tunnels are initiated and terminated mainly by specially equipped routers equipped with the respective hardware and software for establishing VPNs. If the organization at the endpoint of a tunnel needs additional security, the router can be replaced by a firewall router. It is also possible to establish VPNs through firewalls, that is, to tunnel a VPN link through a firewall. In the case of opening a firewall for a VPN tunnel, the instance allowing access to systems behind their firewall has to make sure that the other side deploys at least the same security policy level. If a host establishes a single unprotected connection to the Internet, and is at the same time connected through a VPN to computers behind a firewall, hackers can break in quite easily.

3. SECURITY AND THE INTERNET PROTOCOL

There exists a wide spectrum of technologies securing Internet communication, but most of them are dedicated to specific software applications. In that case, security is provided by the application layer. Good examples are Pretty Good Privacy (PGP) for mail encryption and browser-based authentication as well as Secure Sockets Layer (SSL) for traffic encryption between Web browser and Web server. These restrictions are not consistent with the requests of a large enterprise, and the average ISP that may never know precisely the kind of applications running tomorrow over today's networks.

3.1. Possible Threats on the Internet

VPNs are driven by security threats in the network environment and must fulfill three fundamental requirements:

- *Authentication* — the communicating persons must really be the persons they claim to be.
- *Confidentiality and privacy* — no one shall be able to electronically eavesdrop traffic.
- *Integrity* — the received traffic must not be altered in any way during transmission.

3.1.1. Spoofing. In IP networks it is difficult to know where information really originates. An attack called IP "spoofing" takes advantage of this weakness. Since the source IP address of a packet has no influence on routing, it can easily be forged. In this type of attack, a packet coming from one machine appears to be coming from another one. In fact, an IP source address is not trustable.

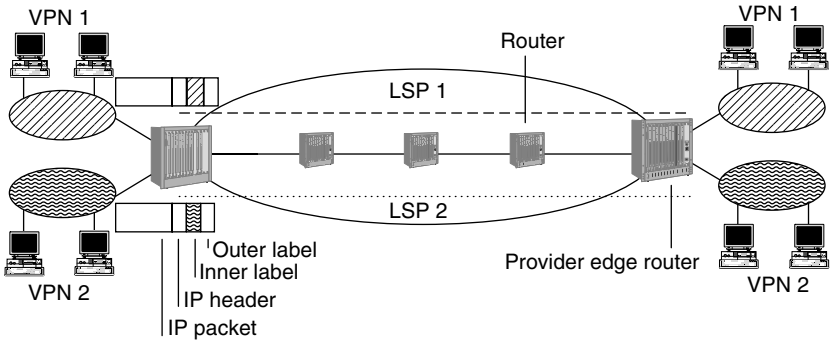


Figure 2. Different VPNs tunneled with MPLS over the same link.

3.1.2. Session Hijacking and "Man in the Middle" Attack. Spoofing makes it possible to take over a connection. Even initial authentication for each communication is no protection against session hijacking. A hacker can take over a session and stay invisible in the middle, pretending to be the respective peer of the two original session partners. The hacker thereby possibly filters and modifies all packets of the session. Identifying the communicating person once does not ensure that it remains the same person throughout the rest of the session. Each data source has to be authenticated throughout the whole session.

3.1.3. Electronic Eavesdropping. A large part of most networks is based on Ethernet LANs. This technology has the advantage of being cheap, universally available, and easy to expand. But it has the disadvantage of making sniffing easy. An even more severe situation nowadays exists in wireless LANs.

In Ethernet networks, every node can read each packet. Conventionally, each network interface card listens and responds only to packets specifically addressed to it. But it is easy to force these devices to collect every packet that passes on the wire. Physically, there is no way to detect from elsewhere on the network which network interface card is working in the so-called promiscuous mode.

Diagnostic tools called "sniffers" get the information out of the collected packets. Such tools can record all the network traffic and are normally used to quickly determine what is happening on any segment of the network. However, in the hands of someone who wants to listen in on sensitive communications, a sniffer is a powerful eavesdropping tool.

The grown Internet structure with the global backbones makes electronic eavesdropping on routers and especially on backbone routers very efficient. Also in virtual LANs that transfer clear text, packets can be eavesdropped easily.

3.2. The Security Architecture for the Internet Protocol (IPSec)

The Internet Engineering Task Force (IETF) standardized IP version 6 (IPv6) [6] to solve pending problems such as address shortage of the current version of the IP protocol (IPv4). A spinoff development of this process was the IP security architecture (IPSec), which introduces per packet security features. While the IP version 6 deployment has been delayed, the security architecture has been adopted by the current IP version (IPv4). A key motivation for this was that IPSec includes all security mechanisms needed to implement VPNs.

The Internet security architecture consists of a family of protocols. IPSec describes IP packet header extensions and packet trailers that provide security functions. The per packet security functions come from two protocols: the Authentication Header (AH) [7], which provides packet integrity, and authenticity and the Encapsulating Security Payload (ESP) [8], which provides privacy through encryption. AH and ESP (Figs. 3–5) are independent protocols that can be used separately and that can be combined. One reason for the separation was that there are countries that have restrictive regulations on encrypted communication. There, IPSec can be deployed solely using AH because authentication mechanisms are not regulated.

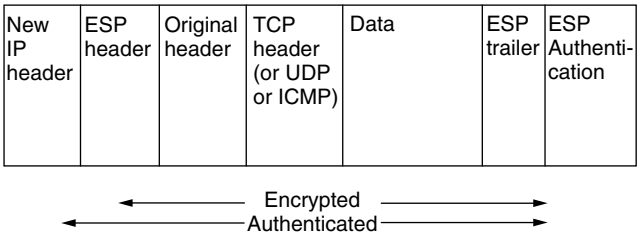


Figure 3. IPSec; IP packet after applying ESP in tunnel mode.

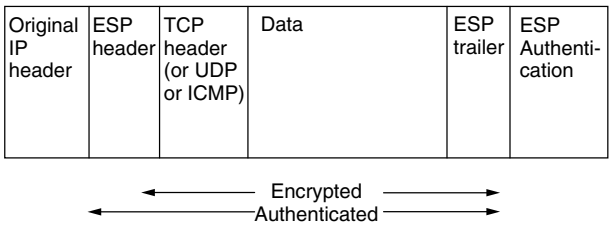


Figure 4. IPSec; IP packet after applying ESP in transport mode.

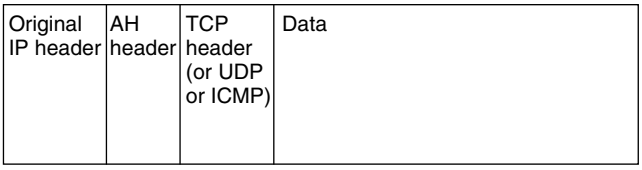


Figure 5. IPSec; IP packet after applying AH in transport mode.

The set of AH and ESP is required in order to guarantee interoperability between different IPSec implementations. Both protocols are specified independently of cryptographic algorithms. A new encryption algorithm, for example, can easily be added to IPSec. Both AH and ESP assume the presence of a secret key. This key material may be installed manually. A better and more scalable approach is to use the third protocol of the IPSec family: the Internet Key Exchange protocol (IKE) [9], described below.

3.2.1. The Encapsulation Security Payload. The Internet Assigned Numbers Authority (IANA) has assigned the protocol number 50 for the IPSec encapsulation security payload. ESP ensures privacy of the IP payload. For that purpose an ESP header and an ESP trailer clamp the IP payload between them. The payload and the trailer are encrypted. The ESP also provides optional authentication. Figure 6 depicts the ESP part of an IP packet transformed by ESP in transport mode. The ESP header is located after the IP header and contains the security parameter index to identify the security association. Furthermore, there is a sequence number that is incremented for each packet. This helps to detect replay attacks, where the attacker records a packet and resends it later.

The ESP trailer is added after the payload. The trailer includes padding that is necessary because the encryption

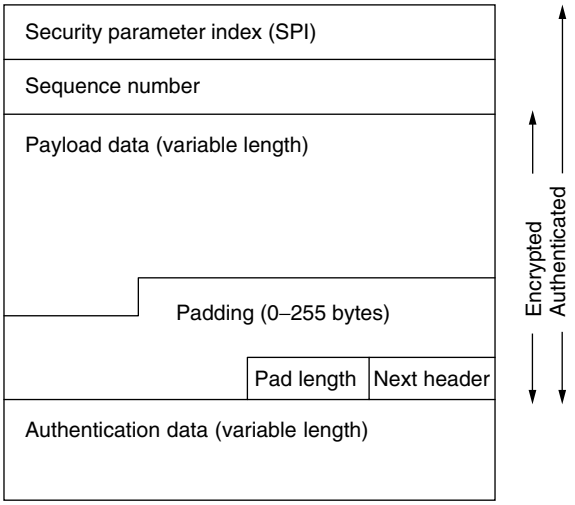


Figure 6. ESP part of an IPSec packet in transport mode.

algorithms often require the payload to be blocks of fixed length (e.g., 8 bytes). The pad length field encodes the length of the padding in bits. The next header field contains the protocol number of the next (eventually higher layer) protocol in the payload (e.g., IP or a concatenated IPSec protocol). Note that the trailer up to here is also encrypted. So, an attacker can, for example, not read what protocol is in the payload data. The ESP trailer may end with optional authentication data. The authentication data consist in a message authentication code (MAC) computed by a secure hash function. The input of the hash is a secret key, the ESP header, the ESP payload, and the rest of the ESP trailer. The MAC does not protect the initial IP header.

ESP supports nearly any kind of symmetric encryption. The default standard built into ESP, which assures basic interoperability, is 56-bit DES. ESP also supports some authentication (as does AH — the two options have been designed with some overlap).

3.2.2. The Authentication Header. The IANA has assigned the protocol number 51 for the IPSec authentication header. AH authenticates the packet so that a receiving IPSec peer can know for sure that the packet originates from the sending peer. Furthermore, the packet integrity is guaranteed. The receiver can verify that nobody has changed the packet while it was in transit between the peers. AH ensures this by calculating authentication data with a secure one-way hash function. The calculation also includes the secret key.

An attacker not knowing this key is neither able to forge a valid packet nor to authenticate the packet. Figure 7 depicts the AH part of an IP packet transformed by AH in transport mode. The AH header includes the next header field and encodes the payload length. The length is necessary because the authentication data are variable in length. The AH header, just like the ESP header, contains a security parameter index and a sequence number. Finally, there is the authentication data (the secure hash value). The authentication of AH also covers the original IP header, in contrast to the optional authentication of ESP. However, some fields of the IP header are excluded

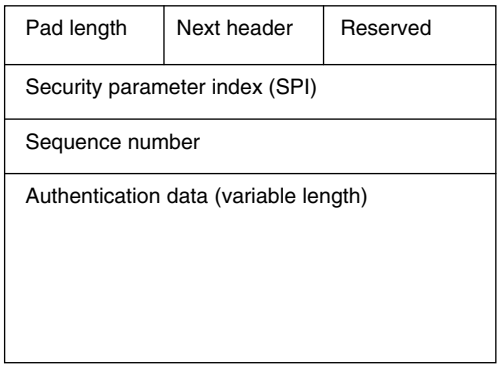


Figure 7. AH part of an IPSec packet.

from the authentication, because their values may change during the forwarding of the packet. These exceptions are the time-to-live field that is decremented by each router and the Differentiated Services Code Point (DSCP) protocol.

The design of the authentication header protocol makes it independent from the higher-level protocol. It can be used with or without ESP. The different fields of the AH are

- The next header field that specifies the higher-level protocol following the AH.
- The pad length field is an 8-bit value specifying the size of the AH.
- The reserved field is reserved for future use and is currently always set to zero.
- The SPI identifies a set of security parameters to be used for this connection.
- The sequence number is incremented for each packet sent with a given security parameter index (SPI).
- Finally, the authentication data are the actual integrity check value (ICV), or digital signature, for the packet. It may include padding to align the header length to an integral multiple of 32 bits (in IPv4) or 64 bits (in IPv6).

To guarantee minimal interoperability, all IPSec implementations must support at least HMAC-MD5 (Keyed-Hash Message Authentication Code for the Message Digest 5 Algorithm) and HMAC-SHA-1 (Keyed-Hash Message Authentication Code for Secure Hash 1 Algorithm) for AH. IPSec, including AH and ESP, has been designed for both IPv4 and IPv6.

3.2.3. Transport and Tunnel Mode. Both ESP and AH have two modes: the transport mode and the tunnel mode. Transport mode just encrypts and authenticates the payload and a part of the IP header. It extends the IP headers by adding new fields. Transport mode allows the user to run IPSec from end-to-end (Fig. 8), while the tunnel mode is ideal for implementing a VPN tunnel at Internet access routers (Fig. 9).

The tunnel mode adds a complete new IP header (plus extension fields). In tunnel mode both AH and ESP can be

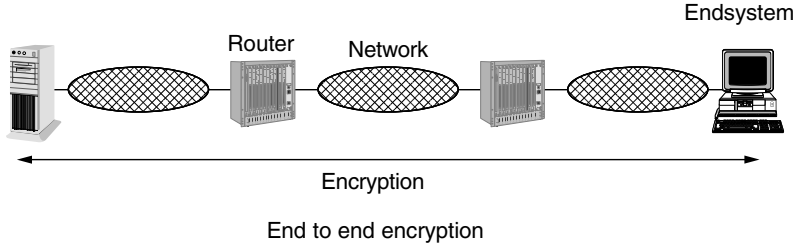


Figure 8. Transport mode.

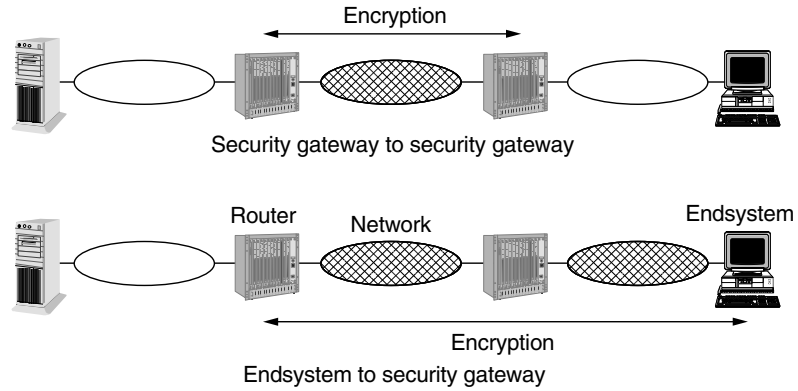


Figure 9. Tunnel mode.

used to implement IP-VPN tunnels. AH and ESP dispose of a small standardized set of cryptographic algorithms to ensure authenticity and privacy. Tunneling takes the original IP packet and encapsulates it within the ESP. Then it adds a new IP header to the packet containing the address of the IPSec gateways. This mode allows passing nonroutable IP addresses or other protocols through a public network as the addresses of the inner header are hidden. Privacy is also given by hiding the original network topology.

3.2.4. Security Association and Security Policy Database. At some point in the network, both AH and ESP perform a transformation to IP packets. The IPSec-compliant nodes always form sender–receiver pairs, where the sender performs the transformation and the receiver reverses it. The relation between sender and receiver is described as a security association (SA). Note that the security association describes just one transformation and its inverse. Concatenated AH and ESP transformations are described by concatenated SAs. SAs can be seen as descriptions of "open" IPSec connections. Both IPSec peering machines store representations of security associations.

Under IPSec, the SA specifies the mode of the authentication algorithm used in the AH and the keys of that authentication algorithm. Also, it specifies the ESP encryption algorithm mode and the respective keys, the presence and size or absence of any cryptographic synchronization to be used in that encryption algorithm, how to authenticate traffic (protocols, encrypting algorithm and key), how to make communication private (again, algorithm and key), how often those keys need to be changed and the authentication algorithm, mode and transform for use in ESP, and the keys to be used by that algorithm. Finally it specifies the key lifetimes, the lifetime of the SA itself, the SA source address, and a sensitivity-level descriptor.

A SA is uniquely identified by a triple consisting of a security parameter index (SPI) (a 32-bit number), the destination IP address, and the IPSec protocol (AH or ESP). The sending party writes the SPI into the appropriate field of the IP protocol extension. The receiver uses this information to identify the correct security association. In that way the receiver is able to invert the transformation and to restore the original packet. Each IPSec-compliant machine may be involved in an arbitrary number of security associations.

Accordingly, a SA is a management construct used to enforce a security policy in the IPSec environment. The policy specifications are stored locally in every IPSec node's security policy database (SPD), which is consulted each time when processing inbound and outbound IP traffic, including non-IPSec traffic. The SPD contains different entries for inbound and outbound traffic. The SPD determines if traffic must be encrypted or can remain clear text or if traffic must be discarded. If traffic is encrypted, the SPD must point to the respective SA by a selector, a set of IP and upper-layer protocol field values to map traffic to a policy.

3.2.5. The Internet Key Exchange Protocol. If two parties would like to communicate using authentication and encryption services, they need to negotiate the protocols, encryption algorithms, and keys to use. Afterward they need to exchange keys (this might include changing them frequently) and keep track of all these agreements.

The Internet Key Exchange (IKE) protocol allows two nodes to securely set up a security association by allowing

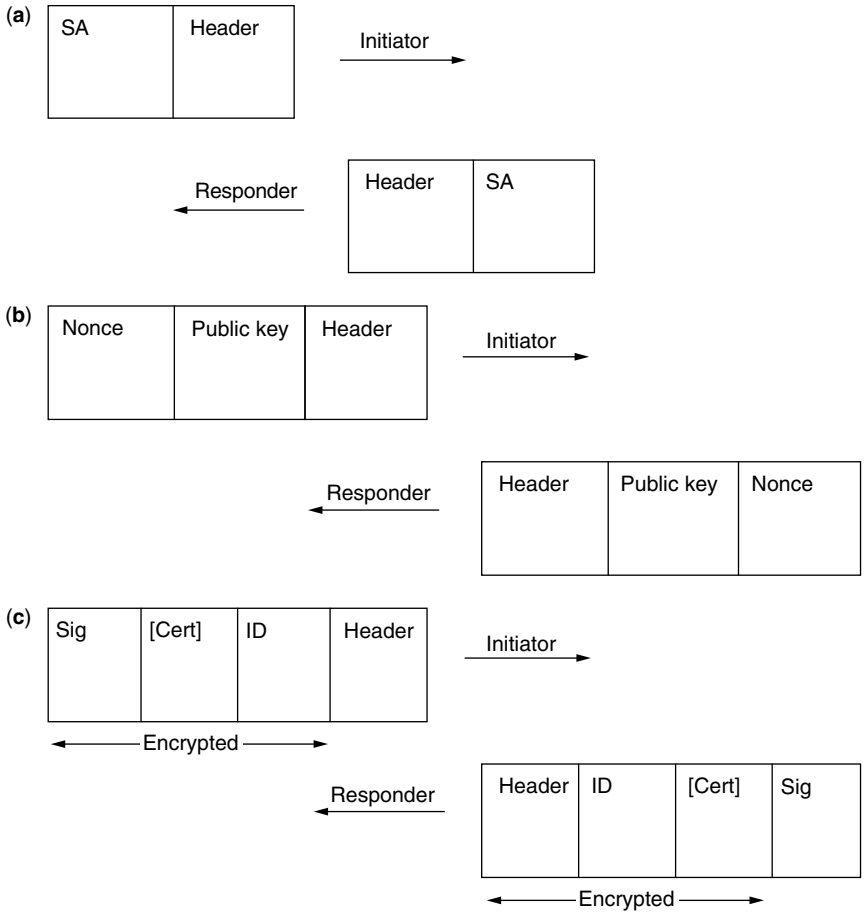


Figure 10. IKE main mode: (**a**) first step; (**b**) second step; (**c**) third step.

these peers to negotiate the protocol (AH or ESP), the protocol mode, and the cryptographic algorithms to be used. Furthermore, IKE allows the peers to renew an established security association.

IKE uses the Internet Security Association and Key Management Protocol (ISAKMP) [10] to exchange messages. ISAKMP provides a framework for authentication and key exchange but does not define a particular key exchange scheme. IKE uses parts of the key exchange schemes Oakley [11] and SKEME [12].

IKE operates in two phases. In phase 1 the two peers establish a secure authenticated communication channel (also called *ISAKMP security association*). In phase 2 security associations can be established on behalf of other services (most prominently IPSec security associations). Phase 2 exchanges require an existing ISAKMP SA. Several phase 2 exchanges can be protected by one ISAKMP SA, and a phase 2 exchange can negotiate several SAs on behalf of other services.

ISAKMP SAs are bidirectional. The following attributes are used by IKE and are negotiated as part of the ISAKMP SA: encryption algorithm, hash algorithm, authentication method, and initial parameters for the Diffie–Hellman algorithm [13].

3.2.5.1. Phase 1 Exchange. IKE defines two modes for phase 1 exchanges: main mode and aggressive mode. The *main mode* consists of three request–response message pairs. The first two messages negotiate the policy (e.g., authentication method) (Fig. 10a); the next two messages exchange Diffie–Hellman public values and ancillary data necessary for the key exchange (Fig. 10b). The last two messages authenticate the Diffie–Hellman exchange (Fig. 10c). The last two messages are encrypted and conceal the identity of the two peers.

The *aggressive mode* of phase 1 consists of only three messages (Fig. 11). The first message and its reply negotiate the policy. Moreover, they exchange Diffie–Hellman public values, ancillary data necessary for the key exchange, as well as identities. In addition the second message authenticates the responder. The third message authenticates the initiator and provides a proof of participation in the exchange. The final message may be encrypted. Aggressive mode securely exchanges authenticated key material and sets up an ISAKMP SA, but it reveals the identities of the ISAKMP SA peers to eavesdroppers. Note, that the choice of the authentication method influences the specific composition of the payload of this exchange. Note also, that IKE assumes security policies that describe what options can be offered during the IKE negotiation.

3.2.5.2. Phase 2 Exchange. A phase 2 exchange negotiates security associations for other services and is protected (encrypted and authenticated) based on an existing ISAKMP security association. The payloads of all phase 2

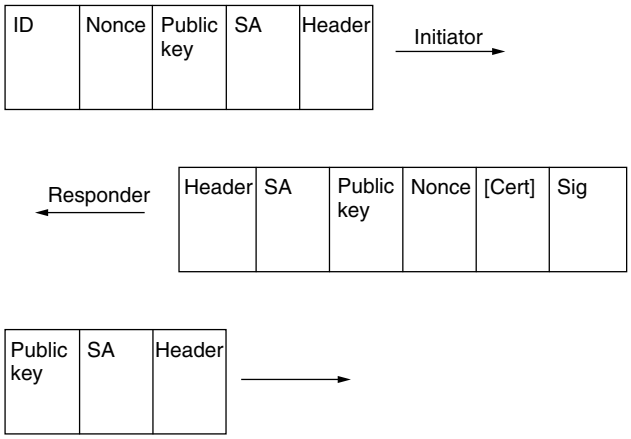


Figure 11. IKE aggressive mode.

messages are encrypted. A phase 2 exchange consists of three messages. The initiator sends a message containing a hash value, the proposed security association parameters and a nonce. The hash value is calculated over ISAKMP SA key material and proves authenticity.

The nonce prevents replay attacks. Optionally, the initial message can also contain key exchange material. Such optional phase 2 key exchange generates key material that is independent from the key material of the ISAKMP SA. If the new SA should be broken, the ISAKMP SA is thus not compromised. The initial message may also contain identifiers in case the new SA is to be established between peers different from the ISAKMP SA peers.

The responder replies with a message of the same structure as the initial message: an authenticating hash value, the selected SA parameters, and a nonce. If the initial message contained optional parameters, then these are also part of the reply. Finally, the initiator acknowledges the exchange with a third and final message containing yet another hash value.

3.2.5.3. Authentication. IKE establishes authenticated keying material. IKE supports four authentication methods to be used in phase 1: preshared secret keys, two forms of authentication with public key encryption, and digital signatures. Today's IKE implementations support X.509 certificates. Two computers not knowing each other can initialize a security association through the help of the commonly trusted third party that verified the certificates.

4. OUTLOOK

The move from legacy-technology-based VPNs such as frame relay and ATM to IP-based VPNs will go on and thereby accelerate the deployment of newer VPN techniques such as generalized MPLS (G-MPLS). GMPLS is being considered as an extension to the MPLS framework to include optical, non-packet-switched technologies. A more recent traffic engineering technology development in the context of G-MPLS is multiprotocol lambda switching (MPλS). The major difference lies in the replacement of the traditional numeric MPLS labels by wavelengths (lambda).

Another trend are mobile devices. Mobile users, as described above, move around and connect through fixed wire dialup lines for example. These users are called "nomadic" users because the from the IP network view they remain locally immobile during a connection time. Roaming users that connect by mobile IP (MIP) require special solutions in the VPN area as with each handover of the mobile node the VPN tunnels need to be reestablished within very short timeframes. An interesting combination of the IPSec suite and the MIP protocols has been described [14] in which, where mobile hosts are allowed access to VPNs that are protected by firewalls from the public Internet.

BIOGRAPHIES

Manuel Günter received his Diploma Degree and his Ph.D. from the University of Berne (Switzerland) in 1998 and 2001, respectively. His research interests are new IP services, network and service monitoring, software engineering, and mobile agents. He now works as a scientific consultant for the Swiss National Bank.

Marc Danzeisen received his Diploma Degree from the University of Bern (Switzerland) in 2001. Since then he has been working as a Ph.D. student in the area of mobile ad hoc networks in collaboration with Swisscom Innovations AG (Switzerland). He is also active in the creation of future mobile network services.

Marc-Alain Steinemann received his Diploma Degree from the University of Bern (Switzerland) in 2000. He is now working as a Ph.D. student in the Computer Networks and Distributed Systems Research Group of the University of Bern. His research interests are remote learning and communication systems for the next-generation Internet.

Torsten Braun received his Diploma Degree and Ph.D. degree from the University of Karlsruhe (Germany) in 1990 and 1993, respectively. From 1994 to 1995 he has been a Guest Scientist with INRIA Sophia-Antipolis (France). From 1995 to 1997 he worked at the IBM European Networking Center Heidelberg (Germany), and at the end of that period served as a project leader and senior consultant. Since 1998, he has been a full Professor of Computer Science at the Institute of Computer Science and Applied Mathematics, heading the Computer Networks and Distributed Systems research group. He is a member of several international conference and workshop program committees and the foundation council of SWITCH (Swiss National Research Network).

BIBLIOGRAPHY

1. Y. Rekhter et al., *Address Allocation for Private Internets*, RFC 1918, Feb. 1996.

2. P. Ferguson and G. Huston, What is a VPN — part I, *Internet Protocol J.* **1**(1): 2–11 (1998).

3. P. Ferguson and G. Huston, What is a VPN — part II, *Internet Protocol J.* **1**(2): 2–18 (1998).

4. B. Gleeson et al., *A Framework for IP Based Virtual Private Networks*, RFC 2764, Feb. 2000.

5. I. Khalil, T. Braun, and M. Günter, Management of quality of service enabled VPNs, *IEEE Commun. Mag.* **39**(5): 90–98 (May 2001).

6. S. Deering and R. Hinden, *Internet Protocol, Version 6* (*IPv6*) *Specification*, RFC 2460, Dec. 1998.

7. S. Kent and R. Atkinson, *IP Authentication Header*, RFC 2402, Nov. 1998.

8. S. Kent and R. Atkinson, *IP Encapsulating Security Payload* (*ESP*), RFC 2406, Nov. 1998.

9. D. Harkins and D. Carrel, *The Internet Key Exchange* (*IKE*), RFC 2409, Nov. 1998.

10. D. Maughhan, M. Schertler, M. Schneider, and J. Turner, *Internet Security Association and Key Management Protocol*, RFC 2408, Nov. 1998.

11. H. Orman, *The Oakley Key Determination Protocol*, RFC 2412, Nov. 1998.

12. F. Knabe, An overview of mobile agent programming, in *Analysis and Verification of Multiple-Agent Languages*, Vol. 1192 of *Lecture Notes in Computer Science*, Springer, June 1996; 5th LOMAPS Workshop.

13. B. Schneier, *Applied Cryptography*, Wiley, New York, 1996.

14. M. Danzeisen and T. Braun, Access of mobile IP users to firewall protected VPNs, *Workshop on Wireless Local Networks at 26th Annual IEEE Conf. Local Computer Networks* (LCN'2001), Tampa, FL, Nov. 15–16, 2001.

VITERBI ALGORITHM

A. J. VITERBI
Viterbi Group
San Diego, California

1. FUNDAMENTALS

The Viterbi algorithm is a computationally efficient technique for determining the most probable path taken through a Markov graph. The graph and underlying Markov sequence is characterized by a finite set of states $\{S_0, S_1, \ldots, S_n, \ldots\}$, state-transition probabilities $\Pr(S_j \to S_i)$, and the output (observable parameter) probabilities $p(y/S_j \to S_i)$ for all i and j, where the observables y are either discrete or continuous random variables. An example of a four-state Markov graph is shown in Fig. 1, where only the nonzero probability transitions are shown. Thus, for example, from S_1 the only nonzero transition probabilities are those to S_2 and S_3, while from S_3 they are those to itself, S_3, and to S_2.

It is also convenient for the description of the algorithm to view the multistep evolution of the path through the graph by means of a multistage replication of the Markov graph known as a *trellis* diagram. Figure 2 is the trellis diagram corresponding to the four-state Markov graph of Fig. 1.

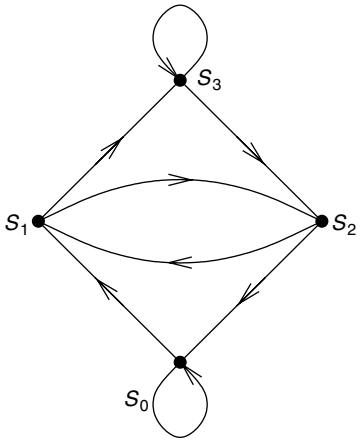


Figure 1. Markov graph example.

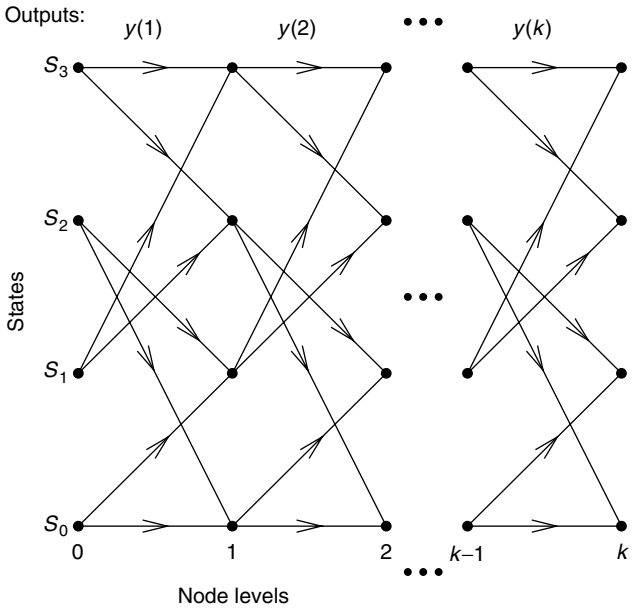


Figure 2. Trellis diagram for Markov graph of Fig. 1.

At the top of Fig. 2 are shown the successive observables $y(1), y(2), \ldots, y(k), \ldots$, each of which may be vectors corresponding to multiple observations per branch. Henceforth we use the notation $y(k)$ to denote the observable(s) for the kth successive branch. Similarly, $S(k)$ will denote any state at the kth successive node level (and we shall dispense with subscripts until necessary).

The goal, then, is to find the most probable path through the trellis diagram. A fundamental assumption is that successive Markov state probabilities $\Pr[S(k-1) \to S(k)]$ are mutually independent for all k, as are the conditional output probabilities $p[y(k)/S(k-1) \to S(k)]$. For any given path from the origin ($k = 0$) to an arbitrary node n, $S(0), S(1), \ldots, S(n)$, the relative path probability (likelihood function) is given by

$$L = \prod_{k=1}^{n} \Pr[S(k-1) \to S(k)]\, p[y(k)/S(k-1) \to S(k)]$$

For computational purposes it is more convenient to consider its logarithm, which is given by the sum

$$\ln(L) = \sum_{k=1}^{n} m[y(k); S(k-1), S(k)]$$

where

$$m[y(k); S(k-1), S(k)] = \ln\{\Pr[S(k-1) \to S(k)] + \ln p[y(k)/S(k-1) \to S(k)]\}$$

which is denoted the *branch metric* between any two states at the $(k-1)$th and kth node levels. We next define the *state metric*, $M_K(S_i)$, of the state $S_i(K)$ to be the maximum over all paths leading from the origin to the ith state at the Kth node level. Thus, again inserting subscripts where necessary, we obtain

$$M_K(S_i) = \max_{\text{all paths } S(0),S(1),\ldots,S(K-1)} \left\{ \sum_{k=1}^{K-1} m[y(k); S(k-1), S(k)] + m[y(K); S(K-1), S_i(K)] \right\}$$

It then follows that to maximize this sum over K terms, it suffices to maximize the sum over the first $K-1$ terms for each state $S_j(K-1)$ at the $(K-1)$th node and then maximize the sum of this and the Kth term over all states $S(K-1)$. Thus

$$M_K(S_i) = \max_{S_j(K-1)} \{M_{K-1}(S_j) + m[y(K); S_j(K-1), S_i(K)]\}$$

This recursion is known as the *Viterbi algorithm*. It is most easily described in connection with the trellis diagram. If we label each branch (allowable transition between states) by its branch metric $m[\,]$ and each state at each node level by its state metric $M[\,]$, the state metrics at node level K are obtained from the state metrics at the level $K-1$ by adding to each state metric at level $K-1$ the branch metrics that connect it to states at the Kth level, and for each state at level K preserving only the largest sum that arrives to it. If additionally at each level we delete all branches other than the one that produces this maximum, there will remain only one path through the trellis leading from the origin to each state at the Kth level, which is the most probable path reaching it from the origin. In typical (but not all) applications, both the initial state (origin) and the final state are fixed to be S_0 and thus the algorithm produces the most probable path through the trellis both initiating and ending at S_0.

2. APPLICATIONS

Numerous applications of this algorithm have appeared since the 1960s or so. The following list represents the most prominent in approximate chronological order:

1. Decoders for convolutional codes on various wireless channels
2. Maximum-likelihood sequence estimation (MLSE) demodulators for intersymbol interference and multipath fading channels
3. Decoders for recorded data
4. Optical character recognition
5. Voice recognition
6. DNA sequence alignment

We shall describe each in the order indicated above.

2.1. Convolutional Codes

The earliest application, for which the algorithm was originally proposed in 1967, was for the maximum likelihood decoding of convolutionally coded digital sequences transmitted over a noisy channel. Currently the algorithm forms an integral part of the majority of wireless telecommunication systems, both involving satellite and terrestrial mobile transmission. The convolutional encoder and channel combination, as shown in Fig. 3, gives rise directly to the Markov graph representation. In the simplest case, one bit at a time enters the L-stage shift register and the n linear combiners, each of which is a modulo-2 adder of the contents of some subset of the L shift register stages, generate n binary symbols. These are serially transmitted, for example, as binary amplitude ($x = +1$ or -1) modulation of a carrier signal. At the receiver, the demodulator generates an output y, which is either a real number or the result of quantizing the latter to one of a finite set of values. The conditional densities $p(y/x)$ of the channel outputs are assumed to be mutually independent, corresponding to a "memoryless channel." A commonly treated example of such is the additive white Gaussian noise (AWGN) channel for which each y is the sum of the encoded symbol x and a Gaussian random noise variable, with all noise variables mutually independent. This channel model is closely approximated by satellite and space communication applications and, with appropriate caution, it can also be applied to terrestrial communication design.

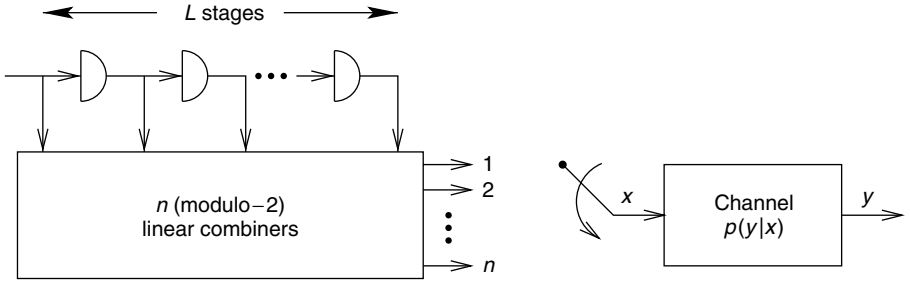


Figure 3. Convolutionally encoded memoryless channel.

The communication system model just described gives rise naturally to the Markov graph representation. The 2^L states correspond to the states of the contents of the L-stage register. Thus S_0 corresponds to the contents being all zeros, S_1 to the first stage containing a one and all the rest zeros, and so on. Since only one input bit changes each time, each state has only two branches both exiting and entering it, each from two other states. For the exiting branches, one corresponds to a zero entering the register and the other to a one. Figure 1 could be used to represent a two-stage encoder, where the state indices are the decimal equivalents of the binary register contents. It is generally assumed that all input bits are equally likely to be a zero or a one, so the state transition probabilities, $P(S_j \to S_i) = \frac{1}{2}$ for each branch. Hence the first term of the branch metric m can be omitted since it is the same for each branch. As for the second term of m, the conditional probability density $p(y/S_j \to S_i) = p(y/x)$, where x is an n-dimensional binary vector generated by the n modulo-2 adders for each new input bit, which corresponds to a state transition, while y is the random vector corresponding to the n noise-corrupted outputs for the n channel inputs represented by the vector x. For the AWGN, $\ln p(y/x)$ is proportional to the inner product of the two vectors x and y.

The convolutional encoder and its Markov graph just described represent a rate $1/n$ code, since each input bit generates n output symbols. To generalize to any rational rate $m/n < 1$, m input bits enter each time and the register shifts in blocks of m. The Markov graph changes only in having each state connected to 2^m other states. Another generalization is to map each binary vector x, not into a vector of n binary values, $+1$ or -1, but into a constellation of points in two or more dimensions. An often employed case is quadrature amplitude modulation (QAM). For example, for $n = 4$, 16 points may be mapped into a two-dimensional (2D) grid and the value in each dimension modulates the amplitude of one of the two quadrature components of the sinusoidal carrier. Here x is the 2D vector representing one of the 16 modulating values and y is the corresponding demodulated channel output. Multiple generalizations of this approach abound in the literature and in real applications. In most cases this multidimensional approach is used to conserve bandwidth at the cost of a higher channel signal-to-noise requirement.

An interesting footnote on this first application is that the Viterbi algorithm was proposed not so much to develop an efficient maximum-likelihood decoder for a convolutional code, but primarily to establish bounds on its error correcting performance.

2.2. MLSE Demodulators for Intersymbol Interference and Multipath Fading Channels

In the previous application, the convolution operation is employed in order to introduce redundancy for the purpose of increasing transmission reliability. But convolution also occurs naturally in physical channels whose bandwidth constraints linearly distort the transmitted digital signal. Treating the channel as a linear filter, it is well known that the output signal is the convolution of the input signal and the filter's impulse response. A discrete model of the combination of signal waveform, channel effects, and receiver filtering is shown in Fig. 4. This combination produces, after sampling at the symbol rate, the discrete convolution $x_k = \sum_{j=0}^{m} h_j u_{k-j}$, where the variables u are the input bit sequence, generally taken to be binary ($+A$ or $-A$). The h_j terms are called the *intersymbol interference* coefficients, since except for the h_0 term, all the other terms of the sum represent interference by preceding symbols on the given symbol. To the output of the discrete filter x_k must be added noise variables n_k to account for the channel noise. While generally these noise variables are not mutually independent, they can be made so by employing at the receiver a so-called whitened matched filter prior to sampling.

A related application with a very similar model is that of multipath fading channels. Here the taps represent multiple delays in a multipath channel. Their relative spacings may be less than one symbol period, in which case each symbol of the input sequence must be repeated several times. Most important, because of random variations in propagation, the multipath coefficients h are now random variables, so the conditional densities of y depend on the statistics of both the additive noise and the multiplicative random coefficients.

Comparing Fig. 4 with Fig. 3, we note that the principal difference is that modulo-2 addition is replaced by real addition and there is one rather than n outputs for each branch. Otherwise, the same Markov graph applies, with two branches emanating from each state when the input sequence is binary ($+A$ or $-A$). Generation of the branch metrics, $\ln p(y/S_j \to S_i)$, is slightly more complex because besides depending on the noise, the y variables depend on a linear combination of the h variables, with their signs

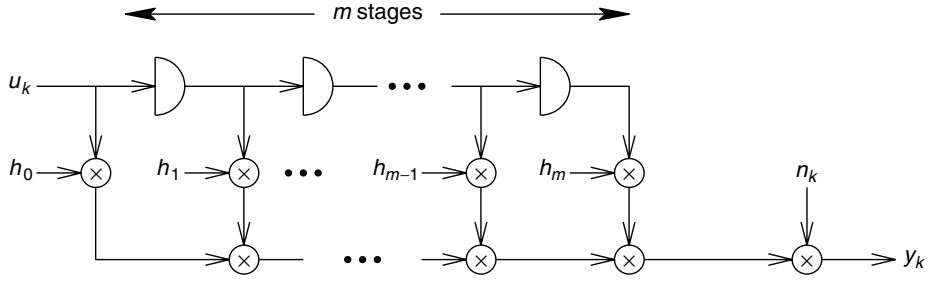


Figure 4. Intersymbol interference or multipath fading channel model.

determined by the register contents involved in the branch transition.

2.3. Partial-Response Maximum-Likelihood Decoders for Recorded Data

A filter model for the magnetic recording medium and reading process is very similar to the intersymbol interference model. The simplest version, known as the "partial response" channel results, when sampled at the symbol rate, in an output depending on just the difference between two input symbols, $x_k = u_k - u_{k-2}$, with additive noise samples also being mutually independent. Note that since inputs u are $+1$ or -1, the outputs x (prior to adding noise) are ternary, $+2$, 0, or -2. This then reduces to the model of Fig. 4 with just two nonzero taps, $h_0 = +1$ and $h_2 = -1$. Often this is described by the polynomial whose coefficients are the tap values; in this case $h(D) = (1 - D^2)$. This can be modeled by a two-stage shift register that gives rise to a four-state Markov graph, as in Figs. 1 and 2. But actually, a simpler model can be used based on the fact that all outputs for which k is odd depend only on odd-indexed inputs, and similarly for even. Thus a two-state Markov graph suffices for each of the odd and even subsets, as shown in Fig. 5. When the recording density is increased, the simple partial-response channel model is replaced by a longer shift register. A generally accepted model has tap coefficients represented by the polynomial $h(D) = (1 - D)(1 + D)^N$, where $N > 1$. For example, for the case of $N = 2$, known as "extended partial response," an eight-state Markov graph applies.

The Markov nature of the preceding three applications is obvious from the system model. For the next three, it is not as obvious and often it is only a tentative model of the phenomenon derived from observations. In such cases the term *hidden Markov model* (HMM) is used. Such models are often empirically derived based on experience in the given discipline. Since background in each field is a prerequisite for full understanding, we give only a superficial description of each of the following.

2.4. Optical Character Recognition

The algorithm has been applied to the automatic character recognition of hand-printed text. For many decades, statistical analysis of English (and other languages) has led to a Markov model whose states are the single letters, digrams, trigrams, or generally N-grams, of English text, as measured by their relative frequency in numerous published texts. Thus given 27 letters, including space as a letter, a Markov graph of 27^N states can be created. Unlike the previous applications, here the branch metric

$$m[y(k); S(k-1), S(k)] = \ln\{\Pr[S(k-1) \to S(k)] + \ln p[y(k)/S(k-1) \to S(k)]\}$$

depends as much on the first term as on the second. The first term, of course, is determined, as just noted, from the predetermined N-gram transition relative frequencies. (The transitions will involve just a change of a single letter as, e.g., transition from THA to HAT.) As for the observables $y(k)$, these may be as simple as the character recognizer's preliminary estimate (without benefit of the Markov character) to measurements on a grid of points, possibly including gray levels, which will result in better performance. In any case, the conditional density of the measurement values given the letter causing the branch transition must also be predetermined to implement the second term.

2.5. Voice Recognition

An increasingly popular application has been to voice recognition for automated dialing or response systems. The situation is very similar to character recognition, except that the language text characters are replaced by *phonemes*, which are voiced fragments of speech. So the states may be the phoneme N-grams and the first term of the branch metrics are the predetermined relative frequencies of transitions between phonemes. The second term then depends on measurements of the recorded voice phoneme and its conditional probability relative to the actual phoneme causing the given transition.

2.6. DNA Sequence Analysis

The most recent and most surprising application has been to the alignment of strands of DNA sequences involved in mapping the genome. DNA sequences consist of four types of nucleotides labeled A, C, G, and T. In addition, in analyzing similarities among sequences, one must accept the possibility of insertions and deletions. Thus the states of the hypothetical hidden Markov model each involve nucleotides, insertions, or deletions or some combination thereof. The Markov state-transition probabilities are initially assigned arbitrarily or based on previous experience. After the maximum-likelihood alignments are found by means of the algorithm, the relative frequencies of the state transitions are counted and used as the transition probabilities for a second iteration of the algorithm. This process may continue

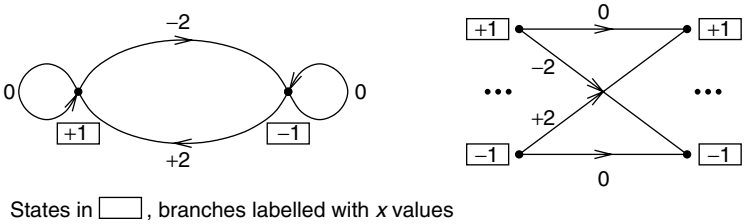


Figure 5. Markov graph and Trellis diagram for even (odd) partial-response states.

through several iterations until the measured relative frequencies of a given iteration provide a good match to those of the hypothesized model from the previous iteration. In this way the accuracy and reliability of the HMM is refined along with the maximum-likelihood alignments.

2.7. Other Applications

Given the pervasiveness of Markov models for a variety of fields, there have been numerous other applications of the Viterbi algorithm throughout the engineering and scientific literature, and there will likely continue to be more. The list is too lengthy and some applications are too obscure to identify. We have provided here the six most often cited.

BIOGRAPHY

Dr. Andrew Viterbi is a cofounder, retired vice chairman, and chief technical officer of QUALCOMM Incorporated. He spent equal portions of his career in industry, having also cofounded a previous company, and in academia as a professor in the Schools of Engineering first at UCLA and then at UCSD, at which he is now professor emeritus. He is currently president of the Viterbi Group, a technical advisory and investment company.

His principal research contribution, the Viterbi algorithm, is used in most digital cellular phones and digital satellite receivers, as well as in such diverse fields as magnetic recording, voice recognition, and DNA sequence analysis. In recent years he has concentrated his efforts on establishing CDMA as the multiple access technology of choice for cellular telephony and wireless data communication.

Dr. Viterbi has received numerous honors both in the United States and internationally. Among these are four honorary doctorates, from the Universities of Waterloo, Rome, Technion, and Notre Dame, as well as memberships in the National Academy of Engineering, the National Academy of Sciences, and the American Academy of Arts and Sciences. He has received the Marconi International Fellowship Award, the IEEE Alexander Graham Bell and Claude Shannon Awards, the NEC C&C Award, the Eduard Rhein Award, and the Christopher Columbus Medal.

VOCODERS

ANDREAS SPANIAS
Arizona State University
Tempe, Arizona

1. INTRODUCTION

Vocoder is a term that is formed from the concatination of the words *voice* and *coder*. Although originally it became associated with a specific class of analysis/synthesis systems for speech bandwidth reduction, such as the channel vocoder, it is now used for a wider class of algorithms for the compression of bit rate of speech signals. Strictly speaking, *vocoders* as opposed to *waveform coders* represent a subcategory of speech coders that make heavy use of speech spectral properties and rely on a source system model for the representation and parametrization of speech. More recently the term *vocoder* has been used more loosely, and essentially it became associated with the speech coding algorithms that are used in cellular phones, streaming speech applications, Internet telephony, secure communications, digital answering machines, portable digital voice recorders, and other applications.

Vocoder algorithms are embedded in several international standards formed for telephony and multimedia applications. The standardization section of the International Telecommunications Union (ITU), formerly CCITT, has been developing compatibility standards for telephony and more recently for Internet and multimedia applications. Other standardization committees, such as the European Telecommunications Standards Institute (ETSI) and the International Standards Organization (ISO), have also drafted requirements for speech (GSM) and audio/video coding (MPEG) standards. In addition to these organizations there are also committees forming standards for private or government applications such as secure telephony, satellite communications, and emergency radio applications. Standard specifications have driven much of the research and development in the speech coding area and several speech and audio coding algorithms have been developed and eventually adopted in international standards. A series of competing speech signal models based on linear predictive coding (LPC) [1–7] and transform-domain analysis–synthesis [8–10] have been proposed since the mid-1980s. On the other hand, high-end audio coding relied heavily on sub-band and transform coding algorithms that use psychoacoustic signal models [11,12].

Speech coding for low-rate applications involves parametric representation of speech using analysis synthesis systems. Analysis can be open-loop or closed-loop. In closed-loop analysis, also called *analysis-by-synthesis*, the parameters are extracted and encoded by minimizing the perceptually weighted difference between the original and reconstructed speech. Speech coding algorithms are evaluated based on speech quality, algorithm complexity, delay, and robustness to channel and background noise. Moreover in network applications coders must perform reasonably well with nonspeech signals such as DTMF tones, voiceband data, and modem tones. Standardization of candidate speech coding algorithms involves evaluation of speech quality using subjective measures such as the *Mean Opinion Score* (MOS), which involves rating speech according to a 5-level quality scale, as shown in Table 1.

A MOS of 4–4.5 is associated with network or toll quality (wireline telephone grade), and scores between 3.5 and 4 imply communications quality (cellular grade). The simplest coder that achieves toll quality is the 64 kbps (kilobits per second) ITU G.711 pulse code modulation (PCM), which has a MOS of 4.3. Several of the new general-purpose algorithms, such as the 8 kbps ITU G.729 [13] and 6.3 kbps G.723.1 [38], also achieve toll quality. Algorithms

Table 1. The MOS Scale

MOS	Subjective Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

for the cellular standard such as the *TIA* IS54 [14], the full-rate ETSI GSM 6.10 [15] achieve communications quality, and the old Federal Standard 1015 (LPC-10e) [16] is associated with synthetic quality. More recent algorithms for cellular standards have near-toll quality performance. Such algorithms include the adaptive multirate (AMR) coder for the GSM system [42,43] and the selectable mode vocoder (SMV) [52] for use in wideband CDMA applications. The new frontier in vocoder standardization targets toll quality at 4 kbps. In fact, ITU is currently evaluating several algorithms [45–51] that aim for toll quality at 4 kbps.

In this article, we will focus on speech coding algorithms. Section 2 presents an introduction to speech properties and the opportunities for bit rate reduction. Section 3 discusses source system representations of speech and the use of linear prediction in speech coding. Section 4 presents open-loop algorithms and the classical LPC-10 algorithm. Section 5 presents closed-loop LP algorithms, and Section 6 focuses on code-excited linear prediction (CELP) and three generations of standardized algorithms based on CELP. Section 7 concludes with a summary of this article.

2. THE SPEECH SPECTRAL PROPERTIES AND VOCODERS

Before we begin our presentation of vocoder algorithms, we discuss some of the important speech spectral properties that provide opportunities for speech parametrization and compression. First, in digital telephony applications speech is typically band-limited to 3.2 kHz and sampled at 8 kHz. The bandwidth of 3.2 kHz preserves both speech intelligibility and the speaker identity. Speech is a nonstationary random signal and is considered as quasistationary only over short segments, typically 5–20 ms. Speech can generally be classified as *voiced* (e.g., /a/, /i/), *unvoiced* (e.g., /sh/), or *mixed*. We note that this course classification into voiced/unvoiced/mixed segments is adequate only for coding applications. Speech recognition and voice synthesis algorithms used a finer and much more precise phonemic classification. Time-domain plots for sample voiced and unvoiced segments are shown in Fig. 1. A segment of voiced speech from an uttered steady vowel is quasiperiodic in the time domain and harmonically structured in the frequency domain. *Unvoiced* speech is randomlike and broadband. In addition, the energy of voiced segments is generally higher than the energy of unvoiced segments.

The short-time spectrum of *voiced* speech is characterized by its fine and formant structure. The fine harmonic structure is due to the periodicity of voiced speech and

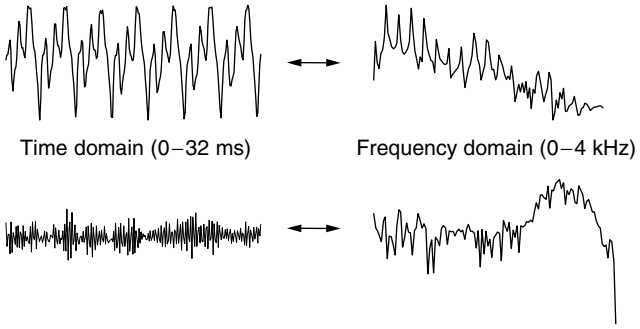


Figure 1. Voiced (top) and unvoiced (bottom) waveforms and their short-time spectra.

is attributed to the activity of the vibrating vocal chords. The *formant* structure (spectral envelope) is due to the interaction of the source and the vocal tract. The spectral envelope shown in Fig. 2 "fits" the short-time spectrum of voiced speech and is associated with the transfer characteristics of the vocal tract and the spectral tilt (6 dB/octave) due to the glottal pulse. The spectral envelope is characterized by a set of peaks called *formants*. The formants are the resonant modes of the vocal tract. For the average vocal tract there are three to five formants below 5 kHz.

The amplitudes and locations of the first three formants, which usually occur below 3 kHz, are quite important in both speech synthesis and perception. Higher formants are also important for wideband and unvoiced speech representations. The properties of speech are related to the physical speech production system as follows. Voiced speech is produced by exciting the vocal tract with quasiperiodic glottal air pulses generated by the vibrating vocal chords. The frequency of the periodic pulses is referred to as the *fundamental frequency* or "pitch." Unvoiced speech is produced by forcing air through a constriction in the vocal tract. Nasal sounds (e.g., /n/) are due to the acoustical coupling of the nasal tract to the vocal tract, and plosive sounds (e.g., /p/) are produced by abruptly releasing air pressure that was built up behind a closure in the tract.

3. SOURCE SYSTEM MODELS AND SHORT-TERM LINEAR PREDICTION

Speech is produced by the interaction of the vocal tract with the vocal chords in the glottis. Engineering models (Fig. 3) for speech production typically model the vocal tract as a time-varying digital filter excited by quasiperiodic waveform when speech is voiced (e.g., as in steady vowels) and random waveforms for unvoiced speech (e.g., as in consonants). The vocal tract filter is estimated using linear prediction (LP) algorithms [17,18].

Linear prediction algorithms are part of several speech coding standards, including ADPCM systems [19–21], open-loop linear predictive coders [16], and *code-excited linear prediction* (CELP) algorithms and other analysis-by-synthesis linear predictive coders [21–26]. In linear prediction the most recent sample of speech is predicted by a linear combination of past samples. This is done using a finite-length impulse response (FIR) filter (Fig. 4) whose

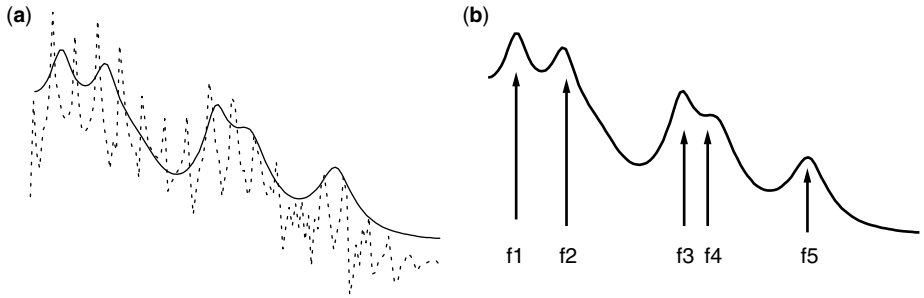


Figure 2. (**a**) Spectral envelope and (**b**) formants.

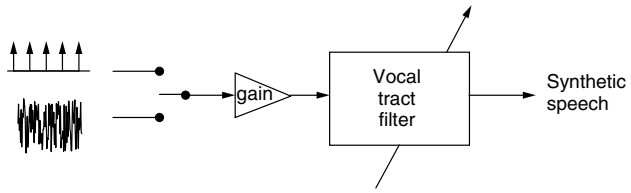


Figure 3. Engineering model for speech synthesis.

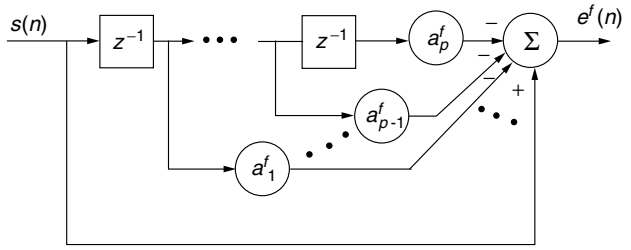


Figure 4. Linear prediction analysis.

output $e(n)$ is minimized. The output of the linear predictor analysis filter is called the linear prediction residual.

The LP coefficients are chosen to minimize the mean square of the LP residual

$$\varepsilon = E[e^2(n)] \tag{1}$$

where the error is the difference of the current speech sample and a linear combination of past samples:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

Because only short-term delays are considered in (2), the linear predictor in Fig. 4 is also known as a *short-term linear predictor*. The inverse of the LP analysis filter is an all-pole filter called the *LP synthesis filter*. The frequency response associated with the short-term synthesis filter captures the *formant* structure of the short-term speech spectrum. The all-pole filter or vocal tract transfer function is given by

$$H(z) = \frac{g}{1 - \sum_{k=1}^{M} a_k z^{-k}} \tag{3}$$

The minimization of ε yields a set of equations involving autocorrelations $[r_{ss}(m)]$ of speech and the LP coefficients are found by inverting an autocorrelation matrix using the Levinson–Durbin algorithm [24].

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ a_M \end{bmatrix} = \begin{bmatrix} r_{xx}(0) & r_{ss}(-1) & r_{ss}(-2) & \cdots & r_{xx}(1-M) \\ r_{ss}(1) & r_{ss}(0) & r_{ss}(-1) & \cdots & r_{ss}(2-M) \\ r_{ss}(2) & r_{ss}(1) & r_{ss}(0) & \cdots & r_{ss}(3-M) \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ r_{ss}(M-1) & r_{ss}(M-2) & r_{ss}(M-3) & \cdots & r_{ss}(0) \end{bmatrix}^{-1} \times \begin{bmatrix} r_{ss}(1) \\ r_{ss}(2) \\ r_{ss}(3) \\ \cdot \\ \cdot \\ r_{ss}(M) \end{bmatrix} \tag{4}$$

There are many efficient algorithms [17,18] for inverting this matrix including algorithms tailored to work well with finite-precision arithmetic [33]. Preconditioning of the speech and autocorrelation data using tapered windows improves the numerical behavior of these algorithms. In addition, bandwidth expansion or scaling of the LP coefficients is very typical in LPC as it reduces distortion during synthesis.

In short-term LP the analysis window is typically 20 ms long. In order to avoid transient effects from large changes in the LP parameters from one frame to the next, the frame is usually divided into subframes (typically 5 ms long), and subframe parameters are obtained by linear interpolation. The direct form LP coefficients a_k are not adequate for quantization and transformed coefficients are typically used in quantization tables. The reflection or lattice prediction coefficients are a byproduct of the Levinson recursion and have better quantization properties than do direct-form coefficients. Some of the early standards such as the LPC-10 [16] and the IS54 VSELP [14] encode reflection coefficients for the vocal tract. Transformation of the reflection coefficients can also lead to a set of parameters that are less sensitive to quantization. In particular, the log area ratios (LARs) and the inverse sine transformation have been used in the early GSM 6.10 algorithm [15] and in the skyphone standard [22]. Most recent LPC-related cellular standards [23–25] quantize line spectrum pairs (LSPs). The main advantage of the LSPs is that they relate directly to frequency-domain information and hence they can be encoded using perceptual criteria. In most recent toll-quality standards such as the wideband CDMA SMV [52], the GSM adaptive multirate vocoder [42,43], the ITU G.723.1 [38], and the

ITU G.729 [13] the LSPs are encoded using split-vector quantization.

3.1. Linear Prediction and ADPCM

One of the simplest scalar quantization schemes that uses short-term LP is the adaptive differential pulse code modulation (ADPCM) coder [19,20]. ADPCM algorithms encode the difference between the current and the predicted speech samples. The prediction parameters are obtained by backward estimation, namely, from quantized data, using a gradient algorithm. The ADPCM 32-kbps algorithm in the ITU G.726 standard (formerly known as CCITT G.721) uses a pole-zero adaptive predictor (Fig. 5). ITU G.726 also accommodates 16, 24, and 40 kbps with individually optimized quantizers. The ITU G.727 has embedded quantizers and was developed for packet network applications. Because of the embedded quantizers, G.727 has the capability to switch easily to lower rates in network congestion situations by dropping bits. The MOS for 32 kbps G.726 is 4.1, and complexity is estimated to be 2 million instructions per second (Mips) on special-purpose chips and about 10 Mips on generic fixed-point DSP chips.

The International Mobile Satellite B (INMARSAT-B) standard [27] uses also a 10MIPS ADPCM coder with a long-term predictor (discussed later) in addition to short-term LP. The INMARSAT-B algorithm operates at 12.8 and 9.6 kbps.

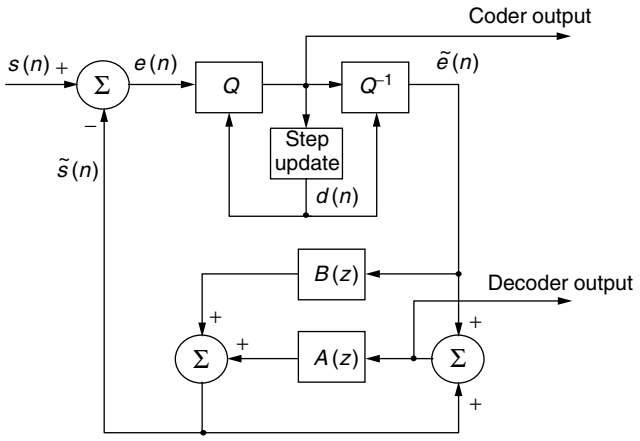


Figure 5. The ITU G.726 ADPCM encoder.

4. SPEECH ANALYSIS–SYNTHESIS USING LINEAR PREDICTION

Unlike waveform ADPCM coders that use LP only for differential quantization, this class of algorithms use LP to represent the vocal tract and make explicit use of the synthesis model shown in Fig. 3. Our discussion of linear prediction is divided in two categories: open-loop analysis–synthesis LP and closed-loop analysis-by-synthesis linear prediction. In the following, we describe two open-loop algorithms, the LPC-10 and the mixed-excitation LPC. Unless otherwise stated, the input to all coders discussed in this section is speech sampled at 8 kHz.

4.1. Open-Loop Linear Prediction

This section describes source system algorithms that use open-loop analysis to determine the excitation sequence. Open-loop linear predictive vocoders are essentially the first-generation LP vocoders. The DOD LPC-10 is a good example of an algorithm that uses open-loop analysis. In 1976 a consortium established by the U.S. Department of Defense (DoD) recommended an LPC algorithm for secure communications at 2.4 kbps, (Fig. 6). The algorithm, known as the LPC-10, eventually became the original Federal Standard FS-1015 [16,28,29]. The LPC-10 uses a 10th-order predictor to estimate the vocal tract parameters. Segmentation and frame processing in LPC-10 depend on voicing. Pitch information is estimated using the average magnitude difference function (AMDF) [16]. Voicing is estimated using energy measurements, zero-crossing measurements, and the maximum to minimum ratio of the AMDF. The excitation signal for voiced speech in the LPC-10 consists of a sequence that resembles a sampled glottal pulse. This sequence is defined in the standard [16] and periodicity is created by a pitch-synchronous pulse repetition process. The LPC-10 produces synthetic speech with a MOS of 2.3. Complexity is estimated at 5–7 Mips.

4.2. Mixed-Excitation Linear Prediction

In 1996 the U.S. government standardized a new 2.4-kbps algorithm called *mixed-excitation LP* (MELP) [7]. The development of mixed-excitation models in LPC was motivated largely by voicing errors in LPC-10 and also by the inadequacy of the two-state excitation model

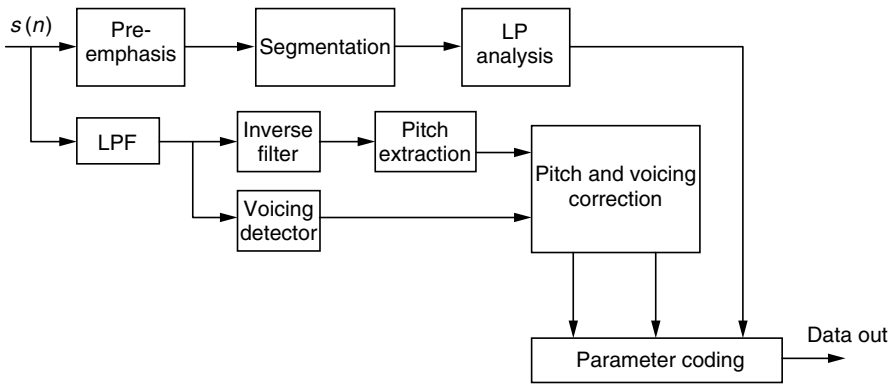


Figure 6. The LPC-10 encoder.

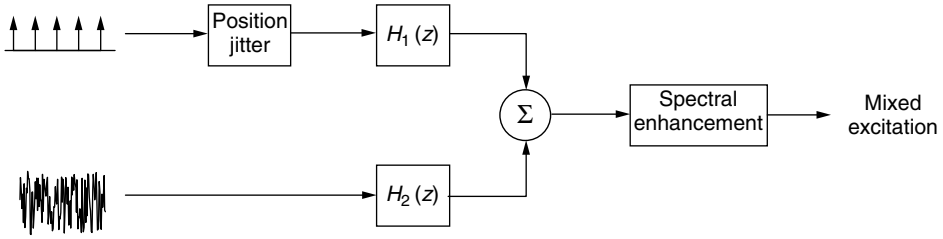


Figure 7. The mixed-excitation LPC coder.

in cases of voicing transitions (mixed voiced–unvoiced frames) [7,30]. This problem is solved using a mixed excitation scheme that combines the lowband impulse train (buzz) and highband noise (Fig. 7). The excitation shaping is done using first-order FIR filters $H_1(z)$ and $H_2(z)$ with time-varying parameters. The mixed source model also uses (selectively) pulse position jitter for the synthesis of weakly periodic or aperiodic voiced speech. An adaptive pole-zero spectral enhancer is used to boost the formant frequencies. Finally a dispersion filter is used after the LP synthesis filter to improve the matching of natural and synthetic speech away from the formants. The 2.4-kbps MELP was based on a 22.5-ms frame, and the algorithmic delay was estimated to be 122.5 ms. An integer pitch estimate is obtained open-loop by searching autocorrelation statistics followed by a fractional pitch refinement process. The LP parameters are obtained using the Levinson–Durbin algorithm and vector quantized as LSPs. MELP outperforms the LPC-10 with an estimated MOS of 3.2 but with higher complexity estimated at 40 Mips.

5. ALGORITHMS BASED ON ANALYSIS-BY-SYNTHESIS LINEAR PREDICTION

We describe here several speech coding standards based on a class of modern source-system coders where system parameters are determined by linear prediction and the excitation sequence is determined by closed-loop or analysis-by-synthesis optimization (Fig. 8). The optimization process determines an excitation sequence that minimizes the weighted difference between the input speech and synthesis speech [1–3]. Strictly speaking, this class of speech coders are not called vocoders but instead they are called *hybrid* coders. This is because closed-loop LP combines the spectral modeling properties of vocoders with the waveform-matching features of waveform coders. The system consists of a short-term LP synthesis filter, a long-term LP synthesis filter for the pitch (fine) structure of speech, a perceptual weighting filter $W(z)$ that shapes the error such that quantization noise is masked by the high-energy formants, and the excitation generator. The three most common excitation models for analysis-by-synthesis LPC are the multipulse model [2,3], the regular pulse excitation model [5], and the vector or code excitation model [1]. These excitation models are described in the context of standardized algorithms.

5.1. Long-Term Prediction (LTP)

Almost all *analysis-by-synthesis* LP algorithms include long-term prediction in addition to short-term prediction. Long term prediction, as opposed to the short-term prediction, is a process that captures the long-term correlation in the speech signal. The LTP provides a mechanism for representing the periodicity in speech and as such it represents the *fine* harmonic structure in the short-term speech spectrum. The LTP requires estimation of two parameters: a delay $a_{\leftrightarrow}$ and a parameter $a_{\leftrightarrow}$. For strongly voiced segments the delay is usually an integer that approximates the pitch period. A transfer function of a simple LTP synthesis filter is given below; more complex LTP filters involve multiple parameters and noninteger (fractional) delays [26]:

$$H_\tau(z) = \frac{1}{1 - a_\tau z^{-\tau}} \tag{5}$$

The LTP can be implemented as open loop or closed loop. The open-loop LTP parameters are typically obtained by searching the autocorrelation sequence over 128 integer delays (20–147 for speech sampled at 8 kHz). The gain is simply obtained by $a_{\leftrightarrow} = r_{ss}(\leftrightarrow)/r_{ss}(0)$. Closed-loop LTP searches produce improved speech quality at the expense of additional complexity. In closed-loop LTP search the

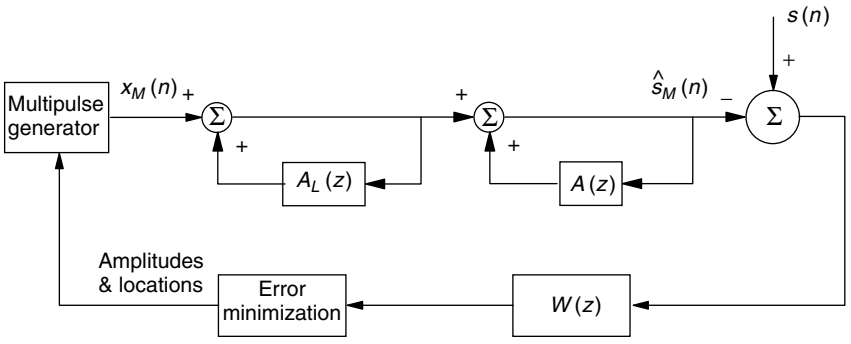


Figure 8. MPLP for the Skyphone standard.

signal is synthesized for a range of candidate LTP lags and the lag that produces the best waveform matching is chosen. Because of the intensive computations in full-search closed-loop LTP, more recent algorithms use open-loop LTP to establish an initial LTP lag, which is then refined using closed-loop search around the neighborhood of the initial estimate. In addition, to reduce complexity further LTP searches are in some cases carried every other subframe.

5.2. Multipulse Excited Linear Prediction

A 9.6-kbps multipulse excited linear prediction (MPLP) algorithm is used in *skyphone* airline applications [22]. The MPLP algorithm forms an excitation sequence that consists of multiple nonuniformly spaced pulses (Fig. 8). During analysis both the amplitude and locations of the pulses are determined (sequentially) one pulse at a time such that the weighted mean-square error is minimized. The MPLP algorithm typically uses 4–6 pulses every 5 ms [2,3]. The weighting filter is given by

$$W(z) = \frac{1 + \sum_{i=1}^{p} \gamma_1^i \, a_i \, z^{-i}}{1 + \sum_{i=1}^{p} \gamma_2^i \, a_i \, z^{-i}} \quad 0 < \gamma_2 < \gamma_1 < 1 \qquad (6)$$

The role of $W(z)$ is to deemphasize the error energy in the formant regions. This deemphasis strategy is based on the fact that in the formant regions quantization noise is partially masked by speech. Excitation coding in the MPLP algorithm is more expensive than in the classical linear predictive vocoder because MPLP encodes both the amplitudes and the locations of the pulses. The British Telecom International skyphone MPLP algorithm accommodates passenger communications in aircraft. The algorithm incorporates both short- and long-term prediction. The LP analysis window is updated every 20 ms and the LTP parameters are obtained using open-loop analysis. The MOS for the skyphone algorithm is 3.4.

5.3. The Regular Pulse Excitation (RPE) Algorithm

RPE coders also employ an excitation sequence which consists of multiple pulses. The basic difference of the RPE algorithm from the MPLP algorithm is that the pulses in the RPE coder are uniformly spaced and therefore their positions are determined by specifying the location of the first pulse within the frame and the spacing between nonzero pulses. The analysis-by-synthesis optimization in RPE algorithms represents the LP residual by a regular pulse sequence that is determined by weighted error minimization [5]. A 13-kbps coding scheme that uses RPE with long-term prediction (LTP) was adopted in 1990 for the full-rate ETSI GSM [22] Pan-European digital cellular standard. The performance of the GSM codec in terms of MOS was reported to be between 3.47 (min) and 3.9 (max), and its complexity is 5 to 6 Mips.

6. CODE-EXCITED LINEAR PREDICTION (CELP) ALGORITHMS

The vector or code-excited linear prediction (CELP) algorithm [1] (Fig. 9) encodes the excitation using vector quantization. The codebook used in a CELP coder contains vector excitation, and in each subframe a vector is chosen using an analysis-by-synthesis process. The "optimum" vector is selected such that the perceptually weighted MSE is minimized. A scaled excitation vector is filtered by the long- and short-term synthesis filters.

This category of analysis-by-synthesis LPC algorithms proved to be the most successful in terms of standards and applications. We describe in the following CELP algorithms as used in a variety of standards. We divide these algorithms in three categories that are also consistent with the chronology of their development: first-generation CELP (1986–1992), second-generation CELP (1993–1998), and third-generation CELP (1999–present).

6.1. First-Generation CELP Coders

Although the initial development of some of these algorithms was funded in part by the U.S. Department of Defense (DoD), the key driving force behind research and development of CELP coders was the demand for vocoders for the first generation digital cellular phones. Many of the first generation CELP algorithms were developed between 1986 and 1992 and work at bit rates of 4.8–16 kbps. These are generally high complexity algorithms and nontoll quality. These algorithms include, the FS-1016 CELP, the IS54 VSELP, the IS96 QCELP, and the G.728 LD-CELP.

A 4.8-kbps CELP algorithm is used by the Department of Defense for possible use in the third-generation secure telephone unit (STU-III) [31,32]. This algorithm

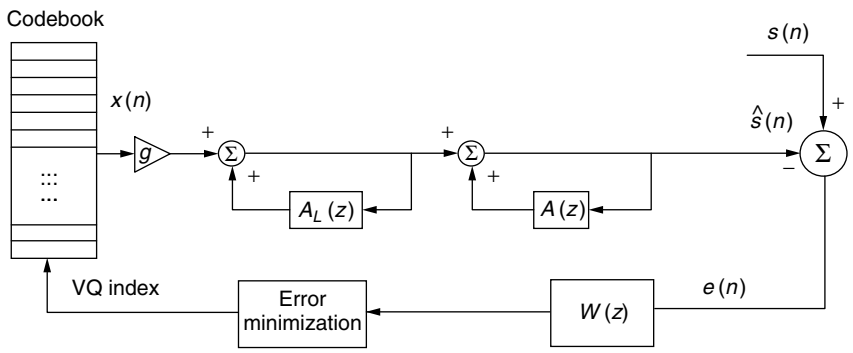


Figure 9. The analysis-by-synthesis CELP algorithm.

is described in the Federal Standard 1016 (FS-1016) and was jointly developed by the DoD and the Bell Labs. Speech in the FS-1016 CELP is sampled at 8 kHz and segmented in frames of 30 ms, and each frame is segmented in subframes of 7.5 ms. The excitation in this CELP is formed by combining vectors from an adaptive (LTP) and a stochastic codebook. The excitation vectors are selected in every subframe and the codebooks are searched sequentially starting with the adaptive codebook. The term *adaptive codebook* is used because the backward estimated LTP lag search can be viewed as an adaptive codebook search where the codebook is defined by previous excitation sequences (LTP state), and the lag τ determines the vector index. The adaptive codebook contains the history of past excitation signals and the LTP lag search is carried over 128 integer (20–147) and 128 noninteger delays. A subset of lags is searched in even subframes to reduce the computational complexity. The stochastic codebook contains 512 sparse and overlapping codevectors. Each codevector consists of sixty samples and each sample is ternary valued $(1, 0, -1)$ to allow for fast convolution. Ten short-term prediction parameters are encoded as LSPs on a frame-by-frame basis. Sub-frame LSPs are obtained by linear interpolation. The computational complexity of the FS1016 CELP was estimated at 16 MPS, and a MOS score of 3.2 has been reported.

Another standardized CELP is the IS54 VSELP, which uses highly structured codebooks that are tailored for reduced computational complexity and increased robustness to channel errors. VSELP excitation is derived by combining excitation vectors from three codebooks, namely, an adaptive codebook and two highly structured stochastic codebooks, (Fig. 10). The 128 Forty-sample vectors in each stochastic codebook are formed by linearly combining seven basis vectors. The weights used for the basis vectors are allowed to take the values of one or minus one. Hence the effect of changing one bit in the codeword, possibly due to a channel error, is not minimal since the *codevectors* corresponding to adjacent (gray codewise) codewords are different only by one basis vector. The search of the codebook is also greatly simplified because the response of the short-term synthesis filter, to codevectors from the stochastic codebook, can be formed by combining filtered basis vectors. The complexity of the 8-kbps VSELP was reported to be around 13.5 MPS, and the MOS reported was 3.45. The vector-sum-excited linear prediction (VSELP) algorithm [6] and its variants are embedded in three digital cellular standards: the 81-kbps TIA IS54 [14], the 6.3-kbps Japanese standard [33], and the 5.6-kbps half-rate GSM [34].

One problem in network applications of speech coding is that coding gain is achieved at the expense of coding delay. The one-way delay is basically the time elapsed from the instant a speech sample arrived at the encoder to the instant that this sample appears at the output of the decoder. This definition of one-way delay does not include channel- or modem-related delays. Roughly speaking, the one-way delay is generally between two and four frames. The ITU G.728 low-delay CELP coder [35,36] achieves low one-way delay by short frames, backward-adaptive predictor, and short excitation vectors (5 samples). In backward-adaptive prediction, the LP parameters are determined by operating on previously quantized speech samples that are also available at the decoder, (Fig. 11). The LD-CELP algorithm does not utilize LTP. Instead, the order of the short-term predictor is increased to 50 to compensate for the lack of a pitch loop.

The frame size in LD-CELP is 2.5 ms, and the subframes are 0.625 ms long. The parameters of the 50th-order predictor are updated every 2.5 ms. The perceptual weighting filter is based on 10th-order LP operating directly on unquantized speech and is updated every 2.5 ms. In order to limit the buffering delay in LD-CELP only 0.625 ms of speech data are buffered at a time. LD-CELP utilizes adaptive short- and long-term postfilters to emphasize the pitch and formant structures of speech. The one-way delay of the LD-CELP is less than 2 ms and MOSs as high as 3.93 and 4.1 were obtained. The speech quality of the LD-CELP was judged to be equivalent or better than the G.726 standard even after three asynchronous tandem encodings. The coder was also shown to be capable of handling voiceband modem signals at rates as high as 2400 baud (provided that perceptual weighting is not used). The coder complexity and memory requirements were found to be 10.6 MPS and 12.4 kB for the encoder and 8.06 MPS and 13.8 kB for the decoder.

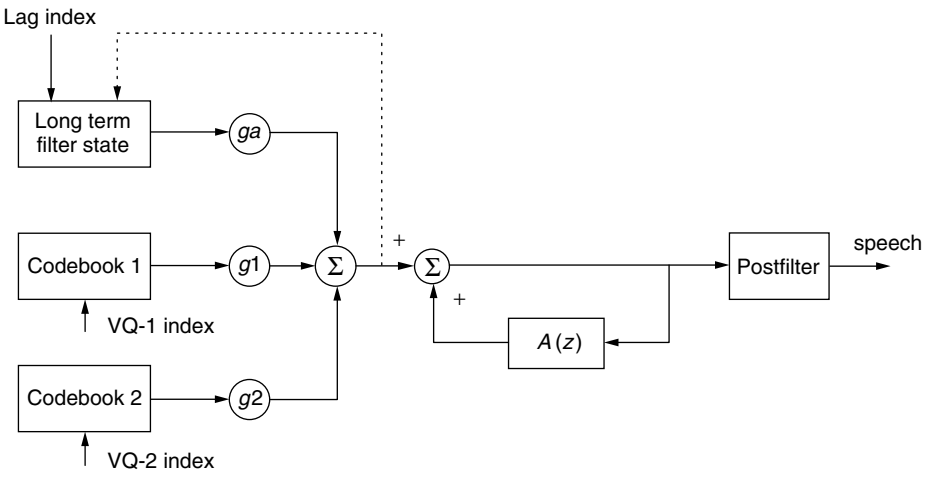


Figure 10. The IS54 VSELP algorithm.

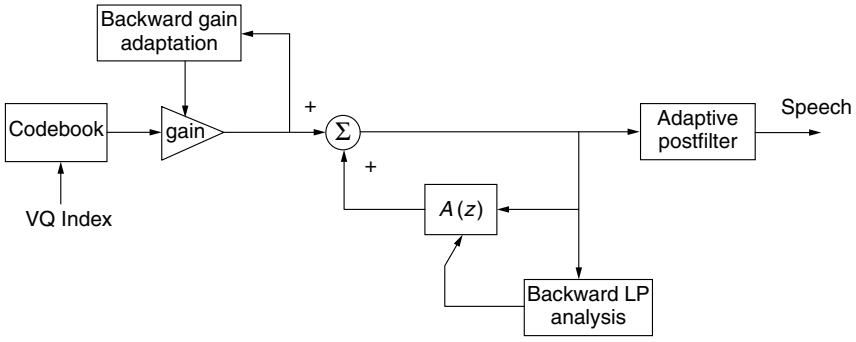


Figure 11. The G.728 low-delay CELP algorithm.

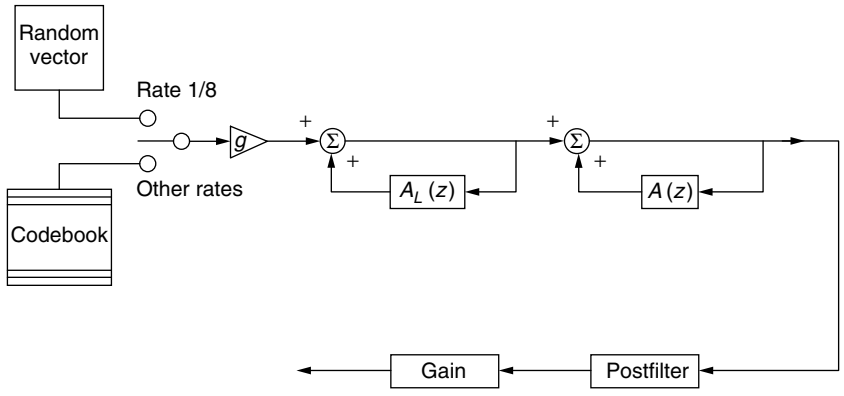


Figure 12. The IS96 QCELP decoder.

6.1.1. Variable-Rate CELP Algorithms for CDMA Applications. The IS96 QCELP [37], (Fig. 12) is a variable bit rate algorithm and is part of the code-division multiple-access (CDMA) standard for cellular communications. The bit rate is variable, with four rates supported" 9.6, 4.8, 2.4, and 1.2 kbps. The rate is determined by speech activity. Rate changes can also be activated upon command from the network. The short-term LP parameters are encoded as LSPs.

Lower rates are achieved by allocating fewer bits to LP parameters and by reducing the number of updates of the LTP and random codebook parameters. At 1.2 kbps (rate $\frac{1}{8}$) the algorithm essentially encodes comfort noise. The MOS for QCELP at 9.6 kbps is 3.33, and the complexity is estimated to be around 15 MPS.

6.2. Second-Generation Near-Toll-Quality CELP Algorithms

The second-generation CELP algorithms were developed between 1993 and 1998 for applications in second-generation cellular phones, PC Internet streaming applications, Voice over Internet Protocol (VoIP), and secure communications. These are also high-complexity algorithms that deliver near-toll-quality speech, with some of the most successful ones certified for toll quality. The speech quality enhancement with these coders is because of the improvement in the coding of excitation, which is done by algebraic codebooks [13], as well as due to the coding of the LPC parameters in using line spectral frequencies and perceptually optimized split-vector quantization. Furthermore interpolative techniques with the LTP (relaxed CELP) [4] also contribute to improvements. Algorithms in this category include the G.729 ACELP, the G.723.1 coder, the GSM EFR, the IS127 RCELP.

6.2.1. CELP Algorithms with Algebraic Codebooks. A low-delay 8 kbps conjugate structure *algebraic CELP* (CS-ACELP) algorithm has been adopted as ITU recommendation G.729 [13]. The G.729 is designed for both wireless and multimedia network applications. CS-ACELP is a low-delay algorithm with a frame size of 10 ms, a lookahead of 5ms, and a total algorithmic delay of 15 ms. The algorithm is based on an analysis-by-synthesis CELP scheme and uses two codebooks for excitation modeling. The short-term prediction parameters are obtained every 10 ms and vector-quantized as LSPs. The algorithm uses an algebraically structured fixed codebook that does not require storage. Each 40-sample (5-ms) codevector contains four nonzero, binary-valued $(-1, 1)$ pulses. Pulses are interleaved and position-encoded, which allows efficient search. Gains for the fixed and adaptive codebooks are jointly vector-quantized in a two-stage, two-dimensional conjugate structured codebook. Search efficiency is enhanced by a preselection process that constrains the exhaustive search to 32 of a possible 128 codevectors. The algorithm comes in two versions: the original G.729 (20 MIPS) and the less complex G.729 Annex A (11 MIPS). The algorithms are interoperable and the lower complexity algorithm has slightly lower quality. The MOS for the G.729 is 4.1 and for the G.729A 3.76. G.729 Annex B defines a silence compression algorithm allowing either the G.729 or the G.729A to operate at lower rates, thereby making them particularly useful in digital simultaneous voice and data (DSVD) applications. Also extensions to G.729 at 6.4 and 12 kbps are planned.

6.2.2. CELP Algorithms for PC Videoconferencing and Voice-over-IP Applications. The ITU G.723.1 [38] is a dual-rate speech coder algorithm intended for audio/videoconferencing/telephony over public phone (POTS) networks. G.723.1 is part of the ITU H.323 and H.324 audio/videoconferencing standards. The standard is dual rate 6.3 and 5.3 kbps. The excitation is selected using an analysis-by-synthesis process, and two excitation schemes are defined: the multipulse maximum-likelihood quantization (MP-MLQ) for the 6.3-kbps mode and the ACELP for 5.3 kbps. Ten short-term LP parameters are computed and vector-quantized as LSPs. A fifth-order LTP is used in this standard, and the LTP lag is determined using a closed-loop process searching around a previously obtained open-loop estimate. The LTP gains are vector-quantized. The high-rate MP-MLQ excitation involves matching of the LP residual with a set of impulses with restricted positions. The lower-rate ACELP excitation is similar but not identical to the excitation scheme used in G.729. G.723.1 provides a toll-quality MOS of 3.98 at 6.3 kbps and has a frame size of 30 ms with a lookahead of 7.5 ms. The estimated one-way delay is 37.5 mS. An option for variable-rate operation using a voice activity detector (silence compression) is also available. The Voice over IP (VoIP) forum, that is part of The International Multimedia Teleconferencing Consortium (IMTC), recommended G.723.1 to be the default audio codec for voice of the network (decision pending).

6.2.3. The ETSI GSM 6.60 Enhanced Full-Rate Standard. The enhanced full-rate (EFR) encoder was developed for use in the full-rate GSM standard [23]. The EFR is a 12.2-kbps algorithm with a frame of 20 ms and a 5-ms lookahead. A tenth-order short-term predictor is used, and its parameters are transformed to LSPs and encoded using split-vector quantization. The LTP lag is determined using a two-stage process where open-loop search provides an initial estimate which is then refined by closed-loop search around the neighborhood of the initial estimate. An algebraic codebook similar to that of the G.729 ACELP is also used. The algorithmic delay for the EFR is 25 ms and the MOS estimated to be around 4.1. The standard has provisions for a voice activity detector and an elaborate error protection scheme is also part of the standard. The North American PCS 1900 standard uses the GSM infrastructure and hence the EFR.

6.2.4. The Use of the EFR Algorithm in the IS641 TDMA Cellular/PCS Standard. The IS641 [25] is a 7.4-kbps EFR algorithm, also known as the Nokia/USH algorithm, and it is a variant of the GSM EFR. IS641 is the speech coding standard that is embedded in the IS136 Digital-AMPS (D-AMPS) North American digital cellular standard. EFR offers improved quality for this service relative to IS54. The algorithm is based on a 20-ms frame and 10th-order LP whose parameters are split-vector-quantized as LSPs. An algebraic codebook (ACELP) is used and the LTP lag is determined using open-loop search followed by closed-loop refinement with a fractional pitch resolution. The complexity of this coder is estimated at 14 MPS, and the Mean Opinion Score is estimated at 3.8.

6.2.5. The Relaxed CELP for the IS127 Enhanced Variable-Rate Coder (EVRC) for CDMA. The IS127 enhanced variable-rate coder (EVRC) [24] is based on relaxed CELP (RCELP), which uses interpolative coding methods [4] as a means for reducing further the bit rate and complexity in analysis-by-synthesis linear predictive coders. The EVRC encodes the RCELP parameters using a variable-rate approach. There are three possible bit rates for EVRC — 8, 4, and 0.8 kbps — or after error protection, 9.6, 4.8, and 1.2 kbps, respectively. The rate is determined using a voice activity detection algorithm that is embedded in the standard. Rate changes can also be initiated on command from the network. At the lowest rate (0.8 kbps) the algorithm does not encode excitation information; hence the decoder essentially produces comfort noise. The other rates are achieved by changing the number of bits allotted to LP and excitation parameters. The algorithm uses a 20-ms frame, and each frame is divided in three 6.75-ms subframes. Tenth-order short-term LP parameters are obtained using the Levinson–Durbin algorithm and split-vector encoded as LSPs. The fixed codebook structure is ACELP. The LTP parameters are estimated using generalized analysis-by-synthesis where instead of matching the input speech, a down-sampled version of a modified LP residual that conforms to a pitch contour is matched. The pitch contour is established using interpolative methods. The standard also specifies an FFT-based speech enhancement pre-processor that is intended to remove background noise from speech. The MOS for EVRC is 3.8 at 9.6 bps, and the algorithmic delay is estimated to be 25 ms.

6.2.6. The Japanese PDC Full-Rate and Half-Rate Standards. The Japanese Research and Development Center for Radio Systems (RCR) has adopted two algorithms for the personal digital cellular (PDC) full-rate (6.3 kbps) and the PDC half-rate (3.45 kbps) standards. The full-rate algorithm [33] is a variant of the IS54 VSELP algorithm described in a previous section. The half-rate PDC coder is a high-complexity pitch-synchronous innovation CELP (PSI-CELP) [39]. As the name implies, the codebooks of PSI-CELP depend on the pitch period. PSI-CELP defaults to periodic vectors if the pitch period is less than the frame size. The complexity of the algorithm is about 50 MIPS and the algorithmic delay is about 50 ms.

6.3. Third-Generation CELP for 3G Cellular Standards

The effort to establish wideband wireless cellular standards have driven further research and development toward algorithms that work at multiple rates and deliver significantly enhanced speech quality. These third-generation algorithms are multimodal as they accommodate several different bit rates. This is consistent with the vision on wideband wireless standards [44] that will operate in different modes, with low mobility, high mobility, indoors, and so on. At least two algorithms have been developed and standardized for these applications. In Europe GSM is looking at the adaptive multirate coder [42,43], and in the United States the TIA has tested the *selectable mode vocoder* (SMV) [45,52] developed by Connexant.

6.3.1. The Adaptive Multirate (AMR) Coder for GSM. An adaptive GSM multirate coder [42,43] has been adopted by ETSI for use in WCDMA. This is an ACELP algorithm that operates at multiple rates: 12.2, 10.2, 7.95, 6.7, 5.9, 5.15, and 4.75 kbps. The bit rate is adjusted according to the traffic conditions. The AMR is based on ACELP with 20-ms frame and 5-ms subframes. It uses a 10th-order short-term LPC and encodes LSPs using split vector quantization. At the highest bit rate it provides toll quality, and at the half rate it provides communications quality.

6.3.2. The Selectable Mode Vocoder. The SMV algorithm was developed to provide higher quality, flexibility, and capacity over the existing IS96C and IS127 EVRC CDMA algorithms. The SMV is based on 4 codecs: full rate at 8.5 kbps, half rate at 4 kbps, quarter rate at 2 kbps, and eighth rate at 800 bps. The full rate and half rate are based on the eXtended CELP (eX-CELP) algorithm, which is based on a combined closed-loop/open-loop-analysis (COLA). In eX-CELP the signal frames are first classified as silence/background noise, nonstationary unvoiced, stationary unvoiced, onset, nonstationary voiced, and stationary voiced. The algorithm includes voice activity detection (VAD) followed by an elaborate frame classification scheme. Silence/background noise and stationary unvoiced frames are represented by spectrum modulated noise and coded at rate $\frac{1}{4}$ or $\frac{1}{8}$. The SMV uses four subframes for full rate and three subframes for half rate. The stochastic (fixed) codebook structure is also elaborate and uses subcodebooks, each tuned for a particular type of speech. The subcodebooks have different degrees of pulse sparseness (more sparse for noise like excitation). SMV scored as high as 4.1 MOS at full rate with clean speech.

7. SUMMARY

In this article, we provided an overview of some of the current linear predictive vocoder methods for speech and audio coding. Speech coding research has come a long way since the early 1990s, and several algorithms are rapidly finding their way in consumer products ranging from wireless cellular telephones to computer multimedia systems. Research and development in code excited linear prediction yielded several algorithms that have been adopted, or are strong candidates for adoption, in international wired and wireless communication standards from 16 down to 4.8 kbps. On the other hand, research activities are concentrating on the development of toll-quality algorithms that operate under 4 kbps. ITU has called for proposals for toll quality at 4 kbps for use in future videophones, personal communications, and third-generation cellular systems. Several proposals have been submitted [45–52] with none selected as of yet. Our coverage focussed on vocoder methods based on linear prediction because they have been the most popular in more recent low-rate communications and media standards. We must mention that in addition to the linear predictive coders, some frequency-domain methodologies have also been used in low-rate speech coding. In particular, sinusoidal analysis–synthesis schemes have been used as alternative to LPC. The sinusoidal model [8] represents speech by a linear combination of sinusoids. A low-rate sinusoidal transform coder (STC) has been considered in federal and TIA standardization competitions and is currently considered for scalable wideband and high-fidelity applications. Another sinusoidal system, the *improved multiband excitation* (IMBE) coder [9,10], became part of the Australian (AUSSAT) mobile satellite standard and the International Mobile Satellite (INMARSAT-M) standard.

BIOGRAPHY

Andreas Spanias is professor in the Department of Electrical Engineering at Arizona State University (ASU). His research interests are in the areas of adaptive signal processing and speech/audio processing. While at ASU, he has developed and taught courses in DSP, adaptive signal processing, and speech coding. Andreas Spanias has received research contracts and grants from the National Science Foundation, Intel Corporation, Sandia National Labs, Motorola Inc., and Active Noise and vibration Technologies. He has also consulted with Inter-Tel Communications, Texas Instruments, and the Cyprus Institue of Neurology and Genetics. He authored the J-DSP educational software package (ISBN 0-9724984-0-0) that is used for on-line labs in DSP. He is a Senior Member of the IEEE and has served as a member on two IEEE technical comittees of the IEEE Signal Processing Society (SPS). He has also served as associate editor of the *IEEE Transactions on Signal Processing* and the *IEEE Signal Processing Letters*. He co-chaired the 1999 International Conference on Acoustics Speech and Signal Processing (ICASSP-99) in Phoenix, and he currently is the IEEE SPS Vice-President for Conferences. Andreas Spanias is the co-recipient of the 2002 IEEE Donald G. Fink prize paper award for the IEEE Proceedings manuscript entitled "Perceptual Coding of Digital Audio."

BIBLIOGRAPHY

1. M. R. Schroeder and B. Atal, Code-excited linear prediction (CELP): High quality speech at very low bit rates, *Proc. ICASSP'85*, Tampa, April 1985, p. 937.
2. S. Singhal and B. Atal, Improving the performance of multipulse coders at low bit rates, *Proc. ICASSP'84*, 1984, p. 1.3.1.
3. B. Atal and J. Remde, A new model for LPC excitation for producing natural sounding speech at low bit rates, *Proc. ICASSP'82*, April 1982, pp. 614–617.
4. W. B. Kleijn et al., Generalized analysis-by-synthesis coding and its application to pitch prediction, *Proc. ICASSP'92*, 1992, pp. 1337–1340.
5. P. Kroon, E. Deprettere, and R. J. Sluyeter, Regular-pulse excitation — a novel approach to effective and efficient multipulse coding of speech, *IEEE Trans.* **ASSP-34**(5) (Oct. 1986).
6. I. Gerson and M. Jasiuk, Vector sum excited linear prediction (VSELP) speech coding at 8 kbits/s, *Proc. ICASSP'90*, New Mexico, April 1990, pp. 461–464.
7. A. McCree and T. Barnwell III, A new mixed excitation LPC vocoder, *Proc. ICASSP'91*, Toronto, May 1991, pp. 593–596.
8. R. McAulay and T. Quatieri, Low-rate speech coding based on the sinusoidal model, in S. Furui and M. M. Sondhi, eds., *Advances in Speech Signal Processing*, Marcel Dekker, New York, 1992, Chap. 6, pp. 165–207.
9. D. Griffin and J. Lim, Multiband excitation vocoder, *IEEE Trans.* **ASSP-36**(8): 1223 (Aug. 1988).

10. J. Hardwick and J. Lim, The application of the IMBE speech coder to mobile communications, *Proc. ICASSP'91*, May 1991, pp. 249–252.

11. P. Noll, Digital audio coding for visual communications, *Proc. IEEE* **83**(6): 925–943 (June 1995).

12. T. Painter and A. Spanias, Perceptual coding of digital audio, *Proc. IEEE* **88**(4): 451–513 (April 2000).

13. ITU Study Group 15 Draft Recommendation G.729, *Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction* (CS-ACELP), International Telecommunication Union, 1995.

14. TIA/EIA-PN 2398 (IS54), *The 8 kbit/s VSELP Algorithm*, 1989.

15. GSM 06.10, *GSM Full-Rate Transcoding*, Technical Report, Version 3.2, ETSI/GSM, July 1989.

16. Federal Standard 1015, *Telecommunications: Analog to Digital Conversion of Radio Voice by 2400 Bit/Second Linear Predictive Coding*, National Communication System — Office Technology and Standards, Nov. 1984.

17. J. Makhoul, Linear prediction: A tutorial review, *Proc. IEEE* **63**(4): 561–580 (April 1975).

18. J. Markel and A. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

19. N. Benevuto et al., The 32Kb/s coding standard, *AT&T Tech. J.* **65**(5): 12–22 (Sept.–Oct. 1986).

20. ITU Recommendation G.726 (formerly G.721), *24, 32, 40 kb/s Adaptive Differential Pulse Code Modulation* (*ADPCM*), Blue Book, Vol. III, Fascicle III.3, Oct. 1988.

21. A. Spanias, Speech coding: A tutorial review, *Proc. IEEE* **82**(10): 1541–1582 (Oct. 1994).

22. I. Boyd and C. Southcott, A speech codec for the skyphone service, *Br. Telecommun. Techn. J.* **6**(2): 51–55 (April 1988).

23. GSM 06.60, *GSM Digital Cellular Communication Standards: Enhanced Full-Rate Transcoding*, ETSI/GSM, 1996.

24. TIA/EIA/IS-127, *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, TIA, 1997.

25. TIA/EIA/IS-641, *Cellular/PCS Radio Interface — Enhanced Full-Rate Speech Codec*, TIA, 1996.

26. P. Kroon and B. Atal, Pitch predictors with high temporal resolution, *Proc. ICASSP'90*, New Mexico, April 1990, pp. 661–664.

27. INMARSAT-B SDM Module 1, Appendix I, Attachment 1, MAC/SDM/BMOD1/ATTACH/ISSUE 3.0.

28. J. Campbell and T. E. Tremain, Voiced/unvoiced classification of speech with applications of the U.S. government LPC-10e algorithm, *Proc. ICASSP'86*, Tokyo, 1986, pp. 473–476.

29. T. E. Tremain, The government standard linear predictive coding algorithm: LPC-10, *Speech Technol.* 40–49 (April 1982).

30. J. Makhoul et al., A mixed-source model for speech compression and synthesis, *J. Acous. Soc. Am.* **64**: 1577–1581 (Dec. 1978).

31. J. Campbell, T. E. Tremain, and V. Welch, The proposed federal standard 1016 4800 bps voice coder: CELP, *Speech Technol.* 58–64 (April 1990).

32. Federal Standard 1016, *Telecommunications: Analog to Digital Conversion of Radio Voice by 4800 Bit/Second Code Excited Linear Prediction* (*CELP*), National Communication System — Office Technology and Standards, Feb. 1991.

33. I. Gerson, Vector sum excited linear prediction (VSELP) speech coding for Japan digital cellular, paper presented at meeting of IEICE, RCS90-26, Nov. 1990.

34. GSM 06.20, *GSM Digital Cellular Communication Standards: Half Rate Speech; Half Rate Speech Transcoding*, ETSI/GSM, 1996.

35. ITU Draft Recommendation G.728, *Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction* (*LD-CELP*), 1992.

36. J. Chen et al., A low-delay CELP coder for the CCITT 16 kB/s speech coding standard, *IEEE Trans. Selec. Areas Commun.* (Special Issue on Speech and Image Coding; N. Hubing, ed.) 830–849 (June 1992).

37. TIA/EIA/IS-96, *QCELP, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, TIA 1992.

38. ITU Recommendation G.723.1, *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, Draft 1995.

39. T. Ohya, H. Suda, and T. Miki, The 5.6 kb/s PSI-CELP of the half-rate PDC speech coding standard, *Proc. IEEE Vehicular Technology Conf.*, 1993, pp. 1680–1684.

40. ITU Recommendation G.722, *7 KHz Audio Coding within 64 kbits/s*, Blue Book, Vol. III, Fascicle III, Oct. 1988.

41. INMARSAT Satellite Communications Services, *INMARSAT-M System Definition*, Issue 3.0 — Module 1: *System Description*, Nov. 1991.

42. ETSI AMR *Qualification Phase Documentation*, 1998.

43. R. Ekudden, R. Hagen, I. Johansson, and J. Svedburg, The adaptive multi-rate speech coder, *Proc. IEEE Workshop on Speech Coding*, 1999, pp. 117–119.

44. D. N. Knisely, S. Kumar, S. Laha, and S. Navda, Evolution of wireless data services: IS-95 to cdma2000, *IEEE Commun. Mag.* 140–146 (Oct. 1998).

45. Conexant Systems, *Conexant's ITU-T 4 kbit/s Deliverables*, ITU-T Q21/SG16 Rapportuer meeting, AC-99-20, Sept. 1999.

46. Matsushita Electric Industrial Co. Ltd., *High Level Description of Matsushita's 4-kbit/s Speech Coder*, ITU-T Q21/SG16 Rapportuer meeting, AC-99-19, Sept. 1999.

47. Mitsubhishi Electric Corp., *High Level Description of Mitsubhishi 4 kb/s Speech Coder*, ITU-T Q21/SG16 Rapportuer meeting, AC-99-016, Sept. 1999.

48. NTT, *High Level Description of NTT 4 kb/s Speech Coder*, ITU-T Q21/SG16 Rapportuer meeting, AC-99-17, Sept. 1999.

49. Rapporteur (Mr. Paul Barrett, BT/UK), *Q.21/16 Meeting Report*, ITU-T Q21/SG16 Rapportuer meeting, Temporary Document E (3/16), Geneva, Feb. 7–18, 2000.

50. Texas Instruments, *High Level Description of TI's 4 kb/s Coder*, ITU-T Q21/SG16 Rapportuer meeting, AC-99-25, Sept. 1999.

51. Toshiba Corp., *Toshiba Codec Description and Deliverables* (*4 kbit/s Codec*), ITU-T Q21/SG16 Rapportuer meeting, AC-99-15, Sept. 1999.

52. Y. Gao et al., The SMV algorithm selected for TIA and 3GPP2 for CDMA applications, *ICASSP'2001*, Salt Lake City, May 5–12, 2002, Vol. 2.

W

WAVEFORM CODING

GÜNEŞ KARABULUT
ABBAS YONGAÇOĞLU
University of Ottawa
School of Information
Technology and Engineering
Ottawa, Ontario, Canada

1. INTRODUCTION

A signal is an entity that changes with time and contains some information. If a signal is continuous in both time and amplitude, it is referred to as an *analog signal* (*waveform*) and can be represented by a physical quantity (e.g., voltage, current, pressure) proportional to it. If it is discrete both in time and amplitude, then it is referred to as a *digital signal* and can be represented in bits. Conversion from one form to the other is performed whenever necessary. For example human speech is an analog signal but what we hear from compact disks are digital signals since they are stored and processed as ones and zeros.

Digital representation of analog signals offers many advantages. Digital signals are less sensitive to various transmission impairments and noise, easier to store and regenerate, and suitable for encryption for greater security. It is easy to multiplex various forms of digital information. Digital signals also enable unification of transmission and switching functions in communications and permit error protection.

Waveform coding is the process of describing analog signals in a digital form. It can also be viewed as the process of associating a mathematical representation with a waveform. The major goal of waveform coding is to obtain the minimum data rate for a given amount of distortion; or conversely, to represent a waveform at a given data rate while causing the minimum amount of distortion.

The area of waveform coding has been very active since the early 1940s, and it is still going strong with the spread of digitization. In this article we will first discuss the basic principles of sampling and quantization, and will then present temporal and spectral waveform coding techniques.

2. DIGITIZATION

A source generates the information to be transmitted. If a source is producing waveforms, then it is an analog source, and the source encoder converts those waveforms to a digital form suitable for transmission or storage. A waveform encoder follows an analog source, and its location in a generic digital communication system is shown in Fig. 1. If the source is already producing digital symbols, then the source encoder compresses the information by taking advantage of its statistical properties.

The source decoder tries to reconstruct the input waveform from the received digital data. The semblance of the input signal to its received estimate obtained from the source decoder determines the performance of the waveform coding system. This performance depends mainly on two factors: quantization noise and channel impairments. *Quantization noise* is introduced in the waveform encoder, and it is the result of representing infinite number of amplitudes an input signal can take by using only a finite number of amplitudes. The degradation due to quantization is measured with a quantity called signal-to-quantization noise ratio (SQNR), as will be defined in later sections. SQNR depends on the input waveform properties and the specific source encoding technique employed.

Channel impairments are experienced while modulated waveform is transmitted through a noisy communication channel. A topic that has gained popularity is *channel optimized coding*, which factors in the noisy nature of the channel in the source coding process [1].

Waveforms are continuous in amplitude and time. In order to produce their digital equivalents, we need to make them discrete in both amplitude and time. The process of time discretization (for two-dimensional signals this corresponds to space discretization) is called *sampling*, and the process of amplitude discretization is called *quantization*. The circuit that performs sampling and quantization is often referred to as an *analog-to-digital converter* (ADC). Block diagram of an ADC is shown in Fig. 2. The prefilter shown is added to bandlimit input waveforms as will be explained later. A waveform $x(t)$, and the corresponding prefilter, sampler, and quantizer outputs are shown in Fig. 3. When sampled at or above the Nyquist rate (defined in the next section), the original analog signal can always be fully recovered from its samples.

2.1. Sampling and Reconstruction

2.1.1. Sampling. Sampling is the process of converting a waveform to a series of samples in time. A sample

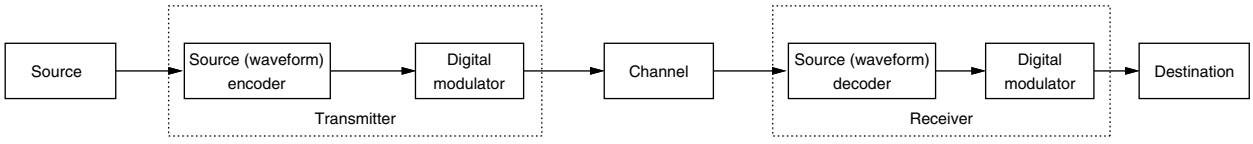


Figure 1. A digital communication system.

Figure 2. Analog-to-digital converter (ADC).

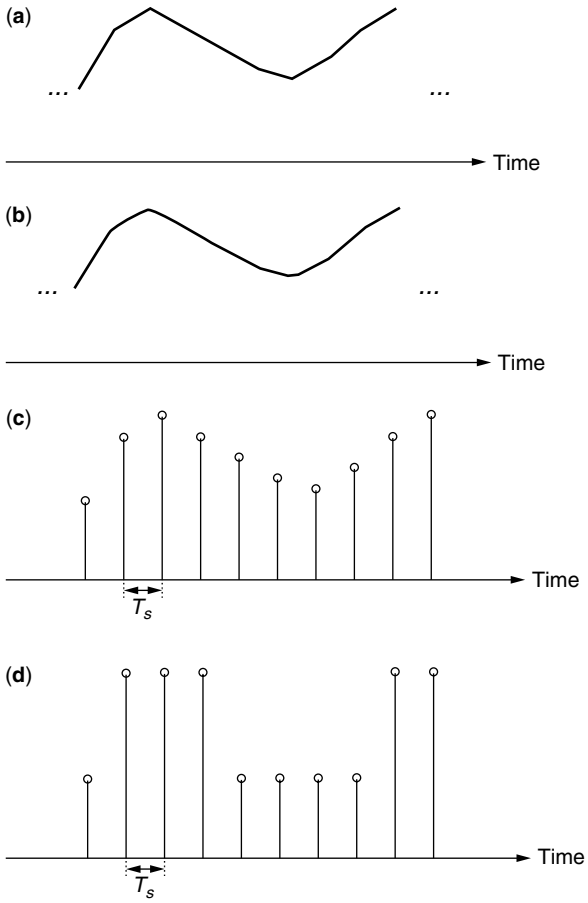


Figure 3. Signal samples of ADC: (**a**) input signal, $x(t)$ amplitude–time plot; (**b**) prefilter output that band-limits the input signal; (**c**) sampler output $x_\delta(t)$; (**d**) quantizer output, x.

is a measure of amplitude of a waveform evaluated over a period of time. When this period is constant for each sample (i.e., samples are equally spaced in time), the process is called *uniform sampling*. The period is termed the sampling period or sampling interval, T_s. The reciprocal of T_s is called the sampling rate or sampling frequency and denoted by f_s.

In the following analysis uniform sampling will be considered because of its widespread usage and convenience. To explain the sampling process, we must first introduce the concept of a band-limited waveform.

2.1.2. Band-Limited Waveform. Let $x(t)$ be a real valued waveform with finite energy

$$\int_{-\infty}^{+\infty} |x(t)|^2\, dt < \infty, \tag{1}$$

so that $x(t)$ is Fourier transformable. Denoting the Fourier transform of $x(t)$ as $X(f)$, the Fourier transform pair equations are given by

$$X(f) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft}\, dt \Leftrightarrow x(t) = \int_{-\infty}^{+\infty} X(f)e^{j2\pi ft}\, df \tag{2}$$

A waveform band-limited to W hertz or $\Omega_W = 2\pi W$ radians per second (rad/s) is defined as

$$X(f) = 0; \quad |f| \geq W = \frac{\Omega_W}{2\pi} \tag{3}$$

2.1.3. Sampling Theorem. Let the finite energy signal $x(t)$ be also band-limited, and let $x_\delta(t)$ denote the sampled version of $x(t)$. Defining the Dirac delta function $\delta(t)$ as $\delta(t) = 0$, for $t \neq 0$ and $\int_{-\infty}^{+\infty} \delta(t)dt = 1$, the relationship between $x(t)$ and $x_\delta(t)$ is given by

$$x_\delta(t) = \sum_{k=-\infty}^{+\infty} x(t)\delta(t - kT_s) \tag{4}$$

where $\delta(t - kT_s)$ is the Dirac delta function positioned at $t = kT_s$. Through the sifting property of Dirac delta function we can modify Eq. (4) as

$$x_\delta(t) = \sum_{k=-\infty}^{+\infty} x(kT_s)\delta(t - kT_s) \tag{5}$$

This equation represents a modified impulse train with weights of $x(kT_s)$ at $t = kT_s$. Therefore uniform sampling can be visualized as the multiplication of $x(t)$ by an infinitely long train of impulse functions that are T_s seconds apart.

The frequency domain view of the sampled signal becomes

$$X_\delta(f) = \sum_{k=-\infty}^{+\infty} x(kT_s)e^{-j2\pi kfT_s} \tag{6}$$

where we can see that the uniform sampling process results in a periodic spectrum with a period of $1/T_s$. If we choose the sampling rate as $T_s = 1/2W$, which is named as the Nyquist rate, we obtain the spectrum as

$$X_\delta(f) = \sum_{k=-\infty}^{+\infty} x\left(\frac{k}{2W}\right) e^{-j\pi kf/W} \tag{7}$$

If we choose a sampling period larger than $1/2W$ (i.e., $f_s < 2W$), we would observe the effect of aliasing. In this case the sampler output is referred to as *undersampled*, and the original waveform $x(t)$ cannot be recovered from $x_\delta(t)$. If we choose a sampling period smaller than $1/2W$ (i.e., $f_s > 2W$), the original waveform $x(t)$ can be

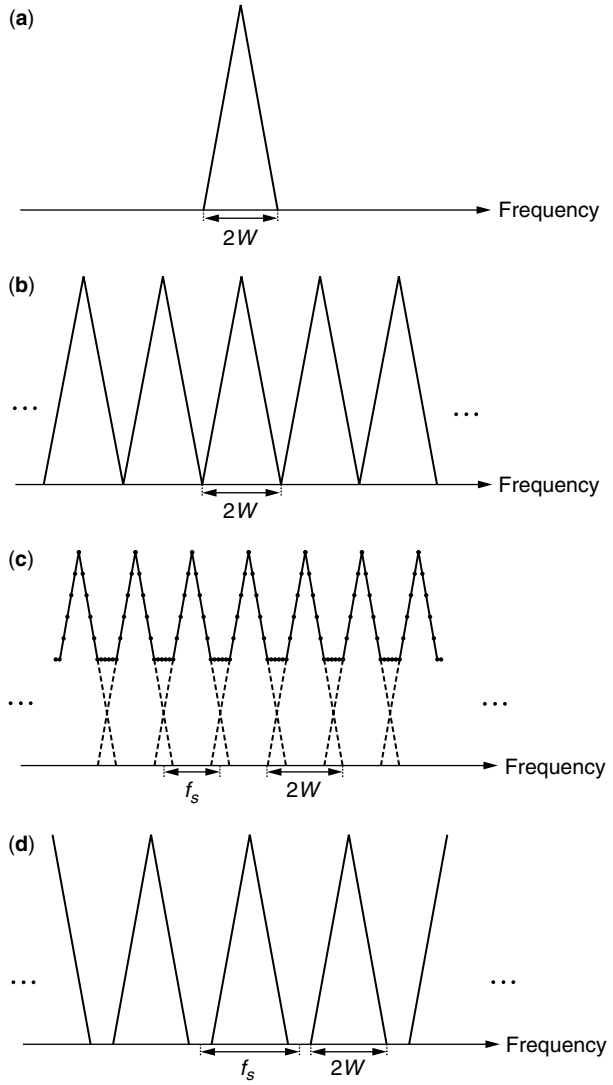


Figure 4. Sampling process in frequency domain: (**a**) frequency spectrum of $x(t)$, $X(f)$; (**b**) sampled at Nyquist rate; (**c**) undersampled, aliasing present; (**d**) oversampled.

recovered, but we will observe bandwidth expansion due to oversampling.

In Fig. 4, the effects of the sampling period in the frequency domain are shown. Frequency spectrum of $x(t)$, $X(f)$ is shown in Fig. 4a. The spectrum shown in Fig. 4b is sampled at the Nyquist rate. In Fig. 4c the signal is undersampled and we can see the effect of aliasing, and in Fig. 4d the signal is oversampled.

The fourier transform of $x(t)$ can also be expressed as

$$X_\delta(f) = f_s \sum_{k=-\infty}^{+\infty} X(f - kf_s) = f_s X(f) + \sum_{\substack{k=-\infty \\ k\neq 0}}^{+\infty} X(f - kf_s) \quad (8)$$

Hence

$$X_\delta(f) = f_s X(f), \quad -W < f < W \quad (9)$$

The reconstruction method of $x(t)$ from this equation will be explained in the following section.

2.1.4. Reconstruction. To recover $x(t)$ from $x_\delta(t)$, we need to use a reconstruction filter, which has the following properties:

$$H_R(f) = \frac{1}{f_s} = T_s \quad |f| \leq W$$
$$H_R(f) = 0 \qquad |f| > W \quad (10)$$

The corresponding impulse response of the reconstruction filter is

$$h_R(t) = \frac{\sin(\pi t/T_s)}{\pi t/T_s} = \text{sinc}\left(\frac{\pi t}{T_s}\right) \quad (11)$$

The multiplication of $X_\delta(f)$ by $H_R(f)$, is equivalent to convolution of $x_\delta(t)$ with $h_R(t)$ in the time domain. Thus, we obtain the interpolation formula as

$$x_\delta(t)^* h_R(t) = \sum_{k=-\infty}^{+\infty} x(kT_s)\,\text{sinc}\left(\frac{t}{T_s} - k\right) = x(t) \quad (12)$$

where * denotes the convolution operation.

When $x(t)$ is band-limited and the sampling rate exceeds $2W$, the sampling and reconstruction operations are error-free. In practice, in order to recover the input waveform exactly from its samples, the waveform must be band-limited. To render $x(t)$ band-limited, a *prefilter* (also known as the *antialiasing filter*) can be inserted before the sampler as shown in Fig. 2. The aim of this filter is to prevent or minimize the effects of aliasing. The distortion caused by the prefilter with an appropriately chosen bandwidth is less objectionable than the effect of aliasing. For band-limited signals, prefiltering operation does not have a distorting effect. The input signal $x(t)$ and its prefiltered version are shown in Figs. 3a and 3b, respectively. The sampled version of prefiltered signal, $x_\delta(t)$, is shown in Fig. 3c. The quantized version of $x_\delta(t)$, $\mathbf{x}$, is shown in Fig. 3d.

Another precaution to prevent aliasing and to fully recover the original signal is sampling at a frequency slightly higher than the Nyquist frequency of $2W$.

2.2. Quantization

The continuous amplitude of a sample can take on a value from an infinitely large set. However, to represent this value digitally, a finite set of discrete amplitude values are used. This process is called *quantization*. The simplest form of quantization is rounding off a number. Unlike sampling, the quantization is an irreversible process; that is, from a quantized value we cannot go back to the original value. Hence it can be viewed as a type of lossy data compression.

Two major parameters in the quantization process are the data rate and distortion. The unit of data rate can be bits per second (bps) or bits per sample. To save transmission and/or storage resources, quantization at the minimal rate is highly desirable. However, rate and distortion are limited by Shannon's rate distortion theory [2]. Hence the other view of the rate distortion tradeoff is to achieve the minimum distortion for a given rate. Distortion can be defined as a subjective

quantity, involving the ratings that are given by experts such as mean opinion score, or it can be an objective quantity defined in mathematical terms. An objective distortion definition resulting from the quantization of signal amplitudes is

$$D = \int_{-\infty}^{+\infty} F[x_\delta(t) - \mathbf{x}]\, dx \tag{13}$$

where $F[\cdot]$, denotes the desired error function.

A quantizer can be optimized according to the probability density function (pdf) of the input; therefore it is waveform specific.

The difference between the input and the output of a quantization process is defined as the *quantization error*, $e(t)$, and is given by

$$e(t) = x_\delta(t) - \mathbf{x} \tag{14}$$

Performance of a quantizer is often measured by the *signal-to-quantization noise ratio* (SQNR), which is defined as

$$\text{SQNR} = \frac{P_x}{\sigma_e^2} \tag{15}$$

where P_x denotes the input signal power and σ_e^2 denotes the variance of the quantization error, $e(t)$.

Two major quantization techniques are scalar quantization and vector quantization.

2.2.1. Scalar Quantization (SQ). *Scalar quantization* (SQ) is the mapping of a sample to the nearest quantization level. *Quantization interval* is the range of amplitudes that are mapped to the same quantization level. The difference between two quantization levels is referred to as the step size. These concepts are shown in Fig. 5. A quantization level should be chosen for each quantization interval. In an n-bit scalar quantizer each sample is represented by n bits, and there can be 2^n different quantization levels.

2.2.1.1. Uniform Quantization. In *uniform quantization*, the quantization intervals are of equal length, and each quantization level is chosen as the midpoint of a quantization interval. The distortion resulting from the quantization process is directly proportional to the square of the step size and therefore is inversely proportional to the number of levels, n. A uniform eight-level midrise quantizer is shown in Fig. 5.

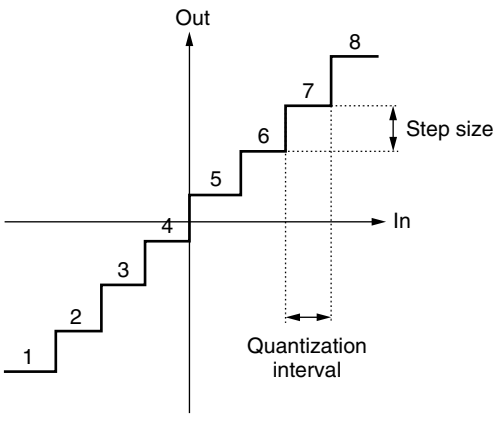


Figure 5. A uniform eight-level midrise quantizer.

2.2.1.2. Adaptive Quantization. *Adaptive quantization* is used for samples with dynamically changing range of amplitudes. They have uniform step sizes, but the length of step sizes changes according to the dynamic range of the input waveform. There are two major approaches [3]:

1. *Offline Adaptive* (*Forward-Adaptive*) *Approach*. In this type of quantization, source output is first divided into blocks. Each input block is individually analyzed and distinct quantizer parameters are assigned. Quantizer parameters should be transmitted as side information.
2. *Online Adaptive* (*Backward-Adaptive*) *Approach*. In this method, adaptation is based on the quantizer output. Since parameters are available both at the transmitter and receiver, no feedback link is necessary.

2.2.1.3. Nonuniform Quantization. Uniform quantization is optimum when the input signal levels are uniformly distributed. Often the distribution of the input signal is nonuniform, and using a nonuniform quantizer instead of a uniform quantizer results in a considerable performance improvement. In nonuniform quantization, the quantizer step sizes are not equal. Usually step sizes close to the mean signal level are smaller and the step sizes become larger as the input signal deviates further from the mean. Having the same number of quantization levels, nonuniform quantizers result in a lower average distortion than uniform quantizers. The expense associated with nonuniform quantizers is their more complex structure.

There are various kinds of nonuniform quantization schemes. One method is to optimize the quantizer levels to produce minimum quantization error according to the source statistics. This can be achieved by compressing the input amplitudes according to a PDF, and then applying the resulting signal to a uniform quantizer. At the decoder a corresponding expander is utilized in the reconstruction process. The mappings that are performed in the blocks are also shown in Fig. 6. This type of quantizer is termed compander, because of *comp*ressor and exp*ander* that exist in the system. A major application area of companders is commercial telephony. There are two frequently used companding techniques named μ law and A law, which are used in the United States and Europe, respectively [4].

2.2.2. Vector Quantization (VQ). Although Shannon has stated that given an average distortion D and rate $R(D)$, a waveform can be reconstructed with a distortion arbitrarily close to D, this is not practically achievable with scalar quantization. Vector quantization (VQ) takes advantage of processing many samples at once for a more accurate representation of source output. VQ works best when input blocks are correlated, but even when the quantized samples are independent VQ performs better than scalar quantization [4].

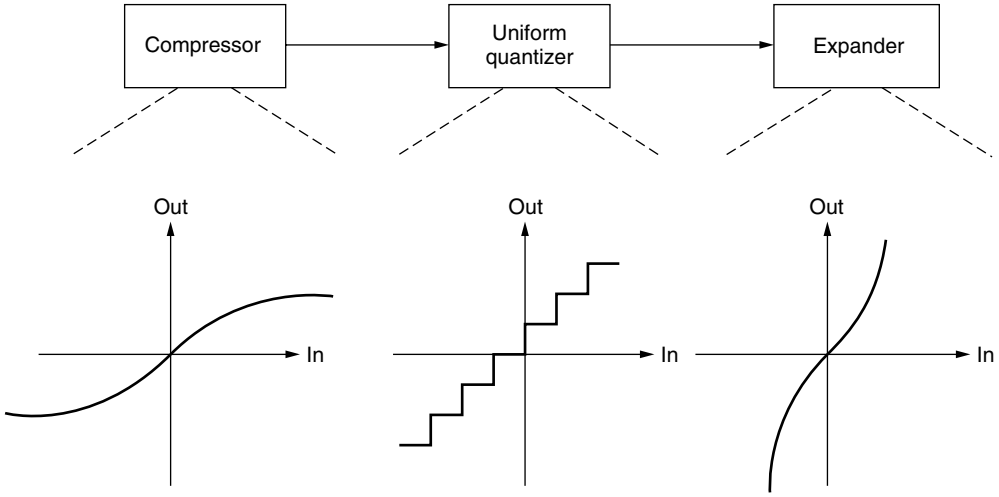


Figure 6. Compander.

In VQ, instead of processing one sample at a time, a group of samples (called a *vector*) is mapped to a codebook index. A *codebook* is a set consisting of a finite number of vectors formed in such a way that each input vector has a corresponding output.

The dimension of a vector quantizer is defined as the number of samples in the vector. The rate of the vector quantizer is defined as

$$R = \frac{\log_2 n}{L} \text{ bits per sample} \tag{16}$$

where n is the size of VQ codebook and L is the dimension of the quantizer [3].

VQ structure introduces increased implementation complexity. To alleviate this problem, tree-structured VQ can be used. Details of various aspects of vector quantization can be found in the literature [4–6].

3. CODING TECHNIQUES

Waveform coding techniques can be classified into several categories. In this article we discuss only major temporal and spectral coding techniques. Further details on various coding techniques can be found in the literature [4,5,7]. Another important category is model based coding, which is not covered here.

3.1. Temporal Coding Techniques

In temporal coding techniques the waveform coding process is performed in the time domain. In the following sections we will discuss the basic temporal coding techniques: pulse code modulation (PCM), differential PCM (DPCM), and delta modulation (DM). Details on these and other temporal coding methods can be found in the literature [4,5,7].

3.1.1. Pulse Code Modulation (PCM). *Pulse code modulation* (PCM) has gained popularity with digital telephony. It is the most straightforward method of digitization as we saw in the last section. Conceptually the encoder performs three functions: (1) sampling the input waveform, (2) quantizing the samples, and (3) representing each quantizer level with a binary index.

A typical PCM system is shown in Fig. 7. An input signal is first prefiltered to ensure band limitation, and then is sampled and quantized. Next, the quantized amplitude values are mapped to electrical signals suitable for transmission over a particular channel. The received signals are demodulated and passed through the reconstruction filter. Finally, an estimate of the input waveform is received by the destination.

Assuming that the quantization noise is uniformly distributed over the quantization intervals, it has been

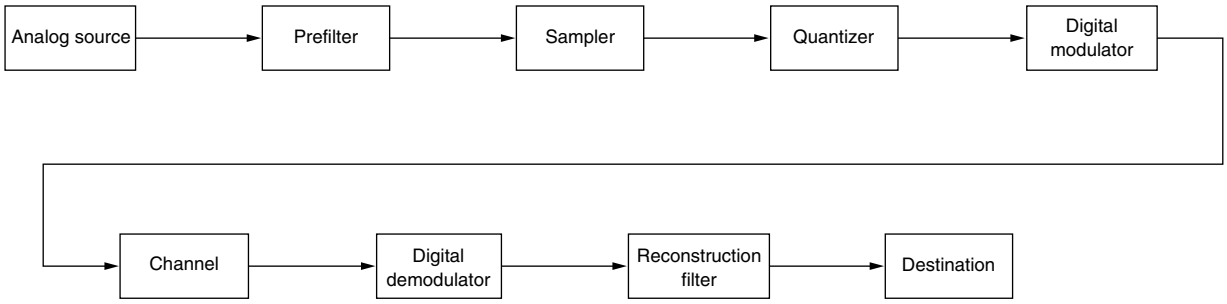


Figure 7. A typical PCM system.

shown [3] that the performance with an n-bit uniform quantizer can be evaluated by

$$(\text{SQNR})_{\text{dB}} = 6.02n + \alpha \tag{17}$$

where $\alpha = 4.07$ for peak SQNR and $\alpha = 0$ for average SQNR. Hence each additional bit improves the performance by about 6 dB.

PCM has several advantages, including immunity to channel impairments such as noise and interference and ease of implementation. Because of its widespread use, PCM has become a standard for comparison of data compression schemes. PCM systems using companding generally provide toll quality speech at a rate of 64 Kbps.

3.1.2. Differential Pulse Code Modulation (DPCM). PCM is designed without taking into account any correlation between successive samples. In practice the successive samples of most waveforms are highly correlated. The key idea behind differential pulse code modulation (DPCM) is to exploit this correlation.

A typical DPCM system is shown in Fig. 8. In this system, an estimate of the current sample value is produced by the predictor through the feedback loop that is located in the encoder part. Next, this estimate is subtracted from the current sample value; the difference is quantized, modulated, and transmitted through the channel. Then in the receiver, demodulated symbols are recovered through the second feedback loop located at the receiver. In this loop, the predicted value is added to the quantized prediction error. The variance of the difference between a sample and its estimate is smaller than or equal to the variance of the original signal. Therefore the quantization noise power of a DPCM system is smaller than that of a PCM system. This results in a higher SQNR than that of a PCM system for the same data rate or the same SQNR value can be achieved with a lower data rate. A typical DPCM system provides toll-quality speech at a rate of 32–48 kbps.

Other versions of the DPCM system where predictor parameters change according to the statistics of the input signal and other parameters can provide additional improvements [5,7].

3.1.3. Delta Modulation (DM). *Delta modulation* (DM) is a simplified version of DPCM. DM is also based on the observation that input signal amplitudes generally do not make very sudden changes. The key point in DM is oversampling the input waveform with a sampling frequency much higher than the Nyquist rate, and increasing the correlation between the samples. DM can be visualized as a derivative approximation of an input waveform.

DM has almost the same structure as DPCM system shown in Fig. 8. The only change required for DM is to replace the predictor by a unit delay. In DM, following the input waveform within a $\pm\Delta$ range, a quantized signal is produced, as shown in Fig. 9. This type of approximation is known as "staircase approximation." When the ratio of step size to sampling period (i.e., Δ/T_s) is larger than the maximum slope of the input waveform, then slope overload occurs. The resulting distortion is named as slope overload distortion. The second type of quantization noise appears in DM when Δ is larger than the slope of the waveform. This is termed the *granular noise*. Both noise types are depicted in Fig. 9. The most crucial parameter in DM is Δ, also known as the *step size*. Large values of Δ permit the approximation of rapid changes, whereas smaller values of Δ help reduce the effect of granular noise. DM systems generally provide a rate of 32–64 kbps for toll-quality speech.

To reduce the effects of the abovementioned quantization noise types, adaptive delta modulation (ADM) can be used. In ADM Δ varies according to the input waveform. Details on ADM can be found in the treatise by Jayant [5].

3.2. Spectral Coding Techniques

In temporal coding techniques, the input is a full-band waveform, that is, not filtered into smaller frequency bands. In spectral waveform coding techniques, the input signal is divided into several frequency components and

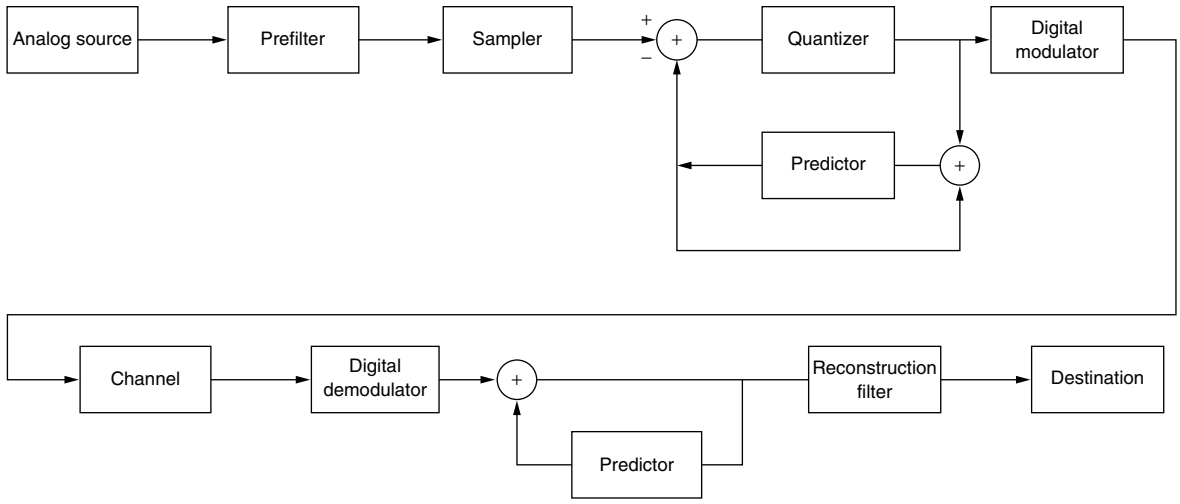


Figure 8. A typical DPCM system.

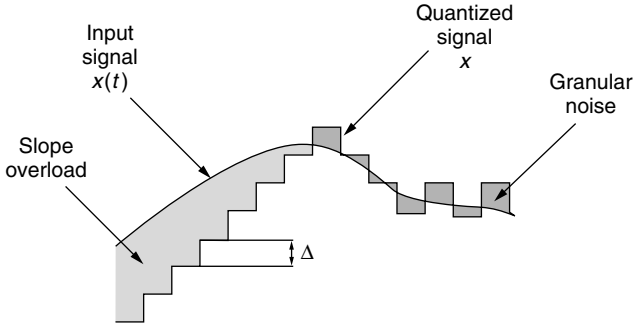


Figure 9. Noise in a DM system.

each component is coded separately. Decomposing the source output into different frequency bands (subbands) is performed by digital filters. Each filter output component is quantized separately so that the resulting quantization noise is contained within its band.

Spectral coding techniques take advantage of human perception of signals such as image or speech. Human senses cannot detect quantization distortion at all frequencies with the same precision. Spectral coding techniques utilize dynamic bit allocation by assigning more bits to more important frequency components, and less bits to the others. Spectral coding techniques also increase efficiency by removing the redundancy between input samples, thus providing uncorrelated channel inputs.

Spectral coding techniques can be divided into two major classes: subband coding (SBC) and transform coding (TC). In practice SBC is generally used for audio signals [5], and TC is dominantly used for image signals [8].

3.2.1. Subband Coding (SBC). *Subband coding* (SBC) is a waveform coding method that first decomposes an input waveform into different subbands and then operates on them. As shown in Fig. 10, division of the signal into subbands is performed with the help of an analysis filterbank. Filter outputs are generally translated to low-pass signals by an operation equivalent to single-sideband modulation. Then the outputs are sampled at Nyquist rate and encoded in the time domain, with the help of one of the temporal coding techniques such as PCM or DPCM. The coding technique for each branch is chosen according to the properties of the corresponding band, or a perceptual criterion. The resulting signals are multiplexed and transmitted through the channel. At the receiver part, the received signal is demultiplexed and decoded. The resulting signal is passed through a synthesis filterbank and the outputs are summed to form the reconstructed signal.

Quadrature mirror filters (QMFs) are the popular choice for the filterbanks. These filters obey certain symmetry conditions and can achieve perfect alias cancellation.

Depending on the nature of a particular application, SBC can operate over equal-width subbands or unequal-width subbands. In speech coding applications usually four to eight subbands are used. A 32-kbps subband coder can provide toll-quality speech.

In SBC, since each filter output is encoded and quantized separately, the system can control and distribute the quantization noise across the spectrum. Therefore the quantization noise in each band can be adjusted independently from that in other bands. Different number of bits can be assigned to each subband. Also errors due to channel impairment affect only the outputs of corresponding frequency bands. Combination of subband coding with adaptive prediction systems can improve system performance [5].

3.2.2. Transform Coding (TC). *Transform coding* (TC) is also known as *block transform coding* or *block quantization*. As shown in Fig. 11, a transform coder encodes a block of a discrete input sequence according to a bit allocation scheme, derived from the perceptual significance of components. The input to the linear transformer can be obtained with a high-resolution temporal coder.

The efficiency of a TC technique depends on the type of linear transform applied and the bit allocation scheme used in the quantization of transform coefficients.

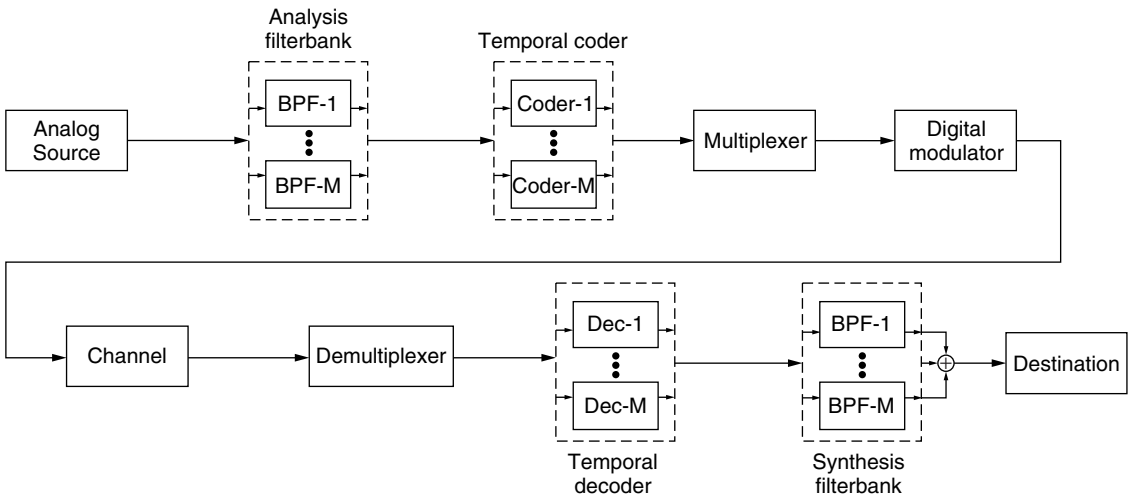


Figure 10. A typical subband coder.

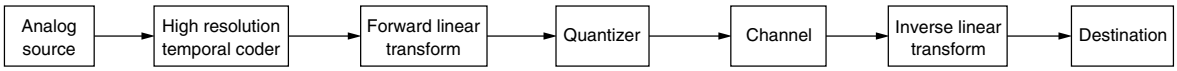


Figure 11. A typical transform coder.

The main objective of the linear transformation in TC is to provide uncorrelated samples. In this sense, the *Karhunen–Loeve transform* (KLT), also known as the *Hotelling transform* or the *eigenvalue transform*, is optimum since it transforms discrete variables into uncorrelated coefficients. KLT is based on the statistical properties of vector representations of the input signal and minimizes the mean-square error between the input and its approximation. KLT has the maximum capacity to perform data compression.

A major disadvantage of the KLT is the high computational complexity. This problem can be solved by using a suboptimal but more practical transform method such as discrete Fourier transform (DFT), discrete cosine transform (DCT), discrete Hadamard transform (DHT), or discrete Walsh transform (DWT).

An important design issue in TC is how to implement the bit allocation scheme. Optimum bit allocation, which yields the best SQNR, depends on the variance of quantization coefficients but it is too complex for most practical purposes.

Zonal sampling is a simple bit allocation method. In zonal sampling zero bits are allocated to less important (i.e., out-of the zone) coefficients. In order to obtain a good performance from zonal sampling, the input power spectral density should be highly structured.

A more complex version of zonal sampling is threshold sampling, where a zone is defined as a variable region depending on the amplitudes of the coefficients.

TC provides a better performance than does SBC at the expense of increased coding delay and more complex structure. The number of subbands in TC is typically greater than the number of subbands used in SBC. This number is also referred to as the order of the transform.

A more complex version of TC is adaptive transform coding (ATC). It includes adaptive bit allocation from window to window while keeping the total data rate constant. ATC results in a significant increase in the SQNR with the expense of further increased complexity. More details on ATC can be found in Refs. 5 and 7.

BIOGRAPHIES

Güneş Karabulut was born in Istanbul, Turkey in 1979. She received the B.Sc. degree in Electrical and Electronics Engineering from Boğaziçi University, Istanbul, Turkey and the M.A.Sc. degree in Electrical Engineering from the University of Ottawa, Ontario, Canada. Currently she is pursuing the Ph.D. degree at the University of Ottawa.

From 1999 to 2000 she worked on motion estimation algorithms in Boğaziçi University Signal and Image Processing Laboratory. Since September 2000 she is employed as a Research Assistant at Communications and Signal Processing Group, University of Ottawa. Her research interests include coding theory, coded modulation schemes, and image coding. Ms. Karabulut is a member of IEEE Information Theory Society.

Abbas Yongaçoğlu received the B.Sc. degree from Boğaziçi University, Turkey, in 1973, the M.Eng. degree from the University of Toronto, Canada, in 1975, and the Ph.D. degree from the University of Ottawa, Canada, in 1987, all in Electrical Engineering.

He worked as a researcher and a system engineer at TUBITAK Marmara Research Institute in Turkey, Philips Research Labs in Holland and Miller Communications Systems in Ottawa. In 1987 he joined the University of Ottawa as an assistant professor. He became an associate professor in 1992, and a full professor in 1996. His area of research is digital communications with emphasis on modulation, coding, equalization and multiple access for wireless and high speed wireline communications.

BIBLIOGRAPHY

1. N. Farvardin and V. Vaishampayan, Optimal quantizer design for noisy channels: An approach to combined source channel coding, *IEEE Trans. Inform. Theory* **33**(6): 827–838 (Nov. 1987).
2. C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE Conv. Rec.* **7**: 142–163 (1959).
3. T. S. Rappaport, *Wireless Communication: Principles & Practice*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
4. K. Sayood, *Introduction to Data Compression*, Morgan Kaufmann, San Francisco, 2000.
5. N. S. Jayant, *Digital Coding of Waveforms Principles and Applications to Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
6. R. M. Gray, Vector quantization, *IEEE Acoust. Speech Signal Process. Mag.* **1**(2): 4–29 (April 1984).
7. J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 2001.
8. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison Wesley, New York, 1993.

FURTHER READING

Gibson J. D., ed., *The Mobile Communication Handbook*, CRC Press, Boca Raton, FL, 1996.

Haykin S., *Communication Systems*, Wiley, New York, 1994.

Ortega A. and K. Ramchandran, Rate-distortion methods for image and video compression, *IEEE Signal Process. Mag.* **15**(6): 23–50 (Nov. 1998) (gives details on rate distortion theory).

Effros M., Optimal modeling for complex system design, *IEEE Signal Process. Mag.* **15**(6): 51–73 (Nov. 1998) (gives details on rate distortion theory).

Goyal V. K., Theoretical foundations of transform coding, *IEEE Signal Process. Mag.* **18**(5): 9–21 (Sept. 2001).

Oppenheim A. V., R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1999 (an excellent source for more details on sampling and reconstruction processes).

WAVELENGTH-DIVISION MULTIPLEXING OPTICAL NETWORKS

EYTAN MODIANO
Massachusetts Institute of Technology
Cambridge, Massachusetts

1. INTRODUCTION

Communication networks were first developed for providing voice telephone service. Early networks were deployed using copper wire as the medium over which traffic was sent in the form of electromagnetic waves. As demand for communication increased, networks began to use optical fiber cables over which information was sent in the form of lightwaves. Thanks to the relatively low attenuation losses of optical fiber, transmitting information over fiber allowed for a significant increase in the transmission capacity of networks. For example, while transmission over copper was limited to only a few tens of thousands of bits per second (kbps), optical fiber transmission enabled data rates exceeding hundreds of millions of bits per second (Mbps). However, the relatively recent development of the Internet has resulted in a tremendous increase in demand for transmission capacity. More recently, developments in optical transmission technology have achieved data rates that exceed many billions of bits per second (Gbps). Even with these enormous data rates, demand still far exceeds the available network capacity. As a result, telecommunication equipment companies are constantly trying to develop new technology that can increase network capacity at reduced costs.

While fiberoptic technology resulted in a significant increase in a network's "bandwidth," or the amount of information that the network could send, the creation of the Internet resulted in an even greater demand for bandwidth. As demand for network capacity increased, service providers exhausted their available transmission capacity. One approach to alleviating fiber exhaust is to deploy additional fiber. This solution, however, is not always economically feasible. As a result, new technologies were developed to increase the transmission capacity of existing fiber.

The simplest approach is to increase the rate of transmission over the fiber (i.e., sending more bits per second). Since 1980, fiber transmission rates have increased from a few Mbps to nearly 100 Gbps. Since most users rarely need such high data rates, a network technology called *synchronous optical networks* (SONET) was developed to allow users to share the capacity of a fiber [1].

SONET is a technology for multiplexing a large number of low-rate circuits onto the high-rate fiber channel. The "basic" transmission rate of SONET is 64 kbps for supporting voice communications. SONET multiplexes large numbers of 64-kbps channels onto higher-rate datastreams. SONET defines a family of supported data rates. These data rates are often referred to as (optical carrier) OC-1, OC-3, OC-12, OC-48, and OC-192; where OC-1 corresponds to a data rate of 51.84 Mbps, or 672 voice circuits, OC-3 is 155.52 Mbps or 2016 voice circuits (3 times OC-1, OC-12 is 12 times OC-1, etc.). SONET uses time-division multiplexing (TDM) for combining traffic from multiple sources onto a common output. TDM multiplexes traffic from different sources by interleaving small "slices" of data from each source. Thus, if traffic from three OC-1 sources is being time division multiplexed onto an OC-3 transmission channel, each source would get access to the channel for a short period of time in a round-robin order, as shown in Fig. 1.

However, because of fundamental limits on optical transmission, the transmission capacity of a fiber cannot be increased indefinitely. Hence, to further increase the capacity of a fiber, a technology called wavelength-division multiplexing (WDM) was developed [1]. Wavelength division multiplexing allows transmissions on the fiber to use different colors of light (each color represents a different wavelength over which light propagates). Whereas in the first optical communications networks, light was transmitted through the fiber using a single wavelength, WDM permits light at multiple, different wavelengths, to be transmitted through a single fiber simultaneously. WDM is analogous to frequency-division multiplexing (FDM), which is often used for transmission over the airwaves.

In WDM systems, incoming optical signals are assigned a specific wavelength and then multiplexed onto the fiber. Moreover, such systems are bit-rate- and protocol-independent, meaning that each incoming signal can be carried in its native format and at a different rate. For example, a WDM system may support the transmission of multiple SONET signals on a single fiber, each operating at transmission rates of 10 Gbps (OC-192). As shown in Fig. 2, WDM systems are designed to operate in the low loss region of optical fiber, around the 1.5-μm band. Typically, wavelengths are assigned in this region with a separation of 25–100 GHz; and systems supporting anywhere from 80 to 160 wavelengths are presently being deployed.

The simplest approach to using WDM is to treat each wavelength as if it were on a separate fiber and continue

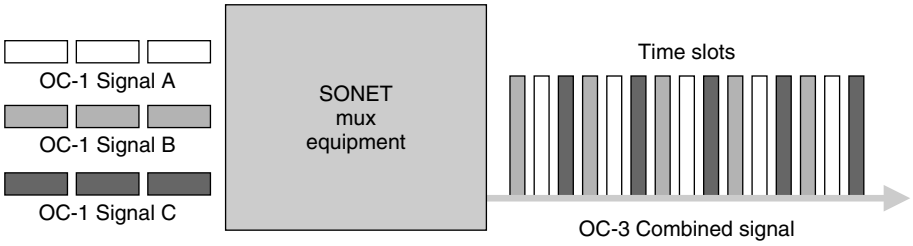


Figure 1. SONET time-division multiplexing.

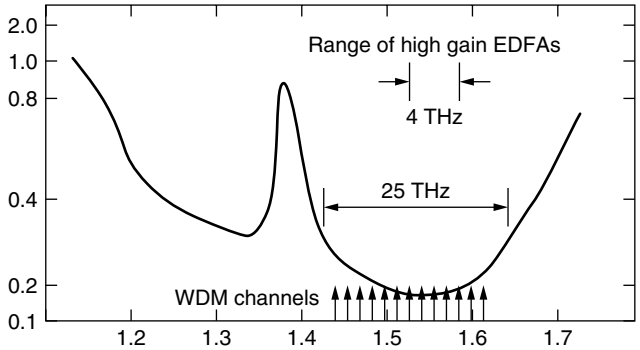


Figure 2. Wavelength-division multiplexing allows the transmission on multiple wavelengths within a single fiber.

to design the network using point-to-point links, as shown in Fig. 3. With this approach the available fiber capacity would in effect be increased by a factor that equals the number of wavelengths, and the network architecture would be largely unchanged. By itself, this approach leads to significant cost savings. First, in many cases existing fiber cannot meet demand, and WDM can help alleviate the fiber exhaust problem. In addition, through the use of erbium-doped fiber amplifiers (EDFAs), as shown in Fig. 3, a number of wavelengths can be simultaneously amplified. This leads to significant cost savings when compared to pre-WDM systems where each fiber would require its own amplification. Since the cost of amplification (and or regeneration) forms a large fraction of the overall network deployment cost, these cost savings through the use of EDFAs play a large role in the commercial success of WDM systems.

As described so far, from a network perspective, WDM systems are not vastly different from any other optical transmission systems. However, in addition to pure transmission technology, a number of WDM network elements have been developed that make it possible to design networks that actually route and switch traffic in the optical domain [2]. While these optical networking functions are rather limited, they have the potential of significantly enhancing network performance. In Section 2, we describe some of the basic WDM network elements. Subsequently, in Section 3 we describe architectures for future WDM optical local-area networks (LANs), and in Section 4 we describe architectural issues in the design of all-optical WDM wide-area networks (WANs). Since all-optical networks are not likely to emerge in the near future, we devote the last section of this article to discussing future WDM-based networks that use a combination of optical and electronic processing.

2. WDM NETWORK ELEMENTS

A number of optical network elements have been developed that allow for simple optical processing of signals. The simplest such device is a "broadcast star," shown in Fig. 4. In a broadcast star, each node is connected using one input and one output fiber. The fibers are then coupled or "fused" together so that a signal coming in from any input fiber will propagate on all output fibers. In this way, all nodes can communicate with each other. Because of its simplicity and broadcast property, a star has often been proposed as a suitable technology for optical LANs. In Section 3 of this article we describe WDM-based LAN architectures utilizing a broadcast star.

A somewhat more sophisticated device is a *wavelength router*, a passive optical device that combines signals from a number of input fibers and "routes" those signals in a static manner, based on the wavelength on which they propagate, to the output fibers. The operation of a wavelength router with four input fibers is shown in Fig. 5. Notice that the signal propagating on wavelength 1 of the first input fiber is routed to the first output fiber, while the signal propagating on wavelength 2 is routed to the second output fiber, and so on. Similarly, the signal propagating on wavelength 1 of the second fiber is routed directly to the second output fiber, while the signal on wavelength 2 is routed to the third output fiber, and so forth. A wavelength router is different from a broadcast star in that it separates the wavelengths onto different output fibers. Notice that the signals propagating on the first input fiber are routed onto the four output fibers so that the first wavelength is routed to the first output, the second to the second output, and so on. In this way, a wavelength router is commonly used in optical networks as a WDM demultiplexer. WDM demultiplexers can be used to form optical add/drop multiplexers (OADMs), that are able to drop or add any number of wavelengths at a node. OADMs play an important role in a network and can be designed in a number of ways (not necessarily using a

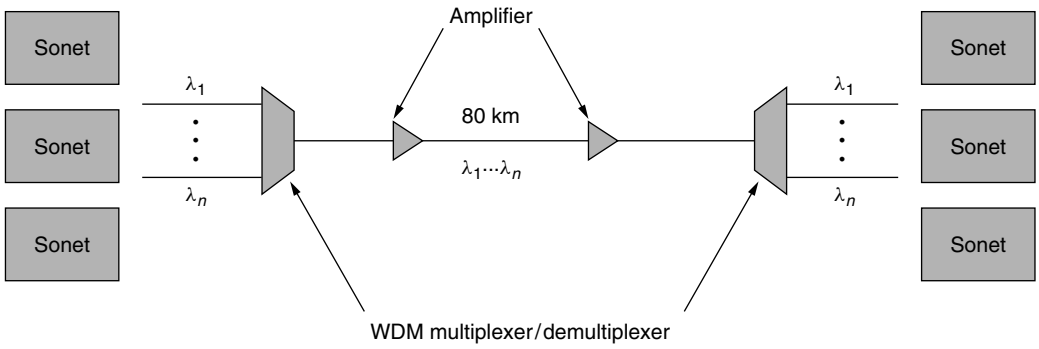


Figure 3. Typical SONET/WDM deployment.

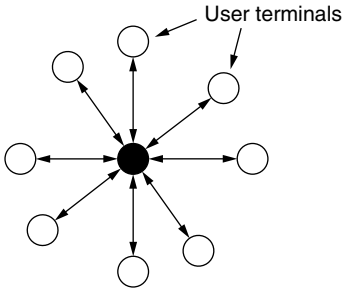


Figure 4. Broadcast star.

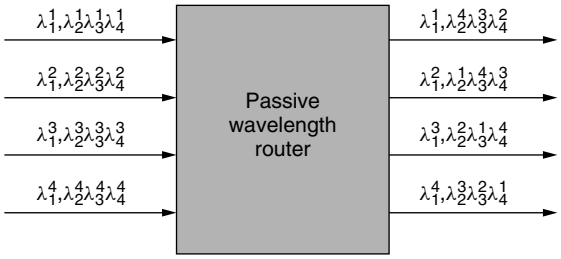


Figure 5. Wavelength router.

wavelength router); the operation of an OADM is shown in Fig. 6.

WDM demultiplexers can be used in conjunction with optical switches to form an even more sophisticated optical switching device known as a frequency-selective switch (FSS). Unlike a wavelength router that routes wavelength from input fibers onto output fibers in a static manner, a FSS is a configurable device that can take any wavelength from any input fiber and switch it onto any output fiber. In this way, a FSS allows for some flexibility in the operation of the network. The basic operation of a FSS is shown in Fig. 7. The signals traveling on each input fiber are demultiplexed into the different wavelengths. Each wavelength is then connected to a switching element that can switch any of the input fibers onto any of the output fibers. The outputs of the switch elements are then connected to WDM multiplexers that combine the signals onto the output fibers.

While a FSS adds significant functionality to the operation of the network, it requires the ability to dynamically switch optical signals from one input fiber onto another. Such switching can be accomplished optically using a number of techniques that are beyond the scope of this article. Optical switching technology, while rapidly progressing, is still relatively immature and costly. Alternatively, the signals can be switched in the electronic domain by first converting the signals traveling on each wavelength to electronics, switching the signals electronically, and retransmitting the signal onto the output fiber. Such optoelectronic switching is often less costly, and results in faster switching times, than does optical switching. However, in order to perform the optoelectronic conversion, the signal format (e.g., modulation technique, bit rate) must be known. This eliminates the signal "transparency" that is so desirable in optical networks.

A FSS adds flexibility to network operations by allowing wavelengths to be dynamically switched among the different fibers. However, notice that wavelength cannot be switched arbitrarily. For example, it is not possible for a FSS to route the signal propagating on a given wavelength from more than one input fiber onto the same output fiber. This limitation can be overcome by a *wavelength converter*, which can convert a signal from one wavelength onto another. Wavelength converters come in a number of variations. The simplest is a *fixed-wavelength converter*, which can convert a given wavelength onto another in a predetermined static fashion. More flexible converters can convert a given wavelength onto one of a number of wavelengths dynamically; such converters are known as *limited-wavelength converters*. The most flexible wavelength converters can convert any wavelength onto any other. Wavelength conversion can be accomplished in either the optical or electronic domain. Optical wavelength conversion is a rather immature technology primarily implemented in experimental laboratories; while electronic wavelength conversion suffers from the need for optoelectronic conversion and the consequent loss of transparency. Hence, while desirable, wavelength conversion in optical networks is still very limited.

Of course, in order to use WDM technology, one must be able to transmit and receive the signal on the different wavelengths. Transmission is accomplished using lasers that operate at a given wavelength, while reception is accomplished using WDM filters and light detectors. Typically lasers and filters are designed to operate at a single, fixed, frequency; such devices are commonly referred to as *fixed tuned devices*. Fixed tuned WDM transmitters and receivers again limit the capability and flexibility of an optical network because a signal that is transmitted on a given wavelength must travel throughout the network, and be received on that wavelength. Hence, without wavelength conversion, a node that has a transmitter that operates on a given frequency can

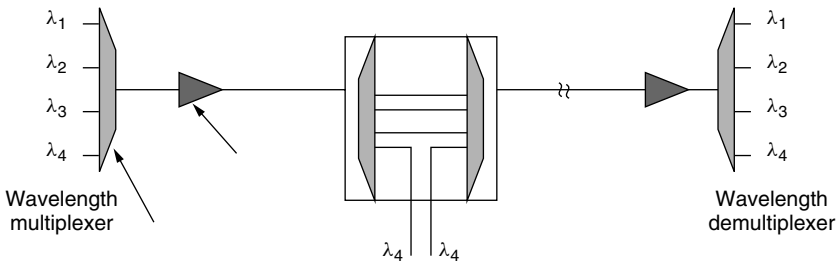


Figure 6. Optical add/drop multiplexer.

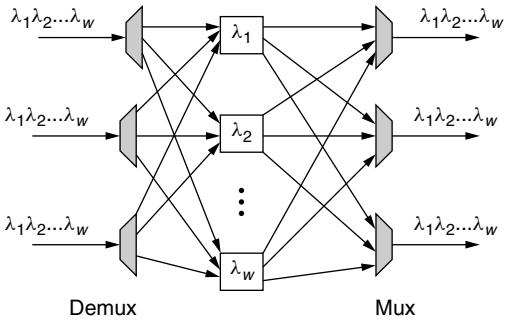


Figure 7. A frequency-selective switch.

communicate only with nodes that are equipped with a receiver for that frequency. Tunable transmitters and receivers are hence very desirable for optical networks; however, much like wavelength conversion, that technology is still at its inception. Tunable transmitters and receivers are often characterized according to the speed with which they can tune to different wavelengths. Slow tuning lasers, which can tune on the order of a few milliseconds, are now becoming commercially available, while fast lasers that can tune in microseconds are emerging.

3. WDM LOCAL AREA NETWORKS

Typically, local-area networks (LANs) span short distances, ranging from a few meters to a few thousands of meters. Because of the relatively close proximity of nodes, LANs are typically designed using a shared transmission medium. In this section we discuss WDM-based LANs, where users share a number of wavelengths, each operating at moderate rates (e.g., 40 wavelengths at 2.5 Gbps each).

Typically WDM-based LANs assume the use of a broadcast star architecture [3]. An optical star coupler is used to connect all the nodes. Each node is attached to the star using a pair of fibers: one for transmission and the other for reception. The star coupler is a passive device that simply connects all the incoming and outgoing fibers so that any transmission, on any wavelength, on an incoming fiber is broadcast on all outgoing fibers. In order for nodes to communicate, they must tune their transmitters and receivers to the appropriate wavelength.

A WDM LAN based on a broadcast star architecture can provide a transmission capacity that can easily exceed 100 Gbps. Perhaps the greatest reason preventing such systems from emerging is the cost of WDM transceivers. In order for a WDM LAN to allow flexible bandwidth sharing, both transmitter and receiver must be rapidly tunable over the available wavelengths. Transceiver tuning times that are smaller than the packet transmission times are desirable if efficient use of the bandwidth is to be obtained. With packets that are just a few thousands of bits in length, this calls for tuning times on the order of microseconds or faster. Present technology for fast tuning lasers is largely at the experimental stage; and while such lasers are slowly becoming commercially available, they are very expensive. Similarly, fast tuning receivers are also complex and expensive.

It is reasonable to expect that as the commercial market for these devices develops, their cost will decrease and they will become more widely available. However, in the near future, if WDM-based LANs are to become a reality they must limit the use of tunable components. WDM-based LANs are usually classified according to the number and tunability of the transmitters and receivers [4]. For example, a system utilizing one tunable transmitter and one tunable receiver is referred to as a TT-TR system. Similarly, a fixed tuned system would be referred to as FT-FR. Obviously, a FT-FR system can only use one wavelength if full connectivity among the nodes is desired. In order to provide full connectivity over multiple wavelengths, it is necessary that either the receivers or the transmitters be tunable. Systems employing either a tunable transmitter and a fixed tuned receiver (TT-FR) or a fixed transmitter and a tunable receiver (FT-TR) have been proposed in the past for the purpose of reducing the network costs.

Particularly attractive is the use of a fixed tuned receiver, because with a fixed tuned receiver all communication to a node is done on a fixed wavelength. Hence, this eliminates the need for any coordination before the transmission takes place. Of course, having a fixed tuned receiver means that nodes will have to be assigned to wavelengths in some fashion. For example, in an N node–W wavelengths network, N/W nodes can be assigned to receive on each wavelength. This, of course, creates a number of complications. First, when nodes are assigned to wavelengths in such a fixed manner, it is possible that certain wavelengths will be carrying a larger load than others and so, while some wavelengths may be lightly loaded, others may be overly saturated. In addition, such a network is complicated to administer because whenever adding a new node, care must be taken to determine on which wavelength it must be added, and a transceiver card tuned to that wavelength must be used.

In order to obtain the full benefit of the WDM bandwidth, a WDM-based LAN must have a TT-TR architecture. With this architecture, some form of transmission coordination is necessary for three reasons: (1) if two nodes transmit on the same wavelength simultaneously, their transmissions will interfere with each other (collide) and so some mechanism must be employed to prevent such collisions; (2) if two or more nodes transmit to the same node at the same time (albeit on different wavelengths), and if that node has only a single receiver, it will be able to receive only one of the transmissions; and (3) for a node to receive a transmission on a wavelength, it must know in advance of the upcoming transmission so that it can tune its receiver to the appropriate wavelength.

Most proposed WDM LANs use a separate control channel for the purpose of pretransmission coordination. Often, these systems use an additional fixed tuned transceiver for the control channel. Alternatively, the control and data channels can share a transceiver, as shown in Fig. 8.

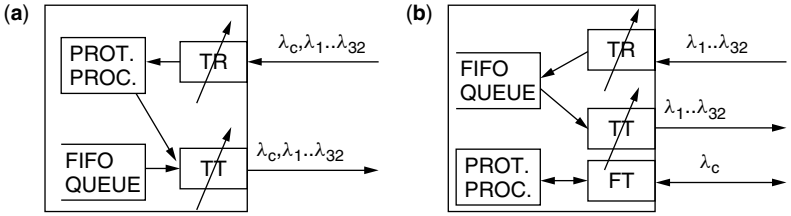


Figure 8. User terminals: (**a**) single turnable transceiver; (**b**) two-transceiver configuration.

In order for one node to send a packet to another, it must first choose a wavelength on which to transmit, and then inform the receiving node, on the control channel, of that upcoming transmission. A number of medium access control (MAC) protocols have been proposed to accomplish this exchange [5]. These protocols are more complicated than single-channel MAC protocols because they must arbitrate among a number of shared resources: the data channels, the control channel, and the receivers.

Early MAC protocols for WDM broadcast networks attempted to use ALOHA for sharing the channels [6]. With ALOHA, nodes transmit on a channel without attempting to coordinate their transmissions with any of the other nodes. If no other node transmits at the same time, the transmission is successful; however, if two nodes transmit simultaneously, their transmissions "collide" and both nodes must retransmit their packets. To reduce the likelihood of repeated collisions, nodes wait a random delay before attempting retransmission. When the load on the network is light, the likelihood of such a collision is low; however, with increased load such collisions occur more often, limiting network throughput. Single-channel versions of ALOHA have a maximum throughput of approximately 18%. A slotted version of ALOHA, where nodes are synchronized and transmit on slot boundaries, can achieve a throughput of 36%.

In a WDM system using a control channel, a MAC protocol must be used both for the control and the data channels. Early MAC protocols attempted to use a variation of ALOHA on both the control and the data channels [7]. In order for a transmission to be successful, the following sequence of events must take place: (1) the transmission on the control channel must be successful (i.e., no control channel collision), (2) the receiving node must not be receiving any other transmission at the same time (i.e., no receiver collision), and (3) the transmission on the chosen data channel must also be successful. In a system that uses ALOHA for both the control and data channels, it is clear that throughput will be very limited. It has been shown that systems using slotted ALOHA for both the data and control channels achieve a maximum utilization of less than 10% [7].

In view of the discussion above, a number of MAC protocols that attempt to increase utilization by coordinating and scheduling the transmissions more carefully have been proposed [4]. For example, the protocol described by Modiano and Barry [8] uses a simple master/slave scheduler as shown in Fig. 9. All nodes send their requests to the scheduler on a dedicated control wavelength, λ_C. The scheduler, located at the hub, schedules the requests and informs the nodes on a separate wavelength, $\lambda_{C'}$, of their turn to transmit.

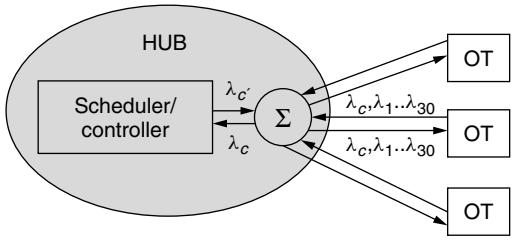


Figure 9. A WDM-based LAN using a broadcast star and a scheduler.

On receiving their assignments, nodes immediately tune to their assigned wavelength and transmit. Hence nodes do not need to maintain any synchronization or timing information. By measuring the amount of time that nodes take to respond to the assignments, the scheduler is able to obtain an estimate of each node's round-trip delay to the hub. This estimate is used by the scheduler to overcome the effects of propagation delays. The system uses simple scheduling algorithms that can be implemented in real time. Unicast traffic is scheduled using first-come–first-serve input queues and a window selection policy to eliminate head-of-line blocking, and multicast traffic is scheduled using a random algorithm [9].

We should point out, however, that despite their appeal, WDM-based LANs still face significant economic challenges. This is because the cost of WDM transceivers (especially tunable) is far greater than the typical cost of today's LAN interfaces. Since tunable WDM transceivers are just beginning to emerge in the marketplace, it is difficult to provide an accurate cost estimate for these devices, but it is certainly in the thousands of dollars. While a 100-Gbps LAN is very attractive, few would be willing to pay thousands of dollars for such LAN interfaces. Hence, in the near term, it is reasonable to expect WDM LANs to be used only in experimental settings or in networks requiring very high performance. However, as the cost of transceivers declines, it is not unlikely that this technology will become commercially viable.

4. WDM WIDE-AREA NETWORKS

In the WDM broadcast LAN architecture described above, nodes communicate by tuning their transmitter and receiver to common wavelengths. In a wide-area network (WAN), where a broadcast architecture is not scalable, traffic must be switched and routed at various communication nodes throughout the network. In electronic networks, this switching is accomplished using either circuit switching or packet switching techniques.

With packet switching, each network node must process each packet's header to determine the destination of the packet and make suitable routing decisions; with circuit switching, circuits are set up in advance of the communication and routing and switching decisions are predetermined for the duration of the call. Hence, with circuit switching there is no need for nodes to process the incoming data.

Optical packet switching involves a number of rather complex functions, such as header recognition, packet synchronization, and optical buffering. These technologies are rather crude and largely experimental at present [10]. Hence most efforts at optical networking have focused on circuit-switched networks. With WDM, much of the effort has been on the design of wavelength-routed networks, where connections between end nodes in the network utilize a full wavelength. There are a number of challenges in the design of an optical WDM network including the choice of a network architecture, performing the functions of routing and switching wavelengths, as well as assigning wavelengths to the various connections.

Since optical network elements are relatively expensive and of limited capabilities, the choice of a network architecture is particularly critical. Early efforts at designing all-optical WDM networks have been focused on a hierarchical architecture where different network elements are employed at different levels of the hierarchy. For example, as shown in Fig. 10, an optical star may be used in the local areas of the network and frequency-selective switches, optical amplifiers, wavelength converters, and other components may be used in the backbone of the network.

An early prototype of an all-optical-network is the all-optical network (AON) testbed developed by scientists at MIT, AT&T, and Digital Equipment Corporation (DEC) [2]. The AON testbed used the hierarchical architecture shown in Fig. 11. The lowest level in the

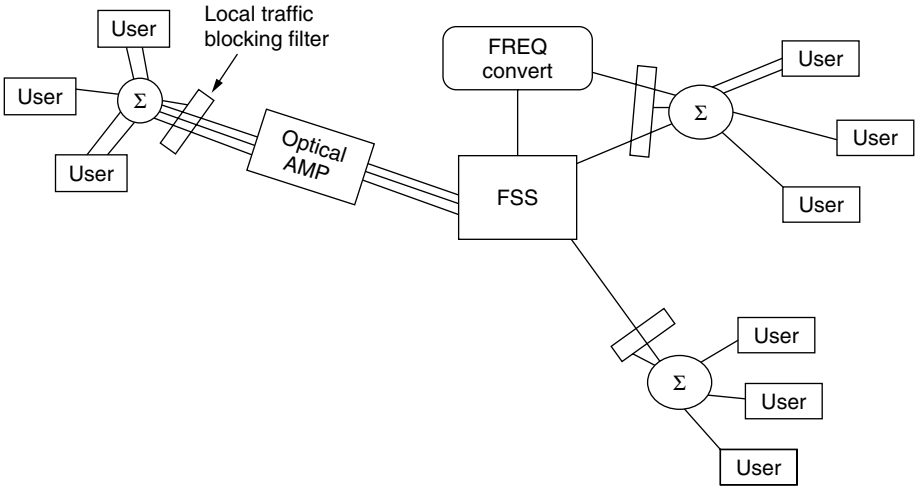


Figure 10. A partitioned all-optical WDM network.

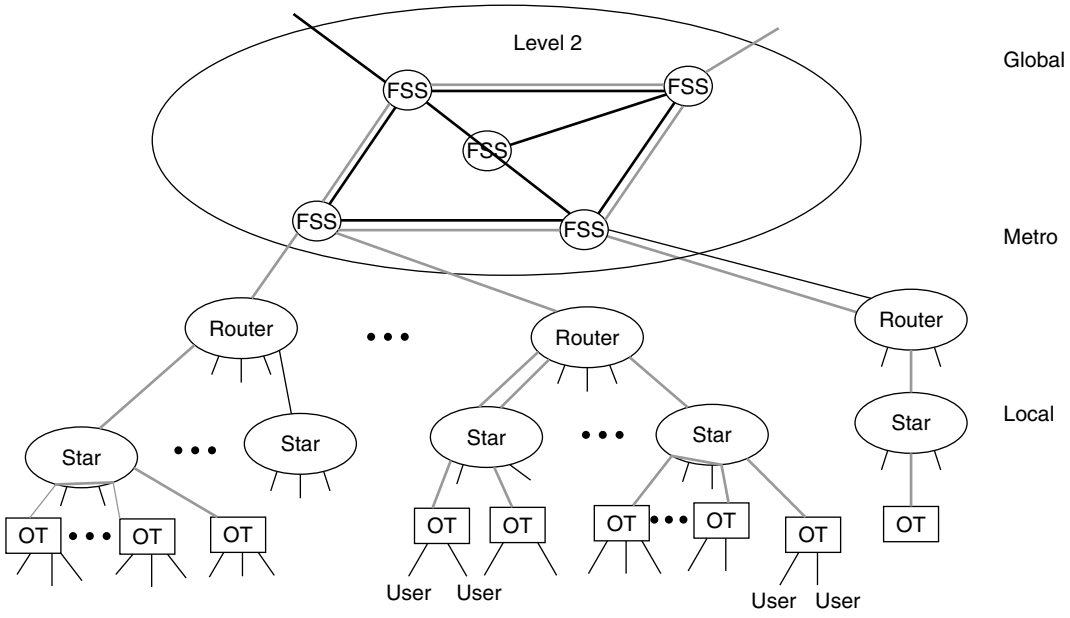


Figure 11. The AON architecture.

hierarchy was the LAN employing a broadcast star. A number of wavelengths were allocated for use within the local area, and those were separated from the other levels of the hierarchy using a wavelength blocking filter. Separating the local wavelengths from the rest of the network allows for those wavelengths to be reused at different local areas. The next level of the hierarchy used a wavelength router for connecting (in a static manner) different local areas. Finally, a FSS was used for providing connectivity in the wide area. A prototype of the architecture consisting of the lowest two levels was deployed in the Boston area connecting between facilities at MIT, MIT Lincoln Laboratory, and DEC.

The AON testbed supported two primary types of services: (1) a circuit-switched wavelength service that can establish wavelength connectivity between different nodes and (2) a circuit-switched time-slotted service by which a fraction of a wavelength can be assigned to a connection. The time-slotted service allows the flexibility of provisioning at the subwavelength level. However, implementing such a service requires very precise synchronization between the nodes so that the different circuits can be aligned on time-slot boundaries. The AON testbed demonstrated a 20-wavelength network, separated by 50 GHz and transmitting at rates of up to 10 Gbps per wavelength. AON also employed tunable transceivers. The transmitter was implemented using a DBR (distributed Bragg reflector) laser that can tune between wavelengths in 10 ns.

The early wavelength routing networks raised a number of architectural questions for all-optical networks. Perhaps the one that received the most attention is that of dealing with wavelength conflicts. Without using a wavelength converter, the same wavelength must be available for use on all links between the source and the destination of the call. A wavelength conflict may occur when each link on the route may have some free wavelengths, but the same wavelength is not available on all of the links. This situation can be dealt with through the use of wavelength converters that can switch between the wavelengths. However, because of the high cost of wavelength conversion, a number of studies quantifying the benefits of wavelength conversion in a network have been published [11,12]. Others have considered the possibility of placing the wavelength converters only at some key nodes in the networks [13]. However, the detailed results of these studies are beyond the scope of this article.

Another promising approach for dealing with wavelength conflicts is the use of a good wavelength assignment algorithm that attempts to reduce the likelihood of a wavelength conflict occurring. The wavelength assignment algorithm is responsible for selecting a suitable wavelength among the many possible choices for establishing the call. For example, the three calls illustrated in Fig. 12 can be established using three wavelengths ($\lambda_1\lambda_2\lambda_3$) as shown on the left or just two wavelengths as shown on the right. By choosing the assignment on the right, λ_3 remains free for use by future potential calls. A number of wavelength assignment schemes have been proposed [14,15], and the subject remains an active area of research.

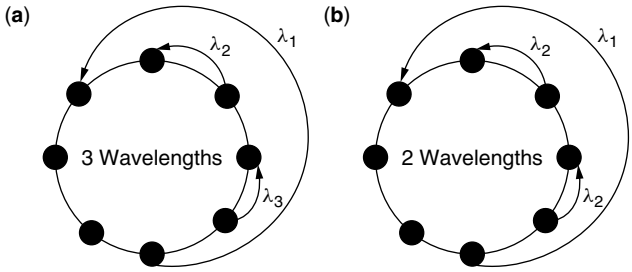


Figure 12. Two possible wavelength assignments for three calls on a ring: (**a**) bad and (**b**) good assignments.

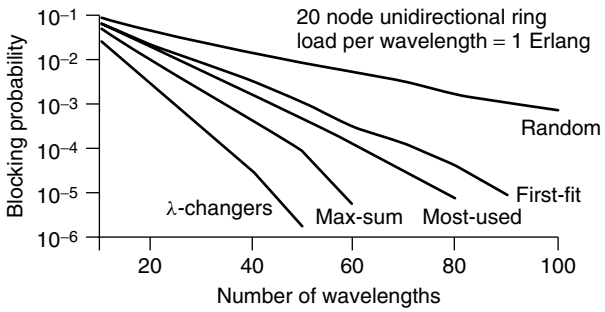


Figure 13. Performance of wavelength assignment algorithms in a ring network.

Figure 13 compares the performance of some proposed wavelength assignment algorithms. The simplest algorithm is to randomly select a wavelength from among the available wavelengths along the path. Clearly such an algorithm would be very inefficient, and, as can be seen from the figure, the random algorithm results in the highest blocking probability. A first-fit heuristic assigns the first available (i.e., lowest index number) wavelength that can accommodate the call. The most frequently used heuristic assigns the wavelength that is used on the most number of fibers in the network and lastly, the max-sum algorithm assigns the wavelength that maximizes the number of paths that can be supported in the network after the wavelength has been assigned [16]. All of these algorithms attempt to pack the wavelengths as much as possible, leaving free wavelengths open for future calls. Also shown in Fig. 12 is the blocking probability that results when wavelength changers are used. This represents an upper bound on the performance of any wavelength assignment algorithm. The significance of this illustration is that a good wavelength assignment algorithm can result in a blocking probability that is nearly as low as if wavelength changers are employed. Hence, a significant reduction in network costs can be obtained by using a good wavelength assignment algorithm.

Wavelength assignment was the first fundamental architectural problem in the design of all-optical networks. It has received much attention in the literature, and remains an active area of research. Beyond wavelength assignment, other important areas of research include the use of wavelength conversion (e.g., where wavelength converters should be deployed), the use of optical switching, and mechanisms for providing protection from

failures. While a meaningful discussion of these topics is beyond the scope of this article, Ref. 1 provides a recent overview on most of these topics.

5. JOINT OPTICAL AND ELECTRONIC NETWORKS

Since all-optical networks are not likely to become a reality in the near future, the current trend in networking is to design networks that use a combination of optical and electronic techniques. A simple example would be a SONET-over-WDM network where the nodes in a SONET ring are connected via wavelengths rather than point-to-point fiber links. As we explained in the introduction, this use of WDM transmission is beneficial because it both reduces network cost and increases network capacity due to the large number of wavelengths. In fact, using WDM at the optical layer also introduces an additional flexibility in the design of the network.

Consider, for example, the networks in Fig. 14, where the optical topology consists of optical nodes [e.g., optical switches or ADMs (add/drop multiplexes)] that are connected via fiber and the electronic topology consists of electronic nodes (e.g., SONET multiplexers) that are connected using electronic links. Without WDM, the electronic topology shown in the figure cannot possibly be realized on the optical topology because the optical topology does not have a fiber link between nodes 1 and 3. However, with WDM an electronic link can be established between nodes 1 and 3 using a wavelength that is routed through node 2. The optical switch (or ADM) at node 2 can be configured to pass that wavelength through to node 3, creating a virtual link between nodes 1 and 3. This approach allows for various electronic topologies to be realized on optical topologies that do not necessarily have the same structure. Electronic nodes can be connected via wavelengths that are routed on the optical topology.

This approach can be used to realize a variety of electronic networks, such as ATM, IP, or SONET [17]. The connectivity of the electronic nodes determines the required wavelength connection that must be established. In other words, each link in the electronic topology requires a wavelength connection (also referred to as *lightpath*) between the optical nodes. In order to realize a particular electronic topology, the corresponding set of wavelength connections must be realized on the optical topology. This very practical problem leads to another version of the routing and wavelength assignment (RWA) problem known as the *batch RWA*. Given a set of lightpaths that must be established, a RWA must be found such that each lightpath must use the same wavelengths along its route from the source to the destination (assuming no wavelength conversion) and no two lightpaths can use the same wavelength on a given link. This problem is closely related to the well-known NP-complete graph coloring problem, and in fact Chlamtac et al. [17] showed that the static RWA problem is indeed NP-complete by suitable transformation from graph coloring.

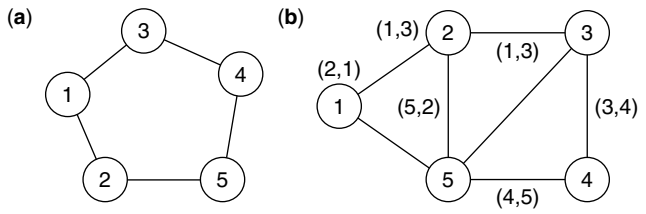


Figure 14. The electronic (**a**) and optical (**b**) topologies of a network.

WDM allows the electronic (logical) topology to be different from the physical topology over which it is implemented. This ability created the interesting opportunity for "logical topology design"; that is, given the traffic demand between the different nodes in the network, what is the best logical topology for supporting that demand. For example, suppose that one is to implement a ring logical topology as in Fig. 14. If a large amount of traffic is being carried between nodes 1 and 4, and virtually no traffic between 1 and 3, it makes more sense to connect the ring in the order 1–4–3 rather than 1–3–4. In this way, the length of the path that the traffic must traverse is reduced, and consequently, the load on the links is also reduced. Designing logical topologies for WDM networks has typically been formulated as an Integer Programming problem, solutions to which are obtained using a variety of search heuristics [18,19]. Furthermore, with configurable WDM nodes (e.g., wavelength switches), it is even possible to reconfigure the logical topology in response to changes in traffic conditions [20].

6. FUTURE DIRECTIONS

Since their inception, the premise of optical networks has been to eliminate the electronic bottleneck. Yet, so far, all optical networks have failed to emerge as a viable alternative to electronic networks. This can be attributed to the tremendous increase in the processing speeds and capacity of electronic switches and routers. Nonetheless, the enormous capacity and configurability of WDM can be used to reduce the cost and complexity of the electronic network, paving the way to even faster networks.

While in the near term optical networking will be used only as a physical layer beneath the electronic network, researchers are aggressively pursuing all-optical packet-switched networks. Such networks may first emerge in the local area, where a broadcast architecture can be used without the need for optical packet switching. As all-optical packet switching technology matures, all-optical networks may become viable. However, as of this writing, it is very unclear whether truly all-optical networks will ever become a reality.

BIOGRAPHY

Eytan Modiano received his B.S. degree in electrical engineering and computer science from the University of Connecticut at Storrs in 1986 and his M.S. and Ph.D. degrees, both in electrical engineering, from the University of Maryland, College Park, Maryland, in 1989 and 1992, respectively. He was a Naval Research Laboratory fellow between 1987 and 1992 and a National Research Council postdoctoral fellow during 1992–1993, while he was

conducting research on security and performance issues in distributed network protocols.

Between 1993 and 1999 he was with the Communications Division at MIT Lincoln Laboratory where he designed communication protocols for satellite, wireless, and optical networks and was the project leader for MIT Lincoln Laboratory's Next Generation Internet (NGI) project. Since 1999, he has been a member of the faculty of the Aeronautics and Astronautics Department and the Laboratory for Information and Decision Systems (LIDS) at MIT, where he conducts research on communication networks and protocols with emphasis on satellite and hybrid networks, and high speed optical networks.

BIBLIOGRAPHY

1. R. Ramaswami and K. Sivarajan, *Optical Networks*, Morgan Kaufmann, San Francisco, CA, 1998.
2. I. P. Kaminow, A wideband all-optical WDM network, *IEEE JSAC/JLT* **14**(5): 780–799 (June 1996).
3. E. Modiano, WDM-based packet networks, *IEEE Commun. Mag.* **37**(3): 130–135 (March 2000).
4. B. Mukherjee, WDM-based local lightwave networks part I: Single-hop systems, *IEEE Network* **6**(3): 12–27 (May 1992).
5. G. N. M. Sudhakar, M. Kavehrad, and N. D. Georganas, Access Protocols for passive optical star networks, *Comput. Networks ISDN Syst.* 913–930 (1994).
6. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
7. N. Mehravari, Performance and protocol improvements for very high-speed optical local area networks using a star topology, *IEEE/OSA J. Lightwave Technol.* **8**(4): 520–530 (April 1990).
8. E. Modiano and R. Barry, A novel medium access control protocol for WDM-based LAN's and access networks using a master/slave scheduler, *IEEE J. Lightwave Technol.* **18**(4): 2–12 (April 2000).
9. E. Modiano, Random algorithms for scheduling multicast traffic in WDM broadcast-and-select networks, *IEEE/ACM Trans. Network.* **7**(3): 425–434 (June 1999).
10. V. Chan, K. Hall, E. Modiano, and K. Rauchenbach, Architectures and technologies for optical data networks, *J. Lightwave Technol.* **16**(12): 2146–2186 (December 1998).
11. R. Barry and P. Humblet, Models of blocking probability in all-optical networks with and without wavelength changers, *IEEE J. Select. Areas Commun.* **14**(5): 858–867 (June 1996).
12. S. Subramanian, M. Azizoglu, and A. Somani, Connectivity and sparse wavelength conversion in wavelength-routing networks, *IEEE INFOCOM* 148–155 (1996).
13. R. Ramaswami and G. Sasaki, Multiwavelength optical networks with limited wavelength conversion, *IEEE INFOCOM* 490–499 (1997).
14. R. Ramaswami and K. N. Sivarajan, Routing and wavelength assignment in all-optical networks, *IEEE/ACM Trans. Network.* 489–500 (Oct. 1995).
15. S. Subramanian and R. Barry, Wavelength assignment in fixed routing wdm networks, ICC '97, Montreal, June 1997.
16. I. Chlamtac, A. Ganz, and G. Karmi, Lightpath communications: An approach to high-bandwidth optical WANs, *IEEE Trans. Commun.* **40**(7): 1171–1182 (July 1992).
17. J. Bannister, L. Fratta, and M. Gerla, Topological design of WDM networks, *Proc. Infocom '90*, Los Alamos, CA, 1990.
18. R. Ramaswami and K. Sivarajan, Design of logical topologies for wavelength routed optical networks, *IEEE J. Select. Areas Commun.* **14**(5): 840–851 (June 1996).
19. J. P. Labourdette and A. Acampora, Logically rearrangeble multihop lightwave networks, *IEEE Trans. Commun.* **39**(8): 1223–1230 (Aug. 1991).
20. A. Narula-Tam and E. Modiano, Dynamic load balancing in WDM packet networks with and without wavelength constraints, *IEEE J. Select. Areas Commun.* **18**(10): 1972–1979 (Oct. 2000).

WAVELETS: A MULTISCALE ANALYSIS TOOL

HAMID KRIM
ECE Department, North Carolina State University
Centennial Campus
Raleigh, North Carolina

One's first and most natural reflex when presented with an unfamiliar object is to carefully look it over, and hold it up to the light to inspect its different facets in the hope of recognizing it, or of at least relating any of its aspects to a more familiar and well-known entity. This almost innate strategy pervades all science and engineering disciplines.

Physical phenomena (e.g., earth vibrations) are monitored and observed by way of measurement in the form of temporal and/or spatial data sequences. Analyzing such data is tantamount to extracting information useful to further understand the underlying process (e.g., frequency and amplitude of vibrations may be an indicator for an imminent earthquake). Visual or manual analysis of typically massive amounts of acquired data (e.g., in remote sensing) are impractical, causing one to resort to adapted mathematical tools and analysis techniques to better cope with potential intricacies and complex structure of the data. Among these tools figure a variety of functional transforms (e.g., Fourier transform) that in many cases may facilitate and simplify an analytical track of a problem, and frequently (and just as importantly) provide an alternative view of, and a better insight into, the problem. (This, in some sense, is analogous to exploring and inspecting data under a "different light.") An illustration of such a "simplification" is shown in Fig. 1, where a rather intricate signal $x(t)$ shown in the leftmost figure may be displayed or viewed in a different space as two elementary tones. In Fig. 2, a real bird chirp is similarly displayed as a fairly rich signal which, when considered in an appropriate space, is reduced and "summarized" to a few "atoms" in the time–frequency (TF) representation. Transformed signals may formally be viewed as *convenient* representations in a different domain that is itself described by a set of vectors/functions $\{\phi_i(t)\}_{i=\{1,2,\ldots,N\}}$. A contribution of a signal $x(t)$ along a direction "$\phi_i(t)$" (its projection) is given by the following inner product

$$C_i(x) = < x(t), \phi_i(t) > = \int_{-\infty}^{\infty} x(t)\phi_i(t)\,dt \qquad (1)$$

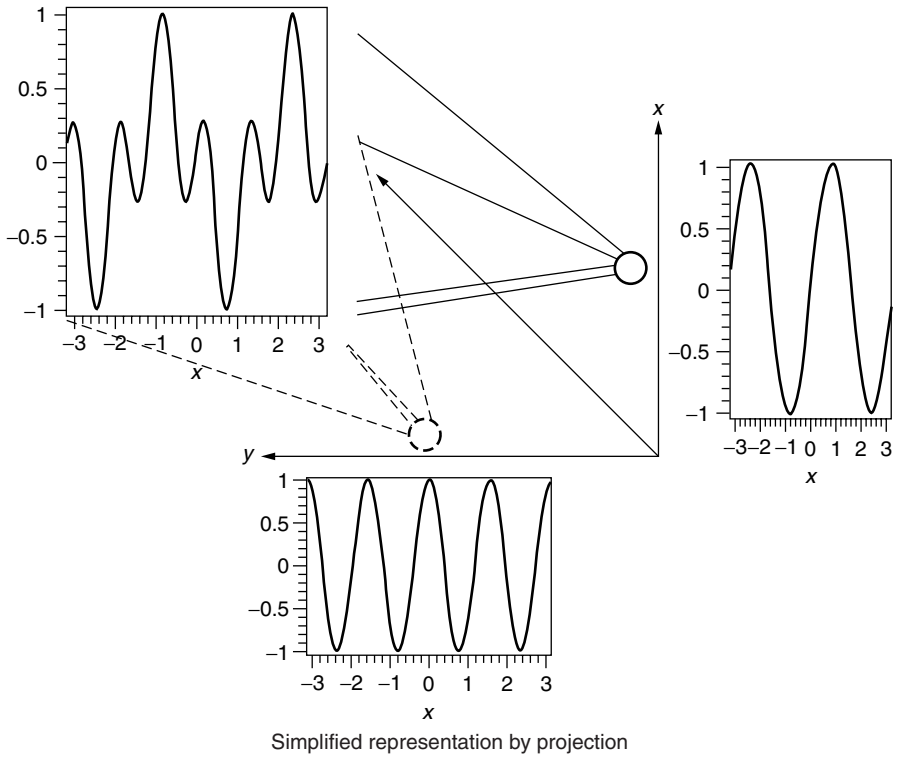


Figure 1. A canonical function-based projection to simplify a representation.

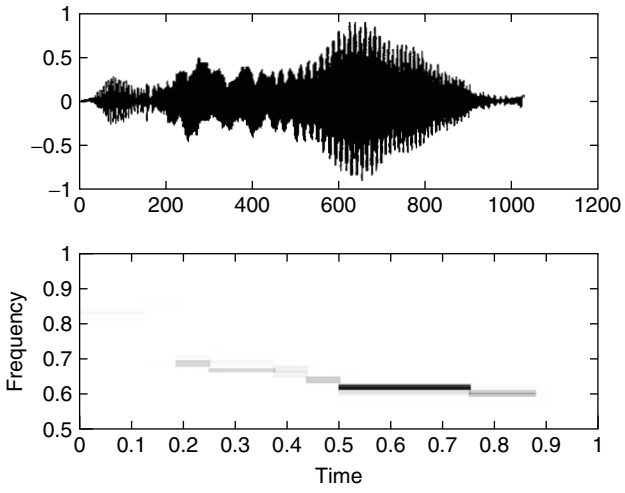


Figure 2. Bird chirp with a simplified representation in an appropriate basis.

where the compact notation $< \cdot, \cdot >$ for an inner product is used. The choice of functions in the set is usually intimately tied to the nature of information that we are to extract from the signal of interest. A Fourier function for example, is specified by a complex exponential "$e^{j\omega t}$" and reflects the harmonic nature of a signal, and provides it with a simple and an intuitively appealing interpretation as a "tone." Its spectral (frequency) representation is a Dirac impulse that is well localized in frequency. A Fourier function is hence well adapted to represent a class of signals with a fairly well localized "spectrum," and is highly inadequate for a transient signal that, in contrast, tends to be temporally well localized. This thus calls for a different class of analyzing functions, namely, those with good temporal compactness. A relatively recent introduction of a function class that balances between the two extremes will be the focus of this chapter. The wavelet transform, by virtue of its mathematical properties, indeed provides such an analysis framework that in many ways is complementary to that of the Fourier transform. As the goal of this article is of a tutorial nature, we cannot help but provide a somewhat high-level exposition of this theoretical framework while attempting to provide a working knowledge of the tool as well as its potential for applications in information sciences in general.

The remainder of this article is organized as follows. In Section 1 we review some of the background starting with a generic functional basis with illustrations from the familiar Fourier basis. In Section 2 we discuss the wavelet transform. In Section 3 we define a multiscale analysis based on a wavelet basis and elaborate on their properties and their implications, as well as on their applications. In Section 4 we discuss a specific estimation technique that is believed to be sufficiently generic and general to be useful in a number of different applications of information sciences in general and of signal processing and communications in particular.

1. SIGNAL REPRESENTATION IN FUNCTIONAL BASES

As noted above, a proper selection of a set of analyzing functions heavily impacts the resulting representation of an observed signal in the dual space, and hence the resulting insight as well.

When considering finite-energy signals [these are also said to be $L^2(\mathbb{R})$]

$$\int_{-\infty}^{\infty} |x(t)|^2 \, dt < \infty, \tag{2}$$

a convenient functional vector space is one that is endowed with a norm inducing inner product, namely, a Hilbert space, which will also be our assumed space throughout.

Definition 1. *A vector space endowed with an inner product, which in turn induces a norm, and such that every sequence of functions $x_n(t)$ whose elements are asymptotically close is convergent, (this convergence is in the Cauchy sense whose technical details are deliberately omitted to streamline the flow of the article) is called a Hilbert space.*

Remark. An $\mathrm{L}^2(\mathbb{R}^d), d = 1, 2$ space is a Hilbert space.

Definition 2. *A set of vectors or functions $\{\phi_i(t)\}, i = 1, \ldots,$ which are linearly independent and which span a vector space, is called a basis. If in addition these elements are orthonormal, then the basis is said to be an orthonormal basis.*

Among the desirable properties of a selected basis in such a space are *adaptivity* and a *capacity* to preserve and reflect some key signal characteristics (e.g., signal smoothness). Fourier bases have in the past been at the center of all science and engineering applications. They have in addition, provided an efficient framework for analyzing periodic as well as nonperiodic signals with dominant modes.

1.1. Fourier Bases

If a canonical function of a selected basis is $\phi(t) = e^{j\omega t}$, where $\omega \in \mathbb{R}$, then we speak of a Fourier basis that yields a Fourier series (FS) for periodic signals and a Fourier transform (FT) for aperiodic signals.

A FT essentially measures the spectral content of a signal $x(t)$ across all frequencies ω:

$$\hat{X}(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} \, dt. \tag{3}$$

The latter, also referred to as the *Fourier integral*, is defined for a broad class of signals [1], and may also be specialized to derive the Fourier series (FS) of a periodic signal. This is easily achieved by noting that a periodic signal $\tilde{x}(t)$ may be evaluated over a period T, which in turn leads to

$$\tilde{x}(t) = x(t + T) = \sum_{n=-\infty}^{\infty} \alpha_n^x e^{jn\omega_0 t} \tag{4}$$

with $\omega_0 = 2\pi/T$ and $\alpha_n^x = 1/T \int_0^T x(t) e^{-jn\omega_0 t} \, dt$. Because it is an orthonormal transform, the FT enjoys a number of interesting properties [2], including the energy preservation in the dual space (transform domain):

$$\int_{-\infty}^{\infty} |x(t)|^2 \, dt = \int_{-\infty}^{\infty} |\hat{X}(f)|^2 \, df.$$

This is referred to as the *Plancherel–Parseval property* [2].

In applications, however, $x(t)$ is measured and sampled at discrete times, requiring that the aforementioned transform be extended to obtain a spectral representation that closely approximates the theoretical truth and that remains just as informative. Toward that end, we proceed to define a discrete-time Fourier transform (DFT) as

$$\hat{X}(e^{j\omega}) = \sum_{i=-\infty}^{\infty} x(i) e^{-j\omega i} \tag{5}$$

This expression may also be extended for finite-observation (finite-dimensional) signals via the fast Fourier transform [2]. While these transforms have been and remain crucial in many applications, they show limitations in problems requiring a good "time–frequency" localization as often encountered in transient analysis. This may easily be understood by reinterpreting Eq. (5) as a weighted transform where each of the time samples is equally weighted in the summation for $\hat{X}(e^{j\omega})$. Gabor [3] first proposed to use a different weighting window leading to the so-called windowed FT, which served well in many practical applications [4].

1.2. Windowed Fourier Transform

As just mentioned, when our goal is to analyze very local features, such as those present in transient signals, for instance, it then makes sense to introduce a focusing window as follows

$$W_{\mu,\omega}(t) = e^{j\omega t} W(t - \mu),$$

where $\|W_{(\mu,\omega)}\| = 1 \, \forall (\mu, \omega) \in \mathbb{R}^2$ and where $W(\cdot)$ is typically a smooth function with a compact support. This yields the following parameterized transform:

$$\hat{X}_W(\omega, \mu) = \langle x(t), W_{\mu,\omega}(t) \rangle. \tag{6}$$

The selection of a proper window is problem-dependent and is ultimately resolved by the desired spectrotemporal tradeoff which is itself constrained by the Heisenberg uncertainty principle [4,5]. From the TF (time–frequency) distribution perspective, and as discussed by Gabor [3] and displayed in Fig. 3, the Gaussian window may be shown to have minimal temporal as well as spectral support of any other function. It hence represents the best compromise between temporal and spectral resolutions. Its numerical implementation entails a discretization of the modulation and translation parameters and results in a uniform partitioning of the TF plane as illustrated in Fig. 4. Different windows result in various TF distribution of elementary atoms, favoring either temporal or spectral resolution as may be seen for the different windows in Fig. 5. While representing an optimal time–frequency compromise, the uniform coverage of the TF plane by the Gabor transform falls short of adequately resolving a signal whose components are spectrally far apart. This may easily and convincingly be illustrated by the study case in Fig. 6, where we note the number of cycles that may be enumerated within a window of fixed time width. It is readily

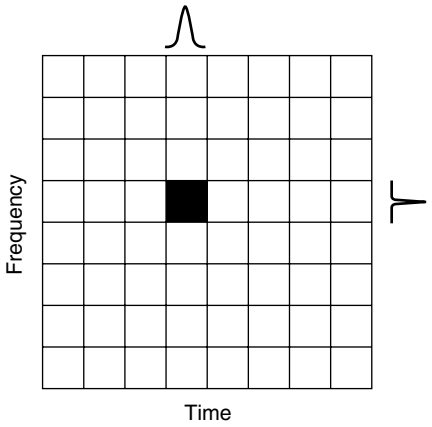


Figure 3. A Gaussian waveform results in a uniform analysis in the time–frequency plane.

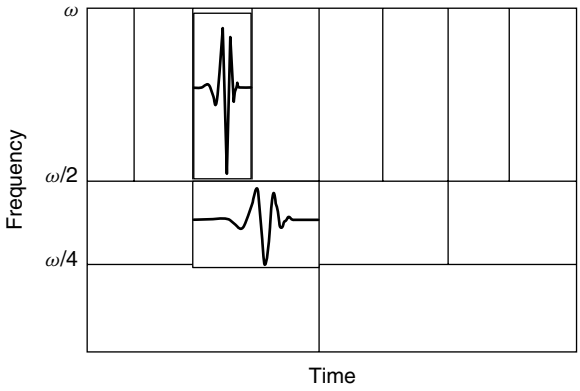


Figure 4. A time–frequency tiling of dyadic wavelet basis by a proper subsampling of a wavelet frame.

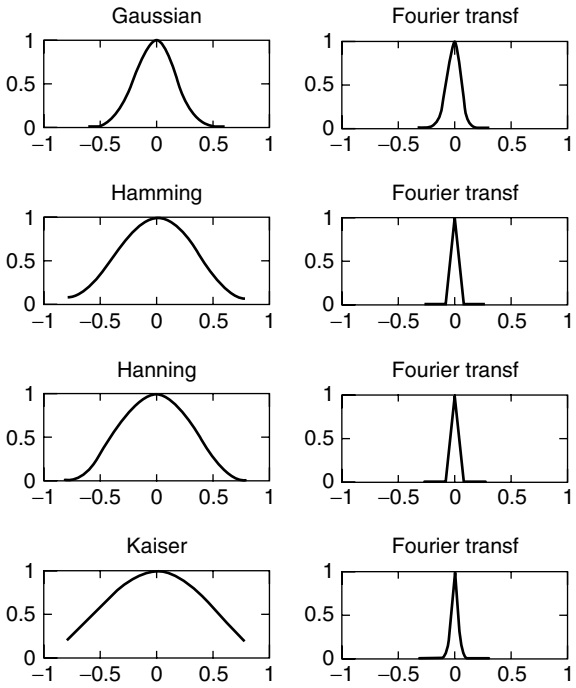


Figure 5. Tradeoff resulting from windows.

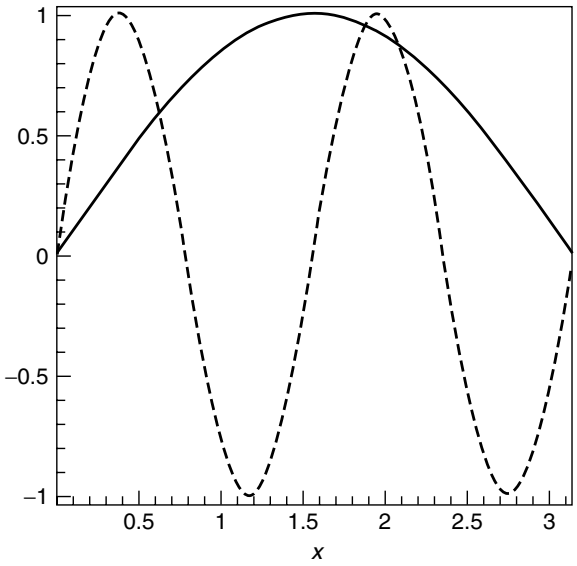


Figure 6. Time windows and frequency tradeoff.

seen that while the selected window (shown grid) may be adequate for one fixed frequency component, it is inadequate for another lower frequency component. An analysis of a spectrum exhibiting such a time-varying behavior is ideally performed by way of a frequency-dependent time window as we elaborate in the next section. The wavelet transform described next, offers a highly adaptive window that is of compact support, and that, by virtue of its dilations and translations, covers different spectral bands at all instants of time.

2. WAVELET TRANSFORM

Much like the FT, the WT is based on an elementary function, which is well localized in time and frequency. In addition to a compactness property, a function has to satisfy a set of properties to be admissible as a wavelet. The first fundamental property is stated next.

Definition 3. *A wavelet is a finite-energy function* $\psi(\cdot)$ *[i.e.,* $\psi(\cdot) \in L^2(\mathbb{R})$*] with zero mean [4,6,7]*:

$$\int_{-\infty}^{\infty} \psi(t)\, dt = 0. \tag{7}$$

Commonly normalized so that $\|\psi\| = \int |\psi|^2\, dt = 1$, it also constitutes a fundamental building block in the construction of functions (atoms) spanning the time–frequency plane by way of dilation and translation parameters. We hence write

$$\psi_{\mu,\xi}(t) = \frac{1}{\sqrt{\xi}} \psi\left(\frac{t-\mu}{\xi}\right)$$

where the scaling factor $\xi^{1/2}$ ensures an energy invariance of $\psi_{\mu,\xi}(t)$ over all dilations $\xi \in \mathbb{R}^+$ and translations $\mu \in \mathbb{R}$. With such a function in hand, and toward mitigating the limitation of a windowed FT, we proceed to define a

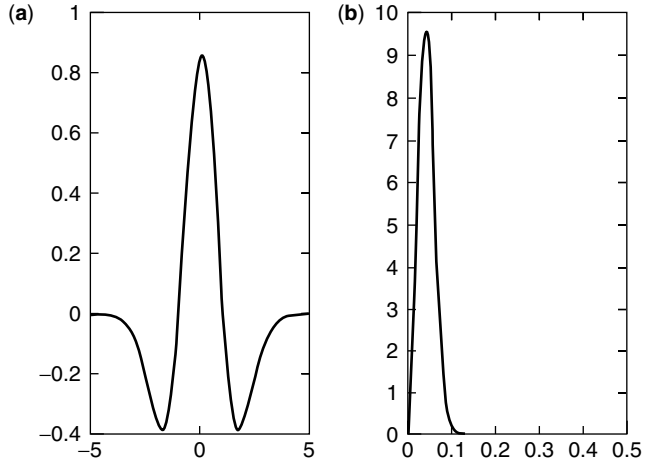


Figure 7. Admissible Mexican hat wavelet: (**a**) Mexihat wavelet; (**b**) spectrum of Mexihat.

Wavelet transform (WT) with such a capacity. A scale-dependent window with good time localization properties as shown in Fig. 7 yields a transform for $x(t)$ given by

$$\mathcal{W}_x(\mu,\xi) = \int_{-\infty}^{\infty} x(t)\frac{1}{\sqrt{\xi}}\psi^*\left(\frac{t-\mu}{\xi}\right)dt \tag{8}$$

where the asterisk denotes complex conjugate. This is, of course, in contrast to the Gabor transform whose window width remains constant throughout. A time–frequency plot of a continuous wavelet transform is shown in Fig. 8 for a corresponding $x(t)$.

2.1. Inverting the Wavelet Transform

Similar to the weighted FT, the WT is a redundant representation, which, with a proper normalization factor, leads to a reconstruction formula

$$x(t) = \frac{1}{C_\psi}\int_0^{\infty}\int_{-\infty}^{\infty}\mathcal{W}_x(\mu,\xi)\frac{1}{\sqrt{\xi}}\psi\left(\frac{t-\mu}{\xi}\right)\frac{d\xi}{\xi^2}\,d\mu \tag{9}$$

with $C_\psi = \int_0^{+\infty}[\psi(\hat{\omega})/\omega]\,d\omega$.

While the direct and inverse WT have been successfully applied in a number of different problems [4], their computational cost due to their continuous and redundant nature is considered as a serious drawback. An immediate and natural question arises about a potential reduction in the computational complexity and hence in the WT redundancy. Clearly, this has to be carefully carried out to guarantee a sufficient coverage of the TF plane and thereby ensure a proper reconstruction of a transformed signal. On discretizing the scale and dilation parameters, a dimension reduction relative to a continuous transform is achieved. The desired adaptivity of the window width with varying frequency naturally results by selecting a geometric variation of the dilation parameter $\xi = \xi_0^m, m \in \mathbb{Z}$ (*set of all positive and negative integers*). To obtain a systematic and consistent coverage of the TF plane, our choice of the translation parameter μ should be in congruence with the fact that at any scale m, the coverage of the whole line $\mathbb{R}$ (e.g.) is complete, and the translation parameter be in step with the chosen wavelet $\psi(t)$, that is, $\mu = n\mu_0\xi_0^m$ with $n \in \mathbb{Z}$. This hence gives the following scale and translation adaptive wavelet

$$\psi(t)_{m,n} = \xi_0^{-m/2}\psi\left(\frac{t}{\xi_0^m} - n\mu_0\right) \tag{10}$$

where the factor $\xi_0^{-m/2}$ ensures a unit energy function. This reduction in dimensionality yields a redundant discrete wavelet transform endowed with a structure that lends itself to an iterative and fast inversion or reconstruction of a transformed signal $x(t)$. The set of resulting wavelet coefficients $\{< x(t), \psi_{mn}(t) >\}_{(m,n)\in Z^2}$ completely characterizes $x(t)$, and hence leads to a stable reconstruction [4,5] if $\forall x(t) \in L^2(\mathbb{R})$ (or finite energy signal) the following condition on the energy holds:

$$A\|x(t)\|^2 \le \sum_{m,n}|\langle x(t), \psi_{mn}(t)\rangle|^2 \le B\|x(t)\|^2. \tag{11}$$

Such a set of functions $\{\psi_{mn}(t)\}_{(m,n)\in Z^2}$ then constitutes a frame. The energy inequality condition intuitively suggests that the redundancy should be controlled to avoid

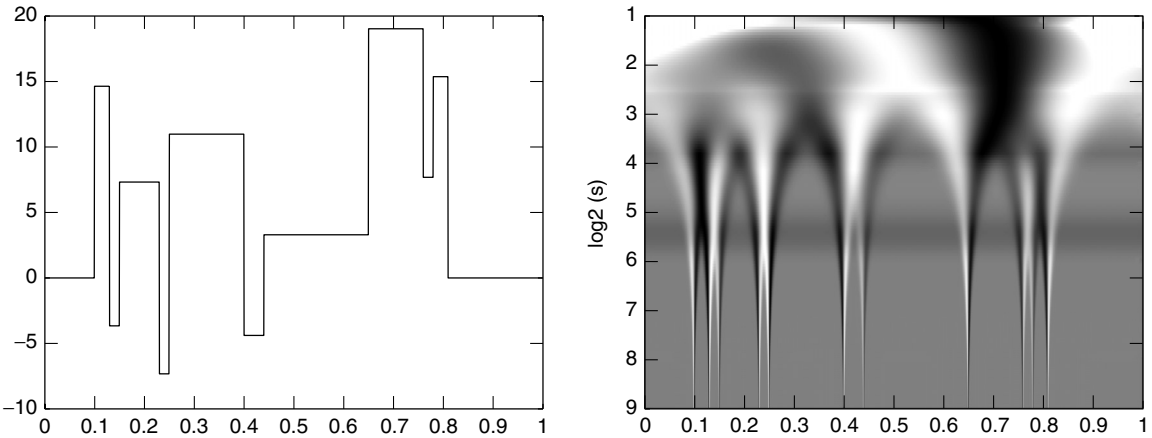


Figure 8. Continuous signal with corresponding wavelet transform with cones of influence around singularities.

instabilities in the reconstruction (i.e., too much redundancy makes it more likely that any perturbation of coefficients will yield an unstable/inconsistent reconstruction). If the frame bounds A and B are equal, the corresponding frame is said to be tight, and if furthermore $A = B = 1$, it is an orthonormal basis. Note, however, that for any $A \neq 1$ a frame is not a basis. In some cases, the frame bounds specify a redundancy ratio that may help guide one in inverting a frame representation of a signal, since a unique inverse does not exist. An efficient computation of an inverse (for reconstruction), or more precisely of a pseudoinverse, may only be obtained by a judicious manipulation of the reconstruction formula [4,5]. This is, in a finite-dimensional setting, similar to a linear system of equations with a rank-deficient matrix whose column space has an orthogonal complement (the union of the two subspaces yields all the Hilbert space), and hence whose inversion is ill conditioned. The size of this space is determined by the order of the deficiency (number of linearly dependent columns), and explains the potential for instability in a frame projection-based signal reconstruction [4]. This type of "effective rank" utilization is encountered in numerical linear algebra applications (e.g., signal subspace methods [8], model order identification,). The close connection between the redundancy of a frame and the rank deficiency of a matrix suggests that solutions available in linear algebra [9] may provide insight in solving our frame-based reconstruction problem. A well-known iterative solution to a linear system and based on a gradient search solution has been described [10] (and in fact coincides with a popular iterative solution to the frame algorithm for reconstructing a signal) for solving

$$\mathbf{f} = \mathbf{L}^{-1}\mathbf{b}$$

where $\mathbf{L}$ is a matrix operating on $\mathbf{f}$ and $\mathbf{b}$ is the data vector. Starting with an initial guess $\mathbf{f}_0$ and iterating on it leads to

$$\mathbf{f}_n = \mathbf{f}_{n-1} + \alpha(\mathbf{b} - \mathbf{L}\mathbf{f}_{n-1}) \tag{12}$$

When α is appropriately selected, the latter iteration may be shown to yield [4]

$$\lim_{n \to +\infty} \mathbf{f}_n = \mathbf{f}. \tag{13}$$

Numerous other good solutions have also been proposed and described in detail in the literature [4,5,11,12].

3. WAVELETS AND MULTIRESOLUTION ANALYSIS

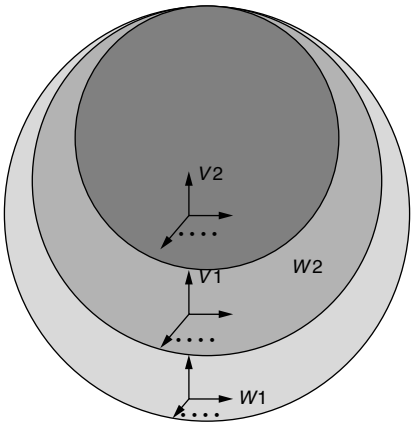
While a frame representation of a signal is of lower dimension than that of its continuous counterpart, a complete elimination of redundancy is possible only by orthonormalizing a basis. A proper construction of such a basis ensures orthogonality within scales as well as across scales. The existence of such a basis with a dyadic scale progression was first shown, and an explicit construction was given by Meyer and Daubechies [5,13]. The connection to subband coding discovered by Mallat [14] resulted in a numerically stable and efficient implementation that helped propel wavelet analysis at the forefront of computational sciences. (See Daubechies and Meyer [15] for a comprehensive historical as well as technical development [5], which also led to Daubechies' celebrated orthonormal wavelet bases.)

To help maintain the smooth flow of this article and achieve the goal of endowing an advanced undergraduate with a working knowledge of the multiresolution analysis (MRA) framework, we give a high-level, albeit comprehensive, introduction, and defer most of the technical details to the sources in the bibliography. For example, the tradeoff in time-frequency resolution advocated earlier lies at the heart of the often technical wavelet design and construction. The balance between the spectral and temporal decay, which constitutes one of the key design criteria, has led to a wealth of new functional bases, and this follows the introduction of the now classical multiresolution analysis framework [4,16]. The pursuit of a more refined analysis arsenal resulted in the nonlinear multiresolution analysis introduced in 1997 and 1998 [17–19].

3.1. Multiresolution Analysis

The MRA theory developed by Mallat and Meyer [13,20], may be viewed as a functional analytic approach to subband signal analysis, which had previously been introduced in and applied to engineering problems [21,22]. The clever connection between an orthonormal wavelet decomposition of a signal and its filtering by a bank of corresponding filters led to an aximonatic formalization of the theory and subsequent equivalence. This as a result, opened up a new wide avenue of research in the development and construction of new functional bases as well as filter banks [7,23–25]. The construction of a telescopic set of nested approximation and detail subspaces $\{V_j\}_{j \in Z}$ and $\{W_j\}_{j \in Z}$ each endowed with an orthonormal basis, as shown in Fig. 9, is a key step of the analysis. An inter- and intrascale orthogonality of the wavelet functions, as noted above, is preserved, with the interscale orthogonality expressed as

$$W_i \perp W_j \, \forall i \neq j. \tag{14}$$



Nested wavelet subspaces and corresponding bases

Figure 9. Hierarchy of wavelet bases.

By replacing the discrete parameter wavelet $\psi_{ij}(t)$ where $(i,j) \in \mathbb{Z}^2$ (*set of all positive and negative 2-tuple integers*) respectively denote the translation and the scale parameters, in Eq. (8) (i.e., $\mu = 2^j i, \xi = 2^J$) we obtain the orthonormal wavelet coefficients as

$$C_j^i(x) = \langle x(t), \psi_{ij}(t) \rangle. \tag{15}$$

The orthogonal complementarity of the scaling subspace and that of the residuals and details amount to synthesizing the following higher-resolution subspace:

$$V_i \oplus W_i = V_{i-1}.$$

Iterating this property, leads to a reconstruction of the original space where the observed signal lies and that, in practice, is taken to be V_0: the observed signal at its first and finest resolution. (This may be viewed as implicitly accepting the samples of a given signal as the coefficients in an approximation space with a scaling function corresponding to that on which the subsequent analysis is based.) The dyadic scale progression has been thoroughly investigated, and its wide acceptance and popularity is due to its tight connection with subband coding whose practical implementation is fast and simple. Other nondyadic developments have also been explored [e.g., 7].

The qualitative characteristics of the MRA we have thus far discussed may be succinctly stated as follows.

Definition 4. *A sequence $\{V_i\}_{i\in Z}$ of closed subspaces of $L^2(\mathbb{R})$ is a multiresolution approximation if the following properties hold [4]*:

- $\forall (j,k) \in \mathbb{Z}^2, x(t) \in V_j \leftrightarrow x(t - 2^j k) \in V_j$
- $\forall j \in \mathbb{Z}, V_{j+1} \subset V_j$
- $\forall j \in \mathbb{Z}, x(t) \in V_j \leftrightarrow x\left(\frac{t}{2}\right) \in V_{j+1}$
- $\lim_{j\to-\infty} V_j = \bigcap_{j=-\infty}^{\infty} V_j = \{0\}$
- $\lim_{j=-\infty} = \overline{\bigcup_{j=-\infty}^{\infty} V_j} = L^2(\mathbb{R})$
- There exists a function $\phi(t)$ such that $\{\phi(t-n)\}_{n\in Z}$ is a Riesz basis of V_0, where the overbar denotes closure of the space.

3.2. Properties of Wavelets

A wavelet analysis of a signal assumes a judiciously preselected wavelet and hence a prior knowledge about the signal itself. As stated earlier, the properties of an analyzing wavelet have a direct impact on the resulting multiscale signal representation. Carrying out a useful and meaningful analysis is hence facilitated by a good understanding of some fundamental wavelet properties.

3.2.1. Vanishing Moments. Recall that one of the fundamental admissibility conditions of a wavelet is that its first moment be zero. This is intuitively understood in the sense that a wavelet focuses on residuals or oscillating features of a signal. This property may in fact be further exploited by constructing a wavelet with an arbitrary number of vanishing moments. We say that a wavelet has n vanishing moments if

$$\int \psi(t) t^i \, dt = 0, i = \{0, 1, \ldots, n-1\} \tag{16}$$

Reflecting a bit on the properties of a Fourier Transform of a function [2], it is easy to note that the number of zero moments of a wavelet reflects the behavior of its Fourier Transform around zero. This property is also useful in applications such as compression, where it is highly desirable to maximize the number of small or negligible coefficients, and preserve only a minimal number of large coefficients. The associated cost with increasing the number of vanishing moments is that of an increased support size for the wavelet, hence that of the corresponding filter [4,5], and hence of some of its localizing potential.

3.2.2. Regularity and Smoothness. The smoothness of a wavelet $\psi(t)$ is important for an accurate and parsimonious signal approximation. For a large class of wavelets (those relevant to applications), the smoothness (or regularity) property of a wavelet, which may also be measured by its degree of differentiability ($d^\alpha \psi(t)/dt^\alpha$) or equivalently by its "Lipschitzity" γ, is also reflected by its number of vanishing moments [4]. The larger the number of vanishing moments, the smoother the function. In applications, such as image coding, a smooth analyzing wavelet is useful for not only compressing the image but for controlling the visual distortion due to errors as well. The associated cost (i.e., some tradeoff is in order) is again a size increase in the wavelet support, which may in turn make it more difficult to capture local features, such as important transient phenomena.

3.2.3. Wavelet Symmetry. At the exception of a Haar wavelet, compactly supported real wavelet are asymmetric around their centerpoint. A symmetric wavelet clearly corresponds to a symmetric filter that is characterized by a linear phase. The symmetry property is important for certain applications where symmetric features are crucial (e.g., symmetric error in image coding is better perceived). In many applications, however, it is generally viewed as a property secondary to those described above. When such a property is desired, truly symmetric biorthogonal wavelets have been proposed and constructed [5,26] with the slight disadvantage of using different analysis and synthesis mother wavelets. In Fig. 10, we show some illustrative cases of symmetric wavelets. Other nearly symmetric functions (referred to as *symlets*) have also been proposed, and a detailed discussion of the pros and cons of each is deferred to Daubechies [5].

3.3. A Filter Bank View: Implementation

As noted earlier, the connection between a MRA of a signal and its processing by a filter bank was not only of intellectual interest, but was of breakthrough proportion for general applications as well. It indeed provided a highly

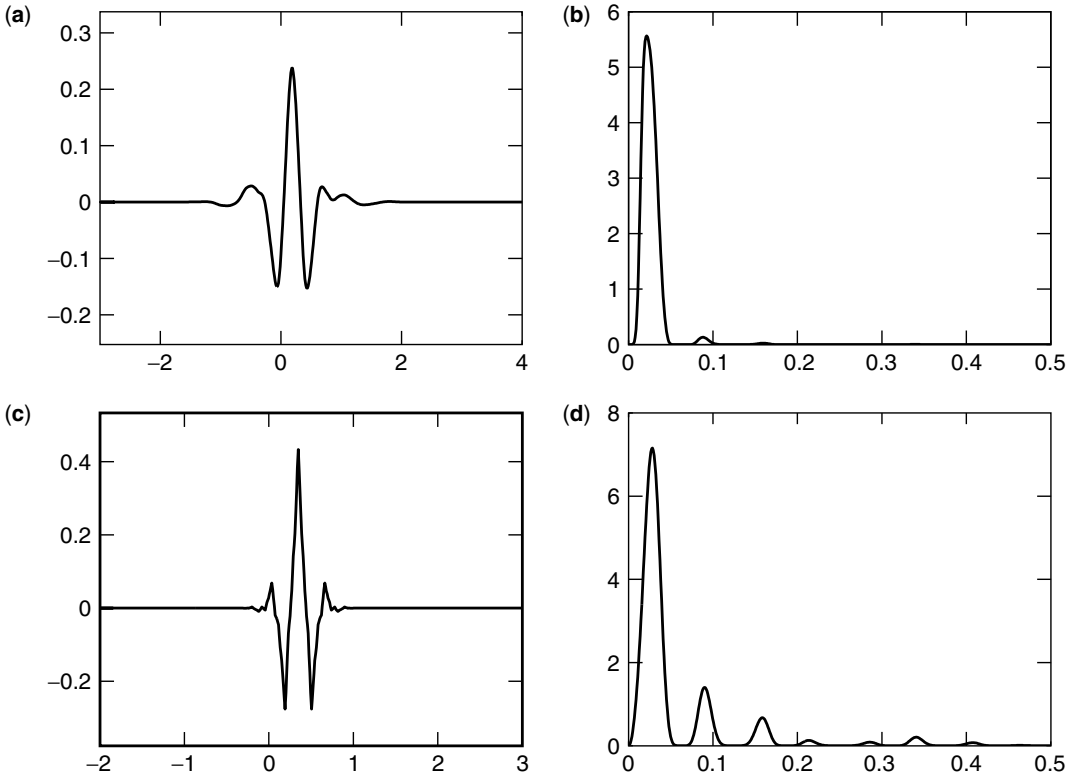


Figure 10. Biorthogonal wavelets preserve symmetry and symlets nearly do: (**a**) symmlet; (**b**) symmlet spectrum; (**c**) spline biorthogonal wavelet; (**d**) spectrum of spline wavelet.

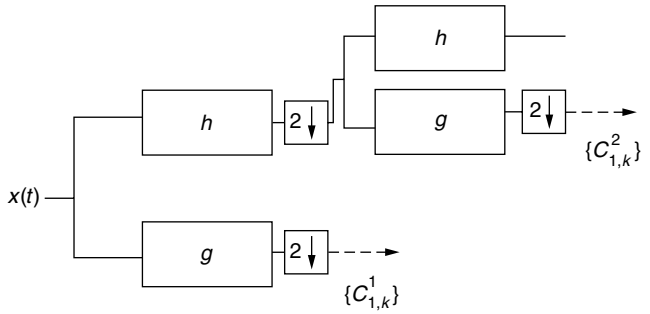


Figure 11. Filter bank implementation of a wavelet decomposition.

efficient numerical implementation for a theoretically very powerful methodology. Such a connection is most easily established by invoking the nestedness property of MR subspaces. Specifically, it implies that if $\phi(2^{-j}t) \in V_j$ and $V_j \subset V_{j-1}$, we can hence write

$$\frac{1}{2^{j/2}}\phi(2^{-j}t) = \frac{1}{2^{(j-1)/2}} \sum_{k=-\infty}^{\infty} h(k)\phi(2^{-j+1}t - k) \tag{17}$$

where $h(k) =< \phi(2^{-j}t), \phi(2^{-j+1}t - k) >$, that is, the expansion coefficient at time shift k. By taking the FT of Eq. (17), we obtain

$$\Phi(2^j\omega) = \frac{1}{2^{1/2}} H(2^{j-1}\omega)\Phi(2^{j-1}\omega) \tag{18}$$

which, when iterated through scales, leads to

$$\Phi(\omega) = \prod_{p=-\infty}^{\infty} \frac{h(2^{-p}\omega)}{\sqrt{2}}\Phi(0). \tag{19}$$

The complementarity of scaling and detail subspaces noted earlier, stipulates that any function in subspace W_j may also be expressed in terms of $V_{j-1} = \text{Span}\{\phi_{j-1,k}(t)\}_{(j,k)\in Z^2}$, or

$$\frac{1}{2^{j/2}}\psi(2^{-j}t) = \sum_{i=-\infty}^{\infty} \frac{1}{2^{(j-1)/2}} g(i)\phi(2^{-j+1}t - i). \tag{20}$$

In the Fourier domain, this leads to

$$\Psi(2^j\omega) = \frac{1}{\sqrt{2}} G(2^{j-1}\omega)\Phi(2^{j-1}\omega) \tag{21}$$

whose iteration also leads to an expression of the wavelet FT in terms of the transfer function $G(\omega)$, as given for $\Phi(\omega)$ in Eq. (19). In light of these equations, it is clear that the filters $\{G(\omega)$. This is illustrated for Daubechies wavelet in Fig. 12. $H(\omega)\}$ may be used to compute functions at successive scales. This in turn implies that the coefficients of any function $x(t)$ may be similarly obtained as they are merely the result of an inner product of the signal of interest $x(t)$ with a basis function:

$$\begin{aligned}\langle x(t), \psi_{ij}(t)\rangle &= \sum \langle x(t), g(k)\frac{1}{2^{(j-1)/2}}\phi(2^{-j+1}(t - 2^{-j}i) - k)\rangle \\ &= m_k g(k)\mathcal{A}^{j+1}_{i-k}(x)\end{aligned} \tag{22}$$

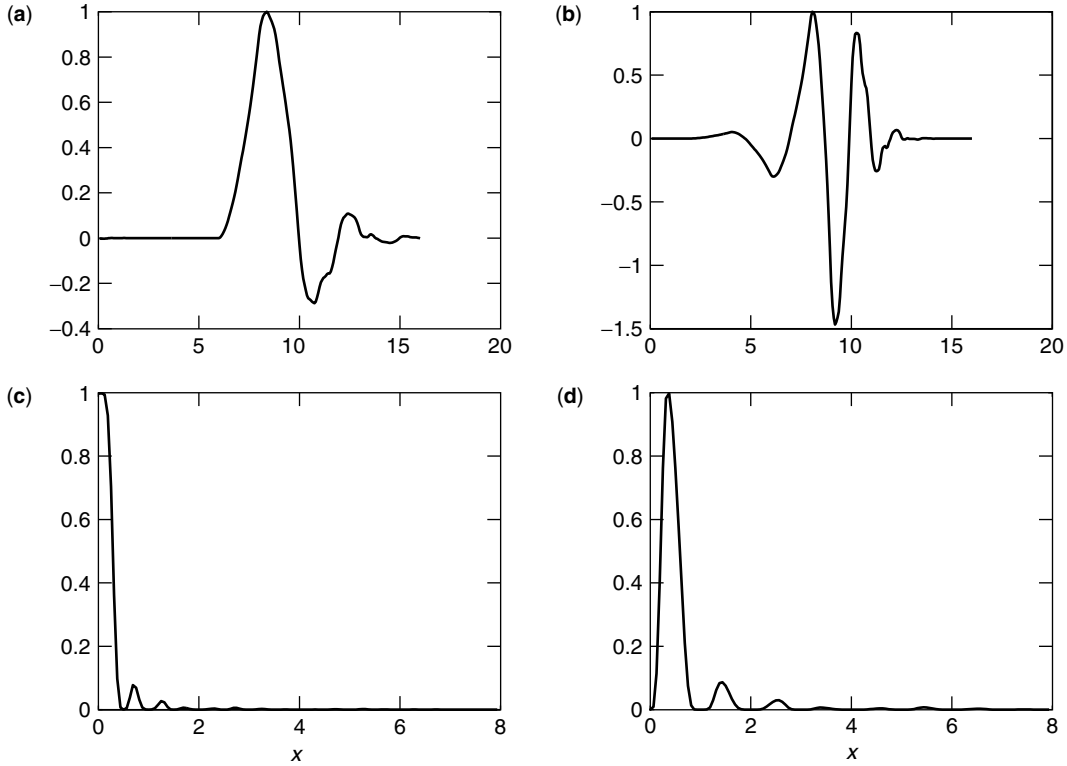


Figure 12. A Daubechies- 8 (D-8) function and its spectral properties: (**a**) D-8 scaling function; (**b**) D-8 wavelet function; (**c**) D-8 scaling Fourier transform; (**d**) D-8 wavelet Fourier transform.

To complete the construction of the filter pair $\{H(\cdot), G(\cdot)\}$, the properties between the approximation subspaces ($\{V_j\}_{j\in Z}$) and detail subspaces ($\{W_j\}_{j\in Z}$) are exploited to derive the design criteria for the discrete filters. Specifically, the same scale orthogonality property between the scaling and detail subspaces in

$$\sum_{k\in Z} \Phi(2^j\omega + 2k\pi)\Psi(2^j\omega + 2k\pi) = 0, \forall j \tag{23}$$

is the first property that the resulting filters should satisfy. For the sake of illustration, fix $j = 1$ and use

$$\sqrt{2}\Phi(2\omega) = H(\omega)\Phi(\omega) \tag{24}$$

Using Eq. (24) together with the orthonormality property of jth scale basis functions $\{\Phi_{jk}(t)\}$ [4], namely, $\sum |\Phi(\omega + 2k\pi)|^2 = 1$, where we separate even and odd terms, yields the first property of one of the so-called conjugate mirror filters:

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1. \tag{25}$$

Using the nonoverlapping property expressed in Eq. (23), together with the evaluations of $\Psi(2\omega)$ and $\Phi(2\omega)$ and making a similar argument to that advanced in the preceding equation yield the second property of conjugate mirror filters:

$$H(\omega)G^*(\omega) + H(\omega + \pi)G^*(\omega + \pi) = 0. \tag{26}$$

The combined structure of the two filters is referred to as "a conjugate mirror filter bank," and their respective impulse responses completely specify the corresponding wavelets (see numerous references, e.g., Ref. 6 for additional technical details). Note that the literature in the MR studies tends to follow one of two possible strategies:

- A more functional analytic approach, which is mostly followed by applied mathematicians/mathematicians and scientists [4,5,16] (we could not possibly do justice to the numerous good texts now available; the author cites only what he is most familiar with.)
- A more filtering-oriented approach widely popular among engineers and some applied mathematicians [7,24,27].

3.4. Refining a Wavelet Basis: Wavelet Packet and Local Cosine Bases

A selected analysis wavelet is not necessarily well adapted to any observed signal. This is particularly the case when the signal is time-varying and has a rich spectral structure.

One approach is to then partition the signal into homogeneous spectral segments and by the same token find it an adapted basis. This is tantamount to further refining the wavelet basis, and resulting in what is referred to as a *wavelet packet basis*. A similar adapted partitioning may be carried out in the time domain by way of an orthogonal local trigonometric basis (sine or cosine). The two formulations are very similar, and the solution to the search for an adapted basis in both cases is resolved in

precisely the same way. In the interest of space, we focus in the following development on only wavelet packets.

3.4.1. Wavelet Packets. We maintained above that a selection of an analysis wavelet function should be carried out in function of the signal of interest. This, of course, assumes that we have some prior knowledge about the signal at hand. While plausible in some cases, this assumption is very unlikely in most practical cases. Yet it is highly desirable to select a basis that might still lead to an adapted representation of an apriorily "unknown" signal. Coifman and Meyer [28] proposed to refine the standard wavelet decomposition. Intuitively, they proposed to further partition the spectral region of the details (i.e., wavelet coefficients) in addition to partitioning the coarse and/or scaling portion as ordinarily performed for a wavelet representation. This, as shown in Fig. 13, yields an overcomplete representation of a signal, in other words, a dictionary of bases with a tree structure. The binary nature of the tree as further discussed below, affords a very efficient search for a basis which is best adapted to a given signal in the set.

Formally, we accomplish a partition refinement of, say a subspace U_j by way of a new wavelet basis which includes both its approximation and its details, as shown in Fig. 13. This construction due to Coifman, and Meyer [28] may be simply understood as one's ability to find a basis

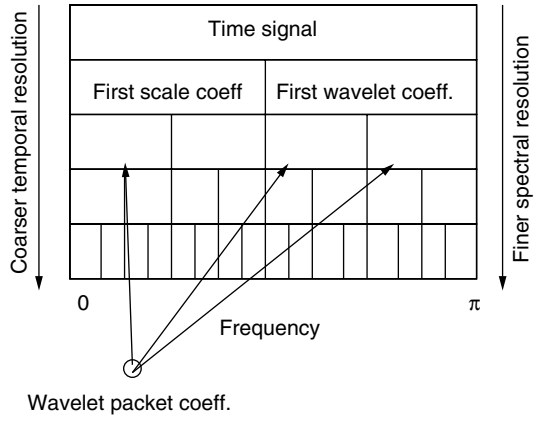


Figure 13. Wavelet packet bases structure.

for all the subspaces $\{V_j\}_{j\in Z}$ and $\{W_j\}_{j\in Z}$ by iterating the nestedness property. This then amounts to expressing two new functions $\tilde{\psi}_j^0(t)$ and $\tilde{\psi}_j^1(t)$ in terms of $\{\tilde{\psi}_{j-1,k}\}_{(j,k)\in Z^2}$, an orthonormal basis of a generic subspace U_{j-1} (which in our case may be either V_{j-1} or W_{j-1})

$$\tilde{\psi}_j^0(t) = \sum_{k=-\infty}^{\infty} h(k)\tilde{\psi}_{j-1}(t - 2^{j-1}k) \tag{27}$$

$$\tilde{\psi}_j^1(t) = \sum_{k=-\infty}^{\infty} g(k)\tilde{\psi}_{j-1}(t - 2^{j-1}k) \tag{28}$$

where $h(\cdot)$ and $g(\cdot)$ are the impulse responses of the filters in a corresponding filter bank and the combined family $\{\tilde{\psi}_j^0(t - 2^j k), \tilde{\psi}_j^1(t - 2^j k)\}_{(j,k)\in Z^2}$ is an orthonormal basis of U_j. This, as illustrated in Fig. 13, is graphically represented by a binary tree where, at each of the nodes reside the corresponding wavelet packet coefficients. The implementation of a wavelet packet decomposition is a straightforward extension of that of a wavelet decomposition, and consists of an iteration of both $H(\cdot)$ and $G(\cdot)$ bands (see Fig. 14) to naturally lead to an overcomplete representation. This is reflected on the tree by the fact that, with the exception of the root and bottom nodes, which have only respectively two children nodes and one ancestor node, each node bears two children nodes and one ancestor node.

3.4.2. Basis Search. To identify the best adapted basis in an overcomplete signal representation, as noted above, we first construct a criterion that, when optimized, will reflect desired properties intrinsic to the signal being analyzed. The earliest proposed criterion applied to a wavelet packet basis search is the so-called entropy criterion [29]. Unlike Shannon's information theoretic criterion, this is additive and makes use of the coefficients residing at each node of the tree in lieu of computed probabilities. The presence of more complex features in a signal necessitates such an adapted basis to ultimately achieve an ideally more parsimonious or succinct representation. As pointed out earlier, when searching for a wavelet packet or local cosine best basis, we typically have a dictionary $\mathcal{D}$ of possible bases with a binary tree structure. Each node (j, j') (where $j \in \{0, \ldots, J\}$ represents the depth and $j' \in \{0, \ldots, 2^j - 1\}$ represents the

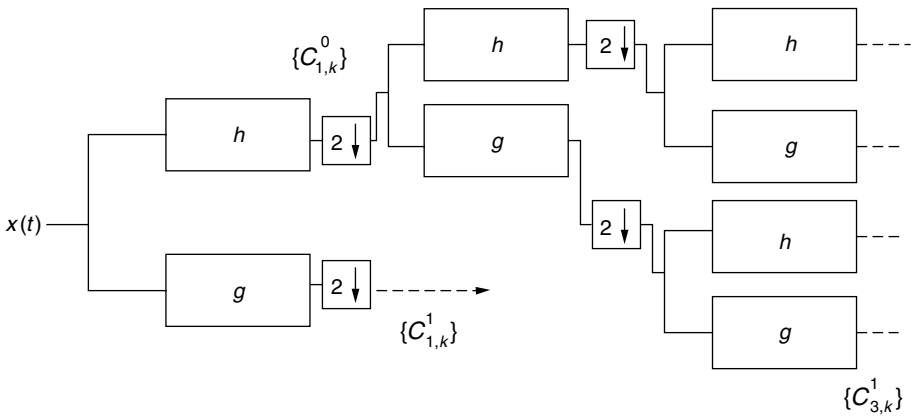


Figure 14. Wavelet packet filter bank realization.

branches on the jth level) of the tree then corresponds to a given orthonormal basis $\mathcal{B}_{j,j'}$ of a vector subspace of $\ell^2(\{1,\ldots,N\})$ [$\ell^2(\{1,\ldots,N\})$ is a Hilbert space of finite-energy sequences]. Since a particular partition $p \in \mathcal{P}$ of $[0,1]$ is composed of intervals $I_{j,j'} = [2^{-j}j', 2^{-j}(j'+1)]$, an orthonormal basis of $\ell^2(\{1,\ldots,N\})$ is given by $\mathcal{B}^p = \cup_{\{(j,j')|I_{j,j'} \in \mathcal{P}\}} \mathcal{B}_{j,j'}$. By taking advantage of the property

$$\text{Span}\{\mathcal{B}_{j,j'}\} = \text{Span}\{\mathcal{B}_{j+1,2j'}\} \overset{\perp}{\oplus} \text{Span}\{\mathcal{B}_{j+1,2j'+1}\} \tag{29}$$

where $\oplus$ denotes a subspace direct sum, we associate to each node a cost $\mathcal{C}(\cdot)$. We can then perform a bottom−up comparison of children versus parent costs (thereby, in effect, eliminating all redundand or inadequate leaves from the tree) and ultimately prune the tree.

Our goal is to then choose the basis that leads to the *best* approximation of $\{x[t]\}$ among a collection of orthonormal bases $\{\mathcal{B}^p = \{\Psi x_i p\}_{1 \le i \le N} | p \in \mathcal{P}\}$, where the term x_i emphasizes that it is adapted to $\{x[t]\}$. Trees of wavelet packet bases studied by Coifman and Wickerhauser [29] are constructed by quadrature mirror filter banks and constitute functions that are well localized in time and frequency. This family of orthonormal bases, partitions the frequency axis into intervals of different sizes, with each set corresponding to a specific wavelet packet basis. Another family of orthonormal bases, studied by Malvar [31], and Coifman and Meyer [28], can be constructed with a tree of windowed cosine functions, and correspond to a division of the time axis into intervals of dyadically varying sizes.

For a discrete signal of size N (e.g., the size of the WP tableau shown in Fig. 13), one can show that a tree of wavelet packet bases or local cosine bases has $P = N(1 + \log_2 N)$ distinct vectors but includes more than $2^{N/2}$ different orthogonal bases. One can also show that the signal expansion in these bases is computed with algorithms that require $O(N \log_2 N)$ operations. Coifman and Wickerhauser [29] proposed that for any signal $\{x[m]\}$ and an appropriate functional $\mathcal{K}(\cdot)$, one finds the best basis $\mathcal{B}^{p_0}$ by minimizing an "additive" cost function

$$\text{Cost}(\mathbf{x}, \mathcal{B}^p) = \sum_{i=1}^{N} \mathcal{K}(|\langle \mathbf{x}, \Psi x_i p \rangle|^2) \tag{30}$$

over all bases. As a result, the basis that results from minimizing this cost function corresponds to the "best" representation of the observed signal. The resulting pruned tree of Fig. 15 bears the coefficients at the remaining leaves and nodes.

4. MR APPLICATIONS IN ESTIMATION PROBLEMS

The computational efficiency of a wavelet decomposition together with all its properties have triggered unprecedented interest in their application in the area of information sciences [e.g., 32−37]. Specific applications have ranged from compression [38−41] to signal or image modeling [42−45], and from signal or image enhancement to communications [46−53]. The literature in statistical applications as a whole has seen an explosive growth, and

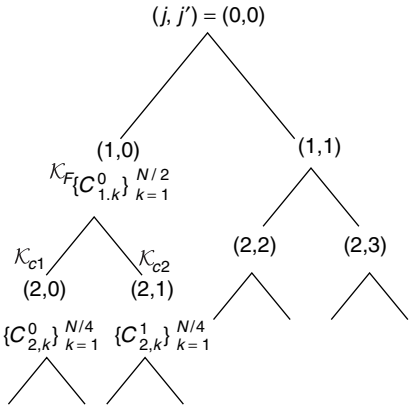


Figure 15. Tree pruning in search for a "best" basis.

in the interest of space, we will focus our discussion on a somewhat broader perspective that may, in essence, be usefully reinterpreted in a number of different instances.

4.1. Signal Estimation Denoising and Modeling

Denoising may be heuristically interpreted as a quest for parsimony of a representation of a signal. Wavelets as described above, have a great capacity for energy compaction, particularly at or near singularities. Given a particular wavelet, its corresponding basis, as noted above, is not universally optimal for all signals and particularly not for a noisy one; this difficulty may be lifted by adapting the representation to the signal in a "best" way possible and according to some criterion. The first of the two possible ways is to pick an optimal basis in a wavelet packet dictionary and discard the negligible coefficients [54]. The second, which focuses on reconstruction of a signal in noise, and which we discuss here, accounts for the underlying noise statistics in the multiscale domain to separate the "mostly" signal part from the "mostly" noise part [55−57]. We opt here to discuss a more general setting that assumes unknown noise statistics and where a signal reconstruction is still sought in some optimal way, as we elaborate below. This approach is particularly appealing in that it may be reduced to the setting in earlier developments.

4.1.1. Problem Statement. Consider an additive noise model

$$x(t) = s(t) + n(t) \tag{31}$$

where $s(t)$ is an unknown but deterministic signal corrupted by a zero-mean noise process $n(t)$, and $x(t)$ is the observed, that is, noisy, signal. The objective is to recover the signal $\{s(t)\}$ based on the observations $\{x(t)\}$.

The underlying signal is modeled with an orthonormal basis representation

$$s(t) = \sum_i C_i^s \psi_i(t)$$

and similarly the noise is represented as

$$n(t) = \sum_i C_i^n \psi_i(t)$$

By linearity, the observed signal can also be represented in the same fashion, and its coefficients are given by

$$C_i^x = C_i^s + C_i^n.$$

A key assumption we make is that for certain values of i, $C_i^s = 0$; in other words, the corresponding observation coefficients C_i^x represent "pure noise," rather than signal corrupted by noise. As shown by Krim and Pesquet [58], this is a reasonable assumption in view of the spectral and structural differences between the underlying signal $s(t)$ and the noise $n(t)$ across scales. Given this assumption, wavelet-based denoising consists of determining which wavelet coefficients represent primarily signal, and which mostly capture noise. The goal is to then localize and isolate the "mostly signal" coefficients. This may be achieved by defining an information measure as a function of the wavelet coefficients. It identifies the "useful" coefficients as those whose inclusion improves the data explanation. One such measure is Rissanen's information-theoretic approach [or minimum description length (MDL)] [59]. In other words, the MDL criterion is utilized for resolving the tradeoff between model complexity (each retained coefficient increases the number of model parameters) and goodness of fit [each truncated coefficient decreases the fit between the received (i.e., noisy) signal and its reconstruction].

4.1.2. The Coding Length Criterion. Wavelet thresholding is essentially an order estimation problem, one of balancing model accuracy against overfitting, and one of capturing as much of the "signal" as possible, while leaving out as much of the "noise" as possible. One approach to this estimation problem is to account for any prior knowledge available on the signal of interest, which usually is of probabilistic nature. This leads to a Bayesian estimation approach as developed in Refs. 60 and 61. While generally more complex, it does provide a regularization capacity which is much needed in low-SNR environments.

A parsimony-driven strategy, which we expound on here, addresses the problem of modeling in general, and that of compression in particular. It provides in addition a fairly general and interesting framework where the signal is assumed deterministic and unknown, and results in some intuitively sensible and mathematically tractable techniques [54]. Specifically, it brings together Rissanen's work on stochastic complexity and coding length [59,62], and Huber's work on minimax statistical robustness [63,64].

Following Rissasen, we seek the data representation that results in the shortest encoding of both observations and complexity constraints. As a departure from the commonly assumed Gaussian likelihood, we rather assume that the noise distribution f of our observed sequence is a (possibly) scaled version of an unknown member of the family of ε-contaminated normal distributions

$$\mathcal{P}_\varepsilon = \{(1-\varepsilon)\Phi + \varepsilon G : G \in \mathcal{F}\}$$

where Φ is the standard normal distribution, $\mathcal{F}$ is the set of all suitably smooth distribution functions, and $\varepsilon \in (0, 1)$ is the known fraction of contamination. (This is no loss of generality, since ε may always be estimated if unknown.) Note that this study straightforwardly reduces to the additive Gaussian noise case, by setting the mixture parameter $\varepsilon = 0$, and is in that sense more general.

For fixed model order, the expectation of the MDL criterion is the entropy, plus a penalty term that is independent of both the distribution and the functional form of the estimator. In accordance with the minimax principle, we seek the least favorable noise distribution and evaluate the MDL criterion for that distribution. In other words, we solve a minimax problem where the entropy is maximized over all distributions in $\mathcal{P}_\varepsilon$, and the description length is minimized over all estimators in $\mathcal{S}$. The saddle point (provided its existence) yields a minimax robust version of MDL, which we call the *minimax description length* (MMDL) criterion.

Krim and Schick [54], show that the least favorable distribution in $\mathcal{P}_\varepsilon$, which also maximizes the entropy, is one that is Gaussian in the center and Laplacian ("double exponential") in the tails, and switches from one to the other at a point whose value depends on the fraction of contamination ε.

Proposition 1. *The distribution $f_H \in \mathcal{P}_\varepsilon$ that minimizes the negentropy is*

$$f_H(c) = \begin{cases} (1-\varepsilon)\phi_\sigma(a)e^{(1/\sigma^2)(ac+a^2)} & c \le -a \\ (1-\varepsilon)\phi_\sigma(c) & -a \le c \le a \\ (1-\varepsilon)\phi_\sigma(a)e^{(1/\sigma^2)(-ac+a^2)} & a \le c \end{cases} \tag{32}$$

where ϕ_σ is the normal density with mean zero and variance σ^2 and a is related to ε by the equation

$$2\left(\frac{\phi_\sigma(a)}{a/\sigma^2} - \Phi_\sigma(-a)\right) = \frac{\varepsilon}{1-\varepsilon} \tag{33}$$

4.2. Coding for Worst-Case Noise

Let the set of wavelet coefficients obtained from the observed signal be denoted by $\mathcal{C}^N = \{C_1^x, C_2^x, \ldots, C_N^x\}$ as a time series without regard to the scale, and where the superscript indicates the corresponding process. Let exactly K of these coefficients contain signal information, while the remainder only contain noise. If necessary, we reindex these coefficients so that

$$C_i^x = \begin{cases} C_i^s + C_i^n & i = 1, 2, \ldots, K \\ C_i^n & \text{otherwise} \end{cases} \tag{34}$$

By assumption, the set of noise coefficients $\{C_i^n\}$ is a sample of independent, identically distributed random variates drawn from Huber's distribution f_H. It follows, by Eq. (34), that the observed coefficients C_i^x obey the distribution $f_H(c - C_i^s)$ when $i = 1, 2, \ldots, K$, and $f_H(c)$ otherwise. Thus, the likelihood function is given by

$$\ell(\mathcal{C}^N; K) = \prod_{i \le K} f_H(C_i^x - C_i^s) \prod_{i > K} f_H(C_i^x)$$

Since f_H is symmetric and unimodal with a maximum at the origin, this expression is maximized (with respect to the signal coefficient estimates $\{\hat{C}_i^s\}$) by setting

$$\hat{C}_i^s = C_i^x$$

for $i = 1, 2, \ldots, K$. It follows that the maximized likelihood (given K) is

$$\ell^*(\mathcal{C}^N; K) = \prod_{i \le K} f_H(0) \prod_{i > K} f_H(C_i^x)$$

Thus, the problem is reduced to choosing the optimal value of K, in the sense of minimizing the MDL criterion:

$$\begin{aligned} \mathcal{L}(\mathcal{C}^N; K) &= -\log \ell^*(\mathcal{C}^N; K) + K \log N \\ &= -\sum_{i \le K} \log f_H(0) - \sum_{i > K} \log f_H(C_i^x) + K \log N \end{aligned} \quad (35)$$

Neglecting terms independent of K, this is equivalent to minimizing

$$\tilde{\mathcal{L}}(\mathcal{C}^N; K) = \frac{1}{2\sigma^2} \sum_{i > K} \eta(C_i^x) + K \log N$$

where

$$\eta(c) = \begin{cases} c^2 & \text{if } |c| < a \\ a|c| - a^2 & \text{otherwise} \end{cases}$$

is proportional to the exponent in Huber's distribution f_H. This can simply be achieved by a thresholding scheme [54].

Proposition 2. *When* $\log N > a^2/2\sigma^2$, *the coefficient* $|C_i^x|$ *is truncated if*

$$|C_i^x| < \frac{a}{2} + \frac{\sigma^2}{a} \log N \qquad (case\ 1)$$

When $\log N \le \dfrac{a^2}{2\sigma^2}$, *the coefficient* $|C_i^x|$ *is truncated if*

$$|C_i^x| < \sigma \sqrt{2 \log N} \qquad (case\ 2)$$

Remarks. More ample details may be found in the paper by Krim and Schick [54]. Note however, when $\sigma^2 \to 0$, the thresholding scheme reduces to case 2, and C_i^x is never truncated; since this represents the no-noise case, it is reasonable that all coefficients should be retained in the reconstruction. On the other hand, for large σ^2, the thresholding scheme reduces to case 1, which is more conservative. For $\sigma^2 \to \infty$, the signal-to-noise ratio becomes zero and the best one can do is to estimate the signal as identically zero.

Similarly, when $a \to \infty$, the noise distribution becomes purely Gaussian, and the thresholding scheme reduces to case 2, as expected. The resulting threshold of this particular noise case coincides with the results of Donoho and others [55,56] and is qualitatively similar to that derived by Saito [57]. On the other hand, when $a \to 0$, the noise distribution becomes purely Laplacian, and the thresholding scheme reduces to Case 1.

Finally, when $N \to 1$, the thresholding scheme reduces to case 2, suggesting that outliers are unlikely to occur in a small sample, and it is hence more reasonable to assume purely Gaussian noise. On the other hand, for large N, the thresholding scheme reduces to case 1, since outliers are highly likely to occur in a large sample.

It is important to distinguish the minimax error result obtained by Donoho and Johnstone [55], which was achieved over a signal smoothness class, from those discussed here and derived by Krim and Schick [54], which are obtained over a family of noise distributions.

4.2.1. Numerical Experiments. In the examples that follow, we demonstrate the performance of the robust thresholding procedure described above, and compare it with that of the thresholding scheme based on the assumption of normally distributed noise.

Example 1. Using WAVELAB (available from the Stanford Statistics Department, courtesy of D. L. Donoho and I. M. Johnstone), we synthesized a broken ramp signal of length $N = 1024$. This signal admits an efficient representation in a wavelet basis, that is, one with very few nonzero coefficients. The noise is additive and independent identically distributed, obeying a $N(0, \sigma^2)$ distribution contaminated by a fraction $\varepsilon = 10\%$ of white Gaussian noise with distribution $N(0, 9\sigma^2)$. The overall signal-to-noise ratio (SNR) was maintained at 10 dB (see Fig. 16, top).

We implemented two estimators, the first of which is based on a purely Gaussian noise assumption (i.e., $\varepsilon = 0$), and where the thresholding scheme due to [58,65] was used. The second was the MMDL robust estimator described above. The reconstructions based on each estimator appear in Fig. 16. As may easily be observed, and in contrast to the MMDL technique, the Gaussian assumption induces a high susceptibility to outliers.

Monte Carlo simulations were carried out to evaluate the reconstruction performance over a range of SNRs. At each value of the SNR, 100 experiments were conducted, and the cumulative reconstruction error is displayed in Fig. 17. The robust estimator uniformly outperforms the *classic* estimator in both L_1 and L_2 errors over a wide range of SNRs. Furthermore, the performances of the Gaussian and robust estimators become indistinguishable at high SNRs, that is, with small noise variance, showing that robustness does not come at the cost of reduced efficiency.

Bounding the Reconstruction Error. Although the robust estimator-based reconstruction error is much improved it is still potentially unbounded. As discussed further by Krim and Schick, [54], and because of the compactness of wavelets, unbounded noise will still result in unbounded reconstruction error, a property that may be considered undesirable. This problem may be circumvented by making the assumption that the signal has bounded energy; in that case, one of at least two alternatives is possible:

1. In practice, the signal is known to be bounded, and prior knowledge of the physical properties of the signal may be used to determine the $\|\cdot\|_\infty$ of the sequence of signal wavelet coefficients $\{C_i^s\}$.

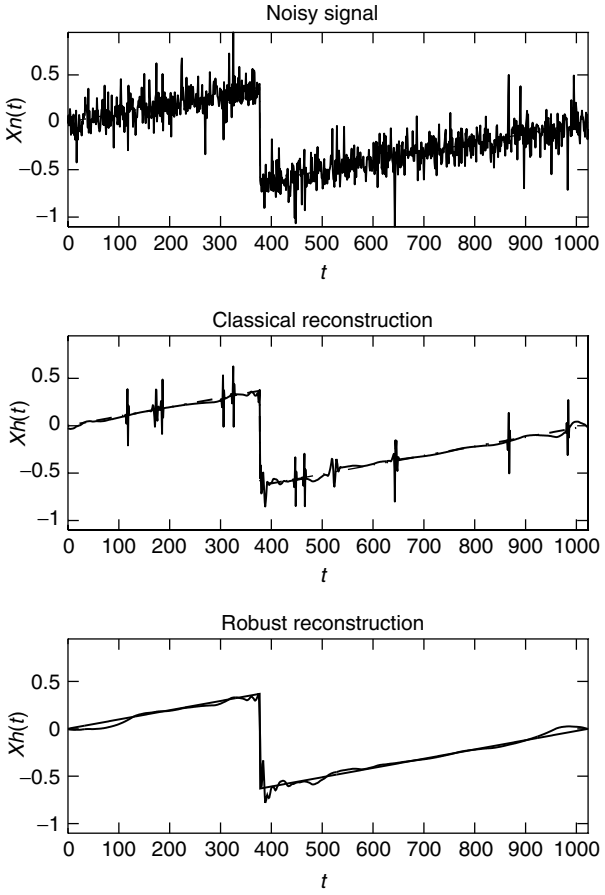


Figure 16. Noisy ramp signal and its Gaussian and robust reconstructions.

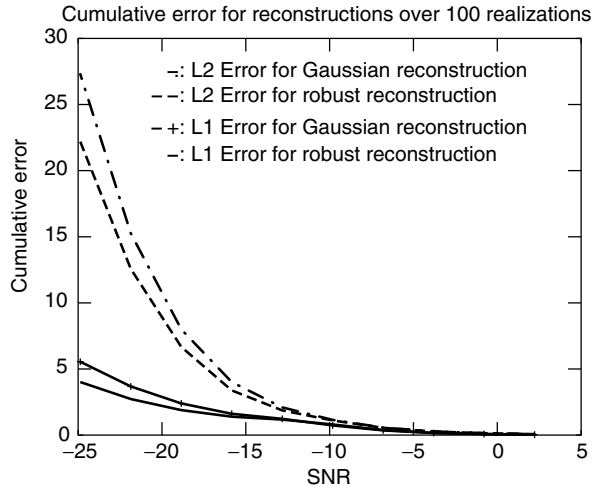


Figure 17. L_1 and L_2 error performance versus SNR, for the Gaussian and robust estimators.

This information may be used to truncate observed coefficients $\{C_i^x\}$ not only below, as discussed earlier, but also above.

2. In the absence of such prior knowledge, it may still be possible to bound the reconstruction error through an adaptive supremum–secondary thresholding scheme based on some representation criterion, such as entropy.

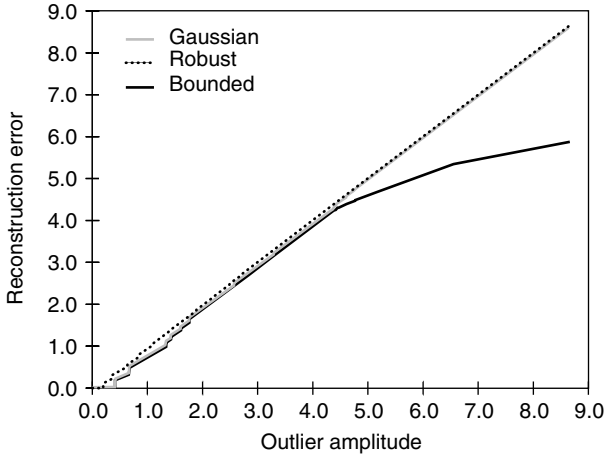


Figure 18. Absolute reconstruction error versus outlier amplitude for three thresholding schemes.

The first of these approaches is illustrated in Fig. 18, which uses the following modified thresholding. Let $\alpha > 0$ be an upper bound on the magnitude of the signal coefficients; then

$$\tilde{C}_i^s = \begin{cases} 0 & \text{if } |C_i^x| \le \dfrac{a}{2} + \dfrac{\sigma^2}{a}\log N \\ \hat{C}_i^s = C_i^x & \text{if } \dfrac{a}{2} + \dfrac{\sigma^2}{a}\log K \le |C_i^x| \le \alpha \\ \alpha\,\mathrm{sgn}(C_i^x) & \text{if } \alpha \le |C_i^x| \end{cases}$$

provided $\log N > a^2/2\sigma^2$ and $\alpha > (a/2) + (\sigma^2/a)\log N$. The graph shows that although the robust estimator's reconstruction error initially grows more slowly than that of the Gaussian estimator, the two errors soon converge as the variance of the outliers grows. The reconstruction error for the bounded-error estimator, however, levels off past a certain magnitude of the outlier, as expected.

Sensitivity Analysis for the Fraction of Contamination. Although such crucial assumptions as the normality of the noise or exact knowledge of its variance σ^2 usually go unremarked, it is often thought that Huber-like approaches are limited on account of the assumption of known ε. We demonstrate the resilience and robustness of the approach by studying the sensitivity of the estimator to changes in the assumed value of ε.

Figure 19 shows the total reconstruction error as a function of variation in the true fraction of contamination ε. In other words, an abscissa of 0 corresponds to an assumed fraction of contamination equal to the true fraction; larger abscissas correspond to outliers of larger magnitude than assumed by the robust estimator, and vice versa. Clearly, the Gaussian estimator assumes zero contamination throughout. Figure 19 shows that the reconstruction error for the Gaussian estimator grows very rapidly as the true fraction of contamination increases, whereas that of the robust estimator is nearly flat over a broad range. This should not come as a surprise: outliers

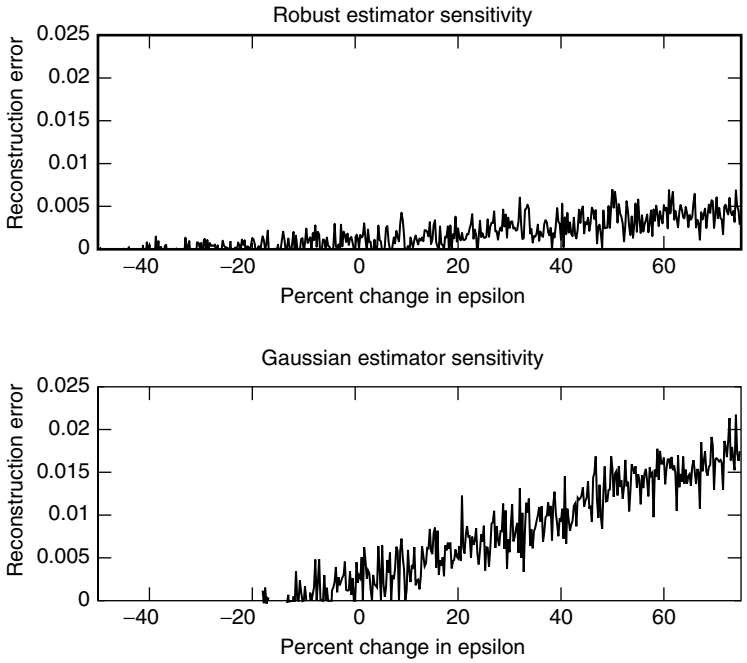


Figure 19. Error stability versus variation in the mixture parameter ε.

are, by definition, rare events, and for a localized procedure such as wavelet expansion, the precise frequency of outliers is much less important than their existence at all.

Example 2. Example 1 assumed a fixed wavelet basis. As discussed above, however, highly nonstationary signals can be represented most efficiently by using an adaptive basis that automatically chooses resolutions as the behavior of the signal varies over time. Using a L^2 error criterion, we search for the best basis of a chirp and show the reconstruction in Fig. 20 (see Krim et al. [66] for more details).

5. CONCLUSION

We have given an overview of an already broad area of research and discussed its application in information sciences, and more specifically an estimation technique that we believe captures the essence of many encountered problems. The article assumes an advanced undergraduate knowledge in electrical engineering applications and mathematics and may also play a role of a primer for a first-time reader on the topic and is aimed at providing a working knowledge of the tools, deferring most of the technical details to the references. While the bibliography is certainly incomplete (too large for it to be exhaustive), it hopefully provides a sufficient overview for the interested reader who may want to probe further.

BIOGRAPHY

Hamid Krim received his degrees in electrical engineering from University of Washington and Northeastern University. In 1991 he became a NSF postdoctoral scholar

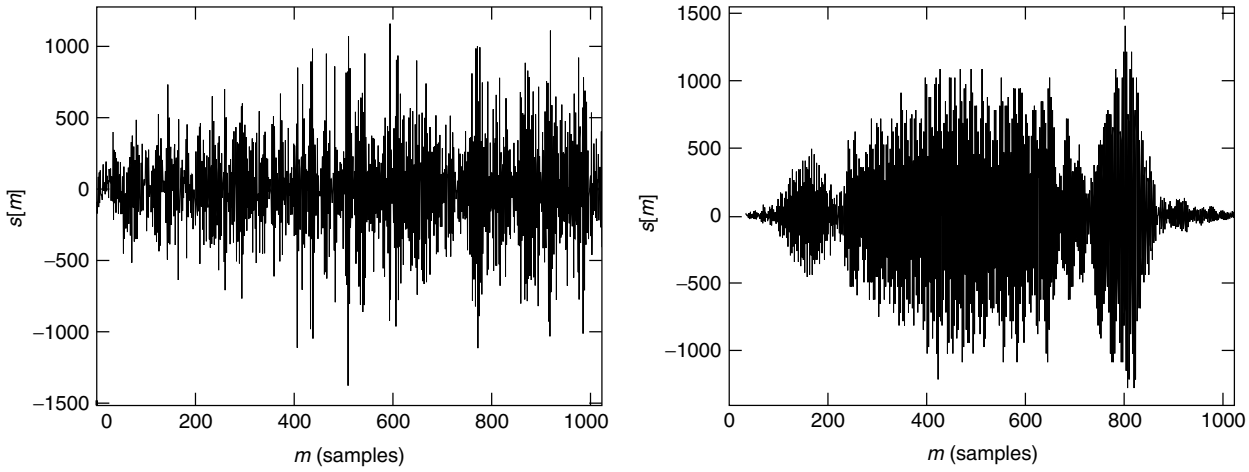


Figure 20. A noisy chirp reconstructed from its best basis and denoised as shown in the figure on the right.

at Foreign Centers of Excellence (LSS Supelec/Univ. of Orsay, Paris, France). He subsequently joined the Laboratory for Information and Decision Systems, MIT, Cambridge, Massachusetts, as a research scientist performing/supervising research in his area of interest and was an original contributor to the Center for Imaging Science sponsored by ARO. He then joined the faculty in the ECE Department at North Carolina State University in Raleigh, North Carolina in 1998. He also is a recipient of the NSF Career Young Investigator Award and an associate editor of the *IEEE Trans. On SP*. As a member of technical staff at AT&T Bell Labs, he has worked in the area of telephony and digital communication systems/subsystems. His research interests are in statistical estimation and detection and mathematical modeling with a keen emphasis on applications.

BIBLIOGRAPHY

1. E. Brigham, *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
2. A. Papoulis, *Signal Analysis*, McGraw-Hill, New York, 1977.
3. D. Gabor, Theory of communication, *J. IEE* **93**: 429–457 (1946).
4. S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Boston, 1997.
5. I. Daubechies, S. Mallat, and A. Willsky, Special edition on wavelets and applications, *IEEE Trans. on IT*, 1992.
6. I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF, SIAM, Philadelphia, 1992.
7. M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
8. H. Krim, On the distribution of optimized multiscale representations, in *ICASSP*, Vol. V, (Munich, Germany), IEEE, May 1997.
9. G. H. Golub and C. F. VanLoan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1984.
10. D. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1968.
11. K. Gröchenig, Acceleration of the frame algorithm, *IEEE Trans. Signal Process.* **41**: 3331–3340 (Dec. 1993).
12. A. B. Hamza and H. Krim, Image Denoising: A Nonlinear Robust Statistical Approach, *IEEE, Trans. SP* **SP-49**(12): 3045–3054 (2001).
13. Y. Meyer, *Ondelettes et opérateur*, Vol. 1, Hermann, Paris, 1990.
14. S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Trans. Pattern Anal. Mach. Int.* **PAMI-11**: 674–693 (July 1989).
15. Y. Meyer, *Wavelets and Applications*, SIAM, Philadelphia, 1992.
16. Y. Meyer, *Ondelettes et opérateur*, Vol. 1, Hermann, Paris, 1990.
17. F. J. Hampson and J.-C. Pesquet, *A Nonlinear Decomposition with Perfect Reconstruction, IEEE, Trans. on IP 7-11*, 1998, 1547–1560.
18. P. L. Combettes and J.-C. Pesquet, Convex multiresolution analysis, *IEEE Trans. Pattern Anal. Mach. Int.* **20**: 1308–1318 (Dec. 1998).
19. W. Sweldens, The lifting scheme: A construction of second generation wavelets, *SIAM J. Math. Anal.* **29**(9): 511–546 (1997).
20. S. Mallat, Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$, *Trans. Am. Math. Soc.* **315**: 69–87 (Sept. 1989).
21. J. Woods and S. O'Neil, Sub-band coding of images, *IEEE Trans. ASSP* **34**: 1278–1288 (May 1986).
22. M. Vetterli, Multi-dimensional subband coding: some theory and algorithms, *Signal Process.* **6**: 97–112 (April 1984).
23. F. Meyer and R. Coifman, Brushlets: a tool for directional image analysis and image compression, *Appl. Comput. Harm. Anal.* 147–187 (1997).
24. G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Boston, 1996.
25. P. P. Vaidyanathan, Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial, *Proc. IEEE* **78**: 56–93 (Jan. 1990).
26. A. Cohen, I. Daubechies, and J. C. Feauveau, Biorthogonal bases of compactly supported wavelets, 1992. Unpublished Technical Report.
27. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
28. R. Coifman and Y. Meyer, Remarques sur l'analyse de Fourier à fenêtre, *C. R. Acad. Sci. Série I* 259–261 (1991).
29. R. R. Coifman and M. V. Wickerhauser, Entropy-based algorithms for best basis selection, *IEEE Trans. Inform. Theory* **IT-38**: 713–718 (March 1992).
30. M. V. Wickerhauser, INRIA lectures on wavelet packet algorithms, in *On-delettes et paquets d'ondelettes* INRIA T-R (Roquencourt), June 17–21, 1991, pp. 31–99.
31. H. Malvar, Lapped transforms for efficient transform subband coding, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-38**: 969–978 (June 1990).
32. A. Aldroubi and E. M. Unser, *Wavelets in Medicine and Biology*, CRC Press, Boca Raton, FL, 1996.
33. A. Akansu and E. M. J. Smith, *Subband and Wavelet Transforms*, Kluwer, 1995.
34. A. Arneodo, F. Argoul, J. E. E. Bacry, and J. Muzy, *Ondelettes, Multifractales et Turbulence*, Diderot, Paris, 1995.
35. A. Antoniadis and E. G. Oppenheim, *Wavelets and Statistics*, Vol. LNS 103, *Lecture Notes in Statistics*, Springer-Verlag, 1995.
36. P. Mueller and E. B. Vidakovic, *Bayesian Inference in Wavelet-Based Models*, Vol. LNS 141, *Lecture Notes in Statistics*, Springer-Verlag, 1999.
37. B. Vidakovic, *Statistical Modeling by Wavelets*, Wiley, New York, 1999.
38. K. Ramchandran and M. Vetterli, Best wavelet packet bases in a rate-distorsion sense, *IEEE Trans. Image Process.* **2**: 160–175 (April 1993).
39. J. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. Signal Process.* **41**: 3445–3462 (1993).
40. P. Cosman, R. Gray, and M. Vetterli, Vector quantization of image subbands: a survey, *IEEE Trans. Image Process.* **5**: 202–225 (Feb. 1996).

41. A. Kim and H. Krim, Hierarchical stochastic modeling of sar imagery for segmentation/compression, *IEEE Trans. SP* 458–468 (Feb. 1999).

42. M. Basseville et al., Modeling and estimation of multiresolution stochastic processes, *IEEE Trans. Inform. Theory* **IT-38**: 529–532 (March 1992).

43. M. Luettgen, W. Karl, and A. Willsky, Likelihood calculation for multiscale image models with applications in texture discrimination, *IEEE Trans. Image Process.* **3**(1): 41–64 (1994).

44. E. Fabre, New fast smoothers for multiscale systems, *IEEE Trans. Signal Process.* **44**: 1893–1911 (Aug. 1996).

45. P. Fieguth, W. Karl, A. Willsky, and C. Wunsch, Multiresolution optimal interpolation and statistical ananlysis of topex/poseidon satellite altimetry, *IEEE Trans. Geosci. Remote Sens.* **33**: 280–292 (March 1995).

46. M. K. Tsatsanis and G. B. Giannakis, Principal component filter banks for optimal multiresolution analysis, *IEEE Trans. Signal Process.* **43**: 1766–1777 (Aug. 1995).

47. M. K. Tsatsanis and G. B. Giannakis, Optimal linear receivers for ds-cdma systems: A signal processing approach, *IEEE Trans. Signal Process.* **44**: 3044–3055 (Dec. 1996).

48. A. Scaglione, G. B. Giannakis, and S. Barbarossa, Redundant filterbank precoders and equalizers, parts i and ii, *IEEE Trans. Signal Process.* **47**: 1988–2022 (July 1999).

49. A. Scaglione, S. Barbarossa, and G. B. Giannakis, Filterbank transceivers optimizing information rate in block transmissions over dispersive channels, *IEEE Trans. Inform. Theory* **45**: 1019–1032 (April 1999).

50. R. Learned, H. Krim, A. Willsky, and W. Karl, Wavelet-packet-based multiple access communication, *Wavelet Appl. Signal Image Proc.* **II**: 246–259 (Oct. 1994).

51. R. Learned, A. Willsky, and D. Boroson, Low complexity optimal joint detection for oversaturated multiple access communications, *IEEE Trans. Signal Process.* **45**(1): 113–123 (1997).

52. A. Lindsey, Wavelet packet modulation for orthogonally multiplexed communication, *IEEE Trans. Signal Process.* **45**(5): 520–524 (1997).

53. K. Wong, J. Wu, T. Davidson, and Q. Jin, Wavelet packet division multiplexing and wavelet packet design under timing error effects, *IEEE Trans. Signal Process.* **45**: 2877–2886 (Dec. 1997).

54. H. Krim and I. Schick, Minimax description length for signal denoising and optimized representation, *IEEE Trans. Inform. Theory* (April 1999). (Eds. H. Krim, W. Willinger, A. Iouditski and D. Tse).

55. D. L. Donoho and I. M. Johnstone, *Ideal Spatial Adaptation by Wavelet Shrinkage*, preprint, Dept. Statistics, Stanford Univ., June 1992.

56. H. Krim, S. Mallat, D. Donoho, and A. Willsky, Best basis algorithm for signal enhancement, in *ICASSP'95*, Detroit, MI, IEEE, May 1995.

57. N. Saito, *Local Feature Extraction and Its Applications Using a Library of Bases*, Ph.D. thesis, Yale Univ., Dec. 1994.

58. H. Krim and J.-C. Pesquet, On the statistics of best bases criteria, in *Wavelets in Statistics, Lecture Notes in Statistics*, Springer-Verlag, July 1995.

59. J. Rissanen, Modeling by shortest data description, *Automatica* **14**: 465–471 (1978).

60. B. Vidakovic, Nonlinear wavelet shrinkage with bayes rules and bayes, *J. Am. Stat. Assoc.* **93**: 173–179 (1998).

61. D. Leporini, J.-C. Pesquet, and H. Krim, *Best Basis Representation with Prior Statistical Models, Lecture Notes in Statistics*, LNS 141 Springer-Verlag, 1999, Chap. 11.

62. J. Rissanen, Stochastic complexity and modeling, *Ann. Stat.* **14**: 1080–1100 (1986).

63. P. Huber, Robust estimation of a location parameter, *Ann. Math. Stat.* **35**: 1753–1758 (1964).

64. P. Huber, *Théorie de l' inférence statistique robuste*, technical report, Univ. Montreal, Montreal, Quebec, Canada, 1969.

65. D. Donoho and I. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, *J. Am. Stat. Assoc.* **90**: 1200–1223 (Dec. 1995).

66. H. Krim, D. Tucker, S. Mallat, and D. Donoho, Near-optimal risk for best basis search, *IEEE Trans. Inform. Theory* **45**(7): 2225–2238 (Nov. 1999).

WDM METROPOLITAN-AREA OPTICAL NETWORKS

IOANNIS TOMKOS
Athens Information Technology
Peania, Greece

1. INTRODUCTION TO WDM OPTICAL NETWORKS

The role of a telecommunications network is to transmit information in the most efficient, reliable, and cost-effective way. To do this, a communications channel with enough bandwidth to satisfy the traffic demands is needed. From all the known transmission media, the optical fiber has the largest available bandwidth (>50 THz) and can satisfy in the most efficient way the traffic demands. Therefore, fiberoptic links are the backbone of current high-speed communication networks.

To take full potential of the available fiber bandwidth, several multiplexing techniques have been introduced in optical communication systems in accordance with conventional digital communications systems. Among all of them, the most popular is the *wavelength-division multiplexing* (WDM) technique [1]. WDM allows multiple datastreams from various application protocols to be combined and transmitted in parallel at different wavelengths, thereby multiplying the capacity of a single fiber strand. The different wavelengths, separated from one another at a fixed spacing frequency, are multiplexed together using a WDM multiplexer and are then transmitted over the same optical fiber. The signals at the output end of the fiber are demultiplexed and are redistributed into the various applications. Depending on the application, WDM can be deployed in a coarse version (CWDM) with 16 or fewer wavelengths having relatively wide spacing or in a dense version (DWDM) with up to hundreds of wavelengths. The state-of-the-art commercially available DWDM systems utilize 80 10-Gbps (gigabit per second) channels spaced at 50-GHz-frequency separation. The next generation of DWDM systems will eventually utilize 160 channels spaced at 25 GHz.

The first commercial systems that used fiber to transmit signals instead of WDM utilized space division-multiplexing (SDM) through the use of multiple fibers and single channel transmission per fiber (Fig. 1a). To overcome losses, optoelectronic (OEO) regenerators were frequently used, resulting in huge system costs. Several installed systems based on *synchronous optical network/synchronous digital hierarchy* (SONET/SDH) are still using such a technique. The breakthrough for the optical transmission technology came with the use of WDM and the invention of optical amplifiers. Tens of optical channels can be amplified simultaneously, leading to an enormous reduction of cost, by eliminating many fibers and more importantly OEO regenerators (Fig. 1b).

Since the introduction of point-to-point WDM systems, the potential for WDM all-optical networking, without the use of optoelectronic conversions, has been exploited by many research groups [2–6], and is now becoming a commercial reality. Beyond a higher capacity in a single fiber, WDM optical networks offer the capability for transparent wavelength routing [5,6]. Their key network elements are optical switching devices (optical add/drop

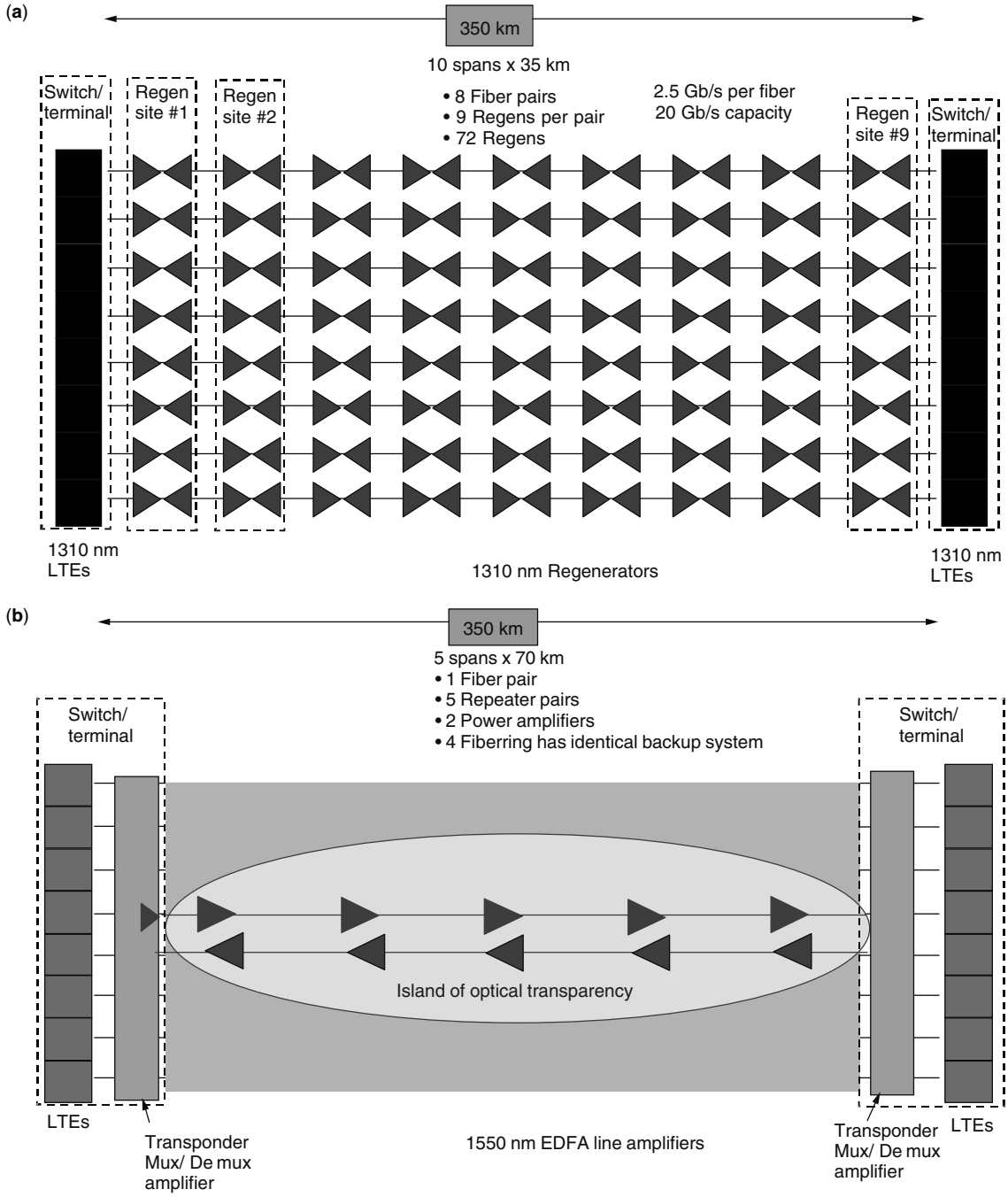


Figure 1. Layout of (**a**) early systems that used fiber to transmit signals, and (**b**) WDM point-to-point systems with optical amplification.

multiplexers and optical cross-connects) that enable routing at the wavelength level without the use of OEO conversion [7].

Optical add/drop multiplexers (OADMs) are used to add/remove signals from a WDM "comb" transmitted along a fiber connection. (Fig. 2a). The OADM architectures should introduce minimum impairment to the passthrough signals while enabling access to all the channels in the transmitted fiber with minimum cost [8]. The use of OADMs in transparent chains [9] or in single transparent rings [10] depending on the application, has been discussed. *Optical cross-connects* (OXCs) are under development and in the future will enable transparent ring interconnection and mesh optical networking [5–7]. OXCs are a more generalized form of OADMs, and besides their use for add/drop of individual signals from a WDM comb, they are used to cross-connect signals coming from different fibers (Fig. 2b). The main functions of the OXC will be to dynamically reconfigure the network — at the wavelength level — for restoration purposes, or to accommodate changes in bandwidth demand.

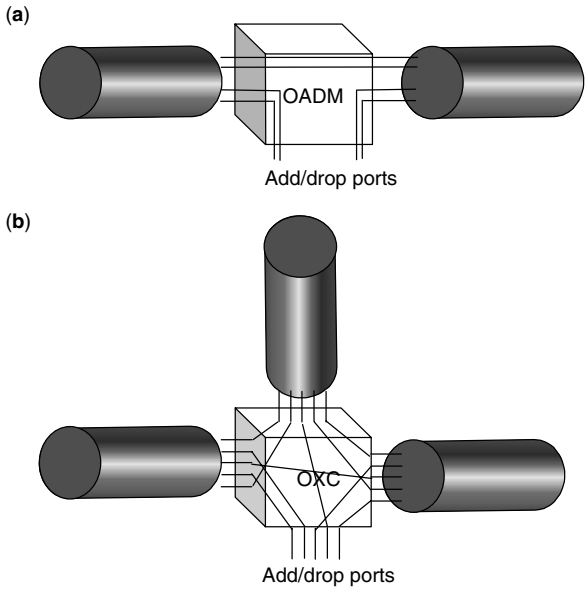


Figure 2. Schematic representation of the functionality of an (**a**) OADM and (**b**) OXC.

In the near future, *wavelength-routed* WDM optical networks will span across all network segments. In a general end-to-end connectivity picture, the current networks consists of three major segments, the long-haul network, the metropolitan-area network, and the residential access network. Figure 3 presents a schematic representation of the layout of the various network segments. A metropolitan-area network (MAN) is simply defined as the part of the network that interfaces between the end users ("residential access" or "last-mile" networks) and the backbone long-haul network [10].

Significant network growth has been observed in metropolitan areas, mainly due to the increased growth in data and IP (Internet Protocol) traffic demands enabled by the introduction of broadband access technologies (e.g., cable modems, ADSL/VDSL) to end users. Driven mainly by the increasing traffic demand in metropolitan areas [11], WDM is now beginning to expand from a network core technology toward the metropolitan and access network arenas. The focus of research has shifted toward WDM metropolitan networks [10–17], with the goal of bringing the benefits (cost and network efficiency) of the optical networking revolution toward the end users. From both technical and economic perspectives, the ability to provide potentially unlimited transmission capacity is the most obvious advantage of DWDM technology. However, the use of WDM as simply a network infrastructure tool that provides increased system capacity is not as compelling for short-haul networks, and must bring more benefits to the operator to gain acceptance. Bandwidth aside, the most compelling technical advantages of WDM networking for metro (MAN) applications can be summarized as follows:

- *Transparency*. A real catalyst behind the use of WDM in short-haul networks may be its promise of "transparency" in offering new high-end wavelength–based services. Several "shades" of transparency have been envisioned, spanning the spectrum from "full" transparency (format, protocol, bit rate) to some subset. WDM can transparently support at their native bit rate, both time-division multiplexing (TDM) and data formats such as *asynchronous transfer mode* (ATM), *Gigabit Ethernet* (GbE), *Fiber Distributed Data Interface* (FDDI), and

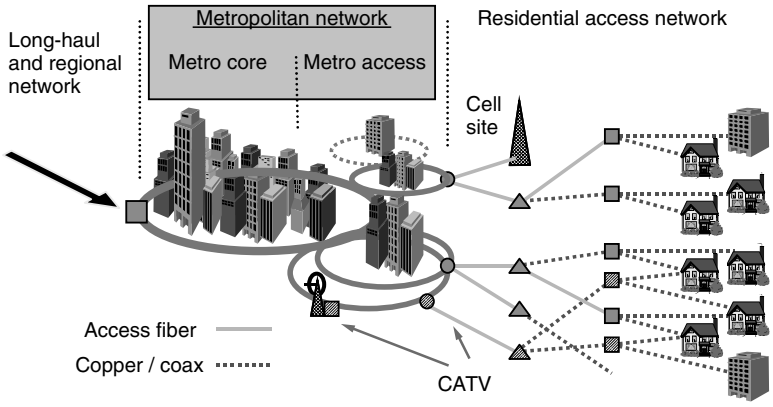


Figure 3. Schematic representation of the layout of the various networks segments.

Enterprise System Connectivity (ESCON). Networks would potentially have the flexibility to transport any kind of data without regard for the restrictions of the SONET digital hierarchy. Full (or analog) transparency at this level of the network would "future-proof" the infrastructure against bit rate increase and new traffic types, and would reduce the equipment load in the signal path, resulting in significant cost advantage.

- *Scalability*. DWDM can leverage the abundance of dark fiber in many metropolitan-area networks to quickly meet demand of capacity on point-to-point links and on spans of existing SONET/SDH rings.
- *Dynamic Provisioning*. Fast, simple and dynamic provision of network connections gives providers the ability to provide high-bandwidth services to enterprises/end-users in days rather than months.
- *Bandwidth Allocation*. Ability to drop individual wavelengths to the building and utilization of the available bandwidth more efficiently.

In the following text we review the requirements, architectures, and performance issues related with WDM metropolitan area optical networks. We present our considerations about the evolution of MANs. We also discuss the optical impairments that limit the transparency in metropolitan WDM networks and present the characteristics of MAN-optimized optical amplifiers, fibers, and architectures of optical add/drop multiplexers that enable flexible and highly performing metropolitan-area networks.

2. METROPOLITAN OPTICAL NETWORKS: CHARACTERISTICS, DEFINITIONS, AND REQUIREMENTS

The main role of the metropolitan-area network segment is to provide traffic grooming and aggregation of a full range of client protocols from enterprise/private customers in access networks to backbone service provider networks. In addition, since the majority of the traffic stays within the same area, metro networks need to provide efficient networking capabilities within the metropolitan-area.

The design of metropolitan-area networks presents many engineering challenges, especially in light of the large existing base of legacy SONET/SDH infrastructure prevalent in current MANs. These traditional TDM networks were originally designed to transport a limited set of traffic types, mainly multiplexed voice and private line services. SONET/SDH systems are not able to transport efficiently data-optimized protocols (e.g., to carry a GbE signal at 1.25 Gbps, a full 2.488-Gbps SONET/SDH channel is needed — a waste of bandwidth). However, today's MAN market is being driven by the need to streamline network efficiencies under rapidly growing capacity demands and increasingly variable traffic patterns. Hence there is a strong desire to migrate from the current SONET/SDH-based network architecture into a more proactive (dynamic and intelligent), multiservice optical network. This will allow service providers to circumvent the need to perform forklift upgrades or lay more fiber (which is time- and cost-intensive), thereby cost-effectively migrating toward a "futureproof" network.

From an architecture point of view, the conventional metropolitan networks can be separated in two different segments with distinct roles: the "metro-access" (or "collector"), and "metro-core" [or "interoffice" (IOF)] networks (Fig. 4). "Metro-access" networks are responsible for collecting the traffic and forward it to a hub node of the "inter office" network, which in turn will act to network the traffic between hub nodes and redirect it to the backbone long-haul network. The typical length of longest path in IOF networks is ~300 km and ~100 km in "metro-access" networks. IOF networks are designed mostly as physical rings with a meshed traffic pattern as shown in Fig. 4b. The dots on the schematic indicate network node sites, which perform aggregation of high-bandwidth optical connections, cross-connect, and handoff to the core network. Cross-connects at these nodes are quite likely to employ electronic switch cores to support subwavelength "grooming" of circuits. Grooming allows for the most efficient utilization of network bandwidth by performing a mix of space switching (port-to-port) and time-domain switching (time slot to time slot), which cannot be implemented easily in the optical domain. Metro-access networks are designed mostly as physical rings with a hubbed traffic pattern as shown in Fig. 4c. The hub node has access to all traffic present in the network, representing an aggregation point and a connection to the IOF network. "Edge boxes" sitting at the access network edge, perform traffic aggregation and WDM multiplexing of low-bandwidth services.

Metropolitan networks are subject to specific requirements and traffic demands that are quite different from those of the backbone long-haul network. Therefore, metro network carriers face additional challenges due to the distinct characteristics of MANs:

- They are very *cost-sensitive*, since the overall network cost is divided into a smaller number of customers than in the long-haul network.
- They are very *sensitive to space and power consumption* characteristics of the network equipment, since the network carriers are struggling to reduce the cost of operating and maintaining the network.
- They are characterized by more *rapidly changing traffic patterns* requiring fast provisioning and therefore ability for network scalability, modularity, fast reconfiguration, and availability of capacity.
- Since the metro network interfaces with a wide variety of end customers, it needs to support very *diverse types of traffic* (SONET/SDH, Ethernet, ESCON, Fiber Channel, ATM, IP, etc.). Therefore network nodes — especially at the edges of the metro network should perform traffic aggregation.
- The *bit rates of the data tributaries accessing the "metro-access" network can also vary quite significantly* (e.g., from OC-3 to OC-196 for SONET and from 100 Mbps to 10 Gbps for Ethernet traffic). Therefore network nodes should also perform traffic grooming to improve the efficiency of the transport

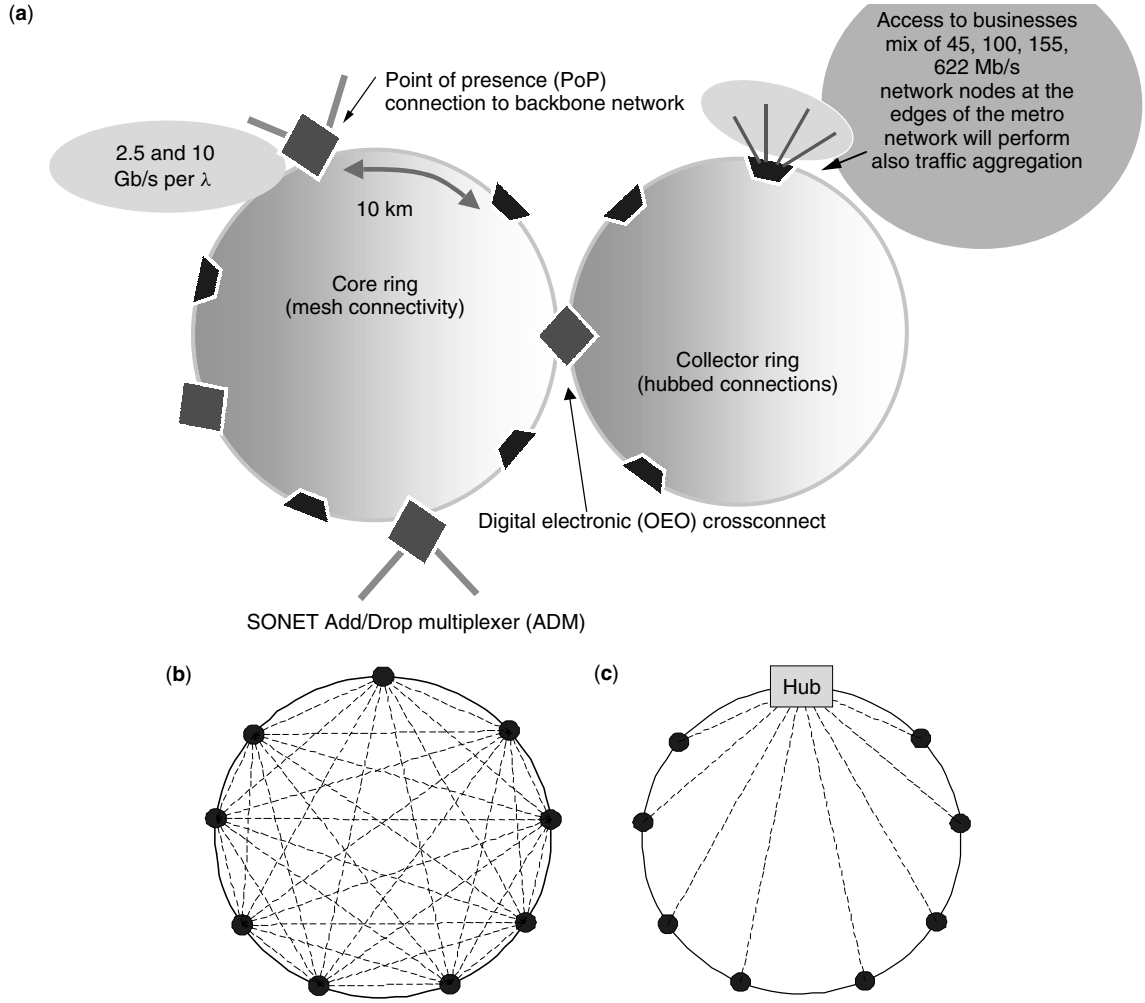


Figure 4. Metro network reference architecture (**a**) and traffic pattern for the IOF (**b**) and metro-access (**c**) network segments

network by combining the low-bit-rate signals to a high-bit-rate wavelength channel. Moreover, the network should be able to support bandwidth provisioning with various levels of granularity.

- Metropolitan-area optical networks should transport information while satisfying the service-layer agreements for *quality of service and high signal integrity*.

3. SYSTEM OFFERINGS FOR METRO NETWORKS

On the basis of the abovementioned requirements, we can conclude that the system offerings for metro networks should

- Utilize cost-effective devices with small size and low power consumption. The devices may trade these characteristics for lower performance, since the size of the metro networks is relatively small and high signal quality can be preserved.
- Provide many different protocol interfaces and the capability of traffic grooming and aggregation.
- Utilize technologies (e.g., optical switching) that will enable the network carriers to move from costly time-consuming static network provisioning (the step-by-step process of making an optical connection between two sites) to fast automated provisioning that requires minimum cost. With the proper network management system, the operator should "see" the network resources from a single screen, change the connections through software tools, and receive instant verification that the connection is intact (*point-and-click provisioning*).
- Provide network protection and restoration as major requirements that ensure quality of service [18]. Next-generation optical networks will eventually need to support optical protection switching capabilities, which have proved to considerably reduce the network costs as opposed to protection switching in the electronic domain [19]. The simplest protection option is called "dedicated" or "$1 + 1$," where two copies of the optical signals are counterpropagated through the network on different fibers and both are delivered to the customer equipment, where the signal having the best quality is detected [20].

More advanced protection schemes perform reuse of capacity (i.e., wavelengths) [19]. For the implementation of optical protection there are several enabling technologies such as fast and reliable optical switches, signal-quality monitors, and transient gain-controlled optical amplifiers [19–20].

Several companies are currently researching and developing systems optimized for use in metropolitan applications. The system offerings that are considered for deployment in metro networks can be separated in broad categories that are listed below:

1. *Legacy SONET/SDH*. This is the dominant method of optical transport in metro networks today. Roughly 80,000 metro SONET rings exist only in North America today. It is a highly reliable ring-based solution, which offers service protection within 50 ms in case of network failure. It was optimized for voice traffic and is therefore inflexible and inefficient with data traffic. Network connectivity is established with SONET add/drop multiplexers (ADMs) and digital electronic cross-connects (DXCs) for interconnecting the rings (Fig. 4a). The SONET ADMs and DXCs require optical–electrical–optical (O-E-O) conversion, which makes the technology costly. The increasing traffic demands in metropolitan areas have been satisfied by increasing the SONET channel bit rate (i.e., from OC-3 to OC-196) or the number of fibers connecting the nodes, introducing in that case network overlays.

2. *Next-Generation SONET/SDH*. This is the response of SONET/SDH system integrators to the evolving characteristics of metro networks. These are "data-optimized" SONET/SDH systems that are fully interoperable with legacy SONET/SDH networks. They are very compelling solutions since they simplify the network provisioning and improve network efficiency by integrating traffic grooming and aggregation equipment.

3. *Metro-Core Point-to-Point DWDM*. These systems transport multiple streams (32–128 wavelengths) of optical signals on a single fiber. These conventional WDM solutions have simply translated the SONET ring model to a WDM platform with multiple wavelengths per fiber and possibly optical amplifiers to overcome losses (Fig. 5a). They are typically used to "connect" large carrier offices within the metropolitan area in a point-to-point fashion. The network savings arising from these system offerings can be translated in a first approximation to a reduction of the number of required optical fibers and SONET network overlays that are needed to accommodate the traffic demands. The additional cost of WDM system offerings are in the use of multiplexers/demultiplexers, amplifiers, and devices for signal conditioning management (e.g., variable optical attenuators, dispersion compensating modules).

4. *Metro Core Ring DWDM*. These systems utilize optical add/drop multiplexers, which replace SONET add/drop multiplexers to build physical rings carrying WDM signals (Fig. 5b). OADMs, through the wavelength routing concept, can transform a physical ring topology into any type of network logical topology (ring, mesh, or star). These systems are capable to transport multiple streams (e.g., 32–128 wavelengths) of optical signals on a single fiber. Some systems also promise the use of OXC for transparent ring interconnection (Fig. 5c). In such case, no time-division multiplexing/demultiplexing will be performed at the XCs. This will result in inability to perform grooming, as opposed to DXC, and may delay their introduction in metro networks.

5. *Metro-Access WDM*. These systems transport optical signals between large enterprise sites and carrier central offices. Shorter distance (typically 5–20 km) and lower capacity requirements mean that these systems need lower optical performance than do metro-core DWDM systems. *Coarse WDM* (CWDM) systems with channel spacing as high as 20 nm is a suitable technology for such application. Efficient aggregation and transport of data traffic is again a key objective.

6. *Next-Generation Data Transport Systems*. These are data-only devices (layer 2 switches and layer 3 routers) that will eliminate SONET/SDH systems altogether. They achieve more efficient transport of data traffic, at the expense of limited voice capabilities and questionable reliability. They are based mostly on *Ethernet*, although *packet-over-SONET* and other transport protocols have been proposed [21]. Ethernet is a transport solution that is almost 100% predominant within the local-area network (LAN) market. Its widespread extension into the metro area is considered to be quite reasonable in the near future. New technologies, such as *multiprotocol label switching* (MPLS) and *resilient packet rings* (RPRs), which have been introduced to deliver efficient routing and improved quality of service, are showing significant promise. RPR is a *medium-access control* (MAC) protocol designed to optimize bandwidth utilization and facilitate services over a ring network (unlike Ethernet), while providing carrier-class attributes such as 50-ms ring-protection switching. It is worth pointing out that WDM systems are interoperable with data transport systems and can be used for network upgrades. For example, the network operator can keep TDM voice services over SONET on one wavelength, while deploying the data-oriented technology on another wavelength.

7. *All-Optical Packet Switching*. This is a data-optimized transport technology that promises to provide high throughput, packet-level switching, rich routing functionality, and excellent flexibility, making this system offering ideal for metro networks [22–25]. It will utilize OADMs and OXCs that, on top of wavelength switching, will be able to perform packet-level switching. Although this type of system has been researched extensively in the past by many groups and consortia [22–25], it remains the most forward-looking solution.

4. ENGINEERING OF TRANSPARENT METRO NETWORKS

Clearly, optical switching through the use of OADMs and OXCs is a promising technology for provisioning, protection, and restoration of high-bandwidth services in the optical layer. However, the adoption of optical switching technology and optically transparent network architectures will not be automatic by all network carriers. The reduction of O-E-O interfaces presents a special set

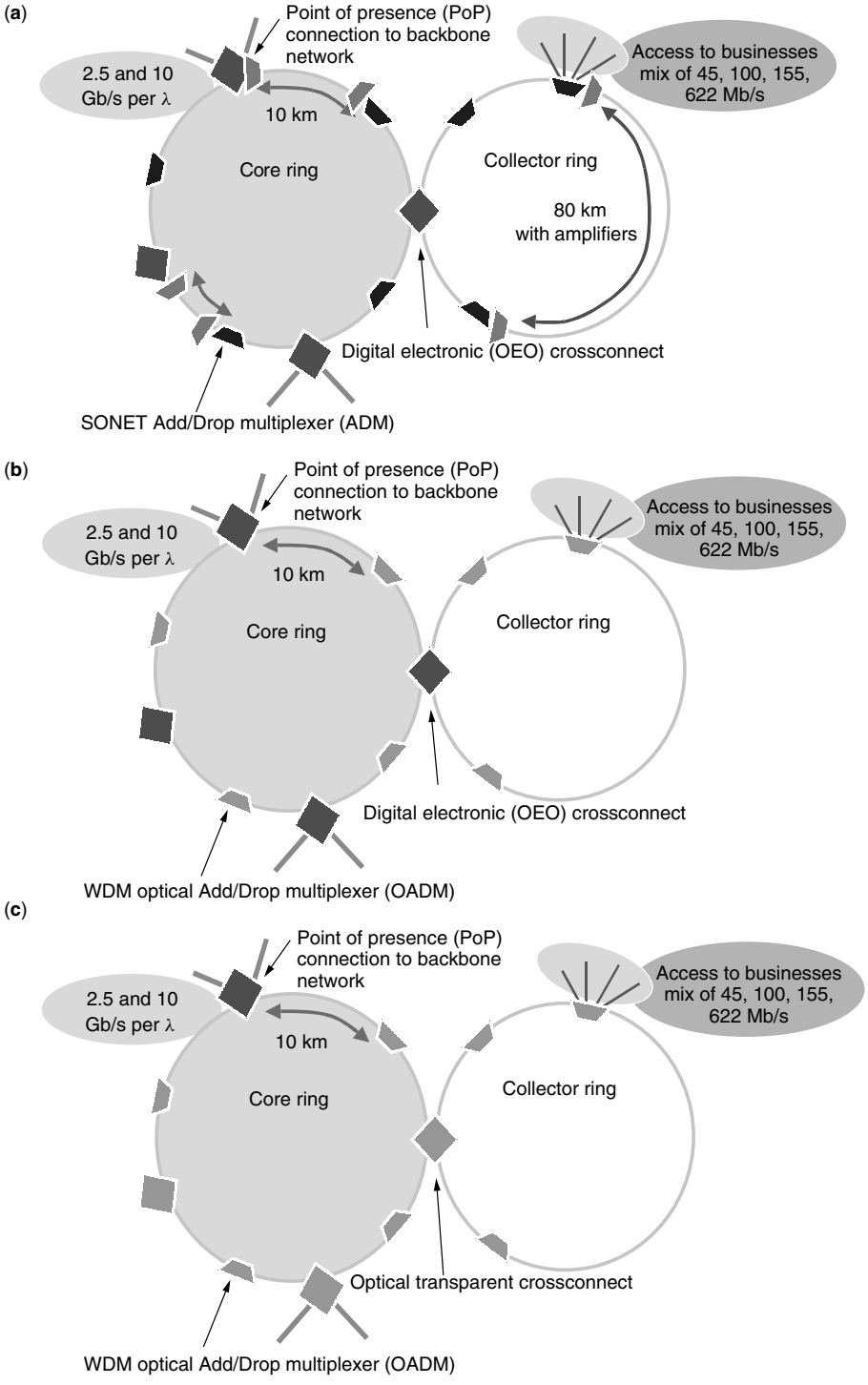


Figure 5. Metro network evolution: (**a**) SONET rings, WDM on selected routes; (**b**) WDM OADM Rings interconnected with opaque digital XC, optical amplifiers used frequently; (**c**) WDM OADM rings interconnected with transparent optical XC, optical amplifiers used frequently.

of challenges to network designers. The transparency that WDM systems offer limits the scalability (in terms of the number of channels, bit rate, etc.) or geographic extent of the network. The limitations arise from

- *Accumulation of Transmission and Networking Impairments*. Several effects, unique in optical transmission and networking systems, limit the maximum physical size of the network that can be supported [10,14,15,26–32]. In a transparent optical network the WDM signals pass through many optical components that may introduce deterioration of signal quality.
- *Difficulty in Engineering the Network*. Indeed, the network would have to be engineered for the "worst case" [15,26]. As a consequence, the worst path (e.g., the longer restoration path) will have to be known from the outset, and all components must be specified for this worst path (and the specifications will be tighter, the greater the network reach). Furthermore,

transparency requires that the whole network be engineered at once. And once engineered, the network cannot be extended beyond its intended design.

- *Difficulty in Performance Monitoring*. Performance monitoring of the WDM signals at the bit level is difficult without optoelectronic conversions [33,34]. Also, in-band management information is difficult to add and monitor, which makes it difficult to manage networks with transparent all-optical nodes.

A significant consideration in metro optical DWDM networks is the total transmission length and number of cascaded nodes that can be satisfied by system optics without resorting to the cost and complexity of electrical regeneration. The use of cost-effective devices (e.g., directly modulated transmitters) to reduce the overall cost the need for network reconfiguration render the engineering of a transparent metro network a nontrivial task. Special care should be taken to reduce the degradation of the signal quality due to the use of such cost-effective technologies. Moreover, depending on the network and network node architecture, the impact of some of the effects can be more pronounced. In the following paragraphs we discuss the main effects that make the design of system offerings for metro networks not a trivial task.

Signal attenuation from fiber/component loss could be overcome using optical amplifiers. The *noise* that these amplifiers introduce in metropolitan area networks can be managed since the size of the network is relatively small and short amplifier spans are used [high optical signal-to-noise ratio (SNR) can be preserved]. *Power divergence* among the WDM channels caused by *component ripple* as well as *polarization-depended loss/gain* effects can degrade system performance, but they can be combated using static/dynamic spectral equalization [10,15]. *Signal transients* occur in metro networks because of the increased number of channel add/drops and in the case of protection switching, but the use of dynamic gain-controlled amplifiers and loss-controlled attenuators can reduce their impact [10,20,27]. *Filter concatenation*-induced distortion is a more severe problem since, in most cases, it cannot be compensated [28,29]. *In-band crosstalk* might also limit the size of network that can be supported. However, proper network design can reduce crosstalk-induced limitations [10,15]. *Polarization mode dispersion* is not considered to be a severe degradation, due to the low bit rates used in metropolitan-area networks [1]. *Fiber nonlinearities* are seldom probed in metro networks since the channel spacing is large and the injected power per channel in the fiber can remain low because of the small amplifier spacing [1]. *Chromatic dispersion* is another degrading effect in metro systems, especially with the use of low-cost transmitters (e.g., directly modulated lasers/DMLs and electroabsorption modulator-integrated DFB lasers/EA-DFBs) [32]. These lasers present *high-frequency chirp* (i.e., the optical frequency of the emitted signal varies rapidly with time depending on the changes of the optical powers) [1], which limits the uncompensated reach [32]. The dispersion/chirp impairment can be overcome by using dispersion-compensating modules or properly engineered optical fibers with special dispersion characteristics [16,32].

Some of the abovementioned degrading effects could be assigned to the transmission line (fiber) and others to the network node equipment. It is then critical to optimize the design of the transmission fiber, the network node architecture and the equipment used in the nodes in order to achieve optimum performance. In the following, we will see the specific desirable features that the metro network requirements impose to equipment (e.g., amplifiers), fibers and OADMs.

4.1. Metropolitan Amplifiers

The ever-changing traffic demands lead to the requirement for changes in the total number of WDM channels and also in the percentage of channels added or dropped at the OADM network nodes. Such dynamic network reconfigurations will more likely occur more frequently in metropolitan than in long-haul networks. Furthermore, protection and restoration mechanisms may force the termination of some channels or traffic in the network. The discussion above indicates that the amplifiers designed for metropolitan networks should have some unique requirements. The amplifiers have to deal with very dynamic networks where a large number of channels can be added and dropped, which means that amplifiers have to respond rapidly to these events by keeping their gain, and consequently the per channel power at the amplifier output, at a constant level independent of the number of wavelengths present in the network. Another requirement for metro amplifiers is operation over a wide gain range, since they have to deal with extremely wide variations in span length and node losses.

Consequently, key features for metro amplifiers are (1) variable-gain operation and (2) fast transient gain control. Such amplifiers will enable fast provisioning in dynamically reconfigurable networks, optical protection switching, and either longer system reach or more OADM nodes per ring, or larger percentage of add/drop capability per OADM node. In amplified optical networks, the use of erbium-doped fiber amplifiers (EDFAs) with dynamic gain-control capability [35] can significantly reduce the signal transients, and allow for tunable gain characteristics. Other amplification technologies based on semiconductor optical amplifiers (SOAs) have shown potential for use in metro networks [36].

4.2. MAN-Optimized Optical Fibers

Depending on the size, capacity, application, and terminal equipment used in metro networks, several different fiber types could be considered. Although advanced WDM metropolitan networks borrow heavily from concepts and practices originally developed and evaluated for their long-haul cousins, there are significant differences between the two application spaces, as we have pointed out. Among these is the necessity to support a broad spectrum of services and data rates in the MAN while keeping costs as low as possible. More specifically, metro fibers need to support:

- *1310-nm operation* — since the majority of legacy SONET systems operate at 1310 nm, the fiber type

used in metro networks should be able to operate at that wavelength range.

- *Full-spectrum (1260–1630-nm) coarse WDM (CWDM)* — CWDM continues to foster growing interest as a cost-effective means of enabling efficient bandwidth usage without the complexity and tight tolerances on optics associated with DWDM systems. The ability to have future CWDM compatibility built into metro networks reflects intelligent and proactive planning.
- *Large uncompensated/unregenerated reach for DWDM systems at 1550 nm* — long uncompensated/unregenerated reach with low-cost transmitters will enable the reduction of OEO regenerators and dispersion compensation modules and consequently will minimize the cost of the network.

Only selected fiber types can meet all the requirements for a metro fiber. For example, the most widely deployed single-mode fiber type (standard ITU G.652 fiber) does not satisfy all the requirements. Such fiber type has attenuation and dispersion characteristics, illustrated in Fig 6. Early versions of this fiber type may not be used for CWDM because of excess attenuation around 1380 nm. For networks requiring wavelength-band service differentiation and ability for CWDM, a fiber type with reduced attenuation at that wavelength range (e.g., Corning's SMF-28e fiber, OFS's Allwave fiber) should be the optimum choice.

A fiber with reduced water peak attenuation and optimized dispersion characteristics across the available fiber bandwidth should enable the use of cost-effective transmitters while ensuring compatibility with both CWDM and DWDM implementations. For example, the use of fibers with negative dispersion at the operating wavelength (Fig. 6b) enables the deployment of cost-effective directly modulated 10-Gbps transmitters for uncompensated reach comparable to that achieved by the best externally modulated sources over standard single-mode fiber [31]. Such a choice may enable cost-effective long-reach 10-Gigabit Ethernet transport solutions.

Nonzero dispersion-shifted fibers (NZ-DSF) also have great potential for use in metro networks. The lower absolute value of dispersion and the low attenuation across the operating wavelength window (see Fig. 6) will enable long uncompensated reach, legacy SONET 1310-nm operation, and simultaneous DWDM and CWDM operation.

4.3. OADM Architectures

Currently most of the metro-system integrators are building equipment that will support WDM OADM rings.

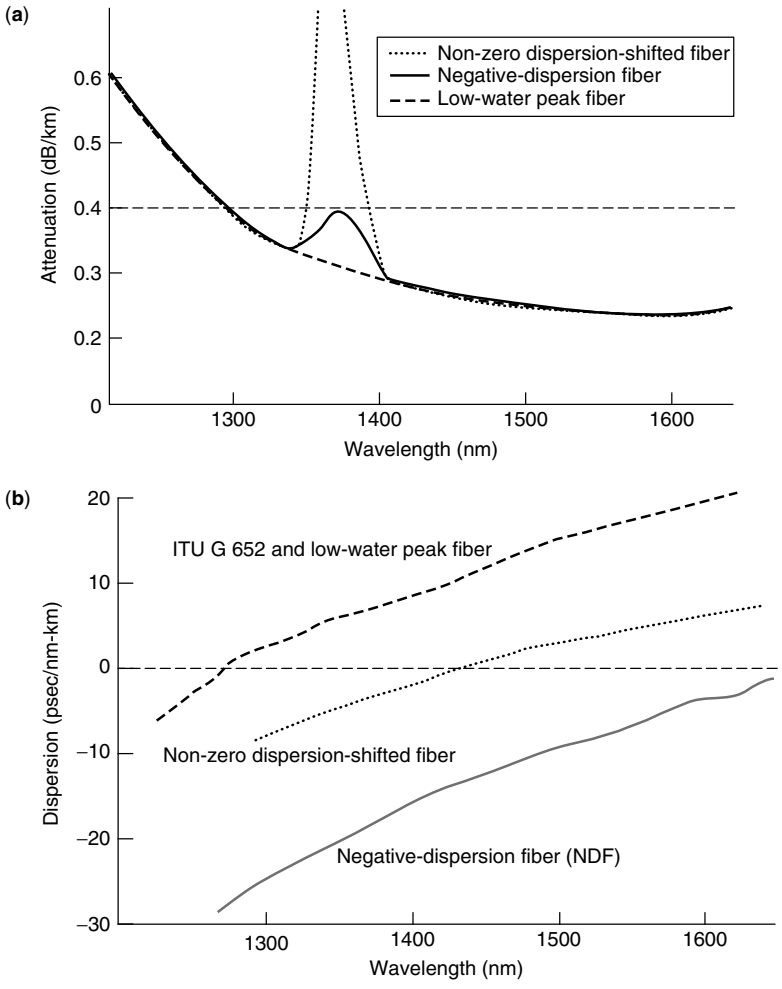


Figure 6. (**a**) Attenuation and (**b**) Dispersion characteristics of nonzero dispersion-shifted fiber, negative-dispersion fiber, and low-water peak fiber.

Because of the dynamic nature of data traffic originating in metro networks, OADM architectures should be designed for capability of large percentage of traffic add/drop at the network nodes. Although only a few channels may actually be added or dropped at each OADM, access to a large number of channels will allow for mesh connectivity of many nodes and will help avoid wavelength blocking issues with minimum pre-planning. Remote reconfiguration of the OADMs is also desirable [8].

Typically, OADM metro ring networks are built according to a banded wavelength channel plan [10,12]. The wavelengths are split into individual bands (typically 3–8 wavelengths each), with a guard-band spacing between the wavelength bands. The channels within the band can be spaced at frequencies of 25, 50, 100, or 200 GHz. The use of wavelength banding enables hierarchical multiplexing/demultiplexing at each network node, optical node bypassing on a per band basis, and scalable capacity upgrades and consequently reduced first installed costs [12]. The idea of node bypassing is very attractive since demultiplexing every wavelength at each node can be expensive, especially with many wavelengths per fiber and possibly many fibers per node. In addition to cost reduction of switching equipment, the wavelength banding will result in improved optical transmission performance since the insertion loss at each node for the passthrough channels will be reduced and also less filter concatenation effects are expected. The wavelength band-layered architecture consists of wavelength-band filters in combination with single-channel filters, and proper amplification to compensate for OADM losses (Fig. 7a). After the signals enter the OADM, some wavelength bands can be dropped using band-drop filters. The other bands (passthrough) experience a small loss by the band filters as they pass transparently through the OADM. The channels corresponding to the dropped bands are demultiplexed using single-channel demultiplexers and are directed to the receivers. In the add path a combination of single-channel multiplexers and band-add filters is used to add new traffic to the available channel slots. Variable optical attenuators (not shown in Fig. 7) can be used to match the power of the added channels to that of the passthrough channels. Optical amplifiers are placed at the input and the output of the OADM to compensate for losses. Wavelength-band layered OADM architectures meet most of the requirements of metro networks. However, when using such OADM architecture once the network is provisioned, the configuration is fixed, which induces limitation on the network flexibility. Furthermore, the wavelength-band layered OADM structure limits the maximum number of network nodes that can be supported for mesh connectivity.

More recently, a "broadcast & select" OADM architecture (Fig. 7b) has been proposed as an alternative to OADM applications where a large number of channels must be accessed [8,20]. The B&S OADM architecture consists of a 1×1 wavelength-selective device [e.g., Corning's PurePath dynamic spectral equalizer (DSE)] in combination with 1×2 power splitters/combiners to perform traffic add/drop, and proper amplification to compensate

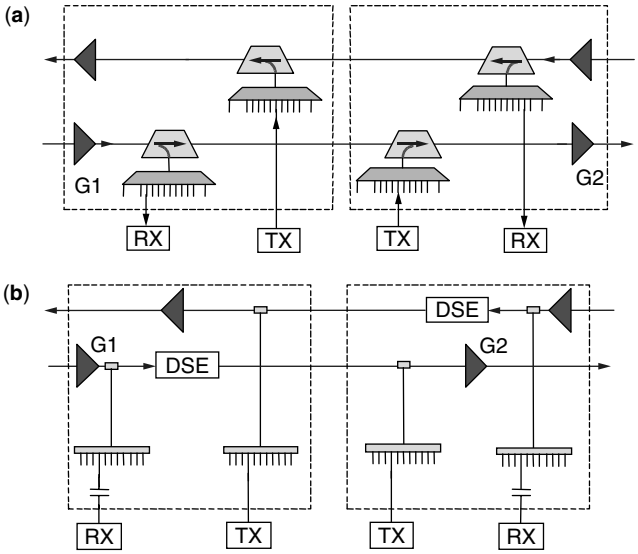


Figure 7. (**a**) Wavelength band layered and (**b**) broadcast & select OADM architectures.

for OADM losses. In this architecture (Fig. 7b), all incoming traffic is split into two paths for drop and passthrough. In the drop path, the dropped traffic is selected by a combination of a power splitter ($1 \times N$, where N is the number of simultaneously accessible DWDM channels) and tunable filters. In the passthrough path, the dropped channels are blocked by the DSE and the available channel slots can be filled by signals coming from the add path. The add path consists of N tunable transmitters and a $N \times 1$ power combiner. An EDFA is used to compensate for losses in the add path, and the amplifier noise in the empty channel slots is then filtered by another DSE. Both DSEs in this architecture can perform signal blocking for some of the channels and simultaneous power leveling for the pass through and added traffic. EDFAs are placed at the input and the output of the OADM to compensate for OADM losses. Such architecture has the advantages of keeping the passthrough loss to a minimum, while adding flexibility to the number of channels to be accessed. The B&S OADM architecture allows an arbitrary number of add/drop channels at each node and is dynamically reconfigurable [8].

5. SUMMARY

We have presented the characteristics, requirements, and architectures of optical metropolitan-area networks. The main features of system offerings proposed for metro networks were also discussed. We outlined our considerations regarding the evolution of metropolitan-area networks from SONET rings, to transparent wavelength-routed WDM rings. We also described the major transport impairments that limit the performance of transparent WDM metropolitan-area networks, and we presented the characteristics of MAN-optimized optical amplifiers, fibers, and OADMs. More recent research results demonstrate that a properly engineered network, utilizing application-optimized components and fiber, will

enable the buildup of cost-effective metrooptimized WDM optical networks that satisfy all the requirements and demands.

BIOGRAPHY

Ioannis Tomkos received the B.Sc. degree in Physics from University of Patras, Greece, and the M.Sc. degree in Telecommunications Engineering and Ph.D. degree in Optical Telecommunications from the University of Athens, Greece. His Ph.D. work focused on novel all-optical wavelength conversion technologies.

In 1996, he joined the Optical Communications Group of the University of Athens, Greece, as a Research Fellow, where he researched technologies for all-optical networks and digital transmission systems for access networks.

In January 2000, he joined the Photonics Research and Test Center of Corning Inc. as a Senior Research Scientist. He studied extensively the performance and design issues of metropolitan-area optical networks and successfully led several related projects.

Since September 2002, he has been an Associate Professor at the Athens Information Technology Center of Excellence in Research and Graduate Education, Athens, Greece. He is also an Adjunct Associate Professor at Carnegie–Mellon University, Pennsylvania (USA).

Professor Tomkos received the Best Paper Award from IEEE-LEOS in 1998. In 2002, he received the 2001 Corning Research Outstanding Publication Award. He has co-authored about 70 contributed and invited papers, has published in international journals and conference proceedings, and has several patent applications pending. He is a member of IEEE-LEOS and a member of the Optical Fiber Communication Conference (OFC) Technical Program Committee.

BIBLIOGRAPHY

1. G. P. Agrawal, *Fiber Optic Communication Systems*, Wiley, New York, 1992.
2. C. A. Brackett et al., A scalable multiwavelength multihop optical network: A proposal for research on all-optical networks, *J. Lightwave Technol.* **11**(5): 736–753 (May–June 1993).
3. S. B. Alexander et al., A precompetitive consortium on wide-band all-optical networks, *J. Lightwave Technol.* **11**(5): 714–735 (May–June 1993).
4. R. E. Wagner, R. C. Alferness, A. M. Saleh, and M. S. Goodman, MONET: Multiwavelength optical networking, *J. Lightwave Technol.* **14**(6): 1349–1355 (June 1996).
5. T. E. Stern and K. Bala, *Multiwavelength Optical Networks*, Prentice-Hall, Englewood Cliffs, NJ, 2000.
6. B. Mukherjee, *Optical Communication Networking*, McGraw-Hill, New York, 1997.
7. N. V. Srinivasan, Add-drop multiplexers and cross-connects for multiwavelength optical networking, *Proc. Optical Fiber Communication Conf., OFC '98*, 1998, 57–58.
8. A. Boscovic, M. Sharma, N. Antoniades, and M. Lee, Broadcast and select OADM nodes: Application and performance trade-offs, *Proc. Optical Fiber Communication Conf., OFC'02*, Paper TuX2, 2002, 158–159.
9. I. Tomkos et al., 80×10.7 Gb/s ultra-long-haul (4200 km) DWDM network with dynamic operation of broadcast & select reconfigurable OADMs, *Proc. Optical Fiber Communication Conf., OFC'02*, Paper FC-1, 2002.
10. I. Tomkos et al., Transport performance of an 80 Gb/s WDM ring network utilizing directly modulated lasers and uncompensated negative dispersion fiber, *IEEE/OSA J. Lightwave Technol.* **20**(4): 562–573 (April 2002).
11. M. D. Vaughn and R. E. Wagner, Metropolitan network traffic demand study, *Proc. Lasers and Electro-Optics Society IEEE Annual Meeting, LEOS 2000*, 2000, Vol. 1, pp. 102–103.
12. A. A. M. Saleh and J. M. Simmons, Architectural principles of optical regional and metropolitan access networks, *J. Lightwave Technol.* **17**(12): 2431–2448 (Dec. 1999).
13. D. Stoll, P. Leisching, H. Bock, and A. Richter, Metropolitan DWDM: A dynamically configurable ring for the KomNet field trial in Berlin, *IEEE Commun. Mag.* **39**(2): 106–113 (Feb. 2001).
14. P. Leisching et al., All-optical-networking at 0.8 Tb/s using reconfigurable optical add/drop multiplexers, *IEEE Photon. Technol. Lett.* **12**(7): 918–920 (July 2000).
15. N. Antoniades et al., Performance engineering and topological design of metro WDM optical networks using computer simulation, *IEEE J. Select. Areas Commun.* (Special Issue on WDM Based Network Architectures) **20**(1): 149–165 (Jan. 2002).
16. J.-P. Faure et al., A scalable transparent waveband-based optical metropolitan network, *Proc. Eur. Conf. Optical Communication, ECOC '01*, 2001, Vol. 6, 64–65.
17. K. C. Reichmann et al., An eight-wavelength 160-km transparent metro WDM ring network featuring cascaded erbium-doped waveguide amplifiers, *IEEE Photon. Technol. Lett.* **13**(10): 1130–1132 (Oct. 2001).
18. P. Arijs et al., Architecture and design of optical channel protected ring networks, *J. Lightwave Technol.* **19**(1): 11–22 (Jan. 2001).
19. D. Tebben et al., Two-fiber optical shared protection ring with bi-directional 4/spl times/4 optical switch fabrics, *Proc. Annual Meeting of Lasers and Electro-Optics Society, LEOS 2001*, 2001, Vol. 1, pp. 228–229.
20. J.-K. Rhee et al., A novel 240-Gbps channel-by-channel dedicated optical protection ring network using wavelength selective switches, *Proc. Optical Fiber Communication Conf., OFC'01*, Paper PD-38, 2001.
21. P. Bonenfant and A. Rodriguez-Moral, Framing techniques for IP over fiber, *IEEE Network* **15**(4): 12–18 (July–Aug. 2001).
22. Yao Shun, S. J. B. Yoo, B. Mukherjee, and S. Dixit, All-optical packet switching for metropolitan area networks: Opportunities and challenges, *IEEE Commun. Mag.* **39**(3): 142–148 (March 2001).
23. K. V. Shrikhande et al., HORNET: A packet-over-WDM multiple access metropolitan area ring network, *IEEE J. Select. Areas Commun.* **18**(10): 2004–2016 (Oct. 2000).
24. A. Jourdan et al., The perspective of optical packet switching in IP dominant backbone and metropolitan networks, *IEEE Commun. Mag.* **39**(3): 136–141 (March 2001).

25. N. Le Sauze et al., A novel, low cost optical packet metropolitan ring architecture, *Proc. Eur. Conf. Optical Communication, ECOC '01*, 2001, Vol. 6, pp. 66–67.

26. N. Antoniades, M. Yadlowsky, and V. L. daSilva, Computer simulation of a metro WDM interconnected ring network, *IEEE Photon. Technol. Lett.* **12**(11): 1576–1578 (Nov. 2000).

27. J.-K. Rhee, I. Tomkos, P. Iydroose, and M.-J. Li, Optical transient effect of dedicated optical channel protection in a two fiber ring network, *Proc. Lasers and Electro-Optics Society IEEE Annual Meeting, LEOS 2001*, 2001, Vol. 1, pp. 226–227.

28. J. Downie et al., Effects of filter concatenation for directly modulated transmission lasers at 2.5 and 10 Gb/s, *IEEE/LEOS J. Lightwave Technol.* **20**(2): 218–228 (Feb. 2002).

29. I. Tomkos et al., Filter concatenation penalties for 10 Gb/s sources suitable for WDM metropolitan area networks, *IEEE Photon. Technol. Lett.* **14**(4): 564–566 (April 2002).

30. C.-C. Wang et al., Negative dispersion fibers for uncompensated metropolitan networks, *Proc. Eur. Conf. Optical Communication, ECOC'00*, Munich, Germany, Sept. 2000, Vol. 1, pp. 70–71.

31. I. Tomkos, R. Hesse, R. Vodhanel, and A. Boskovic, A 320 Gb/s metropolitan area ring network utilizing 10 Gb/s directly modulated lasers, *IEEE Photon. Technol. Lett.* **14**(3): 408–410 (March 2002).

32. I. Tomkos et al., Demonstration of negative dispersion fibers for DWDM metropolitan area networks, *IEEE/LEOS J. Select. Top. Quant. Electron.* (Special Issue on Specialty Fibers) **7**(3): 439–460 (May–June 2001).

33. A. Richter et al., Optical performance monitoring in transparent and configurable DWDM networks, *IEE Proc. Optoelectron.* **149**(1): 1–5 (Feb. 2002).

34. E. Park, Error monitoring for optical metropolitan network services, *IEEE Commun. Mag.* **40**(2): 104–109 (Feb. 2002).

35. M. Lelic et al., Smart EDFA with embedded control, *Proc. Annual Meeting of the Lasers and Electro-Optics Society, LEOS 2001*, 2001, Vol. 2, pp. 419–420.

36. D. A. Francis, S. P. DiJaili, and J. D. Walker, A single-chip linear optical amplifier, *Proc. Optical Fiber Communication Conf., OFC 2001*, 2001, Vol. 4, pp. PD13, pp. 1–3.

WIDEBAND CDMA IN THIRD-GENERATION CELLULAR COMMUNICATION SYSTEMS

JON W. MARK
University of Waterloo
Waterloo, Ontario, Canada

SHIHUA ZHU
Xian Jiaotong University
Xian, Shaanxi
People's Republic of China

1. INTRODUCTION

The second-generation (2G) CDMA standard, IS-95, is a narrowband CDMA standard. The third-generation (3G) CDMA systems will have a target transmission rate of 2 Mbps (megabits per second). ITU (International Telecommunications Union), the international standards body, has adopted both UMTS/IMT-2000 (Universal Mobile Telecommunications System/International Mobile Telecommunications by the year 2000) and CDMA-2000 as 3G network access technologies.

The impetus behind the popularization of wireless communications is the flexibility it offers for mobile users to roam. For wireless systems to be economically feasible, they must be able to deploy low-power transmitters and offer high system capacity, that is, be able to support a large population with a high transmission rate. These are the reasons why radio cells in wireless systems are relatively small (e.g., microcells) and arranged as a cellular structure. Communications by mobile users in a cellular environment encounters a number of challenging problems. These include the need to (1) expand the spectral width of the transmission channel, (2) manage the available resources efficiently, and (3) manage the user mobility effectively and efficiently. The 3G standards have specifications to address the above-mentioned problems. Also, in CDMA systems, there is a need for power control, at least to combat the near–far effects [6,7,11]. Because of space limitation, in this article we mainly focus attention on the network architecture and signaling strategies of IMT-2000 wideband CDMA (WCDMA).

2. NETWORK ARCHITECTURE OF IMT-2000

The UMTS/IMT-2000 architecture is shown in Fig. 1. The symbols used in Fig. 1 are listed in Table 1. The core network (CN) includes at least the circuit-switched mobile switching center (MSC) and the packet-switched packet radio service support node [2].

2.1. Functional Characteristics

In the UMTS/IMT-2000 system, the functional layering introduces the concepts of access stratum and nonaccess stratum. All the functional blocks in the UTRAN belong to the access stratum. The UTRAN handles all radio-specific procedures, whereas the core network handles the service-specific procedures, including mobility management and call control. The other functional characteristics are as follows:

- The link Uu (see Fig. 1) connects the two physical layers (UE and UTRAN) to provide a transparent digital signal path to the upper layers. The physical layer occupies certain bandwidth, and is responsible for synchronization.
- The MAC layer resolves the contention resulting from the radioband sharing between a number of users. It also maps the data from the link control layer to the physical layer, and vice versa.
- The RLC provides reliable data packet transmission over the radio link. The signaling from the control plane can be regarded as a special data packet requiring higher priority and short delay.
- The LAC provides protection to the user data.
- The RRC performs the radio resource management, and allocates the transmission speed and power for each connection.

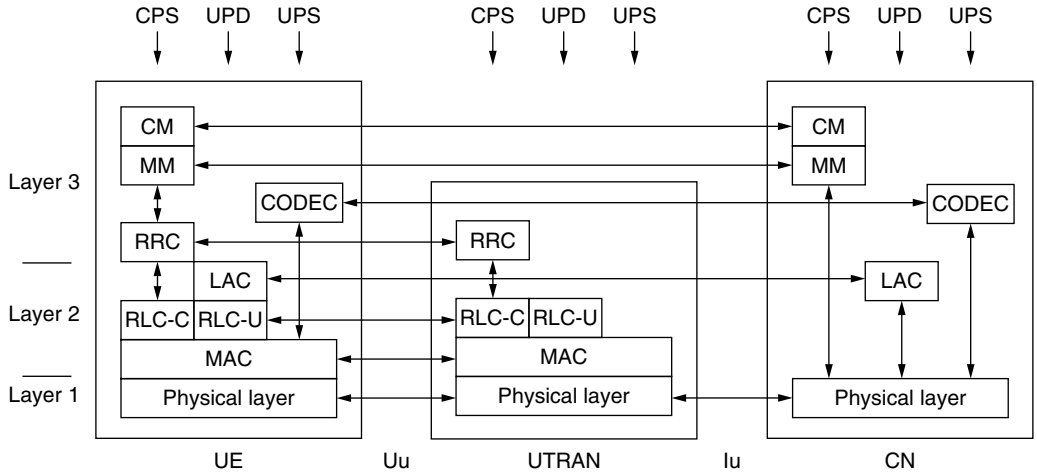


Figure 1. UMTS/IMT2000 system architecture and protocol layering.

Table 1. Definitions of Symbols Used in Fig. 1

Symbol	Description
UE	User equipment
UTRAN	UMTS terrestrial radio-access network
CN	Core network
Uu	Radio interface between UTRAN and UE
Iu	Interface between UTRAN and CN
CPS	C plane, signaling
UPD	U plane, data
UPS	U plane, speech
CM	Connection management
MM	Mobility management
LAC	Link-access control
RLC-C	Radio-link control for the control plane
RLC-U	Radio-link control for the user plane
RRC	Radio resource management and rate allocation
MAC	Medium-access control

- The MM maintains a record of the positions of all the users in the system and provides the updated user position information during a call.
- The CM grants or rejects an incoming call, either newly generated or handed over from an adjacent cell, based on the channel resources available at the time of call arrival.

3. PRINCIPLE OF CDMA

3.1. Orthogonal Spreading

It is well known [9] that in CDMA, if the spreading signals, $\phi_i(t) \in L^2(T), i = 1, 2, \ldots, M$, are orthogonal:

$$\int_0^T \phi_k(t)\phi_l(t)\,dt = \delta_{kl} \triangleq \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases} \tag{1}$$

where T is the signal duration; the spread signals $x_i\phi_i(t)$, $i = 1, 2, \ldots, M$, will also be orthogonal to each other. Here x_i is the data bit of the ith user, which is kept constant during T. Consequently, the interference from any other signal, say, $x_j\phi_j(t)$, on the wanted signal, say, $x_i\phi_i(t)$, is theoretically zero after despreading:

$$I_{ij} = \int_0^T x_j\phi_j(t)\phi_i(t)\,dt = x_j \int_0^T \phi_j(t)\phi_i(t)\,dt = 0, \quad i \neq j \tag{2}$$

where $\phi_i(t)$ is the despreading function used to extract the desired data sequence, $\{x_i\}$, transmitted by the ith user.

It is not difficult to design a set of orthogonal functions or codes. Some good examples are Hadamard–Walsh codes [4]. However, when these codes are not properly aligned, the cross-correlation of these codes will be nonzero, and may be relatively large compared to the autocorrelation; that is, the signals are no longer orthogonal.

Thus, to eliminate the interference from unwanted signals, the local despreading code is required to be accurately aligned with the arriving wanted code. This demands that all the signals arriving at every receiver be synchronized at the bit level and with the spreading code. In other words, if it is desired to avoid interferences from other users using the orthogonality property, all the received signals must be bit-synchronized with the local timing that is locked to the spreading code of the incoming wanted signal. For practical systems in which mobile terminals are constantly in motion, this is obviously a very costly requirement. It is desirable to have a type of code that can separate channels and yet requires no synchronization.

To date, no orthogonal code requiring no synchronization has been found. In frequency diversity methods such as FDMA, if a certain amount of interference from adjacent bands is tolerable, a less sharp cutoff filter can be employed, which can greatly simplify the filter design. Similarly, in CDMA if nonperfect channelization is tolerable, nonorthogonal codes that require no synchronization may be used.

3.2. Nonorthogonal Spreading

The nonorthogonal codes when used for spreading must require no synchronization at the bit level and with the

spreading code. To reduce the interference from other transmissions to a minimum, the cross-correlation, Γ_c, between any pair of codes of the code set at any time shift should be small; thus, we require

$$\Gamma_c = \int_0^T \phi_k(t)[\phi_l(t - T + \tau) + \phi_l(t + \tau)]\, dt \ll 1 \quad 0 \le \tau < T \tag{3}$$

Besides, for reasons to be explained in the next section, the codes are required to be balanced; that is, the number of ones must approximately equal the number of zeros. Gold codes and Kasami codes have such properties [4,5,10]. Kasami codes consist of two classes: the large Kasami code sets and the small Kasami code sets. The large Kasami code sets are of more interest to 3G CDMA applications.

3.3. Two-Layer Spreading

In two-layer spreading, the transmitter signal is spread in the first layer using orthogonal codes. The output from the first layer is then spread again using nonorthogonal codes. Orthogonal codes are used in the first layer spreading for reasons that synchronization between channels can be easily maintained. Nonorthogonal codes are used for the second layer to reduce implementation complexity.

Two-layer spreading is illustrated in Fig. 2, where the set of parallel signals, $\{x_i(t), i = 1, 2, \ldots, K\}$, belongs to one transmitter. All the channels originating from a single transmitter can be readily synchronized. Therefore, orthogonal codes $(C_{o1}, C_{o2}, \ldots, C_{oK})$, are used to separate these channels. These channels are then linearly combined and multiplied by a transmitter-specific nonorthogonal code.

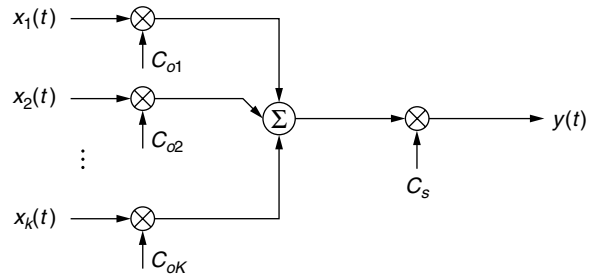


Figure 2. Two-layer spreading.

In two-layer spreading, every transmitter uses the same orthogonal code set. Let the data signal on the kth channel from the mth transmitter be represented by

$$x_{mk}(t) = \sum_i b_{mki} g_{T_b}(t - iT_b) \tag{4}$$

where T_b is the bit duration, $b_{mki} \in \{-1, 1\}$ is the ith bit of the kth channel from the mth transmitter, and $g_{T_b}(t)$ is a rectangular pulse of duration T_b:

$$g_{T_b}(t) = \begin{cases} 1 & 0 \le t < T_b \\ 0 & \text{elsewhere} \end{cases} \tag{5}$$

Let $c_{okj} \in \{-1, 1\}$ be the jth chip of the kth orthogonal code. Then, the orthogonal spreading code used for the kth channel can be expressed as

$$c_{ok}(t) = \sum_{j=1}^{N} c_{okj} g_{T_c}(t - jT_c) \tag{6}$$

where T_c is the chip width and N is the code length. Normally, the values of T_b and T_c are chosen to yield $T_b/T_c = N$, resulting in the same code repeatedly multiplied onto each data bit of the channel.

The length of the nonorthogonal codes is normally many times the length of orthogonal codes. They can be a set of codes, or a set of segments of some long code. Their chip rate is the same as that of the orthogonal code: $1/T_c$. The nonorthogonal signal can be written as

$$c_{sm}(t) = \sum_{l=1}^{N'} c_{sml} g_{T_c}(t - lT_c) \tag{7}$$

where $c_{sml} \in \{-1, 1\}$ is the lth chip of the nonorthogonal code for the mth transmitter and N' is the code length. The output from the two-layer spreading is then given by

$$y_m(t) = \sum_{k,i} b_{mki} c_{ok}(t - iT_b) \sum_p c_{sm}(t - pT') \tag{8}$$

where $T' = N'T_c$ is the nonorthogonal code duration. Figure 3 shows the timing of the waveforms in a two-layer spreading system, where $x_{mk}(t)$ is the kth channel signal from the mth transmitter.

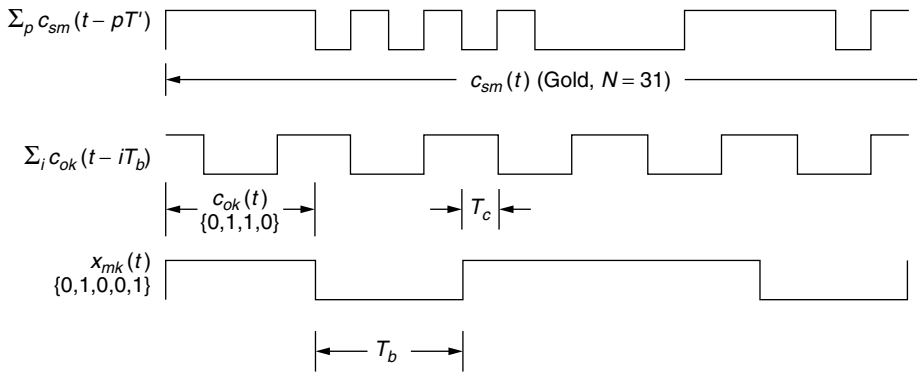


Figure 3. Timing of two-layer spreading.

In a CDMA system, many users transmit signals in the same band. The received signal (in the absence of channel impairments and background noise) is given by

$$r(t) = \sum_m \sum_{k,i} b_{mki} c_{ok}(t - \tau_m - iT_b) \sum_p c_{sm}(t - \tau_m - pT') \tag{9}$$

where τ_m is the transmission delay from the mth transmitter to the target receiver.

To receive the desired signal, say, $x_1(t)$ of transmitter 1, the receiver must be synchronized to the nonorthogonal code frame; that is, when the receiver achieves synchronization, $\tau_1 = 0$. The output from the nonorthogonal despreading is

$$\begin{aligned} z'(t) &= r(t) \sum_p c_{s1}(t - pT') \\ &= \sum_{k,i} b_{1ki} c_{ok}(t - iT_b) \\ &\quad + \sum_{m \neq 1} \sum_{k,i} b_{mki} c_{ok}(t - \tau_m - iT_b) \\ &\quad \times \sum_p c_{sm}(t - \tau_m - pT') c_{s1}(t - pT') \end{aligned} \tag{10}$$

After orthogonal despreading, the output becomes

$$\begin{aligned} z(t) &= z'(t) \sum_i c_{o1}(t - iT_b) \\ &= \sum_i b_{11i} + \sum_{k \neq 1, i} b_{1ki} c_{ok}(t - iT_b) c_{o1}(t - iT_b) \\ &\quad + \sum_{m \neq 1} \sum_{k,i} b_{mki} c_{ok}(t - \tau_m - iT_b) c_{o1}(t - iT_b) \\ &\quad \cdot \sum_p c_{sm}(t - \tau_m - pT') c_{s1}(t - pT') \end{aligned} \tag{11}$$

The first term on the RHS of the second equality in (11) is the wanted signal. The second term represents the interferences arising from the other channels of the same transmitter, which vanishes after integration due to the orthogonality between $c_{o1}(t - iT_b)$ and $c_{ok}(t - iT_b), k \neq 1$. The third term is the interference contribution from the other transmitters. Since the transmitters are not synchronized, the first part of the product may not be zero after integration; it may even be 1 (suppose $k = 1$ and $\tau_m = iT$). This is equivalent to that, at any time t, the probability of the signal being 1 and being -1 is not equal. But if $c_{sm}(t - \tau_m - pT')$ and $c_{s1}(t - pT')$ have small correlation values and are balanced, namely, at any time t, the probability of the signal being 1 is approximately the same as it being -1. After multiplication, the product is approximately balanced and approaches zero after integration. This is the scrambling technique normally used in data transmission. For this reason, the nonorthogonal codes are called *scrambling codes*. Similarly, the orthogonal codes are called *channelization codes*.

It can be seen from Fig. 3 that when the spreading factor is small, there is a possibility that the segment of the scrambling code corresponding to a bit interval (or a channelization code period) is far from balanced. If we assume that the number of interfering transmitters is large and the interferences are independent of each other, the statistics of the interferences can still be regarded as random noise. It is certain, however, that a smaller spreading factor will give a worse scrambling effect than will a larger spreading factor.

3.4. Spreading in IMT-2000

UMTS/IMT-2000 employs two-layer spreading. For the downlink, every base station uses the same set of Hadamard codes or OVSF (orthogonal variable spreading factor) codes as its channelization codes. Each cell uses a cell-specific Gold code as its scrambling code.

For the uplink, every active mobile UE establishes a connection with the base station. The connection may consist of one or several channels. Every UE or connection uses the same Hadamard/OVSF codes as its channelization code. Each connection is then distinguished by using a user-specific Gold code or Kasami code as its scrambling code.

Obviously, since there are many more downlink channels from a single base station than uplink channels from a single UE, a much larger set of channelization codes for the downlink transmission is needed. Similarly, since the number of UEs in a cellular system is much larger than that of base stations, a larger scrambling code set for the uplink transmission is required.

In UMTS/IMT-2000, the same OVSF codes are used for both links. Since the maximum capacity of OVSF codes has been proved to be bounded by the transmission rate, special attention must be paid to the downlink channelization code usage, so that more connections can be accommodated by the given channelization code set. This has led to different structures for the uplink and downlink physical channels (see Section 4.2).

The scrambling codes used for the downlink is a set of computer-selected 10-ms segments of a $(2^{18} - 1)$-chip-long Gold sequence. Each segment is equivalent to 40,960 chips for the recommended 4.096-Mchip/s[1] transmission rate at a carrier spacing of 5 MHz. For the uplink, a larger number of scrambling codes is required. They can be either a set of the extended 256-chip-long large Kasami codes, or optionally, a set of 10 ms (40,960 chips) segments selected from a $(2^{41} - 1)$-chip-long Gold sequence [4].

4. PHYSICAL CHANNELS

4.1. Physical Channel Types

UMTS/IMT-2000 defines the following physical channel types (see Fig. 1 for references to protocol layers):

[1] The originally recommended chip rate for IMT-2000 was 4.096 Mchips/s at a bandwidth of 5 MHz. In the move toward harmonized global 3G (G3G), ITU adopted a compromised chip rate of 3.84 Mchips/s for 3G. For the purpose of discussing IMT-2000 spreading in this article, we continue to use 4.096 Mchips/s as the reference. For the 3.84-Mchip/s rate, proper scaling can easily be accommodated.

- Synchronization channel (SCH) — used for initial cell search and link synchronization by the UE so that the data and control channels can be properly despreaded (downlink only).
- Dedicated physical data/control channel (DPDCH/DPCCH) — each connection between a dedicated UE and the network is allocated one DPCCH and zero, one, or several DPDCHs. The DPDCH is used to carry the data generated at layer 2 and above, dedicated to a single UE; the DPCCH is used to carry layer 1 control information relevant to the DPDCH.
- Common control physical channel (CCPCH) — used to broadcast common control information from layers 2 and 3 to all UEs of a cell or the whole system (downlink only).
- Physical random-access channel (PRACH) — provides the UEs fast access of short data packets to the network. This is a supplement to the DPDCH/DPCCH and is particularly valuable for control information and signaling transmission (uplink only).

4.2. Dedicated Physical Data/Control Channels

The *frame structures* of IMT-2000 are shown in Figs. 4 and 5, with the following parameter values:

- Data sequence is divided into frames of length $T_f = 10$ ms.
- Each frame is further divided into 16 slots of equal duration $T_s = 0.625$ ms.
- The uplink DPDCH and DPCCH each occupies a separate channel, or one frame per 10 ms. Each slot of the DPCCH is divided into three fields, for the known pilot bits, transmit-power-control (TPC) command, and transport format indicator (TFI), respectively, as shown in Fig. 4. The pilot bits are used for downlink channel estimation. The reason for a dedicated pilot instead of a common pilot is to support the use of adaptive antenna arrays. The TPC is used to perform closed-loop power control at a frequency of 1600 Hz. The TFI (optional) is used to inform the receiver which transmit format the layer 2 data is used in the current DPDCH.
- The downlink DPDCH and DPCCH are multiplexed within each radio frame as shown in Fig. 5. The main purpose of this multiplexing is to let the two channels share a single channelization code.

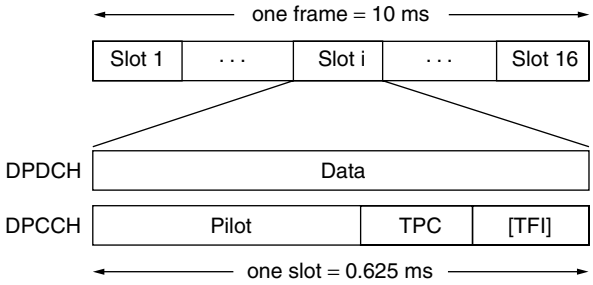


Figure 4. Uplink DPDCH/DPCCH frame structure.

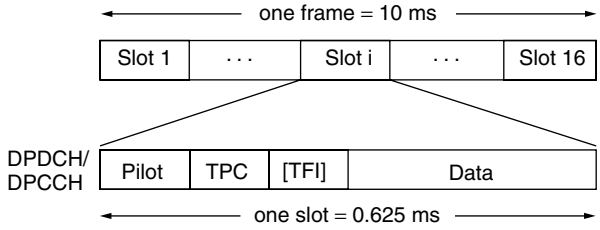


Figure 5. Downlink DPDCH/DPCCH frame structure.

4.2.1. Variable Orthogonal Spreading

- On the uplink (Fig. 6), the allowable data bits carried by each slot are $n_{su} = 10,20,40,80,160,320,640$ $(10 \times 2^k, k = 0, 1, \ldots, 6)$ bits. These correspond to channel data rates $R_{bu} = n_{su}/T_s = 16,32,64,128,256,512,1024$ $(16 \times 2^k, k = 0, 1, \ldots, 6)$ kbps (kilobits per second). The spreading output is kept at a constant rate $R_c = 4.096$ Mchips/s, resulting in variable spreading factors (SF) of $R_c/R_{bu} = 256,128,64,32,16,8,4$ $(256/2^k, k = 0, 1, \ldots, 6)$.
- On the downlink, since the DPDCH and DPCCH are multiplexed in each slot (see Fig. 5), the number of bits carried by each slot must be doubled in order to offer the same transmission capacity for both links. Consequently, the allowable data bits carried by each slot should be $n_{sd} = 20,40,80,160,320,640,1280$ $(20 \times 2^k, k = 0, 1, \ldots, 6)$ bits. These correspond to channel data rates $R_{bd} = n_{sd}/T_s = 32,64,128,256,512,1024,2048$ $(32 \times 2^k, k = 0, 1, \ldots, 6)$ kbps. The spreading output rate is constant: $R_c = 4.096$ Mchips/s. To obtain the same spreading factors as in the uplink, the serial datastream is first converted to two parallel sequences (see Fig. 7) before being sent for spreading. The resulting variable spreading factors are again $R_c/(R_{bd}/2) = 256,128,64,32,16,8,4$ $(256/2^k, k = 0, 1, \ldots, 6)$.
- The channelization codes available for each connection on the uplink and for each base station on the downlink are the same set of OVSF codes. Within an uplink connection, each DPDCH or DPCCH requires a unique channelization code for orthogonal spreading. On the downlink, each multiplexed DPDCH/DPCCH requires a unique channelization code. These codes are assigned by the network and

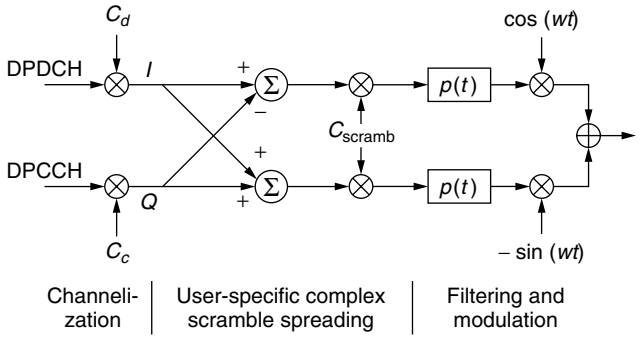


Figure 6. Uplink spreading and modulation.

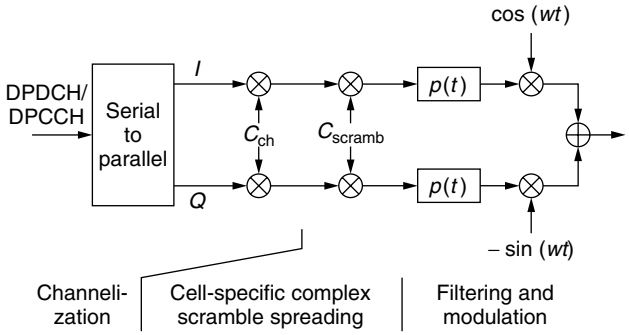


Figure 7. Downlink spreading and modulation.

may be changed dynamically, frame by frame, during the connection.

4.2.2. Scrambling Code Spreading. For the uplink, complex (m phase and quadrature) scrambling code spreading, as shown in Fig. 6, is used. The reason for using the complex spreading is to balance the loads on the I and Q branches. To control the amount of overhead introduced by using the DPCCH, the power of the DPCCH is to be kept to a minimum. Consequently, the power of the DPCCH is normally smaller than that of the DPDCH. Besides, this difference can vary with the traffic type carried on the DPDCH. For instance, the relative power difference between the DPCCH and DPDCH for speech and 384-kbps data are 3 and 10 dB, respectively. Another advantage of this complex spreading is that additional DPDCH can be easily added to the connection by adding it to both I and Q branches after the independent channelization spreading.

The scrambling code for both I and Q branches is the same code. It can be either 256-chip-long extended codes from the VL Kasami set[2] of length 255, or a 40,960-chip (10-ms) segment of a Gold code of length $2^{41} - 1$ (when an ordinary RAKE receiver is used). The scrambling code on the uplink is user-specific. It is assigned by the network and may be changed dynamically during the connection.

For the downlink, because of the scarcity of channelization codes, the DPDCH and DPCCH are first multiplexed, as shown in Fig. 5, and then spread by a single channelization code, as shown in Fig. 7. The serial DPDCH/DPCCH datastream is converted into two parallel paths, I and Q paths, so that the data rate on each path is halved, resulting in the same data rates as in the uplink case. The two branches are then spread by the same channelization code and scrambling code.

The scrambling code for the downlink is a 40,960-chip (10-ms) segment of a Gold code of length $2^{18} - 1$. There are 512 different segments, which are divided into 32 groups, each consisting of 16 codes. Each cell is assigned a specific downlink code at the initial deployment [3].

4.2.3. Filtering and Modulation. The filter $p(t)$ is a root-raised cosine function with a rolloff factor of 0.22. The modulation scheme is quaternary phase shift keying (QPSK).

[2] The VL Kasami code is simpler and the cross-correlation properties are maintained between symbols, but has worse interference averaging properties. Its use is more appropriate when multiuser detection is used.

4.3. Synchronization Channels

To access the network, a mobile UE must be synchronized to the desired cell to detect the information. From the description in the last section, the UE must first establish the scrambling code timing in order to decode the information sent from the cell site. Since the scrambling code boundary is aligned with the bit or channelization code boundary, orthogonal despreading timing can be readily obtained, once the scrambling code timing is derived.

The most straightforward method of acquisition of the scrambling code timing is to correlate the locally generated scrambling sequence with the incoming signal. The search can be performed by shifting the local sequence phase one-half a chip at a time, until the complete period is searched. In the worst case, the mobile UE may need to search all the possible scrambling codes used by the base stations. Since the scrambling codes must be sufficient to provide cell-site separation, the number of scrambling codes is usually very large. This makes the above searching algorithm extremely time-consuming, usually too slow to be acceptable for normal operation. To cope with the difficulty, a special mechanism has been developed for UMTS/IMT-2000.

4.3.1. Synchronization Channel Types. The system employs two types of synchronization channels (SCHs): the primary SCH and the secondary SCH. The primary SCH transmits a 256-chip-long Gold sequence, called the *primary synchronization code* (PSC), in each slot (see Fig. 8). The sequence is system-specific and is predetermined for the whole system; that is, it is transmitted by every base station in the system. It is not spread further by any other spreading code but is superimposed on the scrambled downlink datastream. At the start of a slot, the channel transmits 0.0625 ms of this 256-chip sequence (256 bits/4096 kbps), pauses, and transmits the code again when the next slot starts. The code is transmitted time-aligned with the slot boundary so that by acquiring the PSC, the mobile UE acquires slot synchronization to the target base station.

The system allocates 16 Gold codes of 256 chip length, called the *secondary synchronization code* (SSC), for indicating the 16 scrambling code groups used for downlink data and control channels (see Section 4.2.1).

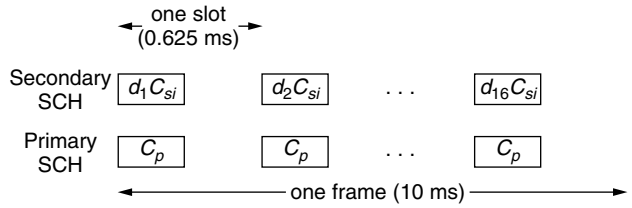


Figure 8. Synchronization channel frames (*key*: C_p — primary synchronization code; C_{si} — the ith code of the 16 possible secondary synchronization codes; d_j — the jth bit of the 16-bit sequence).

Each base station is assigned a SSC by the network. The code is transmitted once per slot on its secondary SCH. As in the primary SCH, each code occupies 0.0625 ms (see Fig. 8). The SSC is further modulated by a 16-bit system-specific sequence in such a way that each 256-chip code in a slot is modulated by one bit of the 16-bit sequence, resulting in a repeat pattern every 16 slots or one frame. By correlating the local code to the SSC, the UE can determine which code group the target cell uses, and by demodulating the 16-bit sequence, the frame timing is obtained.

4.3.2. Initial Cell Search and Synchronization. The process of the initial cell search and synchronization by a mobile UE is done in three steps:

1. The UE receiver has a matched filter, matched to the primary SCH code, C_p. The output of the filter responding to the C_p sent from any cell is therefore a sequence of narrow peaks spaced one slot (0.625 ms) apart. Since all the cells in the system are transmitting the same C_p, the matched-filter output is a superposition of many such sequences. The UE selects the one with the largest peak, and takes the corresponding cell as its working base-station cell. This peak sequence also provides the UE with the slot timing for that cell.
2. With the local slot boundary aligned to the cell slot boundary obtained in step 1, the UE tries to correlate each of the 16 possible secondary SCH codes, $C_{si}, i = 1, 2, \ldots, 16$, with the received signal. The correlation is carried over a frame, and for each possible SSC the searching process is done by shifting one slot at a time. This results in a maximum of 16 correlations per code, or 256 correlations in total, for the UE to search through all the possible SSCs and the possible 16-bit sequence phases. When the code selected by the UE is matched to the secondary SCH code used by its working cell, and the local 16-bit sequence is matched to the incoming one, the search is complete. The code group that the cell uses for its downlink data and control channels is determined from the matched Gold code, while the frame timing is obtained by the 16-bit sequence phase.
3. Once the code group used by the cell for its downlink data and control channels is determined, the UE tries all the 32 codes within the group. (*Note*: These are 40,960-chip segments of a Gold code, which should not be confused with the 16 codes used by the secondary SCH, see the last section.) This is done by correlating each code with the primary CCPCH. The primary CCPCH uses a system-specific channelization code and has a fixed predefined data rate, which makes the scrambling code detection the simplest task among all the data and control channels. The search is performed through symbol-by-symbol correlation over a frame. Once the correlation operation is finished, the downlink scrambling code used by the target cell is determined and information can then be retrieved.

4.4. Common Control Physical Channels

There are two types of common control physical channels: the primary common control physical channel (primary CCPCH) and the secondary common control physical channel (secondary CCPCH). The primary CCPCH is used to broadcast system-specific information, while the secondary CCPCH is used to broadcast cell-specific information.

The framing, channelization, and scrambling for the CCPCHs are done in the same way as for the downlink time-multiplexed DPDCH/DPCCH, but with the following exceptions:

- There is no TPC field (see Fig. 5). Since it is a common point-to-multipoint broadcast channel, no closed-loop power control can be applied. Also, since both the primary and secondary CCPCHs have a fixed data rate and transport format, there is also no need for TFI. As a result, the layer 1 control information of the downlink CCPCHs (equivalent to the DPCCH of the dedicated physical channel) conveys pilot bits only.
- The primary CCPCH uses a predefined system-specific channelization code of length 256 chips common to all cells. The secondary CCPCH uses a cell-specific code. A cell can have one or several secondary CCPCHs. The data rate for each channel is fixed during the transmission but may vary for different secondary CCPCHs within the cell and between cells. The channelization code used and the data rate carried by each secondary CCPCH is indicated by the information broadcast on the primary CCPCH (instead of on the TFI field of the layer 1 control information channel).

4.5. Physical Random-Access Channels

The physical random-access channel (PRACH) is used to carry random-access bursts and short packets in the uplink [2]. It is a common channel shared among all users in the cell. The random-access burst structure of the PRACH is shown in Fig. 9.

The preamble part is a 16-bit word spread by a cell-specific 256-chip-long Gold code that is indicated on the primary CCPCH. The rest is called the data part of the burst consisting of a UE ID, a "requested service" field, an optional user data packet, and a CRC field. Spreading and modulation of the data part is similar to the uplink dedicated channels. (So the length of the whole burst is preample $+n \times 10$ ms $= 16 \times 256/4096 + 10n$ ms $= 1 + 10n$ ms.) The operation of the PRACH will be described in Section 5.3.

5. TRANSPORT CHANNELS

5.1. Transport Channel Types

The physical layer offers information transfer services to the MAC layer. These services are accomplished by

Preamble	Mobile station ID	Requested service	User packet	CRC

Figure 9. Ransom-access burst structure of the PRACH.

conveying MAC layer channels on the physical channels. These MAC layer channels, including those listed below, are denoted as transport channels (TrCh's).

- *Broadcast channel* (BCCH) — used for downlink only. With a fixed rate, it is used to broadcast system information to all the mobile UEs in its cell.
- *Paging channel* (PCH) — used for downlink only; also used for paging in the whole cell.
- *Forward-access channel* (FACH) — employed for downlink only. It is used to convey data to one or more UEs, which may be over a part of the cell using beamforming.
- *Random-access channel* (RACH) — used for uplink only. It is used by the UE to transmit short user data packets and control packets (e.g., for initiating packet transfer on the DCHs).

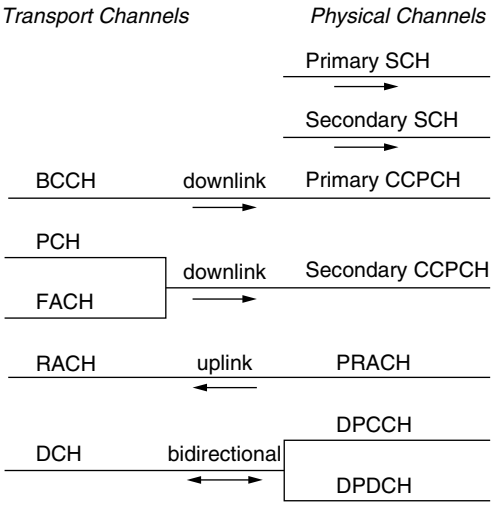


Figure 10. Mapping of transport channels to physical channels in IMT-2000.

- *Dedicated channel* (DCH) — a bidirectional channel. It is a point-to-point channel used to convey data to/from a UE. Beamforming can be used to achieve this.

5.2. Mapping of Transport Channels to Physical Channels

5.2.1. Mapping Relation. The mapping relation of the transport channels and the physical channels is shown in Fig. 10. The SCH is provided solely for physical layer use. One or more DCHs can be mapped onto a DPDCH/DPCCH pair. Two multiplexing schemes are proposed in Sections 5.2.2 and 5.2.3.

5.2.2. Coded Composite TrCh (CC-TrCh). In this scheme (see Fig. 11) several TrCh's are coded and interleaved individually and then multiplexed to form a coded composite TrCh (CC-TrCh). The physical layer allocates a DPDCH for the CC-TrCh and generates additionally an associated DPCCH for the connection.

5.2.3. Code-Multiplexed TrCh. In this scheme (see Fig. 12) each TrCh is coded and interleaved, and the result is sent to the physical layer which occupies a single DPDCH. So multiple DPDCHs and a single DPCCH will be generated at the physical layer for the connection.

Obviously, the coded composite TrCh is more efficient in channelization code usage but has poorer performance because it produces a high data rate, which results in a smaller spreading factor. As explained before, it is more suitable for downlink transmission. In contrast, the code multiplexed TrCh is less efficient in channelization code usage but has better performance, so it is more favorable in uplink transmission.

The processing of each block in Figs. 11 and 12 is as follows [2]:

1. *Channel Coding and Interleaving*. Depending on the specific requirements in terms of error rates, delay,

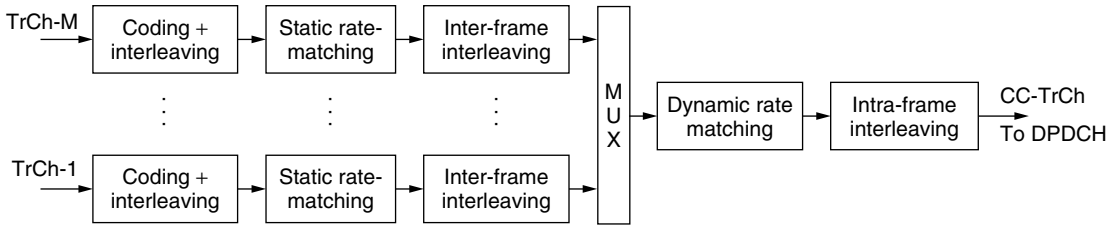


Figure 11. Coded composite TrCh in the IMT-2000.

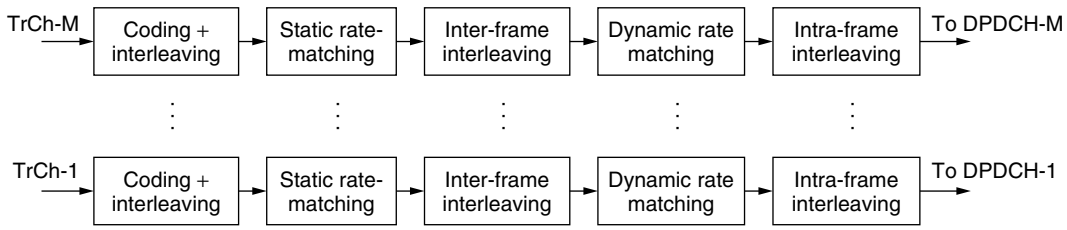


Figure 12. Code-multiplexed TrCh's in the IMT-2000.

and other factors, the coding can take one of the following forms:
Rate-$\frac{1}{3}$ convolutional coding for low-delay services with moderate error-rate requirements, such as voice.
A concatenation of rate-$\frac{1}{3}$ convolutional coding and outer Reed−Solomon coding + interleaving for high-quality services, such as data.
Turbo codes for high-rate high-quality services, such as data.

2. *Rate Matching*. Rate matching, which can be fixed or dynamic, is used to match the bit rate of the CC-TrCh to one of the available bit rates of the physical channels.
Static rate matching is carried out only when we add, remove, or redefine a TrCh. It is applied after channel coding, and uses code puncturing (decreasing rate) to adjust either the uplink or downlink rate to the physical channel rate in such a way that approximately the same SIR is achieved. But on the downlink, the rate should be adjusted to the closest lower physical channel rate to avoid the overallocation of the orthogonal codes.
Dynamic rate matching is carried out every 10 ms to match the TrCh rate to the physical channel rate by symbol repetition (increasing rate). It applies to the *uplink* only. For the downlink, discontinuous transmission within each slot is used when the instantaneous rate of the CC-TrCh does not exactly match the physical channel rate. When the data to be sent in a slot are sent off, the transmission stops until the next slot starts.

The parameters of all these processings are contained in the TFI field of each slot of the DPCCH. All the preceding processings are performed at the physical layer, but under the control of the radio-resource controller at the RRC layer (see Fig. 1).

5.3. Random Access

To facilitate burst data packet transmission, The UMTS/IMT-2000 has special arrangement for random-access capability.

5.3.1. Random-Access Procedure. The random-access procedure is based on slotted ALOHA and works as follows:

1. The UE acquires chip and frame synchronization with the target cell using the initial cell-search procedure described in Section 4.3.
2. The BCCH is read to retrieve information about the random-access scramble code and channelization code(s) used in the target cell.
3. The downlink path loss is estimated from the pilot bits of the primary CCPCH (see Section 4.4). The result is used to calculate the required transmit power of the random-access burst.
4. A random-access burst is transmitted on the RACH with a random time offset. The time offset is a multiple of 1.25 ms relative to the received frame boundary.
5. The base station responds with an acknowledgment on the FACH.
6. If the UE receives no acknowledgement, it selects a new time offset and tries again.

5.3.2. Random-Access Services. There are three different service classes:

1. *Packet Data Services*. The packet data can be transmitted in three different ways:
   a. If a small amount of data is to be sent, the data is simply appended to the access burst (see Fig. 13).
   b. If the data packet is large, the access burst is sent on the RACH while the data are transmitted on the DCH. The UE first sends a "resource request" (Res_Req) message on the RACH, which indicates the message type to be transmitted. If the network has the necessary resource, it transmits on the FACH a "resource allocation" (Res_All) message which contains a set of TFs. Exactly which TF the UE may use on the DCH and at what time the UE may initiate its transmission is indicated by a "capacity allocation" (Cap_All) message, which may be sent together with the Res_All message (if traffic is light), or in a separate Cap_All message at a later time (if traffic is heavy). The TF can be changed within the given TF set during the connection and this change is indicated by a TF_Change message, which contains the new TF to be used, on the DCH.
   c. A third method is used when the UE already has a dedicated channel at its disposal. In case when the UE has only a small amount of data to transmit, it can just start transmitting. If the UE has a large amount of data to transmit, it sends a Cap_Req message on the DCH before transmitting the data.
2. *Real-Time Services*. The real-time data transmission is very similar to the second way of packet data transmission, but with the following exceptions:
   a. The UE starts transmitting immediately after the Res_All message is received. In contrast, in the case of packet data, a further Cap_All message must be received before the UE can start transmission.
   b. Any TF in the TF set allocated in the Res_All message is allowed to be used by the UE. This makes the UE capable of supporting variable bit rate services.
   c. The TF set can be limited to a smaller subset by a "resource limit" (Res_Limit) message if the

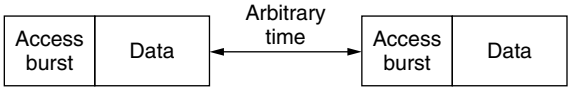


Figure 13. Packet random access on the RACH.

network resources are insufficient. The set can be made fully available again when later the network resources become sufficient.

3. *Mixed Services*. For the mixed services, such as data and voice, two TF sets are assigned to the UE. The UE uses the voice TF set the same way as in the single-service case, while the data TF is adapted to the voice TF usage in such a way that the aggregate output power and rate should never exceed the threshold set by the MAC.

6. SUMMARY

Wideband CDMA is a de facto air interface for third-generation mobile communications systems. The IMT-2000 WCDMA adopted by ITU has many salient features:

1. It employs a two-layer spreading; no synchronization is required between different transmitters.
2. It provides a wide range of services such as voice, video, data, image, and multimedia. The bit rate per channel can vary from a few kbps to 2 Mbps, either constant bit rate or variable bit rate. The information can be real-time or non-real-time. The transfer mode can be circuit-, packet-, STM-, or ATM-switched.
3. It provides high quality of services: toll quality for speech; BER less than 10^{-6} for data; low packet loss, delay, and delay jitter; soft handoff and macro diversity performance. Physical random-access channels are provided for fast access of short data packets from UE to the network.
4. It is flexible in system deployment and service provision, requiring simplified frequency planning; allowing macro-, micro-, picocell and employing advanced approaches of soft handoff.

In this article, because of space limitation we have focused attention mainly on signal transmission and reception strategies in the physical layer. Managing radio resources, handoff procedures, call admission control, and so on in the link layer is equally important for 3G deployment. IMT-2000 also has specifications in place to address these issues.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Grant no. RGPIN7779.

BIOGRAPHIES

Jon W. Mark received his Ph.D. degree in electrical engineering from McMaster University, Hamilton, Ontario, Canada, in 1970.

He was a professor of electrical and computer engineering at the University of Waterloo from 1970 to 2001. He is currently a distinguished professor emeritus and director of the Centre for Wireless Communications at the University of Waterloo. He had previously been at Westinghouse Canada Ltd.; IBM Thomas J. Watson Research Center, USA; Bell Laboratories, USA; Universite Pierre et Marie Curie, Paris, France; and the National University of Singapore.

He had previously worked in the areas of sonar signal processing, adaptive equalization, image and video coding, spread spectrum communications, computer communication networks, ATM switch design, and traffic management. His current research interests are in broadband wireless communications, and wireless/wireline interworking. His recently coauthored text entitled *Wireless Communications and Networking* is currently under production by Prentice-Hall.

Dr. Mark is an IEEE fellow, a former editor of *IEEE Transactions on Communications*, currently a member of the Inter-Society Steering Committee of the IEEE/ACM Transactions on Networking, an editor of *Wireless Networks*, and an associate editor of *Telecommunication Systems*.

Shihua Zhu received his B.Sc. degree in radio techniques in 1982 from Xi'an Jiaotong University, China, his M.Sc. degree in telecommunication systems in 1984, and his Ph.D. degree in electric systems engineering in 1987, both from the University of Essex, United Kingdom. He joined Xi'an Jiaotong University in 1987 and is currently a professor and the dean of the School of Electronic and Information Engineering. In the summers of 1993, 1994, and 1995, he visited the Chinese University of Hong Kong where he involved in a wireless digital transceiver development. He also worked as a visiting professor at University of Waterloo between October 1998 and October 1999 where he carried out a research on resource management for wideband CDMA systems. His research interests are multiuser detection, nonlinear multipath channel modeling, resource management in mobile communications systems, and power control in wideband CDMA systems.

BIBLIOGRAPHY

1. F. Adachi, M. Sawahashi, and H. Suda, Wideband DS-CDMA for next-generation mobile communications systems, *IEEE Commun. Mag.* 56–69 (Sept. 1998).
2. E. Dahlman et al., WCDMA — The radio interface for future mobile multimedia communications, *IEEE Trans. Vehic. Technol.* **47**(4): 1105–1118 (Nov. 1998).
3. E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, UMTS/IMT-2000 based on wideband CDMA, *IEEE Commun. Mag.* 70–80 (Sept. 1998).
4. E. H. Dinan and B. Jabbari, Spreading codes for direct sequence CDMA and wideband CDMA cellular networks, *IEEE Commun. Mag.* 48–54 (Sept. 1998).
5. R. Gold, Optimal binary sequences for spread spectrum multiplexing, *IEEE Trans. Inform. Theory* **IT-13**: 619–621 (Oct. 1967).
6. P. R. Larijani, J. W. Chinneck, and R. H. Hafez, Nonlinear power assignment in multimedia CDMA wireless networks, *IEEE Commun. Lett.* **2**: 251–253 (Sept. 1998).
7. J. W. Mark and S. Zhu, Power control and rate allocation in multirate wideband CDMA systems, *Proc. IEEE Wireless Communications and Networking Conf.*, 2000, pp. 168–172.

8. T. Ojanpera and R. Prasad, An overview of air interface multiple access for IMT-2000/UMTS, *IEEE Commun. Mag.* 82–95 (Sept. 1998).

9. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, Theory of spread-spectrum communications — a tutorial, *IEEE Trans. Commun.* **COM-30**: 855–884 (May 1982).

10. D. V. Sarwate and M. B. Pursley, Crosscorrelation properties of pseudorandom and related sequences, *Proc. IEEE* **68**: 598–619 (May 1980).

11. R. D. Yates, A framework for uplink power control in cellular radio systems, *IEEE J. Select. Areas Commun.* **13**: 1341–1347 (Sept. 1995).

WIRELESS AD HOC NETWORKS

Zygmunt J. Haas
Jing Deng
Ben Liang
Panagiotis Papadimitratos
S. Sajama
Cornell University
Ithaca, New York

1. INTRODUCTION

1.1. The Notion of the Ad Hoc Networks

A mobile ad hoc network (MANET) is a network architecture that can be rapidly deployed without relying on preexisting fixed network infrastructure. The nodes in a MANET can dynamically join and leave the network, frequently, often without warning, and possibly without disruption to other nodes' communication. Finally, the nodes in the network can be highly mobile, thus rapidly changing the node constellation and the presence or absence of links. Examples of the use of the MANETs are

- *Tactical operation* — for fast establishment of military communication during the deployment of forces in unknown and hostile terrain
- *Rescue missions* — for communication in areas without adequate wireless coverage
- *National security* — for communication in times of national crisis, where the existing communication infrastructure is nonoperational due to a natural disaster or a global war
- *Law enforcement* — for rapid establishment of communication infrastructure during law enforcement operations
- *Commercial use* — for setting up communication in exhibitions, conferences, or sales presentations
- *Education* — for operation of wall-free (virtual) classrooms
- *Sensor networks* — for communication between intelligent sensors [e.g., microelectromechanical systems (MEMS)] mounted on mobile platforms

Nodes in the MANET exhibit nomadic behavior by freely migrating within some area, dynamically creating and tearing down associations with other nodes. Groups of nodes that have a common goal can create formations (clusters) and migrate together, similar to military units on missions or to guided tours on excursions. Nodes can communicate with each other at any time and without restrictions, except for connectivity limitations and subject to security provisions. Examples of network nodes are pedestrians, soldiers, or unmanned robots. Examples of mobile platforms on which the network nodes might reside are passenger cars, trucks, buses, tanks, trains, planes, helicopters, or ships.

MANETs are intended to provide a data network that is immediately deployable in arbitrary communication environments and is responsive to changes in network topology. Because ad hoc networks are intended to be deployable anywhere, existing infrastructure may not be present. The mobile nodes are thus likely to be the sole elements of the network. Differing mobility patterns and radio propagation conditions that vary with time and position can result in intermittent and sporadic connectivity between adjacent nodes. The result is a time-varying network topology.

MANETs are distinguished from other ad hoc networks by rapidly changing network topologies, influenced by the network size and node mobility. Such networks typically have a large span and contain hundreds to thousands of nodes. The MANET nodes exist on top of diverse platforms that exhibit quite different mobility patterns. Within a MANET, there can be significant variations in nodal speed (from stationary nodes to high-speed aircraft), direction of movement, acceleration/deceleration, or restrictions on paths (e.g., a car must drive on a road, but a tank does not). A pedestrian is restricted by built objects, while airborne platforms can exist anywhere in some range of altitudes. In spite of such volatility, the MANET is expected to deliver diverse traffic types, ranging from pure voice to integrated voice and image, and even possibly some limited video.

1.2. The Communication Environment and the MANET Model

The following are a number of assumptions about the communication parameters, the network architecture, and the network traffic in a MANET:

- Nodes are equipped with portable communication devices. Lightweight batteries may power these devices. Limited battery life can impose restrictions on the transmission range, communication activity (both transmitting and receiving) and computational power of these devices.
- Connectivity between nodes is *not* a transitive relation; if node A can communicate directly with node B and node B can communicate directly with node C, then node A *may not*, necessarily, be able to

communicate directly with node C. This leads to the hidden-terminal problem [2].

- A hierarchy in the network routing and mobility management procedures could improve network performance measures, such as the latency in locating a mobile. However, a physical hierarchy may lead to areas of congestion and is very vulnerable to frequent topological reconfigurations.
- We assume that nodes are identified by fixed IDs (e.g., based on IP [3] addresses).
- All the network nodes have equal capabilities. This means that all nodes are equipped with identical communication devices and are capable of performing functions from a common set of networking services. However, all nodes do not necessarily perform the same functions at the same time. In particular, nodes may be assigned specific functions in the network, and these roles may change over time.
- Although the network should allow communication between any two nodes, it is envisioned that a large portion of the traffic will be between geographically close nodes. This assumption is clearly justified in a hierarchical organization. For example, it is much more likely that communication will take place between two soldiers in the same unit, rather than between two soldiers in two different brigades.

A MANET is a *peer-to-peer* network that allows *direct* communication between any two nodes, when adequate radio propagation conditions exist between these two nodes and subject to transmission power limitations of the nodes. If there is no direct link between the source and the destination nodes, *multihop* routing is used. In multihop routing, a packet is forwarded from one node to another, until it reaches the destination. Of course, appropriate routing protocols are necessary to discover routes between the source and the destination, or even to determine the presence or absence of a path to the destination node. Because of the lack of central elements, distributed protocols have to be used.

The main challenges in the design and operation of the MANETs, compared to more traditional wireless networks, stem from the lack of a centralized entity, the potential for rapid node movement, and the fact that *all* communication is carried over the wireless medium. In standard cellular wireless networks, there are a number of centralized entities [e.g., the base stations, the mobile switching centers (MSCs), the home location register (HLR), and the visitor location register (VLR)]. In ad hoc networks, there is no preexisting infrastructure, and these centralized entities do not exist. The centralized entities in the cellular networks perform the function of coordination. The lack of these entities in the MANETs requires distributed algorithms to perform these functions. In particular, the traditional algorithms for mobility management, which rely on a centralized HLR/VLR, and the medium access control schemes, which rely on the base-station/MSC support, become inappropriate.

All communications between all network entities in ad hoc networks are carried over the wireless medium. Because the radio communications are vulnerable to propagation impairments, connectivity between network nodes is not guaranteed. In fact, intermittent and sporadic connectivity may be quite common. Additionally, as the wireless bandwidth is limited, its use should be minimized. Finally, as some of the mobile devices are expected to be handheld with limited power sources, the required transmission power should be minimized as well. Therefore, the transmission radius of each mobile (device) is limited, and channels assigned to mobiles are typically spatially reused. Consequently, since the transmission radius is much smaller than the network span, communication between two nodes often needs to be relayed through intermediate nodes; thus, multihop routing is used.

Because of the possibly rapid movement of the nodes and variable propagation conditions, network information, such as a route table, becomes obsolete quickly. Frequent network reconfiguration may trigger frequent exchanges of control information to reflect the current state of the network. However, the short lifetime of this information means that a large portion of this information may never be used. Thus, the bandwidth used for distribution of the routing update information is wasted. In spite of these attributes, the design of the MANETs still needs to allow for a high degree of reliability, survivability, availability, and manageability of the network.

On the basis of the discussion above, we require the following features for the MANETs:

- *Robust routing and mobility management algorithms* to increase the network's reliability and availability (e.g., to reduce the chances that any network component is isolated from the rest of the network)
- *Adaptive algorithms and protocols* to adjust to frequently changing radio propagation, network, and traffic conditions
- *Low-overhead algorithms and protocols* to preserve the radio communication resource
- *Multiple (distinct) routes between a source and a destination* to reduce congestion in the vicinity of certain nodes, and to increase reliability and survivability
- *Robust network architecture* to avoid susceptibility to network failures, congestion around certain nodes, and the penalty due to inefficient routing

In this article, we present a survey of techniques used to establish communications in MANETs. In particular, we concentrate on four areas: the medium access control (MAC) schemes, the routing protocols, the multicasting protocols, and the security schemes.

2. MAC-LAYER PROTOCOLS FOR AD HOC NETWORKS

Applicability of the existing MAC-layer protocol, in particular the family of the *carrier sense multiple access* (CSMA), to the radio environment is limited by

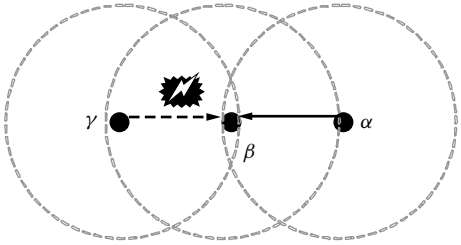


Figure 1. An example of the hidden-terminal problem.

the following two interference mechanisms: the hidden-terminal and the exposed-terminal problems.

The *hidden-terminal* problem occurs because the radio network, as opposed to other networks, such as a LAN, for instance, does not guarantee a high degree of connectivity. Thus, two nodes, which maintain connectivity to a third node, cannot, necessarily hear each other. Consider the situation in Fig. 1. Node α is in communication with node β. Node α is currently transmitting. Node γ wishes to communicate with node β as well. Following the CSMA protocol, node γ listens to the medium, but since there is an obstruction between node α and node γ, node γ does not detect node's α transmission, declaring the medium is free. Consequently, γ accesses the medium, causing collisions at β.

The second problem, the *exposed-terminal* problem, is depicted in Fig. 2. In the figure, node α is transmitting to node β, while node γ wants to transmit to node δ. Following the CSMA protocol, node γ listens to the medium, hears that node α transmits and defers from accessing the medium. However, there is no reason why node γ cannot transmit concurrently with the transmission of node α, as the transmission of node γ would not interfere with the reception at node β due to the distance between the two. The culprit here is, again, the fact that the collisions occur at the receiver, while the CSMA protocol checks the status of the medium at the transmitter.

In general, the hidden-terminal problem reduces the capacity of a network due to increasing the number of collisions, while the exposed-terminal problem reduces the network capacity due to the unnecessarily deferring nodes from transmitting.

Several attempts have been made in the literature to reduce the adverse effect of these two problems. The necessity of a dialog between the transmitting and the receiving nodes that preempts the actual transmission and that is referred to as the *RTS/CTS dialog*, has been

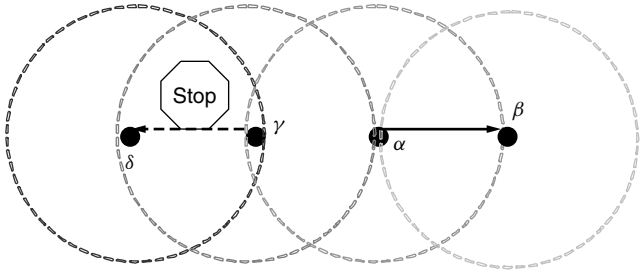


Figure 2. An example of the exposed-terminal problem.

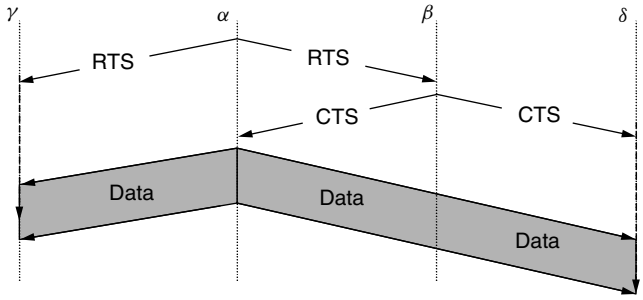


Figure 3. The RTS/CTS dialog reduces the chances of collision.

generally accepted. The RTS/CTS dialog is depicted in Fig. 3. A node ready to transmit a packet, sends a short control packet, the *request to send* (RTS), with all nodes that hear the RTS defer from accessing the channel for the duration of the RTS/CTS dialog. The destination, on reception of the RTS, responds with another short control packet, the *clear to send* (CTS). All nodes that hear the CTS packet defer from accessing the channel for the duration of the DATA packet transmission. The reception of the CTS packet at the transmitting node acknowledges that the RTS/CTS dialog has been successful and the node starts the transmission of the actual data packet. Although the RTS/CTS dialog does not eliminate the hidden- and the exposed-terminal problems, it does provide some degree of improvement over the traditional CSMA schemes.

In what follows, we present a number of attempts to further improve the performance of the MAC-layer protocols for ad hoc networks.

2.1. The Multiple Access Collision Avoidance (MACA) Scheme

In multiple access collision avoidance (MACA), Karn [4] proposed the use of RTS/CTS dialog for collision avoidance on the shared channel. Through the use of the RTS/CTS dialog, the MACA scheme reduces the probability of data packet collisions caused by hidden terminals.

The function of the RTS packet in the MACA scheme is similar function to that of the packet preamble in the receiver initiated busy-tone multiple access (RI-BTMA) scheme [4.5]. The RI-BTMA scheme does not have the CTS packet, because it uses a "busy" tone to notify the communication initiator. Since the CTS packet may suffer from packet collisions, the notification from CTS packets is not as safe as that from the busy tone in the RI-BTMA scheme. An example is the reception failure of CTS packet at some hidden nodes because of transmissions from other nodes. These hidden nodes, without receiving any CTS packet notification, may transmit new RTS packets when the CTS packet sender is receiving its data packet. This leads to data packet collisions. It is clear that additional continuous notification is necessary to protect data packets.

2.2. The Multiple Access Collision Avoidance Wireless (MACAW) Scheme

Bharghavan [5] suggested the use of the RTS-CTS-DS-DATA-ACK message exchange for a data packet

transmission in the MACAW protocol. Two new control packets were added to the packet train: DS and ACK packets. When the transmitter receives the CTS packet from its intended destination, it sends out a DS (data sending) packet before it transmits the data packet. The DS packet notifies neighbor nodes of the fact that a RTS/CTS dialog has been successful and a data packet will be sent. The ACK packet was implemented for immediate acknowledgment and the possibility of fast retransmission of collided data packets instead of upper-layer retransmission.

A new backoff algorithm, the multiple increase and linear decrease (MILD) algorithm, was also proposed in the paper to address the unfairness problem in accessing the shared channel. In the MILD backoff algorithm, successful nodes decrease their backoff interval by one step and unsuccessful nodes increase their backoff interval by multiplying them with 1.5. Backoff interval is also put into the header of the transmitted packet, so that the nodes overhearing successful packet transmission can copy the backoff interval on the packet into a local variable. Compared with the binary exponential backoff algorithm, the MILD algorithm has milder oscillation of the backoff intervals. Additional features of the MILD algorithm, such as multiple backoff intervals for different destinations, further improve the fairness performance of MACAW. The drawback of the MACAW scheme is inherited from the MACA scheme: the RTS/CTS packet collisions in a network with hidden terminals degrade its performance.

2.3. The Floor Acquisition Multiple Access (FAMA) Schemes

Fullmer and Garcia-Luna-Aceves [6] proposed the floor acquisition multiple access (FAMA) scheme. In FAMA, each ready node has to acquire the channel (the "floor") before it can use the channel to transmit its data packets. FAMA uses both carrier sensing and RTS/CTS dialog to ensure the acquisition of the "floor" and the successful transmission of the data packets. FAMA performs as well as MACA, when hidden terminals are present and as well as CSMA otherwise. In Ref. 7, FAMA was extended to FAMA-NPS (FAMA nonpersistent packet sensing) and FAMA-NCS (FAMA nonpersistent carrier sensing). FAMA-NPS requires nodes sensing packets to backoff. FAMA-NCS uses carrier sensing to keep neighbor nodes from transmitting while the channel is being used for data packet transmission. The length of the CTS packet is longer than that of the RTS packet, maintaining the dominance of CTS packets in the situation of collisions. Nodes can sense the carrier of the CTS packet when there is a collision between an RTS and a CTS packet and keep quiet; hence the data packet is protected at the receiver.

It was quantitatively shown [7] that FAMA-NPS did not perform well in situations with hidden terminals present, unless multiple transmissions of the CTS packet are used. The reason is the possible packet collisions resulting from hidden terminals. FAMA-NCS, by combining the carrier sensing and floor acquisition schemes together, outperforms nonpersistent CSMA and previous FAMA schemes in multihop networks.

2.4. The Dual-Busy-Tone Multiple Access (DBTMA) Scheme

In the DBTMA scheme [8], in addition to the use of an RTS packet, two out-of-band busy tones are used to notify neighbor nodes of the channel status. When a node is ready to transmit, it sets up its *transmit busy tone* and sends out an RTS packet to its intended receiver. On reception of the RTS packet, the receiver sets up a busy tone (the *receive busy tone*) and waits for the incoming data packet. The receive busy tone operates similarly to the busy tone of the RI-BTMA scheme. However, with the help of the second busy tone (the transmit busy tone), the probability of RTS packets colliding is decreased and the performance is improved.

The DBTMA scheme completely solved the hidden-terminal problems and the exposed-terminal problems. It forbids the hidden terminals to send any packet on the channel while the receiver is receiving the data packet. It allows the exposed terminals to initiate transmission by sending out the RTS packets. Furthermore, it allows the hidden terminals to reply to RTS packets by setting up the receive busy tone and initiate data packet reception.

3. ROUTING PROTOCOLS FOR AD HOC NETWORKS

Traditionally, the network routing protocols could be divided into proactive protocols and reactive protocols. *Proactive protocols* continuously learn the topology of the network by exchanging topological information among the network nodes. Thus, when there is a need for a route to a destination, such route information is available immediately. The early protocols that were proposed for routing in ad hoc networks were proactive distance vector protocols based on the *Distributed Bellman–Ford* (DBF) algorithm [9]. To address the problems of the DBF algorithm — convergence and excessive control traffic, which are especially an issue in resource-poor ad hoc networks — modifications were considered [10–12]. Yet another approach taken to address the convergence problem is the application of the link-state protocols to the ad hoc environment. An example of the latter is the optimized link-state routing protocol (OLSR) [13]. Yet another approach taken by some researchers is the proactive path-finding algorithms. In this approach, which combines the features of the distance vector and link-state approaches, every node in the network constructs a minimum spanning tree (MST), using the information of the MSTs of its neighbors, together with the cost of the link to its neighbors. The path-finding algorithms allow to reduce the amount of control traffic, to reduce the possibility of temporary routing loops, and to avoid the "counting-to-infinity" problem. An example of this type of routing protocols is the wireless routing protocol (WRP) [14,15].

The main issue with the application of proactive protocols to the ad hoc networking environment stems from the fact that as the topology continuously changes, the cost of updating the topological information may be prohibitively high. Moreover, if the network activity is low, the information about the actual topology may even not be used and the investment of limited transmission and computing resources in maintaining the topology is lost.

On the other end of the spectrum are the *reactive* routing protocols, which are based on some type of "query–reply" dialog. Reactive protocols do not attempt to continuously maintain the up-to-date topology of the network. Rather, when the need arises, a reactive protocol invokes a procedure to find a route to the destination; such a procedure involves some sort of flooding the network with the route query. As such, such protocols are often also referred to as on-demand. Examples of reactive protocols include the temporally ordered routing algorithm (TORA) [16], the dynamic source routing (DSR) [17], and ad hoc on-demand distance vector (AODV) [18]. In TORA, the route replies use controlled flooding to distribute the routing information through a form of a *directed acyclic graph* (DAG), which is rooted at the destination. The DSR and the AODV protocols, on the other hand, use unicast to route the reply back to the source of the routing query, along the reverse path of the query packet. The reversed path is "inscribed" into the query packet as "accumulated" route in the DSR and is used for source routing. In AODV, the path information is stored as the "next hop" information within the nodes on the path. Although the reactive approach can lead to less control traffic, as compared with proactive distance vector or link-state schemes, in particular, when the network activity is low and the topological changes frequent, the amount of traffic can still be significant at times. Moreover, due to the networkwide flooding, the delay associated with reactive route discovery may be considerable as well.

Thus, both of the routing "extremes," the proactive and the reactive schemes, may not perform best in a highly dynamic networking environment, such as in ad hoc networks. Although proactive protocols can produce the required route immediately, they may waste too much of the network resources in the attempt to always maintain the updated network topology. The reactive protocol, on the other hand, may reduce the amount of used network resources, but may encounter excessive delay in the flooding of the network with routing queries. Another approach to address the routing problem is through the *hybrid* protocols, which incorporate some aspects of the proactive and some aspects of the reactive protocols. The zone routing protocol (ZRP) [19] is an example of the hybrid approach. In ZRP, each node proactively maintains the topology of its close neighborhood only, thus reducing the amount of control traffic relative to the proactive approach. To discover routes outside its neighborhood, the node reactively invokes a generalized form of controlled flooding, which reduces the route discovery delay, as compared with purely reactive schemes. The size of the neighborhood is a single parameter that allows optimizing the behavior of the protocol based on the degree of nodal mobility and the degree of network activity.

In what follows, we present a number of examples of routing protocols that were developed for the ad hoc networking environment.

3.1. Single-Scope Routing Protocols

3.1.1. Advantages and Disadvantages. The main advantage of the single-scope routing protocols, in comparison with the multiscope routing protocols, is their lower complexity. There is no distinction of nearby or faraway nodes, and there is no need to maintain a hierarchical structure. Therefore, they are generally simpler to implement, both in simulations and in practical systems. The current activities within the IETF MANET group predominantly involve single-scope routing.

However, inefficient resource management can result from treating nodes equally, regardless of their relative location. For example, it may not be necessary for a node to maintain very accurate link-state tables or route caching information of faraway nodes. Therefore, the single-scope routing protocols may not scale well as the network size increases.

The single-scope routing protocols can be categorized into reactive (or on-demand) and proactive (or table-driven) ones. The main advantage of the reactive protocols is that no routing-table updating is required, unless a route is used. Therefore, battery power and wireless bandwidth can be conservatively utilized. However, when a route is needed, the source node needs to query for the route. That can lead to routing delay. Furthermore, an efficient route querying mechanism is required in order to prevent overloading the network with query packets.

On the other hand, the proactive protocols generally provide a source node with readily available routes to all other nodes. They incur no routing delay or query traffic. The disadvantage of the proactive protocols is that they may incur unnecessary control traffic in maintaining up-to-date topology information, whether that information is needed for routing or not.

3.1.2. Reactive/On-Demand Routing Protocols

3.1.2.1. Ad Hoc On-Demand Distance Vector Routing (AODV). AODV [19] incorporates the destination sequence number technique of destination-sequenced distance vector (DSDV) routing into an on-demand protocol. (DSDV is discussed in the sequel.)

Each node keeps a next-hop routing table containing the destinations to which it currently has a route. A route expires, if it is not used or reactivated for a threshold amount of time.

If a source has no route to a destination, it broadcasts a route request (RREQ) packet using an *expanding ring search* procedure, starting from a small time-to-live value (maximum hop count) for the RREQ, and increasing it if the destination is not found. The RREQ contains the last seen sequence number of the destination, as well as the source node's current sequence number. Any node that receives the RREQ updates its next-hop table entries with respect to the source node. A node that has a route to the destination with a higher-sequence number than the one specified in the RREQ unicasts a route reply (RREP) packet back to the source. Upon receiving the RREP packet, each intermediate node along the RREP routes updates its next-hop table entries with respect to the destination node, dropping the redundant RREP packets and those RREP packets with a lower destination sequence number than one seen previously.

When an intermediate node discovers a broken link in an active route, it broadcasts a route error (RERR)

packet to its neighbors, which in turn propagate the RERR packet upstream toward all the nodes that have an active route using the broken link. The affected source can then reinitiate route discovery, if the route is still needed.

3.1.2.2. Dynamic Source Routing (DSR). DSR [17] is a source routing on-demand protocol with various efficiency improvements. In DSR, each node keeps a *route cache* that contains full paths to known destinations. If a source has no route to a destination, it broadcasts a route request packet to its neighbors. Any node receiving the route request packet and without a route to the destination appends its own ID to the packet and rebroadcasts the packet. If a node receiving the route request packet has a route to the destination, the node replies to the source with a concatenation of the path from the source to itself and the path from itself to the destination. If the node already has a route to the source, the route reply packet will be sent over that route. Otherwise, depending on the underlining assumption of the directionality of links, the route reply packet can be sent over the reversed source-to-node path, or piggybacked in the node's route request packet for the source.

When an intermediate node discovers a broken link in an active route, it sends a route error packet to the source, which may reinitiate route discovery if an alternate route is not available.

DSR has efficiency improving features. One such feature is the *promiscuous* mode, in which a node listens to route request, reply, or error messages not intended to itself and updates its route cache correspondingly. Another DSR feature is the *expanding ring search* procedure, in which the route request packets are sent with a maximum hop count, which can be increased if the destination is not found within the hop-count limit. Finally, adding *jitter* in sending the route reply messages to prevent *route reply storms* and *packet salvaging* to extract correct routes from route error packets, are yet two other features that improve DSR performance.

3.1.2.3. Temporally Ordered Routing Algorithm (TORA). TORA [20] is a merger of the proactive link-reversal algorithm for destination-oriented directional acyclic graph creation [21] and the on-demand query–reply mechanism of lightweight mobile routing (LMR) [22].

In TORA, routes to a destination are defined by a directional acyclic graph (DAG) rooted at the destination. Each link in the network is assumed to be bi-directional, but in order to form the DAG with respect to a destination, a logical direction of the link is defined by giving *height* values to the two nodes at the ends of the link. Since time is part of the height value, TORA requires synchronized clocks across all nodes.

If a source has no route to a destination (i.e., the source node has no outgoing edge in the DAG), it broadcasts a route query packet (QRY), which is propagated outward by its neighbors. After receiving the QRY, a node that has a route to the destination broadcasts a route update packet (UPD) containing its own height. Receiving the UPD, each node that doesn't have a route to the destination updates its height to reflect the creation of an outgoing edge.

Route maintenance is achieved through height adjustment and UPD exchange. Network partition can be detected by a node receiving UPDs reflected from the partition boundary, in which case a clear message (CLR) is used to update all routes within the partition.

TORA also supports a proactive mode, in which the destination initiates the route creation process by sending a packet that is processed and forwarded by the neighboring nodes.

3.1.3. Proactive/Table-Driven

3.1.3.1. Destination-Sequenced Distance-Vector Routing (DSDV). DSDV [12] provides improvements over the conventional Bellman–Ford distance vector protocol. It eliminates route looping, increases convergence speed, and reduces control message overhead.

In DSDV, each node maintains a next-hop table, which it exchanges with its neighbors. There are two types of next-hop table exchanges: periodic full-table broadcast and event-driven incremental updating. The relative frequency of the full-table broadcast and the incremental updating is determined by the node mobility.

In each data packet sent during a next-hop table broadcast or incremental updating, the source node appends a sequence number. This sequence number is propagated by all nodes receiving the corresponding distance vector updates, and is stored in the next-hop table entry of these nodes. A node, after receiving a new next-hop table from its neighbor, updates its route to a destination only if the new sequence number is larger than the recorded one, or if the new sequence number is the same as the recorded one, but the new route is shorter.

In order to further reduce the control message overhead, a *settling time* is estimated for each route. A node updates its neighbors with a new route only if the settling time of the route has expired and the route remains optimal.

3.1.3.2. Wireless Routing Protocol (WRP). This protocol [15] provides improvements over the Bellman–Ford distance vector protocol. It reduces the amount of route looping, and has a mechanism to ensure the reliable exchange of update messages.

In WRP, each node maintains a distance table matrix, which contains all destination nodes, and, for each destination node, all neighbors through which the destination node can be reached. For each neighbor–destination pair, if a route exists, the route length is recorded. Also recorded is the *predecessor*, the last node along a route before the destination node.

Each node's neighbor broadcasts its current best route to selected destinations on an event-driven incremental basis. After a broadcast, acknowledgments are expected from all neighbor nodes. If some acknowledgments are missing, the broadcast will be repeated, with a *message retransmission list* specifying the subset of neighbors that need to respond. A node, after receiving the route updating packets from a neighbor, updates its own routing table only if the consistency of the new information is checked against the predecessor information from all its neighbors.

3.2. Multiscope Routing Protocols

3.2.1. Advantages and Disadvantages. The multiscope routing protocols distinguish nodes by their relative positions. More resource is devoted to maintaining the topology information of more nearby, and hence more frequently used, parts of the network. Therefore, scalability is the main advantage of the multiscope routing protocols.

Their disadvantage is their relative complexity in comparison with the single-scope routing protocols. Ranking mechanisms that distinguish the nodes are required. Furthermore, they generally need to be reconfigurable, in order to adapt to the changing network topology and the varying node traffic and movement patterns.

Multiscope routing can be categorized into the flat protocols and the hierarchical protocols. The main advantage of the flat protocols, in comparison with the hierarchical ones, is that they do not require specialized nodes. All nodes serve the same set of functions. Therefore, they are relatively simple to implement, and they avoid the control message overhead and nonuniform loading involved in node specialization. However, since the flat structure does not have special nodes that can provide locally centralized functionality, the nodes between nearby local scope exchange link information in a strictly distributive manner. Thus, the lack of coordination can lead to inefficiency.

On the other hand, the hierarchical protocols utilize specialized nodes, such as the cluster heads, group leaders, or the route gateways, to coordinate the dissemination of local link information. Furthermore, the relative position of the specialized nodes can provide directional guidance to routing between the regular nodes. However, the dynamic maintenance of the hierarchy can potentially consume a large amount of the battery power and wireless bandwidth from routing itself, especially when the network is highly mobile. Furthermore, mechanisms are needed to avoid overloading the local controllers and to alleviate the traffic hot spots.

3.2.2. Flat Routing Protocols

3.2.2.1. Zone Routing Protocol (ZRP). This protocol [18] provides a *hybrid* routing framework that is locally proactive and globally reactive. Each node proactively advertises its link state within a fixed number of hops, called the *zone radius*. These local advertisements give each node an updated view of its routing zone — the collection of all nodes and links that are reachable within the zone radius. The routing zone nodes that are at the minimum distance of the zone radius are called *peripheral nodes*. The peripheral nodes represent the boundary of the routing zone and play an important role in zone-based route discovery. Each node has an associated routing zone, and routing zones of neighboring nodes overlap.

ZRP uses the knowledge of the routing zone connectivity to guide its global route discovery. Rather than blindly broadcasting route queries from a node to all its neighbors, ZRP employs a service called *bordercasting*, which directs the route request from a node to its peripheral nodes via multicast. Special query control mechanisms are used to identify those peripheral nodes that have been covered by the route query (i.e., that belong to the routing zone of a node that already has bordercast the query) and prune them from the bordercast's query distribution tree. This encourages the query to propagate outward, away from its source and away from covered regions of the network.

Routing zones also help improve the quality and survivability of discovered routes, by making them more robust to changes in network topology. Once routes have been discovered, routing zones offer enhanced, real-time, route maintenance. Multiple hop paths within the routing zone can bypass link failures. Similarly, sub optimal route segments can be identified and traffic can be rerouted along shorter paths.

3.2.2.2. Optimized Link State Routing (OLSR). OLSR [23] is a link-state protocol, where the link information is disseminated through an efficient flooding technique.

The key concept in OLSR is *multipoint relay* (MPR). A node's MPR set is a subset of its neighbors, whose combined radio range covers all nodes two hops away. Heuristics are proposed for each node to determine its minimum MPR set based on its two-hop topology. Each node obtains the two-hop topology through its neighbors' periodic broadcasting of "Hello" packets containing the neighbors' lists of neighbors.

As with a conventional link-state protocol, a node's link information update is propagated throughout the network. However, in OlSR, when a node forwards a link updating packet, only those neighbors in the node's MPR set participate in forwarding the packet (similar to ZRP's border-casting with *one-hop* zone radius).

Furthermore, a node only originates link updates concerning those links between itself and the nodes in its MPR set. Therefore, routes are computed using a node's partial view of the network topology.

3.2.2.3. Fisheye State Routing (FSR). The fisheye routing concept is based on the premise that changes in a network region's topology have less effect on a router's packet forwarding decisions as the distance (in hops) between the router and the network increases. This relationship can be exploited in order to reduce routing traffic by relaying topology updates for distant regions less often than updates for nearby regions. Given an approximate view of the distant parts of the network, a node can forward a packet in the proper direction toward the destination. As the packet progresses toward the destination, the view of the destination's region becomes more accurate, providing for more precise packet forwarding.

This fisheye technique is applied in the fisheye state routing (FSR) protocol [24], an adaptation of the global state routing (GSR) [25]. In the original GSR protocol, link-state information is propagated through the network by periodic link state table exchanges between neighbors. In FSR, a node exchanges individual link state table entries at different rates, depending on the distance to the link's source. In particular, FSR defines *scopes* of increasing radii (in hops) around each node. A node relays a link state table entry if the link's source lies within

the largest scope covered by the current table exchange. The first level (innermost) scope is covered by every table exchange. The kth-level scope is covered by every X_kth interval, where X_k is an integer multiple of X_{k-1}. This relationship ensures that an exchange covering a level kth scope coincides with the more frequent updates of all the interior scopes.

3.2.3. Hierarchical Routing Protocols

3.2.3.1. Core-Extraction Distributed Ad Hoc Routing (CEDAR). CEDAR [26] employs a set of core nodes, at least one of which is within one hop of each node, in its routing mechanism. The core nodes are selected using a highest-degree scheme. A core node dominates each noncore node. Through periodic updating, the noncore nodes maintain a list of the IDs of their neighbors and their neighbors' respective dominators. The state information of each link is disseminated toward the core nodes away from the link, and the higher is the capacity of link, the further the information travels. Each core node keeps a local link-state table containing only the stable, high capacity, and nearby links.

Global route search is carried out reactively. Similar to CBRP, the dominator of a source node determines a *core path* to the dominator of the destination node by an efficient flooding over the core. Using its local link state, the dominator of the source computes a "shortest–widest–furthest" QoS-admissible path to an intermediate node, along the core path toward the destination. It then sends a route forwarding request to the dominator of the intermediate node, which then starts the same QoS-admissible path search using its own local link-state table. The process continues until the QoS-admissible path reaches the destination. Source routing is then carried out to forward the data packets.

3.2.3.2. Zone-Based Hierarchical Link State (ZHLS). In ZHLS [27], the system coverage area is divided into nonoverlapping physical zones. The nodes are equipped with geographic location devices such as the GPS receivers, so that each node can determine its zone membership by comparing its physical location with the zone map. Furthermore, if the nodes within a zone are partitioned, logical subzones are created, each containing one of the partitions. Every node maintains an intrazone routing table and an interzone routing table. The intrazone routing table enables a node to reach all the other nodes within the zone. Interzone communications are carried out through the *gateway* nodes near the zone edges. The gateway nodes broadcast the status of the virtual links between zones to the entire network. A node aggregates all gateway broadcasts to form the interzone routing table.

When a source node needs to transmit data to a destination node outside the source node's zone, global zone query is used to determine the zone identity of the destination node. Using its interzone routing table, the source node sends the query to all zones in the network. After receiving the query message, the gateway nodes, whose zone contains the destination node sends back to the source node the destination node's zone identity. The source node then sends out the data packets with the destination node's zone ID and node ID specified in their headers. The packets are then forwarded to the destination node according to both the interzone and the intrazone routing tables at the intermediate nodes.

3.2.3.3. Landmark Ad Hoc Routing (LANMAR). The original landmark scheme for wired networks was proposed by Tsuchiya [28]. LANMAR [29] adopts that scheme for ad hoc network routing. In LANMAR, the network consists of predefined logical subnets, each with a preselected *landmark*. All nodes in a subnet are assumed to move as a group, and they remain connected to each other via *fisheye state routing*.

The routes to the landmarks, and hence the corresponding subnets, are proactively maintained by all nodes in the network through the exchange of distance vectors. Every node has a lifetime hierarchical address, identifying the subnet where it belongs. A source node specifies the hierarchical address of a destination node in the data packet headers. The packets are then forwarded toward the landmark of the subnet, where the destination node belongs. When a packet reaches a node in the subnet, where the destination node belongs, the node forwards the packet to the destination node using its subnet routing table.

3.3. Geographically Routed Protocols

3.3.1. Location-Aided Routing (LAR). In LAR [30], a source node estimates the range of a destination's location, based on the destination's last reported velocity, and broadcasts route request only to nodes within a geographically defined *request zone*. LAR requires each node to obtain its geographic location through external devices such as GPS.

3.3.2. Distance Routing Effect Algorithm for Mobility (DREAM). In DREAM [31], each node obtains its geographic location through external devices such as GPS, and periodically transmits its location coordinates to other nodes in the network. The period of location transmission depends on the node's velocity and the geographic distance to nodes to which the location information is intended.

A source sends a data packet to a subset of its neighbors in the direction of the destination. The intermediate nodes similarly forward the data packet towards the destination.

4. MULTICASTING PROTOCOLS FOR AD HOC NETWORKS

Multicasting is an efficient communication tool for use in multipoint applications. Many of the proposed multicast routing protocols, both for the Internet and for ad hoc networks, construct trees over which information is transmitted. Using trees is evidently more efficient than brute-force approach of sending the same information from the source individually to each receiver. Another benefit of using trees is that routing decisions at the intermediate nodes become very simple; a router in a multicast tree that receives a multicast packet over an in-tree interface forward the packet over the rest of its in-tree interfaces.

Multicast routing algorithms in the Internet [32] can be classified into three broad categories:

- *Shortest-path tree* algorithms [9]
- *Minimum-cost tree* algorithms [33,34]
- *Constrained tree* algorithms [35,36]

There are two fundamental approaches in designing multicast routing — one is to minimize the distance (or cost) from the sender to each receiver individually (shortest path tree algorithms), and the other is to minimize the overall (total) cost of the multicast tree. Practical considerations lead to a third category of algorithms, which try to optimize both constraints using some metric (minimum cost trees with constrained delays). The majority of multicast routing protocols in the Internet is based on shortest-path trees, because of their ease of implementation. Also, they provide minimum delay from sender to receiver, which is desirable for most real-life multicast applications. However shared trees are used in some more recent protocols (e.g., PIM [37] and CBT [38]), in order to minimize the state stored in the routers.

Multicasting in ad hoc networks is more challenging than in the Internet, because of the need to optimize the use of several resources simultaneously: (1) nodes in ad hoc networks are battery-power-limited and data travel over the air where wireless resources are scarce; (2) there is no centralized access point or existing infrastructure (as in the cellular network) to keep track of the node mobility; and (3) the status of communication links between nodes is a function of their positions, transmission power levels, and so on. The mobility of routers and randomness of other connectivity factors lead to a network with a potentially unpredictable and rapidly changing topology. This means that by the time a reasonable amount of information about the topology of the network is collected and a tree is computed, there may be very little time before the computed tree becomes useless.

Work on multicast routing in ad hoc networks gained momentum in the mid-1990s. Some early approaches to provide multicast support in ad hoc networks consisted of adapting the existing Internet multicasting protocols; for example, Shared Tree Wireless network Multicast [39]. On the other hand, ODMRP [40], AMRIS [41], CAMP [42], and others [43–51] have been designed specifically for ad hoc networks. ODMRP is a mesh-based, on-demand protocol that uses a soft state approach for maintenance of the message transmission structure. It exploits the robustness of mesh structure to frequent route failure and gains stability at the expense of bandwidth. The core-assisted mesh protocol (CAMP) attempts to remedy this excessive overhead, while still using a mesh by using a core for route discovery. AMRIS constructs a shared delivery tree rooted at a node, with ID numbers increasing as they radiate from the source. Local route recovery is made possible due to this property of ID numbers, hence reducing the route recovery time and also confining route recovery traffic to the region of link failures.

In what follows, we present a number of examples of multicasting protocols that were developed for the ad hoc networking environment.

4.1. Core Assisted Mesh Protocol (CAMP)

CAMP [42] builds and maintains a multicast mesh for information distribution within each multicast group. A router is allowed to accept unique packets coming from any neighbor in the process of forwarding packets through the mesh. Because a member router of a mesh has redundant paths to any other router in the mesh, this protocol is more resilient to topology changes than tree based protocols.

Cores are used to limit the control traffic needed for receivers to join multicast groups. In contrast to CBT [38], one or multiple cores can be defined for each mesh and cores need not be part of the mesh of their group. Routers can join a group even if all associated cores become unreachable using an expanded ring search. CAMP ensures that all reverse shortest paths between sources and receivers are part of groups mesh by means of "heartbeat" messages. In the event of link failure and partition, the operation of mesh components continues. Different components merge by sending join requests to cores as soon as connectivity with a core is reestablished.

CAMP is designed to support very dynamic ad hoc networks. According to the performance analysis presented in [42], this article, CAMP performs better than the on-demand multicast routing protocol (ODMRP) in terms of percentage of packets lost by routers, average packet delay, and total number of control packets received by each router. (ODMRP is described below.)

4.2. Multicast Operation of Ad Hoc On-Demand Distance Vector Routing Protocol

This is an extension of AODV to support multicasting and it builds multicast trees on demand to connect group members. Route discovery in MAODV [43] follows a route request/route reply discovery cycle. As nodes join the group, a multicast tree composed of group members is created. Multicast group membership is dynamic and group members are routers in the multicast tree. Link breakage is repaired by downstream node broadcasting a route request message. The control of a multicast tree is distributed, so there is no single point of failure. One big advantage claimed is that since AODV offers both unicast and multicast communication; route information, when searching for a multicast route can also increase unicast routing knowledge and vice versa.

In [43], an ad hoc network consists of laptops in a room (50–100 m wide, 10-m range) talking to each other, moving at a rate of 1 m/s. The results presented only verify working of AODV and do not compare performance with other multicasting protocols. [43] shows that AODV attains a high output ratio and is able to offer this communication with a minimum of control packet overhead. It also demonstrates its operation under frequent network partitions.

4.3. AMRIS: A Multicast Protocol for Ad Hoc Wireless Networks

AMRIS [41] is an on-demand protocol that constructs a shared delivery tree to support multiple senders and receivers within a multicast session. Each participant in the multicast session has a session-specific multicast

session member id (*msm-id*). These msm-ids increase in numerical value as they radiate away from a central node known as SID. Tree initialization is done by the SID broadcasting a new session message. All nodes of the network calculate their msm-id to be larger than the msm-id of the node they received the new session message from. There are beacon messages exchanged between nodes, which help a node to calculate its new msm-id after it moves to a new location.

AMRIS does not depend on the unicast routing protocol to provide routing information to other nodes, since it maintains a neighbor-status table. It is the child's responsibility to reconnect to the tree if a link failure occurs. If it has potential parents, that is, neighboring nodes with lower msm-ids, it sends a join request to them, which in turn try to join the tree in the same way. If there are no potential parents, the node transmits a join request message.

In the simulation given by Wu and Tay [41], the authors vary membership from 50 to 100 and the speed of nodes is up to 20 m/s. The simulation results presented study various performance parameters in terms of network conditions. The paper studied the effect of beacon intervals, membership sizes and mobility on packet delivery ratio and concludes that there is an optimum beacon interval. Control overhead is verified to be higher, when the beacon interval is small. They also show that the relationship between end-to-end packet delay and packet delivery ratio is robust with respect to membership.

4.4. AMRoute: Ad Hoc Multicast Routing Protocol

AMRoute [44] presents an approach for robust IP multicast in ad hoc networks by exploiting user-multicast trees and dynamic logical cores. It creates a bidirectional shared tree for data distribution using only group senders and receivers as tree nodes. Unicast tunnels are used as tree links to connect neighbors in the user-multicast tree. Hence the AMRoute protocol does not need to be supported by other nodes in the network. Also the tree structure does not change even in the case of dynamic network topology and hence reduces signaling. Each node is aware of its tree neighbors only and forwards data on the tree links to its neighbors. This saves node resources.

Certain tree nodes are designated by AMRoute as logical cores and are responsible for initiating and managing the signaling component of AMRoute, such as detection of group members and tree setup. Unlike CBT and PIM-SM, they are not a central point for data distribution and can migrate dynamically among member nodes. Hence there is no single point of failure. Like DVMRP, AMRoute provides robustness by periodic flooding for tree construction. However, AMRoute periodically floods a small signaling message instead of data.

AMRoute simulations were done with TORA as the underlying unicast protocol. The mobility of the network was emulated by keeping the node location fixed and breaking/connecting links between neighbors. The simulation results show that broadcasting signaling traffic generated by AMRoute is independent of the group size and inversely proportional to the network mobility. Unicast signaling traffic is proportional to the group size and inversely proportional to the network mobility. Total signaling traffic is independent of the data rate. Both signaling traffic and join latency are relatively low for typical group sizes. They verify that group members receive a high proportion of data multicast by a sender, even in the case of a dynamic network.

4.5. ODMRP: On-Demand Multicast Routing Protocol

ODMRP [40] is a mesh based, rather than a conventional tree based, scheme and uses a forwarding group concept (only a subset of nodes forwards the multicast packets via scoped flooding). By maintaining a mesh instead of a tree, the drawbacks of multicast trees in ad hoc networks, like frequent tree reconfiguration and non–shortest path in a shared tree, are avoided. ODMRP applies on-demand routing techniques to avoid channel overhead and to improve scalability.

The source starts a session by flooding a "join data" control packet with data payload attached, which is subsequently broadcast at regular intervals to the entire network to refresh membership information and update the routes. The mesh is created by the replies of receivers to this packet received via various paths. When receiving a multicast data packet, a node forwards it only when it is not a duplicate, hence minimizing traffic overhead.

Because the nodes maintain soft state, finding the optimal flooding interval is critical to ODMRP performance. ODMRP uses location and movement information to predict the duration of time that routes will remain valid. With the predicted time of route disconnection, a "join data" packet is flooded, when route breaks of ongoing data sessions are imminent. Lee et al. [40] compare DVMRP with ODMRP and show that the latter is better suited for ad hoc networks in terms of bandwidth utilization.

4.6. MCEDAR: Multicasting Core-Extraction Distributed Ad Hoc Routing

This scheme [50] is a multicast extension of CEDAR, which was a routing scheme proposed for unicast communication in ad hoc networks. MCEDAR relies on the *core extraction* and the *core broadcast* components of the CEDAR architecture. The core-extraction algorithm used is a distributed heuristic for finding a good approximation to a minimum dominating set. Each core node has the following state stored in it: its nearby core nodes and the nodes it dominates (i.e., each core node has enough local information to reach the domain of its nearby nodes and set up virtual links). Core broadcast is used instead of flooding in order to discover a route to the destination. This is done by making each node cache every RTS and CTS packet that it hears on the channel for core broadcast packets. So a node knows whether a packet has been received by the destination already. If it has, it does not transmit the packet and hence suppresses the duplicate transmission.

The infrastructure for a multicast group resides entirely within the core broadcast mechanism, which is used to perform data forwarding. Each multicast group extracts a subgraph of the core graph to function as "mgraph." Data

forwarding is done on the mgraph using the core broadcast mechanism. In this way the forwarding is tree-based, although the structure is robust, because it is a mesh.

5. SECURITY OF AD HOC NETWORKS

The provision of security services in the MANET context faces a set of challenges specific to this new technology. The insecurity of the wireless links, energy constraints, relatively poor physical protection of nodes in a hostile environment, and the vulnerability of statically configured security schemes have been identified in the literature [52,53] as such challenges. Nevertheless, the single most important feature that differentiates MANET is the absence of a fixed infrastructure. No part of the network is dedicated to support individually any specific network functionality; routing (topology discovery, data forwarding) is the most prominent example. Additional examples of functions that cannot rely on a central service, and that are also of high relevance to this work, are naming services, certification authorities (CAs), directory, and other administrative services.

Even if such services were assumed, their availability would not be guaranteed, due either to the dynamically changing topology that could easily result in a partitioned network, or to congested links close to the node acting as a server. Furthermore, performance issues, such as delay constraints on acquiring responses from the assumed infrastructure, would pose an additional challenge.

The absence of infrastructure and the consequent absence of authorization facilities impede the usual practice of establishing a line of defense, separating nodes into trusted and nontrusted. Such a distinction would have been based on a security policy, the possession of the necessary credentials and the ability for nodes to validate them. In the MANET context, there may be no ground for an a priori classification, since all nodes are required to cooperate in supporting the network operation, while no prior security association can be assumed for all the network nodes. Additionally, in MANET, freely roaming nodes form transient associations with their neighbors; join and leave MANET sub-domains independently and without notice. Thus it may be difficult in most cases to have a clear picture of the ad hoc network membership. Consequently, especially in the case of a large-size network, no form of established trust relationships among the majority of nodes could be assumed.

In such an environment, there is no guarantee that a path between two nodes would be free of malicious nodes, which would not comply with the employed protocol and attempt to harm the network operation. The mechanisms currently incorporated in MANET routing protocols cannot cope with disruptions due to malicious behavior. For example, any node could claim that it is one hop away from the sought destination, causing all routes to the destination to pass through itself. Alternatively, a malicious node could corrupt any in-transit route request (reply) packet and cause data to be misrouted.

The presence of even a small number of adversarial nodes could result in repeatedly compromised routes, and, as a result, the network nodes would have to rely on cycles of timeout and new route discoveries to communicate. This would incur arbitrary delays before the establishment of a noncorrupted path, while successive broadcasts of route requests would impose excessive transmission overhead. In particular, intentionally falsified routing messages would result in a denial-of-service (DoS) experienced by the end nodes.

Despite the fact that security of MANET routing protocols is envisioned to be a major roadblock in commercial application of this technology, only a limited number of works has been published in this area. Below, we review some schemes related to the problem of incorporating security provisions within the context of ad hoc communication.

5.1. Overview of Security Schemes for Ad Hoc Networks

Efforts to incorporate security measures in the ad hoc networking environment have concentrated mostly on the aspect of data forwarding, disregarding the aspect of topology discovery. On the other hand, solutions that target route discovery have been based on approaches for fixed-infrastructure networks, defying the particular MANET challenges.

For the problem of secure data forwarding, two mechanisms that (1) detect *misbehaving* nodes and report such events and (2) maintain a set of metrics reflecting the past behavior of other nodes [54] have been proposed to alleviate the detrimental effects of packet dropping. Each node may choose the "best" route, composed of relatively well behaved nodes, namely, nodes that do not have history of avoiding forwarding packets along established routes. Among the assumptions of the abovementioned work [54] are a shared medium, bidirectional links, use of source routing (i.e., packets carry the entire route that becomes known to all intermediate nodes), and no colluding malicious nodes. Nodes operating in promiscuous mode overhear the transmissions of their successors and may verify whether the packet was forwarded to the downstream node and check the integrity of the forwarded packet. On detection of a misbehaving node, a report is generated and nodes update the rating of the reported misbehaving node. The ratings of nodes along a well-behaved route are periodically incremented, while reception of a misbehavior alert dramatically decreases the node rating.[1] When a new route is required, the source node calculates a path metric equal to the average of the ratings of the nodes in each of the route replies and selects the route with the highest metric.

The detection mechanism exploits two features that frequently appear in MANET: the use of a shared channel and source routing. Nevertheless, the plausibility of this solution could be questioned for several reasons, and, indeed, the authors provide a short list of scenarios of incorrect detection. The possibility of falsely detecting misbehaving nodes could easily create a situation with

[1] The initial rating, 0.5, is increased by 0.01 every 200 ms. Suspected nodes have a rating equal to -100, with the option for a long timeout period after which the negative rating is changed back to a positive value.

many nodes falsely suspected for a long period of time. In addition, the metric construction may lead to a route choice that includes a suspected node, if, for example, the number of hops is relatively high, so that a low rating is "averaged out." Finally, the most important vulnerability is the proposed feedback itself; there is no way for the source, or any other node that receives a misbehavior report to validate its authenticity or correctness. Consequently, the simplest attack would be to generate fake alerts and eventually disable the network operation altogether. The protocol attempts new route discoveries, when none of the route replies is free of suspected nodes, with the excessive route request traffic degrading the network performance. At the same time, the adversary can falsely accuse a significant fraction of nodes within the timeout period related to reinstating from a negative rating and, essentially, partition the network.

A different approach [55] is to provide incentive to nodes, so that they comply with protocol rules to properly relay user data. The concept of fictitious currency is introduced, in order to *endogenize* the behavior of the assumed greedy nodes, which would forward packets in exchange for currency. Each intermediate node purchases from its predecessor the received data packet and sells it to its successor along the path to the destination. Eventually the destination pays for the received packet.[2] This scheme assumes the existence of an overlaid geographic routing infrastructure and a public key infrastructure (PKI). All nodes are preloaded with an amount of currency, have unique identifiers, are associated with a pair of private/public keys, and all cryptographic operations related to the currency transfers are performed by a physically tamper-resistant module. The applicability of the scheme, which targets wide-area MANET, is limited by the assumption of an online certification authority in the MANET context. Moreover, nodes could flood the network with packets destined to nonexistent nodes and possibly lead nodes unable to forward purchased packets to starvation. The practicality of the scheme is also limited by its assumptions, the high computational overhead (hop-by-hop public key cryptography, for each transmitted packet), and the implementation of physically tamper-resistant modules.

The protection of the route discovery process has been regarded as an additional Quality-of-Service (QoS) issue [56], by choosing routes that satisfy certain quantifiable security criteria. In particular, nodes in a MANET subnet are classified into different trust and privilege levels. A node initiating a route discovery sets the sought security level for the route; the required minimal trust level for nodes participating in the query/reply propagation. Nodes at each trust level share symmetric encryption and decryption keys. Intermediate nodes of different levels cannot decrypt in-transit routing packets, or determine whether the required QoS parameter can be satisfied, and simply drop them. Although this scheme provides protection (e.g., integrity) of the routing protocol traffic, it does not eliminate false routing information provided by malicious nodes. Moreover, the proposed use of symmetric cryptography allows any node to corrupt the routing protocol operation within a level of trust, by mounting virtually any attack that would be possible without the presence of the scheme. Finally, the assumed supervising organization and the fixed assignment of trust levels does not pertain to the MANET paradigm. In essence, the proposed solution transcribes the problem of secure routing in a context, where nodes of a certain group are assumed to be trustworthy, without actually addressing the global secure routing problem.

An extension of the ad hoc on-demand distance vector (AODV) [18] routing protocol has been proposed [57] to protect the routing protocol messages. The secure AODV (S-AODV) scheme assumes that each node has certified public keys of all network nodes, so that intermediate nodes can validate all in-transit routing packets. The basic idea is that the originator of a control message appends an RSA signature [58] and the last element of a *hash chain* [59] (i.e., the result of n consecutive hash calculations on a random number). As the message traverses the network, intermediate nodes cryptographically validate the signature and the hash value, generate the kth element of the hash chain, where k is the number of traversed hops, and place it in the packet. The route replies are provided either by the destination or intermediate nodes having an active route to the sought destination, with the latter mode of operation enabled by a different type of control packets.

The use of public key cryptography imposes a high processing overhead on the intermediate nodes and can be considered unrealistic for a wide range of network instances. Furthermore, it is possible for intermediate nodes to corrupt the route discovery by pretending that the destination is their immediate neighbor, advertising arbitrarily high sequence numbers and altering (either decreasing by one or arbitrarily increasing) the actual route length. Additional vulnerabilities stem from the fact that the *IP* portion of the *S-AODV* traffic can be trivially compromised, since it is not (and cannot be, due to the AODV operation) protected, unless additional hop-by-hop cryptography and accumulation of signatures is used. Finally, the assumption that certificates are bound with *IP* addresses is unrealistic; roaming nodes joining MANET sub-domains will be assigned *IP* addresses dynamically (e.g., DHCP [60]) or even randomly (e.g., zero configuration [61]).

A different approach is taken by the secure message transmission (SMT) [62] protocol, which, given a topology view of the network, determines a set of diverse paths connecting the source and the destination nodes. Then, it introduces limited transmission redundancy across the paths, by dispersing a message into N pieces, so that successful reception of any M out of N pieces allows the reconstruction of the original message at the destination. Each piece is equipped with a cryptographic header that provides integrity and replay protection, along with origin authentication, and is transmitted over one of the paths. Upon reception of a number of pieces, the destination generates an acknowledgment

[2] An alternative implementation, with each packet carrying a purse of fictitious currency from which nodes remove their reward, faces different challenges as well.

informing the source of which pieces, and thus routes, were intact. In order to enhance the robustness of the feedback mechanism, the small-sized acknowledgments are maximally dispersed (i.e., successful reception of at least one piece is sufficient) and are protected by the protocol header as well. If less than M pieces were received, the source retransmits the remaining pieces over the intact routes. If too few pieces were acknowledged or too many messages remain outstanding, the protocol adapts its operation, by determining a different path set, reencoding undelivered messages and reallocating pieces over the path set. Otherwise, it proceeds with subsequent message transmissions.

The protocol exploits MANET features such as the topological redundancy, interoperates widely with accepted techniques such as source routing, relies on a security association between the source and the destination, and makes use of highly efficient symmetric key cryptography. It does not impose processing overhead on intermediate nodes, while the end nodes make the routing decisions based on the feedback provided by the destination and the underlying topology discovery and route maintenance protocols. The fault tolerance of SMT is enhanced by the adaptation of parameters such as the number of paths and the dispersion factor (i.e., the ratio of required pieces to the total number of pieces). SMT can yield 100% successful message reception, even if 10–20% of the network nodes are malicious. Moreover, algorithms for the selection of path sets with different properties, based on different metrics and the network feedback, can be implemented by SMT. SMT provides a flexible, end-to-end, secure traffic engineering scheme, tailored to the MANET characteristics.

It is noteworthy that SMT provides a limited protection against the use of compromised topological information, although its main focus is to safeguard the data forwarding operation. The use of multiple routes compensates for the use of partially incorrect routing information [52], rendering a compromised route equivalent to a route failure. Nevertheless, the disruption of the route discovery can still be the most effective way for adversaries to consistently compromise the communication of one or more pairs of nodes.

Another approach to secure the route discovery procedures, the secure routing protocol (SRP), has been proposed [63]. The scheme guarantees that a node initiating a route discovery will be able to identify and discard replies providing false topological information, or avoid receiving them. The novelty of the scheme, as compared with other MANET secure routing schemes, is that false route replies, as a result of malicious node behavior, are discarded partially by benign nodes while in transit toward the querying node, or deemed invalid on reception. The security goals are achieved with the existence of a security association between the pair of end nodes *only*, without the need for intermediate nodes to cryptographically validate control traffic.

The widely accepted technique in the MANET context of route discovery based on broadcasting query packets is the basis of the SRP protocol. More specifically, as query packets traverse the network, the relaying intermediate nodes append their identifier (e.g., *IP* address) in the query packet header. When one or more queries arrive at the sought destination, replies that contain the accumulated routes are returned to the querying node; the source then may use one or more of these routes to forward its data.

Reliance on this basic route query broadcasting mechanism allows SRP to be applied as an extension of a multitude of existing routing protocols. In particular, the dynamic source routing (DSR) [17] and the IERP [64] of the zone routing protocol (ZRP) [65] framework are two protocols that can be extended in a natural way to incorporate SRP. Furthermore, other protocols such as ABR [66], for example, could be combined with SRP with minimal modifications to achieve the security goals of the SRP protocol.

In SRP, only the end nodes have to be securely associated, and there is no need for cryptographic validation of control traffic at intermediate nodes; two factors that render the scheme efficient and scalable. SRP places the overhead on the end nodes, an appropriate choice for a highly decentralized environment, and contributes to the robustness and flexibility of the scheme. Moreover, SRP does not rely on the state stored in intermediate nodes, thus is immune to malicious acts not directed against the nodes that wish to communicate in a secure manner. Finally, SRP provides one or more route replies, whose correctness is verified by the route "geometry" itself.

6. SOME CONCLUDING THOUGHTS

The ad hoc networking technology has stimulated substantial research activity since the early 1990s. The rather interesting fact is that although the military has been experimenting and even using this technology since 1970s, the research community has been coping with the rather frustrating task of finding a "killer" nonmilitary application for ad hoc networks. A major challenge that has been perceived as a possible "show stopper" for technology transfer is the fact that commercial applications do not necessarily conform to the "collaborative" environment that the military communication environment does. In other words, why should a user forward someone else's transmission, depleting his/her own battery power and, thus, possibly restricting his/her use of the network in the future? This question may relate to the issue of billing — if billing is possible (and, in fact, desirable), then nodes that serve as "good citizens" could be rewarded. But billing is, by itself, a significant challenge in ad hoc networks.

Other challenges in deployment of ad hoc networks relate to the issues of manageability, security, and availability of communication through this type of technology.

Realizing that technology transfer has not been the motivating factor, the most recent research interest in ad hoc networks could be, most probably, attributed to the intellectual challenges that are part of this type of communication environment.

Nevertheless, we believe that there is substantial commercial potential of ad hoc networks. Future extensions of the cellular infrastructure could be carried out using this

type of technology and may very well be the basis for the fourth-generation (4G) of wireless systems. Other possible applications include sensing systems (also referred to as *sensor networks)* or augmentation to the wireless LAN technology.

BIOGRAPHIES

Zygmunt J. Haas received his B.Sc. in EE in 1979 and M.Sc. in EE in 1985. In 1988, he earned his Ph.D. from Stanford University, Stanford, California, and subsequently joined AT&T Bell Laboratories in the Network Research Department. There he pursued research on wireless communications, mobility management, fast protocols, optical networks, and optical switching. From September 1994 to July 1995 Dr. Haas worked for the AT&T Wireless Center of Excellence, where he investigated various aspects of wireless and mobile networking, concentrating on TCP/IP networks. As of August 1995, he joined the faculty of the School of Electrical and Computer Engineering at Cornell University, Ithaca, New York.

Dr. Haas is an author of numerous technical papers and holds 15 patents in the fields of high-speed networking, wireless networks, and optical switching. He has organized several workshops, delivered tutorials at major IEEE and ACM conferences, and serves as editor of several journals and magazines, including the *IEEE Transactions on Networking*. He has been a guest editor of three IEEE JSAC issues ("Gigabit Networks," "Mobile Computing Networks," and "Ad-Hoc Networks"). Dr. Haas is a senior member of IEEE, a voting member of ACM, and the chair of the IEEE Technical Committee on Personal Communications. His interests include: mobile and wireless communication and networks, personal communication service, and high-speed communication and protocols. His e-mail is: haas@ece.cornell.edu and his URL is: http://wnl.ece.cornell.edu.

Jing Deng received his B.E. in 1994 and M.E. in 1997, both in EE, from Tsinghua University in Beijing, P. R. China. He earned his Ph.D. degree from Cornell University, Ithaca, New York, in 2002. Dr. Deng's interests include protocol design, evaluation, and optimization for mobile ad-hoc networks. His e-mail address is: jing@ece.cornell.edu.

Dr. Ben Liang received his B.Sc. and M.Sc. degrees, both in electrical engineering, from Polytechnic University in 1997. In 2001, he received his Ph.D. degree in electrical engineering from Cornell University, Ithaca, New York. He is currently a postdoctoral research associate and visiting lecturer in the School of Electrical and Computer Engineering at Cornell University. He has pursued research in wireless networking, mobility management, distributed fault-tolerance, and image processing. His current research interests are in developing theories, algorithms, and protocols for wireless and mobile communication systems. Dr. Liang is a member of IEEE and Tau Beta Pi.

Panagiotis Papadimitratos is a Ph.D. candidate in the Electrical and Computer Engineering Department at Cornell University, Ithaca, New York. He is currently a graduate research assistant, member of the Wireless Network Laboratory, working under the supervision of Professor Z. J. Haas. His research examines issues in the general areas of mobile and wireless networking. His work focuses on the design and evaluation of secure and fault tolerant routing protocols to support mobile ad-hoc networking. Mr. Papadimitratos joined the Ph.D. program at Cornell University in 1998, having received his Diploma degree from the Computer Engineering and Informatics Department at the University of Patras, Patras, Greece. His Diploma thesis examined issues on adaptive and blind equalization for cellular mobile networks.

S. Sajama received her bachelor of technology from the Indian Institute of Technology, Bombay, India, in 1998. Subsequently, she has joined the School of Electrical and Computer Engineering at Cornell University, Ithaca, New York, obtaining her masters degree in August 2001. Her thesis concentrated on multicast routing protocols with application to ad-hoc networks. Her e-mail address is: sajama@ece.cornell.edu

BIBLIOGRAPHY

1. C. E. Perkins, ed., *Ad Hoc Networking*, Addison-Wesley Longman, 2001.
2. F. A. Tobagi and L. Kleinrock, Packet switching in radio channels: Part II — The hidden terminal problem in carrier multiple-access and the busy-tone solution, *IEEE Trans. Commun.* **COM-23**(12): 1417–1433 (Dec. 1975).
3. J. Postel, *Internet Control Message Protocol*, RFC 792, IETF, Sept. 1981.
4. P. Karn, MACA — a new channel access method for packet radio, *ARRL/CRRL Amateur Radio 9th Computer Networking Conf.*, 1990, pp. 134–140.

4.5 Wu C. and Li V. O. K., Receiver-initiated busy-tone multiple access in packet radio networks, *Proc. ACM SIGCOMM'87*, 1987, pp. 336–342.

5. V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, MACAW: A media access protocol for wireless LAN's, *Proc. ACM SIGCOMM'94*, 1994, pp. 212–225.
6. C. L. Fullmer and J. J. Garcia-Luna-Aceves, Floor acquisition multiple access (FAMA) for packet-radio networks, *Proc. ACM SIGCOMM'95*, 1995, pp. 262–273.
7. C. L. Fullmer and J. J. Garcia-Luna-Aceves, Solutions to hidden terminal problems in wireless networks, *Proc. ACM SIGCOMM'97*, 1997, pp. 39–49.
8. Z. J. Haas and J. Deng, Dual busy tone multiple access (DBTMA): A multiple access control scheme for ad hoc networks, *IEEE Trans. Commun.* (in press).
9. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice-Hall, 1992.
10. C. Cheng, R. Reley, S. P. R. Kumar, and J. J. Garcia-Luna-Aceves, A loop-free extended Bellman-Ford routing protocol without bouncing effect, *ACM Comput. Commun. Rev.* **19**(4): 224–236 (1989).
11. J. J. Garcia-Luna-Aceves, Loop-free routing using diffusing computations, *IEEE/ACM Trans. Network.* **1**(1): 130–141 (Feb. 1993).

12. C. E. Perkins and P. Bhagwat, Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers, *ACM SIGCOMM Comput. Commun. Rev.* **24**(4): 234–244 (Oct. 1994).

13. P. Jacquet, P. Muhlethaler, and A. Qayyum, *Optimized Link State Routing Protocol*, IETF MANET, Internet Draft, Nov. 1998.

14. S. Murthy and J. J. Garcia-Luna-Aceves, A routing protocol for packet radio networks, *Proc. ACM Mobile Computing and Networking Conf., MOBICOM'95*, Nov. 14–15, 1995.

15. S. Murthy and J. J. Garcia-Luna-Aceves, An efficient routing protocol for wireless networks, *ACM Mobile Networks Appl. J.* **1**(2): 187–197 (Oct. 1996).

16. V. D. Park and M. S. Corson, A highly adaptive distributed routing algorithm for mobile wireless networks, *IEEE INFOCOM '97*, Kobe, Japan, 1997.

17. D. B. Johnson and D. A. Maltz, Dynamic source routing in ad hoc wireless networks, in T. Imielinski and H. Korth, eds., *Mobile Computing*, Kluwer Academic Publishers, 1996, pp. 153–181.

18. C. E. Perkins and E. M. Royer, Ad-hoc on-demand distance vector routing, *Proc. 2nd IEEE Workshop on Mobile Computing Systems and Applications*, Feb. 1999, pp. 90–100.

19. M. R. Pearlman and Z. J. Haas, Determining the optimal configuration for the zone routing protocol, *IEEE J. Select. Areas Commun.* **17**(8): 1395–1414 (Aug. 1999).

20. V. Park and M. S. Corson, IETF MANET Internet Draft, *draft-ietf-MANET-tora-spec-03.txt*, Nov. 2000.

21. E. Gafni and D. Bertsekas, Distributed algorithms for generating loop-free routes in networks with frequently changing topology, *IEEE Trans. Commun.* **29**(1): 11–15 (Jan. 1981).

22. M. S. Corson and A. Ephremides, A distributed routing algorithm for mobile wireless networks, *ACM/Baltzer Wireless Networks* **1**(1): 61–81 (Feb. 1995).

23. P. Jacquet et al., IETF MANET Internet Draft, *draft-ietf-MANET-olsr-02.txt*, July 2000.

24. A. Iwata et al., Scalable routing strategies for ad hoc wireless networks, *IEEE J. Select. Areas Commun.* **17**(8): 1369–1379 (Aug. 1999).

25. T.-W. Chen and M. Gerla, Global state routing: A new routing scheme for ad-hoc wireless networks, *IEEE ICC* 171–175 (June 1998).

26. R. Sivakumar, P. Sinha, and V. Bharghavan, CEDAR: A core-extraction distributed ad hoc routing algorithm, *IEEE J. Select. Areas Commun.* **17**(8): 1454–1465 (Aug. 1999).

27. M. Joa-Ng and I.-T. Lu, A peer-to-peer zone-based two-level link state routing for mobile ad hoc networks, *IEEE J. Select. Areas Commun.* **17**(8): 1415–1425 (Aug. 1999).

28. P. F.Tsuchiya, The landmark hierarchy: A new hierarchy for routing in very large networks, *Comput. Commun. Rev.* **18**(4): 35–42 (Aug. 1988).

29. G. Pei, M. Gerla, and X. Hong, LANMAR: Landmark routing for large scale wireless ad hoc networks with group mobility, *1st Annual Workshop on Mobile and Ad Hoc Networking and Computing* (*MobiHOC*), Aug. 2000.

30. Y.-B. Ko and N. H. Vaidya, Location-aided routing (LAR) in mobile ad hoc networks, *ACM/IEEE MobiCom*, Dallas, TX, 1998.

31. S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, A distance routing effect algorithm for mobility (DREAM), *ACMIEEE MobiCom*, Dallas, TX, 1998.

32. S. Paul, *Multicasting on the Internet and its Applications*, Kluwer, 1998.

33. C.-H. Chow, On multicast path finding algorithms, *Proc. IEEE INFOCOM'91*, 1991, pp. 1274–1283.

34. B. M. Waxman, Performance evaluation of multipoint routing algorithms, *Proc. IEEE INFOCOM'93*, 1993, pp. 980–986.

35. B. Kadaba and J. M. Jaffe, Routing to multiple destinations in computer networks, *IEEE Trans. Commun.* **COM-31**(3): 343–351 (March 1983).

36. V. P. Kompella, J. C. Pasquale, and G. C. Polyzos, Multicast routing for multimedia communication, *IEEE/ACM Trans. Network.* **1**(3): 286–292 (June 1993).

37. S. E. Deering et al., An architecture for wide-area multicast routing, *IEEE/ACM Trans. Network.* **4**(2): 153–162 (April 1996).

38. A. Ballardie, *Core Based Trees* (*CBT Version 2*) *Multicast Routing — Protocol Specification*, RFC-2189, Sept. 1997.

39. C. Chiang, M. Gerla, and L. Zhang, Shared tree wireless network multicast, *IEEE International Conf. Computer Communications and Networks* (*ICCCN'97*), Sept. 1997.

40. S.-J. Lee, M. Gerla, and C.-C. Chiang, On-demand multicast routing protocol, *IEEE WCNC'99*, New Orleans, LA, Sept. 1999, pp. 1298–1304.

41. C. W. Wu and Y. C. Tay, AMRIS: A multicast protocol for ad hoc wireless networks, *IEEE MILCOM'99*, Atlantic City, NJ, Nov. 1999.

42. J. J. Garcia-Luna-Aceves and E. L. Madruga, The Core-assisted mesh protocol, *IEEE J. Select. Areas Commun.* **17**(8): (Aug. 1998).

43. E. Royer and C. E. Perkins, Multicast operation of ad-hoc, on-demand distance vector routing protocol, *ACM/IEEE MobiCom'99*, Aug. 1999.

44. E. Bommaiah, A. McAuley, R. Talpade, and M. Liu, AMRoute: Ad hoc multicast routing protocol, Internet-Draft, IETF, Aug. 1998.

45. S. Lee and C. Kim, Neighbor supporting ad hoc multicast routing protocol, *2000 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing*, 2000, pp. 37–44.

46. L. Briesemeister and G. Hommel, Role-based multicast in highly mobile but sparsely connected ad hoc networks, *2000 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing*, 2000, pp. 45–50.

47. H. Zhou and S. Singh, Content based multicast in ad hoc networks, *2000 1st Annual Workshop on Mobile and Ad Hoc Networking and Computing*, 2000, pp. 51–60.

48. G. D. Kondylis, S. V. Krishnamurthy, S. K. Dao, and G. J. Pottie, Multicasting sustained CBR and VBR traffic in wireless ad hoc networks, *Int. Conf. Communications*, 2000, pp. 543–549.

49. C. Sankaran and A. Ephremides, Multicasting with multiuser detection in ad-hoc wireless networks, *Conf. Information Sciences and Systems* (*CISS*), 1998, pp. 47–54.

50. P. Sinha, R. Sivakumar, and V. Bharghavan, MCEDAR: Multicast core-extraction distributed ad hoc routing, *Wireless Communications and Networking Conf.*, 1999, pp. 1313–1318.

51. J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, Algorithms for energy-efficient multicasting in ad hoc wireless networks, *IEEE Military Communications Conf.* (*MILCOM*), 1999, pp. 1414–1418.

52. L. Zhou and Z. J. Haas, Securing ad hoc networks, *IEEE Network Mag.* **13**(6): (Nov./Dec. 1999).

53. F. Stajano and R. Anderson, The resurrecting duckling: Security issues for ad hoc wireless networks, *Security Protocols, 7th Int. Workshop*, LNCS, Springer-Verlag, 1999.

54. S. Marti, T. J. Giuli, K. Lai, and M. Baker, Mitigating routing misbehavior in mobile ad hoc networks, *6th MOBICOM Conf.*, Boston, Aug. 2000.

55. L. Buttyan and J. P. Hubaux, Enforcing service availability in mobile ad hoc WANs, *1st MobiHoc conf.*, Boston, Aug. 2000.

56. S. Yi, P. Naldurg, and R. Kravets, *Security-Aware Ad-Hoc Routing for Wireless Networks*, UIUCDCS-R-2001-2241 Technical Report, Aug. 2001.

57. M. Guerrero, Secure AODV, Internet draft sent to *manet@itd.nrl.navy.mil* mailing list, Aug. 2001.

58. R. Rivest, A. Shamir, and L. Adleman, A method for obtaining Digital Signatures and public key cryptosystems, *Commun. ACM* **21**(2): 120–126 (Feb. 1978).

59. L. Lamport, Password authentication with insecure communication, *Commun. ACM* **24**(11): 770–772 (Nov. 1981).

60. R. Droms, *Dynamic Host Configuration Protocol*, RFC 2131, IETF, March 1997.

61. M. Hattig, ed., Zero-conf IP host requirements, *draft-ietf-zeroconf-reqts-09.txt*, IETF MANET Working Group, Aug. 31, 2001.

62. P. Papadimitratos and Z. J. Haas, *Secure message transmission in mobile ad hoc networks* (in preparation).

63. P. Papadimitratos and Z. J. Haas, Secure routing for mobile ad hoc networks, *SCS Communication Networks and Distributed Systems Modeling and Simulation Conference* (*CNDS 2002*), San Antonio, TX, Jan. 27–31 2002.

64. Z. J. Haas, M. Perlman, and P. Samar, The interzone routing protocol (IERP) for ad hoc networks, *draft-ietf-manet-zone-ierp-01.txt*, MANET Working Group, IETF, June 1, 2001.

65. Z. J. Haas and M. Perlman, The performance of query control schemes of the zone routing protocol, *IEEE/ACM Trans. Network.* **9**(4): 427–438 (Aug. 2001).

66. C. K. Toh, Associativity-based routing for ad-hoc mobile networks, *Wireless Pers. Commun.* **4**(2): 1–36 (March 1997).

FURTHER READING

Asokan N. and P. Ginzboorg, Key agreement in ad hoc networks, *Comput. Commun.* **23**(17): 1627–1637 (Nov. 2000).

Bellovin S. M. and M. Merritt, Encrypted key exchange: Password-based protocols secure against dictionary attacks, *IEEE SympSecurity and Privacy* (May 1992).

Bellur B. and R. G. Ogier, A reliable, efficient topology broadcast protocol for dynamic networks, *IEEE INFOCOM*, New York, March 1999.

Bianchi G., Performance analysis of the IEEE 802.11 distributed coordination function, *IEEE JSAC* **18**(3): 535–547 (March 2000).

Binkley J. and W. Trost, Authenticated ad hoc routing at the link layer for mobile systems, *ACM/Baltzer Wireless Networks* 139–145 (March 2001).

Blaze M., J. Feigenbaum, J. Ioannidis, and A. D. Keromytis, *The KeyNote Trust-Management System*, RFC 2074, IETF, Sept. 1999.

Broch J., D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, A performance comparison of multi-hop wireless ad hoc network routing protocols, *ACM/IEEE MobiCom*, 1998, pp. 85–97.

Chiang C.-C., H.-K. Wu, W. Liu, and M. Gerla, Routing in clustered multihop, mobile wireless networks with fading channel, *IEEE Singapore Int. Conf. on Networks*, 1997.

Das S. R., C. E. Perkins, and E. M. Royer, Performance comparison of two on-demand routing protocols for ad hoc networks, *IEEE INFOCOM* **1**: 3–12 (March 2000).

Dube R., C. D. Rais, K.-Y. Wang, and S. K. Tripathi, Signal stability-based adaptive routing (SSA) for ad hoc mobile networks, *IEEE Pers. Commun.* **4**(1): 36–45 (Feb. 1997).

Ephremeides A., J. E. Wieselthier, and D. J. Baker, A design concept for reliable mobile radio networks with frequency hopping signaling, *Proc. IEEE* **75**(1): (Jan. 1987).

Feeney L. M., B. Ahlgren, and A. Westerlund, Spontaneous networking: An application-oriented approach to ad hoc networking, *IEEE Commun. Mag.* **39**(6): 176–181 (June 2001).

FedInfProcStandards, *Secure Hash Standard*, Publication 180, NIST, May 1993.

Gerla M. and T. C. Tsai, Multiuser, mobile multimedia radio network, *ACM/Balzer J. Wireless Networks* **1**(3): 255–265 (1995).

Gerla M., K. Tang, and R. Bagrodia, TCP performance in wireless multi-hop networks, *Mobile Comput. Syst. Appl.* 41–50, 25–26 (Feb. 1999).

Garcia-Luna-Aceves J. J. and M. Spohn, Source-tree routing in wireless networks, *IEEE ICNP 99: 7th Int. Conf. Network Protocols*, Toronto, Canada, Oct. 1999.

http://www.scienceforum.com/glomo/.

Hauser R., T. Przygenda, and G. Tsudik, Lowering security overhead in link state routing, *Comput. Networks* **31**: 885–894 (1999).

IEEE STD 802.11, *Wireless LAN Media Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE, 1999.

Jiang M., J. Li, and Y. C. Tay, IETF MANET Internet Draft, *draft-ietf-MANET-cbrp-spec-01.txt*, Aug. 1999.

Johansson P. et al., Scenario-based performance analysis of routing protocols for mobile ad-hoc networks, *ACM/IEEE MobiCom*, Aug. 1999, pp. 195–206.

Jubin J. and J. D. Tornow, The DARPA packet radio network protocols, *Proc. IEEE* **75**: 21–32 (Jan. 1987).

Kershenbaum A. and W. Chou, A unified algorithm for designing multidrop teleprocessing networks, *IEEE Trans. Commun.* **COM-22**: 1762–1772 (Nov. 1974).

Kondylis G. D., S. V. Krishnamurthy, S. K. Dao, and G. J. Pottie, Multicasting sustained CBR and VBR traffic in wireless ad hoc networks, *Int. Conf. Communications*, 2000, pp. 543–549.

Nikaein N., H. Labiod, and C. Bonnet, DDR-distributed dynamic routing algorithm for mobile ad hoc networks, *1st Annual Workshop on Mobile and Ad Hoc Networking and Computing* (*MobiHOC*), Aug. 2000.

Perkins C., *IP Mobility Support*, RFC 2002, IETF, Oct. 1996.

Ramanathan R. and M. Steenstrup, Hierarchically-organized, multihop mobile wireless networks for quality-of-service support, *ACM/Baltzer Mobile Networks Appl.* **3**(1): 101–119 (1998).

Royer E. M. and C.-K. Toh, A review of current routing protocols for ad hoc mobile wireless networks, *IEEE Pers. Commun.* (April 1999).

Ruppe R. and S. Griswald, Near term digital radio (NTDR) system, *IEEE MILCOM* **3**: 1282–1287 (Nov. 1997).

Santivanez C., *Asymptotic Behavior of Mobile Ad Hoc Routing Protocols with Respect to Traffic, Mobility and Size*, Technical Report TR-CDSP0052, Dept. Electrical and Computer Engineering, Northeastern Univ., Boston, 2000.

Shacham N. and J. Westcott, Future directions in packet radio architectures and protocols, *Proc. IEEE* **75**: 83–99 (Jan. 1987).

Sharony J., An architecture for mobile radio networks with dynamically changing topology using virtual subnets, *ACM Mobile Networks Appl. J.* **1**(1): 75–86 (1996).

Stajano F., The resurrecting duckling — what next? *Security Protocols, 8th International Workshop*, LNCS, Springer-Verlag, 2000.

Xu S. and T. Saadawi, Does the IEEE 802.11 MAC Protocol Work Well in Multihop Wireless Ad Hoc Networks? *IEEE Communi. Magazine* 130–137 (June 2001).

WIRELESS APPLICATION PROTOCOL (WAP)

ALESSANDRO ANDREADIS
GIOVANNI GIAMBENE
University of Siena
Siena, Italy

1. INTRODUCTION

It is expected that within a few years the number of mobile devices accessing the Internet will exceed the number of *personal computers* (PCs). The use of mobile terminals is becoming attractive, since they allow the access on the move and require a shorter setup time than the initial power-on of a PC. However, Internet services have not been developed for mobile devices, as they are not suitable for small displays and they are not personalized or location-dependent. A solution to these needs is represented by the *Wireless Application Protocol* (WAP), which is an open, global specification that empowers mobile users with wireless devices for easy Internet access [1,2].

WAP is the result of the WAP Forum's efforts to promote industrywide specifications for developing applications and services that operate over wireless communication networks [3]. The WAP Forum was formed after a U.S. network operator, Omnipoint, issued a tender for the supply of mobile information services in early 1997. It received several responses from different suppliers using proprietary techniques for delivering the information such as *Smart Messaging* from Nokia and *Handheld Device Markup Language* (HDML) from Phone.com. Omnipoint informed the tender responders that it would not accept a proprietary approach and recommended that various vendors get together to explore defining a common standard. After all, there was not a great deal of difference between the different approaches, which could be combined and extended to form a powerful standard. These events triggered the development of WAP, with Ericsson and Motorola joining Nokia and Phone.com as the founder members of the WAP Forum. At present, the WAP Forum encompasses more than 500 members.

WAP is the de facto standard for porting Internet services to wireless devices, such as mobile phones and *Personal Digital Assistants* (PDA). WAP contains a lightweight protocol stack (based on the Internet one) well suited for the wireless scenario and an application environment that allows delivering information services to mobile phones. WAP can be built on many mobile phone operating systems, including: PalmOS, EPOC, Windows CE and JavaOS [3]. WAP also contains a microbrowser according to which the information received is interpreted in the handset and presented to the user. WAP is designed to work with most wireless networks such as CPDC, CDMA, GSM, PDC, PHS, TETRA, and DECT [3]. WAP devices, despite the current rather limited user interface, provide a valuable means to access corporate and public services. Examples of applications are

- Use of a corporate application in less time than it takes to boot the laptop
- Availability of the device on the move
- Access to services from several countries
- Mobile browsing and mobile access to email

2. WAP SYSTEM BUILDING BLOCKS

The WAP network architecture envisages both WAP servers, hosting pages designed in a suitable markup language and WAP gateways between the wireless network and the wireline Internet (see Fig. 1) [4]. WAP 1.0 was based on *Wireless Markup Language* (WML), a subset of the *eXtensible Markup Language* (XML). The basic markup language in WAP 2.0, namely, WML2, is based on the *eXtensible HyperText Markup Language* (XHTML), as defined by the *World Wide Web Consortium* (W3C) [5]. By using the XHTML modularization approach, the WML2 language is very extensible, permitting additional language elements to be added as needed.

The WAP protocol architecture is based on a client/server model, as sketched in Fig. 2, where we can see the mobile client, that is the mobile user, the WAP proxy and the Web server. The client Web browser makes a request for a Webpage. This request is sent to the WAP proxy that acts as a gateway for the Internet [6]. Through a protocol conversion, a *HyperText Transfer Protocol* (HTTP) request is thus sent to the appropriate Web

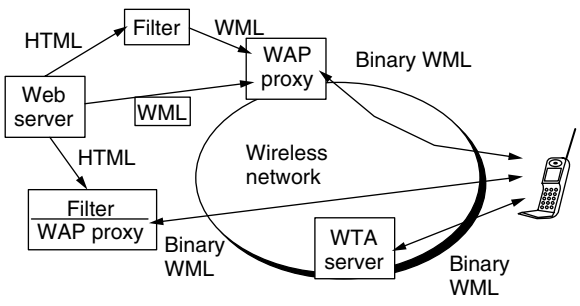


Figure 1. WAP system architecture.

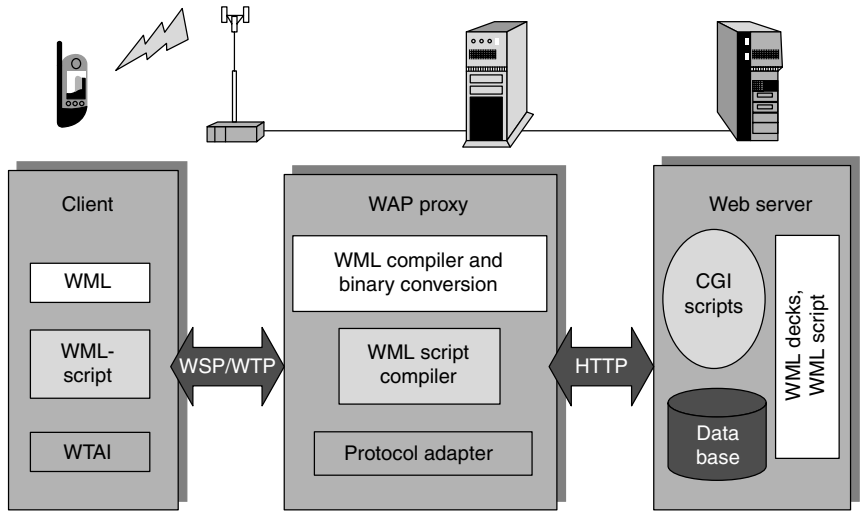


Figure 2. The interoperation of the WAP elements.

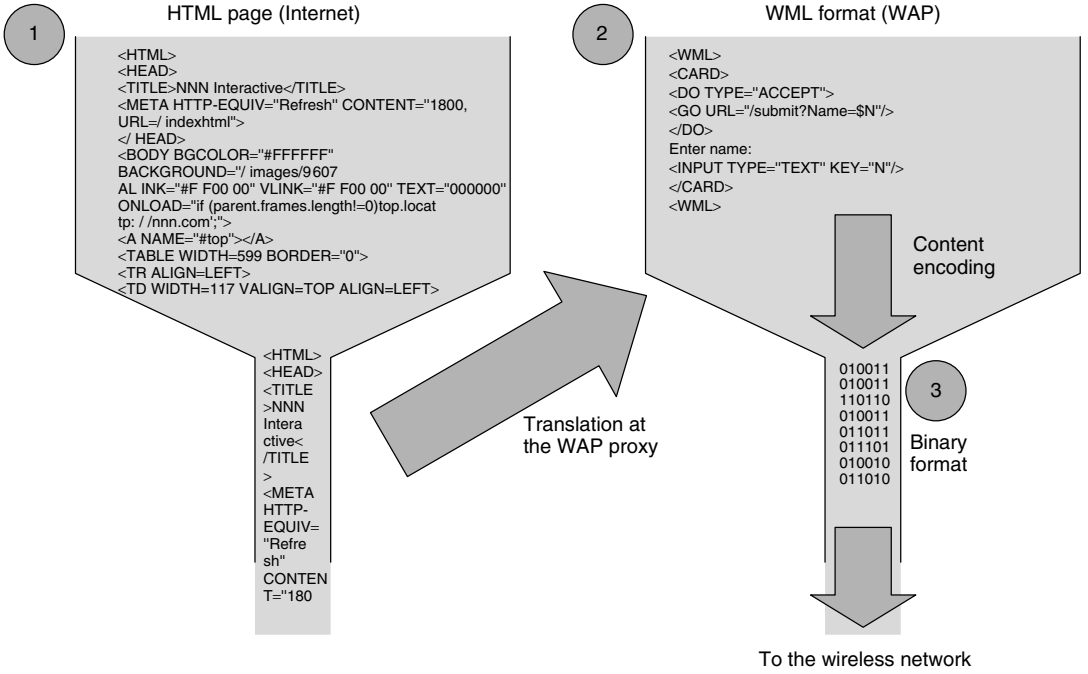


Figure 3. WML compiling and encoding.

server. The response is a bytestream of ASCII text, which is a *HyperText Markup Language* (HTML) Webpage. The use of *Common Gateway Interface* (CGI) programs or Java servlets allows for the dynamic creation of HTML pages using content stored in a database. The Webpage is sent to the WAP proxy that first translates it from HTML to the WML language (see Fig. 3, step 2). Finally, a page conversion is performed in a compact binary representation that is suitable for wireless networks (see Fig. 3, step 3) [7]. In the case that there is a WAP server (i.e., hosting pages in the WML format) in the mobile network, the mobile client directly receives WML pages from the server without involvement of the WAP gateway.

The infrastructure required to deliver WAP services to the mobile terminal is similar to that of the existing WWW model. The Web server is the same product; the one most commonly used is *Apache*. The Web server needs to be configured to serve the pages written in WML as well as HTML.

Although reusing of existing Internet content by means of on-the-fly adaptations and translations is an explicit goal, test realizations of the WAP gateway and proxy servers show that the creation of new content that is explicitly designed for presentation using WML is a more effective option.

Since a mobile user cannot use a QWERTY keyboard or a mouse, WML documents are structured into a set of well-defined units of user interactions called *cards*. Each card may contain instructions for gathering user input, information to be presented to the user, and similar (see Fig. 4). A single collection of cards is called a *deck*, which is the unit of content transmission, identified by

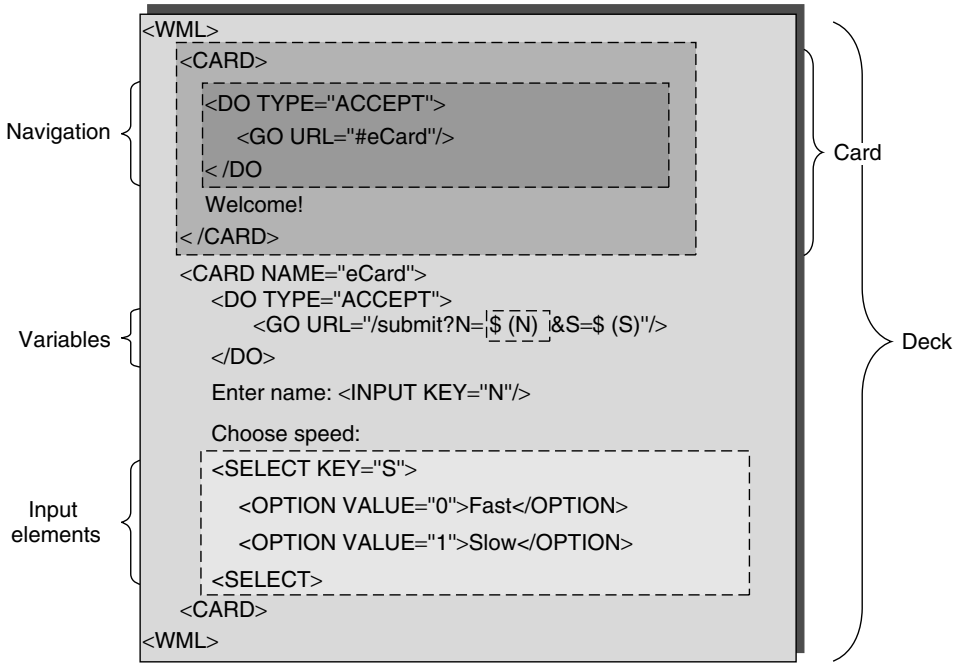


Figure 4. Internal organization of a WML page (=deck) in cards, with different tags.

a *Uniform Resource Locator* (URL) [8]. After browsing a deck, the WAP-enabled phone displays the first card; then, the user decides whether to proceed to the next card of the same deck. WML content is scalable from a two-line text display on a basic device to a full graphic screen on the latest smart phones and communicators. WML supports:

- Text (bold, italics, underlined, line breaks, tables)
- Black-and-white images (wireless bitmap format, WBMP)
- User input
- Variables
- Navigation and history stack
- Scripting (WMLScript), a lightweight script language, similar to JavaScript

In particular, WML includes support for managing user agent state by means of variables and for tracking the history of the interaction. Moreover, WMLScripts are sent separately from decks and are used to enhance the client *man–machine interface* (MMI) with sophisticated device and peripheral interactions.

The *Wireless Telephony Application* (WTA) of WAP contains a client-side WTA programming library and a WTA server (see Figs. 1 and 2); together they allow the WAP session to control the voice channel. WTA and its interface, *WTA-Interface* (WTAI), provide the access and the programming interface to telephony services. The WTA server generates WTA events interpreted by the WAP gateway that sends the resulting WML to the WAP mobile phone. The WTA server then initiates and controls any voice connections that are required.

3. WAP PROTOCOL STACK

WAP protocol and its functions are layered similarly to the OSI Reference Model [9]. In particular, the WAP protocol stack is analogous to the Internet one (see Fig. 5). Each layer is accessible by the layers above, as well as by other services and applications. The WAP layered architecture enables other services and applications to utilize the features of the WAP stack through a set of well-defined interfaces. Figure 6 compares Internet and WAP protocol stacks. A brief survey of the protocols at the different WAP layers is provided below.

3.1. Wireless Application Environment (WAE)

WAE specifies an application framework for wireless devices such as mobile phones, pagers, and PDAs. WAE specifies the markup languages and acts as a container for applications such as a microbrowser. In particular, WAE encompasses the following parts:

- WML Microbrowser
- WMLScript Virtual Machine
- WMLScript Standard Library
- Wireless Telephony Application Interface (i.e., telephony services and programming interfaces)
- WAP Content Types

The two most important formats defined in WAE are the WML and WMLScript byte-code formats. A WML encoder at the WAP gateway, or "tokenizer," converts a WML deck into its binary format (see Fig. 3, step 3) [7] and a WMLScript compiler takes a script into byte-code. This process allows a significant compression of the data to be transmitted on the air interface, thus making more efficient the transmission of WML and WMLScript data.

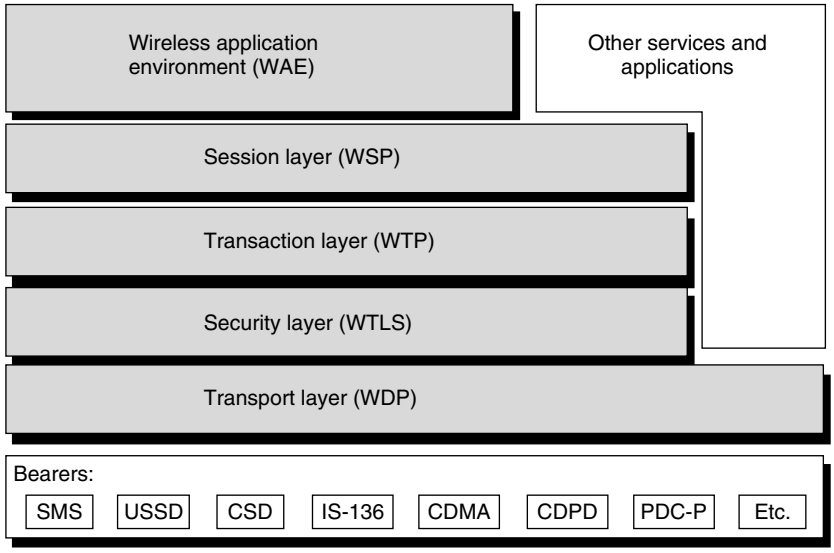


Figure 5. WAP 1.0 protocol stack.

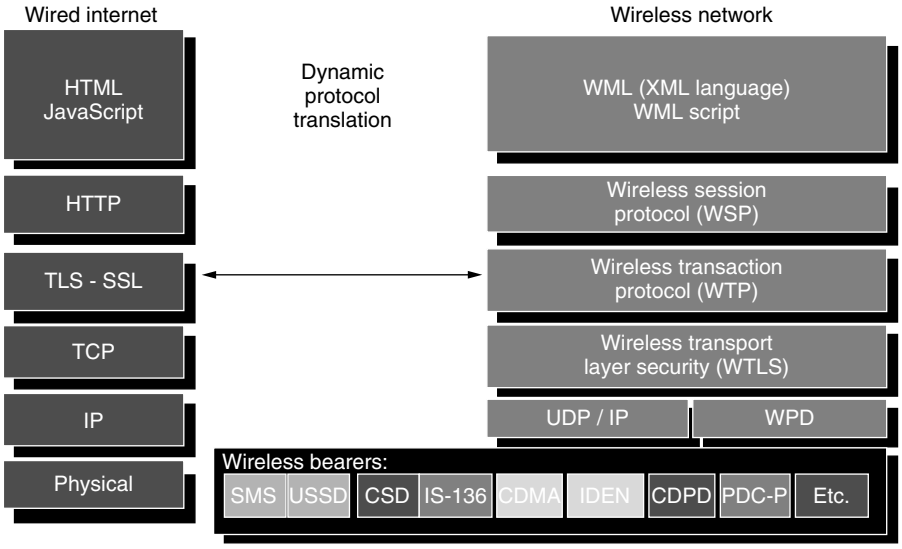


Figure 6. Internet and WAP 1.0 protocol stacks.

3.2. Wireless Session Protocol (WSP)

WSP provides the application layer of WAP (i.e., WAE) with a consistent interface for two-session services. The first is a connection-oriented service above the *Wireless Transaction Protocol* (WTP). The second is a connectionless service operating above a secure or nonsecure datagram service [*Wireless Datagram Protocol* (WDP)]. WSP is the equivalent of the HTTP protocol in both the Internet and WAP 2.0 release that supports the TCP/IP levels in the protocol stack.

3.3. Wireless Transaction Protocol (WTP)

WTP runs on top of a datagram service [such as the *User Datagram Protocol* (UDP)] and provides a lightweight transaction-oriented protocol that is suitable for implementation in mobile terminals. WTP offers three classes of transaction services: unreliable one-way request, reliable one-way request and reliable two-way request respond.

3.4. Wireless Transport Layer Security (WTLS)

WTLS is a security protocol based on the industry-standard *Transport Layer Security* (TLS) protocol. WTLS is intended for use with the WAP transport protocols and has been optimized for wireless communication networks. It includes data integrity checks, privacy on the WAP gateway-to-client leg, and authentication.

3.5. Wireless Datagram Protocol (WDP)

WDP is transport-layer protocol in WAP [10]. WDP supports connectionless reliable transport and bearer independence. WDP offers consistent services to the upper-layer protocols of WAP and operates above the data-capable bearer services supported by various air interface types. Since the WDP protocols provide a common interface to upper-layer protocols, the security, session, and application layers are able to operate independently of the underlying wireless network. At the mobile terminal, the WDP protocol consists of the common WDP elements

plus an adaptation layer that is specific for the adopted air interface bearer. The WDP specification lists the bearers that are supported and the techniques used to allow WAP protocols to operate over each bearer [3]. The WDP protocol is based on UDP. UDP provides port-based addressing, and IP provides *Segmentation And Reassembly* (SAR) in a connectionless datagram service. When the IP protocol is available over the bearer service, the WDP datagram service offered for that bearer will be UDP.

3.6. Bearers on the Air Interface

Let us refer to the *Global System for Mobile communications* (GSM) network, where the following bearer services can be adopted to support WAP traffic [11]:

- *Unstructured Supplementary Services Data* (USSD)
- *circuit-switched Traffic CHannel* (TCH)
- *Short Message Service* (SMS)
- *General Packet Radio Service* (GPRS), plain data traffic
- *Multimedia Messaging Service* (MMS) over GPRS

Let us compare these different options to support WAP traffic. TCH has the disadvantage of a 30–40s connection delay between the WAP client and the gateway, thus making it less suitable for mobile subscribers. SMS and USSD are inexpensive bearers for WAP data with respect to TCH, leaving the mobile device free for voice calls. SMS and USSD are transported by the same air interface channels. SMS is a store-and-forward service that relies on a *Short Message Service Center* (SMSC), whereas USSD is a connection-oriented (no store-and-forward) service, where the *Home Location Register* (HLR) of the GSM network receives and routes messages from/to the users. The SMS bearer is well suited for WAP *push* applications (available from WAP release 1.2), where the user is automatically notified each time an event occurs. USSD is particularly useful for supporting *transactions* over WAP. Finally, GPRS radio transmissions allow a high capacity [$\leq$170 kbps (kilobits per second) using all the slots of a GSM carrier with the lightweight coding scheme] that is shared among mobile phones according to a packet-switching scheme. Hence, GPRS can provide an efficient scheme for WAP contents delivery.

The WAP protocol layers at the client, at the gateway, and at the Web server are detailed in Fig. 7. Figure 8 gives further details on different WAP protocol stack possibilities on the client side. In particular, the leftmost stack represents a typical example of a WAP application, namely, a WAE user agent running over the complete

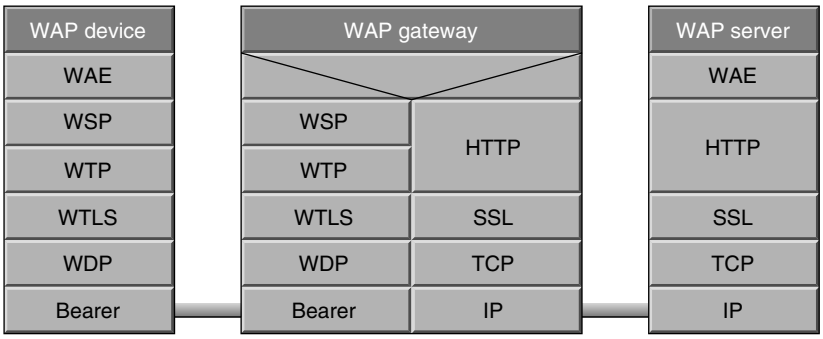


Figure 7. WAP 1.0 protocol architecture at different interfaces.

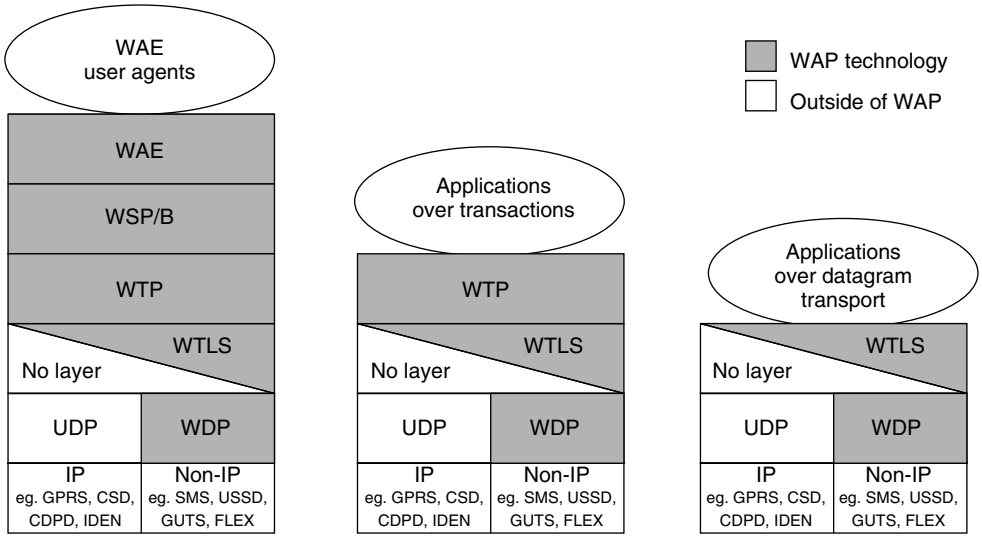


Figure 8. Different possibilities for the WAP protocol stack on the client side.

portfolio of WAP technology. The middle stack is intended for applications and services that require transaction services with or without security. The rightmost stack is intended for applications and services that only require datagram transport with or without security.

4. COMPARISON BETWEEN WAP PROTOCOL RELEASES

The differences between the different releases of the WAP protocol are detailed below:

- WAP 1.0: first version of software for mobile clients, first adoption of WML, WBMP image format
- WAP 1.1: WTAI — "clickable" phone numbers, support of tables, boldface types, encrypted communication
- WAP 1.2: support of push applications, telephone identification, certificate handling
- WAP 2.0: latest WAP release, WML replaced by XHTML, colour screens, banners, MP3 and MP4 audio files, Internet radio, Bluetooth, remote control, integration with *Mobile Positioning System* (MPS) for locating the users (location-aware services)

5. TOOLS AND APPLICATIONS

The WAP programming model is similar to the WWW programming one. This fact provides several benefits to the application developer community, including a proven architecture and the ability to leverage existing tools (e.g., Web servers, XML tools). Optimizations and extensions have been made in order to match the characteristics of the wireless environment. Different WAP browsers can be found in Ref. 12; they are useful tools for developing WAP-based services for mobile users. WAP allows customers to easily reply to incoming information on the phone by adopting new menus to access mobile services.

Existing mobile operators have added WAP support to their offering by either developing their own WAP interface or, more often, partnering with one of the WAP gateway suppliers. WAP has also given new opportunities to allow the mobile distribution of existing information contents. For example, CNN and Nokia teamed up to offer CNN Mobile. Moreover, Reuters and Ericsson teamed up to provide Reuters Wireless Services.

New mobile applications that can be made available through a WAP interface include:

- Location-aware services
- Web browsing
- Remote local-area network access
- Corporate email
- Document sharing/collaborative working
- Customer service
- Remote monitoring such as meter reading
- Job dispatch
- Remote point of sale
- File transfer
- Home automation
- Home banking and trading on line

Another group of important applications are based on the WAP push service that allows contents to be sent or "pushed" to devices by server-based applications via a push proxy. Push functionality is especially relevant for real-time applications that send notifications to their users, such as messaging, stock prices, and traffic update alerts. Without push functionality, these types of applications would require the devices to poll application servers for new information or status. In cellular networks such polling activities would cause an inefficient and wasteful use of the resources. WAP push functionality provides control over the lifetime of pushed messages, store-and-forward capabilities at the push proxy, and control over the bearer choice for delivery.

Interesting WAP applications are made possible by the creation of dynamic WAP pages by means of the following different options:

- Microsoft ASP
- Java and servlets or *Java Server Pages* (JSPs) for generating WAP decks
- *XSL Transformation* (XSLT) for generating WAP pages adapted for displays of different characteristics and sizes

Alternative approaches to the use of WAP for mobile applications could be as follows:

- *Subscriber Identity Module* (*SIM*) *Toolkit* — the use of SIMs or smart cards in wireless devices is already widespread.
- *Windows CE* — a multitasking, multithreaded operating system from Microsoft designed for including or embedding mobile and other space-constrained devices.
- *JavaPhone* — Sun Microsystems is developing PersonalJava and a JavaPhone *Application Programming Interface* (API), which is embedded in a Java virtual machine on the handset. Thus, cellular phones can download extra features and functions over the Internet.

SIM Toolkit and Windows CE are present days technologies as well as WAP. SIM Toolkint implies the definition of a set of services "embedded" on the SIM that allow users to contact several service providers through the mobile phone network. The Windows CE solution is based on an operating system developed for mobile devices and that may support different applications. Finally, JavaPhone will be the most sophisticated option for the development of device-independent applications.

Within the *European Telecommunications Standards Institute* (ETSI) and *3rd Generation Partnership Project* (3GPP), standardization activities are in progress for the realization of mobile services. Accordingly, a new standard, called *Mobile station application Execution Environment* (MExE), has been defined [13]. In order to

ensure the portability of a variety of applications, across a broad spectrum of multivendor mobile terminals, a dynamic and open architecture has been standardized for both the *Mobile Station* (MS) and the SIM, that is a common set of APIs and development tools. MExE is based on the idea to specify a terminal-independent execution environment on the client device (i.e., MS and SIM) for nonstandardized applications and to implement a mechanism that allows the negotiation of supported capabilities (taking into account available bandwidth, display size, processor speed, memory, MMI). The key concept of the MExE service environment to make applications mobile-aware (i.e., aware of MS capabilities, network bearer characteristics, and user preferences) is the introduction of MExE classmarks that have been standardized as follows:

- *MExE classmark 1* — service based on WAP; requires limited input and output facilities (e.g., as simple as a 3-line × 15-character display and a numeric keypad) on the client side and is designed to provide quick and cheap information access even over narrow and slow data connections.
- *MExE classmark 2* — service based on PersonalJava; provides and utilizes a run-time system requiring more processing, storage, display, and network resources, but supports more powerful applications and more flexible MMIs. MExE classmark 2 also includes support for MExE classmark 1 applications (via the WML browser).

6. CONCLUSIONS

With the advent of the information society there is a growing need for network operators to support the mobile access to the Internet and its most popular applications such as Web browsing, email, file transfer, and remote login. The WAP protocol proposed by the WAP Forum is a first solution for allowing mobile access to the Internet. Despite its limitations, due to both the use of inadequate radio bearers (i.e., circuit-switched traffic channels) and the inefficient translation from HTML to WML (WAP 1 releases), WAP permits the mobile provision of services and contents. WAP can make available to users many information services that will be adequately supported by future-generation packet-switched bearers on the air interface.

BIOGRAPHIES

Alessandro Andreadis is assistant professor at the Department of Information Engineering of Siena University, Italy, since 1998. In 1993, he received the graduate degree in electronic engineering at the University of Florence, Italy. In the same year he won a research grant at the public administration of Regione Toscana, for a two-year research program on broadband networks based on SMDS and DQDB protocols. His work was funded for two further years, toward the development and diffusion of telematic services, via MAN networks, to small and medium enterprises of the territory. He held the courses of "Systems and Technologies for Communications" at the Department of Communication Science (University of Siena) and of "Telecommunication Networks" at the Faculty of Engineering (University of Siena). Here, he is presently teaching the course on transmission and processing of information in multimedia systems. Since 1995, he has been working at various international projects funded by the European Commission, in the Advanced Communication Technologies and Services (ACTS) Information Society Technologies (e IST) programs. His research interests focus on adaptive multimedia applications for mobile environments, traffic modeling, WAP services, TCP/IP on wireless and mobile networks.

Giovanni Giambene received the Dr. Ing. degree in electronics and a Ph.D. degree in telecommunications and informatics from the University of Florence, Italy, in 1993 and in 1997, respectively. From 1994 to 1996 he was technical external secretary of the European Community project COST 227 *Integrated Space/Terrestrial Mobile Networks*. He also contributed to the resource management activity of the Working Group 3000 within the RACE Project called *Satellite Integration in the Future Mobile Network* (SAINT, RACE 2117). From 1997 to 1998 he was with OTE of the Marconi Group, Florence, Italy, where he was involved in a GSM development program. In the same period he also contributed to the COST 252 Project (*Evolution of Satellite Personal Communications from Second to Future Generation Systems*) research activities. Since 1999, he has been a research associate at the Information Engineering Department, University of Siena, Italy, where he is involved in the activities of the *Personalised Access to Local Information and services for tOurists* (PALIO) IST Project within the fifth Research Framework of the European Commission. His research interests include third-generation mobile communication systems, medium access control protocols, traffic scheduling algorithms, and queuing theory.

BIBLIOGRAPHY

1. F. Harvey, The Internet in your hand, *Sci. Am.* (Oct. 2000).
2. K. J. Bannan, The promise and perils of WAP, *Sci. Am.* (Oct. 2000).
3. WAP Forum Website with address: *http://www.wapforum.org/*.
4. B. Hu, Wireless portal technology — an overview and perspective, *Proc. 1st First On Line Symp. Electrical Engineers*, Oct. 2000.
5. WAP Forum, *Wireless Application Protocol*, *WAP 2.0 Technical White Paper*, available at the WAP forum site, *http://www.wapforum.org/*, Aug. 2001.
6. Example of WAP gateway characteristics: *Nokia WAP Gateway*, available at the address (date of access: Dec. 2001), *http://www.nokia.com/corporate/wap/gateway.html*.
7. Wireless Application Protocol Forum, Ltd., WAP-154, *Binary XML Content Format Specification*, version 1.2, Nov. 4, 1999.
8. S. Lee and N.-O. Song, Experimental WAP (Wireless Application Protocol) traffic modeling on CDMA based mobile wireless

network, *Proc. 54th Vehicular Technology Conf., 2001, VTC 2001*, 2001, pp. 2206–2210.

9. D. Ralph and H. Aghvami, Wireless application Protocol overview, *Wireless Commun. Mobile Comput. J.* **1**(2): 125–140 (April–June 2001).
10. Wireless Application Protocol Forum Ltd., WAP-158, *Wireless Datagram Protocol Specification*, Nov. 5, 1999.
11. A. Andreadis, G. Benelli, G. Giambene, and B. Marzucchi, Analysis of the WAP Protocol over SMS in GSM networks, *Wireless Commun. Mobile Comput. J.* **1**(4): 381–395 (Oct.–Dec. 2001).
12. WAP browsers:
Nokia, *http://www.nokia.com*
Ericsson, *http://www.ericsson.se/WAP*
UP.Browser from Phone.com, *http://updev.phone.com*
WinWAP, *http://www.slobtrot.com*
Motorola, *http://www.motorola.com*
Gelon.net, *http://www.gelon.net*
WAPman from Palm, *http://palmsoftware.tucows.com*.
13. 3GPP, *Technical Specification Group Terminals*; *Mobile Station Application Execution Environment* (MExE); *Functional Description; Stage 2*. (3G TS 23.057).

WIRELESS ATM

Nikos Passas
Lazaros Merakos
University of Athens
Panepistimiopolis, Athens, Greece

1. INTRODUCTION

Broadband and mobile communications are presently the two major drivers in the telecommunications industry. *Asynchronous transfer mode* (ATM) is considered the most suitable transport technique for the Broadband Integrated Services Digital Network (BISDN), because of its ability to flexibly support a wide range of services with quality-of-service (QoS) guarantees. These services are categorized in five classes according to their traffic generation rate pattern: constant bit rate (CBR), real-time variable bit rate (RTVBR), non-real-time variable bit rate (NRTVBR), available bit rate (ABR), and unspecified bit rate (UBR). On the other hand, wireless communications are enjoying a large growth in the last decade. Wireless local-area networks (LANs) in particular are becoming popular for indoor data communications because of their tetherless feature and increasing transmission speed. The combination of wireless communications and ATM, referred to as *wireless ATM*, aims at providing freedom of mobility with service advantages and QoS guarantees.

Wireless ATM is mainly considered for wireless access to a fixed ATM network; in this sense, it is applicable mostly to wireless LANs. A typical wireless ATM network (Fig. 1) includes the following main components:

- *Mobile terminals* (MTs), the end user equipment, which are basically ATM terminals with a radio adapter card for the air interface
- *Access points* (APs), the base stations of the cellular environment, which the MTs access to connect to the rest of the network
- An *ATM switch* (SW) to support interconnection with the rest of the ATM network
- A *control station* (CS), attached to the ATM switch, containing mobility specific software, to support mobility related operations, such as handover,[1] which are not supported by the ATM switch

In many proposals, the CS is considered integrated with the ATM switch in one network module, referred to as *switch workstation* (SWS). Even though this is the most common architecture, other schemes are possible. For example, APs could be equipped with switching and buffering capabilities, as proposed by Veeraraghavan et al. [1]. This, in principle, could expedite mobility and call control operations, but could also increase the overall cost of the system significantly, since the APs need to be more complicated, implementing the full signaling ATM stack.

The main challenge for wireless ATM is to harmonize the development of broadband wireless systems with BISDN/ATM, and offer similar advanced multimedia, multiservice features for the support of time-sensitive voice communications, LAN data traffic, video, and desktop multimedia applications to the wireless user. A sensible quality degradation is unavoidable, due to the reduced bandwidth of the wireless channel and the presentation capabilities of the MTs, but the network should be able to guarantee a minimum acceptable quality. Toward this direction, there are several problems to be faced, mainly because of the incompatibilities of the ATM protocol and the wireless channel:

1. ATM was originally designed for reliable, point-to-point optical fiber links. On the contrary, the wireless channel is a multiple access channel that suffers from high, time-varying, bit error rates, mainly due to fading and interference. This leads to the need for advanced multiple access control and error control mechanisms, for the efficient and reliable sharing of the scarce available bandwidth of the wireless channel, among different kinds of connections.
2. ATM was also designed for large bandwidth environments, following a bandwidth consuming policy to attain simplicity and fast switching of data packets. This leads to a packet header (ATM cell header, in the ATM terminology), which consumes approximately 10% of the available bandwidth (5 of 53 bytes). For gigabit-per-second (Gbps) optical fibers used in BISDN, this is not considered a drawback, compared to fast switching and packet delivery. But for a wireless channel of tens of megabits per second (Mbps), this can be vital for the overall performance. As shown later in this article, the usual practice is to perform header compression to reduce overhead as much as possible.

[1] Mobility issues will be explained in detail later in this article.

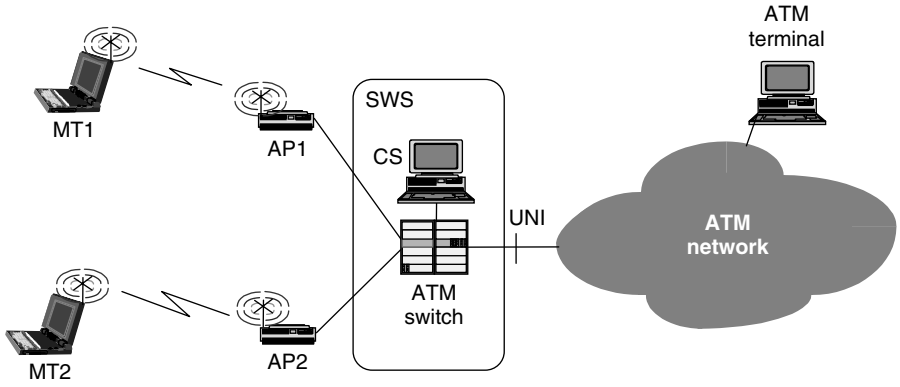


Figure 1. A typical wireless ATM network.

3. ATM signaling enhancements are definitely an important subject for wireless ATM, mainly for mobility. To support it, several additional functions and signaling need to be added in traditional ATM, for registration, location update, handover, and other applications. Particularly for handover, the comparatively high transmission speed, combined with the requirements of some real-time applications (e.g., videoconference), ask for fast and efficient handover techniques.

All these mobility-related functions are usually implemented in the CS shown in Fig. 1 to leave the conventional ATM switches intact. Mobility issues are discussed in detail later in this article. Additionally, some standard call control procedures of fixed ATM need also to be enhanced to cover the particularities of the wireless channel. Especially connection setup requires advanced call admission control algorithms that consider the instabilities of the wireless channel.

The rest of the article is organized in two main sections. Section 2 describes the basic issues and solutions for the medium access control, concluding with the most important standards. Some important protocols are discussed, and their effectiveness in servicing ATM traffic is analyzed. In Section 3, the required signaling enhancements for call and mobility control are discussed. The section starts with a basic signaling architecture, and continues with connection setup (especially call admission control), and handover. Finally, Section 4 contains our conclusions.

2. MEDIUM ACCESS CONTROL (MAC)

2.1. MAC Protocol Structure

In wireless ATM networks, an advanced MAC protocol is required, able to provide adequate support to the traffic classes defined by ATM standards, together with efficient use of the scarce radio bandwidth. Additionally, this protocol should be adaptive to frequent variations of channel quality.

MAC protocols can be grouped, in general, into five classes [2]: (1) fixed assignment, (2) random access, (3) centrally controlled demand assignment, (4) demand assignment with distributed control, and (5) adaptive strategies. *Fixed-assignment techniques* permanently reserve one constant capacity subchannel for each connection for its whole duration and they perform very well with constant bit rate connections in terms of both service quality and channel efficiency. However, their performance decreases dramatically when they are asked to support many infrequent users with variable-rate connections. In such cases, *random-access protocols* usually perform better. A typical example of such a protocol is Aloha, which permits users to transmit at will; whenever a collision occurs, collided packets are retransmitted after some random delay. It is well known that, although ALOHA-type protocols are easy to implement and attain minimum delays under light load, they suffer from long delays and instability under heavy traffic load. Enhancements of ALOHA include collision resolution techniques that increase the maximum achievable stable throughput. *Centrally controlled demand* assignment protocols reserve a variable portion of bandwidth for each connection, adjustable to its needs. Unlike random-access techniques, these protocols operate in two phases: reservation and transmission. In the reservation phase, the user requests from the system the portion of bandwidth required for its transmission needs, and the system responds by reserving the bandwidth and informing the user, while in the second phase the actual transmission takes place. *Demand assignment* protocols are usually complex, but are also stable and perform well under a wide range of conditions, although the reservation phase results in time and bandwidth consumption. With distributed control, the users themselves schedule their transmissions, based on broadcast information. Finally, *adaptive schemes* combine elements from techniques 1–4, and aim at supporting many different types of traffic [3].

Concerning the multiple access technique, the proposed protocols for the radio interface of wireless ATM networks are in general based on frequency-division multiple access (FDMA), code-division multiple access (CDMA), or time-division multiple access (TDMA), or combinations of these techniques. The scarcity of available frequencies, and the requirement for dynamic bandwidth allocation, especially for VBR connections render the use of FDMA inefficient. On the other hand, CDMA limits the peak bit rate of a connection to a relatively low value, which is a problem for broadband applications (>2 Mbps). Accordingly, most of the proposed protocols use an adaptive TDMA scheme,

due to its ability to flexibly accommodate a connection's bit rate needs, by allocating a variable number of time slots, depending on current traffic conditions.

Beyond this general choice of a TDMA-based scheme, the MAC protocols proposed in the literature differ in the technique used to build the required adaptivity in the TDMA scheme. The three main techniques used, alone or in combinations, are *contention*, *reservation*, and *polling*.

Contention-based random-access protocols are simple and require minimal scheduling. An example is the slotted ALOHA with exponential backoff protocol presented by Porter and Hopper [4]. Functionality that can be omitted from the MAC layer, such as handover and wireless call admission control, is pushed to the upper layers. These protocols, attain good delay performance under light traffic, and fit well with the statistical multiplexing philosophy of ATM. Nevertheless, their performance is questionable under heavy traffic conditions, or when multiple traffic classes must be supported with guaranteed QoS.

Another group of protocols uses reservation techniques, mainly through reservation/allocation cycles, to dynamically allocate the available bandwidth to connections, based on their current needs and traffic load. A well-designed representative protocol of this group can be found in the article by Raychaudhuri et al. [5]. It is a TDMA time-division duplex (TDD) protocol, where time is divided in constant length frames and every frame is subdivided in a request subframe and a data subframe. The request subframe is accessed by MTs, through a simple slotted-ALOHA protocol, in order to declare their transmission needs, while the data subframe is used for user data transmission. The allocation of data slots is performed by the AP, based on a scheduling algorithm, and the MTs are informed through broadcast messages. These protocols are more complex and introduce some extra delays, due to the required reservation phase; on the other hand, they are stable under a wide range of traffic loads and can guarantee a predictable quality of service, which is very important in wireless ATM networks. Their performance depends to a large extent on the scheduling mechanism used for the allocation of the available bandwidth. A number of scheduling algorithms has been proposed in the literature, which try to separate real-time and non-real-time connections. For example, a minimum bandwidth can be allocated to non-real-time connections, while real-time connections are served as soon as possible. A delay-oriented scheduling algorithm, referred to as prioritized regulated allocation delay oriented scheduling (PRADOS), has been proposed to meet the requirements of the various traffic classes defined by the ATM architecture [6]. In order for PRADOS to maximize the fraction of ATM cells that are transmitted before their deadlines, each ATM cell is initially scheduled for transmission as close to its deadline as possible. After that, a packetization process ensures that no time slots will be left empty.

A third group of protocols uses adaptive polling to distribute bandwidth among connections [e.g., 7]. A slot is given periodically to each connection, without request, based on its expected traffic. Compared to reservation-based protocols, these protocols are simpler, since there is no reservation phase, but their performance depends on the algorithm that determines the polling period for each connection. If the polling period is shorter than needed, then such protocols might suffer from low utilization, since many slots will be empty. On the other hand, if the polling period is longer than needed, they result in increased delays and poor QoS. The problem becomes more difficult for variable-bit-rate bursty connections. Several proposals suggest an adaptive algorithm to decide on the polling period of each connection, based on total traffic load, expected traffic for each connection, and required QoS [7].

Finally, to improve performance, a combination of the abovementioned schemes is possible; for example, a protocol that is based mainly on reservation, but has also a random-access part for urgent traffic. A typical representative of this category is mobile access scheme based on contention and reservation for ATM (MASCARA) [8]. The multiple access technique used in MASCARA for uplink (from the MTs to the AP of their cell) and downlink (from the AP to its MTs) is based on TDMA/TDD, where a time slot is equal to the time required to transmit an ATM cell. The MASCARA time frame is divided into a DOWN period for downlink data traffic, an UP period for uplink data traffic, and an uplink CONTENTION period used for MASCARA control information. Each of the three periods has a variable length, depending on the traffic to be carried on the wireless channel. The AP schedules the transmission of its uplink and downlink traffic and allocates bandwidth dynamically, based on traffic characteristics and QoS requirements, as well as the current bandwidth needs of all connections. The current needs of an uplink connection from a specific MT are sent to the AP through MT "reservation requests," which are either piggybacked in the data MPDUs (mobile power distribution units), where the MT sends in the UP period, or contained in special "control MPDUs" sent for that purpose in the CONTENTION period. Protocols belonging to the same category can be found in the literature [5,9].

To minimize overhead added by the ATM header, header compression techniques can be used. A straightforward solution is the replacement of the 3-byte-long VPI/VCI (virtual path identifier/virtual channel identifier), used for addressing in ATM, with a shorter MAC specific identifier (MAC_ID), whose length is at most 1 byte, depending on the environment. The MAC_ID is used only for wireless channel transmission, and after this it is replaced with the original VPI/VCI.

2.2. Error Control

In wireless ATM, fulfilling the strict QoS requirements of ATM over an unreliable wireless channel is a challenging problem, and error control is very important. The error control mechanisms used can be thought of as belonging to a sublayer of the MAC layer (usually the upper part), referred to as *wireless data-link control* (WDLC) sublayer. WDLC is responsible for recovering from occasional quality degradations of the wireless channel, and for providing an interface to the ATM layer in terms of frame format and required QoS.

Error control techniques, in general, can be divided in two main categories: *automatic repeat request* (ARQ)

and *forward error correction* (FEC). In ARQ techniques, the receiver detects the erroneously received data and requests retransmission from the transmitter. Since retransmissions imply increased delays, ARQ is efficient for non-real-time data. ARQ techniques are conceptually simple and provide high system reliability at the expense of some extra delay and bandwidth consumption due to retransmissions. FEC, on the other hand, is efficient for real-time data. A number of bits is added in every transmitted data unit, using a predetermined error-correction code, which allows the receiver to detect and correct errors up to a predetermined number per data unit, without requesting any additional information from the transmitter. It is clear that FEC techniques are fast at the expense of lower bandwidth utilization because of the transmission of additional bits.

In wireless ATM, where both real-time and non-real-time data must be supported, a hybrid scheme combining ARQ and FEC is usually used. According to this, for real-time connections (e.g., CBR, RTVBR) FEC bits are included in the header of every MAC data unit, to allow the receiver (AP or MT) to correct most of the errors. For non-real-time connections (e.g., NRTVBR), no extra bits are included, and the AP (MT in the downlink) requests from the MT (AP in the downlink) the retransmission of erroneously transmitted MAC data units.

2.3. MAC Standards

Currently, the MAC technology for wireless ATM is served mainly by two standards, both based on TDMA. The 802.11 standard [10], developed by the IEEE 802 LAN standards organization, and the high-performance radio LAN type 2 (HIPERLAN/2) [11], defined by the European Telecommunications Standards Institute (ETSI) RES-10 Group. Although both standards were designed mainly for conventional LAN traffic, they can definitely serve as a medium for passing ATM traffic, with the proper QoS guarantees. Here we focus more on HIPERLAN/2 because it provides more flexibility for ATM traffic. IEEE 802.11 operates at 2.4 GHz and considers data traffic up to 2 Mbps. The medium can alternate between a contention mode, known as the *contention period* (CP), and a contention-free mode, based on polling, known as the *contention-free period* (CFP). IEEE 802.11 supports three different kinds of frames: management, control, and data. A management frame is used for MT association/deassociation, timing, synchronization, and authentication/deauthentication. A control frame is used for handshaking and positive acknowledgments during a CP, and to end a CFP. Finally, a data frame is used for transmission of data during a CP or CFP. On the horizon there is the need for higher data rates, for applications requiring wireless connectivity at 10 Mbps and higher. This will allow 802.11 to match the data rates of most wired LANs. There is no current definition of the characteristics for the higher data rate signal. However, for many of the options available to achieve it, there is a clear upgrade path for maintaining interoperability with 2-Mbps systems, while providing higher data rates as well.

HIPERLAN/2 systems, on the other hand, operate at the 5.2 GHz unlicensed band and attain transmission rates ranging from 6 to 54 Mbps (a typical value is 25 Mbps). In that sense, it serves better the desired transmission speed for ATM applications. The MAC protocol of HIPERLAN/2 is based on a TDMA/TDD scheme. Time is divided in MAC frames, which are further divided into time slots. Time slots are allocated to the connections dynamically and adaptively depending on the current needs of each connection. Slot allocation is performed by a MAC scheduler that takes into account QoS requirements of each connection. A MAC scheduling algorithm has not yet been specified by the HIPERLAN/2 standards. An efficient algorithm that will be able to meet the requirements of different connections should be developed. The duration of each MAC frame is fixed to 2 ms. Each frame comprises transport channels for broadcast control, frame control, access control, downlink and uplink data transmission, and random access. All data between the AP and the MTs are transmitted in the dedicated time slots, except for the random access channel where contention for the same time slot is allowed. The length of the broadcast control field is fixed, while the length of the other field may vary according to the current traffic needs.

HIPERLAN/2 error control entity supports three different modes of operation: acknowledged mode, repetition mode, and unacknowledged mode. *Acknowledged mode* provides for reliable transmissions using retransmissions to compensate for the poor link quality. The retransmissions are based on acknowledgments from the receiver. The ARQ protocol that is used is selective-repeat (SR) allowing various transmission window sizes to be used depending on the requirements of each connection. In order to support QoS for delay critical applications (e.g., voice, real-time video), error control may also utilize a discard mechanism for discarding data units that have exceeded their lifetime. *Repetition mode* provides for reliable transmission by repeating data units. In repetition mode, the transmitter transmits new data units consecutively, and is allowed to make arbitrary repetitions of each data unit. No feedback is provided by the receiver. Finally, *unacknowledged mode* provides for unreliable, low-latency transmissions. In unacknowledged mode, data flow only from the transmitter to the receiver. No ARQ retransmission control or discard messages are supported.

From the above short description of the two standards, it is clear that HIPERLAN/2 provides more alternatives to better satisfy the requirements of different ATM connections. Nevertheless, we should note that, mainly as a result of increased complexity, HIPERLAN/2 products are not yet available in the market, while there is a wide range of 802.11 equipment from a number of vendors.

3. SIGNALING ENHANCEMENTS

Terminal mobility in wireless ATM requires a number of additional operations not supported in fixed ATM networks. These operations include the following:

Registration/Deregistration. When a MT is switched on, it needs to inform the network and be accepted by it to be able to send and receive calls. This

operation is called *registration*. An important part of registration is authentication, where the MT is recognized as authentic and it is permitted to continue registering. The operation opposite to registration, when the MT is switched off, is called *deregistration*, and informs the network that the MT is no longer available.

Location Update. When a MT has no active connections, it is practically untraceable by the network. So a passive operation is required, in which the system periodically records the current location of the MT in some database that it maintains, in order to be able to forward an incoming connection, when a new connection setup request arrives.

Handover. (Also referred to as "handoff.") It is the operation that allows a connection in progress to continue as the MT changes channels in the same cell, or moves between cells. In a multichannel system, handovers within the same cell, where the connections are transferred to new radio channels, are referred to as *intracell handovers*. The case where the MT connections are transferred to an adjacent cell is referred to as an *intercell handover*. One of the key issues in wireless ATM is maintaining the QoS of different connections during a handover.

Connection Setup. Standard protocols for connection setup in fixed ATM networks assume that the terminal's address implicitly identifies its attachment point to the network. However, this is not the case in wireless ATM. Thus, the ATM connection setup protocols must be augmented to dynamically resolve a MT endpoint location. Additionally, connection admission control (CAC), as part of the connection setup process, is much more difficult in wireless ATM. This is because the wireless channel quality varies in time, due to temporary interference or fading, so the available resources are not fixed. A proposal for wireless ATM CAC can be found in Yu and Leung [12].

Registration/deregistration and location update solutions are more or less generic and do not have extra requirements in a wireless ATM environment. Below we elaborate on connection setup and handover, and analyze their requirements and constraints, starting with the signaling architecture.

3.1. Signaling Architecture

Current trends in designing the access network (AN) part of fixed BISDN aim at concentrating the traffic of a number of different user–network interfaces (UNIs) and routing this traffic to the appropriate service node (SN) through a broadband V interface (referred to as VB). The main objective in AN design is to provide cost-effective implementations without degrading the agreed QoS, while achieving high utilization of network resources. This is reflected in both the reduction of the AN physical equipment and in the limitations imposed on the AN functionality, such as the inability to interpret the full ATM layer control information and signaling. The use of only low-level operations in AN forces the establishment of several internal mechanisms that are used to unambiguously identify the connection an ATM packet belongs to, and to convey only those connection parameters that are absolutely necessary for traffic handling.

In this framework, a fast control protocol running over a universal VB interface can be introduced [13], which serves a number of AN internal functions while preserving the highest possible degree of transparency at the SN. The protocol is based on the local exchange access network interaction protocol (LAIP), which was developed to accommodate the SN-to-AN communication requirements, as identified in the early study and design of the dynamic $VB_{5.2}$ interface, namely, the interface between the fixed ATM AN and the SN. In the relevant standardization bodies, the presence of such a protocol has been firmly decided and has been given the name Broadband Bearer Channel Control Protocol (B-BCCP). The services of the $VB_{5.2}$ control protocol enable the dynamic AN operation by conveying the necessary connection-related parameters required for dynamic resource allocation, traffic policing, and routing in the AN, as well as information on the status of the AN before a new connection is accepted by the SN.

The signaling access architecture for wireless ATM considered here is an extension of the broadband V interface, where an enhanced version of the $VB_{5.2}$ control protocol is used to enable the dynamic operation of the AN and to serve the AN internal functions. It is assumed that a mobility-enhanced version of the existing B-ISDN UNI call control (CC) signaling is employed to provide the basic call control function and to support the handover-related functions. In addition, pure ATM signaling access techniques, based on metasignaling, are adopted for the unique identification and control of signaling channels. These features allow us to minimize the changes required to the signaling infrastructure used in the wired network, and, in this respect, they can guarantee the integration of the wireless ATM access system with fixed B-ISDN. However, when striving for full integration, the mobile-specific requirements imposed by the radio access part need to be taken into account.

In today's wired ATM environment, the user–network interface is a fixed port that remains stationary throughout the lifetime of a connection. The current B-ISDN UNI protocol stack uses a single protocol over fixed point-to-point or point-to-multipoint interfaces. On the other hand, in wireless ATM, mobility causes the user access point to the wired network to change constantly, and the mobile terminal connections must be transferred from access point to access point, through a handover process. The support of the handover functionality assumes that the fixed network of the access part has the capability to dynamically set-up and release bearer connections during the call. A well-accepted methodology to support these features is the call and bearer separation at the UNI. The use of the extended $VB_{5.2}$ interface control protocol for wireless ATM access systems serves for the setup and reconfiguration of fixed bearer connections of the same call, supporting in this way the call and bearer control separation in the AN part.

On the basis of the terminology described above, the following types of signaling interaction for the communication of peer entities can be identified [14]:

- *Mobile Call Control Signaling* (*MCCS*). This includes an enhanced B-ISDN call control signaling protocol (denoted as Q.2931*), based on the ITU (International Telecommunication Union) recommendation Q.2931, for the setup, modification, and release of calls between the MT and the CS. The enhancements required in the current signaling standards are related to the support of the handover function (e.g., inclusion of handover-specific messages).
- *Mobility Management Signaling* (*MMS*). This is responsible for the MT registration/authentication and tracking procedures.
- *Bearer Channel Control Signaling* (*BCCS*). This serves for providing the traffic parameters to the AP, and handles the establishment, modification/reconfiguration, and release of fixed ATM connections between the AP and the CS.
- *Radio Channel Control Signaling (RCCS)*. This deals with low-level signaling related to the radio interface consisting of messages between the MT and the AP (MAC and physical layer specific messages).

At the user plane, the MT has a typical ATM protocol stack on top of a radio-specific physical layer and a MAC layer. The AP acts as a simple interworking unit that extracts the encapsulated ATM cells from the MAC frame, and forwards them to the CS through a proper ATM virtual connection. The MAC functionality realized at the AP is based on a MAC scheduler, which, on the basis of the ATM connection characteristics declared at connection setup and current transmission requests, allocates the radio bandwidth according to the declared QoS requirements and service type of each connection. As already mentioned, such a mechanism provides a degree of transparency to a subset of broadband/ATM services, and achieves efficient sharing of the scarce radio bandwidth among the mobile users. The CS realizes the typical B-ISDN protocol functionality of the U plane.

3.2. Connection Setup

Connection setup procedures used in traditional ATM networks assume (1) reliable gigabit links with fixed capacity and (2) stationary users. Accordingly, CAC algorithms do not need to be constantly informed about the available resources and the users' attachment points. But this is not the case in wireless ATM. The wireless channel impairments and MAC layer overheads can result in lower bandwidth than the theoretically available, while a mobile users' attachment point with the network can change anytime. Below we describe typical connection setup scenarios in wireless ATM, focusing on the differences with fixed ATM.

When a MT initiates a new call, its signaling channel transparently conveys a standard connection SETUP_REQUEST signaling message to the CS. Upon receipt of this request, the CS identifies the calling MT and the called terminal, and contacts the location server to track the location of the calling MT and the called terminal (if it is mobile). An initial connection acceptance decision is made, based on the user service profile data and on the QoS requirements set by the MT.

In case the request is accepted, the AP of the calling MT should be notified by the CS (using BCCS) on the expected new traffic so that it can decide on the admission in the wireless channel, and allocate radio resources accordingly. To this end, the traffic parameters of the new connection, or at least a useful subset of them, should be communicated to the AP of the calling MT. This information makes it possible to exercise a policing functionality at the AP, implemented implicitly by its radio bandwidth allocator. It also protects the CS from the unlikely case where, although the CS expects availability of radio resources, these are exhausted due to additional overheads of the MAC layer, or a temporary reduction in radio link quality. The latter is useful in case the CAC of the CS does not take into account issues specific to the wireless access. Since the final CAC decision is taken at the CS, it is possible to implement a connection acceptance algorithm customized to the specific wireless access system. Traffic characteristics will appear at the AP together with the QoS requirements, declared as the class of service (CoS) that the specific connection will support. This enables the MAC to implement a set of priorities according to the connection to which an ATM cell belongs. To be able to recognize the particular connection class, it is necessary to declare also the VPI/VCI values that will be used.

The task of the AP-CS communication and bearer channel establishment in the fixed access network is very important in this case. An ALLOC message is generated and forwarded to the AP, through BCCS. The AP will reply with an ALLOC_COMPLETE or an ALLOC_REJECT message indicating whether it agrees with the CAC decision. The latter implies that the call is rejected at the AP. On receipt of an ALLOC_COMPLETE, the CS returns a CALL_PROCEEDING message to the calling MT and initiates the connection establishment procedures toward the core network (B-ISUP IAM message) if the called terminal is a fixed one.

In case the called terminal is another MT (i.e., intra-CS call), the call processing module of the CS forwards the setup request towards the AP of the called terminal, where functions similar to those described above take place. The signal exchanges for this case are shown in Fig. 2. In the fixed-to-MT (incoming) connection setup scenario, the CS receives an incoming SETUP_REQUEST message, identifies the called MT, tracks its location, draws an initial CAC decision, and asks the corresponding AP of the called MT [14].

In all cases, the ALLOC message transfers to the AP all the connection-related information required for the AP operation. This includes the bandwidth requested by the connection, the service class, the QoS parameter values, etc. An improvement, in the case where the requested bandwidth or the QoS cannot be supported by the radio part of the communication path, is for the AP to generate an ALLOC_MODIFY message indicating

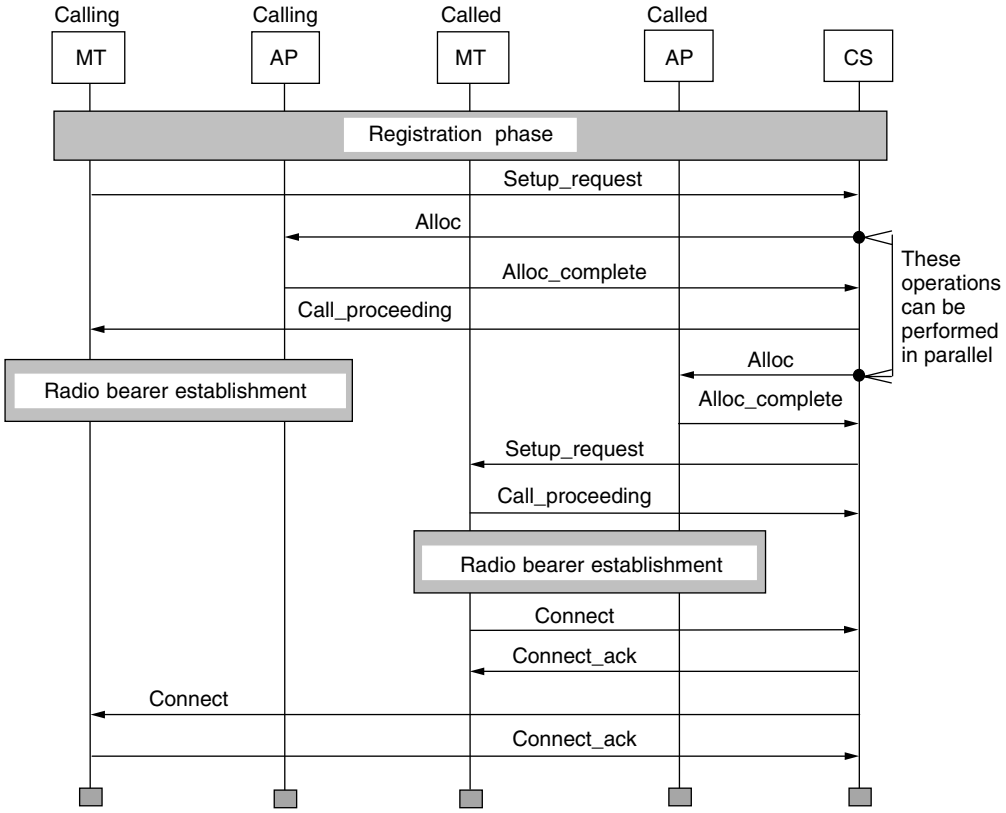


Figure 2. Connection setup procedure between two MTs.

this situation and suggesting a QoS degradation needed for the connection to be accepted. This useful "fallback" mechanism intends to set up connections with the highest available bandwidth. However, such a capability is useless if the standard ATM signaling does not support QoS negotiation to let the CS and the MT negotiate the new situation. In all scenarios, we have implicitly assumed that MTs remain stationary at connection setup. If we assume that a MT may move during connection setup, the setup might not succeed. In this case, the new location of MT is determined and another setup should be attempted following the same procedures. The calling or called party can initiate the release of a call. On receipt of a RELEASE message, the CS releases all the resources associated with that call and triggers the release toward the AP, the core network, or the MT.

3.3. Handover

Among mobile-specific operations, handover is probably the most difficult to perform, due to the diversity of requirements of different kinds of connections, and the constraints imposed by the wireless channel. In any case, an unavoidable period of time is required, during which the end-to-end connection data path is incomplete. This means that some data might get lost or should be buffered for later delivery. The effect that this increase of losses or delays has on each application depends on the nature of the application and the duration of the disruption. Current proposed protocols for handover in wireless ATM may be grouped in four categories [15]:

Full-Connection Rerouting. This is the simplest kind of handover, where the system establishes a completely new end-to-end route for each handover — as if it were a new connection. Clearly, this kind of handover is simple in terms of implementation, but can result in unacceptable delays and losses, depending on the distance between the two parties.

Route Augmentation. In this case, the original connection is extended with an additional hop to the MT's next location. For users with limited mobility, this solution can result in low delays and very limited or no losses, since no actual rerouting is performed. But if the MT begins to change cells more often, the additional extensions will result in a very long connection path, increasing delays and reducing network utilization.

Partial-Connection Rerouting. This kind of handover attempts to perform a more efficient rerouting, by preserving as much of the old connection path as possible and rerouting the rest. The key issue here is to locate the nearest ATM network node that is common to both the old and the new data paths. Then the common node will handle the tearing down of the old part and establishing the new, also taking care of the data that are on the way in the old part, when the switching is performed. Temporary buffering before switching or temporary rerouting after switching can be used to minimize losses. Partial connection rerouting is the most common handover type found in the literature.

Multicast Connection Rerouting. In this kind of handover, more than one connection paths are maintained at a time, although only one path is operational. When the MT moves to a new cell, data can immediately start flowing toward the new direction. This eliminates the need for establishing a new path during handover (partial or full) and leads to lower delays and losses. On the other hand, since the system cannot maintain a path for every cell a MT can move to, an intelligent algorithm is required to predict the MT's movement and preestablish paths on the neighboring cells, while at the same time, paths that are no longer needed are canceled. An extension of this kind of handover could also permit the same data to flow in more than one data path when the MT is at the threshold between two cells (this is also referred to as *macrodiversity*). This allows a MT with multiple receivers (antenna diversity) to get data from more than one AP, and keep only the correctly transmitted information, reducing in this way the bit error rate.

Another categorization in wireless ATM handover is based on who performs what. In general, a handover mechanism involves a continuous procedure of channel measurements, and starts with a handover request initiation. In that sense, there are three fundamentally different categories of handover mechanisms: network-controlled, mobile-assisted, and mobile-controlled. In *network-controlled handover*, the MT is completely passive. All measurements are performed by the network (basically the AP) and the handover request is initiated by the AP. This is a simple solution, which does not perform well in the case where the signal received by the AP is good, while the signal received by the MT is bad. This weakness is overcome by *mobile-assisted handover*, where both the AP and the MT are measuring the strength of the received signal; however, the handover request is initiated by the AP. The MT can only send its measurements to the BS in order for it to have a better picture of the situation. Finally, in *mobile-controlled handover* all measurements and handover requests are executed at the MT. If the handover request is executed via the "old" AP (the AP that the MT is leaving), we have a *backward handover*, and if it is executed via the "new" AP, we have a *forward handover*. Backward handovers are in general more seamless than forward, so the usual practice is for the MT to prefer backward handover, and, only if this is not possible (in case of an abrupt signal strength reduction), to perform forward. Mobile-controlled handover can operate either alone or in conjunction with network-controlled or mobile-assisted handover.

No matter what handover algorithm is used, the main target should be to maintain the QoS of active connections, not only during, but also after handover in the new cell. During handover, temporary buffering can be used at the switching point to ensure delivery of loss-sensitive data. For delay-sensitive data that cannot be buffered, the only solution is to ensure simple and fast handover operation. On the other hand, maintaining QoS in the new cell is not always possible. In fixed ATM, if an efficient CAC algorithm decides that a connection can be accepted with the requested QoS, then the network can guarantee this QoS throughout the duration of the connection. The same cannot be said for a connection to a MT, which can be rerouted when the MT is handed over to a new cell. For example, if this new cell is overcrowded, there might not be enough resources to support the QoS of the connections of the newly arrived MT. In this case, the smallest possible number of the MT's connections should be rejected, to leave enough resources for the rest. This decision is usually taken by the CS, because it has a more global view of the system. A more advanced solution is to renegotiate the QoS in the new cell in order to avoid connection rejection as much as possible. In this case, the MT will be asked to reduce its requirements if it wants to maintain its connections in the new cell. In the following paragraphs we describe a simple but typical mobile-controlled handover procedure.

When the MT decides that a handover should be performed, it sends a HANDOVER_REQUEST message toward the CS, transparently via the old AP. This message contains identification of the MT, the call, and the target AP. The MT may have multiple active connections at the same time, as multimedia applications are to be supported. If this is the case, during the request for handover the MT could also indicate the priorities of the different connections in case the new AP cannot accommodate all of them.

A fast control protocol between AP and CS is required for the release/establishment of the old/new bearers in the fixed network part, and for performing possible QoS renegotiations during handover. On receipt of the HANDOVER_REQUEST, the CS identifies the MT, and initiates a state machine for the handover. Similar procedures to those described for connection setup are performed between the CS and the new AP. In this way, the CS informs the target AP about the expected QoS and bandwidth requirements to allocate radio resources accordingly.

When the CS receives the response from the new AP (ALLOC_COMPLETE), it sends a HANDOVER_RESPONSE message to the MT to inform it about the handover results and possible QoS modifications, and reconfigures the ATM connections of the ATM switch toward the new AP. After receiving the HANDOVER_RESPONSE, the MT releases its radio connection with the old AP, and establishes a radio link with the new AP. Special ATM (and lower) layer cell relay functions take place at the MT and the CS to coordinate the switching of traffic, and to guarantee the transport of user data at an agreed QoS level in terms of cell loss, ordering, and delay. Finally, the CS updates the location server about the new location of the MT, and sends a RELEASE message to the old AP to notify it that the connection no longer exists and to de-allocate the corresponding radio resources.

The handover process described above is expected to be fast. In the unlikely case that a MT moves again before the handover is accomplished, handover is again attempted to the current destination AP, until it eventually succeeds. The forward handover scenario is similar to the backward one. The MT releases the old radio connection and communicates directly with the new AP. Since all signaling

is passed through this new AP, a dynamic signaling channel allocation scheme is employed, in order for the MT to obtain a signaling channel for passing the messages to the CS.

4. CONCLUSIONS

The design of wireless ATM systems to offer ATM services to wireless users has attracted considerable attention during the past few years, and a large number of proposals exist in the literature dealing with specific design issues. The most important of these issues are the medium access control and the signaling enhancements.

Medium access control is much more demanding in wireless ATM than in traditional wireless networks, owing to the, often conflicting, requirements of the various ATM traffic types. The current trend is for flexible, TDMA/TDD protocols with variable time frame, enabled with a sophisticated traffic scheduling algorithm that adjusts the bandwidth given to a connection to its time-varying requirements, without violating the contract made with other active connections.

On the other hand, enhancements are required to standard ATM signaling, to cover issues such as wireless call admission control and handover. Wireless call admission control is part of the overall call admission control process, handling available resources in the wireless link. Since the available wireless bandwidth is time-varying, due to temporary deterioration of the radio signal, the procedure should always have up-to-date information on the current status of the radio link. Finally, handover is a completely new issue for wireless ATM signaling. Handover mechanisms should be fast and efficient, in order to minimize losses or delays, which could influence the QoS provided to the user. The usual practice is to introduce a special-purpose control station, centrally located in the wireless ATM network, which handles all the extra signaling and implements wireless call admission control and handover mechanisms.

As a final comment we can say that the future trends in wireless communications tend to be toward wireless IP-based systems with QoS provision (i.e., IPv6), rather than wireless ATM. Nevertheless, the issues and problems are more or less the same, so techniques and mechanisms developed for wireless ATM can, with proper adjustments, be used in wireless IP as well.

BIOGRAPHIES

Nikos Passas received his B.S. degree in computer engineering in 1992 from the University of Patras, Patras, Greece, and a Ph.D. degree in computer engineering from the University of Athens, Athens, Greece, in 1997. In 1995, he joined the Greek National Research Center "Demokritos" as a network engineer, where he worked on network management. Since 1997, he has been a senior researcher in the Communication Networks Laboratory, where he has been working on wireless networks. His areas of interest are multiple access control, quality of service of wireless networks, and performance analysis of wireless communications.

Lazaros Merakos received the Diploma in electrical and mechanical engineering from the National Technical University of Athens, Greece, in 1978, and his M.S. and Ph.D. degrees in electrical engineering from the State University of New York, Buffalo, in 1981 and 1984, respectively. From 1983 to 1986 he was on the faculty of the Department of Electrical Engineering and Computer Science at the University of Connecticut, Storrs, Connecticut. From 1986 to 1994 he was on the faculty of the Electrical and Computer Engineering Department at Northeastern University, Boston, Massachusetts Between 1993 and 1994 he served as director of the communications at the Digital Signal Processing Research Center at Northeastern University. In 1994, he joined the faculty of the University of Athens, Athens, Greece, where he is presently a professor in the Department of Informatics and Telecommunications, director of the Communication Networks Laboratory and the Networks Operations and Management Center. His research interests are in the design and performance analysis of broadband networks and wireless/mobile networks and services. He is the author of over 120 papers in the above areas. He was the recipient of the Guanella Award for the Best Paper presented at the 1994 International Zurich Seminar on Mobile Communications.

BIBLIOGRAPHY

1. M. Veeraraghavan, M. J. Karol, and K. Y. Eng, Mobility and connection management in a wireless ATM LAN, *IEEE J. Select. Areas Commun.* **15**(1): 50–68 (Jan. 1997).
2. F. A. Tobagi, Multiaccess link control, in P. E. Green, Jr., ed., *Computer Network Architectures and Protocols*, Plenum Press, New York, 1982.
3. E. Ayanoglu, K. Y. Eng, and M. J. Karol, Wireless ATM: Limits, challenges, and proposals, *IEEE Pers. Commun. Mag.* **3**(4): 18–34 (Aug. 1996).
4. J. Porter and A. Hopper, *An ATM-Based Protocol for Wireless LANs*, Olivetti Research Ltd. Technical Report 94.2, April 1994, available at *ftp://ftp.cam-orl.co.uk/pub/docs/ORL/tr.94.2.ps.Z*.
5. D. Raychaudhuri et al., WATMnet: A prototype wireless ATM system for multimedia personal communication, *IEEE J. Select. Areas Commun.* **15**(1): 83–95 (Jan. 1997).
6. N. Passas and L. Merakos, Traffic scheduling in wireless ATM networks, *Proc. IEEE ATM '97 Workshop*, Lisbon, Portugal, May 1997.
7. C.-S. Chang, K.-C. Chen, M.-Y. You, and J.-F. Chang, Guaranteed quality-of-service wireless access to ATM networks, *IEEE J. Select. Areas Commun.* **15**(1): 106–118 (Jan. 1997).
8. F. Bauchot et al., MASCARA: A MAC protocol for wireless ATM, *Proc. Int. Conf. Telecommunications*, Melbourne, Australia, April 1997.
9. F. D. Priscoli, Adaptive parameter computation in a PRMA, TDD based medium access control for ATM wireless networks, *Proc. IEEE GLOBECOM'96*, London, Nov. 1996, Vol. 3, pp. 1779–1783.
10. IEEE 802.11D5, *Wireless LAN Medium Access Control (MAC) and PHysical Layer (PHY) Specifications*, July 1996.

11. ETSI TR 101 683 (V1.1.1), *Broad Radio Access Network (BRAN); HIgh Performance Radio Local Area Networks (HIPERLAN) Type 2; System Overview*, Feb. 2000.

12. O. T. W. Yu and V. C. M. Leung, Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN, *IEEE J. Select. Areas Commun.* **15**(7): 1208–1225 (Sep. 1997).

13. I. S. Venieris et al., Architectural and control aspects of the multi-host ATM subscriber loop, *J. Network Syst. Manage.* **5**: 55–71 (1997).

14. N. Loukas, N. Passas, L. Merakos, and I. Venieris, A signaling architecture for wireless ATM access networks, *ACM Wireless Networks J.* **6**: 145–159 (Nov./Dec. 2000).

15. I. F. Akyildiz et al., Mobility management in current and future communication networks, *IEEE Network* **12**(4): 39–49 (July/Aug. 1998).

WIRELESS COMMUNICATIONS SYSTEM DESIGN

CHINTHA TELLAMBURA
Monash University
Clayton, Victoria, Australia

A. ANNAMALAI
Virginia Tech
Blacksburg, Virginia

1. INTRODUCTION

The wireless communications industry has been experiencing phenomenal annual growth rates exceeding 50% since the late 1990s. This degree of growth reflects the tremendous demand for commercial untethered communications services such as paging, analog and digital cellular telephony, and emerging personal communications services (PCS), including high-speed data, full-motion video, Internet access, on-demand medical imaging, real-time roadmaps, and anytime, anywhere videoconferencing. By 2002 subscriber rates for personal wireless services are expected to reach 70% of the population for industrial nations, and by 2004 these rates are expected to reach 17% of the population worldwide. Of these subscribers, it is expected that by 2005 half will have data-capable handsets, creating an even greater demand for wireless data services. Since wired broadband services such as digital subscriber loop (DSL) and cable modems have been slow to market, this will drive even more customers to wireless alternatives; by 2003 more than 34% of homes and 45% of businesses in the United States will be served by wireless broadband services. The first-generation cellular and cordless telephone networks, which were based on analog technology with frequency modulation, have been successfully deployed throughout the world since the early and mid-1980s. Second-generation (2G) wireless systems employ digital modulation and advanced call processing capabilities. Third-generation (3G) wireless systems will evolve from mature 2G networks, with the aim of providing universal access and global roaming. Introduction of wideband packet data services for wireless Internet up to 2 Mbps (megabits per second) will probably be the main attribute of 3G systems.

To meet this increasing demand, new wireless techniques and architectures must be developed to maximize capacity and quality of service (QoS) without a large penalty in the implementation complexity or cost. This provides many new challenges to system designers, one of which is ensuring the integrity of the data is maintained during transmission. The largest obstacle facing designers of wireless communications systems is the nature of the propagation channel. The wireless channel is nonstationary and typically very noisy as a result of fading and interference. The sources of interference could be natural (e.g., thermal noise in the receiver) or synthetic (human-made; e.g., hostile jammer, overlay communication), while the most common type of fading is caused by multipath effects, in which multiple copies of a signal arrive out of phase at the receiver and destructively interfere with the desired signal. Another problem imposed by multipath effects is delay spread, in which the multiple copies of a signal arriving at different times spread out each data symbol in time. The stretched-out data symbols will interfere with the symbols that follow, causing intersymbol interference. All of these effects can significantly degrade the performance and QoS of a wireless system. Another critical issue in wireless system development is channel capacity. The Shannon channel capacity may be conveniently expressed in terms of the channel characteristics as

$$C = B \log_2(1 + \gamma |H|^2) \tag{1}$$

where γ is the signal-to-noise ratio (SNR), B denotes the channel bandwidth, and $|H|^2$ is the normalized channel power transfer characteristic. The ratio C/B, called *spectral efficiency*, is the information rate per hertz, is directly related to the modulation of a signal. To illustrate this, the analog AMPS cellular telephone system has a spectral efficiency of 0.33 bps/Hz while the digital GSM system has a spectral efficiency of 1.35 bps/Hz and the IS 54 system has 1.6 bps/Hz.

To overcome the problems mentioned above and to increase spectral efficiency, many techniques are employed, including

- The use of a set of signals that fade independently is referred to as diversity combining. Diversity techniques include selective combining, switched combining, maximal ratio combining, and equal gain combining. The effectiveness of diversity combining is limited by the degree of independence of fading within the set of signals. A measure of this can be obtained from calculating correlations between pairs of signals.
- Diversity can also be sought through the use of coding techniques, multiple frequency bands, and multiple antennas.
- "Smart" or directional antennas allow the energy transmitted toward the significant scatterers to be reduced and hence reduce far-out echoes.

- Adaptive filters and equalizers can be used to flatten the channel response for wideband fading channels.
- The delay spread affects high-data-rate systems. The required data can be simultaneously transmitted on a large number of carriers, each with a low data rate, and the total data rate can be high. This concept, which is known as *orthogonal frequency-division multiplexing* (OFDM), is used in digital broadcasting.

The system designers need to assess the efficacy of such techniques to determine the most appropriate choice of complexity and implementation constraints. One may use Monte Carlo simulations or develop an analytic framework for system design. The analytic approach has three advantages over the Monte Carlo approach; it

- Facilitates rapid computation of the system performance
- Provides insight as to how different design parameters affect the overall system performance
- Provides some ability to optimize the design parameters

Nevertheless, analytic solutions are governed by a set of simplifying assumptions needed for analytic tractability, and hence care must be exercised when extrapolating from analytic results to real-world designs. However, as this approach can identify viable design options before further computer simulations are undertaken, it can be the first step of the design process of communication systems.

1.1. Fading Channels

A generic communication system is shown in Fig. 1. The information from the source is converted into a signal suitable for sending by the transmitter and is then sent over the channel. The channel is a description of how the communications medium alters the signal that is being transmitted. Finally the receiver takes the signals that have been altered by the channel, and attempts to recover the information that was sent by the source. The estimate of this information is passed to the sink as the received information. The channel modifies the signal in ways that may be unpredictable to the receiver, so the receiver must be designed on the basis of statistical principles to estimate the information and to deliver the information to the receiver with as few errors as possible. In our case, the channel is the wireless channel that includes all the antenna and propagation effects within it.

We now briefly describe the statistical models of fading multipath channels, which are frequently used in the analysis and design of wireless communications systems.

1.2. Fading Channel Characterization

In the most general setting, a fading multipath channel is characterized as a linear, time-varying system having an (equivalent lowpass) impulse response $h(t, \tau)$ (or a time-varying frequency response) $H(t, f)$, which is a wide-sense stationary random process. Time variations in $h(t, \tau)$ or $H(t, f)$ result in frequency spreading, which is known as *Doppler spreading*, of the signal transmitted through the channel. Multipath propagation results in spreading the transmitted signal in time. Consequently, a fading multipath channel may be generally characterized as a doubly spread channel in time and frequency.

The channel output y at time t can be found from the convolution of the input signal $x(t)$ with the impulse response $h(t, \tau)$ (also known as the *input delay spread function*) of the channel at time t. We then have

$$y(t) = \int_{-\infty}^{\infty} h(t, \tau)x(t - \tau)\, d\tau \tag{2}$$

where τ is the delay variable.

Assuming that the multipath signals propagating through the channel at different delays are uncorrelated, we can characterize a doubly spread channel by the delay Doppler spread function, which is obtained by transforming $h(t, \tau)$ with respect to time. The scattering function $S(\tau, \nu)$ is a measure of the power spectrum of the channel at delay τ and frequency offset ν (relative to the carrier frequency). From the scattering function, we obtain the delay power spectrum of the channel (also called the *multipath intensity profile*) by simply averaging over

$$S_c(\tau) = \int_{-\infty}^{\infty} S(\tau, \nu)\, d\nu \tag{3}$$

This spectrum expresses the average power received for a transmitted pulse as a function of time delay, τ. The range of values over which the delay power spectrum $S_c(\tau)$ is nonzero is defined as the multipath spread of the channel T_m. Similarly, the Doppler power spectrum is

$$S_c(\nu) = \int_{-\infty}^{\infty} S(\tau, \nu)\, d\tau \tag{4}$$

This spectrum expresses the average power received for a transmitted pulse as a function of frequency offset, ν.

1.2.1. Doppler Spectrum. When a mobile moves at a certain velocity, as pathlengths between transmitter and receiver change, the Doppler effect results in a change of the apparent frequency of the arriving wave. The amount of this change is known as the Doppler shift. The maximum Doppler shift f_d is given by

$$f_d = \frac{vf_c}{c}$$

Figure 1. Generic communication model.

where v is the velocity of the mobile, f_c is the communication frequency, and c is the velocity of propagation of light.

Example 1. A mobile system operates at 900 MHz. What is the maximum Doppler shift observed by a mobile traveling at 80 km/h?

The maximum Doppler shift is

$$f_d = f_c \frac{v}{c} = 900 \times 10^6 \times \frac{80 \times 10^3}{60 \times 60 \times 3 \times 10^8} = 67 \text{ Hz}$$

With multipath propagation, the copies of a signal arrive from several directions and each copy has its own Doppler frequency. Thus, the exact shape of the resulting spectrum $S_c(\nu)$ depends on the relative amplitudes and directions of each of the incoming signals. The range of values over which the $S_c(\nu)$ is nonzero is defined as the Doppler spread f_d of the channel. The exact expression for $S_c(\nu)$ cannot be obtained without making some assumptions of the arrival angle of the multipath signals. Most commonly, the arriving multipath signals at the mobile are assumed to be equally likely to come from any horizontal angle. The classic Doppler spectrum is then given by

$$S_c(\nu) = \frac{1.5}{\pi f_d \sqrt{1 - (\nu/f_d)^2}} \quad \text{for } |\nu| < f_d \tag{5}$$

and $S_c(\nu) = 0$ for $|\nu| \geq f_d$. This function is sharply limited to $\pm f_d$. The width of the Doppler spectrum is known as the *fading bandwidth*. This function (Fig. 2) is used as the basis of many simulators of mobile radio channels. The assumption of uniform angle-of-arrival distribution may not hold over short distances where propagation is dominated by the effect of a particular local scatterers, but this is a good reference model for the long-term average Doppler spectrum.

The Doppler power spectrum cannot be measured accurately in practice. The level crossing rate (LCR) and the average fade duration (AFD) are directly related to a given Doppler spectrum. These parameters are easier to measure. The LCR is the number of positive-going crossings of a reference level r in unit time and the AFD is the average time between negative and positive level crossings. For the classic Doppler spectrum, the LCR is given by

$$N_r = \sqrt{2\pi} f_d \bar{r} e^{-\bar{r}^2} \tag{6}$$

where $\bar{r} = r/r_{\text{rms}}$. The average fade duration for a signal level of $\bar{r}$ is given by

$$\tau_{\bar{r}} = \frac{e^{-\bar{r}^2} - 1}{\sqrt{2\pi} f_d \bar{r}} \tag{7}$$

These two parameters are plotted in Figs. 3 and 4. Note that the signal spends most of its time crossing signal levels just below r_{rms}, and that fades below this level have short duration.

1.2.2. Signal Correlation. The Doppler spread provides a measure of how rapidly the channel impulse response varies in time. The inverse Fourier transform of the Doppler power spectrum (5) is the autocorrelation function (ACF), which expresses the correlation between a signal at time t and $t + \tau$. For a classic spectrum (5) with Rayleigh fading, the correlation function is

$$\rho(\tau) = J_0(2\pi f_d \tau) \tag{8}$$

where $J_0(x)$ is the Bessel function of the first kind and zeroth order. This is plotted in Fig. 5. For large f_d, the correlation can decrease rapidly. To express this temporal relationship, the channel coherence time T_c for a channel is defined as the time over which the channel can be assumed constant. This is assured if the ACF remains close to unity for this duration. The coherence time is therefore inversely proportional to the Doppler spread of the channel:

$$T_c \propto \frac{1}{f_d} \tag{9}$$

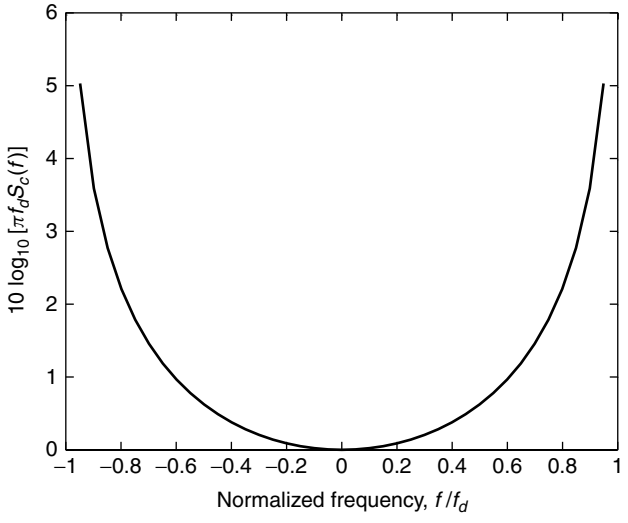


Figure 2. The classic Doppler spectrum.

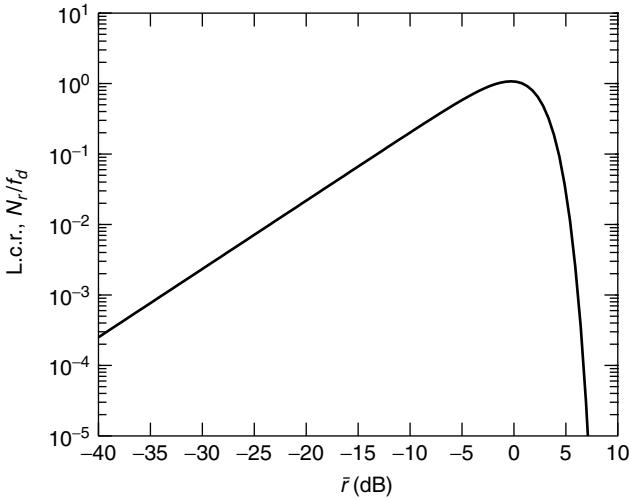


Figure 3. The normalized LCR for the classic Doppler spectrum.

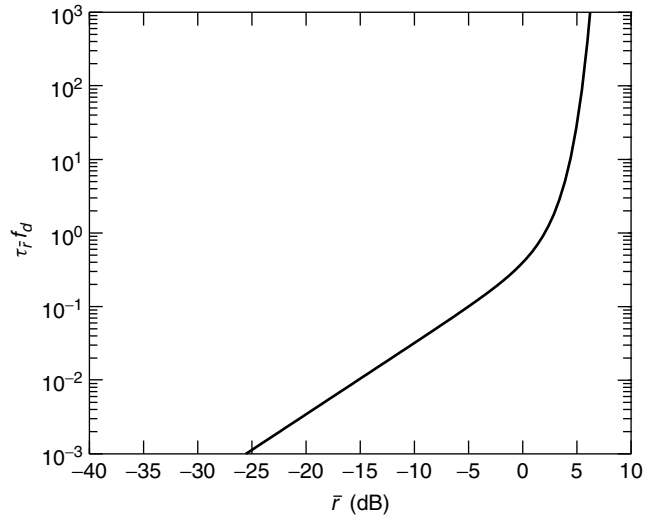


Figure 4. The AFD for the classic Doppler spectrum.

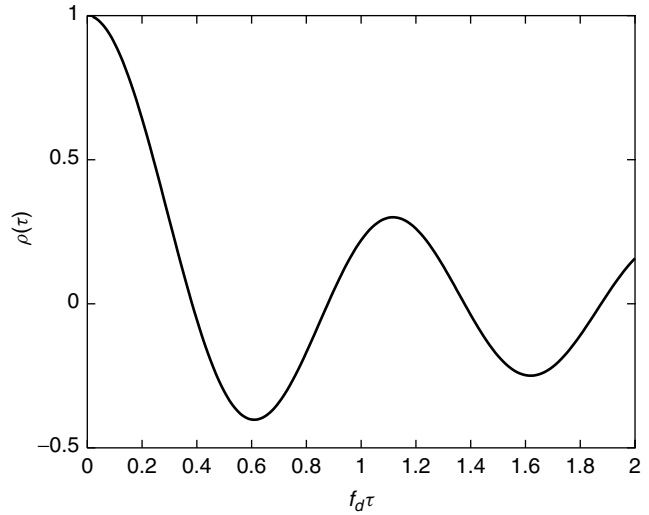


Figure 5. ACF for the classic Doppler spectrum.

Thus a slowly fading channel has a large coherence time, and a rapidly fading channel has a small coherence time. To determine the proportionality constant, a threshold level of correlation for the complex envelope has to be chosen. A useful approximation to the coherence time for the classic channel is

$$T_c \approx \frac{9}{16\pi f_d} \tag{10}$$

Example 2. A mobile system operates at 900 MHz. The maximum speed of a mobile is 80 km/h. What is the minimum symbol rate to avoid the effects of Doppler spread?

From the previous example, the maximum Doppler shift is 67 Hz. The coherence time is therefore

$$T_c \approx \frac{9}{16\pi \times 67} = 2.7 \text{ ms}$$

So the minimum symbol rate for undistorted symbols is, the reciprocal of this, 500 symbols per second. As most systems have data rates exceeding this, the correlation effect is negligible on most practical systems.

The channel coherence bandwidth is defined as the reciprocal of the multipath spread

$$B_c = \frac{1}{T_m} \tag{11}$$

If the correlation is examined for two signals at the same time, then the frequency separation for which the correlation equals 0.5 is termed the *coherence bandwidth* of the channel. This measures the width of the band of frequencies that are similarly affected by the channel response, that is, the width of the frequency band over which the fading is highly correlated.

The product $T_m f_d$ is called the spread factor of the channel. A spread factor smaller than unity results in an underspread channel. A spread factor greater than unity results in an overspread channel. For severely underspread channels ($T_m f_d \ll 1$), $h(t, \tau)$ can be measured by the use of nondata symbols. These symbols (pilot symbols) are known to the receiver and are regularly spread in time and/or frequency domains. Pilot symbols constitute an overhead. Channel measurements can be used at the receiver to demodulate the received signal and at the transmitter to optimize the transmitted signal.

1.3. Flat Fading

We define the time-varying transfer function of the channel as

$$H(f, t) = \int h(t, \tau) e^{-j2\pi f \tau} \, d\tau \tag{12}$$

Using this, the channel output (2) for a band-limited input signal $x(t)$ can be expressed as

$$y(t) = \int_{f \in f_x} H(f, t) X(f) e^{j2\pi f t} \, df \tag{13}$$

where f_x denotes the frequency range over which $X(f)$ is not zero [for $f \notin f_x$, $X(f) = 0$]. If the bandwidth of the signal is much less than the coherence bandwidth of the channel, $H(f, t)$ does not change appreciably over the integration interval above. In other words, all the frequency components of $x(t)$ are subject to the same attenuation and phase shift in transmission through the channel. Such a channel is called *frequency-nonselective*, *narrowband*, or *flat fading*. The effect of the channel on the signal is thus multiplicative and the channel output can be written as

$$y(t) = \alpha(t) x(t) \tag{14}$$

where $\alpha(t)$ is the complex fading coefficient at time t.

A frequency-nonselective channel is slowly fading if the time duration of a transmitted symbol T_s is much smaller than the coherence time of the channel ($T_s \ll T_c$). Equivalently, $T_s \ll 1/f_d$ or $f_d \ll 1/T_s$. A slowly fading, frequency-nonselective channel is normally underspread.

Table 1. Channel Types

	Flat	Selective
Slow	$T_s < 1/f_d$ $T_m < T_s$	$T_s < 1/f_d$ $T_m > T_s$
Fast	$T_s > 1/f_d$ $T_m < T_s$	$T_s > 1/f_d$ $T_m > T_s$

A rapidly fading channel is defined by the condition $T_s \geq T_c$. Table 1 shows several channel types.

Example 3. In the GSM mobile cellular system, which operates at around 900 MHz, data are sent in bursts of duration approximately 0.5 ms. The maximum speed of a mobile is 80 km/hr. Is this a rapidly or slowly fading channel? The TETRA digital private mobile radio system, which operates at around 400 MHz, with a burst duration of around 14 ms. Is this a rapidly or slowly fading channel?

For the GSM case, $T_s f_d = 0.5 \times 10^{-3} \times 64 = 0.034$. For the TETRA case, the maximum Doppler is 40 Hz. So $T_s f_d = 14 \times 10^{-3} \times 40 = 0.5$.

1.4. Frequency-Selective Fading

When considering mobile/wireless systems for voice and low-bit-rate data applications, it is customary to use narrowband channels. But the wideband mobile radio channel has assumed increasing importance as the emergence of data rates to support multimedia services.

When the transmitted signal has a bandwidth greater than the coherence bandwidth of the channel, the signal suffers *frequency-selective* fading. Such channels also include time-selective fading. The standard model for wideband channel models is a tapped-delay line with complex-valued, time-varying tap gains. In the most general model, we have

$$h(t, \tau) = \sum_n \alpha_n(t)\delta(t - \tau_n(t)) \tag{15}$$

The tap gains $\alpha_n(t)$ are usually modeled as stationary mutually uncorrelated random processes having not necessarily identical ACFs and Doppler power spectra. Thus each resolvable multipath component may be modeled with its own appropriate Doppler power spectrum (5) and corresponding Doppler spread.

1.5. Fading Distribution Models

A transmitter and receiver are surrounded by objects that reflect and scatter signals. For a large number of such objects, we can apply the central limit theorem to model $h(t, \tau)$ as a Gaussian random process. If this process has a mean of zero, the envelope of the channel impulse response at time t has a Rayleigh probability distribution and the phase is uniformly distributed; that is, the envelope

$$R = |h(t, \tau)| \tag{16}$$

has the probability density function (PDF)

$$f(r) = \frac{2r}{\Omega} e^{-r^2/\Omega}, \quad r \geq 0 \tag{17}$$

where $\Omega = E(r^2)$. The Rayleigh distribution is characterized by this single parameter. For the frequency-nonselective channel, the envelope is simply the magnitude of the channel multiplicative gain [Eq. (14)]. For the frequency-selective channel model, each of the tap gains $\alpha_n(t)$ [Eq. (15)] has a magnitude that can be modeled as Rayleigh fading.

1.5.1. Nakagami *m* Distribution. The Nakagami distribution (m distribution) is a versatile statistical distribution that can accurately model a variety of fading environments. It has greater flexibility in matching some empirical data than do the Rayleigh, lognormal, or Rice distributions owing to its characterization of the received signal as the sum of vectors with random moduli and random phases. It also includes the Rayleigh and the one-sided Gaussian distributions as special cases. Moreover, the m distribution can closely approximate the Rice distribution. The PDF for this distribution is [10]

$$f(r) = \frac{2}{\Gamma(m)} \left(\frac{m}{\Omega}\right)^m r^{2m-1} e^{-mr^2/\Omega}, \quad r \geq 0 \tag{18}$$

where the parameter m is defined as the ratio of moments, called the *fading figure*:

$$m = \frac{\Omega^2}{E(R^2 - \Omega)^2)}, \quad m \geq \frac{1}{2} \tag{19}$$

1.5.2. Rice Distribution. The Rice distribution is used to characterize the signal in a line-of-sight (LoS) channel. The received signal consists of a multipath component, whose amplitude is described by the Rayleigh distribution, and a LoS component (also called the *specular component*) that has constant power. The PDF for the Rice distribution is

$$f(r) = \frac{r}{\sigma^2} e^{-(r^2+s^2)/2\sigma^2} I_0\left(\frac{rs}{\sigma^2}\right) \tag{20}$$

where s^2 represents the power in the nonfading (specular) signal components and σ^2 is the variance of the corresponding zero-mean Gaussian components. If s is set to zero, this reduces to the Rayleigh PDF.

2. ANALYSIS TECHNIQUES

In this section, we illustrate several analysis techniques via examples. We will consider diversity reception, outage analysis, and trellis codes. We will show how to obtain theoretical expressions for parameters such as the error rate, the average output SNR, and the outage probability.

2.1. Diversity Reception

Diversity methods (implemented at either the receiver, the transmitter, or both) can be effective for combating the effects of multipath fading. Performance and complexity can be traded off against each other when implementing

diversity techniques. For instance, consider the design of an antenna array receiver for millimeter-wave communications. Since the wavelength is less than 1 cm, several tens of array elements can be placed on the surface of a portable receiver. Classic signal combining techniques such as maximal-ratio combining (MRC), equal-gain combining (EGC), and selection combining (SC) may not be used with a large number of antenna elements (say, N) because of the need for N independent receivers, which is expensive and obeys the law of diminishing returns. An alternative is switched diversity combining (SDC), but the performance is worse. Thus, suboptimal receiver structures may exploit ordered statistics or a partitioned diversity combining scheme can be used to achieve the performance comparable to the optimal receiver but with considerably fewer electronics (hardware) and power consumption. While performance analysis of such schemes is beyond the scope of this article, the following techniques are a good starting point.

The basic premise of diversity is that the receiver processes multiple copies of the transmitted signal, where each copy is received through a distinct channel. If these channels are independent, then the chance of a deep fade occurring on all the channels simultaneously is small. Indeed, if a chance of a fade in a channel is p, the chance of a fade among N independent channels is p^N, which can be very small. This method requires N receiver circuits in the combiner. Each channel and the corresponding receiver circuit is called a *branch*. Two conditions are necessary for obtaining a high degree of improvement from a diversity combiner: (1) the fading in individual branches should have low cross-correlation — if the correlation is high, then deep fades in the branches can occur simultaneously, which negates diversity gain; and (2) the mean power from each branch should be almost equal.

2.2. Selection Combining

We will next show how selection combining can be analyzed for Rayleigh fading channels. The selection diversity combiner selects the branch that instantaneously has the highest SNR. The mathematical expression for the output SNR is simply

$$\gamma_{sc} = \max(\gamma_1, \gamma_2, \ldots, \gamma_N) \tag{21}$$

where γ_i is the SNR for the ith branch. For Rayleigh fading, using Eq. (17), the probability that a branch having an SNR less that γ can be found as

$$\Pr(\text{SNR} < \gamma) = \left(1 - e^{-\gamma/\Gamma}\right) \tag{22}$$

If all the fading branches are independent, the probability of the output of the selection combiner having an SNR less than γ is just the abovementioned probability raised to the power N. Thus, we have

$$\Pr(\gamma_{sc} < \gamma) = \left(1 - e^{-\gamma/\Gamma}\right)^N \tag{23}$$

where Γ is the SNR at the input of each branch, assumed to be the same for all branches. If γ is very small compared to the mean input SNR Γ, we have

$$\Pr(\gamma_{sc} < \gamma) \approx \left(\frac{\gamma}{\Gamma}\right)^N \tag{24}$$

Hence, the probability of a fade is simply the equivalent for a single-branch Rayleigh raised to the power N. Diversity gain is defined as the decrease in mean SNR to achieve a given probability of signal exceedance with and without diversity. The preceding shows that diversity gain increases with N. To further analyze performance, we need the probability density function (PDF) of the output. By differentiating (23) with respect to γ, we obtain

$$f_{\gamma_{sc}}(\gamma) = \frac{1}{\Gamma}\sum_{k=1}^{N} k(-1)^{k+1}\binom{N}{k} e^{-k\gamma/\Gamma} \tag{25}$$

This PDF can be used to derive expressions for various statistical parameters of the output. For example, the average output SNR is obtained as

$$\begin{aligned}\overline{\gamma}_{sc} &= \int_0^\infty \gamma f_{\gamma_{sc}}(\gamma)\, d\gamma \\ &= \Gamma \sum_{r=1}^{N} \frac{1}{r}\end{aligned} \tag{26}$$

The bit error rate for optimum detection of binary phase shift keying (BPSK), differential phase shift keying (DPSK), coherent frequency shift keying (CFSK), and noncoherent frequency shift keying (NCFSK) in Gaussian noise can be given as

$$\begin{aligned} P_e(\gamma) &= Q\left(\sqrt{2a\gamma}\right) \begin{cases} a = 1 & \text{BPSK} \\ a = \frac{1}{2} & \text{CFSK} \end{cases} \\ P_e(\gamma) &= \frac{1}{2}\exp(-a\gamma) \begin{cases} a = 1 & \text{DPSK} \\ a = \frac{1}{2} & \text{NCFSK} \end{cases} \end{aligned} \tag{27}$$

where γ is the instantaneous SNR and

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

We model the instantaneous SNR as a random variable with the PDF in (25). So the average output error rate is obtained by the formula

$$\overline{P}_e = \int_0^\infty P_e(\gamma) f_{sc}(\gamma)\, d\gamma \tag{28}$$

Thus, we obtain

$$\begin{aligned} \overline{P}_e &= \frac{1}{2}\sum_{k=1}^{N}(-1)^{k+1}\binom{N}{k}\left[1 - \sqrt{\frac{a\Gamma}{a\Gamma + k}}\right] \\ &\qquad \text{for BPSK and CFSK} \\ \overline{P}_e &= \frac{1}{2}\sum_{k=1}^{N} k(-1)^{k+1}\binom{N}{k}\frac{1}{k + a\Gamma} \\ &\qquad \text{for DPSK and NCFSK.} \end{aligned} \tag{29}$$

This averaging technique can be extended to other higher-order modulation schemes. We refer the reader to many papers on such topics.

2.2.1. Unequal Fading Branches. If the power in the fading branches is unequal (but independent), we need to modify the above analysis. Thus we can show that

$$\Pr(\gamma_{sc} < \gamma) = \prod_{k=1}^{N} \left(1 - e^{-\gamma/\Gamma_k}\right) \tag{30}$$

where Γ_k is the SNR at the kth input branch. Although equal branch powers (where Γ_k is constant) are needed to obtain maximum diversity benefit, better results are achievable in this case. The error rate performance of the above modulation methods can be derived similarly.

2.2.2. Dual-Branch SC Performance in Correlated Rayleigh Fading. The discussion above is premised on the assumption of independent fading. However, the branch signals in practical diversity systems can often be correlated. So the effects of correlation in fading among diversity branches on the error rates of digital receivers is of interest to the designers. Fairly comprehensive results have been developed for maximal-ratio combining (MRC), with arbitrary orders of diversity. The performance of MRC depends on the distribution of a sum $\sum \gamma_l$ of correlated signals, which is known for many cases. Unfortunately, performance analysis of selection diversity combiner in correlated fading is much more difficult.

For the dual-branch case with correlated Rayleigh fading, we can write the cumulative distribution function (CDF) of the SC output as

$$P(\gamma_{sc} \le \gamma) = 1 - \exp\left(-\frac{\gamma}{\Gamma}\right)\left[1 - Q(a,b) + Q(b,a)\right] \tag{31}$$

where $Q(a,b)$ is the Marcum Q function, defined as

$$Q(a,b) = \int_b^\infty \exp\left(-\frac{a^2+x^2}{2}\right) I_0(ax)\,dx$$

and

$$a = \sqrt{\frac{2\gamma\rho}{\Gamma(1-\rho)}} \quad \text{and} \quad b = \sqrt{\frac{2\gamma}{\Gamma(1-\rho)}}$$

where ρ is the normalized envelope covariance between the two branches. By differentiating (31) with respect to γ, we find

$$f_{\gamma_{sc}}(\gamma) = \frac{2}{\Gamma}\exp\left(-\frac{\gamma}{\Gamma}\right)\left[1 - Q(a,b)\right] \tag{32}$$

This PDF can be used to obtain performance statistics. For example, the average output SNR is obtained as

$$\begin{aligned}\overline{\gamma}_{sc} &= \int_0^\infty \gamma f_{\gamma_{sc}}(\gamma)\,d\gamma \\ &= 2\Gamma - \frac{\Gamma}{2}(1-\rho)^2 \\ &\quad \times \int_0^{2\pi} \frac{1}{(1-\sqrt{\rho}\cos\theta)(1+\rho-2\sqrt{\rho}\cos\theta)}\frac{d\theta}{2\pi} \\ &= \Gamma\left[1 + \frac{1}{2}\sqrt{1-\rho}\right].\end{aligned} \tag{33}$$

Note that for heavily correlated branches (e.g., $\rho \approx 1$), the average output SNR is simply the single-branch input SNR; that is, there is no diversity gain.

Using a technique similar to the derivation of (29), we can show that

$$\overline{P}_e = \frac{1}{2(1+a\Gamma)}\left(1 - \frac{a\Gamma(1-\rho)}{\sqrt{[2+a\Gamma(1-\rho)]^2 - 4\rho}}\right) \quad \text{for DPSK and NCFSK} \tag{34}$$

Again for heavily correlated branches (e.g., $\rho \approx 1$), the average output BER is simply that of the single-branch case; thus, there is no diversity gain.

2.3. Dual-Branch EGC Performance in Correlated Rayleigh Fading

EGC is of practical interest because it provides performance comparable to the optimal MRC technique but with greater simplicity. However, analyzing EGC receiver performance in fading is much more difficult. This is due to the difficulty of finding the PDF of the EGC output SNR, which depends on the square of a sum of N fading amplitudes. A closed-form solution to the PDF of this sum has been elusive for nearly 100 years (dating back to Lord Rayleigh), and indeed, even for the case of Rayleigh fading (mathematically simplest distribution), no solution exists for $N > 2$.

In an EGC combiner, the output of different diversity branches are first cophased and weighted equally before being summed to give the resultant output. The instantaneous SNR at the output of the EGC combiner is

$$\begin{aligned}\gamma_{\text{egc}} &= \frac{\gamma_1 + \gamma_2 + 2\sqrt{\gamma_1\gamma_2}}{2} \\ &= \tfrac{1}{2}(R_1 + R_2)^2\end{aligned} \tag{35}$$

where γ_1 and γ_2 are the SNRs on individual branches and R_1 and R_2 denote to the signal amplitudes divided by the factor $\sqrt{2N_0}$ (i.e., normalized with respect the noise voltage).

We next show how the BER performance of EGC reception in correlated fading can be analyzed. For exact analysis of EGC, we need the characteristic function of $R_1 + R_2$ and therefore

$$\begin{aligned}\phi_\gamma(\omega) &= E\left\{e^{j\omega(R_1+R_2)}\right\} \\ &= (1-\rho)e^{-\omega^2(1-\rho)\Gamma/4}\sum_{k=1}^{\infty}\rho^{k-1} \\ &\quad \times \left[\frac{(2k-1)!}{2^{k-1}(k-1)!}D_{-2k}\left(-j\omega\sqrt{\frac{\Gamma(1-\rho)}{2}}\right)\right]^2\end{aligned} \tag{36}$$

where $D_p(z)$ is the parabolic cylinder function of order p. The average BER can be expressed as

$$\overline{P}_e = E\left\{Q\left(\sqrt{2a\gamma}\right)\right\} = E\left\{Q\left[\sqrt{a}(R_1 + R_2)\right]\right\} \tag{37}$$

where $a = 1$ for BPSK and $a = \frac{1}{2}$ for CFSK. Using an infinite series for the error function, we obtain

$$\overline{P}_e = \frac{1}{2} - \frac{2}{\pi}\sum_{\substack{n=1\\ n\text{odd}}}^{\infty} \frac{\exp(-n^2\omega_0^2/2)}{n}\,\mathrm{Im}\left[\phi_\gamma(n\omega_0\sqrt{a})\right] \tag{38}$$

where ω_0 is a suitably small parameter. This is the exact solution for the performance of EGC reception in a correlated Rayleigh fading environment.

2.4. Mobile Outage Undershadowing

Consider computing the image function defined as

$$\phi_\alpha(s) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \frac{e^{-x^2}}{1 + se^{\alpha x}}\,dx \tag{39}$$

where s, $\alpha > 0$. The Laplace transform of a Suzuki PDF is a special case with $\alpha = \sqrt{2}\sigma/4.34$, where σ is the standard deviation of shadowing in decibels. The range of interest may be $3 < \sigma \le 12$ and $0 < s \le 10^3$. This image function has extensive applications in evaluating the outage performance of multiuser mobile radio networks.

Using some analytic techniques (see listed references), we can show that

$$\phi_\alpha(s) = \frac{h}{\sqrt{\pi}}\sum_{n=-\infty}^{\infty} \frac{e^{-(nh - \ln s/\alpha)^2}}{1 + e^{nh\alpha}} + E_c \tag{40}$$

where h is a small parameter controlling the correction term E_c. The value of h should not be too large or too small. It is found that a h value between 0.2 and 0.4 is sufficient for this application.

2.5. Outage Probability

Consider evaluating the probability of outage (outage) in a mobile fading environment. The instantaneous signal powers are modeled as random variables (RVs) $\mathbf{p}_k$, $k = 0, \ldots, L$, with mean $\overline{p}_k$. Subscript $k = 0$ denotes the desired signal and $k = 1, \ldots, L$ are for interfering signals. The outage is given by

$$\mathrm{P}_{\text{out}} = \Pr\{qI > \mathbf{p}_0\} \tag{41}$$

where $I = \mathbf{p}_1 + \cdots + \mathbf{p}_L$ and q is the power protection ratio, which is fixed by the type of modulation and transmission technique employed and the quality of service desired. Typically, $9 < q < 20$ (dB). On introducing $\gamma = qI - \mathbf{p}_0$, we can readily find the moment-generating function (MGF) $\phi_\gamma(s)$.

Since the outage probability is $\mathrm{p}_{\text{out}} = \Pr(\gamma < 0)$, we can show that

$$\mathrm{P}_{\text{out}} = \frac{1}{2n}\sum_{i=1}^{n} \tilde{\phi}\left[\frac{(2i-1)\pi}{2n}\right] + R_n \tag{42}$$

where $\tilde{\phi}(\theta) = \mathrm{Re}\left[(1 - j\tan(\theta/2))\phi_\gamma(c + jc\tan(\theta/2))\right]$ and the remainder term R_n vanishes rapidly. Although c can be anywhere between 0 and $a_{\min}$, the optimal location ensures that $|\phi_\gamma(c + j\omega)|$ decays as rapidly as possible for $|\omega| \to \infty$. This rapid decay occurs if $s = c$ is the saddle point; thus, at $s = c$, $s^{-1}\phi_\gamma(s)$ achieves its minimum on the real axis. While this optimal c requires a numerical search, it is sufficient to use $c = a_{\min}/2$. This formula can be used to compute the outage probability for various mobile systems and fading channel configurations.

2.6. Trellis-Coded PSK

The performance of convolutional codes, Turbo codes, and trellis-coded modulation (TCM) schemes over wireless channels has received much attention. There are several methods to analyze the performance of such codes. Here we describe the evaluation of the union bound.

The union bound technique is based on

$$P_b \le \frac{1}{k}\sum_{\mathbf{z},\hat{\mathbf{z}}\in\mathcal{C}} a(\mathbf{z} \to \hat{\mathbf{z}})P(\mathbf{z} \to \hat{\mathbf{z}}) \tag{43}$$

where k is the number of input bits per encoding interval, $P(\mathbf{z} \to \hat{\mathbf{z}})$ is the pairwise error probability (PEP), $a(\mathbf{z} \to \hat{\mathbf{z}})$ is the number of associated bit errors, and $\mathcal{C}$ is the set of all legitimate code sequences. But the evaluation of even the union bound is difficult since $P(\mathbf{z} \to \hat{\mathbf{z}})$ requires complex calculations. Thus, bounds on $P(\mathbf{z} \to \hat{\mathbf{z}})$ itself are used to compute (43), resulting in a weaker union bound. We next show a more general method to evaluate the union bound exactly, and this method is applicable to practical schemes such as differential detection and pilot-tone-aided detection. This approach can also be extended to schemes such as Turbo coding and spacetime codes.

2.6.1. System Model. The received complex sample at time n is

$$y_n = \alpha_n z_n + v_n$$

where α_n is the channel gain and v_n is an additive Gaussian noise sample. The following is used throughout the presentation:

A1. z_n is a q-ary phase shift keying (PSK) symbol (i.e., $z_n \in \{e^{j2\pi k/q} \mid k = 0, 1, \ldots, q-1\}$ and $j = \sqrt{-1}$).
A2. Each α_n is a zero-mean, complex, Gaussian random variable (RV). The α_n terms are independent (i.e., ideal interleaving/deinterleaving) and identically distributed RVs.
A3. Each α_n remains constant during a symbol interval (i.e., nonselective slow fading).
A4. The receiver has some form of channel measurements given by $\hat{\alpha}_n$ that is a complex Gaussian RV. The correlation coefficient between α and $\hat{\alpha}_n$ is μ.

If $\mu = 1$, ideal channel measurements exist. For practical channel estimators $|\mu| \le 1$. The more μ deviates from unity, the larger is the performance penalty.

2.6.2. Pairwise Error Event Probability. Consider two codewords $\mathbf{z} = \{z_1, z_2, \ldots, z_N\}$ and $\hat{\mathbf{z}} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_N\}$ of length N. The Viterbi decoder computes the path metrics and selects $\mathbf{z}$ over $\hat{\mathbf{z}}$ according to the path metric difference D. The characteristic function of D, $\phi(j\omega) = E[e^{j\omega D}]$, can be written as

$$\phi(j\omega) = \prod_{n\in\eta} \frac{\Delta_n}{\omega^2 - j\omega + \Delta_n} \tag{44}$$

where $\Delta_n \triangleq [1 + (1 - |\mu|^2)\gamma_s]/(|\mu|^2|z_n - \hat{z}_n|^2\gamma_s)$, $\eta \triangleq \{n \mid z_n \neq \hat{z}_n, n = 1, \ldots, N\}$, and $\gamma_s = \overline{E}_s/N_0$ is the average signal-to-noise ratio. It can be shown that

$$P(\mathbf{z} \to \hat{\mathbf{z}}) = \frac{-1}{2\pi j} \int_{-\infty+j\varepsilon}^{\infty+j\varepsilon} \frac{\phi(j\omega)}{\omega}\, d\omega \tag{45}$$

where ε is a small positive number. An explicit expression for $P(\mathbf{z} \to \hat{\mathbf{z}})$, which cannot be used with the transfer function approach, can be obtained by solving for the residues of the contour integral. However, a transfer function can be defined with the *factors* of $\phi(j\omega)$.

2.6.3. Union Bound. Consider the evaluation of the union bound (43). Let $\mathbf{Z} = (Z_1, Z_2, \ldots)$ be a vector of formal variables. Define the generating function of the form

$$T(\mathbf{Z}, I) = \sum_{\mathbf{z},\hat{\mathbf{z}}\in\mathcal{C}} I^{a(\mathbf{z}\to\hat{\mathbf{z}})} \prod_{n\in\eta} Z_n \tag{46}$$

where I is another formal variable. Moreover, let

$$D_n(\omega) \triangleq \frac{\Delta_n}{\omega^2 - j\omega + \Delta_n} \tag{47}$$

The number of distinct values that $D_n(\omega)$ can take depends on the size of the signal constellation. The transfer function $T(\mathbf{D}(\omega), I)$ can be determined by a signal flow graph with the branch labels $I^v D_n(\omega)$ for uniform trellis codes. By contrast, for the union–Chernoff bound, the branches are labeled with $I^v(1 + 1/(4\Delta_n))^{-1}$.

Combining (43), (45) and (46), and using the standard analysis, it follows that

$$P_b \le \frac{-1}{j2\pi k} \frac{\partial}{\partial I} \left\{ \int_{-\infty+j\varepsilon}^{\infty+j\varepsilon} \frac{T(\mathbf{D}(\omega), I)}{\omega}\, d\omega \right\}\Bigg|_{I=1} \tag{48}$$

where the partial derivative can be computed. While this integral has no analytical solution in general, its numerical computation poses little difficulty. Since $|T(\mathbf{D}(\omega), I)| \to 0$ as $|\omega| \to \infty$, simple techniques such as the Simpson method are adequate.

Example 4. Consider the performance of the two-state trellis-coded QPSK (Fig. 6) in Rayleigh fading. Using branch label gains, $D_n(\omega)$, the transfer function becomes

$$T(\mathbf{D}(\omega), I) = \frac{I\Delta_2\Delta_4}{(\omega^2 - j\omega + \Delta_4)(\omega^2 - j\omega + (1 - I)\Delta_2)}$$

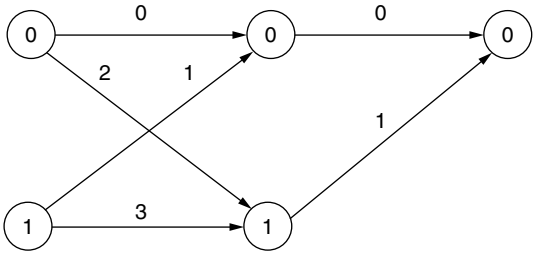


Figure 6. State diagram.

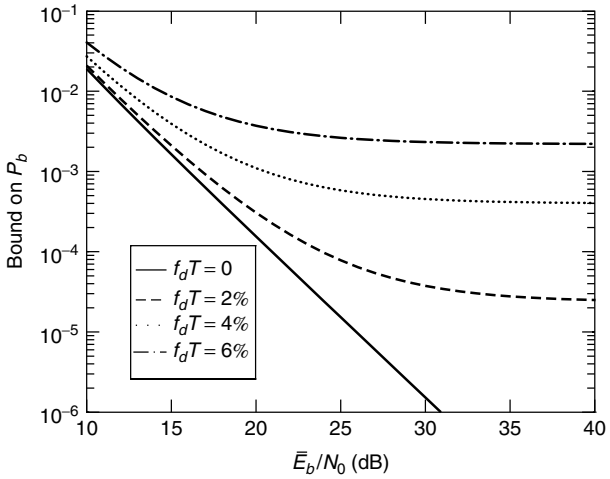


Figure 7. Bit error performance of rate-$\frac{1}{2}$ trellis-coded QPSK (differential detection) for fast Rayleigh fading.

where Δ_2 and Δ_4 are obtained with $|z_n - \hat{z}_n|^2$ equal to 2 and 4, respectively. Substituting this in (48), carrying out the integration, and evaluating the derivative at $I = 1$, one has

$$P_b \le 1 - \frac{(1 + (1 - |\mu|^2)\gamma_s)}{2|\mu|^2\gamma_s} + \frac{3(1 + (1 - |\mu|^2)\gamma_s)^2}{4|\mu|^4\gamma_s^2} - |\mu|\sqrt{\frac{\gamma_s}{1 + \gamma_s}}. \tag{49}$$

This is the exact union bound for this TCM scheme. If differential detection is used in a flat fading land mobile channel, the correlation coefficient μ is given in Ref. 19, Eq. (32) as a function of the normalized maximum Doppler frequency f_dT. Figure 7 shows Eq. (49) for several f_dT values. Unless $f_dT = 0$, an error floor exists.

Example 5. Consider the trellis-coded 8-PSK scheme given in Ref. 20, Fig. 5. Its transfer function Ref. 20, Eq. (19) can be defined with the weight profiles obtained using $D_n(\omega)$. Consider pilot-tone-aided detection with μ given in Ref. 19, Eq. (40). Assume that the bandwidth of the pilot tone filter is $2f_d$ and the power-split ratio is $\sqrt{2f_dT}$. Figure 8 shows the simulation results and the union bound (48). At $P_b \approx 10^{-4}$, the simulation results are within 0.2, 1.0, and 1.5 dB of the union bound for f_dT of 0, 0.02, and 0.04, respectively.

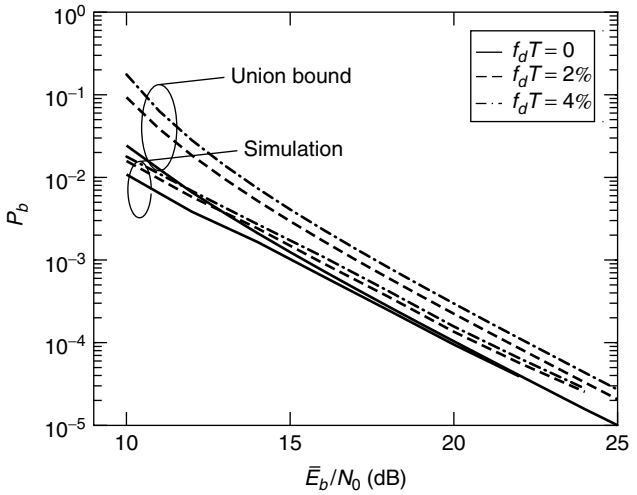


Figure 8. Bit error performance of rate-$\frac{2}{3}$, 4-state, trellis-coded 8PSK for fast Rayleigh fading and pilot-tone-aided detection.

BIOGRAPHIES

C. Tellambura received his B.Sc. degree with honors from the University of Moratuwa, Sri Lanka, in 1986, his M.Sc. in electronics from the King's College, UK, in 1988, and his Ph.D. in electrical engineering from the University of Victoria, Canada, in 1993. He was a postdoctoral research fellow with the University of Victoria and the University of Bradford. Currently, he is a senior lecturer at Monash University, Australia. He is an editor for the *IEEE Transactions on Communications* and the *IEEE Journal on Selected Areas in Communications* (Wireless Communications Series). His research interests include coding, communications theory, modulation, equalization, and wireless communications.

A. Annamalai received his B.E. degree in electrical and computer engineering in 1993 from University of Science of Malaysis, and his M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Victoria in 1997 and 1999, respectively. He was with Motorola Inc. as an RF design engineer from 1993 to 1995 and a postdoctoral research fellow at the University of Victoria in 1999. Since December 1999, he has been an assistant professor at the Virginia Polytechnic Institute and State University where he has been working on smart antennas and communication receiver designs. Dr. Annamalai has published over 60 papers in the field of wireless communications. He was the recipient of the 1997 Lieutenant Governor's medal, 1998 IEEE Daniel E. Noble Graduate Fellowship, 2000 NSERC Doctoral Prize, 2000 CAGS/UMI Doctoral Dissertation Award, and the 2001 IEEE Leon K. Kirchmayer Prize Paper Award for his work on diversity systems. He is an editor for the *IEEE Transactions on Wireless Communications* and *Wiley's International Journal on Wireless Communications and Mobile Computing*, an associate editor for the *IEEE Communications Letters* and is the technical program chair for VTC'2002 (Fall). His research interests are in high-speed data transmission on wireless links, smart antennas, multicarrier communications, mathematical modeling of radio channels, and wireless communications theory.

BIBLIOGRAPHY

1. M. Zeng, A. Annamalai, and V. K. Bhargava, Harmonization of global third generation mobile systems, *IEEE Commun. Mag.* **38**: 94–104 (Dec. 2000).
2. M. Zeng, A. Annamalai, and V. K. Bhargava, Recent advances in cellular wireless communications, *IEEE Commun. Mag.* **37**: 128–138 (Sept. 1999).
3. S. R. Saunders, *Antennas and Propagation for Wireless Communication Systems*, Wiley, 1999.
4. E. Biglieri, J. Proakis, and S. Shamai, Fading channels: information-theoretic and communications aspects, *IEEE Trans. Inform. Theory* **44**: 2619–2692 (Oct. 1998).
5. H. Hashemi, The indoor radio propagation channel, *IEEE Proc.* **81**: 943–967 (July 1993).
6. B. Sklar, Rayleigh fading channels in mobile digital communication systems. Part I: Characterization, *IEEE Commun. Mag.* **35**: 136–146 (Sept. 1997).
7. B. Sklar, Rayleigh fading channels in mobile digital communication systems. Part II: Mitigation, *IEEE Commun. Mag.* **35**: 148–155 (Sept. 1997).
8. J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill Series in Electrical Engineering: Communications and Signal Processing, McGraw-Hill, New York, 1995.
9. W. C. Jakes, ed., *Microwave Mobile Communications*, Wiley, New York, 1974.
10. M. Nakagami, The m-distribution, a general formula of intensity distribution of rapid fading, in W. G. Hoffman, ed., *Statistical Methods in Radio Wave Propagation*, Pergamon, Oxford, 1960.
11. N. C. Beaulieu and A. A. Abu-Dayya, Analysis of equal gain diversity on Nakagami fading channels, *IEEE Trans. Commun.* **39**: 225–234 (Feb. 1991).
12. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill, New York, 1966.
13. N. C. Beaulieu, A simple series for personal computer computation of the error function Q(.), *IEEE Trans. Commun.* **37**: 989–991 (Sept. 1989).
14. A. Annamalai, C. Tellambura, and V. K. Bhargava, Equal-gain diversity receiver performance in wireless channels, *IEEE Trans. Commun.* **48**: 1732–1745 (Oct. 2000).
15. J.-P. M. G. Linnartz, Exact analysis of the outage probability in multiple-user mobile radio, *IEEE Trans. Commun.* **40**: 20–23 (Jan. 1992).
16. J.-P. Linnartz, *Narrowband Land-Mobile Radio Networks*, Artech, House, Boston, 1993.
17. A. Annamalai, C. Tellambura, and V. K. Bhargava, Simple and accurate methods for outage analysis in cellular mobile radio systems-a unified approach, *IEEE Trans. Commun.* **49**: 303–316 (Feb. 2001).
18. C. Tellambura, Evaluation of the exact union bound for trellis coded modulations over fading channels, *IEEE Trans. Commun.* **44**: 1693–1699 (Dec. 1996).
19. J. K. Cavers and P. Ho, Analysis of the error performance of trellis coded modulations in Rayleigh fading channels, *IEEE Trans. Commun.* **40**: 74–83 (Jan. 1992).

20. R. G. McKay, E. Biglieri, and P. J. McLane, Error bounds for Trellis-Coded MPSK on a fading mobile satellite channel, *IEEE Trans. Commun.* **39**: 1750–1761 (Dec. 1991).

WIRELESS INFRARED COMMUNICATIONS

JEFFREY B. CARRUTHERS
Boston University
Boston, Massachusetts

1. INTRODUCTION

Wireless infrared communications refers to the use of free-space propagation of lightwaves in the near-infrared band as a transmission medium for communication [1–3], as shown in Fig. 1. The communication can be between one portable communication device and another or between a portable device and a tethered device, called an *access point* or *base station*. Typical portable devices include laptop computers, personal digital assistants, and portable telephones, while the base stations are usually connected to a computer with other networked connections. Although infrared light is usually used, other regions of the optical spectrum can be used (hence the term "wireless optical communications" instead of "wireless infrared communications" is sometimes used).

Wireless infrared communication systems can be characterized by the application for which they are designed or by the link type, as described below.

1.1. Applications

The primary commercial applications are as follows:

- Short-term cableless connectivity for information exchange (business cards, schedules, file sharing) between two users. The primary example is Infrared Data Association (IRDA) systems (see Section 4).
- Wireless local-area networks (WLANs) provide network connectivity inside buildings. This can either be an extension of existing LANs to facilitate mobility, or to establish ad hoc networks where there is no LAN. The primary example is the IEEE 802.11 standard (see Section 4).
- Building-to-building connections for high-speed network access or metropolitan- or campus-area networks.

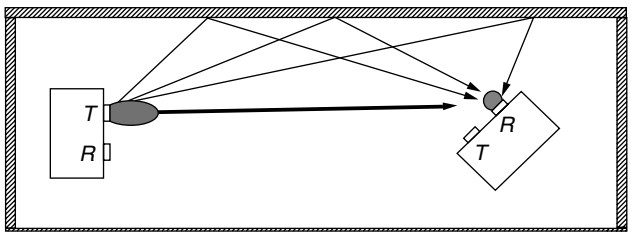


Figure 1. A typical wireless infrared communication system.

- Wireless input and control devices, such as wireless mice, remote controls, wireless game controllers, and remote electronic keys.

1.2. Link Type

Another important way to characterize a wireless infrared communication system is by the "link type," which means the typical or required arrangement of receiver and transmitter. Figure 2 depicts the two most common configurations: the point-to-point system and the diffuse system.

The simplest link type is the point-to-point system. There, the transmitter and receiver must be pointed at each other to establish a link. The *line-of-sight* (LoS) path from the transmitter to the receiver must be clear of obstructions, and most of the transmitted light is *directed* toward the receiver. Hence, point-to-point systems are also called *directed LoS systems*. The links can be temporarily created for a data exchange session between two users, or established more permanently by aiming a mobile unit at a base station unit in the LAN replacement application.

In diffuse systems, the link is always maintained between any transmitter and any receiver in the same vicinity by reflecting or "bouncing" the transmitted information-bearing light off reflecting surfaces such as ceilings, walls, and furniture. Here, the transmitter and receiver are *nondirected*; the transmitter employs a wide transmit beam and the receiver has a wide field of view (FoV). Also, the LoS path is not required. Hence, diffuse systems are also called *nondirected non-LoS systems*. These systems are well suited to the wireless LAN application, freeing the user from knowing and aligning with the locations of the other communicating devices.

1.3. Fundamentals and Outline

Most wireless infrared communications systems can be modeled as having an output signal $Y(t)$, and an input signal $X(t)$, which are related by

$$Y(t) = X(t) \otimes c(t) + N(t) \tag{1}$$

where $\otimes$ denotes convolution, $c(t)$ is the impulse response of the channel and $N(t)$ is additive noise. This article is organized around answering key questions concerning the system as represented by this model.

In Section 2, we consider questions of optical design. What range of wireless infrared communications systems does this model apply to? How does $c(t)$ depend on the electrical and optical properties of the receiver and transmitter? How does $c(t)$ depend on the location, size, and orientation of the receiver and transmitter? How do $X(t)$ and $Y(t)$ relate to optical processes? What wavelength is used for $X(t)$? What devices produce $X(t)$ and $Y(t)$? What is the source of $N(t)$? Are there any safety considerations? In Section 3, we consider questions of communications design. How should a data symbol sequence be modulated onto the input signal $X(t)$? What detection mechanism is best for extracting the information about the data from the received signal $Y(t)$? How can one measure and improve the performance of the system? In Section 4, we consider

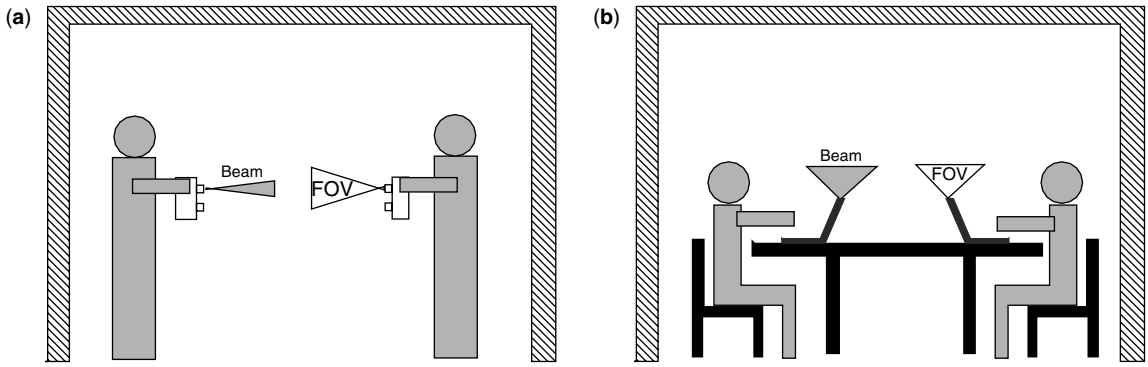


Figure 2. Common types of infrared communication systems: (**a**) point-to-point system; (**b**) diffuse system.

the design choices made by existing standards such as IRDA and IEEE 802.11. Finally, in Section 5, we consider how these systems can be improved in the future.

2. OPTICAL DESIGN

2.1. Modulation and Demodulation

What characteristic of the transmitted wave will be modulated to carry information from the transmitter to the receiver? Most communication systems are based on phase, amplitude, or frequency modulation, or some combination of these techniques. However, it is difficult to detect such a signal following nondirected propagation, and more expensive narrow-linewidth sources are required [2]. An effective solution is to use *intensity* modulation, where the transmitted signal's intensity or power is proportional to the modulating signal.

At the demodulator (usually referred to as a *detector* in optical systems), the modulation can be extracted by mixing the received signal with a carrier lightwave. This *coherent detection* technique is best when the signal phase can be maintained. However, this can be difficult to implement and additionally, in nondirected propagation, it is difficult to achieve the required mixing efficiency. Instead, one can use *direct detection* using a photodetector. The photodetector current is proportional to the received optical signal intensity, which for intensity modulation, is also the original modulating signal. Hence, most systems use intensity modulation with direct detection (IM/DD) to achieve optical modulation and demodulation.

In a free-space optical communication system, the detector is illuminated by sources of light energy other than the source. These can include ambient lighting sources, such as natural sunlight, fluorescent lamp light, and incandescent lamp light. These sources cause variation in the received photocurrent that is unrelated to the transmitted signal, resulting in an additive noise component at the receiver.

We can write the photocurrent at the receiver as

$$Y(t) = X(t) \otimes Rh(t) + N(t)$$

where R is the responsivity of the receiving photodiode [in amperes per watt (A/W)]. Note that the electrical impulse response $c(t)$ is simply R times the optical impulse response $h(t)$. Depending on the situation, some authors use $c(t)$ and some use $h(t)$ as the impulse response.

2.2. Receivers and Transmitters

A transmitter or *source* converts an electrical signal to an optical signal. The two most appropriate types of device are the light-emitting diode (LED) and semiconductor laser diode (LD). LEDs have a naturally wide transmission pattern, and so are suited to nondirected links. Eye safety is much simpler to achieve for an LED than for a laser diode, which usually has very narrow transmit beams. The principal advantages of laser diodes are their high energy-conversion efficiency, their high modulation bandwidth, and their relatively narrow spectral width. Although laser diodes offer several advantages over LEDs that could be exploited, most short-range commercial systems currently use LEDs.

A receiver or *detector* converts optical power into electrical current by detecting the photon flux incident on the detector surface. Silicon *p-i-n* photodiodes are ideal for wireless infrared communications as they have good quantum efficiency in this band and are inexpensive [4]. Avalanche photodiodes are not used here since the dominant noise source is background light-induced shot noise rather than thermal circuit noise.

2.3. Transmission Wavelength and Noise

The most important factor to consider when choosing a transmission wavelength is the availability of effective, low-cost sources and detectors. The availability of LEDs and silicon photodiodes operating in the 800–1000-nm range is the primary reason for the use of this band. Another important consideration is the spectral distribution of the dominant noise source: background lighting.

The noise $N(t)$ can be broken into four components: photon noise or shot noise, gain noise, receiver circuit or thermal noise, and periodic noise. Gain noise is only present in avalanche-type devices, so we will not consider it here.

Photon noise is the result of the discreteness of photon arrivals. It is due to background light sources, such as sunlight, fluorescent lamp light, and incandescent lamp

light, as well as the signal-dependent source $X(t) \otimes c(t)$. Since the background light striking the photodetector is normally much stronger than the signal light, we can neglect the dependency of $N(t)$ on $X(t)$ and consider the photon noise to be additive white Gaussian noise with two-sided power spectral density $S(f) = qRP_n$ where q is the electron charge, R is the responsivity, and P_n is the optical power of the noise (background light).

Receiver noise is due to thermal effects in the receiver circuitry, and is particularly dependent on the type of preamplifier used. With careful circuit design, it can be made insignificant relative to the photon noise [5].

Periodic noise is the result of the variation of fluorescent lighting due to the method of driving the lamp using the ballast. This generates an extraneous periodic signal with a fundamental frequency of 44 kHz with significant harmonics to several megahertz. Mitigating the effect of periodic noise can be done using highpass filtering in combination with baseline restoration [6], or by careful selection of the modulation type, as discussed in Section 3.1.

2.4. Safety

There are two safety concerns when dealing with infrared communication systems. Eye safety is a concern because of a combination of two effects. First, the cornea is transparent from the near violet to the near IR. Hence, the retina is sensitive to damage from light sources transmitting in these bands. Secondly, however, the near IR is outside the visible range of light, and so the eye does not protect itself from damage by closing the iris or closing the eyelid. Eye safety can be ensured by restricting the transmit beam strength according to IEC or ANSI standards [7,8].

Skin safety is also a possible concern. Possible short-term effects such as heating of the skin are accounted for by eye safety regulations (since the eye requires lower power levels than does the skin). Long-term exposure to IR light is not a concern, as the ambient light sources are constantly submitting our bodies to much higher radiation levels than these communication systems do.

3. COMMUNICATIONS DESIGN

Equally important for achieving the design goals of wireless infrared systems are communications issues. In particular, the modulation signal format together with appropriate error control coding is critical to achieving power efficiency. Channel characterization is also important for understanding performance limits.

3.1. Modulation Techniques

To understand modulation in IM/DD systems, we must look again at the channel model

$$Y(t) = X(t) \otimes c(t) + N(t)$$

and consider its particular characteristics. First, since we are using intensity modulation, the channel input $X(t)$ is optical intensity and we have the constraint $X(t) \geq 0$. The average transmitted optical power P_T is the time average of $X(t)$. Our goal is to minimize the transmitted power required to attain a certain probability of bit error P_e, also known as a bit error rate (BER).

It is useful to define the signal-to-noise ratio (SNR) as

$$\mathrm{SNR} = \frac{R^2 H^2(0) P_t^2}{R_b N_0}$$

where $H(0)$ is the DC gain of the channel, i.e., it is the Fourier transform of $h(t)$ evaluated at zero frequency, so

$$H(0) = \int_{-\infty}^{\infty} h(t)\, dt.$$

The transmitted signal can be represented as

$$X(t) = \sum_{n=-\infty}^{\infty} s_{a_n}(t - nT_s).$$

The sequence $\{a_n\}$ represents the digital information being transmitted, where a_n is one of L possible data symbols from 0 to $L-1$. The function $s_i(t)$ represents one of L pulseshapes with duration T_s, the symbol time. The data rate (or bit rate) R_b, bit time T, symbol rate R_s, and symbol time T_s are related as follows: $R_b = 1/T$, $R_s = 1/T_s$, and $T_s = \log_2(L)T$.

There are three commonly used types of modulation schemes: on/off keying (OOK) with non-return-to-zero (NRZ) pulses, OOK with return-to-zero (RZ) pulses of normalized width δ (RZ-δ), and pulse position modulation with L pulses (L-PPM). OOK and RZ-δ are simpler to implement at both the transmitter and receiver than L-PPM. The pulse shapes for these modulation techniques are shown in Fig. 3. Representative examples of the resulting transmitted signal $X(t)$ for a short data sequence are shown in Fig. 4.

We compare modulation schemes in Table 1 by looking at measures of *power efficiency* and *bandwidth efficiency*. Bandwidth efficiency is measured by dividing the zero-crossing (ZC) bandwidth by the data rate. Bandwidth-efficient schemes have several advantages — the receiver and transmitter electronics are cheaper, and the modulation scheme is less likely to be affected by multipath distortion. Power efficiency is measured by comparing the required transmit power to achieve a target probability of error P_e for different modulation techniques. Both RZ-δ and PPM are more power-efficient than OOK, but at the cost of reduced bandwidth efficiency. However, for a given bandwidth efficiency, PPM is more power-efficient than RZ-δ, and so PPM is most commonly used. OOK is most useful at very high data rates, say 100 Mbps (megabits per second) or greater. Then, the effect of multipath distortion is the most significant effect and bandwidth efficiency becomes of paramount importance [9].

3.2. Error Control Coding

Error control coding is an important technique for improving the quality of any digital communication system. We concentrate here on forward error correction channel

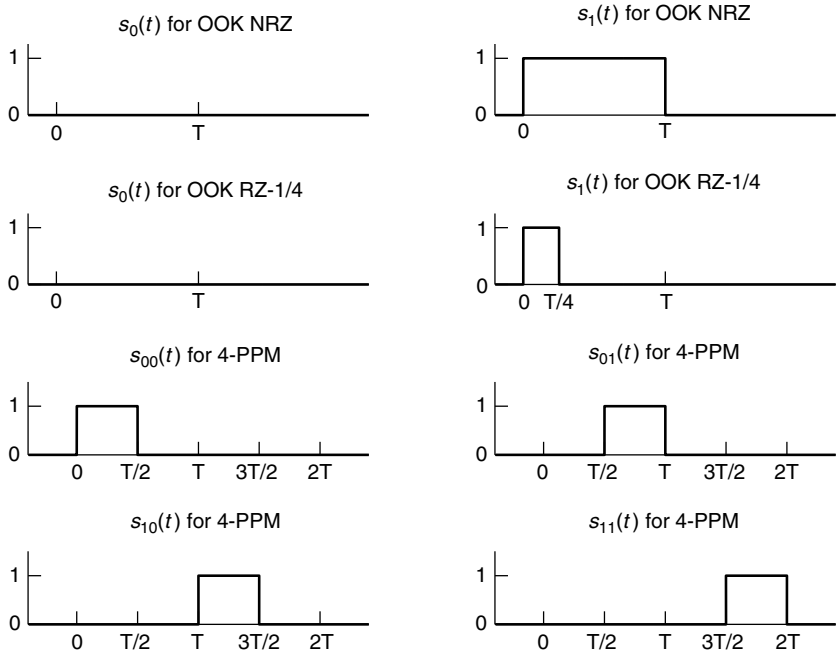


Figure 3. The pulse shapes for OOK, RZ-0.25, and 4-PPM.

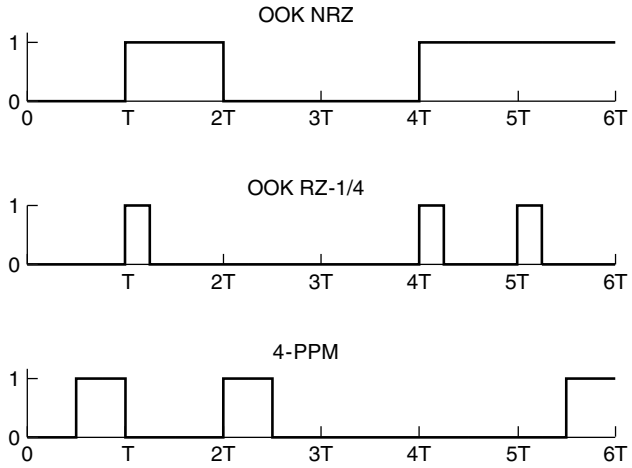


Figure 4. The transmitted signal for the sequence 010011 for OOK, RZ-0.25, and 4-PPM.

coding, as this specifically relates to wireless infrared communications; source coding and ARQ (automatic repeat request) coding are not considered here.

Trellis-coded PPM has been found to be an effective scheme for multipath infrared channels [10,11]. The key technique is to recognize that although on a distortion-free channel, all symbols are orthogonal and equidistant in signal space, this is not true on a distorting channel. Hence, trellis coding using set partitioning designed to separate the pulse positions of neighboring symbols is an effective coding method. Coding gains of 5.0 dB electrical have been reported for rate $\frac{2}{3}$-coded 8-PPM over uncoded 16-PPM, which has the same bandwidth [11].

3.3. Channel Impulse Response Characterization

Impulse response characterization refers to the problem of understanding how the impulse response $c(t)$ in Eq. (1) depends on the location, size, and orientation of the receiver and transmitter. There are basically three classes of techniques for accomplishing this: measurement, simulation, and modeling. Channel measurements have been described in several studies [2,9,12], and these form the fundamental basis for understanding the channel properties. A particular study might generate a collection of hundreds or thousands of example impulse responses $c_i(t)$ for configuration i. The collection of measured impulse responses $c_i(t)$ can then be studied by looking at scatterplots of path loss versus distance, scatterplots of delay spread versus distance, the effect of transmitter and receiver orientations, robustness to shadowing, and so on.

Simulation methods have been used to allow direct calculation of a particular impulse response based on a site-specific characterization of the propagation

Table 1. Comparison of Modulation Schemes on Ideal Channels

Modulation Type	P_e	ZC Bandwidth
On/off keying (OOK)	$Q(\mathrm{SNR}^{1/2})$	R_b
OOK RZ-δ	$Q(\delta^{-1/2}\ \mathrm{SNR}^{1/2})$	$\frac{1}{\delta}R_b$
L-PPM	$Q\left((0.5L * \log_2(L))^{1/2}\mathrm{SNR}^{1/2}\right)$	$\frac{L}{\log_2 L}R_b$

environment [13,14]. The transmitter, the receiver, and the reflecting surfaces are described and used to generate an impulse response. The basic assumption is that most interior surfaces reflect light diffusely in a Lambertian pattern, that is, all incident light, regardless of incident angle, is reflected in all directions with an intensity proportional to the cosine of the angle of the reflection with the surface normal. The difficulty with existing methods is that accurate modeling requires extensive computation.

A third technique attempts to extract knowledge gained from experimental and simulation-based channel estimations into a simple-to-use model. In Ref. 15, for example, a model using two parameters (one for path loss, one for delay spread) is used to provide a general characterization of all diffuse IR channels. Methods for relating the parameters of the model to particular room characteristics are given, so that system designers can quickly estimate the channel characteristics in a wide range of situations.

4. STANDARDS AND SYSTEMS

We examine the details of the two dominant wireless infrared technologies, IRDA and IEEE 802.11, and other commercial applications.

4.1. Infrared Data Association Standards (IRDA)

The Infrared Data Association [16], an association of about 100 member companies, has standardized low-cost optical data links. The IRDA link transceivers or "ports," appear on many portable devices, including notebook computers, personal digital assistants, and also computer peripherals such as printers.

The series of IRDA transmission standards are described in Table 2. The current version of the physical-layer standards is IrPHY 1.3. Data rates ranging from 2.4 kbps to 4 Mbps are supported. The link speed is negotiated by starting at 9.6 kbps.

Most of the transmission standards are for short-range, directed links with an operating range from 0 to 1 m. The transmitter half-angle must be between 15° and 30°, and the receiver field-of-view half-angle must be at least 15°. The transmitter must have a peak-power wavelength between 850 and 900 nm.

4.2. IEEE 802.11 and Wireless LANs

The IEEE has published a set of standards for wireless LANs, IEEE 802.11 [17]. The IEEE 802.11 standard is designed to fit into the structure of the suite of IEEE 802 LAN standards. Hence, it determines the physical layer (PHY) and medium-access control layer (MAC) leaving the logical link control (LLC) IEEE to 802.2. The MAC layer uses a form of carrier-sense multiple access with collision avoidance (CSMA/CA).

The original standard supports both radio and optical physical layers with a maximum data rate of 2 Mbps. The IEEE 802.11b standard adds a 2.4-GHz radio physical layer at up to 11 Mbps and the IEEE 802.11a standard adds a 5.4-GHz radio physical layer at up to 54 Mbps.

The two supported data rates for infrared IEEE 802.11 LANs are 1 and 2 Mbps. Both systems use PPM but share a common chip rate of 4 Mchips/s, as explained below. Each frame begins with a preamble encoded using 4 Mbps OOK. In the preamble, a 3-bit field indicates the transmission type, either 1 or 2 Mbps (the six other types are reserved for future use). The data are then transmitted at 1 Mbps using 16-PPM or 2 Mbps using 4-PPM. 16-PPM carries $\log_2(16)/16 = \frac{1}{4}$ bits/chip, and 4-PPM carries $\log_2(4)/4 = \frac{1}{2}$ bit/chip, resulting in the same chip time for both types.

The transmitter must have a peak-power wavelength between 850 and 950 nm. The required transmitter and receiver characteristics are intended to allow for reliable operation at link lengths up to 10 m.

4.3. Building-to-Building Systems

Long-range (>10 m) infrared links must be directed LoS systems in order to ensure a reasonable path loss. The emerging products for long-range links are typically designed to be placed on rooftops [18,19], as this provides the best chance for establishing line-of-sight paths from one location to another in an urban environment. These high-data-rate connections can then be used for enterprise network access or metropolitan- or campus-area networks.

Several design issues are specific to these systems that are unique to these long-range systems [3]: (1) *atmospheric path loss*, which is a combination of clean-air absorption from the air and absorption and scattering from particles in the air, such as rain, fog, and pollutants;

Table 2. IRDA Data Transmission Standards

Version	Link Type	Link Range (m)	Data Rate	Modulation
1.3	Point-to-point	1	2.4–115.2 kbps	RZ-$\frac{3}{16}$
1.3	Point-to-point	1	576 kbps	RZ-$\frac{1}{4}$
			1152 kbps	RZ-$\frac{1}{4}$
1.3	Point-to-point	1	4 Mbps	4-PPM
VFIR[a]/1.4	Point-to-point	1	16 Mbps	OOK
AIR[b]/proposed	Network	4	4 Mbps	—
		8	250 kbps	—

[a]VIFR = Very Fast Infrared.
[b]AIR = Advanced Infrared.

(2) *scintillation*, which is caused by temperature variations along the LoS path, causing rapid fluctuations in the channel quality; and (3) *building sway*, which can affect alignment and result in signal loss unless the transceivers are mechanically isolated or active alignment compensation is used.

4.4. Other Applications

Wireless infrared communication has found several markets in and around the home, car, and office that fall outside the traditional telecommunications markets of voice and data networking. These can be classified as either wireless input devices or wireless control devices, depending on one's perspective. Examples include wireless computer mice and keyboards, remote controls for entertainment equipment, wireless videogame controllers, and wireless door keys for home or vehicle access. All such devices use infrared communication systems because of the attractive combination of low cost, reliability, and light weight in a transmitter/receiver pair that achieves the required range, data rate, and data integrity required.

5. TECHNOLOGY OUTLOOK

In this section, we discuss how competition from radio and developments in research will impact the future uses of wireless infrared communication systems.

5.1. Comparison to Radio

Wireless infrared communication systems enjoy significant advantages over radio systems in certain environments. First, there is an abundance of unregulated optical spectrum available. This advantage is shrinking somewhat as the spectrum available for licensed and unlicensed radio systems increases with the modernization of spectrum allocation policies.

Radio systems must make great efforts to overcome or avoid the effects of multipath fading, typically through the use of diversity. Infrared systems do not suffer from time-varying fades because of the inherent diversity in the receiver, thus simplifying design and increasing operational reliability.

Infrared systems provide a natural resistance to eavesdropping, as the signals are confined within the walls of the room. This also reduces the potential for neighboring wireless communication systems to interfere with each other, which is a significant issue for radio-based communication systems.

In-band interference is a significant problem for both types of systems. A variety of electronic and electrical equipment radiates in transmission bands of current radio systems; microwave ovens are a good example. For infrared systems, ambient light, either human-made (synthetic) or natural, is a dominant source of noise.

The primary limiting factor of infrared systems is their limited range, particularly when no good optical path can be made available. For example, wireless communication between conventional rooms with opaque walls and doors cannot be accomplished; one must resort to using either a radio-based or a wireline network to bypass the obstruction.

5.2. Research Challenges

Various techniques have been considered to improve on the performance of wireless infrared communication systems.

At the transmitter, the radiation pattern can be optimized to improve performance characteristics such as range. Some optical techniques for achieving this are diffusing screens, multiple-beam transmitters, and computer-generated holographic images.

At the receiver, performance is ultimately determined by signal collection (limited by the size of the photodetector) and by ambient noise filtering. Optical interference filters can be used to reduce the impact of background noise; the primary difficulty is in achieving a wide field of view. This can be done using nonplanar filters or multiple narrow-FoV receiving elements.

Some recent developments and research programs are described in Ref. 20, and an on-line resource guide is maintained in Ref. 21.

6. CONCLUSIONS

Wireless infrared communication systems provide a useful complement to radio-based systems, particularly for systems requiring low cost, lightweight, moderate data rates, and only requiring short ranges. When LoS paths can be ensured, range can be dramatically improved to provide longer links.

Short-range wireless networks are poised for tremendous market growth in the near future, and wireless infrared communications systems will compete in a number of arenas. Infrared systems have already proved their effectiveness for short-range temporary communications and in high-data-rate, longer-range point-to-point systems. It remains an open question whether infrared will successfully compete in the market for general-purpose indoor wireless access.

BIOGRAPHY

Jeffrey B. Carruthers received his B.Eng. degree in computer systems engineering from Carleton University, Ottawa, Canada, in 1990, and M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, Berkeley, California, in 1993 and 1997, respectively. Since 1997, he has been an assistant professor in the Department of Electrical and Computer Engineering at Boston University, Boston, Massachusetts. Previously, he was with SONET Development Group of Bell-Northern Research, Ottawa, Canada, from 1990 to 1991. From 1992 to 1997 he was a research assistant at the University of California, Berkeley. Dr. Carruthers was an NSERC 1967 Postgraduate Scholar, and he received the National Science Foundation CAREER Award in 1999. His research interests include wireless infrared communications, wireless networking, and digital

communications. He is a member of the IEEE and the IEEE Communications Society.

BIBLIOGRAPHY

1. F. R. Gfeller and U. H. Bapst, Wireless in-house data communication via diffuse infrared radiation, *Proc. IEEE* **67**: 1474–1486 (Nov. 1979).
2. J. M. Kahn and J. R. Barry, Wireless infrared communications, *Proc. IEEE* **85**: 265–298 (Feb. 1997).
3. D. Heatley, D. Wisely, I. Neild, and P. Cochrane, Optical wireless: The story so far, *IEEE Commun. Mag.* 72–82 (Dec. 1998).
4. J. R. Barry, *Wireless Infrared Communications*, Kluwer, Boston, 1994.
5. J. R. Barry and J. M. Kahn, Link design for non-directed wireless infrared communications, *Appl. Opt.* **34**: 3764–3776 (July 1995).
6. R. Narasimhan, M. D. Audeh, and J. M. Kahn, Effect of electronic-ballast fluorescent lighting on wireless infrared links, *IEE Proc.-Optoelectron.* **143**: 347–354 (Dec. 1996).
7. International Electrotechnical Commission, *CEI/IEC 825-1: Safety of Laser Products*, 1993.
8. ANSI-Z136-1, *American National Standard for the Safe Use of Lasers*, 1993.
9. J. M. Kahn, W. J. Krause, and J. B. Carruthers, Experimental characterization of non-directed indoor infrared channels, *IEEE Trans. Commun.* **43**: 1613–1623 (Feb.–April 1995).
10. D. Lee and J. Kahn, Coding and equalization for PPM on wireless infrared channels, *IEEE Trans. Commun.* 255–260 (Feb. 1999).
11. D. Lee, J. Kahn, and M. Audeh, Trellis-coded pulse-position modulation for indoor wireless infrared communications, *IEEE Trans. Commun.* 1080–1087 (Sept. 1997).
12. H. Hashemi et al., Indoor propagation measurements at infrared frequencies for wireless local area networks applications, *IEEE Trans. Vehic. Technol.* **43**: 562–576 (Aug. 1994).
13. J. R. Barry et al., Simulation of multipath impulse response for indoor wireless optical channels, *IEEE J. Select. Areas Commun.* **11**: 367–379 (April 1993).
14. M. Abtahi and H. Hashemi, Simulation of indoor propagation channel at infrared frequencies in furnished office environments, *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications* (*PIMRC*) 1995, 306–310.
15. J. B. Carruthers and J. M. Kahn, Modeling of nondirected wireless infrared channels, *IEEE Trans. Commun.* 1260–1268 (Oct. 1997).
16. *http://www.irda.gov.*
17. IEEE Standard 802.11-1999, *Wireless LAN Medium Access Control* (*MAC*) *and Physical Layer (PHY) specifications*, 1999.
18. *http://www.canobeam.com.*
19. *http://www.terabeam.com.*
20. Special issue of *IEEE Communications Magazine* on Optical Wireless Systems and Networks (Dec. 1998).
21. *http://www.bu.edu/wireless/irguide.*

WIRELESS IP TELEPHONY

DAVID R. FAMOLARI
Telecordia Technologies
Morristown, New Jersey

1. INTRODUCTION

No two innovations have done more in the 1990s to advance communications than the wireless cellular telephone network and the Internet. Wireless cellular telephony extended access to the Public Switched Telephone Network (PSTN), the conventional landline telephone system, to users with small handheld radio terminals. These users could then establish and maintain voice calls anywhere within radio coverage, which for practical purposes in the United States is nationwide. This ushered in an era of *mobile* communications that disassociated a telephone user with a geographic location. No longer did dialing a particular number ensure that a caller would find (or not find) the intended recipient "at home," "at work," or elsewhere; with wireless telephony they could in fact be anywhere. And so could the caller. The desire to be mobile, coupled with falling subscription charges and rising voice quality, motivate an increasing number of customers to flock to the cellular telephone network. The reliability and quality of wireless phone service has even prompted some to forego landline service all together. Simply put, wireless telephony has made good on the promise of anywhere/anytime voice communications.

Based on the unifying, packet switching Internet Protocol (IP) [1], the Internet has had no less a dramatic impact on how society communicates. Its advent has interconnected devices over vast distances at limited costs and made possible rich multimedia content exchanges, most notably with the emergence of the World Wide Web (WWW). The open and standard protocols of the Internet allow equipment vendors to easily produce infrastructure products that interoperate with other vendor's products, increasing competition and fostering economies of scale. This has led to the pronounced deployment of Internet architecture throughout the world and, consequently, extended the reach of the Internet on a global scale. Because of its reach and its ability to support a variety of services, the Internet has evolved from a loose interconnection of computers employed by researchers to exchange files into a global network infrastructure that is an essential underpinning to commercial, governmental, and private communication.

Both the Internet and wireless telephony have created unprecedented operating freedoms and are forcing paradigm shifts in business and social communication practices. Their impact, however, has emerged from opposite design philosophies. The wireless telephone network provides ubiquitous access to speech applications; the Internet, on the other hand, delivers a wide variety of application types to fixed locations. The freedom and mobility offered by wireless networks provides a perfect compliment to the economies of scale and flexibility offered by the IP-based Internet.

Now, at the beginning of the twenty-first century, the two most defining communications breakthroughs of the late twentieth century are starting to merge. As IP becomes an important driver of both core and access networks, it must support wireless voice at comparable levels of spectrum efficiency and voice quality as cellular telephony. Similarly, wireless network practices must be modified to accommodate IP-based packet protocols. It is our aim here to describe the design challenges and technical hurdles facing successful deployment of wireless IP telephony.

2. BACKGROUND AND BASIC CONCEPTS

The design goals of the current breed of wireless telephony standards, commonly referred to as *second-generation* (2G)[1] *standards*, were to carry digital voice conversations to and from the PSTN with packet data networking only as an afterthought. As a consequence, current cellular systems have only limited packet data capabilities, usually to exchange brief text messages such as the Short Message Service (SMS). However new breeds of cellular communications systems are promising advanced features for packet data networking, along with faster transmission speeds and higher capacity. These new communications systems are widely referred to as *third-generation* (3G) *systems*. It is in 3G systems that the first steps are being taken towards integrating IP transport with wide-area wireless networks.

The new breed of 3G systems were initiated by the International Telecommunications Union (ITU), a United Nations governing body, under a directive called "International Mobile Telecommunications for the year 2000" (IMT-2000). IMT-2000 is a family of guideline specifications and recommendations serving as an umbrella standard to harmonize the evolution of the disparate 2G radio technologies currently in place; most notably the CDMA-based IS95 systems in the United States and the GSM systems in Europe and elsewhere. Regionalized activities within the IMT-2000 family emerged based on these two technologies. These include the Third Generation Partnership Project (3GPP), centered on evolving the GSM standard under the name Universal Mobile Telecommunications Service (UMTS) [2], and the Third Generation Partnership Project 2 (3GPP2), focused on evolving the IS95 standard, referred to as "cdma2000" [3].[2] Both groups are actively mapping out architectures and protocol references to support high-rate packet data services that will be used to carry IP telephony over their respective networks.

The challenges that these groups will face are to accommodate three fundamental issues in successful delivery of wireless voice. These issues are service quality of the perceived voice output stream, efficient use of the band-limited medium, and mobility of wireless users.

1. *Service Quality*. Wireless telephone systems have been efficiently designed to carry one type of application: voice, a time-sensitive, error-tolerant communication service. Its perceived quality degrades more rapidly when the delay characteristics of the underlying delivery change than as the error characteristics change. The destination of each call is usually a human being who can compensate for imperfections in the received signal. This is akin to deriving the proper meaning of a sentence when a word or two is mispronounced. When sounds arrive with noticeable variation in their inter−arrival times (called *jitter*), or when there are long lapses in the conversation (*delay*), conversation can become tedious. Circuit switched principles, at the core of wireless telephone systems, are well suited for voice since they offer timely and regular access to the transport medium. The Internet, on the other hand, thrives on the principles of packet switching. Consequently IP packets often take different routes through the network and can arrive out of sequence with variable delays. Therefore the basic IP protocols are not well suited to deliver time-sensitive applications such as voice.

2. *Efficiency*. Wireless telephony contends with limited spectrum and an unpredictable, time-varying physical channel that is subject to path loss, fading, and multiple-access interference. In order to support high capacity and revenue, service providers must make best utilization of their given spectrum. To this end, cellular telephony employs techniques that make best use of scarce, unreliable resources to deliver near-toll-quality voice. These techniques include employing low-bit-rate voice codecs to compress speech and reducing frame sizes to make voice packets less vulnerable to rapid fluctuations in the channel. The protocols and architecture of the Internet give prominence to end-to-end principles over centralized approaches. While offering flexibility, this demands that each packet contain a greater degree of control information than circuit-switched packets. The additional control information is embedded in the detailed protocol headers that are included in each IP data packet. Consequently, squeezing these heavyweight protocol headers into the relatively small packet sizes of the wireless channel gives rise to a number of performance problems related to efficiency.

3. *Mobility*. A fundamental design concept of wireless telephony is to allow users to seamlessly change their point of attachment to the network at any time. This must occur without explicit reconfiguration or significant performance loss. IP routing, however, associates an IP address with a fixed attachment to a router. When this association is no longer valid, as is the case when wireless users move, standard IP routing procedures are unable to deliver packets to and from the mobile terminal. Furthermore, re-establishing a valid association under the standard IP procedure may require manual intervention and explicit reconfiguration that will terminate any ongoing communications. Therefore these routing protocols require

[1] Second-generation systems represented a significant leap from the first generation Analog Mobile Phone System (AMPS), including the use of digital modulation techniques, enhanced security features, better spectral efficiency and longer mobile terminal battery life.

[2] For more detail on 3G wireless standards, the reader is referred to Ref. 4.

extensions and additional control architectures to transfer associations in midsession. In other words, mobility solutions for IP will need to offer seamless transfers in order to support wireless voice effectively.

In the remainder of this chapter we focus on each of these three principles and discuss modifications to the IP protocols and architecture to successfully offer IP telephony services.

3. SERVICE QUALITY FOR WIRELESS IP TELEPHONY

Service quality can be characterized by three parameters; packet loss, delay, and delay variability. In the wired network, heavy traffic at ingress points can overload IP routers, causing packets to be dropped. Even in the absence of heavy traffic, the detrimental effects of the wireless channel can corrupt voice packets so that they cannot be recovered at the receiver and are considered lost. Packet loss is therefore a considerable problem on wireless links. One-way delay is the time from when a voice packet enters the encoder at the source terminal to when it exits the decoder at the destination terminal. As mentioned earlier, the variability in packet arrival times is referred to as jitter. These three parameters are interdependent and all contribute to the overall service quality. In wireless networks where packet loss can be high, delay and jitter will increase correspondingly. Human perceptions can accommodate reasonable delays, but service quality becomes degraded when those delays have a high variation from packet to packet.

Wireless IP telephony service quality is dependent upon two issues: the service quality that voice receives over the air interface and the service quality that it receives while traversing the wired backbone network of the service provider. *Wired IP service quality* management has been a topic of much research. Many efforts have been geared toward supporting classes of service above and beyond the best-effort service of the typical Internet. These technologies attempt to provide more reliable delay and jitter performance by classifying traffic and giving priority treatment to certain traffic classes. We focus here on the service quality aspects particular to *wireless* transmission and refer the reader to Refs. 5–7 for more detailed information on wired service quality measures.

3.1. Wireless Service Quality

3.1.1. Link-Layer Solutions. Information transfer over wireless links is subject to impairments and environmental constraints that landline communications are free from. Factors affecting the wireless channel include interference, shadowing, path loss, and fading. These affects can have dramatic negative impacts on the signal-to-interference ratio (SIR). Lower SIR levels lead to increases in the frame error rates (FERs). Higher FER, in turn, increases packet loss. Furthermore, these wireless channel effects vary over small distances making signal quality very sensitive to terminal mobility. Simply increasing the bit rate cannot overcome the impediments of the wireless channel. Though all 3G systems will offer much higher available bandwidths for packet data [supporting rates of 144 kbps (kilobits per second) at high mobility rates, 384 kbps at pedestrian mobility rates, and 2 Mbps for indoor systems], voice service quality is still primarily dependent on achieving timely delivery of good-quality voice packets.

In order to improve the quality of the voice packets, wireless networks have focused on making voice packets more robust to transmission errors. In addition to physical-layer solutions that seek to improve the SIR performance of receivers, wireless system providers employ coding techniques such as forward error correction (FEC). FEC allows the receiver to locally reconstruct packets corrupted by bit errors without forcing retransmissions [8]. This scheme works by chopping frames into a smaller number of *codewords* and inserting a series of parity bits into each codeword that helps the receiver recover from a small number of bit errors; the greater the number of parity bits, the greater the recovery ability. Inherent in FEC, therefore, is a tradeoff between efficiency and resiliency. A resilient codeword contains more parity bits than does a less resilient one and thus has more transmission overhead, that is, less efficient use of the limited bandwidth. On the other hand, a less resilient code has a higher probability of unrecoverable error and can lead to retransmissions. This is detrimental in both efficient use of bandwidth and delay performance. Typical coding rates — the ratios of information bits to the total size of the codeword — used in wireless systems are usually on the order of $\frac{1}{4}$ to $\frac{2}{3}$. Some systems even offer dynamic FEC coding strategies that change the coding rate on the basis of the perceived error characteristics of the channel. The proper use of FEC techniques, therefore, can improve the integrity of the transmission sequence and reduce delays incurred on error-prone wireless links by reducing retransmissions.

The effectiveness of FEC coding is improved by interleaving [9]. Often wireless channels do not exhibit statistically independent error properties from one bit to the next. This means that bit errors tend to be lumped together in bursts. FEC codes can correct only a small number of errors per codeword. As such, they can be ineffective when dealing with very bursty error channels. Interleaving works to scramble the bit orderings before transmission and then place them back in proper order at the receiver before decoding. Thus bits that travel consecutively over the air will not be consecutive when presented to the FEC decoder. When deep fades corrupt a series of bits in transit, the interleaving process at the receiver will disperse those errors throughout the frame over multiple codewords, improving the effectiveness of the FEC scheme. This again implies that more packets can be corrected without retransmissions thereby reducing delay.

Another FEC technique employed in wireless voice systems to ensure good-quality characteristics is unequal error protection [10]. Unequal error protection is the practice of classifying certain bits in the voice frame as more important then other bits. If such essential bits are corrupted in transmission, then the frame cannot be used because too much fundamental information about the voice sample will have been lost. These bits, as a result,

are protected with FEC codes while the others are not. The nonessential bits in the voice sample can be delivered to the higher layers even if they contain bit errors. Although this reduces the sound quality of the voice sample, the end effect is still tolerable for the listener. The benefit of this approach is twofold: (1) the overhead associated with FEC coding is reduced and (2) usable (albeit less-than-perfect) voice information is delivered in cases where it would otherwise have been dropped or retransmitted.

In order to combat jitter and ensure smooth playback, real-time media systems buffer received voice packets. This solution to combat poor delay performance exists in both the wireless and wired telephone systems. With a buffered voicestream the receiver presents packets to the decoder at regular intervals even if they arrive at irregular intervals. Buffering at the receiver also helps establish correct packet ordering by allowing the receiver to collect a number of unordered packets and place them in the proper sequential order before they are played. This concept is illustrated in Fig. 1. Jitter buffers, however, add to the one-way delay already incurred due to packet encoding and transmission. Jitter buffer lengths must be designed with this delay penalty in mind. The International Telecommunications Union (ITU) states that one-way delay for voice samples must be lower than 400 ms for acceptable voice quality for almost all applications (with the exception of certain long-haul satellite communications). Furthermore, the ITU recommends that delays be below 150 ms [11]. Jitter buffer lengths are an important piece of the overall delay budget that includes codec, medium access, network, and transmission delays.

In addition to these techniques, all 3G systems will also have advanced methods for guaranteeing dataflows greater degrees of access to their air interfaces. Most notably these include forms of traffic classification and priority scheduling at the MAC layer that can dedicate resources to real-time packet streams [12]. These guarantees often provide a minimum bit rate and/or delay. Additionally, they provide intelligent voice-friendly queue management policies. The UMTS, in particular, has specified two real-time traffic classes, *conversational* and *streaming* [13]. The conversational class is the most delay-sensitive and is expected to handle wireless IP telephony. Certain deliverability attributes are associated with each class, such as maximum and guaranteed bit rates, maximum transfer delays, handling priorities, and whether packets with errors should be forwarded to higher layers. Using these attributes to define the service quality requirements for various traffic flows will allow 3G wireless systems to service time-sensitive voice packets before delay-tolerant data packets.

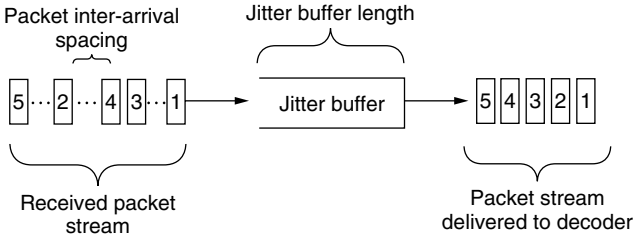


Figure 1. Jitter buffering.

3.1.2. Transport-Layer Solutions. While these link-level improvements will go a long way toward improving the service quality of wireless voice, performance is also highly dependent on the transport protocols used to deliver samples from the mobile terminal to their final destination. Reliable end-to-end transport protocols that promise ordered delivery of error-free packets, such as *Transmission Control Protocol* (TCP) [14], respond to link-level errors by requesting retransmissions. Such retransmissions hold up the timely delivery of subsequent packets. Furthermore, TCP operates on a principle that packet loss is due to congestion in the network and will thus respond to errors on the channel by generating less traffic. TCP will at first throttle the bit rate and then slowly increase it as the signs of congestion dissipate. This has dire consequences on the performance of voice applications by slowing down the service unnecessarily and imposing unacceptable delays. These issues are further exacerbated in wireless networks where link-level errors are frequent and seldom the result of congestion. It is clear that the application and performance characteristics of wireless voice are at odds with the design goals of TCP. The emphasis on timely delivery of voice over error-free reception cannot be reconciled with the TCP design philosophy, which sacrifices delays to achieve an error-free result.

As a result, the use of TCP to carry wireless voicestreams is not likely. User Datagram Protocol (UDP) [15], the connection-less alternative to TCP, in conjunction with the Real-Time Transport Protocol (RTP) [16], will be the most prominent implementation of voice over IP. UDP requires no retransmissions and is not session based, meaning that it carries no state and treats each packet individually independent of packets before or after it. This statelessness is not necessarily well suited for isochronous media where many packets are often carried with similar characteristics in a stream and it is advantageous to make use of state information. However, protocols such as RTP, which are used in conjunction with UDP, provide functionality to real-time voice above what simple UDP can provide.

RTP was developed to use the basic UDP datagram in order to provide end-to-end support for time-sensitive applications. RTP offers no congestion control and no promises of reliability; however, in the world of voice this is a boon. Without reliability and congestion control, the transport protocol will not mistake link-level errors for congestion and will not delay packets while waiting for retransmissions. End devices can use the information provided in RTP packets to determine the real-time characteristics of the received packet. RTP provides a timestamp field that allows receivers to determine whether an incoming packet is "fresh" and should be played or is "stale" and should be dropped. Sequence numbers are used to identify gaps in the reception of packets. This may signal that an error has occurred or that packets have been received out of order. RTP provides the necessary information to the receiver to

reconstruct the original stream when packets are received out of sequence. RTP is also particularly well suited for delivering multimedia traffic and provides mechanisms by which media streams can be synchronized and multiplexed, such as audio and video for a videoconference. RTP is by far the most frequently used protocol in the wired Internet for transferring real-time information. Because of its widespread use and special features designed for real-time traffic, RTP is the natural transport protocol for wireless IP telephony.

Wireless IP telephony providers must strive to offer service qualities that are comparable to the cellular voice services that customers are accustomed to. Strict requirements on delay, jitter, and packet loss create a challenge in ensuring service quality. In addition to intelligent service quality management of wired backbone links, this challenge will have to be met by coordinated efforts between radio-level mechanisms operating over the wireless link and transport protocols operating end to end.

4. EFFICIENCY AND WIRELESS IP TELEPHONY

Wide-area wireless channels typically employ bandwidths much smaller than those found on landline networks. Furthermore the unpredictable nature of the wireless channel makes the use of small link-level packet sizes advantageous. Small packet sizes are less vulnerable to channel fluctuations and help the radio network recover more gracefully from packet losses. As an example, typical cellular systems today, as well as future 3G systems, employ basic packet sizes that are on the order of 20 ms.

These small packet sizes represent a major challenge to the use of IP-type protocols that employ large headers. The resulting overhead can have detrimental effects on the efficiency of the system. Efficient use of wireless resources allows service providers to support higher capacities. For wireless IP telephony providers to match the level of spectral efficiency of traditional cellular networks, the resultant overhead of IP transport must be reduced. Header compression techniques are the most effective way to reduce IP packet overheads, and several such compression schemes have been defined for compressing a variety of protocol types. Below we discuss the most relevant aspects of header compression to wireless IP telephony.

4.1. Header Compression

As indicated earlier, voice over IP networks will be supported by the Real-Time Transport Protocol, which runs over UDP. Thus the typical protocol layering for IP voice looks like RTP/UDP/IP.[3] Each of these protocols introduces its own overheads in the form of required headers. Typically this value totals 40 bytes, including 20 bytes for RTP, 8 bytes for UDP, and 12 bytes for IP. The protocol header fields for RTP/UDP/IP are shown in Fig. 2, where the numbers represent bit positions and are used to indicate the length of the header fields. We discuss only a few of the header fields below; for more detailed information on the RTP/UDP/IP header fields the reader is referred to the literature [1,15,16].

This 40-byte RTP/UDP/IP overhead represents a significant portion of available wireless bandwidth. Current wireless voice packetization schemes, including

[3] This is taken to mean that RTP is on top of UDP, which is on top of IP, indicating that a packet must travel down the protocol stack, traversing first the RTP layer, then the UDP and IP layers. Other references write the protocol layering as IP/UDP/RTP, indicating that IP is outside UDP that is outside RTP. This latter notion emphasizes that IP protocol headers are followed by UDP and RTP protocol headers.

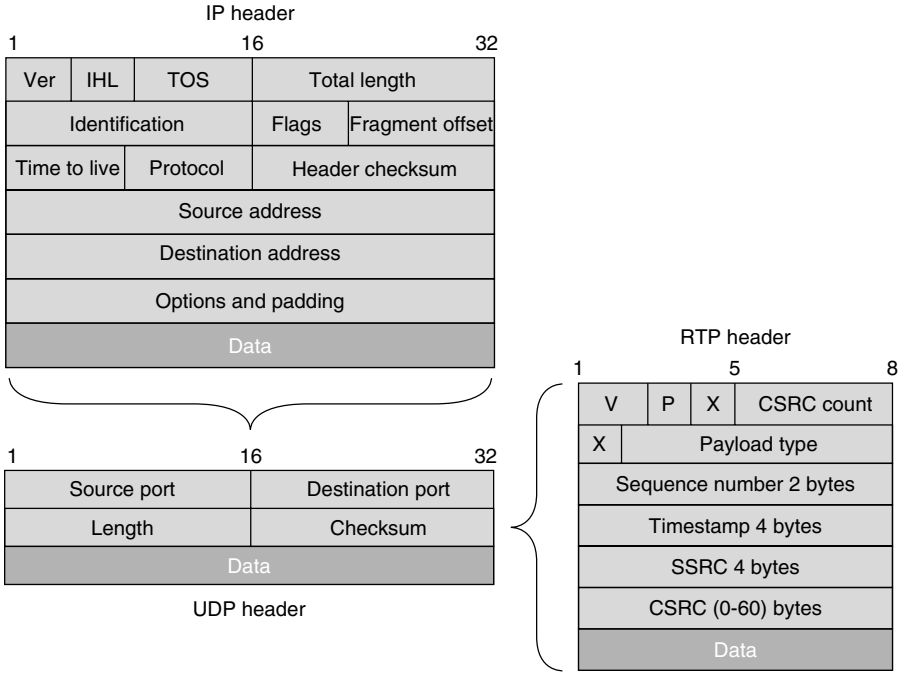


Figure 2. RTP/UDP/IP protocol headers.

the ITU standard G.729 [17], employ 8-kbps codecs where 20 bytes of voice information is sent every 20 ms. Therefore voice packets carried under the RTP/UDP/IP regime will incur a 200% overhead price. Clearly deep RTP/UDP/IP header compression is required for wireless IP telephony to achieve the efficiency that cellular operators need to service a growing number of subscribers.

4.1.1. Compression State. Header compression works on the principle that many RTP/UDP/IP packet header fields change either predictably or very infrequently from packet to packet. For example, the IP addresses of the two correspondents never change during the course of a typical session; however, the 64 bits (8 bytes) used to denote source and destination IP addresses, is included in each UDP header. A similar situation exists for the source and destination port addresses, which account for 2 bytes. Addressing and port assignments, which remain largely static over the lifetime of a call, account for roughly 25% of the total overhead. Header compression schemes leverage the predictability of packet headers from packet to packet to achieve compression levels on the order of 95–97%; in some instances they reduce 40-byte overhead to 1 or 2 bytes.

Header compression schemes are able to achieve these levels of compression by first eliminating well-known a priori information or information that can be inferred from other mechanisms, such as the link layer. Fields such as the IP and RTP version numbers are expected to be well known, other values such as the IP header and payload lengths can be successfully inferred from the link layer. Moreover, sending non–a priori, but static, information only once at the onset of the session reduces overhead considerably. Values such as the abovementioned 8-byte IP addresses and 2-byte port numbers, among others, can be sent once and will remain constant over a large number of packet headers. Sending these values once helps the compressor and decompressor establish a *context*, or compression state, by which future compressed packets can be evaluated. Compression state is the knowledge necessary to reconstitute the full header from a compressed header. After the initial sending, it is necessary to only provide delta values, or updates, in header fields that change instead of the absolute values. This significantly reduces the amount of overhead transmission.

Figure 3 shows the general architecture of a header compression scheme where a *compressor* takes full RTP/UDP/IP headers and generates compressed headers that are sent over the wireless channel. On the receiving side the decompressor attempts to reconstruct the original header by applying the corresponding decompression scheme using the current header context. When the context becomes lost or loses synchronization between the compressor and the decompressor, as is the result of link-layer errors, the compression scheme must be able to restore context.

4.1.2. Error Recovery. An important consideration in the design of a header compression scheme for the wireless environment is its ability to recover from errors. Cassner

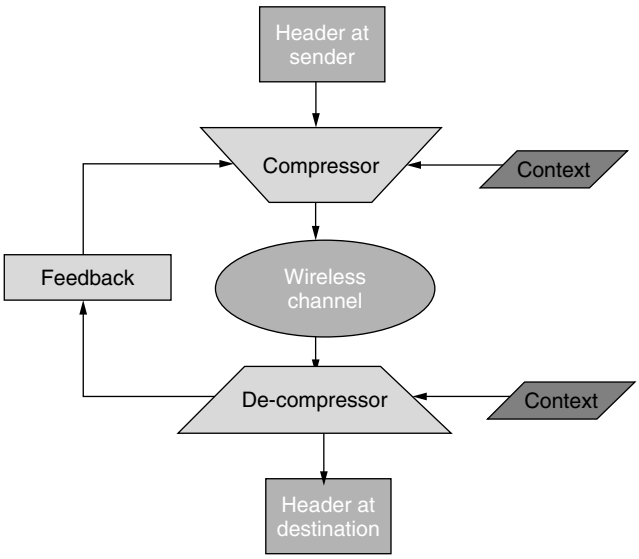


Figure 3. Header compression architecture.

and Jacobson proposed the earliest header compression technique for RTP/UDP/IP, *compressed RTP* (CRTP), [18] which became a draft standard in 1999. While this scheme worked well for telephone dialup connections and other low-loss link layers, its design did not perform well in a highly variable and error-prone wireless channel. When packets were received in error, the header compression states at the compressor and the receiver would lose context and full packet headers had to be transferred to reestablish synchronized header state. The effect of long round-trip times, as is the case with wireless IP telephony, has dramatic effects on the error recovery performance. When links have long round-trip times the context cannot be regained as quickly and many compressed packets will require retransmissions or be dropped. Performance lags in CRTP over wireless links became evident [19], and efforts were made to design more robust header compression schemes that could adapt to imperfections in the channel and gracefully recover from errors.

4.1.3. Robust Header Compression. To meet the error performance requirements necessitated by volatile wireless links, header compression schemes have to be reliable and robust. The *ro*bust *h*eader *c*ompression (ROHC) scheme [20] was created to make header context less sensitive to packet loss and delay as well as make context recovery faster. This approach repairs context locally, thereby eliminating the need to send update information over the wireless link.

The key factor in ROHC is that cyclic redundancy check (CRC) codes are computed on the uncompressed headers and are sent along with the compressed header. CRC codes are the result of passing a string of bits through a generator polynomial. The CRC codes are then sent along with the bits used to generate the codes. The receiver applies the same function to the received bit string, generating a local copy of the CRC code. If the local copy and the received CRC code are not equal, the receiver is sure that a transmission

error has occurred. CRC codes provide a reliable error-detection mechanism. The decompressor, after generating its copy of the full header from the received compressed header, will perform a CRC check on the full header. This allows the decompressor to reliably determine whether the decompression process was successful. In addition to being able to repair context locally, ROHC is also capable of withstanding a number of consecutive packet errors without losing context. This is important as wireless environments seldom have bit-independent channels and errors often arrive in bursts. ROHC can support upwards of 24 consecutive packet errors without losing context [21].

4.1.4. Performance of Header Compression. Header compression schemes can be evaluated along three distinct performance attributes: compression efficiency, robustness, and compression reliability. In terms of compression efficiency, ROHC can optimally reduce the 40-byte RTP/UDP/IP header to a single byte. Under bit error rates typical of wireless channels, it can achieve an average of 2.27 bytes of overhead [22]. Furthermore, compression efficiency is enhanced by the ability to restore context locally, reducing the transmission of noncompressed headers over the wireless link. Compression reliability, the ability to ensure that decompressed headers are accurate representations of the uncompressed headers, is achieved through the use of CRC codes. This provides a highly reliable way for the decompressor to determine whether the decompressed packet is correct. Finally, ROHC provides a high degree of robustness by correcting loss of context locally and maintaining context in the presence of multiple consecutive errors. Performance results [22] show that under a simulated WCDMA channel operating at a BER of 0.0002 the FER achieved with CRTP is 1.10% while the FER achieved with ROHC was 0.12%. At a higher BER value of 0.001, the frame error rates were 4.06% and 0.81% for CRTP and ROHC, respectively.

It is clear that this new class of robust header compression will be critical to improving the efficiency and performance of wireless IP telephony.

5. MOBILITY

Mobility creates problems with IP routing protocols and can break ongoing sessions. However, wireless IP telephony must be as seamless as present cellular telephony. This requires the ability to change points of attachment to the wireless network while maintaining connectivity with minimal disruption. Cellular solutions address mobility by monitoring channel assignments, code allocations, and received power levels. The introduction of IP transport, however, requires additional mobility solutions that are not addressed by the traditional link-layer cellular mobility techniques. These requirements for *network* mobility extend beyond the *physical*, or *link-level*, mobility offered in cellular networks. The basic functions needed to support mobile access to IP-based networks include

- *Detection* — terminals learn when they have entered new network areas.
- *Registration* — users indicate their presence and requirements to the network.
- *Configuration* — network adapts nodes to the particular network characteristics, including IP address assignment and configuration of the default router.
- *Authentication, authorization, and accounting* (AAA) — validate users and their permission and record their usage for billing and management purposes.
- *Dynamic address binding* — provides a dynamic mapping of old network addresses with new network addresses.

To become a full network participant a user must first have means of physically detecting and connecting to a network. This entails establishing a valid link with the appropriate physical layer protocols, after which a terminal can format information in a contextually meaningful manner. When basic connectivity has been established the mobile and the network can begin to perform parameter negotiations. This occurs during the registration process where the network learns of a terminal's presence and requirements.

Configuration involves fulfilling any registration requests and providing information to enable the mobile to properly orientate itself to the new network surroundings. This may include assigning a new IP address and passing the locations of default routers and network servers. After configuration the next step is for the network to grant the user access to network resources based on AAA measures. The network arrives at these decisions based on negotiations using credentials passed between the terminal and/or user, either explicitly or implicitly. Additionally the network may validate these credentials with third parties located in outside networks. Finally, once the user and terminal have been properly authenticated, dynamic address binding creates an association between the new configuration and the old configuration. This allows mobile terminals to be found after they change networks and allows active sessions to be maintained across different points of attachment transparently.

5.1. IP Mobility for Wireless Telephony

Current cellular networks have mechanisms in place to successfully support link-level mobility. Additionally roaming agreements between service providers allow mobile subscribers from one provider to access another provider's network. In this sense, current cellular networks already have in place mechanisms to support registration, configuration, and AAA functions associated with mobility. However, since mobile telephone numbers are constant and do not change, there is no need to perform dynamic address binding. Also since addressing is valid over the entire network, there is no need to detect when an address change is required. Therefore detecting when address changes are required and dynamically binding those addresses represents a major challenge for current cellular networks to provide mobility.

Supporting dynamic address binding for wireless IP telephony users poses some unique challenges. The IP protocol inherently links physical location with network

representation. In other words, an IP address represents a host's physical location on the network as well as its identity within the network. IP uses this association to route packets efficiently to destination addresses. When this association is broken, or invalid, packets can no longer reach their destinations.

5.2. Mobile IP

The industry standard regarding IP mobility is called *mobile IP* [23] and is actively being designed into the all-IP architectures of next-generation networks including the 3GPP and the 3GPP2 [24]. Mobile IP creates a level of indirection within the network so that on-going communications are not interrupted due to the IP address changes of mobile terminals. This is achieved by associating two addresses for the mobile terminal: a permanent *home address* that represents the terminal's IP address within its home network and a temporary locally assigned *care-of address* that is valid within the visited network. A *home agent* (HA) located in the terminal's home network keeps an association between a terminal's home address and its care-of address. A *foreign agent* (FA) in the visited network provides routing and support services to visiting mobile terminals. The elements of a typical Mobile IP architecture are depicted in Fig. 4.

Both foreign and home agents can advertise their presence on their respective local networks by issuing agent advertisement messages that inform terminals of their availability. Likewise, a mobile terminal may solicit these messages by broadcasting agent solicitation messages on entering a new network to learn about that network's mobility support. A mobile terminal can then use these advertisement messages to detect if it has migrated into a new IP subnet.

Once a mobile terminal learns that it is on a new IP subnet, it will attempt to obtain a care-of address for the visited subnet. This can be done in one of two ways. It may be given a care-of address by the FA, called a *foreign agent care-of address*, which is associated with a network interface on the FA. The mobile terminal may also obtain a *collocated care-of address* associated with one of its own network interfaces. After receiving the care-of address, the mobile terminal will then register this address with its HA via a registration request message. When this registration is accepted, the HA will respond with a registration response message and store the association between the mobile terminal's home address and care-of address.

Packets that are destined for the mobile terminal, that is, those that have the mobile terminal's home address in the destination field of the IP header, always arrive inside the mobile terminal's home network and are intercepted by the HA. The packets are then *tunneled* to the mobile terminal by the HA. The tunneling process involves encapsulating [25] the sender's original IP packet inside the body of another IP packet generated by the HA that contains the mobile terminal's care-of address in the destination field. Since the outer header of the encapsulated IP packet contains the care-of address, it will be forwarded to the visited network where the mobile terminal currently resides. When foreign agent care-of addresses are used, the tunneled packets arrive at the FA who decapsulates them by stripping off the encapsulated packet header and sends the original IP packet to the mobile terminal. Mobile terminals that have collocated care-of addresses will receive incoming encapsulated packets directly and be responsible for decapsulating them.

Bandwidth-constricted wireless networks will most likely support the foreign agent care-of address model, as is the case with the 3GPP and 3GPP2. This mode of operation is more spectrally efficient since only the original IP packet, and not the extra headers associated with the encapsulated packet, is sent over the air interface. Additionally this mode conserves IP address space; one FA can service multiple mobile terminals, and therefore only one care-of address per FA is needed.

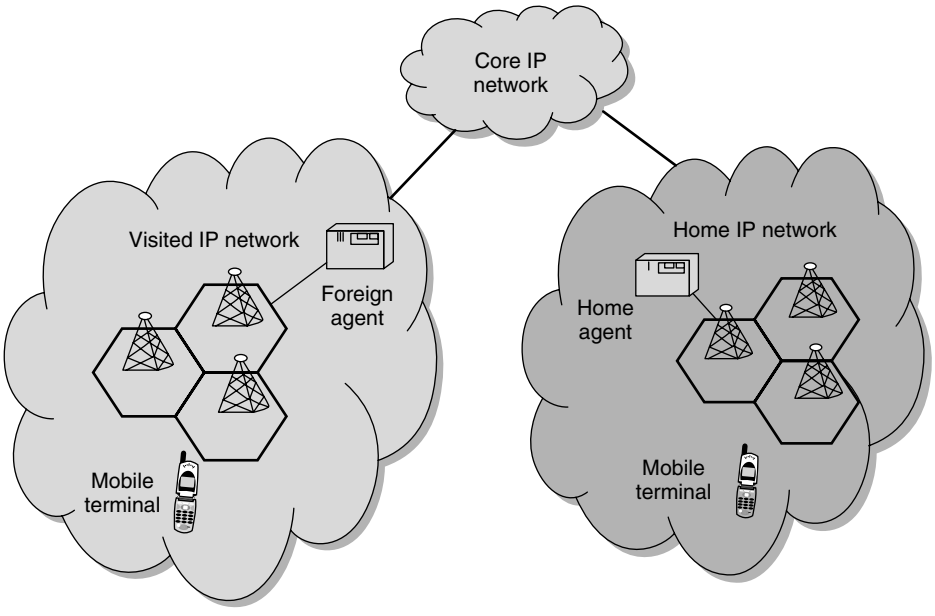


Figure 4. Mobile IP architecture.

The mobile terminal can send outbound IP packets on the visited network using normal IP routing without any modifications. The mobile terminal will insert its home address into the source address field of all IP data packets it generates. The mobile IP agents effectively hide mobility events from correspondent hosts so that IP address changes are transparent. Thus correspondent hosts are completely unaware that the mobile terminal has moved. This feature of mobile IP allows all network terminals, regardless of whether they support mobile IP, to communicate with mobile terminals that do.

5.2.1. Route Optimization. Mobility implementations in wireless IP telephony are judged on their ability to provide seamless handoffs with minimal interruption to user sessions. Packet loss and handoff delay therefore must be minimized in order to provide quality voice service. A vulnerability of the basic mobile IP approach discussed above is that traffic streams are required to go to the mobile terminal's HA and then to the mobile terminal, creating what is called the *triangular routing problem*. This is particularly problematic when a mobile terminal is very far from its home network and is communicating with a correspondent host local to the visited network as shown in Fig. 5. As an example, consider a New Yorker visiting a friend in Los Angeles. Packets from the friend's terminal would have to travel across the country to the HA in New York and then be tunneled back to the present location of the mobile terminal in Los Angeles. This needlessly introduces two cross-country trip times. In addition, it consumes unnecessary resources within the wide-area network, especially when compared to direct delivery on the local Los Angeles network. Efforts within the mobile IP community have addressed this problem by creating a modified mobile IP approach that uses route optimization [26] and eliminates unnecessary triangular routing.

Route optimization works by making correspondent hosts aware of the current care-of address of the mobile terminal. The host then stores those associations in a binding cache. Home agents, on receipt of a packet destined for a mobile host in a visited network, will send the originator of that packet a message containing the mobile terminal's current care-of address. The originator will then store this association in its binding cache and can begin to send packets directly to the mobile host without unnecessary involvement of the HA. This, in turn, reduces latency and frees network resources. Implementing route optimization in mobile IP can help drastically eliminate latencies and improve quality for wireless IP telephony.

5.3. Mobility Architectures

A design challenge for implementing mobile IP in wireless environments is to define proper placement and demarcations of areas serviced by the mobile IP elements. This helps balance signaling overhead and delay while allowing seamless connectivity. Advertisement and solicitation messages generated in mobile IP are a way of determining whether new physical connections

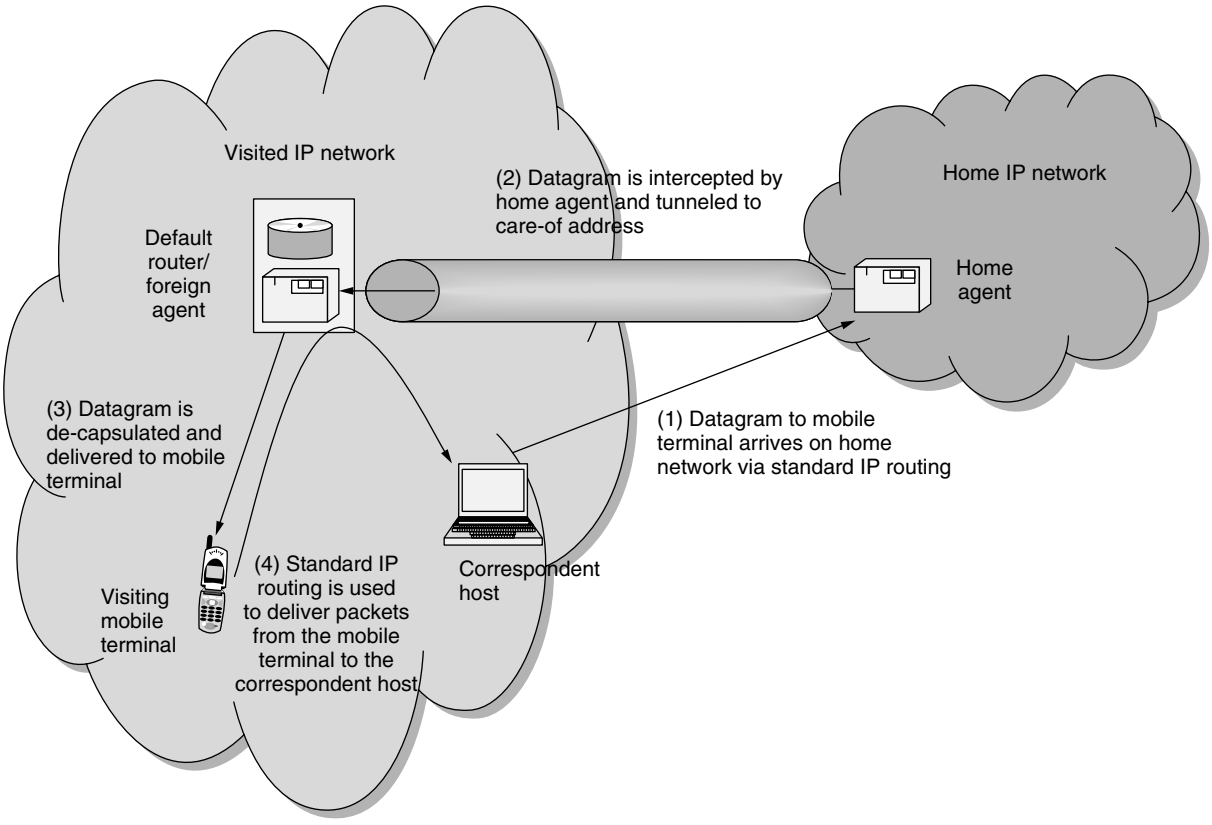


Figure 5. Mobile IP triangular routing.

require IP-level mobility procedures. However, when cell radii decrease and terminals travel at higher speeds, the mobility rate increases. This triggers more and more solicitation and registration messages that could begin to have a detrimental overhead effect in the network. Furthermore, the registration messages must travel to the mobile terminal's home network and may introduce unacceptable delays in establishing new network connections. Designing wireless networks with subnets that do not cover a lot of area may lead to an undesirable amount of signaling and delay in current mobile IP implementations.

Additionally, mobile agents need to balance the frequency with which they broadcast advertisements to suit the signaling capabilities of the wireless network. More frequent signaling allows for faster detection of network-layer mobility and therefore shorter handoff times. It comes, however, at the price of greater signaling overhead. There exists a design tradeoff that trades signaling overhead for responsiveness of the mobility protocol. As radio-level resources are at a premium, optimal solutions will employ as little signaling overhead as possible to achieve the required level of responsiveness.

Industry efforts have been focused on this problem, and new breeds of mobility strategies have emerged that attempt to reduce the delays and overhead caused by excessive signaling and frequent mobility [27–29]. Many of these strategies introduce levels of hierarchy so that registration messages do not need to travel all the way to the home network every time there is a mobility event. These types of strategies help reduce the latencies and packet losses due to IP mobility.

6. CONCLUSION

Since 1980 wireless voice service has grown into a reliable mainstay for personal and business communications. In the same period the Internet has flourished to unprecedented levels; enjoying economies of scale and ease of deployment. New wireless network architectures will take advantage of the service and management flexibility offered by IP, allowing service providers to offer multimedia and data content to their wireless subscribers. As a consequence, voice strategies must be amended from their traditional circuit-switched roots to perform comparably over IP. The greatest challenges facing the successful deployment of wireless IP telephony are threefold: securing reliable guarantees of service quality on par with traditional cellular systems, obtaining spectral efficiencies over the wireless channel that will not hinder system capacity or service quality, and effectively providing seamless connections to mobile users.

BIOGRAPHY

David Famolari received his B.S and M.S degrees in electrical engineering from Rutgers University, New Jersey, in 1996 and 1999 respectively. In 1996, he joined the Wireless Information Network Laboratory (WINLAB), at Rutgers University, as a research assistant where he worked on radio resource management protocols and parameter optimizations for third generation (3G) cellular systems. Since 1998, he has been a member of the Applied Research Department at Telcordia Technologies, Morristown, New Jersey, where he has worked on emerging mobile computing technologies, wireless networking protocols, and residential networking. David was awarded the Telcordia Technologies CEO Award in 2000 for his contributions in wireless IP networking. He is currently the cochair of the Open Services Gateway Initiative (OSGi) Device Expert Group, a leading industry consortium producing open specifications to promote the delivery of broadband services into home, automotive, and other similar networks. His current research interests include wireless local area network (WLAN) technologies and systems, mobility management, wireless computing, and personal area networks and systems. He can be reached by e-mail at fam@research.telcordia.com.

BIBLIOGRAPHY

1. J. Postel, *Internetwork Protocol*, RFC 791, Sept. 1981.
2. H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access For Third Generation Mobile Communications*, Wiley, New York, 2000.
3. TIA/EIA/IS-2000-2-A, *Physical Layer Standard for cdma2000 Spread Spectrum Systems*.
4. Special Issue, IMT-2000: Standards Efforts of the ITU, *IEEE Pers. Commun.* **4**(4): (Aug. 1997).
5. S. Blake et al., *An Architecture for Differentiated Services*, RFC 2475, Dec. 1998.
6. V. Jacobson, K. Nichols, and K. Poduri, *An Expedited Forwarding PHB*, RFC 2598, June 1999.
7. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, *Assured Forwarding PHB Group*, June 1999.
8. D. J. Goodman and C. E. Sundberg, Transmission errors and forward error correction in embedded differential PCM, *Bell Syst. Tech. J.* **62**: 2735–2764 (Nov. 1983).
9. S. H. Lim, D. M. An, and D. Y. Kim, Impact of cell unit interleaving on header error control performance in wireless ATM, *Proc. IEEE GLOBECOM'96*, Nov. 1996, pp. 1705–1709.
10. A. Nazer and F. Alajaji, *Unequal Error Protection and Source Channel Decoding of CELP Speech Over Very Noisy Channels*, Technical Report, Mathematics and Engineering Communications Laboratory, Queens Univ., 1999.
11. International Telecommunications Union, *One-Way Transmission Time*, Recommendation G.114, June 1996.
12. C. Comaniciu, N. B. Mandayam, D. Famolari, and P. Agrawal, QoS guarantees for third generation (3G) CDMA systems via admission and flow control, *Proc. IEEE Vehicular Technology Conf.* (*VTC*), Boston, Sept. 2000.
13. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects, *QoS Concept and Architecture*, 3GPP TS 23.107 v3.4.0.
14. J. Postel, ed., *Transmission Control Protocol*, RFC-793, Sept. 1981.
15. J. Postel, *User Datagram Protocol*, RFC-768, Aug. 1980.
16. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, RFC 1889, Jan. 1996.

17. ITU-T, *Coding of speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction* (*CS-ACELP*), Recommendation G.729, 1996.

18. S. Casner and V. Jacobson, *Compressing IP/UDP/RTP Headers for Low-Speed Serial Links*, RFC 2508, Jan. 1999.

19. M. Degermark, H, Hannu, L. Jonsson, and K. Svanbro, Evaluation of CRTP performance over cellular radio links, *IEEE Pers. Commun.* (Aug. 2000).

20. C. Bormann et al., *RObust Header Compression* (*ROHC): Framework and Four Profiles: RTP, UDP, ESP, and Uncompressed*, RFC 3095, July 2001.

21. L. Jonsson, M. Degermark, H. Hannu, and K. Svanbro, RObust checksum-based header COmpression (ROCCO), Internet draft (June 2000), *<draft-ietf-rohc--rtp-rocco-01.txt>*.

22. H. Hannu, K. Svanbro, and L.-E. Jonsson, ROCCO performance evaluation, Internet draft (May 2000), *<draft-ietf-rohc-rtp-rocco-performance-00.txt>* (for performance results in ROHC).

23. C. Perkins, *IP Mobility Support*, RFC 2002, Oct. 1996.

24. G. Patel and S. Dennet, The 3GPP and 3GPP2 movements toward an all-IP mobile Network, *IEEE Pers. Commun.* (Aug. 2000).

25. C. Perkins, *IP Encapsulation within IP*, RFC 2003, Oct. 1996.

26. C. Perkins and D. Johnson, Route optimization in Mobile IP, Internet draft (Sept. 2001), *<draft-ietf-mobileip-optim-11.txt>*.

27. E. Gustafsson, A. Jonsson, and C. Perkins, *Mobile IP Regional Registration*, Internet draft (March 2001), *<draft-ietf-mobileip-reg-tunnel-04.txt>*.

28. A. Campbell et al., Cellular IP, Internet draft (Dec. 1999), *<draft-ietf-mobileip-cellularip-00.txt>*.

29. R. Ramjee et al., IP micro-mobility support using HAWAII, Internet draft (June 1999), *<draft-ietf-mobileip-hawaii-00.txt>*.

WIRELESS LAN STANDARDS

RICHARD VAN NEE
Woodside Networks
Breukelen, The Netherlands

1. INTRODUCTION

Since the early 1990s, wireless local-area networks (WLANs) for the 900-MHz, 2.4-GHz, and 5-GHz ISM (industrial–scientific–medical) bands have been available for a range of proprietary products. In June 1997, the Institute of Electrical and Electronics Engineers approved an international interoperability standard (IEEE 802.11 [1]). The standard specifies both medium-access control (MAC) procedures and three different physical layers (PHY). There are two radio-based PHYs using the 2.4-GHz band. The third PHY uses infrared light. All PHYs support a data rate of 1 Mbps (megabit per second) and optionally 2 Mbps. The 2.4 GHz band is available for license exempt use in Europe, the United States and Japan. Table 1 lists the available frequency bands and the restrictions to devices that use this band for communications.

Table 1. 2.4- and 5-GHz Bands

Location	Regulatory Range (GHz)	Maximum Output Power
North America	2.400–2.4835	1000 mW
Europe	2.400–2.4835	100 mW (EIRP[a])
Japan	2.400–2.497	10 mW/MHz
USA (UNII lower band)	5.150–5.250	Minimum of 50 mW or $4\ \text{dBm} + 10\log_{10} B$[b]
USA (UNII middle band)	5.250–5.350	Minimum of 250 mW or $11\ \text{dBm} + 10\log_{10} B$
USA (UNII upper band)	5.725–5.825	Minimum of 1000 mW or $17\ \text{dBm} + 10\log_{10} B$

[a] EIRP = effective isotropic radiated power.
[b] B is the −26-dB emission bandwidth in MHz.

User demand for higher bit rates and the international availability of the 2.4-GHz band has spurred the development of a higher-speed extensions to the 802.11 standard. In 1999, the IEEE 802.11b standard was finished, and describes a PHY providing rates of 5.5 and 11 Mbps [2]. IEEE 802.11b is an extension of the direct-sequence 802.11 standard, using the same 11 MHz chip rate, such that the same bandwidth and channelization can be used.

In parallel to IEEE 802.11b, the IEEE 802.11a standard was developed to provide high bit rates in the 5-GHz band. This development was motivated by an amendment to Part 15 of the U.S. Federal Communications Commission in January 1997. The amendment made available 300 MHz of spectrum in the 5.2-GHz band, intended for use by a new category of unlicensed equipment called "unlicensed national information infrastructure" (UNII) devices. Table 1 lists the frequency bands and the corresponding power restrictions.

In July 1998, the IEEE 802.11 standardization group decided to select orthogonal frequency-division multiplexing (OFDM) [3] as the basis for their new 5-GHz standard, targeting a range of data rates from 6 to 54 Mbps. This standard is the first one to use OFDM in packet-based communications, while the use of OFDM previously was limited to continuous transmission systems like digital audiobroadcasting (DAB) and digital videobroadcasting (DVB). Following the IEEE 802.11 decision, the European HIPERLAN type 2 [4] standard and the Japanese Multimedia Mobile Access Communication (MMAC) standard also adopted OFDM. The three bodies have worked in close cooperation since then to minimize differences between the various standards, thereby enabling the manufacturing of equipment that can be used worldwide.

Regulatory issues played an important role in the development of wireless LAN standards. One of the key factors in the choice of modulation schemes for the 2.4-GHz band has been the FCC spreading requirement for unlicensed devices in the ISM bands, where wireless LANs

are predominantly used. According to the FCC spreading rules, transmission in the ISM bands have to use either direct sequence, spread spectrum, or frequency hopping. Frequency-hopping devices have to use at least 75 hopping channels with a maximum dwell time of 400 ms. Direct-sequence devices have to demonstrate at least 10 dB processing gain in a narrowband jammer test, which basically shows that there is a gap of at least 10 dB between signal-to-noise ratio and signal-to-interference ratio requirements for a certain bit error ratio. In the early days of wireless LAN, many people interpreted the spreading rule as a requirement for at least 10 chips per symbol; hence the 11 chips spreading sequence in the 802.11 standard. Later, a less strict interpretation was adopted, purely based on meeting the narrowband jammer test. This is clearly visible in the IEEE 802.11b standard. The 802.11b standard uses complementary code keying (CCK), which can be viewed as direct-sequence spread-spectrum modulation with multiple spreading codes with a length of 8 chips. Despite the less strict interpretation, the spreading rule formed a barrier for really high data rates. It blocked the use, for instance, of OFDM in the 2.4-GHz band. In order not to avoid further technological progress in the 2.4-GHz band, in May 2001 the FCC decided to allow digital transmissions without any spreading requirement [5]. This opened the way to higher data rates using OFDM in the 2.4-GHz band. The 802.11 committee took advantage of this rule change by selecting the OFDM based 802.11a standard as basis for the 802.11g standard, extending the data rates in the 2.4-GHz band up to 54 Mbps.

In the following sections we describe the various IEEE wireless LAN standards, and mention the differences with HIPERLAN and MMAC. Because of length limitations, the scope of this article is restricted to the most predominantly used parts of the standards. More details can be found in the references listed at the end of this article.

2. IEEE 802.11 MAC

The IEEE 802.11 MAC standard consists of one mandatory and two optional modes [1]. All modes use time-division duplex (TDD), so the medium is shared in time between different users and/or access points. The mandatory part is the *distributed coordination function* (DCF), which uses carrier sense multiple access with collision avoidance (CSMA/CA). Figure 1 shows the timing diagram of a DCF packet transmission. Before starting a transmission, the channel is sensed to see if it is available. If no other signal is received above a certain defer threshold, a packet is send. After successful reception, the recipient sends an acknowledgment back. After receiving the acknowledgement, the first user has to wait for a time DIFS plus a random backoff time before transmitting another packet. DIFS is the *distributed interframe spacing*, which is equal to the *short interframe spacing* between packet and acknowledgment plus 2 slot times.

Optional modes in the 802.11 MAC are the request-to-send/clear-to-send (RTS/CTS) protocol, and the point coordination function (PCF). PCF is a centralized MAC, where an access point polls stations to see if they have packets to transmit. PCF can be used to guarantee a minimum packet delay, but it can do this only in the absence of interference from other cells. With the RTS/CTS protocol, prior to a data packet, first a short request packet is sent. The receiver answers with a CTS packet, which contains a net allocation vector (NAV) that tells all users how long the current RTS/CTS cycle will take. The effect of this is that all users that can receive the CTS packet will not try to try to compete for the channel for the duration indicated by the NAV. This solves the hidden-node problem of DCF without RTS/CTS, because in that case, only users that can receive the transmitter will stop competing for the channel. So, it can happen that a packet from user A to B is interfered by user C, who does not have a good link to A, but it does have a good link to B. With RTS/CTS, this situation is avoided, because user C will hear the CTS coming from B.

3. IEEE 802.11 DSSS

The IEEE 802.11 Direct-Sequence Spread-Spectrum standard is based on the transmission of 11-chip Barker codes at a 11 MHz chip rate. Data rates of 1 and 2 Mbps are achieved using BPSK or QPSK modulation of the Barker codes, respectively. The 11-chip Barker code is defined as $\{1, -1, 1, 1, -1, 1, 1, 1, -1, -1, -1\}$. Its primary use is to satisfy the FCC spreading requirements, as well as providing robustness against multipath propagation and narrowband interferers. Robustness against multipath is obtained by the ideal aperiodic autocorrelation properties that define a Barker code — a *Barker code* is a code for which the absolute autocorrelation sidelobes are equal to or less than one (≤ 1) for all nonzero delays, compared to L for a zero delay, where L is the codelength. Because of the low-autocorrelation sidelobes, effects of intersymbol interference are greatly suppressed, while a simple RAKE receiver is able to significantly benefit from multipath diversity in frequency selective channels.

The 802.11 packet structure is shown in Fig. 2. The complete packet (PPDU) has three segments. The first segment is the preamble, which is used for signal detection and sychronization. The second segment is the header, which contains data rate and packet length information. The third segment (MPDU) contains the information bits.

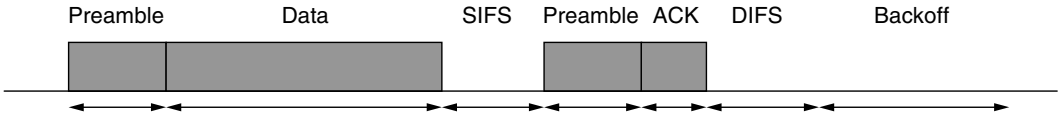


Figure 1. Timing diagram of a single-packet transmission using DCF.

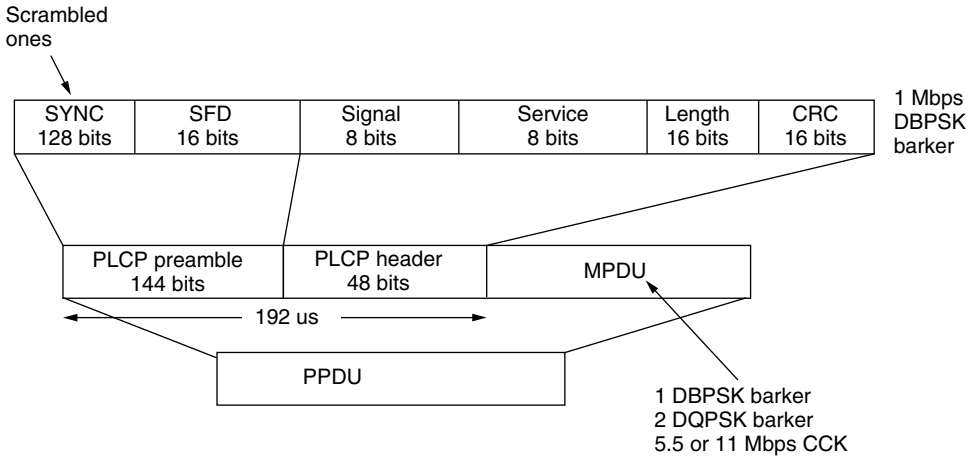


Figure 2. The packet structure used for 802.11 DSSS 1 and 2 Mbps, with the extension to 5.5 and 11 Mbps shown.

The preamble and header are transmitted at 1 Mbps, while the data portion is sent at one out of four possible rates.

The preamble is formed from a SYNC field and a SYNC field delimiter (SFD). The SYNC field is generated using 128 scrambled ones. The SYNC field is used for clear channel assessment, signal detection, timing acquisition, frequency acquisition, multipath estimation, and descrambler synchronization.

4. IEEE 802.11b

In July 1998, the IEEE 802.11b working group adopted complementary code keying (CCK) as the basis for the high-rate physical-layer extension to deliver data rates of 5.5 and 11 Mbps [6]. This high-rate extension was adopted in part because it provided an easy path for interoperability with the existing 1- and 2-Mbps networks by maintaining the same bandwidth and utilizing the same preamble and header as shown in Fig. 1. An optional short preamble with a 56-bit SYNC field is specified to increase the net data throughput.

Complementary codes were originally conceived by M. J. E. Golay for infrared multislit spectrometry [7]. However, their properties also make them useful in radar applications and more recently for discrete multitone communications and OFDM [8]. The original publication [7] defines a complementary series as a pair of equally long sequences composed of two types of elements that have the property that the number of pairs of like elements with any given separation in one series is equal to the number of pairs of unlike elements with the same separation in the other series. Another way to define a pair of complementary codes is to say that the sum of their aperiodic autocorrelation functions is zero for all delays except for a zero delay.

The CCK codes that were selected as the basis for IEEE 802.11b were first published in 1996 [8]. More background information on these codes can be found in Halford et al. [9]. The following equation represents the 8 complex chip values for the CCK code set, with the phase variables being QPSK phases:

$$c = \{e^{j(\varphi_1+\varphi_2+\varphi_3+\varphi_4)}, e^{j(\varphi_1+\varphi_3+\varphi_4)}, e^{j(\varphi_1+\varphi_2+\varphi_4)}, -e^{j(\varphi_1+\varphi_4)}, e^{j(\varphi_1+\varphi_2+\varphi_3)}, e^{j(\varphi_1+\varphi_3)}, -e^{j(\varphi_1+\varphi_2)}, e^{j(\varphi_1)}\} \quad (1)$$

Basically, the three phases φ_2, φ_3 and φ_4, define 64 different codes of 8 chips, where φ_1 gives an extra phase rotation to the entire codeword. Actually, the latter phase is differentially encoded across successive codewords, equivalent to the 1- and 2-Mbps DSSS differential phase encoding. This feature allows the receiver to use differential phase decoding, eliminating a carrier tracking PLL, if desired. Each of the four phases φ_1 to φ_4 represents 2 bits of information, so a total of 8 bits is encoded per 8-chip CCK codeword.

At 5.5 Mbps, the processing is similar. Four information bits are consumed per 8-chip CCK codeword transmission. The codeword rate is still 1.375 MHz, since the chip rate is 11 Mchips/s. Two bits select 1-of-4 CCK subcodes. The other two information bits quadriphase-modulate (rotate) the whole codeword. The 4 CCK subcodes are contained in the larger 64 subcode set of 11 Mbps. At the receiver, the CCK codes can be decoded by using a modified fast Walsh transform as described by Grant and van Nee [10].

5. IEEE 802.11a

IEEE 802.11a provides data rates of 6–54 Mbps in the 5-GHz band using orthogonal frequency-division multiplexing (OFDM). The basic principle of OFDM is to split a high-rate datastream into a number of lower-rate streams that are transmitted simultaneously over a number of subcarriers. Since the symbol duration increases for the lower-rate parallel subcarriers, the relative amount of time dispersion caused by multipath delay spread is decreased. Intersymbol interference is eliminated almost completely by introducing a guard time in every OFDM symbol. In the guard time, the OFDM symbol is cyclically extended to avoid intercarrier interference. Figure 3 shows an example of 4 subcarriers

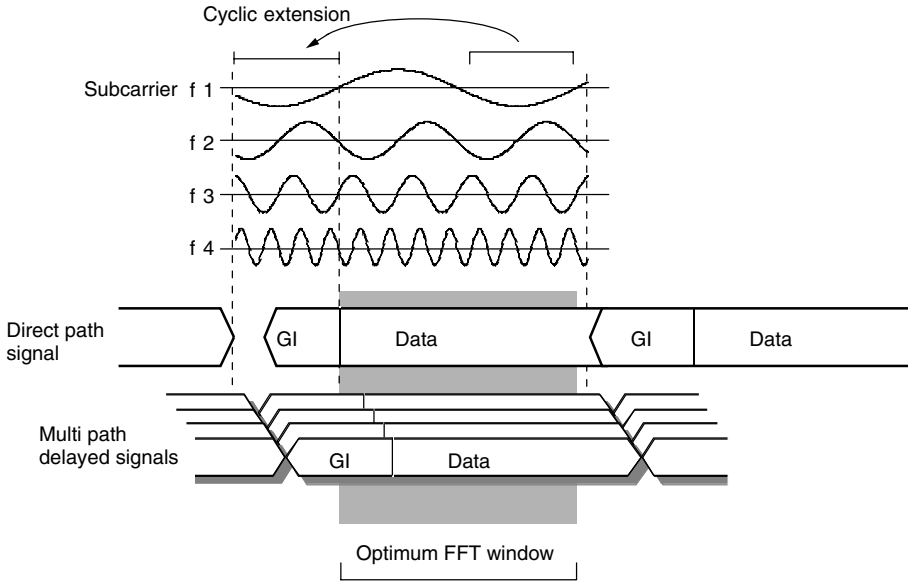


Figure 3. OFDM symbol with cyclic extension.

from one OFDM symbol. In practice, the most efficient way to generate the sum of a large number of subcarriers is by using inverse fast fourier transform (IFFT). At the receiver side, FFT can be used to demodulate all subcarriers. It can be seen in Fig. 3 that all subcarriers differ by an integer number of cycles within the FFT integration time, which ensures the orthogonality between the different subcarriers. This orthogonality is maintained in the presence of multipath delay spread, as illustrated by Fig. 3. Because of multipath, the receiver sees a summation of time-shifted replicas of each OFDM symbol. As long as the delay spread is smaller than the guard time, there is no intersymbol interference nor intercarrier interference within the FFT interval of an OFDM symbol. The only remaining effect of multipath is a random phase and amplitude of each subcarrier, which has to be estimated in order to do coherent detection. In order to deal with weak subcarriers in deep fades, forward error correction across the subcarriers is applied.

5.1. OFDM Parameters

Table 2 lists the main parameters of the IEEE 802.11a OFDM standard. A key parameter that largely determined the choice of the other parameters is the guard interval of 800 ns. This guard interval provides robustness to root-mean-squared delay spreads up to several hundreds of nanoseconds, depending on the coding rate and modulation used. In practice, this means that the modulation is robust enough to be used in any indoor environment, including large factory buildings. It can also be used in outdoor environments, although directional antennas may be needed in this case to reduce the delay spread to an acceptable amount and to increase the range.

In order to limit the relative amount of power and time spent on the guard time to 1 dB, the symbol duration was chosen to be 4 μs. This also determined the subcarrier spacing to be 312.5 kHz, which is the inverse of the symbol duration minus the guard time. By using 48 data subcarriers, uncoded data rates of 12–72 Mbps can be achieved by using variable modulation types from BPSK to 64-QAM. In addition to the 48 data subcarriers, each OFDM symbol contains an additional 4 pilot subcarriers, which can be used to track the residual carrier frequency offset that remains after an initial frequency correction during the training phase of the packet.

Table 2. Main Parameters of the OFDM Standard

Data Rate	6, 9, 12, 18, 24, 36, 48, 54 Mbps
Modulation	BPSK, QPSK, 16-QAM, 64-QAM
Coding rate	$\frac{1}{2}, \frac{2}{3}, \frac{1}{3}$
Number of subcarriers	52
Number of pilots	4
OFDM symbol duration	4 μs
Guard interval	800 ns
Subcarrier spacing	312.5 kHz
−3-dB bandwidth	16.56 MHz
Channel spacing	20 MHz

In order to correct for subcarriers in deep fades, forward error correction across the subcarriers is used with variable coding rates, giving coded data rates of 6–54 Mbps. Convolutional coding is used with the industry standard rate-$\frac{1}{2}$, constraint length 7 code with generator polynomials (133,171). Higher coding rates of $\frac{2}{3}$ and $\frac{3}{4}$ are obtained by puncturing the rate-$\frac{1}{2}$ code.

5.2. Channelization

For the 200-MHz-wide spectrum in the lower and middle UNII bands, 8 OFDM channels are available with a channel spacing of 20 MHz. The outermost channels are spaced 30 MHz from the band edges in order to meet the stringent FCC-restricted band spectral density requirements. The FCC also defined an upper UNII band from 5.725 to 5.825 GHz, which carries another 4 OFDM channels. For this upper band, the guard spacing from the band edges is only 20 MHz, since the out-of-band spectral requirements for the upper band are less severe than

those of the lower and middle UNII bands. In Europe, the same spectrum as the lower and middle UNII band is available, plus an extra band from 5.470 to 5.725 GHz. In Japan, a 100-MHz-wide band from 5.15 to 5.25 is available. This band contains 4 OFDM channels with 20 MHz guard spacings from both band edges.

5.3. OFDM Signal Processing

The general block diagram of the baseband processing of an OFDM transceiver is shown in Fig. 4. In the transmitter path, binary input data are encoded by a standard rate-$\frac{1}{2}$ convolutional encoder. The rate may be increased to $\frac{2}{3}$ or $\frac{3}{4}$ by puncturing the coded output bits. After interleaving, the binary values are converted into QAM values. To facilitate coherent reception, 4 pilot values are added to each 48 data values, so a total of 52 QAM values is reached per OFDM symbol, which are modulated onto 52 subcarriers by applying the inverse fast Fourier transform (IFFT). To make the system robust to multipath propagation, a cyclic prefix is added. Further, windowing is applied to get a narrower output spectrum. After this step, the digital output signals can be converted to analog signals, which are then upconverted to the 5 GHz band, amplified, and transmitted through an antenna.

The OFDM receiver basically performs the reverse operations of the transmitter, together with additional training tasks. First, the receiver has to estimate frequency offset and symbol timing, using special training symbols in the preamble. Then, it can do a FFT for every symbol to recover the 52 QAM values of all subcarriers. The training symbols and pilot subcarriers are used to correct for the channel response as well as remaining phase drift. The QAM values are then demapped into binary values, after which a Viterbi decoder can decode the information bits.

Figure 5 shows the time-frequency structure of an OFDM packet, where all known training values are marked in gray. It illustrates how the packet starts with 10 short training symbols, using only 12 subcarriers, followed by a long training symbol and data symbols, with each data symbol containing 4 known pilot subcarriers that are used for estimating the reference phase. The preamble, which is contained in the first 16 μs of each packet, is essential to perform start-of-packet detection, automatic gain control, symbol timing, frequency estimation, and channel estimation. All of these training tasks have to be performed before the actual data bits can be successfully decoded. More detailed information on OFDM signal processing as well as performance results can be found in Ref. 11.

5.4. Differences Between IEEE, ETSI, and MMAC

The main differences between IEEE 802.11 and HIPERLAN type 2 — which is standardized by ETSI BRAN — are in the medium access control (MAC). IEEE 802.11 uses a distributed MAC based on carrier sense multiple access with collision avoidance (CSMA/CA), while HIPERLAN type 2 uses a centralized and scheduled MAC, based on wireless ATM. MMAC supports both of these MACs. As far as the physical layer is concerned, there are only a few minor differences, summarized as follows:

- HIPERLAN uses extra puncturing to accommodate the tail bits in order to keep an integer number of OFDM symbols in 54-byte packets [12].
- In the case of 16-QAM, HIPERLAN uses rate $\frac{9}{16}$ instead of rate $\frac{1}{2}$ — giving a bit rate of 27 instead of 24 Mbps — in order to get an integer number of OFDM symbols for packets of 54 bytes. The rate $\frac{9}{16}$ is made by puncturing 2 out of every 18 coded bits.
- HIPERLAN uses different training sequences. The long training symbol is the same as for IEEE 802.11, but the preceding sequence of short training symbols is different. A downlink transmission starts with 10 short symbols as in IEEE 802.11, but the first 5 symbols are different in order to enable detection of the start of the downlink frame. Uplink packets may use 5 or 10 identical short symbols, with the last short symbol inverted.

6. IEEE 802.11g

The IEEE 802.11g standard extends the 802.11b standard with higher data rates for the 2.4-GHz band [13]. It

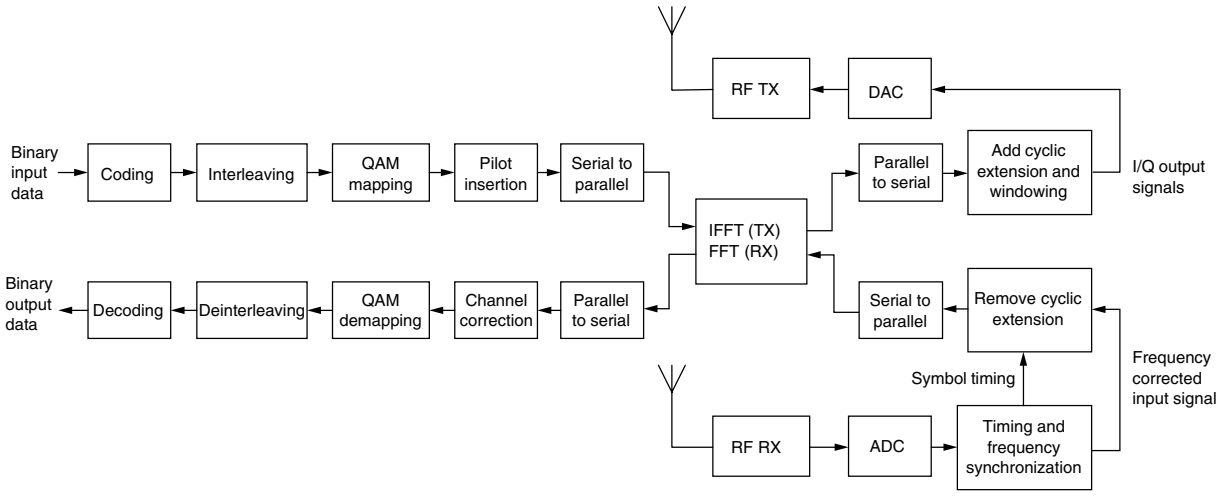


Figure 4. Block diagram of OFDM transceiver.

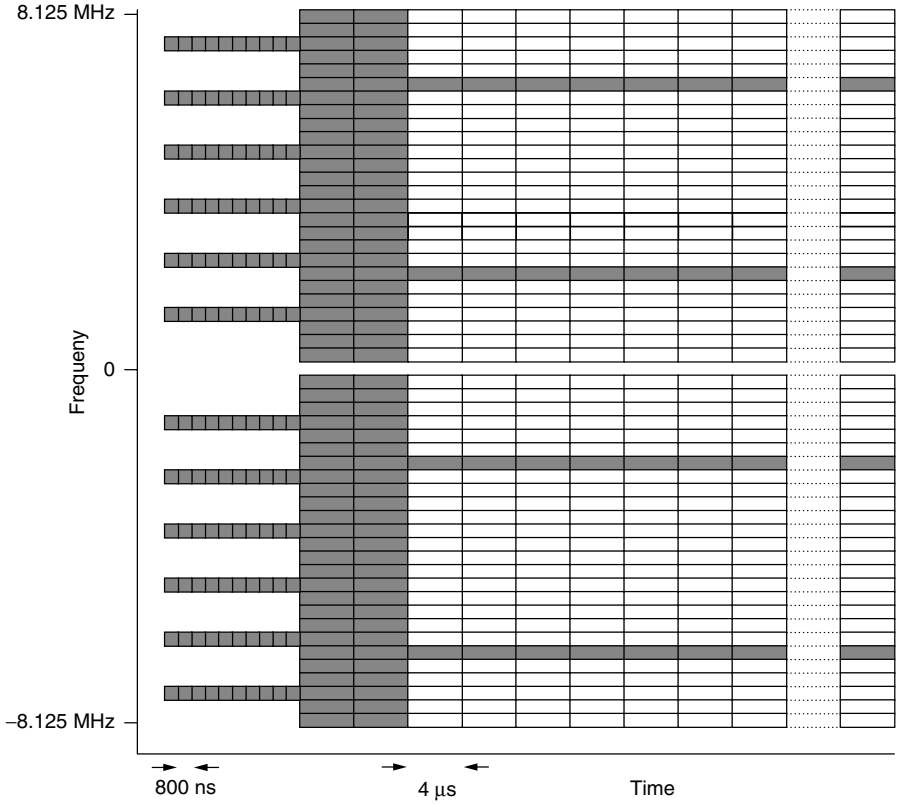


Figure 5. Time–frequency structure of an OFDM packet. Gray-shaded subcarriers contain known training values.

achieves this by simply allowing IEEE 802.11a OFDM transmissions in the 2.4-GHz band, while 802.11a was originally targeted at the 5-GHz band. To remain coexistent with legacy 802.11b devices, the 802.11 RTS-CTS mechanism can be used to claim airtime for high-rate OFDM packets. If no legacy devices are present in a network, it is possible to transmit 802.11a packets without the RTS-CTS mechanism to maximize user throughput. The IEEE 802.11g standard also defines two optional modulation schemes. The first is CCK-OFDM, where each OFDM packet is preceded by an 802.11b header to provide full coexistence with legacy 802.11b devices without the need for RTS-CTS. The second option is *packet binary convolutional coding* (PBCC), which provides a 22-Mbps raw data rate using coded 8-PSK together with a standard 802.11b header.

With the advent of IEEE802.11g, OFDM has become the single solution for high data rates in both the 2.4- and 5-GHz bands, thereby facilitating the production of dual-band devices. While OFDM is ideal for high data rates, it is expected that the old 1- and 2-Mbps 802.11 rates in the 2.4-GHz band will remain important for providing the largest possible coverage range.

BIOGRAPHY

Richard van Nee before his employment as director of WLAN Product Engineering at Woodside Networks, Dr. Van Nee was a key member of the technical staff at Lucent Technologies/Bell Labs in the Netherlands. Dr. Van Nee was among those who proposed the OFDM-based physical layer, which was selected for standardization in IEEE 802.11, MMAC, and ETSI HiperLAN. He was involved in the design of the OFDM modems for the European Magic WAND project. Together with NTT, he made the original OFDM-based proposal that lead to the IEEE 802.11a wireless LAN high-rate extension for the 5 GHz band, with data rates up to 54 Mbps. He also was one of the original proposers of the 11 Mbps IEEE 802.11b extension for the 2.4 GHz band, which is based on Complementary Code Keying as described in one of his papers from 1996. Together with Prof. Ramjee Prasad from Aalborg University, Denmark, he wrote the book *OFDM for Mobile Multimedia Communications*, which is a well-read reference for anyone involved with the new IEEE 802.11a standard. He received his Ph.D. in electrical engineering from Delft University, and his M.Sc. in electrical engineering from Twente University.

BIBLIOGRAPHY

1. IEEE, 802.11, *Wireless LAN Medium Access Control* (*MAC*) *and Physical Layer* (*PHY*) *Specifications*, Nov. 1997.
2. IEEE 802.11, *Supplement to IEEE Standard for Information Technology — Telecommunications and Information Exchange between Systems — LAN/MAN Specific Requirements — Part 11: Wireless MAC and PHY Specifications: Higher Speed Physical Layer in the 2.4 GHz Band*, IEEE Standard 802.11b, Jan. 2000.
3. IEEE 802.11, *Supplement to IEEE Standard for Information Technology — Telecommunications and Information Exchange between Systems — LAN/MAN Specific Requirements — Part 11: Wireless MAC and PHY Specifications: High Speed Physical Layer in the 5 GHz Band*, IEEE Standard 802.11a, Dec. 1999.

4. ETSI, *Radio Equipment and Systems, HIgh PErformance Radio Local Area Network (HIPERLAN) Type 1*, European Telecommunication Standard, ETS 300–652, Oct. 1996.
5. Federal Communications Commission Notice of Proposed Rulemaking and Order, FCC 01–158, ET Docket 99–231, May 11, 2001.
6. M. Webster, C. Andren, J. Boer, and R. van Nee, *Harris/Lucent TGb Compromise CCK (11 Mbps) Proposal*, Document IEEE P802.11-98/246, July 1998.
7. M. J. E. Golay, Complementary series, *IRE Trans. Inform. Theory* 82–87 (April 1961).
8. R. van Nee, OFDM codes for peak-to-average power reduction and error correction, *IEEE Global Telecommunications Conf.* Nov. 18–22, 1996, pp. 740–744.
9. K. Halford et al., *Complementary Code Keying for Rake-Based Indoor Wireless Communication*, IEEE ISCAS '99, Orlando, FL.
10. A. Grant and R. van Nee, Efficient maximum likelihood decoding of Q-ary modulated Reed-Muller codes, *IEEE Commun. Lett.* **2**(5): 134–136 (May 1998).
11. R. van Nee and R. Prasad, *OFDM for Mobile Multimedia Communications*, Artech House, Boston, Jan. 2000.
12. ETSI BRAN, *HIPERLAN Type 2 Functional Specification Part 1—Physical Layer*, DTS/BRAN030003-1, June 1999.
13. S. Halford et al., *Proposed Draft Text: B + A = G High Rate Extension to the 802.11b Standard*, draft IEEE 802.11G Standard, Document IEEE 802.11-01/644r0, November 2001.

WIRELESS LOCAL LOOP STANDARDS AND SYSTEMS

HOMAYOUN HASHEMI
Sharif University of Technology
Teheran, Iran

1. INTRODUCTION

Communication plays a vital role in economic development of nations, and in prosperity and well-being of their citizens. A communication plant consists of three main segments: (1) local access plant, the "last mile" of communication, where a subscriber (home or office) is connected to the telephone company's central office; (2) switching facilities, the mechanism that switches and routes calls to their final destination; and (3) long-distance transmission lines, through which calls are transferred within remote areas of the same nation, or between different countries. Segments 2 and 3 have undergone drastic changes in the >125-year history of telephony. The manual switches of early era were transformed into the more advanced electromechanical switches, which then evolved to today's high-speed electronic (digital) switches. "Long distance" lines of 100 years ago consisted primarily of copper wires installed on tens of thousands of telephone poles between cities. Today, millions of calls are transferred nationally and internationally every day by vast and complicated interconnection of high-capacity microwave lines, fiberoptic networks, and low- and high-orbit satellites. The local access technology, however, remains fundamentally unchanged since the times of Alexander Graham Bell. The dominant local loop up to the mid 1990s was a twisted pair of copper line buried underground to connect the home to the nearest telephone exchange.

Telecommunication systems have gone through three phases of evolution [1]. In the era of interconnection (1876–1950) homes and businesses across cities were wired up to telephone central offices (COs), and COs within a city were connected by wire. Limited long-distance lines were established between cities and nations, again, dominantly by copper wire. In the era of networks (1950–1990) the core network was transformed by expanding, automating, and expediting the switching, call processing, and transport functions. Explosive growth of telephony in this era was fueled by invention and successful deployment of equipment and systems such as digital computers, digital switches, satellites, fiberoptics, Integrated Services Digital Network (ISDN), asynchronous transfer mode (ATM) switches, and the Internet. These inventions revolutionized telecommunications, and drastically changed the way people live and work. Great expansion of long-distance telephony, automatic dialing, and introduction of intelligent networking functions are among the achievements of this era. As an example, only 1% of the "core network" in the United States in 1990 was developed in the first 75 years of "interconnection era." The balance of 99% was developed in the 40 years of "network era" [1].

The invention and rapid deployment of cellular radio systems in the 1980s has paved the way for a new era in telecommunications, the "era of access," which started roughly in 1990. This era is manifested by gradual replacement of the primitive, costly, and difficult to install and maintain wired (copper)-based local loops, to the modern, relatively inexpensive, and easy to install wireless local loops (WLLs). It is anticipated that in the near future the majority of telephone services in the world will be based on wireless access (fixed or mobile). Replacement of wired local loop in a sense removes the bottleneck of communications, paves the way for quick, efficient, and economical introduction of other services such as data and video, in addition to the traditional speech communications, a process that expedites transformation of our society into the Information Age.

Basic principles of WLL, including its technical, economical, and regulatory aspects, are described in the literature [1–6].

2. WLL PRINCIPLES

A wireless access system, which is also known as WLL, fixed cellular radio, fixed wireless access (FWA), radio in the local loop (RLL), is emerging as a modern access method. Basic wireline and wireless access system models are shown in Fig. 1. Inspection of this figure shows that a multiple access radio system replaces the wires in the loop.

WLL systems consist of four basic building blocks: (1) a "radio terminal" or "subscriber station," the visible transceiver device carried by or located near the user; (2) a "radio base station," which provides the "wireless" air interface between the subscriber and the network;

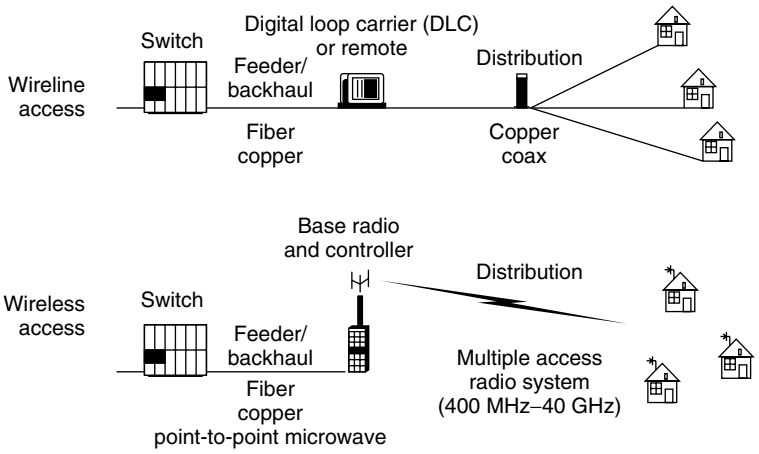


Figure 1. Basic wireline and wireless access system models. (*Source*: Ref. 2.)

(3) a "network control subsystem," which controls the wireless access system; and (4) a "fixed network," the wired infrastructure to which the wireless system provides access.

This scenario can be implemented at each location, independent of other locations, to provide access to limited number of subscribers. The large-scale implementation in an area, however, requires application of spectrum-efficient techniques. Modern WLL is therefore based on principles of cellular radio.

2.1. The Cellular Radio Concept

Cellular radio has been developed for mobile telecommunications, out of a need to increase system capacity and at the same time conserve the scare radio spectrum. The concept is very simple; a large geographic area is partitioned into smaller local areas called "cells." The block of radio spectrum (normally consisting of hundreds of channels) is also partitioned into smaller subblocks or "sets." Channel sets are then assigned to each cell. Two distant cells can use the same channel sets without excessive interference with each other. Repeated reuse of the same channel sets throughout a service area results in drastic increases in system capacity [7]. A mobile subscriber initializes a call while moving in a given cell. When he crosses the cell boundary, calls are "handed off" to the new base station serving the new cell. This process of changing carrier frequency is performed automatically, without any subscriber action.

When telephone traffic in a service area increases, the system capacity can be increased accordingly by "cell splitting," where new fixed antennas or base stations are placed half-way between existing antennas, splitting large calls into smaller cells. This process increases frequency reuse and therefore system capacity. The cell splitting process is gradual and nonuniform to reflect the nonuniformity of telephone traffic and differences in growth rate throughout a service area.

Cellular radio, originally developed for providing voice communication to mobile (vehicular) subscribers, has evolved since the early 1980s into a sophisticated wireless engine capable of providing variety of services to both fixed and mobile users. Basic principles are described by Lee [8] and Rappaport [9].

2.2. Differences Between Fixed and Mobile Access

Application of cellular radio to WLL is similar to mobile radio, with the exception that in WLL both ends of transmission are normally fixed. This brings a number of advantages [5]:

1. The handoff procedure is not implemented, resulting in simplifications of system design and resource management, and in reduction of system controller's processing capability requirements. It should be noted that even if the user moves around moderately, this advantage still applies since the coverage area (or cell) does not change.
2. Both antennas are higher, resulting in a line-of-sight link most of the time. Propagation loss is, therefore, smaller, and coverage area is larger.
3. Fixed subscriber transceiver makes higher transmission powers (as compared to mobile) possible.
4. Directional antennas can be implemented at subscriber site, as well as at the base station. This results in higher gain transmission to the desired location, and limited interference to other sites using the same frequency. Smaller overall interference increases frequency reuse and system capacity.
5. Properties 2–4 (above) result in an increase in coverage area. This reduces number of required cells in an area, a great advantage for low-density sparsely populated rural locations where system capacity is not an issue.
6. Unlike mobile access, the nature of traffic is not dynamic. This makes frequency planning easier and more efficient.
7. Stationarity of the subscriber unit results in the absence of short-term multipath fading channel impairments. This results in better quality of service.

In spite of a number of differences between WLL and mobile wireless access, since both are based on the cellular concept, a number of standards originally developed for mobile subscribers have been successfully used for WLL applications.

2.3. WLL Subscriber Base and Types of Service

WLL can provide service to both developed industrial nations, and developing countries.

In developed countries, where a reasonably extensive wireline-based communication infrastructure is in place and telephone penetration is high, WLL can be used to provide cost-effective and easy-to-implement service to low-density remote areas. It can also provide additional low-cost, quick-to-implement service to high-density urban areas.

In developing countries where telephone penetration in normally very low, building a copper-based infrastructure is very expensive and time-consuming. WLL can provide a suitable answer to the basic needs of developing countries by injecting hundreds of thousands of lines into each metropolitan area in a short span of time. Rapid improvements in communication facilities of these countries improve standards of living and reduce economic gaps with developed nations [5]. Cellular radio layout with large cells can also be implemented in remote areas of developing countries to provide single-line service to each village lacking telecommunication privileges. This is the fastest and cheapest method to connect the rural population to the communication network.

It should be noted that communication requirements of developing and developed countries are different. Developing countries mostly need voice telephone lines. Service requirement emphasizes low cost and high capacity, even if voice quality is to be sacrificed. In developed industrial countries emphasis is on quality of service and capability to provide new services. Although WLL principles are the same for both, system design aspects are different [10,11].

Low-capacity WLL systems, not based on cellular, have been used for many decades to provide telecommunication (voice) services to isolated and remote areas of both developed and developing countries. More recently, however, new communication services such as fax, data, and video have grown tremendously. The explosive growth of the Internet technology and services has resulted in an increase in data transmission requirements of modern offices and homes. Such services, a number of which are broadband in nature, have changed telecommunication needs of the subscribers, and have brought new requirements on telecommunication facilities and infrastructure. WLL is also capable of providing these new services. For broadband applications, however, availability of spectrum is an issue. Design concepts are also different.

2.4. Advantages of WLL

Deployment of WLL is gaining momentum for providing service to densely populated urban areas, as well as sparsely-populated remote and rural areas [2–5]. Replacing the wires in the "last mile" of communication with WLL results in the following major advantages:

1. *Speed of Implementation*. Installation of wired local loops involves burial of wires underground, a time-consuming process, especially for longer paths. In WLL, on the other hand, radio units can be installed quickly at both ends of transmission. The speed of implementation within the radio coverage area is insensitive to distance. WLL can be implemented 5–10 times faster than copper-based systems [2].
2. *Small Initial Investment*. Wired local loops require large lumpy investments, as shown in Fig. 2. A

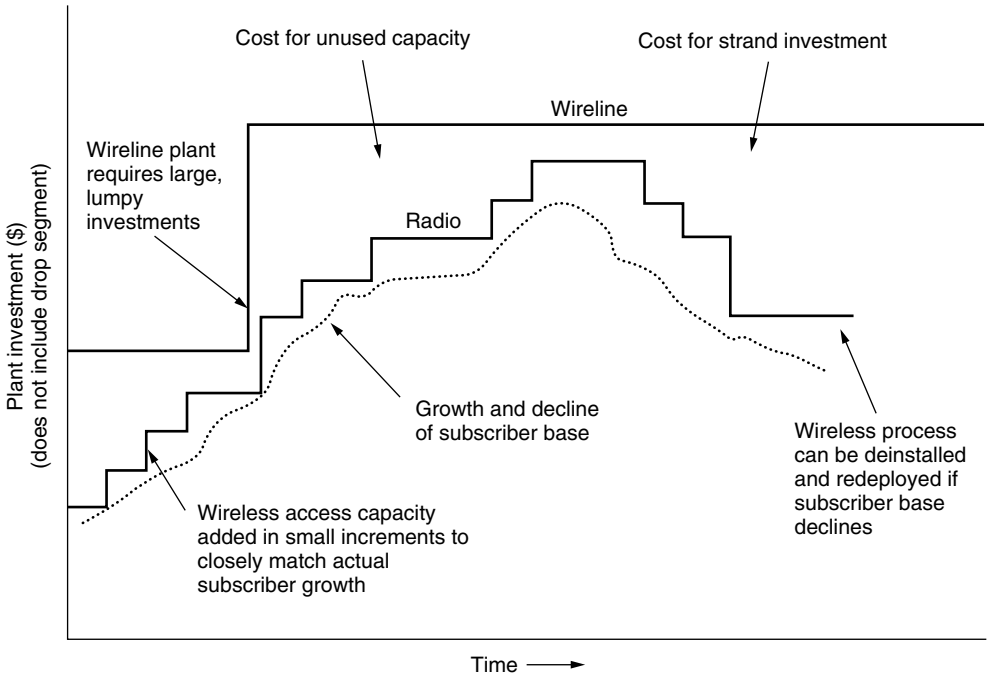


Figure 2. Costs of wireline and wireless access systems versus time. (*Source*: Ref. 1.)

large initial investment is required to lay down main distribution cables, a capacity that remains unused for long periods of time. Subsequent major expansions also involve other lumps of added investments, which bring no immediate return. WLL, on the other hand, requires investments in small increments to match growth in traffic (Fig. 2). Roughly 20% of installation costs are related to the infrastructure; the remaining 80% is spent at the time when subscriber receives service, which is followed by revenues and immediate return on investment. These advantages make competition feasible for small startup companies with limited capital.

3. *Cheap and Easy Maintenance*. Operational costs of WLL are considerably less than those of wireline loops. One study shows a reduction of 25% per subscriber per year [2]. This study shows that over 30% of trouble reports for wired networks are related to distribution cable, drop wires, and in-home wiring, which all can be eliminated in WLL-based systems. Theft and vandalism of wired loops are other types of loss for telephone companies, which can be avoided by WLL.
4. *Fast and Easy Substitution of Faulty Equipment*. Telecommunication facilities in WLL-based systems are located either at the central office (CO) or in customer premises. In wired loops underground copper wires substitute a major portion of the hardware. Repairs and substitutions in WLL are therefore much faster.
5. *Possibility of Deinstallation and Redeployment*. If subscriber demand in an environment declines, wired loops are simply abandoned, with great capital losses (Fig. 2). In WLL, on the other hand, equipment consisting of radio units at both ends of transmission can be removed and redeployed in other places.
6. *Insensitivity to Subscriber's Exact Location*. Implementation of copper loops requires knowledge of exact subscriber location, in contrast with WLL, in which the subscriber only needs to be within the radio coverage area.
7. *Mobility of Subscriber*. The emphasis of WLL is on providing service to fixed terminals. The radio loop, however, results in the added advantage of subscriber mobility.
8. *Variety of Services*. The twisted pair of copper is a primitive, very-low-capacity transmission medium that has been used traditionally for single-channel voice communication. Effective line capacity of wired loops has been increased by introducing sophisticated (and relatively expensive) digital subscriber loop (DSL) modems at both ends of the loop. Still, the wired loop provides a bottleneck for delivery of broadband video and high bit rate data to subscribers. In radio loops, however, such broadband services can be offered by deploying line-of-sight (LoS) transmission at higher frequencies. Availability of spectrum is, of course, an issue in WLL.

3. WLL TECHNOLOGY

Since large-scale deployment of WLL is based on cellular radio, most techniques and technologies developed for mobile radio applications can be successfully applied to WLL.

3.1. Access Methods

Three major channel access methods for cellular wireless communications are frequency-division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access (CDMA) [8,9].

Under an FDMA scheme the allocated band is divided into a number of distinct channels, each one to be used as a single-voice channel. Signals are therefore separate in frequency but mixed in time. In FDMA, when all channels in a cell are occupied, a new call is blocked. FDMA with frequency modulated voice was applied successfully to large-scale mobile telephony in the 1980s. Typically 500–1000 narrowband duplex channels were allocated to a system.

In TDMA, which should better be labeled FDMA/TDMA, the allocated band is first divided into a number of distinct physical channels. Unlike FDMA, however, each physical channel is now used for time multiplexing of several messages. Each user is assigned one of a number of nonoverlapping time slots during which he/she can send or receive digitized messages. Signals are therefore separate in time but mixed in frequency. Transmitter and receiver should have precise time synchronization to avoid interchannel crosstalk. A number of mobile cellular standards operating on the FDMA/TDMA principle were developed in the 1980s and implemented in the 1990s.

CDMA is a spread-spectrum technique, and therefore benefits from associated antinoise, antiinterference capabilities. In CDMA signals are mixed both in time and frequency. Each user in a cell is assigned a distinct code that has a large bandwidth. Codes have good orthogonal properties. At the receiver a correlator is used to correlate the received signal with a replica of the transmitted code (for that particular user). The original message is therefore reconstructed. One great advantage of CDMA is its high capacity. In cellular CDMA, which is also FDMA/CDMA in nature, every physical channel is assigned to each cell. This process eliminates the need for frequency planning, in addition to increasing capacity. CDMA enjoys a "soft capacity limit," in which an additional user can always be accommodated at the expense of small added interference to all other users in the same cell. Successful implementation of CDMA requires accurate synchronization and appropriate power control capabilities. Performance and capacity of CDMA for mobile radio applications were subjects of intense debate in the 1990s. Only one major mobile radio standard based on CDMA emerged in the 1990s. More recently developed wireless communication standards are, however, based mostly on CDMA technology.

TDMA and CDMA can each be either narrowband or wideband. This refers to the width of each physical channel. Wideband schemes have higher capacities and are capable of accommodating new high-bit-rate

nonvoice services. Implementation, however, is more difficult. Newly emerging wireless standards are wideband-oriented.

WLL systems can operate in both frequency-division duplex (FDD) and time-division duplex (TDD) modes. In FDD a pair of duplex channels, widely separated in frequency, are assigned to a user for two-way transmission. In TDD the same physical channel is used for both downlink (base-to-subscriber) and uplink (subscriber-to-base) transmissions. Each subscriber using the physical channel, however, sends and receives messages at different time slots. With exceptions, most currently working wireless systems operate on FDD mode. The emerging systems use both FDD and TDD.

3.2. Resource Management

Successful deployment of large-scale WLL systems depends on efficient use of scarce radio spectrum, which in turn requires good channel assignment policies. There are two major techniques: fixed channel assignment (FCA), and dynamic channel assignment (DCA). Each technique contains a number of variations, and a combination of the two has also been suggested.

In FCA the available channels are partitioned into blocks or "sets." During initial planning of the system, channel sets are assigned to each cell according to forecast traffic requirements. Although initial planning of FCA is sophisticated and requires knowledge of traffic distributions throughout the service area, its later operation is relatively simple. The great disadvantage of FCA is its lower capacity (compared to that of DCA). This is particularly true where traffic distribution is nonuniform.

In DCA all channels are assigned to every cell. Channel assignment for each call is performed by a central processor on an individual basis, after taking interference limitation requirements of the system into account. The advantage of DCA is minimal initial planning and high capacity, especially where traffic distribution is nonuniform in space, and changing with time. The major disadvantage of DCA is elaborate call supervision, which puts a heavy burden on the central processor for small-cell high-capacity systems. With exceptions, current mobile wireless systems use FCA. The emerging standards, however, take advantage mostly of DCA.

All currently available wireless (fixed or mobile) systems are based on circuit switching, where a dedicated circuit is allocated to a user throughout the connection. The circuit may be an FDMA channel, a time slot of a TDMA channel, or a CDMA orthogonal code. On termination of the call the circuit is marked "idle," and later assigned to a new user. The newly emerging standards operate on both circuit-switching and packet-switching modes. In packet switching there is no permanent connection for a user. A number of calls are made using the same channel. This channel is assigned to a number of users on a temporary basis. Each user sends its information in "packets" at assigned intervals. This scheme increases efficiency, and hence capacity, but requires great call supervision and sophisticated processing. Although packet switching can be used for speech, it is more suitable for data transmission applications, which are bursty in nature.

4. WLL STANDARDS

In principle any wireless personal communication standard can be used for WLL systems. The standards, however, are grouped in two major categories: standards based on existing and emerging digital mobile radio systems, and those based on proprietary radio technologies. The first category covers open standards developed by recognized standardization bodies. They provide network operators with the freedom of supplying different subsystems from different manufacturers. In the second category the entire system is normally developed by a specific manufacturer based on proprietary-developed standards.

4.1. Mobile-Radio-Based Standards

Wireless mobile communication is the fastest-growing sector of the telecommunications industry, providing service to over one billion customers worldwide. The exponential growth in number of subscribers started in early 1980s with commercial introduction of systems based on the cellular radio principles.

Wireless cellular communications has undergone three distinct phases of expansion, known as first, second, and third generations:

First-generation (1G) *systems*, designed in the 1970s, and commercially introduced in early 1980s, are based on single-channel analog FM technology with channel spacings of 25 or 30 kHz. Such systems, which are still operating in parts of the world, are used almost exclusively for duplex voice transmissions.

Second-generation (2G) *systems* were designed primarily in 1980s and commercially deployed in 1990s. They are all based on digital technology, making compact and power-efficient transceivers feasible. The primarily vehicular-mounted units of the first generation were, therefore, transformed into personal portable units of the second generation. Low-bit-rate data communication services are also introduced in 2G.

Third-generation (3G) *systems* also operate on digital technology principles. Higher bandwidths allocated to 3G, combined with more sophisticated signal processing techniques, have greatly improved capacity and capabilities of wireless services. 3G is the gateway to personal multimedia, in which standards are capable of providing speech, data, video, and Internet services to wireless (mobile or fixed) units. International coordination for standardization of 3G has been performed by the ITU (International Telecommunications Union) in the framework of IMT-2000 (International Mobile Telecommunications) plan. Implementation of 3G systems started in 2002.

4.1.1. Major Second-Generation Standards. Major second-generation mobile radio standards that are also candidates for WLL are GSM, IS136, IS95 (high-power systems, originally developed for providing service to high-speed vehicular subscribers moving in outdoor large cell

environments), and DECT, PACS, and PHS (low-power microcellular systems, originally intended to provide coverage to low-speed outdoor pedestrians, and indoor users) [2]. Second-generation standards are reviewed by Black [12].

Detailed evaluations of DECT, PACS, and PHS standards have shown the suitability of all three for WLL applications [6]. Basic parameters of second generation standards are summarized below.

4.1.1.1. GSM (Global System for Mobile Communication). GSM, developed by CEPT (Conference Europenne des Postes et Telecommunications) in the 1980s to serve as a pan-European unified standard, evolved into a "model" digital mobile communications standard with the largest subscriber base in the world. GSM was standardized by ETSI (European Telecommunications Standard Institute). It operated in the FDD mode with 25 MHz bandwidth (935–960 MHz for downlink, and 890–915 MHz for uplink). Each band consists of 125 physical channels, each 200 kHz wide. Every physical channel operates in the TDMA mode with 8 user channels, each 0.577 ms. wide, forming 4.615-ms-long frames. Speech coding is RPE-LTP (regular pulse excited–long-term prediction) with 13 kbps (kilobits per second) for each subscriber. A combination of CRC and rate-$\frac{1}{2}$ convolution channel coding results in a gross bit rate per channel of 22.8 kbps. Modulation is Gaussian minimum shift keying (GMSK). GSM channel assignment is fixed (FCA).

The success of original GSM standard, and insufficiency of the original 900-MHz band resulted in introduction of DCS-1800 (Digital Communication System). This standard occupies a duplex pair of bands, each 75 MHz wide (1710–1785 MHz uplink, 1805–1880 MHz downlink). Each band, therefore, accommodates 375 channels with channel spacing of 200 kHz. All other parameters of DCS-1800 are identical to those of GSM. More details on GSM standard can be found in studies by Black [12] and Mehrotra [13].

4.1.1.2. North American TDMA Digital Cellular (IS136). The TDMA-based IS136 standard was developed by TR45.3, a subcommittee of the EIA/TIA (Electronic Industry Association/Telecommunications Industry Association) in the United States. IS136 is compatible with the analog, first-generation system AMPS (advance mobile phone service). It operates in two frequency bands (869–894 MHz uplink, 824–849 MHz downlink). There are 832 physical channels with a channel spacing of 30 kHz (the same as analog FM channels used in AMPS). Channel assignment is FCA. Each channel accommodates three users in TDMA mode with a channel bit rate of 48.6 kbps. TDMA frame length is 40 ms, speech coding is VSELP (vector sum excited linear predictive), channel coding is rate-$\frac{1}{2}$ convolution, and modulation is $\pi/4$-DQPSK (differential quadrature phase shift keying). The standard provides flexibility to accommodate six users in the same 30-kHz band. Black has reported the details of this standard [12].

4.1.1.3. North American CDMA Digital Cellular CdmaOne (IS95). This is the first CDMA digital cellular standard in the world, developed by Qualcomm Inc. in the United States, and standardized by Subcommittee TR45.5 of the EIA/TIA in 1993. The standard uses the same 800-MHz band as analog AMPS and digital TDMA IS136 standards. Each duplex band of 25 MHz is, however, divided into 20 channels, each 1.25 MHz wide. Each channel provides service to a number of users, which are each assigned a distinct code. The technique is based on direct-sequence spread spectrum in which the 9.6-kbps user data are converted to 1.2288 Mcps (million chips per second), occupying one 1.25-MHz physical channel. Frame length is 20 ms, speech coding is QCELP (Qualcomm code excited linear predictive), channel coding is rate-$\frac{1}{2}$/$\frac{1}{3}$ convolution in the downlink/uplink paths, and modulation is OQPSK (offset QPSK). Channel assignment of IS95 is DCA. More details of the standard are reported by Black [12].

4.1.1.4. DECT (Digital Enhanced Cordless Telecommunications). DECT is a European standard also developed by ETSI. It is designed to operate in the 1880–1900-MHz frequency band, with flexibility to use other close bands. It is based on the TDMA-TDD principle. The number of carriers is 10, and carrier separation is 1726 kHz. The transmission rate is 1152 kbps, and the number of TDMA channels for each carrier is 12. Therefore, the total number of voice channels is 120 (10 carriers $\times$ 12 time slots per carrier). Speech coding is 32 kbps ADPCM (adaptive differential pulse code modulation), and modulation method is GFSK (Gaussian frequency shift keying). Channel assignment is dynamic. Maximum transmission power of the base and portable is 250 mW, where dynamic power control reduces it down to 60 mW. This, however, is peak power used during the transmission of a time slot. Average power is $\leq$10 mW, resulting in long battery usage before recharge. Normal cell radius in DECT is several hundred meters. For each voice connection two time slots are used for two-way transmission. In the other 22 time slots the portable unit scans and evaluates other channels for handover to a better channel when available. More information about DECT can be found in the article by Yu et al. [14].

4.1.1.5. PACS (Personal Access Communication System). PACS has been developed in the United States and was standardized by the JTC (Joint Technical Committee) in 1994. It operates in two wide duplex bands 1850–1910 MHz (uplink) and 1930–1990 MHz (downlink). These bands were allocated by the FCC (Federal Communications Commission) in three paired 5-MHz and three paired 15-MHz bands for licensed wideband PCS applications. Also a 10-MHz band (1920–1930 MHz) has been allocated for unlicensed TDD operation. The air interface of PACS allows FDD operation in the licensed band and TDD operation in the unlicensed band [15].

The PACS standard is based on FDD-TDMA (frequency-division duplex) with 200 channels (carrier separation of 300 kHz). Modulation and speech coding are $\pi/4$-QPSK (quadrature phase shift keying) and 32 kbps ADPCM (adaptive pulse code modulation), respectively. Bit rate per channel is 384 kbps.

Table 1. Basic Parameters of the Second-Generation Wireless Standards

	High-Power Macrocellular			Low-Power Microcellular		
System	GSM	IS136	IS95	DECT	PACS	PHS
Frequency band (MHz)	935–960 890–915	869–894 824–849	869–894 824–849	1880–1900	1850–1910 1930–1990	1895–1918
Standardization body	ETSI	EIA/TIA TR45.3	EIA/TIA TR45.5	ETSI	JTC	ARIB
Duplex method	FDD	FDD	FDD	TDD	FDD	TDD
Access method	FDMA/TDMA	FDMA/TDMA	FDMA/CDMA	FDMA/TDMA	FDMA/TDMA	FDMA/TDMA
Number of carriers	124	832	20	10	200	77
Carrier separation (kHz)	200	30	1250	1728	300	300
Modulation	GMSK	$\pi/4$-DQPSK	QPSK	GFSK	$\pi/4$-QPSK	$\pi/4$-QPSK
Rate per channel (kbps)	270.83	48.6	1228.8	1152	384	384
Frame time (ms)	4.615	40	20	$5+5$	2.5	$2.5+2.5$
Slots per frame	8/16	3/6	1	$12+12$	8	$4+4$
Speech coding type and rate (kbps)	RPE-LTP, 13	VSELP, 7.95	QCELP, 9.6	ADPCM, 32	ADPCM, 32	ADPCM, 32
Channel coding	Rate-$\frac{1}{2}$ convolution	Rate-$\frac{1}{2}$ convolution	$\frac{1}{2}$ forward $\frac{1}{3}$ reverse	CRC	CRC	CRC
Channel assignment	FCA	FCA	DCA	DCA	QSAFA	DCA
Modulation efficiency (bps/Hz)	1.35	1.62	0.98	0.67	1.28	1.28
Handoff strategy	Mobile-assisted	Mobile-assisted	Mobile-assisted	Mobile-controlled	Mobile-controlled	Mobile-assisted

Channel assignment is quasistatic autonomous frequency assignment (QSAFA/DCA) [15]. The standard is designed for low mobility applications. However, operation at high speed (several tens of kilometers per hour) is also possible. Maximum transmission power of the portable unit is 200 mW, and average power is 25 mW. More details about PACS have been reported [14,15].

4.1.1.6. PHS (Personal Handy-Phone System). PHS is the Japanese-developed standard operating in the 1895–1918-MHz band. PHS was envisioned as an efficient low-cost cordless and portable phone system. In late 1993 RCR (Research and development Center for Radio systems), currently known as ARIB (Association of Radio Industries and Businesses), approved the RCR STD-28 standard. The interface for connection to the network was subsequently completed by the TTC (Telecommunication Technology Committee), and trial systems started operation. The first commercial system was implemented in mid-1995. Since then, PHS has experienced explosive growth in Japan, reaching a market size of over 8 million in 1999.

PHS is based on the TDMA-TDD principle with 77 channels (carrier separation of 300 kHz). Bit rate is also 384 kbps, and modulation is $\pi/4$-QPSK. Speech coding is 32 kbps ADPCM and channel assignment is dynamic. Each physical channel can be used as four traffic channels in the TDMA mode. Details of PHS are reported in Ref. 16.

Table 1 summarizes parameters of the six major high-power and low-power second-generation digital cellular standards.

4.1.2. Major Third-Generation Standards. The unparalleled success and exponential growth in first-generation mobile communication systems in the early 1980s necessitated initiation of collective efforts in developing standards that could be used internationally to realize the slogan of PCS (personal communication services), wireless access of "any kind, to any one, and at any where." Such efforts were initiated at the ITU in 1985 in the framework of FPLMTS (future public land mobile telephone systems), which was later renamed IMT-2000. The goals set at IMT-2000 was to provide flexible and spectrum efficient voice and data services to wireless users. The minimum bit rate requirements for outdoor high-speed macrocellular, outdoor pedestrian microcellular, and indoor picocellular environments are 144 kbps, 384 kbps, and 2 Mbps, respectively.

To satisfy the needs of a large international subscriber base requires two-way transmission of low-to-high bit rate data and video in addition to conventional voice telephony. The ITU allocated 230 MHz in the 2 GHz band at WARC'92 (World Administrative Radio Conference). The frequency bands are 1885–2025 MHz and

2110–2200 MHz. To provide global coverage and roaming capabilities, both terrestrial and satellite links were considered.

Many proposals were submitted to the ITU, and a number of standards capable of fulfilling the IMT-2000 vision emerged in the late 1990s. The three major standards, two of them based on wideband CDMA, and one on wideband TDMA, are described in this section. The degree of suitability of each standard for WLL applications is yet to be determined. However, the anticipated large-scale deployment of equipment based on these standards, which results in introduction of economically attractive systems, coupled with the capability to provide new higher bit rate services, make these standards suitable candidates for future WLL systems.

Major third-generation standards have been described [17,18].

4.1.2.1. European-Based WCDMA (Wideband CDMA).

This standard, which is also supported by the Japanese wireless industry, has been developed by ETSI in the framework of the European UMTS (Universal Mobile Telecommunication Systems) project. UMTS is the European version of IMT-2000. WCDMA operates in the paired band of the IMT-2000 spectrum based on FDD. It uses 5-, 10-, 15-, and 20-MHz-wide channels, and operates at chip rates of 1.024, 4.096, 8.192, and 16.384 Mcps. Frame length is 10 ms. Provisions are made for multirate and packet data services. The standard is backward-compatible with GSM. Yang has described the basic principles and applications of CDMA [19]. Detailed descriptions of WCDMA are provided by other authors [20,21].

4.1.2.2. North American WCDMA Standard Cdma2000.

The cdma2000 standard was finalized by the Subcommittee TR45.5 of the TIA Engineering Committee TR45 in the United States in March 1998. It is backward-compatible with the cdmaOne (IS95) standard. Provisions for packet data transmission are provided, channel bandwidths are $N \times 1.25$ MHz ($N = 1, 4, 8, 12, 16$), and chip rates are 1.2288, 3.6864, 7.3728, 11.0593, and 14.7456 Mcps. Frame length is 20 ms.

4.1.2.3. North American UWC-136 (Universal Wireless Communications).

The UWC-136 standard, prepared by Subcommittee TR45.3 of the TIA, is a family of technologies based on TDMA. It consists of (1) the IS136 standard with 30-kHz channels for speech and data below 28.8 kbps; (2) IS136+, which again uses 30-kHz channels for speech and data (bit rates, however, are increased to 64 kbps applying M-ary modulation; (3) IS136 HS (high speed) for outdoor vehicular applications using 200 kHz wide channels (data rates of $\leq$384 kbps are possible; duplex policy is FDD); and (4) IS136 HS indoor, in which the 1.6-MHz-wide channels make data rates up to 2 Mbps feasible. FDD and TDD methods are both used. IS136 HS is also compatible with GSM, using the same frame length of 4.615 ms.

The basic parameters of major 3G standards are summarized in Table 2. Further details can be found in the literature [17,18].

4.2. Proprietary Radio Interface Standards

These standards are developed by private organizations to replace the existing wired loops. Such standards cover a wide range of carrier frequencies, radio interface technology, transmission rates, range, performance, and types of service. Major standards cited by the ITU are Nortel Proximity I-Series, SR Telecom's SR 500, and TRT/Lucent Technologies IRT [2]. Other major systems cited by Webb [3] are Innowave Multigain, Airspan, Lucent AirLoop, Interdigital Broadband CDMA, and Granger CD2000. An overview of these standards and

Table 2. Basic Parameters of the Third-Generation Wireless Standards

			UWC-136		
System	WCDMA	Cdma2000	IS136+	IS136 HS Outdoor/Vehicular	IS136HS Indoor
Backward compatibility	GSM	CdmaOne (IS95)		IS136	
Standardization body	ETSI	EIA/TIA TR45.5		EIA/TIA TR45.3	
Access method	WCDMA	WCDMA		TDMA	
Carrier separation	5, 10, 20 MHz	1.25, 5, 10, 15, 20 MHz	30 kHz	200 kHz	1.6 MHz
Maximum user bit rate per code or channel	480, 960, 1920 Kbps	1.0368 Mbps	64 Kbps	384 Kbps	2 Mbps
Maximum user bit rate (multicode)	2 Mbps	2 Mbps		N/A	
Chip rate (Mcps)	1.024, 4.096, 8.192, 16.384	1.2288, 3.6864, 7.3728, 11.0593 direct spread $n \times 1.2288$ (n = 1,3,6,9,12) for multicarrier		N/A	
Frame length (ms)	10	20	40	4.615	4.615

details of their basic parameters have been presented by the ITU [2] and Webb [3].

5. WLL SPECTRUM AND CAPACITY

Large-scale deployment of WLL depends on availability and efficient use of the radio spectrum. Wireless access systems, fixed and mobile, operate in the wide frequency range of 400 MHz–40 GHz. A number of analog FM systems based on European standards operate at 400 MHz. IS136 mobile cellular uses the 800-MHz band, while GSM operates at 900 MHz. Point-to-multipoint radio in the loop has been designed at 1.4 GHz. The 1.8–1.9-GHz bands are used for GSM, DECT, and PHS standards. A wide band at 2 GHz has been allocated by the ITU for global implementation of IMT-2000 standards. Multipoint distribution systems (MDSs) and point-to-multipoint radio also operate at 2.4, 2.5, 2.6, and 10.5 GHz. Finally, 28- and 40-GHz bands are used for local multipoint communications and distribution systems (LMCS/LMDS) [2]. Most of these frequencies, especially lower bands, have been developed for mobile users; however, they are either being used, or have the potential of being used for WLL-type applications.

Capacity is the most critical issue in wireless personal communications, fixed or mobile. Cellular radio architecture provides high capacities because of its spectrum efficiencies [22]. It has been shown by an example that even the relatively low capacity cellular analog FM standards are capable of providing millions of fixed WLL-type telephone lines in a metropolitan area if a reasonably large frequency band is allocated [5]. Capacity of digital cellular systems is higher than analog because (1) speech coding reduces bandwidth per subscriber, and (2) channel coding results in smaller signal-to-interference requirements, which in turn decreases minimum reuse distance, and further increases capacity.

Performance and capacity of DECT, PACS, and PHS standards for WLL applications have been reported [6]. Detailed qualitative and quantitative evaluations indicate that all three standards provide satisfactory performance for WLL applications. For low-traffic environments, PACS which can employ larger cells performs better than the other two standards. In suburban areas where in addition to coverage capabilities capacity is an issue, DECT has better performance. For high-traffic-density urban areas with great capacity requirements, the three standards all have good performance [6].

TDMA systems have higher capacity than do FDMA systems. In CDMA more interference can be tolerated, which boosts capacity even higher. The CDMA-versus-TDMA capacity of cellular mobile radio systems was a subject of intense debate in the 1990s. Real capacity evaluations are difficult due to a large number of parameters and assumptions. It has been suggested [3] that capacity of CDMA for WLL applications is 1.4–2.5 times that of TDMA. Cell sectorization results in even further CDMA capacity improvements.

Capacities of cellular mobile and fixed WLL have been compared with reference to detailed calculations [10,11]. In one study, the capacities of the IS136 (TDMA), GSM (TDMA), and IS95 (CDMA) standards were evaluated in detail and compared for both fixed and mobile access [10]. CDMA WLL capacity (measured in erlangs per cell per MHz) was found to be about 2.5 times that of IS136, and about 5.4 times that of GSM. Detailed analysis [11] has shown that capacity of WLL is not necessarily higher than that of mobile if FDMA or TDMA is used. It is higher if CDMA is applied. For WLL applications alone, CDMA always provides higher capacity than does TDMA [11].

Many advantages of CDMA have made it the major access method for the emerging third-generation wireless mobile standards. CDMA is also expected to be access of the choice for WLL applications in years to come.

6. ECONOMICS OF WLL

Successful deployment and operation of any communication system depends on its economic viability. In a typical system the operator makes an initial capital investment to build the initial infrastructure for a system that will grow and provide service to a large number of subscribers in the future. Interest should be paid on the capital sum until the network becomes profitable. Revenues increase as customer base increases. There are also ongoing maintenance, upgrade, and marketing costs associated with operation of the system.

A typical communication system consists of a core network of switching and transmission equipment for backhaul connections, and the access part, which connects the subscriber to the network. The network cost is approximately the same for wired and wireless loops. The per subscriber cost of the core network, estimated at \$30–\$60 for a large network of half a million users [3], is insignificant as compared to the cost of access. Detailed cost evaluations are complex for both wired and wireless access (especially wired) because of the large number of interrelated parameters. It is also changing with time as a result of inflation and changes in technology. However, one can compare the two by elaborating on the involved parameters of each, and by looking at their trends.

6.1. Cost of Wired Loops

The copper-based wired loops require lumps of investments to lay down main cables to support a future large subscriber base. The cost of unused capacity and the interest payments on the initial large capital makes competition stiff for small-size new operators. The cost per line of the access varies greatly, depending on the number of customers, length of loop, and type of terrain. As an example, analysis of cable costs for 30 rural area sites in the United States in 1990 showed a per line cost as low as \$834 and as high as \$50,000. The average per line cost for all 30 sites covering 8282 subscribers was \$3102 [1]. Looking at the trends, however, it can be shown that cost per access line increases almost linearly as a function of length of access line, up to a point where it takes a jump as a result of required loading coils. From that point it increases almost linearly, although with a larger slope due to the required shift to larger gauge cable. Increasing the distance results in another jump for insertion of additional loading coils,

and a subsequent exponential increase afterward due to even larger gauge cable, and due to the fact that very long loops normally involve higher costs because of difficult terrain [1].

The per line cost of wired loops in a flat area normally decreases as penetration (subscriber density) increases. Simple calculations by Webb [3] indicate that in an environment where houses are located adjacent to each other, cost per subscriber increases from $450 to $1650 when penetration decreases from 100% to 20%.

6.2. Cost of Wireless Loops

In wireless local loops, investments are made in small increments, matching growth in traffic (Fig. 2). The large idle initial investment can, therefore, be avoided. Subsequent interest payment on investment is also considerably smaller. Installation costs are divided typically 20% on infrastructure (initial investment for base station equipment) and 80% on subscriber (paid when the customer receives service, which is immediately followed by generated revenue). These factors make WLL particularly attractive for low-capital new-entrant operators.

The cost of wireless access also depends on a number of factors, including type of access and subscriber density. Per subscriber costs are lowest for low-power microcellular systems operating in dense outdoor urban areas and in indoor environments. The cost is highest for megacellular satellite systems providing coverage to low-density rural areas. The most important aspect of WLL cost is that it is distance-insensitive. If a subscriber is within the coverage area of a base station, its distance to the base does not affect cost.

The cost of WLL can be divided into "shared" and "dedicated." Shared costs are those allocated to a number of subscribers, such as base station equipment. Dedicated costs are per subscriber costs such as terminal transceiver and antenna. Calhoun [1] subdivided WLL costs into the following five major categories:

1. *Common equipment costs*, such as power supplies, base station antennas, and central processor for system control. This type of cost does not vary with the size of the system.
2. *Traffic-sensitive equipment costs* — costs mainly for RF channel hardware (base station transceiver or radio units), the cost of which depends on the number of subscribers and average traffic statistics.
3. *Per subscriber equipment costs* — costs related to equipment purchased one unit per subscriber. This includes subscriber transceiver, and line card to interface the central office.
4. *Ancillary materials and labor* — includes costs for the pole, antenna, and power supply, all installed in customer premises, and dedicated to a single customer, or shared by a number of customers at the same location.
5. *Overhead charges* — costs charged as a standard percentage of the total value of a project.

This cost structure is depicted in Fig. 3. A simple but effective formula to estimate cost of wireless loops is also provided by Calhoun [1]. The total project cost R is given by $R = AX + BY + C$, where X is the number of subscribers and Y is the number of radio channels. A, B, and C are equipment cost of the per subscriber equipment (item 3 above), equipment cost of per RF-channel equipment (item 2), and cost of the common equipment (item 1), respectively. The relationship between X and Y is not fixed; it depends on the user calling habits (average holding times) and on the grade of service (blocking probability).

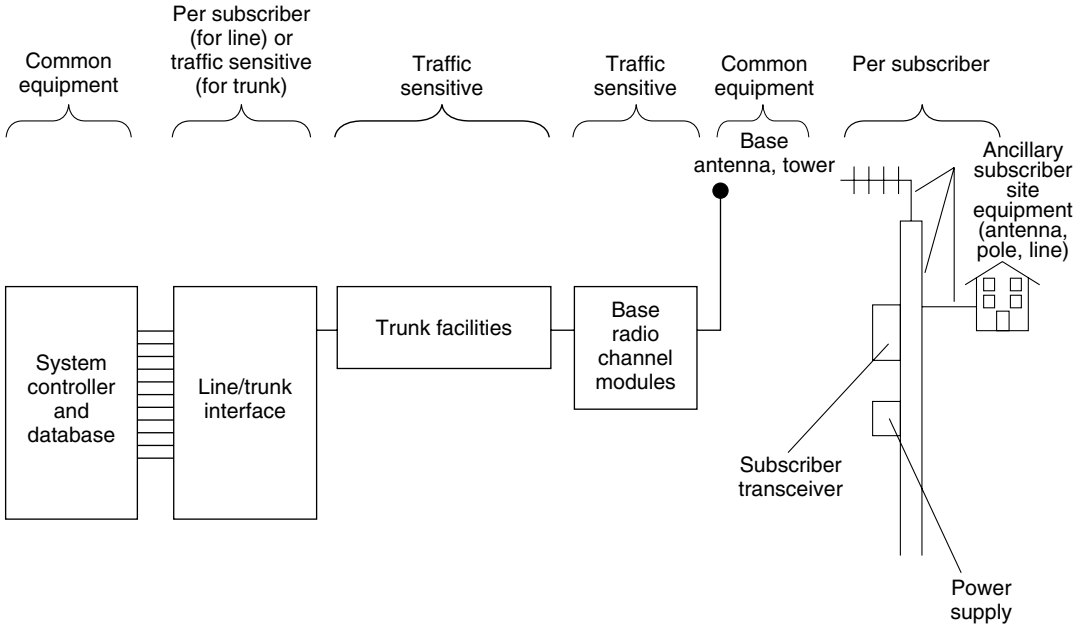


Figure 3. Major components of WLL for cost categorization. (*Source*: Ref. 1.)

Using this approach, one can deduce that total per subscriber cost is high for small X; it decreases rapidly as X increases, and reaches a lower limit (saturation point) where cost of subscriber station plus traffic-sensitive base station cost are dominant cost factors.

6.3. Comparison Between Wired and Wireless Costs

The costs of wired and wireless loops are both decreasing functions of subscriber density (number of subscribers per square kilometer). Comparison of wireless and wireline systems in terms of capital (installation) cost is provided in Fig. 4. Cost per subscriber of WLL in the 1990s was lower than in the 1980s because of mass production of cellular equipment and the "negative inflation" phenomena associated with electronics. Figure 4 shows a breakpoint around 100 subscribers per square kilometer. As time passes, this breakpoint is expected to shift even further to the right (i.e., WLL becomes economically superior for even larger subscriber densities). The reason is that wired loop economy depends on the cost of raw material (copper) and labor, both of which increase with time. WLL economy on the other hand, is governed by the "negative inflation" phenomena associated with electronics.

A case study comparing economy of two access methods is provided by Webb [3]. Three flat areas representing a range of housing densities (high-density, medium-density, and low-density) were considered. It has been shown that in each case when penetration (fraction of houses subscribing to the service) decreases from 25% to 10%, cost per subscriber of cable access increases almost linearly with a modest slope. When penetration decreases below 10%, slope of the cost curve rises sharply. The cost of WLL, however, remains almost the same for penetrations of 5–25%. This comparative study also indicates that per subscriber cost of wired loops is much more sensitive to housing density, as compared to the cost of wireless loops.

Another study shows that a combination of wired and wireless loops provides best economic solutions for a number of cases [1]. In this study relative cost is plotted as a function of percentage of wireless loops, ranging from 0 to 100%. There is an optimum percentage point where the cost is lowest. The optimum point, however, is very sensitive to the individual case, and to the governing parameters and assumptions. It is important to note that capital (installation) cost is only one aspect of economic feasibility. A valid comparison of the two access methods should be based on "annual lifetime cost," which contains capital cost, operating cost, and replacement cost. Such analysis points further at the superiority of WLL [2]. It is estimated that up to 80% of a telephone company's total maintenance cost is allocated to the local loop. WLL represents great savings in maintenance costs since the expensive operations associated with digging the ground and replacing wires are totally eliminated. Operating expenses can be reduced by as much as 25% per subscriber per year in WLL [2]. Considering lifetime cost per subscriber instead of installation costs alone makes WLL superior to wireline-based networks for subscriber densities below approximately 200–400 subscribers per square kilometer. The exact position of the crossover point depends on specific assumptions, distribution of subscribers, and traffic levels.

A major conclusion is that "Wireless access based systems are cost-effective for the provision of telephony services to typical residential subscribers, particularly in rural and suburban areas, or for new entrants in competitive urban markets" [2]. Economic feasibility, combined with other advantages described before, makes WLL access the access of choice in most cases. Large-scale deployment, however, is subject to availability of radio spectrum.

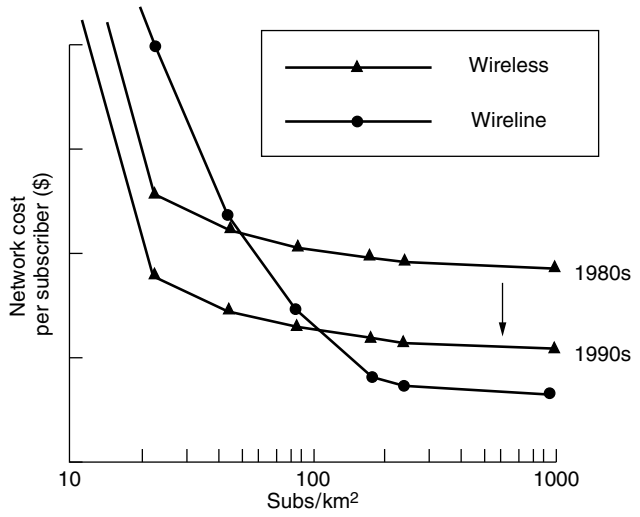


Figure 4. Cost versus subscriber density for wireline and wireless access systems. (*Source*: Ref. 2.)

7. BROADBAND APPLICATIONS OF WLL

For over 100 years telecommunication systems have been dominated by "telephony," that is, two-way transmission of voice. More recently, however, delivery of broadband services such as high-speed Internet and video to the home and office has become increasingly important. Fixed wireline operators are considering digital subscriber techniques such as ADSL (asymmetric digital subscriber line) to increase capacity of copper lines [23]. Very-high-bit-rate services to subscribers can also be delivered by fiberoptic-based techniques such as FTTC (fiber to the curb) and FTTH (fiber to the home). These techniques, however, share many disadvantages with the copper-based local loops.

Subject to allocation of sufficient radio spectrum, WLL is also capable of providing broadband services. It should be noted that "broadband" is not a well-defined term. In voice-oriented mobile telephony, transmission rates above 100 Kbps are considered as "high bit rates." For cable operators, broadband is 8 Mbps or higher.

Technical and economical aspects of broadband WLL, along with a forecast of future are provided by Webb [3]. Third-generation mobile wireless standards provide a peak data rate of 2 Mbps in indoor picocellular environments at the 2-GHz band. This rate provides limited

multimedia capabilities. The emerging new broadband services, however, require higher data rates. As an example, in 1997 the ETSI established the BRAN (Broadband Radio Access Network) project in Europe [24]. The purpose of the project was to utilize broadband LAN (local-area network) technology and broadband fixed radio access to provide mechanisms for delivery of multimedia services to subscribers. In the framework of the BRAN project HiperAccess was suggested for WLL systems. The BRAN project came to the conclusion that a peak data rate of 25 Mbps is sufficient for most broadband user-oriented applications (1.5–6 Mbps for video applications, 2 Mbps for Web browsing, 10 Mbps for corporate access, and up to 25 Mbps for LAN connections). Such high bit rates, however, require transmission at much higher frequencies (say, above 10 GHz), for which technology is evolving. Simple calculations [3] have shown that for economic viability of broadband WLL services, assignment of radio spectrum at least 10 times the bandwidth offered to a user is required. A major conclusion is that "for typical assignments in the frequency bands of 10 GHz and above, bandwidths of 10 MHz per subscriber, using WLL would seem readily achievable" [3]. A number of standards for broadband WLL are being developed, and a number of broadband proprietary WLL products are being introduced to the market [3]. It is safe to expect widespread deployment of broadband WLL-based systems and services in the near future.

8. CONCLUSIONS

In this tutorial presentation the basic principles and applications of fixed wireless access, or WLL, were reviewed. WLL technology, standards, spectrum efficiency, capacity, and economics were described.

A major conclusion is that WLL is an efficient and economically feasible alternative for wired local loops. Major second-generation mobile radio standards, as well as proprietary radio interface standards, have been deployed in WLL applications. Third-generation mobile radio standards and emerging broadband WLL standards, mostly based on wideband CDMA, are potential candidates for providing voice and data services to subscribers in high-density urban areas, as well as sparsely populated rural areas.

BIOGRAPHY

Homayoun Hashemi received the B.S.E.E. degree from the University of Texas at Austin in 1972, and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Sciences, and the M.A. degree in Statistics, all from the University of California at Berkeley, in 1974, 1977, and 1977, respectively. He joined Bell Telephone Laboratories, Holmdel, New Jersey in 1977, where he was involved in system design for high-capacity mobile telephone systems. Since 1979 he has been a faculty member at Sharif University of Technology in Teheran, Iran, where he is currently a full Professor of Electrical Engineering. Dr. Hashemi has done research on different aspects of wireless communications, with emphasis on propagation modeling. His channel simulator package SURP has been used internationally in the design of digital cellular radio communication systems. He spent one year at NovAtel Communications Ltd. in Calgary, Canada in 1990, the summers of 1992 and 1994 at the Electrical Engineering Department of the University of Ottawa, and the summer of 1993 at TRLabs in Calgary. During these visits he defined and supervised three major projects; in each of these projects the largest propagation database of its kind in the world, even to this date, was set up and analyzed. He received the "Best Paper Award" at the IEEE, VTC'99 Conference, and the IEEE Vehicular Technology Society's Neal Shepherd Best Propagation Paper Awards in 2000 and 2001.

BIBLIOGRAPHY

1. G. Calhoun, *Wireless Access and the Local Telephone Network*, Artech House, Norwood, MA, 1992.
2. International Telecommunications Union, *Handbook on Land Mobile (Including Wireless Access),* Vol. I: *Wireless Access Local Loop*, Radio Communication Bureau, 1996.
3. W. Webb, *Introduction to Wireless Local Loop,* 2nd ed., *Broadband and Narrowband Systems*, Artech House, 2000.
4. ITU, Draft Revision of Recommendation ITU-R F.757, *Basic System Requirements and Performance Objectives for Cellular Type Mobile Systems Used as Fixed Systems* (*Fixed Wireless Local Loop Applications of Cellular Type Mobile Technologies*), Document 9B/73, Jan. 1997.
5. H. Hashemi, K. Anvari, and M. Tabiani, Application of cellular radio to telecommunication expansion in developing countries, *Proc. IEEE GLOBECOM'92 Conf.*, Orlando, FL, Dec. 6–9, 1992, pp. 984–988.
6. O. Momtahan and H. Hashemi, A comparative evaluation of DECT, PACS, and PHS standards for wireless local loop applications, *IEEE Commun. Mag.* **39**(5): 156–163 (May 2001).
7. V. H. MacDonald, The cellular concept, *Bell Syst. Tech. J.* **58**(1): 15–41 (Jan. 1979).
8. W. C. Y. Lee, *Mobile Cellular Telecommunications Systems, Analog and Digital*, McGraw-Hill, New York, 1995.
9. T. S. Rappaport, *Wireless Communications, Principle and Practice*, Prentice-Hall, 1996.
10. V. K. Garg and E. L. Sneed, Digital wireless local loop system, *IEEE Commun. Mag.* **34**: 112–115 (Oct. 1996).
11. W. C. Y. Lee, Spectrum and technology of a wireless local loop system, *IEEE Pers. Commun. Mag.* **5**: 49–54 (Feb. 1998).
12. U. Black, *Second Generation Mobile & Wireless Technologies*, Prentice-Hall, 1998.
13. A. Mehrotra, *GSM System Engineering*, Artech House, 1997.
14. Ch. C. Yu et al., Low-tier wireless local loop systems — Part I: Introduction, *IEEE Commun. Mag.* **35**: 84–92 (March 1997).
15. A. R. Noerpel, Y. B. Lin, and H. Sherry, PACS: Personal communications system — a tutorial, *IEEE Pers. Commun. Mag.* **3**: 32–43 (June 1996).
16. S. Sampei, *Application of Digital Wireless Technologies to Global Wireless Communications*, Prentice-Hall, 1997.
17. P. Stavroulakis, *Third Generation Mobile Telecommunication System: UMTS & IMT-2000*, Springer-Verlag, 2000.

18. R. Prasad, *Third Generation Mobile Communication Systems*, Artech House, 2000.

19. S. C. Yang, *CDMA RF System Engineering*, Artech House, 1998.

20. T. Ojanpera and R. Prasad, *Wideband CDMA for Third Generation Mobile Communications*, Artech House, 1998.

21. H. Holma and A. Toskala, eds., *WCDMA for UMTS, Radio Access for Third Generation Mobile Communications*, Wiley, New York, 2000.

22. W. C. Y. Lee, Spectrum efficiency in cellular, *IEEE Trans. Vehic. Technol.* **38**(2): 69–75 (May 1989).

23. P. Kyees et al., ADSL: A new twisted pair access to the information highway, *IEEE Commun. Mag.* **33**: 52–60 (April 1995).

24. J. Haine, HiperAccess: An access system for the information age, *IEE Electron. Commun. Eng. J.* **10**(5): 229–235 (Oct. 1998).

WIRELESS LOCATION

ALI H. SAYED
NABIL R. YOUSEF
Adaptive Systems Laboratory
University of California
Los Angeles, California

1. DEFINITION

Wireless location refers to obtaining the position information of a mobile subscriber in a cellular environment. Such position information is usually given in terms of geographic coordinates of the mobile subscriber with respect to a reference point. Wireless location is also commonly termed *mobile positioning, radiolocation*, and *geolocation*.

2. APPLICATIONS

Wireless location is an important public safety feature of future cellular systems since it can add a number of important services to the capabilities of such systems. Among these services and applications of wireless location are [e.g., 1–11]:

1. *E-911*. A high percentage of emergency 911 (E-911) calls nowadays come from mobile phones [1,2]. However, these wireless E-911 calls do not get the same quality of emergency assistance that fixed-network E-911 calls enjoy. This is due to the unknown location of the wireless E-911 caller. To face this problem, the Federal Communications Commission (FCC) issued an order on July 12, 1996 [1], which required all wireless service providers to report accurate mobile station (MS) location to the E-911 operator at the public safety answering point (PSAP). According to the FCC order, it is mandated that within 5 years from the effective date of the order, October 1, 1996, wireless service providers must convey to the PSAP the location of the MS within 100 m of its actual location for at least 67% of all wireless E-911 calls.[1] It is also expected that the FCC will further tighten the required location accuracy level in the near future [3]. This FCC mandate has motivated research efforts toward developing accurate wireless location algorithms and in fact has led to significant enhancements to the wireless location technology [e.g., 4–11].

2. *Location-Sensitive Billing*. Using accurate location information of wireless users, wireless service providers can offer variable-rate call plans that are based on the caller location. For example, the cell-phone call rate might vary according to whether the call was made at home, in the office, or on the road. This will enable wireless service providers to offer competitive rate packages to those of wire-line phone companies.

3. *Fraud Protection*. Cellular phone fraud has attained a notorious level, which serves to increase the usage and operation costs of cellular networks. This cost increase is directly passed to the consumer in the form of higher service rates. Furthermore, cellular fraud weakens the consumer confidence in wireless services. Wireless location technology can be effective in combating cellular fraud since it can enable pinpointing perpetrators.

4. *Person/Asset Tracking*. Wireless location technology can provide advanced public safety applications including locating and retrieving lost children, Alzheimer patients, or even pets. It could also be used to track valuable assets such as vehicles or laptops that might be lost or stolen. Furthermore, wireless location systems could be used to monitor and record the location of dangerous criminals.

5. *Fleet Management*. Many fleet operators, such as police force, emergency vehicles, and other services including shuttle and taxicab companies, can make use of the wireless location technology to track and operate their vehicles in an efficient way in order to minimize response times.

6. *Intelligent Transportation Systems*. A large number of drivers on road or highways carry cellular phones while driving. The wireless location technology can serve to track these phones, thus transforming them into sources of real-time traffic information that can be used to enhance transportation safety.

7. *Cellular System Design and Management*. Using information gathered from wireless location systems, cellular network planners could improve the cell planning of the wireless network based on call/location statistics. Improved channel allocation could be based on the location of active users [9,10].

8. *Mobile Yellow Pages*. According to the available location information, a mobile user could obtain road information of the nearest resource that the user might need such as a gas station or a hospital. Thus, a cellular phone will act as smart handy mobile yellow pages on demand. Cellular users could obtain real-time traffic information according to their locations.

[1] The original FCC requirement was 125 m and was then tightened to 100 m.

3. WIRELESS LOCATION TECHNOLOGIES

Wireless location technologies fall into two main categories: mobile-based and network-based techniques. In mobile-based location systems, the mobile station determines its own location by measuring signal parameters of an external system, which can be the signals of cellular base stations or satellite signals of the Global Positioning System (GPS). On the other hand, network-based location systems determine the position of the mobile station by measuring its signal parameters when received at the network cellular base stations. Thus, in the later type of wireless location systems, the mobile station plays a minimal or no role in the location process.

3.1. Mobile-Based Wireless Location

3.1.1. GPS Mobile-Based Location Systems. In GPS-based location systems, the MS receives and measures the signal parameters of at least four different satellites of a currently existing network of 24 satellites that circle the globe at an altitude of 20,000 km and which constitute the Global Positioning System. Each GPS satellite transmits a binary code, which greatly resembles a code-division multiple-access (CDMA) code. This code is multiplied by a 50-Hz unknown binary signal to form the transmitted satellite signal. Each GPS satellite periodically transmits its location and the corresponding timestamp, which it obtains from a highly accurate clock that each satellite carries. The satellite signal parameter, which the MS measures for each satellite, is the time the satellite signal takes until it reaches the MS. Cellular handsets usually carry a less accurate clock than the satellite clock. To avoid any errors resulting from this clock inaccuracy, the MS timestamp is often added to the set of unknowns that need to be calculated, thus making the number of unknowns equal to four (three MS position coordinates plus timestamp). This is why four satellite signal parameters have to be measured by the MS. Further information on the GPS systems is available in the literature [12,13].

After measuring the satellite signal parameters, the MS can proceed in one of two manners. The first is to calculate its own position and then broadcast this position to the cellular network. Processing the measured signal parameter to obtain a position estimate is known as *data fusion*. In the other scenario, the MS broadcasts the unprocessed satellite signal parameters to another node (or server) in which the data fusion process is performed to obtain an estimate of the MS position. The later systems are known as *server-aided* GPS systems, while the first are known as "pure" GPS systems [14,15].

A general scheme for server-aided GPS systems is shown in Fig. 1. The server-aided GPS approach is successful in a microcell cellular environment, where the diameter of cellular cells is relatively small (a few hundred meters to a few kilometers). This environment is common in urban areas. On the other hand, in macrocell environments, which are common in suburban or rural areas, base stations, and thus servers, are widely spread out. This increases the average distance between the MS and the server leading to ineffective

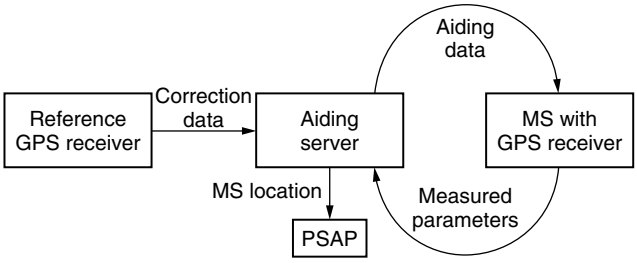


Figure 1. Server-aided GPS location.

correction information. This is why, in many mobile-based GPS location system designs, handsets have to support both server-aided GPS and pure GPS location modes of operation [e.g., 14].

GPS-based mobile location systems have the following advantages. GPS receivers usually have a relatively high degree of accuracy, which can reach less than 10 m with differential GPS server-aided systems [16]. Moreover, the GPS satellite signals are available all over the globe, thus providing global location information. Finally, GPS technology has been studied and enhanced for a relatively long time and for various applications, and is a rather mature technology. Despite these advantages, wireless service providers may be unwilling to embrace GPS fully as the principal location technology due to the following disadvantages of GPS-based location systems:

1. Embedding a GPS receiver in the mobile handset directly leads to increased cost, size, and battery consumption of the mobile handset.
2. The need to replace hundreds of millions of handsets that are already in the market with new GPS-aided handsets. This will directly impact the rates that the wireless carriers offer their users and can cause considerable inconvenience to both users and carriers during the replacement period.
3. The degraded accuracy of GPS measurements in urban environments, when one or more satellites are obscured by buildings, or when the mobile antenna is located inside a vehicle.
4. The need for handsets to support both server-aided and pure GPS modes of operation, which increases the average cost, complexity, and power consumption of the mobile handset. Furthermore, the power consumption of the handset can increase dramatically when used in the pure GPS mode. Moreover, the need to deploy GPS aiding servers in wireless base stations adds up to the total cost of GPS-aided location systems.
5. GPS-based location systems face a political issue raised by the fact that the GPS satellite network is controlled by the U.S. government, which reserves the right to shut GPS signals off to any given region worldwide. This might make some wireless service providers outside the United States unwilling to rely solely on this technology.

3.1.2. Cellular Mobile-Based Location Systems. Cellular mobile-based wireless location technology is similar to

GPS based location technology, in the sense that the MS uses external signals to determine its own location. However, in this type of location systems, the MS relies on wireless signals originating from cellular base stations. These signals could be actual traffic cellular signals or special-purpose probing signals, which are specifically broadcast for location purposes. Although this approach, which is also known as *forward-link wireless location*, avoids the need for GPS technology, it has the same disadvantages that GPS location systems have, which is the need to modify existing handsets, and may even have increased handset power consumption over that of the GPS solution. In addition, this solution leads to lower location accuracy than that of the GPS solution. This makes cellular mobile-based location systems less favorable to use by wireless service providers.

3.2. Network-Based Wireless Location

Network-based location technology depends on using the current cellular network to obtain wireless user location information. In these systems, the base stations (BSs) measure the signals transmitted from the MS and relay them to a central site for processing and calculating the MS location. The central processing site then relays the MS location information to the associated PSAP, as shown in Fig. 2. Such a technique is also known as *reverse-link wireless location*. Reverse-link wireless location has the main advantage of not requiring any modifications or specialized equipment in the MS handset, thus accommodating a large cluster of handsets already in use in existing cellular networks. The main disadvantage of network-based wireless location is its relatively lower accuracy, when compared to GPS-based location methods [3].

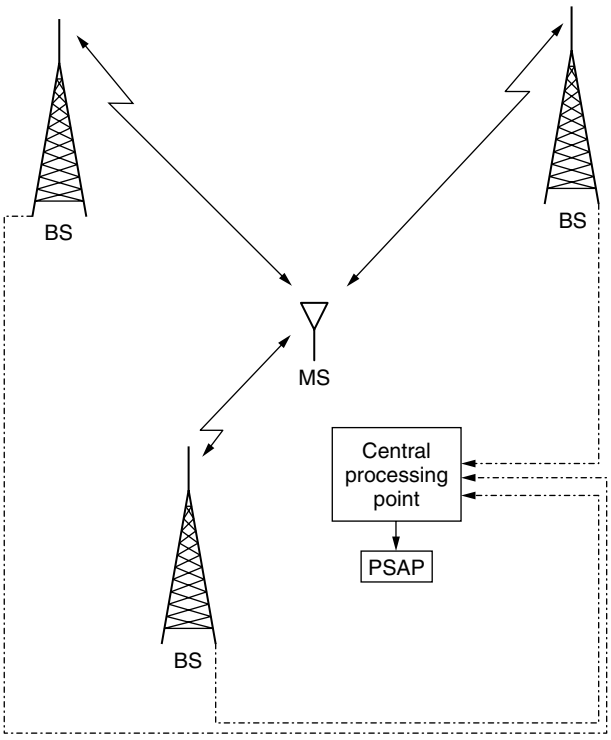


Figure 2. Network-based wireless location.

Network-based wireless location techniques have the significant advantage that the MS is not involved in the location-finding process; thus these systems do not require any modifications to existing handsets. Moreover, they do not require the use of GPS components, thus avoiding any political issue that may arise from their use. However, unlike GPS location systems, many aspects of network-based location are not fully studied yet. This is due to the relatively recent introduction of this technology. In most of the rest of this article, we will focus on network-based wireless location. First, we will review the MS signal parameters that need to be estimated by the cellular base stations and how these signals are combined to obtain a MS location estimate, in data fusion, defined earlier. We will also discuss the sources of error that limit the accuracy of network-based location. Finally, we study different MS signal parameter estimation techniques along with some hardware implementation issues. Here, we may add that although many of the studied aspects apply to both GPS-based location and forward-link location, we will focus on reverse link network-based location. From this point on until the end of the article, we will refer to network-based wireless location simply as *wireless location*.

4. DATA FUSION METHODS

Data fusion for wireless location refers to combining signal parameter estimates obtained from different base stations to obtain an estimate of the MS location. We will study the conventional data fusion methods. The MS location coordinates in a Cartesian coordinate system are denoted by (x_0^o, y_0^o), with the superscript 'o' used to denote quantities that are unknown and which we wish to estimate. These coordinates can be estimated from measured MS signal parameters, when measured at three or more base stations (BSs). The coordinates of the nearest three BSs to the MS, denoted by BS_1, BS_2, and BS_3, are (x_1, y_1), (x_2, y_2), and (x_3, y_3), respectively. Without loss of generality, the origin of the Cartesian coordinate system is set to those of BS_1:

$$(x_1, y_1) = (0, 0)$$

We will denote the time instant at which the MS starts transmission as time instant t_0^o. This MS signal reaches the three BSs involved in the MS location process at instants t_1^o, t_2^o, and t_3^o, respectively. The amplitudes of arrival of the MS signal at the main and adjacent sectors of BS_i are respectively denoted by A_{i1}^o and A_{i2}^o, for $i = 1, 2, 3$.[3] Data fusion methods obtain estimates for the MS coordinates, say, (x_0, y_0), by combining the MS signals through

$$(x_0, y_0) = g(t_0, t_i, A_{i1}, A_{i2}) \tag{1}$$

[3] In cellular systems, a sectored antenna structure is very common. Each BS usually contains three different antennas, with the main lobe of each antenna facing a different direction, and with an angle of 120° between each of the directions. The sector whose antenna faces a specific MS is termed the *main sector* serving this MS. The sector next to the main sector from the MS side is termed the *adjacent sector*.

where $\{t_0, t_i, A_{i1}, A_{i2}\}$ are estimates of $\{t_0^o, t_i^o, A_{i1}^o, A_{i2}^o\}$ and the function g depends on the data fusion method. The resulting location error from the data fusion operation is thus given by

$$e = \sqrt{(x_0 - x_0^o)^2 + (y_0 - y_0^o)^2} \tag{2}$$

One performance index, which is used to compare the accuracy of data fusion methods, is the location mean-square error (MSE), defined by

$$\text{MSE} = Ee^2 = E[(x_0 - x_0^o)^2 + (y_0 - y_0^o)^2] \tag{3}$$

Another performance index for data fusion methods is the value below which the error magnitude, $|e|$, lies for 67% of the time. In other words, it is the value of the error, $e_{67\%}$, at which the error cumulative density function (CDF) is equal to 0.67. The 67% error limit is the performance index that is used by the FCC to set the required location accuracy. Here we may add that for zero-mean Gaussian errors of variance σ_e^2, we have

$$e_{67\%} = \sigma_e = \sqrt{\text{MSE}} \tag{4}$$

Several wireless location data fusion techniques have been introduced since the late 1990s, all of which are based on combining estimates of the time and/or amplitude of arrival of the MS signal when received at various BSs. These methods fall into the following categories:

- Time of arrival (ToA)
- Time difference of arrival (TDoA)
- Angle of arrival (AoA)
- Hybrid techniques

4.1. Time of Arrival (ToA)

The time of arrival (ToA) data fusion method is based on combining estimates of the time of arrival of the MS signal, when arriving at three different BSs. Since the wireless signal travels at the speed of light (C), thus the actual distance between the MS and BS_i, r_i, is given by

$$r_i^o = (t_i^o - t_0^o)C \tag{5}$$

where t_0^o is the actual time instant at which the MS starts transmission and t_i^o is the actual time of arrival of the MS signal at BS_i. Each ToA estimate, t_i, serves to form an estimate of the distance between the MS and the corresponding BS as

$$r_i = (t_i - t_0)C \tag{6}$$

These estimated distances between the MS and each of the three BSs are then used to obtain (x_0, y_0) by solving the following set of equations:

$$r_1^2 = x_0^2 + y_0^2 \tag{7}$$

$$r_2^2 = (x_2 - x_0)^2 + (y_2 - y_0)^2 \tag{8}$$

$$r_3^2 = (x_3 - x_0)^2 + (y_3 - y_0)^2 \tag{9}$$

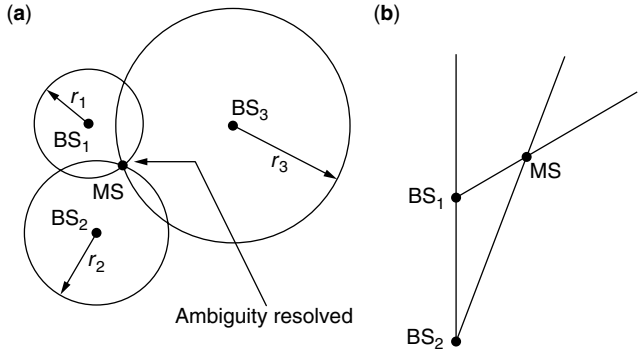


Figure 3. Some wireless location techniques: (**a**) TOA and (**b**) AOA. The MS is positioned at the intersection of the loci.

Without loss of generality, it can be assumed that $r_1 < r_2 < r_3$.

Now, a conventional way of solving this overdetermined nonlinear system of equations is as follows. First, equations (7) and (8) are solved for the two unknowns (x_0, y_0) to yield two solutions. As shown in Fig. 3a, each equation defines a locus on which the MS must lie. Second, the distance between each of the two solutions and the circle, whose equation is given by (9) is calculated. Finally, the solution that results in the shortest distance from the circle (9) is chosen to be an estimate of the MS location coordinates [4].

Although this method will help resolve the ambiguity between the two solutions resulting from solving Eqs. (7) and (8), it does not combine the third measurement r_3 in an optimal way. Furthermore, it is not possible to combine more ToA measurements from BSs more than three.

This can be solved by combining all the available set of measurements using a least-squares approach into a more accurate estimate. This approach can be summarized as follows. Subtracting (7) from (8), we obtain

$$r_2^2 - r_1^2 = x_2^2 - 2x_2x_0 + y_2^2 - 2y_2y_0$$

Similarly, subtracting (7) from (9), we obtain

$$r_3^2 - r_1^3 = x_3^2 - 2x_3x_0 + y_3^2 - 2y_3y_0$$

Rearranging terms, the previous two equations can be written in matrix form as

$$\begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} K_2^2 - r_2^2 + r_1^2 \\ K_3^2 - r_3^2 + r_1^2 \end{bmatrix} \tag{10}$$

where

$$K_i^2 = x_i^2 + y_i^2 \tag{11}$$

Equation (10) can be rewritten as

$$\boldsymbol{Hx} = \boldsymbol{b} \tag{12}$$

where

$$\boldsymbol{H} = \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \boldsymbol{b} = \tfrac{1}{2} \begin{bmatrix} K_2^2 - r_2^2 + r_1^2 \\ K_3^2 - r_3^2 + r_1^2 \end{bmatrix}$$

The solution of (12) is given by

$$\boldsymbol{x} = \boldsymbol{H}^{-1}\boldsymbol{b}$$

If more than three ToA measurements are available, it can be verified that (12) still holds, with

$$\boldsymbol{H} = \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \\ \vdots & \vdots \end{bmatrix}, \boldsymbol{b} = \frac{1}{2}\begin{bmatrix} K_2^2 - r_2^2 + r_1^2 \\ K_3^2 - r_3^2 + r_1^2 \\ K_4^2 - r_4^2 + r_1^2 \\ \vdots \end{bmatrix}$$

In this case, the least-squares solution of (12) is given by

$$\boldsymbol{x} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{b} \tag{13}$$

The ToA method requires accurate synchronization between the BSs and MS clocks. Many of the current wireless system standards only mandate tight timing synchronization among BSs [e.g., 17]. However, the MS clock might have a drift that can reach a few microseconds. This drift directly reflects into an error in the location estimate of the ToA method.

4.2. Time Difference of Arrival (TDoA)

Another widely used technique that avoids the need for MS clock synchronization is based on time difference of arrival (TDoA) of the MS signal at two BSs. Each TDoA measurement forms a hyperbolic locus for the MS. Combining two or more TDoA measurements results in a MS location estimate that avoids MS clock synchronization errors [e.g., 18–21].

We now illustrate how a closed-form location solution can be obtained from TDoA measurements in the case of three BSs involved in the MS location. The TDoA measurement between BS_i and BS_1 is defined by

$$\begin{aligned} r_{i,1} &\triangleq r_i - r_1 \\ &= (t_i - t_0)C - (t_1 - t_0)C = (t_i - t_1)C \end{aligned} \tag{14}$$

Note that TDoA measurements are not affected by errors in the MS clock time (t_0) as it cancels out when subtracting two ToA measurements. Equation (8) can be rewritten, in terms of the TDoA measurement $r_{2,1}$, as

$$(r_{2,1} + r_1)^2 = K_2^2 - 2x_2x_0 - 2y_2y_0 + r_1^2$$

Expanding and rearranging terms, we get

$$-x_2x_0 - y_2y_0 = r_{2,1}r_1 + \frac{1}{2}(r_{2,1}^2 - K_2^2)$$

Similarly, we can write

$$-x_3x_0 - y_3y_0 = r_{3,1}r_1 + \frac{1}{2}(r_{3,1}^2 - K_3^2)$$

Rewriting these equations in matrix form we get

$$\boldsymbol{H}\boldsymbol{x} = \boldsymbol{c}r_1 + \boldsymbol{d} \tag{15}$$

where

$$\boldsymbol{c} = \begin{bmatrix} -r_{2,1} \\ -r_{3,1} \end{bmatrix}, \boldsymbol{d} = \frac{1}{2}\begin{bmatrix} K_2^2 - r_{2,1}^2 \\ K_3^2 - r_{3,1}^2 \end{bmatrix}$$

This equations can be used to solve for $\boldsymbol{x}$, in terms of the unknown r_1, to get

$$\boldsymbol{x} = \boldsymbol{H}^{-1}\boldsymbol{c}r_1 + \boldsymbol{H}^{-1}\boldsymbol{d}$$

Substituting this intermediate result into (7), we obtain a quadratic equation in r_1. Substituting the positive root back into the above equation yields the final solution for $\boldsymbol{x}$.

If more than three BSs are involved in the MS location, Eq. (15) still holds with

$$\boldsymbol{H} = \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \\ \vdots & \vdots \end{bmatrix}, \boldsymbol{c} = \begin{bmatrix} -r_{2,1} \\ -r_{3,1} \\ -r_{4,1} \\ \vdots \end{bmatrix}, \boldsymbol{d} = \frac{1}{2}\begin{bmatrix} K_2^2 - r_{2,1}^2 \\ K_3^2 - r_{3,1}^2 \\ K_4^2 - r_{4,1}^2 \\ \vdots \end{bmatrix}$$

which yields the following least-squares intermediate solution

$$\boldsymbol{x} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T(\boldsymbol{c}r_1 + \boldsymbol{d}) \tag{16}$$

Combining this intermediate result with (7), the final estimate for $\boldsymbol{x}$ is obtained. A more accurate solution can be obtained as in Ref. 19 if the second-order statistics of the TDoA measurement errors are known.

4.3. Angle of Arrival (AoA)

In cellular systems, AoA estimates can be obtained by using antenna arrays. The direction of arrival of the MS signal can be calculated by measuring the phase difference between the antenna array elements or by measuring the power spectral density across the antenna array in what is known as *beamforming* (see, e.g., reference [22] and the works cited therein). Combining the AoA estimates of two BSs, an estimate of the MS position can be obtained (see Fig. 3b). Thus the number of BSs needed for the location process is less than that of ToA and TDoA methods by one. Another advantage of AoA location methods is that they do not need any BS clock synchronization. However, one disadvantage of using antenna-array-based location methods is that antenna array structures do not currently exist in second-generation (2G) cellular systems. Deploying antenna arrays in all existing BSs may lead to high cost burdens on wireless service providers. The use of antenna arrays is planned in some third-generation (3G) cellular systems, such as Universal Mobile Telecommunications System (UMTS) networks [e.g., 23,24], which will use antenna arrays to provide directional transmission in order to improve the network capacity.

AoA estimates can also be obtained using sectored multibeam antennas, which already exist in current cellular systems, using the technique described in reference [25]. In this technique, an estimate of the AoA ($\hat{\theta}$ — see Fig. 4) is obtained based on the difference between the measured signal amplitude of arrival (AmpoA) at the main beam (beam 1) and the corresponding AmpoA

measured at the adjacent beam (beam 2).[4] This difference is denoted by $A_1 - A_2$ in Fig. 5, where A_1 and A_2 are the measured amplitude levels in decibels. The measured AmpoA at the third beam may be used to resolve any ambiguity that might result from antenna sidelobes. One main challenge facing this technique is the relatively low signal-to-noise ratio (SNR) of the received MS signal at the adjacent beam, especially in cases where the AoA is close to a null in the adjacent beam field pattern (e.g., θ close to 0 degrees in Figs. 4 and 5). This significantly limits the AmpoA estimation accuracy at the adjacent beam.

4.4. Hybrid Techniques

In ToA, TDoA, and AoA methods, two or more BSs are involved in the MS location process. In situations where the MS is much closer to one BS (serving site) than the other BSs, the accuracy of these methods is significantly degraded because of the relatively low SNR of the received MS signal at one or more BSs. Such accuracy is further reduced due to the use of power control, which requires the MS to reduce its transmitted power when it approaches a BS, causing what is known as the *hearability* problem [26]. Such problems will be discussed in the next section. In these cases, an alternate location procedure is to obtain an angle of arrival estimate (AoA) from the serving site and combine it with a ToA estimate of the serving site [27]. Combining ToA and AoA estimates from one BS leads to one well-defined MS position estimate, which corresponds to the intersection of a circle and a straight line that starts at the center of the circle. The precision of this hybrid technique is limited by the accuracy of the ToA measurement, which is dictated by the accuracy of the MS clock. Many other hybrid location data fusion techniques can be used, such as combining TDoA and AoA measurements [28].

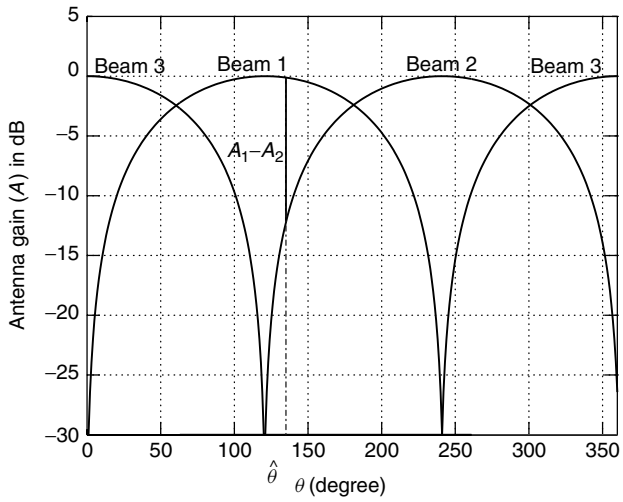


Figure 4. Sectored-antenna field pattern.

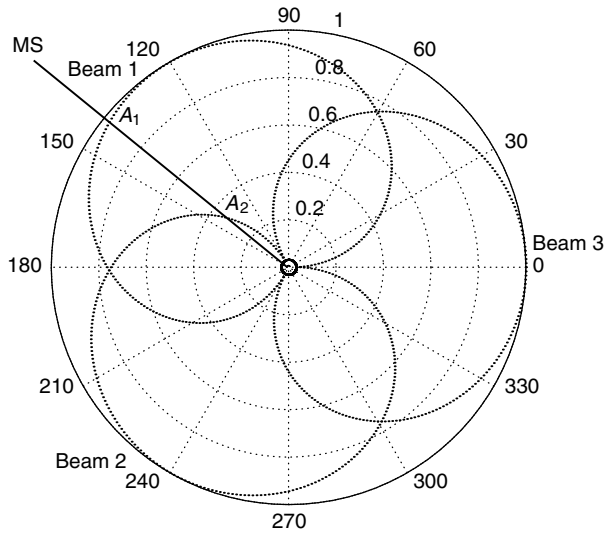


Figure 5. Measured AmpOA level patterns (in dB) for a three-beam antenna versus the AOA (θ).

[4] Here, *main beam* denotes the beam with the highest received signal level and *adjacent beam* refers to the beam that receives the second highest signal level.

5. SIGNAL PARAMETER ESTIMATION

From the previous discussion, we can see that the wireless location methods depend on combining estimates of the ToA and/or AoA of the received signal at/from different BSs. Although estimating the time and amplitude of arrival of wireless signals has been studied in many works since 1990 as it is needed in many cellular systems for online signal decoding purposes [29], parameter estimation for wireless location is actually a different estimation problem in many respects. This makes the success of using conventional estimation algorithms very limited in wireless location problems. In this section, we will illustrate the differences between signal parameter estimation for conventional signal decoding and wireless location. We will then discuss some particular system issues that makes signal estimation for wireless location different from one cellular system to the other (e.g., GSM, 2G and 3G CDMA systems).

Signal parameter estimation for wireless location purposes is different than that for online signal decoding in the following aspects:

1. *Lower SNRs*. Cellular systems usually suffer from high multiple-access interference levels that degrade the SNR of the received signal, thus degrading the signal parameter estimation accuracy in general. Moreover, for network-based wireless locations, the ability to detect the MS signal at multiple base stations is limited by the use of power control algorithms, which require the MS to decrease its transmitted power when it approaches the serving BS. This significantly decreases the received MS signal power level, when received at other BSs involved in the location process. This scenario is shown in Fig. 6, where the received SNRs at BS_1 and BS_2 are significantly reduced as the target MS approaches BS_3. In a typical CDMA IS95 cellular environment, the received SNR of the serving BS is in the order of −15 dB. Conventional signal estimation algorithms are usually designed to work at this SNR level. However, the received SNR at BSs other than

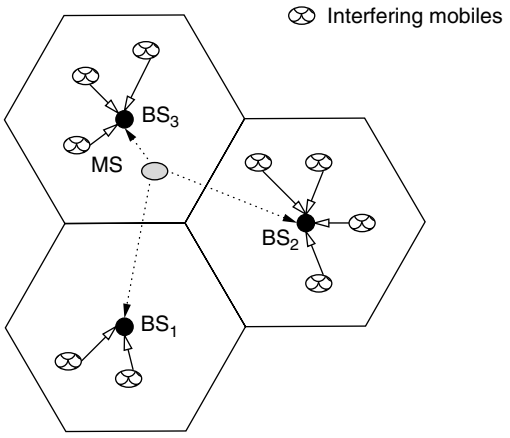


Figure 6. Multiple-access interference among adjacent cells in cellular systems. The letters BS indicate the base stations and the letters MS denote the target mobile station.

the main serving BS can be as low as −40 dB, which poses a challenge for wireless location in such environments.

2. *Almost Perfect Knowledge of Transmitted Signals*. In conventional signal parameter estimation for online signal decoding, the transmitted MS bits are unknown. This forces signal estimation algorithms to perform a squaring operation to remove any bit ambiguity. The squaring operation limits the period over which coherent signal integration (averaging) is possible to the bit period. Further signal integration is only possible in a noncoherent manner, that is averaging after squaring. In wireless location applications, signal estimation algorithms can have almost perfect knowledge of the MS signal in many cases. For example, at the serving site, the MS signal is decoded with reasonably high accuracy (within a 1% frame error rate). The decoded bits become ready for use after a delay that is equal to the decoded frame period used in the cellular system (20 ms for IS95 systems). Because of the nature of wireless location applications, such a delay is not critical. Thus, the received MS signal can be *buffered* or delayed until the decoded bits become available through the conventional decoding process. Moreover, in many cellular systems, a cyclic redundancy check (CRC) feature is used. This enables the decoder to point out the erroneous frames after the decoding process. These erroneous frames can be ignored in the signal estimation process. The decoded bit information, obtained from the main sector of the serving site, can also be used by other adjacent sectors of the same site. Furthermore, this bit information can be transmitted through the network infrastructure to other BSs involved in locating the MS. This is known as *tape recording* of the MS signal. Another technique that avoids the tape recording process is known as the *powerup function* (PuF), which requires the MS in emergency situations to override the power control commands and raise its transmitted power level above the conventional level. Moreover, the MS transmits *known* probing bit sequences instead of its regular unknown bit sequence for a part or all of the transmission period. Although this solution overcomes many of the difficulties encountered at far BSs, it requires modifying the existing handsets or at least the used power control algorithms. Furthermore, it can cause a decrease in the overall network capacity [26].

3. *Channel Fading*. Channel fading is considered constant during the relatively short estimation period of conventional signal parameter algorithms for online signal decoding, and is thus ignored in the design of such algorithms. This assumption cannot be made for wireless location applications where the estimation period could be *considerably longer* (might reach a few seconds). Furthermore, coherent integration periods are no longer limited by the bit duration, much longer coherent averaging periods could be achieved in wireless location applications [30,31]. In this case, the coherent integration period is limited by the received signal phase rotation. Thus, unlike the case of online channel estimators, channel fading plays an important role in any successful design of signal parameter estimators for wireless location. In many cases, the system parameters have to be adapted to the available knowledge of the channel fading characteristics.

4. *Need to Resolve Overlapping Multipath*. Multipath propagation is often encountered in wireless channels (see, e.g., the paper [32] and the references cited therein). In wireless location systems, the accurate estimation of the time and amplitude of arrival of the *first arriving ray* of the multipath channel is vital. In general, the first arriving (prompt) ray is assumed to correspond to the most direct path between the MS and the BS. However, in many wireless propagation scenarios, the prompt ray is succeeded by a multipath component that arrives at the receiver within a short delay from the prompt ray. If this delay is smaller than the duration of the pulse-shape used in the wireless system, these two rays overlap causing significant errors in the prompt ray time and amplitude of arrival estimation. Resolving these overlapping multipath components becomes rather difficult in low SNR and rapid channel fading situations. On the other hand, resolving these overlapping components is not vital for signal decoding applications as it does not significantly affect the performance of the signal decoding operation, for which *coarse* estimates for the channel time delays and amplitudes are sufficient.

Figure 7 shows an example for the combined impulse response of a two-ray channel and a conventional pulseshape for a conventional CDMA IS95 system in two cases (a,b). In case (a), the delay between the two channel rays is equal to twice the chip duration ($2T_c$). It is clear that the peaks of both rays are resolvable, by a simple peak picking procedure, thus allowing for relatively accurate estimation of the prompt ray time and the amplitude of arrival. However, in case (b), both multipath components overlap and are *nonresolvable* via peak picking. This can lead to significant errors in the prompt ray time and amplitude of arrival estimation. These errors cannot be tolerated for wireless location applications, especially in the case of a relatively wide pulseshaping waveform.

5.1. Parameter Estimation Schemes

We now elaborate on some schemes that are used to estimate the wireless signal time and amplitude of arrival.

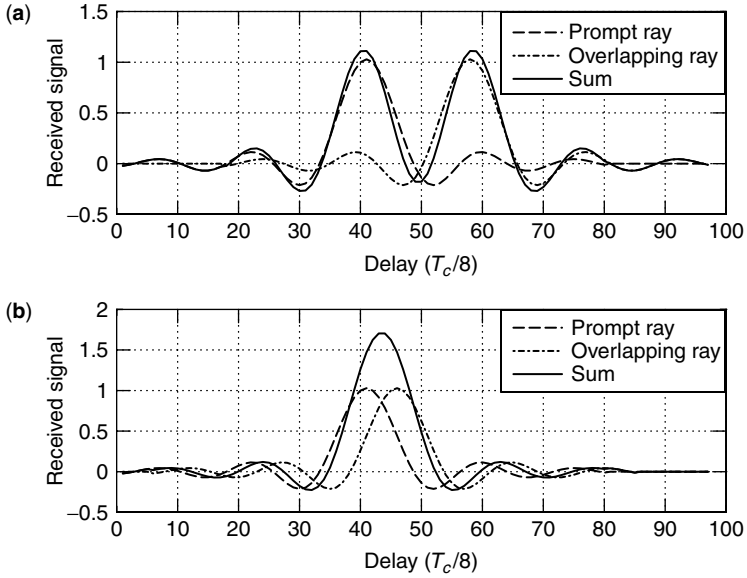


Figure 7. Overlapping rays: (**a**) delay $= 2T_c$; (**b**) delay $= T_c/2$.

The aim of such schemes is to estimate an unknown constant discrete-time delay, τ^o, of a known real-valued sequence $\{s(n)\}$. The signal is transmitted over a single-path time-varying channel, and the designer has access to a measured sequence $\{r(n)\}_{n=1}^{K}$ that relates to $\{s(n)\}$ via

$$r(n) = Ax^o(n)s(n - \tau^o) + v(n) \tag{17}$$

where $v(n)$ is additive white Gaussian noise, and $\{x^o(n)\}$ accounts for the time-varying nature of the *fading* channel gain over which the sequence $\{s(n)\}$ is transmitted, while A is a constant *unknown* received signal amplitude that accounts for both the gain of the *static* channel if fading were not present and the antenna beam gain. Multipath issues are considered later in this section.

A conventional estimation scheme for τ^o for online bit decoding purposes is shown in Fig. 8. In this scheme, the received sequence, $r(n)$, is correlated with replicas of $\{s(n - \tau_i\}$ over a grid of τ values, say, $\{\tau_1, \tau_2, \ldots, \tau_F\}$. The coherent averaging period, N, is set to the bit interval. The outputs of the correlation process are squared to remove any bit ambiguity and then noncoherently averaged over the rest of the available estimation period.

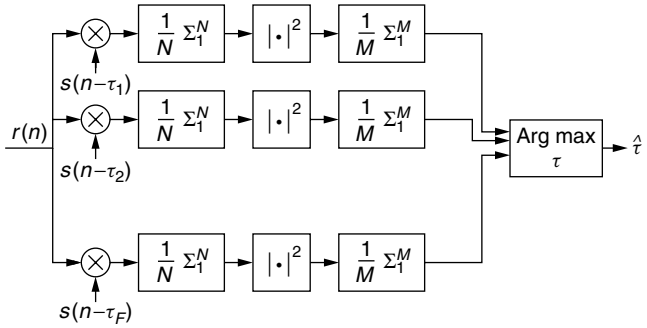


Figure 8. Conventional time-delay estimation for single-path channels.

Figure 9 shows a block diagram of a wireless location ToA/AoA estimation scheme [30]. In this scheme, the received sequence $\{r(n)\}$ is also multiplied by a replica of the transmitted sequence $\{s(n - \tau)\}$ for different values of τ. The resulting sequence is then averaged coherently over an interval of N samples, and further averaged noncoherently for M samples to build a power delay profile, $J(\tau)$. The averaging intervals N and M are positive integers that satisfy $K = NM$, and the value of N is picked adaptively in an optimal manner by using an estimate of the maximum Doppler frequency of the fading channel ($\hat{f}_D$), which can be estimated using some suggested techniques [e.g., 33].

The searcher picks the maximum of $J(\tau)$, which is given by

$$J(\tau) = \frac{1}{M}\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{n=(m-1)N+1}^{mN} r(n)s(n-\tau)\right|^2 \tag{18}$$

and assigns its index to the ToA estimate, according to

$$\widehat{\tau^o} = \arg\max_{\tau} J(\tau) \tag{19}$$

The optimal value of the coherent averaging period (N_{opt}) is obtained by maximizing the SNR gain at the output of the estimation scheme with respect to N which leads to [30]

$$\sum_{i=1}^{N_{\text{opt}}-1} iR_x(i) = 0 \tag{20}$$

where $R_x(i)$ is the autocorrelation function of the sequence $\{x(n)\}$. For a Rayleigh fading channel, $R_x(i)$ is given by

$$R_x(|i|) = J_0(2\pi f_D T_s i)$$

where $J_0(\cdot)$ is the first-order Bessel function, T_s is the sampling period of the received sequence $\{r(n)\}$, and f_D is

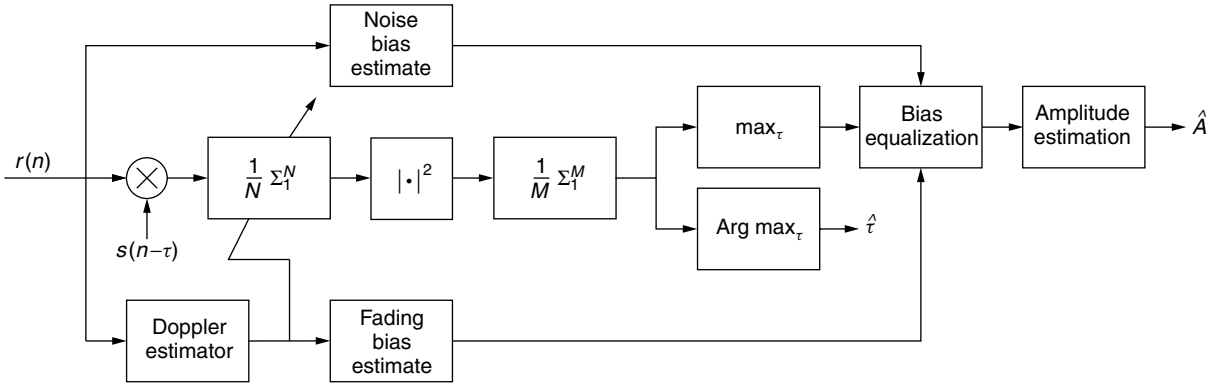


Figure 9. A time-delay estimation scheme for single-path fading channels.

the maximum Doppler frequency of the Rayleigh fading channel. Equation (20) shows that the coherent averaging interval N *should be adapted* according to the channel autocorrelation function.

It has been shown [30,31] that when coherent/noncoherent averaging estimation schemes are used for wireless location applications, where an extended coherent averaging interval is used, two biases arise at the output of the estimation scheme. Both biases affect the accuracy of the amplitude estimate significantly. The first bias is an additive noise bias that increases with the noise variance and is given by

$$B_n = \frac{\sigma_v^2}{N} \tag{21}$$

The second bias is a multiplicative fading bias that depends on the autocorrelation function and is given by

$$B_f = \frac{R_x(0)}{N} + \sum_{i=1}^{N-1} \frac{2(N-i)R_x(i)}{N^2} \tag{22}$$

It is clear that B_f is less than or equal to unity (it is unity for static channels, which explains why previous conventional designs ignored this bias as fading was not considered in these designs [29]; the value of B_f is also unity for $N = 1$).

To correct for these biases, the searcher equalizes the peak value of $J(\tau)$ by subtracting two fading and noise biases, which are estimated by means of the upper and lower branches of the scheme of Fig. 9. The output of this correction procedure is taken as an estimate for the amplitude of arrival, which is given by

$$A = \sqrt{C_f[J(\tau^o) - B_n]}$$

The value of C_f (the fading correction factor) is $C_f = 1/B_f$:

$$C_f = \left[\frac{R_x(0)}{N} + \sum_{i=1}^{N-1} \frac{2(N-i)R_x(i)}{N^2}\right]^{-1} \tag{23}$$

For a Rayleigh fading channel, this correction factor increases with the maximum Doppler frequency of the fading channel. When f_D is estimated, we actually end up with an estimate for C_f. For the case of CDMA systems, the quantity B_n can be estimated as follows. Note first that the noise variance σ_v^2 can be estimated *directly* from the received sequence $\{r(n)\}$ since, for CDMA signals, the SNR is typically very low. In other words, we can get an estimate for σ_v^2 as follows:

$$\widehat{\sigma_v^2} = \frac{1}{K}\sum_{i=1}^{K} |r(i)|^2$$

Then, an estimate for B_n is given, from (21), by

$$\widehat{B}_n = \frac{\widehat{\sigma_v^2}}{N} = \frac{1}{NK}\sum_{i=1}^{K} |r(i)|^2 \tag{24}$$

With $\{\widehat{B}_n, C_f\}$ so computed, we obtain an estimate for A via the expression

$$\widehat{A} = \sqrt{C_f[J(\widehat{\tau^o}) - \widehat{B}_n]} \tag{25}$$

More details on this scheme and simulation results can be found in the literature [30,31,34].

5.2. Overlapping Multipath Resolving

As mentioned before, wireless propagation usually suffers from severe multipath conditions. In situations where the prompt ray overlaps with a successive ray, a significant error in both the time and amplitude of arrival estimation is encountered.

Overlapping multipath components can be modeled by considering the relation

$$r(n) = c(n) * p(n) * h(n) + v(n) \tag{26}$$

where $\{r(n)\}$ continues to denote the received sequence, $\{c(n)\}$ is a known binary sequence, $\{p(n)\}$ is a known pulse-shape impulse response sequence, $v(n)$ is additive white Gaussian noise of variance σ_v^2, and $h(n)$ now refers to a multipath channel that is described by

$$h(n) = \sum_{l=1}^{L} \alpha_l x_l(n)\delta(n - \tau_l^o) \tag{27}$$

Here α_l, $\{x_l(n)\}$, and τ_l^o are respectively the unknown gain, the normalized amplitude sequence, and the time of arrival of the lth multipath component (ray). The above model assumes that there is a multipath component at each delay with corresponding amplitude α_l. In practice, most of these amplitudes will be zero or insignificant. For this reason, a common procedure is to estimate the amplitudes at all delays and to compare them to a threshold value that is proportional to the noise variance. If the amplitude α_l, at a specific delay τ_l^o, is larger than the threshold, then it is declared to correspond to a multipath component. The time and amplitude of arrival are then taken as the time and amplitude of the earliest ray higher than this threshold.

In this regard, the required estimation problem is one of estimating the vector of amplitudes at all possible delays, which is given by

$$\boldsymbol{h} \triangleq \text{col}[\alpha_1, \alpha_2, \ldots, \alpha_L]$$

Several least-squares-type methods have been suggested for this purpose [35–37]. These methods exploit the known transmitted pulse-shape to resolve overlapping rays. For example, it has been shown [37] that, under some reasonable assumptions, the vector $\boldsymbol{h}$ can be estimated by means of the following procedure. The received sequence is multiplied by delayed replica of the known transmitted sequence, $\{s(n-\tau)\}$. Each N sample of the resulting sequence is coherently averaged and the resulting averages at all delays are collected into a vector, say, $\boldsymbol{r}$. An estimate of $\boldsymbol{h}$ is then obtained from $\boldsymbol{r}$ by solving a least-squares problem, which leads to

$$\hat{\boldsymbol{h}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{r} \tag{28}$$

where $\boldsymbol{A}$ denotes a convolution matrix that is constructed from the pulse-shaping waveform. A general block diagram for such least-squares based techniques is shown in Fig.10. Alternative so-called *superresolution* techniques are also available that are based on methods known as ESPRIT and MUSIC (see the paper [22] and references cited therein).

Least-squares multipath resolving techniques, however, suffer from noise boosting, which is usually caused by the ill conditioning of the matrices involved in the LS operation. This ill-conditioning magnifies the noise at the output of the LS stage. For wireless location finding applications, where the received signal-to-noise ratio (SNR) is relatively low, noise magnification leads to significant errors in the time and amplitude of arrival estimates, which in turn result in low location precision. Other modified LS techniques that attempt to avoid matrix ill conditioning — such as regularized least-squares, total least-squares, and singular value decomposition methods — lack the required fidelity to resolve overlapping multipath components. Furthermore, applying least-squares methods may produce unnecessary errors in the case of single-path propagation.

An adaptive filtering technique for multipath resolving that avoids the aforementioned difficulties has been discussed [38]. Although adaptive filters do not suffer from noise amplification, they can still suffer from slow convergence and also divergence in some cases. These problems can be addressed by using knowledge about the channel autocorrelation and the fact that each channel ray fades at a different Doppler frequency.

6. HARDWARE IMPLEMENTATION ISSUES

It is clear from the previous considerations that signal parameter estimation for wireless location purposes often requires performing an extensive search over a dense grid of the estimated parameter (e.g., ToA estimation). The hardware implementation of these search schemes requires special attention as they might introduce a dramatic increase in the overall system hardware complexity and power consumption. In this section we review two hardware architectures for implementing ToA estimation schemes. The first scheme depends on combing the hardware of both channel and location searchers, while the second involves a Fast Fourier Transform (FFT)-based estimation scheme. Both architectures aim at reducing the overall hardware complexity.

6.1. Combined Channel/Location Searchers

A main hardware block in CDMA receivers is the conventional RAKE receiver, which consists of a dedicated channel searcher and a minimum of three RAKE fingers. Channel searchers obtain coarse estimates of the time and amplitude of arrival of the strongest multipath components of the MS signal. This information is then used by the receiver RAKE fingers and delay-locked loops (DLLs) to lock onto the strongest channel multipath components, which are combined and used in bit decoding. Estimates of the time and amplitude of arrival of the strongest rays are continuously fed from the channel searcher to the RAKE fingers.

Although the location searcher and RAKE receiver differ in purpose, structure, and estimation period, several basic building hardware blocks used in each of them are common. This fact can be exploited to combine both searchers into a single architecture that serves to save hardware blocks with added design flexibility.

Figure 11 shows the scheme, proposed in another paper [39], for the combined searcher architecture. The

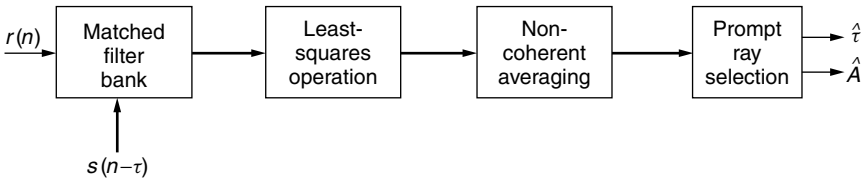


Figure 10. Multipath least-squares searcher.

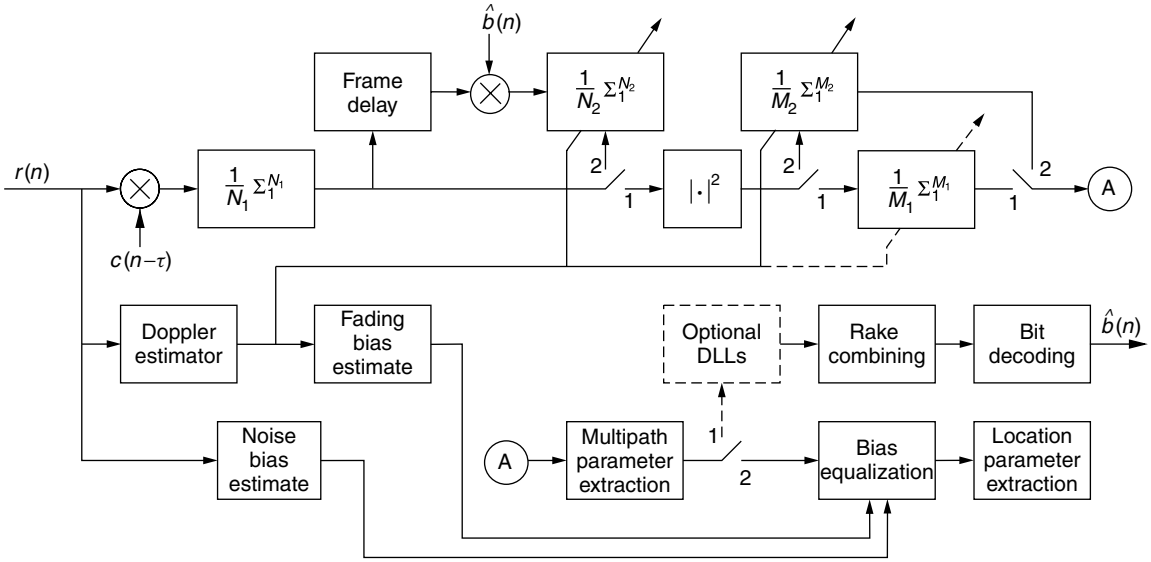


Figure 11. Combined architecture for location searcher and RAKE receiver.

scheme is formed from L_l data branches. Each data branch starts with a correlator over N_1 samples (despreader), where N_1 is the number of chips per symbol multiplied by the number of data samples per chip (4, 8, or 16). The output of the correlator is then multiplexed between two paths, marked "1" and "2" in Fig. 11. In path 1, which corresponds to data path of the channel searcher or RAKE finger branches, the despread signal is squared and noncoherently averaged over M_1 samples, where M_1 is *optionally* adapted to an estimate of the maximum Doppler frequency of the fading channel. For path 2, which is needed for the location parameter estimation, the despread sequence is delayed for a frame period, multiplied by an estimate of the transmitted bit sequence, coherently averaged over N_2 samples, squared, and noncoherently averaged over M_2 samples. Both N_2 and M_2 are adapted according to an estimate of the maximum Doppler frequency. The output of either paths is used to extract the channel multipath parameters.

The dynamic operation of the scheme is as follows. The received sequence is despread by multiplying by delayed code replica $c(n-\tau)$ and averaged over N_1 samples after which the N_1 register is reset. For online bit decoding, samples of the despread sequence are squared and noncoherently averaged. The average of every M_1 samples is passed to the multipath parameter extraction block and the M_1 register is then reset. Coarse multipath rays information are continuously fed to the L_f RAKE fingers, which use $3L_f$ branches of the scheme to obtain *early, on-time, and late* correlations over M_1 symbols. Such correlations are needed to advance or delay the sampling timing to lock onto the correct sampling point. This is done according to the difference between the early and late correlations [40]. The outputs of these fingers are combined, and used in bit decoding. Optional DLLs can be used to further enhance the tracking performance of the RAKE fingers. For location parameter estimation, the despread sequence is delayed, multiplied by an estimate of the transmitted bit sequence $\hat{b}(n)$, and continuously averaged over N_2 symbols. Every N_2 symbols, the N_2 register is reset and its output is squared using the shared squaring circuits and averaged over M_2 samples. After the total location estimation period ($N_2 \times M_2$ symbols), the time and amplitude of arrival of the prompt ray are equalized for fading and noise biases and used to extract needed location parameters.

This architecture has the following advantages:

1. Saving a large number of hardware building blocks via multiplexing basic hardware blocks between $3L_f + L_c$ location searcher and RAKE receiver branches.
2. Improving the performance of the RAKE receiver by continuously adapting the estimation period M_1 to an estimate of the maximum Doppler frequency. This period is conventionally adjusted to track a fading channel in the worst (fastest) case, which restricts this period to a small value (around 6 symbols for IS95 systems). Adapting the estimation period of the channel searcher has two advantages: (a) it will increase the accuracy of the delay and amplitude estimates, and (b) it will help save power as it will reduce the number of times the RAKE fingers need to change their lock point, especially for low maximum Doppler frequency cases.
3. Reducing the hardware complexity significantly by eliminating the need to use DLLs for fine tracking in the cases where the accuracy of the used combined architecture is $T_c/8$ or higher (which is typical for location applications). In such cases the accuracy of the RAKE receiver will be adequate for online bit decoding without the use of DLLs. Hardware implementation of DLLs is extremely complex, especially with regard to the analog front end [40].

Further details of the operation of this architecture is given in [39], along with performance simulation results.

6.2. FFT-Based Searchers

In mobile-based wireless location systems, a maximum-likelihood searcher is embedded in the MS handset. It is very common for such searchers to involve multiple *correlation* operations of the received signals, from cellular BSs or GPS satellites, with local delayed replica of the transmitted signals. Performing these extensive correlations in the time domain may be a burden for the MS hardware. Often these correlations are performed using a general-purpose DSP processor, which is embedded in the MS to perform many other tasks, including the correlation process. DSP processors have many advantages, such as low cost, versatility, and design flexibility. An efficient way of implementing correlation operations in this case is through the use of fast Fourier transform (FFT). FFT-based location searchers have shown significant efficiency when implemented on DSP processors [e.g., 15,41].

We now review the basic principles of operation of FFT-based location searchers. The output of the correlation operation, say, $y(\tau)$, between two sequences, $\{r(n)\}$ and $\{s(n)\}$, can be viewed as a sum of the form

$$y(\tau) = \sum_{n=1}^{N} r(n)s(n-\tau)$$

Evaluating this sum requires N multiplication operations for every value of the delay τ. Thus, computing $y(\tau)$ in the time domain needs N^2 multiplications. On the other hand, the correlation operation can be viewed as a multiplication of two sequences in the frequency domain, which has the form

$$Y(\omega) = R(\omega) \cdot S(\omega)$$

where $Y(\omega)$, $R(\omega)$, and $S(\omega)$ are the Fourier transforms of $y(\tau)$, $r(n)$, and $s(n)$, respectively. Thus, an alternate way of computing $y(\tau)$, is via

$$y(\tau) = \mathcal{F}^{-1}[R(\omega) \cdot S(\omega)].$$

Thus, the total number of multiplications needed to perform this operation is the number of multiplications needed to obtain the FFT of the sequences $r(n)$ and $s(n)$, the product of $R(\omega)$ and $S(\omega)$, and finally the inverse FFT of this product. Notice that if the sequence $s(n)$ is perfectly known, its FFT can also be known and stored instead of storing the signal $s(n)$ itself. Thus the total number of multiplications needed is given by $(N + N\log_2 N)$. When N is relatively large, this number of multiplications can be significantly less than N^2. For example, correlating GPS signals of length 1024 using the FFT approach is faster than performing the correlation in the time domain by a factor of 64 [15,41]. This directly reflects to a huge saving in the complexity and power consumption of the MS handset. We may also note that this procedure works only for sequences whose length is a power of 2. The general case can still be efficiently treated using the chirp-z transform (CZT), which can handle sequences whose length is not a power of two (See Refs. 15 and 41 for more details).

7. CONCLUDING REMARKS

As can be inferred from the discussions in the body of the article, and from the extended list of references below, wireless location is an active field of investigation with many open issues and with a variety of possible approaches and techniques. The final word is yet to come, which opens the road to much further work and, ultimately, to tremendous benefits.

Acknowledgments

This work was partially supported by the National Science Foundation under Award numbers CCR-9732376 and ECS-9820765. The authors would also like to thank Dr. L. M. A. Jalloul for his input, insights and collaboration in this area of research.

BIOGRAPHIES

Ali H. Sayed received his B.S. and M.S. degrees in electrical engineering from the University of Sao Paulo, Brazil, and his Ph.D. degree in electrical engineering in 1992 from Stanford University. He is professor of electrical engineering at the University of California, Los Angeles. He has over 170 publications, is the coauthor of two published books and the coeditor of a third book. He sits on the editorial boards of several journals including the *IEEE Transactions on Signal Processing*, the SIAM *Journal on Matrix Analysis and Its Applications*, and the *International Journal of Adaptive Control and Signal Processing*. He is also a member of the technical committees on signal processing theory and methods (SPTM) and on signal processing for communications (SPCOM), both of the IEEE Signal Processing Society. He has contributed several articles to engineering and mathematical encyclopedias and handbooks, and has served on the program committees of several international meetings. He has also consulted with industry in the areas of adaptive filtering, adaptive equalization, and echo cancellation. His research interests span several areas including adaptive and statistical signal processing, filtering and estimation theories, equalization techniques for communications, interplays between signal processing and control methodologies, and fast algorithms for large-scale problems. Dr. Sayed is a recipient of the 1996 IEEE Donald G. Fink Award and is a fellow of IEEE.

Nabil R. Yousef received his B.S. and M.S. degrees in electrical engineering from Ain Shams University, Cairo, Egypt, in 1994 and 1997, respectively, and his Ph.D. in electrical engineering from the University of California, Los Angeles, in 2001. He is currently a senior research staff member at Broadcom Corp., Irvine, California. His research interests include adaptive filtering, equalization, CDMA systems, and wireless location. He is a recipient of a 1999 Best Student Paper Award at an international meeting for work on adaptive filtering, and of a 1999 NOKIA Fellowship Award.

BIBLIOGRAPHY

1. FCC Docket 94-102, *Revision of the Commissions Rules to Insure Compatibility with Enhanced 911 Emergency Calling Systems*, Technical Report RM-8143, July 1996.

2. State of New Jersey, *Report on the New Jersey Wireless Enhanced 911 Terms: The First 100 Days*, Technical Report, June 1997.
3. J. H. Reed, K. J. Krizman, B. D. Woerner, and T. S. Rappaport, An overview of the challenges and progress in meeting the E-911 requirement for location service, *IEEE Commun. Mag.* **36**(4): 30–37 (April 1998).
4. J. J. Caffery and G. L. Stuber, Overview of radiolocation in CDMA cellular systems, *IEEE Commun. Mag.* **36**(4): 38–45 (April 1998).
5. J. J. Caffery and G. L. Stuber, Radio location in urban CDMA microcells, *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Toronto, Canada, Sept. 1995, Vol. 2, pp. 858–862.
6. J. M. Zagami, S. A. Parl, J. J. Bussgang, and K. D. Melillo, Providing universal location services using a wireless E911 location network, *IEEE Commun. Mag.* **36**(4): 66–71 (April 1998).
7. T. S. Rappaport, J. H. Reed, and B. D. Woerner, Position location using wireless communications on highways of the future, *IEEE Commun. Mag.* **34**(10): 33–41 (Oct. 1996).
8. L. A. Stilp, Carrier and end-user application for wireless location systems, *Proc. SPIE*, Philadelphia, Oct. 1996, Vol. 2602, pp. 119–126.
9. I. Paton et al., Terminal self-location in mobile radio systems, *Proc. 6th Int. Conf. Mobile Radio and Personal Communications*, Coventry, UK, Dec. 1991, Vol. 1, pp. 203–207.
10. A. Giordano, M. Chan, and H. Habal, A novel location-based service and architecture, *Proc. IEEE PIMRC*, Toronto, Canada, Sept. 1995, Vol. 2, pp. 853–857.
11. J. J. Caffery and G. L. Stuber, Vehicle location and tracking for IVHS in CDMA microcells, *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, Amsterdam, Netherlands, Sept. 1994, Vol. 4, pp. 1227–1231.
12. B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins, *Global Positioning System: Theory and Practice*, 2nd ed., Springer-Verlag, New York, 1993.
13. Special issue on GPS, *Proc. IEEE* **87**: 3–172 (Jan. 1999).
14. S. Chakrabarti and S. Mishra, A network architecture for global wireless position location services, *Proc. IEEE Int. Conf. Communications*, Vancouver, BC, Canada, June 1999, Vol. 3, pp. 1779–1783.
15. U.S. Patent 5,663,734 (Sept., 1997), N. Krasner, GPS receiver and method for processing GPS signals.
16. P. K. Enge, R. M. Kalafus, and M. F. Ruane, Differential operation of the Global Positioning System, *IEEE Commun. Mag.* **26**(7): 48–60 (July 1988).
17. Telecommunications Industry Association, The CDMA2000 ITU-R RTT Candidate Submission VO. 18, July 1998.
18. K. C. Ho and Y. T. Chan, Solution and performance analysis of geolocation by TDOA, *IEEE Trans. Aerospace Electron. Syst.* **29**(4): 1311–1322 (Oct. 1993).
19. Y. T. Chan and K. C. Ho, A simple and efficient estimator for hyperbolic location, *IEEE Trans. Signal Process.* **42**(8): 1905–1915 (Aug. 1994).
20. R. Schmidt, Least squares range difference location, *IEEE Trans. Aerospace Electron. Syst.* **32**(1): 234–242 (Jan. 1996).
21. B. T. Fang, Simple solutions for hyperbolic and related position fixes, *IEEE Trans. Aerospace Electron. Syst.* **26**(5): 748–753 (Sept. 1990).
22. H. Krim and M. Viberg, Two decades of array signal processing research: The parametric approach, *IEEE Signal Process. Mag.* **13**(4): 67–94 (July 1996).
23. T. Ojanpera and R. Prasad, *Wideband CDMA for Third Generation Mobile Communications*, Artech House, Boston, 1998.
24. R. Prasad, W. Mohr, and W. Konhauser, *Third Generation Mobile Communication Systems*, Artech House, Boston, 2000.
25. K. Kuboi et al., Vehicle position estimates by multibeam antennas in multipath environments, *IEEE Trans. Vehic. Technol.* **41**(1): 63–68 (Feb. 1992).
26. A. Ghosh and R. Love, Mobile station location in a DS-CDMA system, *Proc. IEEE Vehicular Technology Conf.*, May 1998, Vol. 1, pp. 254–258.
27. K. J. Krizman, T. E. Biedka, and T. S. Rappaport, Wireless position location: fundamentals, implementation strategies, and sources of error, *Proc. IEEE Vehicular Technology Conf.*, New York, May 1997, Vol. 2, pp. 919–923.
28. G. P. Yost and S. Panchapakesan, Automatic location identification using a hybrid technique, *Proc. IEEE Vehicular Technology Conf.*, May 1998, Vol. 1, pp. 264–267.
29. S. Glisic and B. Vucetic, *Spread Spectrum CDMA Systems for Wireless Communications*, Artech House, Boston, 1997.
30. N. R. Yousef and A. H. Sayed, A new adaptive estimation algorithm for wireless location finding systems, *Proc. Asilomar Conf.*, Pacific Grove, CA, Oct. 1999, Vol. 1, pp.491–495.
31. N. R. Yousef, L. M. A. Jalloul, and A. H. Sayed, Robust time-delay and amplitude estimation for CDMA location finding, *Proc. IEEE Vehicular Technology Conf.*, Amsterdam, Netherlands, Sept. 1999, Vol. 4, pp. 2163–2167.
32. R. Ertel et al., Overview of spatial channel models for antenna array communication systems, *IEEE Pers. Commun. Mag.* **5**(1): 10–22 (Feb. 1998).
33. A. Swindlehurst, A. Jakobsson, and P. Stoica, Subspace-based estimation of time delays and Doppler shifts, *IEEE Trans. Signal Process.* **46**(9): 2472–2483 (Sept. 1998).
34. G. Gutowski et al., Simulation results of CDMA location finding systems, *Proc. IEEE Vehicular Technology Conf.* Houston, TX, May 1999, Vol. 3, pp. 2124–2128.
35. Z. Kostic, M. I. Sezan, and E. L. Titlebaum, Estimation of the parameters of a multipath channel using set-theoretic deconvolution, *IEEE Trans. Commun.* **40**(6): 1006–1011 (June 1992).
36. T. G. Manickam and R. J. Vaccaro, A non-iterative deconvolution method for estimating multipath channel responses, *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, April 1993, Vol. 1, pp. 333–336.
37. N. R. Yousef and A. H. Sayed, Overlapping multipath resolving in fading conditions for mobile positioning systems, *Proc. IEEE 17th Radio Science Conf.*, Minuf, Egypt, Feb. 2000, Vol. 1, No. C-19, pp. 1–8.
38. N. R. Yousef and A. H. Sayed, Adaptive multipath resolving for wireless location systems, *Proc. Asilomar Conference on Signals, Systems, and computers*, Pacific Grove, CA, Nov. 2001.
39. N. R. Yousef and A. H. Sayed, A new combined architecture for CDMA location searchers and RAKE receivers, *Proc. Int. Symp. Circuits and Systems*, Geneva, Switzerland, May 2000, Vol. 3, pp. 101–104.

40. S. Sheng and R. Brodersen, *Low-Power CMOS Wireless Communications. A Wideband CDMA System Design*, Kluwer, Boston, 1998.

41. R. G. Davenport, FFT processing of direct sequence spreading codes using modern DSP microprocessors, *Proc. IEEE Natl. Aerospace and Electronics Conf.*, Dayton, OH, May 1991, Vol. 1, No. 3, pp. 98–105.

42. T. Rappaport, *Wireless Communications; Principles and Practice*, Prentice-Hall, Englewood Cliffs, NJ, 1996.

WIRELESS MPEG-4 VIDEOCOMMUNICATIONS*

MADHUKAR BUDAGAVI
Texas Instruments, Incorporated
Dallas, Texas

1. INTRODUCTION

With the success of personal mobile wireless phones for voice communications, there is now wide commercial interest and activity in extending the capabilities of the mobile phone to support videocommunications. Addition of video functionality to mobile phones leads to several new applications of the mobile phone — these include videotelephony, streaming video, video e-postcards and messaging, surveillance, and distance learning and collaboration. Mobile videotelephony enables users to not only talk to each other anywhere and at any time they want to, but it also allows them to see each other at the same time. Streaming video turns the mobile phone into a mobile entertainment device — it enables users to watch news and sports clips, music videos, and movie clips at any place and at any time they want to. It also allows mobile phone users to watch the video streaming from their home camera for surveillance purposes. Support for sending video e-postcards and messages in mobile phones enables users, for example, to send their vacation videos and photos directly from their vacation spots itself. Support for receiving instant video messages enables users to be immediately notified of any security event detected on their home surveillance camera. Users can take a look at the video of the security event enclosed with the instant video message and decide what action to take. Users can also use their video-enabled mobile phones to look at real-time educational lectures and videos and get trained on their long commute to work on trains. They can also use the video-enabled mobile phone to remotely collaborate from a worksite; for instance, they can use their mobile videophone to send real-time images of ongoing construction to their colleagues in their office and collaborate. One can similarly think of many other applications of a video-enabled mobile phone or device.

* Portions reprinted, with permission, from M. Budagavi, W. R. Heinzelman, J. Webb, and R. Talluri, "Wireless MPEG-4 video communication on DSP chips," *IEEE Signal Processing Magazine*, Vol. 17, No. 1, pp. 36–53, January 2000.

Wireless videocommunications is a multifaceted problem covering the fields of signal processing, wireless communications, data compression, transport protocols, and microelectronics. Supporting video transmission on wireless channels involves many technical challenges. Raw digital video data require a large amount of bandwidth; for instance, even a low-resolution (176×144-pixel) color video sequence at 15 frames per second (fps) requires 4.5 megabits per second (Mbps). The bandwidth available on current wireless channels is limited and also expensive. Hence it becomes important that the video data be compressed prior to transmission over wireless channels. Video sequences have redundancies in both the temporal (i.e., between adjacent video frames) and the spatial (i.e., within a video frame) domains. Video compression is achieved by removing these redundancies. Standard video compression algorithms usually make use of the following three steps to achieve efficient compression:

1. Predict the current video frame from the previous video frame (by using motion vectors) to remove temporal redundancy.
2. Then use the energy-compacting discrete-cosine transform (DCT) to encode spatial redundancy.
3. Finally, use entropy coding (variable-length coding) to encode the various parameters resulting from steps 1 and 2.

Advances in low-bit-rate video coding now enable a 176×144-pixel resolution color video sequence at 15 fps (which requires about 4.5 Mbps to be transmitted in the uncompressed form) to be compressed to about 32–64 kbps while still maintaining adequate viewing quality. The amount of compression that can be achieved is strongly dependent on the content in the video sequence. Video sequences with low motion can be usually compressed more efficiently than video sequences with high motion.

Current second-generation wireless systems provide data rates of only about 9.6–13 kbps. This amount of bandwidth is not enough for acceptable quality videocommunications. Advances in wireless technology and increased spectrum availability have lead to the development of third-generation (3G) wireless systems, which provide bandwidths of $\leq$384 kbps outdoors and $\leq$2 Mbps indoors. With this increased bandwidth availability, wireless videocommunications becomes possible. In fact, the first widely deployed mobile wireless videocommunication service was started recently in Japan [1]. This service, called *Freedom of Mobile multimedia Access* (FOMA), is based on the International Telecommunications Union (ITU)'s International Mobile Telephony (IMT) 2000 3G mobile communication standard. FOMA provides a 64-kbps circuit-switched wireless connection for videoconferencing and a 384-kbps downlink packet-switched wireless connection for streaming video.

A concern while transmitting compressed video data over wireless channels is that the wireless channel is a noisy channel and error bursts are commonly encountered on it because of multipath fading. The effect of channel errors on compressed video data can be deleterious.

Compressed video data are more sensitive to channel interference because of the absence of redundancy in the data. When the video bitstreams get corrupted on the wireless channel, predictive coding causes errors in the reconstructed video to propagate in time to future frames of video, and the variable-length codewords cause the decoder to easily lose synchronization with the encoder in the presence of bit errors. The end result is that the received video soon becomes unusable. Hence it becomes important that a good transport mechanism or protocol, one that provides adequate error protection to the compressed video bit stream, be used while transmitting the video data. Techniques such as forward error correction (FEC) channel coding and/or Automatic Repeat reQuest (ARQ) [2] are usually used for error protection when transporting video data over wireless channels. These techniques introduce redundancy in the transmitted data, thereby giving up some coding efficiency gains achieved by video compression. They also introduce additional delays in the system. In practice, depending on the bandwidth and delay constraints of the system, channel coding can be used to provide only a certain level of error protection to the video bit stream and it becomes necessary for the video decoder to accept some level of errors in the bit stream. Thus, it becomes essential to use error resilience techniques in the video coding scheme so that the video decoder performs satisfactorily in the presence of these errors.

Another challenge faced in wireless videocommunications is that compression and decompression of video and audio data are computationally very complex. Therefore, the processors used in the mobile phones must have high performance — they must be fast enough to play out and/or encode video in real time. Digital signal processors (DSPs) and application-specific integrated circuits (ASICs) are well suited for this task. The processors must also have low power consumption to avoid excessive battery drain and they must also be small enough to fit into compact form factors of mobile phones. Progress in microelectronics has enabled processors to satisfy these conflicting requirements. Note that the power consumption of the displays used to view the video also becomes important and it also needs to be low enough.

International standardization has also played an equally important part in facilitating wireless videocommunications. Standardization of video compression algorithms and communication systems and protocols allow devices from different manufacturers to interoperate — this brings economies of scale and mass production of equipment into picture, thereby facilitating cost-effective services.

In this article, we will focus on describing the relevant international standards that have made wireless videocommunications possible. We will cover standards that specify both the video compression as well as the *systems* for wireless videocommunications. In the next section, we start off by providing an overview of wireless videocommunication systems. We describe three categories of wireless videocommunication systems: messaging video systems, streaming video systems, and conferencing video systems. Wireless videocommunication systems are actually wireless multimedia communication systems since video is usually transmitted along with speech, audio, other multimedia data such as still pictures and documents, and control signals. In order to design a good videocommunication system, it is important to understand the interplay of video with the other components in the system. Therefore, we also provide an overview of the various components of a wireless multimedia system. The overviews provided in Section 2 will help in understanding why the various parts of wireless multimedia communication standards are required when we explain the standards in a later section. In Section 3, we talk about the Motion Pictures Experts Group (MPEG)-4 video compression standard [3]. MPEG-4 has been standardized by the International Standards Organization(ISO)/International Electrotechnical Commission (IEC). The MPEG-4 video coding standard caters to a wide range of multimedia applications covering a variety of storage media and transmission channels. We describe only those parts of the video coding standard that are suited for mobile wireless communications — this subset of the MPEG-4 video coding standard is called the *Simple Profile*. In Section 4, we describe systems standards specified by the Third Generation Partnership Project (3GPP) for messaging, streaming, and conferencing over 3G wireless networks. We conclude the article with discussions in Section 5.

2. OVERVIEW

There are basically three categories of wireless videocommunication systems: messaging video systems, streaming video systems, and conferencing video systems. One of the main factors that separates these three systems is the amount of playout delay in the receiver. Playout delay is the amount of time between the reception of video data in receiver and the playout of the received video data in the receiver. The playout delay determines whether a high-delay or a low-delay wireless connection is required. Another factor that separates these systems is whether both the video encoder and decoder or only the video decoder is used in the mobile phone — this determines whether a two-way or a one-way wireless video connection is required.

Figure 1 shows the block diagrams of the three wireless videocommunication systems. In *messaging video systems* (see Fig. 1a), a mobile phone wishing to send a video message (e.g., a vacation video clip) first creates the video message by capturing and encoding the video. It then uploads the complete video message onto a *multimedia messaging service center* (MMSC) along with the address of the mobile phone to which the video message is directed to. The mmsc then notifies the recipient mobile phone of the video message that has been sent to it. The recipient mobile phone then downloads the whole video message before playing it out. Video email is a variation of the messaging scheme explained above. In *video email*, the recipient mobile phone polls the email server to see if it has any email. If there is an email on the server, the mobile phone downloads it fully before playing it out. The playout

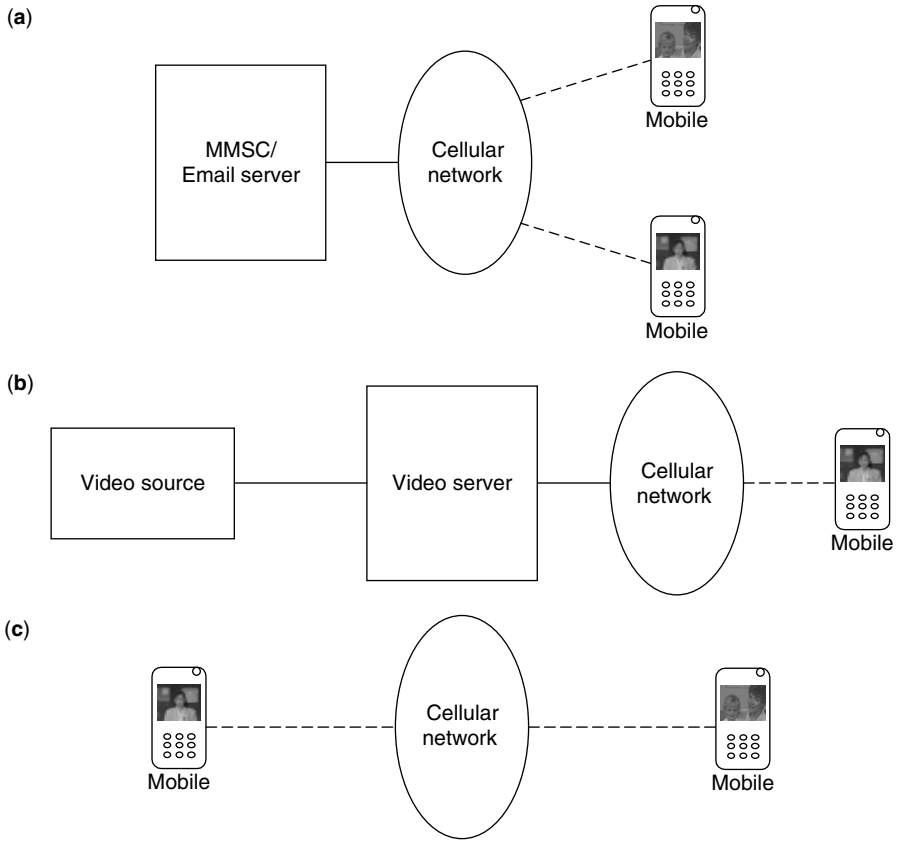


Figure 1. Three basic categories of wireless video systems: (**a**) messaging systems; (**b**) streaming systems; (**c**) conferencing systems. A solid line indicates a wireline connection and a dashed line indicates a wireless connection.

delay in the case of messaging video systems is the time taken to download the entire video message. Note that this playout delay may not be visible to the end user, since the downloading could be occurring in the background and the end user could be notified of the message only when the download is complete. Because of their download-and-play nature, it is not necessary for video messaging systems to have a low-delay connection. Also, the video encoder is not required if the mobile phone wants to have the capability of only receiving video messages.

In *streaming video systems* (see Fig. 1b), the video data received from the video server are buffered for a small amount of time (e.g., ~3 s) before being played out. This small buffer absorbs the delay jitters experienced by the video data sent on the wireless channel. The end result is that even if there is delay variation on the wireless channel, the video playout will still be smooth without any breakups. The playout delay in the case of streaming video systems is equal to the initial buffering delay. Note that streaming video systems require a connection that has a delay less than the initial buffering delay in order to have a smooth playout without any breaks. The streaming video data is sent in only one direction — from the video server to the mobile phone. In streaming video systems, a video encoder is not required on the mobile phone. The streaming video data can come from prestored video clips on the server, or they can come from live feeds of news and entertainment events. Note that the video is distributed from the source location to the video servers using a content distribution network, which is typically a wireline network.

Conferencing video systems (see Fig. 1c), which are used mainly for videoconferencing, have very strict delay requirements. The end-to-end delay must be less than 150 ms (though somewhat higher delays might still be acceptable) for the videoconference to be natural. Hence the video data that are received is played out as soon as possible. Also because of the two-way nature of the videoconference, videoconferencing systems require both the video encoder and decoder, and a two-way wireless channel for simultaneously transmitting and receiving the video data.

It is important to note that Fig. 1 illustrates wireless systems, but not wireless applications. In many cases, wireless applications can run on one or more of the wireless systems shown in Fig. 1. For example, a streaming video player *application* can be built on top of a conferencing video system or a streaming video system. The behavior of the streaming video player application will be different on the streaming and the conferencing video system. If the streaming video player application is built on top of a conferencing video system, since there is a very low-delay connection, the streamed video can be played out sooner. The uplink wireless channel (from the mobile phone to the cellular network) will remain unused in this case.

In each of three wireless video systems, video is usually transmitted along with speech/audio and other multimedia data such as images and documents. Therefore the mobile phone used for wireless videocommunications consists of various other components as shown in Fig. 2. The video codec, the audio codec, and the multimedia data blocks process (compress/decompress)

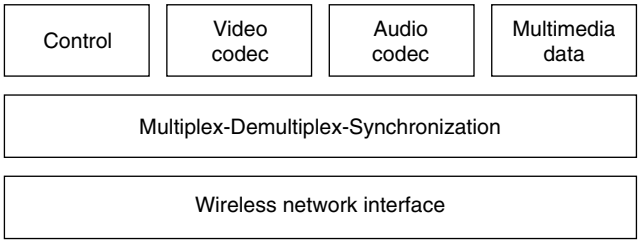


Figure 2. Components of a general wireless multimedia phone.

the video, audio, and multimedia data used in the multimedia communication session. In addition to these blocks, we have two other important blocks: control and multiplex–demultiplex–synchronization (MDS) blocks. The control block is used to initiate and teardown the multimedia communication session. It is also used to decide the audio and video compression methods to use and the data rates to use. The MDS block is used to combine the audio, video, multimedia data, and control signals into a single stream before transmission on the wireless network. In the receiver, it is used to demultiplex the received stream to obtain the audio, video, multimedia data, and control signals which are then passed on to their respective processing blocks. The MDS block is also used to synchronize and schedule the presentation of audio, video and other multimedia data.

In the next section, we describe the video codec block and in Section 4, we will look at the various manifestations of Fig. 2 as applied to messaging, streaming, and conferencing systems standards.

3. MPEG-4 SIMPLE PROFILE VIDEO COMPRESSION[1]

The *Simple Profile* of the MPEG-4 video standard [3] uses compression techniques similar to H.263 [5], with some additional tools for error detection and recovery. The scope of MPEG-4 simple profile is schematically shown in Fig. 3. As in H.263, video is encoded using a hybrid block motion compensation (BMC)/discrete-cosine transform (DCT) technique. Figure 4 illustrates a standard hybrid BMC/DCT video coder configuration. Pictures are coded in either *intraframe* (INTRA) or *interframe* (INTER) mode, and are called *I frames* or *P frames*, respectively. For intracoded I frames, the video image is encoded without any relation to the previous image, whereas for intercoded P frames, the current image is predicted from the previous reconstructed image using BMC, and the difference between the current image and the predicted image (referred to as the *residual image*) is encoded.

[1] This section and the figures appearing in this section have been taken from Ref. 4 with some modifications

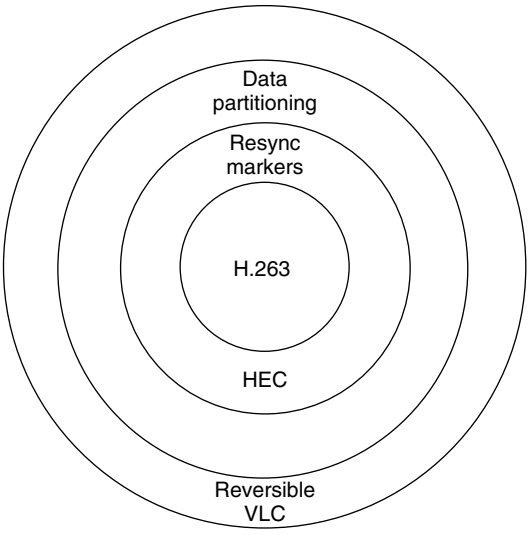


Figure 3. MPEG-4 simple profile includes error resilience tools for wireless applications. The core of MPEG-4 simple profile is the H.263 coder. Resynchronization markers, header extension code (HEC), data partitioning, and reversible VLCs provide error resilience support. From Fig. 4 of [4]. © 2000 IEEE with permission.

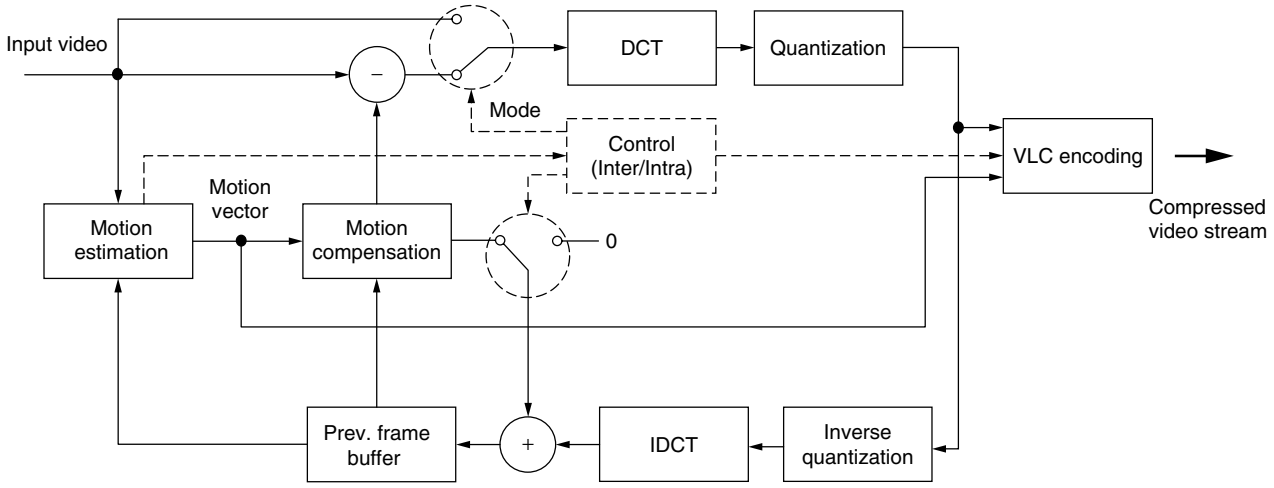


Figure 4. A standard videocoder based on block motion compensation and DCT. From Fig. 5 of [4]. © 2000 IEEE with permission .

The basic unit of information which is operated on is called a macroblock and is the data (both luminance and chrominance) corresponding to a block of 16×16 pixels. Motion information, in the form of motion vectors, is calculated for each macroblock in a P frame. MPEG-4 allows the motion vectors to have half-pixel resolution and also allows for four motion vectors per macroblock. Note that individual macroblocks within a P frame can be coded in INTRA mode. This is typically done if BMC does not give a good prediction for that macroblock. All macroblocks must also be INTRA-refreshed periodically to avoid the accumulation of numerical errors, but the INTRA refresh can be implemented asynchronously among macroblocks.

Depending on the mode of coding (INTER or INTRA) used, the macroblocks of either the image or the residual image are split into blocks of size 8×8, which are then transformed using the DCT. The resulting DCT coefficients are quantized, run-length-encoded, and finally variable-length-coded (VLC) before transmission. Since residual image blocks often have very few nonzero quantized DCT coefficients, this method of coding achieves efficient compression. For INTER-coded macroblocks, motion information is also transmitted. Since a significant amount of correlation exists between neighboring macroblocks' motion vectors, the motion vectors are themselves predicted from already transmitted motion vectors, and the motion vector prediction error is encoded. The motion vector prediction error and the mode information are also variable-length-coded before transmission to achieve efficient compression. In the decoder, the process described above is reversed to reconstruct the video signal. Each video frame is also reconstructed in the encoder, to mimic the decoder, and to use for motion estimation of the next frame.

Because of the use of VLC, compressed video bit streams are particularly sensitive to channel errors. In VLC, the boundary between codewords is implicit. Transmission errors typically lead to an incorrect number of bits being used in VLC decoding, causing loss of synchronization with the encoder. Also, because of VLC, the location in the bit stream where the decoder detects an error is not the same as the location where the error has actually occurred. This is illustrated in Fig. 5. Once an error is detected, all the data between the resynchronization points are typically discarded. The error resilience tools in MPEG-4 simple profile basically help in minimizing the amount of data that has to be discarded whenever errors are detected.

The error resilience tools included in the *simple* profile to increase the error robustness are

- Resynchronization markers
- Data partitioning
- Header extension codes (HECs)
- Reversible variable-length codes (RVLCs)

In addition to these tools, error concealment [6] should be implemented in the decoder. Also, the encoder can be implemented to limit error propagation using an adaptive INTRA refresh technique [7].

3.1. Resynchronization Markers

As mentioned earlier, a video decoder that is decoding a corrupted bit stream typically loses synchronization with the encoder due to the use of variable-length codes. MPEG-4 adopted a resynchronization strategy referred to as the "video packet" approach. Packetization allows the receiver to resynchronize with the transmitter when a burst of transmission errors corrupts too much data in an individual packet. A video packet consists of a resynchronization marker, a video packet header, and macroblock data, as shown in Fig. 6. The resynchronization marker is a unique code, consisting of a sequence of zero bits followed by a 1-bit, that cannot be emulated by the variable-length codes used in MPEG-4. Whenever an error is detected in the bit stream, the video decoder jumps to the next resynchronization marker to establish synchronization with the encoder. The video packet header contains information that helps in restarting the decoding process, such as the absolute macroblock number of the first macroblock in the video packet and the initial quantization parameter used to quantize the DCT coefficients in the packet. A third field, labeled HEC, is discussed in Section 3.4. The macroblock data part of the video packet consists of the motion vectors, DCT coefficients, and mode information for the macroblocks contained in the video packet.

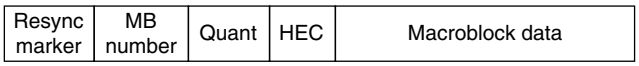


Figure 6. Resynchronization markers help in localizing the effect of errors to an MPEG-4 video packet. The header of each video packet contains all the necessary information to decode the macroblock data in the packet. From Fig. 7 of [4]. © 2000 IEEE. with permission.

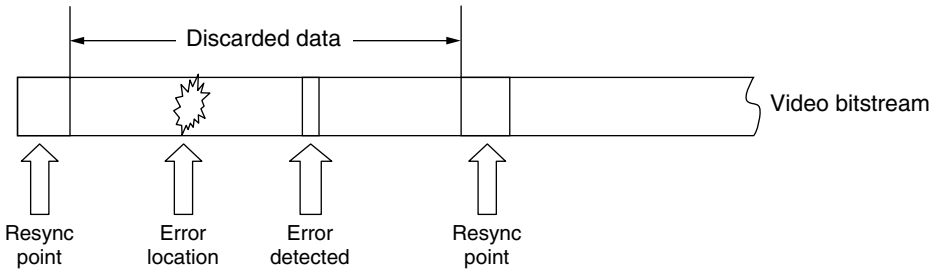


Figure 5. At the decoder, it is seldom possible to detect the error at the actual error occurrence location. From Fig. 6 of [4]. © 2000 IEEE with permission.

The predictive encoding methods are modified so that there is no data dependency between the video packets of a frame. Each video packet can be independently decoded irrespective of whether the other video packets of the frame are received correctly. A video packet always starts at a macroblock boundary. The exact size of a video packet is not fixed by the MPEG-4 standard (the standard does specify the maximum size that a video packet can take); however, it is recommended that the size of the video packets (and hence the spacing between resynchronization markers) be approximately equal.

3.2. Data Partitioning

The data partitioning mode of MPEG-4 partitions the macroblock data within a video packet as shown in Fig. 7. For I frames, the first part contains the coding mode and six dc DCT coefficients for each macroblock (4 for luminance and 2 for chrominance) in the video packet, followed by a dc_marker (DCM) to denote the end of the first part, as shown in Fig. 7a. (Note that the zeroeth DCT coefficient is called the *dc* DCT coefficient, the remaining 63 DCT coefficients are called *ac* coefficients.) The second part contains the ac coefficients. The DCM is a 19-bit marker whose value is 110 1011 0000 0000 0001. If only the ac coefficients are lost, the dc values can be used to partially reconstruct the blocks. For P frames, the macroblock data is partitioned into a motion part and a texture part (DCT coefficients) separated by a unique motion_marker (MM), as shown in Fig. 7b. All the syntactic elements of the video packet that are required to decode motion related information are placed in the motion partition and all the remaining syntactic elements that relate to the DCT data are placed in the texture partition. The MM indicates to the decoder the end of the motion information and the beginning of the DCT information. The MM is a 17-bit marker whose value is 1 1111 0000 0000 0001. If only the texture information is lost, data partitioning allows the use of motion information to conceal errors in a more effective manner. Data partitioning thus provides a mechanism to recover more data from a corrupted video packet.

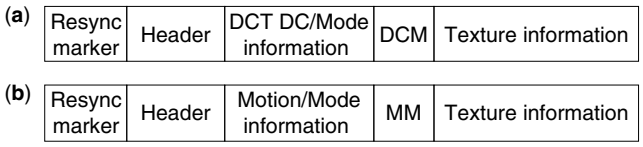


Figure 7. MPEG-4 data partitioned video packet for (**a**) I frames and (**b**) P frames. Data partitioning uses additional markers (DCM and MM) and puts the most important information in the first partition of video packet, for better error concealment. From Fig. 8 of [4]. © 2000 IEEE with permission.

3.3. Reversible Variable-Length Codes (RVLCs)

Reversible VLCs can be used with data partitioning to recover more DCT data from a corrupted texture partition. Reversible VLCs are designed such that they can be decoded both in the forward and the backward direction. Figure 8 illustrates the steps involved in two-way decoding of RVLCs in the presence of errors. While decoding the bit stream in the forward direction, if the decoder detects an error, it can jump to the next resynchronization marker and start decoding the bit stream in the backward direction until it encounters an error. Based on the two error locations, the decoder can recover some of the data that would have otherwise been discarded. Because the error may not be detected as soon as it occurs, the decoder may conservatively discard additional bits around the corrupted region. Note that if RVLCs were not used, more data in the texture part of the video packet would have to be discarded. RVLCs thus enable the decoder to better isolate the error location in the bit stream. Note that RVLC can be used only when data partitioning is enabled.

3.4. Header Extension Code (HEC)

Important information that remains constant over a video frame, such as the spatial dimensions of the video data, the timestamps associated with the decoding and the presentation of these video data, and the type of the current frame (INTER coded/INTRA coded), are transmitted in the header at the beginning of the video frame data. If some of this information is corrupted due to channel errors, the decoder has no other recourse but to discard all the information belonging to the current

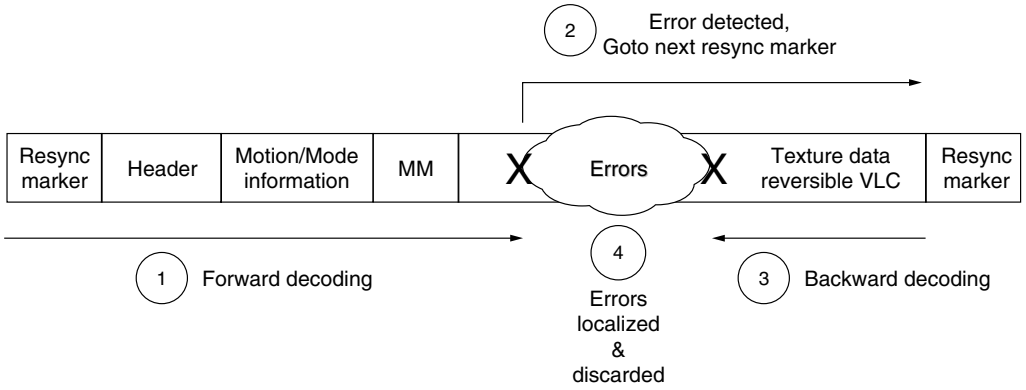


Figure 8. Reversible VLCs can be parsed in both the forward and backward directions, making it possible to recover more DCT data from a corrupted texture partition. From Fig. 9 of [4]. © 2000 IEEE with permission.

video frame. In order to reduce the sensitivity of this data, a 1-bit field called HEC is used in the video packet header. When HEC is set, the important header information that describes the video frame is repeated in the bits following the HEC. This duplicate information can be used to verify and correct the header information of the video frame. The use of HEC significantly reduces the number of discarded video frames and helps achieve a higher overall decoded video quality.

3.5. Error Concealment

The MPEG-4 standard does not specify what action the decoder should take when an error is detected. Several error concealment techniques have been developed based on temporal, spatial, or frequency-domain prediction of the lost data [6]. The simplest temporal concealment technique is macroblock copy. Under this procedure, corrupted macroblocks are replaced with collocated macroblocks from the previous frame. In practice this technique works quite satisfactorily when the amount of motion in video sequences is low, such as the head and shoulder video sequence type that arises in videoconferencing. More sophisticated temporal concealment techniques use the motion vector of the macroblock to copy the motion compensated macroblock from the previous frame. Sometimes the motion vector is available when data partitioning is used. In cases where the motion vector of the macroblock is lost, it is estimated from the motion vector of the neighboring macroblocks. However, temporal concealment cannot be used for the first frame (which is an I frame), and may yield poor results for intracoded macroblocks or areas of high motion. In such cases spatial domain error concealment techniques, wherein lost blocks are interpolated from correctly received neighboring blocks in the video frame, have to be used. Concealment in the spatial domain typically involves more computation because of the use of pixel-domain interpolation. In some cases, frequency-domain interpolation may be more convenient, by estimating the dc value and possibly some low-order ac DCT coefficients.

3.6. Adaptive INTRA Refresh (AIR)

AIR is a standard-compatible encoder technique for limiting error propagation by using nonpredictive INTRA coding [7]. INTRA refresh forcefully encodes some macroblocks in INTRA mode to flush out possible errors. INTRA refresh is very effective in stopping the propagation of errors, but it comes at the cost of a large overhead; coding a macroblock in INTRA mode typically requires many more bits than coding in INTER mode. Hence the INTRA refresh technique has to be used judiciously.

AIR adaptively performs INTRA refresh based on the motion in the scene. For areas with low motion, simple temporal error concealment works quite effectively. Since the high motion areas can propagate errors to many macroblocks, any persistent error in the high motion area becomes very noticeable. The AIR technique of MPEG-4 INTRA refreshes the motion areas more frequently, thereby allowing the possibly corrupted high motion areas to recover quickly from errors.

4. WIRELESS VIDEOCOMMUNICATION SYSTEM STANDARDS

In this section we briefly describe the wireless videocommunication system standards recommended by 3GPP for messaging, streaming, and conferencing. The standards are Multimedia Messaging Services (MMS) standard [8] for messaging, the Real-time Streaming Protocol (RTSP) standard [9,10] for streaming, the Session Initiation Protocol (SIP) standard [11,12] for videoconferencing over packet-switched networks, and the 3G-324 standard [13] for videoconferencing over circuit-switched networks.

It is useful to understand the characteristics of the network types used for these standards first before reading about the standards. The MMS, RTSP, and SIP standards are used over packet-switched networks whereas the 3G-324 standard is used over circuit-switched networks. Circuit-switched networks allocate a dedicated amount of bandwidth to the connection and hence they provide a predictable-delay connection. On the other hand, on packet-switched networks, data is packetized and transmitted over shared bandwidth and thus a predictable timing of data delivery cannot be guaranteed. The types of channel impairments observed on these two types of networks are different. On circuit-switched networks, the transmission errors experienced are in the form of bit errors, whereas on packet-switched networks, the transmission errors experienced are in the form of packet losses. On packet-switched networks, the predominantly used network layer protocol is the Internet Protocol (IP). There are two transport layer protocols developed for use with IP: the Transmission Control Protocol (TCP) and the User Datagram Protocol (UDP). TCP provides a reliable point-to-point service for delivery of packet information in proper sequence, whereas UDP simply provides a service for delivering packets to the destination without guarantee. TCP uses retransmission of lost packets to guarantee delivery. TCP is found to be inappropriate for real-time transport of audio video information because retransmissions may result in indeterminate delays leading to discernible distortions and gaps in the real-time playout of the audio/videostreams. In contrast, UDP does not have the problem of indeterminate delays because it does not use retransmission. However, the problem in using UDP for transmitting media is that UDP does not provide sequence numbers to transmitted packets. Hence UDP packets can get delivered out-of-order and packet loss might go undetected.

Before proceeding ahead in this section, it is also useful to revisit Fig. 2 and Section 2. All the standards follow the overall architecture of Fig. 2. They all differ in the type of control and multiplex–demultiplex–synchronization blocks they use.

4.1. Multimedia Messaging Services (MMS) Standard

Figure 9 shows the MMS [8] protocol stack for multimedia messaging over wireless IP networks. Each MMS multimedia message consists of a MMS header and an optional MMS body. The MMS header is used for signaling information between the mobile phone and the multimedia messaging service center (MMSC), and the MMS body

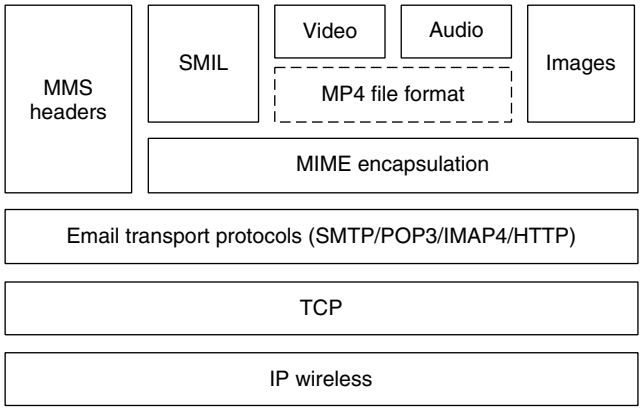


Figure 9. MMS protocol stack for multimedia messaging.

is used to carry the actual multimedia message data. The MMS header includes information on the type of the MMS message, specifically, whether it is a request from the mobile phone to the MMSC, a notification from the MMSC to the mobile phone, or a confirmation response to a request/notification. The header also contains the destination address, the date when the multimedia message was created, the address of the originating mobile phone, the version of the MMS protocol, and more such fields. The MMS body contains the media data and also the layout information, that is, information on where the various media components should be displayed on the screen and also as to when they should be played out and in what order. The presentation layout is specified using the Synchronized Multimedia Integration Language (SMIL) [14]. If audio and video must be played out in a synchronized fashion, then they must be packaged together using the MPEG-4 file format [15] first. The multiple media elements and the SMIL description are combined into a single composite entity using the Multipurpose Internet Mail Extensions (MIME) multipart format [16]. This final MIME encapsulated data forms the MMS body. The whole MMS message (MMS headers and the MMS body) is transmitted using transport protocols used for emails. The email transport protocols are layered on TCP which provides a reliable connection. TCP can be used in messaging systems because messaging systems can tolerate delays. At the receiving end, after the entire MMS message has been received, the mobile phone extracts the multimedia data and the SMIL description from the MIME encapsulated MMS body, and plays out the media according to the presentation information in the SMIL description.

Comparing Fig. 9 to Fig. 2 at a high level, the MMS headers forms the control block and the combination of SMIL, MPEG-4 file format, and MIME forms the multiplex–demultiplex-synchronization block.

4.2. Real-Time Streaming Protocol (RTSP)

The protocol stack for a RTSP-based [9,10] streaming media player is shown in Fig. 10. RTSP specifies a text-based protocol for exchanging control information with the video server. Control messages are used to establish the

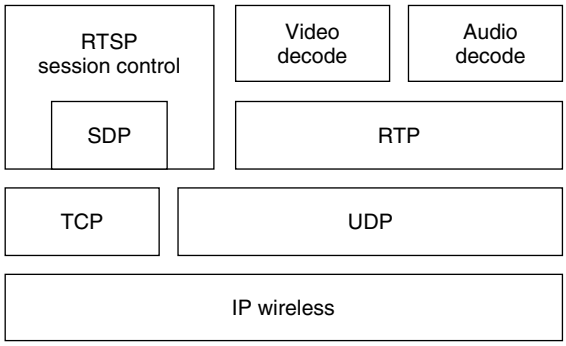


Figure 10. RTSP protocol stack for streaming video.

streaming session and to signal the type and format of the media to be used in the streaming session. They are also used for functions such as pausing, fast forwarding, and stopping the media playout. At the beginning of the streaming session, the mobile first sends a RTSP message to the video server specifying the media clip that it wants to watch. This is similar to sending a Webpage address to the Webserver for downloading the Webpage from the Webserver. The video server replies back, providing information on the type and format of the media in the media clip. The Session Description Protocol (SDP) [17] is used to provide this information. The SDP information is enclosed in the response from the server. The mobile phone looks at the enclosed SDP message and decides if it is able to decode the types of media present in the media clip. If it can, it then proceeds and issues a RTSP request to the video server to start streaming the media clip. The video server then starts streaming the media to the mobile phone. The mobile phone typically buffers the media for a short duration of time (e.g., 3 s) before playing them out. At anytime the streaming has to be stopped or paused, the mobile phone sends a RTSP request informing the video server to do so. The RTSP messages are usually transmitted reliably by using TCP.

The media are usually transmitted using UDP for prompt delivery. UDP does not provide timestamp information that is required in the playout of the media. Also as was stated earlier, by using UDP, media packets can get lost and can be delivered out-of-order. Hence the Real-time Transport Protocol (RTP) [18] is used on top of UDP to overcome the shortcomings of UDP. The RTP packet headers contain timestamps and sequence number information. The sequence number enables detection of packet loss and out-of-order packet delivery, and the timestamp provides timing information required in the playout of media.

Comparing Fig. 10 to Fig. 2 at a high level, the combination of RTSP and SDP forms the control block. Synchronization is carried out by RTP. There is no explicit multiplex–demultiplex block since the multiplexing and demultiplexing is carried out at the network level below IP.

4.3. Session Initiation Protocol (SIP)

Figure 11 shows the protocol stack for a SIP-based [11,12] mobile phone for videoconferencing over wireless IP

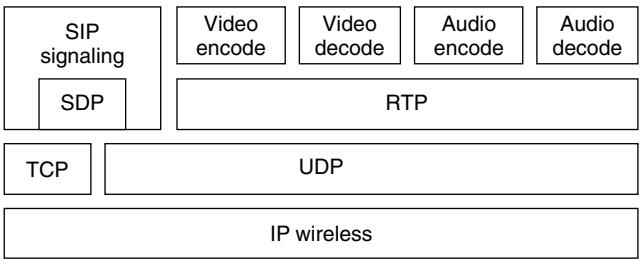


Figure 11. SIP protocol stack for two-way videoconferencing over packet-switched networks.

networks. As in the case of the RTSP streaming media player, RTP is used for transporting media. The SIP videoconferencing terminal contains both the encoder and the decoder for audio/video, unlike the RTSP streaming media player, which contains only the audio/video decoders.

The SIP protocol is used for signaling call setup and teardown. The SIP protocol can be layered on top of TCP or UDP. SIP also uses textual encoding of control messages. A SIP-based mobile phone wishing to place a call sends a message to the remote end inviting it to a call. Along with the invite message, the mobile phone also sends a SDP message describing the types and formats of media that it can receive and transmit during the call. If the remote end is ready to accept the call, it sends an acknowledgment to the invite. In the acknowledgment, the remote end sends back a SDP message indicating the media types and formats it can receive and transmit. Using the SDP messages, both endpoints can then decide on the media format that each will use for transmission and then start transmitting the media using RTP. Either end can end the call by sending a "Bye" message.

Comparing Fig. 11 to Fig. 2 at a high level, the combination of SIP and SDP forms the control block. Synchronization is carried out by RTP. There is no explicit multiplex–demultiplex block since the multiplexing and demultiplexing is carried out at the network level below IP.

4.4. 3G.324

For videoconferencing over circuit-switched wireless networks, 3GPP specifies the use of the 3G-324 standard [13]. The 3G-324 standard is based on the ITU standard for circuit-switched multimedia communications — H.324 [19]. The protocol stack for a 3G-324 videoconferencing terminal is shown in Fig. 12. Video, audio, and the control information are sent on distinct logical channels. The H.245 standard [20] is used to negotiate the media to be used in the call and also to set up the logical channels for the media. The H.223 standard [21] determines the way in which the logical channels are mixed into a single bit stream before transmission over the wireless channel. The H.223 standard was originally designed for operation over the benign public switched telephone networks. H.223 was later extended for operation over error-prone wireless channels. H.223 consists of two layers: the adaptation layer and multiplex layer. The adaptation layer allows for the use of both FEC and ARQ to protect the media being transmitted. This error protection is in addition to what might be provided by the wireless network. The multiplex layer is responsible for multiplexing the various logical channels into a single bit stream.

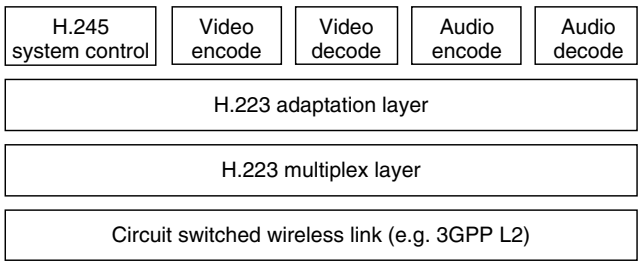


Figure 12. 3G.324 protocol stack for two-way videoconferencing over circuit-switched networks.

Comparing Fig. 12 to Fig. 2 at a high level, H.245 forms the control block and H.223 forms the multiplex–demultiplex-synchronization block.

5. DISCUSSION

In this article we provided an overview of wireless videocommunications and described the various international standards that have made it possible. In addition to video compression, error resilience techniques for video and efficient mechanisms for video transport are important in wireless videocommunications. We discussed the *Simple Profile* of MPEG-4 video coding standard which introduces several error resilience tools aimed at containing the effect of transmission errors in the video bit stream. In practice, video is usually transmitted along with speech, audio, multimedia data such as images and documents, and control signals to form a complete multimedia communication system. It is important to understand the interplay of video with the other components of the multimedia communication system. So we also provided an overview of the following systems standards that have been recommended by 3GPP for use over third generation wireless networks: MMS for multimedia messaging, RTSP for streaming, SIP for videoconferencing over packet-switched wireless networks, and 3G.324 for videoconferencing over circuit-switched wireless networks. It should be noted that though these standards were specifically described for use on wireless networks, many of them can be used and are being used on wireline networks too.

In addition to the standards described in this article, there are other video coding standards and proprietary techniques that can/will be used on wireless networks. ITU has also extended the H.263 to provide support for error resilience. In 3GPP standards, baseline H.263 is the mandatory video coder that has to be supported. Both MPEG-4 and the extensions to H.263 (called H.263++) are optional. The first commercial wireless video deployment — FOMA [1] — uses MPEG-4. For wireless streaming applications, proprietary video compression standards such as Quicktime, RealVideo and Microsoft Windows Media Video might also be used because of the wide availability of existing content in those formats on the World Wide Web.

Wireless videocommunications is a relatively new field, and there is a lot of ongoing research activity for improving the overall video quality. Techniques such as joint-source channel coding, where the amount of bits allocated to source coding and channel coding are adaptively varied according to channel conditions, and unequal error protection (UEP), where the level of error protection is varied per the importance of the data, are being studied. Layered video coding, where the video is split into a base layer and several enhancement layers that provide incremental quality improvement over the layers below them, can be used in conjunction with UEP with the base layer being protected the most. Low-complexity (and hence low-power) video compression is another important field of research. In addition to research in video compression, research activities in the fields of low-power semiconductors and displays, wireless communications and networking, are all simultaneously enabling wireless video to become a compelling application on mobile phones.

BIOGRAPHY

Madhukar Budagavi received the B.E. degree (first class with distinction) in electronics and communications engineering from the Regional Engineering College, Trichy, India, in 1991, and the M.Sc.(Engg.) degree in electrical engineering from Indian Institute of Science, Bangalore, India, in 1993, and the Ph.D. degree in electrical engineering from Texas A&M University, College station, Texas (USA), in 1998.

From 1993 to 1995, he was first a Software Engineer and then a Senior Software Engineer in Motorola India Electronics Ltd., primarily developing DSP software and algorithms for the Motorola DSP chips. Since 1998, he has been a Member of Technical Staff with the Texas Instruments DSP Solutions R&D center, working on MPEG-4 and protocols for wireless videocommunications. His research interests include video coding, speech coding, and wireless and Internet multimedia communications.

BIBLIOGRAPHY

1. NTT DoCoMo, *Freedom of Mobile Multimedia Access* (online), NTT DoCoMo (2002); *http://foma.nttdocomo.co.jp/english/catalog/network/index.html* (March 15, 2002).
2. S. Lin and D. J. Costello Jr., *Error control coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
3. International Standards Organization(ISO)/International Electrotechnical Commission (IEC), *Information Technology—Coding of Audio-Visual Objects—Part 2: Visual*, ISO/IEC 14496-2, 1999.
4. M. Budagavi, W. Rabiner Heinzelman, J. Webb, and R. Talluri, Wireless MPEG-4 video communication on DSP chips, *IEEE Signal Process. Mag.* **17**: 36–53 (Jan. 2000).
5. International Telecommunications Union (ITU), *Video Coding for Low Bitrate Communication*, Recommendation H.263, 1996.
6. Y. Wang and Q.-F. Zhu, Error control and concealment for video communications: A review, *Proc. IEEE* **86**: 974–997 (May 1998).
7. K. Imura and Y. Machida, Error resilient video coding schemes for real-time and low-bitrate mobile communications, *Signal Process. Image Commun.* **14**: 519–530 (May 1999).
8. 3rd Generation Partnership Project, Multimedia Messaging Services (MMS): *Functional Description*, Technical Specification 23.140, V5.1.0, 2001.
9. 3rd Generation Partnership Project, *Transparent End-to-End Packet Switched Streaming Service (PSS): Protocols and Codecs*, Technical Specification 26.234, V4.2.0, 2001.
10. Internet Engineering Task Force, *Real Time Streaming Protocol (RTSP)*, RFC 2326.
11. 3rd Generation Partnership Project, *Packet Switched Conversational Multimedia Applications: Default Codecs*, Technical Specification 26.235, V5.0.0, 2001.
12. Internet Engineering Task Force, SIP, *Session Initiation Protocol*, RFC 2543.
13. 3rd Generation Partnership Project, *Codec for Circuit Switched Multimedia Telephony Service: Modification to H.324*, Technical Specification 26.111, V3.4.0, 2000.
14. W3C, *Synchronized Multimedia Integration Language (SMIL 2.0)*, *http://www.w3.org/TR/2001/REC-smil20-20010807/* (Aug. 2001).
15. International Standards Organization(ISO)/International Electrotechnical Commission (IEC), *Information Technology—Coding of Audio-visual Objects—Part 1: Systems*, ISO/IEC 14496-1, 2001.
16. Internet Engineering Task Force, *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*, RFC 2046.
17. Internet Engineering Task Force, SDP, *Session Description Protocol*, RFC 2327.
18. Internet Engineering Task Force, RTP, *A Transport Protocol for Real-Time Applications*, RFC 1889.
19. International Telecommunications Union—Telecommunications Standardization Sector, *Terminal for Low Bit Rate Multimedia Communications*, Recommendation H.324, 1998.
20. International Telecommunications Union—Telecommunications Standardization Sector, *Control Protocol for Multimedia Communication*, Recommendation H.245, 1999.
21. International Telecommunications Union—Telecommunications Standardization Sector, *Multiplexing Protocol for Low Bitrate Multimedia Communication*, Recommendation H.223 and Annex A, B, C, 1997.

WIRELESS PACKET DATA

KRISHNA BALACHANDRAN
KENNETH BUDKA
WEI LUO
Lucent Technologies
Bells Laboratories
Holmdel, New Jersey

1. INTRODUCTION

Wireless packet data is defined as the transfer of information over wireless links that are shared dynamically

by multiple users without dedicating wireless resources to individual users during periods of inactivity.

In traditional circuit-mode cellular voice services, radio resources (e.g., frequencies, time slots, or codes) are assigned exclusively to a single user for the duration of a call. As a result, resources remain assigned during periods of inactivity. Circuit mode voice service is fairly wasteful of network resources. Studies of conversational patterns, for example, indicate that the amount of time speakers actually speak accounts for only 40–50% of the duration of the call. The remaining time is consumed by pauses in conversation. Due to the bursty nature of data transmissions, periods of activity for data services users tend to be much lower (e.g., 10–15%). Under these traffic assumptions, packet-mode operation allows much more efficient utilization of radio resources than circuit mode. In packet mode, a stream of bits is broken down into smaller units, or packets, which are individually transferred through the network. The communication resource may be dynamically shared by multiple users at any given time. Circuit mode operation wastes resources during periods of inactivity. With packet mode operation, the communication resource is shared by multiple users, and an active user can take advantage of the inactive periods of other users.

The most commonly used electromagnetic wireless access media include radio frequency (RF) waves and lightwaves. When the frequency is below 1 THz (terahertz), the electromagnetic waves are referred to as RF, and above 1 THz, they are called lightwaves. The most commonly used RF band for wireless communications is from 800 MHz to 100 GHz. The most commonly used lightwave wavelengths for wireless communications are infrared, with wavelengths ranging from 870 to 900 nm. The range of RF spectrum less than 800 MHz has been used by radio, TV broadcast, and transportation. Generally speaking, the higher the frequency, the more directional the propagation of electromagnetic waves and the shorter the range of the signal. At one extreme, lightwaves are usually limited to short range line-of-sight communications, such as indoor wireless and point-to-point inter-building communications. At the other extreme, RF spectrum at about 1 or 2 GHz is widely used for cellular wireless communications over large coverage areas because this spectrum does not experience interference from other services and it also provides good coverage in both rural and urban areas without any line-of-sight requirements. As a result, multiple users at different locations can easily share the radio channel. This is well suited for point-to-multipoint, broadcast or multicast communication.

The transmission characteristics of wireless channels are random, time-varying, and difficult to predict. Consequently, the data rate that can be achieved varies depending on the prevailing channel quality. Therefore, it is important to employ link adaptation algorithms that can measure the channel and quickly adapt to the physical-layer parameters in order to maximize the throughput under the prevailing channel conditions. Furthermore, wireless packet data channels are shared by multiple users, and the user throughput performance or system capacity is strongly dependent on the resource allocation and scheduling schemes employed. In general, resource allocation, scheduling, and link adaptation schemes should be carefully designed in order to maximize the user throughput and/or the system capacity.

Portability and mobility are two major advantages of wireless packet data. Wireless local area networks (WLANs), for example, IEEE 802.11 [1], were originally conceived to serve the purpose of portability within the office. Most lightwave-based wireless packet data services are also used to avoid wiring. Unlike WLANs, cellular packet data systems typically support lower data rates but provide wide-area coverage and support mobility. Third Generation (3G) networks will also be able to coarsely track user locations and provide location based services. Cellular packet data networks have evolved from cellular voice networks, and have inherited mobility support from voice networks. WLANs, however, were conceived in order to provide Ethernet-like connectivity without wires and have focused more on providing wireless access to the wired network. The support of mobility is a task left to Internet Engineering Task Force (IETF) protocols such as mobile IP. The distinction between cellular packet data networks and WLANs may blur and finally disappear as the convergence of the two types of networks becomes more evident.

The main purpose of wireless packet data is to connect mobile users to the fixed network and to connect mobile users to each other. In the wired network, packets generated by end users are directly forwarded to the destination if the destination is within the same local area network (LAN). Otherwise, packets are forwarded through a predetermined route or are routed hop by hop to the final destination. Wireless access usually refers to the "last mile" connection of the end-user devices to the network. Wireless access can be provided using a traditional cellular approach or alternatively, using wireless LANs. The cellular approach assumes that a mobile terminal communicates with a base station over a radio link. The base stations are connected to each other and to the Internet through a wired network. The wireless LAN approach assumes a more ad hoc structure, in which terminals may communicate with each other directly without the help of a base station or wired network. However, if a wireless terminal needs access to the Internet, the wireless LAN should provide an access point that is connected to the wired network. The cellular network model and ad hoc wireless LAN model are illustrated in Fig. 1.

2. GENERIC WIRELESS PACKET DATA SYSTEM ARCHITECTURE

Figure 2 shows a generic wireless packet data system protocol stack. Layer 1 represents the physical layer, which is responsible for radio signal transmission and reception. Layer 2 includes the medium access control (MAC) and radio link control (RLC) layers. The MAC layer allocates airlink resources to mobiles, thus allowing them to share the wireless link. The RLC protocol provides reliable in-sequence packet delivery when configured to operate in an acknowledged mode. Layer 3 provides

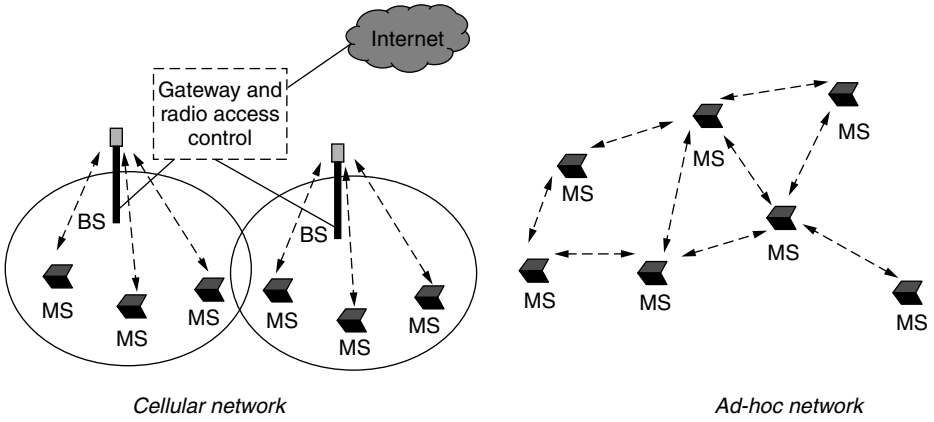


Figure 1. Two basic network connection structures. In cellular networks, mobile stations (MSs) are connected to the base station (BS) via a wireless link. The BS is connected to a gateway that controls the data transfer between the radio access network and the rest of the wired network. In ad hoc networks, MSs are connected to each other directly through radio or infrared links. A terminal can route data to other terminals with which radio links can be established so that terminals that are a few hops away can communicate with each other.

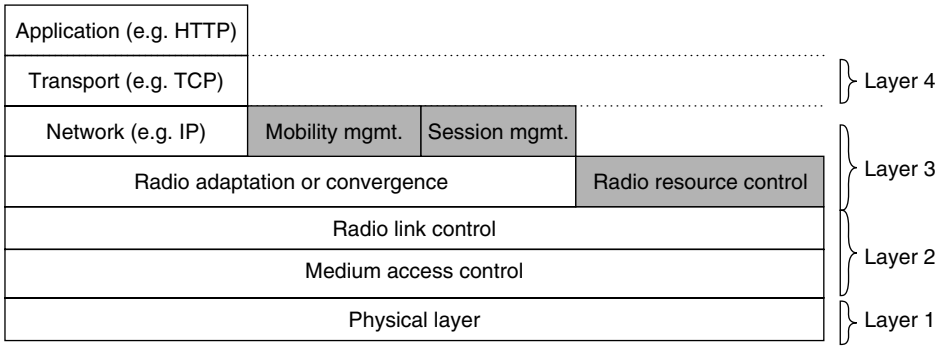


Figure 2. Protocol stack for wireless packet data (signaling plane protocols are shaded).

functions that connect radio access networks to the wired network, such as radio resource control, mobility management and routing within and beyond the radio access networks. In addition, layer 3 may also perform header and/or payload adaptation. The radius adaptation or convergence sublayer is also referred to as the logical link central (LLC) layer. Ciphering may be performed in layer 2 or in layer 3. Layer 4, the transport layer, provides reliable end-to-end data transfer and end-to-end flow control, if desired. The application layer denotes end user applications (e.g., HTTP for Web browsing).

All four layers of the protocol stack are implemented in mobile stations. On the fixed network side, however, portions of the protocol stack are implemented on different network elements. The choice of how to partition the protocol stack is influenced by technology requirements and/or constraints imposed by legacy systems. In General Packet Radio Service (GPRS) and Universal Mobile Telecommunications System (UMTS) systems, for example (Fig. 3) [2], the base station is responsible for physical-layer functions such as channel coding/decoding, modulation/demodulation, pulseshaping and transmission; MAC, RLC, and some radio resource control functions are in a remote radio network controller, which controls several base stations. The Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN) are responsible for session management (i.e., packet data service and context activation, authentication, and charging), mobility management and routing. Any two mobile stations that are served by the same GGSN can communicate with each other directly through the wired backbone (i.e., network of SGSNs) of the wireless packet data network. In a WLAN configuration, the base station (or access point) includes all user plane protocols up to layer 3. There is no dedicated wired backbone network in this case. Instead, the access point acts like any IP router within the wired network, and IP-based protocols specified by the IETF are used for mobility and session management (mobile IP, AAA (Authentication, Authorization and Accounting) protocols, etc.).

Ideally, wireless packet data systems should provide at least the same services that are currently provided over

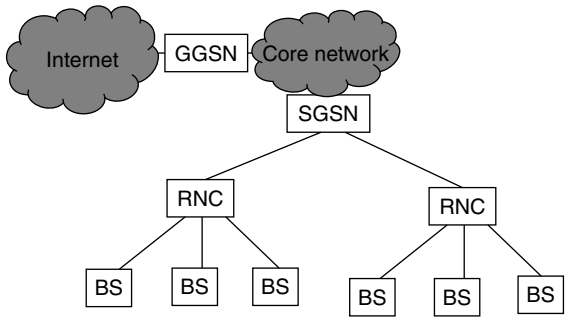


Figure 3. UMTS system architecture. A radio network controller (RNC) controls several base stations. Several RNCs are connected to a Serving GPRS Support Node (SGSN). The wired network of SGSNs is referred to as the "backbone" or "core network." The Gateway GPRS Support Node (GGSN) serves as a gateway between the wired backbone and the Internet or other virtual private networks.

wired networks. In addition, there may be services that are specifically targeted at mobile users. From an application perspective, wireless packet data systems should be able to provide these services with acceptable quality. From a business perspective, it is desirable to provide these services to as many users as possible. These objectives need to be satisfied under constraints on spectrum availability, cost and complexity. Furthermore, services targeted at mobile terminals should be enabled through small and inexpensive terminals with long battery life. These issues pose several challenging problems for wireless packet data system design, as will be discussed later.

3. QUALITY OF SERVICE REQUIREMENTS

Wireless packet data technologies support data transfer services with a wide range of quality of service (QoS) levels. End-to-end QoS requirements (e.g., loss and delay) can be broken down into corresponding requirements on the wireless data network (i.e., user equipment to a wireless gateway node) and on the fixed network. The QoS classes and attributes employed by the wireless data network should span a wide range of possible applications. Because of the limited amount of bandwidth available at the air interface, support of QoS must not involve significant overhead or complexity and should allow efficient utilization of resources.

The third-generation partnership project (3GPP) specifications define four QoS classes [3]:

- *Conversational Class*. The requirements for this class are similar to those for conventional telephony. In order to maintain acceptable quality, there are strict limits on transfer delay and loss rate (1–3%). Furthermore, the time relation between information entities (e.g., speech frames) needs to be preserved.
- *Streaming Class*. Streaming audio and video applications fall into this category. Streaming applications are more tolerant of packet transfer delays than conversational applications. The time relations (variation) between information entities (i.e., samples, packets) within a flow are to be preserved. In addition, loss rates must be low (1–3%).
- *Interactive Class*. Web browsing, database queries and other applications that follow a request/response pattern are considered interactive. Payload content must be preserved (i.e., lossless transfer). In addition, limits are placed on acceptable round-trip delay
- *Background Class*. This class of traffic is delay insensitive (i.e., best effort). Examples include electronic mail (email) and short message service (SMS). This class requires that the payload content be preserved but does not have stringent delay requirements.

Because of signaling and packet transfer delays, airlink error detection and recovery takes time. The stringent delay requirements for voice services, however, do not allow sufficient time for error recovery. Due to advances in speech coding, the rates demanded by voice services are quite low ($<12-16$ kbps). The typical approach is to pad transmissions with enough redundancy to allow operation under poor channel conditions. The channel coding is fixed, and no retransmissions are allowed. This error avoidance approach is not well suited to the support of data services. Since the delay requirements for data services are typically more relaxed than voice, error detection and recovery techniques can be used. It is inefficient to use a fixed amount of redundancy independent of the actual delay requirements and the prevailing channel quality. Higher data rates can be supported if error recovery is carried out using an automatic repeat request (ARQ) scheme.

For best-effort data services, the radio link control (RLC) layer typically allows *full recovery* [i.e., service data units (SDUs) are not discarded and there is no limit on the maximum number of retransmissions] and delivers data in sequence to the higher layer. This scheme is best suited to the support of best-effort data services over wireless links, and not to applications such as streaming where the time relation between information entities needs to be preserved. For streaming, a selective ARQ scheme with SDU discard and limited retransmission (i.e., partial recovery) capability can achieve better performance.

4. WIRELESS PACKET DATA TECHNOLOGIES

4.1. Performance Measures

4.1.1. Delay. From the user perspective, it is meaningful to consider measures such as the transfer delay for a speech frame, file, web object or webpage. The air interface is often the largest contributor to "end-to-end delay" perceived at the application layer. A widely used measure of performance for protocol performance in wireless data systems is the SDU delivery delay over the air interface (i.e., delay between peer RLC protocol layers).

The RLC SDU delivery delay may be defined as the time between SDU arrival at the transmitter RLC to in-sequence delivery by the receiver RLC. The RLC SDU delivery serves as a good measure of performance over the air interface for streaming, interactive, and background data services.

For data traffic with varying SDU sizes, a more appropriate measure of performance is the ratio of the SDU delivery delay to SDU size. This is typically known as the *normalized* SDU delivery delay.

SDU delay is a random quantity. Delay jitter (variance) is important when constant delivery rate is required. If the transmission control protocol (TCP) is used for end-to-end error detection and recovery, for example, the delay jitter must kept as small as possible. This will be elaborated on later. Another metric is delay percentiles, specifying the percentage of transferred packets that experience a delay exceeding a certain value. This is often used in the case when a packet delay exceeds a certain threshold and the packet is dropped.

4.1.2. Throughput. For a given choice of physical-layer parameters (coding, spreading, modulation), an upper bound on the expected user throughput over the air interface is given by $R \cdot (1 - b)$, where R denotes the peak

data rate and b denotes the expected error rate of each RLC packet. Throughput can be improved further when selective hybrid ARQ (i.e., incremental redundancy, which will be discussed later) is employed for radio link control.

There are usually two types of throughput measures mentioned in the literature. One is *user-perceived throughput*, which is usually used as one of the metrics to quantify the quality of service provided to the mobile user. The user-perceived throughput is the individual user's data throughput when the user has data to transmit. Wireless packet data channels are shared by multiple users and the user-perceived throughput computation should additionally account for the queuing time. In the literature, the data rate quoted for the system usually refers to user-perceived throughput. The other type of throughput is *aggregated throughput* or *system throughput*, which is usually used to quantify the system capacity or system load. The aggregated throughput is the amount of bits successfully transmitted per unit time in the whole system that includes all users. User-perceived throughput usually should be used together with system aggregated throughput, because in many cases, system aggregated throughput is proportional to the system loading and the user-perceived throughput decreases when the system loading increases.

4.1.3. Packet Loss and In-Sequence Delivery. Wireless packet data systems typically require in-sequence SDU delivery over the air interface. If selective ARQ is employed, this translates into a requirement on the RLC protocol. Note that in-sequence delivery does not imply lossless data transfer.

SDU loss can occur in one of the following ways:

- Buffer overflow at the transmitter.
- SDU discard at the transmitter if the delivery delay requirements cannot be met.

The loss rate requirement depends on the QoS requirements of the application.

4.1.4. Coverage and Mobility Support. It is desirable to satisfy a minimum throughput or a maximum SDU delivery delay (or normalized SDU delivery delay) for a large fraction (90–95%) of users in the network.

Most cellular data networks allow users to roam over a wide area and attempt to provide an "always on" experience for the end user (i.e., mobility management is handled without significant impact on the service). In addition, for real-time services, handovers should occur seamlessly as a mobile terminal moves from the coverage area of one sector to another. This is achieved by imposing very stringent delay requirements on a handover. Generally speaking, there is a tradeoff between providing high-throughput service and providing large coverage and good mobility support. For example, at the time when this article is written, WLANs provide a data rate as high as 11 Mbps (with future standards targeting rates up to 54 Mbps), but coverage is typically limited to indoor environments with no mobility. On the contrary, the 3G wireless networks can support only 384 kbps, optimistically, but with almost seamless coverage and mobility support.

4.1.5. Security. Wireless packet data systems must provide safeguards against unauthorized network access and protect users' privacy. This is achieved through security protocols that carry out authentication and ciphering of user data and protect the integrity of control information.

To prevent unauthorized access, users need to be authenticated before they can access the wireless network. This typically involves comparison of a user-provided unique equipment ID number with ID numbers stored and maintained by the network. The network may also ask for a password to check the user's identity.

The content of the user's data, location, identity, and data usage pattern must all be protected. The use of wireless links makes it easier for casual eavesdroppers to infringe on a user's privacy. The objective of the security provided through wireless packet data is to provide security at levels comparable to wireline data service. This requires that user data be encrypted before they are sent over the air. In addition, the network should try to reduce the times that the mobile user sends its unique ID number over the air. This prevents eavesdroppers from deriving a user's location and data usage pattern. Once a user is registered with the network, the network can assign a temporary ID for the user and the temporary ID number can be used for authentication. It should be emphasized that the wireless data networks do not provide end-to-end security. Such security must be provided by the users/applications themselves or IP-based security protocols (e.g., IPSec).

4.1.6. Energy Efficiency. Many mobile terminals are powered by batteries and energy efficiency is quite important. Low mobile terminal power consumption is critical in order to lengthen communication time and to prevent mobile devices from overheating.

The definition of energy efficiency is not as simple as it appears to be. At first glance, energy efficiency can be defined as the amount of energy expended per information bit received. But this definition does not capture the data rate. Generally speaking, the higher the data rate, the higher the amount of energy expended per unit of information bits received. This is due mainly to two effects: (1) the higher data rate implies higher receiver processing requirements and (2) increasing data rate over the air interface requires higher transmission power per bit. According to Shannon theory, the required transmission power is an exponentially increasing function of the supported data rate. If the wireless link is shared by multiple users in certain ways, increasing one user's data rate leads to a higher interference level being experienced by other users. Therefore, the other users need more transmission power to combat the increased interference level. From a system perspective, if multiuser interference is a concern, the higher the aggregated throughput of the system, the higher the energy requirement per information bit. A

good indication of energy efficiency, therefore, is a power consumption function corresponding to the supported data rate.

Reduction of power consumption in wireless packet data involves optimization of all aspects of the system and circuit design, from hardware and software to communication protocols. For example, on the RF portion of a mobile terminal's circuitry, the use of low peak-to-average modulation schemes and nonlinear amplifiers improve amplifier efficiency and saves energy. On the digital logic part of a mobile terminal's circuitry, low-voltage and low-clock-frequency circuits are preferred for power reduction. Integration of the system into a small number of chips is desirable to reduce I/O power consumption. Discontinuous transmission and reception[1] modes that allow the mobile to go idle if there are no data to send or receive are very useful in reducing power consumption. All the power reduction approaches discussed must ensure that they do not compromise performance.

4.2. Technology Methods

4.2.1. Link Adaptation and Incremental Redundancy. Link adaptation is a technique that uses channel quality measurements in order to select the physical-layer parameters such as modulation, coding, and spreading that are needed in order to achieve the highest throughput under delay constraints [4]. The RLC block sizes and physical-layer parameters should be chosen in such a way that blocks can be retransmitted without significant overhead even if the physical-layer parameters used for the initial transmission are different from those used for retransmission.

Incremental redundancy is a technique that can be applied in conjunction with link adaptation in order to achieve higher throughput under a wide range of operating conditions. For each set of physical-layer parameters, a set of puncturing schemes (P1, P2, P3, etc.) achieving the same code rate are defined. All puncturing is carried out on the same rate $1/N$ "mother code." The initial transmission of a block consists of the bits obtained by applying the puncturing scheme P1 (for the chosen physical-layer parameters) to the rate $1/N$ encoded data. On receiving a negative acknowledgment for the RLC block, additional coded bits (i.e., the output of the rate $1/N$ encoded data punctured with scheme, P2, corresponding to the prevailing set of physical-layer parameters) are transmitted. If all the punctured versions of the encoded data block have been transmitted, the cycle is repeated, starting again with P1. If the receiver does not have sufficient memory for incremental redundancy operation, it can attempt to decode the data by using the information corresponding to just P1, or P2, or P3. This corresponds to the pure link adaptation case. For incremental redundancy operation, the receiver must have sufficient memory in order to store soft information corresponding to RLC blocks that have not yet been decoded successfully. Each time the receiver obtains additional coded bits, it attempts soft-decision decoding using this redundant information in addition to previously stored soft information corresponding to the same RLC block(s).

The individual puncturing schemes are designed to be as disjunctive (or nonoverlapping) as possible in order to achieve good performance with incremental redundancy. In addition, to achieve good performance with pure link adaptation, the puncturing schemes are designed to ensure that the individual schemes (P1, P2, P3) achieve comparable error performance.

4.2.2. Peak Picking Scheduling. In a cellular environment, different users sharing the same airlink will observe different link performance. In addition, the performance observed by each user may vary in time due to changes in interference levels and shadow fading.

Figure 4 shows an idealized plot of logical link control (LLC)-layer efficiency variations as a function of time as link adaptation tracks the changes in link quality.

To increase overall system capacity, a scheduler can dynamically allocate larger portions of airlink resources to those users with high link efficiencies, a technique known as "peak picking." To help understand the role peak picking plays in the design of airlink schedulers, it is helpful to look at two extremes:

- One naive scheduling algorithm is to allocate all network resources to the mobile with the highest link efficiency. Such an algorithm maximizes the total amount of data carried over the airlink. However, such an algorithm is woefully impractical. One problem with such a scheduler is fairness. Using

[1] Discontinuous reception is commonly referred to as "sleep mode."

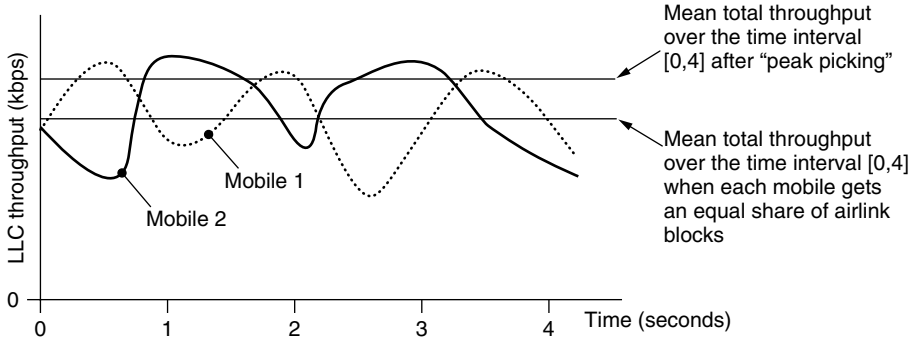


Figure 4. Peak picking can substantially improve overall system capacity. The curves represent the throughputs that mobile 1 and mobile 2 would receive if they were given all airlink blocks. With peak picking, network operators will be able to generate more revenue from the airlink.

such a scheme, it is possible that one user in the cell is granted all airlink bandwidth, while all other users starve. Such bandwidth starvation can have disastrous effects on the performance of higher-layer protocols, such as TCP. Such starvation can result in frequent TCP retransmissions and TCP connection failures: focusing on maximizing airlink throughput alone may end up causing transmission inefficiency at higher layers. In addition, it is often necessary for the network to periodically schedule transmissions to/from a mobile so that it has current information needed for power control and link adaptation to function properly.

- At the other extreme is simple round-robin scheduling, in which each user is given an equal fraction of airlink blocks. Such a scheduling scheme does not take advantage of peak picking.

Effective schedulers employing peak picking lie somewhere between these two extremes. The design of such schedulers is currently an active area of research.

4.2.3. Header Compression. Wireless data networking technologies are designed to make most efficient use of airlink spectrum. Link adaptation and incremental redundancy techniques are one way of doing this. Another way of increasing link efficiency is to shrink the size of user data packets carried over the wireless data network's airlink. Two additional tools are used for this purpose: packet header compression and packet payload compression.

Each TCP/IP packet contains a 20-byte TCP header and a 20-byte IP header. For small payloads, 40 bytes of overhead is a large price to pay. For this reason, wireless data networks typically employ techniques to compress the headers of each network-layer packet carried over the airlink. Such schemes are based on the packet header compression scheme developed by Van Jacobsen [5]. Taking advantage of the fact that many of the fields in TCP/IP packet change little over the lifetime of a TCP connection, Van Jacobsen's header compression algorithm can reduce the amount of data needed to carry a TCP/IP packet header from 40 to 3 bytes. Similar techniques are also being proposed to reduce the headers used by UDP/IP headers to enable efficient carrying of audio streaming and voice over Internet Protocol (VoIP) applications, which typically have payloads of less than 100 bytes.

Packet payload compression schemes can also increase airlink efficiency. V.42bis data compression is a common compression scheme used by wireless data networks. V.42bis is based on the string compression algorithm of Ziv–Lempel [6]. V.42bis maps strings appearing in the uncompressed input data to a set of codewords. Both the compressor and decompressor dynamically construct a dictionary mapping codewords to strings. By sending shorter-length codewords over the airlink instead of the longer length strings the codewords represent, V.42bis can achieve favorable compression ratios and make more efficient use of the airlink. V.42bis, however, may be used only over links providing lossless, in-sequence packet delivery.

4.2.4. Mobility Management. Wireless data networks employ several mobility management techniques to give mobile terminals access to the Internet. Some wireless data network technologies leverage the use of mobility management architecture already deployed for the service of circuit-switched traffic (e.g., GPRS/EGPRS, UMTS). Other wireless data technologies (e.g., CDMA2000) have opted instead for mobility management based on mobile IP.

4.2.4.1. Mobility Management in GPRS/EGPRS and UMTS. Figure 5 shows the high-level mobility management architecture employed by GPRS/EGPRS and UMTS networks. In these technologies, each user is assigned a "home" service provider, the provider who bills for their service. Mobiles are also permitted to use networks other than those owned by their home service provider, a process known as "roaming."

Mobility management in GPRS/EGPRS and UMTS networks uses the following network entities:

- *Visitor Location Register* (*VLR*). This is a database containing information on the roaming mobiles currently active in a service provider's network. The VLR contains information on the roaming current mobiles' locations, the identities of the mobiles' home service provider, information needed to authenticate users, and the capabilities of roaming mobiles and other information. This database also contains information on roaming circuit-switched customers currently using the service provider's network.
- *Home Location Register* (*HLR*). Similar to the VLR, this database contains information on all mobiles that receive service from the service provider. This database also contains information on the service provider's circuit-switched customers.

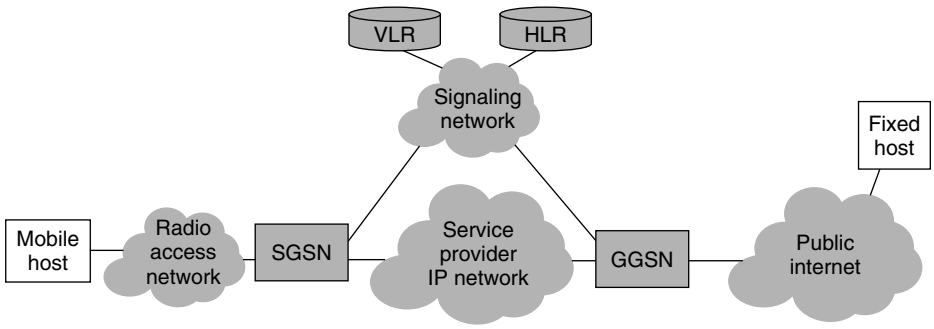


Figure 5. Mobility management architecture employed by GPRS/EGPRS and UMTS networks.

- *GGSN*. The GGSN is a gateway node used to mask the mobility of mobile hosts from Internet applications. All data traffic sent between a mobile host and fixed host are routed via the GGSN. The GGSN collects usage information that can be later used to bill users for wireless data service. A GGSN may access the VLR/HLR to obtain information on mobile capabilities and special features that a user has subscribed to. The GGSN tunnels network layer traffic to the SGSN that a mobile host is currently attached to.
- *SGSN*. The SGSN tracks a mobile as it moves from cell to cell, delivering the network-layer packets it receives from the GGSN. The SGSN may compress the header and payload of the packets it transmits. The SGSN tunnels packets it receives from the mobile host to the GGSN. The SGSN also communicates with the VLR/HLR to retrieve mobile capabilities and update location information and mobile state. The SGSN also collects mobile-specific usage data that can be later used for billing.

Mobility management "procedures" are used to activate or deactivate data service and track mobiles as they move from cell to cell. During each procedure, signaling messages are exchanged between the mobile host and network entities:

- *Attach*. Before being able to transfer data over a wireless data network, a user must "attach." During the attach procedure, signaling messages are exchanged between the mobile host and the SGSN to identify and authenticate the mobile. Credentials supplied by the mobile may be compared with credentials the SGSN retrieves from the mobile's HLR. The SGSN updates the mobile's HLR with its new location. If the state information in the HLR shows that the mobile was attached with another SGSN, the new SGSN informs the old SGSN that the mobile will no longer need service from the old SGSN. The SGSN assigns the mobile a temporary link layer address which will be used to identify the mobile for as long as it remains attached.
- *Detach*. The detach procedure may be initiated by a mobile (called an "explicit detach"), or initiated in response to a lack of routing area update messages (called an "implicit detach"). The detach procedure informs the wireless data network that a mobile is no longer available to send and receive data over the wireless data network. During a detach, the VLR is updated to indicate that the mobile is now idle.
- *Packet Data Protocol* (*PDP*) *Context Activation*. PDP context activation makes a mobile "visible" to the public Internet. Signaling messages sent between the mobile and the SGSN identify the type of service that the mobile is requesting. The SGSN may perform optional procedures to determine whether the mobile is allowed access to the type of service it is requesting. If the SGSN determines the context activation should be allowed, it informs a GGSN to create a PDP context for the mobile. Once a PDP context has been created, IP traffic can flow between a mobile and the public Internet.
- *Packet Data Protocol* (*PDP*) *Context Deactivation*. PDP context deactivation is used to signal the end of a data session. Once a mobile's PDP context has been successfully deactivated, the SGSN and GGSN that were serving the mobile are free to assign network resources (memory, link bandwidth, etc.) to other mobiles. At the end of PDP context deactivation, a mobile is no longer reachable from the public internet.
- *Routing Area Update*. To help manage the amount of signaling traffic needed to track a mobile as it moves from cell to cell, contiguous clusters of cells are grouped together to form a "routing area." Mobiles are required to inform the SGSN any time they begin to receive service in a cell with a new routing area identifier, a process known as a *routing area update*. Routing areas are also used to control the amount of paging traffic that must be generated to deliver traffic to mobiles in standby mode. Paging messages are only sent in those cells belonging to the routing area the mobile was last known to be in. Routing area updates are sent periodically by mobiles in standby mode. If a mobile does not send periodic updates, the network will detach the mobile, freeing up resources for other mobiles.
- *Paging*. Constant reception and decoding of airlink channels drains a mobile's battery. To increase battery life, during times of inactivity, mobiles enter a "standby mode." While in standby mode, mobiles periodically decode paging channels to determine whether the network wished to send traffic to them. Periodic decoding of paging channels can substantially increase batter, life, since data transfer tends to be sporadic.

Once a mobile host has attached and activated a PDP context, it is able to exchange IP packets with the public Internet. A "ping" [Internet Control Message Protocol (ICMP) echo] message sent from a mobile host to a fixed host is carried over the radio access network to the SGSN currently serving the mobile. The source address of the IP packet carrying the ping message is the IP address assigned to the mobile during PDP context activation. The SGSN tunnels the ping message to the GGSN currently serving the mobile's PDP context. The GGSN passes the ping message to the public Internet, where traditional IP routing is used to deliver the message to the fixed host. The fixed host echoes the ping message back to the mobile host. The destination IP address used by the fixed host is the IP address assigned to the mobile host during context activation. IP packets using the IP address of the mobile host are routed using traditional IP routing to the GGSN serving the mobile host's PDP context. The GGSN tunnels the IP packet to the SGSN serving the mobile host. The SGSN then forwards the packet to the cell currently being used by the mobile host.

4.2.4.2. Mobile IP. Mobile IP is a protocol defined by the Internet Engineering Task force to deliver

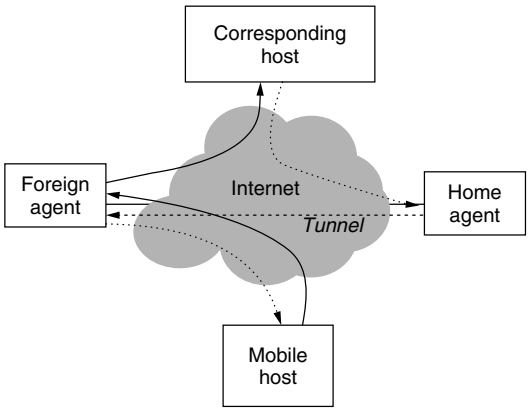


Figure 6. Mobile IP mobility management architecture.

network layer packets to mobile hosts [7]. Here we highlight a few key differences between the mobile IP mobility management approach and the approach used in GPRS/EGPRS and UMTS networks.

Mobile IP high-level mobility management architecture is shown in Fig. 6:

- *Foreign Agent*. The foreign agent plays a role analogous to the SGSN in GPRS/EGPRS/UMTS networks. The foreign agent is responsible for sending and receiving network layer packets to the mobile host.
- *Home Agent*. The home agent plays a role analogous to the GGSN in GPRS/EGPRS/UMTS. Data packets sent to a mobile host are first routed to the home agent. The home agent then tunnels the packets to the foreign agent currently serving the mobile host. The foreign agent "detunnels" packets forwarded by the home agent and delivers them to the mobile host.

Mobile IP employs a technique known as "triangular routing" to transfer packets to and from mobile hosts. Network-layer packets sent from a corresponding host to a mobile host, are always routed via the home agent. However, packets sent by a mobile host to the corresponding hosts bypass the home agent, and are routed directly to the corresponding host. This is in contrast to the route traveled by packets sent by mobile hosts in GPRS/EGPRS/UMTS, where packets are always routed via the GGSN, regardless of the direction they are being sent — so-called bidirectional tunnels.

4.2.5. TCP over Wireless. TCP is the predominant wireline transport layer protocol used by internet applications [8]. TCP was designed to guarantee end-to-end reliable data delivery and end-to-end flow control over wireline networks. One key assumption underlies the design of TCP's flow control and error recovery mechanisms — packet loss and delay are caused primarily by network congestion. In wireless data networks, however, packet loss and delay are caused primarily by airlink transmission errors. TCP flow control and error recovery mechanisms can cause severe end-to-end performance degradation when used over wireless links.

Two problems are encountered when TCP is used on wireless data networks. TCP has its own ARQ mechanism that often conflicts with the ARQ mechanism employed by a wireless data network's RLC layer. TCP segments are typically divided into several RLC PDUs, which are then scheduled for transmission. When a wireless link is bad, many retransmissions occur in the RLC layer, increasing the delay experienced by individual TCP segments. If a TCP segment is not received within certain period of time, the TCP transmitter assumes that the segment is lost and retransmits it. However, the retransmission of the TCP segment is not necessary because the first TCP segment was not lost, but delayed as a result of airlink errors. These spurious TCP retransmissions waste airlink bandwidth and reduce TCP throughput.

Another problem with the use of TCP over wireless data networks is TCP's end-to-end flow control mechanism. When TCP sees that a segment is delayed too much or lost, TCP "thinks" that there is congestion on the path between the transmitter and the receiver. Therefore, TCP reduces the transmission rate. When TCP receives a segment, it may think that the congestion has disappeared and then increases the transmission rate. However, the reason for packet delay fluctuation and loss is due mainly to the fluctuation of wireless link qualities and their induced network reactions. The way in which TCP reacts to packet delay and loss — quickly reducing transmission rates in response to excessive delay, and slowly increasing transmission rates when the delay decreases — is not well suited to the delays observed over wireless links. As a result, TCP transmission rate reductions may be too aggressive, and the wireless link may end up never fully utilized.

To solve the performance problems caused by using TCP over wireless data networks, the packet delay seen at the RLC layer should be made as smooth as possible. In this way, TCP will have an accurate estimate on the round-trip delay and then will not unnecessarily retransmit a packet or reduce the transmission rate. The delay jitter seen at the RLC layer may come from various sources, such as RLC retransmission and resequencing delay, connection teardown and resetup in the middle of a TCP session, and wireless scheduling. Although all of those causes may not be possible to eliminate, at least the designer should be aware of this and trade off the TCP problems against other considerations.

5. CONCLUSIONS

As a result of the interplay of each layer of the wireless data protocol stack, the need for spectrally efficient data transmission, the limitations placed by mobile terminal battery life, the design and engineering of wireless data networks offers many unique challenges. The wireless data networking technologies discussed in this article provide a variety of different approaches used to solve these problems.

BIOGRAPHIES

Wei Luo received a B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1995 and

M.S. and Ph.D. degrees in electrical and computer engineering from University of Maryland, College Park, Maryland, in 1997 and 1999, respectively. Dr. Luo joined Bell Labs, Lucent Technologies in 1999 working on the performance, protocol and algorithm designs in wireless communication systems.

Kenneth C. Budka received his B.S. (summa cum laude) degree in electrical engineering from Union College in Schenectady, New York, in 1987 and M.S. and Ph.D. degrees in engineering science from Harvard University in Cambridge, Massachusetts, in 1988 and 1991, respectively. During his undergraduate and graduate studies, he was an exchange scholar at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland, and Columbia University in New York, New York. Dr. Budka joined AT&T Bell Labs in 1991 working on control, resource allocation, scheduling, and performance issues arising in second and third generation wireless voice, and data communications technologies. He was named a Bell Labs Distinguished Member of Technical Staff in 1999, and currently works as a technical manager in Lucent Technologies Bell Labs High Performance Communication Systems Laboratory. A senior member of the Institute of Electrical and Electronics Engineers, Dr. Budka holds two patents in the area of wireless voice and data networks, with more than 10 others pending.

Krishna Balachandran received his B.E. (Hons) degree in electrical and electronics engineering from Birla Institute of Technology and Science, Pilani, India, M.S. in computer and systems engineering, M.S. in mathematics and Ph.D. in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, New York. He joined Bell Labs, Lucent Technologies in 1996 as a member of technical staff and is currently an acting technical manager in the Networking Technologies and Performance Department at Bell Labs. His research interests include the design, analysis and simulation of advanced physical layer, media access control and radio link control protocols, link adaptation/hybrid ARQ techniques, resource allocation, and scheduling techniques for wireless systems. Dr. Balachandran has published over 30 journal and conference papers in related areas. He also holds three patents (several pending) and has contributed significantly to standardization activities in the TIA, ETSI, and 3GPP.

BIBLIOGRAPHY

1. IEEE std 802.11b-1999, Supplements to ANSI/IEEE std 802.11, 1999 ed.
2. 3GPP TS 25.401, *UTRAN Overall Description*, v.3.1.0, Jan. 2000.
3. 3GPP TS 23.107, *QoS Concept and Architecture*, v.3.2.0, March 2000.
4. S. Nanda, K. Balachandran, and S. Kumar, Adaptation techniques in wireless packet data services, *IEEE Commun. Mag.* **38**(1): (Jan. 2000).
5. V. Jacobsen, *Compressing TCP/IP Headers for Low-Speed Serial Links*, IETF RFC 1144, 1990.
6. J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans. Inform. Theory* **23**(3): 337–343 (May 1977).
7. C. Perkins, *IP Mobility Support*, IETF RFC 2002, Oct. 1996.
8. W. R. Stevens, TCP/IP Illustrated, Vol. 1, *The Protocol*, Addison-Wesley, Reading, MA, 1994.

WIRELESS SENSOR NETWORKS

Seapahn Megerian
Miodrag Potkonjak
University of California at Los Angeles
West Hills, California

1. INTRODUCTION

Since the early 1990s, the sustained high pace of technological advances paved the way for the exponential growth of the Internet. We can trace the development of two implementation technologies as prime enablers of this growth. The first was the dramatic reduction in the cost of disks, namely, massive long-term storage. The second was the huge reduction in the cost of optical communication and its simultaneous capacity increase. More specifically, since 1991, the capacity of a $100 disk increased by a factor of 1200, while during the same period, the bandwidth of optical cable doubled every 9 months. The Internet, as we know it today, is an exceptional educational, research, entertainment, and economic resource, which enables information to be available at the touch of a mouse. There is a wide consensus that the Internet will continue to grow rapidly in both quantitative and qualitative terms. At the same time, it appears that we are on the brink of the next technological revolution that may have even more profound impact on our lives. This revolution, that will enable any time, anywhere, communication and connection between the physical and computational worlds, is due to the advancement of wireless communication technology and sensors. While in the early 1990s wireless technology was mainly stagnant, since 1996, it has experienced an exponential growth. Wireless bandwidth in industrial offerings has increased by a factor of 28 from 1997 to 2002. On the other hand, recent progress in fabrication of micro-electromechanical systems-based (MEMS) sensors has opened new vistas in terms of cost, reliability, accuracy, and low energy requirements. While most of the MEMS-based sensors are still in the research phase, a boom in government funding in this area has resulted in amazing progress. For this field, the total funding was $2 million in 1991 and $35 million in 1995, while in 2001 it was estimated to have been $300 million worldwide. With such advancement, there is currently a need for methodologies and technologies that will enable efficient and effective use of wireless embedded sensor network applications. The motivational factors pushing for these applications include the mobility of computational devices, such as cellular phones and personal digital assistants (PDAs), and the ability to embed these devices into the physical world.

Almost all of the modern science and engineering has been built using compound experiment–theory iteration steps. Typically, the experiments have been the expensive and slow components of the iterations. Thus, the existences of flexible yet economic experimentation platforms often result in great conceptual and theoretical breakthroughs. For example, advanced optical and infrared telescopes enabled spectacular progress in the understanding of large scale cosmology theory. Particle accelerators and colliders enabled great progress in the understanding the ultra small world of elementary particles. Furthermore, the progresses in computer science, information theory, and nonparametric statistics have been greatly facilitated by the ability to compile and execute programs quickly on general-purpose computers. Sensor networks will enable the same type of progress in better understanding many other sciences, not just by information processing, but also through new connections between the sciences and the physical, chemical, and biological worlds.

Sensor networks consist of a set of sensor nodes, each equipped with one or more sensors, communication subsystems, storage and processing resources, and in some cases actuators. The sensors in a node observe phenomena such as thermal, optic, acoustic, seismic, and acceleration events, while the processing and other components analyze the raw data and formulate answers to specific user requests. The recent advances in technology mentioned above, have paved the way for the design and implementation of new generations of sensor network nodes, packaged in very small and inexpensive form factors with sophisticated computation and wireless communication abilities. Once deployed, sensor nodes begin to observe the environment, communicate with their neighbors (i.e., nodes within communication range), collaboratively process raw sensory inputs, and perform a wide variety of tasks specified by the applications at hand. The key factor that makes wireless sensor networks so unique and promising in terms of both research and economic potentials is their ability to be deployed in very large scales without the complex preplanning, architectural engineering, and physical barriers that wired systems have faced in the past. The term "ad hoc" generally signifies such a deployment scenario where no structure, hierarchy, or network topology is defined a priori. In addition to being ad hoc, the wireless nature of the communication subsystems that rely on radio frequency (RF), infrared (IR), or other technologies, enable usage and deployment scenarios that were never before possible.

To illustrate the key concepts and a possible application of wireless ad hoc sensor networks (WASNs), consider the environmental monitoring requirements of large office buildings. Such buildings typically contain hundreds of environmental sensors (such as thermostats) that are wired to central air conditioning and ventilation systems. The significant wiring costs limit the complexity of current environmental controls and their reconfigurability. Furthermore, in highly dynamic corporate environments, cubicle offices may continuously be added, removed, and restructured, which makes environmental control rewiring an intractable task. However, replacing the hard-wired monitoring units with inexpensive ad hoc wireless sensor nodes will easily improve the quality and energy efficiency of the environmental system while allowing unlimited reconfiguration and customization in the future. In addition to the classic temperature sensing, senor nodes with multiple modalities (i.e., equipped with several different types of sensors) can significantly enhance the abilities of such a system. For example, motion or light sensors can detect the presence of people and even adjust the environmental controls using actuators, according to prespecified user preferences. In many instances, the savings in the initial wiring costs alone may justify the use of such wireless sensor nodes.

Although the environmental monitoring example above is an application of WASNs to a task that has existed for a long time, many new applications have also started to emerge as direct consequences of WASN developments. Such applications range from early forest fire detection and sophisticated earthquake monitoring in dense urban areas, to highly specialized medical diagnostic tasks where tiny sensors may even be ingested or administered into the human body. As mentioned above, personal spaces such as offices and living rooms can be customized to each individual by sensors that detect the presence of a nearby person and command the appropriate actuators to execute actions according to that person's preferences. In essence, WASNs provide the final missing link connecting our physical world to the computational world and the Internet. Although many of these sensor technologies are not new, technological barriers and physical laws governing the energy requirements of performing wireless communications have limited their feasibility in the past. A few highlights and benefits of the newer, more capable sensor nodes are their abilities to

- Form very large-scale networks (thousands or more nodes).
- Implement sophisticated networking protocols as well as distributed and localized analysis algorithms.
- Reduce the amount of communication required to perform tasks by distributed and/or localized computations.
- Implement complex power-saving modes of operation depending on the environment, current tasks, and the state of the network.

In the following sections, we describe the generic components that form a wireless sensor network and highlight the key issues and characteristics that differentiate sensor networks from traditional peer-to-peer and ad hoc wireless communication networks. Section 2 lists the architectural and hardware related components, while in Section 3 the focus is on higher-level services and software issues. Section 4 provides a brief overview of the state of the art and the challenges ahead.

2. ARCHITECTURE AND HARDWARE

Similar to classical computer architectures, the main components of the physical architecture of WSN nodes

can be classified into four major groups: (1) processing, (2) storage, (3) communication, and (4) sensing and actuation [input/output (I/O)]. The following is a brief summary of the main issues involved and some related topics for each of these components.

2.1. Processing

Two key constraints for processing components are energy and cost. Essentially all current WSN processors are those used for mass markets. This is due in large part to the advantages of the economies of scale and the availability of comprehensive and mature software development environments for such processors. Since the processing in a node has to address a variety of different tasks, many nodes have several types of processors: microprocessors and/or microcontrollers, low-power digital signal processors (DSPs), communication processors, and application-specific integrated circuits (ASICs) for certain special tasks. The standard complementary metal oxide semiconductor (CMOS) process will be the technology of choice for sensor node processors at least until 2012.

2.2. Storage

Currently, sensor nodes have relatively small storage components. They most often consist of standard dynamic random access memory (DRAM) and relatively large quantities of nonvolatile (flash) memory. Since the communication is a dominant component of the overall energy consumption is wireless sensor networks, we expect that the amount of local storage at a node will continue to increase. This expectation is further enforced by the fact that since 1992, the cost of memory was declining much faster than the cost of processors. We also expect that new technologies, in particular magnetoresistive random-access memory (MRAM), will soon be widely used for this type of storage.

2.3. Communication

The communication paradigms often associated with the current generations of wireless sensor networks are multihop communication. Several current results indicate that multihop communications scale very well and can significantly reduce the energy consumption in large sensor networks [1]. A number of new projects are currently targeting low power communication. This is an area where it is most difficult to predict how technology will impact future architectures, since commercial wireless communication is a relatively new field. It is very important to note here that in typical low-power radios used in WASN communication, listening often requires as much energy as transmitting. This is in sharp contrast to the assumptions made in most previous work in ad hoc multihop networking, where sending a message was believed to have been the major consumer of energy. These new constraints indicate that the study of complex power saving modes of operation, such as having multiple different sleep states, will be crucial in this field.

2.4. Sensors and Actuators

One can envision the sensors as the eyes of the sensor network, and the actuators, as its muscles. Although MEMS technology has been making steady progress since the early 1960s, it is obvious that it is still in its early phases where development is mainly sustained by research funding and not yet commercial. However, significant results have already been obtained. A good starting point for learning more about sensor systems is the article by Mason et al. [2].

3. SYSTEM SOFTWARE AND APPLICATIONS

As described above, the recent advent of WASNs has required completely new approaches for building system software and optimization algorithms, as well as the adaptation of existing techniques. It is interesting and important to analyze why the already existing distributed algorithm techniques were not directly applicable to WASNs. There are at least five major reasons: (1) WASNs are intrinsically related to the physical and geometric world and therefore have very special properties — the uses of local and geographic information, for example, play key roles in designing efficient, robust, and scalable sensor networks; (2) relative communication costs are much higher than they were assumed to be in all previous distributed computing research — since WASN nodes are severely energy constrained, the cost of communication becomes an extremely important factor in the design of WASN software; (3) accuracy of physical measurements is intrinsically limited and therefore there is little advantage on insisting on completely accurate results; (4) energy consumption is a critical system constraint; and (5) data acquisition is naturally distributed and error-prone, implying a strong need for new sensing, computation, and communication models.

The relative communication delay in sensor networks is significantly larger than in traditional computational systems. It is interesting to note that in modern deep submicrometer (DSM) chip designs, delay on a single system-on-chip will be up to 20 clock cycles. However, even the fastest communication protocols in WASNs will have delays in millions of cycles due to technological and physical limitations as well as system software overhead. Furthermore, communication generally dominates both sensing and computation in terms of energy (currently, image and video sensors are exceptions). Again, it is interesting to draw parallels with DSM designs: In DSM, communication will also dominate power consumption, maybe eventually by as high as a 10 : 1 ratio with respect to computations. In WASNs, technology trends are much more difficult to predict, yet at least in current and pending technologies, this ratio is much higher, often estimated at 1000 : 1.

Interestingly, several new hardware and architectural characteristics have also come into play that strongly influence WASN communication costs. For example, we have already mentioned that in many of the current low-power radios used in WASN nodes, the power requirements for listening or receiving messages is

about the same as when transmitting. This is in sharp contrast to the assumptions made in numerous wireless communication research efforts in the past, where transmitting a message was almost always assumed to have required much more power than listening or receiving a message. Consequently, in order to be truly effective, WASN system software must try to maximize the duration of the times when the communication subsystems can be turned off or placed in sleep modes, thus saving precious reserve energy resources.

In addition to placing nodes in sleep modes to conserve energy, one can expect that fault tolerance and autonomous operation will be essentially mandatory for large scale WASNs, due to wide-scale deployment and the relatively high cost of servicing nodes. During the useful lifetime of a typical WASN, it is not unreasonable to expect that at least some nodes will exhaust their energy supply. Even if latency (real-time constraints), energy consumption, and fault tolerance were not an issue, security and privacy issues would very often mandate that only a subset of nodes participate in a specific task. In addition, sensor nodes are often deployed outside strictly controlled environments, communicate using wireless (insecure) media, and hence are highly susceptible to security attacks. This further indicates that expecting all nodes to always be able to sense, communicate, and compute is not realistic. Moreover, as WASNs evolve into an Internet-like scale and organization and span the whole earth and beyond, the only realistic possibility for all tasks will be execution in highly localized scenarios. In localized computation models, only a subset of nodes, which are almost always within geographic proximity, collaborate and participate in formulating results to specific application tasks.

The challenges outlined above can be classified into three major categories: (1) strict constraints; (2) new modes of operation; and (3) interface between physical world, computation, and information theory. The "strict constraint" challenges include problems related to the need for low cost, long life, and reliable infrastructures. Low-power operation, wireless bandwidth efficiency, reliability, fault tolerance, high availability, error recovery, distributed synchronization, and real-time operation in unpredictable environments are all important factors that influence the design decisions at this step. In this direction, the current key problem is learning how to scale the already available techniques to the next levels of strictness of constraints.

There are two main research direction related to the "new modes of operation" of WSNs, due to their distributed and multihop natures: localized algorithms and autonomous continuous operation. *Localized algorithms* are algorithms implemented on sensor networks in such a way that only a limited number of nodes communicate, therefore reducing overall energy consumption and bandwidth requirements. Consequently, localized algorithms often operate with incomplete information, noisy data, and almost always under very strict communication and energy constraints.

One way of modeling localized algorithms in WASNs is as follows: One or more nodes initiate a request for a computation (a query). The result of the query is to be sent to one or more sink nodes. Each node can obtain its required information either through its sensors or by communication with neighboring nodes. The goal is to maximize an objective function for the optimization task at hand, in such a way that all constraints are satisfied and the communication cost is minimal. The first and most important difference between the localized algorithm and other traditional methods is the amount of information available to the processing units. In conventional scenarios, the processors have all the information that is needed for their computation tasks. However, in localized approaches, the required information is not complete and thus the communication between components should be interleaved with the computations in different parts so that they compensate for the insufficient information. The other interesting aspect of localized procedures is that although there are many processing units in a pervasive computing environment such as WASNs, for most of the applications, only a few processors are sufficient to carry out the required calculations. This is in contrast to the classical distributed computing paradigms where all processors involved in a computation are actively computing all the time. For centralized algorithms, of course, one processing unit handles all the computations and control. In addition to the localized nature of the optimization algorithms in WASNs, autonomous closed-loop modes of operation are a must for effective use of such networks. Essentially, the applications must execute with minimal or no intervention of a human operator.

Traditional wired and wireless computer communication network designers have typically followed (although often loosely) the International Organization for Standardization (ISO[1]) Open System Interconnection (OSI) Reference Model as the basis for their protocol stack design. The OSI Reference Model specifies seven protocol layers: physical, data link, network, transport, session, presentation, and application [3]. The following subsections briefly describe two main WASN protocol stack layer functions, namely, medium access control (MAC) and routing, which are equivalent to what the OSI model classifies as data-link- and network-layer functions, respectively. The subsequent sections then describe sensor network specific tasks and problems such as location discovery and coverage.

3.1. Medium Access Control (MAC)

Wireless communication media are almost always broadcast in nature and thus are shared among the participants. For example, RF transmissions of one node can be "heard" by any other node that is within communication range. If two nodes that are close together transmit at the same time, their transmissions will most likely "collide" and interfere with each other. *Medium access*

[1] ISO is not an acronym. ISO is an international standardization organization with members from more than 75 countries.

control refers to the process by which nodes determine when and how to utilize a shared communication medium. In WASNs where network communications are multihop (often require intermediate nodes to forward packets), the MAC layer is also where specific self-organization and autonomous configuration abilities can be introduced into the network.

Traditional MAC designs have followed two distinct philosophies: dedicated and contention-based. In the *dedicated* scenarios, each node receives the shared resource according to a prespecified scheme. Time-division multiple access (TDMA) is one such scheme where each node may only transmit within a small, periodic, time slot. Such MAC strategies are typically not well suited for ad hoc networks that have no predefined organization and can be very dynamic in nature; that is, nodes can join, move, or leave the network at any time. In *contention-based* schemes, nodes attempt to "grab" the medium and transmit when needed. Often, nodes have abilities to sense that a channel is in use and thus determine that they must wait. References 4 and 5 provide an overview of existing techniques and propose new MAC layer schemes that are designed specifically for WASNs.

3.2. Ad Hoc Routing

Routing refers to the process of finding ways to deliver a message from a source to its destination. In ad hoc, multihop networking scenarios, routing is an especially difficult problem since the nodes must discover the destination and the routes to the destinations subject to extreme energy consumption limitations. Existing works from ad hoc wireless networking domains provide a solid foundation for WASN routing problems. However, WASNs have unique features that make traditional routing philosophies less relevant. In traditional wired and wireless data communication networks, connections are peer-to-peer. This means that the user at a specific source node must send data (usually in forms of packets) to another user at a specific destination. Consequently, the endpoints of communication typically have unique names and specific communications are identified by the source and destination names (addresses). In WASNs, however, such peer-to-peer communications are less meaningful. Typically, nodes that sense events, analyze the data, collaborate with neighbors, and communicate processed information to one or more sink nodes. In addition, the information may be processed further along the path to the destination which makes the definition of "routing" very vague in WASNs compared to traditional data communication networks.

"Flooding" is a well-known basic scheme that can be used for routing in any network. During flooding, each intermediate node that receives a packet simply forwards it to all its neighbors until it reaches the destination. In a connected network, the packet will most likely reach the destination, although packet losses due to interference and other transmission errors are always possible. Although for broadcast messages flooding is very effective, the overhead for point-to-point communication is extremely high. Other more sophisticated approaches have been proposed such as dynamic source routing (DSR) and ad hoc on-demand distance vector (AODV) routing, which try to discover routes and maintain information about the network topology to eliminate flooding overheads. Reference 6 provides an overview and detailed analysis of several ad hoc network routing protocols. However, as stated above, routing schemes from ad hoc networking do not necessarily work well in sensor networks.

Several schemes have been proposed for routing in WASNs that leverage on sensor network specific characteristics such as geographic information and application requirements. Because of the immaturity of the field, none have established themselves as definitive solutions to WASN routing. Directed diffusion is one example of a generic scheme for managing the data communication requirements (and thus routing) in WASNs. The basic scheme in directed diffusion proposes the naming of data as opposed to naming sources and destinations of data. Data are "named" using attribute–value pairs. Data are requested by name as "interests" in the network. The request (dissemination) sets up "gradients" so that the named data (or events) can be "drawn." In traditional IP-style communication, nodes are identified as "endpoints" and the communication is layered as an end-to-end service. In directed diffusion, in contrast, named data flow toward the originators of their corresponding interests along multiple paths with the network "reinforcing" one or multiple such paths [7]. However, as stated above, WASN nodes may process the data at intermediate steps and the specific routing solution may be tightly coupled with application tasks (as opposed to layered).

3.3. Location Discovery

Geographic information is an integral attribute of any physical measurement. Thus, the knowledge of node locations is fundamental in proper operations of sensor networks, especially for WASNs. The ad hoc nature of WASN deployment necessitates that each node determine its location through a location discovery process. The Global Positioning System (GPS) is one method that was designed and is controlled by the United States Department of Defense for this purpose. The GPS system consists of at least 24 satellites in orbit around the earth, with at least four satellites viewable from any point, at a given time, on earth. They each broadcast time-stamped messages at periodic intervals. Any device that can hear the messages from four or more satellites can estimate its distance from each satellite and thus perform trilateration to compute its position.

Although GPS is an elegant solution to the location discovery process, it has several limitations that hinder its use in WASN applications: (1) GPS is costly in terms of both hardware and power requirements and (2) it requires line-of-sight communication between the receiver and the satellites and thus does not work well when obstructions such as buildings, trees, and mountains block the direct "view" to the satellites. Thus, other techniques have been proposed to dynamically compute the locations of the nodes in WASNs. In several location

discovery schemes, the received signal strength indicator (RSSI) of RF communication is used as a measure of distance between nodes. In other schemes, the time difference in arrival of RF and acoustic (ultrasound) signals are used to approximate node distances. Once nodes in a WASN have the ability to estimate distances between each other (ranging), they can then compute their locations using the simple trilateration method. In order for a trilateration to be successful, a node must have at least three neighbors who already know their locations. This requires that at least a subset of nodes determine their locations through other means such as by using GPS, manual programming, or deterministic deployment (placing nodes at specified coordinates). References 8 and 9 provide detailed discussions on location discovery techniques and algorithms.

3.4. Coverage

Several different coverage formulations arise naturally in many domains. The "art gallery problem," for example, deals with determining the number of observers necessary to cover an art gallery room such that every point is seen by at least one observer. This problem has several applications such as for optimal antenna placement problems in wireless communication. The art gallery problem was solved optimally in two dimensions (2D) [10] and was shown to be computationally intractable in the 3D case. Coverage in the context of sensor networks can have very new semantics. The main question at the core of coverage is trying to answer how well the sensors observe a physical space. References 11 and 12 present several formulations of sensor coverage in sensor networks. The formulations include calculations based on best- and worst-case coverages for agents moving in a sensor field and exposure-based methods. In the best- and worst-case formulations, the distance of the agent to the closest sensors are of importance, while in exposure-based methods the detection probability (observability) in the sensor field is represented by a path-dependent integral of multiple sensor intensities. In both of these schemes, the types of actions that the agent performs impact the coverage metric. For example, the sensor field may have a different coverage level if an agent is traveling west to east as opposed to north to south, or along any other arbitrary paths. The actual physical characteristics and abilities of WASN nodes will play crucial roles in building practical, accurate, and useful coverage models and analysis algorithms.

4. FUTURE DIRECTIONS

We conclude by summarizing some important future challenges in wireless sensor networks:

QoS. For quality of service (QoS), one can define both syntactic and semantic interpretations. On the syntactic level, one can consider dimensions such as coverage, exposure, latency, measurement and communication errors, and event detection confidence. On the semantic level, one can define utility and cost functions to enable the analysis of how particular data can help in the construction of more accurate models of the physical world or more efficient algorithms.

Scaling. Scaling has been the key metric in analyzing both graph-theoretic and physical phenomena. The goal will be to develop new methods that are based on statistical techniques instead of traditional probabilistic ones. Existing techniques such as state transitions and percolation will be key factors in analyzing and building very large systems and optimizing their performance.

Profilers, Recommenders, and Search Engines. Profilers, recommenders, and search engines rapidly emerged as mandatory systems enabling efficient use of the World Wide Web (WWW) and the Internet. There are clear needs to develop such systems for sensor networks. New dimensions and challenges include ways to include information and knowledge, not just about physical location and physical time, but also about the physical, chemical, and biological worlds. There are needs for profilers of events, objects, areas, sensors, and users, among other things.

Foundations and Theory. There is a need to develop new theoretical foundations, new models, new algorithmic complexity theory and practice, new programming models, and languages for embedded sensor networks. For example, new models of sensor networks will encompass the already existing Markov models, interacting particle models (e.g., the Ising model), bifurcation-based models, fractals, oscillations, and space pattern models. In addition, there will be a need to create new models unique to wireless sensor networks. As another example, the VLSI (very large scale integration) theory field was built based on two lasting premises: (1) that integrated circuits are planar and (2) that features are of small size, yet limited in quantity. There is a need to explore such lasting features in sensor networks. Examples of such rule-based modeling are "energy spent on communication is dominant and distance-dependent," "all measurements have intrinsic errors," and "storage space on nodes is very limited."

Other potential research directions include validation and debugging, data compression and aggregation, real-time constraints, distributed scheduling and assignment, pricing, and privacy of actions.

BIOGRAPHIES

Seapahn Megerian (**Meguerdichian**) is currently a Ph.D. student in the Computer Science Department at the University of California, Los Angeles. He received his Computer Science and Engineering B.S. degree in 1998 and Computer Science M.S. degree in 1999 from UCLA. His primary focus is in the design and development of efficient algorithms for deployment,

performance, and coverage analysis; decision support; operation optimization; security; and privacy in wireless ad hoc sensor networks. In addition, his research includes high-performance communication systems, system-on-chip network design, application-specific compilers, and computational security. He was the recipient of the 7th Annual International Conference on Mobile Computing and Networking (MobiCom 2001) Best Student Paper Award for the paper titled "Exposure in Wireless Ad Hoc Sensor Networks."

Miodrag Potkonjak is a Professor in the Computer Science Department at the University of California, Los Angeles. He received his Ph.D. degree in Electrical Engineering and Computer Science from University of California, Berkeley in 1991. In 1991, he joined C&C Research Laboratories, NEC USA in Princeton, New Jersey. Since 1995, he has been with UCLA. He has received the NSF CAREER award, OKAWA foundation award, UCLA TRW SEAS Excellence in Teaching Award, and a number of best-paper awards. His research intersets include complex distributed systems, communication system design, embedded systems, computational security, practical optimization techniques, and intellectual property protection.

BIBLIOGRAPHY

1. J. M. Rabaey et al., PicoRadio supports ad hoc ultra-low power wireless networking, *Computer* **33**: 42–48 (July 2000).
2. A. Mason et al., A generic multielement microsystem for portable wireless applications, *Proc. IEEE* **86**: 1733–1745 (Aug. 1998).
3. *http://webopedia.internet.com/quick_ref/OSI_Layers.html*.
4. W. Ye, J. Heidemann, and D. Estrin, An energy-efficient MAC protocol for wireless sensor networks, *IEEE Infocom* (in press).
5. A. Woo and D. Culler, A transmission control scheme for media access in sensor networks, *Proc. ACM/IEEE Int. Conf. Mobile Computing and Networking* (*MobiCOM 2001*).
6. J. Broch et al., A performance comparision of multi-hop wireless ad hoc network routing protocols, *Proc. ACM/IEEE Int. Conf. Mobile Computing and Networking*, Oct. 1998, pp. 85–97.
7. C. Intanagonwiwat, R. Govindan, and D. Estrin, Directed diffusion: A scalable and robust communication paradigm for sensor networks. *Proc. 6th Annual Int. Conf. Mobile Computing and Networking* (*MobiCOM 2000*), 2000, pp. 56–67.
8. P. Bahl and V. N. Padmanbhan, RADAR: An in-building RF-based user location and tracking system, *Proc. IEEE Infocom 2000*, April 2000, pp. 775–784.
9. A. Savvides, C. C. Han, and M. B. Srivastava, Dynamic fine-grained localization in ad-hoc networks of sensors, *Proc. 7th Annual Int. Conf. Mobile Computing and Networking* (*MobiCOM 2001*), July 2001.
10. J. O'Rourke, Computational geometry column 15 (open problem from art gallery solved), *Int. J. Comput. Geom. Appl.* **2**: 215–217 (June 1992).
11. S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. Srivastava, Coverage problems in wireless ad-hoc sensor networks, *IEEE Infocom 2001* **3**: 1380–1387 (April 2001).
12. S. Meguerdichian, F. Koushanfar, G. Qu, and M. Potkonjak, Exposure in wireless ad-hoc sensor networks, *Proc. 7th Annual Int. Conf. Mobile Computing and Networking* (*MobiCOM 2001*), July 2001, pp. 139–150.

INDEX

Note: Boldface numbers indicate illustrations and tables.